CHAPTER 6

CHARTS AND GRAPHS

THE CAPACITIES AND FUNCTIONS OF CHARTS AND GRAPHS ●

While a "picture may be worth a thousand words," graphs and charts rarely speak entirely for themselves. When joined by statistics and text, however, graphs and charts can help produce principled and persuasive argument, strengthen storytelling, and aid in pattern recognition. Graphs, of course, can also mislead, and we will display several examples in this chapter where graphs either wittingly or unwittingly do just that.

As we look at examples of good and bad graphical display in this chapter, we will also provide guidelines, based on good practice and research on perception and cognition, on which we should base the production of graphs and by which we can evaluate the quality of others' graphs. Aspects of graphs (and tables as well) can be chosen in ways that highlight what's important in the story we seek to tell. Conversely, violations of these principles obscure important points, highlight sideshows, confuse readers, and even misinform.

As you will see in the exercises you complete, statistical software programs such as SPSS and Excel easily produce graphs and charts. Unfortunately, the "default" graphs they produce invariably violate the principles of good graphical display. Fortunately, these software packages also provide editing tools for you to bring their graphs, charts, and tables into closer compliance with the guidelines for effective graphical display that this chapter will provide. These software packages also entice the user with numerous graphic bells and whistles, many of which get in the way of effective communication. You will discover in the pages that follow that there are plenty of defaults to derail and options to avoid in producing good graphs and charts.

Graphs Complement Statistics, Detecting Patterns Where Statistics Alone Fail

Consider Anscombe's Quartet, named after its inventor, F. J. Anscombe (1973). He begins with four sets of data (displayed in Tufte, 2001, pp. 13–14), as shown in Table 6.1.

Clearly, it's difficult to deduce much from this assortment of numbers by simply eyeballing them. What's interesting about them is that they produce the same results if each is entered into a regression equation. We'll turn to regression analysis in Chapter 11, so don't worry about the meaning of particular numbers below except to note that these statistics are the same for each set of paired numbers in Table 6.1.

Number of observations (n) = 11
Mean of the xs ($\bar{x}$) = 9.0
Mean of the ys ($\bar{y}$) = 7.5
Regression coefficient (b_1) of y on x = .5
Equation of regression line: y = 3 + .5x
Sum of squares of $x - \bar{x}$ = 110.0
Regression sum of squares = 27.50 (1 *df*)
Residential sum of squares of y = 13.75 (9 *df*)
Estimated standard error of b_1 = 0.118
Multiple R^2 = .667

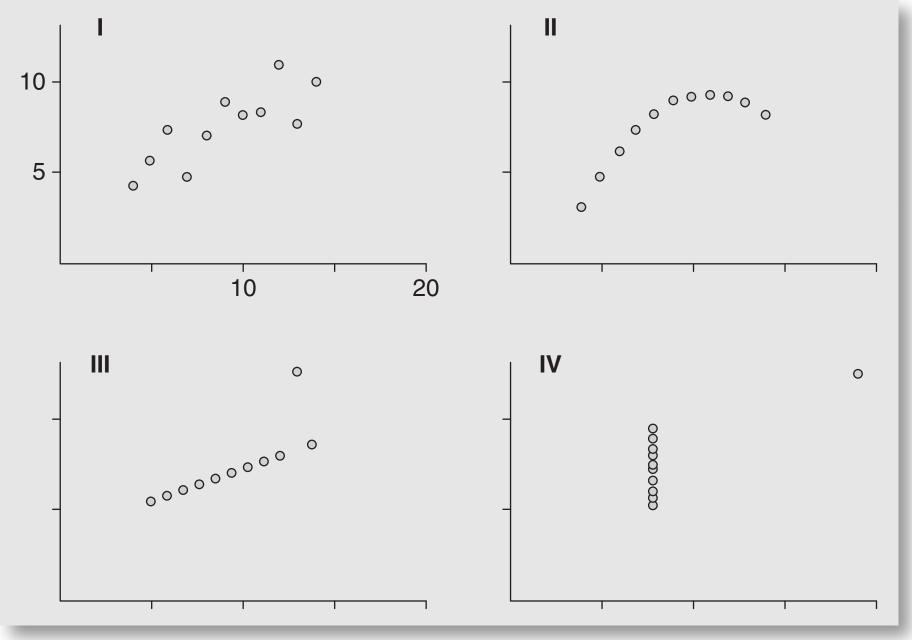
Table 6.1 Anscombe's Quartet Design

I		*II*		*III*		*IV*	
X	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

SOURCE: Tufte (2001, pp. 13–14). Used by permission from Graphics Press.

Note too, however, what happens when we produce a graph of each set of paired numbers. We see that these four sets of numbers, although characterized exactly the same by a number of descriptive and regression statistics, reveal quite different patterns and thus tell quite different stories in Figure 6.1 (Tufte, 2001, p. 14).

Figure 6.1 The Graphical Display of Anscombe's Quartet



SOURCE: Tufte (2001, p. 14). Used by permission from Graphics Press.

The point of these four graphs is to show that statistics alone do not always adequately describe the patterns that we look for in the data. The story is in the picture *and* the statistics. Don't tell only part of a story by relying on statistics (or graphs) alone.

Graphs Can Help Diagnose Problems in the Data

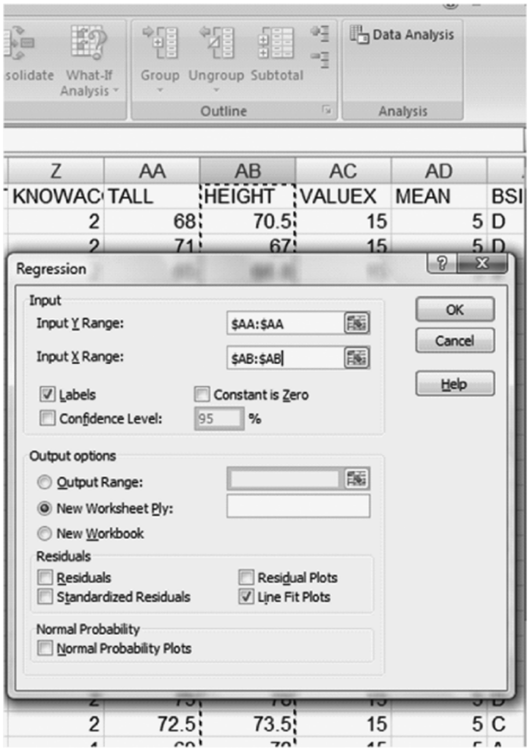
I ask students in my statistics class to complete the survey found in Appendix C at the beginning of each semester. The purpose of this exercise is to illustrate some of the problems and pitfalls that face those who write questions and design questionnaires. But one pair of questions is intended to demonstrate that all measurement errors aren't as bad as one might believe from the discussion of the complexities of asking and answering questions. In this latter regard, I ask two versions of the same question about the height of each student. One is self-reported; the other is an "objective" measure of height taken by a fellow student. This latter measure is arrived at by using tape measures that I tape to the wall at four different locations throughout the classroom. I position the tape measures to begin at 36 inches above the floor, which will factor into the story's conclusion below.

To demonstrate to myself and to the students in my class one year that these two measures would not be exactly alike but pretty close to each other, I ran a regression equation in Excel, asking for the correlation between the two measures. (A correlation is a measure of the extent to which two variables are linearly related to each other and can take on values of +1.0 if perfectly, and positively, related; .0 if not related at all; and –1 if perfectly, and negatively, related.)

Here's how I produced these results in Excel[1]:

Step 1: Select Tools/Data Analysis/Regression, and click OK.

Step 2: Enter the values as shown in the screen shot below.



Z	AA	AB	AC	AD	
KNOWAC	TALL	HEIGHT	VALUEX	MEAN	BSI
2	68	70.5	15	5	D
2	71	67	15	5	D
2	72.5	73.5	15	5	C

Regression

Input

Input Y Range: $AA:$AA

Input X Range: $AB:$AB

☑ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

☐ Residuals ☐ Residual Plots

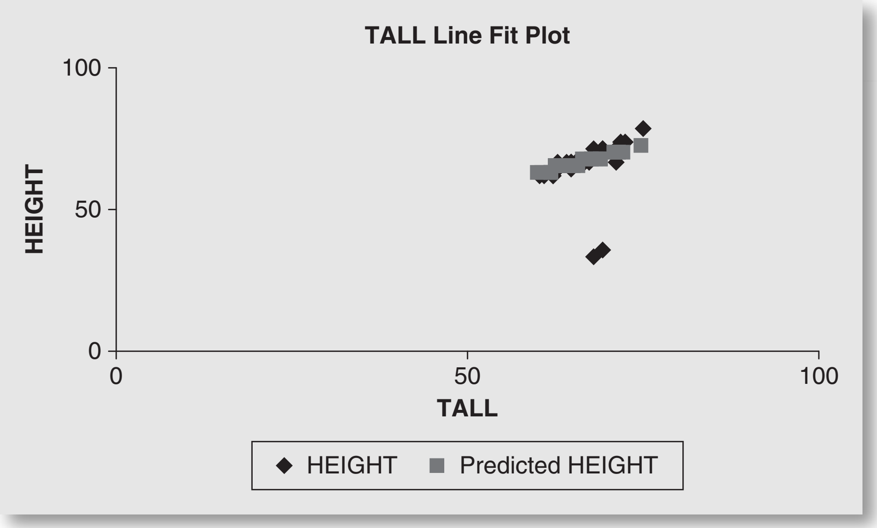
☐ Standardized Residuals ☑ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

I was shocked to discover that the correlation between height as measured in these two ways was only .30, far less than I had expected. I then turned to the scattergram that accompanies the regression analysis (because I specified "Line Fit Plots" in the dialog box above). The resulting default graph (Figure 6.2) is quite uninformative and directs attention to the regression line or predicted values of measured height, given the measures of self-reported height.

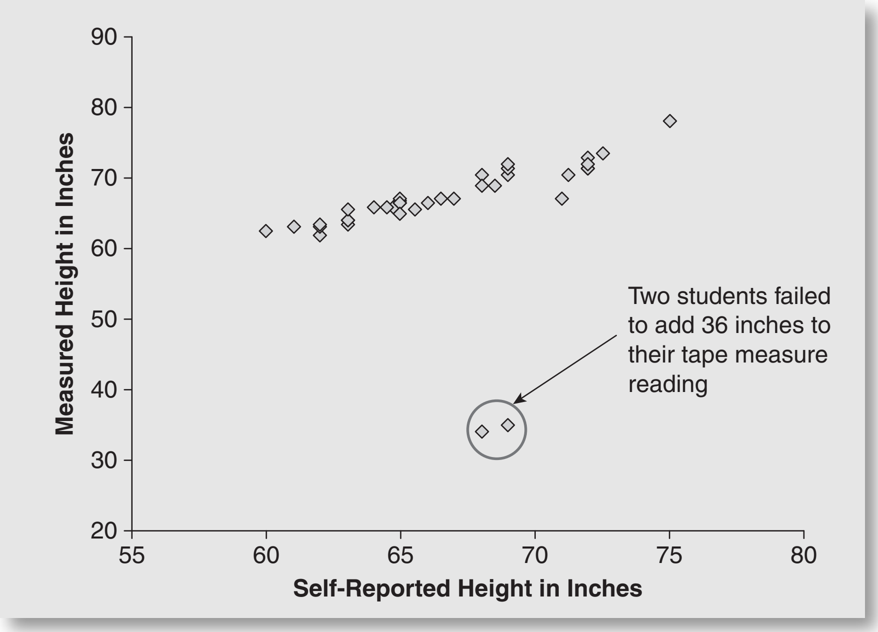
Figure 6.2 Excel's Default (and uninformative) Scattergram of Student Heights



This graph violates nearly every principle of graphical design that we'll examine later in this chapter. Here's what the graph should look like (Figure 6.3) after overriding the defaults and applying several principles of graphical design.

It is fairly clear from this picture that two students reported their measured height without adding the 36 inches that the tape measure was hung above the floor. The graph enabled me to spot the problem quickly. (By the way, these 2 observations represent what we'll later call **bivariate** (i.e., two-variable) **outliers**. These stray observations can be easily corrected in the spreadsheet and the above analysis rerun, which would show a correlation of .94 instead of .30. I was right after all, but I could only demonstrate this with clean data.

Figure 6.3 Graphically Identifying the Source of Unexpected Statistics

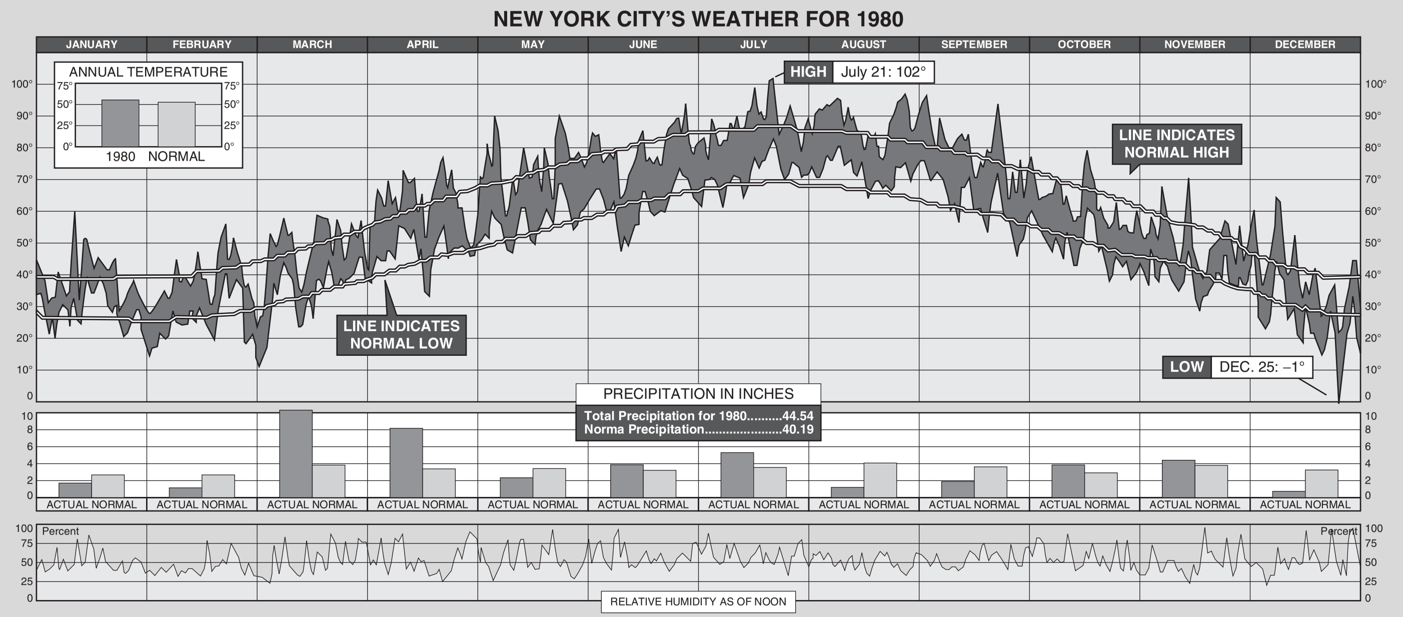


Graphs Can Display Tons of Data

The power of graphs lies in part in their ability to display a lot of data. Tufte (2001, p. 30) demonstrates this point with a graph of New York City's weather for 1980 (see Figure 6.4). The graph displays the temperature, precipitation, and relative humidity for every day of that year. It displays 1,888 different numbers. Can you imagine what the tables with these numbers would look like? Surely, we could more precisely identify specific values in such tables but not their patterns as easily as the graph enables. Indeed, a table is more appropriate than a graph if your purpose or need is to find and compare precise values. Graphs are better at identifying patterns and irregular observations.

