

1 Hello data

Scientists seek to answer questions using rigorous methods and careful observations. These observations – collected from the likes of field notes, surveys, and experiments – form the backbone of a statistical investigation and are called **data**. Statistics is the study of how best to collect, analyze, and draw conclusions from data. In this first chapter, we focus on both the properties of data and on the collection of data.

1.1 Case study: Using stents to prevent strokes

In this section we introduce a classic challenge in statistics: evaluating the efficacy of a medical treatment. Terms in this section, and indeed much of this chapter, will all be revisited later in the text. The plan for now is simply to get a sense of the role statistics can play in practice.

An experiment is designed to study the effectiveness of stents in treating patients at risk of stroke ([Chimowitz et al. 2011](#)). Stents are small mesh tubes that are placed inside narrow or weak arteries to assist in patient recovery after cardiac events and reduce the risk of an additional heart attack or death.

Many doctors have hoped that there would be similar benefits for patients at risk of stroke. We start by writing the principal question the researchers hope to answer:

Does the use of stents reduce the risk of stroke?

The researchers who asked this question conducted an experiment with 451 at-risk patients. Each volunteer patient was randomly assigned to one of two groups:

- **Treatment group.** Patients in the treatment group received a stent and medical management. The medical management included medications, management of risk factors, and help in lifestyle modification.
- **Control group.** Patients in the control group received the same medical management as the treatment group, but they did not receive stents.

Researchers randomly assigned 224 patients to the treatment group and 227 to the control group. In this study, the control group provides a reference point against which we can measure the medical impact of stents in the treatment group.

Researchers studied the effect of stents at two time points: 30 days after enrollment and 365 days after enrollment. The results of 5 patients are summarized in Table 1.1. Patient outcomes are recorded as `stroke` or `no event`, representing whether the patient had a stroke during that time period.

The `stent30` data and `stent365` data can be found in the [openintro](#) R package.

patient	group	30 days	365 days
1	treatment	no event	no event
2	treatment	stroke	stroke
3	treatment	no event	no event
4	treatment	no event	no event
5	control	no event	no event

Table 1.1: Results for five patients from the stent study.

On this page

- [1 Hello data](#)
- [1.1 Case study: Using stents to prevent strokes](#)
- [1.2 Data basics](#)
 - [1.2.1 Observations, variables, and data matrices](#)
 - [1.2.2 Types of variables](#)
 - [1.2.3 Relationships between variables](#)
 - [1.2.4 Explanatory and response variables](#)
 - [1.2.5 Observational studies and experiments](#)
- [1.3 Chapter review](#)
 - [1.3.1 Summary](#)
 - [1.3.2 Terms](#)
- [1.4 Exercises](#)

[View source](#)

[Edit this page](#)

It would be difficult to answer a question on the impact of stents on the occurrence of strokes for **all** study patients using these *individual* observations. This question is better addressed by performing a statistical data analysis of *all* observations. Table 1.2 summarizes the raw data in a more helpful way. In this table, we can quickly see what happened over the entire study. For instance, to identify the number of patients in the treatment group who had a stroke within 30 days after the treatment, we look in the leftmost column (30 days), at the intersection of treatment and stroke: 33. To identify the number of control patients who did not have a stroke after 365 days after receiving treatment, we look at the rightmost column (365 days), at the intersection of control and no event: 199.

Group	30 days		365 days	
	Stroke	No event	Stroke	No event
Control	13	214	28	199
Treatment	33	191	45	179
Total	46	405	73	378

Table 1.2: Descriptive statistics for the stent study.

GUIDED PRACTICE

Of the 224 patients in the treatment group, 45 had a stroke by the end of the first year. Using these two numbers, compute the proportion of patients in the treatment group who had a stroke by the end of their first year. (Note: answers to all Guided Practice exercises are provided in footnotes!)³

We can compute summary statistics from the table to give us a better idea of how the impact of the stent treatment differed between the two groups. A **summary statistic** is a single number summarizing data from a sample. For instance, the primary results of the study after 1 year could be described by two summary statistics: the proportion of people who had a stroke in the treatment and control groups.

- Proportion who had a stroke in the treatment (stent) group: $45/224 = 0.20 = 20\%$.
- Proportion who had a stroke in the control group: $28/227 = 0.12 = 12\%$.

These two summary statistics are useful in looking for differences in the groups, and we are in for a surprise: an additional 8% of patients in the treatment group had a stroke! This is important for two reasons. First, it is contrary to what doctors expected, which was that stents would *reduce* the rate of strokes. Second, it leads to a statistical question: do the data show a “real” difference between the groups?

This second question is subtle. Suppose you flip a coin 100 times. While the chance a coin lands heads in any given coin flip is 50%, we probably won’t observe exactly 50 heads. This type of variation is part of almost any type of data generating process. It is possible that the 8% difference in the stent study is due to this natural variation. However, the larger the difference we observe (for a particular sample size), the less believable it is that the difference is due to chance. So, what we are really asking is the following: if in fact stents have no effect, how likely is it that we observe such a large difference?

While we do not yet have statistical tools to fully address this question on our own, we can comprehend the conclusions of the published analysis: there was compelling evidence of harm by stents in this study of stroke patients.

Be careful: Do not generalize the results of this study to all patients and all stents. This study looked at patients with very specific characteristics who volunteered to be a part of this study and who may not be representative of all stroke patients. In addition, there are many types of stents, and this study only considered the self-expanding Wingspan stent (Boston Scientific). However, this study does leave us with an important lesson: we should keep our eyes open for surprises.

1.2 Data basics

Effective presentation and description of data is a first step in most analyses. This section introduces one structure for organizing data as well as some terminology that will be used throughout this book.

1.2.1 Observations, variables, and data matrices

Table 1.3 displays six rows of a dataset for 50 randomly sampled loans offered through Lending Club, which is a peer-to-peer lending company. This dataset will be referred to as `loan50`.

The `loan50` data can be found in the [openintro](#) R package.

Each row in the table represents a single loan. The formal name for a row is a **case** or **observational unit**. The columns represent characteristics of each loan, where each column is referred to as a **variable**. For example, the first row represents a loan of \$22,000 with an interest rate of 10.90%, where the borrower is based in New Jersey (NJ) and has an income of \$59,000.

GUIDED PRACTICE

What is the grade of the first loan in Table 1.3? And what is the home ownership status of the borrower for that first loan? Reminder: for these Guided Practice questions, you can check your answer in the footnote.⁴

In practice, it is especially important to ask clarifying questions to ensure important aspects of the data are understood. For instance, it is always important to be sure we know what each variable means and its units of measurement. Descriptions of the variables in the `loan50` dataset are given in Table 1.4.

	loan_amount	interest_rate	term	grade	state	total_income	homeownership
1	22,000	10.90	60	B	NJ	59,000	rent
2	6,000	9.92	36	B	CA	60,000	rent
3	25,000	26.30	36	E	SC	75,000	mortgage
4	6,000	9.92	36	B	CA	75,000	rent
5	25,000	9.43	60	B	OH	254,000	mortgage
6	6,400	9.92	36	B	IN	67,000	mortgage

Table 1.3: Six observations from the `loan50` dataset

Variable	Description
loan_amount	Amount of the loan received, in US dollars.
interest_rate	Interest rate on the loan, in an annual percentage.
term	The length of the loan, which is always set as a whole number of months.
grade	Loan grade, which takes values A through G and represents the quality of the loan and its likelihood of being repaid.
state	US state where the borrower resides.
total_income	Borrower's total income, including any second income, in US dollars.
homeownership	Indicates whether the person owns, owns but has a mortgage, or rents.

Table 1.4: Variables and their descriptions for the `loan50` dataset.

The data in Table 1.3 represent a **data frame**, which is a convenient and common way to organize data, especially if collecting data in a spreadsheet. A data frame where each row is a unique case (observational unit), each column is a variable, and each cell is a single value is commonly referred to as **tidy data** [Wickham \(2014\)](#).

When recording data, use a tidy data frame unless you have a very good reason to use a different structure. This structure allows new cases to be added as rows or new variables as new columns and facilitates visualization, summarization, and other statistical analyses.

GUIDED PRACTICE

The grades for assignments, quizzes, and exams in a course are often recorded in a gradebook that takes the form of a data frame. How might you organize a course's grade data using a data frame? Describe the observational units and variables.⁵

GUIDED PRACTICE

We consider data for 3,142 counties in the United States, which includes the name of each county, the state where it resides, its population in 2017, the population change from 2010 to 2017, poverty rate, and nine additional characteristics. How might these data be organized in a data frame?⁶

The data described in the Guided Practice above represents the `county` dataset, which is shown as a data frame in Table 1.5. The variables as well as the variables in the dataset that did not fit in Table 1.5 are described in Table 1.6.

name	state	pop2017	pop_change	unemployment_rate	median_edu
Autauga County	Alabama	55,504	1.48	3.86	some_college
Baldwin County	Alabama	212,628	9.19	3.99	some_college
Barbour County	Alabama	25,270	-6.22	5.90	hs_diploma
Bibb County	Alabama	22,668	0.73	4.39	hs_diploma
Blount County	Alabama	58,013	0.68	4.02	hs_diploma
Bullock County	Alabama	10,309	-2.28	4.93	hs_diploma

Table 1.5: Six observations and six variables from the `county` dataset.

