Data Wrangling and Analysis: WeRateDogs Twitter Activity

By: Daniel Mayer

Introduction:

Source: Google Images 'WeRateDogs'

WeRateDogs (@dog_rates) is a popular Twitter account that rates people's dogs and provides a humorous comment about the dog. Ratings are out of 10, but many of the dogs intentionally receive scores that exceed 10/10 (i.e. 11/10, 12/10).

In this project I will be wrangling, analyzing and visualizing the WeRateDogs Twitter archive data. Note that the tweet data used in this project are from 2015-2017 and the popularity of the account has grown a lot in the years since, now having nearly 9M followers. This report contains the final insights and visualizations from the project. For more information on the wrangling process, including the steps taken to gather, assess, and clean the data, see the *Wrangle Report.*
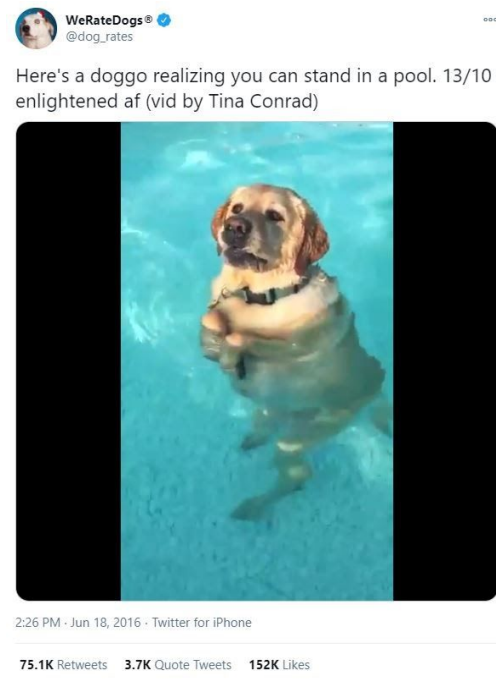
Section 1: Popularity and Ratings

Most Popular Tweet

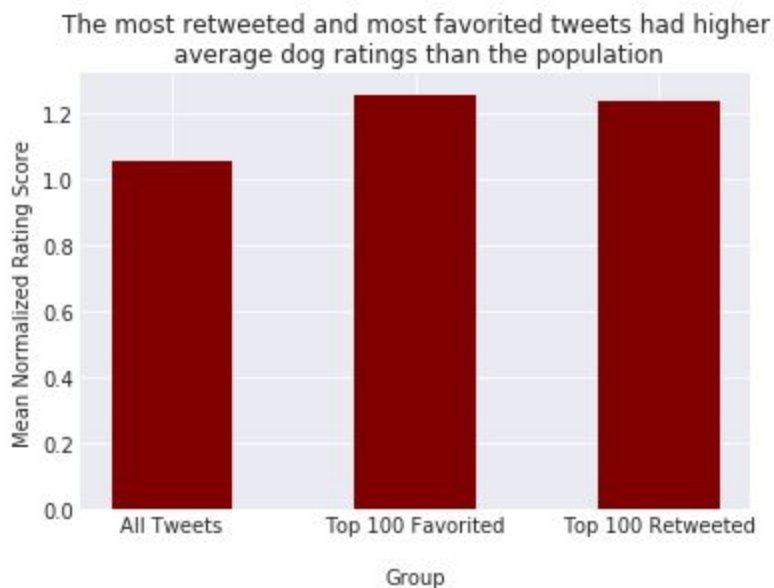This tweet of a 'doggo' standing in a pool was both the account's most retweeted and most favorited tweet:

Retweets: 75,173
Favorites: 152,086

Source: Twitter @dog_rates

Comparing Average Rating Scores of the Most Popular Tweets vs Baseline
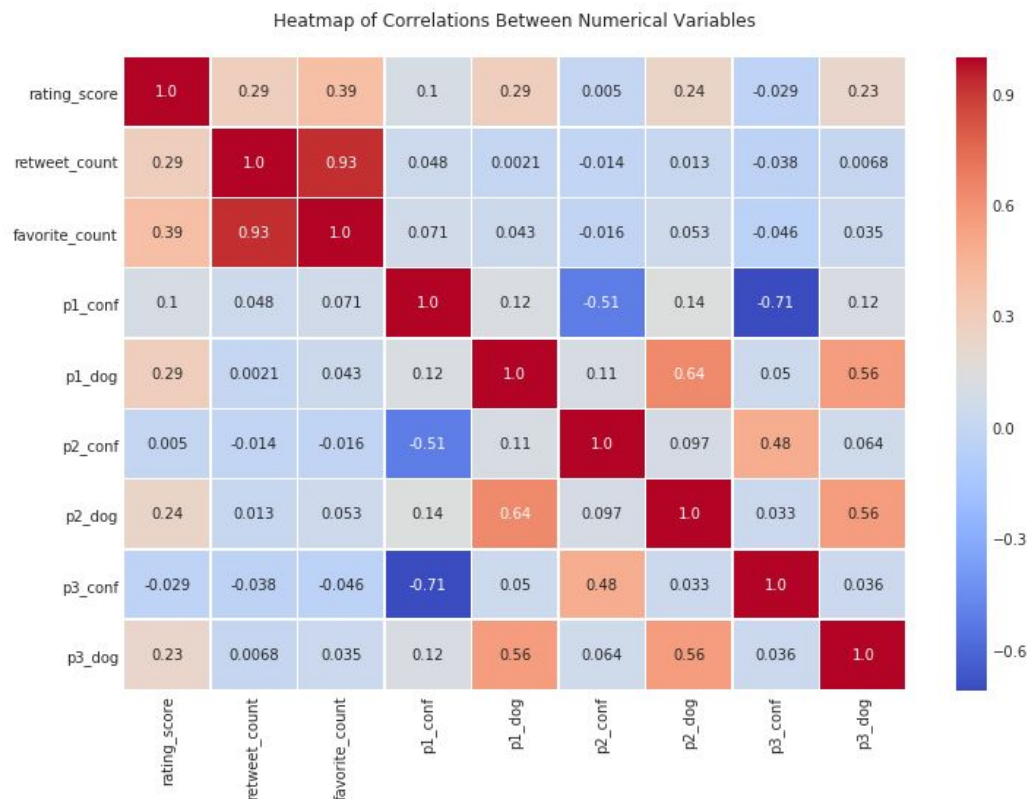

The most retweeted and most favorited tweets in the sample had received on average a 0.2 higher rating score from the WeRateDogs Twitter account. For context, this is equivalent to an

average rating of 12.4/10 (top 100 most retweeted) and 12.5/10 (top 100 most favorited) versus 10.5/10 (entire sample)

Variable Correlations:

The heatmap shows us which variables are closely tied together. Red indicates a positive relationship (as one variable increases, the other increases as well) and blue indicates a negative relationship (as one variable increases, the other decreases). The darker colours indicate a stronger relationship, with 1.0 being a perfect positive correlation and -1.0 being a perfect negative correlation.

Heatmap of Correlations Between Numerical Variables

|	rating_score	retweet_count	favorite_count	p1_conf	p1_dog	p2_conf	p2_dog	p3_conf	p3_dog
rating_score	1.0	0.29	0.39	0.1	0.29	0.005	0.24	-0.029	0.23
retweet_count	0.29	1.0	0.93	0.048	0.0021	-0.014	0.013	-0.038	0.0068
favorite_count	0.39	0.93	1.0	0.071	0.043	-0.016	0.053	-0.046	0.035
p1_conf	0.1	0.048	0.071	1.0	0.12	-0.51	0.14	-0.71	0.12
p1_dog	0.29	0.0021	0.043	0.12	1.0	0.11	0.64	0.05	0.56
p2_conf	0.005	-0.014	-0.016	-0.51	0.11	1.0	0.097	0.48	0.064
p2_dog	0.24	0.013	0.053	0.14	0.64	0.097	1.0	0.033	0.56
p3_conf	-0.029	-0.038	-0.046	-0.71	0.05	0.48	0.033	1.0	0.036
p3_dog	0.23	0.0068	0.035	0.12	0.56	0.064	0.56	0.036	1.0

