his entire focus to a single variable: the number of people in the flow. This variable sees only one type of variation—*a sharp and steady decline*. It seems to have been this potent and poignant message that made these two maps (and particularly the Napoleon one) so successful in telling a story about the cataclysm of war.

CONCLUSION

In this chapter we surveyed the promises and perils of visualizing geographic data. Sometimes the varying sizes of geographic units may distort the data. Other times sizing the geography to the data may make a familiar geography look foreign.

When working with geographic data, your instinct may be to create a map. But take a moment to consider: Is a map the best way to present your data? Does your reader need to see the exact differences between data values? If so, the aggregation problem inherent in many maps may make that difficult. Are there clear geographic patterns to be seen in the data? If not, then the map may not actually help the reader see your point.

If a data map *is* the right approach, carefully consider the map projection you use and whether the standard choropleth map is the best choice. Maybe some kind of cartogram—even with all its flaws—would be a better fit for your context and reader.

You may also determine that the best approach is to *combine* visualization types. Depending on your final publication type, you might use multiple visualizations, say, a map with a bar chart or table. This approach can help give your readers a familiar visualization type in which they can identify themselves and their location, but also help them gain a better, more detailed view of the actual data.

8

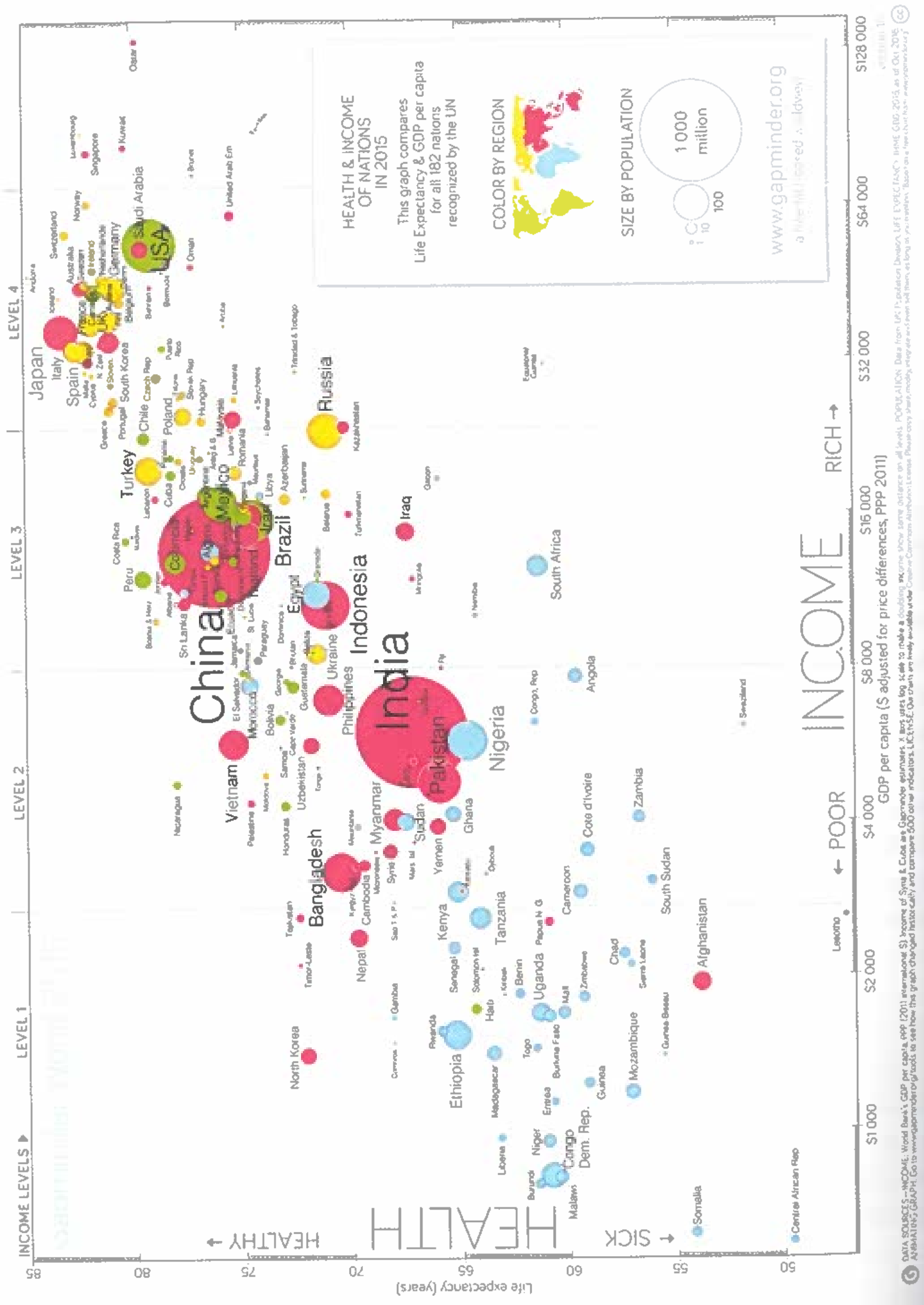
RELATIONSHIP

The charts in this chapter show relationships and correlations between two or more variables. Perhaps the most familiar chart type in this class is the scatterplot, a chart in which the data are encoded to a single horizontal and vertical axis. Other shapes and objects can also be used to visualize the relationship between two or more variables—a parallel coordinates plot uses lines, while a chord diagram uses arcs within a circle. These charts can show the reader correlations and even causal relationships.

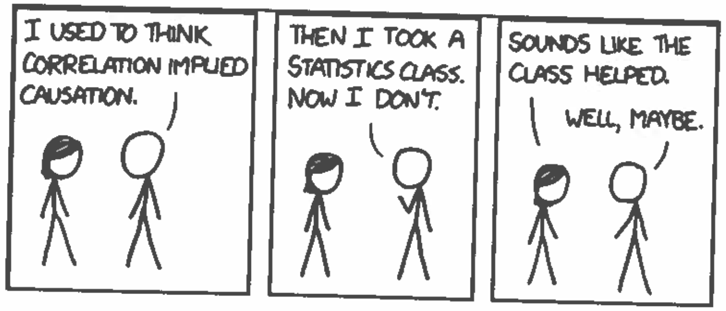
For this chapter, I use the color palette and font from a famous bubble chart created by the Swedish academic Hans Rosling and his colleagues at Gapminder, a foundation dedicated to visualizing statistics. Rosling's Gapminder project didn't lay out a specific data visualization style guide, but the visualizations in this chapter use the basic colors and font (Bariol), with additional styles based on other visualizations from the Gapminder website.

SCATTERPLOT

The scatterplot is perhaps the most common visualization to illustrate correlations (or lack thereof) between two variables—one variable is plotted along a horizontal axis, and the other along a vertical axis. The specific observations are plotted in the created space. Unlike a bar chart, the scatterplot axes do not necessarily need to start at zero, especially if zero is not a possible value for the data series.



Source: Based on a free chart from www.gapminder.org


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

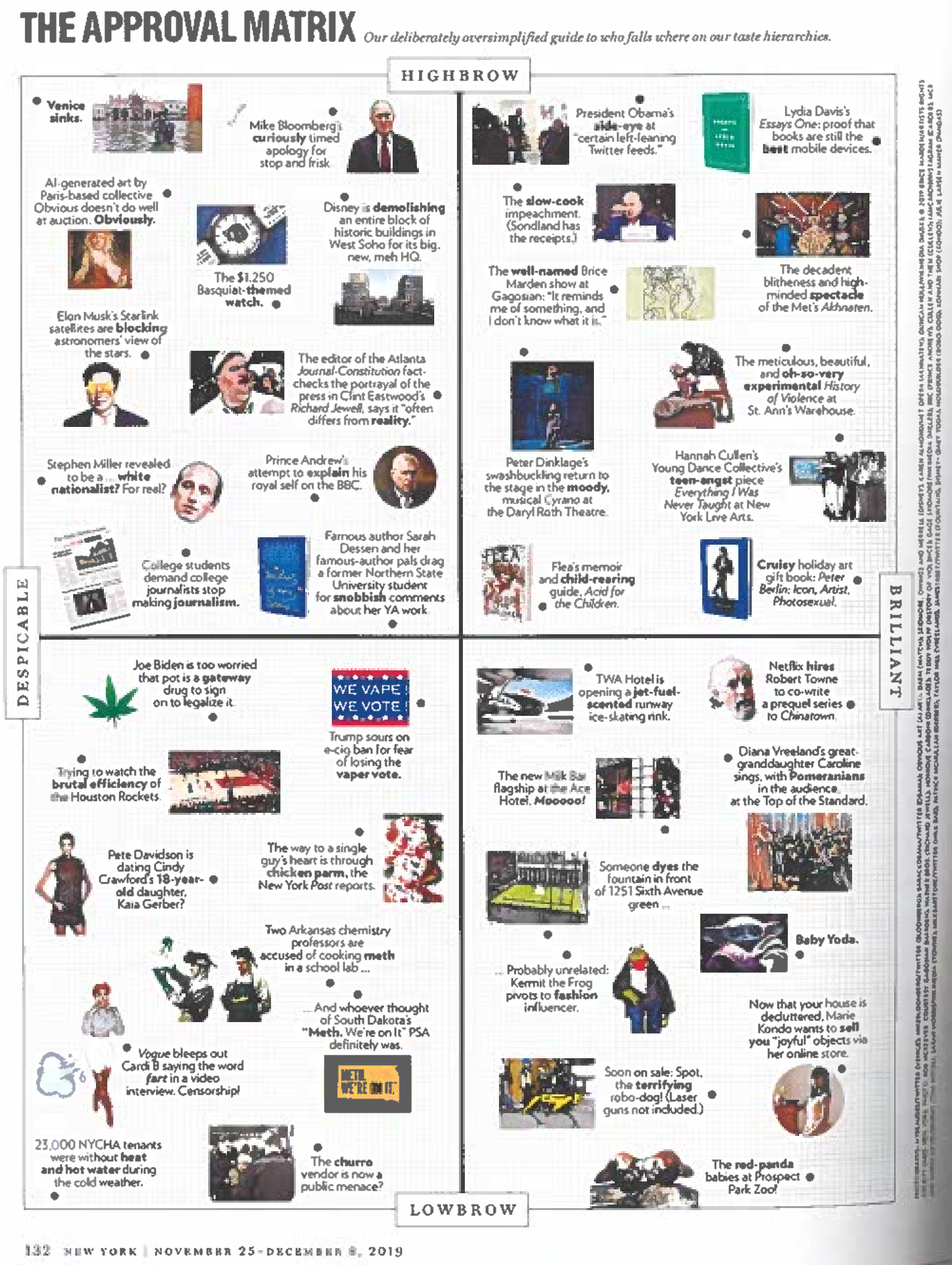
Source: XKCD

One of the most famous graphs that shows the relationship between two variables is the set of scatterplots created by Rosling and his colleagues at Gapminder. A physician by training who spent roughly two decades studying public health in rural areas across Africa, Rosling is perhaps best known for engaging presentations and data visualizations, and for promoting the use of data to explore issues around international development. In Rosling's 2014 TED Talk, he showed an animated scatterplot of the relationship between the fertility rate (number of births per woman) and life expectancy at birth from 1962 to 2003 for countries around the world.

As I've noted previously, there are many nonstandard graph types—scatterplots included—with which your reader may be unfamiliar. This doesn't mean you can't use these visualizations, but it does mean you should be mindful that your reader may be unfamiliar with them, and consider how to prepare them to understand your graphs.

Some readers may be familiar with scatterplots (or other nonstandard graph types for that matter) even if they are not familiar with reading *data* in scatterplots. *New York Magazine's* weekly *Approval Matrix* on the next page, for example, is their "deliberately oversimplified guide to who falls where on our taste hierarchies." Bits of text, images, and icons are plotted in a space defined by the "Highbrow-Lowbrow" vertical axis and "Despicable-Brilliant" horizontal axis. Though it's a light-hearted way to list popular news tidbits, it is, at its core, a scatterplot.

Moving to scatterplots that are a bit more, shall we say, data-driven, these two scatterplots on page 253 show net immigration (defined as the number of people migrating into a region divided by the total number of migrants moving in and out of a region) plotted along the horizontal axis and per capita gross domestic product (GDP) along the vertical axis. The version on the left uses a single color with a slight transparency (or "opacity") so the reader



THE APPROVAL MATRIX *Our deliberately oversimplified guide to who falls where on our taste hierarchies.*

Venice sinks.

Mike Bloomberg's **curiously** timed apology for stop and frisk.

AI-generated art by Paris-based collective Obvious doesn't do well at auction. **Obviously.**

The $1,250 Basquiat-**themed watch.**

Disney is **demolishing** an entire block of historic buildings in West Soho for its big, new, meh HQ.

Elon Musk's Starlink satellites are **blocking** astronomers' view of the stars.

The editor of the Atlanta *Journal-Constitution* fact-checks the portrayal of the press in Clint Eastwood's *Richard Jewell*, says it "often differs from **reality.**"

Stephen Miller revealed to be a … **white nationalist?** For real?

Prince Andrew's attempt to **explain** his royal self on the BBC.

College students demand college journalists stop making **journalism.**

Famous author Sarah Dessen and her famous-author pals drag a former Northern State University student for **snobbish** comments about her YA work.

President Obama's **side-eye** at "certain left-leaning Twitter feeds."

Lydia Davis's *Essays One*: proof that books are still the **best** mobile devices.

The **slow-cook** impeachment. (Sondland has the receipts.)

The **well-named** Brice Marden show at Gagosian: "It reminds me of something, and I don't know what it is."

The decadent blitheness and high-minded **spectacle** of the Met's *Akhnaten.*

The meticulous, beautiful, and **oh-so-very experimental** *History of Violence* at St. Ann's Warehouse.

Peter Dinklage's swashbuckling return to the stage in the **moody,** musical *Cyrano* at the Daryl Roth Theatre.

Hannah Cullen's Young Dance Collective's **teen-angst** piece *Everything I Was Never Taught* at New York Live Arts.

Flea's memoir and **child-rearing** guide, *Acid for the Children.*

Cruisy holiday art gift book: *Peter Berlin: Icon, Artist, Photosexual.*

Joe Biden is too worried that pot is a **gateway** drug to sign on to legalize it.

Trump sours on e-cig ban for fear of losing the **vaper vote.**

Trying to watch the **brutal efficiency** of the Houston Rockets.

Pete Davidson is dating Cindy Crawford's **18-year-old daughter,** Kaia Gerber?

The way to a single guy's heart is through **chicken parm,** the *New York Post* reports.

Two Arkansas chemistry professors are accused of cooking **meth** in a school lab …

… And whoever thought of South Dakota's "**Meth.** We're on It" PSA definitely was.

Vogue bleeps out Cardi B saying the word *fart* in a video interview. Censorship!

23,000 NYCHA tenants were without **heat and hot water** during the cold weather.

The **churro** vendor is now a public menace?

TWA Hotel is opening a **jet-fuel-scented** runway ice-skating rink.

Netflix **hires** Robert Towne to co-write a prequel series to *Chinatown.*

The new Milk Bar flagship at the Ace Hotel. ***Mooooo!***

Diana Vreeland's great-granddaughter Caroline sings, with **Pomeranians** in the audience, at the Top of the Standard.

Someone **dyes** the fountain in front of 1251 Sixth Avenue green …

Baby Yoda.

… Probably unrelated: Kermit the Frog pivots to **fashion** influencer.

