

Dealing with Messy Data

4 min

All happy datasets are alike; each unhappy dataset is unhappy in its own way.

- Leo Tolstoy (if he had written a Data Science book)

Ok, Tolstoy was writing about families in Anna Karenina, but families are just like data. Clean datasets are all alike, but every messy dataset is messy in its own unique way. That's why cleaning data involves a lot of critical thinking when considering the nuances of the dataset you are working with.

Fortunately, there are some patterns in what can go wrong, and the first step in cleaning data is knowing what to look for.

What is a messy dataset?

Imagine we are outside collecting the data about our trees. We have our iPad and our tape measure. Our fingers are cold, we are distracted by a beautiful bird 🐦, and we're ready for lunch 🍲, but we just have to measure and categorize these last 3 trees. The last 3 entries look like this:

Tree Census						
ID	Height (ft)	Species	Location	Type	Single	Prettiness
21246	0.60	Tuuullip	Highway		no	1
11222	nan	Pin Oak	Highway		no	1
18564	6.10	Pin Oak	Highway		no	three

Yikes! What a mess. But we're hungry, so we decide to fix the issues after lunch. They never get fixed. Six weeks later, we are back at our desk ready to analyze our data. Oh no! We have over 10,000 observations and quite a few problems.

Messy Data Problems

Different problems need to be handled differently, so let's categorize them:

- Typos like Tuuullip for Tulip
- Missing data like the Pin Oak (tree ID 11222) that doesn't have a height

- Inconsistent coding like the Pin Oak (tree 18564)'s Prettiness value is 'three' rather than '3' and the Single value for all of our trees is 'no' rather than '0'.

If we don't fix these issues, we will likely end up with problems in our analysis. For example:

- Tulip trees might be divided into 2 categories
- We might get an inaccurate average height for Pin Oaks because we are missing a data point
- Our computer might return an error message when we try to group trees into their Prettiness value or find all of the trees that grow alone.




Finding and solving these problems requires detective work. For now, we will fix these issues manually, but know that if you work with data, you will see these issues again. We cover how to deal with these issues and more in [How to Clean Data with Python](#) and in the course of our Data Scientist Career Paths.

Instructions

Correct the data to the right.

Note that you will have to use some detective skills for two of the data points. For the typo, what is the most likely typo that would result in a height of .90? For the missing data, we actually cannot know what the value should be. One best guess is to take the mean of the other Pin Oaks.

Tree Census

ID	Height (ft)	Species	Location Type	Single	Prettiness
 Correct!	6.0	Pin Oak	Highway	0	3
11239	10.0	Pin Oak	High way	0	5
 Correct!	12.3	Honeylocust	Highway	0	2
21149	1.90	Honeylocust	Highway	0	1
 Best guess!	8.0	Pin Oak	Highway	0	1