Wrangling Report on the WeRateDogs Twitter Archive Using Twitter's API.

Data Wrangling, the most important part of the Data Analysis but also the most tedious, is about gathering the data, assessing it and cleaning the unwanted data also fixing some errors in data.

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. The goal of the project was to analyze and know the humour level of each breed of dog. Each dog is rated differently based on humour level with almost common rating denominator of 10 but with weird levels of rating numerator, some higher than 10, but that is all part of the fun of the @dog_rates twitter account.

Data Gathering: The Udacity's *twitter-archive-enhanced.csv* file was gotten through manual download as stipulated while the *image-predictions.tsv* file was downloaded programmatically. I also downloaded the *twitter_json.txt* file from Twitter's API and it was a bot challenging since this was my first time working with APIs. The process of getting my keys were, in itself, another load of stress. From creating the Developer's account to answering cycles of questions but I finally got it. Once credentials were gotten and verified, a program had to be written to get the Likes and Retweet Counts of the Tweet IDs in the *twitter-archive-enhanced.csv* file.

Data Assessing: The goal was to find, at minimum, eight quality and two tidiness issues. Mine were as follows:

Quality

- Remove retweets
- Dropping source column
- Renaming rating_numerator column
- Incorrect ratings
- Missing values for in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp
- Incorrect data types for tweet_id, in_reply_to_status_id, in_reply_to_user_id, timestamp, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp
- Incorrect datatype for tweet_id
- Renaming every column
- Incorrect datatype for tweet_id
- Investigate outlier in rating_denominator column

Tidiness

- Single variable split into multiple variables
- Combine likes and retweets

Data Cleaning: The approach to take to clean the aforementioned issues was the *define-code-test* method. It entails defining the problem, coding up the solution and the testing the results which were done individually for each problem

Data Storing: After Cleaning, the new master data frame was stored as a *twitter-master-archive.csv* file

Data Analysis and Visualization: After everything, since the basis of this project was Data Wrangling, I did some basic analysis and visualization on the new master data frame