Alternative chart types are useful when you have too many data series to track in a single graph. Try sparklines, a small multiples approach, cycle charts, or horizon charts when you have a lot of data to visualize. For some of these approaches, enabling the reader to discern exact values is less important than showing them the overall trend or pattern.

Other graph types, like flow charts and timelines, have infinite varieties and styles. Horizontal layouts may work for some people, content, and platforms, while vertical layouts may be better online where it matches the natural scrolling motion. Compact layouts are best for mobile platforms.

Whichever graph you use to plot your data, consider how much detail your reader needs and how you can guide them to the point you wish to convey. Many of these chart types are well-known and understood, so our challenge is to make them engaging and interesting without sacrificing accuracy.

DISTRIBUTION

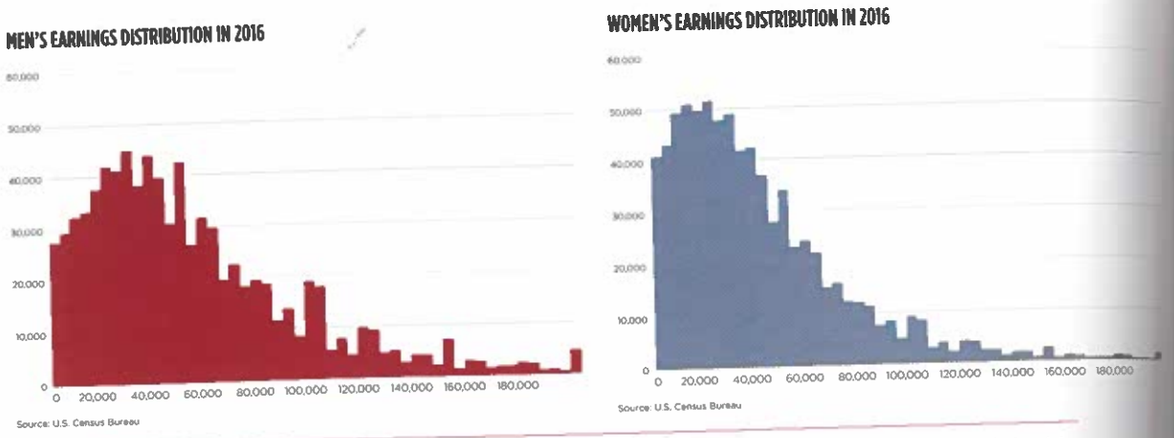
This chapter covers visualizations of data distributions and statistical uncertainties. These may be inherently difficult for many readers because they may not be as familiar with the statistical terminology or the graphs themselves, which may look quite different from the standard graphs they are used to seeing.

Charts like the fan chart and the box-and-whisker plot show statistical measures like confidence intervals and percentiles. Violin plots, which depict entire distributions, may look so foreign that your reader will need detailed explanations to understand them. This doesn't mean that these charts are inherently *bad* at visualizing data—proper labeling and design can make even the most esoteric box-and-whisker plot interesting—but the hurdle of statistical literacy may make such graphs difficult for many readers.

Graphs in this chapter follow the guidelines published by the *Dallas Morning News* in 2005. The *News*'s guidelines include instructions for specific fonts and colors, as well as ways to design and style different graphs, tables, maps, icons, and a summary of the newsroom workflow. The guide uses two fonts, Gotham and Miller Deck, depending on the size and purpose; I use the Montserrat font, which is similar to Gotham.

HISTOGRAM

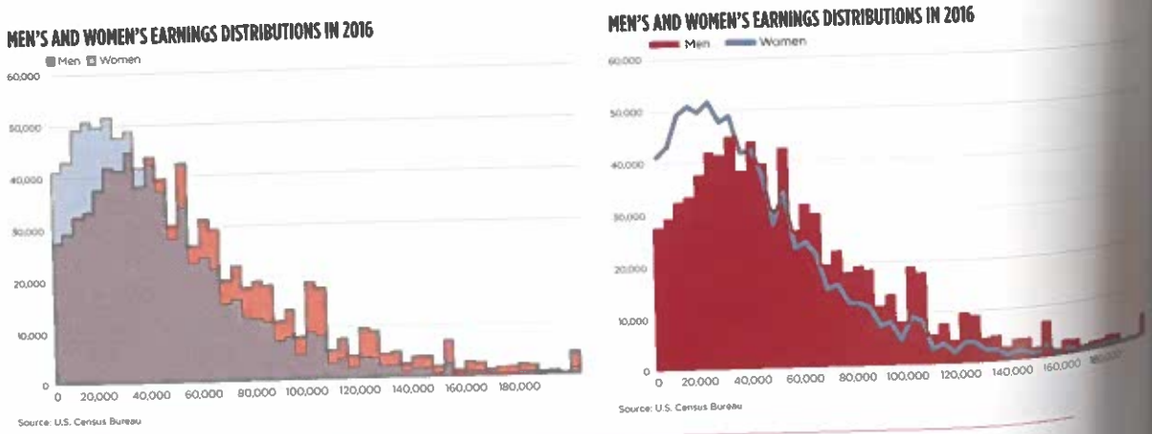
The histogram is the most basic graph type for visualizing a distribution. It is a specific kind of bar chart that presents the *tabulated frequency* of data over distinct intervals, called *bins*, that sum to the



Histograms divide the entire sample into intervals (also called "bins"). The height of the bin shows the number of observations within it.

total distribution. The entire sample is divided into these bins, and the height of each bar shows the number of observations within each interval. Histograms can show where values are concentrated within a distribution, where extreme values are, and whether there are any gaps or unusual values.

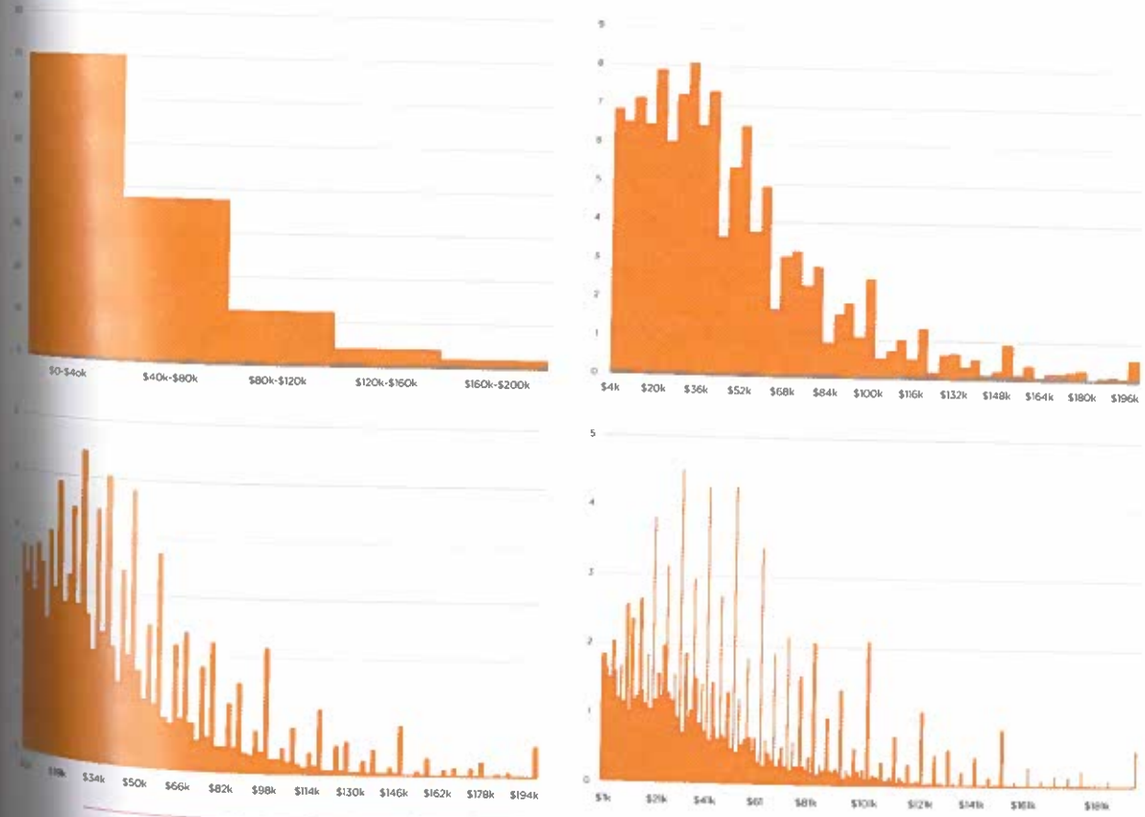
We can layer histograms together to show how different distributions compare. The two histograms above depict the distribution of earnings for men and women working in the United States in 2016. We can make some general comparisons between them, but that task is made easier in the next two graphs, where the distributions are placed on top of each other.



Histograms can be layered together by using saturated colors (left) or different encodings like bars and lines (right).

The graph on the left uses two column charts and transparent colors so both are visible. The graph on the right combines a column chart and line chart, which has the advantage of not using transparent colors, but the balance of how the two groups are presented is now unequal. You might also notice that the line intersects the *middle* of each bin as opposed to spanning the entire bin as the columns do—it's a minor difference but one that you may want to keep in mind.

A key consideration in creating a histogram is how wide to set the bins. Bins that are too wide may hide patterns in the distribution, while bins that are too narrow may obscure the overall shape of the distribution. While there is no "correct" number of bins, there are a number of statistical tests (using, for example, square roots, logarithmics, or cube roots) that



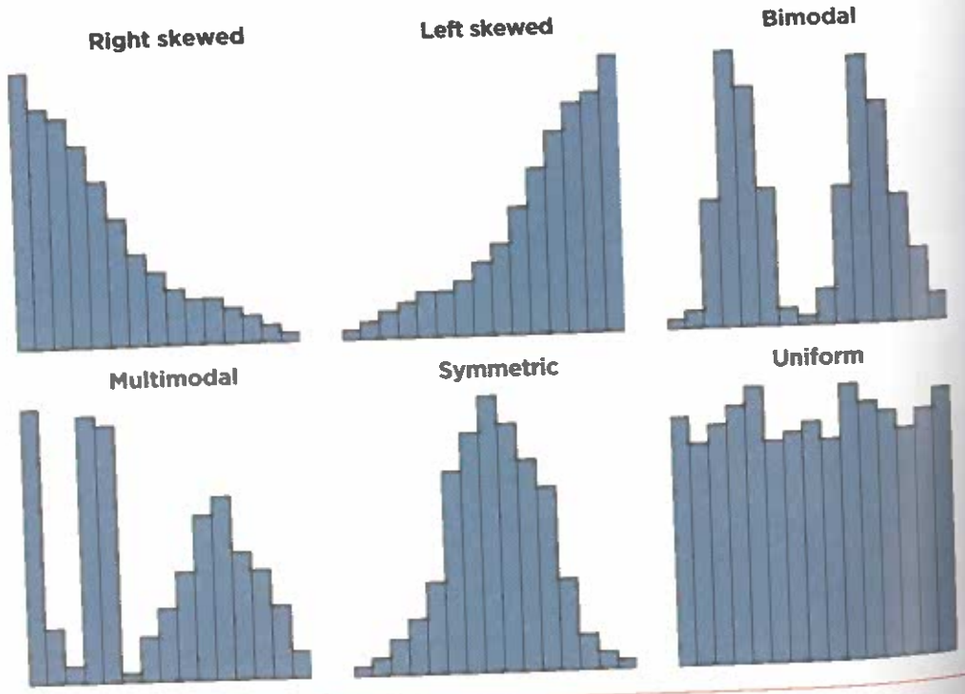
A data distribution will look different depending on the number bins, as it does here as the number of bins increases from 5 to 30 to 50 to 120.

can help determine the optimal bin width. In this example, notice how the distribution looks different as the number of bins increases from 5 to 30 to 50 to 120.

