Project: Wrangling (and analyzing and visualizing) the tweet archive of Twitter user @dog_rates, also known as WeRateDogs.

About:

The dataset contains 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.

Goals:

Step 1: Gathering data

Step 2: Assessing data

Step 3: Cleaning data

Step 4: Storing data

Step 5: Analyzing, and visualizing data

Gathering Data

I gathered all three parts of the required data from the following sources:

- **Enhanced Twitter Archive**

The WeRateDogs Twitter archive provided contains basic tweet data for all 5000+ of their tweets including each tweet's text, rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo). Of the 5000+ tweets, there are only 2356 with ratings.

- **Additional Data via the Twitter API**

Retweet count and favorite count are two of the notable columns.

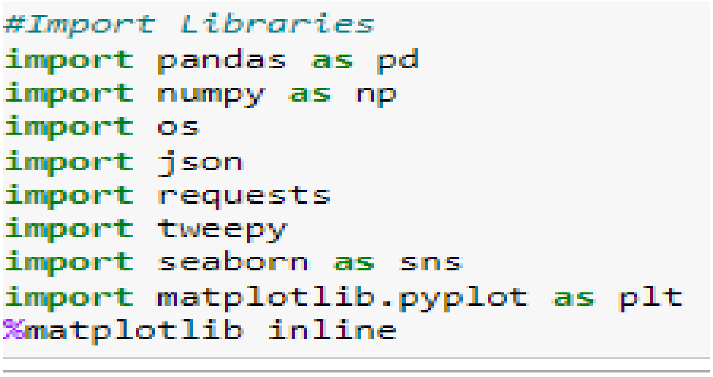
Image Predictions File

A table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images). *NOTE*: *tweet_id is the last part of the tweet URL after "status/", p1 is the algorithm's #1 prediction for the image in the tweet, p1_conf is how confident the algorithm is in its #1 prediction, p1_dog is whether or not the #1 prediction is a breed of dog, p2 is the algorithm's second most likely prediction , p2_conf is how confident the algorithm is in its #2 prediction, p2_dog is whether or not the #2 prediction is a breed of dog, etc*.

Wranging process:

Gathering the data

I imported the needed libraries



```
#Import Libraries
import pandas as pd
import numpy as np
import os
import json
import requests
import tweepy
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

- Then downloaded the data provided in the ‘twitter_archives.csv” file and loaded it into a dataframe.
- I programmatically downloaded the image predictions file hosted on udacity servers via the link ‘https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv' and loaded it into a pandas dataframe
- Since i found it difficult to access twitter’s API i made use of the file provided and downloaded the json text file which contained the favourite counts and retweet counts of the tweets and loaded it into the “tweet_statistics” dataframe.

Assessing Data:

I assessed the data visually and programmatically using the .info() method on the tables and recorded the following quality and tidiness issues:

quality Issues

Re:archive Table

- The timestamp data type is object instead of date time
- The tweet_id data type is integer instead of string
- in_reply_to_status_id has a large number of missing values
- in_reply_to_user_id has a large number of missing values
- retweeted_status_id has missing values
- retweeted_status_user_id has missing values
- retweeted_status_timestamp has missing values

Re:image_prediction Table

- tweet_id datatype should be string and not integer

Re:tweet_statistics Table

-
- the id datatype should be string and not integer
- the id column name does not match that of the other dataframes

Tidiness issues

- doggo,pupper,puppo,floofer should be in a single "dog_stage" column
- tweet_statistics should be included in archive datafraame
- extranneous columns are not needed

Cleaning the data:

Broke down the cleaning of the dataset into the Define,Code and Test categories and then performed cleaning operations on all the issues that were identified in the assessing stage.

Storing the data:

I merged the three tables into one and then stored it as ‘twitter_master_archive.csv”