CHAPTER 11 VISUALIZING YOUR DATA

Data visualization has been all the rage recently, as Web-based tools have made it easier to display data and share them with others. But data visualization has been used for decades as a means to help understand data patterns and communicate those to others. The goal of this section of the book is to provide a solid footing in the fundamentals of data visualization. This section will touch on some best practices that have been refined over the years and some options for visualizing data accurately. Additionally, in this section we'll learn how to create a number of different types of visualizations using Excel's charting functions and online services offered by Google. We start this section with Chapter 11, where we will explore some principles that will help us better understand data visualization.

DATA VISUALIZATION DEFINED

Visualization is simply the act of creating charts based on data. We can easily visualize a small data set, such as a spreadsheet file. Data scientists, using other tools, can visualize terabytes of more-complex data (sometimes called "big data") from social networks, such as Twitter and Facebook, to better understand the flow of information about disasters or news events.

At their simplest, data visualizations can be barebones representations of data. Many times, analysts visualize data just for themselves, so they can get a better understanding of their data. Sometimes it's difficult to detect interesting or meaningful patterns by looking at columns and rows of data. In fact, data analysts will often create many different visualizations that can help show the data from different perspectives, or they will show a subset of the data. The techniques of **exploratory data analysis**, as promoted by statistician John Tukey, use different types of graphs to get a better understanding (Tukey, 1977). Some statistical programs, such as SPSS, JMP, MATLAB and the open source R, have exploratory data analysis functions. Excel's charting tools allow us to accomplish many of the same basic goals, albeit with more effort.

If we want to communicate to a specific audience, we might choose to create an **information graphic**—a more elegant representation. We could use a graphic design program like Adobe Illustrator to create a static information graphic, or we could use an interactive visualization tool like Tableau Desktop. The art of crafting information graphics is outside the scope of this book. To learn more, read *The Functional Art* by Alberto Cairo (2013), *Visualize This* by Nathan Yau (2011), or the books of Edward Tufte (1983, 2006).

Regardless of whether we're creating a visualization for ourselves or for others, our goal is to emphasize the content, and to help it tell a story.

"Graphics, charts, and maps are not just tools to be seen, but to be read and scrutinized. The first goal of an infographic is not to be beautiful just for the sake of eye appeal, but, above all, to be understandable first, and beautiful after that; or to be beautiful because of its exquisite functionality," wrote Cairo (2013, xx).

In the words of information design guru Tufte, "Graphics reveal data" (1983, 13). Many information designers have gravitated toward Tufte's ideas, which emphasize simplicity. In 1983, he issued his theory of graphical excellence, which includes instructions to

> Show the data
>
> Induce the viewer to think about the substance rather than about the methodology, graphic design, the technology of the graphic production, or something else
>
> Avoid distorting what the data have to say
>
> Present many numbers in a small space
>
> Make large data sets coherent
>
> Encourage the eye to compare different pieces of data
>
> Reveal the data at several levels of detail, from broad overview to the fine structure
>
> Serve a reasonably clear purpose: description, exploration, tabulation or decoration
>
> Be closely integrated with the statistical and verbal descriptions of a data set (Tufte, 1983).

Reprinted with permission by Edward R. Tufte, Graphics Press Cheshire, CT.

SOME BEST PRACTICES

Here, then, are some quick guidelines for creating visualizations, whether they're for yourself, another person or a broader audience.

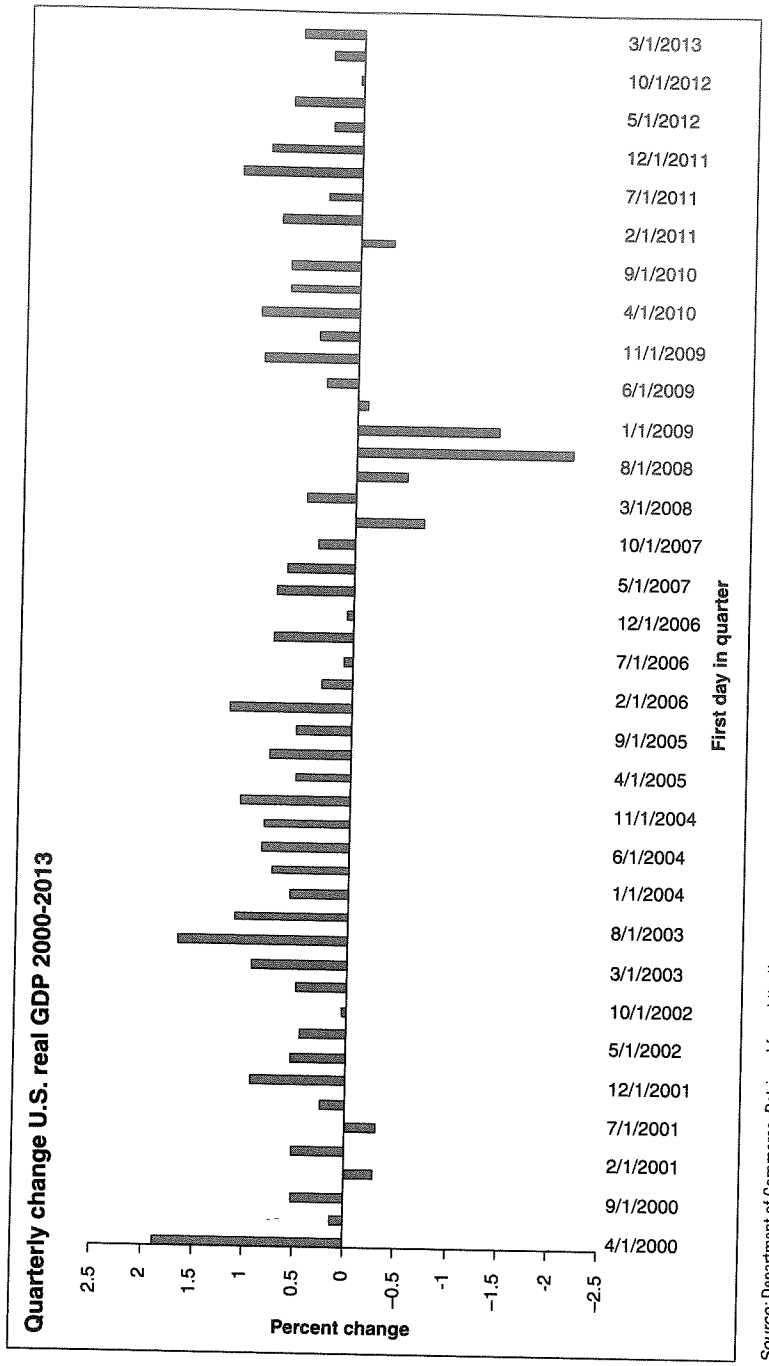
Give your chart a title. Use just a few words to describe what's displayed. Include the time frame for your data (e.g., "State unemployment rate, 2000–2012").

Label your chart elements. That includes the vertical and horizontal axes, legends and any other elements.

Include a source line. A source line reminds you about where you obtained the data. If you're sharing your visualization, it shows others where you got the data. This is important because you want to give others the ability to try to replicate or even challenge your results.

Display enough data to provide context. For instance, showing only three years of unemployment data might mask the fact that the rate had been rising in earlier years.

Let's look at this example of an Excel column chart that was created relatively quickly by the author. Using data downloaded from the St. Louis Federal Reserve Bank's FRED service, it shows how the U.S. GDP has changed from 2000 through early 2013. The GDP



Source: Department of Commerce, Retrieved from http://research.stlouisfed.org/fred2/graph/?id=GDPC1.
Note: Excel column chart.

is an important measure that's used to determine whether the U.S. economy is growing and includes "the output of goods and services produced by labor and property located in the United States" (Bureau of Economic Analysis, 2014). The U.S. Department of Commerce's Bureau of Economic Analysis releases the data quarterly. We can easily see by looking at this chart how the GDP in the United States began shrinking in 2007, as the global financial crisis began.

Note that the chart title says, "Quarterly change U.S. Real GDP 2000–2013." This communicates to viewers that the data are reported quarterly and expressed in terms of real dollars. We prefer to use data based on real dollars because doing so takes into account the effects of inflation. The title also signals to viewers they are looking at data spanning the years 2000 to 2013.

Our source line, placed just below the title, tells the viewers that the BEA produced these data and gives them the link on FRED, in the event that they'd like to retrieve the data themselves.

The vertical axis label shows viewers that the change is reported as percent, with the axis scale ranging from 2.5 percent growth to 2.5 percent loss.

