DATA WRANGLING REPORT

In this project, I was working with the dataset from the tweet archive of Twitter user @dog_rates or WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc.

I effectively wrangled the data which involved gathering additional data from other datasets, assessing the data for any quality and tidiness issues and cleaning the data where I found any fault.

Gathering Data

I began by downloading the Tweet archive which was in csv format hence not much formatting was required to display it in pandas as a data frame.

I then programmatically downloaded the image predictions dataset which was in tsv format, using the requests library in Python Pandas. The URL was provided in the project description.

Finally, I read a tweet JSON data text file line by line into a Pandas data frame with the tweet ID, favorite count and retweet count columns. The JSON data file was queried from the Twitter API using Python's Tweepy library.

Assessing data

After gathering all the data I required, I assessed it both visually and programmatically for any issues and errors that required cleaning. These are the issues I found.

Quality issues

1. Missing values in retweeted_status_id, retweeted_user_id and retweeted_timestamp columns.
2. expanded_urls columns has some missing records.
3. Invalid entries in name column.
4. in_reply_to_status_id and in_reply_to_user_id columns have null values.
5. p1, p2 and p3 columns not quite neat and consistent.
6. tweet id column should be in string format and not integer in all datasets.
7. The timestamp column should be in datetime format and not object.
8. tweet id column should be similar in all datasets.

Tidiness issues

1. Columns doggo, pupper, puppo and floofer should be in one column in the twitter archive dataset.
2. All tables should be combined into one dataset for analysis.

Cleaning

I created copies of the original datasets to perform cleaning.

I began with the quality issues by dropping the columns that had many missing values and were not relevant to my analysis. These columns were retweeted_status_id, retweeted_user_id and retweeted_timestamp and expanded_urls in the twitter archive dataset.

I then combined the doggo, pupper, puppo and floofer column into one column and called it look for tidiness reasons.

F

Storing the clean data

After performing all cleaning operations, I merged all three datasets into one dataset and saved it into a file named twitter_archive_master.csv .