Analyze Data WeRateDogs Tweets

Hanin Falatah | Nanodegree Data Analyst | 12/30/19

Introduction

The dataset that I will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also wikipedia known as in 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.

> WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for you to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

Analysis

This is analysis of the tweets rating dogs by the account @dog_rates, that is concise analysis and they are provided to @dog_rates by their direct massages. And there are some replies changing the ratings and there in the analysis all the replies were ignore, and this analysis is from df_tweet_data, df_predict, and df_archive_tweet the DataFrames.

	rating_numerator	rating_denominator	img_num	p1_conf	p2_conf	p3_conf	favorite_count	retweet_count
count	1819.000000	1819.000000	1819.000000	1819.000000	1.819000e+03	1.819000e+03	1819.000000	1819.000000
mean	12.723474	10.517317	1.216603	0.599781	1.355054e-01	6.060288e-02	9465.238043	2922.585487
std	43.297103	7.132573	0.577664	0.269589	1.013052e-01	5.155760e-02	12608.289684	4854.004383
min	0.000000	2.000000	1.000000	0.044333	1.011300e-08	1.740170e-10	81.000000	16.000000
25%	10.000000	10.000000	1.000000	0.369419	5.340615e-02	1.605405e-02	2181.500000	652.500000
50%	11.000000	10.000000	1.000000	0.596882	1.192430e-01	4.921790e-02	4550.000000	1466.000000
75%	12.000000	10.000000	1.000000	0.848609	1.967120e-01	9.351970e-02	12207.000000	3433.000000
max	1776.000000	170.000000	4.000000	1.000000	4.676780e-01	2.734190e-01	132810.000000	79515.000000

Figure 1: the describe() methods

All the posts that are analysis are 1819 posts, and from the post the describe() methods shows:

1- The maximum number of Likes for tweet that were analysis 132810 likes.
2- The maximum number of Retweets for tweet that were analysis 132810 retweets.
3- The average of retweet 2922.58 on a post.
4- the average of favorite 9465.23 on a post.
5- The average of rating is 12.7.

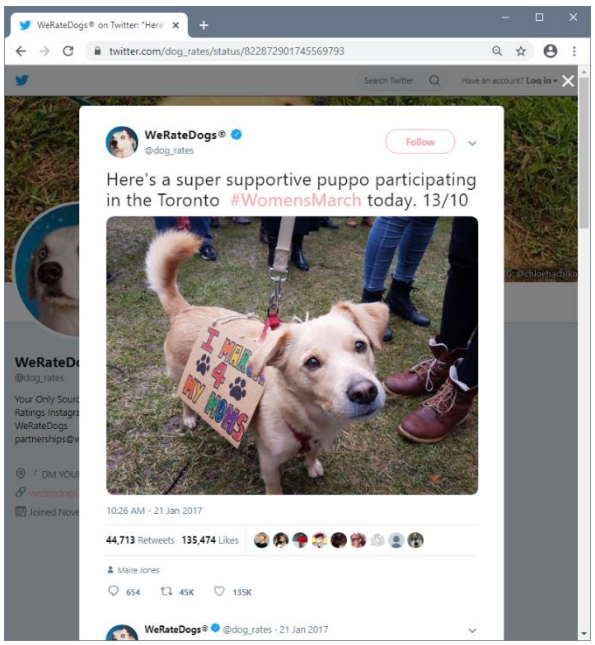


Figure 2: Most Liked post.

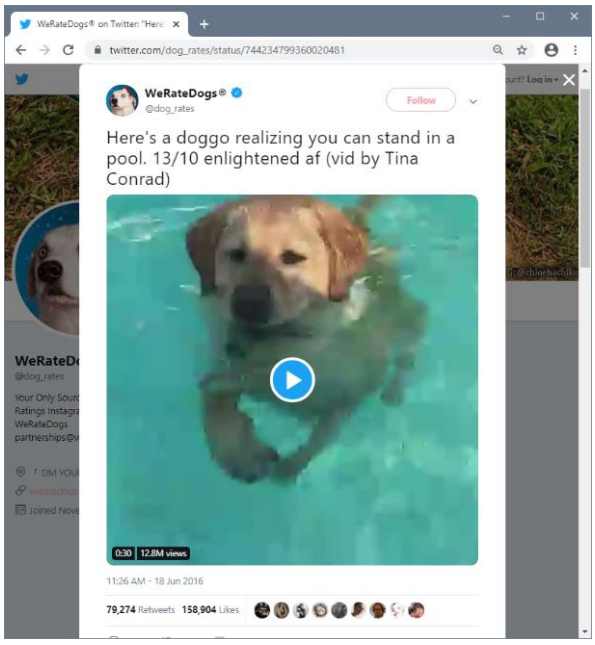


Figure 3: Most Retweeted Post.

