Chapter 6

Visualizing Data

> At their best, graphics are instruments for reasoning.
>
> – Edward Tufte

Effective data visualization is an important aspect of data science, for at least three distinct reasons:

- *Exploratory data analysis*: What does your data really look like? Getting a handle on what you are dealing with is the first step of any serious analysis. Plots and visualizations are the best way I know of to do this.
- *Error detection*: Did you do something stupid in your analysis? Feeding unvisualized data to any machine learning algorithm is asking for trouble. Problems with outlier points, insufficient cleaning, and erroneous assumptions reveal themselves immediately when properly visualizing your data. Too often a summary statistic (77.8% accurate!) hides what your model is really doing. Taking a good hard look what you are getting right vs. wrong is the first step to performing better.
- *Communication*: Can you present what you have learned effectively to others? Meaningful results become actionable only after they are shared. Your success as a data scientist rests on convincing other people that you know what you are talking about. A picture is worth 1,000 words, especially when you are giving a presentation to a skeptical audience.

You have probably been making graphs and charts since grade school. Ubiquitous software makes it easy to create professional-looking images. So what is so hard about data visualization?

To answer, I offer a parable. A terrible incident during my youth concerned an attack on an ice skating champion. A thug hit her on the knee with a stick, hoping to knock her out of the upcoming Olympic games. Fortunately, he missed the knee, and the skater went on to win the silver medal.

S.S. Skiena, *The Data Science Design Manual,*
Texts in Computer Science, https://doi.org/10.1007/978-3-319-55444-0_6

But upon getting to know his client, the thug's lawyer came up with an interesting defense. This crime, he said, was clearly too complex for his client to have conceived on his own. This left an impression on me, for it meant that I had underestimated the cognitive capacity necessary to rap someone on the leg with a stick.

My intended moral here is that many things are more complicated than they look. In particular, I speak of the problem of plotting data on a graph to capture what it is saying. An amazingly high fraction of the charts I have seen in presentations are *terrible*, either conveying no message or misconveying what the data actually showed. Bad charts can have negative value, by leading you in the wrong direction.

In this section, we will come to understand the principles that make standard plot designs work, and show how they can be misleading if not properly used. From this experience, we will try to develop your sense of when graphs might be lying, and how you can construct better ones.

6.1 Exploratory Data Analysis

The advent of massive data sets is changing in the way science is done. The traditional scientific method is *hypothesis driven*. The researcher formulates a theory of how the world works, and then seeks to support or reject this hypothesis based on data. By contrast, *data-driven* science starts by assembling a substantial data set, and then hunts for patterns that ideally will play the role of hypotheses for future analysis.

Exploratory data analysis is the search for patterns and trends in a given data set. Visualization techniques play an important part in this quest. Looking carefully at your data is important for several reasons, including identifying mistakes in collection/processing, finding violations of statistical assumptions, and suggesting interesting hypotheses.

In this section, we will discuss how to go about doing exploratory data analysis, and what visualization brings to the table as part of the process.

6.1.1 Confronting a New Data Set

What should you do when encountering a new data set? This depends somewhat upon why you are interested in it in the first place, but the initial steps of exploration prove almost application-independent.

Here are some basic steps that I encourage doing to get acquainted with any new data set, which I illustrate in exploring the body measurement data set NHANES, available at `https://www.statcrunch.com/app/index.php?dataid=1406047`. This is tabular data, but the general principles here are applicable to a broader class of resources:

- *Answer the basic questions*: There are several things you should know about your data set before you even open the file. Ask questions like:

  - *Who constructed this data set, when, and why?* Understanding how your data was obtained provides clues as to how relevant it is likely to be, and whether we should trust it. It also points us to the right people if we need to know more about the data's origin or provenance. With a little digging, I discovered that it came from the National Health and Nutrition Examination Survey 2009–2010, and who was responsible for posting it.

  - *How big is it?* How rich is the data set in terms of the number of fields or columns? How large is it as measured by the number of records or rows? If it is too big to explore easily with interactive tools, extract a small sample and do your initial explorations on that. This data set has 4978 records (2452 men and 2526 women), each with seven data fields plus gender.

  - *What do the fields mean?* Walk through each of the columns in your data set, and be sure you understand what they are. Which fields are numerical or categorical? What units were the quantities measured in? Which fields are IDs or descriptions, instead of data to compute with? A quick review shows that the lengths and weights here were measured using the metric system, in centimeters and kilograms respectively.

- *Look for familiar or interpretable records*: I find it extremely valuable to get familiar with a few records, to the point where I know their names. Records are generally associated with a person, place, or thing that you already have some knowledge about, so you can put it in context and evaluate the soundness of the data you have on it. But if not, find a few records of special interest to get to know, perhaps the ones with the maximum or minimum values of the most important field.

  If familiar records do not exist, it sometimes pays to create them. A clever developer of a medical records database told me that he had used the top 5000 historical names from *Who's Bigger* to serve as patient names while developing the product. This was a much more inspired idea than making up artificial names like "Patient F1253." They were fun enough to encourage playing with the system, and memorable enough that outlier cases could be flagged and reported: e.g. "There is something seriously wrong with Franz Kafka."

- *Summary statistics*: Look at the basic statistics of each column. Tukey's *five number summary* is a great start for numerical values, consisting of the extreme values (max and min), plus the median and quartile elements.

  Applied to the components of our height/weight data set, we get:

	Min	25%	Median	75%	Max
Age	241	418	584	748	959
Weight	32.4	67.2	78.8	92.6	218.2
Height	140	160	167	175	204
Leg Length	23.7	35.7	38.4	41	55.5
Arm Length	29.5	35.5	37.4	39.4	47.7
Arm Circumference	19.5	29.7	32.8	36.1	141.1
Waist	59.1	87.5	97.95	108.3	172

This is very informative. First, what's the deal that the median age is 584? Going back to the data, we learn that age is measured in months, meaning the median is 48.67 years. Arm and leg length seem to have about the same median, but leg length has much greater variability. *I never knew that.* But suddenly I realize that people are more often described as long/short legged than long/short armed, so maybe this is why.

For categorical fields, like occupation, the analogous summary would be a report on how many different label types appear in the column, and what the three most popular categories are, with associated frequencies.

- *Pairwise correlations*: A matrix of correlation coefficients between all pairs of columns (or at least the columns against the dependent variables of interest) gives you an inkling of how easy it will be to build a successful model. Ideally, we will have several features which strongly correlate with the outcome, while not strongly correlating with each other. Only one column from a set of perfectly correlated features has any value, because all the other features are completely defined from any single column.

	Age	Weight	Height	Leg Length	Arm Length	Arm Circum	Waist
Age	1.000						
Weight	0.017	1.000					
Height	−0.105	0.443	1.000				
Leg_Len	−0.268	0.238	0.745	1.000			
Arm_Len	0.053	0.583	0.801	0.614	1.000		
Arm_Circ	0.007	0.890	0.226	0.088	0.444	1.000	
Waist	0.227	0.892	0.181	−0.029	0.402	0.820	1.000

These pairwise correlations are quite interesting. Why is height *negatively* correlated with age? The people here are all adults (241 months = 20.1 years), so they are all fully grown. But the previous generation was shorter than the people of today. Further, people shrink when they get older, so together this probably explains it. The strong correlation between weight and waist size (0.89) reflects an unfortunate truth about nature.

- *Class breakdowns*: Are there interesting ways to break things down by major categorical variables, like gender or location? Through summary

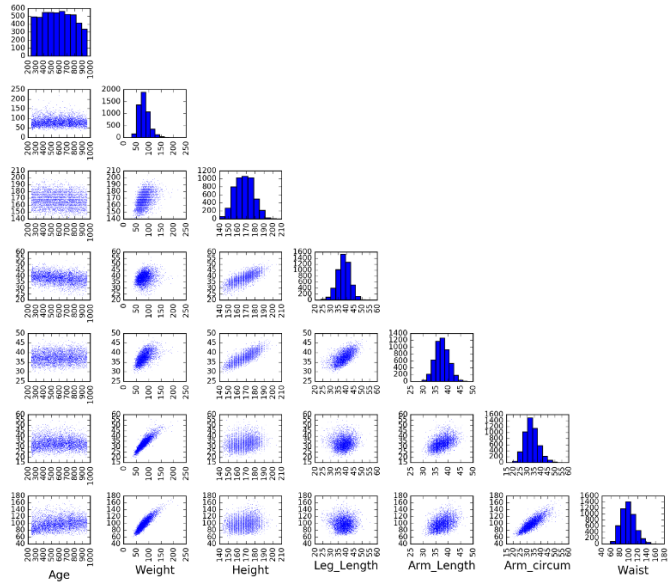

Figure 6.1: The array of dot plots of variable pairs provides quick insight into the distributions of data values and their correlations.

statistics, you can gauge whether there is a difference among the distributions when conditioned on the category. Look especially where you think there *should* be differences, based on your understanding of the data and application.

The correlations were generally similar by gender, but there were some interesting differences. For example, the correlation between height and weight is stronger for men (0.443) than women (0.297).

- *Plots of distributions*: This chapter focuses on visualization techniques for data. Use the chart types we will discuss in Section 6.3 to eyeball the distributions, looking for patterns and outliers. What is the general shape of each distribution? Should the data be cleaned or transformed to make it more bell-shaped?

  Figure 6.1 shows the power of a grid of dot plots of different variables. At a glance we see that there are no wild outliers, which pairs are correlated, and the nature of any trend lines. Armed with this single graphic, we are now ready to apply this data set to whatever is the challenge at hand.

6.1.2 Summary Statistics and Anscombe's Quartet

There are profound limits to how well you can understand data without visualization techniques. This is best depicted by Anscombe's quartet: four two-dimensional data sets, each with eleven points and shown in Figure 6.2. All four

	I		II		III		IV	
	x	**y**	**x**	**y**	**x**	**y**	**x**	**y**
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.31	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Var.	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
Corr.	0.816		0.816		0.816		0.816	

Figure 6.2: Four data sets with identical statistical properties. What do they look like?

data sets have identical means for the x and y values, identical variances for the x and y values, and the exact same correlation between the x and y values.

These data sets must all be pretty similar, right? Study the numbers for a little while so you can get an idea of what they look like.

Got it? Now peak at the dot plots of these data sets in Figure 6.3. They all look different, and tell substantially different stories. One trends linear, while a second looks almost parabolic. Two others are almost perfectly linear modulo outliers, but with wildly different slopes.

The point here is that you can instantly appreciate these differences with a glance at the scatter plot. Even simple visualizations are powerful tools for understanding what is going on in a data set. Any sensible data scientist strives to take full advantage of visualization techniques.

6.1.3 Visualization Tools

An extensive collection of software tools are available to support visualization. Generally speaking, visualization tasks fall into three categories, and the right choice of tools depends upon what your mission really is:

- *Exploratory data analysis*: Here we seek to perform quick, interactive explorations of a given data set. Spreadsheet programs like Excel and notebook-based programming environments like iPython, R, and Mathematica are effective at building the standard plot types. The key here is hiding the complexity, so the plotting routines default to doing something reasonable but can be customized if necessary.

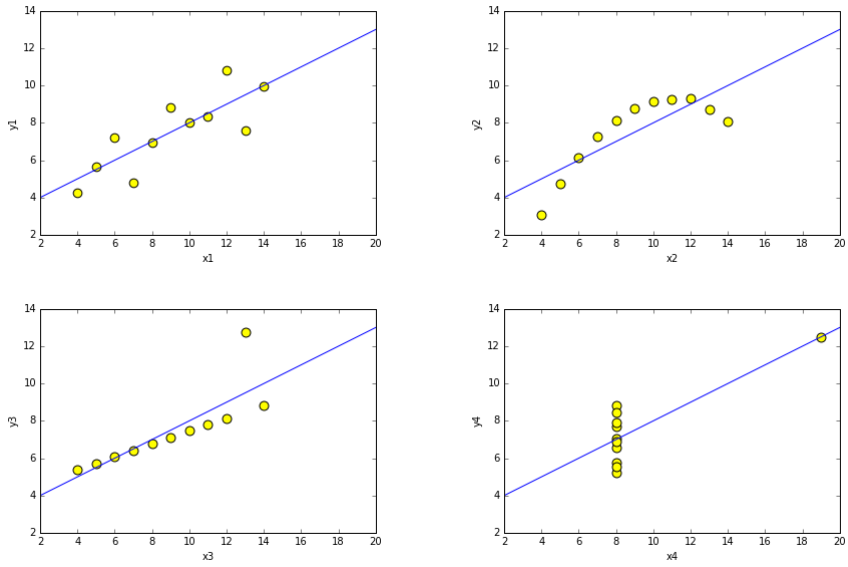

Figure 6.3: Plots of the Ascombe quartet. These data sets are all dramatically different, even though they have identical summary statistics.

- *Publication/presentation quality charts*: Just because Excel is very popular does not mean it produces the best possible graphs/plots. The best visualizations are an interaction between scientist and software, taking full advantage of the flexibility of a tool to maximize the information content of the graphic.

