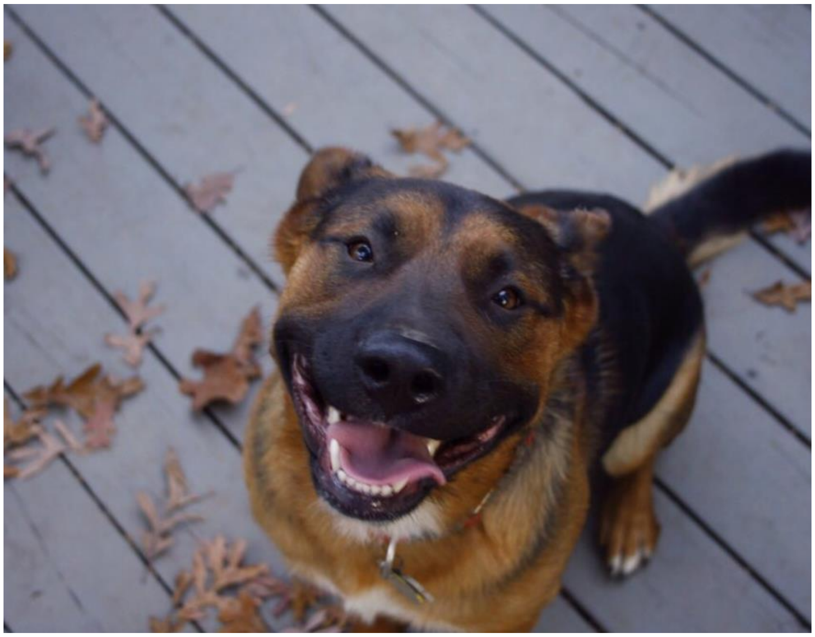


WE RATE DOGS

Udacity

Tash Bettridge

Data Wrangling Report

Introduction

The goal of this project is to wrangle WeRateDogs Twitter data to analyse the findings retrieved from the data from Twitter by analysing trends and creating data visualizations.

Data

The data will be analysed from the following:

- Enhanced Twitter Archive
- Data retrieved by the Twitter API
- Image Predictions File

Process

The data wrangling steps that were involved in this project were:

- Step1. Gathering data
- Step2. Assessing data
- Step3. Cleaning data

Data Analysis, storing data and data visualization of the wrangled data

Step1. Gathering Data

In step 1, I gathered the data from manually downloading the twitter-archive-enhanced.CSV from Udacity.

The second dataset was programed downloaded from the Udacity server using the request:

data = requests.get('https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv')

The third dataset was gathered from the Python request library called image-predictions.tsv.

Finally, the last dataset using the Tweepy library and stored as a JSON file.

Step2. Assessing Data

In this step, I assessed the data visually using the head(), info(), describe() and programmatically find quality and tidiness issues. The following were found during this analysis:

Quality Issues

- There are some disaprities in the tweet_data column names where some columns have geo ... quoted_status quoted_status_id, quoted_status_id_str or quoted_status_permalink with None or NaN
- The source column contains the html tag <a href="http://twitter.com/download/iphone" r...

Dataframe twitter_archive Table:

- There are retweeted_user_id and retweeted_status_id column: there are some retweets
- rating_numerator column has values less than 10.
- There are some disparities in the rating_denominator column which has values other than 10.
- There are some records that have more than one dog.
- There are some retweets and duplicates.
- There are some html tags in the source column: href=""http://twitter.com/download/iphone"" rel=""nofollow"">Twitter for iPhone
- There are duplicates in the dog name table. There maybe an error since "a" is counted 55 times.
- There was also "an" as the the dogs names in the table.

Tidiness¶

- The image_predictions and the tweet_data should be combined with the twitter_archive table this would make the data cleaner to read.
- doggo, floofer, pupper and puppo columns in twitter_archive table should be merged into one column.
- Columns that contain duplicate information of entities should be extended.
- The breed column should be added in twiiter_archive table as its values are based on p1_conf and p1_dog columns of img_df image_predictions table.

Step3. Cleaning Data

Prior to cleaning the data, I created copies of each dataset. Then I tried to fix quality issues that were showing missing, duplicated or incorrect data. Some issues were fixed in this step, some records whose rating_denominator which was less than 10 and divisible by 10 were supplying an invalid output. That had to be cleaned.

