4

COMPARING CATEGORIES

The graphs in this chapter are intended to help our readers compare values across categories. Bars, lines, and dots can all let our readers compare within and between groups. In some cases, we want our reader to see both levels *and* change, or some other variable combination; in other cases, we want to focus their attention on one comparison or another.

The challenge when comparing categorical data is deciding what we want the chart to convey. Is there a primary argument or story? Is there something you can identify as the most important comparison you want the reader to make? As chart creators, we need to prioritize what we want our charts to do. By putting *every* bar or dot in the graph, we can obscure the point we wish to convey.

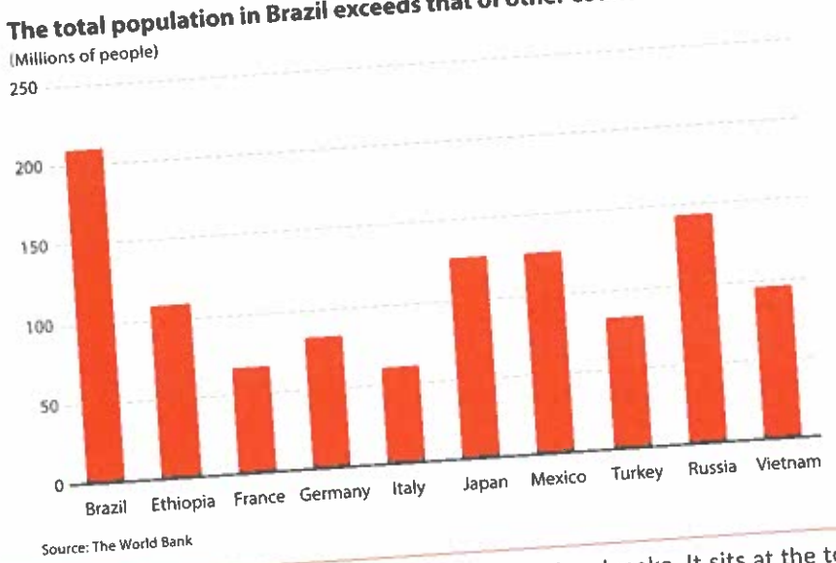
This chapter starts with the bar chart. Like the line chart that will kick off the next chapter, the bar chart is familiar to most readers, which makes it a convenient choice to guide readers as they compare categories or view changes over time. It also sits at the top of the perceptual ranking diagram. It's not necessarily the case that we must *always* give our readers the exact values, but when we do, the bar chart is an excellent choice.

Graphs in this chapter are styled roughly following the guidelines published by Eurostat, the statistical office of the European Union. Eurostat's seventy-six-page style guide covers everything from color, typography, logos, tables, layout, and more elements of a comprehensive style guide that we will discuss in Chapter 12.

BAR CHARTS

One of the most familiar data visualizations, the length or height of the rectangular bars in bar and column charts depict the value of your data. The rectangles can be arranged along the vertical axis so that the bars lie horizontally (often called a bar chart) or vertically on the horizontal axis (often called a column chart). For the sake of brevity, and the fact that whichever way you align them they are still bars, I call these bar charts throughout the book.

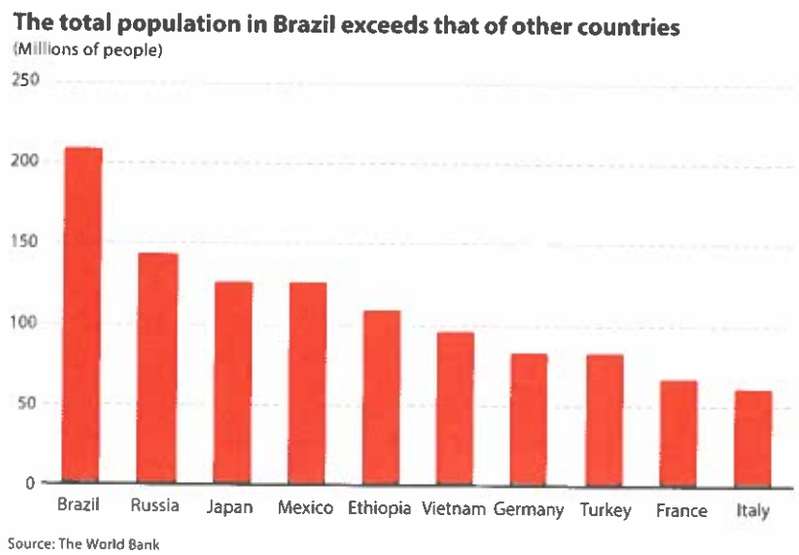
Bar charts sit at the top of the perceptual rankings list. With rectangles sitting on the same straight axis, it's easy to compare the values quickly and accurately. Bar charts are also easy to make, even with pen and paper. This one shows the total population in ten countries from around the world. It's easy to find the least (Italy) and most (Brazil) populous countries in the group, even when they are not labeled with the exact values.



The bar chart is a familiar chart that's easy to read and make. It sits at the top of the perceptual ranking matrix.

Data Source: The World Bank.

It's even easier to see the highest and lowest values when the data are sorted according to their data values. This strategy doesn't always work, however. If, for example, I was showing population levels for sixty countries, I might sort the values alphabetically, so that readers



When possible, sort the data in your bar charts. This makes it easier for your reader to find the highest and lowest values.

Data Source: The World Bank.

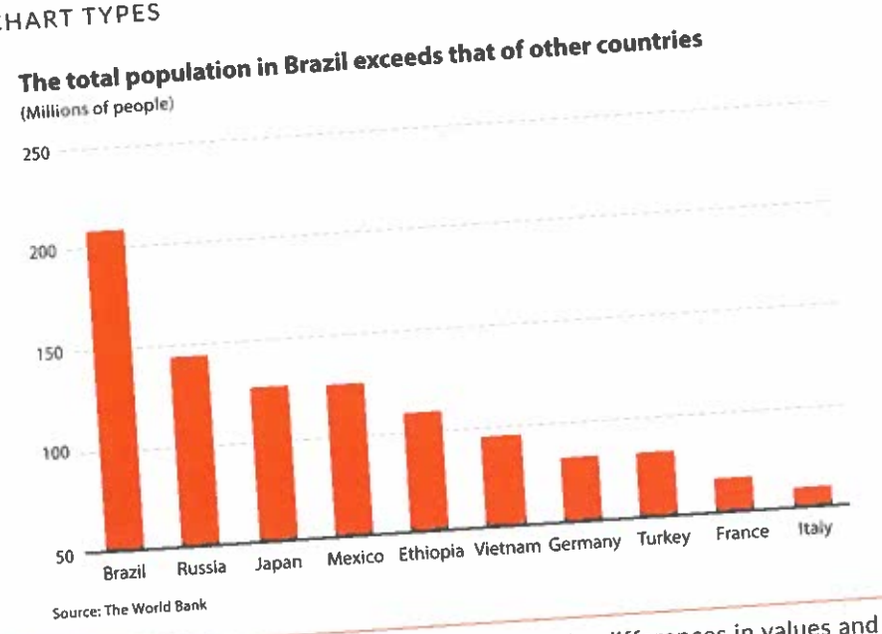
could more easily find the bar for a specific country. But if I was making an argument about the population level in a specific country or set of countries, I might sort the data so that the country or countries of interest are at one end of the graph. Alternatively, I could simply use a different color to highlight whichever bar or bars I want to set apart from the rest.

There are a few strategies to creating bar charts, many of which will apply to other charts in this chapter as well.

START THE AXIS AT ZERO

Starting the axis of bar charts at zero is a rule of thumb upon which many data visualization experts and authors agree. Because we perceive the values in the bar chart from the length of the bars, starting the axis at something other than zero may overemphasize the difference between the bars and skew our perception.

Take the bar chart of population. Because none are lower than fifty million, we might be tempted to start the axis at fifty million. After all, this would emphasize the difference between the values.

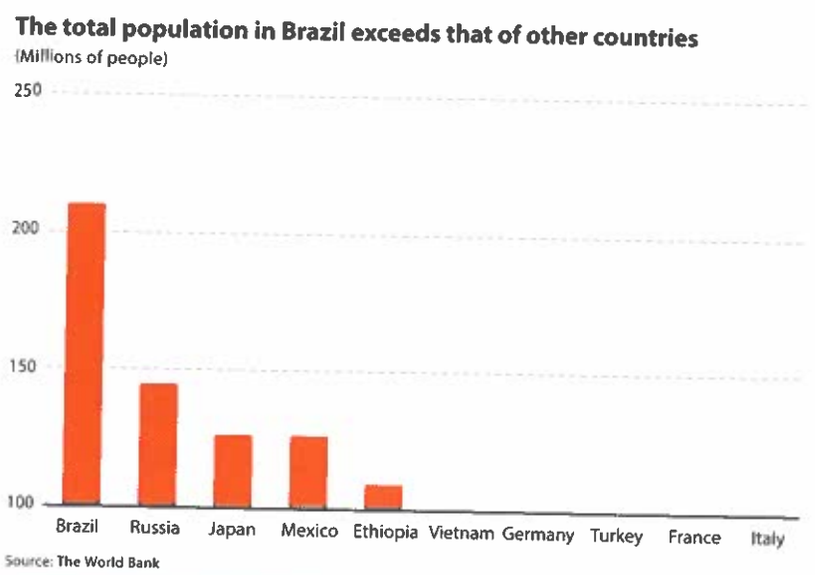


Starting the vertical axis at 50 million overmphasizes the differences in values and skews our perception of the data.

But notice what happens when we do that. The differences in values are emphasized—in fact, they are *over*emphasized. Here, it looks as though Brazil is orders of magnitude larger than Italy, when, in fact, it is only about three-and-a-half times greater. This isn't a matter of moving from accurate perception to general perception—it's a matter of moving from accurate to inaccurate.

If you want to take a more extreme view of this, imagine starting the graph at a hundred million—and why not? If starting at fifty is OK, then we can pick any arbitrary number. Now at a glance it looks like nobody lives in half of these countries!

There is emerging research in this area that suggests that perhaps starting bar charts at something other than zero does not bias our perception of the data. In one recent study, participants were better able to assess the sensitivity of the results (e.g., no effect, small effect, medium effect, or big effect) and more accurate (e.g., the size of the effect) when the vertical axis was set at a range more consistent with the variation of the data. Until more research is conducted, however, my preference is to start the axis in bar charts at zero to avoid any confusion or possibility of visual bias.



If starting the y-axis at fifty is OK, then why not one hundred?
Data Source: The World Bank.

DON'T BREAK THE BAR

Another cardinal sin of data visualization is what is called "breaking the bar"—that is, using a squiggly line or shape to show that you've cropped one or more of the bars. It's tempting to do this when you have an outlier (see Box on page 74), but it distorts the relative values between the bars.

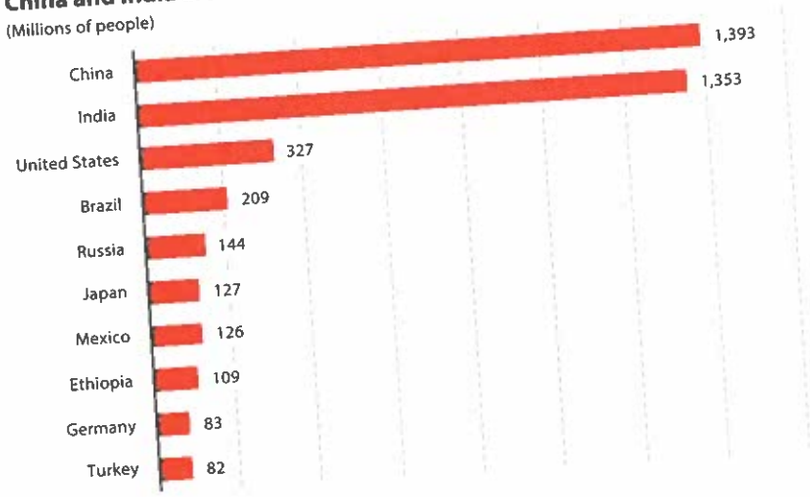
Let's create a bar chart of population in the ten most populous countries of the world. In 2018, China and India were the most populous countries on the planet with 1.39 billion and 1.34 billion people, respectively, followed by the United States with 327 million people. We can see how dramatically larger China and India are relative to the rest of these countries in the top graph on the next page. If we wanted to make the differences between the less populous countries larger, we could break the bars, but this makes China and India look much less populous than they are. Chopping the lengths of the bars is completely arbitrary—I can place those squiggly lines wherever I like to zoom in on the other differences. But that's not being honest with the data.

If you run into a case where you have outliers but want to show the detailed differences between the smaller values, try using more graphs. You might think of this as

a "zoom in" and "zoom out" approach—show all of your data so your reader can see the magnitude of the largest values, and then zoom in for a detailed look that omits the outliers. On the next page, I've highlighted the less populous countries to show the

China and India are the most populous countries in the world

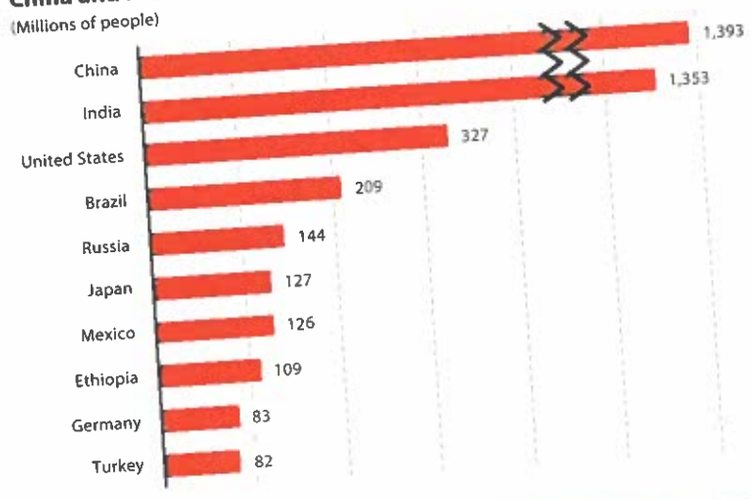
(Millions of people)



Source: The World Bank

China and India are the most populous countries in the world

(Millions of people)



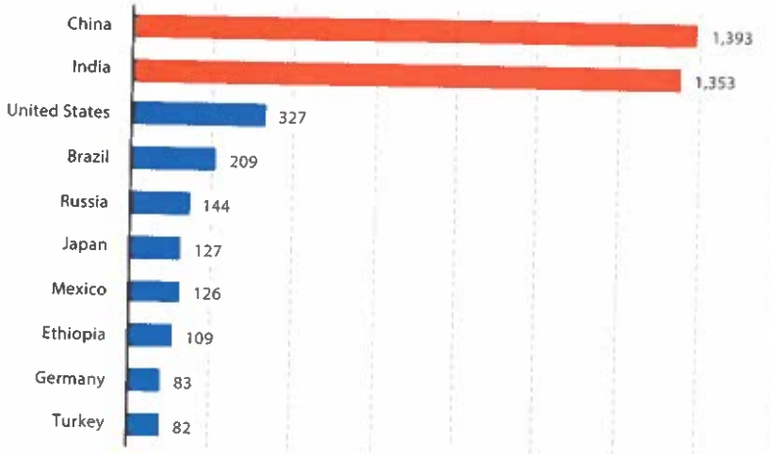
Source: The World Bank

Don't break the bar in your bar charts. The break can be arbitrarily set anywhere and distort our perception of the data.

differences between them, which we can't quite see in the main graph. Adding labels and an active title is another good way to communicate the differences between smaller values to the reader.

China and India are the most populous countries in the world

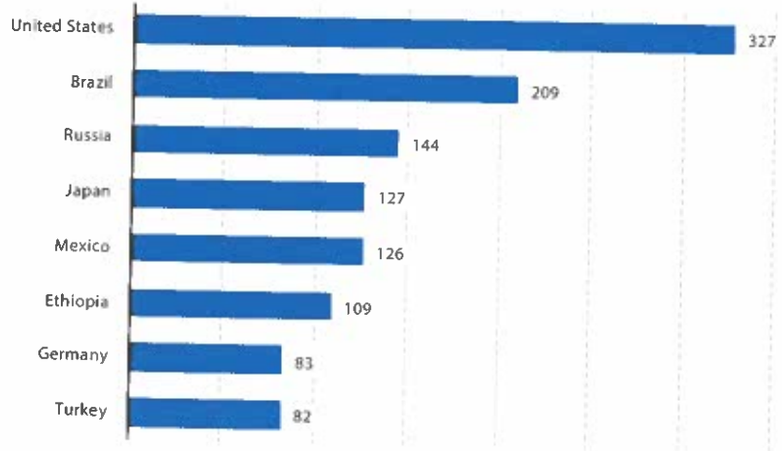
(Millions of people)



Source: The World Bank

Total population in these countries ranges from 82 million to 327 million

(Millions of people)



Source: The World Bank

In cases where you have large values or outliers but want to show the detailed differences between the smaller values, try using more graphs.

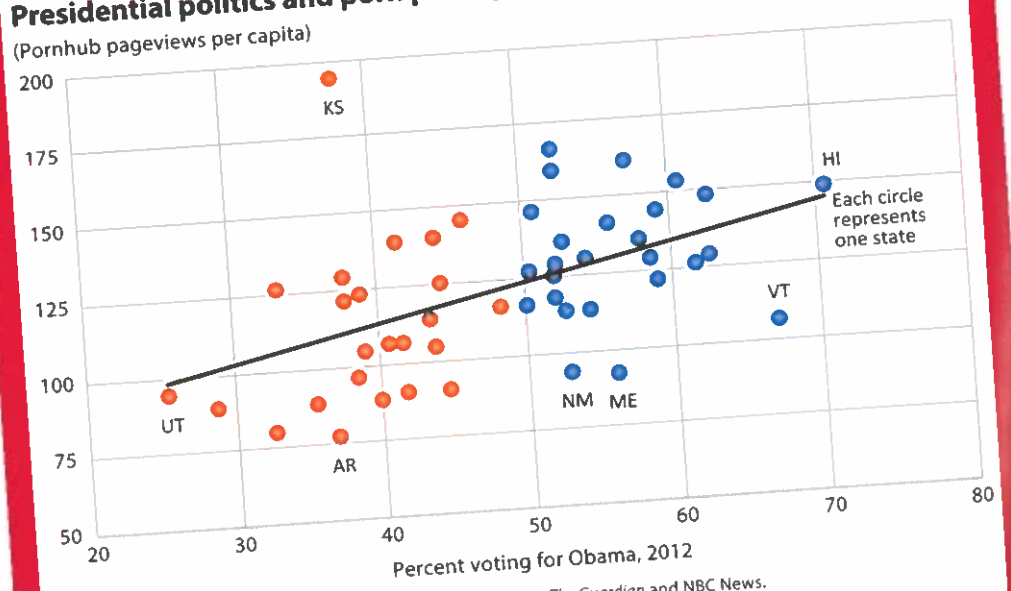
EXTREME VALUES OR OUTLIERS

An outlier is a data point that is far away from other observations in your data. It may be due to random variability in the data, measurement error, or an actual anomaly. Outliers are both an opportunity and a warning. They potentially give you something very interesting to talk about, or they may signal that something is wrong in the data.

In 2014, Buzzfeed teamed up with the website Pornhub to look at pornography viewing by state. Using geolocation data of people accessing their site, Pornhub calculated the number of page views per person in each state. People in Kansas, they found, watched far, far more pornography than any other state in the country: 194 page views per person. Nevada was second with 166 page views.

The data went into the scatterplot below, comparing blue-state and red-state porn consumption. You can clearly see Kansas as an outlier in page views. Do people in Kansas really watch that much more porn?

Presidential politics and porn per capita



Source: Pornhub views from Buzzfeed; Voting percentages from *The Guardian* and NBC News.
Scatterplot originally created by Christopher Ingraham.

Turns out the answer is no. Apparently, Pornhub's methodology assigned missing geolocation data to the geographic center of the United States, which, as it turns out, is Kansas.

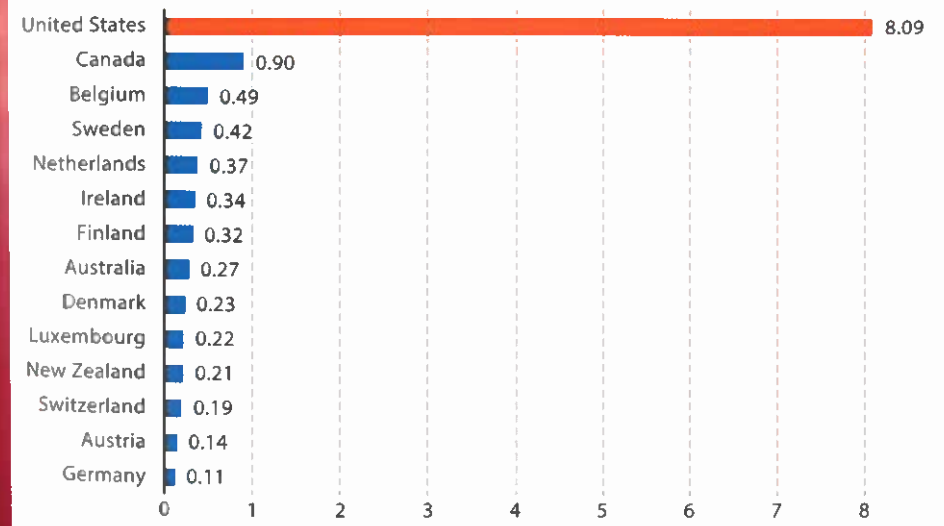
Not all outliers are mistakes, however. As just one example, we can look at the rate of physical violence from a firearm across advanced countries. In 2017, more than 8 per 100,000 people were victims of firearm violence in the United States, compared with 0.90 per 100,000 people in Canada and 0.49 per 100,000 in Belgium. In some cases, outliers are truly outliers.

There are lots of ways to test for outliers in your data, some more complex than others. One way is to simply *look* at your data. Exploring your data does not need to start with complex math and statistics—you should always visually inspect your data.