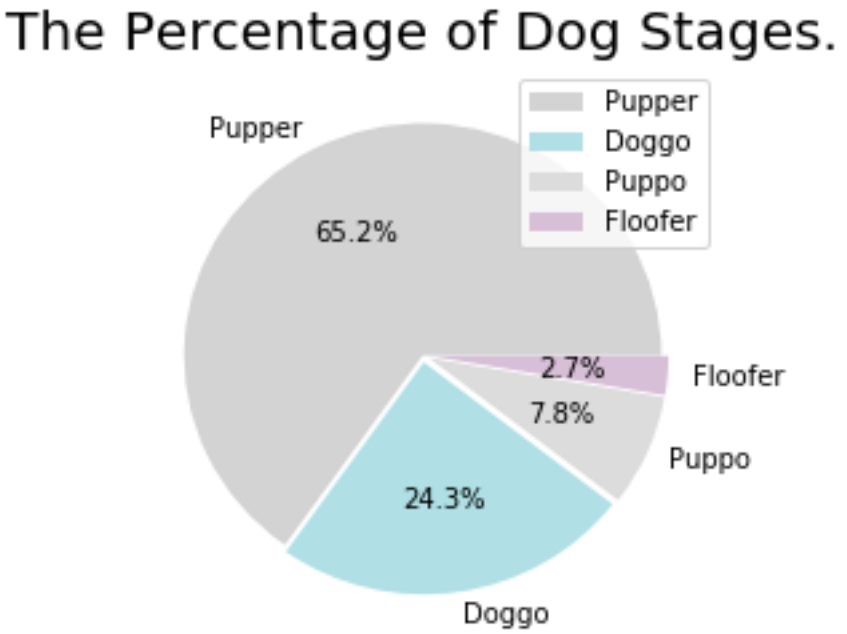
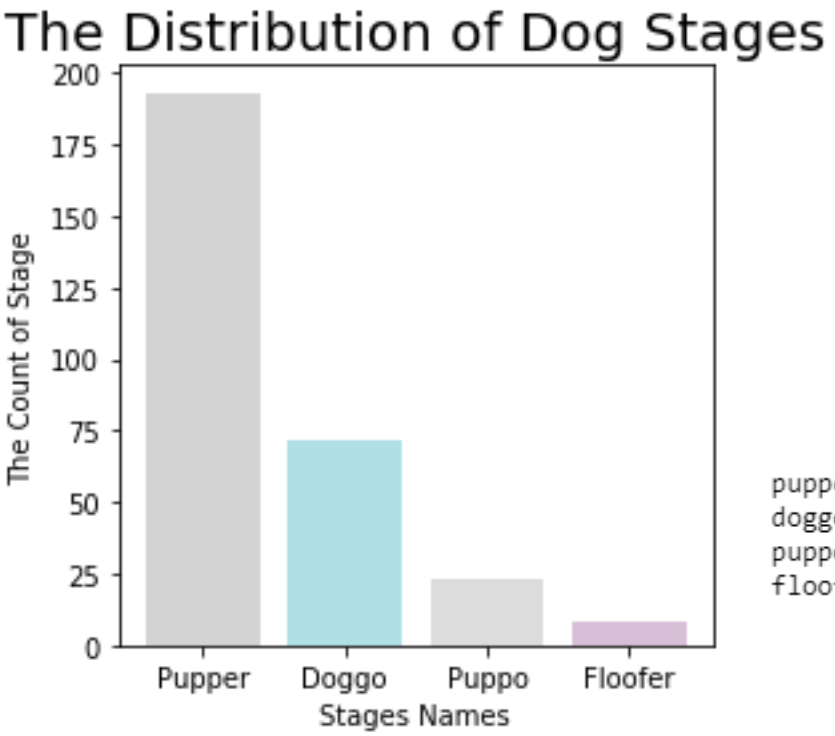


Figure 4: the Percentage of dog stages

There is a major amount of the dog stage that is not provieded and the represent 82.7% of the stages. but after ignoring the None in the dog stages the two plot shows:

1- That most rated stage is represent as 65.2% of the provided stages is Pupper with the count of 193.

2- That 24.3% of the provided stages is Doggo with the count of 72.



3- That 7.8% of the provided stages is Doggo with the count of 23.

4- That 2.7% of the provided stages is Floofer with the count of 8.

pupper	193
doggo	72
puppo	23
floofer	8

Which means the main dog stage is Pupper and it 65.2%, and the rare dog stage is Floofer and it 2.7%.

Figure 5: The Distribution of Dog Stages

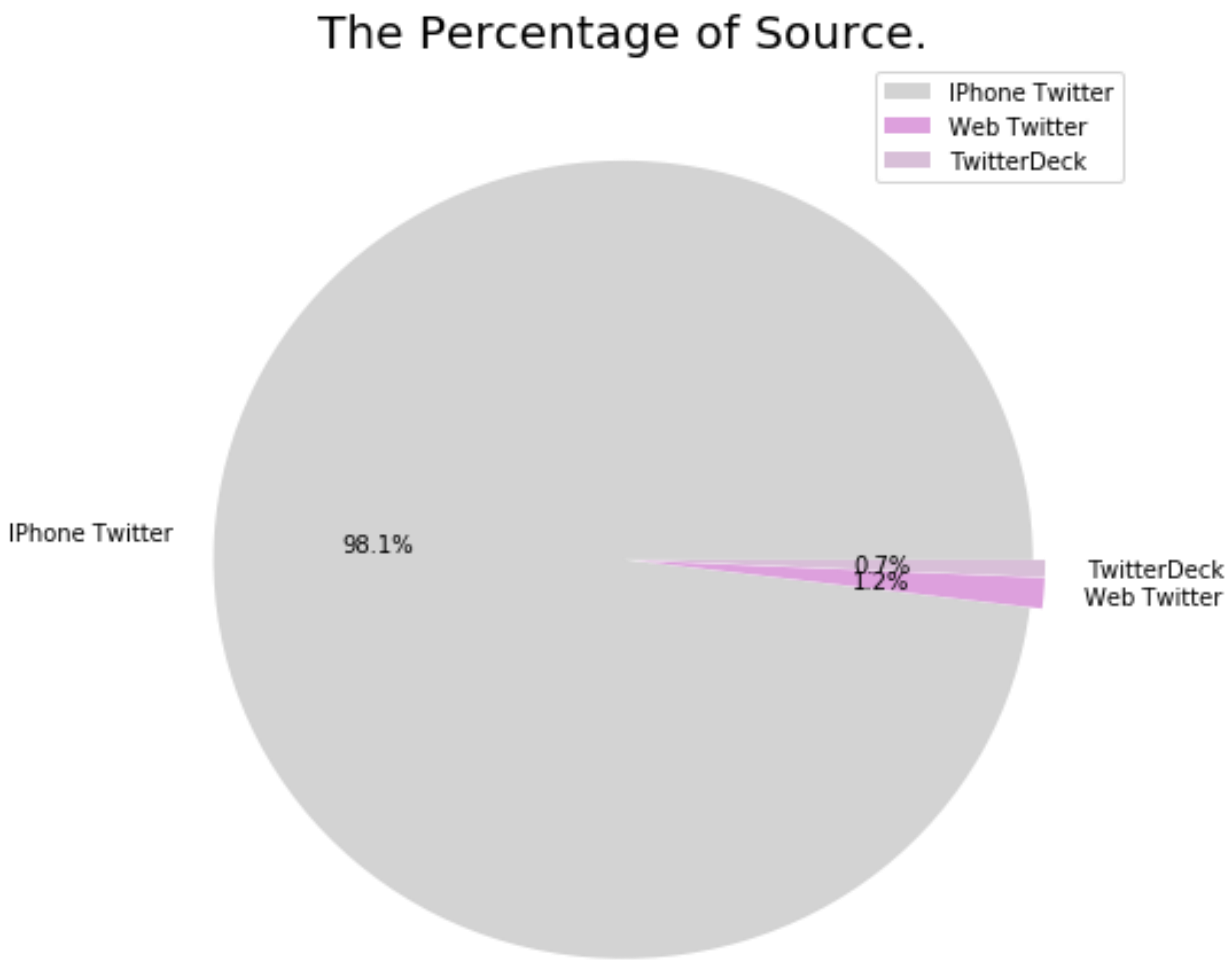
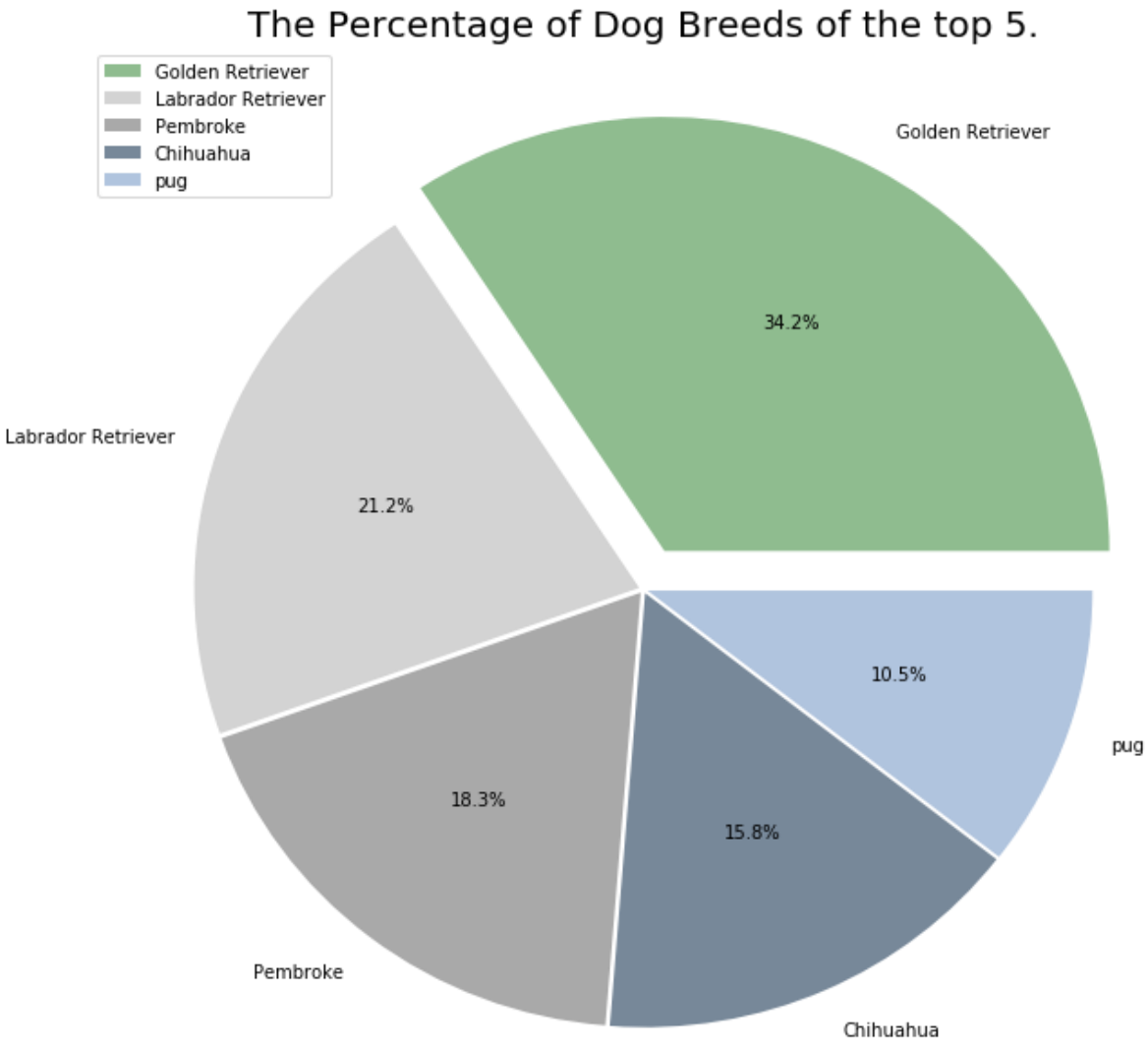