There were also duplicate names and names like "a" which was represented 55 times and "an" which was represented as a name 6 times, had to replace with nan instead.

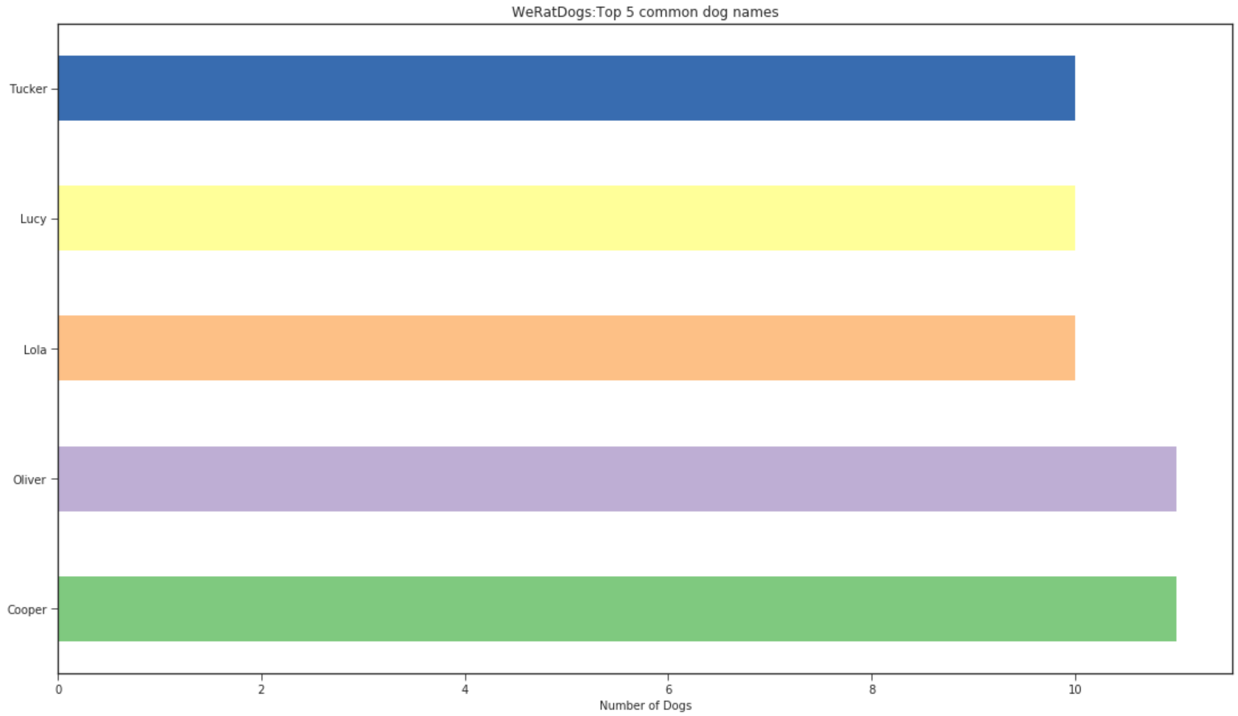
After the tidiness issues were fixed, the datasets were merged to fix the remainder quality sets. I could have gone further by breaking down the analysis by dog breed as well.

Storing data

After cleaning the dataset, the dataset was stored in a CSV file.

Data Visualization

What are the top 5 most common dog names?