Variable	Description
name	Name of county.
state	Name of state.
pop2000	Population in 2000.
pop2010	Population in 2010.
pop2017	Population in 2017.
pop_change	Population change from 2010 to 2017 (in percent).
poverty	Percent of population in poverty in 2017.
homeownership	Homeownership rate, 2006-2010.
multi_unit	Multi-unit rate: percent of housing units that are in multi-unit structures, 2006-2010.
unemployment_rate	Unemployment rate in 2017.
metro	Whether the county contains a metropolitan area, taking one of the values yes or no.

Variable	Description
median_edu	Median education level (2013-2017), taking one of the values below_hs, hs_diploma, some_college, or bachelors.
per_capita_income	Per capita (per person) income (2013-2017).
median_hh_income	Median household income.
smoking_ban	Describes the type of county-level smoking ban in place in 2010, taking one of the values none, partial, or comprehensive.

Table 1.6: Variables and their descriptions for the `county` dataset.

The `county` data can be found in the [usdata](#) R package.

1.2.2 Types of variables

Examine the `unemployment_rate`, `pop2017`, `state`, and `median_edu` variables in the `county` dataset. Each of these variables is inherently different from the other three, yet some share certain characteristics.

First consider `unemployment_rate`, which is said to be a **numerical** variable since it can take a wide range of numerical values, and it is sensible to add, subtract, or take averages with those values. On the other hand, we would not classify a variable reporting telephone area codes as numerical since the average, sum, and difference of area codes does not have any clear meaning. Instead, we would consider area codes as a categorical variable.

The `pop2017` variable is also numerical, although it seems to be a little different than `unemployment_rate`. This variable of the population count can only take whole non-negative numbers (0, 1, 2, ...). For this reason, the population variable is said to be **discrete** since it can only take numerical values with jumps. On the other hand, the unemployment rate variable is said to be **continuous**.

The variable `state` can take up to 51 values after accounting for Washington, DC: AL, AK, ..., and WY. Because the responses themselves are categories, `state` is called a **categorical** variable, and the possible values (states) are called the variable's **levels** (e.g., DC, AL, AK, etc.).

Finally, consider the `median_edu` variable, which describes the median education level of county residents and takes values `below_hs`, `hs_diploma`, `some_college`, or `bachelors` in each county. This variable seems to be a hybrid: it is a categorical variable, but the levels have a natural ordering. A variable with these properties is called an **ordinal** variable, while a regular categorical variable without this type of special ordering is called a **nominal** variable. To simplify analyses, any categorical variable in this book will be treated as a nominal (unordered) categorical variable.

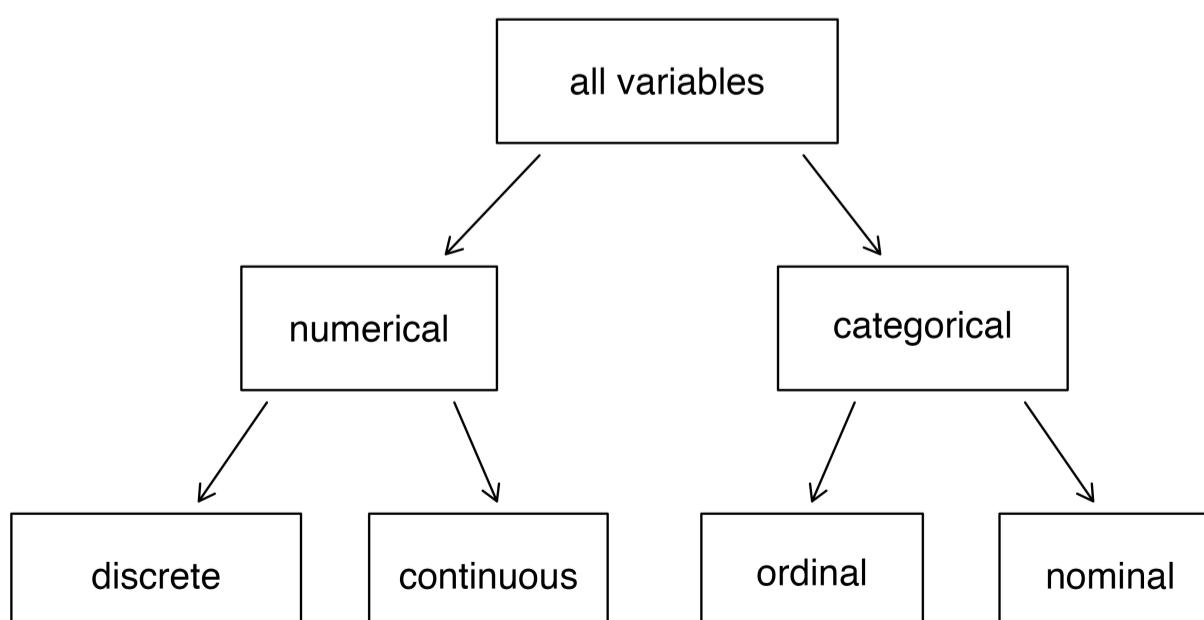


Figure 1.1: Breakdown of variables into their respective types.

EXAMPLE:

Data were collected about students in a statistics course. Three variables were recorded for each student: number of siblings, student height, and whether the student had previously taken a statistics course. Classify each of the variables as continuous numerical, discrete numerical, or categorical.

The number of siblings and student height represent numerical variables. Because the number of siblings is a count, it is discrete. Height varies continuously, so it is a continuous numerical variable. The last variable classifies students into two categories – those who have and those who have not taken a statistics course – which makes this variable categorical.

GUIDED PRACTICE

An experiment is evaluating the effectiveness of a new drug in treating migraines. A `group` variable is used to indicate the experiment group for each patient: treatment or control. The `num_migraines` variable represents the number of migraines the patient experienced during a 3-month period. Classify each variable as either numerical or categorical?⁷

1.2.3 Relationships between variables

Many analyses are motivated by a researcher looking for a relationship between two or more variables. A social scientist may like to answer some of the following questions:

Does a higher-than-average increase in county population tend to correspond to counties with higher or lower median household incomes?

If homeownership in one county is lower than the national average, will the percent of housing units that are in multi-unit structures in that county tend to be above or below the national average?

How much can the median education level explain the median household income for counties in the US?

To answer these questions, data must be collected, such as the `county` dataset shown in Table 1.5. Examining **summary statistics** can provide numerical insights about the specifics of each of these questions. Alternatively, graphs can be used to visually explore the data, potentially providing more insight than a summary statistic.

Scatterplots are one type of graph used to study the relationship between two numerical variables. Figure 1.2 displays the relationship between the variables `homeownership` and `multi_unit`, which is the percent of housing units that are in multi-unit structures (e.g., apartments, condos). Each point on the plot represents a single county. For instance, the highlighted dot corresponds to County 413 in the `county` dataset: Chattahoochee County, Georgia, which has 39.4% of housing units that are in multi-unit structures and a homeownership rate of 31.3%. The scatterplot suggests a relationship between the two variables: counties with a higher rate of housing units that are in multi-unit structures tend to have lower homeownership rates. We might brainstorm as to why this relationship exists and investigate each idea to determine which are the most reasonable explanations.

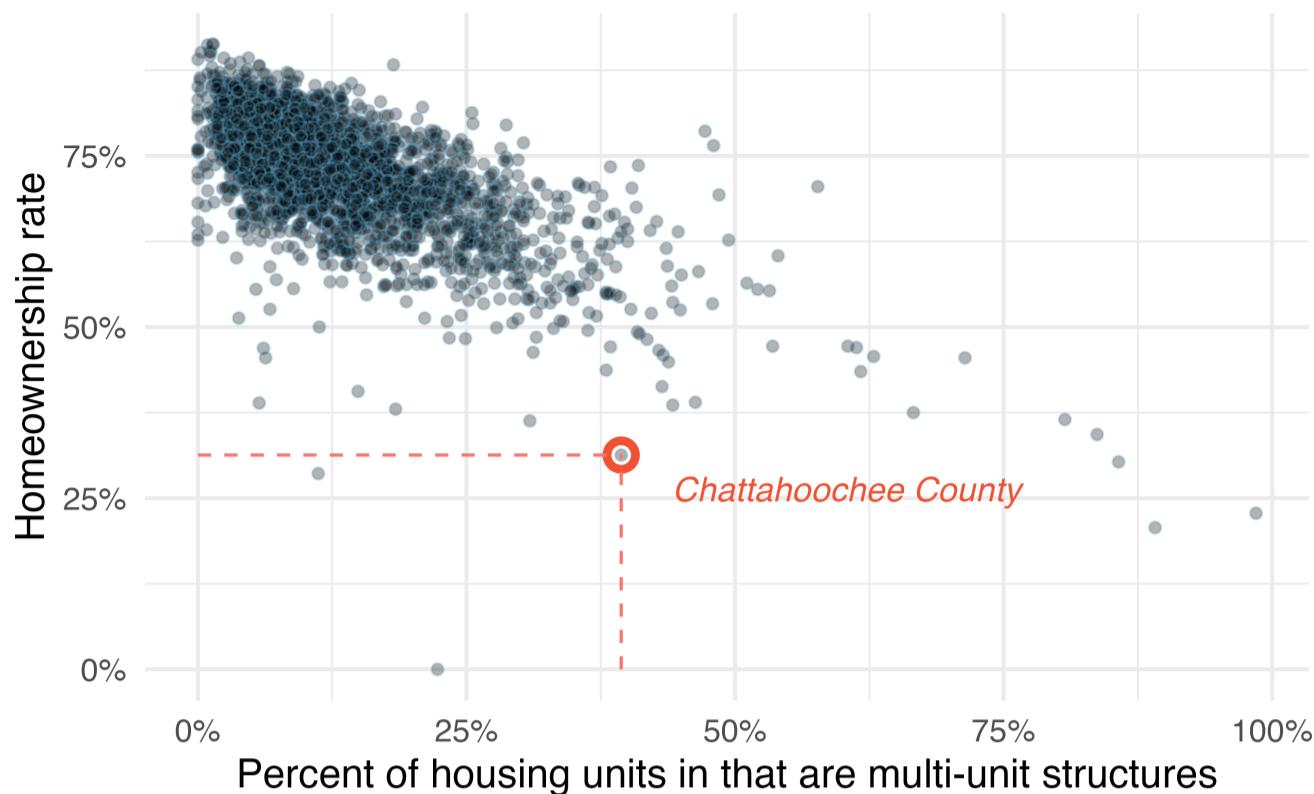


Figure 1.2: A scatterplot of homeownership versus the percent of housing units that are in multi-unit structures for US counties. The highlighted dot represents Chattahoochee County, Georgia, which has a multi-unit rate of 39.4% and a homeownership rate of 31.3%.

The multi-unit and homeownership rates are said to be associated because the plot shows a discernible pattern. When two variables show some connection with one another, they are called **associated** variables.

GUIDED PRACTICE

Examine the variables in the `loan50` dataset, which are described in Table 1.4. Create two questions about possible relationships between variables in `loan50` that are of interest to you.⁸

EXAMPLE:

This example examines the relationship between the percent change in population from 2010 to 2017 and median household income for counties, which is visualized as a scatterplot in Figure 1.3. Are these variables associated?

The larger the median household income for a county, the higher the population growth observed for the county. While it isn't true that every county with a higher median household income has a higher population growth, the trend in the plot is evident. Since there is some relationship between the variables, they are associated.

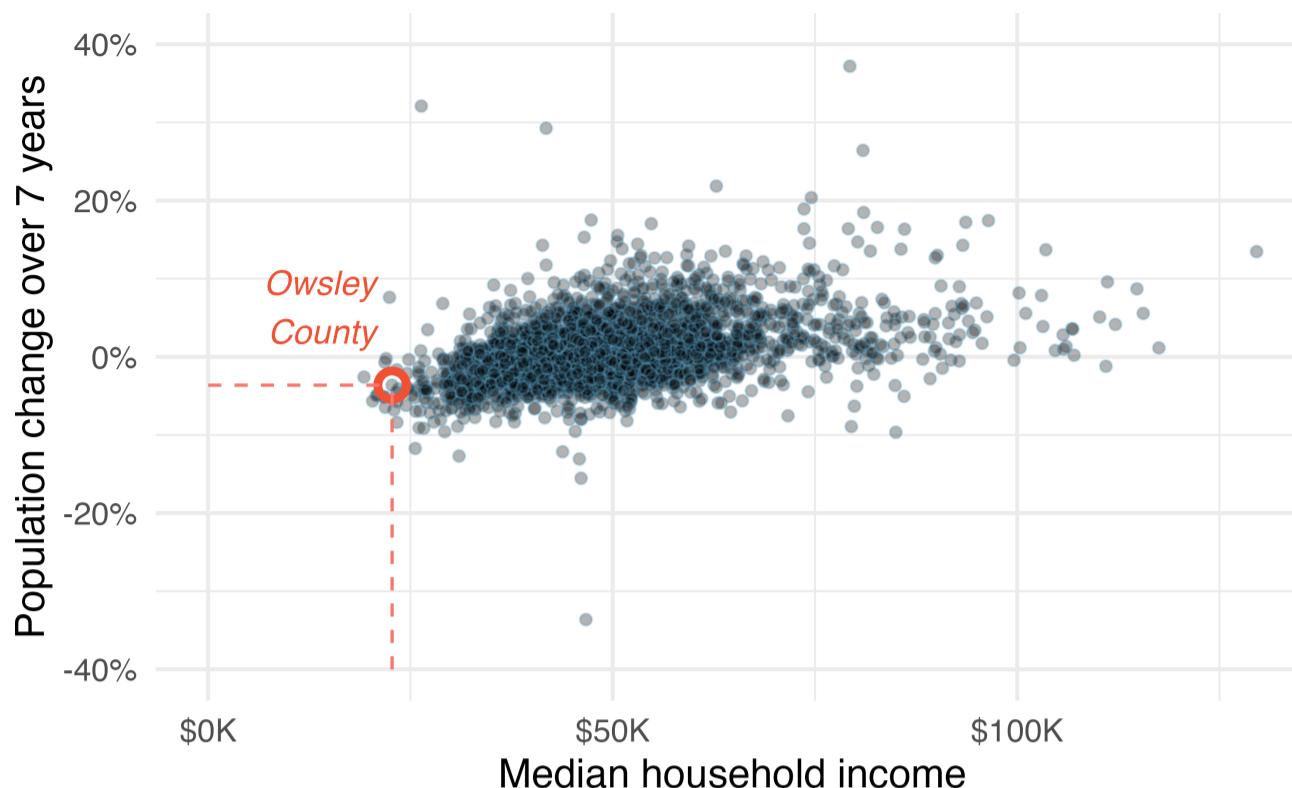


Figure 1.3: A scatterplot showing population change against median household income. Owsley County of Kentucky is highlighted, which lost 3.63% of its population from 2010 to 2017 and had median household income of \$22,736.

Because there is a downward trend in Figure 1.2 – counties with more housing units that are in multi-unit structures are associated with lower homeownership – these variables are said to be **negatively associated**. A **positive association** is shown in the relationship between the `median_hh_income` and `pop_change` variables in Figure 1.3, where counties with higher median household income tend to have higher rates of population growth.

If two variables are not associated, then they are said to be **independent**. That is, two variables are independent if there is no evident relationship between the two.

Associated or independent, not both.

A pair of variables are either related in some way (associated) or not (independent). No pair of variables is both associated and independent.

1.2.4 Explanatory and response variables

When we ask questions about the relationship between two variables, we sometimes also want to determine if the change in one variable causes a change in the other. Consider the following rephrasing of an earlier question about the `county` dataset:

If there is an increase in the median household income in a county, does this drive an increase in its population?

In this question, we are asking whether one variable affects another. If this is our underlying belief, then *median household income* is the **explanatory variable**, and the *population change* is the **response variable** in the hypothesized relationship.⁹

Explanatory and response variables.

When we suspect one variable might causally affect another, we label the first variable the explanatory variable and the second the response variable. We also use the terms **explanatory** and **response** to describe variables where the **response** might be predicted using the **explanatory** even if there is no causal relationship.

explanatory variable → *might affect* → response variable

For many pairs of variables, there is no hypothesized relationship, and these labels would not be applied to either variable in such cases.

Bear in mind that the act of labeling the variables in this way does nothing to guarantee that a causal relationship exists. A formal evaluation to check whether one variable causes a change in another requires an experiment.

1.2.5 Observational studies and experiments

There are two primary types of data collection: experiments and observational studies.

When researchers want to evaluate the effect of particular traits, treatments, or conditions, they conduct an **experiment**. For instance, we may suspect drinking a high-calorie energy drink will improve performance in a race. To check if there really is a causal relationship between the explanatory variable (whether the runner drank an energy drink or not) and the response variable (the race time), researchers identify a sample of individuals and split them into groups. The individuals in each group are *assigned* a treatment. When individuals are randomly assigned to a group, the experiment is called a **randomized experiment**. Random assignment organizes the participants in a study into groups that are roughly equal on all aspects, thus allowing us to control for any confounding variables that might affect the outcome (e.g., fitness level, racing experience, etc.). For example, each runner in the experiment could be randomly assigned, perhaps by flipping a coin, into one of two groups: the first group receives a **placebo** (fake treatment, in this case a no-calorie drink) and the second group receives the high-calorie energy drink. See the case study in Section [1.1](#) for another example of an experiment, though that study did not employ a placebo.

