WeRateDogs Data Analysis Report

Introduction

Using data from WeRateDogs twitter handle, I wanted to get a deeper understanding of the data by observing trends and possible correlations between different metrics. To do this we did statistical calculations, sample distributions and multivariate analysis. The aim was to see if patterns would emerge and how that can be used to better understand user behavior. The popularity of the account has been increasing over time and provides insight into where interests lie and how its evolving. As a result, this work is able to identify popular dog breeds, their ratings and other user behavior on retweets and favorites.

Dataset

The wrangling of the WeRateDogs data is part of my project for Udacity Data Analysis NanoDegree program. The project is meant to test the ability of students to use Python and its libraries to gather data from a variety of sources and in different formats, assess it quality and tidiness then clean it and analyze it.

The data we are using for this Analysis is from 3 sources:

1. The WeRateDogs Twitter archive which contains basic tweet data for all 5000+ of their tweets. it was manually downloaded through a link given in the Udacity classroom in csv format.

2. Data gathered from Twitter's API for the 3000 most recent tweets by @weratedogs in JSON file format

3. The Image Predictions data which was gotten by running every image in the WeRateDogs Twitter archive through a neural network that can classify breeds of dogs. Its hosted on Udacity servers in tsv format and can be downloaded programmatically using the requests library in python

Quality and tidiness issues were identified and cleaned in all datasets. All 3 were then combined into a master dataset to serve as the basis of our analysis.

Features and Pre-analysis

For our main features, we considered all the features available for each tweet: from the average rating of the dog, dog stage, retweet and favorite counts to timestamps. The features were analyzed individually and in relation to one another. We only used original tweets that have ratings and images.

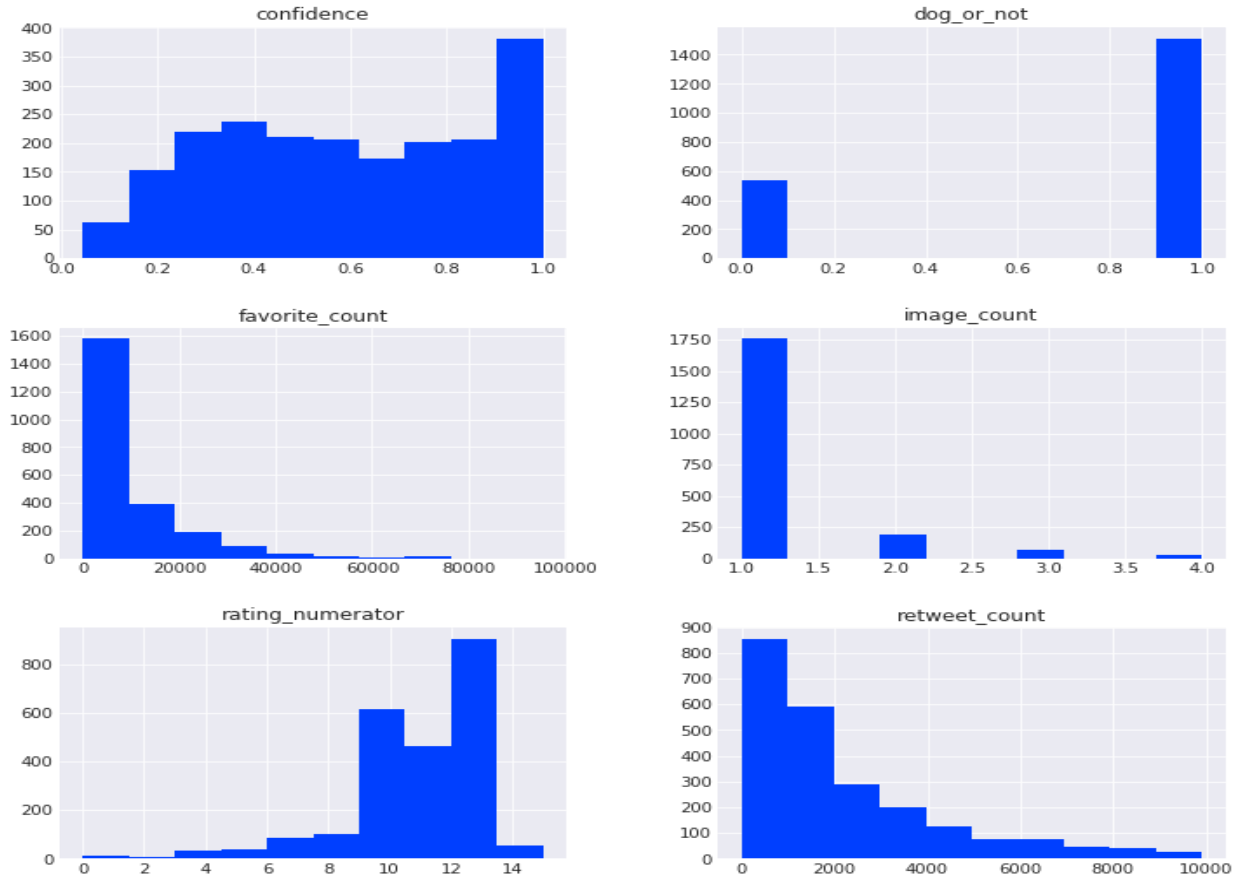
WeRateDogs has a unique rating system. The ratings almost always have a denominator of 10, so all rating that were not 10 were discarded for uniformity while the numerators are mostly greater than 10. Only the most relevant columns for analysis were kept in the master data set.