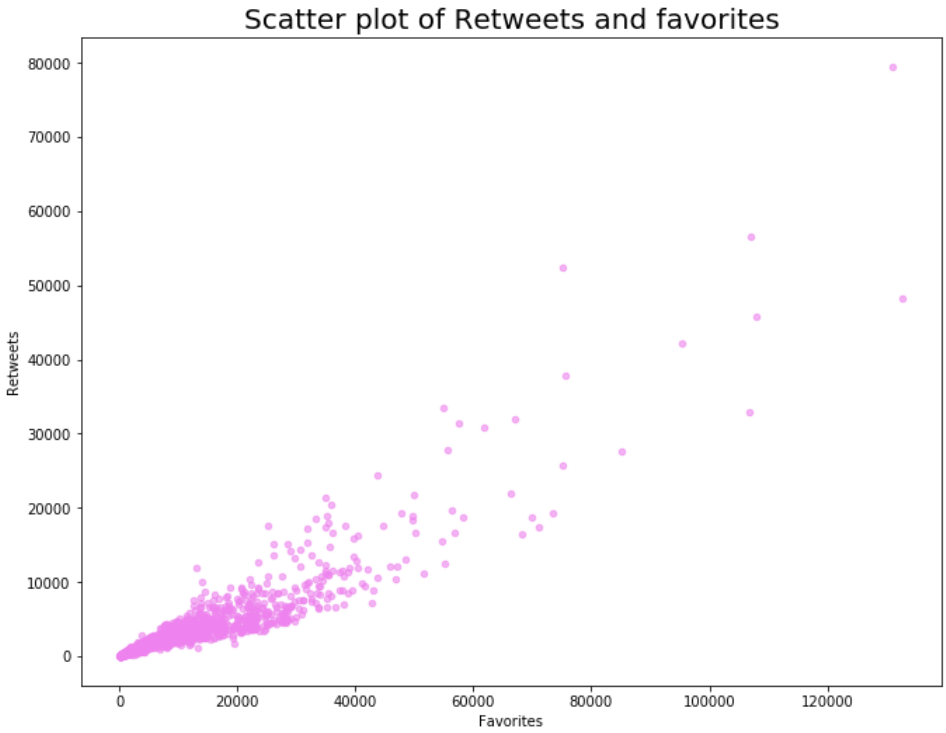
Figure 6: the tweets sources

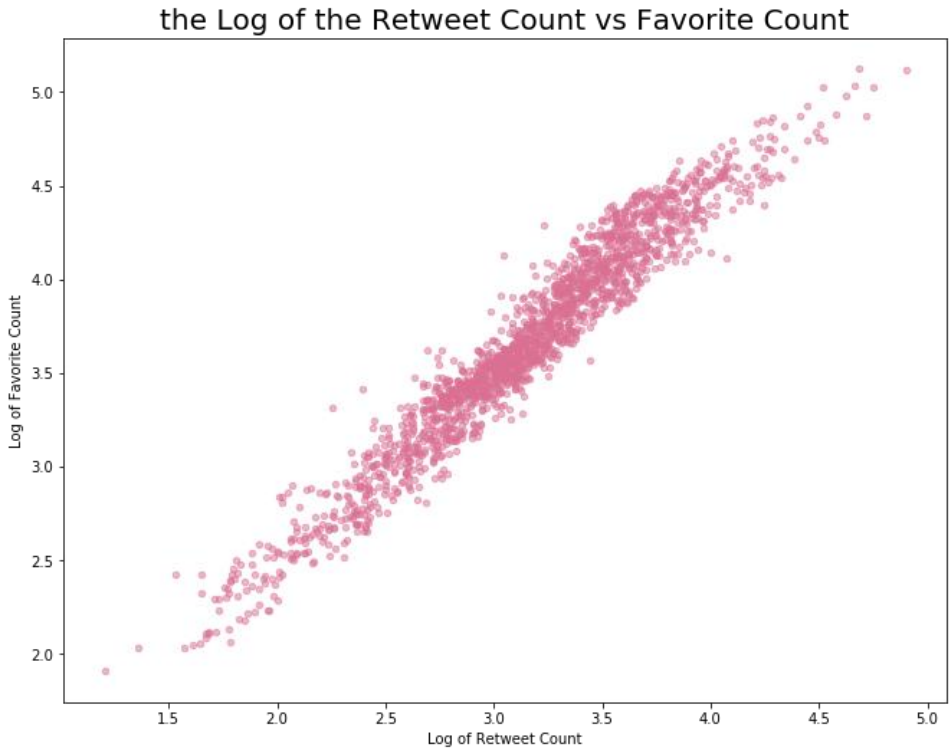
The above plot shows the source of tweest in the DataFrame, shows the 98.1% represent the tweet from IPhone Twitter app, and that 1.2% represent the tweets are from the Twitter Web Client page, and that 0.7% represent the tweet from TweetDesk. Which means the main source is the iphone twitter app, and the rare source is TweetDesk.



The above plots shows the top 5 dogs breads which are the Golden Retriever, Labrador Retriever, Pembroke, Chihuahua, and