  Plotting libraries like MatPlotLib or Gnuplot support a host of options enabling your graph to look exactly like you want it to. The statistical language R has a very extensive library of data visualizations. Look through catalogs of the plot types supported by your favorite library, to help you find the best representation for your data.

- *Interactive visualization for external applications*: Building *dashboards* that facilitate user interaction with proprietary data sets is a typical task for data science-oriented software engineers. The typical mission here is to build tools that support exploratory data analysis for less technically-skilled, more application-oriented personnel.

  Such systems can be readily built in programming languages like Python, using standard plotting libraries. There is also a class of third-party systems for building dashboards, like Tableau. These systems are programmable at a higher-level than other tools, supporting particular interaction paradigms and linked-views across distinct views of the data.

Figure 6.4: Which painting do you like better? Forming intelligent preferences in art or visualizations depend upon having a distinctive visual aesthetic.

6.2 Developing a Visualization Aesthetic

Sensible appreciation of art or wine requires developing a particular taste or aesthetic. It isn't so much about whether you like something, but figuring out why you like it. Art experts talk about the range of a painter's palate, use of light, or the energy/tension of a composition. Wine connoisseurs testify to the fragrance, body, acidity, and clarity of their favorite plonk, and how much oak or tannin it contains. They always have something better to say than "that tastes good."

Distinguishing good/bad visualizations requires developing a design aesthetic, and a vocabulary to talk about data representations. Figure 6.4 presents two famous landmarks in Western painting. Which one is better? This question is meaningless without a sense of aesthetics and a vocabulary to describe it.

My visual aesthetic and vocabulary is largely derived from the books of Edward Tufte [Tuf83, Tuf90, Tuf97]. He is an artist: indeed I once got to meet him at his former art gallery across from Chelsea Piers in Manhattan. He has thought long and hard about what makes a chart or graph informative and beautiful, basing a design aesthetic on the following principles:

- *Maximize data-ink ratio*: Your visualization is supposed to show off your data. So why is so much of what you see in charts the background grids, shading, and tic-marks?
- *Minimize the lie factor*: As a scientist, your data should reveal the truth, ideally the truth you want to see revealed. But are you being honest with your audience, or using graphical devices that mislead them into seeing something that isn't really there?
- *Minimize chartjunk*: Modern visualization software often adds cool visual effects that have little to do with your data set. Is your graphic interesting because of your data, or in spite of it?
- *Use proper scales and clear labeling*: Accurate interpretation of data depends upon non-data elements like scale and labeling. Are your descriptive

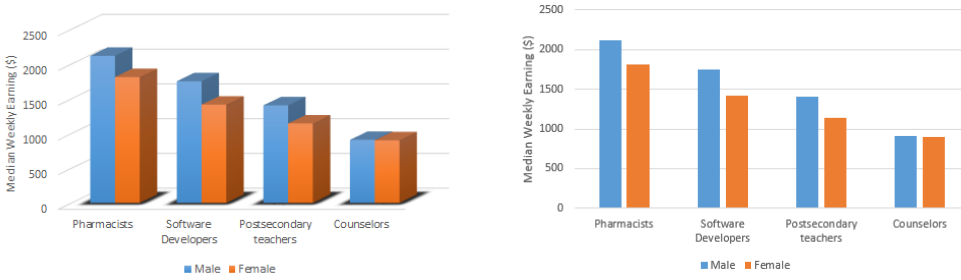

Figure 6.5: Three-dimensional monoliths casting rendered shadows (l) may look impressive. But they really are just chartjunk, which serves to reduce the clarity and data-ink ratio from just showing the data (r).

materials optimized for clarity and precision?

- *Make effective use of color*: The human eye has the power to discriminate between small gradations in hue and color saturation. Are you using color to highlight important properties of your data, or just to make an artistic statement?

- *Exploit the power of repetition*: Arrays of similar graphics with different but related data elements provide a concise and powerful way to enable visual comparisons. Are your chart multiples facilitating comparisons, or merely redundant?

Each of these principles will be detailed in the subsections below.

6.2.1 Maximizing Data-Ink Ratio

In any graphic, some of the ink is used to represent the actual underlying data, while the rest is employed on graphic effects. Generally speaking, visualizations should focus on showing the data itself. We define the *data-ink ratio* to be:

$$\text{Data-Ink Ratio} = \frac{\text{Data-Ink}}{\text{Total ink used in graphic}}$$

Figure 6.5 presents average salary by gender (Bureau of Labor Statistics, 2015), and helps clarify this notion. Which data representation do you prefer? The image on the left says "Cool, how did you make those shadows and that three-dimensional perspective effect?" The image on the right says "Wow, women really are underpaid at all points on the income spectrum. But why is the gap smallest for counselors?"

Maximizing the data-ink ratio lets the data talk, which is the entire point of the visualization exercise in the first place. The flat perspective on the right permits a fairer comparison of the heights of the bars, so the males do not look

like pipsqueaks to the women. The colors do a nice job enabling us to compare apples-to-apples.

There are more extreme ways to increase the data-ink ratio. Why do we need bars at all? The same information could be conveyed by plotting a point of the appropriate height, and would clearly be an improvement were we plotting much more than the eight points shown here. Be aware that less can be more in visualizing data.

6.2.2 Minimizing the Lie Factor

A visualization seeks to tell a true story about what the data is saying. The baldest form of lie is to fudge your data, but it remains quite possible to report your data accurately, yet deliberately mislead your audience about what it is saying. Tufte defines the *lie factor* of a chart as:

$$\text{lie factor} = \frac{\text{(size of an effect in the graphic)}}{\text{(size of the effect in the data)}}$$

Graphical integrity requires minimizing this lie factor, by avoiding the techniques which tend to mislead. Bad practices include:

- *Presenting means without variance*: The data values $\{100, 100, 100, 100, 100\}$ and $\{200, 0, 100, 200, 0\}$ tell different stories, even though both means are 100. If you cannot plot the actual points with the mean, at least show the variance, to make clear the degree to which the mean reflects the distribution.

- *Presenting interpolations without the actual data*: Regression lines and fitted curves are effective at communicating trends and simplifying large data sets. But without showing the data points it is based on, it is impossible to ascertain the quality of the fit.

- *Distortions of scale*: The aspect ratio of a figure can have a huge effect on how we interpret what we are seeing. Figure 6.6 presents three renderings of a given financial time series, identical except for the aspect ratio of the chart.

  In the bottom rendering, the series looks flat: there is nothing to worry about here. On the right, profits have fallen off a cliff: the sky is falling! The left corner plot presents a serious decline, but with signs of an autumn rebound.

  Which plot is right? People are generally used to seeing plots presented according to the Golden ratio, implying that the width should be about 1.6 times the height. Give this shape to them, unless you have well-developed reasons why it is inappropriate. Psychologists inform us that 45 degree lines are the most readily interpretable, so avoid shapes that substantially amplify or mute lines from this objective.

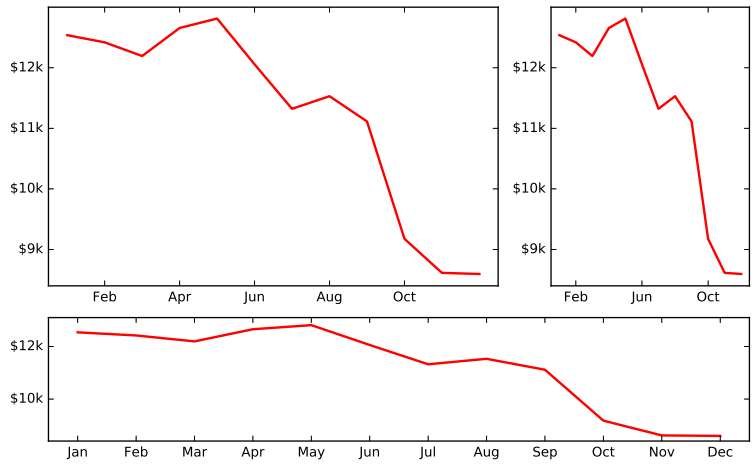

Figure 6.6: Three renderings of the same financial time series. Which most accurately represents the situation?

- *Eliminating tick labels from numerical axes*: Even the worst scale distortions can be completely concealed by not printing numerical reference labels on axes. Only with the numerical scale markings can the actual data values be reconstructed from the plot.

- *Hide the origin point from the plot*: The implicit assumption in most graphs is that the range of values on the y-axis goes from zero to $y_{\max}$. We lose the ability to visually compare magnitudes if the y-range instead goes from $y_{\min} - \epsilon$ to $y_{\max}$. The largest value suddenly looks many times larger than the smallest value, instead of being scaled to the proper proportion.

  If Figure 6.5 (right) were drawn with a tight y-range $[900, 2500]$, the message would be that counselors were starving, instead of earning salaries as close to teachers as software developers are to pharmacists. Such deceptions can be recognized provided the scales are marked on the axis, but are hard to catch.

Despite Tufte's formula, the lie factor cannot be computed mechanically, because it requires understanding the agenda that is behind the distortion. In reading any graph, it is important to know who produced it and why. Understanding their agenda should sensitize you to potentially misleading messages encoded in the graphic.

6.2.3 Minimizing Chartjunk

Extraneous visual elements distract from the message the data is trying to tell. In an exciting graphic, the data tells the story, not the chartjunk.

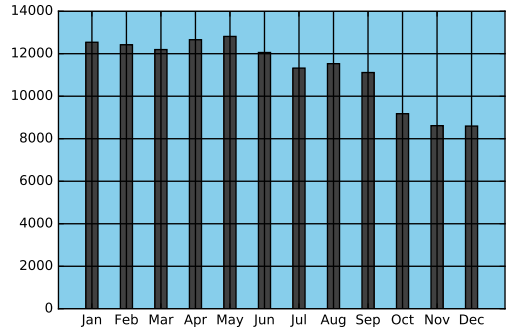

Figure 6.7: A monthly time series of sales. How can we improve/simplify this bar chart time series?

Figure 6.7 presents a monthly time series of sales at a company beginning to encounter bad times. The graphic in question is a *bar plot*, a perfectly sound way to represent time series data, and is drawn using conventional, perhaps default, options using a reasonable plotting package.

But can we simplify this plot by removing elements to make the data stand out better? Think about this for a minute before peeking at Figure 6.8, which presents a series of four successive simplifications to this chart. The critical operations are:

- *Jailbreak your data (upper left)*: Heavy grids imprison your data, by visually dominating the contents. Often graphs can be improved by removing the grid, or at least lightening it.

  The potential value of the data grid is that it facilitates more precise interpretation of numerical quantities. Thus grids tend to be most useful on plots with large numbers of values which may need to be accurately quoted. Light grids can adequately manage such tasks.

- *Stop throwing shade (upper right)*: The colored background here contributes nothing to the interpretation of the graphic. Removing it increases the data-ink ratio, and makes it less obtrusive.

- *Think outside the box (lower left)*: The bounding box does not really contribute information, particularly the upper and rightmost boundaries which do not define axes. Take them out, and let more air into your plots.

- *Make missing ink work for you (lower right)*: The effect of the reference grid can be recovered by removing lines from the bars instead of adding elements. This makes it easier to compare the magnitude of biggest numbers, by focusing attention on big changes in the relatively small top piece, instead of small changes in the long bar.

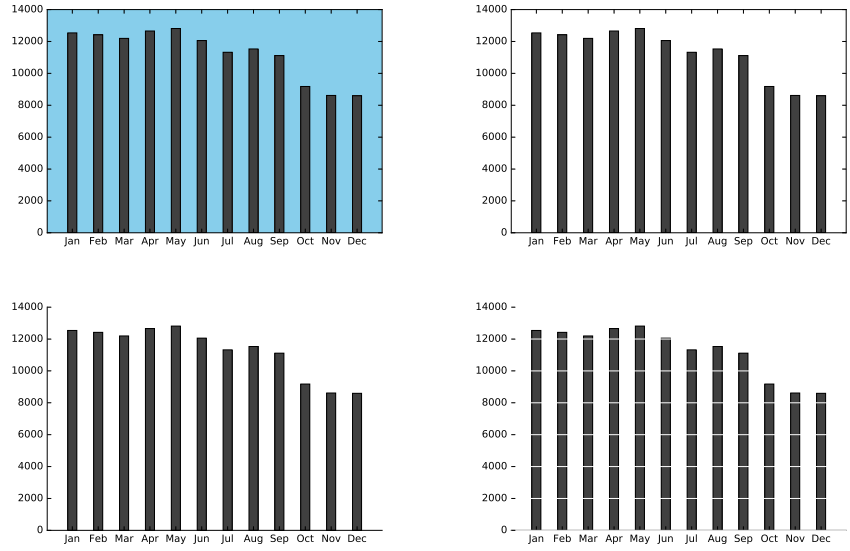

Figure 6.8: Four successive simplifications of Figure 6.7, by removing extraneous non-data elements.

The architect Mies van der Rohe famously said that "less is more." Removing elements from plots often improves them far more than adding things. Make this part of your graphical design philosophy.

