Analysis of We Rate Dog Tweets through August 1, 2017

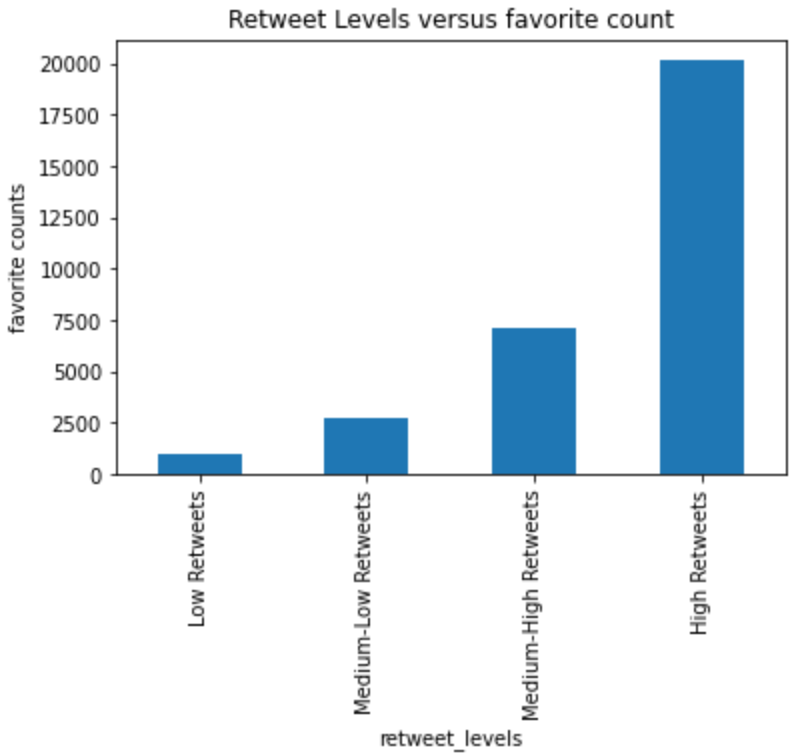
We analyzed the following from the We Rate Dogs tweet through August 1, 2017:

1. The relationship between retweet_count (how many times a tweet is retweeted) and favorite_count (how many times a tweet is favorited).
2. The relationship between whether a post mentions a “doggo” or a puppy (we combined the “puppo” and “pupper” columns into one “puppy” column, since these two terms seem to be used interchangeably to describe puppies) and how many retweets the post receives.
3. Which dog breeds receive the most retweets. For this one, we used the automated image recognition software’s best-guess about the breeds of the dogs featured in the tweets.

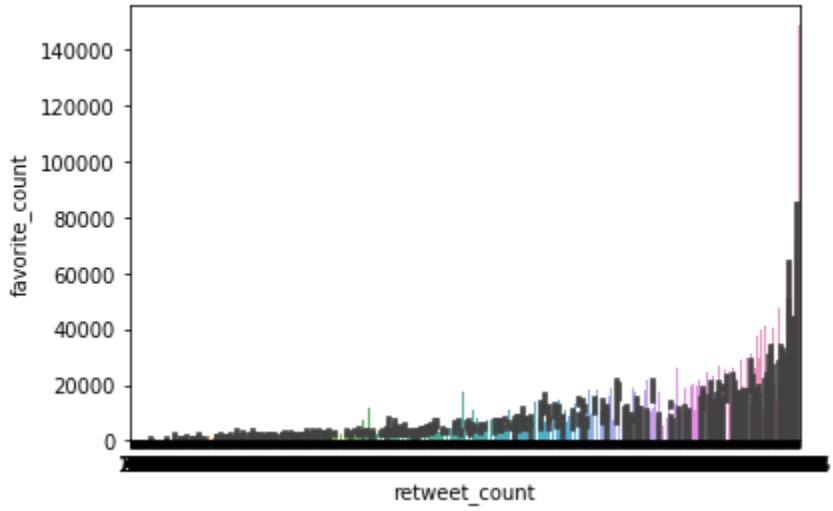
1. The relationship between retweet_count (how many times a tweet is retweeted) and favorite_count (how many times a tweet is favorited).

