Data Wrangling Report

Data Gathering:

For the data gathering, I did not put too much more effort than following the project instruction. The tweet_json.txt was also made by following the instruction and I managed to collect 2354 tweets ranged from 2015 to 2017.

Data assessment:

Here are the issues I found in the three tables after the data gathering:

Quality issue:

For issues found in 1-8, all are from table *twitter_archive_enhanced.csv.*

1. Incorrect retrieved values in “rating_numerator” as compared with ratings in “text”. For example, in row index 45, in the “text”, the rating is “13.5/10”, but in “rating_numberator”, it is 5 instead of 13.5. Also the rating_numerator should be float instead of integer.
2. Duplicated tweets when dog_rates retweeted its own tweets before. For example, row index 68, which the dog’s name is Emmy, is a retweet for tweet in row index 76. Therefore, retweets should be deleted.
3. Multiple website urls in column “expanded_urls”.
4. Missing dog names. Row index: 149, name in “text” should be Pipsy, while is None in “name’ column.
5. ‘None’ values in “name”, “doggo”, “floofer”, “pupper”, “puppo” columns are strings instead of actual null values.
6. 23 rows do not have a rating_denominator of 10, and most of them are wrong either because of wrong rating information extracted or no rating exist.
7. The dog stage for row 433 should be “floofer” but None was recorded in neither of the 4 stage columns. Similar for row 1228 and 1351, seems like the word filter for stage columns only contained the exact word but no plural forms or derived words such as floofs.
8. In timestamp column, the last ‘+0000’ is useless and the cell should be converted to datetime type instead of current string type.
9. Row numbers of three tables are different (2075 for image_prediction, 2356 for twitter_archive_enhanced and 2354 for twitter_json_data), and should be merged based on image_prediction data.

Tidiness issue:

1. In *twitter-archive-enhanced* table, the columns of “doggo”,”floofer”,”pupper” and “puppo” should under one column “stages”.
2. Text in *twitter_archive_enhanced* table should not include the urls.

Data Cleaning:

In data cleaning, first I focused on the quality issue #9 because it will first merge all three tables together and easier to continue other cleaning after the combination. After the combination, some of the row numbers recorded in the “quality issue” section above might have been changed to other row numbers. Although it is not difficult to match the original row numbers in the original table and then locate in the new table using the tweet_id parameter, it is better next time to record the tweet_id in the first place.

Data Storing, Analyzing and Visualizing:

The cleaned table was stored as “twitter_archive_master.csv”.

For the analysis and visualization, I mainly used statistical methods of linear regression to find out the correlation among parameters, as well as logistic regression to find out if any parameters are related to dog prediction in the images.