Hagop Bozawglanian
Udacity Nanodegree - Data Analyst

Data Wrangle Report

There are multiple steps to this project but I will be outlining the steps that I took into wrangling the data. There are three data sources that I had to pull data from: the csv for the archived tweets, the image files from the udacity link and the live data from Twitter using the tweepy library and the twitter APIs.

Pulling the data from the archived csv file and the url request were simple and I got them on the first try, however the twitter API was giving me some issues. I realized it was because I had refreshed the credentials and had not updated those credentials in the pandas code. The documentation provided was sufficient to go through and set up the developer account and use the tweepy library. It was also difficult and frustrating to pull the data because of the timeouts needed with a free twitter developer account, it would take a very long time for the code to run because of these limitations.

Assessing the data was somewhat simple, but it is really hard to do visual assessment of large data sets, and for some of the data here it was difficult to do a programmatic assessment. It is also difficult to assess for me when you have multiple tables all referring to the same data.

Cleaning the data was definitely tougher than anticipated. For instance the work with the rating numerator and denominator. It was difficult to parse out the values as correct or incorrect because the ratios used are not “correct” for instance people tweeting “ the dog is 12/10” isn’t a “correct” fraction but it is one that is used in our vocabulary. I dropped a couple columns I didn’t see any value in, made some simplifications like removing camelcase, and converted data types for correct representation of the data. The most difficult column to fix was the source column, I am already very familiar with the python regex syntax but wanted to try a different approach, so I decided on using the Beautiful Soup library. It worked really well since it is normally used for pulling data out of HTML so it was simple to pull values out.

When I pulled the data back into the final dataframe after storing to csv I noticed some columns lost their changed data type(timestamp and category) so I had to redo those to the end to use for visualizations. I analyzed the data and noticed a few categories for sources, type of dogs, and accuracy of the prediction. I also learned that the predictions were the same each time, but with varying confidences. I was able to use bar charts, horizontal bar charts, and pie graphs to show relationships between the fields being analyzed.

This project was definitely the most challenging one so far. Not only because of the different levels of understanding you must have on how to wrangle, clean, and analyze, but you have to do so yourself without prompts being given. This to me is the most difficult part, looking at data and really thinking “what needs to be done here? Or what does this mean?”