6.2.4 Proper Scaling and Labeling

Deficiencies in scaling and labeling are the primary source of intentional or accidental misinformation in graphs. Labels need to report the proper magnitude of numbers, and scale needs to show these numbers to the right resolution, in a way to facilitate comparison. Generally speaking, data should be scaled so as to fill the space allotted to it on the chart.

Reasonable people can differ as to whether to scale the axes over the full theoretical range of the variable, or cut it down to reflect only the observed values. But certain decisions are clearly unreasonable.

Figure 6.9 (left) was produced by a student of mine, presenting the correlation between two variables for almost a hundred languages. Because the correlation ranges between $[-1, 1]$, he forced the plot to respect this interval. The vast sea of white in this plot captures only the notion that we might have done better, by getting the correlation closer to 1.0. But the chart is otherwise unreadable.

Figure 6.9 (right) presents exactly the same data, but with a truncated scale. *Now* we can see where there are increases in performance as we move from left to right, and read off the score for any given language. Previously, the bars

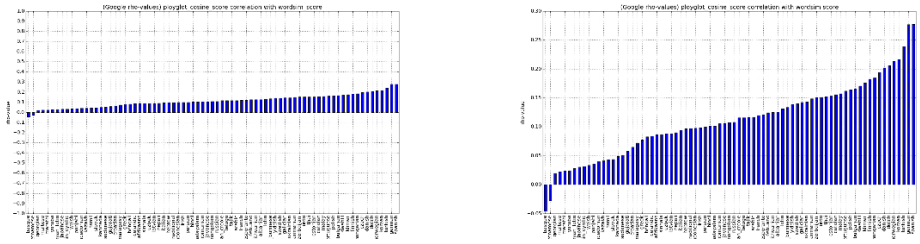

Figure 6.9: Scaling over the maximum possible range (left) is silly when all it shows is white space. Better scaling permits more meaningful comparisons (right).

were so far removed from the labels that it was difficult to make the name–bar correspondence at all.

The biggest sin of truncated scales comes when you do not show the whole of each bar, so the length of the bar no longer reflects the relative value of the variable. We show the $y = 0$ line here, helping the reader to know that each bar must be whole. Getting the data out of its prison grid would also have helped.

6.2.5 Effective Use of Color and Shading

Colors are increasingly assumed as part of any graphical communication. Indeed, I was pleased to learn that my publisher's printing costs are now identical for color and black-and-white, so you the reader are not paying any more to see my color graphics here.

Colors play two major roles in charts, namely marking class distinctions and encoding numerical values. Representing points of different types, clusters, or classes with different colors encodes another layer of information on a conventional dot plot. This is a great idea when we are trying to establish the extent of differences in the data distribution across classes. The most critical thing is that the classes be easily distinguishable from each other, by using bold primary colors.

It is best when the colors are selected to have mnemonic values to link naturally to the class at hand. Losses should be printed in red ink, environmental causes associated with green, nations with their flag colors, and sports teams with their jersey colors. Coloring points to represent males as blue and females as red offers a subtle clue to help the viewer interpret a scatter plot, as shown in Figure 9.17.

Selecting colors to represent a numerical scale is a more difficult problem. Rainbow color maps are perceptually non-linear, meaning it is not obvious to anyone whether purple lies before or after green. Thus while plotting numbers in rainbow colors groups similar numbers in similar colors, the relative magnitudes are imperceptible without explicitly referencing the color scale. Figure 6.10 presents several color scales from Python's MatPlotLib, for comparison.

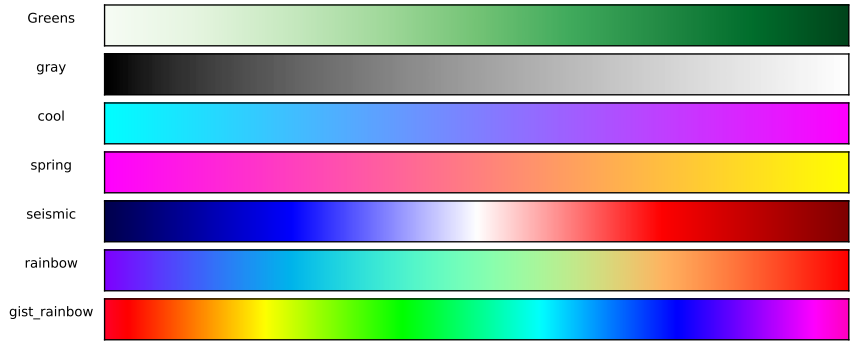

Figure 6.10: Color scales from Python's MatPlotLib, varying hue, saturation, and brightness. Rainbow color maps are perceptually non-linear, making it difficult to recognize the magnitudes of differences.

Much better are color scales based on a varying either brightness or saturation. The *brightness* of a color is modulated by blending the hue with a shade of gray, somewhere between white and black. *Saturation* is controlled by mixing in a fraction of gray, where 0 produces the pure hue, and 1 removes all color.

Another popular color scale features distinct positive/negative colors (say, blue and red, as in the seismic color scale of Figure 6.10) reflected around a white or gray center at zero. Thus hue tells the viewer the polarity of the number, while brightness/saturation reflects magnitude. Certain color scales are much better for color-blind people, particularly those avoiding use of red and green.

As a general rule, large areas on plots should be shown with unsaturated colors. The converse is true for small regions, which stand out better with saturated colors. Color systems are a surprisingly technical and complicated matter, which means that you should always use well established color scales, instead of inventing your own.

6.2.6 The Power of Repetition

Small multiple plots and tables are excellent ways to represent multivariate data. Recall the power of grids showing all bivariate distributions in Figure 6.1.

There are many applications of small multiple charts. We can use them to break down a distribution by classes, perhaps plotting separate but comparable charts by region, gender, or time period. Arrays of plots facilitate comparisons: what has changed between different distributions.

Time series plots enable us to compare the same quantities at different calendar points. Even better is to compare multiple time series, either as lines on the same plot, or multiple plots in a logical array reflecting their relationship.

6.3 Chart Types

In this section, we will survey the rationale behind the primary types of data visualizations. For each chart, I present best practices for using them, and outline the degrees of freedom you have to make your presentation as effective as possible.

Nothing says "Here's a plot of some data" like a thoughtlessly-produced graphic, created using the default settings of some software tool. My students present me with such undigested data products way too often, and this section is somewhat of a personal reaction against it.

> *Take-Home Lesson*: You have the power and responsibility to produce meaningful and interpretable presentations of your work. Effective visualization involves an iterative process of looking at the data, deciding what story it is trying to tell, and then improving the display to tell the story better.

Figure 6.11 presents a handy decision tree to help select the right data representation, from Abela [Abe13]. The most important charts will be reviewed in this section, but use this tree to better understand *why* certain visualizations are more appropriate in certain contexts. We need to produce the right plot for a given data set, not just the first thing that comes to mind.

6.3.1 Tabular Data

Tables of numbers can be beautiful things, and are very effective ways to present data. Although they may *appear* to lack the visual appeal of graphic presentations, tables have several advantages over other representations, including:

- *Representation of precision*: The resolution of a number tells you something about the process of how it was obtained: an average salary of \$79,815 says something different than \$80,000. Such subtleties are generally lost on plots, but nakedly clear in numerical tables.
- *Representation of scale*: The digit lengths of numbers in a table can be likened to bar charts, on a logarithmic scale. Right-justifying numbers best communicates order-of-magnitude differences, as (to a lesser extent) does scanning the leading digits of numbers in a column.

left	center	right
1	1	1
10	10	10
100	100	100
1000	1000	1000

  Left justifying numbers prevents such comparisons, so always be sure to right justify them.

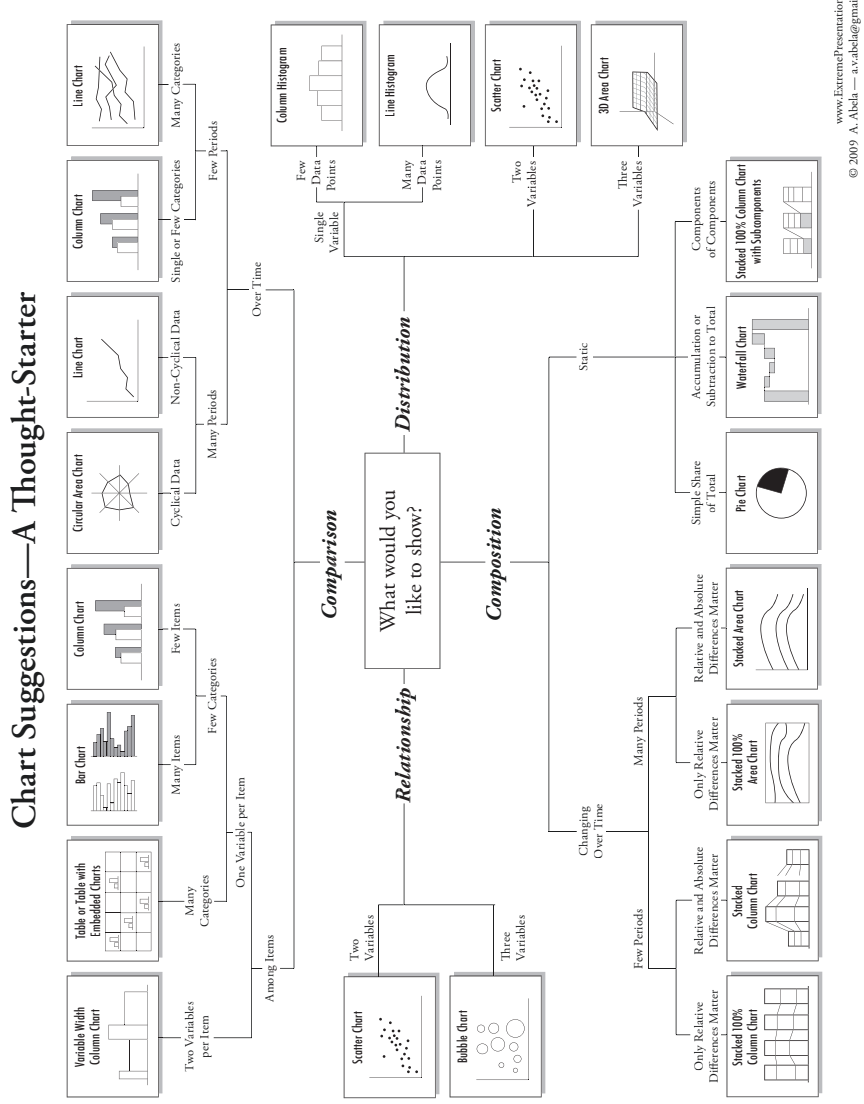

Figure 6.11: A clever decision tree to help identify the best visual representation for representing data. Reprinted with permission from Abela [Abe13].

- *Multivariate visualization*: Geometry gets complicated to understand once we move beyond two dimensions. But tables can remain manageable even for large numbers of variables. Recall Babe Ruth's baseball statistics from Figure 1.1, a table of twenty-eight columns which is readily interpretable by any knowledgeable fan.
- *Heterogeneous data*: Tables generally are the best way to present a mix of numerical and categorical attributes, like text and labels. Glyphs like emojis can even be used to represent the values of certain fields.
- *Compactness*: Tables are particularly useful for representing small numbers of points. Two points in two dimensions can be drawn as a line, but why bother? A small table is generally better than a sparse visual.

Presenting tabular data seems simple to do ("just put it in a table"), akin to rapping a leg with a stick. But subtleties go into producing the most informative tables. Best practices include:

- *Order rows to invite comparison*: You have the freedom to order the rows in a table any way you want, so take advantage of it. Sorting the rows according to the values of an important column is generally a good idea. Thus grouping the rows is valuable to facilitate comparison, by putting likes with likes.

  Sorting by size or date can be more revealing than the name in many contexts. Using a canonical order of the rows (say, lexicographic by name) can be helpful for looking up items by name, but this is generally not a concern unless the table has many rows.
- *Order columns to highlight importance, or pairwise relationships*: Eyes darting from left-to-right across the page cannot make effective visual comparisons, but neighboring fields are easy to contrast. Generally speaking, columns should be organized to group similar fields, hiding the least important ones on the right.
- *Right-justify uniform-precision numbers*: Visually comparing 3.1415 with 39.2 in a table is a hopeless task: the bigger number has to look bigger. Best is to right justify them, and set all to be the same precision: 3.14 vs. 39.20.
- *Use* **emphasis**, font, *or color to highlight important entries*: Marking the extreme values in each column so they stand out reveals important information at a glance. It is easy to overdo **this**, however, so strive for subtlety.
- *Avoid excessive-length column descriptors*: White ribbons in tables are distracting, and usually result from column labels that are longer than the values they represent. Use abbreviations or multiple-line word stacking to minimize the problem, and clarify any ambiguity in the caption attached to the table.

To help illustrate these possible sins, here is a table recording six properties of fifteen different nations, with the row and column orders given at random. Do you see any possible ways of improving it?