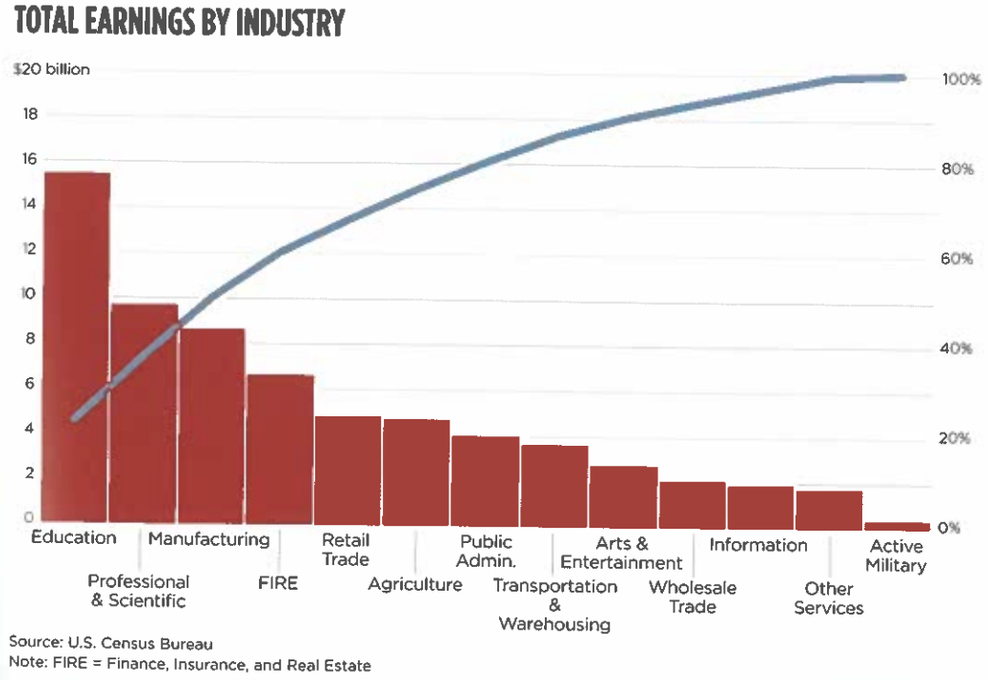
Histograms can help us understand whether our data *lean* to one side or another. A distribution in which more data are pushed to the left is known as *right-skewed*. A histogram with more observations to the right is called *left-skewed*. Distributions that have two peaks are called *bimodal* and distributions with multiple peaks are called *multimodal*. *Symmetric* distributions are those with a roughly equal number of observations on either side of a central value and *uniform* distributions in which the observations are roughly equally distributed.

When we understand the distribution of our data and its possible skew, we're better prepared to conduct more accurate statistical tests. Two completely different distributions may have the same mean and median, but if we don't understand the spread and structure of our data, our results may not paint a complete picture. This is where visualizing our data can be invaluable.

A modification to the basic histogram is the Pareto chart, named after the Italian engineer and economist Vilfredo Pareto. The Pareto chart (next page) consists of bars that represent



Histograms can help us understand the shape of the distribution of our data. Here we see six such forms of distribution.



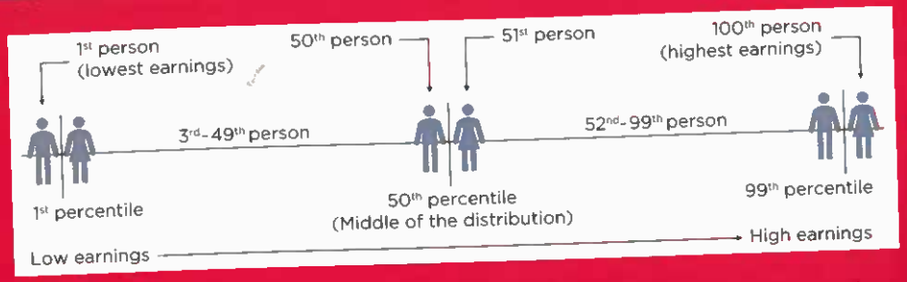
The Pareto chart shows values for each group (usually in bars) and the cumulative total as a line.

individual data points and a line that represents the cumulative total. The Pareto chart may be the exception to the rule against dual-axis charts because its *purpose* is to show the complementary distributions overlaid on each other. Of course, the two metrics are really not two different measures—it's the same metric, one as a marginal distribution (separate values of each group) and one as a cumulative distribution (where the values sum to a total).

This Pareto chart shows total earnings in thirteen different major industries in the United States—the bars show the total earnings in each industry and the line shows how the shares add up to total earnings in the economy.

UNDERSTANDING PERCENTILES

Imagine one hundred people standing on an auditorium stage. You stand in the audience as they line up from your left to right, arranged by their earnings. The person with the lowest earnings stands at the left side of the stage, and the person with



the highest earnings stands at the right. Together, they represent the entire earnings distribution.

The first person in the line has the lowest earnings. At their position, 99 percent of people—all those to their left—have higher earnings. They are said to be in the 1st percentile. Similarly, there is a worker on the other side of the stage in the 100th position. To their right stands 99 other people—99 percent of all people on the stage with lower earnings. They are in the 99th percentile of the distribution, in the top 1 percent. Finally, there is a point in the middle of the stage that splits everyone into two equal groups, 50 percent of the distribution on either side. That point (or more precisely, the *earnings* at that point) represents the 50th percentile or the median of the distribution.

The mathematics of increasing the number of people on stage from 100 to 200 to 1,000 to 150 million does not change—the middle of the ordered distribution is still the median and the person standing at the 10 percent position is at the 10th percentile. Because percentiles are independent of the population, you can compare them across any group such as country or industry.

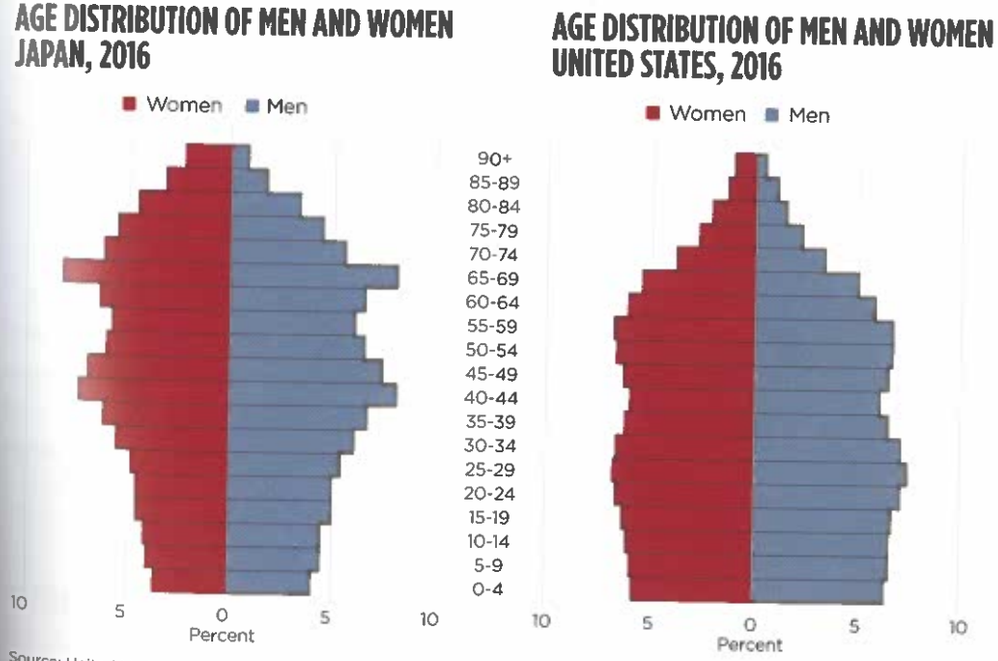
While percentiles identify a specific location in the distribution, there are other metrics that will give you an overall measure of the distribution. The *mean* or *average* is equal to the sum of all values divided by the number of observations. Because we add up all the data, large values can generate a distorted picture of the true distribution. In the example above, the mean would change dramatically if we replaced one of the people on stage with someone who earned $100 million, but—take note—the median would *not* change because that person would still stand on the far right edge of the stage and the rest of the people in line would stay in the same position.

The *variance* is another metric of a distribution and measures how far each observation in a data set is spread out from the mean value. A large variance indicates the values in the data are far from the mean and from each other; a small variance, by contrast, indicates the opposite. A full decomposition of the variance and related formulas are beyond the scope of this book, but if you are working with data and creating data visualizations, it is certainly worth a bit of time and study to understand how they can be used to better understand the shape and scope of your data.

PYRAMID CHART

Most often used to show changes in population-based metrics such as birth rates, mortality rates, age, or overall population levels, pyramid charts put two groups on either side of a center vertical axis. The pyramid chart is a subcategory of the diverging bar chart (see page 92), but the name is reserved for comparing distributions, most often ages. As with the diverging bar chart, the layout may cause some confusion, because your reader may assume the leftward bars represent negative values, and the rightward bars represent positive values.

The advantage of the pyramid chart is that we can assess the overall shape of the distribution because both groups sit on the same vertical baseline. While many pyramid charts use different colors for the two groups, that's not a necessary characteristic.



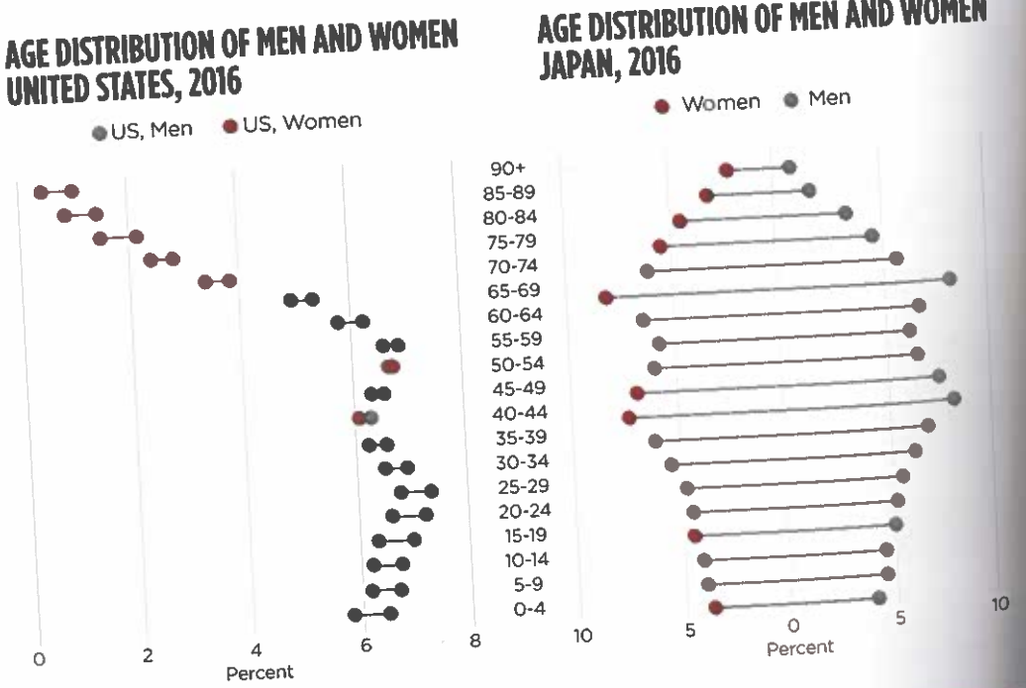
Source: United Nations

Pyramid charts are a type of diverging bar chart, typically used to show population-based metrics like birth rates, mortality rates, age, or overall population levels.

These pyramid charts show the distribution of ages in the United States and Japan in 2016. In both graphs, women are represented on the left branch of the vertical axis and men on the right. Each row represents a different age group: 0–4 years old, 5–9 years old, and so forth. The shape of the chart means we can immediately see that Japan has a greater share of older people, while there is a larger share of younger people in the United States.

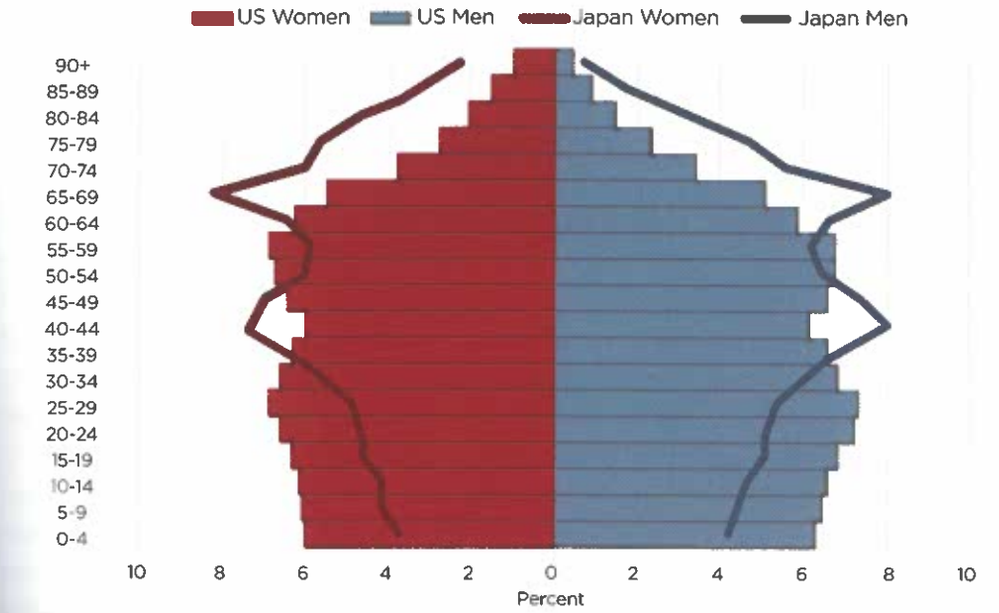
Because the bars are not next to each other, it is difficult to compare the total shares of men and women. But again, whether that's a problem depends on the goal of your visual. If you want your reader to compare the shares of the two genders, then a different chart type—such as a paired bar chart or dot plot—would be a better choice. But if you want your reader to see the overall shape of the distribution, the pyramid chart is perfect.

A natural alternative to the pyramid chart is the dot plot or the lollipop chart. Dots for each gender are positioned along the horizontal axis corresponding to the data value and connected by a line. Or, as shown on the right, we could simply use a lollipop chart (also see page 80), replacing the bars with lines and dots. With either approach, we can still use different colors or just use a single color for the entire graph.



Source: United Nations

Alternatives to the basic pyramid chart are the dot plot or lollipop chart.



Source: United Nations

Combining distributions in one chart can be accomplished by adjust colors and combining different encodings.

One challenge with the pair of pyramid charts is to accurately compare the age distributions across the two countries. We can infer the main relationship, that Japan is, on average, older than the United States, but it's harder to make a more detailed comparison. Placing the charts atop each other—a technique we saw earlier with the histogram—makes that task easier. But be mindful that in using this approach the data for the two countries are shown with different encodings (bars and lines), which can risk emphasizing one set over the other.

VISUALIZING STATISTICAL UNCERTAINTY WITH CHARTS

There are lots of types of uncertainty in data and statistics. Before we learn the different ways to visualize uncertainty, it is worth pausing to understand what we mean by the term. Even if you're not a statistician or mathematician, it's important to understand how such uncertainty and measurement error can affect our results and, ultimately our

Source: Scott Adams

visualizations. Accounting for uncertainty—and making it clear you've done so—builds the reader's trust in your work.

We can think of the term uncertainty in two main ways. One is *uncertainty from randomness*, which applies to the statistical confidence in our statistical models and results. As an example, consider the standard margin of error built into political polling data: "Candidate Smith has a 54 percent approval rating with a margin of error of plus-or-minus 4 percentage points." Another kind is what we might call *uncertainty from unknowns*, where our data are inaccurate, untrusted, imprecise, or even unknown. A very simple example might be something like a data set that includes infants' ages in months instead of weeks. Using statistical and probabilistic models enables us to confront the first kind of uncertainty, which we can therefore visualize; the second kind of uncertainty concerns unknowns that can't necessarily be resolved through more data.

A thorough treatment of *uncertainty from randomness* (error margins, confidence intervals, and the like) are beyond the scope of this book. But *uncertainty from unknowns* is something that many readers can easily relate to. To illustrate, let's consider the data I use in this chapter: worker's earnings by industry and state. The data set used for this analysis is the 2016 U.S. Census Bureau's American Community Survey. The survey includes demographic and economic information for about 3.5 million people per year. For the data in this chapter, I examine individual earnings for more than a million people.

Now, imagine all the reasons why someone might tell the Census Bureau the wrong answer when they ask about their earnings. They might lie. They might round their earnings to the nearest dollar—or nearest hundred dollars, or nearest thousand dollars. Maybe they work side jobs they didn't mention. They may be asked about their spouse's or partner's earnings and have to hazard a guess. There are all sorts of reasons they may get it wrong, and recent economic research

shows that reporting error (especially in government program participation) in some of the largest, most trusted government household surveys has been increasing over the past several years.

We must also recognize that for this survey, the Census Bureau only asks a *share* of Americans (so our calculations from these data also suffer from *uncertainty from randomness*). Consider all the reasons why that "sample" may not be truly representative. Maybe some people don't want to answer the survey. Maybe they moved and didn't get the form, or changed their phone number and didn't get the call.

Whenever we work with data, we should consider how these kinds of uncertainty may lead to some "error" around our final estimates. Not mistakes, but uncertainty. This error is fundamental to being careful with data and ultimately visualizing and explaining our results carefully. Line charts and bar charts suggest certainty with their sharp boundaries and crisp edges, but that certainty is rarely actually the case.

In his book *How Charts Lie*, Alberto Cairo remarks that, "Uncertainty confuses many people because they have the unreasonable expectation that science and statistics will unearth precise truths, when all they can yield is imperfect estimates that can always be subject to changes and updates." We should not expect flawless data, and we should be ready to explain those imperfections to our readers as best we can.

▸ ▸ ▸ ▸ ▸

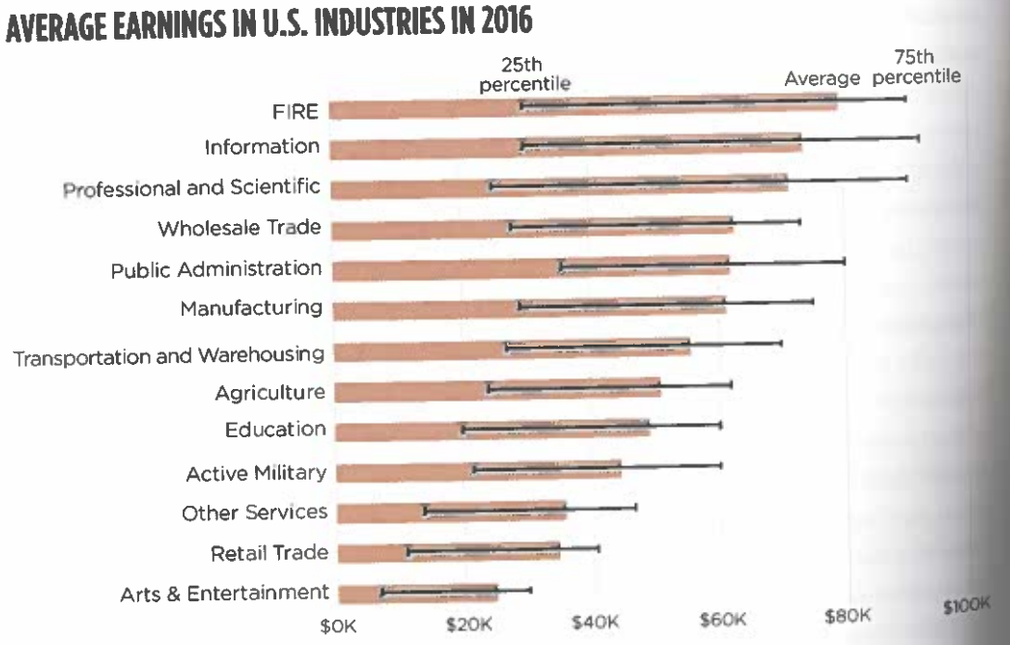
We now move to the specific challenge of conveying uncertainty around data estimates or results from statistical models. This is a common problem: In a survey of ninety data visualization authors and developers, information visualization researcher Jessica Hullman found that graph creators did not include uncertainty in their work for four main reasons. First, they did not want to confuse or overwhelm viewers. Second, they did not have access to information about the uncertainty in their data. Third, they did not know how to calculate uncertainty. And fourth, they did not want to make the data appear questionable. Hullman argues that visualizing uncertainty is important because "a central problem is that authors often omit or downplay information, such that data are interpreted as being more credible than they are." More effectively conveying such uncertainty—especially when making statistical claims—builds trust and credibility.

This section will familiarize you with visual signals of uncertainty. There are many chart types that can be used to show uncertainty around a central estimate, and this section introduces a few of them: error bar charts, confidence interval charts, gradient charts, and fan charts.

ERROR BARS

Perhaps the simplest and most common way to visualize uncertainty is to use error bars: small markers that denote the error margin or confidence interval. Error bars are not really a visualization on their own, but are an addition to other charts, often bar or line charts. The ends of the error bars can correspond to any value you choose: percentiles, the standard error, the 95-percent confidence interval, or even a fixed number. And because error bars can convey these multiple statistical measures, recent research has shown that this can invite confusion on the part of the reader, making incorrect conclusions about the data. We must therefore clearly label the intervals, either in a chart note or, preferably, on the chart itself.

This bar chart shows average earnings in each of thirteen industries in 2016. Error bars denote the 25th and 75th percentiles.



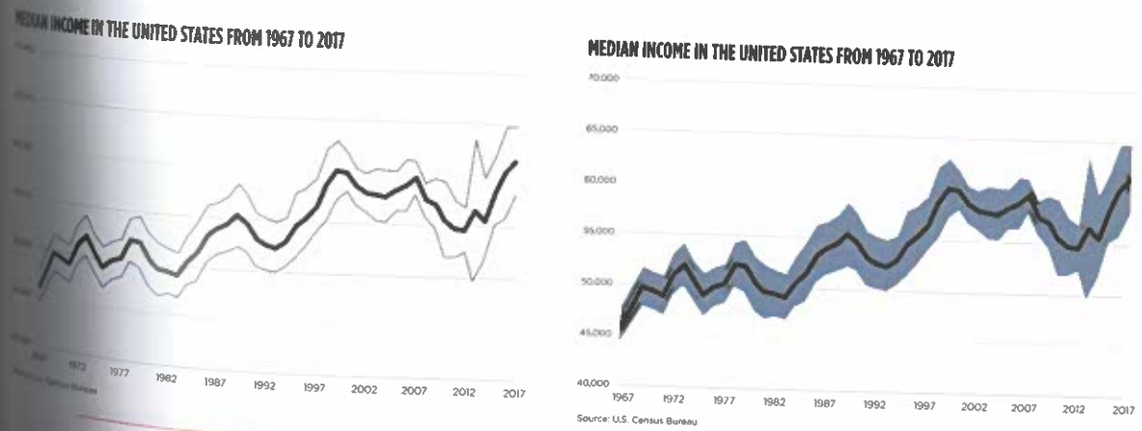
The simplest and most common way to visualize uncertainty or distributions is to use error bars, small markers that denote the margin of error or confidence interval.

Applying error bars to bar charts raises a potentially interesting complication: some research suggests that we tend to judge the points that fall *within* the bar as more likely than those *outside* the bar ("within-the-bar" bias). In the previous chart, that would mean a reader is more likely to assume that the salary of a worker in the Finance, Insurance, and Real Estate (FIRE) sector is less than $80,000 and not more than $80,000. Other research has found that we can better judge uncertainty and the distribution with other types of graphs, such as the violin plot, stripe plot, or gradient plot.

While it may be a familiar approach for many readers and requires less data than some of these other charts, existing research shows that we are not particularly good at assessing uncertainty through these kinds of visual approaches.

CONFIDENCE INTERVAL

A confidence interval chart typically uses lines or shaded areas to depict ranges or amounts of uncertainty, often over time. The basic confidence interval chart is literally a line chart with three lines: one for the central estimate, one for the upper confidence interval value, and another for the lower confidence interval value (these upper and lower lines can be confidence intervals, standard errors, or a fixed number). The lines may be solid, dashed, or colored, but if the central estimates are the primary numbers of interest, that line should be thicker or darker to highlight it against the confidence interval values.



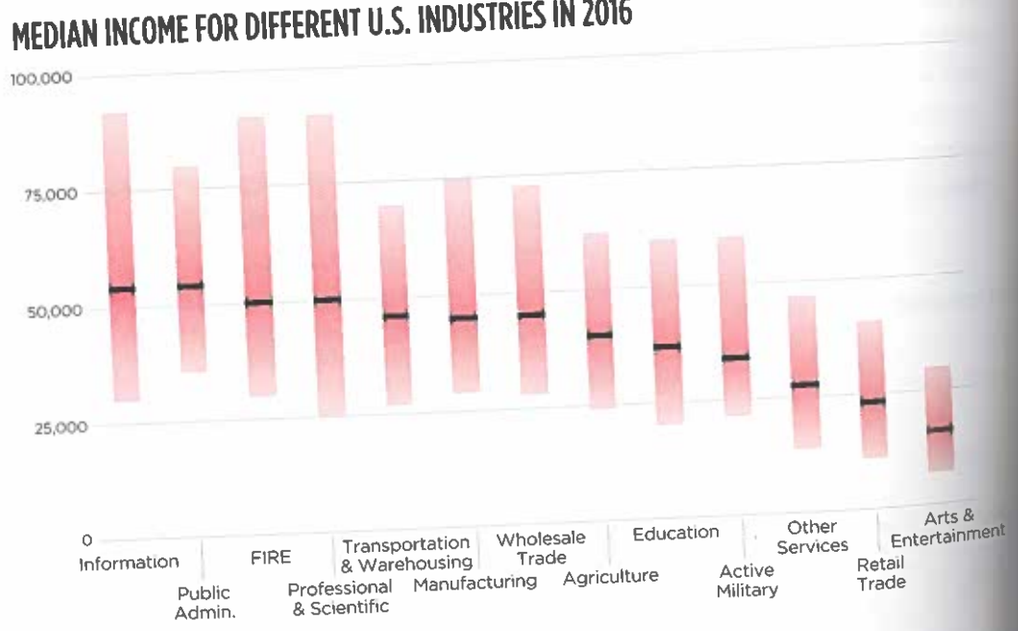
Lines or shaded areas around a central line can visualize ranges of uncertainty.

The two charts on the previous page show median earnings in the United States between 1967 and 2017. The standard error around those estimates is depicted by two separate lines in the left graph and by a shaded area in the right.

GRADIENT CHART

A gradient chart (sometimes called a stripe plot) shows distributions or differences in uncertainty. There are many ways to use the gradient plot, but the basic technique is to plot the primary number of importance and add a color gradient on one or both sides to visually demonstrate the measure of uncertainty around that single point. The plot is named not necessarily after the *shape* of the graph but after the use of the color gradient.

The gradient plot can show changes over time or, as in the graph here, the distribution around individual observations. This gradient plot shows the exact same data as the error

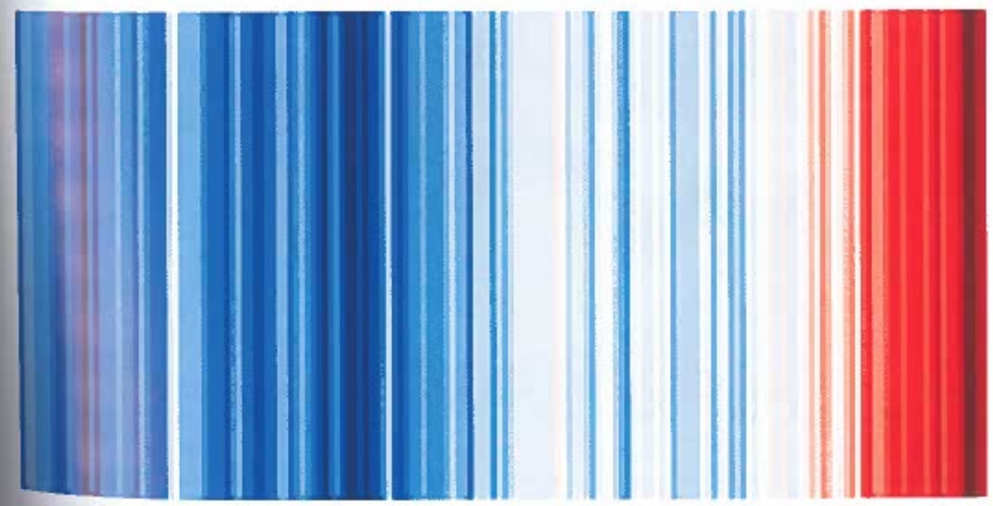


Gradient charts use a color gradient on one or both sides of the primary number of interest to show distributions or uncertainty.

bar chart above, but instead of bars with error lines jutting out in both directions, average earnings are encoded with the dark horizontal line and the 25th to 75th percentiles of the distribution are shown with the gradient. The color gradient might illustrate, for example, multiples of the standard error, which can be a signal to the reader that the outcomes are less certain the further they are from the central estimate line.

Stripe charts can also be an effective way to show changes over time. A strong example of this is the series of stripe charts created by Dr. Ed Hawkins, a climate scientist at the University of Reading, that showed temperature changes from 1850 to 2018. Each bar (stripe) showed a different temperature level, ranging from cooler blues to hotter reds. Together, readers could quickly and easily see the marked increase in temperatures around the world as a whole and in their specific region of the world by using an online tool. These stripe charts were published by a multitude of websites, television stations, and even became a cover of the *Economist* magazine.

In a 2019 interview on the *Data Stories* podcast, Hawkins said that he "was looking for a way to communicate to audiences that aren't used to seeing graphs, or axes, or labels—things that we see day-to-day, but are complicated to them. It may look too mathematical to them, so it turns them off straight away." In a later interview on that same podcast, Jennifer Christiansen, the senior graphics editor at *Scientific American*, described them as, "every region's version of the climate stripe pattern progresses from cool blue to a warm red. No labels

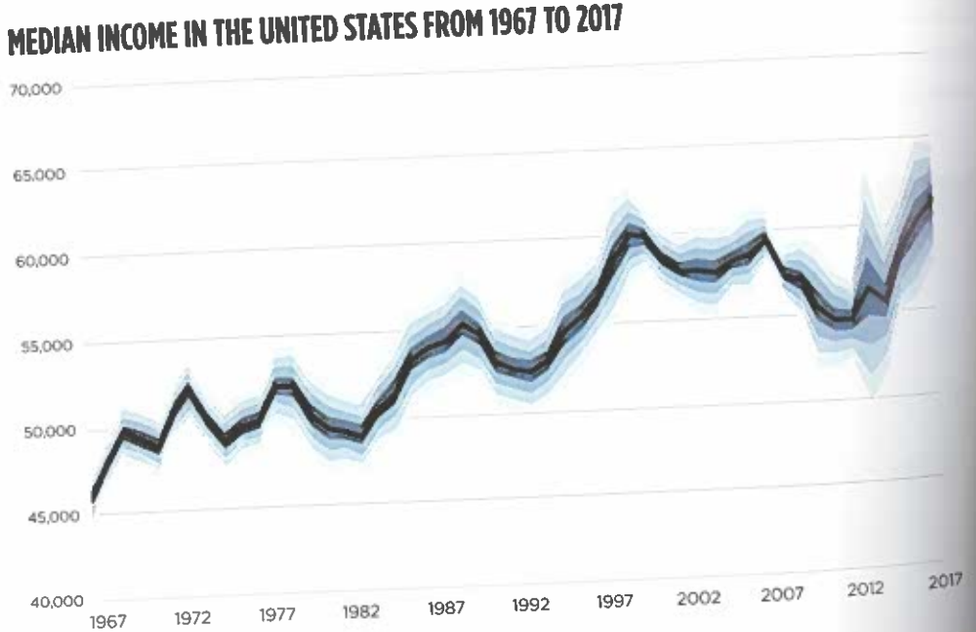


This stripe chart from ShowYourStripes.com shows global temperatures from 1850 to 2018. The simple colored stripes are easy to see and understand.

are needed; no caption is needed. It's a visceral and accessible nod to our warming planet with color representing annual temperature. And it prints legibly on everything from social media profiles, to pins, neckties, magazine covers, mugs, and concert screens."

FAN CHARTS

If the color or saturation of the shaded area between the confidence interval lines changes based on the value, it is often called a fan chart. Fan charts are like gradient plots for line charts, and they are great for visualizing changes in uncertainty over time. In the fan chart, values closest to the central estimate are the darkest and they lighten as the values move outward. The use of color distinguishes the move from higher levels of statistical confidence to lower levels. The advantage here is that it signals to the reader how the estimates become less certain the further they are from the central estimate.

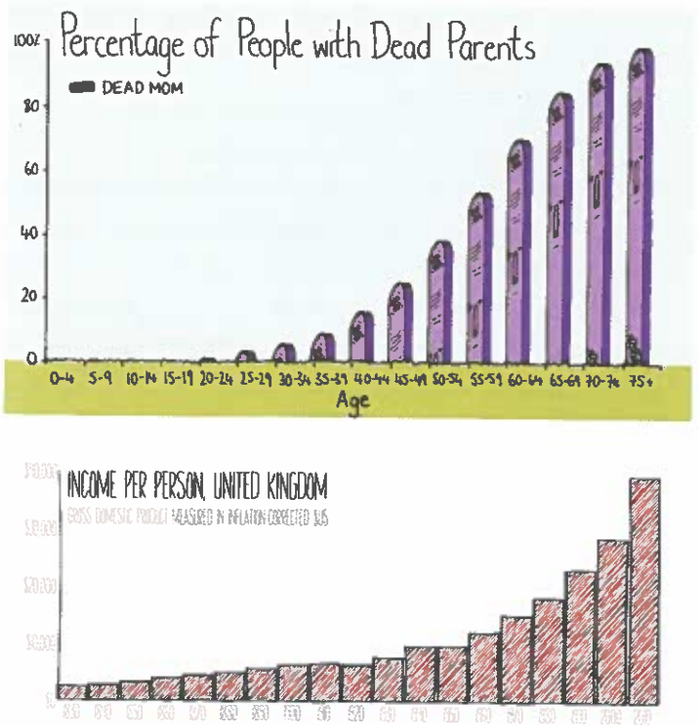


Like the gradient chart, the fan chart shows distributions (in this case, standard errors) around a central estimate.

This fan chart shows the change in median household income over the last fifty years. The color bands show the standard error divided into eight segments, though they could also show bands of percentiles or other measures. Similar to the gradient chart, the change in color saturation can show multiples of the standard error and signal less certainty as we move farther from the central estimate, which is here represented by the black line.

THE HAND-DRAWN LOOK

One last strategy to suggest uncertainty is not a visualization technique per se but a design technique. The hand-drawn, "sketchy," "gooey," or "painty" techniques can be used to add an



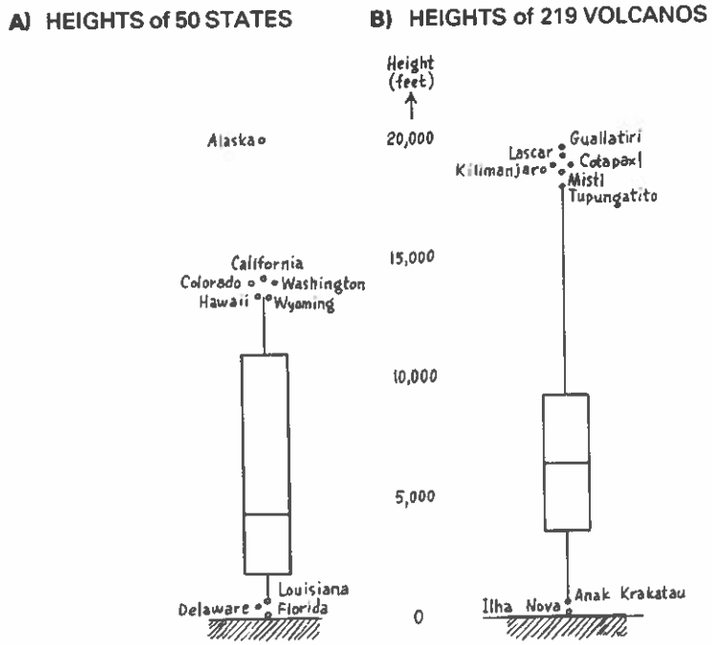
Hand-drawn, "sketchy," or "gooey" design effects use uneven edges to communicate a sense of uncertainty or imprecision.

Sources: Copyright Mona Chalabi (top) and Jo Wood, giCentre, University of London (bottom).

uneven edge or fuzziness to graph objects that will create a sense of uncertainty or imprecision. Research suggests that sketchy graphs generate more engagement in the graph and that we can "tie sketchiness to uncertainty or significance values." These two examples, the first from journalist Mona Chalabi at the *Guardian* and the second from Jo Wood at the University of London, both demonstrate these techniques in action.

BOX-AND-WHISKER PLOT

When you visualize the distribution of your data, you can show the entire distribution or just specific points within it. The box-and-whisker plot (or boxplot), originally called a *schematic plot* by its inventor John W. Tukey, uses a box and line markers to show specific percentile values within a distribution. You can also add markers to show outliers or other interesting data points or values. It is a compact summary of the data distribution, though it displays less detail than a histogram or violin chart.