But that approach is hardly mathematical. A standard method is to compare data values to 1.5 times the interquartile range (IQR). The IQR is a simple summary of your data and is the difference between the third and first quartiles (see the Box in Chapter 6 on percentiles).

The United States has the highest rate of physical violence by firearm among advanced countries

(Rate per 100,000 people, 2017)



Source: Institute for Health Metrics and Evaluation

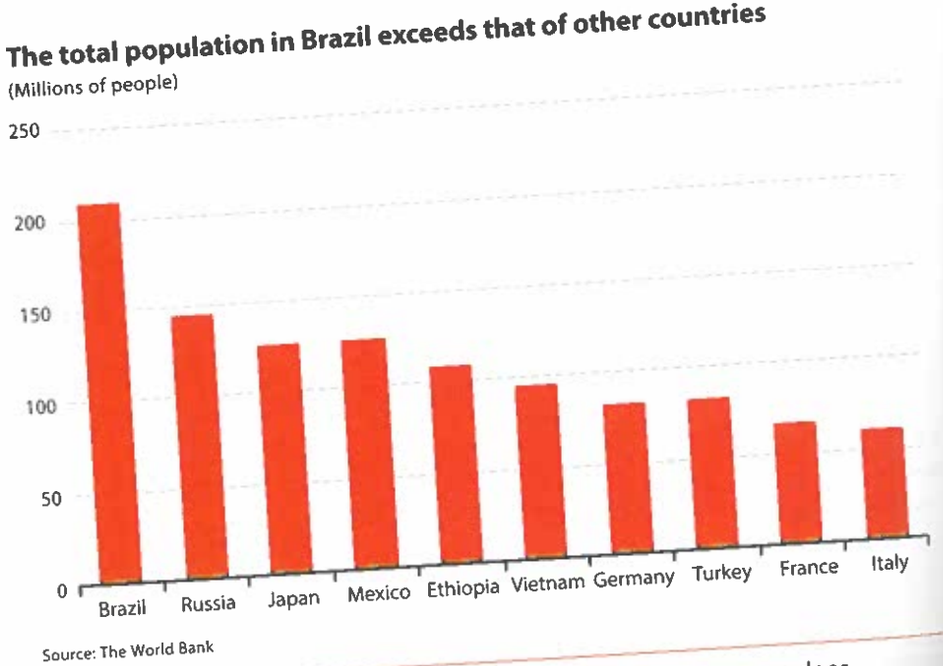
USE TICK MARKS AND GRIDLINES JUDICIOUSLY

Bar charts don't need tick marks between the bars. White space is an effective separator and deleting the tick marks reduces clutter.

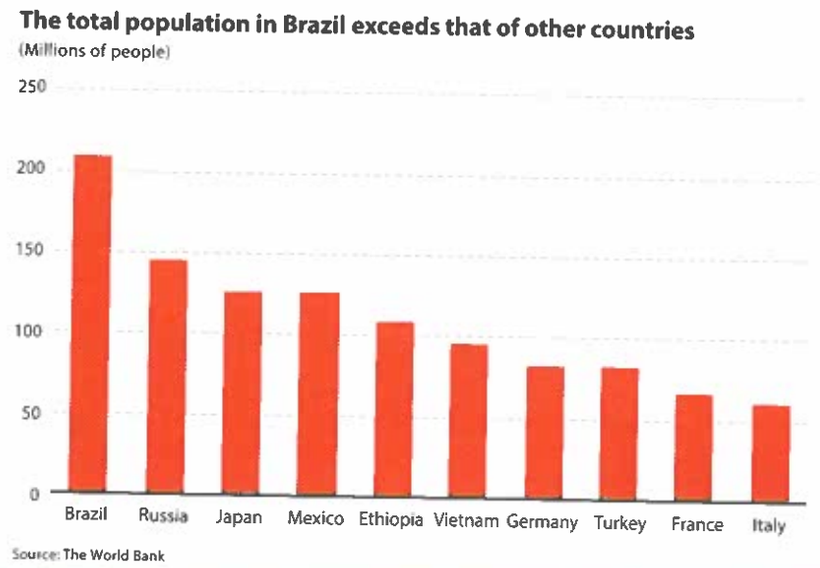
One exception is if you have a "major" category label that spans multiple bars. In such cases, larger tick marks can be helpful to group the labels (see the bottom chart on the next page).

Gridlines help the reader see the specific values for each bar and are especially useful for the bars farthest from the axis label. Because they serve as a visual guide, they can be rendered in a lighter color so the reader's eye stays on the data.

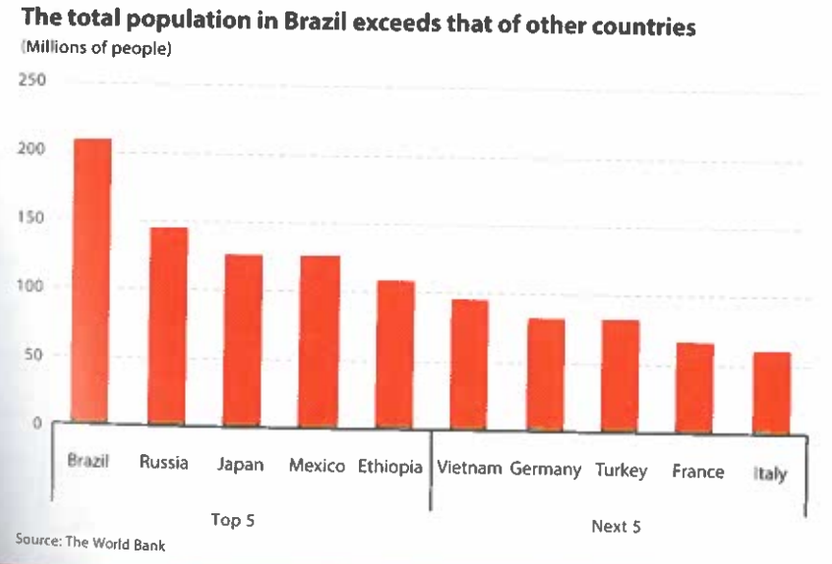
When it's important for the reader to know the *exact* values, you can add data labels to the chart. My preference is to forgo the gridlines and axis lines altogether in these cases.



In bar charts, tick marks are not necessary. The white space does the job of separating the bars.

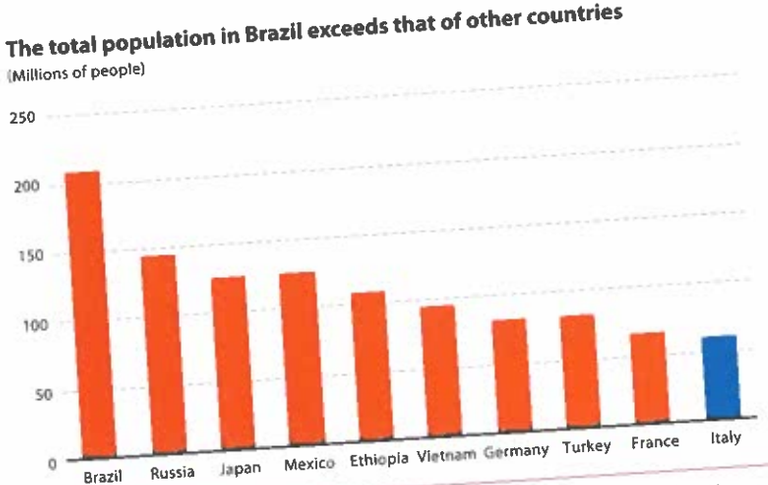


Omitting tick marks is part of removing as many non-data elements as possible.



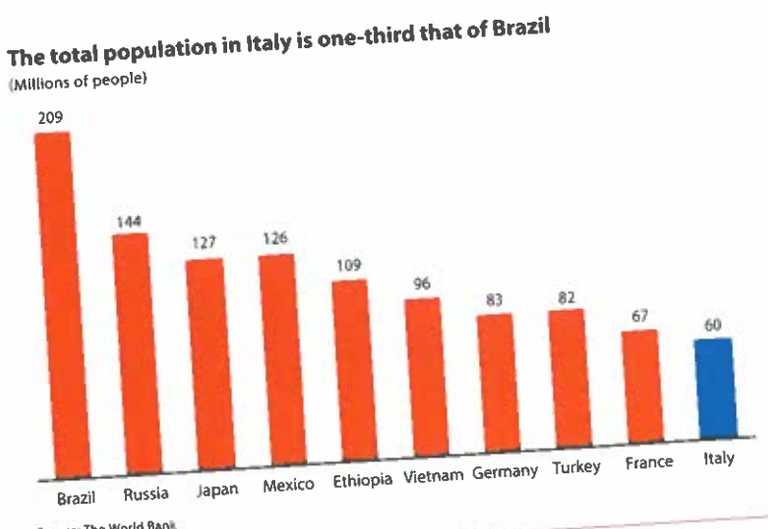
Tick marks may be necessary when you have a "major" category.

Consider Italy in the next two graphs (highlighted in blue). Without the labels, the gridline helps us see that there are more than fifty million people living in the country; with the label, it is clear that it's sixty million people and thus the gridlines are probably not necessary.



These horizontal gridlines help the reader see that, for example, there are more than fifty million people living in Italy.

Data Source: The World Bank.



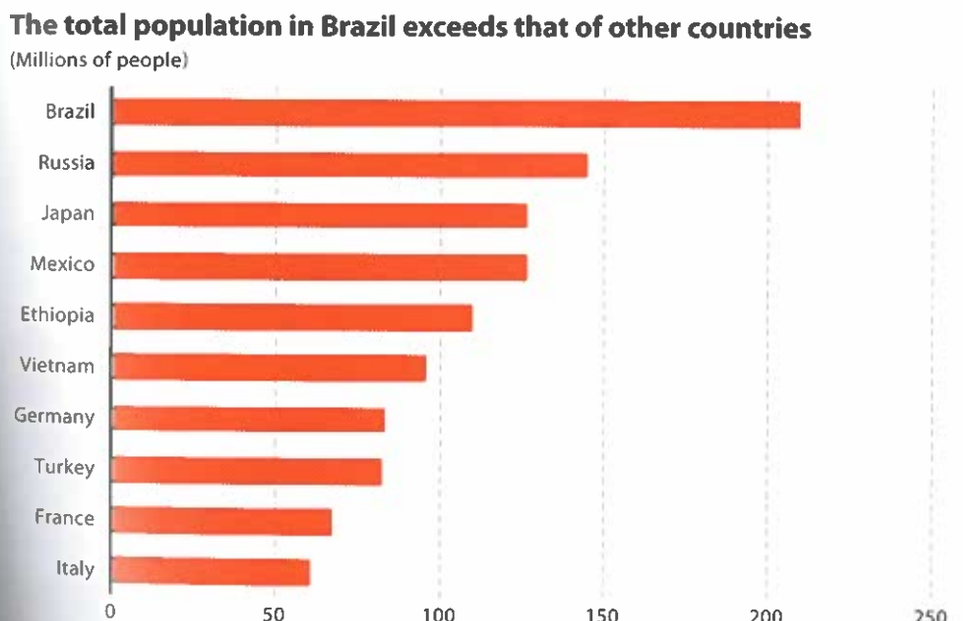
Data labels make gridlines redundant and, by extension, the vertical axis.

The graph may look too cluttered with labels if I had fifty or maybe even twenty countries, so I might include a separate table or an appendix. Deleting the gridlines when including data labels is primarily an aesthetic choice and as you continue to work with data and make your own graphs, you will develop your own style for these graphic elements.

ROTATE LONG AXIS LABELS

The default solution for long horizontal axis labels is to run the text vertically, as on the spine of a book. But this approach forces your reader to turn their head to the side. One solution is to rotate them 45 degrees, but the reader still has to turn their head. Another approach is to shrink the font size so they are aligned horizontally—though this usually makes them too small.

The most elegant solution is to simply rotate the entire graph. This still uses the same pre-attentive attribute—the length of the bars—but the axis labels are now aligned horizontally; they are easy to read with no effect on data comprehension.



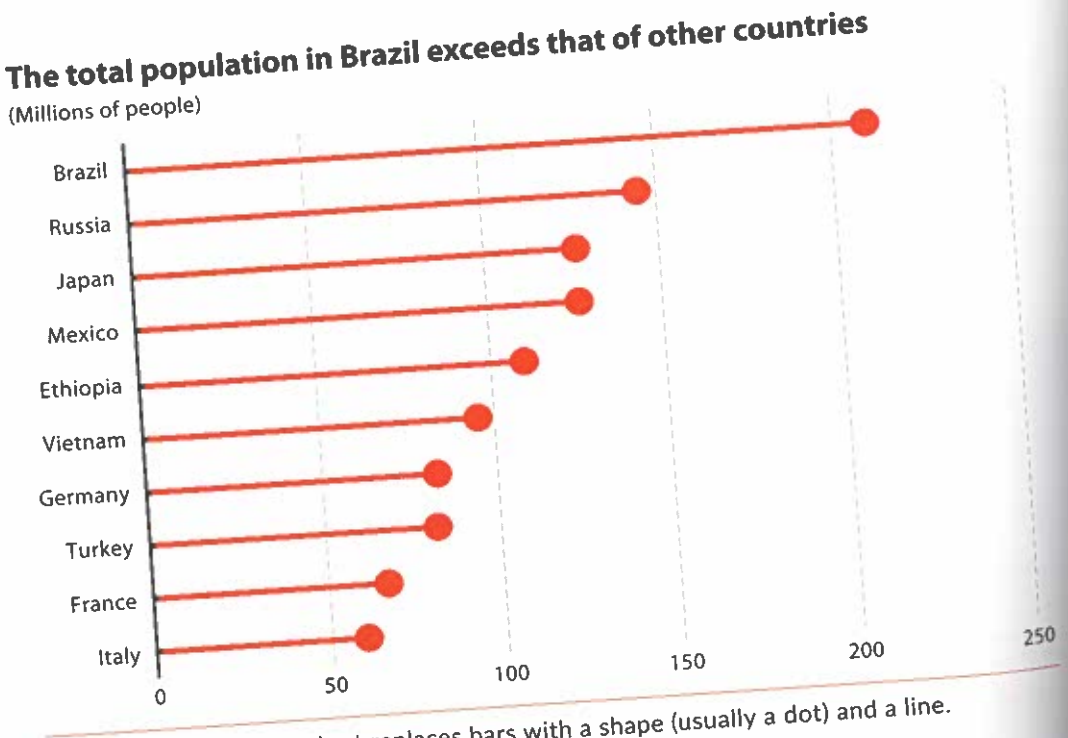
With long axis labels, consider rotating the chart to make the labels horizontal and easier to read.

Data Source: The World Bank.

VARIATIONS ON THE BAR CHART

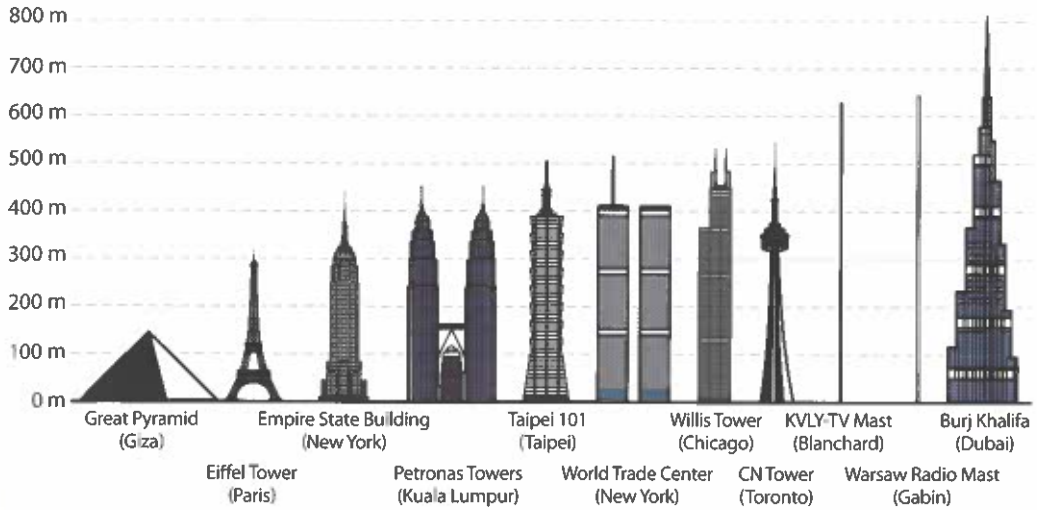
There are countless ways to modify the standard bar chart. One simple variation is to use other shapes in lieu of bars. The lollipop chart, for example, replaces the bar with a line and a dot at the end. This version lives a hair below the bar chart on our perceptual rankings, because it's not exactly clear which part of the circle encodes the value. But it removes a lot of ink from the page and gives you more white space to add labels or other annotation.

This is just one example of an alternative shape. Triangles, squares, and arrows are other options, as are bar-shaped images that reinforce your data. A chart showing data on urban growth may use building-shaped bars, and a chart on climate change may use trees for bars. Be careful with this approach, however, as readers may confuse the total *area* of the icons as a value indicator rather than just the height.



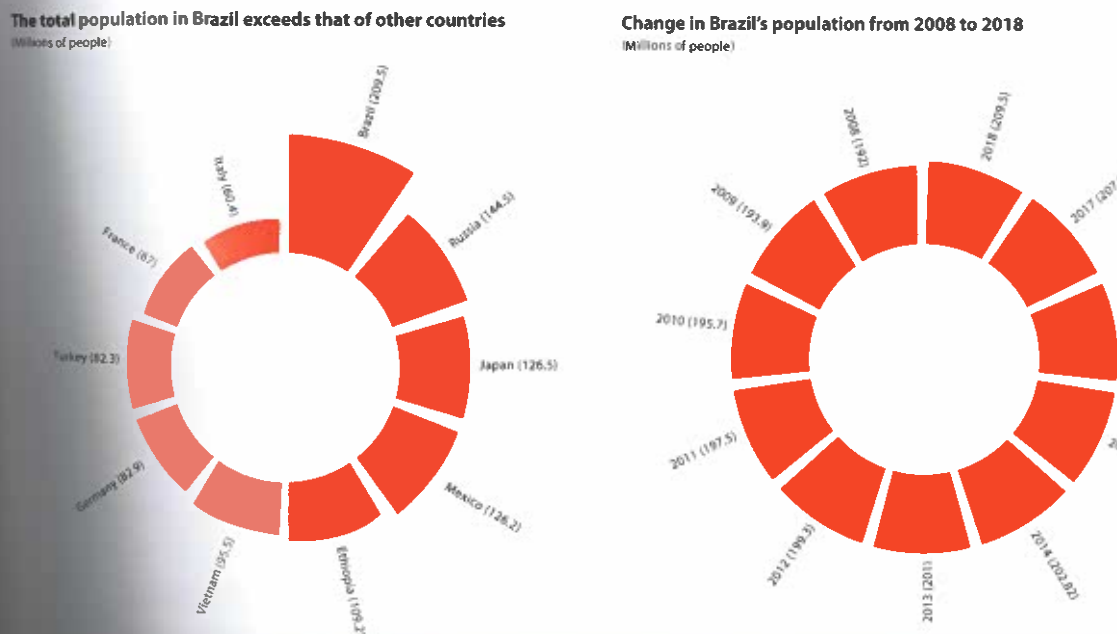
The lollipop chart replaces bars with a shape (usually a dot) and a line.

Data Source: The World Bank.



Alternative shapes, like buildings or people, can be used in lieu of the basic bar shape.

Source: Based on Wikimedia user BurjKhalifaHeight Petronas Towers



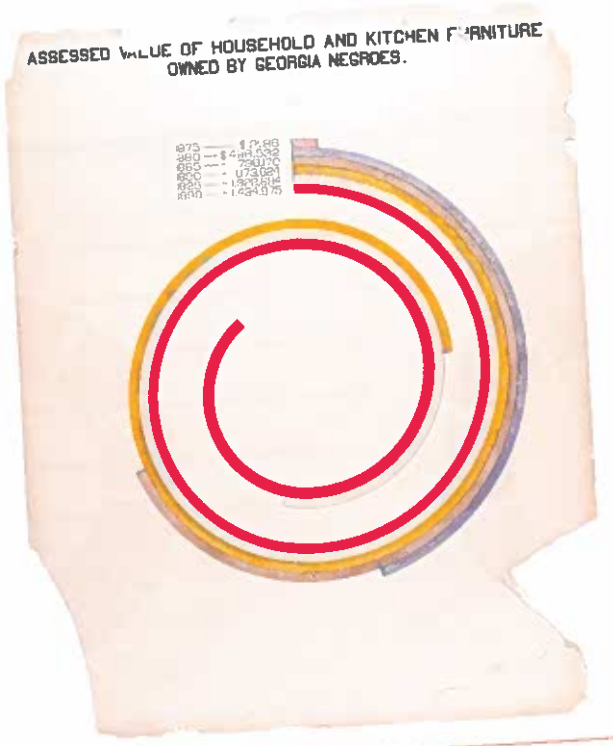
A radial bar chart wraps the standard bar chart around a circle. This chart type moves down the perceptual ranking list because it is harder to compare the heights of the bars.

Data Source: The World Bank.

Another approach to the basic bar chart is to abandon the usual grid and instead place the bars in a circle, called a radial layout. There are two common ways to do this: the radial bar chart and the circular bar chart.

The radial bar chart, also called the polar bar chart, arranges the bars to radiate outward from the center of a circle. This graph lies lowers on the perceptual ranking list because it is harder to compare the heights of the bars arranged around a circle than when they are arranged along a single flat axis. But this layout does allow you to fit more values in a compact space, and makes the radial bar chart well-suited for showing more data, frequent changes (such as monthly or daily), or changes over a long period of time.

