REPORT

ON

DATA WRANGLING

OF

WE RATE DOGS DATA

DATA GATHERING:

We were given a dataset named ***twitter-archive-enhanced.csv*** which I downloaded manually and also a link which contains image predictions of dogs in Udacity's server. The image predictions were downloaded programmatically with the help of ***request library*** and the downloaded data was stored in a file which I named ***image_predictions.tsv.***

I was supposed to also use ***tweepy library*** to extract data from WeRateDogs twitter account with the aid of twitters API, but I was unable to do that because my developer's account was not yet approved by twitters team. In other to move on, I downloaded the json file on the Udacity's classroom manually and read the file programmatically on my workspace where I selected only needed data such as ***tweet id, retweet count, favourite count and followers count*** then I formed a pandas data frame ***API_tweets*** where I stored the collected information of my json file.

DATA ASSESSMENT AND CLEANING:

The data were assessed visually using excel and MS Word for ***twitter-archive-enhanced.csv*** and ***tweet_json.txt.*** Also, the data was assessed programmatically using info () method and other necessary functions. I was able to detect 8 quality issues and 2 tidiness issues related to the dataset. Some of the issues and the proffered solutions include:

Quality issues:

1. The source column was represented in HTML format so I had to remove the html text and convert the column to string.
2. Some of the dogs had incorrect name, so I had to fix this issue by changing all incorrect names to None since we can't just randomly choose a name.
3. The tweet id column was in int, so I changed it to string to avoid losing information when working with it as some software might round up the numbers.
4. Timestamp column had its data type as object, so I converted it to datetime.
5. Rating numerator and denominator column was in int, so I converted it to float since some of its values had decimals.
6. Some values in rating numerator's column were wrong so I had to update them.
7. There were different dog type columns such as doggo, floofer, pupper and puppo which I merged together to form a single column named ***dog_type.***
8. Some columns were not really relevant so I drooped them from the dataset.

Tidiness issues:

1. There were three different datasets, so I merged them together to form a single dataset.
2. Some tweets do not have images, so I removed such observations.

DATA STORING:

After cleaning the data, the data was stored in a csv file named ***twitter_archive_master.csv*** then I began my analysis on the data.