دانشگاه علوم پزشکی مشهد
معاونت پژوهش و فناوری
۵۵۳

مبانی مصوّرسازی داده‌ها

اصول اولیه برای ترسیم نمودارهای گویا و گیرا

کلاوس او. ویلکه

مترجمین

مجید خادم رضاییان
دانشیار پزشکی اجتماعی دانشگاه علوم پزشکی مشهد

محمد معصوم وند
دانشجوی کارشناسی ارشد تغذیه، دانشگاه علوم پزشکی مشهد


بسم الله الرحمن الرحيم

سرشناسه	:	ویلکه، کلاوس، ۱۹۷۲– م. Wilke, C.(Claus), 1972–
عنوان و نام پدیدآور	:	مبانی مصورسازی داده‌ها؛ اصول اولیه برای ترسیم نمودارهای گویا و گیرا/ [کلاوس ویلکه]؛ مترجم مجید خادم رضائیان، محمد معصوم‌وند.
مشخصات نشر	:	مشهد: دانشگاه علوم پزشکی و خدمات بهداشتی درمانی مشهد، انتشارات، ۱۴۰۲.
مشخصات ظاهری	:	۳۸۲ ص.: نمودار.
فروست	:	انتشارات دانشگاه علوم پزشکی مشهد؛ ۵۵۳.
شابک	:	۹۷۸–۶۰۰–۳۶۹–۱۹۲–۶
وضعیت فهرست‌نویسی	:	فیپا
یادداشت	:	عنوان اصلی: Fundamentals of data visualization : a primer on making informative and compelling figures, 2019.
یادداشت	:	کتابنامه: ص.[۳۷۹] – ۳۸۱.
موضوع	:	مصورسازی اطلاعات Information visualization تحلیل بصری Visual analytics تجسم -- داده‌پردازی Visualization -- Data processing
شناسه افزوده	:	خادم رضائیان، مجید، ۱۳۶۵ -، مترجم
شناسه افزوده	:	معصوم‌وند، محمد، ۱۳۷۶-، مترجم
شناسه افزوده	:	دانشگاه علوم پزشکی و خدمات بهداشتی درمانی مشهد. انتشارات
رده بندی کنگره	:	۹۱۵۲/QA۷۶
رده بندی دیویی	:	۴۲۲۶۲۳/۰۰۱
شماره کتابشناسی ملی	:	۹۵۴۸۴۳۵

انتشارات دانشگاه علوم پزشکی مشهد

مبانی مصورسازی داده‌ها:
اصول اولیه برای ترسیم نمودارهای گویا و گیرا

شماره اثر ۵۵۳

مترجمان	:	دکتر مجید خادم رضائیان، محمد معصوم وند
طراح جلد	:	الهه نوری
نوبت چاپ	:	اول، بهار ۱۴۰۳
قطع	:	وزیری، ۳۸۲ صفحه
شابک	:	۹۷۸-۶۰۰-۳۶۹-۱۹۲-۶

دفتر انتشارات :
مشهد، خیابان دانشگاه، دانشگاه ۱۸،
دانشکده بهداشت
واحد کتب و نشریات دانشگاه
ص.پ | ۹۱۳۷۶۷۳۱۱۹
تلفن | ۰۵۱-۳۸۵۱۴۳۰۰
فکس | ۰۵۱-۳۸۵۱۳۴۰۰
http://pub.mums.ac.ir
pub@mums.ac.ir

مبانی مصوّرسازی داده‌ها

اصول اولیه برای ترسیم نمودارهای گویا و گیرا

دکتر مجید خادم رضائیان
محمد معصوم وند

یادداشت ناشر

بسمه تعالی

انتشارات دانشگاه علوم پزشکی مشهد به عنوان یک نهاد علمی – فرهنگی، چاپ و نشرِ کتاب‌های مورد استفاده استادان و دانشجویان رشته‌های علوم پزشکی را از وظایف اساسی خود می‌داند و بر آن است تا با انتشار متون مربوط به علوم پزشکی، هم بر اعتبار و اعتلای آن‌ها بیفزاید و هم موجبات دل‌گرمی و شوق و نشاط ایشان در امر تحقیق را فراهم آورد. به یقین چاپ و نشر آثار معتبر علمی، امکان دست‌یابی اهل تحقیق به منابع مختلف و گوناگون علمی مورد نیاز به بهترین صورت ممکن خواهد ساخت و راه را برای تحقیقات بعدی هموار خواهد کرد.

گستردگی فعالیت‌های ناشران خصوصی، نه تنها از مسئولیت مراکز معتبر دانشگاهی برای کوشش در این زمینه‌ها با تکیه بر جنبه‌های علمی و فرهنگی موضوع نمی‌کاهد، بلکه به دلایل گوناگون از جمله لزوم بررسی‌های دقیق کارشناسی، ویرایش‌های علمی و فنی و نیز توجه بیشتر به توان مالی مخاطبان آثار به ویژه دانشجویان، باید دانشگاه‌های دولتی دامنه خدمات خود را با تکیه بر جنبه‌های کیفی و گاه صرف نظر از بازده اقتصادی گسترش دهند.

مسئولان انتشارات دانشگاه علوم پزشکی مشهد با اشتیاق و اخلاص تمام می‌خواهند جدیدترین آثار مربوط به دانش‌های پزشکی را به بهترین شیوه فراهم آورند و می‌کوشند تا آثار منتشر شده با کیفیت مطلوب عرضه شود. در عین حال انتشارات مترصد دریافت نظرها و انتقادهای به جا و سازنده اهل نظر است، تا از این طریق بتواند بر کمال آثار منتشره شده خویش بیفزاید.

معاونت پژوهش و فناوری دانشگاه علوم پزشکی مشهد با استعانت از الطاف خداوندی امیدوار است که نشر این اثر با نظارت همه جانبه این معاونت و با تکیه بر ابتکار و کوشش اعضا و کارشناس دانش‌آموخته خویش، همانند دیگر آثار این انتشارات، موجب ارتباط محکم و پایدار با جامعه پزشکی و دانشجویان پژوهشگر در جهت ارتقای کیفی و تنوع موضوعات مورد نیاز دانش‌پژوهان رشته‌های علوم پزشکی شود و مورد استفاده دانشجویان و مقبول طبع صاحب نظران و دانشگاهیان نکته‌بین قرار بگیرد.

انتشارات دانشگاه علوم پزشکی مشهد

فهرست مطالب

مقدمه مترجم

به ترجمه کتاب "Fundamentals of Data Visualization" خوش آمدید! همانطور که عصر دیجیتال به تکامل خود ادامه می‌دهد، اهمیت درک و برقراری ارتباط موثر با داده‌ها از طریق مصوّرسازی بسیار مهم شده است. در این کتاب، به اصول و شیوه‌های اصلی مصوّرسازی داده ها پرداخته شده است و راهنمای جامعی را هم برای مبتدیان و هم برای پژوهشگران باتجربه ارائه می‌دهد.

توصیف داده‌ها یکی از مراحل مهم در تحلیل داده‌ها و علم آمار است. این مرحله به ما کمک می‌کند تا با داده‌های خود آشنا شویم و با استفاده از آن‌ها تحلیل‌های دقیق‌تری انجام دهیم. بررسی توزیع داده‌ها، شناسایی پارامترهای مختلف، اعتبارسنجی داده‌ها و ارائه گزارش و توضیحات از دلایل اهمیّت توصیف داده‌ها در علم آمار به‌شمار می‌روند. نوع و طرز توصیف داده‌ها هم از اهمیت بالایی برخوردار است و باید با موضوع مورد پژوهش انطباق داشته باشد. آن‌ها باید به گونه ای باشند که قابل فهم و استفاده باشند.

در عصری که داده‌ها تصمیم‌گیری را در زمینه‌های مختلف هدایت می‌کنند، توانایی ارائه داده‌ها به روشی معنادار یک مهارت ارزشمند است. هدف این کتاب تجهیز خوانندگان با دانش و ابزار لازم برای ایجاد نمایش‌های بصری «گویا و گیرا» است که درک و بینش را تسهیل می‌کند. ما امیدواریم از طریق این ترجمه بتوانیم محتوا را برای مخاطبان گسترده‌تری در دسترس قرار دهیم تا پژوهشگران بیشتری بتوانند درک خود را از نحوه صحیح مصوّرسازی داده‌ها و کاربرد آن در زمینه‌های مختلف گسترش دهند.

در ایـن سـفر بـه مـا بپیوندیـد تـا مفـاهیم اساسـی، بهتـرین شـیوه‌ها و تکنیک‌هـای عملـی را کـه زیربنای مصوّرسـازی مـؤثر داده‌هـا هسـتند، بررسـی کنـیم. چـه علاقه‌منـد بـه داده هسـتید، چـه یـک دانش‌آموز یا دانشجو، چه یـک حرفـه‌ای یـا هـر کسـی کـه بـه قـدرت داسـتان‌گویی بصـری از طریـق داده‌ها علاقه دارد، این کتاب برای الهام بخشیدن و توانمندسازی شما طراحی شده است.

پـس از چنـدین سـال فعالیـت دانشـگاهی، متوجـه شـدیم کـه کتـابی جـامع و کامـل در زمینـه مصوّرسازی داده‌ها به زبـان فارسـی بـه‌نـدرت یافـت مـی‌شـود. بـه خـاطر همـین موضـوع مـا تصـمیم گـرفتیم یکـی از جـامع‌ترین منـابع موجـود کـه بـه بررسـی همـه جانبـه اصـول توصـیف داده‌هـا بـا نگـاهی کـاربردی مـی‌پـردازد را ترجمـه کنـیم و در اختیـار شـما عزیـزان قـرار دهـیم. هرچنـد بـا بازنگری و ویرایش‌های متعدد تـلاش شـده اسـت کـه مجموعـه حاضـر تـا حـد امکـان عـاری از خطـا باشد، با این حـال مشـتاقانه منتظـر دریافـت نقطـه نظـرات و بازخوردهـای ارزشـمند خواننـدگان عزیـز (پست الکترونیک: KhademRM@mums.ac.ir) هستیم.

دکتر مجید خادم رضاییان، محمد معصوم‌وند

زمستان ۱۴۰۲

پیشگفتار

اگر شما محقق، تحلیلگر، مشاور یا هر مسئولیت دیگری دارید که باید اسناد یا گزارش‌های فنّی را تهیه کند، یکی از مهم‌ترین مهارت‌هایی که باید داشته باشید، توانایی نمایش داده‌ها به شیوه‌ای قانع‌کننده و عمدتاً به صورت اشکال می‌باشد. اشکال معمولاً بار استدلال‌های شما را به دوش می‌کشند. آن‌ها باید واضح، جذّاب و متقاعدکننده باشند. تفاوت بین اشکال خوب و بد می‌تواند تفاوت بین یک مقالهٔ بسیار تأثیرگذار یا مبهم، برنده شدن یا از دست دادن کمک مالی یا قرارداد، یک مصاحبهٔ شغلی خوب یا ضعیف باشد. با این حال، منابع کمی وجود دارد که به شما آموزش می‌دهد چگونه ترسیم‌های قانع‌کننده‌ای داشته باشید. تعداد کمی از دانشگاه‌ها دوره‌هایی در این زمینه ارائه می‌دهند و کتاب‌های زیادی نیز در این زمینه وجود ندارد (البته بدیهی است که منابع محدودی وجود دارند). آموزش‌های نرم‌افزارهای مربوط به نمایش داده‌ها معمولاً بر چگونگی دستیابی به جلوه‌های بصری خاص تمرکز می‌کنند تا اینکه توضیح دهند چرا انتخاب‌های خاصی ارجح هستند. در کار روزانه، از شما انتظار می‌رود که بدانید چگونه اشکال خوب ترسیم کنید، و اگر خوش شانس باشید، یک مشاور صبور دارید که در حین نوشتن اولین مقالهٔ علمی‌تان، چند ترفند را به شما آموزش می‌دهد.

در زمینهٔ نوشتن، ویراستاران با تجربه درباره «گوش» صحبت می‌کنند، یعنی توانایی شنیدن (درونی، در حین خواندن یک متن) که آیا نوشته خوب است یا خیر. وقتی صحبت از شکل‌ها و سایر ترسیم‌ها به میان می‌آید، به طور مشابه به «چشم» نیاز داریم، یعنی توانایی نگاه کردن به یک شکل و دیدن اینکه آیا متعادل، واضح و قانع‌کننده است یا خیر. درست مانند نوشتن، توانایی دیدن اینکه آیا یک شکل مناسب است یا خیر را می‌توان آموخت. داشتن چشم در

درجهٔ اوّل به این معنی است که شما از مجموعهٔ بزرگتری از قوانین ساده و اصول ترسیم خوب آگاه هستید و به جزئیات کوچکی توجه می‌کنید که دیگران ممکن است به آن‌ها بی‌توجه باشند.

بر اساس تجربه، باز هم مانند نوشتن، با خواندن یک کتاب در آخر هفته، «چشم»هایتان رشد نمی‌کند. این یک فرآیند مادام‌العمر است و مفاهیمی که امروز برای شما بسیار پیچیده یا بسیار ظریف هستند، ممکن است پنج سال بعد بسیار منطقی‌تر باشند. من می‌توانم در مورد خودم بگویم که همچنان در حال تکامل درک خود از نحوه آماده‌سازی اشکال هستم. من معمولاً سعی می‌کنم خودم را در معرض رویکردهای جدید قرار دهم و به انتخاب‌های بصری و طراحی که دیگران در شکل‌هایشان بکار می‌برند، توجّه می‌کنم. من همچنین آماده تغییر نظرم هستم. ممکن است امروز شکلی را عالی بدانم، اما ماه آینده ممکن است دلیلی برای انتقاد از آن پیدا کنم. پس با در نظر گرفتن این موضوع، لطفاً هر چیزی را که می‌گویم به عنوان وحی الهی در نظر نگیرید. در مورد استدلال من برای انتخاب‌های خاص تفکّر نقّاد داشته باشید و تصمیم بگیرید که آیا می‌خواهید آن‌ها را بپذیرید یا خیر.

در حالی که مطالب این کتاب با یک توالی منطقی تنظیم شده است، اکثر فصل‌ها می‌توانند به طور مستقل مطالعه شوند و نیازی به خواندن کتاب از اوّل تا آخر نیست. راحت باشید و بخش خاصی که در حال حاضر به آن علاقه دارید، یا بخشی که طراحی خاص مدنظرتان را توضیح می‌دهد، انتخاب کنید. در واقع، اگر کتاب را یک‌باره نخوانید، بلکه آن را تکه‌تکه و در طول زمان طولانی‌تری خوانده و تلاش کنید برخی از مفاهیم فراگرفته شده را در طراحی شکل‌هایتان بکار بگیرید و بعداً برگردید تا در مورد مفاهیم دیگر مطالعه کنید یا بخش‌هایی از مفاهیمی را که مدتی قبل در مورد آن‌ها آموخته‌اید دوباره بخوانید، بیشترین بهره را از این کتاب خواهید برد. اگر بعد از گذشت چند ماه دوباره فصلی را بخوانید، ممکن است متوجّه شوید که همان فصل به شما چیزهای جدیدی می‌گوید.

اگرچه تقریباً تمام شکل‌های این کتاب با R و ggplot2 ترسیم شده‌اند، اما ما این مجموعه را به عنوان یک کتاب R نمی‌دانیم. ما در این کتاب در مورد اصول کلی آماده‌سازی شکل صحبت می‌کنیم. نرم‌افزار مورد استفاده برای ساخت شکل‌ها انتخابی است. شما می‌توانید از هر نرم‌افزار ترسیمی که می‌خواهید برای تولید انواع شکل‌هایی که در اینجا نشان می‌دهیم استفاده کنید. با این حال، ggplot2 و بسته‌های مشابه، بسیاری از روش‌هایی را که استفاده می‌کنیم بسیار ساده‌تر از سایر نرم‌افزارها رسم می‌کنند. نکتهٔ مهم این است که چون این کتاب R

نیست، در هیچ جای این کتاب درباره کد یا روش‌های برنامه‌نویسی بحث نمی‌کنیم. ما می‌خواهیم شما روی مفاهیم و شکل‌ها تمرکز کنید، نه روی کد. اگر کنجکاو هستید که چگونه هر یک از شکل‌ها ساخته شده است، می‌توانید کد منبع کتاب را در مخزن GitHub آن بررسی کنید.

نرم‌افزار و فرایند آماده‌سازی شکل

من بیش از دو دهه تجربه در مورد تهیه اشکال برای انتشارات علمی دارم و هزاران شکل ساخته‌ام. اگر در این دو دهه یک مساله ثابت وجود داشته است، آن مساله، تغییر در فرایند آماده‌سازی شکل بوده است. هر چند سال یک بار، یک کتابخانه جدید ایجاد شده یا یک الگوی جدید پدید می‌آید، و گروه‌های بزرگی از دانشمندان به سمت استفاده از ابزار جدید می‌روند. من با استفاده از gnuplot، Xfig، Mathematica، Matlab، matplotlib در پایتون، پایه R، ggplot2 در R و احتمالاً موارد دیگری که در حال حاضر نمی‌توانم به یاد بیاورم، شکل‌هایی ساخته‌ام. رویکرد ترجیحی فعلی من ggplot2 در R است، اما من انتظار ندارم که تا زمانی که بازنشسته شوم به استفاده از آن ادامه دهم.

این تغییر مداوم در بن‌سازه‌های نرم‌افزاری یکی از دلایل کلیدی است که چرا این کتاب یک کتاب برنامه‌نویسی نیست و چرا تمام نمونه‌های کد حذف شده‌اند. ما می‌خواهیم این کتاب صرف‌نظر از اینکه از کدام نرم‌افزار استفاده می‌کنید برای شما مفید باشد، و حتی زمانی که همه ggplot2 را کنار گذاشته‌اند و از نرم‌افزار جدیدی استفاده می‌کنند، ارزشمند باقی بماند. این انتخاب ممکن است برای برخی از کاربران ggplot2 که مایلند بدانند چگونه یک شکل معیّن را ساخته‌ام، ناامیدکننده باشد. با این حال، هر کسی که در مورد روش‌های کدنویسی من کنجکاو است، می‌تواند کد منبع کتاب را بخواند. همچنین، در آینده ممکن است یک سند تکمیلی را منتشر کنیم که فقط بر روی کد متمرکز است.

یکی از چیزهایی که در این سال‌ها آموخته‌ام این است که خودکارسازی دوست شماست. من فکر می‌کنم اشکال باید به‌عنوان بخشی از فرایند تجزیه و تحلیل داده‌ها (که باید خودکار نیز باشند) تولید شوند و آماده ارسال به چاپگر باشند، و نیازی به پردازش دستی پس از آن نداشته باشند. بسیاری از کارآموزان را می‌بینم که پیش‌نویس‌های ناموزونی از شکل‌های خود را به طور خودکار تولید می‌کنند، و سپس آن‌ها را برای اصلاح به نرم‌افزار ایلاستریتور وارد می‌کنند. دلایل مختلفی وجود دارد که چرا این ایده بد است. اوّل، لحظه‌ای که شما به صورت دستی

یک شکل را ویرایش می‌کنید، شکل نهایی شما قابلیّت بازتولید نخواهد داشت. شخص دیگری نمی‌تواند دقیقاً همان شکلی را که شما رسم کردید، ایجاد کند. در حالی که اگر تنها کاری که انجام می‌دهید تغییر قلم برچسب‌های محور باشد، ممکن است این اقدام اهمیّت چندانی نداشته باشد، با این حال مرزها محو بوده و به راحتی می‌توان به منطقه‌ای رفت که همه چیز واضح نیست. به عنوان مثال، فرض کنید می‌خواهید به‌صورت دستی برچسب‌های عناوین را با برچسب‌های قابل خواندن‌تر جایگزین کنید. ممکن است شخص دیگری نتواند تأیید کند که جایگزینی برچسب مناسب بوده است یا خیر. دوّم، اگر پردازش‌های پسین دستی زیادی را به فرایند آماده‌سازی شکل خود اضافه کنید، در آن صورت تمایلی به ایجاد تغییر جدید یا بازگرداندن اصلاحات نخواهید داشت. بنابراین، ممکن است درخواست‌های معقول برای اعمال تغییرات که توسط همکاران یا دوستان ارائه شده است را نادیده بگیرید، یا ممکن است وسوسه شوید که از یک شکل قدیمی مجدداً استفاده کنید، حتی اگر واقعاً همه داده‌ها را بازسازی کرده باشید. سوّم، ممکن است حتی خودتان فراموش کنید که دقیقاً برای تهیّه یک شکل معیّن چه کاری انجام داده‌اید، یا ممکن است نتوانید در آینده شکلی برای داده‌های جدید ایجاد کنید که دقیقاً از نظر بصری با شکل قبلی شما مطابقت داشته باشد. این‌ها مثال‌های ساختگی نیستند و همهٔ آن‌ها در واقع اتفاق افتاده‌اند.

با توجه به همه دلایل ذکر شده، برنامه‌های تعاملی نمودار ایده بدی هستند. آن‌ها ذاتاً شما را مجبور می‌کنند که شکل‌های خود را به صورت دستی آماده کنید. در واقع، احتمالاً بهتر است پیش‌نویسی از شکل را به صورت خودکار تولید کنید و آن را در نرم‌افزار ایلاستریتور بسازید تا اینکه کل شکل را با دست در برخی از برنامه‌های تعاملی شکل بسازید. لطفاً توجه داشته باشید که اکسل نیز یک برنامه تعاملی شکل است و برای تهیه شکل (یا تجزیه و تحلیل داده‌ها) توصیه نمی‌شود.

یکی از مؤلفه‌های مهم در کتابی مرتبط با مصوّرسازی داده، امکان‌سنجی ترسیم‌های پیشنهادی است. ابداع نوع جدیدی از مصوّرسازی خوب است، امّا اگر کسی نتواند به راحتی با استفاده از این ترسیم‌ها اشکالی ایجاد کند، کاربرد زیادی نخواهد داشت. به عنوان مثال، زمانی که Tufte برای اولین بار خطوط جرقه[1] را پیشنهاد کرد، هیچ کس راه آسانی برای ساخت آن‌ها نداشت. در حالی که ما به رویاهایی نیاز داریم که جهان را به جلو می‌برند، ترجیح می‌دهیم که این کتاب عملی باشد و مستقیماً برای دانشمندانی که اشکال را برای انتشارات

1. sparklines

خـود آمـاده می‌کننـد، قابـل اسـتفاده باشـد. بنـابراین، ترسـیم‌هایی کـه در فصـل‌های بعـدی پیشـنهاد مـی‌کنیم را می‌تـوان بـا چنـد خـط کـد R از طریـق ggplot2 و بسـته‌های الحـاقی کـه بـه آسـانی در دسترس هستند، تولیـد کـرد. در واقـع، تقریبـاً هـر شـکلی در ایـن کتـاب، بـه اسـتثنای چنـد شـکل در فصـل‌های ۲۶، ۲۷ و ۲۸، دقیقـاً همـانطور کـه نشـان داده شـده اسـت، بـه طـور خودکـار تولیـد شـده است.

علایم راهنما

این علامت نشان دهندهٔ یک نکته یا توصیه است.

این علامت نشان دهندهٔ یک قاعده کلی است.

این علامت نشان دهندهٔ یک هشدار یا احتیاط است.

منابع تکمیلی مربوط به این کتاب از آدرس زیر قابل دریافت است:

https://github.com/clauswilke/dataviz

فصل ۱

مقدمه

نمایش داده‌ها ترکیبی از هنر و علم است. چالش موجود این است که هنر را به درستی به کار بریم بدون اینکه اشتباه علمی داشته باشیم و بالعکس. نمایش داده‌ها در درجهٔ اول باید به درستی داده‌ها را ارائه نموده و نباید گمراه کننده یا تحریف شده باشد. اگر یک عدد دو برابر عدد دیگر باشد، امّا در نمایش داده‌ها تقریباً یکسان به نظر برسد، شکل مربوطه نادرست است. در عین حال، نمایش داده‌ها باید از نظر زیبایی‌شناسی جذاب باشد. ارائهٔ مناسب داده‌ها از نظر بصری منجر به انتقال بهتر پیام موجود در داده‌ها می‌شود. اگر یک شکل رنگ‌های ناهمخوان، عناصر بصری نامتعادل، یا سایر ویژگی‌های پرت کنندهٔ حواس را داشته باشد، بررسی شکل و تفسیر صحیح آن برای بیننده دشوارتر خواهد بود.

بر اساس تجربه، دانشمندان اغلب (البته نه همیشه!) می‌دانند چگونه داده‌ها را بدون اینکه گمراه‌کننده باشند، نمایش دهند. با این حال، ممکن است حس زیبایی شناختی بصری به اندازهٔ کافی مدنظر قرار نگیرد و ممکن است ناخواسته دست به انتخاب‌هایی در رسم شکل‌ها بزنند که پیام مورد نظرشان را به خوبی منتقل نکند. از سوی دیگر، طراحان ممکن است شکل‌هایی را تهیه کنند که زیبا به نظر می‌رسند امّا ارتباط قوی و مستندی با داده‌ها ندارند. هدف این کتاب آن است که اطّلاعات مفیدی به هر دو گروه ارائه کند.

این کتاب تلاش می‌کند تا اصول، روش‌ها و مفاهیم کلیدی مورد نیاز برای نمایش داده‌های قابل انتشار، گزارش یا سخنرانی را پوشش دهد. از آنجایی که نمایش داده‌ها حوزه وسیعی است، و در گسترده‌ترین تعریف آن می‌تواند موضوعاتی مانند نقشه‌های فنی شماتیک، پویانمایی‌های سه بُعدی و رابط‌های کاربری را در بر گیرد، الزاماً بایستی دامنه موضوع محدود شود. در اینجا به طور خاص در خصوص نمودارهای ثابت که به صورت چاپی، برخط یا اسلاید ارائه می‌شوند، صحبت خواهد شد. این کتاب به جز در یک بخش کوتاه در فصل ۱۶، تصاویر یا فیلم‌های تعاملی را پوشش نمی‌دهد. بنابراین، در سرتاسر این کتاب، از واژه‌های «مصوّرسازی[1]» و «شکل[2]» به جای یکدیگر استفاده خواهیم کرد. همچنین این کتاب هیچ دستورالعملی در مورد چگونگی ساخت شکل با نرم‌افزارهای نمایش یا کتابخانه‌های برنامه‌نویسی موجود ارائه نمی‌دهد. کتابشناسی مشروح در پایان کتاب، متون مناسبی ارائه می‌دهد که این موضوعات را پوشش می‌دهند.

این کتاب به سه بخش تقسیم شده است. اوّلین بخش، «از داده‌ها تا نمودار[3]» انواع مختلف نمودارها مانند نمودار میله‌ای، نمودار پراکنش، و نمودار دایره‌ای را معرفی می‌کند. تاکید اولیه آن بر علم مصوّرسازی است. در این بخش، به جای تلاش برای ارائهٔ پوشش دایره‌المعارفی از تمام رویکردهای مصوّرسازی ممکن، مجموعه‌ای از نمودارهایی را که احتمالاً در مقالات منتشر شده و/یا در نگارش مقالهٔ خود به آن‌ها نیاز دارید، مورد بحث قرار می‌گیرد. در تنظیم این بخش، سعی کرده‌ایم نمودارها را به جای نوع داده‌ای که نمایش می‌دهند، بر اساس نوع پیامی که منتقل می‌کنند، گروه‌بندی کنیم. کتب مرجع آماری اغلب تجزیه و تحلیل و نمایش داده‌ها را بر اساس نوع داده‌ها ارائه می‌دهند، لذا مطالب بر اساس تعداد و نوع متغیّرها (یک متغیّر پیوسته، یک متغیّر گسسته، دو متغیّر پیوسته، یک متغیّر پیوسته و یک متغیّر گسسته و غیره) بیان می‌شوند. به نظر می‌رسد تنها متخصصین آمار این روش را مفید می‌دانند. سایر افراد به نمودار در چارچوب پیام منتقل شده می‌نگرند، مانند اینکه بزرگی یک کمیّت چقدر است، اجزای آن چیست، چگونه با کمیّت دیگری ارتباط دارد و ...

بخش دوّم، «اصول طراحی شکل[4]» مسائل مختلفی پیرامون طراحی که در هنگام مصوّرسازی داده‌ها با آن‌ها روبرو می‌شویم را مورد بحث قرار می‌دهد. تاکید اصلی امّا نه انحصاری این فصل، بر جنبهٔ زیبایی‌شناسی مصوّرسازی داده است. هنگامی که نمودار مناسب را برای

1. visualization
2. figure
3. From Data to Visualization
4. Principles of Figure Design

مجموعهٔ داده خود انتخاب کردیم، باید در مورد عناصر بصری، مانند رنگ‌ها، نمادها و اندازه قلم، انتخاب‌هایی مبتنی بر اصول زیبایی‌شناسی داشته باشیم. این انتخاب‌ها می‌توانند بر وضوح و زیبایی ظاهری نمودار تاثیر بسزایی داشته باشد. فصل‌های این بخش به رایج‌ترین مسائلی می‌پردازد که مکرراً در عمل با آن روبرو می‌شوید.

بخش سوم، «موضوعات متفرقه[1]» چند موضوع باقی مانده را پوشش می‌دهد که در دو بخش اول نمی‌گنجیدند. این بخش در مورد قالب فایل‌هایی که معمولاً برای ذخیره‌سازی تصاویر و نمودارها استفاده می‌شوند، نکاتی در مورد انتخاب نرم‌افزار مصوّرسازی، و نحوهٔ قرار دادن شکل‌های منفرد در بدنهٔ یک متن بزرگ‌تر توضیحاتی ارائه می‌دهند.

اشکال زشت، بد، و اشتباه

در سرتاسر این کتاب، عمدتاً نسخه‌های مختلفی از شکل‌های یکسان را نشان می‌دهیم، برخی به عنوان نمونه‌هایی از چگونگی ایجاد یک نمودار خوب و برخی برعکس. برای ارائهٔ یک دستورالعمل بصری ساده در مورد اینکه کدام نمونه مناسب است و از کدام نمونه باید اجتناب شود، شکل‌های مشکل‌دار را به‌صورت «زشت»، «بد» یا «اشتباه» برچسب زده‌ایم (شکل ۱–۱).

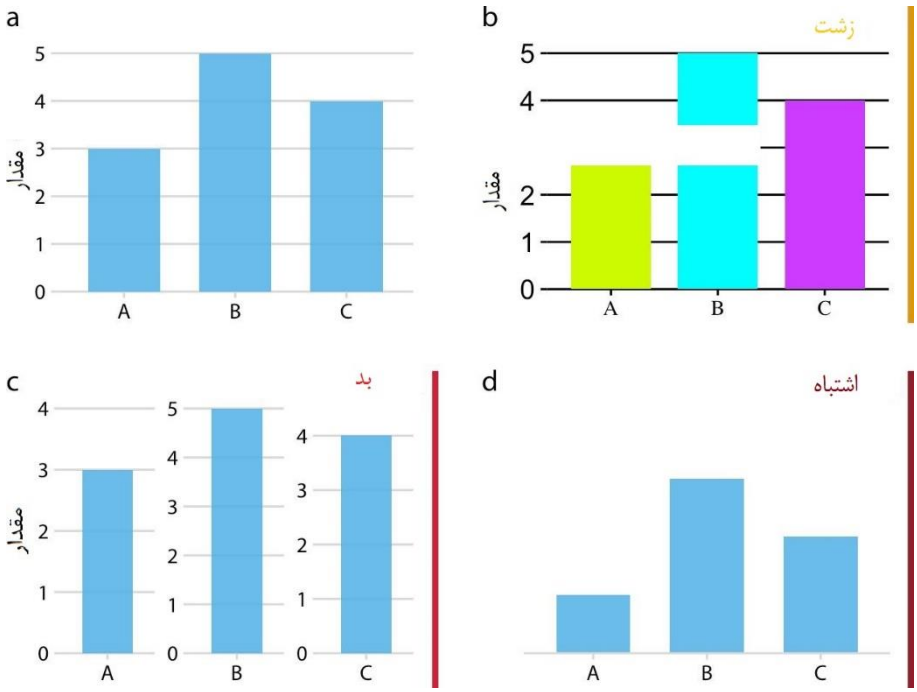
زشت: شکلی که مشکلات زیبایی‌شناسی دارد امّا با این حال واضح و آموزنده است.

بد: شکلی که درک آن مشکل است؛ ممکن است غیرواضح، گیج‌کننده، بیش از حد پیچیده یا فریبنده باشد.

اشتباه: شکلی که مشکلاتی در حوزهٔ محاسبات و ریاضی دارد، لذا به وضوح نادرست است.

ما شکل‌های خوب را برچسب نمی‌زنیم. هر شکلی که به عنوان ناقص برچسب‌گذاری نشده باشد، باید حداقل به عنوان شکلی قابل قبول تلقّی شود. این شکلی است که آموزنده خواهد بود، جذّاب به نظر می‌رسد و می‌تواند به همان صورت چاپ شود. توجّه داشته باشید که در بین شکل‌های خوب، همچنان تفاوت‌هایی در کیفیت وجود خواهد داشت و برخی از شکل‌های خوب بهتر از سایرین می‌باشند.

1. Miscellaneous Topics



شکل ۱–۱. نمونه‌هایی از شکل های زشت، بد و اشتباه. (الف) نمودار میله‌ای که سه مقدار را نشان می‌دهد (A = 3، B = 5 و C = 4). این یک نمودار معقول و بدون خطای عمده است. ب) نسخهٔ زشت قسمت (الف). اگرچه نمودار از نظر فنّی درست است، امّا از نظر زیبایی‌شناسی جذّاب نیست. رنگ‌ها خیلی روشن هستند و کاربردی نیز نیستند. شبکه پس زمینه بیش از حد برجسته است. متن با استفاده از سه قلم مختلف در سه اندازه مختلف نمایش داده شده است. ج) نسخهٔ بد قسمت (الف). هر میله با مقیاس محور y مخصوص به خود نشان داده شده است. از آنجایی که مقیاس‌ها هم تراز نیستند، شکل گمراه‌کننده می‌باشد. به راحتی می‌توان به اشتباه این تصوّر را داشت که این سه مقدار بیشتر از آنچه در واقعیّت هست، به هم نزدیک می‌باشند. (د) نسخهٔ اشتباه قسمت (الف). بدون یک مقیاس مشخّص در محور y، مقادیر مربوط به هر میله را نمی‌توان مشخّص کرد. به نظر می‌رسد میله‌ها دارای طول ۱، ۳، و ۲ هستند، در حالی که مقادیر واقعی ۳، ۵، و ۴ باشد.

ما عمدتاً منطق خود را برای رتبه‌بندی‌های خاص ارائه می‌کنیم، امّا برخی از آن‌ها سلیقه‌ای است. به طور کلی، رتبه‌بندی «زشت» انتزاعی‌تر از رتبه‌بندی «بد» یا «اشتباه» است. علاوه بر این، مرز بین «زشت» و «بد» تا حدودی سیّال است. گاهی اوقات انتخاب‌های ضعیف طراحی می‌تواند در ادراک انسان اختلال ایجاد کند تا جایی که رتبه‌بندی «بد» مناسب‌تر از رتبه‌بندی «زشت» باشد. در هر صورت، ما شما را تشویق می‌کنیم که قوهٔ قضاوت خود را تقویت و انتخاب‌های ما را ارزیابی نقّادانه کنید.

فصل ۲

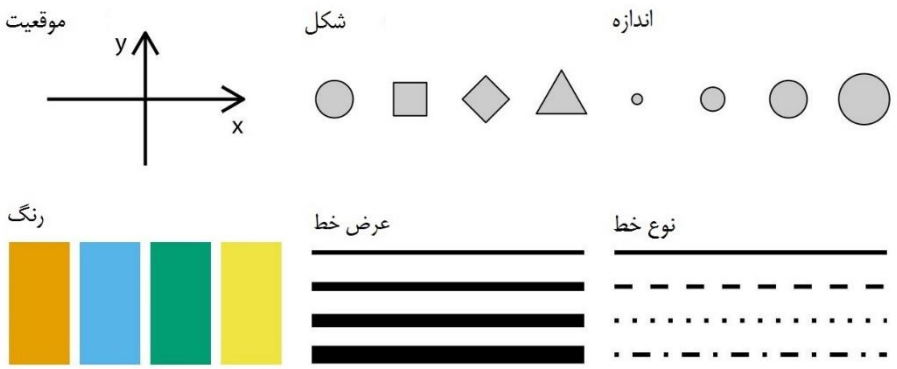
نمایش داده‌ها:
نگاشت داده‌ها بر زیبایی‌شناسی

زمانی که داده‌ها را نمایش می‌دهیم، مقادیر داده‌ها را به روشی نظام‌مند و منطقی به عناصر بصری که نمودار نهایی را تشکیل می‌دهند، تبدیل می‌کنیم. اگرچه انواع مختلفی از نمایش داده‌ها وجود دارد، و در نگاه اول به نظر نمی‌رسد که نمودار پراکنش، نمودار دایره‌ای و نقشه حرارتی نقاط مشترک زیادی با هم داشته باشند. همه این نمودارها را می‌توان با یک زبان مشترک توصیف کرد که نشان می‌دهد چگونه مقادیر داده‌ها به حباب‌های جوهر روی کاغذ یا پیکسل‌های رنگی روی صفحه تبدیل می‌شوند. نکته کلیدی این است: در تمام انواع مصوّرسازی داده‌ها، مقادیر مربوطه در قالب ویژگی‌های قابل اندازه‌گیری در شکل حاصل ارائه می‌شوند. ما از این ویژگی‌ها به عنوان زیبایی‌شناسی یاد می‌کنیم.

زیبایی‌شناسی و انواع داده‌ها

زیبایی‌شناسی تمام جنبه‌های یک عنصر گرافیکی معیّن را توصیف می‌کند. چند نمونه در شکل ۲-۱ ارائه شده است. یک جزء مهم هر عنصر گرافیکی، موقعیّت آن است، که محل استقرار عنصر مربوطه را نشان می‌دهد. در گرافیک‌های دو بُعدی استاندارد، موقعیّت‌ها را با مقدار x و y توصیف می‌کنیم، امّا سیستم‌های مختصّات دیگر و ترسیم نمودارهای یک یا سه بُعدی نیز امکان‌پذیر است. در مرحلهٔ بعد، تمام عناصر گرافیکی دارای شکل، اندازه و رنگ هستند. حتّی اگر در حال آماده‌سازی یک طراحی سیاه و سفید هستیم، عناصر گرافیکی باید رنگ داشته باشند تا قابل مشاهده باشند: برای مثال، سیاه اگر پس‌زمینه سفید است یا سفید

اگر پس‌زمینه سیاه است. در نهایت، در جایی که از خطوط برای مصوّرسازی داده‌ها استفاده می‌کنیم، این خطوط ممکن است دارای عرض‌های متفاوت یا الگوهای خط چین یا نقطه-نقطه باشند. فراتر از نمونه‌های نشان داده شده در شکل ۲-۱، بسیاری نکات زیبایی‌شناسی دیگر نیز وجود دارد که ممکن است در نمایش داده‌ها با آن‌ها مواجه شویم. برای مثال، اگر بخواهیم متنی را نمایش دهیم، ممکن است مجبور باشیم خانوادهٔ قلم، نوع قلم و اندازهٔ قلم را مشخّص کنیم، و اگر اجزای گرافیکی با هم همپوشانی داشته باشند، ممکن است مجبور باشیم میزان شفافیت هر یک را مشخّص کنیم.



شکل ۲-۱. انواع زیبایی‌شناسی‌های رایج در نمایش داده‌ها: موقعیّت، شکل، اندازه، رنگ، عرض خط، نوع خط. برخی از این زیبایی‌شناسی‌ها می‌توانند داده‌های پیوسته و گسسته (موقعیّت، اندازه، عرض خط، رنگ) را نشان دهند، در حالی که برخی دیگر معمولاً فقط می‌توانند داده‌های گسسته (شکل، نوع خط) را نشان دهند.

همهٔ زیبایی‌شناسی‌ها در یکی از این دو گروه قرار می‌گیرند: آن‌هایی که می‌توانند داده‌های پیوسته را نشان دهند و آن‌هایی که نمی‌توانند. داده‌های پیوسته، مقادیری هستند که بین هر دو مقدار، مقدار جدیدی قابل تصوّر است. به عنوان مثال، مدّت زمان یک مقدار پیوسته است. بین هر دو مدّت زمان، مثلاً ۵۰ ثانیه تا ۵۱ ثانیه، مقدارهای زیادی وجود دارد، مانند ۵۰٫۵ ثانیه، ۵۰٫۵۱ ثانیه، ۵۰٫۵۰۰۰۱ ثانیه و... در مقابل، تعداد افراد حاضر در یک اتاق یک مقدار گسسته است. یک اتاق می‌تواند ۵ نفر یا ۶ نفر را در خود جای دهد، امّا نه ۵٫۵ نفر. برای مثال‌های شکل ۲-۱، موقعیّت، اندازه، رنگ و عرض خط می‌توانند داده‌های پیوسته و شکل و نوع خط معمولاً فقط می‌توانند داده‌های گسسته را نشان دهند.

در مرحلهٔ بعد انواع داده‌هایی را که می‌توانیم در نمودار خود نشان دهیم، در نظر خواهیم گرفت. ممکن است داده‌ها را تنها به عنوان اعداد در نظر بگیرید، امّا مقادیر عددی تنها دو مورد از

انواع مختلف داده‌هایی هستند که ممکن است با آن‌ها روبرو شویم. علاوه بر مقادیر عددی پیوسته و گسسته، داده‌ها می‌توانند به صورت دسته‌بندی‌های گسسته، به صورت تاریخ یا زمان و به صورت متن باشند (جدول ۲–۱). وقتی داده‌ها عددی هستند آن را کمّی و زمانی که دسته‌بندی شده هستند آن‌ها را کیفی می‌نامیم. متغیّرهای مبتنی بر داده‌های کیفی عامل‌ها هستند و دسته‌های مختلف را سطوح می‌نامند. سطوح یک عامل معمولاً بدون ترتیب هستند (مانند سگ، گربه، ماهی در جدول ۲–۱)، امّا زمانی که ترتیب ذاتی بین سطوح عامل وجود داشته باشد، می‌توان عوامل را نیز مرتّب کرد (مانند خوب، متوسط، ضعیف در جدول ۲–۱).

جدول ۲–۱. انواع متغیّرهایی که در سناریوهای نمایش داده‌های معمولی با آن مواجه می‌شوید.

نوع متغیّر	مثال‌ها	مقیاس مناسب	توصیف
کمی/ عددی پیوسته	۱٫۳، ۵٫۷، ۸۳، ۱٫۵X10^{-2}	پیوسته	مقادیر عددی دلخواه که می‌توانند اعداد صحیح، اعداد گویا یا اعداد واقعی باشند.
کمی/ عددی گسسته	۱، ۲، ۳، ۴	گسسته	اعداد در واحدهای گسسته. این‌ها معمولاً امّا نه لزوماً اعداد صحیح هستند. به عنوان مثال، اعداد ۰٫۵، ۱٫۰، ۱٫۵ نیز می‌توانند به عنوان گسسته در نظر گرفته شوند اگر مقادیر بینابینی در مجموعه داده وجود نداشته باشد.
کیفی/ طبقه‌بندی‌شده بدون ترتیب	سگ، گربه، ماهی	گسسته	دسته‌بندی بدون ترتیب. این‌ها دسته‌بندی‌های گسسته و منحصر به فردی هستند که نظم ذاتی ندارند. به این متغیّرها عوامل نیز گفته می‌شود.
کیفی/ طبقه‌بندی‌شده دارای ترتیب	خوب، متوسط، ضعیف	گسسته	دسته‌بندی دارای ترتیب. این‌ها دسته‌بندی‌های گسسته و منحصر به فرد دارای ترتیب هستند. برای مثال، «متوسط» همیشه بین «خوب» و «ضعیف» قرار دارد. به این متغیّرها *عوامل دارای ترتیب* نیز گفته می‌شود.
تاریخ یا زمان	۵ ژانویه ۲۰۱۸، ۸:۰۳ صبح	پیوسته یا گسسته	روزها و/یا زمان‌های خاص. همچنین تاریخ‌های عمومی، مانند ۴ جولای یا ۲۵ دسامبر (بدون سال).
متن	روباه قهوه‌ای سریع از روی سگ تنبل می‌پرد.	هیچکدام، یا گسسته	متن آزاد. در صورت نیاز می‌توان به عنوان دسته‌بندی شده با آن‌ها برخورد کرد.

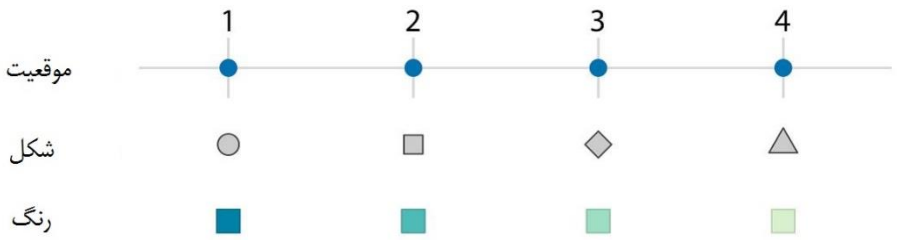
برای بررسی یک مثال عینی از این انواع مختلف داده‌ها، به جدول ۲-۲ نگاهی بیندازید. این جدول چند ردیف اول یک مجموعه داده را نشان می‌دهد که دمای طبیعی روزانه (متوسط دمای روزانه در یک بازه ۳۰ ساله) را برای چهار مکان ایالات متحده نشان می‌دهد. این جدول شامل پنج متغیّر است: ماه، روز، مکان، شناسه ایستگاه و دما (بر حسب درجهٔ فارنهایت). ماه یک عامل دارای ترتیب، روز یک مقدار عددی گسسته، مکان یک عامل بدون ترتیب، شناسه ایستگاه به طور مشابه یک عامل بدون ترتیب و دما یک مقدار عددی پیوسته است.

جدول ۲-۲. ۸ ردیف اول یک مجموعهٔ داده که دمای طبیعی روزانه را برای چهار ایستگاه هواشناسی فهرست می‌کند. منبع داده: اداره ملی اقیانوسی و جوی (NOAA)

ماه	روز	محل	کد ایستگاه	دما (درجه فارنهایت)
ژانویه	۱	شیکاگو	USW000141819	۲۵٫۶
ژانویه	۱	سن دیگو	USW00093107	۵۵٫۲
ژانویه	۱	هیوستون	USW00012918	۵۳٫۹
ژانویه	۱	دره مرگ	USC00042319	۵۱٫۰
ژانویه	۲	شیکاگو	USW00014819	۲۵٫۵
ژانویه	۲	سن دیگو	USW00093107	۵۵٫۳
ژانویه	۲	هیوستون	USW00012918	۵۳٫۸
ژانویه	۲	دره مرگ	USC00042319	۵۱٫۲

مقیاس، ارزش داده‌ها را بر روی زیبایی‌شناسی نگاشت می‌کند

برای نگاشت مقادیر داده‌ها بر روی زیبایی‌شناسی، باید مشخّص کنیم که کدام مقادیر داده با کدام ارزش‌های زیبایی‌شناسی خاص مطابقت دارد. به عنوان مثال، اگر شکل ما محور x داشته باشد، باید مشخّص کنیم که کدام مقادیر داده در موقعیّت‌های خاصی در امتداد این محور قرار می‌گیرند. به طور مشابه، ممکن است لازم باشد مشخّص کنیم که کدام مقادیر داده با اشکال یا رنگ‌های خاص نشان داده می‌شوند. این نگاشت بین مقادیر داده‌ها و ارزش‌های زیبایی‌شناسی از طریق مقیاس‌ها ایجاد می‌شود. مقیاس یک نگاشت منحصر به فرد بین داده و زیبایی‌شناسی را تعریف می‌کند (شکل ۲-۲). نکته مهم این است که مقیاس باید تناظر یک به یک داشته باشد، به طوری که برای هر مقدار داده خاص دقیقاً یک مقدار زیبایی‌شناسی وجود داشته باشد و بالعکس. اگر مقیاس تناظر یک به یک نداشته باشد، نمایش داده‌ها مبهم می‌شود.



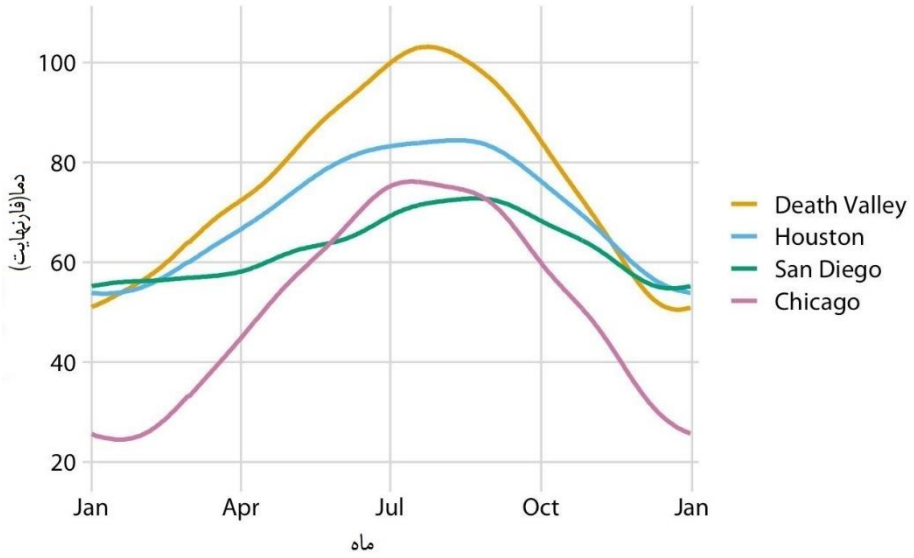
شکل ۲–۲. مقیاس‌ها، مقادیر داده‌ها را به زیبایی‌شناسی پیوند می‌دهند. در اینجا، اعداد ۱ تا ۴ بر روی یک مقیاس موقعیّت، یک مقیاس شکل و یک مقیاس رنگ نگاشت شده‌اند. برای هر مقیاس، هر عدد مربوط به یک موقعیّت، شکل یا رنگ منحصر به فرد است و بالعکس.

بیایید ببینیم این مساله در عمل به چه صورت است. می‌توانیم مجموعه داده‌های نشان‌داده‌شده در جدول ۲–۲ را در نظر بگیریم، دما را روی محور y، روز سال را روی محور x، و مکان را با رنگ ترسیم کنیم، و این زیبایی‌شناسی را با خطوط ثابت ترسیم کنیم. نتیجهٔ کار یک نمودار خطی استاندارد است که دمای طبیعی را در چهار مکان با تغییر آن‌ها در طول سال نشان می‌دهد (شکل ۲–۳).

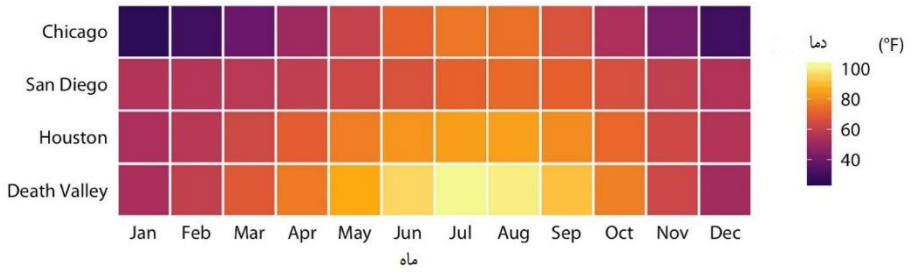
شکل ۲–۳ یک نمایش نسبتاً استاندارد برای منحنی دما است و احتمالاً نموداری است که اکثر دانشمندان حوزهٔ داده ابتدا آن را انتخاب می‌کنند. با این حال، این به ما بستگی دارد که کدام متغیّرها را در کدام مقیاس ترسیم کنیم. به عنوان مثال، به جای نگاشت دما بر روی محور y و مکان روی رنگ، می‌توانیم برعکس عمل کنیم. از آنجا که اکنون متغیّر کلیدی مورد نظر (دما) با رنگ نشان داده شده است، باید مناطق رنگی به اندازه کافی بزرگی را نشان دهیم تا رنگ‌ها اطّلاعات مفیدی را منتقل کنند [Stone, Albers Szafir, and Setlur 2014]. بنابراین، برای این نمودار، مربع را به جای خطوط، یکی برای هر ماه و مکان انتخاب نموده، و آن‌ها را بر اساس میانگین دمای طبیعی برای هر ماه رنگ کرده‌ایم (شکل ۲–۴).

باید تأکید کنیم که شکل ۲–۴ از دو مقیاس موقعیّت (ماه در امتداد محور x و مکان در امتداد محور y) استفاده می‌کند، امّا هیچ یک مقیاس پیوسته نیست. ماه یک عامل دارای ترتیب با ۱۲ سطح و مکان یک عامل بدون ترتیب با ۴ سطح است. بنابراین، هر دو مقیاس موقعیّت، گسسته هستند. برای مقیاس‌های موقعیّت گسسته، معمولاً سطوح مختلف عامل را با فاصله مساوی در امتداد محور قرار می‌دهیم. اگر عامل دارای ترتیب باشد (همانطور که در اینجا برای ماه این چنین است)، باید سطوح را با ترتیب مناسب و منطقی قرار داد. اگر عامل بدون ترتیب

باشد (همانطور که در اینجا برای مکان این چنین است)، ترتیب چیدمان سطوح به دلخواه است و می‌توانیم هر ترتیبی را که می‌خواهیم انتخاب کنیم. ما مکان‌ها را از سردترین (شیکاگو) تا گرم‌ترین (دره مرگ) مرتب کرده‌ایم تا نمایش رنگ‌ها جذاب باشد. با این حال، می‌توانستیم هر ترتیب دیگری را انتخاب کنیم و نمودار همچنان معتبر بود.



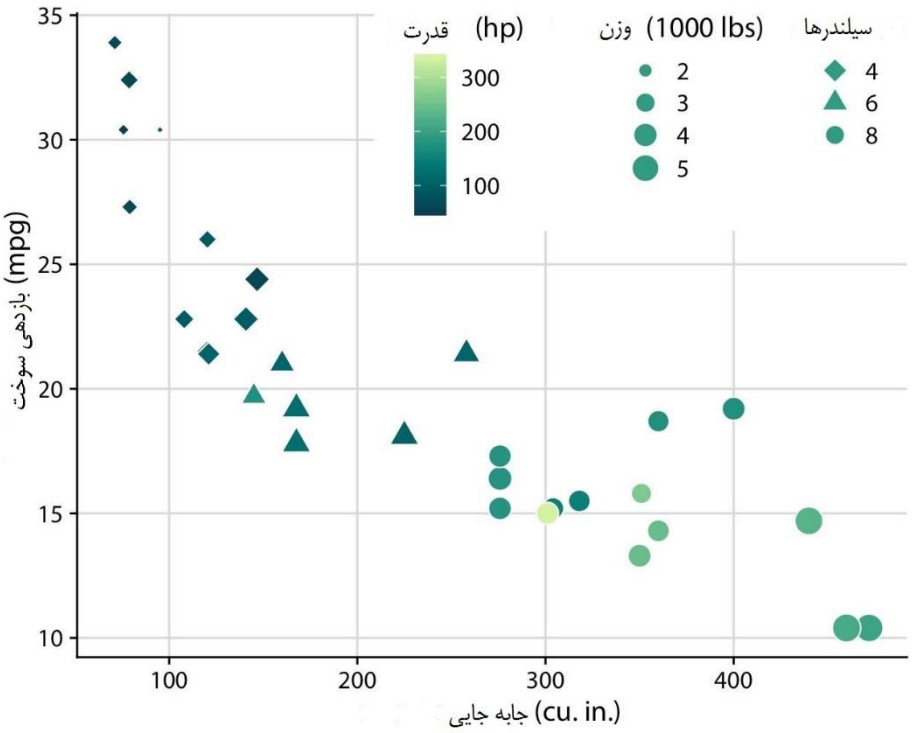
شکل ۲–۳. دمای طبیعی روزانه برای چهار مکان انتخاب شده در ایالات متحده. دما در محور y، روز سال در محور x و مکان با رنگ خط نگاشت شده است. منبع داده: NOAA



شکل ۲–۴. میانگین دمای طبیعی ماهانه برای چهار مکان در ایالات متحده. منبع داده:NOAA

هر دو شکل ۲–۳ و ۲–۴ در مجموع از سه مقیاس استفاده کرده‌اند: دو مقیاس موقعیّت و یک مقیاس رنگی. برای یک نمودار اولیه، این تعداد مقیاس کاملاً مرسوم است، امّا می‌توانیم بیش

از سه مقیاس را نیز همزمان استفاده کنیم. شکل ۲–۵ از پنج مقیاس استفاده می‌کند: دو مقیاس موقعیّت، یک مقیاس رنگ، یک مقیاس اندازه، و یک مقیاس شکل و هر مقیاس یک متغیّر متفاوت را از مجموعه داده نشان می‌دهد.



شکل ۲–۵. بازدهی سوخت در مقابل جابجایی، برای ۳۲ خودرو (مدل‌های ۱۹۷۳–۱۹۷۴). این شکل از پنج مقیاس مجزا برای نمایش داده‌ها استفاده می‌کند: (۱) محور x (جابجایی)، (۲) محور y (بازدهی سوخت)، (۳) رنگ نقاط داده (قدرت)؛ (۴) اندازه نقاط داده (وزن)، و (۵) شکل نقاط داده (تعداد سیلندرها). چهار متغیّر از پنج متغیّر نمایش داده شده (جابجایی، بازدهی سوخت، قدرت و وزن) عددی پیوسته هستند. متغیّر باقی مانده (تعداد سیلندرها) را می‌توان به صورت عددی گسسته یا کیفی ترتیبی در نظر گرفت. منبع داده: Motor Trend، 1974

فصل ۳

سیستم‌های مختصّات و محورها

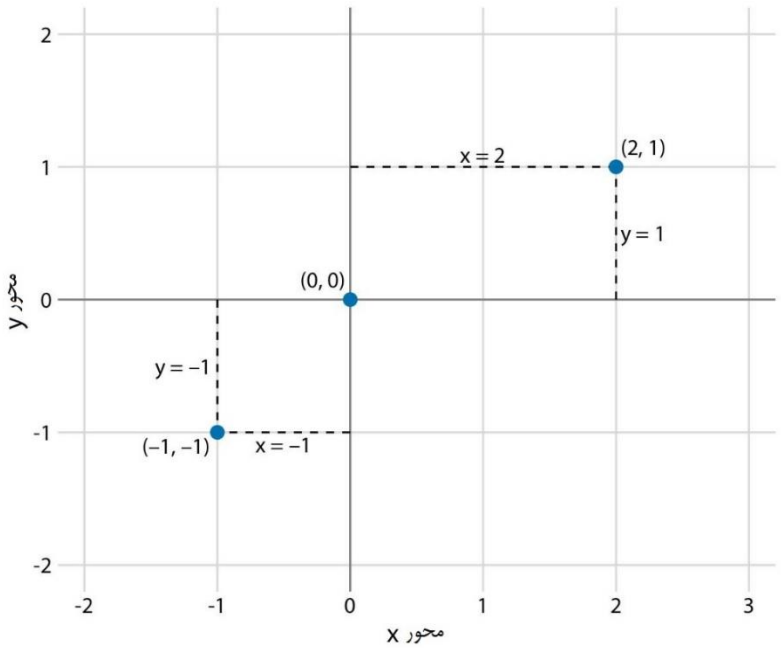
برای نمایش داده‌ها، ابتدا باید مقیاس‌های موقعیّت را تعریف کنیم. این مقیاس تعیین می‌کند در یک نمودار مقادیر مختلف داده در کجا قرار دارند. ما نمی‌توانیم داده‌ها را بدون قرار دادن نقاط داده مختلف در مکان‌های مختلف نمایش دهیم، حتی اگر آن‌ها را در کنار یکدیگر در امتداد یک خط مرتّب کنیم. برای نمودارهای دو بُعدی معمولی، دو عدد برای مشخّص کردن یک نقطه مورد نیاز است و بنابراین به دو مقیاس موقعیّت نیاز داریم. این دو مقیاس معمولاً، امّا نه لزوماً، محورهای x و y نمودار هستند. همچنین باید آرایش هندسی نسبی این مقیاس‌ها را مشخّص کنیم. به طور معمول، محور x به صورت افقی و محور y به صورت عمودی قرار می‌گیرند، امّا می‌توانیم الگوهای دیگری را نیز انتخاب کنیم. به عنوان مثال، می‌توانیم محور y را با زاویه حادّه نسبت به محور x رسم کنیم، یا می‌توانیم یک محور را در یک دایره و دیگری را به صورت شعاعی رسم کنیم. ترکیب مجموعه‌ای از مقیاس‌های موقعیّت و آرایش هندسی نسبی آن‌ها را سیستم مختصّات می‌گویند.

مختصّات دکارتی[1]

پرکاربردترین سیستم مختصّات برای نمایش داده‌ها، سیستم مختصّات دکارتی دوبعدی است، که در آن هر مکان به طور منحصر به فرد با یک مقدار x و یک مقدار y مشخّص می‌شود. محورهای x و y به صورت عمود بر هم قرار داشته و مقادیر داده در یک فاصله مشخّص در

1. Cartesian Coordinates

امتداد هر دو محور قرار می‌گیرند (شکل ۳-۱). این دو محور مقیاس‌های موقعیّت پیوسته هستند و می‌توانند هم اعداد حقیقی مثبت و هم منفی را نشان دهند. برای مشخّص کردن کامل سیستم مختصّات، باید محدودهٔ اعدادی را که هر محور پوشش می‌دهد، مشخّص کنیم. در شکل ۳-۱، محور x از ۲٫۲- تا ۳٫۲ و محور y از ۲٫۲- تا ۲٫۲ امتداد دارد. هر مقدار داده بین این دو سر طیف محور در محل مناسب در نمودار قرار می‌گیرد. هر مقدار داده خارج از محدودهٔ محور نادیده گرفته می‌شود.

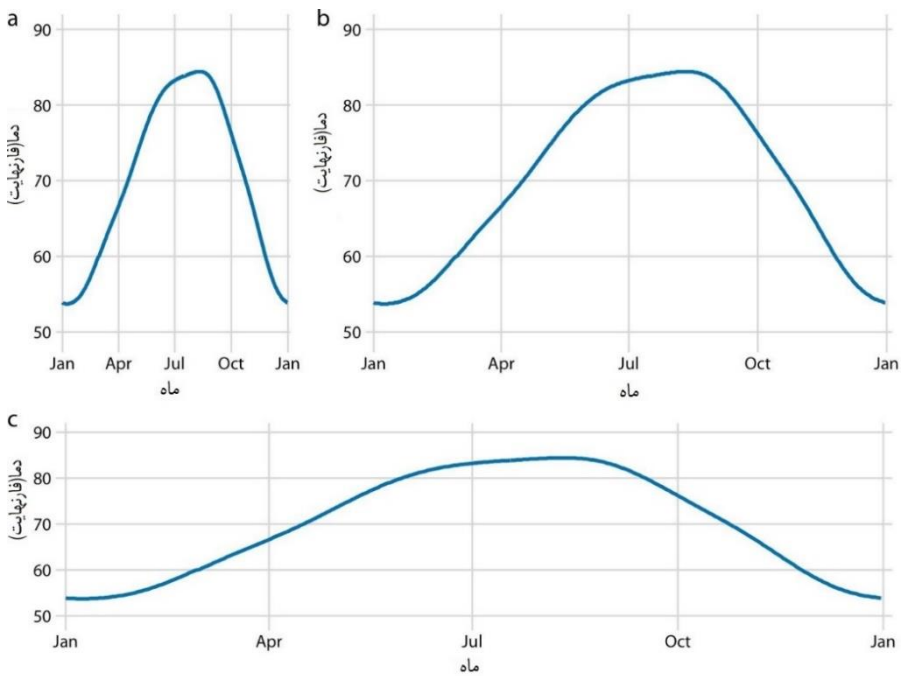


نمودار ۳-۱. سیستم مختصّات دکارتی استاندارد، محور افقی معمولاً x و محور عمودی y نامیده می‌شود. دو محور یک شبکه با فاصله‌های مساوی را تشکیل می‌دهند. در اینجا، خطوط شبکه x و y با واحدی معادل یک از هم جدا شده‌اند. نقطه (۲، ۱) دو واحد در محور x در سمت راست و یک واحد در محور y بالاتر از مبدا (۰، ۰) قرار دارد. نقطه (۱-، ۱-) یک واحد در محور x در سمت چپ مبدا و یک واحد در محور y در زیر مبدا قرار دارد.

با این حال، مقادیر داده‌ها معمولاً فقط مقادیر عددی نیستند و دارای واحد نیز می‌باشند. برای مثال، اگر در حال اندازه‌گیری دما هستیم، ممکن است مقادیر بر حسب درجهٔ سانتی‌گراد یا فارنهایت اندازه‌گیری شوند. به طور مشابه، اگر مسافت را اندازه‌گیری کنیم، مقادیر ممکن است بر حسب کیلومتر یا مایل و اگر مدّت زمان را اندازه‌گیری کنیم، مقادیر بر حسب دقیقه، ساعت یا روز اندازه‌گیری می‌شوند. در یک سیستم مختصّات دکارتی، فاصله بین خطوط شبکه در امتداد یک محور نشان‌دهندهٔ فواصل گسسته در این واحدهای داده است. برای مثال، در

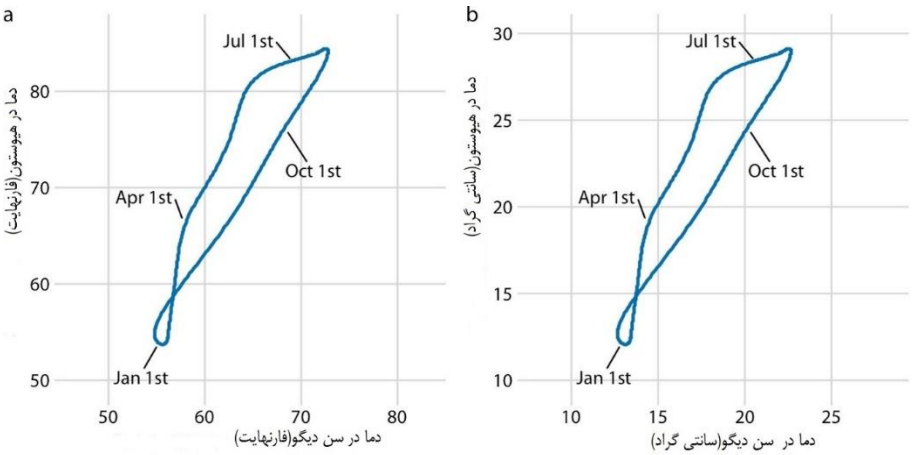
مقیاس دما، ممکن است برای هر ۱۰ درجهٔ فارنهایت یک خط شبکه داشته باشیم، و در مقیاس مسافت، ممکن است برای هر ۵ کیلومتر یک خط شبکه داشته باشیم.

یک سیستم مختصّات دکارتی می‌تواند دو محور داشته باشد که دو واحد متفاوت را نشان دهد. زمانی که ما دو نوع متغیّر مختلف را بر x و y نگاشت می‌کنیم، این وضعیّت معمولاً به وجود می‌آید. به عنوان مثال، در شکل ۲–۳، دما را بر اساس روزهای سال ترسیم کردیم. محور y در شکل ۲–۳ بر حسب درجهٔ فارنهایت، با یک خط شبکه در هر ۲۰ درجه، و محور x بر حسب ماه، با یک خط شبکه در اوّلین ماه هر فصل ترسیم شده است. هر زمان که دو محور با واحدهای مختلف اندازه‌گیری شوند، می‌توانیم یکی را نسبت به دیگری کشیده‌تر یا فشرده‌تر کنیم و نمودار معتبری از داده‌ها را داشته باشیم (شکل ۳–۲). اینکه کدام نسخه ارجح است به داستانی که می‌خواهیم روایت کنیم بستگی دارد. شکل بلند و باریک بر تغییر در امتداد محور y تاکید دارد و شکل کوتاه و پهن برعکس بر تغییر در امتداد محور x تاکید دارد. در حالت ایده‌آل، ما می‌خواهیم نسبتی از طول و عرض نمودار را انتخاب کنیم تا اطمینان حاصل شود که تفاوت‌های مهم در موقعیّت‌های مختلف قابل شناسایی باشند.



نمودار ۳–۲. دمای روزانه برای هیوستون، تگزاس. دما در محور y و روز سال در محور x ترسیم شده است. قسمت‌های (الف)، (ب) و (ج) یک نمودار را در نسبت‌های مختلف نشان می‌دهند. هر سه بخش نمایش معتبری از داده‌های دما هستند. منبع داده: NOAA

از طرف دیگر، اگر محورهای x و y با واحدهای یکسان اندازه‌گیری شوند، فاصله‌های شبکه برای دو محور باید برابر باشد، به طوری که فاصله یکسان در امتداد محور x یا y با تعداد واحدهای یکسان داده مطابقت داشته باشد. به عنوان مثال، می‌توانیم دما را در هیوستون، تگزاس، در برابر دمای سن دیگو، کالیفرنیا، برای هر روز از سال رسم کنیم (شکل ۳-۳ الف). از آنجایی که مقدار یکسانی در امتداد هر دو محور رسم شده است، باید مطمئن شویم که خطوط شبکه مربع‌های کاملی را تشکیل می‌دهند، همانطور که در شکل ۳-۳ الف وجود دارد.



نمودار ۳-۳. دمای روزانه برای هیوستون، تگزاس، در مقابل دمای مربوطه در سن دیگو، کالیفرنیا رسم شده است. روزهای اول ماه‌های ژانویه، آوریل، جولای و اکتبر برای ارائه یک مرجع موقّت برجسته شده‌اند. (الف) دما بر حسب درجهٔ فارنهایت نشان داده شده است. (ب) دما بر حسب درجهٔ سانتیگراد نشان داده شده است. منبع داده: NOAA

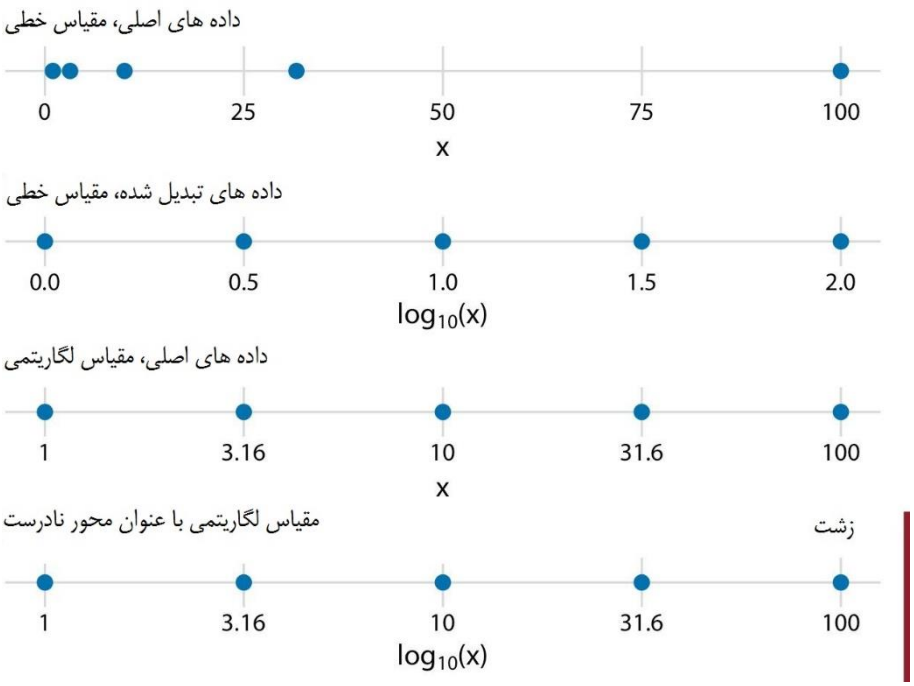
ممکن است بخواهید بدانید که اگر واحدهای داده را تغییر دهید چه اتفاقی می‌افتد. به هر حال، واحدها دلخواه هستند و ترجیحات شما ممکن است با ترجیحات دیگران متفاوت باشد. تغییر در واحدها یک تبدیل خطی است که در آن یک عدد را به همه مقادیر داده‌ها اضافه یا از آن‌ها کم می‌کنیم یا همه داده‌ها را در عدد دیگری ضرب می‌کنیم. خوشبختانه، سیستم مختصّات دکارتی تحت چنین تبدیل‌های خطی ثابت می‌مانند. بنابراین، می‌توانید واحدهای داده‌های خود را تغییر دهید و تا زمانی که هر دو محور را بر این اساس تغییر دهید، نمودار حاصل تغییر نخواهد کرد. به عنوان مثال، نمودارهای ۳-۳ الف و ۳-۳ ب را مقایسه کنید. هر دو، دادهٔ یکسانی را نشان می‌دهند، امّا در قسمت (الف) واحد دما درجهٔ فارنهایت و در قسمت (ب) درجهٔ سانتی‌گراد هستند. اگرچه خطوط شبکه در مکان‌های متفاوتی قرار دارند و اعداد در امتداد محورها متفاوت هستند، این دو نمودار دقیقاً یکسان به نظر می‌رسند.

محورهای غیرخطی

در سیستم مختصّات دکارتی، فاصله‌گذاری خطوط شبکه در امتداد یک محور هم در واحدهای داده و هم در نمودار حاصل یکسان است. به اینگونه مقیاس‌های موقعیّت در این سیستم‌های مختصّات، خطی اطلاق می‌شود. در حالی که مقیاس‌های خطی عموماً نمایش دقیقی از داده‌ها را ارائه می‌دهند، سناریوهایی وجود دارد که مقیاس‌های غیر خطی ارجح هستند. در مقیاس غیر خطی، فاصله یکنواخت در واحدهای داده منجر به فاصله غیر یکنواخت در نمودار می‌شود، یا برعکس فاصله یکنواخت در نمودار مطابق با فاصله غیر یکنواخت در واحدهای داده است.

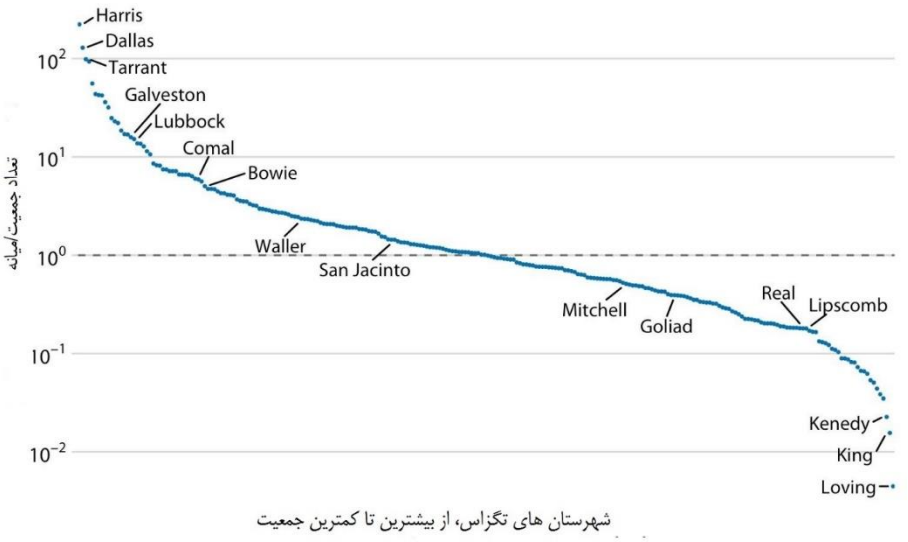
رایج‌ترین مقیاس غیر خطی مورد استفاده، مقیاس لگاریتمی است. مقیاس‌های لگاریتمی در ضرب به صورت خطی هستند، به طوری که افزایش یک واحد در مقیاس با ضرب یک مقدار ثابت در آن مقدار مطابقت دارد. برای ایجاد یک مقیاس لگاریتمی، باید مقادیر داده‌ها را تبدیل لگاریتمی کنیم و اعدادی که در امتداد خطوط شبکه محور نشان داده شده‌اند را تبدیل نمایی کنیم. این فرآیند در شکل ۳-۴ نشان داده شده است که اعداد ۱، ۳٫۱۶، ۱۰، ۳۱٫۶ و ۱۰۰ را در مقیاس‌های خطی و لگاریتمی نشان می‌دهد. اعداد ۳٫۱۶ و ۳۱٫۶ ممکن است انتخاب‌های عجیبی به نظر برسند، امّا به این دلیل انتخاب شده‌اند چون دقیقاً در وسط اعداد ۱ تا ۱۰ و ۱۰ تا ۱۰۰ در مقیاس لگاریتمی قرار دارند. ما می‌توانیم این را با مشاهده $\sqrt{10} = 10^{0.5} \approx$ ۳٫۱۶، و معادل $10^{0.5} \times 10 = 10^{1.5} \approx$ ۳۱٫۶ × ۳٫۱۶ ≈ ۱۰ ببینیم. به طور مشابه، $10^{0.5}$.

از نظر ریاضی، هیچ تفاوتی بین رسم داده‌های تبدیل شده با لگاریتم در مقیاس خطی یا ترسیم داده‌های اصلی در مقیاس لگاریتمی وجود ندارد (شکل ۳-۴). تنها تفاوت در برچسب‌گذاری برای تیک‌های محور و برای کل محور نهفته است. در بیشتر موارد، برچسب‌گذاری برای مقیاس لگاریتمی ترجیح داده می‌شود، زیرا بار ذهنی کمتری را بر خواننده تحمیل می‌کند تا اعداد نشان داده شده را به عنوان برچسب‌های تیک محور تفسیر کند. همچنین خطر سردرگمی کمتری در مورد پایهٔ لگاریتم وجود دارد. هنگام کار با داده‌های تبدیل‌شده لگاریتمی، ممکن است در مورد اینکه داده‌ها با استفاده از لگاریتم طبیعی یا لگاریتم در مبنای ۱۰ تبدیل شده‌اند سردرگم شویم. حتی ممکن است برچسب‌گذاری هم مبهم باشد، مثلاً log(x)، که اصلاً مبنای لگاریتم را مشخّص نمی‌کند، لذا توصیه می‌کنیم همیشه هنگام کار با داده‌های تبدیل شدهٔ لگاریتمی، مبنا را مشخّص کنید. هنگام ترسیم داده‌های تبدیل شدهٔ لگاریتمی، همیشه مبنا را در برچسب‌گذاری محور مشخّص کنید.



شکل ۳–۴. رابطه بین مقیاس‌های خطی و لگاریتمی. نقاط مربوط به مقادیر داده ۱، ۳٫۱۶، ۱۰، ۳۱٫۶، و ۱۰۰ است که اعدادی با فاصلهٔ مساوی در مقیاس لگاریتمی هستند. ما می‌توانیم این نقاط داده را در مقیاس خطی نمایش دهیم، می‌توانیم آن‌ها را تبدیل لگاریتمی کنیم و سپس آن‌ها را در مقیاس خطی نشان دهیم، یا می‌توانیم آن‌ها را در مقیاس لگاریتمی نشان دهیم. دقّت نمایید که عنوان صحیح محور برای مقیاس لگاریتمی، نام متغیّر می باشد، نه لگاریتم آن متغیّر.

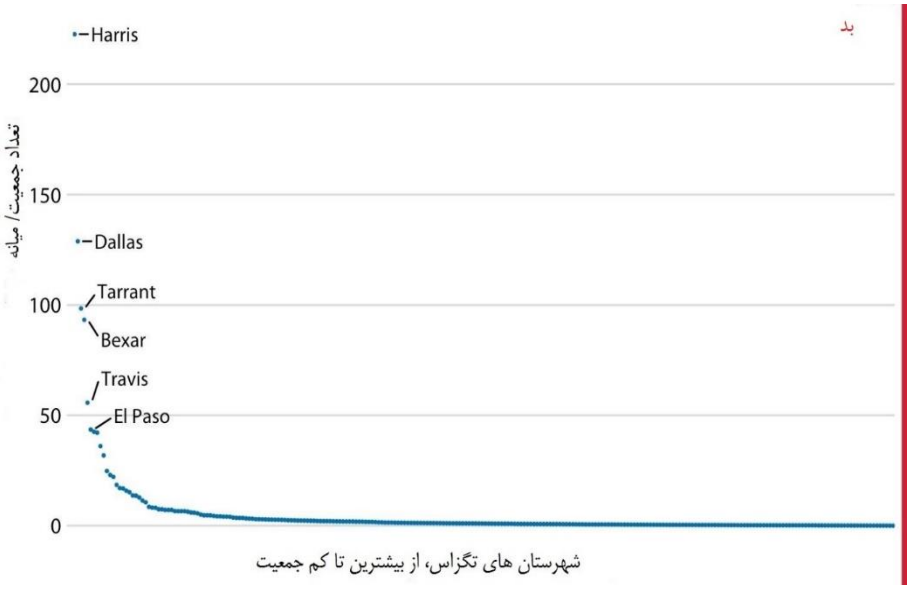
از آنجایی که ضرب در مقیاس لگاریتمی مانند جمع در مقیاس خطی به نظر می‌رسد، مقیاس‌های لگاریتمی انتخاب مناسبی برای داده‌هایی است که با ضرب یا تقسیم به دست آمده‌اند. به طور خاص، نسبت‌ها باید به طور کلی در مقیاس لگاریتمی نشان داده شوند. به عنوان مثال، من تعداد ساکنان هر شهرستان در تگزاس را در نظر گرفتم و آن را بر میانهٔ ساکنان در تمام شهرستان‌های تگزاس تقسیم کردم. نسبت حاصل عددی است که می‌تواند بزرگتر یا کوچکتر از ۱ باشد. عدد ۱ نشان می‌دهد که شهرستان مربوطه دارای میانه تعداد ساکنان است. هنگام نمایش این نسبت‌ها در مقیاس لگاریتمی، می‌بینیم که تعداد جمعیّت در شهرستان‌های تگزاس به طور متقارن در حول و حوش میانه توزیع شده است، و اینکه پرجمعیّت‌ترین شهرستان‌ها بیش از ۱۰۰ برابر بیشتر از میانهٔ ساکنان جمعیّت دارند در حالی که ساکنین شهرستان‌های کم جمعیّت ۱۰۰ برابر کمتر از میانهٔ کل جمعیّت است. (شکل ۳–۵).



نمودار ۳–۵. جمعیّت شهرستان‌های تگزاس نسبت به مقدار میانه. نام شهرستان‌های منتخب ذکر شده است. خط چین نسبت ۱ را نشان می‌دهد که مربوط به شهرستانی با تعداد جمعیّت برابر مقدار میانهٔ جمعیّت است. پرجمعیّت ترین شهرستان‌ها تقریباً ۱۰۰ برابر بیشتر از میانه سکنه دارند و شهرستان‌های کم جمعیّت تقریباً ۱۰۰ برابر کمتر از میانه سکنه دارند. منبع داده: سرشماری دهه ۲۰۱۰ ایالات متحده.

در مقابل، برای همان داده‌ها، یک مقیاس خطی تفاوت‌های بین یک شهرستان با میانهٔ جمعیّت و یک شهرستان با تعداد جمعیّت بسیار کمتر از میانه را پنهان می‌کند (شکل ۳–۶).

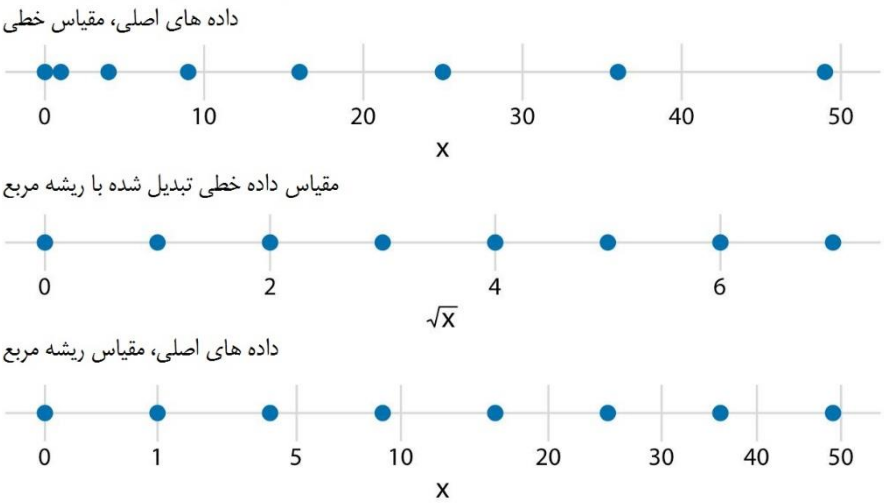
در مقیاس لگاریتمی، مقدار ۱ نقطه میانی طبیعی است، مشابه مقدار صفر در مقیاس خطی. ما می‌توانیم مقادیر بزرگتر از ۱ را نشان دهندهٔ ضرب و مقادیر کمتر از ۱ را نشان‌دهندهٔ تقسیم در نظر بگیریم. به عنوان مثال، می‌توانیم بنویسیم ۱۰ = ۱ × ۱۰ و ۰٫۱ = ۱/۱۰. از طرف دیگر، مقدار صفر هرگز نمی‌تواند در مقیاس لگاریتمی گزارش شود. در حقیقت این مقدار بی نهایت با ۱ فاصله دارد. یکی از راه‌های در نظر گرفتن آن، این است که log(0)=-∞. به بیان دیگر در نظر بگیرید که برای رفتن از ۱ به صفر یا باید تعداد بی نهایت تقسیم بر یک مقدار محدود (مثلاً ۰ =/۱۰/۱۰/۱۰/۱۰/۱۰/۱۰/...) یا یک تقسیم بر بی‌نهایت انجام شود (یعنی ۰=۱/∞).



نمودار ۳–۶. اندازهٔ جمعیّت شهرستان‌های تگزاس نسبت به مقدار میانه. با نمایش نسبت در مقیاس خطی، بر نسبت‌های > ۱ بیش از حد تأکید شده و نسبت‌های < ۱ پوشانده شده است. به عنوان یک قاعدهٔ کلی، نسبت‌ها نباید در مقیاس خطی نمایش داده شوند. منبع داده: سرشماری دهه ۲۰۱۰ ایالات متحده.

مقیاس‌های لگاریتمی اغلب زمانی استفاده می‌شود که مجموعهٔ داده حاوی اعدادی با مقادیر بسیار متفاوت باشد. برای شهرستان‌های تگزاس که در شکل‌های ۳–۵ و ۳–۶ نشان داده شد، پرجمعیّت‌ترین شهرستان (هریس[1]) در سرشماری ایالات متحده در سال ۲۰۱۰ دارای ۴۰۹۲۴۵۹ نفر بود در حالی که کم جمعیّت‌ترین شهرستان (لاوینگ[2]) دارای ۸۲ نفر بود. لذا، حتی اگر اعداد جمعیّت را بر میانهٔ آن‌ها تقسیم نکرده باشیم تا آن‌ها را به نسبت تبدیل کنیم، همچنان مقیاس لگاریتمی مناسب خواهد بود. امّا اگر شهرستانی با جمعیّت صفر وجود داشت، چه می‌کردیم؟ این شهرستان را نمی‌توان در مقیاس لگاریتمی نشان داد، زیرا در منهای بی‌نهایت قرار می‌گیرد. در چنین حالاتی، توصیه می‌شود که از مقیاس ریشهٔ دوّم استفاده شود که از تبدیل ریشهٔ دوّم به جای تبدیل لگاریتمی استفاده می‌کند (شکل ۳–۷). درست مانند مقیاس لگاریتمی، مقیاس ریشهٔ دوّم اعداد بزرگتر را در محدودهٔ کوچکتری فشرده می‌کند، امّا بر خلاف مقیاس لگاریتمی، امکان حضور مقدار صفر را فراهم می‌کند.

1. Harris
2. Loving

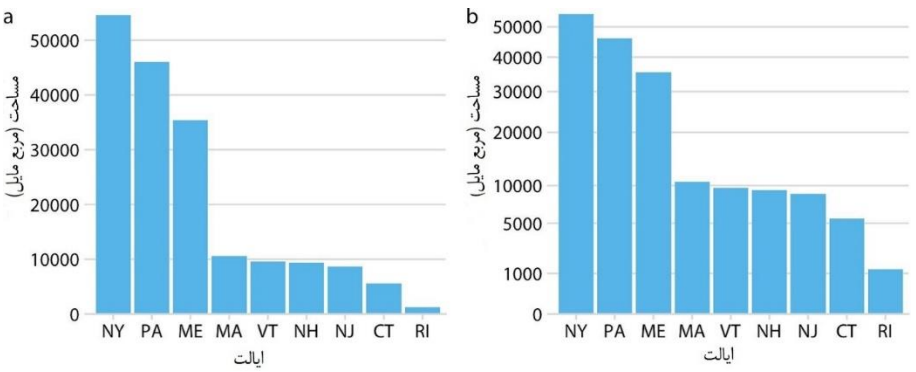


شکل ۳–۷. رابطهٔ بین مقیاس‌های خطی و ریشهٔ دوم. نقاط مربوط به مقادیر داده ۰، ۱، ۴، ۹، ۱۶، ۲۵، ۳۶ و ۴۹ هستند که اعدادی با فاصلهٔ مساوی در مقیاس ریشهٔ دوّم هستند، زیرا آن‌ها مجذور اعداد صحیح از ۰ تا ۷ هستند. می‌توانیم این نقاط داده را در مقیاس خطی نمایش دهیم، می‌توانیم آن‌ها را با ریشهٔ دوّم تبدیل کنیم و سپس آن‌ها را در مقیاس خطی نشان دهیم، یا می‌توانیم آن‌ها را در مقیاس ریشهٔ دوّم نشان دهیم.

دو مشکل در مقیاس‌های ریشه دوّم وجود دارد. اولاً، در حالی که در مقیاس خطی، یک واحد تغییر مربوط به جمع یا تفریق یک مقدار ثابت است، و در مقیاس لگاریتمی با ضرب یا تقسیم بر یک مقدار ثابت مطابقت دارد، چنین قانونی برای مقیاس ریشه دوّم وجود ندارد. معنای یک واحد تغییر در مقیاس ریشهٔ دوّم به مقدار مقیاسی که در آن شروع می‌کنیم بستگی دارد. دوّم، مشخّص نیست که چگونه می‌توان تیک‌های محور را در مقیاس ریشه دوّم به بهترین شکل قرار داد. برای به دست آوردن تیک هایی با فاصله یکسان، باید آن‌ها را در محل مجذورها قرار دهیم، امّا تیک‌های محوری، به عنوان مثال، در موقعیّت‌های ۰، ۴، ۲۵، ۴۹ و ۸۱ (مجذورهای دوّم) غیر قابل درک است. از طرف دیگر، می‌توانیم آن‌ها را در فواصل خطی (۱۰، ۲۰، ۳۰، و غیره) قرار دهیم، امّا این باعث می‌شود که تیک‌های محوری بسیار کمی در نزدیکی انتهای پایینی مقیاس و یا تعداد بیش از حدی در نزدیکی انتهای بالایی ایجاد شود. در شکل ۳–۷، تیک‌های محور را در موقعیّت‌های ۰، ۱، ۵، ۱۰، ۲۰، ۳۰، ۴۰ و ۵۰ در مقیاس ریشه دوّم قرار داده‌ایم. این‌ها مقادیر دلخواه هستند امّا پوشش معقولی از محدودهٔ داده را ارائه می‌دهند.

با وجود این مشکلات در مقیاس‌های ریشهٔ دوّم، آن‌ها مقیاس‌های موقعیّت معتبری هستند و احتمال اینکه کاربردهای مناسبی داشته باشند کم نیست. به عنوان مثال، درست مانند مقیاس

لگاریتمی که مقیاس طبیعی برای نسبت‌ها است، می‌توان استدلال کرد که مقیاس ریشهٔ دوّم مقیاس طبیعی برای داده‌هایی است که به صورت مجذور هستند. یکی از سناریوهایی که در آن داده‌ها به طور طبیعی مجذور هستند، در زمینه مناطق جغرافیایی است. اگر نواحی مناطق جغرافیایی را در مقیاس ریشهٔ دوّم نشان دهیم، وسعت خطی مناطق را از شرق به غرب یا شمال به جنوب برجسته می‌کنیم. برای مثال، اگر بخواهیم بدانیم چقدر طول می‌کشد تا در یک منطقه رانندگی کنیم، این محدوده‌ها می‌توانند مرتبط باشند. شکل ۳–۸ ایالت‌های شمال شرقی ایالات متحده را در مقیاس خطی و ریشه دوّم نشان می‌دهد. با وجود اینکه مساحت این ایالت‌ها کاملاً متفاوت است (شکل ۳–۸ الف)، زمان نسبی‌ای که برای گذر ار هر ایالت لازم است با دقّت بیشتری در مقیاس ریشه دوّم (شکل ۳–۸ ب) نسبت به مقیاس خطی نشان داده شده است (شکل ۳–۸ الف).



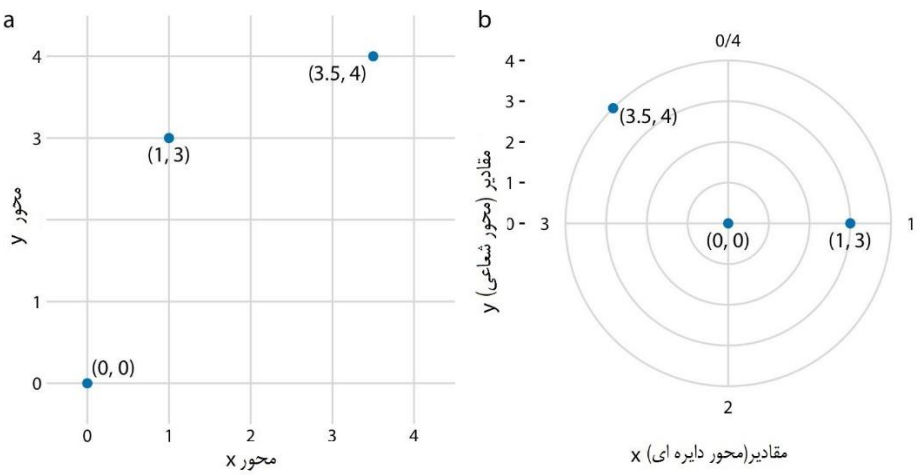
شکل ۳–۸. مساحت ایالت‌های شمال شرقی ایالات متحده. (الف) مساحت نشان داده شده در مقیاس خطی. (ب) مساحت نشان داده شده در مقیاس ریشهٔ دوّم.منبع داده: گوگل

سیستم‌های مختصّات با محورهای منحنی

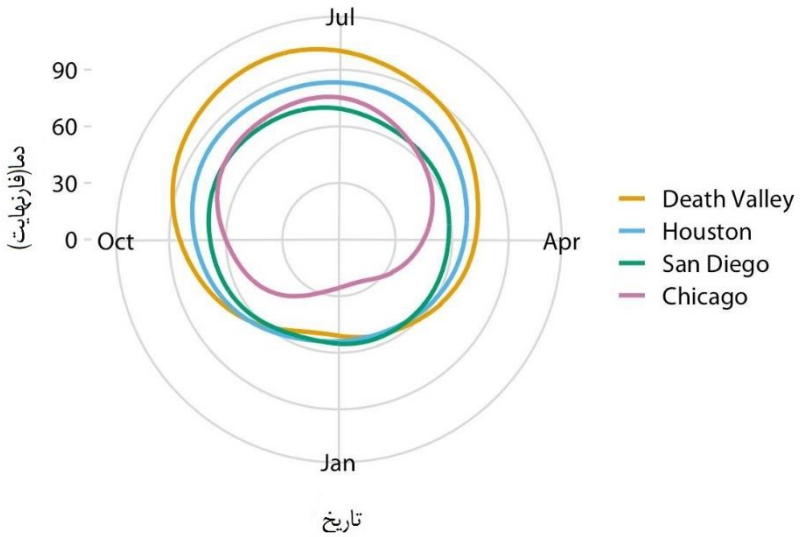
تمام سیستم‌های مختصّاتی که تاکنون با آن‌ها مواجه شده‌ایم، از دو محور مستقیم استفاده کرده‌اند که در یک زاویه قائم نسبت به یکدیگر قرار گرفته‌اند، حتی اگر خود محورها نگاشتی غیرخطی از مقادیر داده‌ها در موقعیّت‌ها ایجاد کنند. با این حال، سیستم‌های مختصّات دیگری وجود دارند که خود محورها منحنی هستند. به طور خاص، در سیستم مختصّات قطبی، موقعیّت‌ها را از طریق یک زاویه و فاصله شعاعی از مبدا مشخّص می‌کنیم، بنابراین زاویه محور دایره‌ای است (شکل ۳–۹).

مختصّات قطبی می‌تواند برای داده‌هایی با ماهیت تناوبی مفید باشد، به گونه‌ای که مقادیر داده در یک انتهای مقیاس را می‌توان به طور منطقی به مقادیر داده در انتهای دیگر متصل کرد. به عنوان مثال، روزهای یک سال را در نظر بگیرید. ۳۱ دسامبر آخرین روز سال است، امّا یک روز قبل از اوّلین روز سال بعدی است. اگر بخواهیم نشان دهیم که چگونه مقداری در طول سال تغییر می‌کند، می‌توان از مختصّات قطبی با مختصّات زاویه‌ای که هر روز را مشخّص می‌کند استفاده نمود. بیایید این مفهوم را برای داده‌های دمایی شکل ۲–۳ اعمال کنیم. از آنجایی که دمای نرمال دمای متوسطی است که به هیچ سال خاصی وابسته نیست، می‌توان ۳۱ دسامبر را ۳۶۶ روز دیرتر از اول ژانویه (دمای نرمال شامل ۲۹ فوریه) و همچنین ۱ روز زودتر از آن در نظر گرفت.

با ترسیم دما در یک سیستم مختصّات قطبی، بر این خاصیت چرخه‌ای تأکید می‌کنیم (شکل ۳–۱۰). در مقایسه با شکل ۲–۳، نمودار قطبی نشان می‌دهد که چقدر دما در دره مرگ، هیوستون و سن دیگو از اواخر پاییز تا اوایل بهار مشابه است. در سیستم مختصّات دکارتی، این واقعیّت مغفول می‌ماند، زیرا مقادیر دما در اواخر دسامبر و اوایل ژانویه در قسمت‌های مخالف نمودار نشان داده می‌شوند و بنابراین یک واحد بصری یکپارچه را تشکیل نمی‌دهند.

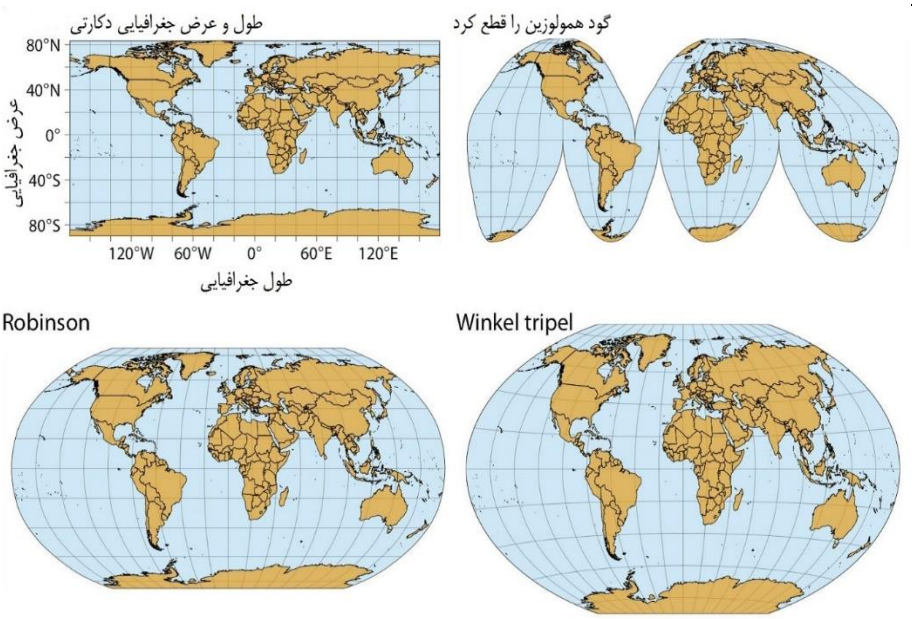


شکل ۳–۹. رابطه بین مختصّات دکارتی و قطبی.(الف) سه نقطهٔ داده در یک سیستم مختصّات دکارتی نشان داده شده است.(ب) همان سه نقطهٔ داده در یک سیستم مختصّات قطبی نشان داده شده است. مختصّات x را از قسمت (الف) گرفته و به عنوان مختصّات زاویه‌ای و مختصّات y را از قسمت (الف) گرفته و به عنوان مختصّات شعاعی استفاده کرده‌ایم. محور دایره‌ای در این مثال از ۰ تا ۴ است و بنابراین $x = 0$ و $x = 4$ مکان‌های مشابهی در این سیستم مختصّات هستند.



شکل ۳–۱۰. دمای روزانه برای چهار مکان انتخاب شده در ایالات متحده، که در مختصّات قطبی نشان داده شده است. فاصلهٔ شعاعی از نقطه مرکزی، دمای روزانه را برحسب درجهٔ فارنهایت نشان می‌دهد و روزهای سال در خلاف جهت عقربه‌های ساعت مرتّب شده‌اند که از اوّل ژانویه در موقعیّت ۶:۰۰ شروع می‌شود. منبع داده: NOAA

شرایط دیگری که در آن با محورهای منحنی مواجه می‌شویم، در زمینه داده‌های مکانی، یعنی نقشه‌ها است. مکان‌های روی کرهٔ زمین بر اساس طول و عرض جغرافیایی مشخّص می‌شوند. امّا از آنجا که زمین یک کره است، ترسیم طول و عرض جغرافیایی به عنوان محورهای دکارتی گمراه‌کننده است و توصیه نمی‌شود (شکل ۳–۱۱). در عوض، ما از انواع مختلفی از روش‌های غیرخطی استفاده می‌کنیم که تلاش می‌کنند تا خطاها را به حداقل برسانند و تعادل را بین مناطق یا زوایای حفظ شده نسبت به خطوط شکل واقعی روی کرهٔ زمین ایجاد می‌کنند (شکل ۳–۱۱).



شکل ۳–۱۱. نقشهٔ جهان، در چهار طرح مختلف نشان داده شده است. سیستم طول و عرض جغرافیایی دکارتی، طول و عرض جغرافیایی هر مکان را بر روی یک سیستم مختصّات دکارتی منظّم ترسیم می‌کند. این نقشه‌برداری باعث ایجاد اعوجاج قابل توجّهی هم در ناحیه و هم در زوایا نسبت به مقادیر واقعی آن‌ها در کرهٔ سه بُعدی می‌شود. برون‌یابی هومولوسین منقطع گود، به قیمت تقسیم برخی از توده‌های خشکی به قطعات جداگانه، به ویژه گرینلند و قطب جنوب، به خوبی سطوح مناطق واقعی را نشان می‌دهد. طرح رابینسون و طرح ریزی سه گانه وینکل، تعادلی بین اعوجاج زاویه‌ای و ناحیه‌ای برقرار می‌کنند و معمولاً برای نقشه‌های کل کرهٔ زمین استفاده می‌شوند.

فصل ۴

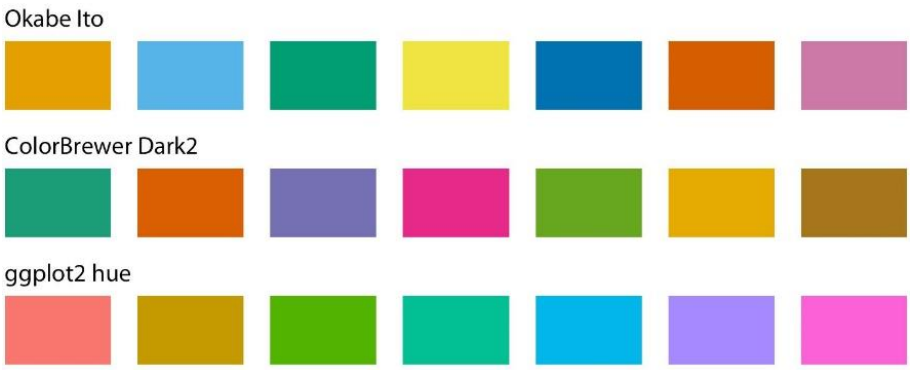
مقیاس‌های رنگی

سه کاربرد اساسی برای استفاده از رنگ در ترسیم نمودارها وجود دارد: تشخیص گروه‌های داده از یکدیگر، نمایش مقادیر داده‌ها و برجسته‌سازی. انواع رنگ‌هایی که استفاده می‌شود و نیز روش استفاده از آن‌ها برای این سه مورد، کاملاً متفاوت است.

رنگ به عنوان ابزاری برای تمایز

اغلب از رنگ به عنوان وسیله‌ای برای تشخیص اقلام یا گروه‌های مجزا که نظم ذاتی ندارند، مانند کشورهای مختلف روی نقشه یا تولیدکنندگان مختلف یک محصول خاص، استفاده می‌کنیم. در این حالت مقیاس رنگی کیفی به کار می‌رود. چنین مقیاسی شامل مجموعه‌ای محدود از رنگ‌های خاص است که به‌طور واضح از یکدیگر متمایز به نظر می‌رسند و در عین حال معادل یکدیگر هستند. شرط دوّم مستلزم آن است که هیچ رنگی نسبت به رنگ‌های دیگر برجسته نباشد. همچنین، رنگ‌ها نباید وجود ترتیب را القا کنند مانند حالتی که در آن رنگ‌ها به صورت متوالی روشن‌تر می‌شود، چنین رنگ‌هایی یک نظم ظاهری را در میان اشکال رنگ‌شده ایجاد می‌کنند که طبق تعریف هیچ ترتیبی ندارند.

بسیاری از مقیاس‌های رنگی کیفی مناسب به راحتی در دسترس هستند. شکل ۴-۱ سه نمونه مناسب را نشان می‌دهد. به طور خاص، پروژه ColorBrewer انتخاب خوبی از مقیاس‌های رنگی کیفی شامل رنگ‌های نسبتاً روشن و نسبتاً تیره را ارائه می‌دهد.[Brewer 2011]

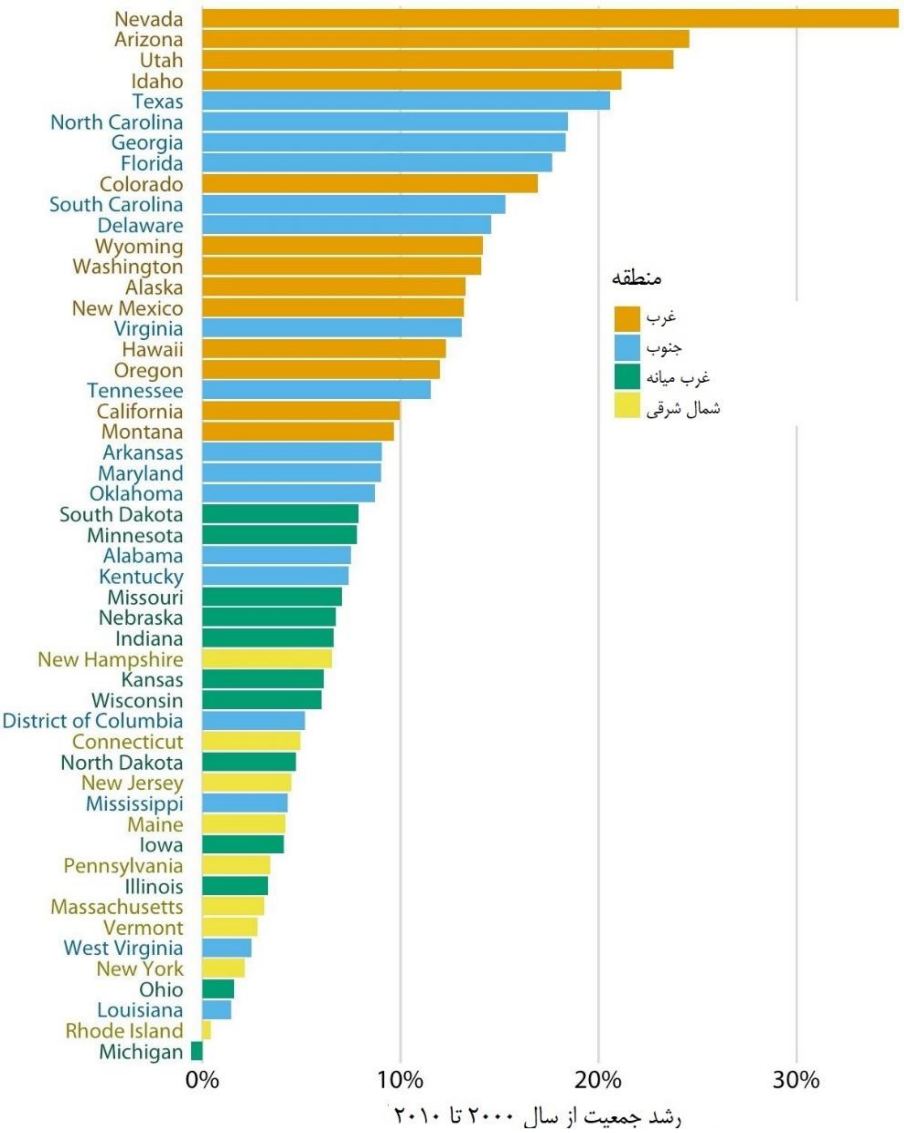


شکل ۴–۱. نمونه مقیاس‌های رنگی کیفی. مقیاس Okabe Ito مقیاس پیش فرض مورد استفاده در این کتاب است. [Okabe and Ito 2008] مقیاس ColorBrewer Dark2 توسّط پروژه ColorBrewer ارائه شده است. [Brewer 2017] مقیاس رنگی ggplot2 مقیاس کیفی پیش فرض در نرم‌افزار پرکاربرد ترسیم نمودار ggplot2 است.

به عنوان نمونه‌ای از نحوه استفاده از مقیاس‌های رنگی کیفی، نمودار ۴–۲ را در نظر بگیرید. این نمودار درصد رشد جمعیّت از سال ۲۰۰۰ تا ۲۰۱۰ را در ایالت‌های ایالات متحده نشان می‌دهد. ایالت‌ها به ترتیب رشد جمعیّت آن‌ها مرتّب شده و بر اساس منطقه جغرافیایی رنگ‌بندی شده‌اند. این رنگ‌آمیزی نشان می‌دهد که ایالت‌هایی که در منطقه جغرافیایی یکسانی قرار داشته‌اند، رشد جمعیّتی مشابهی را تجربه کرده‌اند. به ویژه، ایالت‌های غربی و جنوبی بیشترین افزایش جمعیّت را داشته‌اند، در حالی که ایالت‌های غرب میانه و شمال شرق رشد جمعیّت بسیار کمتری داشته‌اند.

رنگ برای نمایش مقادیر داده‌ها

رنگ همچنین می‌تواند برای نشان دادن مقادیر کمی داده مانند درآمد، دما یا سرعت استفاده شود. در این حالت از مقیاس رنگی متوالی استفاده می‌شود. چنین مقیاسی شامل طیفی از رنگ‌ها است که به وضوح نشان می‌دهد کدام مقادیر بزرگتر یا کوچکتر از یکدیگر هستند و دو مقدار خاص چقدر از یکدیگر فاصله دارند. مساله دیگر آن است که مقیاس رنگ باید به طور یکنواخت در کل محدوده آن تغییر کند.



نمودار ۴–۲. رشد جمعیّت در ایالات متحده از سال ۲۰۰۰ تا ۲۰۱۰. ایالت‌های غربی و جنوبی بیشترین افزایش را داشته‌اند، در حالی که ایالت‌های غرب میانه و شمال شرق شاهد افزایش بسیار کمتری (یا حتی، در مورد میشیگان، کاهش) بوده‌اند. منبع داده: اداره سرشماری ایالات متحده

مقیاس‌های متوالی را می‌توان بر اساس یک رنگ واحد (به عنوان مثال، از آبی تیره تا آبی روشن) یا چندین رنگ (به عنوان مثال، از قرمز تیره تا زرد روشن) (شکل ۴–۳) به کار برد.

مقیاس‌های چندرنگی تمایل به پیروی از طیف‌های رنگی موجود در طبیعت دارند، مانند قرمز تیره، سبز، یا آبی تا زرد روشن، یا بنفش تیره تا سبز روشن. معکوس آن (به عنوان مثال، زرد تیره تا آبی روشن) غیرطبیعی به نظر می‌رسد و مقیاس متوالی مفیدی ایجاد نمی‌کند.

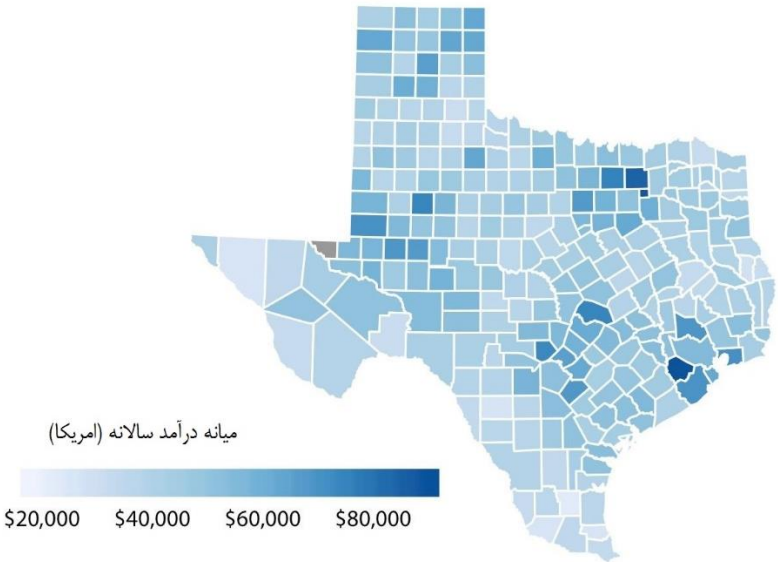
نمودار ۴–۳. نمونه‌ای از مقیاس‌های رنگی متوالی. مقیاس آبی ColorBrewer یک مقیاس تک رنگ است که از آبی تیره تا آبی روشن متغیّر است. مقیاس‌های Heat و Viridis مقیاس‌های چندرنگی هستند که به ترتیب از قرمز تیره تا زرد روشن و از آبی تیره تا سبز و در نهایت زرد روشن متفاوت هستند.

نمایش مقادیر داده‌ها به صورت رنگ به ویژه زمانی مفید است که هدف نمایش تغییر مقادیر داده‌ها در مناطق جغرافیایی متفاوت باشد. در این حالت می‌توان نقشه‌ای از مناطق جغرافیایی ترسیم نمود و آن‌ها را بر اساس مقادیر داده رنگ‌آمیزی کرد. به چنین نقشه‌هایی ناحیه–مقدار[1] می‌گویند. نمودار ۴–۴ مثالی را نشان می‌دهد که میانهٔ درآمد سالانه در هر شهرستان در تگزاس بر روی نقشه آن شهرستان‌ها ترسیم شده است.

در برخی موارد، باید انحراف مقادیر داده در یک جهت نسبت به یک نقطهٔ میانی خنثی نمایش داده شود. یک مثال ساده مجموعه داده‌ای است که شامل اعداد مثبت و منفی است. ممکن است بخواهیم آن‌ها را با رنگ‌های مختلف نشان دهیم، به طوری که فوراً مشخّص شود که یک مقدار مثبت است یا منفی و همچنین تا چه اندازه در هر جهت از صفر انحراف دارد. مقیاس رنگی مناسب در این شرایط، مقیاس رنگی واگرا است. می‌توان مقیاس واگرا را به صورت دو مقیاس متوالی که در یک نقطه میانی مشترک به هم متصل شده‌اند، تصوّر نمود که این نقطهٔ میانی معمولاً با رنگ روشن نشان داده می‌شود (نمودار ۴–۵). مقیاس‌های واگرا باید

1. choropleths

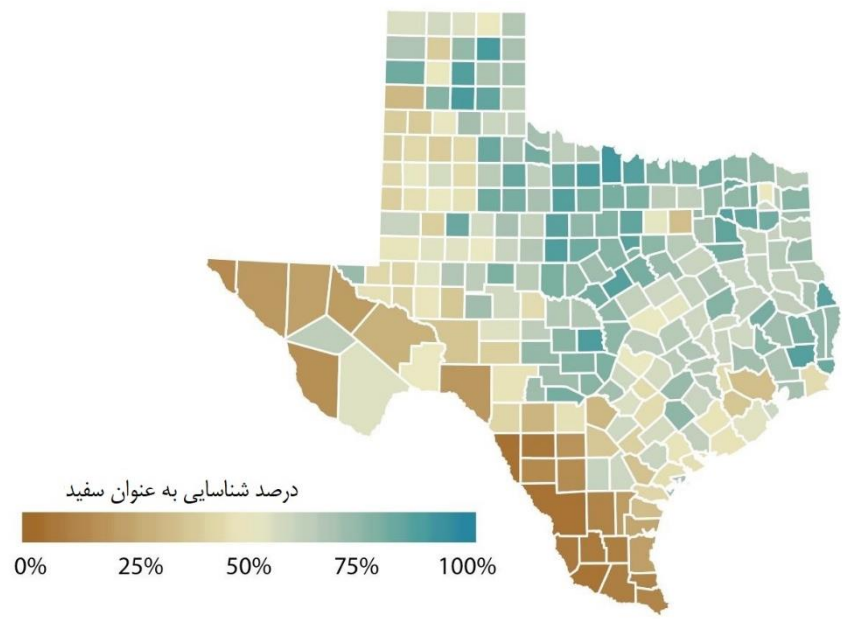
متعادل باشند، به طوری که تغییر رنگ از رنگ‌های روشن در مرکز به رنگ‌های تیره در خارج مرکز تقریباً در هر دو جهت تقریباً یکسان باشد. در غیر این صورت، بزرگی درک شده از یک مقدار داده بستگی به این دارد که بالاتر از مقدار نقطه میانی قرار می‌گیرد یا پایین‌تر.



نمودار ۴–۴. میانهٔ درآمد سالانه در شهرستان‌های تگزاس. بالاترین میانهٔ درآمد مربوط به کلانشهرهای اصلی تگزاس، به ویژه در نزدیکی هیوستون و دالاس می‌باشد. هیچ برآوردی از میانهٔ درآمد برای شهرستان لاوینگ در غرب تگزاس در دسترس نیست، بنابراین این شهرستان به رنگ خاکستری نشان داده شده است. منبع داده: نظرسنجی پنج ساله جامعه آمریکا در سال ۲۰۱۵.

نمودار ۴–۵. نمونه‌ای از مقیاس‌های رنگی واگرا. مقیاس‌های واگرا را می‌توان به صورت دو مقیاس متوالی در نظر گرفت که در یک رنگ میانی مشترک به هم متصل شده‌اند. رنگ‌های رایج برای مقیاس‌های واگرا عبارتند از: قهوه‌ای تا آبی مایل به سبز، صورتی تا زرد–سبز و آبی تا قرمز.

به عنوان مثالی از یک مقیاس رنگی واگرا، نمودار ۴-۶ که درصد افراد سفیدپوست ساکن شهرستان‌های تگزاس را نشان می‌دهد، در نظر بگیرید. اگرچه درصد همیشه یک عدد مثبت است، استفاده از مقیاس واگرا در اینجا توجیه‌پذیر است، زیرا ۵۰ درصد مقدار میانی معنی‌دار است. اعداد بالای ۵۰ درصد نشان دهندهٔ این است که اکثریت ساکنین سفیدپوست هستند و برعکس. نمودار به وضوح نشان می‌دهد که در کدام شهرستان‌ها سفیدپوستان در اکثریت هستند، در کدام شهرستان‌ها در اقلیت هستند، و در کدام شهرستان‌ها سفید پوستان و غیرسفید پوستان تقریباً مساوی هستند.



نمودار ۴-۶. درصد افرادی که در شهرستان‌های تگزاس به عنوان سفیدپوست شناخته می‌شوند. اکثریت ساکنین در شمال و شرق تگزاس سفیدپوستان می‌باشند امّا در تگزاس جنوبی یا غربی اینطور نیست. منبع داده: سرشماری ده ساله ایالات متحده در سال ۲۰۱۰.

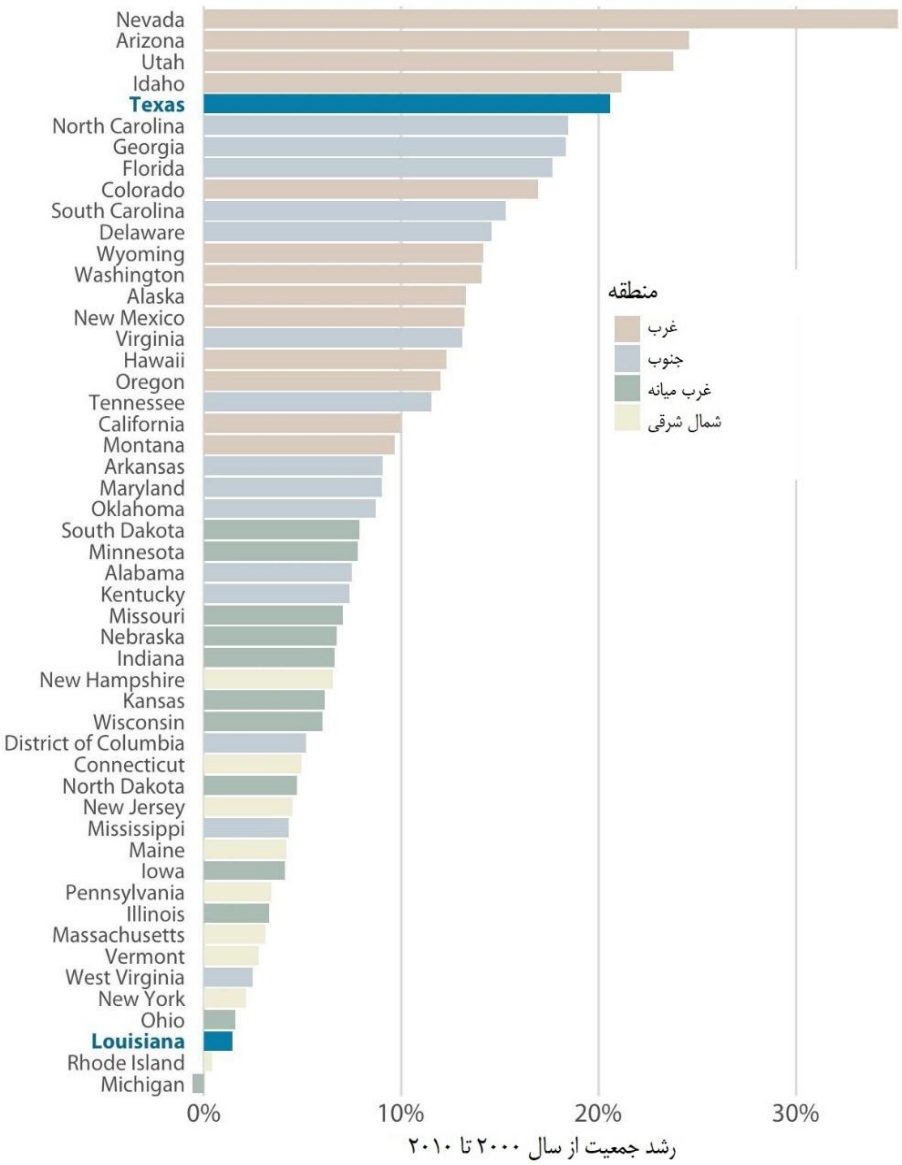
رنگ به عنوان ابزاری برای برجسته سازی

رنگ می‌تواند ابزار موثری برای برجسته سازی عناصر خاصی در داده‌ها باشد. ممکن است طبقات یا مقادیر خاصی در مجموعه داده وجود داشته باشد که حاوی اطّلاعات کلیدی دربارهٔ داستانی که قصد ارائهٔ آن را داریم باشند، لذا می‌توان با تأکید بر عناصر مرتبط در نمودار مربوطه داستان را برای خواننده شفاف‌تر نمود. یک راه آسان برای رسیدن به این هدف، رنگ‌آمیزی این عناصر با مجموعه‌ای از رنگ‌ها است که به وضوح در مقایسه با بقیه برجسته

باشند. این اثر را می‌توان با مقیاس‌های رنگ تاکیدی به دست آورد، که مقیاس‌های رنگی هستند که هم مجموعه‌ای از رنگ‌های کم‌رنگ و هم مجموعه‌ای از رنگ‌های تیره‌تر، پررنگ‌تر، و/یا رنگ‌های با اشباع بیشتر را در بر می‌گیرند. (نمودار ۴–۷)

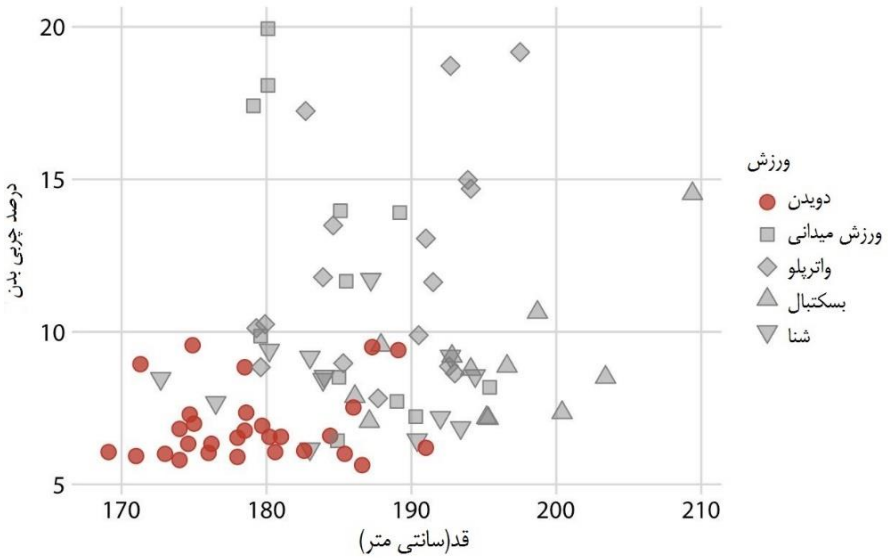
نمودار ۴–۷. نمونه‌ای از مقیاس رنگ تاکیدی، که هرکدام دارای چهار رنگ پایه و سه رنگ تاکیدی هستند. مقیاس‌های رنگ تاکیدی را می‌توان به روش‌های مختلفی استخراج کرد: (بالا) می‌توان از یک مقیاس رنگی موجود استفاده نمود (به عنوان مثال، مقیاس Okabe Ito، (نمودار ۴–۱) و برخی از رنگ‌ها را روشن و/یا تا حدی غیراشباع و برخی دیگر را تیره نمود.(وسط)؛ می‌توان مقادیر خاکستری را انتخاب و آن‌ها را با رنگ‌ها جفت نمود. (پایین)؛ می‌توان از یک مقیاس رنگ تاکیدی موجود استفاده نمود (به عنوان مثال، یکی از موارد موجود در پروژه ColorBrewer)

به عنوان نمونه‌ای از اینکه چگونه داده‌های یکسان می‌توانند از داستان‌های مختلف توسّط رویکردهای رنگ‌آمیزی متفاوت پشتیبانی کنند، نسخه‌ای از نمودار ۴–۲ رسم شده که در آن دو ایالت خاص، تگزاس و لوئیزیانا برجسته شده‌اند (نمودار ۴–۸). هر دو ایالت در جنوب بوده و همسایه هستند، با این حال یک ایالت (تگزاس) پنجمین ایالت از نظر سریع‌ترین نرخ رشد در ایالات متحده از سال ۲۰۰۰ تا ۲۰۱۰ بود، در حالی که دیگری سوّمین ایالت با کندترین نرخ رشد بود.



نمودار ۴–۸. طی سال ۲۰۰۰ تا ۲۰۱۰، دو ایالت جنوبی همسایه، تگزاس و لوئیزیانا، جزو ایالت‌های با بالاترین و پایین‌ترین نرخ رشد جمعیّت در سراسر ایالات متحده بودند. منبع داده ها: اداره سرشماری ایالات متحده

هنگام کار با رنگ‌های تاکیدی، بایستی دقّت شود که رنگ‌های پایه تقریبا مشابه بوده و برای جلب توجّه رقابت نکنند. توجّه کنید که رنگ‌های پایه در شکل ۴–۸ چقدر یکنواخت و کسل‌کننده هستند، امّا به عنوان رنگ تاکیدی عملکرد قابل قبولی دارند. ممکن است به سادگی در دام استفاده از رنگ‌های پایه رنگارنگ افتاد، به طوری که این رنگ‌ها در نهایت برای جلب توجّه خواننده در برابر رنگ‌های تاکیدی رقابت می‌کنند. با این حال، یک راه حل آسان هم وجود دارد: تمام رنگ‌ها از تمام عناصر موجود در شکل حذف شده و فقط رنگ دسته‌ها یا نقاط داده‌های مورد نظر حفظ شوند. نمونه‌ای از این راهبرد در نمودار ۴–۹ ارائه شده است.



نمودار ۴–۹. ورزشکاران دونده از جمله کوتاه قدترین و لاغرترین ورزشکاران حرفه‌ای مرد هستند که در ورزش‌های محبوب شرکت می‌کنند. منبع داده: Telford and Cunningham 1991

فصل ۵

فهرست راهنمای نمودارها

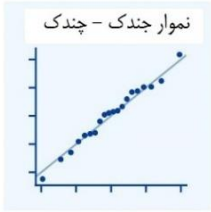
این فصل یک نمای کلی بصری سریع از نمودارهای مختلفی که معمولاً برای تجسم انواع مختلف داده‌ها استفاده می‌شوند، ارائه می‌دهد. در صورتی که به دنبال ترسیم نمودار خاصی هستید که ممکن است نام آن را ندانید، این فصل به‌عنوان فهرست مطالب کمک‌کننده خواهد بود. همچنین در صورتی که نیاز به جایگزین‌هایی برای نمودارهایی که به‌طور معمول ترسیم می‌کنید دارید، این فصل برایتان منبع الهام بخشی خواهد بود.

مقادیر

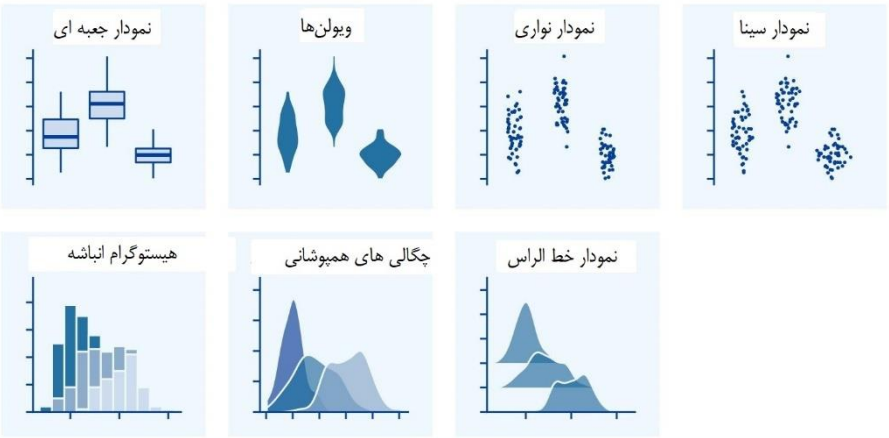
رایج‌ترین رویکرد برای تجسم مقادیر (یعنی مقادیر عددی نشان داده شده برای برخی از دسته‌ها) استفاده از چینش عمودی یا افقی میله‌ها است (فصل ۶). با این حال، به جای استفاده از میله‌ها، می‌توانیم نقطه‌هایی را در محلی که میلهٔ مربوطه به پایان می‌رسد، قرار دهیم (فصل ۶).

اگر دو یا چند مجموعه از دسته‌ها وجود دارد که می‌خواهیم مقادیر را برای آن‌ها نشان دهیم، می‌توانیم میله‌ها را گروه‌بندی یا روی هم قرار دهیم (فصل ۶). همچنین می‌توانیم دسته‌ها را بر روی محورهای x و y ترسیم کنیم و مقادیر را با رنگ، توسّط نقشه حرارتی نشان دهیم (فصل ۶).

توزیع‌ها

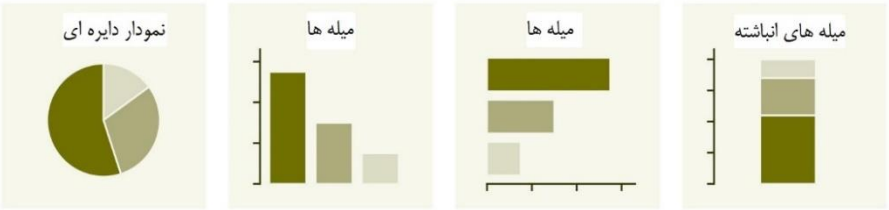


هیستوگرام‌ها و نمودارهای چگالی (فصل ۷) شهودی‌ترین نحوهٔ نمایش یک توزیع را ارائه می‌دهند، امّا هر دو نیازمند انتخاب سلیقه‌ای برخی پارامترها بوده و می‌توانند گمراه‌کننده باشند. چگالی تجمّعی و نمودارهای چندک–چندک (q-q) (فصل ۸) همیشه داده‌ها را صادقانه نشان می‌دهند، امّا تفسیر آن‌ها دشوارتر است.



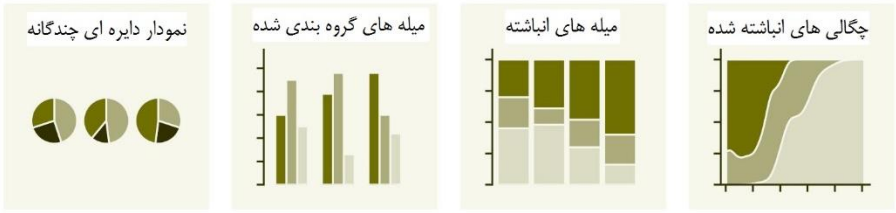
نمودار جعبه‌ای، طرح‌های ویولن، نمودارهای نواری و نمودارهای سینا زمانی مفید هستند که می‌خواهیم توزیع‌های زیادی را به‌طور همزمان نمایش دهیم و/یا اگر در درجهٔ اوّل به تغییرات کلی بین توزیع‌ها علاقه‌مندیم (به «نمایش توزیع‌ها در امتداد محور عمودی» مراجعه کنید). هیستوگرام‌های انباشته و چگالی‌های همپوشان امکان مقایسهٔ عمیق‌تر تعداد کمتری از توزیع‌ها را فراهم می‌کنند، اگرچه تفسیر هیستوگرام‌های انباشته ممکن است دشوار باشد و بهتر است از آن‌ها اجتناب شود (به «نمایش همان زمان چند توزیع» مراجعه کنید). نمودارهای خط الراس می‌توانند جایگزین مفیدی برای نمودارهای ویولن باشند و اغلب هنگام نمایش توزیع‌های متعدد یا تغییرات در توزیع‌ها در طول زمان مفید هستند (به «نمایش توزیع‌ها در امتداد محور افقی» مراجعه کنید).

نسبت‌ها



نسبت‌ها را می‌توان به صورت نمودار دایره‌ای، میله‌های کنار هم، یا میله‌های روی هم نمایش داد (فصل ۱۰). همانند نحوهٔ نمایش مقادیر، وقتی نسبت‌ها را با میله‌ها نمایش می‌دهیم،

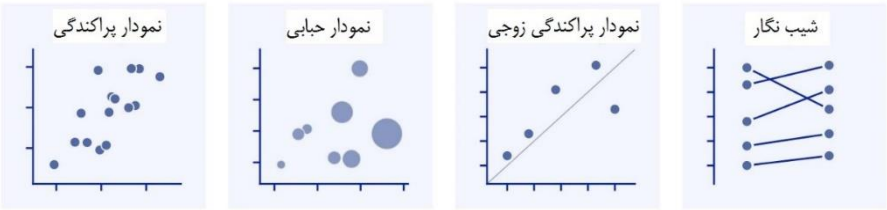
میله‌ها را می‌توان به صورت عمودی یا افقی ترسیم کرد. نمودارهای دایره‌ای بر این مساله که تک تک طبقات یک کل را می‌سازند، تاکید دارند و کسرهای ساده را برجسته می‌کنند. با این حال، تک تک طبقات با میله‌های کنار هم راحت‌تر قابل مقایسه خواهند بود. میله‌های روی هم برای یک مجموعه از نسبت‌ها نامناسب به نظر می‌رسند، امّا هنگام مقایسهٔ چندین مجموعه از نسبت‌ها مفید هستند.



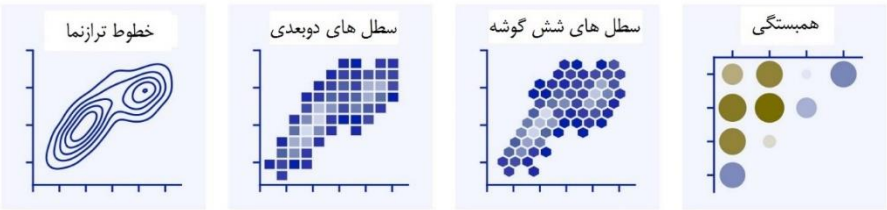
هنگام نمایش مجموعه‌های متعددی از نسبت‌ها یا تغییرات در نسبت‌ها، نمودارهای دایره‌ای معمولاً ناکارآمد بوده و اغلب روابط موجود را می‌پوشانند. میله‌های گروه‌بندی شده تا زمانی که تعداد وضعیت‌های مقایسه شده متوسط باشد، کارایی مناسبی دارند، در حالی که میله‌های روی هم می‌توانند برای مقایسه وضعیت‌های متعدد مفید باشند. چگالی‌های انباشته (فصل ۱۰) زمانی مناسب هستند که نسبت‌ها مبتنی بر یک متغیّر پیوسته تغییر کنند.

وقتی نسبت‌ها بر اساس متغیّرهای گروه‌بندی چندگانه مشخّص می‌شوند، نمودارهای موزاییکی، نقشه‌های درختی یا مجموعه‌های موازی رویکردهای مفیدی هستند (فصل ۱۱). پیش فرض نمودارهای موزاییکی بر این است که هر سطح از یک متغیّر گروه‌بندی را می‌توان با هر سطح از متغیّر گروه‌بندی دیگر ترکیب کرد، در حالی که نقشهٔ درختی چنین پیش فرضی را ندارند. حتی اگر زیربخش‌های یک گروه کاملاً از زیربخش‌های گروه دیگر متمایز باشند، نقشه‌های درختی همچنان مناسب هستند. هنگامی که بیش از دو متغیّر گروه‌بندی وجود دارد، عملکرد مجموعه‌های موازی بهتر از طرح‌های موزاییکی یا نقشه‌های درختی می‌باشد.

روابط x-y



زمانی که می‌خواهیم یک متغیّر کمی را نسبت به دیگری نشان دهیم، نمودارهای پراکنش (فصل ۱۲) نمایش کهن الگو را فراهم می‌کنند. اگر سه متغیّر کمی داشته باشیم، می‌توانیم یک متغیّر را روی اندازهٔ نقاط سوار نموده، و نوعی از نمودار پراکنش به نام نمودار حباب ایجاد کنیم. برای داده‌های زوجی، زمانی که متغیّرهای محورهای x و y با واحدهای یکسان اندازه‌گیری می‌شوند، افزودن یک خط نشان دهندهٔ x = y مفید است (به «داده‌های زوجی» مراجعه کنید). داده‌های زوجی را می‌توان به صورت یک شیب‌نگار از جفت نقاطی که با خطوط مستقیم به هم متصل شده‌اند نشان داد.



برای تعداد زیادی از نقاط، نمودارهای پراکنش معمولی به دلیل ترسیم بیش از حد اطلاعات، ممکن است ناکارآمد شوند. در این حالت، خطوط ترازنما[1]، سطل‌های دوبعدی[2]، یا سطل‌های شش ضلعی[3] ممکن است جایگزین مناسبی باشند (فصل ۱۸). از سوی دیگر، وقتی می‌خواهیم بیش از دو کمیّت را نمایش دهیم، ممکن است نمایش ضرایب همبستگی در قالب یک همبستگی نگار[4] به جای داده‌های خام زیربنایی، ارجح باشد (به «همبستگی نگار» مراجعه کنید).

1. contour lines
2. 2D bins
3. hex bins
4. correlogram

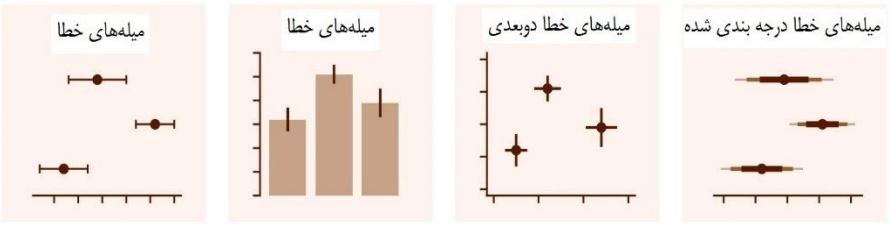
هنگامی که محور x نشان‌دهندهٔ زمان یا یک کمیّت صرفاً فزاینده مانند دوز درمان است، معمولاً نمودارهای خطی ترسیم می‌کنیم (فصل ۱۳). اگر توالی زمانی از دو متغیّر پاسخ داشته باشیم، می‌توانیم یک نمودار پراکنش متصل رسم کنیم، که در آن ابتدا دو متغیّر پاسخ را در یک نمودار پراکنش رسم نموده و سپس نقاط مربوط به بازه‌های زمانی مجاور را به هم وصل می‌کنیم (به «سری‌های زمانی دو یا چند متغیّر پاسخ» مراجعه کنید.) همچنین می‌توانیم از خطوط صاف برای نمایش روندها در مجموعه داده‌های بزرگتر استفاده کنیم (فصل ۱۴).

داده‌های جغرافیایی

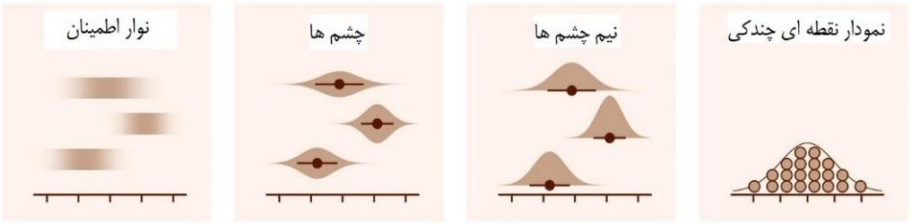
حالت اولیه نمایش داده‌های مکانی به صورت نقشه است (فصل ۱۵). نقشه، مختصّات روی کرهٔ زمین را گرفته و آن‌ها را روی یک سطح صاف ترسیم می‌کند، به طوری که اشکال و فواصل روی کرهٔ زمین تقریباً با اشکال و فواصل در نمایش دو بُعدی متناسب است. علاوه بر این، می‌توانیم مناطق مختلف را با رنگ‌آمیزی مبتنی بر مقادیر یک متغیّر دیگر نشان دهیم. چنین نقشه‌ای ناحیه–مقدار نامیده می‌شود (به «نقشه برداری ناحیه–مقدار» مراجعه کنید). در برخی موارد، تغییر مناطق مختلف بر اساس متغیّر دیگری (مثلاً تعداد جمعیّت) یا ساده کردن هر منطقه به شکل یک مربع ممکن است مفید باشد. به این گونه نمودارها، نقشه آماری[1] می‌گویند (به «نقشه آماری» مراجعه کنید).

1. cartograms

عدم قطعیّت



میله‌های خطا برای نشان دادن محدودهٔ مقادیر محتمل برای تخمین‌ها یا اندازه‌گیری‌ها هستند. آن‌ها به صورت افقی و/یا عمودی از نقطه مرجعی که تخمین یا اندازه‌گیری را نشان می‌دهد گسترش می‌یابند (فصل ۱۶). نقاط مرجع را می‌توان به روش‌های مختلفی مانند نقطه یا میله نشان داد. میله‌های خطای درجه‌بندی شده چندین محدوده را به طور همزمان نشان می‌دهند، که در آن هر محدوده مربوط به دامنهٔ اطمینان متفاوتی است. آن‌ها در واقع میله‌های خطای متعددی با ضخامت‌های متفاوت هستند که روی هم ترسیم شده‌اند.



برای دستیابی به نمایش دقیق‌تری نسبت به آنچه با میله‌های خطا یا میله‌های خطا درجه‌بندی شده امکان‌پذیر است، می‌توانیم اطمینان واقعی یا توزیع‌های پسین را ترسیم کنیم (فصل ۱۶). نوارهای اطمینان حس بصری عدم قطعیّت را فراهم می‌کنند امّا خواندن دقیق آن‌ها دشوار است. چشم‌ها و نیم‌چشم‌ها نوارهای خطا را با رویکردهایی برای نمایش توزیع‌ها (به ترتیب ویولن و خطوط برآمدگی) ترکیب می‌کنند و بنابراین هم محدودهٔ دقیق سطوح اطمینان و هم توزیع کلی عدم قطعیّت را نشان می‌دهند. نمودار چندک نقطه‌ای می‌تواند به عنوان یک روش نمایش جایگزین برای توزیع عدم قطعیّت به کار رود (به «کادربندی احتمالات به صورت فراوانی» مراجعه کنید). از آنجایی که نمودار چندک نقطه‌ای توزیع را در واحدهای گسسته نشان می‌دهد، چندان دقیق نیست، امّا خواندن آن نسبت به توزیع پیوسته نشان داده شده توسّط نمودار ویولن یا خط الراس آسان‌تر است.

معادل میله خطا برای نمودارهای خط صاف، نوار اطمینان است (به «نمایش عدم قطعیّت در برازش‌های منحنی» مراجعه کنید). نوار اطمینان محدوده‌ای از مقادیری را نشان می‌دهد که خط می‌تواند در سطح اطمینان مشخصی در آن قرار گیرد. همانند میله‌های خطا، می‌توانیم نوارهای اطمینان درجه‌بندی شده‌ای را ترسیم کنیم که چندین سطح اطمینان را به طور همزمان نشان می‌دهند. همچنین می‌توانیم به‌جای نوارهای اطمینان یا علاوه بر آن، خطوط منفرد برازش شده را ترسیم کنیم.

فصل ٦

ترسیم مقدارها

در بسیاری از مواقع، می‌خواهیم بزرگی مجموعه‌ای از مقادیر را نشان دهیم. برای مثال، ممکن است بخواهیم حجم کل فروش برندهای متفاوت خودرو، یا تعداد کل افرادی که در شهرهای مختلف زندگی می‌کنند و یا سن بازیکنان المپیک که ورزش‌های مختلفی انجام می‌دهند را نمایش دهیم. در تمام این موارد، یک متغیّر کیفی (مانند: برند ماشین، شهرها، ورزش‌ها) و یک متغیّر کمی برای هر دسته داریم. به این موارد ترسیم مقدارها اطلاق می‌کنیم زیرا تاکید اصلی در این نمودارها بر اندازه متغیّر کمی است. انتخاب استاندارد در این شرایط، نمودار میله‌ای است که تنوع زیادی دارد و شامل نمودار میله‌ای ساده، گروه‌بندی شده و انباشته می‌باشد. از جایگزین‌های نمودار میله‌ای، می‌توان به نمودار نقطه‌ای یا نقشه حرارتی اشاره کرد.