I organized the retweet_count column into Low (bottom 25% number of retweets [aka: retweet_count]), Medium-Low (25%-50% retweet_count), Medium-High (50%-75%), and High (75%-100%).

The graph below shows that there’s a pretty even linear relationship between the retweet_levels and favorite_count (how many times a post was favorited) for the first three levels, but then the High Retweet level shows an explosion of favorites.

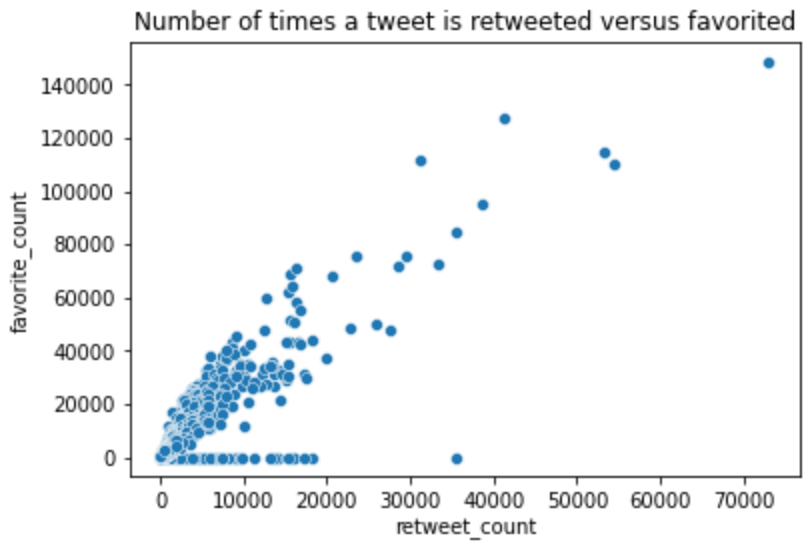


I then plotted retweet_count (without breaking that column down into levels) versus favorite_count.



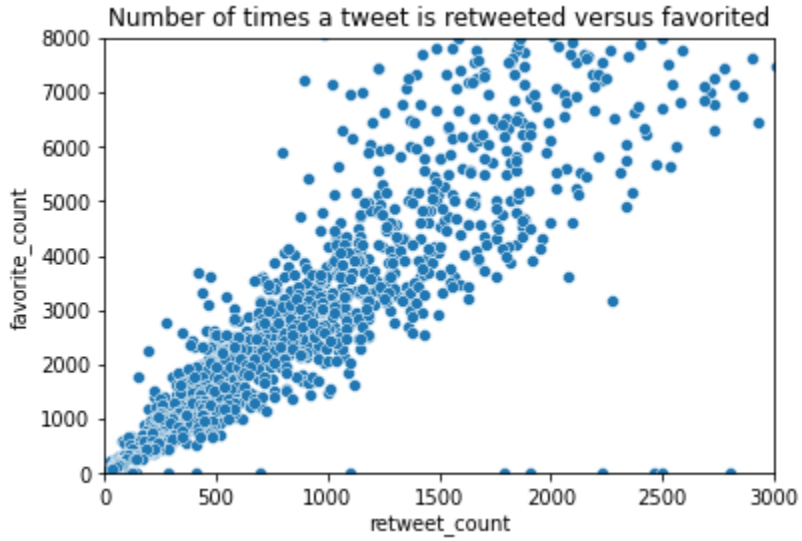
Again we see a pretty even linear relationship until we get to the upper end of retweet_count, at which point the favorite_count explodes.

Here we use a scatter plot, the more standard plot for showing correlation.



Again fairly linear, although there does seem to be an increasing variation in favorite_count as retweet_count increases. And notice that there is a line of values at the bottom which seems to indicate that there's a not insignificant number of tweets that received no favorites, but were retweeted quite a bit.

Here we zoom in on the area with the most activity.



The mean number of retweets per post is 2467.
The mean number of favorites per post is 7759.

I decided to adopt retweet_count over favorite_count as the best metric for measuring a tweet's popularity.

My arguments for this conclusion are:

1. The average number of favorites per post is a little more than three times as many as the average number of retweets per post. The more a post is retweeted, the more people see it and the more chance it has of being favorited.