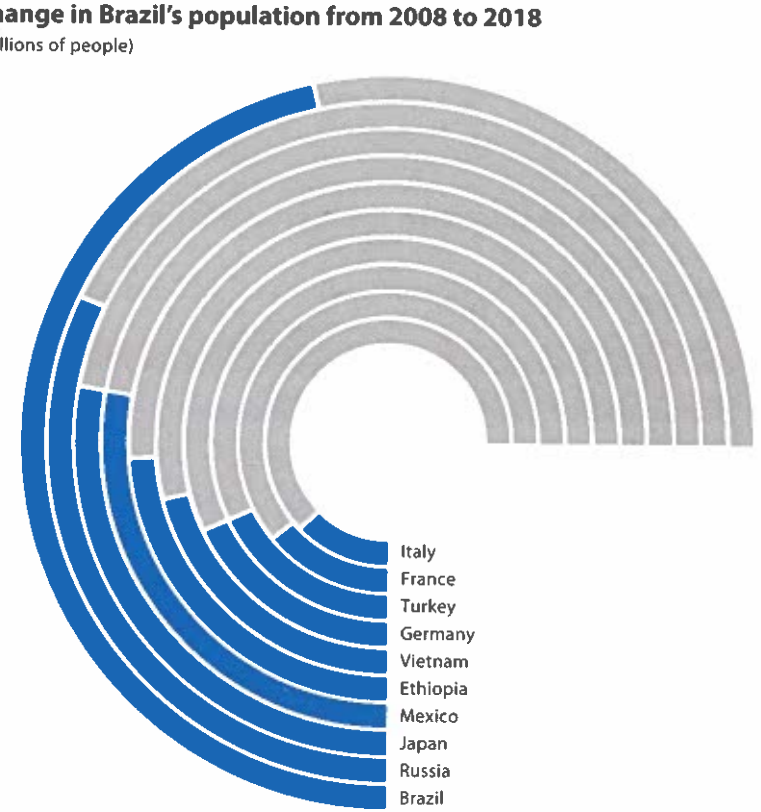
W. E. B. Du Bois used a circular bar chart in his famous *Exposition des Negres d'Ameriique* at the 1900 Paris Exposition. He included this radial bar chart in his set of infographics for



Source: W. E. B. Du Bois, Assessed Value of Household and Kitchen Furniture Owned by Georgia Negroes (1900) via Library of Congress Prints and Photographs Division.

The Georgia Negro: A Social Study, which shows the dollar value of household and kitchen furniture held by African Americans in Georgia in six years (1875, 1880, 1885, 1890, 1895, and 1899). "The end result," wrote Whitney Battle-Baptiste and Britt Rusert in their book about Du Bois's graphics, "is simultaneously easy to read and hypnotic."

Perceptually speaking, the circular bar chart is problematic because it distorts our perception of the data—in this case, the lengths of the bars don't correspond to their actual value. Consider the case where the values of two bars are the same—the ends of the bars will line up in the same position, but the lengths of the bars are not actually the same because they lie along the circumference of two different circles. Author and data visualization expert Andy



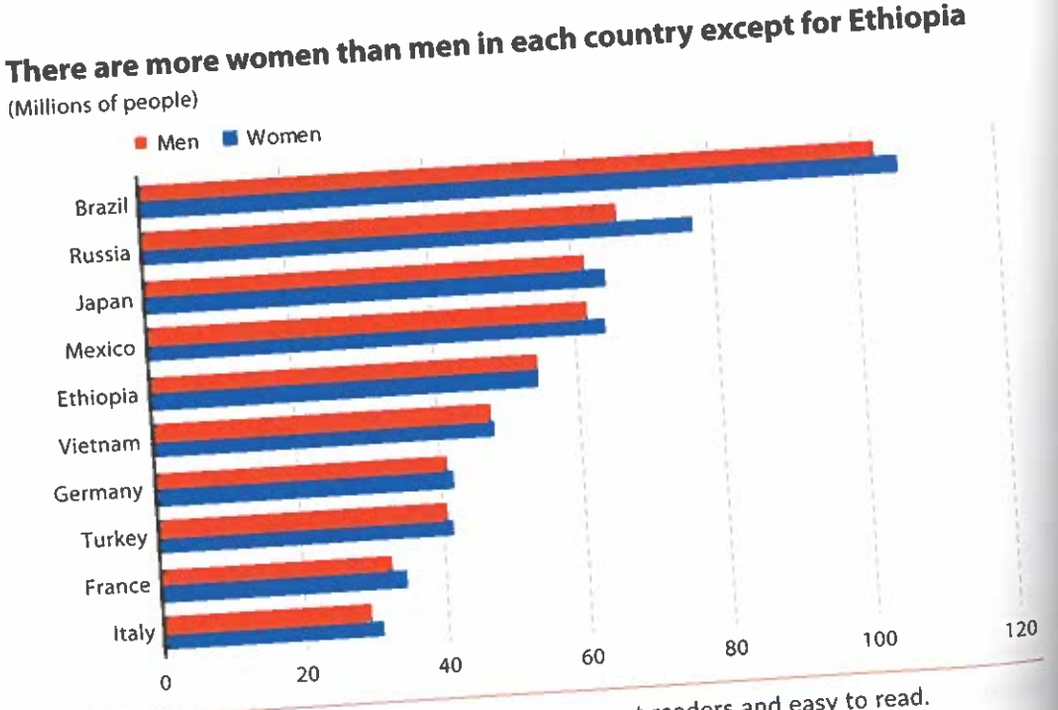
Perceptually speaking, the circular bar graph is problematic because it distorts our perception of the data. In this case, the lengths of the bars don't correspond to their actual value.

Data Source: The World Bank.

Kirk uses an Olympic footrace as a metaphor. Runners start at staggered positions on the track, but they all end up running the same distance because the runner on the outside lane has more distance to cover than the runner on the inside lane. Here, the visualization doesn't move down the perceptual ranking, but off of it altogether because it distorts the data and for that reason, I recommend avoiding them altogether.

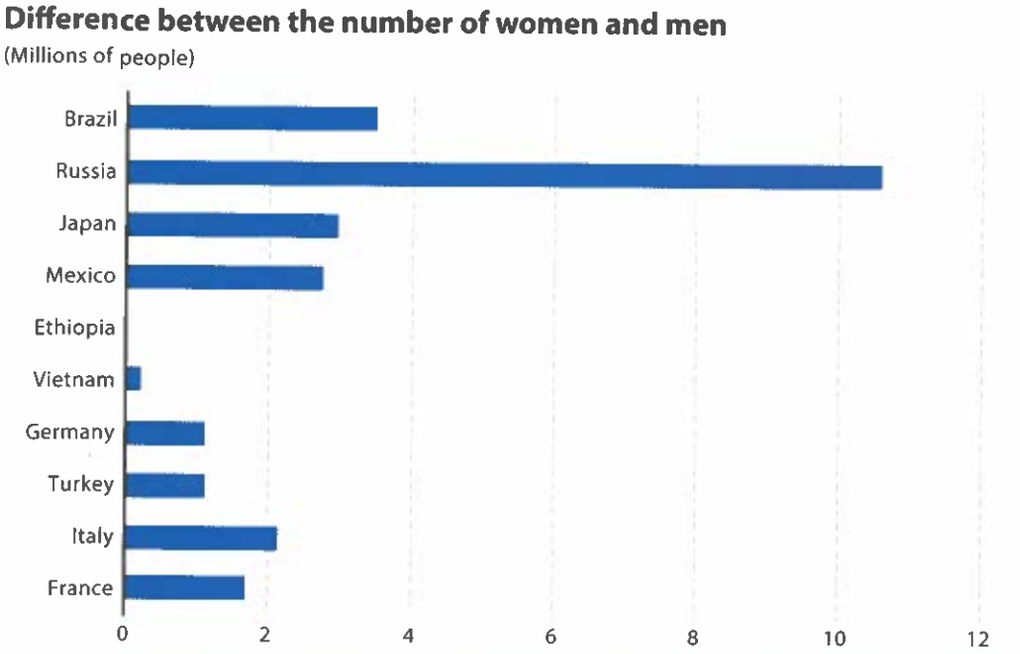
PAIRED BAR

A simple bar chart is perfect for making comparisons across categories, like comparing populations across countries. If I want to show comparisons not just across but also *within* countries, the paired bar chart is a good option. The paired bar chart will be familiar to most readers and is easy to read, and the shared baseline makes it easy to make comparisons.



A simple paired bar chart is familiar to most readers and easy to read.

Data Source: The World Bank.



Instead of showing both data values, we could show the difference between them.

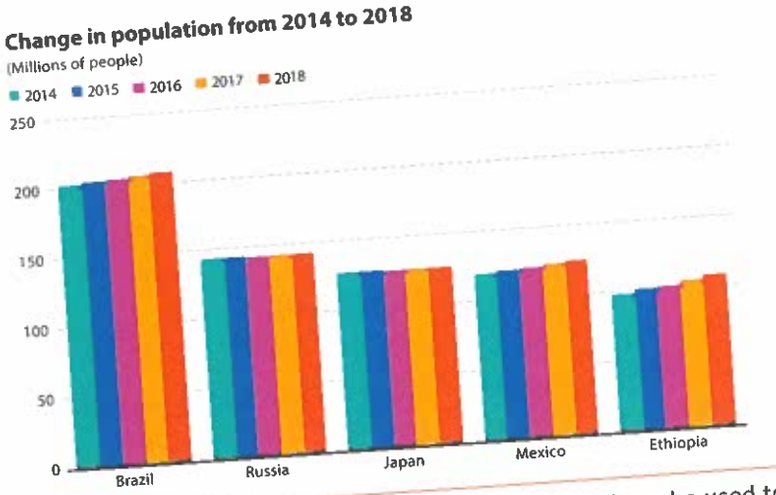
Data Source: The World Bank.

Say we want to show the number of men and women in each country in our sample. A paired bar chart allows us to do so.

Note that the paired bar chart directs the reader's attention not just to the levels, but also to the *difference*. If it's important that readers see both, this is a good option.

But if our goal is for the reader to focus only on the *difference* between the two values within each category, this isn't the most direct way to do so, because we are asking them to compare the difference in lengths. Instead, we could just show the difference between the two values in a single bar, like the one above.

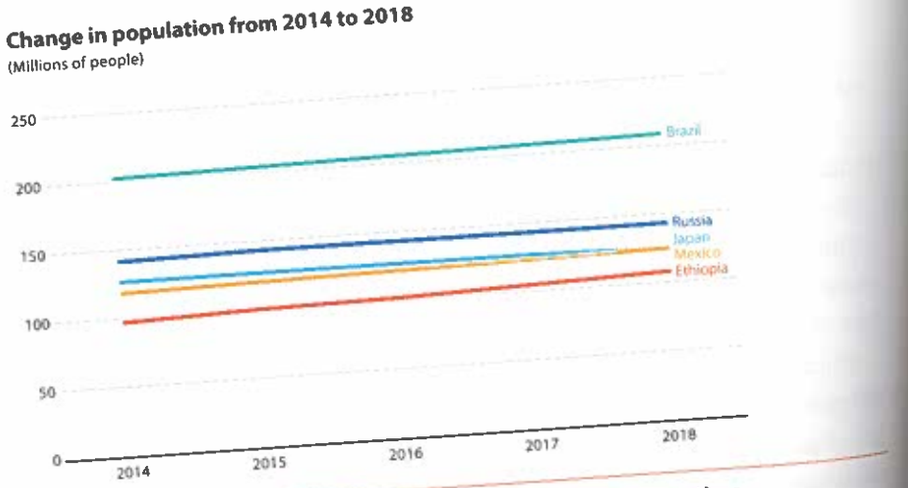
In cases where you want the reader to see the level *and* the difference, you may need a different chart entirely. I prefer the parallel coordinates plot (see page 263), the slope chart (for data that vary over time; see page 150), and the dot plot (see page 97). Remember to ask yourself, What is the goal of this graph? That question will guide you to the best way to visualize your data.



The paired bar chart can be used to show changes over time and can be used to examine changes within and between countries.

Data Source: The World Bank.

Another use for the paired bar chart is to show changes over time. And although I include the word *pair* in the title, these charts can have more than two values. The chart below, for example, shows the population in five of our countries from 2014 to 2018. This



The line chart is a more familiar way to show changes over time.

Data Source: The World Bank.

allows the reader to examine the population change *within* countries and the differences *across* countries.

The patterns in your data can also drive your chart type selection. If the values decline evenly across the different years for all categories, a paired bar chart may look fine. But if the values move around over time, a line chart (as shown earlier) or cycle chart (see Chapter 5) may make for better comparisons over time within and across each group.

There are two instances when I prefer to use bar charts rather than line charts to show changes over time. First, when there are few data points—for example, only five years—the extra ink in the five bars gives the graph more visual weight. Second, when I have discrete time intervals (and few observations), such as the first quarter of the year.

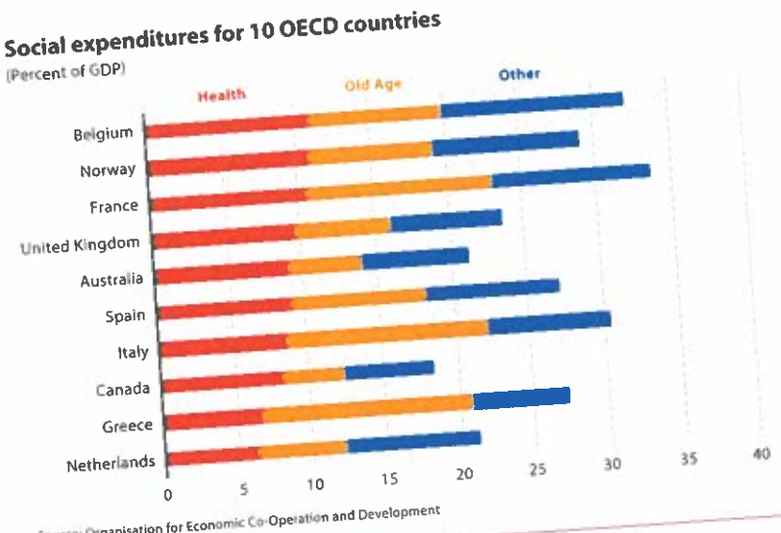
Clutter is the main issue to keep in mind when assessing whether a paired bar chart is the right approach. With too many bars, and especially when there are more than two bars for each category, it can be difficult for the reader to see the patterns and determine whether the most important comparison is between or within the different categories.

When it comes to whether a paired bar chart is too cluttered, trust your eyes and your instincts. Put yourself in your readers' shoes—try to imagine where their eyes will go when they look at the graph for the first time. If there's too much going on, you may need to break up your data, use a different chart type, or try a small multiples approach.

STACKED BAR

Another variation on the bar chart is the stacked bar chart. While the paired bar chart shows two or more data values for each category, this chart subdivides the data within each category. The categories could sum to the same total, say, 100 percent, so that the total length of the bar is the same for every group. Or the totals may differ across the groups, in which case the total length of each bar may differ. Above, I've plotted the share of gross domestic product (GDP) each of ten countries spends on support for health care, old age, and other programs. The entire length of the bars shows how much each country spends on these programs as a share of GDP.

As with the bar charts we've looked at thus far, the stacked bar chart is familiar, easy to read, and easy to create. The biggest challenge, however, is that it can be difficult to compare

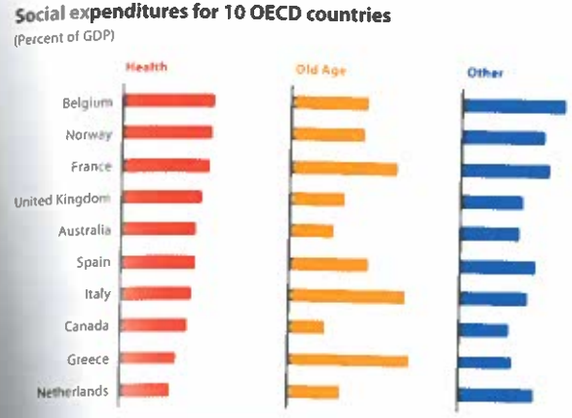


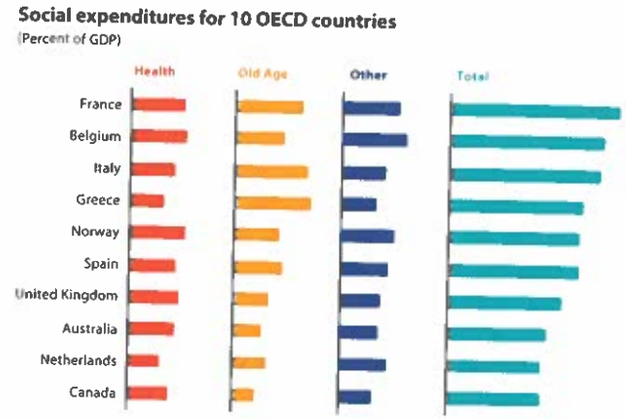
The stacked bar charts shows how different categories sum to a total. The interior series in the chart, however, are harder to compare with one another because they do not sit on the same baseline.

Data Source: Organisation for Economic Co-Operation and Development.

the different values of the segments *within* the chart. In the example above, it's easy to compare the values across the countries for the Health category, because the bar segments share the same vertical baseline. But that's harder to do with the two other series because they do not share a baseline. Which country spends more on old-age programs, Italy or Greece? You can quickly see that Italy spends more on health programs than Greece, because those segments are left-aligned on the vertical axis, but it's much harder to determine with the segments for the other categories.

One way to address the changing baseline is to break the graph apart so that each series sits on its own vertical baseline. This is a small multiples graph, arranged side by side. It's now easier to see that Greece spends more on Old Age programs than Italy. The tradeoff is that it is harder (if not impossible) to see the *total* values. But that too can be overcome: You can still break up the stacked graph and add a final segment that represents the total amount (this is not an issue when all of the series sum to 100 percent because the summed segments will all have the same length).

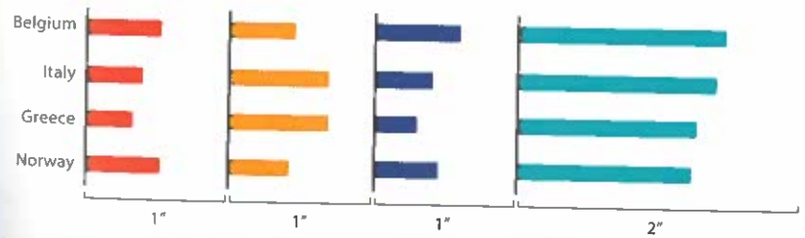




Instead of stacking all the data series together, we can break them up (either with or without the totals) to create a sort of small-multiples approach. Here, we move up to the top of the perceptual ranking list because each series sits on its own baseline.

Data Source: Organisation for Economic Co-Operation and Development.

In both versions, the horizontal spacing for each segment should be the same width, otherwise it might appear that a segment takes up a larger proportion of the space than it really does. In cases where you add the total, the width does not need to be the same as the other groups, but the increments along the axis should be the same. In other words, if the width of each segment above in which the data range from 0 percent to 50 percent is one inch wide, the total category that spans 0 percent to 100 percent should be two inches wide.

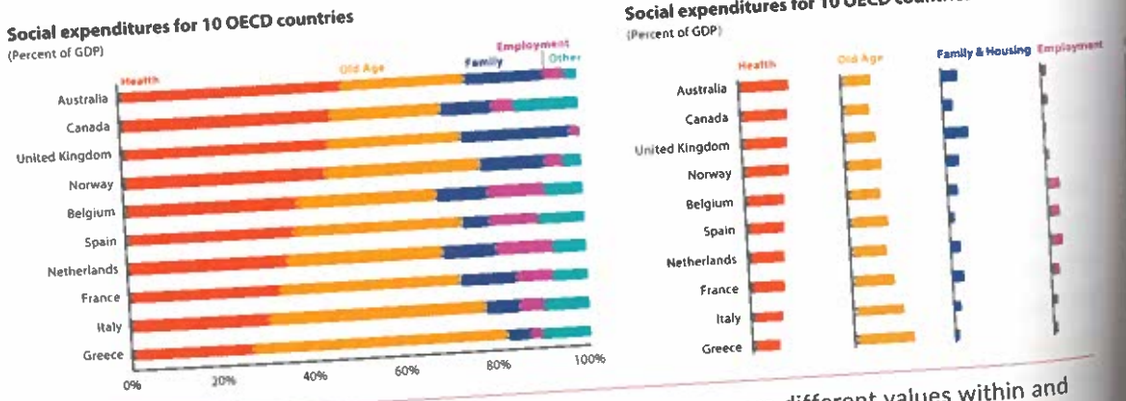


When creating these sorts of small multiples bar charts, be sure that each segment has the same width.

Even though different baselines in the standard stacked bar chart can make it more difficult to compare values, there are cases when the stacked bar chart is preferable. In this stacked bar chart, I've included more spending categories and divided them into shares of the total so the graph highlights the *distribution*. In this view, it becomes clear that around three-quarters of total government spending in these countries goes to programs for health care and old age programs. That observation is harder to see in the version on the right, where each category is placed on its own vertical baseline. Even though it is easier to compare differences in each category across countries, you don't see large differences between them.

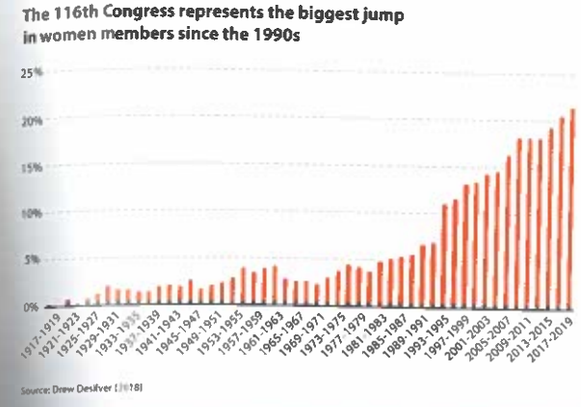
As always, identify what you want to show and where you'd like to focus your reader's attention. In these examples, the *Health* category is emphasized because the data are sorted according to those values (shown as a percent of total spending on these specific programs) and it is situated along the vertical baseline. In this layout, the other segments become secondary in comparison to health spending.