We can see that certain variables like retweets and favorites have a strong positive correlation, meaning a tweet with many retweets will also have many favorites. There is also a moderate positive correlation between a tweet's rating score and the number of favorites it receives (0.390) and a slightly milder positive correlation with the number of retweets it received (0.292).

We can also see from this table that an image being classified as a dog (p1_dog, p2_dog, and p3_dog) correlated positively with rating. It's too bad for all the other creatures and things posted in photos, but this account is impartial to dogs!
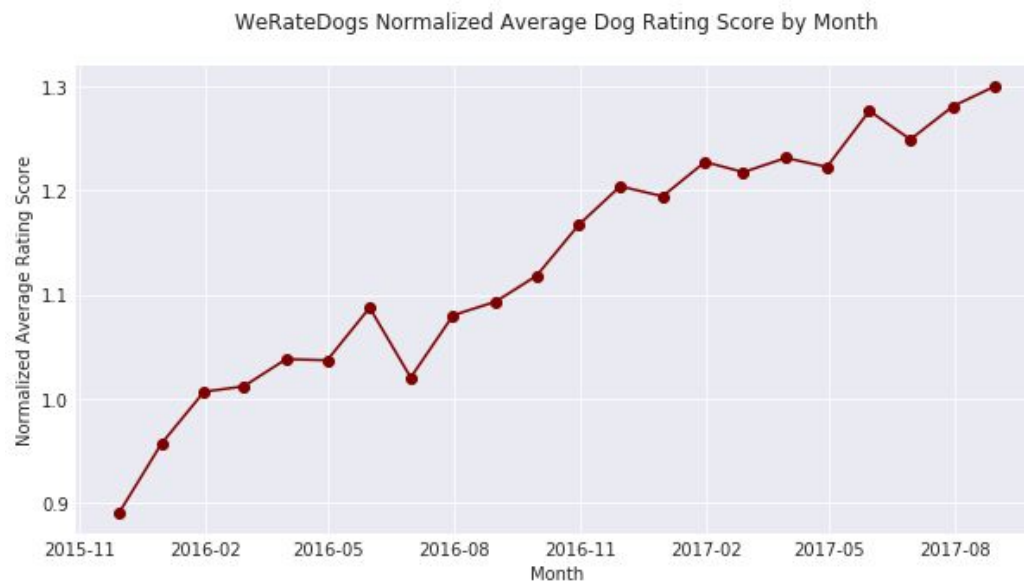
Section 2: Activity Patterns over Time

