Analyzing and visualizing WeRateDogs Dataset

For this project, the goal was to capitalize on Twitter's vast amounts of tweet dara, utilizing the Twitter API to exploit the Twitter data of the user @dog_rates. Here is some background information just in case you didn't hear about WeRateDogs. WeRateDogs is a very popular Twitter account with over 4 million followers and has received international media coverage. It rates people's dogs with humorous comments.

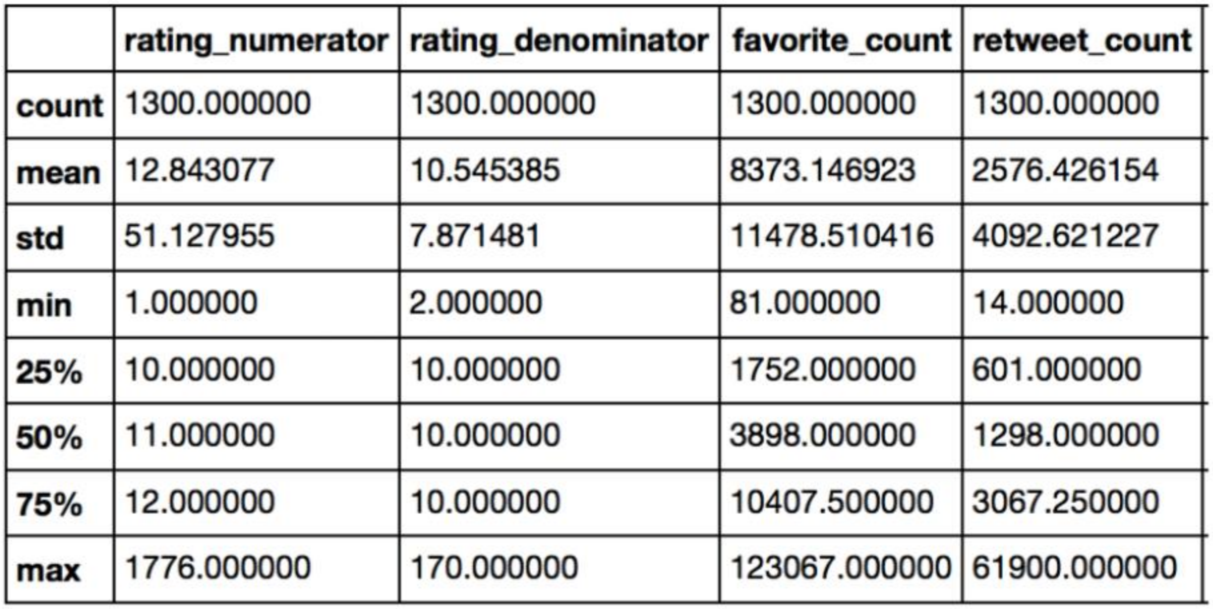
For this analysis, I gathered data from three different sources. WeRateDogs gave Udacity exclusive access to their Twitter archive for this project in the form of a csv file. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1,2017. Each tweet image was run through a convolutional neural network with the purpose of analyzing the images to correctly identify the dog breeds. This can be downloaded using the Requests Python library as a tsv file. And finally, using the tweet IDs from the WeRateDogs archive I queried the Twitter API for each tweet's JSON data using the Python's Tweepy library I stored each tweet's entire set of JSON data.

Before diving into the analysis, I began answer myself some basic questions such as

1) What is the most common dog names in this dataset?
2) what is the comment for the highest rating?
3) What is the average rating for all of the dogs?