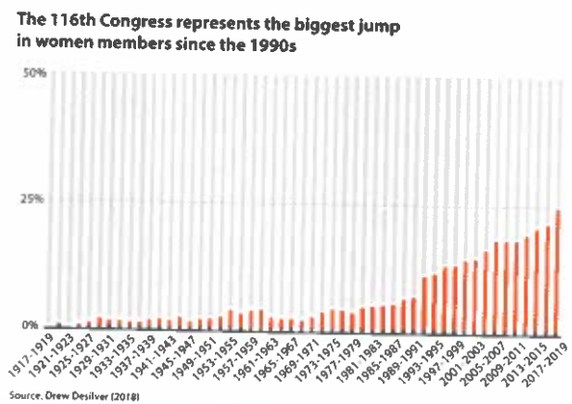
There is one other stacked bar chart that you may have come across that shows a single set of data values and the gap between them and another value (often the total). The graph on the next page uses this approach to show the share of women elected to the U.S. House of Representatives from 1917 to 2018. The version on the left shows the raw percentages; the vertical axis ranges from 0 to 25 percent. Here, you see a dramatic increase in the share of



In these examples, we can see how our ability to compare different values within and across countries varies between these two views.

Data Source: Organisation for Economic Co-Operation and Development.





A somewhat rare case where a stacked bar chart is used to focus attention on one series. This technique may be particularly valuable where the relative proportion is as important as the change.

women in Congress. The version on the right shows the same data, but stacks a gray series on top of the data values to 50 percent. In this version, we can emphasize that the share of women is still small even though that share is rising. It is in these cases—where the relative proportion is as important as the change—that this technique may be particularly valuable.

PERCENT CHANGE VS. PERCENTAGE POINT CHANGE

There is an important distinction between *percent change* and *percentage point change*, and it's a mistake that many often make.

Percent change compares an initial value *OLD* to a final value *NEW* according to this simple formula:

((NEW-OLD)/OLD) × 100.

Positive percent changes (that is, *NEW>OLD*) mean there is a percent (or percentage) increase. Negative changes (*NEW<OLD*) mean there is a decline. You can calculate differences over time or between groups; all that really matters is that you follow the formula and know that you are comparing the change relative to the initial value of *OLD*.

Now, *percentage point change* is specific to looking at raw differences in percentages. The *percentage point change* is a simpler formula:

NEW-OLD

where both are already percentages.

These are very different things. Let's take a simple example. According to the U.S. Census Bureau, there were 40.6 million people in poverty in 2016 and 39.6 million people in poverty in 2017. The poverty rate (the number of people in poverty as a percent of the total population) was 12.7 percent in 2016 and 12.3 percent in 2017.

The number of people in poverty fell by 2.3 percent. The *percent change* was

$$[(39{,}698{,}000 - 40{,}616{,}000)/40{,}616{,}000]\times 100 = [-0.023]\times 100 = -2.3\%$$

But the poverty *rate* fell by 0.4 *percentage points* over the two years:

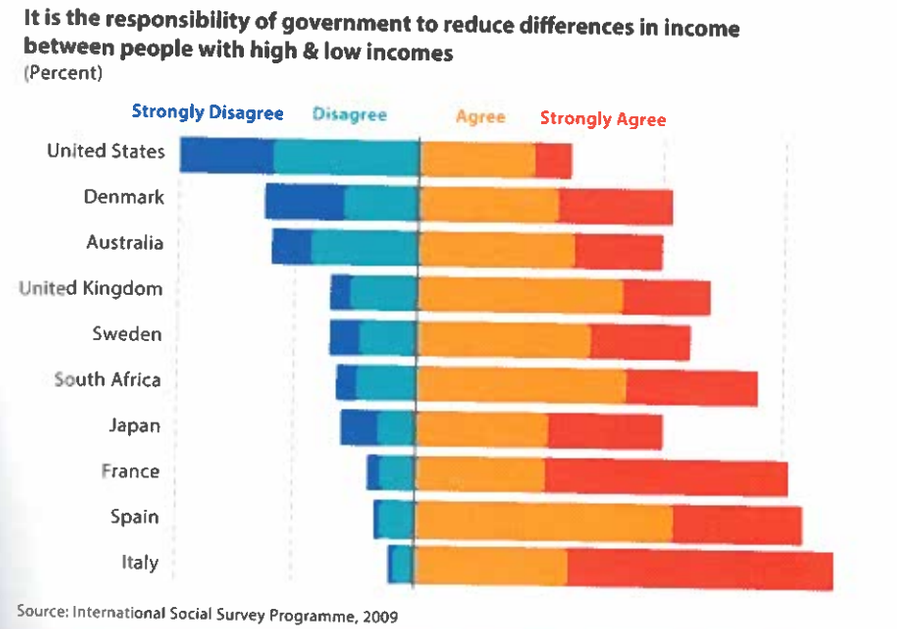
12.3% - 12.7% = -0.4 percentage points

Obviously, those are two very different numbers, but people confuse them all the time. Clearly representing your data starts with clearly understanding your data, how they were collected, and how to calculate basic descriptive statistics.

DIVERGING BAR

A variation on the stacked bar chart is one in which the stacks diverge from a central baseline in opposite directions. These are often found in surveys where the responses are arrayed in ranges from, for instance, *strongly disagree* to *strongly agree*. These are often called "Likert Scales," named after the psychologist Rensis Likert, who invented the scales in the early 1930s.

In this example, drawing on data from the International Social Survey Programme, survey respondents were asked whether they believe it is the government's responsibility to reduce income inequality. By grouping the "disagrees" and the "agrees" together on either side of a central baseline, we can compare the *total* sentiment across the different countries.



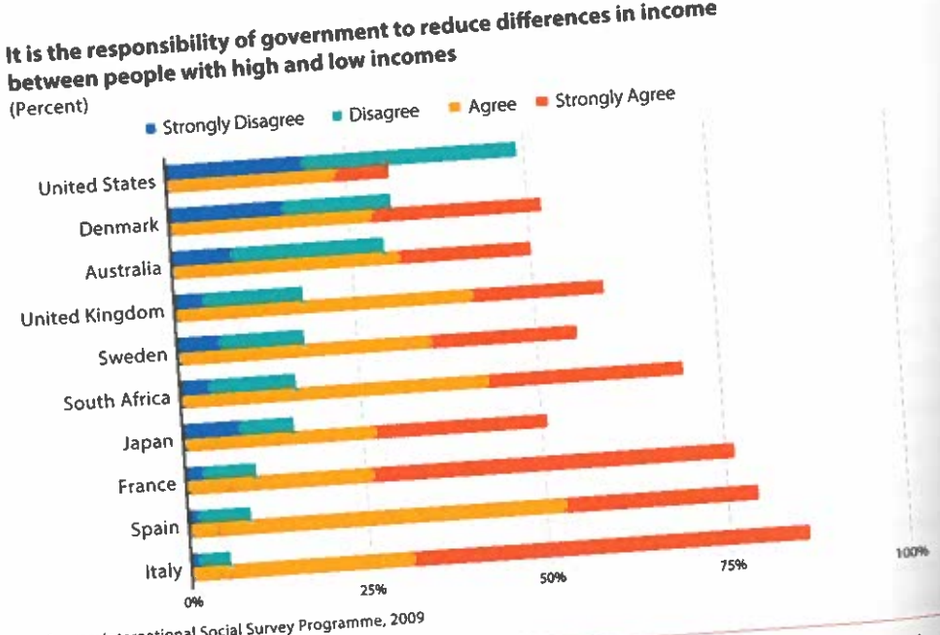
The diverging bar chart can show differences in opposing sentiments or groups, such as "agree/disagree" or "true/false."

One advantage of this chart is that the sentiments are clearly presented—the Disagrees jut out to the left (in what we might typically think of as a negative direction) and the Agrees out to the right. This works well if your audience is most interested in the *total* sentiment of each side and not necessarily comparisons between each individual component. If the individual comparisons are the primary focal point, then a paired bar chart could do the job just as well.

Why do we perceive these values to the left as negative? Throughout western history, the concept of left—and even left-handed people—has been plagued with negative connotations.

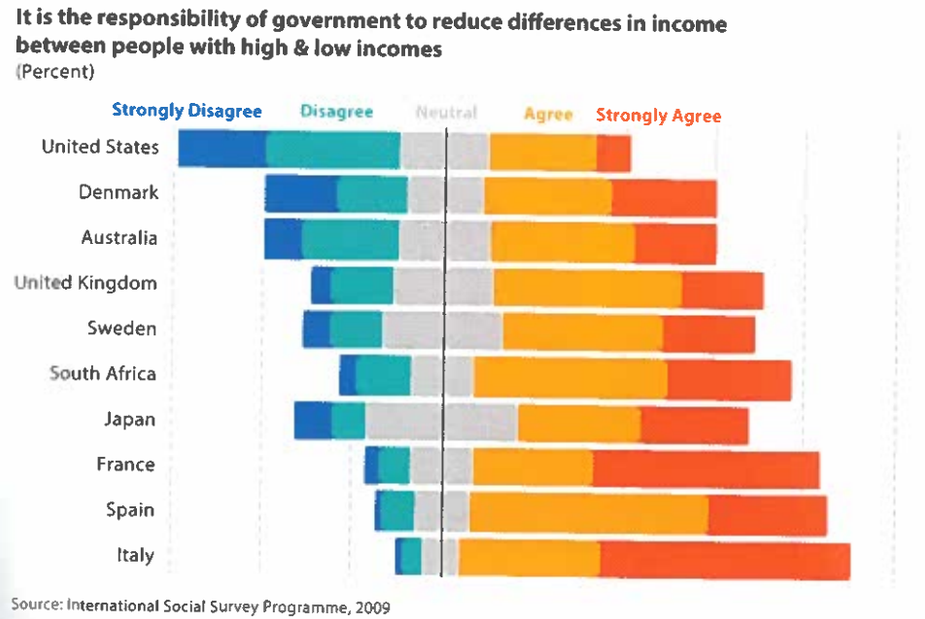
Consider the etymology of the word: *left* is derived from the Old English word *lyft*, which means "weak." In Latin the word *sinister* means the left or left-hand direction. The word *right* comes from the Old English *riht*, whose original meaning was "straight" and thus not bent or crooked. And this is why we have phrases such as "standing upright" or "do the right thing" or "the right answer," all of which connote goodness and correctness. You can also see this in other languages: In Spanish, for example, the word *derecha* means "right" and the closely-derived *derecho* means "straight."

As with the stacked bar chart, the challenge with visualizing these kinds of data is that we are comparing within *and* across the categories. Arranging the bars in opposite directions makes it difficult to compare the totals of the two groups. In other words, it's difficult to compare the *total share* of people who disagree with the *total share* of people who agree. That task is slightly easier in the paired bar chart, but then you lose the positive-negative connotation of the diverging chart. Depending on the patterns in your data and the number of categories and groups, you might find this chart looks cluttered and busy.



Taking the opposite sides of the diverging bar chart and placing them in a more standard paired bar chart approach can also work and allows us to more accurately compare the totals.

You must be especially careful using a diverging bar chart when you have a "neutral" category. By definition, the neutral survey response is neither agree nor disagree, and should therefore be grouped with neither category.

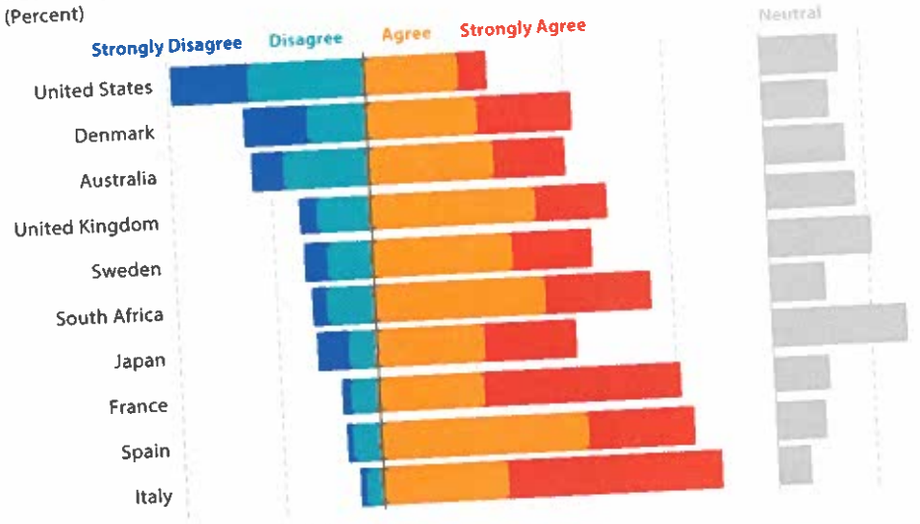


Placing the *Neutral* category of a diverging bar chart in the middle wrongly implies that the neutral responses are split between the two sentiments.

Placing the neutral category in the middle of the chart along the vertical baseline creates a misalignment between the two groups and implies the neutral responses are split between the two sentiments. It also means that none of the segments sit on a vertical baseline. Placing it to the side of the chart is a better strategy because the disagree, agree, and neutral categories now all sit on their own vertical axes, even though the neutral category is somewhat emphasized as it sits to the side (see next page).

Another alternative—regardless of whether you have a neutral category—is the stacked bar chart as shown on the next page. In this view, the different categories sum to 100 percent, and one can more easily compare the totals between the countries. A good strategy is to mark specific aggregate values to guide the reader. Here, for example, I have marked the 50-percent position to make it clear for which countries the total "agree" and "disagree" sentiments are at least half of the total.

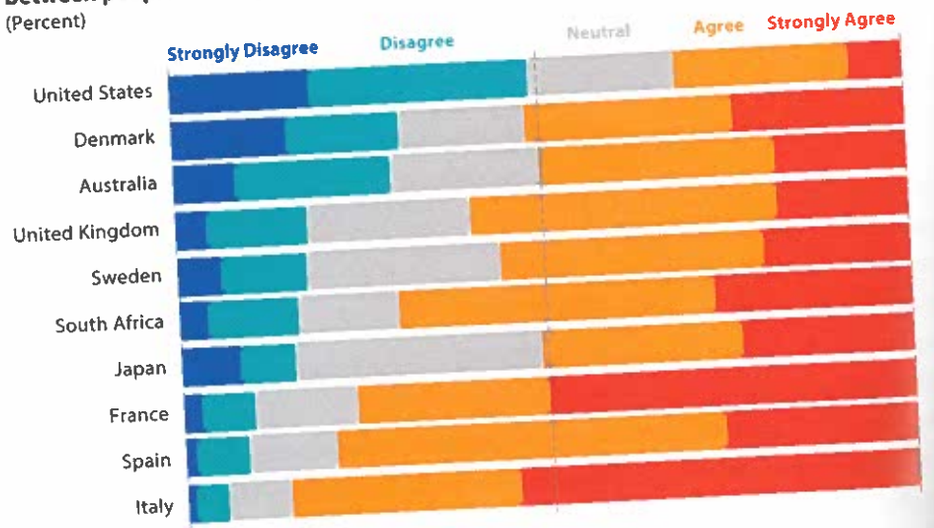
It is the responsibility of government to reduce differences in income between people with high & low incomes
(Percent)



Source: International Social Survey Programme, 2009

A better approach is to place the *Neutral* category off to the side of the graph.

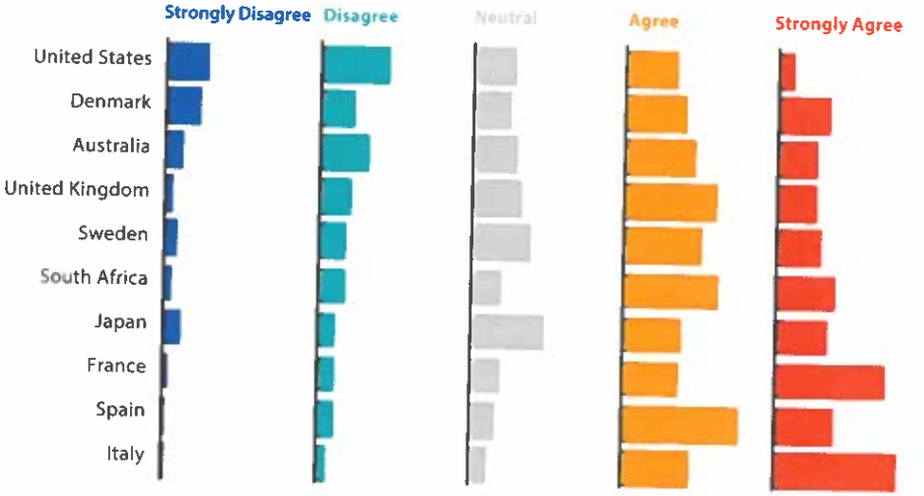
It is the responsibility of government to reduce differences in income between people with high & low incomes
(Percent)



Source: International Social Survey Programme, 2009

A stacked bar chart can be used to show these kinds of Likert scales.

It is the responsibility of government to reduce differences in income between people with high & low incomes
(Percent)



Source: International Social Survey Programme, 2009

The small multiples bar chart is yet another way to visualize these kinds of data.

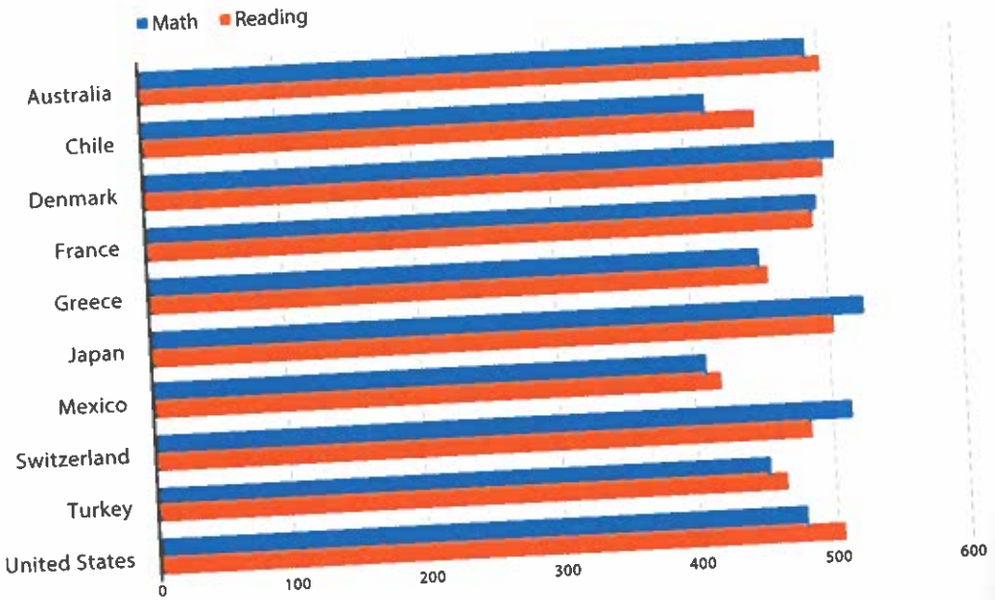
You could take this further and break the chart into its components, as we discussed in the previous section. In general, as with many graphs, which variation you choose will depend on your goals.

DOT PLOT

The dot plot (sometimes called a dumbbell chart, barbell chart, or gap chart) is one of my favorite alternatives to a paired or stacked bar chart. Developed by William Cleveland, one of the early pioneers in data visualization research, the dot plot uses a symbol—often but not always a circle—corresponding to the data value, connected by a line or arrow. The data values correspond to one axis and the groups to the other, which do not necessarily need to be ordered in a specific way, though sorting can help.

The dot plot is an easy way to compare categories—especially many categories—when bars might add too much ink and clutter to the page. For this example, let's look at scholastic test

PISA scores for math and reading among 10 OECD countries



Source: Programme for International Student Assessment

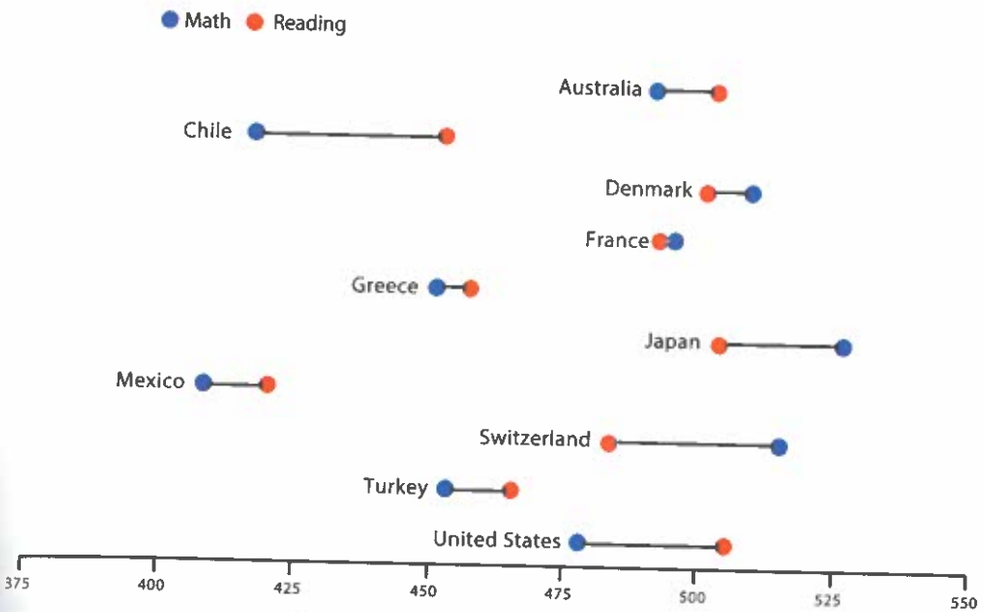
This simple bar chart shows math and reading scores across multiple countries. Bar charts are often the default visual for these kind of data, but it looks heavy and dense.

scores around the world from the Programme for International Student Assessment (PISA), an international set of achievement tests taken by fifteen-year-old students in reading, mathematics, and science. We can easily plot the mathematics and reading scores for a set of countries using a simple bar chart, but the twenty bars makes the chart heavy and dense.