Tweets Per Month

#WeRateDogs Number of Unique Tweets by Month
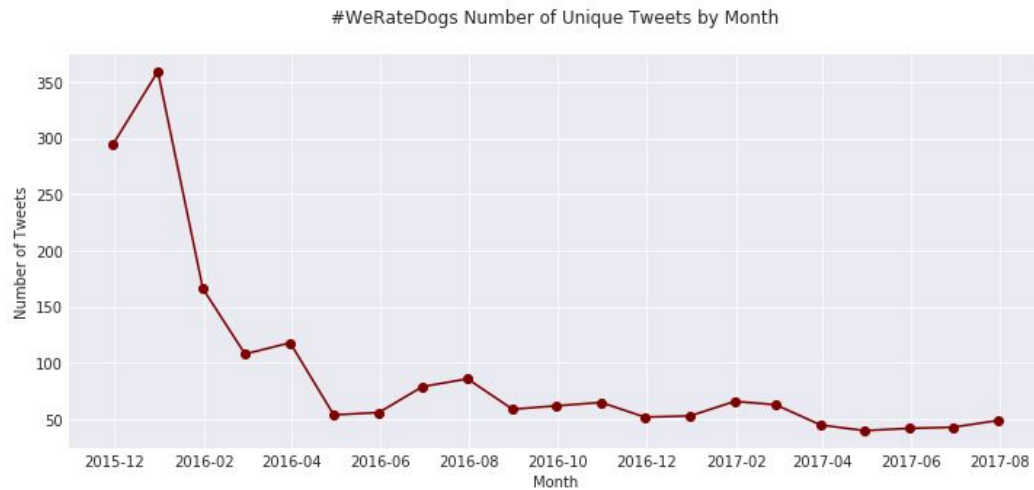
After peaking at over 350 tweets in January 2016, the activity levels dropped and have since held steady at around 50 tweets per month.
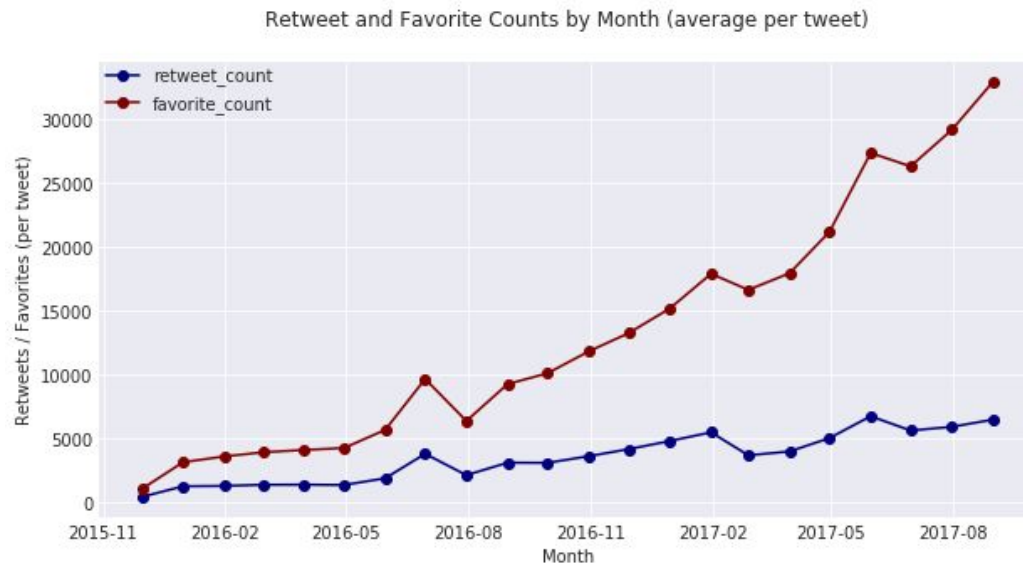
WeRateDogs Rating Scores

WeRateDogs Normalized Average Dog Rating Score by Month

The WeRateDogs staff are getting more generous with their rating scores over time, with the average score approaching 13/10 in summer 2017, compared to just under 9/10 at the end of 2015.

This may also help explain the positive correlations between rating scores-retweets and rating_scores-favorites. The account has gained many followers over time, meaning the more recent tweets (which also happen to be rated higher) will have more retweets and favorites. This assumption can be validated in the plot below that analyzes the mean numbers of retweets and favorites per month. For a more in-depth analysis, I could also go back and pull the WeRateDogs follower numbers over time from the Twitter API.

Retweets and Favorite Counts Over Time

Retweet and Favorite Counts by Month (average per tweet)

Section 3: Image Classification Algorithm Statistics

Breeds that Received the Highest Average Rating from WeRateDogs

	dog_breed	rating_score
0	Saluki	1.250000
1	briard	1.233333
2	Tibetan_mastiff	1.225000
3	Border_terrier	1.214286
4	standard_schnauzer	1.200000
5	silky_terrier	1.200000
6	Eskimo_dog	1.177778
7	Gordon_setter	1.175000
8	Irish_setter	1.175000
9	Samoyed	1.174359

Saluki. *Source: TheSprucePets*

## Images that the algorithm predicted most confidently:
(Note - no dog breeds!)

The first photo is of the so-called 'ping-pong ball' that the algorithm identified with 99.999% confidence...clearly some room for improvement. Though it did do a great job correctly identifying the peacock.