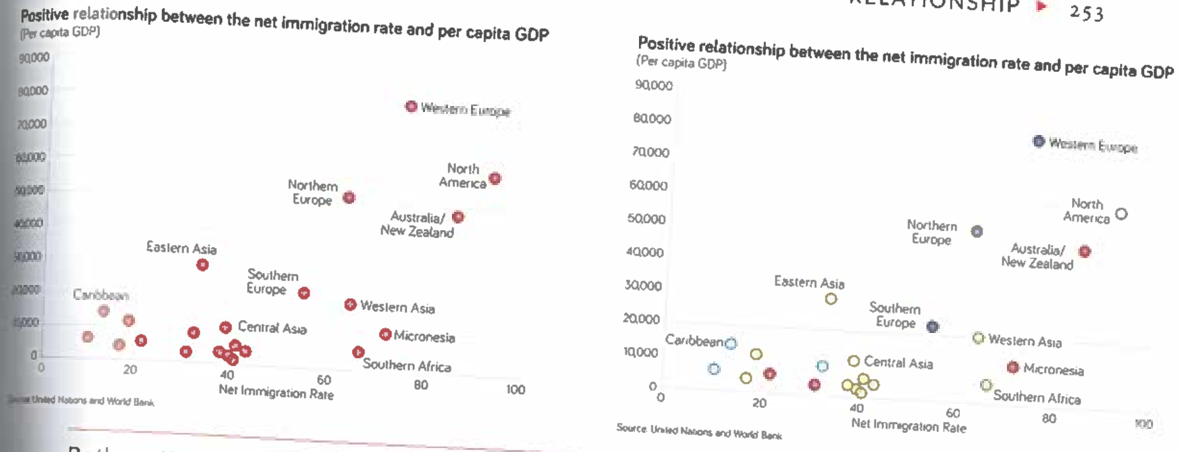
Now that your house is decluttered, Marie Kondo wants to **sell you** "joyful" objects via her online store.

Soon on sale: Spot, the **terrifying** robo-dog! (Laser guns not included.)

The **red-panda** babies at Prospect Park Zoo!

132 NEW YORK | NOVEMBER 25–DECEMBER 8, 2019

New York magazine's "Approval Matrix" is a scatterplot. Not truly data driven, but a scatterplot nonetheless.



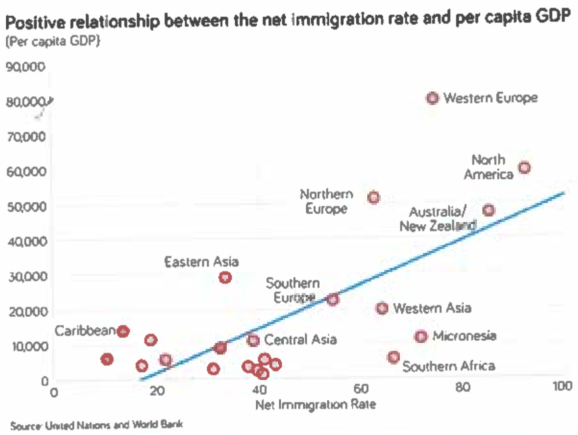
Both scatterplots show the association between net immigration and per capita GDP, using either a single transparent color (left) or different colors for regions of the world (right).

can see overlapping values. The same transparent effect is used in the scatterplot on the right, but this time colors capture the different regions of the world.

A scatterplot can help the reader see whether two variables are associated with one another. If the two variables move in the same direction—to the right along the horizontal axis and up along the vertical axis—they are said to be *positively correlated.* In other words, when both variables get bigger or smaller simultaneously, they are positively correlated. If they move in opposite directions, they are said to be *negatively correlated.* And if there is no apparent relationship, then they are not correlated (see Box on the next page). In the two scatterplots above, you get a visual sense that the two metrics are positively correlated—that net immigration is higher for regions with higher per capita GDP, in particular Western and Northern Europe, Australia/New Zealand, and North America.

One way to make the correlation even clearer is to add what statisticians call a *line of best fit* to the scatterplot. These are also called "regression lines" or "trendlines," and they show the general direction of the relationship. The statistical calculations to create lines of best fit are beyond the scope of this book, but the point is that you can make it even clearer to the reader in what direction (and to what magnitude) the two variables are correlated by calculating and including this line.

While the scatterplot is becoming a more common chart type, readers may still have difficulties reading and understanding it. A 2016 Pew Research Center survey showed that about 60 percent of people could correctly identify what they were seeing in a scatterplot.



Positive relationship between the net immigration rate and per capita GDP

Source: United Nations and World Bank

A *line of best fit* visualizes the correlation between the two variables.

CORRELATIONS

You have probably heard the old adage that "Correlation does not imply causation." We hear this so often because people regularly assign a causal relationship between variables that is actually coincidental. People eat more ice cream when it's hot outside, but that doesn't mean that more ice cream consumption *causes* the temperature to rise. As you look at your data and visualize the relationship between the observations within, be careful to understand when something may be correlated and when it might be causal. The less we know, the more we observe correlations and not causation.

Correlation is a measure of the strength of the linear association between two quantitative variables. The most common measure of correlation is the *Pearson correlation coefficient*, which measures the linear association between variables and is typically denoted with the Greek letter rho (ρ).

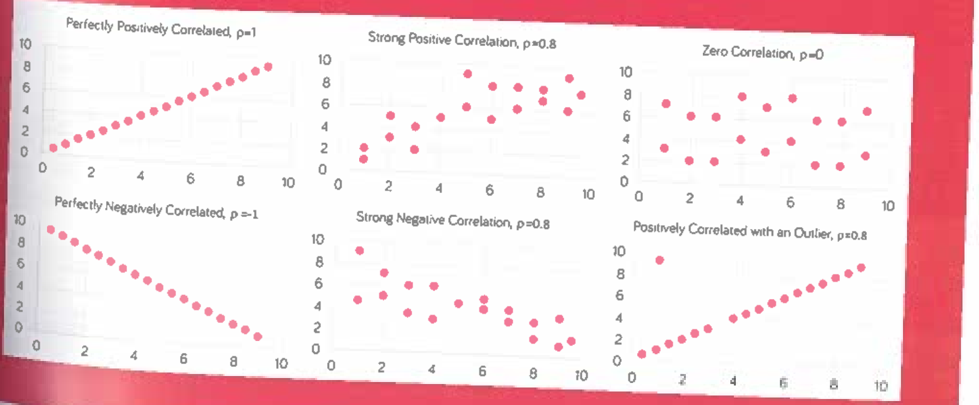
The sign and value of a linear correlation coefficient describes the direction and magnitude of the relationship between the variables. The value of the correlation coefficient ranges between −1 and +1. Values of −1 signal a perfectly negative correlation, +1 signals a perfectly positive correlation, and 0 represents no linear correlation. Positive correlation coefficients denote a positive correlation, which means that if one variable gets bigger, the other variable also gets bigger; negative coefficients mean the variables move in opposite directions, one variable gets bigger as the other gets smaller.

This discussion is related to these linear associations (or relationships) and it is also possible for two variables to have a *nonlinear* relationship. A linear association is a statistical term that describes a straight-line relationship between one variable and another. A simple example is how we might calculate distance as rate times time. In this case, if we were driving sixty miles per hour for two hours, we would travel 120 miles. The driving speed does not change over time, so the relationship is linear.

A nonlinear association, by comparison, refers to patterns in data that curve or break from the straight linear trend. Consider, as an example, the profit a company makes from a new product. When it is first released, there is little competition and sales grow. As sales continue to rise, public awareness increases, and profits start to roll in. Competing companies then start making their own version of the product and prices for the original fall to keep up, so profits decline. The company then develops a new version, and the whole cycle starts again.

As you can see in the images below, the two data values lie on a single diagonal line when they are perfectly positively or negatively correlated. When the two values move in tandem, either up or down, they are said to be positively or negatively correlated. The data in the bottom-right graph demonstrates the impact outliers have on this measure of correlation—moving a single point to the top-left part of the graph reduces the correlation from +1.0 to +0.8.

These visuals reinforce the importance of looking at our data as we conduct our analysis. Data visualizations help us not only communicate our work to our readers, but also enable us to explore our data. They can reveal patterns and relationships we wouldn't otherwise see. It's important not to leave this to the end of the workflow.



Perfectly Positively Correlated, ρ=1

Strong Positive Correlation, ρ=0.8

Zero Correlation, ρ=0

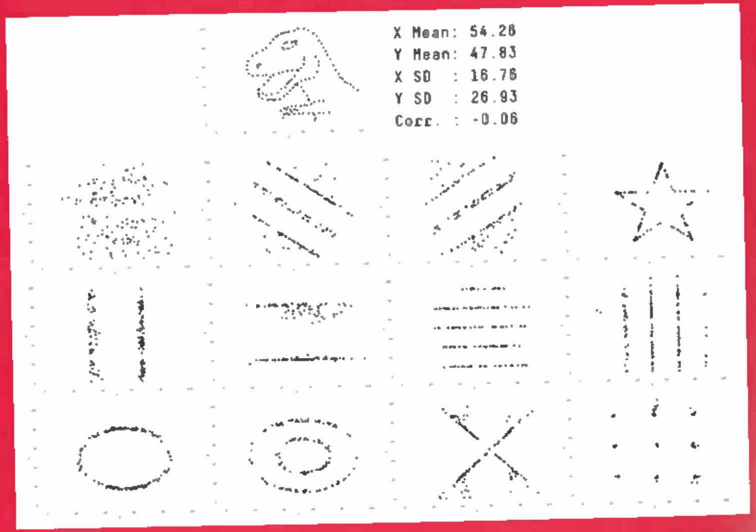
Perfectly Negatively Correlated, ρ =-1

Strong Negative Correlation, ρ=0.8

Positively Correlated with an Outlier, ρ=0.8

In 2016, University of Miami journalism professor Alberto Cairo drew a dinosaur with points in a scatterplot and dubbed it the "Datasaurus." His goal was to show the importance of visualizing your data in the exploratory phase. Imagine if you were teaching a data visualization class and asked your students to draw a scatterplot with 142 points, an average *x* value of 54.26, an average y value of 47.83, accompanying standard deviations, and a Pearson correlation of −0.06. Do you think anyone would draw a dinosaur?

In a 2017 paper, researchers Justin Matejka and George Mitzfaurice took the "Datasaurus" one step further and generated twelve alternatives that maintained the same summary statistics (mean, standard deviation, and correlation). The message of Cairo's "Datasaurus," Matejka and Fitzmaurice's paper, and Anscombe's quartet from Chapter 1, is that we should never rely on summary statistics alone but also on visuals of the data.



Source: Matejka and Fitzmaurice, 2017

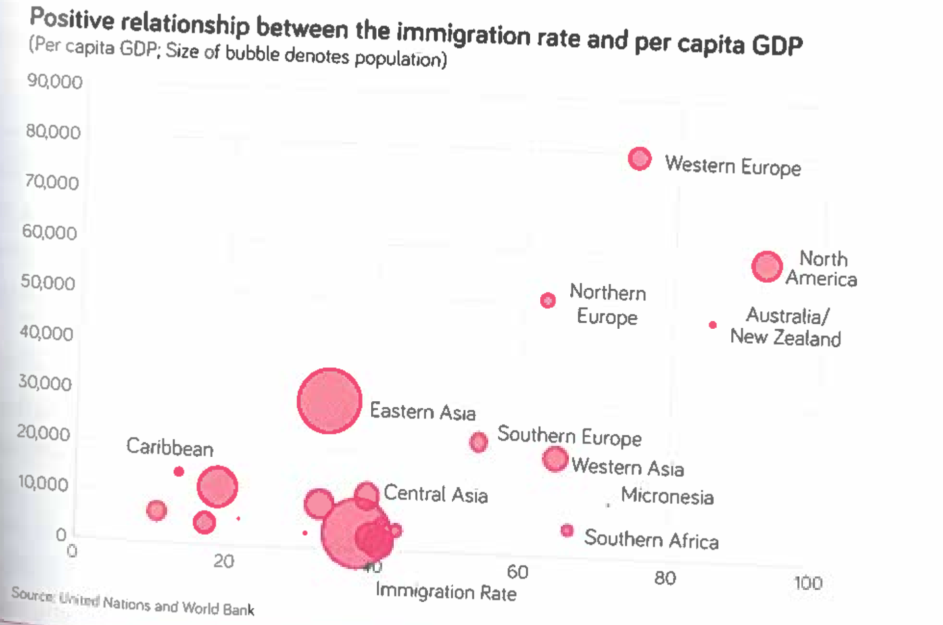
BUBBLE PLOT

The scatterplot can be transformed into a bubble plot (or bubble scatterplot) by varying the sizes of the circles according to a third variable. The data points don't have to be circles, they can be any other shape that doesn't distort our perception of the data. As mentioned in the section on bubble charts, the circles should be sized by area, not radius (see page 123). Color can help group or highlight certain points or direct the reader's attention to different parts of the graph.

The circles in this bubble plot are scaled according to the population of each region. The same positive relationship is still evident, but you can now see the relative size of each area.

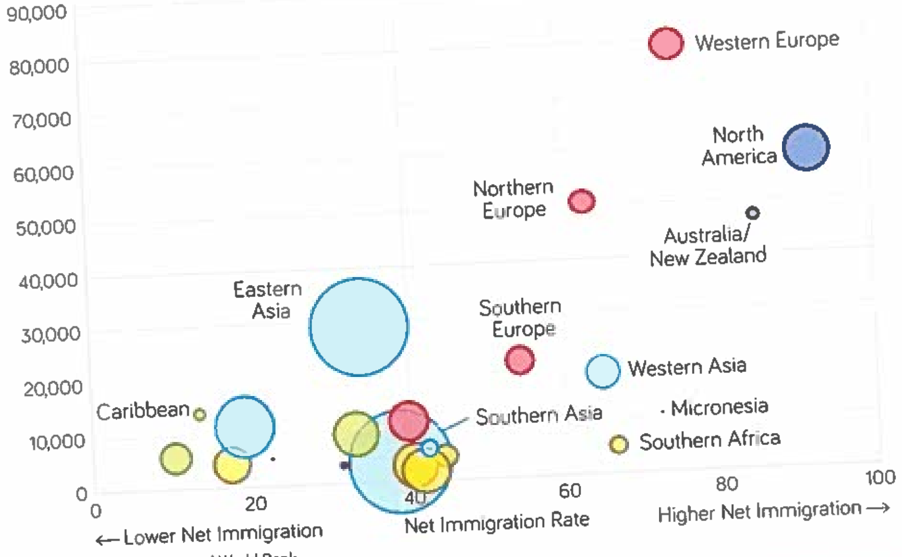
Because it is a more uncommon graph, be especially mindful of how labeling and annotation can guide readers through the chart and its content. One strategy is to label each axis and the direction of the change along the axis. The bubble plot on the next page has a centrally located horizontal axis that reads "Net Immigration" and two other labels, "Higher Net Immigration" and "Lower Net Immigration."

To further guide the reader, we could add a 45-degree line, on which the values are equal. We could also highlight specific points with color or outlines, or we could add text to explain what a point or set of points means. Properly labeled, these elements can lead the reader through the graph and content. Even people who know how to read scatterplots can struggle for a moment to understand what is going on when there are lots of points.



A bubble chart adds a third variable to the typical scatterplot. Here, the size of the circles corresponds to the population in each region.

