CHAPTER

1

# Data and Decisions

## H&M

Even if you haven't bought something from H&M recently, chances are good that you've passed by one of their stores. With over 4000 stores in 64 markets worldwide, they are one of the largest and fastest-growing clothing retailers in the world. Over the past decade, H&M has built new stores at an astounding rate of over 10% a year. Thanks to this growth, the CEO, Karl-Johan Persson, grandson of the founder, is now the richest person in Sweden.

Like most companies, H&M's online presence has been increasing as well. Of their 64 worldwide markets, 35 offer e-commerce where customers can shop 24 hours a day, 7 days a week, with just the click of a mouse. H&M now reaches their customers in ways no one could even imagine just a generation ago. But what of the future? Will the company be better off continuing to grow brick and mortar stores at the same pace, or should they devote more resources into the digital space?[1]

[1] We developed this hypothetical example in late 2017 based on our business and consulting experience. As we were going to press, the news caught up with us. It turns out that indeed H&M had been struggling with their balance of online sales vs. brick and mortar inventory. Perhaps if this book had been published a year earlier, they could have solved the problem: www.nytimes.com/2018/03/27/business/hm-clothes-stock-sales.html

A few generations ago, many store owners knew their customers and their business well. With that knowledge, they could forecast growth, see trends, and even personalize their suggestions to customers, guessing which items that particular customers might like. Businesses today rely on similar information to make decisions, but most never meet their customers. With 4000 different stores and thousands of online customers, H&M has to obtain and analyze their data in other ways.

The key to turning data into information and knowledge is Statistics—the collection of tools that extract information from data. These tools that you will learn also provide the foundation for more advanced methods like data mining and analytics. According to CEO Karl-Johan Persson, "advanced analytics provide an important support for our operations. The algorithms we have started to use will contribute to improvements within everything from assortment planning and logistics to sales."[2] Using statistical methods to turn data into information, information into knowledge, and knowledge into smart business decisions is the key to all successful modern business enterprises.

And it all starts . . . with data.

Thomasine has just landed her first job out of school as a marketing and strategy analyst working for H&M. Her team's first assignment is to decide whether to build more brick and mortar stores or invest more in online operations. To help make the decision, they investigate store sales data over the past ten years and display them in the following graph:

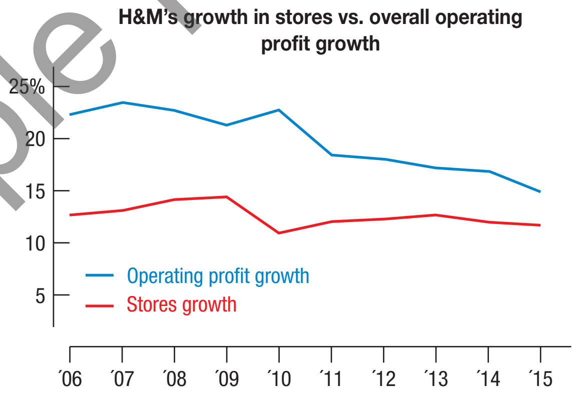


**FIGURE 1.1** H&M's store growth has remained steady at just over 10% a year, but operating profit growth seems to be coming down.

Thomasine wonders if the decline she sees in the stores' profit growth (the blue line in Figure 1.1) means she should recommend putting more resources into online sales instead of just building more stores.

Displays like this, called *data visualizations*, can summarize large amounts of data in a concise way that helps make good business decisions, and can often reveal things that weren't expected.

## IN PRACTICE 1.1 Business insights from visualizations

One of the authors was consulting for a large multinational firm and was given access to their sales data. Management wondered if there might be sales opportunities around the world and where they might be. Because the company sold many consumer items to individuals, the consultant decided that rather than focus on the total sales

[2] 2016 H&M Group annual report, about.hm.com/en/media/news/financial-reports/2017/1/2441626.html

(in dollars) in each country he should divide the total sales in each country by the population size, creating the new variable *Sales per Capita*. When he displayed this variable on a map, management was shocked:[3]

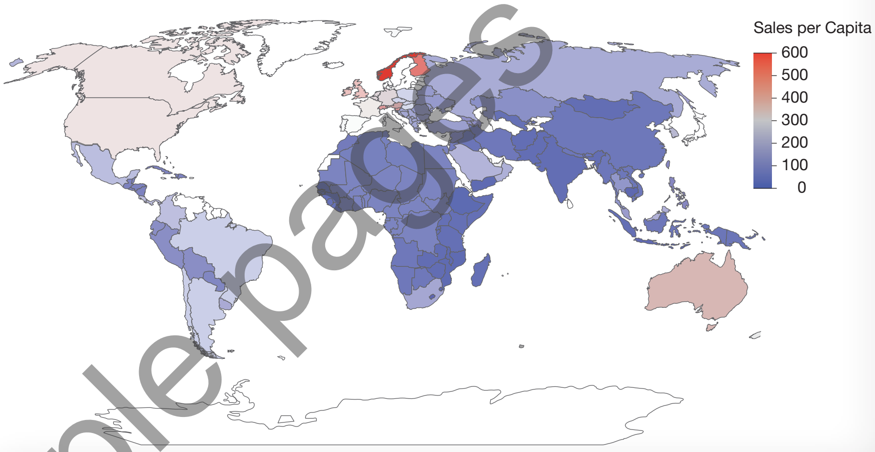
**MANAGER** We know that we sell more in the United States than anywhere else in the world, but why are some countries redder than the U.S.?

**CONSULTANT** In this color scheme, low *Sales per Capita* ($ spent per customer) is indicated by dark blue, average by white (grey) and higher than average by red. The countries in the brightest red are the ones with the highest sales per person.

**MANAGER** You mean we sell more per person in Norway, Finland, and even Australia than in the U.S.?

**CONSULTANT** Exactly. Norway has the highest sales at more than $600 per person, compared with the U.S. at $364.

**MANAGER** Wow! I had no idea. I never would have guessed that. Thank you for the insight!



We will be using visualizations, summaries, and models of data to understand, explain, and predict throughout the course. Along the way we will encounter many types of data and corresponding ways to visualize, model, and analyze the data we collect. And because it all starts with data, we'll spend the rest of this chapter getting to know more about the nature of data.

## 1.1 Data

Every time you make an online purchase, more information is captured than just the details of the purchase itself. What pages did you search to get to your purchase? How much time did you spend looking at each? These recorded values, whether numbers or labels, together with their context are called **data**. They are recorded and stored electronically, in vast digital repositories called **data warehouses**. Businesses have always relied on data to make good decisions, but today, more than ever before, companies use data to make decisions about virtually all aspects of their business, from inventory to advertising to website design.

[3] This is based on a true story. We can't reveal the name of the company due to a non-disclosure agreement.

Every swipe of your credit card and every click of your mouse has helped these data warehouses grow. The challenges of collecting, managing, storing, and curating all of this information collectively fall under the term **Big Data**.

But data alone can't make good decisions. To start the process of turning data into useful information, you first need to know what decisions you want to make. Without a question, you have no idea what might be interesting about the data. Should you look at the time of transactions, their location, their price, which products were bought, or something else? Your knowledge of the business issues and the questions you want to answer will help guide your search for insights from the data, and help you harness data to make better decisions.

Once you have data and a clear vision of the problem, the statistics techniques in this book can empower your decision making. They will help you in two ways: You'll learn how to estimate the likely values needed for your decisions and—possibly more important—you'll learn how to quantify the *uncertainty* of those estimates.

**Why are you taking this course?**

The typical answer is "because it's required." But why is it required? Because these are the tools that will help you leverage your business domain knowledge with data.

Albert Einstein is credited with saying "If I had one hour to save the world, I'd spend 55 minutes defining the problem and 5 minutes solving it."[4] The wisdom of using your business acumen to define your question will be clear throughout this book.