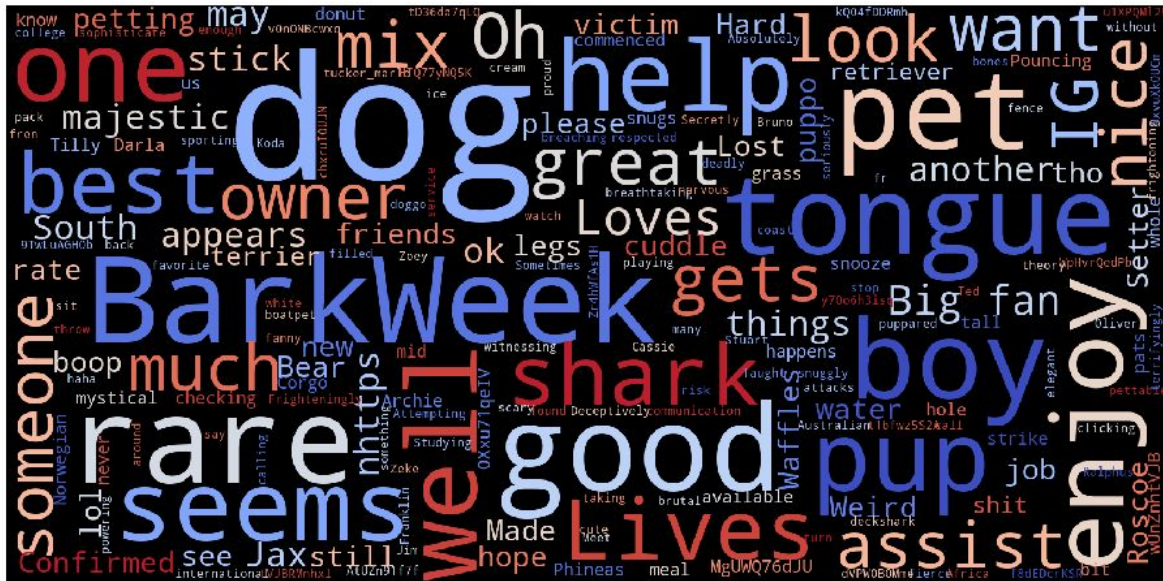|	image_prediction	confidence_level
0	ping-pong_ball	0.999945
1	peacock	0.999924
2	school_bus	0.999833
3	bib	0.998814
4	slug	0.998075
5	zebra	0.997673
6	fountain	0.997509
7	china_cabinet	0.996031
8	flamingo	0.992710
9	fiddler_crab	0.992069

Source: Twitter @dog_rates

Source: Twitter @dog_rates
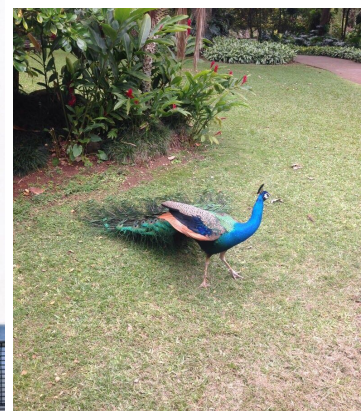
Section 4: Word Clouds

A summary of the most common words across the text of all WeRateDogs tweets between November 2015 and August 2017

And just for fun a word cloud in the shape of a dog...

Sources:
https://weratedogs.com/
https://www.thesprucepets.com/saluki-dog-breed-profile-4685967
https://pbs.twimg.com/media/CU3FbQgVAAACdCQ.jpg
https://pbs.twimg.com/media/CU3qHNTWsAApGr0.jpg