INTERNAL DOCUMENT (THE WRANGLE PROCESS)

The dataset that I will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

GATHERING DATA

A total of three datasets were gathered for this analysis:

- **twitter-archive-enhanced.csv**: this dataset was downloaded manually from https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv
- **image_predictions.tsv:** this dataset was downloaded programmatically using the url: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
- **tweet._json.txt:** this dataset was downloaded manually due to my unsuccessful application of Twitter Developer account from https://video.udacity-data.com/topher/2018/November/5be5fb7d_tweet-json/tweet-json.txt

ASSESSING DATA

The three datasets were assessed both visually and programmatically to find out quality and tidiness issues that need to be cleaned. Below are some of the issues spotted during the assessment phase of the data wrangling process.

- There were a few tweets in the “image_predictions.tsv” dataset that did not have images of dogs
- There were 66 duplicate “jpg_url” in the “image_prediction.tsv” dataset
- The datatype for timestamp in “twitter-archive-enhanced.csv” was incorrectly formatted
- The dataset contained retweets and replies
- A few columns in the dataset were not going to be useful for my analysis
- There were four (4) separate columns that represented dog stage
- Nine separate columns represented predictions of dogs and the confidence levels of each prediction

CLEANING DATA

After spotting the above quality and tidiness issues, and attempt was made to clean the datasets. Copies of the original datasets were made before the cleaning process begun. The steps below outlines how the quality and tidiness issues stated above were cleaned.

- The tweets without dog images were dropped using pandas.drop()
- Dropped the 66 duplicated jpg urls
- Converted the timestamp datatype from string to datetime using pandas.to_datetime()
- Dropped the retweeted and reply to tweets
- Columns from the three datasets that were not useful for my analysis were dropped
- Converted the four separate columns that represent the dog stages into one column
- Using numpy.select(), the nine separate columns for dog predictions were converted into two columns: prediction and confidence
- Merged all three datasets into one and saved as a csv file.