Positive relationship between the net immigration rate and per capita GDP
(Per capita GDP; Size of bubble denotes population)



Source: United Nations and World Bank

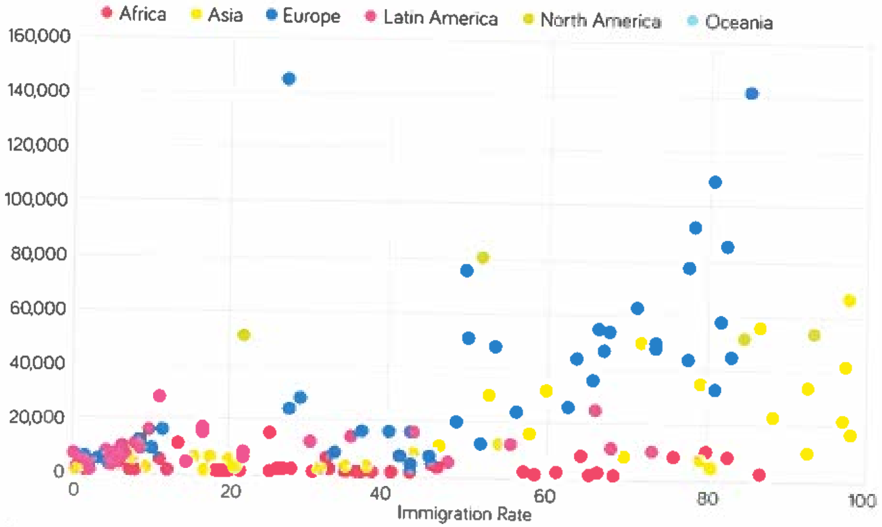
As before, more colors can be added to denote another variable, such as region of the world.

Labeling certain points or groups of points—with text, color, or enclosing shapes—can help the reader navigate the chart and draw their attention. The plot above added color to denote regions of the world. The next two employ the same strategy and include the more than two hundred countries around the world. Using color like this lets the reader identify certain regions. If instead we want to highlight one specific region, we might use a single color for a region of interest and use gray to push the others to the background.

Two final points about scatterplots:

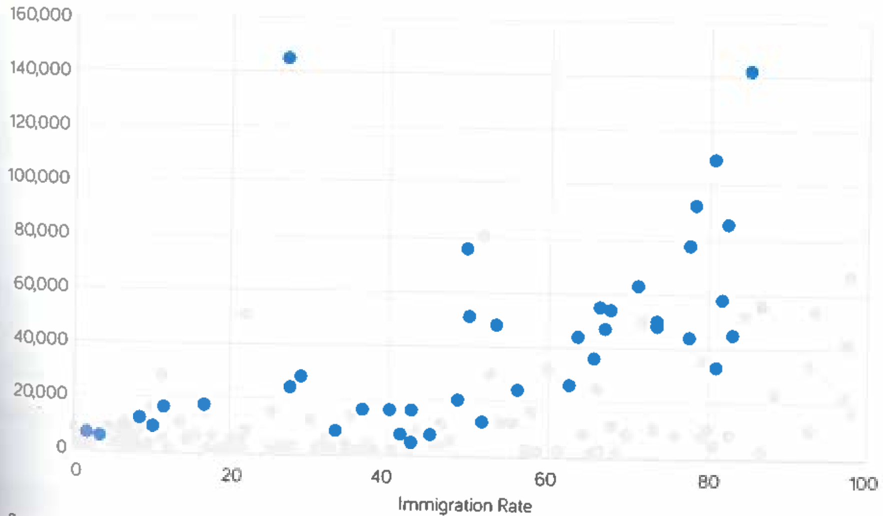
First, you will often see scatterplots that include labels for every single point, like the one on page 260. The end result is overwhelming clutter, with overlapping labels that are impossible to read. Luckily, we are far beyond the time where labels are the only way to convey information. If you believe there are readers who want to know the exact position of some of the points you didn't explicitly label, you can post a data file online, or create an interactive version using a tool like Tableau or PowerBI. Many academic researchers, for example, have an author page or webpage on their university website, as do many academic journals. These are great places to post the underlying data for your graphs.

Positive relationship between the immigration rate and per capita GDP
(Per capita GDP)



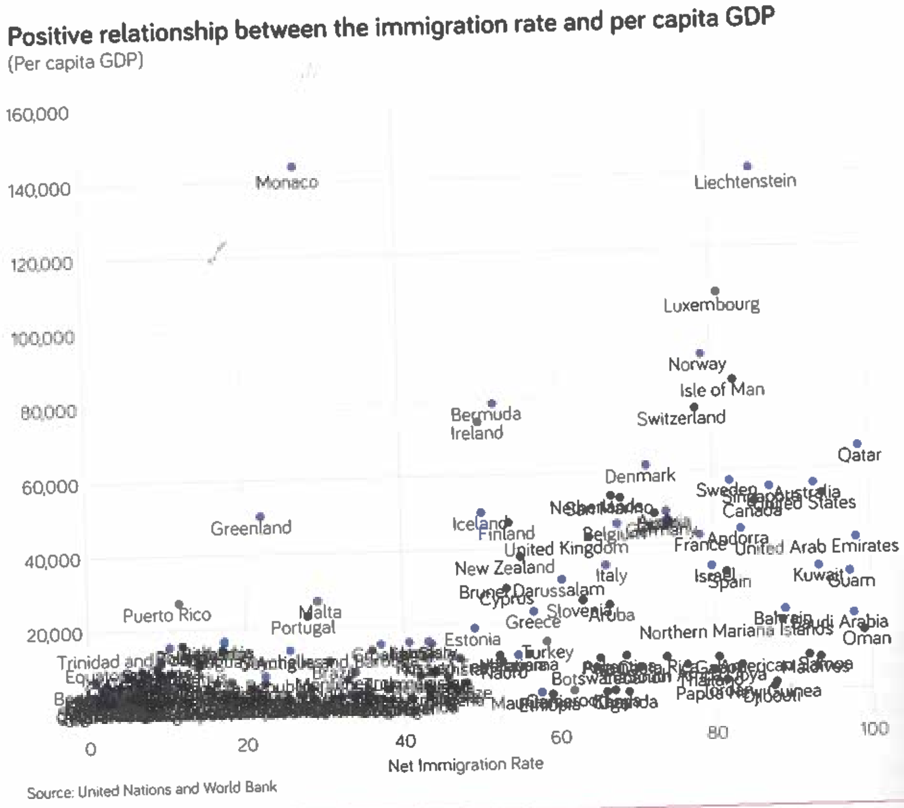
Source: United Nations and World Bank

European countries tend to have higher per capita GDP and immigration rates
(Per capita GDP)



Source: United Nations and World Bank

As we've seen elsewhere, color can be used strategically to highlight different groups (for example, regions of the world as in the top graph) or to highlight a single group or data point (as in the bottom graph).

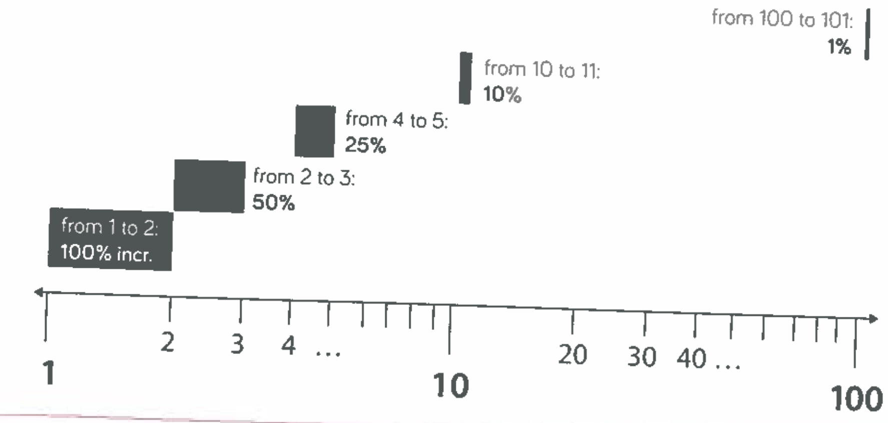


There's rarely a case where labeling all of the data points is necessary. Reduce clutter like this so your readers can better see the data.

Second, there may be times when normalizing your data, calculating percent change, or taking the logarithm of your data may improve the visual clarity of your graph. This is especially true when our data are clustered too densely in a visual. The logarithm (or log for short), is, simply put, an exponent written in a different way. Using mathematical laws of exponents, the log transformation shows relative values instead of absolute ones. Visualizing log data can make highly skewed distributions appear less so.

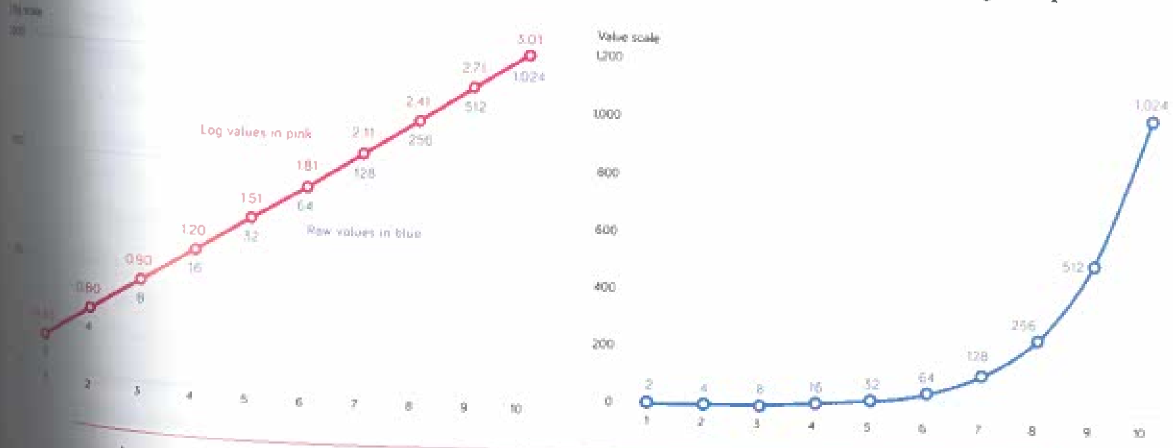
In a log scale, the fact that 101 minus 100 and 2 minus 1 are the same doesn't matter. Instead, what matters is that going from 100 to 101 is a 1 percent increase and from 1 to 2 is a 100 percent increase. Thus, on a log scale, going from 100 to 101 is about 1 percent of the distance as going from 1 to 2.

Another way to understand the differences between absolute (level) and relative (log) values is to graph some sample data. The graph on the left shows a simple doubling of each number—2, 4, 8, 16, 32, and so on. In this linear scale, we see what is called an "exponential" curve as the difference between each sequential value gets farther and farther apart. By contrast, in a logarithmic plot (the graph on the right), each gridline represents a tenfold increase over the previous one: 1, 10, 100, and 1,000. In this representation, the same numbers appear as a straight line as opposed to a curved one, even though the growth rate is the same.

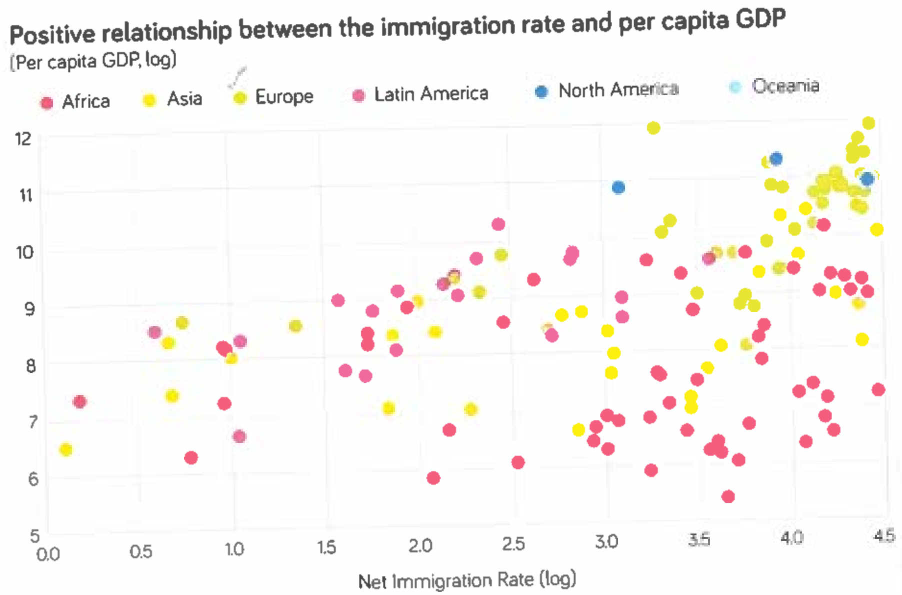


On a log scale, going from 100 to 101 is about 1 percent of the distance as going from 1 to 2. Based on Lisa Charlotte Rost (2018).

In the GDP-immigration scatterplot, there are many countries clustered around the origin, with both low per-capita GDP and net immigration rates. By taking the logs of both variables, the data are more spread out across the graphic space. The tradeoff is that logs (or other transformations, for that matter) are not immediately intuitive. Knowing that per



Logarithmic values are useful to show relative values rather than absolute values.

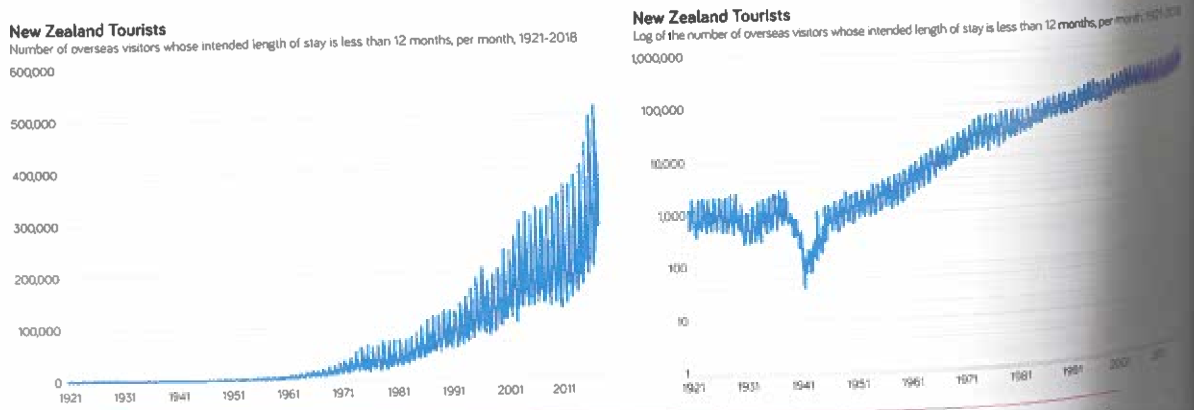


Source: United Nations and World Bank

By taking the logs of both variables, the data are more spread out across the graph.

capita GDP in Luxembourg is $107,865 (in U.S. dollars) is a number we can all understand, but we don't as easily grasp that if we write it as the log per capita GDP: $11.59.

Here's another example of how to use a log scale. This one uses time series data so we can see *relative* changes over time. Lisa Charlotte Rost, a designer and blogger at the online data visualization tool Datawrapper, used New Zealand tourism data to demonstrate how the log transformation works and how it can affect our view and understanding of data showing changes over time. In the line chart on the left, she's plotted the number of tourists per month from 1921 to 2018. We can see the relatively flat pattern from the beginning of the period to about 1970, when the number of visitors starts to rise. The version on the right uses log values and thus shows the *relative* number of visitors. Here, we can see a clear drop in the early 1940s during World War II. In the graph on the left, there isn't as clear a decline in *absolute* numbers (the number of tourists fell from about 2,000 in early 1939 to fewer than 100 in 1942) but there was a sharp decline in *relative* numbers.



We can't see the drop in the *number* of visitors to New Zealand during World War II, but when we convert the data to logs, the *change* becomes apparent. Based on Lisa Charlotte Rost (2018).

