ACT_REPORT

The goal of this project was to wrangle WeRateDogs Twitter data using Python, starting with collecting the data, cleaning it, analysing and then visualizing trends from the data.

@dog_rates also known as WeRateDogs is a twitter account with over 9 million followers that has received international media coverage. The account that rates people's dogs with humorous comments about the dog. These ratings almost always have a denominator of 10. The numerators, are almost always greater than 10. Why? **Because "they're good dogs Brent."**

Let's get a little technical, what is *Data Wrangling*?

- It is the process of cleaning, organizing, and transforming raw- often messy and complex- data into the desired format for use and analysis, visualization and in turn decision-making. Every minute the amount of data and data sources available are rapidly increasing, it is extremely important now more than ever for data analysts to know how to organize these large amounts of available data for analysis.
- Wrangling occurs in 3 stages: **Gathering data, Assessing data and Cleaning data**

Gathering

The data for this project was in three different formats and they were gathered in three different methods as mentioned below:

- Twitter Archive: This file was provided as twitter_archive_enhanced.csv by Udacity and downloaded from their server. The archive contains basic tweet data (tweet ID, timestamp, text, etc.) for over 5000 tweets from the WeRateDogs twitter account as it stood on August 1, 2017.
- Image Prediction File: This file contains predictions for the dog breeds in each tweet after being run through a neural network. The image predictions were stored in the image_predictions.tsvfilw which was hosted on Udacity's servers and I downloaded it programmatically using the requests library

- JSON File from Tweeter API: Using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data was written to its own line. Then I read the .txt file line by line into a pandas data frame, extracting the retweet count and favourite count of each tweet..

Assessing and Cleaning

During assessment I looked for quality and tidiness issues.

I visually assessed the data in pandas data frames, excel spreadsheets and text editors after which I programmatically assessed the data frames.

- I found a few quality issues and tidiness issues which I cleaned using the 3 step cleaning processes
    - Define: data issues observed during assessment are converted into cleaning tasks.
    - Code: the cleaning task is converted into working code.
    - Test: code is written to test cleaning efforts to make sure they worked.

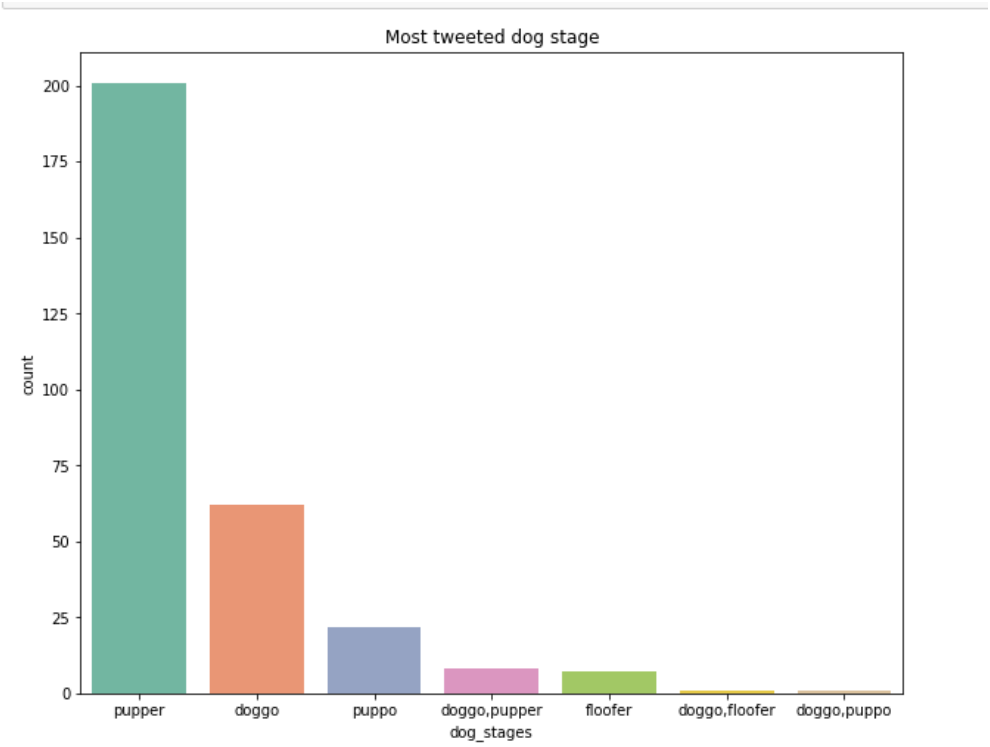
After cleaning the data, I combined the three cleaned datasets using the duplicated tweet_id attribute, and then saved to a master dataset in a CSV file named twitter_archive_master.csv.

- The assessing and cleaning processes are documented in detail in my wrangle report.

Analysis

Most tweeted dog stages

The most tweeted dog stage was the **Pupper stage with 198 tweets**.



Pupper: A pupper is a small doggo, usually younger. Can be equally if not more mature than some doggos.

