Working with Missing Data

4 min

In our dataset, we had some missing values. There are various types of "missing-ness" that affect how we treat the missing data.

If we remember when we were collecting our data, we were hungry, our fingers were cold, and we were distracted. Even though there's a reason we didn't enter some values (we were hungry and tired), it's not a systemic reason. There's no deeper meaning to why the data is missing: it just wasn't entered properly. This kind of missing is **Missing Completely at Random.**

However, we don't always know if there is a deeper meaning, so we have to treat missing data like a mystery to solve. For example, we might notice that all of the Redwood trees are missing Height values. Well, that's interesting! We can predict if a tree is missing its Height value based on what Species it is.

More generally *we can predict if one value is missing based on the value in another variable*. This kind of missing is called **Missing at Random.** It is a confusing label because it's not really missing at "random" in the normal meaning of the word. If we dig a little deeper into how the data was collected, we might uncover a story about the data collection or about Redwoods. For example, our tape measures might have been too short to measure them.

Finally, data can be structurally missing, meaning that we wouldn't expect a value there to begin with. For example, let's say we are also collecting data about fruit on our trees. Some trees will have visible fruit. For those trees, we can count how many fruits are visible. If there's no visible fruit, we can't count how many there are. The number of fruits will be **Structurally Missing**.

What should I do about missing data?

Well, for structurally missing data, we can just ignore it, we don't expect there to be values there anyhow. For Missing at Random and Missing Completely at Random, there is an entire science behind what to do with these values. Learn more in our course on Handling Missing Data.

When trying to recover missing data or work around it, the most important thing to consider is that anything you do will affect your analysis. Once data goes missing, it can't be recovered, so whatever decision you make becomes a part of your result and your analysis (even doing nothing will affect the analysis). Because of this, it is best practice to keep track of which values were missing just in case you ever need to revisit your data.

Instructions

Take a look at the dataset to the right. The 'Distance (ft)' variable refers to the distance between the trees that are growing in groups.

Data is missing in many different ways. Some data is structurally missing while other data is Missing at Random or Missing Completely at Random.

Take a moment to think about the ways that the data is missing. Is there anything that you should do about the missing values?

Tree Census

ID	Height (ft)	Species	Location Type	Single	Distance (ft)	Prettiness
11246	5.90	Tulip	City	0	nan	3
11562	14.10	Red Oak	Highway	0	nan	5
11584	6.10	Honeylocust	Highway	0	nan	3
12139	10.40	Pin Oak	Undeveloped Area	nan	3.10	3
13280	nan	Red Oak	Highway	nan	nan	2
13281	nan	Honeylocust	City	1	3.20	5
15156	9.50	nan	Undeveloped Area	0	nan	3
19325	12.20	Tulip	City	0	nan	nan
20143	6.70	American Linden	nan	0	nan	nan
21110	27.50	London Plane	City	1	3.30	1