The WeRateDogs Project

An analysis of data wrangled from Twitter

By Savahnna L Cunningham
Date: November 26, 2017

According to an article by The Globe and Mail, Twitter today has almost 200 million users worldwide. Approximately 460,000 new Twitter accounts are opened daily. More than 140 million tweets are sent daily. That's one billion weekly tweets! For this project, the goal was to capitalize on Twitter's vast amounts of tweet data, utilizing the Twitter API to exploit the Twitter data of the user @dog_rates, aka WeRateDogs. WeRateDogs is a very popular Twitter account with over 4 million followers and has received international media coverage. WeRateDogs gained its popularity by rating people's dogs with a good-natured comment about the dog. The rating system is based on a fraction, with the denominator fixed at 10 and the numerator is almost always a number greater than 10 ($\frac{13}{10}$,$\frac{14}{10}$, etc.) Why, you ask? Because "they're good dogs Brent."

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. 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 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, 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 dataframe is there a picture of the dog with the lowest rating and was the dog classifier able to accurately predict the dog's breed?

I discovered the most common dog names within the WeRateDogs dataset, excluding the NaN values, are Charlie, Lucy, Oliver and Cooper.

The tweet with the lowest rating said, "When you're so blinded by your systematic plag...". The poor doggo's picture is on the right. Interestingly, the dog breed

classifier wasn't able to predict this dog's breed. The neural network was able to identify the swing but miss identified the dog as an American Staffordshire Terrier, not a Labrador Retriever.

Now let's dive into a statistical analysis of the Dog Ratings!

`Descriptive statistics of Dog Ratings:`

	rating_numerator	rating_denominator
count	2175.000000	2175.000000
mean	13.215172	10.492874
std	47.725696	7.019084
min	0.000000	0.000000
25%	10.000000	10.000000
50%	11.000000	10.000000
75%	12.000000	10.000000
max	1776.000000	170.000000