ANALYSIS

The summary statistics for the master dataframe reveals a few things about the dataset. The maximum rating_numerator is 15 and minimum is 0. The retweet and favorite counts vary vastly between maximum(9907, 95450) and minimum(0, 52) respectively

The retweet and favorite counts are right skewed(see figure 1). The rating numerator is left skewed showing most dogs having a rating of 10 and above as expected. I found the most popular dog name to be Charlie even though it loses its significance because a lot of tweets didn't contain names(721 entries were missing). I also found the top rated breed were Saluki but a closer inspection of the mean ratings for other breeds in the top 10 shows there is very little difference in ratings . This suggests the ratings by users are mostly arbitrary and subjective. This is different for the dog breed with the most favorites. The Golden Retriever sits at the top with 1,888,078 while the closest to it had 1164351. A comparison of the top 10 rated breeds and top 10 favorites shows none of the breeds it present in either group.

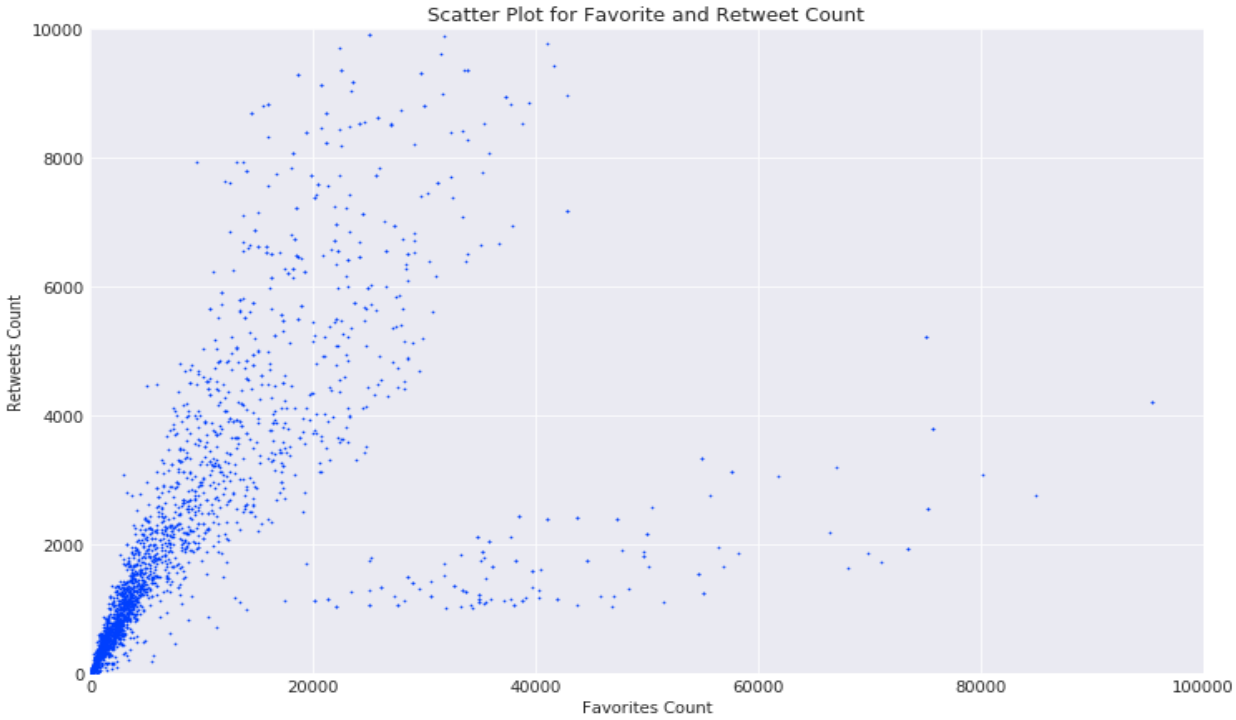
Figure 1



The retweets and favorite counts look to have a strong relationship from the scatter plot(see figure 2). So, a regression line was fitted to further explore the relationship. It found that for every 1 unit increase in favorite count we will predict that the retweet count will increase by 0.1003. The other relationship examined was the rating versus the favorite. I found that for every 1 unit increase in rating numerator we will predict that the favorite count will increase by 2175.27. Graphically it can

be seen, the ratings between 10 and 14 have the highest favorite count(See Figure 3). The p-value shows the rating numerator and favorite count are statistically significant in predicting the favorite and retweet count respectively.

Figure 2



Next we explored the effect of time on tweets, retweets, favorites and rating. We found the number of tweets have declined over time on the account but the retweets and favorite counts have risen consistently with only a few irregular dips(See Figure 4). This shows the continued popularity of the WeRateDogs twitter account despite reduced content. Mondays, Tuesdays and Wednesdays are days were people retweet and favorite the most. The ratings have also changed significantly over time. At the beginning in November 2015, there was an average monthly rating of 8.94 but by August 2017 it has risen to 13.0(See Figure 3).