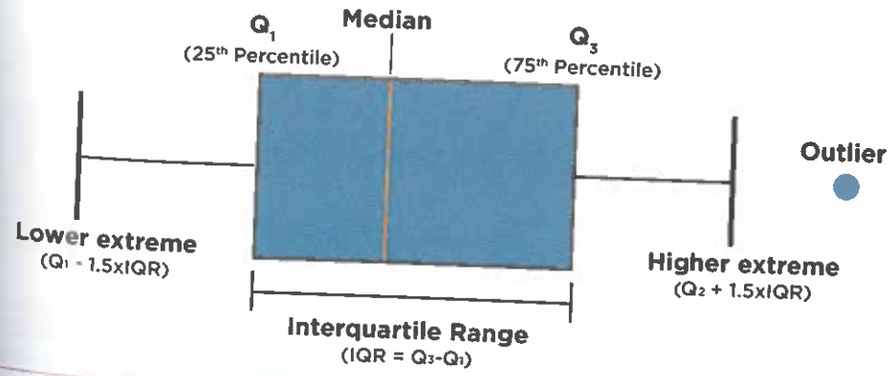
The original boxplot from Tukey 1977

The basic box-and-whisker plot consists of a rectangle (the *box*), two lines (the *whiskers*) that emanate from the top and bottom of the box, and dots for outliers or other specific data points. Most standard box-and-whisker plots have five major components:

1. The *median*, encoded by a single horizontal line inside the box.
2. Two *hinges*, which are the upper and lower edges of the box and typically correspond to the first quartile (or the 25th percentile) and third quartile (75th percentile). The difference between these two points is called the *Interquartile Range* or *IQR*.
3. The higher and lower extremes (sometimes the maximum and minimum) are placed at a position 1.5 times the *IQR* (recall the Box on page 74).
4. Two *whiskers* (the lines) connect the hinges to a specific observation or percentile.
5. *Outliers* are individual data points that are further away from the median than the edges of the whiskers.

Each of these components can vary depending on which parts of the distribution we wish to show. Some creators replace outliers with fixed quantiles such as the minimum and maximum values or the 1st and 99th percentiles. Some use the semi-interquartile range $(Q_3-Q_1)/2$, which can generate asymmetric whiskers. And some add other descriptive statistics like the mean or standard error. We can also vary the color, line thickness, and how and which parts of the chart are labeled.

As a practical example, the box-and-whisker plots on the next page show the distribution of earnings in our thirteen industries. The vertical line in the middle of each box represents



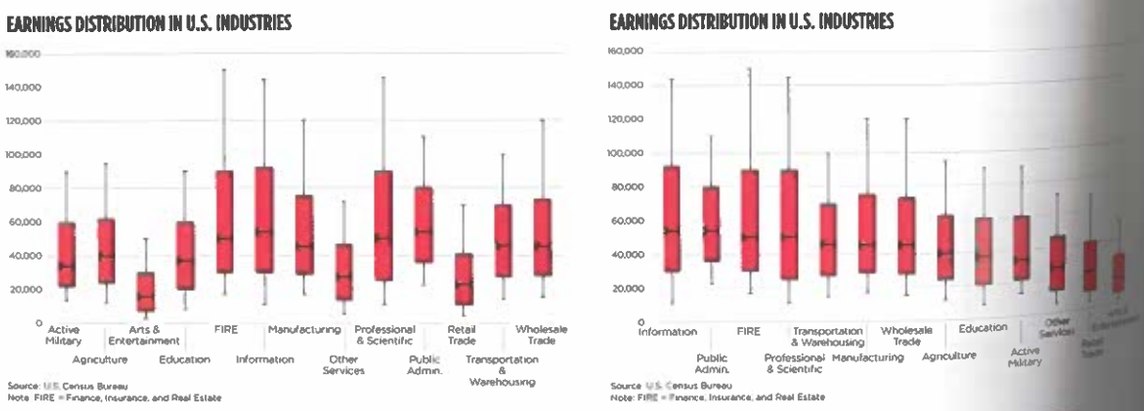
The basic box-and-whisker plot.

the median and the edges of the box represent the 25th and 75th percentiles. The ends of the lines (the whiskers) show the 10th and 90th percentiles.

The graph on the left sorts the industries alphabetically while the one on the right sorts them by the median value. I generally find it preferable to present data sorted by their values as opposed to alphabetical or some other arbitrary sorting. There may be cases when alphabetical sorting is best—for example, if we were presenting earnings across all fifty U.S. states, readers would find it easier to find individual states if they're sorted alphabetically. If, however, the goal of that graph is to discuss the high/low earnings of a particular state, we would sort the data by their values to make those comparisons easier.

As with all of these visualizations, showing percentiles and statistical or data uncertainty will often depend on the experience, interest, and expertise of our audience. In scientific or research applications, for example, communicating uncertainty is especially important to demonstrate whether a finding is statistically meaningful. But in other cases, where our data only have a single value for each observation—say, a single estimate of per capita GDP in the United States—we may not be able to visualize parts of the distribution.

In the case of the box-and-whisker plot, by plotting these specific percentile points, we are explicitly deciding *not* to visualize the entire distribution. This may not be entirely problematic, especially if other percentile points aren't particularly important, or if the data follow a fairly standard distribution. But we must always fully explore the data we're certain we're not hiding important patterns from our reader—or ourselves!

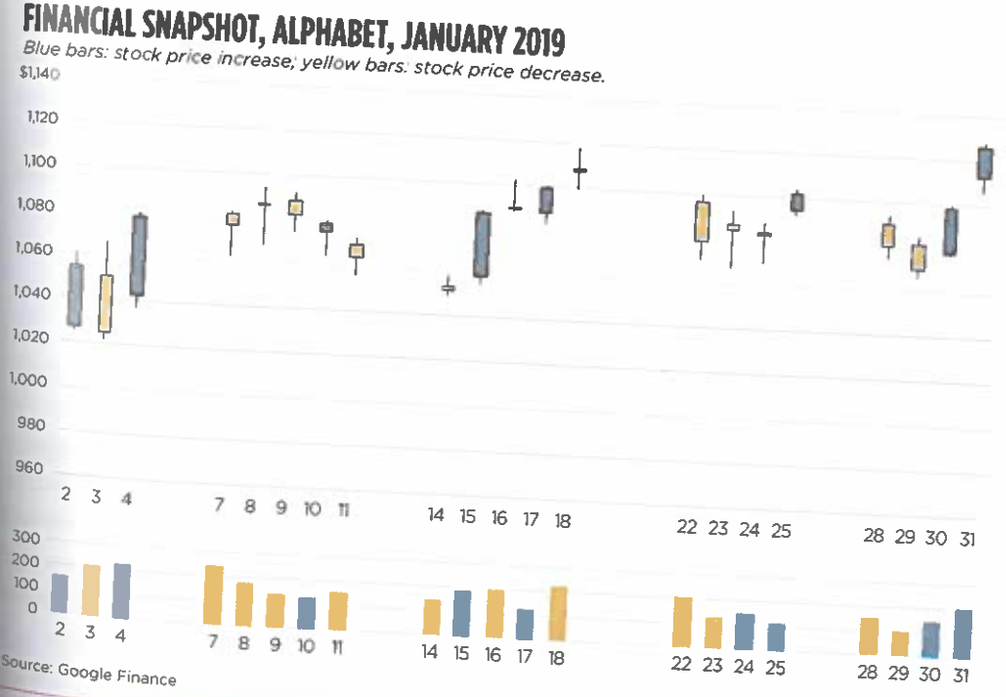


These charts show the distribution of earnings in thirteen industries either sorted alphabetically (left) or by median value (right). The edges of the box show the 25th and 75th percentiles and the whiskers show the 10th and 90th percentiles.

CANDLESTICK CHART

Candlestick or stock charts look like box-and-whisker plots, but they visualize different content. Whereas box-and-whisker plots visualize uncertainty or a distribution, candlestick charts visualize changes in the prices of stocks, bonds, securities, and commodities over time. Bars and lines show opening and closing prices and highs and lows in a day, plotted along a horizontal axis that measures time.

There are two elements of the candlestick chart. The central box—sometimes called the "real body"—shows the gap between the opening and closing prices. The lines that extend upward and downward from the real body—sometimes call the "wick"—show the low and high value for the day. Like the box-and-whisker chart, the candlestick chart includes specific points and does not show *all* of the activity during the day, such as price volatility.



The candlestick chart is like a box-and-whisker chart but is typically reserved to refer to prices of stocks, bonds, securities, and commodities over time.

Specific characteristics of the candlestick chart can vary in some obvious ways: color can be changed to differentiate between a drop in price during the day (i.e., the closing price is greater than or less than the opening price) and icons or other symbols can identify the high and low prices. Although I've placed this chart in this chapter, because of its relation to box-and-whisker plots, it could easily have also appeared in the *Time* chapter or even the *Comparing Categories* chapter.

The candlestick chart on the previous page shows overall daily trading patterns for shares of Alphabet, Inc.—the parent company of the Google search engine—from January to February 2018. The bar underneath shows trading volume. In both graphs, blue bars signal an increase in price over the day and yellow bars signal a decrease. Notice how the two graphs are stacked together as opposed to using a dual axis chart, which might be confusing or just plain cluttered.

VIOLIN CHART

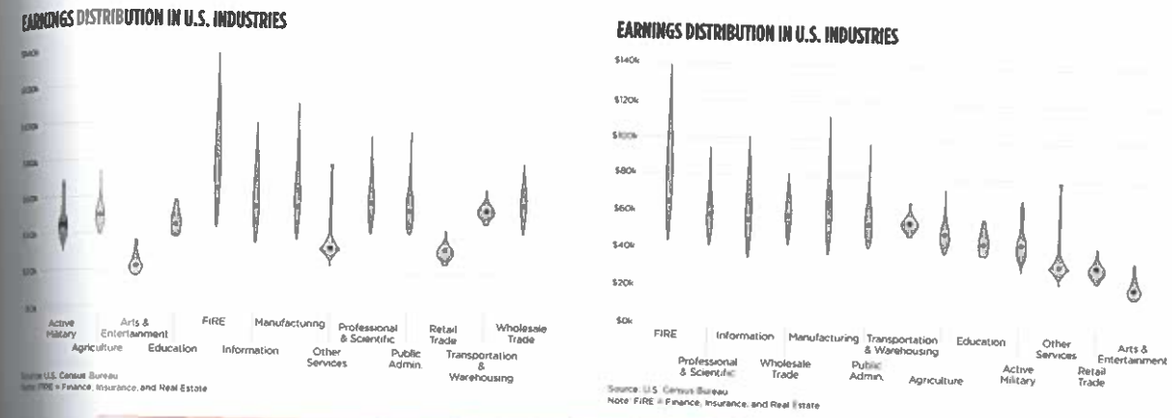
Instead of showing selected percentiles from the distribution, the goal of the next set of charts is to show the *entire* distribution. Unlike the box-and-whisker plot, in which we choose specific points in the distribution, or the histogram in which values are grouped together into intervals, the violin chart shows the shape of the whole distribution.

These violin charts use the same data as above, the average earnings in thirteen industries in 2016. The thicker areas mean that there are more values in those sections while the thinner parts imply lower frequency of observations. I've added a dot in the middle to mark average earnings within each industry. Notice again the differences in the view when the chart is sorted alphabetically (on the left) versus by the mean income (on the right).

KERNEL DENSITY

One consideration in creating this chart type is that it requires estimating what is called the *kernel density* of each distribution. Kernel densities are a way to estimate the distribution of a variable—akin to a histogram—but can be smoothed or made to look more continuous using different algorithms. For the violin plot, those density estimates are plotted to mirror each other around an invisible central line.

Think of it this way: A histogram plots a summary view of a distribution along a single axis. The violin plot mirrors a smoothed version of the histogram on either side of that single



Instead of showing select points (percentiles) in a data distribution, the violin chart shows the estimated shape of the entire distribution using kernel densities.