   Therefore, the explosion in favorites at the high end of the retweet_count levels can probably be at least partly explained by the fact that many more people are seeing the most retweeted posts. (Recall that at the lower levels of retweet_count, the relationship between retweet_count and favorite_count was a linear one with a fairly consistent slope, indicating that at those levels, the two metrics are pretty much interchangeable.) It seems plausible that the most retweeted posts get disproportionately more favorites due to the great number of people seeing the tweets, and this skews the favorite_count data unfairly.

2. It is less of a commitment to favorite a tweet than to retweet it. Favoriting it just means that you admit you like it. Retweeting it means that you are willing to go out on a limb and tell your friends that this tweet is worth looking at.

There is one obvious counterargument to this conclusion:

People who retweet are already more committed to the entire enterprise of publicly declaring their love for cute tweets about dogs. Perhaps this makes retweet_count a worse indicator of the public's mood than favorite_count.

However:
The general public, with only a casual lust for and desultory participation in this kind of excessive devotion to the trivialization/kitschification of the canine, is probably greatly influenced by the tweets the more committed are forwarding them ("what's this? Oh, cool: Favorite. Moving on …"). I therefore maintain that the mechanism of the more devoted retweeting and the less devoted favoriting seems a reasonable explanation for the way the favorite_count begins to build past the initial linear relationship between the two metrics in the 50%-75% range and then explodes beyond that initial linear relationship at the 75%-100% range.

To be safe, I kept the the favorite_count information in the tables I created. Perusing these tables, I noticed no terribly anomalous behavior (ie: nothing like x breed of dog gets way more retweets than y breed, and yet y breed gets way more favorites than x breed). But there were definitely cases where dog breed x ranked higher in average tweets but lower in average favorites than dog breed y. It is perhaps therefore best to conclude that accurately describing the relationship between retweet_count, favorite_count, and individual and group psychology is beyond what my time, knowledge, and interest currently allow for. I'll accept using retweet_count for measuring popularity as a current best-guess, but will not cling to the metric in a dogmatic, dogged, or otherwise dirty-dog fashion.

2. The relationship between whether a post mentions a “doggo” or a puppy (we combined the “puppo” and “pupper” columns into one “puppy” column, since these two terms seem to be used interchangeably to describe puppies) and how many retweets the post receives.

I had assumed that—dogs being merely cute but puppies adorable—the tweets in which a puppy was mentioned would receive more retweets than the tweets mentioning mere dogs.

In this I was mistaken.

Voici the mean number of retweets and favorites per post for the various dog ages (let’s ignore the relatively rare floofer and doggo, pupper categories):

dog_stage_2	retweet_count	favorite_count
doggo	6737.800000	17208.892308
doggo, pupper	7464.200000	14358.200000
floofer	4100.857143	11510.428571
puppy	2455.722467	7613.453744

However, let’s look at how many total tweets each dog stage received:

puppy	227
doggo	65
floofer	7
doggo, pupper	5

There are four times as many tweets in which a moderator mentions a puppy (aka: a “puppo” or a “pupper”) than the number of tweets in which a moderator mentions a dog (aka: a “doggo”)*.

This implies that the moderators of We Rate Dogs are four times as likely to speak of puppies as of dogs. Perhaps they only speak of dogs when the tweet is particularly charming, whereas their threshold for speaking of puppies is somewhat (four times?) lower. If so, it could be that the tweets in which dogs are mentioned are on the whole more eye-catching and retweet-winning than the ones in which puppies are mentioned.

No one would suggest that people prefer dogs to puppies (the moderators’ predilection for puppy-centered posts argues well for the common wisdom: puppies are king). However, perhaps winning a retweet is more about a tweet’s overall charm than whether or not a puppy or a dog is figured most prominently in the tweet.

