WRANGLE REPORT: WE RATE DOGS TWEET ANALYSIS

The wrangling of the WeRateDogs Twitter account was a very challenging and time-consuming task. I got to implement what I had learned from the wrangling class and I also developed an intuition about the various functions that can be combined together to perform what might seem as really undoable tasks.

I started the project by uploading the 'Twitter-archive-enhanced.csv' file manually. Then, I downloaded 'image-predictions.tsv' programmatically from Udacity's server using the requests library. Next, I wrote it into image_predictions.tsv. 'twitter_data' was created by accessing and downloading Twitter's JSON data using the tweepy library. Firstly, I extracted a list of tweet IDs from the 'Twitter-archive-enhanced.csv' file, then looped through each ID and queried Twitter's API with the ID to get each tweet's JSON data. Subsequently, I recorded the data in a text file named 'tweet-json.txt', with each tweet's data written in a new line. After the query was completed and all the data was written in the text file, I read the text file line by line, obtained each tweet's information (tweet ID, retweet count, favorite count, and followers count) using the json library, and appended the information into an empty list. Finally, I convert the list of dictionaries to a pandas DataFrame and saved it into 'twitter_data'.

During the Cleaning and accessing, I identified some quality and tidiness issues in the three tables, some of which are:

Quality issues:

1. Most of the values in the in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp columns are null
2. Incorrect data types for tweet_id, timestamp, and rating_numerator in the tweet_archive table.
3. Incorrect data type for tweet id in the image_predictions table.
4. Incorrect data type for the tweet id in the tweet count table.
5. Underscores in place of spaces in the p1, p2, and p3 prediction columns.
6. Inaccurate dog names in the name column and NaN values are represented by the word 'None'.
7. Inaccurate values in the rating_numerator and rating_denominator columns.
8. The values in the columns p1_conf, p2_conf and p3_conf should be percentages instead of proportions.

Tidiness issues:

1. Doggo, floofer, pupper, puppo should be column values but are instead column headers.
2. Joining this table and the Twitter archives table.

After the tidiness and quality issues were solved, the tables were merged together to form "twitter_archive_master.csv" which I used in performing my analysis