Reporting: wragle_report

- **Gather:** Three sets of data was gathered from different sources, the ‘twitter_archive_enhanced.csv’ file was on hand and easily downloaded from the archives, the ‘image_predictions.tsv’ file was downloaded programmatically using the ‘Request’ python library from the servers and finally the ‘tweet_json.txt’ file was gotten from twitter using the twitter API (Application Programming Interface) and the Tweepy python library.
- **Access:** Visual assessment was done by displaying the first 5 entries for each dataset using the .head() function and displaying 5 random entries for each dataset using the .sample() function. The programmatic assessment was done using the .info() function to display the column names, number of columns, the number of non null values (observations), the data types of each column and total number of observations in each dataset. The .value_counts() function was used the display the number of non null values in the ‘doggo’, ‘floofer’, ‘pupper’ and ‘puppo’ columns of the ‘twitter_archive_enhanced.csv’ dataset which is named ‘tweet_arch’ after reading it into jupyter notebook. During the assessment the following issue were discovered:

  Quality Issues

  tweet_arch table

  - retweet data is included in dataset which can skew analysis on original data
  - 'retweeted_status_id', 'retweeted_status_user_id'and 'retweeted_status_timestamp' have only 181 entries
  - 'in_reply_to_status_id' and 'in_reply_to_user_id' have only 78 entries
  - 'tweet_id' in all 3 datasets are integers
  - 'expanded_urls' has a few missing entries
  - 'timestamp' is a string type

- 'rating_denominator' has a 0 value

image_pred table

- predictions don't march in all 3 predictions column

Tidiness Issues

- The tweet_arch table and the tweet_json (read from ‘tweet_json.txt’) table have the same type of observational unit
- The dog stages in the tweet_arch table are in seperate columns

- **Cleaning:** A copy of all three datasets was made to prevent loss and allow for recovery if a mistake was made. The cleaning process was done using the Define-Code-Test framework. Beginning with quality issues, the tweet_arch dataset contained retweet data which was removed, the 'retweeted_status_id', 'retweeted_status_user_id' and the 'retweeted_status_timestamp' columns had only 181 obsevation to begin and after removing retweet data from the tweet_arch dataset they had only null values so, they were dropped, the 'in_reply_to_status_id' column and the 'in_reply_to_user_id' column of the tweet_arch table had only 78 obervations and was dropped using the .drop() method likewise, the 'expanded_urls' had about 58 missing row which was filled using the .fillna() method and values from the 'expanded_url' column of the tweet_j table. The ‘timestamp’ column was converted from a string to a datetime type. The tweet_id column in all 3 dataset (ie tweet_arch, image_pred and tweet_j) had the integer data type and was changed to an object data type. The rating_denominator column of the tweet_arch dataset had a 0 value which was drop. The image_pred table had discrepancies between the dog prediction values for it 3 predictions hence, only observations with all 3 predictions as ‘True’ was selected. Thereafter, tidiness issues was handled by merging the tweet_arch and tweet_j tables using the merge() function, converting the doggo, floofer, puppo and pupper columns into 1 column using the melt() function and removing duplicates that came about as a result of the melt() function.