Researchers perform an **observational study** when they collect data in a way that does not directly interfere with how the data arise. For instance, researchers may collect information via surveys, review medical or company records, or follow a **cohort** of many similar individuals to form hypotheses about why certain diseases might develop. In each of these situations, researchers merely observe the data that arise. In general, observational studies can provide evidence of a naturally occurring association between variables, but they cannot by themselves show a causal connection as they do not offer a mechanism for controlling for confounding variables.

Association \neq Causation.

In general, association does not imply causation. An advantage of a randomized experiment is that it is easier to establish causal relationships with such a study. The main reason for this is that observational studies do not control for confounding variables, and hence establishing causal relationships with observational studies requires advanced statistical methods (that are beyond the scope of this book). We will revisit this idea when we discuss experiments later in the book.

1.3 Chapter review

1.3.1 Summary

This chapter introduced you to the world of data. Data can be organized in many ways but tidy data, where each row represents an observation and each column represents a variable, lends itself most easily to statistical analysis. Many of the ideas from this chapter will be seen as we move on to doing full data analyses. In the next chapter you're going to learn about how we can design studies to collect the data we need to make conclusions with the desired scope of inference.

1.3.2 Terms

We introduced the following terms in the chapter. If you're not sure what some of these terms mean, we recommend you go back in the text and review their definitions. We are purposefully presenting them in alphabetical order, instead of in order of appearance, so they will be a little more challenging to locate. However, you should be able to easily spot them as **bolded text**.

associated	experiment	ordinal
case	explanatory variable	placebo
categorical	independent	positive association
cohort	level	randomized experiment
continuous	negative association	response variable
data	nominal	summary statistic
data frame	numerical	tidy data
dependent	observational study	variable
discrete	observational unit	

1.4 Exercises

Answers to odd numbered exercises can be found in Appendix [A.1](#).

1. **Marvel Cinematic Universe films.** The data frame below contains information on Marvel Cinematic Universe films through the Infinity saga (a movie storyline spanning from Ironman in 2008 to Endgame in 2019). Box office totals are given in millions of US Dollars. How many observations and how many variables does this data frame have?¹⁰

Title	Length			Gross		
	Hrs	Mins	Release Date	Opening Wknd US	US	World
1 Iron Man	2	6	5/2/2008	98.62	319.03	585.8
2 The Incredible Hulk	1	52	6/12/2008	55.41	134.81	264.77
3 Iron Man 2	2	4	5/7/2010	128.12	312.43	623.93
4 Thor	1	55	5/6/2011	65.72	181.03	449.33
5 Captain America: The First Avenger	2	4	7/22/2011	65.06	176.65	370.57
...
23 Spiderman: Far from Home	2	9	7/2/2019	92.58	390.53	1131.93

2. **Cherry Blossom Run.** The data frame below contains information on runners in the 2017 Cherry Blossom Run, which is an annual road race that takes place in Washington, DC. Most runners participate in a 10-mile run while a smaller fraction take part in a 5k run or walk. How many observations and how many variables does this data frame have?¹¹

Bib	Name	Sex	Age	City / Country	Time				
					Net	Clock	Pace	Event	
1	6	Hiwot G.	F	21	Ethiopia	3217	3217	321	10 Mile

Bib	Name	Sex	Age	City / Country	Time				Event
					Net	Clock	Pace		
2	22	Buze D.	F	22	Ethiopia	3232	3232	323	10 Mile
3	16	Gladys K.	F	31	Kenya	3276	3276	327	10 Mile
4	4	Mamitu D.	F	33	Ethiopia	3285	3285	328	10 Mile
5	20	Karolina N.	F	35	Poland	3288	3288	328	10 Mile
...
19961	25153	Andres E.	M	33	Woodbridge, VA	5287	5334	1700	5K

3. Air pollution and birth outcomes, study components. Researchers collected data to examine the relationship between air pollutants and preterm births in Southern California. During the study air pollution levels were measured by air quality monitoring stations. Specifically, levels of carbon monoxide were recorded in parts per million, nitrogen dioxide and ozone in parts per hundred million, and coarse particulate matter (PM_{10}) in $\mu g/m^3$. Length of gestation data were collected on 143,196 births between the years 1989 and 1993, and air pollution exposure during gestation was calculated for each birth. The analysis suggested that increased ambient PM_{10} and, to a lesser degree, CO concentrations may be associated with the occurrence of preterm births. ([Ritz et al. 2000](#))

- a. Identify the main research question of the study.
- b. Who are the subjects in this study, and how many are included?
- c. What are the variables in the study? Identify each variable as numerical or categorical. If numerical, state whether the variable is discrete or continuous. If categorical, state whether the variable is ordinal.

4. Cheaters, study components. Researchers studying the relationship between honesty, age and self-control conducted an experiment on 160 children between the ages of 5 and 15. Participants reported their age, sex, and whether they were an only child or not. The researchers asked each child to toss a fair coin in private and to record the outcome (white or black) on a paper sheet and said they would only reward children who report white. ([Bucciol and Piovesan 2011](#))

- a. Identify the main research question of the study.
- b. Who are the subjects in this study, and how many are included?
- c. The study's findings can be summarized as follows: "*Half the students were explicitly told not to cheat, and the others were not given any explicit instructions. In the no instruction group probability of cheating was found to be uniform across groups based on child's characteristics. In the group that was explicitly told to not cheat, girls were less likely to cheat, and while rate of cheating didn't vary by age for boys, it decreased with age for girls.*" How many variables were recorded for each subject in the study in order to conclude these findings? State the variables and their types.

5. Gamification and statistics, study components. Gamification is the application of game-design elements and game principles in non-game contexts. In educational settings, gamification is often implemented as educational activities to solve problems by using characteristics of game elements. Researchers investigating the effects of gamification on learning statistics conducted a study where they split college students in a statistics class into four groups: (1) no reading exercises and no gamification, (2) reading exercises but no gamification, (3) gamification but no reading exercises, and (4) gamification and reading exercises. Students in all groups also attended lectures. Students in the class were from two majors: Electrical and Computer Engineering (n = 279) and Business Administration (n = 86). After their assigned learning experience, each student took a final evaluation comprised of 30 multiple choice questions and their score was measured as the number of questions they answered correctly. The researchers considered students' gender, level of studies (first through

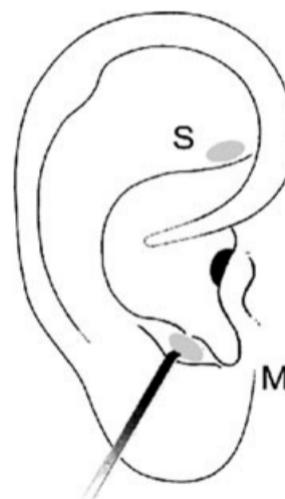
fourth year) and academic major. Other variables considered were expertise in the English language and use of personal computers and games, both of which were measured on a scale of 1 (beginner) to 5 (proficient). The study found that gamification had a positive effect on student learning compared to traditional teaching methods involving lectures and reading exercises. They also found that the effect was larger for females and Engineering students. ([Legaki et al. 2020](#))

- a. Identify the main research question of the study.
- b. Who were the subjects in this study, and how many were included?
- c. What are the variables in the study? Identify each variable as numerical or categorical. If numerical, state whether the variable is discrete or continuous. If categorical, state whether the variable is ordinal.

6. Stealers, study components. In a study of the relationship between socio-economic class and unethical behavior, 129 University of California undergraduates at Berkeley were asked to identify themselves as having low or high social class by comparing themselves to others with the most (least) money, most (least) education, and most (least) respected jobs. They were also presented with a jar of individually wrapped candies and informed that the candies were for children in a nearby laboratory, but that they could take some if they wanted. After completing some unrelated tasks, participants reported the number of candies they had taken. ([Piff et al. 2012](#))

- a. Identify the main research question of the study.
- b. Who were the subjects in this study, and how many were included?
- c. The study found that students who were identified as upper-class took more candy than others. How many variables were recorded for each subject in the study in order to conclude these findings? State the variables and their types.