Before H&M introduces a new product they usually test market it to a small sample of customers and collect data on the product's performance before committing to it worldwide. Statistics helps them make the leap from a sample to an understanding of the world at large. We hope this text will empower you to draw conclusions from data and make valid business decisions in response to such questions as:

- Will the new design of our website increase click-through rates and result in more sales?
- What is the effect of advertising on sales?
- Do aggressive, "high-growth" mutual funds really have higher returns than more conservative funds?
- Is there a seasonal cycle in your firm's profits?
- What is the relationship between shelf location and cereal sales?
- Do students around the world perceive issues in business ethics differently?
- Are there common characteristics about your customers and why they choose your products?—and, more importantly, are those characteristics the same among those who aren't your customers?

Your ability to answer questions such as these and make sound business decisions with data depends largely on your ability to take a business problem, *translate* it into a question that data can answer, and *communicate* that answer to others. The steps to follow are shown in the box in the margin. The **Plan**, **Do**, and **Report** strategy is found throughout the book. The main headings will stay the same although the specific subparts will vary slightly depending on the topic we're learning.

**Plan (1–2)**

1. **Define** the problem.
2. **Collect** and/or find data and **identify** the variables.

**Do (3–6)**

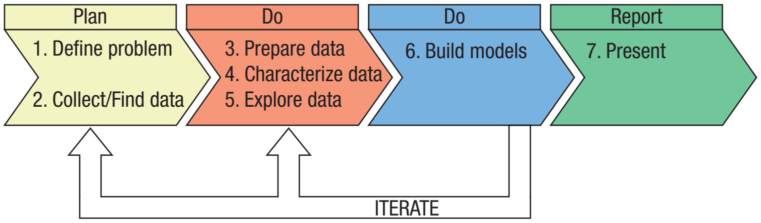
3. **Prepare** and wrangle data.
4. **Characterize** the data.
5. **Explore** the data.
   Summarize
   Visualize
6. **Model** (if appropriate).
   Check conditions and assumptions for modeling.
   Fit the model and make the necessary calculations.

**Report (7)**

7. Communicate and **present**.

Rarely does the journey from problem definition to solution proceed straight from Step 1 to Step 7. As you learn more about your data you'll probably want to rethink earlier steps, possibly even modifying the original question itself. Or you may decide to collect different data after you see the limitations of your current model. But bearing this process in mind will help you to strategize your data analytics process and keep you on the road toward the goal of delivering good decisions.

[4] According to quoteinvestigator.com there is "no substantive evidence that Einstein ever made a remark of this type." It appeared in a paper by William H. Markle, who credited an unnamed Yale professor. But many people, including those at goodreads.com, still give the credit to Einstein.



## 1.2 The Role of Data in Decision Making

**Q:** What is analytics?

**A:** Analytics is the term for extracting information from data.

**Q:** Is there really a difference between statistics and analytics?

**A:** Essentially no. We'll use the terms interchangeably. Some use the term "advanced analytics" to include modern machine learning methods not traditionally found in statistics. (See Chapter 21)

When companies try to obtain actionable information from data that may have been collected in the course of doing business (such as records of transactions or a customer database) it is usually called **data mining**. Sometimes the analysis is called **predictive analytics** if it focuses on future performance. The more general term, **business analytics** (or sometimes simply analytics), refers to any use of data and statistical analysis to inform business decisions. Leading companies are embracing analytics to extract value from their data. As Clive Humby, author of *The Loyalty Myth*, said, "data is the new oil." For example, Zillow recently offered $1,200,000 to improve the prediction of home sale prices from publicly available data. As of June 2017, 791 teams were competing for the prize.

Companies use data to make decisions about nearly every aspect of their business. By studying the past behavior of customers and predicting their responses, they hope to better serve their customers and to compete more effectively.

eBay collected data and used analytics to examine its own use of computer resources. Although not obvious to its own technical people, once they crunched the data they found huge inefficiencies. According to Forbes, eBay was able to "save millions in capital expenditures within the first year."

THE W'S:
WHO
WHAT
WHEN
WHERE
WHY

Data come in many forms. Some are numerical (consisting only of numbers), others are alphabetic (consisting only of letters), and yet others are alphanumerical (mixed numbers and letters). But data are useless unless we know their **context**. Newspaper journalists know that the lead paragraph of a good story should establish the "Five W's": *who*, *what*, *when*, *where*, and (if possible) *why*. Often, we add *how* to the list as well. Answering these questions connects the data to the business problem at hand. The answers to the first two questions are essential. If we don't know *who* and *what*, we don't have any useful information.

We can make the meaning clear if we add the context of *who* the data are about and *what* was measured and organize the values into a **data table**. Table 1.1 shows part of a data table of purchase records from an online music retailer. Each row represents a purchase of a music album. The most general term for a row of a data table is **case** or **record**. Each column of the table records some characteristic

| Order Number | Name | State/Country | Price | Area Code | Album Download | Gift? | Stock ID | Artist |
|---|---|---|---|---|---|---|---|---|
| 105-2686834-3759466 | Katherine H. | Ohio | 5.99 | 440 | Identity | N | B00000I5Y6 | James Fortune & Flya |
| 105-9318443-4200264 | Samuel P. | Illinois | 9.99 | 312 | Port of Morrow | Y | B000002BK9 | The Shins |
| 105-1872500-0198646 | Chris G. | Massachusetts | 9.99 | 413 | Up All Night | N | B000068ZVQ | Syco Music UK |
| 103-2628345-9238664 | Monique D. | Canada | 10.99 | 902 | Fallen Empires | N | B000001OAA | Snow Patrol |
| 002-1663369-6638649 | Katherine H. | Ohio | 11.99 | 440 | Sees the Light | N | B002MXA7Q0 | La Sera |

**TABLE 1.1** Example of a data table. The variable names are in the top row. Typically, the *Who* of the table are found in the leftmost column.

of the cases. The columns are called **variables**. You'll usually find the name of the variable at the top of the column as in Table 1.1.

We call cases by different names, depending on the situation. Individuals who answer a survey are referred to as **respondents**. People on whom we experiment are **subjects** or (in an attempt to acknowledge the importance of their role in the experiment) **participants**, but animals, plants, websites, and other inanimate subjects are often called **experimental units**. Often we call cases just what they are: for example, *customers*, *economic quarters*, or *companies*. When referring to a transaction, rows are often called *records*. In Table 1.1, the rows are the individual orders, or purchase records. A common place to find the *who* of the table is the leftmost column. It's often an identifying variable for the cases, in this example, the order number.

## JUST CHECKING

**1** What is the "*who*" of Table 1.1? That is, does each row refer to a) a person or b) an order? How can you tell?

### Metadata

Metadata became a common term when the National Security Agency (NSA) claimed that they weren't collecting Americans' phone calls but only the information about the phone calls, the phone numbers of the caller and recipient, the time and duration of the call and any bank information used to make the call—in other words—the metadata.

If you collect the data yourself, you'll know what the cases are and how the variables are defined. But, often, you'll be looking at data that someone else collected. The information about the data, called the metadata, might have to come from the company's database administrator or from the information technology department of a company. **Metadata** typically contains information about *how*, *when*, and *where* (and possibly *why*) the data were collected; *who* each case represents; and the definitions of all the variables.

A general term for a data table like the one shown in Table 1.1 is a **spreadsheet**, a name that comes from bookkeeping ledgers of financial information. The data were typically spread across facing pages of a bound ledger, the book used by an accountant for keeping records of expenditures and sources of income. For the accountant, the columns were the types of expenses and income, and the rows were transactions, typically invoices or receipts. These days, it is common to keep modest-size datasets in a spreadsheet even if no accounting is involved. It is usually easy to move a data table from a spreadsheet program to a program designed for statistical graphics and analysis, either directly or by copying the data table and pasting it into the statistics program.

Although data tables and spreadsheets are great for relatively small data sets, they are cumbersome for the complex data sets that companies must maintain on a day-to-day basis. Try to imagine a spreadsheet from a company the size of Amazon with customers in the rows and products in the columns. Amazon has hundreds of millions of customers and millions of products. But very few customers have purchased more than a few dozen items, so almost all the entries in the spreadsheet would be blank—not a very efficient way to store information. For that reason, various other database architectures are used to store data. The most common is a relational database.