Whether it's appropriate to transform your data is largely a function of the question you want to answer. Are you after relative or absolute values? Percent changes or levels? There is no right or wrong answer to this question, but each has their tradeoffs.

PARALLEL COORDINATES PLOT

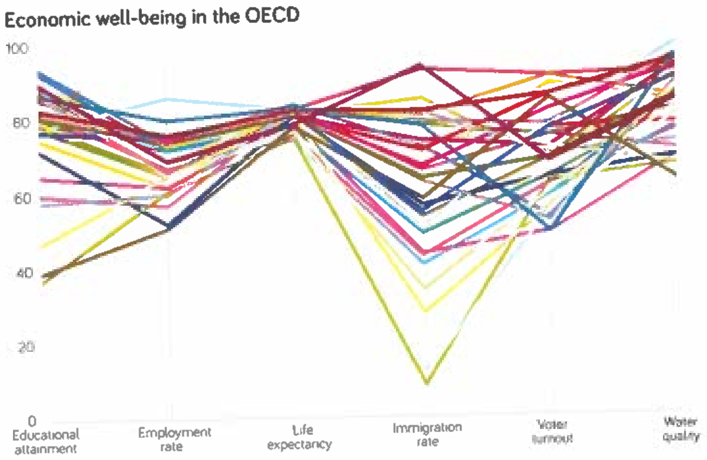
A scatterplot has data along two variables represented by a horizontal axis and vertical axis. Sometimes, though, we have more variables to visualize. That's where the parallel coordinates plot comes in.

In these charts, the data values are plotted along multiple vertical axes and connected by lines. As in the scatterplot, the axes can have different units of measurement, or they can be normalized—for example, as percentages—to keep the scales uniform. Thus, instead of visualizing a single correlation between two variables, the parallel coordinates plot permits multiple correlations within a single view.

As an example, the parallel coordinates plot on the next page shows correlations across six different variables related to migration for thirty-two countries around the world. Each vertical axis represents a different variable—like educational attainment, employment rate, and life expectancy—and each line represents a different country.

But this graph is really, really hard to read! There are too many lines, all different colors, and all crossing at different places. But before we try to address the challenges of the full parallel coordinates plot, let's try simplifying.

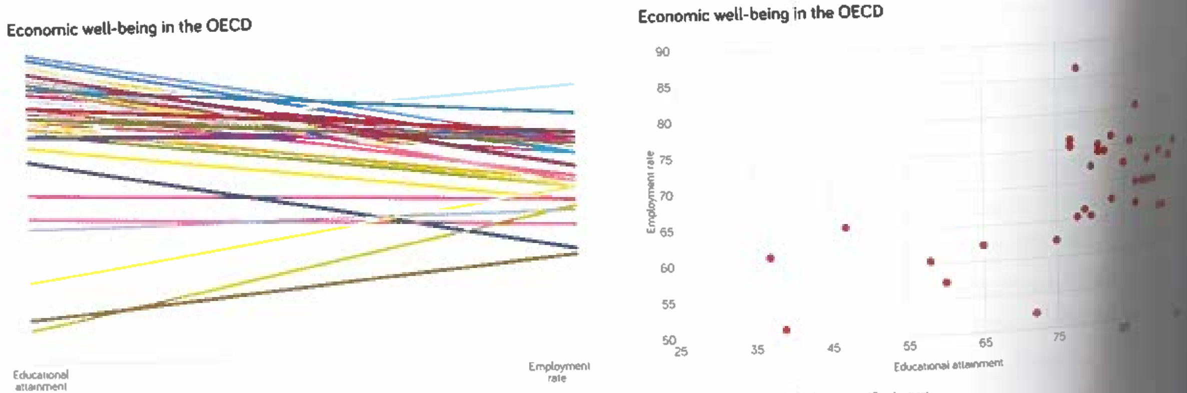
If we zoom in for a moment on the first two axes, we see perhaps the simplest parallel coordinates plot, one that resembles a slope chart. (I differentiate this from the slope chart in Chapter 5 because slope charts show changes over time, while parallel coordinate plots



Source: Organisation for Economic Co-Operation and Development

The parallel coordinates plot shows correlations between two or more variables across multiple vertical axes.

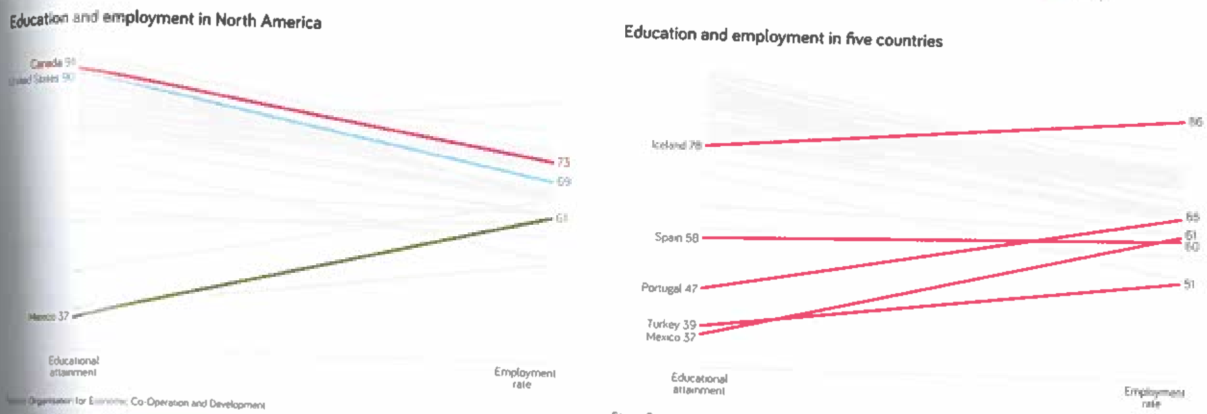
are used to compare different variables.) Here, I plot the relationship between educational attainment and the employment rate. Because the lines (countries) at the top of the left axis (education) are also near the top of the right axis (employment rate), these two variables are positively correlated. (It's not that the lines slope down, but the relative position of the points on each axis.) This is also clear in the scatterplot shown to the right. As always, which chart you use depends on your purpose and audience.



Source: Organisation for Economic Co-Operation and Development

Source: Organisation for Economic Co-Operation and Development

It's much easier to see the relationship between two variables in a parallel coordinates plot with just two axes (similar to a slope chart). An alternative visual approach is the scatterplot.



Source: Organisation for Economic Co-Operation and Development

Source: Organisation for Economic Co-Operation and Development

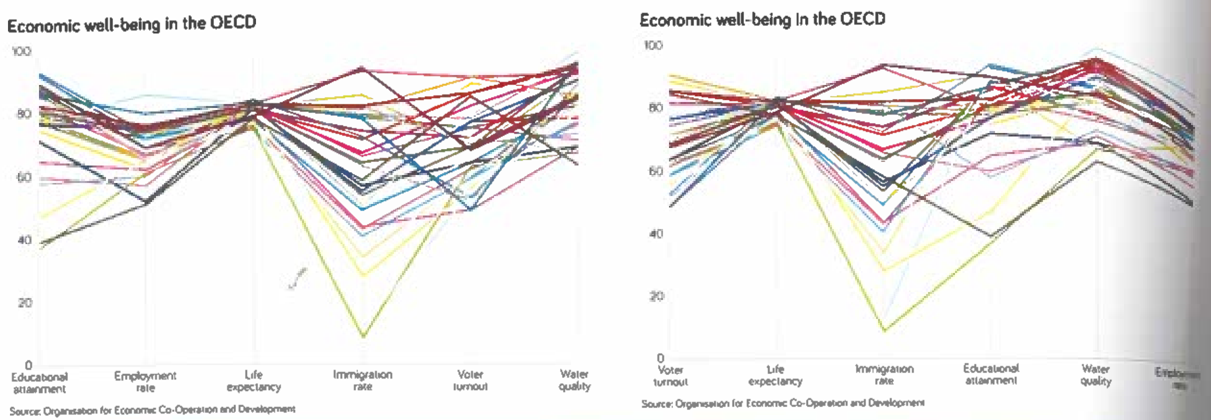
As with the slope chart, you can use different colors, line thicknesses, or other visual elements to highlight areas or values.

As with the slope chart, you can use different colors, line thicknesses, or other visual or textual elements to highlight certain areas or values. You could, for example, highlight the North American countries (left graph) or maybe just those lines that are upward sloping (right graph).

Back to the parallel coordinates plot with all six variables shown at the top of the next page. Now that you understand how to read the chart, you can see the positive correlation between education and the employment rate in the first two axes. You can also see the positive correlation between the employment rate and life expectancy in the second and third axes. Our view of the data and the specific correlations we can most clearly identify are a function of how we organize the axes. The plot on the right changes the order of the vertical axes so we can now see the positive correlation between voter turnout and life expectancy in the first two axes, which we could not see in the original plot.

Placing all six metrics on the same vertical range also has the effect of suppressing the range (or variance) in some of these measures. For example, life expectancy varies only slightly, from 74.6 years to 83.9 years, while net migration varies more widely, from 8.6 percent to 93.9 percent. Allowing the ranges along the axes to fluctuate from one to the next (as in the middle chart on the next page) requires more labeling along each axis, but it also gives a better view of the data. The advantage of the two plots at the top is that you don't need to label every line. The disadvantage is that you suppress the variation within each variable. Notice, however, that this parallel coordinate plot—in which the range of each axis differs—looks more variable.

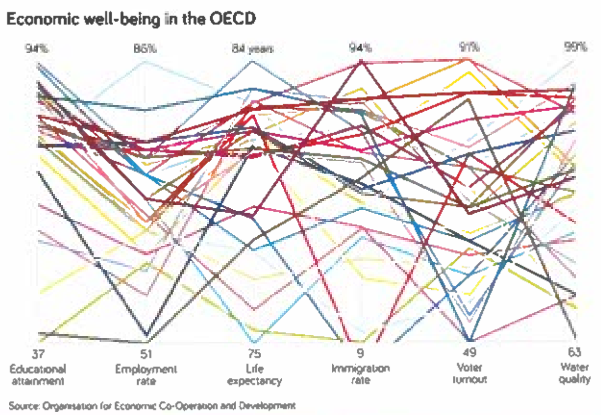
In sum, the challenge with many parallel coordinates plots is that they quickly become cluttered. With lots of observations (lines) and multiple axes, readers may have trouble finding the



Source: Organisation for Economic Co-Operation and Development

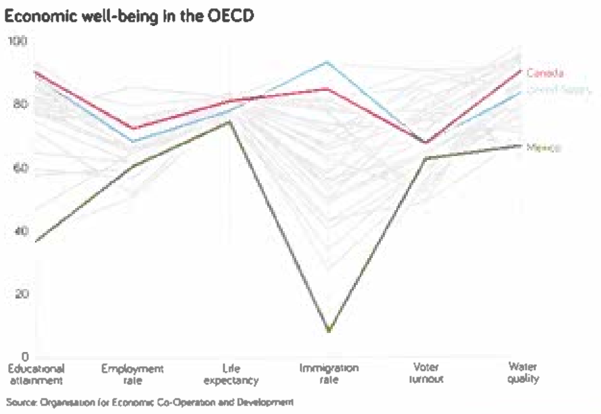
Source: Organisation for Economic Co-Operation and Development

Parallel coordinates plots with too many observations quickly become cluttered.



Source: Organisation for Economic Co-Operation and Development

The axes in parallel coordinate plots can differ based on the minimum and maximum of the metric.



Source: Organisation for Economic Co-Operation and Development

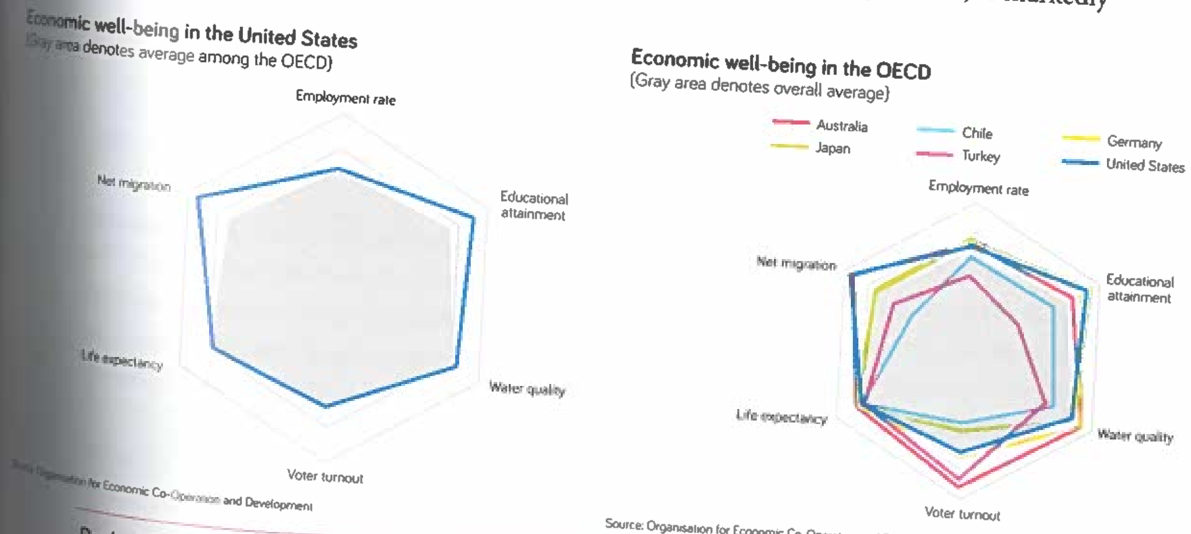
As we've seen previously, one way to simplify dense graphs like these is to use the "start with gray" strategy and add color to only a select number of observations.

correlations and picking out specific values. One way to alleviate this difficulty is to remember the "Start with Gray" guideline: Color a group of lines gray and highlight just a subset of the data.

RADAR CHARTS

Radar charts are like parallel coordinate plots, but the lines wrap around a circle instead of being arranged parallel to one another. These are also sometimes called *spider charts* or *star charts*, and they're a good way to show multiple comparisons within a relatively compact space. Data values are plotted along separate axes that radiate from the center (the axes themselves may or may not be shown) and are connected by lines or areas to show the relationships between the different variables.