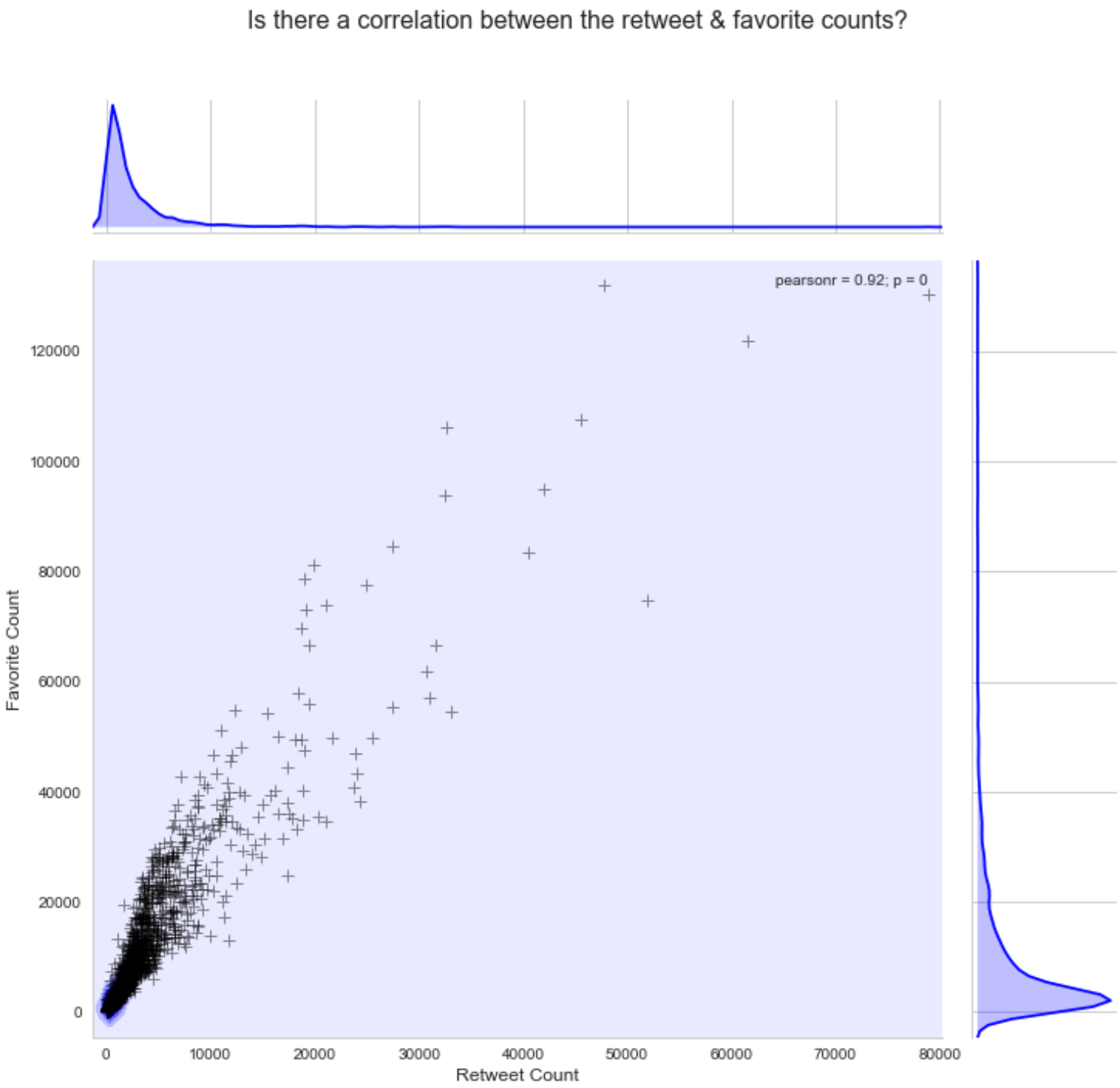
The mean numerator value is 13.2. The most interesting result is the rating_numerator maximum value of 1776. Recall, July 4, 1776 is the date the United States of America declared its independence from Britain, so let's explore this outlier further.

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.

Now it's time to dive into the favorite and retweet data. The statistical analysis denotes a large positive (right) skewed distribution in both the favorite and retweet count data. The results also indicate people will favorite a tweet more often then they will retweet the tweet.