نمودار میله‌ای[1]

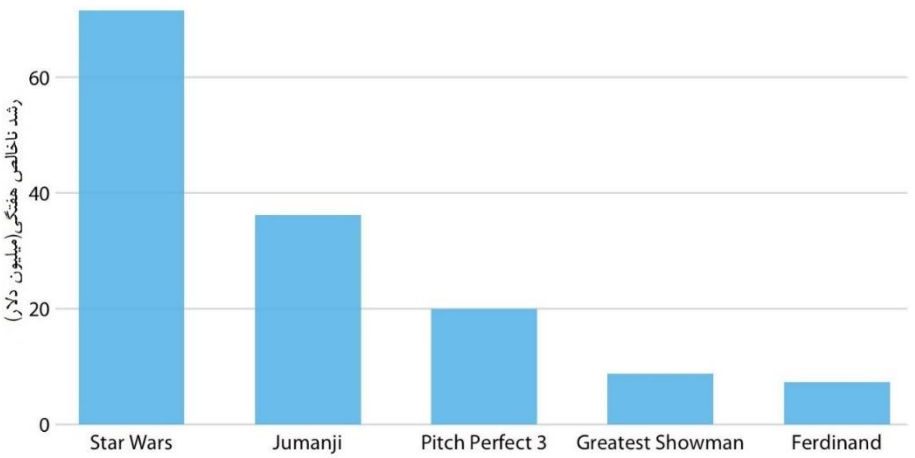
برای تصوّر بهتر نمودار میله‌ای، تعداد کل فروش بلیت‌های معروف‌ترین فیلم آخر هفته را در نظر بگیرید. در جدول ۶-۱، پنج فیلم با بیشترین فروش در تعطیلات قبل از کریسمس ۲۰۱۷ نشان داده شده است. پر فروش‌ترین فیلم با فاصله زیاد، فیلم جنگ ستارگان است که فروش آن تقریبا ۱۰ برابر دو فیلم آخر در رتبه‌های چهارم و پنجم لیست می‌باشد.

1. Bar Plots

جدول ۶–۱ لیست پر فروش‌ترین فیلم‌های هفته منتهی به ۲۲–۲۴ دسامبر ۲۰۱۷. منبع داده: Box Office Mojo

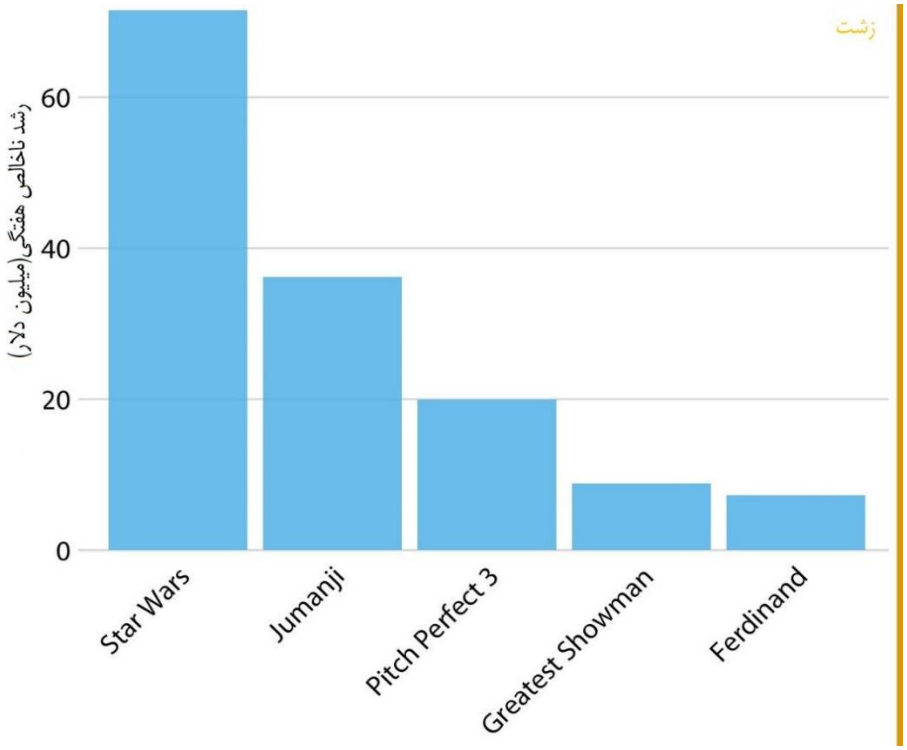
Rank	Title	Weekend gross
1	*Star Wars: The Last Jedi*	$71,565,498
2	*Jumanji: Welcome to the Jungle*	$36,169,328
3	*Pitch Perfect 3*	$19,928,525
4	*The Greatest Showman*	$8,805,843
5	*Ferdinand*	$7,316,746

این نوع داده معمولاً با میله‌های عمودی رسم می‌شوند. برای هر فیلم یک ستون رسم می‌شود که از صفر شروع شده و تا مقدار فروش آن بر حسب دلار بالا می‌رود (شکل ۶–۱). شکل حاصل، نمودار میله‌ای می‌باشد.



نمودار ۶–۱. لیست پر فروش‌ترین فیلم‌های هفته منتهی به ۲۲–۲۴ دسامبر ۲۰۱۷ که به صورت نمودار میله‌ای نشان داده شده است. منبع داده: Box Office Mojo

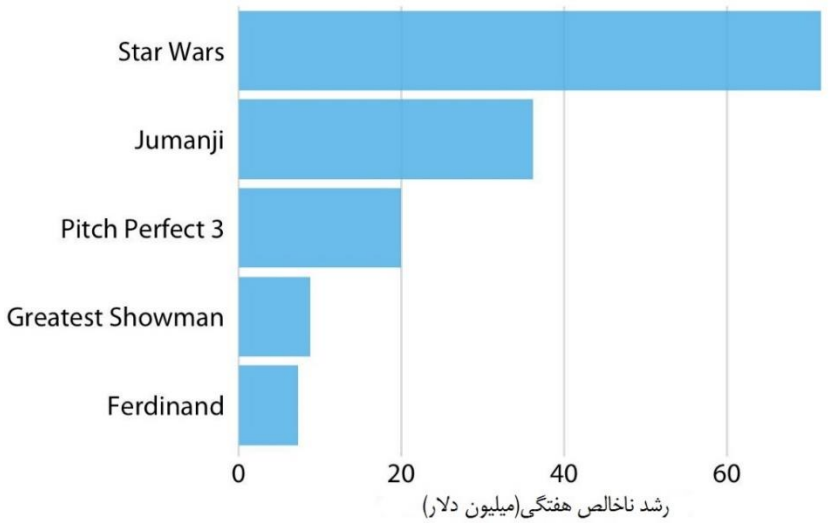
یک مشکل رایج در ترسیم نمودار میله‌ای عمودی این است که نام میله‌ها فضای افقی زیادی را اشغال می‌کنند. در شکل ۶–۱ فضای زیادی بین میله‌ها خالی می‌ماند تا بتوان نام فیلم‌ها را در زیر آن‌ها نوشت. برای حفظ فضای افقی، می‌توان میله‌ها را نزدیک‌تر به هم گذاشت و عنوان آن‌ها را چرخاند (شکل ۶–۲). البته این حالت پیشنهاد نمی‌شود، ظاهر نمودار نامناسب به نظر می رسد و خواندن آن را سخت می‌کند. طبق تجربه، وقتی عنوان‌ها آنقدر طولانی هستند که نمی‌توان آن‌ها را به صورت افقی نوشت، چرخاندن آن‌ها هم نمای مناسبی به دست نخواهد داد.



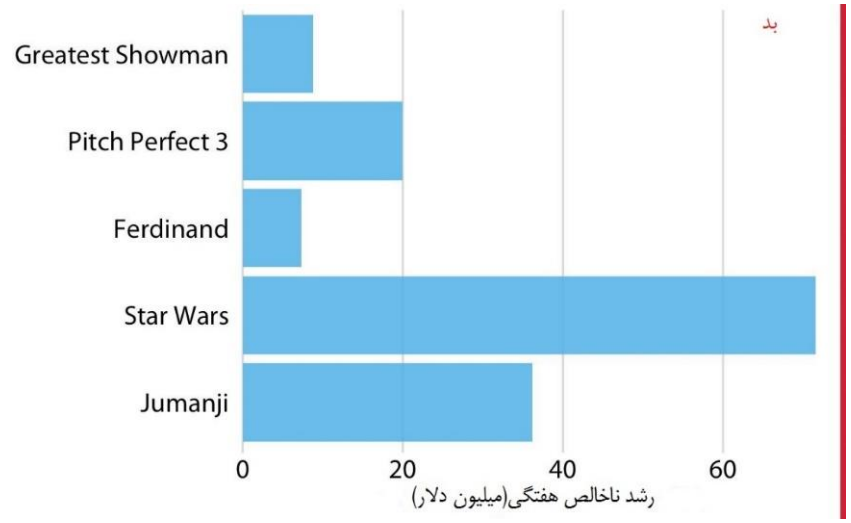
نمودار ۶–۲. لیست پر فروش‌ترین فیلم‌های هفته منتهی به ۲۲–۲۴ دسامبر ۲۰۱۷ که به صورت نمودار میله‌ای با چرخش عناوین محور افقی نشان داده شده است. خواندن عناوین چرخانده شده در محور افقی دشوار بوده و نیازمند فضای زیادی در پایین محور افقی می‌باشد. به این دلایل عمدتاً نمودارهای این چنینی زیبایی بصری ندارند.
منبع داده: Box Office Mojo

بهترین راه حل برای عنوان‌های طولانی، عوض کردن جای محور x و y است تا میله‌ها افقی شوند (شکل ۶–۳). بعد از جا به جا کردن محورها، شکلی فشرده به دست می‌آید که در آن تمام اجزای بصری از جمله متن‌ها، در جهت افقی می‌باشند. در نتیجه خواندن نمودار از نمودار ۶–۲ و حتّی ۶–۱ راحت‌تر می‌باشد.

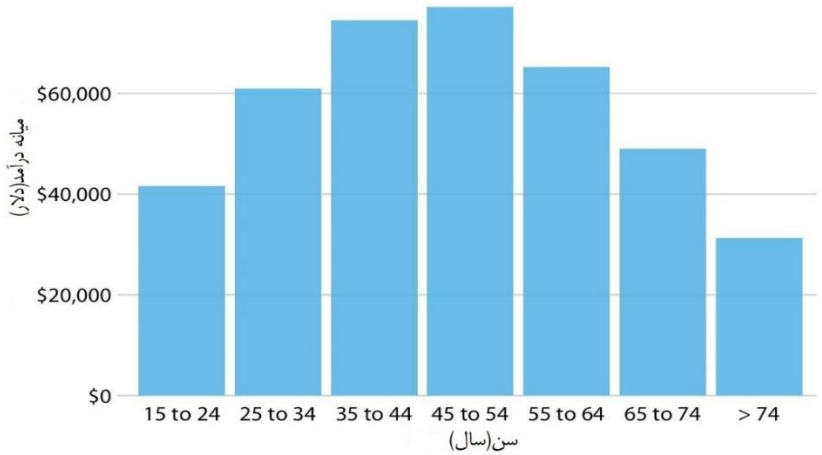
جدا از اینکه میله‌ها به صورت افقی یا عمودی هستند، باید به ترتیبی که میله‌ها قرار می‌گیرند توجّه کرد. معمولاً میله‌ها به صورت دلخواه چیده می‌شوند یا بر اساس معیاری که با توجّه به مضمون نمودار بدون معنی است. بعضی برنامه‌های ترسیم نمودار، میله‌ها را بر اساس حروف الفبایی عنوان آن‌ها یا سایر روش‌ها ترسیم می‌کنند (شکل ۶–۴). این گونه نمودارها نسبت به نمودارهایی که میله‌ها براساس ترتیب اندازه‌شان مرتّب شده‌اند، گیج‌کننده‌تر و نامفهوم‌تر می‌باشند.



نمودار ۶-۳. لیست پر فروش‌ترین فیلم‌های هفته منتهی به ۲۲-۲۴ دسامبر ۲۰۱۷ که به صورت نمودار میله‌ای افقی نشان داده شده است. منبع داده: Box Office Mojo

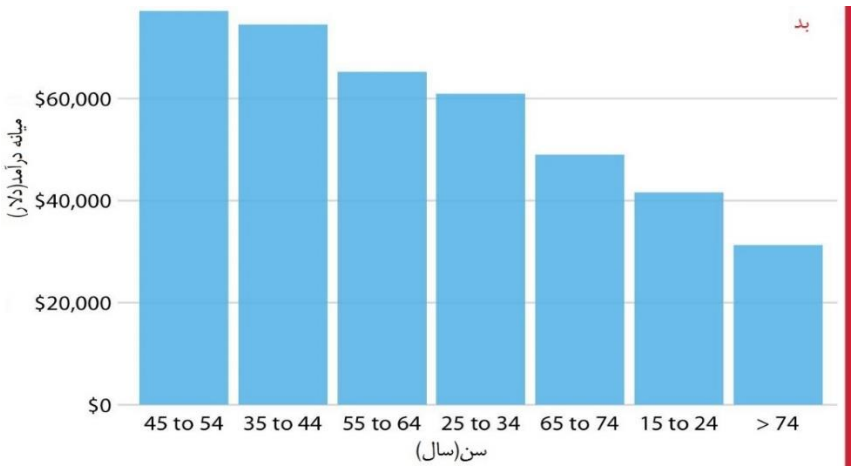


نمودار ۶-۴. لیست پر فروش‌ترین فیلم‌های هفته منتهی به ۲۲-۲۴ دسامبر ۲۰۱۷ که به صورت نمودار میله‌ای افقی نشان داده شده است. در اینجا میله‌ها به ترتیب نزولی طول عنوان فیلم مرتّب شده‌اند. این ترتیب دلخواهی بوده و هدف مشخصی را دنبال نمی‌کند و نمودار حاصل نسبت به نمودار ۶-۳ مبهم‌تر است.
منبع داده: Box Office Mojo



نمودار ۶–۵. میانهٔ درآمد سالانه خانوار در ایالات متحده امریکا در سال ۲۰۱۶ به تفکیک گروه سنی. گروه سنی ۴۵ تا ۵۴ سال بالاترین میانهٔ درامدی را دارند. منبع داده: اداره سرشماری ایالات متحده امریکا

وقتی ترتیب ذاتی در طبقه‌بندی میله‌ها وجود نداشته باشد، باید میله‌ها مجددا مرتّب شوند. وقتی ترتیب ذاتی وجود داشته باشد (مثلاً وقتی متغیّر کیفی یک عامل طبقه‌بندی شده است) باید طبقه‌بندی مذکور در ترسیم نمودار مدنظر باشد. برای مثال، شکل ۶–۵ میانهٔ درآمد سالانه در آمریکا را بر اساس گروه سنی نشان می‌دهد. در این مورد، میله‌ها باید بر اساس افزایش سن مرتّب شوند. مرتّب سازی میله‌ها بر اساس ارتفاع آن‌ها که سبب بهم ریختن گروه‌های سنی می‌شود، منطقی نیست (شکل ۶–۶).



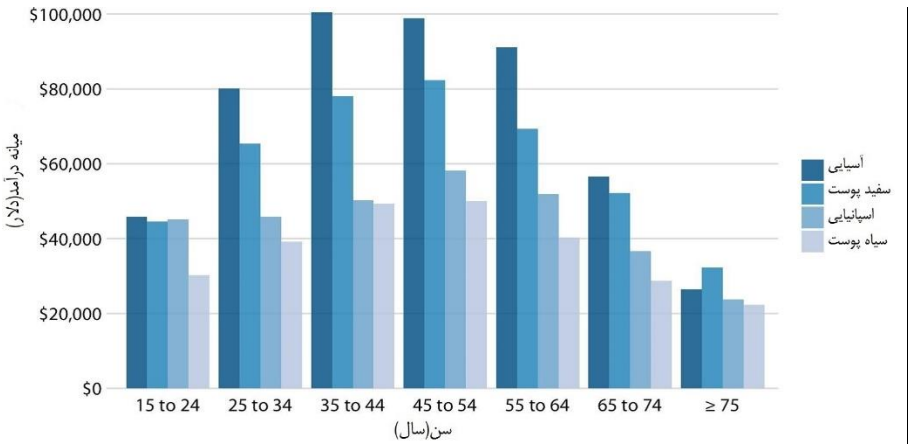
نمودار ۶–۶. میانهٔ درآمد سالانه خانوار در ایالات متحده امریکا در سال ۲۰۱۶ به تفکیک گروه سنی که بر اساس درآمد مرتّب شده است. در حالی که ترتیب نزولی میله‌ها به نظر منطقی می‌رسد امّا ترتیب گروه سنی گیج‌کننده است. منبع داده: اداره سرشماری ایالات متحده امریکا

به ترتیب میله‌ها توجّه کنید. اگر میله‌ها نشان دهنده متغیّری غیرترتیبی هستند، آن‌ها را به ترتیب صعودی یا نزولی مقادیر داده‌ها مرتّب کنید.

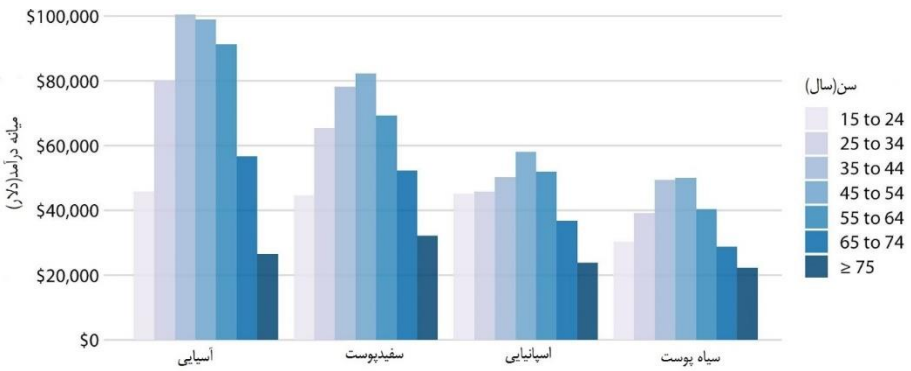
نمودارهای میله‌ای گروه‌بندی شده و انباشته

در مثال‌های پیش دیدیم که چگونه یک مقدار کمّی بر اساس یک متغیّر کیفی تغییر می‌کرد. اگرچه عمدتاً مایل به بررسی دو متغیّر طبقه‌بندی شده به صورت همزمان هستیم. برای مثال، ادارهٔ سرشماری آمریکا، میانهٔ سطح درآمد را بر اساس سن و نژاد گزارش می‌کند. این داده‌ها را می‌توان با نمودار میله‌ای گروه‌بندی شده رسم نمود (شکل ۶–۷). در نمودار میله‌ای گروه‌بندی شده، گروهی از میله‌ها در محور x بر اساس یک متغیّر کیفی مشخّص می‌شود. سپس میله‌ها را در هر گروه بر اساس متغیّر کیفی دیگر طبقه‌بندی می‌شوند.

نمودار میله‌ای گروه‌بندی شده، اطّلاعات زیادی را در یک نگاه نشان می‌دهد که می‌تواند گیج‌کننده باشد. اگرچه در شکل ۶–۷ از کلمه بد یا زشت استفاده نشده امّا خواندن آن سخت است. در واقع مقایسه کردن میانهٔ درآمد در گروه‌های سنی مختلف برای یک گروه نژادی خاص سخت است. پس این نمودار تنها مناسب زمانی است که فقط می‌خواهیم تفاوت سطوح درآمد گروه‌های نژادی مختلف را در یک گروه سنی خاص مقایسه کنیم. اگر بیشتر به الگوی کلی سطوح درآمد در گروه‌های نژادی اهمیت می‌دهیم، بهتر است نژاد را در محور x و سن را با میله‌های مجزا درون گروه‌های نژادی نمایش دهیم (شکل ۶–۸).



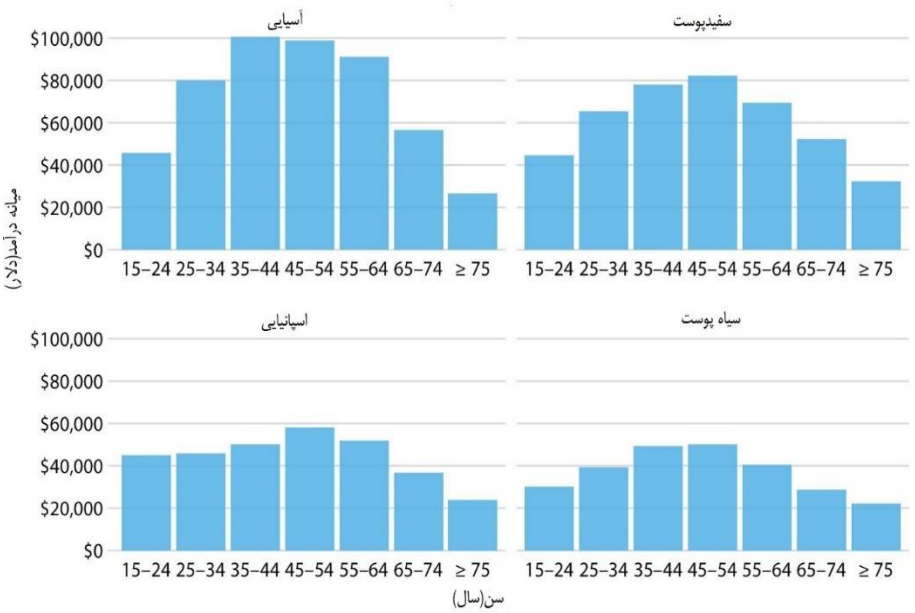
نمودار ۶–۷. میانهٔ درامد سالانهٔ خانوار در ایالات متحده امریکا در سال ۲۰۱۶ به تفکیک گروه سنی و نژاد. گروه‌های سنی در امتداد محور x رسم شده و برای هر گروه سنی ۴ میله رسم شده است که به ترتیب نشان دهندهٔ میانهٔ درآمد در نژاد آسیایی، سفیدپوست، اسپانیایی و سیاه پوست می‌باشد. منبع داده: اداره سرشماری ایالات متحده امریکا



نمودار ۶–۸ میانهٔ درآمد سالانهٔ خانوار در ایالات متحده امریکا در سال ۲۰۱۶ به تفکیک گروه سنی و نژاد. برخلاف نمودار ۶–۷ در اینجا نژاد در امتداد محور x نشان داده شده و برای هر نژاد، ۷ میله متناظر با گروه‌های سنی رسم شده است. منبع داده: اداره سرشماری ایالات متحده امریکا

هر دو شکل ۶–۷ و ۶–۸ یک متغیّر ترتیبی را در محور x نشان داده و متغیّر دیگر را بر اساس رنگ میله مشخّص می‌کنند. در هر دو مورد، خواندن بر اساس موقعیّت راحت‌تر از رنگ است که نیاز به تلاش فکری بیشتری دارد زیرا باید رنگ میله‌ها را بر اساس راهنما شناسایی کنیم. می‌توان از این بار فکری، با ساختن ۴ نمودار میله‌ای بجای یک نمودار میله‌ای گروه‌بندی شده پیشگیری کرد (شکل ۶–۹). این انتخاب سلیقه‌ای است امّا انتخاب شکل ۶–۹، مساله نیاز به رنگ‌های مختلف را برطرف می‌کند.

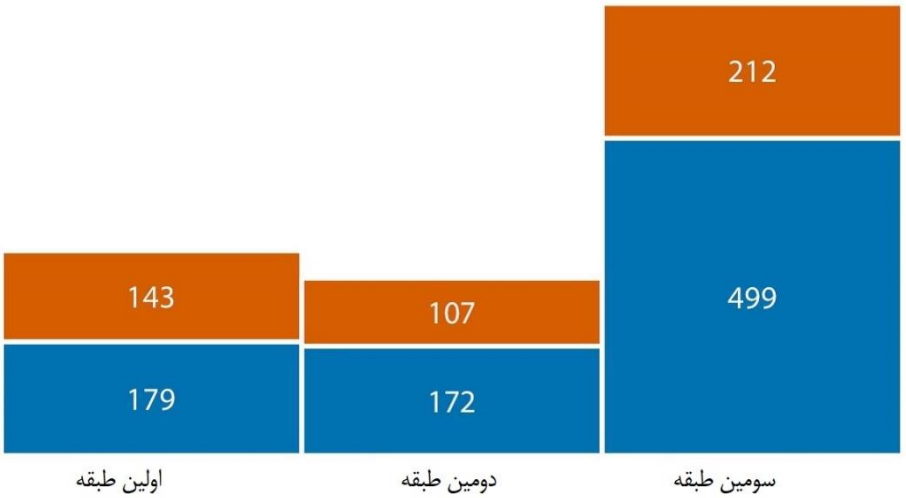
بجای رسم گروهی از میله‌ها در کنار هم، گاهی بهتر است آن‌ها را روی هم انباشته کرد. انباشته کردن وقتی خوب است که جمع مقادیر میله‌ها با هم، یافته معناداری باشند. پس اگرچه انباشته کردن مقادیر میانهٔ درآمد در شکل ۶–۷ بر روی هم معنی‌دار نیست (جمع دو میانهٔ درآمد، مقدار معنی‌داری نیست) امّا انباشته کردن مقادیر فروش آخر هفته در شکل ۶–۱ منطقی می‌باشد (جمع مقدار فروش آخر هفته برای ۲ فیلم مقدار معنی‌داری است). همچنین انباشته کردن مناسب زمانی است که هر میلهٔ مجزا نمایانگر فراوانی است. برای مثال، در مجموعه داده‌ای از افراد، می‌توان مرد و زن را جداگانه یا با هم شمارش کرد. اگر یک ستون را که نمایانگر تعداد زنان است را بر ستون نمایانگر تعداد مردان قرار دهیم، ارتفاع میله انباشته، تعداد کل افراد را فارغ از جنسیّت نشان می‌دهد.



نمودار ۶–۹. میانهٔ درآمد سالانهٔ خانوار در ایالات متحده امریکا در سال ۲۰۱۶ به تفکیک گروه سنی و نژاد. به جای نمایش داده‌ها به صورت نمودار میله‌ای گروه‌بندی شده همانند نمودارهای ۶–۷ و ۶–۸، نمودار در قالب چهار نمودار میله‌ای معمولی مجزا نشان داده شده است. این انتخاب این مزیّت را دارد که دیگر نیازی به استفاده از رنگ برای نمایش متغیّر کیفی نیست. منبع داده: اداره سرشماری ایالات متحده امریکا

این اصل را با اطّلاعات مسافران کشتی اقیانوس اطلس پیمای تایتانیک که در ۱۵ آوریل ۱۹۱۲ غرق شد بررسی می‌کنیم. این کشتی بجز خدمه، ۱۳۰۰ نفر مسافر داشت که در سه طبقهٔ اقتصادی اوّل، دوّم و سوّم قرار داشتند و تعداد مردان تقریبا دو برابر زنان بود. برای رسم تقسیم‌بندی مسافران بر اساس طبقهٔ اقتصادی و جنسیّت، می‌توان میله‌های مختلف برای هر طبقهٔ اقتصادی و جنسیّت رسم نمود و برای هر طبقهٔ اقتصادی، میله‌های مربوط به زنان را در بالای میله‌های مربوط به مردان انباشته کرد (شکل ۶–۱۰). ارتفاع هر میله تعداد کل مسافران هر طبقهٔ اقتصادی را نشان می‌دهد.

شکل ۶–۱۰ از این جهت که محور y مشخصی وجود ندارد، با سایر نمودارهای میله‌ای متفاوت است. به‌جای آن مقادیر عددی مربوط به هر میله نشان داده شده است. وقتی نموداری قرار است تعداد کمی از مقادیر مختلف را نشان دهد، منطقی است که مقادیر واقعی آن‌ها درون نمودار نوشته شود. این کار می‌تواند قویاً اطلاعاتی که توسّط نمودار منتقل می‌شود را بدون افزودن بار بصری مازاد افزایش داده و نیاز به محور y مجزا را از بین می‌برد.



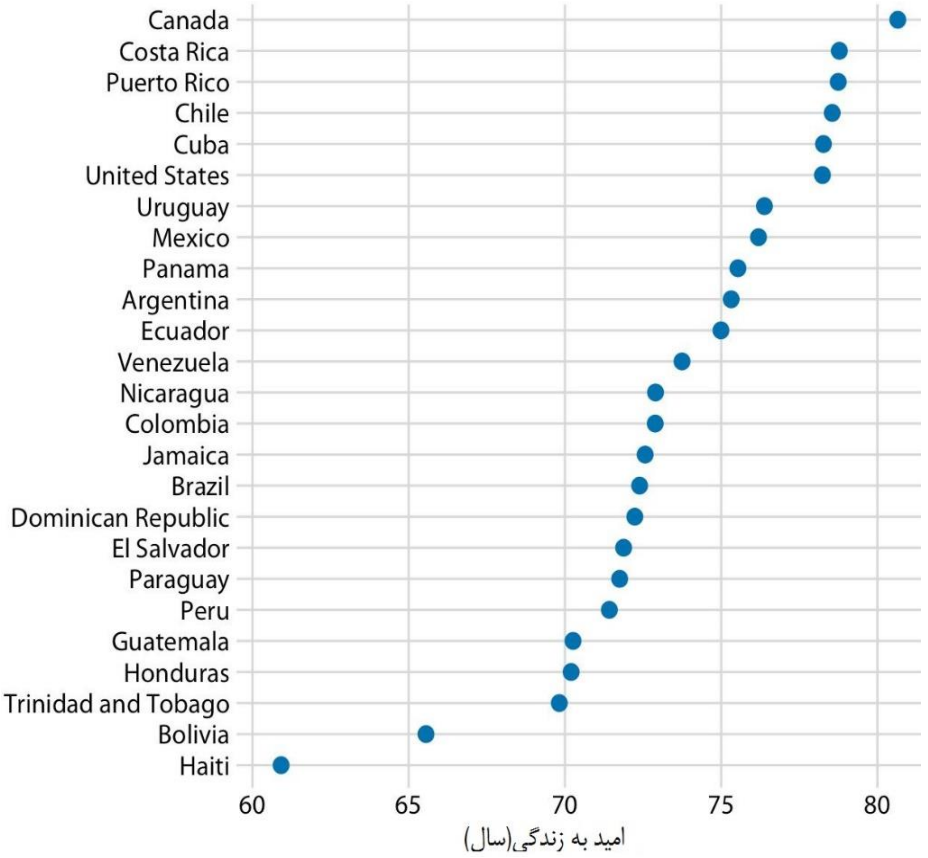
نمودار ۱۰–۶. تعداد مسافران مرد و زن در کشتی تایتانیک به تفکیک طبقهٔ اقتصادی یک، دو و سه. منبع داده: Encyclopedia Titanica

نمودار نقطه‌ای و نقشه حرارتی[1]

میله‌ها تنها راه ترسیم مقادیر نیستند. یکی از محدودیت‌های مهم میله‌ها این است که باید از صفر شروع شوند تا طول میله متناسب با مقدار نشان داده شده باشد. برای برخی داده‌ها، این کار غیرعملی بوده یا ممکن است ویژگی‌های اصلی داده‌ها را مخفی کند. در این شرایط می‌توان مقادیر را با نقطه در موقعیّت مناسب نسبت به x و y نشان داد.

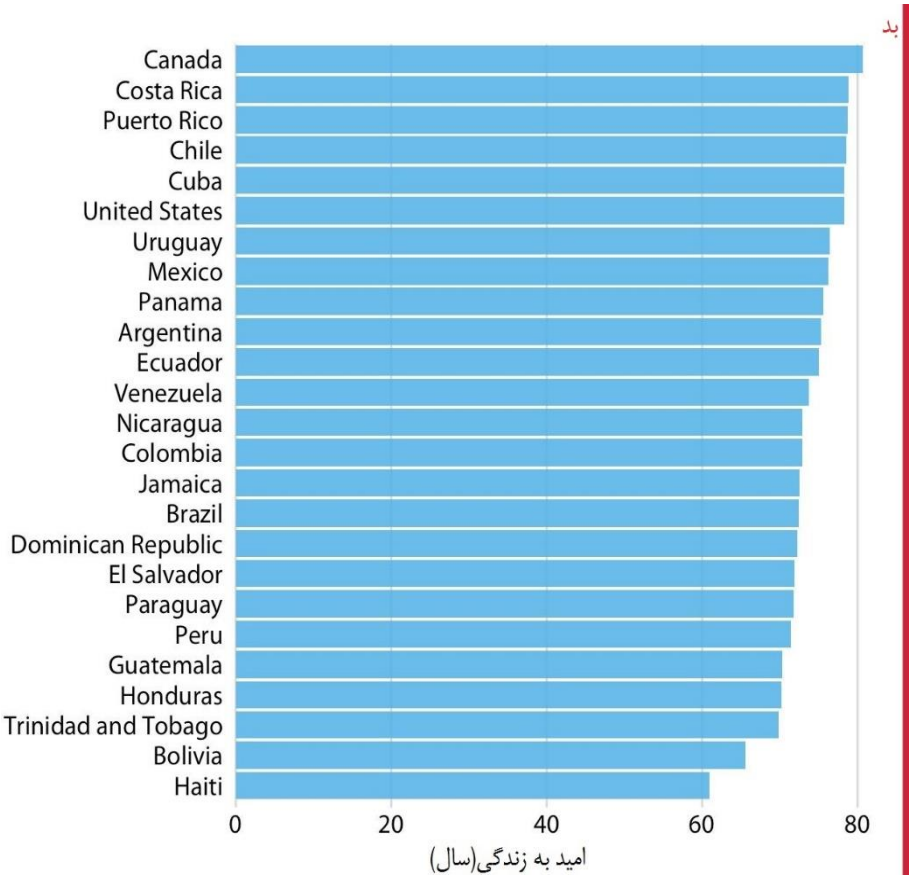
شکل ۶–۱۱ این رویکرد رسم نمودار را برای داده‌های امید به زندگی در ۲۵ کشور در قارهٔ آمریکا نشان می‌دهد. امید به زندگی شهروندان این کشورها بین ۶۰ تا ۸۱ سال است و مقدار امید به زندگی هر کدام با نقطه آبی در موقعیّت مناسب نسبت به محور x رسم شده است. با محدود کردن دامنهٔ محور به ۶۰ تا ۸۱ سال، شکل حاصل نکات کلیدی این داده‌ها را نمایان می‌کند: کانادا بیشترین و بولیوی و هائیتی کمترین امید به زندگی را در بین کشورهای مورد بررسی دارند. اگر از میله به جای نقطه استفاده می‌شد (شکل ۶–۱۲)، شکل حاصل وضوح کمتری داشت. از آنجایی که میله‌ها خیلی بلند بوده و تقریبا طول یکسانی دارند، چشم را به جای انتهای میله‌ها، به وسط آن‌ها می‌کشاند و پیام مدنظر منتقل نمی‌شود.

1. Dot Plots and Heatmaps



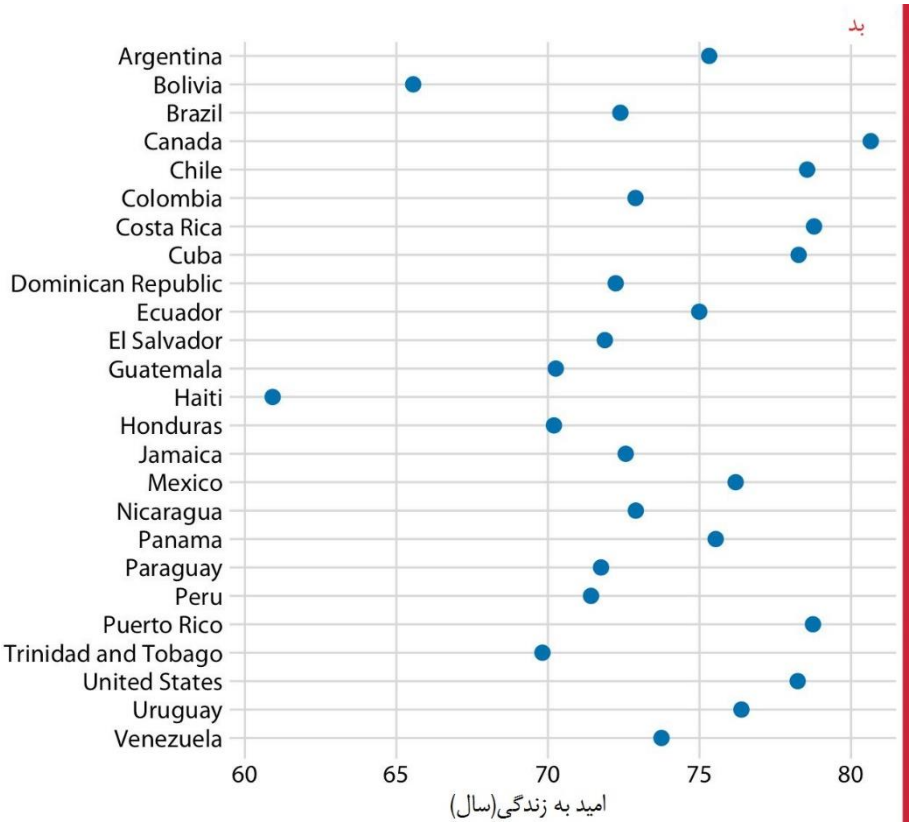
نمودار ۶–۱۱. امید به زندگی در کشورهای قارهٔ امریکا در سال ۲۰۰۷. منبع داده: Gapminder

صرف نظر از اینکه از میله یا نقطه استفاده می‌شود، باید به ترتیب مقادیر توجّه نمود. در شکل ۶–۱۱ و ۶–۱۲ کشورها به ترتیب نزولی بر اساس امید به زندگی مرتّب شده‌اند. اگر بر اساس حروف الفبا مرتّب شوند، با ابری از نقاط نامنظم مواجه می‌شویم که گیج کننده بوده و پیام واضحی را منتقل نمی‌کند (شکل ۶–۱۳).



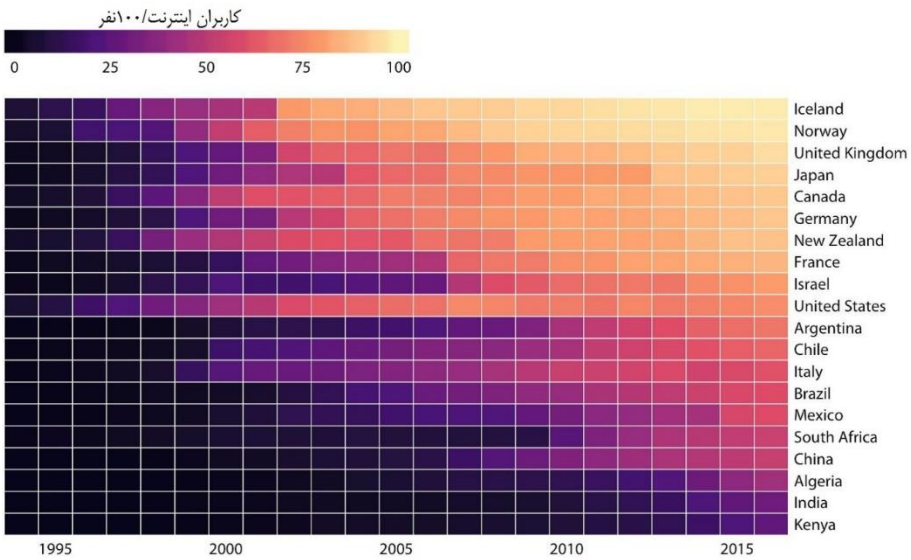
نمودار ۶–۱۲. امید به زندگی در کشورهای قارهٔ امریکا در سال ۲۰۰۷ که به صورت میله‌ای نشان داده شده است. این داده‌ها برای نمایش توسّط میله‌ها مناسب نیستند. میله‌ها خیلی طولانی بوده و توجّه را از خصوصیت اصلی داده‌ها که تفاوت در امید به زندگی در کشورهای مختلف است، منحرف می‌کند. منبع داده: Gapminder

تاکنون نمایش داده در تمام مثال‌ها مبتنی بر موقعیّت آن‌ها در طول محور بوده است، چه به صورت محل انتهای میله و یا قرارگیری نقطه. برای داده‌های خیلی زیاد، هیچ کدام از این دو روش مناسب نیستند زیرا نمودار نهایی خیلی شلوغ خواهد بود. قبلا در شکل ۶–۷ دیدیم که تنها ۷ گروه که هر کدام ۴ داده دارند، نموداری پیچیده به دست می آید که خواندن آن آسان نیست. اگر ۲۰ گروه متشکل از ۲۰ داده داشتیم، نمودار حاصل کاملاً گیج‌کننده می‌شد.



شکل ۶–۱۳. امید به زندگی در کشورهای قارهٔ امریکا در سال ۲۰۰۷. در اینجا کشورها به ترتیب حروف الفبا مرتّب شده‌اند که ماحصل آن ابری از نقاط نامنظم است. خواندن این نمودار سخت است و لذا برچسب "بد" خورده است. منبع داده: Gapminder

بجای نمایش مقادیر داده‌ها با موقعیّت بوسیله میله و نقطه، می‌توان با رنگ آن‌ها را نمایش داد که به چنین شکلی نقشه حرارتی اطلاق می‌شود. در شکل ۶–۱۴ این روش برای نشان دادن درصد کاربران اینترنت در ۲۰ کشور و در طول ۲۳ سال از سال ۱۹۹۴ تا ۲۰۱۶، به کار رفته است. در حالی که این روش مشخّص کردن مقدار دقیق اطّلاعات نشان داده شده را سخت‌تر می‌کند (مثلاً درصد دقیق کاربران اینترنت در ایالات متحده در سال ۲۰۱۵ چقدر بوده است؟). با این حال برای تاکید و برجسته کردن روندهای بزرگتر خیلی مفید است. می‌توان دید که در کدام کشور استفاده از اینترنت زودتر شروع شده و در کدام کشور دیرتر و نیز می‌توان مشاهده نمود که در کدام کشورها در آخرین سال مورد بررسی (یعنی ۲۰۱۶)، نفوذ اینترنت بیشتر بوده است.

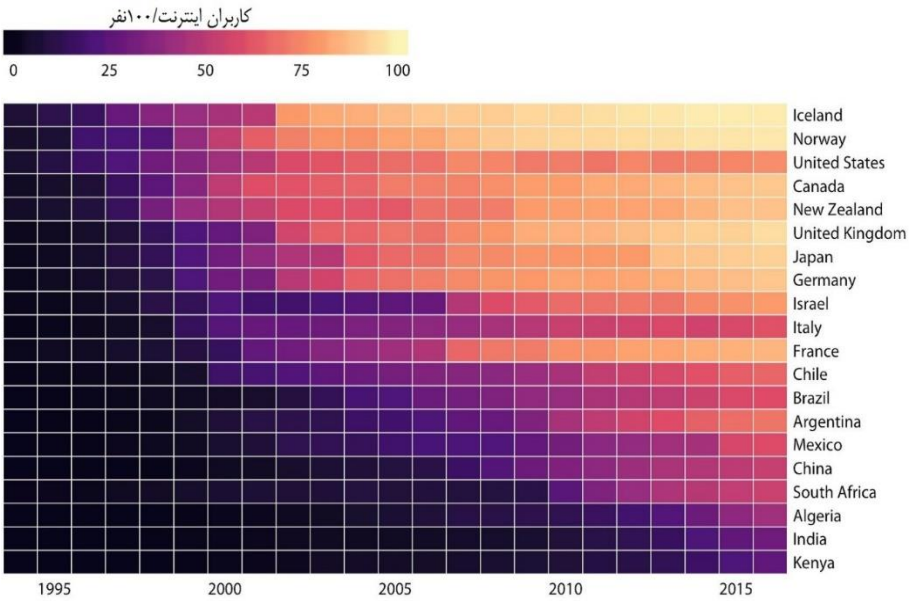


نمودار ۶–۱۴. استفاده از اینترنت در طول زمان در کشورهای منتخب. رنگ‌ها نشان‌دهندهٔ درصد کاربران اینترنت در هر کشور و در هر سال مشخّص می‌باشد کشورها بر اساس درصد کاربران اینترنت در سال ۲۰۱۶ مرتّب شده‌اند. منبع داده: بانک جهانی

در این روش هم مانند تمام رویکردهای دیگری که در این فصل بحث شد، باید به مرتّب‌سازی صحیح متغیّر ترتیبی هنگام رسم نقشه حرارتی توجّه شود. در شکل ۶–۱۴ کشورها بر اساس درصد کاربران اینترنت در سال ۲۰۱۶ مرتّب شده‌اند. این طبقه‌بندی، انگلیس، ژاپن، کانادا و آلمان را بالاتر از آمریکا قرار می‌دهد زیرا تمام این کشورها نفوذ اینترنت بیشتری در سال ۲۰۱۶ نسبت به آمریکا داشته‌اند، اگرچه در سال‌های قبل‌تر آمریکا کاربران بیشتری داشته است. همچنین می‌توان کشورها را بر اساس اینکه چقدر زود شروع به مصرف قابل توجّه اینترنت کرده‌اند مرتّب کرد. در شکل ۶–۱۵ کشورها بر اساس سالی که در آن استفاده از اینترنت به بالای ۲۰٪ رسیده، مرتّب شده‌اند. در این شکل، آمریکا جایگاه سوّم را به دست می‌آورد که نمایانگر مصرف نسبتاً کم اینترنت در سال ۲۰۱۶ است، در مقایسه با اینکه چقدر زود استفاده از اینترنت را شروع کرده است. الگوی مشابهی برای ایتالیا دیده می‌شود. برعکس آن، هرچند سرزمین‌های اشغالی فلسطین و فرانسه به نسبت دیر شروع کردند امّا میدان را سریع به دست گرفتند.

هر دو شکل ۶–۱۴ و ۶–۱۵ نمایش صحیحی از داده‌ها هستند. انتخاب یکی از آن‌ها وابسته به داستانی است که قرار است تعریف شود. اگر داستان در مورد استفاده از اینترنت در سال ۲۰۱۶

است، شکل ۶-۱۴ احتمالاً انتخاب بهتری است. از سوی دیگر اگر داستان در مورد این است که استفاده فعلی اینترنت چقدر با شروع استفاده زودهنگام یا دیرهنگام از آن مرتبط است، نمودار ۶-۱۵ انتخاب بهتری است.



نمودار ۶-۱۵. استفاده از اینترنت در طول زمان در کشورهای منتخب. کشورها بر اساس سالی که برای اوّلین بار درصد کاربران اینترنت از مرز ۲۰% فراتر رفت مرتب شده‌اند. منبع داده: بانک جهانی

فصل ۷

نمایش توزیع‌ها: هیستوگرام[1] و نقشهٔ تراکمی[2]

ما مکرراً با شرایطی مواجه می‌شویم که می‌خواهیم نحوهٔ توزیع یک متغیّر را در مجموعهٔ داده‌های خود مشاهده کنیم. به عنوان مثال مسافرین کشتی تایتانیک را به عنوان یک مجموعهٔ داده که در فصل ۶ مرور کردیم، در نظر بگیرید، حدودا ۱۳۰۰ مسافر (بدون احتساب خدمه) در این کشتی حضور داشتند و سن ۷۵۶ نفر از آن‌ها را داریم. ممکن است بخواهیم تعداد مسافرین در هر گروه سنی را بدانیم (یعنی تعداد کودکان، نوجوانان، میانسال‌ها، سالخورده‌ها و ...). سهم نسبی مسافران از سنین مختلف را توزیع سن مسافران می‌نامیم.

نمایش یک توزیع منفرد

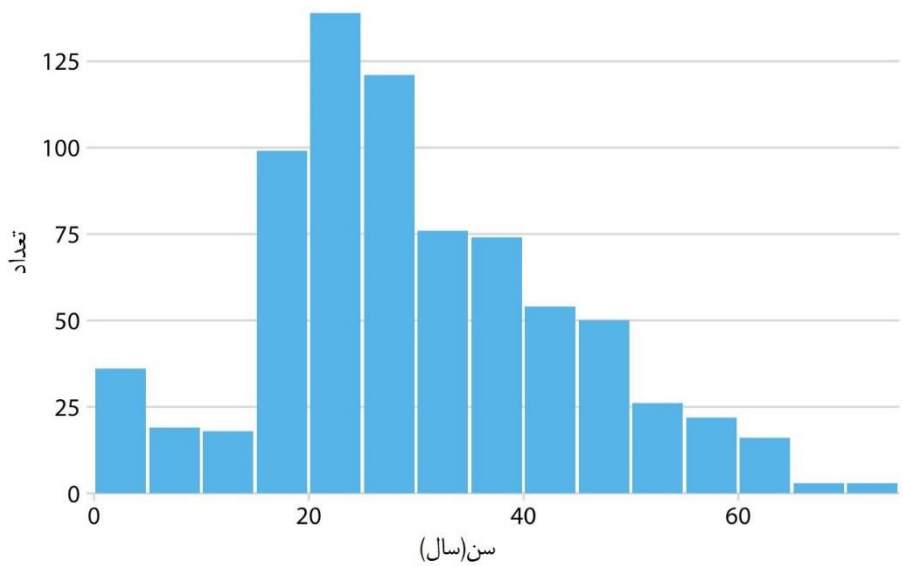
ما می‌توانیم با گروه‌بندی مسافرین به دسته‌هایی با سنین قابل مقایسه و سپس شمارش تعداد مسافرین در هر دسته به توزیع سن در میان مسافرین دست پیدا کنیم. نتایج در جدول ۷-۱ نمایش داده شده است.

1. Histograms
2. Density Plots

جدول ۷–۱. تعداد مسافران با سن مشخّص در کشتی تایتانیک

دامنه سنی	فراوانی	دامنه سنی	فراوانی	دامنه سنی	فراوانی
۰–۵	۳۶	۳۱–۳۵	۷۶	۶۱–۶۵	۱۶
۶–۱۰	۱۹	۳۶–۴۰	۷۴	۶۶–۷۰	۳
۱۱–۱۵	۱۸	۴۱–۴۵	۵۴	۷۱–۷۵	۳
۱۶–۲۰	۹۹	۴۶–۵۰	۵۰		
۲۱–۲۵	۱۳۹	۵۱–۵۵	۲۶		
۲۶–۳۰	۱۲۱	۵۶–۶۰	۲۲		

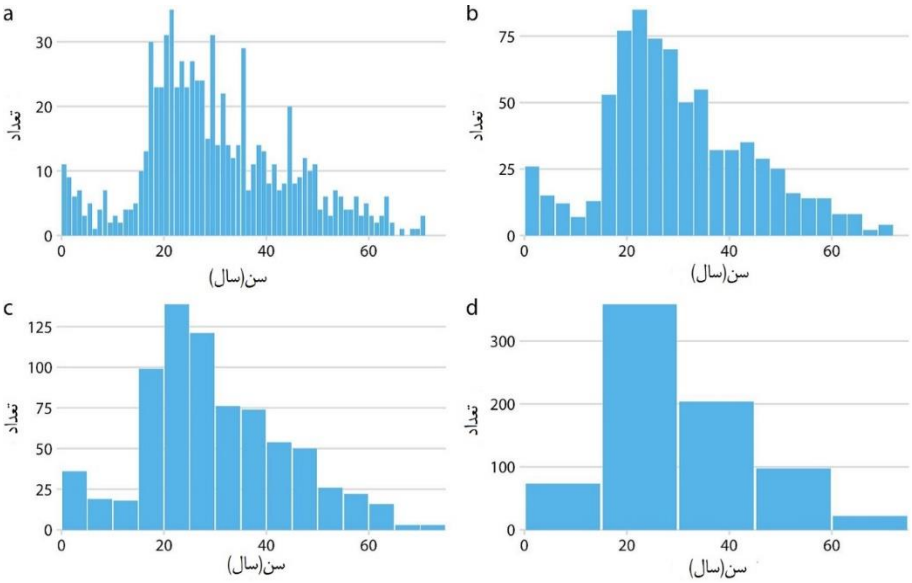
ما می‌توانیم این جدول را به صورت مستطیل‌های توپری که ارتفاع آن‌ها بیانگر تعداد افراد و عرض آن‌ها نشان‌دهندهٔ گروه سنی است به تصویر بکشیم (شکل ۷–۱). این نوع نمودار را هیستوگرام می‌گویند (در نظر داشته باشید که تمام دسته‌ها باید عرض یکسانی داشته باشند که بتوان آن را به عنوان هیستوگرام صحیحی در نظر گرفت).



نمودار ۷–۱. هیستوگرام سن برای مسافران کشتی تایتانیک

از آنجایی که هیستوگرام با دسته‌بندی داده‌ها رسم می‌شود، ظاهر اصلی آن‌ها بستگی به عرض دسته‌بندی‌های انجام شده دارد. اکثر نرم‌افزارها که هیستوگرام‌ها را رسم می‌کنند، به طور پیش فرض عرض دسته را انتخاب می‌کنند، امّا ممکن است عرض های انتخاب شده توسّط برنامه دقیقاً آن چیزی که شما می‌خواستید نباشد. در نتیجه ضروریست که همیشه عرض‌های مختلف دسته را امتحان کنید تا مطمئن شوید که بهترین حالت برای نمایش داده‌ها را انتخاب کرده‌اید. به طور کلی اگر عرض دسته‌ها خیلی کوچک باشد هیستوگرام ظاهری شلوغ و قلّه‌های متعدد خواهد داشت و ممکن است روند اصلی داده‌ها پنهان شود. از طرف دیگر اگر عرض دسته‌ها بسیار زیاد باشد، ممکن است مشخصات کوچک موجود در توزیع داده‌ها از بین بروند؛ مانند توزیع فراوانی در اطراف گروه سنی ۱۰ سال در نمودار ۷–۱.

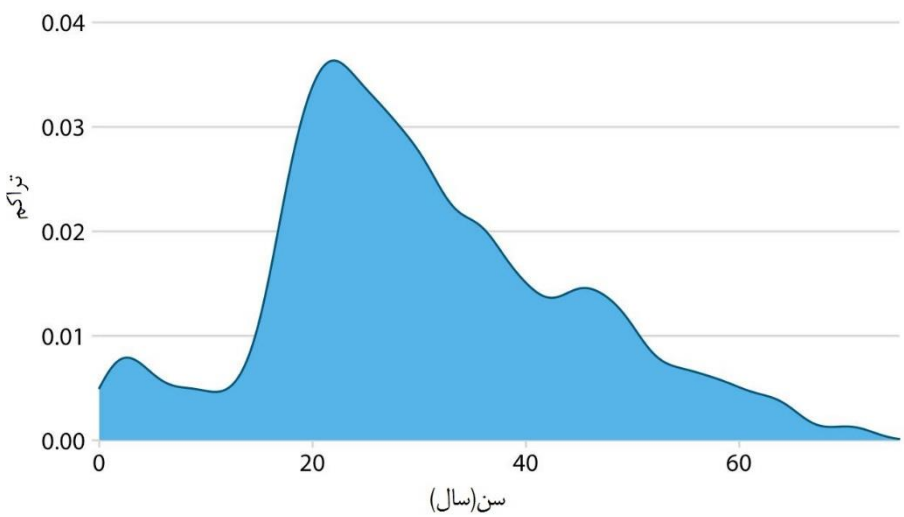
برای توزیع سن در خصوص مسافران کشتی تایتانیک می‌دانیم که عرض دسته ۱ سال بسیار کوچک است و عرض دسته ۱۵ سال بسیار بزرگ است، به نظر می‌رسد عرض دسته در حدود ۳ تا ۵ سال مناسب باشد (نمودار ۷–۲).



نمودار ۷–۲. هیستوگرام به انتخاب عرض دسته وابسته است. در اینجا توزیع سنی مسافران کشتی تایتانیک با ۴ عرض دسته مختلف نمایش داده شده است: (الف) ۱ سال، (ب) ۳ سال، (ج) ۵ سال، (د) ۱۵ سال

همیشه هنگام رسم هیستوگرام، عرض‌های مختلف دسته‌ها را امتحان کنید.

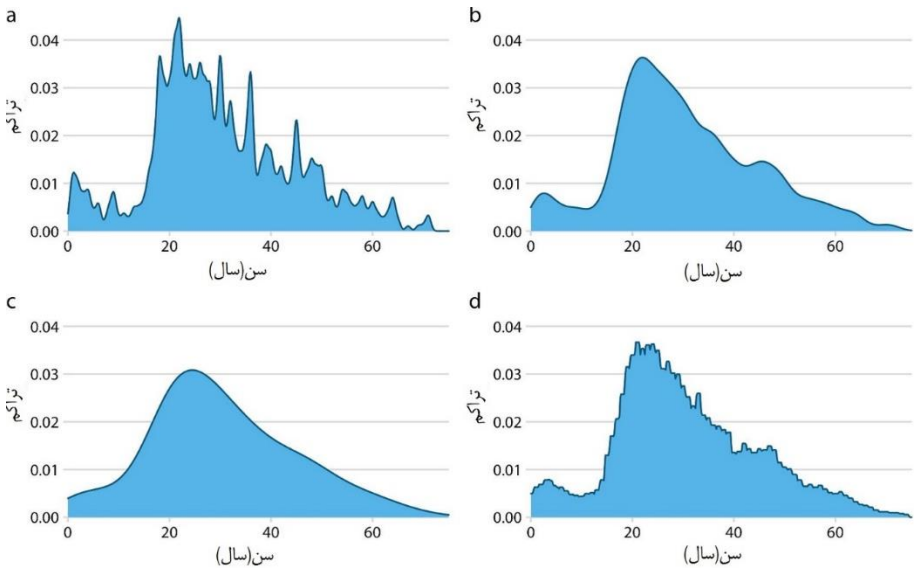
هیستوگرام‌ها از قرن ۱۸ میلادی یکی از انتخاب‌های محبوب برای به تصویر کشیدن توزیع فراوانی بودند زیرا به راحتی با دست رسم می‌شوند. اخیراً به علّت اینکه قدرت محاسباتی بالا از طریق لپ‌تاپ‌ها و موبایل‌ها در دسترس همه قرار دارد، نمودارهای تراکمی به سرعت جای هیستوگرام‌ها را می‌گیرند. در نمودار تراکمی تلاش می‌شود که توزیع احتمال داده‌های موجود با استفاده از منحنی پیوسته مناسب رسم شود (نمودار ۷-۳). این منحنی باید بر اساس تخمینی از داده‌ها رسم شود و یکی از رایج‌ترین روش‌های استفاده شده برای این فرآیند، تخمین تراکم کرنل[1] نام دارد. در تخمین تراکم کرنل یک منحنی پیوسته (منحنی کرنل) با یک عرض کوچک (که با پارامتری به نام پهنای باند کنترل می‌شود) در محل متناظر هر داده رسم می‌شود و سپس تمام منحنی‌ها به هم اضافه شده تا تخمین تراکم نهایی به دست آید. شایع‌ترین کرنل مورد استفاده، کرنل گوسی (زنگوله گوسی) است امّا انتخاب‌های متعدد دیگری نیز وجود دارند.



نمودار ۷-۳. تخمین تراکم کرنل برای سن مسافران در کشتی تایتانیک. ارتفاع منحنی به گونه‌ای تنظیم شده که سطح زیر منحنی معادل ۱ شود. تخمین تراکمی با استفاده از کرنل گوسی و پهنای باند ۲ رسم شده است.

1. kernel density estimation

دقیقاً مانند هیستوگرام‌ها، شکل اصلی نمودار تراکمی بسته به کرنل و پهنای باند انتخابی دارد (شکل ۷–۴). پارامتر پهنای باند مانند عرض دسته در هیستوگرام عمل می‌کند. اگر پهنای باند بسیار کوچک باشد نمودار تخمین تراکمی بسیار شلوغ با قلّه‌های فراوان خواهد بود و ممکن است روند اصلی داده‌ها از دست برود. از طرف دیگر اگر پهنای باند بسیار بزرگ باشد، روندهای کوچک که در توزیع داده‌ها وجود دارد ناپدید خواهند شد. علاوه بر این انتخاب کرنل، بر شکل نهایی منحنی تراکم تاثیرگذار است. به عنوان مثال یک کرنل گوسی تمایل خواهد داشت که نمودار تخمین تراکمی را به شکل گوسی (نرمال) با لبه های صاف و دم‌های باریک تبدیل کند. از سوی دیگر، کرنل مستطیلی می‌تواند نمای پله‌ای به نمودار تراکمی بدهد (نمودار ۷–۴). به طور کلی هرچه تعداد داده‌ها بیشتر باشد، انتخاب نوع کرنل اهمیت کمتری خواهد داشت. در نتیجه، نمودارهای تراکمی برای پایگاه‌های داده بزرگ بسیار قابل اعتماد و آگاهی بخش می‌باشند امّا برای پایگاه‌های داده کوچک می‌تواند گمراه کننده باشد.

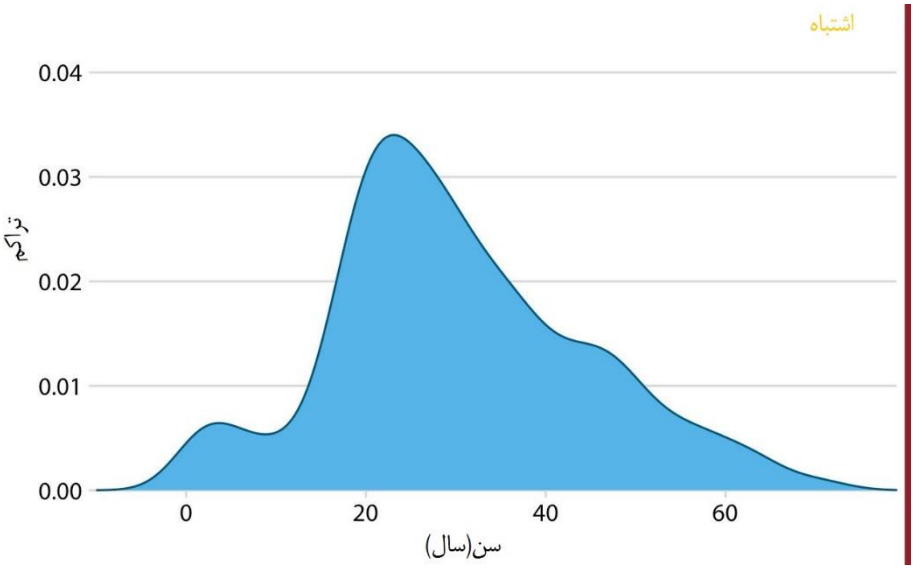


نمودار ۷–۴. تخمین تراکمی کرنل، به کرنل و پهنای باند انتخاب شده وابسته است. در اینجا توزیع سنی مسافران کشتی تایتانیک با ۴ ترکیب از پارامترهای روبرو رسم شده است: (الف) کرنل گوسی و پهنای باند ۰/۵، (ب) کرنل گوسی و پهنای باند ۲، (ج) کرنل گوسی و پهنای باند ۵، (د) کرنل مستطیلی و پهنای باند ۲

منحنی های تراکمی به گونه‌ای تنظیم می‌شوند که سطح زیر منحنی برابر یک شود. این مساله می‌تواند مقیاس محور y را گیج کننده نماید، زیرا مقادیر آن به واحد محور x وابسته

خواهد بود. مثلاً در مورد توزیع سن، دامنه داده‌های محور x از ۰ تا حدود ۷۵ است، در نتیجه انتظار داریم که میانگین ارتفاع منحنی تراکم برابر ۰/۰۱۳ = ۱/۷۵ باشد. در واقع هنگام بررسی منحنی تراکمی سن متوجّه می‌شویم که مقادیر y از صفر تا تقریبا ۰/۰۴ تغییر می‌کند که میانگین آن حدود ۰/۰۱ است.

تخمین های تراکمی کرنل یک اشکال دارند که باید از آن آگاه باشیم: آن‌ها تمایل دارند الگویی از داده‌هایی را مخصوصاً در دم‌ها تشکیل دهند که در حقیقت وجود ندارند. در نتیجه استفاده بی مورد از تخمین تراکمی می‌تواند منجر به رسم نمودارهایی شود که منطقی نیستند. به عنوان مثال، اگر به این مساله دقّت نشود ممکن است منحنی توزیع سنی رسم شود که شامل سنین منفی هم باشد (نمودار ۷–۵).



نمودار ۷–۵. تخمین تراکمی کرنل ممکن است دم‌های منحنی توزیع را تا نواحی ادامه دهد که هیچ داده‌ای نداشته و حتّی منطقاً هم صحیح نیست. در اینجا تخمین تراکمی برای سن مسافران کشتی تایتانیک تا گروه سنی منفی گسترش یافته است. این مساله منطقی نیست و باید از آن پرهیز شود.

همواره دقّت نمایید که نمودار تخمین تراکمی پیشگویی‌کنندهٔ مقادیر داده غیر منطقی نباشد.

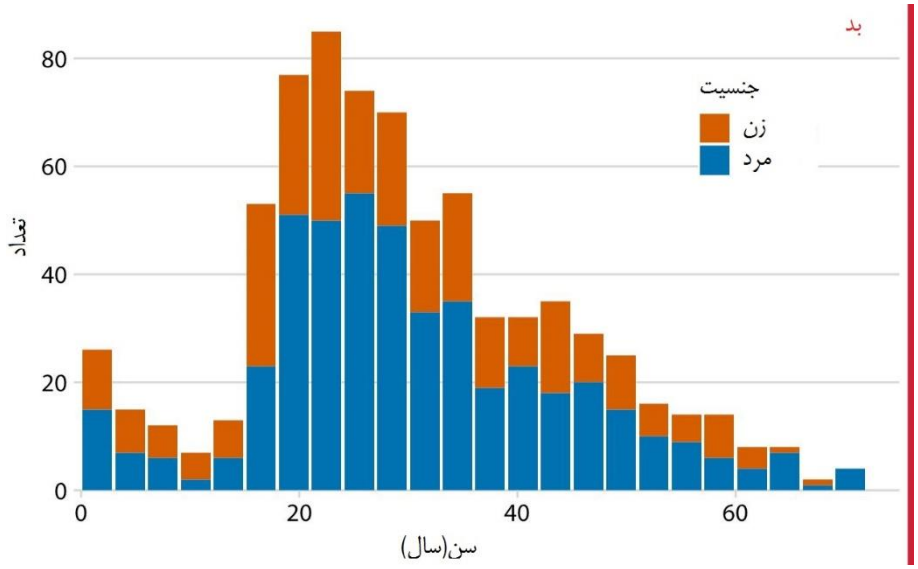
بالاخره برای نمایش توزیع بهتر است از هیستوگرام استفاده شود یا نمودار تراکمی؟ بحث‌های داغی پیرامون این موضوع می‌توان مطرح کرد. برخی افراد به شدّت مخالف نمودارهای تراکمی هستند و معتقدند آن‌ها گمراه‌کننده و ساختگی هستند. برخی دیگر معتقدند که نمودارهای هیستوگرام می‌تواند گمراه‌کننده و ساختگی باشند. به نظر می‌رسد انتخاب نهایی بین این دو بیشتر به سلیقهٔ افراد وابسته است، امّا بعضی اوقات یکی از این دو به صورت دقیق‌تری می‌تواند خصوصیات داده‌های موجود را نمایش دهد. همچنین می‌توان از این دو استفاده نکرد و در عوض از توابع تراکم تجمّعی تجربی یا نمودارهای q-q بهره برد (فصل ۸). هرچند این اعتقاد وجود دارد که نمودارهای تخمین تراکمی، به خصوص وقتی هدف نمایش همزمان چند توزیع است، برتری ذاتی نسبت به هیستوگرام‌ها دارند.

نمایش چند توزیع به صورت همزمان

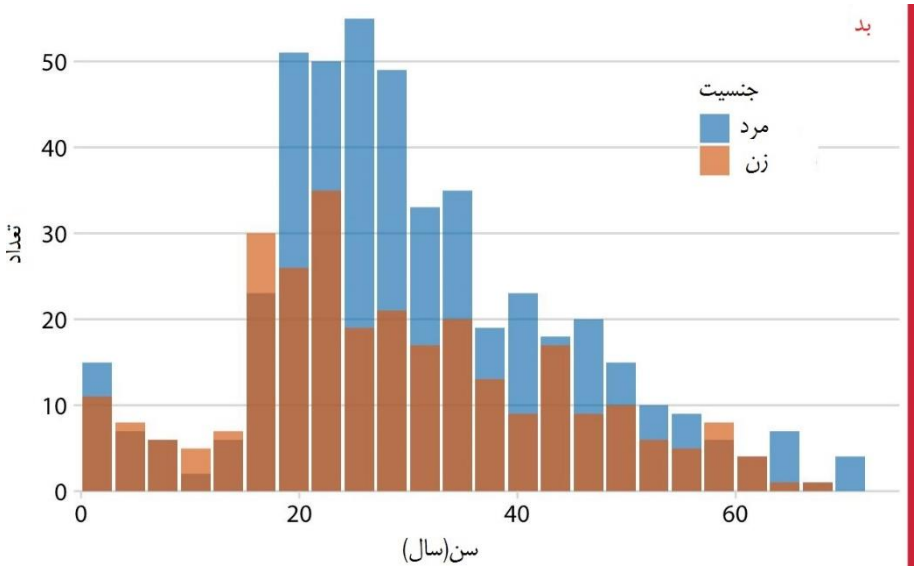
در بسیاری مواقع با چندین توزیع مواجه هستیم و هدف نمایش آن‌ها به صورت همزمان است. مثلاً فرض کنید می‌خواهیم نحوهٔ توزیع سن در مسافران کشتی تایتانیک را بین خانم‌ها و آقایان بررسی نماییم. آیا سن خانم‌ها و آقایان به طور کلی مشابه است یا تفاوتی بین آن‌ها وجود دارد؟ یک راهبرد برای نمایش این حالت استفاده از هیستوگرام انباشته است که در آن ستون مربوط به هیستوگرام زنان با رنگی متفاوت روی ستون هیستوگرام مربوط به مردان رسم می‌شود (نمودار ۷-۶).

به نظر می‌رسد از این نحوه نمایش باید پرهیز شود. زیرا دو مشکل اساسی وجود دارد. اوّل، با نگاه کردن به این نمودار، هیچگاه نمی‌توان متوجّه شد که ستون‌ها دقیقاً از کجا شروع شده است. آیا از جایی که رنگ تغییر می‌کند شروع می‌شوند یا از نقطه صفر؟ به بیان دیگر آیا حدود ۲۵ خانم در گروه سنی ۱۸ تا ۲۰ سال وجود دارد یا حدود ۸۰ نفر؟ (پاسخ حالت اوّل است) دوّم، طول ستون‌های مربوط به خانم‌ها را نمی‌توان به طور مستقیم با یکدیگر مقایسه نمود، زیرا نقطه شروع متفاوتی دارند. به عنوان مثال، مردان به طور متوسط مسن‌تر از خانم‌ها هستند، با این حال این مساله به هیچ عنوان در نمودار ۷-۶ مشخّص نیست.

پس می‌توانیم برای حل این مشکلات تمام ستون‌ها را از نقطه صفر شروع کرده و ستون‌ها را مقداری شفاف نماییم (نمودار ۷-۷).



شکل ۷–۶. هیستوگرام توزیع سن مسافران کشتی تایتانیک طبقه‌بندی شده بر اساس جنسیت. این نمودار به عنوان «بد» برچسب زده شده است زیرا هیستوگرام انباشه به راحتی با هیستوگرام هایی که همپوشانی دارند، اشتباه می‌شوند (نمودار ۷–۷ را ببینید). علاوه بر این ارتفاع ستون‌های مربوط به مسافران خانم به راحتی با بقیّه قابل مقایسه نیست.

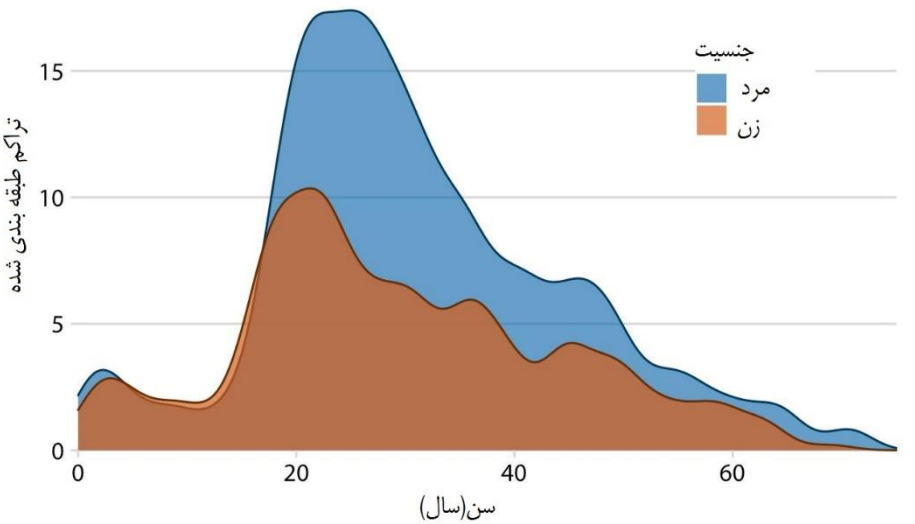


شکل ۷–۷. توزیع سن مسافران کشتی تایتانیک به صورت دو هیستوگرام همپوشان. این نمودار به عنوان «بد» برچسب زده شده است زیرا هیچ سرنخ بصری وجود ندارد که تمام ستون‌های آبی از نقطه صفر شروع شده‌اند.

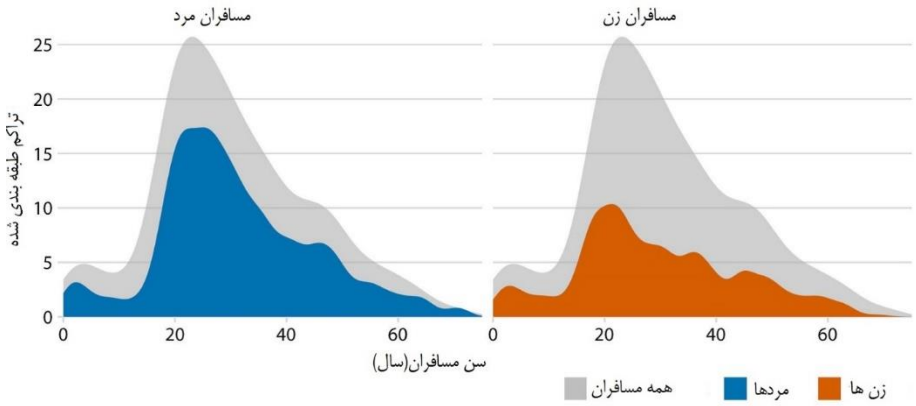
با این حال، این رویکرد مشکلات جدیدی ایجاد می‌کند. الان به نظر می‌رسد که در حقیقت سه گروه وجود دارد و همچنان مشخّص نیست که هر ستون از چه نقطه‌ای شروع و در چه نقطه‌ای پایان می‌یابد. همپوشانی هیستوگرام‌ها ایده خوبی نیست زیرا ستون‌های نیمه شفّاف که همپوشانی دارند، به صورت ستونی با رنگ جدید دیده می‌شوند.

معمولاً نمودارهای تراکمی همپوشان مشکلاتی که هیستوگرام‌ها دارند را ندارند، زیرا پیوستگی خطوط منحنی تراکمی باعث می‌شود چشم بتواند تمایز توزیع‌ها را به وضوح تشخیص دهد. با این حال برای داده‌های مثال حاضر، توزیع سن برای مردان و زنان تا حدود سن ۱۷ سال یکسان است و سپس واگرا می‌شود، بنابراین نمودار به دست آمده ایده‌آل نخواهد بود (نمودار ۷–۸).

راه حلی که برای این داده‌ها وجود دارد این است که توزیع سنی زنان و مردان را به طور جداگانه و به صورت نسبتی از توزیع کلی سنی نشان دهیم (نمودار ۷–۹). این نمودار به وضوح نشان می‌دهد که تعداد زنان در سنین ۲۰ تا ۵۰ سال نسبت به مردان بسیار کمتر است.

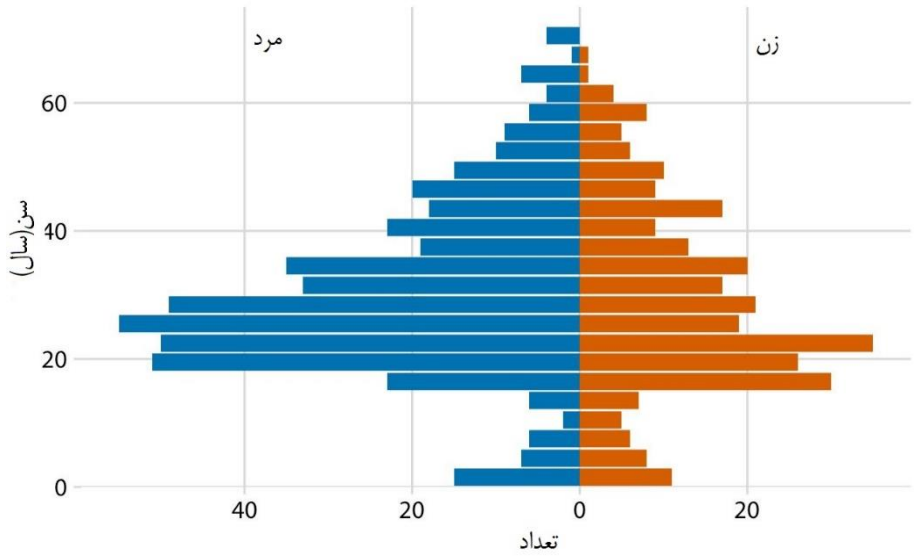


نمودار ۷–۸. تخمین تراکمی سن مسافران زن و مرد کشتی تایتانیک. برای تاکید بر اینکه تعداد مردان نسبت به زنان بیشتر بوده است، منحنی تراکمی به گونه‌ای رسم شده است که سطح زیر منحنی با تعداد کل مسافران مرد و زن متناسب باشد (۴۶۸ مرد و ۲۸۸ زن).



نمودار ۷–۹. توزیع سنی مسافران مرد و زن کشتی تایتانیک به صورت نسبتی از کل مسافران. ناحیه رنگی نشان دهندهٔ تخمین تراکمی مسافران مرد و زن می‌باشد. ناحیه خاکستری نشان دهندهٔ توزیع سنی کل مسافران است.

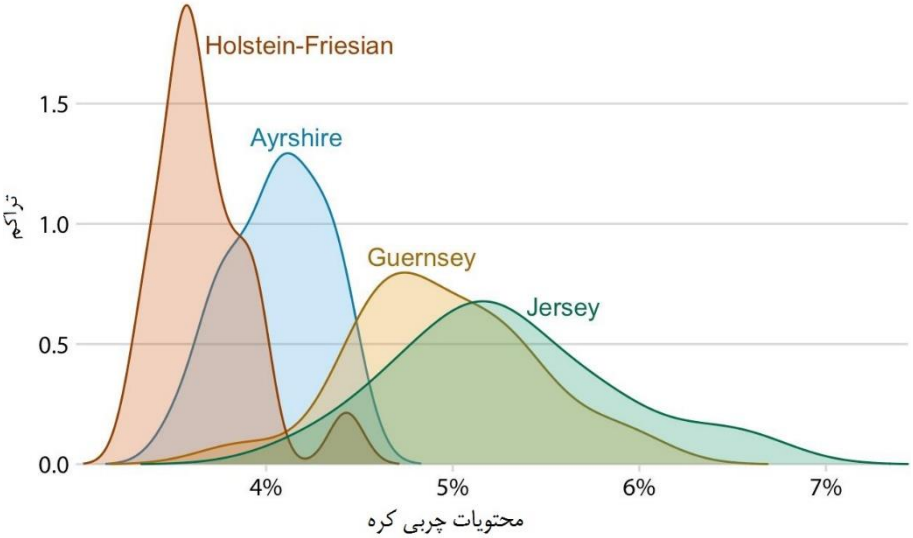
در نهایت وقتی بخواهیم دقیقاً هر دو توزیع نمایش داده شود، می‌توان دو هیستوگرام مجزا رسم نموده، آن‌ها را۹۰ درجه چرخانده و هیستوگرام‌ها را با یک نقطه شروع مشترک امّا با میله‌هایی در جهت‌های مخالف تنظیم نمود. این روش عمدتاً برای نمایش توزیع سن استفاده می‌شود و به آن هرم سنی اطلاق می‌شود (نمودار ۷–۱۰).



نمودار ۷–۱۰. توزیع سنی مسافران مرد و زن کشتی تایتانیک که به صورت هرم سنی ارائه شده است.

باید توجّه داشت کـه وقتـی می‌خـواهیم بیشـتر از دو توزیـع را بـه طـور همزمـان بـه نمـایش بگـذاریم این روش کاربرد ندارد. بـرای چنـدین توزیـع، هیسـتوگرام‌ها پیچیـده و گـیج‌کننـده می‌شـوند، امّـا اگـر توزیع‌هـا پیوسـته و مجـزا باشـند، نمودارهـای تراکمـی مناسـب خواهنـد بـود. بـه عنـوان مثـال، بـرای نمـایش توزیـع درصـد چربی‌هـای شـیر گـاو از ۴ نـژاد مختلـف، نمودارهـای تراکمـی مناسـب هسـتند (نمودار ۷–۱۱).

برای نمایـش چندین تـوزیع به طور هـمزمان، نمودارهای تـراکمی کرنل بهـتر از هیستوگرام‌ها هستند.



نمودار ۷–۱۱. تخمین تراکمی درصد چربی در شیر چهار نژاد مختلف گاو.

فصل ۸

نمایش توزیع‌ها:
توابع توزیع فراوانی تجمّعی و نمودارهای Q-Q

در فصل ۷، توضیح دادیم که چگونه می‌توان توزیع متغیّرها را با هیستوگرام یا نمودار تراکمی نشان داد. هر دوی این رویکردها شهودی و از نظر بصری جذّاب هستند. با این حال، همانطور که در آن فصل بحث شد، هر دوی آن‌ها در این محدودیت مشترک‌اند که شکل حاصل تا حد زیادی بستگی به پارامترهایی دارد که کاربر باید انتخاب کند، مانند عرض میله برای هیستوگرام‌ها و پهنای باند برای نمودارهای تراکمی. در نتیجه، هر دو باید به‌عنوان تفسیری از داده‌ها در نظر گرفته شوند تا نمایش مستقیم خود داده‌ها.

به عنوان جایگزینی برای استفاده از هیستوگرام‌ها یا نمودارهای تراکمی، می‌توان به سادگی تمام نقاط داده را به صورت جداگانه به عنوان ابری از نقاط داده نشان داد. با این حال، این رویکرد برای داده‌های بسیار بزرگ سخت می‌باشد، و در هر صورت روش‌های تجمّعی ارزشمندتر است زیرا خصوصیات توزیع داده‌ها را برجسته می‌کند و نه داده‌های منفرد را. برای حل این مشکل، متخصصین آمار توابع توزیع فراوانی تجمّعی تجربی[1] (ECDFs) و نمودارهای چندک–چندک[2] (q-q) را ابداع کرده‌اند. این نمودارها نیازی به انتخاب پارامتر ندارند و همه داده‌ها را همزمان نشان می‌دهند. متأسفانه، آن‌ها کمی کمتر از هیستوگرام یا نمودار چگالی

1. empirical cumulative distribution functions
2. quantile-quantile

شهودی هستند و جز در مجلات تخصصی، به ندرت دیده می‌شود. با این حال، این نمودارها در میان متخصصین آمار بسیار محبوب هستند، و بنابراین معتقدیم که هر کسی که به نمایش داده‌ها علاقه دارد باید با این تکنیک‌ها آشنا باشد.

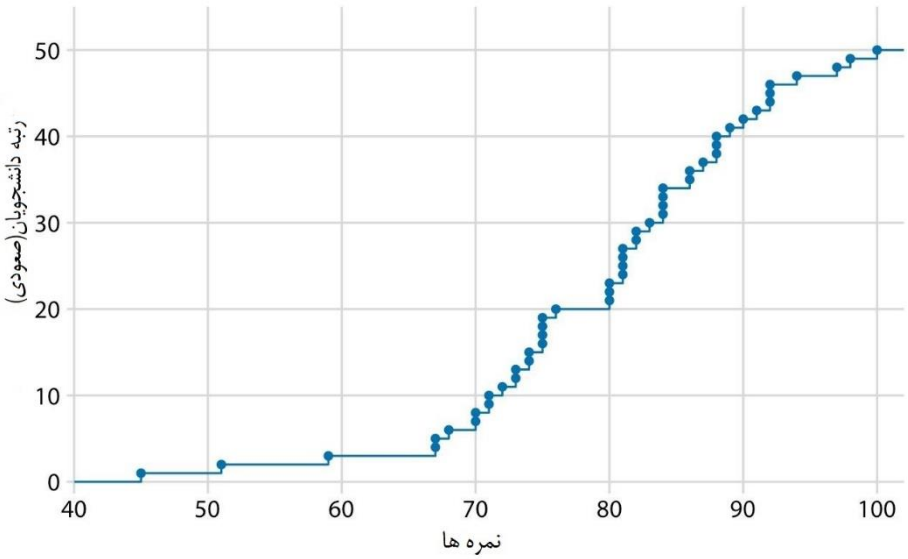
توابع توزیع فراوانی تجمّعی تجربی

برای توضیح ECDF، با یک مثال فرضی شروع می‌کنیم که دقیقاً از چیزی که به‌عنوان یک استاد در کلاس درس با آن سروکار داریم، مدل‌سازی شده است: مجموعه داده‌های نمرات دانشجویان. تصوّر کنید کلاس فرضی ما ۵۰ دانشجو دارد و دانشجویان به تازگی امتحانی را به پایان رسانده‌اند که در آن می‌توانند نمره‌ای بین ۰ تا ۱۰۰ کسب کنند. چگونه می‌توانیم عملکرد کلاس را به بهترین نحو نمایش دهیم، مثلاً برای تعیین محدوده مناسب نمرات؟

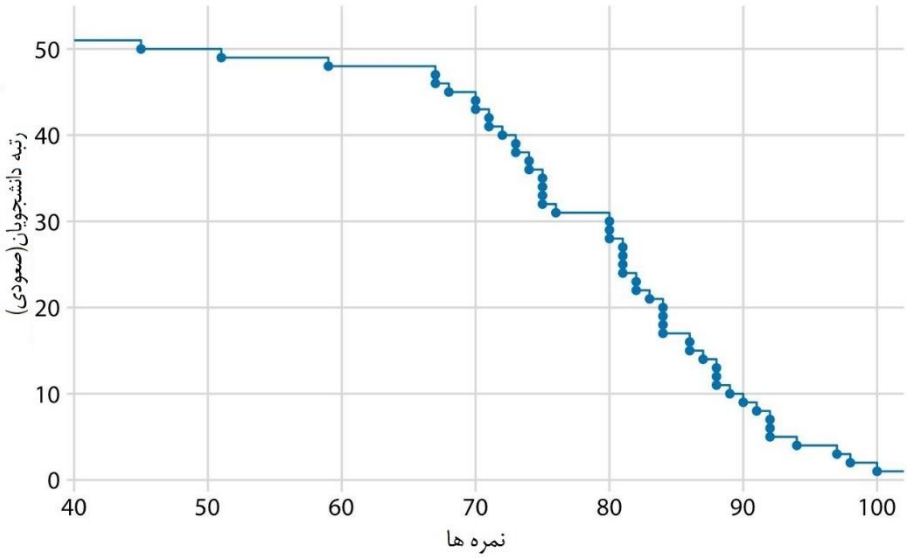
می‌توان نمودار کل دانشجویانی را که بیشترین نمره را دریافت کرده‌اند در مقابل تمام نمرات ممکن ترسیم نمود. این نمودار یک تابع صعودی خواهد بود که از تعداد صفر برای نمره صفر شروع می‌شود و به تعداد ۵۰ برای نمره ۱۰۰ ختم می‌شود. این نمودار را می‌توان به این صورت هم تصوّر کرد: می‌توان همهٔ دانشجویان را بر اساس نمره‌هایی که به دست آورده‌اند به ترتیب صعودی رتبه‌بندی نمود (بنابراین دانشجویانی که کمترین نمره را دارند کمترین رتبه و دانشجویانی که بیشترین نمره را دارند بالاترین رتبه را دریافت می‌کنند) و سپس رتبه، در مقابل نمرهٔ واقعی به دست آمده رسم شود. نتیجه حاصل یک تابع فراوانی تجمّعی تجربی یا همان فراوانی تجمّعی است. هر نقطه نشان‌دهندهٔ یک دانشجو است و خطوط نشان دهندهٔ بالاترین رتبهٔ مشاهده شده برای هر نمره ممکن را به تصویر می‌کشند (نمودار ۸–۱).

ممکن است این سوال پیش آید که اگر دانشجویان را به صورت برعکس و به ترتیب نزولی رتبه‌بندی کنیم چه اتفاقی می‌افتد. این رتبه‌بندی صرفا تصویر آینه‌ای از نمودار نسبت به راس آن ایجاد می‌کند. نمودار حاصل همچنان یک تابع فراوانی تجمّعی تجربی است، امّا حالا خطوط نشان‌دهندهٔ پایین‌ترین رتبه مشاهده شدهٔ دانشجو برای هر نمره است (شکل ۸–۲).

توابع فراوانی تجمّعی صعودی رایج‌تر و شناخته شده‌تر از توابع نزولی هستند، امّا هر دو کاربردهای مهمی دارند. توابع فراوانی تجمّعی نزولی زمانی خیلی کاربردی است که بخواهیم توزیع‌های با چولگی زیاد را نمایش دهیم. در این خصوص در ادامه بحث خواهد شد.

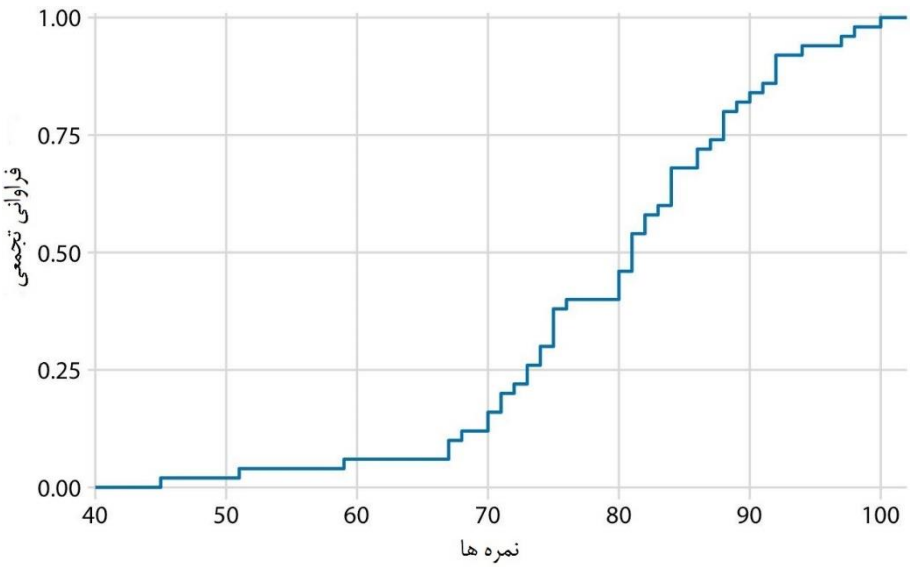


نمودار ۸–۱. نمودار تابع فراوانی تجمّعی تجربی برای نمرات در کلاس فرضی با ۵۰ دانشجو



نمودار ۸–۲. توزیع نمرات داشجویان که در قالب یک نمودار تابع فراوانی تجمّعی تجربی نزولی رسم شده است.

در عمل، عمدتاً ECDF بدون برجسته کردن نقاط داده رسم می‌شود و رتبه‌های فردی بر اساس رتبه حداکثر استاندارد می‌شود، در نتیجه محور y فراوانی تجمّعی را نشان می‌دهد (شکل ۸–۳).



نمودار ۸–۳. نمودار تابع فراوانی تجمّعی تجربی برای نمرات دانشجویان. رتبهٔ دانشجویان بر اساس تعداد کل دانشجویان استاندارد شده است لذا نقاط محور عمودی متناظر با تعداد دانشجویانی است که حداکثر آن نمره را دارند.

می‌توان مستقیماً ویژگی‌های کلیدی توزیع نمرات دانشجویان را در این نمودار بررسی کرد. به عنوان مثال، تقریباً یک چهارم دانشجویان (۲۵٪) نمره کمتری از ۷۵ داشته‌اند. مقدار میانه (متناظر با فراوانی تجمّعی ۰٫۵) ۸۱ است. تقریباً ۲۰ درصد از دانشجویان نمره ۹۰ یا بیشتر گرفته‌اند.

نمودارهای ECDF برای تعیین مرز نمرات مفید هستند زیرا کمک می‌کنند تا نقاط دقیقی را پیدا کنیم که دغدغه‌های دانشجویان را به حداقل می‌رساند. مثلاً در مثال فوق، یک خط افقی نسبتاً طولانی درست زیر نمره ۸۰ وجود دارد و به دنبال آن یک صعود تند در نمره ۸۰ مشهود است. این حالت به این خاطر است که سه دانشجو در امتحان نمره ۸۰ گرفته‌اند در حالی که بالاترین نمره بعدی ۷۶ و مربوط به یک دانشجو بود. در این سناریو، ممکن است تصمیم

بگیریم که هرکسی که نمره ۸۰ یا بیشتر دارد، قبول و هرکسی که نمره ۷۹ یا کمتر دارد، مردود شود. سه دانشجوی با نمره ۸۰ بسیار خوشحال هستند زیرا توانسته‌اند قبول شوند و دانشجوی با نمره ۷۶ متوجّه می‌شود که برای قبول شدن باید عملکرد بسیار بهتری داشته باشد. اگر نقطه برش روی ۷۷ تنظیم شود، فراوانی قبولی دقیقاً یکسان خواهد بود، امّا ممکن است دانشجوی با نمره ۷۶ به اتاق مدرّس مراجعه نماید تا در مورد نمرهٔ خود مذاکره کند. به همین ترتیب، اگر نقطه برش روی ۸۱ تنظیم شود، احتمالاً سه دانشجوی با نمره ۸۰ به اتاق مدرّس مراجعه خواهند کرد تا در مورد نمرهٔ خود مذاکره کنند.

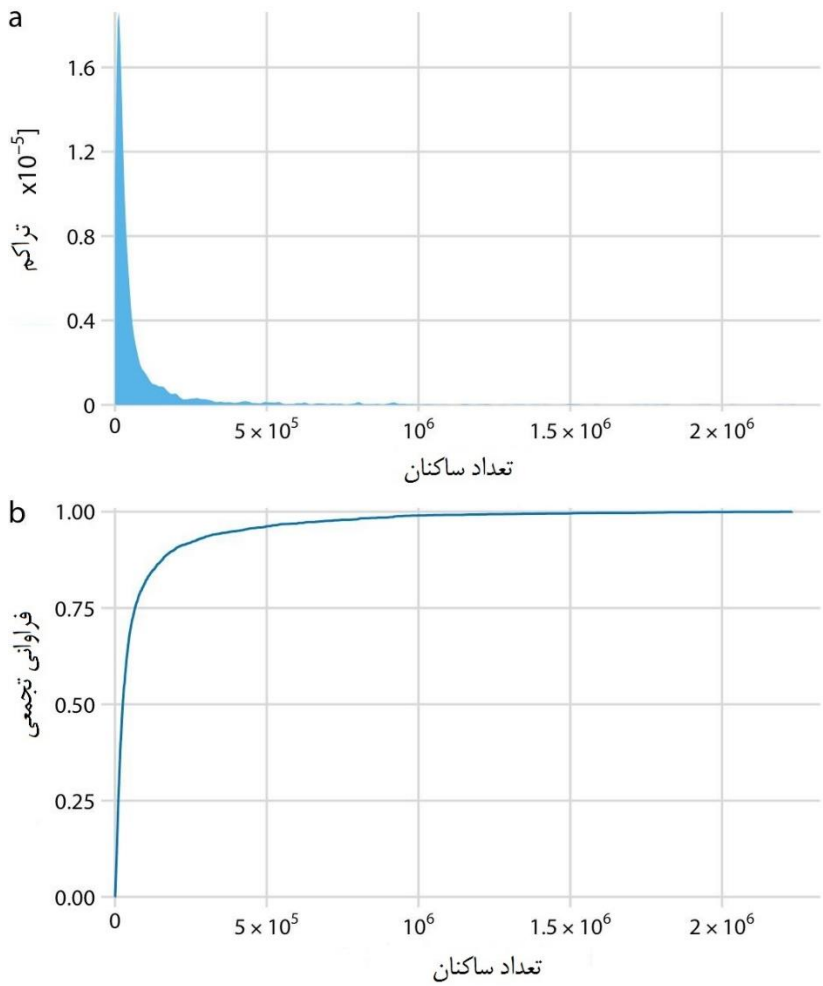
توزیع‌های با چولگی زیاد

توزیع بسیاری از داده‌های تجربی چولگی زیادی دارند، به ویژه با دم‌های بلند در سمت راست، و نمایش این توزیع‌ها می‌تواند چالش برانگیز باشد. نمونه‌هایی از این توزیع‌ها عبارتند از: تعداد افرادی که در شهرها یا شهرستان‌های مختلف زندگی می‌کنند، تعداد مخاطبین در یک شبکه اجتماعی، فراوانی ظاهر شدن کلمات در یک کتاب، تعداد مقالات دانشگاهی نوشته شده توسّط نویسندگان مختلف، دارایی خالص افراد، و تعداد تعامل پروتئین‌ها در شبکه‌های تعامل پروتئین-پروتئین [Clauset, Shalizi, and Newman 2009] . همه این توزیع‌ها وجه اشتراکی دارند که دم سمت راست آن‌ها کندتر از یک تابع نمایی تحلیل می‌رود. در عمل، این بدان معنی است که مقادیر بسیار بزرگ آنقدرها هم نادر نیستند، حتی اگر میانگین توزیع کوچک باشد. یک دسته مهم از چنین توزیع‌هایی، توزیع توانی[1] هستند، که در آن احتمال مشاهدهٔ مقداری که x برابر بزرگ‌تر از یک نقطه مرجع است، با توان x کاهش می‌یابد. برای ارائهٔ یک مثال عینی، دارایی خالص در ایالات متحده را در نظر بگیرید که منطبق بر توزیع توانی با توان ۲ می‌باشد. در هر سطح معینی از دارایی خالص (مثلاً ۱ میلیون دلار)، افرادی با نصف این مقدار دارایی خالص چهار برابر بیشتر هستند، و فراوانی افرادی که دو برابر این دارایی خالص را داشته باشند، یک چهارم می‌شود. مهم این است که اگر از ۱۰ هزار دلار یا از ۱۰۰ میلیون دلار به عنوان نقطه مرجع استفاده کنیم، همین رابطه برقرار است. به همین دلیل، توزیع‌های توانی را توزیع‌های بدون مقیاس نیز می‌نامند.

در ادامه، تعداد افرادی که بر اساس سرشماری سال ۲۰۱۰ در شهرهای مختلف ایالات متحده زندگی می‌کنند را نمایش می‌دهیم. این توزیع دارای یک دم بسیار بلند به سمت راست است. گرچه اکثر شهرها ساکنان نسبتاً کمی دارند (میانه ۲۵۸۵۷ است)، شهرهای محدودی ساکنان

1. power-law distributions

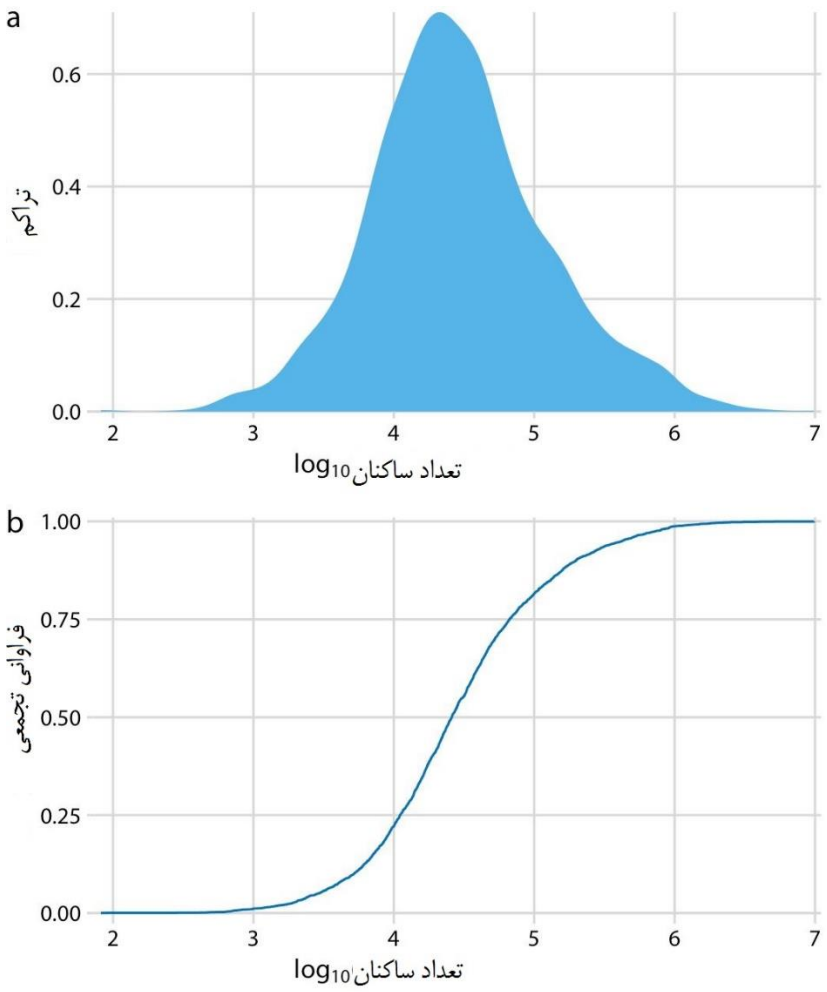
زیادی دارند (به عنوان مثال لس آنجلس با ۹۸۱۸۶۰۵ نفر). اگر بخواهیم توزیع جمعیّت را به صورت نمودار تراکمی یا ECDF رسم کنیم، شکل‌هایی خواهیم داشت که اساساً بی فایده هستند (نمودار ۸–۴).



نمودار ۸–۴. توزیع ساکنین در مناطق مختلف ایالات متحده امریکا (الف) نمودار تراکمی (ب) نمودار تابع فراوانی تجمّعی تجربی

نمودار تراکمی (شکل ۸–۴ الف) یک قلّه تیز را درست در نقطه صفر نشان می‌دهد و عملاً هیچ جزئیاتی از توزیع قابل مشاهده نیست. به طور مشابه نمودار تابع فراوانی تجمعّی تجربی (شکل ۸–۴ ب) افزایش سریعی را نزدیک به نقطه صفر نشان می‌دهد، و دوباره هیچ جزئیاتی

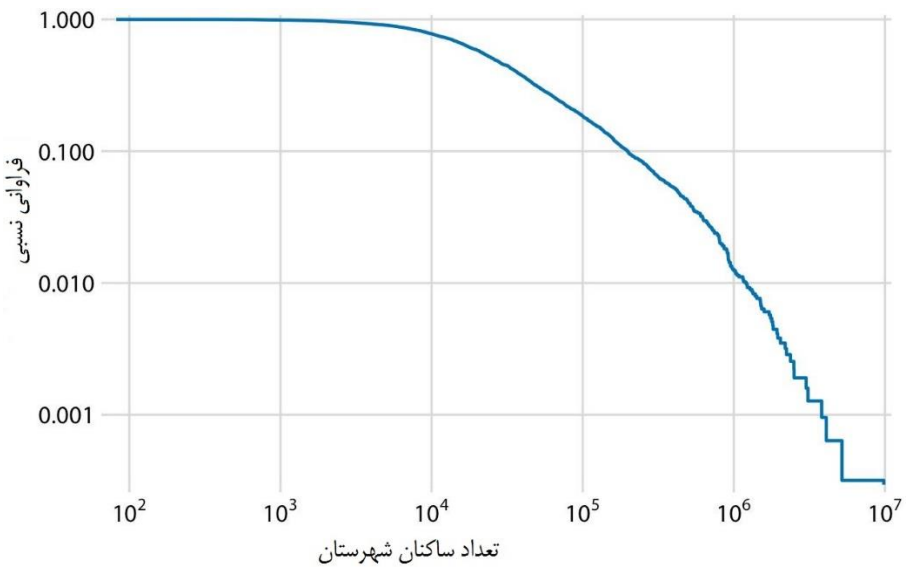
از توزیع قابل مشاهده نیست. برای این داده‌های خاص، می‌توان داده‌ها را تبدیل لگاریتمی کرد و توزیع این مقادیر را رسم نمود. این تبدیل در اینجا قابل استفاده است زیرا توزیع جمعیّت در مناطق مختلف در واقع یک توزیع توانی نیست، بلکه تقریباً یک توزیع نرمال لگاریتمی کامل است (به نمودار چندک-چندک مراجعه کنید). در واقع، نمودار تراکمی برای داده‌های تبدیل شده در مقیاس لگاریتمی کاملا زنگوله‌ای بوده و نمودار ECDF مربوطه نیز شکل سیگموئیدی خوبی را نشان می‌دهد (نمودار ۸-۵).



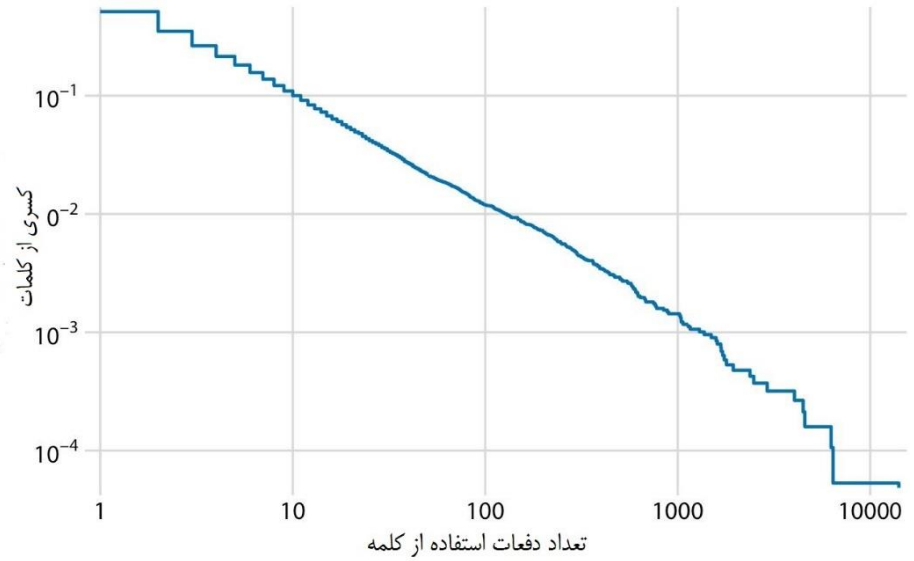
نمودار ۸-۵. توزیع لگاریتم تعداد ساکنین در مناطق مختلف ایالات متحده امریکا (الف) نمودار تراکمی (ب) نمودار فراوانی تجمّعی تجربی

برای نمایش اینکه این توزیع توانی نیست، این داده‌ها به صورت یک ECDF نزولی با محورهای x و y لگاریتمی رسم گردید. در این نمودار، توزیع توانی به صورت یک خط مستقیم ظاهر می‌شود. برای تعداد جمعیّت در شهرها، دم سمت راست تقریباً امّا نه به صورت کامل یک خط مستقیم در نمودار نزولی log-log ECDF تشکیل می‌دهد (نمودار ۸–۶).

به عنوان مثال دوّم، به توزیع فراوانی کلماتی که در رمان Moby Dick استفاده شده است، توجّه نمایید. این توزیع کاملا از توزیع توانی پیروی می‌کند. هنگامی که به صورت یک ECDF نزولی با محورهای لگاریتمی ترسیم می‌شود، تقریباً یک خط مستقیم قابل مشاهده است (شکل ۸–۷).



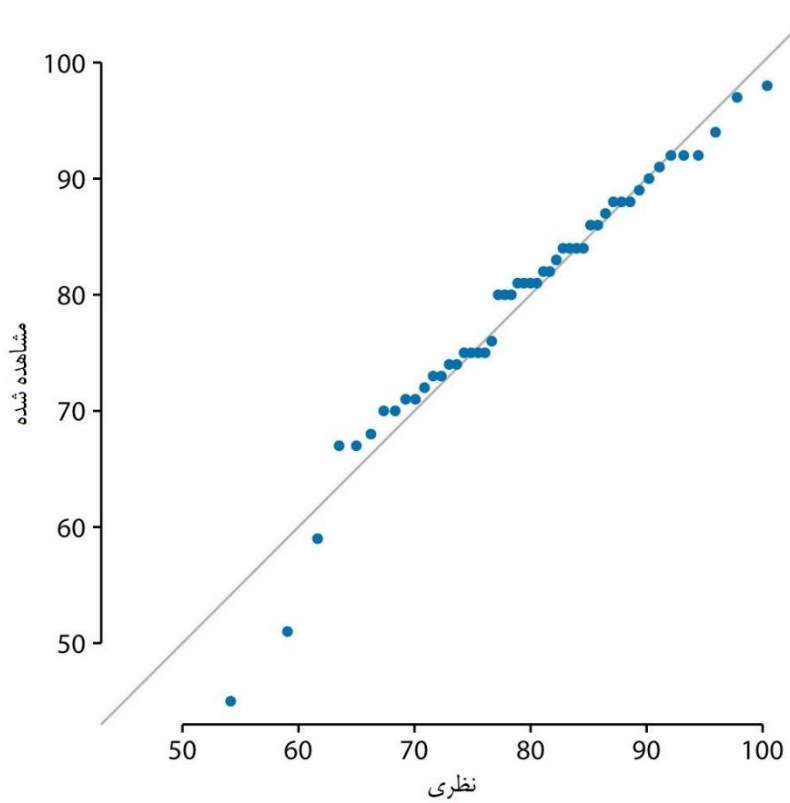
نمودار ۸–۶. فراوانی نسبی شهرهای با حداقل تعداد مشخصی از ساکنین در برابر تعداد ساکنین شهرها



نمودار ۸–۷. توزیع کلمات در رمان Moby Dick. فراوانی نسبی لغات استفاده شده حداقل به دفعات مورد نظر در برابر تعداد دفعاتی که لغات استفاده شده، نمایش داده شده است.

نمودارهای چندک–چندک (q-q)

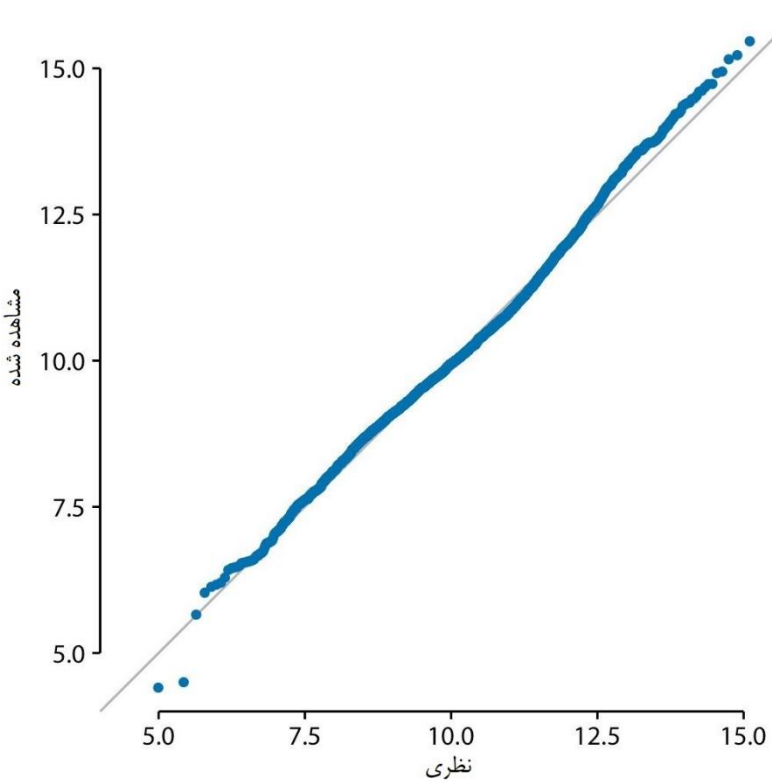
نمودارهای چندک–چندک (q-q) برای تعیین اینکه که مقادیر داده‌ها تا چه اندازه از یک توزیع مشخّص پیروی می‌کنند، مفید هستند. دقیقا همانند ECDF ها، نمودارهای q-q نیز بر اساس رتبه‌بندی داده‌ها و ترسیم رابطه بین رتبه‌ها و مقادیر واقعی است. با این حال، در نمودارهای q-q رتبه‌ها به صورت مستقیم ترسیم نمی‌شوند. بلکه از آن‌ها برای پیش‌بینی اینکه اگر داده‌ها از توزیع مشخصی پیروی کنند، یک داده مشخّص در کجا قرار خواهد گرفت، استفاده می‌شود. معمولاً نمودارهای q-q با استفاده از توزیع نرمال به عنوان توزیع مرجع رسم می‌شوند. به عنوان یک مثال عینی، فرض کنید که مقادیر واقعی داده‌ها دارای میانگین ۱۰ و انحراف معیار ۳ است. سپس، با فرض توزیع نرمال، انتظار داریم که یک داده با رتبه صدک ۵۰ در موقعیّت مقدار ۱۰ (میانگین)، یک داده در صدک ۸۴ در موقعیّت مقدار ۱۳ (یک انحراف استاندارد بالاتر از میانگین) و یک داده در صدک ۲٫۳ در موقعیّت مقدار ۴ (دو انحراف استاندارد کمتر از میانگین) قرار گیرند. می‌توان این محاسبه را برای تمام داده‌ها انجام داده و سپس مقادیر مشاهده شده (یعنی مقادیر در مجموعهٔ داده‌ها) را در برابر مقادیر نظری (یعنی مقادیر مورد انتظار با توجّه به رتبه هر داده و توزیع مرجع فرضی) رسم نمود. وقتی این روش را برای توزیع نمرات دانشجویان ابتدای این فصل انجام می‌دهیم، شکل ۸–۸ به دست می‌آید.



نمودار ۸–۸. نمودار q-q برای نمرات دانشجویان فرضی

خط مستقیم در این نمودار خط رگرسیون نیست، بلکه نقاطی را نشان می‌دهد که x برابر با y است، یعنی جایی که مقادیر مشاهده شده برابر با مقادیر نظری است. تا جایی که نقاط روی آن خط قرار می‌گیرند، داده‌ها از توزیع فرضی پیروی می‌کنند (در اینجا، توزیع نرمال). همانطور که مشاهده می‌شود نمرات دانشجویان عمدتاً از توزیع نرمال پیروی می‌کند، با کمی انحراف در پایین و بالای توزیع (چند دانشجو در هر دو سر طیف بدتر از حد انتظار عمل کردند). انحراف از توزیع در انتهای بالایی ناشی از حداکثر مقدار نمره ۱۰۰ در آزمون فرضی است. فارغ از اینکه بهترین دانشجو چقدر خوب است، دانشجویان حداکثر می‌توانند نمره ۱۰۰ کسب کنند. همچنین می‌توان از نمودار q-q برای آزمایش ادعای مطرح شده در ابتدای این فصل مبنی بر اینکه جمعیّت در مناطق مختلف ایالات متحده از توزیع نرمال لگاریتمی پیروی می‌کند، استفاده نمود. اگر این داده‌ها توزیع نرمال لگاریتمی داشته باشند، آنگاه مقادیر تبدیل شده به لگاریتم آن‌ها توزیع نرمال خواهند داشت و بنابراین باید دقیقا روی خط x = y قرار گیرند. بعد

از ترسیم این نمودار، مشاهده می‌شود که توافق بین مقادیر مشاهده شده و مورد انتظار استثنایی است (شکل ۸–۹). این نشان می‌دهد که توزیع جمعیّت در شهرهای مختلف واقعاً نرمال است.



نمودار ۸–۹. نمودار q-q برای لگاریتم تعداد ساکنین شهرهای مختلف ایالات متحده امریکا

فصل ۹

ترسیم همزمان چندین توزیع

حالات مختلفی وجود دارد که بخواهیم چندین توزیع را به طور همزمان ترسیم کنیم. به عنوان مثال، داده‌های آب و هوا را در نظر بگیرید. ممکن است بخواهیم نحوهٔ تغییر دما در ماه‌های مختلف را ترسیم کنیم و در عین حال توزیع دمای مشاهده شده را در هر ماه نشان دهیم. این حالت مستلزم نمایش همزمان ۱۲ توزیع دما (یکی برای هر ماه) است. هیچ یک از نمودارهای مورد بحث قبلی (فصل ۷ یا ۸) برای این مثال مناسب نیستند. در عوض، رویکردهای قابل اجرا شامل نمودار جعبه‌ای[1]، نمودار ویولن[2] و نمودار خط الراس[3] هستند.

هر زمان که با توزیع‌های زیادی سروکار داریم، بهتر است به متغیّرها در قالب متغیّر پاسخ و یک یا چند متغیّر گروه‌بندی نگاه کنیم. متغیّر پاسخ، متغیّری است که می‌خواهیم توزیع‌های آن را نشان دهیم. متغیّرهای گروه‌بندی، زیرمجموعه‌هایی از داده‌ها را با توزیع متمایزی از متغیّر پاسخ نشان می‌دهد. برای مثال، برای نمایش توزیع دما در ماه‌های مختلف، دما متغیّر پاسخ و ماه متغیّر گروه‌بندی است. تمام روش‌های مورد بحث در این فصل، متغیّر پاسخ را در امتداد یک محور و متغیّر(های) گروه‌بندی را در امتداد محور دیگر ترسیم می‌کنند. در بخش‌های بعدی، ابتدا رویکردهایی را توضیح خواهیم داد که متغیّر پاسخ را در امتداد محور عمودی نشان

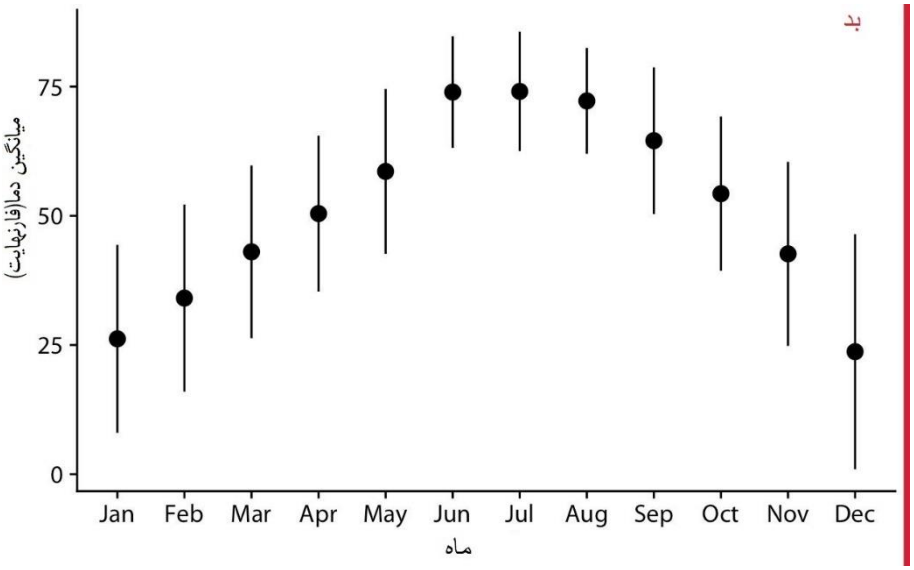
1. boxplot
2. violin plot
3. ridgeline plot

می‌دهند و سپس رویکردهایی را توضیح خواهیم داد که متغیّر پاسخ را در امتداد محور افقی نشان می‌دهند. در تمام مواردی که مورد بحث قرار خواهد گرفت، می‌توان محورها را جابجا نمود و به نمودار جایگزینی رسید. در اینجا اشکال متعارف نمودارهای مختلف را نشان می‌دهیم.

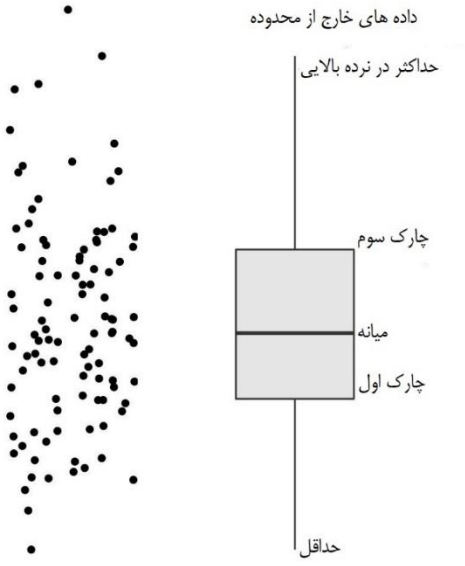
ترسیم توزیع‌ها در امتداد محور عمودی

ساده‌ترین روش برای نشان دادن همزمان توزیع‌های متعدد، استفاده از میانگین یا میانه به‌عنوان نقاط داده به همراه شاخصی از پراکندگی حول میانگین یا میانه به صورت میله‌های خطا است. شکل ۹–۱ این رویکرد را برای توزیع دمای ماهانه در لینکلن، نبراسکا، در سال ۲۰۱۶ نشان می‌دهد. این نمودار به عنوان «بد» برچسب خورده است زیرا مشکلات متعددی در این رویکرد وجود دارد. اوّل، با نمایش هر توزیع تنها با یک نقطه و دو میله خطا، اطّلاعات زیادی را در مورد داده‌ها از دست می‌دهیم. دوّم، به سرعت نمی‌توان فهمید که این نقاط نشان‌دهندهٔ چیست، حتّی اگر اکثر خوانندگان حدس بزنند که آن‌ها میانگین یا میانه را نشان می‌دهند. سوّم، مشخّص نیست که میله‌های خطا نشان‌دهندهٔ چیست. آیا آن‌ها نشان‌دهندهٔ انحراف معیار داده‌ها هستند یا خطای استاندارد میانگین، فاصله اطمینان ۹۵ درصد یا چیز دیگر؟ هیچ استاندارد پذیرفته شده‌ای وجود ندارد. با خواندن زیرنویس شکل ۹–۱، می‌بینیم که آن‌ها در اینجا دو برابر انحراف معیار میانگین دمای روزانه را نشان می‌دهند، به این معنی که محدوده‌ای را نشان می‌دهند که تقریباً ۹۵ درصد از داده‌ها را شامل می‌شود. با این حال، میله‌های خطا معمولاً برای ترسیم خطای استاندارد (یا دو برابر خطای استاندارد که معادل فاصلهٔ اطمینان ۹۵ درصد می‌باشد) استفاده می‌شوند و خوانندگان به آسانی ممکن است خطای استاندارد را با انحراف معیار اشتباه بگیرند. خطای استاندارد نشان‌دهندهٔ این است که تخمین ما از میانگین چقدر دقیق است، در حالی که انحراف معیار نشان‌دهندهٔ پراکندگی داده‌ها حول میانگین می‌باشد. ممکن است یک مجموعه داده خطای استاندارد میانگین بسیار کوچک و از سوی دیگر انحراف معیار بسیار بزرگی داشته باشد. چهارم، اگر چولگی در داده‌ها وجود داشته باشد، میله‌های خطای متقارن، گمراه‌کننده هستند، که در اینجا نیز صادق است و این حالت تقریباً همیشه برای مجموعه داده‌های دنیای واقعی نیز مصداق دارد.

ما می‌توانیم هر چهار نقص شکل ۹–۱ را با استفاده از یک روش سنّتی و رایج برای نمایش توزیع‌ها، یعنی نمودار جعبه‌ای، برطرف کنیم. نمودار جعبه‌ای داده‌ها را به چارک‌ها تقسیم می‌کند و آن‌ها را به شیوه‌ای استاندارد نمایش می‌دهد (شکل ۹–۲).



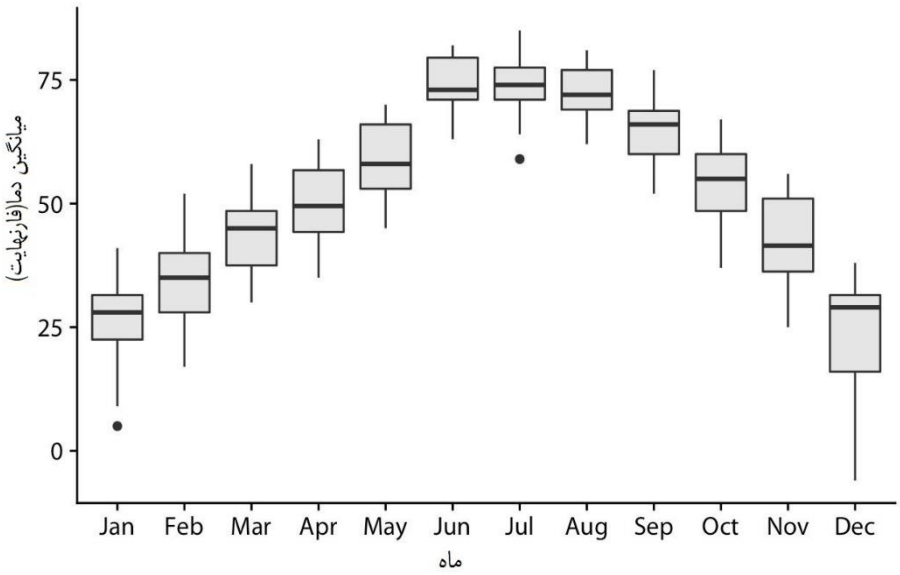
شکل ۹–۱. میانگین دمای روزانه در لینکلن، نبراسکا در سال ۲۰۱۶. نقاط نشان‌دهندهٔ میانگین روزانه دما برای هر ماه، که ماحصل میانگین در تمام روزهای ماه می‌باشد، و میله‌های خطا نشان‌دهندهٔ دو برابر انحراف معیار میانگین دمای روزانه در هر ماه است. این نمودار به عنوان «بد» برچسب خورده است زیرا میله‌های خطا معمولاً برای نمایش عدم قطعیّت یک تخمین استفاده می‌شود، نه تغییرپذیری در یک جمعیّت. منبع داده: آب و هوای زیرزمینی.



شکل ۹–۲. آناتومی یک نمودار جعبه‌ای. ابری از نقاط (چپ) و نمودار جعبه‌ای متناظر (راست) نشان داده شده است.

فقط مقادیر محور y در نمودار جعبه‌ای در شکل ۹–۲ نشان داده شده است. خط وسط جعبه نمودار نشان‌دهندهٔ میانه است و جعبه ۵۰ درصد وسط داده‌ها را در بر می‌گیرد. خطوط عمودی که از جعبه به سمت بالا و پایین امتداد می‌یابند سبیل[1] نامیده می‌شوند. سبیل‌های بالا و پایین بر اساس اینکه کدام یکی از این دو حالت، خطوط کوتاه‌تری را به دست می‌دهد رسم می‌شود: حداکثر و حداقل مقادیر داده‌ها یا حداکثر یا حداقل مقادیری که در بازه ۱٫۵ برابری ارتفاع جعبه قرار داشته باشد. به فواصل ۱٫۵ برابر ارتفاع جعبه در هر جهت نرده بالا و پایین می‌گویند. نقاط داده‌ای که فراتر از نرده‌ها قرار می‌گیرند به عنوان نقاط پرت در نظر گرفته شده و معمولاً به صورت نقاط مجزا نشان داده می‌شوند.

نمودارهای جعبه‌ای، ساده و در عین حال آموزنده هستند و وقتی در کنار یکدیگر ترسیم می‌شوند، می‌توان توزیع‌های متعددی را همزمان نمایش داد. برای داده‌های دمای لینکلن، از نمودار جعبه‌ای در شکل ۹–۳ استفاده شده است. در این شکل، می‌توان دید که دما در ماه دسامبر توزیعی اریب دارد (اکثر روزها نسبتاً سرد و تعداد کمی از آن‌ها بسیار سرد هستند) و در برخی از ماه‌های دیگر، مانند جولای، اصلاً اریب نیست.

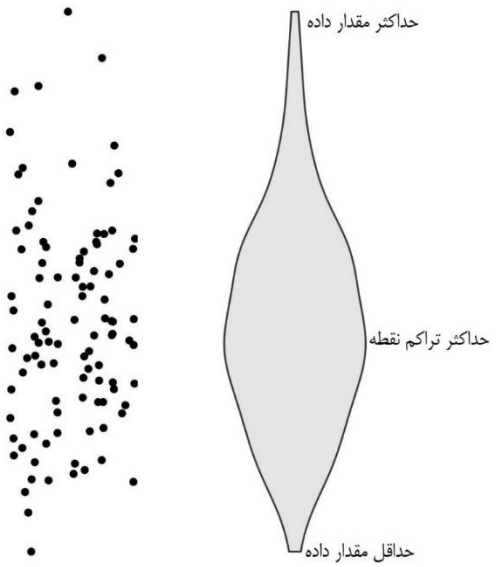


شکل ۹–۳. میانگین دمای روزانه در لینکلن، نبراسکا، که توسّط نمودارهای جعبه‌ای نمایش داده شده است. منبع داده: آب و هوای زیرزمینی.

1. whiskers

نمودار جعبه‌ای توسّط آماردان جان توکی[1] در اوایل دهه ۱۹۷۰ اختراع شد و به سرعت محبوبیّت یافتند زیرا بسیار آموزنده بوده و در عین حال به راحتی ترسیم می‌شدند، روش ترسیم با دست در آن دورهٔ زمانی روش مرسوم رسم نمودارها بود. با این حال، با محاسبات مدرن و قابلیت‌های جدید ترسیم نمودار، ما دیگر محدود به ترسیم نمودار با دست نیستیم. لذا، اخیراً شاهد جایگزینی نمودارهای جعبه‌ای با نمودار ویولن هستیم (شکل ۹-۴). برای هر داده‌ای که بتوان نمودار جعبه‌ای ترسیم کرد، نمودار ویلون نیز می‌توان ترسیم نمود و این نمودارها تصویر بسیار دقیق‌تری از داده‌ها ارائه می‌دهد. به طور خاص، نمودار ویولن می‌تواند داده‌های با دو نما (دو قله‌ای) را به طور دقیق نشان می‌دهند، در حالی که نمودار جعبه‌ای این امکان را ندارد.

فقط مقادیر محور y در طرح ویولن ترسیم می‌شود. عرض ویولن در یک مقدار معین محور y نشان‌دهندهٔ چگالی نقطه به ازای آن مقدار y است. از نظر فنّی، طرح ویولن یک تخمین چگالی است که ۹۰ درجه چرخیده و سپس تصویر آینه‌ای آن روی خودش رسم شده است (فصل ۷). بنابراین ویولن‌ها متقارن هستند. نمودار ویولن با حداقل مقدار داده‌ها شروع شده و با حداکثر مقدار داده‌ها پایان می‌یابند. ضخیم‌ترین قسمت ویولن منطبق بر پرتراکم‌ترین نقطه در مجموعه داده است.



شکل ۹-۴. آناتومی نمودار ویولن. ابری از نقاط (چپ) و طرح ویولن متناظر آن (راست) نشان داده شده است.

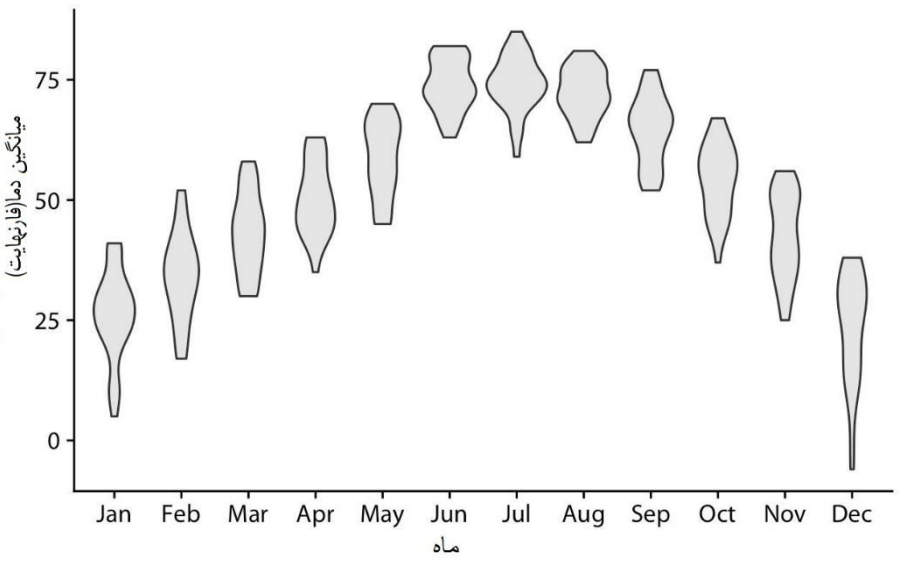
1. John Tukey

قبل از استفاده از نمودار ویولن برای ترسیم توزیع‌ها، مطمئن شوید که در هر گروه به اندازهٔ کافی نقاط داده دارید تا تراکم نقاط را به صورت خطوط صاف نشان دهید.

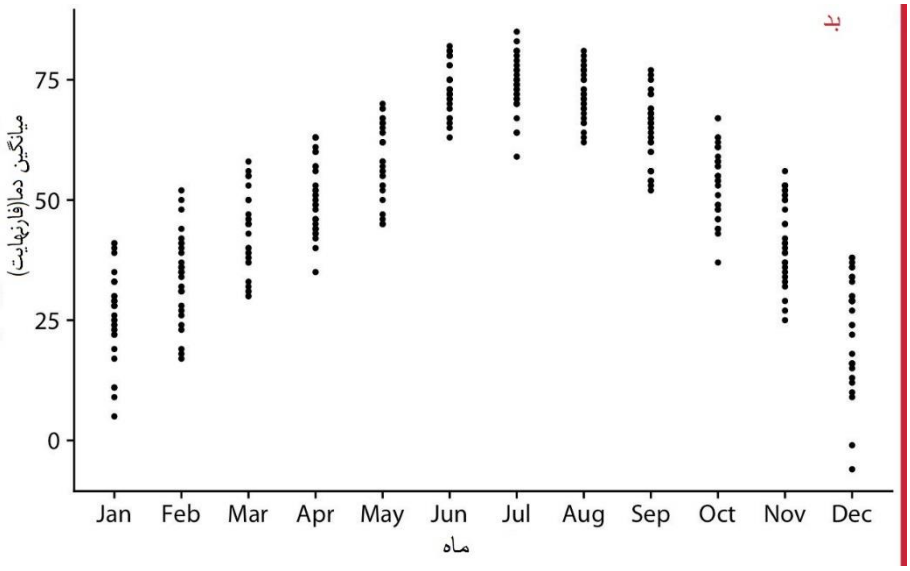
هنگامی که داده‌های دمای لینکلن را با ویولن ترسیم می‌کنیم، شکل ۹–۵ به دست می‌آید. اکنون می‌توان دید که برخی از ماه‌ها داده‌های نسبتاً دو نمایی دارند. به عنوان مثال، به نظر می‌رسد ماه نوامبر دارای دو خوشهٔ دمایی بوده است، یکی در محدودهٔ ۵۰ درجهٔ فارنهایت و دیگری در محدودهٔ ۳۵ درجهٔ فارنهایت.

از آنجایی که نمودارهای ویولن از تخمین‌های تراکمی به دست می‌آیند، دارای کاستی‌های مشابهی هستند. به طور مشخص، آن‌ها می‌توانند در جایی که داده‌ای وجود ندارد یا داده‌ها پراکنده هستند به اشتباه تجمّع داده‌ها را نشان دهند. ما می‌توانیم این مسائل را با ترسیم مستقیم تمام نقاط داده به صورت نقطه‌ای اصلاح کنیم (شکل ۹–۶). به نمودار حاصل نمودار نواری[1] می‌گویند. نمودارهای نواری اصولاً مناسب هستند، به شرطی اطمینان حاصل کنیم که نقاط زیادی را روی هم ترسیم نمی‌کنیم. یک راه حل ساده برای مساله ترسیم بیش از حد این است که نقاط را تا حدی در امتداد محور x با اضافه کردن مقداری تغییر[2] تصادفی در این محور پخش نمود (شکل ۹–۷). به این روش لرزانش[3] اطلاق می‌شود.

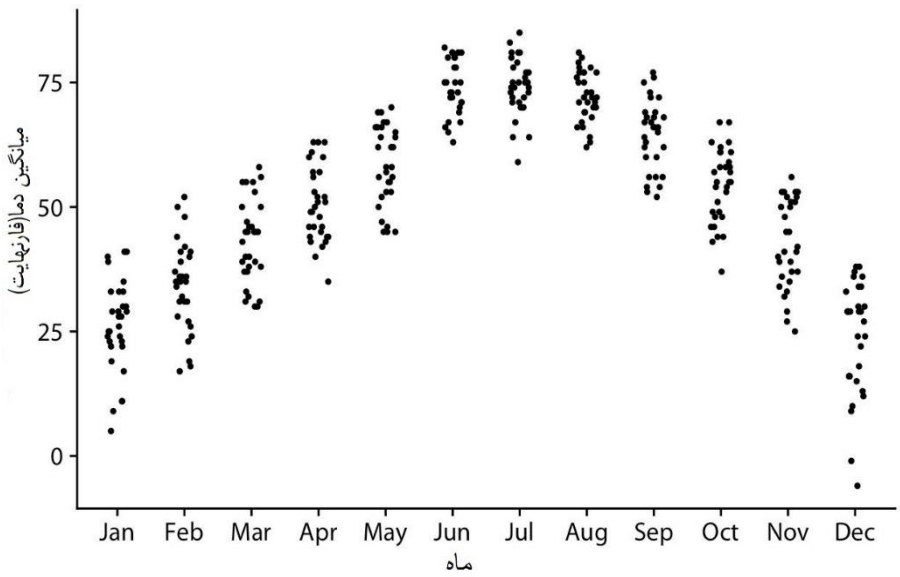
1. strip chart
2. noise
3. jittering



شکل ۹–۵. میانگین دمای روزانه در لینکلن، نبراسکا که توسّط نمودار ویولن نمایش داده شده است. منبع داده: آب و هوای زیرزمینی.



شکل ۹–۶. میانگین دمای روزانه در لینکلن، نبراسکا، که توسّط نمودارهای نواری نمایش داده شده است. هر نقطه نشان‌دهندهٔ میانگین دمای یک روز است. این نمودار به عنوان «بد» برچسب زده شده است زیرا نقاط زیادی روی هم ترسیم شده است که نمی‌توان مشخّص کرد کدام دما در هر ماه بیشترین فراوانی را داشته است. منبع داده: آب و هوای زیرزمینی.

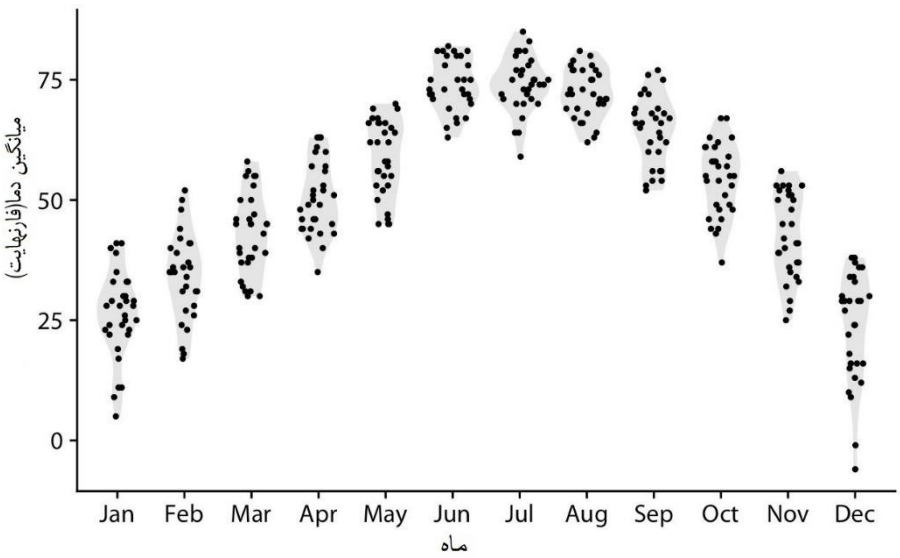


شکل ۹–۷. میانگین دمای روزانه در لینکلن، نبراسکا، که توسّط نمودارهای نواری نمایش داده شده است. نقاط در امتداد محور x لرزانده شده‌اند تا چگالی نقاط را در هر مقدار دما بهتر نشان دهند. منبع داده: آب و هوای زیرزمینی.

هر زمان که مجموعهٔ داده برای ترسیم نمودار ویولن بسیار کم باشد، رسم داده‌های خام به صورت نقاط جداگانه گزینهٔ جایگزین خواهد بود.

در نهایت، می‌توانیم دو نمودار فوق را ادغام کنیم: پخش کردن نقاط به نسبت چگالی نقطه‌ای در محور y. این روش نمودار سینا[1] نام دارد که آن را می‌توان به عنوان ترکیبی بین نمودار ویولن و نقاط لرزان در نظر گرفت و علاوه بر اینکه تک تک نقاط را نشان می‌دهد، توزیع‌ها را نیز نمایش می‌دهد. در شکل ۹–۸، نمودارهای سینا در بالای ویولن‌ها ترسیم شده است تا رابطه بین این دو روش را ملاحظه کنیم.

۱. نمودار سینا به افتخار سینا هادی سوهی که دانشجوی دانشگاه کوپنهاگن در دانمارک بود، نامگذاری شده است. او اولین نسخه از کدهایی را نوشت که محققین برای ترسیم این نمودار استفاده نمودند. (Frederik O.Bagger)



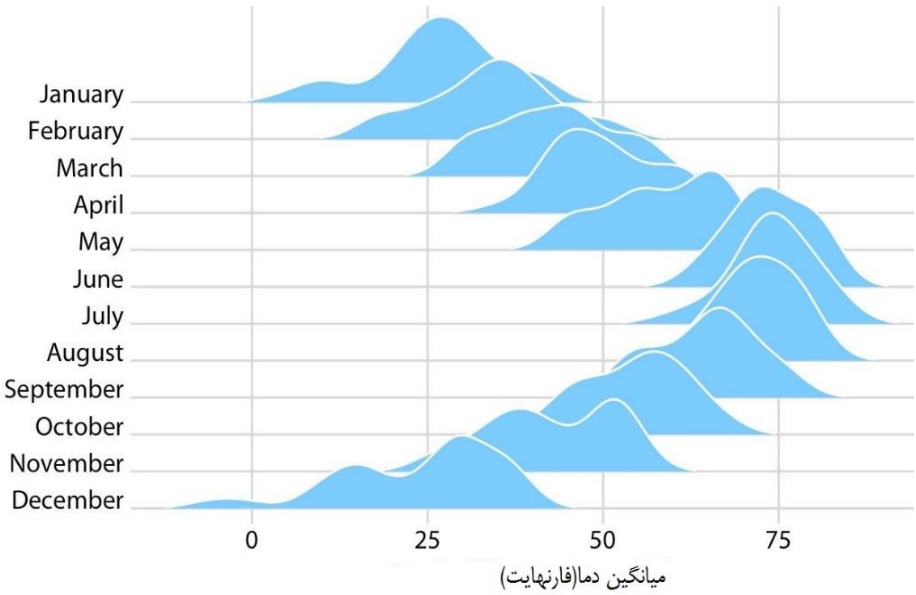
شکل ۹–۸. میانگین دمای روزانه در لینکلن، نبراسکا، که توسّط نمودارهای سینا (ترکیبی از نقاط منفرد و ویولن) نمایش داده شده است. نقاط در امتداد محور x متناسب با چگالی نقطه در دمای مربوطه لرزانده شده‌اند. در اینجا، نمودارهای سینا روی نمودار ویولن قرار گرفته‌اند. منبع داده: آب و هوای زیرزمینی.

ترسیم توزیع‌ها در امتداد محور افقی

در فصل ۷، توزیع‌ها را در امتداد محور افقی با استفاده از نمودارهای هیستوگرام و چگالی ترسیم کردیم. در اینجا، این ایده را با ترسیم نمودارهای توزیع در جهت عمودی گسترش خواهیم داد. نمودار حاصل را نمودار خط الراس می‌نامند، زیرا این نمودارها شبیه خط الراس کوه هستند. اگر بخواهید روند توزیع‌ها را در طول زمان نشان دهید، نمودارهای خط الراس گزینهٔ بسیار مناسبی هستند.

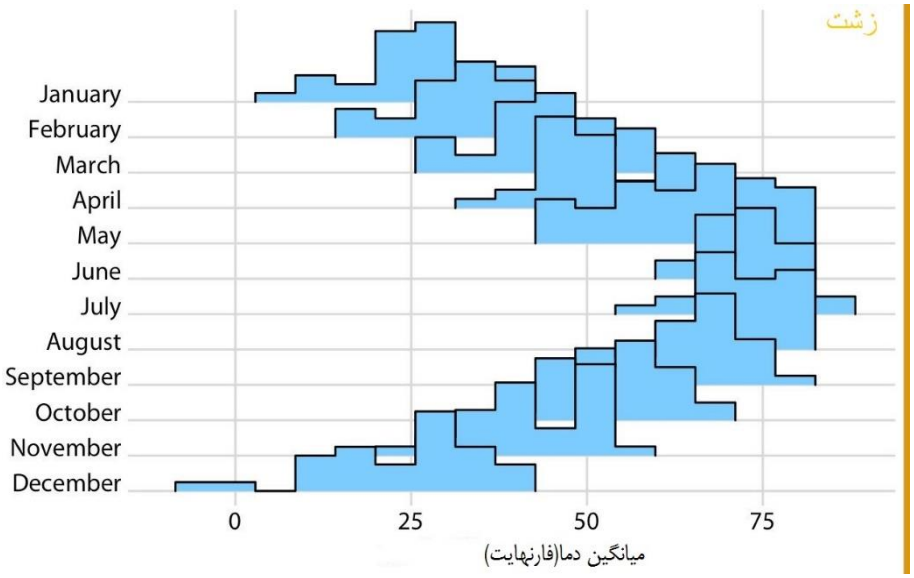
نمودار خط الراس استاندارد از تخمین‌های چگالی استفاده می‌کند (شکل ۹–۹). این نمودار کاملاً با طرح ویولن مرتبط است، امّا اغلب امکان درک شهودی‌تری از داده‌ها را فراهم می‌کند. به عنوان مثال، دو خوشهٔ دما در محدوده ۳۵ و ۵۰ درجهٔ فارنهایت در ماه نوامبر در شکل ۹–۹ بسیار واضح‌تر از شکل ۹–۵ هستند.

از آنجایی که محور x متغیّر پاسخ و محور y متغیّر گروه‌بندی را نشان می‌دهد، هیچ محور جداگانه‌ای برای تخمین چگالی در نمودار خط الراس وجود ندارد. تخمین تراکمی در کنار متغیّر گروه‌بندی نشان داده شده است. این کار هیچ تفاوتی با طرح ویولن ندارد، که در آن تراکم‌ها نیز در کنار متغیّر گروه‌بندی، بدون مقیاس جداگانه نشان داده می‌شوند. در هر دو مورد، هدف نمودار نشان دادن مقادیر چگالی خاص نیست، بلکه در عوض امکان مقایسهٔ آسان اشکال چگالی و ارتفاع نسبی در گروه‌ها را فراهم می‌کند.



