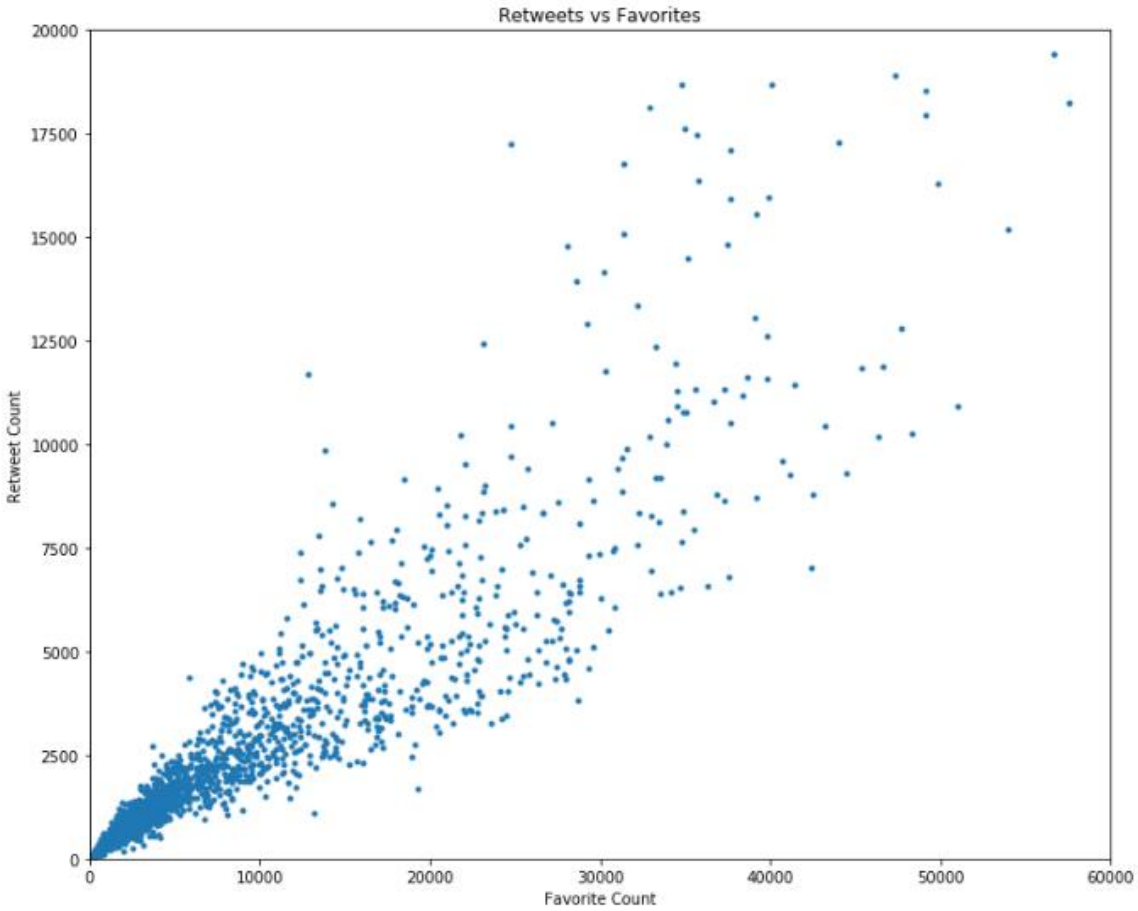
Analysis of WeRateDogs

Jin Lim, Udacity Data Analyst Nanodegree

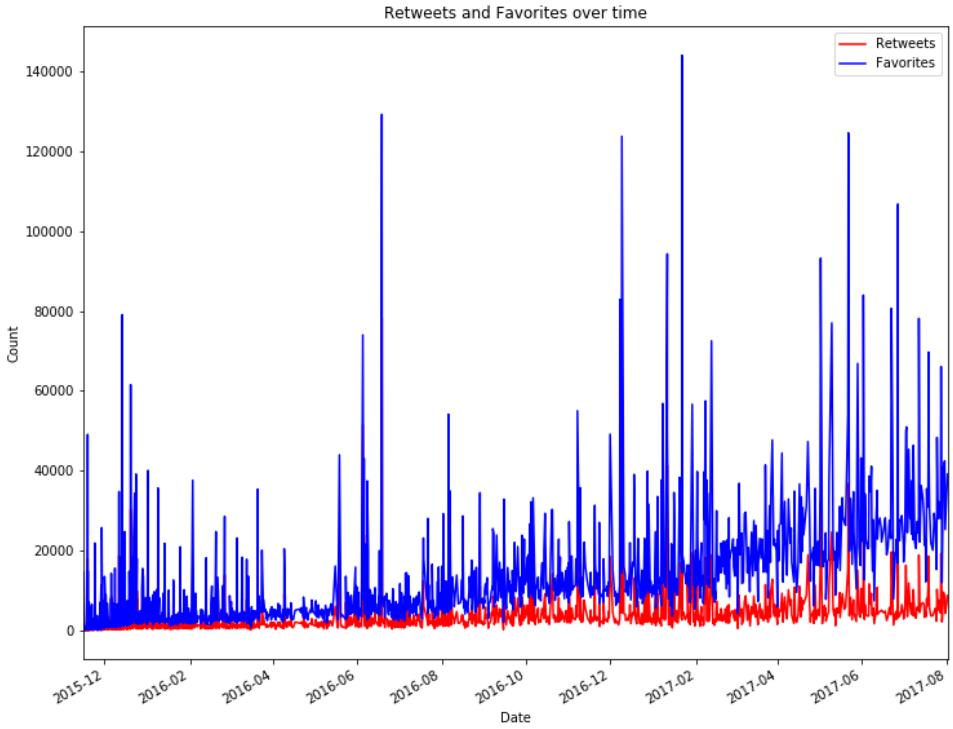
First, I returned some basic summary statistics.

	rating_numerator	rating_denominator	favorite_count	retweet_count	img_num	p1_conf	p2_conf	p3_conf
count	1994.000000	1994.000000	1994.000000	1994.000000	1994.000000	1994.000000	1.994000e+03	1.994000e+03
mean	12.237101	10.532096	8911.892678	2749.108325	1.203109	0.593941	1.344195e-01	6.024848e-02
std	41.471197	7.320710	12634.336616	4739.609935	0.560777	0.271954	1.006807e-01	5.089067e-02
min	0.000000	2.000000	80.000000	14.000000	1.000000	0.044333	1.011300e-08	1.740170e-10
25%	10.000000	10.000000	1949.250000	615.500000	1.000000	0.362857	5.393988e-02	1.619283e-02
50%	11.000000	10.000000	4078.500000	1334.000000	1.000000	0.587635	1.174550e-01	4.950530e-02
75%	12.000000	10.000000	11224.500000	3160.500000	1.000000	0.846285	1.951377e-01	9.159438e-02
max	1776.000000	170.000000	144132.000000	78215.000000	4.000000	1.000000	4.880140e-01	2.734190e-01

From this, I could see that the favorite counts and retweet counts have heavy skews, with much higher means than their medians (I did not find the numerator ratings significant as there were many “joke” ratings, such as 1776 for a dog in American regalia). I deduced that there was some kind of “snowball effect” where tweets surpassing a certain threshold of favorites and retweets “go viral”, such that the tweet itself is passed on to others on the buoying effect of its own popularity. Similarly, I guessed that favorites and retweets have their own self-reinforcing effect.



And yes, the two are highly correlated with a coefficient of around 0.91. Highly favorited tweets are more likely to be retweeted, and/or vice-versa.