Country	Area	Density	Birthrate	Population	Mortality	GDP
Russia	17075200	8.37	99.6	142893540	15.39	8900.0
Mexico	1972550	54.47	92.2	107449525	20.91	9000.0
Japan	377835	337.35	99.0	127463611	3.26	28200.0
United Kingdom	244820	247.57	99.0	60609153	5.16	27700.0
New Zealand	268680	15.17	99.0	4076140	5.85	21600.0
Afghanistan	647500	47.96	36.0	31056997	163.07	700.0
Israel	20770	305.83	95.4	6352117	7.03	19800.0
United States	9631420	30.99	97.0	298444215	6.5	37800.0
China	9596960	136.92	90.9	1313973713	24.18	5000.0
Tajikistan	143100	51.16	99.4	7320815	110.76	1000.0
Burma	678500	69.83	85.3	47382633	67.24	1800.0
Tanzania	945087	39.62	78.2	37445392	98.54	600.0
Tonga	748	153.33	98.5	114689	12.62	2200.0
Germany	357021	230.86	99.0	82422299	4.16	27600.0
Australia	7686850	2.64	100.0	20264082	4.69	29000.0

There are many possible orderings of the rows (countries). Sorting by any single column is an improvement over random, although we could also group them by region/continent. The order of the columns can be made more understandable by putting like next to like. Finally, tricks like right-justifying numbers, removing uninformative digits, adding commas, and highlighting the biggest value in each column makes the data easier to read:

Country	Population	Area	Density	Mortality	GDP	Birth Rate
Afghanistan	31,056,997	647,500	47.96	**163.07**	700	36.0
Australia	20,264,082	7,686,850	2.64	4.69	29,000	**100.0**
Burma	47,382,633	678,500	69.83	67.24	1,800	85.3
China	**1,313,973,713**	9,596,960	136.92	24.18	5,000	90.9
Germany	82,422,299	357,021	230.86	4.16	27,600	99.0
Israel	6,352,117	20,770	305.83	7.03	19,800	95.4
Japan	127,463,611	377,835	**337.35**	3.26	28,200	99.0
Mexico	107,449,525	1,972,550	54.47	20.91	9,000	92.2
New Zealand	4,076,140	268,680	15.17	5.85	21,600	99.0
Russia	142,893,540	**17,075,200**	8.37	15.39	8,900	99.6
Tajikistan	7,320,815	143,100	51.16	110.76	1,000	99.4
Tanzania	37,445,392	945,087	39.62	98.54	600	78.2
Tonga	114,689	748	153.33	12.62	2,200	98.5
United Kingdom	60,609,153	244,820	247.57	5.16	27,700	99.0
United States	298,444,215	9,631,420	30.99	6.50	**37,800**	97.0

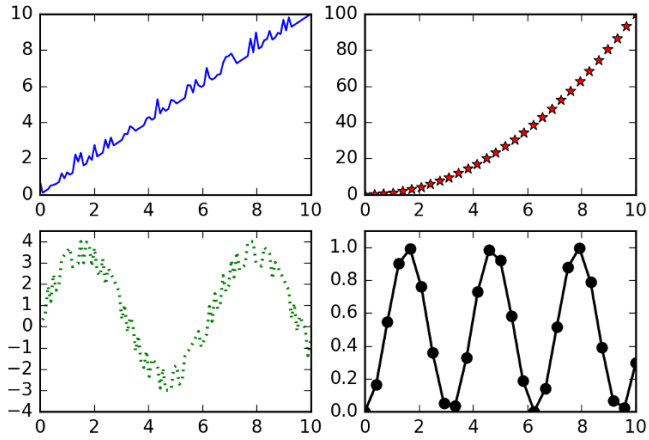

Figure 6.12: Many of the line chart styles that we have seen are supported by Python's MatPlotLib package.

6.3.2 Dot and Line Plots

Dot and line plots are the most ubiquitous forms of data graphic, providing a visual representation of a function $y = f(x)$ defined by a set of (x, y) points. Dot plots just show the data points, while line plots connect them or interpolate to define a continuous function $f(x)$. Figure 6.12 shows several different styles of line plots, varying in the degree of emphasis they give the points vs. the interpolated curve. Advantages of line charts include:

- *Interpolation and fitting*: The interpolation curve derived from the points provides a prediction for $f(x)$ over the full range of possible x. This enables us to sanity check or reference other values, and make explicit the trends shown in the data.

  Overlaying a fitted or smoothed curve on the same graph as the source data is a very powerful combination. The fit provides a model explaining what the data says, while the actual points enable us to make an educated judgment of how well we trust the model.

- *Dot plots*: A great thing about line plots is that you don't actually have to show the line, resulting in a *dot plot*. Connecting points by line segments (polylines) proves misleading in many situations. If the function is only defined at integer points, or the x-values represent distinct conditions, then it makes no sense at all to interpolate between them.

  Further, polylines dramatically swing out of their way to capture outliers, thus visually encouraging us to concentrate on exactly the points we should most ignore. High frequency up-and-down movement distracts us from

seeing the broader trend, which is the primary reason for staring at a chart.

Best practices with line charts include:

- *Show data points, not just fits*: It is generally important to show the actual data, instead of just the fitted or interpolated lines.

  The key is to make sure that one does not overwhelm the other. To represent large numbers of points unobtrusively, we can (a) reduce the size of the points, possibly to pinpricks, and/or (b) lighten the shade of the points so they sit in the background. Remember that there are fifty shades of gray, and that subtlety is the key.

- *Show the full variable range if possible*: By default, most graphic software plots from $x_{\min}$ to $x_{\max}$ and $y_{\min}$ to $y_{\max}$, where the mins and maxes are defined over the input data values. But the logical min and max are context specific, and it can reduce the lie factor to show the full range. Counts should logically start from zero, not $y_{\min}$.

  But sometimes showing the full range is uninformative, by completely flattening out the effect you are trying to illustrate. This was the case in Figure 6.9. One possible solution is to use a log scale for the axis, so as to embed a wider range of numbers in a space efficient way. But if you must truncate the range, make clear what you are doing, by using axis labels with tick marks, and clarifying any ambiguity in the associated caption.

- *Admit uncertainty when plotting averages*: The points appearing on a line or dot plot are often obtained by averaging multiple observations. The resulting mean better captures the distribution than any single observation. But means have differing interpretations based on variance. Both $\{8.5, 11.0, 13.5, 7.0, 10.0\}$ and $\{9.9, 9.6, 10.3, 10.1, 10.1\}$ average to be 10.0, but the degree to which we trust them to be precise differs substantially.

  There are several ways to confess the level of measurement uncertainty in our plots. My favorite simply plots all the underlying data values on the same graph as the means, using the same x-value as the associated mean. These points will be visually unobtrusive relative to the heavier trend line, provided they are drawn as small dots and lightly shaded, but they are now available for inspection and analysis.

  A second approach plots the standard deviation σ around y as a whisker, showing the interval $[y - \sigma, y + \sigma]$. This interval representation is honest, denoting the range with 68% of the values under a normal distribution. Longer whiskers means you should be more suspicious of the accuracy of the means, while short whiskers implies greater precision.

  Box plots concisely record the range and distribution of values at a point with a box. This box shows the range of values from the quartiles (25% and 75%), and is cut at the median (50th percentile). Typically whiskers

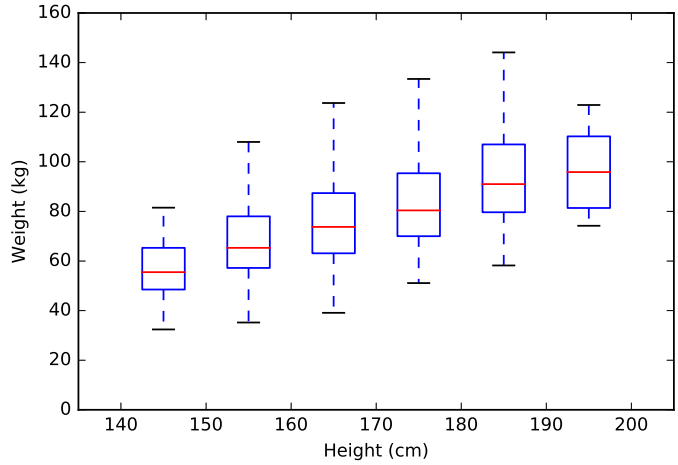

Figure 6.13: Box and whisker plots concisely show the range/quartiles (i.e. median and variance) of a distribution.

(hairs) are added to show the range of the highest and lowest values. Figure 6.13 shows a box-and-whisker plot of weight as a function of height in a population sample. The median weight increases with height, but not the maximum, because fewer points in the tallest bucket reduces the chance for an outlier maximum value.

Real scientists seem to love box-and-whisker plots, but I personally find them to be overkill. If you really can't represent the actual data points, perhaps just show the contour lines flanking your mean/median at the 25th and 75th percentiles. This conveys exactly the same information as the box in the box plot, with less chartjunk.

- *Never connect points for categorical data*: Suppose you measure some variable (perhaps, median income) for several different classes (say, the fifty states, from Alabama to Wyoming). It might make sense to display this as a dot plot, with $1 \leq x \leq 50$, but it would be silly and misleading to connect the points. Why? Because there is no meaningful adjacency between state i and state $i+1$. Indeed, such graphs are better off thought of as bar charts, discussed in Section 6.3.4

  Indeed, connecting points by polylines is very often chartjunk. Trend or fit lines are often more revealing and informative. Try to show the raw data points themselves, albeit lightly and unobtrusively.

- *Use color and hatching to distinguish lines/classes*: Often we are faced representing the same function $f(x)$ drawn over two or more classes, perhaps income as a function of schooling, separately for men and women.

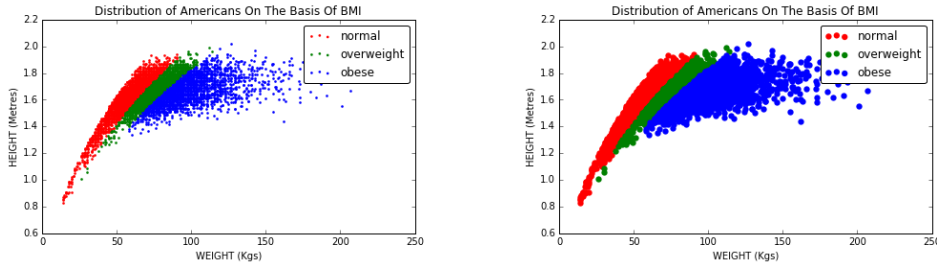

Figure 6.14: Smaller dots on scatter plots (left) reveal more detail than the default dot size (right).

These are best handled by assigning distinct colors to the line/points for each class. Line hatchings (dotted, dashed, solid, and bold) can also be used, but these are often harder to distinguish than colors, unless the output media is black and white. In practice, two to four such lines can be distinguished on a single plot, before the visual collapses into a mess. To visualize a large number of groups, partition them into logical clusters and use multiple line plots, each with few enough lines to be uncluttered.

6.3.3 Scatter Plots

Massive data sets are a real challenge to present in an effective manner, because large numbers of points easily overwhelm graphic representations, resulting in an image of the black ball of death. But when properly drawn, scatter plots are capable of showing thousands of bivariate (two-dimensional) points in a clear understandable manner.

Scatter plots show the values of every (x, y) point in a given data set. We used scatter plots in Section 4.1 to represent the body mass status of individuals by representing them as points in height–weight space. The color of each point reflected their classification as normal, overweight, or obese. Best practices associated with scatter plots include:

- *Scatter the right-sized dots*: In the movie *Oh G-d*, George Burns as the creator looks back on the avocado as his biggest mistake. Why? Because he made the pit too large. Most people's biggest mistake with scatter plots is that they make the points too large.

  The scatter plots in Figure 6.14 show the BMI distribution for over 1,000 Americans, with two different sizes of dots. Observe how the smaller dots show finer structure, because they are less likely to overlap and obscure other data points.

  Now we see a fine structure of a dense core, while still being able to detect the light halo of outliers. The default dot size for most plotting programs

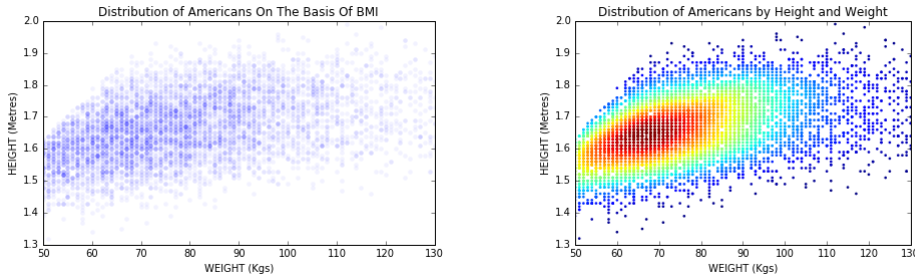

Figure 6.15: Overlapping dots can obscure scatter plots, particularly for large data sets. Reducing the opacity of the dots (left) shows some of the fine structure of the data (left). But a colored heatmap more dramatically reveals the distribution of the points (right).

is appropriate for about fifty points. But for larger data sets, use smaller dots.

- *Color or jiggle integer points before scatter-plotting them*: Scatter plots reveal grid patterns when the x and y values have integer values, because there are no smooth gradations between them. These scatter plots look unnatural, but even worse tend to obscure data because often multiple points will share exactly the same coordinates.

  There are two reasonable solutions. The first is to color each point based on its frequency of occurrence. Such plots are called *heatmaps*, and concentration centers become readily visible provided that a sensible color scale is used. Figure 6.15 shows a heatmap of height–weight data, which does a much better job of revealing point concentrations than the associated simple dot plot.

  A related idea is to reduce the *opacity* (equivalently, increase the *transparency*) of the points we scatter plot. By default, points are generally drawn to be opaque, yielding a mass when there are overlapping points. But now suppose we permit these points to be lightly shaded and transparent. Now the overlapping points show up as darker than singletons, yielding a heatmap in multiple shades of gray.

  The second approach is to add a small amount of random noise to each point, to jiggle it within a sub-unit radius circle around its original position. Now we will see the full multiplicity of points, and break the distracting regularity of the grid.

- *Project multivariate data down to two dimensions, or use arrays of pairwise plots*: Beings from our universe find it difficult to visualize data sets in four or more dimensions. Higher-dimensional data sets can often be projected down to two dimensions before rendering them on scatter plots, using techniques such as principal component analysis and self-organizing

maps. A pretty example appears a few chapters ahead in Figure 11.16, where we project a hundred dimensions down to two, revealing a very coherent view of this high-dimensional data set.

The good thing about such plots is that they can provide effective views of things we couldn't otherwise see. The bad thing is that the two dimensions no longer mean anything. More specifically, the new dimensions do not have variable names that can convey meaning, because each of the two "new" dimensions encodes properties of all the original dimensions.

An alternative representation is to plot a grid or lattice of all pairwise projections, each showing just two of the original dimensions. This is a wonderful way to get a sense of which pairs of dimensions correlate with each other, as we showed in Figure 6.1.

- *Three-dimensional-scatter plots help only when there is real structure to show*: TV news stories about data science always feature some researcher gripping a three-dimensional point cloud and rotating it through space, striving for some important scientific insight. They never find it, because the view of a cloud from any given direction looks pretty much the same as the view from any other direction. There generally isn't a vantage point where it suddenly becomes clear how the dimensions interact.

  The exception is when the data was actually derived from structured three-dimensional objects, such as laser scans of a given scene. Most of the data we encounter in data science doesn't fit this description, so have low expectations for interactive visualization. Use the grid of all two-dimensional projections technique, which essentially views the cloud from all orthogonal directions.

- *Bubble plots vary color and size to represent additional dimensions*: Modulating the color, shape, size, and shading of dots enables dot plots to represent additional dimensions, on bubble plots. This generally works better than plotting points in three dimensions.

  Indeed Figure 6.16 neatly shows four dimensions (GDP, life expectancy, population, and geographic region) using x, y, size, and color, respectively. There is much to read into such a bubble chart: it clearly reveals the correlation between GDP and health (by the straight line fit, although note that the x-values are not linearly-spaced), the new world is generally richer than the old world, and that the biggest countries (China and India) generally sit in the middle of the pack. Certain rich but sick nations are doing abysmally by their people (e.g. South Africa), while countries like Cuba and Vietnam seem to be punching above their weight.

6.3.4 Bar Plots and Pie Charts

Bar plots and pie charts are tools for presenting the relative proportions of categorical variables. Both work by partitioning a geometric whole, be it a

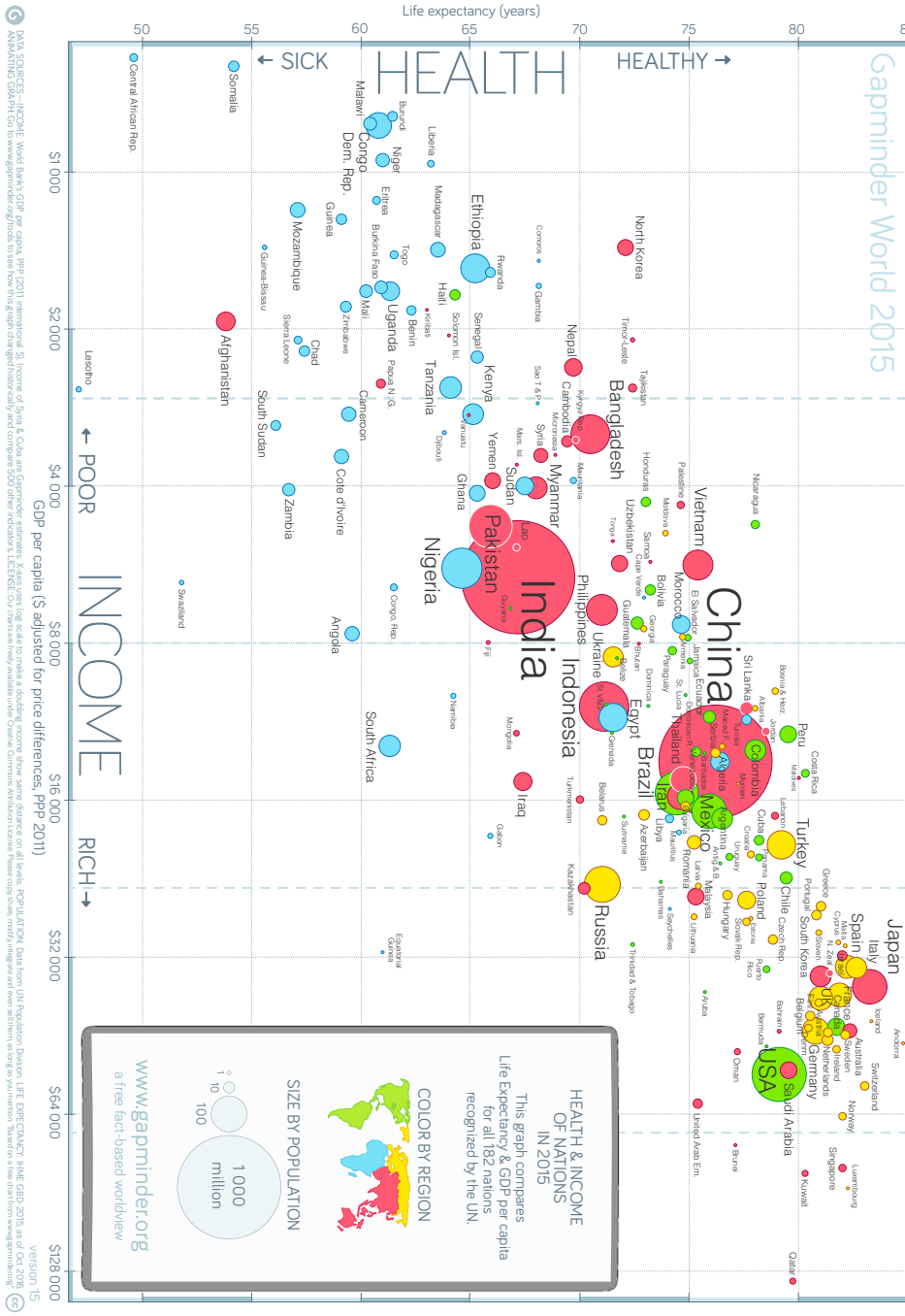

Figure 6.16: Adding size and color to a scatter plot leads to natural four-dimensional visualizations, here illustrating properties of the world's nations. Based on a free chart from `http://www.gapminder.org`.

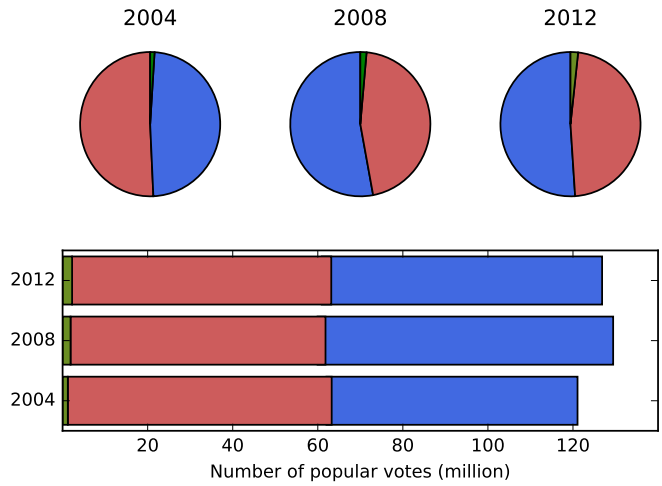

Figure 6.17: Voter data from three U.S. presidential elections. Bar plots and pie charts display the frequency of proportion of categorical variables. Relative magnitudes in a time series can be displayed by modulating the area of the line or circle.

bar or a circle, into areas proportional to the frequency of each group. Both elements are effective in multiples, to enable comparisons. Indeed, partitioning each bar into pieces yields the *stacked bar chart*.

Figure 6.17 shows voter data from three years of U.S. presidential elections, presented as both pie and bar charts. The blue represents Democratic votes, the red Republican votes. The pies more clearly shows which side won each elections, but the bars show the Republican vote totals have stayed fairly constant while the Democrats were generally growing. Observe that these bars can be easily compared because they are left-justified.

Certain critics get whipped into an almost religious fever against pie charts, because they take up more space than necessary and are generally harder to read and compare. But pie charts are arguably better for showing percentages of totality. Many people seem to like them, so they are probably harmless in small quantities. Best practices for bar plots and pie charts include:

- *Directly label slices of the pie*: Pie charts are often accompanied by legend keys labeling what each color slice corresponds to. This is very distracting, because your eyes must move back and forth between the key and the pie to interpret this.

  Much better is to label each slice directly, inside the slice or just beyond the rim. This has a secondary benefit of discouraging the use of too many slices, because slivers generally become uninterpretable. It helps to group the slivers into a single slice called *other*, and then perhaps present a

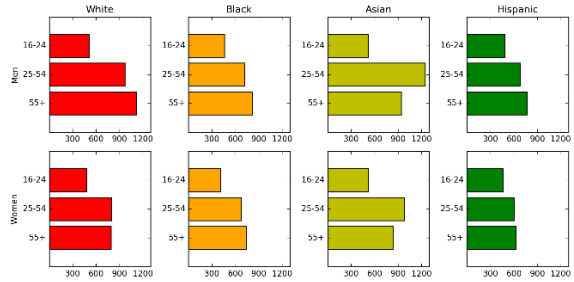

Figure 6.18: Small multiple bar plots/tables are excellent ways to represent multivariate data for comparison.

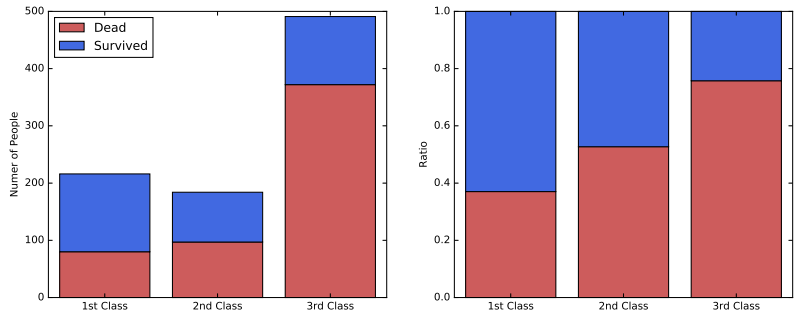

Figure 6.19: Stacked bar charts illustrating the survivorship rate on the doomed ship Titanic, by ticket class. The histogram (left) informs us of the size of each class, but scaled bars (right) better capture proportions. Primary conclusion: you were better off not to be traveling steerage (third class).

second pie chart decomposed into the major *other* components.

- *Use bar charts to enable precise comparisons*: When anchored on a fixed line, arrays of bars make it easy to identify the minimum and maximum values in a series, and whether a trend is increasing or decreasing.

  Stacked bar charts are concise, but harder to use for such purposes. Presenting an array of small bar charts, here one for each gender/ethnic group, empowers us to make such fine comparisons, as shown in Figure 6.18.

- *Scale appropriately, depending upon whether you seek to highlight absolute magnitude or proportion*: Pie charts exist to represent fractions of the whole. In presenting a series of pie or bar charts, your most critical decision is whether you want to show the size of the whole, or instead the fractions of each subgroup.

  Figure 6.19 shows two stacked bar charts presenting survivorship statistics

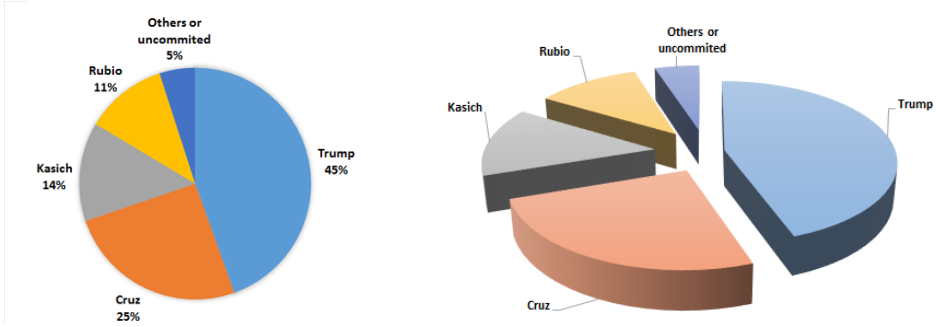

Figure 6.20: Pie charts of delegates to the 2016 Republican convention by candidate. Which one is better and why?

on the doomed ship *Titanic*, reported by ticket class. The histogram (left) precisely records the sizes of each class and the resulting outcomes. The chart with equal length bars (right) better captures how the mortality rate increased for lower classes.

Pie charts can also be used to show changes in magnitude, by varying the area of the circle defining the pie. But it is harder for the eye to calculate area than length, making comparisons difficult. Modulating the radius to reflect magnitude instead of area is even more deceptive, because doubling the radius of a circle multiplies the area by four.

Bad Pie Charts

Figure 6.20 shows two pie charts reporting the distribution of delegates to the 2016 Republican convention, by candidate. The left pie is two-dimensional, while the chart on right has thick slices neatly separated to show off this depth. Which one is better at conveying the distribution of votes?

It should be clear that the three-dimensional effects and separation are pure chartjunk, that only obscures the relationship between the size of the slices. The actual data values disappeared as well, perhaps because there wasn't enough space left for them after all those shadows. But why do we need a pie chart at all? A little table of labels/colors with an extra column with percentages would be more concise and informative.

6.3.5 Histograms

The interesting properties of variables or features are defined by their underlying frequency distribution. Where is the peak of the distribution, and is the mode near the mean? Is the distribution symmetric or skewed? Where are the tails? Might it be bimodal, suggesting that the distribution is drawn from a mix of two or more underlying populations?

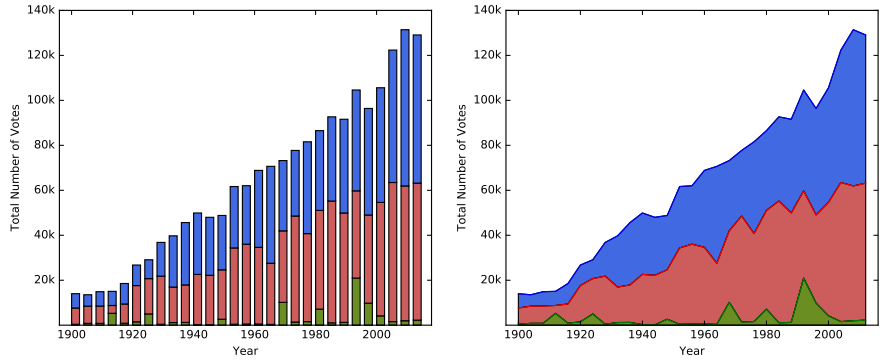

Figure 6.21: Time series of vote totals in U.S. presidential elections by party enable us to see the changes in magnitude and distribution. Democrats are shown in blue, and Republicans in red. It is hard to visualize changes, particularly in the middle layers of the stack.

Often we are faced with a large number of observations of a particular variable, and seek to plot a representation for them. *Histograms* are plots of the observed frequency distributions. When the variable is defined over a large range of possible values relative to the n observations, it is unlikely we will ever see any exact duplication. However, by partitioning the value range into an appropriate number of equal-width bins, we can accumulate different counts per bin, and approximate the underlying probability distribution.

The biggest issue in building a histogram is deciding on the right number of bins to use. Too many bins, and there will be only a few points in even the most popular bucket. We turned to binning to solve exactly this problem in the first place. But use too few bins, and you won't see enough detail to understand the shape of the distribution.

Figure 6.22 illustrates the consequences of bin size on the appearance of a histogram. The plots in the top row bin 100,000 points from a normal distribution into ten, twenty, and fifty buckets respectively. There are enough points to fill fifty buckets, and the distribution on right looks beautiful. The plots in the bottom row bin have only 100 points, so the thirty-bucket plot on right is sparse and scraggy. Here seven bins (shown on the left) seems to produce the most representative plot.

It is impossible to give hard and fast rules to select the best bin count b for showing off your data. Realize you will never be able to discriminate between more than a hundred bins by eye, so this provides a logical upper bound. In general, I like to see an average of 10 to 50 points per bin to make things smooth, so $b = \lceil n/25 \rceil$ gives a reasonable first guess. But experiment with different values of b, because the right bin count will work much better than the others. You will know it when you see it.

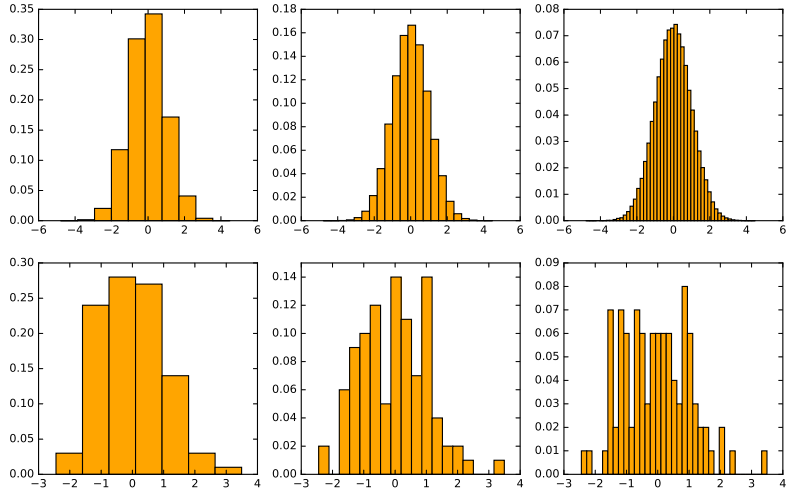

Figure 6.22: Histograms of a given distribution can look wildly different depending on the number of bins. A large data set benefits from many bins (top). But the structure of a smaller data set is best shown when each bin a non-trivial number of element.

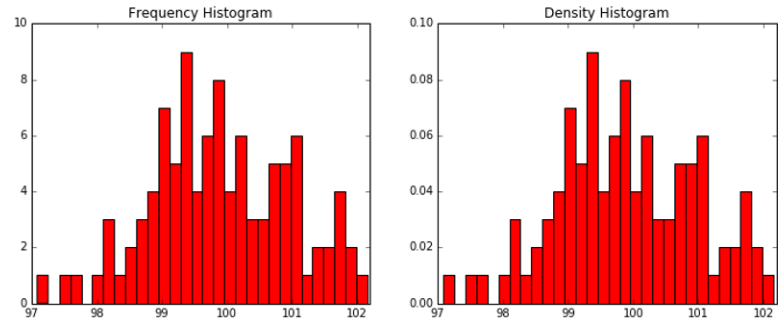

Figure 6.23: Dividing counts by the total yields a probability density plot, which is more generally interpretable even though the shapes are identical.

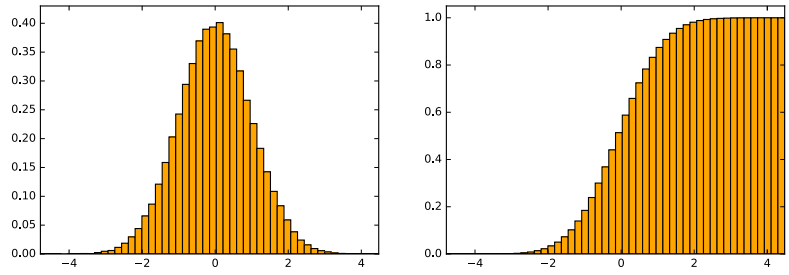

Figure 6.24: Generally speaking, histograms are better for displaying peaks in a distribution, but cdfs are better for showing tails.

Best practices for histograms include:

- *Turn your histogram into a pdf*: Typically, we interpret our data as observations that approximate the probability density function (pdf) of an underlying random variable. If so, it becomes more interpretable to label the y-axis by the fraction of elements in each bucket, instead of the total count. This is particularly true in large data sets, where the buckets are full enough that we are unconcerned about the exact level of support.

  Figure 6.23 shows the same data, plotted on the right as a pdf instead of a histogram. The shape is exactly the same in both plots: all that changes is the label on the y-axis. Yet the result is easier to interpret, because it is in terms of probabilities instead of counts.

- *Consider the cdf*: The cumulative density function (cdf) is the integral of the pdf, and the two functions contain exactly the same information. So consider using a cdf instead of a histogram to represent your distribution,

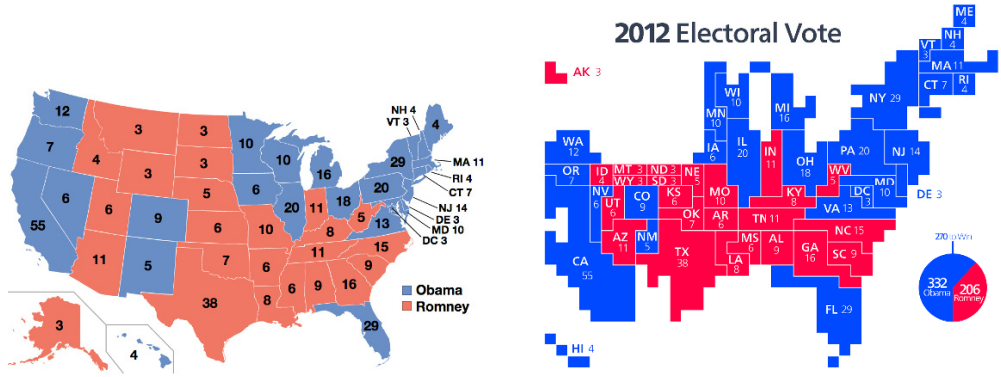

Figure 6.25: Maps summarizing the results of the 2012 U.S. presidential election. The recipient of each state's electoral votes is effectively presented on a data map (left), while a cartogram making the area of each state proportional to the number of votes (right) better captures the magnitude of Obama's victory. *Source*: Images from Wikipedia.

as shown in Figure 6.24.

One great thing about plotting the cdf is that it does not rely on a bin count parameter, so it presents a true, unadulterated view of your data. Recall how great the cdfs looked in the Kolmogorov-Smirnov test, Figure 5.13. We draw a cdf as a line plot with $n + 2$ points for n observations. The first and last points are $(x_{\min} - \epsilon, 0)$ and $(x_{\max} + \epsilon, 1)$. We then sort the observations to yield $S = \{s_1, \ldots, s_n\}$, and plot $(s_i, i/n)$ for all i.

Cumulative distributions require slightly more sophistication to read than histograms. The cdf is monotonically increasing, so there are no peaks in the distribution. Instead, the mode is marked by the longest vertical line segment. But cdfs are much better at highlighting the tails of a distribution. The reason is clear: the small counts at the tails are obscured by the axis on a histogram, but accumulate into visible stuff in the cdf.

6.3.6 Data Maps

Maps use the spatial arrangement of regions to represent places, concepts, or things. We have all mastered skills for navigating the world through maps, skills which translate into understanding related visualizations.

Traditional data maps use color or shading to highlight properties of the regions in the map. Figure 6.25 (left) colors each state according to whether they voted for Barack Obama (blue) or his opponent Mitt Romney (red) in the 2012 U.S. presidential election. The map makes clear the country's political divisions. The northeast and west coasts are as solidly blue as the Midwest and south are red.

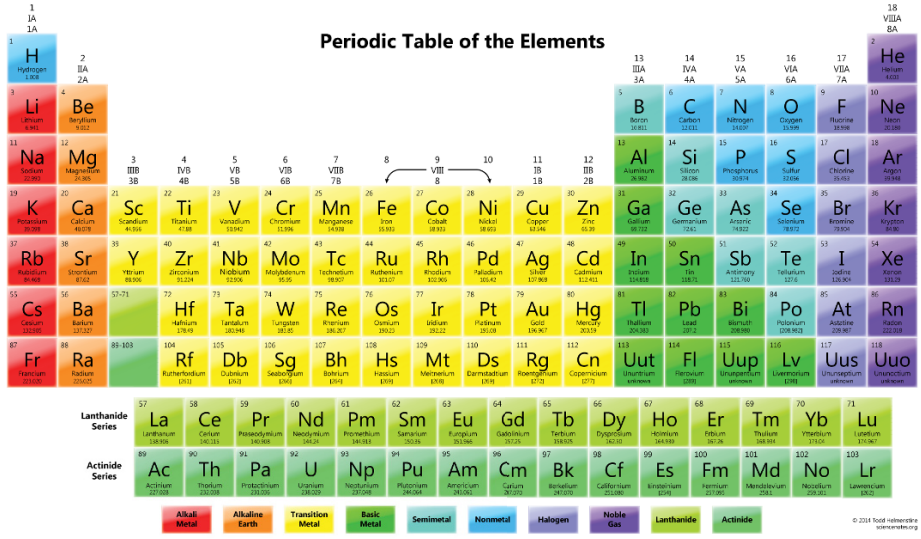

Figure 6.26: The periodic table maps the elements into logical groupings, reflecting chemical properties. *Source*: `http://sciencenotes.org`

Maps are not limited to geographic regions. The most powerful map in the history of scientific visualization is chemistry's periodic table of the elements. Connected regions show where the metals and Nobel gases reside, as well as spots where undiscovered elements had to lie. The periodic table is a map with enough detail to be repeatedly referenced by working chemists, yet is easily understood by school children.