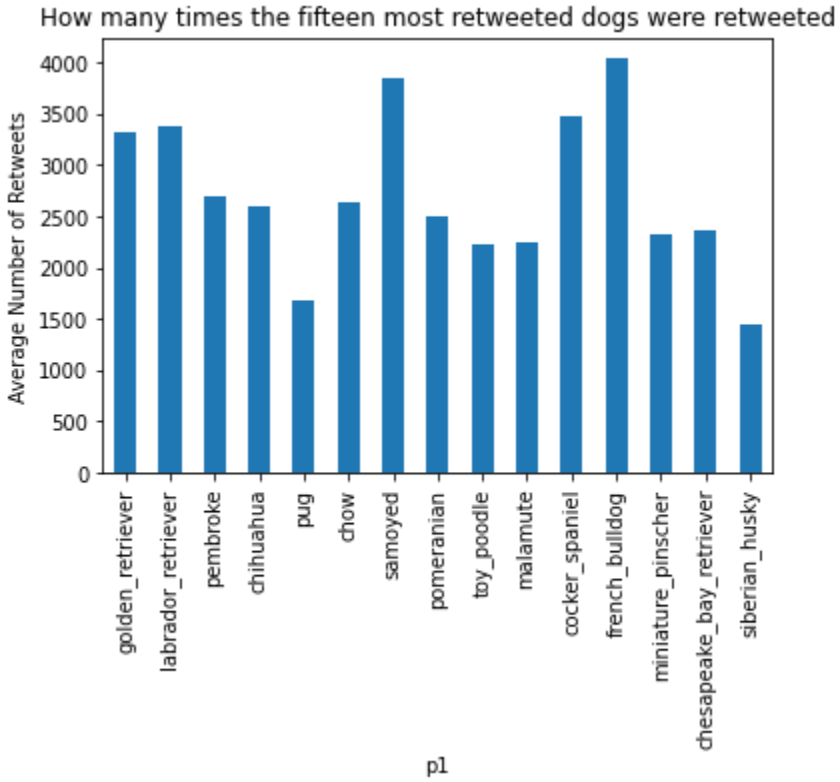
While finding a workable system for quantifying the human experience of charm would be a worthy endeavor, it lies outside the scope of this exercise. So we will simply conclude this section with a glance at the total number of retweets and favorites for each dog stage:

Dog_stage_2	retweet_count	favorite_count
Doggo	437957	1118578
doggo, pupper	37321	71791
Floofer	28706	80573
Puppy	557449	1728254

As you'd guess from the means and total tweet counts, puppy has more total retweets and favorites than doggo, but only a little more.

3. **Which dog breeds receive the most retweets. For this one, we used the automated image recognition software's best-guess about the breeds of the dogs featured in the tweets.**

Here is a graph comparing the retweet averages for the posts about the fifteen most retweeted dogs.



As you can see, the dogs with the highest average retweet counts were, from highest to lowest:

French Bulldog, Samoyed, Cocker Spaniel, Labrador Retriever, and Golden Retriever.

Here's a table of those five breeds that shows total count, average number of retweets, and average number of favorites.

Dog Breed	Retweet Count	Favorite Count	Total
golden_retriever	3316.916667	10307.638889	144
labrador_retriever	3379.053763	9872.301075	93
samoyed	3858.550000	10342.675000	40
cocker_spaniel	3473.428571	9353.785714	28
french_bulldog	4045.000000	16751.320000	25

If we measure popularity in total tweets, Golden Retriever is far and away the favorite, with more than 1.5 times as many tweets as Labs, more than 3 times as many tweets as Samoyeds, more than 5 times as many tweets as Cocker Spaniels, and almost six times as many tweets as French Bulldogs.

Measured by average number of favorite counts, the order is French Bulldog, Samoyed, Golden Retriever (close to Samoyed), Lab Retriever, and Cocker Spaniel.

Measured by average number of retweets (as in the graph), the order is French Bulldog, Samoyed, Cocker Spaniel, Labrador Retriever, and Golden Retriever.

I would argue that Golden Retriever is the most popular dog in this dataframe. It is tweeted the most by far, and as you go up in total number of tweets, the likelihood of a few relatively-ignored tweets increases. It is true that the likelihood of a few explosively popular tweets also increases, but as the number of total tweets goes down, you have a better chance of a lucky combination (a few big hits and relatively few duds), which I think can explain French Bulldog's retweet count (at 25 total tweets, French Bulldog is ranked #11 out of 15 on this chart in total tweets). Furthermore, Golden Retriever is almost tied for the number two spot in average number of favorites.