7. Migraine and acupuncture. A migraine is a particularly painful type of headache, which patients sometimes wish to treat with acupuncture. To determine whether acupuncture relieves migraine pain, researchers conducted a randomized controlled study where 89 individuals who identified as female diagnosed with migraine headaches were randomly assigned to one of two groups: treatment or control. Forty-three (43) patients in the treatment group received acupuncture that is specifically designed to treat migraines. Forty-six (46) patients in the control group received placebo acupuncture (needle insertion at non-acupoint locations). Twenty-four (24) hours after patients received acupuncture, they were asked if they were pain free. Results are summarized in the contingency table below. Also provided is a figure from the original paper displaying the appropriate area (M) versus the inappropriate area (S) used in the treatment of migraine attacks.¹² ([Allais et al. 2011](#))



Group	Pain free?	
	No	Yes
Control	44	2
Treatment	33	10

- a. What percent of patients in the treatment group were pain free 24 hours after receiving acupuncture?
- b. What percent were pain free in the control group?
- c. In which group did a higher percent of patients become pain free 24 hours after receiving acupuncture?

d. Your findings so far might suggest that acupuncture is an effective treatment for migraines for all people who suffer from migraines. However, this is not the only possible conclusion. What is one other possible explanation for the observed difference between the percentages of patients that are pain free 24 hours after receiving acupuncture in the two groups?

e. What are the explanatory and response variables in this study?

8. **Sinusitis and antibiotics.** Researchers studying the effect of antibiotic treatment for acute sinusitis compared to symptomatic treatments randomly assigned 166 adults diagnosed with acute sinusitis to one of two groups: treatment or control. Study participants received either a 10-day course of amoxicillin (an antibiotic) or a placebo similar in appearance and taste. The placebo consisted of symptomatic treatments such as acetaminophen, nasal decongestants, etc. At the end of the 10-day period, patients were asked if they experienced improvement in symptoms. The distribution of responses is summarized below.¹³ ([Garbutt et al. 2012](#))

Improvement		
Group	No	Yes
Control	16	65
Treatment	19	66

a. What percent of patients in the treatment group experienced improvement in symptoms?

b. What percent experienced improvement in symptoms in the control group?

c. In which group did a higher percentage of patients experience improvement in symptoms?

d. Your findings so far might suggest a real difference in the effectiveness of antibiotic and placebo treatments for improving symptoms of sinusitis. However, this is not the only possible conclusion. What is one other possible explanation for the observed difference between the percentages of patients who experienced improvement in symptoms?

e. What are the explanatory and response variables in this study?

9. **Daycare fines, study components.** Researchers tested the deterrence hypothesis which predicts that the introduction of a penalty will reduce the occurrence of the behavior subject to the fine, with the condition that the fine leaves everything else unchanged by instituting a fine for late pickup at daycare centers. For this study, they worked with 10 volunteer daycare centers that did not originally impose a fine to parents for picking up their kids late. They randomly selected 6 of these daycare centers and instituted a monetary fine (of a considerable amount) for picking up children late and then removed it. In the remaining 4 daycare centers no fine was introduced. The study period was divided into four: before the fine (weeks 1–4), the first 4 weeks with the fine (weeks 5–8), the last 8 weeks with fine (weeks 9–16), and the after fine period (weeks 17–20). Throughout the study, the number of kids who were picked up late was recorded each week for each daycare. The study found that the number of late-coming parents increased significantly when the fine was introduced, and no reduction occurred after the fine was removed.¹⁴ ([Gneezy and Rustichini 2000](#))

center	week	group	late_pickups	study_period
1	1	test	8	before fine
1	2	test	8	before fine
1	3	test	7	before fine
1	4	test	6	before fine
1	5	test	8	first 4 weeks with fine
...

center week group late_pickups study_period

1	1	test	8	before fine
1	2	test	8	before fine
1	3	test	7	before fine
1	4	test	6	before fine
1	5	test	8	first 4 weeks with fine
1	6	test	9	first 4 weeks with fine
1	7	test	9	first 4 weeks with fine
1	8	test	12	first 4 weeks with fine
1	9	test	13	last 8 weeks with fine
1	10	test	13	last 8 weeks with fine
1	11	test	15	last 8 weeks with fine
1	12	test	13	last 8 weeks with fine
1	13	test	14	last 8 weeks with fine
1	14	test	16	last 8 weeks with fine
1	15	test	14	last 8 weeks with fine
1	16	test	15	last 8 weeks with fine
1	17	test	16	after fine
1	18	test	13	after fine
1	19	test	15	after fine
1	20	test	17	after fine
2	1	test	6	before fine
2	2	test	7	before fine
2	3	test	3	before fine
2	4	test	5	before fine
2	5	test	2	first 4 weeks with fine
2	6	test	11	first 4 weeks with fine
2	7	test	14	first 4 weeks with fine
2	8	test	9	first 4 weeks with fine
2	9	test	16	last 8 weeks with fine

center	week	group	late_pickups	study_period
2	10	test	12	last 8 weeks with fine
2	11	test	10	last 8 weeks with fine
2	12	test	14	last 8 weeks with fine
2	13	test	14	last 8 weeks with fine
2	14	test	16	last 8 weeks with fine
2	15	test	12	last 8 weeks with fine
2	16	test	17	last 8 weeks with fine
2	17	test	14	after fine
2	18	test	10	after fine
2	19	test	14	after fine
2	20	test	15	after fine
3	1	test	8	before fine
3	2	test	9	before fine
3	3	test	8	before fine
3	4	test	9	before fine
3	5	test	3	first 4 weeks with fine
3	6	test	5	first 4 weeks with fine
3	7	test	15	first 4 weeks with fine
3	8	test	18	first 4 weeks with fine
3	9	test	16	last 8 weeks with fine
3	10	test	14	last 8 weeks with fine
3	11	test	20	last 8 weeks with fine
3	12	test	18	last 8 weeks with fine
3	13	test	25	last 8 weeks with fine
3	14	test	22	last 8 weeks with fine
3	15	test	27	last 8 weeks with fine
3	16	test	19	last 8 weeks with fine
3	17	test	20	after fine
3	18	test	23	after fine

center week group late_pickups study_period

3	19	test	23	after fine
3	20	test	22	after fine
4	1	test	10	before fine
4	2	test	3	before fine
4	3	test	14	before fine
4	4	test	9	before fine
4	5	test	6	first 4 weeks with fine
4	6	test	24	first 4 weeks with fine
4	7	test	8	first 4 weeks with fine
4	8	test	22	first 4 weeks with fine
4	9	test	22	last 8 weeks with fine
4	10	test	19	last 8 weeks with fine
4	11	test	25	last 8 weeks with fine
4	12	test	18	last 8 weeks with fine
4	13	test	23	last 8 weeks with fine
4	14	test	22	last 8 weeks with fine
4	15	test	24	last 8 weeks with fine
4	16	test	17	last 8 weeks with fine
4	17	test	15	after fine
4	18	test	23	after fine
4	19	test	25	after fine
4	20	test	18	after fine
5	1	test	13	before fine
5	2	test	12	before fine
5	3	test	9	before fine
5	4	test	13	before fine
5	5	test	15	first 4 weeks with fine
5	6	test	10	first 4 weeks with fine
5	7	test	27	first 4 weeks with fine

center	week	group	late_pickups	study_period
5	8	test	28	first 4 weeks with fine
5	9	test	35	last 8 weeks with fine
5	10	test	10	last 8 weeks with fine
5	11	test	24	last 8 weeks with fine
5	12	test	32	last 8 weeks with fine
5	13	test	29	last 8 weeks with fine
5	14	test	29	last 8 weeks with fine
5	15	test	26	last 8 weeks with fine
5	16	test	31	last 8 weeks with fine
5	17	test	26	after fine
5	18	test	35	after fine
5	19	test	29	after fine
5	20	test	28	after fine
6	1	test	5	before fine
6	2	test	8	before fine
6	3	test	7	before fine
6	4	test	5	before fine
6	5	test	5	first 4 weeks with fine
6	6	test	9	first 4 weeks with fine
6	7	test	12	first 4 weeks with fine
6	8	test	14	first 4 weeks with fine
6	9	test	19	last 8 weeks with fine
6	10	test	17	last 8 weeks with fine
6	11	test	14	last 8 weeks with fine
6	12	test	13	last 8 weeks with fine
6	13	test	10	last 8 weeks with fine
6	14	test	15	last 8 weeks with fine
6	15	test	14	last 8 weeks with fine
6	16	test	16	last 8 weeks with fine

center week group late_pickups study_period

6	17	test	6	after fine
6	18	test	12	after fine
6	19	test	17	after fine
6	20	test	13	after fine
7	1	control	7	before fine
7	2	control	10	before fine
7	3	control	12	before fine
7	4	control	6	before fine
7	5	control	4	first 4 weeks with fine
7	6	control	13	first 4 weeks with fine
7	7	control	7	first 4 weeks with fine
7	8	control	8	first 4 weeks with fine
7	9	control	5	last 8 weeks with fine
7	10	control	12	last 8 weeks with fine
7	11	control	3	last 8 weeks with fine
7	12	control	5	last 8 weeks with fine
7	13	control	6	last 8 weeks with fine
7	14	control	13	last 8 weeks with fine
7	15	control	7	last 8 weeks with fine
7	16	control	4	last 8 weeks with fine
7	17	control	7	after fine
7	18	control	10	after fine
7	19	control	4	after fine
7	20	control	6	after fine
8	1	control	12	before fine
8	2	control	9	before fine
8	3	control	14	before fine
8	4	control	18	before fine
8	5	control	10	first 4 weeks with fine

