Data Wrangling Report

The tasks of this project were:
1. Gathering data
2. Assessing Data
3. Cleaning data

Gathering data for this project

1. **The WeRateDogs Twitter archive:** this data was provided by Udacity
2. **The tweet image predictions:**, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and the provided link.
3. **Twitter API & JSON**: Using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data was written to its own line. Then this .txt file was read line by line into a pandas DataFrame with tweet ID, retweet count, favorite count and retweeted status.

These were then exported to their corresponding tables.

Assessing data for this project

By using programmatic methods, I assessed the three dataframes looking for inconsistencies, errors, duplicates, missing data and other problems with the data. These included:

- .head()
- .mean()
- .info()
- .duplicated()
- .value_counts()
- .sum()

I thus made a list of 11 quality and 2 tidiness issues to be fixed. These were the following:

Quality issues
In table tweet_archive

1. Data type in 'timestamp' column is string instead of datetime, and there should be columns for year, month and day
2. There are 181 retweets - we only want original tweets with images
3. Columns [doggo, floofer, pupper, puppo] have "None" values instead of null values
4. Column 'name' has values 'None', 'a', 'an', 'the' and other wrong ones
5. There are rating_denonimator values other than 10 and wrong numerator values
6. Drop columns that won't be used for the analysis

In table tweet_predictions
7. Names in columns p1, p2 and p3 have different capitalisations
8. Only use first (stronger) prediction and rename the columns to 'prediction', 'confidence' and 'is_dog'
9. Drop duplicated jpg_url rows
10. Drop columns that won't be used for analysis

In table tweet_popularity
11. Drop 'retweeted' column

Tidiness issues
In table tweet_archive

1. Mutliple columns for dog type variable (doggo, floofer, pupper, puppo)
2. Merge all tables

Cleaning data for this project

After creating a copy of each dataframe, I went on to use common programmatic methods to clean and fix those issues. Most of them were simple methods like

- pd.to_datetime
- .replace()
- .drop() and drop_duplicates()
- .merge()
- Etc.

However, some of them were more challenging and included creating “loops” that analysed each row and altered the data according to “if” statements. A particular one involved using the index value of each row to iterate through other columns and replace the values in one column according to the values in other columns. I am not sure yet if this is the most efficient way of doing it but I did learn quite a bit in the process.

I also had to look up a lot of times for answers in communities like Stackoverflow to find out how to do what I was wanting to do, since many methods and details were new to me.