DATA WRANGLING REPORT

1. Motivations and Objectives

The motivation for this project is to achieve a wrangled and cleaned WeRateDogs Twitter data that could be used for analysis and visualization to generate insights.

The project main objectives were:

1. Perform data wrangling (gathering, assessing and cleaning) on the provided sources of data.
2. Store, analyze, and visualize the wrangled data.
3. Reporting on
   - data wrangling efforts.
   - data analyses and visualizations.

2. Data

The dataset I perform wrangling on was the twitter data of @dog_rates also known as WeRateDogs.

- The twitter_archive_enhanced: The twitter_archive_enhanced.csv file was provided to me. Specifically, the data contains basic tweet data like tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.
- The tweet image predictions: It presents the breed of dog (or other object, animal, etc.) that is presented in each tweet according to a neural network. This file is hosted on Udacity's servers and was downloaded programmatically using the Requests library from the given URL
- The tweet json dataset: This was gathered using the tweet IDs in the WeRateDogs Twitter archive to query the Twitter API for each tweet's JSON data using the supporting twiiter-api.py code that uses Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data was written to its own line. Then read line by line into a pandas DataFrame with columns like tweet ID, retweet count, and favorite count present.

3. Assessing Data:

After all the data were gathered, they were viewed or displayed for visual assessment and then further assessed Programmatically. The following is the summary of all assessments done:

Quality Issues

1. twitter-archive-enhanced dataset
   - The timestamp column has the object datatype instead of datetime
   - The name column has None instead of NaN and also too many invalid values.
   - The interest only is on original tweet and not retweet
   - Many columns were constituted of over 90% missing values
   - The rating_numerator column contains some wrongly extracted values
   - The doggo, floofer, pupper and puppo columns has None for missing values.
2. image_predictions dataset
   - Too many not really important columns.
   - The prediction dog breeds involve both uppercase and lowercase for the first letter.
3. tweet_json
   - A lot of columns are useless in the dataset as they were mostly made up of missing values.
   - The source column has not been properly extracted as it contains the html tag <a>.
   - The id column name did not match that of twitter achive enhanced

Tidiness Issues

1. The doggo, floofer, pupper and puppo columns should be made into one categorical variable
2. The number of records in image predictions was less than the number of tweets in the twitter-archive-enhanced dataset.
3. The twitter-archive-enhanced, image_predictions, and tweet_json dataframes should be made into a dataset.

4. Cleaning Data:

Using the define-code-test format, the following cleaning were made:

- Converted timestamp to datedime data type using pandas to_datetime function.
- Replaced None and invalid names in the all datasets with np.nan.
- Extracted Original tweets.
- Extracted the rating score correctly and converted it to float
- The html attributes in the source column were removed and left with the actual source values
- Dropped all columns of no interest (retweets) in all datasets
- The doggo, floofer, pupper and puppo columns were made into one categorical variable
- Extracted new breed name and prediction confidence using the highest prediction confidence from all predictions and dropped columns that are of no interest.
- Made all id column name to match

- Combined all the 3 datasets into one dataset and saved as a csv file (twitter_archive_master.csv)