The case of the Golden Retriever makes me think that we should perhaps reevaluate our metric for measuring tweet popularity. The best metric would seem to be one that somehow factored in total tweet number, retweet count, and favorite count. But how to organize and weight these various metrics? We will leave that exercise for another day, or perhaps for never ever. Most likely for never ever.

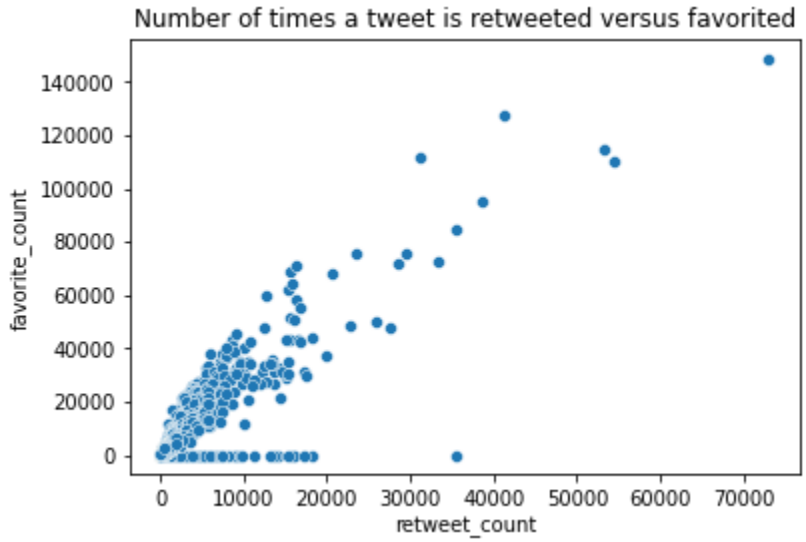
Postscript: A Final Note on retweet_count and favorite_count

I got a little carried away with the correlation between retweet_count and favorite_count.

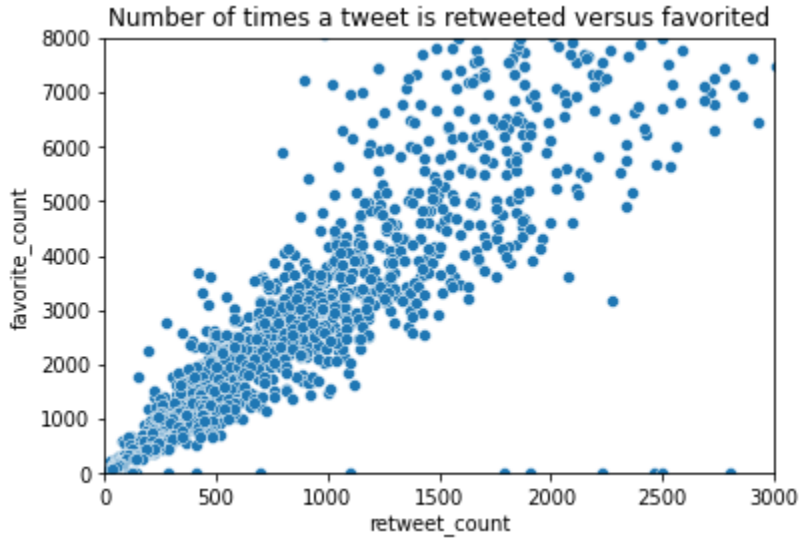
I did a Pearson correlation analysis between the two variables: about 86%

I thought that if I threw out the most-retweeted or the most-favorited tweets, I'd get an even higher correlation. Because if you zoom in on the graph, the points seem to bunch up closer together.

