Project Overview:

In this project, Data Wrangling(analyzing & Visualization) of WeRateDogs data has been done. WeRateDogs is a twitter account that rates people's dog with hilarious comments. The rating always has a denominator of 10 and Numerator greater than 10. The analysis has been done on Jupyter Notebook using different python libraries.

I have gathered 3 datasets for this purpose, assessed them for quality & tidiness issues and cleaned them for analysis.

3 Datasets:

1. Twitter-archive-enhanced.csv => This dataset was downloaded manually
2. Image_predictions.tsv => this dataset was hosted on Udacity server & was downloaded programmatically using the request library
3. Tweet_json.txt => Twitter API for each tweet's JSON data was queried using Python's Tweepy library and store it in tweet_json.txt file

Gathering Data:

To start with the project, firstly I imported all the necessary Python libraries required for analysis & Visualization.

After importing libraries, I started data gathering from 3 different datasets mentioned above. 'Image_predictions.tsv' data was downloaded using Udacity Server programmatically using Request Library.

'Twitter-archive-enhanced.csv' file was provided by udacity. Reading the file just by using pd.read_csv command.

'Tweet_json.txt' was created by accessing & downloading twitter JSON's data using tweepy library.

To get a list of tweet id from 'twitter-archive-enhanced.csv', was looped through each tweet ID and query and query Twitter's API with the ID to get each tweet's JSON data. The data was recorded in a text file named 'tweet_json.txt' with each tweet's data written in a new line.
After this, we stored tweet_id, retweet_count, favourite_count variables to the tweet_info data frame.

Assessing & Cleaning & Storing Data:

After gathering the necessary datasets, all the data frames are read using pandas read function.

Before cleaning, I made a copy of the dataset to work with.

Programmatical assessment was done to assess data using Pandas functions like info(), describe() and value_counts() function.

There were different Quality & Tidiness issues in the dataset. I have noticed 8 Quality & 2 Tidiness issues mainly like incorrect data types, missing values etc.

Each issue was defined, code and tested during the cleaning process and it is clearly documented in the dataset.

After the dataset is cleaned enough for the analysis, it is saved into a CSV file named ‘twitter_archive_master.csv’ and used for my analysis accordingly.