for collecting data.[8] Throughout this book, whenever we introduce data, we'll provide a margin note listing the W's (and H) of the data. Identifying the W's is a habit we recommend.

The first step of any data analysis is to know what you are trying to accomplish and what you want to know. To help you use Statistics to understand the world and make decisions, we'll lead you through the entire process of *thinking* about the problem, *showing* what you've found, and *telling* others what you've learned. Every guided example in this book is broken into these three steps: *Think*, *Show,* and *Tell.* Identifying the problem and the *who* and *what* of the data is a key part of the *Think* step of any analysis. Make sure you know these before you proceed to *Show* or *Tell* anything about the data.

More About Variables (*What*?)

The Amazon data table displays information about several variables: *Order Number*, *Name*, *State/Country*, *Price*, and so on. These identify *what* we know about each individual. Variables such as these can play different roles, depending on how we plan to use them. While some are merely identifiers, others may be categorical or quantitative. Making that distinction is an important step in our analysis.

Identifiers

Amazon wants to know who you are when you sign in again and it doesn't want to confuse you with some other customer. So it assigns you a unique identifying number.[9] Amazon also wants to send you the right product, so it assigns a unique Amazon Standard Identification Number (ASIN) to each item it carries. Both of these numbers are useful to Amazon, but they aren't measurements of anything. They're generated by Amazon and *assigned* uniquely to customers and products. Data like these values are called **identifier variables**. Other examples are student ID numbers and Social Security numbers. Identifier variables are typically used to identify single cases, not to look for patterns in collections of data, so they are seldom used in data analysis.

PRIVACY AND THE INTERNET

You have many identifiers: a Social Security number, a student ID number, possibly a passport number, a health insurance number, and probably a Facebook account name. Privacy experts are worried that Internet thieves may match your identity in these different areas of your life, allowing, for example, your health, education, and financial records to be merged. Even online companies such as Facebook and Google are able to link your online behavior to some of these identifiers, which carries with it both advantages and dangers. The National Strategy for Trusted Identities in Cyberspace (www.wired.com/images_blogs/threatlevel/2011/04/NSTICstrategy_041511.pdf) proposes ways that we may address this challenge in the near future.

Categorical Variables

Some variables just tell us what group or category each individual belongs to. Are you fluent in Spanish? Do you have any piercings? What's your favorite music genre? We call variables like these **categorical variables**.[10] Some variables are clearly categorical, like the variable *State/Country*. Its values are text and those values tell us what category the particular case falls into. Descriptive responses to questions are often categories. For example, the responses to the questions "Who is your cell phone provider?" or "What is your marital status?" yield categorical values. But numerals are sometimes used to label categories, so categorical data can also be numerals. ZIP codes, for example, are numerals but they convey geographic information, which is categorical.

Quantitative Variables

When a variable contains *measured* numerical values, we call it a **quantitative variable**. Quantitative variables typically have **units**, such as centimeters or years, but they may also be *counts* of something associated with each case, such as the number of siblings a person

[8]Coming attractions: to be discussed in Part III. We sense your excitement.

[9]Or sometimes a code containing numerals and letters.

[10]You may also see them called *qualitative* variables.

has, or how many languages they speak. Whenever you look at a quantitative variable, be sure you know what its units are. Without units, the values of a measured variable have no meaning. It does little good to be promised a raise of 5000 a year if you don't know whether it will be paid in euros, dollars, pennies, yen, or bronze knuts.[11]

Either/Or?

Some variables with numeric values can be treated as either categorical or quantitative depending on what we want to know. Amazon could record your *Age* in years. That seems quantitative, and it would be if the company wanted to know the average age of those customers who visit their site after 3 A.M. But suppose Amazon wants to decide which album to feature on its site when you visit. Then thinking of your age in one of the categories Child, Teen, Adult, or Senior might be more useful. So, sometimes whether a variable is treated as categorical or quantitative is more about the question we want to ask than an intrinsic property of the variable itself.

