DATA WRANGLING REPORT

1. Goal

The goal for this project is to wrangle WeRateDogs Twitter data, store, analyze and visualize it. The wrangle process entails:

a) Gathering
b) Assessing
c) Cleaning

a) Gathering

This entailed collecting data from different sources to beef up my analysis. The following are the datasets:.

- The twitter_archive_enhanced:
  - This .csv dataset has over 5000 tweets and has columns such as Source, timestamp, text, tweetID etc.
- The image predictions:
  - This .tsv dataset has columns such as tweetID, predictions, prediction confidence etc. This prediction data is a result of a neural network that was used to classify breeds of dogs. This dataset is hosted on Udacity and I programmatically downloaded it using Requests library with the help of the given URL.
- The tweet_extension json:
  - This .json dataset was supposed to be generated with the help of twitter API using the tweetIDs in the WeRateDogs twitter_archive_enhanced dataset. However I had challenges with my developer account, therefore I did not programmatically gather the data as expected. As soon as the account is approved I'll be sure to update that bit and fetch the data programmatically. It has columns such as TweetID, favorite count and retweet count.

b) Assessing Data:

After gathering data, I both visually and programmatically assessed the data to identify quality issues and tidiness issues as required; I was able to gather the following:

Quality Issues

image_prediction dataframe

1. Entries on columns P1, P2 and P3 have mixed upper and lowercases
2. The column names are not as descriptive as they should
3. The image number column should be dropped since it has no significance

tweet_archive dataframe

4. The column source contains "a" html tags
5. Missing data on columns in_reply_to_status_id, in_reply_to_user_id, retweeted_status_timestamp, retweeted_status_user_id & retweeted_status_id
6. Wrong data type for the timestamp column
7. There are retweets on our dataset, yet we don't need retweets for this analysis
8. The column "name" has "None" entries instead of "NaN"

Tidiness Issues

1. The three datasets, tweet_archive, image_prediction and tweet_extension should be combined into one dataset

2. The columns "doggo", "floofer", "pupper" and "puppo" should be converted to one variable

c) Cleaning Data:

After assessing the data and picking out the quality and tidiness issues, I get set to perform the cleaning tasks. But even before that a quick rule of thumb dictates that we first deal with the issue of completeness, that is missing data in any of the tables; then we solve the tidiness issues before we proceed with other quality issues.

Therefore the following is the order in which I performed my cleaning:

1. Used pandas "drop" method to drop in_reply_to_status_id, in_reply_to_user_id, retweeted_status_timestamp & retweeted_status_user_id columns since a bigger percentage of the data is missing
2. Unpivoted the columns using pandas Dataframe "melt" by creating "dog_stage" column and a "stage" column, then Used Drop to drop the intermediate "stage" column and finally dropped duplicated rows that came as a result of unpivoting the columns
3. Used Pandas DataFrame "Merge" function to merge the 3 datasets
4. Used Pandas DataFrame .str.lower() function to convert all the string to lower case
5. Dropped all columns of no interest (retweets) in all datasets
6. Used pandas.DataFrame.rename to rename some of the non-descriptive columns

7. Used pandas DataFrame.drop() to drop the “img_num” column since it had no significance in my analysis
8. Used pandas DataFrame.str.extract to remove the html tag on “source” column
9. Used pandas.to_datetime to convert the timestamp column into datetime format
10. use pandas.DataFrame.query function to extract rows with missing data on retweeted_status_id so that we can remain with the original tweets
    .