4)the relationship between count of retweets and count of favorites
5)the trend in popularity over time of the account
6)analysis of the rating scores over time

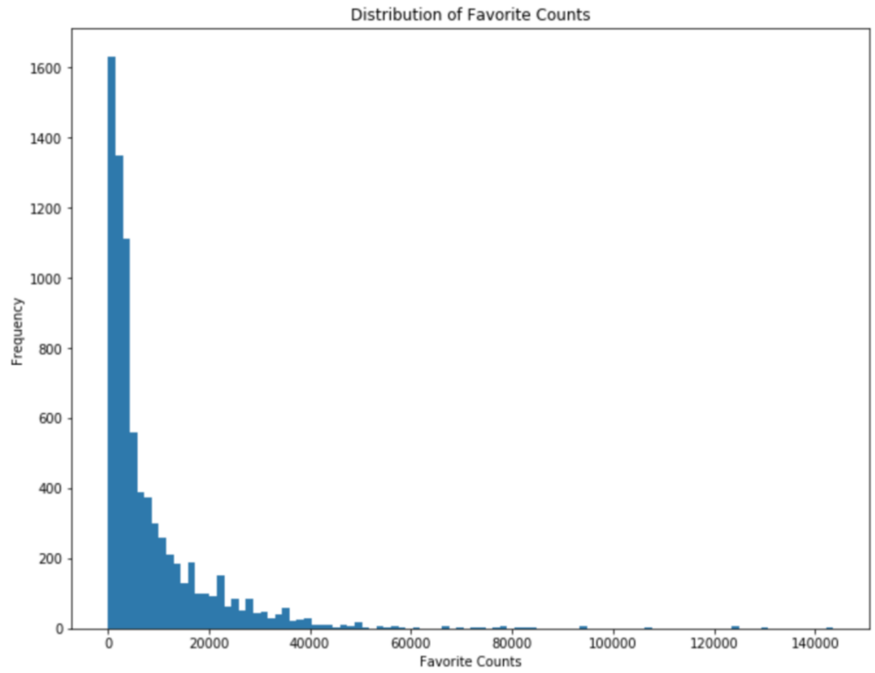
I figure out the most common dog names in this dataset, not including the NaN values, are Oliver, Winston, Tucker and Penny.

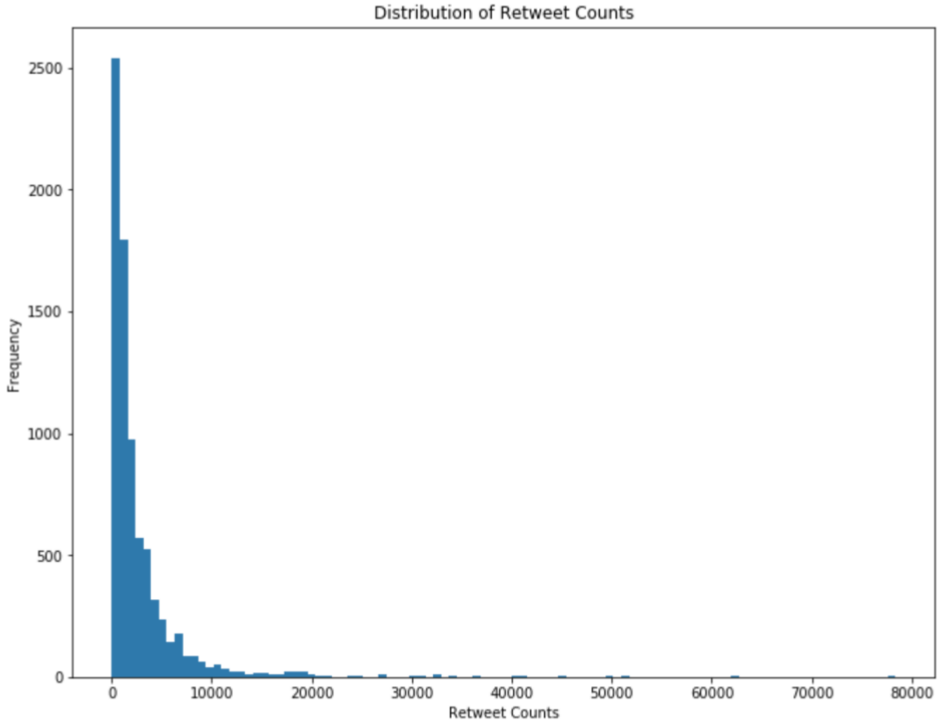


	rating_numerator	rating_denominator	favorite_count	retweet_count
count	1300.000000	1300.000000	1300.000000	1300.000000
mean	12.843077	10.545385	8373.146923	2576.426154
std	51.127955	7.871481	11478.510416	4092.621227
min	1.000000	2.000000	81.000000	14.000000
25%	10.000000	10.000000	1752.000000	601.000000
50%	11.000000	10.000000	3898.000000	1298.000000
75%	12.000000	10.000000	10407.500000	3067.250000
max	1776.000000	170.000000	123067.000000	61900.000000