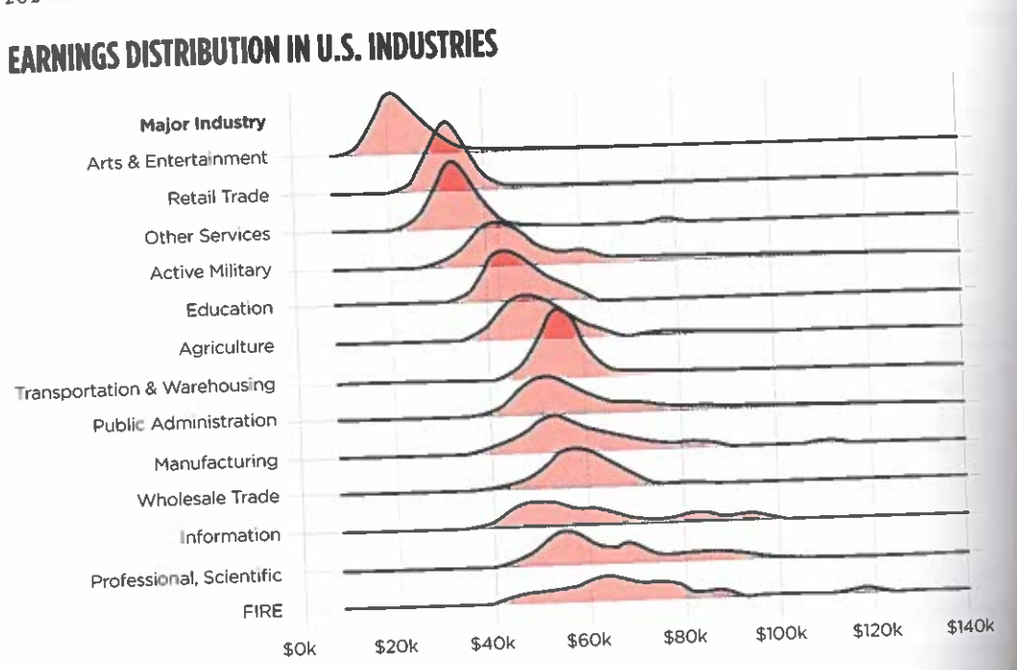
axis. How that smoothing is accomplished will depend on what sort of kernel density estimator you choose, which can vary based on the data, underlying function, and more.

Violin charts, then, are richer than the box-and-whisker plot, but can also be more difficult to create and for our audience to understand. In modern versions of Excel, for example, the box-and-whisker plot is a default graph, while a violin plot requires manually calculating the probability densities and then finding a graphing solution.

RIDGELINE PLOT

The ridgeline plot is a series of histograms or density plots shown for different groups aligned along the same horizontal axis and presented with a slight overlap along the vertical axis. In a basic sense, the ridgeline plot is like a small multiples histogram or a horizon chart where the histograms are aligned in particular way.

The ridgeline plot on the next page shows the earnings distribution across the thirteen different industries. The horizontal axis is shared across all thirteen industries and the distributions sometimes overlap along the vertical dimension. Depending on the color scheme and density of the data, there may be more or less overlap between the series, but as we have seen in other graph types (for example, sparklines and horizon charts), sometimes showing the overall pattern is more important than the reader being able to pick out all of the specific



Source: U.S. Census Bureau

The ridgeline plot is a series of histograms shown for different groups aligned along the same horizontal axis and presented with a slight overlap along the vertical axis.

values. These overlaps can be a problem, but one that is mitigated by how quickly and easily readers see how the different distributions line up with one another along the same axis.

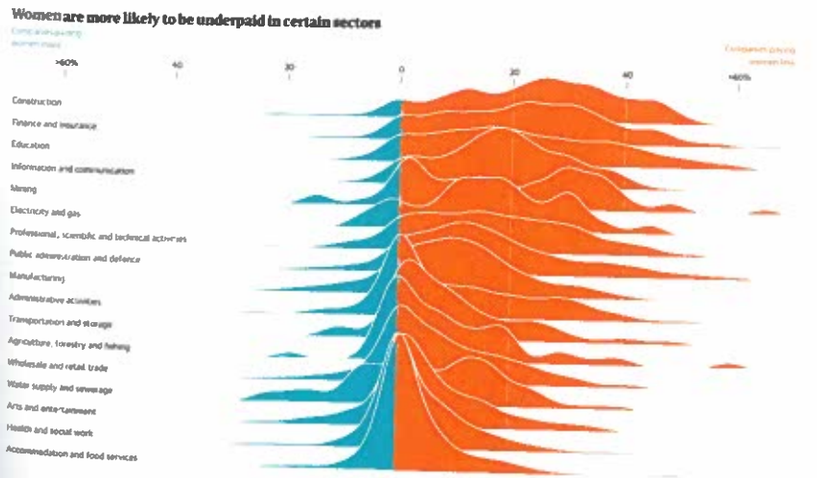
The most famous ridgeline plot is one that you didn't even know was a ridgeline plot: The album cover for *Unknown Pleasures*, the 1979 debut album of English post-punk band Joy Division, which had white lines on a black background with no band name, album title, or other identifiers.

In 2015, Jen Christiansen, the Senior Graphics Editor at *Scientific American*, tracked down the original image to the 1970 doctoral dissertation of Harold D. Craft, Jr., a radio astronomer at Cornell University. The original chart graphed the distribution of consecutive radio pulses emanating from a pulsar, a type of neutron star. The album cover designer, Peter Saville, called it "a wonderfully enigmatic symbol for a record cover."

Data that lend themselves well to a ridgeline plot are those in which the distributions differ from one category (row) to another so that the reader can see the shift up and down the page. With data in hand, variations on color, font, and layout can help engage and interest your reader. The ridgeline plot on the next page was published by the *Guardian* in 2018 and

Source: Photograph by Jen Christiansen (featuring figure 5.37 from "Radio Observations of the Pulse Profiles and Dispersion Measures of Twelve Pulsars," by Harold D. Craft Jr., September 1970).

shows the distribution of the gap in pay between men and women in more than ten thousand companies and public bodies in the United Kingdom. Using different colors on either side of the vertical zero-percent line (perfect equality) and sorting the data from industries that have the highest pay gaps (e.g., Construction) to smallest pay gaps (e.g., Accommodation and Food Services) also helps direct the eye.



This ridgeline plot from the *Guardian* shows earnings distributions for men and women in different industries.

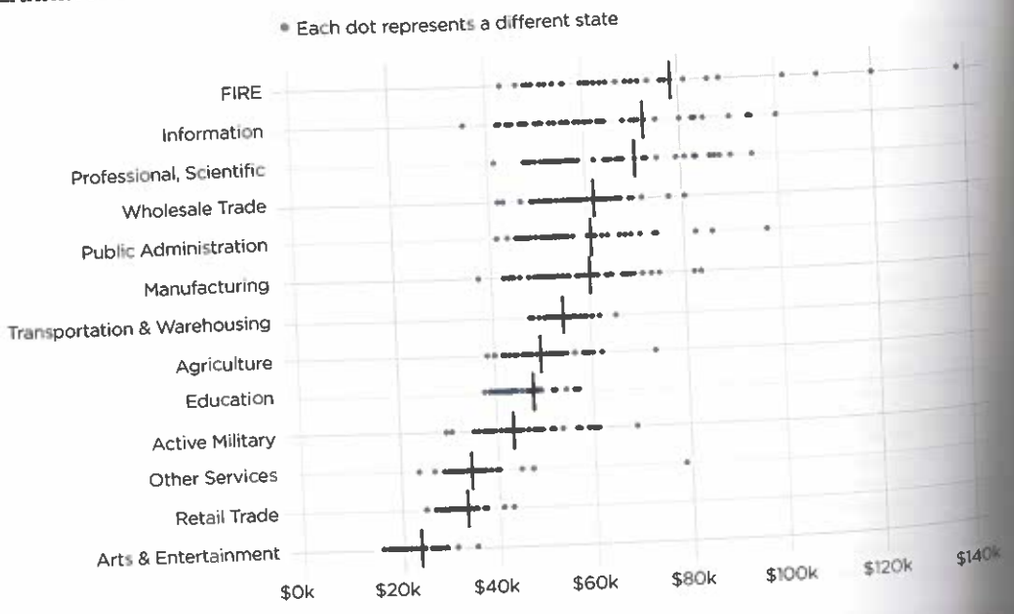
VISUALIZING UNCERTAINTY BY SHOWING THE DATA

The graphs presented in this chapter so far summarize data distributions with lines, points, bars, and color. To different degrees, this is what the histogram, violin, and ridgeline plots all do. Another way to visualize the distribution of your data is to just *show* the data.

STRIP PLOT

The basic way to show your data is with what is known as a *strip plot*. In this graph type, the data points are plotted along a single horizontal or vertical axis.

In this example, average earnings for each state are shown for each of the thirteen industries (the U.S. average is denoted with the vertical black line). We have already seen similar data presented in the box-and-whisker plot and violin chart, but here you can see individual

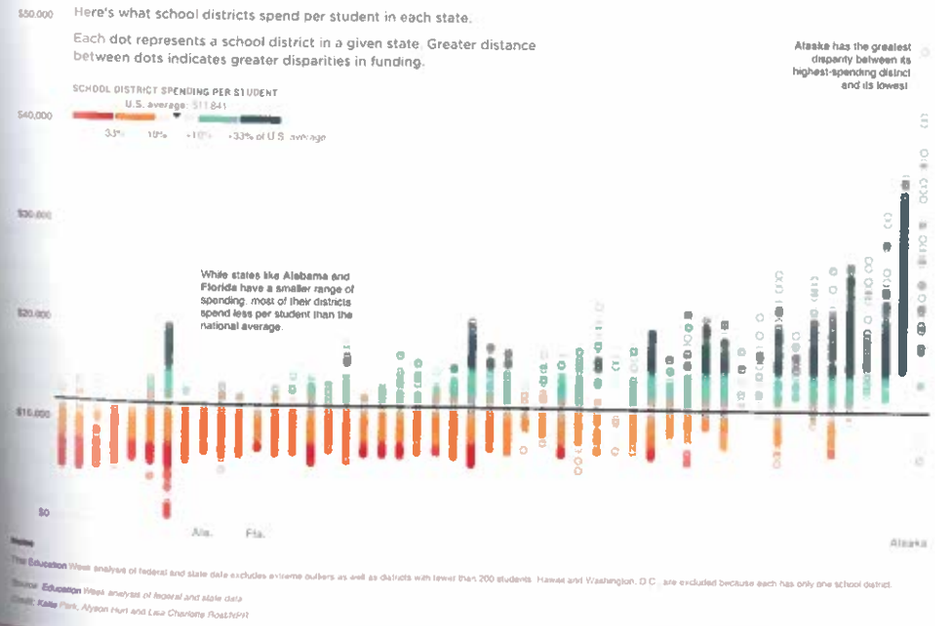


In a strip plot, data points are plotted along a single horizontal or vertical axis. This strip plot encodes the data with circles, but small lines are also often used.

points. Notice, however, that I'm showing average earnings for the fifty states, not for individual workers. Plotting earnings for everyone in my data (some 1.3 million people) would look like a single, dark line. There are too many values.

It's true that some of the data is obscured, but, especially by virtue of the overlapping transparent colors that make the patterns darker, it becomes clearer where the bulk of earnings lie in the distribution. There's no rule for how many data points are *too many*, but as you plot your data, you can always tell when you've passed that threshold.

This static image from an interactive strip plot from NPR is a good example of how this visualization can be richer than a standard bar chart or histogram. Here, they plot a point for every school district in every state. Darker orange dots (and below the black horizontal line) are districts in which spending per student is below the national average. Darker green dots are districts in which spending per student is above the national average. An interesting and useful design choice to make the dots transparent (with a solid border) lets us see where districts are close enough to overlap. Also notice that only Alabama, Florida, and Alaska are



This strip plot from NPR shows the distribution of spending in different school districts around the United States.

labeled along the horizontal axis. Those three states are explicitly annotated in the chart. Other state labels only appear along the horizontal axis when, in the interactive version, the user hovers over a stack of circles.

BEESWARM PLOT

If we want to plot individual data points rather than distributions, one way to make the data more visible is to use a technique called "jittering." This is when we alter individual values slightly so that the data points don't lie on top of one another.