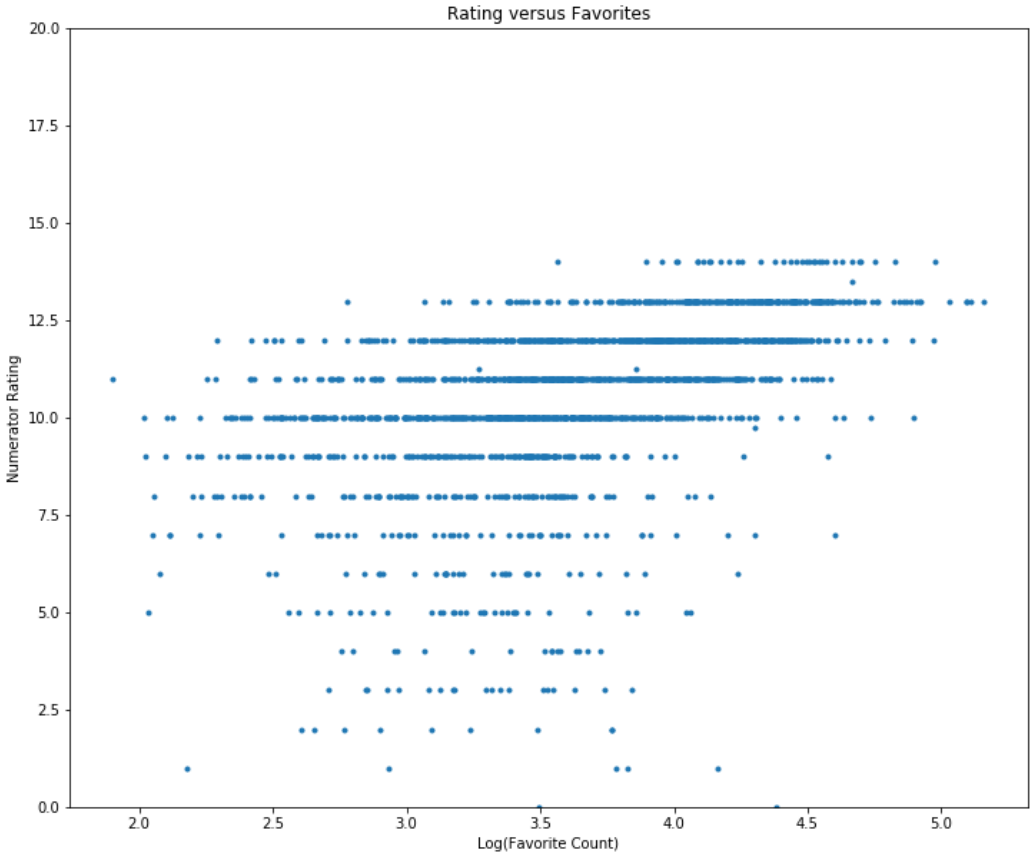
And here is a graph of WeRateDog's tweets' overall favorites and retweets over time. There is an upward trend, indicating that the account has only become more popular over time. Retweets are also used a lot less frequently than favorites.

Here is a summary of ratings grouped over dog types:

dog_type	count	mean	std	min	25%	50%	75%	max
doggo	72.0	14.305556	22.729666	5.0	11.0	12.0	13.0	204.0
floofer	3.0	12.666667	0.577350	12.0	12.5	13.0	13.0	13.0
pupper	223.0	12.929417	15.051251	3.0	10.0	11.0	12.0	143.0
puppo	28.0	12.142857	1.208436	9.0	12.0	12.5	13.0	14.0

Not every dog has a dog_type (in fact, precious few do), but for the ones that do, it would seem “doggos” are rated the highest by a significant margin. Even when assessing the data visually, this is significant because the next eight or so highest ratings (that have a dog type) are all “puppers” (despite doggo having the highest max). This must mean that doggos are consistently rated higher.

And finally, a plot relating ratings to number of favorites:



It’s not as clear a trend as the others, but as a general rule, the higher a dog’s rating, the more favorites it receives.