It is also the case that graphs can often convey information more efficiently than text can. Here again is another example to illustrate that point from Tufte (2001, p. 37). Figure 6.5 displays counts of the outgoing mail from the U.S. Congress, arrayed in a time series for the years 1967 through 1972. The peaks occur in October of even-number years. Why might that be? You guessed it. The peaks are the months prior to congressional elections.

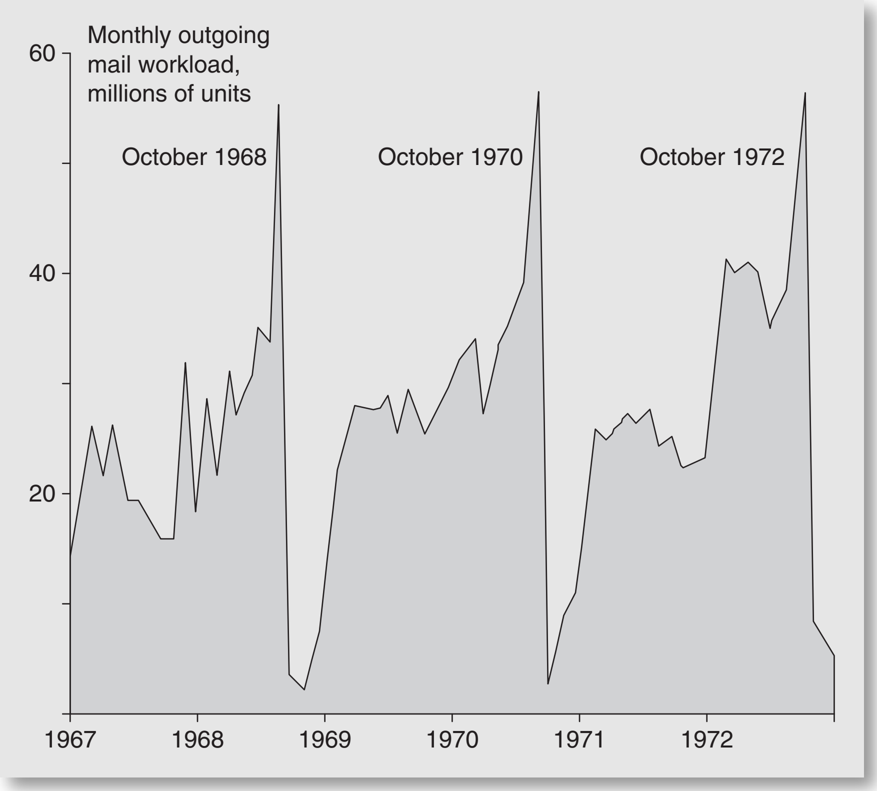
Figure 6.4 Graphical Display of 1,888 Different Numbers to Reveal Patterns



New York Times, January 11, 1981, p. 32.

SOURCE: Tufte (2001, p. 30). Used by permission from Graphics Press.

Figure 6.5 Graphical Display of Congressional Outgoing Mail



SOURCE: Tufte (2001, p. 37). Used by permission from Graphics Press.

● SPOTTING DISTORTIONS AND LIES IN GRAPHS

Graphs, as well as statistics and words, can be used to deceive and mislead. Distortion occurs when a graph's visual representation is inconsistent with its numeric representation. Tufte (2001, p. 57) provides a formal calculation of such distortions, which he calls the **Lie Factor Quotient**. He calculates this number by using the following formula:

$$\text{Lie Factor} = \frac{\text{Size of effect in the graph}}{\text{Size of effect in the data}}.$$

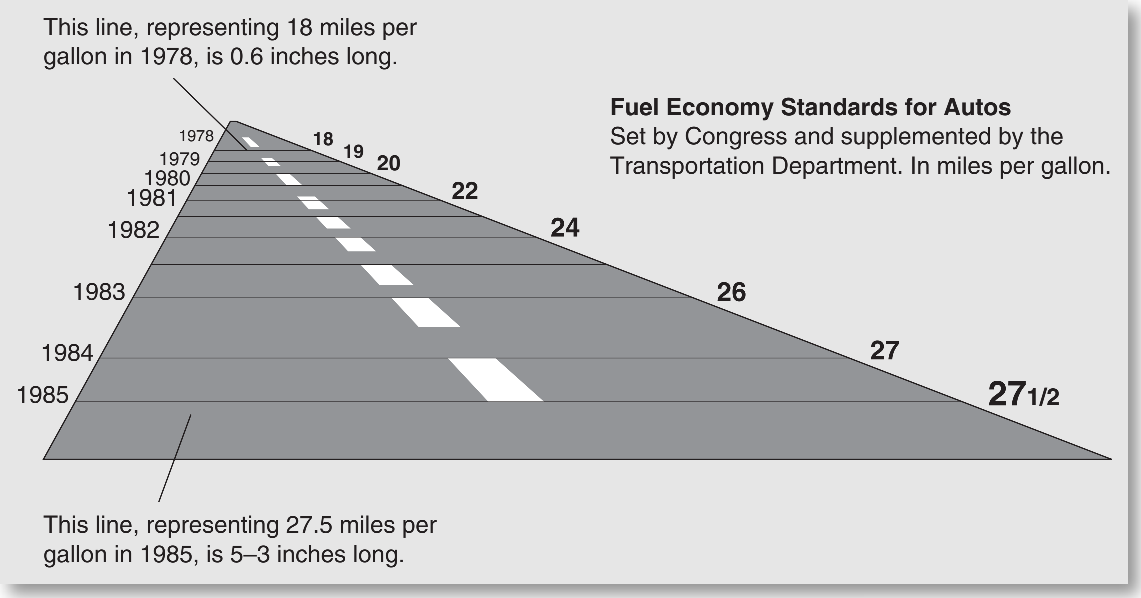
Distortion quotients less than 0.95 and greater than 1.05 constitute substantial distortion in Tufte's view. Let's examine an illustration and calculate the Lie Factor for it.

Consider the graphical depiction (Figure 6.6) of fuel economy standards for new automobiles set by the U.S. Congress in 1978 (reproduced in Tufte, 2001, p. 57, from *New York Times*, August 9, 1978, p. D-2.)

Fuel standards for new cars were to increase from 18 miles/gallon in 1978 to 27.5 by 1985, an increase of 53% ((27.5 − 18.0)/18.0). The graph as originally displayed, however, represented 18 miles/gallon by a line 0.6 inches long. It represented the 27.5 mile/gallon target in 1985 with a line 5.3 inches long. This graphical increase was 783% ((5.3 − 0.6)/0.6). The Lie Factor was therefore

$$\frac{783}{53} = 14.8.$$

Figure 6.6 Graphical Distortions of Fuel Economy Standards for Automobiles



This line, representing 18 miles per gallon in 1978, is 0.6 inches long.

Fuel Economy Standards for Autos
Set by Congress and supplemented by the Transportation Department. In miles per gallon.

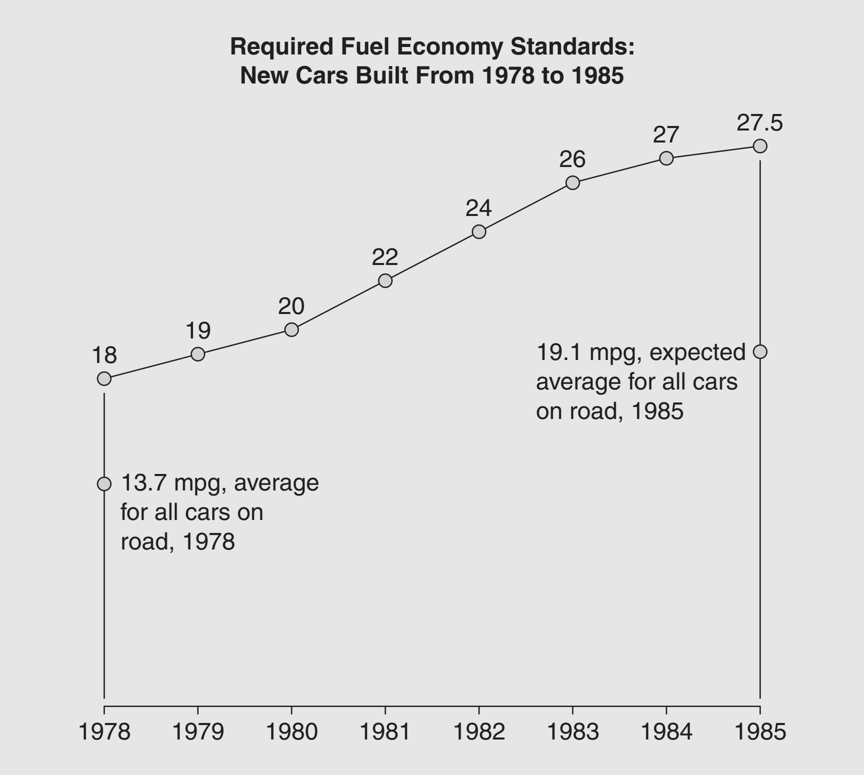
This line, representing 27.5 miles per gallon in 1985, is 5–3 inches long.

SOURCE: Tufte (2001, p. 57). Used by permission from Graphics Press.

It's a creative graph, but one that leaves an impression substantially at variance with the facts. Tufte provides a simpler and more honest presentation of the same data (with the added and useful figures for the expected average fuel economy for all automobiles in 1978 and 1985). It's not as artistic but certainly more truthful (see Figure 6.7).

Another example of graphical distortion is provided by the National Science Foundation (NSF), the agency charged with supporting basic research in the United States. NSF issues biennial chartbooks of statistics about science and technology. But it also must go before Congress each year and request and typically justify larger requests for funds. The temptation is to massage the data in ways that make a more compelling case for increased funds. Would a precipitous drop in the number of Nobel Prizes for science awarded to U.S. scientists help? Someone must have thought so, because the 1974 *Science*

Figure 6.7 Honest Graphical Depiction of Fuel Economy Standards

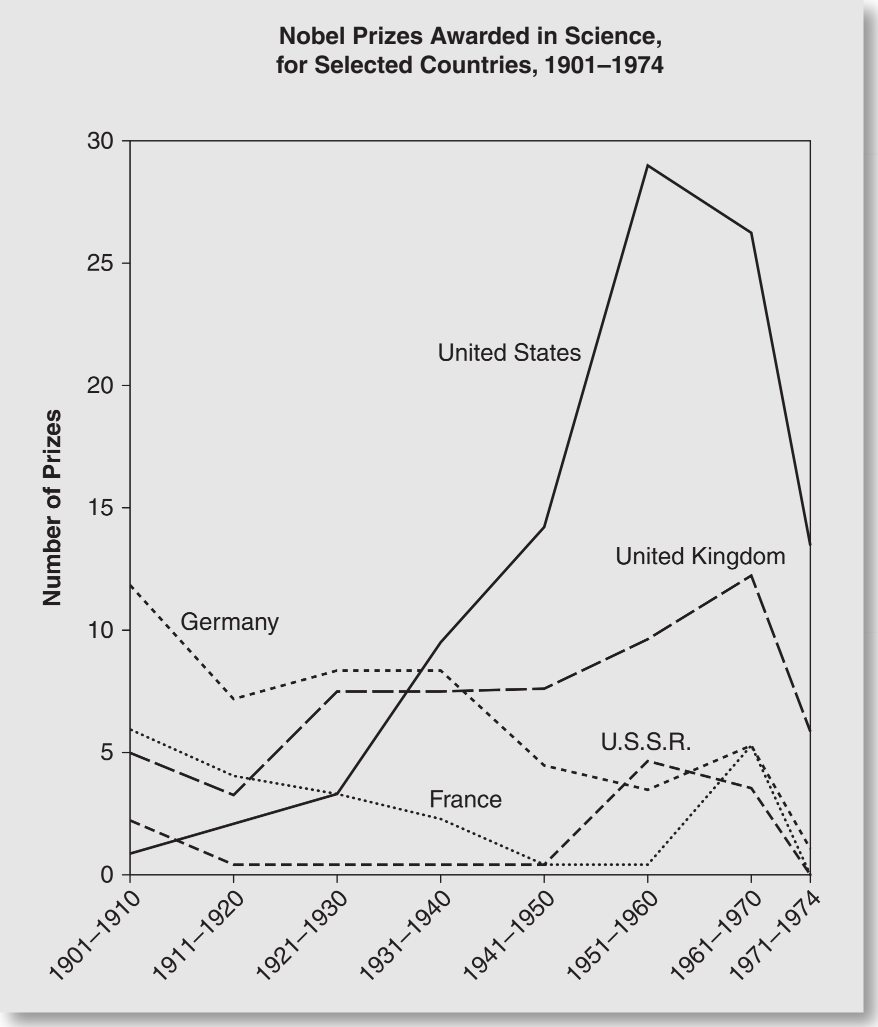


SOURCE: Tufte (2001). Used by permission from Graphics Press.

Indicators chartbook included a graph of Nobel prizes in science awarded to the citizens of selected countries (reproduced in Tufte, 2001; see Figure 6.8).

Wow, U.S. science looked like it was in trouble, didn't it? Can you spot the sleight of hand in this chart? The graph changes the total number of years in the final data point. While all others represent a 10-year period, the last on

Figure 6.8 Misleading Graphical Display of Nobel Prizes Awarded for Selected Countries

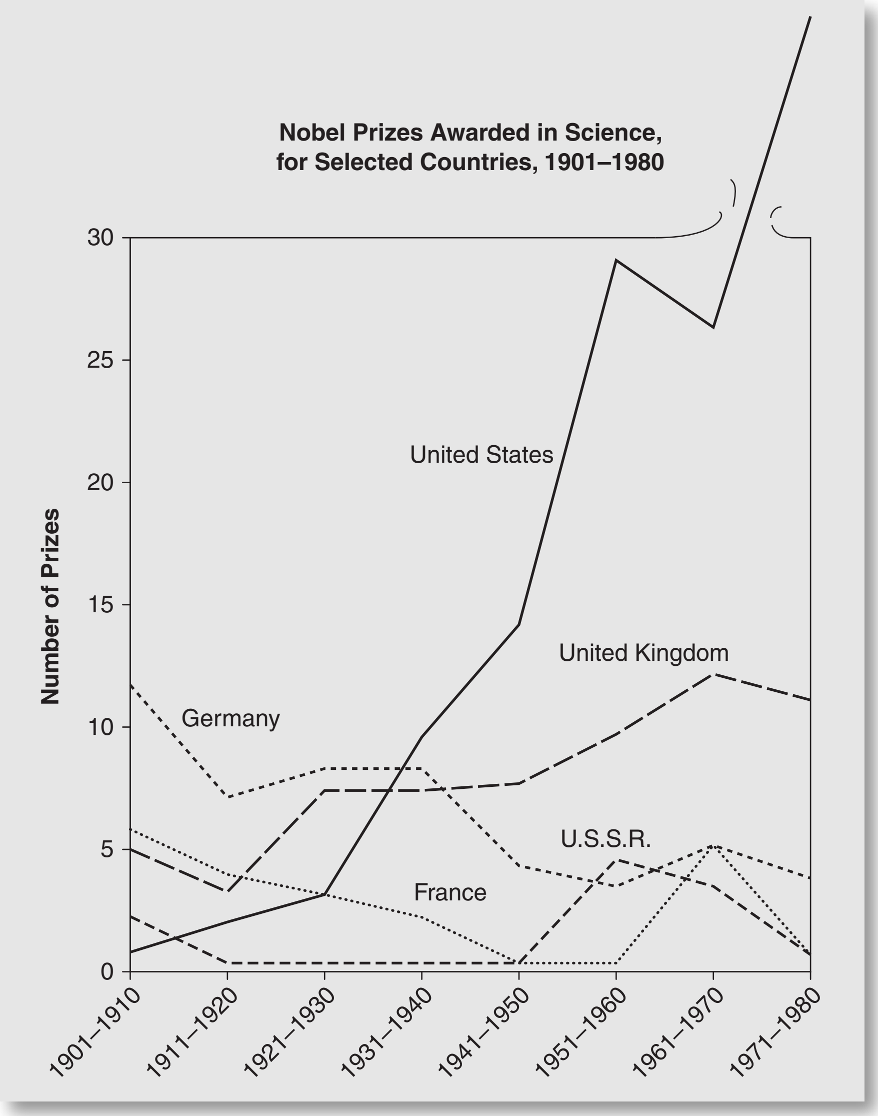


SOURCE: Tufte (2001). Used by permission from Graphics Press.

this chart includes the number of Nobel prizes in science awarded during a 4-year period. Of course, there will be fewer of them during this truncated period, as is the case with every country represented here.

What do the data show when the remaining 6 years of awards are included? Tufte (2001, p. 60) cleverly displays the results in a graph (see Figure 6.9).

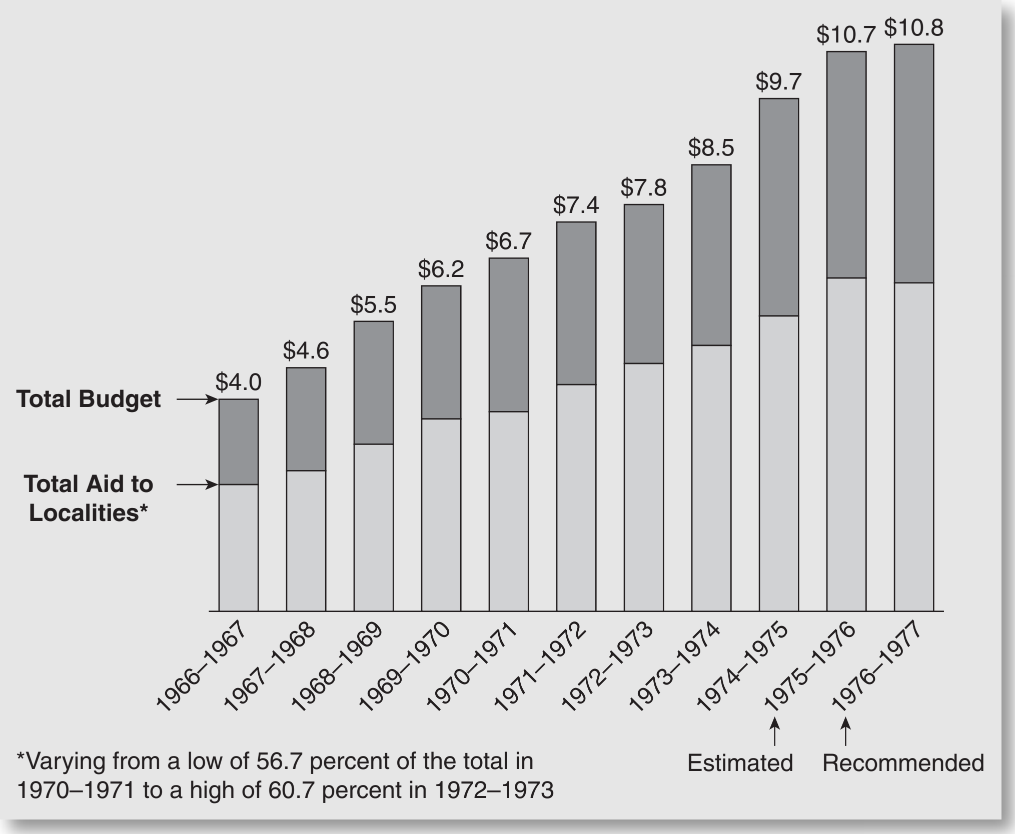
Figure 6.9 Honest Graphical Display of Nobel Prizes for Selected Countries



SOURCE: Tufte (2001, p. 60). Used by permission from Graphics Press.

Our third example (Figure 6.10) is a chart of the aid to localities provided by the state of New York during the 11-year period from 1966–1967 to 1976–1977 (Tufte, 2001, p. 61).

Figure 6.10 Misleading Graphical Display of Budget Amounts

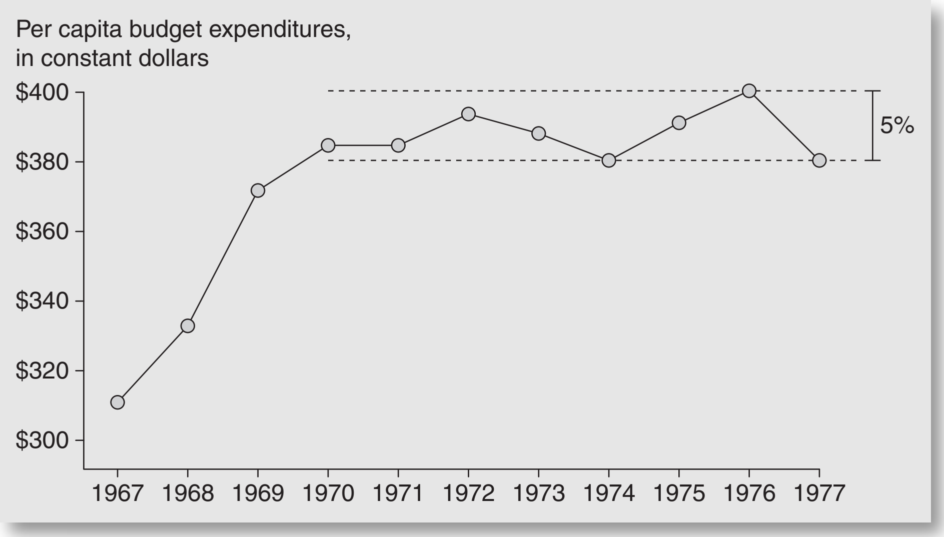


*Varying from a low of 56.7 percent of the total in 1970–1971 to a high of 60.7 percent in 1972–1973

SOURCE: Tufte (2001, p. 61). Used by permission from Graphics Press.

Another example of out-of-control government spending, isn't it? That's certainly the impression I get. But which two factors may be causing much of this apparent increase in aid to localities? Remember from Chapter 4's presentation of index construction that time-series financial data should always be adjusted for changes in the value of money and population size. Not doing so conflates real change with the change brought about, in this case, by inflation and population growth. How do you, therefore, correct the distortions contained in Figure 6.10? (The steps for doing this can be found in Chapter 4.) Figure 6.11 (Tufte, 2001, p. 68) provides a more truthful story by transforming the data into per capita, constant budget expenditures. The

Figure 6.11 Honest Graphical Display of Budget Figures



SOURCE: Tufte (2001, p. 68). Used by permission from Graphics Press.

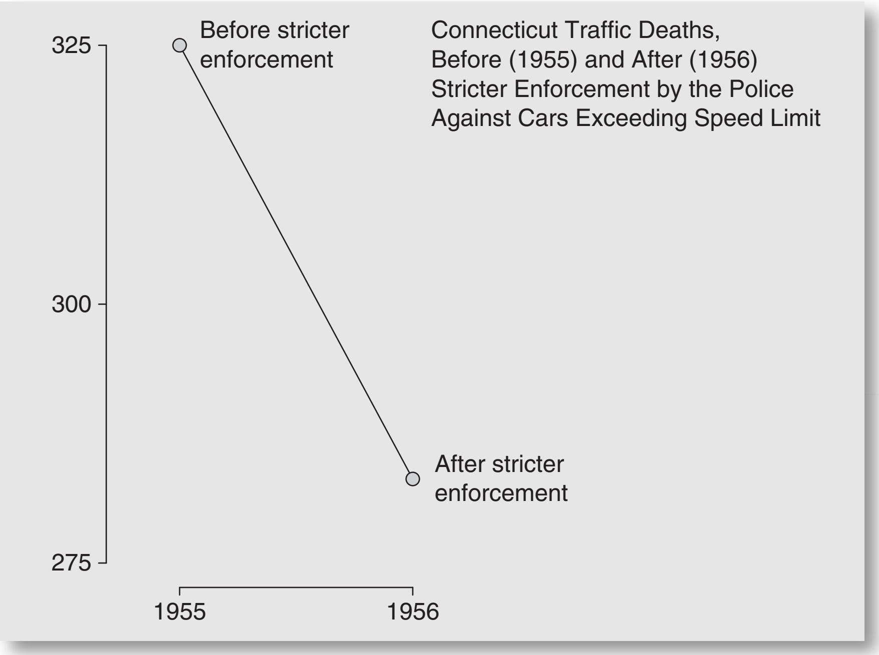
story is one of substantial growth in revenues to localities between 1967 and 1970 but relatively stable revenue streams thereafter. Note also the helpful visual cues Tufte (2001) provides in telling this graph's story by providing the two dotted lines and translating this range into the percentage figure of 5%. These attributes satisfy two of the graphic guidelines we'll see later: (1) don't require your viewer to perform the mathematical operations that help interpret the chart and (2) direct the observer's attention to the point you want the data to reveal by enclosing the data in some fashion.

Finally, imagine yourself a staunch supporter of stricter enforcement of highway speed limits because you believe, as did Governor Abraham Ribicoff of Connecticut in the mid-1950s, that such a crackdown would result in fewer traffic fatalities. You'd be rather pleased to share the graph shown in Figure 6.12 with those who would look and listen (Tufte 2001, p. 74, as originally presented in Campbell & Ross, 1968).

You'd also be engaged in a form of graphical tomfoolery. Why? There are at least four reasons:

1. The scale on the vertical *Y*-axis (i.e., the number of traffic fatalities) is severely truncated, exaggerating the 1-year change.

Figure 6.12 Misleading Graphical Display of Traffic Deaths in Connecticut



SOURCE: Campbell and Ross (1968, Figure d, p. 38). Used by permission. Copyright John Wiley & Sons.

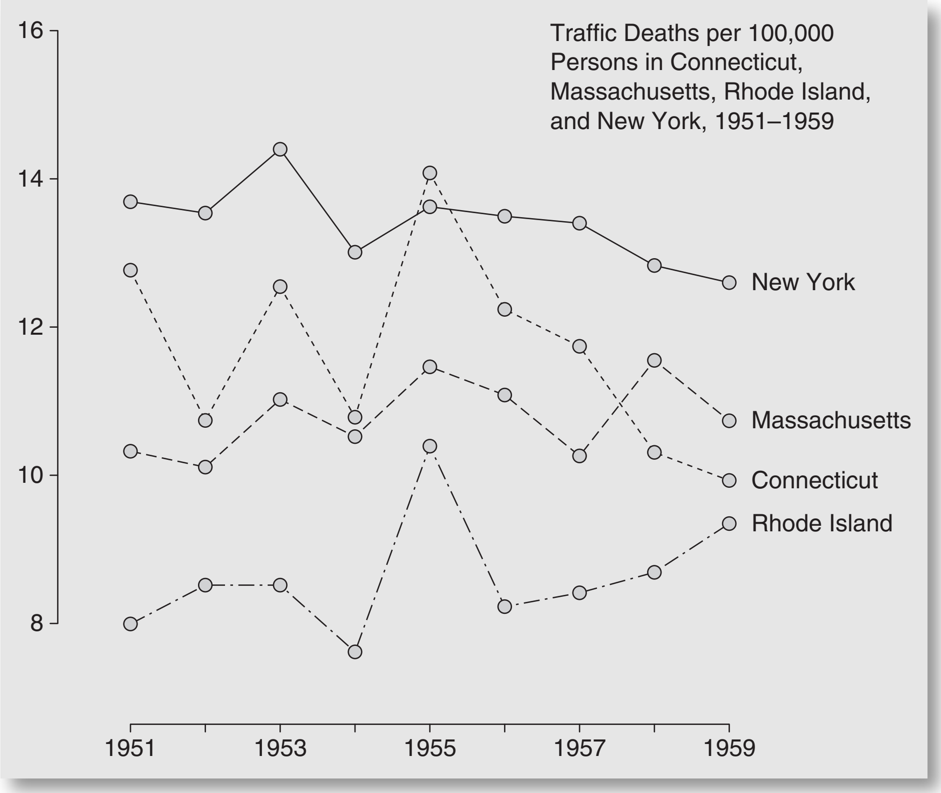
2. The graph includes only two data points yet wants to argue that this is a trend caused by stricter enforcement. Chance or random fluctuation is inherent in nearly every human phenomenon. Two successive observations of anything are likely to vary. Two observations do not make a trend.

3. The graph lacks context both in terms of a longer series of observations for Connecticut traffic fatalities and in comparisons with other states that, during this period, did not crack down on speeding.

4. The graph doesn't take population size into account. The more the people, the more the drivers, the more the miles driven, and the more likely you'll have traffic fatalities. Increases in fatalities may be an artifact of such increases. Decreases may be an artifact of population decline and/or fewer miles driven.

Figure 6.13 corrects for these possible distortions (Tufte, 2001, p. 75).

Figure 6.13 Contextualized Traffic Fatality Data



SOURCE: Tufte (2001, p. 75). Used by permission from Graphics Press.

You can note several features of this more useful graph that might shake your faith in stricter enforcement, although we would need to know much more about this time period than this graph alone provides. Assuming that the other three Northeastern states displayed here did not crack down on speeding in 1955, we see that all states experienced reductions in traffic fatalities. What might have caused this? Bad weather, a sharp rise in gasoline prices, or any number of things that might have reduced traffic overall, which could have led to declines in traffic fatalities. If Connecticut continued its stricter enforcement, however, the graph displays a continuing decline in fatalities, which the other states did not experience, so your faith in stricter enforcement may not be quite so shaken.

This graph illustrates another point to which we will return later in this book. To make a convincing case (e.g., stricter enforcement of speeding laws produces fewer traffic fatalities) requires that you rule out other possible causes for the patterns you observe (e.g., weather, increasing gasoline prices). Physicians might call this "diagnosis through exclusion." You do so in a statistical sense by "controlling for" these other factors, to which we will turn in the chapters ahead. Of course, you can't control for other possible causes unless you measure them. Statisticians give this the fancy name of **full model specification**, a concept to which we will return in later chapters.

Having viewed some bad and misleading graphs, let's consider next how our understanding of perception, cognition, and memory helps us design better graphs, charts, and tables.

PERCEPTION, COGNITION, AND MEMORY INFLUENCE THE INTERPRETATION OF GRAPHS ●

Applied policy research like the kind that this book examines is predicated, first, on asking what we need to know to help make better decisions. This will invariably involve a search for answers in the existing literature. Although research on the questions you're asking may be extensive, it will often come up short of answering the question in the specific context you face. In such cases, additional research may be needed, often involving the collection of data from which the statistics in this text help you extract the answers or tell a story. The narrative of such a story should arise from the data and your analysis of them. But there are at least three challenges facing you after the data have revealed their story:

1. Your primary audiences are likely to be very busy people with limited knowledge of statistics. This requires you to determine what the story line is, that is, what are the important points that come from your data and analysis. A memorandum and PowerPoint presentations are likely to be the forms of such communication. (This communication might also be a 30-second elevator speech, but we won't address that type of communication in this text.)

2. Your secondary audiences are other data analysts, who will look carefully at your methodology to make sure that your arguments are based on principled statistical analysis. This is especially likely to be the case if the

questions you're answering are important ones with powerful stakeholders, some of whom may not benefit from your conclusions. A technical appendix is the likely form for such communication.

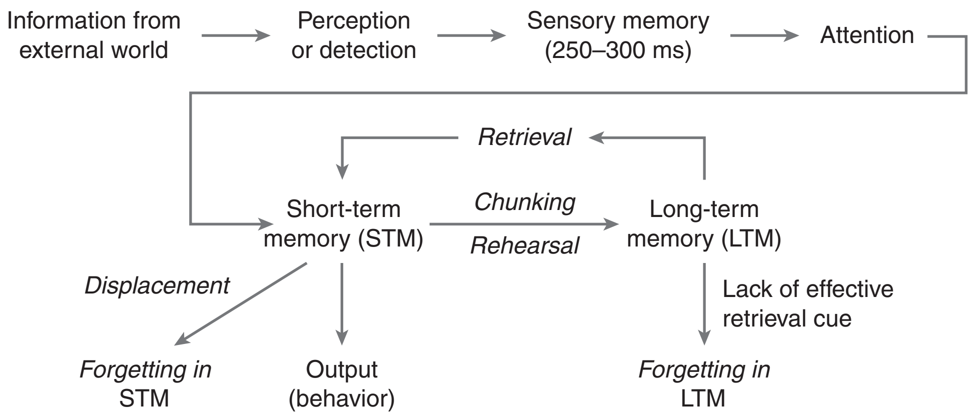
3. In both instances, you have to determine which point(s) you should make, how to direct or draw the reader (or listener) to those points, and how to make them stick.

Graphs and charts can help accomplish these objectives, especially if you understand what grabs peoples' attention and helps them cognitively process (i.e., understand, remember, and retrieve) the points you're trying to make. Research on perception and cognition provides some helpful clues here that will inform the guidelines for effective graphical display. Of course, there's a flip side to this coin. To wit: How do you reduce or eliminate things that get in the way of achieving these objectives? We'll provide some "don'ts" as well as "dos" in the guidelines you'll find below.

As with asking and answering questions, presenting and understanding a written argument depends on recognizing or paying attention to the message and processing that information in ways that aid understanding, memory, judgment, and retrieval. Graphs are a visual means of communication and are subject to the mechanisms and processes of sight. It is useful, therefore, to understand how those mechanisms and processes work.

The process begins with sensation; that is, we sense some kind of stimuli, whether words on a written page, the voice of a colleague down the hall, or a graph in a memorandum we're reading to help determine how to reduce teenage pregnancy or increase high school completion rates. When we focus on a page, our eyes can distinguish as many as 625 separate points in a square inch (Tufte 2001, p. 161). But that focus is fleeting, fixing on a particular spot for less than half a second before jumping to another. While our eyes are receiving millions of bits of information, our attention and perception attend to a limited amount of that information (Few, 2004, p. 95).

The brain processes the stimuli first in iconic or sensory memory, analogous in a computer to a keyboard buffer that serves as a (very) temporary waiting room. The electrical signals may then pass into short-term memory, analogous to a computer's working memory. Stimuli are translated here into meaningful terms to interpret and respond to the stimuli. Under certain conditions (e.g., rehearsal or repetition), these stimuli may be stored in what is called long-term memory, to be retrieved (with the right cues) at a later time.



Sensory memory is automatic and unconscious and, for these reasons, is considered as pre-attentive processing, attention, or recognition. We'll later see that there are aspects of graphs that are more (or less) likely to "grab your attention" through, say, the location of an item or the color of an object on a page. For example, recall how your attention was drawn in Figure 6.3 to the two errant height measurements because I enclosed them in a circle and pointed an arrow at them.

Short-term memory has two attributes that affect the principles of graphical display. Short-term memory is both (1) temporary and (2) limited in storage capacity.

The second of these is perhaps the more important consideration for graphical design. Many believe that people can store only seven plus or minus two items in short-term memory at any one time (Miller, 1956). (This conclusion may more precisely be the case for the short-term memory of college students recalling lists of digits, but we don't need to bother ourselves with such details or with such debates in cognitive psychology about whether short-term and long-term memory are really two different structures.) If your short-term memory register is full, moving a new item into it requires that an item already there be moved to long-term memory or forgotten. There are at least two implications of this fact for graphical design:

1. A legend with 10 or more colors or shapes for different categories will tax the reader to continually return to the legend.

2. Combining information into coherent patterns may help combine multiple bits of data into a single "chunk," thus saving space in short-term memory. You may recall that remembering the names of the Great Lakes is made easier by the mnemonic "HOMES," a chunk representing the first letters of Huron, Ontario, Michigan, Erie, and Superior.

Although we saw in our measurement chapter the implications that long-term memory harbors for asking and answering questions (e.g., use of language in which prior events were encoded and retrieval cues to help recall), it does not inform the design of graphs and tables as much as do pre-attentive visual attributes and short-term memory characteristics. Let's turn to these two forms of memory and flesh out their implications for the design of graphs and tables.

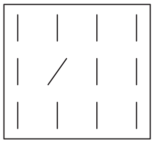
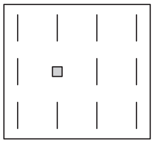
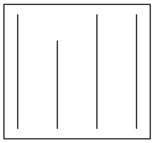
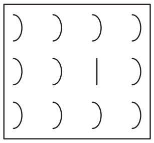
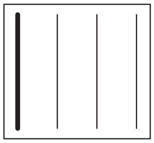
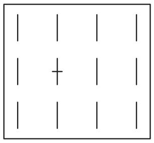
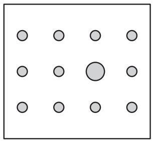
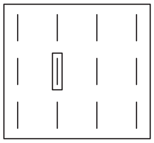
Recall (have you stored this in long-term memory yet?) that sensory memory or pre-attentive processing is unconscious and automatic. It may be an evolutionary relic of our distant ancestors' need to quickly recognize danger in their environment and take flight or fight. We may be hardwired to have our eyes move toward those stimuli that are different or that take on certain attributes. We modify below Colin Ware's (2000) and Stephen Few's (2004) specification of categories and attributes that have lessons for the design of tables and graphs.

Category	*Attribute*
Form	Orientation Line length Line width Size Shape Curvature Added marks Enclosure
Color	Hue Saturation Lightness
Spatial position	Two-dimensional positioning Proximity
Context	Contrast

Attributes of Form

Differences in each of the attributes of form are illustrated below. Note how your eye is drawn to those attributes that differ from those with which they are enclosed (Few, 2004, p. 99).

Attributes of Form

Attribute	*Illustration*	*Attribute*	*Illustration*
Orientation		Shape	
Line Length		Curvature	
Line Width		Added Marks	
Size		Enclosure	

SOURCE: Few (2004, p. 99).

Attributes of Color

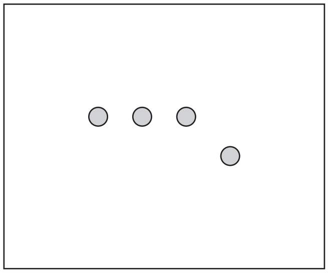
Color is actually composed of three different attributes: hue, saturation, and lightness. (It's threes again.) We usually think of hue as color (e.g., red, green, blue), but we'll call it hue because of the different pre-attentive properties of saturation and lightness. Saturation is the degree to which a hue displays its quality (Few, 2004, p. 101).

Lightness is the degree to which a color is characterized by being fully black or white.

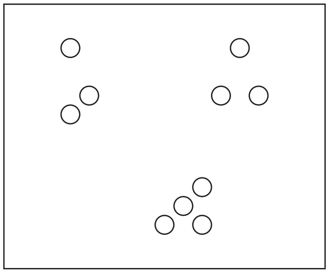
Consider the difference between hue and, say, line length as they relate to graphical design. We naturally ascribe greater value to longer line lengths than to shorter ones. In other words, line length attributes are encoded quantitatively. But is that the case with hue? Is green perceived to be larger than red? No. Different hues are perceived as different categorically but not quantitatively. Different hues, therefore, are of little assistance if we mean to represent quantitative differences. Saturation and lightness, however, can be perceived quantitatively and used as an aid in our graphs in expressing differences of degree. Therefore, use hues in your graphs to demark different categories or draw attention to objects or numbers, not to suggest differences in quantity. Also, remember that colors are more expensive to print and don't necessarily photocopy well (unless using a color copier). In other words, stick with shades of gray to convey quantitative differences in graphs, a guideline to which we'll return below.

Attributes of Spatial Position

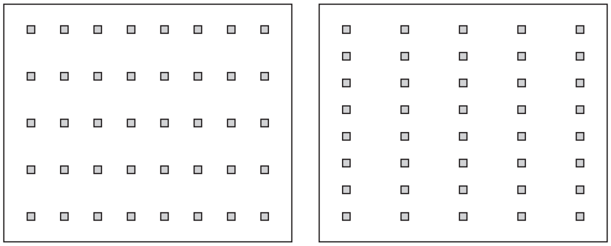
Two-dimensional (2-D) spatial positioning on a page can be illustrated by the following box. Here, the eye gravitates toward the one object that differs in position from the others.



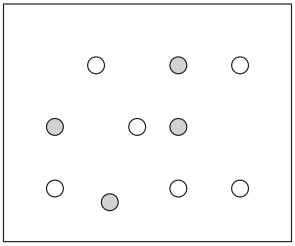
The eye's attention will also be drawn to the grouping of objects in proximity to each other. We perceive the 10 objects in the following box, for example, as belonging to three groups because of their spatial proximity.



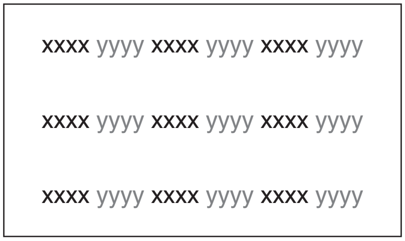
This *principle of proximity* (Few, 2004) is based on work by the Gestalt School of Psychology in the early 20th century, as are the other principles noted in this section. We can use these principles, for example, to subtly instruct a reader's eyes to scan the rows or columns of a table by grouping the observations more closely in a row or in a column, as the following figure demonstrates. Your strong visual inclination is to read the left box as rows and the right box as columns as a consequence of the principle of proximity.



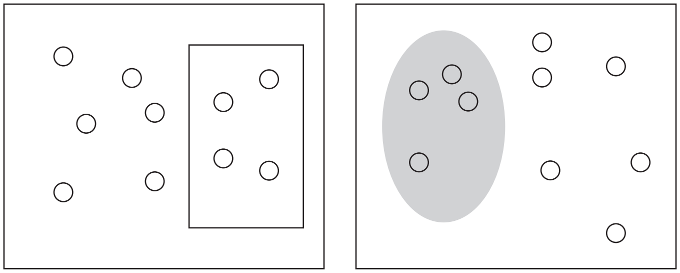
We also perceive objects that are similar in size, color, shape, and orientation as belonging to the same group. This is referred to as the *similarity principle*. This principle or effect is illustrated in the figure below. The gray circles are visually recorded as members of the same group because of their similarity of color, no matter what their proximity is.



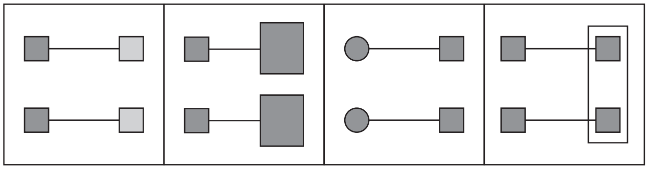
We can again use this principle (e.g., using similarity of both color and shape) to guide the reader's eye, as the following figure does in causing the eye to see columns instead of rows.


xxxx yyyy xxxx yyyy xxxx yyyy
xxxx yyyy xxxx yyyy xxxx yyyy
xxxx yyyy xxxx yyyy xxxx yyyy

The *principle of enclosure* achieves a similar result. It uses visual borders to set off elements in the enclosure from those outside it, again suggesting group membership. It also achieves the effect of drawing your eye first to those elements inside the enclosure, which in graphs and tables can be used to highlight what is important or what you want the reader to view first, as illustrated in the following figures:



The *principle of connection* argues that objects that are connected by a line are conceived of as part of the same group. Indeed, this effect trumps the proximity and similarity effects, although not the enclosure effect, as you can see in the following figure (Few 2004, p. 114).



SOURCE: Few (2004, p. 114).

Points on a graph that are not connected by lines are difficult for us to connect without their aid. They reveal the shape of the data as well as link the objects. They are commonplace (essential) in graphs of time-series data.

Attributes of Context and Contrast

Every attribute of visual perception is influenced by its context. Objects can seem to be brighter when lit in a dark room than in one flush with sunlight. Hues can be made to appear lighter or darker depending on the saturation of the surrounding colors. In the context of a table or graph, for example, it is easier to read black text on a white background.

Our visual senses evolved in ways that cause us to be particularly aware of differences, which are themselves defined by the contrast of an object to its context. Our brain is not only drawn to these differences but also tries to make some sense of them. Contrast is especially effective if one item in a group of others has a distinguishing attribute in contrast to all others. The more items with that attribute, however, the less any one of them will stand out.

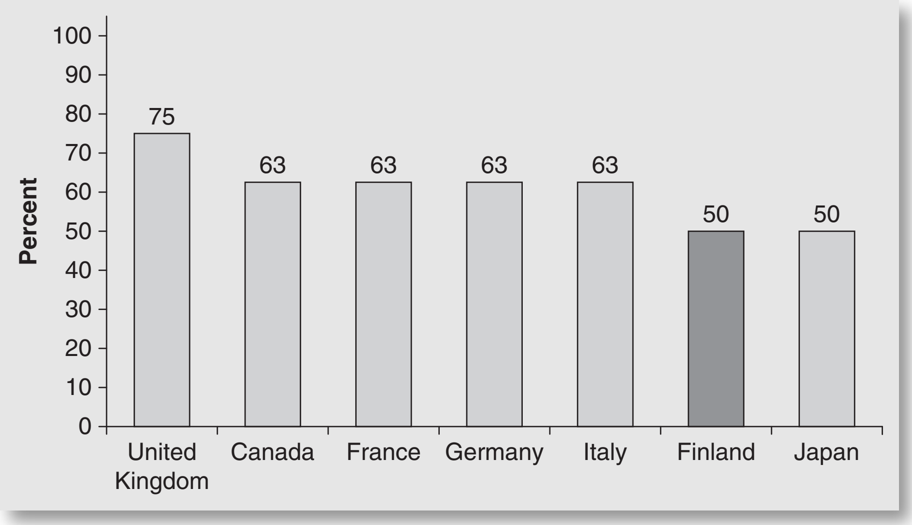
Don't make your reader work at the comparisons by requiring additional calculations to make the point you want the reader to come away with. If you want the reader, for example, to focus on the rates of child mortality in the United States in contrast with selected developed countries, as shown in Figure 6.14, you could do so by expressing other countries' rates as a percentage of the U.S. rate.

Tables and graphs should provide the reader with "affordances," that is, characteristics that reveal their purpose and use. A handle on a door tells you implicitly to pull; a plate, to push. Coloring Finland dark gray in the child mortality graph (Figure 6.14) signals to the reader that I'm going to devote particular attention in the text to that country in trying to explain why its child mortality rate is half that of the United States.

Limits to the Perception of Attributes

There are, of course, limits to our abilities to perceive attributes. Ware (2000), for example, reports that people can distinguish no more than about eight different hues in a graph, about four different orientations, and about four different sizes. Few (2004, p. 106) suggests that, with the exception of hue and shape, it's best to limit the number of different attributes to no more than four. In addition, pre-attentive processing tends to be limited to one attribute at a time. That is to say, it's okay to vary lightness (especially if, say,

Figure 6.14 Under-5-Year-Old Child Mortality Rates as a Percentage of the U.S. Rate (8 per 1,000), 2003 Estimates



SOURCE: World Health Organization, www.who.int/healthinfo/statistics/mortchildmortality/en/index.html, accessed June 18, 2008. The chart was constructed using Excel.

the four variations of gray are quite distinct) but not to add differences in shape at the same time.

Tufte (1990, p. 90) recommends the use of hues that are found in nature, especially the lighter shades of gray, yellow, green, and blue, which are easy on the eye. This is particularly useful if you don't want to draw attention to numbers or objects. Conversely, objects painted in deeply saturated versions of any hue will draw the eye toward those items.

● GUIDELINES FOR CREATING EFFECTIVE GRAPHS

We have already provided a few guidelines for graphical display above in the context of aspects of visual perception. We'll repeat them here and add more that are consistent with our understanding of sensory and short-term memory and that also build on experience and good practice. You might even use these dos and don'ts as a checklist in creating graphs and tables.

Don'ts

- Do not use three dimensions to represent one. A third dimension, say, in a 3-D bar chart will exaggerate visual differences because you're now comparing volumes instead of line lengths or areas. Three-dimensional graphs also make actual values of lines or bars difficult to compare with grid values. In short, don't use 3-D graphs.

- Do not use severely truncated ranges on your vertical axis. This causes the visual impression to depart substantially from what the data say. So too does fake perspective. Don't use either.

- Do not select arbitrary starting points or change the scale within a graph.

- Do not enclose an entire graph by using a box or color because it causes the reader to attend to those elements in contrast to the data (Few 2004, p. 113).

- Do not use vivid line fill patterns in bar graphs (e.g., cross-hatching) because they can cause a dizzying visual effect called "moiré vibration" (Few 2004, p. 64).

- Do not introduce elements in a graph that puzzle the reader (e.g., unusual interval values on the *Y*- or *X*-axis). This may inspire mystery but lead the reader to look for clues to solve the mystery rather than understand what the data say.

- Do not lead a reader to focus on an element by contrasting it with others in size, shape, orientation, and so on, unless you want the reader to attend to those objects or numbers.

- Do not make your reader work by requiring additional calculations to make the point you want the reader to come away with.

- Do not manipulate the aspect ratio of your graph (i.e., the ratio of height to width) to exaggerate or minimize the pattern or slope of the data.

- Do not design graphs for the purpose of showing off your technical proficiency. If your graphs cause the reader to think, "Gee, I wonder how she did that," you've led the reader off the point. Graphs should help tell the story, not *be* the story.

Dos

- Devote most of the graph's ink to the data instead of the supporting components, such as grids (Tufte, 2001). Many nondata elements are unnecessary to the meaning or impact you want the graph to have and can get in the way of that message by moving the eye away from the data. Delete or mute the nondata elements.

- Avoid clutter; it taxes our sensory and cognitive capacities. For example, use no more than five different distributions in a line graph. Use multiple boxplots when displaying more than five distributions.

- Label in plain English, not with mnemonics or abbreviations (ditto for tables and explanations in text).

- Use black on white for text.

- Label data elements directly on the graph itself. Minimize the use of legends whenever possible. Legends require storing the meaning of each line or bar in short-term memory, which is limited in time and space.

- Because our eyes are practiced in detecting deviations from the horizon, graphics should tend toward the horizontal, where "cause" is displayed on the horizontal (*X*) axis and effect, vertically (on the *Y*-axis). In general, make your graphs wider than they are tall.

- Use different hues in your graphs to demark different categories or draw attention to objects or numbers, not to suggest differences in quantity.

- The width of bars in a bar chart should be approximately equal to the white space between them. (There's no scientific evidence to support this prescription that I know of. I just think it looks clean and clearly displays each bar.)

- Provide enough context to accurately present the data (e.g., more than two time points, comparisons with similar units).

- Tables and graphs should provide the reader with "affordances," that is, characteristics that reveal their purpose and use.

- Rotate bar charts so that categorical labels and bars run horizontally from high to low if your purpose is to make such comparisons and to emphasize those categories that are at the top and the bottom of such rankings.

- Adjust for inflation in graphs that include time-series financial data. Adjust for population change in data that can be reasonably supposed to be influenced by population size.

- Sort, group, organize, sequence (e.g., top to bottom, left to right), and prioritize the data (through the use of contrasting visual attributes such as more saturated hues) in ways that serve the graph's purpose.

- The top and left positions of a graph in Western cultures tend to be considered the beginning. Place the categories you want to receive the greatest emphasis in those positions.
- Use simple geometric forms, such as straight lines, circles, and squares.
- Locate explanatory text as close as possible to the data to which it refers.
- Add good titles that clearly tell the reader what the graph's story line is.
- Annotate extensively, including the sources of data.
- Colors do not have a natural visual hierarchy; varying shades of gray show quantities better and are more likely to photocopy accurately.
- When creating a related series of graphs, use the same characteristics in each (e.g., scale, aspect ratios, colors, shapes, fonts).
- Editing and revision are as essential in producing good graphs and tables as they are in good writing.
- Always think about the "yo' momma" rule: Would your mother understand what you are trying to communicate in this graph?
- Finally, apply none of these principles rigidly. Use good judgment. Think about the purpose of your graph and how the elements work to achieve that purpose.

THE TOOLS OF GRAPHICAL DISPLAY ●

Histograms

Histograms display the distribution of continuous variables. Histograms are, therefore, useful for at least three different purposes:

1. To visually display the extent to which the distribution can be characterized as "normal," which is a characteristic that many statistics assume variables have

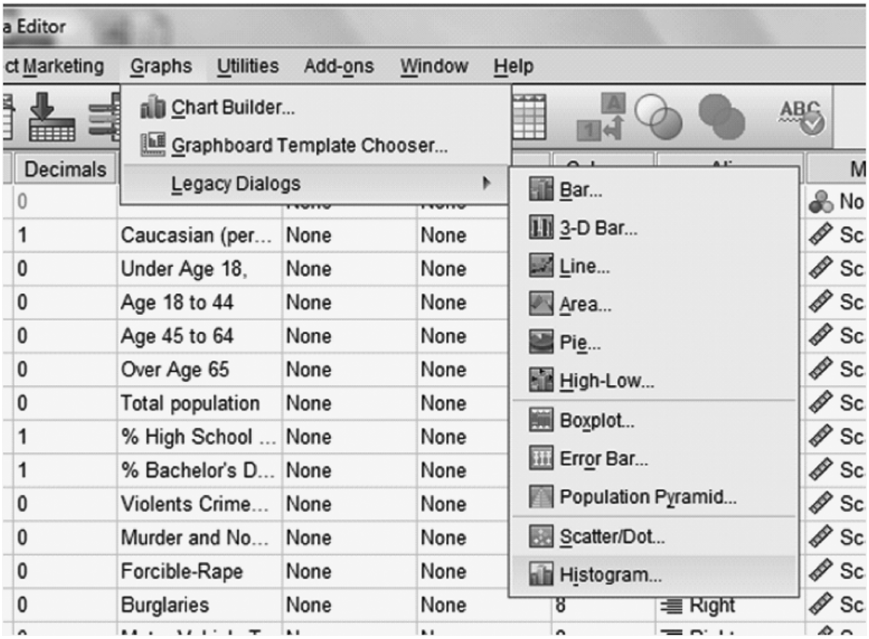
2. To draw substantive conclusions about the overall pattern of the distribution or individual observations within that distribution (e.g., half of the cities in your study have child mortality rates of less than 4 in every 1,000 births; a particular city falls within the upper half of the set of similar cities in terms of its per capita homicide rate)

3. To visually identify unusual observations or outliers

Let us create a histogram of the frequency or number of violent crimes in the largest cities in the United States in 2004.

Step 1: Open the Community Indicators data set in SPSS.

Step 2: Click on Graphs/Legacy Dialogs/Histogram, as shown on the screen save below:

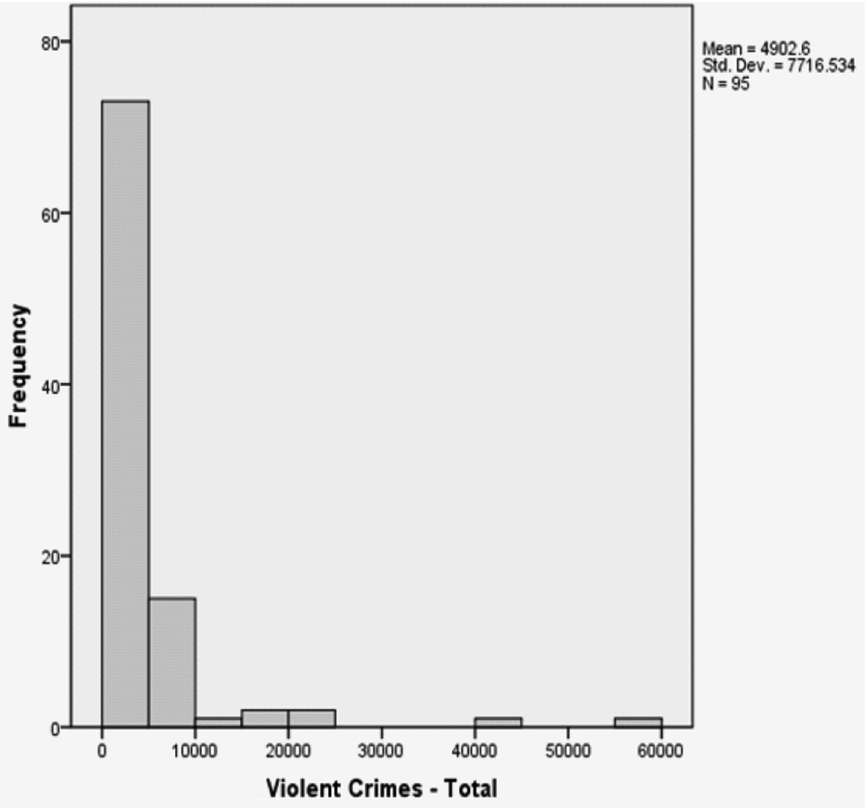


Decimals						M
0						No
1	Caucasian (per...	None	None			Sc
0	Under Age 18,	None	None			Sc
0	Age 18 to 44	None	None			Sc
0	Age 45 to 64	None	None			Sc
0	Over Age 65	None	None			Sc
0	Total population	None	None			Sc
1	% High School ...	None	None			Sc
1	% Bachelor's D...	None	None			Sc
0	Violents Crime...	None	None			Sc
0	Murder and No...	None	None			Sc
0	Forcible-Rape	None	None			Sc
0	Burglaries	None	None	8	Right	Sc

Step 3: Move VIOLENTCRIME (Violent Crimes – Total) into the variable box, and hit OK.

The histogram that SPSS produces has a range for each bar that SPSS defines for you. These are not always helpful or informative. Fortunately, we can change them, as we can nearly all the attributes of a graph produced by SPSS. We may not need to do so in a preliminary analysis of the data, but we almost always have to do so before presenting the charts to a target audience. The default features of these charts invariably violate our guidelines for graphical display.

Here's what your histogram for the number of violent crimes in 95 of these 98 cities will look like:



How might this graph be improved? You probably shouldn't bother. It's a fatally flawed graph because it doesn't take into account the population size of each city, thus violating an important graphic design guideline.

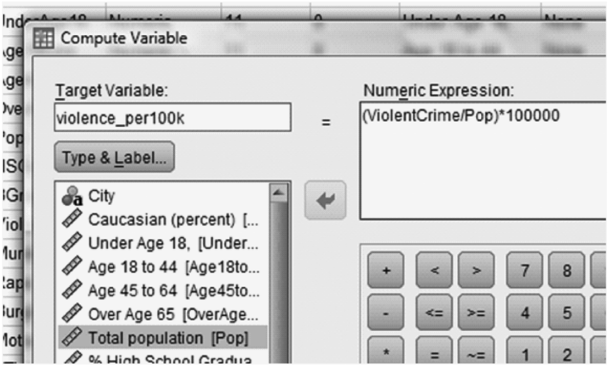
We should transform the variable before asking for its histogram. One of the more easily interpretable and communicable transformations is to create a variation of per capita for total violent crimes. We don't, however, merely want to divide, say, the total number of violent crimes by the total population of each city. We would find, for example, that there were 0.007 violent crimes per each resident of New York City in 2004 (55,688/8,008,278) and 0.008 violent crimes per each resident of St. Paul, Minnesota (2,408/287,151).

