Data Wrangling

Gathering Data

There were 3 sources of data provided in this project

1) The WeRateDogs Twitter archive was provided as a downloadable file and it was straightforward in reading it into a data frame df_twitter_archive using the pandas read_csv method
2) The tweet image predictions were downloaded using the Requests library from the provided website and the contents were written to a .tsv file. This .tsv file was then read into a dataframe df_predictions, using the pandas read_csv method
3) The retweet count and the favourite count had to obtained directly from the Twitter API, using the tweet ids (converted to string) from the Twitter archive provided in step 1. Using the tweet ids, the individual tweets were written into a file in json format. A pandas dataframe df_tweet_json was created for the tweets from the json file using the pandas read_json method.

Assessing the Data

df_twitter_archive

The following issues were noted while assessing this dataframe

Tidiness Issues Summary

1) There were individual columns for the dog stages doggo, puppo, pupper, floofer. They could be merged into one column dog_stage instead of 4 columns.
2) The retweet and favourite count columns would have to be merged from df_tweet_json dataframe.

Quality Issues Summary

1) As per the instruction the retweets would have to be removed from the df_twitter_archive dataframe
2) Some entries have invalid names like just, such, a, an etc. Some of them had a name specified preceded with "name is".
3) Some tweet texts had no valid name, but the name column had an invalid name.
4) There were some tweet texts that had multiple dog names. However, the name field had only one name.
5) There were some tweet texts that had multiple fractions. The right fraction had to be used as ratings
6) Some tweet numerators were float and the rating numerator had the decimal part as the numerator.

While the issues are summarized above, some entries needed individual handling as they did not follow a fixed pattern.

df_predictions

Tidiness Issues Summary

1) The tweet_id column entries were integer type.

df_tweet_json

Tidiness Issues Summary

1) There were too many columns that were not relevant for analysis. Only the retweet_count and the favorite_count was relevant.
2) The retweet entries were to be removed.

Quality Issues Summary

1) There were 2 columns id and id_str that were meant to be tweet ids. But in some tweet entries the 2 values were different. Which would be the right one?

Cleaning the Data

The Tidiness issues for df_predictions and df_tweet_json were pretty straight forward and I am not going into them in detail

df_tweet_json

Cleaning Quality Issues

- The id and id_str issue was not explicitly cleaned. id was deemed correct as when doing a join with df_twitter_archive using id_str over 700 tweet ids did not match between df_tweet_json and df_twitter_archive. Whereas when id was used only a handful of tweets were lost

df_twitter_archive

Cleaning Quality Issues

- The retweets were removed by removing any row that had retweeted_status column entry that was not NaN.
- Quality issues 2, 3, 5 and 6, required appropriate python functions to be written and the dataframe apply method was used make the changes.
- While a big chunk of the rows that had the problem highlighted in 5), had multiple fractions where the 2nd fraction was the correct rating, and were fixed, there were others where this pattern did not hold. I didn't bother fixing those as the cleaning would be too much for the scope of this project.
- Issues falling under the 4) category had to be handled on an individual basis. I fixed only 2 of them in the interest of time.

Cleaning Tidiness Issues

- As for issue 1), the individual dog stage columns were merged into a single dog_stage column using the dataframe apply function and using an appropriate lambda function to join the column entries.
- The favourite count and retweet column counts were added to df_twitter_archive through a merge using the twitter_id column of df_twitter_archive and id column of df_tweet_json as

keys. An inner join was used so that only entries with retweet and favourite counts were present.