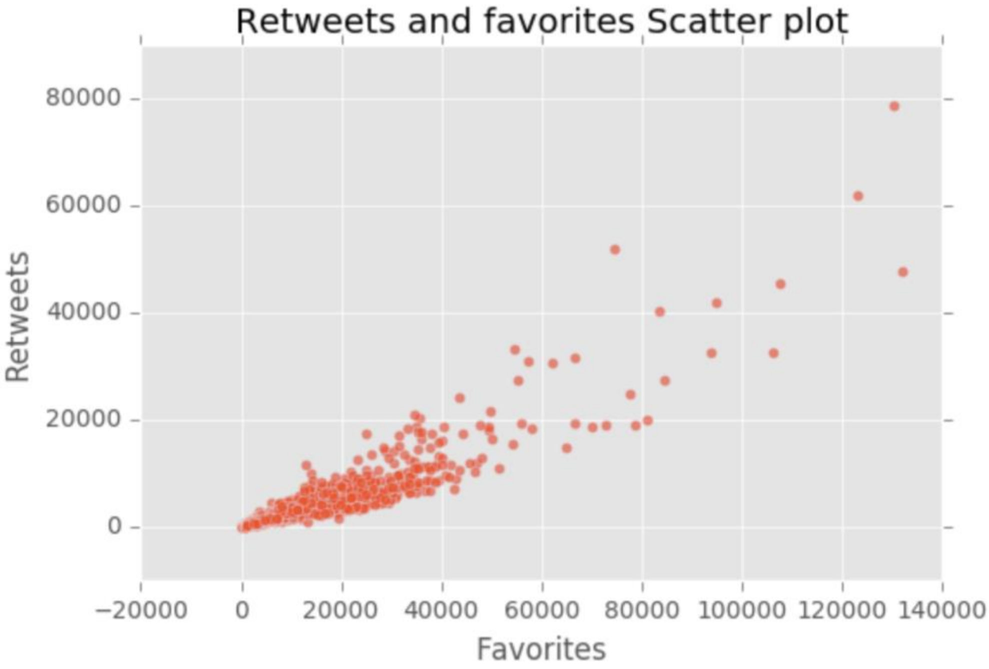
The average rating is 12.84. And the max of rating in this dataset is 1776. What happened for this rating? This outlier obviously is too abnormal. After a series of investigation, I found out that the outlier dog named Atticus. The Tweet was sent on July 4,2016.

In my analysis, I recognized a trend in the favorites and retweets over time. This trend increased, presumably as the account became more popular. In the chart below, we see an upward trend for both retweets and favorites. There is a more noticeable increase in the number of favorites when compared to retweets as well as several large outliers in favorites for extremely popular tweets.

