Wrangling and Analyzing WeRateDogs

By Patrick Amaral
July 1, 2019

Introduction

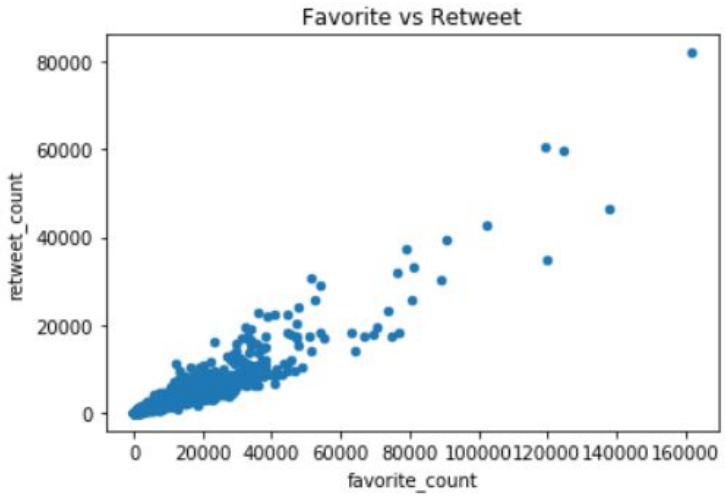
First, WeRateDogs is a Twitter page, that rate people's dogs with curious comments, but these rating numerators are almost always greater than 10 and sometimes the denominator are too big like 15/10, 13/10 or 10/300. In theory, the rates should be 1 to 10. However, this is what made the WeRateDogs popular, today, the WeRateDogs has 8 million followers.

So, about data, this analysis was used three datasets, two of them is available by Udacity's resources and one them I need to get using a python library called tweep, that allows me to access the Twitter API. The file that Udacity get is exclusive, their contain approximately 3000+ tweets with (tweet_id, text, timestamp, rating_numerator, rating_denominator, expanded_urls, name, etc) and another Udacity's file is the result of predictions using the neural network on each image posted, and the last file is programmatically downloaded using python libraries request and tweep, I needed just two additional information, quantity that tweet was favorited and retweeted.

Before beginning the analysis, I needed wrangling data using technics like gathering, accessing and cleaning data. So how I said, to gathering the missing data I used tweep and to read these datasets was used normal functions of Pandas, so in accessing step I analyze using some programmatically methods, I found some problems of tidiness and quality like 'Missing some expanded_urls, many names aren't real names, sometimes is just a letter or syllable, join the dataframe twitter with twitter_plus'. The final step is cleaning, all points of tidiness and quality fixed up.

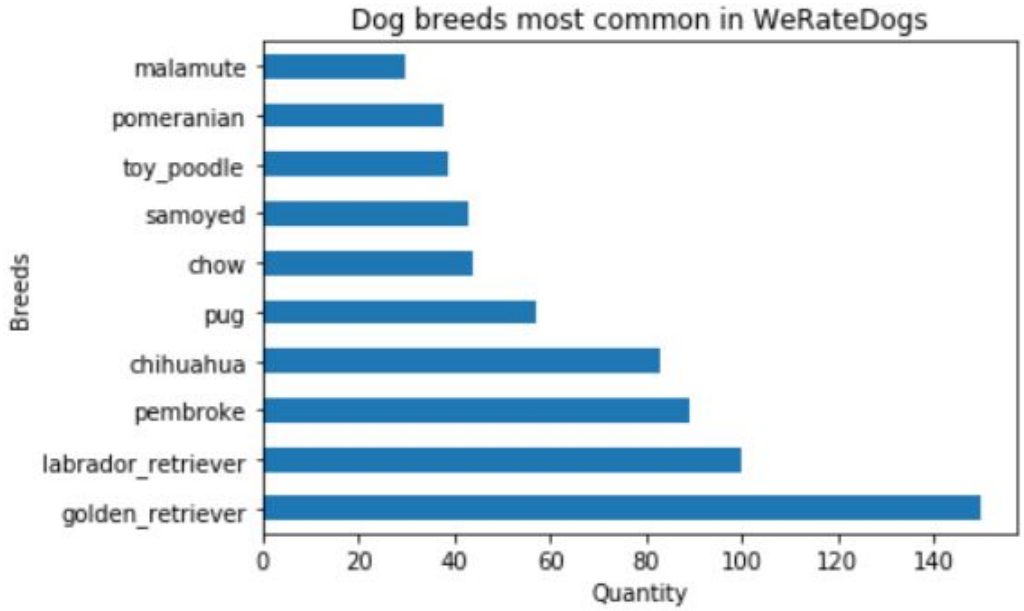
Analyzing & Visualizing

Analyzing if had correlation between the number of retweets and number of favorites, I found that both increase together, so if it has a great number of favorites is probably that has a great number of retweets. It's better illustrated in the scatter plot below.