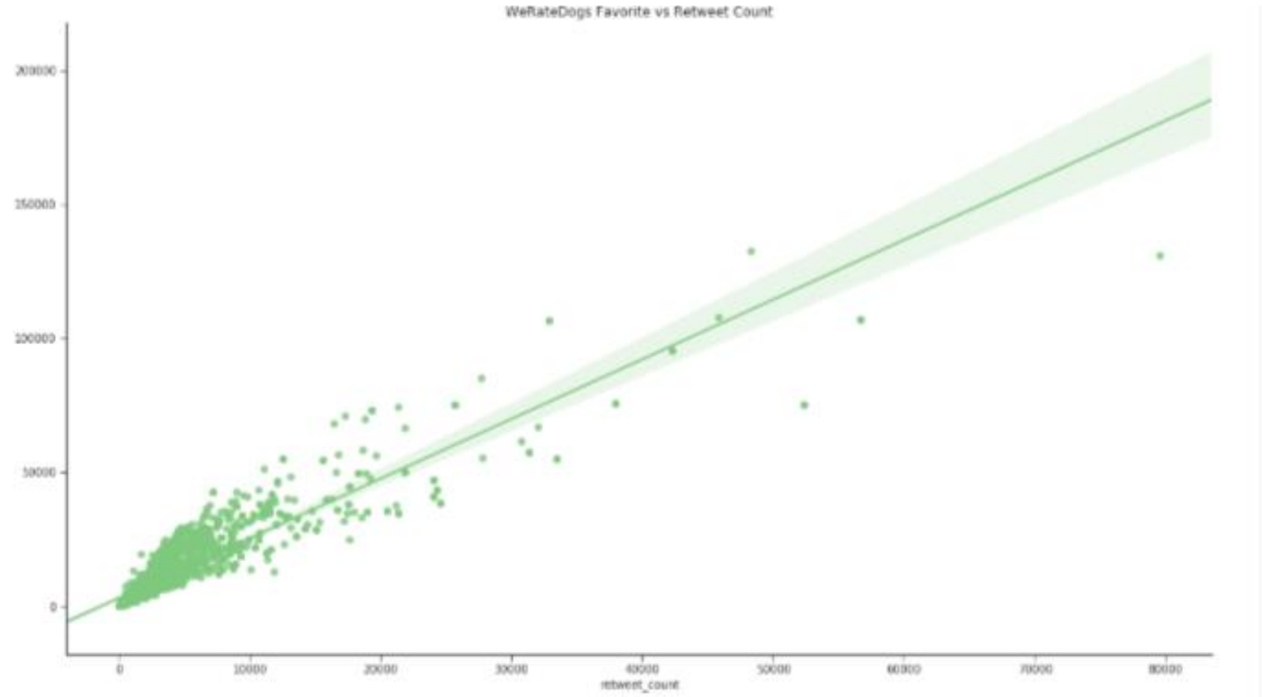


Based on the findings:

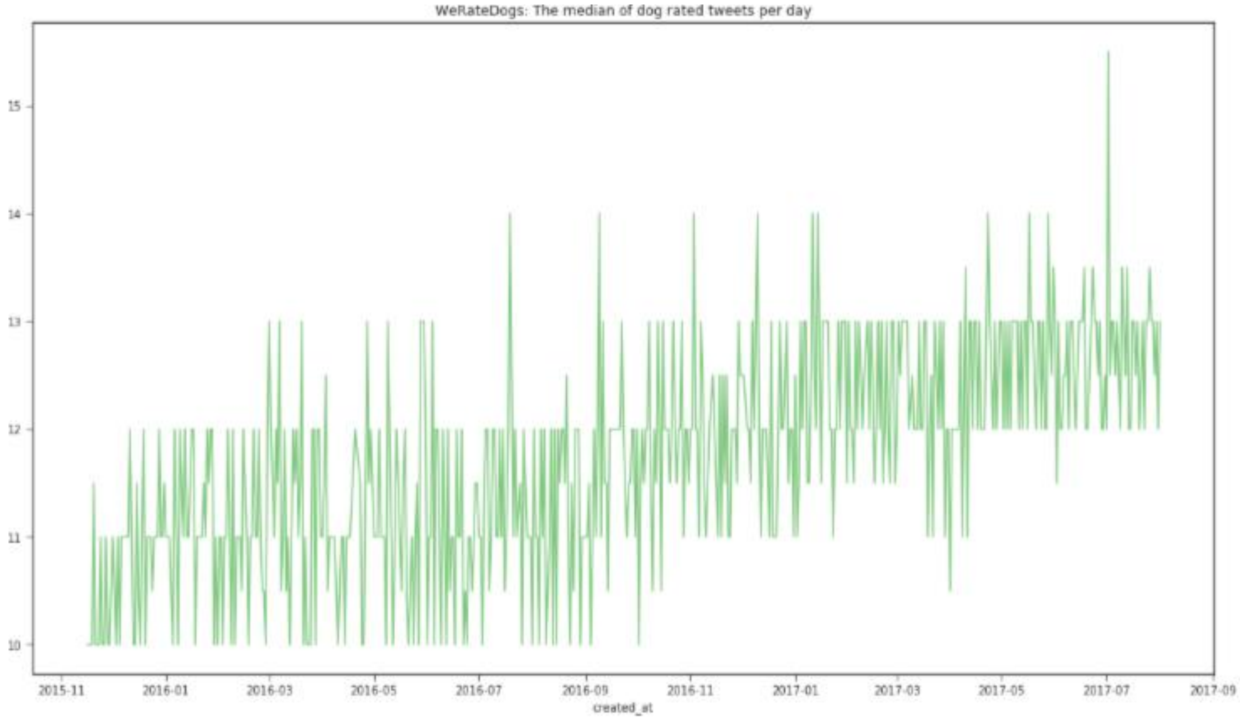
Charlie is the most common dog name with Charlie 12