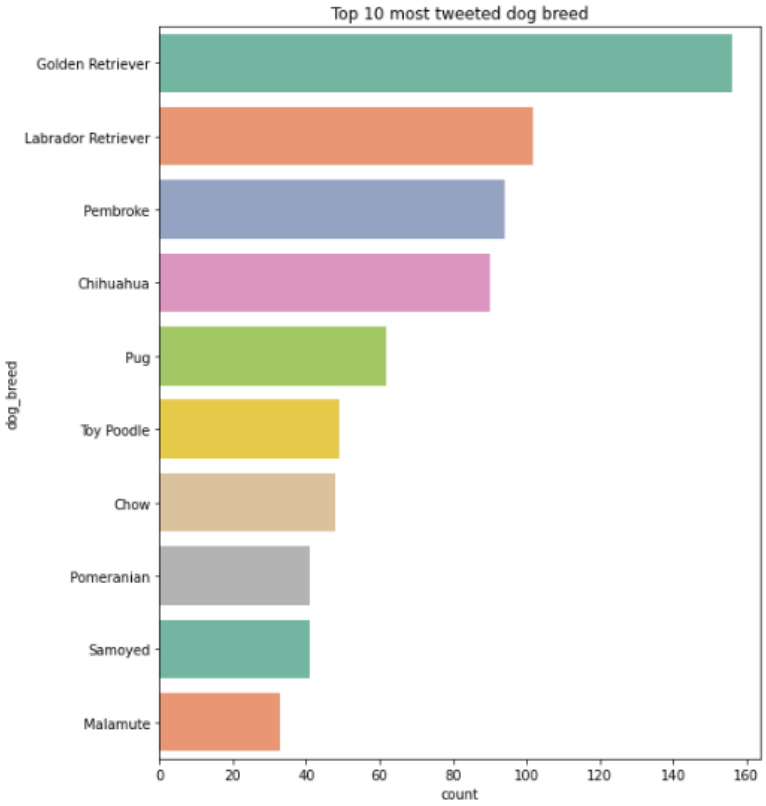
Doggo: A doggo is a big pupper, usually older. It appears to have its life in order. Probably understands taxes and whatnot.

Puppo: A puppo is a transitional phase between pupper and doggo. Easily understood as the dog's equivalent of a teenager.

- ## Most tweeted dog breeds

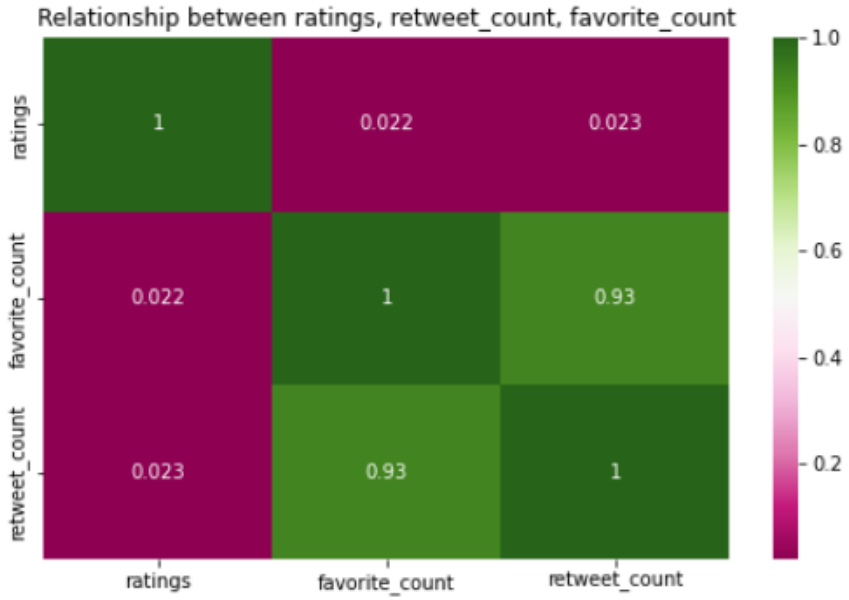
In my analysis, I controlled for outliers by only using dog breeds that had as many or more tweets than the mean of approximately 14

The most tweeted dog breed is the **Golden Retriever**



- ## Relationship between Ratings, Retweet and Favorite count

There is a strong positive correlation (0.93) between Favorite count and retweet count, on the other hand there is no relationship between ratings and favourite count and ratings and retweet count



- Dog breed with highest rating **Pomeranian**
- Dog breed with highest favourite count **French Bulldog**
- Dog breed with retweet favourite count **French Bulldog**
- Dog stage with highest rating **Puppo**
- Dog stage with highest favourite count **Puppo**
- Dog stage with retweet favourite count **Doggo**

- The Top 5 most popular dog names in order are **Charlie, Cooper, Oliver, Lucy, Penny.**

Conclusion

Ratings doesn't impact the popularity of tweets. The most important factors for popularity are the Dog Breed and Dog Stages. If the dog breed is French Bulldog and the dog stage is either doggo or puppo, the tweet has higher chances of extreme popularity through favourites and retweets.