In a **relational database**, two or more separate data tables are linked together so that information can be merged across them. Each data table is a *relation* because it is about a specific set of cases with information about each of these cases for all (or at least most) of the variables ("fields" in database terminology). For example, a table of H&M customers, along with demographic information on each, is such a relation. A data table of all the items sold by the company, including information on price, inventory, and past history, is another relation. Transactions may be held in a third "relation" that references each of the other two relations. Table 1.2 shows a small example.

In statistics, analyses are typically performed on a single relation because all variables must refer to the same cases. But often the data must be retrieved from a relational database. Retrieving data from these databases may require specific expertise with that software. In the rest of the book, we'll assume that the data have been retrieved and placed in a data table or spreadsheet with variables listed as columns and cases as the rows.

**Customers**

| Customer Number | Name | City | State | ZIP Code | Customer since | Gold Member? |
|---|---|---|---|---|---|---|
| 473859 | R. De Veaux | Williamstown | MA | 01267 | 2007 | No |
| 127389 | N. Sharpe | New York City | NY | 10021 | 2000 | Yes |
| 335682 | P. Velleman | Ithaca | NY | 14580 | 2003 | No |
| . . . | | | | | | |

**Items**

| Product ID | Name | Price | Currently in Stock? |
|---|---|---|---|
| 42-8719 | Resort Shirt | 24.99 | Yes |
| 73-2671 | Lace Dress | 69.99 | No |
| 35-0518 | Cashmere Sweater | 129.00 | Yes |
| 72-9665 | Leather Derby Shoes | 69.00 | Yes |

**Transactions**

| Transaction Number | Date | Customer Number | Product ID | Quantity | Shipping Method | Free Ship? |
|---|---|---|---|---|---|---|
| T23478923 | 9/15/17 | 473859 | 42-8719 | 1 | UPS 2nd Day | N |
| T23478924 | 9/15/17 | 473859 | 35-0518 | 1 | UPS 2nd Day | N |
| T63928934 | 10/20/17 | 335682 | 73-2671 | 3 | UPS Ground | N |
| T72348299 | 12/22/17 | 127389 | 72-9665 | 1 | Fed Ex Ovnt | Y |

**TABLE 1.2** A relational database shows all the relevant information for three separate relations linked together by customer and product numbers.

## IN PRACTICE 1.2 Gaining insight from data by identifying variables and the W's

Carly is an analyst at a credit card issuer. Her manager wants to know if an offer mailed 3 months ago has affected customers' use of their cards. To answer, Carly asks the IT department to assemble some data on recent customer spending. The IT department sends her a spreadsheet. The first six rows look like this:

| Account ID | Pre Spending | Spending | Age | Segment | Enroll? | Offer | Segment Spend |
|---|---|---|---|---|---|---|---|
| 393371 | $2,698.12 | $6,261.40 | 25–34 | Travel/Ent | NO | None | $887.36 |
| 462715 | $2,707.92 | $3,397.22 | 45–54 | Retail | NO | Gift Card | $5,062.55 |
| 433469 | $800.51 | $4,196.77 | 65+ | Retail | NO | None | $673.80 |
| 462716 | $3,459.52 | $3,335.00 | 25–34 | Services | YES | Double Miles | $800.75 |
| 420605 | $2,106.48 | $5,576.83 | 35–44 | Leisure | YES | Double Miles | $3,064.81 |
| 473703 | $2,603.92 | $7,397.50 | <25 | Travel/Ent | YES | Double Miles | $491.29 |

(*continued*)

**MANAGER** Thanks for the information. I'm not quite sure about the structure. Can you tell me what each row represents and what was measured on each?

**ANALYST (CARLY)** The cases are individual customers. The data are from our internal records for the past 6 months (3 months before and 3 months after an offer was sent to the customers). The variables include the account ID of the customer (*Account ID*), and the amounts charged on the card before (*Pre Spending*) and after (*Post Spending*) the offer was sent out. We also have the customer's *Age*, marketing *Segment*, whether they enrolled on the website (*Enroll?*), what offer they were sent (*Offer*), and how much they charged on the card in their marketing segment (*Segment Spend*). (The marketing *Segment* classifies cardholders based on their spending patterns.)

## 1.3 Variable Types

When the values of a variable are simply the names of categories we call it a **categorical**, or **qualitative**, **variable**. When the values of a variable are measured numerical quantities, we call it a **quantitative variable**.

Descriptive responses to questions are often categories. For example, the responses to the questions "What type of mutual fund do you invest in?" or "What kind of advertising does your firm use?" yield categorical values. An important special case of categorical variables is one that has only two possible responses (usually "yes" or "no"), which arise naturally from questions like "Do you invest in the stock market?" or "Do you make online purchases from this website?"

| Question | Categories or Responses |
|---|---|
| Do you invest in the stock market? | __ Yes __ No |
| What kind of advertising do you use? | __ Newspapers __ Internet __ Direct mailings |
| What is your class at school? | __ Freshman __ Sophomore __ Junior __ Senior |
| I would recommend this course to another student. | __ Strongly Disagree __ Slightly Disagree __ Slightly Agree __ Strongly Agree |
| How satisfied are you with this product? | __ Very Unsatisfied __ Unsatisfied __ Satisfied __ Very Satisfied |

TABLE 1.3 Some examples of categorical variables.

**Categorical or Quantitative?**

Dates can be confusing. Depending on how a date is used, it may be categorical or quantitative. For example, *Day of the Week* has no units, and is categorical. What about a date such as October 31, 2017 (which is a string of characters)? Most software will treat this as categorical. However, many statistics programs can add and subtract dates to determine that there are 60 days between 10/30/17 and 12/30/17, and that 11/30/17 falls exactly in the middle of this date range. Most programs can convert any date into the number of seconds, minutes or hours past a given starting date. If this is the case, then dates may be treated as a quantitative variable, but be sure to specify the units.

Many measurements are quantitative. In a purchase record, price, quantity, and time spent on the website are all quantitative values with **units** (dollars, count, and seconds). For quantitative variables, the units tell how each value has been measured. Even more important, units such as yen, cubits, carats, angstroms, nanoseconds, miles per hour, or degrees Celsius tell us the *scale* of measurement, so we know how far apart two values are. Without units, the values of a measured variable have no clear 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, yen, or Estonian krooni. An essential part of a quantitative variable is its units. Some quantitative variables, however, don't have obvious units. The Dow Jones Industrial "Average" has units (points?) but no one talks about them. Percentages are ratios of two quantities and so the units "cancel out," but they are still percentages of something. So, although it isn't imperative that a quantitative variable have explicit units, when they are not explicit, be careful to think about whether adding their values, averaging them, or otherwise treating them as numerical, makes sense.

The distinction between categorical and quantitative variables seems clear, but there are reasons to be careful. First, some variables can be considered as either categorical or quantitative, depending on the kind of questions we ask about them. For example, the variable *Age* would be considered quantitative if the responses were numerical and they had units. A doctor would certainly consider *Age* to be quantitative. The units could be years, or for infants, the doctor would want even more precise units, like months, or even days. On the other hand, a retailer might lump together the values into categories like “Child (12 years or less),” “Teen (13 to 19),” “Adult (20 to 64),” or “Senior (65 or over).” For many purposes, like knowing which song download coupon to send you, that might be all the information needed. Then *Age* would be a categorical variable.

How to classify some variables as categorical or quantitative may seem obvious. But be careful. Area codes may look quantitative, but are really categories. What about ZIP codes? They are categories too, but the numbers do contain information. If you look at a map of the United States with ZIP codes, you’ll see that as you move West, the first digit of ZIP codes increases, so treating them as quantitative might make sense for some questions. Area codes present a similar set of issues; see the sidebar.

