Wrangling and Analyzing Twitter @Dog\_Rates

# Introduction

There is a popular saying that dogs are man’s best friend. In accordance with this love of dogs, someone created a famous Twitter account with the handle @dog\_rates and the name “WeRateDogs™.” As of March 12, 2018, WeRateDogs™ has 6,719 tweets, around 6 million followers, and a Klout score of 80. This project analyzes a selection of those tweets.



Figure : Example Tweet (Source: https://twitter.com/dog\_rates/status/966106502858756096)

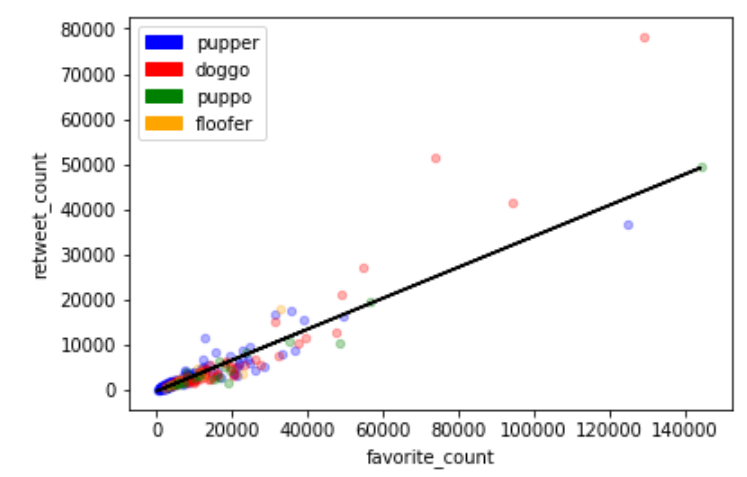
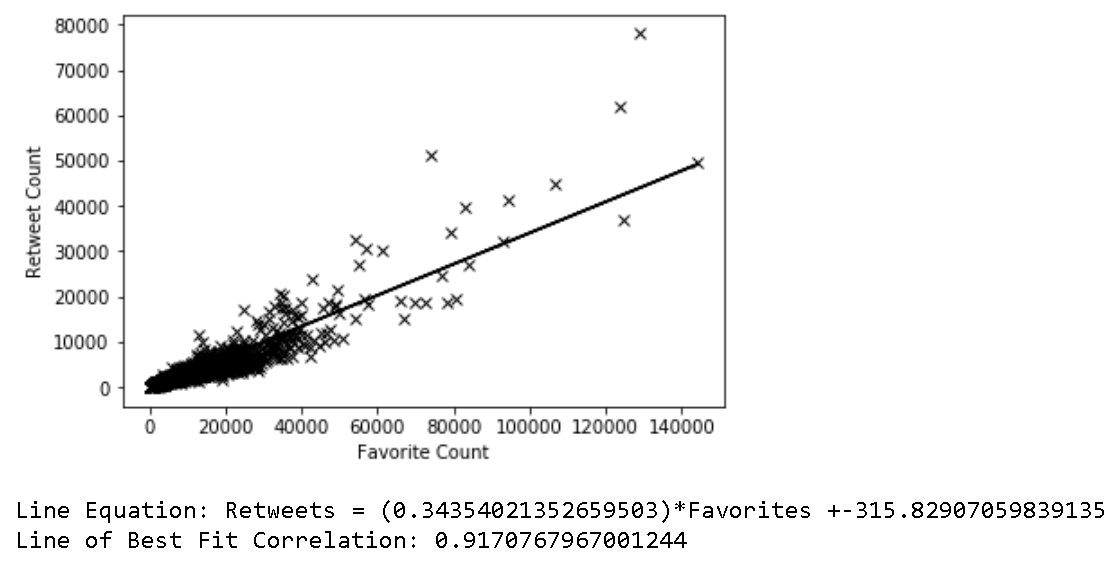
# Dataset

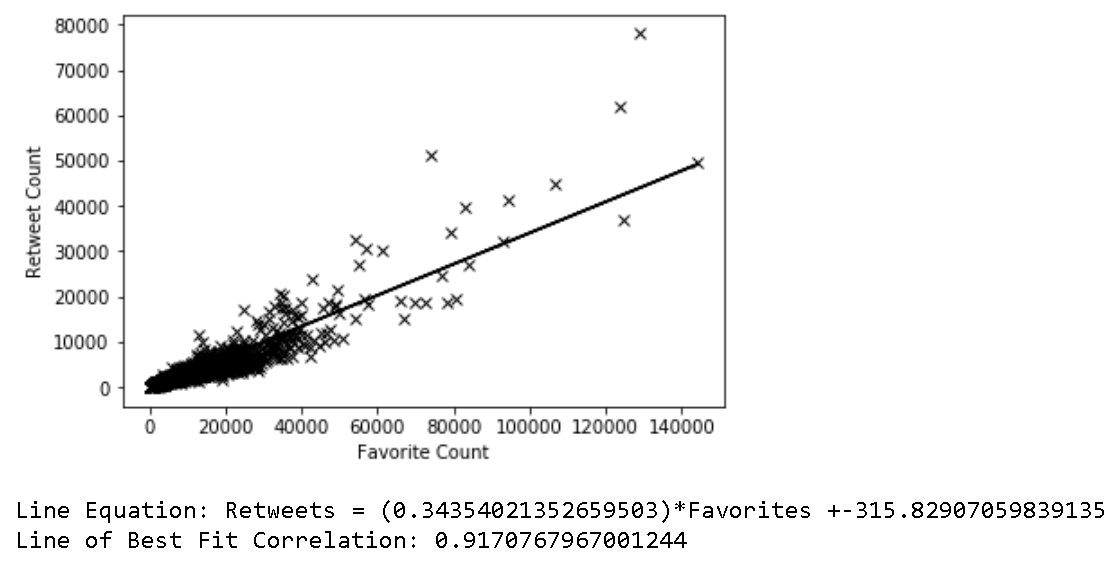
Three different sources were utilized in the process of analyzing the tweets from WeRateDogs™. The first source is an archive of the past tweets provided via a CSV from Udacity. The second source is from the Twitter API which was used to retrieve the number of retweets and favorites. The third data source predicted the dog breed in each tweet's image programmatically from a neural network, which was also provided by Udacity. After combining all these datasets, 1994 tweets remained. These tweets were from before August 2017.

# Analysis

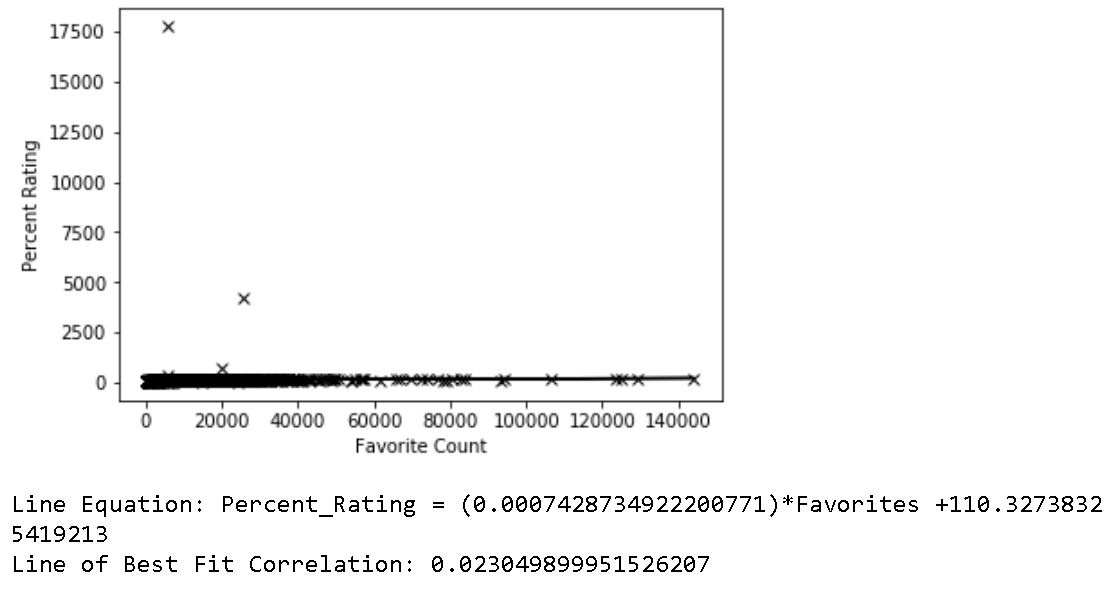
The first aspect analyzed was which dog breeds and dog categories were tweeted about the most. The most tweeted dog categories were pupper (11.3%, 68.6% without “None” category) then doggo (3.16%, 19.2% without “None” category). Ignoring non-identified dogs, the most tweeted dog breeds were Golden Retriever (9.4%), Labrador Retriever (6.4%), and Pembroke (5.6%).