شکل ۹–۹. دما در لینکلن، نبراسکا، در سال ۲۰۱۶، که توسّط نمودار خط الراس نشان داده شده است. برای هر ماه، توزیع میانگین دمای روزانه اندازه‌گیری شده برحسب درجهٔ فارنهایت نشان داده شده است. ایده اصلی نمودار: Wehrwein 2017. منبع داده: آب و هوای زیرزمینی.

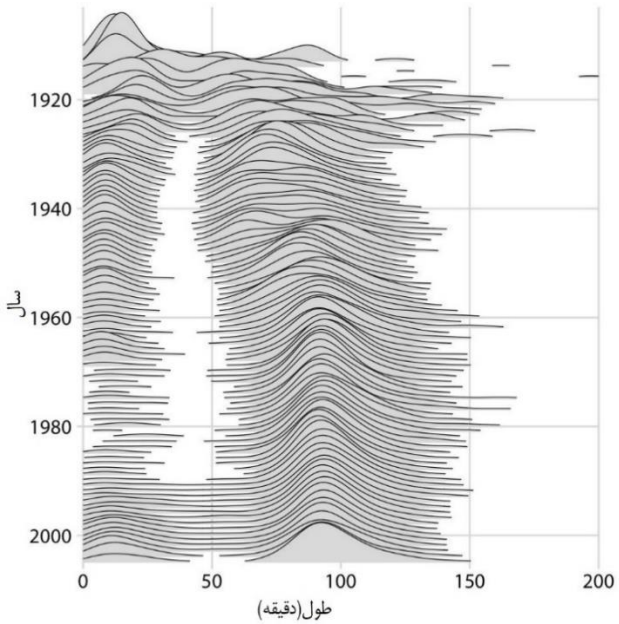
در اصل، می‌توانیم از هیستوگرام به جای نمودارهای چگالی در ترسیم نمودار خط الراس استفاده کنیم. با این حال، نمودار حاصل اغلب خیلی خوب به نظر نمی‌رسد (شکل ۹–۱۰). مشکلات مشابه نمودار حاصل، شبیه مشکلات هیستوگرام‌های انباشته یا همپوشان هستند (بخش نمایش همزمان چندین توزیع را ببینید). از آنجایی که خطوط عمودی در این هیستوگرام‌های خط الراس دقیقاً در مقادیر x یکسانی ظاهر می‌شوند، میله‌های هیستوگرام‌های مختلف به طرز گیج‌کننده‌ای با یکدیگر همسو می‌شوند. لذا بهتر است چنین هیستوگرام‌های همپوشانی ترسیم نشود.



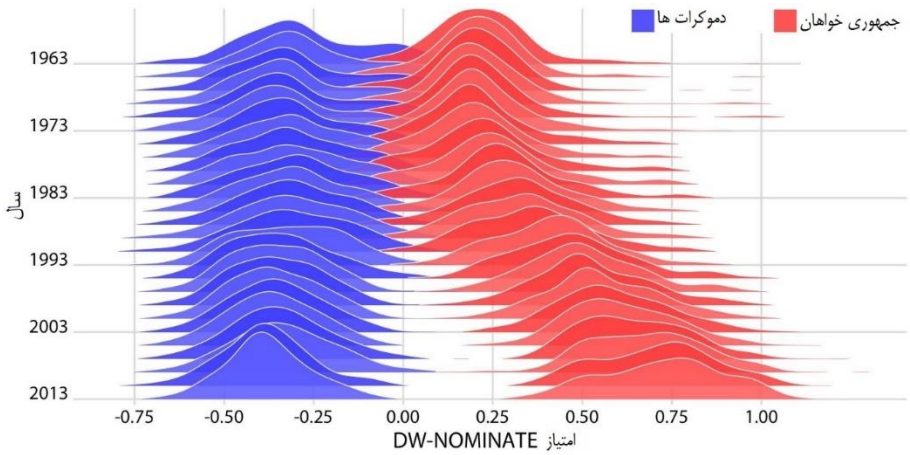
شکل ۹–۱۰. دما در لینکلن، نبراسکا، در سال ۲۰۱۶، که توسّط نمودار هیستوگرام خط الراس نمایش داده شده است. هیستوگرام‌ها از نظر بصری به خوبی از هم جدا نمی‌شوند و نمودار کلی بسیار شلوغ و گیج‌کننده است. منبع داده: آب و هوای زیرزمینی.

نمودارهای خط الراس می‌توانند برای نمایش توزیع‌های بسیار زیادی به کار روند. به عنوان مثال، شکل ۹–۱۱ توزیع مدّت زمان فیلم‌ها را از سال ۱۹۱۳ تا ۲۰۰۵ نشان می‌دهد. این شکل شامل تقریباً ۱۰۰ توزیع متمایز است و در عین حال خواندن آن بسیار آسان است. می‌توان دید که مدّت زمان فیلم‌ها در دهه ۱۹۲۰ بسیار متنوّع بوده است، امّا از حدود سال ۱۹۶۰ استاندارد طول فیلم تقریباً ۹۰ دقیقه شده است.

اگر بخواهیم دو روند را در طول زمان مقایسه کنیم، نمودارهای خط الراس گزینهٔ مناسبی هستند. مثلاً اگر بخواهیم الگوهای رأی اعضای دو حزب مختلف را تحلیل کنیم، این نمودار قابل استفاده است. می‌توان این مقایسه را با تناوبی کردن توزیع‌ها به صورت عمودی بر حسب زمان و استفاده از رنگ متفاوت برای هر حزب در هر نقطه زمانی انجام داد (شکل ۹–۱۲).



شکل ۹–۱۱. تکامل مدّت زمان فیلم‌ها در طول زمان. از دهه ۱۹۶۰، طول اکثر فیلم‌ها تقریباً ۹۰ دقیقه بوده است. منبع داده‌ها: پایگاه اینترنتی فیلم‌های اینترنتی (IMDB).



شکل ۹–۱۲. الگوهای رأی دهی در مجلس نمایندگان ایالات متحده به طور فزاینده‌ای قطبی شده است. امتیازات DW-NOMINATE اغلب برای مقایسه الگوهای رأی دهی نمایندگان بین احزاب و در طول زمان استفاده می‌شود. در اینجا، توزیع امتیاز برای کنگره از سال ۱۹۶۳ تا ۲۰۱۳ به طور جداگانه برای دموکرات‌ها و جمهوری خواهان نشان داده شده است. داده‌ها مربوط به سال اول هر کنگره است. ایده اصلی نمودار: [McDonald 2017]. منبع داده: کیت پول.

فصل ۱۰

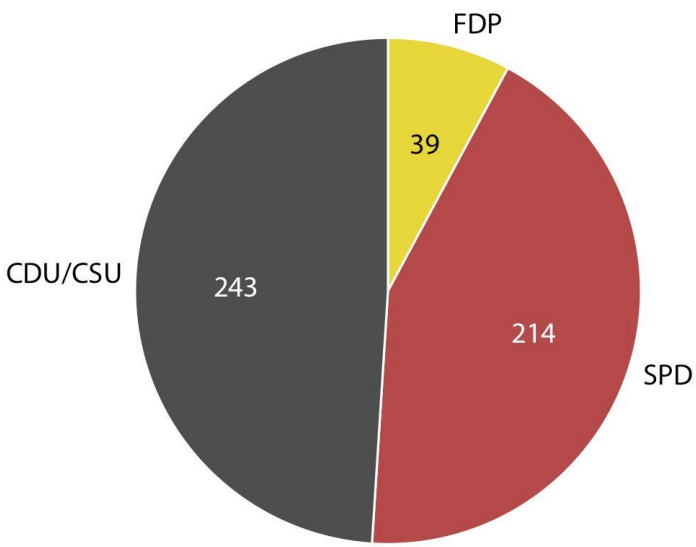
نمایش نسبت‌ها

ما اغلب می‌خواهیم نحوه تقسیم شدن یک گروه، دسته یا مقدار را به بخش‌های منفرد که هر یک نسبتی از کل را نشان می‌دهند، نمایش دهیم. نمونه‌های رایج عبارتند از نسبت مردان و زنان در گروهی از مردم، درصد رأی‌دهندگان به احزاب سیاسی مختلف در یک انتخابات یا سهم بازار برای هر شرکت. کهن الگوی این نمایش، نمودار دایره‌ای است که در هر سخنرانی تجاری وجود دارد امّا در میان دانشمندان در حوزه داده خوش نام نیست. همانطور که خواهیم دید، نمایش نسبت‌ها می‌تواند چالش برانگیز باشد، به ویژه زمانی که کل داده‌ها به قسمت‌های مختلفی تقسیم می‌شود یا زمانی که می‌خواهیم تغییرات نسبت‌ها را در طول زمان یا در شرایط مختلف مشاهده کنیم. هیچ نمایش ایده‌آل واحدی وجود ندارد که بتوان همواره از آن استفاده کرد. برای نشان دادن این موضوع، چند سناریو مختلف را مورد بحث قرار خواهیم داد که هر کدام نیازمند نوع متفاوتی از نحوه نمایش است.

به یاد داشته باشید، همیشه باید نحوه نمایشی را انتخاب کنید که بیشترین تناسب را با مجموعه داده‌های شما داشته و ویژگی‌های کلیدی داده‌ها را که قصد نمایش آن را دارید، به بهترین نحو برجسته کند.

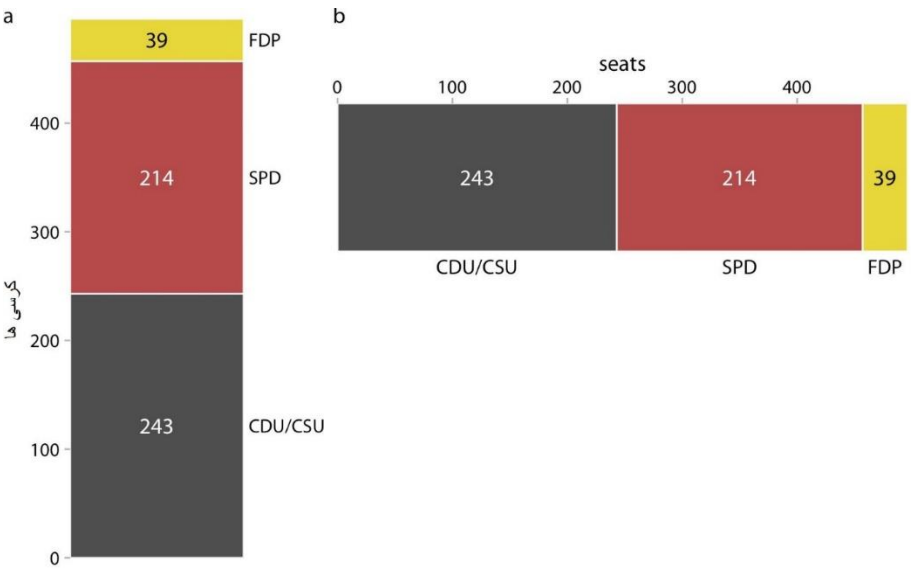
نمونه‌ای برای نمودارهای دایره‌ای

از سال ۱۹۶۱ تا ۱۹۸۳، پارلمان آلمان (به نام بوندستاگ) از اعضای سه حزب مختلف CDU/CSU، SPD و FDP تشکیل شده بود. در بیشتر این مدّت زمان،CDU/CSU و SPD تعداد کرسی‌های تقریباً مشابهی داشتند، در حالی که FDP عمدتاً تنها بخش کوچکی از کرسی‌ها را در اختیار داشت. مثلاً در بوندستاگ هشتم، از ۱۹۷۶ تا ۱۹۸۰، حزب CDU/CSU ۲۴۳ کرسی، حزب SPD ۲۱۴ و حزب FDP ۳۹ کرسی (در مجموع ۴۹۶ کرسی) را به خود اختصاص داده بودند. چنین داده‌های پارلمانی معمولاً به صورت نمودار دایره‌ای نمایش داده می‌شوند (نمودار ۱۰–۱).



نمودار ۱۰–۱. ترکیب حزب هشتمین بوندستاگ آلمان، ۱۹۷۶–۱۹۸۰، که به صورت نمودار دایره‌ای نشان داده شده است. این نمودار نشان می‌دهد که ائتلاف حاکم SPD و FDP برتری اندکی بر حزب CDU/CSU داشته است.

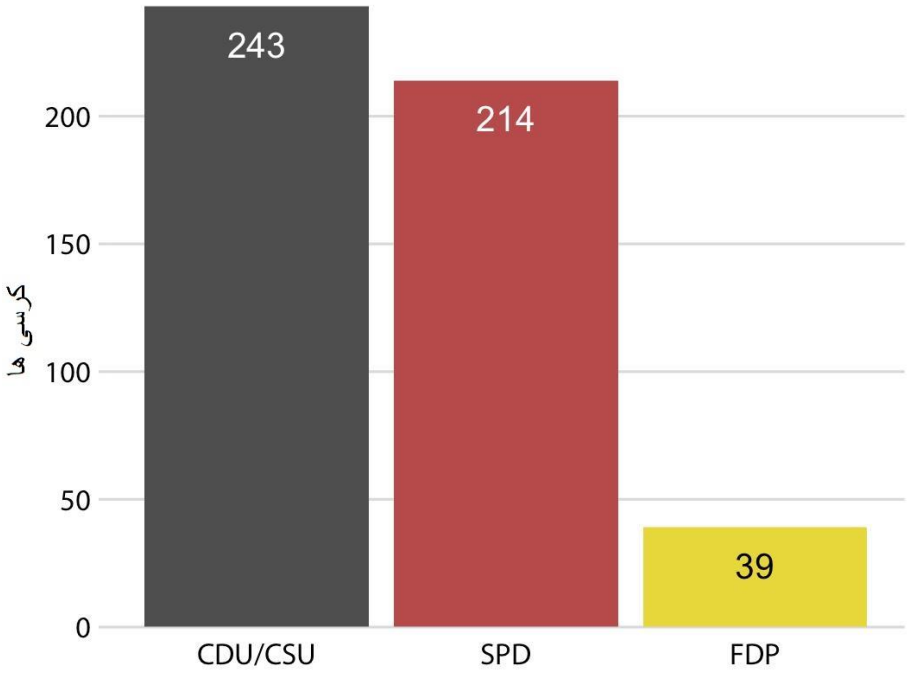
در نمودار دایره‌ای، دایره به قطاع‌هایی تقسیم می‌شود که مساحت هر کدام متناسب با نسبت آن گروه از کل می‌باشد. همین روش را می‌توان روی یک مستطیل انجام داد، و نتیجه یک نمودار میله‌ای انباشته است (نمودار ۱۰–۲). بر اساس اینکه میله به صورت عمودی یا افقی برش خورده باشد، به ترتیب نمودار میله‌ای انباشته عمودی (نمودار ۱۰–۲ الف) یا نمودار میله‌ای انباشته افقی (نمودار ۱۰–۲ ب) خواهیم داشت.



نمودار ۱۰–۲. ترکیب حزب هشتمین بوندستاگ آلمان، ۱۹۷۶–۱۹۸۰، که به صورت نمودار میله‌ای انباشته ترسیم شده است. (الف) میله‌ها به صورت عمودی روی هم چیده شده اند. (ب) میله‌ها به صورت افقی روی هم چیده شده‌اند. به سرعت نمی‌توان متوجّه شد که SPD و FDP مشترکاً کرسی‌های بیشتری نسبت به CDU/CSU داشتند.

همچنین می‌توان میله‌ها را از نمودار ۱۰–۲ الف برداشت و به جای آنکه آن‌ها را روی هم چید، در کنار هم قرار داد. این نحوه نمایش امکان مقایسهٔ مستقیم سه گروه را آسان‌تر می‌نماید، گرچه برخی از جنبه‌های دیگر داده را مخفی خواهد کرد (نمودار ۱۰–۳). مهمتر از همه، در یک نمودار که میله‌های کنار هم دارد، ارتباط هر میله با کل داده‌ها به صورت بصری مشخّص نیست.

بسیاری از نویسندگان قاطعانه با ترسیم نمودارهای دایره‌ای مخالفت می‌کنند و از نمودار میله‌ای به صورت انباشته یا میله‌های کنار هم حمایت می‌کنند. بعضی دیگر استفاده از نمودار دایره‌ای در برخی برنامه‌های کامپیوتری یا موبایلی دفاع می‌کنند. به نظر می‌رسد هیچ یک از این نمودارها برتری ذاتی بر بقیه ندارد. بلکه بر اساس ویژگی‌های مجموعهٔ داده و داستان خاصی که قصد بیان آن را دارید، می‌توانید هر کدام از دو رویکرد مذکور را انتخاب کنید. در مورد بوندستاگ هشتم آلمان، به نظر می‌رسد نمودار دایره‌ای بهترین گزینه است. این نوع نمودار نشان می‌دهد که ائتلاف حاکم SPD و FDP مشترکاً برتری مختصری نسبت به CDU/CSU داشتند (نمودار ۱۰–۱). این حقیقت از نظر بصری در هیچ یک از نمودارهای دیگر مشخّص نیست (نمودارهای ۱۰–۲ و ۱۰–۳).



نمودار ۱۰–۳. ترکیب حزب هشتمین بوندستاگ آلمان، ۱۹۷۶–۱۹۸۰، که به صورت میله‌های کنار هم ترسیم شده است. همانند نمودار ۱۰–۲، به سرعت نمی‌توان متوجّه شد که SPD و FDP مشترکاً کرسی‌های بیشتری نسبت به CDU/CSU داشتند.

به طور کلی، نمودارهای دایره‌ای زمانی ارجحیّت دارند که هدف تأکید بر نسبت‌های ساده باشد، مانند یک دوّم، یک سوّم یا یک چهارم. همچنین این نمودارها برای مجموعه داده‌های بسیار کوچک مناسب هستند. یک نمودار دایره‌ای تکی، مانند نمودار ۱۰–۱، به نظر مناسب است، امّا یک نمودار میله‌ای انباشته مانند نمودار ۱۰–۲ الف، نامناسب به نظر می‌رسند. از سوی دیگر میله‌های انباشته امکان مقایسه چند وضعیّت یا مقایسه در طول زمان را فراهم می‌کنند. میله‌های کنار هم به خصوص زمانی که هدف مقایسه مستقیم نسبت‌ها است، ارجحیّت دارند. خلاصه‌ای از مزایا و معایب نمودارهای دایره‌ای، میله‌های انباشته شده و میله‌های کنار هم در جدول ۱۰–۱ ارائه شده است.

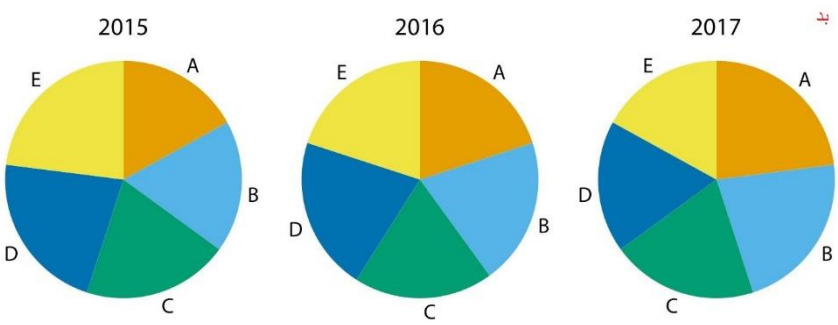
جدول ۱۰–۱. مزایا و معایب رویکردهای رایج برای نمایش نسبت ها: نمودارهای دایره ای، میله‌های انباشته و میله‌های کنار هم.

	نمودار دایره‌ای	میله‌های انباشته	میله‌های کنار هم
به وضوح داده‌ها را به صورت کسری از کل نمایش می‌دهد	✓	✓	×
امکان مقایسه بصری آسان نسبت‌ها را فراهم می‌کند	×	×	✓
به صورت بصری بر کسرهای ساده تاکید می‌کند مانند ۱/۲، ۱/۳، ۱/۴	✓	×	×
حتی برای مجموعه‌های کوچک داده نیز از نظر بصری مناسب است	✓	×	✓
وقتی مجموعه داده‌ها به قسمت‌های زیادی شکسته می‌شود، مناسب است	×	×	✓
برای نمایش مجموعه متنوعی از نسبت‌ها یا بررسی آن‌ها در طول زمان مناسب است	×	✓	×

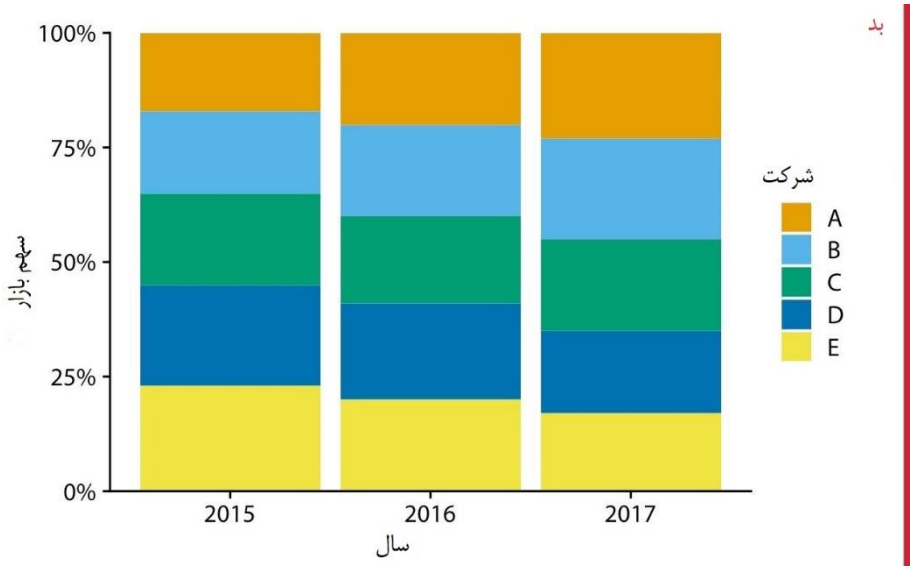
نمونه‌ای برای میله‌های کنار هم

حال بیایید مثالی را بررسی کنیم که نمودار دایره‌ای برای آن مناسب نیست. این مثال بعد از اینکه نقدهایی در خصوص نمودار دایره‌ای در ویکی پدیا ثبت شد، مطرح شده است. سناریوی فرضی را برای پنج شرکت A، B، C، D و E در نظر بگیرید که در آن همگی آن‌ها سهم تقریباً مشابه و حدوداً ۲۰ درصد از بازار را دارند. مجموعهٔ داده فرضی حاوی سهم بازار هر شرکت در سه سال متوالی است. وقتی این مجموعهٔ داده را با نمودار دایره‌ای نمایش می‌دهیم، رویت روندهای خاص در داده‌ها دشوار است (نمودار ۱۰–۴). به نظر می‌رسد سهم بازار شرکت A در حال رشد و سهم شرکت E در حال کاهش است، امّا فراتر از این تفسیر، اطّلاعات اضافه‌تری نمی‌توانیم ارائه دهیم. به طور خاص، معلوم نیست سهم بازار شرکت‌های مختلف در هر سال در مقایسه با هم به چه صورت است.

اگر از نمودار میله‌ای انباشته استفاده شود، تصویر کمی واضح‌تر خواهد شد (نمودار ۱۰–۵). اکنون روند رشد سهم بازار برای شرکت A و کاهش سهم بازار برای شرکت E به وضوح قابل مشاهده است. با این حال، همچنان مقایسهٔ سهم نسبی بازار برای شرکت‌ها در هر سال دشوار است. همچنین مقایسهٔ سهم بازار شرکت‌های B، C و D در طول سال‌ها دشوار است، زیرا میله‌ها نسبت به یکدیگر در طول سال‌ها تغییر می‌کنند. این مشکل عمده نمودارهای میله‌ای انباشته است و به همین دلیل به طور معمول این نوع نمودار توصیه نمی‌شود.

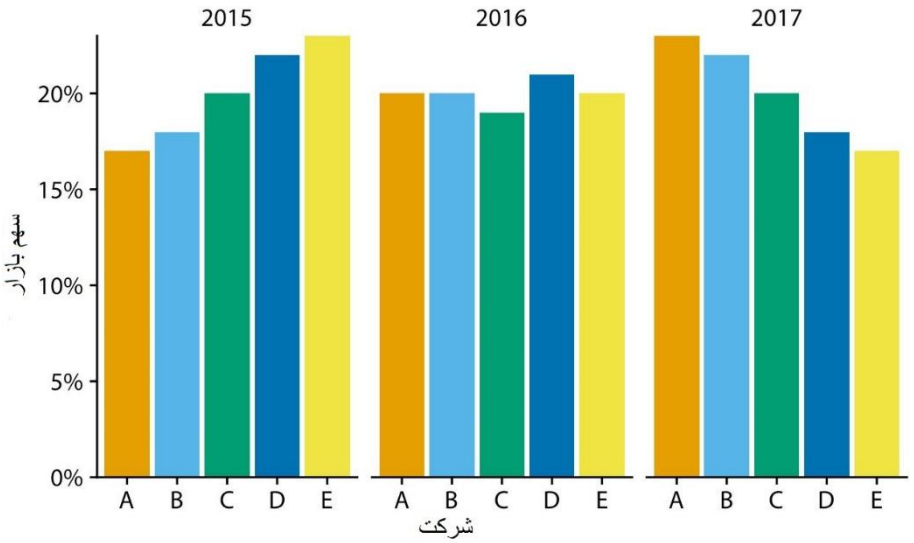


نمودار ۱۰–۴. سهم بازار پنج شرکت فرضی، A-E، در سال‌های ۲۰۱۵–۲۰۱۷، که به صورت نمودار دایره‌ای نمایش داده شده است. این نمودار دو اشکال عمده دارد: (۱) مقایسهٔ سهم نسبی بازار در شرکت‌های مختلف در هر سال تقریباً غیرممکن است، و (۲) تغییرات در سهم بازار هر شرکت در طی سال‌ها به سختی قابل مشاهده است.



نمودار ۱۰–۵. سهم بازار پنج شرکت فرضی برای سال‌های ۲۰۱۵–۲۰۱۷، که به صورت میله‌های انباشته نمایش داده شده است. این نمودار دو مشکل عمده دارد: (۱) مقایسهٔ سهم نسبی بازار برای شرکت‌های مختلف در هر سال دشوار است، و (۲) تغییرات در سهم بازار در طی سال‌های مختلف برای شرکت‌های میانی (B، C و D) دشوار است زیرا موقعیّت مکانی میله‌ها در هر سال تغییر می‌کند.

برای این مجموعه دادهٔ فرضی، میله‌های کنار هم بهترین انتخاب هستند (نمودار ۱۰–۶). این نمودار نشان می‌دهد که هر دو شرکت A و B سهم خود از بازار را طی سال های ۲۰۱۵ تا ۲۰۱۷ افزایش داده‌اند، در حالی که سهم هر دو شرکت D و E کاهش یافته است. همچنین مشاهده می‌شود که سهم بازار به ترتیب از شرکت A به E در سال ۲۰۱۵ افزایش یافته و به طور مشابه با همین ترتیب در سال ۲۰۱۷ کاهش یافته است.

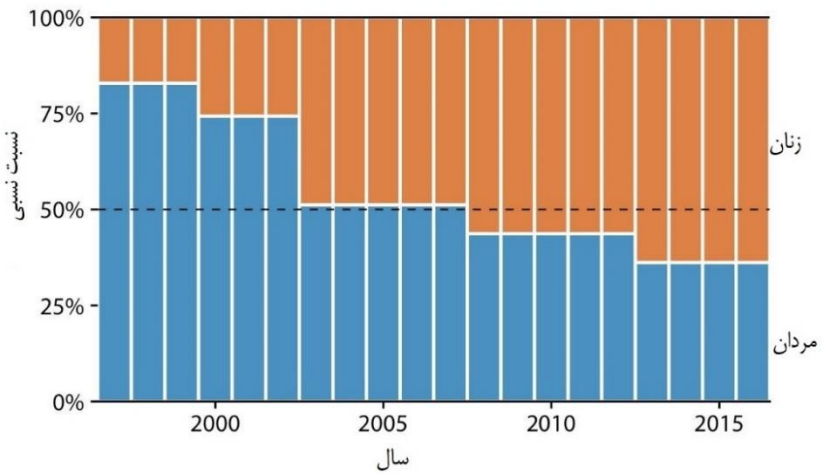


نمودار ۱۰–۶. سهم بازار پنج شرکت فرضی در سال‌های ۲۰۱۵–۲۰۱۷، که به صورت میله‌های کنار هم نمایش داده شده است.

نمونه‌ای برای میله‌های انباشته و چگالی‌های انباشته

در بخش قبل، بیان شد که معمولاً توالی چند میلهٔ انباشته توصیه نمی‌شود زیرا مکان میله‌های میانی در طول توالی تغییر می‌کند. با این حال، مشکل جابجایی میله‌های میانی در صورتی که فقط دو میله وجود داشته باشد از بین می‌رود. در این موارد نمودار حاصل می‌تواند کاملاً واضح باشد. به عنوان مثال، نسبت زنان در پارلمان ملّی یک کشور را در نظر بگیرید. به طور خاص کشور آفریقایی رواندا را بررسی خواهیم کرد که از سال ۲۰۱۶ در صدر فهرست کشورهای با بیشترین تعداد نمایندگان زن در پارلمان قرار دارند. اکثریت پارلمان در رواندا از سال ۲۰۰۸ در اختیار زنان بود و از سال ۲۰۱۳ تقریباً دو سوم اعضای پارلمان را زنان تشکیل می‌دهند. برای نمایش نحوه تغییر نسبت زنان در پارلمان رواندا در طول زمان می‌توان دنباله‌ای از میله‌ای انباشته ترسیم کرد (شکل ۱۰–۷). این نمودار یک نمایش بصری فوری از تغییر نسبت زنان

پارلمان در طول زمان ارائه می‌کند. برای کمک به خواننده برای درک اینکه دقیقاً چه زمانی اکثریّت پارلمان را زنان تشکیل می‌دادند، یک خط چین افقی در ۵۰ درصد اضافه گردید. بدون وجود این خط، تعیین اینکه از سال ۲۰۰۳ تا ۲۰۰۷ اکثریت را کدام گروه تشکیل می‌دادند، تقریباً غیرممکن خواهد بود. برای پیشگیری از شلوغ شدن نمودار، از افزودن خطوط مشابهی در ۲۵ و ۷۵ درصد پرهیز شده است.

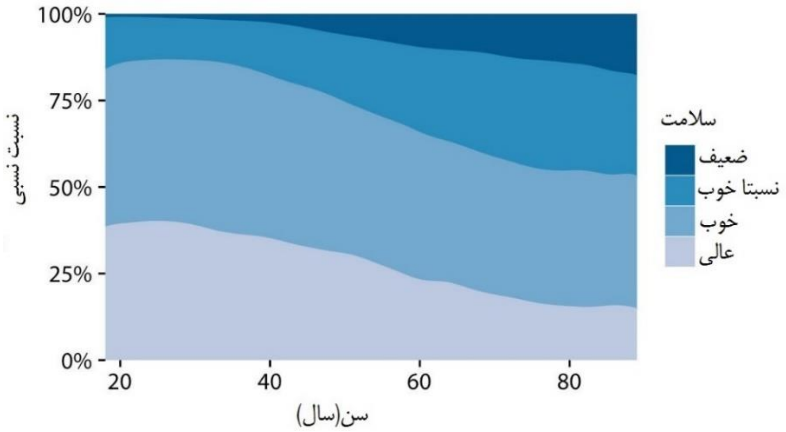


نمودار ۱۰–۷. تغییر در ترکیب جنسیتی پارلمان رواندا در طول سال‌های ۱۹۹۷ تا ۲۰۱۶. منبع داده: اتحادیه بین پارلمانی (IPU).

اگر هدف نمایش نحوه تغییر نسبت‌ها در پاسخ به یک متغیّر پیوسته باشد، می‌توان میله‌های انباشته را به چگالی‌های انباشته تبدیل نمود. چگالی‌های انباشته را می‌توان به صورت تعداد بی‌شماری از میله‌های انباشته بسیار بسیار کوچک که در کنار هم چیده شده‌اند، در نظر گرفت. چگالی‌های استفاده شده در نمودار چگالی انباشته معمولاً از کرنل برآورد چگالی به دست می‌آیند، همانگونه که در فصل ۷ توضیح داده شد. برای مطالعهٔ بحث کلی در مورد نقاط قوت و ضعف این روش به فصل ۷ رجوع نمایید.

برای ارائهٔ مثالی که در آن چگالی‌های انباشته ممکن است مناسب باشند، وضعیّت سلامت افراد را به صورت تابعی از سن در نظر بگیرید. سن را می‌توان یک متغیّر پیوسته در نظر گرفت، و نمایش داده‌ها با این روش از نظر منطقی صحیح است (نمودار ۱۰–۸). با اینکه در مثال چهار گروه سلامتی داریم و همانطور که قبلاً اشاره شد انباشتن چند گروه نیز معمولاً توصیه نمی‌شود امّا برای این مثال، نمودار قابل قبول می‌باشد. همانطور که این نمودار نشان

می‌دهد با افزایش سن وضعیّت کلی سلامتی افت می‌کند، از سوی دیگر می‌توان مشاهده نمود که علیرغم وجود روند مذکور بیش از نیمی از افراد حتّی تا سن‌های بسیار بالا همچنان از وضعیّت سلامتی خوب یا عالی برخوردارند.



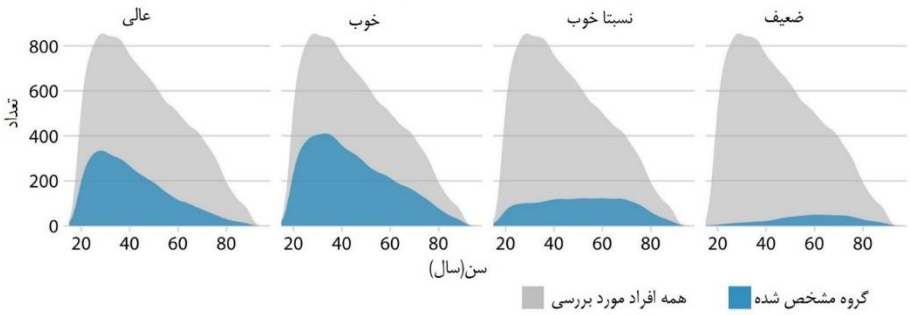
نمودار ۱۰–۸. وضعیّت سلامت بر اساس سن منبع داده‌ها: نظرسنجی عمومی اجتماعی (GSS)

با این وجود، این نمودار یک محدودیت عمده دارد: با نمایش نسبت‌های چهار وضعیّت سلامتی به صورت درصدی از کل، این مساله مغفول می‌ماند که تعداد افراد جوان در مجموعهٔ داده بسیار بیشتر از افراد مسن است. بنابراین، اگرچه درصد افرادی که وضعیّت سلامتی خوب دارند در طی هفت دهه تقریباً بدون تغییر باقی مانده است، تعداد مطلق افرادی که وضعیّت سلامتی خوب دارند همسو با کاهش تعداد کل افراد در یک سن معین کاهش می‌یابد. برای حل این مشکل در بخش بعدی راه حلی ارائه خواهد شد.

نمایش جداگانه نسبت‌ها به صورت بخشی از کل

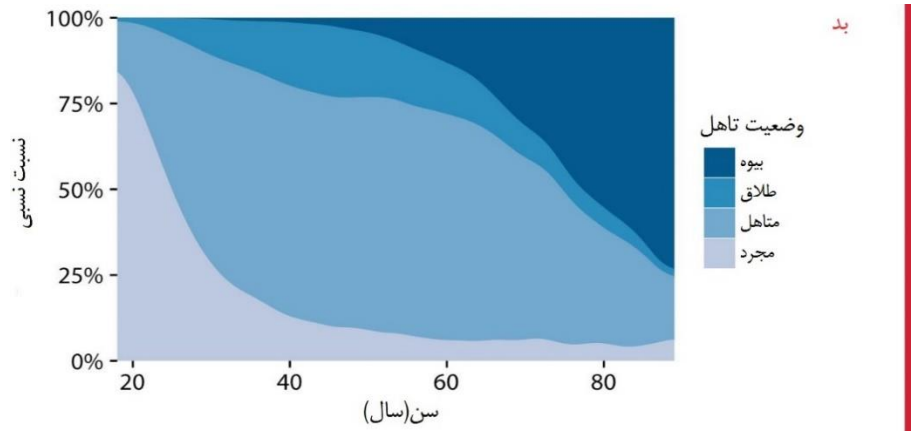
نقطه ضعف میله‌های کنار هم این است که اندازهٔ هر قسمت را نسبت به کل نشان نمی‌دهند و نقطه ضعف میله‌های انباشته این است که به دلیل خطوط پایه متفاوت، نمی‌توان آن‌ها را به راحتی با هم مقایسه کرد. این دو مساله را می‌توان به این صورت حل کرد که برای هر قسمت یک نمودار جداگانه رسم نمود که در آن وضعیّت آن قسمت نسبت به کل نمایش داده شود. انجام این کار برای مجموعه داده‌های سلامت مربوط به نمودار ۱۰–۸، منجر به رسم نمودار ۱۰–۹ می‌شود. توزیع کلی سن در مجموعه داده به صورت مناطق سایه‌دار خاکستری نشان داده شده است، و توزیع سنی برای هر وضعیّت سلامتی به رنگ آبی نشان داده شده

است. این نمودار نشان می‌دهد که به صورت مطلق، تعداد افراد دارای وضعیّت سلامتی عالی یا خوب بعد از سنین ۳۰ تا ۴۰ سالگی کاهش می‌یابد، در حالی که تعداد افراد با سلامت متوسط تقریباً در تمام سنین ثابت می‌ماند.



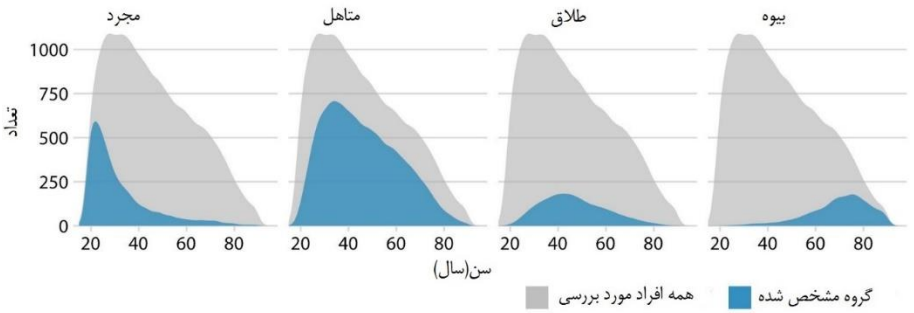
نمودار ۱۰–۹. وضعیّت سلامت بر اساس سن، که به صورت نسبت تعداد کل افراد شرکت کننده در مطالعه نشان داده شده است. نواحی رنگی تخمین تراکمی از سن افراد با وضعیّت سلامت مربوطه و مناطق خاکستری توزیع کلی سن را نشان می‌دهد. منبع داده ها: GSS

به عنوان مثالی دیگر، بیایید متغیّر متفاوتی از همان نظرسنجی را در نظر بگیریم: وضعیّت تاهل. وضعیّت تاهل نسبت به وضعیّت سلامتی با افزایش سن با شدت بیشتری تغییر می‌کند، و نمودار چگالی انباشته وضعیّت تاهل در مقابل سن چندان کاربردی نیست (نمودار ۱۰–۱۰).



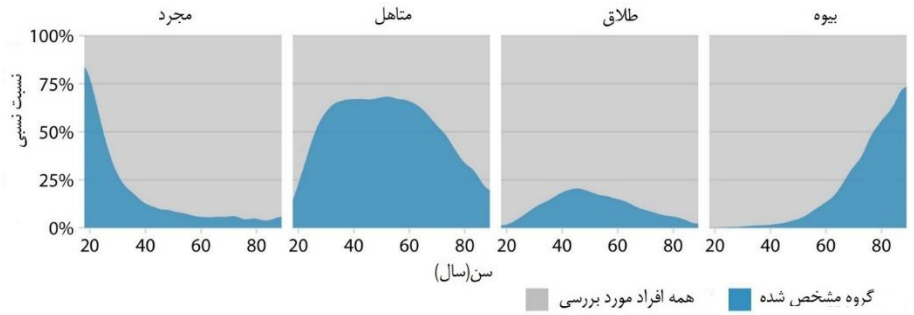
نمودار ۱۰–۱۰. وضعیّت تاهل بر اساس سن. برای ساده کردن نمودار، تعداد کمی از افراد که «زندگی مجزا» داشتند حذف شده است. این نمودار برچسب «بد» خورده است زیرا فراوانی افرادی که هرگز ازدواج نکرده‌اند یا بیوه هستند با افزایش سن به شدّت تغییر می‌کند. از سوی دیگر توزیع سنی افراد متاهل و مطلقه به شدّت مخدوش بوده و تفسیر آن دشوار است. منبع داده: GSS

همان مجموعه داده اگر به صورت نمودار تراکم جزئی ترسیم شود بسیار واضح‌تر خواهد بود (نمودار ۱۰–۱۱). به طور مشخص، قابل مشاهده است که نسبت افراد متاهل در اواخر دهه ۳۰، نسبت افراد مطلقه در اوایل دهه ۴۰ و نسبت بیوه‌ها در اواسط دهه ۷۰ به اوج خود می‌رسد.



شکل ۱۰–۱۱. وضعیّت تاهل بر اساس سن، که به صورت نسبتی از کل شرکت کنندگان در مطالعه نشان داده شده است. نواحی رنگی تخمین تراکمی سن افراد با وضعیّت تاهل مربوطه و مناطق خاکستری توزیع کلی سنی را نشان می‌دهد.

با این حال، یک نقطه ضعف نمودار ۱۰–۱۱ این است که این نمودار امکان مقایسه نسبت‌ها را در یک بازهٔ زمانی مشخّص فراهم نمی‌کند. به عنوان مثال، اگر هدف دانستن این مساله بود که در چه سنی بیش از ۵۰ درصد افراد مورد بررسی ازدواج کرده‌اند، پاسخ آن را نمی‌توان به راحتی از نمودار ۱۰–۱۱ استخراج کرد. برای پاسخ به این سوال می‌توان از همان نمودار قبلی استفاده کرد با این تفاوت که در محور عمودی به جای فراوانی مطلق از فراوانی نسبی استفاده نمود (نمودار ۱۰–۱۲). اکنون می‌توان دید که از اواخر دههٔ ۲۰ افراد متاهل و از اواسط دههٔ ۷۰ افراد بیوه در اکثریت هستند.



نمودار ۱۰–۱۲. وضعیّت تاهل بر اساس سن، که به صورت نسبت تعداد کل افراد شرکت کننده در مطالعه نشان داده شده است. نواحی آبی رنگ درصد افراد در سن معین با وضعیّت تاهل مربوطه را نشان می‌دهد، و نواحی خاکستری رنگ درصد افراد با سایر وضعیّت‌های تاهل را نشان می‌دهد منبع داده: GSS

فصل ۱۱

ترسیم نسبت‌های لانه‌گزیده

در فصل قبل، حالت‌هایی را مورد بحث قرار دادیم که در آن یک مجموعه داده بر اساس یک متغیّر گروه‌بندی شده، مانند حزب سیاسی، شرکت یا وضعیّت سلامتی به دسته‌هایی تقسیم می‌شود. با این حال، ممکن است که بخواهیم عمیق‌تر بررسی کرده و یک مجموعه داده را با چندین متغیّر به طور همزمان گروه‌بندی کنیم. به عنوان مثال، در مورد کرسی‌های پارلمان، می‌توان نسبت کرسی‌ها را بر اساس حزب و جنسیّت نمایندگان گروه‌بندی نمود. به طور مشابه، در مورد وضعیّت سلامت افراد، می‌توانیم بپرسیم که چگونه وضعیّت سلامت بر اساس وضعیّت تاهل گروه‌بندی می‌شود. به این سناریوها به عنوان نسبت‌های لانه‌گزیده اطلاق می‌شود، زیرا هر متغیّری که افزوده می‌شود، زیرمجموعهٔ کوچکتری از داده‌ها که در نسبت قبلی جای دارد را نشان می‌دهد. چندین رویکرد مناسب برای ترسیم چنین نسبت‌های لانه‌گزیده‌ای وجود دارد، از جمله طرح‌های موزاییکی، نقشه‌های درختی و مجموعه‌های موازی.

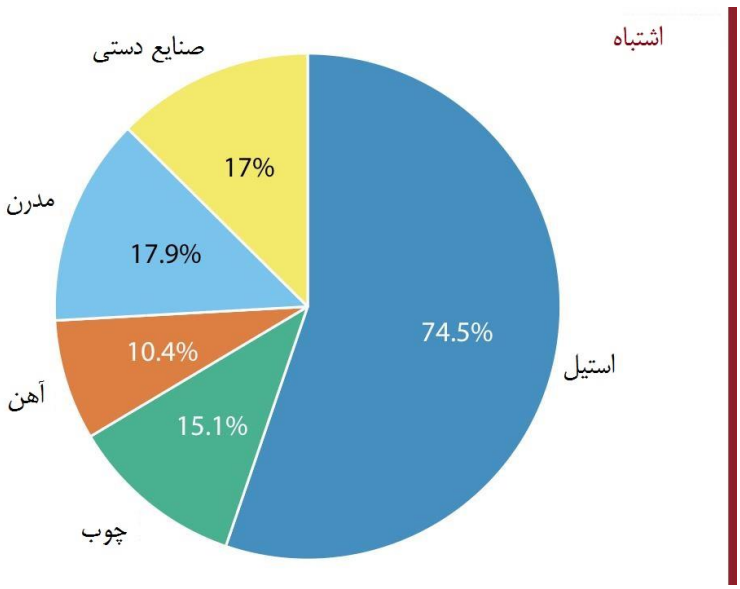
نسبت‌های لانه گزیده اشتباه

در ابتدا دو رویکرد اشتباه برای ترسیم نسبت‌های لانه‌گزیده را نشان می‌دهیم. در حالی که این رویکردها ممکن است برای متخصصین علوم داده بی معنی به نظر برسد، ما با آن‌ها برخورد داشته‌ایم و لذا نیاز به بحث دارند. در طول این فصل، با مجموعه داده‌ای از ۱۰۶ پل در پیتسبورگ کار خواهیم کرد. این مجموعه داده اطّلاعات مختلفی در مورد پل‌ها دارد، مانند مواد سازنده پل (استیل، آهن یا چوب) و سال ساخت. بر اساس سال ساخت، پل‌ها به دسته‌های

مجزا تقسیم شده‌اند، مانند پل‌های سنّتی که قبل از سال ۱۸۷۰ ساخته شده‌اند و پل‌های مدرن که پس از سال ۱۹۴۰ ساخته شده‌اند.

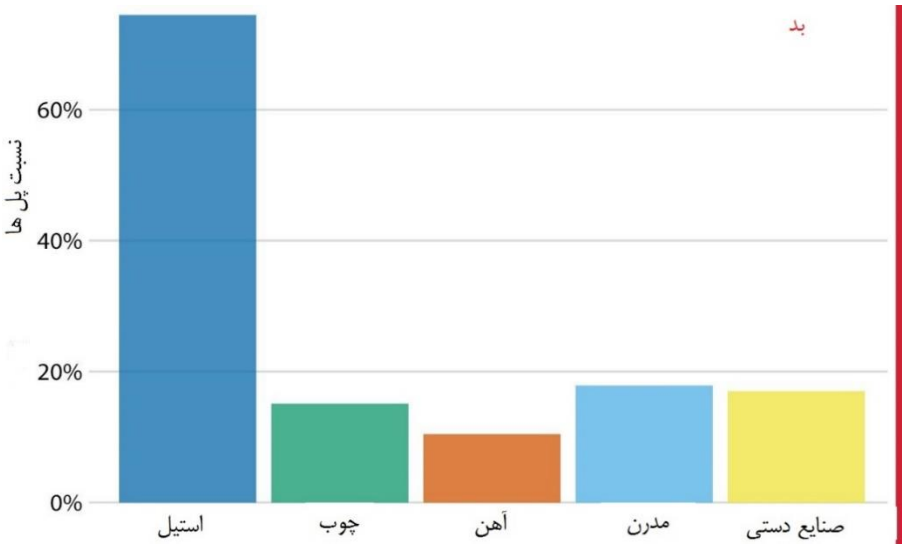
بیایید فرض کنیم که می‌خواهیم هم نسبتی از پل‌های ساخته شده از استیل، آهن یا چوب را ترسیم کنیم و هم نسبتی را که سنّتی یا مدرن هستند. ممکن است رسم نمودار دایره‌ای ترکیبی ما را وسوسه کند (شکل ۱۱–۱). با این حال، این نمودار معتبر نیست. حاصل جمع تمام برش‌های یک نمودار دایره‌ای باید ۱۰۰ درصد باشد امّا در اینجا حاصل جمع برش‌ها ۱۳۵ درصد است.

مجموع قطاع‌ها بیش از ۱۰۰ درصد می‌شود، زیرا دو بار پل‌ها را می‌شماریم. هر پل در مجموعه داده‌ها از استیل، آهن یا چوب ساخته شده است، بنابراین این سه برش در حال حاضر ۱۰۰ درصد از پل‌ها را نشان می‌دهد. هر پل سنّتی یا مدرن نیز یک پل استیلی، آهنی یا چوبی است و از این رو در نمودار دایره‌ای دو بار محاسبه می‌شود.



شکل ۱۱–۱. دسته‌بندی پل‌ها در پیتسبورگ بر اساس مصالح ساختمانی (استیل، چوب، آهن) و تاریخ ساخت (سنتی، قبل از ۱۸۷۰، و مدرن، پس از ۱۹۴۰)، که به صورت نمودار دایره‌ای نشان داده شده است. اعداد نشان دهندهٔ درصد پل‌های یک گروه معیّن در بین همهٔ پل‌ها هستند. این نمودار نامعتبر است، زیرا مجموع درصدها بیش از ۱۰۰ درصد است. بین مصالح ساختمانی و تاریخ ساخت همپوشانی وجود دارد. به عنوان مثال، تمام پل‌های مدرن از استیل و اکثر پل‌های سنّتی از چوب ساخته شده‌اند. منبع داده: Yoram Reich و Steven J. Fenves، از طریق [Dua and Karra Taniskidou 2017] UCI Machine Learning Repository.

اگر نموداری را انتخاب کنیم که الزامی برای اینکه که مجموع نسبت‌ها ۱۰۰ درصد شود، وجود نداشته باشد، شمارش دوباره لزوماً مشکل‌ساز نخواهد بود. همانطور که در فصل قبل بحث شد، میله‌های کنار هم این معیار را برآورده می‌کنند. می‌توانیم نسبت‌های مختلف پل‌ها را به صورت نمودار میله‌ای نشان دهیم، و این نمودار از نظر فنّی اشتباه نیست (شکل ۱۱–۲). با این وجود، این نمودار به عنوان «بد» برچسب خورده است زیرا نشان نمی‌دهد که بین برخی از گروه‌بندی‌های نشان داده شده همپوشانی وجود دارد. یک خوانندهٔ معمولی ممکن است از شکل ۱۱–۲ نتیجه بگیرد که پنج دسته مجزا از پل‌ها وجود دارد و برای مثال، پل‌های مدرن نه از استیل ساخته شده‌اند و نه از چوب یا آهن.



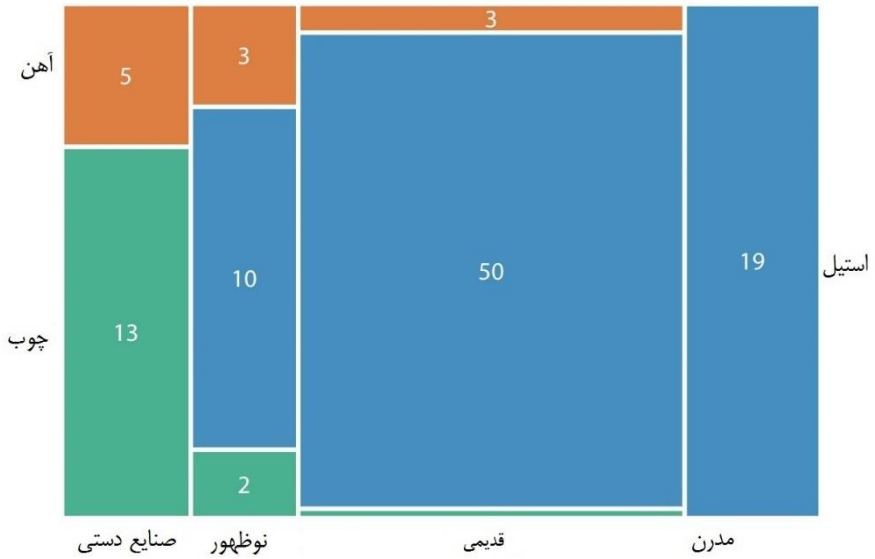
شکل ۱۱–۲. دسته‌بندی پل‌ها در پیتسبورگ بر اساس مصالح ساختمانی (استیل، چوب، آهن) و بر اساس تاریخ ساخت (سنتی، قبل از سال ۱۸۷۰، و مدرن، پس از ۱۹۴۰)، که به صورت نمودار میله‌ای نشان داده شده است. برخلاف شکل ۱۱–۱، این ترسیم از نظر فنّی اشتباه نیست، زیرا به این معنی نیست که مجموع ارتفاع میله باید ۱۰۰ درصد شود. با این حال، به وضوح نشان‌دهندهٔ همپوشانی بین گروه‌های مختلف هم نیست، و بنابراین به عنوان «بد» برچسب خورده است. منبع داده: Yoram Reich و Steven J. Fenves

نمودارهای موزاییکی[1] و نقشه‌های درختی[2]

هر زمان دسته‌هایی داریم که همپوشانی دارند، بهتر است به صراحت نحوه ارتباط آن‌ها با یکدیگر را نشان دهیم. این کار را می‌توان با نمودار موزاییکی انجام داد (شکل ۱۱–۳). در نگاه

1. Mosaic Plots
2. Treemaps

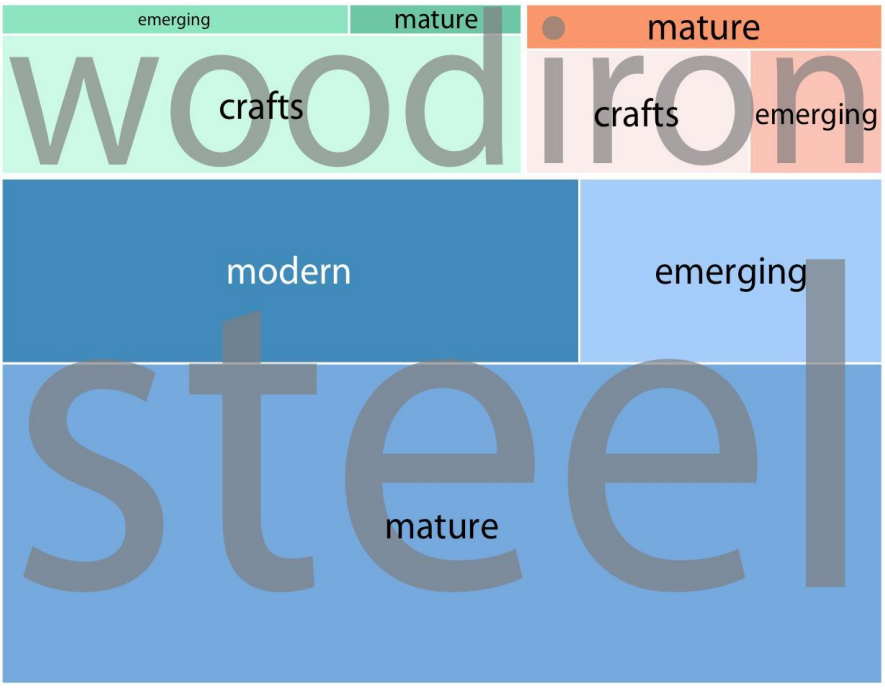
اول، نمودار موزاییکی شبیه نمودار میله‌ای انباشته به نظر می‌رسد (مانند شکل ۱۰–۵). با این حال، در نمودار موزاییکی بر خلاف نمودار میله‌ای انباشته، هم ارتفاع و هم عرض مناطق رنگی متغیّر است. توجّه داشته باشید که در شکل ۱۱–۳، دو دورهٔ ساخت و ساز اضافی، در حال ظهور (از ۱۸۷۰ تا ۱۸۸۹) و بلوغ (۱۸۹۰ تا ۱۹۳۹) را می‌بینیم. در ترکیب با سنّتی و مدرن، این دوره‌های ساخت و ساز تمام پل‌های مجموعهٔ داده را پوشش می‌دهند، همانند سه نوع مصالح ساختمانی. این یک شرط حیاتی برای نمودار موزاییکی است: هر متغیّر کیفی باید تمام مشاهدات موجود در مجموعه داده را پوشش دهد.



شکل ۱۱–۳. دسته‌بندی پل‌ها در پیتسبورگ بر اساس مصالح ساختمانی (استیل، چوب، آهن) و دورهٔ ساخت و ساز (سنتی، نوظهور، بلوغ، مدرن)، که به صورت نمودار موزاییکی نشان داده شده است. عرض هر مستطیل متناسب با تعداد پل‌های ساخته شده در آن دورهٔ زمانی و ارتفاع آن متناسب با تعداد پل‌های ساخته شده از آن ماده است. اعداد نشان‌دهندهٔ تعداد پل‌ها در هر دسته هستند. منبع داده: Yoram Reich و Steven J. Fenves

برای ترسیم یک نمودار موزاییکی، یک متغیّر گروه‌بندی را در امتداد محور x (در اینجا، دوران ساخت پل) قرار داده و محور x را به نسبتی که دسته‌ها را تشکیل می‌دهند، تقسیم می‌کنیم. سپس متغیّر دیگر را در امتداد محور y قرار می‌دهیم (در اینجا، مصالح ساختمانی) و در هر دسته در امتداد محور x، محور y را به نسبتی که دسته‌های متغیّر y را تشکیل می‌دهند، تقسیم می‌کنیم. نتیجه حاصل مجموعه‌ای از مستطیل‌ها است که مساحت آن‌ها متناسب با تعداد مواردی است که تمام ترکیب‌های ممکن از دو متغیّر گروه‌بندی شده را نشان می‌دهد.

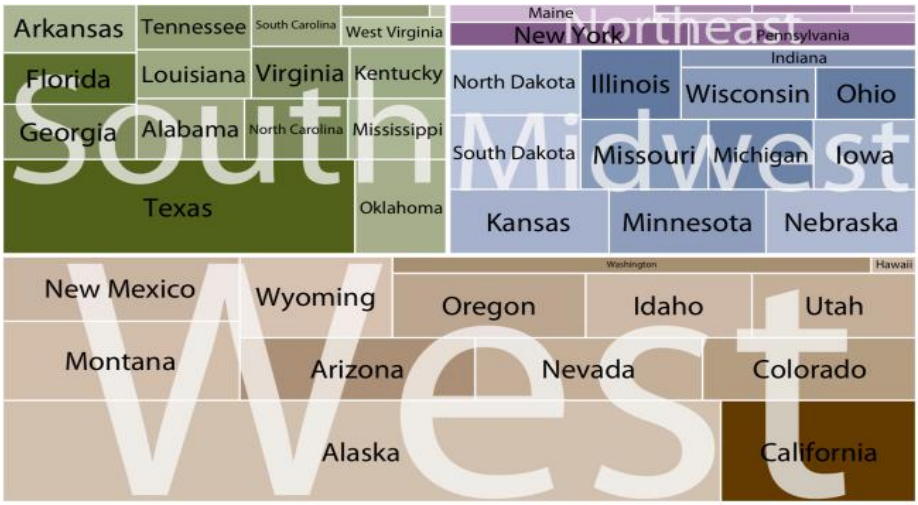
مجموعه داده مربوط به پل‌ها را می‌توان در قالبی مشابه امّا متمایز به نام نقشهٔ درختی ترسیم نمود. در نقشهٔ درختی، درست مانند نمودار موزاییکی، یک مستطیل محصور را در نظر می‌گیریم و آن را به مستطیل‌های کوچکتر تقسیم می‌کنیم که مساحت آن‌ها نسبت‌ها را نشان می‌دهد. با این حال، روش قرار دادن مستطیل‌های کوچک‌تر در مستطیل‌های بزرگتر در مقایسه با نمودار موزاییکی متفاوت است. در نقشهٔ درختی، مستطیل‌ها را به صورت لانه‌گزیده درون هم قرار می‌دهیم. به عنوان مثال، در مورد پل‌های پیتسبورگ، ابتدا می‌توانیم کل مساحت را به سه قسمت تقسیم کنیم که نمایانگر سه نوع مصالح ساختمانی، چوب، آهن و استیل است. سپس می‌توانیم هر یک از آن مناطق را بیشتر تقسیم کنیم تا دوره‌های ساخت و ساز مرتبط برای هر نوع از مصالح ساختمانی را نشان دهیم (شکل ۱۱-۴). در اصل می‌توان این کار را همچنان ادامه داد و زیرمجموعه‌های کوچک‌تری را درون یکدیگر قرار داد، امّا نتیجه حاصل به سرعت گیج‌کننده می‌شود.



شکل ۱۱-۴. دسته‌بندی پل‌ها در پیتسبورگ بر اساس مصالح ساختمانی (استیل، چوب، آهن) و دورهٔ ساخت و ساز (سنتی، نوظهور، بلوغ، مدرن)، که به صورت نقشهٔ درختی نشان داده شده است. مساحت هر مستطیل متناسب با تعداد پل‌های آن نوع است. منبع داده: Yoram Reich و Steven J. Fenves. (با هدف نمایش ظاهر اصلی نمودار، کلمات ترجمه نشده است، مترجم)

در حالی که نمودارهای موزاییکی و نقشه‌های درختی ارتباط نزدیکی با هم دارند، امّا بر نکات مختلفی تاکید دارند و حوزه‌های کاربردشان متفاوت است. در اینجا، نمودار موزاییکی (شکل ۱۱-۳) بر تکامل زمانی در استفاده از مصالح ساختمانی از دوران سنّتی تا عصر مدرن تأکید دارد، در حالی که نقشهٔ درختی (شکل ۱۱-۴) بر تعداد کل پل‌های استیلی، آهنی و چوبی تأکید دارد.

به طور کلی‌تر، پیش فرض در نمودارهای موزاییکی این است که تمام نسبت‌های نشان داده شده را می‌توان از طریق ترکیب دو یا چند متغیّر شناسایی کرد. به عنوان مثال، در شکل ۱۱-۳، هر پل را می‌توان با انتخاب مصالح ساختمانی (چوب، آهن، استیل) و دورهٔ زمانی (سنّتی، در حال ظهور، بلوغ، مدرن) توصیف کرد. علاوه بر این، هر ترکیبی از این دو متغیّر ممکن است، هرچند در عمل از آن استفاده نمی‌شود (در اینجا، هیچ پل سنّتی استیلی و هیچ پل مدرن چوبی یا آهنی وجود ندارد). در مقابل، چنین نیازی برای نقشه‌های درختی وجود ندارد. در واقع، نقشه‌های درختی زمانی کارآمد هستند که نتوان نسبت‌ها را با ترکیب چندین متغیّر به طور معناداری توصیف کرد. به عنوان مثال، می‌توانیم ایالات متحده را به چهار منطقه (غرب، شمال شرق، غرب میانه و جنوب) و هر منطقه را به ایالت‌های مجزا تقسیم کنیم، امّا ایالت‌های یک منطقه هیچ ارتباطی با ایالت‌های منطقه دیگر ندارند (شکل ۱۱-۵).



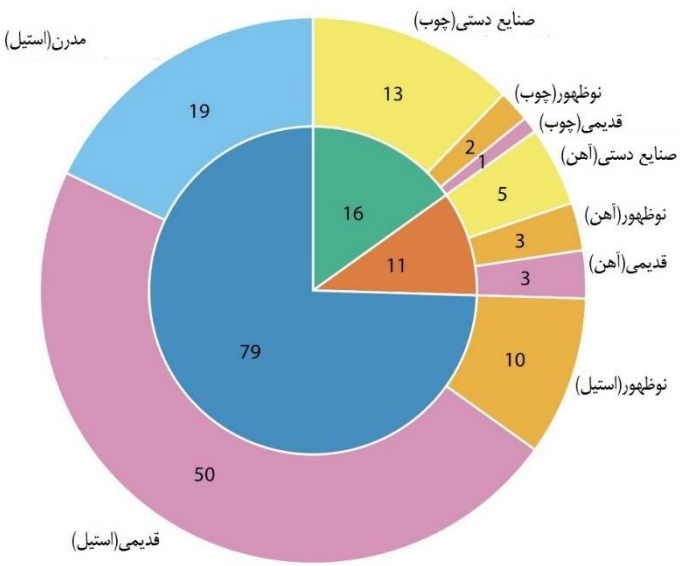
شکل ۱۱-۵. ایالت‌های مختلف در ایالات متحده آمریکا که به صورت نقشهٔ درختی ترسیم شده است. هر مستطیل نشان‌دهندهٔ یک ایالت است و مساحت هر مستطیل متناسب با مساحت آن ایالت است. ایالت‌ها به چهار منطقه غربی، شمال شرقی، غرب میانه و جنوب گروه‌بندی شده‌اند. رنگ‌آمیزی متناسب با تعداد ساکنان هر ایالت است، رنگ‌های تیره‌تر تعداد بیشتری از ساکنان را نشان می‌دهد. منبع داده: سرشماری ده ساله ایالات متحده در سال ۲۰۱۰. (با هدف نمایش ظاهر اصلی نمودار، کلمات ترجمه نشده است، مترجم)

هم نمودارهای موزاییکی و هم نقشه‌های درختی پر کاربرد بوده و اثربخش هستند، امّا محدودیت‌های مشابهی با نمودار میله‌ای انباشته دارند (جدول ۱۰–۱). یعنی، مقایسهٔ مستقیم بین گروه‌ها می‌تواند دشوار باشد، زیرا مستطیل‌های مختلف لزوماً خطوط پایهٔ مشترکی ندارند تا امکان مقایسهٔ بصری فراهم شود. در نمودارهای موزاییکی یا نقشه‌های درختی، این مشکل با این واقعیّت تشدید می‌شود که شکل مستطیل‌ها می‌تواند متغیّر باشد. به عنوان مثال، در دوره‌های نوظهور و بلوغ تعداد یکسانی از پل‌های آهنی (سه پل) وجود دارد، امّا تشخیص این موضوع در طرح موزاییکی دشوار است (شکل ۱۱–۳)، زیرا دو مستطیلی که معرّف این دو گروه می‌باشند اشکال کاملا متفاوتی دارد. لزوماً راه حلی برای این مشکل وجود ندارد؛ ترسیم نسبت‌های لانه‌گزیده می‌تواند گول‌زننده باشد. در صورت امکان، بهتر است تعداد یا درصدهای واقعی را در نمودار نشان دهید، تا خوانندگان بتوانند تأیید کنند که تفسیر بصری آن‌ها از تصویر، درست است.

نمودار دایره‌ای لانه‌گزیده

در ابتدای این فصل، مجموعهٔ داده پل‌ها با یک نمودار دایره‌ای ناکارامد ترسیم شد (شکل ۱۱–۱)، سپس استدلال کردیم که نمودار موزاییکی یا نقشهٔ درختی مناسب‌تر است. با این حال، هر دو نوع نمودار اخیر ارتباط نزدیکی با نمودارهای دایره‌ای دارند، زیرا همه آن‌ها از مساحت برای نمایش مقادیر داده استفاده می‌کنند. تفاوت اصلی در نوع سیستم مختصّات است: مختصّات قطبی در مورد نمودار دایره‌ای در مقابل مختصّات دکارتی در مورد نمودار موزاییکی یا نقشهٔ درختی. این رابطه نزدیک بین این نمودارهای مختلف این سوال را مطرح می‌کند که آیا می‌توان از گونه‌ای از نمودار دایره‌ای برای ترسیم این مجموعه داده استفاده کرد یا خیر.

دو احتمال وجود دارد. ابتدا می‌توانیم نمودار دایره‌ای متشکل از یک دایرهٔ درونی و بیرونی رسم کنیم (شکل ۱۱–۶). دایرهٔ داخلی تقسیم داده‌ها را بر اساس یک متغیّر نشان می‌دهد (در اینجا، مصالح ساختمانی)، و دایره بیرونی تقسیم هر برش از دایره داخلی را بر اساس متغیّر دوّم نشان می‌دهد (در اینجا، دوران ساخت پل). این ترسیم معقول است، امّا با برچسب «زشت» برچسب‌گذاری شده است. مهمترین نکته این است که دو دایرهٔ مجزا این واقعیّت را می‌پوشاند که هر پل در مجموعهٔ داده هم دارای مصالح ساختمانی و هم دورهٔ ساخت است. در واقع، در شکل ۱۱–۶، ما همچنان هر پل را دوبار شمارش می‌کنیم. اگر همهٔ اعداد نشان داده شده در دو دایره را جمع کنیم، مقدار ۲۱۲ به دست می‌آید که دو برابر تعداد پل‌های موجود در مجموعه داده است.

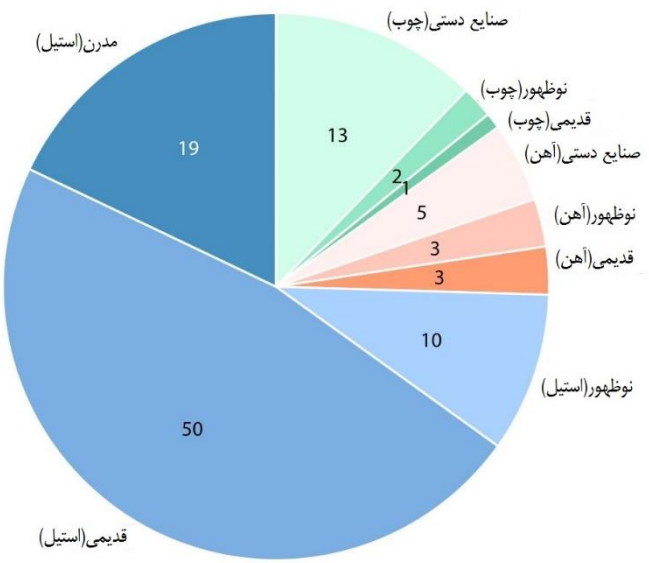


شکل ۱۱–۶. دسته‌بندی پل‌ها در پیتسبورگ بر اساس مصالح ساختمانی (استیل، چوب، آهن؛ دایرهٔ داخلی) و دورهٔ ساخت و ساز (سنتی، در حال ظهور، بلوغ، مدرن؛ دایره بیرونی). اعداد نشان‌دهندهٔ تعداد پل‌ها در هر دسته هستند. منبع داده: Steven J. Fenves و Yoram Reich

از طرف دیگر، می‌توانیم ابتدا نمودار دایره‌ای را به قطعاتی برش دهیم که نشان‌دهندهٔ نسبت‌ها منطبق بر یک متغیّر (مثلاً مصالح ساختمانی) است و سپس این برش‌ها را بر اساس متغیّر دیگر (دوره ساخت و ساز) تقسیم کنیم (شکل ۱۱–۷). به این ترتیب، در واقع یک نمودار دایره‌ای معمولی با تعداد زیادی برش دایره‌ای کوچک می‌سازیم. با این حال، می‌توانیم از رنگ‌آمیزی برای نشان دادن ماهیت لانه‌گزیدهٔ دایره استفاده کنیم. در شکل ۱۱–۷ رنگ سبز نمایانگر پل‌های چوبی، رنگ نارنجی نمایانگر پل‌های آهنی و رنگ آبی نمایانگر پل‌های استیلی هستند. تیرگی هر رنگ نشان‌دهندهٔ دوران ساخت و ساز است، به طوری که رنگ‌های تیره‌تر مربوط به پل‌های اخیراً ساخته شده است. با استفاده از مقیاس رنگ لانه‌گزیده در این روش، می‌توانیم تقسیم داده‌ها را هم توسط متغیّر اولیّه (مصالح ساختمان) و هم توسط متغیّر ثانویّه (دوره ساخت) ترسیم کنیم.

نمودار دایره‌ای شکل ۱۱–۷ نمایش معقولی از مجموعهٔ داده پل‌ها را نشان می‌دهد، امّا در مقایسهٔ مستقیم با نقشهٔ درختی معادل آن (شکل ۱۱–۴) به نظر می‌رسد نقشهٔ درختی به دو

دلیل ارجحیّت دارد. اول اینکه شکل مستطیلی نقشهٔ درختی امکان استفاده بهینه از فضای موجود را فراهم می‌کند. شکل‌های ۱۱–۴ و ۱۱–۷ دقیقاً اندازه یکسانی دارند، امّا در شکل ۱۱–۷ بسیاری از فضای شکل به عنوان فضای خالی هدر می‌رود. شکل ۱۱–۴، نقشهٔ درختی، عملاً فضای خالی اضافی ندارد. این مساله مهم است زیرا امکان قرار دادن برچسب‌ها را در قسمت‌های سایه‌دار در نقشهٔ درختی ممکن می‌کند. برچسب‌های داخلی همیشه یک واحد بصری قوی‌تری با داده‌ها نسبت به برچسب‌های خارجی ایجاد می‌کنند و بنابراین ارجح هستند. دوّم، برخی از برش‌های دایره در شکل ۱۱–۷ بسیار نازک هستند و بنابراین به سختی دیده می‌شوند. در مقابل، هر مستطیل در شکل ۱۱–۴ اندازه معقولی دارد.

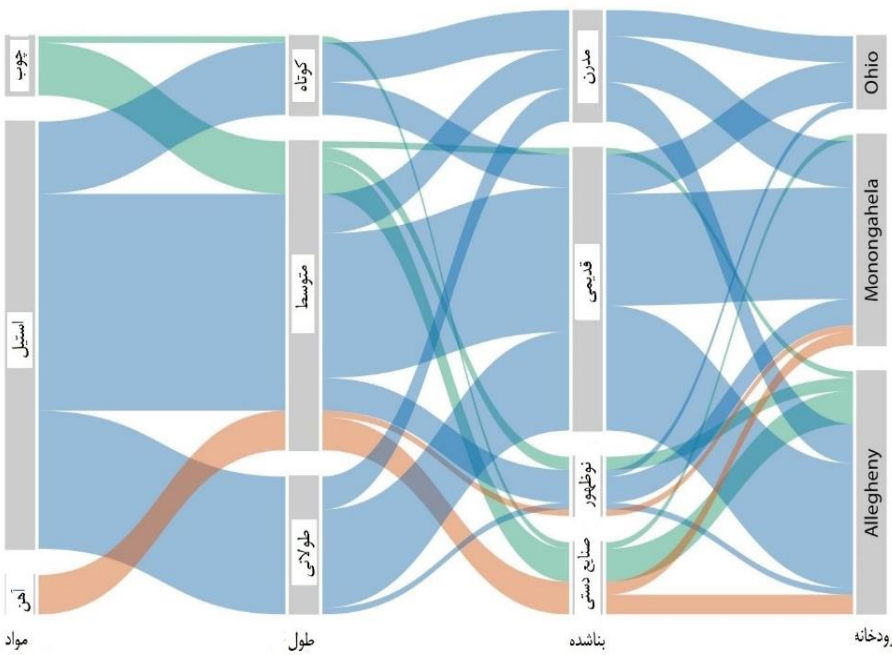


شکل ۱۱–۷. دسته‌بندی پل‌ها در پیتسبورگ بر اساس مصالح ساختمانی (استیل، چوب، آهن) و دورهٔ ساخت و ساز (سنتی، در حال ظهور، بلوغ، مدرن). اعداد نشان‌دهندهٔ تعداد پل‌ها در هر دسته هستند. منبع داده: Yoram Reich و Steven J. Fenves

مجموعه‌های موازی

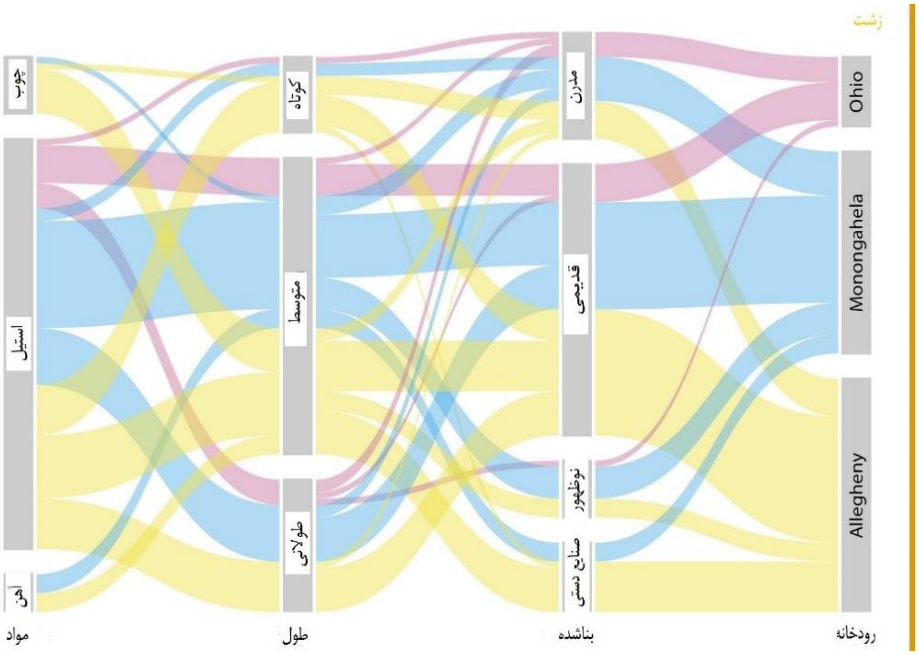
وقتی می‌خواهیم نسبت‌هایی را که توسّط بیش از دو متغیّر گروه‌بندی توصیف می‌شوند ترسیم کنیم، نمودارهای موزاییکی، نقشه‌های درختی و نمودارهای دایره‌ای همگی ناکارآمد می‌شوند. یک جایگزین مناسب در این مورد می‌تواند نمودار مجموعه‌های موازی باشد. در نمودار مجموعه‌های موازی، نشان می‌دهیم که چگونه کل مجموعهٔ داده‌ها بر اساس هر متغیّر

گروه‌بندی تقسیم می‌شود، و سپس نوارهای سایه‌دار را ترسیم می‌کنیم که نشان می‌دهد چگونه زیرگروه‌ها با یکدیگر ارتباط دارند. برای مثال شکل ۱۱–۸ را ببینید. در این شکل، مجموعهٔ داده پل‌ها را بر اساس مصالح ساختمانی (آهن، استیل، چوب)، طول هر پل (بلند، متوسط، کوتاه)، دوره‌ای که طی آن هر پل ساخته شده است (سنتی، نوظهور، بلوغ، مدرن)، و رودخانه‌ای که هر پل بر روی آن قرار دارد (Allegheny، Monongahela، Ohio) تقسیم شده است. رنگ نوارهایی که مجموعه‌های موازی را به هم متصل می‌کنند بر اساس مواد ساختمانی می‌باشند. به عنوان مثال، نمودار نشان می‌دهد که پل‌های چوبی عمدتاً دارای طول متوسط هستند (به همراه چند پل کوتاه)، عمدتاً در دوره سنّتی ساخته شده‌اند (به همراه چند پل با طول متوسط که در دوره‌های نوظهور و بلوغ برپا شده‌اند) و عمدتاً بر روی رودخانه Allegheny (به همراه چند پل سنّتی روی رودخانه Monongahela) قرار دارند. در مقابل، پل‌های آهنی همگی دارای طول متوسط هستند، عمدتاً در دورهٔ صنایع دستی ساخته شده‌اند و به نسبت‌های تقریباً مساوی بر روی رودخانه‌های Allegheny و Monongahela قرار دارند.



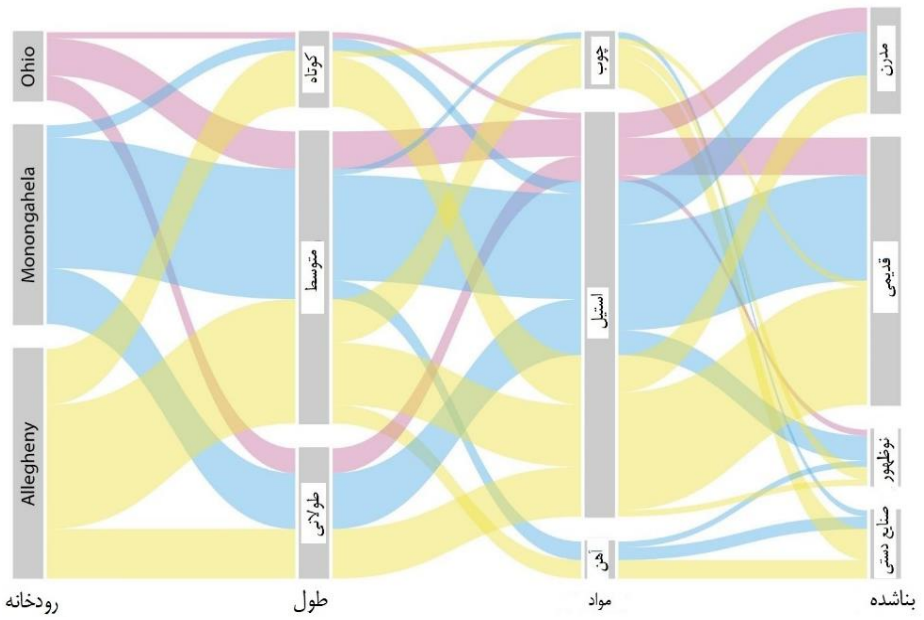
شکل ۱۱–۸. دسته‌بندی پل‌ها در پیتسبورگ بر اساس مصالح ساختمانی، طول، دورهٔ ساخت و رودخانه‌ای که روی آن‌ها قرار گرفته‌اند، که به‌صورت نمودار مجموعه‌های موازی نشان داده شده است. رنگ‌آمیزی نوارها، مصالح ساختمانی پل‌های مختلف را مشخّص می‌کند. منبع داده: Yoram Reich و Steven J. Fenves

اگـر رنـگ آمیـزی بـر اسـاس معیـار دیگـری ماننـد رودخانـه باشـد (شـکل ۱۱–۹)، نمـودار حاصـل متفـاوت خواهـد بـود. ایـن شـکل از نظـر بصـری شـلوغ اسـت و نوارهـای متقـاطع زیـادی دارد، امّـا نمودار حاکی از آن است که تقریباً تمام انواع پل‌ها بر روی هر رودخانه‌ای قرار دارد.



شکل ۱۱–۹. دسته‌بندی پل‌ها در پیتسبورگ بر اساس مصالح ساختمانی، طول، دوران ساخت و ساز و رودخانه‌ای که بر روی آن قرار دارند. این شکل شبیه به شکل ۱۱–۸ است، امّا اکنون رنگ‌آمیزی نوارها، رودخانه‌ای را که پل‌ها روی آن ساخته شده است را مشخّص می‌کند. این شکل به دلیل اینکه چیدمان نوارهای رنگی در وسط شکل بسیار شلوغ است و همچنین به این دلیل که نوارها باید از راست به چپ خوانده شوند، برچسب «زشت» دارد.
منبع داده: Yoram Reich و Steven J. Fenves

شکل ۱۱–۹ بـه عنـوان «زشـت» برچسـب‌گذاری شـده اسـت زیـرا بـیش از حـد پیچیـده و گیج‌کننـده اسـت. اولاً، از آنجـایی کـه مـا بـه خوانـدن از چـپ بـه راسـت عـادت کـرده‌ایم، مجموعـه‌هایی کـه رنگ‌آمیزی را مشخّص می‌کنند باید سـمت چـپ ظـاهر شـوند، نـه در سـمت راسـت. ایـن کـار باعـث می‌شـود کـه ببینیـد رنـگ‌آمیزی از کجـا منشـأ می‌گیـرد و چگونـه در مجموعـهٔ داده جریـان دارد. دوماً، بهتر است که ترتیب مجموعـه‌ها را طـوری تغییـر دهیـد کـه تعـداد نوارهـای متقـاطع بـه حـداقل برسد. با پیروی از این اصول، به شکل ۱۱–۱۰ می‌رسیم که بر شکل ۱۱–۹ ارجح است.



شکل ۱۱–۱۰. دسته‌بندی پل‌ها در پیتسبورگ بر اساس رودخانه، دوران ساخت و ساز، طول و مصالح ساختمانی. این شکل تنها در ترتیب مجموعه‌های موازی با شکل ۱۱–۹ متفاوت است. ترتیب اصلاح شده منجر به شکلی می‌شود که خواندن آن آسان‌تر است. منبع داده: Yoram Reich و Steven J. Fenves

فصل ۱۲

نمایش روابط بین دو یا چند متغیّر کمی

بسیاری از مجموعه‌های داده حاوی دو یا چند متغیّر کمّی هستند و ممکن است بخواهیم نحوه ارتباط این متغیّرها را با یکدیگر بررسی کنیم. برای مثال، ممکن است مجموعه داده‌ای از اندازه‌گیری‌های کمّی در حیوانات مختلف داشته باشیم، مانند قد، وزن، طول و انرژی مورد نیاز روزانه. برای رسم رابطه بین متغیّر مانند قد و وزن، معمولاً از نمودار پراکنش استفاده می‌کنیم. اگر بخواهیم بیش از دو متغیّر را همزمان نشان دهیم، ممکن است نمودار حبابی، ماتریس نمودار پراکنش[1] یا همبستگی نگار[2] را انتخاب کنیم. در نهایت، برای مجموعهٔ داده‌های با ابعاد وسیع، ممکن است کاهش ابعاد مثلاً توسّط تحلیل مولفه‌های اصلی، مفید باشد.

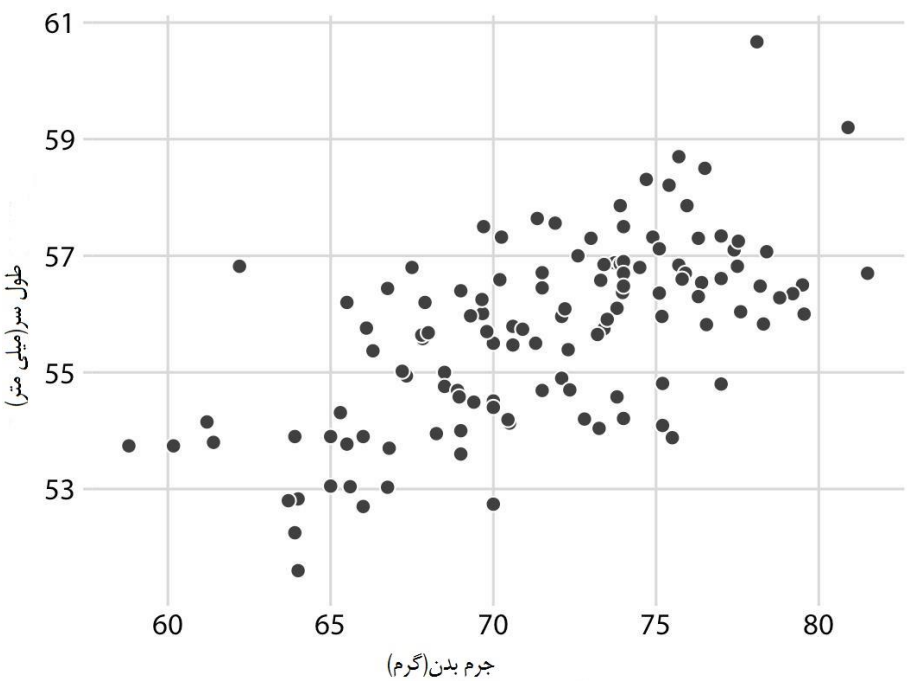
نمودارهای پراکنش

نمودار پراکنش اصلی و زیرشاخه‌های متنوّع آن را با استفاده از مجموعه اندازه‌گیری‌های انجام شده بر روی ۱۲۳ پرنده زاغ آبی نشان خواهیم داد. این مجموعه داده حاوی اطلاعاتی مانند طول سر (اندازه‌گیری از نوک منقار تا پشت سر)، اندازهٔ جمجمه (طول سر منهای طول نوک)، و تودهٔ بدنی هر پرنده است. انتظار داریم که بین این متغیّرها رابطه وجود داشته باشد. به عنوان مثال، انتظار می‌رود پرندگانی که منقارهای بلندتری دارند، اندازهٔ جمجمه بزرگ‌تری داشته

1. scatterplots
2. correlogram

باشند و پرندگان با تودهٔ بدنی بالاتر نسبت به پرندگان با تودهٔ بدنی کمتر، بایستی منقار طویل‌تر و جمجمه‌های بزرگ‌تری داشته باشند.

برای کشف این روابط، بیایید با رسم نمودار طول سر در برابر تودهٔ بدن شروع کنیم (شکل ۱۲–۱). در این نمودار طول سر در امتداد محور y و تودهٔ بدن در امتداد محور x رسم شده و هر پرنده با یک نقطه نشان داده شده است (به اصطلاح به کار رفته توجّه کنید: ما می‌گوییم که متغیّر نشان داده شده را در امتداد محور y در برابر متغیّر نشان داده شده در امتداد محور x رسم می‌کنیم). نقطه‌ها یک ابر پراکنده را تشکیل می‌دهند (از این رو اصطلاح پراکنش برای این نمودار به کار می‌رود)، امّا بدون شک روندی در داده‌ها مشهود است به صورتی که هرچه تودهٔ بدن پرنده بیشتر باشد، طول سر بیشتر خواهد بود. پرنده‌ای با طولانی‌ترین سر نزدیک به پرنده مشاهده شده با بیشترین تودهٔ بدنی است و پرنده با کوتاه‌ترین سر نزدیک به پرندهٔ مشاهده شده با کمترین تودهٔ بدنی است.

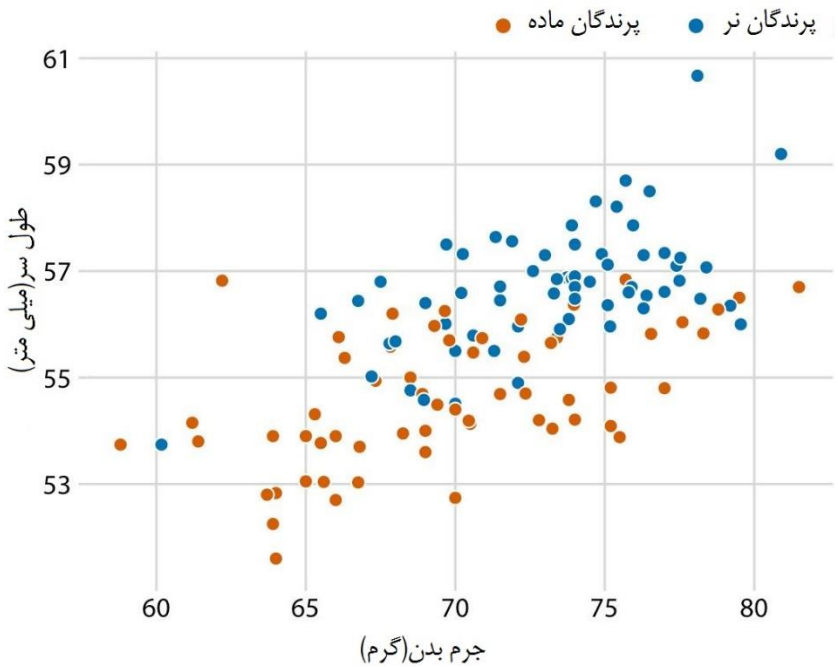


شکل ۱۲–۱. طول سر (اندازه‌گیری شده از نوک منقار تا پشت سر، بر حسب میلی‌متر) در مقایسه با تودهٔ بدن (بر حسب گرم)، برای ۱۲۳ زاغ آبی. هر نقطه مربوط به یک پرنده است. گرایش متوسطی وجود دارد که پرندگان سنگین‌تر، سرهای بلندتری داشته باشند. منبع داده: Keith Tarvin, Oberlin College

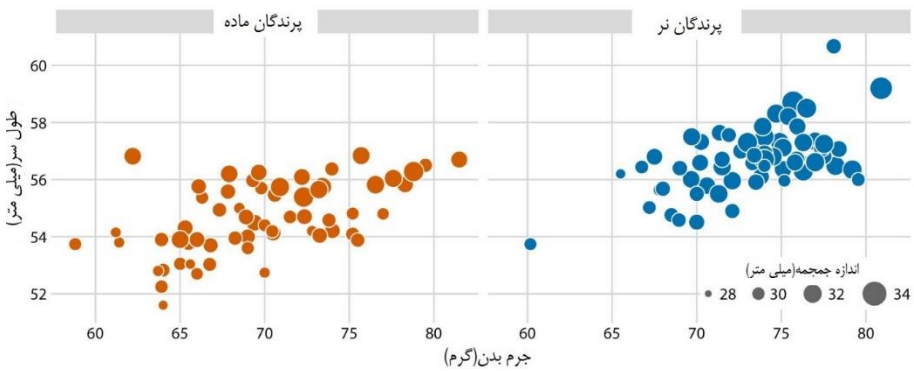
مجموعه داده زاغ آبی شامل پرندگان نر و ماده است و ممکن است بخواهیم بدانیم که آیا رابطهٔ کلی مشاهده شده بین طول سر و تودهٔ بدن در هر جنس نیز وجود دارد یا خیر. برای پاسخ به این سوال، می‌توانیم نقاط موجود در نمودار پراکنش را بر اساس جنسیّت پرنده رنگ‌آمیزی کنیم (شکل ۱۲-۲). این نمودار نشان می‌دهد که روند کلی ارتباط طول سر و تودهٔ بدن حداقل تا حدی به جنسیّت پرندگان بستگی دارد. با تودهٔ بدنی یکسان، پرندگان ماده سرهای کوتاه‌تری نسبت به پرندگان نر دارند. در عین حال، ماده‌ها به طور متوسط سبک‌تر از نرها هستند.

از آنجایی که طول سر به صورت فاصله از نوک منقار تا پشت سر تعریف می‌شود، طول سر بزرگتر می‌تواند به معنای منقار بلندتر، جمجمهٔ بزرگتر یا هر دو باشد. می‌توانیم طول منقار و اندازه جمجمه را با بررسی متغیّر دیگری در مجموعه داده یعنی اندازهٔ جمجمه که شبیه طول سر است امّا طول منقار را حذف می‌کند، جدا کنیم. از آنجایی که در حال حاضر از محور x برای تودهٔ بدن، محور y برای طول سر، و رنگ نقطه‌ها برای جنس پرنده استفاده می‌کنیم، به روش دیگری نیاز داریم تا بتوانیم اندازهٔ جمجمه را با کمک آن نمایش دهیم. یکی از گزینه‌ها استفاده از اندازهٔ نقاط است که منجر به شکل‌گیری نمودار حبابی می‌شود (شکل ۱۲-۳).

نمودارهای حبابی این عیب را دارند که انواع مشابهی از متغیّرها - متغیّرهای کمی - را با دو نوع مقیاس، موقعیّت و اندازه نشان می‌دهند. این امر تشخیص بصری روابط قوی بین متغیّرهای مختلف را دشوار می‌کند. علاوه بر این، درک تفاوت بین مقادیر داده‌های کدگذاری‌شده به صورت اندازهٔ حباب سخت‌تر از درک تفاوت‌های بین مقادیر داده‌های کدگذاری‌شده به صورت موقعیّت است. از آنجایی که حتی بزرگ‌ترین حباب‌ها در مقایسه با اندازهٔ کل نمودار باید تا حدودی کوچک باشند، تفاوت بین بزرگ‌ترین و کوچک‌ترین حباب‌ها لزوماً اندک است. در نتیجه، تفاوت‌های کوچک‌تر در داده‌ها به صورت تفاوت‌های خیلی کوچک‌تر نمایش داده می‌شود که رویت آن‌ها تقریباً غیرممکن است. در شکل ۱۲-۳، از نمایش اندازه‌ها استفاده کرده‌ایم تا به صورت بصری تفاوت بین کوچکترین جمجمه (حدود ۲۸ میلی متر) و بزرگترین جمجمه (حدود ۳۴ میلی متر) تقویت شود، با این حال همچنان تعیین رابطهٔ بین اندازه جمجمه با تودهٔ بدن یا طول سر دشوار است.

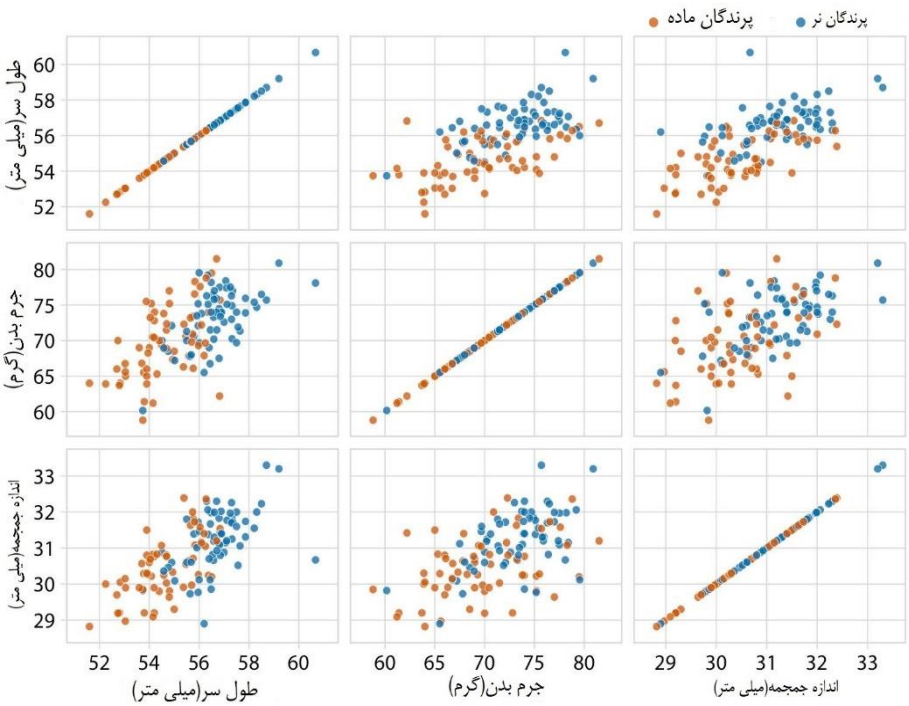


شکل ۱۲–۲. طول سر در مقایسه با تودهٔ بدن برای ۱۲۳ زاغ آبی. جنسیت پرندگان با رنگ مشخّص شده است. با تودهٔ بدنی مشخص، پرندگان نر سرهای بلندتری (و به طور خاص، نوک بلندتری) نسبت به پرندگان ماده دارند.
منبع داده: Keith Tarvin, Oberlin College



شکل ۱۲–۳. طول سر در مقایسه با تودهٔ بدن برای ۱۲۳ زاغ آبی. جنسیّت پرندگان با رنگ و اندازهٔ جمجمه پرندگان با اندازهٔ نماد مشخّص شده است. اندازه‌گیری طول سر شامل طول منقار نیز می‌شود در حالی که اندازه‌گیری اندازهٔ جمجمه اینطور نیست. طول سر و اندازهٔ جمجمه معمولاً با هم همبستگی دارند، امّا برخی از پرندگان با توجّه به اندازه جمجمه‌شان، به طور غیرمعمولی نوک‌های بلندتر یا کوتاه‌تری دارند. منبع داده: Keith Tarvin, Oberlin College

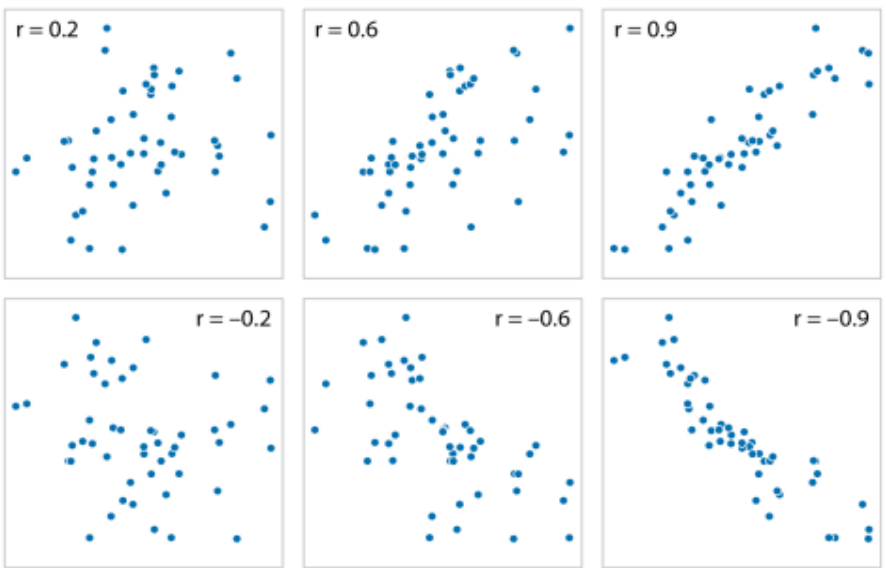
به عنوان جایگزینی برای نمودار حبابی، ممکن است نمایش یک ماتریس همه در برابر همه از نمودارهای پراکنش، که در آن هر نمودار دو بُعد داده را نشان می‌دهد، ارجح باشد (شکل ۱۲–۴). این نمودار به وضوح نشان می‌دهد که رابطهٔ بین اندازه جمجمه و تودهٔ بدن برای پرندگان ماده و نر قابل مقایسه است، با این تفاوت که پرندگان ماده تا حدودی کوچکتر هستند. با این حال، این موضوع برای رابطهٔ بین طول سر و تودهٔ بدن صادق نیست. تمایز واضحی بر اساس جنسیّت وجود دارد. پرندگان نر معمولاً نسبت به پرندگان ماده نوک بلندتری دارند و بقیه موارد برابر هستند.



نمودار ۱۲–۴. ماتریس پراکنش همه در برابر همه برای طول سر، توده بدن، و اندازهٔ جمجمه، برای ۱۲۳ زاغ آبی. این نمودار دقیقاً همان داده‌های نمودار ۱۲–۲ را نشان می‌دهد. از آنجایی که ما موقعیّت را بهتر از اندازهٔ نماد قضاوت می‌کنیم، درک همبستگی بین اندازهٔ جمجمه و دو متغیّر دیگر در نمودارهای پراکنش زوجی آسان‌تر از نمودار ۱۲–۲ است. منبع داده: Keith Tarvin, Oberlin College

همبستگی نگار

وقتی بیش از سه تا چهار متغیّر کمّی داریم، ماتریس‌های پراکنش همه در برابر همه غیرقابل استفاده می‌شوند. در این مورد، کمّی کردن میزان ارتباط بین جفت متغیّرها و نمایش این قدرت ارتباط به جای داده‌های خام مفیدتر است. یکی از راه‌های رایج برای انجام این کار، محاسبهٔ ضرایب همبستگی است. ضریب همبستگی r عددی بین ۱− و ۱+ است که میزان همسو و همگام بودن دو متغیّر را نشان می‌دهد. اگر r معادل صفر باشد به این معنی است که هیچ ارتباطی وجود ندارد و مقدار ۱+ یا ۱− یک همبستگی کامل را نشان می‌دهد. علامت ضریب همبستگی نشان می‌دهد که آیا متغیّرها همسو هستند (مقادیر بزرگتر در یک متغیّر با مقادیر بزرگتر در متغیّر دیگر همراه است) یا غیرهمسو (مقادیر بزرگتر در یک متغیّر با مقادیر کوچکتر در متغیّر دیگر همراه است). برای ارائهٔ مثال‌های بصری از اینکه ضرایب همبستگی مختلف چگونه به نظر می‌رسند، در شکل ۱۲–۵، مجموعه‌های تصادفی از نقاط را ارائه داده‌ایم که در میزان همبستگی مقادیر x و y بسیار متفاوت هستند.



نمودار ۱۲–۵. نمونه‌هایی از همبستگی با بزرگی و جهت مختلف، به همراه ارائهٔ ضریب همبستگی (r) مربوطه. در هر دو ردیف، همبستگی‌ها از سمت چپ به سمت راست قوی‌تر می‌شوند. در ردیف بالا همبستگی‌ها مثبت است (مقادیر بزرگتر برای یک کمیّت با مقادیر بزرگتر برای دیگری همراه است) و در ردیف پایین منفی هستند (مقادیر بزرگتر برای یک کمیّت با مقادیر کوچکتر برای دیگری همراه است). در هر شش حالت، مجموعهٔ مقادیر x و y یکسان هستند، امّا جفت‌های مقادیر x و y برای ایجاد ضرایب همبستگی مشخّص تغییر داده شده‌اند.

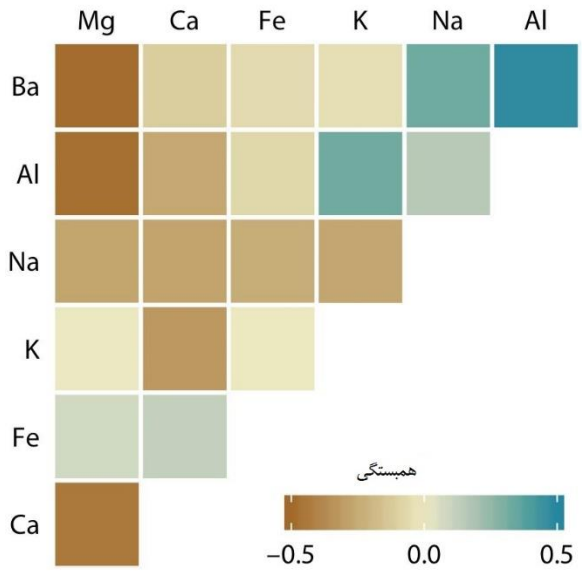
ضریب همبستگی به صورت زیر تعریف می‌شود:

$$r = \frac{\Sigma_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_i (x_i - \bar{x})^2} \sqrt{\Sigma_i (y_i - \bar{y})^2}}$$

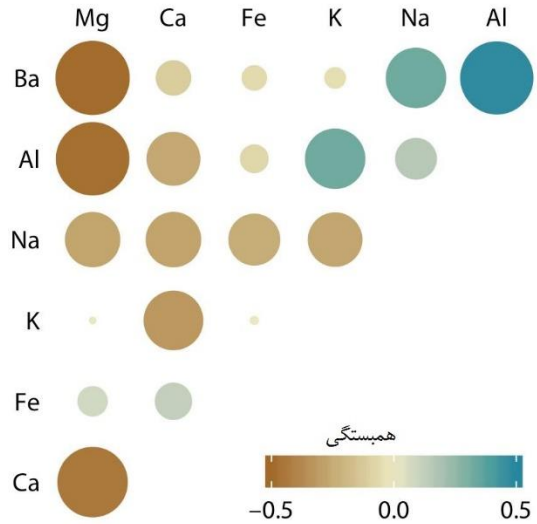
که در آن x_i و y_i دو مجموعه از مشاهدات و $\bar{x}$ و $\bar{y}$ میانگین نمونهٔ متناظر هستند. در خصوص این فرمول می‌توانیم نکات زیادی را بررسی کنیم. اوّل اینکه فرمول در خصوص x_i و y_i متقارن است، بنابراین همبستگی x با y همان همبستگی y با x است. دوّم، مقادیر مجزا x_i و y_i صرفاً در بستر تفاوت از میانگین نمونه مربوطه وارد فرمول می‌شوند، بنابراین اگر کل مجموعهٔ داده را بر اساس مقدار ثابتی تغییر دهیم – برای مثال، اگر x_i را با $x'_i = x_i + C$ جایگزین کنیم، یعنی افزودن مقدار ثابت C – ضریب همبستگی بدون تغییر باقی می‌ماند. سوّم، اگر داده‌ها را مجدداً مقیاس‌بندی کنیم (مثلاً $x'_i = Cx_i$) ضریب همبستگی بدون تغییر می‌ماند، زیرا مقدار ثابت C هم در صورت و هم در مخرج فرمول ظاهر می‌شود و بنابراین می‌توان آن را حذف کرد.

نمایش ضرایب همبستگی را همبستگی نگار می‌نامند. برای نمایش نحوهٔ به‌کارگیری از همبستگی نگار از مجموعهٔ داده حاوی بیش از ۲۰۰ قطعه شیشه که حین کار پزشکی قانونی به دست آمده استفاده خواهیم کرد. برای هر قطعه شیشه، اندازه‌گیری‌هایی در مورد ترکیب داریم که به صورت درصد وزنی اکسیدهای معدنی ارائه شده است. برای هفت اکسید مختلف مقادیر اندازه‌گیری شده را داریم که منجر به ایجاد ۲۱=۱+۲+۳+۴+۵+۶ جفت همبستگی خواهد شد. ما می‌توانیم این ۲۱ همبستگی را به طور همزمان در قالب ماتریسی از کاشی‌های رنگی که هر کدام معرّف یک ضریب همبستگی می‌باشد، نمایش دهیم (نمودار ۱۲–۶). این همبستگی نگار به ما اجازه می‌دهد که به سرعت روندهای کلی موجود در داده‌ها را درک نماییم مثلاً اینکه منیزیم همبستگی منفی با تقریبا تمام سایر اکسیدها داشته در حالی که آلومینیوم و باریوم همبستگی مثبت قوی‌ای دارند.

یکی از نقاط ضعف همبستگی نگار نمایش داده شده در نمودار ۱۲–۶ این است که همبستگی‌های ضعیف – یعنی همبستگی با مقدار مطلق نزدیک به صفر – آنطور که باید از نظر بصری تضعیف نمی‌شوند. برای مثال، منیزیم (Mg) و پتاسیم (K) اصلاً همبستگی ندارند امّا نمودار ۱۲–۶ این موضوع را به خوبی نشان نمی‌دهد. برای غلبه بر این محدودیت، می‌توانیم همبستگی‌ها را به صورت دایره‌های رنگی نشان دهیم و اندازهٔ دایره را مبتنی بر مقدار مطلق ضریب همبستگی ترسیم نماییم (نمودار ۱۲–۷). به این ترتیب همبستگی‌های کوچک تضعیف می‌شوند و همبستگی‌های قوی بهتر نمایش داده می‌شوند.



نمودار ۱۲–۶. همبستگی در محتوای مواد معدنی برای ۲۱۴ نمونه از قطعات شیشه به دست آمده در طول کار پزشکی قانونی. مجموعهٔ داده شامل هفت متغیّر است که مقادیر منیزیم (Mg)، کلسیم (Ca)، آهن (Fe)، پتاسیم (K)، سدیم (Na)، آلومینیوم (Al) و باریوم (Ba) موجود در هر قطعه شیشه را نشان می‌دهد. کاشی‌های رنگی نشان‌دهندهٔ همبستگی دو به دوی این متغیّرها هستند. منبع داده ها: B. German.



نمودار ۱۲–۷. همبستگی در محتوای مواد معدنی برای نمونه‌های شیشه پزشکی قانونی. مقیاس رنگ‌بندی مشابه نمودار ۱۲–۶ است. با این حال، اکنون بزرگی هر همبستگی توسّط اندازهٔ دایره‌های رنگی نیز نمایش داده می‌شود. در این حالت از نظر بصری بر مواردی که همبستگی‌های نزدیک به صفر دارند تأکید کمتری می‌شود. منبع داده‌ها: B. German.

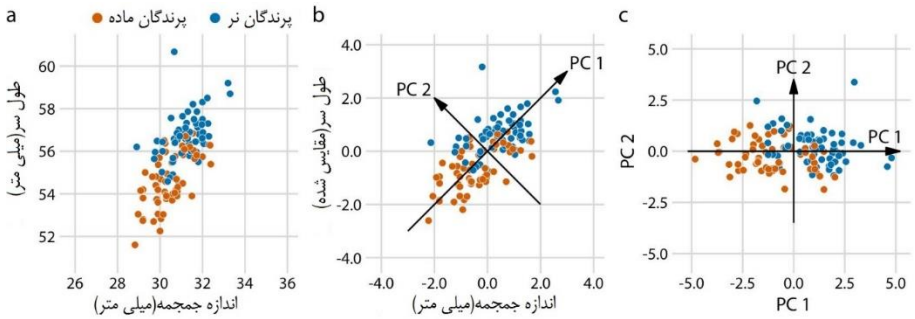
همه همبستگی نگارها یک اشکال مهم دارند: آن‌ها نسبتاً انتزاعی هستند. در حالی که الگوهای مهم داده‌ها را به ما نشان می‌دهند، نقاط دادهٔ زیربنایی را نیز پنهان می‌کنند و ممکن است باعث شوند نتیجه‌گیری نادرستی انجام شود. همیشه بهتر است داده‌های خام را به جای مقادیر انتزاعی مشتق شده از آن‌ها، نمایش دهیم. خوشبختانه، اغلب می‌توانیم حد وسطی بین نشان دادن الگوهای مهم و نمایش داده‌های خام با استفاده از تکنیک‌های کاهش ابعاد پیدا کنیم.

کاهش ابعاد

کاهش ابعاد، متّکی به این مساله کلیدی است که اکثر مجموعه داده‌های با ابعاد بالا متشکل از چندین متغیّر همبسته هستند که حاوی اطّلاعات همپوشان هستند. چنین مجموعه داده‌هایی را می‌توان به تعداد کمتری از ابعاد کلیدی بدون از دست دادن اطّلاعات مهم کاهش داد. به عنوان یک مثال ساده و شهودی، مجموعه‌ای از چندین خصوصیت فیزیکی افراد را در نظر بگیرید، از جمله مقادیری مانند قد و وزن، طول دست‌ها و پاها، دور کمر، باسن، قفسهٔ سینه و غیره. به طور شهودی می‌دانیم که همهٔ این مقادیر در درجهٔ اوّل و عمدتاً با اندازهٔ کلی هر فرد مرتبط است. در صورت مساوی بودن سایر شرایط، یک فرد بزرگتر قاعدتاً قد بلندتر، سنگین وزن‌تر، بازوها و پاهای بلندتر و دور کمر، دور باسن و دور سینه بزرگتری خواهد داشت. بُعد مهم بعدی جنسیّت فرد خواهد بود. متغیّرهای مذکور حتی در مردان و زنان با اندازه‌های مشابه به طور قابل‌توجهی متفاوت است. به عنوان مثال حتی اگر سایر متغیّرها مشابه باشد، دور باسن زنان عمدتاً بیشتر از مردان است.

روش‌های زیادی برای کاهش ابعاد وجود دارد. ما در اینجا فقط یک روش که پرکاربردترین روش نیز می‌باشد را مورد بحث قرار خواهیم داد: تحلیل مولفه‌های اصلی (PCA)[1]. PCA مجموعهٔ جدیدی از متغیّرها را به نام مؤلفه‌های اصلی (PCs) با ترکیب خطی متغیّرهای اصلی در داده‌ها ارائه می‌کند که با میانگین صفر و واحد واریانس استاندارد شده است (نمودار ۱۲–۸ را برای نمونه‌ای در دو بُعد ببینید). مولفه‌های اصلی به‌گونه‌ای انتخاب می‌شوند که همبستگی نداشته باشند، و به گونه‌ای مرتب می‌شوند که اوّلین مؤلفه بیشترین تنوّع ممکن را در داده‌ها شامل شود و مؤلفه‌های بعدی به تدریج معرّف تنوع کمتری در داده‌ها باشند. معمولاً ویژگی‌های کلیدی در داده‌ها را می‌توان تنها با دو یا سه مؤلفه اصلی اوّل مشاهده کرد.

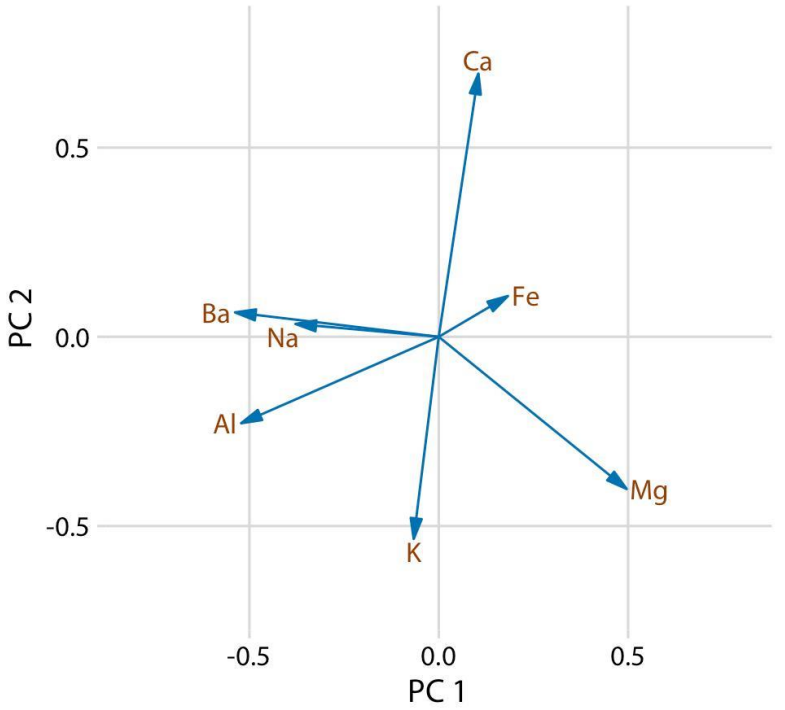
1. principal components analysis



نمودار ۱۲–۸. مثالی برای تحلیل مؤلفه‌های اصلی در دو بُعد. الف) داده‌های اصلی. به عنوان داده‌های نمونه، از اندازه‌گیری‌های طول سر و اندازهٔ جمجمه از مجموعه دادهٔ زاغ آبی استفاده می‌کنیم. پرندگان ماده و نر توسّط رنگ متمایز شده‌اند، امّا این تمایز تأثیری بر PCA ندارد. (ب) به عنوان اوّلین مرحله در PCA، مقادیر داده‌های اصلی را به میانگین و واحد واریانس تبدیل می‌کنیم. سپس متغیّرهای جدید (مؤلفه‌های اصلی) را در امتداد جهت حداکثر تغییرات در داده‌ها تعریف می‌کنیم. (ج) در نهایت، داده‌ها در قالب مختصّات جدید برازش می‌شوند. از نظر ریاضی، این برازش معادل چرخش داده‌ها حول نقطه مبدا است. در مثال دو بُعدی نشان داده شده، داده‌ها در جهت عقربه‌های ساعت ۴۵ درجه چرخیده‌اند. منبع داده: Keith Tarvin, Oberlin College

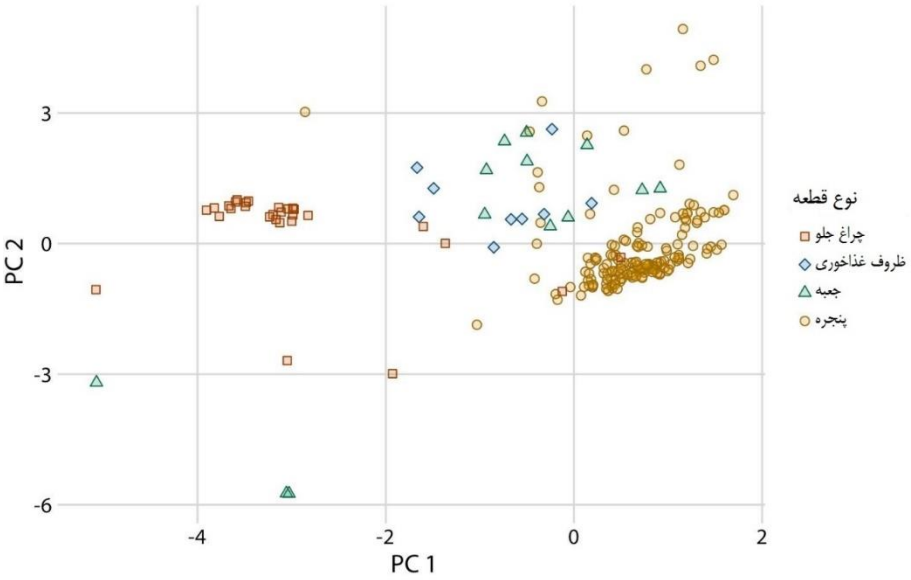
وقتی PCA را انجام می‌دهیم، عموماً به دو بخش اطّلاعات علاقه‌مندیم: ترکیب مؤلفه‌های اصلی و مکان تک تک نقاط داده در فضای مؤلفه‌های اصلی. بیایید به این دو بخش در یک PCA از مجموعه داده‌های پزشکی قانونی نگاه کنیم.

ابتدا به ترکیب مؤلفه‌ها نگاه می‌کنیم (نمودار ۱۲–۹). در اینجا، ما فقط دو مؤلفهٔ اوّل، PC 1 و PC 2 را در نظر می‌گیریم. از آنجایی که مؤلفه‌های اصلی ترکیبی خطی از متغیّرهای اصلی هستند (پس از استانداردسازی)، می‌توانیم متغیّرهای اصلی را به صورت فلش‌هایی نشان دهیم تا ببینیم چه اندازه به مؤلفه‌های اصلی کمک می‌کنند. در اینجا، می‌بینیم که باریوم و سدیم در درجهٔ اوّل به PC 1 کمک می‌کنند و نه به PC 2، کلسیم و پتاسیم در درجهٔ اوّل به PC 2 کمک می‌کنند و نه PC 1، و متغیّرهای دیگر در مقادیر متفاوت به هر دو مؤلفه کمک می‌کنند. طول فلش‌ها متفاوت است زیرا بیش از دو مؤلفه اصلی وجود دارد. به عنوان مثال، فلش آهن به طور مشخّص کوتاه است زیرا در درجهٔ اوّل به مؤلفه‌های اصلی درجه بالاتر کمک می‌کند (نمایش داده نشده است).



نمودار ۱۲–۹. ترکیب دو مؤلفهٔ اوّل در تحلیل مؤلفه‌های اصلی مجموعه داده پزشکی قانونی. مؤلفهٔ اوّل (PC 1) در درجهٔ اوّل مقدار آلومینیوم، باریوم، سدیم و منیزیم را در یک قطعه شیشه اندازه‌گیری می‌کند، در حالی که مؤلفهٔ دوّم (PC 2) در درجهٔ اوّل مقدار کلسیم و پتاسیم و تا حدی مقدار آلومینیوم و منیزیم را اندازه‌گیری می‌کند. منبع داده ها: B. German.

سپس، داده‌های اصلی را در فضای مؤلفه‌های اصلی برازش می‌کنیم (نمودار ۱۲–۱۰). در این نمودار یک خوشه‌بندی تعریف شده از انواع متمایز قطعات شیشه قابل مشاهده است. تکه‌هایی از چراغ‌های جلو و پنجره‌ها در مناطقی مشخصی در مؤلفهٔ اصلی با تعداد کمتری داده پرت قرار می‌گیرند. تکه‌های مربوط به ظروف غذاخوری و ظروف کمی بیشتر پخش شده‌اند، امّا با این وجود به وضوح از تکه‌های چراغ جلو و پنجره متمایز هستند. از مقایسه نمودار ۱۲–۱۰ با نمودار ۱۲–۹، می‌توان نتیجه گرفت که نمونه‌های پنجره دارای محتوای منیزیم بالاتری نسبت به میانگین و میزان باریوم، آلومینیوم و سدیم کمتری نسبت به میانگین می‌باشند، در حالی که عکس این موضوع برای نمونه‌های چراغ جلو صادق است.

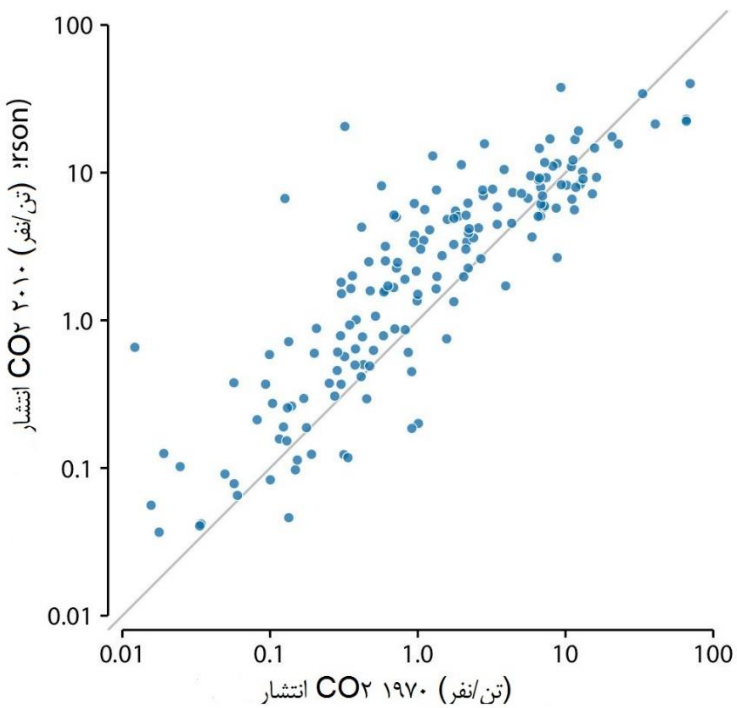


نمودار ۱۲–۱۰. ترکیب تکه‌های شیشه که در فضای مؤلفهٔ اصلی تعریف شده در نمودار ۱۲–۹ نمایش داده شده است. همانگونه که قابل مشاهده است، انواع مختلف نمونه‌های شیشه در مقادیر مشخصی از مؤلفه‌های اصلی ۱ و ۲ جمع شده‌اند. به طور ویژه، قطعات چراغ‌های جلو با مقدار PC 1 منفی مشخّص می‌شوند در حالی که قطعات پنجره تمایل به داشتن مقدار PC 1 مثبت دارند. قطعات ظروف غذاخوری و ظروف دارای مقادیر PC 1 نزدیک به صفر هستند و تمایل به داشتن مقادیر PC 2 مثبت دارند. با این حال، چند استثنا وجود دارد که در آن قطعات ظروف هم مقدار PC 1 منفی و هم مقدار PC 2 منفی دارند. این‌ها قطعاتی هستند که ترکیب آن‌ها کاملاً با سایر قطعات تحلیل شده متفاوت است. منبع داده ها: B. German.

داده‌های زوجی

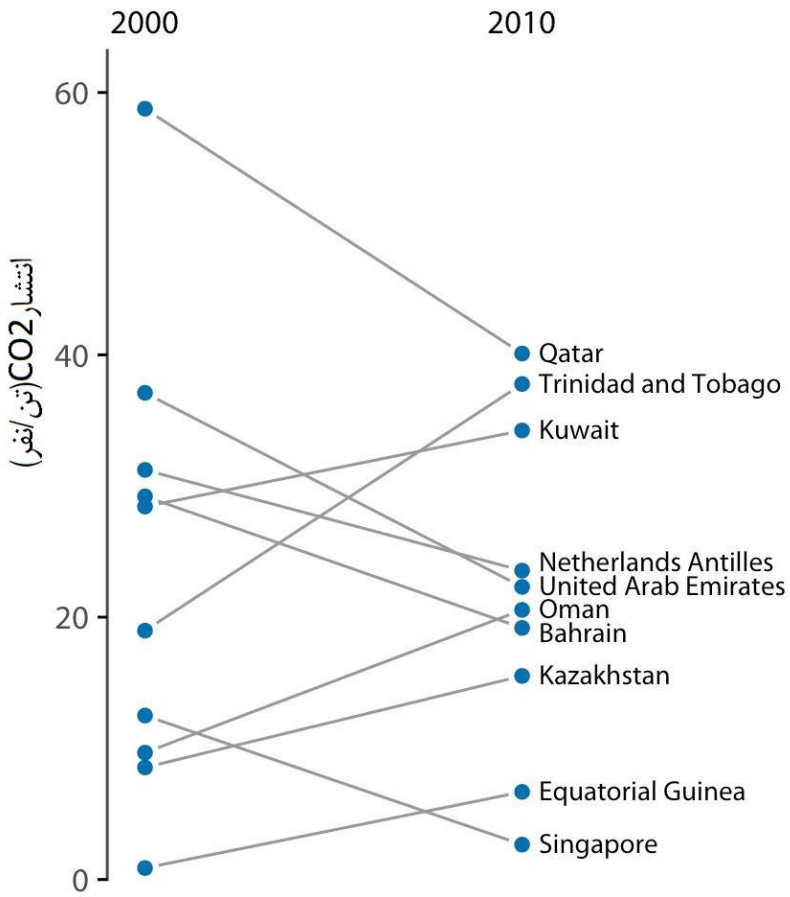
یک مورد خاص از داده‌های کمّی، داده‌های زوجی است: داده‌هایی که در آن دو یا چند اندازه‌گیری از یک واحد مطالعاتی مشخّص در شرایط اندکی متفاوت وجود دارد. مثال‌های آن عبارتند از دو اندازه‌گیری برای هر فرد (به عنوان مثال، طول بازوی راست و چپ یک فرد)، اندازه‌گیری‌های تکراری روی یک فرد در زمان‌های مختلف (به عنوان مثال، وزن یک فرد در دو زمان مختلف در طول سال)، یا اندازه‌گیری در دو فرد بسیار مرتبط (مانند قد دوقلوهای همسان). برای داده‌های زوجی، منطقی است که فرض کنیم اندازه‌گیری‌های متعلق به هر جفت نسبت به اندازه‌گیری‌های متعلق به جفت‌های دیگر، به یکدیگر شبیه‌تر هستند. دوقلوها تقریباً هم قد خواهند بود امّا از این نظر با دوقلوهای دیگر متفاوت خواهند بود. بنابراین، برای داده‌های زوجی، باید نمودارهایی به کار روند که تفاوت‌های بین اندازه‌گیری‌های زوجی را برجسته کند.

یک انتخاب عالی در این مورد ترسیم نمودار پراکنش ساده به همراه خط مورّبی است که $x = y$ را نشان می‌دهد. در چنین نموداری، اگر تنها تفاوت بین دو اندازه‌گیری هر جفت، خطای تصادفی باشد، تمام نقاط به طور متقارن در اطراف این خط پراکنده خواهند شد. در مقابل، هرگونه تفاوت نظام‌مند در اندازه‌گیری‌های زوجی، به صورت جابجایی نظام‌مند داده به بالا یا پایین خط مذکور قابل مشاهده خواهد بود. به عنوان مثال، انتشار دی اکسید کربن (CO_2) به ازای هر نفر را در نظر بگیرید که برای ۱۶۶ کشور در سال ۱۹۷۰ و ۲۰۱۰ اندازه‌گیری شده است (نمودار ۱۲–۱۱). این مثال دو ویژگی مشترک داده‌های زوجی را برجسته می‌کند. اوّل اینکه اکثر نقاط نسبتاً نزدیک به خط مورّب هستند. اگرچه انتشار CO_2 در برخی کشورها چهار برابر کشورهای دیگر است، امّا در هر کشور در یک بازهٔ زمانی ۴۰ ساله نسبتاً ثابت است. دوّم، نقاط به طور نظام‌مند نسبت به خط مورّب به سمت بالا جابجا شده‌اند. اکثر کشورها در این ۴۰ سال شاهد افزایش انتشار CO_2 بوده‌اند.



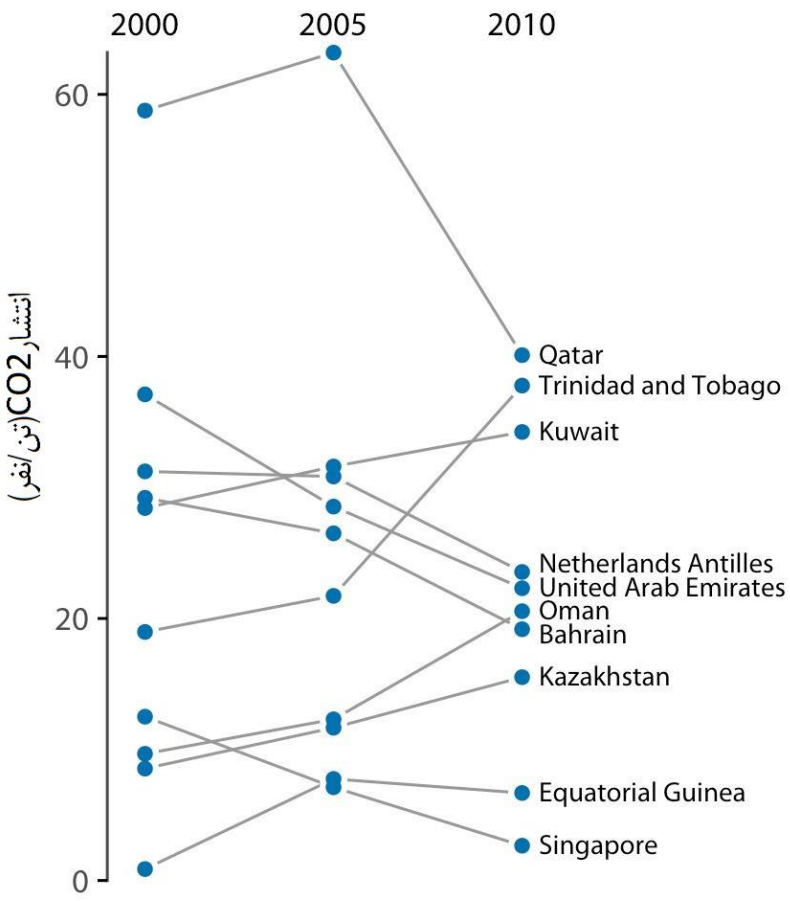
نمودار ۱۲–۱۱. انتشار دی اکسید کربن به ازای هر نفر برای ۱۶۶ کشور مختلف در سال‌های ۱۹۷۰ و ۲۰۱۰. هر نقطه نشان‌دهندهٔ یک کشور است. خط مورّب نشان‌دهندهٔ انتشار CO_2 یکسان در سال‌های ۱۹۷۰ و ۲۰۱۰ است. نقاط به طور نظام‌مند نسبت به خط مورّب به سمت بالا جابجا شده‌اند: در اکثر کشورها، انتشار CO_2 در سال ۲۰۱۰ بیشتر از سال ۱۹۷۰ بود. منبع داده: مرکز تحلیل اطّلاعات دی اکسید کربن.

زمانی که تعداد زیادی داده داریم و/یا علاقه‌مند به بررسی انحراف نظام‌مند مجموعه داده نسبت به فرضیه صفر برابری آن‌ها در بازه‌های زمانی مختلف هستیم، نمودارهای پراکنش مانند نمودار ۱۲–۱۱ مفید هستند. در مقابل، اگر تعداد مشاهدات کم باشد و در درجهٔ اوّل به تغییر وضعیّت هر فرد علاقه‌مند باشیم، شیب‌نگار ممکن است انتخاب بهتری باشد. در شیب‌نگار، اندازه‌گیری‌های فردی را به صورت نقطه‌هایی که در دو ستون چیده شده‌اند ترسیم می‌کنیم و با اتصال نقاط زوجی با یک خط، جفت‌ها را نشان می‌دهیم. شیب هر خط، بزرگی و جهت تغییر را نشان می‌دهد. نمودار ۱۲–۱۲ از این رویکرد برای نشان دادن ۱۰ کشور با بیشترین تغییر در انتشار CO_2 به ازای هر نفر از سال ۲۰۰۰ تا ۲۰۱۰ استفاده نموده است.



نمودار ۱۲–۱۲. انتشار دی اکسید کربن به ازای هر نفر در سال‌های ۲۰۰۰ و ۲۰۱۰، برای ۱۰ کشور با بیشترین تغییر بین این دو سال. منبع داده‌ها: مرکز تحلیل اطّلاعات دی اکسید کربن.

شیب نگارها یک مزیّت مهم نسبت به نمودارهای پراکنش دارند: از آن‌ها می‌توان برای مقایسهٔ بیش از دو اندازه‌گیری استفاده کرد. برای مثال، می‌توانیم نمودار ۱۲–۱۲ را به گونه‌ای تغییر دهیم تا انتشار CO_2 را در سه نقطهٔ زمانی نشان دهیم: سال‌های ۲۰۰۰، ۲۰۰۵ و ۲۰۱۰ (نمودار ۱۲–۱۳). این نمودار کشورهایی که تغییرات زیادی در انتشار CO_2 در طول کل این دهه دارند و همچنین کشورهایی مانند قطر یا ترینیداد و توباگو که تفاوت زیادی در روند مشاهده شده در فاصله پنج ساله اوّل و دوّم وجود دارد، را برجسته می‌کند.



نمودار ۱۲–۱۳. انتشار CO_2 به ازای هر نفر در سال‌های ۲۰۰۰، ۲۰۰۵ و ۲۰۱۰، برای ۱۰ کشور با بیشترین تفاوت بین سال‌های ۲۰۰۰ و ۲۰۱۰. منبع داده: مرکز تجزیه و تحلیل اطّلاعات دی اکسید کربن.

فصل ۱۳

نمایش سری‌های زمانی و سایر توابع یک متغیّر مستقل

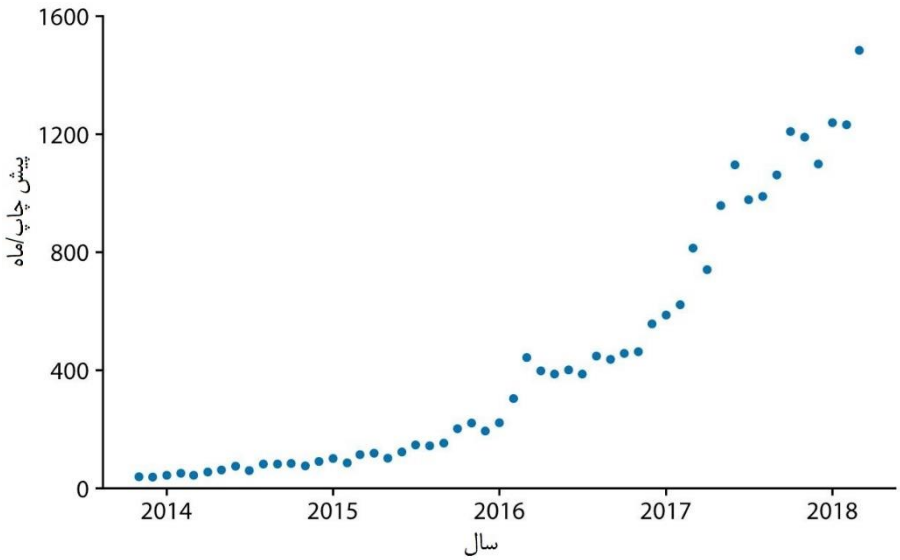
فصل قبل نمودارهای پراکنش را مورد بحث قرار داد، جایی که ما یک متغیّر کمی را در برابر دیگری، رسم می‌کنیم. مورد خاص زمانی پیش می‌آید که یکی از دو متغیّر، زمان باشد، زیرا زمان ساختار اضافی را بر داده‌ها تحمیل می‌کند. در این حالت داده‌ها، ترتیب ذاتی داشته و می‌توان آن‌ها را به ترتیب افزایش زمان مرتّب نمود و برای هر داده یک ماقبل و مابعد تعریف کرد. اغلب می‌خواهیم این ترتیب زمانی را نمایش دهیم و این کار را با نمودارهای خطی انجام می‌دهیم. با این حال نمودارهای خطی محدود به سری‌های زمانی نیستند. این نمودارها زمانی که یک متغیّر، ترتیبی را به داده‌ها تحمیل کند، مناسب هستند. همچنین به عنوان مثال این سناریو، در یک کارآزمایی شاهددار که در آن یک متغیّر درمان به طور هدفمند روی طیفی از مقادیر مختلف تنظیم می‌شود، قابل استفاده است. اگر متغیّرهای متعدد وابسته به زمان وجود داشته باشد، می‌توان نمودارهای خطی جداگانه ترسیم نمود یا اینکه یک نمودار پراکنش منظّم کشیده و سپس خطوطی را برای اتصال نقاط مجاور در زمان، ترسیم کنیم.

سری‌های زمانی منفرد

به عنوان اوّلین نمایش یک سری زمانی، الگوی ماهانه ارسال پیش‌چاپ‌ها[1] در زیست‌شناسی را

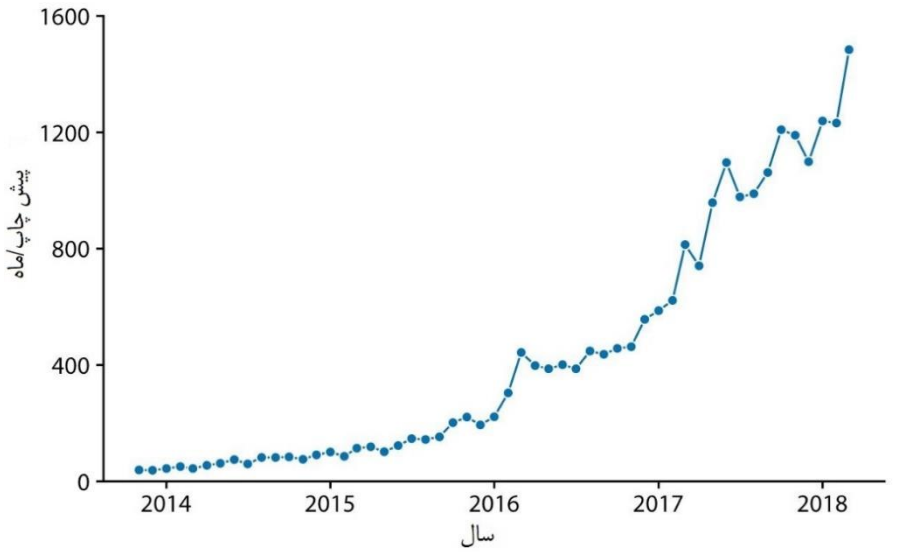
1. Preprints

در نظر خواهیم گرفت. پیش‌چاپ‌ها مقالات علمی هستند که محققان، قبل از داوری همتا و انتشار در یک مجله علمی، آن‌ها را به صورت برخط منتشر می‌کنند. سرور پیش چاپ bioRxiv که در نوامبر ۲۰۱۳ به طور خاص برای محققان علوم زیست‌شناسی تأسیس شد، رشد قابل توجّهی در دریافت این مقالات داشته است. می‌توانیم این رشد را با رسم نمودار پراکنش (فصل ۱۲) نشان دهیم. در اینگونه نمودارها نقاطی را ترسیم می‌کنیم که تعداد ارسال مقالات را در هر ماه نشان می‌دهد (نمودار ۱۳–۱).



نمودار ۱۳–۱، ارسال ماهانهٔ مقالات به سرور پیش‌چاپ bioRxiv از زمان تاسیس آن در نوامبر ۲۰۱۳ تا آوریل ۲۰۱۸. هر نقطه، تعداد ارسال مقالات در هر ماه را نشان می‌دهد. در طول این دورهٔ ۴٫۵ ساله افزایش مستمری در مقالات ارسالی قابل مشاهده است. منبع داده: Jordan Anaya

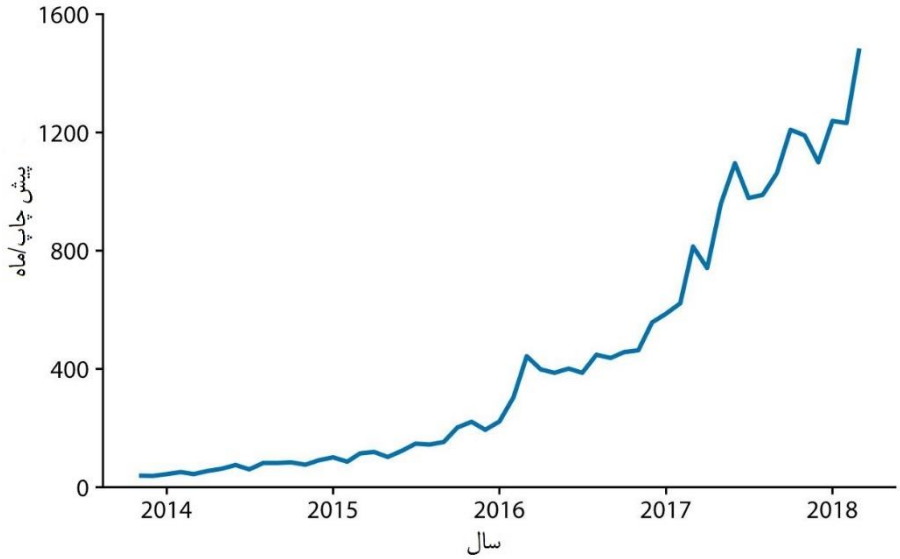
با این حال، تفاوت مهمی بین نمودار ۱۳–۱ و نمودارهای پراکنش که در فصل ۱۲ بحث شد، وجود دارد. در نمودار ۱۳–۱، نقاط به طور یکسانی در امتداد محور x قرار گرفته‌اند و نظم مشخصی در بین آن‌ها وجود دارد. هر نقطه دقیقاً یک همسایه چپ و یک همسایه راست دارد (به جز چپ‌ترین و راست‌ترین نقاط که هر یک فقط یک همسایه دارند). می‌توان با اتصال نقاط همسایه به یکدیگر بر این نظم و ترتیب به صورت بصری تأکید نمود (نمودار ۱۳–۲). چنین نموداری را نمودار خطی می‌نامند.



نمودار ۱۳–۲. ارسال ماهانهٔ مقالات به سرور پیش‌چاپ bioRxiv، که به‌صورت نقطه‌های متصل شده توسّط خطوط، نشان داده شده است. خطوط نشان‌دهندهٔ داده نبوده و فقط به عنوان راهنمای بصری در نظر گرفته شده‌اند. اتصال نقاط منفرد به وسیله خطوط، تأکید می‌کند که بین نقاط، نظمی وجود دارد: هر نقطه یک همسایه دارد که قبل از آن می‌آید و یک همسایه که بعد از آن می‌آید. منبع داده: Jordan Anaya

برخی افراد با رسم خطوط بین نقاط مخالف هستند زیرا خطوط، معرّف داده‌های مشاهده شده نیستند. به ویژه، اگر فقط چند مشاهده با فاصله دور از هم وجود داشته باشد، اگر مشاهداتی در زمان‌های میانی انجام می‌شد، احتمالاً به طور دقیق روی خطوط نشان داده شده قرار نمی‌گرفتند. بنابراین، می‌توان گفت خطوط مصداق داده‌سازی هستند. با این حال، هنگامی که نقاط از هم فاصله زیادی داشته یا این فاصله‌ها نامساوی می‌باشند، ممکن است به درک موضوع کمک کنند. تا حدودی می‌توان این معضل را با اشاره به آن در زیرنویس نمودار حل نمود، مثلاً با نوشتن عبارت «خطوط برای راهنمایی بصری هستند»(به زیرنویس نمودار ۱۳–۲ مراجعه کنید).

با این وجود، استفاده از خطوط برای نمایش سری‌های زمانی به طور کلی پذیرفته شده است و معمولاً همه نقاط نیز حذف می‌شوند (نمودار ۱۳–۳). بدون وجود نقاط، نمودار بیشتر بر روند کلی داده‌ها تاکید می‌کند و کمتر بر مشاهدات فردی. همچنین نمودار بدون نقطه از نظر بصری، شلوغی کمتری دارد. به طور کلی، هر چه سری زمانی متراکم‌تر باشد، نشان دادن مشاهدات فردی به وسیله نقاط، اهمیّت کمتری دارد. برای مجموعه داده پیش‌چاپ که در اینجا نشان داده شده است، به نظر می‌رسد حذف نقطه‌ها ایدهٔ خوبی است.

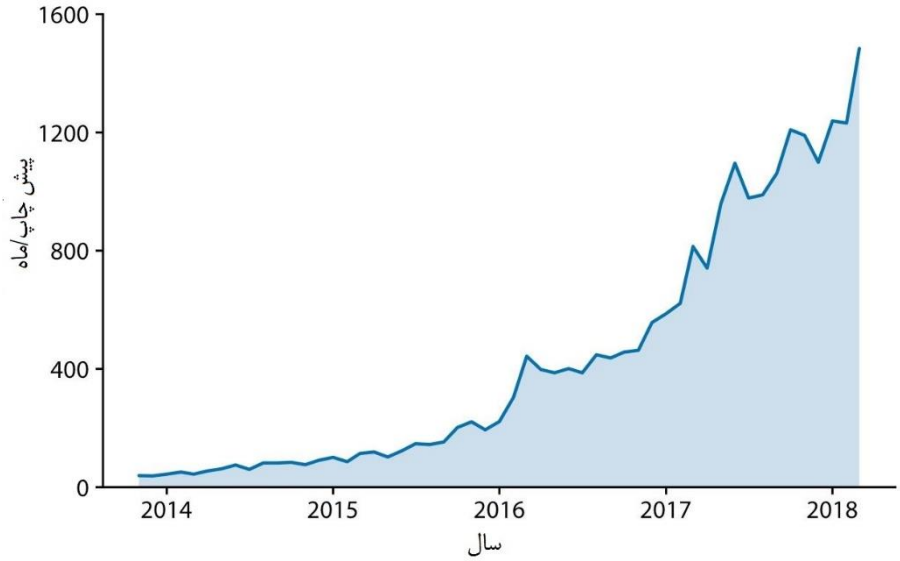


نمودار ۱۳–۳. ارسال ماهانهٔ مقالات به سرور پیش‌چاپ bioRxiv، که به صورت نمودار خطی و بدون نقاط نشان داده شده است. حذف نقاط منجر به افزایش تاکید بر روند کلی زمانی و کاهش تاکید بر مشاهدات فردی در نقاط زمانی خاص می‌شود. این اقدام، به ویژه زمانی مفید است که نقاط زمانی بسیار متراکم هستند. منبع داده: Jordan Anaya

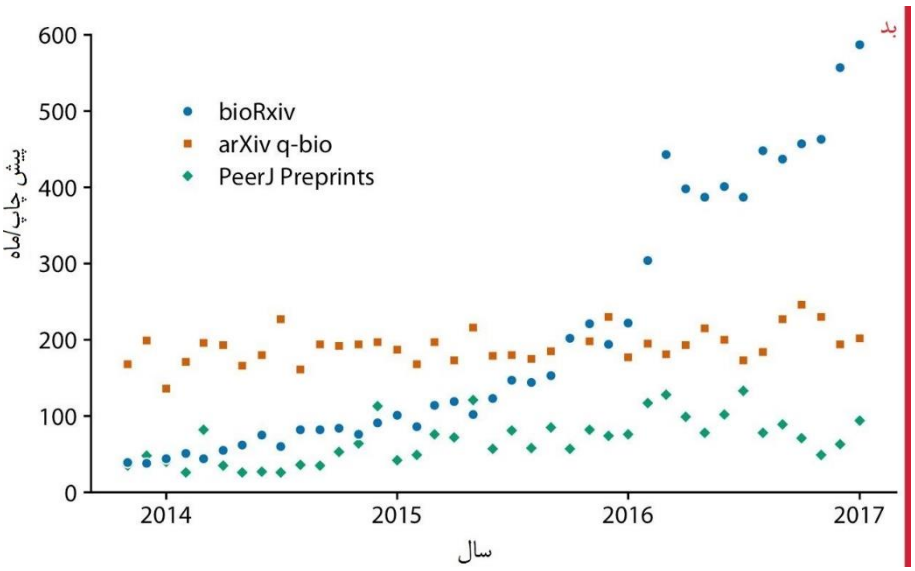
همچنین می‌توان ناحیه زیر منحنی را با یک رنگ ثابت پر نمود (نمودار ۱۳–۴). این روش، بیشتر بر روند تغییرات در داده‌ها تأکید می‌کند، زیرا به صورت بصری، ناحیه بالای منحنی را از ناحیه زیر منحنی جدا می‌کند. با این حال، این نمودار تنها در صورتی معتبر است که محور عمودی از صفر شروع شود، به طوری که ارتفاع ناحیه سایه‌دار در هر نقطهٔ زمانی، نشان‌دهندهٔ مقدار داده در آن نقطهٔ زمانی می‌باشد.

سری زمانی چندگانه و منحنی‌های دوز–پاسخ

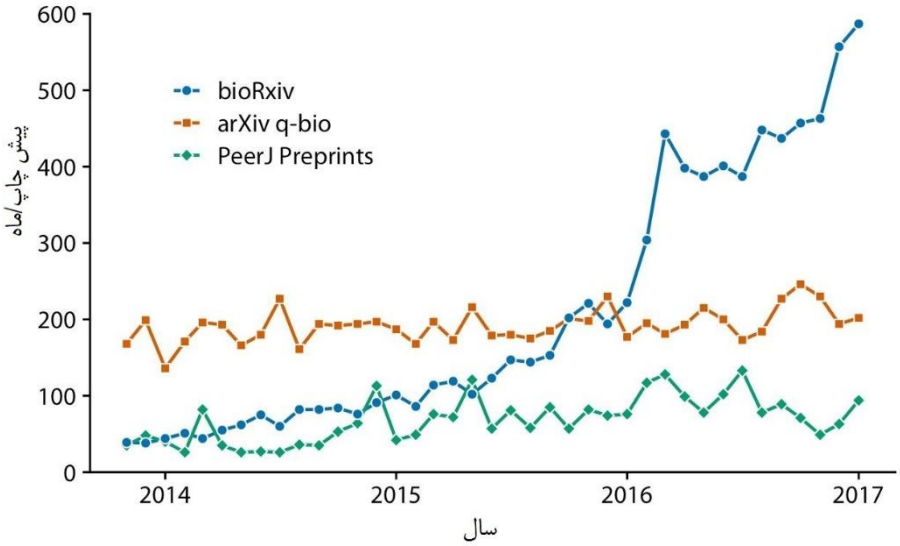
اغلب دوره‌های زمانی متعددی وجود دارد که بایستی به صورت همزمان نمایش داده شود. در این حالت، باید در نحوهٔ رسم داده‌ها دقّت بیشتری صرف شود، زیرا این نمودار می‌تواند گیج‌کننده شده یا خواندن آن سخت شود. به عنوان مثال، اگر هدف نمایش ارسال ماهانه مقالات به چندین سرور پیش‌چاپ باشد، رسم نمودار پراکنش ایده خوبی نیست، زیرا دوره‌های زمانی با یکدیگر همپوشانی دارند (نمودار ۱۳–۵). اتصال نقاط داده با خطوط، این مشکل را برطرف می‌کند (نمودار ۱۳–۶).



نمودار ۱۳–۴. ارسال ماهانهٔ مقالات به سرور پیش‌چاپ bioRxiv، که به صورت نمودار خطی نشان داده شده و ناحیهٔ زیر منحنی پر شده است. با پرکردن ناحیهٔ زیر منحنی در مقایسه با حالتی که فقط نمودار خطی رسم شود، بر روند تغییرات زمانی داده‌ها تاکید بیشتری می‌شود (نمودار ۱۳–۳). منبع داده: Jordan Anaya



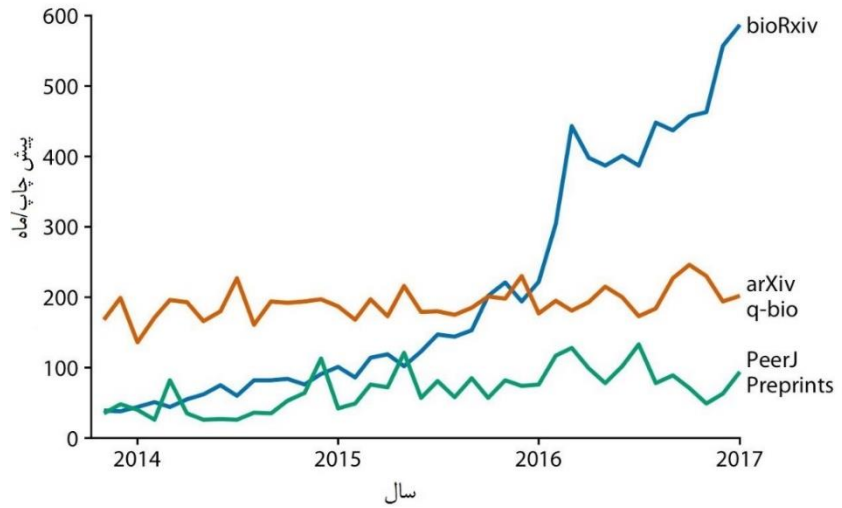
نمودار ۱۳–۵. ارسال ماهانهٔ مقالات به سه سرور پیش‌چاپ که تحقیقات زیست‌پزشکی را پوشش می‌دهند: bioRxiv، بخش q-bio از مجموعه arXiv و PeerJ Preprints. هر نقطه نشان‌دهندهٔ تعداد مقالات ارسال شده در ماه به سرور پیش‌چاپ مربوطه است. این نمودار به صورت «بد» برچسب خورده است زیرا نقاط داده در سه دورهٔ زمانی با یکدیگر تداخل دارند و خواندن آن‌ها دشوار است. منبع داده: Jordan Anaya



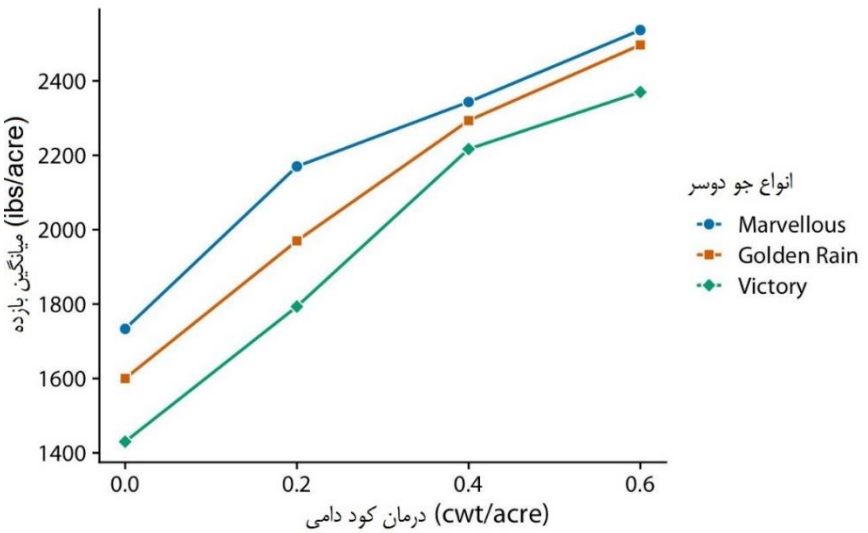
نمودار ۱۳–۶. ارسال ماهانهٔ مقالات به سه سرور پیش‌چاپ که تحقیقات زیست‌پزشکی را پوشش می‌دهند. با اتصال نقاط نمودار ۱۳–۵ توسّط خطوط، به مخاطب کمک می‌شود تا هر دورهٔ زمانی را به صورت مجزا دنبال کند.
منبع داده: Jordan Anaya

نمودار ۱۳–۶ نمایش قابل قبولی از مجموعه داده‌های پیش‌چاپ را نشان می‌دهد. با این حال، ارائه راهنمای نمودار به صورت جداگانه بار شناختی غیر ضروری برای مخاطب ایجاد می‌کند. می‌توان این بار شناختی را با برچسب‌گذاری مستقیم خطوط کاهش داد (نمودار ۱۳–۷). همچنین نقاط انفرادی داده‌ها در این نمودار حذف شده است که در نتیجه نسبت به نمودار اصلی ۱۳–۵ روان‌تر شده و خواندن آن نیز آسان‌تر شده است.

استفاده از نمودارهای خطی به سری‌های زمانی محدود نمی‌شود. هر زمان که نقاط داده، دارای نظم طبیعی باشند، که توسط متغیّر نمایش داده شده در محور افقی منعکس می‌شود، می‌توان نقاط همسایه را با یک خط به یکدیگر متصل کرد. این وضعیّت برای مثال در منحنی‌های دوز–پاسخ، وجود دارد که در آن نحوهٔ تغییر پیامد مورد نظر (پاسخ) بر اساس تغییرات یک پارامتر عددی (دوز) بررسی می‌شود. نمودار ۱۳–۸ یک آزمایش کلاسیک از این نوع را نشان می‌دهد که برداشت جو را در پاسخ به افزایش کوددهی اندازه‌گیری می‌کند. نمودار خطی نشان می‌دهد که منحنی‌های دوز–پاسخ برای سه نوع جو مدنظر، مشابه هستند امّا در نقطه شروع و در غیاب کوددهی، متفاوت می‌باشند (به بیان دیگر، برخی از گونه‌ها در حالت طبیعی محصول بیشتری نسبت به بقیه تولید می‌کنند).



نمودار ۱۳–۷. ارسال ماهانهٔ مقالات به سه سرور پیش‌چاپ که تحقیقات زیست‌پزشکی را پوشش می‌دهد. برچسب زدن مستقیم خطوط به جای ارائه راهنما، بار شناختی را کاهش می‌دهد و حذف راهنما، نیاز به استفاده از شکل‌های مختلف برای نقاط داده را از بین می‌برد. با استفاده از این روش نمودار ۱۳–۶ با حذف نقاط داده ساده‌تر می‌شود. منبع داده: *Jordan Anaya*

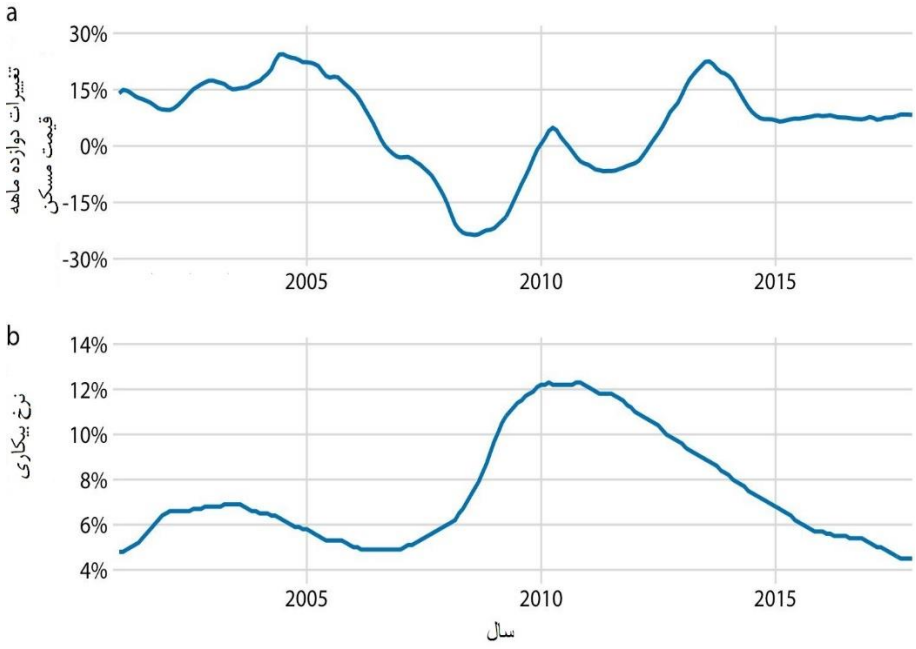


نمودار ۱۳–۸. منحنی دوز–پاسخ که میانگین رشد گونه‌های مختلف جو را پس از کوددهی نشان می‌دهد. کود به عنوان منبع نیتروژن عمل می‌کند و به طور کلی صرف نظر از گونه جو، رشد آن با حضور نیتروژن بیشتر می‌شود. در اینجا، کود در واحد cwt (صد وزن) در هر هکتار اندازه‌گیری شده است. صد وزن یک واحد سلطنتی قدیمی است که برابر با ۱۱۲ پوند یا ۵۰٫۸ کیلوگرم است. منبع داده: Yates 1935

سری زمانی متغیّرهای دو یا چند پاسخی

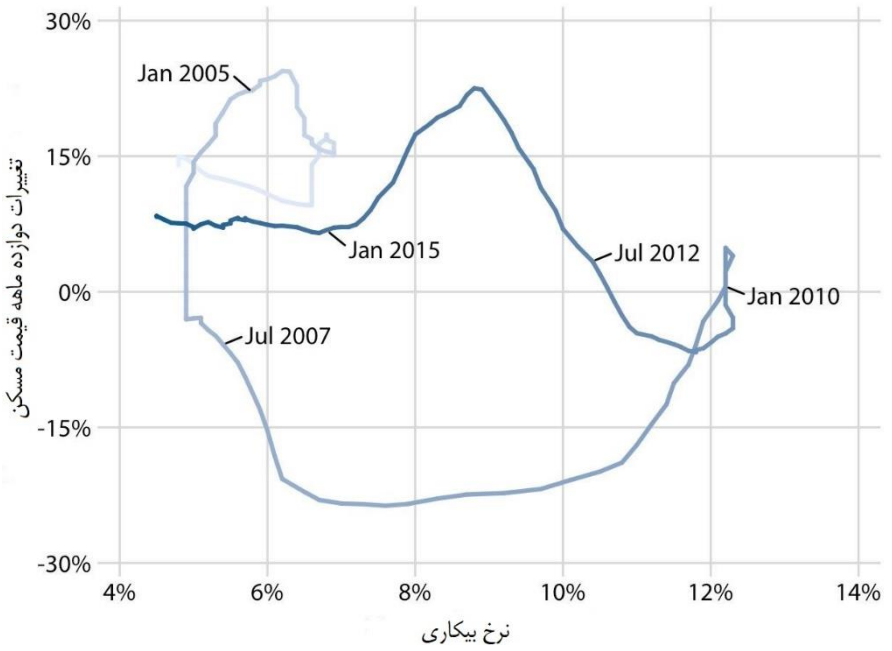
در مثال‌های قبلی، در خصوص دوره‌های زمانی یک متغیّر تک پاسخ صحبت شد. (به عنوان مثال، ارسال مقالات به سرور پیش‌چاپ در هر ماه یا برداشت جو). با این حال، غیر معمول نیست که متغیّرهایی با بیش از یک پاسخ وجود داشته باشند. چنین شرایطی معمولاً در اقتصاد کلان رخ می‌دهد. به عنوان مثال، ممکن است هدف بررسی تغییرات قیمت مسکن از ۱۲ ماه قبل و ارتباط آن با نرخ بیکاری باشد. ممکن است تصوّر شود زمانی که نرخ بیکاری پایین است قیمت مسکن افزایش می‌یابد و بالعکس.

با روش‌های مطرح شده در بخش‌های قبل، می‌توان داده‌ها را به صورت دو نمودار مجزا که بالای هم قرار گرفته‌اند، نشان داد (نمودار ۱۳–۹). این نمودار به طور مستقیم دو متغیّر دلخواه را نشان می‌دهد و تفسیر آن ساده است. با این حال، از آنجایی که دو متغیّر به عنوان نمودارهای خطی جداگانه نشان داده شده‌اند، مقایسه آن‌ها دشوار است. اگر هدف شناسایی نقاط زمانی‌ای باشد که دو متغیّر در یک جهت یا خلاف جهت هم حرکت می‌کنند، باید مرتباً بین این دو نمودار مقایسه‌های متعددی انجام شده و شیب نسبی دو منحنی مقایسه شود.



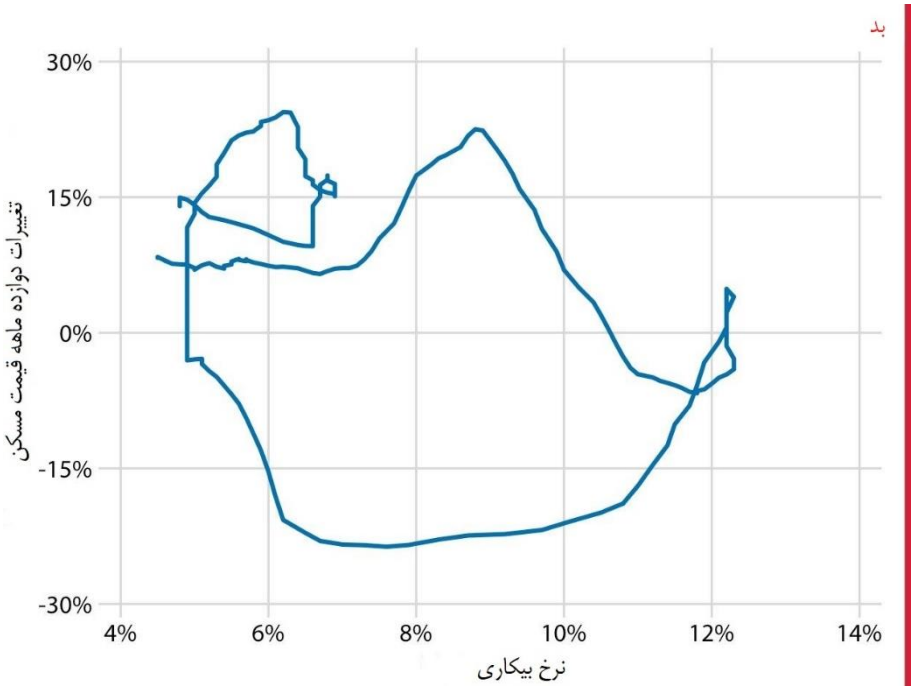
نمودار ۱۳–۹. تغییر دوازده ماهه (الف) قیمت مسکن و (ب) نرخ بیکاری از ژانویه ۲۰۰۱ تا دسامبر ۲۰۱۷. منابع داده : فهرست قیمت مسکن فردی مک، دفتر آمار کار ایالات متحده.

به عنوان جایگزینی برای نشان دادن دو نمودار خطی مجزا، می‌توان دو متغیّر را در برابر یکدیگر رسم نمود. به این منظور مسیری ترسیم می‌شود که از اوّلین نقطه زمانی به آخرین نقطه منتهی خواهد شد (نمودار ۱۳–۱۰). چنین نموداری، نمودار پراکنش متّصل نامیده می‌شود، زیرا در حقیقت یک نمودار پراکنش از دو متغیّر مجزا در مقابل یکدیگر رسم شده و سپس نقاط مجاور به هم متّصل شده است. فیزیک دانان و مهندسان اغلب این نمودار را رخسارهٔ فاز[1] می‌نامند. زیرا در رشتهٔ آن‌ها، از این نمودار معمولاً برای نشان دادن حرکت در فضای فازی استفاده می‌شود. قبلاً در فصل ۳ در خصوص نمودارهای پراکنش متّصل صحبت شده بود، در آنجا دمای روزانه در هیوستون، تگزاس، در مقابل دمای سن دیگو، کالیفرنیا ترسیم گردید (نمودار ۳–۳).



نمودار ۱۳–۱۰. تغییر دوازده ماههٔ قیمت مسکن در مقابل نرخ بیکاری، از ژانویه ۲۰۰۱ تا دسامبر ۲۰۱۷، که به صورت نمودار پراکنش متّصل نشان داده شده است. خطوط تیره‌تر نشان‌دهندهٔ ماه‌های اخیر است. همبستگی معکوس بین تغییر در قیمت مسکن و نرخ بیکاری که در نمودار ۱۳–۹ دیده شد، منجر به ایجاد دو دایره در خلاف جهت عقربه‌های ساعت در نمودار پراکنش متّصل گردیده است. منابع داده: شاخص قیمت مسکن فردی مک، دفتر آمار کار ایالات متحده

1. phase portrait



نمودار ۱۳–۱۱. تغییر دوازده ماههٔ قیمت مسکن در مقابل نرخ بیکاری از ژانویه ۲۰۰۱ تا دسامبر ۲۰۱۷. این یک نمودار به عنوان «بد» نامگذاری شده است زیرا بدون وجود نشانگرهای تاریخ و سایه رنگی نمودار ۱۳–۱۰، نه جهت و نه سرعت تغییر داده‌ها قابل درک نیست. منابع داده: فهرست قیمت خانه فردی مک، دفترآمارکار ایالات متحده

در یک نمودار پراکنش متّصل، حرکت خطوط از ربع پایین- چپ به ربع بالا-راست نشان‌دهندهٔ وجود همبستگی بین دو متغیّر است (وقتی یک متغیّر رشد می‌کند، دیگری نیز رشد می‌کند). از سوی دیگر حرکت خطوط از ربع بالا-چپ به ربع پایین-راست نشان‌دهندهٔ همبستگی معکوس می‌باشد (با رشد یک متغیّر، دیگری کاهش می‌یابد). اگر دو متغیّر رابطه چرخه‌ای داشته باشند، دایره‌ها یا مارپیچ‌هایی در نمودار پراکنش متصّل ظاهر خواهد شد. در نمودار ۱۳–۱۰، یک دایرهٔ کوچک از سال ۲۰۰۱ تا ۲۰۰۵ و یک دایرهٔ بزرگ برای بقیهٔ دورهٔ زمانی می‌بینیم.

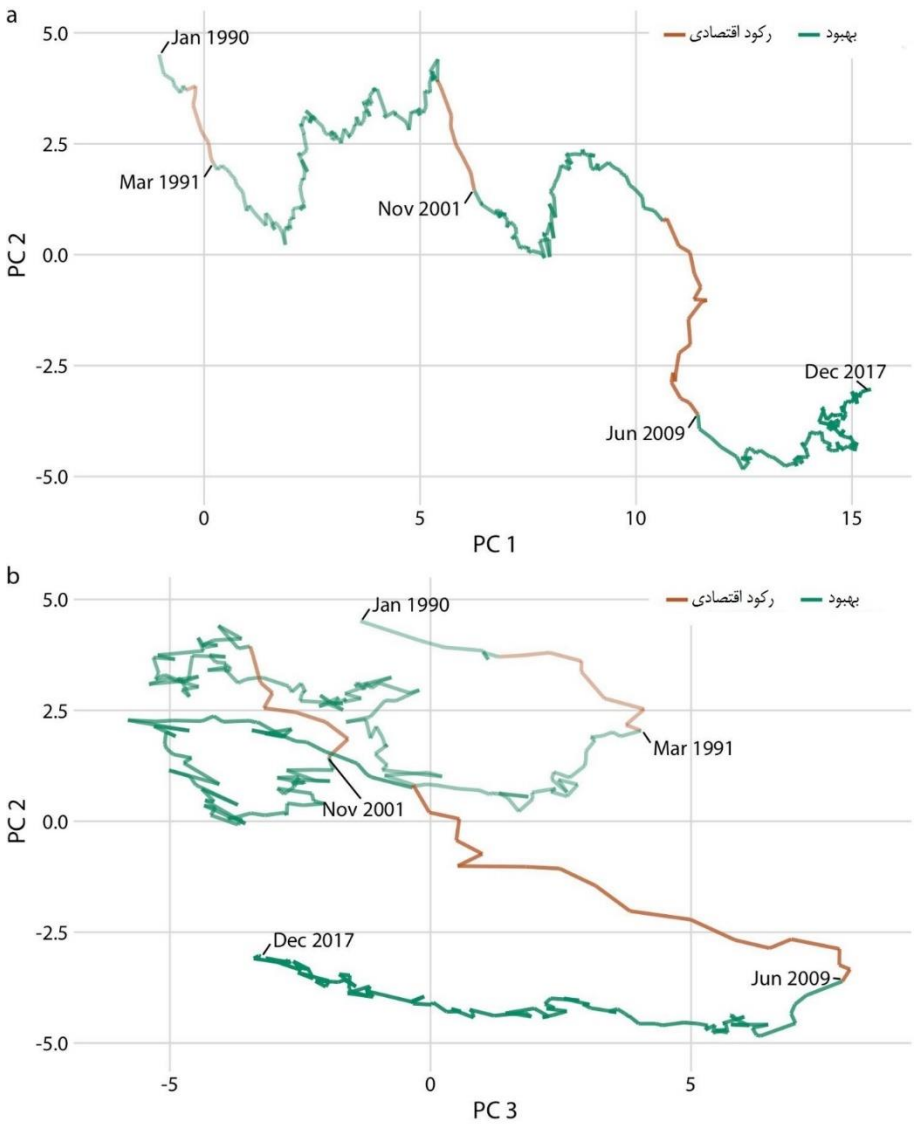
هنگام ترسیم نمودار پراکنش متّصل، مهم است که هم جهت و هم مقیاس زمانی داده‌ها مشخّص شود. بدون ذکر چنین نکاتی، نمودار به یک خط خطی بی معنی تبدیل می‌شود (نمودار ۱۳–۱۱). در نمودار ۱۳–۱۰ از تیره شدن تدریجی برای نشان دادن جهت استفاده شده و به عنوان یک روش جایگزین، می‌توان فلش‌هایی را در امتداد مسیر رسم کرد.

آیا بهتر است از یک نمودار پراکنش متّصل استفاده شود یا از دو نمودار خطی جداگانه؟ نمودارهای خطی مجزا معمولاً راحت‌تر خوانده می‌شوند، امّا زمانی که مردم به نمودارهای پراکنش متّصل عادت کنند، ممکن است بتوانند الگوهای خاصی را استخراج کنند (مانند رفتار چرخه‌ای با مقداری بی‌نظمی) که تشخیص آن در نمودارهای خطی دشوار است. در واقع، تشخیص رابطه چرخه‌ای بین تغییر قیمت مسکن و نرخ بیکاری در نمودار ۱۳–۹ دشوار است امّا مارپیچ خلاف جهت عقربه‌های ساعت در نمودار ۱۳–۱۰ آن را نشان می‌دهد. پژوهش‌ها نشان می‌دهد که احتمال اشتباه خوانندگان در تعیین ترتیب و جهت نمودار پراکنش متّصل، نسبت به نمودارهای خطی بیشتر و احتمال گزارش همبستگی کمتر است. (Haroz, Kosara, and Franconeri 2016) از طرف دیگر، به نظر می‌رسد نمودارهای پراکنش متّصل منجر به تعامل بیشتری شده و لذا چنین نمودارهایی ممکن است ابزارهای مؤثرتری برای جلب خوانندگان به سمت مساله باشند.

هرچند نمودارهای پراکنش متّصل می‌توانند دو متغیّر را در یک زمان نشان دهند، می‌توان از آن‌ها برای نمایش مجموعه داده‌های با ابعاد بالاتر استفاده نمود. ترفندی که باید به کار برد کاهش بُعد است (به فصل ۱۲ مراجعه کنید). سپس می‌توان یک نمودار پراکنش متّصل در فضای کاهش بُعد یافته رسم نمود. به عنوان مثالی برای این رویکرد، یک مجموعه داده از مشاهدات ماهانه بیش از ۱۰۰ شاخص کلان اقتصادی که توسّط بانک فدرال رزرو سنت لوئیس ارائه شده، نمایش داده خواهد شد. ابتدا تحلیل اجزای اصلی برای همه شاخص‌ها انجام شده و سپس نمودار پراکنش متصل برای PC2 در مقابل PC1 (نمودار۱۳–۱۲ الف) و PC2 را در مقابل PC3 (نمودار ۱۳–۱۲ ب) رسم می‌شود.

جالب است که نمودار ۱۳–۱۲ الف تقریباً شبیه یک نمودار خطی معمولی است که سیر زمانی از چپ به راست دارد. این الگو توسّط یک خصوصیّت عمومی تحلیل PCA ایجاد می‌شود: جزء اوّل اغلب اندازهٔ کلی سیستم را اندازه‌گیری می‌کند. در اینجا، PC1 تقریباً اندازهٔ کلی اقتصاد را اندازه‌گیری می‌کند که به ندرت در طول زمان کاهش می‌یابد.

با رنگ‌آمیزی نمودار پراکنش متّصل به صورت رکود و بازیابی، می‌توان مشاهده نمود که رکودها با کاهش PC2 همراه است در حالی که بازیابی‌ها با ویژگی خاصی در PC1 یا PC2 مرتبط نیستند (نمودار ۱۳–۱۲ الف). با این حال، به نظر می‌رسد که بازیابی‌ها با کاهش PC3 مطابقت دارند (نمودار ۱۳–۱۲ ب). علاوه بر این، در نمودار PC2 در مقابل PC3، می‌توان دید که نمودار الگوی حرکت مارپیچی در جهت عقربه‌های ساعت دارد. این الگو بر ماهیت چرخه‌ای اقتصاد با رکودها و بازیابی‌های متناوب تاکید می‌کند.



نمودار ۱۳–۱۲. نمایش یک سری زمانی با ابعاد بالا به صورت نمودار پراکنش متّصل در فضای اجزای اصلی. این مسیر، حرکت بیش از ۱۰۰ شاخص کلان اقتصادی را از ژانویه ۱۹۹۰ تا دسامبر ۲۰۱۷ نشان می‌دهد. دوره‌های رکود و بازیابی از طریق رنگ و نقاط پایانی سه رکود (مارس ۱۹۹۱، نوامبر ۲۰۰۱ و ژوئن ۲۰۰۹) از طریق برچسب‌گذاری مشخّص شده‌اند. الف) PC2 در مقابل PC1 و ب) PC2 در مقابل PC3.

منبع داده.M. W. McCracken, St. Louis Fed

فصل ۱۴

نمایش روندها

هنگام ترسیم نمودارهای پراکنش (فصل ۱۲) یا سری‌های زمانی (فصل ۱۳)، اغلب تمایل بیشتری به بررسی روند کلی داده‌ها وجود دارد تا جزئیات خاص مربوط به تک تک نقاط. با ترسیم روند بر روی یا به جای نقاط دادهٔ واقعی که معمولاً به صورت خط مستقیم یا منحنی می‌باشد، می‌توان نموداری ارائه کرد که خواننده فوراً متوجّه ویژگی‌های کلیدی داده‌ها شود. دو رویکرد اساسی برای تعیین روند وجود دارد: می‌توان تغییرات داده‌ها را با برخی روش‌ها هموار نمود، مانند میانگین متحرّک، یا اینکه می‌توان یک منحنی را بر اساس یک تابع مشخّص برازش نمود. هنگامی که روندی در یک مجموعه داده شناسایی شد، ممکن است بررسی انحراف‌ها از آن روند مهم باشد، یا داده‌ها را به چندین مؤلفه، از جمله روند زمینه‌ای، چرخه‌های موجود، و اجزای فرعی یا خطای تصادفی جدا نمود.

هموارسازی

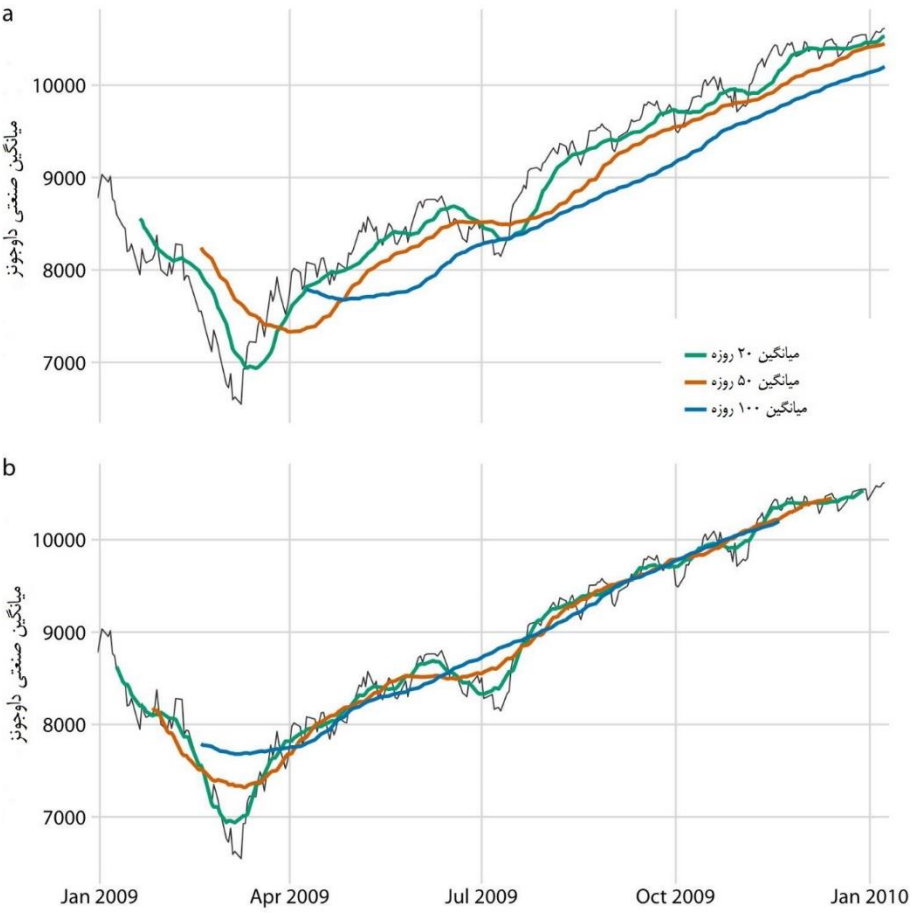
یک سری زمانی از میانگین صنعتی داوجونز[1] که یک شاخص بازار سهام و نشان‌دهندهٔ قیمت ۳۰ واحد بزرگ عمومی دولت ایالات متحده است را در نظر بگیرید. به طور خاص به سال ۲۰۰۹، درست پس از بحران ۲۰۰۸ نگاه خواهیم کرد (نمودار ۱۴–۱). در اواخر بحران، در ۳ ماه اول سال ۲۰۰۹، بازار بیش از ۲۴۰۰ امتیاز (۲۷ درصد) از دست داد. سپس به آرامی تا انتهای سال بهبود یافت. چگونه می‌توان این روندهای بلندمدت را بدون تاکید بر نوسانات کوتاه مدّت ترسیم نمود؟

1. Dow Jones

نمودار ۱۴–۱. مقادیر روزانه میانگین صنعتی داوجونز در سال ۲۰۰۹.

برای این منظور باید از نظر آماری به دنبال راهی برای هموارسازی سری زمانی بورس بود. عمل هموارسازی تابعی را تولید می‌کند که الگوهای کلیدی را در داده‌ها استخراج کرده در حالی که جزئیات کم اهمیت غیر مرتبط یا خطاهای تصادفی را حذف می‌کند. تحلیلگران اقتصادی معمولاً داده‌های بازار سهام را با محاسبهٔ میانگین متحرّک هموارسازی می‌کنند. برای تولید میانگین متحرّک، بایستی یک بازهٔ زمانی در نظر گرفت، مثلاً ۲۰ روز اول سری زمانی، سپس میانگین قیمت در این ۲۰ روز محاسبه شده و بعد پنجرهٔ زمانی یک روز به جلو حرکت می‌کند یعنی بین روز دوّم تا بیست و یکم. سپس میانگین این ۲۰ روز محاسبه خواهد شد و مجدداً پنجرهٔ زمانی یک روز به جلو حرکت کرده و این فرآیند تکرار می‌شود. نتیجه حاصل یک سری زمانی جدید متشکل از دنباله‌ای از میانگین قیمت‌ها است.

برای ترسیم این توالی از میانگین‌های متحرّک، باید تصمیم گرفت که کدام نقطهٔ زمانی خاص را بایستی با میانگین هر پنجرهٔ زمانی مرتبط نمود. تحلیلگران اقتصادی اغلب هر میانگین را با نقطهٔ پایانی پنجرهٔ زمانی مرتبط می‌نمایند. این انتخاب منجر به ایجاد منحنی‌هایی می‌شود که نسبت به داده‌های اصلی با تاخیر آغاز می‌شوند (نمودار ۱۴–۲ الف)، و هرچه پنجرهٔ زمانی بزرگتر باشد این تاخیر بیشتر خواهد بود. از سوی دیگر متخصصین آمار، میانگین را در مرکز پنجرهٔ زمانی رسم می‌کنند، که در نتیجه منحنی‌ای ایجاد می‌شود که به خوبی روی داده‌های اصلی قرار می‌گیرد (نمودار ۱۴–۲ ب).



نمودار ۱۴–۲. مقادیر روزانهٔ میانگین صنعتی داوجونز در سال ۲۰۰۹، که به همراه میانگین متحرّک ۲۰، ۵۰ و ۱۰۰ روزه نشان داده شده است. (الف) میانگین متحرّک در انتهای پنجره‌های زمانی رسم شده است ب) میانگین متحرّک در مرکز پنجره‌های زمانی رسم شده است.

صرف نظر از اینکه سری زمانی هموار شده با یا بدون تاخیر رسم شود، این طول پنجرهٔ زمانی است که تعیین کنندهٔ نوساناتی می‌باشد که در منحنی صاف شده قابل مشاهده باقی می‌مانند. میانگین متحرّک ۲۰ روزه افزایش‌های کوچک و کوتاه مدّت را حذف می‌کند امّا الگوی داده‌های روزانه را به خوبی دنبال می‌کند. از طرف دیگر میانگین متحرّک ۱۰۰ روزه حتی افت‌های نسبتاً قابل توجّه و افزایش‌هایی که در یک بازهٔ زمانی چند هفته‌ای وجود دارند را نیز حذف می‌کند. به عنوان مثال، کاهش شدید به ۷۰۰۰ امتیاز در سه ماههٔ اول ۲۰۰۹ در میانگین متحرّک ۱۰۰ روز قابل مشاهده نیست و آن را با یک منحنی هموار جایگزین نموده که خیلی

پایین‌تر از ۸۰۰۰ امتیاز نمی‌رود (نمودار ۱۴-۲) به طور مشابه، افت موجود در جولای ۲۰۰۹ در میانگین متحرّک ۱۰۰ روزه کاملا ناپدید شده است.

میانگین متحرّک ساده‌ترین رویکرد برای هموارسازی است و برخی محدودیت‌های واضح نیز دارد. اوّل، منجر به ترسیم منحنی هموواری می‌شود که کوتاه‌تر از منحنی اصلی است (نمودار ۱۴-۲). قسمت‌هایی در ابتدا یا انتها یا هر دو وجود ندارند. هر چه سری زمانی بیشتر هموار شود (یعنی پنجره میانگین بزرگتر باشد)، منحنی هموار شده کوتاه‌تر می‌گردد. دوّم، حتی با یک پنجرهٔ میانگین بزرگ، میانگین متحرّک لزوماً خیلی هموار نخواهد بود. هرچند که هموارسازی در مقیاس بزرگی حاصل شده است با این حال منحنی ممکن است ارتعاشات کوچکی از خود نشان دهد (نمودار ۱۴-۲). این ارتعاشات ناشی از نقاط داده‌ای است که به پنجرهٔ میانگین وارد یا خارج می‌شوند. از آنجایی که همهٔ نقاط داده در پنجره زمانی به طور مساوی وزن‌دهی می‌شوند، نقاط دادهٔ پرت می‌توانند تأثیر قابل توجّهی بر میانگین داشته باشند.

متخصصین آمار رویکردهای متعددی را برای هموارسازی طراحی کرده‌اند که باعث کاهش نکات منفی میانگین متحرّک می‌شود. این رویکردها بسیار پیچیده و نیازمند محاسبات سنگین هستند امّا در نرم‌افزارهای آماری مدرن امروزی به راحتی در دسترس قرار دارند. یکی از روش‌های پرکاربرد، هموارسازی نمودار پراکنش با تخمین محلی[1] (LOESS) است [Cleveland 1979]، که چند جمله‌ای‌های درجهٔ پایین را به زیر مجموعه‌ای از داده‌ها برازش می‌کند. نکتهٔ مهم این است که نقاط مرکز هر زیر مجموعه وزن بیشتری نسبت به مرزهای آن خواهند گرفت. این الگوی وزن‌دهی نتیجهٔ بسیار صاف‌تری نسبت به هموارسازی مبتنی بر میانگین وزنی به همراه دارد. منحنی LOESS نشان داده شده در نمودار ۱۴-۳ مشابه میانگین ۱۰۰ روزه در نمودار ۱۴-۲ به نظر می‌رسد، امّا این شباهت نباید بیش از حد تفسیر شود. صافی یک منحنی LOESS با تغییر یک پارامتر قابل تنظیم است و انتخاب پارامترهای مختلف می‌تواند منحنی‌های LOESS شبیه میانگین ۲۰ روز یا ۵۰ روزه ایجاد نماید.

باید توجّه نمود که LOESS به سری‌های زمانی محدود نمی‌شود. می‌توان آن را برای هر نمودار پراکنشی استفاده نمود، همانگونه که از نام آن مشخّص است: هموارسازی نمودار پراکنش با تخمین محلی. برای مثال، می‌توان از LOESS برای بررسی روند رابطه بین ظرفیّت مخزن سوخت خودرو و قیمت آن (نمودار ۱۴-۴) استفاده نمود. خط LOESS نشان می‌دهد که ظرفیت مخزن سوخت تقریباً به صورت خطی با قیمت در خودروهای ارزان قیمت

1. locally estimated scatterplot smoothing

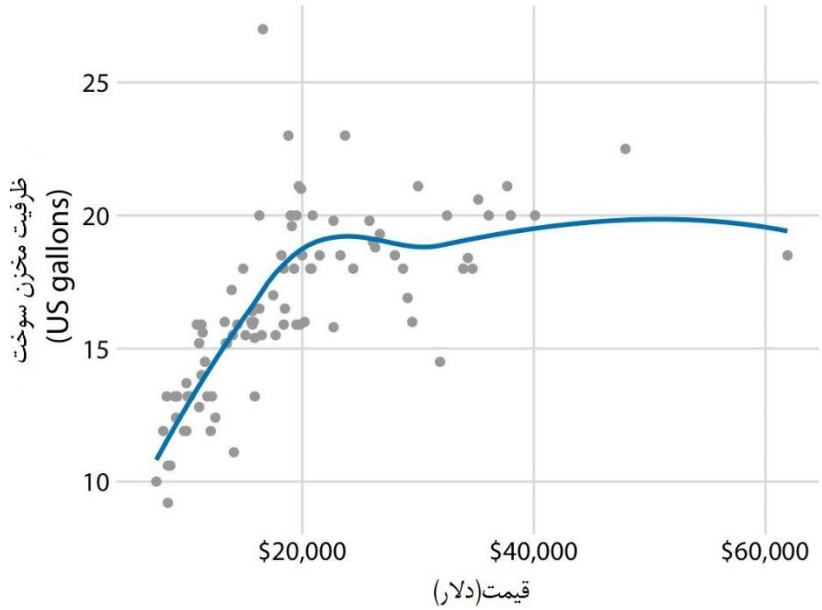
(زیـر ۲۰۰۰۰ دلار) افـزایش می‌یابـد امّـا بـرای خودروهـای گران‌تـر ایـن ارتبـاط خطـی ضـعیف‌تر می‌شـود. بـالاتر از قیمـت تقریبـاً ۲۰۰۰۰ دلار، بـا خریـد خـودروی گـران قیمـت‌تر، مخـزن سـوخت بزرگ‌تری نخواهید داشت.

نمودار ۱۴–۳. مقایسهٔ برازش LOESS با میانگین متحرّک ۱۰۰ روزه برای داده‌های نمودار ۱۴–۲ داو جونز. روند کلی نشان داده شده توسّط LOESS تقریباً با میانگین متحرّک ۱۰۰ روزه یکسان است، امّا منحنی LOESS بسیار هموارتر است و دامنهٔ کامل داده‌ها را پوشش می‌دهد منبع داده: Yahoo finance!

LOESS یـک رویکـرد هموارسـازی بسـیار محبـوب اسـت زیـرا تمایـل بـه ایجـاد نتـایجی دارد کـه برای چشم انسان مناسب است. با ایـن حـال، نیـاز بـه بـرازش مـدل‌های متعـدد رگرسـیونی دارد. ایـن امـر باعـث کنـدی محاسـبهٔ فـوق بـرای مجموعـه داده‌هـای بـزرگ، حتـی بـا اسـتفاده از تجهیـزات محاسباتی مدرن می‌شود.

به عنـوان جـایگزین سـریعتری بـرای LOESS، می‌تـوان از مـدل‌های تکه‌بنـد[1] اسـتفاده نمـود. مـدل تکه‌بند یک تابع چند جمله‌ای بسیار انعطاف‌پـذیر اسـت امّـا همیشـه صـاف بـه نظـر می‌رسـد. هنگـام کار با مدل تکه‌بند بـا اصـطلاح گـره مواجـه خـواهیم شـد. گره‌هـای یـک مـدل تکه‌بنـد نقطـهٔ پایـانی تک‌تـک بخش‌هـای تکه‌بنـد هسـتند. اگـر یـک مـدل تکه‌بنـد را بـا k قطعـه بـرازش داده شـود، بایـد k + 1 گره مشخّص شـود. در حـالی کـه بـرازش مـدل تکه‌بنـد از نظـر محاسـباتی کارآمـد اسـت، بـه ویژه اگر تعداد گره‌ها خیلی‌زیاد نباشـد، بـا این‌حـال مـدل‌های تکه‌بنـد نیـز نقـاط ضـعف خـود را دارنـد.

1. spline models(یک تابع را توسط تکه های چندجمله ای به صورت چندضابطه ای تقریب زدن–مترجم)



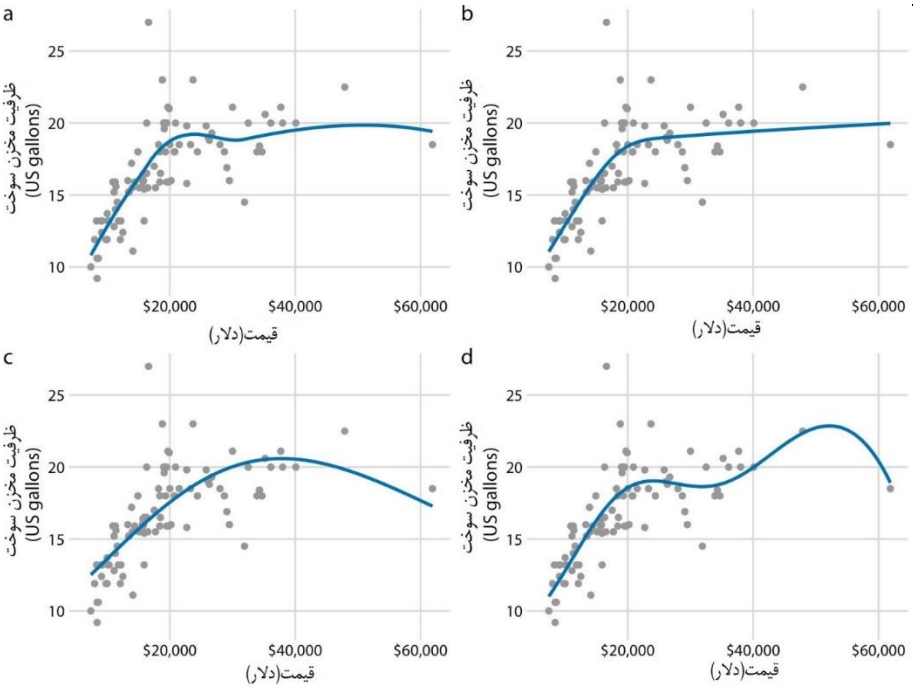
نمودار ۱۴–۴. ظرفیت مخزن سوخت در مقابل قیمت ۹۳ خودرو با مدل سال ۱۹۹۳. هر نقطه مربوط به یک خودرو است. خط توپر نشان‌دهندهٔ برازش LOESS بر داده‌ها است. همانطور که مشاهده می‌شود ظرفیت مخزن سوخت تقریباً به صورت خطی با قیمت افزایش می‌یابد، امّا از قیمت تقریباً ۲۰۰۰۰ دلار تقریباً مسطح می‌شود. منبع داده : Robin H. Lock, St. Lawrence University.

مهم‌ترین نکته منفی مدل تکه‌بند، وجود تعداد متنوعی از آنهاست از جمله مکعبی، نوع ب، صفحه نازک، فرآیند گاوسی و غیره. لذا انتخاب یکی از آن‌ها ممکن است آسان نباشد. انتخاب نوع مدل تکه‌بند و تعداد گره‌های مورد استفاده می‌تواند منجر به نتایج متفاوتی بر روی داده‌های یکسان شود (نمودار ۱۴–۵).

بیشتر نرم‌افزارهای نمایش داده‌ها ابزارهای مخصوص هموارسازی را ارائه می‌کنند که احتمالاً به صورت یک نوع رگرسیون محلی (مانند LOESS) یا یک نوع مدل تکه‌بند پیاده‌سازی شده‌اند. روش هموارسازی ممکن است به عنوان یک مدل افزایشی تعمیم یافته[1] (GAM) نامیده شود که یک ابر مجموعه در بین همهٔ این نوع هموارسازها می‌باشد. باید توجّه نمود که خروجی هموارسازی به مدل خاص GAM که روی داده‌ها برازش شده بستگی دارد. بدون امتحان تعدادی از گزینه‌های مختلف، ممکن است هرگز متوجّه نشوید که تا چه حد نتیجهٔ شما بستگی به انتخاب پیش فرض‌های خاصی دارد که توسّط نرم‌افزار آماری انتخاب شده است.

1. generalized additive model

هنگام تفسیر نتایج یک تابع هموارسازی باید مراقب باشید زیرا یک مجموعهٔ داده مشابه را می‌توان به روش‌های مختلف هموار کرد.

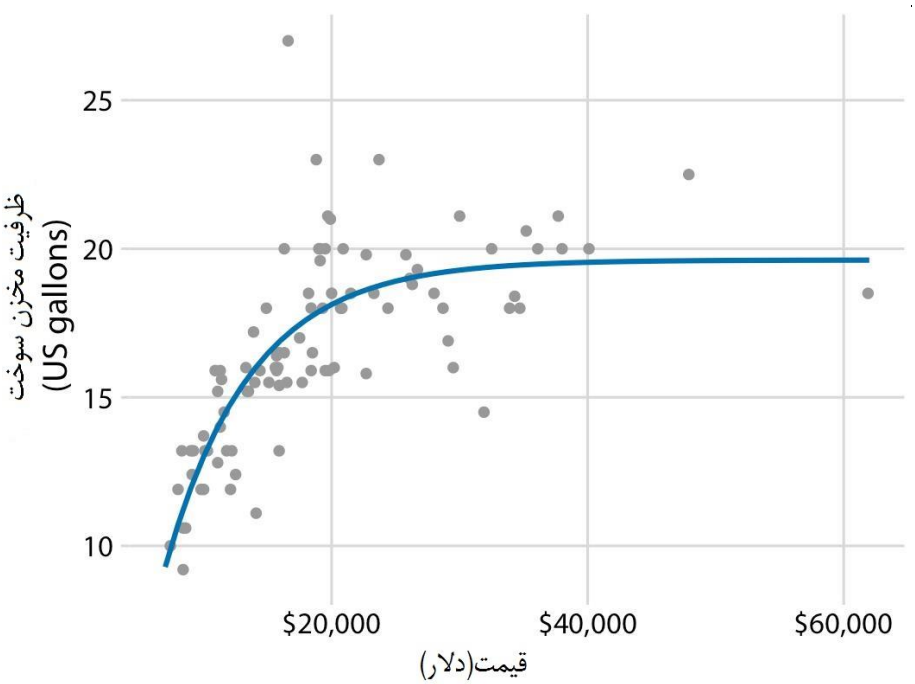


نمودار ۱۴–۵. مدل‌های مختلف هموارسازی، رفتارهای بسیار متفاوتی را به ویژه در خصوص نقاط پرت نشان می‌دهند (الف) LOESS هموارتر مانند نمودار ۱۴–۴ (ب) تکه‌بند رگرسیون مکعبی با ۵ گره. (ج) تکه‌بند رگرسیون صفحه نازک با ۳ گره. (د) تکه‌بند پردازش گوسی با ۶ گره. منبع داده: Robin H. Lock, St. Lawrence University

ترسیم روند با یک تابع تعریف شده

همانطور که در نمودار ۱۴–۵ قابل مشاهده است، رفتار هموارکننده‌های عمومی می‌تواند برای هر مجموعهٔ داده تا حدودی غیر قابل پیش‌بینی باشد. این هموارکننده‌ها تخمینی از پارامترها به طوری که تفسیر معناداری داشته باشند، ارائه نمی‌دهند. بنابراین تا حد ممکن ارجح است که منحنی با تابع عملکردی خاص که برای داده‌ها مناسب بوده و از پارامترهایی با معنای واضح بهره می‌برد، پرازش شود.

برای داده‌های مخزن سوخت، منحنی‌ای نیاز است که ابتدا به صورت خطی بالا رفته، سپس در یک مقدار مشخّص ثابت شود. تابع $y = A - B\exp(-mx)$ ممکن است با الگو مطابقت داشته باشد. در اینجا، A، B، و m مقادیری هستند که برای برازش مناسب منحنی با داده‌ها تنظیم می‌شوند. تابع تقریباً برای مقادیر کوچک x خطی است و مقدار آن تقریباً برابر $y \approx A - B + Bmx$ است با $y \approx A$، امّا برای مقادیر بزرگ x، مقدار آن تقریباً برابر A، خواهد بود و به ازای مقادیر مختلف x رشد بسیار اندکی دارد. نمودار ۱۴–۶ نشان می‌دهد که این معادله حداقل به اندازه هموارکننده‌هایی که قبلاً بحث شد، بر داده‌ها به خوبی برازش می‌شود (نمودار ۱۴–۵).

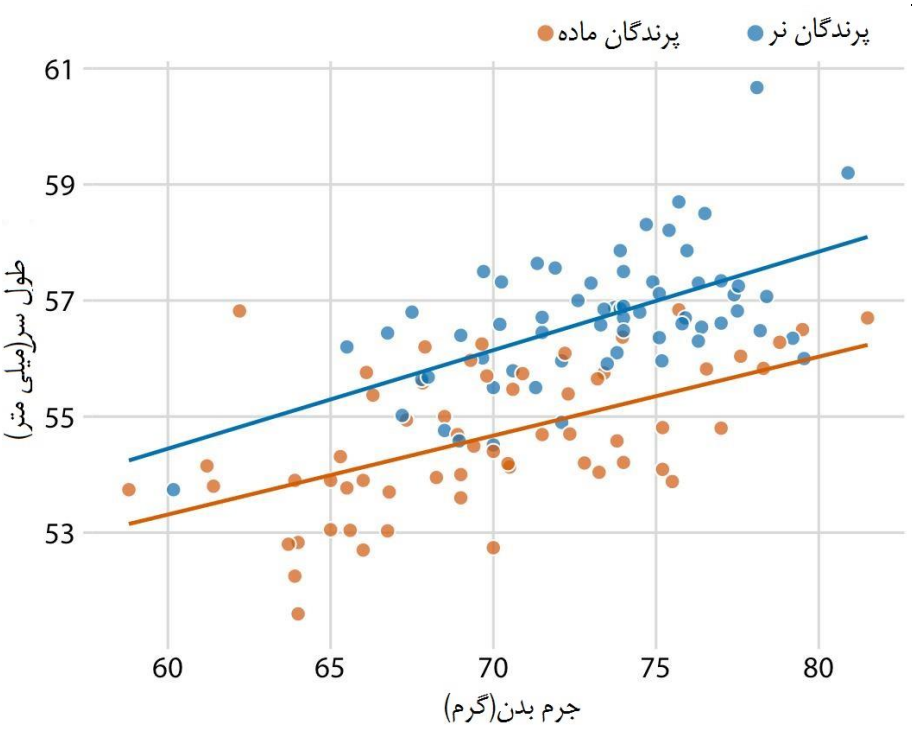


نمودار ۱۴–۶. داده‌های مخزن سوخت با یک مدل تحلیلی صریح نشان داده شده است. خط توپر مربوط به برازش کمترین مربعات فرمول$y = A - B\exp(-mx)$ بر داده‌ها است. پارامترهای برازش شده $A = 19.6$، $B = 29.2$، $m = 0.00015$ هستند. منبع داده Robin H. Lock, St. Lawrence University

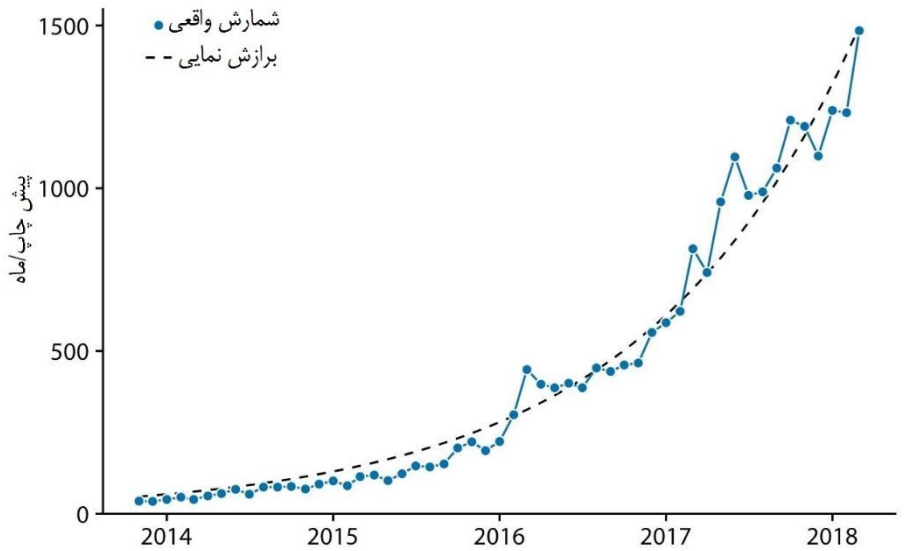
یک تابع که در بسیاری از شرایط قابل استفاده می‌باشد، خط صاف ساده $y = A + mx$ است. به صوت خارق العاده‌ای، روابط تقریباً خطی بین دو متغیّر در مجموعه داده‌های دنیای واقعی بسیار رایج است. به عنوان مثال، در فصل ۱۲، در مورد رابطهٔ بین طول سر و وزن بدن در زاغ آبی توضیح داده شد. این رابطه هم برای پرندگان ماده و هم نر، تقریباً خطی است، و رسم

خطوط روند خطی در بالای نقاط در نمودار پراکنش به خواننده کمک می‌کند تا روندها را بهتر درک کند (نمودار ۱۴–۷).

وقتی داده‌ها یک رابطه غیرخطی را نشان می‌دهند، باید حدس بزنیم که کدام تابع مناسب است. در این صورت می‌توان صحّت حدس خود را با تبدیل محورها به گونه‌ای که یک رابطه خطی پدیدار شود، ارزیابی نمود. برای نشان دادن این اصل، بیایید به ارسال ماهانه مقالات به سرور پیش‌چاپ که در فصل ۱۲ مورد بحث قرار گرفت، برگردیم (bioRxiv). اگر افزایش ارسال مقالات در هر ماه با تعداد ارسال مقالات در ماه قبل متناسب باشد، یعنی اگر ارسال مقالات در هر ماه با یک درصد ثابت رشد کنند، منحنی حاصل نمایی است. به نظر می‌رسد این فرضیه برای داده‌های bioRxiv برآورده شده است، زیرا منحنی با قالب نمایی $y = A \exp(mx)$ به خوبی با داده‌های ارسال مقالات bioRxiv مطابقت دارد (نمودار ۱۴–۸).



نمودار ۱۴–۷. طول سر در مقایسه با وزن بدن برای ۱۲۳ زاغ آبی. جنسیت پرندگان بر اساس رنگ مشخّص شده است. این نمودار معادل نمودار ۱۲–۲ است، با این تفاوت که اکنون خطوط روند روی نقاط داده رسم شده است.
منبع داده: Keith Tarvin, Oberlin College

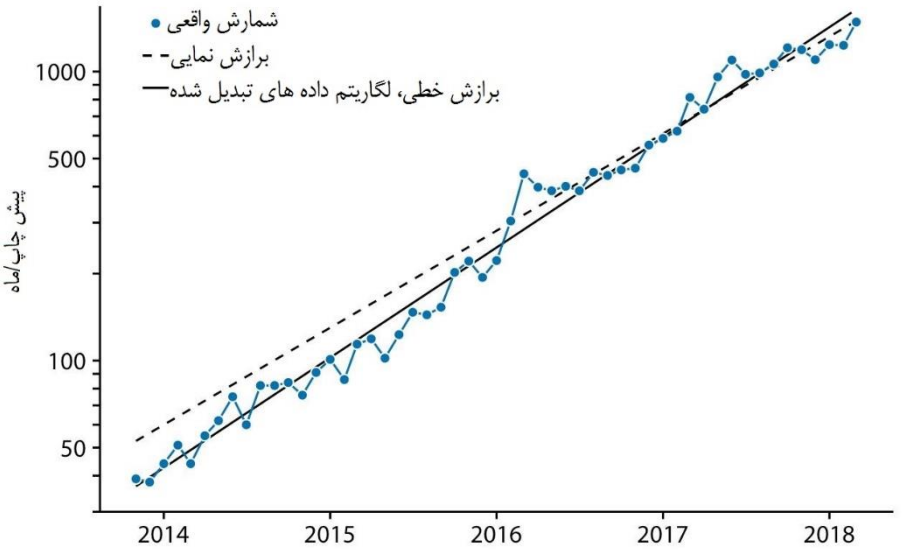


نمودار ۱۴–۸. ارسال ماهانهٔ مقالات به سرور پیش‌چاپ bioRxiv خط ممتد آبی تعداد واقعی مقالات ارسالی به سرور پیش‌چاپ را نشان می‌دهد و خط نقطه‌چین سیاه رنگ نشان‌دهندهٔ برازش نمایی $y = 60 \exp[0.77(x - 2014)]$ بر روی داده‌ها می‌باشد. منبع داده: Jordan Anaya، http://www.prepubmed.org/.

اگر منحنی اصلی نمایی باشد، یعنی $y = A \exp(mx)$، تبدیل لگاریتمی مقادیر y آن‌ها را به یک رابطهٔ خطی تبدیل می‌کند: $\log(y) = \log(A) + mx$. بنابراین، ترسیم داده‌ها با مقادیر y تبدیل لگاریتمی شده (یا به طور معادل، با محور y لگاریتمی) و جستجو برای وجود یک رابطهٔ خطی، راه مناسبی برای تعیین این است که آیا مجموعه داده رشد نمایی دارد یا خیر. در خصوص تعداد مقالات ارسالی به bioRxiv، در صورت استفاده از محور y لگاریتمی، رابطهٔ خطی قابل مشاهده است (نمودار ۱۴–۹).

در نمودار ۱۴–۹، علاوه بر تعداد واقعی ارسال مقالات، برازش نمایی از نمودار ۱۴–۸ و یک برازش خطی برای داده تبدیل لگاریتمی شده نشان داده شده است. این دو برازش مشابه هستند امّا یکسان نیستند. به طور خاص، شیب خط چین تا حدودی نامناسب به نظر می‌رسد. این خط به طور منظم در نیمی از سری زمانی بالای نقاط داده‌ها قرار می‌گیرد. این یک مشکل رایج در برازش‌های نمایی است: مجذور انحرافات نقاط داده از منحنی برازش شده برای مقادیر بزرگ داده‌ها نسبت به مقادیر کوچک، بسیار بزرگتر است. به عبارت دیگر انحرافات داده‌های کوچک سهم کمی در مجموع کلی مربعاتی دارد که مدل آن‌ها را به حداقل می‌رساند. در نتیجه، خط برازش شده به طور منظّم از کوچکترین مقادیر داده‌ها بالاتر یا

پایین‌تر می‌رود. به همین دلیل، به طور کلی توصیه می‌شود از برازش‌های نمایی اجتناب نموده و به جای آن از برازش‌های خطی در داده‌های تبدیل شده لگاریتمی استفاده گردد.



نمودار ۱۴–۹. ارسال ماهانهٔ مقالات به سرور پیش‌چاپ bioRxiv، که در مقیاس لگاریتمی نشان داده شده است. خط ممتد آبی نشان‌دهندهٔ تعداد واقعی ارسال ماهانهٔ مقالات به سرور پیش‌چاپ، خط نقطه‌چین سیاه نشان‌دهندهٔ برازش نمایی از نمودار ۱۴–۸، و خط سیاه ممتد نشان‌دهندهٔ برازش خطی بر روی داده‌های تبدیل شده لگاریتمی بر اساس $y = 43 \exp[0.88(x - 2014)]$ می‌باشد. منبع: Jordan Anaya، http://www.prepubmed.org/

معمولاً بهتر است خط مستقیمی بر داده‌های تبدیل لگاریتمی شده برازش شود تا اینکه منحنی غیرخطی بر داده‌های تبدیل نشده برازش گردد.

نموداری مانند نمودار ۱۴–۹ معمولاً به صورت لگاریتمی–خطی نامیده می‌شود، زیرا محور y لگاریتمی و محور x خطی است. نمودارهای دیگری که ممکن است با آن‌ها مواجه شویم عبارتند از لگاریتمی–لگاریتمی که در آن هر دو محور y و x لگاریتمی هستند، و خطی–لگاریتمی که در آن محور y خطی و محور x لگاریتمی است. در نمودار لگاریتمی–لگاریتمی قوانین توان در قالب $y \sim x^{\alpha}$ به صورت خطوط مستقیم ظاهر می‌شود (به نمودار ۸–۷ مراجعه شود)، و در نمودار خطی–لگاریتمی، روابط لگاریتمی در قالب $y \sim \log(x)$ به صورت خطوط مستقیم ظاهر می‌شوند. سایر توابع را می‌توان با تبدیل‌های تخصصی‌تر به روابط خطی تبدیل کرد. امّا این سه (لگاریتمی–خطی، لگاریتمی–لگاریتمی، خطی–لگاریتمی) طیف وسیعی از کاربردی‌های دنیای واقعی را پوشش می‌دهند.

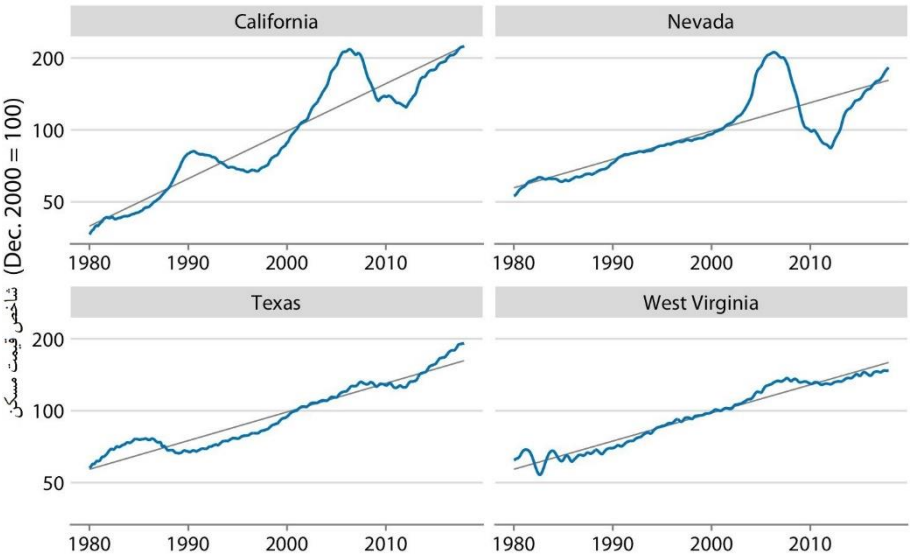
روندزدایی و تجزیهٔ سری زمانی

در هر سری زمانی با روندی بلندمدت، ممکن است حذف این روند با هدف برجسته کردن هرگونه انحراف قابل توجّه مفید باشد. این روش روندزدایی نامیده می‌شود و در ادامه با مثال قیمت مسکن بحث خواهد شد. در ایالات متحده، وام دهندهٔ مسکن «فردی مک»[1] یک شاخص ماهانه به نام شاخص قیمت خانه فردی مک منتشر می‌کند که تغییر قیمت مسکن را در طول زمان ردیابی می‌کند. این شاخص تلاش می‌کند تا وضعیّت کل بازار مسکن را در یک منطقه معیّن به تصویر بکشد، به طوری که به عنوان مثال افزایش ۱۰ درصد در شاخص را می‌توان به عنوان افزایش متوسط ۱۰ درصد قیمت مسکن در سال تفسیر کرد. این شاخص به طور قراردادی در دسامبر ۲۰۰۰ معادل ۱۰۰ در نظر گرفته شده است.

در طول دوره‌های زمانی بلند مدّت، قیمت مسکن تمایل دارد رشد سالانه ثابتی را نشان دهد که تقریباً با تورّم مطابقت دارد. با این حال، روی این روند، حباب‌های مسکن است که منجر به چرخه‌های رونق و رکود شدید در این بازار می‌شود. نمودار ۱۴–۱۰ شاخص واقعی قیمت مسکن و روند بلندمدت آن برای چهار ایالت منتخب ایالات متحده را نشان می‌دهد. همانطور که قابل مشاهده است بین سال‌های ۱۹۸۰ تا ۲۰۱۷، کالیفرنیا دچار دو حباب شد، یکی در سال ۱۹۹۰ و دیگری در اواسط دههٔ ۲۰۰۰. در همان دوره در اواسط دههٔ ۲۰۰۰، نوادا تنها یک حباب را تجربه کرد و قیمت مسکن در تگزاس و ویرجینیای غربی تقریباً به طور کامل از روند بلندمدت تبعیّت نمود. از آنجایی که قیمت مسکن تمایل به افزایش درصدی دارد، یعنی به صورت نمایی، در نمودار ۱۴–۱۰ محور لگاریتمی y انتخاب شده است. خطوط مستقیم مربوط به افزایش سالانه قیمت معادل ۴٫۷ درصد در کالیفرنیا و ۲٫۸ درصد در نوادا، تگزاس، و ویرجینیای غربی است.

حال بیایید قیمت خانه را با تقسیم شاخص قیمت واقعی در هر نقطهٔ زمانی بر روند طولانی مدّت قیمت مسکن، روندزدایی نماییم. از نظر بصری، این تقسیم شبیه تفریق خطوط خاکستری از خطوط آبی در نمودار ۱۴–۱۰ خواهد بود زیرا تقسیم مقادیر تبدیل نشده معادل تفریق مقادیر تبدیل لگاریتمی شده می‌باشد. قیمت‌های مسکن حاصل که روندزدایی شده‌اند، حباب‌های مسکن را با وضوح بیشتری نشان می‌دهند (نمودار ۱۴–۱۱)، زیرا روندزدایی بر تغییرات غیرمنتظره در یک سری زمانی تاکید می‌کند.

1. Freddie Mac

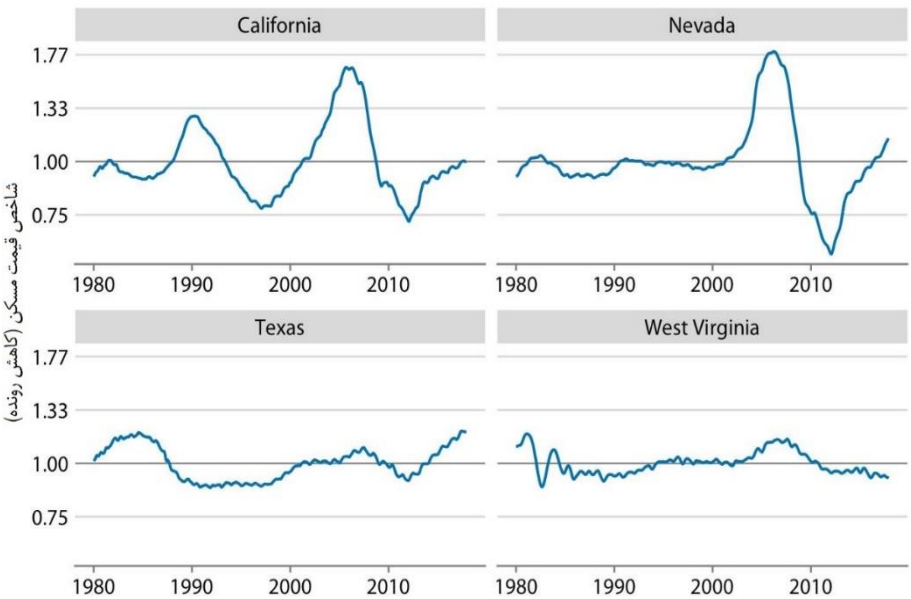


نمودار ۱۴–۱۰. شاخص قیمت خانه فردی مک از سال ۱۹۸۰ تا ۲۰۱۷، برای چهار ایالت منتخب (کالیفرنیا، نوادا، تگزاس و ویرجینیای غربی). شاخص قیمت مسکن یک عدد بدون واحد است که قیمت نسبی خانه را در منطقه جغرافیایی منتخب در طول زمان ردیابی می‌کند. این شاخص بطور قراردادی در دسامبر سال ۲۰۰۰ معادل ۱۰۰ انتخاب شده است. خطوط آبی نوسانات ماهانهٔ شاخص و خطوط خاکستری مستقیم نشان‌دهندهٔ روند طولانی مدّت قیمت در نواحی ذکر شده می‌باشد. توجّه شود که محورهای y لگاریتمی است، بنابراین خطوط خاکستری مستقیم نشان‌دهندهٔ رشد نمایی ثابت است. منبع داده‌ها: شاخص قیمت خانه فردی مک

برای مثال، در سری زمانی اصلی، کاهش قیمت مسکن در کالیفرنیا از سال ۱۹۹۰ تا حدود ۱۹۹۸ به نظر ملایم می‌رسد (نمودار ۱۴–۱۰). با این حال، در طول همان دورهٔ زمانی، بر اساس روند بلندمدت، انتظار افزایش قیمت وجود داشت. نسبت به افزایش مورد انتظار، کاهش قیمت‌ها قابل توجّه بود، و در کمترین نقطه به ۲۵ درصد نیز رسید (نمودار ۱۴–۱۱).

فراتر از روندزدایی، می‌توان یک سری زمانی را به چندین قسمت مجزا جدا نمود به طوری که مجموع آن‌ها سری زمانی اصلی را می‌سازد. به طور کلی، علاوه بر روند بلندمدت، سه مؤلفه مجزا وجود دارد که ممکن است سری زمانی را شکل دهد. اوّل، خطای تصادفی که منجر به نوسانات کوچک و نامنظّم به سمت بالا و پایین می‌شود. این خطای تصادفی در تمام سری‌های زمانی نشان داده شده در این فصل قابل مشاهده است، امّا شاید بیشتر از همه در نمودار ۱۴– ۹ مشهود باشد. دوّم، ممکن است رویدادهای خارجی منحصر به فردی وجود داشته باشد که ردپای آن‌ها در سری زمانی قابل تعقیب باشد مانند حباب‌های مسکن که در

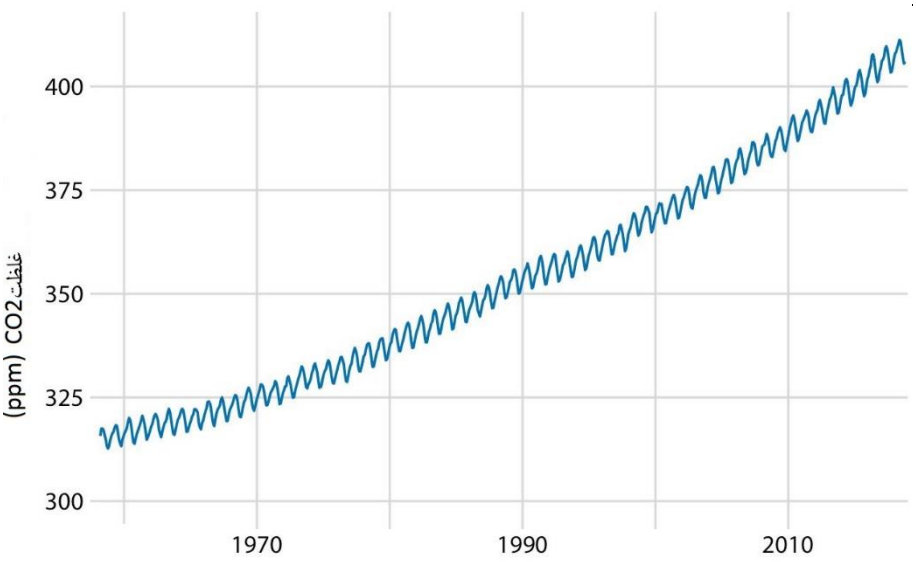
نمودار ۱۴–۱۰ مشاهده شد. سوّم، ممکن است تغییرات چرخه‌ای وجود داشته باشد. به عنوان مثال، دمای محیط بیرون از خانه تغییرات چرخه‌ای روزانه را نشان می‌دهد. بالاترین دما مربوط به اوایل بعد از ظهر و کمترین دما مربوط به اوایل صبح می‌باشد. دمای محیط بیرون از خانه تغییرات چرخه‌ای سالانه را نیز نشان می‌دهد. بدین صورت که در بهار تمایل به افزایش داشته، در تابستان به حداکثر خود رسیده و سپس در پاییز کاهش یافته و در زمستان به حداقل خود می‌رسد (نمودار ۳–۲). برای نشان دادن مفهوم مؤلفه‌های مجزای سری زمانی، در ادامه منحنی کیلینگ[1] که تغییرات مقدار CO_2 را در طول زمان نشان می‌دهد، تجزیه می‌نماییم (نمودار ۱۴–۱۲). از سال ۱۹۵۸، مقدار CO_2 به طور مداوم در مرکز Mauna Loa در هاوایی پایش شده است و در ابتدا تحت مدیریت چارلز کیلینگ[2] بوده است.



نمودار ۱۴–۱۱. نسخهٔ روندزدایی شدهٔ شاخص قیمت مسکن فردی مک که در نمودار ۱۴–۱۰ نشان داده شد. شاخص روندزدایی شده حاصل تقسیم شاخص واقعی (خطوط آبی در نمودار ۱۴–۱۰) بر مقدار مورد انتظار بر اساس روند بلند مدّت (خطوط خاکستری مستقیم در نمودار ۱۴–۱۰) می‌باشد. این تصویر نشان می‌دهد که کالیفرنیا دو حباب مسکن را در حدود سال ۱۹۹۰ و در اواسط دههٔ ۲۰۰۰، تجربه کرده است. این دو حباب از افزایش سریع و سپس کاهش قابل توجّه در قیمت واقعی مسکن نسبت به آنچه بر اساس روند طولانی مدّت انتظار می‌رفت شناسایی شده است. به طور مشابه، نوادا یک حباب مسکن را در اواسط دههٔ ۲۰۰۰ تجربه کرد و نه تگزاس و نه ویرجینیای غربی اصلا حبابی را تجربه نکردند. منبع داده: شاخص قیمت خانه فردی مک

1. Keeling curve
2. Charles Keeling

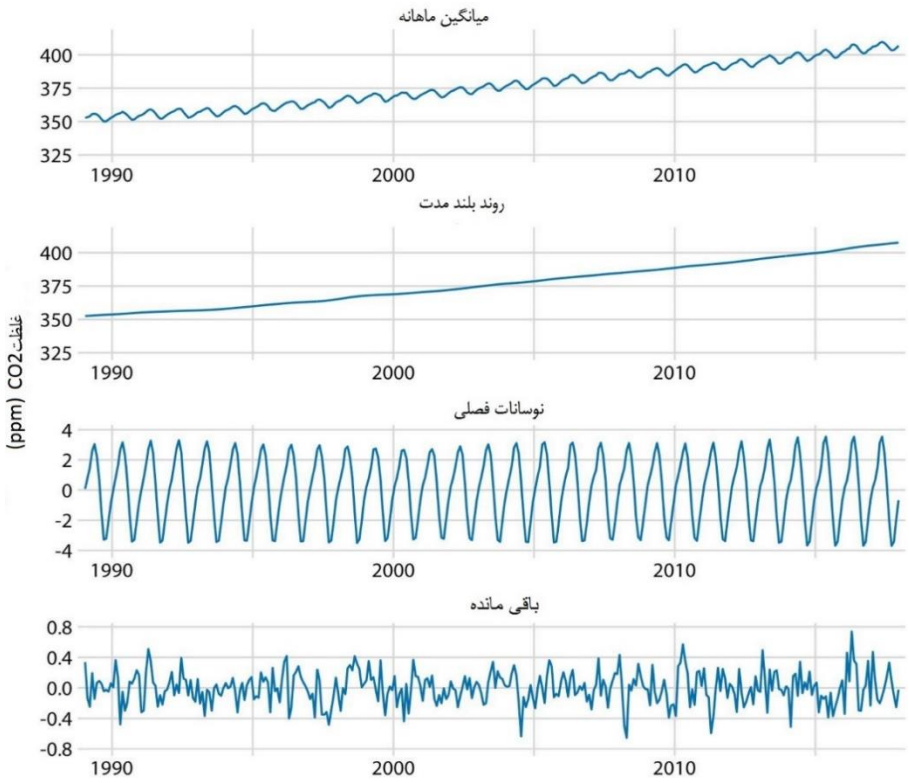
CO_2 بـر حسـب قسـمت در میلیـون[1] (ppm) انـدازه‌گیری می‌شـود. یافتـه‌ها حـاکی از افـزایش طولانی مـدّت فراوانـی CO_2 کـه کمـی سـریعتر از رونـد خطـی اسـت، از کمتـر از ۳۲۵ ppm در دهـهٔ ۱۹۶۰ تا بـالای ۴۰۰ در دهـهٔ دوّم قـرن بیسـت و یکـم می‌باشـد (نمـودار ۱۴–۱۲). همچنـین فراوانـی CO_2 نوسانات سالانه به صورت یـک الگـوی ثابـت حرکـات بـالا و پـایین دارد کـه بـر رونـد افـزایش کلـی سـوار شـده اسـت. نوسـانات سـالانه ناشـی از رشـد گیـاه در نیمکـره شـمالی اسـت. گیاهـان در طـول فتوسـنتز CO_2 مصـرف می‌کننـد. از آنجـایی کـه بیشـترین زمین‌هـای کـرهٔ زمـین در نیمکـرهٔ شمالی قرار دارند، و رشـد گیاهـان نیـز بیشـتر در بهـار و تابسـتان رخ می‌دهـد، شـاهد کـاهش سـالانهٔ CO_2 جهانی هستیم که مصادف با ماه‌های تابستان در نیمکرهٔ شمالی است.



نمودار ۱۴–۱۲. منحنی کیلینگ. منحنی کیلینگ تغییر CO_2 را در جو در طول زمان نشان می‌دهد. در اینجا میانگین مقدار CO_2 ماهانه در واحد قسمت در میلیون (ppm) نشان داده شده است. سالانه CO_2 با تغییر فصول تغییر می‌کند امّا روند افزایش طولانی مدّت ثابتی را نشان می‌دهد. منبع داده: Dr. Pieter Tans, NOAA /ESRL, and Dr. Ralph Keeling ، موسسه اقیانوس شناسی اسکریپس.

می‌تـوان منحنـی کیلینـگ را بـه رونـد بلندمـدت، نوسـانات فصـلی و بـاقی مانـده‌ها (نمـودار ۱۴–۱۳) تجزیه نمود. روش خاصی کـه در اینجـا اسـتفاده شـده اسـت تجزیـه فصـلی سـری‌های زمـانی توسـط LOESS (STL) [Cleveland et al. 1990] نامیـده می‌شـود امّـا بسـیاری از روش‌هـای دیگـر نیز وجود دارد که به اهداف مشابه دست می‌یابند.

1. parts per million



نمودار ۱۴–۱۳. تجزیهٔ سری زمانی منحنی کیلینگ، که میانگین ماهانه (مانند نمودار ۱۴–۱۲)، روند بلندمدت، نوسانات فصلی و باقی مانده را نشان می‌دهد. باقی مانده تفاوت بین مقادیر واقعی و مجموع روند بلند مدّت و نوسانات فصلی است و معرّف خطای تصادفی می‌باشد. در اینجا بر داده‌های ۳۰ سال اخیر تمرکز شده تا بر نوسانات سالانه تاکید شود. منبع داده: ESRL, and Dr. Ralph Keeling/Dr. Pieter Tans, NOAA، موسسه اقیانوس شناسی اسکریپس.

تجزیه نشان می‌دهد که در طول سه دهه گذشته، مقدار CO_2 بیش از ۵۰ ppm افزایش یافته است. از سوی دیگر، نوسانات فصلی تنها منجر به تغییری کمتر از ۸ ppm شده (هرگز باعث افزایش یا کاهش بیش از ۴ ppm نسبت به روند بلند مدّت نشده‌اند) و مقدار باقی مانده کمتر از ۱٫۶ ppm (نمودار ۱۴–۱۳) می‌باشد. باقی مانده تفاوت بین مقدار واقعی و مجموع روند بلندمدت و نوسانات فصلی است و در اینجا با خطای تصادفی مقدار ماهانهٔ CO_2 مطابقت دارد. با این حال، از منظر کلی‌تر، باقی مانده می‌تواند رویدادهای خارجی منحصر به فردی را نیز ثبت کند. به عنوان مثال، اگر یک فوران آتشفشان عظیم اتفاق افتاده و مقادیر قابل توجهی CO_2 آزاد شود چنین رویدادی ممکن است به صورت یک افزایش ناگهانی در باقی مانده‌ها قابل مشاهده باشد. نمودار ۱۴–۱۳ نشان می‌دهد که هیچ رویداد خارجی منحصر به فردی در دهه‌های اخیر تاثیر قابل توجهی بر منحنی کیلینگ نداشته است.

فصل ۱۵

ترسیم داده‌های مکانی

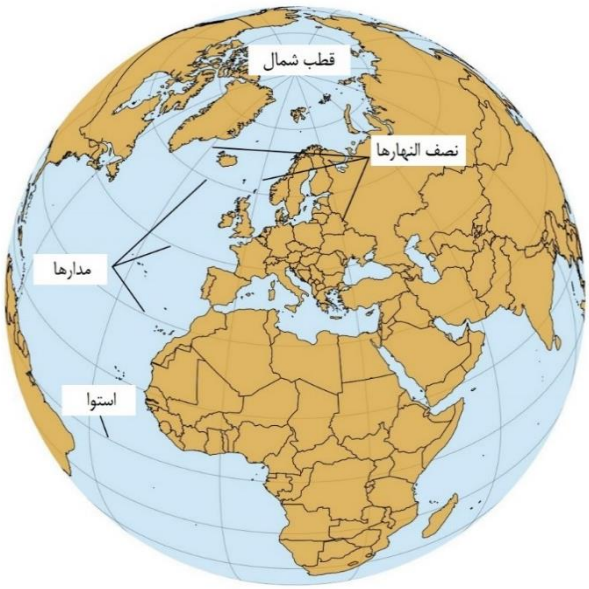
بسیاری از پایگاه‌های داده، حاوی اطّلاعات مرتبط با مکان‌ها در دنیا هستند. به عنوان مثال، در یک مطالعه اکولوژیکی، یک مجموعهٔ داده ممکن است مکان‌هایی را که گیاهان یا حیوانات خاصی پیدا شده‌اند، فهرست کند. به طور مشابه، در حوزه اجتماعی–اقتصادی یا سیاسی، یک مجموعهٔ داده ممکن است حاوی اطلاعاتی در مورد محل زندگی افراد با ویژگی‌های خاص (مانند درآمد، سن، یا پیشرفت تحصیلی) و یا مکان ساخت سازه‌ها (مانند پل‌ها، جاده‌ها، ساختمان‌ها) باشد. در تمام این موارد، ترسیم داده‌ها در زمینهٔ جغرافیایی مناسب، یعنی نمایش داده‌ها بر روی یک نقشهٔ واقعی یا نموداری شبیه نقشه، می‌تواند مفید باشد.

نقشه‌ها برای خوانندگان شهودی هستند، امّا طراحی آن‌ها می‌تواند چالش‌برانگیز باشد. ما باید در مورد مفاهیمی مانند پیش‌بینی نقشه و اینکه آیا برای کاربرد خاص مدنظرمان نمایش دقیق زوایا یا نواحی ضروری است یا خیر، تصمیم بگیریم. یک تکنیک معمول نگاشت، یعنی نقشه ناحیه–مقدار[1]، شامل نمایش مقادیر داده‌ها به صورت مناطق فضایی با رنگ‌های متفاوت است. نقشه‌های ناحیه–مقدار برخی اوقات می‌توانند بسیار مفید و در برخی مواقع کاملاً گمراه‌کننده باشند. به عنوان یک جایگزین، می‌توان نمودارهای نقشه مانندی به نام نقشه آماری[2] ترسیم نمود که ممکن است به طور هدفمند نواحی نقشه را تحریف کرده یا آن‌ها را به شکل الگوی خاص، مثلاً به صورت مربع‌های مساوی نشان دهد.

1. choropleth map
2. cartogram

برون‌یابی[1]

زمین تقریباً یک کره است (شکل ۱۵-۱) و به طور دقیق‌تر یک کرهٔ مایل است که در امتداد محور چرخش خود کمی مسطّح شده است. محل تقاطع محور چرخش با کره، قطب (شمال و جنوب) نامیده می‌شوند. کشیدن خطی به فاصله مساوی از هر دو قطب به دور کره، آن را به دو نیمکره شمالی و جنوبی تقسیم می‌کند. به این خط استوا می‌گویند. برای تعیین یک مکان منحصربه‌فرد روی زمین، به سه داده نیاز داریم: جایی که در امتداد جهت استوا قرار داریم (طول جغرافیایی)، زمانی که عمود بر استوا حرکت می‌کنیم (عرض جغرافیایی) چقدر به هر یک از قطب‌ها نزدیک می‌شویم، و از مرکز زمین (ارتفاع) چقدر فاصله داریم. طول و عرض جغرافیایی و ارتفاع نسبت به یک سیستم مرجع به نام سطح مبنا[2] مشخّص می‌شود. این سطح مبنا ویژگی‌هایی مانند شکل و اندازه زمین و همچنین طول، عرض و ارتفاع صفر را مشخّص می‌کند. یکی از سطوح مبنای پرکاربرد سیستم جهانی ژئودتیک 84 (WGS) است که توسّط سیستم موقعیّت‌یاب جهانی (GPS) استفاده می‌شود.



شکل ۱۵-۱. طرح اورتوگرافیک جهان، اروپا و شمال آفریقا را به گونه‌ای نشان می‌دهد که از فضا قابل مشاهده هستند. خطوطی که از قطب شمال سرچشمه می‌گیرند و به سمت جنوب می‌روند نصف‌النهار و خطوطی که به صورت متعامد روی نصف النهارها حرکت می‌کنند، مدار نامیده می‌شوند. طول همه نصف‌النهارها یکسان است، امّا هر چه به قطب نزدیک‌تر باشیم، مدارها کوتاهتر می‌شوند.

1. Projections
2. datum

در حالی که ارتفاع، یک کمیّت مهم در بسیاری از کاربردهای جغرافیایی است، هنگام ترسیم داده‌های مکانی در قالب نقشه، ما در درجهٔ اوّل به دو بُعد دیگر یعنی طول و عرض جغرافیایی توجّه می‌کنیم. طول و عرض جغرافیایی زوایایی هستند که بر حسب درجه بیان می‌شوند. درجهٔ طول جغرافیایی میزان شرق یا غرب بودن یک مکان را اندازه‌گیری می‌کند. خطوط با طول جغرافیایی مساوی به عنوان نصف‌النهار نامیده می‌شوند و همه نصف‌النهارها به دو قطب ختم می‌شوند (شکل ۱۵–۱). نصف‌النهار مبدا، مربوط به طول جغرافیایی صفر درجه، از روستای گرینویچ[1] در بریتانیا می‌گذرد. نصف‌النهار مخالف نصف‌النهار مبدا در طول جغرافیایی ۱۸۰ درجه (که به آن ۱۸۰ درجهٔ شرقی نیز گفته می‌شود) قرار دارد که معادل طول جغرافیایی ۱۸۰– درجه (که به آن ۱۸۰ درجهٔ غربی نیز گفته می‌شود) در نزدیکی خط تاریخ بین‌المللی است. درجهٔ عرض جغرافیایی میزان فاصله یک نقطه از شمال یا جنوب را اندازه‌گیری می‌کند. خط استوا مربوط به عرض جغرافیایی صفر درجه، قطب شمال مربوط به عرض جغرافیایی ۹۰ درجه (همچنین به عنوان ۹۰ درجهٔ شمالی نامیده می‌شود)، و قطب جنوب مربوط به عرض جغرافیایی ۹۰– درجه (همچنین به عنوان ۹۰ درجهٔ جنوبی نامیده می‌شود) می‌باشند. خطوط با عرض جغرافیایی مساوی به عنوان مدار شناخته می‌شوند، و موازی با خط استوا هستند. تمام نصف‌النهارها طول یکسانی دارند که معادل نیمی از یک دایرهٔ بزرگ در سراسر کرهٔ زمین است، در حالی که طول مدارها به عرض جغرافیایی آن‌ها بستگی دارد (شکل ۱۵–۱). طولانی‌ترین مدار، خط استوا در عرض جغرافیایی صفر درجه است و کوتاه‌ترین مدارها در قطب شمال و جنوب، ۹۰ درجه شمالی و ۹۰ درجه جنوبی قرار دارند و طول آن‌ها صفر است.

چالش در ساخت نقشه این است که باید سطح کروی زمین را بگیریم و آن را صاف کنیم تا بتوانیم آن را روی نقشه نمایش دهیم. این فرآیند که برون‌یابی نامیده می‌شود، ناچاراً تحریف‌هایی را ایجاد می‌کند، زیرا یک سطح منحنی را نمی‌توان دقیقاً روی یک سطح صاف قرار داد. برون‌یابی می‌تواند زوایا یا نواحی را حفظ کند امّا نه هر دو را به طور همزمان. برون‌یابی که اوّلی را انجام می‌دهد همشکل[2] و برون‌یابی‌ای که دوّمی را انجام می‌دهد مساحت مساوی نامیده می‌شود. سایر برون‌یابی‌ها ممکن است نه زوایا و نه مناطق را حفظ کنند، امّا در عوض مقادیر دیگری را حفظ کنند، مانند فواصل تا نقطه یا خط مرجع. در نهایت، برخی از برون‌یابی‌ها تلاش می‌کنند تا بین حفظ زوایا و نواحی تعادلی ایجاد کنند. این پیش‌بینی‌های تعادلی اغلب برای نمایش کل جهان به شیوه‌ای که از نظر بصری زیبا باشد مورد استفاده قرار می‌گیرند و تحریف‌هایی در زاویه یا ناحیه را می‌پذیرند (شکل ۳–۱۱). برای

1. Greenwich
2. conformal

سیستم‌بندی و پیگیری روش‌های مختلف نمایش بخش‌ها یا کل زمین برای نقشه‌های خاص، موسسه و سازمان‌های استاندارد مختلف، مانند گروه پیمایش نفت اروپا (EPSG) و موسسه تحقیقات سیستم‌های محیطی (ESRI)، برون‌یابی‌ها را ثبت می‌کنند. به عنوان مثال، EPSG:4326 مقادیر طول و عرض جغرافیایی برون‌یابی نشده را در سیستم مختصّات WGS 84 که توسّط GPS استفاده می‌شود، نشان می‌دهد. چندین وبسایت دسترسی آزاد به این برون‌یابی‌های ثبت شده را فراهم می‌کنند، از جمله http://spatialreference.org و https://epsg.io.

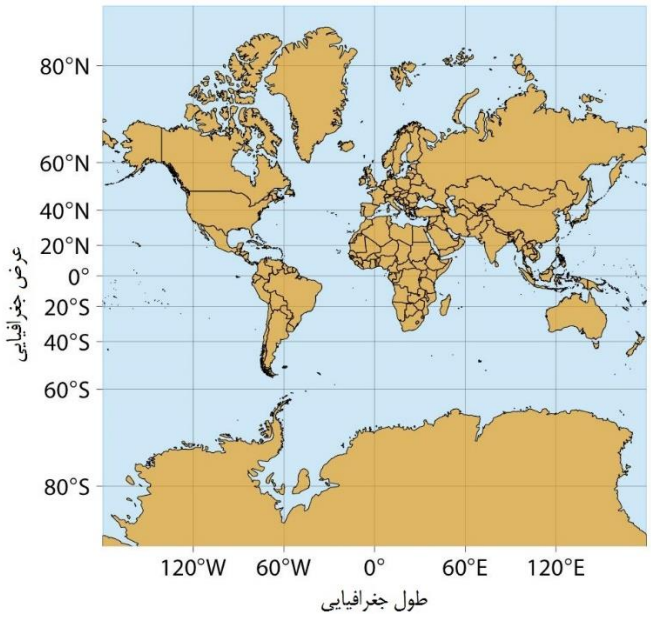
یکی از اوّلین برون‌یابی‌های نقشه در حال استفاده، طرح مرکاتور[1]، در قرن شانزدهم برای ناوبری دریایی توسعه یافت. این یک برون‌یابی همشکل است که به دقّت اشکال را نشان می‌دهد امّا تحریف‌های شدیدی در نواحی نزدیک قطب‌ها ایجاد می‌کند (شکل ۱۵-۲). طرح مرکاتور، کرهٔ زمین را روی یک استوانه ترسیم می‌کند و سپس استوانه را باز می‌کند تا به یک نقشهٔ مستطیل شکل برسد. نصف‌النهارها در این برون‌یابی به صورت خطوط عمودی با فاصلهٔ یکسان هستند، در حالی که مدارها خطوط افقی هستند که فاصله آن‌ها با دور شدن از استوا افزایش می‌یابد. فاصلهٔ بین مدارها به نسبت افزایش کشیدگی با هدف نزدیکی به قطب‌ها افزایش می‌یابد تا نصف‌النهارها به صورت کاملاً عمودی نگه داشته شوند.

به دلیل تحریف‌های شدید منطقه‌ای، تمایل به استفاده از برون‌یابی مرکاتور برای نقشه‌های کل جهان از بین رفته است. با این حال، انواعی از این برون‌یابی همچنان به حیات خود ادامه می‌دهند. برای مثال، برون‌یابی عرضی مرکاتور معمولاً برای نقشه‌های مقیاس بزرگ استفاده می‌شود که مناطق نسبتاً کوچک (که در طول جغرافیایی کمتر از چند درجه قرار دارند) را با بزرگنمایی بالا نشان می‌دهند. نوع دیگر، برون‌یابی وب مرکاتور، توسّط گوگل برای نقشه‌های گوگل معرفی شد و توسّط چندین برنامه نقشه‌برداری برخط استفاده می‌شود.

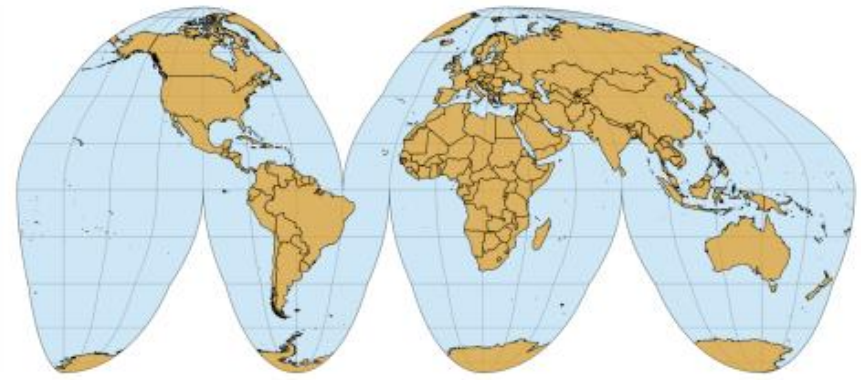
یک برون‌یابی کل جهان که کاملاً منطقه را حفظ می‌کند هموالوزین Goode است (شکل ۱۵-۳). این برون‌یابی معمولاً به شکل منقطع خود نشان داده می‌شود که دارای یک برش در نیمکرهٔ شمالی و سه برش در نیمکرهٔ جنوبی است که با دقّت انتخاب شده است تا توده‌های خشکی اصلی را قطع نکنند (شکل ۱۵-۳). برش‌ها به برون‌یابی اجازه می‌دهد هم مناطق و هم تقریباً زاویه‌ها را حفظ کند، البته به قیمت نمایش غیریکپارچه اقیانوس‌ها، یک برش از وسط گرینلند، و چندین برش در قطب جنوب. در حالی که هموالوزین Goode ظاهر

1. Mercator projection

بصری غیرعادی و نـامی عجیـب دارد، امّـا انتخـاب خـوبی بـرای برنامـه‌های نقشـه‌بـرداری اسـت کـه نیاز به بازتولید دقیق مناطق در مقیاس جهانی دارند.



شکل ۱۵–۲. برون‌یابی مرکاتور از جهان. در این برون‌یابی، مدارها خطوط افقی مستقیم و نصف‌النهارها خطوط عمودی مستقیم هستند. این یک برون‌یابی همشکل است که زوایا را حفظ می‌کند، امّا تحریف‌های شدیدی را در مناطق نزدیک به قطب‌ها ایجاد می‌کند. به عنوان مثال، گرینلند در این برون‌یابی بزرگتر از آفریقا به نظر می‌رسد، در حالی که در واقعیّت آفریقا ۱۴ برابر بزرگتر از گرینلند است (شکل‌های ۱۵–۱ و ۱۵–۳ را ببینید).



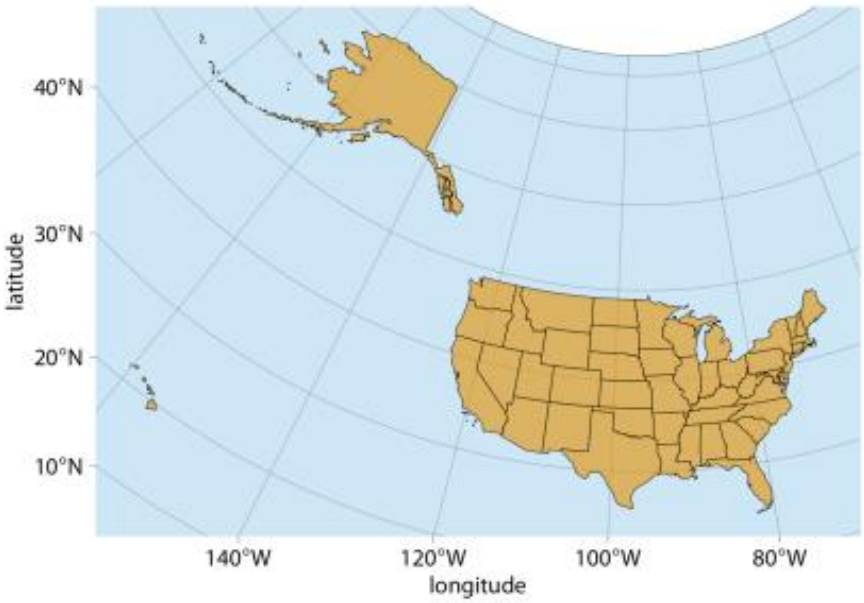
شکل ۱۵–۳. برون‌یابی منقطع هممولوزین Goode از جهان. این طرح به قیمت نمایش غیریکپارچه اقیانوس‌ها و برخی از توده‌های خشکی (گرینلند، قطب جنوب)، مناطق را با دقّت حفظ می‌کند و در عین حال تحریف‌های زاویه‌ای را به حداقل می‌رساند.

تحریف شکل یا ناحیه به دلیل برون‌یابی نقشه‌ها به‌ویژه زمانی که می‌خواهیم نقشه‌ای از کل جهان بسازیم برجسته است، امّا می‌تواند حتّی در مقیاس قاره‌ها یا کشورها نیز باعث ایجاد مشکل شود. به عنوان مثال، ایالات متحده را در نظر بگیرید که از ۴۸ ایالت پایین‌تر (که ۴۸ ایالت به هم پیوسته هستند)، آلاسکا و هاوایی (شکل ۱۵–۴) تشکیل شده است. در حالی که ۴۸ ایالت پایین به راحتی روی نقشه نمایش داده می‌شوند، آلاسکا و هاوایی آن قدر از ۴۸ ایالت پایین فاصله دارند که نمایش همه ۵۰ ایالت روی یک نقشه عجیب می‌شود.

شکل ۱۵–۴. مکان‌های نسبی آلاسکا، هاوایی و ۴۸ ایالت پایین‌تر که روی کرهٔ زمین نشان داده شده‌اند.

شکل ۱۵–۵ نقشه‌ای از تمام ۵۰ حالت را نشان می‌دهد که با استفاده از برون‌یابی آلبرز[1] با مساحت مساوی ساخته شده‌اند. این برون‌یابی نمایش معقولی از اشکال، نواحی و مکان‌های نسبی ۵۰ ایالت ارائه می‌دهد، امّا برخی مسائل همچنان پابرجاست. اوّل، آلاسکا در مقایسه با شکل ظاهری آن، برای مثال، آنچه در شکل‌های ۱۵–۲ یا ۱۵–۴ آمده است، به‌طور عجیبی کشیده به نظر می‌رسد. دوّم، نقشه تحت سلطه اقیانوس/فضای خالی است. بهتر است بزرگ‌نمایی بیشتر شود به طوری که ۴۸ ایالت پایین‌تر نسبت بیشتری از نقشه را اشغال کنند.

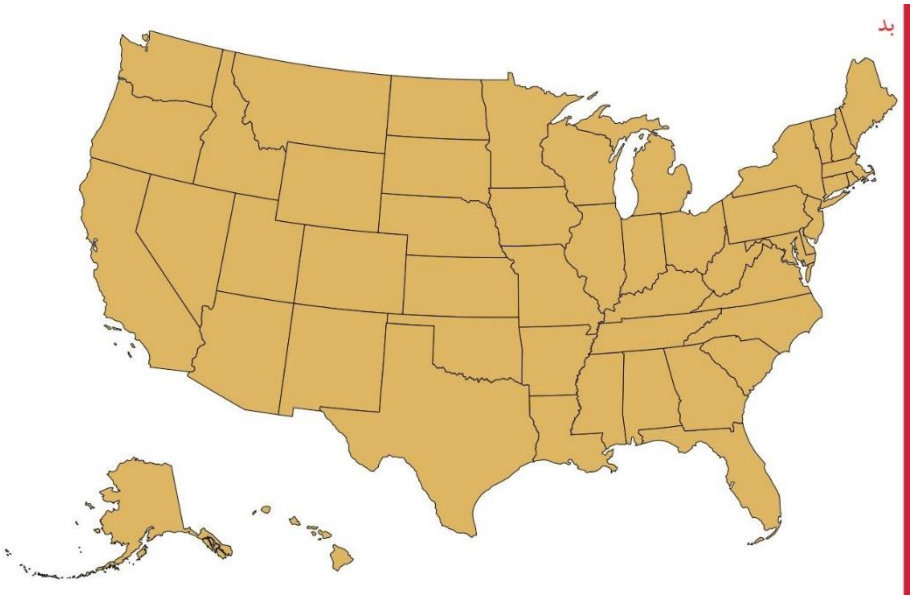
1. Albers projection



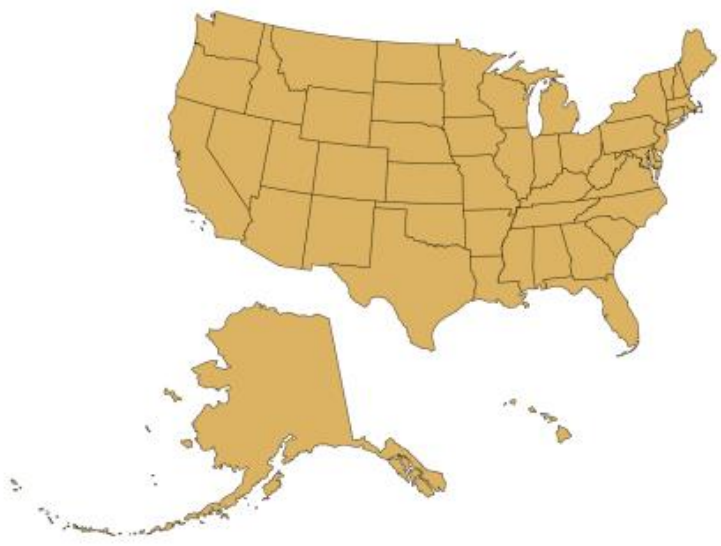
شکل ۱۵–۵. نقشه ایالات متحده آمریکا، با استفاده از برون‌یابی آلبرز که منطقه را حفظ می‌کند (ESRI: 102003، که معمولاً برای نمایش ۴۸ ایالت پایین‌تر استفاده می‌شود). آلاسکا و هاوایی در مکان‌های واقعی خود نشان داده شده‌اند.

برای پرداختن به مشکل فضای خالی غیرجذاب، معمول است که آلاسکا و هاوایی به طور جداگانه ترسیم شود (برای به حداقل رساندن تحریف شکل) و سپس در زیر ۴۸ ایالت پایینی نشان داده شوند (شکل ۱۵–۶). ممکن است در شکل ۱۵–۶ متوجّه شوید که آلاسکا نسبت به ۴۸ ایالت پایینی در شکل ۱۵–۵ بسیار کوچکتر به نظر می‌رسد. دلیل این اختلاف این است که آلاسکا نه تنها جابجا شده است، بلکه به گونه‌ای مقیاس‌بندی شده تا از نظر اندازه با ایالات معمولی غرب میانه یا غربی قابل مقایسه است. این مقیاس‌بندی، هرچند رایج است، امّا گمراه‌کننده بوده و بنابراین این نمودار به عنوان «بد» برچسب خورده است.

به جای اینکه آلاسکا هم جابجا شود و هم مقیاس آن عوض شود، می‌توان آن را بدون تغییر مقیاس حرکت داد (شکل ۱۵–۷). این ترسیم نشان می‌دهد که آلاسکا بزرگ‌ترین ایالت است که دو برابر تگزاس وسعت دارد. ما عادت نداریم که ایالات متحده را به این شکل ببینیم، امّا به نظر این نمایش معقول‌تری از ۵۰ ایالت آمریکا نسبت به شکل ۱۵–۶ است.



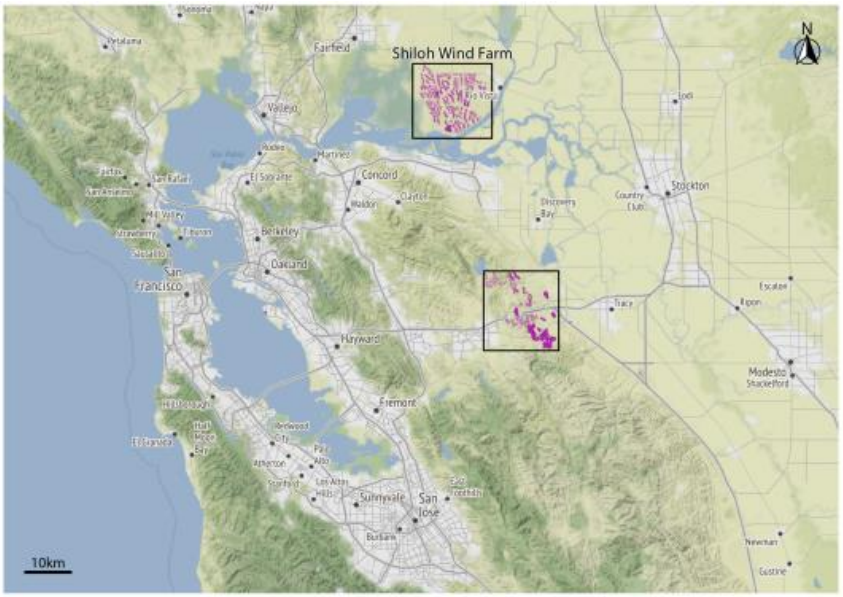
شکل ۱۵–۶. نقشهٔ ایالات متحده، با ایالت‌های آلاسکا و هاوایی که در زیر ۴۸ ایالت پایینی قرار دارند. آلاسکا نیز مقیاس‌بندی شده است، بنابراین وسعت خطی آن تنها ۳۵ درصد از اندازهٔ واقعی ایالت است. (به عبارت دیگر، مساحت ایالت تقریباً به ۱۲ درصد از اندازه واقعی آن کاهش یافته است.) چنین مقیاس‌بندی اغلب در آلاسکا اعمال می‌شود تا از نظر بصری به اندازه ایالت‌های غرب میانه یا غربی به نظر برسد. با این حال، این مقیاس‌بندی گمراه‌کننده است، و بنابراین این نمودار به عنوان «بد» برچسب‌گذاری شده است.



شکل ۱۵–۷. نقشه ایالات متحده، با ایالت‌های آلاسکا و هاوایی که در زیر ۴۸ ایالت پایینی قرار دارند.

لایه‌ها

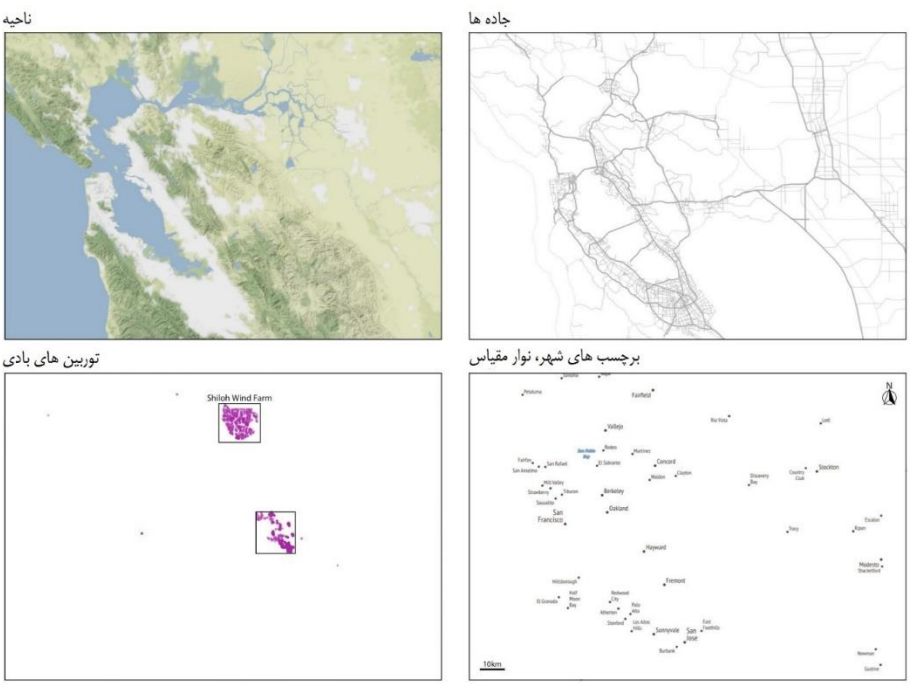
برای ترسیم داده‌های مکانی در زمینهٔ مناسب، معمولاً نقشه‌هایی متشکل از لایه‌های متعددی ایجاد می‌شود که انواع مختلف اطّلاعات را نشان می‌دهد. برای نشان دادن این مفهوم، مکان توربین‌های بادی را در منطقه خلیج سانفرانسیسکو نمایش می‌دهیم. در منطقه خلیج، توربین‌های بادی در دو مکان خوشه‌بندی شده‌اند. یک مکان، که از آن به عنوان مزرعه بادی شیلوه یاد خواهیم کرد، در نزدیکی ریو ویستا و دیگری در شرق هیوارد در نزدیکی تریسی قرار دارد (شکل ۱۵–۸).



شکل ۱۵–۸. توربین‌های بادی در منطقهٔ خلیج سانفرانسیسکو. هر کدام از توربین‌های بادی به صورت نقاط بنفش رنگ نشان داده شده‌اند. دو منطقه با تراکم بالایی از توربین‌های بادی با مستطیل‌های سیاه برجسته شده‌اند. ما مجموعه توربین‌های بادی نزدیک ریو ویستا را به عنوان مزرعه بادی شیلوه می‌نامیم. قطعه‌های نقشه توسّط Stamen Design و تحت CC BY 3.0 ارائه شده است. داده‌های نقشه توسّط OpenStreetMap و تحت ODbL می‌باشد. منبع داده‌های توربین بادی: پایگاه داده‌های توربین بادی ایالات متحده.

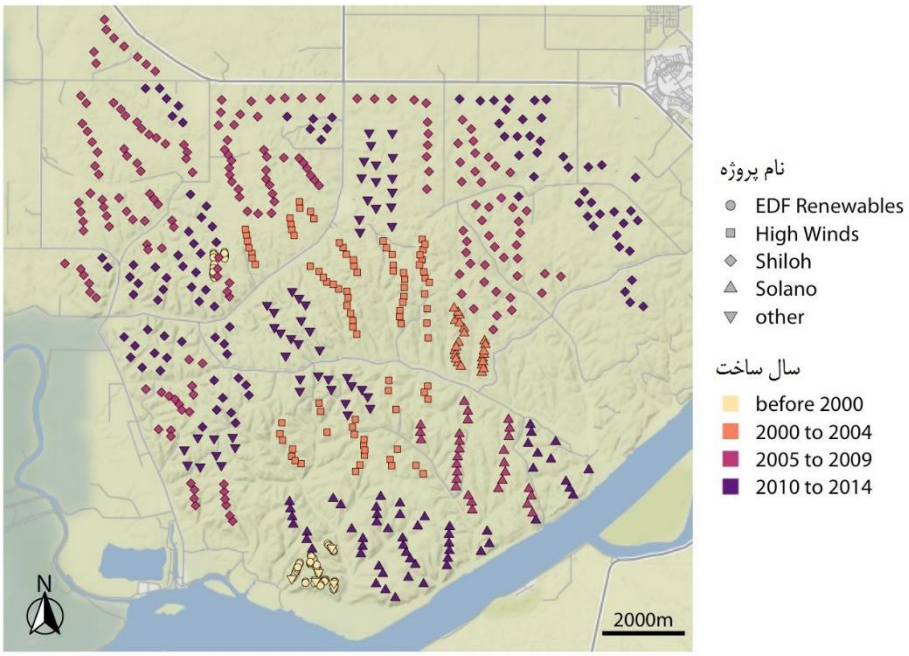
شکل ۱۵–۸ از چهار لایه مجزا تشکیل شده است. در پایین، لایهٔ زمین را داریم که تپه‌ها، دره‌ها و آب را نشان می‌دهد. لایهٔ بعدی شبکهٔ راه را نشان می‌دهد. در بالای لایهٔ جاده، یک لایه قرار داده‌ایم که مکان هر کدام از توربین‌های بادی را نشان می‌دهد. این لایه همچنین شامل دو مستطیل است که اکثر توربین‌های بادی را برجسته می‌کند. در نهایت، لایهٔ بالایی

مکان‌ها و نام شهرها را اضافه می‌کند. این چهار لایه به طور جداگانه در شکل ۱۵–۹ نشان داده شده است. برای هر نقشه‌ای که می‌خواهیم رسم کنیم، ممکن است برخی از این لایه‌ها را اضافه یا حذف کنیم. به عنوان مثال، اگر بخواهیم نقشه‌ای از مناطق رای‌دهنده ترسیم کنیم، ممکن است اطّلاعات زمین را نامربوط و گیج‌کننده بدانیم. از طرف دیگر، اگر بخواهیم نقشه‌ای از مناطق باز یا مسقّف ترسیم کنیم تا پتانسیل ذخیره انرژی خورشیدی را ارزیابی کنیم، ممکن است اطّلاعات زمین را با تصاویر ماهواره‌ای جایگزین کنیم تا مناطق مسقّف و پوشش گیاهی واقعی را نشان می‌دهد. شما می‌توانید به صورت تعاملی این انواع مختلف لایه‌ها را در اکثر نقشه‌های برخط مانند نقشه‌های گوگل امتحان کنید. تأکید می‌کنیم که صرف نظر از اینکه کدام لایه را می‌خواهید نگه دارید یا حذف کنید، به طور کلی توصیه می‌شود یک نوار مقیاس و یک فلش جهت شمال اضافه نمایید. نوار مقیاس به خوانندگان کمک می‌کند تا اندازه ویژگی‌های فضایی نشان داده شده در نقشه را درک کنند، در حالی که فلش شمال جهت نقشه را نشان می‌دهد.



شکل ۱۵–۹. لایه‌های جداگانه شکل ۱۵–۸. از پایین به بالا، این شکل شامل یک لایه زمین، یک لایه جاده، یک لایه نشان‌دهندهٔ توربین‌های بادی، و یک لایه نشان‌دهندهٔ شهرها و اضافه کردن یک نوار مقیاس و فلش شمالی است. قطعه‌های نقشه توسط Stamen Design و تحت CC BY 3.0 ارائه شده است. داده‌های نقشه توسط OpenStreetMap و تحت ODbL می‌باشد. منبع داده‌های توربین بادی: پایگاه داده‌های توربین بادی ایالات متحده.

تمام مفاهیم مورد بحث در فصل ۲ از نگاشت داده‌ها در مورد زیبایی‌شناسی در خصوص نقشه‌ها نیز صدق می‌کند. ما می‌توانیم نقاط داده را در بافت جغرافیایی آن‌ها قرار دهیم و سایر ابعاد داده را از طریق سایر ابزارهای زیبایی‌شناسی مانند رنگ یا شکل نشان دهیم. به عنوان مثال، شکل ۱۵–۱۰ یک تصویر بزرگنمایی شده از مستطیل با برچسب «مزرعه بادی شیلوه» در شکل ۱۵–۸ ارائه می‌دهد. توربین‌های بادی مجزا به صورت نقطه‌ای نشان داده می‌شوند که رنگ آن نشان‌دهندهٔ زمان ساخت و شکل آن نشان‌دهندهٔ پروژه‌ای است که توربین بادی به آن تعلّق دارد. نقشه‌ای مانند این می‌تواند یک نمای کلی از چگونگی توسعه یک منطقه ارائه دهد. به عنوان مثال، در اینجا می‌بینیم که EDF Renewables یک پروژهٔ نسبتاً کوچک است که قبل از سال ۲۰۰۰ ساخته شده است، High Winds یک پروژه با اندازهٔ متوسط است که بین سال‌های ۲۰۰۰ و ۲۰۰۴ ساخته شده است، و Shiloh و Solano بزرگترین پروژه‌های منطقه هستند که هر دو در یک دورهٔ طولانی ساخته شده‌اند.

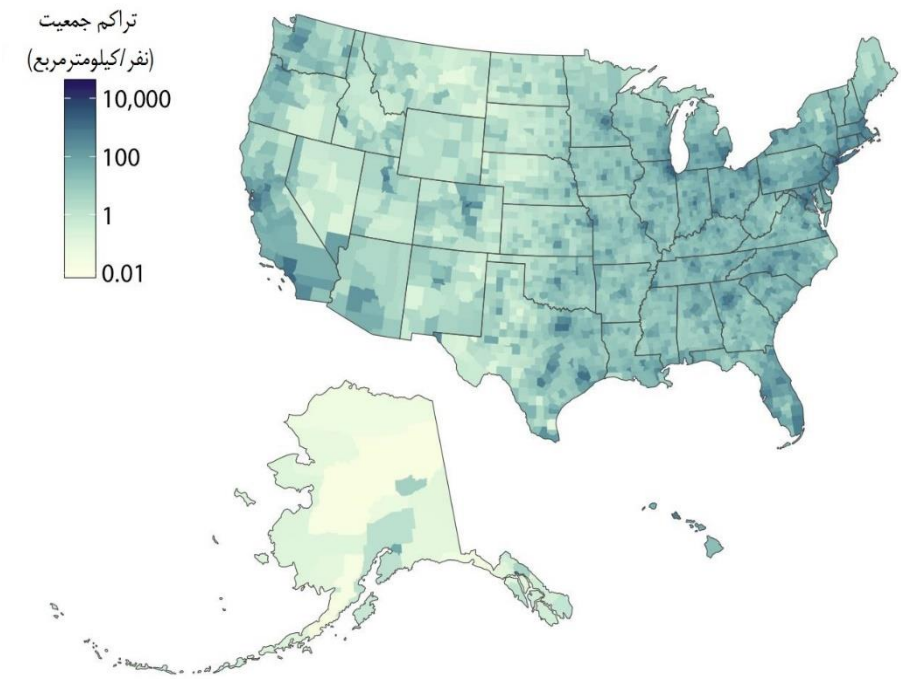


شکل ۱۵–۱۰. مکان توربین‌های بادی مجزا در مزرعه بادی شیلوه. هر نقطه محل یک توربین بادی را مشخّص می‌کند. منطقهٔ نقشه مربوط به مستطیل بالایی در شکل ۱۵–۸ است. نقطه‌ها بر اساس زمان ساخت توربین بادی رنگ شده و شکل نقطه نشان‌دهندهٔ پروژه‌ای است که توربین بادی به آن تعلق دارد. قطعه‌های نقشه توسّط Stamen Design و تحت CC BY 3.0 ارائه شده است. داده‌های نقشه توسّط OpenStreetMap و تحت ODbL می‌باشد. منبع داده‌های توربین بادی: پایگاه داده‌های توربین بادی ایالات متحده.

نقشه برداری ناحیه–مقدار

ما اغلب می‌خواهیم نشان دهیم که چگونه برخی متغیّرها در مکان‌های مختلف، متفاوت است. می‌توانیم این کار را با رنگ‌آمیزی مناطق جداگانه در یک نقشه با توجّه به بُعد داده‌ای که می‌خواهیم نمایش دهیم انجام دهیم. چنین نقشه‌هایی را نقشه‌های ناحیه–مقدار[1] می‌نامند.

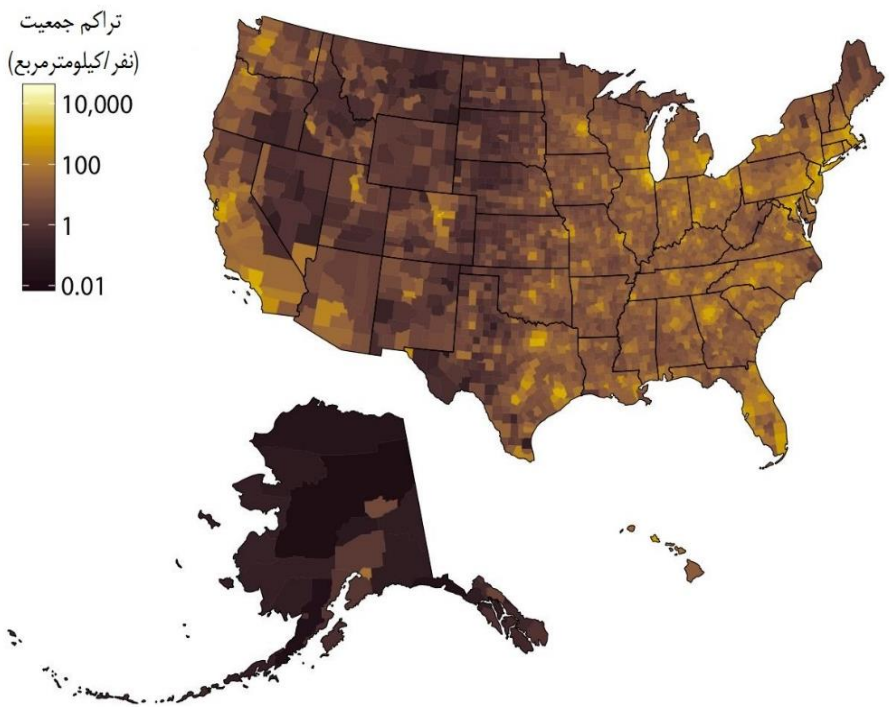
به عنوان یک مثال ساده، تراکم جمعیّت (نفر در هر کیلومتر مربّع) را در ایالات متحده در نظر بگیرید. تعداد جمعیّت هر شهرستان در ایالات متحده را در نظر می‌گیریم، آن را بر مساحت شهرستان تقسیم می‌کنیم، و سپس نقشه‌ای می‌کشیم که در آن رنگ هر شهرستان با نسبت بین تعداد جمعیّت و منطقه مطابقت دارد (شکل ۱۵–۱۱). می‌توانیم ببینیم که شهرهای اصلی در سواحل شرقی و غربی پرجمعیّت‌ترین مناطق ایالات متحده هستند، دشت‌های بزرگ و ایالت‌های غربی تراکم جمعیّت پایینی دارند و ایالت آلاسکا کمترین جمعیّت را دارد.



شکل ۱۵–۱۱. تراکم جمعیّت در هر شهرستان ایالات متحده، که به صورت یک نقشه ناحیه–مقدار نشان داده شده است. تراکم جمعیّت به صورت نفر در هر کیلومتر مربّع گزارش شده است. منبع داده: پیمایش پنج ساله جامعه آمریکا در سال ۲۰۱۵.

1. choropleth map

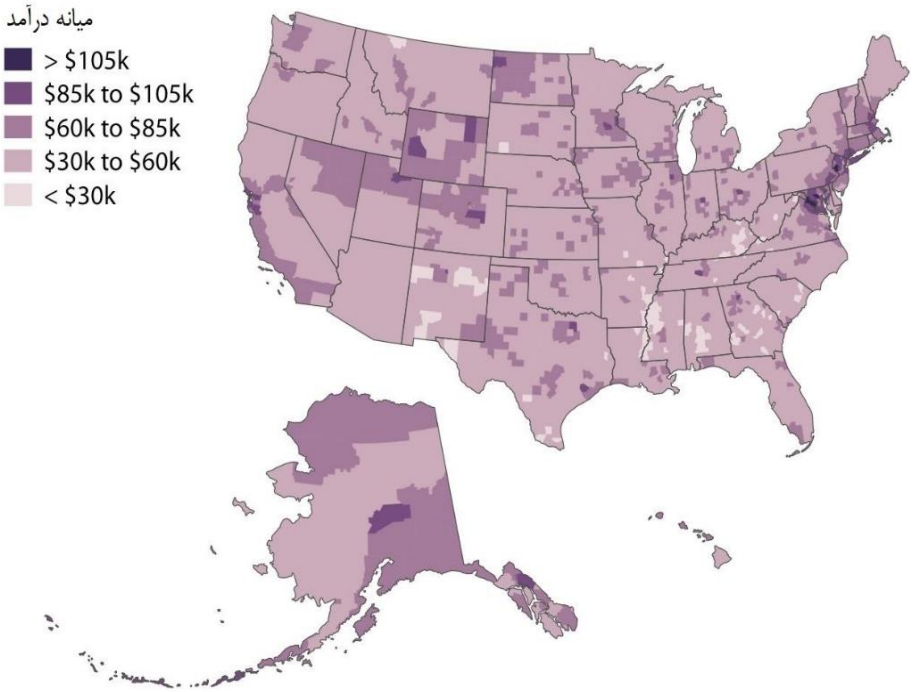
شکل ۱۵–۱۱ از رنگ‌های روشن برای نشان دادن تراکم جمعیّت کم و رنگ‌های تیره برای نشان دادن تراکم بالا استفاده می‌کند، به طوری که مناطق شهری با تراکم بالا به صورت رنگ‌های تیره روی پس‌زمینه رنگ‌های روشن برجسته می‌شوند. وقتی رنگ پس‌زمینه شکل روشن است، رنگ‌های تیره‌تر را با شدت‌های بالاتر مرتبط می‌کنیم. با این حال، می‌توانیم مقیاس رنگی را انتخاب کنیم که در آن مقادیر بالا توسّط رنگ‌های روشن و روی یک پس‌زمینه تاریک، نشان داده می‌شوند (شکل ۱۵–۱۲). تا زمانی که رنگ‌های روشن‌تر در طیف قرمز–زرد قرار می‌گیرند، به طوری که درخشان به نظر می‌رسند، می‌توان آن‌ها را به‌عنوان نمایانگر شدت‌های بالاتر درک کرد. به عنوان یک اصل کلی، زمانی که قرار است نمودارها روی کاغذ سفید چاپ شوند، مناطق پس‌زمینه با رنگ روشن (مانند شکل ۱۵–۱۱) معمولاً بهتر هستند. برای مشاهده برخط یا در پس‌زمینه تیره، مناطق پس‌زمینه رنگی تیره (مانند شکل ۱۵–۱۲) ممکن است ارجح باشند.



شکل ۱۵–۱۲. تراکم جمعیّت در شهرستان‌های ایالات متحده، که به صورت نقشه ناحیه–مقدار نشان داده شده است. این نقشه مشابه شکل ۱۵–۱۱ است با این تفاوت که اکنون مقیاس رنگ از رنگ‌های روشن برای تراکم جمعیّت بالا و از رنگ‌های تیره برای تراکم جمعیّت کم استفاده می‌کند. منبع داده: پیمایش پنج ساله جامعه آمریکا در سال ۲۰۱۵

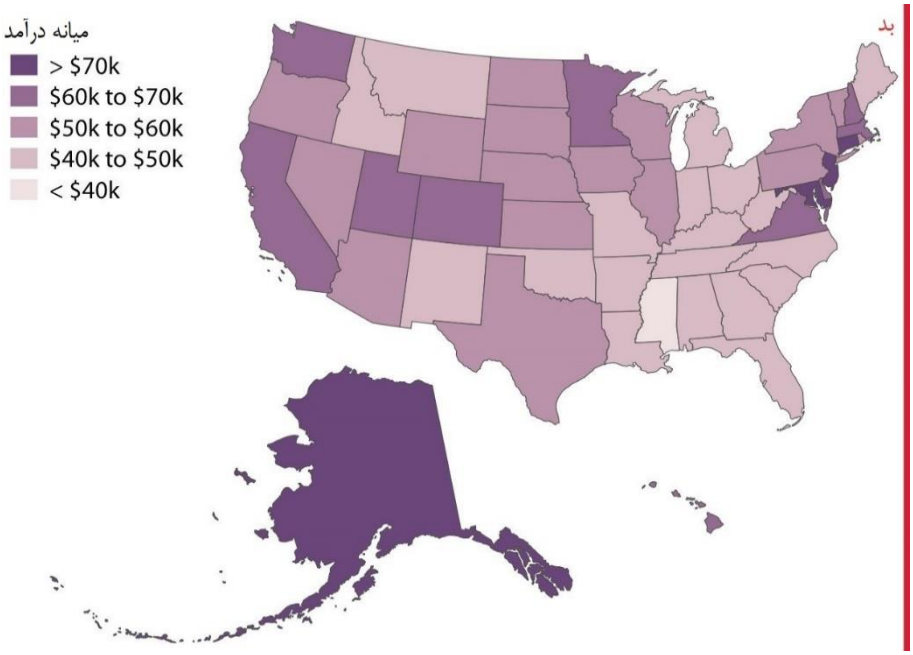
زمانی که از رنگ‌آمیزی برای نمایش چگالی (به عنوان مثال، یک متغیّر کمی تقسیم بر مساحت سطح، مانند شکل‌های ۱۵–۱۱ و ۱۵–۱۲) استفاده می‌شود، نقشه‌های ناحیه–مقدار بهترین کارایی را دارند. ما انتظار داریم مناطق با مساحت بیشتر نسبت به مناطق کوچکتر، مقادیر بیشتری از متغیّر کمی را به خود اختصاص دهند (فصل ۱۷ را نیز ببینید)، و استفاده از چگالی این اثر را تصحیح می‌کند. با این حال، در عمل، اغلب می‌بینیم که نقشه‌های ناحیه–مقدار بر اساس متغیّری رنگ می‌شوند که چگالی نیست. به عنوان مثال، در شکل ۴–۴ نقشه‌ای از میانهٔ درآمد سالانه را در شهرستان‌های تگزاس نشان دادیم. چنین نقشه‌هایی زمانی می‌توانند مناسب باشند که با احتیاط تهیّه شوند. دو وضعیتی که تحت آن‌ها می‌توانیم از رنگ‌آمیزی برای نمایش مقادیری که چگالی نیستند، استفاده کنیم عبارتند از اولاً، اگر تمام قسمت‌هایی که رنگ می‌کنیم تقریباً اندازه و شکل یکسانی داشته باشند، دیگر لازم نیست نگران باشیم که برخی مناطق صرفاً به دلیل اندازه آن‌ها توجّه نامتناسبی را به خود جلب کنند. دوّم، اگر نواحی که رنگ می‌کنیم در مقایسه با اندازهٔ کلی نقشه نسبتاً کوچک باشند و اگر کمیّت آن رنگ در مقیاس بزرگتری نسبت به مناطق تغییر می‌کند، دیگر لازم نیست نگران این باشیم که برخی مناطق صرفاً به دلیل اندازه آن‌ها توجّه نامتناسبی را به خود جلب کنند. تقریباً هر دوی این شرایط در شکل ۴–۴ وجود دارد.

همچنین باید تأثیر مقیاس‌های رنگی پیوسته در مقابل مقیاس‌های رنگی گسسته را در نگاشت ناحیه–مقدار در نظر بگیریم. در حالی که مقیاس‌های رنگی پیوسته از نظر بصری جذاب به نظر می‌رسند (به عنوان مثال، شکل‌های ۱۵–۱۱ و ۱۵–۱۲)، خواندن آن‌ها ممکن است دشوار باشد. ما در تشخیص یک رنگ خاص و تطبیق آن با مقیاس پیوسته خیلی توانمند نیستیم. بنابراین، اغلب بهتر است که مقادیر داده‌ها را در گروه‌های مجزا که با رنگ‌های متمایز نشان داده می‌شوند، قرار دهیم. چهار تا شش طبقهٔ رنگ انتخاب خوبی است. دسته‌بندی برخی از اطّلاعات را قربانی می‌کند، امّا از طرف دیگر، رنگ‌های دسته‌بندی شده را می‌توان به طور منحصر به فرد تشخیص داد. به عنوان مثال، شکل ۱۵–۱۳ نقشه میانهٔ درآمد در شهرستان‌های تگزاس (شکل ۴–۴) را به تمام شهرستان‌های ایالات متحده گسترش می‌دهد و از یک مقیاس رنگی متشکل از پنج گروه درآمدی متمایز استفاده می‌کند.



شکل ۱۵–۱۳. میانهٔ درآمد در شهرستان‌های ایالات متحده، که به صورت نقشه ناحیه–مقدار نشان داده شده است. مقادیر میانهٔ درآمد به پنج گروه مجزا تقسیم شده‌اند، زیرا مقیاس‌های رنگی دسته‌بندی شده معمولاً راحت‌تر از مقیاس‌های رنگی پیوسته خوانده می‌شوند. منبع داده: پیمایش پنج ساله جامعه آمریکا در سال ۲۰۱۵

اگرچه شهرستان‌ها به اندازه تگزاس در سراسر ایالات متحده مساحت مشابه نداشته و هم شکل نیستند، امّا شکل ۱۵–۱۳ همچنان به عنوان یک نقشه ناحیه–مقدار عمل می‌کند. هیچ شهرستانی بیش از حد بر نقشه تسلط ندارد. با این حال، وقتی یک نقشهٔ مشابه در سطح ایالت ترسیم می‌کنیم، همه چیز متفاوت به نظر می‌رسد (شکل ۱۵–۱۴). در اینجا آلاسکا بر نقشه ناحیه–مقدار تسلط پیدا می‌کند و به دلیل وسعت آن، نشان می‌دهد که میانهٔ درآمدی بالای ۷۰۰۰۰ دلار رایج است. با این حال آلاسکا جمعیّت بسیار کمی دارد (شکل‌های ۱۵–۱۱ و ۱۵–۱۲ را ببینید)، و بنابراین سطوح درآمد در آلاسکا فقط برای بخش کوچکی از جمعیّت ایالات متحده اعمال می‌شود. اکثریت شهرستان‌های ایالات متحده، که تقریباً همگی پرجمعیّت‌تر از شهرستان‌های آلاسکا هستند، میانهٔ درآمدی کمتر از ۶۰۰۰۰ دلار دارند.

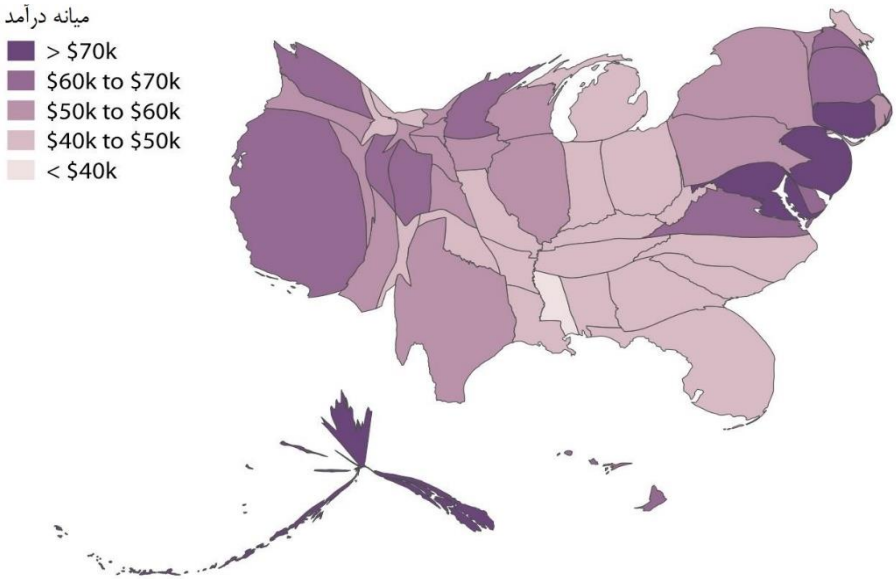


شکل ۱۵–۱۴. میانهٔ درآمد در هر ایالت ایالات متحده، که به صورت نقشه ناحیه–مقدار نشان داده شده است. این نقشه از نظر بصری تحت تسلط ایالت آلاسکا است که میانهٔ درآمد بالا امّا تراکم جمعیّت بسیار پایینی دارد. در عین حال، ایالت‌های پردرآمد و پرجمعیّت در ساحل شرقی در این نقشه چندان برجسته به نظر نمی‌رسند. در مجموع، این نقشه ترسیم ضعیفی از توزیع درآمد در ایالات متحده ارائه می‌دهد، و بنابراین آن را به عنوان «بد» برچسب‌گذاری کرده‌ایم. منبع داده: نظرسنجی پنج ساله جامعه آمریکا در سال ۲۰۱۵

نقشه آماری[1]

هر ترسیم نقشه مانندی لازم نیست برای مفید بودن، لزوماً از نظر جغرافیایی دقیق باشد. به عنوان مثال، مشکل شکل ۱۵–۱۴ این است که برخی از ایالت‌ها منطقه نسبتاً بزرگی را اشغال می‌کنند امّا جمعیّت کمی دارند، در حالی که برخی دیگر منطقه کوچکی را اشغال می‌کنند امّا تعداد زیادی ساکن دارند. چه می‌شود اگر ایالت‌ها را تغییر شکل دهیم تا اندازه آن‌ها متناسب با تعداد ساکنان آن‌ها باشد؟ چنین نقشه اصلاح شده‌ای نقشه آماری نامیده می‌شود و شکل ۱۵–۱۵ نشان می‌دهد که چگونه می‌تواند برای مجموعه داده میانهٔ درآمد به کار رود. ما هنوز هم می‌توانیم تک‌تک ایالت‌ها را تشخیص دهیم، امّا همچنین می‌بینیم که چگونه تعدیل تعداد جمعیّت تغییرات مهمی را ایجاد کرده است. ایالت‌های ساحل شرقی، فلوریدا و کالیفرنیا از نظر وسعت بسیار رشد کرده‌اند، در حالی که سایر ایالت‌های غربی و آلاسکا کوچک شده‌اند.