pug and after seeing the plot the top 5 the so That most populer breed is represent as 34.2% of the provided breeds is Golden Retriever with the count of 176, That 21.2% of the provided breeds is Labrador Retriever with the count of 109, That 18.3% of the provided breeds is Pembroke with the count of 94, That 15.8% of the provided breeds is Chihuahua with the count of 81, and That 10.5% of the provided breeds is pug with the count of 54.

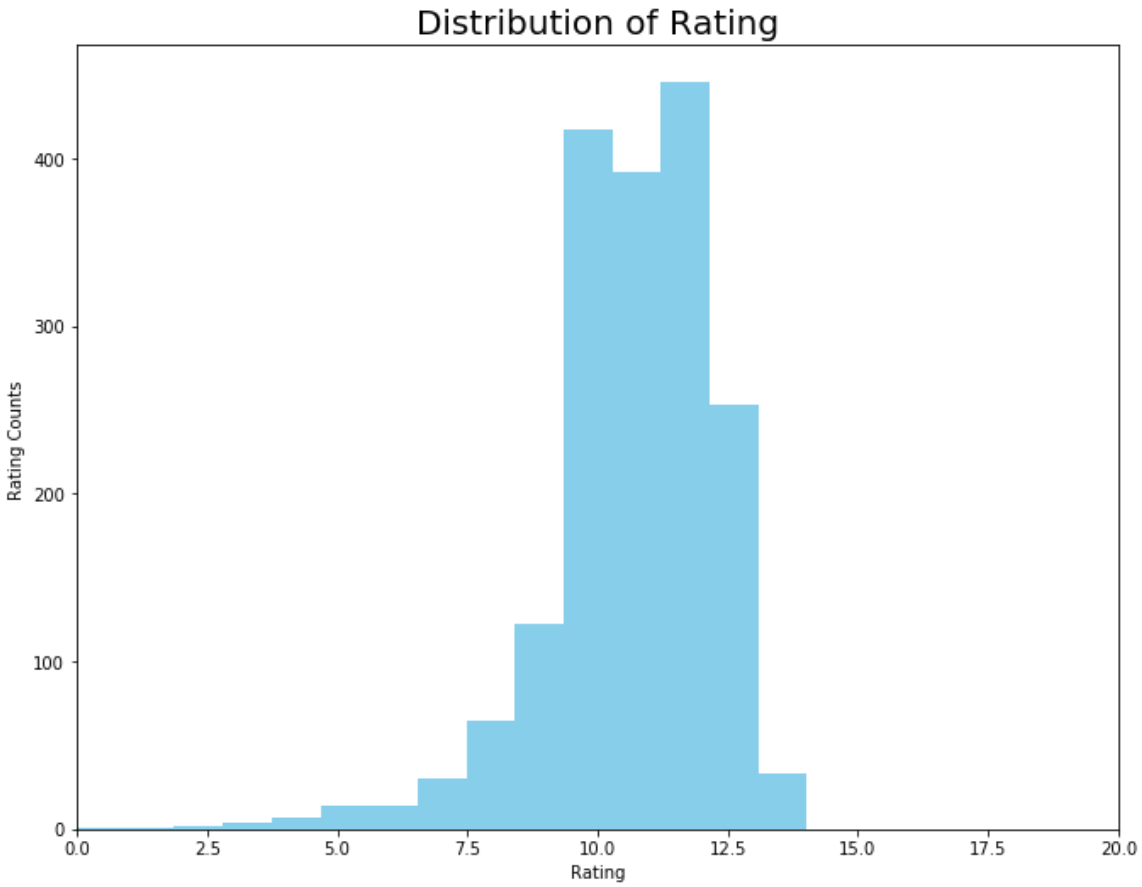
Which means the main dog breeds is Golden Retriever and it 34.2%