By contrast, the dot plot shows the same data with a dot at each data value connected by a line to show the range or difference. The circles use less ink than the bars, which lightens the visual with more empty space. The country labels are placed close to the leftmost dot, though they could also be set off to the left along the vertical axis. If necessary, data values can be placed next to, above, or within each circle.

Dot plots are not restricted to two dots and a connecting line, nor are they restricted to simply comparing different categories. You can use dot plots to show a change between two years, for example. You could use different shapes or icons or arrows instead of lines to denote direction. You can also use more than two objects. For example, we could add science test scores to this plot, but we need to be sure to add sufficient labeling so our reader

PISA scores for math and reading among 10 OECD countries



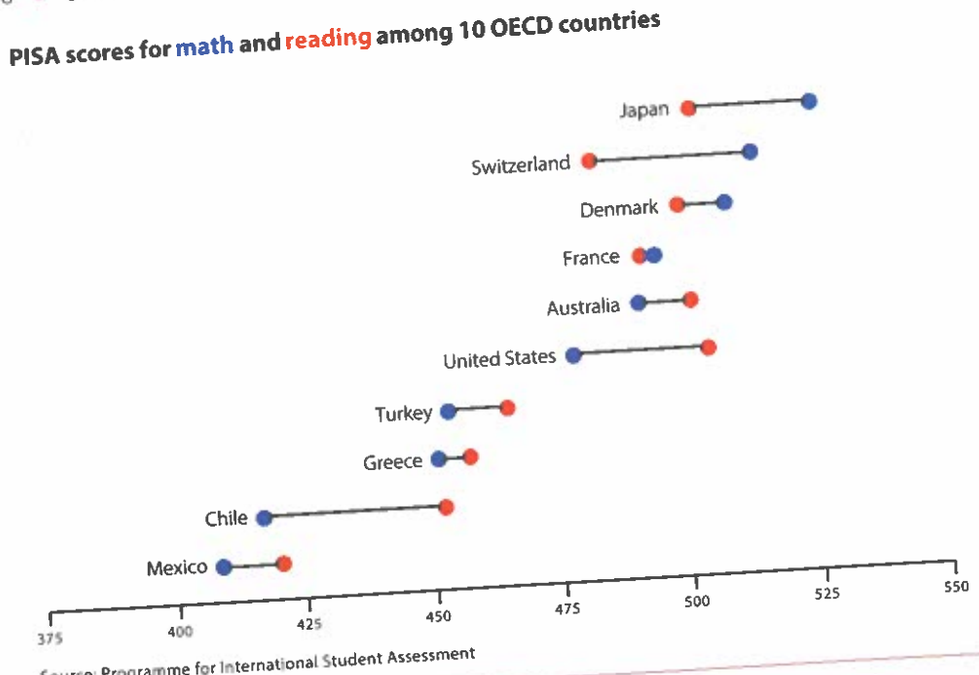
Source: Programme for International Student Assessment

The basic dot plot places a dot for each data point and connects them with a line. Notice how more white space lightens the visualization.

knows what each object on the graph represents. Axes and gridlines can be included or not, depending on how important it is for the reader to determine the exact values.

A few points of caution about the dot pot. First, it's not entirely obvious when the direction of the values change, as in the last chart. Did you notice that math scores were higher than reading scores in four of the countries in the dot plot above? That difference is not immediately evident unless the reader carefully examines the points and their coloring. In this and other cases, we should consider how sufficient annotation, clear labeling, and highlighting colors can help clarify different directions. The data are sorted by math scores in the dot plot on the next page, which helps organize the countries, but it is still not immediately clear that in only the first four countries are math scores higher than reading scores.

One approach is to split the graph into two groups, one for countries in which math scores are higher than reading scores and another for countries where the opposite occurred. In these versions (page 100), the groups are split and then sorted with larger, bold headers to distinguish them. We can also add data values—I will sometimes put them right inside the

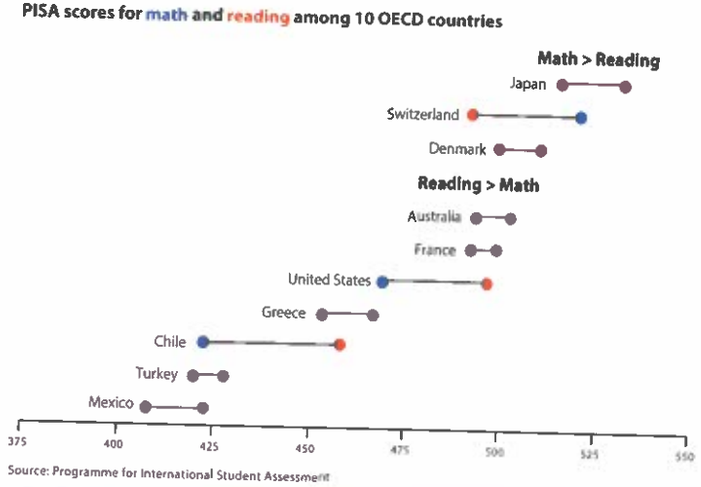


As with the basic bar chart, sorting the data in a dot plot helps organize the space for the reader.

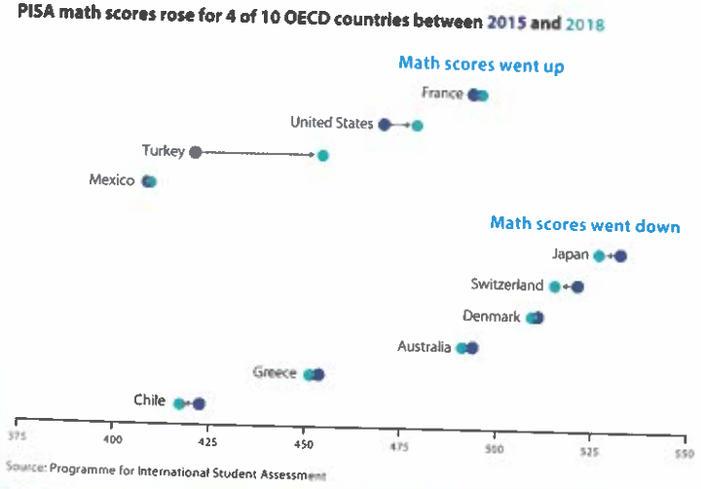
circle—but be careful because the labels can clutter the chart. An alternative is to include vertical gridlines, depending on how precisely we want to communicate the data to the reader.

When using a dot plot to show change over time, I prefer to change the linking lines to arrows, which helps make the direction clear.

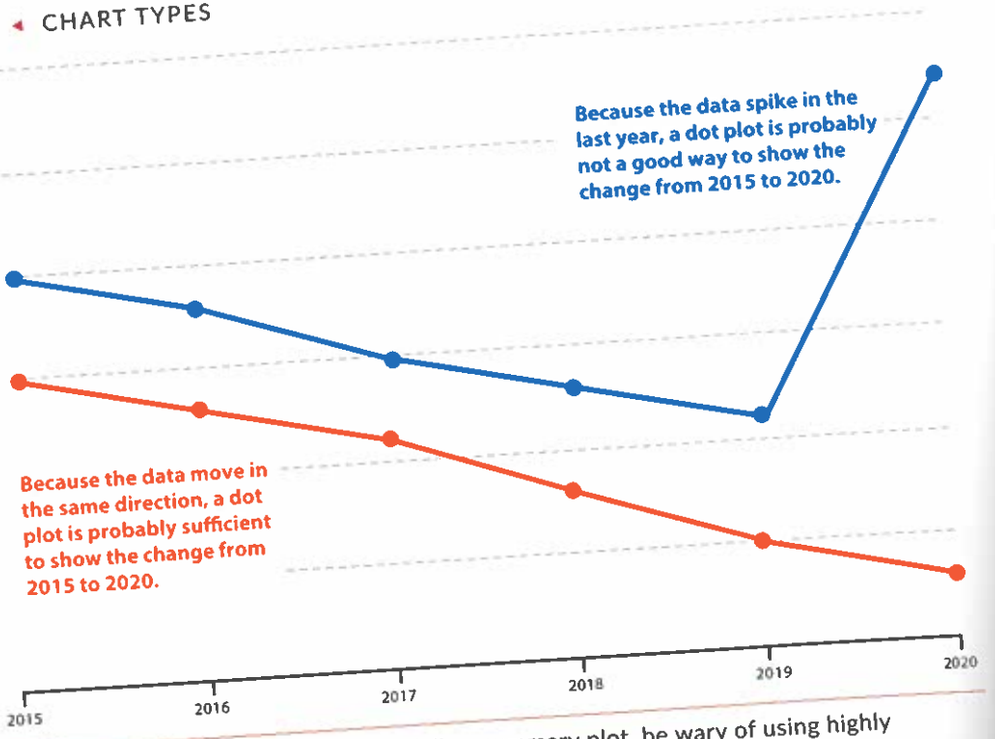
Another word of caution for dot plots that show changes over time. The dot plot is, by definition, a summary chart. It does not show all of the data in the intervening years. If the data between the two dots generally move in the same direction, a dot plot is sufficient. But if the data contain sharp variations year by year, a dot plot will obscure that pattern (as it also does for bar charts). For example, if test scores decreased between 2015 and 2019 and then increased sharply between 2019 and 2020, the dot plot would only show an overall increase, masking the change in the intervening years. In some cases, you may not have a choice—if you are using data from the decennial U.S. Census, by definition you will only have data for every ten years. That's something you can't help, but if you are familiar enough with your content, you'll know whether showing only those points is enough to clearly and accurately make your point.



Labels and annotation can clarify differences in the relationships between the values. Gridlines are not always necessary.



Dot plots can show changes over time. In these cases, I will often make the linking line an arrow to suggest the change over time.



Because dot plots are essentially a summary plot, be wary of using highly variable data with dot plots.

MARIMEKKO AND MOSAIC CHARTS

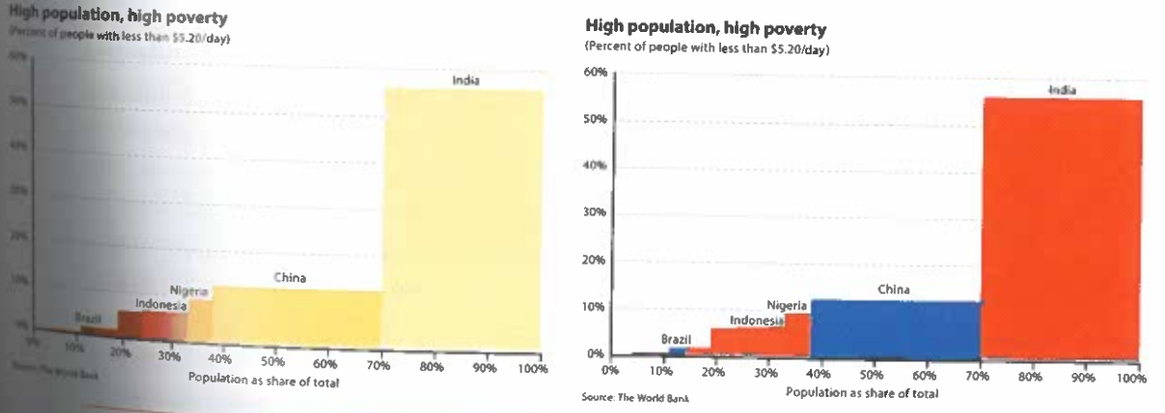
Marimekko charts may look odd at first, but they are just an extension of the bar chart. This type of chart is useful when you want to make comparisons between two variables: one comparing categories and one showing how they sum to a total. The name of the chart comes from the Finnish design firm Marimekko, founded in 1951 by Armi Ratia and her husband, Viljo. Early Marimekko style featured straight, oversized, geometric patterns and bright colors.

In the standard vertical bar (or column) chart, the data are measured along the height of the vertical axis and the widths of the columns are identical. The Marimekko chart takes that standard column chart and expands the width of each bar according to another data value. The Marimekko chart is an easy way to add a second variable to your standard column or bar chart.

In this Marimekko chart, I show two variables for the ten most populous countries: the share of people with less than $5.20 per day and the share of the total population among these countries. The percent of people with less than $5.20 per day is plotted along the vertical axis

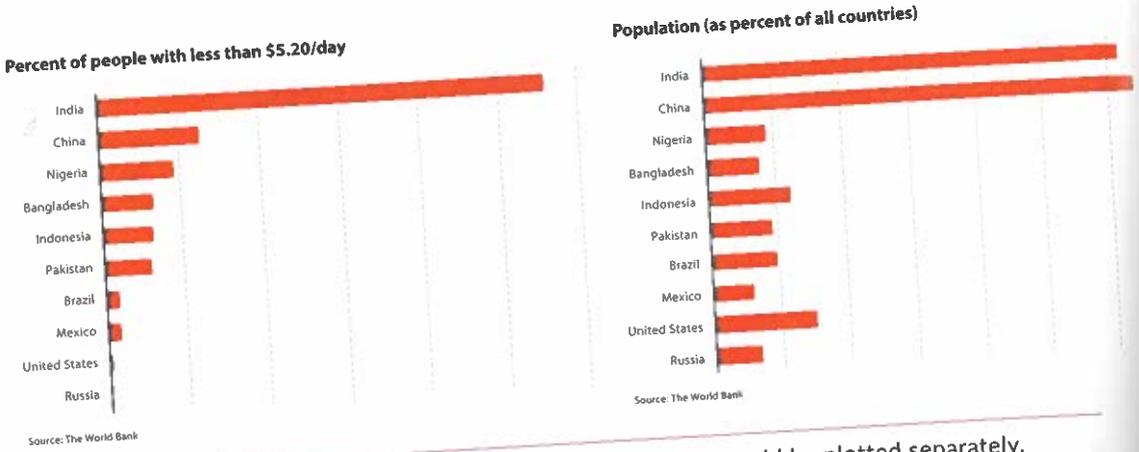
Early Marimekko fabric styles like this one featured straight, oversized, geometric patterns and bright colors.

as in a standard bar chart; the widths of each bar are then scaled according to the share of each country's population summing to 100 percent across these ten countries (an alternative is to show the raw counts rather than percentages). You can see that the most populous countries in this sample—the widest bars—and their distribution of poverty. You can also use color strategically here: if this graph were in an article about poverty in Brazil and China, we might shade all the bars the same color except for those two, as in the graph on the right.



The Marimekko chart (sometimes called a Mekko chart) scales the widths of the bars in a bar chart corresponding to another variable. Color can be used to highlight specific values.

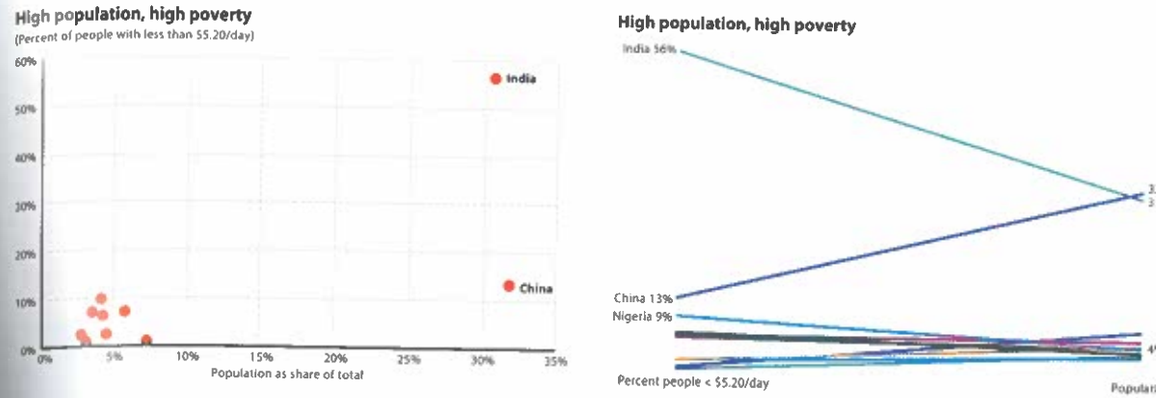
The two variables could also be plotted separately in two bar charts, and while these graphs are familiar and easy to read, they do not communicate the relationship between the two variables as well.



Instead of a Marimekko chart, the two variables could be plotted separately.

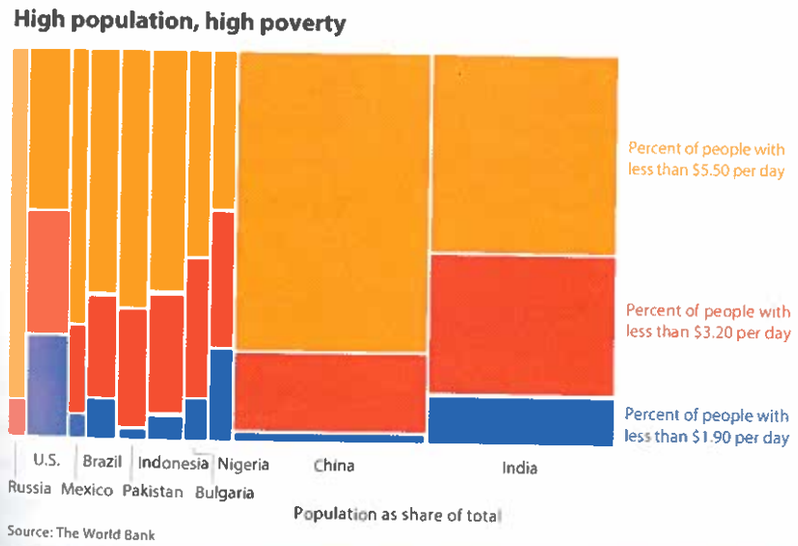
Putting two variables on a chart might get you thinking about the association or relationship between the two variables. If that's the case, you could visualize that relationship with other chart types. I've plotted the same data in this scatterplot (see also page 249). You can see how China and India are outliers, especially along the population axis, but the chart doesn't communicate the part-to-whole picture of population. The parallel coordinates plot on the right (see also page 263) similarly shows how many more people live in China and India, and how a greater share of the population in India lives on $5.20 per day. (One potential issue with the parallel coordinates plot is that the lines might suggest a change over time to some readers when instead, in this case, it is being used to compare the two variables.)

A variation on the Marimekko is to have *both* the heights and the widths of the bars sum to 100 percent. This is sometimes called a mosaic chart, though many people don't differentiate between these two charts and use the terms interchangeably. In this definition of the mosaic chart, you fill the entire graph space and can therefore provide a part-to-whole perspective of the data along both dimensions. In this way, the mosaic chart is also closely related to the treemap (see page 297), but does not necessarily show a hierarchical relationship.

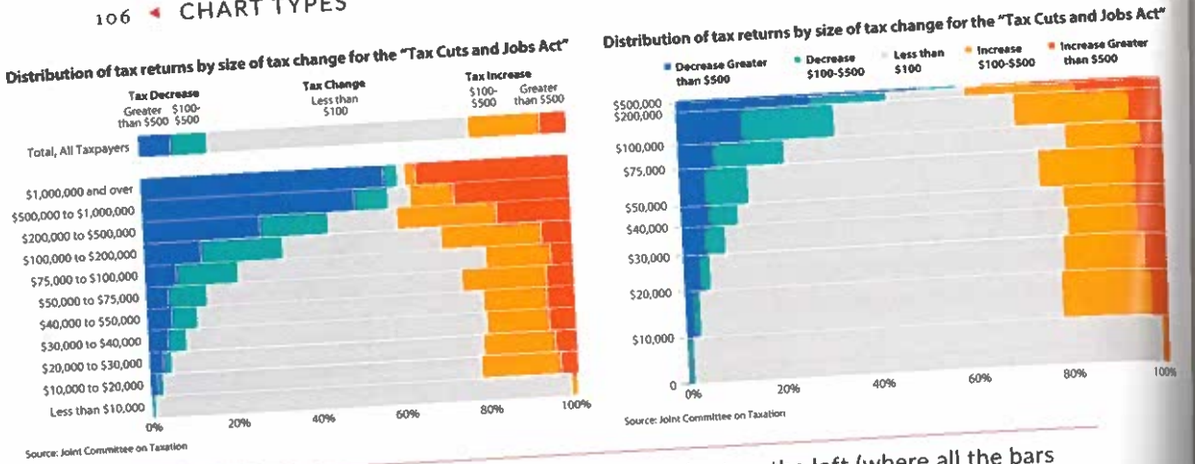


Other ways to plot two data series are a scatterplot (left) or parallel coordinates plot (right). Both are discussed later in this book.

In this example, population is still plotted along the horizontal axis, and the vertical axis now contains three categories for people with low levels of income: share of people with less than $1.90 per day, $3.20 per day, and $5.20 per day.



The mosaic chart is a variation on the Marimekko where both the heights and the widths of the bars sum to 100 percent.



