WRANGLE REPORT: WE RATE DOGS TWEET ANALYSIS

The wrangling of the WeRateDogs Twitter account required a lot of effort and time. I was able to put what I had learned in the wrangle session into practice, and I also gained an understanding of how different functions may be combined to complete tasks that might initially appear impossible.

I began the project by manually transferring the "Twitter-archive-enhanced.csv" file. Finally, utilizing the requests library, I programmatically downloaded "image-predictions.tsv" from Udacity's server. I then entered it into the file image predictions.tsv. Using the tweepy package, 'twitter data' was produced by gaining access to and downloading Twitter's JSON data. I took the Twitter-archive-enhanced.csv file's list of tweet IDs and looped through each one. And since I couldn't access the Twitter API, I grabbed the "tweet-json.txt" file. Following the execution of the query, I read the text file line by line, using the json library to extract each tweet's details (tweet ID, retweet count, favorite count, and followers count), and appended the data to an empty list. I saved the list of dictionaries as a pandas DataFrame with the filename "twitter data" after converting the list of dictionaries.

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

Quality issues

1. columns like in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp has lost of null values
2. wrong 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 data table.
5. The values in the columns p1_conf, p2_conf and p3_conf should be percentages instead of proportions
6. Wrong dog names in the name column and NaN values represented by the word 'None'
7. p1, p2, p3 prediction columns looks untidy, the underscores should be replaced with spaces
8. Inaccurate values in the rating_numerator and rating_denominator columns.
9. We should drop the retweeted tweets to avoid duplications

Tidiness issues

1. Doggo, floofer, pupper, puppo should be column values but are instead column headers.
2. the image prediction table and the twitter archives table should be joined together rather than separate

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 and visualization