Wrangle Report

2/23/19

Jamie Farley

The goal of this project was to gather, assess and clean data obtained from a tweet archive of Twitter user @dog_rates, also known as WeRateDogs. This Twitter account rates people's dogs with a funny comment about the dog. The ratings usually have a denominator of 10 and the numerator can be a lot higher. This dataset included basic tweet data like an ID, timestamp, text, URL, etc. for 5000+ tweets until August 1, 2017.

Gathering Data

The data required for this project came from three different datasets:

- Twitter Archive CSV File – provided by Udacity and downloaded manually
- Tweet Image Predictions TSV File – hosted on Udacity's servers and downloaded programmatically using the Requests library and the URL
- Twitter API & JSON Library – queried Twitter API for each tweet's JSON data using the Tweepy library and stored each tweet's entire set of JSON data and converted to pandas dataframe.

Assessing Data

I visually inspected the data from the .csv file using Excel to determine any errors and then I loaded each data file and assessed them programmatically using python in Jupyter Notebooks. After inspecting and assessing all the data in each dataset, I listed the quality and tidiness problems that needed to be addressed in the cleaning section.

Cleaning Data

After previous experience and practice of wrangling, I decided to divide each problem into three parts. The first part was to define the problem, the second was to code, and then the third part was to test the code. This process made it very concise and easy to keep track of each problem in the dataset. It's also clear and organized throughout the entire notebook.

Cleaning consisted of removing unrelated columns that weren't necessary for analysis. It also included removing incorrect values (None, a, an, the) in the dog name column, dropping duplicates, and merging all the datasets into one dataset.

Conclusion

By wrangling this dataset, I was able to answer some basic questions using visualizations from the resulting dataset after gathering, assessing and cleaning the data obtained from Udacity and the WeRateDogs Twitter account. That being said, there was only a limited amount of information obtained about each tweet, it would be hard to ask some more in-depth questions about the dataset. There was a lot of missing and/or incorrect information as well, which is why the data required a lot of assessing and cleaning. Additional gathering of more data would most likely help answer and solidify some more questions.