Wrangling Act Report

This report covers the procedures and art involved in Data Wrangling WeRateDogs twitter account data. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. In this project we will be working on wrangling the data from WeRateDogs Twitter page.

Tools used in carrying out this activity are not limited to the following:

- pandas
- numpy
- random
- matplotlib
- Seaborn
- Tweepy
- Json

We first imported the archived data into a dataframe using pandas after which we viewed a rough description of the data. We then programmatically saved the image prescriptions to the file system after converting the fetched data to CSV format.
Using Tweepy, we queried Twitter API for each tweet in the Twitter archive and the JSON in a text file *tweet_json.txt* which we read into a JSON file format programmatically. We then created another dataframe called *tweets_df* which have a the *tweet_id, retweet_count and the favorite_count* and then the file was exported to a CSV file *tweets_data_details.csv.*
With these in place, we carried out a quick Exploratory Data Analysis on the dataframes. After exploring the data, we were able to establish that the following issues needed to be addressed:

1. The maximum value for *rating_numerator* imakes1776 which is an outlier
2. The SD of rating numerator is much high compared to the value of the mean, this means that the outliers have a big impact on out values
3. Links in client source list, source_list in archive df
4. Missing values for retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp
5. Duplicate tweet with the same ID

6. Null values exist for retweeted status and in replies
7. Timestamp for the archive is in object format
8. Stopwords (a, an) are used for the name of dogs
9. Inconsistent numerators and denominators
10. Not all the data have images

We addressed all the issues above programmatically which entailed a summary of cleaning the data, converting relevant data types, removing null values, merging repeating variables, removing invalid dog names, merging dataframes in to one and then removing duplicates before saving the cleaned data into the filesystem.

For reference purpose, we have exported all the individual dataframes into the file system at each critical point.