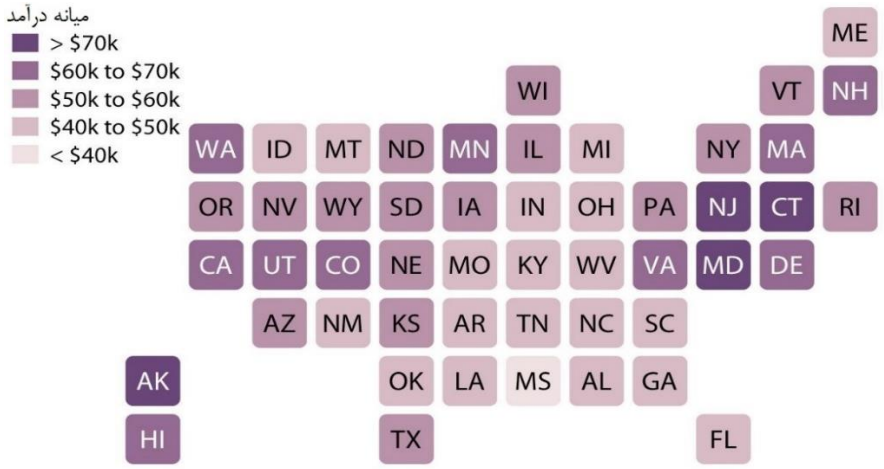
1. Cartograms



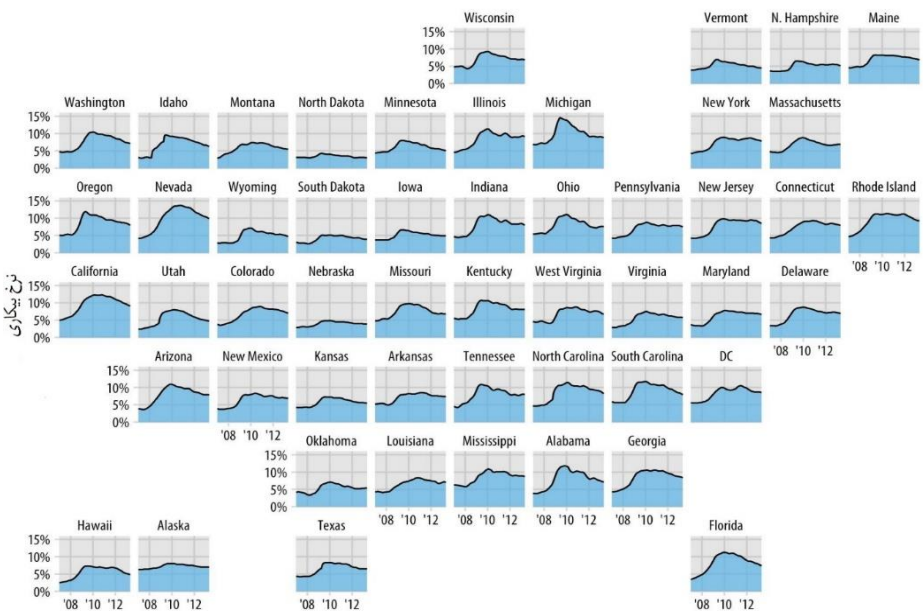
شکل ۱۵–۱۵. میانهٔ درآمد در هر ایالت ایالات متحده، که به صورت نقشه آماری نشان داده شده است. شکل هر ایالت به گونه‌ای اصلاح شده است که مساحت آن‌ها متناسب با تعداد ساکنان آن‌ها باشد. منبع داده: پیمایش پنج ساله جامعه آمریکا در سال ۲۰۱۵

به‌عنوان جایگزینی برای یک نقشه آماری با اشکال تحریف‌شده، می‌توانیم خیلی ساده‌تر یک نقشه حرارتی آماری رسم کنیم، که در آن هر ایالت با یک مربع رنگی نشان داده شده است (شکل ۱۵–۱۶). در حالی که این تصویر تعداد جمعیّت در هر ایالت را تصحیح نمی‌کند، و در نتیجه ایالت‌های پرجمعیّت‌تر را کمتر از مقدار واقعی و ایالت‌های کم جمعیّت را بیش از مقدار واقعی نشان می‌دهد، حداقل با همه ایالت‌ها به یک صورت برخورد می‌کند و به آن‌ها بر اساس شکل یا اندازهٔ آن‌ها وزن نمی‌دهد.

در نهایت، می‌توانیم با قرار دادن نمودارهای جداگانه در محل هر ایالت، نقشه‌های آماری پیچیده‌تری ترسیم کنیم. به عنوان مثال، اگر بخواهیم تکامل نرخ بیکاری را در طول زمان برای هر ایالت ترسیم کنیم، می‌توان نمودار جداگانه برای هر ایالت ترسیم کرد و سپس نمودارها را بر اساس موقعیّت‌های نسبی تقریبی ایالات نسبت به یکدیگر مرتّب کرد (شکل ۱۵–۱۷). برای فردی که با جغرافیای ایالات متحده آشنایی دارد، این ترتیب ممکن است یافتن نمودارها را برای ایالت‌های خاص آسان‌تر از مرتّب کردن آنها، به عنوان مثال، به ترتیب حروف الفبا کند. علاوه بر این، می‌توان انتظار داشت که ایالت‌های مجاور هم الگوهای مشابهی داشته باشند، و شکل ۱۵–۱۷ نشان می‌دهد که واقعاً چنین است.



شکل ۱۵–۱۶. میانهٔ درآمد در هر ایالت ایالات متحده، که به صورت نقشه حرارتی آماری نشان داده شده است. هر ایالت با مربعی با اندازه یکسان نشان داده شده است و مربع‌ها بر اساس موقعیّت تقریبی هر ایالت نسبت به ایالت‌های دیگر مرتّب شده‌اند. این نمایش وزن بصری یکسانی را به هر ایالت می‌دهد. منبع داده: پیمایش پنج ساله جامعه آمریکا در سال ۲۰۱۵.



شکل ۱۵–۱۷. نرخ بیکاری منتهی به بحران مالی ۲۰۰۸ و پس از آن، بر اساس ایالت. هر پانل نرخ بیکاری را برای یک ایالت، از جمله ناحیه کلمبیا (DC) از ژانویه ۲۰۰۷ تا می ۲۰۱۳ نشان می‌دهد. خطوط شبکه عمودی نشانگر ژانویه ۲۰۰۸، ۲۰۱۰، و ۲۰۱۲ هستند. ایالاتی که از نظر جغرافیایی به هم نزدیک هستند روندهای مشابهی را در نرخ بیکاری نشان می‌دهند. منبع داده‌ها: اداره آمار کار ایالات متحده.

فصل ۱۶

ترسیم عدم قطعیّت

یکی از چالش برانگیزترین جنبه‌های ترسیم داده، نمایش عدم قطعیّت است. هنگامی که یک نقطه داده را در یک مکان خاص می‌بینیم، معمولاً آن را به عنوان نمایشی دقیق از مقدار داده واقعی تفسیر می‌کنیم. تصوّر اینکه یک نقطه داده واقعاً در جایی باشد که ترسیم نشده است دشوار است. با این حال، این حالت در ترسیم داده‌ها به وفور وجود دارد. تقریباً هر مجموعه داده‌ای که ما با آن کار می‌کنیم دارای سطوحی از عدم قطعیّت است، و اینکه چگونه نحوه نمایش این عدم قطعیّت را انتخاب کنیم می‌تواند تفاوت عمده‌ای در میزان دقّت مخاطبان ما از درک معنای داده‌ها ایجاد کند.

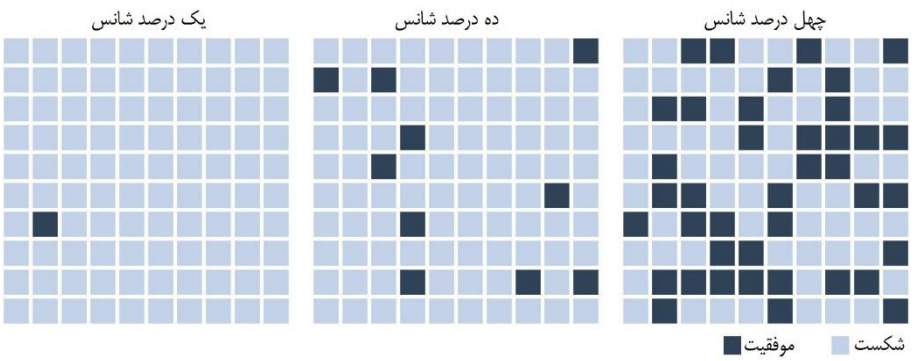
دو رویکرد متداول برای نشان دادن عدم قطعیّت، میلهٔ خطا و نوار اطمینان هستند. این رویکردها در زمینه انتشارات علمی توسعه یافته‌اند و برای تفسیر صحیح نیاز به مقداری دانش تخصصی دارند، امّا دقیق و کارآمد هستند. برای مثال، با استفاده از میلهٔ خطا، می‌توانیم عدم قطعیّت تخمین‌های متغیّرهای مختلف را در یک نمودار نشان دهیم. با این حال، برای یک مخاطب عادی، رویکردهای ترسیم که یک تصوّر شهودی قوی از عدم قطعیّت ایجاد می‌کنند، ارجح هستند، حتی اگر به قیمت کاهش دقّت ترسیم داده یا تصاویری با چگالی کم داده باشد. گزینه‌های موجود عبارتند از قاب‌بندی فراوانی، که در آن به صراحت حالت‌های ممکن مختلف را با نسبت‌های تقریبی یا تصاویری که بین حالت‌های مختلف چرخش می‌کنند، ترسیم می‌کنیم.

قاب‌بندی احتمالات به صورت فراوانی

قبل از اینکه بتوانیم درباره چگونگی ترسیم عدم قطعیّت بحث کنیم، باید آن را تعریف کنیم. ما به راحتی می‌توانیم مفهوم عدم قطعیّت را در زمینه رویدادهای آینده درک کنیم. اگر بخواهیم سکه‌ای را بیاندازیم، از قبل نمی‌دانیم نتیجه چه خواهد بود. نتیجه نهایی قطعی نیست. با این حال، عدم قطعیّت در مورد رویدادهای گذشته نیز صدق می‌کند. مثلاً اگر دیروز دقیقاً دو مرتبه، یک بار در ساعت ۸ صبح و یک بار در ساعت ۴ بعد از ظهر، از پنجره آشپزخانه خود به بیرون نگاه کرده باشید، و ماشین قرمزی را دیده باشید که در ساعت ۸ صبح در آن طرف خیابان پارک شده است امّا در ساعت ۴ بعد از ظهر آنجا نبوده باشد، می‌توانید نتیجه بگیرید که ماشین در بازه زمانی از ساعت ۸ و قبل ساعت ۴ از آنجا رفته است، امّا دقیقاً نمی‌دانید چه زمانی. زمان حرکت آن ماشین در این هشت ساعت ممکن است ساعت ۸:۰۱ صبح، ۹:۳۰ صبح، ۲ بعد از ظهر یا هر زمان دیگری باشد.

از نظر ریاضیات، عدم قطعیّت را با استفاده از مفهوم احتمال بررسی می‌کنیم. تعریف دقیق احتمال پیچیده و فراتر از محدوده این کتاب است. با این حال، می‌توانیم با موفقیت در مورد احتمالات بدون درک همه پیچیدگی‌های ریاضی آن بحث کنیم. برای بسیاری از مشکلات مربوط به ارتباط عملی کافی است که در مورد فراوانی نسبی فکر کنیم. فرض کنید نوعی آزمایش تصادفی انجام می‌دهید، مانند انداختن سکه یا ریختن تاس، و به دنبال یک نتیجه خاص هستید (مثلاً، شیر یا آمدن عدد شش). شما می‌توانید این نتیجه را موفقیت، و هر نتیجهٔ دیگری را شکست بنامید. سپس، احتمال موفقیت تقریباً کسری از دفعاتی است که اگر آزمایش تصادفی را بارها و بارها تکرار کنید، آن نتیجه را مشاهده خواهید کرد. به عنوان مثال، اگر یک پیامد خاص با احتمال ۱۰ درصد رخ دهد، انتظار داریم که در بین بسیاری از کارآزمایی‌های مکرر، آن نتیجه تقریباً در ۱ مورد از هر ۱۰ مورد مشاهده شود.

ترسیم یک احتمال منفرد مشکل است. شانس برنده شدن در بخت آزمایی یا شانس آوردن عدد شش را چگونه نمایش می‌دهید؟ در هر دو حالت احتمال مورد نظر یک عدد منفرد است. لذا می‌توانیم آن عدد را به عنوان یک مقدار در نظر بگیریم و با استفاده از هر یک از روش‌های مورد بحث در فصل ۶، مانند نمودار میله‌ای یا نمودار نقطه‌ای، آن را نمایش دهیم، امّا نتیجه چندان مفید نخواهد بود. اکثر افراد درک شهودی‌ای از چگونگی تبدیل یک مقدار احتمال به واقعیّت تجربه شده ندارند. نشان دادن مقدار احتمال به صورت یک میله یا به صورت نقطه‌ای که روی یک خط قرار می‌گیرد به حل این مشکل کمکی نمی‌کند.

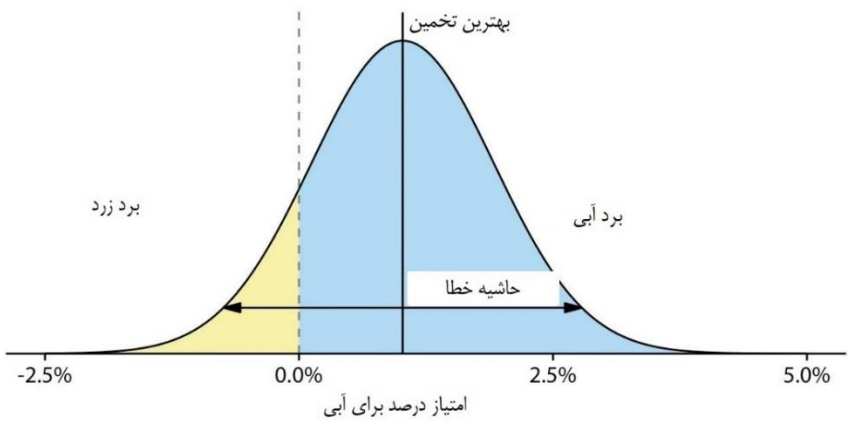


شکل ۱۶–۱. ترسیم احتمال به صورت فراوانی. در هر شبکه ۱۰۰ مربع وجود دارد و هر مربع نشان‌دهندهٔ موفقیت یا شکست در آزمایشات تصادفی است. احتمال ۱ درصد موفقیت با ۱ مربع تاریک و ۹۹ مربع روشن، احتمال موفقیت ۱۰ درصد با ۱۰ مربع تاریک و ۹۰ مربع روشن و احتمال موفقیت ۴۰ درصد با ۴۰ مربع تاریک و ۶۰ مربع روشن نشان داده شده است. با قرار دادن تصادفی مربع‌های تیره در میان مربع‌های روشن، می‌توانیم تصوّری بصری از تصادفی بودن ایجاد کنیم که بر عدم قطعیّت نتیجه یک آزمایش واحد تأکید می‌کند.

ما می‌توانیم با ایجاد نموداری که بر جنبه فراوانی و غیرقابل پیش‌بینی بودن آزمایش تصادفی تأکید می‌کند، مفهوم احتمال را ملموس کنیم، برای مثال با رسم مربع‌هایی با رنگ‌های مختلف در یک آرایش تصادفی. در شکل ۱۶–۱، از این روش برای ترسیم سه احتمال مختلف، ۱ درصد احتمال موفقیت، ۱۰ درصد احتمال موفقیت و ۴۰ درصد احتمال موفقیت استفاده می‌کنیم. برای خواندن این شکل، تصوّر کنید قبل از اینکه بتوانید ببینید کدام یک از مربع‌ها تیره و کدام یک روشن خواهند بود، وظیفهٔ انتخاب مربع تیره به شما داده می‌شود (مثلاً با چشمان بسته به انتخاب مربع فکر کنید). به طور شهودی، احتمالاً می‌دانید که بعید به نظر می‌رسد تا یک مربع تیره را در سناریو احتمال ۱ درصد انتخاب کنید. به طور مشابه، انتخاب یک مربع تیره در سناریو احتمال ۱۰ درصد بسیار کم است. با این حال، در مورد سناریوی احتمال ۴۰ درصد، احتمال چندان بد به نظر نمی‌رسد. این سبک از ترسیم، که در آن نتایج بالقوهٔ خاصی را نشان می‌دهیم، ترسیم پیامد گسسته نامیده می‌شود و عمل تجسّم یک احتمال به صورت فراوانی، قاب‌بندی فراوانی نامیده می‌شود. ما ماهیت احتمالی یک پیامد را بر حسب تناوب‌های نتایجی که به راحتی قابل درک است، قاب‌بندی می‌کنیم.

اگر ما فقط به دو پیامد مجزا علاقه‌مند باشیم، مانند موفقیت یا شکست، شکل ۱۶–۱ کارایی خوبی دارد. با این حال، ما اغلب با حالت‌های پیچیده‌تری سر و کار داریم که در آن پیامد یک آزمایش تصادفی یک متغیّر عددی است. یکی از حالت‌های رایج پیش‌بینی‌های انتخاباتی است که در آن ما نه تنها می‌خواهیم بدانیم چه کسی پیروز می‌شود، بلکه می‌خواهیم مقدار آن را نیز

بدانیم. یک مثال فرضی از انتخابات آینده با دو حزب، زرد و آبی را در نظر بگیرید. فرض کنید از رادیو می‌شنوید که پیش‌بینی می‌شود حزب آبی نسبت به حزب زرد ۱ واحد درصد با حاشیه خطای ۱٫۷۶ واحد درصد برتری دارد. این اطّلاعات در مورد نتیجهٔ احتمالی انتخابات به شما چه می‌گوید؟ شنیدن «حزب آبی پیروز خواهد شد» طبیعت انسان است، امّا واقعیّت پیچیده‌تر است. اوّل، و مهمتر از همه، طیف وسیعی از نتایج ممکن وجود دارد. حزب آبی می‌تواند با برتری دو واحد درصدی برنده شود، یا حزب زرد می‌تواند با برتری نیم واحد درصدی به پیروزی برسد. دامنهٔ نتایج ممکن با احتمالات مرتبط با آن‌ها را توزیع احتمال می‌گویند و می‌توانیم آن را به صورت منحنی صافی ترسیم کنیم که در محدودهٔ نتایج ممکن افزایش و سپس کاهش می‌یابد (شکل ۱۶-۲). هر چه منحنی برای یک نتیجه خاص بلندتر باشد، احتمال آن نتیجه بیشتر است. توزیع‌های احتمال ارتباط نزدیکی با هیستوگرام‌ها و کرنل‌های چگالی مورد بحث در فصل ۷ دارند، و ممکن است بخواهید آن فصل را دوباره بخوانید تا حافظهٔ خود را تازه کنید.

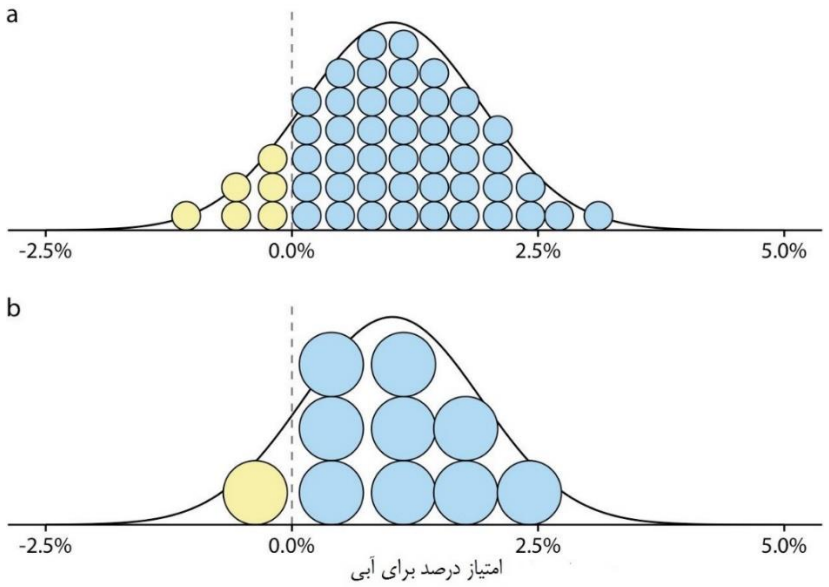


شکل ۱۶-۲ پیش‌بینی فرضی نتیجه یک انتخابات. پیش‌بینی می‌شود که حزب آبی تقریباً با ۱ واحد درصد برتری بر حزب زرد پیروز شود (با برچسب «بهترین تخمین»)، امّا این پیش‌بینی دارای حاشیهٔ خطا است (در اینجا رسم شده تا ۹۵ درصد از نتایج احتمالی را پوشش دهد، ۱٫۷۶ واحد درصد در هر جهت از نقطه بهترین برآورد). ناحیهٔ آبی رنگ که ۸۷٫۱ درصد از کل است، نشان‌دهندهٔ همه نتایجی است که در آن حزب آبی برنده می‌شود. به همین ترتیب، ناحیهٔ زرد رنگ که ۱۲٫۹ درصد از کل است، نشان‌دهندهٔ همه نتایجی است که در آن‌ها حزب زرد برنده می‌شود. در این مثال، حزب آبی ۸۷ درصد احتمال برنده شدن در انتخابات را دارد.

با انجام محاسبات، می‌توانیم ببینیم که برای مثال فرضی، احتمال برنده شدن حزب زرد ۱۲٫۹ درصد است. بنابراین، احتمال برنده شدن زرد کمی بهتر از حالت احتمال ۱۰ درصدی نشان داده شده در شکل ۱۶-۱ است. اگر طرفدار حزب آبی هستید، زیاد نگران نباشید، امّا حزب زرد

بـه انـدازه کـافی احتمـال برنـده شـدن دارد و ممکـن اسـت موفـق باشـد. اگـر شـکل ۱۶–۲ را بـا شـکل ۱۶–۱ مقایسـه کنیـد، متوجّـه می‌شـوید کـه شـکل ۱۶–۱ درک بهتـری از عـدم قطعیّـت در پیامـد را ایجـاد می‌کنـد، هرچنـد کـه منـاطق رنگـی در شـکل ۱۶–۲ بـه طـور دقیـق احتمـال برنـده شدن آبی یا زرد را نشـان دهنـد. ایـن قـدرت ترسـیم پیامـد گسسـته اسـت. تحقیقـات در حـوزهٔ ادراک انسان نشان می‌دهـد کـه مـا در درک، شـمارش و قضـاوت در خصـوص فراوانـی نسـبی اجسـام مجـزا — تـا زمـانی کـه عـدد کلـی آن‌هـا خیلـی بـزرگ نباشـد–بسـیار بهتـر از قضـاوت در مـورد انـدازه‌های نسبی نواحی عمل می‌کنیم.

مـا می‌تـوانیم ماهیـت پیامـد گسسـته شـکل ۱۶–۱ را بـا یـک توزیـع پیوسـته ماننـد شـکل ۱۶–۲ بـا رسم نمودار نقطه‌ای چنـدکی ترکیـب کنـیم. در نمـودار نقطـه‌ای چنـدکی، مسـاحت کـل زیـر منحنـی را به واحدهایی با اندازه مسـاوی تقسـیم مـی‌کنیم و هـر واحـد را بـه صـورت دایـره می‌کشـیم. سـپس دایره‌ها را طـوری روی هـم مـی‌چینیم کـه آرایـش آن‌هـا تقریبـاً منحنـی توزیـع اصـلی را نشـان دهـد (شکل ۱۶–۳).



شکل ۱۶–۳. نمودارهای نقطه‌ای چندکی برای نمایش توزیع نتیجهٔ انتخابات در شکل ۱۶–۲ (الف) توزیع صاف با ۵۰ نقطه که هر کدام نشان‌دهندهٔ احتمال ۲ درصد است. بنابراین، ۶ نقطه زرد معادل احتمال ۱۲ درصد است، که تقریباً نزدیک به مقدار واقعی ۱۲٫۹ درصد است. (ب) توزیع صاف با ۱۰ نقطه که هر کدام احتمال ۱۰ درصد را نشان می‌دهد. بنابراین، ۱ نقطه زرد معادل احتمال ۱۰ درصد است که همچنان به مقدار واقعی نزدیک است. خواندن نمودارهای نقطه‌ای چندکی با تعداد نقاط کمتر آسان‌تر است، بنابراین در این مثال، نسخه ۱۰ نقطه‌ای ممکن است به نسخه ۵۰ نقطه‌ای ترجیح داده شود.

به عنوان یک اصل کلی، نمودارهای نقطه‌ای چندکی باید از تعداد کم تا متوسط نقطه استفاده کنند. اگر نقاط بیش از حد باشد، ما تمایل داریم آن‌ها را به عنوان یک طیف ببینیم و نه به عنوان واحدهای مجزا و گسسته. این مساله مزیت‌های نمودارهای گسسته را منتفی می‌کند. شکل ۱۶–۳ انواع با ۵۰ نقطه (شکل ۱۶–۳ الف) و با ۱۰ نقطه (شکل ۱۶–۳ ب) را نشان می‌دهد. در حالی که نسخهٔ با ۵۰ نقطه توزیع احتمال واقعی را با دقّت بیشتری نشان می‌دهد، تعداد نقاط آن قدر زیاد است که نمی‌توان به راحتی تک‌تک آن‌ها را تشخیص داد. نسخهٔ با ۱۰ نقطه سریعتر احتمال نسبی برنده شدن آبی یا زرد را منتقل می‌کند. یک ایراد نسخهٔ ۱۰ نقطه‌ای ممکن است این باشد که خیلی دقیق نیست. در حقیقت احتمال برد زردها ۲/۹ واحد درصد کمتر نشان داده می‌شود. با این حال، اغلب ارزشمند است که کمی دقّت را به قیمت درک دقیق‌تر از ترسیم حاصل، به ویژه هنگام برقراری ارتباط با یک مخاطب عادی، مبادله کنیم. ترسیمی که از نظر ریاضی درست باشد امّا به درستی درک نشود، در عمل چندان مفید نیست.

ترسیم عدم قطعیّت برای تخمین‌های نقطه‌ای

در شکل ۱۶–۲، «بهترین تخمین» و «حاشیه خطا» را نشان دادیم، امّا توضیحی برای اینکه این مقادیر دقیقاً چه هستند یا چگونه به دست می‌آیند ارائه ندادیم. برای درک بهتر آن‌ها، باید بر مفاهیم اولیهٔ نمونه‌گیری مرور سریعی داشته باشیم. در آمار، هدف اصلی این است که با نگاه کردن به بخش کوچکی از جهان، چیزی در مورد جهان بیاموزیم. در ادامهٔ مثال انتخاباتی، فرض کنید حوزه‌های انتخاباتی زیادی وجود دارد و شهروندان هر ناحیه قرار است به حزب آبی یا زرد رأی دهند. ممکن است بخواهیم نحوه رأی دادن هر ناحیه و همچنین میانگین کلی رأی در مناطق (میانگین) را پیش‌بینی کنیم. برای پیش‌بینی قبل از انتخابات، نمی‌توانیم از تک‌تک شهروندان در هر ناحیه درباره نحوه رأی دادن آن‌ها نظرسنجی کنیم. در عوض، ما باید نمونه‌ای از شهروندان را در زیرمجموعه‌ای از مناطق نظرسنجی کنیم و از آن داده‌ها برای رسیدن به بهترین حدس استفاده کنیم. در زبان آماری به مجموع آرای ممکن همه شهروندان در همه مناطق، جمعیّت گفته می‌شود و زیرمجموعه شهروندان و/یا مناطقی که ما نظرسنجی می‌کنیم نمونه است. جمعیّت نمایانگر وضعیّت واقعی زیربنایی جهان است و نمونه پنجره ما به آن جهان است.

ما معمولاً به مقادیر خاصی علاقه‌مندیم که خصوصیات مهم جمعیّت را خلاصه می‌کند. در مثال انتخاباتی، این‌ها می‌توانند میانگین نتیجهٔ رأی در مناطق یا انحراف معیار بین نتایج مناطق

باشـد. کمیّت‌هـایی کـه جمعیّـت را توصـیف می‌کننـد، پـارامتر نامیـده می‌شـوند و معمـولاً قابـل شناخت نیستند. با این حـال، مـا می‌تـوانیم از یـک نمونـه بـرای حـدس زدن مقـادیر واقعـی پارامترهـا اسـتفاده کنـیم و متخصصـین آمـار ایـن حـدس‌ها را بـا عنـوان تخمـین می‌نامنـد. میـانگین نمونـه تخمینی برای میانگین جامعه می‌باشد کـه یـک پـارامتر است. بـه‌تخمـین مقـادیر پارامترهـای منفـرد، تخمین نقطه‌ای نیز گفته می‌شود، زیرا هر یک را می‌توان با یک نقطه در یک خط نشان داد.

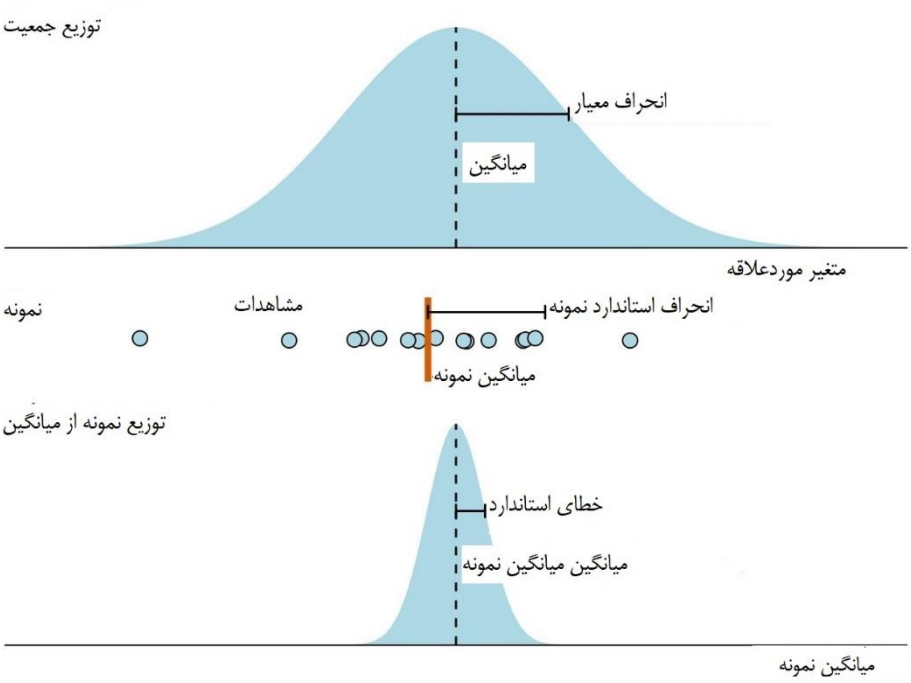
شکل ۱۶-۴ نشان می‌دهد که چگونـه ایـن مفـاهیم کلیـدی بـا یکـدیگر مـرتبط هسـتند. متغیّـر مـورد نظـر (بـه عنـوان مثـال، نتیجـه رأی در هـر منطقـه) دارای تـوزیعی در جمعیّـت اسـت، بـا میـانگین و انحـراف معیـار مشـخّص در سـطح جمعیّـت. یـک نمونـه شـامل مجموعـه‌ای از مشـاهدات خـاص خواهـد بـود. بـه تعـداد مشـاهدات مجـزا در نمونـه، حجـم نمونـه اطـلاق می‌شـود. بـرای نمونـه می‌توانیم میـانگین نمونـه و انحـراف معیـار را محاسـبه کنـیم و این‌هـا عمومـاً بـا میـانگین و انحـراف معیـار جامعـه متفـاوت خواهنـد بـود. در نهایـت، می‌تـوانیم توزیـع نمونـه‌گیری را تعریـف کنـیم، کـه توزیـع تخمین‌هـایی اسـت کـه اگـر فرآینـد نمونـه‌گیری را بارهـا تکـرار کنـیم، بـه دسـت مـی‌آوریم. گسـترهٔ توزیـع نمونـه‌گیری خطـای اسـتاندارد[1] نامیـده می‌شـود و بـه مـا می‌گویـد بـرآورد مـا چقـدر دقیق است. به عبـارت دیگـر، خطـای اسـتاندارد انـدازه‌گیری عـدم قطعیّـت مـرتبط بـا بـرآورد پـارامتر را ارائه می‌دهد. بـه عنـوان یـک قاعـدهٔ کلـی، هـر چـه حجـم نمونـه بزرگتـر باشـد، خطـای اسـتاندارد کوچکتر است و بنابراین قطعیّت تخمین بیشتر خواهد بود.

بسـیار ضـروری اسـت کـه انحـراف معیـار و خطـای اسـتاندارد را اشـتباه نگیـریم. انحـراف معیـار خصوصیت جمعیّت است و بـه مـا می‌گویـد کـه چقـدر بـین مشـاهدات منفـردی کـه می‌تـوانیم انجـام دهـیم، پراکنـدگی وجـود دارد. بـرای مثـال، اگـر جمعیّـت منـاطق رأی‌دهنـده را در نظـر بگیـریم، انحـراف معیـار بـه مـا می‌گویـد کـه منـاطق چقـدر بـا یکـدیگر متفـاوت هسـتند. در مقابـل، خطـای اسـتاندارد بـه مـا می‌گویـد کـه چقـدر یـک پـارامتر را دقیـق تخمـین زده‌ایـم. اگـر بخـواهیم میـانگین نتیجه رأی‌گیری را در تمام منـاطق تخمـین بـزنیم، خطـای اسـتاندارد بـه مـا می‌گویـد کـه بـرآورد مـا برای میانگین چقدر دقیق است.

همـه متخصصـین آمـار از نمونـه‌ها بـرای محاسـبه تخمـین پارامترهـا و عـدم قطعیّـت آن‌هـا اسـتفاده می‌کنند. با این حال، آن‌هـا در نحـوه برخـورد بـا ایـن محاسـبات بـه دو دسـته قضـیهٔ بیـزین[2] و قضـیهٔ فراوانی‌گرایـان[3] تقسـیم می‌شـوند. پیـروان قضـیهٔ بیـزین تصـوّر می‌کننـد کـه دانـش قبلـی در مـورد

1. standard error
2. Bayesians
3. Frequentists

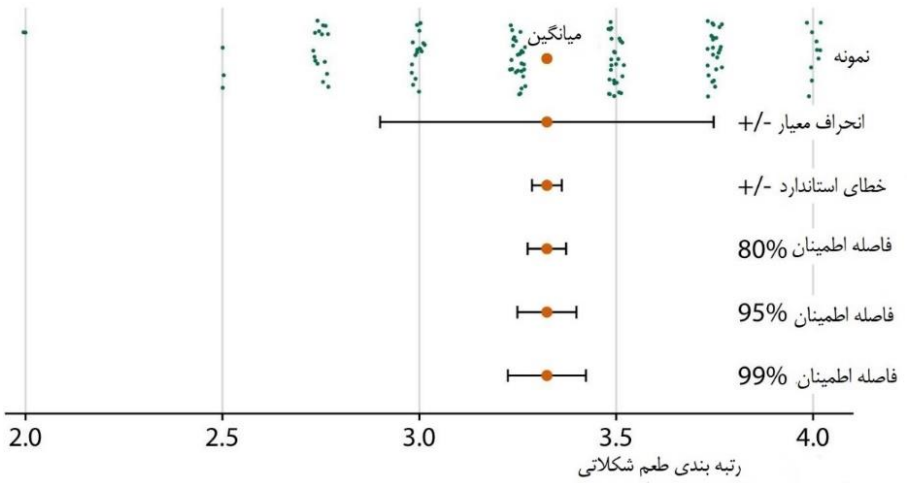
جهـان دارنـد و از نمونـه بـرای بروزرسـانی ایـن دانـش اسـتفاده می‌کننـد. در مقابـل، فراوانـی‌گرایـان سعی می‌کننـد بـدون داشـتن دانـش قبلـی، اظهـارات دقیقـی دربـارهٔ جهـان ارائـه دهنـد. خوشـبختانه، وقتی صحبت از ترسـیم عـدم قطعیّـت بـه میـان می‌آیـد، پیـروان قضـیهٔ بیـزین و فراوانی‌گـرا معمـولاً می‌تواننـد از راهبردهـای یکسـانی اسـتفاده کننـد. در ادامـه، ابتـدا رویکـرد فراوانی‌گرایـی را مـورد بحث قرار می‌دهیم و سپس چند موضـوع خـاص را کـه منحصـر بـه حـوزه قضـیهٔ بیـزین اسـت، شـرح خواهیم داد.



شکل ۱۶–۴. مفاهیم کلیدی نمونه‌گیری آماری. متغیّر مورد نظری که ما در حال مطالعه آن هستیم دارای توزیع واقعی در جامعه با میانگین و انحراف معیار مشخّص در سطح جمعیّت است. هر نمونه محدودی از آن متغیّر دارای میانگین و انحراف معیاری است که با پارامترهای جامعه متفاوت است. اگر به طور مکرر نمونه‌برداری کنیم و هر بار میانگین را محاسبه کنیم، میانگین به دست آمده بر اساس توزیع نمونه‌گیری میانگین توزیع می‌شود. خطای استاندارد اطلاعاتی در مورد عرض توزیع نمونه‌گیری ارائه می‌کند، که به ما می‌گوید که چقدر پارامتر مورد نظر را دقیق تخمین می‌زنیم (در اینجا، میانگین جمعیّت).

فراوانی‌گرایان معمـولاً عـدم قطعیّـت را بـا میلـه‌های خطـا نمـایش می‌دهنـد. در حـالی کـه میلـه‌های خطا می‌توانند بـرای نمـایش عـدم قطعیّـت مفیـد باشـند، همـانطور کـه قـبلاً در فصـل ۹ (شـکل ۹–۱ را ببینید) به آن اشاره کردیم، بـدون مشـکل نیـز نیسـتند. خواننـدگان بـه راحتـی ممکـن اسـت دربـاره

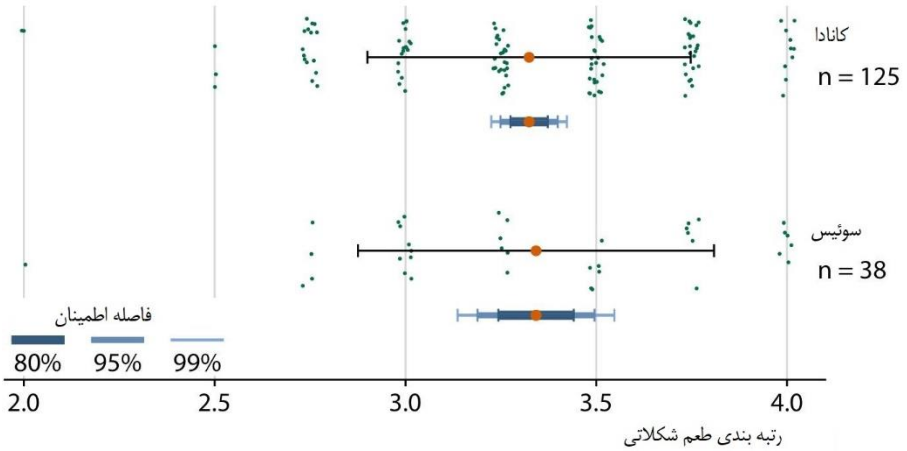
آنچه که در میلهٔ خطا نشان می‌دهد گیج شوند. برای مشخّص کردن این مشکل، در شکل ۱۶–۵ پنج کاربرد مختلف از میله‌های خطا را برای یک مجموعه داده واحد نشان می‌دهیم. مجموعهٔ داده شامل رتبه‌بندی‌های تخصصی در مقیاس ۱ تا ۵ از بسته‌های شکلات است که در کشورهای مختلف تولید شده‌اند. برای شکل ۱۶–۵، تمام رتبه‌بندی‌های شکلات تولید شده در کانادا استخراج شده است. در زیر نمونه، که به صورت نمودار نواری از نقاط لغزنده نشان داده شده است، میانگین نمونه بعلاوه/منهای انحراف معیار نمونه، میانگین نمونه بعلاوه/ منهای خطای معیار و فاصله‌های اطمینان ۸۰ درصد، ۹۵ درصد و ۹۹ درصد را مشاهده می‌کنیم. هر پنج میلهٔ خطا از تنوّع در نمونه مشتق شده‌اند، و همگی از نظر ریاضی مرتبط هستند، امّا معانی متفاوتی دارند. از نظر بصری نیز کاملاً متمایز هستند.



شکل ۱۶–۵. رابطه بین نمونه، میانگین نمونه، انحراف معیار، خطای استاندارد، و فواصل اطمینان، در نمونه‌ای از رتبه‌بندی بسته‌های شکلات. مشاهدات (نشان داده شده به صورت نقاط سبز پراکنده) که نمونه را تشکیل می‌دهند، رتبه‌بندی کارشناسی برای ۱۲۵ بستهٔ شکلات از تولیدکنندگان کانادایی را نشان می‌دهند که در مقیاسی از ۱ (ناخوشایند) تا ۵ (بسیار عالی) رتبه‌بندی شده‌اند. نقطهٔ نارنجی بزرگ نشان‌دهندهٔ میانگین رتبه‌بندی‌ها است. میله‌های خطا، از بالا به پایین، نشان‌دهندهٔ دو برابر انحراف معیار، دو برابر خطای استاندارد (انحراف معیار میانگین)، و فواصل اطمینان ۸۰، ۹۵، و ۹۹ درصد از میانگین را نشان می‌دهند. منبع داده: Brady Brelinski، انجمن شکلات منهتن.

هر زمان عدم قطعیّت را با میلهٔ خطا نمایش می‌دهید، باید مشخّص نمایید که این میلهٔ خطا چه کمیّت و/یا چه فاصله اطمینانی را نمایش می‌دهد.

خطای استاندارد تقریباً با تقسیم انحراف معیار نمونه بر جذر حجم نمونه به دست می‌آید و فواصل اطمینان با ضرب خطای معیار در مقادیر کوچک و ثابت محاسبه می‌شود. به عنوان مثال، فاصله اطمینان ۹۵ درصد تقریباً دو برابر خطای استاندارد در هر جهت از میانگین گسترش می‌یابد. بنابراین، نمونه‌های بزرگتر معمولاً خطاهای استاندارد کوچکتر و فواصل اطمینان باریک‌تری دارند، حتّی اگر انحراف معیار آن‌ها یکسان باشد. ما می‌توانیم این موضوع را در مقایسه رتبه‌بندی شکلات‌های تولید شده در کانادا با شکلات‌های تولید شده در سوئیس ملاحظه کنیم (شکل ۱۶–۶). میانگین امتیاز و انحراف معیار نمونه بین شکلات‌های کانادایی و سوئیسی مشابه هستند، امّا این اعداد حاصل از امتیازدهی برای ۱۲۵ شکلات کانادایی و تنها ۳۸ شکلات سوئیسی است، در نتیجه فاصله‌های اطمینان حول میانگین در مورد شکلات‌های سوئیسی بسیار پهن‌تر است.

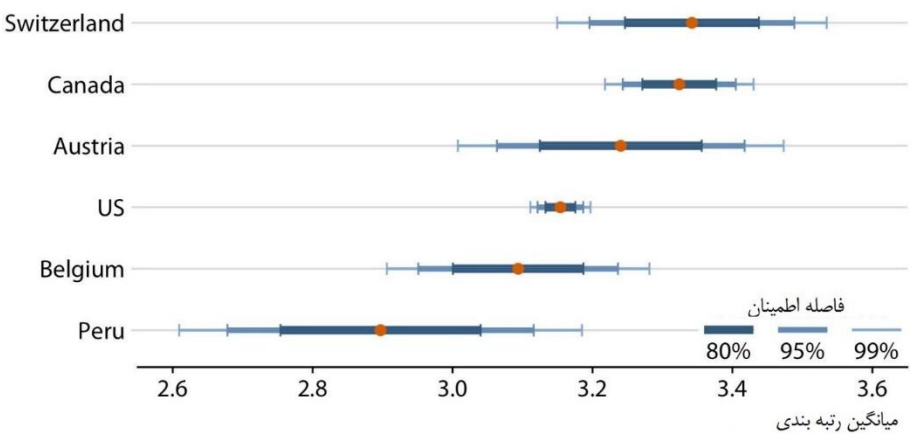


شکل ۱۶–۶. در حجم نمونه‌های کوچکتر، فواصل اطمینان پهن‌تر می‌شود. شکلات‌های کانادایی و سوئیسی دارای میانگین رتبه‌بندی و انحراف معیار مشابهی هستند (که با میله‌های خطای سیاه ساده نشان داده شده‌اند). با این حال، تعداد شکلات‌های رتبه‌دهی شده کانادایی بیش از سه برابر تعداد شکلات‌های رتبه‌دهی شده سوئیسی است، و بنابراین فواصل اطمینان (که با میله‌های خطا با رنگ‌ها و ضخامت‌های مختلف که روی هم کشیده شده، نشان داده شده‌اند) برای میانگین رتبه‌بندی سوئیس به طور قابل‌توجهی پهن‌تر از میانگین رتبه‌بندی کانادا می‌باشد. منبع داده: Brady Brelinski، انجمن شکلات منهتن.

در شکل ۱۶–۶ سه فاصله اطمینان مختلف را با استفاده از رنگ‌های تیره‌تر و خطوط ضخیم‌تر برای فواصلی با سطوح اطمینان پایین‌تر، به طور همزمان نشان می‌دهیم. به این ترسیم‌ها به عنوان میله‌های خطای درجه‌بندی شده اطلاق می‌شود. درجه‌بندی به خواننده کمک می‌کند تا متوجّه شود که طیف وسیعی از احتمالات مختلف وجود دارد. اگر میله‌های خطای ساده (بدون

درجه‌بندی) را به گروهی از افراد نشان دهیم، احتمال کمی وجود دارد که میله‌های خطا را به درستی درک کنند، برای مثال ممکن است آن را به عنوان نمایندهٔ حداقل و حداکثر داده‌ها در نظر گیرند. از طرف دیگر، ممکن است فکر کنند که میله‌های خطا همهٔ محدوده ممکن برای تخمین پارامتر را در بر می‌گیرد، به بیان دیگر تخمین هرگز نمی‌تواند خارج از میله‌های خطا قرار گیرد. این نوع برداشت‌های نادرست، خطاهای تعبیر قطعی نامیده می‌شوند. هر چه بیشتر بتوانیم خطر خطای تعبیر قطعی را به حداقل برسانیم، ترسیم بهتری از عدم قطعیّت خواهیم داشت.

میله‌های خطا کاربردی هستند زیرا به ما امکان می‌دهند تخمین‌های زیادی را به همراه عدم قطعیّت آن‌ها به طور همزمان نشان دهیم. بنابراین، آن‌ها معمولاً در نشریات علمی مورد استفاده قرار می‌گیرند، جایی که هدف اولیه معمولاً انتقال حجم زیادی از اطّلاعات به مخاطبان متخصص است. به عنوان نمونه‌ای از این نوع کاربرد، شکل ۱۶–۷ میانگین رتبه‌بندی شکلات و فواصل اطمینان مرتبط را برای بسته‌های شکلات تولید شده در شش کشور مختلف نشان می‌دهد.



شکل ۱۶–۷. میانگین رتبه‌بندی طعم شکلات و فواصل اطمینان مرتبط برای بسته‌های شکلات از تولیدکنندگان در شش کشور مختلف. منبع داده: Brady Brelinski، انجمن شکلات منهتن.

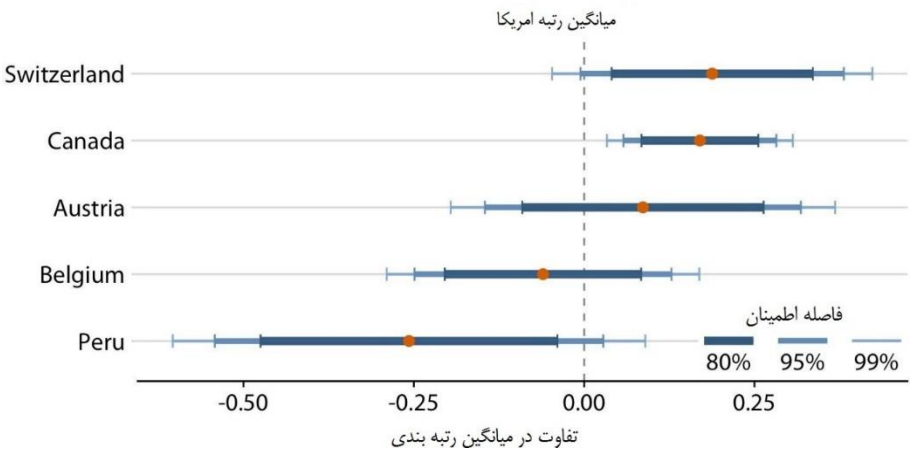
هنگامی که به شکل ۱۶–۷ نگاه می‌کنید، ممکن است بپرسید که این نمودار در مورد تفاوت در میانگین رتبه‌بندی‌ها به ما چه می‌گوید. میانگین رتبه‌بندی‌های شکلات‌های کانادا، سوئیس و اتریش بالاتر از میانگین رتبه‌بندی شکلات‌های امریکا است، امّا با توجّه به عدم قطعیّت در این میانگین رتبه‌بندی‌ها، آیا تفاوت در میانگین‌ها معنی‌دار است؟ کلمه «معنی‌دار» در اینجا

یک اصطلاح فنّی است که توسّط متخصصین آمار استفاده می‌شود. ما تفاوتی را معنی‌دار می‌نامیم که بتوانیم با سطحی از اطمینان این فرض را که تفاوت مشاهده شده ناشی از نمونه‌گیری تصادفی است، رد کنیم. از آنجایی که تنها تعداد محدودی از شکلات‌های کانادایی و امریکایی رتبه‌بندی شده‌اند، ارزیابی‌کنندگان ممکن است به طور تصادفی تعداد بیشتری از شکلات‌های بهتر کانادایی و تعداد کمتری از شکلات‌های بهتر امریکایی را در نظر بگیرند، و این شانس تصادفی ممکن است به صورت برتری نظام‌مند رتبه‌بندی شکلات‌های کانادایی بر شکلات‌های امریکایی به نظر برسد.

ارزیابی اهمیّت از شکل ۱۶–۷ دشوار است، زیرا هم میانگین رتبه‌بندی کانادا و هم میانگین رتبه‌بندی امریکا دارای عدم قطعیّت هستند. هر دو عدم قطعیّت برای پاسخ به این سوال که آیا میانگین‌ها متفاوت هستند یا خیر، اهمیّت دارند. کتاب‌های درسی آمار و آموزش‌های برخط، گاهی قوانین سرانگشتی را دربارهٔ نحوهٔ قضاوت درباره معنی‌داری بر اساس میزان همپوشانی یا عدم همپوشانی میله‌های خطا ارائه می‌کنند. با این حال، این قوانین سرانگشتی قابل اعتماد نیستند و باید از آن‌ها اجتناب شود. روش صحیح برای ارزیابی وجود تفاوت در میانگین رتبه‌بندی، محاسبهٔ فواصل اطمینان برای تفاوت‌ها است. اگر آن فواصل اطمینان صفر را دربرنگیرد، می‌توان گفت که تفاوت در سطح اطمینان مربوطه معنی‌دار است. برای مجموعه داده‌های رتبه‌بندی شکلات، می‌بینیم که فقط شکلات‌های کانادایی به طور معنی‌داری بالاتر از شکلات‌های امریکایی هستند (شکل ۱۶–۸). برای شکلات‌های سوئیسی، فاصله اطمینان ۹۵ درصد تفاوت شامل مقدار صفر می‌شود. بنابراین، احتمال کمی بیشتر از ۵ درصد وجود دارد که تفاوت مشاهده شده بین میانگین رتبه‌بندی بسته‌های شکلات امریکایی و سوئیسی چیزی بیش از تنوّع نمونه‌گیری نباشد. در نهایت، هیچ شواهدی وجود ندارد که نشان دهد شکلات‌های اتریشی به طور نظام‌مند میانگین رتبه‌بندی بالاتری نسبت به شکلات‌های امریکایی دارند.

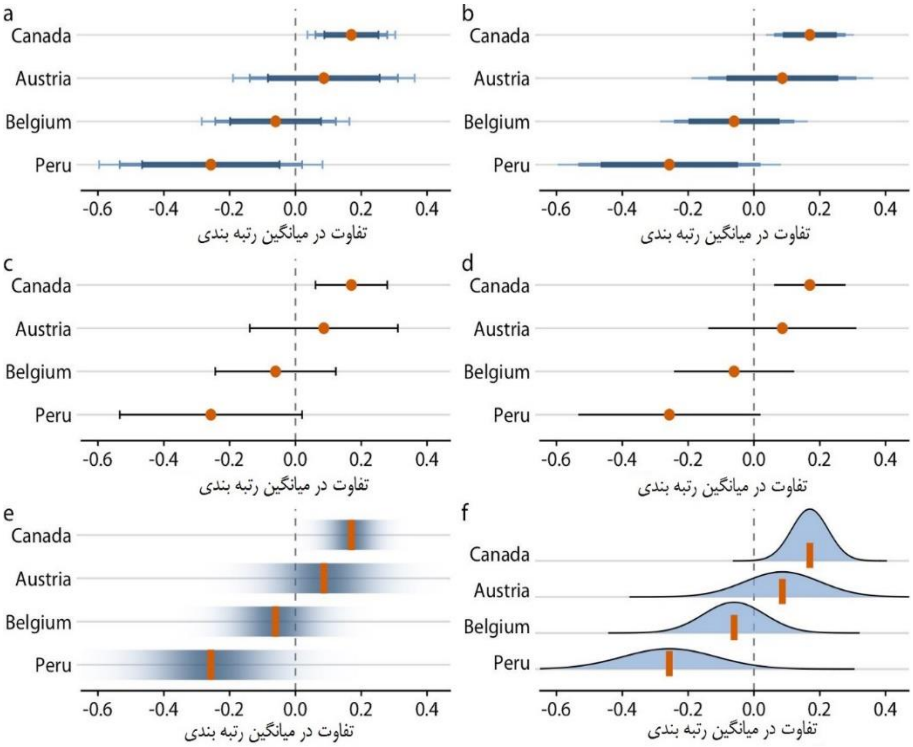
در شکل‌های قبل، از دو نوع مختلف میلهٔ خطا، درجه‌بندی شده و ساده استفاده شده است. تغییرات بیشتر نیز امکان‌پذیر است. برای مثال، می‌توانیم میله‌های خطا را با یا بدون کلاهک انتهایی رسم کنیم (شکل ۱۶–۹ الف، ج در مقابل شکل ۱۶–۹ ب، د). هر کدام از این انتخاب‌ها مزایا و معایبی دارند. میله‌های خطای درجه‌بندی شده وجود محدوده‌های متفاوت مربوط به سطوح اطمینان مختلف را برجسته می‌کنند. با این حال، عیب ارائهٔ این اطّلاعات اضافی ایجاد اختلال بصری است. بر اساس اینکه یک شکل چقدر پیچیده بوده و تراکم اطّلاعات دارد، میله‌های خطای ساده ممکن است به میلهٔ خطای درجه‌بندی شده ترجیح داده

شوند. ترسیم میله‌های خطا با یا بدون کلاهک به سلیقهٔ شخصی وابسته است. کلاهک‌ها محل ختم دقیق میلهٔ خطا را نشان می‌دهند (شکل ۱۶–۹ الف، ج)، در حالی که میلهٔ خطا بدون کلاهک تأکید یکسانی بر کل محدوده فاصله اطمینان دارد (شکل ۱۶–۹ ب، د) همچنین، مجدداً کلاهک‌ها اختلال بصری اضافه می‌کنند، بنابراین در شکلی با میله‌های خطای زیاد حذف کلاهک ممکن است ارجح باشد.



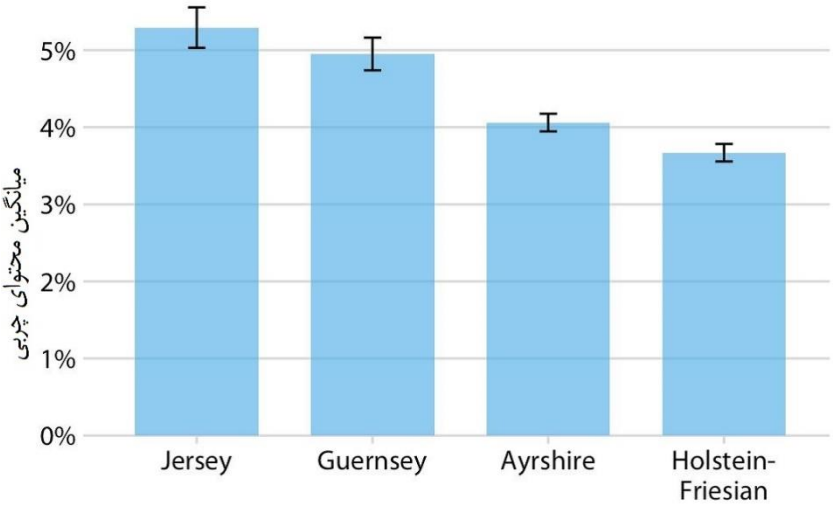
شکل ۱۶–۸. میانگین رتبه‌بندی طعم شکلات برای تولیدکنندگان از پنج کشور مختلف، نسبت به میانگین رتبه‌بندی بسته‌های شکلات امریکایی. رتبه‌بندی شکلات کانادایی به طور معنی‌داری بهتر از شکلات امریکایی است. برای چهار کشور دیگر تفاوت معنی‌داری در میانگین رتبه‌بندی در مقایسه با امریکا در سطح اطمینان ۹۵ درصد وجود ندارد. سطوح اطمینان برای مقایسه‌های چندگانه با استفاده از روش دانت تعدیل شده است. منبع داده: Brady Brelinski، انجمن شکلات منهتن.

به‌عنوان جایگزینی برای میله‌های خطا، می‌توانیم نوارهای اطمینان را ترسیم کنیم که به تدریج محو می‌شوند (شکل ۱۶–۹). نوارهای اطمینان بهتر نشان می‌دهند که مقادیر مختلف چقدر محتمل هستند، امّا خواندن آن‌ها دشوار است. زیرا باید به صورت بصری سایه‌های مختلف رنگ را ادغام کنیم تا مشخّص کنیم که یک سطح اطمینان خاص در کجا ختم می‌شود. از شکل ۹–۱۶ هـ ممکن است نتیجه بگیریم که میانگین امتیاز شکلات‌های پرو به طور معنی‌داری کمتر از شکلات‌های امریکایی است، در حالی که اینطور نیست. هنگامی که توزیع‌های اطمینان را نشان می‌دهیم هم مشکلات مشابهی ایجاد می‌شود (شکل ۱۶–۹ و). ادغام بصری سطح زیر منحنی و تعیین اینکه دقیقاً کجا به یک سطح اطمینان معیّن رسیده دشوار است. این مشکل را می‌توان تا حدودی با رسم نمودارهای چندکی مانند شکل ۱۶–۳ کاهش داد.

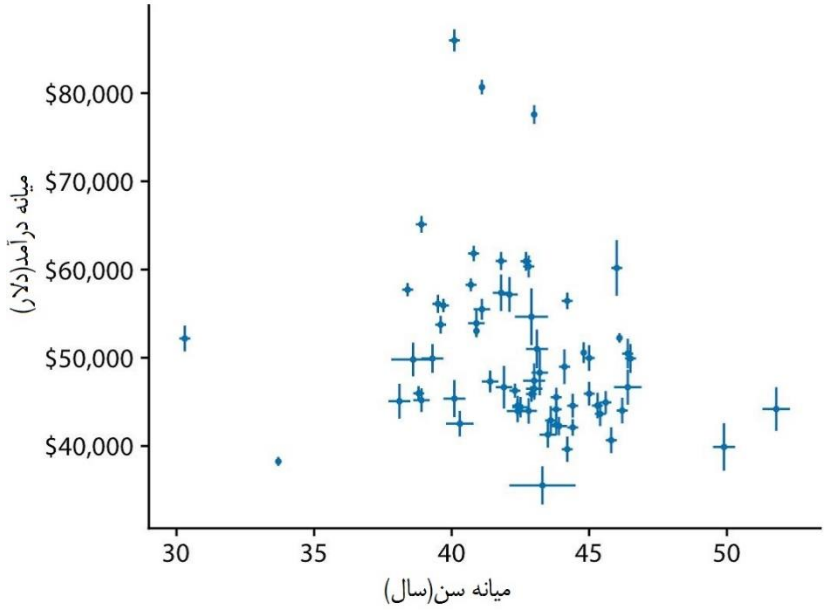


شکل ۱۶–۹. میانگین رتبه‌بندی طعم شکلات برای تولیدکنندگان از چهار کشور مختلف، نسبت به میانگین رتبه‌بندی شکلات‌های امریکایی. هر پانل از رویکرد متفاوتی برای ترسیم اطّلاعات یکسان عدم قطعیّت استفاده می‌کند: (الف) میله‌های خطای درجه‌بندی شده با کلاهک (ب) میله‌های خطای درجه‌بندی شده بدون کلاهک (ج) میله‌های خطای تک بازه‌ای با کلاهک (د) میله‌های خطای تک بازه‌ای بدون کلاهک (ه) نوارهای اطمینان (و) توزیع اطمینان. منبع داده: Brady Brelinski، انجمن شکلات منهتن.

برای شکل‌های دوبعدی ساده، میله‌های خطا نسبت به نمایش‌های پیچیده‌تر عدم قطعیّت یک مزیّت مهم دارند: می‌توان آن‌ها را با بسیاری از نمودارهای دیگر ترکیب کرد. تقریباً برای هر نوع شکلی، می‌توانیم با افزودن میله‌های خطا، نشانه‌ای از عدم قطعیّت اضافه کنیم. برای مثال، می‌توانیم مقادیر را به همراه عدم قطعیّت آن‌ها با رسم نمودار میله‌ای با میله‌های خطا نشان دهیم (شکل ۱۶–۱۰). این نوع ترسیم معمولاً در نشریات علمی استفاده می‌شود. همچنین می‌توانیم میله‌های خطا را در امتداد محور x و محور y در نمودار پراکنش ترسیم کنیم (شکل ۱۶–۱۱).



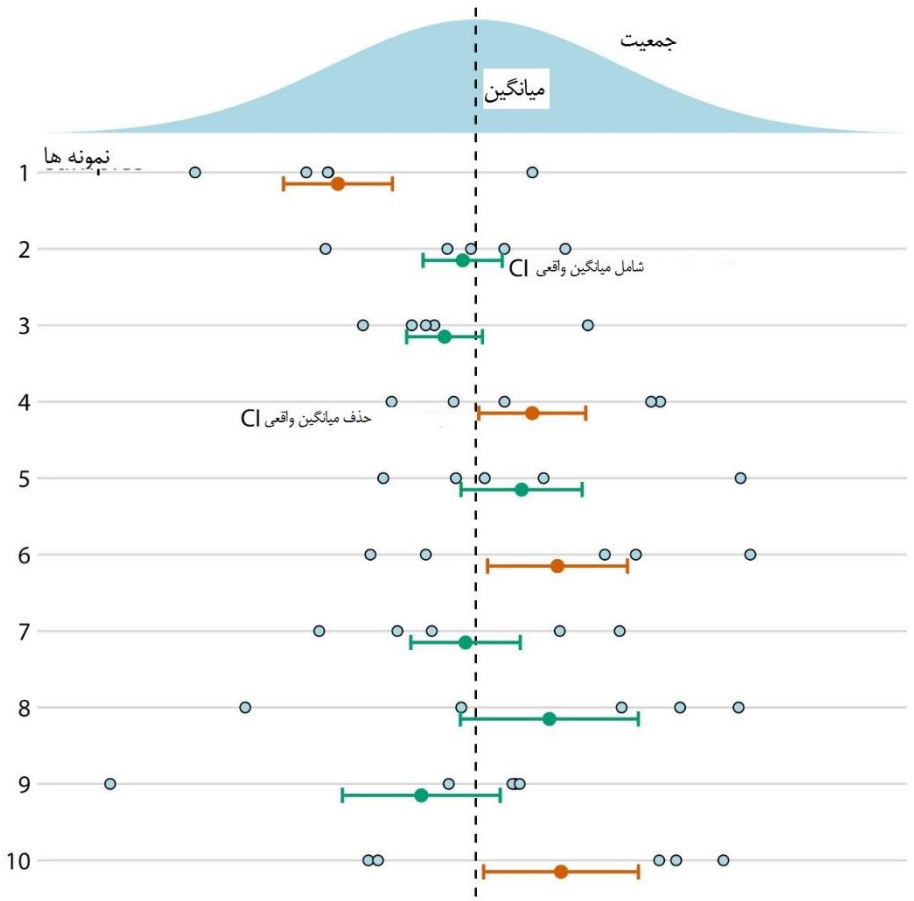
شکل ۱۶–۱۰. میانگین محتوای چربی در شیر چهار نژاد گاو. میله‌های خطا +/– یک خطای معیار میانگین را نشان می‌دهد. ترسیم‌هایی از این نوع اغلب در متون علمی دیده می‌شود. در حالی که آن‌ها از نظر فنّی صحیح هستند، امّا تنوّع در هر دسته و عدم قطعیّت نمونه را به خوبی نشان نمی‌دهند. شکل ۱۱–۷ را برای تنوع در محتوای چربی در هر نژاد مشاهده کنید. منبع داده: رکورد عملکرد کانادایی برای گاوهای شیری خالص.



شکل ۱۶–۱۱. میانهٔ درآمد در مقابل میانهٔ سنی برای ۶۷ شهرستان در پنسیلوانیا. میله‌های خطا نشان‌دهندهٔ فواصل اطمینان ۹۰ درصد هستند. منبع داده: نظرسنجی پنج ساله جامعه آمریکا در سال ۲۰۱۵.

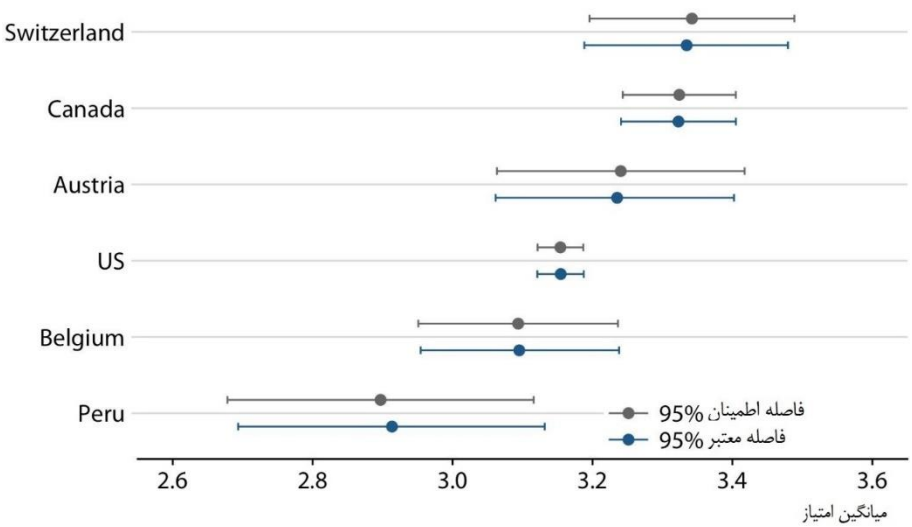
بیایید به موضوع فراوانی‌گرایان و بیزین‌گرایان برگردیم. فراوانی‌گرایان عدم قطعیّت را با فواصل اطمینان ارزیابی می‌کنند، در حالی که بیزین‌گرایان توزیع‌های پسین و فواصل معتبر را محاسبه می‌کنند. توزیع پسین بیزین به ما می‌گوید که تخمین پارامترهای خاص چقدر منجر به ارائه داده‌های ورودی می‌شود. فاصله معتبر محدوده‌ای از مقادیر را نشان می‌دهد که پارامتر مورد نظر با احتمال معینی قرار خواهد داشت، همانطور که از توزیع پسین محاسبه می‌شود. به عنوان مثال، یک فاصله معتبر ۹۵ درصد مربوط به ۹۵ درصد مرکزی از توزیع پسین است. مقدار پارامتر واقعی ۹۵ درصد احتمال دارد که در فاصله معتبر ۹۵ درصد قرار داشته باشد.

اگر شما یک متخصص آمار نیستید، ممکن است از تعریف ارائه شده از فاصله معتبر شگفت‌زده شده باشید. ممکن است فکر کرده باشید که این در واقع همان تعریف فاصله اطمینان است. امّا اینطور نیست. فاصله معتبر بیزین به شما می‌گوید که پارامتر واقعی به احتمال زیاد کجاست، و فاصله اطمینان فراوانی‌گرایان به شما می‌گوید که پارامتر واقعی به احتمال زیاد کجا نیست. در حالی که این تمایز ممکن است مانند بازی با کلمات به نظر برسد، تفاوت‌های مفهومی مهمی بین این دو رویکرد وجود دارد. تحت رویکرد بیزین، شما از داده‌ها و دانش قبلی خود در مورد سیستم مورد مطالعه استفاده می‌کنید (که پیشین نامیده می‌شود) تا توزیع احتمال (که پسین نامیده می‌شود) را محاسبه نمایید که به شما می‌گوید می‌توانید انتظار داشته باشید که مقدار متغیّر واقعی کجا باشد. در مقابل، تحت رویکرد فراوانی‌گرایانه ابتدا فرضی را مطرح می‌کنید که قصد دارید آن را رد کنید. این فرض، فرضیه صفر نامیده می‌شود، و اغلب به سادگی این فرض است که پارامتر برابر با صفر است (به عنوان مثال، هیچ تفاوتی بین دو وضعیّت وجود ندارد). سپس احتمال اینکه نمونه‌گیری تصادفی داده‌هایی مشابه آنچه که در صورت صحت فرضیه صفر مشاهده می‌شد را محاسبه می‌نمایید. فاصله اطمینان نمایشی از این احتمال است. اگر یک فاصله اطمینان معیّن، مقدار پارامتر را تحت فرضیه صفر (یعنی مقدار صفر) حذف کند، می‌توانید فرضیه صفر را در آن سطح اطمینان رد کنید. از طرف دیگر، می‌توانید فاصلهٔ اطمینان را به عنوان فاصله‌ای در نظر بگیرید که مقدار پارامتر واقعی را با احتمال مشخّص شده تحت نمونه‌گیری مکرر در بر می‌گیرد (شکل ۱۶-۱۲). بنابراین، اگر مقدار پارامتر واقعی صفر باشد، فاصله اطمینان ۹۵ درصد صفر را فقط در ۵ درصد از نمونه‌های تجزیه و تحلیل شده حذف می‌کند.



شکل ۱۶–۱۲. تفسیر مبتنی بر فراوانی از فاصلهٔ اطمینان. فواصل اطمینان به بهترین وجه در بستر نمونه گیری مکرر درک می‌شوند. هر نمونه، یک فاصلهٔ اطمینان خاص دارد که می‌تواند دربرگیرنده (سبز) یا غیردربرگیرنده (نارنجی) پارامتر واقعی باشد، که در اینجا میانگین است. با این حال، اگر به طور مکرر نمونه‌گیری کنیم، فواصل اطمینان (در اینجا فواصل اطمینان۶۸ درصد نشان داده شده است که معادل میانگین نمونه +/– خطای استاندارد است) تقریباً در ۶۸ درصد مواقع شامل میانگین واقعی است.

به طور خلاصه، فاصلهٔ معتبر بیزین در مورد مقدار واقعی پارامتر اظهار نظر می‌کند و فاصلهٔ اطمینان فراوانی‌گرایان در مورد فرضیه صفر صحبت می‌کند. با این حال، در عمل، تخمین‌های بیزین‌گرایان و فراوانی‌گرایان اغلب کاملاً مشابه هستند (شکل ۱۶–۱۳). مزیّت مفهومی رویکرد بیزین این است که بر تفکّر در مورد بزرگی اثر تأکید می‌کند، در حالی که رویکرد فراوانی‌گرایانه بر دیدگاه دوتایی از یک اثر، یعنی وجود یا عدم وجود آن، تأکید می‌کند.

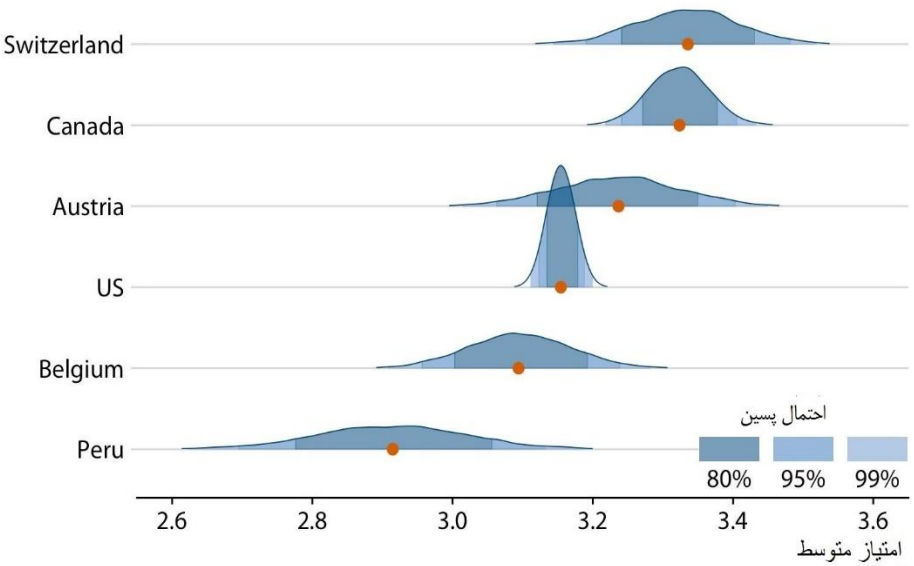


شکل ۱۶–۱۳. مقایسه فواصل اطمینان فراوانی‌گرایان و فواصل معتبر بیزین‌گرایان برای میانگین امتیازدهی شکلات. می‌بینیم که این دو رویکرد نتایج مشابهی دارند هرچند دقیقاً یکسان نیستند. به طور خاص، تخمین‌های بیزین کمی انقباضی است، که ناشی از تعدیل افراطی‌ترین تخمین‌های پارامتر نسبت به میانگین کلی است. (توجّه داشته باشید که چگونه تخمین بیزین‌گرایان نسبت به تخمین متناظر فراوانی‌گرایان برای سوئیس کمی به سمت چپ و تخمین بیزین برای پرو کمی به سمت راست منتقل شده است) تخمین‌های فراوانی‌گرایان و فواصل اطمینان نشان داده شده در اینجا با فاصله لطمینان ۹۵ درصد نشان داده شده در شکل ۱۶–۷ یکسان است. منبع داده: Brady Brelinski، انجمن شکلات منهتن

فاصله معتبر بیزین‌گرایان به این سوال پاسخ می‌دهد که «انتظار داریم پارامتر واقعی کجا قرار داشته باشد؟» فاصله اطمینان فراوانی‌گرایان به این سوال پاسخ می‌دهد: «چقدر مطمئن هستیم که مقدار واقعی پارامتر صفر نیست؟»

هدف اصلی تخمین بیزین بدست آوردن توزیع پسین است. بنابراین معمولاً بیزینی‌ها کل توزیع را به جای ساده کردن آن در یک فاصله معتبر، ترسیم می‌کنند. بنابراین، از نظر تجسم داده‌ها، تمام رویکردهای بصری‌سازی توزیع‌های مورد بحث در فصل‌های ۷، ۸ و ۹ قابل اجرا هستند. به طور خاص، هیستوگرام، نمودار چگالی، نمودار جعبه‌ای، ویولن، و نمودار خط الراس همگی معمولاً برای تجسم توزیع‌های پسین بیزین استفاده می‌شوند. از آنجایی که این رویکردها به طور مفصّل در فصل‌های مربوطه مورد بحث قرار گرفته‌اند، ما در اینجا تنها یک مثال را با استفاده از نمودار خط الراس برای نشان دادن توزیع‌های پسین بیزین برای میانگین

امتیازدهی شکلات (شکل ۱۶-۱۴) نشان می‌دهیم. در این مورد خاص، سایه‌هایی را در زیر منحنی اضافه کرده‌ایم تا نواحی مشخصی از احتمالات پسین را نشان دهد. به عنوان جایگزینی برای سایه‌زنی، می‌توانیم نمودارهای نقطه‌ای چندکی را ترسیم کنیم، یا می‌توانیم میله‌های خطای درجه‌بندی شده را در زیر هر توزیع اضافه نماییم. به نمودارهای خط الراس با میله‌های خطا در زیر آنها، نیمه چشم و به نمودارهای ویولن با میله‌های خطا، نمودار چشم اطلاق می‌شود (به مبحث «عدم قطعیّت» مراجعه کنید).

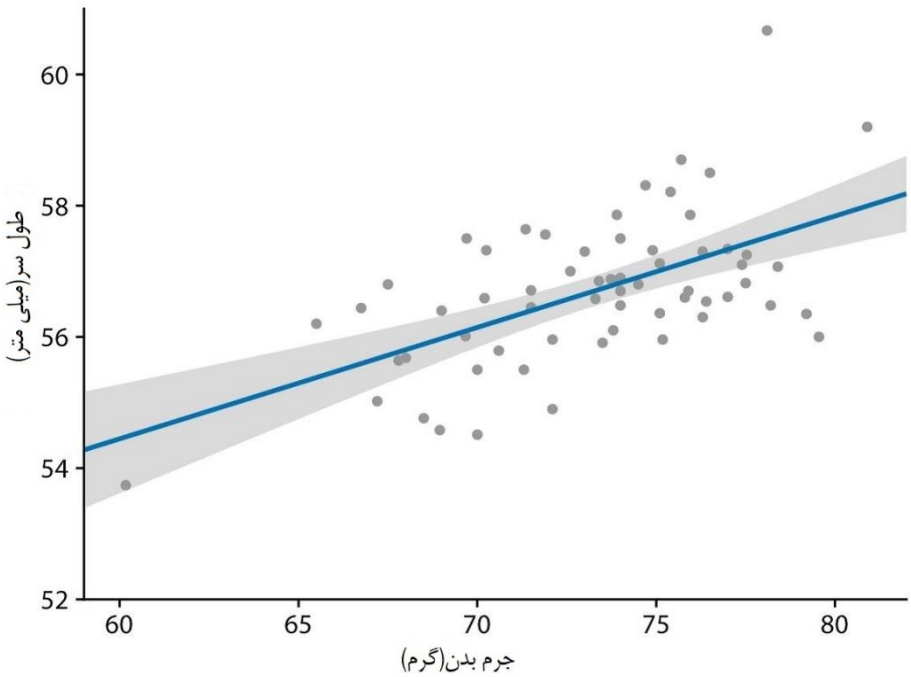


شکل ۱۶-۱۴. توزیع‌های پسین بیزین از میانگین امتیازدهی شکلات، که به صورت نمودار خط الراس نشان داده شده است. نقاط قرمز نشان‌دهندهٔ میانهٔ هر توزیع پسین است. از آنجایی که تبدیل یک توزیع پیوسته به نواحی اطمینان خاص با چشم دشوار است، ما سایه‌هایی را در زیر هر منحنی اضافه کرده‌ایم تا ۸۰، ۹۵، و ۹۹ درصد مرکزی هر توزیع پسین را نشان دهد. منبع داده: Brady Brelinski، انجمن شکلات منهتن

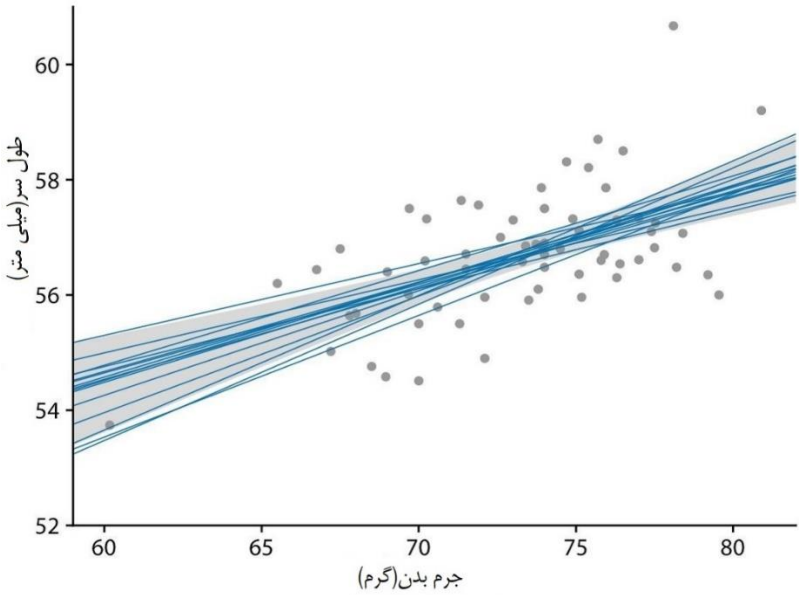
ترسیم عدم قطعیّت برازش منحنی

در فصل ۱۴، نحوه نشان دادن روند در یک مجموعه داده را با برازش یک خط مستقیم یا منحنی بر داده‌ها مورد بحث قرار دادیم. این تخمین‌های روند نیز دارای عدم قطعیّت هستند و مرسوم است که عدم قطعیّت را برای یک خط روند با نوار اطمینان نشان می‌دهند (شکل ۱۶-۱۵). نوار اطمینان طیف وسیعی از خطوط قابل برازش را در اختیار ما قرار می‌دهد. هنگامی که دانشجویان برای اوّلین بار با یک نوار اطمینان مواجه می‌شوند، اغلب از این که

حتّی یک خط برازش شده کاملاً مستقیم، نوار اطمینان منحنی ایجاد می‌کند شگفت‌زده می‌شوند. دلیل انحنا این است که خط برازش شده مستقیم می‌تواند در دو جهت متمایز حرکت کند: می‌تواند به سمت بالا و پایین حرکت کند (یعنی دارای عرض از مبدا مختلف باشد) و می‌تواند بچرخد (یعنی شیب‌های متفاوتی داشته باشد). ما می‌توانیم با رسم مجموعه‌ای از خطوط برازش جایگزین که به‌طور تصادفی از توزیع پسین پارامتر برازش شده ایجاد می‌شوند، به‌طور بصری نحوه تولید نوار اطمینان را نشان دهیم. این کار در شکل ۱۶–۱۶ انجام شده است که ۱۵ خط برازش جایگزین که به طور تصادفی انتخاب شده‌اند را نشان می‌دهد. همانطور که می‌بینید اگرچه هر خط کاملاً مستقیم است، ترکیبی از شیب‌ها و عرض از مبداهای مختلف هر خط یک شکل کلی ایجاد می‌کند که کاملاً شبیه نوار اطمینان است.



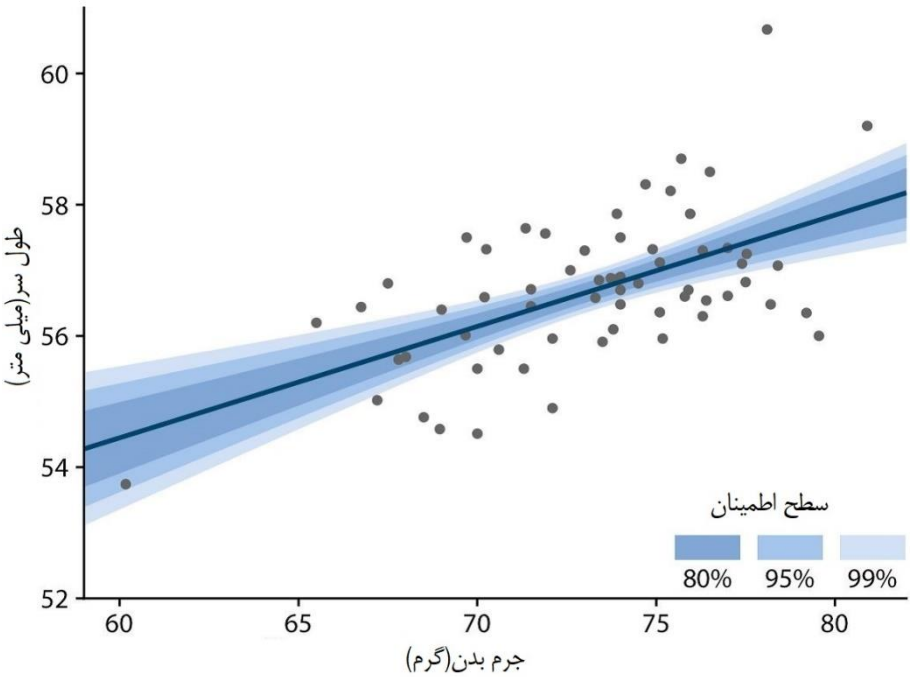
شکل ۱۶–۱۵. طول سر در مقابل تودهٔ بدن برای زاغ نر آبی، آن گونه که در شکل ۱۴–۷ آمده بود. خط مستقیم آبی نشان‌دهندهٔ بهترین برازش خطی با داده‌ها است و نوار خاکستری اطراف خط عدم قطعیّت در برازش خطی را نشان می‌دهد. نوار خاکستری نشان‌دهندهٔ فاصله اطمینان ۹۵ درصد است. منبع داده: Keith Tarvin, Oberlin College



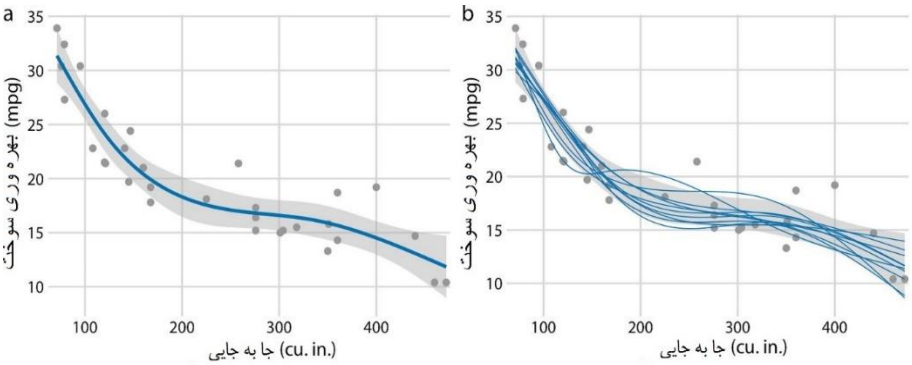
شکل ۱۶–۱۶. طول سر در مقایسه با تودهٔ بدن برای زاغ نر آبی. برخلاف شکل ۱۶–۱۵، خطوط مستقیم آبی اکنون نشان‌دهندهٔ برازش‌های جایگزین با احتمال مشابه هستند که به طور تصادفی از توزیع پسین انتخاب شده است.
منبع داده: Keith Tarvin, Oberlin College

برای ترسیم یک نوار اطمینان، باید یک سطح اطمینان را مشخّص کنیم، و همانطور که برای میله‌های خطا و احتمالات پسین دیدیم، مشخّص کردن فواصل مختلف اطمینان می‌تواند مفید باشد. این کار ما را به نوار اطمینان درجه‌بندی شده می‌رساند که چندین سطح اطمینان را به طور همزمان نشان می‌دهد (شکل ۱۶–۱۷). نوار اطمینان درجه‌بندی شده، حس عدم قطعیّت را در خواننده افزایش می‌دهد و وی را مجبور می‌کند با این احتمال روبرو شود که داده‌ها ممکن است از خطوط روند جایگزین متفاوتی پشتیبانی کنند.

همچنین می‌توانیم نوارهای اطمینان را برای برازش‌های منحنی غیرخطی ترسیم کنیم. چنین نوارهای اطمینانی زیبا به نظر می‌رسند امّا تفسیر آن‌ها دشوار است (شکل ۱۶–۱۸). اگر به شکل ۱۶–۱۸ الف نگاه کنیم، ممکن است فکر کنیم که نوار اطمینان با بالا و پایین رفتن خط آبی ایجاد می‌شود و ممکن است کمی آن را تغییر شکل دهد. با این حال، همانطور که شکل ۱۶–۱۸ ب نشان می‌دهد، نوار اطمینان نشان‌دهندهٔ خانواده‌ای از منحنی‌ها است که همگی نسبت به بهترین برازش کلی که در بخش (الف) نشان داده شده است، پر پیچ و خم تر هستند. این یک اصل کلی برازش منحنی غیرخطی است. عدم قطعیّت نه تنها با حرکت منحنی به سمت بالا و پایین، بلکه با افزایش پیچ و خم همراهی دارد.



شکل ۱۶–۱۷. طول سر در مقایسه با تودهٔ بدن برای زاغ نر آبی. همانطور که در مورد میله‌های خطا داشتیم، می‌توانیم نوارهای اطمینان درجه‌بندی شده را ترسیم کنیم تا عدم قطعیّت در برآورد را برجسته کنیم. منبع داده: Keith Tarvin, Oberlin College



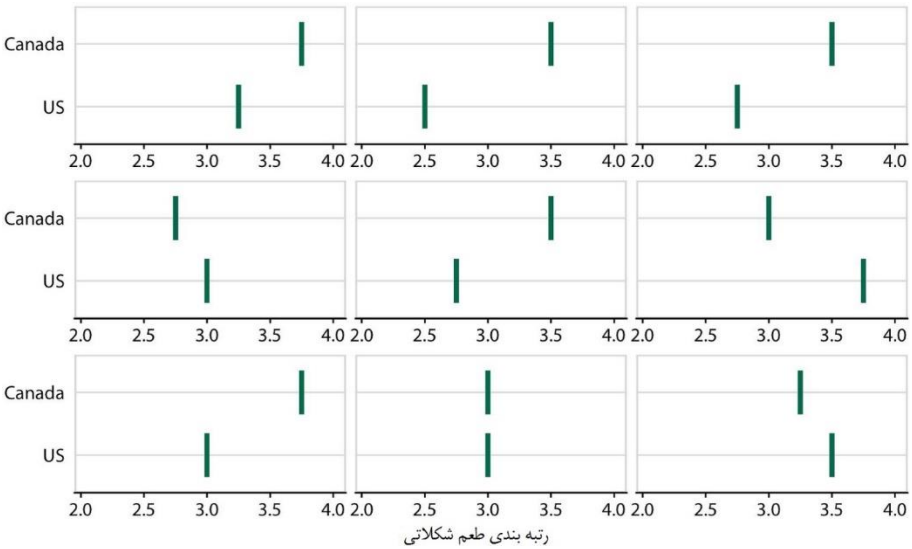
شکل ۱۶–۱۸. بازده سوخت در مقابل جابجایی، برای ۳۲ خودرو (مدل‌های ۱۹۷۳–۱۹۷۴). هر نقطه نشان‌دهندهٔ یک خودرو است، و خطوط صاف با به کارگیری یک تکه‌بند رگرسیون مکعبی با ۵ گره به دست آمده است. (الف) بهترین برازش تکه‌بند و نوار اطمینان. (ب) برازش‌های جایگزین محتمل که از توزیع پسین استخراج شده است. منبع داده: موتور ترند ۱۹۷۴

نمودارهای نتیجه فرضی

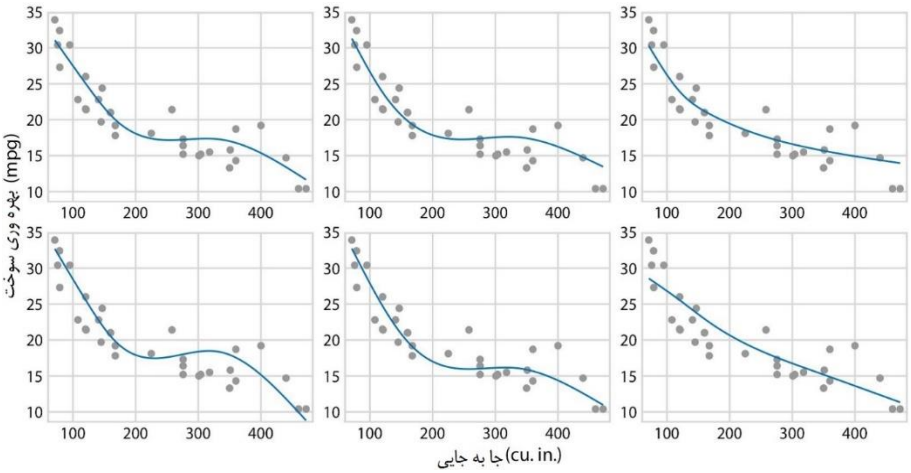
همه ترسیم‌های ثابت عدم قطعیّت از این مشکل رنج می‌برند که بینندگان ممکن است برخی از جنبه‌های ترسیم عدم قطعیّت را به عنوان یک ویژگی قطعی داده‌ها تفسیر کنند (یک خطای ساختاری قطعی، همانطور که قبلاً توضیح داده شد). ما می‌توانیم با ترسیم عدم قطعیّت از طریق انیمیشن، بوسیله چرخش در تعدادی از طرح‌های مختلف امّا با احتمال مشابه، از این مشکل جلوگیری کنیم. این نوع ترسیم، نمودار پیامد فرضی نامیده می‌شود. در حالی که بکارگیری نمودارهای پیامد فرضی در رسانه چاپی امکان‌پذیر نیست، امّا در بستر برخط که نمودارهای متحرّک را می‌توان در قالب‌های تصاویر متحرّک (GIF) یا ویدیو (MP4) ارائه کرد، این نمودارها می‌توانند بسیار موثر باشند. همچنین نمودارهای پیامد فرضی می‌توانند برای ارائهٔ شفاهی بسیار اثربخش باشند.

برای نشان دادن مفهوم پیامد فرضی، اجازه دهید یک بار دیگر به مثال رتبه‌بندی‌های شکلات برگردیم. وقتی در فروشگاه مواد غذایی ایستاده‌اید و به خرید شکلات فکر می‌کنید، احتمالاً به میانگین امتیاز طعم و عدم قطعیّت مرتبط با گروه‌های خاصی از شکلات اهمیت نمی‌دهید. در عوض، ممکن است بخواهید پاسخ یک سوال ساده‌تر را بدانید، مانند: اگر من به‌طور تصادفی یک شکلات ساخت کانادا و امریکا را انتخاب کنم، باید انتظار طعم بهتری از کدام یک از این دو داشته باشم؟ برای رسیدن به پاسخ این سوال، می‌توانیم به‌طور تصادفی یک شکلات کانادایی و آمریکایی را از مجموعه داده‌ها انتخاب کنیم، رتبه‌بندی‌های آن‌ها را با هم مقایسه کنیم، نتیجه را ثبت کنیم، و سپس این فرآیند را بارها تکرار کنیم. اگر این کار را انجام می‌دادیم، متوجّه می‌شدیم که تقریباً در ۵۳ درصد موارد، رتبه شکلات کانادایی بهتر است و در ۴۷ درصد موارد، یا شکلات امریکایی رتبه بهتری دارد یا هر دو شکلات وضعیّت یکسانی دارند. ما می‌توانیم این فرآیند را به صورت بصری با چرخش بین چند جفت از این قرعه‌کشی‌های تصادفی و ارائه رتبه نسبی دو شکلات برای هر جفت نشان دهیم (شکل ۱۶–۱۹).

به عنوان مثال دوّم، تنوع شکل‌ها را در میان خطوط روند محتمل در شکل ۱۶–۱۸ ب در نظر بگیرید. از آنجایی که همه خطوط روند روی یکدیگر ترسیم شده‌اند، ما در درجهٔ اوّل منطقهٔ کلی را که توسّط خطوط روند پوشانده شده است، درک می‌کنیم که شبیه به نوار اطمینان است. درک خطوط روند مجزا دشوار است. با تبدیل این شکل به یک نمودار پیامد فرضی، می‌توانیم خطوط روند مجزا را یکی یکی مشخّص کنیم (شکل ۱۶–۲۰).



شکل ۱۶–۱۹. نمای شماتیک از یک نمودار پیامد فرضی برای رتبه‌بندی شکلات‌های تولید شده در کانادا و امریکا. هر نوار سبز عمودی رتبه‌بندی یک شکلات را نشان می‌دهد، و هر پانل مقایسه‌ای از دو شکلات انتخابی تصادفی، یکی کانادایی و یکی امریکایی را نشان می‌دهد. در یک نمودار پیامد فرضی واقعی، نمایشگر به جای نشان دادن آن‌ها در کنار هم، بین پانل‌های نمودار چرخش می‌کند. منبع داده: Brady Brelinski، انجمن شکلات منهتن



شکل ۱۶–۲۰. نمای شماتیک نمودار پیامد فرضی برای بازده سوخت در مقابل جابجایی. هر نقطه نشان‌دهندهٔ یک خودرو است، و خطوط صاف با بکارگیری یک تکه‌بند رگرسیون مکعبی با ۵ گره به دست آمده است. هر خط در هر پانل نشان‌دهندهٔ برازش یک پیامد جایگزین است که از توزیع پسین پارامترهای برازش شده استخراج شده است. در یک نمودار پیامد فرضی واقعی، نمایشگر به جای نشان دادن آن‌ها در کنار هم، بین پانل‌های نمودار متمایز چرخش می‌کند. منبع داده: موتور ترند، ۱۹۷۴

هنگام تهیه یک نمودار پیامد فرضی، ممکن است بپرسید که آیا بهتر است بین نتایج مختلف یک تغییر ناگهانی داشته باشید (مانند اسلاید پروژکتور) یا به آرامی از یک نتیجه به نتیجه بعدی حرکت کنید (به عنوان مثال، به آرامی خط روند را برای یک نتیجه تغییر شکل دهید تا زمانی که مانند خط روند برای یک نتیجه دیگر به نظر برسد). در حالی که این تا حدی یک سوال باز است که همچنان نیازمند تحقیق بیشتر است، برخی شواهد نشان می‌دهد که انتقال آرام قضاوت در مورد احتمالات ارائه شده را دشوارتر می‌کند. اگر متحرّک‌سازی بین نتایج را در نظر دارید، حداقل بهتر است این پویانمایی‌ها را بسیار سریع بسازید، یا سبک پویانمایی را انتخاب کنید که در آن نتایج به جای تغییر شکل از یکی به دیگری، از محو شدن برای ورود و خروج استفاده می‌کند.

یک نکتهٔ حیاتی وجود دارد که باید هنگام تهیه نمودار پیامد فرضی به آن توجّه کنیم: باید مطمئن شویم که نتایجی که نشان می‌دهیم نمایندهٔ توزیع واقعی تمام نتایج ممکن است. در غیر این صورت، نمودار پیامد فرضی می‌تواند گمراه‌کننده باشد. برای مثال، برگردیم به مثال رتبه‌بندی شکلات، اگر به طور تصادفی ۱۰ جفت شکلات را انتخاب کنیم و در بین آن‌ها، شکلات‌های امریکایی در ۷ مورد بهتر از شکلات کانادایی باشد، نمودار پیامد فرضی به اشتباه این تصوّر را ایجاد می‌کند که شکلات امریکایی احتمالا بهتر از شکلات‌های کانادایی رتبه‌بندی می‌شود. ما می‌توانیم با انتخاب پیامدهای بسیار زیاد از این موضوع جلوگیری کنیم تا سوگیری‌های نمونه‌گیری نامحتمل باشد یا به نوعی از مناسب بودن پیامدهای نشان داده شده اطمینان حاصل کنیم. در زمان ترسیم نمودار ۱۶–۱۹ مطمئن شدیم که تعداد دفعاتی که نشان می‌داد شکلات کانادایی بهتر است نزدیک به مقدار واقعی ۵۳ درصد باشد.

فصل ۱۷

اصل جوهر متناسب

در بسیاری از سناریوها مقادیر داده‌ها با طیف یک عنصر گرافیکی نشان داده می‌شوند. به عنوان مثال، در یک نمودار میله‌ای، میله‌هایی ترسیم می‌شوند که از صفر شروع و به مقدار داده‌ای که معرّف آن هستند، ختم می‌شوند. در این حالت، مقدار داده نه تنها در نقطه پایانی میله، بلکه در ارتفاع یا طول میله نیز اعمال می‌شود. اگر میله‌ای رسم شود که از مقداری متفاوت با صفر آغاز شود، طول میله و نقطه پایانی میله اطّلاعات متناقضی را منتقل می‌کنند. چنین نمودارهایی از نظر ذاتی متناقض هستند، زیرا دو مقدار متفاوت را با عنصر گرافیکی یکسانی نشان می‌دهند. این را با سناریویی مقایسه کنید که در آن مقدار داده با یک نقطه نشان داده می‌شود. در این مورد، مقدار داده فقط در محل نقطه اعمال می‌شود، نه در اندازه یا شکل نقطه.

هنگامی که از عناصر گرافیکی مانند میله‌ها، مستطیل‌ها، مناطق سایه‌دار، یا هر عنصر دیگری که مساحت مشخصی دارد استفاده می‌کنیم که می‌تواند با مقدار داده نشان داده شده سازگار یا ناسازگار باشد، مسائل مشابهی به وجود می‌آیند. در تمام این موارد، باید اطمینان حاصل نمود که هیچ تناقضی وجود ندارد. این مفهوم به عنوان اصل جوهر متناسب[1] نامیده شده است. [Bergstrom and West 2016]:

1. principle of proportional ink

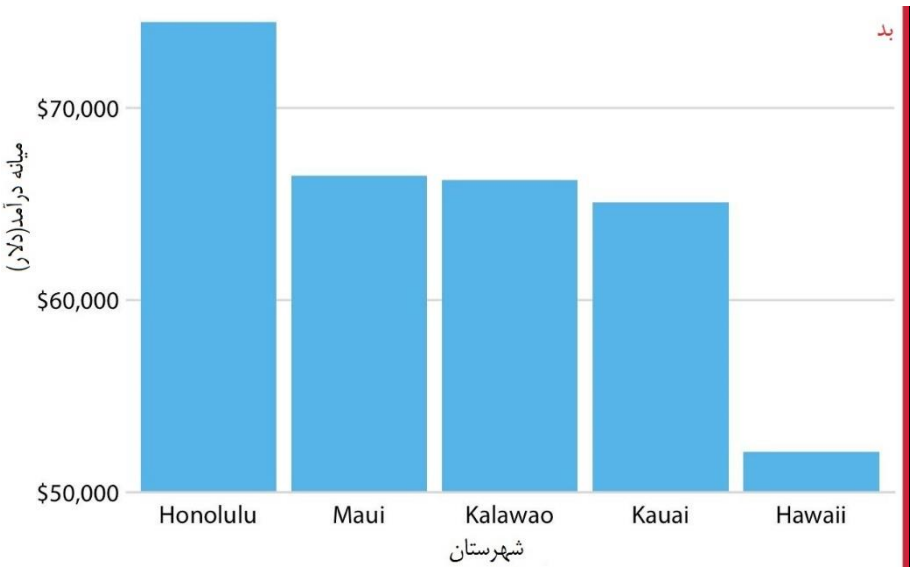
هنگامی که یک ناحیه سایه‌دار برای نشان دادن یک مقدار عددی استفاده می‌شود، مساحت آن ناحیه سایه‌دار باید مستقیماً با مقدار مربوطه متناسب باشد.

(استفاده از کلمه «جوهر» برای اشاره به هر بخشی از نمودار که متفاوت با رنگ پس‌زمینه می‌باشد، معمول است. لذا می‌تواند شامل خطوط، نقاط، مناطق مشترک و متن شود. امّا در این فصل، در درجهٔ اوّل منظور مناطق سایه‌دار است) نقض این اصل به ویژه در مطبوعات عمومی و در دنیای مالی کاملاً رایج است.

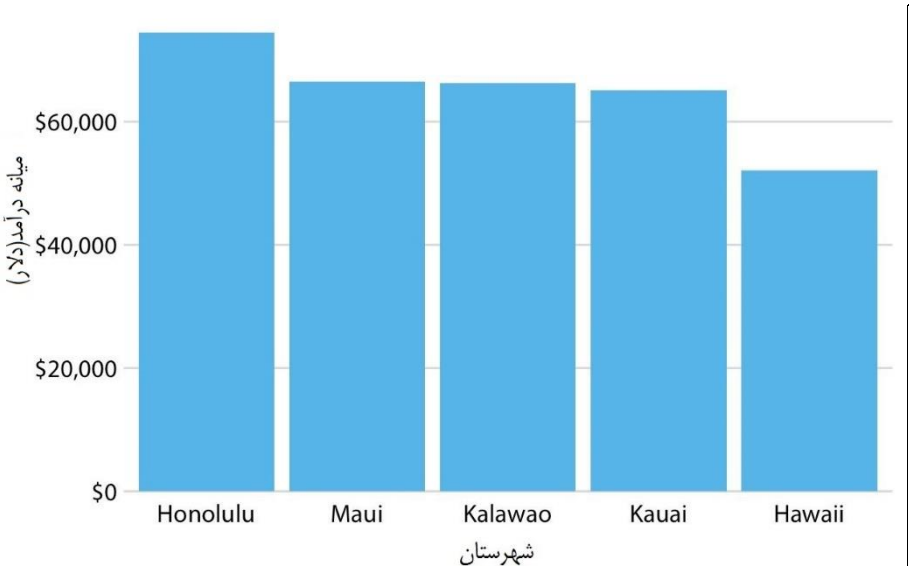
نمایش در امتداد محورهای خطی

در ابتدا رایج‌ترین سناریو را در نظر می‌گیریم، نمایش مقادیر در یک مقیاس خطی. نمودار ۱۷–۱ میانهٔ درآمد را در پنج شهرستان تشکیل‌دهنده ایالت هاوایی نشان می‌دهد. این یک نمودار معمولی است که هر کسی ممکن است در روزنامه با آن روبرو شود. یک نگاه اجمالی به این نمودار نشان می‌دهد که شهرستان هاوایی به طرز باورنکردنی فقیر است در حالی که شهرستان هونولولو بسیار ثروتمندتر از سایر شهرستان‌ها است. با این حال، نمودار ۱۷–۱ کاملاً گمراه‌کننده است، زیرا همهٔ میله‌ها از میانهٔ درآمد ۵۰۰۰۰ دلار شروع می‌شوند. بنابراین، در حالی که نقطه پایانی هر میله به درستی نشان‌دهندهٔ میانهٔ واقعی درآمد در هر شهرستان است، امّا ارتفاع میله نشان‌دهندهٔ میزانی است که میانهٔ درآمد از ۵۰۰۰۰ دلار (یک انتخاب دلخواهی) فراتر می‌رود. ادراک انسان به گونه‌ای است که در زمان نگاه کردن به این نمودار، ارتفاع میله به عنوان کمیّت کلیدی درک می‌شود و نه مکان نقطه پایانی میله نسبت به محور y.

نمایش مناسب این مجموعه داده، هیجان کمتری ایجاد می‌کند (نمودار ۱۷–۲). در حالی که تفاوت‌هایی در میانهٔ درآمد بین شهرستان‌ها وجود دارد، امّا اصلاً به اندازه‌ای که در نمودار ۱۷–۱ پیشنهاد شد، نیست. به طور کلی، میانهٔ درآمد در شهرستان‌های مختلف تا حدودی شبیه یکدیگر است.



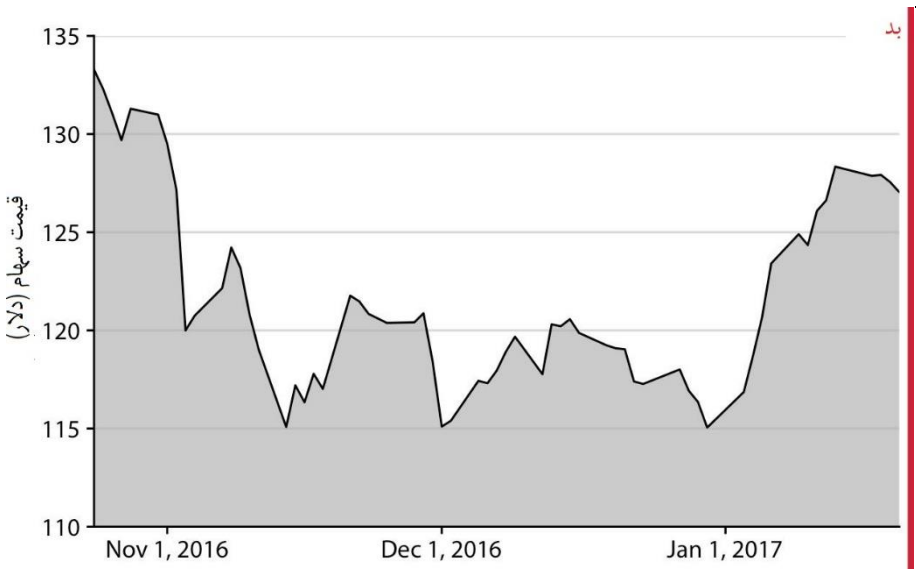
شکل ۱۷–۱. میانهٔ درآمد در پنج شهرستان ایالت هاوایی. این نمودار گمراه‌کننده است، زیرا مقیاس محور y به جای صفر دلار از ۵۰۰۰۰ دلار شروع می‌شود. در نتیجه، ارتفاع میله‌ها با مقادیر نشان داده شده متناسب نیست و تفاوت درآمد بین شهرستان هاوایی و چهار شهرستان دیگر بسیار بزرگ‌تر از آنچه در واقع است به نظر می‌رسد.



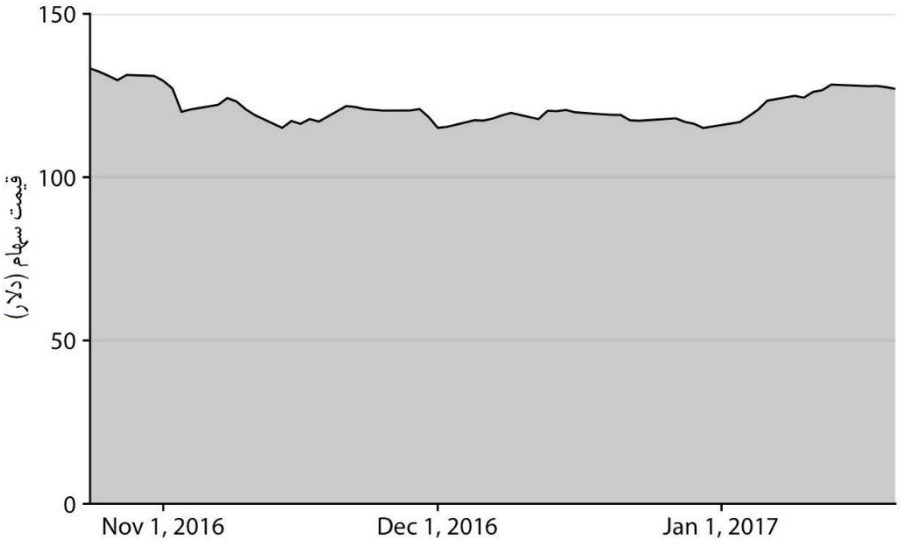
نمودار ۱۷–۲ میانهٔ درآمد در پنج شهرستان ایالت هاوایی. در اینجا، مقیاس محور y از صفر دلار شروع می‌شود و بنابراین نسبت مقادیر میانهٔ درآمد در پنج شهرستان به طور دقیق نشان داده شده است.

در مقیاس خطی میله‌ها همیشه باید از صفر شروع شوند.

مشکلات مشابهی اغلب در نمودارهای سری زمانی، مانند قیمت سهام، ایجاد می‌شود. نمودار ۱۷–۳ نشان دهندهٔ سقوط بزرگ در قیمت سهام فیس بوک در ۱ نوامبر ۲۰۱۶ می‌باشد. در واقع، کاهش قیمت نسبت به قیمت کل سهام متوسط بود (نمودار ۱۷–۴). محدوده محور y در نمودار ۱۷–۳ حتی بدون سایه‌زنی زیر منحنی نیز سوال‌برانگیز خواهد بود. امّا با سایه زدن، نمودار به طور مشخّص مشکل‌ساز می‌شود. سایه بر روی فاصله محل محور x تا مقادیر خاص هر داده در محور y تاکید می‌کند و بنابراین این تصوّر بصری را ایجاد می‌کند که ارتفاع ناحیه سایه‌دار در یک روز معین نشان دهندهٔ قیمت سهام آن روز است. در حالی که در حقیقت فقط تفاوت قیمت سهام را از خط پایه نشان می‌دهد که در نمودار ۱۷–۳ معادل ۱۱۰ دلار است.

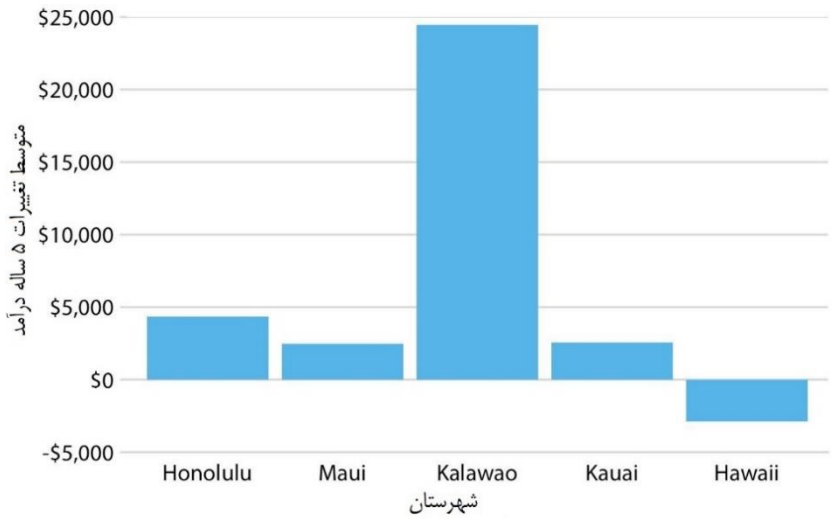


نمودار ۱۷–۳. قیمت سهام فیس بوک از ۲۲ اکتبر ۲۰۱۶ تا ۲۱ ژانویه ۲۰۱۷. به نظر می‌رسد این نمودار تاکید دارد که قیمت سهام فیس‌بوک در حدود ۱ نوامبر ۲۰۱۶ سقوط کرده است. با این حال، این حالت گمراه‌کننده است، زیرا محور y از ۱۱۰ دلار به جای صفر دلار شروع شده است.



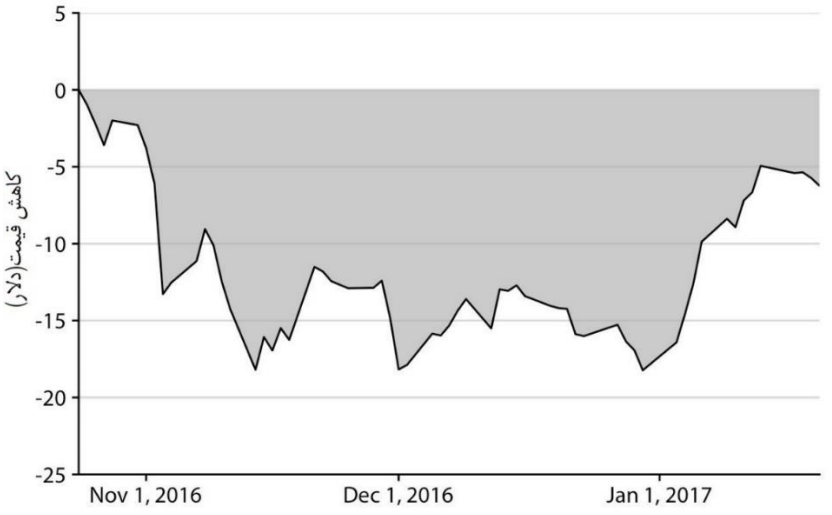
نمودار ۱۷–۴. قیمت سهام فیس‌بوک از ۲۲ اکتبر ۲۰۱۶ تا ۲۱ ژانویه ۲۰۱۷. با نشان دادن قیمت سهام در مقیاس y از صفر تا ۱۵۰ دلار، این نمودار با دقّت بیشتری میزان افت قیمت سهام فیس‌بوک را در حدود ۱ نوامبر ۲۰۱۶ نشان می‌دهد.

مثال‌های ارائه شده در نمودارهای ۱۷–۲ و ۱۷–۴ پیشنهاد می‌کند که میله‌ها و نواحی سایه‌دار برای نشان دادن تغییرات کوچک در طول زمان یا تفاوت بین شرایط مختلف مفید نیستند، زیرا همیشه باید کل میله یا ناحیه از نقطه صفر شروع شود. با این حال این پیشنهاد صحیح نیست. استفاده از میله‌ها یا مناطق سایه‌دار برای نشان دادن تفاوت بین شرایط مختلف کاملاً صحیح است، البته زمانی که مشخّص شود که چه تفاوت‌هایی ارائه شده‌اند. برای مثال، می‌توان از میله‌ها برای نشان دادن تغییر میانهٔ درآمد در شهرستان‌های هاوایی از سال ۲۰۱۰ تا ۲۰۱۵ استفاده نمود (نمودار ۱۷–۵). برای همه شهرستان‌ها به جز کالاوائو، این تغییر کمتر از ۵۰۰۰ دلار است (کالاوائو یک شهرستان غیرمعمول است، از آنجایی که کمتر از ۱۰۰ نفر جمعیّت دارد ممکن است با ورود یا خروج تعداد کمی از افراد نوسانات زیادی را در میانهٔ درآمد تجربه کند) و برای شهرستان هاوایی، این تغییر منفی است یعنی میانهٔ درآمد در سال ۲۰۱۵ کمتر از سال ۲۰۱۰ بوده است. مقادیر منفی با ترسیم میله‌هایی که در جهت مخالف قرار می‌گیرند، نشان داده می‌شوند و از صفر به جای بالا، به سمت پایین حرکت می‌کنند.



نمودار ۱۷–۵. تغییر در میانهٔ درآمد در شهرستان‌های هاوایی از ۲۰۱۰ تا ۲۰۱۵.

به طور مشابه، می‌توان تغییر قیمت سهام فیس‌بوک را در طول زمان به صورت تفاوت از نقطه اوج موقّت آن در ۲۲ اکتبر ۲۰۱۶ ترسیم نمود (نمودار ۱۷–۶). با سایه زدن ناحیه‌ای که نشان‌دهندهٔ فاصله از نقطه اوج است، به طور دقیق بزرگی مطلق افت قیمت بدون اظهار نظر ضمنی در مورد میزان افت قیمت نسبت به کل قیمت سهام قابل نمایش است.



نمودار ۱۷–۶. کاهش قیمت سهام فیس‌بوک نسبت به قیمت ۲۲ اکتبر ۲۰۱۶. بین ۱ نوامبر ۲۰۱۶ و ۱ ژانویه ۲۰۱۷، قیمت تقریباً ۱۵ دلار کمتر از نقطه اوج خود در ۲۲ اکتبر ۲۰۱۶ باقی‌ماند. قیمت در ژانویه شروع به بهبود کرد.

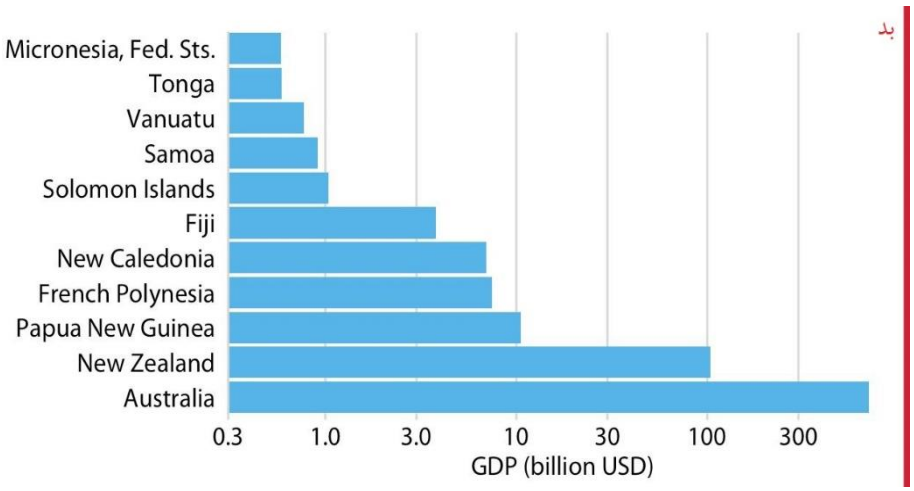
نمایش در امتداد محورهای لگاریتمی

هنگامی که داده‌ها در یک مقیاس خطی نمایش داده می‌شوند، مساحت میله‌ها، مستطیل‌ها یا اشکال دیگر به طور خودکار با مقادیر داده متناسب هستند. اگر از یک مقیاس لگاریتمی استفاده شود، این امر صادق نیست، زیرا مقادیر داده‌ها به صورت خطی در امتداد محور قرار ندارند. بنابراین، می‌توان استدلال کرد که، برای مثال، نمودارهای میله‌ای در مقیاس لگاریتمی ذاتاً ناقص هستند. از طرف دیگر، مساحت هر میله با لگاریتم مقدار داده متناسب خواهد بود، و بنابراین نمودارهای میله‌ای در مقیاس لگاریتمی اصل جوهر متناسب را در مختصّات تبدیل شده لگاریتمی برآورده می‌کنند. در عمل، به نظر می‌رسد هیچ یک از این دو بحث نمی‌تواند مشخّص کند که آیا نمودارهای میله‌ای در مقیاس لگاریتمی مناسب هستند یا خیر. در عوض، سوال این است که آیا هدف نمایش مقدار مطلق است یا نسبت.

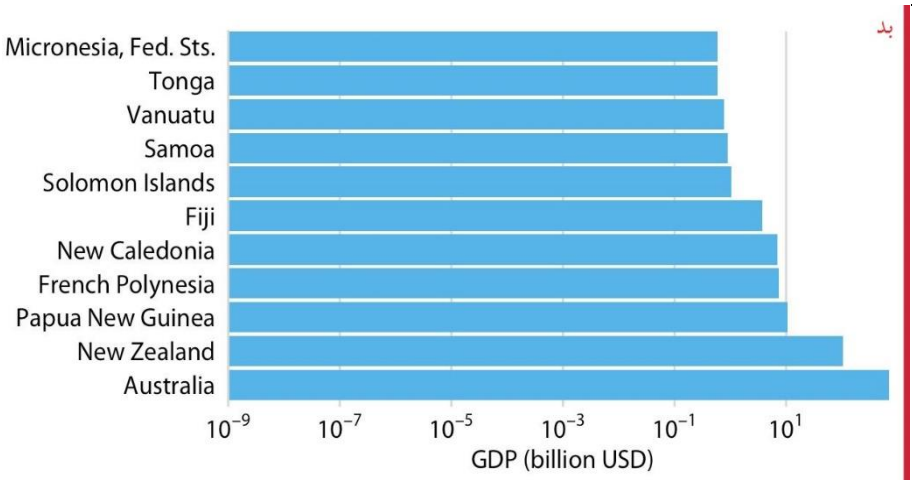
در فصل ۳، توضیح داده شد که مقیاس لگاریتمی، مقیاس طبیعی برای نمایش نسبت‌ها است، زیرا حرکت یک واحدی در امتداد مقیاس لگاریتمی معادل با ضرب یا تقسیم بر یک عامل ثابت می‌باشد. با این حال، در عمل، مقیاس‌های لگاریتمی مشخصاً برای نمایش نسبت‌ها استفاده نمی‌شود، بلکه فقط به این دلیل از آن‌ها بهره برده می‌شود چون اعداد نشان داده شده در طیف بسیار گسترده‌ای هستند. به عنوان مثال، تولید ناخالص داخلی[1] (GDP) کشورهای اقیانوسیه را در نظر بگیرید. در سال ۲۰۰۷، این شاخص از کمتر از یک میلیارد دلار آمریکا تا بیش از ۳۰۰ میلیارد دلار متغیّر بود (شکل ۱۷–۷). نمایش این اعداد در مقیاس خطی مناسب نخواهد بود، زیرا دو کشور با بیشترین تولید ناخالص داخلی (نیوزیلند و استرالیا) نمودار را تسخیر خواهند کرد.

با این‌حال، نمایش با میله در مقیاس لگاریتمی نیز مناسب نیست(نمودار ۱۷–۷). میله‌ها با مقدار دلخواه ۰٫۳ میلیارد دلار شروع می‌شوند و حداقل این نمودار از مشکل مشابه نمودار ۱۷–۱ رنج می‌برد، که طول میله‌ها معرّف مقادیر داده‌ها نبود. مشکل اضافه شده در مقیاس لگاریتمی این است که نمی‌توان به سادگی میله‌ها را از صفر آغاز نمود. در نمودار ۱۷–۷، مقدار صفر در نقطه بی‌نهایت سمت چپ قرار دارد. بنابراین، می‌توان مانند نمودار ۱۷–۸، میله‌های را با حرکت دادن بیشتر و دورتر از منشأ، طولانی نمود. این مشکل همیشه زمانی ایجاد می‌شود که هدف نمایش مقادیر مطلق (مشابه مقادیر تولید ناخالص داخلی) در مقیاس لگاریتمی می‌باشد.

1. gross domestic products

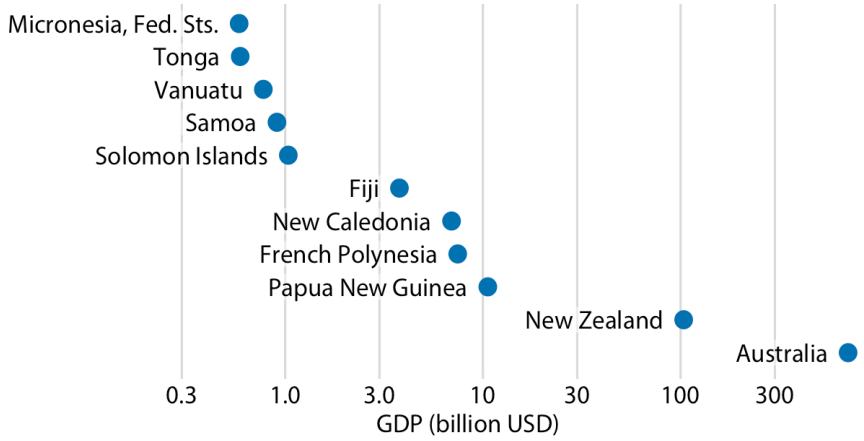


نمودار ۱۷–۷. تولید ناخالص داخلی در سال ۲۰۰۷ در کشورهای اقیانوسیه. طول میله‌ها واقعاً منعکس کنندهٔ مقادیر نشان داده شده نیست، زیرا میله‌ها با مقدار دلخواه ۰٫۳ میلیارد دلار شروع می‌شوند.



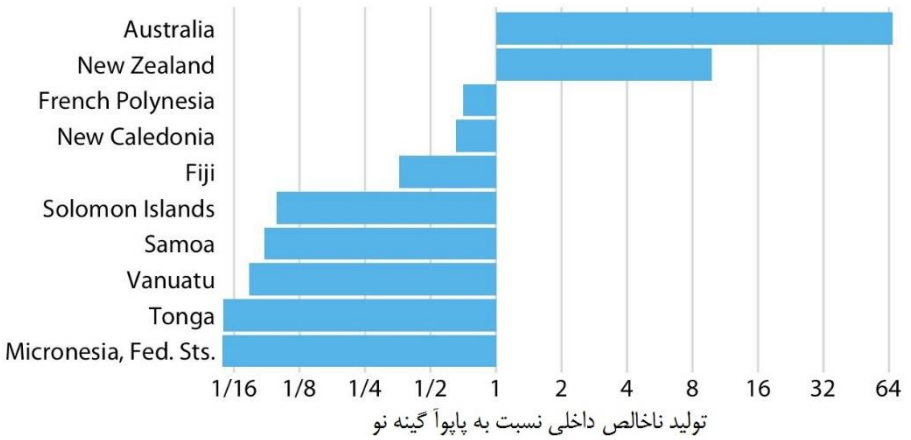
نمودار ۱۷–۸. تولید ناخالص داخلی در سال ۲۰۰۷ در کشورهای اقیانوسیه. طول میله‌ها واقعاً منعکس‌کنندهٔ مقادیر نشان داده شده نیست، زیرا میله‌ها با مقدار دلخواه 10^{-9} میلیارد دلار شروع می‌شوند.

برای داده‌های نمودار ۱۷–۷، به نظر می‌رسد میله‌ها نامناسب هستند. در عوض، می‌توان به راحتی تنها یک نقطه را در مکان مناسب در امتداد مقیاس برای تولید ناخالص داخلی هر کشور قرار داد و از مشکلات ناشی از طول میله‌ها به طور کلی اجتناب نمود (نمودار ۱۷–۹). نکتهٔ مهم این است که با قرار دادن نام کشورها دقیقاً در کنار نقاط مربوطه به جای محور y، از ایجاد ادراک بصری مربوط به بزرگی که به دلیل وجود فاصله نقاط داده از نام کشور ایجاد می‌شود، جلوگیری گردیده است.



نمودار ۱۷–۹. تولید ناخالص داخلی در سال ۲۰۰۷ در کشورهای اقیانوسیه

با این حال، اگر هدف نمایش نسبت‌ها به جای مقادیر باشد، میله‌ها در مقیاس لگاریتمی گزینه مناسبی خواهند بود. در واقع، در این مورد، آن‌ها به میله‌ها در مقیاس خطی ارجحیّت دارند. به عنوان مثال، بیایید ارزش تولید ناخالص داخلی کشورهای اقیانوسیه را نسبت به تولید ناخالص داخلی پاپوآ گینه‌نو رسم نماییم. نمودار به دست آمده به خوبی روابط کلیدی بین تولید ناخالص داخلی کشورهای مختلف را برجسته می‌کند (نمودار ۱۷–۱۰). همانطور که مشاهده می‌شود که تولید ناخالص داخلی نیوزلند بیش از ۸ برابر پاپوآ گینه‌نو و استرالیا بیش از ۶۴ برابر تولید ناخالص داخلی پاپوآ گینه نو می‌باشد، در حالی که تونگا و ایالات فدرال میکرونزی هر کدام کمتر از یک شانزدهم تولید ناخالص داخلی پاپوآ گینه‌نو را دارند. تولید ناخالص داخلی پلی‌نزی فرانسه و کالدونیای نو نزدیک به هم و کمی کمتر از پاپوآ گینه‌نو است.



نمودار ۱۷–۱۰ . تولید ناخالص داخلی در کشورهای اقیانوسیه نسبت به تولید ناخالص داخلی پاپوآ گینه نو در سال ۲۰۰۷

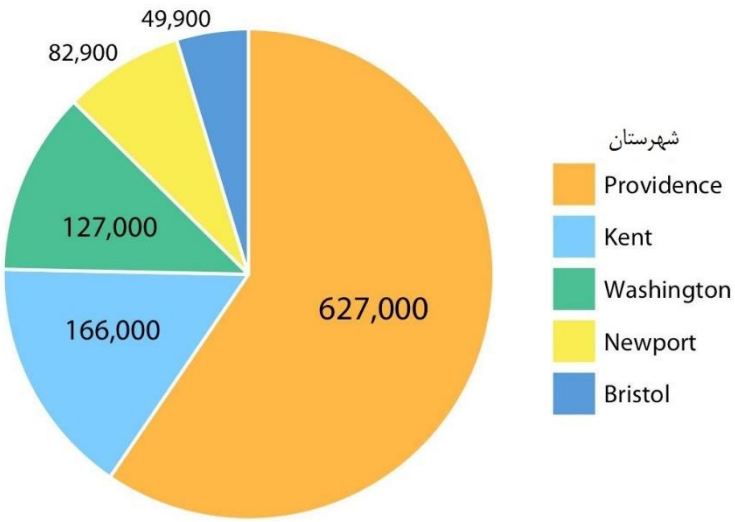
نمودار ۱۷–۱۰ همچنین بر این نکته تاکید می‌کند که نقطه میانی مقیاس لگاریتمی عدد یک است، و میله‌هایی که مقدار بیشتر از یک دارند در یک جهت و میله‌هایی که مقدار کمتر از یک دارند در جهت دیگر قرار دارند. میله‌ها در مقیاس لگاریتمی نشان‌دهندهٔ نسبت‌ها هستند و همیشه باید از یک شروع شوند، در حالی که میله‌ها در مقیاس خطی نشان‌دهندهٔ مقادیر هستند و همیشه باید از صفر شروع شوند.

هنگامی که میله‌ها در مقیاس لگاریتمی رسم می‌شوند، نسبت‌ها را نشان می‌دهند و باید از عدد یک شروع شوند و نه از صفر.

نمایش مستقیم منطقه

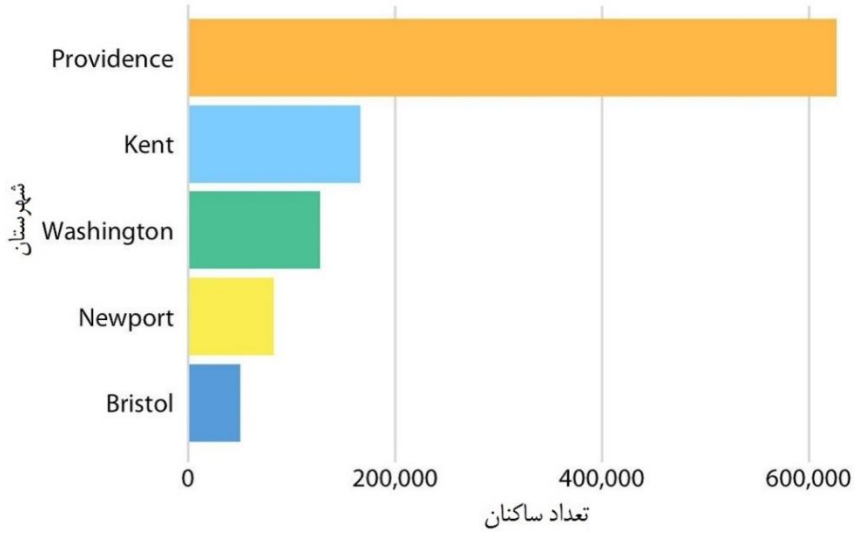
همه مثال‌های قبلی، داده‌ها را در امتداد یک محور خطی تجسم کردند، به طوری که هر مقدار داده هم بر اساس مساحت و هم بر اساس مکان در امتداد محور x یا y ثبت می‌شد. در این موارد می‌توان مساحت مشخّص شده را به صورت اتفاقی و ثانویّه نسبت به ثبت مکان مقدار داده در نظر گرفت. با این حال، سایر رویکردهای نمایش داده‌ها، مقادیر داده‌ها را به صورت اولیّه یا مستقیماً بر اساس مساحت و بدون انتساب نقطه مکانی نشان می‌دهند. رایج‌ترین آن‌ها

نمودار دایره‌ای است (نمودار ۱۷–۱۱). هرچند از نظر فنّی مقادیر داده‌ها بر روی زوایه‌های دایره ترسیم می‌شوند که در حقیقت موقعیّت مکانی در محوری دایره‌ای دارند، در عمل معمولاً زوایای نمودار دایره‌ای مورد قضاوت قرار نمی‌گیرند. در عوض، ویژگی بصری غالبی که مورد توجّه قرار می‌گیرد، مساحت هر قطاع نمودار دایره‌ای است.



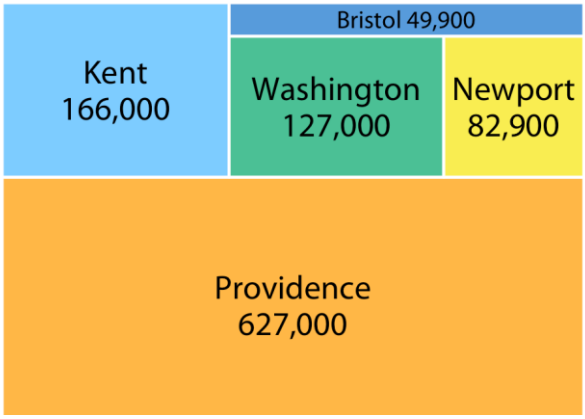
نمودار ۱۷–۱۱. تعداد ساکنان شهرستان‌های رود آیلند، که به صورت نمودار دایره‌ای نشان داده شده است. هم زاویه و هم مساحت هر قطاع نمودار دایره‌ای متناسب با تعداد ساکنان شهرستان مربوطه است.

از آنجایی که مساحت هر قطاع نمودار دایره‌ای با زاویهٔ آن و از سوی دیگر زاویهٔ آن با مقدار داده‌ای که قطاع نشان می‌دهد متناسب است، نمودارهای دایره‌ای از اصل جوهر متناسب پیروی می‌کنند. با این حال، مساحت در نمودار دایره‌ای متفاوت از همان مساحت در نمودار میله‌ای درک می‌شود. دلیل اساسی این است که ادراک انسان در درجهٔ اوّل فاصله‌ها را قضاوت می‌کند و نه مساحت‌ها را. بنابراین، اگر یک مقدار داده به صورت یک فاصله نمایش داده شود، مثلاً نمایش به صورت طول میله، انسان با دقّت بیشتری آن را درک می‌کند نسبت به زمانی که مقدار داده از طریق ترکیبی از دو یا چند فاصله که به طور مشترک یک مساحت را ایجاد می‌کنند، نمایش داده شود. برای مشاهدهٔ این تفاوت، نمودار ۱۷–۱۱را با نمودار ۱۷–۱۲ که داده‌های مشابهی را با میله نشان می‌دهد، مقایسه کنید. تفاوت تعداد ساکنان بین شهرستان پراویدنس و سایر شهرستان‌ها در نمودار ۱۷–۱۲ بزرگتر از نمودار ۱۷–۱۱ به نظر می‌رسد.



نمودار ۱۷–۱۲. تعداد ساکنان شهرستان‌های رود آیلند، که به صورت میله نشان داده شده است. طول هر میله متناسب با تعداد ساکنان شهرستان مربوطه است.

این مشکل که درک انسان در قضاوت فواصل بهتر از قضاوت مساحت‌ها می‌باشد، در نقشه‌های درختی، که می‌توان به عنوان نسخهٔ مربعی نمودارهای دایره‌ای در نظر گرفته شود، نیز صادق است (نمودار ۱۷–۱۳)، باز هم، در مقایسه با نمودار ۱۷–۱۲، تفاوت در تعداد ساکنان در میان شهرستان‌ها در نمودار ۱۷–۱۳ کمتر از حد واقعی به نظر می‌رسد.



نمودار ۱۷–۱۳. تعداد ساکنان شهرستان‌های رود آیلند، که به صورت نقشهٔ درختی نشان داده شده است. مساحت هر مستطیل متناسب با تعداد ساکنان شهرستان مربوطه است.

فصل ۱۸

مدیریت نقاط همپوشان

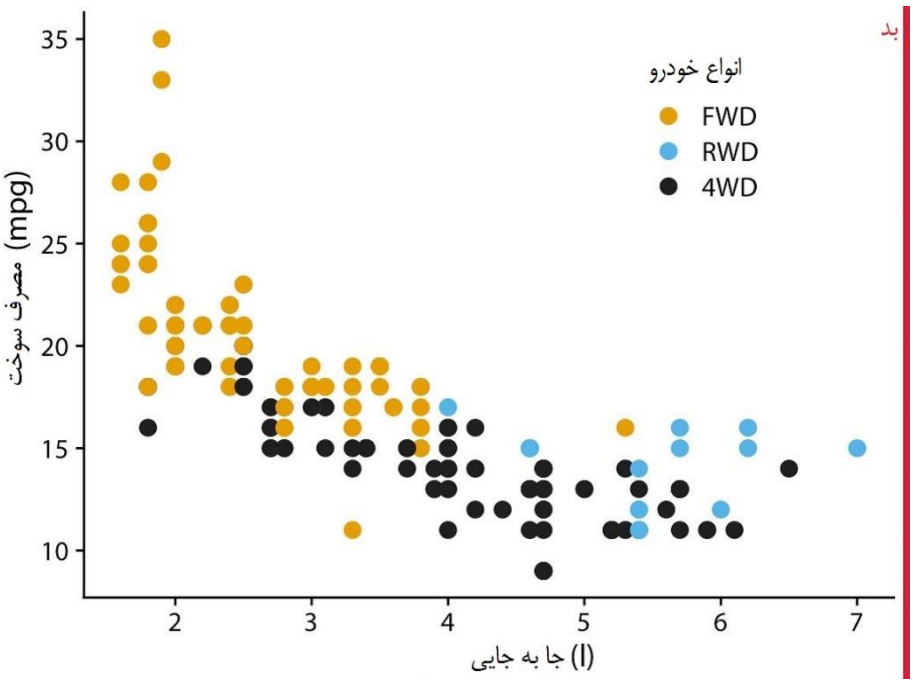
زمانی که هدف نمایش مجموعه داده‌های بزرگ یا بسیار بزرگ است، اغلب این چالش وجود دارد که نمودارهای ساده پراکنش x-y عملکرد مناسبی ندارند، زیرا بسیاری از نقاط روی هم قرار گرفته‌اند و تا حدی یا به طور کامل همپوشانی دارند. همچنین اگر مقادیر داده‌ها با دقّت کم یا به صورت گرد شده ثبت شوند، به طوری که مشاهدات متعدد دقیقاً مقادیر یکسانی داشته باشند، ممکن است حتّی در مجموعه‌های داده کوچک نیز مشکلات مشابهی ایجاد شود. اصطلاح فنّی که معمولاً برای توصیف این وضعیّت استفاده می‌شود همپوشانی نقاط[1] است، به این معنی که نقاط زیادی روی هم ترسیم شده‌اند. در ادامه چندین راهبرد توضیح داده می‌شود که می‌توان در هنگام مواجهه با این چالش‌ها از آن‌ها بهره برد.

شفافیت جزئی و لرزانش

بیایید در ابتدا سناریویی با تعداد متوسطی از نقاط داده که به طور قابل توجهی گرد شده‌اند را در نظر بگیریم. این مجموعه داده شامل مصرف سوخت در حین رانندگی در شهر و حجم موتور برای ۲۳۴ مدل خودروی محبوب است که بین سال‌های ۱۹۹۹ و ۲۰۰۸ عرضه شدند (نمودار ۱۸–۱). در این مجموعهٔ داده، مصرف سوخت بر حسب مایل بر گالن (mpg)

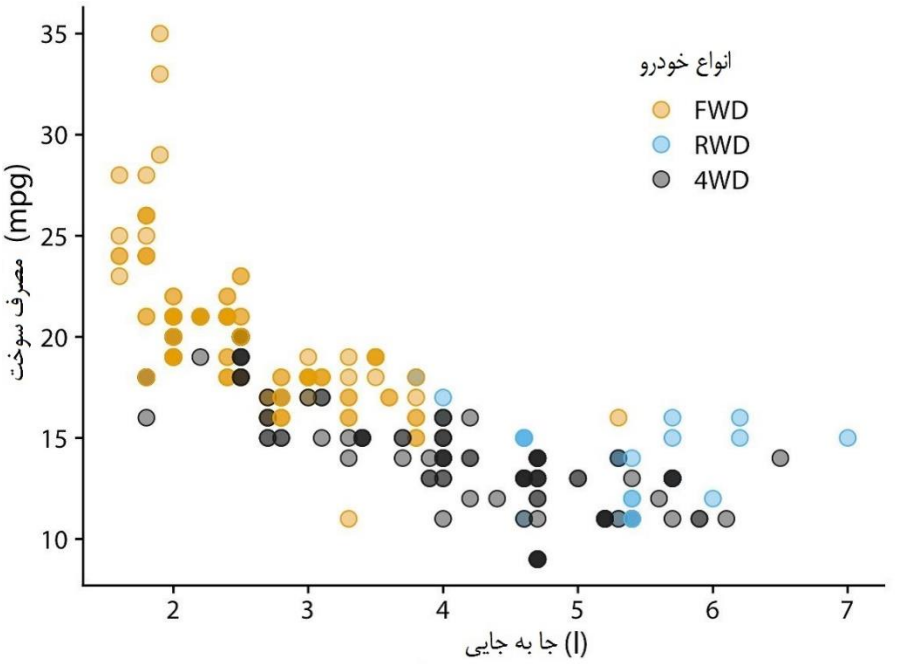
1. overplotting

انـدازه‌گیری می‌شـود و بـه نزدیـک‌تـرین مقـدار صـحیح گـرد می‌شـود. حجـم موتـور بـر حسـب لیتـر اندازه‌گیری می‌شود و به نزدیـک‌ترین دسـی لیتـر گـرد می‌شـود. بـه دلیـل ایـن گـرد کـردن، بسـیاری از مدل‌های خـودرو دارای مقـادیر دقیقـاً یکسـان هسـتند. بـه عنـوان مثـال، در مجمـوع ۲۱ خـودرو بـا حجم موتور ۲٫۰ لیتری وجود دارد و بـه عنـوان یـک گـروه، آن‌هـا تنهـا چهـار مقـدار مصـرف سـوخت متفاوت دارند: ۱۹، ۲۰، ۲۱ یـا ۲۲ مایـل بـر گـالن. بنـابراین، در شـکل ۱۸–۱ ایـن ۲۱ خـودرو تنهـا بـا چهار نقطه متمایز نشـان داده شـده‌اند، بـه طـوری کـه محبوبیـت موتورهـای ۲٫۰ لیتـری بسـیار کمتـر از آنچه که هست، بـه نظـر می‌رسـد. عـلاوه بـر ایـن، مجموعـهٔ داده شـامل دو خـودروی چهـار چـرخ متحرّک با موتورهای ۲٫۰ لیتری است کـه بـا نقـاط سـیاه نشـان داده شـده اسـت. بـا ایـن حـال، ایـن نقاط سیاه به طور کامل توسّط نقـاط زرد پوشـیده شـده اسـت، بـه طـوری کـه بـه نظـر می‌رسـد هـیچ خودروی چهار چرخ متحرّکی با موتور ۲٫۰ لیتری وجود ندارد.

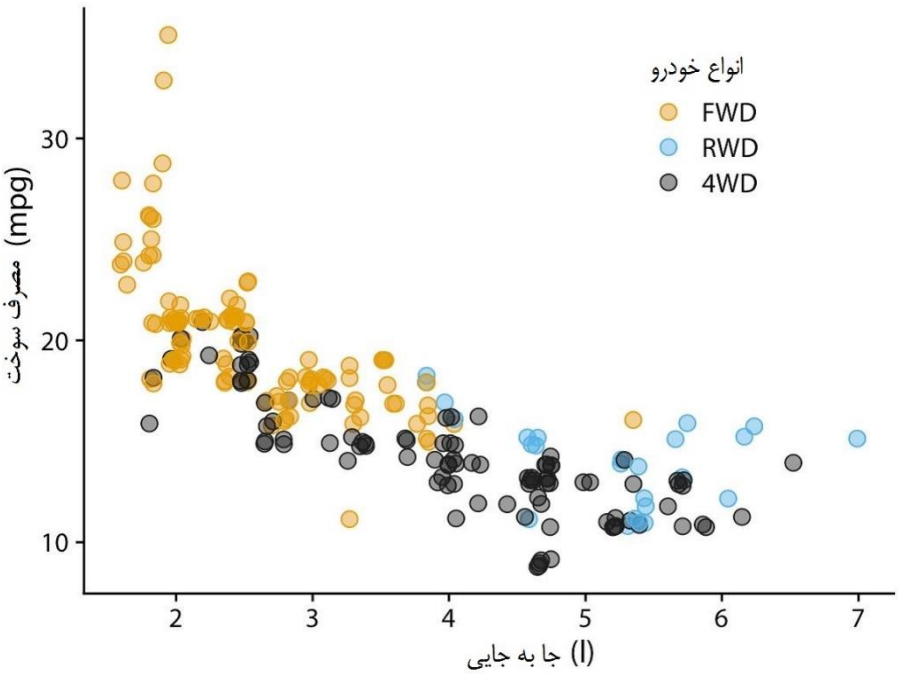


شکل ۱۸–۱. مصرف سوخت شهری در مقابل حجم موتور، برای خودروهای محبوبی که بین سال‌های ۱۹۹۹ و ۲۰۰۸ عرضه شدند. هر نقطه نشان‌دهندهٔ یک خودرو است. رنگ نقطه، الگوی نیروی محرکه را نشان می‌دهد: دیفرانسیل جلو (FWD)، دیفرانسیل عقب (RWD)، یا چهار چرخ متحرّک (WD۴). این شکل برچسب «بد» خورده است زیرا بسیاری از نقاط بر روی نقاط دیگر ترسیم شده و آن‌ها را محو می‌کنند. منبع داده: آژانس حفاظت از محیط زیست ایالات متحده (EPA)، https://fueleconomy.gov.

