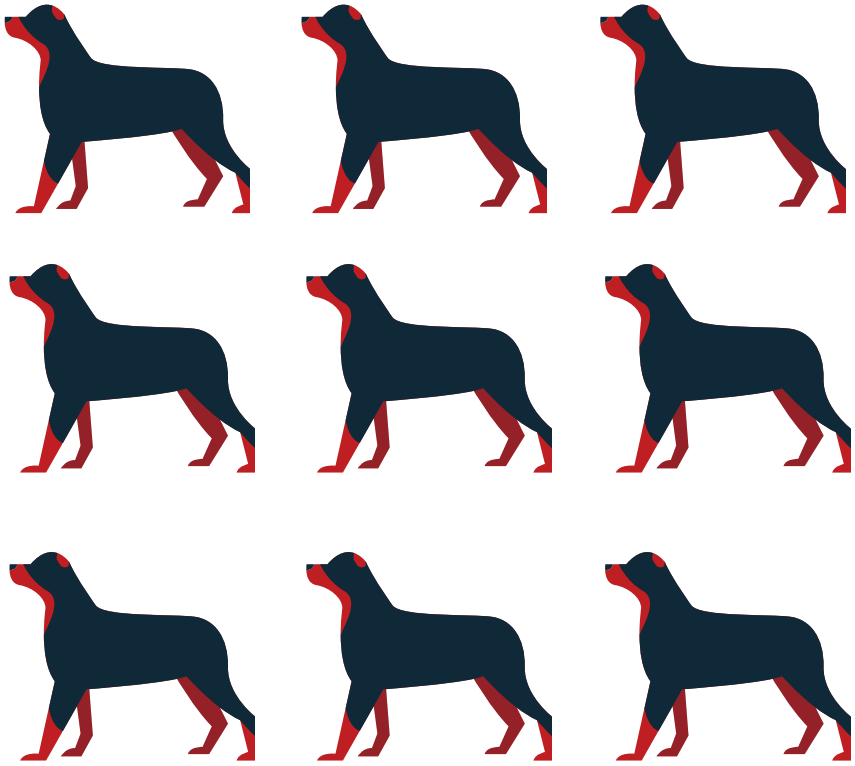

DATA WRANGLING WERATEDOGS TWITTER ARCHIVE & API

Wrangle Report by Duyen Nguyen

For this data wrangling project, data is gathered from three different sources: WeRateDog Twitter archive under the name *twitter-archive-enhanced.csv* provided by WeRateDog, image prediction file named *image_predictions.tsv* from a neural network hosted on Udacity's server and queries of Twitter API contained in a file named *tweet_json.txt*. All of these datasets are acquired with the Requests and Tweepy packages. Once downloaded, all three datasets are loaded to a jupyter notebook using Pandas and the built-in function *open()*.

After all three dataframes are loaded under the names t*witter_archive_og, image_predictions_og, and tweets_original,* I did a quick visual examination of each dataframe and found issues with missing data in three columns and column name in *tweets_original*. Alsmot all data are missing from the three columns *contributors, coordiantes, and place* so I decided not to include these three columns in the final dataset. Column *id* has a fine name, but for the purpose of merge these three datasets into one master dataset, I renamed this column into 'tweet_id' so merging can be done upon this column. The only columns in this dataset I'm interested in using for analysis is *lang, retweet_count and favorite_count*. Therefore, a copy of this dataframe containing *tweet_id, retweet_count, favorite_count and lang* , named *tweets_stats*, was created. Column *tweet_id* was converted to object type since ID of each tweet, even though is numeric, is unique and is not used for any numeric calculation since this is nominal data.

image_predictions is a clean dataset with only one issue in the data type of *tweet_id* column, which is integer. Per the same reasoning with the *tweet_id* column in *tweets_origina*l, this column was also converted to string.

twitter_archive_og is the biggest dataset with 2356 entries. Via visual examination, an issue in tidiness was found. Three columns *doggo, pupper and puppo* contain values of stages of the dog(s) in each tweet: 'doggo', 'pupper' and 'puppo'. These three columns were combined into one single column. If a tweet has more than one dog, thus more than one dog stage, the stages are combined into one single value i.e. 'doggo/pupper' or 'doggo/puppo'. The column *floofer*, in my opinion, reflects a dog's characteristic rather a dog's stage so it was kept separate from *doggo, pupper and puppo*. However, since the values in this column are 'floofer' and 'None', which are binary, I replaced these values with the boolean 'True' and 'False'. Via programmatic examination, other quality issues were found in the *twitter_archive_og* dataset. Datatype is an issue with columns *tweet_id, in_reply_to_status_id, in_reply_to_user_id,* which were originally integer and were eventually converted to object type. Some names in the *name* column were extracted incorrectly so a list of invalid names was extracted manually from a list of all dog names. These invalid names were replaced by np.nan. Retweets that contain 'RT @' were removed from the dataframe, along with columns related to retweets i.e. *retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp.* Finally, column *timestamp* was object type so I converted it to datetime type for easier extraction of day and month, which will be used for analysis and visualization.

Copies of the three dataframes were created before any cleaning step took place to ensure non-destructive editing to the original data. The three cleaned dataframes *tweets_stats, image_predictions and twitter_archive* are then merged upon *tweet_id* to create a master dataframe, which effectively removed tweets beyond August 1, 2017 with no image prediction data. This dataframe is written to a csv file named *twitter_archive_master.csv*