The horizontal axis label tells viewers that each column represents the first day in a quarter. The labels themselves mention the specific dates. Note that the St. Louis Fed adjusts all quarterly dates to the first date of the quarter (Federal Reserve Bank of St. Louis, n.d.). Our data contain more than a decade of information, which is enough to show the big picture of GDP growth and loss during those years.

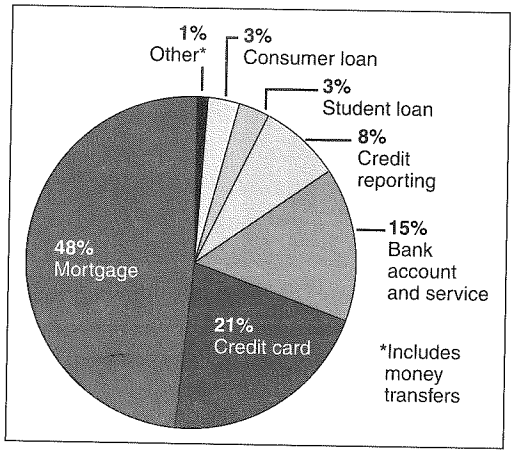
Now that we've seen some of the best practices in action, it's time to learn which chart options are the best for displaying different kinds of data.

CHAPTER 12 CHARTING CHOICES

When creating data visualizations, part of the challenge is picking the right type of chart. Excel gives us around a dozen options, everything from the simple pie chart to sparklines. So, the big question becomes, Which option should we choose? That is, which will be the most appropriate and communicate the best? The answer depends on the kind of data that we want to visualize. Here is a guide to the chart options and the data types that they're best suited to display.

VISUALIZING DATA WITH CHARTS

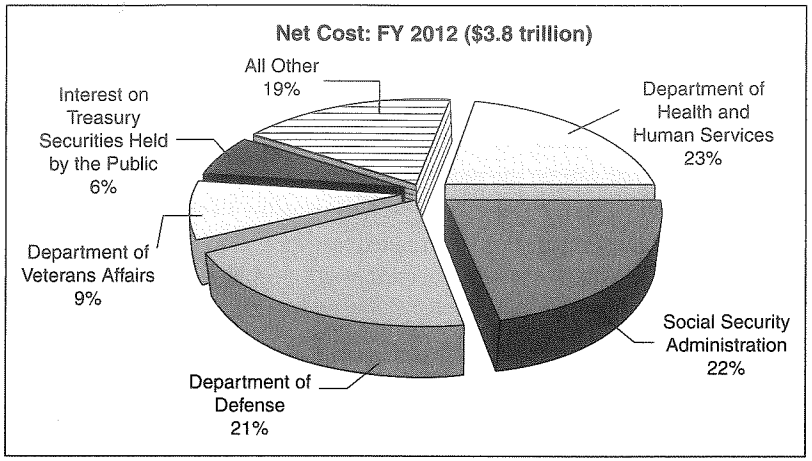
Pie charts are the best tool for showing proportions of the whole. These charts are the visual equivalent of using a spreadsheet to calculate the percent of total. As long as you use a limited number of categories, pie charts can make it easy for people to understand proportions. For instance, this pie chart from a report by the U.S. Consumer Financial Protection Bureau shows us that nearly half of complaints received by the Bureau were about mortgages, at 48 percent. Complaints about credit cards, the next largest category, came in at 21 percent of the total. Student loans made up just a sliver at 3 percent (Consumer Financial Protection Bureau, n.d.a).



Source: A snapshot of complaints received. (n.d.). Consumer Financial Protection Bureau. Retrieved July 11, 2013, from http://www.consumerfinance.gov/reports/a-snapshot-of-complaints-received-3/

Note: Pie chart showing proportion of consumer complaints reported to federal authorities.

Likewise, this pie chart from a federal spending report tells us an awful lot about how the U.S. government spent its money in fiscal year 2012. Roughly two-thirds (66 percent) of all spending was by the Social Security Administration, Department of Health and Human Services and the Department of Defense (Financial Management Service, 2013).



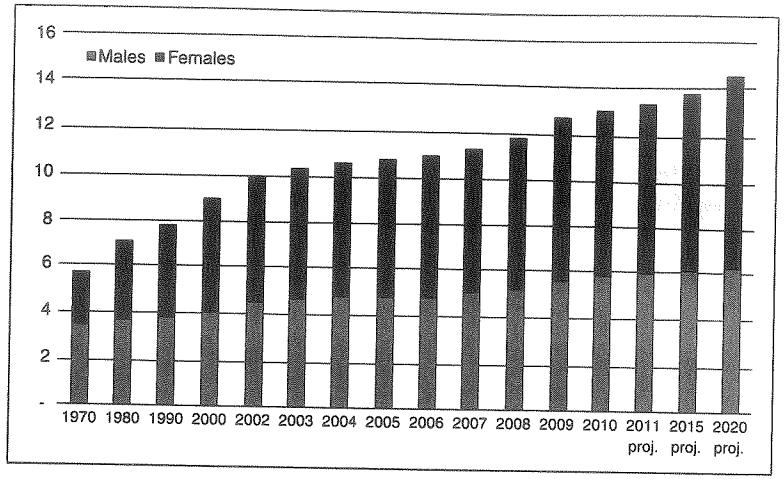
Source: Current report: Combined Statement of Receipts, Outlays and Balances. Financial Management Service. (n.d.) Web. Retrieved from http://www.fiscal.treasury.gov/fsreports/rpt/combStmt/cs2012/outlay.pdf

Note: 3-D pie charts distort data. The slice for the Department of Defense in the front looks bigger than the one for the Department of Health and Human Services in the back, but the Department of Defense's share is actually 2 percentage points less.

Pie charts have come under attack lately because they sometimes fail to make the data more understandable (Hickey, 2013). As Hickey suggests, avoid the temptation to create 3-D pie charts (such as the one above) because they can distort results and make some pie slices appear bigger than they ought to be.

Vertical **column charts** are ideal for showing change over time when you have discrete **time-series data**. Discrete time-series data are reported at defined intervals. Some examples include the quarterly GDP data we saw visualized in Chapter 11, monthly unemployment figures and annual four-year college tuition costs.

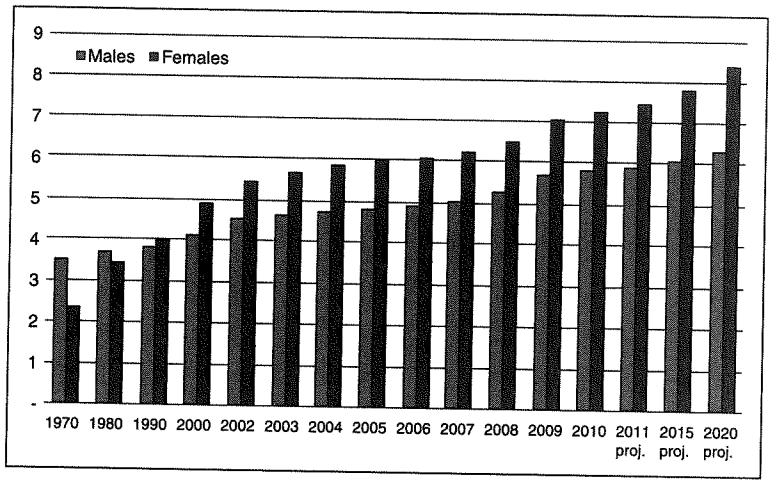
We can even create stacked vertical column charts to show proportions over time. For instance, this chart visualizes not only the increase in enrollment at full-time degree-granting institutions, but also the growing share of women attending college.



Source: National Center for Education Statistics, Department of Education, Retrieved from http://nces.ed.gov/programs/digest/d11/tables/dt1.

Note: Stacked vertical column chart showing proportions over time.

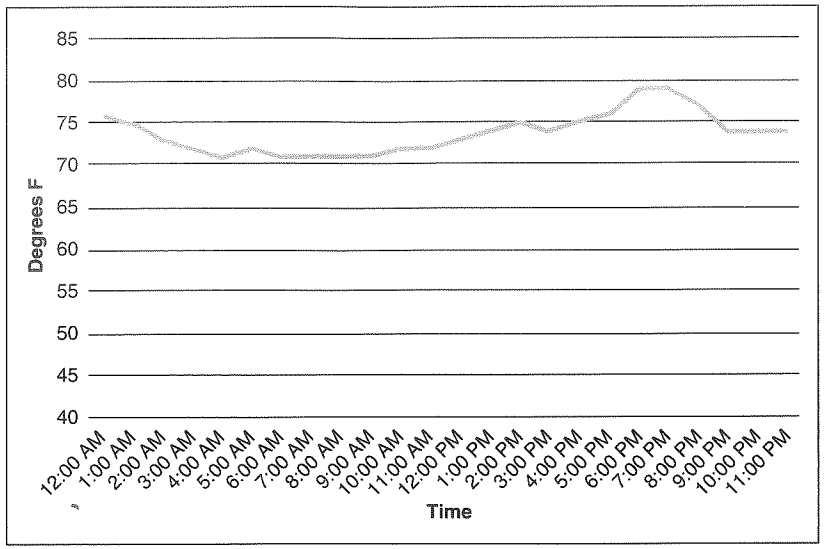
Clustered column charts allow us to compare categories over time by placing columns side by side. This clustered column chart provides a different view of our enrollment data. This one shows even better how the gap between female and male students has been widening.