Figure 3

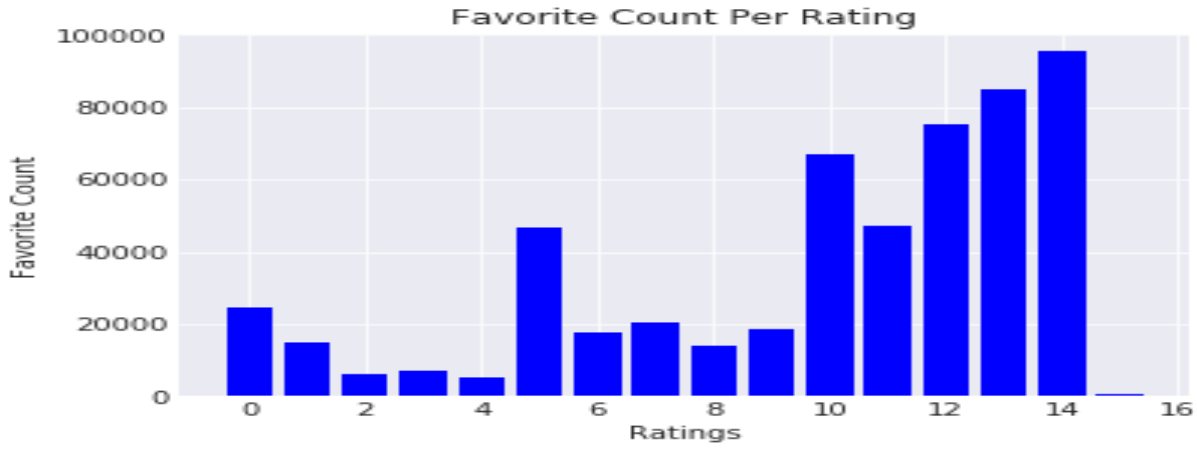
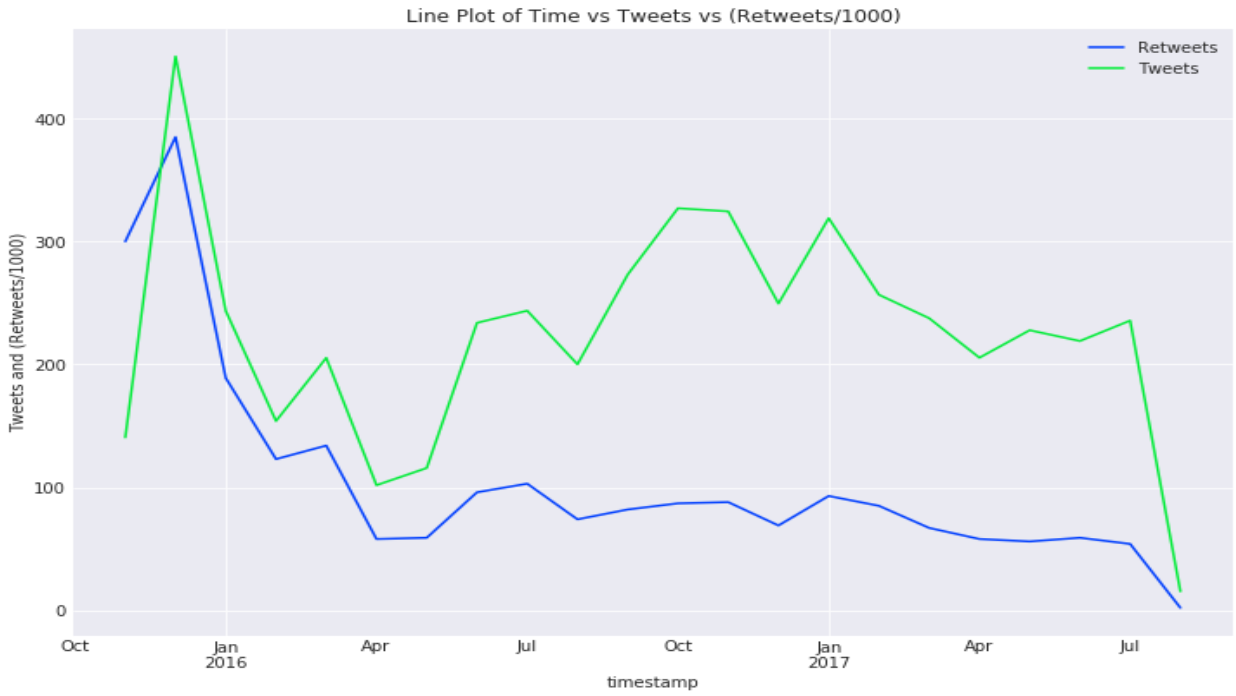


Figure 4



LIMITATIONS

- 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." This shows it's not meant to be a rigorous exercise. Ratings, retweets and favorites are subjective. It's more of a social experiment.
- The dog stage column has only 378 entries with 1943 missing. This limits how much you interpret results involving dog stages.
- The name column also has a lot of missing values making it difficult to read much meaning into it.

CONCLUSIONS

1. The WeRateDogs account really took off and had exponential growth after the initial effort by the owner. The growth looks organic and shows have social media accounts can take a life of their own. So even though the owners are not tweeting as much as they used to, the account still maintains its popularity.
2. Top rated dogs are not also the most favorited. The Golden and Labrador retriever are the 2 most favorited dogs but do not appear on the top rated 10. In fact none of the top rated is also in the most favorited. This shows that users have a very large variety of dogs.

3. Ratings have a moderate correlation with favorite counts while retweets and favorite counts have an even stronger correlation.
4. Rating changed over time, with monthly averages increasing month on month.