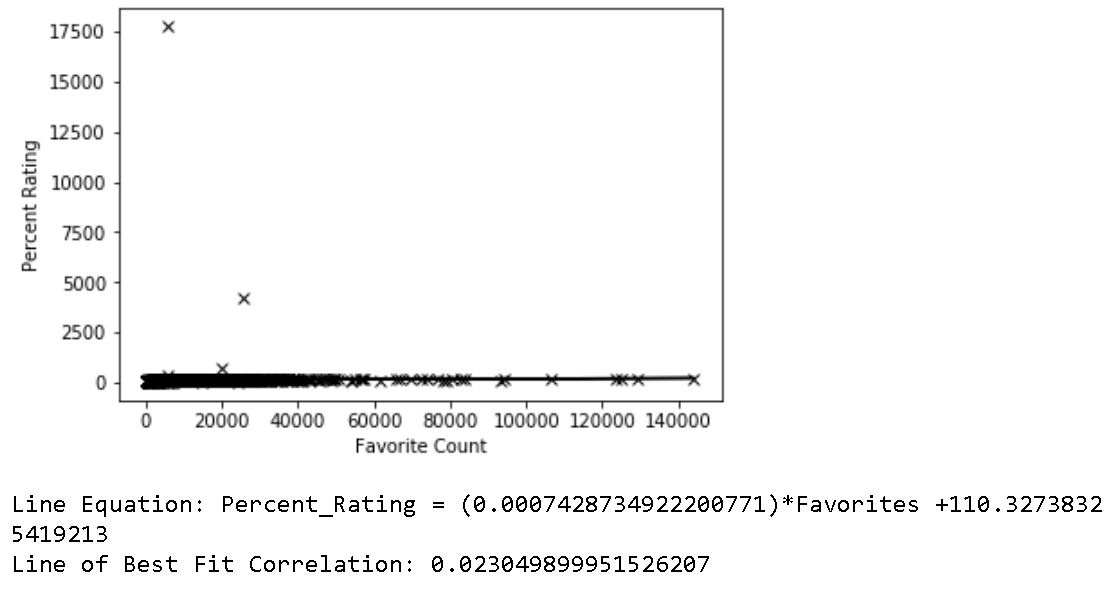
The next aspect investigation was if there was a correlation between retweets and favorite count as expected. There is over a 90% correlation between these two factors.





However, when seeing if there is a correlation between the percent rating of the dog and the retweet count, there was no strong correlation. There was also no strong correlation between percent rating and favorite count, but it was better than retweets. Percent rating is the comical rating given to each dog that is generally greater than 100% and is typically out of ten. In the tweet example shown in the beginning of this project, the percent rating would be 130% because the dog was awarded a 13/10.





The next two graphs show percent rating by dog breeds. While the Golden Retriever was the most tweeted dog, it was not the dog with the highest rating.

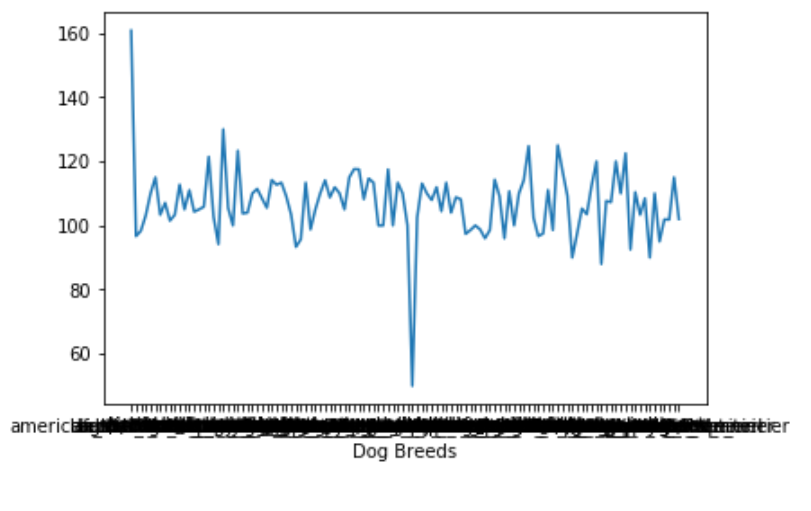


Figure : This graph shows that most dog breeds average a rating between 110 and 130, but there are two outliers.