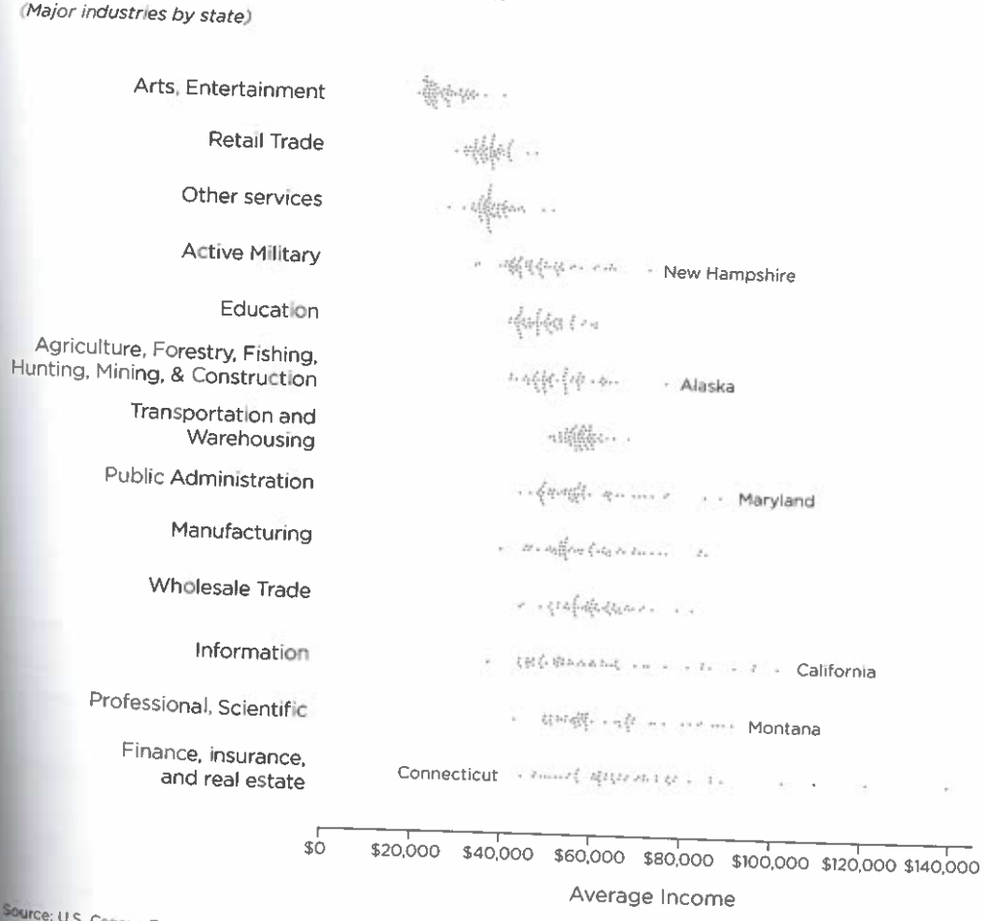
Consider the strip plot on the left, which leaves all of the data along the same horizontal axis. We can see clusters, but not all of the individual values. In the version on the right, the data are jittered along the horizontal and vertical axes to help make each point visible. There are different algorithms and approaches to jittering the data, and the most important consideration is to manipulate the values just enough to make them visible but not so much that it changes the overall view of the distribution. As with choosing a kernel density estimator for violin charts, the choice of jittering technique (for example, do we jitter both x and y variables and if so, do we jitter each variable independently?) will depend on the data and its underlying distribution.

Jittering doesn't work in every case, of course. We are limited by how many points we can plot. With too many data points, jittering would require so much movement that it would modify the underlying distribution. Plotting *everyone's* earnings in each industry, for example, creates a mass of dots that requires so much jittering that it modifies the presentation of the data by moving the points too far away from their true position. But showing average earnings in each of the fifty states across the thirteen industries is not as overwhelming, and you can see where the bulk of the distribution lies in each. Of course, if you're interested in just showing that there are a *lot* of points in your dataset and you can maintain the overall shape of the data, plotting many points may help make your exact argument.

To create what is called a *beeswarm plot*—because the clustering of the data points resembles a swarm of bees—we jitter the values so that they don't overlap and each point is visible. As with other charts in this chapter (and coming up in Chapter 8), there are different calculations

we can use to arrange the dots—for example, arranging the points in increasing order or placing them in a square or hexagonal grid. Here, similar to the ridgeline plot, each industry shares the same horizontal axis so that we can easily compare across the different sectors.

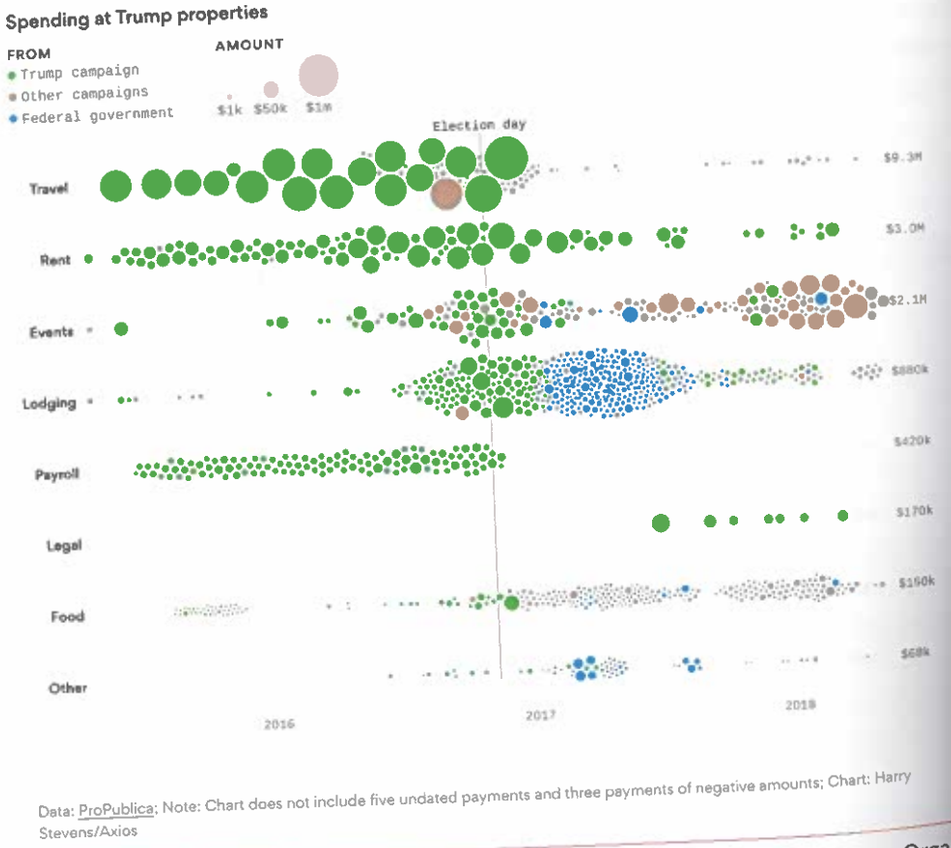
Notice how I have added some simple annotation and labeling to some of the outlier values. These clearly stand out in the graph, and a curious reader will wonder what's going



A beeswarm plot "jitters" all of the points in the data set so they don't overlap and each point is visible.

on with those points. Are they errors? If not, what is that state, and why are earnings so high relative to the rest of the country? I haven't labeled *every* outlier point, but enough to assure the reader I've thought about those values.

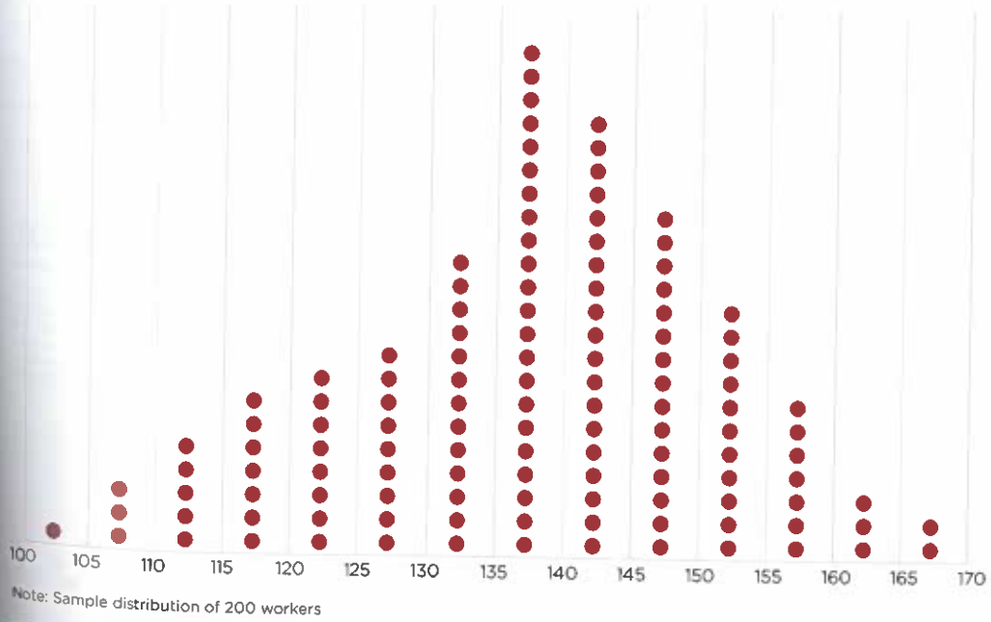
Beeswarm plots can also show changes over time. This beeswarm plot from Axios—really eight beeswarm plots aligned together—shows spending at properties owned by The Trump Organization before and after the 2016 election. The combination of color (spending origin), size (amount of spending), and the density of points (the time dimension) makes this an effective visualization to show the patterns around Election Day.



This beeswarm plot from Axios shows spending at properties owned by The Trump Organization before and after the 2016 election.

WILKINSON DOT PLOTS AND WHEAT PLOTS

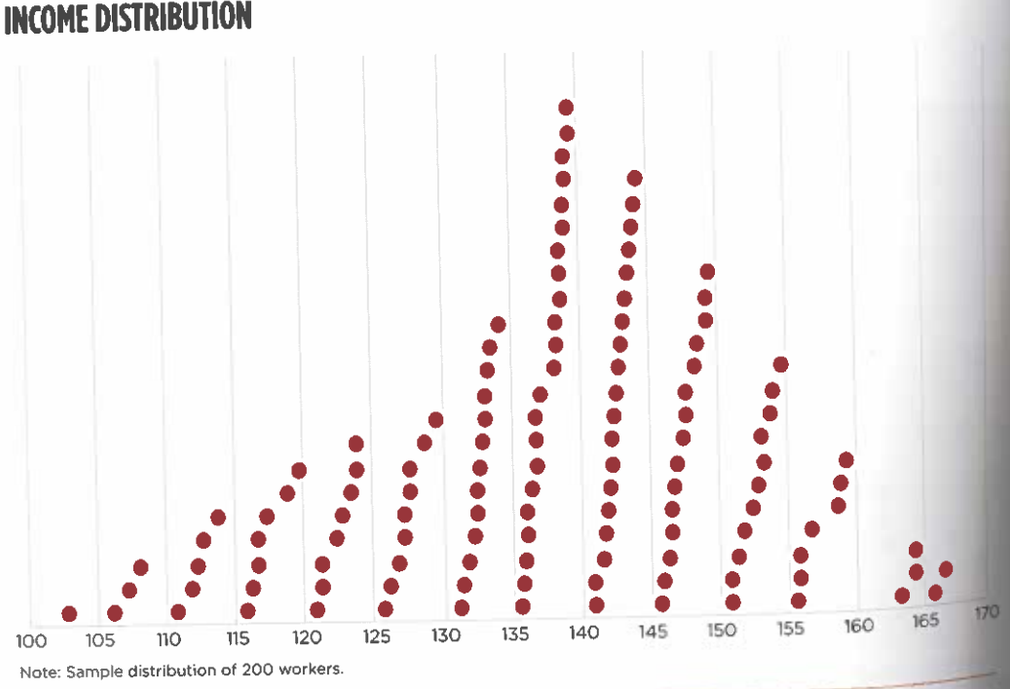
The wheat plot, developed and named by Stephen Few, is a richer version of what is called a dot histogram or a Wilkinson Dot Plot (which is named after Leland Wilkinson author of the seminal data visualization book *The Grammar of Graphics*, though Wilkinson himself actually referred to these charts as *histodot plots*, a name that clearly did not stick). A Wilkinson Dot Plot is like a regular histogram except that instead of showing a single bar encoding all of the observations, data points are stacked on top of each other within their relative bins—something like combining a histogram and unit chart. With this approach, the points do not show their actual *value* (measured along the horizontal axis) because they are stacked in a single column. In other words, each dot represents an observation in each bin, not its actual value.