یکی از راه‌های حل این مشکل استفاده از شفافیت جزئی است. اگر نقاط منفرد تا حدی شفاف شوند، آنگاه نقاط همپوشان به صورت نقاط تیره‌تر ظاهر می‌شوند و بنابراین سایه نقاط، چگالی نقاط را در آن مکان از نمودار منعکس می‌کند (نمودار ۱۸-۲). با این حال، شفاف‌سازی نقاط همیشه برای حل مسئله همپوشانی نقاط کافی نیست. به‌عنوان مثال، حتی اگر در نمودار ۱۸-۲ مشاهده کنیم که برخی از نقاط سایه تیره‌تری نسبت به سایرین دارند، تخمین اینکه در هر مکان چند نقطه روی هم ترسیم شده‌اند، دشوار است. علاوه بر این، در حالی که تفاوت در سایه‌ها به وضوح قابل مشاهده است، آن‌ها گویا نیستند. خواننده‌ای که برای اوّلین بار این نمودار را می‌بیند احتمالاً تعجّب می‌کند که چرا برخی از نقاط تیره‌تر از سایرین هستند و متوجّه نمی‌شود که آن نقاط در واقع چندین نقطه هستند که روی هم چیده شده‌اند. ترفند ساده‌ای که در این شرایط کمک‌کننده است، اعمال مقدار کمی لرزانش بر روی نقاط است — به بیان دیگر بایستی هر نقطه به طور تصادفی مقدار کمی در جهت x یا y یا هر دو جابجا شود. با اعمال لرزانش، به سرعت مشخّص می‌شود که نواحی تیره‌تر از همپوشانی نقاط به وجود می‌آیند (نمودار ۱۸-۳). همچنین اکنون برای اوّلین بار نقاط سیاهی که نشان‌دهندهٔ خودروهای چهارچرخ متحرّک با موتورهای ۲ لیتری هستند نیز دیده می‌شود.

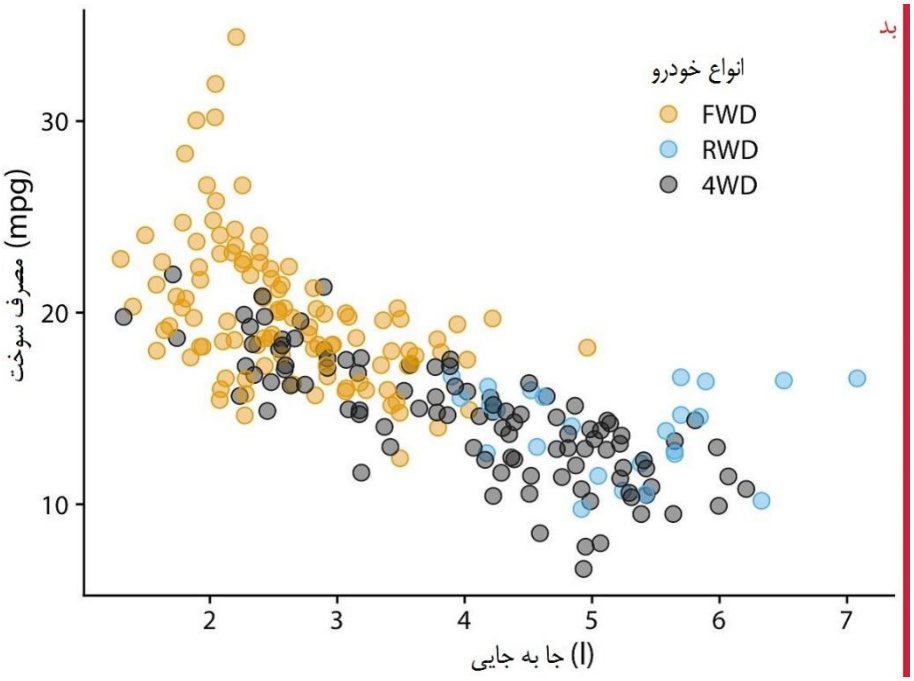


نمودار ۱۸-۲، مصرف سوخت شهری در مقابل حجم موتور. از آنجایی که نقاط تا حدی شفاف شده‌اند، اکنون می‌توان نقاطی که روی نقاط دیگر قرار دارند را از طریق سایه تیره‌ترشان شناسایی کرد. منبع داده: EPA



نمودار ۱۸–۳. مصرف سوخت شهری در مقابل حجم موتور. با افزودن مقدار کمی لرزانش به هر نقطه، می‌توان امکان مشاهده نقاط همپوشان شده را بدون تحریف اساسی پیام شکل افزایش داد. منبع داده: EPA

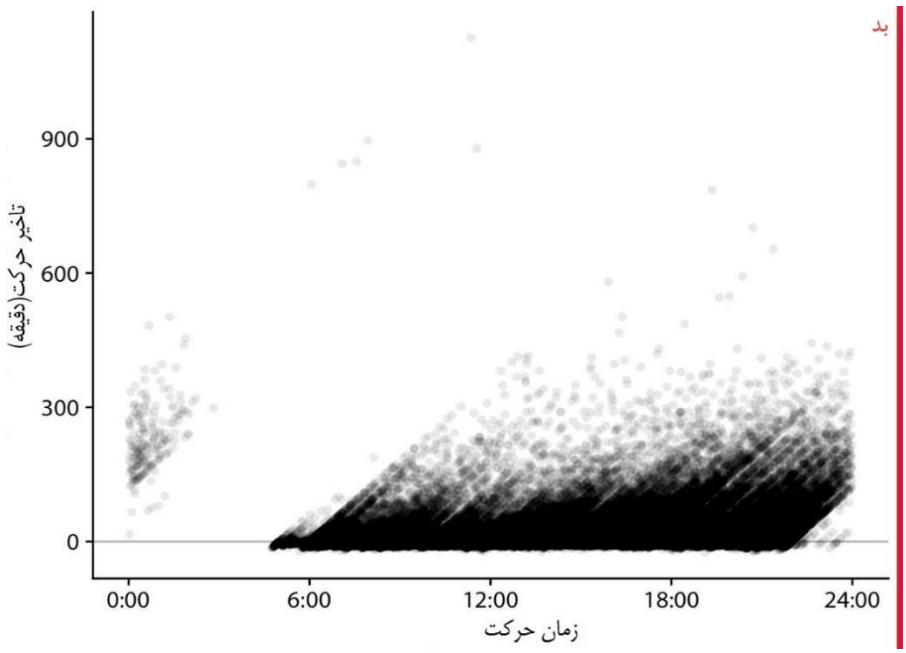
یکی از نکات منفی لرزانش این است که داده‌ها را تغییر می‌دهد و بنابراین باید با دقّت کامل انجام شود. اگر اعمال لرزش بیش از حد باشد، در نهایت نقاط در مکان‌هایی قرار خواهند گرفت که نمایندهٔ مجموعه داده‌های اصلی نیستند. نتیجه نهایی، نموداری گمراه‌کننده از داده‌های موجود است. به عنوان نمونه نمودار ۱۸–۴ را ببینید.



نمودار ۱۸–۴. مصرف سوخت شهری در مقابل حجم موتور. با افزودن لرزانش بیش از حد به نقاط، نموداری ایجاد شده که منعکس کنندهٔ داده‌های مربوطه نیست. منبع داده: EPA

هیستوگرام‌های دوبعدی

وقتی تعداد نقاط منفرد بسیار زیاد می‌شود، شفافیت جزئی (با یا بدون لرزانش) برای حل مشکل همپوشانی نقاط کافی نخواهد بود. آنچه معمولاً اتفاق می‌افتد این است که نواحی با تراکم نقطهٔ بالا به صورت حباب‌های یکنواخت با رنگ تیره ظاهر می‌شوند، در حالی که در مناطق با تراکم پایین، نقاط منفرد به سختی قابل مشاهده هستند (نمودار ۱۸–۵). تغییر در مقدار شفافیت نقاط یکی از این مشکلات را بهبود می‌بخشد و در عین حال دیگری را بدتر می‌کند. در مجموع هیچ سطحی از شفافیت نمی‌تواند هر دو مشکل را به طور همزمان برطرف کند.

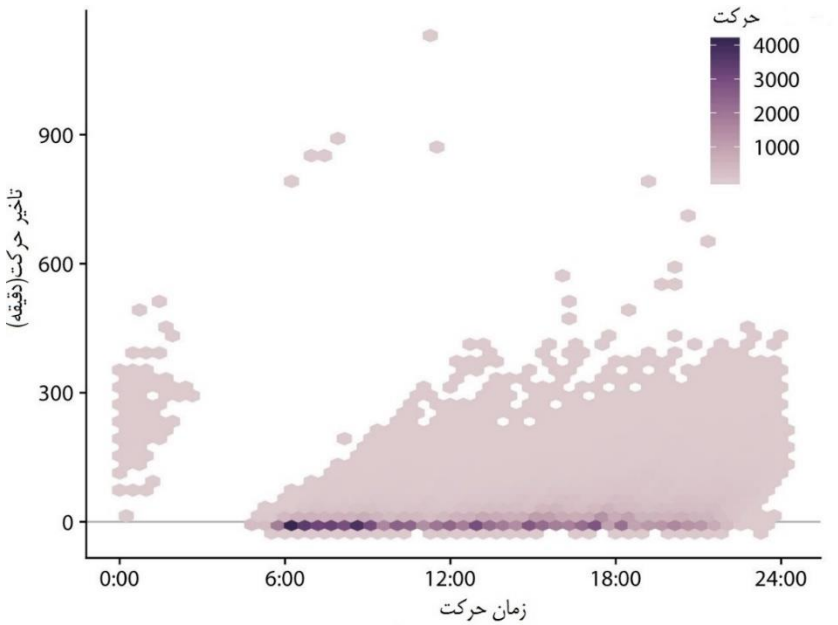


نمودار ۱۸–۵. تأخیر حرکت بر حسب دقیقه در مقابل زمان پرواز برای همه پروازهای خروجی از فرودگاه نیوآرک (EWR) در سال ۲۰۱۳. هر نقطه نشان‌دهندهٔ یک پرواز است. منبع داده: اداره حمل و نقل ایالات متحده، اداره آمار حمل و نقل.

نمودار ۱۸–۵ تأخیر حرکت را برای بیش از ۱۰۰۰۰۰ پرواز مجزا نشان می‌دهد و هر نقطه نشان‌دهندهٔ یک پرواز خروجی است. با وجود اینکه تک‌تک نقاط نسبتاً شفاف شده است، اکثر آن‌ها یک نوار سیاه را برای تاخیر پرواز بین ۰ تا ۳۰۰ دقیقه تشکیل می‌دهند. این نوار مشخّص نمی‌کند که آیا اکثر پروازها تقریباً به موقع یا با تأخیر قابل توجّه (مثلاً ۵۰ دقیقه یا بیشتر) حرکت نموده‌اند. همچنین، اغلب پروازهایی که بیشترین تاخیر را دارند (با تاخیر ۴۰۰ دقیقه یا بیشتر) به دلیل شفاف بودن نقاط به سختی قابل مشاهده هستند.

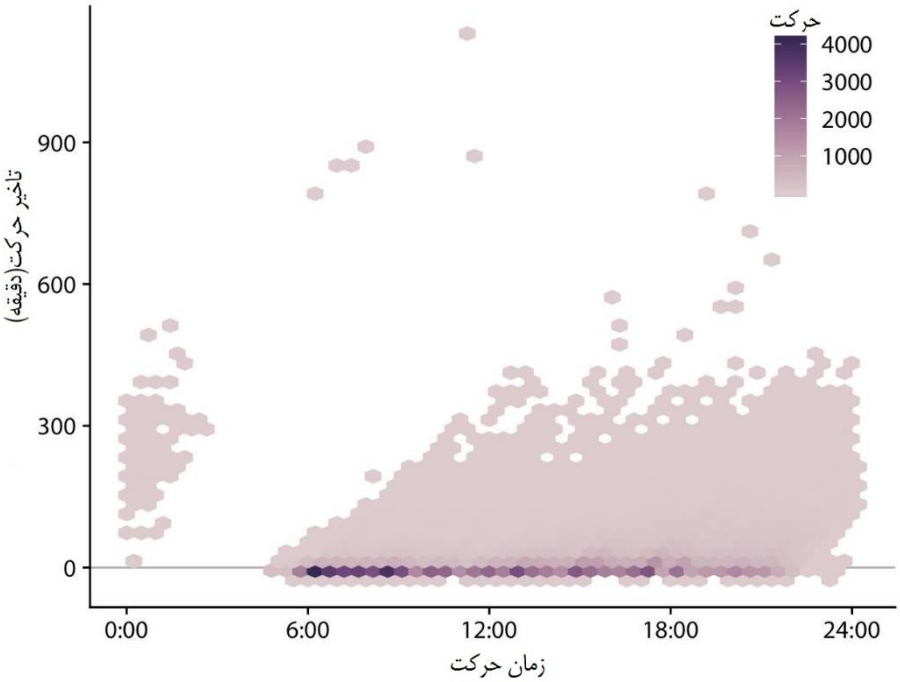
در چنین مواردی، به جای ترسیم نقاط منفرد، می‌توان یک هیستوگرام دو بُعدی رسم نمود. یک هیستوگرام دوبعدی از نظر مفهومی مشابه هیستوگرام یک بُعدی است، همانطور که در فصل ۷ بحث شد، امّا اکنون داده‌ها را در دو بعد رسم می‌کنیم. بدین منظور کل صفحه x-y به مستطیل‌های کوچک تقسیم شده، تعداد مشاهداتی که در هر یک قرار می‌گیرند شمارش شده و سپس مستطیل‌ها بر اساس این تعداد رنگ می‌شوند. نمودار ۱۸–۶ نتیجهٔ این رویکرد را برای داده‌های تاخیر خروج نشان می‌دهد. این نمودار چندین ویژگی مهم داده‌های خروج پرواز را

برجسته می‌کند. اولاً، اکثریّت قریب به اتفاق پروازها در طول روز (از ساعت ۶ صبح تا حدود ۹ شب) در واقع بدون تأخیر یا حتی زودتر از موعد انجام می‌شود (تاخیر منفی). با این حال، برخی از پروازها تاخیر قابل توجّهی دارند. علاوه بر این، هر چه یک هواپیما دیرتر در روز پرواز کند، تاخیر بیشتری می‌تواند داشته باشد. نکته مهم این است که زمان حرکت، زمان واقعی حرکت است، نه زمان برنامه‌ریزی شدهٔ حرکت، بنابراین این نمودار لزوماً به ما نمی‌گوید که هواپیماهایی که قرار است در ساعات اولیه روز حرکت کنند، هرگز با تأخیر مواجه نمی‌شوند. با این حال، چیزی که به ما می‌گوید این است که اگر یک هواپیما زودتر در ساعات اولیه روز پرواز کند، یا تاخیر کمی دارد یا در موارد بسیار نادر، حدود ۹۰۰ دقیقه تاخیر دارد.



نمودار ۱۸–۶. تاخیر حرکت بر حسب دقیقه در مقابل زمان پرواز. هر مستطیل رنگی نشان‌دهندهٔ تمام هواپیماهایی است که در آن زمان با آن تاخیر حرکت پرواز نموده‌اند. رنگ‌آمیزی نشان‌دهندهٔ تعداد پروازهایی است که توسّط آن مستطیل نشان داده شده است. منبع داده: اداره حمل و نقل ایالات متحده، اداره آمار حمل و نقل.

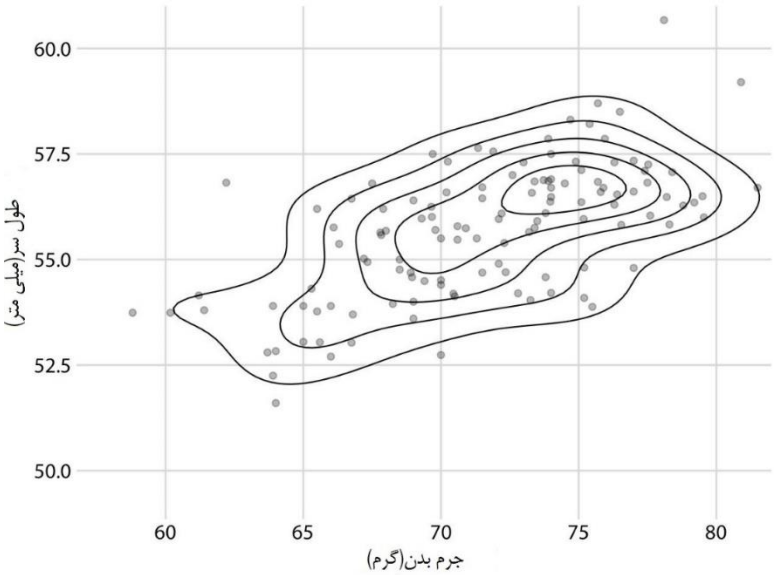
به عنوان جایگزینی برای رسم داده‌ها در مستطیل‌ها، می‌توان از شش ضلعی استفاده کرد [Carr et al. 1987]. این روش این مزیّت را دارد که نقاط یک شش ضلعی به طور متوسط به مرکز شش ضلعی نزدیک‌تر هستند تا نقاط یک مربع به مرکز مربع. بنابراین، شش ضلعی‌های رنگی داده‌ها را کمی دقیق‌تر از مستطیل‌های رنگی نشان می‌دهند. نمودار ۱۸–۷ داده‌های پروازهای خروجی را با رسم شش ضلعی به جای مستطیل نشان می‌دهد.



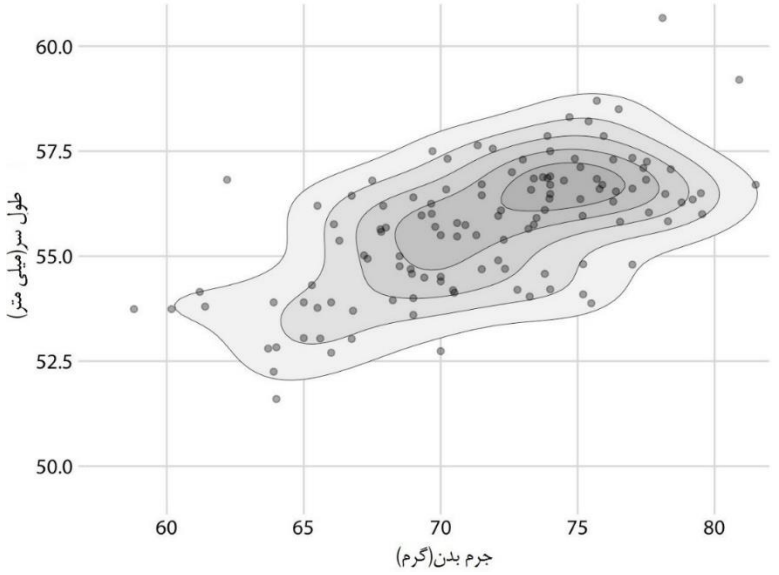
نمودار ۱۸–۷. تاخیر حرکت بر حسب دقیقه در مقابل زمان پرواز. هر شش ضلعی رنگی نشان‌دهندهٔ تمام هواپیماهایی است که در آن زمان با آن تاخیر حرکت، پرواز می‌کنند. رنگ‌آمیزی نشان‌دهندهٔ تعداد پروازهایی است که توسّط آن شش ضلعی نشان داده شده است. منبع داده: اداره حمل و نقل ایالات متحده، اداره آمار حمل و نقل.

خطوط ترازنما

به جای اینکه نقاط داده را در مستطیل یا شش ضلعی قرار دهیم، می‌توان چگالی نقاط را در سراسر نمودار تخمین زده و مناطقی با تراکم مختلف را با خطوط ترازنما مجزا کرد. این روش زمانی عملکرد خوبی دارد که چگالی نقاط داده در هر دو محور x و y به آرامی تغییر کند. به عنوان مثالی برای این رویکرد، به مجموعهٔ داده‌های زاغ آبی از فصل ۱۲ باز می‌گردیم. نمودار ۱۲–۱ رابطهٔ بین طول سر و تودهٔ بدن را برای ۱۲۳ زاغ آبی نشان می‌دهد، و مقداری همپوشانی بین نقاط وجود دارد. می‌توان با کوچک‌تر کردن و تا حدی شفاف‌تر کردن نقاط و ترسیم آن‌ها در بالای خطوط هم ترازی که مناطقی با تراکم نقطه‌ای مشابه را مشخّص می‌کنند، توزیع داده‌ها را واضح‌تر نمود (نمودار ۱۸–۸). همچنین با سایه‌زنی نواحی محصور شده توسّط خطوط هم تراز، با استفاده از رنگ‌های تیره‌تر برای مناطقی که تراکم نقطهٔ بالاتر را نشان می‌دهند، درک مخاطب از تغییرات چگالی نقاط را افزایش داد (نمودار ۱۸–۹).

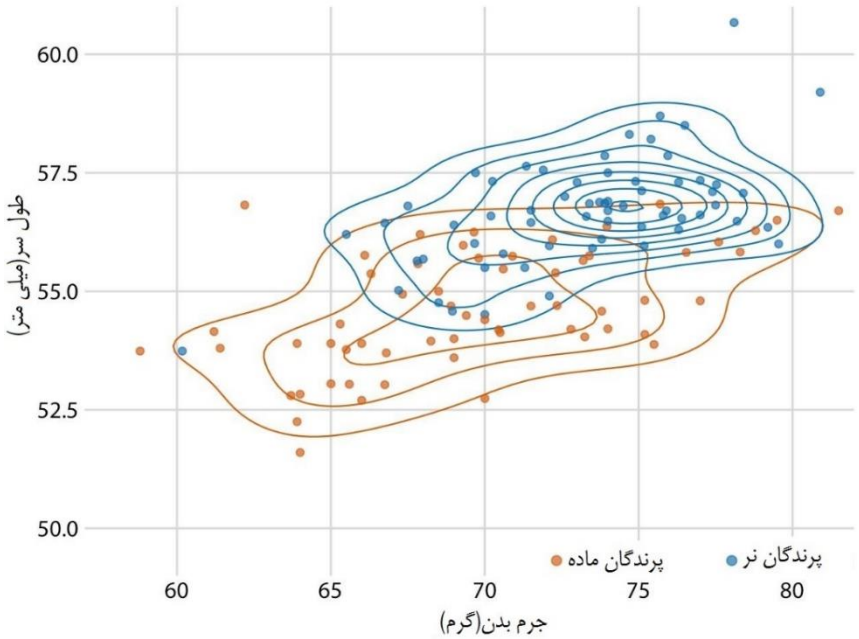


نمودار ۱۸–۸. طول سر در برابر تودهٔ بدن برای ۱۲۳ زاغ آبی، (نمودار ۱۲–۱)، هر نقطه مربوط به یک پرنده است و خطوط مرز مناطقی با تراکم مشابه را نشان می‌دهند. چگالی نقطه‌ای به سمت مرکز نمودار، یعنی جرم بدن ۷۵ گرم و طول سر بین ۵۵ میلی متر تا ۵۷٫۵ میلی متر افزایش می‌یابد. منبع داده: Keith Tarvin, Oberlin College



نمودار ۱۸–۹. طول سر در مقایسه با تودهٔ بدن برای ۱۲۳ زاغ آبی. این نمودار تقریباً مشابه نمودار ۱۸–۸ است، امّا اکنون نواحی محصور شده توسّط خطوط حدفاصل با سایه‌هایی در طیف خاکستری مشخّص‌تر شده‌اند. این سایه‌زنی، تصوّر بصری قوی‌تری از افزایش تراکم نقطه‌ای به سمت مرکز داده‌ها ایجاد می‌کند. منبع داده: Keith Tarvin, Oberlin College

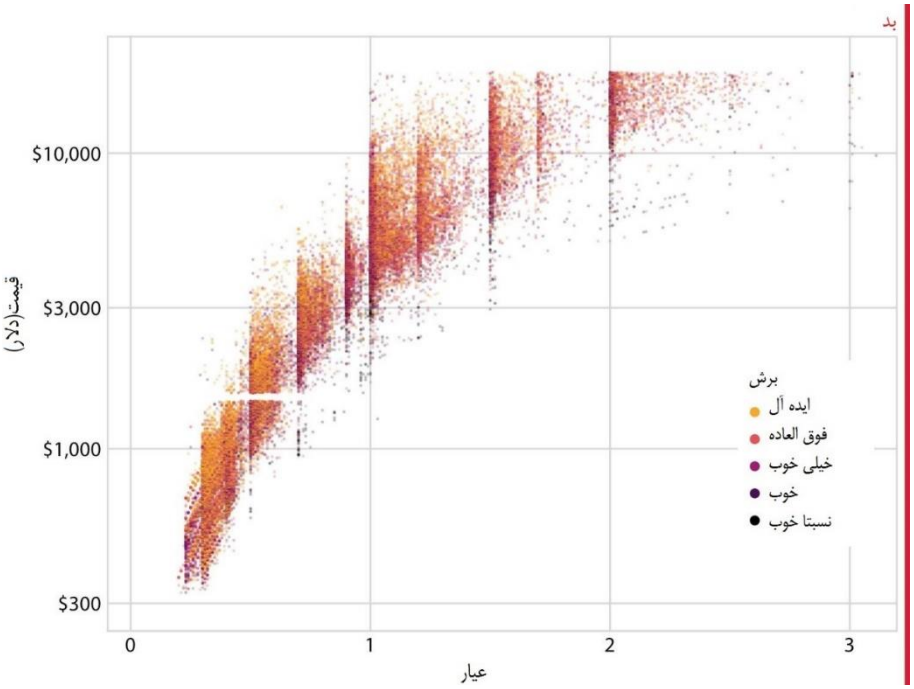
در فصل ۱۲، رابطهٔ بین طول سر و تودهٔ بدن به طور جداگانه برای پرندگان نر و ماده ارائه شد (نمودار ۱۲–۲). می‌توان همین کار را با خطوط ترازنما، با ترسیم خطوط رنگی جداگانه برای پرندگان نر و ماده انجام داد (نمودار ۱۸–۱۰).



نمودار ۱۸–۱۰. طول سر در مقایسه با تودهٔ بدن برای ۱۲۳ زاغ آبی. همانند نمودار ۱۲–۲، هنگام ترسیم خطوط ترازنما نیز می‌توان جنسیّت پرندگان را با رنگ مشخّص نمود. این نمودار نشان می‌دهد که چگونه توزیع داده‌ها برای پرندگان نر و ماده متفاوت است. به طور خاص، پرندگان نر در یک منطقه از کل نمودار متراکم‌تر هستند در حالی که پرندگان ماده الگوی پراکنده‌تری در کل نمودار دارند. منبع داده: Keith Tarvin, Oberlin College

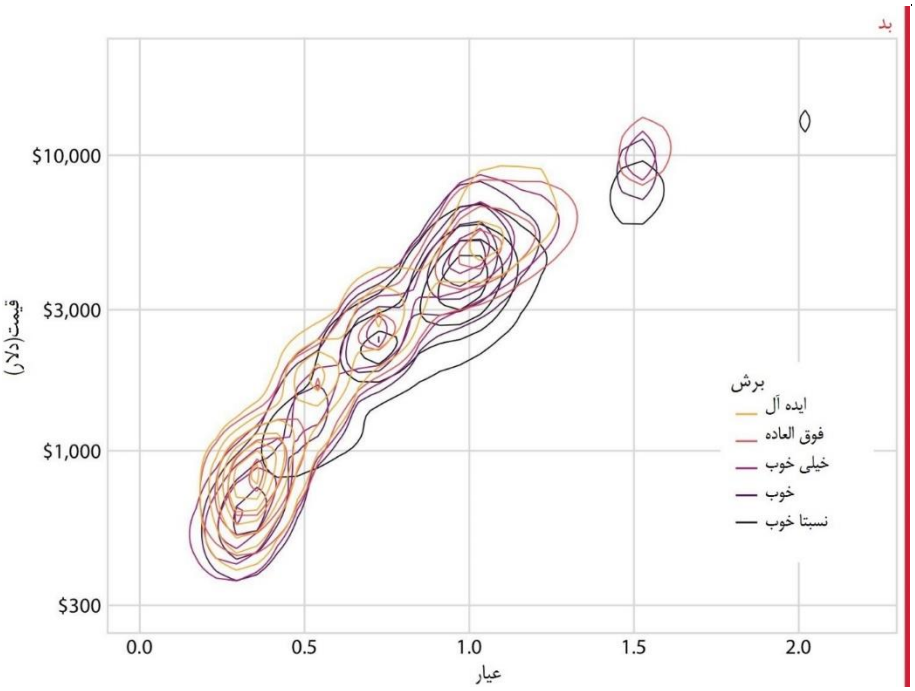
ترسیم مجموعه‌های متعددی از خطوط ترازنما در رنگ‌های مختلف می‌تواند یک راهبرد قدرتمند برای نشان دادن همزمان توزیع چندین ابر از داده‌ها باشد. با این حال، این روش باید با دقّت مورد استفاده قرار گیرد. این روش تنها زمانی کارآیی دارد که تعداد گروه‌ها با رنگ‌های متمایز کم باشد (دو تا سه) و گروه‌ها به وضوح از هم متمایز باشند. در غیر این صورت، ممکن است با یک گلوله مو از خطوطی با رنگ‌های متفاوت روبرو شویم که همگی روی هم افتاده و اصلاً الگوی خاصی را نشان نمی‌دهند.

برای نشان دادن این مشکل بالقوه، مجموعه داده‌های الماس که حاوی اطّلاعات ۵۳۹۴۰ الماس، از جمله قیمت، وزن (قیراط) و نحوه تراش آن‌ها است را بررسی خواهیم کرد. نمودار ۱۸–۱۱ این مجموعه داده را به صورت نمودار پراکنش نشان می‌دهد. این نمودار وضعیّت همپوشانی شدید نقاط را نشان می‌دهد. نقاط رنگارنگ بسیار زیادی روی یکدیگر قرار گرفته‌اند به طوری که استخراج اطّلاعاتی فراتر از طرح کلی ارتباط قیمت–وزن این الماس‌ها غیرممکن است.



نمودار ۱۸–۱۱. قیمت ۵۳۹۴۰ قطعه الماس در برابر وزن (قیراط) آنها، نحوه تراش هر الماس با رنگ مشخّص شده است. این طرح به عنوان «بد» برچسب‌گذاری شده است، زیرا همپوشانی بیش از حد نقاط مانع تشخیص هر گونه الگویی بین انواع مختلف تراش الماس می‌شود. منبع داده: Hadley Wickham، ggplot2.

همانند نمودار ۱۸–۱۰ می‌توان خطوط ترازنما رنگی را برای انواع مختلف تراش رسم نمود. با این حال، در مجموعه داده الماس، پنج رنگ متمایز وجود دارد و گروه‌ها به شدت با هم همپوشانی دارند. بنابراین، نمودار با خطوط ترازنما (نمودار ۱۸–۱۲) خیلی بهتر از نمودار پراکنش اصلی نیست (نمودار ۱۸–۱۱).

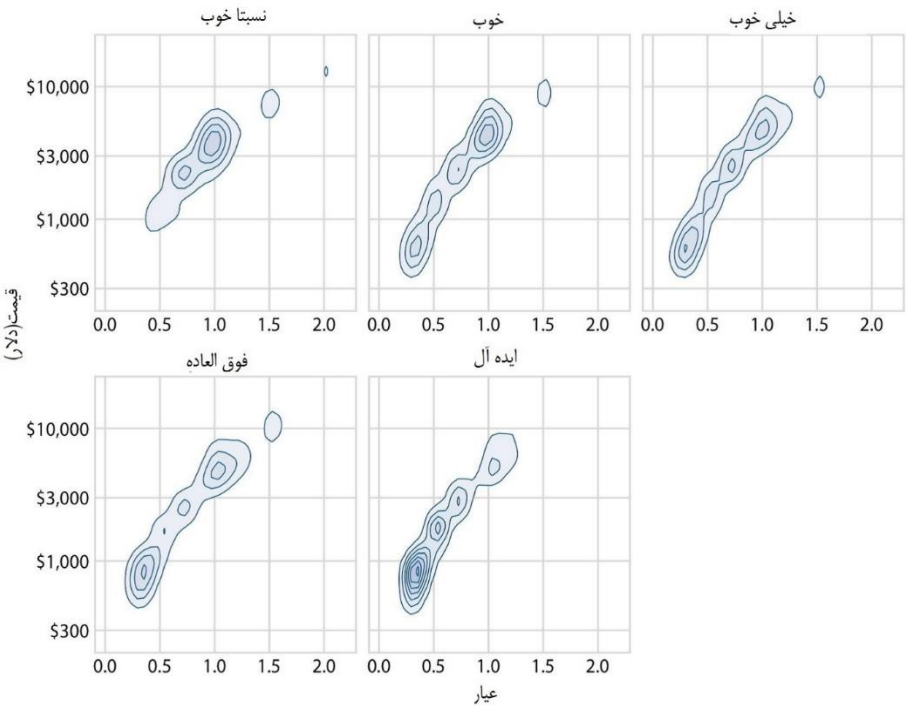


نمودار ۱۸–۱۲. قیمت الماس در برابر وزن (قیراط) آن‌ها. در مقایسه با نمودار ۱۸–۱۱ در اینجا نقاط داده با خطوط ترازنما جایگزین شده‌اند. نمودار حاصل همچنان به عنوان «بد» برچسب‌گذاری شده است، زیرا همهٔ خطوط ترازنما روی هم قرار می‌گیرند. نه توزیع نقطه‌ای برای انواع تراش‌های مختلف و نه توزیع نقطه‌ای کلی را نمی‌توان تشخیص داد. منبع داده: Hadley Wickham، ggplot2.

اقدام کمک کننده در اینجا، ترسیم خطوط ترازنما برای هر نوع تراش در خود نمودار است (نمودار ۱۸–۱۳). هدف از ترسیم همه آن‌ها در یک نمودار ایجاد امکان مقایسه بصری بین گروه‌ها می‌باشد، امّا نمودار ۱۸–۱۲ آنقدر شلوغ است که مقایسه امکان‌پذیر نیست. در عوض، در نمودار ۱۸–۱۳، شبکهٔ پس‌زمینه ما را قادر می‌سازد تا انواع تراش را با توجّه به محل قرارگیری خطوط ترازنما نسبت به خطوط مشبک مقایسه کنیم (اثر مشابهی را می‌توان با ترسیم نقاط منفرد نیمه شفاف به جای خطوط ترازنما در هر نمودار به دست آورد).

حال می‌توان دو روند اصلی را مشخّص نمود. اوّل، تراش‌های بهتر (بسیار خوب، ممتاز، ایده‌آل) نسبت به تراش‌های ضعیف‌تر (متوسط، خوب) وزن (قیراط) کمتری دارند. می‌دانید که قیراط معیاری برای وزن الماس است (۱ قیراط = ۰٫۲ گرم). تراش‌های بهتر منجر به (به طور متوسط) الماس‌های سبک‌تری می‌شود، زیرا برای ایجاد آن‌ها باید مواد بیشتری حذف شود.

دوّم، با وزن (قیـراط) یکسـان، تراش‌هـای بهتـر منجـر بـه قیمت‌هـای بـالاتر می‌شـود. بـرای مشـاهده این الگو، به عنـوان مثـال بـه توزیـع قیمـت در ۰٫۵ قیـراط نگـاه کنیـد. بـرای تراش‌هـای بهتـر توزیـع به سمت بالا متمایل می‌شـود، و ایـن مسـاله بـه ویـژه بـرای الماس‌هـایی بـا تـراش ایـده‌آل بـه طـور قابل توجهی بیشتر از الماس‌هایی با تراش متوسط یا خوب است.



نمودار ۱۸–۱۳. قیمت الماس در برابر وزن (قیراط) آن‌ها، در اینجا، خطوط ترازنما از نمودار ۱۸–۱۲ برداشته و به طور جداگانه برای هر تراش رسم شده است. اکنون می‌توان دید که تراش‌های بهتر (بسیار خوب، ممتاز، ایده‌آل) نسبت به تراش‌های ضعیف‌تر (متوسط، خوب) وزن (قیراط) کمتری دارند، امّا قیمت به ازای هر قیراط بالاتر است. منبع داده: Hadley Wickham، ggplot2.

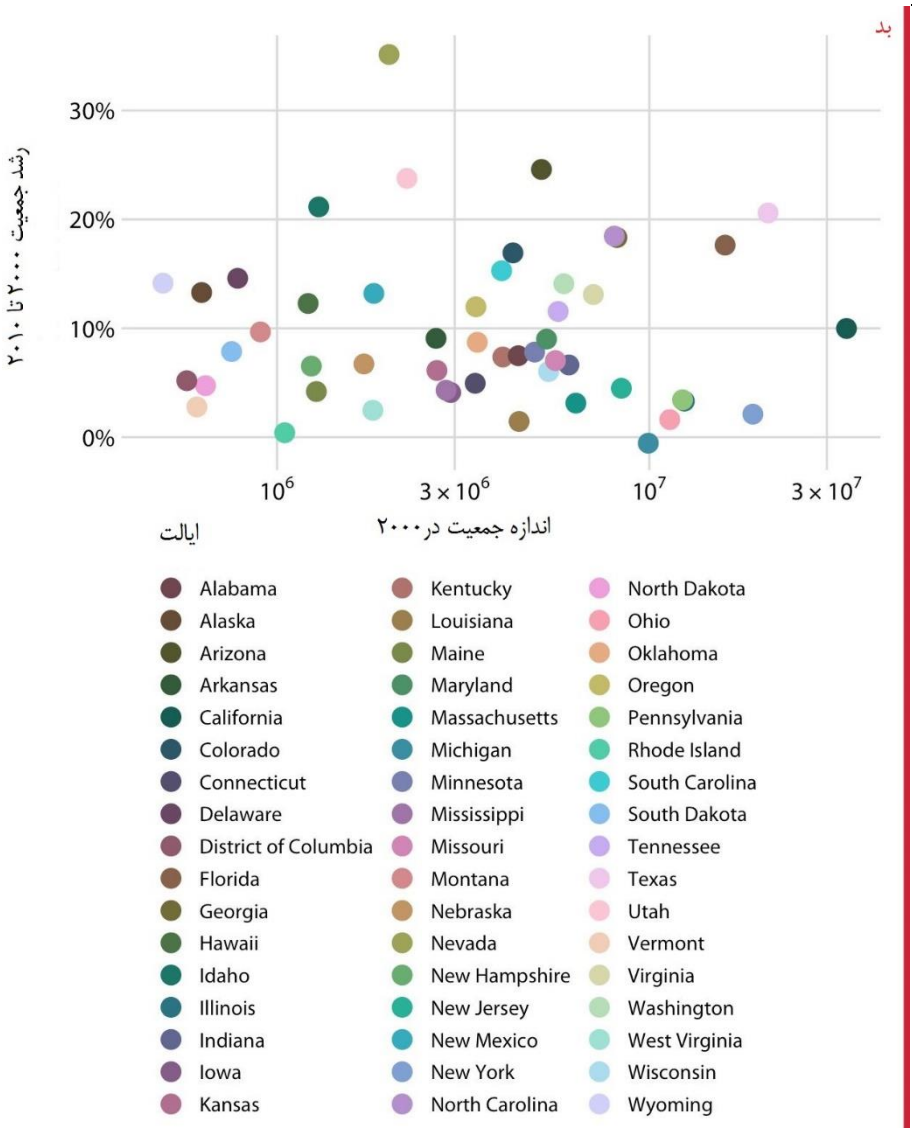
فصل ۱۹

اشتباهات رایج در استفاده از رنگ

رنگ می‌تواند یک ابزار فوق‌العاده برای نمایش داده‌های آماری باشد. در عین حال انتخاب رنگ نامناسب می‌تواند نمایش ایده‌آل داده‌های آماری را به هم بزند. از رنگ باید به صورت هدفمند استفاده شود. به بیان دیگر هدف محقق از استفاده از رنگ بایستی واضح بوده و باعث سردرگمی نشود.

نمایش اطّلاعات بیش از حد یا نامرتبط

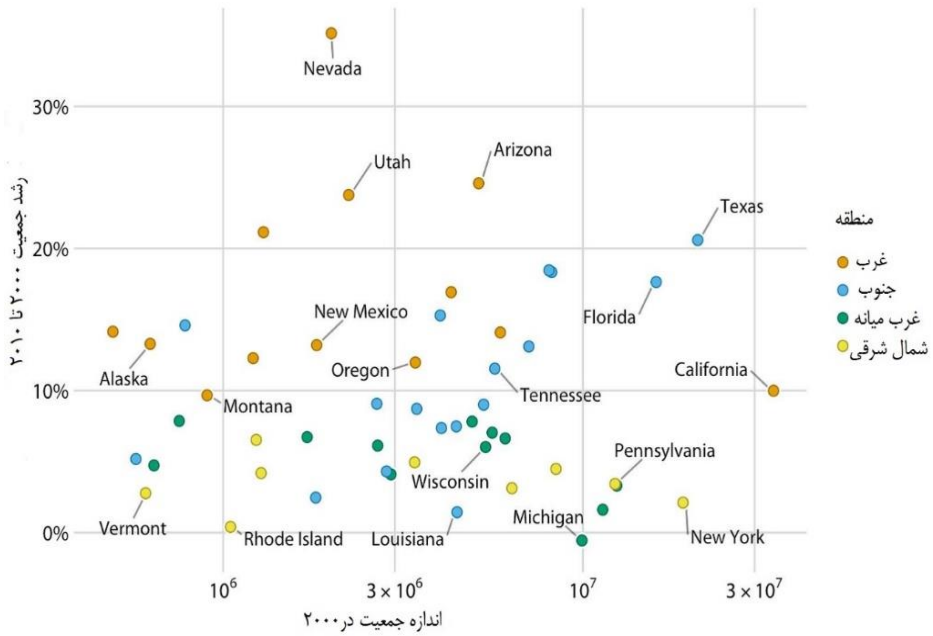
یک اشتباه رایج تلاش برای اختصاص دادن رنگ به داده‌های آماری است که تعدادشان به حدی زیاد است که استفاده از یک رنگ مجزا برای هر یک از داده‌ها منجر به پیچیدگی نمودار می‌شود. برای مثال نمودار ۱۹-۱ را در نظر بگیرید. این نمودار میزان رشد جمعیّت در برابر اندازهٔ جمعیّت را در ۵۰ ایالت آمریکا نشان می‌دهد. تلاش شده برای شناسایی هر ایالت یک رنگ به آن اختصاص داده شود امّا با این حال نتیجه خیلی کارآمد نیست. اگر چه می‌توان با نگاه کردن به نقاط رنگی نمودار و سپس راهنمای نمودار حدس زد که هر رنگ متعلق به کدام ایالت است ولی این کار خیلی سخت است که برای هر ۵۰ ایالت این تطبیق دادن رنگ‌ها در نمودار انجام شود. رنگ‌های زیادی در نمودار وجود دارند و بسیاری از آن‌ها خیلی شبیه هم هستند. اگر چه با تلاش بسیار می‌توان دریافت که هر رنگ متعلق به کدام ایالت است ولی این رنگ‌آمیزی که با هدف نمایش بهتر داده‌های آماری است با شکست مواجهه می‌شود. باید از رنگ‌ها در بهبود نمودارها و آسان ساختن نمایش داده‌های آماری استفاده نمود نه این که به کار بردن رنگ‌های مختلف منجر به یک معمای تصویری و ابهام در داده‌های آماری گردد.



نمودار ۱۹–۱. رشد جمعیّت از ۲۰۰۰ تا ۲۰۱۰ در برابر اندازه جمعیّت در سال ۲۰۰۰ در تمام ۵۰ ایالت آمریکا. هر ایالت با رنگ متفاوتی نشان داده شده است. از آنجایی که تعداد ایالت‌ها زیاد است یافتن ایالت متناظر مربوط به هر رنگ در راهنمای نمودار کار دشواری است. منبع داده: ادارهٔ سرشماری ایالات متحده

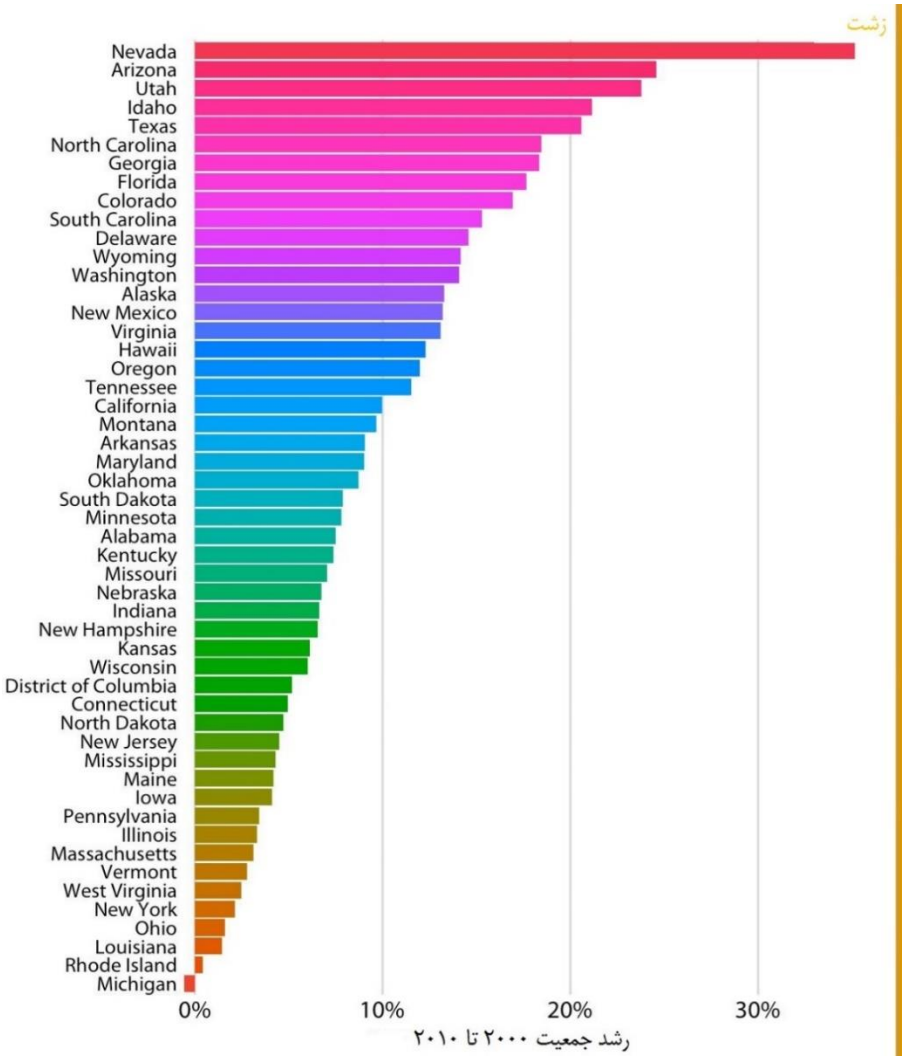
به عنوان یک قانون سرانگشتی، استفاده از مقیاس‌های کیفی رنگی زمانی بیشترین کارایی را دارد که بین سه تا پنج رنگ مختلف برای گروه‌های متفاوت استفاده شود. هنگامی که داده‌های آماری در ۸ تا ۱۰ گروه متفاوت یا بیشتر باشد، انطباق رنگ‌های مختلف متعلق به هر

گروه بسیار مشکل خواهد بود، حتی اگر رنگ‌ها به خوبی قابل تمایز باشند. برای داده‌های آماری نمودار ۱۹–۱ احتمالاً بهترین روش استفاده از یک رنگ خاص برای نشان دادن ایالت‌های هر ناحیه جغرافیایی است و برای شناسایی ایالت‌ها، بایستی نام هر یک در کنار نقطهٔ متناظر آن در نمودار حک شود (نمودار ۱۹–۲). هرچند نگارش نام هر ایالت در کنار نقطهٔ داده متناظر آن منجر به شلوغی نمودار می‌شود، برچسب‌گذاری مستقیم انتخاب صحیحی برای این نمودار است. در کل برای نمودارهایی مثل نمودار ۱۹–۱ نیازی به ثبت نام تک‌تک ایالت‌ها در کنار نقاط داده نیست. بهتر است تنها نام ایالت‌هایی که مشخصاً در متن قرار است در موردشان بحث شود، در نمودار ارائه شود. این امکان همیشه وجود دارد که داده‌های مهم مد نظر در قالب جدول نیز ارائه شود تا اطمینان حاصل شود که خواننده به همهٔ داده‌ها دسترسی دارد.



نمودار ۱۹–۲. رشد جمعیّت از ۲۰۰۰ تا ۲۰۱۰ در برابر اندازه جمعیّت در سال ۲۰۰۰ در تمام ۵۰ ایالت آمریکا. بر خلاف نمودار ۱۹–۱ در اینجا ایالت‌ها بر اساس منطقه رنگ‌بندی شده و نام ایالت‌های منتخب در کنار داده‌ها آمده است. از ذکر نام اغلب ایالت‌ها پرهیز شده تا از شلوغی نمودار پیشگیری شود. منبع داده: ادارهٔ سرشماری ایالات متحده

هنگامی که می‌خواهید بیش از هشت گروه را از هم متمایز نمایید، به جای استفاده از رنگ، نام آن‌ها را در کنار نقاط داده ذکر کنید.



نمودار ۱۹–۳ . رشد جمعیّت در آمریکا از سال ۲۰۰۰ تا ۲۰۱۰. رنگ‌آمیزی رنگین کمانی ایالت‌ها بدون هدف بوده و منجر به حواس پرتی می‌شود. علاوه بر این رنگ‌ها بیش از حد اشباع شده‌اند. منبع داده: ادارهٔ سرشماری ایالات متحده

مشکل رایج دوّم، استفاده از رنگ‌ها فقط جهت رنگ‌آمیزی (زیبایی) و بدون داشتن یک هدف روشن است. برای مثال نمودار ۱۹–۳ که نسخهٔ دیگری از نمودار ۴–۲ است، را در نظر بگیرید. در اینجا به جای اختصاص دادن یک رنگ مشترک به ایالت‌های هر ناحیهٔ جغرافیایی، به هر ایالت رنگی مخصوصی برای همان ایالت اختصاص داده شده که در مجموع یک اثر رنگین

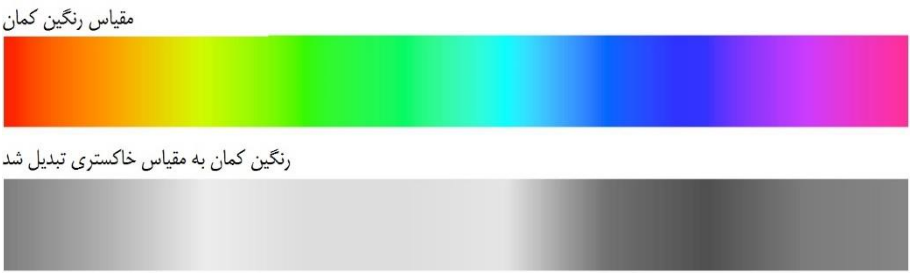
کمان تشکیل می‌دهد. این نوع نمایش داده‌های آماری شاید جالب به نظر برسد امّا بینش جدیدی در خصوص داده‌ها ایجاد نکرده و باعث درک آسان‌تر داده‌های آماری نیز نمی‌شود.

در کنار استفاده نابجا از رنگ‌های مختلف، نمودار ۱۹-۳ یک مشکل دیگر مربوط به رنگ‌آمیزی نیز دارد: رنگ‌های انتخاب شده بیش از حد اشباع شده و پر رنگ هستند. این شدت رنگ، نگاه کردن به نمودار را دشوار می‌کند. برای مثال دشوار است که نام ایالت‌ها را خواند و تلاش کرد از انحراف چشم به میله‌های پررنگ کنار آن جلوگیری نمود. همچنین مقایسهٔ نقطهٔ پایانی میله‌ها با خطوط راهنما نیز دشوار است.

از استفاده از طیف متنوع رنگ‌های اشباع شده پرهیز کنید. این کار مانع می‌شود که خواننده نمودار را به دقّت بررسی کند.

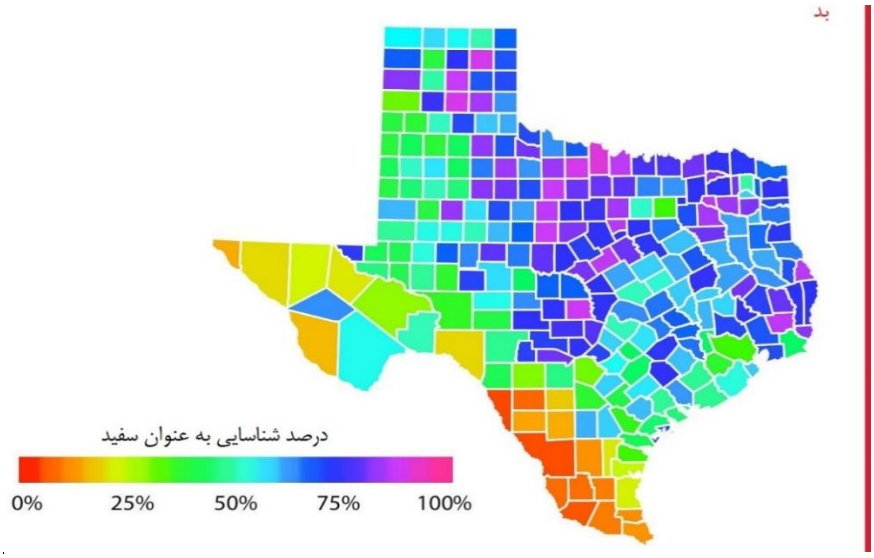
استفاده از رنگ‌های غیر یکنواخت برای نمایش مقادیر داده‌ها

در فصل ۴ دو شرط مهمی که باید در طراحی مقیاس‌های مبتنی بر طیف رنگ به کار برد، اشاره گردید: اوّل اینکه رنگ‌ها باید به وضوح نشان دهند که ارزش کدام داده بزرگ‌تر یا کوچک‌تر از دیگری است. دوّم اینکه اختلاف بین رنگ‌ها باید معرّف اختلاف بین مقادیر داده‌ها باشد. متاسفانه چندین مورد از مقیاس‌های رنگی موجود -از جمله مشهورترین موارد- یک یا هر دو مورد از شروط فوق را نقض می‌کند. معروف‌ترین نوع این مقیاس‌ها، طیف رنگین کمان است (نمودار ۱۹-۴). در این روش از تمام رنگ‌های موجود در طیف رنگ‌ها استفاده می‌شود. این بدان معنی است که رنگ‌ها در این طیف به صورت چرخه‌ای هستند؛ یعنی رنگ‌های ابتدا و انتهای طیف تقریبا مشابه هم (قرمز تیره) هستند. اگر این دو رنگ ابتدا و انتهای طیف در نمودار در کنار یکدیگر قرار بگیرند نمی‌توان آن دو را به عنوان دو داده‌ای که بسیار با هم متفاوت هستند، تشخیص داد. به علاوه این طیف بسیار غیر یکنواخت است. در این طیف مناطقی وجود دارد که میزان تغییر رنگ در آن آرام و در مناطقی دیگر بسیار سریع است. این فقدان یکنواختی رنگ‌ها هنگامی که این طیف به طیف خاکستری تبدیل می‌شود، کاملا مشخّص می‌شود (نمودار ۱۹-۴). این طیف از تیرهٔ متوسط شروع و به روشن، سپس بسیار تیره رسیده و به تیرهٔ متوسط باز می‌گردد و مناطق وسیعی وجود دارد که تغییر رنگ بسیار اندک بوده و برخی مناطق باریک نیز تغییرات قابل توجهی را به همراه دارند.



نمودار ۱۹–۴. مقیاس طیف رنگین کمان به شدت غیریکنواخت است. این مساله زمانی که رنگ‌ها به طیف خاکستری تبدیل می‌شود، واضح‌تر می‌گردد. از چپ به راست، این مقیاس از تیرهٔ متوسط شروع و به روشن و سپس خیلی تیره رسیده و به تیرهٔ متوسط بازمی‌گردد. علاوه بر این تغییرات روشنایی یکدست نیست. روشن‌ترین قسمت مقیاس (متناظر با رنگ‌های زرد، سبز روشن، فیروزه‌ای) تقریبا یک سوم کل مقیاس را تشکیل می‌دهد. در حالی که تیره‌ترین قسمت‌ها (متناظر با رنگ آبی تیره) در قسمت محدودی از طیف متمرکز شده است.

در نمایش داده‌های واقعی طیف رنگین کمان می‌تواند ویژگی‌های داده‌های آماری را مبهم کند یا جنبه‌های دلخواهی از داده‌ها را برجسته نماید (نمودار ۱۹–۵). همچنین رنگ‌ها در طیف رنگین کمان بیش از حد اشباع شده هستند. نگاه کردن برای مدّت زمان طولانی به نمودار ۱۹–۵ آسان نخواهد بود.



نمودار ۱۹–۵. درصد سفیدپوستان شناسایی شده در شهرستان‌های مختلف تگزاس. طیف رنگین کمان برای نمایش داده‌های کمّی پیوسته مناسب نیست زیرا می‌تواند بر جنبه‌های دلخواه داده‌ها تاکید کند. در این جا روی شهرستان‌هایی تاکید شده است که در آن‌ها ۷۵% جمعیّت به عنوان سفیدپوست شناسایی شده‌اند. منبع داده: سرشماری سال ۲۰۱۰ در ایالات متحده

در نظر نگرفتن کوررنگی

زمانی که برای نمایش داده‌ها از رنگ بهره برده می‌شود، باید به یاد داشت که نسبت قابل توجّهی از خوانندگان ممکن است کوررنگی (یعنی نقص در شناسایی رنگ‌ها) داشته باشند. این افراد ممکن است نتوانند رنگ‌هایی را که از نظر عمده مردم کاملاً متمایز هستند را از یکدیگر تفکیک کنند. البته افرادی که این مشکل را دارند در دیدن همهٔ رنگ‌ها ناتوان نیستند. در حقیقت آن‌ها در تشخیص انواع خاصی از رنگ‌ها مشکل دارند مانند سبز و قرمز (کوررنگی سبز–قرمز) یا آبی و سبز (کوررنگی زرد–آبی). اصطلاح علمی برای کسانی که کوررنگی سبز–قرمز دارند عبارت است از دئوترآنومالی/دئوترآنوپیا[1] (برای کسانی که اختلال درک رنگ سبز دارند) و پروتانومالی/پروتانوپیا[2] (برای کسانی که اختلال درک رنگ قرمز دارند). همچنین اصلاح علمی برای کسانی که در تشخیص رنگ آبی – زرد مشکل دارند عبارت است از تریتانومالی/تریتانوپیا[3] (کسانی که اختلال درک رنگ آبی دارند). اصطلاحاتی که پسوند آنومالی[4] دارند به اختلال در درک و تشخیص رنگ مربوطه اشاره دارند و اصطلاحاتی با پسوند آنوپیا[5] به عدم قدرت تشخیص کامل آن رنگ اشاره می‌کنند. تقریبا ۸ درصد مردان و ۰/۵ درصد زنان از نوعی اختلال در تشخیص رنگ رنج می‌برند. دئوترآنومالی رایج‌ترین آن‌ها و ترتانومالی نسبتاً نادر است.

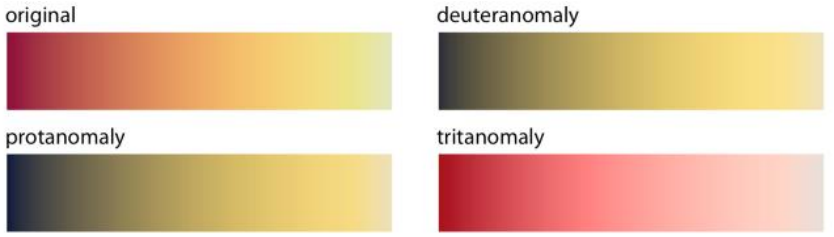
همان طور که در فصل ۴ بیان شد، سه نوع مقیاس اصلی رنگ وجود دارد که در رسم نمودار استفاده می‌شود: مقیاس متوالی، مقیاس واگرا و مقیاس کیفی. از این سه نوع مقیاس اصلی، مقیاس متوالی به طور کلی برای افرادی که کوررنگی دارند مشکلی به وجود نمی‌آورد. زیرا یک مقیاس متوالی با طراحی مناسب یک طیف پیوسته از رنگ‌های تیره تا روشن به نمایش می‌گذارد. نمودار ۱۹–۶ مقیاس حرارت که در نمودار ۴–۳ ارائه شد را در انواع مختلف کوررنگی شبیه‌سازی کرده است. گرچه هیچکدام از موارد شبیه‌سازی شده مشابه مقیاس اصلی نیست، با این حال همه یک شیب واضح از تاریک به روشن را نمایش داده و لذا عملکرد قابل قبولی برای نمایش مفهوم اندازهٔ داده‌ها دارند.

در مورد مقیاس واگرا، اوضاع کمی پیچیده‌تر است زیرا تضاد رنگ‌های معروف برای افراد دچار کوررنگی می‌تواند غیر قابل تشخیص باشد. به طور خاص رنگ‌های قرمز و سبز تقریبا

1. deuteranomaly/deuteranopia
2. protanomaly/protanopia
3. tritanomaly/tritanopia
4. anomaly
5. anopia

بیشترین تضاد را برای افرادی که دید رنگ طبیعی دارند، به همراه خواهند داشت امّا تشخیص این دو رنگ برای افراد با اختلال دید رنگ سبز و قرمز تقریبا ناممکن است (نمودار ۱۹–۷). به طور مشابه تضاد سبز– آبی برای افراد با اختلال دید سبز و اختلال دید قرمز قابل مشاهده و تشخیص است امّا برای افراد با اختلال دید آبی غیر قابل تشخیص است (نمودار ۱۹–۸).

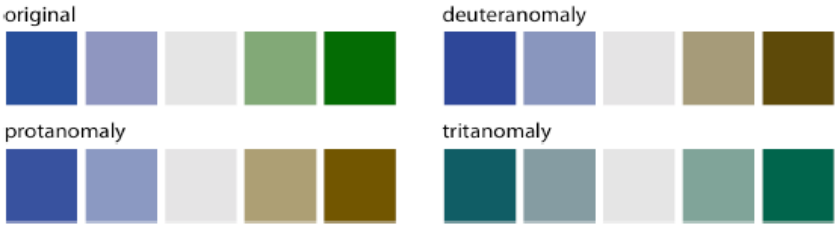
این مثال‌ها بیان می‌کنند که پیدا کردن دو رنگ متضاد که برای تمام انواع کوررنگی مناسب باشد، تقریبا ناممکن است. با این حال شرایط تا این اندازه نیز وخیم نیست. معمولاً این امکان وجود دارد که تغییرات اندکی در رنگ‌ها به وجود آورد به نحوی که هم ویژگی مورد نظر را داشته باشند و هم برای افرادی با اختلال دید رنگ مناسب باشند. مثلاً طیف PiYG[1] (از صورتی تا زرد متمایل به سبز) در نمودار ۴–۵ برای افراد با دید طبیعی رنگ به صورت قرمز–سبز به نظر می‌آید، با این حال برای افراد دچار کوررنگی نیز همچنان قابل تمایز است (نمودار ۱۹–۹).



نمودار ۱۹–۶. شبیه‌سازی کوررنگی در مقیاس متوالی حرارتی از قرمز تیره تا زرد روشن. از چپ به راست و بالا به پایین مقیاس اصلی و مقیاس‌های شبیه‌سازی شده تحت سه نوع کوررنگی نمایش داده شده است. گرچه در هر سه نوع کوررنگی، رنگ‌های خاص متفاوت دیده می‌شوند، امّا در هر مورد می‌توان یک شیب واضح از تیره به روشن دید. بنابراین این طیف رنگی برای خوانندگان با کوررنگی نیز مناسب است.

نمودار ۱۹–۷. تضاد سبز–قرمز برای افراد اختلال دید رنگ قرمز–سبز غیر قابل تشخیص است.

1. pink to yellow-green



نمودار ۱۹–۸. تضاد سبز–آبی برای افراد اختلال دید رنگ آبی–زرد غیر قابل تشخیص است.

نمودار ۱۹–۹. مقیاس رنگ PiYG (صورتی تا سبز - زرد) از نمودار ۴–۵ برای افرادی که اختلال دید رنگ ندارند، به شکل تضاد سبز– قرمز دیده می‌شود امّا برای افراد با هر نوع اختلال دید رنگی نیز همچنان کارآمد است. این طیف به این دلیل کارآمد است که رنگ مایل به قرمز آن در اصل صورتی (ترکیبی از قرمز و آبی) است در حالی که رنگ مایل به سبز آن در خود زرد نیز دارد. اختلاف در جزء آبی رنگ در دو رنگ مذکور برای افراد با اختلال درک سبز و قرمز قابل تمایز بوده و از سوی دیگر اختلاف در جز قرمز رنگ برای افراد با اختلال درک آبی نیز قابل تمایز می‌باشد.

بیش‌ترین پیچیدگی مربوط به مقیاس کیفی است زیرا در آن به رنگ‌های متفاوت زیادی نیاز بوده و همه آن‌ها بایستی برای افراد با انواع مختلف کوررنگی قابل تمایز باشند. مقیاس کیفی که به وفور در قسمت‌های مختلف این کتاب به کار رفته، به طور خاص برای غلبه بر این مشکل انتخاب شده است (نمودار ۱۹–۱۰). با کنار هم قرار دادن هشت رنگ مختلف، این مجموعهٔ رنگ تقریبا برای هر سناریویی کارآمد است. همانطور که در ابتدای این فصل گفته شد نباید بیش از ۸ رنگ متفاوت برای نمایش داده‌های آماری در نمودار استفاده نمود.

نمودار ۱۹–۱۰. مجموعه رنگ کیفی برای پوشش تمام انواع کوررنگی. کدهای حروفی–عددی نمایانگر رنگ‌ها در چارچوب RGB [1] هستند که به صورت شش‌تایی کدگذاری شده‌اند. در بسیاری از مجموعه‌های رنگ و نرم‌افزارهای ترسیم نمودار می‌توان مستقیماً کدهای شش تایی را وارد نمود. اگر نرم‌افزار مستقیماً کدهای شش تایی را قبول نمی‌کند می‌توان از مقادیر جدول ۱۹–۱ استفاده کرد.

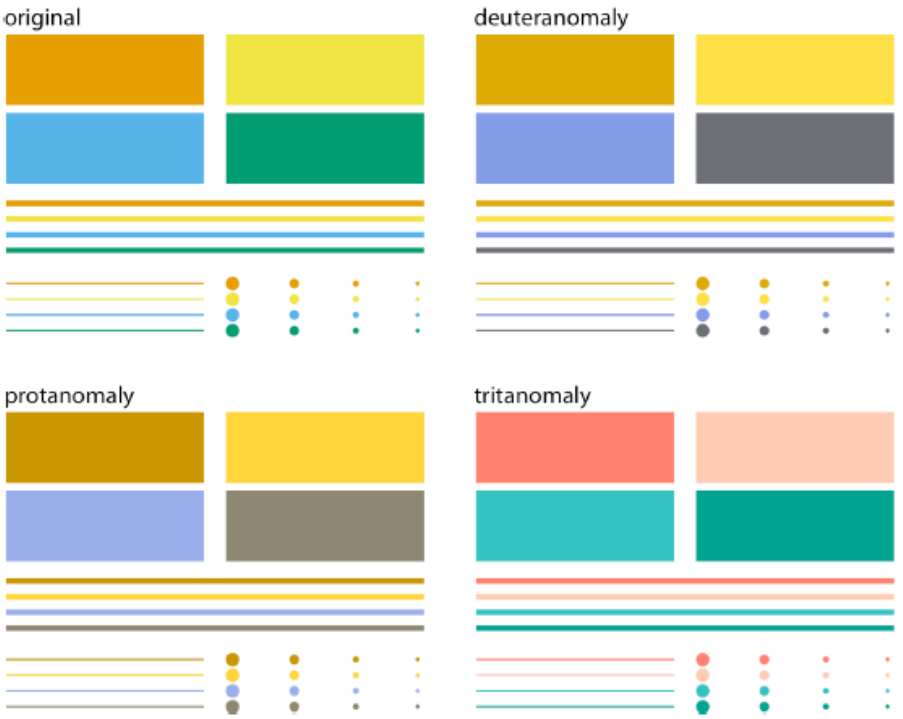
1. Red Green Blue

جدول ۱۹–۱. مقیاس رنگی مناسب برای افراد کوررنگ

نام	کد Hex	Hue	C, M, Y, K (%)	R, G, B (0-255)	R, G, B (%)
نارنجی	#E69F00	41	0,50,100,0	230,159,0	90,60,0
آبی آسمانی	#56B4E9	202	80,0,0,0	86,180,233	35,70,90,
سبز آبی	#009E73	164	97,0,75,0	0,158,115	0,60,50,
زرد	#F0E442	56	10,5,90,0	240,228,66	95,90,25
آبی	#0072B2	202	100,50,0,0	0,114,178	0,45,70,
سرخابی	#D55E00	27	0,80,100,0	213,94,0	80,40,70
بنفش مایل به قرمز	#CC79A7	326	10,70,0,0	204,121,167	80,60,70
سیاه	#000000	N/A	0,0,0,100	0,0,0	0,0,0

با این که مقیاس‌های رنگی متعددی که مناسب افراد کوررنگ باشد در دسترس است، باید توجّه داشت که آن‌ها جادویی نیستند. ممکن است از یک مقیاس رنگی مناسب برای افراد کوررنگ استفاده شود و با این حال فردی با اختلال دید رنگ نتواند آن را درک کند. یکی از پارامترهای اساسی اندازهٔ عناصر گرافیکی رنگ شده است. وقتی رنگ‌ها برای مساحت‌های بزرگ به کار می‌رود نسبت به حالتی که برای مساحت‌های کوچک یا خطوط نازک استفاده می‌شود، بهتر قابل تمایز می‌باشند و این اثر برای افراد با اختلال دید رنگ تشدید می‌شود (نمودار ۱۹–۱۱). علاوه بر ملاحظات مختلفی که در طراحی و استفاده از رنگ‌ها باید مدنظر قرار گیرد و در این فصل و فصل ۴ در موردشان بحث شد، پیشنهاد می‌شود نمودارهای رنگی بر اساس اختلال دید رنگ نیز شبیه‌سازی شده تا بینشی در خصوص نحوه درک آن‌ها توسّط افراد کوررنگ به دست آید. سرویس‌های برخط و نیز نرم‌افزارهای مختلفی امکان شبیه‌سازی نمودارها بر اساس انواع مختلف کوررنگی را فراهم می‌کنند.

برای آنکه مطمئن شوید نمودارهایتان برای افراد مبتلا به اختلال دید رنگ کارآمد است، صرفاً به مقیاس‌های رنگی خاصی اکتفا نکنید. در عوض نمودارتان را در یک شبیه‌ساز اختلال دید رنگ امتحان کنید.



نمودار ۱۹–۱۱. عناصر رنگ‌آمیزی شده در ابعاد کوچک به سختی قابل تمایز هستند. مجموعهٔ سمت بالا چپ که اصلی نام دارد چهار مربع، چهار خط ضخیم، چهار خط نازک و چهار نقطه را نشان می‌دهد که همگی چهار رنگ مشابه دارند. می‌توان دید که هر چه رنگ‌ها در ابعاد کوچک‌تر یا نازک‌تر باشند تمایز آن‌ها سخت‌تر می‌شود. این مشکل در شبیه‌سازی‌های اختلال دید رنگ وخیم‌تر می‌شود زیرا در این حالت تشخیص رنگ‌ها حتی در عناصر با ابعاد بزرگ نیز دشوار است.

فصل ۲۰

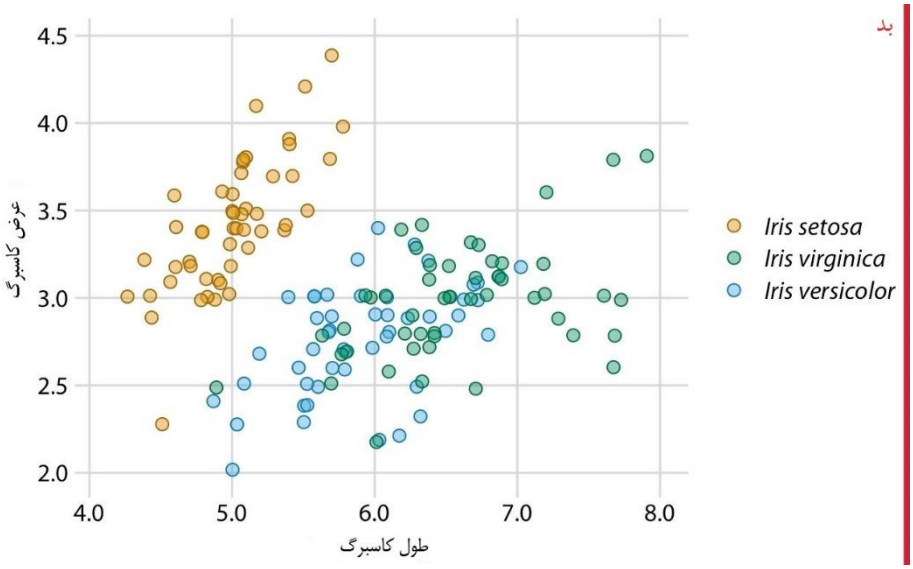
کدگذاری اضافی

در فصل ۱۹ بیان شد که رنگ‌ها به تنهایی نمی‌توانند اطّلاعات را آن طور که انتظار داریم انتقال دهند. نمایش اطّلاعات متفاوت و متعدد ممکن است به وسیله رنگ امکان‌پذیر نباشد. انطباق رنگ‌ها در نمودار با رنگ‌های راهنمای نمودار دشوار خواهد بود (نمودار ۱۹–۱). حتی در مواردی که بایستی فقط دو یا سه مورد متفاوت از هم تمایز داده شوند، چنانچه موارد مذکور اندازهٔ خیلی کوچک داشته باشند، کدگذاری رنگی موثر نخواهد بود (نمودار ۱۹–۱۱). همچنین افرادی که کوررنگی دارند، رنگ‌های نمودار را مشابه هم خواهند دید (نمودار ۱۹–۷ و ۱۹–۸). راه حل کلی برای تمام این حالات استفاده از رنگ‌ها جهت افزایش تاثیر بصری نمودارها می‌باشد، البته بدون اینکه صرفاً به رنگ‌ها برای انتقال اطّلاعات تکیه شود. به این اصول برای طراحی نمودارها اصطلاحاً کدگذاری اضافی اطلاق می‌شود، زیرا فرد را مکرراً به کدگذاری اطّلاعات با استفاده از ابعاد متفاوت زیبایی‌شناسی تشویق می‌کند.

طراحی راهنما با استفاده از کدگذاری اضافی

طراحی نمودارهای پراکنش به صورتی است که نقاط نماینده گروه‌های مختلفی از داده‌ها فقط در رنگ با یکدیگر تفاوت دارند. به عنوان مثال نمودار ۲۰–۱ را در نظر بگیرید که عرض و طول کاسبرگ را در سه گونهٔ مختلف زنبق با هم مقایسه کرده است (کاسبرگ، برگ‌های خارجی گل در گیاهان گل‌دهنده هستند). نقاطی که نشان‌دهندهٔ گونه‌های مختلف زنبق هستند

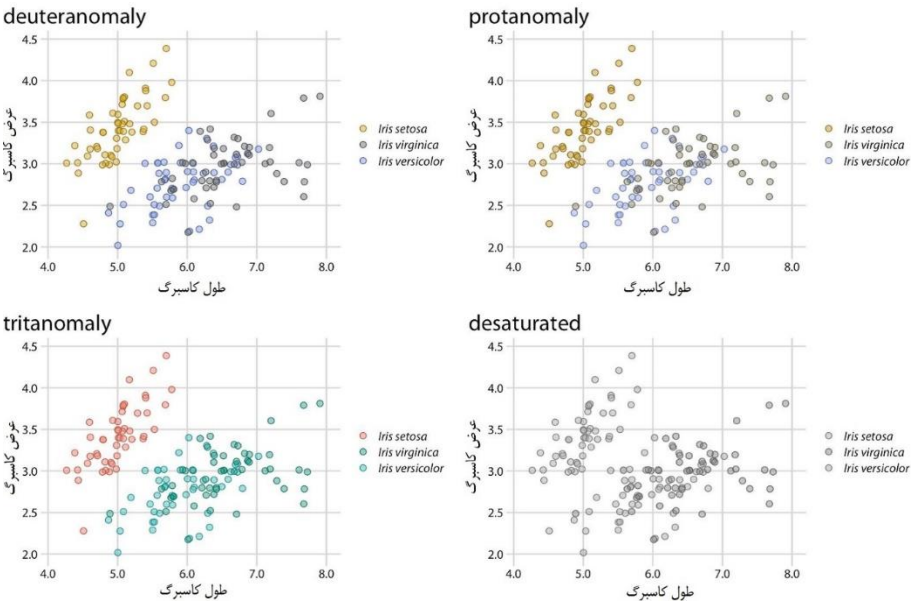
با رنگ‌های مختلف نمایش داده شده‌اند. در غیر این صورت تمام نقاط مشابه یکدیگر بودند. با وجود اینکه این نمودار فقط سه گروه مختلف از نقاط را دربردارد، خواندن آن حتّی برای افرادی که بینایی طبیعی دارند نیز مشکل است. این مشکل ناشی از این است که دو گونهٔ زنبق ویرجینیکا و زنبق ورسیکالر در برخی نقاط با یکدیگر همپوشانی دارند و رنگ‌های مربوط به هر کدام از آن‌ها، یعنی سبز و آبی، به خوبی قابل تشخیص از یکدیگر نیستند.



نمودار ۲۰–۱. مقایسهٔ عرض و طول کاسبرگ در سه گونهٔ مختلف زنبق. (زنبق ستوسا، زنبق ورجینیکا، زنبق ورسیکالر) هر نقطه نماینده اندازه‌گیری‌ها برای یک نمونه گیاه است. به این نمودار برچسب «بد» اختصاص داده شده زیرا نقاط سبز رنگ و آبی رنگ به سختی از یکدیگر قابل افتراق هستند. منبع داده: Fisher 1936

به طور شگفت‌انگیزی افرادی که کور رنگی قرمز–سبز دارند (دئوترانومالی[1] یا پروتوانومالی[2]) در مقایسه با افراد طبیعی می‌توانند نقاط سبز و آبی را به خوبی از هم تشخیص دهند (ردیف بالای نمودار ۲۰–۲ و نمودار ۲۰–۱ را مقایسه کنید). از طرف دیگر برای افرادی که کور رنگی آبی–زرد دارند (تریتانومالی[3]) نقاط آبی و سبز بسیار مشابه به نظر می‌رسند (نمودار ۲۰–۲ سمت پایین–چپ). چنانچه نمودار در مقیاس خاکستری (سیاه - سفید) چاپ شود (یعنی نمودار بدون رنگ شود)، هیچکدام از گونه‌های زنبق قابل افتراق نخواهد بود (نمودار ۲۰–۲ سمت پایین–راست).

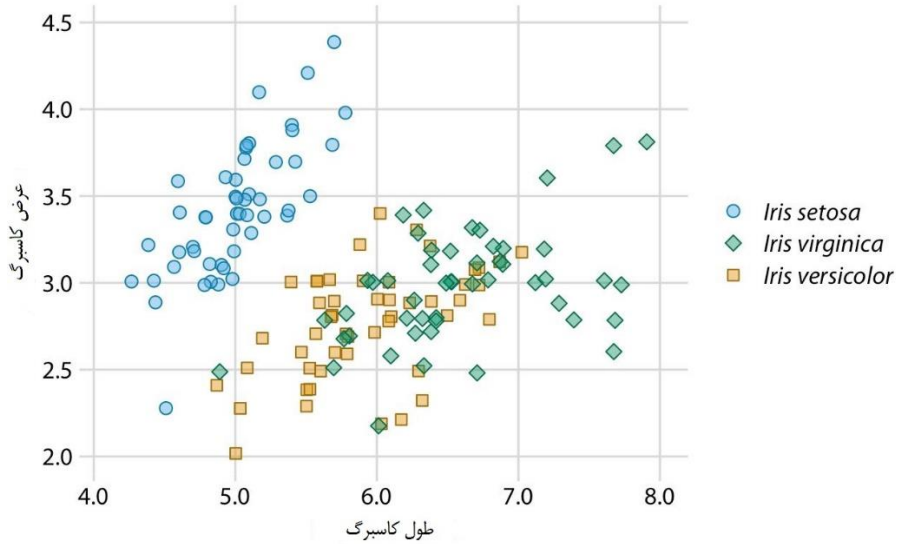
1. deuteranomaly
2. protanomaly
3. tritanomaly



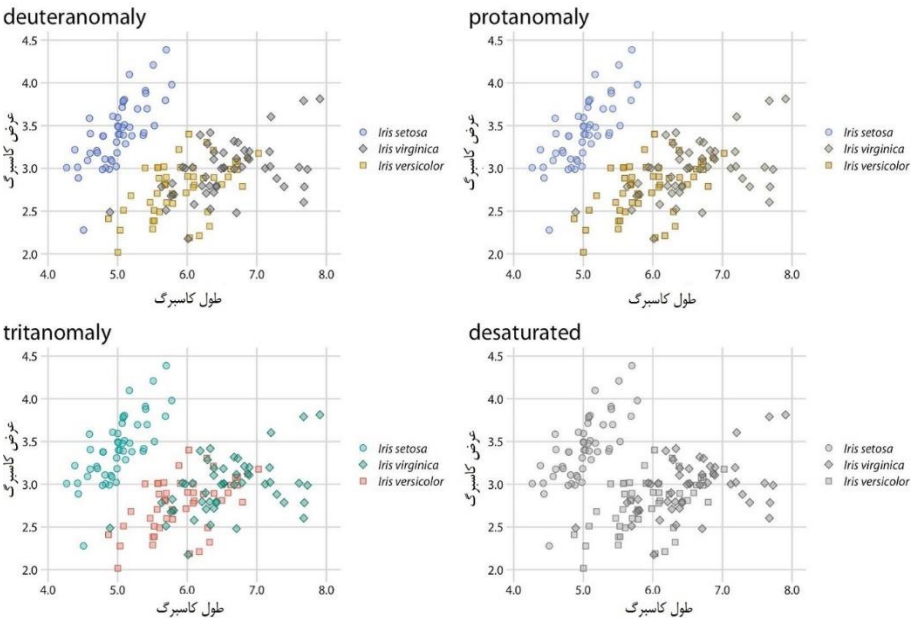
نمودار ۲۰–۲. شبیه‌سازی نمودار ۲۰–۱ برای افرادی که کوررنگی دارند. منبع داده: Fisher 1936

دو روش ساده برای بهبود نمودار ۲۰–۱ وجود دارد تا این مشکلات را کاهش دهد. اوّلین روش این است که رنگ‌های استفاده شده برای دو گونهٔ زنبق ستوسا و زنبق ورسیکالر جا به جا شود. به این ترتیب رنگ آبی دیگر در کنار رنگ سبز نخواهد بود (نمودار ۲۰–۳). در روش دوّم می‌توان از سه شکل متفاوت برای نمایش نقاط گونه‌های مختلف زنبق استفاده کرد. با این دو تغییر، هم نمودار اصلی (نمودار ۲۰–۳) و هم نسخهٔ مربوط به افراد کور رنگ و نیز نسخهٔ سیاه و سفید (نمودار ۲۰–۴) خوانا خواهند بود.

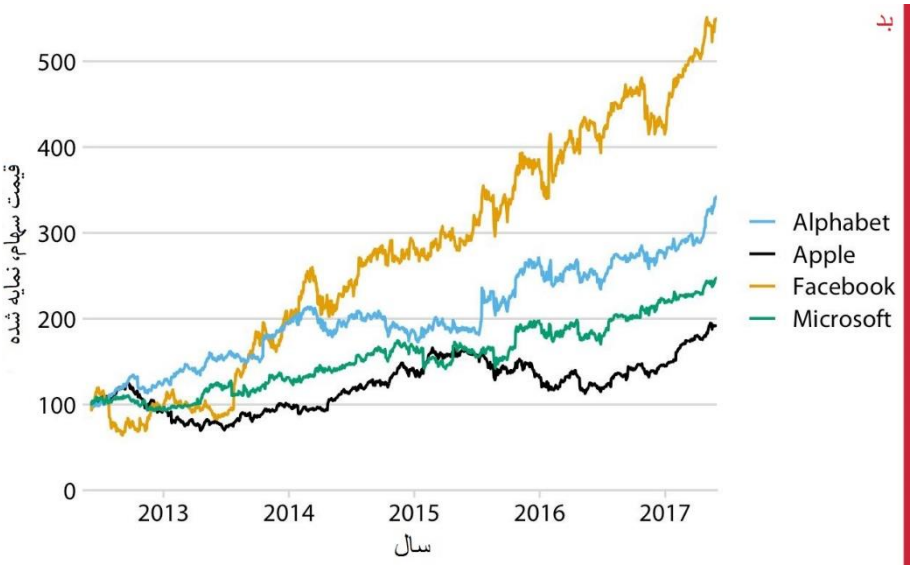
تغییر شکل نقاط، یک راهبرد ساده برای نمودارهای پراکنش است، امّا الزاماً به معنی کاربردی بودن برای سایر نمودارها نیست. در نمودارهای خطی، می‌توان نوع خط را عوض نمود (خطوط توپر، خط‌چین، خطوط نقطه‌ای و ... نمودار ۲–۱ را نیز ببینید)، امّا استفاده از خط تیره یا خطوط نقطه‌ای منجر به نتایج کمتر از حد مطلوب خواهد شد. مخصوصاً خط‌چین یا خطوط نقطه‌ای معمولاً مناسب به نظر نمی‌رسند مگر اینکه کاملاً مستقیم بوده یا به آرامی دچار انحنا شوند و در هر دو صورت باعث اختلال بصری می‌شوند. همچنین در بسیاری از موارد نیازمند تلاش ذهنی زیادی است تا فرد بتواند الگوهای متفاوت نقطه‌ای یا نقطه–خط را در نمودار و راهنمای آن مطابقت دهد. در این صورت با نمودارهایی مانند نمودار ۲۰–۵ که از خطوط برای نمایش تغییرات ارزش بورس در طول زمان برای چهار شرکت بزرگ فناوری استفاده کرده است چه کار باید کرد؟



نمودار ۲۰–۳. مقایسهٔ عرض و طول کاسبرگ در سه گونهٔ متفاوت زنبق. در مقایسه با نمودار ۲۰–۱ رنگ گونهٔ زنبق ستوسا و زنبق ورسیکالر باهم عوض شده‌اند. همچنین به هر گونه زنبق شکل مخصوصی تخصیص داده شده است. منبع داده: Fisher 1936



نمودار ۲۰–۴. شبیه‌سازی نمودار ۲۰–۳ برای افراد کور رنگ. به علّت استفاده از شکل‌های مختلف برای نمایش نقـاط هر کـدام از گونه‌ها، حتّی نسخه حاصل در مقیاس خاکستری هم کاملاً واضح و خواناست. منـبع داده: Fisher 1936

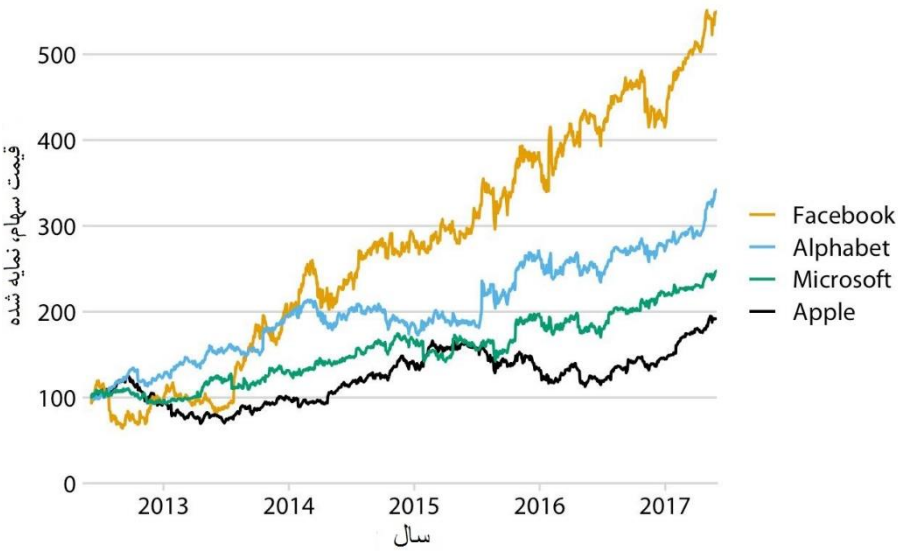


نمودار ۲۰–۵. ارزش سهام چهار شرکت فناوری در طول زمان. ارزش سهام هر شرکت در سال ۲۰۱۲ استاندارد شده و برابر ۱۰۰ در نظر گرفته شده است. به این نمودار برچسب «بد» اختصاص داده شده زیرا باید انرژی ذهنی نسبتاً زیادی برای انطباق نام شرکت‌ها در راهنمای نمودار با خود نمودار صرف کرد. منبع داده: امور مالی شرکت یاهو

نمودار شامل چهار خط است که هر خط نمایندهٔ ارزش سهام برای یک شرکت می‌باشد. هر خط با یک رنگ کدگذاری شده است و رنگ‌ها به صورتی انتخاب شده‌اند تا برای افرادی که کوررنگی دارند نیز مناسب باشد. بنابراین انطباق نام هر شرکت و خط مرتبط با نام آن شرکت باید نسبتاً ساده باشد امّا در عمل اینگونه نیست. مشکل اینجاست که در این گونه نمودارها خطوط ترتیب بصری دارند. رنگ زرد نماینده شرکت فیسبوک، به عنوان بالاترین خط در نمودار می‌باشد. همچنین خط مشکی که مربوط به شرکت اپل می‌باشد، به عنوان پایین‌ترین خط در نمودار می‌باشد. شرکت‌های آلفابت و مایکروسافت نیز بین این دو شرکت به ترتیب ذکر شده قرار دارند. امّا ترتیب این چهار شرکت در راهنما بدین صورت است: آلفابت، اپل، فیسبوک، مایکروسافت (به ترتیب حروف الفبا). بنابراین ترتیب مشاهده شده در نمودار خطی با ترتیب نام شرکت‌ها در راهنمای نمودار متفاوت خواهد بود و تلاش ذهنی زیادی می‌خواهد تا خطوط نمودار با نام شرکت‌ها در راهنما تطبیق داده شود.

این مشکل معمولاً به این علّت است که نرم‌افزار ترسیم نمودار به صورت خودکار ترتیبی را برای راهنما در نظر می‌گیرد. نرم‌افزارها هیچگونه درکی از ترتیب بصری که نمودار در بیننده

القا می‌کند، ندارند. در واقع نرم‌افزار ترتیب چینش راهنما را عمدتاً بر اساس حروف الفبا تنظیم می‌کند. این مشکل به صورت دستی با تغییر چینش راهنما به گونه‌ای که هماهنگ با ترتیب خطوط نمودار باشد، قابل حل است (نمودار ۲۰–۶). نتیجه حاصل نموداری است به همراه راهنمایی که خیلی راحت‌تر قابل درک و انطباق خواهد بود.

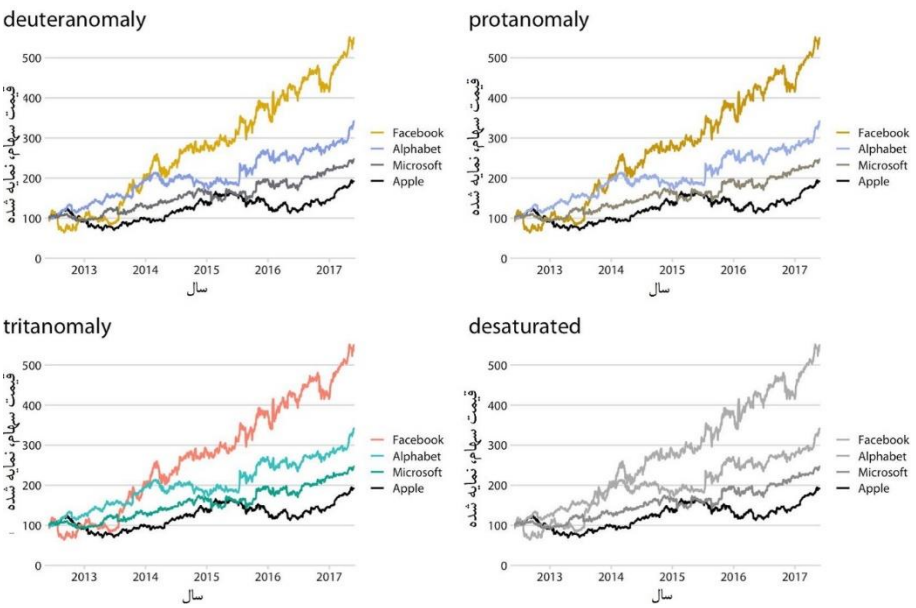


نمودار ۲۰–۶. ارزش سهام چهار شرکت فناوری در طول زمان. در مقایسه با نمودار ۲۰–۵ ترتیب چینش راهنما به صورتی می‌باشد که هماهنگ با ترتیب بصری خطوط نمودار است یعنی فیسبوک بالاترین و اپل پایین‌ترین. منبع داده: امور مالی شرکت یاهو

اگر ترتیب بصری در داده‌ها وجود دارد، اطمینان حاصل کنید که ترتیب چینش راهنمای نمودار با آن تطابق داشته باشد.

تطابق ترتیب راهنمای نمودار با اطّلاعات نمودار همیشه کمک‌کننده است امّا مزایای آن به طور ویژه در نمودارهایی که برای کوررنگی شبیه‌سازی شده مشخّص است (نمودار ۲۰–۷). به عنوان مثال در نسخه تریاتومالی که تشخیص رنگ‌های سبز و آبی از یکدیگر مشکل هستند، کمک‌کننده است (نمودار ۲۰–۷ پایین–چپ). همچنین در نسخهٔ مقیاس خاکستری نیز مفید است (نمودار ۲۰–۷ پایین–راست). با وجود اینکه دو رنگ مربوط به فیسبوک و آلفابت، از

لحاظ بصری تقریباً در یک مقیاس خاکستری هستند، با این وجود به راحتی می‌توان فهمید که رنگ‌های مرتبط با دو شرکت مایکروسافت و اپل تیره‌تر هستند و دو خط آخر را شامل می‌شوند. بنابراین به راحتی می‌توان متوجّه شد که بالاترین خط نمودار مربوط به شرکت فیسبوک و خط پایین آن مربوط به شرکت آلفابت می‌باشد.

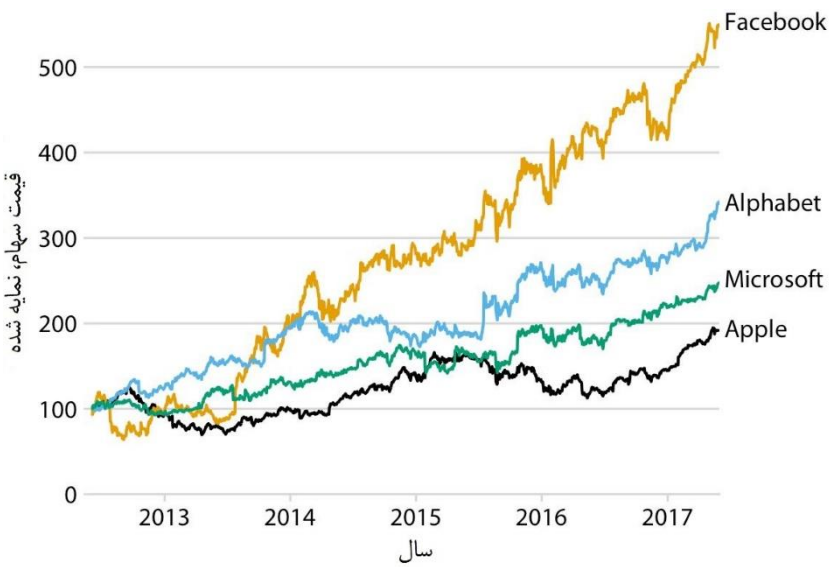


نمودار ۷–۲۰. شبیه سازی کوررنگی نمودار ۲۰–۶. منبع داده: امور مالی شرکت یاهو

طراحی نمودارهای بدون راهنما

با وجود اینکه خواندن راهنماها به وسیله کدگذاری اضافی می‌تواند بهبود پیدا کند، از دیدگاه زیبایی‌شناسی، راهنماها همیشه بار ذهنی اضافی به مخاطب تحمیل می‌کنند. برای خواندن یک راهنما، خواننده باید اطّلاعات را از یک قسمت تصویر برداشته و آن را به قسمت دیگری از همان تصویر منتقل کند. اگر راهنما را به طور کل از نمودار حذف کنیم کمک زیادی به خواننده کرده‌ایم تا بتواند راحت‌تر آن نمودار را تحلیل کند. حذف راهنما بدین معنی نیست که هیچ راهنمایی ارائه نخواهد شد و یا مثلاً در زیرنویس نمودار جملاتی مانند «نقاط زرد رنگ نشان‌دهندهٔ زنبق ورسیکالر می‌باشند» نوشته نمی‌شود. حذف راهنما در واقع به این معنی است که نمودار به روشی طراحی شود که به سرعت بتوان فهمید که هر عنصر گرافیکی نمایندهٔ چه داده‌ای است، حتی اگر هیچ راهنمای صریحی وجود نداشته باشد.

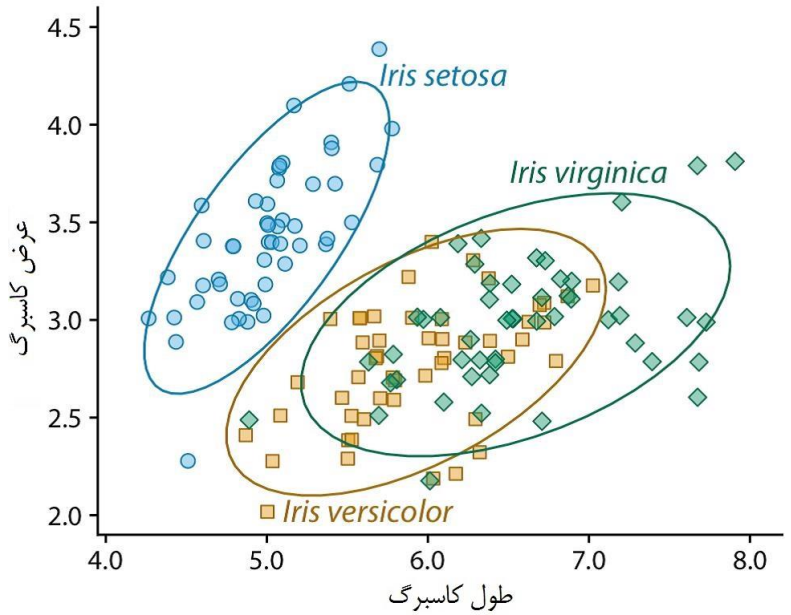
راهبرد کلی برای این موضوع برچسب‌گذاری مستقیم نام دارد که به موجب آن از برچسب‌های متنی مناسب و یا سایر عناصر بصری که مانند یک راهنما هستند در نمودار استفاده می‌شود. قبلاً در فصل ۱۹ (نمودار ۱۹–۲) با برچسب‌گذاری مستقیم به عنوان جایگزینی برای استفاده از راهنمایی با بیش از ۵۰ رنگ متفاوت صحبت شد. برای اعمال برچسب‌گذاری مستقیم برای نمودار ارزش سهام، نام هر شرکت در انتهای خط مرتبط با داده آن قرار داده شد (نمودار ۲۰–۸).



نمودار ۲۰–۸. ارزش سهام در طول زمان برای چهار شرکت اصلی فناوری. ارزش سهام هر شرکت در سال ۲۰۱۲ استاندارد شده و برابر ۱۰۰ در نظر گرفته شده است. منبع داده: امور مالی شرکت یاهو

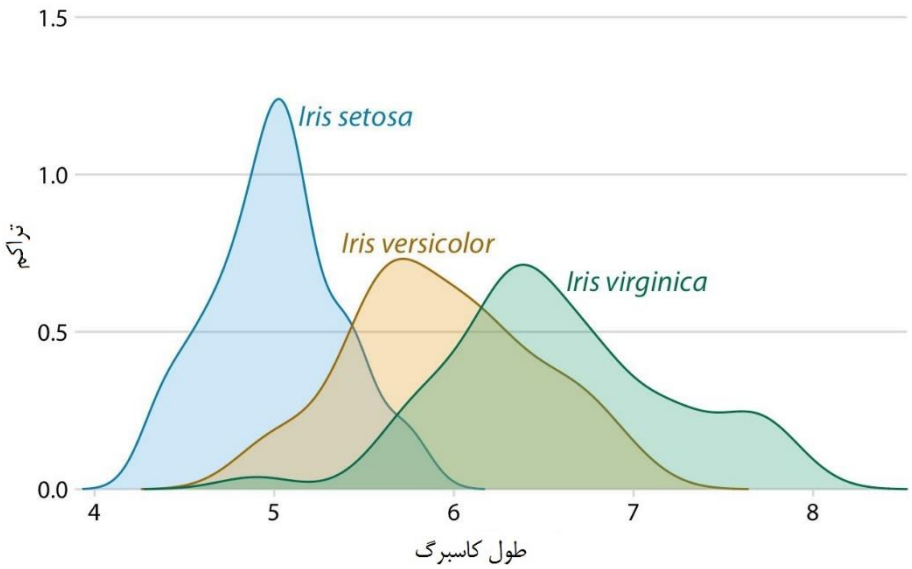
در صورت امکان، نمودارها را طوری طراحی کنید که به راهنمای جداگانه‌ای نیاز نداشته باشند.

برچسب‌گذاری مستقیم را برای نمودار مربوط به گیاه زنبق که در ابتدای فصل آمده بود نیز می‌توان اعمال کرد (نمودار ۲۰–۳). از آنجایی که این نمودار پراکنش شامل تعداد زیادی نقطه است که به سه گروه مختلف تقسیم می‌شوند، بایستی برچسب‌گذاری برای گروه نقاط داده انجام شود و نه نقاط منفرد. یک راه حل، ترسیم بیضی‌هایی است که اکثریت نقاط را در بر بگیرد و سپس هر کدام از این بیضی‌ها برچسب‌گذاری شود (نمودار ۲۰–۹).



نمودار ۲۰–۹. عرض کاسبرگ در مقایسه با طول آن برای سه گونهٔ متفاوت از گیاه زنبق. نقاط معرّف گونه‌های متفاوت این گیاه توسّط بیضی و برچسب‌های متنی رنگی مشخّص شده‌اند. در مقایسه با نمودار ۲۰–۳ خطوط پس زمینه نمودار برای جلوگیری از شلوغی تصویر حذف شده است. منبع داده: Fisher 1936

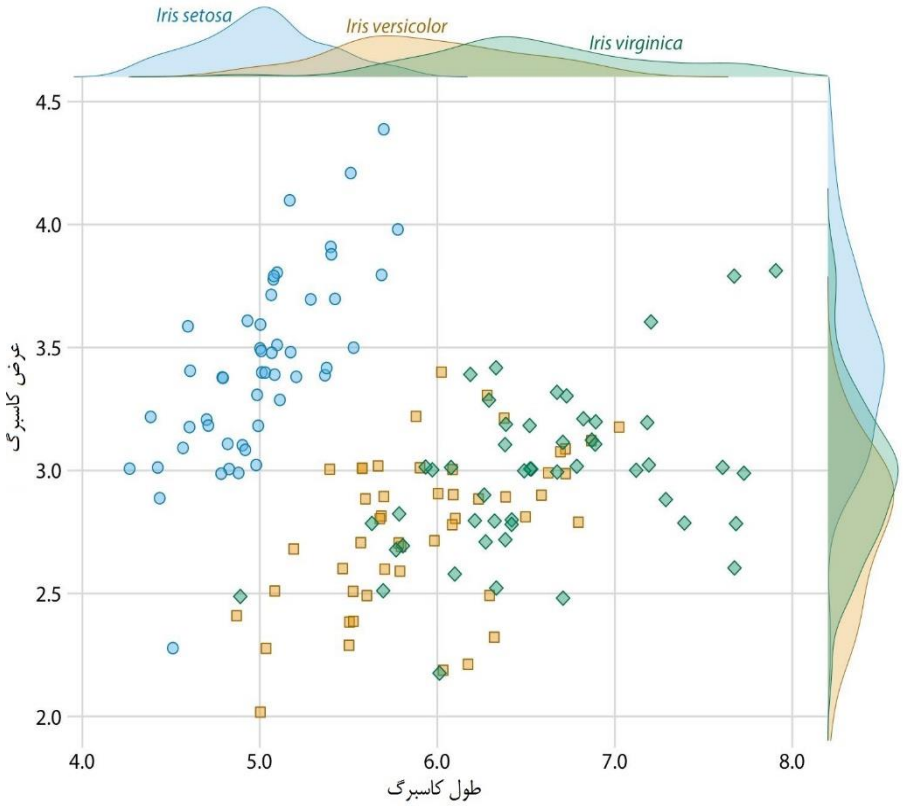
برای نمودارهای تراکمی، نیز می‌توان به جای استفاده از راهنمایی با کدگذاری رنگی از برچسب‌گذاری مستقیم منحنی‌ها استفاده کرد (نمودار ۲۰–۱۰). در هر دو نمودار ۲۰–۹ و ۲۰–۱۰ متن برچسب‌ها همرنگ داده‌ها می‌باشد. برچسب‌های رنگی می‌توانند به میزان زیادی تاثیر برچسب‌گذاری مستقیم را افزایش دهند، امّا ممکن است منجر به تضعیف نمودار نیز بشوند. اگر این برچسب‌های رنگی به هنگام چاپ کردن خیلی روشن باشند، آنگاه خواندن آن‌ها مشکل خواهد بود. همچنین به علّت اینکه این متن‌ها از خطوط نازکی تشکیل شده‌اند، این متون نسبت به ناحیهٔ اطراف آن که همان رنگ را دارد، اغلب روشن‌تر به نظر خواهند رسید. این مشکل را می‌توان با استفاده از دو سایهٔ متفاوت از یک رنگ برطرف کرد، سایه رنگی روشن‌تر برای مناطق پر شده و سایهٔ تاریک‌تر برای خطوط، اطراف کادر و متن. اگر با دقّت نمودار ۲۰–۹ و ۲۰–۱۰ را بررسی نمایید، خواهید دید که هر نقطهٔ داده یا مناطق سایه‌دار با یک رنگ روشن پر شده است و برای خطوط اطراف آن از همان رنگ امّا تیره‌تر استفاده شده است.



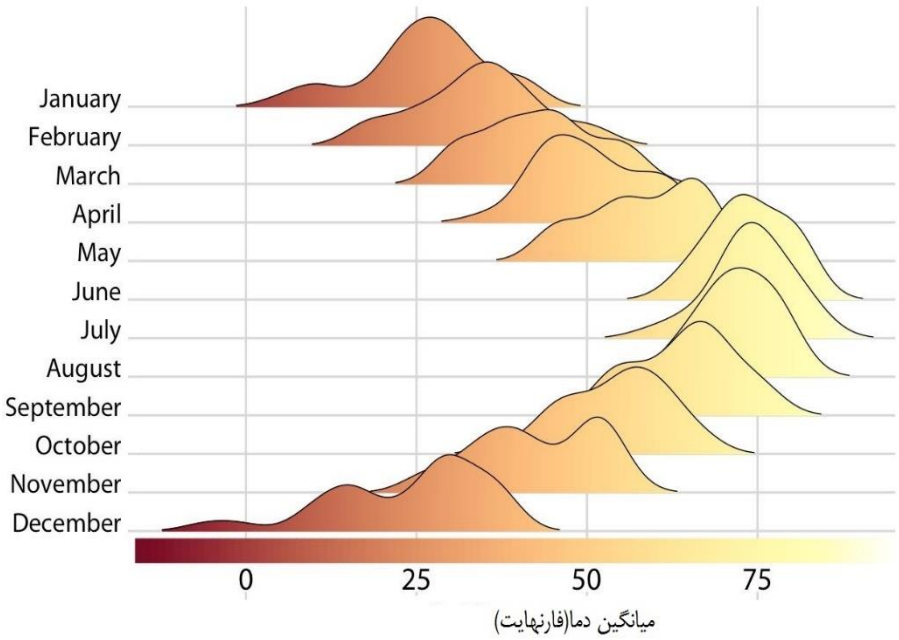
نمودار ۲۰–۱۰. تخمین تراکمی طول کاسبرگ سه گونهٔ متفاوت زتبق. هر تخمین تراکمی مستقیماً با نام گونه مرتبط با خودش برچسب‌گذاری شده است. منبع داده: Fisher 1936

همچنین می‌توان از نمودارهای تراکمی مانند آنچه که در نمودار ۲۰–۱۰ ارائه شد، به عنوان جایگزینی برای راهنما استفاده نمود. به این صورت که نمودارهای تراکمی در لبهٔ نمودار پراکنش قرار گیرد (نمودار ۲۰–۱۱). این کار این امکان را فراهم می‌کند که به جای نمودار پراکنش مرکزی، نمودارهای تراکمی حاشیه‌ای مستقیماً برچسب‌گذاری شود و از این رو نتیجه حاصل نموداری خواهد بود که بهم ریختگی کمتری نسبت به نمودار ۲۰–۹ خواهد داشت.

در نهایت، هرگاه یک متغیّر واحد را با استفاده از چندین شیوهٔ زیبایی‌شناسی کدگذاری کنیم، معمولاً تمایلی برای وجود یک راهنما برای هرکدام از آن‌ها وجود ندارد. بلکه باید فقط یک عنصر بصری به عنوان راهنما وجود داشته باشد، که تمامی اطّلاعات نمودار را به بیننده انتقال دهد. در این صورت زمانی که همان متغیّر در امتداد یک محور اصلی و به صورت رنگی رسم می‌شود، این در واقع بدین معناست که نوار مرجع رنگی باید در امتداد همان محور حرکت نموده و در واقع با آن ادغام شود. نمودار ۲۰–۱۲ موردی را نشان می‌دهد که میزان دما در امتداد محور افقی و به صورت رنگی رسم شده و در نتیجه راهنمای رنگ‌ها نیز با محور افقی ادغام شده است.



نمودار ۲۰–۱۱. مقایسهٔ عرض با طول کاسبرگ برای چهار گونهٔ مختلف زنبق به همراه تخمین تراکمی در حاشیه نمودار برای هر متغیّر و هر کدام از گونه ها. منبع داده: Fisher 1936



نمودار ۲۰–۱۲. دما در لینکین، نبراسکا در سال ۲۰۱۶. این تصویر حالت دیگری از نمودار ۹–۹ می‌باشد. دما به طور همزمان هم در راستای محور افقی و هم به صورت رنگی نمایش داده شده است. نوار رنگی در امتداد محور افقی مقیاسی را به تصویر می‌کشد که میزان دما را بر حسب شدّت رنگ نشان می‌دهد. منبع داده: سازمان هواشناسی

فصل ۲۱

اشکال چند پانلی

وقتی مجموعه داده‌ها بزرگ و پیچیده می‌شوند، اغلب حاوی اطّلاعات بسیار بیشتر از آن هستند که بتوان به صورت منطقی در قالب یک نمودار نشان داد. برای نمایش چنین مجموعه داده‌ای، ایجاد نمودار چند پانلی می‌تواند مفید باشد. این‌ها نمودارهایی هستند که از چندین پانل تشکیل شده‌اند که هر کدام زیرمجموعه‌ای از داده‌ها را نشان می‌دهد. دو دستهٔ مجزا از این نمودارها وجود دارد، چندگانه‌های کوچک[1] و نمودارهای مرکّب[2]. چندگانه‌های کوچک نمودارهایی هستند که از چند پانل تشکیل شده‌اند که در یک شبکه منظّم چیده شده‌اند. هر پانل زیرمجموعهٔ متفاوتی از داده‌ها را نشان می‌دهد امّا همه پانل‌ها از یک نوع نمودار استفاده می‌کنند. نمودارهای مرکّب ترکیبی شامل پانل‌های جداگانه‌ای هستند که در یک آرایش دلخواه (که ممکن است مبتنی بر شبکه باشند یا نباشند) چیده شده‌اند و نمودارهای کاملاً متفاوت یا حتی مجموعه‌های داده متفاوت را نشان می‌دهند.

در بسیاری از قسمت‌های این کتاب با هر دو نوع نمودار چند پانلی برخورد داشته‌ایم. به طور کلی، این نمودارها کاملا شهودی بوده و تفسیر آن‌ها ساده است. با این حال، هنگام تهیه چنین نمودارهایی، باید به چند موضوع توجّه کنیم، مانند مقیاس‌بندی مناسب محور، تراز و سازگاری بین پانل‌های مختلف.

1. Small multiples
2. compound figures

چندگانه‌های کوچک

اصطلاح «چندگانه کوچک» توسّط [Tufte 1990] رایج شد. یک اصطلاح جایگزین، «نمودار داربست» تقریباً در همان زمان توسّط Cleveland و Becker و همکارانش در آزمایشگاه بل رایج شد (Cleveland 1993؛ Becker, Cleveland, and Shyu 1996). صرف نظر از اصطلاحات، ایدهٔ کلیدی این است که داده‌ها را بر اساس یک یا چند بُعد از داده به قطعاتی برش دهیم، هر تکه از داده‌ها را به طور جداگانه ترسیم کنیم، و سپس نمودارهای منفرد را در یک شبکه مرتّب کنیم. ستون‌ها، ردیف‌ها یا پانل‌های مجزا در شبکه با مقادیر ابعادی که برش‌های داده را تعریف می‌کنند، برچسب‌گذاری می‌شوند. اخیراً، این روش گاهی اوقات به عنوان «صورت‌بندی[1]» نیز نامیده می‌شود، که نام آن برگرفته از روش‌هایی است که چنین نمودارهایی را در کتابخانه پرکاربرد ggplot2 ایجاد می‌کند (به عنوان مثال: ggplot2 function facet_grid(). [Wickham2016]

به عنوان مثال اوّل، این روش را در مجموعه دادهٔ مسافران تایتانیک اعمال خواهیم کرد. این مجموعه داده را می‌توانیم بر اساس طبقهٔ اقتصادی مسافران و اینکه آیا مسافر زنده مانده است یا نه، تقسیم کنیم. در هر یک از این شش قطعهٔ داده، مسافران زن و مرد وجود دارند و می‌توانیم تعداد آن‌ها را با استفاده از میله ترسیم کنیم. نتیجه شش نمودار میله‌ای است که در دو میله (یکی برای مسافران فوت شده و دیگری برای کسانی که جان سالم به در برده‌اند) و در سه ردیف (یکی برای هر طبقهٔ اقتصادی) نمایش می‌دهیم (نمودار ۲۱-۱). میله‌ها و ردیف‌ها برچسب‌گذاری شده‌اند، بنابراین به سرعت می‌توان متوجّه شد که کدام یک از شش نمودار مربوط به ترکیب دلخواه از وضعیّت بقا و طبقهٔ اقتصادی مسافری است.

این نمودار، تصویری شهودی و قابل تفسیر از سرنوشت مسافران تایتانیک ارائه می‌دهد. بلافاصله می‌توان دید که بیشتر مردان فوت شده‌اند و بیشتر زنان زنده مانده‌اند. علاوه بر این، در میان زنانی که جان باختند، تقریباً همه در طبقهٔ اقتصادی مسافری سوّم بوده‌اند.

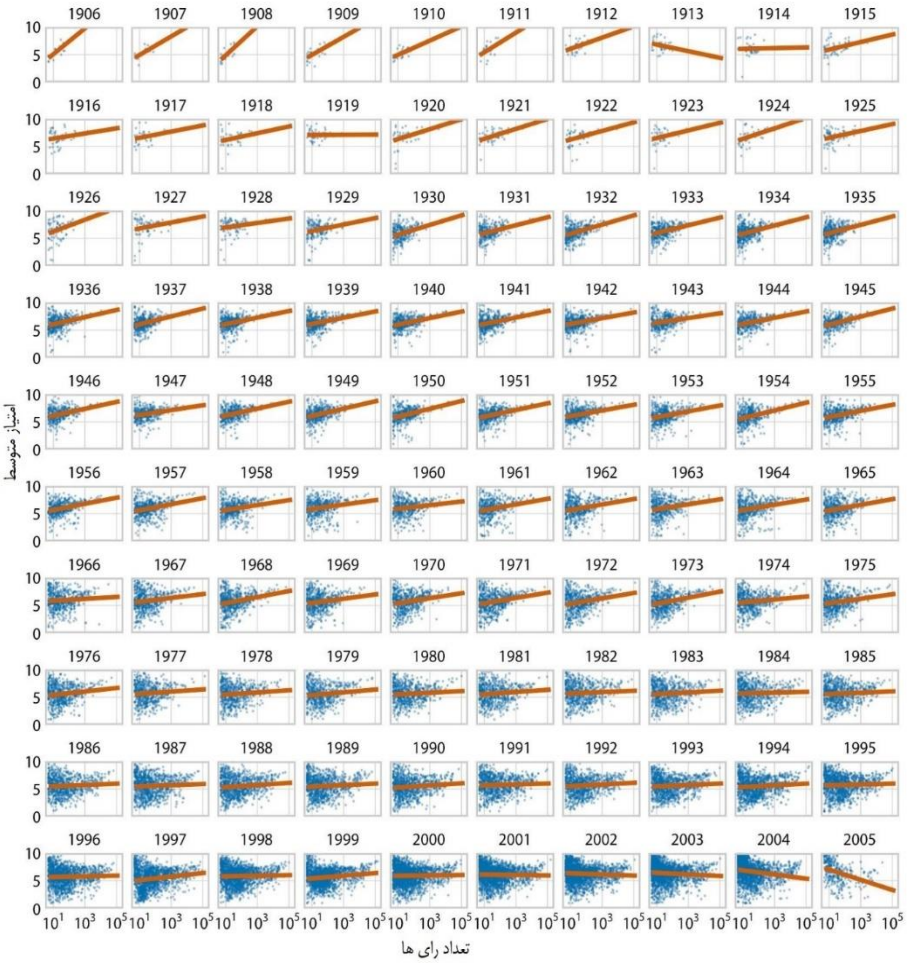
1. faceting

نمودار ۲۱–۱. تفکیک مسافران کشتی تایتانیک بر اساس جنسیّت، بقا و طبقهٔ اقتصادی مسافری (اوّل، دوّم، یا سوّم). منبع داده‌ها: دایره المعارف تایتانیکا.

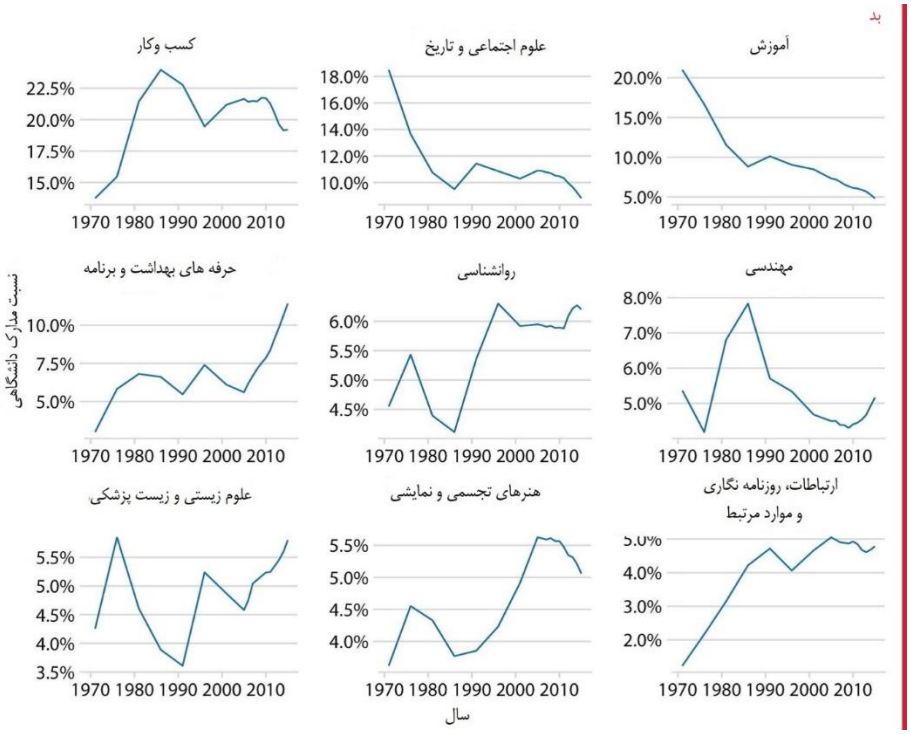
چندگانه‌های کوچک ابزار قدرتمندی برای نمایش همزمان حجم بسیار زیادی از داده‌ها است. نمودار ۲۱–۱ از شش پانل مجزا استفاده می‌کند، امّا می‌توانیم از تعداد بیشتری نیز استفاده کنیم. نمودار ۲۱–۲ رابطهٔ بین میانگین رتبه‌بندی یک فیلم در پایگاه داده اینترنتی فیلم (IMDB) و تعداد رأی‌هایی را که آن فیلم دریافت کرده است، به تفکیک برای فیلم‌هایی که در دورهٔ زمانی ۱۰۰ ساله منتشر شده‌اند، نشان می‌دهد. در اینجا، مجموعه داده تنها بر اساس یک بُعد یعنی سال برش داده شده است و پانل‌های هر سال در ردیف‌هایی از بالا سمت چپ به پایین راست مرتّب شده‌اند. این نمودار نشان می‌دهد که یک رابطهٔ کلی بین میانگین رتبه‌بندی و تعداد آرا وجود دارد، به طوری که فیلم‌هایی با رأی بیشتر، رتبهٔ بالاتری هم دارند. با این حال، قدرت این روند در سال‌های مختلف، متفاوت است و برای فیلم‌هایی که در اوایل دههٔ ۲۰۰۰ منتشر شده‌اند، هیچ رابطه‌ای وجود نداشته یا حتی این رابطه منفی است.

برای اینکه چنین نمودارهای بزرگی به راحتی قابل درک باشند، مهم است که محدودهٔ محور و مقیاس‌بندی پانل‌ها یکسان باشد. ذهن انسان انتظار دارد که چنین باشد. وقتی اینطور نیست، احتمال زیادی وجود دارد که خواننده آنچه را که نمودار نشان می‌دهد اشتباه تفسیر کند. به عنوان مثال، نمودار ۲۱–۳ را در نظر بگیرید، که نشان می‌دهد چگونه نسبت مدارک لیسانس اعطا شده در رشته‌های مختلف در طول زمان تغییر کرده است. نمودار، رشته‌های ۹ گانه را

نشان می‌دهد که به طور متوسّط بیش از ۴ درصد از تمام مدارک بین سال‌های ۱۹۷۱ تا ۲۰۱۵ اعطا شده است. محدودهٔ محور عمودی طوری تنظیم شده است که در هر نمودار مقادیر موجود تمام طول محور عمودی را اشغال کند. یک بررسی گذرا از نمودار ۲۱–۳ نشان می‌دهد که رشته‌های ۹ گانه به یک اندازه محبوب هستند و همه اندازهٔ مشابهی از تنوّع در محبوبیّت را تجربه کرده‌اند.

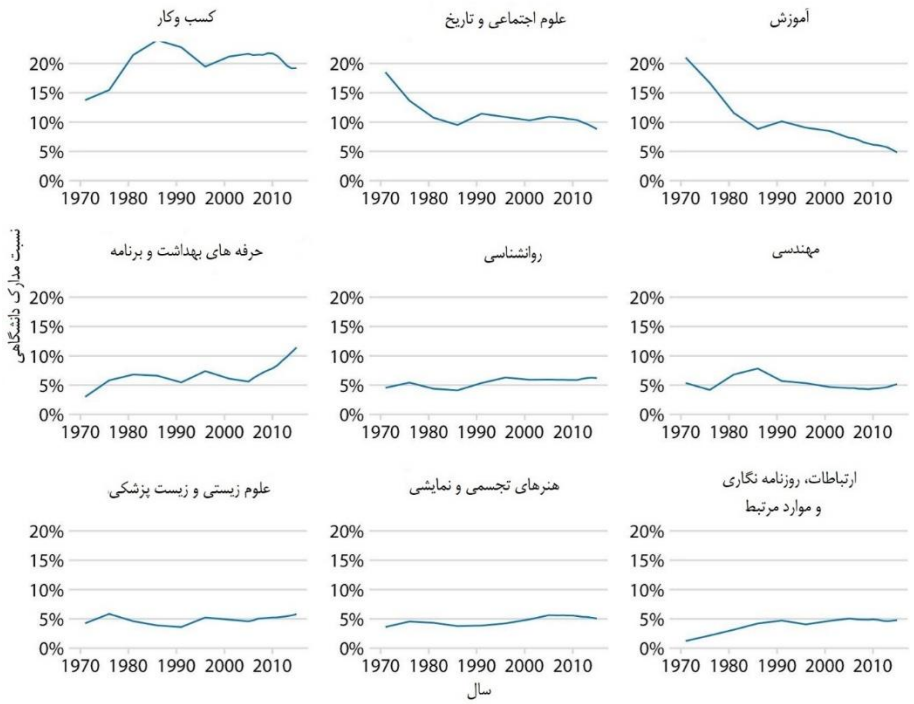


نمودار ۲۱–۲. میانگین رتبهٔ فیلم در مقابل تعداد آرا، برای فیلم‌هایی که از سال ۱۹۰۶ تا ۲۰۰۵ منتشر شده‌اند. نقطه‌های آبی نشان‌دهندهٔ فیلم‌ها و خطوط نارنجی نشان‌دهندهٔ خط رگرسیونی میانگین رتبهٔ هر فیلم در مقابل لگاریتم تعداد آرایی است که فیلم دریافت کرده است. در اغلب سال‌ها، فیلم‌هایی که تعداد رأی بالاتری دارند به طور میانگین رتبهٔ بالاتری نیز دارند. با این حال، این روند در اواخر قرن بیستم تضعیف می‌شود و یک رابطهٔ منفی را می‌توان برای فیلم‌هایی که در اوایل دهه ۲۰۰۰ منتشر شده‌اند مشاهده کرد. منبع داده: IMDB



نمودار ۲۱-۳. روندها در رشته‌های کارشناسی که توسّط موسسات آموزش عالی ایالات متحده ارائه می‌شود. رشته‌های نمایش داده شده به طور متوسط بیش از ۴ درصد تمام رشته‌ها را تشکیل می‌دهند. این نمودار به عنوان «بد» برچسب‌گذاری شده است زیرا محدودهٔ محور عمودی در پانل‌ها متفاوت است. این وضعیّت اندازه‌های نسبی نواحی رشته‌های مختلف را پنهان می‌کند و تغییراتی را که در برخی از رشته‌ها اتفاق افتاده است را اغراق‌آمیز جلوه می‌دهد. منبع داده ها: مرکز ملی آمار آموزش و پرورش

با این حال، قرار دادن تمام پانل‌ها روی یک محور عمودی نشان می‌دهد که این تفسیر گمراه‌کننده است (نمودار ۲۱-۴). برخی از رشته‌ها بسیار محبوب‌تر از سایرین هستند، و به طور مشابه محبوبیّت برخی از رشته‌ها بسیار بیشتر از سایرین افزایش یا کاهش یافته است. به عنوان مثال، محبوبیّت رشته‌های حوزه آموزش به شدت کاهش یافته است، در حالی که نسبت محبوبیّت رشته‌های تجسمی و هنرهای نمایشی تقریباً ثابت مانده است یا شاید افزایش اندکی داشته است.



نمودار ۲۱-۴. روندها در رشته‌های کارشناسی که توسّط موسسات آموزش عالی ایالات متحده ارائه می‌شود. رشته‌های نمایش داده شده به طور متوسّط بیش از ۴ درصد تمام رشته‌ها را تشکیل می‌دهند. منبع داده: مرکز ملی آمار آموزش و پرورش.

توصیه کلی این است که از مقیاس‌بندی مختلف برای محورهای پانل‌ها در نمودار چندگانه‌های کوچک استفاده نشود. با این حال، گاهی اوقات، این مساله اجتناب‌ناپذیر است. اگر با چنین سناریویی مواجه شدید، حداقل باید توجّه خواننده را به این موضوع در زیرنویس نمودار جلب کنید. برای مثال، می‌توان این جمله را اضافه کرد: «توجّه کنید که مقیاس‌گذاری‌های محور عمودی در پانل‌های مختلف این نمودار متفاوت است».

همچنین مهم است که در مورد ترتیب پانل‌های منفرد در نمودار چندگانه‌های کوچک نیز بیاندیشیم. اگر ترتیب پانل‌ها از اصول منطقی پیروی کند، تفسیر نمودار ساده‌تر خواهد بود. در نمودار ۲۱-۱، ردیف‌ها از بالاترین طبقهٔ اقتصادی (اوّل) به پایین‌ترین طبقهٔ اقتصادی (سوّم) مرتّب شده است. در نمودار ۲۱-۲، پانل‌ها با افزایش سال‌ها از سمت چپ بالا به سمت راست پایین مرتّب شده است. در نمودار ۲۱-۴، پانل‌ها بر اساس کاهش میانگین درجه محبوبیّت

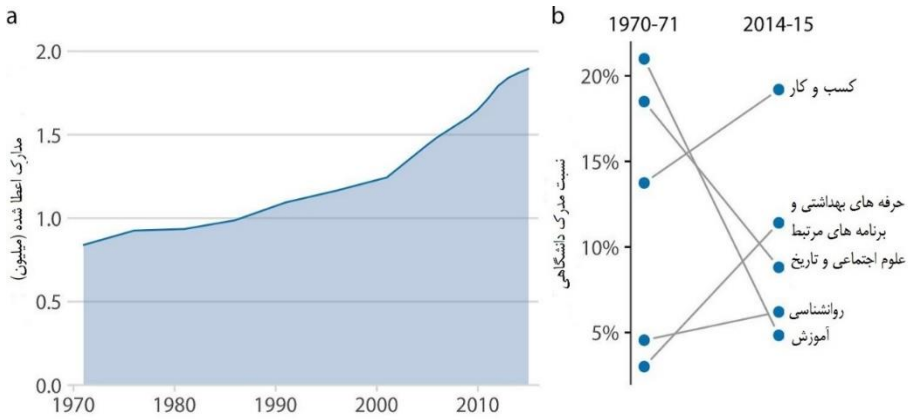
مرتّب شده‌اند، به طوری که محبوب‌ترین رشته‌ها در ردیف بالا و/یا به سمت چپ و کم محبوب‌ترین رشته‌ها در ردیف پایین و/یا به سمت راست قرار دارند.

همیشه پانل‌ها را در نمودار چندگانه‌های کوچک به ترتیب معنی‌دار و منطقی بچینید.

نمودارهای مرکّب

لزوماً هر نمودار با پانل‌های متعدد با الگوی نمودار چندگانه‌های کوچک مطابقت ندارد. گاهی اوقات تنها می‌خواهیم چندین پانل مستقل را در نموداری ترکیب کنیم که یک نکته کلی را بیان کند. در این صورت، می‌توانیم تک‌تک نمودارها را انتخاب کرده و آن‌ها را در ردیف‌ها، ستون‌ها یا ترکیب‌های پیچیده‌تر قرار دهیم و کل آرایش حاصل را یک نمودار بنامیم. به عنوان مثال، نمودار ۲۱–۵ را ببینید، که تحلیل روند رشته‌های کارشناسی که توسّط موسسات آموزش عالی ایالات متحده ارائه می‌شود را نشان می‌دهد. پانل (الف) نمودار ۲۱–۵ رشد تعداد کل مدارک اعطا شده را از سال ۱۹۷۱ تا ۲۰۱۵ نشان می‌دهد، بازهٔ زمانی که در طی آن مدارک تقریباً دو برابر شده است. در عوض، پانل (ب) تغییر در درصد مدارک اعطا شده در یک دورهٔ زمانی مشابه را در پنج رشتهٔ محبوب نشان می‌دهد. می‌توان دید که علوم اجتماعی، تاریخ و آموزش از سال ۱۹۷۱ تا ۲۰۱۵ افت شدیدی را تجربه کرده‌اند، در حالی که اقتصاد و بهداشت رشد قابل‌توجهی داشته‌اند.

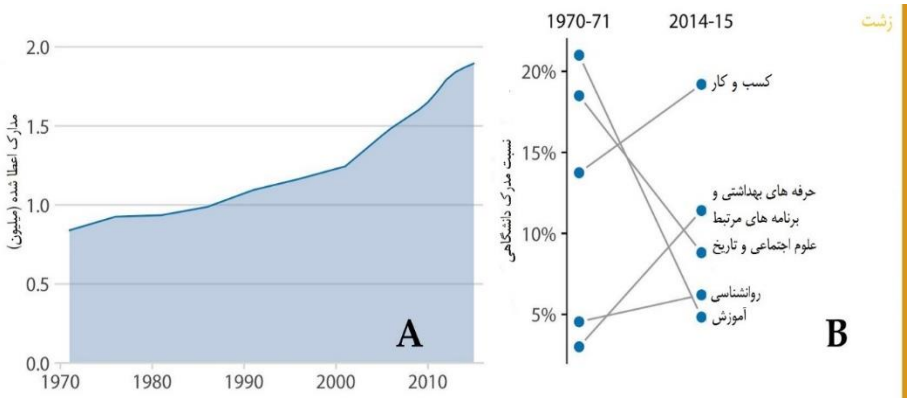
توجّه کنید که برخلاف مثال‌های چندگانه‌های کوچک، پانل‌های مجزای نمودار مرکّب بر اساس حروف الفبا برچسب‌گذاری شده‌اند. استفاده از حروف کوچک یا بزرگ از الفبای لاتین برای این برچسب‌گذاری مرسوم است تا یک پانل خاص به طور منحصربه‌فرد مشخّص شود. به عنوان مثال، هنگامی که می‌خواهیم در مورد بخشی از نمودار ۲۱–۵ صحبت کنیم که تغییرات درصد رشته‌های اعطا شده را نشان می‌دهد، می‌توانیم به پانل (ب) آن نمودار یا به سادگی به نمودار ۲۱–۵ ب اشاره کنیم. بدون برچسب زدن، باید به طرز ناخوشایندی در مورد «پانل سمت راست» یا «پانل چپ» نمودار ۲۱–۵ صحبت کنیم، و اشاره به پانل‌های خاص برای چیدمان‌های پیچیده‌تر حتّی عجیب‌تر خواهد بود. برای نمودار چندگانه‌های کوچک معمولاً برچسب‌گذاری مورد نیاز نیست و انجام هم نمی‌شود، زیرا در آنجا هر پانل به‌طور منحصربه‌فرد توسط متغیّر(های) صورت بندی که به عنوان برچسب‌های نمودار ارائه می‌شوند، مشخّص می‌شود.



نمودار ۲۱–۵. روندها در رشته‌های کارشناسی که توسّط موسسات آموزش عالی ایالات متحده ارائه می‌شود. (الف) از سال ۱۹۷۰ تا ۲۰۱۵، تعداد کل مدارک اعطا شده تقریباً دو برابر شده است. (ب) در میان محبوب‌ترین رشته‌های تحصیلی، علوم اجتماعی، تاریخ و آموزش کاهش عمده‌ای را تجربه کردند، در حالی که محبوبیّت اقتصاد و بهداشت افزایش یافته است. منبع داده: مرکز ملی آمار آموزش و پرورش.

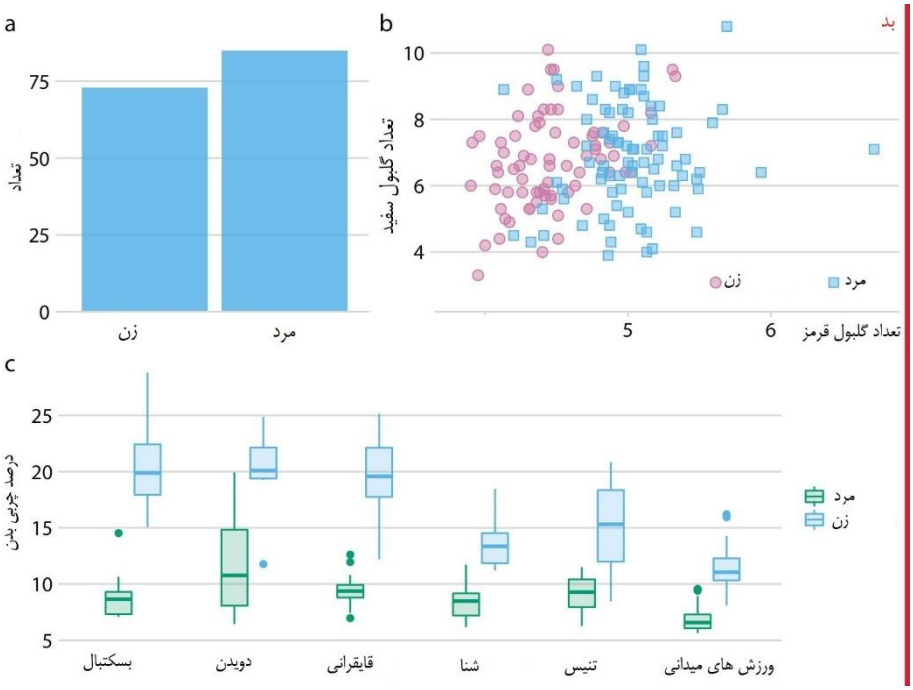
هنگام برچسب زدن پانل‌های مختلف یک نمودار مرکّب، به نحوهٔ تطابق برچسب‌ها با طرح کلی نمودار توجّه کنید. در برخی نمودارها گویی برچسب‌ها توسّط فرد دیگری طراحی شده و روی آن قرار گرفته است. در این حالت، برچسب‌ها بیش از حد بزرگ و برجسته بوده یا در مکانی نامناسب قرار می‌گیرند یا با قلم‌هایی متفاوت از بقیه نمودار تایپ شده‌اند (برای مثال به نمودار ۲۱–۶ مراجعه کنید). برچسب‌ها نباید اوّلین چیزی باشند که هنگام نگاه کردن به یک نمودار مرکّب می‌بینید. در واقع، اصلاً نیازی به برجسته‌سازی آن‌ها نیست. ما معمولاً می‌دانیم که کدام پانل نمودار دارای چه برچسبی است، زیرا به طور قراردادی از گوشه سمت چپ بالا با برچسب «الف» شروع شده و به طور متوالی از چپ به راست و از بالا به پایین برچسب زده خواهد شد. این برچسب‌ها معادل شماره صفحات هستند. معمولاً شماره‌های صفحه را نمی‌خوانیم و جای تعجّب نیست که کدام صفحه چه شماره‌ای دارد، امّا در مواردی استفاده از شماره صفحه برای اشاره به مکان خاصی در کتاب یا مقاله می‌تواند مفید باشد.

همچنین باید به نحوهٔ قرار گرفتن تک‌تک پانل‌های یک نمودار مرکّب با هم توجّه کنیم. ممکن است مجموعه‌ای از پانل‌های نمودار ساخته شود که به صورت جداگانه خوب هستند امّا در ترکیب با هم مناسب نیستند. به طور مشخص، ما باید از یک زبان بصری ثابت استفاده کنیم. منظور از «زبان بصری» رنگ‌ها، نمادها، قلم‌ها و غیره است که برای نمایش داده‌ها استفاده می‌شود. ثابت نگه داشتن زبان، به طور خلاصه، به این معنی است که موارد یکسان در نمودارها به صورت یکسان یا حداقل مشابه باشند.



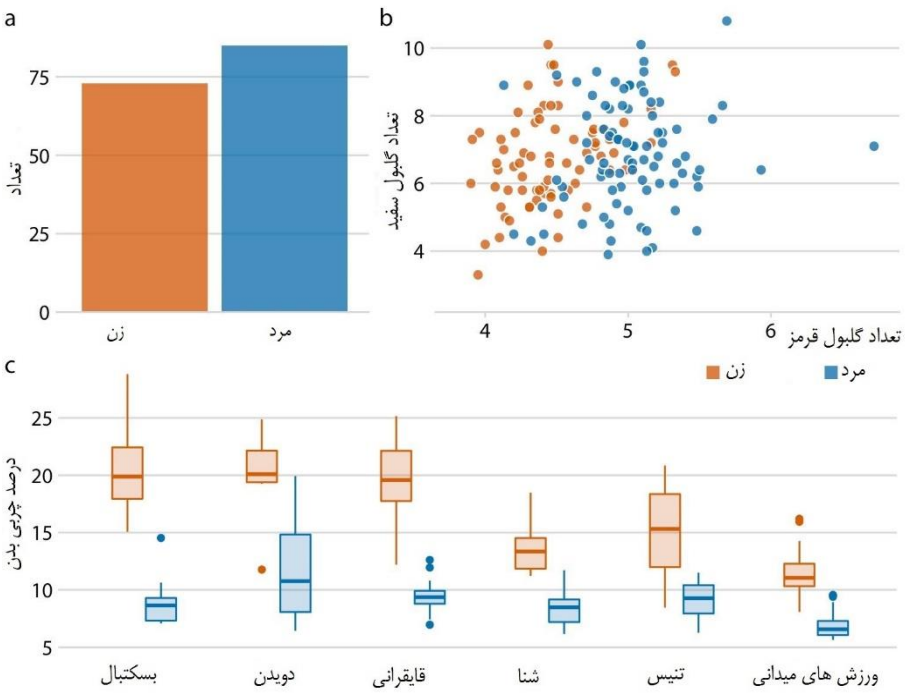
نمودار ۲۱–۶. گونه‌ای از نمودار ۲۱–۵ با برچسب‌گذاری ضعیف. برچسب‌های پانل بیش از حد بزرگ و ضخیم هستند، قلم اشتباه است و در مکان نامناسبی قرار دارد. همچنین، در حالی که برچسب زدن با حروف بزرگ خوب است و در واقع کاملاً رایج است، برچسب زدن باید در تمام نمودارهای یک سند یکسان باشد. در این کتاب، قرارداد این است که نمودارهای چند پانلی از برچسب‌های با حرف کوچک استفاده کنند و بنابراین این نمودار با دیگر نمودارهای این کتاب ناسازگار است. منبع داده: مرکز ملی آمار آموزش و پرورش.

بیایید به مثالی نگاه کنیم که این اصل را نقض می‌کند. نمودار ۲۱–۷ یک نمودار سه پانلی است که مجموعه داده‌ای را در مورد فیزیولوژی و ترکیب بدن ورزشکاران مرد و زن به تصویر می‌کشد. پانل (الف) تعداد مردان و زنان در مجموعه داده را نشان می‌دهد، پانل (ب) تعداد گلبول‌های قرمز و سفید خون را برای مردان و زنان نشان می‌دهد و پانل (ج) درصد چربی بدن مردان و زنان را به تفکیک نوع ورزش نشان می‌دهد. هر پانل به صورت جداگانه یک نمودار قابل قبول است. با این حال، ترکیب این سه پانل نتیجه قابل قبولی ندارد، زیرا آن‌ها زبان بصری مشترکی ندارند. اوّل، پانل (الف) از رنگ آبی یکسانی برای ورزشکاران مرد و زن استفاده می‌کند، پانل (ب) از این رنگ فقط برای ورزشکاران مرد و پانل (ج) از آن فقط برای ورزشکاران زن استفاده می‌کند. علاوه بر این، پانل‌های (ب) و (ج) از رنگ‌های اضافی استفاده می‌کنند، امّا این رنگ‌ها بین دو پانل متفاوت است. بهتر بود از دو رنگ تکرار شونده ثابت برای ورزشکاران زن و مرد در همه نمودارها استفاده می‌شد و در پانل (الف) نیز از طرح رنگ‌آمیزی مشابهی استفاده می‌شد. دوّم، در پانل (الف) و (ب) زنان در سمت چپ و مردان در سمت راست قرار دارند، امّا در پانل (ج) ترتیب برعکس است. ترتیب نمودارهای جعبه‌ای در پانل (ج) باید طوری تغییر کند که با پانل‌های (الف) و (ب) مطابقت داشته باشد.



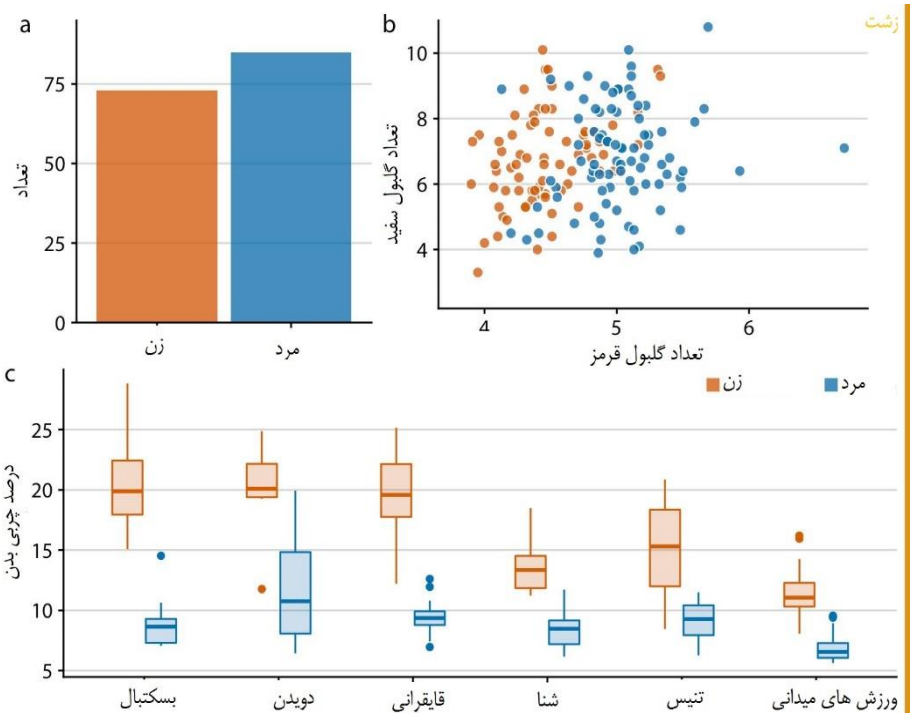
نمودار ۲۱–۷. فیزیولوژی و ترکیب بدنی ورزشکاران زن و مرد. (الف) مجموعه داده شامل ۷۳ ورزشکار حرفه‌ای زن و ۸۵ ورزشکار حرفه‌ای مرد است. (ب) تعداد گلبول‌های قرمز خون (RBC، گزارش شده در واحد 10^{12} در لیتر) در ورزشکاران مرد نسبت به ورزشکاران زن بیشتر است، امّا چنین تفاوتی برای تعداد گلبول‌های سفید خون (WBC، گزارش شده در واحد 10^{9} در لیتر) وجود ندارد. (ج) درصد چربی بدن در ورزشکاران مرد نسبت به ورزشکاران زن که همان ورزش را انجام می‌دهند، کمتر است. این نمودار به عنوان «بد» نامگذاری شده است زیرا بخش‌های (الف)، (ب) و (ج) از یک زبان بصری ثابت استفاده نمی‌کنند.
منبع داده: Telford and Cunningham 1991.

نمودار ۲۱–۸ تمام این مشکلات را برطرف نموده است. در این نمودار، ورزشکاران زن به طور ثابت با رنگ نارنجی و در سمت چپ ورزشکاران مرد با رنگ آبی نشان داده شده‌اند. توجّه داشته باشید که خواندن این نمودار بسیار آسان‌تر از نمودار ۲۱–۷ است. وقتی از یک زبان بصری ثابت استفاده می‌کنیم، تعیین اینکه کدام عناصر بصری در پانل‌های مختلف نشان‌دهندهٔ زنان و مردان هستند، به تلاش ذهنی زیادی نیاز ندارد. از سوی دیگر، نمودار ۲۱–۷ می‌تواند کاملا گیج‌کننده باشد. به طور خاص، در نگاه اوّل ممکن است این تصوّر ایجاد شود که درصد چربی بدن مردان بالاتر از زنان است. همچنین توجّه داشته باشید که ما در نمودار ۲۱–۸ فقط به یک راهنما نیاز داریم امّا در نمودار ۲۱–۷ به دو راهنما نیاز داریم. از آنجایی که زبان بصری با ثبات است، همان راهنما برای پانل‌های (ب) و (ج) کارایی دارد.