|  |  |
| --- | --- |
| Bouvier.JPG  Figure : Bouvier Des Flandres (Source: https://en.wikipedia.org/wiki/Bouvier\_des\_Flandres#/media/File:Bouvier.JPG) | Figure : Japanese Spaniel, also known as a Japanese Chin (Source: https://en.wikipedia.org/wiki/Japanese\_Chin#/media/File:Japanese\_Chin\_adult.jpg) |

The highest average rating given is 160; this rating is the average for no dog breed identification. The lowest average rating is 50. Since the highest average rating is when there is no dog breed identification, two theories are that there is a low count in this category or that the “coolest” images and therefore hardest for the machine learning algorithm to classify the dog breed get the highest ratings. The second highest average dog rating is 130 for a Bouvier Des Flandres. The lowest average rating at 50 was for a Japanese Spaniel.

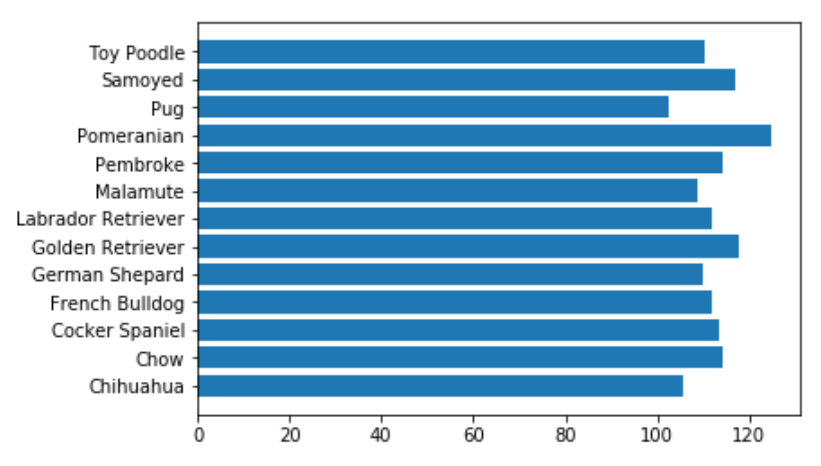
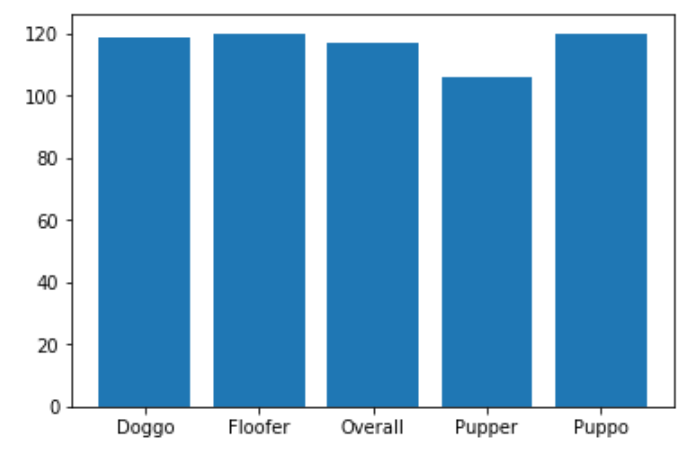


Figure : Dog Ratings for the top 13 tweeted about dogs.

The last graph shows the percent rating by dog category. It shows that there is not much difference between the bars.



The dog with the highest mean favorite count is a Saluki, and the lowest is a Brabancon Griffon. The dog category listed from highest to lowest ignoring combinations are puppo, doggo, floofer, and pupper.

Overall, an important insight from this data is that when a dataset is created out of fun, like the WeRateDogs™ Twitter stream, there might not be any statistically significant relationships. It is likely the owner chooses ratings, dog categories, and which dog breeds to tweet almost randomly.



Figure : An example of how it can be difficult for a computer to identify animals and how the tweets are typically humorous without a consistent structure that could lead to correlations.

# References

2018. Udacity.com

2018. https://twitter.com/dog\_rates