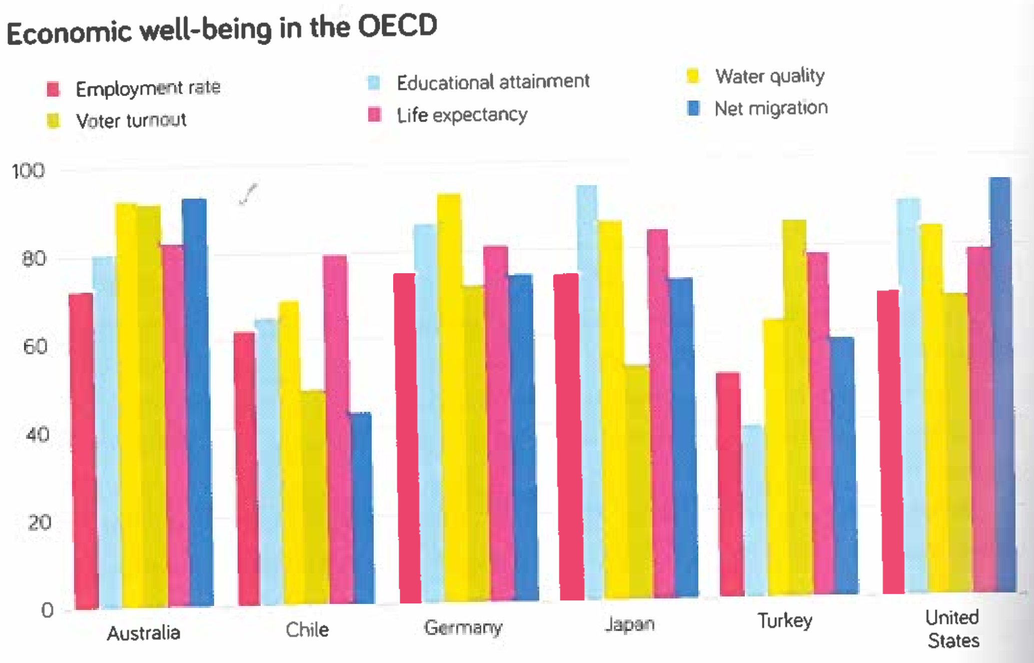
The radar chart on the left shows the same six variables used above—the line for the United States and the gray area behind it the average for the thirty-two countries shown earlier. The version on the right shows the same variables for those six countries as well as the overall average in the gray. Both charts are compact and especially good at highlighting outliers. You can quickly and easily see the shape for Turkey (the pink line) is markedly



Source: Organisation for Economic Co-Operation and Development

Source: Organisation for Economic Co-Operation and Development

Radar charts are like parallel coordinate plots, but the lines wrap around a circle instead of being arranged parallel to one another. The gray interior area represents the overall average for each metric.



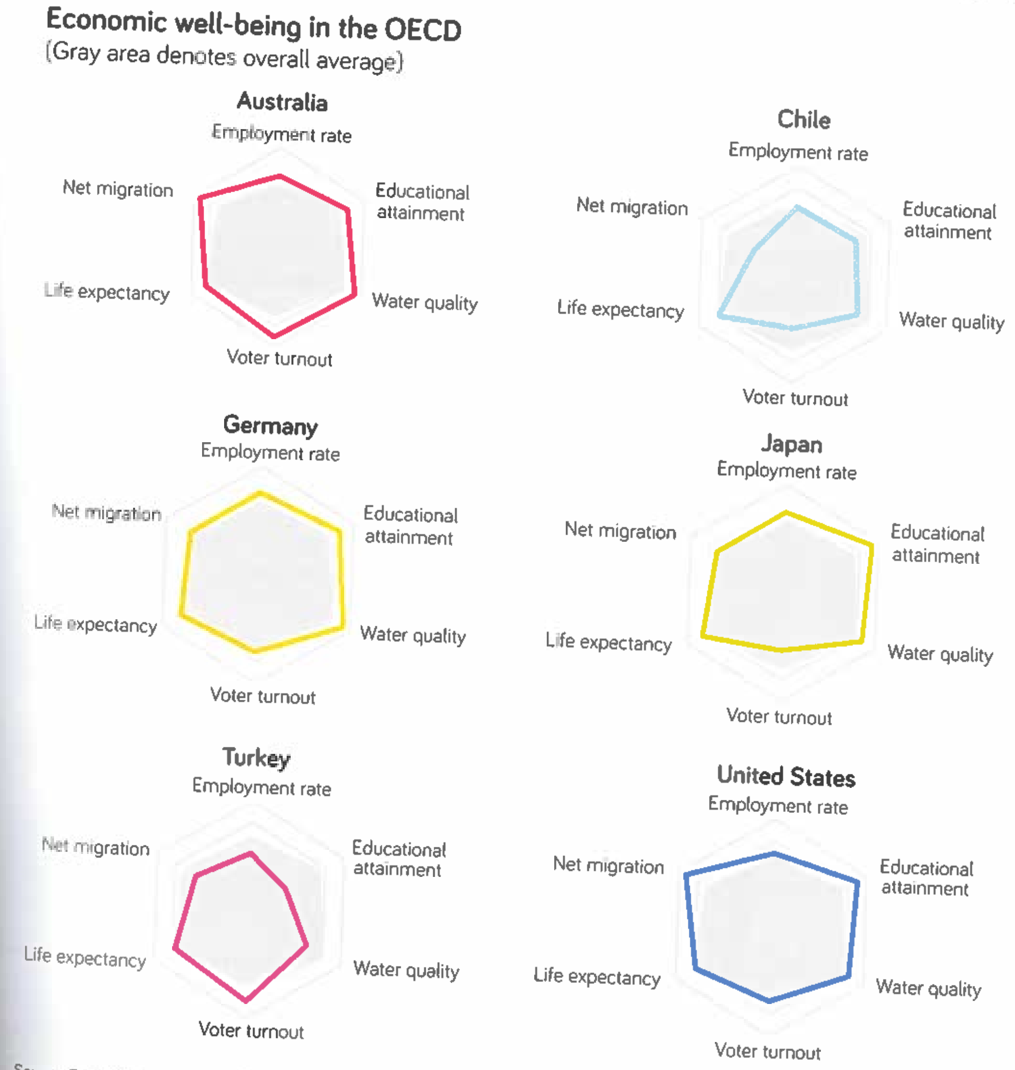
Source: Organisation for Economic Co-Operation and Development

Too many bars in a bar chart like this make it hard to pick out specific observations or patterns.

different than the other countries. It's much harder to make that observation when the data are arrayed in the paired bar chart above.

As with many charts, the radar chart gets more complicated as more lines are added, and the crossing pattern around the circle can make perception even more difficult. As was the case with the parallel coordinates plot, plotting different metrics can also make perception difficult because it requires some normalizing or other modification of the data values—in the multi-country radar chart above, again notice how the values for life expectancy bunch together.

Another strategy is to use the small multiples approach, in which a separate radar chart is created for each country or group. In this case, the small multiples version takes up more space than the original and doesn't necessarily allow easy comparison across specific countries. But it is easier to see the values for each country relative to the overall average.



Source: Organisation for Economic Co-Operation and Development

A small-multiples approach lets us see the values for each country relative to the overall average.

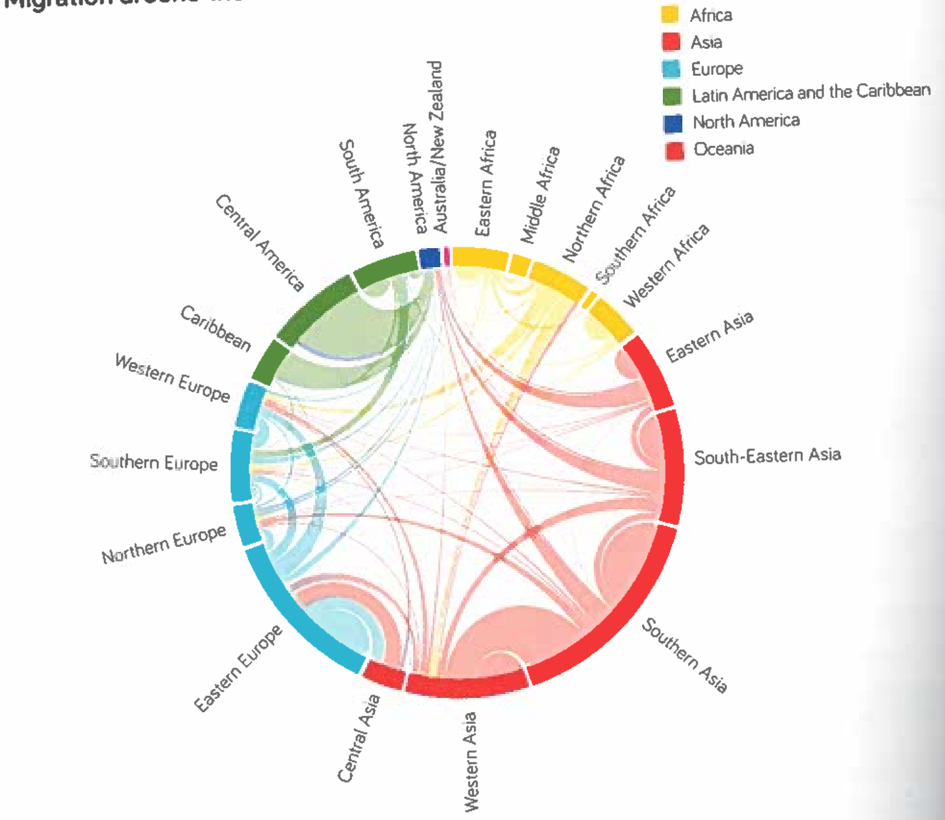
CHORD DIAGRAM

Like the radar chart, the chord diagram is another way to show associations or relationships between observations arrayed in a circle. It is perhaps best used to show how observations have shared characteristics. In chord diagrams, observations (called *nodes*) are located around the circumference of the circle and connected by arcs within the circle to illustrate connections. The thickness of the arcs—often also differentiated by color or the transparency of the color—represent the degree of the connection between the different groups.

This chord diagram uses the same migration data used so far in this chapter to show migration flows between major regions of the world in 2017. Each region is placed along the circumference of the circle and the bands emanating from each correspond to the number of migrants entering or leaving each region. There are more than 110 values plotted in this single graph—though you could pick out specific values in a (very large) table, the chord diagram is clearly more visual and spatially efficient.

In the first chord diagram, you can see the large migrant flows within Asia (the red areas), and the movement between Central and North America (the thick green line to the blue segment near twelve o'clock in the circle). One danger is that the graph can quickly become cluttered and hard for the reader to easily see relationships. Again, we can use the strategy of highlighting specific groups with colors or lines. I've done that in this chord diagram in which the Asian region is red with all other regions in gray. The complexity of the chord diagram (and

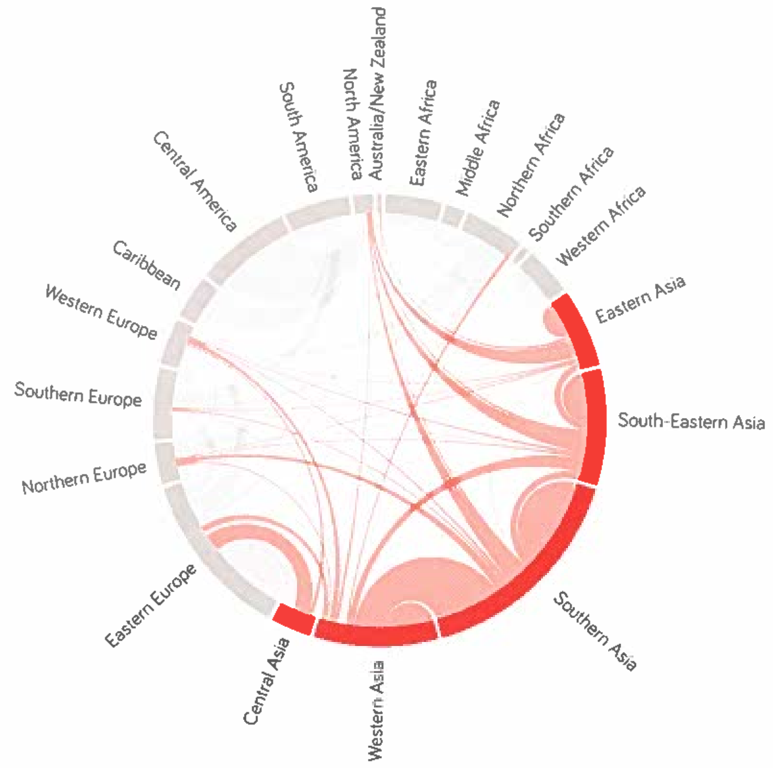
Migration around the world



Source: Organisation for Economic Co-Operation and Development
Note: Data limited to a minimum of 200,000 immigrants or emigrants

In chord diagrams, the observations are located on the circumference of the circle and arcs within illustrate the connections.

Migration from Asia



Source: Organisation for Economic Co-Operation and Development
Note: Data limited to a minimum of 200,000 immigrants or emigrants

Using color strategically—especially with the color gray—can draw attention to groups or points.

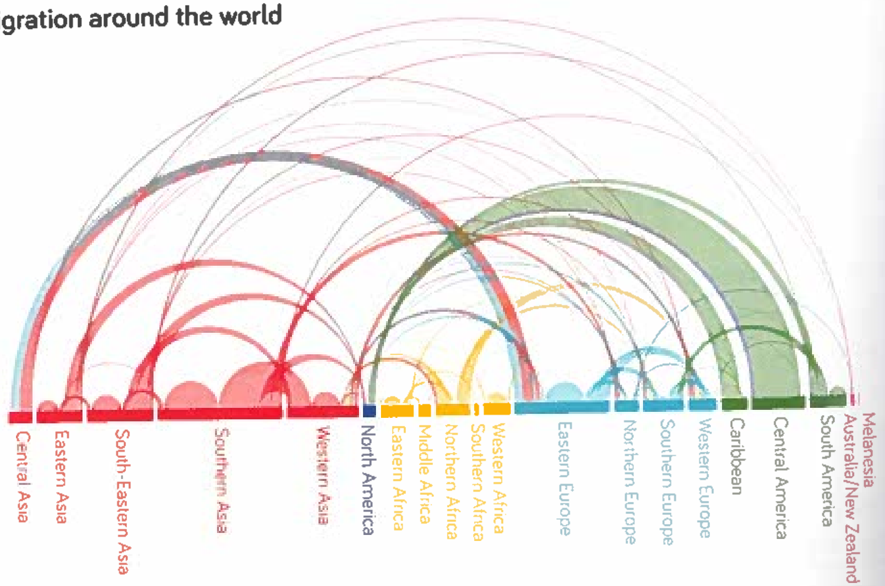
its relative compactness), make them visually intriguing and invites the reader to explore the data in more depth.

ARC CHART

Stretch a chord diagram out along a single horizontal axis and you have an arc chart. In this case, the nodes are placed along a line and are connected by arcing lines. The lines can vary in height, thickness, and color to illustrate the strength of the relationship or correlation. This arc chart shows the same migration flows between regions of the world as in the chord diagram.

A major consideration of the arc chart—and which also applies to many charts in this section—is that the order of the data can influence our perception of the results. Notice the high, wide green arcs stretching from North America to countries in the Caribbean, and Central and South America. If, by contrast, North America is placed to the far-right side of the graph and next to the other countries in the Western Hemisphere, the visualization is substantially changed. There is no longer a tall green arc dominating the view, but a series of red bands that reach across the entire space between North America and countries in Asia. Some of this is obviously a function of the colors used, but the arrangement of the countries also matters. It is worth taking time to experiment with color and node placement to arrange the arc chart in the way that best communicates your argument.

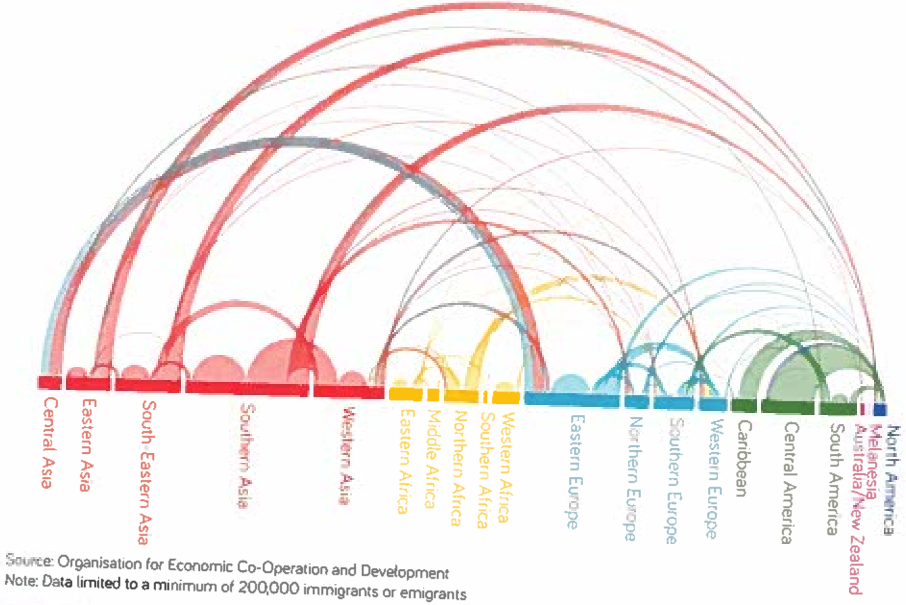
Migration around the world



Source: Organisation for Economic Co-Operation and Development
Note: Data limited to a minimum of 200,000 immigrants or emigrants

The arc chart is like a chord diagram stretched along a single horizontal axis.

