WeRateDogs -

A Journey into their Twitter Data

If you've ever asked yourself where you could get a humorous comment about your dog – WeRateDogs is the answer. WeRateDogs is a twitter account where you can send a picture of your dog to and afterwards WeRateDogs selects some and tweets these pictures with a funny comment and rating. These ratings should vary between 0 to 10 on a 10-point scale but in truth, they often exceed 10

Lucky me, I've got data from their twitter archive (from November 2015 to August 2017), gathered data via Twitters API with tweepy (favorite and retweet count) and got further data from Udacity. An Udacity employee ran every image in the WeRateDogs Twitter archive through a neural network that can classify breeds of dogs. Resulting in a table full of the top three image predictions alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images).

So, I've asked myself some questions. Like: What is the most popular twitter source? What is the monthly number of tweets? How many dogs are rated above 10 and how does the rating distribution look like?

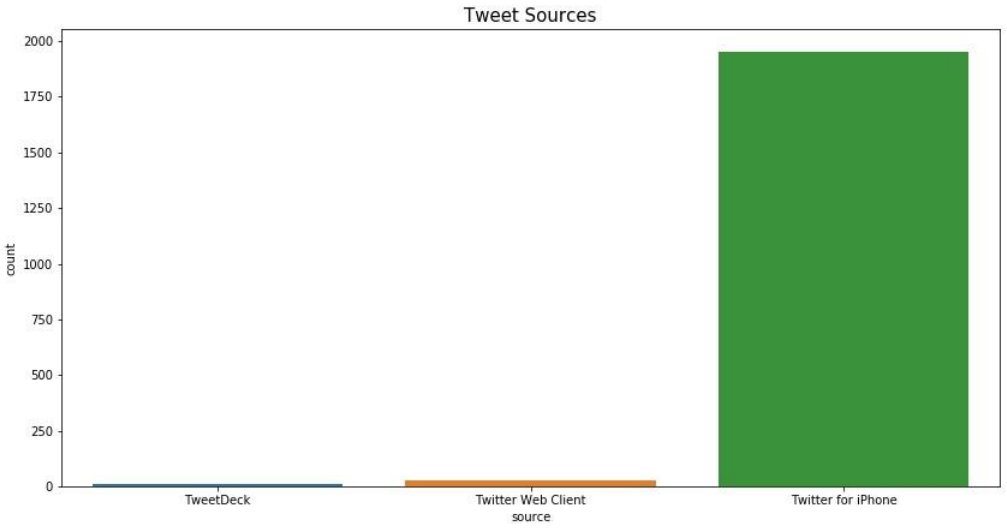
But before I started answering these questions, I wrangled these three datasets, created new variables and merged all variables, that were interesting to me, into one dataset. This final dataset contains 1993 entries and 14 columns. These columns are: tweet_id, timestamp, source, text, rating_denominator, rating_numerator_new, names (of the dogs), dog_stage (pupper, puppo, dogger and floofer), jpg_url, favorite_count, retweet_count, breed_pred (prediction of the dogs breed), pred_confidence, and fraction (which is the rating_numerator_new divided by the rating_denominator).

Some columns have missing values:

Names (597), dog_stage (1666), rating_numerator_new and fraction (27), breed_pred and pred_confidence (308).

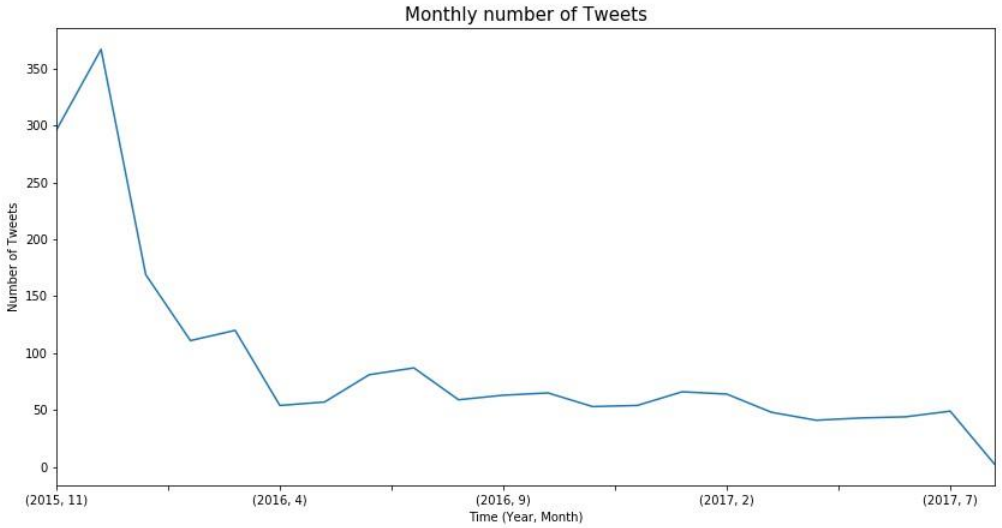
This should be kept in mind while looking at the visualizations below.

