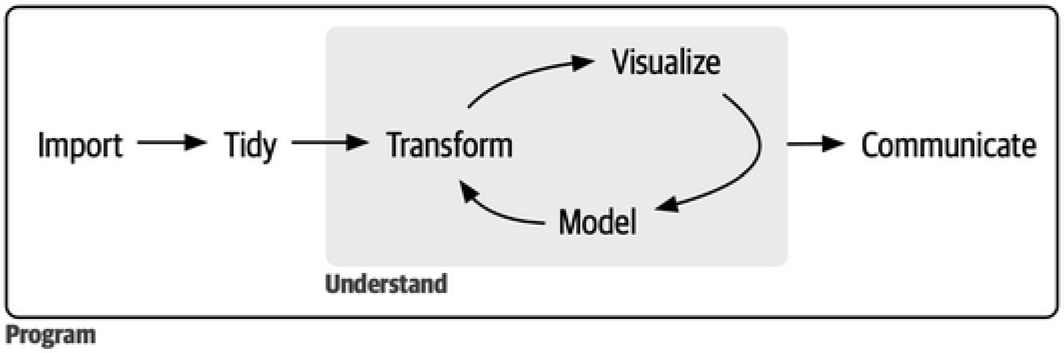
Chapter 1: What is Data Science?



Overview: Statistics is the science of data. In this Chapter, we learn some introductory terms related to data and we look at some examples to help illustrate these terms. We will also begin to formulate questions about finding data which we will eventually answer later on in the course.

Motivating Example: More so now than ever, the world is paying attention to statistics. Daily updates on COVID-19 can include number of new infections, positive test rates, mortality rates, vaccination rates, as well as plots of an infection rate curve which we are trying to flatten. Depending on the source, the information being given can paint very different pictures. Are all of these sources equally reliable? Is the data accurate? Is the data being manipulated in a misleading way?

Whenever confronted with data, one of the first and most important questions that should be answered is: Where does the data come from?

Definitions: We begin by defining **individuals** and **variables**:

- **Individuals or observations**

  For example:

- A **variable**

  For example:

Practice Question: A Starbucks employee decided to collect information about several of their menu items. Consider the resulting data set:

Menu Item	Price	Weight (in g)	Drink? (yes or no)
Cappuccino	$4.75	473	yes
Banana Bread	$3.45	115	no
Dragon Drink	$5.15	473	yes

(a) What are the individuals in this data set? Select all that apply.

(A) Menu Items (B) Price (C) Weight (D) Whether it is a drink

(b) What are the variables in this data set? Select all that apply.

(A) Menu Items (B) Price (C) Weight (D) Whether it is a drink

Definitions: We can classify variables into different types:

- A **categorical variable or an ordinal variable**

  For Example: From the Starbucks data set above,

- A **numerical variable** which is sometimes referred to as a **quantitative variable**

  For Example: From the Starbucks data set above, Question: How do we decide what variables to collect in our data?

Answer: It depends on the question that you are trying to answer using the data.

Example: When trying to judge the recycling habits of a neighbourhood, researchers went around weighing the recycling bins of each household. They found that certain streets (with more expensive houses) had heavier recycling bins and other streets (with less expensive houses) had lighter recycling bins. Does this mean that the more expensive your house the more you recycle?

The individuals:

The variable(s):

Type of variable(s):

Problem with the variable(s):

Definitions: We next define the difference between a **population** and a **sample**.

- A **population** in a statistical study is

- A **sample**

Examples:

(a) You wish to measure the mean weight of 5 month old Koala bears so you find twenty 5-month old Koala bears and weigh them.

Population:

Sample:

(b) You decide to look at all of the uncollected Stat 123 midterms from the Spring 2019 semester. Using this you make a report of the average number of mistakes a Stat 123 student in Spring 2019 made on midterms.

Population:

Sample:

(c) You want to determine the number of wine drinkers in Victoria BC so you loiter inside the Liquor Plus store located at Quadra and MacKenzie on Monday January 11th from 3pm-5pm and count the number of people who buy wine at the store.

Population:

Sample:

Note: Not all samples are good samples. We will discuss this idea further in later sections but if you choose a bad sample, it is not going to accurately describe the population.

Definitions: We now define the different types of studies you can run in order to collect data:

- An **observational study**

- One type of observational study is called a **sample survey**

- A second type of observational study called a **census**

- An **experiment**

Examples:

(a) You are trying to determine the proportion of red cars on the road. You decide to count the number of cars that drive by your house and note how many of them are red.

Type of Study:

(b) You want to determine if having reading break helps students do better on tests so you schedule one test before reading break and one test immediately after reading break and compare the grades on these tests.

Type of Study:

(c) In a (hypothetical) study, you recruited 10 participants. First you measured their alertness (assuming we can) scores. Then you divided the participants into 2 groups of five. For the first group, you gave them a pill that contains some newly developed chemicals to improve alertness. For the second group, you gave them a pill that contains only starch. You waited 15 minutes. Now you measured their alertness again.

Type of Study:

Bonus: What are you studying?

Caution: