Merna Ghaffar

As part of the Data Wrangling Project of my *Udacity Data Analyst Nanodegree*, I looked at the data connected with the WeRateDogs Twitter account and undergone the data wrangling process, which included:

1. Gathering data, where I collected the data using different means
2. Assessing data, where I looked at the data thoroughly and took notes
3. Cleaning data, where the notes taken are addressed and action is taken

▪ Gathering the Data

This was personally the most challenging aspect of the project. The data was collected from three resources: an existing file twitter-archive-enhanced.csv was downloaded and saved as *df*. A second file 'image_prediction.tsv' was downloaded programmatically from Udacity servers using the Requests library and stored as *image_predict.* A third dataset with retweet count and favorite_count was retrieved from Twitter servers in the form of JSON entries, using the Tweepy and saved as *tweet_info.*

▪ Assessing the Data

Assessing the data involves examining both data *quality* and *tidiness.* The following highlights some of the issues that I come across:

Quality Issues, which relates to content issues (completeness, validity, accuracy, consistency) of our data:

1. We only need retweeted_status_id is NAN, with valid tweet_id . Drop everything else.
2. 'in_reply_to_status_id','in_reply_to_user_id','retweeted_status_id','retweeted_statis_user_id' are columns that are not useful to us and need to be dropped.
3. Replace the 'doggo','floofer','pupper','puppo' columns with a 'stage' column.
4. Some of the dog names are set as "a", "the", "an","none". This should be changed into NaN.
5. Tweet_id is an integer and should be a string in all three datasets.
6. Timestamp is set as a string and should be changed into datetime.
7. Most possible breed column with a prediction confidence column should be created to replace the several breed prediction columns.
8. Drop ratings without images.

Tidiness, which relates to structural issues in the dataset:

1. All three dataframes should be merged together (df, image_predict and tweet_info) into one dataframe.
2. Rating numerator and denominator should be in one rating column (numerator/denominator)

- Cleaning the Data

The mentioned points were addressed (in the same order) and tested out to ensure that changes were successfully made. The cleaned data was stored as ‘twitter_archive_master.csv’ and is ready for analysis.