Wrangle Report

For this project, the objective was to analyze Twitter data for the twitter account @dog_rates. The data was spread across three primary areas:

- Twitter data archive from @dog_rates
- Data downloaded from twitter
- Image prediction information provided by Udacity

The twitter data archive from @dog_rates

This data was in csv and was loaded easily into the local dataframe for analysis. The following issues were noticed in the data:

- Quality Issue - Base on the column retweeted_status_id, we noticed that there are around 181 retweets in this dataset. This should be removed.

- Quality Issue - There were tweets with the Text 'We only rate dogs'. These tweets should be excluded from our analysis.

- Quality Issue - There are rows where the rating_denominator is not 10. On checking visually for these rows, it seems that they were not parsed correctly from the text. Re-extract the Rating denominator from the text

- Quality Issue - For the rows where rating_denominator is incorrect, it was also noticed that the rating_numerator is incorrect. This needs to be correctly extracted.

- Tidiness Issue - Currently the rating_numerator and rating_denominator columns are type objects which would be string. We also noticed that some numerator values could be decimals. Change numerator to float and denominator to int

- Tidiness Issue - The puppy stage columns could have been consolidated to have a categorical value. However, there are 2 rows which have both values. For these 2, concatenate using ','.

- Quality Issue - The puppy stage columns have text like 'None' Can be fixed as Blank.

- Quality Issue - The timestamp column has additional details like +0000 which does not look necessary/correct.

- Tidiness Issue - The timestamp, retweeted_status_timestamp columns are object instead of datetime. Change timestamp to datetime type

- Tidiness Issue - in_reply_to_status_id, in_reply_to_user_id expanded_urls contains null values. These can be excluded from analysis.

- Quality Issue - The dog name column has 'None'. Should be Blank instead

- Quality Issue - Are names like all, None, an valid values. On visually checking, For values where it starts with a lowercase letter: a, all, the, actually - again here it seems to be incorrectly extracted. However on visually checking these records, it seems that there are no names in these tweets. Reset these values to Blank

Data downloaded from Twitter

- I was not able to download the data from twitter due to constant timeout issues. I have however included my code base for the same in the jupyter notebook.
- This dataset had information also for retweeted tweets. This was filtered out.

Image Prediction data

- We noticed missing data in this file due to record number mismatch against twitter archive data. However, there is no way to merge this data back again.
- I wanted to get the predicted dog from this dataset. There are 3 predictions. Based on whether the prediction is correct or not and the probability value, I created a score for each of the probabilities. First the p1_dog column was mapped to a 1 or 0 based on True or False respectively. The p1_conf value multiplied with this value gave the score. The final predicted dog was based on the probability which had the highest score. This will help in analyzing the data patterns based on the type of dog in our dataset.

Merging the data together

The twitter data archive and json data were merged. However the merge was done on ‘inner’ join on column ‘tweet_id’. This ensures that the merged dataset would contain common data from both data sets. This was important since the retweet count and favorite count in the json dataset was crucial for the data analysis and would not want them to be blank. Additionally we would be able to ignore tweets that are retweets since we had already cleaned it out of the json dataset.

The final merge was the resultant dataset from the above dataset with the Image prediction data. This was done on a left join on the column tweet_id. This way we ensured we retained all

the data that we had information from json and archive datasets and added the data from image predictions.

Finally, we noticed that there were some columns which we would not be used for our analysis and was removed.