What is the most used source?



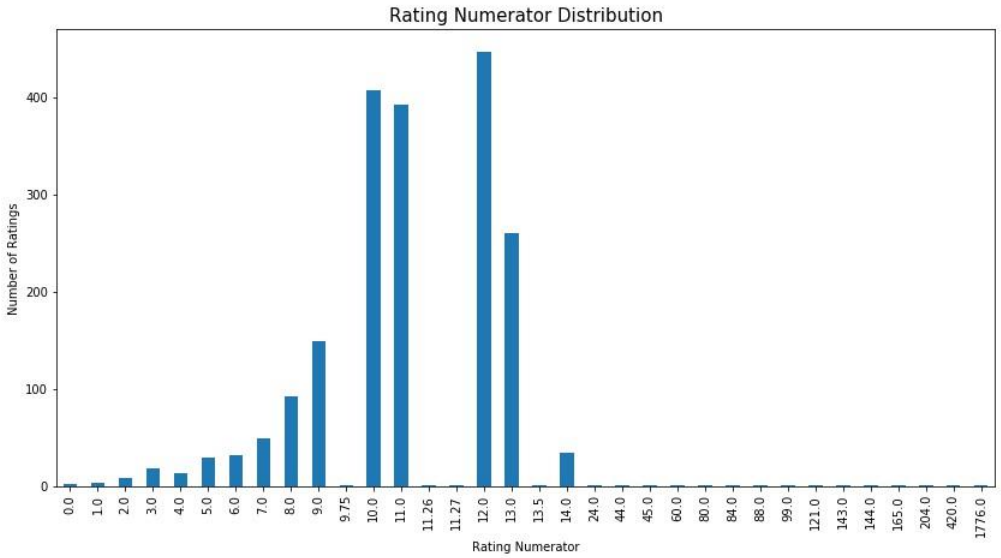
By far the most popular source is Twitter for iPhone (1954 counts) followed by the Twitter Web Client (28) and TweetDeck (11).

What is the monthly number of tweets?



Most tweets were posted in December 2015 (367 tweets). Afterwards the number of tweets decreased rapidly April 2016 and remained fairly constant since then until July 2017.

What does the rating numerator distribution look like?