`Favorited & Retweeted count statistics:`

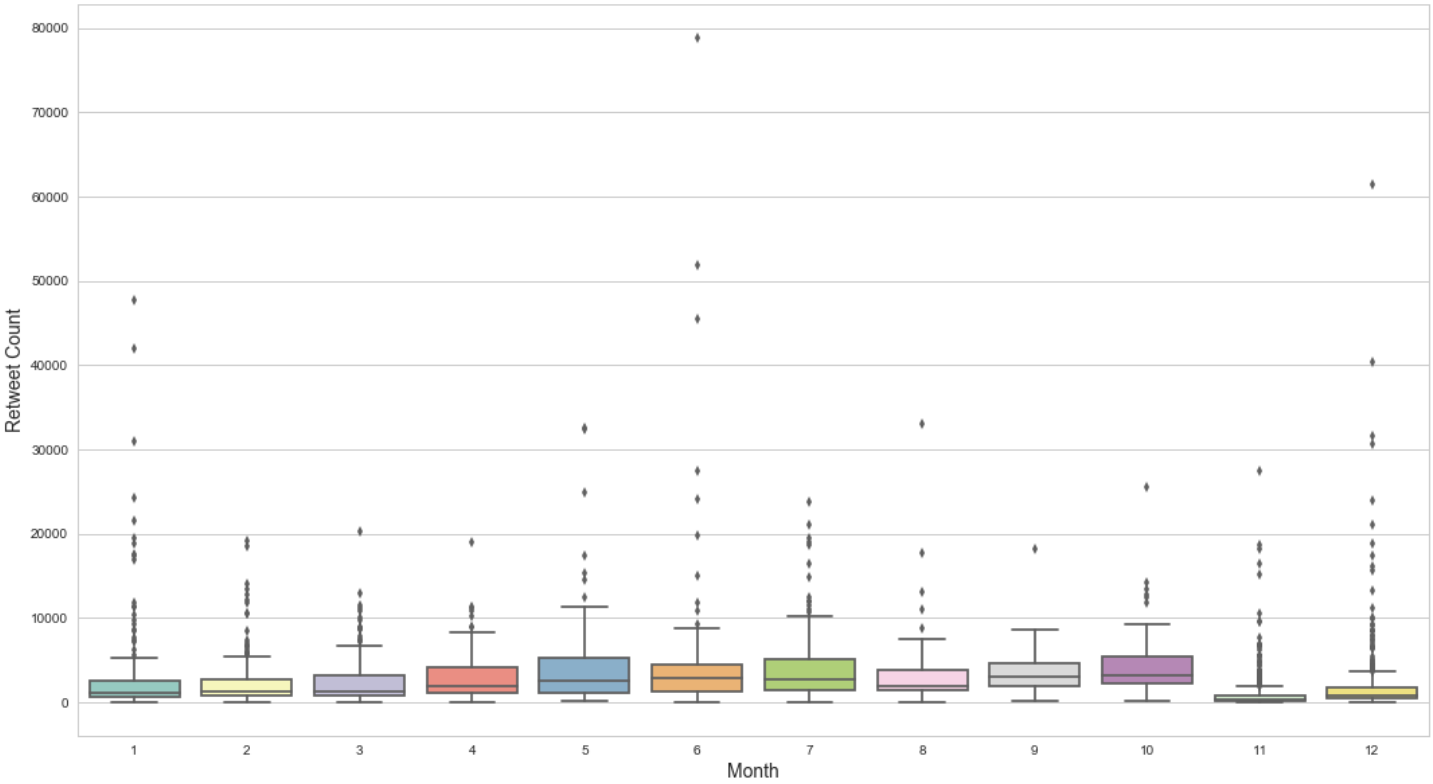
	favorite_count	retweet_count	Month	Day	Year
count	2175.000000	2175.000000	2175.000000	2175.000000	2175.000000
mean	8764.234023	2754.687816	7.028046	15.954483	2015.867126
std	12277.124345	4705.543071	4.125297	8.935383	0.694504
min	51.000000	0.000000	1.000000	1.000000	2015.000000
25%	1902.000000	604.000000	3.000000	8.000000	2015.000000
50%	4011.000000	1331.000000	7.000000	16.000000	2016.000000
75%	11042.500000	3188.000000	11.000000	24.000000	2016.000000
max	131946.000000	78853.000000	12.000000	31.000000	2017.000000



