Data Wrangle Report

This project was aimed at reviewing the tweets made by the twitter account WeRateDogs which has been archived where several dogs are rated on scale of 10 with the dogs been rated above 10 in assurance that they are good dogs

For this project, the data had to be gathered in 3 different ways which included:

1. Direct Download from source
2. Programmatic Download from Udacity Servers
3. Programmatic Download via Twitter API

After gathering the data, each of them were then loaded into a pandas dataframe, taking note of the file types which were in csv, tsv and json format. Once loaded into the dataframes, the data could then be accessed visually and programmatically to find any issues with the data such as messy and dirty data which will then be cleaned later one. From the observations on the datasets, issues surrounding the following were found:

- Missing data
- Column with wrong datatype
- Two datasets needed to be merged into one
- Anomalies with denominator of the rating, higher and lower than 10
- Anomalies with numerator of the ratings
- Unnecessary columns which will need to be dropped
- Rows which are repeats because they are retweets

During the cleaning process, for each of the datasets, a copy was first made which would be the clean version of the dataset and be worked on. All processes or steps done on the datasets were documented using the Define, Code and Test framework, where the issue such as missing values was defined alongside what will be done to to, then the code was written down with comments to explain them which after running will then proceed to test showing that the issue initially defined has now been resolved. In addition to identifying these issues and proceeding to clean the datasets taking them into account, other activities were done on the dataset for easy analysis, such as creating month and year columns from the datetime column, which will be very useful during the analysis.

Once all the cleaning was done, the new clean dataset that can be used for analysis was then exported as csv with the name "twitter_archive_master.csv". This new document can be shared to various individuals to perform their analysis and find insights without having to worry about dirty data and structural issues.