Graph showing full data set:



Zooming in:

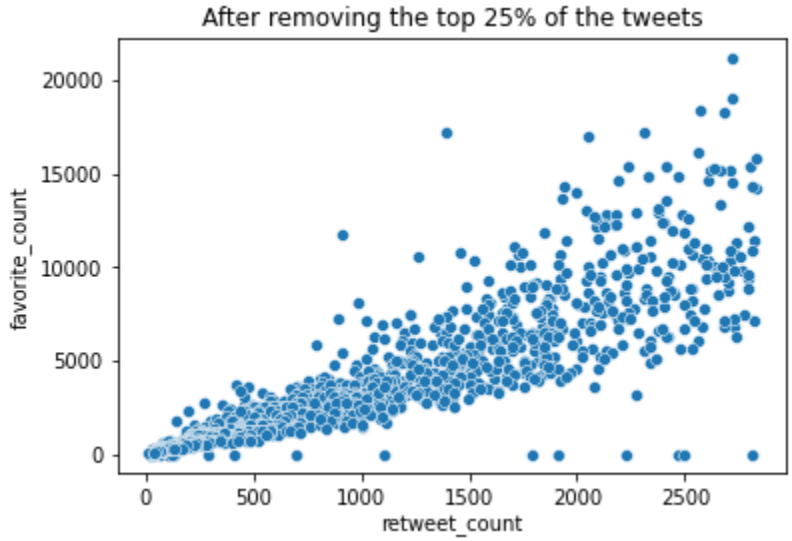


Removing the top 25% of retweeted tweets did result in a slightly better Pearson correlation: about 88% and 90%, respectively. From this, I was able to conclude that those 70 zero-favorited tweets, combined with the much

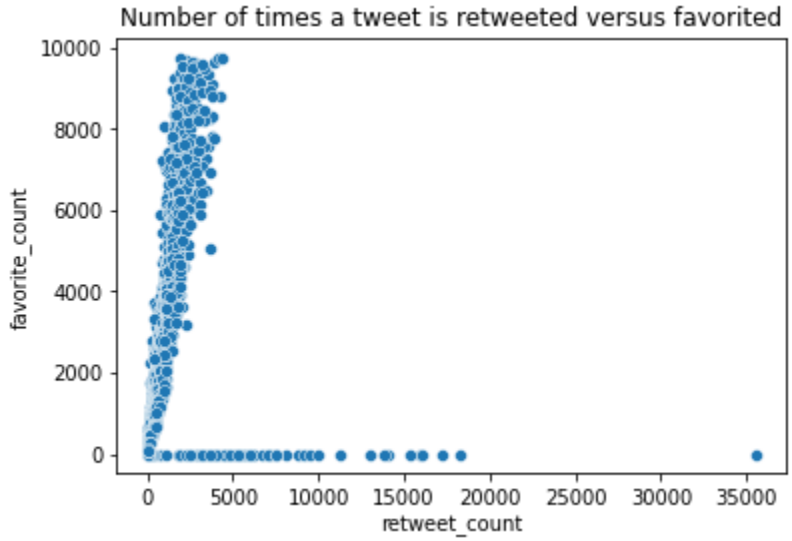
However, removing the top 25% most favorited retweets destroyed the Pearson brought the Pearson correlation down to about 25%

I couldn't fathom why this should be. I also didn't understand why removing the top 25% of either variable greatly changed the graphs. I thought the effect would be akin to zooming in, but it wasn't.

After removing the top 25% of retweet_counts the graph got a more horizontal:



After removing the top 25% of favorite_counts the graph got much more vertical:



I don’t know why this happened.

I decided to remove the 70 tweets with zero favorites to see how that impacted the Pearson correlations.

The whole data set got the highest Pearson score yet: about .93
Dropping the top 25% of retweets or favorites, resulted in scores that were high, but not as high as the full data set (about .88 and .90, respectively).

I concluded that the 70 zero-favorites, combined with the very steep slope of the graph after we removed the 25% most favorited tweets, was enough to give us that terrible .25 Pearson score.

Perhaps retweet_count with the 70 zero-favorited tweets removed from the dataset would’ve been the best popularity metric I could’ve used for this project (barring complicated combinations retweet_count, favorite_count and total counts [of things like tweets highlighting a specific dog breed]). But this project was supposed to be completed long ago, so I will not redo the analysis; I will just note my findings here in the Postscript.

I would welcome an explanation for the changed slope of the graphs.

*Is it lost on the wider world that these labels are all in the diminutive/cutesy form, and are thus all rather demeaning, rather insistent upon the point that dog is not, under any circumstance, a being which one should take seriously, should consider truly sentient, truly souled--at least not in an adult/moral-agent kind of way?