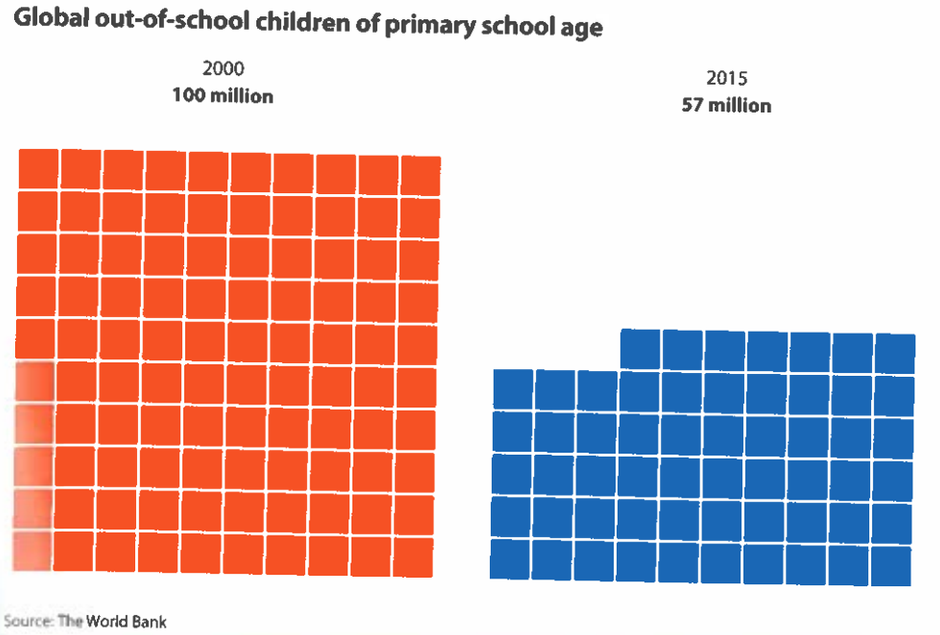
Notice the difference between the stacked bar chart on the left (where all the bars are the same width) and the mosaic chart on the right. While the mosaic chart adds another variable, it is harder to see the details in the top category.

A mosaic chart can also serve as an extension of a stacked bar chart. These graphs show gains from the Tax Cuts and Jobs Act of 2017 for tax units at different points in the United States' income distribution. The stacked bar chart on the left shows five categories of tax gains across eleven income intervals, and each bar shares an equal width. If we scale the widths of the bars to the number of tax units in each income interval—so that the total vertical height of the chart sums to 100 percent—we can create the mosaic chart shown on the right. Notice that the mosaic chart gives a better sense of the distribution of the number of taxpayers in different groups, but because there are relatively few people in the top income group, it's harder to see those values.

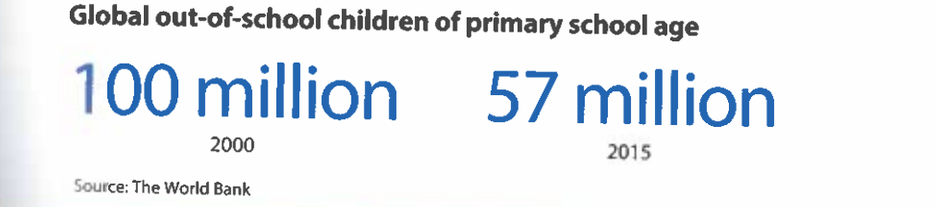
UNIT, ISOTYPE, AND WAFFLE CHARTS

Unit charts show counts of a variable. Each symbol can represent an observation or a number of units. For example, if one symbol represents ten cars and there are ten car icons, the reader mentally multiplies the two for the total of one hundred cars. You can use unit charts to show percentages, dollars, or any other discrete amount. You can arrange them in different directions or break them down into subcategories by using colors or outlines.

Another advantage of these charts is that they can lend themselves to a more human connection. Bar charts, for example, are abstract and impersonal. They collapse all of the people reflected in that data point into a single shape. These charts, on the other hand, offer



Unit charts use symbols to show counts of a variable.



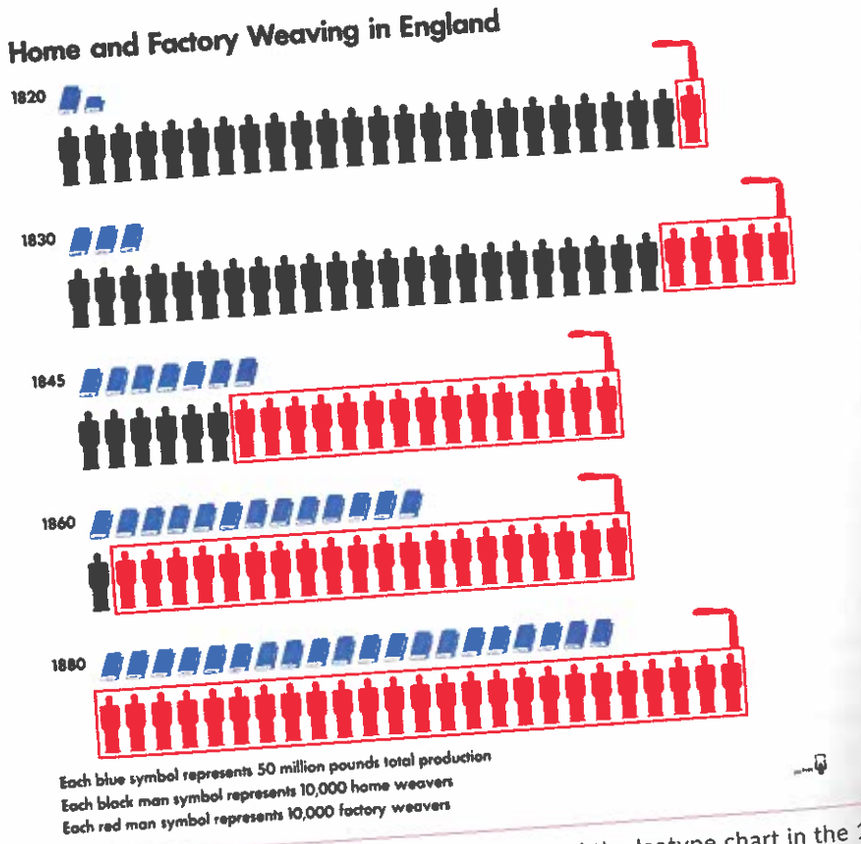
BANs—or Big-Ass Numbers—are a way to just show the values.

an opportunity to connect with the subject by reminding readers of the number of people represented, particularly if each dot represents one person.

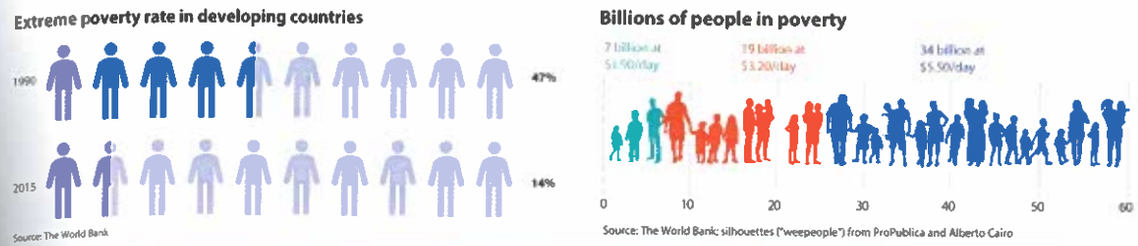
Another simple way to show these kinds of discrete counts is to just show the numbers. In *The Big Book of Dashboards*, authors Steve Wexler, Andy Cotgreave, and Jeff Shaffer call this the BAN approach: "Big-Ass Numbers." BANs might work best in a dashboard, infographic, social media post, or slide deck, but personally, I use them more sparingly in longer reports.

ISOTYPE CHARTS

Isotype charts are a subclass of unit charts that use images or icons instead of simple shapes. The term Isotype—International System of Typographic Picture Education—was coined by German philosopher and political economist Otto Neurath, his wife Marie Neurath, and their collaborator Gerd Arntz in the 1920s. They used the Isotype system to visualize all kinds of data, from workers in different industries, to population density and distribution, to the number of machines used in specific factories. They believed that this kind of visual system would help people communicate demographic, economic, and environmental issues to a broader public regardless of people's educational attainment.



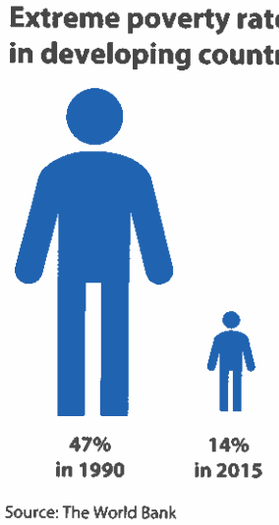
Otto Neurath, Marie Neurath, and Gerd Arntz developed the Isotype chart in the 1920s.



Two different ways to use icons in your data visualizations.

The graphic below is a classic example of their work. Each symbol represents a different count of workers (home or factory) and pounds of production. Aligned along a single vertical axis, it is easy to see how the values change over time.

We can take the same approach with the poverty data we've been using in this chapter. Notice that there's more than one way to use Isotype images in these two charts of extreme poverty rates. The version on the left uses individual icons to show each group of ten percentage points (the lighter icons could be included or not). The version on the right essentially orders the icons atop a bar chart. In either case, the icons connect the subject and content with an immediately recognizable visual image.

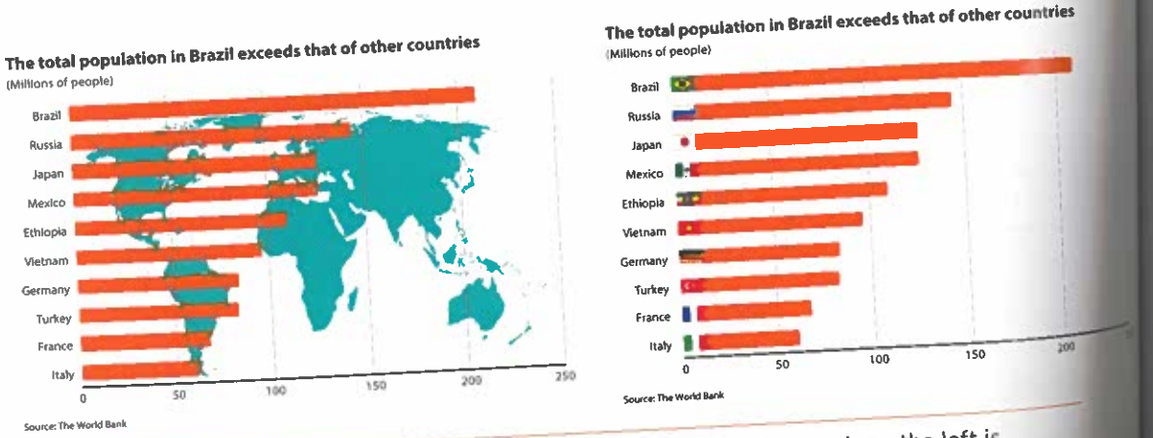


Icons can also be scaled according to their data values, but be careful because it's hard to know whether they are scaled according to the height, width, or area.

Instead of lining up the icons in rows or another arrangement, you can also scale them according to their value. But be careful, as sometimes it's hard to know whether the data are scaled according to the height, width, or area. That may not matter to every audience—it's clear here that the 47 percent is much larger than the 14 percent—but in cases where accuracy is paramount, this icon-scaling approach is inadvisable. Here, the vertical distance represents the data values, but the area of the icon on the left is about 10 times the size as the icon on the right.

These icon-driven graphs may look nice and engage readers, especially for few data values, but they can be difficult for your reader to count and compare. In his 1914 book, *Graphic Methods for Presenting Facts*, Willard Cope Bertin criticized this approach: "Charts of this kind with men represented in different sizes are usually so drawn that the data are represented by the height of the man. Such charts are misleading because the area of the pictured man increases more rapidly than his height." More recently, data visualization author and instructor Stephen Few wrote that unit charts force the reader "to either count, read the numbers, or do our best to compare the areas formed by each, which we can do poorly at best."

But sometimes the downsides of imprecision and slow comprehension may be offset by how memorable the chart is and how it engages readers, an issue that is borne out in recent research. Viewers in one study had clear preferences for graphs that included stacked icons rather than simple bars. They also found that images that sit in the background or are added to a chart but do not depict data are distracting to the reader, so if you choose to use these



Both graphs show the population in ten countries. The graph on the left is cluttered by an unnecessary backdrop of the world, while the one on the right uses flag icons to add identifying detail.

kind of small unit icons or images in your work, be sure to use them only to encode your data and not for gratuitous decoration.

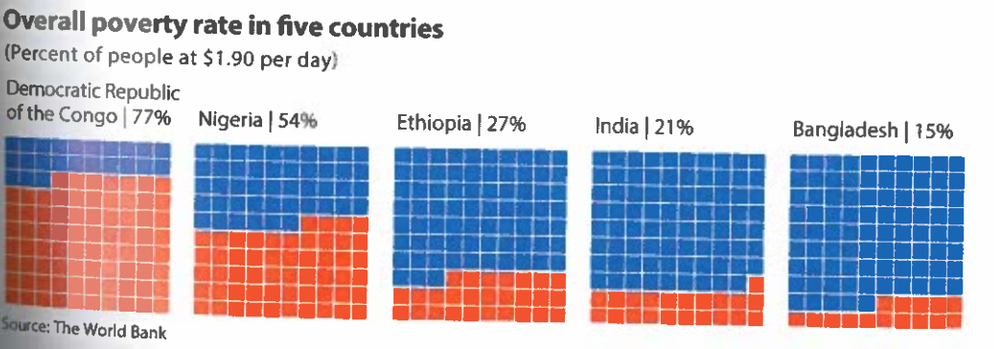
The graph on the left, which shows the population in ten countries, has a backdrop of a world map, a superfluous and distracting decoration. The graph on the right uses flag icons to add identifying detail and some visual engagement to a standard chart type.

Other research suggests that unit visualizations are intuitive and flexible, a way for readers to slowly wade into a visualization. Some have found that "unit visualizations are mostly useful when we want the readers to understand specific data item and encode its value (e.g., a unit, a person, a currency, a region, etc.)." Too much data and too many units, however, can create a visualization that is cluttered and obscures individual data points or the overall argument.

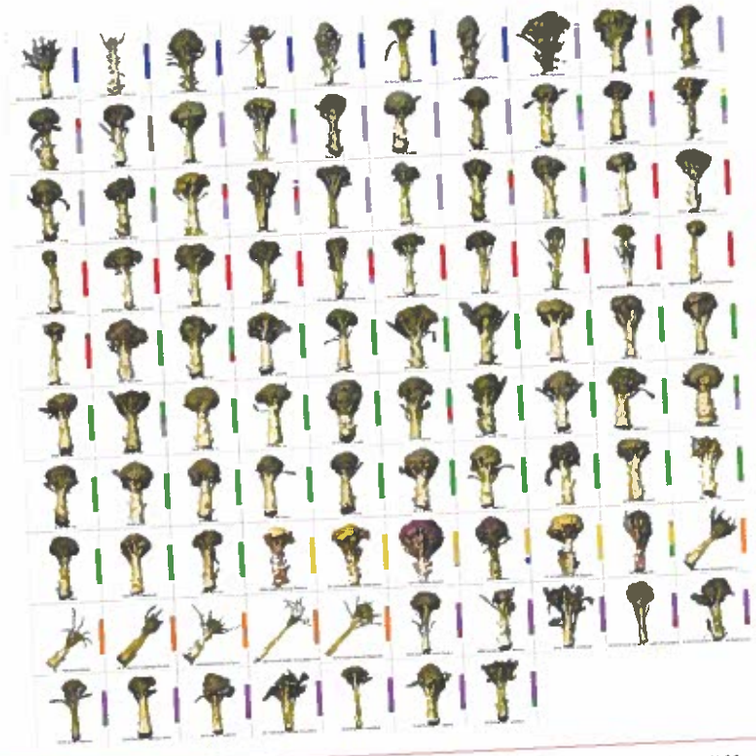
WAFFLE CHARTS

Waffle charts are another subclass of unit charts. They are especially good for visualizing part-to-whole relationships. Waffle charts are arranged in a 10 × 10 grid in which each colored cell represents one percentage point. You can use multiple waffle charts to show separate percentages—so the graph both shows part-to-whole relationships and lets your reader compare across the categories.

When creating unit or waffle charts, especially with icons, be mindful of your audience and how symbols may not appropriately represent your content. If you are visualizing child mortality rates in different countries, for example, an icon of a baby is not appropriate. Using icons of men to represent counts of people may ignore women in your data set. Alternatively,



Waffle charts are a subclass of unit charts and, in this case, arrange the squares in a 10 x 10 grid.



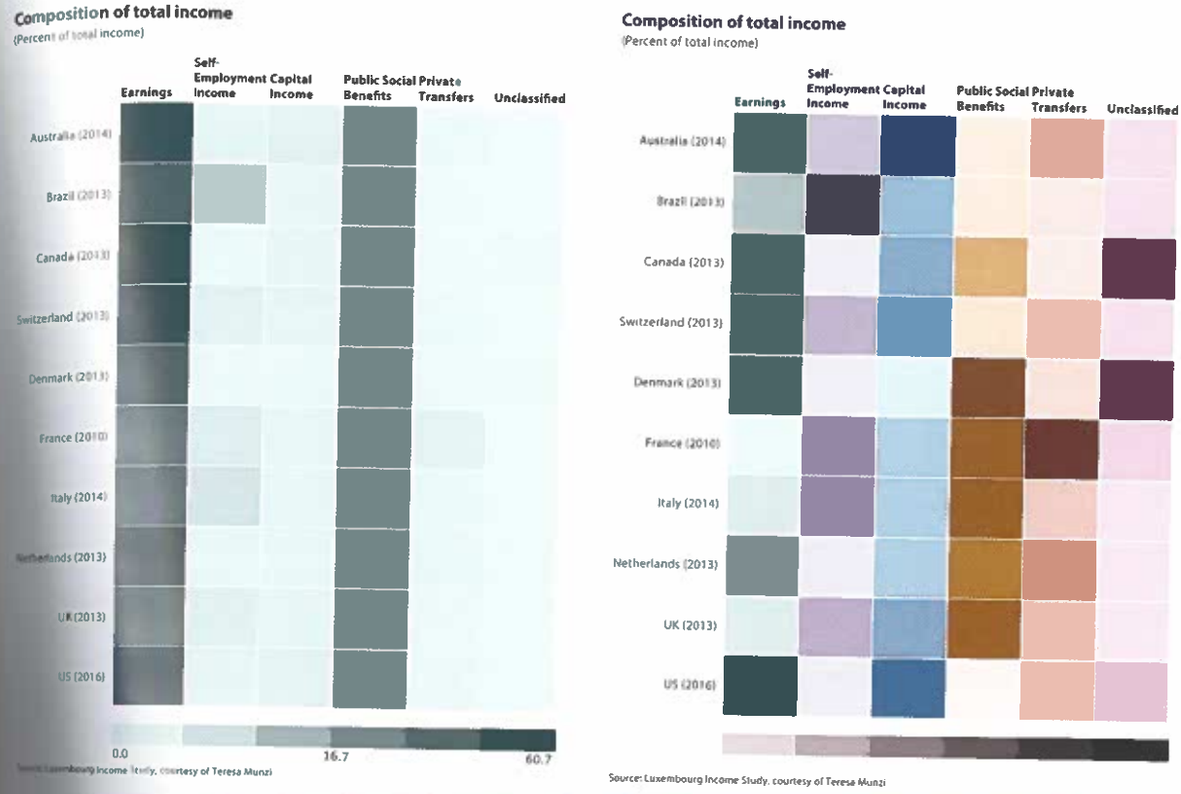
Zachary Stensell created this small multiples visualization of the different types of broccoli he grew in his garden.

if you want to measure the different types of broccoli in your garden, simply line up the pictures, as in this fun project from Zachary Stensell shown above.

HEATMAP

Heatmaps use colors and color saturations to represent data values. Simply put, a heatmap is a table with color-coded cells. They are often used to visualize high-frequency data or when seeing general patterns is more important than exact values.

These two heatmaps show the different components of total income for ten countries using data from the Luxembourg Income Study. The version on the left uses the same color scale for all six categories, where lighter colors encode smaller values and darker values

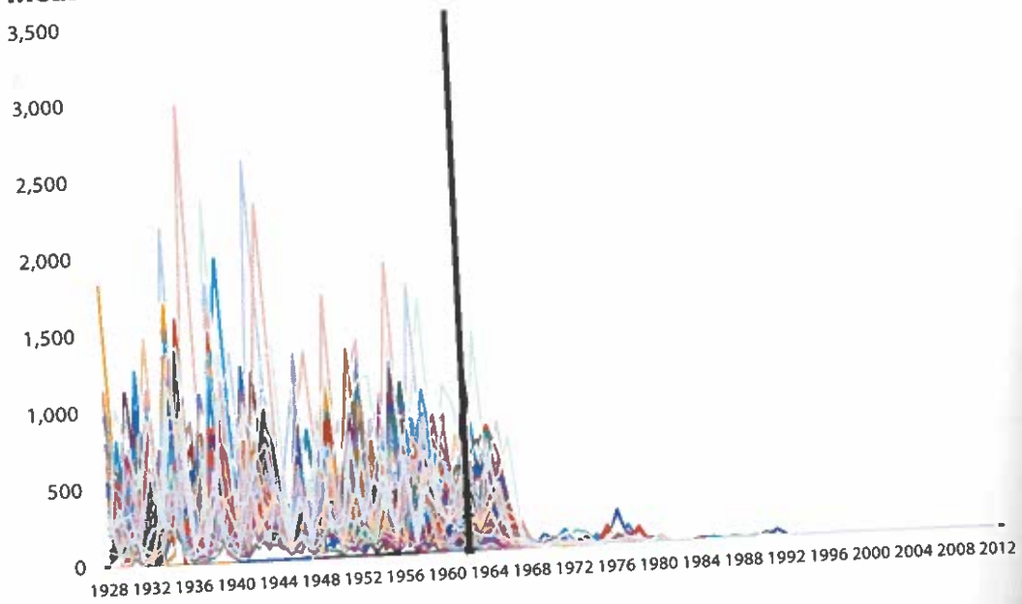


Heatmaps use colors and color saturations to represent data values and can focus the reader's attention along the columns or across the rows.

encode larger values. In this view, you can see that people's earnings account for the greatest share of their total income, and, in most countries, public social benefits appear to be the second-largest share. In the heatmap on the right, each category is assigned its own color scale. Here, you can more clearly see that public social benefits (in the fourth column) account for a smaller share of total income in Australia, Brazil, Switzerland, and the United States. Which one is better depends on your goals. Do you want your reader to compare across all of the values or within each category?