What gives the periodic table such power as a visualization?

- *The map has a story to tell*: Maps are valuable when they encode information worth referencing or assimilating. The periodic table is the *right* visualization of the elements because of the structure of electron shells, and their importance on chemical properties and bonding. The map bears repeated scrutiny because it has important things to tell us.

  Data maps are fascinating, because breaking variables down by region so often leads to interesting stories. Regions on maps generally reflect cultural, historical, economic, and linguistic continuities, so phenomena deriving from any of these factors generally show themselves clearly on data maps.

- *Regions are contiguous, and adjacency means something*: The continuity of regions in Figure 6.26 is reflected by its color scheme, grouping elements sharing similar properties, like the alkali metals and the Nobel gases. Two elements sitting next to each other usually means that they share something important in common.

- *The squares are big enough to see*: A critical decision in canonizing the periodic table was the placement of the Lanthanide (elements 57–71) and Actinide metals (elements 89–103). They are conventionally presented as the bottom two rows, but logically belong within the body of the table in the unlabeled green squares.

  However, the conventional rendering avoids two problems. To do it "right," these elements would either be compressed into hopelessly thin slivers, or else the table would need to become twice as wide to accommodate them.

- *It is not too faithful to reality*: Improving reality for the sake of better maps is a long and honorable tradition. Recall that the Mercator projection distorts the size of landmasses near the poles (yes, Greenland, I am looking at you) in order to preserve their shape.

  Cartograms are maps distorted such that regions reflect some underlying variable, like population. Figure 6.25 (right) plots the electoral results of 2012 on a map where the area of each state is proportional to its population/electoral votes. Only now does the magnitude of Obama's victory become clear: those giant red Midwestern states shrink down to an appropriate size, yielding a map of more blue than red.

6.4 Great Visualizations

Developing your own visualization aesthetic gives you a language to talk about what you like and what you don't like. I now encourage you to apply your judgment to assess the merits and demerits of certain charts and graphs.

In this section, we will look at a select group of classic visualizations which I consider great. There are a large number of terrible graphics I would love to contrast them against, but I was taught that it is not nice to make fun of people.

This is especially true when there are copyright restrictions over the use of such images in a book like this. However, I strongly encourage you to visit `http://wtfviz.net/` to see a collection of startling charts and graphics. Many are quite amusing, for reasons you should now be able to articulate using the ideas in this chapter.

6.4.1 Marey's Train Schedule

Tufte points to E.J. Marey's railroad schedule as a landmark in graphical design. It is shown in Figure 6.27. The hours of the day are represented on the x-axis. Indeed, this rectangular plot is really a cylinder cut at 6AM and laid down flat. The y-axis represents all the stations on the Paris to Lyon line. Each line represents the route of a particular train, reporting where it should be at each moment in time.

Normal train schedules are tables, with a column for each train, a row for each station, and entry (i, j) reporting the time train j arrives at station i. Such tables are useful to tell us what time to arrive to catch our train. But Marey's

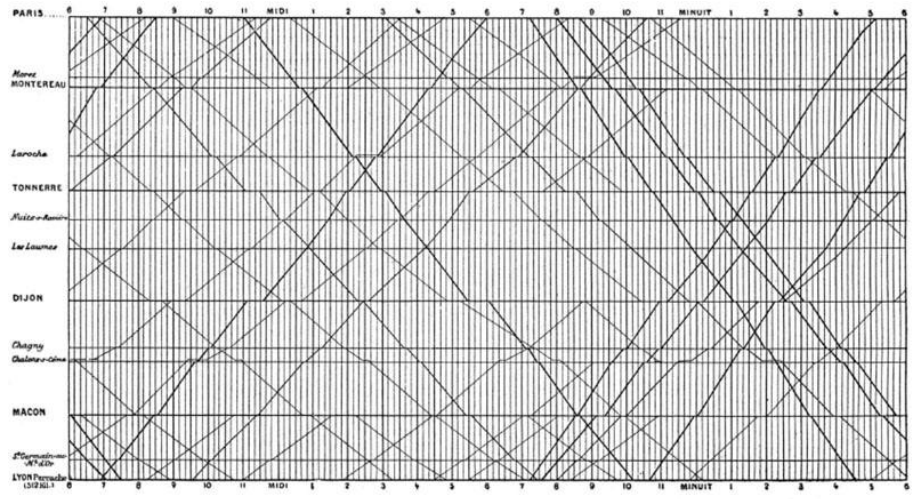

Figure 6.27: Marey's train schedule plots the position of each train as a function of time.

design provides much more information. What else can you see here that you cannot with conventional time tables?

- *How fast is the train moving?* The slope of a line measures how steep it is. The faster the train, the greater the absolute slope. Slower trains are marked by flatter lines, because they take more time to cover the given ground.

  A special case here is identifying periods when the train is idling in the station. At these times, the line is horizontal, indicating no movement down the line.

- *When do trains pass each other?* The direction of a train is given by the slope of the associated line. Northbound trains have a positive slope, and southbound trains a negative slope. Two trains pass each other at the intersection points of two lines, letting passengers know when to look out the window and wave.

- *When is rush hour?* There is a concentration of trains leaving both Paris and Lyon around 7PM, which tells me that must have been the most popular time to travel. The trip typically took around eleven hours, so this must have been a sleeper train, where travelers would arrive at their destination bright and early the next day.

The departure times for trains at a station are also there, of course. Each station is marked by a horizontal line, so look for the time when trains cross your station in the proper direction.

Figure 6.28: Cholera deaths center around a pump on Broad street, revealing the source of the epidemic.

My only quibble here is that it would have been even better with a lighter data grid. Never imprison your data!

6.4.2 Snow's Cholera Map

A particularly famous data map changed the course of medical history. Cholera was a terrible disease which killed large numbers of people in 19th-century cities. The plague would come suddenly and strike people dead, with the cause a mystery to the science of the day.

John Snow plotted cholera cases from an epidemic of 1854 on a street map of London, hoping to see a pattern. Each dot in Figure 6.28 represented a household struck with the disease. What do you see?

Snow noticed a cluster of the cases centered on Broad Street. Further, at the center of the cluster was a cross, denoting a well where the residents got their drinking water. The source of the epidemic was traced to the handle of a single water pump. They changed the handle, and suddenly people stopped getting sick. This proved that cholera was an infectious disease caused by contaminated water, and pointed the way towards preventing it.

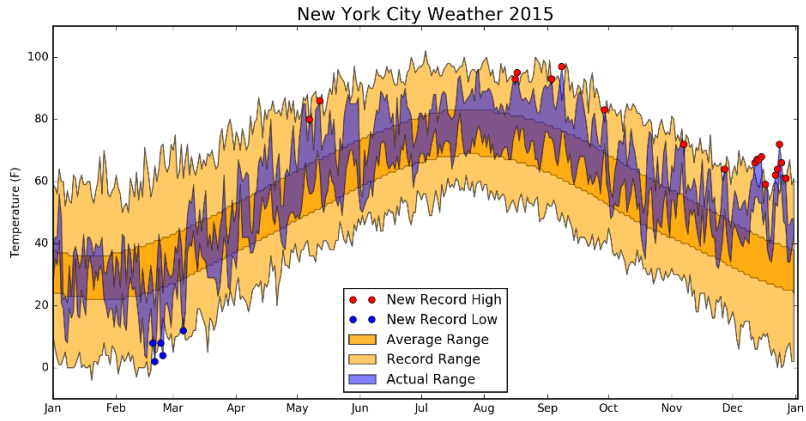

Figure 6.29: This weather year in review displays a clear story in over 2000 numbers.

6.4.3 New York's Weather Year

It is almost worth enduring a New York winter to see a particular graphic which appears in *The New York Times* each January, summarizing the weather of the previous year. Figure 6.29 presents an independent rendition of the same data, which captures why this chart is exciting. For every day of the year, we see the high and low temperature plotted on a graph, along with historical data to put it in context: the average daily high and low, as well as the highest/lowest temperatures ever recorded for that date.

What is so great about that? First, it shows $6 \times 365 = 2190$ numbers in a coherent manner, which facilitates comparisons on the sine curve of the seasons:

- We can tell when there were hot and cold spells, and how long they lasted.
- We can tell what days had big swings in temperature, and when the thermometer hardly moved at all.
- We can tell whether the weather was unusual this year. Was it an unusually hot or cold year, or both? When were record high/lows set, and when did they get close to setting them?

This single graphic is rich, clear, and informative. Be inspired by it.

6.5 Reading Graphs

What you see isn't always what you get. I have seen many graphs brought to me by my students over the years. Some have been amazing, and most others good enough to get the job done.

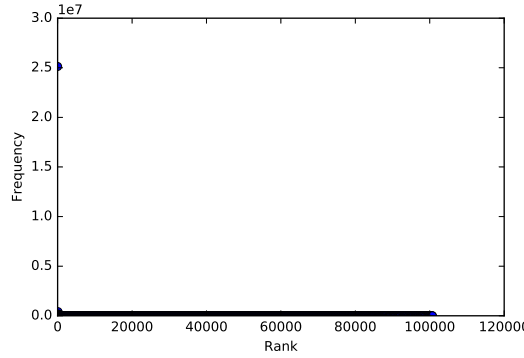

Figure 6.30: A dotplot of word frequency by rank. What do you see?

But I also repeatedly see plots with the same basic problems. In this section, for a few of my pet peeves, I present the original plot along with the way to remedy the problem. With experience, you should be able to identify these kinds of problems just by looking at the initial plot.

6.5.1 The Obscured Distribution

Figure 6.30 portrays the frequency of 10,000 English words, sorted by frequency. It doesn't look very exciting: all you can see is a single point at $(1, 2.5)$. What happened, and how can you fix it?

If you stare at the figure long enough, you will see there are actually a lot more points. But they all sit on the line $y = 0$, and overlap each other to the extent that they form an undifferentiated mass.

The alert reader will realize that this single point is in fact an outlier, with magnitude so large as to shrink all other totals toward zero. A natural reaction would delete the biggest point, but curiously the remaining points will look much the same.

The problem is that this distribution is a power law, and plotting a power law on a linear scale shows nothing. The key here is to plot it on log scale, as in Figure 6.31 (left). Now you can see the points, and the mapping from ranks to frequency. Even better is plotting it on a log-log scale, as in Figure 6.31 (right). The straight line here confirms that we are dealing with a power law.

6.5.2 Overinterpreting Variance

In bioinformatics, one seeks to discover how life works by looking at data. Figure 6.32 (left) presents a graph of the folding energy of genes as a function of their length. Look at this graph, and see if you can make a discovery.

It is pretty clear that something is going on in there. For gene lengths above 1500, the plot starts jumping around, producing some very negative values. Did

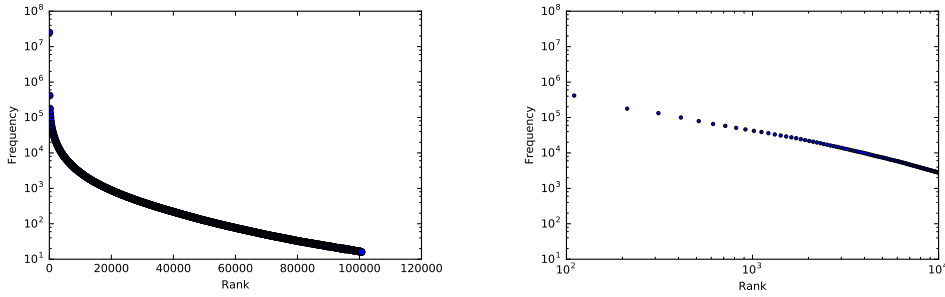

Figure 6.31: Frequency of 10,000 English words. Plotting the word frequencies on a log scale (left), or even better on a log-log scale reveals that it is power law (right).

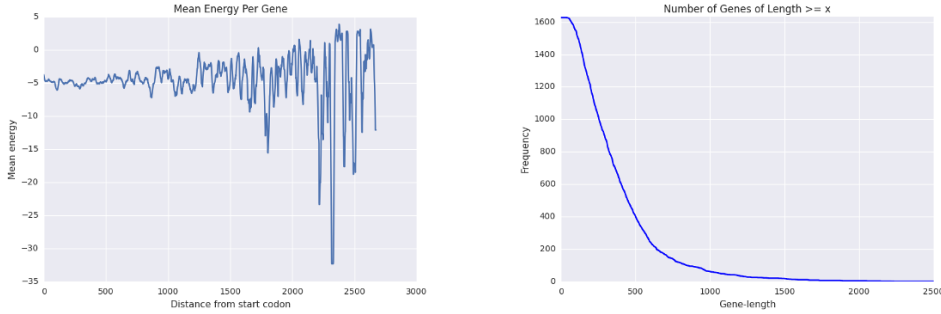

Figure 6.32: Folding energy of genes as a function of their length. Mistaking variance for signal: the extreme values in the left figure are artifacts from averaging small numbers of samples.

we just discover that energy varies inversely with gene length?

No. We just overinterpreted variance. The first clue is that what looked very steady starts to go haywire as the length increases. Most genes are very short in length. Thus the points on the right side of the left plot are based on very little data. The average of a few points is not as robust as the average of many points. Indeed, a frequency plot of the number of genes by length (on right) shows that the counts drop off to nothing right where it starts jumping.

How could we fix it? The right thing to do here is to threshold the plot by only showing the values with enough data support, perhaps at length 500 or a bit more. Beyond that, we might bin them by length and take the average, in order to prove that the jumping effect goes away.

6.6 Interactive Visualization

The charts and graphs we have discussed so far were all static images, designed to be studied by the viewer, but not manipulated. Interactive visualization techniques are becoming increasingly important for exploratory data analysis.

Mobile apps, notebooks, and webpages with interactive visualization widgets can be especially effective in presenting data, and disseminating it to larger audiences to explore. Providing viewers with the power to play with the actual data helps ensure that the story presented by a graphic is the true and complete story.

If you are going to view your data online, it makes sense to do so using interactive widgets. These are generally extensions of the basic plots that we have described in this chapter, with features like offering pop-ups with more information when the user scrolls over points, or encouraging the user to change the scale ranges with sliders.

There are few potential downsides to interactive visualization. First, it is harder to communicate exactly what you are seeing to others, compared with static images. They might not be seeing exactly the same thing as you. Screenshots of interactive systems generally cannot compare to publication-quality graphics optimized on traditional systems. The problem with what you see is what you get (WYSIWYG) systems is that, generally, what you see is all you get. Interactive systems are best for exploration, not presentation.

There are also excesses which tend to arise in interactive visualization. Knobs and features get added because they can, but the visual effects they add can distract rather than add to the message. Rotating three-dimensional point clouds always looks cool, but I find them hard to interpret and very seldom find such views insightful.

Making data tell a story requires some insight into how to tell stories. Films and television represent the state-of-the-art in narrative presentation. Like them, the best interactive presentations show a narrative, such as moving through time or rifling through alternative hypotheses. For inspiration, I encourage you to watch the late Han Rosling's TED talk[1] using animated bubble

[1] `https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen`

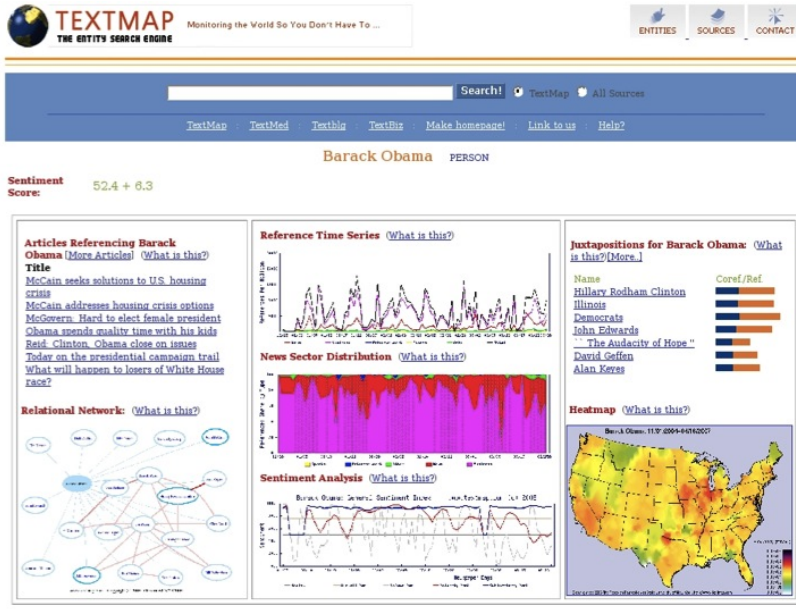

Figure 6.33: The TextMap dashboard for *Barack Obama*, circa 2008.

charts to present the history of social and economic development for all the world's nations.

Recent trends in cloud-based visualization encourage you to upload your data to a site like Google Charts `https://developers.google.com/chart/`, so you can take advantage of interactive visualization tools and widgets that they provide. These tools produce very nice interactive plots, and are easy to use.

The possible sticking point is security, since you are giving your data to a third party. I hope and trust that the CIA analysts have access to their own in-house solutions which keep their data confidential. But feel free to experiment with these tools in less sensitive situations.

6.7 War Story: TextMapping the World

My biggest experience in data visualization came as a result of our large scale news/sentiment analysis system. TextMap presented a news analysis dashboard, for every entity whose name appeared in news articles of the time. Barack Obama's page is shown in Figure 6.33. Our dashboard was made up of a variety of subcomponents:

- *Reference time series*: Here we reported how often the given entity appeared in the news, as a function of time. Spikes corresponded to more important news events. Further, we partitioned these counts according

to appearances in each section of the newspaper: news, sports, entertainment, or business.

- *News sector distribution*: This graph contains *exactly* the same data as in the reference time series, but presented as a stacked area plot to show the fraction of entity references in each section of the paper. Obama clearly presents as a news figure. But other people exhibited interesting transitions, such as Arnold Schwarzenegger when he shifted from acting to politics.

- *Sentiment analysis*: Here we present a time series measuring the sentiment of the entity, by presenting the normalized difference of the number of positive news mentions to total references. Thus zero represented a neutral reputation, and we provided a central reference line to put the placement into perspective. Here Obama's sentiment ebbs and flows with events, but generally stays on the right side of the line.

  It was nice to see bad things happen to bad people, by watching their news sentiment drop. As I recall, the lowest news sentiment ever achieved was by a mom who achieved notoriety by cyberbullying one of her daughter's social rivals into committing suicide.

- *Heatmap*: Here we presented a data map showing the relative reference frequency of the entity. Obama's great red patch around Illinois is because he was serving as a senator from there at the time he first ran for president. Many classes entities showed strong regional biases, like sports figures and politicians, but less so entertainment figures like movie stars and singers.

- *Juxtapositions and relational network*: Our system built a network on news entities, by linking two entities whenever they were mentioned in the same article. The strength of this relationship could be measured by the number of articles linking them together. The strongest of these associations are reported as juxtapositions, with their frequency and strength shown by a bar chart on a logarithmic scale.

  We have not really talked so far about network visualization, but the key idea is to position vertices so neighbors are located near each other, meaning the edges are short. Stronger friendships are shown by thicker lines.

- *Related articles*: We provided links to representative news articles mentioning the given entity.

One deficiency of our dashboard was that it was not interactive. In fact, it was anti-interactive: the only animation was a gif of the world spinning forlornly in the upper-left corner. Many of the plots we rendered required large amounts of data access from our clunky database, particularly the heatmap. Since it could not be rendered interactively, we precomputed these maps offline and showed them on demand.

After Mikhail updated our infrastructure (see the war story of Section 12.2), it became possible for us to support some interactive plots, particularly time series. We developed a new user interface called TextMap Access that let users play with our data.

But when General Sentiment licensed the technology, the first thing it disposed of was our user interface. It didn't make sense to the business analysts who were our customers. It was too complicated. There is a substantial difference between surface appeal and the real effectiveness of an interface for users. Look carefully at our TextMap dashboard: it had "What is this?" buttons in several locations. This was a sign of weakness: if our graphics were really intuitive enough we would not have needed them.

Although I grumbled about the new interface, I was undoubtedly wrong. General Sentiment employed a bunch of analysts who used our system all day long, and were available to talk to our developers. Presumably the interface evolved to serve them better. The best interfaces are built by a dialog between developers and users.

What are the take-home lessons from this experience? There is a lot I still find cool about this dashboard: the presentation was rich enough to really expose interesting things about how the world worked. But not everyone agreed. Different customers preferred different interfaces, because they did different things with them.

One lesson here is the power of providing alternate views of the same data. The reference time series and news sector distribution were exactly the same data, but provided quite different insights when presented in different ways. All songs are made from the same notes, but how you arrange them makes for different kinds of music.

6.8 Chapter Notes

I strongly recommend Edward Tufte's books [Tuf83, Tuf90, Tuf97] to anyone interested in the art of scientific visualization. You don't even have to read them. Just look at the pictures for effective contrasts between good and bad graphics, and plots that really convey stories in data.

Similarly good books on data visualization are Few [Few09] and far between. Interesting blogs about data visualization are `http://flowingdata.com` and `http://viz.wtf`. The first focuses on great visualizations, the second seeking out the disasters. The story of Snow's cholera map is reported in Johnson [Joh07].

The chartjunk removal example from Section 6.2.3 was inspired by an example by Tim Bray at `http://www.tbray.org`. Anscombe's quartet was first presented in [Ans73]. The basic architecture of our Lydia/TextMap news analysis system is reported in Lloyd et al. [LKS05], with additional papers describing heatmaps [MBL+06] and the Access user interface [BWPS10].

6.9 Exercises

Exploratory Data Analysis

6-1. *[5]* Provide answers to the questions associated with the following data sets, available at `http://www.data-manual.com/data`.

(a) Analyze the *movie* data set. What is the range of movie gross in the United States? Which type of movies are most likely to succeed in the market? Comedy? PG-13? Drama?

(b) Analyze the *Manhattan rolling sales* data set. Where in Manhattan is the most/least expensive real estate located? What is the relationship between sales price and gross square feet?

(c) Analyze the *2012 Olympic* data set. What can you say about the relationship between a country's population and the number of medals it wins? What can you say about the relationship between the ratio of female and male counts and the GDP of that country?

(d) Analyze the *GDP per capita* data set. How do countries from Europe, Asia, and Africa compare in the rates of growth in GDP? When have countries faced substantial changes in GDP, and what historical events were likely most responsible for it?

6-2. *[3]* For one or more of the data sets from `http://www.data-manual.com/data`, answer the following basic questions:

(a) Who constructed it, when, and why?

(b) How big is it?

(c) What do the fields mean?

(d) Identify a few familiar or interpretable records.

(e) Provide Tukey's five number summary for each column.

(f) Construct a pairwise correlation matrix for each pair of columns.

(g) Construct a pairwise distribution plot for each interesting pair of columns.

Interpreting Visualizations

6-3. *[5]* Search your favorite news websites until you find ten interesting charts/plots, ideally half good and half bad. For each, please critique along the following dimensions, using the vocabulary we have developed in this chapter:

(a) Does it do a good job or a bad job of presenting the data?

(b) Does the presentation appear to be biased, either deliberately or accidentally?

(c) Is there chartjunk in the figure?

(d) Are the axes labeled in a clear and informative way?

(e) Is the color used effectively?

(f) How can we make the graphic better?

6-4. *[3]* Visit `http://www.wtfviz.net`. Find five laughably bad visualizations, and explain why they are both bad and amusing.

Creating Visualizations

6-5. *[5]* Construct a revealing visualization of some aspect of your favorite data set, using:

(a) A well-designed table.

(b) A dot and/or line plot.

(c) A scatter plot.

(d) A heatmap.

(e) A bar plot or pie chart.

(f) A histogram.

(g) A data map.

6-6. *[5]* Create ten different versions of line charts for a particular set of (x, y) points. Which ones are best and which ones worst? Explain why.

6-7. *[3]* Construct scatter plots for sets of 10, 100, 1000, and 10,000 points. Experiment with the point size to find the most revealing value for each data set.

6-8. *[5]* Experiment with different color scales to construct scatter plots for a particular set of (x, y, z) points, where color is used to represent the z dimension. Which color schemes work best? Which are the worst? Explain why.

Implementation Projects

6-9. *[5]* Build an interactive exploration widget for your favorite data set, using appropriate libraries and tools. Start simple, but be as creative as you want to be.

6-10. *[5]* Create a data video/movie, by recording/filming an interactive data exploration. It should not be long, but how interesting/revealing can you make it?

Interview Questions

6-11. *[3]* Describe some good practices in data visualization?

6-12. *[5]* Explain Tufte's concept of chart junk.

6-13. *[8]* How would you to determine whether the statistics published in an article are either wrong or presented to support a biased view?

Kaggle Challenges

6-14. Analyze data from San Francisco Bay Area bike sharing.
`https://www.kaggle.com/benhamner/sf-bay-area-bike-share`

6-15. Predict whether West Nile virus is present in a given time and place.
`https://www.kaggle.com/c/predict-west-nile-virus`

6-16. What type of crime is most likely at a given time and place?
`https://www.kaggle.com/c/sf-crime`