Analyses

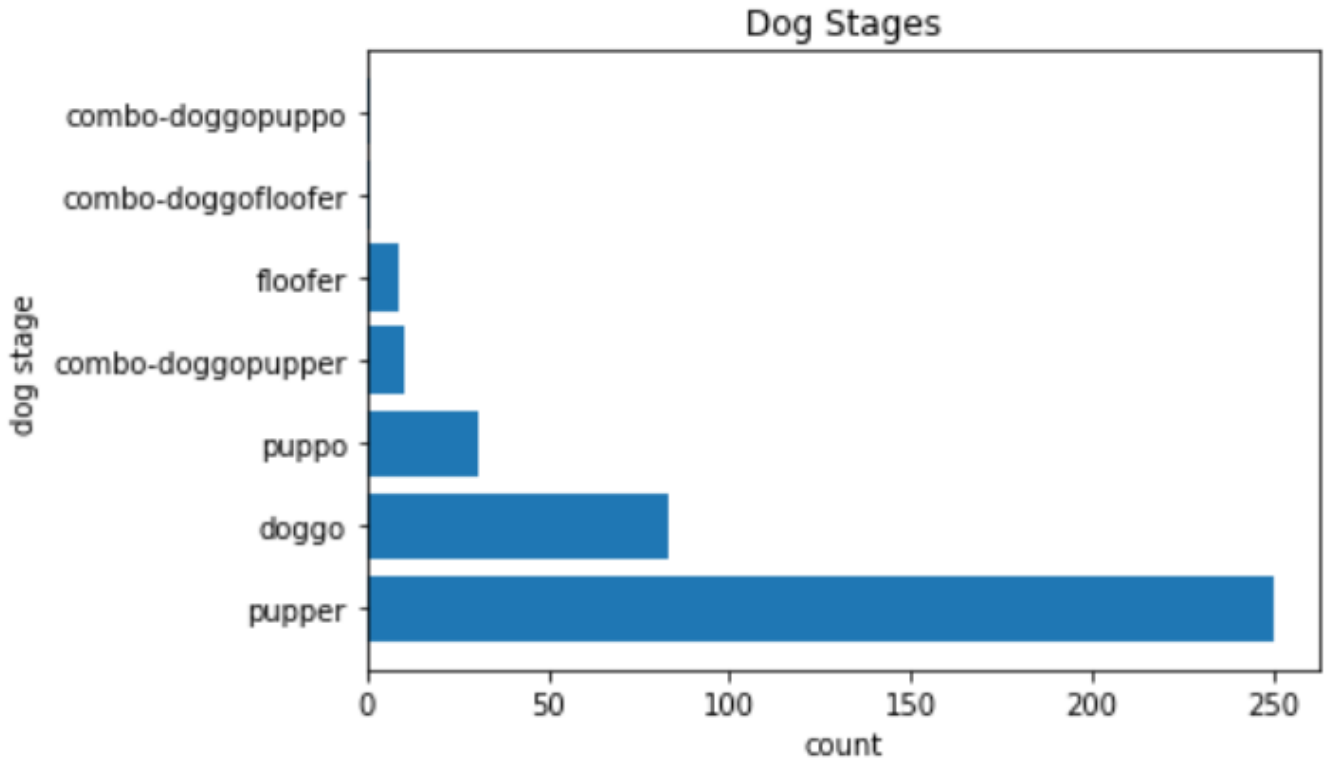
Prelude

Finally, here is an analysis report. It took a long time to get to this point. Initially I was terribly disappointed that we would be using Twitter which I thought was merely a platform used by racist nationalists in United States of America. But apparently, it's also used by dog lovers. I still do not know what Twitter is and its purpose. In spite of this, I worked diligently to get insights on the Twitter archives.

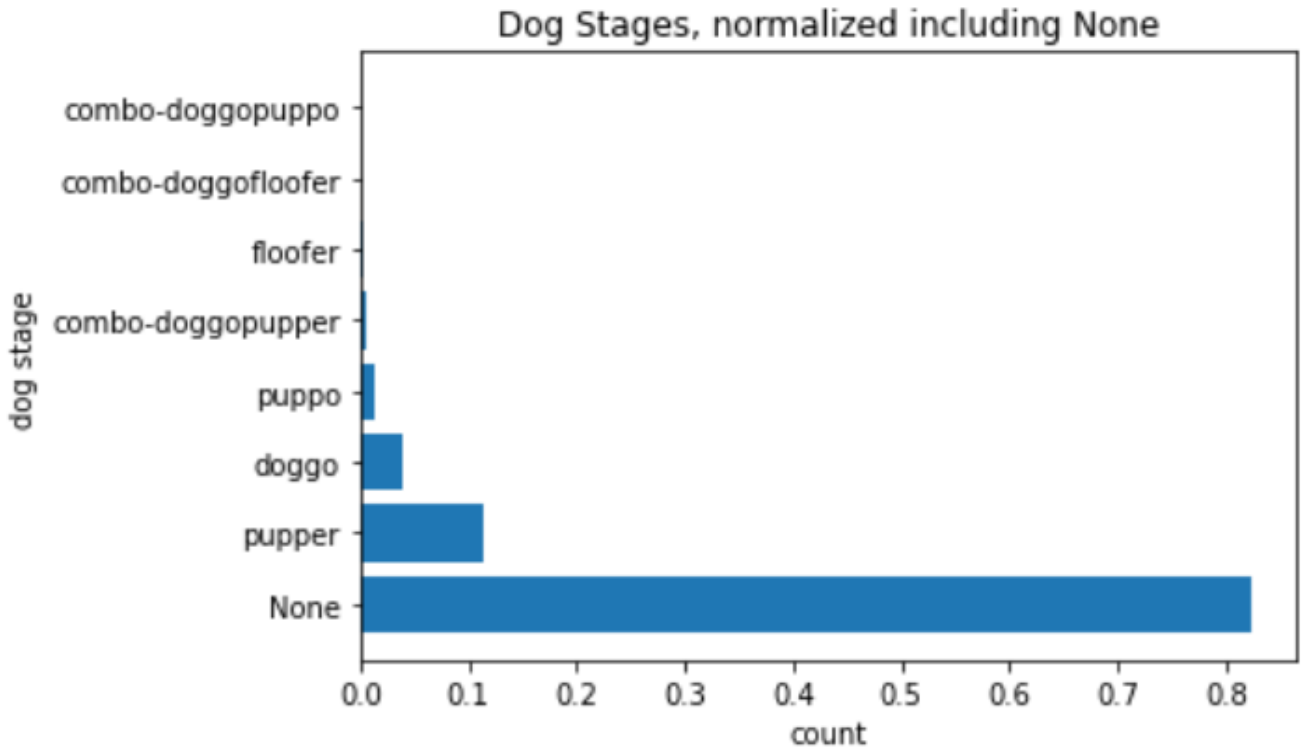
What is surprising is the challenging work was not in any way tedious. As I got more involved with wrangling the data, the more purposeful the work seemed to be. Unfortunately, the wrangling was so time consuming that I was concerned that my analyses would be underwhelming.

Dog Stages

Since it took so long to actually do the task of conflating the dog stages, I must provide some insight no matter how trivial. And here it is... a graphics of frequencies of dog stages.



It is clear that `pupper` is the most frequent followed distantly by `doggo`. However, solely to demonstrate the triviality of that analysis relative to the diligence involved in conflating the "values" columns into one variable, here is another look with `None` included and with normalized numbers. There are more than 80% percent of entries that had no dog stage indicated.



Prediction Confidence

I wish I could have gotten more information from doing the challenging task of combining image predictions and concomitant confidences into 2 variables, namely `dog_breed_prediction` and `dog_breed_confidence`. With my little knowledge of data analysis, the best I could do is that only 25% had a confidence of .75 or higher. It is curious that there is at least one confidence level at .99995. In fact, there are 2 entries with confidence level that high. They happen to have 4 images with which a prediction was made.

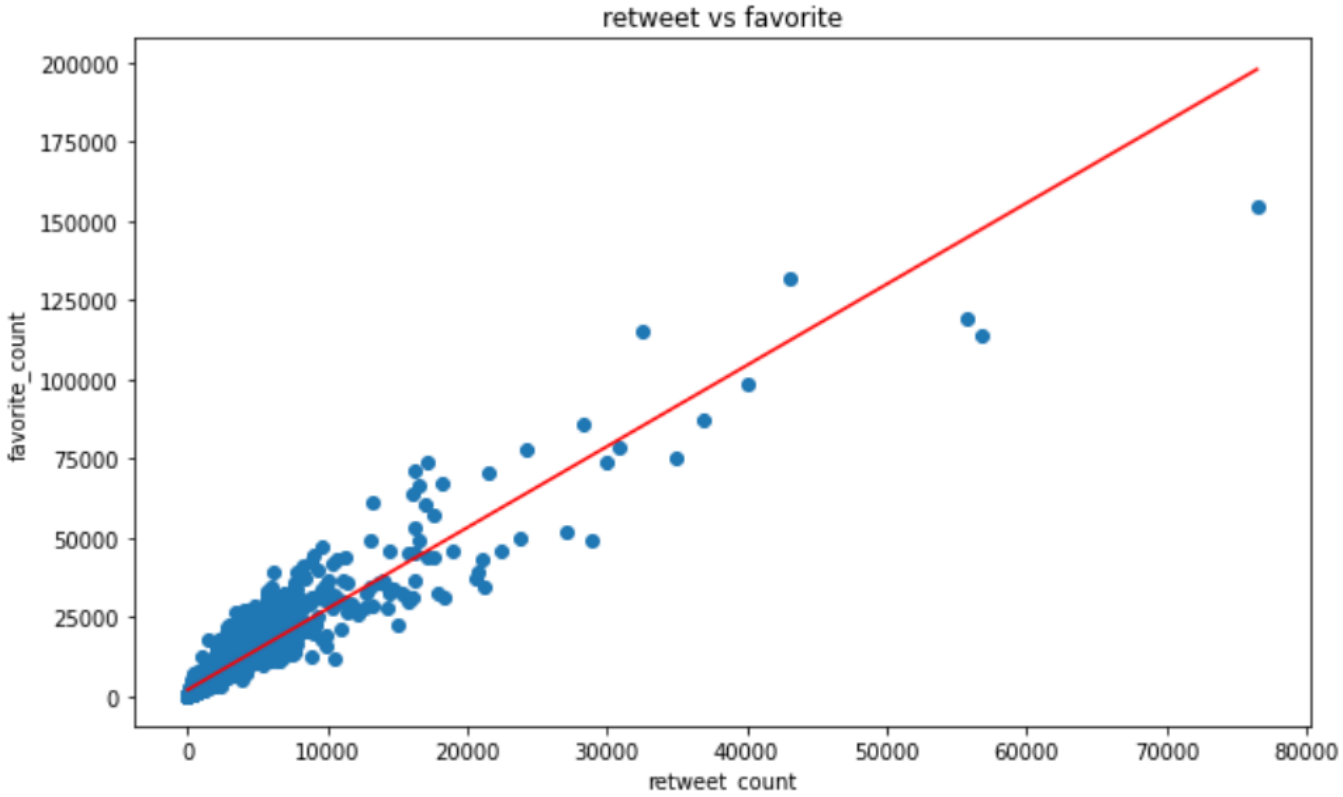
	tweet_id	rating_numerator	rating_denominator	favorite_count	retweet_count	img_num	dog_breed_confidence
count	2.175000e+03	2175.000000	2175.0	2175.000000	2175.000000	2175.000000	2175.000000
mean	7.371205e+17	12.247356	10.0	8006.537471	2401.287356	1.102989	0.426295
std	6.748668e+16	42.466313	0.0	11828.190067	4281.205119	0.631484	0.349498
min	6.660209e+17	0.000000	10.0	0.000000	0.000000	0.000000	0.000000
25%	6.768432e+17	10.000000	10.0	1685.000000	516.000000	1.000000	0.033697
50%	7.098528e+17	11.000000	10.0	3606.000000	1143.000000	1.000000	0.406509
75%	7.894226e+17	12.000000	10.0	9929.000000	2721.500000	1.000000	0.749136
max	8.924206e+17	1776.000000	10.0	154165.000000	76473.000000	4.000000	0.999956

There might be something to that .304 correlation coefficient between `img_num` and `dog_breed_confidence` after all.

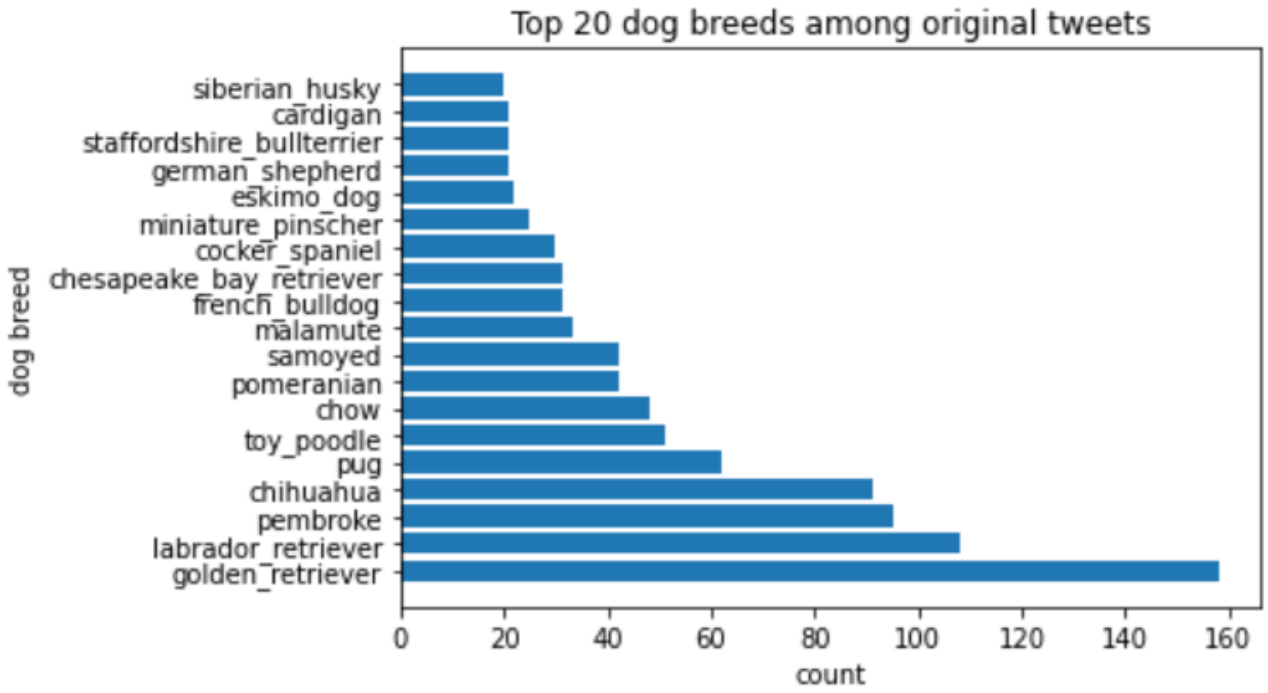
	tweet_id	rating_numerator	rating_denominator	favorite_count	retweet_count	img_num	dog_breed_confidence
tweet_id	1.000000	0.049523	NaN	0.584286	0.360422	0.140386	0.108018
rating_numerator	0.049523	1.000000	NaN	0.014479	0.015269	-0.022185	-0.034274
rating_denominator	NaN	NaN	NaN	NaN	NaN	NaN	NaN
favorite_count	0.584286	0.014479	NaN	1.000000	0.926839	0.128061	0.077764
retweet_count	0.360422	0.015269	NaN	0.926839	1.000000	0.088117	0.026103
img_num	0.140386	-0.022185	NaN	0.128061	0.088117	1.000000	0.304077
dog_breed_confidence	0.108018	-0.034274	NaN	0.077764	0.026103	0.304077	1.000000

Frequency Correlation

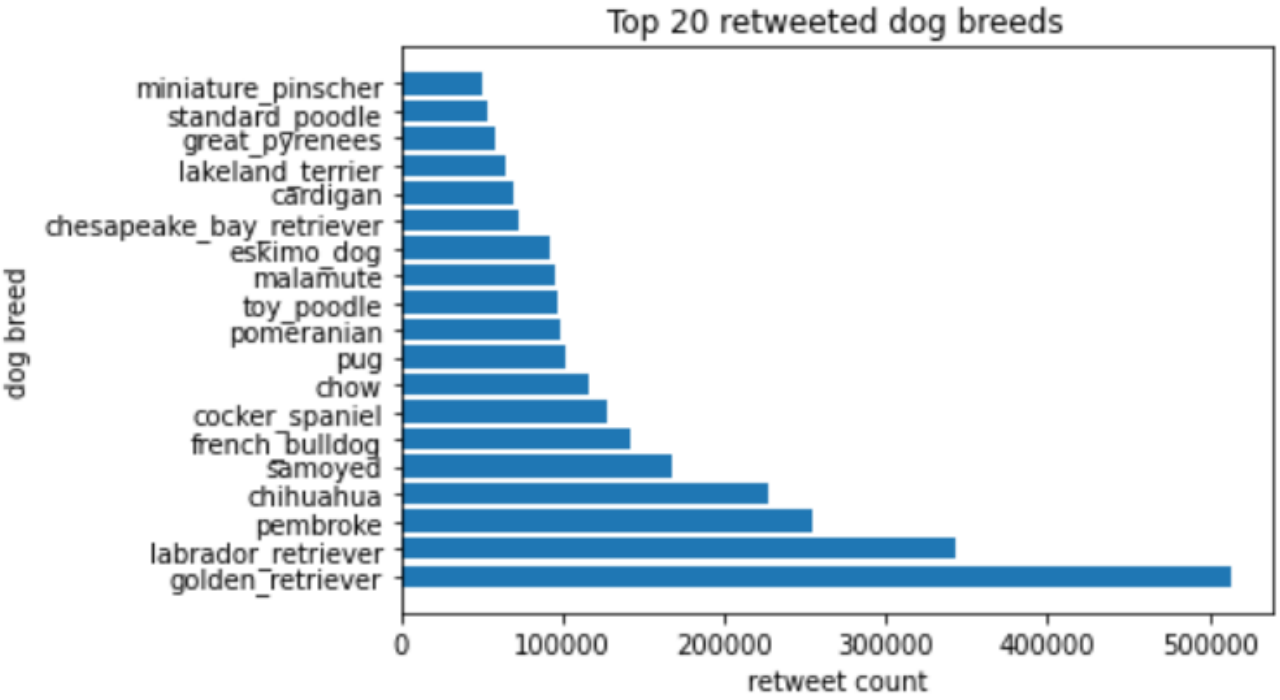
There is a definite high positive correlation between `retweet_count` and `favorite_count` .



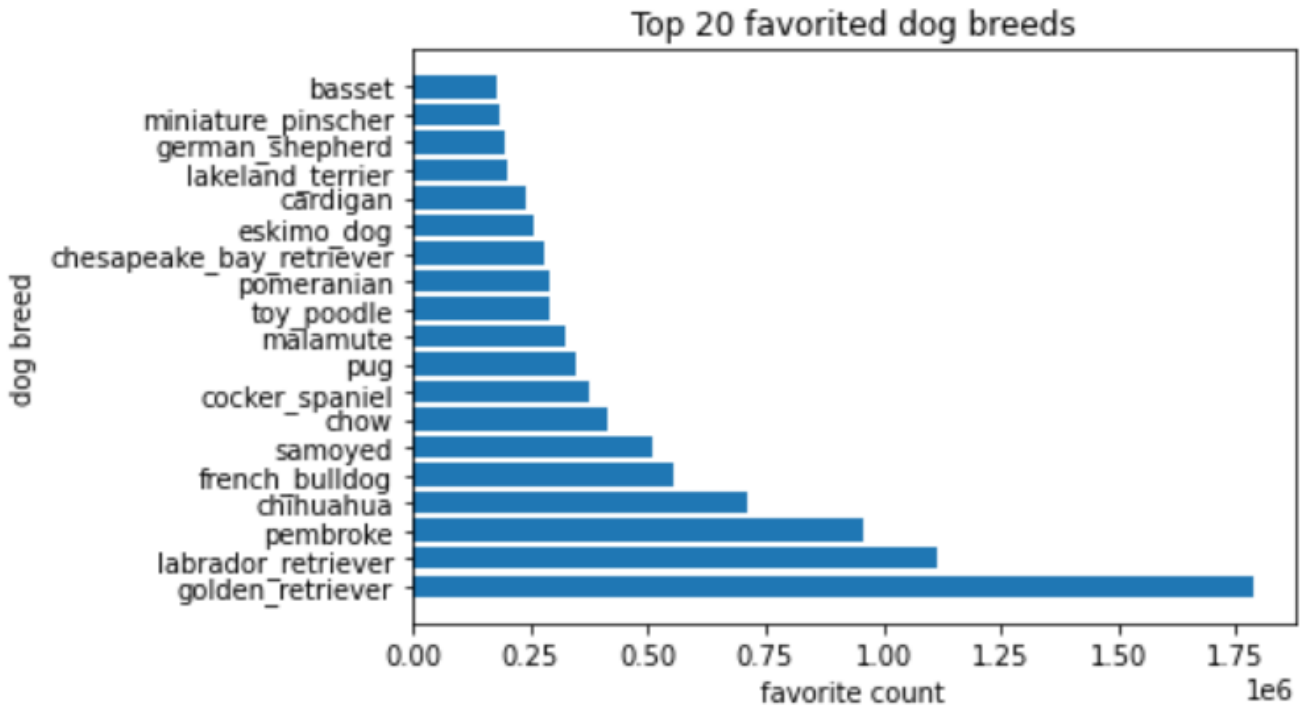
Top 20 originally tweeted dog breeds



Top 20 retweeted dog breeds



Top 20 favorited dog breeds



Upon seeing that the top 4 originally tweeted, retweeted and favorited dog breeds happened to be `golden_retriever`, `labrador_retriever`, `pembroke` and `chihuahua`, I had to do a correlation matrix of those 3 counts.

And I have happy results. They are all highly correlated as the table below illustrates. I have no assumptions since I have no insights on the behavior of tweeterers who rate dogs.

	retweet_count	favorite_count	orig_tweet_count
retweet_count	1.000000	0.995317	0.981340
favorite_count	0.995317	1.000000	0.965946
orig_tweet_count	0.981340	0.965946	1.000000