Data Types and Quality

Introduction to Data Types and Quality

3 min

We chose it because we deal with huge amounts of data. Besides, it sounds really cool.

– *Larry Page, Co-founder of Google*

The founders of Google were playing with the mathematical term, “googol”, which is a 1 with 100 zeros after it (a number so big that it is pretty much incomprehensible to people). And Google knew they were working with an incomprehensible amount of data.

But what does all of that data look like? And what does it mean to work with a dataset?

Data can mean a lot of things, but within data science, it typically means a collection of organized observations.

There are two types of organization: methodology and shape.

The methodology is how the data was collected. We will get more into that later in this lesson.

The most common shape for data is a spreadsheet or table. The things we are measuring (variables) are in the columns, and the individual instances (observations) are in the rows. We can read each column “down” the table (viewing multiple observations), and each row “across” the table (viewing multiple variables).

This isn’t the only way to organize data, but it is the most common.

In this lesson, we’re going to cover some basics of working with data. Some of this might be familiar if you’ve been working with data for a while, but some of this might also be new.

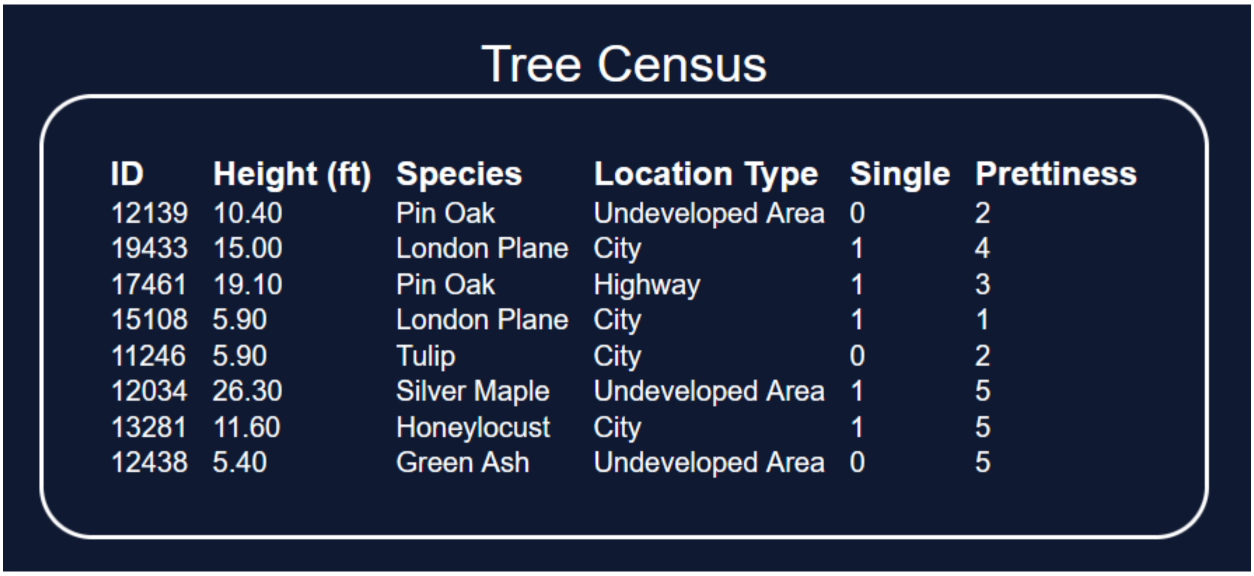
These ideas apply to all datasets you will work with. We will use an example of creating a tree census..

Instructions

Congratulations! You’ve just been hired as a census taker for the North American tree census! This census will be a partnership between Canada and the United States. You’ve agreed to use the Imperial measurements (the American system), so will be collecting the data in feet.

You’ve been assigned to the North Atlantic Region. That includes both the Northeast United States and Eastern Canada. Check out the U.S. Forest Service’s Forest Inventory for details about similar projects.

Take a look at the spreadsheet to the right. Pay special attention to the variables and values. You’ll need a deep understanding of how this data is organized to help us understand our future analysis.



Tree Census

ID	Height (ft)	Species	Location Type	Single	Prettiness
12139	10.40	Pin Oak	Undeveloped Area	0	2
19433	15.00	London Plane	City	1	4
17461	19.10	Pin Oak	Highway	1	3
15108	5.90	London Plane	City	1	1
11246	5.90	Tulip	City	0	2
12034	26.30	Silver Maple	Undeveloped Area	1	5
13281	11.60	Honeylocust	City	1	5
12438	5.40	Green Ash	Undeveloped Area	0	5