These per capita numbers, however, are difficult to communicate in a way that audiences can "get their arms and brains around." And who wants to keep track of decimal places?

Per 100,000 is a potential candidate for the type of transformation we want to perform, and you can help your audience grasp the resulting number by an explicit reference to the number of people watching a college football game on a Saturday in the fall in Pasadena, California (University of Southern California), or in Ann Arbor, Michigan (University of Michigan). It may also be easier to imagine 100 people (e.g., two bus loads of people) or 1,000 (e.g., a small movie theater in a multiplex). But let's use 100,000. (By the way, the number we're about to compute is a yearly figure per 100,000 residents. We could further divide this number by 365 if we wanted, say, to know how many violent crimes would take place on any day (on average) throughout the year, thus creating a daily rate, which may be easier to grasp, although likely to be very small.)

Step 4: A new variable for violent crimes is created by dividing the original variable by the population size for each city and then multiplying that number by 100,000.

In SPSS, this is accomplished by clicking Transform/Compute Variable and entering the following information in the Target Variable and Numeric Expression boxes (as below). Click OK.



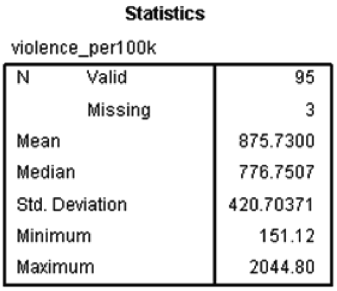
Compute Variable

Target Variable: violence_per100k

Numeric Expression: (ViolentCrime/Pop)*100000

Step 5: After creating this new variable, run summary statistics to make sure the data transformation appears to have executed properly.

Your summary statistics table should look like the following:



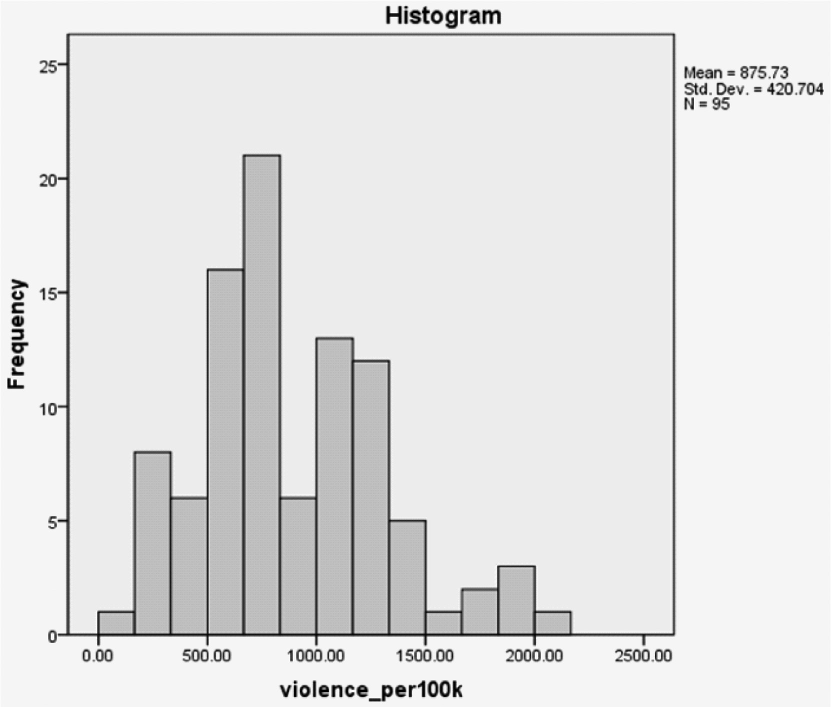
Statistics

violence_per100k

N	Valid	95
	Missing	3
Mean		875.7300
Median		776.7507
Std. Deviation		420.70371
Minimum		151.12
Maximum		2044.80

This transformation may have created not only numbers that are more communicable but also variables that are better "behaved" versions of violent crimes. If we had run descriptive statistics on the old variable, we would have noticed that the means and medians of the newly transformed variables are closer together and the ratios of kurtosis and skewness to their respective standard errors are close to or better than our guideline ratio of 2:1.

The histogram for our transformed violent crime data is presented below. (*Note:* we could skip the step of Graph/Legacy Dialogs/Histogram by asking for this chart within the descriptive statistics command we used to produce the table above. There's a "Charts" button in the Frequencies dialog box that we could have clicked to request a histogram.)



How might this chart be improved?

To illustrate how to change nearly any attribute of a default graph in SPSS, consider the following changes:

1. Add an informative title.

2. Rename the horizontal and vertical titles, repositioning the vertical title to run horizontally.

3. Create narrower ranges for each of the "bins" or what appear here as bars.

4. Remove decimal places from the horizontal scale, and increase their font size.

5. Increase the size, and move the summary statistics inside the chart area.

6. Make the background white (color ink cartridges are expensive, and the background baby blue adds nothing to the story).

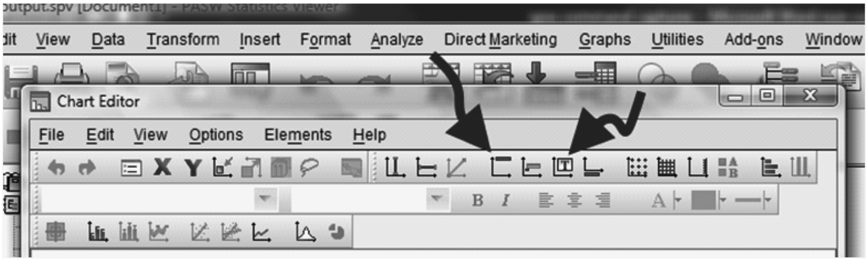
7. Change the color of the bars to gray, which they are not in the original printout.

8. Eliminate the top and right borders.

How would you do all this?

Step 1: Move your cursor to somewhere in this histogram in your output file, right click, and select Edit Content/In Separate Window. This will launch SPSS's Chart Editor, which will enable you to change any of the attributes that you want to change above.

Step 2: Move your cursor to the button on the toolbar to Insert a Title, and click on it, as shown in the arrow on the left below.



Type the following new title in the highlighted box: Violent Crimes in Major U.S. Cities Per 100,000 Residents, 2004. Double click on "Histogram" and delete.

Step 3: Highlight the bottom title on the *X*-axis, and instead of the variable name, type "Violent Crime Rates."

Step 4: Double click on the Y-axis title. In the Properties box that appears, change the Preferred size to 12 points. Hit Apply. Open the tab Text Layout and turn on the radial button for horizontal. Hit apply. Left click on the Y axis and replace "Frequency" with "Number of Cities." If you want to have the Y axis title appear on two lines, move your cursor before "Cities" and hit Shift/Enter.

Step 5: Double click anywhere inside one of the bars. Click on the Fill & Border tab, and highlight one of the gray boxes in the color palette. Click Apply.

Step 6: In the Binning tab, X axis box, activate the Custom radial button Interval width and enter 50. Hit apply. Double click on one of the numbers of the horizontal scale.

- In the Text Style tab, change Preferred Size to "12." Click Apply.
- In the Scale tab, change maximum to "2200." Click Apply.
- In the Number Format tab, insert "0" (i.e., zero) into the Decimal Places box. Click Apply and Close.

Do the same thing for the numbers on the vertical axis.

Step 7: Double click on Mean or Std. Dev to the right of the chart.

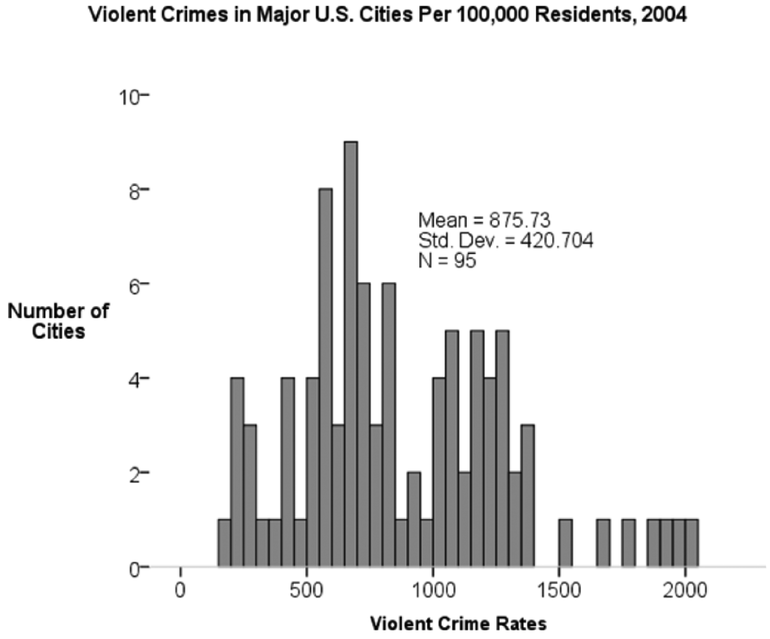
Move your cursor to the border of the box in which these statistics reside until your cursor changes to a figure that looks like the four arrows of a compass, and drag that box into the upper right corner of the chart. Change minimum size to 12 in the Text Style tab. Expand the box (if needed) so that each of these three statistics is on only one line. Click Close.

Step 8: Double click on the blue background.

- In the Fill & Border tab, click on the Fill Box, and click on the white box in the color palette.
- In the Border box within the same tab, select the white or transparent palette. Click Apply and Close.

Step 9: Click on Edit at the top of your screen, and select Copy Chart, which you can paste into an MS-Word document to be submitted to whomever you'd like.

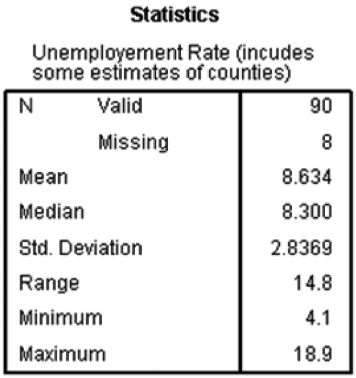
Your chart should look something like this:



Scatterplots

Scatterplots or scattergrams enable the user to display the patterned relationship between two continuous variables. We have already seen this graphical tool in the graph of self-reported student height with measured height.

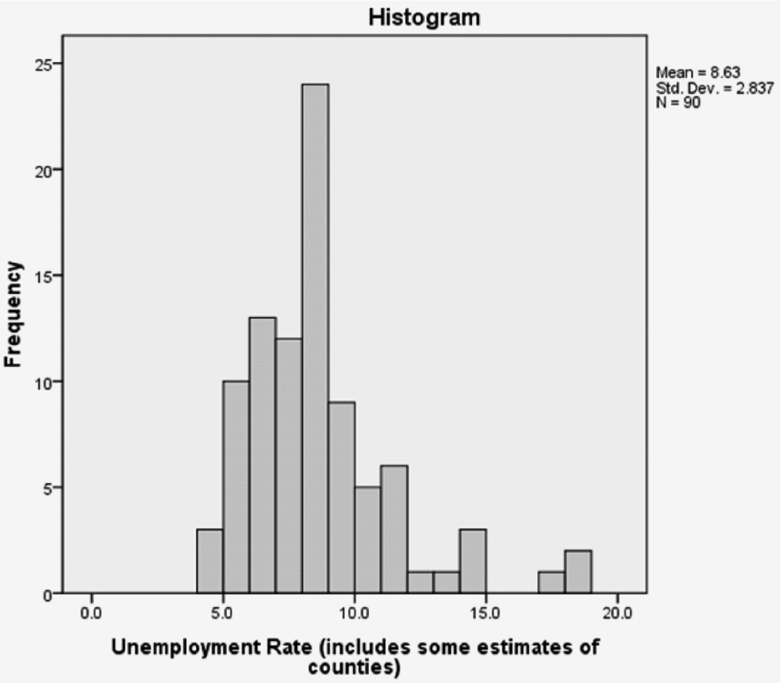
Let's illustrate how to produce a scattergram by examining the relationship between violent crime and cities' unemployment rates. We would, of course, first examine the summary statistics for the second variable, as we did for violent crime rates to make sure there weren't any stray or inexplicable observations. Here's what the summary statistics and histogram (unedited) for unemployment rates look like:



Statistics

Unemployement Rate (incudes some estimates of counties)

N	Valid	90
	Missing	8
Mean		8.634
Median		8.300
Std. Deviation		2.8369
Range		14.8
Minimum		4.1
Maximum		18.9



Some of the rates are high but believable. So we'll move ahead with a request for a scattergram of the unemployment and violent crime rates across the cities for which we have nonmissing values on these two variables.

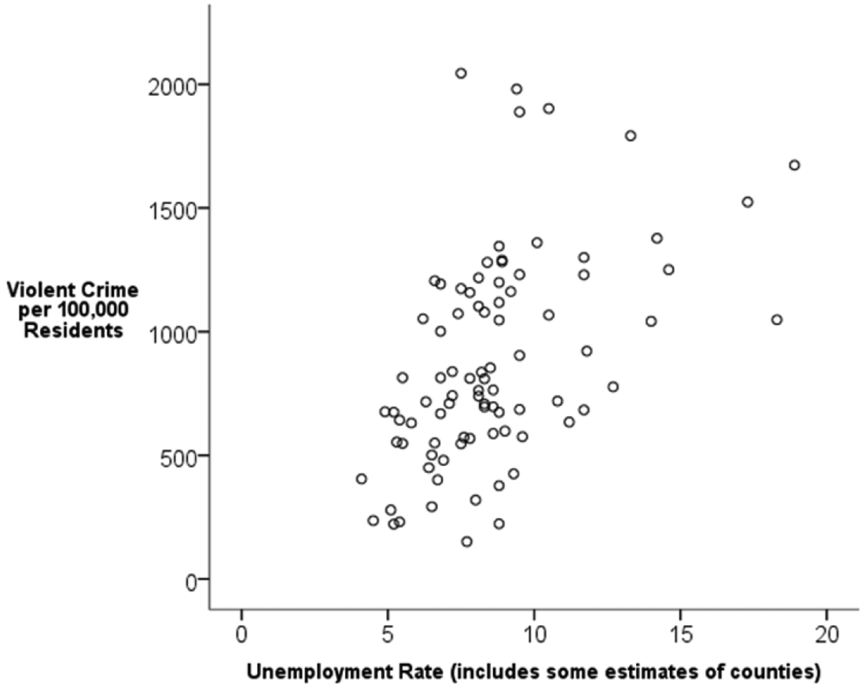
Step 1: Create dichotomous versions of the % of population under 18 and the cities' unemployment rates.

Step 2: Click on Graph/Legacy Dialogs/Scatter/Dot.

Step 3: Select Simple Scatter, and click on the Define button.

Step 4: Move "Crime per 100,000 Residents" into the *Y* and "Unemployment Rate" into the *X*-axis variable boxes. Click OK.

These steps should produce (after some editing) something like the following scattergram:



It would be useful, of course, to augment this graph with statistics that provide summary measures of the degree to which these two variables are related (e.g., an increase of 1 percentage point in unemployment is associated with an average rise of *X* number of violent crimes per every 100,000 residents in a city). But our purpose here is to demonstrate how to produce a scattergram. Its interpretation would be aided by the addition of correlation and regression statistics (as well as descriptive statistics).

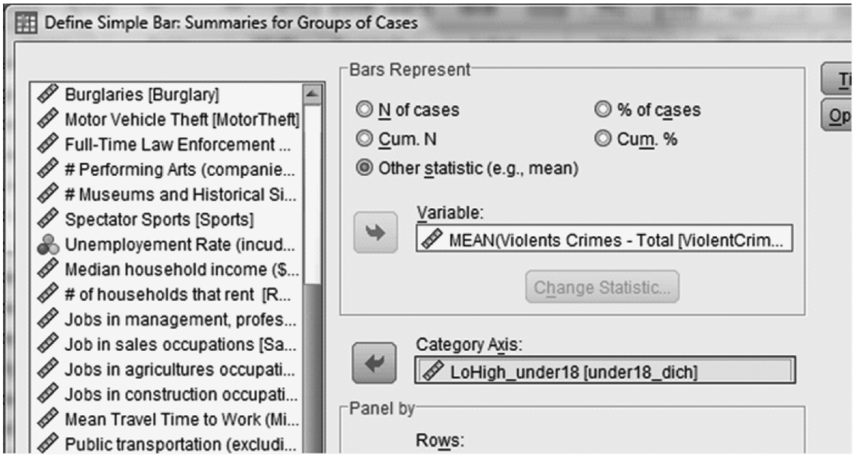
Bar Charts

Bar charts display the frequency or percentage of one or more continuous variables by one or more **categorical variables**. Let's examine the violent crime rate (a continuous variable) by two **dichotomous variables** that roughly divide the 98 cities in the community data set in half according to (1) the percentage of their residents under 18 years of age and (2) the city's unemployment rate. We'll first look just at violent crime rate per 100,000 by whether a city has a low or high proportion of residents under 18 years of age and then create a bar chart that looks simultaneously at both dichotomous variables.

Step 1: Click on Graphs/Legacy Dialogs/Bar. Select Simple, and click on the Define button.

Step 2: In the Bars Represent box, click on "Other statistic" (e.g., mean), and move the variable for violent crime per 100,000 residents into the variable box.

Step 3: Move the dichotomous variable for the proportion of residents under age 18 into the Category Axis box, as indicated in the following screen capture. Click OK.



Define Simple Bar: Summaries for Groups of Cases

Burglaries [Burglary]
Motor Vehicle Theft [MotorTheft]
Full-Time Law Enforcement ...
Performing Arts (companie...
Museums and Historical Si...
Spectator Sports [Sports]
Unemployement Rate (incud...
Median household income ($...
# of households that rent [R...
Jobs in management, profes...
Job in sales occupations [Sa...
Jobs in agricultures occupati...
Jobs in construction occupati...
Mean Travel Time to Work (Mi...
Public transportation (excludi...

Bars Represent

○ N of cases ○ % of cases
○ Cum. N ○ Cum. %
● Other statistic (e.g., mean)

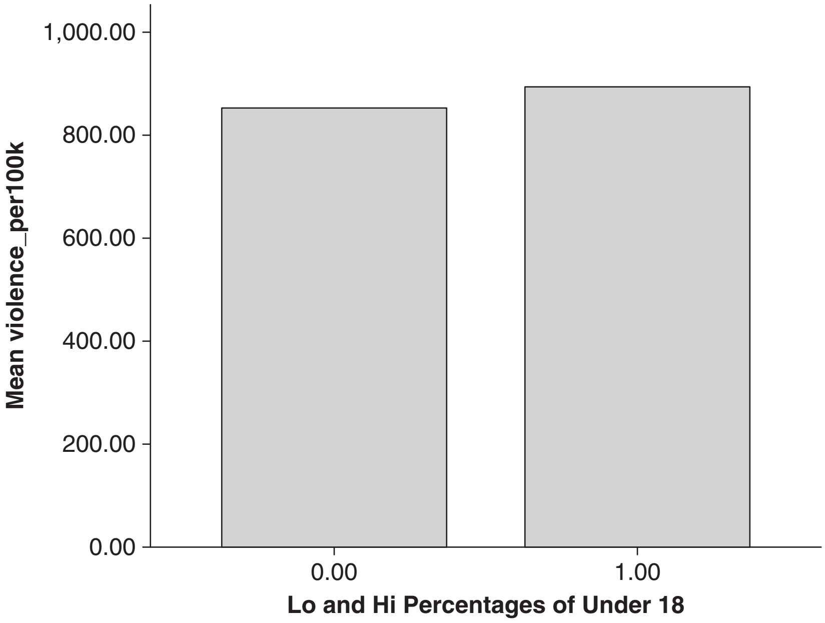
Variable: MEAN(Violents Crimes - Total [ViolentCrim...

Category Axis: LoHigh_under18 [under18_dich]

Panel by

Rows:

These commands produce the unedited bar chart below:

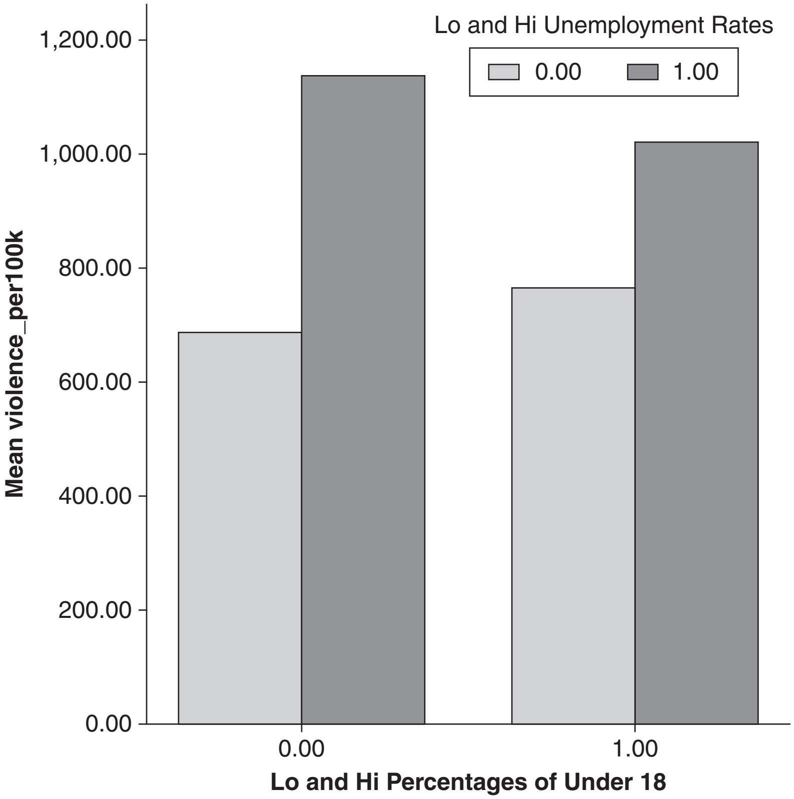


Rates of violent crime appear unaffected by the proportion of a city's residents who are under 18 years of age. We typically wouldn't continue to use this variable in analyzing violent crime rates via bar charts, but let's illustrate how we could display two categorical variables in the same bar chart.

Step 1: Click on Graphs/Legacy Dialogs/Bar. Select Clustered, and click on the Define button.

Step 2: In the Bars Represent box, click on "Other statistic" (e.g., mean), and move the variable for violent crime per 100,000 residents into the variable box.

Step 3: Move the dichotomous variable for the proportion of residents under age 18 into the Category Axis box and the dichotomous variable for unemployment rates into the Define Clusters By box. Click OK.



The unedited bar chart above shows rather pronounced differences in violent crime rates in cities that vary in their unemployment rates, as we had seen earlier in the scattergram of the two continuous versions of these variables.

Bar Charts and Rankings: Vertical Versus Horizontal Display

Consider the 27 items included in the first question asked in the Orange County Public Perception Survey. The respondents were asked how important each of a number of issues is "for you." The response categories included Very Important, Important, Somewhat Important, and Unimportant. You can easily request summary statistics and bar graphs for all 27 of these variables, but how would you present their complete results, say, in a single graph or table?

A typical solution to this problem, and one with some justification in both design and information processing, is to order them from most to least important. Listing them alphabetically requires a reader to scan the list and try to remember the order of the 27 items, a cognitive task that far surpasses our short-term memory capacities.

Interestingly, we have 27 different variables in this instance, not a single variable with 27 different categories. You will have to do a "two step" in order to display all the information contained in responses to these 27 questions. First, calculate, say, the percentage of respondents who judged each item to be "Very Important."

In tabular form, such a table might look like Table 6.2.

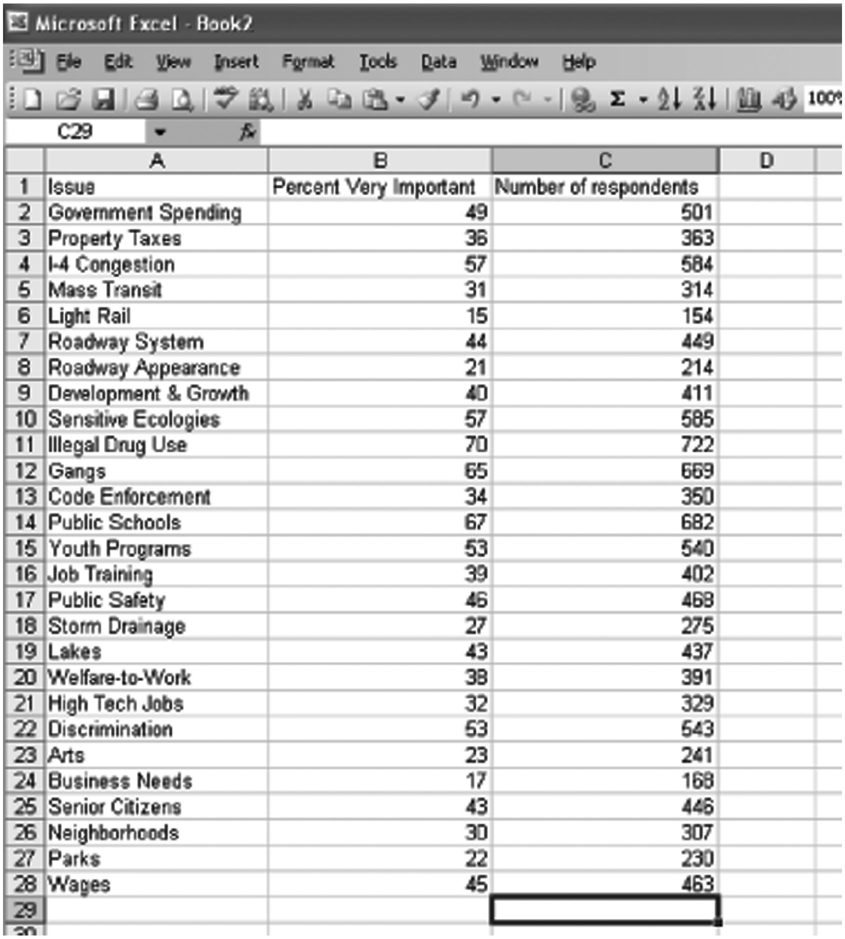
Table 6.2 Rank of Issues Considered "Very Important" by Citizens of Orange County, Florida, 2000

Issue	*Percentage (Number) of Citizens Selecting This Response*
1. Fighting illegal drug use	70 (722)
2. Helping public schools	67 (682)
3. Addressing the problem of gangs	65 (669)
4. Protecting sensitive lands	57 (585)
5. Reducing I-4 congestion	57 (584)
6. Reducing discrimination	53 (543)
7. Youth improvement programs	53 (540)
8. Controlling government spending	49 (501)
9. Public safety	46 (468)
10. Increasing the level of wages	45 (463)
11. Condition of roadway system	44 (449)
12. Senior citizens' needs	43 (446)
13. Water quality of lakes	43 (437)
14. Controlling development and growth	40 (411)
15. Better job training	39 (402)
16. Welfare-to-work programs	38 (391)
17. Cutting property taxes	36 (363)
18. Code enforcement	34 (350)
19. Promoting high-tech jobs	32 (329)
20. Mass transit	31 (314)
21. Helping neighborhoods	30 (307)
22. Storm water drainage	27 (275)
23. Promoting the arts	23 (241)
24. Improving parks	22 (230)
25. Appearance of roadways	21 (214)
26. Addressing business needs	17 (168)
27. Building a light rail	15 (154)

Although not shown here, the top three issues differ in their ranking depending on whether we use just "Very Important" or combine this category with "Important." These differences (or, conversely, the failure to find convergent results from different methods) lead us to be conservative in our declarations. We should not say that the single most pressing issue for the citizens of Orange County is "Fighting illegal drug use" or "Helping public schools," because this ranking depends on a rather arbitrary choice of combining or not combining response categories. In such a circumstance, it would be more prudent for us to report that the top three issues facing respondents were fighting drugs, helping public schools, and addressing gang violence.

Unfortunately, you cannot move from this manually constructed table to a graph easily. Indeed, the information in Table 6.2 must be entered into SPSS or Excel before creating a graph that displays these results.

In Excel, the spreadsheet would look like the following:



	A	B	C	D
1	Issue	Percent Very Important	Number of respondents	
2	Government Spending	49	501	
3	Property Taxes	36	363	
4	I-4 Congestion	57	584	
5	Mass Transit	31	314	
6	Light Rail	15	154	
7	Roadway System	44	449	
8	Roadway Appearance	21	214	
9	Development & Growth	40	411	
10	Sensitive Ecologies	57	585	
11	Illegal Drug Use	70	722	
12	Gangs	65	669	
13	Code Enforcement	34	350	
14	Public Schools	67	682	
15	Youth Programs	53	540	
16	Job Training	39	402	
17	Public Safety	46	468	
18	Storm Drainage	27	275	
19	Lakes	43	437	
20	Welfare-to-Work	38	391	
21	High Tech Jobs	32	329	
22	Discrimination	53	543	
23	Arts	23	241	
24	Business Needs	17	168	
25	Senior Citizens	43	446	
26	Neighborhoods	30	307	
27	Parks	22	230	
28	Wages	45	463	
29				

To sort these from the most to the least important, we would have to

- highlight the contents of the file (A2:C28),
- click on Data/Sort,
- select "Descending" in the dialog box that appears, and
- click on OK.

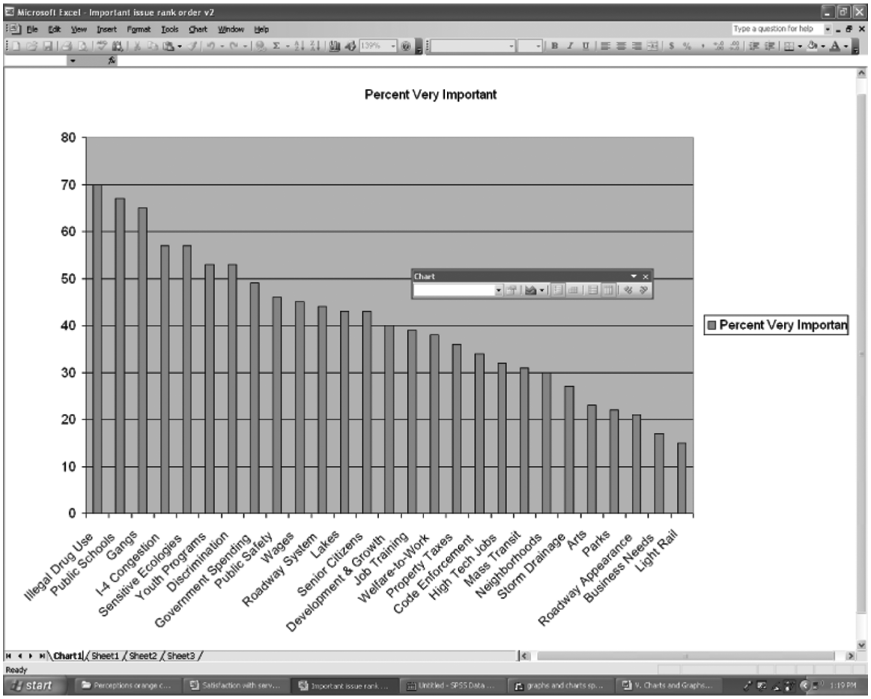
To add a number for the rank order of each issue,

- insert a column to the left of A,
- type the number 1 in cell A2 and the number 2 in cell A3,
- highlight these two cells, and
- grab the box in the lower right corner of these two highlighted cells and drag your cursor downward until you reach the number 27.

To quickly create a chart in Excel,

- highlight Columns B and C and
- click F11.

You should see the following chart appear in a chart sheet (named Chart 1):



Not bad for one keystroke. But you can improve upon this chart by following some of the basic principles of graphical display that we've discussed above.

You may alter elements of this table by right clicking on the element or by using the chart toolbar, which is displayed here. Let's go through a number of steps in Excel to clean up this quick and dirty chart in order to convey our results better to the Orange County mayor, who commissioned this study.

Step 1: Delete the legend by right clicking on Legend and selecting Clear. A legend is usually superfluous if you properly label the chart or table.

Step 2: Flip the chart so that the issue categories are on the left axis, which makes these labels easier to read by right clicking on any bar in the chart to select Chart Type.

Select the Standard Types tab. Click on the horizontal bar chart icon and then on OK to choose the highlighted chart subtype in the upper left corner.

Step 3: Reverse the order from the least important issue at the top of the chart to the most frequently cited issue at the top by right clicking on one of the issue labels and selecting Format Axis.

a. Select the Scale tab, and turn on the Categories in Reverse Order checkbox.
b. Change the number of categories between tick mark label to 1 (if it shows 2) so that we can see all the issue labels, not just every other one.
c. Click on OK.

Step 4: To eliminate the gray background (ink cartridges are expensive!), right click anywhere in that area, and select Clear.

Step 5: To show the actual percentage of respondents who chose a particular issue as "Very Important," right click on any blue bar, and select Format Data Series. Click in the Value box, and then click OK.

Step 6: To eliminate the vertical bars, which are unnecessary insofar as we show the actual percentages that each bar represents, right click on any vertical bar, and select Clear.

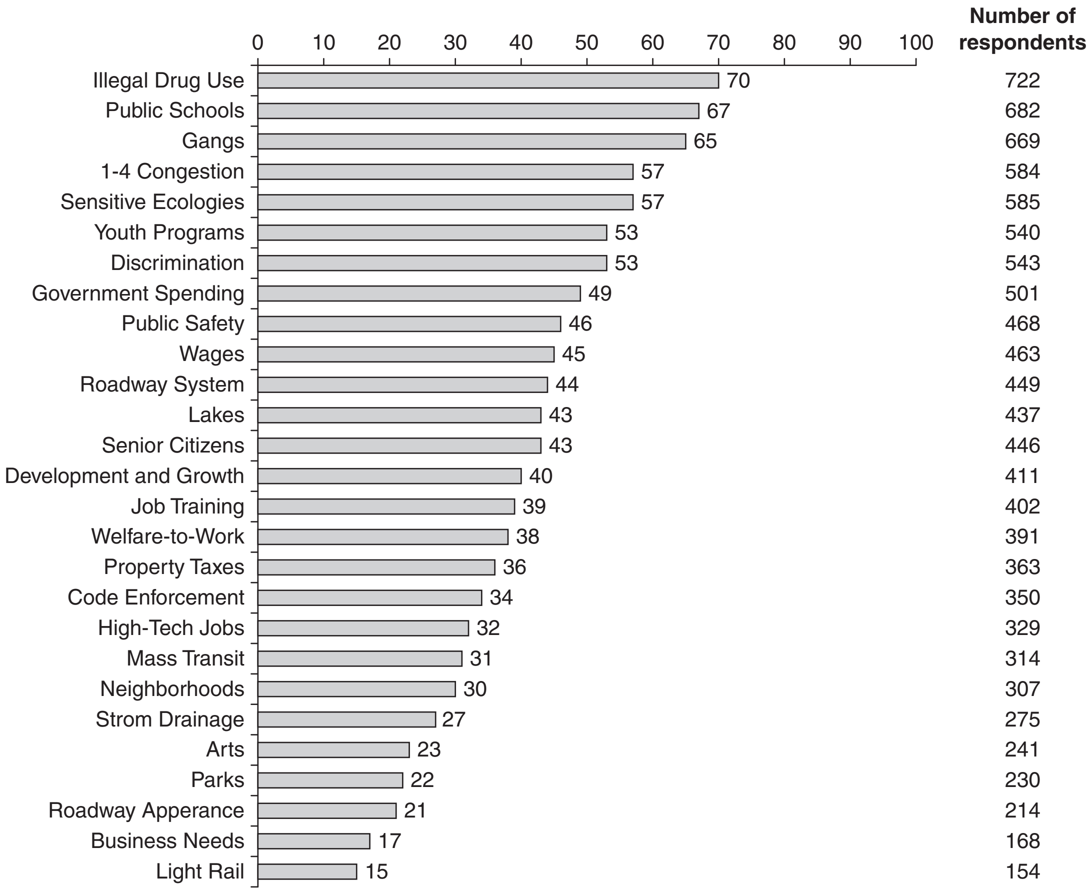
Step 7: To increase the scale of the chart from 80% to 100%, right click on any number on the *X*-axis (0, 10, 20, 30, etc.), and select Format Axis. Choose the Scale tab, and change Maximum to 100. Click on OK.

Step 8: To bring your chart into a Word document, click ALT and Print Screen (assuming the chart is staring you in the face).

Open your Word document to the place where you want to place your graph. Right click on Paste. On the Picture toolbar, click on the Text Wrapping icon (which looks like a dog on a background of horizontal gridlines), and select Behind Text, which permits you to move the chart anywhere in the document you like.

Step 9: To add the actual number of respondents who said "Very Important" to each item, return to the Excel spreadsheet, center the numbers in the column, highlight the column, right click, and copy and paste it into an empty Word document. Then click on Select Table/Delete Gridlines, right click and copy this table, move back to the Word document into which you've pasted your graph, click on Edit/Special Paste/Word Object, and move it to the right of your original chart.

Percentage (and Number) of Orange County Residents Who Reported That an Issue Is Very Important to Them



Boxplots and Outliers

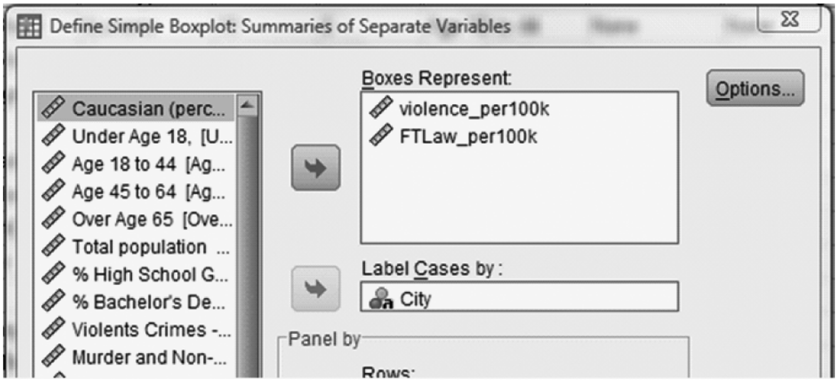
Boxplots provide a visual display of the distributions of variables. Boxplots serve purposes similar to that of a histogram, except that they can do so for multiple variables in the same graph. They also differ from histograms in their attention to displaying certain percentiles in the distribution and identifying specific observations that these charts consider as "outliers" and "extreme points."

Let's use a boxplot to display the distribution of two variables that we may be considering for a later analysis but whose distribution we want to examine at this point in order to see if we might consider any cities in the Community Indicators data set as outliers in terms of these two variables:

- Violent crimes per 100,000 residents
- Full-time law enforcement officers per 100,000 residents (a new variable constructed as we did the violent crimes per 100,000 residents variable)

Step 1: Run a boxplot for these variables by clicking on Graphs/Boxplots. From the dialog box that appears, select Simple and Summaries of Separate Variables, and then click on Define.

Move the two variables into the Boxes Represent box, and move the variable CITY into the Label Cases by box, which will identify by name the cities that may be considered **univariate outliers** on the distribution of these two variables. Click OK.

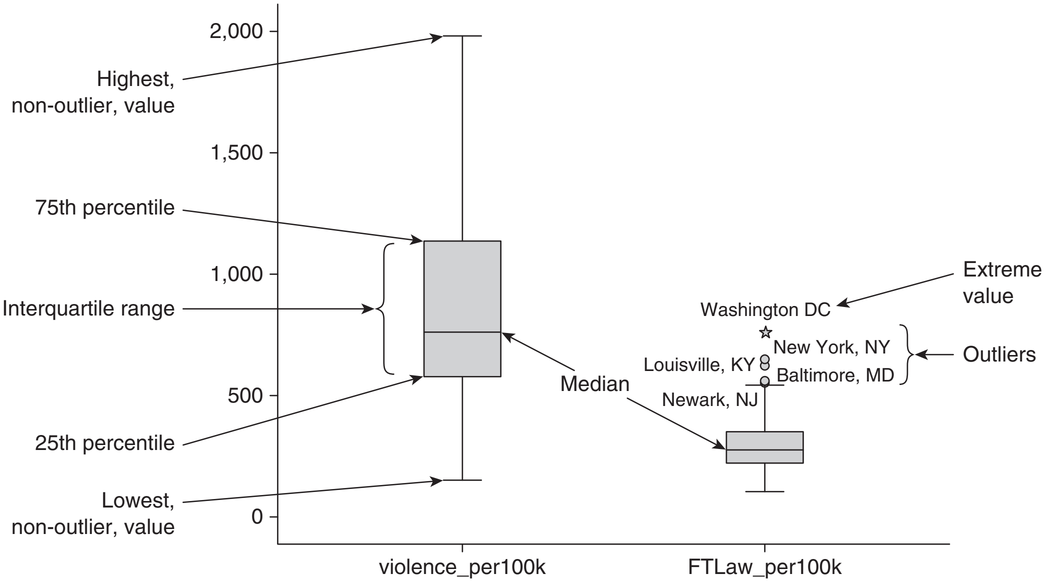


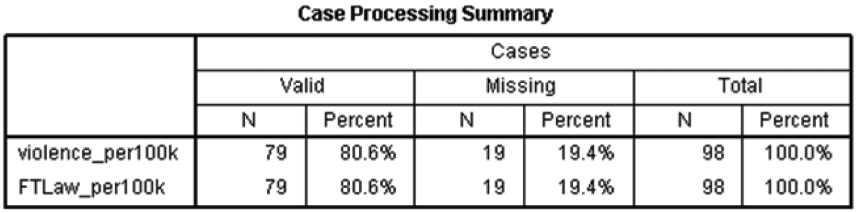
Define Simple Boxplot: Summaries of Separate Variables

Boxes Represent:
violence_per100k
FTLaw_per100k

Label Cases by:
City

The resulting output should look like the following (to which I have added descriptions of the boxplots' elements):





Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
violence_per100k	79	80.6%	19	19.4%	98	100.0%
FTLaw_per100k	79	80.6%	19	19.4%	98	100.0%

How do we interpret these boxplots?

- The dark horizontal line inside each of the boxes is the median value for each variable. Half of all observations fall above that line and half below it.

- The top of each box is the "third quartile," which marks the 75th percentile in the distribution. The bottom of each box is the second quartile and marks the 25th percentile.

- The difference between these two quartiles (i.e., the height of the boxes themselves) is called the interquartile range (IQR). The middle 50% percent of all cases falls within the IQR.

- The stems or "whiskers" that extend from the boxes represent the lowest and highest values that are not considered outliers by SPSS. Boxplots define an observation as being an outlier if it lies more than 1.5 times the IQR above or below the box itself. Any case identified by an open circle is an outlier (e.g., New York, Louisville, Baltimore, and Newark in terms of full-time law officers per 100,000 residents). Any case identified by an asterisk is considered an "extreme" case, falling more than three times the IQR above or below the box.

The case-processing summary indicates that only 79 cities provided valid or nonmissing values for both variables. The default processing rule is listwise deletion of missing values, thus reducing the number of observations on which these distributions are displayed. This can be changed. It's your choice whether you want to exclude about a fourth of all the observations in your data set because of this default rule for missing values. I wouldn't. We're throwing out too much information. Check out what's behind the Options button on the boxplot dialog box to change this default.

● CONCLUSION

Graphs can enlighten or mislead. You now have the tools and insights to produce enlightened graphs that build on an understanding of perception, cognition, and memory. You now know many of the signs of graphical tomfoolery and can formally express graphical distortions of what the data say through calculation of the Lie Factor Quotient. And you can produce histograms, scattergrams, bar charts, and boxplots and can edit them according to the principles of good (and bad) graphical design. Go forth, and with that knowledge, make fishers of men (and women).

To practice and reinforce the lessons of this chapter of the book, turn to Exercise 5 in the *Student Workbook* (available at http://www.sagepub.com/pearsonsp/).

● NOTE

1. If your version of Excel does not include an Analysis ToolPak, you can add it through the following steps:

Step 1: Click the **Microsoft Office Button**, the click **Excel Options.**

Step 2: Click **Add-Ins,** and then in the **Manage** box, **select Excel Add-ins.**

Step 3: Click **Go.**

Step 4: In the **Add-Ins available** box, select the **Analysis ToolPak** check box, and the click **OK.**

If **Analysis ToolPak** is not listed in the **Add-Ins available** box, click **Browse** to locate it.

If you get prompted that the Analysis ToolPak is not currently installed on your computer, click **Yes** to install it.

Step 5: After you load the Analysis ToolPak, the **Data Analysis** command is available in the **Analysis** group on the **Data** tab.