Data Wrangling Report

Gathering Data Phase

The project was started by downloading the 'twitter-archive-enhanced.csv' file manually. This data was stored in a Folder named "Udacity" in my Google drive which was connected to the working Notebook. 'image-predictions.tsv' dataset was downloaded programmatically from Udacity's server using the requests library.

'twitter_data' was created by accessing and downloading Twitter's JSON data using the tweepy library. First, a list of tweet IDs from the 'twitter-archive-enhanced.csv' was extracted through a for loop and query Twitter's API with the ID to get each tweet's JSON data. The data was recorded 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, the file was read line by line, obtained each tweet's information (tweet ID, retweet count, favorite count,url, created_at, source and retweeted_status) using the json library, and appended the information into an empty list.

Finally, the list was converted to a pandas DataFrame and saved to 'tweet_json'

Assessing and Cleaning Data

Some quality and tidiness issues were identified for the three tables. Details of the issues identified and solutions are in the table below:

QUALITY

TWITTER ARCHIVE TABLE

ISSUES	SOLUTIONS
· Keep original ratings (no retweets) that have images	· Delete retweets by filtering the NaN of retweeted_status_user_id
· drop columns not needed for our analysis	· drop columns

· 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, retweeted_status_timestamp, doggo, floofer, pupper, and puppo)	· Convert tweet_id to str from twitter_archive · Convert timestamp to datetime · convert source to category datatype

· Missing values in 'name' and dog stages represented as 'None'	· Change missing values in dog name to unnamed.
· Some records have more than on dog stage	· Separate the dog stages to know which records have more than one dog stage.
· Source column is in HTML formatted string, not a normal string.	· Extract HTML values from source
· Error in dog names (e.g a,an,actually) are not a dog's name.	· Change error name in dog name to None.
· Some values in rating_numerator not showing proper float values.	· Spot those records and confirm changes made.
· Text column includes a text and a short link.	· Remove hyperlinks in tweets.

IMAGE PREDICTION TABLE

ISSUES	SOLUTIONS
· 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 TABLE

ISSUES	SOLUTIONS
• Erroneous datatype (tweet_id)	• Convert tweet_id to str

TIDINESS

TWITTER ARCHIVE TABLE

ISSUES	SOLUTIONS
• doggo, floofer, pupper and puppo columns in twitter_archive table should be merged into one column named "dog_stage"	• Merge columns into one column named ‘dog_stage’

TWITTER API TABLE

ISSUES	SOLUTIONS
• twitter api table columns(retweet_count, favorite_count, followers_count)	• Merge table with twitter archive table.

IMAGE PREDICTION TABLE

ISSUES	SOLUTIONS
• Image predictions table should be added to twitter archive table	• Merge table with twitter archive table.

Storing Cleaned Data

Now the data set is clean and ready for analysis. I saved the master table to twitter_archive_master.csv

Then I started my analysis.