

## **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 favorite count from Twitter using:
  - a) Extracting the tweet IDs in the WeRateDogs Twitter archive and store in another file (tweet\_id.txt)
  - b) Querying 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**

- Some tweets contains retweets and therefore duplicates
- Some tweets are replies and therefore duplicate

#### Twitter Archive Table

- There are lots of missing data in the columns
- Some datatypes are not correct and needs changing e.g timestamp column
- The missing values are represented by None
- Tweet\_id in archive table duplicated in image and tweet tables
- Remove html tags from the source column in the archive table

## Image Prediction Table

- Text columns are not properly formatted
- Sometimes Lowercase is used for P1, P2, and P3
- P1, P2, and P3 should be formatted properly in the image table

## Tweets Table

- Date should be Extracted from the Created\_at column
- Rename the Created\_at column as Timestamp to bridge uniformity

## Tidyness

- doggo, floofer, pupper, and puppo are categorical variable, and can be combined into one column.
- the 'text' column has more than one information in a single column (tweet and URL)
- Merge the 3 tables together

## Data Cleaning

- Deleted rows that have retweets 'retweeted\_status\_id'
- Dropped any tweet with 'in\_reply\_status\_id'
- Replaced inaccurate dog names with an empty string
- in the twitter archive column, meaningless 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
- Standardized Dog ratings by Converting numerator and denominator datatypes to float, remove non expected value of denominator and numerator anything different of 10
- Formating p1, p2, and p3 in the image table, dash separating the words was replaced "-" with space (" ").
- doggo, floofer, pupper, and puppo are categorical variable, and are combined into one column.
- 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