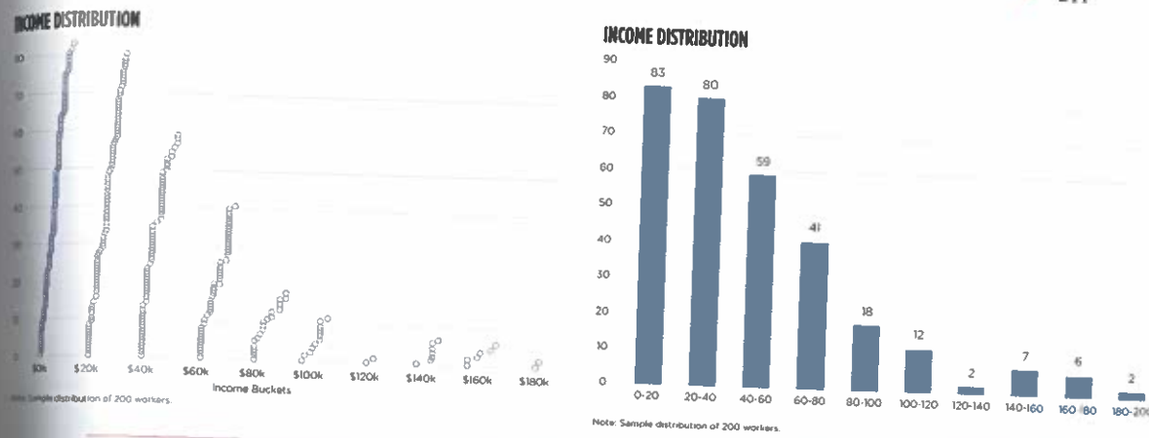
A dot histogram plots individual points within the bins of a histogram. Of course, it can only handle so many individual points.

The Wilkinson Dot Plot can be modified to create a wheat plot in which we show the *actual* data values, stacking them within their separate bins. The actual data values are plotted along the horizontal axis—still grouped into their bins—and stacked vertically to show the total number of observations. Stephen Few writes that, "The curved alignment of the dots is meaningful, for it graphically displays the distribution of values within each interval, based on their positions. Although this looks odd at first glance, it takes only a minute to understand and learn how to read." As with some of the previous distribution graphs, one of the limiting features of the wheat plot is that too many data values may lie on top of each other.

The wheat plot on the next page shows the distribution of earnings for a single industry for about two hundred workers. The histogram on the right is shown as a comparison—you can still see the relative share of observations in each part of the distribution, but not the actual data. There is an obvious tradeoff. On the one hand, the wheat plot shows more detail for the reader to explore and the graph can look more interesting and engaging. On the other hand, the histogram is likely more easily and immediately understandable to readers.

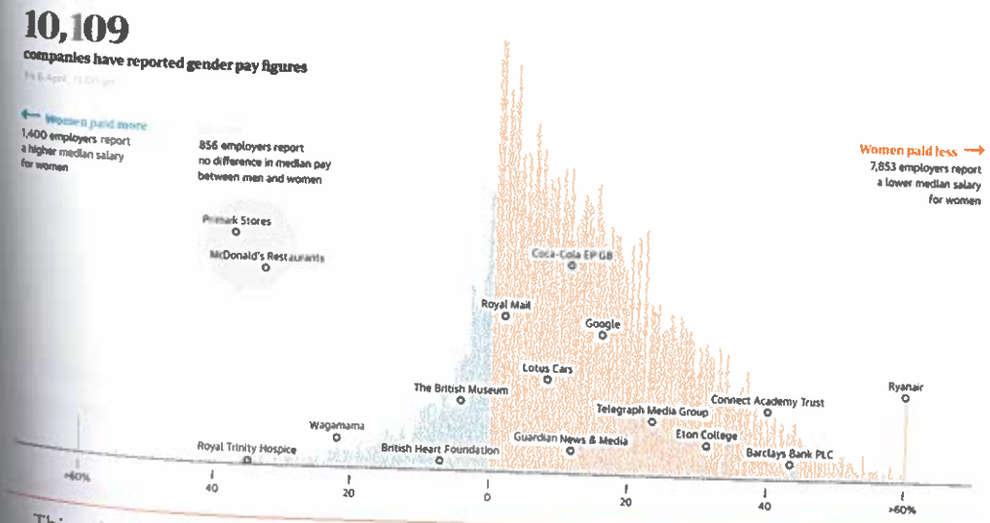


The wheat plot, designed by Stephen Few, adjusts the dot histogram by showing the exact values, still within each bin.



The tradeoff between wheat plots and simple histograms: the wheat plot has more detail but may be harder for people to understand.

We can see the difference between say, a wheat plot and a ridgeline plot in this visualization from the *Guardian*. Leading off the same article that has the ridgeline plot on page 201, this graph includes a dot for every company in their data set. It doesn't quite have the "lean" that a true wheat plot might have (because there are so many points), but you get a good sense of the overall distribution and the greater number of firms towards the right side of the graph. Plotting each individual company also lets the chart creators add labels to highlight



This wheat plot from the *Guardian* includes a dot for every company in their data set.

specific companies, which we couldn't do in a standard histogram or ridgeline plot. It is, however, important to realize that the selected labeled points are arbitrarily chosen along the vertical axis, which is simply stacking the points and not tied to any pay gap data.

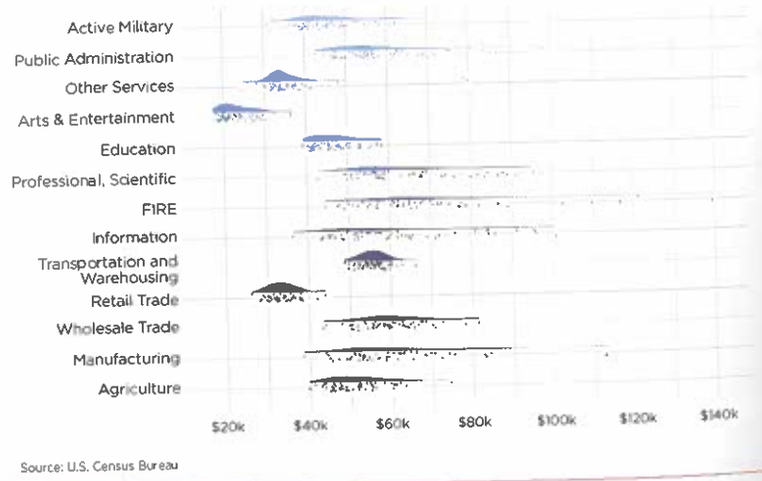
RAINCLOUD PLOT

Sometimes it's useful to show *both* the distribution density of your data and the actual data points. The raincloud plot, perhaps first named by neuroscientist Micah Allan, shows the distribution (think violin chart) with the actual data plotted below. In this arrangement, it looks like a cloud raining data.

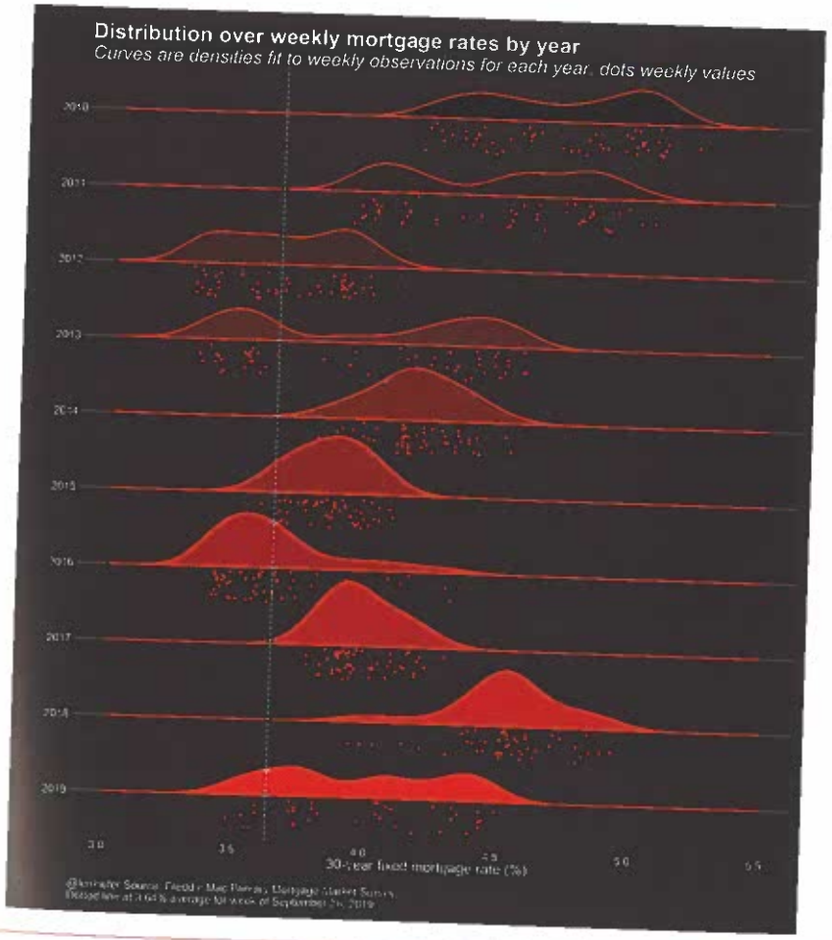
Raincloud plots show us a summary of the data and all the individual data points, so we can spot outliers and patterns. Again, the tradeoff is that this might require more work on the part of the reader to understand how to read the graph.

This raincloud plot shows the distribution of earnings across the fifty states with data values plotted below.

While the raincloud plot may seem like an esoteric chart—and, to be honest, right now it is—there are certainly scenarios and data for which this chart would be a useful choice.



The raincloud plot shows the summary histogram with the actual data plotted below.



An example of a raincloud plot from Len Kiefer that shows the distribution of weekly mortgage rates for different years. The visualization gives you both an overall summary view of the data and shows the specific data points.

This raincloud plot from Len Kiefer, the Deputy Chief Economist at Freddie Mac, shows the distribution of mortgage rates from 2010 to 2019, with weekly observations shown in the dots just below. This view gives us both an overall perspective of the data and the detailed values in the points below.

STEM-AND-LEAF PLOT

The stem-and-leaf plot is a table that shows the place values of each data value. Values are typically shown listed down the "stem" column with the first digit or digits. The rest of the table is reserved for the "leaf" that shows the last digit (or digits).

Stem-and-leaf plots show the place values of each data value. They are sometimes used in transportation schedules, like this train schedule from the Tokorozawa station in Saitama, Japan.

As an example, take a simple dataset with just seven values: 4, 9, 12, 13, 18, 24, and 27. The data are arranged in downward-ascending order with the first digit on the left side and the second (tens) digit on the right. Obviously, for more detailed and complex data, the stem-and-leaf plot may not be a useful approach.

Stem-and-leaf plots are most useful as a reference tool, like a public transportation schedule, or to highlight basic distributions and outliers in a more limited set of data. The Japanese train schedule for the Tokorozawa station in Tokyo on the facing page shows the timing of train arrivals over the course of a day. The hour digit is shown in the far-left column, minutes are shown to the right. The first train begins running at 5:00 am, the next train leaves at 5:08 am, then 5:18 am, and so on. Because the stem-and-leaf plot is a table, it loses some of the advantages of a traditional data visualization, but the leaves illustrate some basic view of the distribution.

CONCLUSION

The collection of graphs in this chapter demonstrates how we can show the distribution of our data or uncertainty around specific values. Some of these charts show summary measures or specific values. We can aggregate the distribution into bins to visualize the distribution in a histogram. Or we might use specific percentiles to generate a box-and-whisker chart, for example, or show stock price variation in a candlestick chart.

With better data visualization tools and faster computers, we can show more data than ever before. Beeswarm charts, wheat plots, and raincloud plots include specific data points. While these visualization types are useful for presenting the full data set to our readers, they have their limits: We can only show so many data points before they begin to overlap and obscure one another.

The graphs in this chapter can introduce challenges to readers who are not familiar with statistical concepts and measures of dispersion. As always, the most important thing you can do when creating your graphs is to remember your audience. If you are a PhD economist presenting your work at your university lunchtime seminar, you don't need to explain the median or variance or even the 95-percent confidence interval. If you were to present the same results to a general audience, however, you would include definitions and annotation. This isn't to say you should avoid presenting these numbers or that you need to dumb things down, but rather that you may need to take time explaining concepts within the visual. The planning, testing, and conceptualizing of your visualization will pay off in the long run as you more effectively communicate your work with your audience.