center	week	group	late_pickups	study_period
8	6	control	11	first 4 weeks with fine
8	7	control	6	first 4 weeks with fine
8	8	control	15	first 4 weeks with fine
8	9	control	14	last 8 weeks with fine
8	10	control	13	last 8 weeks with fine
8	11	control	7	last 8 weeks with fine
8	12	control	12	last 8 weeks with fine
8	13	control	9	last 8 weeks with fine
8	14	control	9	last 8 weeks with fine
8	15	control	17	last 8 weeks with fine
8	16	control	8	last 8 weeks with fine
8	17	control	5	after fine
8	18	control	11	after fine
8	19	control	8	after fine
8	20	control	13	after fine
9	1	control	3	before fine
9	2	control	4	before fine
9	3	control	9	before fine
9	4	control	3	before fine
9	5	control	3	first 4 weeks with fine
9	6	control	5	first 4 weeks with fine
9	7	control	9	first 4 weeks with fine
9	8	control	5	first 4 weeks with fine
9	9	control	2	last 8 weeks with fine
9	10	control	7	last 8 weeks with fine
9	11	control	6	last 8 weeks with fine
9	12	control	6	last 8 weeks with fine
9	13	control	9	last 8 weeks with fine
9	14	control	4	last 8 weeks with fine

center	week	group	late_pickups	study_period
9	15	control	9	last 8 weeks with fine
9	16	control	2	last 8 weeks with fine
9	17	control	3	after fine
9	18	control	8	after fine
9	19	control	3	after fine
9	20	control	5	after fine
10	1	control	15	before fine
10	2	control	13	before fine
10	3	control	13	before fine
10	4	control	12	before fine
10	5	control	10	first 4 weeks with fine
10	6	control	9	first 4 weeks with fine
10	7	control	15	first 4 weeks with fine
10	8	control	15	first 4 weeks with fine
10	9	control	15	last 8 weeks with fine
10	10	control	10	last 8 weeks with fine
10	11	control	17	last 8 weeks with fine
10	12	control	12	last 8 weeks with fine
10	13	control	13	last 8 weeks with fine
10	14	control	11	last 8 weeks with fine
10	15	control	14	last 8 weeks with fine
10	16	control	17	last 8 weeks with fine
10	17	control	12	after fine
10	18	control	9	after fine
10	19	control	15	after fine
10	20	control	13	after fine

a. Is this an observational study or an experiment? Explain your reasoning.

b. What are the cases in this study and how many are included?

c. What is the response variable in the study and what type of variable is it?

d. What are the explanatory variables in the study and what types of variables are they?

10. **Efficacy of COVID-19 vaccine on adolescents, study components.** Results of a Phase 3 trial announced in March 2021 show that the Pfizer-BioNTech COVID-19 vaccine demonstrated 100% efficacy and robust antibody responses on 12 to 15 years old adolescents with or without prior evidence of SARS-CoV-2 infection. In this trial 2,260 adolescents were randomly assigned to two groups: one group got the vaccine ($n = 1,131$) and the other got a placebo ($n = 1,129$). While 18 cases of COVID-19 were observed in the placebo group, none were observed in the vaccine group.¹⁵ (Pfizer 2021)

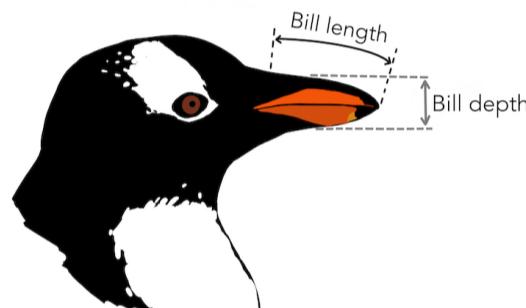
a. Is this an observational study or an experiment? Explain your reasoning.

b. What are the cases in this study and how many are included?

c. What is the response variable in the study and what type of variable is it?

d. What are the explanatory variables in the study and what types of variables are they?

11. **Palmer penguins.** Data were collected on 344 penguins living on three islands (Torgersen, Biscoe, and Dream) in the Palmer Archipelago, Antarctica. In addition to which island each penguin lives on, the data contains information on the species of the penguin (Adelie, Chinstrap, or Gentoo), its bill length, bill depth, and flipper length (measured in millimeters), its body mass (measured in grams), and the sex of the penguin (female or male).¹⁶ Bill length and depth are measured as shown in the image.¹⁷ (Gorman, Williams, and Fraser 2014a)



a. How many cases were included in the data?

b. How many numerical variables are included in the data? Indicate what they are, and if they are continuous or discrete.

c. How many categorical variables are included in the data, and what are they? List the corresponding levels (categories) for each.

12. **Smoking habits of UK residents.** A survey was conducted to study the smoking habits of 1,691 UK residents. Below is a data frame displaying a portion of the data collected in this survey. A blank cell indicates that data for that variable was not available for a given respondent.¹⁸

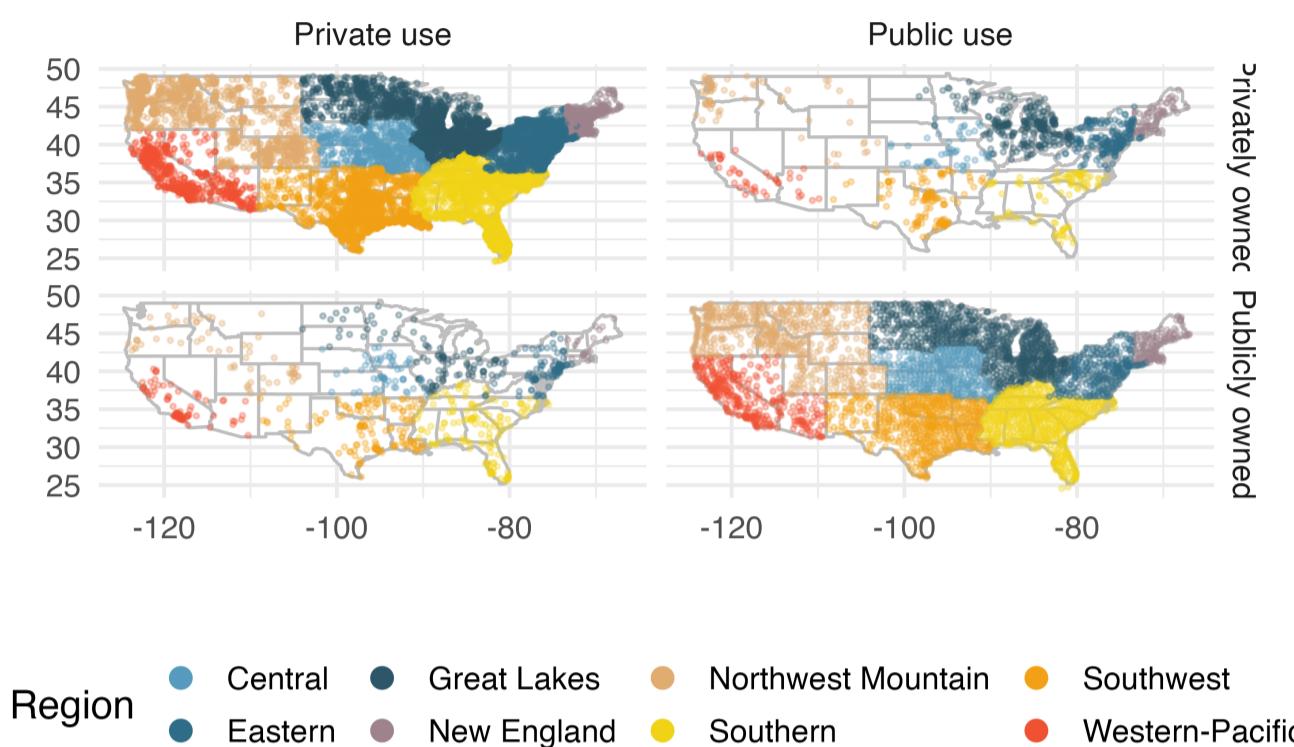
						amount		
	sex	age	marital_status	gross_income	smoke	weekend	weekday	
1	Female	61	Married	2,600 to 5,200	No			
2	Female	61	Divorced	10,400 to 15,600	Yes	5	4	
3	Female	69	Widowed	5,200 to 10,400	No			
4	Female	50	Married	5,200 to 10,400	No			
5	Male	31	Single	10,400 to 15,600	Yes	10	20	
...	
1691	Male	49	Divorced	Above 36,400	Yes	15	10	

a. What does each row of the data frame represent?

b. How many participants were included in the survey?

c. Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

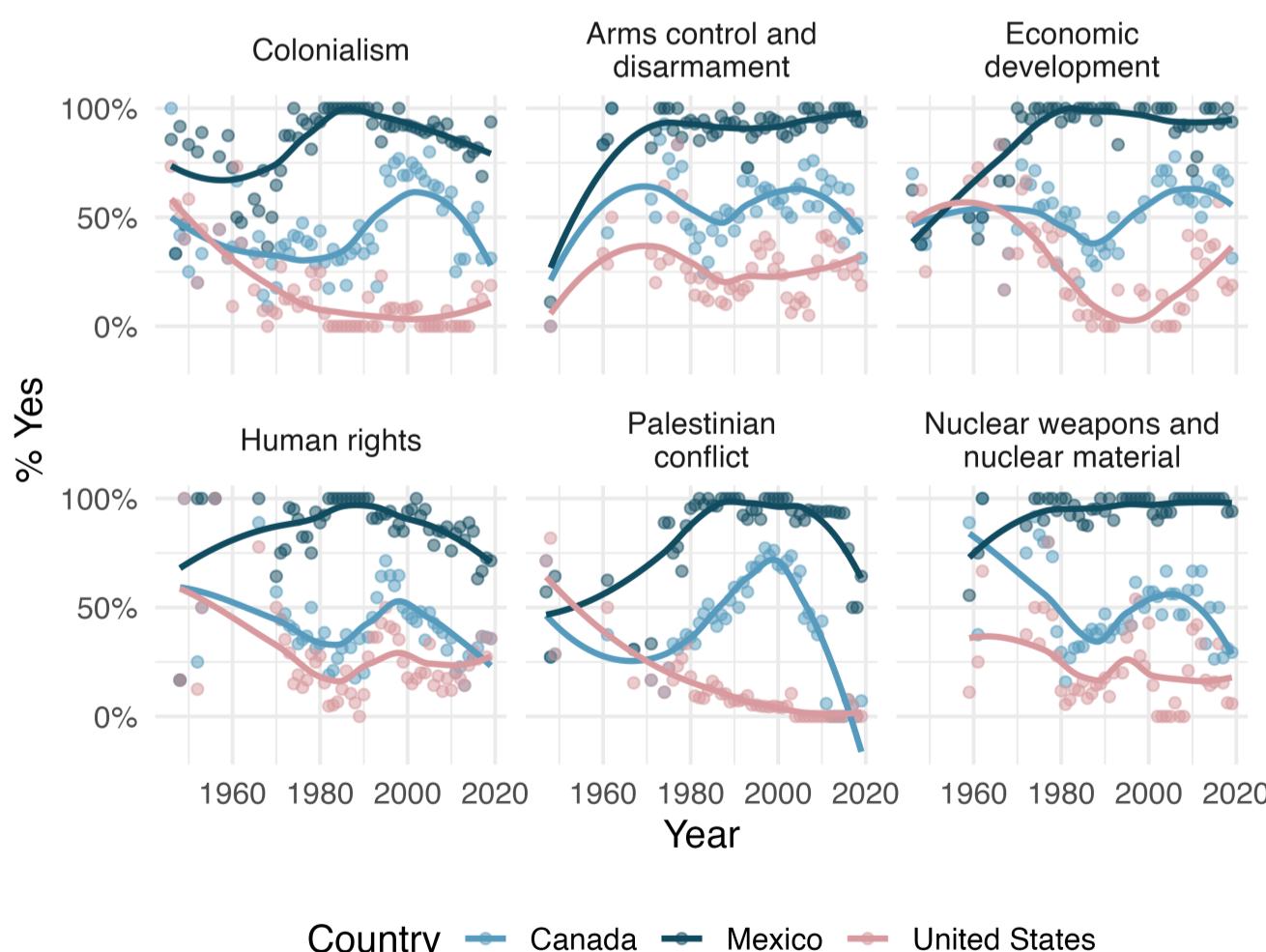
13. **US Airports.** The visualization below shows the geographical distribution of airports in the contiguous United States and Washington, DC. This visualization was constructed based on a dataset where each observation is an airport.¹⁹



a. List the variables you believe were necessary to create this visualization.

b. Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

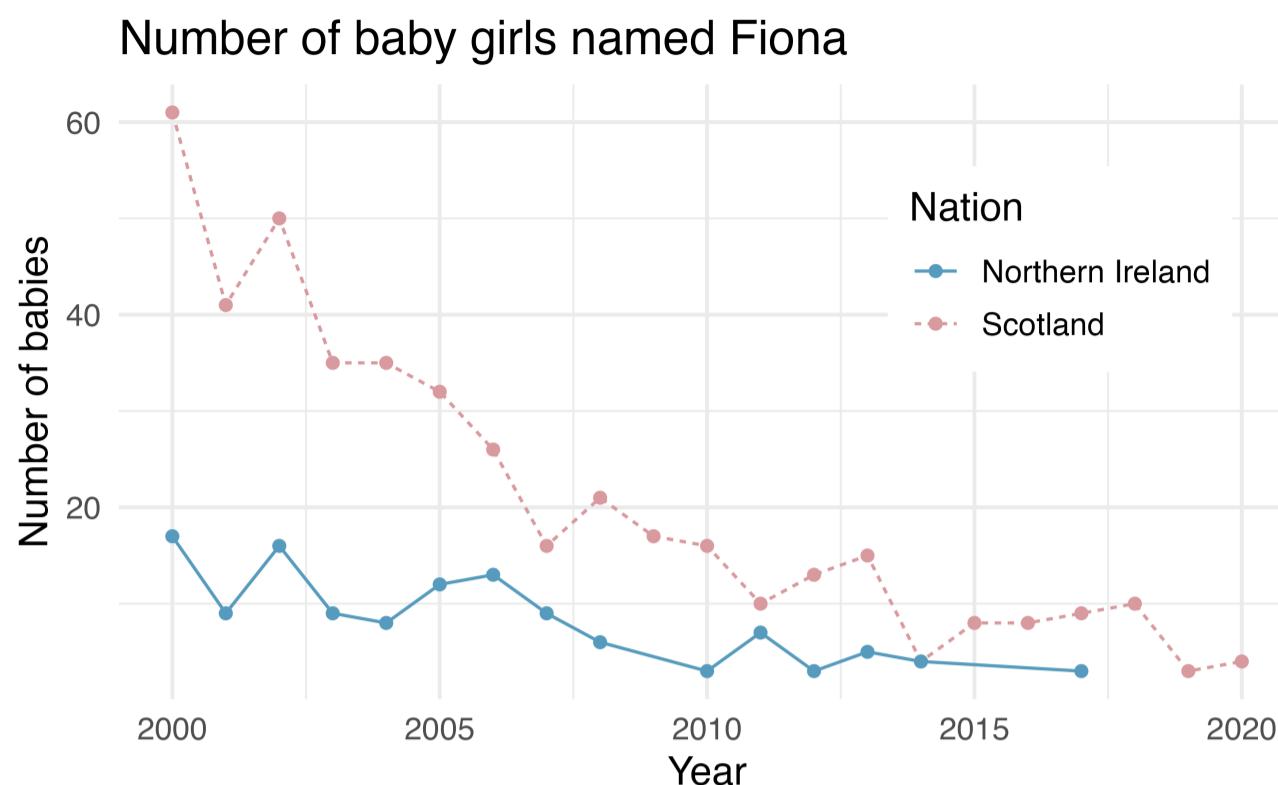
14. **UN Votes.** The visualization below shows voting patterns in the United States, Canada, and Mexico in the United Nations General Assembly on a variety of issues. Specifically, for a given year between 1946 and 2019, it displays the percentage of roll calls in which the country voted yes for each issue. This visualization was constructed based on a dataset where each observation is a country/year pair.²⁰



a. List the variables used in creating this visualization.