Source: National Center for Education Statistics, Department of Education, Retrieved from http://nces.ed.gov/programs/digest/d11/tables/dt1.

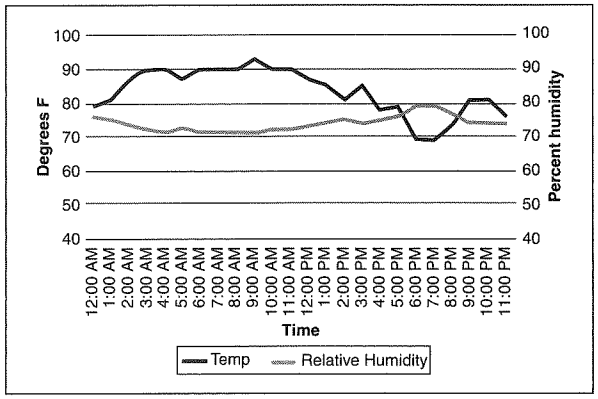
Note: Clustered vertical column chart compares categories over time.

Line charts are a great choice when we have continuous time-series data. Continuous data are those that represent processes or conditions that occur continuously, such as the outdoor temperatures. The use of the line is more appropriate because it suggests an ongoing phenomenon. In this chart, a line represents the hourly temperatures recorded by the National Weather Service in Pittsburgh, Pennsylvania, on September 1, 2013, found at http://www.erh.noaa.gov/pbz/hourlyclimate.htm.



Source: National Weather Service, Retrieved from http://www.erh.noaa.gov/pbz/hourlyclimate.htm.

Note: Line charts show continuous time-series data, such as temperature readings.



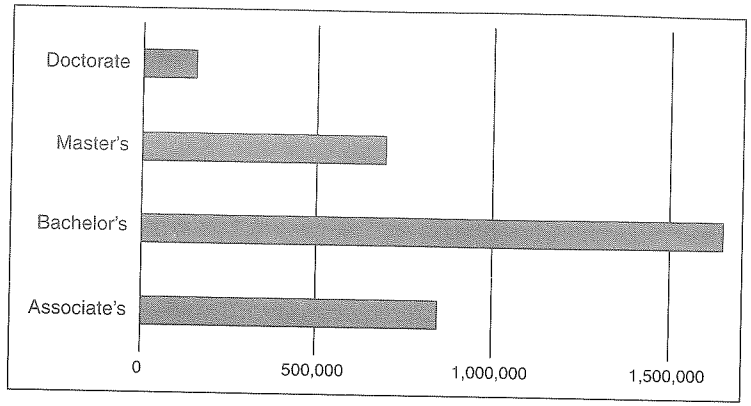
Source: National Weather Service. Retrieved from http://www.erh.noaa.gov/pbz/hourlyclimate.htm.

Note: More than one data element displayed on a line chart.

If we have more than one data element that we'd like to chart to add context or make comparisons, we can add more lines. With this chart, data about relative humidity by the hour in Pittsburgh have been added. Note that these data are plotted on the same scale as the temperature.