You can also use a heatmap to show changes over time. Imagine a spreadsheet that contains infection rates from the measles disease for every state in America from 1928 to 2008. If the spreadsheet had states ordered along the rows and years along the columns, your first instinct might be to create the line chart on the next page.

Measles incidence in the United States from 1928 to 2012



Source: Data from Project Tycho, https://www.tycho.pitt.edu/data

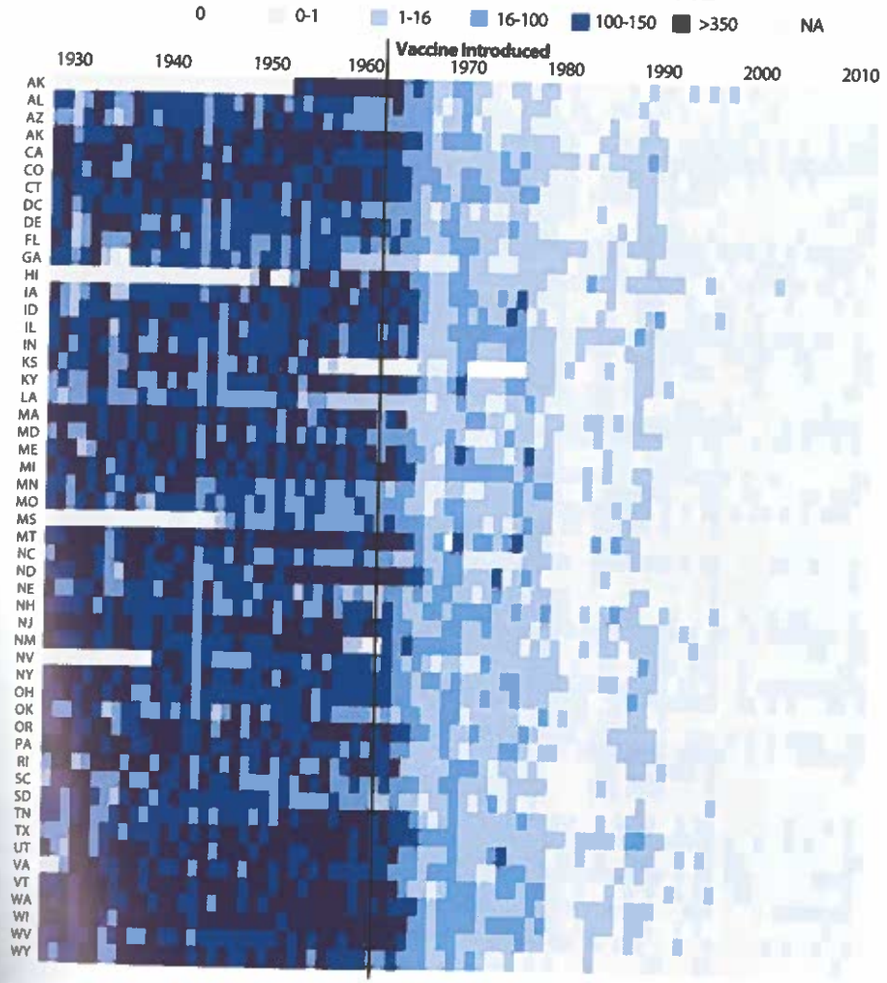
You can see basic patterns in measles infections across the United States in this dense line chart, but it's hard to pick out any specific values.

There's nothing inherently *wrong* with this next line chart—you can see the positive infection rate from 1928 to about 1963 (marked with the black vertical line), the year when the measles vaccine was introduced. Over the next five years or so, infections dropped quickly, and eventually reached around zero within about ten years. What you essentially get from this chart is that there were infections—going up, going down, in a tangle of lines—until about 1963.

The *Wall Street Journal* looked at the same spreadsheet and, instead of creating a dense line chart, they created a heatmap. I've created my own version here using a different color palette and discrete categories of the infection rate. You can see the darker blue cells (mostly above 16 infections per 100,000 people) prior to the introduction of the vaccine, again marked with a black line. After 1963, the colors quickly transition to lighter shades of blue, and ultimately to the lowest rates of infections (zero infections and fewer than 1 infection per 100,000 people).

This chart may not be inherently *better* than the line chart, but it does let you more easily examine each state or year far more easily than picking out a single line from the tangle in the line chart. Also remember that sometimes being different is good in itself. How many

Measles incidence in the United States from 1928 to 2012

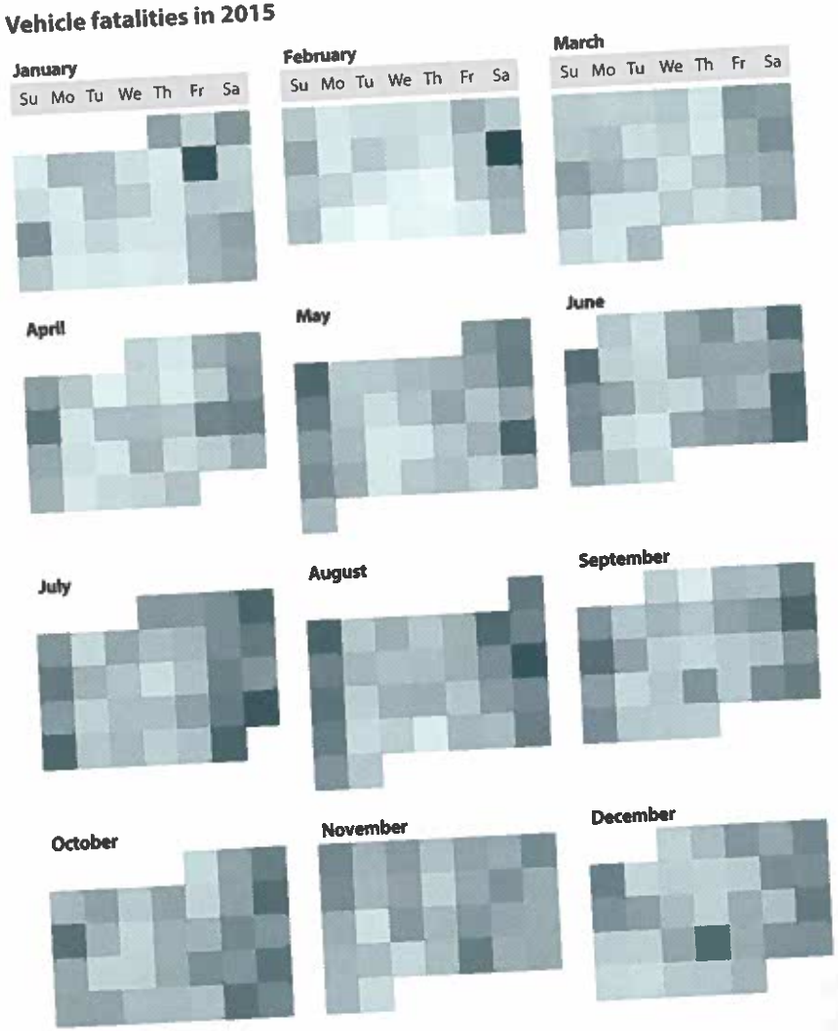


Source: Project Tycho, https://www.tycho.pitt.edu/data

This heatmap may not be inherently better than the line chart at showing measles infections rates, but it does let the reader more easily examine each state or year.

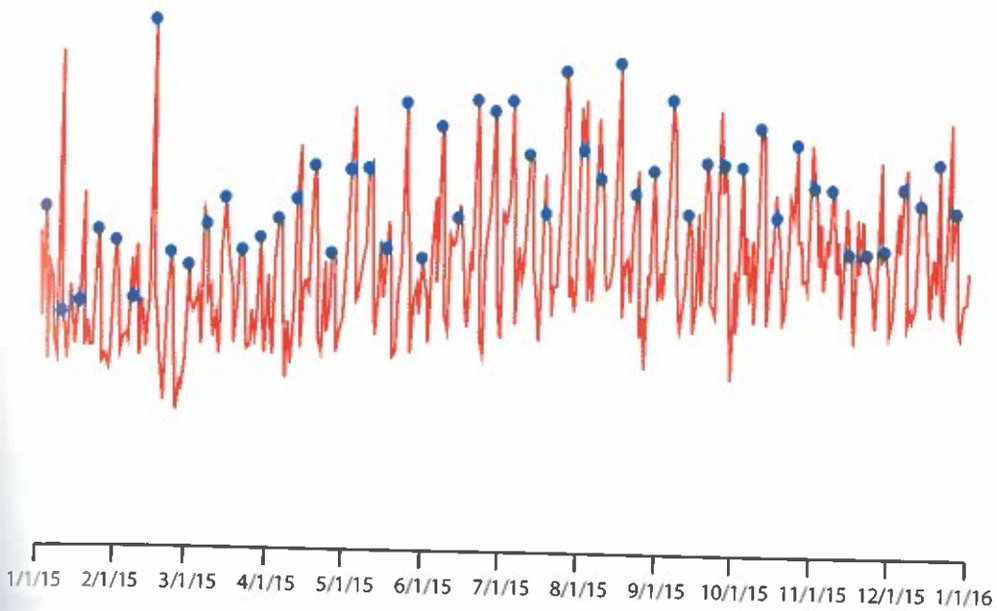
complex line charts have you seen and just immediately skipped over? The heatmap, with its different look and color, can draw readers in. As artist and data visualization expert Giorgia Lupi said, "Beauty is a very important entry point for readers to get interested about the visualization and be willing to explore more. Beauty cannot replace functionality, but beauty and functionality together achieve more."

Another way to use the heatmap is to modify the layout, for example, applying it to a calendar. In this example, vehicle fatalities in 2015 are plotted on heatmaps of months. Notice how easy it is to see the higher fatality rate on Fridays and Saturdays along the right edge of each column.



Source: National Highway Traffic Safety Administration
Note: Inspired by Nathan Yau at FlowingData.com

Another way to use a heatmap is to modify the layout, as in this version that shows vehicle fatalities in 2015.



Source: National Highway Traffic Safety Administration

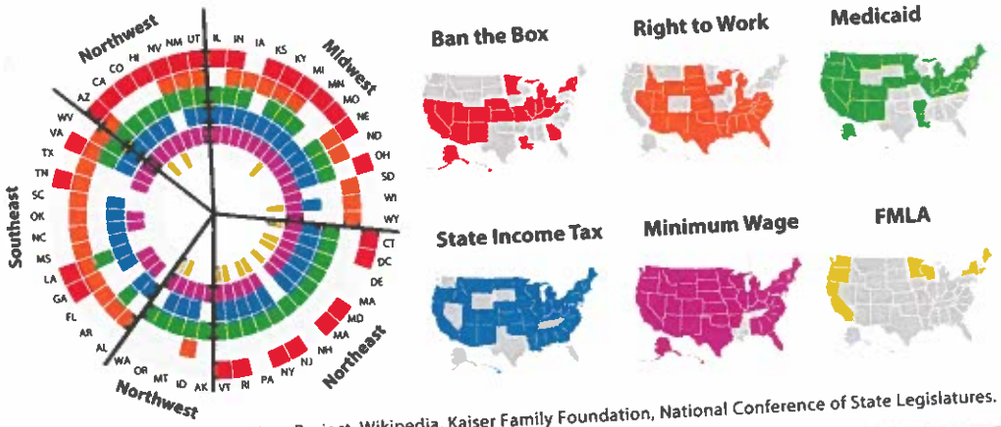
This line chart shows the same data as in the heatmap calendar, but it's more difficult to reach the same conclusion.

By comparison, consider plotting the same data in a line chart. Even with the additional blue circles used to mark Saturdays, it's difficult to reach the same conclusion about more fatalities on the weekends.

Unlike the measles example, where both charts had advantages and disadvantages, in this case the calendar heatmap is a better approach because it does a better job highlighting the important pattern of deaths on the weekends—and it's a more engaging graph placed in a familiar shape.

A final way to modify the heatmap is to change the rectangular layout altogether. On the next page, each of the fifty states is plotted along each radii of a circle, grouped into five geographic regions (separated by the black lines). Each ring represents a different (binary) data type, such as whether the state has right-to-work laws, an income tax, or a minimum wage.

This alternative to a set of six maps has some advantages and some disadvantages. On the one hand, it is a compact representation of six different data series that the reader can quickly and clearly see the categorical and part-to-whole data within and between states. On the other hand, it's not a familiar graph type, and that may turn off some readers. It's also



Source: National Employment Law Project, Wikipedia, Kaiser Family Foundation, National Conference of State Legislatures.

Heatmaps can be arranged in different ways. The radial layout has some advantages and disadvantages, but so does the six-pack of maps.

worth noting that the order of the rings can affect our perception of the data because the (red) squares on the outer ring are by definition larger than the (yellow) squares on the inner ring.

GAUGE AND BULLET CHARTS

The gauge chart (or gauge diagram) looks like the speedometer in your car's dashboard. Typically set up somewhere between a half-circle and a circle, it uses a pointer or needle to indicate where your data fall within a particular range. Sections of the gauge are shaded to illustrate sections such as poor, good, and excellent.

I see gauge diagrams most often in financial planning tools because they give an easy, familiar way to visualize targets or progress towards a goal. They also frequently show up in fundraising campaigns where the entire semicircle represents the goal, and the needle and filled area represent money raised so far. This can be a good example of using a familiar shape to support the metaphor of the visualization—everyone understands that "filling up" the gauge means the fundraising effort has reached its goal.

Gauge diagrams do, however, introduce perceptual challenges because, again, people are not very good at measuring and comparing angles. If you want to give your reader a general

Gauge charts are familiar and easy to read.

sense of the values, the gauge chart is a decent choice. But if enabling your reader to discern the specific values and compare those values to the ranges is of utmost importance, then it is not.

Given the familiarity with the gauge and the obvious metaphor it represents, it's perhaps no surprise that they show up in a variety of settings. As just one example, I once received

This series of gauge charts shows four real estate trends in my Northern Virginia neighborhood.
Source: MountJoy Properties, brokered by Keller Williams Realty.

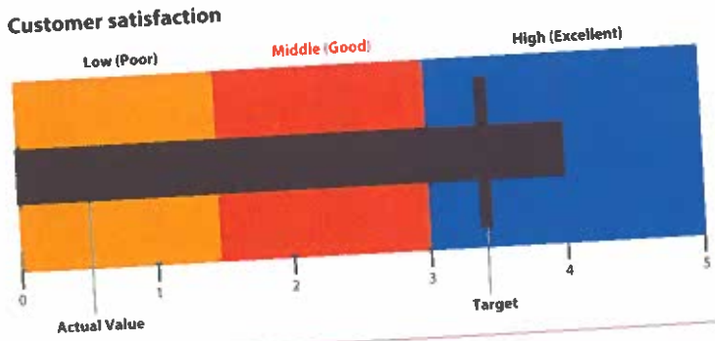
a flyer from Mountjoy Properties (a residential real estate team that serves the DC Metro Area) that consisted of a series of gauge diagrams showing current real estate trends in my neighborhood of Northern Virginia. I could pretty quickly see the current state of the market, but adding more data or more detailed data might be more difficult.

BULLET CHARTS

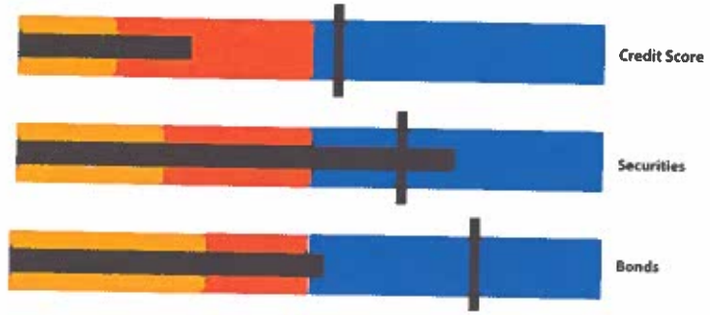
Because of these perceptual challenges with the gauge chart, author Stephen Few invented the bullet chart, which is a linear, more compact way to show similar kinds of data. The basic bullet chart contains three different data elements:

1. First, there is the actual or *observed value*, shown here as the black horizontal bar. In this illustration, the bar represents an average customer satisfaction score of 4.0.
2. Second, there is a *target value*, shown here as the black vertical line. Here, we were aiming for a satisfaction score of 3.5.
3. Finally, there is the *background range*, which shows grades or bands of success, such as poor, good, and excellent. These sit behind the other two series so the reader can compare the actual and target values. Here, poor scores are 1.5 and below, good scores are from 1.5 to 3.0, and excellent scores are anything above 3.0.

The different components of the bullet charts can vary. There might be a scale of five ranges instead of three, or there may not be a target value. The scales can also show the underlying



The bullet chart includes five separate data values.



Combining different bullet charts is a compact way to let the reader make a series of different comparisons.

distribution of the data—for example, showing quartiles or ranges of percentiles (see Box on page 183 for more discussion of percentiles). Because the bullet chart is so compact, it's easy to create multiple versions and stack them together. The bullet charts above show three metrics you might find in a financial report, but they are more compact than gauge charts and, because the rectangles are aligned, it is easier to compare across the different categories.

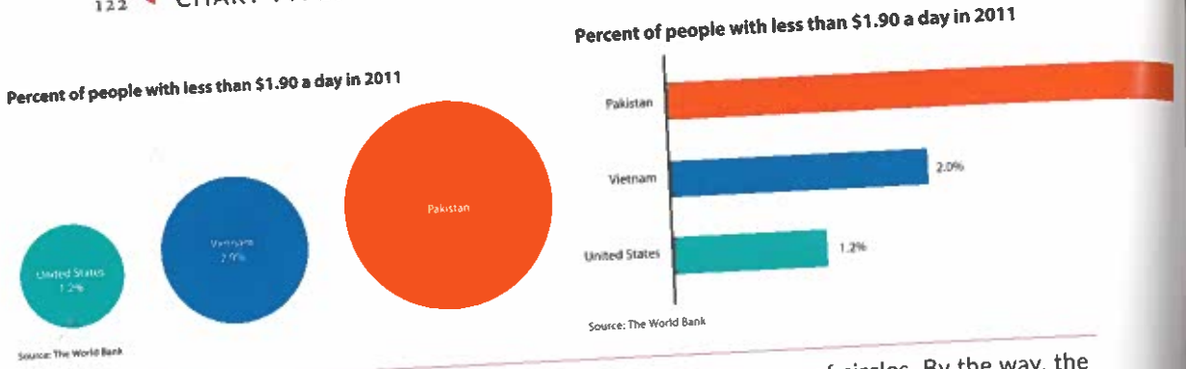
BUBBLE COMPARISON AND NESTED BUBBLES

In the basic bubble graph, circles represent values. Like a bar chart, the purpose of these types of charts is to compare values between categories. But unlike the bar chart, humans are not very good at accurately comparing the sizes of circles (remember the perceptual ranking diagram from page 14). Still, circles may be more visually interesting, they can reinforce a visual or metaphor, and they are a good choice when discerning exact quantities is not paramount.

Another drawback of proportionally sized circles is that you cannot visualize negative values. While bars can go in both directions—typically right or upward for positive and left or downward for negative—that's much harder to visualize with circles.

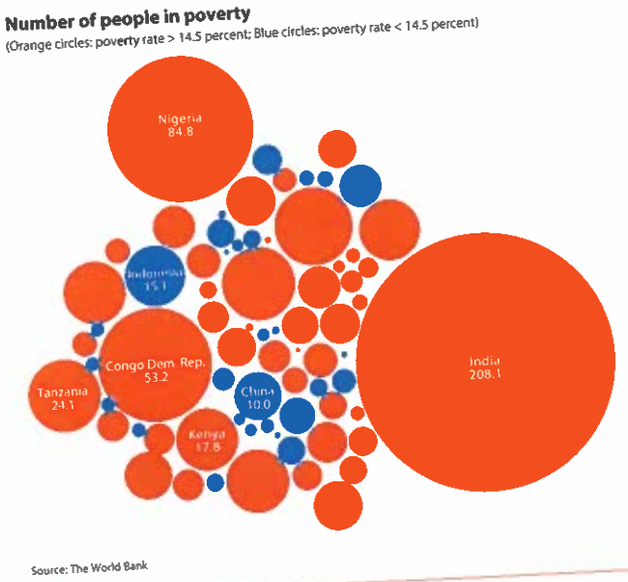
In any case, we are not very good at making accurate estimates from the circles even when they are sized by area. Instead, what I think we try to do is make the comparison based on the diameter of the circle—like in a bar chart—which gives us incorrect conclusions. Take a look at the two graphs on the next page and try to guess the percent of people living on less than $1.90 per day in Pakistan. Do you think that task is easier in the bubble chart or in the bar chart?

This is not to say that you should never use circles. Remember, again, it's all about your audience. A bubble comparison chart, inserted in a short article off to the side with the



We are better at discerning differences from bars than by areas of circles. By the way, the percent of people with less than $1.90 a day in Pakistan is 4.0 percent.

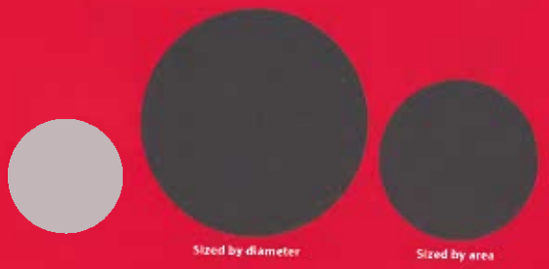
numbers placed prominently in the middle of each circle may be more engaging than a standard bar chart. Too many circles, however, may make it difficult for your audience to discern any quantities or relationships. In this next example, yes, you can see that India, the Democratic Republic of the Congo, and Nigeria have the largest number of people in poverty, but it's difficult to quickly assess *how* different they are or the numbers of the next set of countries.



A bubble comparison chart can be engaging and interesting, but it can also be hard to discern the values.

CALCULATING THE AREA OF A CIRCLE

Remember to size the circles by area, because using the radius or diameter generates circles that overemphasize differences (the radius or diameter scales in a linear way, but the area scales quadratically). The first black circle is sized relative to the gray circle using the diameter while the second black circle uses the area. As you can clearly see, using the diameter skews our perception and makes the difference between the two values look much larger.