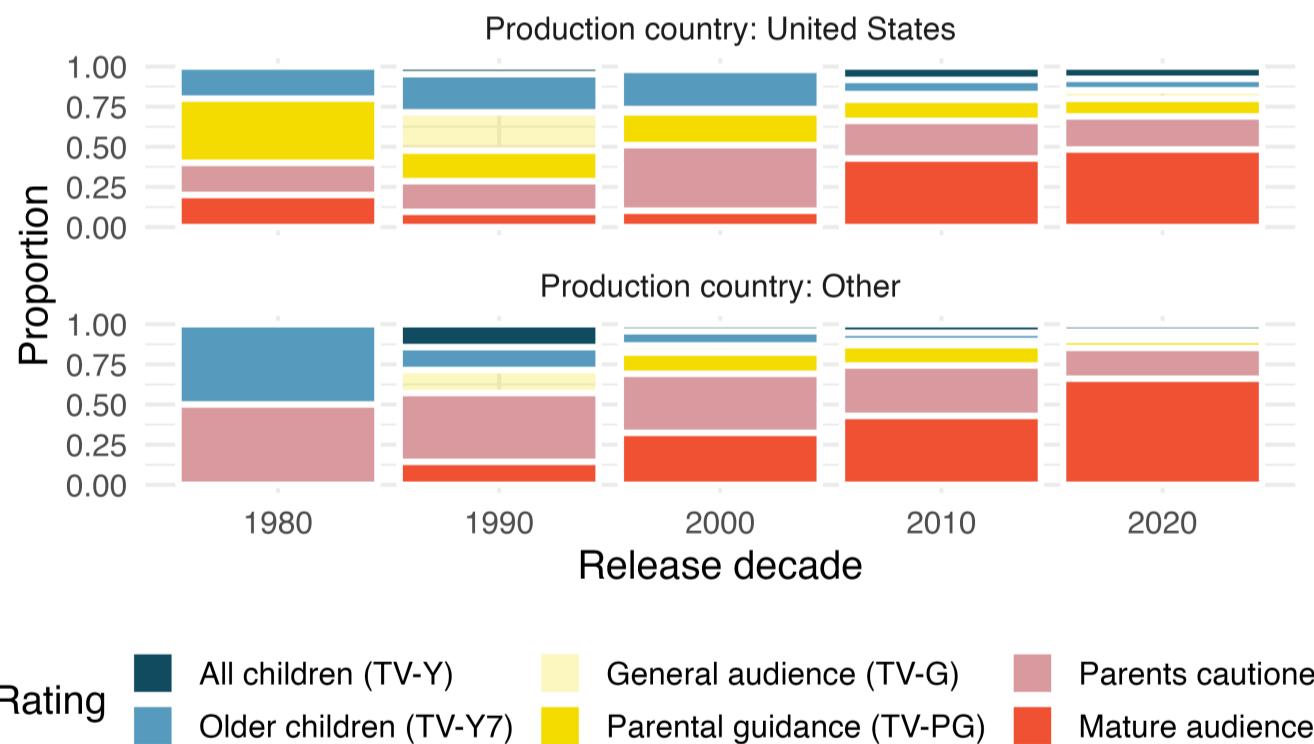
b. Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

15. **UK baby names.** The visualization below shows the number of baby girls born in the United Kingdom (comprised of England & Wales, Northern Ireland, and Scotland) who were given the name "Fiona" over the years.²¹



- List the variables you believe were necessary to create this visualization.
- Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

16. **Shows on Netflix.** The visualization below shows the distribution of ratings of TV shows on Netflix (a streaming entertainment service) based on the decade they were released in and the country they were produced in. In the dataset, each observation is a TV show.²²



- List the variables you believe were necessary to create this visualization.
- Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

17. **Stanford Open Policing.** The Stanford Open Policing project gathers, analyzes, and releases records from traffic stops by law enforcement agencies across the United States. Their goal is to help researchers, journalists, and policy makers investigate and improve interactions between police and the public. The following is an excerpt from a summary table created based off the data collected as part of this project. ([Pierson et al. 2020](#))

County	State	Race / Ethnicity	Driver	Car	
			Arrest rate	Stops / year	Search rate
Apache County	AZ	Black	0.016	266	0.077
Apache County	AZ	Hispanic	0.018	1008	0.053
Apache County	AZ	White	0.006	6322	0.017
Cochise County	AZ	Black	0.015	1169	0.047
Cochise County	AZ	Hispanic	0.01	9453	0.037
Cochise County	AZ	White	0.008	10826	0.024
...
Wood County	WI	Black	0.098	16	0.244
Wood County	WI	Hispanic	0.029	27	0.036
Wood County	WI	White	0.029	1157	0.033

- a. What variables were collected on each individual traffic stop in order to create the summary table above?
- b. State whether each variable is numerical or categorical. If numerical, state whether it is continuous or discrete. If categorical, state whether it is ordinal or not.
- c. Suppose we wanted to evaluate whether vehicle search rates are different for drivers of different races. In this analysis, which variable would be the response variable and which variable would be the explanatory variable?

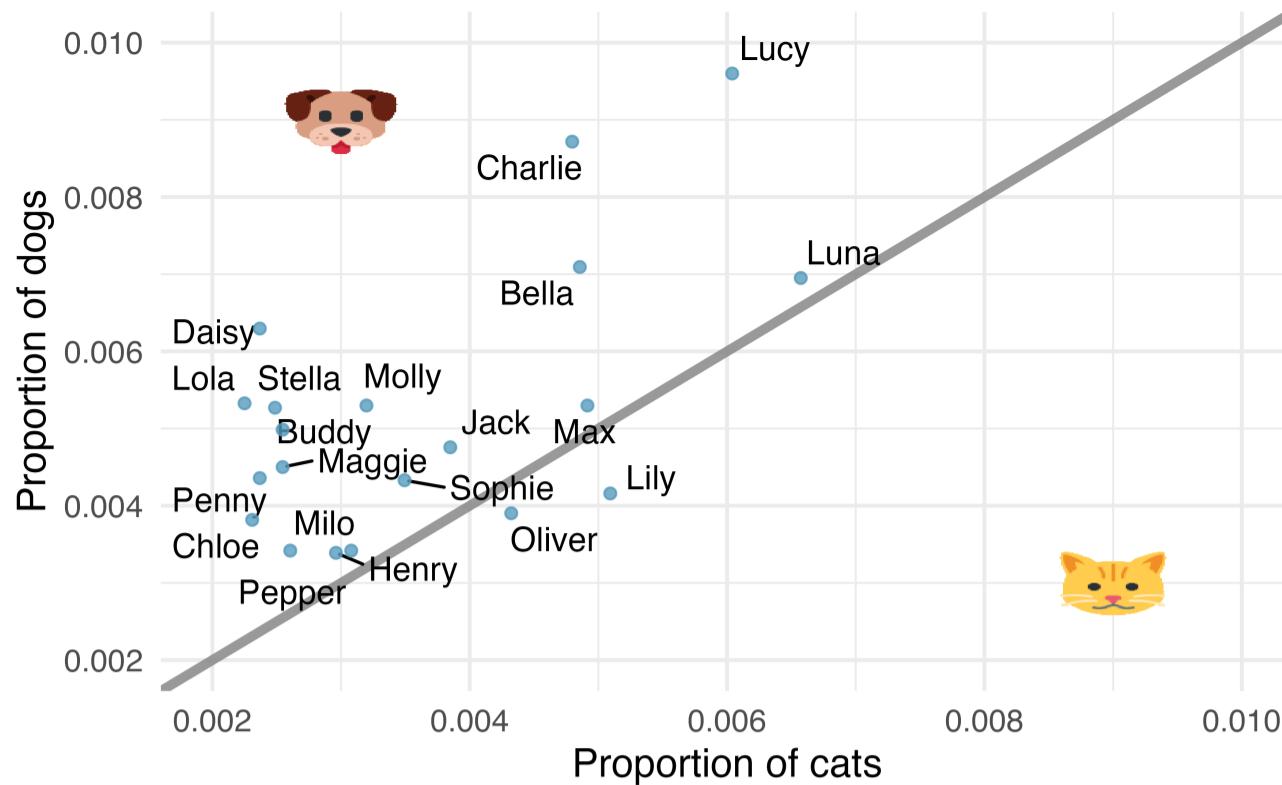
18. **Space launches.** The following summary table shows the number of space launches in the US by the type of launching agency and the outcome of the launch (success or failure).²³

	1957 - 1999		2000-2018	
	Failure	Success	Failure	Success
			Private	State
Private	13	295	10	562
State	281	3751	33	711
Startup	0	0	5	65

- a. What variables were collected on each launch in order to create to the summary table above?
- b. State whether each variable is numerical or categorical. If numerical, state whether it is continuous or discrete. If categorical, state whether it is ordinal or not.
- c. Suppose we wanted to study how the success rate of launches vary between launching agencies and over time. In this analysis, which variable would be the response variable and which variable would be the explanatory variable?

19. **Pet names.** The city of Seattle, WA has an open data portal that includes pets registered in the city.

For each registered pet, we have information on the pet's name and species. The following visualization plots the proportion of dogs with a given name versus the proportion of cats with the same name. The 20 most common cat and dog names are displayed. The diagonal line on the plot is the $x = y$ line; if a name appeared on this line, the name's popularity would be exactly the same for dogs and cats.²⁴



- a. Are these data collected as part of an experiment or an observational study?
- b. What is the most common dog name? What is the most common cat name?
- c. What names are more common for cats than dogs?
- d. Is the relationship between the two variables positive or negative? What does this mean in context of the data?

20. **Stressed out in an elevator.** In a study evaluating the relationship between stress and muscle cramps, half the subjects are randomly assigned to be exposed to increased stress by being placed into an elevator that falls rapidly and stops abruptly and the other half are left at no or baseline stress.

- a. What type of study is this?
- b. Can this study be used to conclude a causal relationship between increased stress and muscle cramps?

[« Preface](#)

[2 Study design »](#)