Another reason to be careful about classifying variables comes from the analysis of Big Data. When analysts want to decide what advertisement to send to the web page you’re looking at, or what the probability is that you’ll renew your phone contract, they use automatic methods involving dozens or even hundreds of variables. Usually the software used to do the analysis has to guess the type of variable from its values. When the variable contains symbols other than numbers, the software will correctly type the variable as categorical, but just because a variable has numbers doesn’t mean it is quantitative. We’ve seen examples (area code, order number) where that’s just not the case. Data miners spend much of their time going back through data sets to correctly retype variables as categorical or quantitative to avoid silly mistakes of misuse.

## Identifiers

A special kind of categorical variable is worth mentioning. **Identifier variables** are categorical variables whose only purpose is to assign a unique identifier code to each individual in the data set. Your student ID number, social security number, and phone number are all identifiers.

Identifier variables are crucial in this era of Big Data because, by uniquely identifying the cases, they make it possible to combine data from different sources and provide unique labels. Your school’s grade transcripts and your bursar bill records are kept separately, but both refer to you. Your student ID is what links them. Most companies keep such relational databases. The identifier is crucial to linking one data table to another in a relational database. The identifiers in Table 1.2 are the *Customer Number*, *Product ID*, and *Transaction Number*. Variables like *UPS Tracking Number* and *Social Security Number* are other examples of identifiers.

## Other Data Types

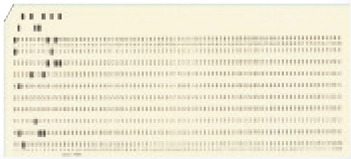
AirBnB, like many travel sites, uses stars to rate their listings. But are star ratings categorical or quantitative? There is certainly an *order* of perceived worth; more stars indicate higher perceived worth. An AirBnB property whose customer responses average around 4 stars is better than one whose average is around 2, but is it *twice* as good? These values are not quantitative, so we can’t really answer that question. When the values of a categorical variable have an intrinsic order, we can say that the variable is **ordinal**. By contrast, a categorical variable with unordered categories is sometimes called **nominal**. Values can be individually ordered (e.g., the ranks of employees based on the number of days they’ve worked for the company) or ordered in classes (e.g., Freshman, Sophomore, Junior, Senior). Ordering is not absolute; how the values are ordered depends on the purpose of the ordering. For example, are the categories Infant, Youth, Teen, Adult, and Senior ordinal?

### Area Codes?

When area codes were first introduced all phones had dials. To reduce wear and tear on the dials and to speed calls, the lowest-digit codes (the fastest to dial—those for which the dial spun the least) were assigned to the largest cities. So, New York City was given 212, Chicago 312, LA 213, and Philadelphia 215, but rural upstate New York was 607, Joliet was 815, and San Diego 619. Back then, the numerical value of an area code could be used to guess something about the population of its region. But after dials gave way to push buttons, new area codes were assigned without regard to population and area codes are now just categories, with no quantitative information.

### Variable Names That Make Sense

A tradition still hangs on in some places to name variables with cryptic abbreviations in uppercase letters. This can be traced back to the 1960s, when computer programs were controlled with instructions punched on cards. The earliest punch card equipment used only uppercase letters, and statistics programs limited variable names to six or eight characters, so variables had names like PRSRF3. Modern programs don’t have such restrictive limits, so there is no reason not to use names that make sense.



Well, if we are ordering on age, they surely are and how to order the categories is clear. But if we are ordering on purchase volume, it is likely that either Teen or Adult will be the top group.[5]

## Cross-Sectional and Time Series Data

The quantitative variable *Total Revenue* in Table 1.4 is an example of a time series. A **time series** is an ordered sequence of values of a single quantitative variable measured at regular intervals over time. Time series are common in business. Typical measuring points are months, quarters, or years, but virtually any consistently spaced time interval is possible. Variables collected over time hold special challenges for statistical analysis, and Chapter 20 discusses these in more detail.

By contrast, most of the methods in this book are better suited for **cross-sectional data**, where several variables are measured at the same time point. If we collect data on sales revenue, number of customers, and expenses for last month at *each* Starbucks (more than 25,000 locations as of 2017) at one point in time, these would be cross-sectional data. Cross-sectional data may contain some time information (such as dates), but they aren't a time series because they aren't measured at regular intervals. Because different methods are used to analyze these different types of data, it is important to be able to identify both time series and cross-sectional data sets.

| Year | Total Revenue (in $B) |
|---|---|
| 2007 | 9.44 |
| 2008 | 10.38 |
| 2009 | 9.77 |
| 2010 | 10.71 |
| 2011 | 11.70 |
| 2012 | 13.30 |
| 2013 | 14.79 |
| 2014 | 16.44 |
| 2015 | 19.16 |
| 2016 | 21.31 |

TABLE 1.4 Starbucks's total revenue (in $B) for the years 2007 to 2016.

### IN PRACTICE 1.3 Identifying the types of variables

**MANAGER** I want to understand the data we've collected. How should we begin?

**ANALYST (CARLY)** First we must classify each variable. Here is a list of the variables and their descriptions.

**Account ID** – categorical (nominal, identifier)

**Pre Spending** – quantitative (units $)

**Post Spending** – quantitative (units $)

**Age** – categorical (ordinal). Could be quantitative if we had more precise information

**Segment** – categorical (nominal)

**Enroll?** – categorical (nominal)

**Offer** – categorical (nominal)

**Segment Spend** – quantitative (units $)

All these data are cross-sectional. We do not have successive values over time.

## 1.4 Data Sources: Where, How, and When

We must know *who*, *what*, and *why* to analyze data. Without knowing these three, we don't have enough to start. Of course, we'd always like to know more because the more we know, the more we'll understand and the better our decisions will be. If possible, we'd like to know the *where*, *how*, and *when* of data as well.

[5] Some people differentiate quantitative variables according to whether their measured values have a defined value for zero. This is a technical distinction and usually not one we'll need to make. (For example, it isn't correct to say that a temperature of 80°F is twice as hot as 40°F because 0° is an arbitrary value. On the Celsius scale those temperatures are 26.67°C and 4.44°C—a ratio of 6.) The term *interval scale* is sometimes applied to quantitative variables that lack a defined zero, and the term *ratio scale* is applied to measurements for which such ratios are appropriate.

Values recorded in 1947 may mean something different than similar values recorded last year. Values measured in Abu Dhabi may differ in meaning from similar measurements made in Mexico. Conclusions drawn about data collected from a store in Singapore last spring may not apply to other locations, or other seasons. Knowing *how*, *when*, *where*, and *why* the data were collected can mean the difference between valid and spurious conclusions.

*How* the data are collected can make the difference between insight and nonsense. As we'll see later, data that come from a voluntary survey on the Internet are almost always worthless. In a recent Internet poll, 84% of respondents said "no" to the question of whether subprime borrowers should be bailed out. While it may be true that 84% of those 23,418 respondents did say that, it's dangerous to assume that that group is representative of any larger group. To make inferences from the data you have at hand to the world at large, you need to ensure that the data you have are representative of the larger group. Chapter 8 discusses sound methods for *designing* a *survey* or poll to help ensure that the inferences you make are valid.

Another way to collect valid data is by performing an experiment in which you actively manipulate variables (called factors) to see what happens. Most of the "junk mail" credit card offers that you receive are actually experiments done by marketing groups in those companies. They may make different versions of an offer to selected groups of customers to see which one works best before rolling out the winning idea to the entire customer base. Chapter 9 discusses both the design and the analysis of experiments like these.

