Wrangling Report

This project was designed to test our knowledge of the data wrangling process. We would have to successfully gather, assess, and clean data from Twitter's WeRateDog's page. The data initially given to us comes from the twitter archive for this page. However this data is not complete, as it does not contain retweet data and favorite data about each tweet. In order to get past this problem, we have to query Twitter's API using python's Tweepy library. To get further insights, we would also use another dataset that tried to predict what breed of dog the tweet had pictured using a neural network. By utilizing all these datasets, and cleaning them in a way that makes them easy to analyze, we can have some real insights into the trends in the data.

The first thing I had to do was gather all the data to work with locally. The archive data was given as a csv file, and importing was simple. For the image data, we had to download them off of Udacity's servers, which involved using the requests library. But the most difficult part was gathering the retweet and favorite data, as this involved querying Twitter's archive. After setting up a twitter account and gaining access to the required keys, I was able to query twitter's API for the required data and dump the data into a Json file.

Once all of the data was gathered I could begin assessing it. Most of the data issues were quality issues, as the data was not complete. There was missing data for the images, retweet counts, and favorite counts, many dog names were not real dog names, there were many erroneous data types, and many of the dog ratings were incorrect. All 3 groups of data were moved into one dataset to tidy up our data as well. With these problems recognized, a combination of programmatic and individual corrections was applied to clean the data, using many of the methods learned throughout this section. With the data now cleaned thoroughly, I was easily able to analyze the data to find trends.