نمودار ۲۱–۸. فیزیولوژی و ترکیب بدنی ورزشکاران زن و مرد. این نمودار دقیقاً همان داده‌های نمودار ۲۱–۷ را نشان می‌دهد، امّا اکنون از یک زبان بصری ثابت استفاده می‌کند. داده‌های مربوط به ورزشکاران زن همیشه در سمت چپ داده‌های مربوط به ورزشکاران مرد نشان داده شده است، و دو جنسیّت به صورت با ثباتی در تمام عناصر نمودار کدگذاری شده‌اند. منبع داده: Telford and Cunningham 1991.

در نهایت، باید به تراز بودن پانل‌های مجزا در یک نمودار مرکّب توجّه نمود. محورها و سایر عناصر گرافیکی هر پانل باید همه با هم تراز باشند. تراز نمودن صحیح می‌تواند بسیار سخت باشد، به ویژه اگر پانل‌ها به طور جداگانه، و احتمالاً توسّط افراد مختلف و/یا در برنامه‌های مختلف، آماده شده و سپس در یک برنامه ویرایش تصویر به هم چسبانده شده باشند. برای جلب توجّه شما به چنین مسائلی در حوزه تراز کردن، نمودار ۲۱–۹ ویرایشی از نمودار ۲۱–۸ را نشان می‌دهد که در آن همهٔ عناصر نمودار کمی خارج از تراز هستند. خطوط محور به تمام پانل‌های نمودار ۲۱–۹ اضافه شده تا بر این مشکلات ناهم‌ترازی تأکید شود. توجّه کنید که چگونه هیچکدام از خطوط محوری در پانل‌های مختلف با هم تراز نیستند.



نمودار ۲۱–۹. ویرایشی از نمودار ۲۱–۸ که در آن تمام پانل‌های نمودار کمی ناهم‌تراز هستند. ناهم‌ترازی‌ها نمودار را زشت می‌کنند و باید از آن‌ها اجتناب کرد. منبع داده: Telford and Cunningham 1991

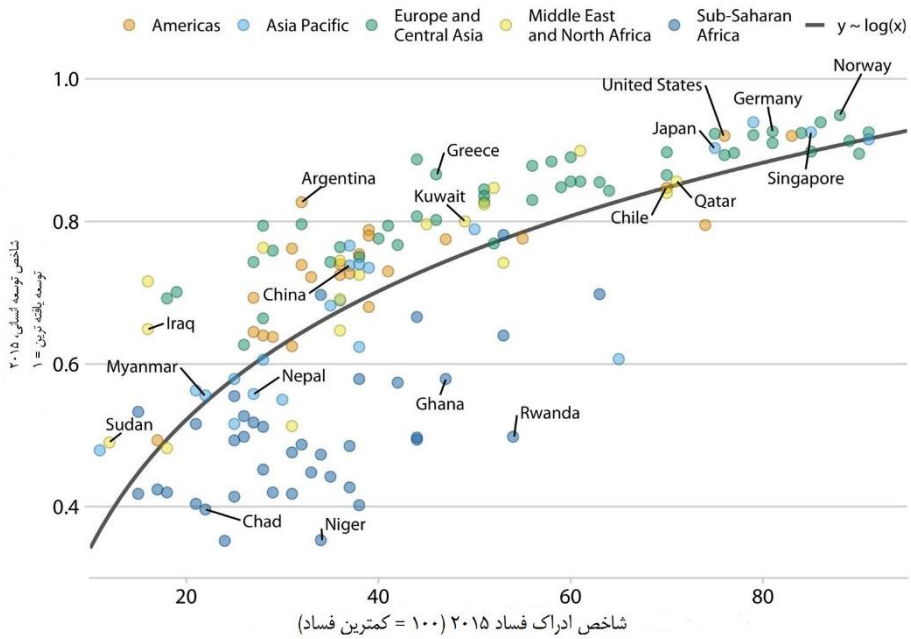
فصل ۲۲

عناوین، توضیحات و جداول

مصوّرسازی داده‌ها یک اثر هنری نیست که فقط به دلیل ویژگی‌های زیبایی‌شناسی آن مورد توجّه قرار گیرد. بلکه، هدف آن انتقال اطّلاعات و انتقال یک مفهوم است. برای دستیابی مطمئن به این هدف هنگام تهیه ترسیم‌ها، باید داده‌ها را در بستر موضوع مربوطه مدنظر قرار دهیم و عناوین، توضیحات و سایر حاشیه‌نویسی‌های همراه را ارائه دهیم. در این فصل، نحوهٔ صحیح تنظیم عنوان و برچسب‌گذاری شکل‌ها را مورد بحث قرار خواهیم داد. همچنین در مورد نحوهٔ ارائه داده‌ها به شکل جدول بحث خواهیم کرد.

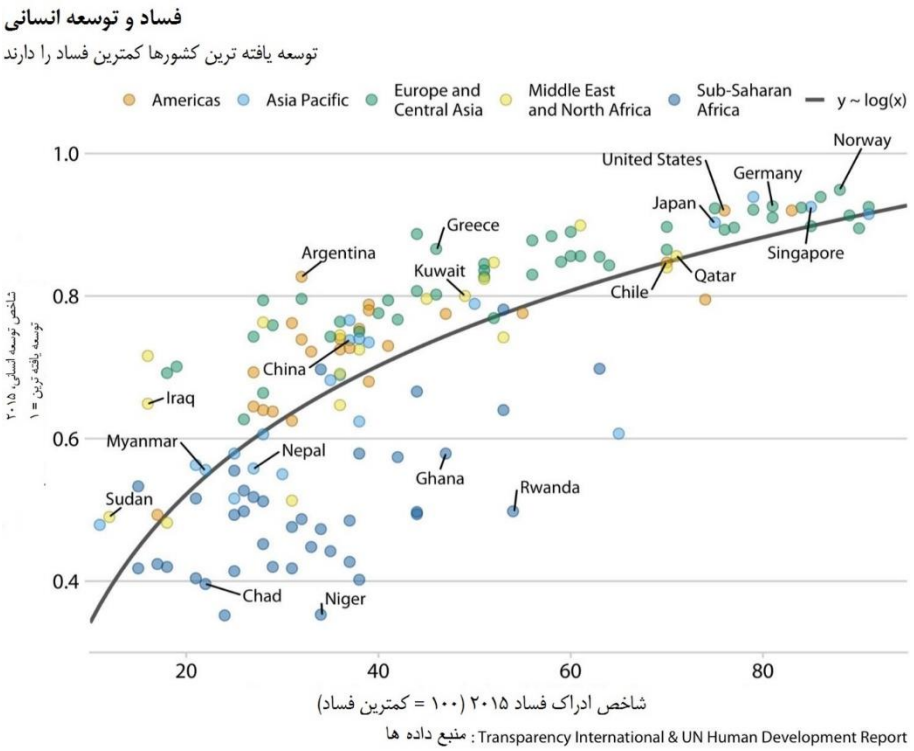
عناوین و توضیحات شکل‌ها

یکی از اجزای مهم هر شکل عنوان است. هر شکلی به یک عنوان نیاز دارد. وظیفهٔ عنوان این است که به طور دقیق به خواننده بگوید که این شکل در مورد چه چیزی است و چه نکته‌ای را نشان می‌دهد. با این حال، عنوان شکل ممکن است لزوماً در جایی که انتظار مشاهده آن را دارید، ظاهر نشود. شکل ۲۲-۱ را در نظر بگیرید. عنوان آن عبارت است از «فساد و توسعهٔ انسانی: توسعه یافته‌ترین کشورها کمترین فساد را تجربه می‌کنند». این عنوان در بالای شکل نشان داده نشده است. بلکه، عنوان در اوّلین قسمت از توضیحات، در زیر شکل ارائه شده است. این سبکی است که در سراسر این کتاب استفاده شده است. ما اشکال را بدون عناوین منسجم و با شرح‌های جداگانه نشان داده‌ایم (یک استثنا، نمونه‌های طرح‌ریزی شده در فصل ۵ هستند که عنوان دارند و بدون زیرنویس هستند).



شکل ۲۲–۱. فساد و توسعهٔ انسانی: توسعه یافته‌ترین کشورها کمترین فساد را تجربه می‌کنند. ایده اصلی شکل: [اکونومیست آنلاین ۲۰۱۱]. منابع داده: گزارش شفافیت بین‌الملل و توسعهٔ انسانی سازمان ملل.

از طرف دیگر، می‌توانیم عنوان شکل و همچنین سایر اجزای توضیحات مانند منبع داده را در قسمت اصلی شکل بگنجانیم (شکل ۲۲–۲). در یک مقایسهٔ مستقیم، ممکن است شکل ۲۲–۲ جذاب‌تر از شکل ۲۲–۱ به نظر برسد، و ممکن است تعجّب کنید که چرا از سبک دوّم در سراسر این کتاب استفاده شده است. زیرا این دو سبک دارای حوزه‌های کاربردی متفاوتی هستند و اشکال با عنوان‌های منسجم برای چیدمان کتاب‌های معمولی مناسب نیستند. اصل اساسی این است که یک شکل می‌تواند تنها یک عنوان داشته باشد، یا عنوان در قسمت اصلی شکل ادغام شده است یا به عنوان اوّلین عنصر توضیحات در زیر شکل ارائه می‌شود و اگر یک نشریه طوری تنظیم شده باشد که هر شکل واجد یک بخش توضیحات در زیر قسمت اصلی باشد، باید عنوان در قسمت توضیحات ارائه شود. به همین دلیل، در حوزهٔ انتشار معمول کتاب یا مقاله، ما معمولاً عناوین را در شکل‌ها ادغام نمی‌کنیم. با این حال، اگر قرار باشد از شکل‌ها به عنوان اینفوگرافیک مستقل استفاده شود یا در رسانه‌های اجتماعی یا در یک صفحه وب بدون توضیحات منتشر شود، اشکال با عناوین، عناوین فرعی (توضیحات) و منبع داده یکپارچه مناسب هستند.



شکل ۲۲–۲. نسخهٔ اینفوگرافیک شکل ۲۲–۱. عنوان، عنوان فرعی، و منبع داده در شکل گنجانده شده است. این شکل را می‌توان آنطور که هست در وب منتشر کرد یا بدون نیاز به بخش توضیحات جداگانه، استفاده کرد.

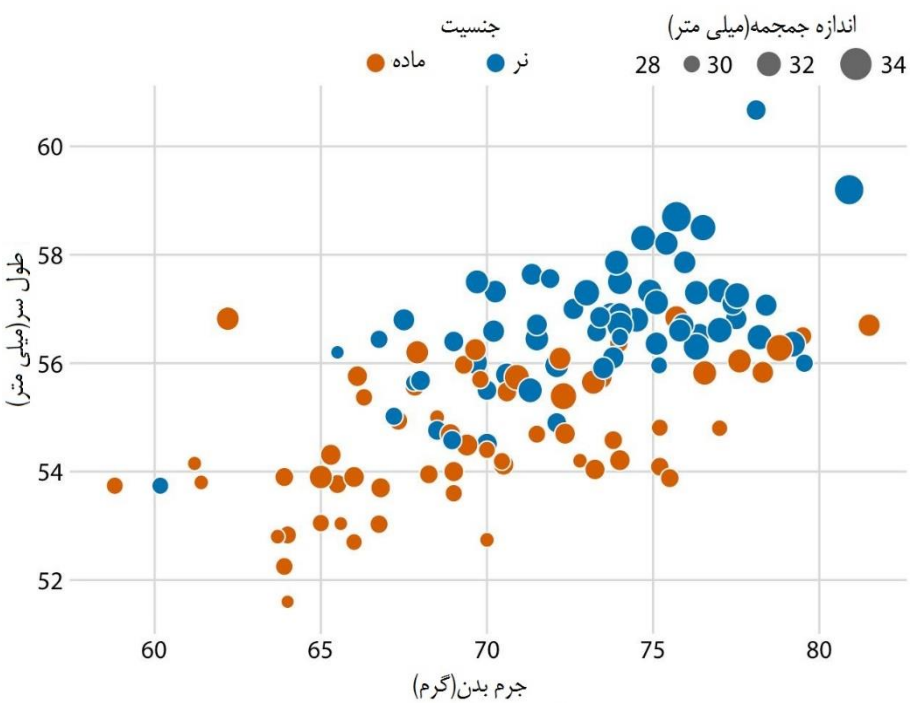
اگر طرح‌بندی سند شما از توضیحات در زیر هر شکل استفاده می‌کند، عناوین شکل‌ها را به‌عنوان اوّلین عنصر هر توضیحات قرار دهید، نه در بالای شکل‌ها.

یکی از رایج‌ترین اشتباهاتی که در توضیحات تصاویر می‌بینیم، حذف عنوان مناسب شکل به عنوان اوّلین عنصر توضیحات است. مجدداً نگاهی به توضیحات تصویر ۲۲–۱ بیندازید. با «فساد و توسعهٔ انسانی» شروع می‌شود. توضیحات با «این شکل نشان می‌دهد که فساد چگونه با توسعهٔ انسانی مرتبط است» آغاز نمی‌شود. قسمت اوّل توضیحات همیشه عنوان است، نه توصیفی از محتوای شکل. لازم نیست عنوان یک جمله کامل باشد، اگرچه جملات کوتاه که مطلب واضحی را بیان می‌کنند، می‌توانند برای عنوان به کار روند. به عنوان مثال، برای شکل ۲۲–۱، عنوانی مانند «توسعه یافته‌ترین کشورها کمترین فساد را دارند» مناسب است.

عناوین محور و راهنما

همانطور که هر نمودار بـه عنـوان نیـاز دارد، محورهـا و راهنماهـا نیـز بـه عنـوان نیـاز دارنـد (عنـاوین محورهـا اغلـب بـه صـورت محـاوره‌ای بـه عنـوان برچسـب‌های محـور نامیـده می‌شـوند). عنـاوین و برچسب‌های محورهـا و راهنماهـا توضـیح می‌دهنـد کـه مقـادیر داده‌هـای نمـایش داده شـده چیسـت و چگونه آن‌ها طرح زیبایی‌شناسی را ترسیم می‌کنند.

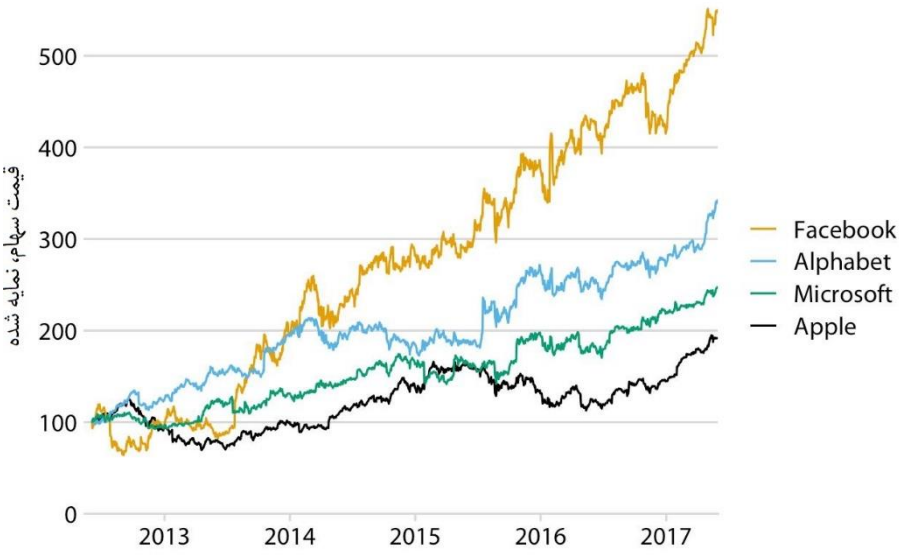
بـــرای ارائـــه نمونـــه‌ای از نمـــوداری کـــه در آن همـــه محورهـــا و راهنماهـــا بـــه‌طور مناســـب برچسـب‌گذاری و عنوان‌گـذاری شـده‌اند، مجموعـه داده زاغ آبـی کـه در فصـل ۱۲ بـه‌طور مفصّـل مـورد بحـث قـرار گرفـت، را در قالـب یـک نمـودار حبـابی ترسـیم کـرده‌ایم (شـکل ۲۲-۳). در ایـن نمودار، عنـاوین محورهـا نشـان می‌دهـد کـه محـور x وزن بـدن را بـر حسـب گـرم و محـور y طـول سر را بر حسب میلی‌متر نشان می‌دهد.



شکل ۲۲-۳. طول سر در مقایسه با وزن بدن برای ۱۲۳ زاغ آبی. جنسیّت پرندگان با رنگ و اندازهٔ جمجمه پرندگان با اندازه نماد مشخّص می‌شود. اندازه‌گیری طول سر شامل طول منقار است در حالی که اندازه‌گیری جمجمه اینطور نیست. منبع داده: Keith Tarvin, Oberlin College

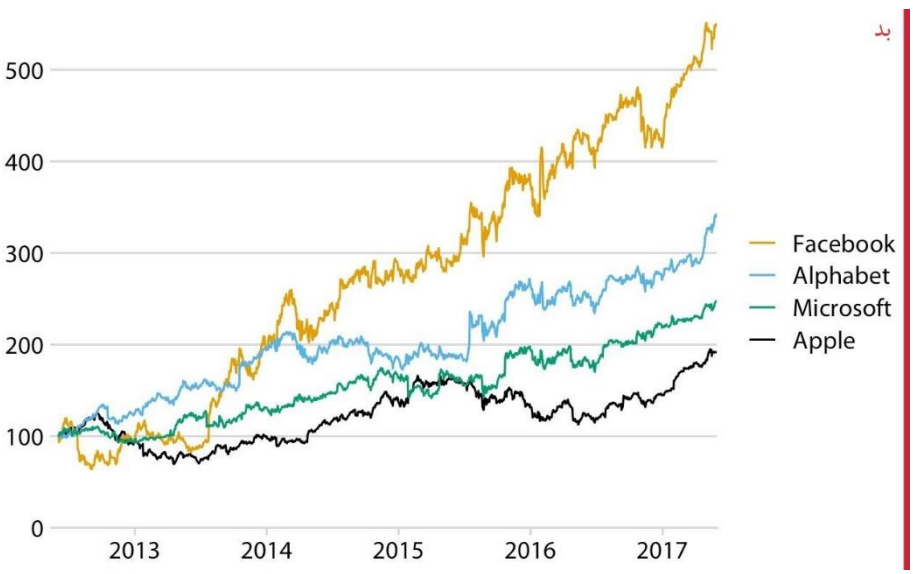
به طور مشابه، عناوین راهنما نشان می‌دهد که رنگ‌آمیزی نقطه‌ها جنسیّت پرندگان را نشان می‌دهد و اندازهٔ نقطه نشان‌دهندهٔ اندازهٔ جمجمه پرندگان برحسب میلی‌متر است. تأکید می‌کنیم که برای همه متغیّرهای عددی (وزن بدن، طول سر و اندازهٔ جمجمه) عناوین مربوطه نه تنها متغیّرهای نشان داده شده را بیان می‌کنند، بلکه واحدهایی را که متغیّرها بر اساس آن‌ها اندازه‌گیری شده‌اند نیز ارائه می‌دهند. این روش خوبی است و باید هر وقت مقدور بود، انجام شود. متغیّرهای گروه‌بندی شده (مانند جنس) نیازی به واحد ندارند.

با این حال، مواردی وجود دارد که عناوین محور یا راهنما را می‌توان حذف کرد، مثلاً زمانی که خود برچسب‌ها کاملاً شفاف هستند. برای مثال، در راهنمایی که دو نقطه با رنگ‌های متفاوت با برچسب‌های «مونث» و «مذکر» را نشان می‌دهد، مشخّص است که رنگ در حال کدگذاری جنسیّت می‌باشد. عنوان «جنسیت» برای روشن شدن این واقعیّت لازم نیست، و در واقع در سراسر این کتاب، اغلب عنوان راهنما را در مواردی که جنسیّت را نشان می‌دهند حذف کرده‌ایم (به عنوان مثال، به شکل‌های ۶–۱۰، ۱۲–۲، یا ۲۱–۱ مراجعه کنید). به طور مشابه، نام کشورها (شکل ۶–۱۱)، فیلم (شکل ۶–۱) یا سال (شکل ۲۲–۴) معمولاً نیازی به عنوان ندارند.



شکل ۲۲–۴. قیمت سهام در طول زمان برای چهار شرکت بزرگ فناوری. قیمت سهام برای هر شرکت در ژوئن ۲۰۱۲ معادل عدد ۱۰۰ نرمال شده است. این شکل نسخهٔ کمی تغییر یافته از شکل ۲۰–۶ در فصل ۲۰ است. در اینجا، محور x که نشان‌دهندهٔ زمان است عنوان ندارد زیرا مشخّص است که اعداد ۲۰۱۳، ۲۰۱۴ و غیره مربوط به سال هستند. منبع داده: امور مالی شرکت یاهو

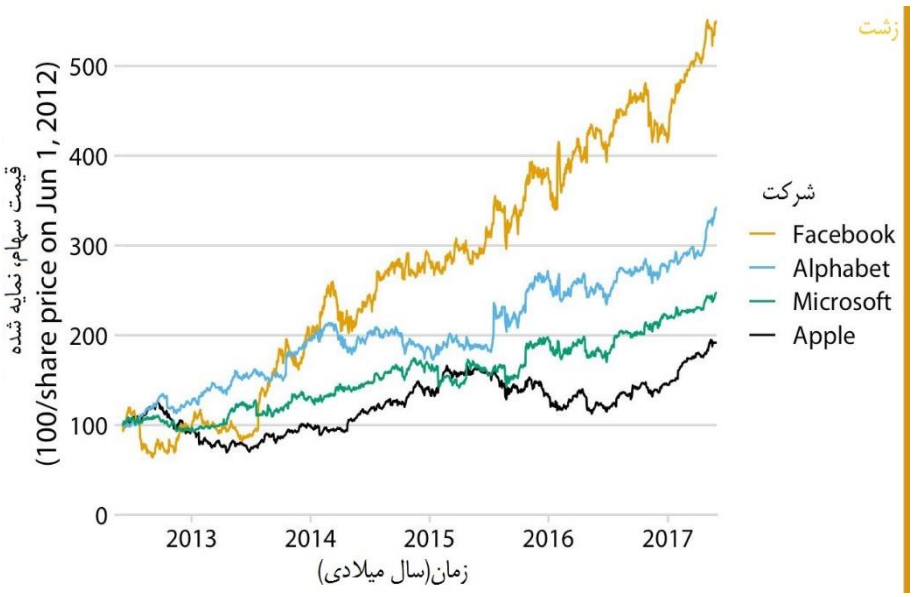
با این حال، هنگام حذف عناوین محور یا راهنما باید مراقب باشیم، زیرا به راحتی می‌توان موارد مشخّص و نامشخّص را به اشتباه قضاوت کرد. اغلب نمودارهایی را در مطبوعات رایج می‌بینیم که حذف عناوین محورها را به نقطه‌ای می‌رسانند که باعث ناخرسندی مخاطب می‌شوند. برای مثال، برخی از نشریات ممکن است شکلی مانند شکل ۲۲–۵ ارائه دهند، با این فرض که معنای محورها از عنوان شکل و عنوان فرعی قابل درک است (در اینجا: «قیمت سهام در طول زمان برای چهار شرکت بزرگ فناوری» و «قیمت سهام برای هر شرکت در ژوئن ۲۰۱۲ معادل عدد ۱۰۰ نرمال شده است»). ما با این دیدگاه مخالفیم که زمینهٔ موضوعی، محورها را تعریف می‌کند. از آنجایی که یک عنوان معمولاً شامل کلماتی مانند «محور x/y ... را نشان می‌دهد» نیست، برای تفسیر شکل همیشه مقداری حدس و گمان لازم است. تجربه نشان‌دهندهٔ آن است که اشکال بدون محورهای برچسب‌گذاری شده صحیح باعث می‌شوند که با احساس ناخوشایند عدم اطمینان مواجه شویم – حتی اگر ۹۵ درصد مطمئن باشیم که آنچه نشان داده شده را متوجّه می‌شویم، با این حال اطمینان ۱۰۰ درصد نداریم. به عنوان یک اصل کلی، به نظر نامناسب است که انتظار داشته باشید خوانندگان منظورتان را حدس بزنند. چرا می‌خواهید در خوانندگان خود احساس عدم اطمینان ایجاد کنید؟



شکل ۲۲–۵. قیمت سهام در طول زمان برای چهار شرکت بزرگ فناوری. قیمت سهام برای هر شرکت در ژوئن ۲۰۱۲ معادل عدد ۱۰۰ نرمال شده است. این نسخه از شکل ۲۲–۴ به عنوان «بد» برچسب‌گذاری شده است زیرا محور y اکنون عنوانی ندارد و اینکه مقادیر نشان داده شده در امتداد محور y چه چیزی را نشان می‌دهند، به سرعت قابل درک نیست. منبع داده: امور مالی شرکت یاهو

از طرف دیگر، ممکن است در برچسب‌گذاری زیاده‌روی کنیم. اگر راهنما نام چهار شرکت معروف را فهرست کند، عنوان راهنمای «شرکت» اضافی است و هیچ چیز مفیدی اضافه نمی‌کند (شکل ۲۲–۶). به طور مشابه، هرچند به طور کلی باید واحدها را برای همه متغیّرهای کمی گزارش کنیم، اگر محور x چند سال اخیر را نشان دهد، نام‌گذاری آن به صورت «زمان (سال‌های پس از میلاد)» نامناسب است.

در برخی موارد، نه تنها حذف عنوان محور بلکه حذف کل محور قابل قبول است. نمودارهای دایره‌ای (به عنوان مثال، شکل ۱۰–۱) و همچنین نقشهٔ درختی (شکل ۱۱–۴) معمولاً دارای محورهای واضحی نیستند. در صورتی که معنای نمودار واضح باشد، نمودارهای موزاییکی یا ستونی را می‌توان بدون یک یا هر دو محور ترسیم کرد (شکل‌های ۶–۱۰ و ۱۱–۳). حذف محورها با تیک‌های مربوطه و برچسب‌های تیک به خواننده این پیام را می‌دهد که ویژگی‌های کیفی نمودار مهم‌تر از مقادیر داده‌های خاص است.



شکل ۲۲–۶. قیمت سهام در طول زمان برای چهار شرکت بزرگ فناوری. قیمت سهام برای هر شرکت در ژوئن ۲۰۱۲ معادل عدد ۱۰۰ نرمال شده است. این نسخه از شکل ۲۲–۴ به دلیل برچسب‌گذاری بیش از حد به عنوان «زشت» برچسب‌گذاری شده است. به طور خاص، ارائه واحد («سال پس از میلاد») برای مقادیر در امتداد محور xنامناسب و غیر ضروری است. منبع داده: امور مالی یاهو

جداول

جداول ابزار مهمی برای ترسیم داده‌ها هستند. با این حال، به دلیل سادگی ظاهری، ممکن است همیشه توجهی که شایسته آن است را دریافت نکنند. جداول متعددی در سراسر این کتاب نشان داده شده است از جمله جداول ۶–۱، ۷–۱، و ۱۹–۱. لطفاً کمی وقت بگذارید و مکان این جداول را پیدا کنید، به نحوهٔ کادربندی آن‌ها نگاه کنید و آن‌ها را با جدولی که اخیراً خود یا یکی از همکارانتان ترسیم کرده‌اید مقایسه کنید. به احتمال زیاد، تفاوت‌های مهمی وجود دارد. با عدم آموزش مناسب برای کادربندی جدول، تعداد کمی از افراد به طور غریزی کادربندی درست را انجام می‌دهند. در اسناد شخصی، جداول با کادربندی ضعیف حتی بیشتر از شکل‌های طراحی شده نامناسب هستند. همچنین، اکثر نرم‌افزارهایی که معمولاً برای ایجاد جداول استفاده می‌شوند، پیش فرض‌هایی را ارائه می‌دهند که توصیه شده نیستند. به عنوان مثال، نسخه Microsoft Word ۱۰۵ سبک جدول از پیش تعریف شده را ارائه می‌دهد، و حداقل ۷۰ یا ۸۰ مورد از آن‌ها برخی از قوانین جدولی را که در اینجا بحث می‌کنیم، نقض می‌کنند. بنابراین، اگر یک طرح‌بندی جدول Microsoft Word را به‌طور تصادفی انتخاب کنید، تقریباً ۸۰ درصد احتمال دارد که کادربندی جدولی را انتخاب کنید که مشکل دارد. اگر هم حالت پیش فرض را انتخاب کنید، هر بار با یک جدول با قالب‌بندی ضعیف مواجه خواهید شد.

برخی از قوانین کلیدی برای کادربندی جدول به شرح زیر است:

۱. از خطوط عمودی استفاده نکنید.
۲. از خطوط افقی بین ردیف‌های داده استفاده نکنید (خطوط افقی به عنوان جداکننده بین ردیف عنوان و ردیف اول داده یا به عنوان یک قاب برای کل جدول مناسب هستند).
۳. ستون‌های متنی باید تراز چپ شوند (از چپ هم راستا شوند).
۴. ستون‌های اعداد باید تراز راست شوند (از راست هم راستا شوند) و باید از تعداد ارقام اعشاری مشابهی استفاده کنند.
۵. ستون‌های حاوی نویسه‌های مجزا باید در مرکز قرار گیرند.
۶. سرستون‌ها باید منطبق بر داده‌های خود تراز باشد. به عنوان مثال، عنوان یک ستون متنی تراز چپ و عنوان یک ستون عددی در تراز راست قرار می‌گیرد.

شکل ۲۲–۷ جدول ۶–۱ را به چهار روش مختلف بازتولید می‌کند که دو مورد از آن‌ها (الف، ب) چندین مورد از این قوانین را نقض می‌کنند و دو مورد از آن‌ها (ج، د) نقض نمی‌کنند.

a

Rank	Title	Amount
1	*Star Wars: The Last Jedi*	$71,565,498
2	*Jumanji: Welcome to the Jungle*	$36,169,328
3	*Pitch Perfect 3*	$19,928,525
4	*The Greatest Showman*	$8,805,843
5	*Ferdinand*	$7,316,746

b

Rank	Title	Amount
1	*Star Wars: The Last Jedi*	$71,565,498
2	*Jumanji: Welcome to the Jungle*	$36,169,328
3	*Pitch Perfect 3*	$19,928,525
4	*The Greatest Showman*	$8,805,843
5	*Ferdinand*	$7,316,746

c

Rank	Title	Amount
1	*Star Wars: The Last Jedi*	$71,565,498
2	*Jumanji: Welcome to the Jungle*	$36,169,328
3	*Pitch Perfect 3*	$19,928,525
4	*The Greatest Showman*	$8,805,843
5	*Ferdinand*	$7,316,746

d

Rank	Title	Amount
1	*Star Wars: The Last Jedi*	$71,565,498
2	*Jumanji: Welcome to the Jungle*	$36,169,328
3	*Pitch Perfect 3*	$19,928,525
4	*The Greatest Showman*	$8,805,843
5	*Ferdinand*	$7,316,746

شکل ۲۲–۷. نمونه‌هایی از جداول با کادربندی ضعیف و مناسب، با استفاده از داده‌های جدول ۶–۱ در فصل ۶. (الف) این جدول قوانین متعدد کادربندی مناسب جدول، از جمله استفاده از خطوط عمودی، استفاده از خطوط افقی بین ردیف‌های داده، و استفاده از ستون‌های داده با تراز وسط را نقض می‌کند. (ب) این جدول تمام مشکلات (الف) را دارد و همچنین با جابجایی بین ردیف‌های بسیار تاریک و بسیار روشن اختلال بصری ایجاد می‌کند. همچنین سرستون‌های جدول به اندازهٔ کافی از بدنهٔ جدول جدا نیست. (ج) این جدول با قالب مناسب با طراحی ساده است. (د) از رنگ‌ها می‌توان به طور موثر برای گروه‌بندی داده‌ها در ردیف‌ها استفاده کرد، امّا تفاوت رنگ‌ها باید اندک باشد. سرستون جدول را می‌توان با استفاده از رنگ متمایزتری تنظیم کرد. منبع داده:Box Office Mojo . استفاده پس از کسب اجازه.

وقتی نویسندگان جداولی با خطوط افقی بین ردیف‌های داده ترسیم می‌کنند، معمولاً هدف این است که به چشم کمک کند تا ردیف‌های جداگانه را دنبال کند. با این حال، تنها در مواردی که جدول بسیار گسترده و پراکنده باشد، معمولاً به این کمک بصری نیاز است. در یک متن معمولی نیز خطوط افقی بین ردیف‌ها نمی‌کشیم. هزینه خطوط افقی (یا عمودی) بهم ریختگی بصری است. قسمت‌های (الف) و (ج) شکل ۲۲–۷ را مقایسه کنید. خواندن قسمت (ج) بسیار راحت‌تر از قسمت (الف) است. اگر احساس کنیم که کمک بصری برای جدا کردن ردیف‌های جدول ضروری است، در آن صورت سایه‌های روشن‌تر و تیره‌تر ردیف‌ها بدون ایجاد اختلال بصری زیاد کاربردی خواهد بود (شکل ۲۲–۷ د).

در نهایت، یک تمایز کلیدی بین شکل‌ها و جداول در مورد مکانی که توضیحات نسبت به بخش اصلی قرار داده می‌شود، وجود دارد. برای شکل‌ها معمولاً توضیحات پایین قرار می‌گیرد، در حالی که برای جداول معمولاً آن را در بالا قرار می‌دهند. این قرار دادن توضیحات بر اساس

روشی است که خوانندگان شکل‌ها و جداول را پردازش می‌کنند. در مورد شکل‌ها، خوانندگان تمایل دارند ابتدا به تصویر نگاه کنند و سپس توضیحات را برای درک موضوع بخوانند، از این رو توضیحات در پایین شکل معنا می‌یابد. در مقابل، جداول عمدتاً مانند متن، از بالا به پایین پردازش می‌شوند، و خواندن محتوای جدول قبل از خواندن توضیحات اغلب مفید نخواهد بود. از این رو، توضیحات در بالای جدول قرار می‌گیرند.

فصل ۲۳

متعادل‌سازی داده‌ها و زمینه

در یک تقسیم‌بندی کلی، می‌توان عناصر گرافیکی در هر مصوّرسازی را به عناصری که نشان‌دهندهٔ داده‌ها هستند و عناصری که نشان‌دهندهٔ داده‌ها نیستند دسته‌بندی کرد. نوع اوّل شامل عناصری مثل نقاط در نمودار پراکنش، میله‌ها در هیستوگرام یا نمودار میله‌ای و یا منطقه رنگی در نقشه‌های حرارتی می‌باشند. نوع دوّم عناصر شامل محورهای نمودار، تیک‌ها و برچسب‌های محور، عنوان محورها، راهنما و یادداشت‌های نمودار می‌شوند. این عناصر به طور کلی زمینه‌ای برای داده‌ها و/یا ساختار بصری برای آن فراهم می‌کنند. هنگام طراحی یک نمودار توجّه به مقدار جوهری (فصل ۱۷) که برای ارائه داده‌ها و زمینهٔ آن استفاده می‌شود حائز اهمیت است. یک توصیه رایج کاهش رنگ بدون داده است و پیروی از آن می‌تواند با کاهش شلوغی نمودار و افزایش آراستگی تصویر همراه باشد. در عین حال زمینه و ساختارهای بصری مهم هستند و بیش از حد کوچک کردن عناصری که شامل آن‌ها می‌شود منجر به ایجاد شکل‌هایی می‌شود که خواندن آن‌ها سخت و گیج‌کننده بوده و جذّاب نیستند.

ارائه مقدار مناسبی از زمینه

ایدهٔ اینکه تمایز بین رنگ داده‌ها و رنگ فاقد داده‌ها ممکن است سودمند باشد در ابتدا توسّط ادوارد توفت[1] در کتابش به نام نمایش بصری اطّلاعات کمّی ارائه شد. توفت مفهوم «نسبت

1. Edward Tufte

داده به جوهر» را معرفی کرد که عبارت است از «نسبت جوهر گرافیک اختصاص داده شده به نمایش غیرزاید اطّلاعات داده‌ها». سپس می‌نویسد:

نسبت داده به جوهر را «با دلیل» افزایش دهید.

ما بر عبارت «با دلیل» تأکید کرده‌ایم چون مهم است و مرتباً فراموش می‌شود. در حقیقت، تصوّر ما بر این است که توفت نیز در بقیه کتابش این مطلب را فراموش می‌کند، زیرا او بیش از حد از طراحی‌های کمینه‌ای[1] دفاع می‌کند که به نظر نه زیبا هستند و نه درک آن‌ها آسان است. اگر عبارت «افزایش نسبت داده به جوهر» را به معنی «برطرف کردن آشفتگی، شلوغی و درهمی و تلاش برای طراحی‌های مرتّب و زیبا» ترجمه کنیم، بنظر می‌رسد توصیه‌ای بسیار به جا و منطقی است. امّا اگر آن را به «هر کاری می‌توانید انجام دهید تا رنگ فاقد داده‌ها را حذف کنید» ترجمه کنیم، این باعث طراحی‌هایی با کیفیت کم می‌شود. اگر در هر یک از این دو مدل افراط شود، نتیجه حاصل نمودارهای خوبی نخواهد بود. هرچند که به جز موارد افراطی، محدودهٔ وسیعی از طراحی‌ها قابل قبول وجود دارد و ممکن است در زمینه‌های مختلف مناسب و قابل استفاده باشند.

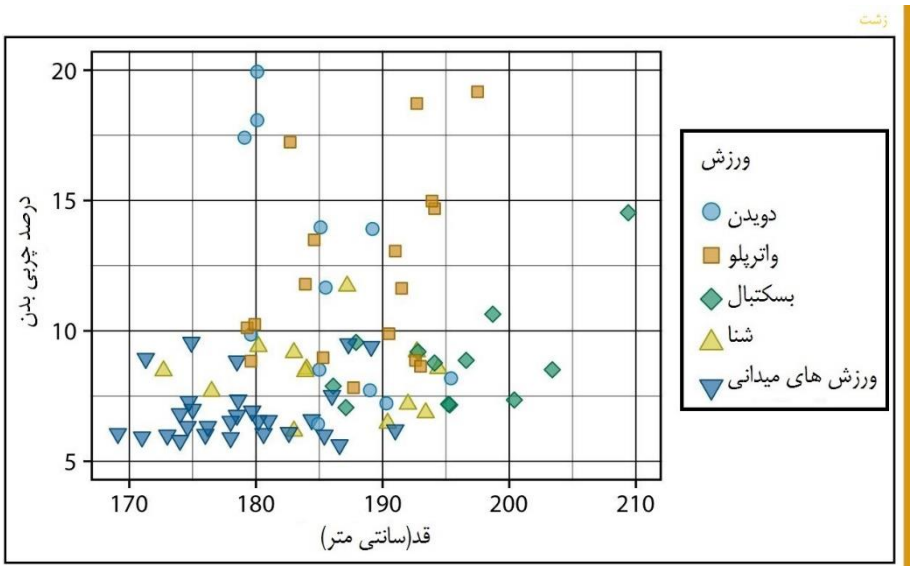
برای بررسی موارد افراطی شکلی را در نظر بگیرید که مقادیر بسیار زیادی رنگ فاقد داده دارد (شکل ۲۳-۱). نقاط رنگی روی صفحه (ناحیهٔ مرکزی که شامل نقطه داده‌ها است) رنگ داده‌ها هستند. همه چیزهای دیگر رنگ فاقد داده است. رنگ فاقد داده‌ها شامل یک قاب در اطراف شکل، یک قاب در اطراف صفحه نمودار و یک قاب در اطراف راهنما می‌شود. به هیچ یک از این قاب‌ها نیازی نیست. همچنین یک شبکه برجسته و پس زمینه‌ای متراکم می‌بینیم که توجّه را از نقطه‌های داده واقعی دور می‌کند. با حذف قاب‌ها و خطوط شبکه فرعی و رسم خطوط شبکه اصلی به رنگ خاکستری روشن، به شکل ۲۳-۲ می‌رسیم. در این نسخه از تصویر، نقاط داده بسیار برجسته‌تر هستند و به عنوان مهمترین اجزای نمودار شناخته می‌شوند.

در مثال دیگری از موارد افراطی ممکن است به شکلی مثل شکل ۳-۲۳ برسیم که نسخه کمینه‌ای از شکل ۲۳-۲ می‌باشد. در این شکل برچسب‌های محورها و عنوان‌ها بسیار محو شده‌اند و به سختی قابل دیدن هستند. اگر یک نگاه کلی به نمودار داشته باشیم متوجّه نخواهیم شد که چه اطلاعاتی نمایش داده شده‌اند. فقط نقاطی را خواهیم دید که در فضا معلّق هستند. فراتر از آن، توضیحات راهنما بسیار کمرنگ است تا حدی که نقاط آن‌ها ممکن است

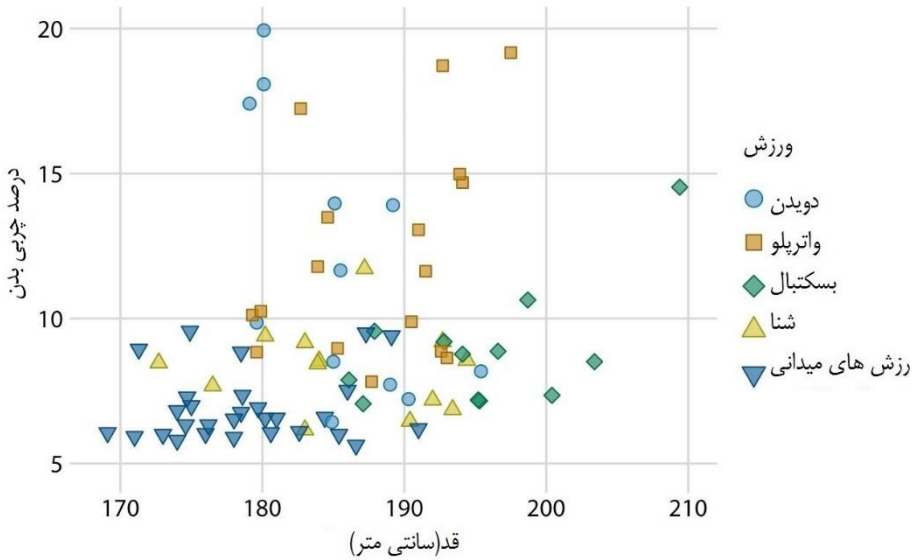
1. minimal

با نقاط داده اشتباه گرفته شوند. این مساله به دلیل اینکه هیچ جداسازی‌ای بین راهنما و ناحیه نمودار انجام نگرفته، تشدید شده است. در شکل ۲۳-۲ به این نکته توجّه کنید که چطور شبکه، هم نقاطی که در فضا هستند را فرا گرفته است و هم مرز بین نمودار و راهنما را جدا می‌کند. هیچکدام از این نکات در شکل ۲۳-۳ مشاهده نمی‌شوند.

در شکل ۲۳-۲ از یک شبکهٔ بدون پس زمینه استفاده شده و محورهای مختصّات فاقد خطوط بوده و کادری هم در اطراف ناحیه نمودار وجود ندارد. این طراحی به بیننده این مفهوم را القا می‌کند که دامنه داده‌های ممکن می‌تواند فراتر از محدوده محورهای مختصّات باشد. مثلاً گرچه شکل ۲۳-۲ هیچ ورزشکاری با قد بلندتر از ۲۱۰ سانتی‌متر را نشان نمی‌دهد، در حقیقت چنین فردی می‌تواند وجود داشته باشد. با این حال برخی نویسندگان علاقه‌مند هستند که محدودهٔ ناحیه نمودار را با کشیدن کادری در اطراف آن مشخّص کنند (شکل ۲۳-۴). هر دو این انتخاب‌ها منطقی هستند و اینکه کدام مناسب‌تر است، بیشتر نظر شخصی است. از مزیّت‌های مدل دارای کادر این است که راهنما را از ناحیه نمودار جدا می‌کند.

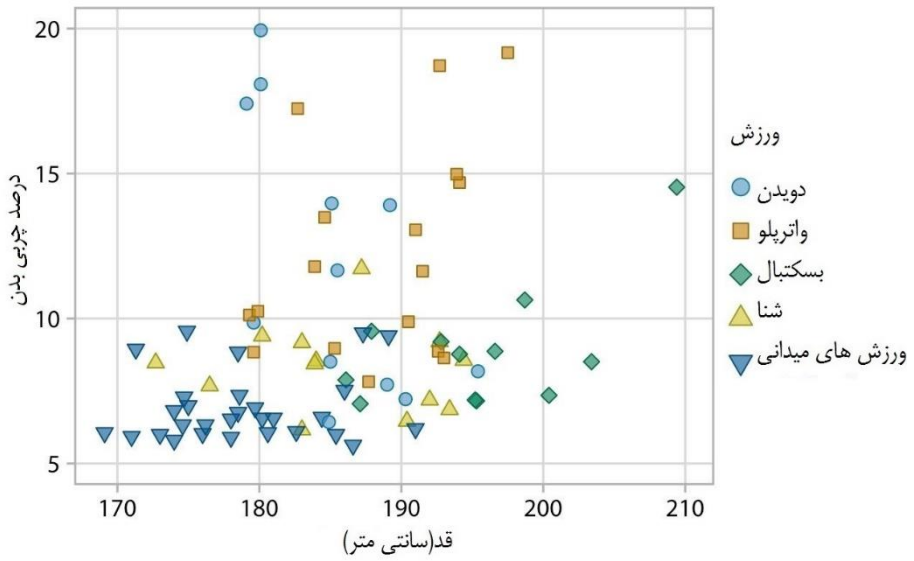


شکل ۲۳-۱. درصد چربی بدن در مقابل قد در مردان ورزشکار حرفه‌ای استرالیایی. هر نقطه نشان‌دهندهٔ یک نفر است. این شکل رنگ فاقد داده زیادی دارد. قاب‌های غیرضروری در اطراف شکل، اطراف نمودار و اطراف راهنما وجود دارد. خطوط شبکه بسیار برجسته هستند و توجّه را از نقاط داده منحرف می‌کنند. منبع داده: Telford and Cunningham 1991.



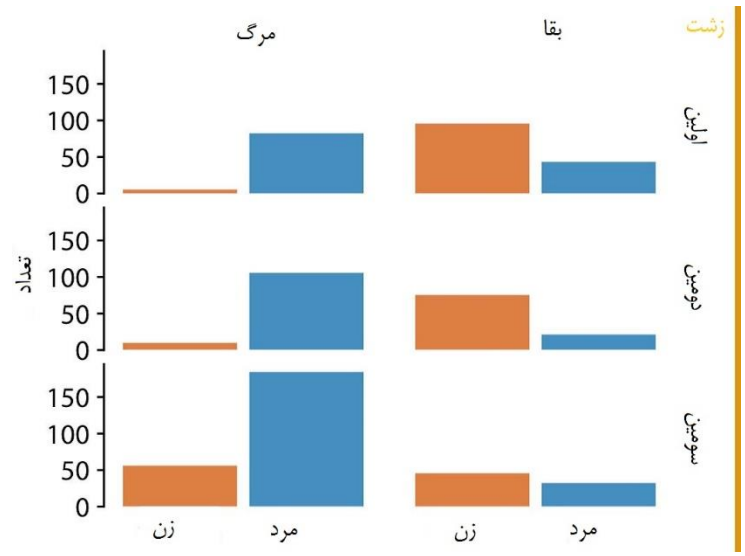
شکل ۲۳–۲. درصد چربی بدن در مقابل با قد در مردان ورزشکار حرفه‌ای استرالیایی. این تصویر نسخهٔ اصلاح شده‌ای از شکل ۲۳–۱ است. قاب‌های غیرضروری و شبکه‌های فرعی حذف شدند و خطوط شبکه‌های اصلی به رنگ خاکستری روشن ترسیم شدند. منبع داده: Telford and Cunningham 1991.

شکل ۲۳–۳. درصد چربی بدن در مقابل قد در مردان ورزشکار حرفه‌ای استرالیایی. در این مثال مفهوم حذف رنگ‌های فاقد داده بیش از حد اعمال شده است. برچسب و عنوان محورها بسیار کمرنگ هستند تا حدی که به ندرت قابل مشاهده هستند. به نظر می‌رسد نقاط در فضا معلّق هستند. نقاط مربوط به راهنما به اندازه کافی از نقاط داده‌ها جداسازی نشده‌اند و بیننده ممکن است تصوّر کند که بخشی از داده هستند. منبع داده: Telford and Cunningham 1991.



شکل ۲۳–۴. درصد چربی بدن در مقابل قد در مردان ورزشکار حرفه‌ای استرالیایی. این شکل یک کادر در اطراف نمودار ۲۳–۳ اضافه کرده است و این قاب به جداسازی راهنما از داده‌ها کمک می‌کند.

نقطه ضعف تصاویری که رنگ فاقد داده کمی دارند این است که عناصر این شکل‌ها به نظر در فضا معلق هستند، و فاقد ارتباط یا ارجاع مشخّص می‌باشند. این مشکل به طور کلی در نمودار چندگانه‌های کوچک شدیدتر است. شکل ۲۳–۵ یک نمودار چندگانهٔ کوچک است که ۶ نمودار میله‌ای متفاوت را مقایسه می‌کند، امّا بیشتر شبیه به یک اثر هنری مدرن است تا یک تصویرسازی مفید از داده‌ها. میله‌ها به خط مبدأ وصل نشده‌اند و محدودهٔ هیچ یک از سطوح نمودار به خوبی مشخّص نشده است. می توانیم این مشکلات را با اضافه کردن پس‌زمینه‌ای به‌رنگ خاکستری روشن و خطوط شبکه افقی باریک به هر محور برطرف کنیم (شکل ۲۳–۶).



شکل ۲۳–۵. مسافران نجات یافته در تایتانیک، بر اساس جنسیّت و طبقه اقتصادی. این نمودار چندگانه‌های کوچک بسیار مینیمالیستی هستند. محورها کادر ندارند و بنابراین تعیین اینکه کدام بخش از شکل مربوط به کدام محور است، دشوار است. فراتر از آن، ستون‌ها به هیچ خط پایه‌ای متصل نشده‌اند و به نظر معلّق هستند. منبع داده: دایره المعارف تایتانیکا

شکل ۲۳–۶. مسافران نجات یافتهٔ تایتانیک بر اساس جنسیت و طبقه اقتصادی. این شکل نسخهٔ تصحیح شده‌ای از شکل ۲۳–۵ است. زمینهٔ خاکستری در هر قسمت به طور واضح ۶ گروهی که این نمودار را تشکیل می‌دهد (نجات یافته یا فوت کرده در طبقه‌های اقتصادی اوّل، دوّم یا سوّم)، از هم جدا کرده است. خطوط افقی باریک در پس زمینه مرجع خوبی برای طول میله‌ها هستند و مقایسهٔ ارتفاع میله‌ها را در قسمت‌های مختلف نمودار تسهیل می‌کند. همچنین می‌توانستیم یک قاب در اطراف هر نمودار بکشیم و از میله‌های خاکستری برای برجسته کردن متغیّرهای گروه‌ها استفاده کنیم (شکل ۲۱–۱). منبع داده ها: دایره المعارف تایتانیکا

شبکهٔ پس زمینه

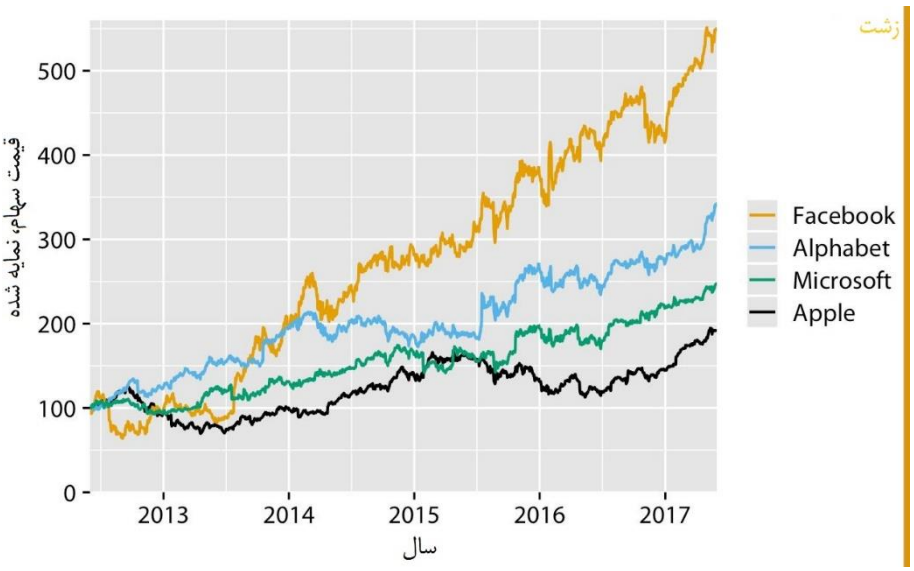
خطوط شبکه‌ای به خواننده کمک می‌کند تا به ارزش داده‌ها پی برده و بتواند آن‌ها را در یک قسمت از نمودار با قسمتی دیگر مقایسه کند. در عین حال خطوط شبکه می‌توانند اختلال بصری ایجاد کنند به خصوص زمانی که قطور باشند یا به شکل متراکمی کنار هم قرار گرفته باشند. نظر افراد مختلف در خصوص استفاده از خطوط شبکه ممکن است متفاوت باشد و در صورت استفاده، نحوهٔ به کارگیری آن‌ها از نظر قطر و تراکم نیز باید مدنظر باشد. در این کتاب از انواع مختلفی از خطوط شبکه استفاده نموده‌ایم تا تاکید کنیم که فقط یک گزینهٔ ایده‌آل وجود ندارد.

نرم‌افزار R ggplot۲ سبکی را معرفی کرده که از شبکه‌های نازک با خطوط سفید در زمینه خاکستری تشکیل شده است. شکل ۲۳–۷ مثالی از این مدل است. این شکل تغییر ارزش سهام ۴ شرکت بزرگ فناوری را در بازه‌ای ۵ ساله از ۲۰۱۲ تا ۲۰۱۷ نشان می‌دهد. با عذرخواهی از نویسندهٔ ggplot۲ هادلی ویکهام[1] که احترام زیادی برای او قایل هستیم، این شبکهٔ خطوط سفید روی زمینهٔ خاکستری هیچ جذابیّتی ندارد. به نظر می‌رسد پس زمینه خاکستری باعث می‌شود داده‌های نمودار جلوهٔ کمتری داشته باشند و شبکه خطوط اصلی و فرعی بسیار متراکم هستند. همچنین مربع‌های خاکستری در قسمت راهنما گیج‌کننده می‌باشند.

استدلال‌هایی که از زمینهٔ خاکستری دفاع می‌کنند شامل این است که آن‌ها هم باعث می‌شوند که نمودار به عنوان یک جزء بصری واحد باشد و هم اینکه باعث می‌شود نمودار به شکل یک جعبهٔ سفید احاطه شده با متن مشکی نباشد.[Wickham 2016] ما کاملاً با بخش اول موافقیم و به همین دلیل بود که در شکل ۲۳–۶ از پس زمینهٔ خاکستری استفاده کردیم. در مورد قسمت دوّم باید احتیاط کنیم که درک رنگ تیره از متن به نوع و اندازه قلم، و فواصل بین خطوط بستگی دارد و درک رنگ تیره از شکل به مقدار و رنگ جوهری که استفاده شده از جمله رنگ داده‌ها بستگی دارد. حروف‌چینی یک مقاله علمی به صورت فشرده و قلم Times New Roman و اندازه ۱۰ بسیار تیره‌تر از حروف‌چینی یک کتاب عمومی با قلم Palatino و اندازه ۱۴ و فاصله یک و نیم خط می‌باشد. به همین صورت یک نمودار پراکنش شامل ۵ نقطه داده با رنگ زرد، بسیار روشن‌تر از یک نمودار پراکنش با ۱۰،۰۰۰ نقطه داده با رنگ مشکی است. اگر می‌خواهید رنگ خاکستری در پس زمینه به کار ببرید، شدّت رنگ عناصر پیش

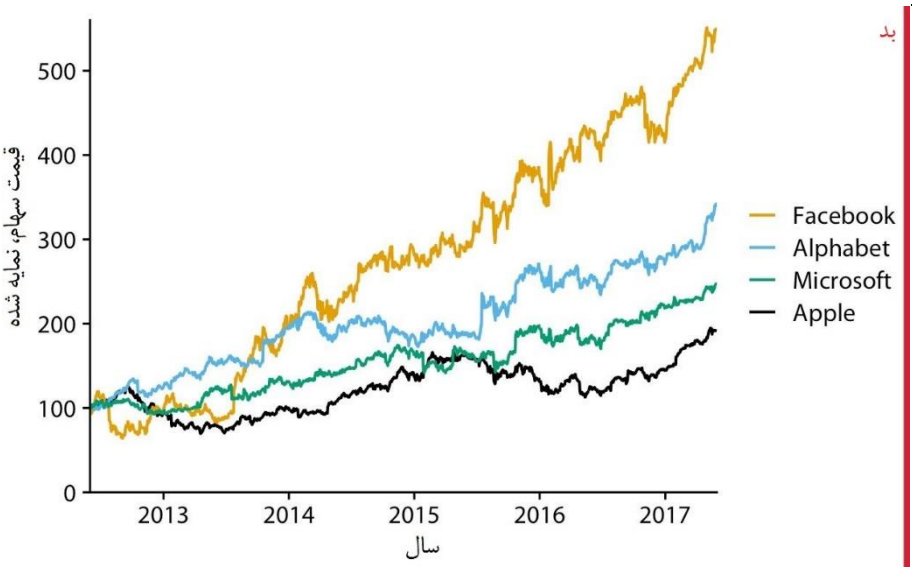
1. Hadley Wickham

زمینه، صفحه‌آرایی متنی که در اطراف شکل است را مدنظر داشته باشید و بر طبق آن رنگ پس زمینه را انتخاب و تنظیم کنید. در غیر این صورت نمودار به شکل یک جعبهٔ تیره خواهد شد که در میان متن روشنی احاطه شده است. همچنین در نظر داشته باشید رنگ‌هایی که در نمودار استفاده می‌کنید باید با پس‌زمینه همخوانی داشته باشند. برداشت ما از رنگ‌ها در پس‌زمینه‌های مختلف متفاوت است و یک پس‌زمینهٔ خاکستری نسبت به یک پس‌زمینهٔ سفید، نیازمند رنگ‌های تیره‌تری در پیش‌زمینه است.



شکل ۲۳–۷. ارزش سهام در طول زمان در ۴ شرکت بزرگ فناوری. ارزش سهام برای هر شرکت معادل عدد ۱۰۰ در ماه ژوئن سال ۲۰۱۲ استاندارد شده است. این تصویر مشابه پیش فرض نمودار ggplot ۲ است و دارای خطوط شبکه‌ای اصلی و فرعی در پس زمینهٔ خاکستری می‌باشد. در این مثال خاص، به نظر می‌رسد خطوط شبکه بر خطوط داده غلبه نموده‌اند و نتیجهٔ حاصل نموداری غیرمتعادل است که تأکید کافی بر داده‌ها ندارد. منبع داده: امور مالی یاهو

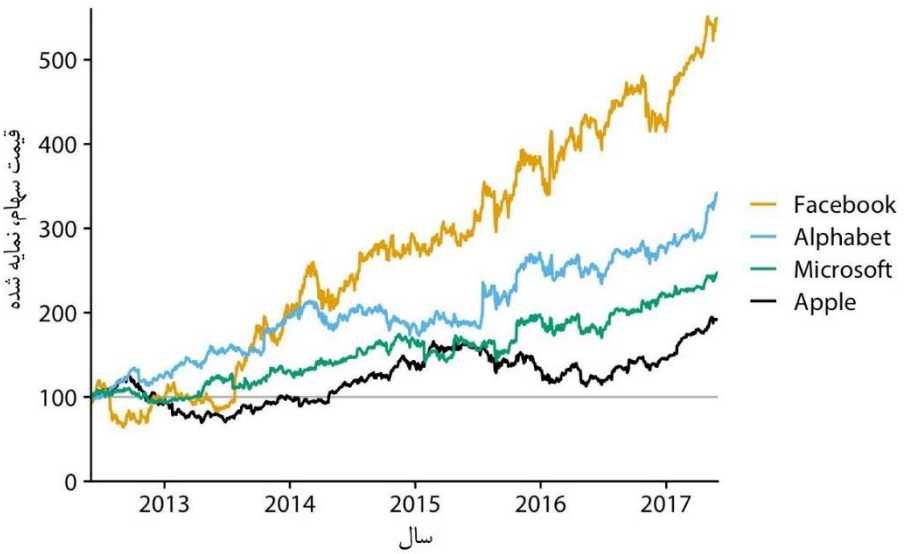
می‌توانیم تمام این مراحل را برعکس کرده و هم رنگ پس‌زمینه و هم خطوط شبکه را حذف کنیم (شکل ۲۳–۸). در این صورت باید خطوط محورها واضح باشند تا نمودار قالب خود را داشته و به عنوان یک واحد بصری مشاهده شود. برای این نمودار خاص، به نظر این گزینه اصلاً مناسب نیست و لذا به صورت «بد» برچسب‌گذاری شده است. در نبود هرگونه خطوط شبکه پس‌زمینه، گویی خطوط منحنی داده در فضا معلّق هستند. همچنین انتساب مقادیر انتهایی داده در قسمت راست نمودار مبتنی بر مقادیر مرجع در سمت چپ دشوار است.



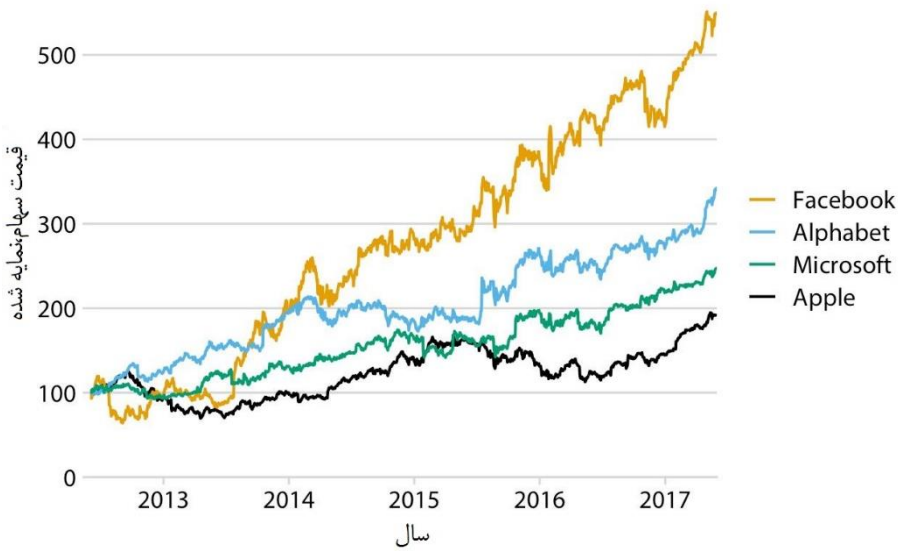
شکل ۲۳–۸. ارزش سهام ۴ شرکت بزرگ فناوری در طول زمان. در این نسخه از نمودار ۲۳–۷ خطوط داده‌ها نقطهٔ اتکایی مناسبی ندارند. در نتیجه درک اینکه در انتهای بازهٔ زمانی چقدر از مقدار مرجع ۱۰۰ فاصله گرفته‌اند دشوار است. منبع داده: امور مالی یاهو

حداقل کار این است که یک خط مرجع افقی اضافه شود. از آنجا که ارزش سهام در شکل ۲۳–۸ در ژوئن ۲۰۱۲ معادل ۱۰۰ در نظر گرفته شده، مشخّص کردن این عدد با یک خط افقی در y=۱۰۰ بسیار کمک‌کننده است (شکل ۲۳–۹). همچنین می‌توانیم از خطوط شبکه افقی کمینه‌ای استفاده کنیم. در نموداری که تغییرات در محور yها حائز اهمیت است، نیازی به خطوط عمودی نیست. خطوط شبکه اصلی اغلب کافی هستند، و خط محور می‌تواند حذف شده و یا بسیار باریک شود چون خطوط افقی حدود نمودار را تعیین می‌کنند (شکل ۲۳–۱۰).

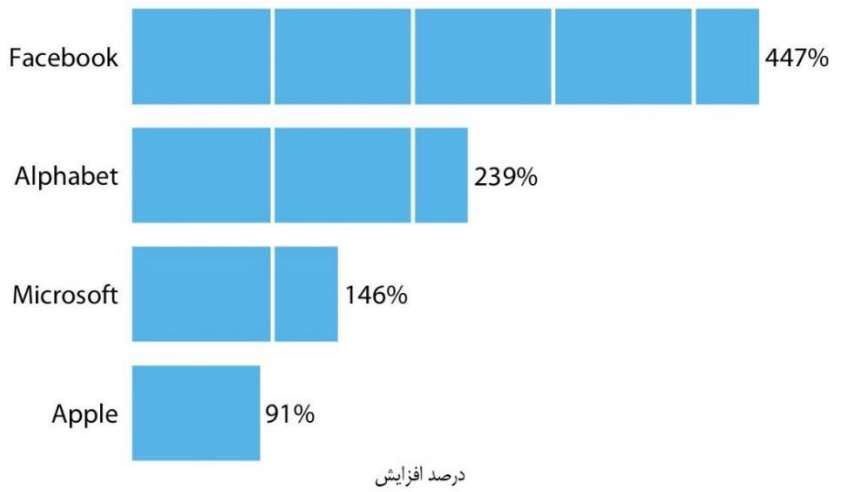
برای چنین شبکه‌های کمینه‌ای، معمولاً خطوط را به شکل عمود در جهتی که مقادیر مورد نظر تغییر می‌کنند رسم می‌کنیم. به همین دلیل اگر به جای آنکه ارزش سهام را در گذر زمان روی نمودار رسم کنیم، افزایش ۵ ساله را روی نمودار با میله‌های افقی رسم کنیم، باید از خطوط عمودی استفاده کنیم (شکل ۲۳–۱۱).



شکل ۲۳–۹. ارزش سهام ۴ شرکت بزرگ فناوری در طول زمان. افزودن یک خط افقی نازک در مقدار شاخص ۱۰۰ به شکل ۲۳–۸ به ایجاد یک مرجع مهم در طول کل بازه زمانی نمودار کمک می‌کند. منبع داده: امور مالی یاهو



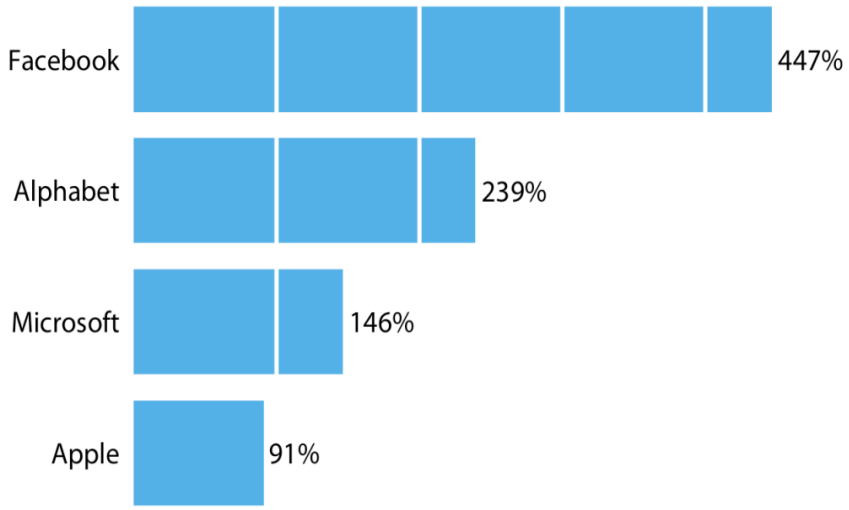
شکل ۲۳–۱۰. ارزش سهام ۴ شرکت بزرگ فناوری در طول زمان. افزودن خطوط افقی باریک در تمام تیک‌های اصلی محور y خطوط مرجع بهتری نسبت به اینکه فقط یک خط افقی مانند شکل ۲۳–۹ داشته باشیم فراهم می‌کند. همچنین این طراحی نیاز خطوط برجسته محوری y و x را برطرف می‌کند زیرا خطوط افقی متوازن قابی برای منطقه نمودار ایجاد می‌کنند. منبع داده: امور مالی یاهو



شکل ۲۳–۱۱. درصد افزایش در ارزش سهام از ماه ژوئن سال ۲۰۱۲ تا ژوئن ۲۰۱۷ برای ۴ شرکت بزرگ فناوری. از آنجایی که میله‌ها افقی هستند، خطوط شبکه عمودی مناسب هستند. منبع داده: امور مالی یاهو

خطوط شبکه‌ای که عمود بر متغیّر مورد مطالعه رسم می‌شوند، بیشترین کارایی را دارند.

برای نمودارهای میله‌ای مثل شکل ۲۳–۱۱، توفت پیشنهاد می‌دهد خطوط شبکه سفید روی میله‌ها رسم شود، بجای آنکه خطوط تیره زیر آن‌ها رسم شود [Tufte 2001] این خطوط سفید میله‌ها را به اجزایی با طول یکسان تقسیم می‌کند (شکل ۲۳–۱۲). ما در مورد این سبک دو دل هستیم. از یک طرف مطالعه روی ادراک انسان‌ها نشان می‌دهد که شکستن میله‌ها به اجزای جدا از هم به بیننده در درک طول میله‌ها کمک می‌کند. [Haroz, Kosara, and Franconeri 2015] از طرف دیگر به نظر اینطور است که میله‌ها از هم جدا هستند و یک واحد بصری را تشکیل نمی‌دهند. در حقیقت ما این سبک را تعمّداً در شکل ۶–۱۰ استفاده کردیم تا میله‌هایی را که نشان‌دهندهٔ مسافران زن و مرد بود را از هم جدا کنیم. اینکه کدام اثر غالب خواهد بود بستگی به عرض میله‌ها، فاصلهٔ بین میله‌ها و ضخامت خطوط سفید دارد. در نتیجه اگر قصد استفاده از این سبک را دارید پیشنهاد می‌شود که آنقدر این متغیّرها را تغییر دهید تا به شکل دلخواهتان از نظر بصری برسید.



شکل ۲۳–۱۲. درصد افزایش ارزش سهام از ژوئن ۲۰۱۲ تا ژوئن ۲۰۱۷ برای ۴ شرکت بزرگ فناوری. خطوط شبکه سفید روی میله‌ها به خواننده کمک می‌کند تا طول نسبی میله‌ها را درک کند. به طور همزمان این خطوط می‌تواند این مفهوم را القا کنند که میله‌ها در حال جدا شدن هستند. منبع داده: امور مالی یاهو

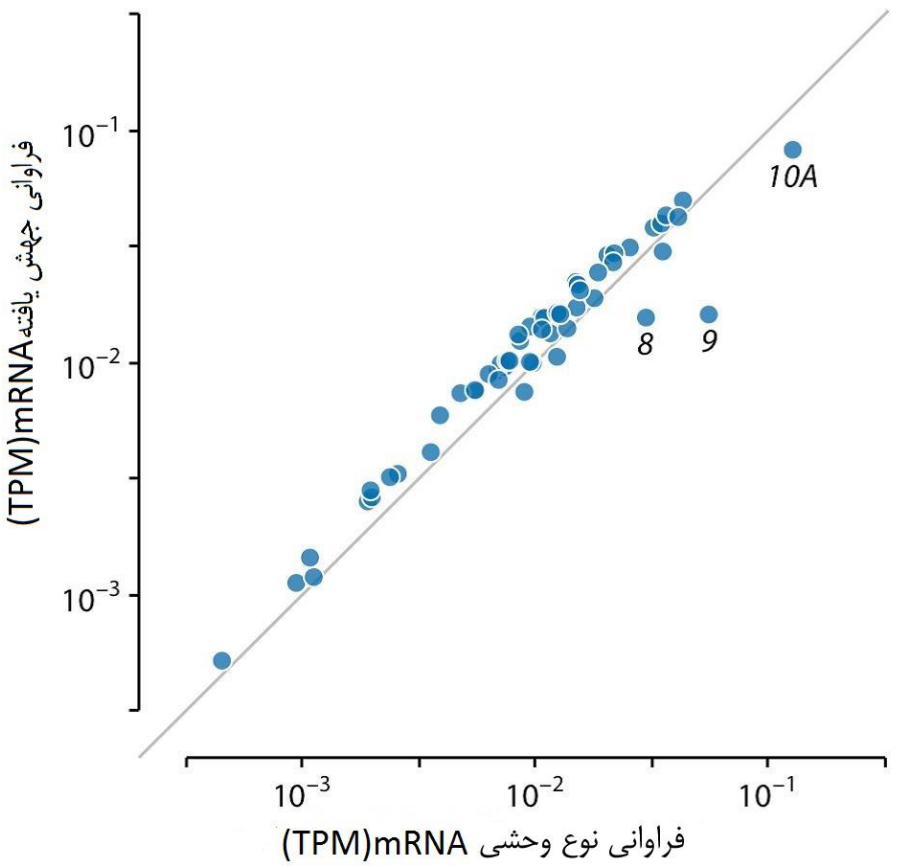
یک مشکل دیگر شکل ۲۳–۱۲ آن است که مجبور شدیم درصد مقادیر را خارج از میله‌ها قرار دهیم، زیرا برچسب‌ها در قسمت انتهایی بسیاری از میله‌ها جای نمی‌گرفتند. این کار از نظر دیداری باعث افزایش طول میله‌های نمودار شده و باید در صورت امکان از آن پرهیز شود.

برای نمودارهای پراکنش که هیچ محور ترجیحی ندارند، رسم شبکه‌های پس‌زمینه در هر دو جهت کمک‌کننده خواهند بود. شکل ۲۳–۲ در آغاز این فصل یک نمونه از آن است. وقتی یک نمودار دارای شبکهٔ پس‌زمینه کامل است، عموماً نیازی به رسم خطوط محورها نیست.

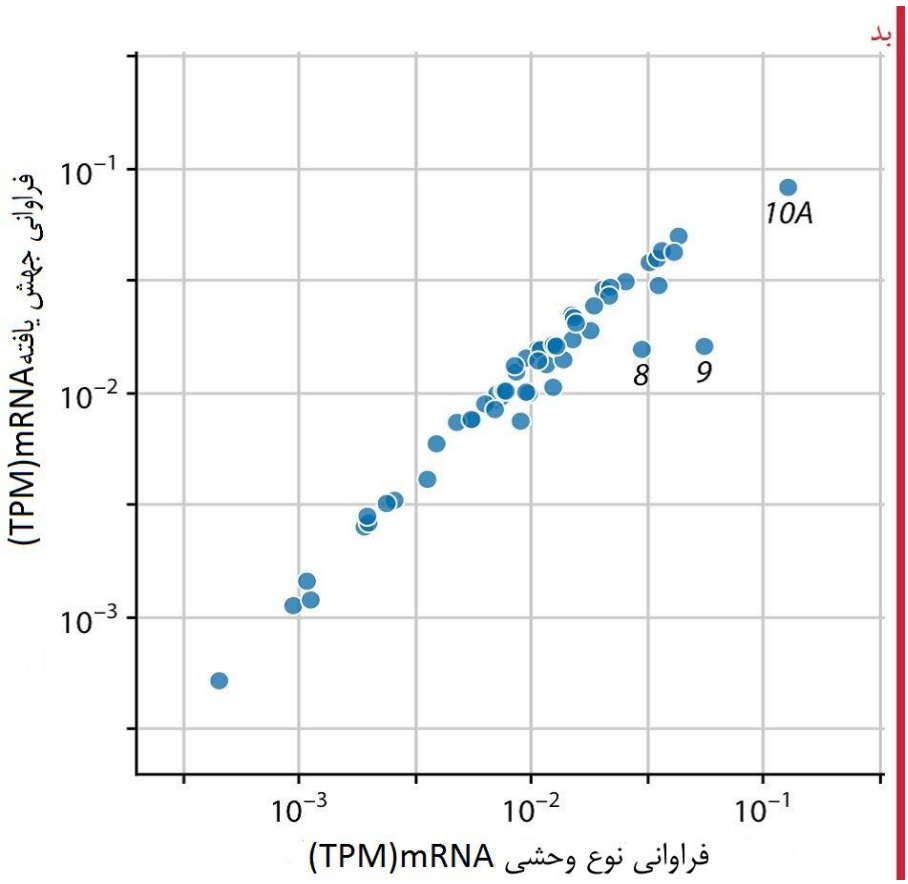
داده‌های زوجی

در شکل‌هایی که مبنای مقایسهٔ اطّلاعات خط x=y است، مثل نمودارهای پراکنش برای داده‌های زوجی، ترجیح بر این است به جای شبکه از یک خط مورّب استفاده کنیم. مثلاً شکل ۲۳–۱۳ را در نظر بگیرید که سطوح بیان ژن را در یک ویروس جهش یافته با یک گونهٔ جهش نیافته (گونهٔ وحشی) مقایسه می‌کند. خط مورّب به ما این امکان را می‌دهد تا سریع متوجّه شویم چه ژن‌هایی در گونهٔ جهش یافته نسبت به گونهٔ جهش نیافته، بیشتر یا کمتر بیان شده‌اند. چنین قضاوتی وقتی شکل دارای پس‌زمینهٔ شبکه‌ای و فاقد خط مورّب باشد، مشکل

است (شکل ۲۳–۱۴). بنابراین اگرچه شکل ۲۳–۱۴ زیبا به نظر می‌آید، به صورت «بد» برچسب‌گذاری شده است. به طور مشخّص سطح بیان ژن A۱۰ که در گونهٔ جهش یافته نسبت به گونهٔ جهش نیافته به طور قابل ملاحظه‌ای کاهش یافته (شکل ۲۳–۱۳)، از نظر بصری در شکل ۲۳–۱۴ واضح نیست.

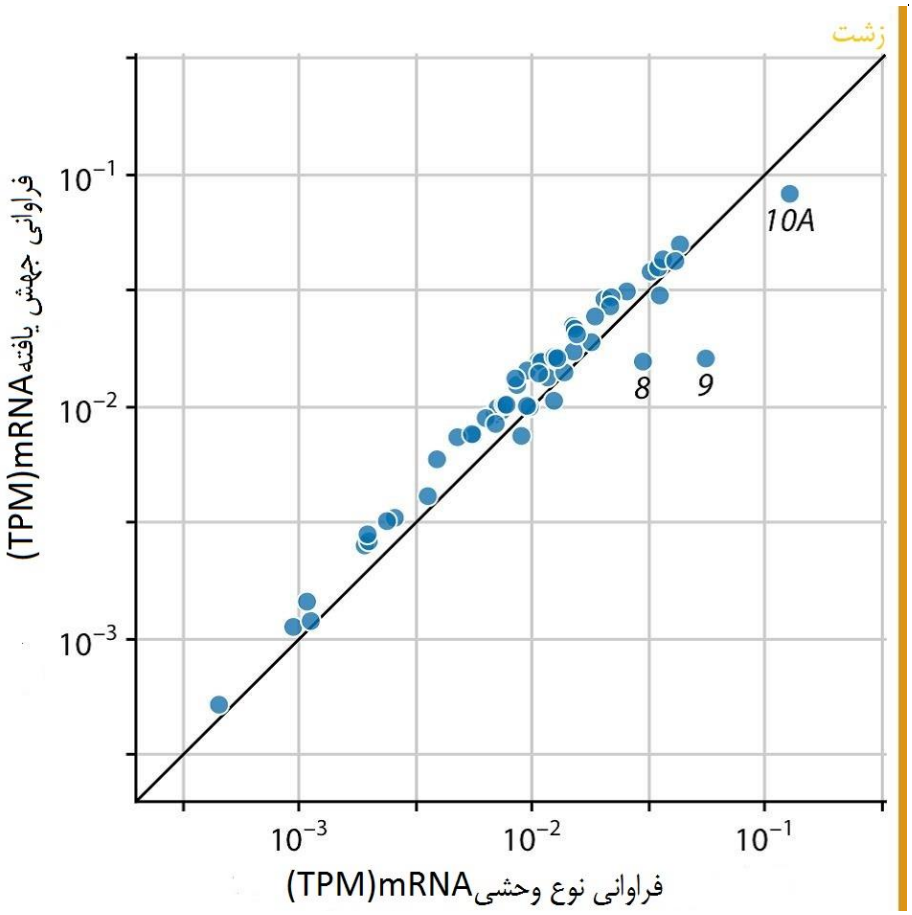


شکل ۲۳–۱۳. سطوح بیان ژن در باکتریوفاژ T7جهش یافته نسبت به گونهٔ وحشی. سطوح بیان ژن توسّط فراوانی mRNA در واحد رونویسی در هر میلیون TPM اندازه‌گیری شده‌اند. هر نقطه نمایندهٔ یک ژن است. در گونهٔ جهش یافته پروموتور ابتدای ژن شماره ۹ حذف شده و این باعث کاهش مقادیر mRNA ناشی از ژن شماره ۹ و ژن‌های مجاور یعنی ۸ و A۱۰ شده است. منبع داده: Paff et al. 2018



شکل ۲۳–۱۴. سطوح بیان ژن ار یک باکتریوفاژ جهش یافته T۷ در مقایسه با گونهٔ وحشی. با ترسیم این داده‌ها روی شبکه به جای استفاده از خط مورّب، اینکه کدام ژن‌ها در گونهٔ جهش یافته نسبت به گونهٔ وحشی بیشتر یا کمتر هستند، مبهم می‌شود. منبع داده: Paff et al. 2018.

البته می‌توانستیم خط مورّب در شکل ۲۳–۱۳ را روی پس‌زمینه شبکه‌ای شکل ۲۳–۱۴ اضافه کنیم تا مطمئن شویم که مبنای مقایسه‌ای وجود دارد. امّا شکل نهایی کمی شلوغ می‌شود (شکل ۲۳–۱۵). ما مجبور شدیم خط مورّب را تیره‌تر کنیم تا مشخّص‌تر از پس‌زمینه شبکه‌ای باشد. اما در حال حاضر به نظر می‌آید که نقاط داده‌ها در پس‌زمینه محو شده‌اند. می‌توانیم این مشکل را از طریق بزرگ‌تر یا تیره‌تر کردن نقاط داده‌ها برطرف کنیم. اما با در نظر گرفتن همهٔ این‌ها ما شکل ۲۳–۱۳ را ترجیح می‌دهیم.



شکل ۲۳–۱۵. سطوح بیان ژن در گونهٔ جهش یافته باکتریوفاژ T7 در مقایسه با گونهٔ وحشی. این شکل خطوط شبکه پس‌زمینه شکل ۲۳–۱۴ را با خط مورّب شکل ۲۳–۱۳ ترکیب نموده است. به نظر این شکل از نظر ظاهری نسبت به شکل ۲۳–۱۳ بسیار شلوغ‌تر است و ما شکل ۲۳–۱۳ را ترجیح می‌دهیم. منبع داده: Paff et al. 2018

خلاصه

هم اشباع کردن شکل با رنگ‌های فاقد داده و هم پاک کردن افراطی آن‌ها می‌تواند منجر به طراحی ضعیف نمودار شود. باید یک حد متوسط را در این میان پیدا کرد، جایی که بیشترین تاکید نمودار روی نقاط داده باشد. در حالی که زمینهٔ کافی در مورد اینکه چه داده‌ای در حال نمایش است، موقعیّت نسبی نقاط نسبت به یکدیگر به چه صورت است و معنی آن‌ها چیست، ارائه شود.

در خصوص پس‌زمینه و شبکه پس‌زمینه، هیچ تک گزینه‌ای مناسب تمام شرایط نیست. پیشنهاد این است که از شبکهٔ خطوط عاقلانه استفاده کنید. با دقّت در نظر بگیرید که کدام یک از شبکه‌ها یا خطوط راهنما بیشترین اطّلاعات را در نمودار شما نشان خواهد داد و سپس فقط از آن‌ها استفاده کنید. از آنجایی که سفید، رنگی خنثی و پیش فرض کاغذ است و تقریباً از هر رنگی روی پیش‌زمینه پشتیبانی می‌کند، توصیهٔ ما استفاده از شبکه‌های کمینه‌ای با رنگ روشن در پس‌زمینه سفید است. با این حال یک پس‌زمینه سایه‌دار کمک می‌کند تا نمودار به صورت یک مجموعهٔ جداگانه ظاهر شود و این مساله مخصوصاً در نمودار چندگانه‌های کوچک کمک‌کننده است. در آخر باید بدانیم چطور همه این انتخاب‌ها به هویّت بخشی بصری اثر کمک می‌کند. بسیاری از مجلات و وبسایت‌ها تمایل دارند سبکی اختصاصی داشته باشند که سریعاً از طرف خواننده شناسایی شود و وجود پس‌زمینه سایه‌دار و انتخاب خاص از شبکه‌های پس‌زمینه می‌تواند به ساختن یک هویّت خاص بصری کمک کند.

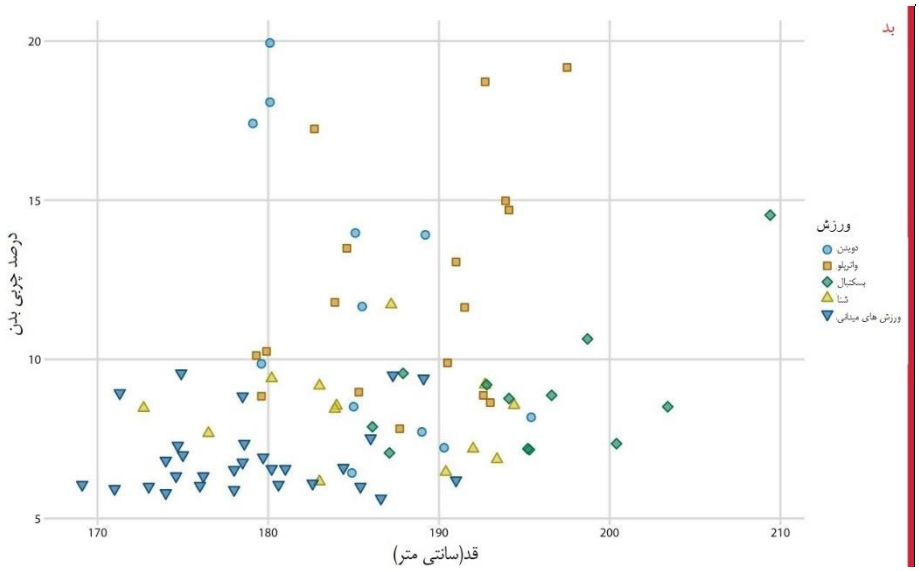
فصل ۲۴

استفاده از برچسب‌های بزرگ‌تر برای محورها

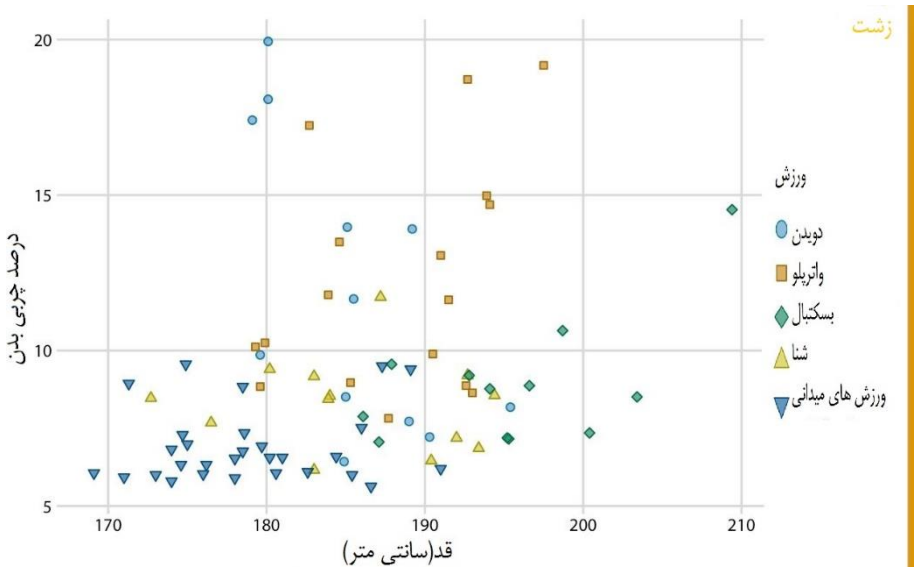
اگر فقط یک درس را از این کتاب به یاد می‌سپارید، آن درس باید این باشد: به برچسب‌های محور، برچسب‌های تیک محورها و سایر حاشیه‌نویسی‌های مختلف نمودار توجّه کنید. به احتمال زیاد آن‌ها خیلی کوچک هستند. تقریباً همه نرم‌افزارهای ترسیم نمودار و کتابخانه‌های نموداری، پیش‌فرض‌های ضعیفی دارند. اگر از مقادیر پیش‌فرض استفاده می‌کنید، به احتمال زیاد انتخاب ضعیفی دارید.

به عنوان مثال، شکل ۲۴-۱ را در نظر بگیرید. ما همیشه چنین نمودارهایی را می‌بینیم. برچسب‌های محور، برچسب‌های تیک محور و برچسب‌های راهنما همگی بسیار کوچک هستند. ما به سختی می‌توانیم آن‌ها را ببینیم، و ممکن است مجبور باشیم برای خواندن حاشیه‌نویسی‌های علائم و اختصارات، صفحه را بزرگ‌نمایی کنیم.

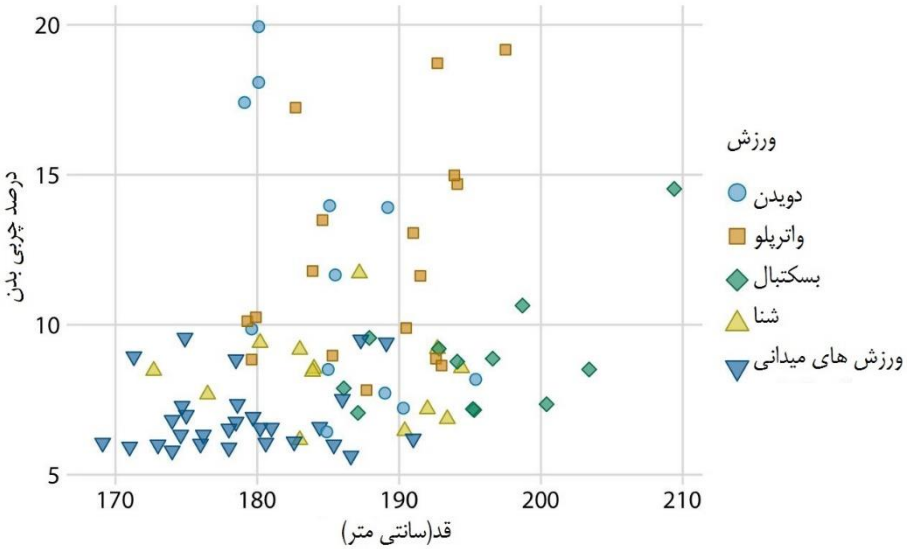
یک نسخه تا حدودی بهتر از این شکل به صورت شکل ۲۴-۲ خواهد بود. اندازهٔ قلم‌ها هنوز خیلی کوچک هستند، و به همین دلیل است که این شکل را «زشت» نامیده‌ایم. با این حال، ما در مسیر درستی حرکت می‌کنیم. این نمودار ممکن است تحت شرایطی قابل قبول باشد. انتقاد اصلی در اینجا خوانا نبودن برچسب‌ها نیست، بلکه متوازن نبودن شکل است. عناصر متن در مقایسه با بقیهٔ شکل بسیار کوچک هستند.



شکل ۲۴–۱. درصد چربی بدن در مقابل قد در ورزشکاران حرفه‌ای مرد استرالیایی. (هر نقطه نشان‌دهندهٔ یک ورزشکار است) مشکل این شکل این است که عناصر متن بسیار کوچک هستند و به سختی خوانده می‌شوند. منبع داده: Telford and Cunningham 1991



شکل ۲۴–۲. درصد چربی بدن در مقابل قد در ورزشکاران مرد. این شکل نسبت به شکل ۲۴–۱ بهبود یافته است، امّا عناصر متن همچنان بسیار کوچک هستند و شکل متعادل نیست. منبع داده: Telford and Cunningham 1991

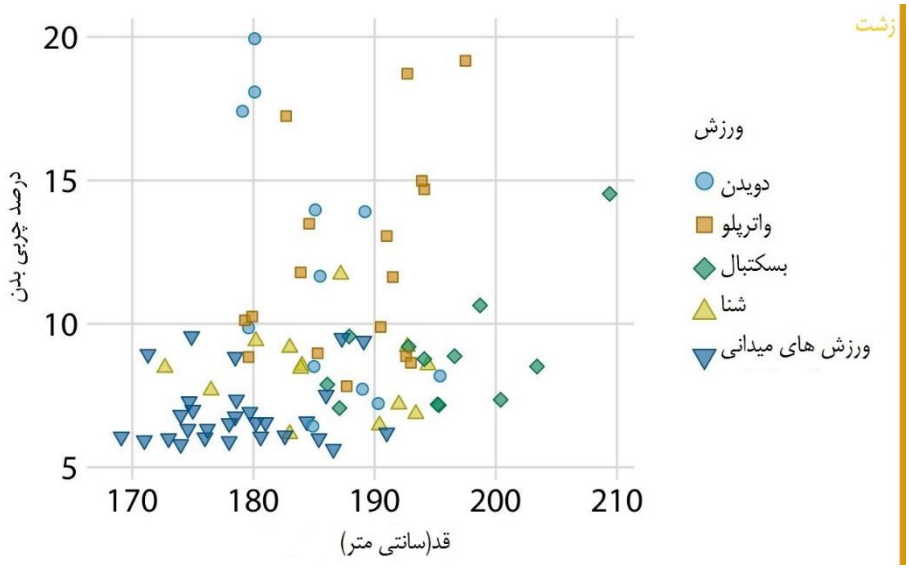


شکل ۲۴–۳. درصد چربی بدن در مقابل قد در ورزشکاران مرد. مقیاس همه عناصر شکل متناسب شده است.
منبع داده: Telford and Cunningham 1991

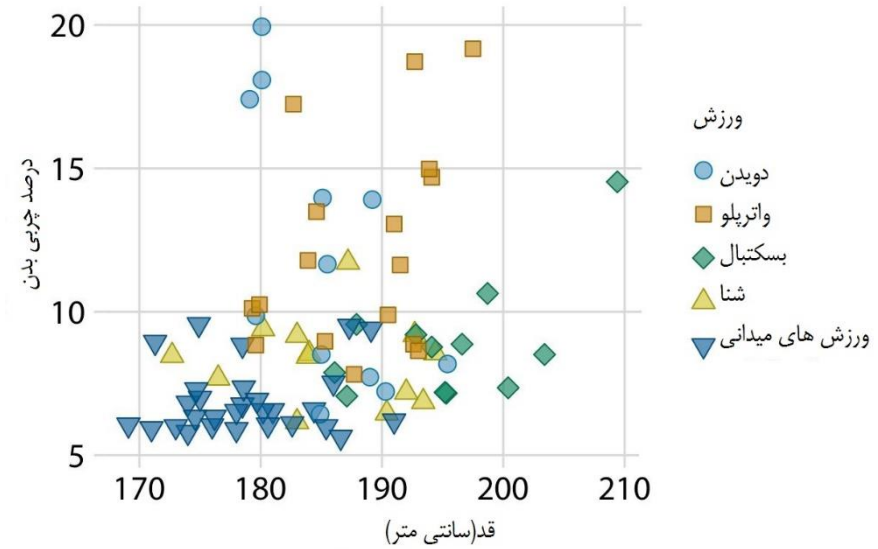
شکل ۲۴–۳ از تنظیمات پیش‌فرضی استفاده می‌کند که در سراسر این کتاب اعمال شده است. به نظر شکل متعادل است، متن خوانا بوده و با اندازهٔ کلی شکل مطابقت دارد.

نکتهٔ مهم این است که می‌توانیم زیاده‌روی کنیم و برچسب‌ها را خیلی بزرگ کنیم (شکل ۲۴–۴). گاهی اوقات به برچسب‌های بزرگ نیاز داریم– برای مثال، اگر قرار است اندازهٔ شکل کاهش یابد – بایستی عناصر مختلف شکل (به ویژه، متن برچسب و نمادهای طرح) با هم هماهنگ باشند. در شکل ۲۴–۴، نقاط مورد استفاده برای نمایش داده‌ها نسبت به متن بسیار کوچک هستند. پس از رفع این مشکل، نمودار دوباره قابل قبول می‌شود (شکل ۲۴–۵).

ممکن است به شکل ۲۴–۵ نگاه کنید و همه چیز را خیلی بزرگ ببینید. با این حال، به خاطر داشته باشید که قرار است کوچک شود. شکل را به گونه‌ای کوچک کنید که عرض آن فقط دو تا سه اینچ باشد و به نظر خوب برسد. در واقع، در آن مقیاس، این تنها شکل در این فصل است که خوب به نظر می‌رسد.



شکل ۲۴–۴. درصد چربی بدن در مقابل قد در ورزشکاران مرد. عناصر متن نسبتاً بزرگ هستند و اگر قرار باشد شکل در مقیاس بسیار کوچک بازتولید شود، اندازهٔ آن‌ها ممکن است مناسب باشد. با این حال، نمودار از نظر کلی متعادل نیست. نقاط نسبت به عناصر متن خیلی کوچک هستند. منبع داده: Telford and Cunningham 1991



شکل ۲۴–۵. درصد چربی بدن در مقابل قد در ورزشکاران مرد. اندازهٔ تمام عناصر به گونه‌ای است که شکل متعادل است و می‌توان آن را در مقیاس کوچک بازتولید کرد. منبع داده: Telford and Cunningham 1991

همیشه به نسخه‌های کوچک شده شکل‌های خود نگاه کنید تا مطمئن شوید که برچسب‌های محور اندازهٔ مناسبی دارند.

به نظر می‌رسد دلیل روان‌شناسی ساده‌ای وجود دارد که چرا ما به‌طور معمول شکل‌هایی می‌سازیم که برچسب‌های محور آن‌ها خیلی کوچک است، و این به نمایشگرهای کامپیوتری بزرگ و با وضوح بالا مربوط می‌شود. ما به طور معمول پیش‌نمایش شکل‌ها را بر روی صفحهٔ نمایش رایانه مشاهده می‌کنیم، و اغلب این کار را در حالی انجام می‌دهیم که شکل، فضای زیادی را روی صفحه اشغال می‌کند. در این حالت مشاهدهٔ حتی متن نسبتاً کوچک، کاملاً خوب و خوانا به نظر رسیده و متن بزرگ می‌تواند ناخوشایند و نامتناسب به نظر برسد. در واقع، اگر اوّلین شکل را از این فصل بگیرید و آن را تا جایی بزرگ کنید که کل صفحه شما را پر کند، احتمالاً فکر می‌کنید که ظاهر خوبی دارد. راه حل این است که همیشه اطمینان حاصل کنید که به شکل‌های خود در اندازه‌ای که قرار است به طور واقعی چاپ شوند نگاه می‌کنید. می‌توانید کوچک‌نمایی کنید تا عرض آن‌ها روی صفحه نمایش فقط سه تا پنج اینچ باشد، یا به خوبی عقب بایستید و بررسی کنید که آیا از فاصله‌ای قابل توجّه هنوز هم ظاهر آن‌ها خوب به نظر می‌رسد یا خیر.

فصل ۲۵

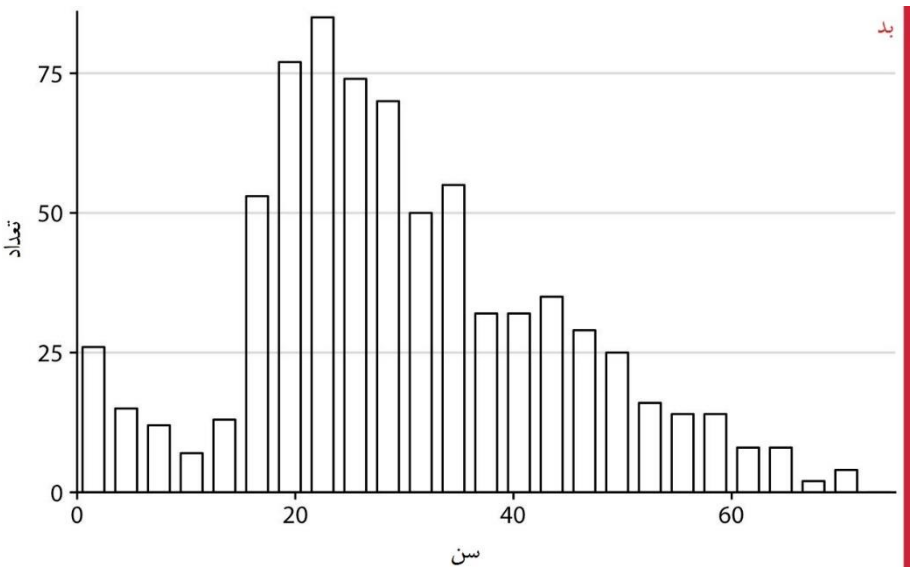
پرهیز از ترسیم خطوط

در صورت امکان، داده‌ها را با اشکال رنگی یک‌دست نمایش دهید و نه با خطوطی که آن اشکال را مشخّص می‌کنند. اشکال با رنگ یک‌دست نسبت به خطوط محیطی راحت‌تر به عنوان اجسام منسجم درک می‌شوند، کمتر احتمال دارد که خطاهای بصری ایجاد کنند، و سریع‌تر منجر به درک مقادیر داده‌ها می‌شوند. به نظر می‌رسد نمودارهایی که از اشکال یک‌دست استفاده می‌کنند نسبت به نسخه‌های معادلی که از نقاشی‌های خطی استفاده می‌کنند، واضح‌تر و قابل قبول‌تر هستند. بنابراین، پیشنهاد می‌شود تا حد امکان از ترسیم خطوط اجتناب شود. با این حال تأکید می‌شود که این توصیه، جایگزین اصل جوهر متناسب نیست (فصل ۱۷).

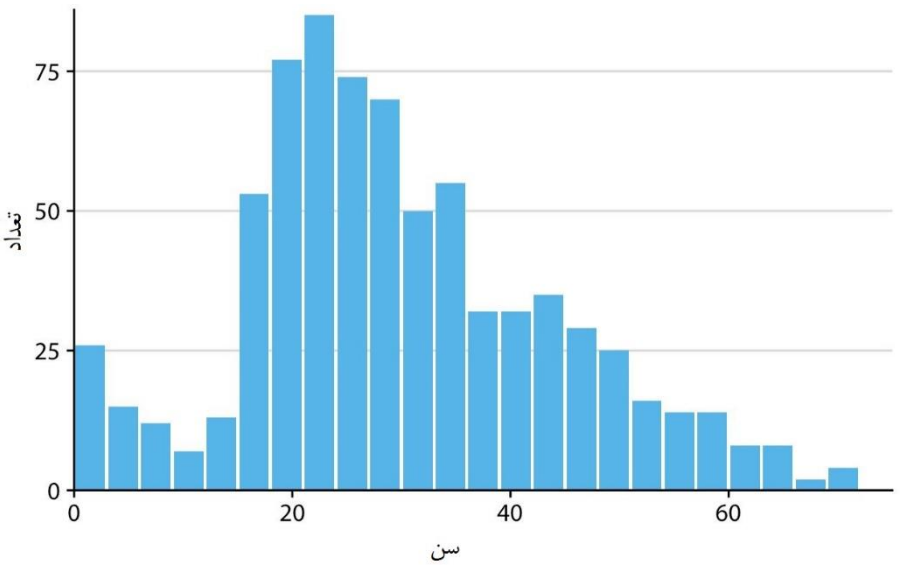
ترسیم خطوط سابقه‌ای طولانی در حوزهٔ نمایش داده‌ها دارد، زیرا در بیشتر قرن بیستم، نمودارهای علمی با دست ترسیم می‌شدند و باید به صورت سیاه و سفید تکثیر می‌شدند. این امر مانع از پر کردن مناطق با رنگ‌های تک رنگ، از جمله تکمیل با رنگ خاکستری شد. در عوض، نواحی پر شده گاهی اوقات با استفاده از الگوهای دریچه، متقاطع یا پایه شبیه‌سازی می‌شدند. نرم‌افزارهای اولیه، ترسیم با دست را شبیه‌سازی می‌کردند و به‌طور مشابه از ترسیم خطوط، الگوهای خطچین یا نقطه‌چین، و دریچه‌ای استفاده گسترده‌ای می‌کردند. در حالی که ابزارهای جدید ترسیم نمودار و بسترهای جدید چاپ و انتشار هیچ یک از محدودیت‌های قبلی

را ندارند، بسیاری از برنامه‌های کاربردی فعلی همچنان به جای مناطق توپر، از خطوط محیطی و اشکال خالی به صورت پیش‌فرض استفاده می‌کنند. برای افزایش آگاهی شما از این موضوع، در ادامه چندین نمونه از همین نمودارها را که با خطوط محیطی و مناطق توپر ترسیم شده‌اند، بررسی خواهد شد.

رایج‌ترین و در عین حال نامناسب‌ترین نحوه استفاده از ترسیم خطوط در نمودارهای هیستوگرام و میله‌ای دیده می‌شود. مشکل ترسیم میله با خطوط محیطی این است که نمی‌توان به سرعت فهمید که کدام طرف هر خط معین در داخل یک میله و کدام طرف در خارج آن است. در نتیجه، به ویژه هنگامی که بین میله‌ها شکاف وجود دارد، با یک الگوی بصری گیج‌کننده مواجه می‌شویم که مخاطب را از پیام اصلی نمودار دور می‌کند (نمودار ۲۵-۱). پر کردن میله‌ها با رنگ روشن، یا با رنگ خاکستری در صورت عدم امکان چاپ رنگی، از این مشکل جلوگیری می‌کند (نمودار ۲۵-۲).

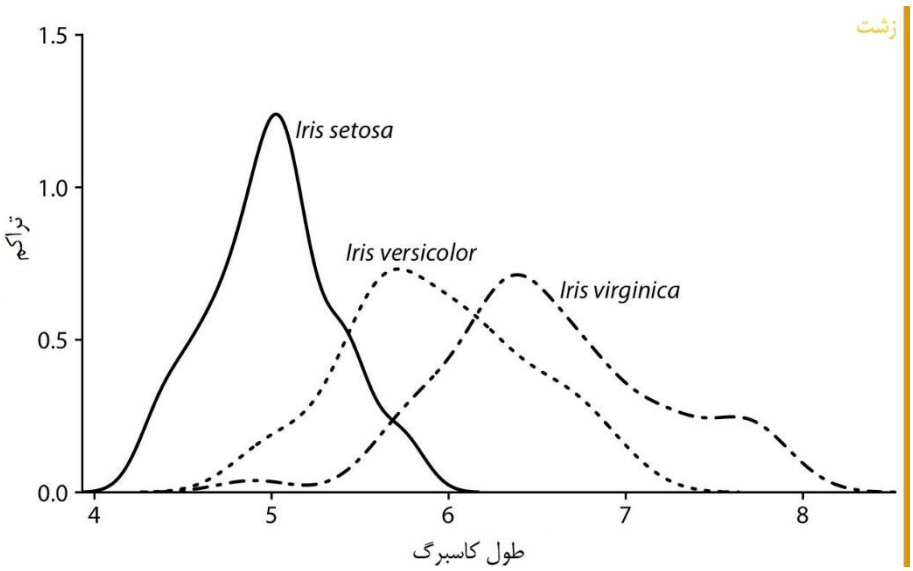


نمودار ۲۵-۱. هیستوگرام سن مسافران تایتانیک که با میله‌های خالی ترسیم شده است. میله‌های خالی یک الگوی بصری گیج‌کننده ایجاد می‌کنند. در مرکز هیستوگرام، تشخیص اینکه کدام قسمت‌ها در داخل میله‌ها و کدام قسمت‌ها بیرون از آن هستند دشوار است. منبع داده: دایره المعارف تایتانیک



نمودار ۲۵–۲. هیستوگرام سن مسافران تایتانیک. این همان هیستوگرام نمودار ۲۵–۱ است که اکنون با میله‌های پر شده ترسیم شده است. شکل توزیع سنی در این نمودار بسیار راحت‌تر قابل تشخیص است. منبع داده: دایره المعارف تایتانیکا.

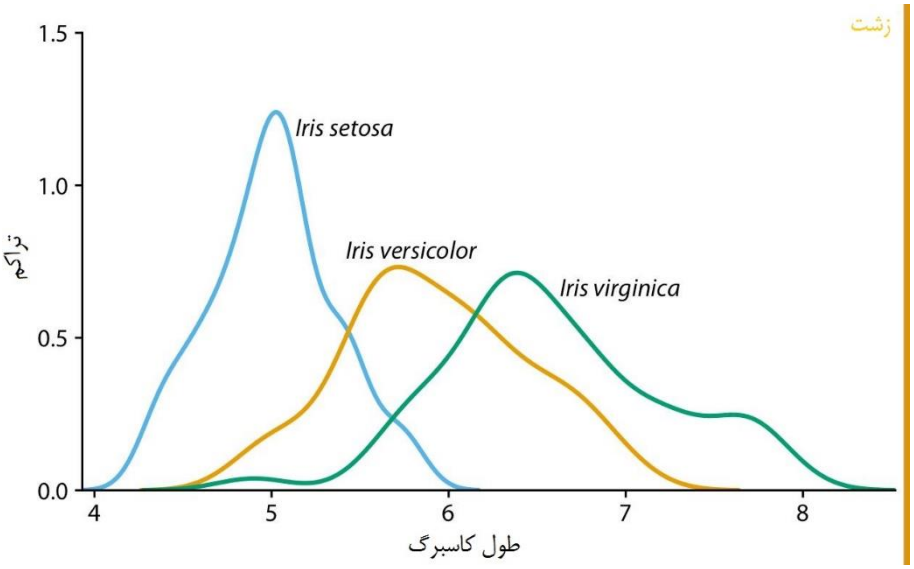
در مرحلهٔ بعد، بیایید نگاهی به نمودار تراکمی قدیمی بیاندازیم. نمودار تخمین تراکمی برای توزیع طول کاسبرگ سه گونهٔ زنبق به صورت کاملاً سیاه و سفید و توسّط ترسیم خطوط نشان داده شده است (نمودار ۲۵–۳). توزیع‌ها فقط با خطوط محیطی نشان داده شده‌اند، و از آنجایی که نمودار سیاه و سفید است، از سبک‌های مختلف خط برای متمایز کردن آن‌ها استفاده گردیده است. این نمودار دو مشکل اساسی دارد. اوّل، سبک‌های خط‌چین تمایز واضحی بین ناحیه زیر منحنی و ناحیه بالای آن ایجاد نمی‌کنند. در حالی که سیستم بینایی انسان در اتصال عناصر مجزای یک خط و تبدیل آن به یک خط پیوسته بسیار خوب عمل می‌کند، با این وجود خط‌چین‌ها متخلخل به نظر می‌رسند و به عنوان یک مرز خوب برای ناحیه محصور عمل نمی‌کنند. دوّم، از آنجایی که خطوط همدیگر را قطع می‌کنند و مناطقی که آن‌ها را محصور می‌کنند سایه ندارند، جدا کردن چگالی‌های مختلف از شش طرح قابل مشاهده دشوار است. اگر برای هر سه توزیع از خطوط یکپارچه به جای خط‌چین استفاده می‌شد، این اثر حتی قوی‌تر نیز می‌بود.



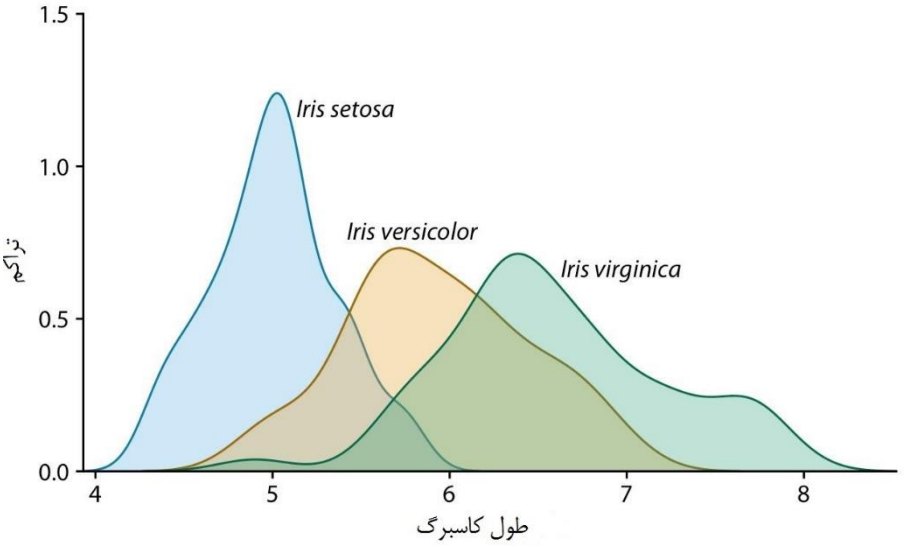
نمودار ۲۵–۳. تخمین تراکمی طول کاسبرگ سه گونهٔ مختلف زنبق. سبک‌های خط شکسته مورد استفاده برای زنبق ورسیکالر و زنبق ویرجینیکا این تصوّر را که مناطق زیر منحنی‌ها از نواحی بالای آن‌ها متمایز هستند، کاهش می‌دهد.
منبع داده: Fisher 1936

برای حل مشکل مرزهای متخلخل می‌توان از خطوط رنگی به جای خطچین استفاده نمود (نمودار ۲۵–۴). با این حال، نواحی تراکم در نمودار حاصل هنوز نمای بصری کمی دارند. به طور کلی، نسخه با مناطق پر شده (نمودار ۲۵–۵) واضح‌ترین و شهودی‌ترین حالت است. با این حال، مهم است که مناطق پر شده تا حدی شفاف شوند، به طوری که توزیع کامل برای هر گونه قابل مشاهده باشد.

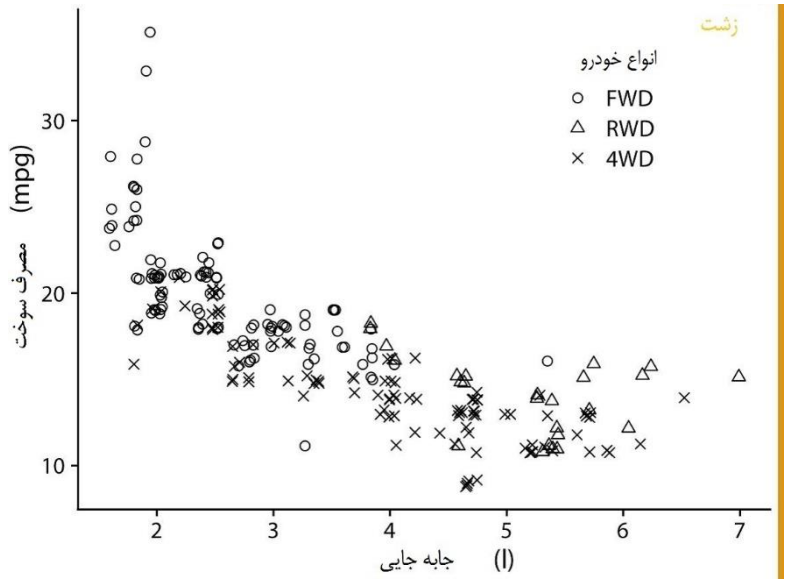
مساله ترسیم خطوط در خصوص نمودارهای پراکنش، که در آن انواع نقاط مختلف به صورت دایره، مثلث، یا صلیب‌های توخالی ترسیم می‌شوند، نیز به وجود می‌آیند. به عنوان مثال، نمودار ۲۵–۶ را در نظر بگیرید. این نمودار حاوی خطای بصری زیادی است و انواع مختلف نقاط به خوبی از یکدیگر قابل تفکیک نیستند. ترسیم همان نمودار با اشکال رنگی توپر این مشکل را برطرف می‌کند (نمودار ۲۵–۷).



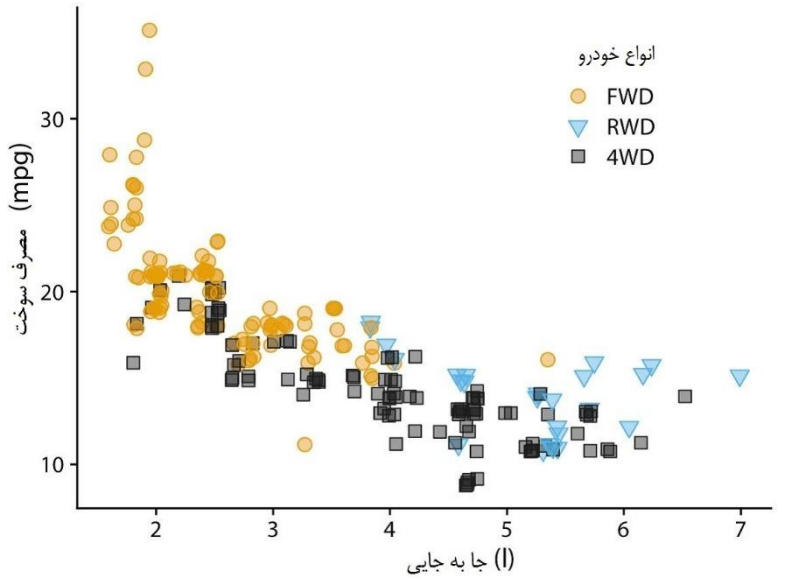
نمودار ۲۵–۴. تخمین تراکم طول کاسبرگ سه گونهٔ مختلف زنبق. با استفاده از خطوط یکدست و رنگی، مشکل نمودار ۲۵–۳. که به نظر می‌رسید نواحی بالا و پایین خطوط به هم متصل هستند، حل شده است. با این حال، هنوز درک قوی از اندازهٔ منطقهٔ زیر هر منحنی وجود ندارد. منبع داده: Fisher 1936



نمودار ۲۵–۵. تخمین تراکم طول کاسبرگ سه گونهٔ مختلف زنبق، که به صورت مناطق سایه‌دار نیمه شفّاف نشان داده شده‌است.سایه‌زنی کمک می‌کند تا تراکم سه منحنی به‌عنوان سه عنصر مجزا درک شود. منبع داده: Fisher 1936



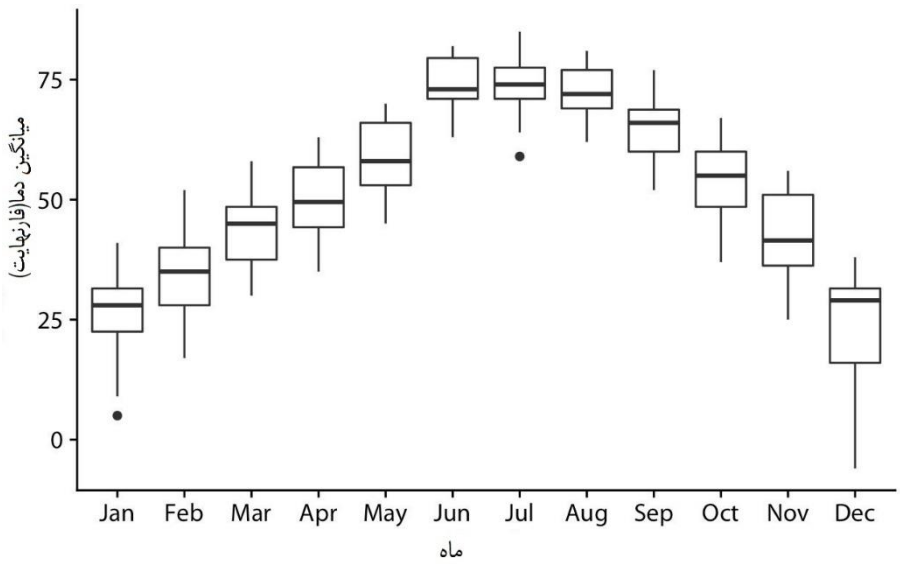
نمودار ۲۵–۶. مصرف سوخت شهری در مقابل جابجایی موتوری، برای خودروهای با دیفرانسیل جلو (FWD)، دیفرانسیل عقب (RWD) و تمام چرخ متحرّک (WD۴). سبک‌های مختلف نقطه، که همه به صورت نمادهای خطی سیاه و سفید هستند، خطای بصری قابل توجّهی ایجاد کرده و خواندن نمودار را دشوار می‌کند. منبع داده: آژانس حفاظت از محیط زیست ایالات متحده (EPA)



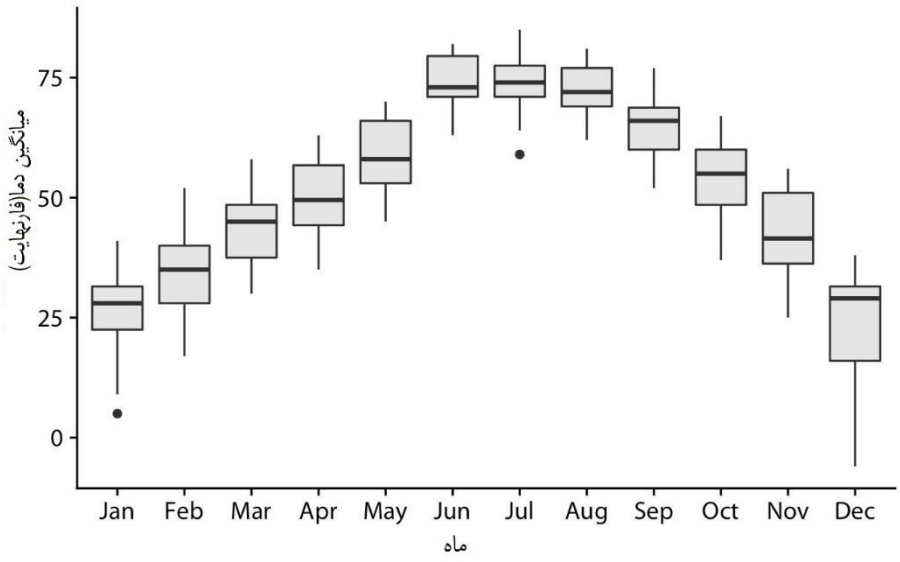
نمودار ۲۵–۷. مصرف سوخت شهری در مقابل جابجایی موتوری. با استفاده از رنگ‌ها و اشکال مختلف توپر برای انواع مختلف خودرو، این نمودار به صورت بصری گونه‌های خودرو را جدا کرده و در صورت نیاز در مقیاس خاکستری نیز قابل چاپ است. منبع داده: EPA

نقـاط تـوپر نسـبت بـه نقـاط توخـالی ارجـح هسـتند، زیـرا نقـاط تـوپر نمـای بصـری بهتـری دارنـد. اسـتدلالی کـه گـاهی اوقـات بـه نفـع نقـاط توخـالی شـنیده می‌شـود ایـن اسـت کـه آن‌هـا بـه درک همپوشـانی داده‌هـا کمـک می‌کننـد، زیـرا منـاطق خـالی در وسـط هـر نقطـه اجـازهٔ مشـاهدهٔ نقـاط دیگری که ممکن است در زیر آن نهفته باشـد را فـراهم می‌کنـد. بـا ایـن وجـود بـه نظـر می‌رسـد بـه طور کلی مزیّت دیـدن نقـاط همپوشـان، بیشـتر از ضـرر خطـای بصـری مـرتبط بـا نمادهـای توخـالی نیست. رویکردهای دیگـری نیـز بـرای مقابلـه بـا همپوشـانی نقـاط وجـود دارد کـه برخـی از آن‌هـا در فصل ۱۸ ارائه شده است.

در نهایـت، بیاییـد نمودارهـای جعبـه‌ای را در نظـر بگیـریم. نمودارهـای جعبـه‌ای معمـولاً ماننـد نمودار ۲۵–۸ بـا کادرهـای خـالی ترسـیم می‌شـوند. بـه نظـر می‌رسـد یـک سـایه روشـن بـرای جعبـه مناسـب باشـد (نمـودار ۲۵–۹). سـایه‌زنی، جعبـه را از پس‌زمینـه طـرح جـدا می‌کنـد، و بـه‌ویژه زمـانی بسـیار کمک‌کننـده اسـت کـه چنـدین جعبـه دقیقـاً در کنـار یکـدیگر نشـان داده شـده باشـند (نمودارهـای ۲۵–۸ و ۲۵–۹). در نمـودار ۲۵–۸، تعـداد زیـاد جعبـه‌ها و خطـوط می‌توانـد مجـدداً ایـن تـوهّم را ایجـاد کنـد کـه منـاطق پس‌زمینـهٔ خـارج از جعبـه‌ها در واقـع در داخـل شـکل دیگـری هسـتند، همـانطور کـه در نمـودار ۲۵–۱ بحـث شـد. ایـن مشـکل در نمـودار ۲۵–۹ برطـرف شـده است. گـاهی اوقـات ایـن انتقـاد مطـرح می‌شـود کـه سـایه‌زنـی داخـل جعبـه بـه ۵۰ درصـدی میـانی داده‌ها وزن بیشـتری می‌دهـد. امـا در حقیقـت، در ذات نمـودار جعبـه‌ای ایـن مسـاله نهفتـه اسـت کـه به ۵۰ درصد میانی داده‌ها وزن بیشـتری می‌دهـد، چـه جعبـه سـایه‌دار باشـد و چـه غیـر سـایه‌دار. اگـر این تاکید مدنظر محقق نیسـت، نبایسـتی از نمـودار جعبـه‌ای اسـتفاده نمایـد. بـه جـای آن، بایسـتی از نمودار ویولن، نقاط لرزان، یا نمودار سینا استفاده گردد (فصل ۹).



نمودار ۲۵–۸. توزیع میانگین دمای روزانه در شمال شرقی لینکلن، در سال ۲۰۱۶. جعبه‌ها به روش سنّتی و بدون سایه ترسیم شده‌اند. منبع داده: آب و هوای زیرزمینی



نمودار ۲۵–۹. توزیع میانگین دمای روزانه در شمال شرقی لینکلن، در سال ۲۰۱۶. با سایه‌زنی خاکستری روشن جعبه‌ها، قابلیّت تمایز آن‌ها نسبت به پس‌زمینه افزایش می‌یابد. منبع داده: آب و هوای زیرزمینی.

فصل ۲۶

پرهیز از نمای سه بُعدی

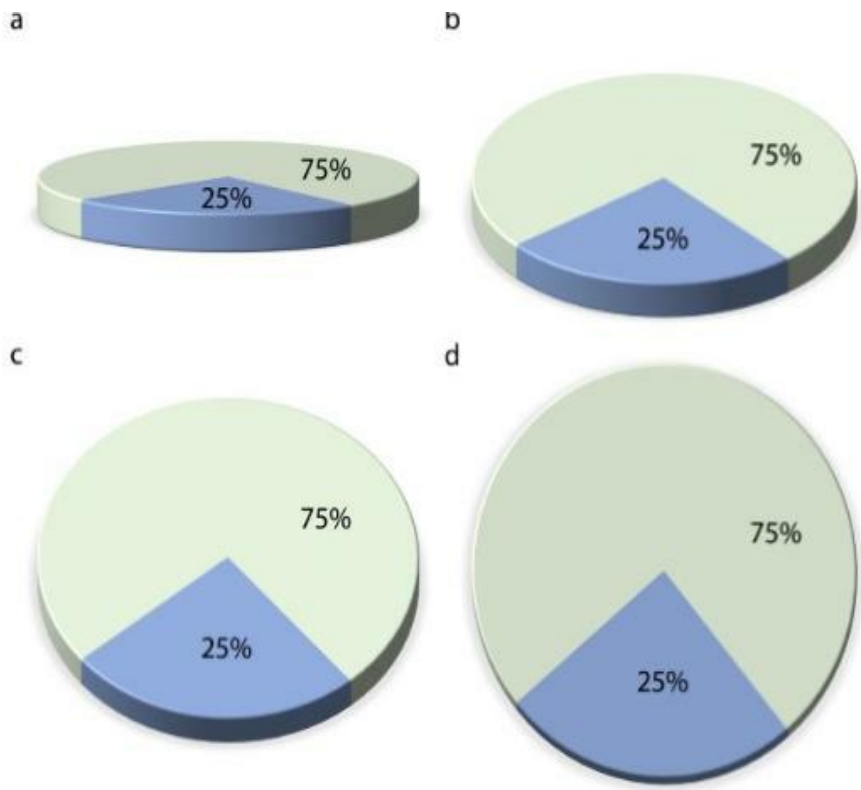
طرح‌های سه بُعدی به ویژه در ارائه‌های تجاری و همچنین در بین دانشگاهیان بسیار محبوب هستند. امّا تقریباً همیشه به طور نامناسب استفاده می‌شوند. به ندرت پیش می‌آید که نموداری سه بُعدی ببینم که با تبدیل آن به نمودار دو بُعدی معمولی قابل ارتقا نباشد. در این فصل، توضیح خواهیم داد که چرا نمودارهای سه بُعدی مشکل دارند، چرا به طور کلی به آن‌ها نیازی نیست، و در چه شرایط خاصی نمودارهای سه بُعدی ممکن است مناسب باشند.

اجتناب از استفاده بی‌دلیل از نمودارهای سه بُعدی

بسیاری از ابزارهای ترسیم به شما امکان می‌دهند با تبدیل عناصر گرافیکی نمودارها به اجزای سه بُعدی، نمودارهای خود را زیباتر کنید. معمولاً نمودارهای دایره‌ای را می‌بینیم که به دیسک‌هایی تبدیل می‌شوند که در فضا چرخیده‌اند، نمودارهای میله‌ای به ستون‌ها و نمودارهای خطی به نوار تبدیل می‌شوند. قابل ذکر است که در هیچ یک از این موارد، بُعد سوم، هیچ داده واقعی را منتقل نمی‌کند. نمای سه بُعدی صرفاً برای تزئین نمودار استفاده می‌شود که ما آن را استفاده بی‌دلیل از نمودار سه بُعدی می‌نامیم. بدون شک این کار صحیح نیست و باید از واژگان بصری دانشمندان داده پاک شود.

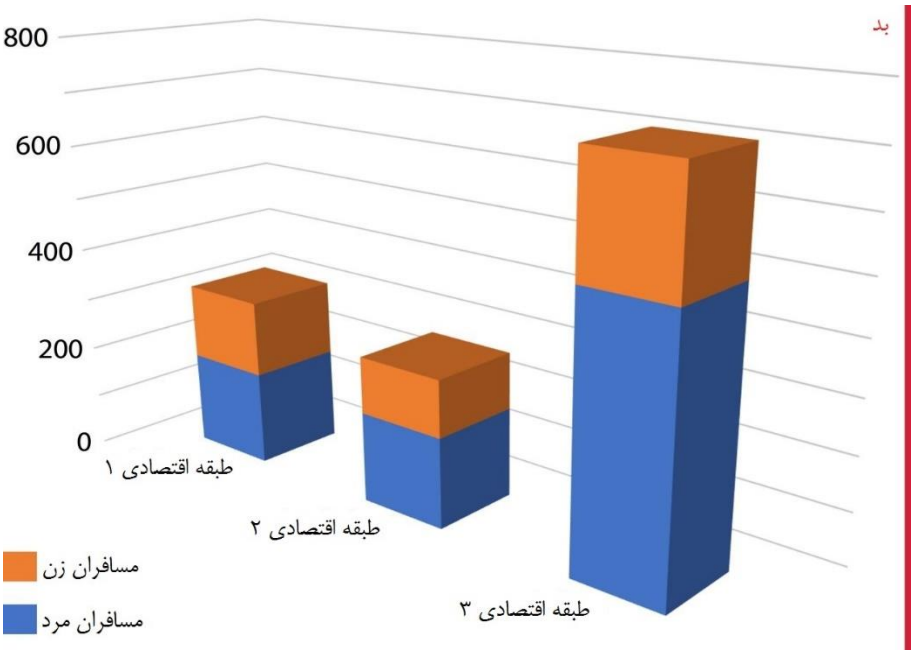
مشکل استفاده بی دلیل از نمای سه بُعدی این است که نمایش سه بُعدی عناصر بر محیط دو بُعدی برای چاپ یا نمایش بر روی نمایشگر، داده‌ها را دچار اختلال می‌کند. سیستم بینایی

انسان سعی می‌کند این اختلال را تصحیح کند، به این صورت که تصویر دو بُعدی یک عنصر سه بُعدی را در فضای سه بُعدی نگاشت می‌کند. با این حال، این اصلاح جزئی است. به عنوان مثال، بیایید یک نمودار دایره‌ای ساده با دو قطاع، یکی ۲۵ درصد از داده‌ها و دیگری ۷۵ درصد را در نظر بگیریم و این دایره را در فضا بچرخانیم (شکل ۲۶-۱). با تغییر زاویه‌ای که به قطاع‌ها نگاه می‌کنیم، به نظر می‌رسد اندازهٔ هر قطاع نیز تغییر می‌کند. به طور خاص، قطاع ۲۵ درصد که در جلوی دایره قرار دارد، به نظر می‌رسد نسبت به زمانی که از یک زاویه صاف به دایره نگاه می‌کنیم، بیشتر از ۲۵ درصد از مساحت را اشغال می‌کند (شکل ۲۶-۱ الف).



شکل ۲۶-۱. نمودار دایره‌ای سه بُعدی که از چهار زاویه مختلف نشان داده شده است. چرخاندن دایره به بُعد سوّم باعث می‌شود قطاع‌های جلوتر بزرگتر از آنچه هستند و قطاع‌های عقب‌تر کوچکتر از آنچه هستند، به نظر برسند. در اینجا، در بخش‌های (الف)، (ب) و (ج)، قطاع آبی مربوط به ۲۵ درصد داده‌ها، بیش از ۲۵ درصد از ناحیه نشان‌دهندهٔ دایره را از نظر بصری اشغال می‌کند. تنها بخش (د) نمایش دقیقی از داده‌ها است.

مشکلات مشابهی برای انواع دیگر نمودارهای سه بُعدی ایجاد می‌شود. شکل ۲۶-۲ تفکیک مسافران تایتانیک را بر اساس طبقهٔ اقتصادی و جنسیّت با استفاده از میله‌های سه بُعدی نشان می‌دهد. به دلیل نحوهٔ چیدمان میله‌ها نسبت به محورها، میله‌ها کوتاه‌تر از آنچه واقعاً هستند به نظر می‌رسند. به عنوان مثال، در مجموع ۳۲۲ مسافر در طبقهٔ اقتصادی اوّل مسافرت می‌کردند، با این حال شکل ۲۶-۲ نشان می‌دهد که این تعداد کمتر از ۳۰۰ نفر است. این گمان به این دلیل به وجود می‌آید چون ستون‌های نشان‌دهندهٔ داده‌ها از دو سطح پشتی که خطوط افقی خاکستری وجود دارد، فاصله دارند. برای مشاهدهٔ این اثر، لبهٔ پایینی یکی از ستون‌ها را امتداد دهید تا به پایین‌ترین خط خاکستری، که نشان‌دهندهٔ صفر است، برسد. سپس همین کار را برای لبه‌های بالایی انجام دهید، می‌بینید که ستون‌ها بلندتر از آن چیزی هستند که در نگاه اوّل به نظر می‌رسند (شکل ۶-۱۰ را در فصل ۶ برای نسخه دو بُعدی منطقی‌تر از این شکل ببینید).



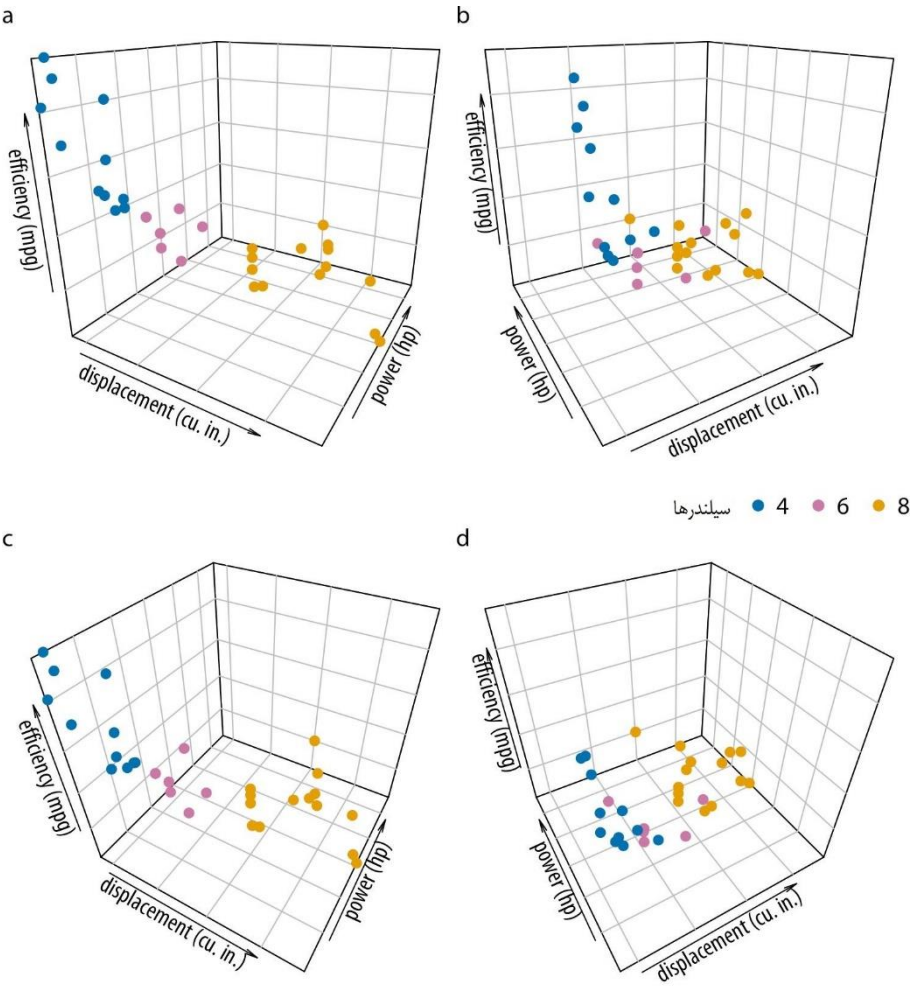
شکل ۲۶-۲. تعداد مسافران زن و مرد تایتانیک که در طبقهٔ اقتصادی ۱، ۲، و ۳ سفر می‌کنند، به صورت نمودار میله‌ای انباشته سه بُعدی نشان داده شده است. تعداد کل مسافران در طبقهٔ اقتصادی ۱، ۲ و ۳ به ترتیب ۳۲۲، ۲۷۹ و ۷۱۱ نفر است (شکل ۶-۱۰ را ببینید). با این حال، در این نمودار، به نظر می‌رسد که ستون طبقهٔ اقتصادی ۱ کمتر از ۳۰۰ مسافر را نشان می‌دهد، ستون طبقهٔ اقتصادی ۳ به نظر می‌رسد کمتر از ۷۰۰ مسافر را نشان می‌دهد، و ستون طبقهٔ اقتصادی ۲ به نظر می‌رسد نزدیک به ۲۱۰ مسافر و نه ۲۷۹ مسافر واقعی را نشان می‌دهد. علاوه بر این، ستون طبقهٔ اقتصادی ۳ از نظر بصری بر کل شکل غلبه دارد و باعث می‌شود تعداد مسافران در این طبقه بزرگتر از آنچه هست به نظر برسد.

اجتناب از مقیاس‌های موقعیّت سه بُعدی

در حالی که استفاده بی‌دلیل از نمای سه بُعدی را می‌توان به راحتی به عنوان بد برچسب زد، در مورد استفاده از سه مقیاس موقعیّت (x، y، و z) برای نمایش داده‌ها مساله جای بحث دارد. در این مورد، استفاده از بُعد سوّم یک هدف واقعی را دنبال می‌کند. با این وجود، تفسیر نمودارهای حاصل اغلب دشوار است و به نظر باید از آن‌ها اجتناب کرد.

یک نمودار پراکنش سه بُعدی از بازده سوخت در مقابل جابجایی و قدرت برای ۳۲ خودرو را در نظر بگیرید. ما این مجموعه داده را قبلاً در فصل ۲ دیدیم (شکل ۲–۵). در اینجا جابجایی را در امتداد محور x، قدرت را در امتداد محور y و بازده سوخت را در امتداد محور z رسم می‌کنیم و هر خودرو را با یک نقطه نشان می‌دهیم (شکل ۲۶–۳). اگرچه این نمودار سه بُعدی از چهار منظر مختلف نشان داده شده است، تصوّر اینکه نقاط دقیقاً چگونه در فضا توزیع شده‌اند دشوار است. به نظر می‌رسد قسمت (د) شکل ۲۶–۳ بسیار گیج‌کننده است. تقریباً به نظر می‌رسد که این قسمت مجموعهٔ داده متفاوتی را نشان می‌دهد، در حالی که هیچ چیزی به جز زاویه‌ای که از آن به نقاط نگاه می‌کنیم، تغییر نکرده است.

مشکل اساسی نمودارهای سه بُعدی این چنینی این است که به دو تبدیل داده جداگانه و متوالی نیاز دارند. اوّلین تبدیل، داده‌ها را از فضای داده به فضای ترسیم سه بُعدی، همانطور که در فصل های ۲ و ۳ در زمینه مقیاس‌های موقعیّت مورد بحث قرار گرفت، نگاشت می‌کند. دوّمین تبدیل، داده‌ها را از فضای ترسیم سه بُعدی به فضای دو بُعدی نمودار نهایی نگاشت می‌کند (بدیهی است که این تبدیل دوّم برای ترسیم‌هایی که در یک محیط سه بُعدی واقعی نشان داده می‌شوند، مانند زمانی که به صورت مجسمه‌های فیزیکی یا اشیاء چاپ شده با چاپ سه بُعدی نشان داده می‌شوند، رخ نمی‌دهد. مشکل اصلی در اینجا ترسیم‌های سه بُعدی نشان داده شده در نمایشگرهای دو بُعدی است). تبدیل دوّم غیرقابل برگشت است. زیرا هر نقطه در صفحه نمایش دو بُعدی با خطی از نقاط در فضای ترسیم سه بُعدی مطابقت دارد. بنابراین، ما نمی‌توانیم به طور منحصربه‌فرد تعیین کنیم که نقطه داده‌ای خاص در چه قسمتی از فضای سه بُعدی قرار دارد.

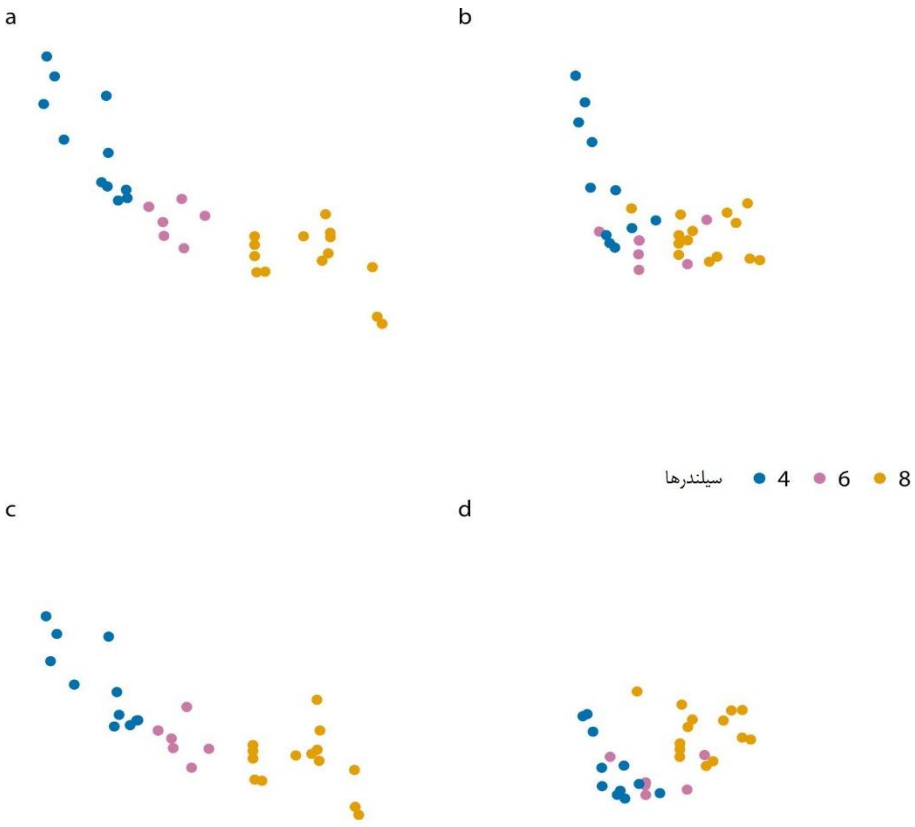


شکل ۲۶–۳. بازده سوخت در مقابل جابجایی و قدرت برای ۳۲ خودرو (مدل های ۱۹۷۳–۱۹۷۴). هر نقطه نشان‌دهندهٔ یک خودرو است و رنگ نقطه نشان‌دهندهٔ تعداد سیلندرهای خودرو است. همهٔ چهار قسمت داده‌های یکسانی را نشان می‌دهند امّا از منظرهای متفاوت. منبع داده: موتور ترند، ۱۹۷۴.

با این وجود سیستم بصری ما تلاش می‌کند تا تبدیل سه بُعدی به دو بُعدی را معکوس کند. با این حال، این فرآیند غیرقابل اعتماد، مملو از خطا، و به شدت وابسته به سرنخ های مناسب در تصویر است که نوعی حس سه بُعدی بودن را منتقل می‌کند. وقتی این سرنخ‌ها را حذف می‌کنیم، معکوس کردن کاملاً غیرممکن می‌شود. این را می‌توان در شکل ۲۶–۴ مشاهده کرد، که با شکل ۲۶–۳ یکسان است، با این تفاوت که تمام سرنخ‌های مربوط به عمق حذف

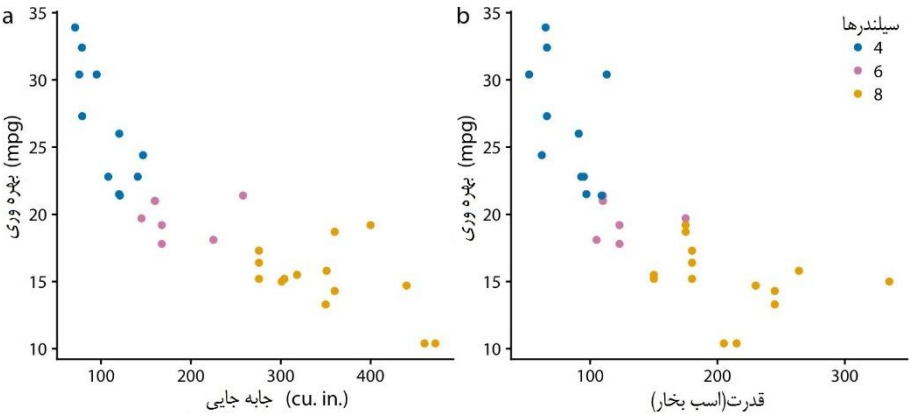
شده‌اند. نتیجه حاصل چهار آرایش تصادفی از نقاط است که ما اصلاً نمی‌توانیم آن‌ها را تفسیر کنیم و حتّی با یکدیگر قابل ارتباط نیستند. آیا می‌توانید بگویید کدام نقاط قسمت (الف) با کدام نقاط از قسمت (ب) مطابقت دارد؟

به جای اعمال دو تبدیل داده جداگانه، که یکی از آن‌ها غیرقابل برگشت است، عموماً بهتر است فقط یک تبدیل مناسب و قابل برگشت را اعمال کنیم و داده‌ها را مستقیماً در فضای دو بُعدی ترسیم کنیم. به ندرت لازم است یک بُعد سوّم به عنوان مقیاس موقعیّت اضافه شود، زیرا متغیّرها را می‌توان با استفاده از مقیاس‌های رنگ، اندازه یا شکل نیز ترسیم کرد. به عنوان مثال، در فصل ۲، پنج متغیّر از مجموعه داده‌های بازده سوخت را به طور همزمان ترسیم کرده‌ایم، امّا تنها از دو مقیاس موقعیّت استفاده نموده‌ایم (شکل ۲–۵).