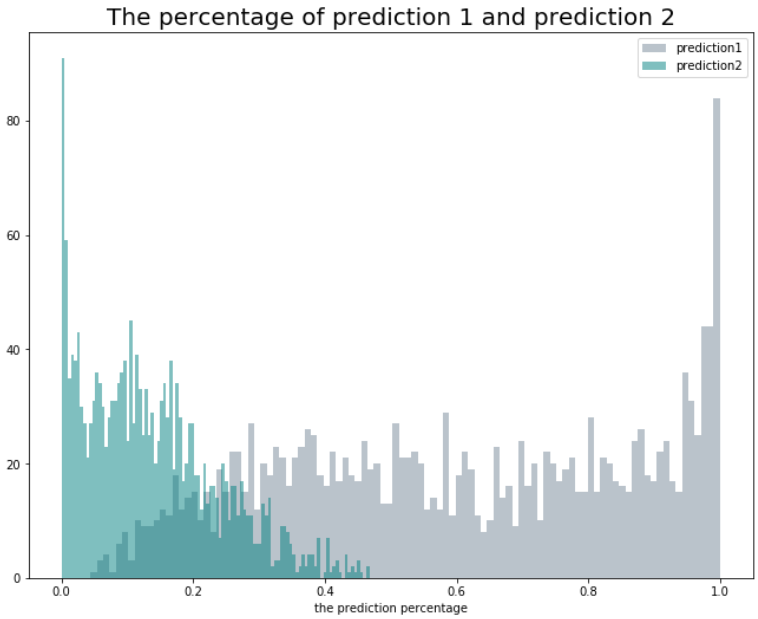


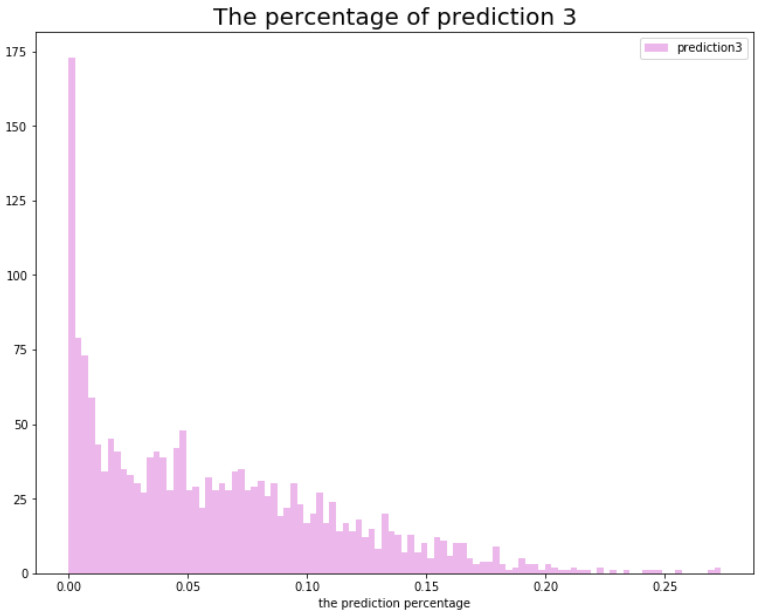
5

The above maps shows the correlation between the Favorites and Retweets which shows a postive strong relationships between the Favorites and Retweets and it's 0.97.