- Followed by Oliver (11), Lucy (11), Cooper (11), Lola (10) and Tucker (10)

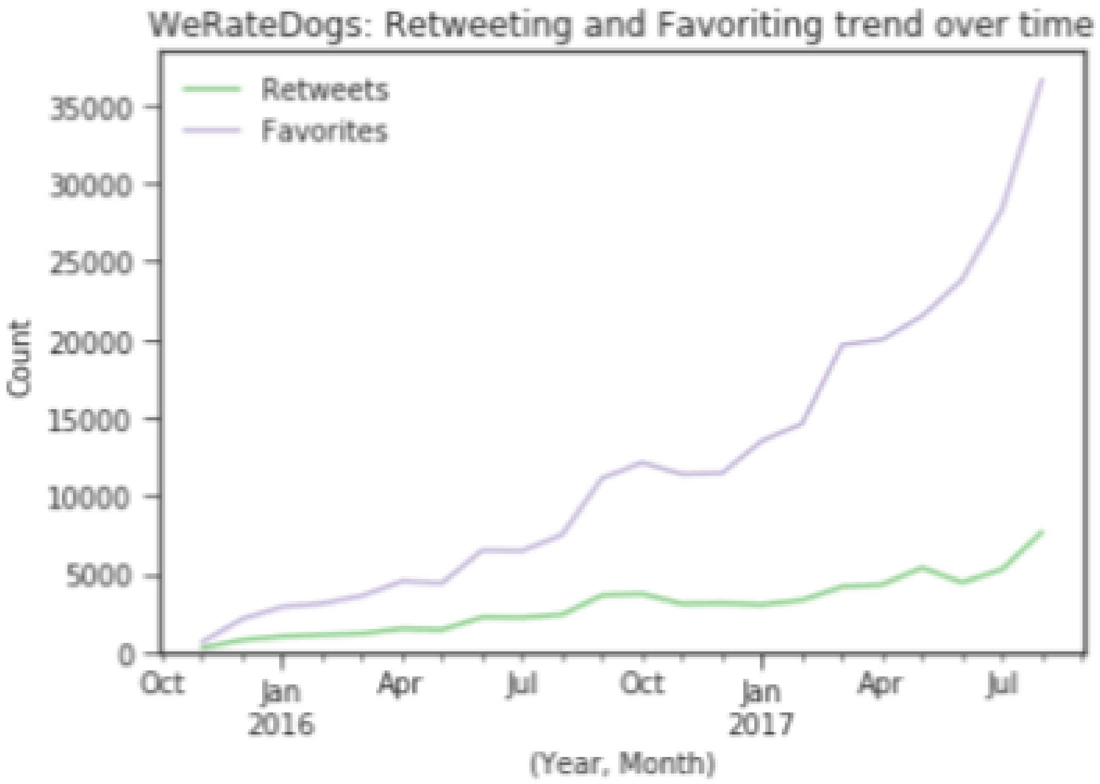
Are there similarities in favourite and retweet counts?



What is the median of dog rated tweets per day?



What are the retweeting and favourite trends over time?



References

- https://matplotlib.org/api/_as_gen/matplotlib.pyplot.subplots.html
- https://grantpatience.com/2019/08/29/project-who-are-the-goodest-doggos-wrangling-analysing-weratedogs-tweets-to-find-the-goodest-floofs/
- https://stackabuse.com/reading-and-writing-json-to-a-file-in-python/
- https://stackoverflow.com/questions/47925828/how-to-create-a-pandas-dataframe-using-tweepy
- S. Github: YashMotwani,https://github.com/YashMotwani/We-Rate-Dogs-Data-Wrangling/blob/master/act_report.pdf