We can choose horizontal **bar charts** when

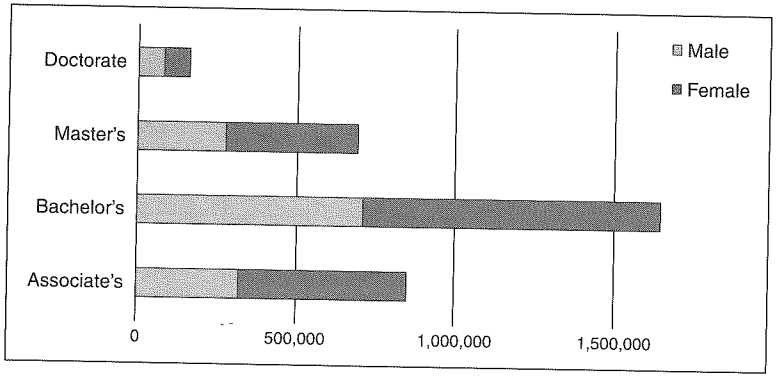
we want data that are categorized in just a few ways. For instance, this bar chart shows us the number of higher-education degrees granted in the United States 2009–2010. This chart works well because we have a small number of categories. It could get difficult to understand if we had too many.



Source: National Center for Education Statistics, Department of Education. Retrieved from http://nces.ed.gov/programs/digest/d11/tables/dt11_283.asp?referrer=report.

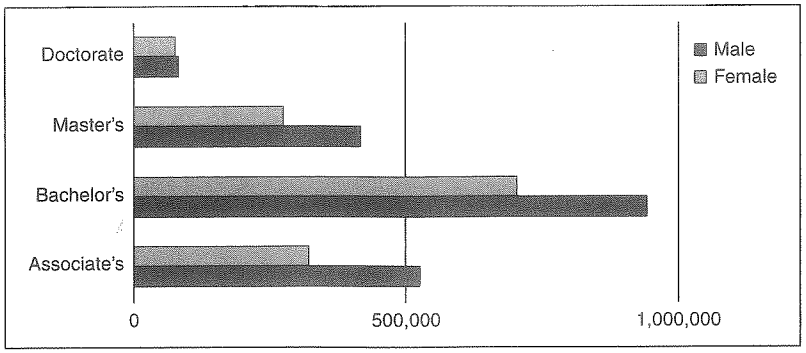
Note: Data categories shown on a horizontal bar chart.

As we did with the vertical columns chart earlier, we can show proportions by using stacked bars. The first chart below shows not only the total number of degrees, but also the proportion of women and men who earned them. We could also use a clustered bar chart to compare the degrees earned by men and women in yet another way.



Source: National Center for Education Statistics, Department of Education. Retrieved from http://nces.ed.gov/programs/digest/d11/tables/dt11_283.asp?referrer=report.

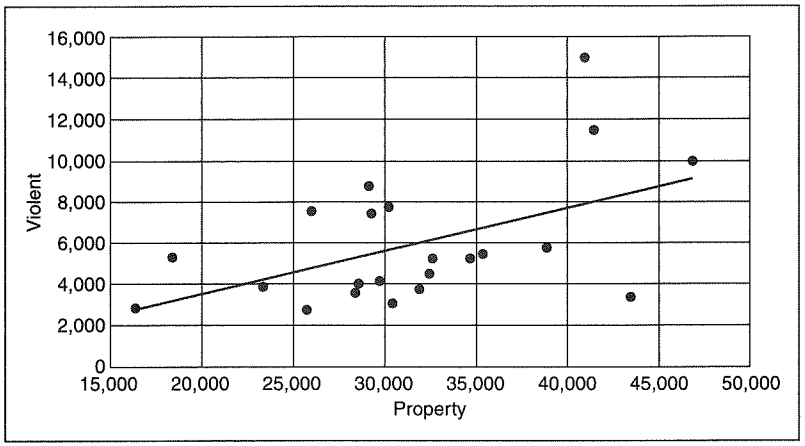
Note: Showing proportion with a stacked bar chart.



Source: National Center for Education Statistics, Department of Education. Retrieved from http://nces.ed.gov/programs/digest/d11/tables/dt11_283.asp?referrer=report.

Note: Comparing categories with a clustered bar chart.