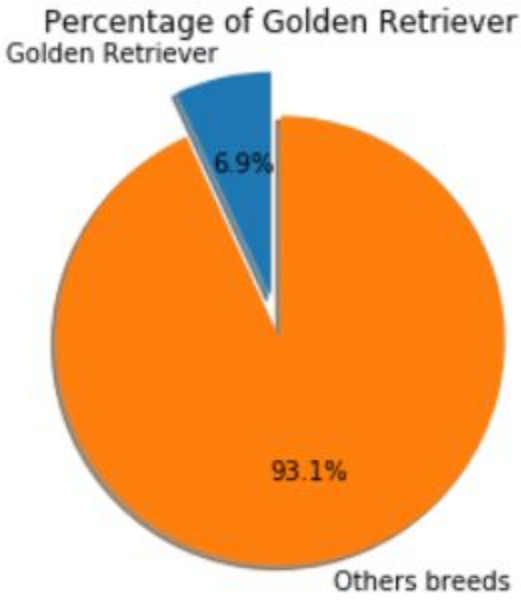


Now another question is which dog name is most common on WeRateDogs? Well, a get at least four dog name because in this dataset are many null values. So Lucy, Charlie, Oliver, and Cooper have the same quantity, to make clear, each name has only ten registers and the null register has seven hundred eighty-four.

Okay, let's see if there is a dog breed most popular in horizontal bar plot below.



Looking at the horizontal bar plot above the Golden Retriever is the most popular and has some advantage with almost fifty registers of difference, so there is no doubt about your popularity, but looking at all data we can see the same thing?



Yes, isn't a big result but 7% is a good percentage over 1532 registers, so this is only an additional point to prove the popularity of Golden Retriever.

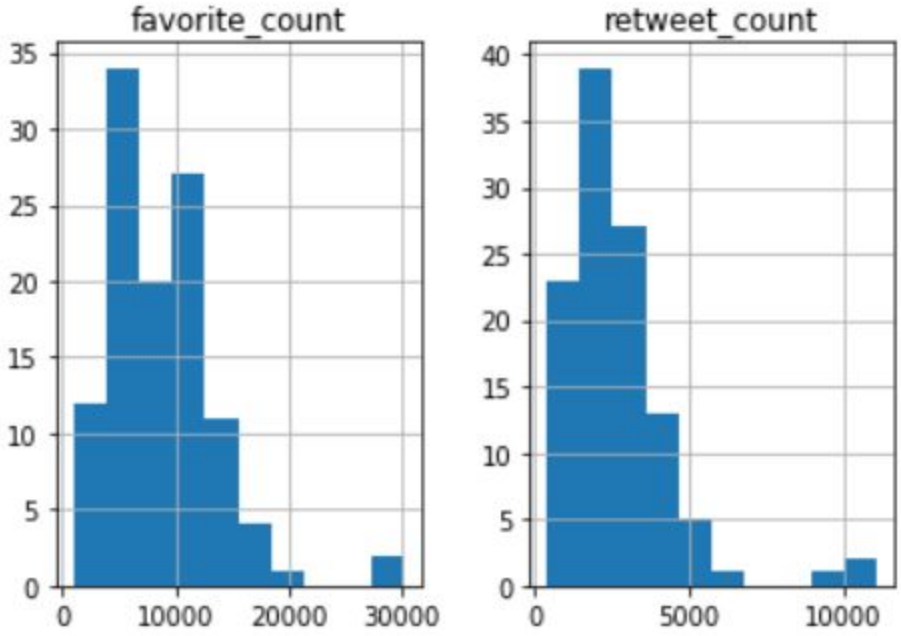
Well, we see that Golden Retriever is the most popular but which one is the least popular? In this analysis, I found that:

- Scotch Terrier
- Entlebucher
- Clumber
- Silky Terrier
- Groenendael
- Japanese Spaniel
- Standard Schnauzer

Are the least popular but a curious point is that Japanse Spaniel has the biggest average of favorited and retweet seems below.

	tweet_id	rating_numerator	rating_denominator	favorite_count	retweet_count
breeds					
japanese_spaniel	7.887659e+17	12.000000	10.000000	28891.000000	11065.000000
wire_haired_fox_terrier	6.877818e+17	10.500000	10.000000	17282.500000	10186.500000
irish_terrier	7.716884e+17	11.500000	10.000000	30215.500000	9402.166667
rhodesian_ridgeback	7.648635e+17	11.000000	10.000000	20126.000000	5962.250000
schipperke	7.239415e+17	10.300000	10.000000	11720.100000	5302.400000
miniature_pinscher	7.414337e+17	10.608696	10.000000	14006.217391	5137.565217
miniature_schnauzer	7.945474e+17	12.000000	10.000000	16391.250000	5103.000000
basset	7.307612e+17	10.384615	10.000000	12992.153846	4986.000000
saint_bernard	7.444181e+17	10.428571	10.142857	14118.857143	4940.142857
greater_swiss_mountain_dog	7.186492e+17	8.666667	10.000000	13539.333333	4607.333333

Yes, something is strange here, yes but it is because japanese_spaniel has just one register and this cause the irregular average, looking at a histogram can be more clear.



Summarizing using this analysis we can see which dogs are most shared in the page and why like your breeds, and how retweets and favorited are strongly correlated.