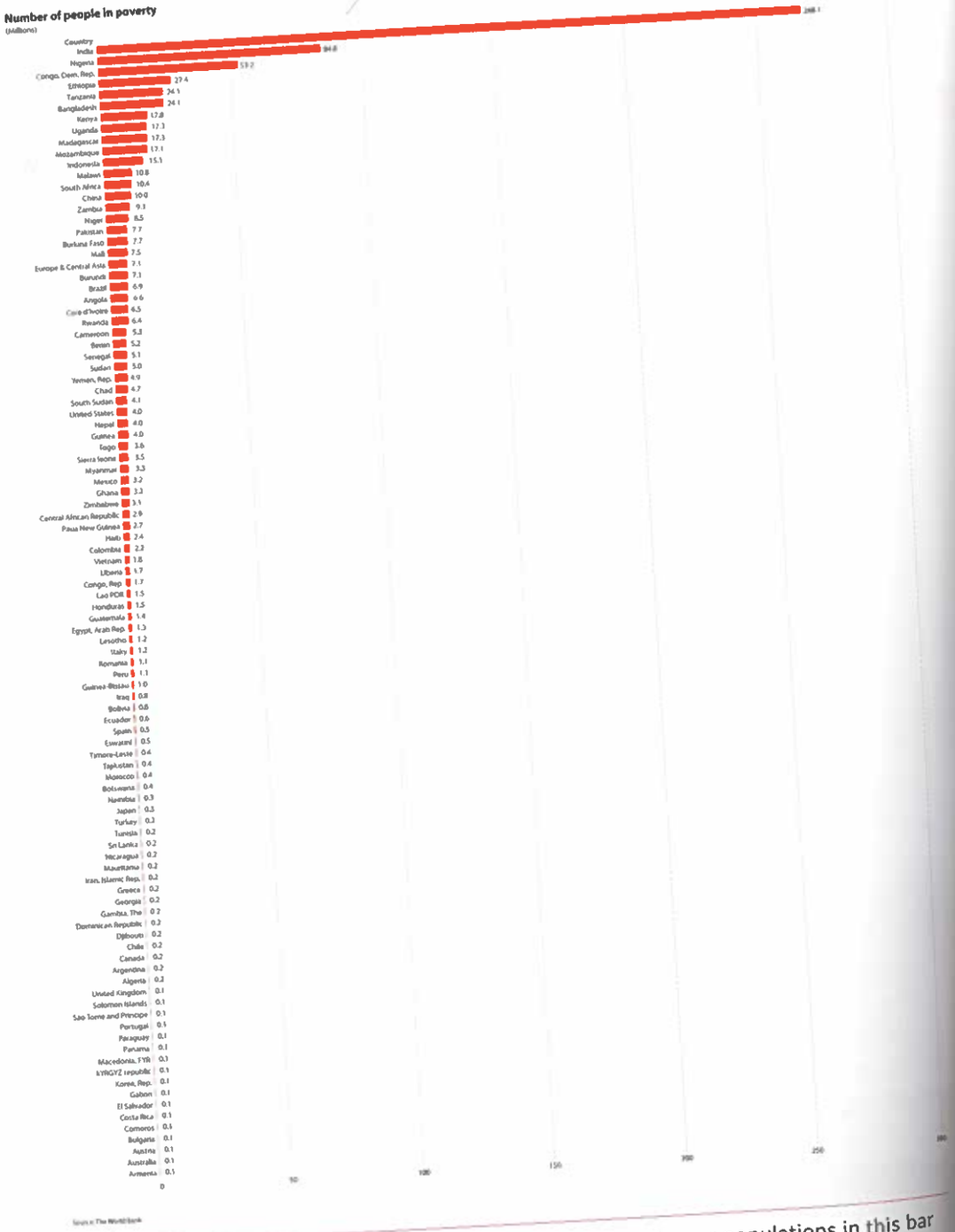


A simple example will demonstrate the importance of using the area rather than the radius/diameter for sizing these circles. In case you don't remember your middle-school math, the diameter is any straight line that passes through the middle of the circle. The radius (r) is half the diameter. And the area (A) is equal to the constant pi (π) times the radius-squared, or $A = \pi r^2$.

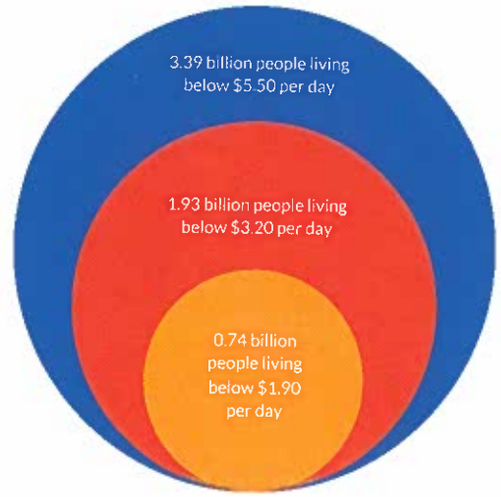
So, say the data value for the gray circle is 1 and for the black circle is 2. If we start with the gray circle and set the radius equal to 1, we can find the area is equal to: $A_G = \pi r^2 = \pi 1^2 = \pi$.

To find the size of the black circle the correct way, we say that the area of that circle is twice the size of the gray circle, corresponding to the difference in their data values. So, if we double the area of the black circle so that it is now 2π, we can then rearrange the formula and find the radius of the black circle to draw it: $r_B = \sqrt{A_B / \pi} = \sqrt{2\pi / \pi} = \sqrt{2}$.

Let's do this the wrong way and use the radius instead of the area. In this case, the radius of the gray circle is still 1, so let's make the radius of the black circle 2, again corresponding to the difference in their values. This makes the area of the black circle now $A_B = \pi r^2 = \pi(2)^2 = 4\pi$. In other words, the area of the black circle sized in this way is four times the size of the gray circle instead of twice the size, as the data suggest.



It's easier to pick out the countries with the most and least poor populations in this bar chart, but it takes up the entire page.



Source: The World Bank

The *nested bubble chart* can sometimes mask circles in the back—but it can also make for easier comparisons.

If we were to use a more perceptually accurate representation of these same data in, say, a bar chart, the visual becomes much larger as on the previous page. Each bar is labeled here, so a reader could find Madagascar, but is that important? Again, is the goal to show all of the countries or just a subset? As always, consider your goals and whether your reader needs a perceptually accurate view to understand your argument.

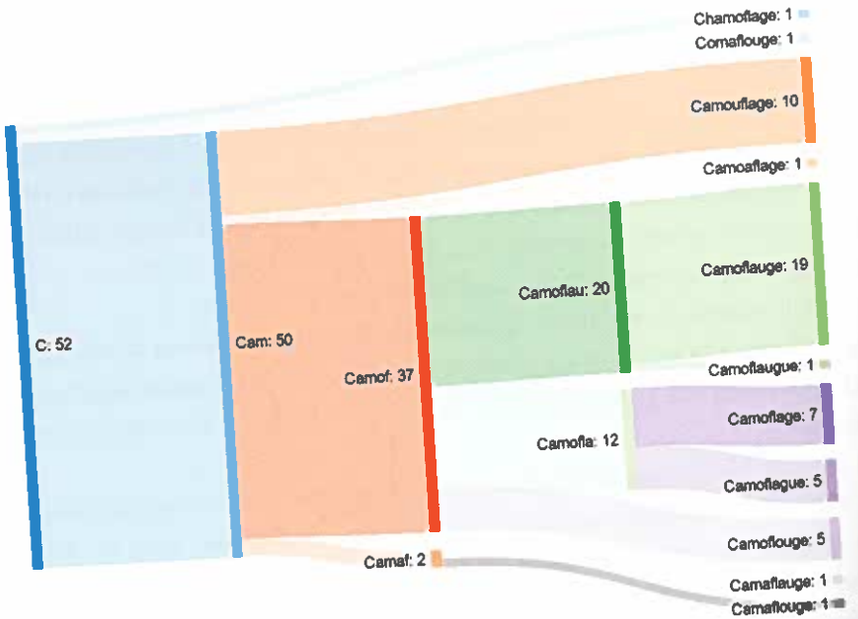
The bubble charts shown above are known as bubble comparison charts. Layering circles on top of each other, as in the graph above, are often called nested bubble charts. The nested bubble chart can sometimes mask circles in the back, but it can also make for easier comparisons.

You can use bubbles to demonstrate correlations (see the bubble chart in Chapter 8) or add bubbles to a map to encode another variable (see the point map in Chapter 7). In general, while there are perceptual issues with using circles and bubbles in data visualization, they can also be more engaging and enjoyable than yet one more bar or line chart. As Amanda Cox, Data Editor at the *New York Times* said, "There's a strand of the data viz world that argues that everything could be a bar chart. That's possibly true but also possibly a world without joy."

SANKEY DIAGRAM

Sankey diagrams—named for their creator Matthew Henry Phineas Riall Sankey in 1898—are especially useful for showing how categories compare to one another and flow into other states or categories. Arrows or lines display the transition from one state to another, and the width of the lines denote the magnitude of each transition. Changes can occur over time or as comparisons between categories. For example, a Sankey diagram could be used show how different companies in an industry have merged, broken apart, or failed in different years.

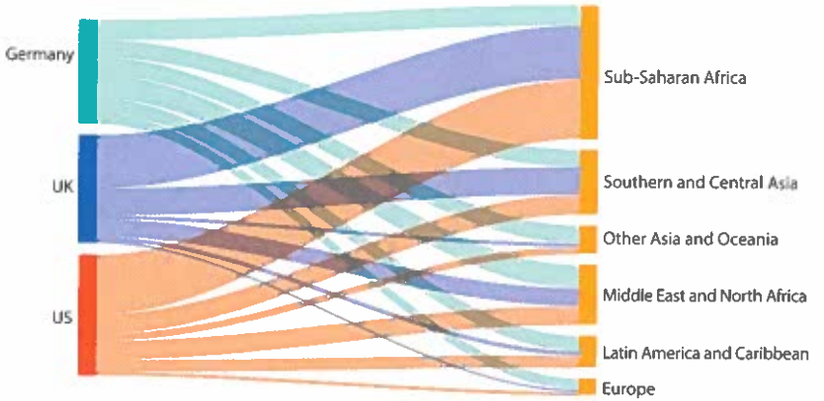
This Sankey diagram shows how fifty-two students tried to spell the word *camouflage*. The first blue segment shows that all fifty-two students started with the letter "C", fifty then went to "Cam", followed by thirty-seven with "Camof", and so on. Ten students spelled the word correctly, shown in the orange segment near the top of the graph.



One of my favorite Sankey diagrams—it shows how fifty-two students tried to spell the word *camouflage*. Graphic by Tim Bennett, data collected by Reddit user iheartdna.

Financial support flows from Germany, the United States, and the United Kingdom to different areas of the world

(Percent of total support)



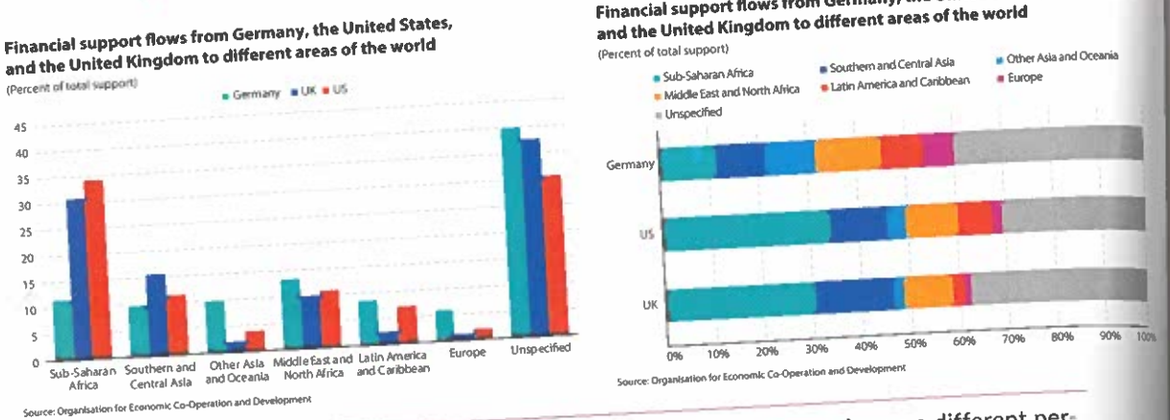
Source: Organisation for Economic Co-Operation and Development

This Sankey diagram shows flows in federal aid from the United States, United Kingdom, and Germany to regions around the world.

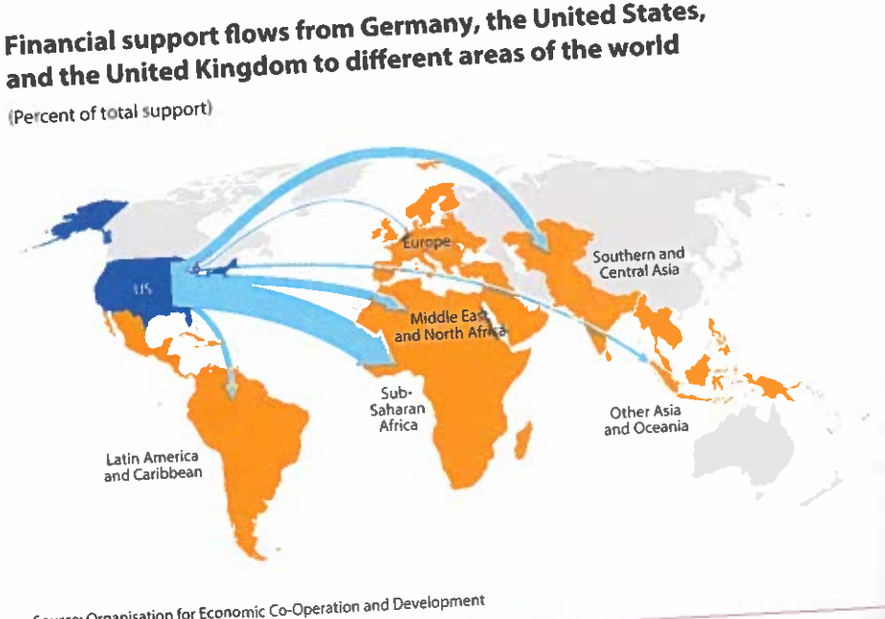
More applicable to the sort of content we've looked at thus far in this chapter, the graph above shows flows in federal aid from the United States, United Kingdom, and Germany to regions around the world. You can see that countries in Sub-Saharan Africa and unspecified countries receive the bulk of the aid and that the United Kingdom and United States contribute more to Sub-Saharan African countries than does Germany.

Presenting these data as paired or stacked bar charts give us a different perspective. In the paired bar chart on the left, the obvious first comparison is across the funder countries for each region of the world. In the stacked bar chart, by comparison you're more likely to compare funding across the recipient regions—that, for example, the greatest share of spending on Sub-Saharan African countries is from the United States. The Sankey draws our attention in a different way, reading horizontally across the page in a way that mixes these two comparisons, privileging neither. Neither view is "right" or "wrong," but may serve different audiences differently, highlight different patterns, and answer different questions.

Sankey diagrams can be layered together with other chart types. In this example, the aid flows from the United States are presented on a world map. This provides a geographic view of the data, but also simplifies things—imagine how cluttered the visual would look if it included flows from the United States, United Kingdom, and Germany.

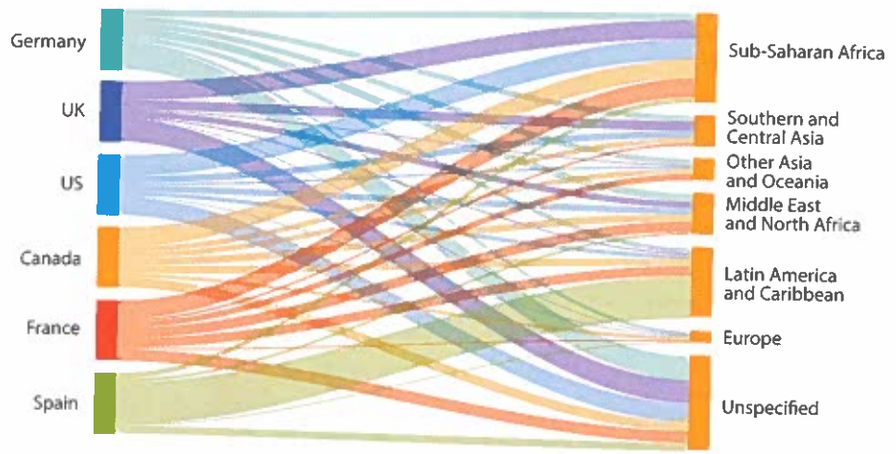


Presenting the financial flow data as a paired or stacked bar chart give us a different perspective than in the Sankey diagram.



This flow map provides a geographic view of the financial flow data, but it also simplifies things by only showing flows from the United States.

The biggest problem with Sankey diagrams—and many charts for that matter—is plotting too many series, as in this version that includes more countries. With too many groups or too many crossings, the chart becomes difficult to navigate. If you find yourself with too



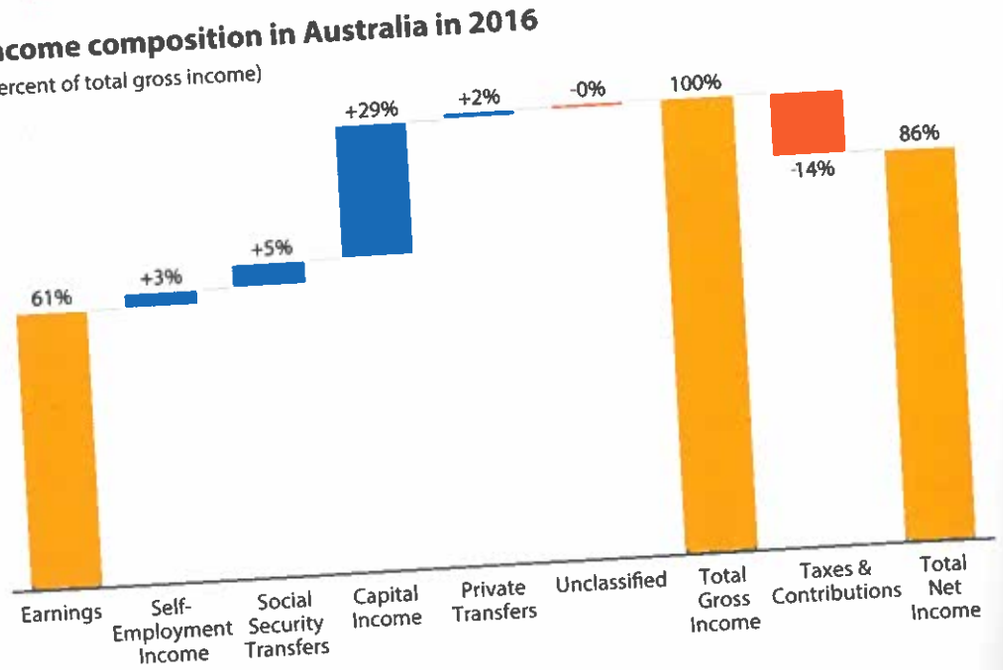
The biggest problem with Sankey diagrams—and many charts for that matter—is plotting too many series makes it difficult to identify any patterns or trends.

many lines or crossings, try simplifying your data, using multiple Sankey diagrams, or using a different chart type.

WATERFALL CHART

A waterfall chart shows a basic mathematical equation: adding or subtracting values from some initial value to produce a final amount. It is essentially a bar chart, but each subsequent bar starts where the previous one left off, showing how they accumulate across the graph. Typically, negative values are given a different color than positive values, and so are the totals at the beginning and end. Including lines that connect the bars can guide the reader through the visualization. Because the lines are guides and not actual data, they should be lighter and thinner than the other elements.

This next chart shows contributions to total gross income and total net income for Australia in 2016. The data are the same as those used in the heatmap example from earlier, but in that approach, I could fit data for ten countries in the same view. Imagine trying to do the



Source: Luxembourg Income Study, courtesy of Teresa Munzi

A waterfall chart shows a basic mathematical equation: adding or subtracting values from some initial value to produce a final amount.

same with a waterfall chart—we would need ten different graphs—something that might be useful under certain scenarios, but is certainly less compact than the heatmap.

Waterfall charts can also show changes over time. You might show, for example, contributions to total GDP from one year to the next, and how different values contribute to the change over the course of the year. Any data series that are added or subtracted to one another can be presented in this way, though, again, it is a nonstandard chart type and may require your reader more time to navigate it.

CONCLUSION

From single bars to groups of bars to stacks of bars, the bar chart is one of the most familiar data visualizations for showing categorical comparisons. It also ranks at the top of our

perceptual ranking scale from earlier. But the bar chart also poses certain challenges: too many bars can make the visual seem overwhelming and cluttered, and stacking the series on top of one another makes it more difficult to compare series that are not aligned on the same axis.

The basic bar shape can be organized in many ways. They can sit next to each other or diverge from a central baseline. They can be stacked on top of one another on a horizontal or vertical dimension, or both as in a mosaic chart. They can also be arranged to show simple mathematical equations, as in a waterfall chart. We are generally good at discerning the data values from the lengths of the bars, so many of these chart types will make it easy for your reader to perceive the exact value.

There are other ways to let your reader make comparisons. I'm especially fond of dot plots because they remove a lot of the heavy ink from a standard bar chart and free up space to add annotation and labels. Using icons, squares, or other shapes can engage our audience in ways that standard charts may not, but may be less data dense.

While bar charts sit at the top of the perceptual ranking list, let's be honest: They can be very boring. We see bar charts every day. As chart creators, sometimes our challenge is to find ways to engage our audience, and deploying less common chart types from our data visualization toolbox can do just that. It's up to us to determine where we want to focus our reader's attention, on the level or the difference, the single or multiple comparison, or the relative or total values.