Kyle Wilbert

Data Analyst Nanodegree

01/15/19

Data Wrangling Report

In the Gather stage, I was able to gather the first two data sources easily based on the exercises and documentation given in the course. For the JSON scrape I definitely used the files included in the Twitter API lesson of the project module as guidance as well as a thread in the Udacity Data Science Slack channel. I did run into a hiccup the first time I tried to scrape the data because I forgot to set the arguments to wait for the rate limit. Otherwise the instructions for setting up the Twitter developer account were helpful and easy to follow.

The Assess stage was relatively straight forward. I found mostly Quality issues and only a few Tidiness issues. Perhaps I misclassified some of my issues or simply overlooked other Tidiness issues. Also, I listed more Quality issues than I addressed.

The Clean stage was the most challenging section for me. I used a couple outside sources to help me isolate the overlapping records across multiple datasets so that I ended up with one dataset with complete information. Otherwise, I was able to drop a number of columns with data that was incomplete or seemed unlikely to be useful for the analysis. Also, I manually cleaned up a few of the dog ratings that I visually found to be incorrect.

To clean up the Source column, I had originally tried to use regular expressions. Upon further investigation, most sources suggested using Beautiful Soup. In fact, I found one source that helped me with a simple implementation of it, and it cleaned the HTML beautifully, with way less code than a regular expression.

I noticed when I reloaded the data from the master file that I needed to change types. I had originally listed my type changes in the Clean section, but it made more sense for me to move them to the analysis section, since that was the dataset I was going to use for my brief analysis.

After changing types, I ran a few simple analyses of the data. I practiced filtering data to reveal the percentages of tweets that had dog stages. Next, I created a simple bar graph to show the distribution of the sources of tweets. As an aside, I definitely prefer plotting in R to plotting in matplotlib. Finally, I did a brief analysis of the confidence levels of the image predictions.

Overall, I found this project challenging. The code in the scraping phase of Gather is still a bit beyond me and requires further study. Also, I'd need to practice a lot of plotting in matplotlib to get comfortable with it and create anything beyond simply plots.