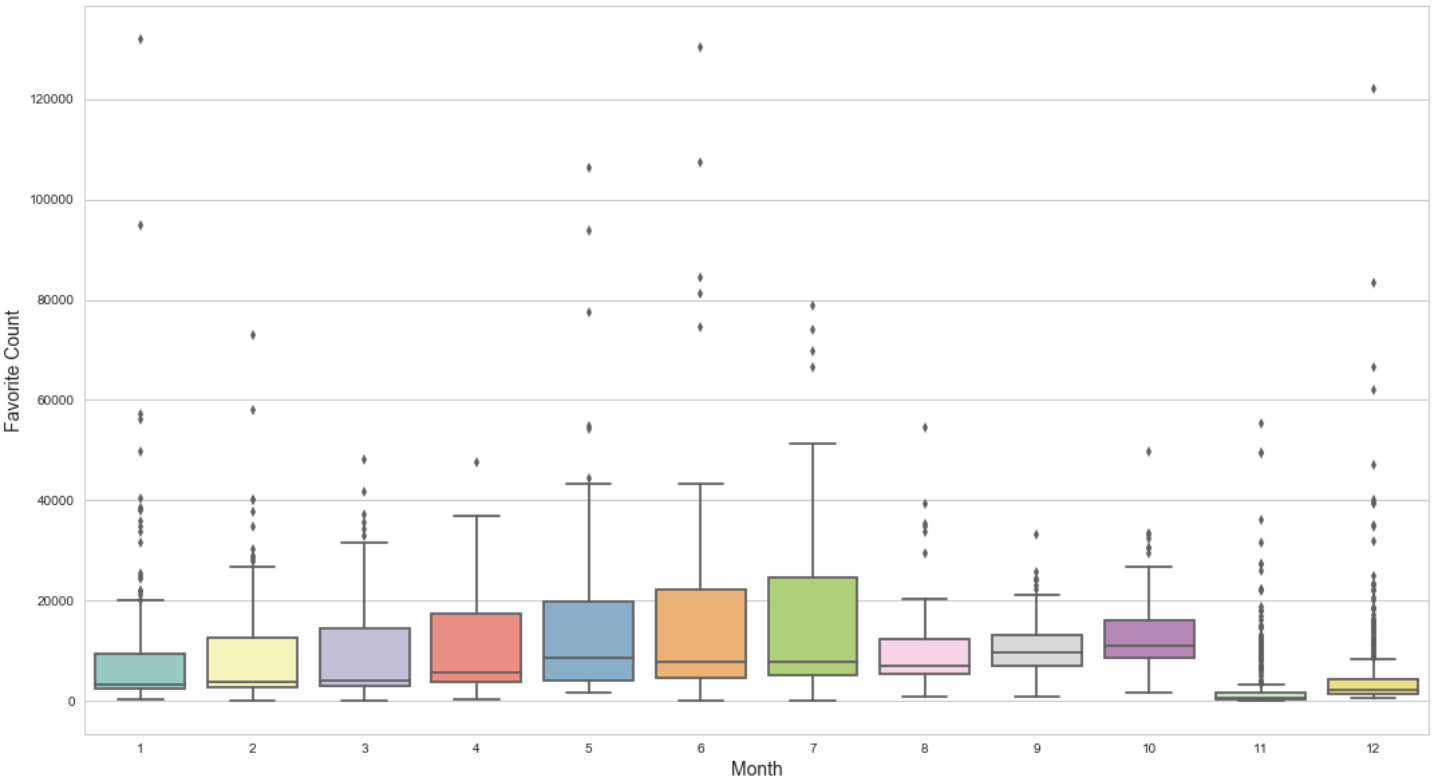
From the visualization above you will see a strong correlation between the favorite and retweet data, with a Pearson correlation coefficient, r, equal to 0.92. It makes logical sense that popular tweet will have a high favorite and retweet count.

With this new insight into the tweet data I was curious, what time of year is the most active time for people to tweet about their beloved companion?

Retweet Count per Month



Favorite Count per Month



As you can see from the visualizations above, for any given month the majority of retweets averages less than 10,000. Activity spikes in the summer months with June containing the largest outlier with ~80,000 retweets, while May has the highest occurrence of retweets. There are a greater number of people “liking” a tweet as compared to the number of retweets. Notice the steady increase in activity from January until it reaches it's maximum in the summer months. The maximum for retweets is reached in June, while the favorite count has a maximum in July. I’m curious as why the favorite count drops steeply below 20,000 in August, with a less noticeable drop in the retweet graph. In both graphs, November and December have the lowest tweet activity, with December having the largest statistical spread in both visualizations.

In summary, the data analysis of the Twitter user WeRateDogs, showed the most common dog is Charlie. The favorite and retweet data is strongly correlated and the most active time of year for people to tweet about dogs is in the summer months. An adorable Labrador puppo has the lowest rating, and a doggo named Attitus fooled the neural network dog classier by having the best dressed doggo outfit at the 4th of July BBQ!