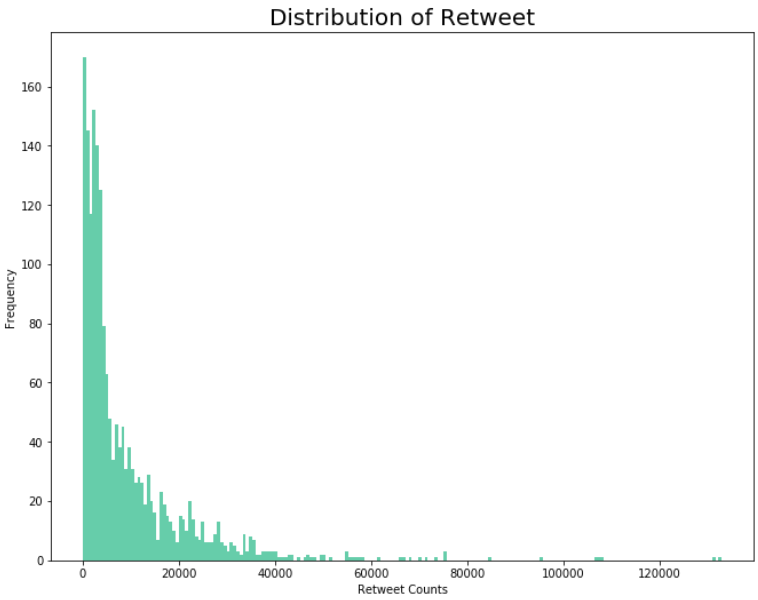
The above histogram shows the most used rating are from the 9 to 14 and the rating is not precise by any means there are some rating as 1776 out of 10 and there some rating that are not from 10.