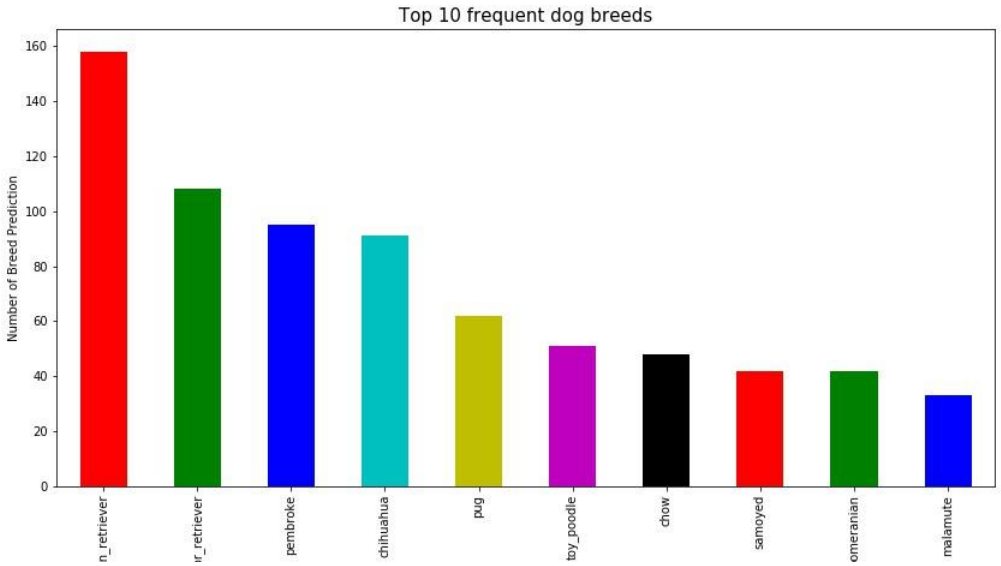


12 was the most often assigned numerator and most rating numerators was above 10 (for 1155 dogs out of 1966 dogs with a valid rating). And 1155 dogs were rated above 10.

How many dog breeds could be predicted by the applied algorithm?

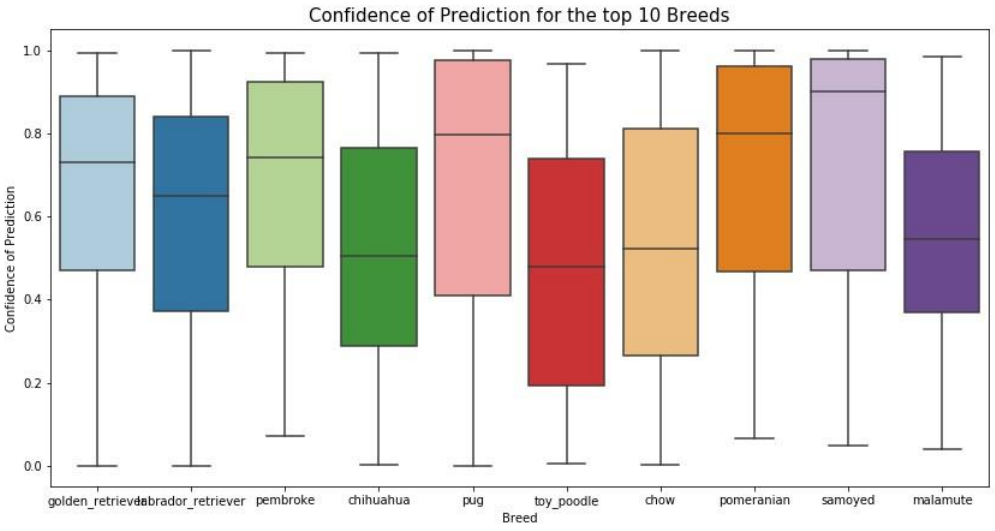
The algorithm could predict the breed of 1685 dogs and failed in 308 cases.

What are the 10 most frequent predicted dog breeds?



The most frequent predicted breed was golden retriever with 158 predictions.

How confident was the algorithm for these top 10 breeds?



These are the top 10 predicted breeds with the corresponding confidence for their prediction. Clearly, the confidence varies between these dog breeds. For some breeds, the algorithm was pretty sure (i.e., pug, pomeranian, samoyed) and for others rather unsure (median confidence around 50%), like for chihuahua and toy_poodle.

What is the difference of the retweet and favorite count depending on the presence of a dog's name and the rating numerator?

The mean Retweet Count for Dog Rating Numerators above 10 is 3730.0.

The mean Retweet Count for Dog Rating Numerators under 10 is 3139.0.

The mean Favorite Count for Dog Rating Numerators above 10 is 12695.0.

The mean Favorite Count for Dog Rating Numerators under 10 is 10443.0.

There is only a slight difference in the retweet count between ratings with a numerator above or under 10. The difference between them regarding the favorite count is more visible. So, if your dog got a rating numerator above 10 there is a good chance your dog will get more favorites ;).