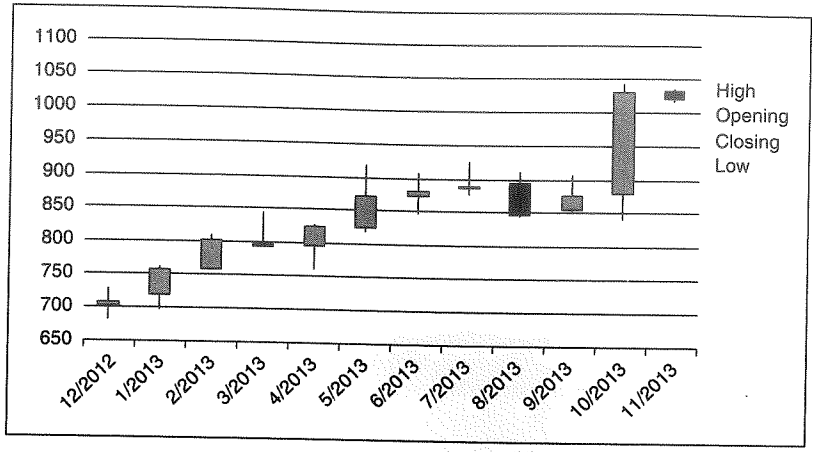
Scatterplots help us show whether we might have a relationship between two data variables. For instance, we could plot SAT or ACT scores against college freshman grade point averages to see whether there is any relationship between high scores on those standardized tests and student performance. The scatterplot below uses the crime data from Chapter 9 and shows the relationship between property crimes (the horizontal axis) and violent crimes (the vertical axis). In general, we see that as the number of property crimes increases, so does the number of violent crimes. The chart includes a linear **trend line** that shows the central tendency of the data. Any point that's well above the trend line represents a city whose violent crime numbers are higher than expected. Excel uses a statistical calculation called **linear regression** to determine the position of the linear trend line.



Source: Federal Bureau of Investigation. Retrieved from http://www.fbi.gov/about-us/cjis/ucr/ucr-publications#Crime.

Note: Scatterplots for comparing variables.

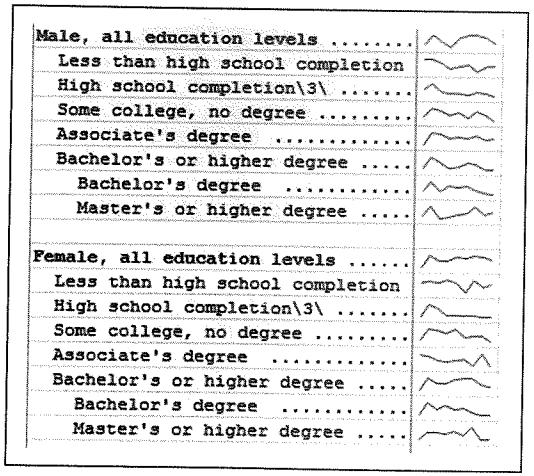
Stock charts—or boxplots—can be used to visualize the performance of stocks over time. Many times, we see stock price data visualized as a line chart, but the stock chart can provide more detail than a simple line chart. Here we see monthly Google stock opening and closing prices, as represented by the inner bars. The outer lines display the high and low prices.



Source: Yahoo Finance. Retrieved from http://finance.yahoo.com/q/hp?s=GOOG&a=07&b=19&c=2004&d=10&e=6&f=2013&g=m&z=66&y=66.

Note: Stock chart, otherwise known as a boxplot.

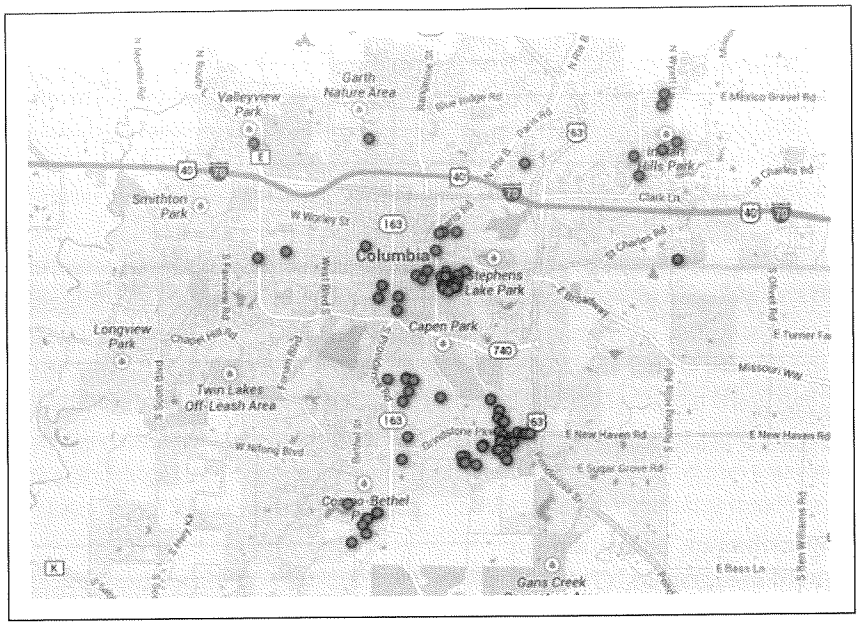
Sparklines are "intense, simple, word-sized graphics" (Tufte, 2006, 47) that are a relatively new charting option in Excel spreadsheets. Sparklines are placed inside individual cells and take the place of numbers or text. In Excel, we can create line, bar and win-loss sparklines. The sparklines in the spreadsheet here show how median income has changed from 1995 to 2011 for men and women, and by level of education. A sparkline is an excellent tool for helping compare patterns in large data sets.