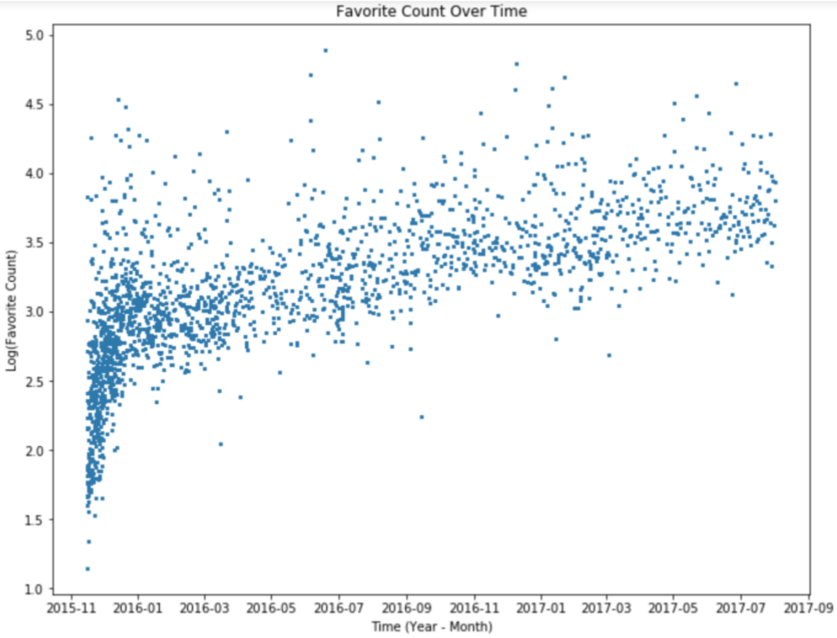


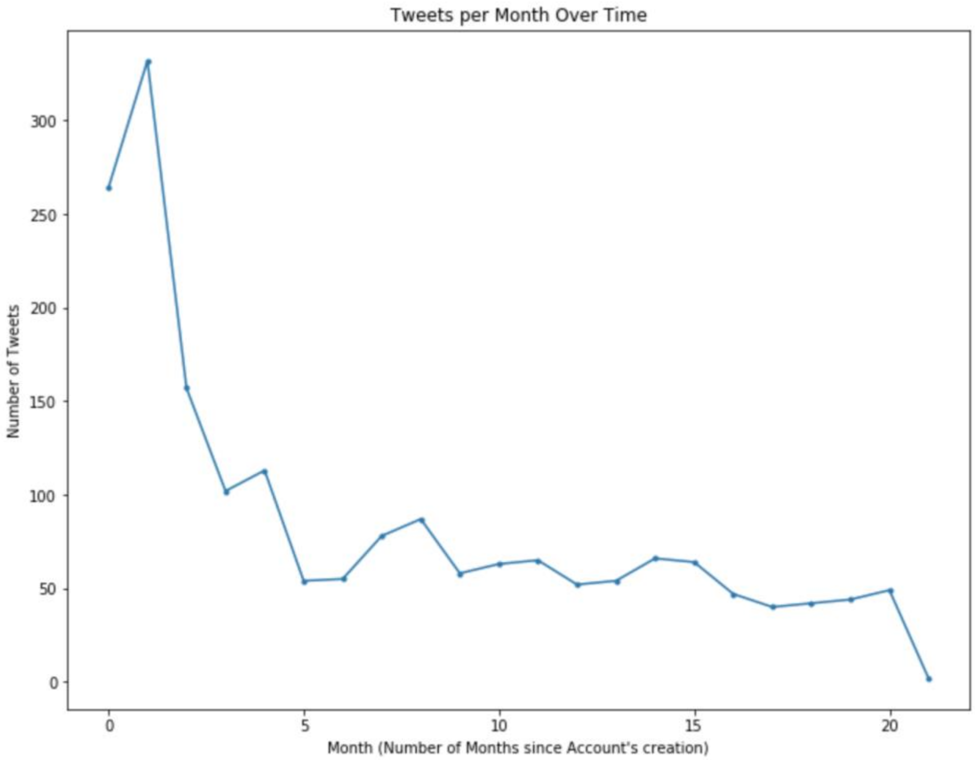


The distribution of retweet_count and favorite_count look similar, but retweet_count is lower than favorite_count. I would guess that these numbers are related to timestamp.



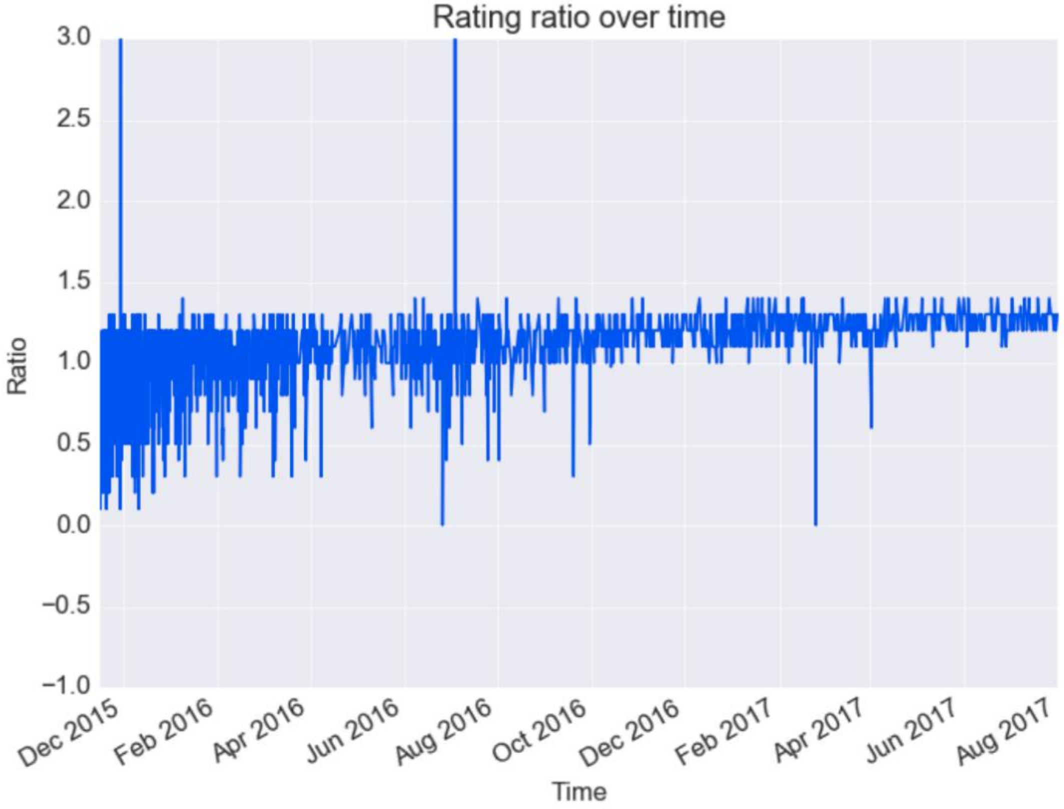
The graph seems show that if the count of retweet is high, then the count of favorites goes high as well. More retweet means more favorites.



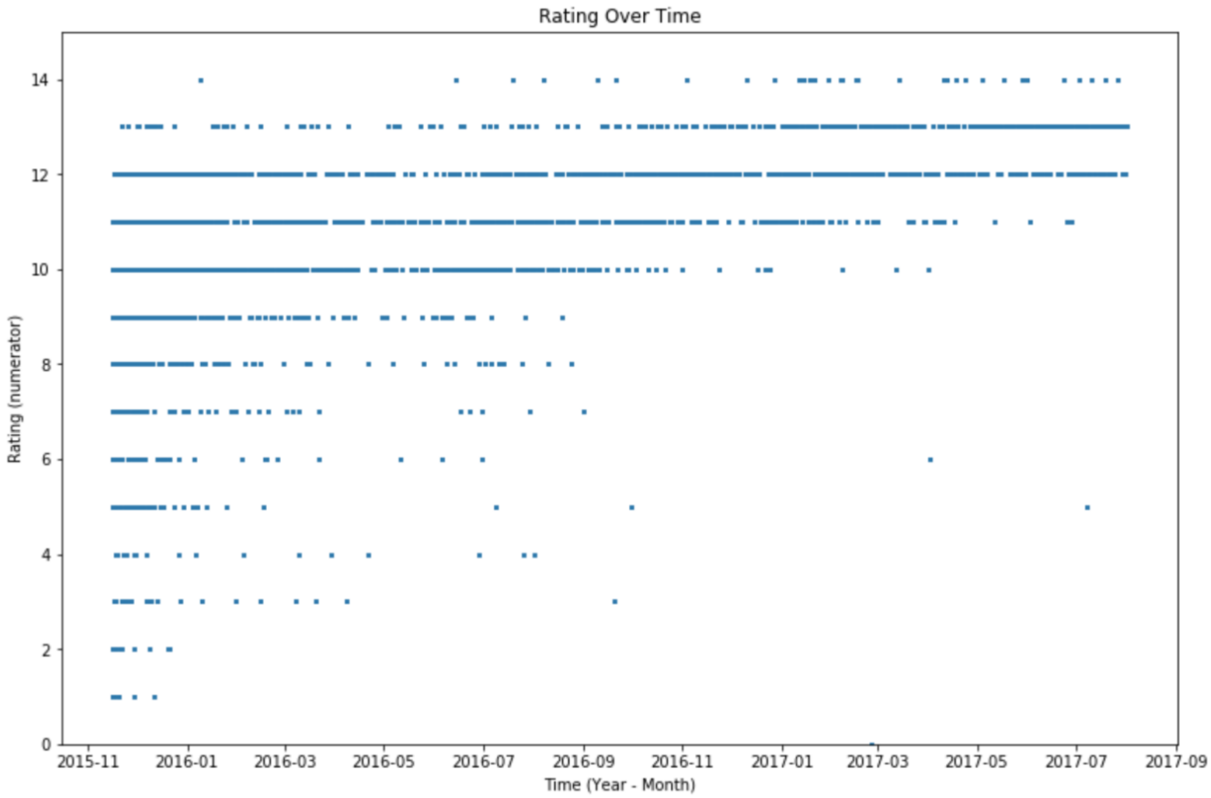


dog_rates tweets much more popular within the first three months after opening this account.

dog_rates started to gain popularity very quickly upon first opening the account.



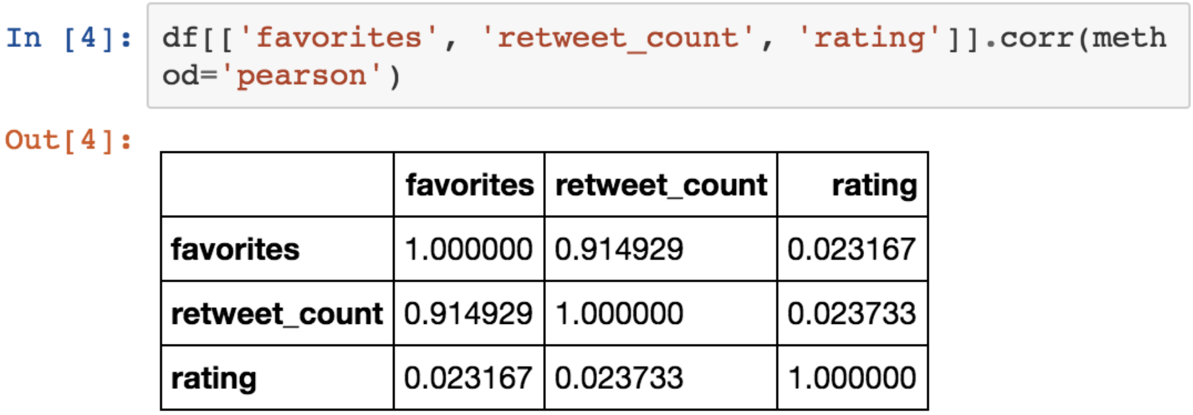
A few dogs received zero scores, or scores close to zero.



Mean :10.5
Std :2.2
25%:10
50%:11
75%:12

Conclusion:

- More than 75% of the data has more than 12 /10 as rating.
- The page starts with small rating than they adopt the system of rating numerator more than the denominator. It seems that ratings getting higher with np specific reason.
- It looks like as time went on dog_rates stopped rating dogs under 10.



In [4]:

```
df[['favorites', 'retweet_count', 'rating']].corr(meth
od='pearson')
```

Out[4]:

	favorites	retweet_count	rating
favorites	1.000000	0.914929	0.023167
retweet_count	0.914929	1.000000	0.023733
rating	0.023167	0.023733	1.000000

So, I ran a correlation to see if there is a correlation between favorites and retweets. Yet there is no correlation between rating and retweets or favorites. There are a few possible explanations. One is that the dogs are not actually getting better. The other is that the 'lower quality' dogs are given funnier captions. In those cases, it is the caption getting retweets and favorites, not the dog itself.