Wrangle Report

The first step gathering the data :

in this project I gather the data from 3 different resources in 3 different file format the data I gather are :

1- **Enhanced Twitter Archive: this archive contains** the basic tweet data for all 5000+ of tweets weRateDog**,** the file name, and format (twitter-archive-enhanced.csv).

2- **Image Predictions File:** a table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction.The file name and format **(**image_predictions.tsv**).**

3- **Additional Data via the Twitter API:** because I don't have access for Tweepy library I use (tweet_json.txt) that contains the data gathered from Twitter API.

The second step Assessing the data:

The data assessed by using both visual assessment and programmatic assessment with pandas' methods such as df. info(),df. shape, df. describe(),…etc. And I found some quality and tidiness issues in data :

❖ Quality issues:

issues in twitter_archive_df are :

- Tweet _id should be a string instead of int.
- timestamp should be DateTime instead of a string.
- there are some invalid dog names like a, an,...etc.
- the none value in the name column should be NaN.
- there alot of missing values in retweeted_status_id, retweeted_status_user_id ,retweeted_status_timestamp in_reply_to_status_id,and in_reply_to_user_id.
- the rating_denominator should be 10 always.
- there are tweets created after August 1st, 2017.
- there are some missing values in the expanded_urls column.

issues in image_predictions_df are :

- the names of column not clear so change p1,p1_conf,p2,p2_conf,... to prediction1,prediction1_confident.
- tweet_id should be a string instead of int.
- there are some missing values since there are 2075 rows whereas twitter_archive_df contains 2356 rows.

issues in tweet_json_df are :

- tweet_id should be a string instead of int.

❖ Tidiness issues:

- in twitter_archive_df doggo, pupper, puppo, and floofer are dog stage.
- the data in 3 separate data frames.

- we only consider tweets, so we do not need the replies and the retweets columns.

The third step Clean the data:

After assessing the data, I start cleaning the issues the I notice and I start to clean the tidiness issues first so I create a column contain the dog stage and fill it with the values, then I merge the 3 data frame, and the drop the replies and the retweets columns. The process of cleaning tidiness issues solve some quality issues like :

- the problem of missing values in the expanded_urls column.
- the problem of tweets created after August 1st, 2017 since image_predictions_clean contains only tweet created before August 1st, 2017.
- the problem of missing values of image_predictions_df since all data in this data frame are before August 1st, 2017.

I start to clean the quality issues like fix the data types and rename the columns and convert the None values into Nan values.