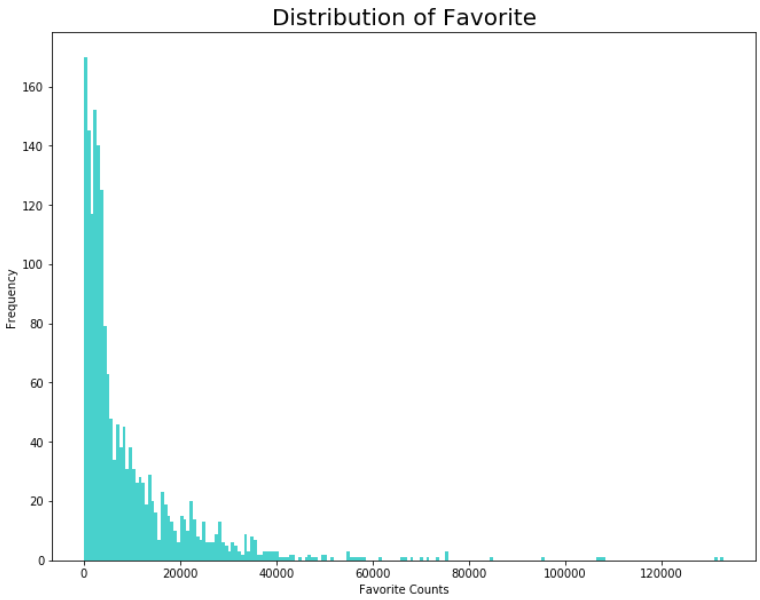


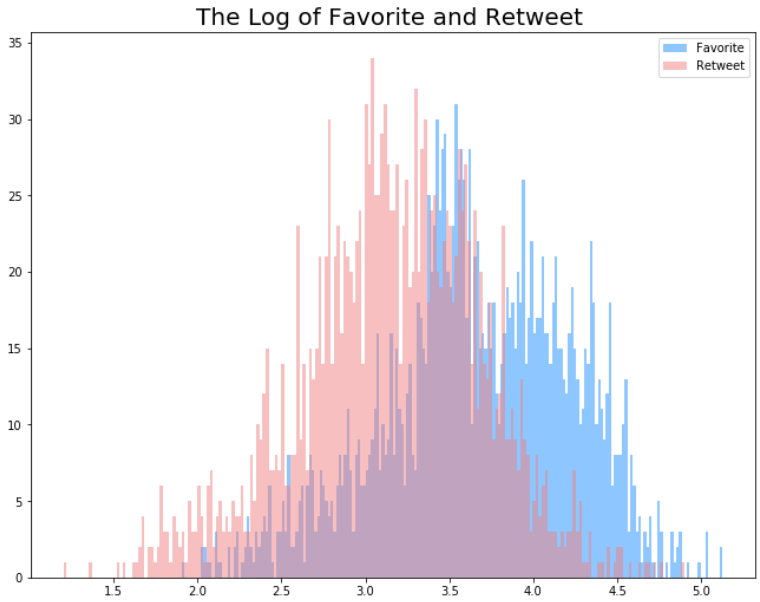


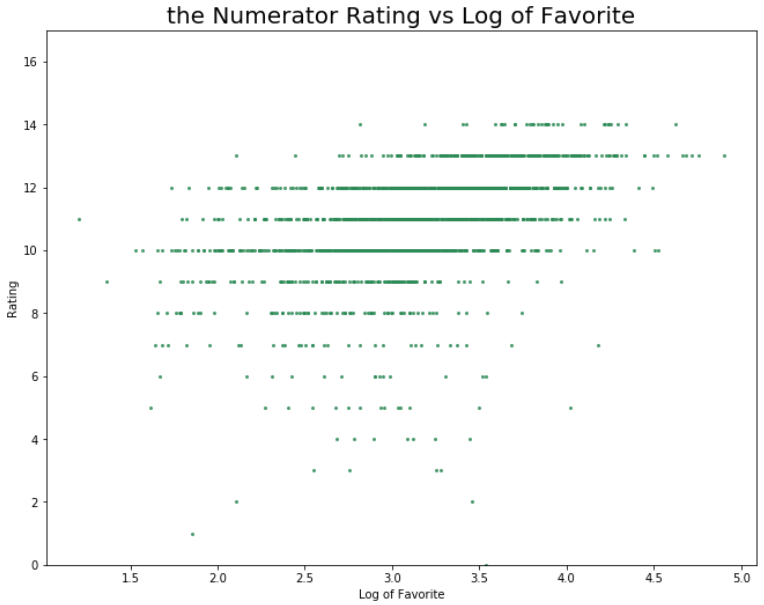
The plots show the prediction 1, prediction 2, and prediction 3, the prediction 1 values between 0.1 to 1.0 but most of the values are closer to 1.0, and prediction 2 1 values between 0.1 to 0.5 but most of the values are closer to 0, and the prediction 3 values between 0.0 to 0.25 but most of the values are closer to 0.0 and the prediction is small that is not relevant.

The most trustworthy of predictions is prediction 1.

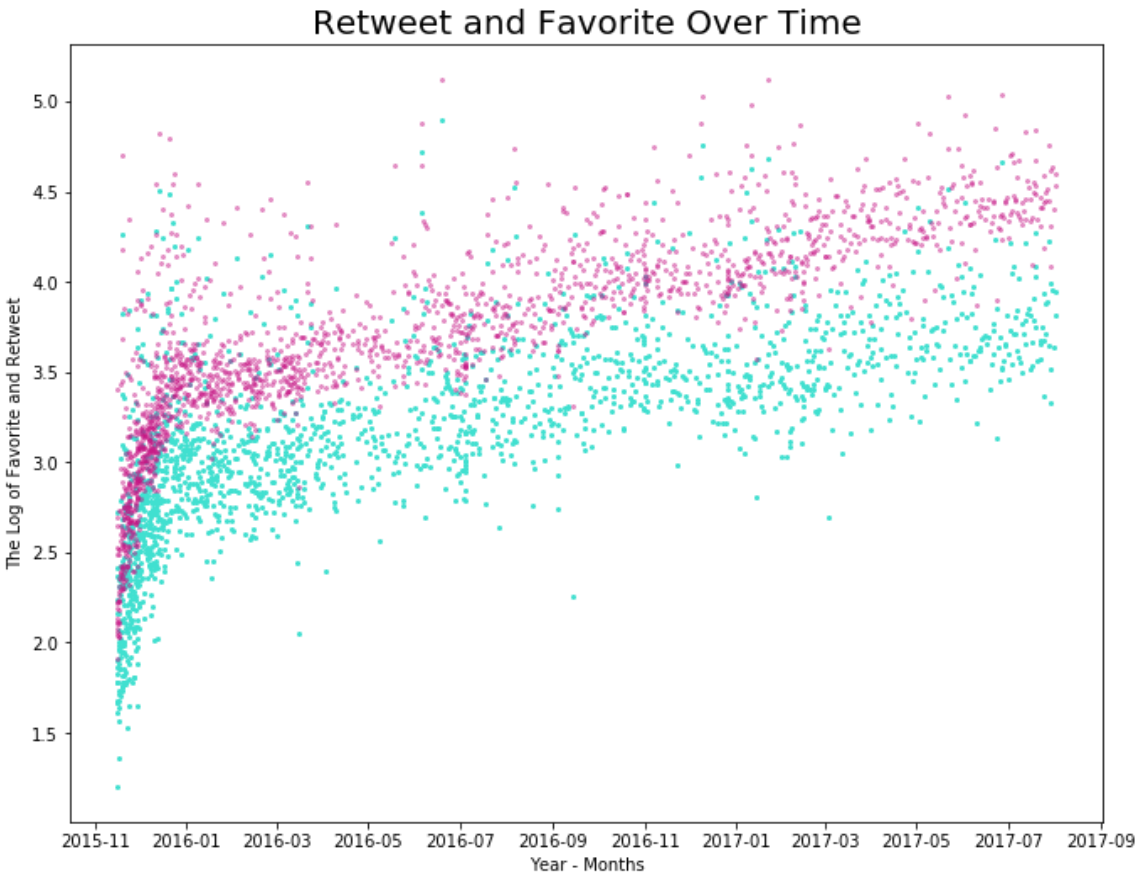








The plots show summarizes the similarities between the rewteet_count, and favorite_count, and most likely the twitter account become verified and viral on twitter. Then a this means retweet and favorite become higher.



From the above scatter plot, you can see how the twitter account got huge amount of retweet and likes then in small amount of time, then the retweet and likes slowly got