The WeRateDogies Project

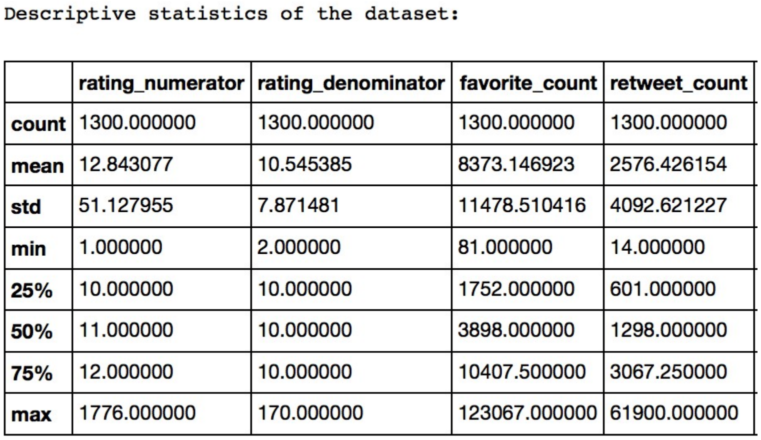
An analysis Report of data wrangled from Twitter

By Bibhash Kalita

For this analysis I gathered data from three different sources. WeRateDogies 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. The convolutional neural network predictions were programmatically downloaded using the Requests Python library as a tsv file. And finally, using the tweet IDs from the WeRateDogies 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, which I would later use to analyze the tweet's retweet and favorite (i.e. "like") counts.

Before diving into the statistical analysis, I began by answering some basic questions. What are the most common dog names in the dataset? What does the tweet say about the dog with the lowest rating (i.e. 0/10)? Using the Dog Breed Classifier, what do the dogs with the lowest rating look like and was the classifier able to accurately predict the dog's breed?

I discovered the most common dog names within the WeRateDogies dataset, excluding the NaN values, are Oliver, Winston,Tucker and Penny.Now let's dive into a statistical analysis of the Dog Ratings!



```
Descriptive statistics of the dataset:
```

	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

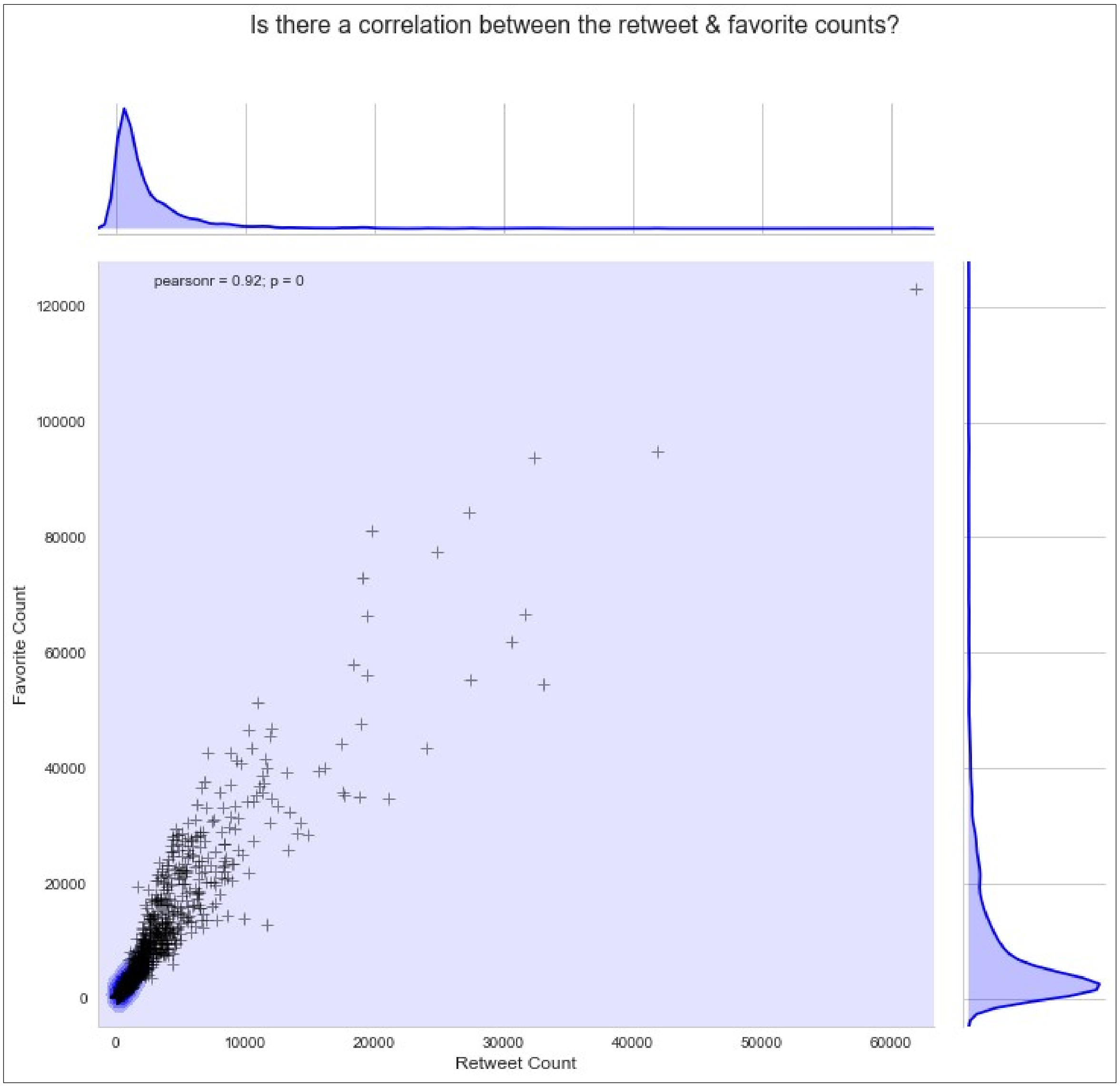
The mean numerator value is 12.84. The most interesting result is the rating_numerator maximum value of 1776. Recall, July 4, 1776 is the date the United States of America declared its independence from Britain, so let's explore this outlier further.

The rating_ numerator outlier is a dog named Atticus. The tweet was sent on July 4, 2016, stating, “This is Atticus. He’s quite simply America af…” Atticus’s picture is on the right. Not surprising, the dog breed classifier wasn’t able to predict Atticus’s breed due to the bowtie and sunglasses.

The dogs with the highest favorite count also has the maximum retweet count. On further investigation I found out that his name is Stephan; he had a rating of 13/10. The tweet said, “This is Stephan. He just wants to help”. The Dog Classifier did really well in predicting Stephan’s breed. Stephan appears to be a Chihuahua/Corgi mix and the classifier pegged Stephan as a Chihuahua with a predication confidence equal to 0.51. Below is a picture of Stephan; the most popular do in the dataset.

Now it’s time to dive into the favorite and retweet count data. The statistical analysis denotes a large positive (right) skewed distribution in both categories indicated by the large standard deviations. The

results also indicate people will favorite a tweet more often then they will retweet the original tweet as shown by the larger favorite count.



From the visualization above, you will see a strong correlation between the favorite and retweet data with a Pearson correlation coefficient, r, equal to 0.92. The strong correlation makes logical sense because the popular a tweet the higher the favorite and retweet count should be.