Ordinal Variables

Suppose a course evaluation survey asks, "How valuable do you think this course will be to you?" 1 = Worthless; 2 = Slightly; 3 = Somewhat; 4 = Reasonably; 5 = Invaluable. Is *Educational Value* categorical or quantitative? A teacher might just count the number of students who gave each response for her course, treating *Educational Value* as a categorical variable. Or if she wants to see whether the course is improving, she might treat the responses as the *amount* of perceived value—in effect, treating the variable as quantitative.

But what are the units? There is certainly an *order* of perceived worth: Higher numbers indicate higher perceived worth. A course that averages 4.2 seems more valuable than one that averages 2.1. But is it *twice* as valuable? Does that even mean anything? Variables like this that have a natural order but no units are often called *ordinal variables*. Other examples are college class (freshman, sophomore, junior, or senior) and hurricane level (1, 2, 3, 4, or 5). Ordinal variables can be a little tricky to analyze and for the most part they are not considered in this text.

FOR EXAMPLE Identifying the *What* and *Why* of Tablets

RECAP: A *Consumer Reports* article about 126 tablet computers lists each tablet's manufacturer, cost, battery life (hours), operating system (Android, IOS, or Windows) and overall performance score (0–100).

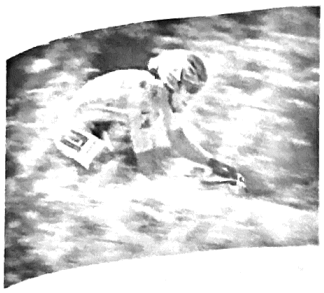
QUESTION: Are these variables categorical or quantitative? Include units where appropriate, and describe the *Why* of this investigation.

ANSWER: The variables are:

- manufacturer (categorical)
- cost (quantitative, $)
- battery life (quantitative, hours)
- operating system (categorical)
- performance score (quantitative, with no units—essentially an ordinal variable)

Why? The magazine hopes to provide consumers with information to help them choose a tablet to meet their needs.

[11] Wizarding money. Just seeing who's paying attention.



JUST CHECKING

In the 2004 Tour de France, Lance Armstrong made history by winning the race for an unprecedented sixth time. In 2005, he became the only 7-time winner and set a new record for the fastest average speed—41.65 kilometers per hour (about 26 mph)—that stands to this day. Then in 2012, Armstrong was banned for life for doping offenses, stripped of all his titles and his records expunged. Here are the first three and last nine lines of a data table of all Tour de France races. Keep in mind that the dataset has over 100 entries.

1. List as many of the W's as you can for this dataset.
2. Classify each variable as categorical or quantitative; if quantitative, identify the units.

Year	Winner	Country of Origin	Age	Team	Total Time (h/min/s)	Avg. Speed (km/h)	Stages	Total Distance Ridden (km)	Starting Riders	Finishing Riders
1903	Maurice Garin	France	32	La Française	94.33.00	25.7	6	2428	60	21
1904	Henri Cornet	France	20	Cycles JC	96.05.00	25.3	6	2428	88	23
1905	Louis Trousseller	France	24	Peugeot	112.18.09	27.1	11	2994	60	24
...										
2012	Bradley Wiggins	Great Britain	32	Sky	87.34.47	39.83	20	3488	198	153
2013	Christopher Froome	Great Britain	28	Sky	83.56.40	40.55	21	3404	198	169
2014	Vincenzo Nibali	Italy	29	Astana	89.56.06	40.74	21	3663.5	198	164
2015	Christopher Froome	Great Britain	30	Sky	84.46.14	39.64	21	3660.3	198	160
2016	Christopher Froome	Great Britain	31	Sky	89.04.48	39.62	21	3529	198	174
2017	Christopher Froome	Great Britain	32	Sky	86.34	40.997	21	3540	198	167
2018	Geraint Thomas	Great Britain	32	Sky	83.28	40.210	21	3349	176	145
2019	Egan Bernal	Colombia	22	INEOS	82.57.00	40.576	21	3365.8	176	155
2020	Tadej Pogacar	Slovenia	21	UAE Team Emirates	87.20.05	39.872	21	3482.2	176	146

THERE'S A WORLD OF DATA ON THE INTERNET

These days, one of the richest sources of data is the Internet. With a bit of practice, you can learn to find data on almost any subject. Many of the datasets we use in this book were found in this way. The Internet has both advantages and disadvantages as a source of data. Among the advantages is the fact that often you'll be able to find even more current data than those we present. The disadvantages include the fact that references to Internet addresses can "break" as sites evolve, move, and die.

Our solution to these challenges is to offer the best advice we can to help you search for the data, wherever they may be residing. We usually point you to a website. We'll sometimes suggest search terms and offer other guidance.

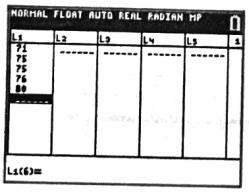
Some words of caution, though: Data found on Internet sites may not be formatted in the best way for use in statistics software. Although you may see a data table in standard form, an attempt to copy the data may leave you with a single column of values. You may have to work in your favorite statistics or spreadsheet program to reformat the data into variables. You will also probably want to remove commas from large numbers and extra symbols such as money indicators ($, ¥, £); few statistics packages can handle these.

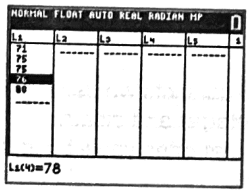
WHAT CAN GO WRONG?

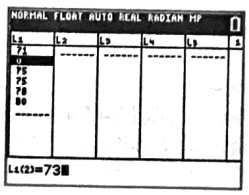
- **Don't label a variable as categorical or quantitative without thinking about the question you want it to answer.** The same variable can sometimes take on different roles.
- **Just because your variable's values are numbers, don't assume that it's quantitative.** Categories are often given numerical labels. Don't let that fool you into thinking they have quantitative meaning. Look at the context.
- **Always be skeptical.** One reason to analyze data is to discover the truth. Even when you are told a context for the data, it may turn out that the truth is a bit (or even a lot) different. The context colors our interpretation of the data, so those who want to influence what you think may slant the context. A survey that seems to be about all students may in fact report just the opinions of those who visited a fan website. The question that respondents answered may be posed in a way that influences responses.

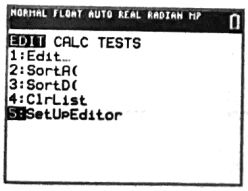
TI TIPS

Working with Data









You'll need to be able to enter and edit data in your calculator. Here's how:

TO ENTER DATA: Hit the STAT button, and choose EDIT from the menu. You'll see a set of columns labeled L1, L2, and so on. Here is where you can enter, change, or delete a set of data.

Let's enter the heights (in inches) of the five starting players on a basketball team: 71, 75, 75, 76, and 80. Move the cursor to the space under L1, type in 71, and hit ENTER (or the down arrow). There's the first player. Now enter the data for the rest of the team.

TO CHANGE A DATUM: Suppose the 76″ player grew since last season; his height should be listed as 78″. Use the arrow keys to move the cursor onto the 76, then change the value and ENTER the correction.

TO ADD MORE DATA: We want to include the sixth man, 73″ tall. It would be easy to simply add this new datum to the end of the list. However, sometimes the order of the data matters, so let's place this datum in numerical order. Move the cursor to the desired position (atop the first 75). Hit 2ND INS (for "insert"), then ENTER the 73 in the new space.

TO DELETE A DATUM: The 78″ player just quit the team. Move the cursor there. Hit DEL. Bye.

TO CLEAR THE DATALIST: Finished playing basketball? Move the cursor atop the L1. Hit CLEAR, then ENTER (or down arrow). You should now have a blank datalist, ready for you to enter your next set of values.

LOST A DATALIST? Oops! Is L1 now missing entirely? Did you delete L1 by mistake, instead of just *clearing* it? Easy problem to fix: buy a new calculator. No? OK, then simply go to the STAT EDIT menu, and run SetUpEditor to return lists L1 through L6 to the STAT EDIT screen.