Source: National Center for Education Statistics, Department of Education.

Note: Sparklines are repetitive charts that fit inside cells.

Finally, a map is often the best way to visualize **geographic data**. Viewers can orient themselves and understand what's happening near them. Out of the box, Excel is unable to create maps but can do so using plug-in programs. We can use **geographic information system** (GIS) programs, such as ArcGIS or Quantum GIS, to create data maps. Or we can use online programs, such as Google Fusion Tables. This Fusion Table map shows the location of residences where the Columbia (Missouri) Police Department responded to reports of nuisance parties. Most of the points are concentrated in areas with large amounts of student housing.



Source: Google maps; City of Columbia, Missouri.

Note: A Google Fusion Table map.

Now that we know which charts are the best for our data, we're going to learn in the next chapter how to build these charts in Excel.

ON YOUR OWN

Find three charts or information graphics in print or on the Internet. Write a critique of each: How easy is it for you to understand? Was the chart the best choice for the data? Why or why not? Did the chart have the necessary elements? Make sure you provide a copy of each chart or a URL for it.

CHAPTER 13 CHARTING IN EXCEL

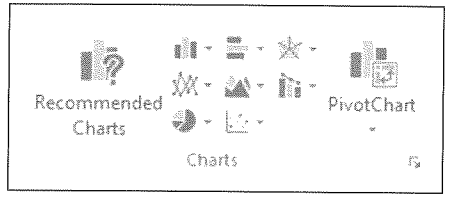
Now we will use Excel to create the charts that we saw in Chapter 12. Download the charting.xlsx file from the website for this book and open it. It contains seven worksheets, which have labels on their tabs. The first one is Degrees conferred. This sheet, like all of the others, has data that have been edited and formatted so it's easier to create the charts. Often, the data we want to use for our charts are stored in columns or rows that aren't neighbors. That makes it challenging to highlight just what we need. The data sources are noted at the bottom of each sheet and include URLs for download.

PIE CHART

We'll start by creating a pie chart to show the proportion of degrees conferred in 2009–2010, using data from the National Center for Education Statistics. The first two rows hold the data we need to build the chart. It has the degree category in the first row and the numbers for each category in the second.

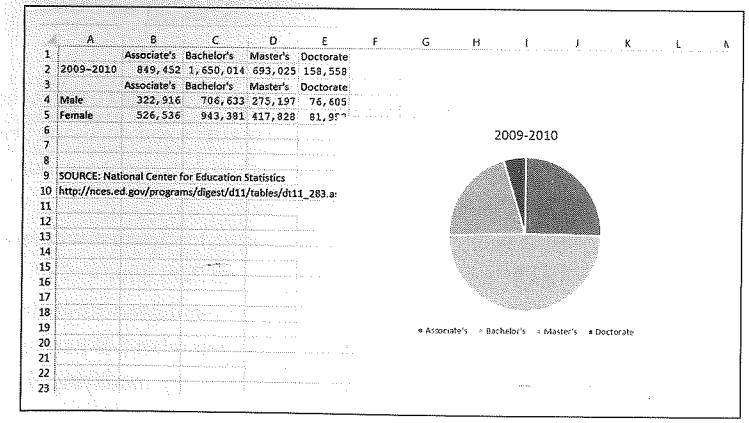
Highlight cells A1 through E2. Then select the Insert tab and click on the drop-down arrow for the pie chart button.

Select the 2-D Pie option and Excel creates the chart just like that.



Source: Microsoft Excel for Windows 2013.

Note: Excel chart type selector.



Source: National Center for Education Statistics, Department of Education.

Note: Creating a pie chart.