Data Wrangling Report

By **Tamizharasu Gangatharan**

The data was gathered from three different sources for this data analysis. With WeRateDogs' access to their Twitter archive tweet data such as tweet ID, timestamp, text is obtained for over 5000+ tweets. Each tweet image was run through a convolutional neural network to analyze the images of dogs and correctly identify their breeds. The convolutional neural network predictions were programmatically downloaded using the Requests Python library as a tsv file. And, using the tweet IDs from the WeRateDogs archive I queried the Twitter API for each tweet's JSON data using the Python’s Tweepy library I stored each tweet’s entire set of JSON data, which will be later used to analyze the tweet, retweets and favorite count.

The data gathering process for this project was time consuming particularly querying the Twitter API. The Twitter API syntax was refered online to make use of the custom generated key to parse the data using 'tweepy'. Appending tweet information to the list took way longer than I expected and I had to redo this everytime I restart the kernel and I had to wait for the 'df_list' to get properly generated so that I can proceed further.

Once I had successfully gathered all the data, I copied the files for the assessment and data cleansing processes. After a clear look at the basic structure that the data has been provied with, I evaluated the dataframes looking for quality and tidiness issues and then started to think of ways to improve the dataset. I began the cleaning process by addressing missing data and mislabeled information. There were 2075 rows in the images dataframe compared to 2356 rows in the archive dataframe. This is because of the inclusion of tweets and retweets without pictures. Several columns had empty values, such as in_reply_to_status,in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp. After removing them, and I also converted columns to a proper data format, primarily changing the timestamp data into datetime object. The text column was then parsed to include gender of the dogs and hashtags.

The columns predicting the dog breed is then condensed. And the unnecessary 'Unnamed: 0' column is ignored. The dog 'stages' statistics had values as columns, instead of one column filled with the values and this was addressed by the end of the cleansing process.

The improved version of the dataset is then saved as 'twitter_archive_save.csv'.