The mean Retweet Count for Dogs without a name is 2854.0

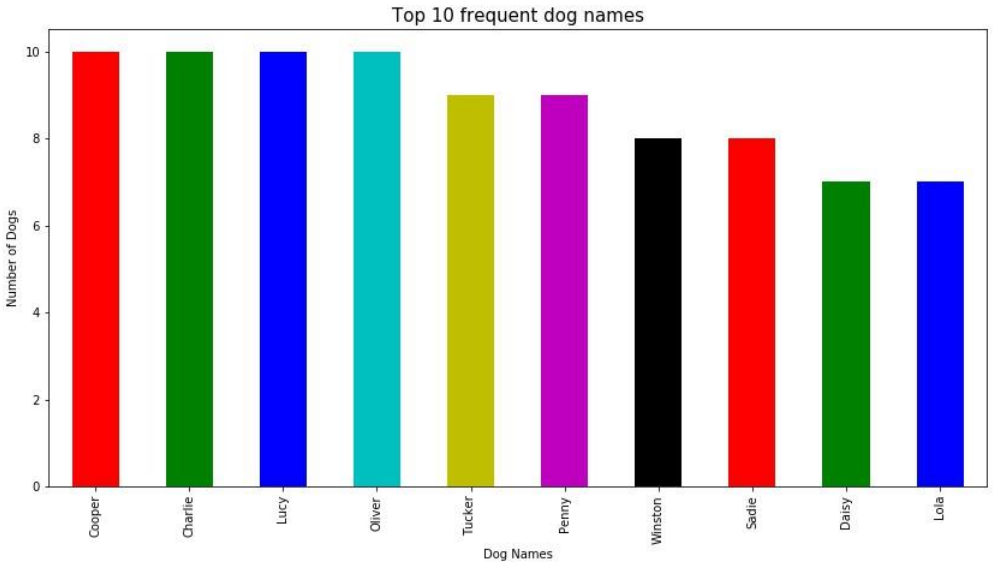
The mean Retweet Count for Dogs with a name is 2593.0

The mean Favorite Count for Dogs without a name is 8236.0

The mean Favorite Count for Dogs with a name is 9032.0

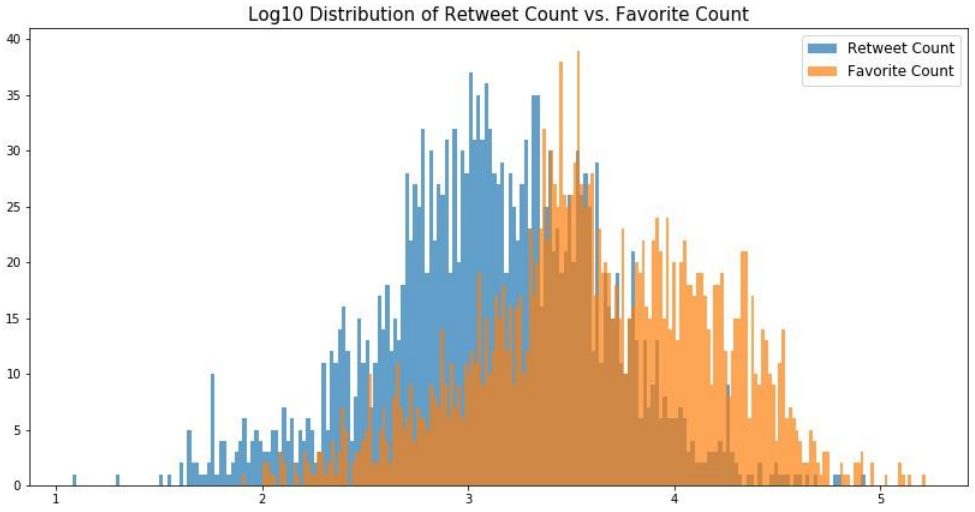
Tweets including dog names are more likely to be favorited but slightly less likely to be retweeted.

What are the 10 most frequent dog names?