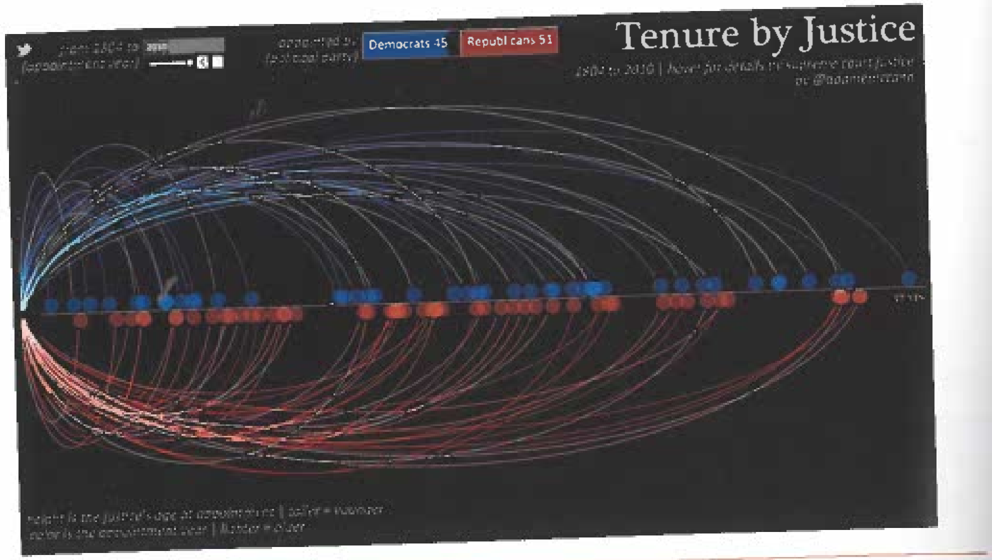
Migration around the world



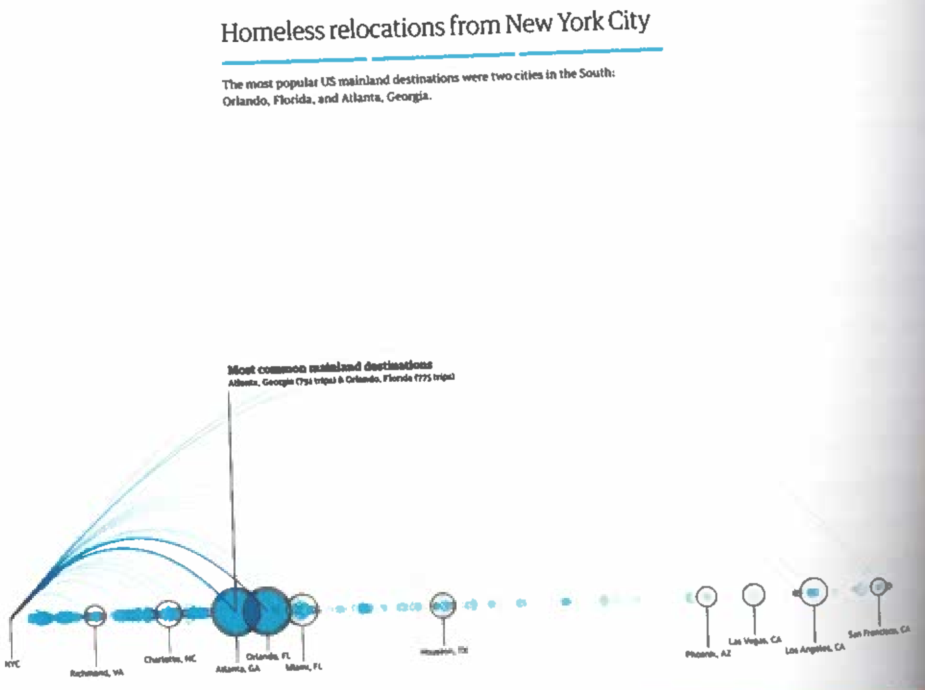
Source: Organisation for Economic Co-Operation and Development
Note: Data limited to a minimum of 200,000 immigrants or emigrants

Like many graphs in this chapter, the organization of the data along the horizontal axis can affect our perception of the data. Compare this arc chart to the previous arc chart—same data, different shape.

A variation on the arc chart is an *arc-time chart* or *arc-connection chart*, in which connections over time are plotted in the same way. Instead of illustrating the correlation or relationship between two distinct variables, the nodes denote time. The arc-connection chart can also be thought of as an alternative to a timeline or flow chart, which we saw in Chapter 5. On the next page, the arc chart from Adam McCann shows the tenure of all Supreme Court justices in the United States since 1804. The origin (the left-most point) shows when each judge started his or her tenure, and the arc stretches to their retirement age. The height of each arc represents the age at the time of appointment (taller is younger) and the color represents their political party and year they were appointed (lighter shades are earlier years). An



Adam McCann used an arc chart to show the tenure of all US Supreme Court justices since 1804.



Arc charts can also show geographic data, as in this one from the *Guardian* showing where New York City sends their homeless population.

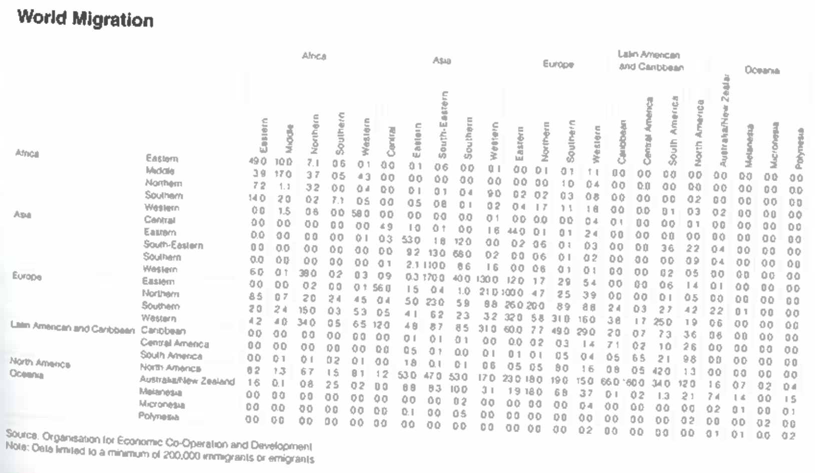
alternative chart type, like a bar chart or a heatmap, could be used to show the same changes, but there is something arresting about the shapes in this view.

Another way to use the arc chart is to plot distances. The arc chart from the *Guardian* shows where New York City sends their homeless population. The cities are organized by distance from New York—Richmond, Virginia, on the left and San Francisco, California, on right. The height and thickness of the arcs and the size of the circles show how many people go to each city.

CORRELATION MATRIX

A correlation matrix is a table with the variables listed along the horizontal and vertical axes. Numbers in each cell represent the strength of that relationship, often as a Pearson's correlation coefficient (see Box on page 254).

The *correlation matrix graph* uses the same layout but instead of numbers it uses shapes—often circles—to show the strength of the correlation, and sometimes color and shades to organize the table. The correlation matrix is a cousin of the heatmap, and we can also think of it as a way to add a visual element to a standard table, a topic we will visit in Chapter 11.

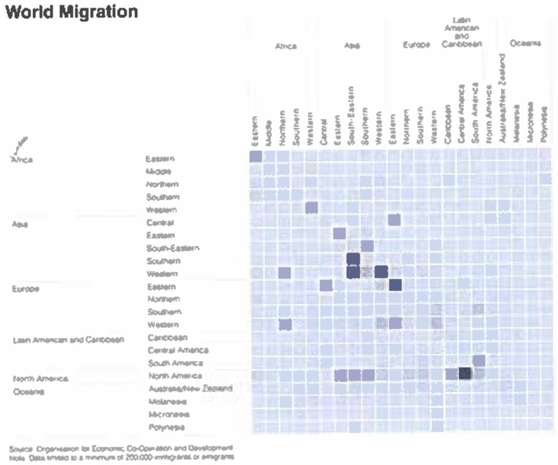


World Migration

		Africa					Asia					Europe				Latin American and Caribbean				Oceania			
		Eastern	Middle	Northern	Southern	Western	Central	Eastern	South-Eastern	Southern	Western	Eastern	Northern	Southern	Western	Caribbean	Central America	South America	North America	Australia/New Zealand	Melanesia	Micronesia	Polynesia
Africa	Eastern	490	100	7.1	06	01	00	01	06	00	01	00	01	01	11	00	00	00	00	00	00	00	00
	Middle	39	170	37	05	43	00	00	00	00	00	00	00	10	04	00	0.0	00	00	00	00	00	0.0
	Northern	72	1.1	32	00	04	00	01	01	04	90	02	02	03	08	00	00	00	02	00	00	00	0.0
	Southern	140	20	02	7.1	05	00	05	08	01	02	04	17	11	18	00	0.0	01	03	02	00	00	00
	Western	00	1.5	08	00	580	00	00	00	0.0	01	00	0.0	00	04	01	00	00	01	00	00	00	00
Asia	Central	00	00	00	00	00	49	10	01	00	16	440	01	01	24	00	00	00	00	00	00	00	00
	Eastern	0.0	00	00	00	01	03	530	18	120	00	02	06	01	03	00	00	36	22	04	00	00	00
	South-Eastern	00	0.0	00	00	00	00	92	130	680	02	00	06	01	02	00	00	00	09	04	00	00	00
	Southern	0.0	00	00	00	00	01	2.1	1100	86	16	00	06	01	01	00	00	02	05	00	00	00	00
	Western	6.0	01	380	02	03	09	0.3	1700	400	1300	120	17	29	54	00	00	06	14	01	00	00	00
Europe	Eastern	00	00	02	00	01	560	15	04	1.0	210	1000	47	25	39	00	00	01	05	00	00	00	00
	Northern	85	07	20	24	45	04	50	230	59	88	260	200	89	88	24	03	27	42	22	01	00	00
	Southern	20	24	150	03	53	05	41	62	23	32	320	58	310	160	38	17	250	19	06	00	00	00
	Western	42	40	340	05	65	120	48	87	85	310	60.0	77	490	290	20	07	73	36	08	00	00	00
Latin American and Caribbean	Caribbean	00	00	00	00	00	00	01	01	01	00	0.0	02	03	14	71	02	10	26	00	00	00	00
	Central America	00	00	00	00	00	00	05	01	0.0	01	01	01	05	04	05	65	21	98	00	00	00	00
	South America	00	01	01	02	01	00	18	0.1	01	06	05	05	80	16	08	05	420	13	00	00	00	00
North America	North America	82	1.3	67	15	81	12	530	470	530	170	230	180	190	150	66.0	600	340	120	16	07	02	04
Oceania	Australia/New Zealand	16	0.1	08	25	02	00	88	83	100	31	19	180	68	37	01	02	1.3	21	74	14	00	15
	Melanesia	00	00	00	00	00	00	00	00	02	00	00	00	00	04	00	00	00	00	02	01	00	01
	Micronesia	00	0.0	00	00	00	00	0.1	00	05	00	00	00	00	00	00	00	00	02	00	00	02	00
	Polynesia	00	00	00	00	00	00	00	00	00	00	00	00	00	02	00	00	00	00	01	01	00	02

Source: Organisation for Economic Co-Operation and Development
Note: Data limited to a minimum of 200,000 immigrants or emigrants

The basic correlation matrix is a table with numbers that show the strength of the relationship between observations.


World Migration

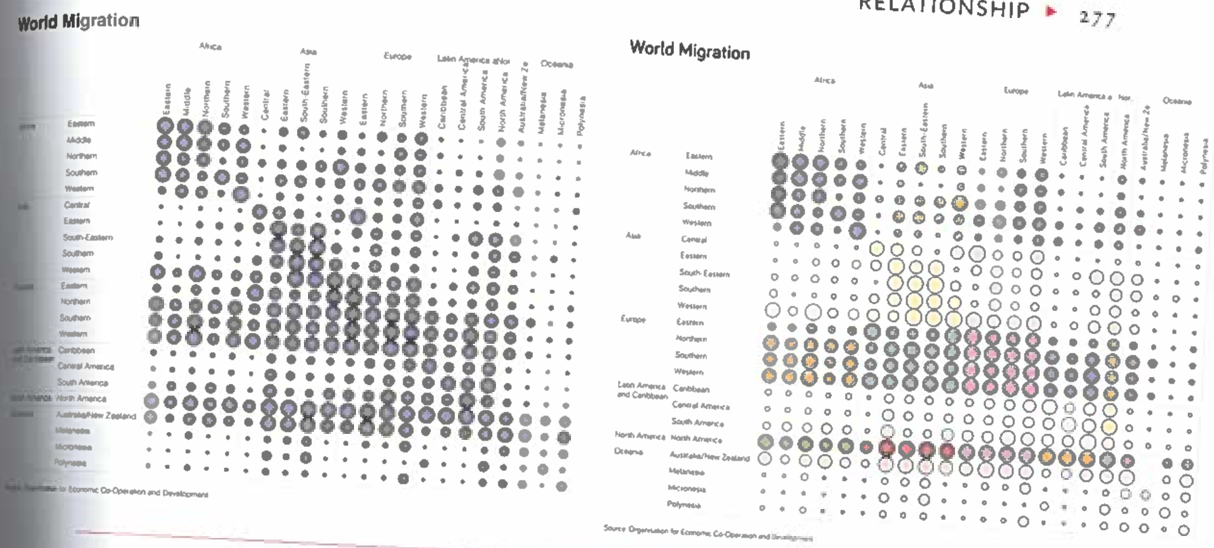
Source: Organisation for Economic Co-Operation and Development
Note: Data limited to a minimum of 200,000 immigrants or emigrants

A heatmap approach to the matrix makes the patterns clear without (in this case) showing the numbers.

These two matrices show the relationship between immigrants (those entering each region) and emigrants (those leaving each region) across the world in 2017. The view on the left is the standard correlation matrix shown as a table. This gives us all the detail we would need to understand the exact correlation between different variables, but it's difficult to navigate. There are a lot of numbers, and the important values don't stand out. The matrix on the right is a heatmap (see page 112), which loses the detail of the table version in favor of highlighting the stronger (positive) correlations, especially within Asia. We could include both the colors and numbers, but the view might end up looking cluttered and busy.

The next two visuals display the data as standard correlation matrix graphs. Circles represent the strength of the relationship, and color (in the version on the right) helps organize each area, though in this case there may be too many colors. It can be hard for the reader to clearly see differences because humans are not very good at assessing quantities from the sizes of circles. In both cases, the circles are sized to fit within each cell, but that doesn't necessarily need to be the case. We could make the circles larger to fill the entire space and use transparent colors when they overlap.

