DATA WRANGLING PROJECT

We Rate Dogs Data

I started by importing the necessary libraries like pandas, numpy, seaborn, json, requests and tweepy.

Gathering Data Phase:

My project began with a manual download of the 'twitter-archive-enhanced.csv' file. Then I made a folder called 'image_predictions' and used the requests library to retrieve 'image-predictions.tsv' from Udacity's server. I then saved it as image-predictions.tsv.

'twitter_data' was produced by utilizing the tweepy package to retrieve and download Twitter's JSON data. To acquire the JSON data for each tweet, I first extracted a list of tweet IDs from the 'twitter-archive-enhanced.csv' file, then looped over each ID and queried Twitter's API using the ID.

I then saved the information in a text file called 'tweet-json.txt,' with each tweet's data typed on a separate line. After the query was finished and all of the data was saved to the text file, I read the text file line by line and used the json library to get the information for each tweet (tweet ID, retweet count, like count, and followers count) and appended it to an empty list.

Finally, I saved the list of dictionaries in the 'twitter data' folder as a pandas DataFrame.

Accessing and Cleaning Data:

The three tables were found to have some quality and tidiness concerns. The table below lists the issues and their remedies in detail:

Twitter Archive Dataframe

PROBLEMS	CLEAN(SOLUTIONS)
▪ Keep original ratings with images(no retweets).	▪ Delete retweets by filtering the NaN of retweeted_status_user_id
▪ Erroneous datatypes in these columns (tweet_id, in_reply_to_status_id, in_reply_to_user_id, timestamp, retweeted_status_id, source, retweeted_status_user_id,	▪ Convert tweet_id to str ▪ Convert timestamp to datetime ▪ Convert source to category datatype

retweeted_status_timestamp, doggo, floofer, pupper, and puppo)	
▪ Name more than 745 records do not contain a valid name	▪ Alll names should start with a capital letter.
▪ Source column is in HTML-formatted string in the source column, not a normal string	▪ Extract HTML values from source
▪ doggo, floofer, pupper, puppo columns contain 'None' value.	▪ Nan should be used, I replaced it with Nan values
▪ Error in dog names (e.g a,an,actually) are not a dog's name.	▪ Change error name in dog name to None. Then to Nan values
▪ Drop columns that is not needed for our analysis	▪ Drop columns

Image Prediction Dataframe

PROBLEMS	CLEAN(SOLUTIONS)
▪ p1, p2, and p3 are inconsistent in a way capital and small letters are used in values.	○ Capitalize the first letter of all items in p1, p2, and p3 columns.
▪ Erroneous datatype (tweet_id)	▪ Convert tweet_id to str
▪ Missing images (only 2075 counts out of possible 2356)	▪ Drop rows with missing Images

Twitter API Dataframe:

PROBLEMS	CLEAN(SOLUTIONS)

▪ Erroneous datatype (tweet_id)	▪ Convert tweet_id to str

TIDINESS

PROBLEMS	CLEAN(SOLUTIONS)
▪ Image predictions table should be added to twitter archive table ▪ Twitter data api columns(retweet_count, favorite_count, followers_count) should be added to twitter archive table.	▪ Merge twitter api clean dataframe and image prediction into twitter archive table.

Storing Cleaned Data:

The data is now clean and ready to be analyzed. The master table was saved as twitter archive master.csv.