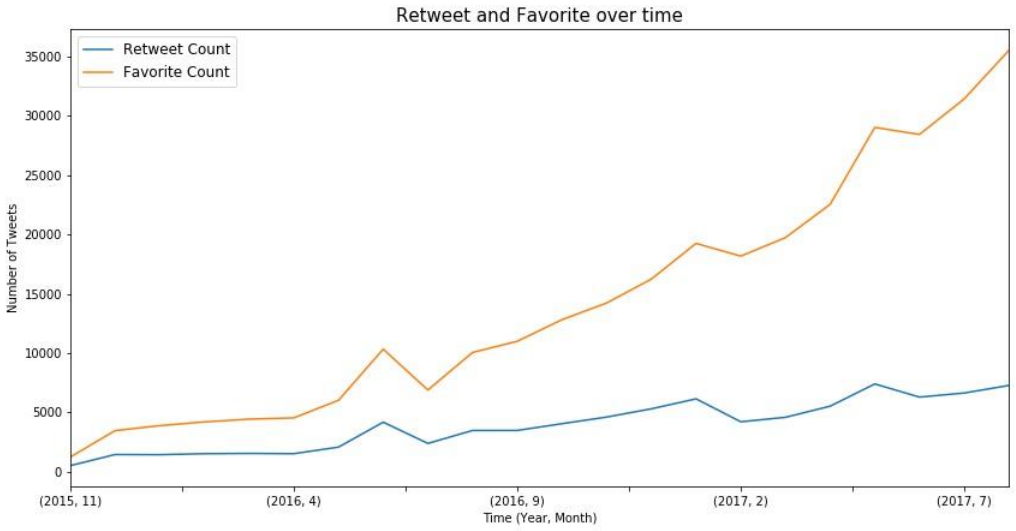
Cooper (10), Charlie (10), Lucy (10), Oliver (10), Penny (9), Tucker (9), Sadie (8), Winston (8), Daisy (7), and Lola (7) are the most frequent dog names in this dataset.

How looks the Distribution of Favorite Count compared to Retweet Count?



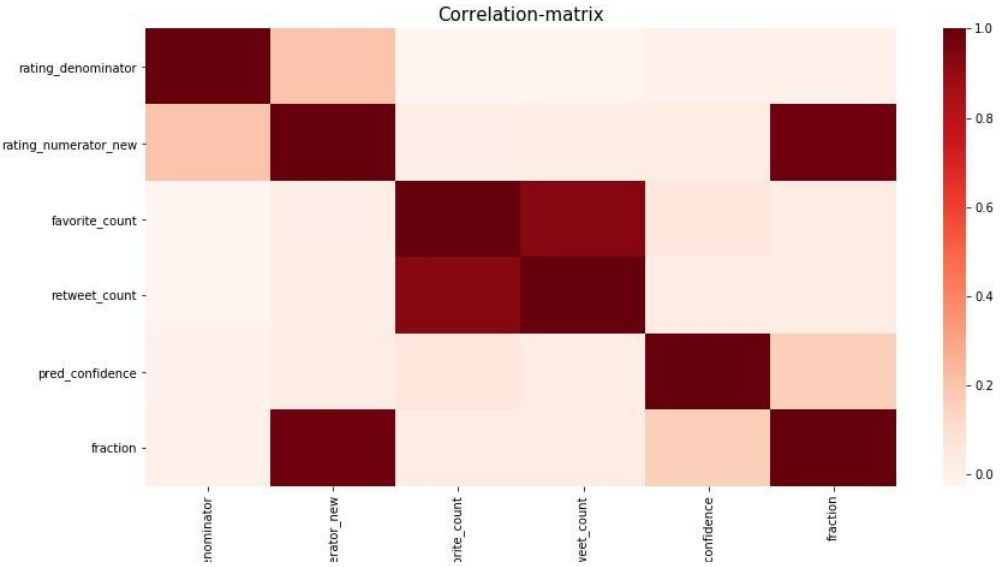
We can see that the distribution of favorite counts is located to the right of the distribution of retweet count. Thus, people favor the tweets more often than retweet them.

How changed the Retweet and Favorite Count over time?



The first thing that caught my eye is that there are far more favorites than retweets. Both, favorites and retweets, increased over the time. While the favorite count increases strongly with the number of tweets, the retweet count seems almost indepedent of the number of tweets.

How are the variables correlated with each other?



There are some weak and strong correlations. It is not surprising that fraction highly correlates with rating_numerator_new (as its calculated by this variable). Furthermore, retweet_count correlates highly with favorite_count (which is also no surprise).