One final consideration with any correlation matrix or table is that the values along the diagonal are, by definition, equal to one. That is, migration between Eastern Africa and Eastern Africa is the same. This means they are often left out because they can visually dominate or clutter the visual.


World Migration

World Migration

Source: Organisation for Economic Co-Operation and Development

Source: Organisation for Economic Co-Operation and Development

An alternative to the correlation matrix table is to use circles or other shapes, to which color can be added to visually organize the space.

NETWORK DIAGRAMS

We now enter a class of graphs for which I use the term *diagram* instead of graph, plot, or chart, largely because some of the decisions about layout and structure are not always determined by math or the data but by what looks best and is most clear. These diagrams are used to show hierarchies and connections within and across groups and systems. The thickness of the lines and size of the points can be sized according to data values to signal the strength of those relationships, and arrows can visualize movements inside groups and communities. Consider a family tree: the lines show links between parents, siblings, spouses, and children, but the connecting lines and the pictures or names of family members are not scaled according to a data value.

We start with the standard network diagram, which shows connections between people, groups, or other units. Generally speaking, the points in a network diagram (called *nodes* or *vertices*) denote the individual person or observation, and the lines (called *edges*) link them together and show the relationship. The position of the nodes and the length (and sometimes thickness) of the linking lines illustrates the strength of the relationship. While nodes are often depicted as circles, you could also use icons, symbols, or pictures.

The ultimate appearance and organization of a network diagram depends on the kind of network we want to visualize and the method with which we arrange the nodes and edges. When creating a network diagram, we must be careful about how edges cross and nodes overlap. In general, we want to achieve some kind of visual harmony in the visualization by finding a uniform and meaningful length of the edges and some symmetry for the entire graph.

To start, we can distinguish between four different kinds of network diagrams:

1. *Undirected and Unweighted*

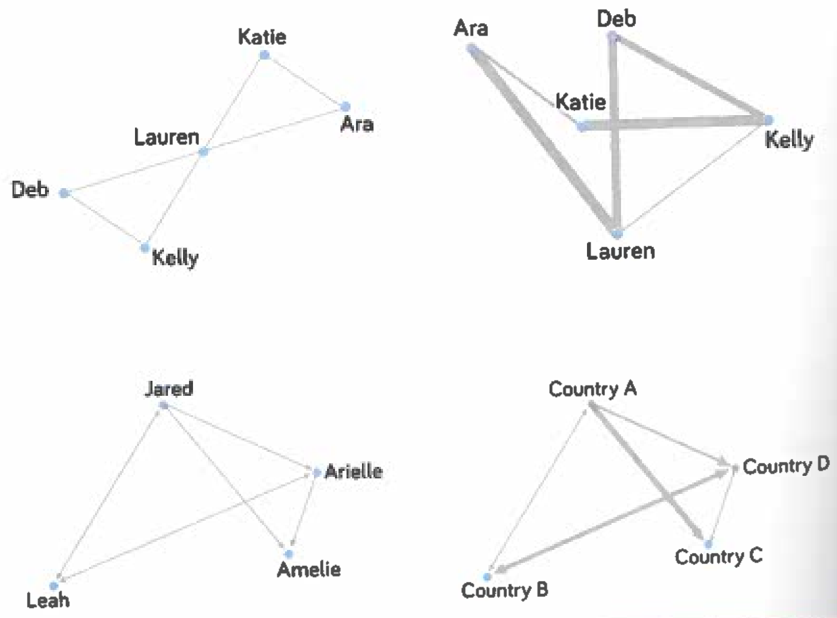
   Lauren, Ara, and Katie are friends. Lauren is also friends with Deb and Kelly.

2. *Undirected and Weighted*

   Researchers in this diagram are connected if they published a paper together. The thickness of the line is the number of times they have published together.

3. *Directed and Unweighted*

   Jared follows Leah, Amelie, and Arielle on Twitter, but only Leah follows him back. Leah and Arielle follow each other, and Arielle follows Amelie. The connection is not weighted—they are either connected (in one or more directions) or not.

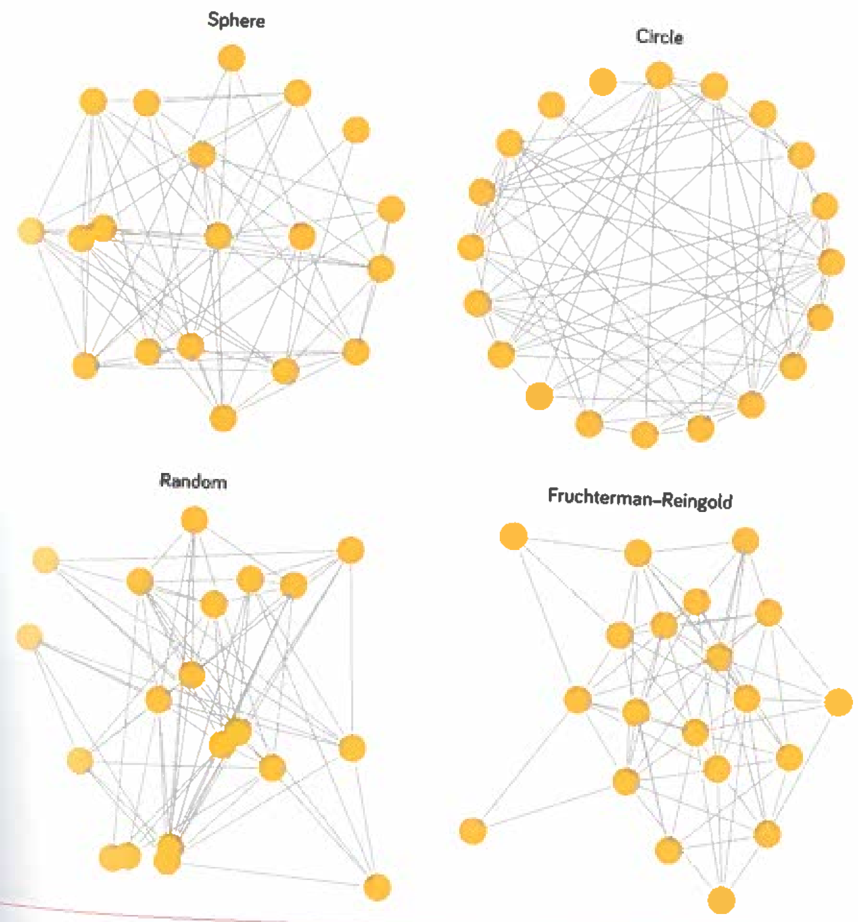


Four kinds of network diagrams, clockwise from top-left: Undirected and Unweighted; Undirected and Weighted; Directed and Unweighted; Directed and Weighted.

4. *Directed and Weighted*

   People migrate from one country to another. The thickness of the line is the number of people migrating and the direction is the destination.

There are many algorithms we can choose from to lay out the nodes and edges in a network diagram. Usually, network algorithms try to minimize how often the edges cross one another and prevent overlap of the nodes. Generally, we want the edges in a network diagram to be of roughly uniform length and the vertices to be distributed evenly. Using example data

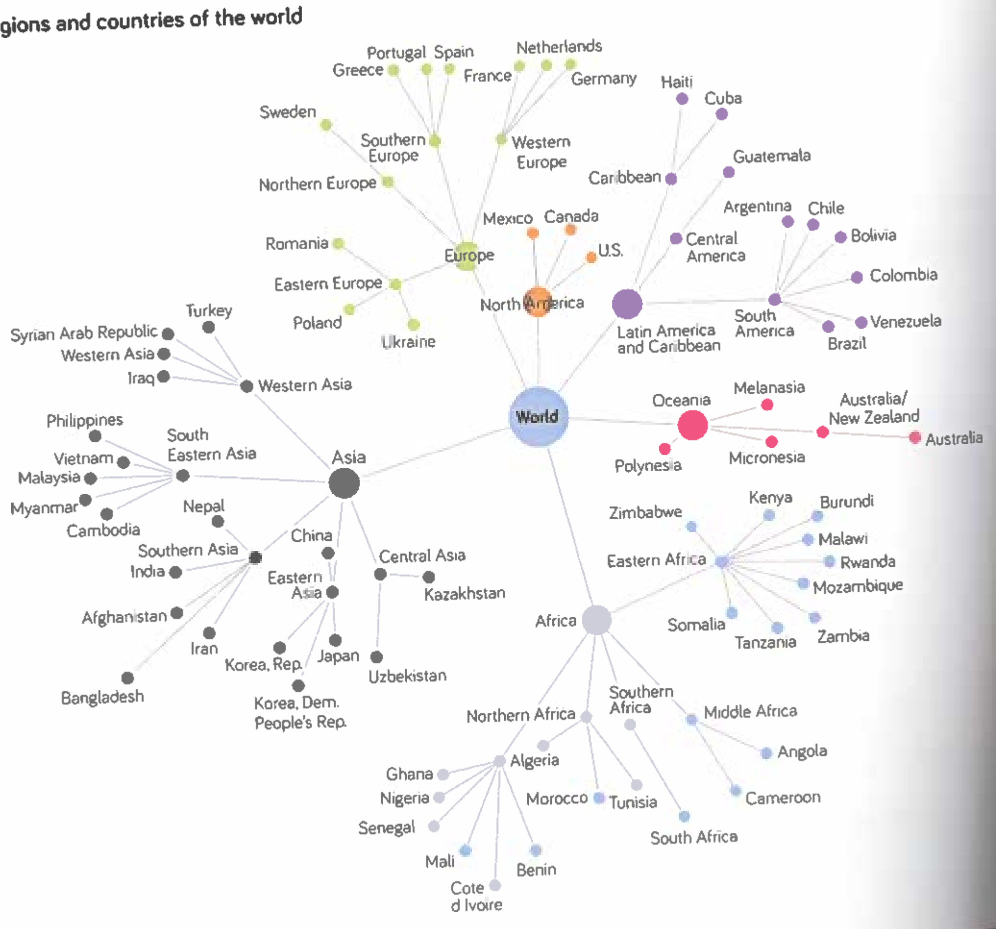


Four types of algorithms to create a network diagram.

Source: Based on the R Graph Gallery

around twenty points, these four network diagrams demonstrate how select organizing algorithms will generate different views of the network and the relationship between the points.

This network diagram shows the relationship of the seventy-five or so most populous countries in the world (those with more than one million people) within their different geographic regions. I'm not arguing that this network diagram is a better visualization than a standard geographic map, but I show it here because you can easily understand the content

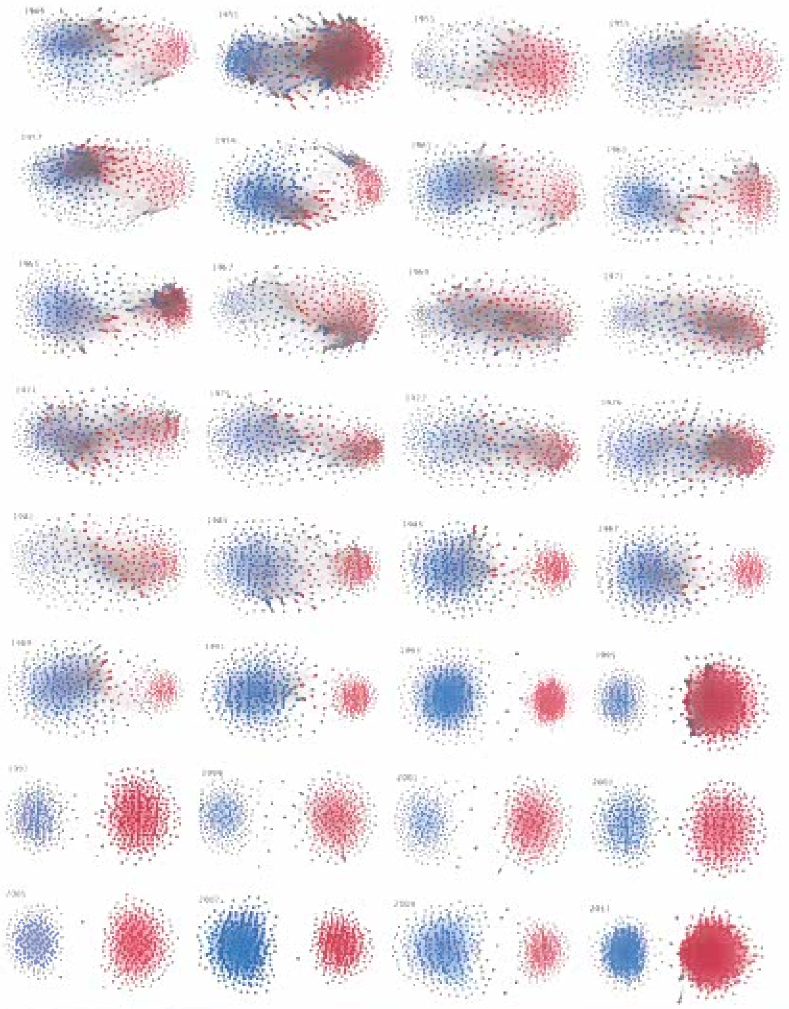

Regions and countries of the world

Source: United Nations

A simple network diagram that shows the arrangement of countries with more than one million people.

and see how the diagram works. Imagine showing the links between people in your Twitter or Facebook network, grouped by family, friends, and coworkers.

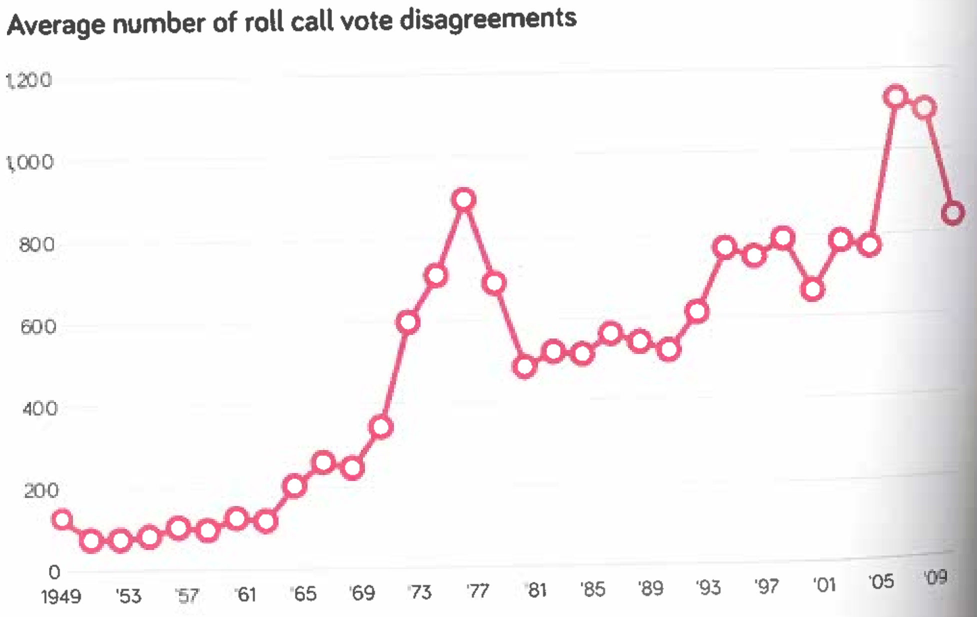
Network diagrams are ideal for showing the structure and relationship between different agents in a system. In some cases, groupings or concentrations become clear as specific nodes cluster near one another. We can use color or other shapes to highlight specific groups within the larger network.



This small multiples set of network diagrams from Andris et al. (2015) shows voting behavior in the United States Congress.