Sometimes, the answer to a question you may have can be found in data that someone or some organization has already collected. Internally, companies may analyze data from their own databases or data warehouse. They may also supplement or rely entirely on data collected by others. Many companies, nonprofit organizations, and government agencies collect vast amounts of data via the Internet. Some organizations may charge you a fee for accessing or downloading their data. The U.S. government collects information on nearly every aspect of life in the United States, both social and economic (see, for example, www.census.gov, or more generally, www.usa.gov), as the European Union does for Europe (see ec.europa.eu/eurostat). International organizations such as the World Health Organization (www.who.org) and polling agencies such as Pew Research (www.pewresearch.org) offer information on a variety of current social and demographic trends. Data like these are usually collected for different purposes than to answer your particular business question. So you should be cautious when generalizing from data like these. Unless the data were collected in a way that ensures that they are representative of the population in which you are interested, you may be misled. Chapter 21 discusses data mining, which attempts to use Big Data to make hypotheses and draw insights.

### 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. We found many of the data sets used in this book by searching on the Internet. The Internet has both advantages and disadvantages as a source of data. Among the advantages are the fact that often you'll be able to find even more current data than we present. One disadvantage is that references to Internet addresses can "break" as sites evolve, move, and die. Another disadvantage is that important metadata—information about the collection, quality, and intent of the data—may be missing.

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 such extra symbols as money indicators ($, ¥, £, €); few statistics packages can handle these.

Throughout this book, we often provide a margin note for a new dataset listing some of the W's of the data. When we can, we also offer a reference for the source of the data. It's a habit we recommend. The first step of any data analysis is to know why you are examining the data (what you want to know), whom each row of your data table refers to, and what the variables (the columns of the table) record. These are the *Why*, the *Who*, and the *What*. Identifying them is a key part of the *Plan* step of any analysis. Make sure you know all three before you spend time analyzing the data.

## IN PRACTICE 1.4 Identifying data sources

On the basis of her initial analysis, Carly asks her colleague Ying Mei to e-mail a sample of customers from the Travel and Entertainment segment and ask about their card use and household demographics. Carly asks another colleague, Gregg, to design a study about their double miles offer. In this study, a random sample of customers receives one of three offers: the standard double miles offer; a double miles offer good on any airline; or no offer.

**MANAGER** It looks like we have three sources of data for our customer spending analysis. How are they different and will you analyze them all the same way?

**ANALYST (CARLY)** My data set was derived from routine data that we collect on each customer transaction. The data were not part of a survey or experiment. I will need to be careful in drawing conclusions, but my initial exploration of the data helped me decide that we needed more information. Ying Mei's data come from a designed survey, so it should be representative of our customers. Gregg's data come from a designed experiment, which may allow us to decide which of these offers will work the best.

## JUST CHECKING

**An insurance company that specializes in commercial property insurance has a separate database for their policies that involve churches and schools. Here is a small portion of that database.**

2 List as many of the W's as you can for this data set.

3 Classify each variable as to whether you think it should be treated as categorical or quantitative (or both); if quantitative, identify the units.

| Policy Number | Years Claim Free | Net Property Premium ($) | Net Liability Premium ($) | Total Property Value ($000) | Median Age in ZIP Code | School? | Territory | Coverage |
|---|---|---|---|---|---|---|---|---|
| 4000174699 | 1 | 3107 | 503 | 1036 | 40 | FALSE | AL580 | BLANKET |
| 8000571997 | 2 | 1036 | 261 | 748 | 42 | FALSE | PA192 | SPECIFIC |
| 8000623296 | 1 | 438 | 353 | 344 | 30 | FALSE | ID60 | BLANKET |
| 3000495296 | 1 | 582 | 339 | 270 | 35 | TRUE | NC340 | BLANKET |
| 5000291199 | 4 | 993 | 357 | 218 | 43 | FALSE | OK590 | BLANKET |
| 8000470297 | 2 | 433 | 622 | 108 | 31 | FALSE | NV140 | BLANKET |
| 1000042399 | 4 | 2461 | 1016 | 1544 | 41 | TRUE | NJ20 | BLANKET |
| 4000554596 | 0 | 7340 | 1782 | 5121 | 44 | FALSE | FL530 | BLANKET |
| 3000260397 | 0 | 1458 | 261 | 1037 | 42 | FALSE | NC560 | BLANKET |
| 8000333297 | 2 | 392 | 351 | 177 | 40 | FALSE | OR190 | BLANKET |
| 4000174699 | 1 | 3107 | 503 | 1036 | 40 | FALSE | AL580 | BLANKET |

## WHAT CAN GO WRONG?

- **Don't label a variable as categorical or quantitative without thinking about the data and what they represent.** The same variable can sometimes take on different roles.
- **Don't assume that a variable is quantitative just because its values are numbers.** 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.

## ETHICS IN ACTION

Sarah Potterman, a doctoral student in educational psychology, is researching the effectiveness of various interventions recommended to help children with learning disabilities improve their reading skills. One particularly intriguing approach is an interactive software system that uses analogy-based phonics.

Sarah contacted the company that developed this software, RSPT Inc., to obtain the system free of charge for use in her research. RSPT Inc. expressed interest in having her compare its product with other intervention strategies and was quite confident that its approach would be the most effective. Not only did the company provide Sarah with free software, but RSPT Inc. also generously offered to fund her research with a grant to cover her data collection and analysis costs.

- **Identify the ethical dilemma in this scenario.**
- **What are the undesirable consequences?**
- **Propose an ethical solution that considers the welfare of all stakeholders.**

Jim Hopler is operations manager for a local office of a top-ranked full-service brokerage firm. With increasing competition from both discount and online brokers, Jim's firm has redirected attention to attaining exceptional customer service through its client-facing staff, namely brokers. In particular, management wished to emphasize the excellent advisory services provided by its brokers.

Results from surveying clients about the advice received from brokers at the local office revealed that 20% rated it *poor,* 5% rated it *below average,* 15% rated it *average,* 10% rated it *above average,* and 50% rated it *outstanding*. With corporate approval, Jim and his management team instituted several changes in an effort to provide the best possible advisory services at the local office. Their goal was to increase the percentage of clients who viewed their advisory services as *outstanding*.

Surveys conducted after the changes were implemented showed the following results: 5% *poor,* 5% *below average,* 20% *average,* 40% *above average,* and 30% *outstanding*. In discussing these results, the management team expressed concern that the percentage of clients who considered their advisory services *outstanding* fell from 50% to 30%.

One member of the team suggested an alternative way of summarizing the data. By coding the categories on a scale from 1 = poor to 5 = outstanding and computing the average, they found that the average rating increased from 3.65 to 3.85 as a result of the changes implemented. Jim was delighted to see that their changes were successful in improving the level of advisory services offered at the local office. In his report to corporate, he only included average ratings for the client surveys.

- **Identify the ethical dilemma in this scenario.**
- **What are the undesirable consequences?**
- **Propose an ethical solution that considers the welfare of all stakeholders.**

# CHAPTER 1 FROM LEARNING TO EARNING

## LEARNING OBJECTIVES

**Understand the business context of the data and the problem you are trying to solve to be successful when making decisions from data.**

- *Who*, *what*, *why*, *where*, *when* (and *how*)—the W's—help nail down the context of the data.
- We must know *who*, *what*, and *why* to be able to say anything useful based on the data. The *who* are the cases (or records or rows). The *what* are the variables. A variable gives information about each of the cases. The *why* helps us decide which way to treat the variables.
- Stop and identify the W's whenever you have data, and be sure you can identify the cases and the variables.

**Identify whether a variable is being used as categorical or quantitative.**

- Categorical variables identify a category for each case. Usually we think about the counts of cases that fall in each category. (An exception is an identifier variable that just names each case.)
- Quantitative variables record measurements or amounts of something; they must have units.
- Sometimes we may treat the same variable as categorical or quantitative depending on what we want to learn from it, which means some variables can't be pigeonholed as one type or the other.

**Consider the source of your data and the reasons the data were collected. That can help you understand what you might be able to learn from the data.**

## TERMS

**Big Data** The collection and analysis of data sets so large and complex that traditional methods typically brought to bear on the problem would be overwhelmed.

**Business analytics** The process of using statistical analysis and modeling to drive business decisions.

**Case** A case is an individual about whom or which we have data. Also called a record or row.

**Categorical (or qualitative) variable** A variable that names categories (whether with words or numerals) is called categorical or qualitative.