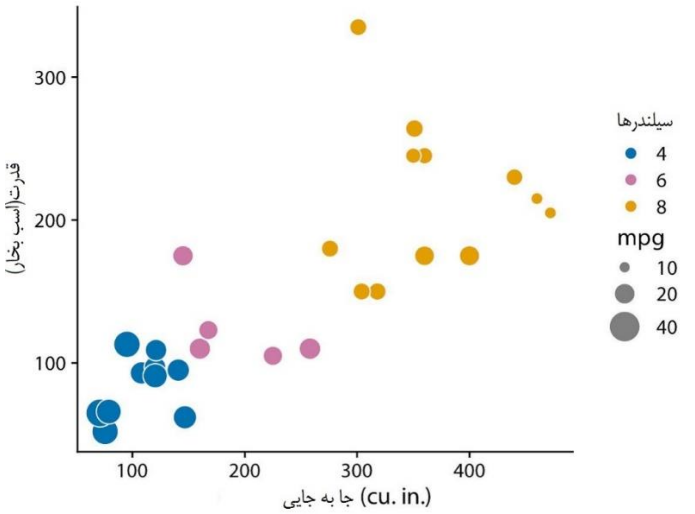


شکل ۲۶–۴. بازده سوخت در مقابل جابجایی و قدرت برای ۳۲ خودرو (مدل های ۱۹۷۳–۱۹۷۴). چهار قسمت فعلی منطبق بر همان قسمت شکل ۲۶–۳ هستند، امّا تمام خطوط شبکه‌ای که سرنخ‌های عمق را ارائه می‌دهند حذف شده‌اند. منبع داده: موتور ترند، ۱۹۷۴.

در ادامه، دو روش جایگزین برای ترسیم دقیق متغیّرهای استفاده شده در شکل ۲۶-۳ نشان داده شده است. اولاً، اگر اساساً به بازده سوخت به عنوان متغیّر پاسخ اهمیت دهیم، می‌توانیم آن را دو بار ترسیم کنیم، یک بار در برابر جابجایی و یک بار در برابر قدرت (شکل ۲۶-۵). دوّم، اگر ما بیشتر به چگونگی ارتباط جابجایی و قدرت با یکدیگر علاقه‌مند باشیم، و بهره‌وری سوخت به عنوان یک متغیّر ثانویه مدنظر باشد، می‌توانیم قدرت را در مقابل جابجایی ترسیم کنیم و بازده سوخت را بر روی اندازهٔ نقاط نگاشت کنیم (شکل ۲۶-۶). هر دو شکل نسبت به شکل ۲۶-۳ مفیدتر و کمتر گیج‌کننده هستند.

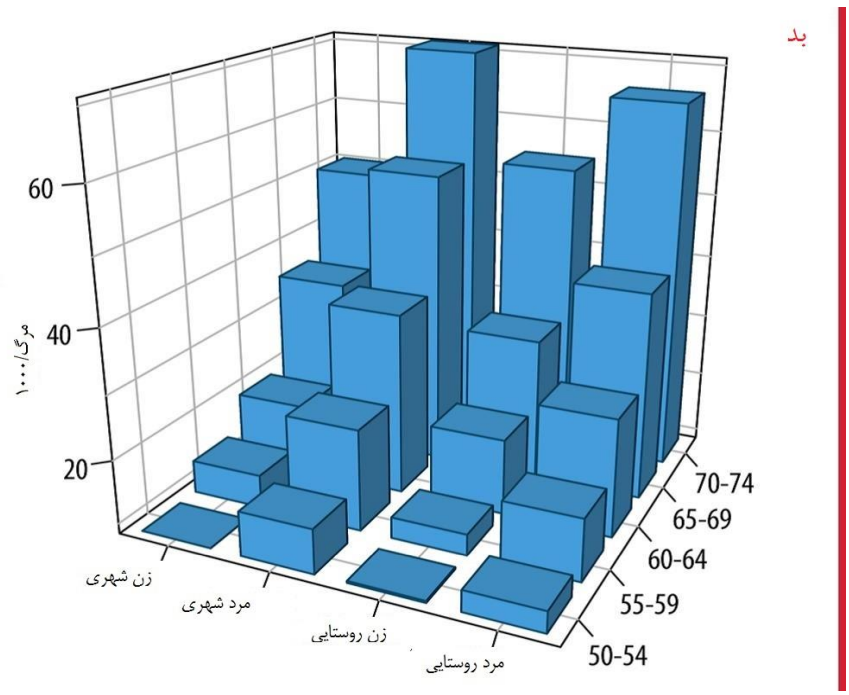


شکل ۲۶-۵. بازده سوخت در مقابل جابجایی (الف) و قدرت (ب) برای ۳۲ خودرو. منبع داده: موتور ترند، ۱۹۷۴



شکل ۲۶-۶. قدرت در مقابل جابجایی برای ۳۲ خودرو، و بازده سوخت که با اندازه نقاط نشان داده شده است. منبع داده: موتور ترند، ۱۹۷۴.

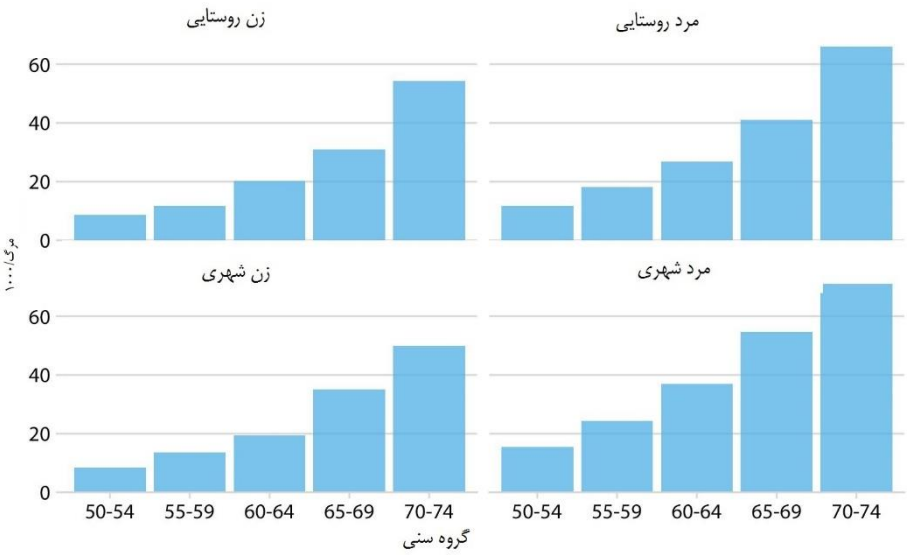
ممکن است سوال کنید که آیا مشکل نمودار پراکنش سه بُعدی این است که نمایش داده‌های واقعی- نقاط- خود هیچ اطّلاعات سه بُعدی را منتقل نمی‌کنند. برای مثال، اگر به جای آن از میله‌های سه بُعدی استفاده کنیم، چه اتفاقی می‌افتد؟ شکل ۲۶-۷ یک مجموعه داده معمولی را نشان می‌دهد که می‌توان با میله‌های سه بُعدی ترسیم کرد، نرخ مرگ و میر در ویرجینیا در سال ۱۹۴۰ بر اساس گروه سنی و جنسیّت و محل سکونت طبقه‌بندی شده است. می‌بینیم که در واقع میله‌های سه بُعدی در تفسیر نمودار به ما کمک می‌کند. بعید است که یک میله در پیش زمینه را با یکی در پس زمینه اشتباه بگیرید یا برعکس. با این وجود، مشکلات مورد بحث در زمینه شکل ۲۶-۲ در اینجا نیز وجود دارد. قضاوت در مورد ارتفاع تک‌تک میله‌ها دشوار است و همچنین مقایسهٔ مستقیم میله‌ها دشوار است. به عنوان مثال، آیا میزان مرگ و میر زنان شهری در گروه سنی ۶۵ تا ۶۹ سال بیشتر از مردان شهری در گروه سنی ۶۰ تا ۶۴ سال است؟



شکل ۲۶-۷. نرخ مرگ و میر در ویرجینیا در سال ۱۹۴۰، که به صورت نمودار میله‌ای سه بُعدی ترسیم شده است. نرخ مرگ و میر برای چهار گروه از مردم (زنان و مردان شهری و روستایی) و پنج گروه سنی (۵۰-۵۴، ۵۵-۵۹، ۶۰-۶۴، ۶۵-۶۹، ۷۰-۷۴) و با واحد مرگ به ازای هر ۱۰۰۰ نفر گزارش شده است. این شکل به عنوان «بد» برچسب‌گذاری شده است زیرا نمای سه بُعدی خواندن نمودار را دشوار می‌کند.

منبع داده: Molyneaux, Gilliam, and Florant ۱۹۴۷

به طور کلی بهتر است به جای ترسیم سه بُعدی از نمودار چندگانه‌های کوچک (فصل ۲۱) استفاده نمود. مجموعه داده مرگ و میر ویرجینیا زمانی که به صورت نمودار چندگانه‌های کوچک نشان داده می‌شود تنها به چهار قسمت نیاز دارد (شکل ۲۶–۸). این شکل واضح و به راحتی قابل تفسیر است. بلافاصله مشخّص می‌شود که میزان مرگ و میر در بین مردان بیشتر از زنان بوده و همچنین مردان شهری نسبت به مردان روستایی نرخ مرگ و میر بیشتری داشته‌اند، در حالی که چنین روندی برای زنان شهری و روستایی مشهود نیست.



شکل ۲۶–۸. نرخ مرگ و میر در ویرجینیا در سال ۱۹۴۰، که به صورت نمودار چندگانه‌های کوچک نشان داده شده است. نرخ مرگ و میر برای چهار گروه از مردم (زنان و مردان شهری و روستایی) و پنج گروه سنی (۵۰–۵۴، ۵۵–۵۹، ۶۰–۶۴، ۶۵–۶۹، ۷۰–۷۴) و با واحد مرگ به ازای هر ۱۰۰۰ نفر گزارش شده است. منبع داده: Molyneaux, Gilliam, and Florant 1947

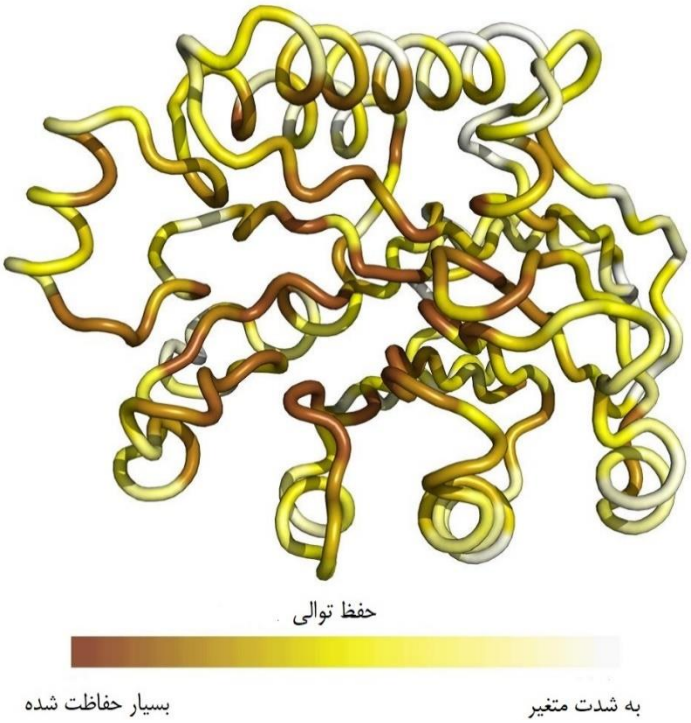
استفاده مناسب از ترسیم سه بُعدی

با این حال، ترسیم با استفاده از مقیاس‌های موقعیّت سه بُعدی می‌تواند گاهی اوقات مناسب باشد. اولاً، اگر ترسیم تعاملی باشد و بیننده بتواند آن را بچرخاند، یا اگر در یک محیط واقعیّت مجازی یا واقعیّت افزوده نشان داده شود که در آن می‌توان نمودار را از زوایای مختلف بررسی کرد، مسائلی که در بخش قبل توضیح داده شد کمتر نگران‌کننده هستند. دوّم، حتی اگر ترسیم تعاملی نباشد، نشان دادن آن در حال چرخش آهسته، به جای استفاده ار یک تصویر ایستا از یک نما، به بیننده این امکان را می‌دهد که تشخیص دهد عناصر گرافیکی مختلف در کجای

فضای سه بُعدی قرار دارند. مغز انسان در بازسازی یک صحنه سه بُعدی از مجموعه‌ای از تصاویر گرفته شده از زوایای مختلف بسیار خوب عمل می‌کند و چرخش آهستهٔ گرافیک دقیقاً این تصاویر را ارائه می‌دهد.

در نهایت، زمانی که می‌خواهیم اشیاء سه بُعدی واقعی و/یا داده‌های نگاشت شده روی آن‌ها را نشان دهیم، استفاده از ترسیم‌های سه بُعدی منطقی است مثلاً برای نشان دادن نقشهٔ برجستهٔ توپوگرافی یک جزیره کوهستانی (شکل ۲۶-۹). به طور مشابه، برای ترسیم حفظ توالی تکاملی یک پروتئین که روی ساختار آن نگاشت شده است، منطقی است که ساختار را به عنوان یک شی سه بُعدی نشان دهیم (شکل ۲۶-۱۰). هر چند در هر دو صورت، اگر این ترسیم‌ها به عنوان پویانمایی‌های چرخشی نشان داده شوند، باز هم تفسیر آن‌ها آسان‌تر خواهد بود. در حالی که در نشریات چاپی سنّتی این امکان وجود ندارد، می‌توان به راحتی از این قابلیت در هنگام انتشار نمودار در وب یا هنگام ارائه سخنرانی استفاده نمود.

شکل ۲۶-۹. برجستگی‌های جزیره کورسیکا در دریای مدیترانه. منبع داده: سرویس نظارت بر زمین کوپرنیک.



شکل ۲۶–۱۰. الگوهای تغییرات تکاملی در یک پروتئین. لوله رنگی نشان‌دهندهٔ استخوان‌بندی پروتئین اگزونوکلئاز III از باکتری اشرشیاکلی است. رنگ‌آمیزی نشان‌دهندهٔ حفظ تکاملی مکان‌های منفرد در این پروتئین است که رنگ تیره نشان‌دهنده اسیدهای آمینه حفظ شده و رنگ‌آمیزی روشن نشان‌دهندهٔ اسیدهای آمینه متغیّر است. منبع داده: Marcos and Echave 2015

فصل ۲۷

آشنایی با رایج‌ترین قالب‌های فایل تصویری

هر کسی که برای نمایش داده‌ها از نمودار استفاده می‌کند، در نهایت باید چند نکته در مورد نحوهٔ ذخیره‌سازی نمودارها در رایانه بداند. قالب‌های بسیار متفاوتی برای فایل تصویری وجود دارد و هر کدام مزایا و معایب خاص خود را دارند. انتخاب قالب فایل و گردش کار مناسب می‌تواند از بسیاری از مشکلات مربوط به آماده‌سازی نمودار پیشگیری کند.

ترجیح ما به طور کلی استفاده از PDF برای فایل‌های آماده انتشار با کیفیت بالا، استفاده از PNG برای اسناد برخط و سایر سناریوهایی که گرافیک بیت مپ مورد نیاز است، و استفاده از JPEG به عنوان راه حل نهایی اگر فایل‌های PNG بیش از حد حجیم باشند، می‌باشد. در بخش‌های بعدی، تفاوت‌های کلیدی بین این قالب‌های فایل و مزایا و معایب مربوط به آن‌ها را توضیح می‌دهیم.

بیت مپ و گرافیک بُرداری

مهم‌ترین تفاوت بین قالب‌های گرافیکی مختلف این است که آن‌ها بیت مپ یا بُرداری هستند (جدول ۲۷-۱). *بیت مپ یا گرافیک شطرنجی*[1] تصویر را به صورت شبکه‌ای از نقاط مجزا (که پیکسل نامیده می‌شوند) ذخیره می‌کند که هر کدام دارای رنگ مشخصی هستند. در مقابل،

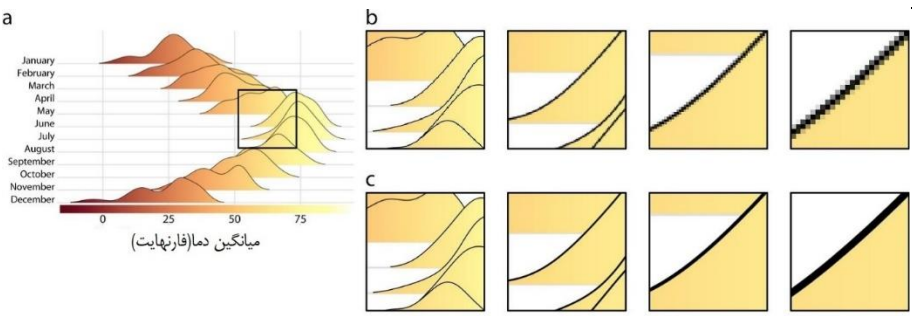
1. raster graphics

گرافیک بُرداری آرایش هندسی عناصر گرافیکی منفرد را در تصویر ذخیره می‌کند. بنابراین، یک تصویر بُرداری حاوی اطلاعاتی مانند «یک خط سیاه از گوشه بالا سمت چپ به گوشه پایین سمت راست، و یک خط قرمز از گوشه پایین سمت چپ به گوشه سمت راست بالا وجود دارد» بوده و تصویر واقعی همانطور که بر روی صفحه نمایش داده می‌شود یا چاپ می‌شود، دوباره ایجاد می‌شود.

جدول ۲۷–۱. قالب‌های فایل تصویری رایج

مخفف	نام کامل	نوع	کاربرد
PDF	Portable Document Format	بُردار	استفاده عمومی
EPS	Encapsulated PostScript	بُردار	استفاده عمومی، قدیمی است؛ از پی دی اف استفاده کنید
SVG	Scalable Vector Graphics	بُردار	استفاده برخط
PNG	Portable Network Graphics	بیت مپ	بهینه شده برای نقاشی خطی
JPEG/JPG	Joint Photographic Experts Group	بیت مپ	بهینه شده برای تصاویر عکاسی
TIFF	Tagged Image File Format	بیت مپ	چاپ، بازتولید دقیق رنگ
RAW	Raw Image File	بیت مپ	عکاسی دیجیتال نیاز به پردازش ثانویّه
GIF	Graphics Interchange Format	بیت مپ	برای نمودارهای ایستا منسوخ شده امّا برای پویانمایی مناسب است

گرافیک‌های بُرداری را «مستقل از رزولوشن» نیز می‌نامند، زیرا می‌توان آن‌ها را تا اندازه دلخواه بدون از دست دادن جزئیات یا وضوح بزرگنمایی کرد. شکل ۲۷–۱ را ببینید.



شکل ۲۷–۱. تصویری از تفاوت کلیدی بین گرافیک بُرداری و بیت مپ. (الف) تصویر اصلی. مربع سیاه نشان‌دهندهٔ ناحیه‌ای است که در قسمت‌های (ب) و (ج) بزرگنمایی شده است. (ب) افزایش بزرگنمایی ناحیه مشخّص شده از قسمت (الف) زمانی که تصویر به صورت گرافیک بیت مپ ذخیره شده است. می‌توان دید که چگونه با افزایش بزرگنمایی، تصویر به طور فزاینده‌ای پیکسلی می‌شود. (ج) افزایش بزرگ‌نمایی یک نمایش بُرداری از تصویر. تصویر وضوح کامل را در سطوح بزرگنمایی دلخواه حفظ می‌کند.

گرافیک بُرداری دو جنبه منفی دارد که اغلب باعث ایجاد مشکل در زمان کاربرد آن‌ها می‌شود. اوّل، از آنجایی که گرافیک‌های بُرداری به سرعت توسّط برنامه گرافیکی که با آن نمایش داده می‌شوند، در لحظه ترسیم می‌شوند، ممکن است تفاوت‌هایی در ظاهر یک گرافیک مشابه در دو برنامهٔ مختلف یا در دو رایانهٔ متفاوت وجود داشته باشد. این مشکل اغلب در متن رخ می‌دهد، برای مثال زمانی که قلم مورد نیاز در دسترس نیست و نرم‌افزار قلم دیگری را جایگزین می‌کند. جایگزینی قلم معمولاً به بیننده اجازه می‌دهد متن را بخواند، امّا تصویر حاصل به ندرت خوب به نظر می‌رسد. راه‌هایی برای جلوگیری از این مشکلات وجود دارد، مانند مشخّص کردن یا جاسازی تمام قلم‌ها در فایل PDF، هرچند ممکن است برای این کار نیاز به نرم‌افزار خاص و/یا دانش فنّی باشد. در مقابل، تصاویر بیت مپ همیشه یکسان هستند.

دوّم، برای شکل‌های بسیار بزرگ و/یا پیچیده، گرافیک‌های بُرداری می‌تواند منجر به افزایش چشمگیر حجم فایل شود و ترسیم آن کند شود. برای مثال، یک نمودار پراکنش با میلیون‌ها نقطه داده حاوی مختصّات x و y هر نقطه است، و هر نقطه باید هنگام نمایش تصویر ترسیم شود، حتی اگر نقاط با هم همپوشانی داشته باشند و/یا توسّط سایر عناصر گرافیکی پنهان شوند. در نتیجه، حجم فایل ممکن است چندین مگابایت باشد و ممکن است مدّتی طول بکشد تا نرم‌افزار شکل را نمایش دهد. در اوایل دههٔ ۲۰۰۰ که دانشجوی پسادکترا بودم، یک فایل PDF ایجاد کردم که در آن زمان نمایش آن در Acrobat Reader تقریباً یک ساعت طول کشید. در حالی که رایانه‌های مدرن بسیار سریع‌تر هستند و زمان‌های رندر چند دقیقه‌ای این روزها بی‌سابقه است، حتّی اگر می‌خواهید شکل خود را در یک سند بزرگ‌تر جاسازی کنید و نرم‌افزار مربوطه متوقف شود، حتّی زمان رندر چند ثانیه‌ای هر بار که به صفحه حاوی آن شکل می‌رسید، می‌تواند مختل‌کننده باشد. البته، از طرف دیگر، اشکال ساده با عناصر کم (مثلاً چند نقطه داده و مقداری متن) اغلب به صورت گرافیک بُرداری بسیار کوچکتر از بیت مپ هستند، و نرم‌افزار مربوطه حتّی ممکن است چنین اشکالی را نسبت به تصاویر بیت مپ متناظر سریع‌تر نمایش دهد.

فشرده‌سازی با و بدون افت در گرافیک‌های بیت مپ

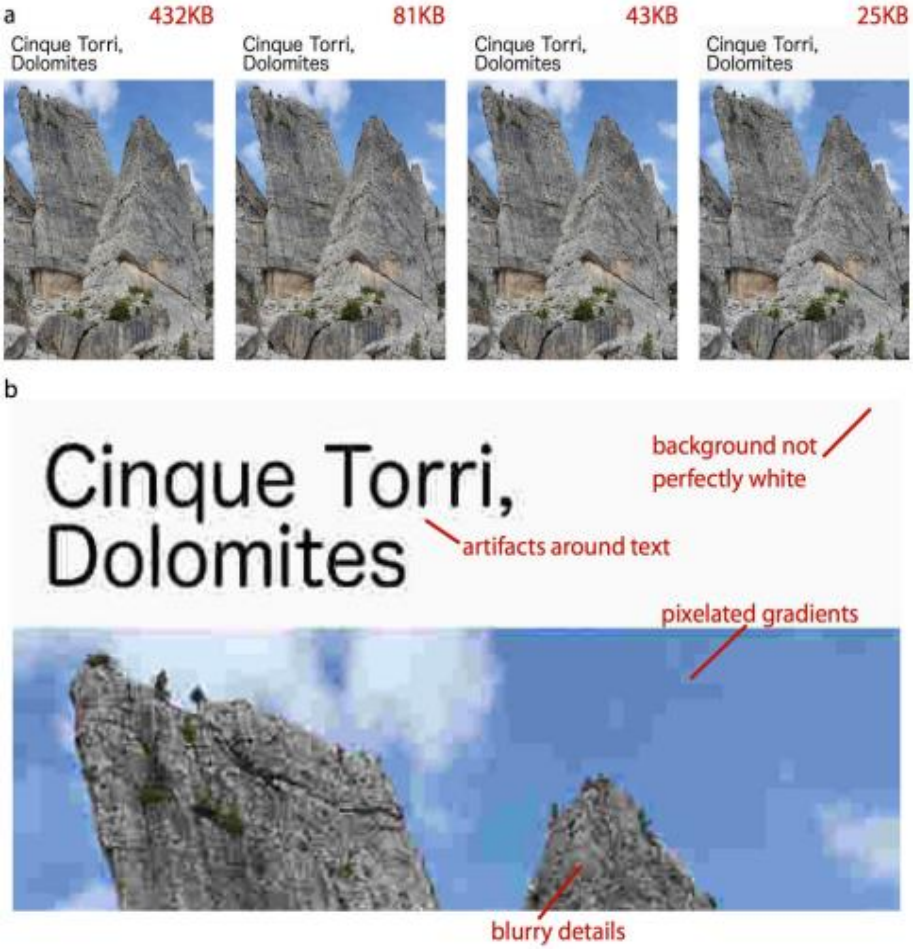
اکثر قالب‌های فایل بیت مپ از نوعی فشرده‌سازی داده استفاده می‌کنند تا اندازه فایل را کنترل کنند. دو نوع اساسی فشرده‌سازی وجود دارد: با افت و بدون افت. فشرده‌سازی بدون افت تضمین می‌کند که تصویر فشرده شده پیکسل به پیکسل مشابه تصویر اصلی است، در حالی که فشرده‌سازی با افت در ازای حجم کوچک‌تر فایل، مقداری افت کیفیت تصویر را می‌پذیرد.

برای درک اینکه چه زمانی استفاده از فشرده‌سازی بدون افت یا فشرده‌سازی با افت مناسب است، درک نحوهٔ کار این الگوریتم‌های فشرده‌سازی مفید است. بیایید ابتدا فشرده‌سازی بدون افت را در نظر بگیریم. تصویری را با پس‌زمینه سیاه تصوّر کنید، که مناطق بزرگی از تصویر سیاه است و بنابراین بسیاری از پیکسل‌های سیاه دقیقاً در کنار یکدیگر قرار می‌گیرند. هر پیکسل سیاه را می‌توان با سه صفر در یک ردیف نشان داد (یعنی ۰ ۰ ۰) که نشان‌دهندهٔ شدت صفر در کانال‌های رنگی قرمز، سبز و آبی تصویر است. نواحی پس زمینه سیاه در تصویر با هزاران صفر در فایل تصویر مطابقت دارد. حالا فرض کنید جایی در تصویر ۱۰۰۰ پیکسل سیاه متوالی دارد و در نتیجه ۳۰۰۰ صفر. به جای نوشتن تمام این صفرها، می‌توانیم به سادگی تعداد کل صفرهای مورد نیاز خود را ذخیره کنیم، مثلاً با نوشتن ۳۰۰۰ ۰. به این ترتیب، ما دقیقاً همان اطّلاعات را تنها با دو عدد، یعنی تعداد (در اینجا، ۳۰۰۰) و مقدار (اینجا، ۰) منتقل کرده‌ایم. در طول سال‌ها، بسیاری از ترفندهای هوشمندانه در این راستا توسعه یافته‌اند و قالب‌های تصویر بدون افت مدرن (مانند PNG) می‌توانند داده‌های بیت مپ را با کارایی چشمگیر ذخیره کنند. با این حال، همه الگوریتم‌های فشرده‌سازی بدون افت زمانی بهترین عملکرد را دارند که تصاویر دارای مناطق وسیعی از رنگ یکنواخت باشند و بنابراین جدول ۲۷-۱ PNG را به‌عنوان بهینه‌سازی شده برای نقاشی‌های خطی فهرست می‌کند.

تصاویر عکاسی به ندرت دارای چندین پیکسل با رنگ و روشنایی یکسان در کنار یکدیگر هستند. در عوض، آن‌ها دارای طیف رنگی و سایر الگوهای نسبتاً منظّم در مقیاس‌های مختلف هستند. بنابراین، فشرده‌سازی بدون افت این تصاویر اغلب خیلی کارایی ندارد و فشرده‌سازی با افت به عنوان یک جایگزین توسعه داده شده است. ایدهٔ کلیدی فشرده‌سازی با افت این است که برخی از جزئیات در یک تصویر برای چشم انسان بسیار ظریف هستند و می‌توان آن‌ها را بدون کاهش واضح در کیفیت تصویر حذف نمود. به عنوان مثال، یک طیف ۱۰۰۰ پیکسلی را در نظر بگیرید که هر کدام رنگ کمی متفاوت دارد. اگر این طیف فقط با ۲۰۰ رنگ ترسیم شود و هر گروه از ۵ پیکسل مجاور رنگ مشابهی داشته باشد، طیف رنگ تقریباً یکسان به نظر می‌رسد.

پرکاربردترین قالب تصویر با افت JPEG است (جدول ۲۷-۱) و در واقع خروجی تصویر بسیاری از دوربین‌های دیجیتال به صورت پیش فرض JPEG می‌باشد. فشرده‌سازی JPEG به‌طور شگفت‌انگیزی برای تصاویر عکاسی کارایی خوبی دارد، و کاهش قابل توجّه در حجم فایل اغلب با کاهش بسیار کمی در کیفیت تصویر همراه است. با این حال، زمانی که تصاویر

دارای لبه‌های تیز هستند، مانند لبه‌های ایجاد شده توسّط خطوط خطی یا متن، فشرده‌سازی JPEG با شکست مواجه می‌شود. در این موارد، فشرده‌سازی JPEG می‌تواند منجر به اختلالات بصری قابل توجهی شود (شکل ۲۷-۲).



شکل ۲۷-۲. نمایش اختلالات JPEG. (الف) یک تصویر چندین بار با استفاده از فشرده‌سازی شدید JPEG تکرار شده است. اندازهٔ فایل حاصل با متن قرمز بالای هر تصویر نشان داده شده است. کاهش حجم فایل با ضریب ۱۰، از ۴۳۲ کیلوبایت در تصویر اصلی به ۴۳ کیلوبایت در تصویر فشرده نهایی، تنها منجر به کاهش جزئی قابل درک در کیفیت تصویر شده است. با این حال، کاهش بیشتر در اندازه فایل با ضریب ۲، یعنی تنها ۲۵ کیلوبایت، منجر به اختلالات قابل مشاهده متعددی می‌شود. (ب) بزرگنمایی روی تصویر بسیار فشرده، اختلالات فشرده‌سازی مختلف را نشان می‌دهد. عکاس: Claus O. Wilke.

حتی اگر اختلالات JPEG به اندازهٔ کافی ظریف باشند به طوری که با چشم غیرمسلح قابل مشاهده نباشند، می‌توانند مثلاً در چاپ مشکل ایجاد کنند. بنابراین، ایدهٔ خوبی است که تا حد امکان از قالب JPEG اجتناب شود. به طور خاص، نباید از این قالب برای تصاویر حاوی نقاشی خط یا متن و نیز مصوّرسازی داده‌ها یا عکس از صفحه نمایش استفاده شود. قالب مناسب برای این تصاویر PNG یا TIFF است. ما از قالب JPEG به طور انحصاری برای تصاویر عکاسی استفاده می‌کنیم. اگر یک تصویر هم حاوی عناصر عکاسی و هم نقاشی خط یا متن باشد، همچنان باید از قالب PNG یا TIFF استفاده شود. بدترین حالت در مورد این قالب‌ها این است که حجم فایل‌های تصویری زیاد می‌شوند، در حالی که بدترین سناریو در مورد JPEG این است که محصول نهایی شما زشت به نظر برسد.

تبدیل قالب‌های تصویر

به طور کلی امکان تبدیل هر قالب تصویری به هر قالب تصویر دیگری وجود دارد. برای مثال، در مک، می‌توانید یک تصویر را باز کنید و سپس در تعدادی قالب مختلف دیگر خروجی بگیرید. با این حال، در این فرآیند، اطّلاعات مهم ممکن است از بین بروند و این اطّلاعات هرگز قابل بازیابی نیستند. به عنوان مثال، پس از ذخیره یک گرافیک بُرداری در قالب بیت مپ (مثلاً یک فایل PDF به صورت JPEG)، استقلال وضوح که یکی از ویژگی‌های کلیدی گرافیک بُرداری است از بین رفته است. برعکس، ذخیره یک تصویر JPEG در قالب فایل PDF به طور جادویی تصویر را به یک گرافیک بُرداری تبدیل نمی‌کند. تصویر همچنان یک تصویر بیت مپ خواهد بود که فقط در داخل فایل PDF ذخیره شده است. به طور مشابه، تبدیل یک فایل JPEG به یک فایل PNG هیچ کدام از اختلالاتی که ممکن است توسّط الگوریتم فشرده‌سازی JPEG اعمال شده باشد را حذف نمی‌کند.

بنابراین، یک قانون خوب این است که همیشه تصویر اصلی را در قالبی ذخیره کنید که حداکثر وضوح، دقّت و انعطاف‌پذیری را حفظ کند. بنابراین، برای ترسیم داده‌ها، یا نمودارهای خود را به صورت PDF ایجاد کنید و سپس در صورت لزوم آن‌ها را به PNG یا JPEG تبدیل کنید، یا آن‌ها را به عنوان PNG با وضوح بالا ذخیره کنید. به طور مشابه، برای تصاویری که فقط به صورت بیت مپ در دسترس هستند، مانند عکس‌های دیجیتال، آن‌ها را در قالبی ذخیره کنید که از فشرده‌سازی با افت استفاده نکند – یا اگر این کار امکان‌پذیر نیست، آن‌ها را تا حد امکان کم فشرده کنید. همچنین، تصاویر را تا حد امکان با وضوح بالا ذخیره کنید و در صورت نیاز مقیاس آن را کاهش دهید.

فصل ۲۸

انتخاب نرم‌افزار مصوّرسازی مناسب

در طول این کتاب، به طور هدفمند از یک سوال مهم در خصوص مصوّرسازی داده‌ها اجتناب کرده‌ایم: از چه ابزارهایی برای ترسیم نمودارهای خود استفاده کنیم؟ این سوال می‌تواند بحث‌های داغی ایجاد کند، زیرا بسیاری از افراد پیوندهای عاطفی قوی با ابزار خاصی که با آن‌ها آشنا هستند، دارند. اغلب دیده‌ایم که افراد به جای صرف زمان برای یادگیری یک رویکرد جدید، به شدت از ابزارهای ترجیحی خود دفاع می‌کنند، حتی اگر رویکرد جدید دارای مزایای واضحی باشد. ما معتقدیم که چسبیدن به ابزارهایی که می‌شناسید اصلاً غیر منطقی نیست. یادگیری ابزار جدید مستلزم صرف زمان و تلاش است، و شما باید یک دورهٔ انتقال دردناک را پشت سر بگذارید که در آن انجام کارها با ابزار جدید بسیار دشوارتر از ابزار قدیمی است. اینکه آیا گذراندن این دوره ارزش تلاش را دارد یا نه، معمولاً تنها پس از سرمایه‌گذاری برای یادگیری ابزار جدید و به صورت گذشته‌نگر قابل پاسخ است. بنابراین، صرف نظر از مزایا و معایب ابزارها و رویکردهای مختلف، اصل اساسی این است که شما باید ابزاری را انتخاب کنید که برای شما مناسب باشد. اگر بتوانید نموداری که می‌خواهید را بدون تلاش بیش از حد بسازید، این تنها مساله‌ای است که اهمیت دارد.

بهترین نرم‌افزار آن است که به شما امکان دهد نمودار مورد نظرتان را رسم کنید.

با این حال، ما معتقدیم می‌توان از یکسری اصول کلی برای تعیین ارزش نسبی رویکردهای مختلف برای تولید نمودار استفاده کنیم. این اصول عمدتاً مبتنی بر میزان تکرارپذیری نمودارها، سهولت کاوش سریع داده‌ها، و قابلیت تغییرپذیری ظاهر بصری خروجی می‌باشند.

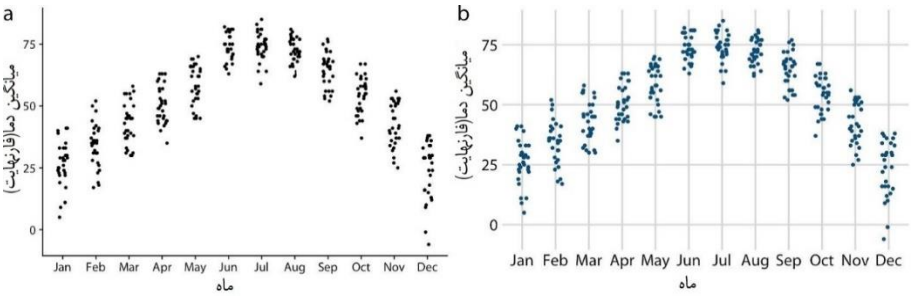
بازتولیدی[1] و تکرارپذیری[2]

در زمینه آزمایش‌های علمی، اگر یک گروه تحقیقاتی متفاوت همان نوع مطالعه را انجام دهند، در صورتی که یافته‌های علمی کلی کار بدون تغییر باقی بماند، کار را بازتولیدی می‌دانیم. به عنوان مثال، اگر یک گروه تحقیقاتی متوجّه شود که یک داروی ضد درد جدید بدون ایجاد عوارض جانبی قابل توجه، سردرد را به طور قابل توجّهی کاهش می‌دهد و گروه دیگری متعاقباً همان دارو را روی یک گروه بیمار متفاوت مطالعه کنند و یافته‌های یکسانی به دست آورند، کار بازتولید شده است. در مقابل، اگر نتایج اندازه‌گیری‌های بسیار مشابه یا یکسان توسّط همان فرد با همان روش اندازه‌گیری و توسّط همان تجهیزات قابل دستیابی باشد، کار تکرارپذیر است. به عنوان مثال، اگر سگم را وزن کنم و متوجّه شوم که ۴۱ پوند وزن دارد و سپس دوباره او را روی همان ترازو وزن کنم و دوباره متوجّه شوم که وزن او ۴۱ پوند است، این اندازه‌گیری تکرارپذیر است.

با تغییرات جزئی، می‌توانیم این مفاهیم را در مصوّرسازی داده‌ها اعمال کنیم. در صورتی که داده‌ها موجود بوده و هرگونه تبدیلی که ممکن است قبل از رسم روی داده‌ها اعمال گردیده باشد دقیقاً مشخّص باشد، تصویرسازی بازتولیدی است. به عنوان مثال، اگر یک نمودار رسم کنید و سپس داده‌هایی را که ترسیم کرده‌اید برای من بفرستید، می‌توانم نموداری تهیه کنم که اساساً مشابه باشد. ممکن است از قلم یا رنگ یا اندازه‌های نقطه متفاوتی برای نمایش داده‌های یکسان استفاده کرده باشیم، لذا ممکن است این دو نمودار دقیقاً یکسان نباشند، امّا نمودار شما و من پیام یکسانی را منتقل می‌کنند و بنابراین بازتولید یکدیگر هستند. از سوی دیگر، اگر بتوان همان ظاهر بصری را تا آخرین پیکسل، از داده‌های خام دوباره ایجاد کرد، این نمودار تکرارپذیر است. به بیان دقیق، تکرارپذیری مستلزم آن است که حتی اگر عناصر تصادفی مانند لرزانش (فصل ۱۸) در نمودار وجود داشته باشد، آن عناصر به روشی قابل تکرار مشخّص شده باشند و بتوان در آینده آن‌ها را دوباره تولید نمود. برای داده‌های تصادفی، تکرارپذیری عموماً مستلزم آن است که یک مولّد اعداد تصادفی خاص را مشخّص کنیم که برای آن کد اختصاصی تنظیم و ثبت کنیم.

1. Reproducibility
2. Repeatability

در سرتاسر این کتاب، نمونه‌های زیادی از نمودارهایی دیده‌ایم که بازتولید می‌شوند، امّا تکرار نمودارهای دیگر نیستند. به عنوان مثال، فصل ۲۵ چندین مجموعه از نمودارها را نشان می‌دهد که داده‌های یکسانی را نشان می‌دهند امّا تا حدودی متفاوت به نظر می‌رسند. به طور مشابه، نمودار ۲۸–۱ الف تا حد لرزانش تصادفی که به هر نقطه داده اعمال می‌شده تکرار نمودار ۹–۷ است، در حالی که شکل ۲۸–۱ب تنها بازتولید آن شکل است. نمودار ۲۸–۱ ب نسبت به نمودار ۹–۷ دارای لرزانش متفاوتی است و همچنین از طراحی بصری متفاوتی استفاده می‌کند تا حدی که این دو نمودار کاملاً متمایز به نظر می‌رسند، حتی اگر اطّلاعات یکسانی را در مورد داده‌ها منتقل کنند.



شکل ۲۸–۱. تکرار و بازتولید یک نمودار. قسمت (الف) تکرار نمودار ۹–۷ است. این دو نمودار تا حد لرزانش تصادفی که در هر نقطه اعمال شده، یکسان هستند. در مقابل، بخش (ب) یک بازتولید است امّا تکرار نیست. به طور خاص، لرزانش در قسمت (ب) با لرزانش در قسمت (الف) یا در نمودار ۹–۷ متفاوت است. منبع داده: آب و هوای زیرزمینی.

زمانی که ما با نرم‌افزارهای تعاملی کار می‌کنیم، دستیابی به تکرارپذیری و بازتولید می‌تواند بسیار سخت باشد. بسیاری از برنامه‌های تعاملی به شما امکان می‌دهند تا داده‌ها را تغییر داده یا به روش دیگری دست‌کاری کنید، امّا سوابق تغییراتی که انجام می‌شود را ثبت نمی‌کند. اگر با استفاده از این نوع برنامه یک نمودار بسازید، و سپس کسی از شما بخواهد که نمودار را بازتولید کنید یا یک نمودار مشابه با مجموعه دادهٔ متفاوت ایجاد کنید، ممکن است در انجام این کار مشکل داشته باشید. در طول سال‌هایی که در پسادکتری و استادیار جوان بودم، برای رسم تمام نمودارهای علمی از یک برنامه تعاملی استفاده می‌کردم و این موضوع دقیقاً چندین بار اتفاق افتاد. مثلاً چند نمودار برای یک دستنوشته علمی رسم کرده بودم. وقتی چند ماه بعد می‌خواستم دستنوشته را اصلاح کنم و لازم بود نسخه‌ای کمی تغییر یافته از یکی از نمودارها را بازتولید کنم، مطمئن نبودم که چگونه نمودار اصلی را رسم کرده‌ام. این تجربه به من

آموخت که تا حد امکان از برنامه‌های تعاملی دوری کنم. من اکنون با نوشتن کد (اسکریپت) و از طریق برنامه‌نویسی نمودار را از داده‌های خام تولید می‌کنم. نمودار ایجاد شده از طریق برنامه‌نویسی معمولاً توسّط هر کسی که به کدهای تولیدکننده و زبان برنامه‌نویسی و کتابخانه‌های خاص مورد استفاده دسترسی داشته باشد، قابل تکرار است.

کاوش داده در مقابل ارائه داده

دو مرحله مجزا برای تجسم داده‌ها وجود دارد که هر کدام الزامات بسیار متفاوتی دارند. اولی کاوش داده است. هر زمان که کار با یک مجموعه دادهٔ جدید را شروع می‌کنید، باید از زوایای مختلف به آن نگاه کنید و راه‌های مختلفی را برای نمایش آن امتحان کنید، تا درک درستی از ویژگی‌های کلیدی مجموعه داده داشته باشید. در این مرحله سرعت و کارامدی از اهمیت بالایی برخوردار است. شما باید انواع مختلف نمودار، روش‌های مختلف تبدیل و زیرمجموعه‌های متفاوت داده را امتحان کنید. هرچه سریعتر بتوانید با روش‌های مختلف به داده‌ها نگاه کنید، بیشتر کاوش خواهید کرد و احتمال اینکه متوجّه یک ویژگی مهم در داده‌ها شوید بیشتر است. مرحله دوّم ارائه داده‌ها است. هنگامی وارد این مرحله می‌شوید که مجموعه داده خود را درک کرده باشید و بدانید که چه جنبه‌هایی از آن را می‌خواهید به مخاطبان خود نشان دهید. هدف اصلی در این مرحله، تهیه یک نمودار با کیفیت بالا و آماده برای انتشار است که می‌تواند در یک مقاله یا کتاب چاپ شود، در یک ارائه گنجانده شود، یا در فضای مجازی ارسال شود.

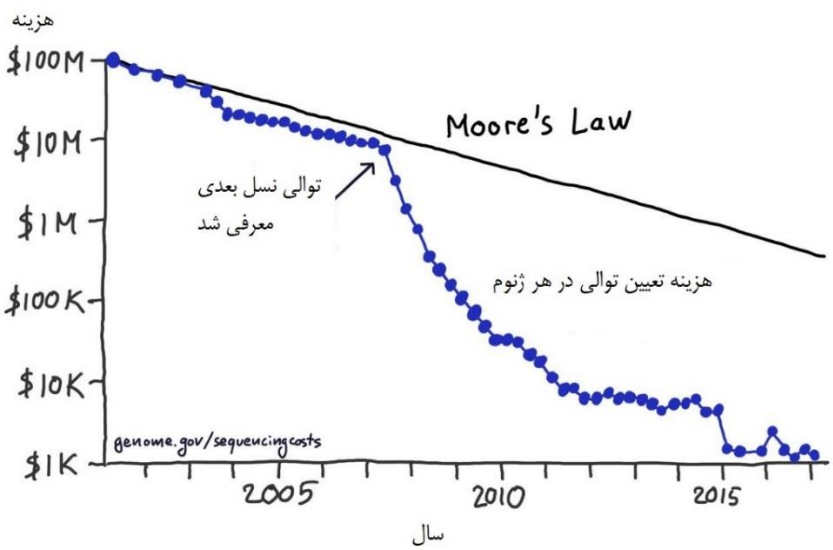
در مرحله اکتشاف، اینکه آیا نمودارهایی که می‌سازید جذّاب به نظر می‌رسند یا نه، در درجهٔ دوّم اهمیت قرار دارد. تا زمانی که بتوانید الگوهای مختلف را در داده‌ها ارزیابی کنید، اگر برچسب محورها را فراموش کنید، راهنما به هم ریخته باشد، یا نمادها خیلی کوچک باشند، اشکالی ندارد. با این حال، نکتهٔ حیاتی آن است که تغییر نحوهٔ نمایش داده‌ها چقدر برای شما آسان است. برای کاوش واقعی داده‌ها، باید بتوانید به سرعت از یک نمودار پراکنش به نمودارهای توزیع چگالی همپوشان و به نمودارهای جعبه‌ای و به یک نقشه حرارتی جهش کنید. در فصل ۲ دیدیم که تمام نمودارها شامل نگاشتی از داده‌ها به زیبایی‌شناسی است. یک ابزار کاوش داده که به خوبی طراحی شده است به شما این امکان را می‌دهد که به راحتی مشخّص کنید که کدام متغیّرها بر روی کدام الگوی زیبایی‌شناسی نگاشت شده‌اند و طیف گسترده‌ای از گزینه‌های مصوّرسازی مختلف را در یک چارچوب منسجم ارائه می‌دهد. با این حال، بر اساس تجربهٔ ما، بسیاری از ابزارهای مصوّرسازی (و به ویژه کتابخانه‌هایی برای تولید

نمودار مبتنی بر برنامه‌نویسی) بر این اساس تنظیم نشده‌اند. در عوض، آن‌ها بر اساس نوع نمودار سازماندهی می‌شوند، که در آن هر نوع نمودار به داده‌های ورودی متفاوتی نیاز داشته و رابط خاص خود را دارد. چنین ابزارهایی می‌توانند مانع کاوش کارآمد داده‌ها شوند، زیرا به خاطر سپردن نحوه عملکرد نمودارهای مختلف دشوار است. من شما را تشویق می‌کنم تا به دقّت ارزیابی کنید که آیا نرم‌افزار شما امکان کاوش سریع داده‌ها را فراهم می‌کند یا خیر. اگر نرم‌افزار این امکان را فراهم نمی‌کند، ممکن است از بررسی گزینه‌های جایگزین سود ببرید.

زمانی که مشخّص کردیم دقیقاً چگونه می‌خواهیم داده‌های خود را نمایش دهیم، چه تبدیل‌هایی را می‌خواهیم انجام دهیم، و از چه نوع نموداری استفاده کنیم، می‌توانیم نموداری با کیفیت بالا برای انتشار آماده کنیم. در این مرحله، چندین راه مختلف پیشِ رو داریم. اوّل اینکه می‌توانیم با استفاده از همان بستر نرم‌افزاری که برای اکتشاف اولیه استفاده کردیم، نمودار را نهایی کنیم. دوّم، می‌توانیم به بستری برویم که کنترل دقیق‌تری بر محصول نهایی به ما می‌دهد، حتی اگر آن بستر کاوش را سخت‌تر کند. سوّم، می‌توانیم پیش‌نویسی از نمودار با نرم‌افزار تولید کنیم و سپس به‌صورت دستی آن را با یک برنامه ویرایش تصویر یا مصوّرسازی مانند فتوشاپ یا ایلاستریتور اصلاح کنیم. چهارم، می‌توانیم به صورت دستی کل نمودار را از ابتدا دوباره ترسیم کنیم، یا با قلم و کاغذ یا با استفاده از یک برنامه مصوّرسازی.

همه این راه‌ها معقول هستند. با این حال، ما می‌خواهیم نسبت به ترسیم دستی نمودار در فرآیند تجزیه و تحلیل‌های معمول داده‌ها یا برای انتشارات علمی هشدار دهیم. آماده‌سازی دستی نمودار، تکرار یا بازتولید آن را بسیار دشوار و وقت‌گیر می‌کند. بر اساس تجربه‌مان از کار در علوم طبیعی، به ندرت فقط یک بار نموداری را رسم می‌کنیم. در طول یک مطالعه، ممکن است آزمایش‌ها را دوباره انجام دهیم، مجموعه داده اصلی را گسترش دهیم یا یک آزمایش را چندین بار با شرایط کمی متفاوت تکرار کنیم. بارها دیده‌ایم که در اواخر فرآیند انتشار، زمانی که فکر می‌کنیم همه چیز انجام شده و کار نهایی شده است، در نهایت یک اصلاح کوچک در نحوهٔ تجزیه و تحلیل داده‌ها مورد نیاز بوده و در نتیجه همه نمودارها باید دوباره ترسیم شوند. همچنین، در موقعیّت‌های مشابه، دیده‌ایم که تصمیم نویسندگان این بوده که تحلیل‌ها دوباره انجام نشود یا نمودارها دوباره ترسیم نشود، یا به دلیل تلاش‌های مورد نیاز یا به این دلیل که افرادی که نمودارهای اوّلیه را تنظیم کرده‌اند در دسترس نیستند. در تمام این سناریوها، یک فرآیند مصوّرسازی داده پیچیده غیرضروری و غیرقابل تکرار در تولید بهترین شواهد ممکن، اختلال ایجاد می‌کند.

با این حال، ما هیچ نگرانی اساسی در مورد نمودارهای ترسیم شده با دست یا نمودارهایی که به صورت دستی پس از رسم مثلاً برای تغییر برچسب محور، افزودن حاشیه‌نویسی، یا تغییر رنگ اصلاح شده‌اند، نداریم. این رویکردها می‌توانند نمودارهای زیبا و منحصربه‌فردی را خلق کنند که به هیچ روش دیگری نمی‌توان آن‌ها را ایجاد کرد. در واقع، همانطور که نمودارهای پیچیده و مرتّب تولید شده توسّط کامپیوتر به طور فزاینده‌ای در حال رواج هستند، می‌بینیم که نمودارهایی که به صورت دستی ترسیم شده‌اند تا حدودی در حال احیا شدن هستند (مثلاً به نمودار ۲۸-۲ مراجعه کنید). به نظر می‌رسد دلیل آن این مساله باشد که چنین نمودارهایی برداشتی منحصر به فرد و شخصی است که در غیر این صورت ممکن است به صورت یک ارائه عقیم و معمولی از داده‌ها رخ دهد.



نمودار ۲۸-۲. با معرفی روش‌های توالی‌یابی نسل بعدی، هزینه توالی‌یابی به ازای هر ژنوم بسیار سریعتر از پیش‌بینی قانون مور کاهش یافته است. این نمودار دستی تصویر معروفی که توسّط مؤسسه ملی بهداشت تهیه شده است را بازتولید می‌کند. منبع داده: موسسه ملی تحقیقات ژنوم انسانی.

جداسازی محتوا و طراحی

یک نرم‌افزار مصوّرسازی خوب باید به شما این امکان را بدهد که به طور جداگانه در مورد محتوا و طراحی نمودار خود فکر کنید. منظورمان از محتوا، مجموعه داده نشان داده شده، تبدیل‌های اعمال شده (در صورت وجود)، نگاشت‌های زیبایی‌شناسی داده‌ها، مقیاس‌ها، دامنه محورها، و نوع نمودار (پراکنش، خطی، میله‌ای، جعبه‌ای و غیره) می‌باشد. از سوی دیگر،

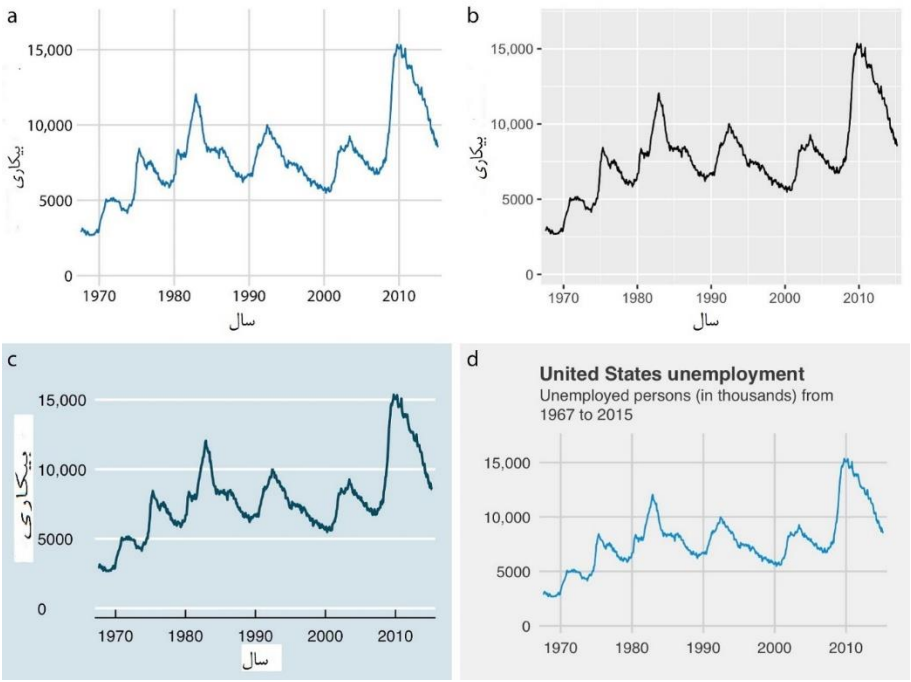
منظـور از طراحـی ویژگی‌هـایی ماننـد رنگ‌هـای پیش‌زمینـه و پس‌زمینـه، خصوصـیات قلـم (ماننـد اندازه، ظاهر و خانواده)، شـکل‌ها و انـدازه نمادهـا، اینکـه آیـا نمـودار دارای شـبکه پس‌زمینـه اسـت یـا نـه، و محـل قرارگیـری راهنمـا، تیک‌هـای محـور، عنـاوین محـور و عنـوان نمـودار می‌باشـد. وقتـی روی یک نمودار جدید کـار مـی‌کنیم، معمـولاً ابتـدا بـا اسـتفاده از کـاوش سـریعی کـه در بخـش قبـل توضیح داده شد، تعیین می‌کنیم کـه محتـوا چـه بایـد باشـد. هنگـامی کـه محتـوا تنظـیم شـد، ممکـن است طرح را تغییر دهـیم، یـا بـه احتمـال زیـاد طرحـی از پـیش تعریـف شـده‌ای کـه دوسـتش داریـم و/یا در بستر کل کار به نمودار ظاهر با ثباتی می‌دهد، را اعمال کنیم.

در ggplot2، نرم‌افزاری که بـرای ایـن کتـاب اسـتفاده کـرده‌ایم، تفکیـک محتـوا و طراحـی از طریـق قالب‌هـا بـه دسـت آمـده اسـت. قالـب ظـاهر بصـری یـک نمـودار را مشـخّص می‌کنـد و بـه راحتـی می‌تـوان یـک قالـب را روی نمودارهـای مختلـف اعمـال کـرد (نمـودار ۲۸-۳). قالب‌هـا می‌تواننـد توسّـط اشـخاص ثالـث نوشـته شـده و بـه صـورت بسـته‌های R توزیـع شـوند. از طریـق ایـن فراینـد، یک زیست بوم پـر رونـق از قالب‌هـای الحـاقی بـرای ggplot2 ایجـاد شـده اسـت و طیـف وسـیعی از سـبک‌ها و سـناریوهای کـاربردی مختلـف را پوشـش می‌دهـد. اگـر نمودارتـان را بـا ggplot2 می‌سـازید، تقریبـاً همـواره می‌توانیـد قالـب موجـودی کـه نیازهـای طراحـی شـما را بـرآورده کنـد، بیابید.

تفکیـک محتـوا و طراحـی بـه دانشـمندان و طراحـان اجـازه می‌دهـد تـا هـر یـک بـر بهتـرین حـوزه عملکـرد خـود تمرکـز کننـد. اکثـر دانشـمندان طـراح نیسـتند، و بنـابراین نگرانـی اصـلی آن‌هـا بایـد داده‌ها باشـد، نـه طراحـی نمـودار. بـه همـین ترتیـب، اکثـر طراحـان دانشـمند نیسـتند و بایـد بتواننـد بدون نگرانی در مـورد داده‌هـای خـاص، تبـدیل‌های مناسـب و غیـره، یـک زبـان بصـری منحصـر بـه فرد و جذّاب بـرای نمودارهـا ارائـه دهنـد. همـین اصـل تفکیـک محتـوا و طراحـی از دیربـاز در دنیـای انتشارات کتـاب، مجـلات، روزنامـه‌ها و وب‌سـایت‌ها دنبـال می‌شـود، جـایی کـه نویسـندگان محتـوا را ارائه می‌دهند، امّـا صـفحه‌آرایی و طراحـی توسّـط گـروه جداگانـه‌ای از افـراد متخصـص در ایـن زمینـه انجـام می‌شـود و تضـمین می‌کننـد کـه نشـریه بـا ظـاهر جـذّاب و متعـادلی منتشـر شـود. ایـن اصـل منطقی و مفید است، امّا هنوز در دنیای مصوّرسازی داده‌ها به آن اندازه گسترش نیافته است.

به طور خلاصه، هنگام انتخاب نرم‌افـزار خـود، بـه ایـن فکـر کنیـد کـه چقـدر راحـت می‌توانیـد نمـودار را بازتولید کرده و آن‌ها را با مجموعـه داده‌هـای بـه روز شـده یـا تغییـر داده شـده دوبـاره رسـم کنیـد، آیـا می‌توانیـد بـه سـرعت نمودارهـای مختلـف از یـک داده را کـاوش کنیـد، و تـا چـه حـد می‌توانیـد صرفاً طراحی بصری نمودار را بـه طـور جداگانـه تغییـر دهیـد. بـر اسـاس سـطح مهـارت و آشـنایی‌تان

با برنامه‌نویسی، ممکن است استفاده از ابزارهای مصوّرسازی مختلف در مراحل کاوش داده و ارائه داده‌ها مفید باشد و ممکن است ترجیح دهید بهینه‌سازی تصویر نهایی را به صورت تعاملی یا دستی انجام دهید. اگر مجبورید به صورت تعاملی و به ویژه با نرم‌افزاری که تمام تغییرات داده‌ها و ترفندهای بصری اعمال شده را رهگیری و ثبت نمی‌کند، نمودار را ترسیم کنید، فراموش نکنید تا از تمام مراحل ساخت هر نمودار به دقّت یادداشت‌برداری کنید، تا تمام کارهایتان قابلیّت بازتولید داشته باشد.



نمودار ۲۸–۳. تعداد افراد بیکار در ایالات متحده از سال ۱۹۷۰ تا ۲۰۱۵. همین نمودار با استفاده از چهار قالب مختلف ggplot2 نمایش داده شده است: (الف) قالب پیش فرض برای این کتاب. (ب) قالب پیش‌فرض ggplot2، نرم‌افزاری که برای ساختن تمام نمودارهای این کتاب استفاده کرده‌ایم. (ج) قالبی که نمودارهای نشان داده شده در اکونومیست را تقلید می‌کند. (د) قالبی که نمودارهای نشان داده شده توسّط FiveThirtyEight را تقلید می‌کند. FiveThirtyEight اغلب از برچسب‌های محوری به نفع عنوان و زیرنویس صرف نظر می‌کند، و بنابراین ما نمودار را بر این اساس تنظیم کرده‌ایم. منبع داده: اداره آمار کار ایالات متحده.

فصل ۲۹

بیان داستان و انتقال مفهوم مورد نظر

اغلب، ترسیم داده‌ها با هدف برقراری ارتباط انجام می‌شود. ما بینشی در مورد یک مجموعه داده داریم و مخاطبان بالقوه‌ای نیز وجود دارند و می‌خواهیم بینش خود را به مخاطبان منتقل کنیم. برای انتقال موفقیت‌آمیز بینش خود، باید داستانی معنادار و هیجان‌انگیز را به مخاطب ارائه کنیم. نیاز به یک داستان ممکن است برای دانشمندان و مهندسان آزاردهنده به نظر برسد، زیرا احتمالاً آن را با ساختن چیزها، چرخش چیزها یا نتایج فروش بیش از حد برابر بدانند. با این حال، این دیدگاه نقش مهمی را که داستان‌ها در استدلال و حافظه ایفا می‌کنند، از قلم می‌اندازد. وقتی یک داستان خوب می‌شنویم هیجان‌زده می‌شویم و وقتی داستان بد است یا وقتی اصلاً داستانی وجود ندارد خسته می‌شویم. علاوه بر این، هر ارتباطی داستانی را در ذهن مخاطب ایجاد می‌کند. اگر خودمان یک داستان واضح ارائه نکنیم، مخاطب ما داستان خود را می‌سازد. در بهترین حالت، داستانی که آن‌ها می‌سازند به طور منطقی به دیدگاه ما از مطالب ارائه شده نزدیک است. با این حال، وضعیّت اغلب بسیار بدتر است. داستان ساخته شده می‌تواند «این خسته کننده است»، «نویسنده اشتباه می‌کند» یا «نویسنده بی‌کفایت است» باشد.

هدف شما از گفتن داستان باید استفاده از حقایق و استدلال منطقی باشد تا مخاطبان خود را علاقمند و هیجان‌زده کنید. بگذارید داستانی درباره فیزیکدان نظریه‌پرداز، استیون هاوکینگ برایتان تعریف کنیم. او در سن ۲۱ سالگی و یک سال پس از اخذ مدرک دکترای خود به

بیماری نورون حرکتی مبتلا شد و فرصت زندگی‌اش ۲ سال بود. هاوکینگ تمام انرژی خود را صرف گسترش علم کرد. او در نهایت تا ۷۶ سالگی زندگی کرد و یکی از بهترین فیزیکدان‌های عصر خود شد. او تمام کارهای شاخص خود را زمانی که به شدّت ناتوان بود انجام داد. ما معتقدیم این یک داستان جذّاب و در عین حال مبتنی بر حقایق و واقعیّت است.

داستان چیست؟

قبل از اینکه بتوانیم در مورد راهبردهای تبدیل ترسیم‌ها به داستان بحث کنیم، باید بفهمیم که یک داستان در واقع چیست. داستان مجموعه‌ای از مشاهدات، حقایق یا رویدادهای واقعی یا اختراعی است که به ترتیب خاصی ارائه می‌شود، به طوری که واکنش احساسی را در مخاطب ایجاد می‌کند. واکنش عاطفی از طریق ایجاد تنش در ابتدای داستان و سپس نتیجه آن تا انتهای داستان ارائه می‌شود. به حرکت از تنش تا نتیجه *قوس داستانی*[1] اطلاق می‌شود، و هر داستان خوب یک قوس داستانی واضح و قابل شناسایی دارد.

نویسندگان باتجربه می‌دانند که الگوهای استانداردی برای داستان‌سرایی وجود دارد که با طرز فکر انسان‌ها همخوانی دارد. برای مثال، می‌توانیم یک داستان را با استفاده از قالب آغاز–چالش–اقدام–نتیجه تعریف کنیم. در واقع این قالبی است که برای داستان هاوکینگ استفاده نمودیم. داستان با معرفی موضوع، فیزیکدان استیون هاوکینگ، شروع شد. در مرحله بعد چالش ارائه شد: تشخیص بیماری نورون حرکتی در سن ۲۱ سالگی. سپس اقدام ارائه شد: تعهّد شدید او به علم. سرانجام نتیجه را ارائه کردیم، هاوکینگ زندگی طولانی و موفقی داشت و در نهایت به یکی از تاثیرگذارترین فیزیکدانان زمان خود تبدیل شد. قالب‌های داستانی دیگر نیز معمولاً استفاده می‌شوند. مقالات روزنامه اغلب از قالب سرنخ–توسعه–نتیجه پیروی می‌کنند، یا حتی کوتاه‌تر، فقط سرنخ–توسعه، که در آن سرنخ، نکتهٔ اصلی را بیان می‌کند و مطالب بعدی جزئیات بیشتری را ارائه می‌دهند. اگر بخواهیم داستان هاوکینگ را در این قالب تعریف کنیم، می‌توانیم با این جمله شروع کنیم: «فیزیکدان تأثیرگذار استیون هاوکینگ، که درک ما از سیاه‌چاله‌ها و کیهان‌شناسی را متحول کرد، ۵۳ سال بیشتر از پیش‌بینی پزشکان خود عمر کرد. تأثیرگذارترین کارهای او زمانی بود که به شدت ناتوان بود.» این سرنخ است. در توسعه، می‌توانیم شرح عمیق‌تری از زندگی، بیماری و وفاداری هاوکینگ به علم را دنبال کنیم. قالب دیگر اقدام–پس‌زمینه–توسعه– اوج–پایان است که داستان را کمی سریع‌تر از قالب

1. story arc

آغاز–چالش–اقدام–نتیجه بیان می‌کند امّا نه به سرعت قالب سرنخ–توسعه. در این قالب، می‌توانیم با جمله‌ای مانند «استیون هاوکینگ جوان که با یک وضعیّت ناتوان‌کننده و احتمال مرگ زودهنگام مواجه بود، تصمیم گرفت تمام تلاش خود را در حوزهٔ علم بکار گیرد، و مصمّم بود تا هر زمانی که می‌تواند تاثیرگذار باشد» هدف این قالب جلب توجّه مخاطب و ایجاد ارتباط عاطفی زودهنگام است، امّا نتیجه نهایی را بیان نمی‌کند.

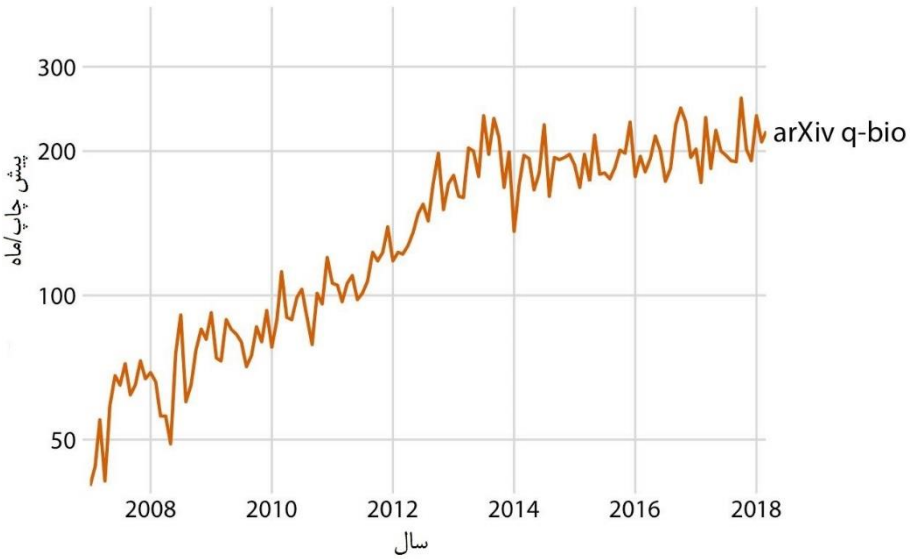
هدف ما در این فصل این نیست که این قالب‌های استاندارد داستان‌سرایی را با جزئیات بیشتر توصیف کنیم. منابع بسیار خوبی وجود دارد که این مطالب را پوشش می‌دهد. برای دانشمندان و تحلیل‌گران، کتاب «علم نوشتن[1]» به قلم جاشوا شیمل [Schimel 2011] توصیه می‌شود. در عوض، می‌خواهیم در مورد اینکه چگونه می‌توان مصوّرسازی داده‌ها را در قوس داستان بیاوریم، بحث کنیم. مهم‌تر از همه، باید بدانیم که یک ترسیم واحد (ایستا) به ندرت کل داستان را بیان می‌کند. یک ترسیم ممکن است آغاز، چالش، اقدام یا نتیجه را به تصویر بکشد، امّا بعید است که تمام این بخش‌های داستان را همزمان در برگیرد. برای بیان یک داستان کامل، معمولاً به ترسیم‌های متعدد نیاز داریم. به عنوان مثال، هنگام ارائه یک سخنرانی، ممکن است ابتدا مقداری پیش‌زمینه یا مطالب انگیزشی را نشان دهیم، سپس شکلی که چالش ایجاد می‌کند، و در نهایت شکل دیگری که نتیجه را ارائه می‌دهد. به همین ترتیب، در یک مقاله تحقیقاتی، ما ممکن است مجموعه‌ای از نمودارها را ارائه کنیم که به طور مشترک یک قوس داستانی قانع‌کننده ایجاد می‌کنند. با این حال، می‌توان کل یک قوس داستان را در یک شکل واحد تجمیع کرد. چنین شکلی باید همزمان شامل یک چالش و یک نتیجه باشد و با قوس داستانی که حاوی بخش آغاز است قابل مقایسه می‌باشد.

برای ارائه یک مثال عینی از گنجاندن اَشکال در داستان، اکنون داستانی را بر اساس دو شکل تعریف می‌کنیم. اوّلی چالش را ایجاد می‌کند و دوّمی به عنوان نتیجه عمل می‌کند. زمینه داستان رشد پیش‌چاپ‌ها[2] در علوم زیستی است (به فصل ۱۳ نیز مراجعه کنید). پیش‌چاپ‌ها دست‌نوشته‌هایی هستند که به صورت پیش‌نویس بوده و دانشمندان قبل از داوری همتا و انتشار رسمی، آن‌ها را با همکاران خود به اشتراک می‌گذارند. از زمانی که دست‌نوشته‌های علمی وجود داشته‌اند، دانشمندان پیش‌نویس‌های آن‌ها را به اشتراک می‌گذاشته‌اند. با این حال، در اوایل دهه ۱۹۹۰ و با ظهور اینترنت، فیزیکدانان دریافتند که ذخیره و توزیع پیش‌نویس دست‌نوشته‌ها در یک مخزن مرکزی بسیار کارآمدتر است. آن‌ها سرور پیش‌چاپ را اختراع

1. Writing Science
2. preprint

کردنـد، یـک وب سـرور کـه در آن دانشـمندان می‌تواننـد پیش‌نـویس دست‌نوشـته‌ها را بارگـذاری، دریافت و جستجو کنند.

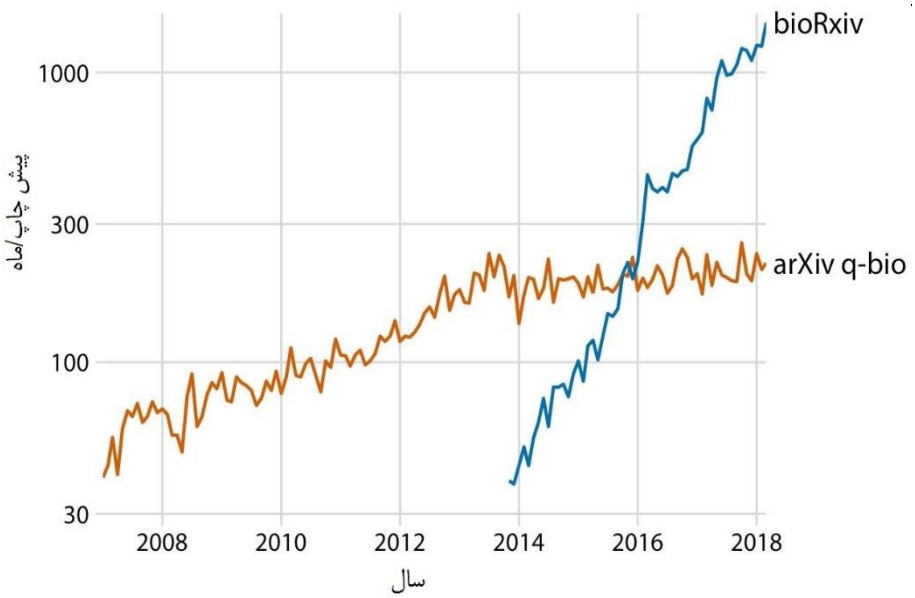
سـرور پیش‌چـاپی کـه فیزیکـدانان توسـعه داده‌انـد و هنـوز از آن اسـتفاده می‌کننـد، arXiv.org نامیـده می‌شـود. مـدّت کوتـاهی پـس از تأسـیس، arXiv.org شـروع بـه گسـترش کـرد و محبوبیّـت آن در زمینـه‌های علـوم مـرتبط، از جملـه ریاضـیات، نجـوم، علـوم کـامپیوتر، آمـار، اقتصـاد و زیست‌شناسـی افـزایش پیـدا کـرد. در اینجـا، ارسـال‌های پیش‌چـاپ بـه بخـش زیست‌شناسـی arXiv.org (q-bio) را بررسـی مـی‌کنیم. تعـداد ارسـال‌های ماهانـه از سـال ۲۰۰۷ تـا اواخـر ۲۰۱۳ به طور تصـاعدی افـزایش یافـت، امّـا ناگهـان ایـن رشـد متوقـف شـد (شـکل ۲۹–۱). بایـد در اواخـر سـال ۲۰۱۳ اتفـاقی افتـاده باشـد کـه چشـم‌انداز ارسـال‌های پیش‌چـاپ در حـوزهٔ زیست‌شناسـی را بـه طور اساسی تغییر داد. چه چیزی باعث این تغییر شدید در رشد ارسال‌ها شد؟



شکل ۲۹–۱. رشد ارسال‌های ماهانه به بخش زیست‌شناسی (q-bio) سرور پیش‌چاپ arXiv.org. یک گذار شدید در نرخ رشد را می‌توان در حدود سال ۲۰۱۴ مشاهده کرد. در حالی که رشد تا سال ۲۰۱۴ سریع بود، تقریباً هیچ رشدی از سال ۲۰۱۴ تا ۲۰۱۸ رخ نداد. توجّه داشته باشید که محور y لگاریتمی است، بنابراین افزایش خطی در y منطبق با رشد نمایی در ارسال‌های پیش‌چاپ می‌باشد. منبع داده: جردن آنایا، http://www.prepubmed.org/.

استدلال ما ایـن اسـت کـه اواخـر سـال ۲۰۱۳ نقطـهٔ زمـانی اسـت کـه پیش‌چاپ‌هـا در زیست‌شناسـی بـه اوج رسـیدند، و از قضـا ایـن باعـث شـد کـه آرشـیو q-bio رشـد خـود را آهسـته کنـد. در نـوامبر

۲۰۱۳، سرور پیش‌چاپ مخصوص زیست‌شناسی bioRxiv توسّط انتشارات CSHL[1] راه‌اندازی شد. CSHL Press ناشری است که در بین زیست‌شناسان بسیار مورد قبول است. پشتوانه CSHL Press کمک بزرگی به پذیرش پیش‌چاپ‌ها به طور کلی و bioRxiv به طور خاص در بین زیست‌شناسان نمود. همان زیست‌شناسانی که به arXiv.org کاملاً مشکوک بودند، با bioRxiv بسیار راحت بودند. در نتیجه bioRxiv به سرعت در بین زیست‌شناسان مقبولیت یافت، به درجه‌ای که arXiv هرگز موفق به این کار نشده بود. در واقع، bioRxiv بلافاصله پس از راه‌اندازی رشد سریع و تصاعدی در ارسال‌های ماهانه را تجربه کرد، و کاهش سرعت ارسال‌های q-bio دقیقاً با شروع رشد نمایی bioRxiv همزمان است (شکل ۲۹–۲). به نظر می‌رسد که بسیاری از زیست‌شناسان که می‌خواستند پیش‌چاپی را به q-bio ارسال نمایند، تصمیم گرفتند در عوض آن را به bioRxiv ارسال کنند.



شکل ۲۹–۲. کاهش رشد ارسال به q-bio همزمان با معرفی سرور bioRxiv بود. رشد ارسال‌های ماهانه به بخش q-bio سرور پیش‌چاپ با هدف عمومی arXiv.org و سرور اختصاصی پیش‌چاپ زیست‌شناسی bioRxiv نشان داده شده است. سرور bioRxiv در نوامبر ۲۰۱۳ فعال شد و از آن زمان میزان ارسال به آن به طور تصاعدی افزایش یافته است. به نظر می‌رسد که بسیاری از زیست‌شناسان که می‌خواستند پیش‌چاپی را به q-bio ارسال نمایند، تصمیم گرفتند در عوض آن را به bioRxiv ارسال کنند. منبع داده: Jordan Anaya، http://www.prepubmed.org

1. Cold Spring Harbor Laboratory

این داستان ما در مورد پیش‌چاپ در زیست‌شناسی است. ما عمداً آن را با دو شکل گفتیم، در حالی که اوّلی (شکل ۲۹-۱) به طور کامل در دوّمی گنجانده شده است (شکل ۲۹-۲). به نظر می‌رسد این داستان بیشترین تاثیر را زمانی دارد که به دو قسمت شکسته شود و لذا در یک سخنرانی آن را به این صورت ارائه خواهیم کرد. با این حال، شکل ۲۹-۲ به تنهایی می‌تواند برای بیان کل داستان استفاده شود، و نسخهٔ تک شکلی ممکن است برای رسانه‌هایی که مخاطبان دامنه توجّه کوتاهی دارد (مانند رسانه‌های اجتماعی) مناسب‌تر باشد.

برای ژنرال‌ها شکل بسازید

در ادامهٔ این فصل، در مورد راهبردهایی برای ساختن شکل‌های منفرد و مجموعه‌هایی از شکل‌ها بحث خواهیم کرد که به مخاطب شما کمک می‌کند تا با داستان شما ارتباط برقرار کند و در کل قوس داستان شما درگیر بماند. اوّل و مهم‌تر از همه، باید به مخاطبان خود اشکالی را نشان دهید که بتوانند درک کنند. کاملاً امکان‌پذیر است که تمام توصیه‌هایی که در سراسر این کتاب ارائه شده را دنبال کنید و همچنان اشکالی تهیه کنید که گیج‌کننده باشند. وقتی این اتفّاق می‌افتد، ممکن است قربانی دو تصوّر اشتباه رایج شده باشید: اوّل اینکه مخاطب می‌تواند نمودارهای شما را ببیند و فوراً نکاتی را که می‌خواهید بیان کنید استنباط کند. دوّم، اینکه مخاطب می‌تواند به سرعت ترسیم‌های پیچیده را پردازش کند و روندها و ارتباط‌های کلیدی را درک کند. هیچ یک از این فرضیّات درست نیست. ما باید تمام تلاش خود را انجام دهیم تا خوانندگان معنای ترسیم‌های ما را درک کنند و همان الگوهایی را در داده‌ها ببینند که ما می‌بینیم. این معمولاً به این معنی است که هرچه کمتر باشد، بهتر است. اشکال خود را تا حد امکان ساده کنید. تمام ویژگی‌های حاشیه‌ای داستان خود را حذف کنید. فقط نکات مهم باید باقی بماند. ما از این مفهوم به عنوان «شکل‌سازی برای ژنرال‌ها» یاد می‌کنیم.

من چندین سال مسئول یک پروژه تحقیقاتی بزرگ بودم که بودجهٔ آن توسّط ارتش آمریکا تامین می‌شد. برای گزارش‌های پیشرفت سالانه‌مان، مدیران برنامه به من دستور دادند که اشکال زیادی را درج نکنم، و هر شکلی که ارائه می‌کنم باید به وضوح نشان دهد که پروژه ما چگونه موفق بوده است. مدیران برنامه به من گفتند که یک ژنرال باید بتواند به هر شکل نگاه کند و فوراً ببیند که چگونه کاری که ما انجام می‌دادیم بر اساس توانایی‌های قبلی، بهبود یافته یا از آن فراتر رفته است. با این حال، زمانی که همکارانم که بخشی از این پروژه بودند، اشکالی را برای گزارش پیشرفت سالانه برای من ارسال کردند، بسیاری از اشکال این معیار را

برآورده نمی‌کردند. اشکال معمولاً بیش از حد پیچیده بودند، با عبارات فنی گیج‌کننده برچسب‌گذاری شده بودند، یا اصلاً به هیچ نکته واضحی اشاره نمی‌کردند. اکثر دانشمندان برای ایجاد اشکال برای ژنرال‌ها آموزش ندیده‌اند.

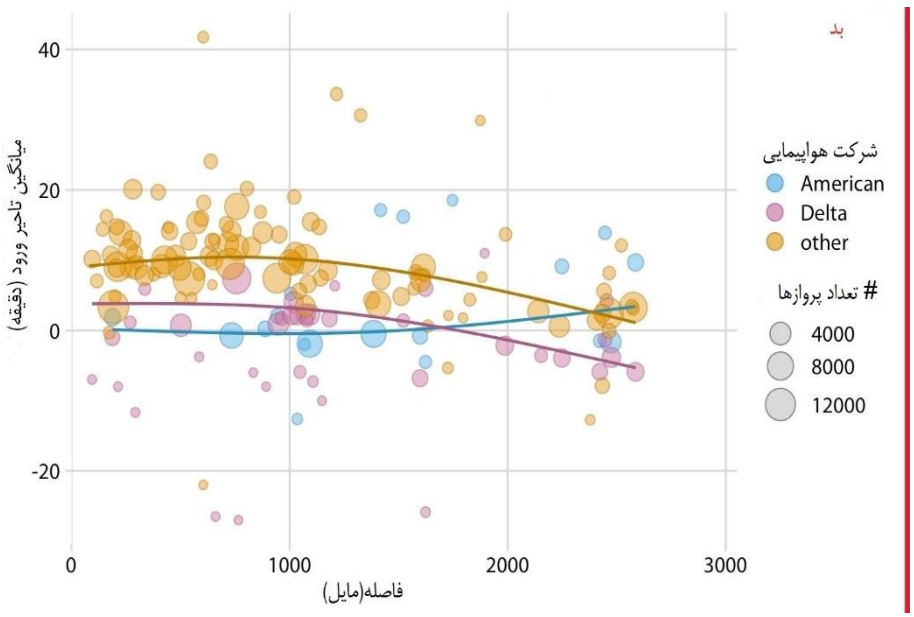
هرگز تصوّر نکنید که مخاطب شما می‌تواند تصاویر پیچیده را به سرعت پردازش کند.

برخی ممکن است این داستان را بشنوند و به این نتیجه برسند که ژنرال‌ها خیلی باهوش نیستند یا خیلی از مطالب علمی سر در نمی‌آورند. امّا این پیام اشتباه است. ژنرال‌ها صرفاً خیلی سرشان شلوغ است. آن‌ها نمی‌توانند ۳۰ دقیقه را صرف رمزگشایی یک نمودار مرموز کنند. وقتی میلیون‌ها دلار از وجوه مالیات‌دهندگان را به دانشمندان می‌دهند تا تحقیقات بنیادی انجام دهند، کمترین چیزی که در ازای آن می‌توانند انتظار داشته باشند نمودارهای واضحی است که نشان دهد کاری ارزشمند و جالب انجام شده است. البته این داستان نباید صرفاً در مورد بودجه نظامی تعبیر شود. ژنرال استعاره از هر کسی است که می‌خواهید ترسیم خود را به او ارائه دهید: داور علمی برای مقاله یا پیش‌نویس درخواست بودجه، سردبیر روزنامه، سرپرست شما یا رئیس سرپرست شما در شرکتی که در آن کار می‌کنید. اگر می‌خواهید داستان شما تاثیرگذار باشد، باید شکلی را بسازید که برای ژنرال‌هایتان مناسب باشد.

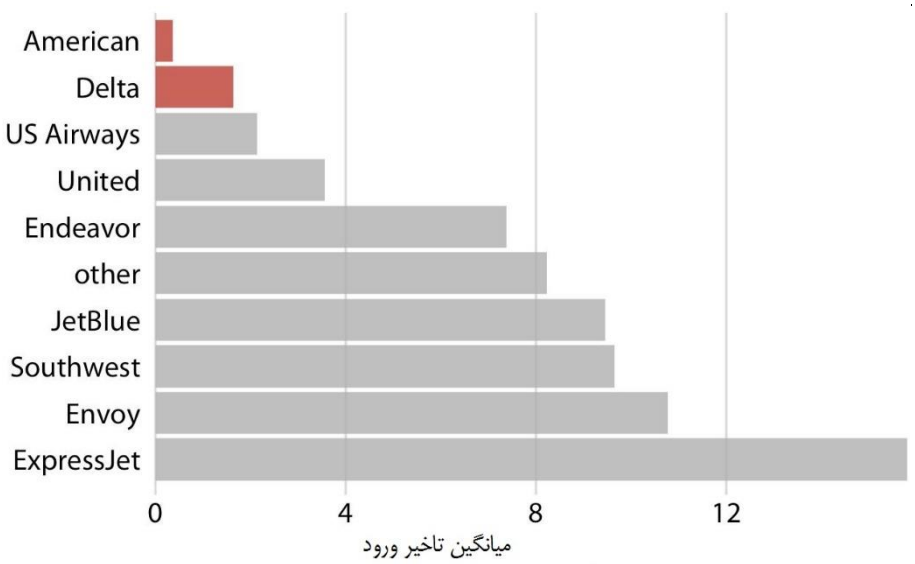
اوّلین چیزی که مانع شکل‌سازی برای ژنرال می‌شود، از قضا، سهولتی است که نرم‌افزارهای مدرن به ما می‌دهد تا ترسیم‌های پیچیده داده را انجام دهیم. با قدرت ترسیم تقریباً بی حد و حصر، تجمیع ابعاد بیشتری از داده‌ها وسوسه‌انگیز می‌شود. در واقع، در دنیای مصوّرسازی داده‌ها شاهد روندی از تمایل به سمت پیچیده‌ترین و چندوجهی‌ترین ترسیم‌ها هستیم. این ترسیم‌ها ممکن است بسیار چشمگیر به نظر برسند، امّا بعید است که داستان معناداری را بیان کنند. شکل ۲۹-۳ را در نظر بگیرید، این شکل تاخیرهای ورود را برای همهٔ پروازهایی که در سال ۲۰۱۳ از شهر نیویورک حرکت می‌کنند، نشان می‌دهد.

به نظر ما مهمترین ویژگی شکل ۲۹-۳ این است که امریکن و دلتا کمترین تاخیر ورود را دارند. این بینش در یک نمودار میله‌ای ساده بسیار بهتر منتقل می‌شود (شکل ۲۹-۴). بنابراین، شکل ۲۹-۴ شکل صحیحی برای نشان دادن این است که آیا داستان مربوط به تاخیرهای ورود خطوط هوایی است، حتی اگر ایجاد آن نمودار مهارت‌های ترسیم داده‌های شما را به

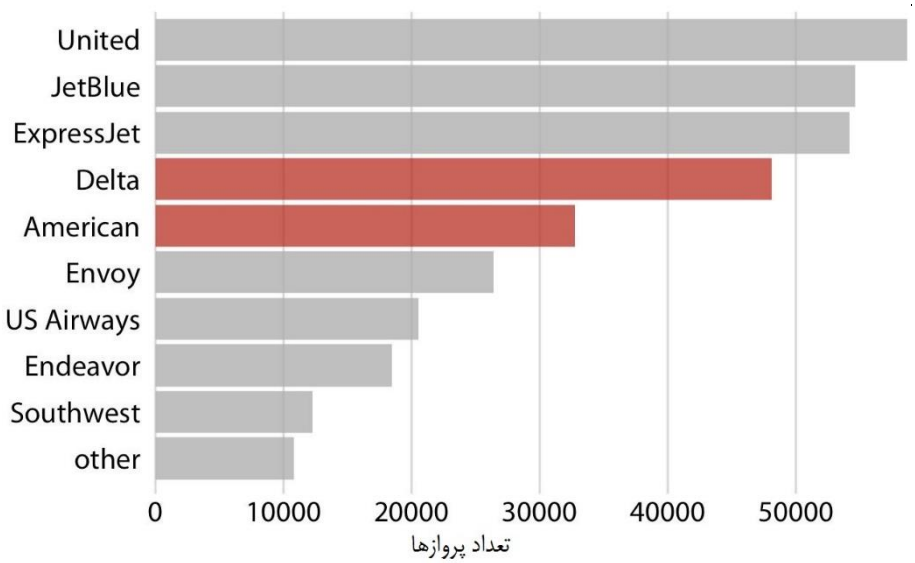
چالش نکشد. اگر از خود می‌پرسید که آیا این خطوط هوایی تاخیرهای کمی دارند زیرا آن‌قدرها به خارج از شهر نیویورک پرواز نمی‌کنند، می‌توانید نمودار ستونی دومی را ارائه دهید که نشان می‌دهد هر دو شرکت آمریکایی و دلتا حامل‌های اصلی در این منطقه هستند (شکل ۵-۲۹). هر دوی این نمودارهای میله‌ای متغیّر فاصله نشان داده شده در شکل ۲۹-۳ را کنار می‌گذارند. این مساله مشکلی ندارد. ما نیازی به ترسیم ابعادی از داده که در حاشیهٔ داستان هست، نداریم، حتی اگر آن‌ها را داشته باشیم و حتی اگر بتوانیم شکلی بسازیم که آن‌ها را نشان می‌دهد. ساده و واضح بهتر از پیچیده و گیج‌کننده است. زمانی که سعی می‌کنید داده‌های زیادی را به صورت همزمان نشان دهید، ممکن است در نهایت اصلاً چیزی را نشان ندهید.



شکل ۲۹-۳. میانگین تاخیر ورود در مقابل فاصله از شهر نیویورک. هر نقطه نشان‌دهندهٔ یک مقصد است، و اندازه هر نقطه تعداد پروازهای یکی از سه فرودگاه اصلی شهر نیویورک (نیوارک، جی‌افکی، یا لاگواردیا) به آن مقصد را در سال ۲۰۱۳ نشان می‌دهد. تاخیرهای منفی نشان می‌دهد که پرواز زودتر رسیده است. خطوط ممتد نشان‌دهندهٔ روند میانگین بین تاخیر ورود و مسافت طی شده است. دلتا بدون در نظر گرفتن مسافت طی شده، به طور مداوم تاخیر کمتری نسبت به سایر خطوط هوایی دارد. امریکن به طور متوسط برای مسافت‌های کوتاه کمترین تاخیر را دارد، امّا برای مسافت‌های طولانی‌تر، بیشترین تاخیر را دارد. این نمودار به عنوان «بد» برچسب‌گذاری شده است زیرا بیش از حد پیچیده است. برای اکثر خوانندگان این نمودار گیج‌کننده است و به طور شهودی متوجّه نمی‌شوند که این شکل چه چیزی را نشان می‌دهد. منبع داده: اداره حمل و نقل ایالات متحده، اداره آمار حمل و نقل.



شکل ۲۹–۴. میانگین تاخیر ورود برای پروازهای خروجی از شهر نیویورک در سال ۲۰۱۳، به تفکیک خطوط هوایی. امریکن و دلتا کمترین میانگین تأخیر ورود را در بین تمام خطوط هوایی که از شهر نیویورک پرواز می‌کنند، دارند. منبع داده: اداره حمل و نقل ایالات متحده، اداره آمار حمل و نقل.

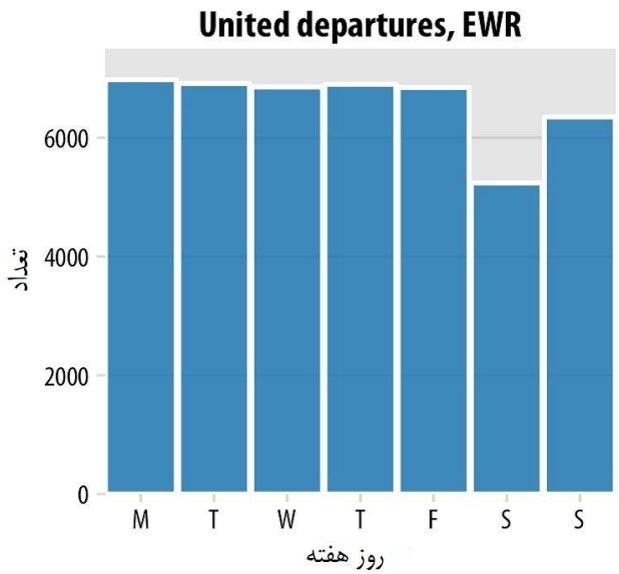


شکل ۲۹–۵. تعداد پروازهای خروجی از شهر نیویورک در سال ۲۰۱۳، به تفکیک خطوط هوایی. دلتا و آمریکن چهارمین و پنجمین شرکت حمل و نقل بزرگ با پروازهای خروجی از شهر نیویورک هستند. منبع داده: اداره حمل و نقل ایالات متحده، اداره آمار حمل و نقل.

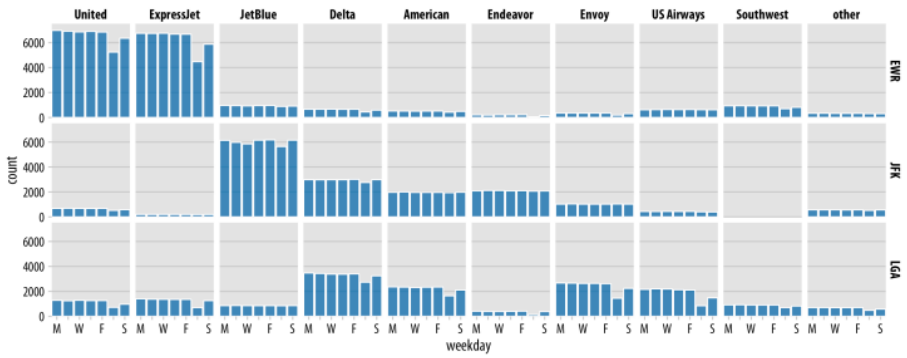
حرکت به سمت اشکال پیچیده

با این حال، گاهی اوقات می‌خواهیم اشکال پیچیده‌تری را نشان دهیم که حاوی مقدار زیادی اطّلاعات هستند. در این موارد، اگر ابتدا یک نسخهٔ ساده شده از شکل را قبل از نمایش نسخه نهایی با پیچیدگی کامل نشان دهیم، می‌توانیم کار را برای خوانندگان آسان‌تر کنیم. همین رویکرد برای ارائه‌ها نیز به شدت توصیه می‌شود. هرگز مستقیماً به سمت یک شکل بسیار پیچیده نروید. ابتدا زیرمجموعه‌ای از اطّلاعات که به راحتی قابل درک می‌باشند را نشان دهید.

این توصیه به ویژه در صورتی مصداق دارد که شکل نهایی یک نمودار چندگانه‌های کوچک باشد (فصل ۲۱) که شبکه‌ای از نمودارهای فرعی را با ساختار مشابه نشان می‌دهد. اگر مخاطب ابتدا یک طرح فرعی را به تنهایی دیده باشد، درک شبکه کامل بسیار آسان‌تر است. به عنوان مثال، شکل ۲۹-۶ تعداد کل خروجی خطوط هوایی یونایتد از فرودگاه نیوآرک (EWR) در سال ۲۰۱۳ را به تفکیک روزهای هفته نشان می‌دهد. هنگامی که این شکل را دیدیم و درک کردیم، پردازش اطّلاعات مرتبط برای ۱۰ شرکت هواپیمایی و ۳ فرودگاه به طور همزمان بسیار آسان‌تر است (شکل ۲۹-۷).



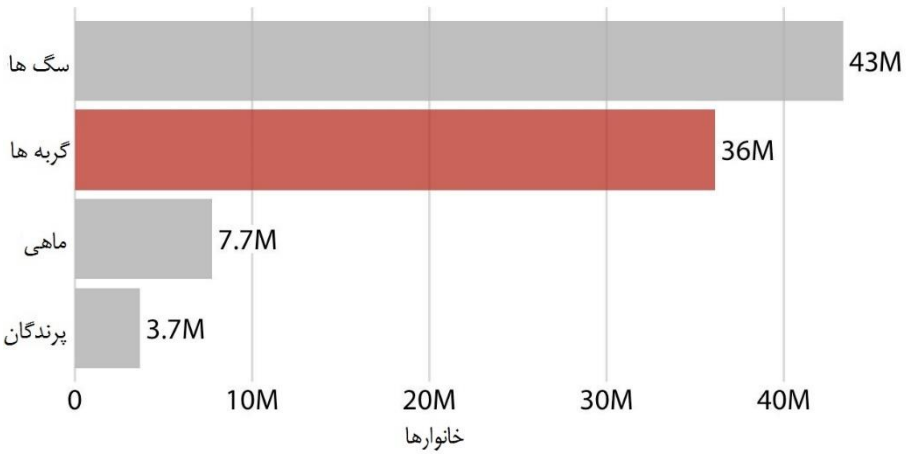
شکل ۲۹-۶. پروازهای خروجی خطوط هوایی یونایتد در سال ۲۰۱۳ از فرودگاه نیوآرک (EWR) به تفکیک روزهای هفته. اکثر روزهای هفته تقریباً تعداد حرکت یکسانی را نشان می‌دهند، اما در آخر هفته‌ها تعداد حرکت‌های کمتری وجود دارد. منبع داده: اداره حمل و نقل ایالات متحده، اداره آمار حمل و نقل



شکل ۲۹–۷. پروازهای خروجی در شهر نیویورک در سال ۲۰۱۳، به تفکیک خطوط هوایی، فرودگاه و روزهای هفته. خطوط هوایی یونایتد و اکسپرس جت بیشتر پروازهای خروجی از فرودگاه نیوآرک (EWR) را تشکیل می‌دهند. جت بلو، دلتا، امریکن و اندور اکثر خروجی های جان اف کندی را تشکیل می‌دهند. و دلتا، امریکن، انوی و ایرویز بیشتر پروازهای خروجی لاگاردیا (LGA) را تشکیل می‌دهند. بسیاری از خطوط هوایی، امّا نه همهٔ آن‌ها، در تعطیلات آخر هفته پروازهای کمتری نسبت به روزهای کاری دارند. منبع داده: اداره حمل و نقل ایالات متحده، اداره آمار حمل و نقل.

نمودارهای خود را به یاد ماندنی کنید

شکل‌های ساده و واضح مانند نمودارهای میله‌ای ساده این مزیّت را دارند که مانع حواس‌پرتی می‌شوند، به راحتی خوانده می‌شوند و به مخاطب شما اجازه می‌دهند تا روی مهم‌ترین نکاتی که می‌خواهید بیان کنید، تمرکز کند. با این حال، سادگی ممکن است با یک نقطهٔ ضعف همراه باشد: اشکال ممکن است عمومی شود. آن‌ها هیچ ویژگی مشخصی ندارند که آن‌ها را خاص و به یاد ماندنی کند. اگر ۱۰ نمودار میله‌ای را پشت سر هم نشان دهیم، به سختی می‌توانید آن‌ها را از هم تفکیک کنید و سپس آنچه را دیدید به خاطر بسپارید. به عنوان مثال، اگر نگاهی گذرا به شکل ۲۹–۸ بیاندازید، متوجّه شباهت بصری آن به شکل ۲۹–۵ که قبلاً در این فصل دیدید، خواهید شد. با این حال، این دو شکل به جز اینکه نمودارهای میله‌ای هستند، هیچ وجه اشتراکی ندارند. شکل ۲۹–۵ تعداد پروازهای خروجی خطوط هوایی را از شهر نیویورک نشان می‌دهد، در حالی که شکل ۲۹–۸ محبوب‌ترین حیوانات خانگی را در ایالات متحده نشان می‌دهد. هر دو این شکل‌ها فاقد عنصری هستند که به شما کمک کند تا به طور شهودی درک کنید که شکل چه موضوعی را پوشش می‌دهد، و بنابراین هیچ کدام از آن‌ها به طور خاص به یاد ماندنی نیستند.

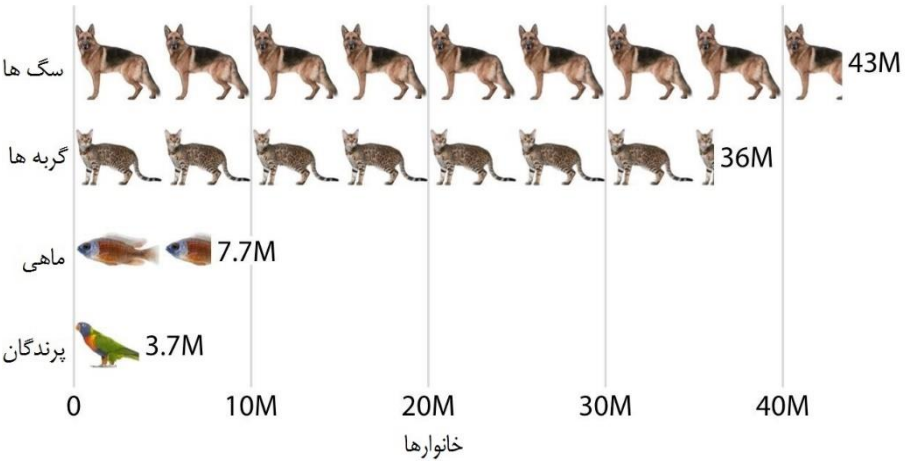


شکل ۲۹–۸. تعداد خانوارهایی که یک یا چند مورد از محبوب‌ترین حیوانات خانگی را دارند: سگ، گربه، ماهی یا پرنده. این نمودار میله‌ای کاملاً واضح است امّا لزوماً به یاد ماندنی نیست. ستون «گربه‌ها» صرفاً برای ایجاد شباهت بصری با شکل ۲۹–۵ رنگی شده است. منبع داده: کتاب منبع مالکیت و جمعیّت حیوانات خانگی ایالات متحده ۲۰۱۲ ، انجمن دامپزشکی آمریکا.

تحقیقات روی ادراک انسان نشان می‌دهد که هر چه نمودارهای بصری پیچیده‌تر و منحصر به فردتر باشند، به یاد ماندنی‌تر هستند [Bateman et al. 2010]؛ [Borgo et al. 2012]. با این حال، منحصربه‌فرد بودن و پیچیدگی بصری فقط بر به یاد ماندن تأثیر نمی‌گذارد، زیرا ممکن است مانع از ایجاد یک دید کلی از اطّلاعات شود یا تشخیص تفاوت‌های کوچک در مقادیر داده را دشوار کند. لذا از یک سو، یک شکل می‌تواند بسیار به یاد ماندنی امّا کاملا گیج‌کننده باشد. چنین شکلی ترسیم خوبی برای داده‌ها نخواهد بود، حتی اگر به عنوان یک اثر هنری خیره‌کننده به نظر برسد. از سوی دیگر، اشکال ممکن است بسیار واضح، امّا فراموش‌شدنی و خسته‌کننده باشند، و آن اشکال ممکن است تاثیری را که ما انتظارش را داریم نداشته باشند. به طور کلی، ما می‌خواهیم بین این دو حد تعادل ایجاد کنیم و نمودارهایی داشته باشیم که هم به یاد ماندنی و هم واضح باشند (مخاطبان مورد نظر نیز مهم هستند، اگر شکلی برای یک نشریهٔ علمی–تخصصی در نظر گرفته شده باشد، به طور کلی کمتر نگران به خاطره ماندن خواهیم بود تا زمانی که این شکل برای یک روزنامه یا وبلاگ پرخواننده در نظر گرفته شده باشد).

ما می‌توانیم با افزودن عناصر بصری که ویژگی‌های داده‌ها را منعکس می‌کنند، مانند نقاشی‌ها یا تصویرنگاشت‌های اجسام یا اشیایی که مجموعه داده دربارهٔ آن‌ها هستند، نمودار را به یاد

ماندنی‌تر کنیم. یکی از روش‌هایی که معمولاً به کار می‌رود نمایش مقادیر داده‌ها در قالب تصاویر تکراری می‌باشد، به طوری که هر کپی از یک تصویر با مقدار مشخصی از متغیّر نمایش داده شده مطابقت دارد. به عنوان مثال، می‌توانیم میله‌های شکل ۲۹–۸ را با تصاویر مکرر سگ، گربه، ماهی و پرنده جایگزین کنیم که در مقیاسی ترسیم شوند که هر حیوان کامل معادل ۵ میلیون خانوار باشد (شکل ۲۹–۹). بنابراین، از نظر بصری، شکل ۲۹–۹ همچنان به عنوان نمودار میله‌ای عمل می‌کند، امّا اکنون مقداری پیچیدگی بصری را اضافه کرده‌ایم که شکل را به یاد ماندنی‌تر می‌کند، و همچنین داده‌ها را با استفاده از تصاویر نشان داده‌ایم که مستقیماً معنای داده‌ها را منعکس می‌کند. پس از یک نگاه کوتاه به شکل، ممکن است بتوانید به یاد بیاورید که تعداد سگ‌ها و گربه‌ها بسیار بیشتر از ماهی یا پرندگان است. نکتهٔ مهم این است که در چنین ترسیم‌هایی، می‌خواهیم از تصاویر برای نمایش داده‌ها استفاده کنیم، نه اینکه از تصاویر صرفاً برای تزئین ترسیم یا حاشیه‌نویسی محورها استفاده کنیم. در آزمایش‌های روان‌شناسی، انتخاب‌های مذکور بیشتر باعث حواس‌پرتی می‌شوند تا اینکه مفید باشند [Haroz, Kosara, and Franconeri 2015].



شکل ۲۹–۹. تعداد خانوارهایی که یک یا چند مورد از محبوب‌ترین حیوانات خانگی دارند که به صورت نمودار ایزوتایپ نشان داده شده است. هر حیوان کامل نشان‌دهندهٔ ۵ میلیون خانوار است که چنین حیوان خانگی دارند. منبع داده: کتاب منبع مالکیت و جمعیّت حیوانات خانگی ایالات متحده ۲۰۱۲، انجمن دامپزشکی آمریکا.

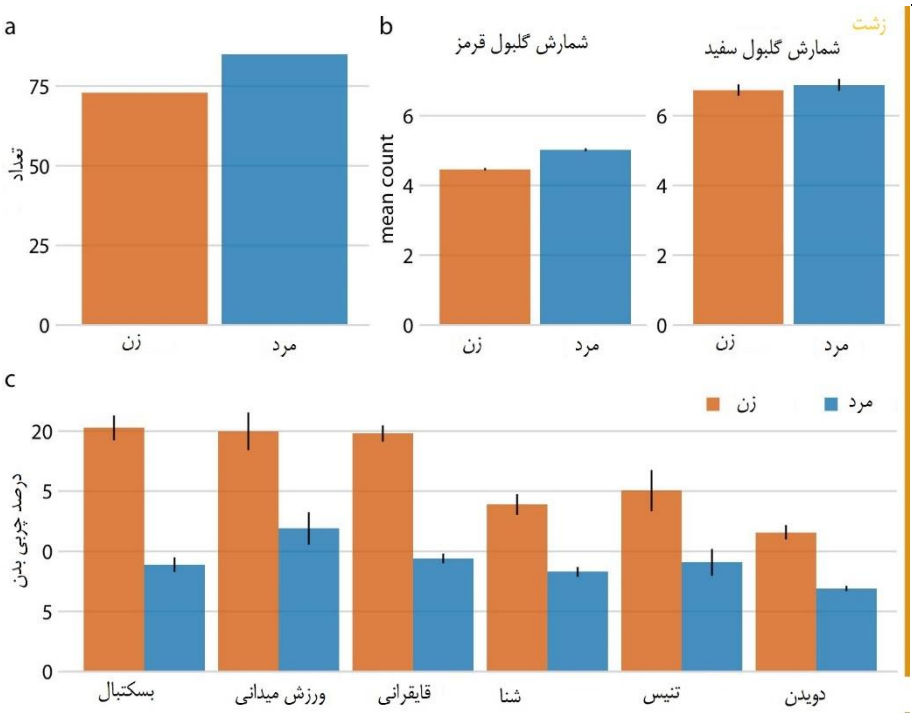
ترسیم‌هایی مانند شکل ۲۹-۹ اغلب نمودارهای ایزوتایپ[1] نامیده می‌شوند. کلمه «ایزوتایپ» به عنوان مخفف سیستم بین المللی آموزش تصویری تایپوگرافیک[2] معرفی شد و به تصویرنگاشت‌های ساده لوگو مانندی اشاره دارد که اشیا، حیوانات، گیاهان یا افراد را نشان می‌دهد [Haroz, Kosara, and Franconeri 2015]. با این حال، منطقی است که از اصطلاح نمودار ایزوتایپ به طور گسترده‌تر استفاده کنیم تا هر نوع ترسیمی که در آن کپی‌های مکرر از یک تصویر برای نشان دادن بزرگی یک مقدار استفاده می‌شود، را در برگیرد. در مجموع، پیشوند «iso» به معنای «همان» است و «type» می‌تواند به معنای یک نوع، طبقه یا گروه خاص باشد.

ثابت قدم باشید امّا تکراری نباشید

هنگام بحث در مورد شکل‌های ترکیبی در فصل ۲۱، اشاره شد که استفاده از زبان بصری ثابت برای بخش‌های مختلف یک شکل بزرگتر مهم است. در مورد خود اشکال نیز همین امر صادق است. اگر سه نمودار ترسیم کنیم که همگی بخشی از یک داستان بزرگتر باشند، باید آن‌ها را طوری طراحی کنیم که به هم مرتبط باشند. با این حال، استفاده از یک زبان بصری ثابت به این معنی نیست که همه چیز باید دقیقاً یکسان باشد. برعکس، مهم است که اشکالی که تحلیل‌های مختلف را توصیف می‌کنند، از نظر بصری متمایز به نظر برسند، به طوری که مخاطبان شما بتوانند به راحتی تشخیص دهند که یک تحلیل کجا ختم می‌شود و تحلیل دیگر از کجا شروع می‌شود. این مساله به بهترین وجه با استفاده از رویکردهای مصوّرسازی متفاوت برای بخش‌های مختلف داستان به دست می‌آید. اگر قبلاً از نمودار میله‌ای استفاده کرده‌اید، در مرحلهٔ بعدی از نمودار پراکنش، نمودار جعبه‌ای یا نمودار خطی استفاده کنید. در غیر این صورت، تحلیل‌های مختلف با هم در ذهن مخاطب شما محو می‌شود و مخاطب شما برای افتراق بخش‌های مختلف داستان از یکدیگر مشکل خواهد داشت. به عنوان مثال، اگر شکل ۲۱-۸ از «شکل‌های مرکب» را بازطراحی کنیم به طوری که فقط از نمودارهای میله‌ای استفاده شود، نتیجه حاصل واضحاً وضوح کمتر و گیج‌کنندگی بیشتری دارد (شکل ۲۹-۱۰).

هنگام تهیه یک ارائه یا گزارش، سعی کنید از انواع مختلف نمودار برای تحلیل‌های مجزا استفاده کنید.

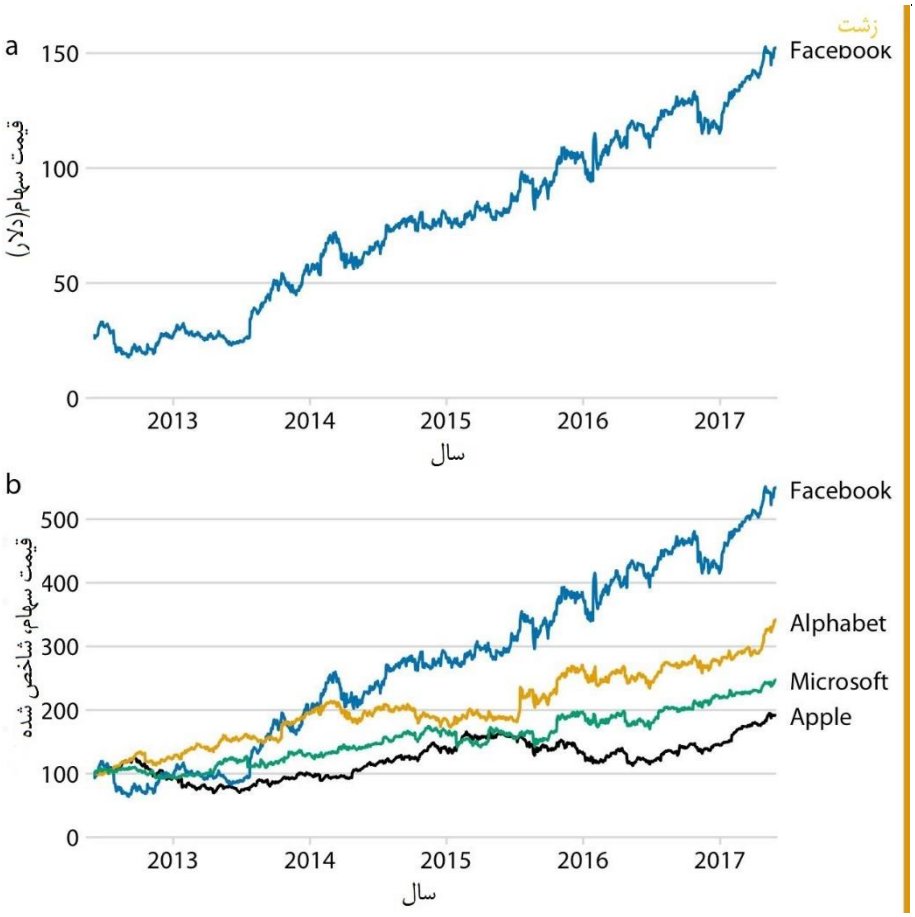
1. isotype
2. International System Of TYpographic Picture Education



شکل ۲۹–۱۰. فیزیولوژی و ترکیب بدنی ورزشکاران زن و مرد. میله‌های خطا نشان‌دهندهٔ خطای استاندارد میانگین است. این شکل بیش از حد تکراری است. این نمودار داده‌های مشابه شکل ۲۱–۸ را نشان می‌دهد و از یک زبان بصری ثابت استفاده می‌کند، امّا همه شکل‌ها از یک نوع مصوّرسازی (نمودار میله‌ای) استفاده می‌کنند. لذا برای خواننده دشوار خواهد بود که متوجّه شود بخش‌های (الف)، (ب) و (ج) نتایج کاملاً متفاوتی را نشان می‌دهند. منبع داده: Telford and Cunningham 1991 .

مجموعه‌ای از اشکال تکراری اغلب نتیجهٔ داستان‌های چند قسمتی است که در آن هر بخش مبتنی بر نوع مشابهی از داده‌های خام است. در این سناریوها، استفاده از نمودارهای مشابه برای هر قسمت می‌تواند وسوسه‌انگیز باشد. با این حال و در مجموع، این اشکال توجّه مخاطب را جلب نخواهد کرد. به عنوان مثال، اجازه دهید داستانی را در مورد قیمت سهام فیسبوک در دو بخش در نظر بگیریم: (۱) قیمت سهام فیسبوک از سال ۲۰۱۲ تا ۲۰۱۷ به سرعت افزایش یافته است، و (۲) افزایش قیمت از سایر شرکت‌های بزرگ فناوری پیشی گرفته است. ممکن است بخواهید این دو عبارت را با دو نمودار که قیمت سهام را در طول زمان نشان می‌دهد، ترسیم کنید، همانطور که در شکل ۲۹–۱۱ نشان داده شده است. با این حال، در حالی که شکل ۲۹–۱۱ الف یک هدف مشخّص را دنبال می‌کند و باید همانطور که هست باقی بماند، شکل ۲۹–۱۱ ب تکراری است و نکتهٔ اصلی را مبهم می‌کند. ما به تحوّلات

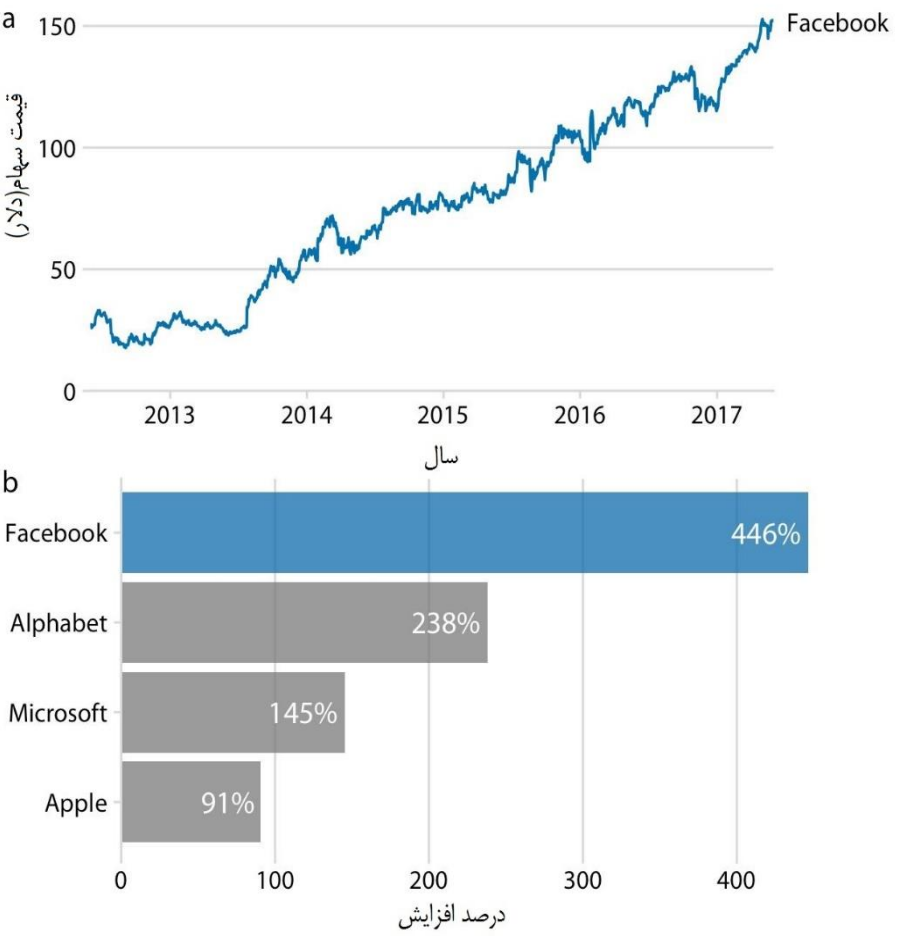
دقیـق قیمـت سـهام آلفابـت، اپـل یـا مایکروسـافت در طـول زمـان اهمیـت نمـی‌دهیم. مـا فقـط می‌خواهیم تأکید کنیم که رشد سهام این شرکت‌ها کمتر از رشد سهام فیسبوک بوده است.



شکل ۲۹–۱۱. رشد قیمت سهام فیسبوک در یک بازه زمانی پنج ساله و مقایسه با سایر شرکت‌های فناوری. (الف) قیمت سهام فیسبوک از حدود ۲۵ دلار برای هر سهم در اواسط سال ۲۰۱۲ به ۱۵۰ دلار برای هر سهم در اواسط سال ۲۰۱۷ افزایش یافت. (ب) قیمت سهام سایر شرکت‌های بزرگ فناوری در همان دورهٔ زمانی به طور قابل مقایسه‌ای افزایش نیافته است. قیمت‌ها در تاریخ ۱ ژوئن ۲۰۱۲ به عدد ۱۰۰ استاندارد شده‌اند تا امکان مقایسهٔ آسان را فراهم کنند. این شکل به عنوان «زشت» برچسب‌گذاری شده است زیرا قسمت‌های (الف) و (ب) تکراری هستند. منبع داده: امور مالی شرکت یاهو.

توصیه می‌شود قسمت (الـف) بـه همـان صـورتی کـه هسـت بـاقی بمانـد امّـا قسـمت (ب) بـا نمـودار میلـه‌ای کـه درصـد افـزایش را نشـان می‌دهـد جـایگزین شـود (شـکل ۲۹–۱۲). اکنـون دو شـکل

متمایز داریم که هر کدام یک نکتهٔ منحصر به فرد را نشان می‌دهند و در ترکیب با هم به خوبی کل داستان را روایت می‌کنند. بخش (الف) به خواننده اجازه می‌دهد با داده‌های خام و زیربنایی آشنا شود و بخش (ب) بزرگی اثر را در عین حذف هرگونه اطّلاعات حاشیه‌ای برجسته می‌کند.



شکل ۲۹-۱۲. رشد قیمت سهام فیسبوک در یک بازهٔ زمانی پنج ساله و مقایسه با سایر شرکت‌های فناوری. (الف) قیمت سهام فیسبوک از حدود ۲۵ دلار برای هر سهم در اواسط سال ۲۰۱۲ به ۱۵۰ دلار به ازای هر سهم در اواسط سال ۲۰۱۷ افزایش یافت که تقریباً ۴۵۰ درصد افزایش داشت. (ب) قیمت سهام سایر شرکت‌های بزرگ فناوری در همان دورهٔ زمانی به طور قابل مقایسه‌ای افزایش نیافته است. افزایش قیمت از حدود ۹۰ درصد تا تقریبا ۲۴۰ درصد متغیّر بود. منبع داده: امور مالی شرکت یاهو.

شکل ۲۹-۱۲ یک اصل کلی را نشان می‌دهد که هنگام تهیه مجموعه‌ای از شکل‌ها برای بیان یک داستان باید از آن پیروی کنیم: با شکلی شروع کنید که تا حد امکان به نمایش داده‌های خام نزدیک است و در شکل‌های بعدی تحلیل‌های جزیی‌تر را نشان دهید. تحلیل‌های جزیی‌تر (مانند درصد افزایش، میانگین‌ها، ضرایب مدل‌های برازش شده و غیره) برای خلاصه کردن روندهای کلیدی در مجموعهٔ داده‌های بزرگ و پیچیده مفید هستند. با این حال، از آنجایی که تحلیل اختصاصی‌تر هستند، کمتر شهودی بوده و اگر قبل از نمایش داده‌های خام، تحلیل اختصاصی را نشان دهیم، درک آن برای مخاطبان دشوار خواهد بود. از طرف دیگر، اگر بخواهیم همهٔ روندها را با نمایش داده‌های خام نشان دهیم، در نهایت به نمودارهای متعددی نیاز داریم و درگیر نمودار تکراری خواهیم شد.

برای بیان داستان خود باید از چند شکل استفاده کنید؟ پاسخ این سوال به محل انتشار بستگی دارد. برای مطلبی کوتاه در وبلاگ یا یک توییت، از یک نمودار استفاده کنید. برای مقالات علمی، استفاده از سه تا شش نمودار پیشنهاد می‌شود. اگر بیش از شش نمودار برای یک مقالهٔ علمی وجود داشته باشد، ممکن است لازم باشد برخی از آن‌ها به بخش پیوست یا بخش ضمائم منتقل شوند. خوب است که تمام شواهدی را که جمع آوری کرده‌ایم مستند کنیم، امّا نباید با ارائه بیش از حد شکل‌هایی که عمدتاً شبیه به هم هستند، مخاطبان خود را خسته کنیم. در زمینه‌های دیگر، تعداد بیشتری از نمودارها ممکن است مناسب باشد. با این حال، در آن زمینه‌ها، ما معمولاً داستان‌های متعدد یا داستانی فراگیر با زیرشاخه‌های فرعی را تعریف می‌کنیم. به عنوان مثال، اگر از من خواسته شود که یک سخنرانی علمی یک ساعته ارائه دهم، معمولاً هدفم این است که سه داستان متمایز را بگویم. به همین ترتیب، یک کتاب یا پایان‌نامه حاوی بیش از یک داستان است و در واقع ممکن است هر فصل یا بخش، یک داستان داشته باشد. در آن سناریوها، هر خط داستانی متمایز نباید بیش از سه تا شش نمودار داشته باشد. در این کتاب ما این اصل را در سطح بخش‌های هر فصل رعایت نمودیم. هر بخش تقریباً مستقل است و معمولاً حاوی بیش از شش شکل نیست.

مشروح کتابشناسی

هیچ کتابی نمی‌تواند به تنهایی همهٔ مطالبی که باید دربارهٔ یک موضوع بدانیم را پوشش دهد. ما شما را تشویق می‌کنیم که متون دیگر را نیز در مورد ترسیم داده‌ها بخوانید تا درک خود را عمیق‌تر کرده و مهارت‌های فنّی خود را در ایجاد شکل‌ها توسعه دهید. در اینجا، مجموعهٔ محدودی از کتاب‌هایی را ارائه می‌دهیم که به نظرمان جالب، قابل تأمل یا مفید می‌باشند. کتاب‌هایی که در بخش اوّل فهرست شده‌اند، به کتاب حاضر بسیار شبیه هستند و ممکن است دیدگاه‌های مکمّل یا جایگزینی در مورد موضوعاتی که بحث شد را ارائه دهند. کتاب‌های فهرست‌شده در بخش «کتاب‌های برنامه‌نویسی» به موضوع مهّم چگونگی ایجاد نمودار با استفاده از رویکردهای برنامه‌نویسی و کتابخانه‌های نرم‌افزاری موجود می‌پردازند. سایر بخش‌ها، کتاب‌های دیگری را فهرست می‌کنند که دانش شما را در مورد مصوّرسازی داده‌ها گسترش داده و به شما در برقراری ارتباط با تصاویر و داده‌ها کمک می‌کنند.

تفکّر در مورد داده‌ها و مصوّرسازی

کتاب‌های زیر فرآیندهای فکری و تصمیم‌گیری مورد نیاز برای تبدیل داده‌ها به نمودار را مورد بحث قرار می‌دهند. آن‌ها به عنوان متون مقدماتی در مورد چگونگی انتخاب نمودار مناسب و مشکلاتی که باید مدنظر قرار گیرند، کمک‌کننده هستند:

Alberto Cairo. The Truthful Art. New Riders, 2016.

مقدمه‌ای عالی برای ترسیم داده‌ها، به ویژه برای روزنامه‌نگاران. این کتاب بسیاری از مفاهیم مهم ترسیم داده‌ها، مانند نحوه ترسیم توزیع‌ها، روندها، عدم قطعیّت و نقشه‌ها را پوشش

می‌دهد. همچنین در بسیاری از فصول، به عنوان مقدمه‌ای بر اصول اولیّه آماری، توضیح مفاهیمی مانند جمعیّت، نمونه و سطوح اطمینان عمل می‌کند.

Stephen Few. Show Me the Numbers. Analytics Press, 2012.

کتابی در مورد ترسیم داده‌ها برای افراد حرفه‌ای در حوزهٔ تجارت. از نظر دامنه و مخاطب هدف مشابه مرجع زیر است، اما حاوی مطالب بیشتری است و بسیاری از موضوعات را با عمق بیشتری پوشش می‌دهد. با این حال، به اندازهٔ کتاب زیر خوب نوشته و منتشر نشده است.

Cole Nussbaumer Knaflic. Storytelling with Data. John Wiley & Sons, 2015

کتابی که در مورد چگونگی تبدیل داده‌ها به تصاویر بصری که به خوبی نوشته شده و به خوبی هم منتشر شده است. مخاطبان اصلی کتاب افرادی هستند که گرافیک تجاری می‌سازند و مرجع عالی برای موضوعاتی است که پوشش می‌دهد. با این حال، بسیاری از موضوعات مهم برای دانشمندان مانند مصوّرسازی توزیع‌ها، روندها یا عدم قطعیّت را پوشش نمی‌دهد.

کتاب‌های برنامه‌نویسی

منابع زیر همگی کتاب‌هایی هستند که رویکردهای برنامه‌نویسی برای مصوّرسازی داده‌ها را آموزش می‌دهند:

Kieran Healy. Data Visualization: A Practical Introduction. Princeton University Press, 2018.

مقدمه‌ای بر استفاده از ggplot2 برای مصوّرسازی داده‌ها. به عنوان متن مکمّل پس از مطالعه کتاب Wickham و Grolemund (که در ادامه در این لیست ذکر شده است) توصیه می‌شود.

Scott Murray. Interactive Data Visualization for the Web: An Introduction to Design- ing with D3. 2nd ed. O'Reilly Media, 2017.

مقدمه‌ای بر ایجاد مصوّرسازی‌های برخط تعاملی با D3، با استفاده از HTML، CSS، جاوا اسکریپت و SVG.

Jake VanderPlas. Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media, 2016.

مقدمه‌ای بر استفاده از زبان برنامه‌نویسی پایتون برای علم داده. دارای مطالب گسترده‌ای در مورد مصوّرسازی داده‌ها با استفاده از Matplotlib و Seaborn پایتون است.

Hadley Wickham, Garrett Grolemund. R for Data Science. O'Reilly Media, 2017

مقدمه‌ای همه جانبه برای استفاده از زبان برنامه نویسی R برای علم داده. شامل چندین فصل در مورد استفاده از ggplot2 برای مصوّرسازی داده‌ها است.

متون آماری

متون مقدماتی در آمار به طور کلی حاوی مطالبی در مورد مصوّرسازی داده‌ها هستند که موضوعاتی مانند نمودارهای پراکندگی، هیستوگرام‌ها، نمودارهای جعبه‌ای و نمودارهای خطی را پوشش می دهند. متون مختلفی در این حوزه وجود دارد و می‌توان آنها را فهرست کرد. در اینجا، فقط به چند مورد که اخیراً منتشر شده‌اند و ارزش مطالعه دارند، اشاره می‌شود:

David M. Diez, Christopher D. Barr, Mine Çetinkaya-Rundel. OpenIntro Statistics. 3rd ed. OpenIntro, Inc., 2015.

کتاب درسی آمار مقدماتی به صورت منبع باز. کل متن کتاب و نیز فایل‌های LaTeX و کد R که برای جمع‌آوری کتاب و ایجاد شکل‌ها استفاده شده است، به‌طور رایگان در دسترس است.

Susan Holmes, Wolfgang Huber. Modern Statistics for Modern Biology. Cambridge University Press, 2018.

یک متن آماری که بر ابزارهای محاسباتی مورد نیاز برای زیست‌شناسی مدرن تاکید دارد. کل کتاب به صورت رایگان در دسترس است و کد R برای همه مثال‌ها ارائه شده است.

Chester Ismay, Albert Y. Kim. Modern Dive—An Introduction to Statistical and DataSciences via R. https://moderndive.com

یک کتاب درسی مقدماتی که فقط به صورت برخط در دسترس است و آمار پایه و علم داده را آموزش می‌دهد. این کتاب هم مفاهیم نظری و هم رویکردهای عملی را با استفاده از R را پوشش می‌دهد.

متون تاریخی

کتاب‌های این بخش عمدتاً به دلایل تاریخی مورد توجّه هستند. آن‌ها در زمان انتشارشان تأثیرگذار بودند، اما اکنون می‌توان مطالب مشابهی را در جاهای دیگر یا به شکل مدرن‌تر یافت:

William S. Cleveland. The Elements of Graphing Data. 2nd ed. Hobart Press, 1994.

یکی از اولین کتاب‌هایی که در مورد کار با داده‌ها برای متخصصین آمار نوشته شده است. این کتاب شامل نمونه‌های فراوانی از نمودارهای پراکندگی، نمودارهای خطی، هیستوگرام‌ها و نمودارهای جعبه‌ای است و آن‌ها را در چارچوب تحلیل داده‌ها و مدل‌سازی آماری مورد بحث قرار می‌دهد. همچنین این کتاب طرح نقطه‌ای کلیولند را رایج کرد.

William S. Cleveland. Visualizing Data. Hobart Press, 1993.

کتاب همراه برای کتاب ذکر شده قبلی توسط همان نویسنده. این یکی بیشتر ریاضی است و در مورد درک انسان صحبت نمی‌کند.

Edward R. Tufte. Envisioning Information. Graphics Press, 1990.

این کتاب مفهوم چندگانه‌های کوچک را رایج کرد.

Edward R. Tufte. The Visual Display of Quantitative Information. 2nd ed. Graphics Press, 2001.

این کتاب اولین بار در سال ۱۹۸۳ منتشر شد و در زمینه مصوّرسازی داده‌ها بسیار تأثیرگذار بوده است. مفاهیمی مانند ضایعات نمودار، نسبت داده به جوهر و خطوط جرقه را معرفی کرد. این کتاب همچنین اولین شیب‌نگار را نشان داد (اما نامی بر آن نگذاشت). با این حال، حاوی چندین توصیه است که در آزمون زمان تاب نیاوردند. به طور خاص، این کتاب طراحی بیش از حد کمینه‌ای نمودار را توصیه می‌کند.

کتاب‌هایی در مورد موضوعات مرتبط

کتاب‌های زیر به طور کلی در مورد موضوع مصوّرسازی داده‌ها و نحوه برقراری ارتباط مؤثر هستند:

Joshua Schimel. Writing Science. Oxford University Press, 2011.

نحوه نوشتن در مورد موضوعات علمی و فنّی دیگر را به شیوه‌ای جذاب و با گفتن داستان آموزش می‌دهد. اگرچه در اصل کتابی در مورد مصوّرسازی داده‌ها نیست، اما برای هر کسی که نیاز به نوشتن مقاله و/یا طرح پژوهشی دارد، متنی ضروری است.

Jonathan Schwabish. Better Presentations. Columbia University Press, 2016

راهنمای کوتاه و آموزنده برای تنظیم یک ارائه. خواندن آن برای هر کسی که به طور معمول از اسلایدها برای سخنرانی یا ارائه استفاده می‌کند، ضروری است.

Maureen C. Stone. A Field Guide to Digital Color. A K Peters, 2003.

راهنمای جامعی برای آشنایی با چگونگی ضبط، پردازش و تکثیر رنگ‌ها توسط رایانه.

Colin Ware. Information Visualization. 3rd ed. Morgan Kaufmann, 2012.

کتابی در مورد اصول مصوّرسازی، به طور خاص به موضوعاتی مانند نحوهٔ عملکرد سیستم بینایی انسان و نحوهٔ درک الگوهای گرافیکی مختلف می‌پردازد. این کتاب سناریوهای مختلف مصوّرسازی، از جمله رابطهای کاربری و دنیای مجازی را پوشش می‌دهد، اما تاکید نسبتاً کمتری بر مصوّرسازی داده‌ها در قالب شکل‌های دوبعدی دارد.

نکات فنی

کل کتاب در R Markdown با استفاده از بسته‌های bookdown، rmarkdown و knitr نوشته شده است. همه شکل‌ها با ggplot2 و با کمک چندین بستهٔ الحاقی از جمله cowplot، geofacet، ggforce، ggmap، ggrepel، ggridges، hexbin، patchwork، sf، statebins، tidybayes و treemapify ساخته شده‌اند. دستکاری رنگ با بسته‌های colorspace و colorblindr انجام شده است. برای بسیاری از این بسته‌ها، نسخهٔ توسعه‌دهنده فعلی برای جمع‌آوری تمام قسمت‌های کتاب مورد نیاز است.

کد منبع کتاب در https://github.com/clauswilke/dataviz موجود است. همچنین این کتاب به یک بستهٔ پشتیبانی R، dviz.supp نیاز دارد که کد آن در https://github.com/clauswilke/dviz.supp موجود است.

آخرین محیط مورد استفاده برای این کتاب عبارت است از:

```
## R version 3.5.0 (2018-04-23)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Sierra 10.12.6
##
## Matrix products: default
## BLAS: /Library/Frameworks/ ... /libRblas.0.dylib
## LAPACK: /Library/Frameworks/ ... /libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/ ... /C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats graphics grDevices utils datasets methods base
##
## other attached packages:
## [1] nycflights13_1.0.0 gapminder_0.3.0 RColorBrewer_1.1-2
## [4] gganimate_1.0.0.9000 ungeviz_0.1.0 emmeans_1.3.1
```

```
## [7] mgcv_1.8-24 nlme_3.1-137 broom_0.5.1
## [10] tidybayes_1.0.3 maps_3.3.0 statebins_2.0.0
## [13] sf_0.7-1 maptools_0.9-4 sp_1.3-1
## [16] rgeos_0.3-28 ggspatial_1.0.3 geofacet_0.1.9
## [19] plot3D_1.1.1 magick_1.9 hexbin_1.27.2
## [22] treemapify_2.5.0 gridExtra_2.3 ggmap_2.7.904
## [25] ggthemes_4.0.1 ggridges_0.5.1 ggrepel_0.8.0
## [28] ggforce_0.1.1 patchwork_0.0.1 lubridate_1.7.4
## [31] forcats_0.3.0 stringr_1.3.1 purrr_0.2.5
## [34] readr_1.1.1 tidyr_0.8.2 tibble_1.4.2
## [37] tidyverse_1.2.1 dviz.supp_0.1.0 dplyr_0.8.0.9000
## [40] colorblindr_0.1.0 ggplot2_3.1.0 colorspace_1.4-0
## [43] cowplot_0.9.99
##
## loaded via a namespace (and not attached):
## [1] rjson_0.2.20 deldir_0.1-15
## [3] class_7.3-14 rprojroot_1.3-2
## [5] estimability_1.3 ggstance_0.3.1
## [7] rstudioapi_0.7 farver_1.0.0.9999
## [9] ggfittext_0.6.0 svUnit_0.7-12
## [11] mvtnorm_1.0-8 xml2_1.2.0
## [13] knitr_1.20 polyclip_1.9-1
## [15] jsonlite_1.5 png_0.1-7
## [17] compiler_3.5.0 httr_1.3.1
## [19] backports_1.1.2 assertthat_0.2.0
## [21] Matrix_1.2-14 lazyeval_0.2.1
## [23] cli_1.0.1.9000 tweenr_1.0.1
## [25] prettyunits_1.0.2 htmltools_0.3.6
## [27] tools_3.5.0 misc3d_0.8-4
## [29] coda_0.19-2 gtable_0.2.0
## [31] glue_1.3.0 Rcpp_1.0.0
## [33] cellranger_1.1.0 imguR_1.0.3
## [35] xfun_0.3 strapgod_0.0.0.9000
## [37] rvest_0.3.2 MASS_7.3-50
## [39] scales_1.0.0 hms_0.4.2
## [41] yaml_2.2.0 stringi_1.2.4
## [43] e1071_1.7-0 spData_0.2.9.4
## [45] RgoogleMaps_1.4.3 rlang_0.3.0.1
## [47] pkgconfig_2.0.2 bitops_1.0-6
## [49] geogrid_0.1.1 evaluate_0.11
## [51] lattice_0.20-35 tidyselect_0.2.5
## [53] plyr_1.8.4 magrittr_1.5
## [55] bookdown_0.7 R6_2.3.0
## [57] generics_0.0.2 DBI_1.0.0
## [59] pillar_1.3.0 haven_1.1.2
## [61] foreign_0.8-71 withr_2.1.2.9000
## [63] units_0.6-1 modelr_0.1.2
## [65] crayon_1.3.4 arrayhelpers_1.0-20160527
## [67] rmarkdown_1.10 progress_1.2.0.9000
## [69] jpeg_0.1-8 rnaturalearth_0.1.0
## [71] grid_3.5.0 readxl_1.1.0
## [73] digest_0.6.18 classInt_0.2-3
## [75] xtable_1.8-3 munsell_0.5.0
## [77] concaveman_1.0.0
```

منابع

Bateman, S., R. Mandryk, C. Gutwin, A. Genest, D. McDine, and C. Brooks. 2010.″ Useful Junk? The Effects of Visual Embellishment on Comprehension and Memorability of Charts″ *ACM Conference on Human Factors in Computing Systems*, 2573–82. doi:10.1145/1753326.1753716.

Becker, R. A., W. S. Cleveland, and M.-J. Shyu. 1996. ″The Visual Design and Control of Trellis Display.″ *Journal of Computational and Graphical Statistics* 5: 123–55.

Bergstrom, C. T., and J. West. 2016. ″The Principle of Proportional Ink.″ *http://callingbullshit.org/tools/tools_proportional_ink.html*.

Borgo, R., A. Abdul-Rahman, F. Mohamed, P. W. Grant, I. Reppa, and L. Floridi. 2012. ″An Empirical Study on Using Visual Embellishments in Visualization.″ *IEEE Transactions on Visualization and Computer Graphics* 18: 2759–68. doi:10.1109/TVCG.2012.197.

Brewer, Cynthia A. 2017. ″ColorBrewer 2.0. Color Advice for Cartography.″ *http://www.ColorBrewer.org*.

Carr, D. B., R. J. Littlefield, W. L. Nicholson, and J. S. Littlefield. 1987. ″Scatterplot Matrix Techniques for Large N.″ *Journal of the American Statistical Association* 82:424–36.

Clauset, A., C. R. Shalizi, and M. E. J. Newman. 2009. ″Power-Law Distributions in Empirical Data″ *SIAM Review* 51: 661–703.

Cleveland, R. B., W. S. Cleveland, J. E. McRae, and I. Terpenning. 1990. ″STL: A Seasonal-Trend Decomposition Procedure Based on Loess.″ *Journal of Official Statistics* 6: 3–73.

Cleveland, W. S. 1979. "Robust Locally Weighted Regression and Smoothing Scatterplots." *Journal of the American Statistical Association* 74: 829–36.

———. 1993. *Visualizing Data*. Summit, New Jersey: Hobart Press.

Dua, D., and E. Karra Taniskidou. 2017. "UCI Machine Learning Repository." University of California, Irvine, School of Information; Computer Sciences. *https://archive.ics.uci.edu/ml*

Fisher, R. A. 1936. "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics* 7: 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x.

Haroz, S., R. Kosara, and S. L. Franconeri. 2015. "ISOTYPE Visualization: Working Memory, Performance, and Engagement with Pictographs." *ACM Conference on Human Factors in Computing Systems*, 1191–1200. doi:10.1145/2702123.2702275.

———. 2016. "The Connected Scatterplot for Presenting Paired Time Series." *IEEE Transactions on Visualization and Computer Graphics* 22: 2174–86. doi:10.1109/TVCG.2015.2502587.

Hullman, J., P. Resnick, and E. Adar. 2015. "Hypothetical Outcome Plots Outperform Error Bars and Violin Plots for Inferences About Reliability of Variable Ordering." *PLOS ONE* 10: e0142444. doi:10.1371/journal.pone.0142444.

Kale, A., F. Nguyen, M. Kay, and J. Hullman. 2018. "Hypothetical Outcome Plots Help Untrained Observers Judge Trends in Ambiguous Data." *IEEE Transactions on Visualization and Computer Graphics* 25: 892–905. doi:10.1109/TVCG.2018.2864909.

Kay, M., T. Kola, J. Hullman, and S. Munson. 2016. "When (Ish) Is My Bus? User-Centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems." CHI Conference on Human Factors in Computing Systems, 5092–5103. doi: 10.1145/2858036.2858558.

Marcos, M. L., and J. Echave. 2015. "Too Packed to Change: Side-Chain Packing and Site-Specific Substitution Rates in Protein Evolution." *PeerJ* 3: e911.

McDonald, Ian. 2017. "DW-NOMINATE Using ggjoy." *http://rpubs.com/ianrmcdonald/293304*.

Molyneaux, L., S. K. Gilliam, and L. C. Florant. 1947. "Differences in Virginia Death Rates by Color, Sex, Age, and Rural or Urban Residence." *American Sociological Review* 12: 525–35.

Okabe, M., and K. Ito. 2008. "Color Universal Design (CUD): How to Make Figures and Presentations That Are Friendly to Colorblind People." *http://jfly.iam.utokyo.ac.jp/color/*.

Paff, M. L., B. R. Jack, B. L. Smith, J. J. Bull, and C. O. Wilke. 2018. ″Combinatorial Approaches to Viral Attenuation.″ bioRxiv, 29918. doi:10.1101/299180.

Schimel, J. 2011. *Writing Science: How to Write Papers That Get Cited and Proposals That Get Funded*. Oxford: Oxford University Press.

Sidiropoulos, N., S. H. Sohi, T. L. Pedersen, B. T. Porse, O. Winther, N. Rapin, and F. O. Bagger. 2018. ″SinaPlot: An Enhanced Chart for Simple and Truthful Representation of Single Observations over Multiple Classes.″ *Journal of Computational and Graphical Statistics* 27: 673–76. doi:10.1080/10618600.2017.1366914.

Stone, M., D. Albers Szafir, and V. Setlur. 2014. ″An Engineering Model for Color Difference as a Function of Size.″ 22nd Color and Imaging Conference, 253–258.

Telford, R. D., and R. B. Cunningham. 1991. ″Sex, Sport, and Body-Size Dependency of Hematology in Highly Trained Athletes.″ *Medicine and Science in Sports and Exercise* 23: 788–94.

The Economist online. 2011. ″Corrosive Corruption.″ *https://www.economist.com/graphic-detail/2011/12/02/corrosive-corruption*.

Tufte, E. R. 1990. *Envisioning Information*. Cheshire, Connecticut: Graphics Press.

———. 2001. *The Visual Display of Quantitative Information*. 2nd ed. Cheshire, Connecticut: Graphics Press.

Wehrwein, A. 2017. ″It Brings Me ggjoy.″ *http://austinwehrwein.com/datavisualization/it-brings-me-ggjoy/*.

Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. 2nd ed. New York: Springer.

Wikipedia, User:Schutz. 2007. ″File:Piecharts.svg.″ *https://en.wikipedia.org/wiki/File:Piecharts.svg*.

Yates, F. 1935. ″Complex Experiments.″ *Supplement to the Journal of the Royal Statistical Society* 2: 181–247. doi:10.2307/2983638.

Mashhad University of
Medical Sciences
Vice Chancellor for
Research and Technology

Fundamentals of Data Visualization

A Primer on Making Informative and Compelling Figures

Claus O. Wilke

Translated by:

Majid Khadem Rezaiyan

Mohammad MasoumVand