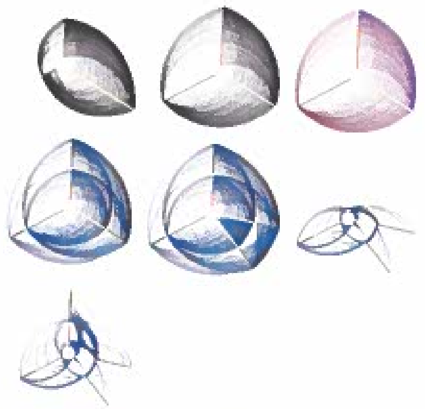
As with all other charts, including too much data in a network diagram can clutter it and make them difficult to read. But unlike many charts, sometimes the goal of a network diagram *is* to show the dense clustering. The set of thirty-two network diagrams from Clio Andris and her collaborators shows the polarization of voting behavior in the United States Congress. The authors created network diagrams for each U.S. House of Representatives from 1949 (top-left) to 2011 (bottom-right). Republican members are represented by red dots, and Democrats by blue dots. The lines denote how often members of Congress voted with one another. Even though each network diagram is very dense, the use of these small multiples makes it clear that the two parties were much less likely to vote together in 2011 than in the past.

By comparison, the line chart below shows the measure of disagreement between the parties in a simple, straightforward way. Though it's immediately informative, it's not as visually stunning as the network view.



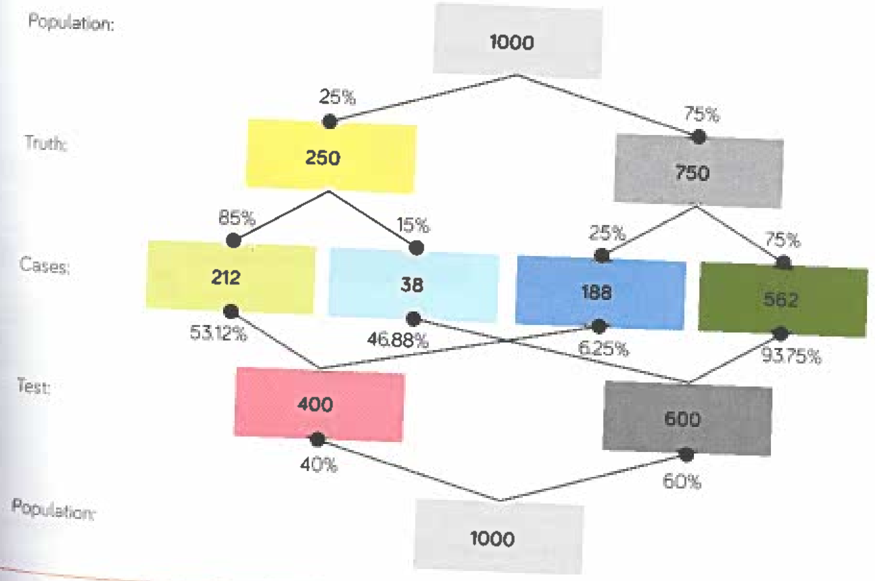
Source: Andris et al. 2015

We can make a similar case about voting behavior in the United States Congress using summary data from Andris et al. (2015), but the line chart probably doesn't grab you the way the small multiples network diagrams did.



Martin Krzywinski's hive charts are another way to show networks.
Source: Canada's Michael Smith Genome Sciences Center.

Because network diagrams can look like hairballs with too many edges and nodes to make the visualization readable, some researchers have developed alternative visualization types. The *hive plot*, for example, first organizes the space along linear axes emanating outward from a single central point. Nodes are placed along three or more axes (possibly divided into



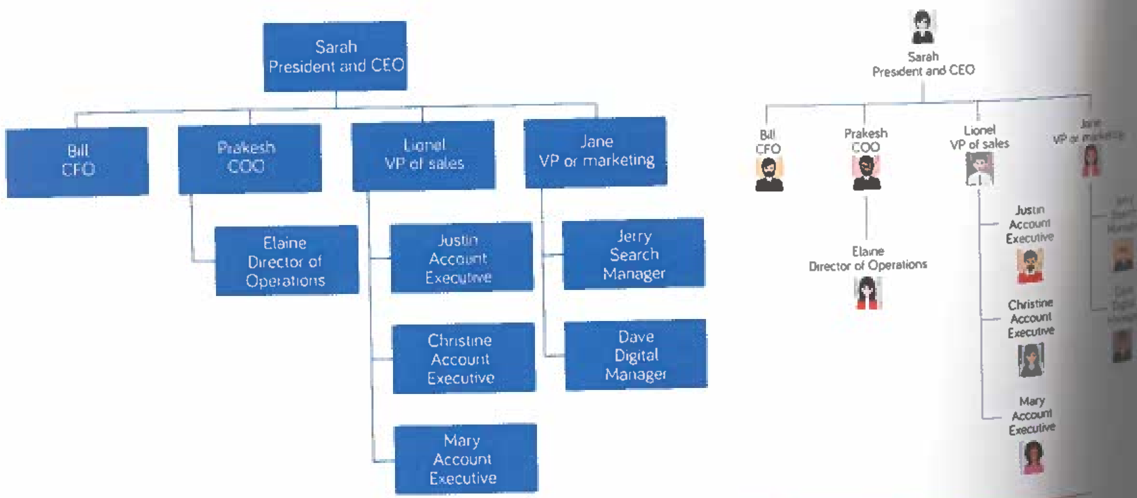
Some fields use network diagrams to visualize a process.

segments) and edges are drawn as curved links connecting the points. Martin Krzywinski, inventor of this visualization, writes that, "The hive plot is itself founded on a layout algorithm. However, its output is not based on aesthetics but network structure. In this sense, the layout is rational—it depends on networks features that you care about."

It's also worth noting that some fields refer to network diagrams in different ways. Instead of plotting how individual values correlate to one another, some network diagrams show flows or processes, similar to a flow chart or timeline. Examples might include how a computer network, a staff directory, or even a logic model of probabilities, as shown on the previous page.

TREE DIAGRAMS

Like the flow chart in Chapter 5, tree diagrams show levels of a hierarchy in a system or group. To imagine the basic tree diagram, think of a hierarchical organizational (or "org") chart. Nodes branch outward from an initial root connected by lines called *links*, *link lines*, or *branches*. The initial node is called the *root* and is the *parent* to all other nodes, some of which have child nodes of their own. Nodes who are not parent nodes are called *leaf nodes*.

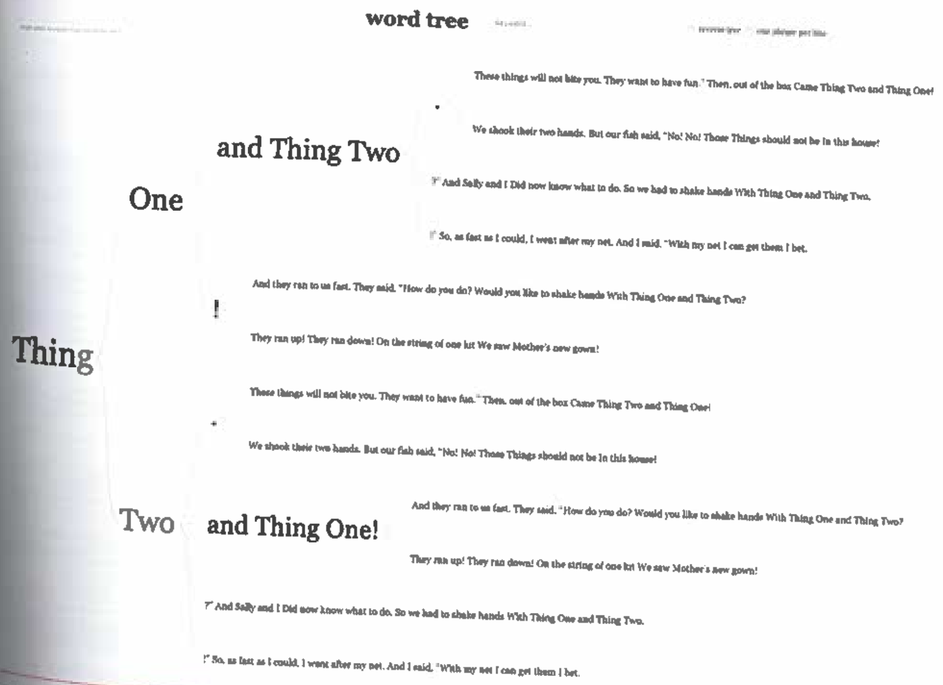


Tree diagrams describe hierarchies. These two org charts could be used for different purposes, depending on whether we think the reader would like a more designed version.

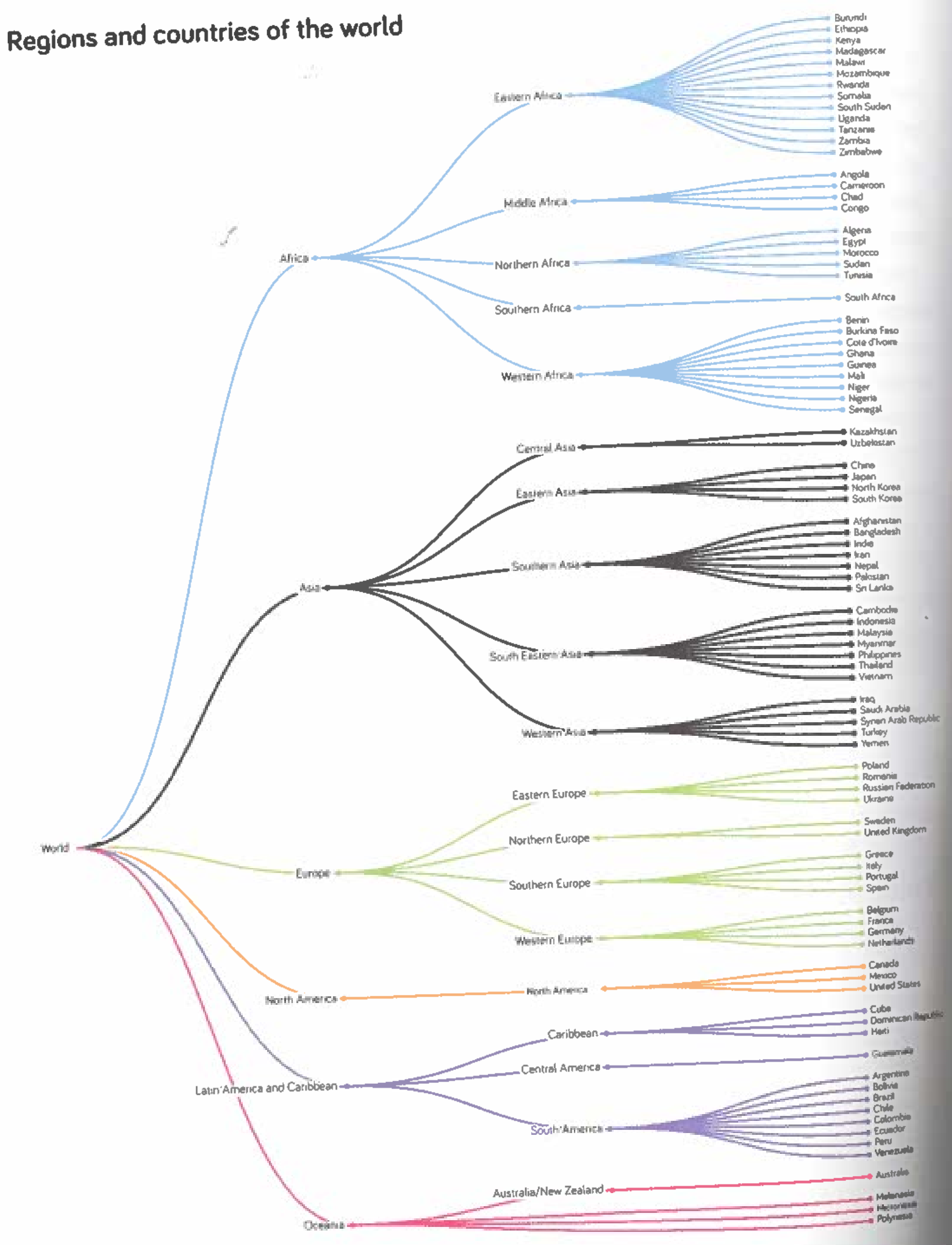
As with many of the visuals we have discussed so far, design is especially important here because there may not be much or even any data to define the elements in the diagram. Take these two imaginary org charts, for example. The chart on the left consists only of names, while the one on the right includes icons. Which one you would use depends entirely on your purpose. The one on the left might work well for the company board meeting or formal presentation; the one on the right might work better in a marketing campaign or website.

While the basic tree diagram will often branch downward, starting with the CEO or president at the top, there are lots of other ways to show these kinds of relationships. We could create a family tree that branches upward instead of downward, or a horizontal arrangement to show a different kind of hierarchy or taxonomy.

Another type of tree diagram is the word tree. Developed by Martin Wattenberg and Fernanda Viegas in 2007, the word tree is a visual representation of text in a book, article, or other passage (also see Chapter 10 on qualitative data visualization). The visualization is


word tree
Thing
One
and Thing Two
Two
and Thing One!
These things will not bite you. They want to have fun." Then, out of the box Came Thing Two and Thing One!
We shook their two hands. But our fish said, "No! No! Those Things should not be in this house!
" And Sally and I Did now know what to do. So we had to shake hands With Thing One and Thing Two.
" So, as fast as I could, I went after my net. And I said, "With my net I can get them I bet.
And they ran to us fast. They said. "How do you do? Would you like to shake hands With Thing One and Thing Two?
They ran up! They ran down! On the string of one kit We saw Mother's new gown!

This word tree is a visual representation of the text of Dr. Seuss's book, *The Cat in the Hat*.
Source: Jason Davies



Regions and countries of the world

Source: United Nations

A simple tree diagram that shows the breakdown of regions into countries.

typically ordered horizontally with a word on the left or right that branches out to show the different contexts in which it appears. These contexts are arranged in a treelike structure so the reader can uncover themes and phrases. Individual words, which act here as the nodes, are often sized by the frequency in which they appear.

There are lots of different kinds of trees to visualize quantitative or qualitative data. They can show complex data, like the human genome, or simple data, like the breakdown of countries into regions. What kind of tree you create and the design touches you include will—as always—depend on your purpose and audience. The tree on the previous page, for example, shows the same data as the network diagram shown earlier. On the one hand, it's not particularly efficient, but on the other hand, it might be easier to navigate than the network diagram.

CONCLUSION

In this chapter, we surveyed charts and diagrams that visualize relationships between variables, individuals, or groups. We often want to understand how two or more things are related, but remember that because two variables might be correlated does not mean there is a causal relationship. Clearly understanding how elements in your data are related before presenting them to your reader or audience is of utmost importance.

This class of graphs uses different strategies and shapes to communicate these relationships, and there are advantages and disadvantages to each approach as they trade off between clarity, order, and compactness. Scatterplots have a single horizontal and vertical axis; bubble plots add a third variable. Parallel coordinate plots are defined by using two or more vertical axes. Radar charts pull the axes together and radiate outward from a center point of a circle while a chord diagram wraps everything around the circumference of a circle. An arc chart then stretches everything out along a single horizontal axis and a correlation matrix uses a square or rectangular format. Network and tree diagrams can be used to show relationships between individuals or groups or passages of text.

As with the graphs in previous chapters, some of the graphs in this chapter may be unfamiliar, even difficult for you or your readers to understand. This doesn't require you to dumb things down or leave things out, but it should prompt you to consider how to best communicate the content of nonstandard graphs. Use labels, annotations, active titles, and helpful pointers to teach them how to read the graph so that they can more easily understand the content.