**Context** The context ideally tells *who* was measured, *what* was measured, *how* the data were collected, *where* the data were collected, and *when* and *why* the study was performed.

**Cross-sectional data** Data taken from situations that vary over time but measured at a single time instant are said to be a cross-section of the time series.

**Data** Recorded values, whether numbers or labels, together with their context.

**Data mining (or predictive analytics)** The process of using a variety of statistical tools to analyze large databases or data warehouses.

**Data table** An arrangement of data in which each row represents a case and each column represents a variable.

**Data warehouse** A large database of information collected by a company or other organization usually to record transactions that the organization makes, but also used for analysis via data mining.

**Experimental unit** An individual in a study for which or for whom data values are recorded. Human experimental units are usually called subjects or participants.

**Identifier variable** A categorical variable that records a unique value for each case, used to name or identify it.

**Metadata** Auxiliary information about variables in a database, typically including *how*, *when*, and *where* (and possibly *why*) the data were collected; *who* each case represents; and the definitions of all the variables.

**Nominal variable** The term "nominal" can be applied to a variable whose values are used only to name categories.

**Ordinal variable** The term "ordinal" can be applied to a variable whose categorical values possess some kind of order.

**Participant** A human experimental unit. Also called a subject.

**Quantitative variable** A variable in which the numbers are values of measured quantities with units.

**Record** Information about an individual in a database.

**Relational database** A relational database stores and retrieves information. Within the database, information is kept in data tables that can be "related" to each other.

**Respondent** Someone who answers, or responds to, a survey.

**Spreadsheet** A spreadsheet is a layout designed for accounting that is often used to store and manage data tables. Excel is a common example of a spreadsheet program.

**Subject** A human experimental unit. Also called a participant.

**Time series** Data measured over time. Usually the time intervals are equally spaced or regularly spaced (e.g., every week, every quarter, or every year).

**Units** A quantity or amount adopted as a standard of measurement, such as dollars, hours, or grams.

**Variable** A variable holds information about the same characteristic for many cases.

## TECH SUPPORT Entering Data

These days, nobody does statistics by hand. We use technology: a programmable calculator or a statistics program on a computer. Professionals all use a *statistics package* designed for the purpose. We will provide many examples of results from a statistics package throughout the book. Rather than choosing one in particular, we'll offer generic results that look like those produced by all the major statistics packages but don't exactly match any of them. Then, in the Tech Support section at the end of each chapter, we'll provide hints for getting started on several of the major packages.

If you understand what the computer needs to know to do what you want and what it needs to show you in return, you can figure out the specific details of most packages pretty easily.

For example, to get your data into a computer statistics package, you need to tell the computer:

- Where to find the data. This usually means directing the computer to a file stored on your computer's disk or to data on a database. Or it might just mean that you have copied the data from a spreadsheet program or Internet site and it is currently on your computer's clipboard. Usually, the data should be in the form of a data table with cases in the rows and variables in the columns. Most computer statistics packages prefer the *delimiter* that marks the division between elements of a data table to be a tab character (comma is another common delimiter) and the delimiter that marks the end of a case to be a *return* character.
- Where to put the data. (Usually this is handled automatically.)
- What to call the variables. Some data tables have variable names as the first row of the data, and often statistics packages can take the variable names from the first row automatically.
- Excel is often used to help organize, manipulate, and prepare data for other software packages. Many of the other packages take Excel files as inputs. Alternatively, you can copy a data table from Excel and Paste it into many packages, or export Excel spreadsheets as tab delimited (.txt) or comma delimited files (.csv), which can be easily shared and imported into other programs. All data files provided with this text are in tab-delimited text (.txt) format.

### EXCEL

To open a file containing data in Excel:

- Choose **File > Open**.
- Browse to find the file to open. Excel supports many file formats.
- Other programs can import data from a variety of file formats, but all can read both tab delimited (.txt) and comma delimited (.csv) text files.
- You can also copy tables of data from other sources, such as Internet sites, and paste them into an Excel spreadsheet. Excel can recognize the format of many tables copied this way, but this method may not work for some tables.
- Excel may not recognize the format of the data. If data include dates or other special formats ($, €, ¥, etc.), identify the desired format. Select the cells or columns to reformat and in the **Home** menu choose **Home** > **Format** > **Format Cells** (found under Cells). Often, the General format is the best option for data you plan to move to a statistics package.

### JMP

To import a text file:

- Choose **File > Open** and select the file from the dialog. At the bottom of the dialog screen you'll see **Open As:**—be sure to change to **Data with Preview**. This will allow you to specify the delimiter and make sure the variable names are correct. (JMP also allows various formats to be imported directly, including .xls files.)

You can also paste a data set in directly (with or without variable names) by selecting:

- **File > New > Data Table** and then **Edit > Paste** (or **Paste with Column Names** if you copied the names of the variables as well).

Finally, you can import a data set from a URL directly by selecting:

- **File > Internet Open** and pasting in the address of the website. JMP will attempt to find data on the page. It may take a few tries and some edits to get the data set in correctly.

### MINITAB

To import a text or Excel file:

- Choose **File > Open**. From **Files of type**, choose **Text (*.txt)** or **Excel (*.xls; *xlsx)**.
- Browse to find and select the file.
- In the lower right corner of the dialog, choose **Open** to open the data file.
- Click **Open**.

### R

**R** can import many types of files, but text files (tab or comma delimited) are easiest. If the file is tab delimited and contains the variable names in the first row, then:

```
> mydata = read.delim(file.choose())
```

will give a dialog where you can pick the file you want to import. It will then be in a data frame called mydata. If the file is comma delimited, use:

```
> mydata = read.csv(file.choose())
```

#### COMMENTS

RStudio provides an interactive dialog that may be easier to use. For other options, including the case that the file does not contain variable names, consult **R** help.

### SPSS

To import a text file:

- Choose **File > Open > Data**. Under "Files of type," choose **Text (*.txt,*.dat)**. Select the file you want to import. Click **Open**.
- A window will open called Text Import Wizard. Follow the steps, depending on the type of file you want to import.

### STATCRUNCH

StatCrunch offers several ways to enter data. Click **MyStatCrunch > My Data**. Click a dataset to analyze the data or edit its properties.

Click a data set link to analyze the data or edit its properties to import a new data set.

- Choose **Select a file on my computer**,
- Enter the URL of a file,
- Paste data into a form, or
- Type or paste data into a blank data table.

For the "select a file on my computer" option, StatCrunch offers a choice of space, comma, tab, or semicolon delimiters. You may also choose to use the first line as the names of the variables.

After making your choices, select the **Load File** button at the bottom of the screen.

BRIEF CASE

## Credit Card Bank

Like all credit and charge card companies, this company makes money on each of its cardholders' transactions. Thus, its profitability is directly linked to card usage. To increase customer spending on its cards, the company sends many different offers to its cardholders, and market researchers analyze the results to see which offers yield the largest increases in the average amount charged.

The dataset (**Credit Card Bank**) is part of a much larger database actually used by the researchers. For each customer, it contains several variables in a spreadsheet. Information on the variables is found in the file **Credit Card Bank Info**.

Examine the data in the data file. List as many of the W's as you can for these data and classify each variable as categorical or quantitative. If quantitative, identify the units.

# CHAPTER 1 EXERCISES

## SECTION 1.2

**1.** A headhunter company collected information on a group of job hunters. The first six applicants' data appear below. The columns correspond to the job hunter's name, gender, age, position applied to, earliest date of joining, and salary expectation (in U.S. dollars).

