Wrangling Report

Data Gathering

The data used for this project was gathered through three different methods.

1. Twitter Archive file: This I downloaded manually from the Udacity server and imported to the project file using pandas
2. Tweet Image Prediction file: This was downloaded programmatically using get Request into the project folder and imported using pandas
3. Tweets: I extracted the retweet count favoriterite count fTwittertter using:
    a. Extracting the tweet IDs in the WeRateDogs Twitter archive and store in another file (tweet_id.txt)
    b. Quering the Twitter API for each tweet's JSON data using Python's Tweepy library and store the data in another file (tweet_json.txt)

Data Quality Issues

- Twitter Archive Table

- There are a lot of missing data in the columns

- Some datatypes are not correct and need changing e.g timestamp column

- The missing values are represented by None

- Image Prediction Table

- Text columns are not properly formatted

- Sometimes Lowercase is used for P1, P2, and P3

- Tweets Table

- Date should be Extracted from the Created_at column

- Rename the Created_at column as Timestamp to bridge uniformity

Data Tidiness

- Tweet_id in archive table duplicated in image and tweet tables

- P1, P2, and P3 should be formatted properly in the image table

- Remove html tags from the source column in the archive table

Data Cleaning

- in the twitter archive column, meaniless columns are dropped
- change timestamp to datetime in twitterarchive table
- change tweet_id in twitterarchive table to a string not integer
- change P2_dog to boolean in imageprediction table
- Change Created_at to datetime in tweets table
- Timestamp and Source in Twitter Archive & Tweets Table
- I will remove html tags from the source column
- Timestamp will contain year, month and day only
- Formating p1, p2, and p3 in the image table, dash separating the words was replaced "-" with space (" ").
- After merging the table to create the master table, two columns, timestamp_x and timestamp_y, were found to be the same and one was dropped.
- Lastly, two additional columns were created as it was dimmed fit that they will be required in answering the research questions