Data Wrangling

Adittya
18 October 2018

Introduction:

Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis. With the amount of data and data sources rapidly growing and expanding, it is getting more and more essential for the large amounts of available data to be organized for analysis.

This process typically includes manually converting/mapping data from one raw form into another format to allow for more convenient consumption and organization of the data.

The dataset I am wrangling is the WeRateDogs dataset, which is a tweet archive of the twitter user @dog_rates. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

The ratings have a denominator of 10, while the denominators are always greater than 10 (since they are all good puppers).

Methods:

This was a challenging assignment as the data had to be obtained from 3 different sources:

- the csv provided by Udacity
- A link to the csv, which had to be programmatically downloaded
- Twitter API

The twitter data was obtained using a Python package called Tweepy. I was not able to obtain a few tweets from the twitter API as these statuses no longer exist. The names of the dogs were missing in many cases. I assumed that all lower-case names values were invalid and converted them to a programmable NA value in Python.

Rating is one of the variables which I expected to be interesting in this dataset. However, most of the rating were above the denominator. I left the numerators (with invalid values over 10) as is, because it appears to an intentional and humorous decision.

I removed the rows which were unnecessary to the analysis and then combined the three tables into a master dataframe.

The master dataframe was then exported to csv format using Python.