**a)** What does a row correspond to in this data table? How would you best describe its role: as a participant, subject, case, respondent, experimental unit, or something else?
**b)** How many variables are measured in each row?

| Name | Gender | Age | Position Applied | Earliest Joining Date | Year(s) of Experience | Expected Salary ($) |
|---|---|---|---|---|---|---|
| Patrick Martin | M | 35 | Assistant Manager | 5/11/2020 | 10 | 8,000 |
| Mohamad Leo Said | M | 25 | Accountant | 6/1/2020 | 6 | 5,500 |
| Miranda Scott | F | 20 | Junior Executive | 5/15/2020 | 1 | 3,000 |
| Justin Phang Kok Min | M | 23 | Management Officer | 8/1/2020 | 2 | 3,500 |
| Chrystal Ng Jit | F | 18 | General Clerk | 6/1/2020 | 0 | 2,800 |
| Siti Harley | F | 30 | Sales Executive | 5/11/2020 | 5 | 3,000 |

**2.** A local bookstore is keeping a database of its customers to find out more about their spending habits so that the store can start to make personal recommendations based on past purchases. Here are the first five rows of their database:

**a)** What does a row correspond to in this data table? How would you best describe its role: as a participant, subject, case, respondent, or experimental unit?
**b)** How many variables are measured for each case?

| Transaction ID | Customer ID | Date | ISBN Number of Purchase | Price | Coupon? | Gift? | Quantity |
|---|---|---|---|---|---|---|---|
| 29784320912 | 4J438 | 11/12/2009 | 345-23-2355 | $29.95 | N | N | 1 |
| 26483589001 | 3K729 | 9/30/2009 | 983-83-2739 | $16.99 | N | N | 1 |
| 26483589002 | 3K729 | 9/30/2009 | 102-65-2332 | $9.95 | Y | N | 1 |
| 36429489305 | 3U034 | 12/5/2009 | 295-39-5884 | $35.00 | N | Y | 1 |
| 36429489306 | 3U034 | 12/5/2009 | 183-38-2957 | $79.95 | N | Y | 1 |

## SECTION 1.3

**3.** Referring to the headhunter company's data table of Exercise 1,

**a)** For each variable, would you describe it as primarily categorical, or quantitative? If quantitative, what are the units? If categorical, is it ordinal or simply nominal?
**b)** Are these data a time series, or are these cross-sectional? Explain briefly.

**4.** Referring to the bookstore data table of Exercise 2,

**a)** For each variable, would you describe it as primarily categorical, or quantitative? If quantitative, what are the units? If categorical, is it ordinal or simply nominal?
**b)** Are these data a time series, or are these cross-sectional? Explain briefly.

## SECTION 1.4

**5.** For the headhunter company data of Exercise 1, do the data appear to have come from a designed survey or experiment? What concerns might you have about drawing conclusions from this data set?

**6.** A student finds data on an Internet site that contains financial information about selected companies. He plans to analyze the data and use the results to develop a stock investment strategy. What kind of data source is he using? What concerns might you have about drawing conclusions from this data set?

## CHAPTER EXERCISES

*For each description of data in Exercises 7 to 26, identify the W's, name the variables, specify for each variable whether its use indicates it should be treated as categorical or quantitative, and for any quantitative variable identify the units in which it was measured (if they are not provided, give some possible units in which they might be measured). Specify whether the data come from a designed survey or experiment. Are the variables time series or cross-sectional? Report any concerns you have as well.*

**7. The news.** Find a newspaper or magazine article in which some data are reported (e.g., see *The Wall Street Journal*, *Financial Times*, *Business Week*, or *Fortune*). For the data discussed in the article, answer the questions above. Include a copy of the article with your report.

**8. The Internet.** Find an Internet site on which some data are reported. For the data found on the site, answer as many of the questions above as you can. Include a copy of the URL with your report.

**9. Survey.** A four-star hotel provides a rating form to its customers during their checkout to evaluate their satisfaction with the hotel's customer service, restaurant service, bar service, room service, and housekeeping. The rating scale is from 1 = very unsatisfied to 5 = very satisfied.

**10. Product testing.** A marketing department wants to conduct a market research to know what customers think about their new model of vacuum cleaner. The new vacuum model will be examined by collecting customer feedback on first impressions, quality of product, innovativeness, and the value for money using the 5-point Likert scale (1 = very bad to 5 = very good).

**11. Evaluation.** A company is using the Employee Performance Evaluation Excel template available at the ExcelDataPro website (exceldatapro.com/employee-evaluation-template) to evaluate its current employees. Answer the questions above for the following indicators: designation, evaluation purpose, review period, functional skills, and interpersonal skills.

**12. Pets' meal.** Based on the needs of pets sent to a veterinary clinic, the clinic will provide each animal with different types of meals. The meals provide different amounts of calories (in kcal), fats (in grams), proteins (in grams), and fibers (in grams) that suits each individual pet.

**13. Spending behavior.** A study was conducted to investigate the factors that influence students' spending behavior toward brand equity in the fashion industry. An online questionnaire was circulated in Klang Valley, Malaysia. The questions included specifying the sex of each respondent; the age of each respondent; how strongly students are aware of brand awareness, brand loyalty, and perceived quality while making their fashion purchases.

**14. Dining experience.** A popular local café would like to know about its customers' dining experience and the satisfaction levels related to the food quality, service quality, and restaurant environment. A satisfaction feedback form had been prepared for the customers to fill up while they are paying their bill.

**15. Planting.** Two groups of plants were used in a study in which one group is given fertilizer and the other is not. If there are any differences between the fertilized plant group and the unfertilized plant group, they may be due to the addition of the fertilizer.

**16. World Values Survey (WVS).** To study changing values and their impact on society and politics, the WVS (www.worldvaluessurvey.org) has designed a cultural map of the world, with nine cultural regions, such as Confucian and Orthodox, instead of the seven continents. Countries are also assigned numerical scores based on two major dimensions of cross-cultural variations: traditional values versus secular-rational values and survival values versus self-expression values.

**17. Olive oil producers.** A local farmers' association hoping to provide better services to its olive oil producers sent out a questionnaire to a random sample of producers requesting information about gross sales, percent profit, unit price, varieties, age, locality, and average production per olive tree.

**18. OECD better life initiative.** Established in Paris in 1961, the Organization for Economic Cooperation and Development (OECD) collects information on many economic and social aspects of countries around the world. The OECD Better Life Initiative is an attempt to compare well-being across the OECD member countries. Central to this initiative is an interactive tool known as "Your Better Life Index." It lets one rank their country based on criteria such as education, jobs, income, and work–life balance. Each of these are described by one or more indicators. For example, education is an average of scores that official statistics assign to educational attainment, reading skills, and years in education (www.oecdbetterlifeindex.org/).

**19. EPA.** The Environmental Protection Agency (EPA) tracks fuel economy of automobiles. Among the data EPA analysts collect from the manufacturer are the manufacturer (Ford, Toyota, etc.), vehicle type (car, SUV, etc.), weight, horsepower, and gas mileage (mpg) for city and highway driving.

**20. Manufacturing.** A solar panel manufacturing firm conducts a study to determine whether their newly designed solar panel would be cost competitive to their current model. The manufacturer records the costs for the materials, components, energy, machines, and labor utilized.

**21. Hair coloring.** A hair coloring products company is testing the quality and effectiveness of one of its newly launched products. There are 50 participants involved in the experiment of whom 25 have black hair and 25 have white hair. The company records the gender, age, time taken to color the participants' hair, effectiveness of the color on the participants' hair, and the participants' satisfaction with the result.

**22. Courier.** LEX is a courier company that provides standard courier services, overnight courier services, same-day express courier services, and international courier services. Based on the type of service, LEX wants to determine the accuracy of its delivery duration across different locations. In order to do so, it retrieves delivery records from its internal database.

**23. Mobile Apps.** Kate is a mobile apps developer. She conducts a survey to get feedback on her most recently developed mobile app in order to improve it. Her survey questions cover the gender of each user, the age of each user, whether the app runs smoothly or not, whether users like the app design, if it helps users achieve their goals, and if users would recommend this app to their friends. Apart from specifying their gender and age, user responses are based on a 5-point Likert scale (1 = strongly disagree to 5 = strongly agree).

