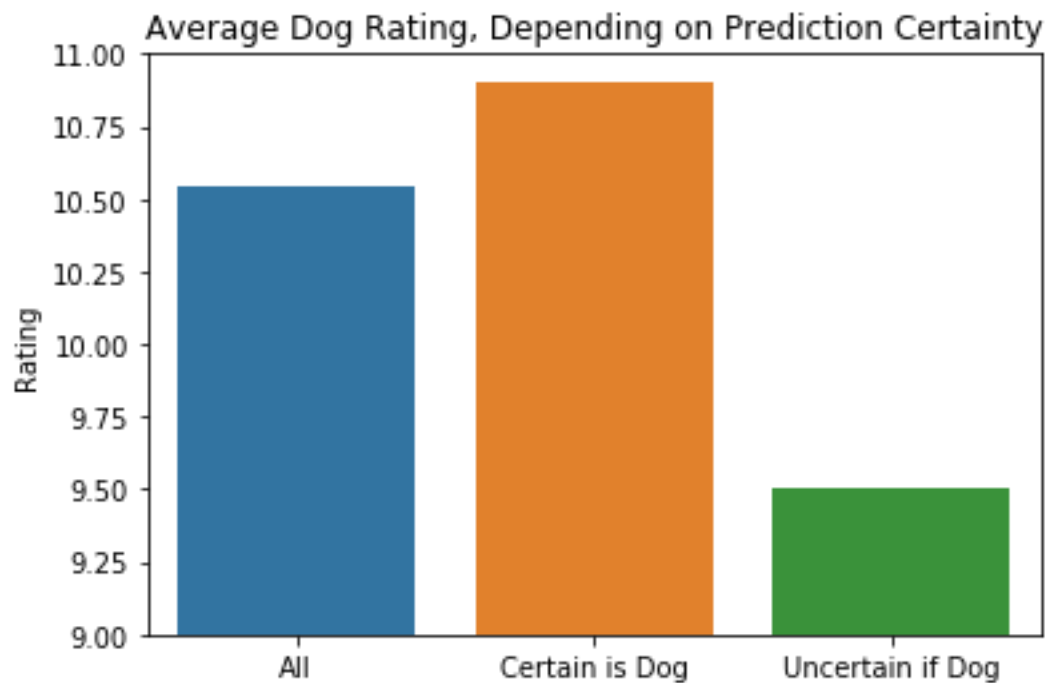


I was recently tasked with the project of extracting data from [WeRateDogs](#) on Twitter. After going through a data wrangling process, I was able to analyze that data to find some interesting insights.

First off, not all posts are of actual dogs, or easily identified as such. So, I was curious if the average rating given would be different for those posts where a dog was identified by a machine learning program. The following figure shows these results; the average overall rating given was around 10.5/10 while the average rating of posts where a dog was identified was 10.9/10 and where a dog was not identified was 9.5/10. So, if a machine thinks the post is a dog, that may be correlated with a higher rating!

Figure 1:



I was also curious as to the effect of the rating given on the number of Retweets and times the post was Favorited. I ran some regressions and found the following results:

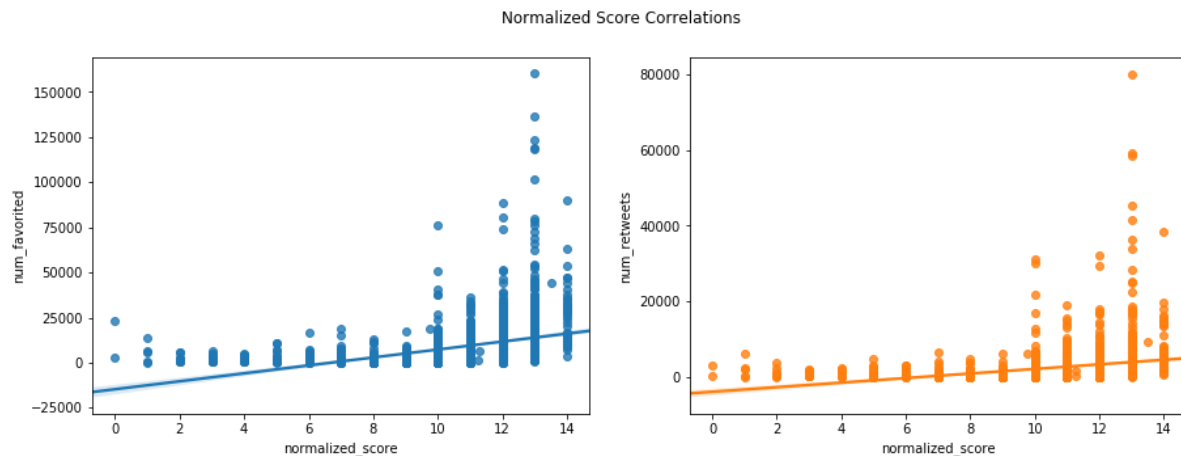
Figure 2: The number of times a post is favorited:

	coef	std err	t	P> t	[0.025	0.975]
intercept	-1.472e+04	1260.857	-11.677	0.000	-1.72e+04	-1.23e+04
normalized_score	2199.6783	117.102	18.784	0.000	1970.022	2429.335

Figure 3: The number of times a post was retweeted:

	coef	std err	t	P> t	[0.025	0.975]
intercept	-3832.8137	474.492	-8.078	0.000	-4763.369	-2902.258
normalized_score	602.9354	44.068	13.682	0.000	516.510	689.361

Figure 4:



So, according to the models, each 1 increase in rating out of 10 is correlated with being favorited 2200 more times and retweeted 603 more times.

The final insight I looked at may be a more general question about twitter as a whole, what's the relationship between retweets and times a post is favorited? For the WeRateDogs data, we can see how these variables are related based on this figure.

The theory that the relationship is linear isn't bad, but there appears to be a slightly exponential relationship where retweets and favorites feed off of each-other in such a way that the relationship is no longer linear.

Figure 5:

