The Shape of Data

4 min

For your new role as a tree census taker, you'll start with height and species. 'Height' and 'Species' are our **variables**. The height of each tree can "vary" from one tree to another (hence the name).

Each individual tree is called an **entity**, **observation**, or **instance** (there are a lot of names for this). We'll stick with observations, but know that these three terms are used interchangeably.

In a well-organized dataset, the variables describe a characteristic of our entities. However, it can be surprisingly difficult to define good variables. Good variables measure only one characteristic and should not be a characteristic themselves. Let's look at an example.

For example, in our tree dataset, we are interested in the type of environment the tree is in. For example, we are looking at trees along city streets, highways, and in undeveloped areas. We also want to know if trees are standing alone or with others.

There are many ways to organize this. We could:

1. Make 3 new variables: 'City', 'Highway', 'Undeveloped' and input 'alone' or 'group' in the values.
2. Make 2 new variables: 'Location' and 'Single' and input the location type in the 'Location' variable and 0 or 1 in the 'Single' variable.

Option 1 might seem ok during the collection phase, but it will be difficult when we start trying to analyze the data. For example, finding all of the 'City' or 'Highway' trees and then segmenting them by alone would be a challenge.

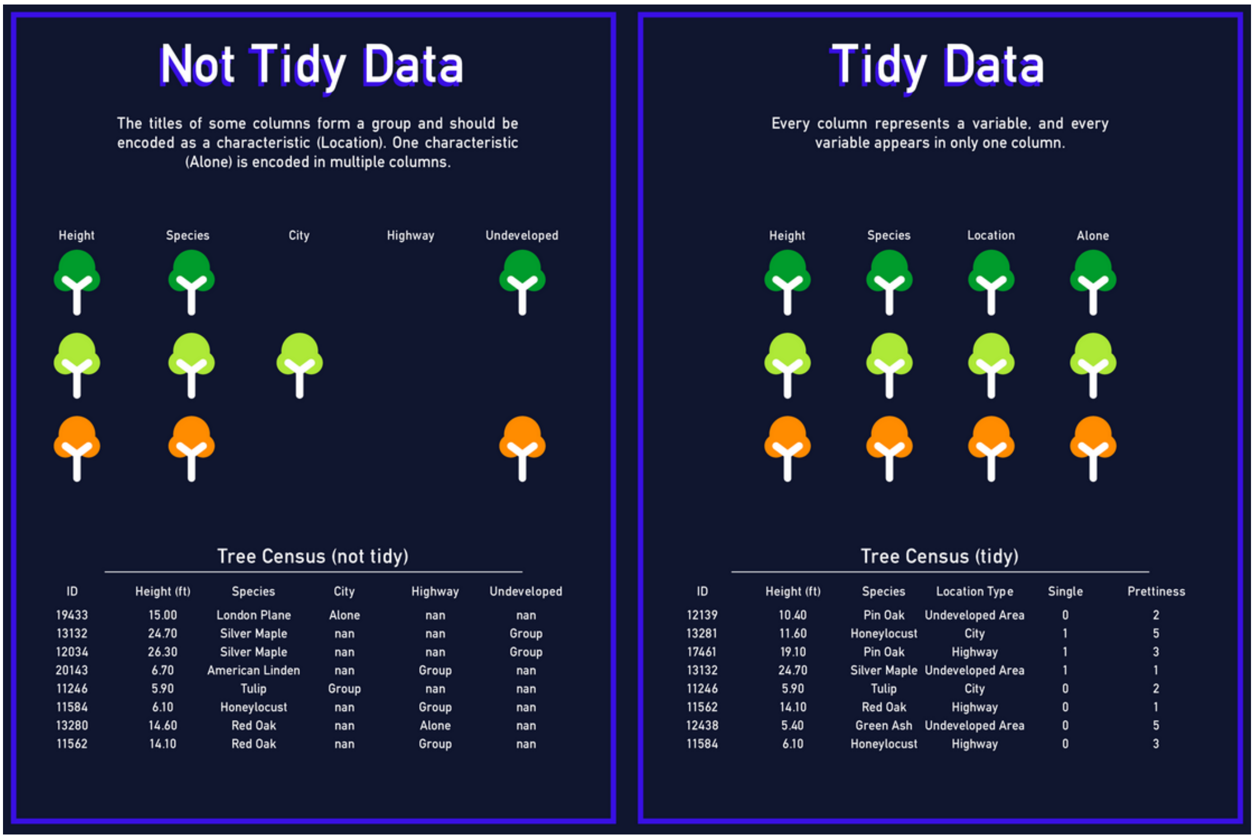
You may have already noticed that 'City', 'Highway', and 'Undeveloped' can be grouped together as a characteristic (and there are categories like 'Park' or 'Yard' that are missing). Rather than naming our variables for the categories themselves, we are better off having one variable named 'Location Type' and entering all the possible values. This will make analysis easier later on, and we can add new categories if we need to (like 'Park').

Looks like Option (2) is the better organization for us.

But what about 'Alone' and 'Group'? Well, we will talk more about this later, but for now, just know that the variable name will be 'Single', and we will fill it in with 1 for True/Yes and 0 for False/No.

Instructions

Compare the organization of the datasets to the right. Notice that with the tidy dataset, every variable has a value for every observation. This isn't always possible, but it is ideal.



Not Tidy Data

The titles of some columns form a group and should be encoded as a characteristic (Location). One characteristic (Alone) is encoded in multiple columns.

Tree Census (not tidy)

ID	Height (ft)	Species	City	Highway	Undeveloped
19433	15.00	London Plane	Alone	nan	nan
13132	24.70	Silver Maple	nan	nan	Group
12034	26.30	Silver Maple	nan	nan	Group
20143	6.70	American Linden	nan	Group	nan
11246	5.90	Tulip	Group	nan	nan
11584	6.10	Honeylocust	nan	Group	nan
13280	14.60	Red Oak	nan	Alone	nan
11562	14.10	Red Oak	nan	Group	nan

Tidy Data

Every column represents a variable, and every variable appears in only one column.

Tree Census (tidy)

ID	Height (ft)	Species	Location Type	Single	Prettiness
12139	10.40	Pin Oak	Undeveloped Area	0	2
13281	11.60	Honeylocust	City	1	5
17461	19.10	Pin Oak	Highway	1	3
13132	24.70	Silver Maple	Undeveloped Area	1	1
11246	5.90	Tulip	City	0	2
11562	14.10	Red Oak	Highway	0	1
12438	5.40	Green Ash	Undeveloped Area	0	5
11584	6.10	Honeylocust	Highway	0	3