**24. Health.** A people's association intends to know the diet habits, physical fitness, and the kinds of exercise the residents do in a particular state. A campaign has been kick-started across the state to collect this information.

**25. Taxi data.** The market share analysis of rides in New York City was based in part on the data from the New York City Taxi and Limousine Commission. Data on each ride can be found on the website www.nyc.gov/html/tlc/html/about/trip_record_data.shtml. Because there are over 10,000,000 rides a month, the monthly files are large—up to 2 GB each. For each ride they contain information on the location, costs, and date, among other variables. Here are 13 columns and the first 11rows (of 11,934,338!) for Yellow Taxis in the month of April 2016. (For a fascinating analysis of taxi data in New York and Chicago, visit the website toddwschneider.com/)

| VendorID | TPEP_ pickup | TPEP_ dropof | Passenger_ count | Trip_ distance | Pickup_ longitude | Pickup_ latitude | Dropoff_ longitude | Dropoff_ latitude | Fare_ amount | Tip_ amount | Tolls_ amount | Total_ amount |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 4/1/16 0:00 | 4/1/16 0:01 | 1 | 0.5 | −73.976883 | 40.7584953 | −73.977669 | 40.7539024 | 3.5 | 0 | 0 | 4.8 |
| 1 | 4/1/16 0:00 | 4/1/16 0:12 | 2 | 2.2 | −73.985207 | 40.7572937 | −73.989288 | 40.7326584 | 10 | 2.25 | 0 | 13.55 |
| 2 | 4/1/16 0:00 | 4/1/16 0:10 | 2 | 0.96 | −73.979202 | 40.7588692 | −73.990677 | 40.7513199 | 8.5 | 0 | 0 | 9.8 |
| 2 | 4/1/16 0:00 | 4/1/16 0:10 | 5 | 1.54 | −73.984856 | 40.7677231 | −73.990829 | 40.7511864 | 8.5 | 1.96 | 0 | 11.76 |
| 2 | 4/1/16 0:00 | 4/1/16 0:00 | 2 | 10.45 | −73.863739 | 40.7694702 | −73.976814 | 40.7752838 | 34 | 8.07 | 5.54 | 48.41 |
| 1 | 4/1/16 0:00 | 4/1/16 0:15 | 1 | 3.5 | −73.973373 | 40.7570763 | −73.933472 | 40.766304 | 14 | 3 | 0 | 18.3 |
| 1 | 4/1/16 0:00 | 4/1/16 0:08 | 1 | 4.4 | −73.790092 | 40.6470833 | −73.793915 | 40.6673737 | 13.5 | 0 | 0 | 14.8 |
| 1 | 4/1/16 0:00 | 4/1/16 0:03 | 1 | 0.6 | −73.988899 | 40.7454262 | −73.991821 | 40.7384453 | 4.5 | 1.15 | 0 | 6.95 |
| 2 | 4/1/16 0:00 | 4/1/16 0:03 | 2 | 0.81 | −73.985275 | 40.747364 | −73.985657 | 40.7550812 | 4.5 | 1 | 0 | 6.8 |
| 1 | 4/1/16 0:00 | 4/1/16 0:04 | 1 | 0.8 | 0 | 0 | −73.977692 | 40.7538757 | 5 | 0 | 0 | 6.3 |
| 1 | 4/1/16 0:00 | 4/1/16 0:06 | 2 | 1.8 | −73.979752 | 40.7809486 | −73.966621 | 40.8028374 | 7.5 | 0 | 0 | 8.8 |

| ID | Age | Plan to Purchase Car? | City or Rural? | Mobile Device | Education | Gender | Latitude | Longitude | Country | Town Size | Household Size |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 00001700-9f84-0134-2d50-0aaafcbd6b1f | 29 | yes | city | mobile | high | female | 1.2855 | 103.8565 | Singapore | City with 5 million–10 million people | 4 |
| 0001a0c0-a08e-0134-cf68-0a62e1402143 | 39 | yes | city | mobile | high | female | −12.0678 | −77.0886 | Peru | Town with 1 000–50 000 people | 5 or more |
| 000497f0-9efa-0134-9d13-0aaafcbd6b1f | 19 | no | city | mobile | low | female | −33.0422 | −71.3733 | Chile | City with 250 000–1 million people | 5 or more |
| 000501c0-ba4c-0134-49bc-0aeaf0818377 | 27 | no | city | desktop | medium | female | 4.6492 | −74.0628 | Colombia | City with more than 10 million people | 2 |
| 0007a460-9ef0-0134-0666-0a62e1402143 | 35 | yes | city | desktop | high | female | 41.0214 | 28.9684 | Turkey | City with more than 10 million people | 5 or more |
| 000861c0-a289-0134-30ba-0a62e1402143 | 53 | no | rural | mobile | high | male | 33.7208 | 130.6997 | Japan | Town with 1 000–50 000 people | 2 |
| 0008dbc0-ac5e-0134-cb4a-0aaafcbd6b1f | 21 | yes | city | tablet | high | female | 43.122 | −79.805 | Canada | City with 250 000–1 million people | 4 |
| 000920d0-9f42-0134-54fc-0aaafcbd6b1f | 19 | no | city | mobile | high | female | 3.0833 | 101.5333 | Malaysia | City with 50 000–250 000 people | 2 |
| 000ae030-b561-0134-aeb9-0aaafcbd6b1f | 44 | yes | city | mobile | high | female | 1.2855 | 103.8565 | Singapore | City with 50 000–250 000 people | 2 |
| 000cb8d0-9df1-0134-4a2a-0aaafcbd6b1f | 20 | no | city | mobile | high | female | 14.5955 | 120.9721 | Philippines | City with 250 000–1 million people | 3 |

**26. Dalia Research.** The data set from Dalia Research (you'll learn more about Dalia in Chapter 2) contains 763 variables, including demographic information on the 43,034 respondents and survey questions on a variety of topics from intention to purchase a car to quantity of bottled water consumption. Above is a small piece of that data set.

*When you organize data in a spreadsheet, it is important to lay it out as a data table. For each of these examples in Exercises 27 to 30, show how you would lay out these data. Indicate the headings of columns and what would be found in each row.*

**27. Mortgages.** For a study of mortgage loan performance: loan number, last 4 digits of borrower's social security number, amount of the loan, the name of the borrower.

**28. Restaurant satisfaction.** Data collected to determine the customers' satisfaction of a restaurant: receipt number, age, quality of food rating (1–10), service rating (1–10), cleanliness rating (1–10), and parking (1–10).

**29. Education in the Better Life Initiative.** The OECD 2020 data file contains the following data related to education: country name, educational attainment, student skills (average score), and years in education.

**30. Company annual dinner.** Data collected to identify employees who are going to attend the company annual dinner: Employee ID, department, bringing partner/spouse (Yes / No), and vegetarian (Yes / No).

*For the following examples in Exercises 31 to 34, indicate whether the data are time series or cross-sectional.*

**31. Road accidents.** Road accidents reported to the Royal Malaysia Police in April 2020.

**32. Transport registration.** New monthly registrations of motor vehicles by type in the year 2019.

**33. OECD well-being.** Comparison of the well-being indicators in the OECD 2020 data file for 40 different countries.

**34. Developments in well-being.** The 2017 OECD Better Life Initiative data for Spain in the second wave compared to the OECD 2020 data of the first wave.

## JUST CHECKING ANSWERS

**1** The row refers to an order. You can tell because there is a unique purchase order number (first column), but not unique customers. Some customers made more than one purchase.

**2** Who—policies on churches and schools

What—policy number, years claim free, net property premium ($), net liability premium ($), total property value ($000), median age in ZIP code, school?, territory, coverage

How—company records

When—not given

**3** Policy number: identifier (categorical)

Years claim free: quantitative

Net property premium: quantitative ($)

Net liability premium: quantitative ($)

Total property value: quantitative ($)

Median age in ZIP code: quantitative

School?: categorical (true/false)

Territory: categorical

Coverage: categorical