ACT Report: WeRateDogs Twitter

An analysis of the @dog_rates twitter from 2017

Finding patterns in retweet, favorite, and ratings

Abstract

Three data sources were provided for analysis, however each had to be accessed in a different way, image_predictions.tsv was obtained via requests, twitter-archive-enhanced was provided directly and imported via quick upload, and the twitter API JSON file were read in and supported. The code for the twitter API was provided as a comment and not an active section in the Jupyter Notebook. Several questions were raised about relationships between different data points in two of the three datasets, while third, image_predictions, was produced by a neural network piece of software, and not directly related to analysis of the other two. Without meaningful connection between that dataset and some other, it would be difficult if not impossible to make any educated analysis. After the analysis we discovered that retweet and favorites were related with significant strength, R-squared ~.64, p_v << 0.00, favorites were impacted by the dog having a name, p_val ~ .016, and favorites were linked to retweets, p_val << 0.00. In addition two 'types' of dog, 'doggo' and 'puppo' were also shown to have an improvement in retweet and favorites, though the effect is small in both cases.

Method

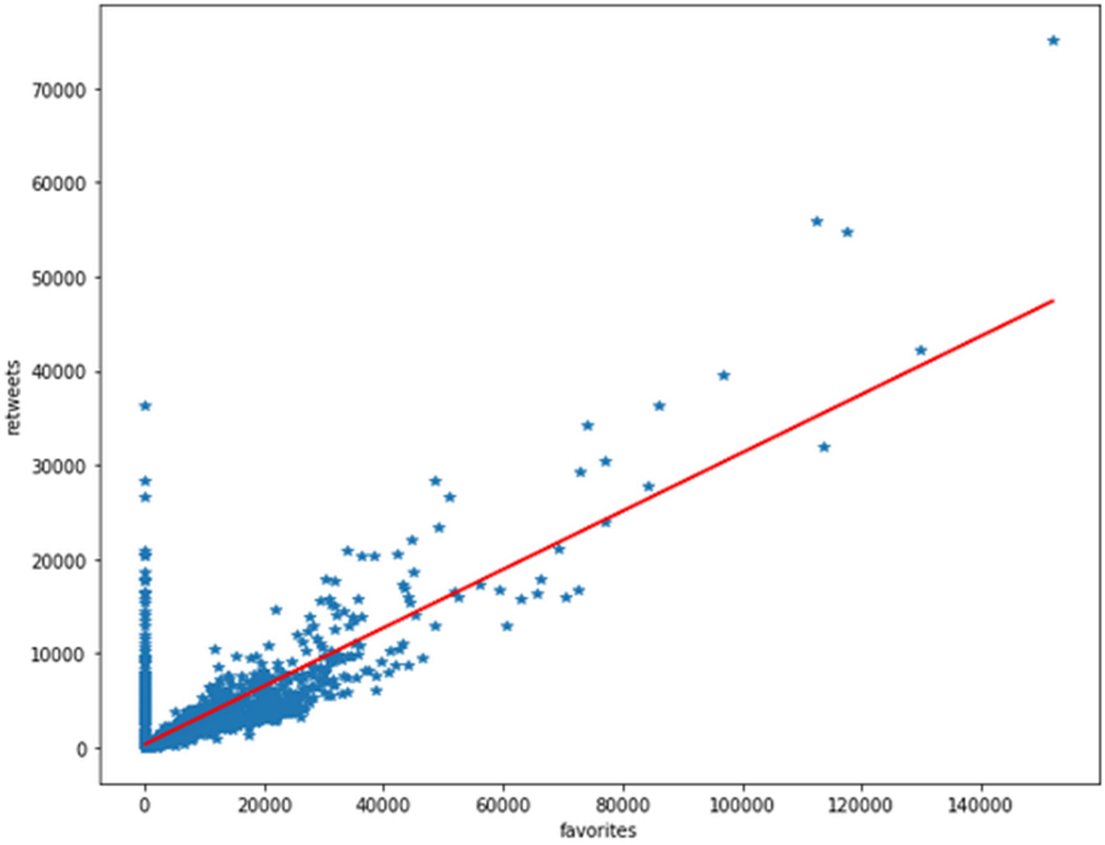
For each set of observations and variables a graph was set up and visually analyzed to see if there was any evidence of an impact. If there was sufficient visual evidence, or in cases where no graph could provide such a visual, hypothesis testing was done. When a hypothesis test was run, the null and alternative are provided as the first statement in the section, with detailed analysis following. Please note, the questions, and observations, are ordered here by level of impact, and not in the order that they were tested in the system. With the most impactful being first, and the least impactful being last. In addition this report contains only a brief description of questions that were asked but for which the answer was that the null hypothesis held. Each observation will include the name of the data frame, where applicable, and any graphs provided.

Analysis

Does retweet impact favorites?

Are favorites retweeted more than non-favorited first tweets, is our alternative hypothesis for this scenario.

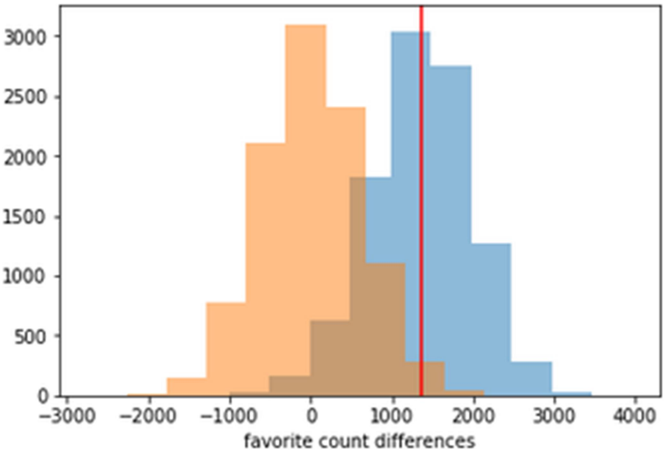
$$N_0 \geq N_1$$



In order to compare retweet and favorite impact on one another, I first created dfa9 from df3_c. I ran a linear regression on the data, which gave an R-Squared of .642, meaning approximately 64% of change can be attributed to one or the other factor. The above linear graph shows a clear moderate relationship. The red line being the actual trend. The line of points rising up from 0,0 represents a number of outlier data points in the dataset. The calculated p_value is 0.000 meaning a very high probability of them being related.

Does having a name improve favorite?

This is a similar comparison to the above for retweets where our null can be stated as saying that having no name gets at least as many, if not more favorites than a dog with a name: $N_0 \geq N_1$

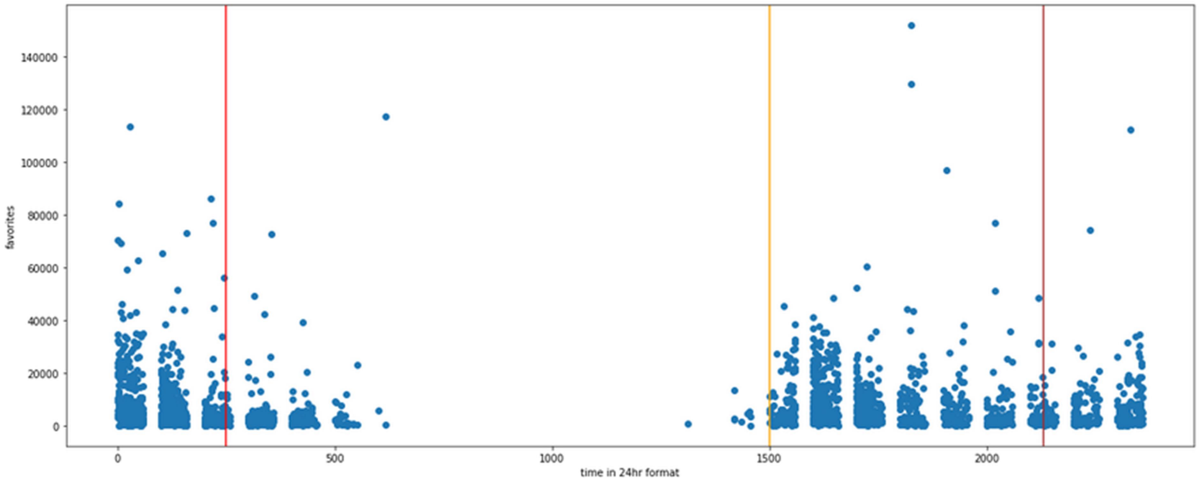


I ran 10000 random selections from the dataset provided and put into analysis data frame dfa6, based on the null assumption, shown in orange in the above histogram comparison. The blue data set is from our actual data with the red line representing the mean of the actual data. Based on this we get a p_value of .016 for there being an actual influence on obtaining a favorite if the dog has a name, vs. if the dog has no name. This is sufficient to reject our null hypothesis in favor of the alternative.

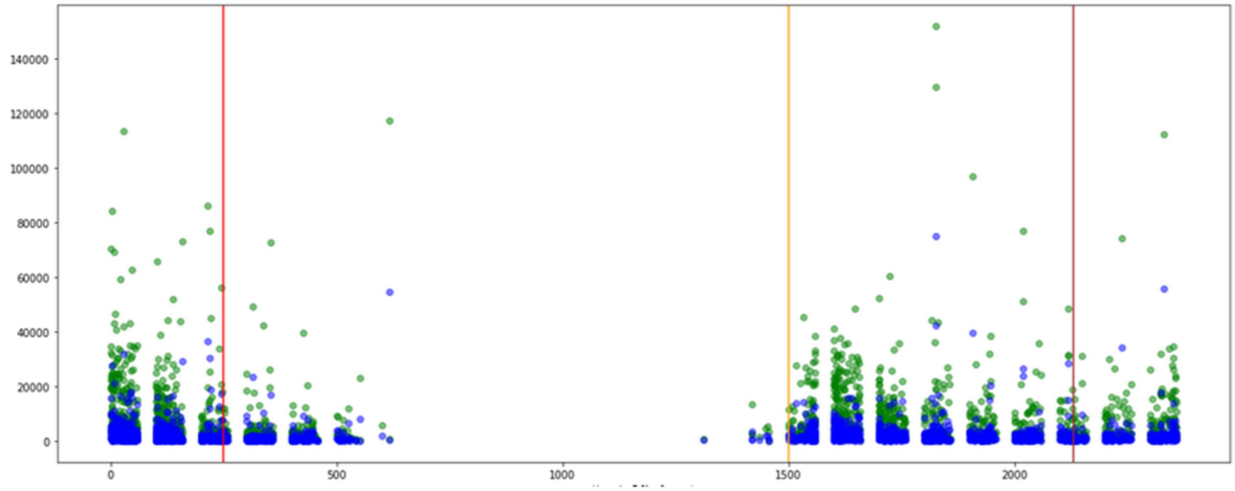
Does tweet time improve favorites?

As with retweets, we will compare favorites to timestamp

$$N_0 = N_1$$



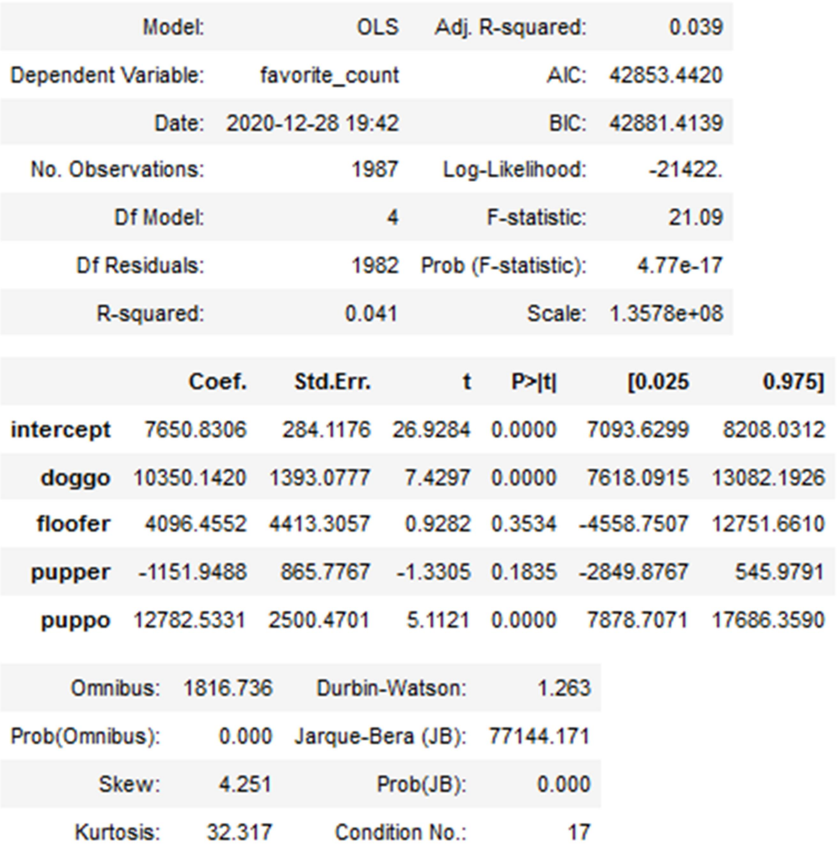
For this question I created dfa5 from df3_c_f and df1_c. While there is a large amount of noise, a pattern that fits a bimodal distribution can be seen, with the first one ending its decline at the redline, 0230, a new spike forming at the yellow line at 1500, descending until about 2100, and a new increase at the brown line, 2100, this spike then increases until peak around 0000, and then descends to the red line at 0230. Time does seem to have some minor impact on if a tweet will become a favorite. This is inconsistent with the pattern in retweeting we saw in the retweet vs time graph, which showed no significant such pattern, in the below graph blue is retweets and green is favorites.



Does type of dog influence favorites?

$$N_0 = N_1$$

The null can be best stated as dog type ‘none’ is equal to any other dog type.



Model:	OLS	Adj. R-squared:	0.039
Dependent Variable:	favorite_count	AIC:	42853.4420
Date:	2020-12-28 19:42	BIC:	42881.4139
No. Observations:	1987	Log-Likelihood:	-21422.
Df Model:	4	F-statistic:	21.09
Df Residuals:	1982	Prob (F-statistic):	4.77e-17
R-squared:	0.041	Scale:	1.3578e+08

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
intercept	7650.8306	284.1176	26.9284	0.0000	7093.6299	8208.0312
doggo	10350.1420	1393.0777	7.4297	0.0000	7618.0915	13082.1926
floofer	4096.4552	4413.3057	0.9282	0.3534	-4558.7507	12751.6610
pupper	-1151.9488	865.7767	-1.3305	0.1835	-2849.8767	545.9791
puppo	12782.5331	2500.4701	5.1121	0.0000	7878.7071	17686.3590

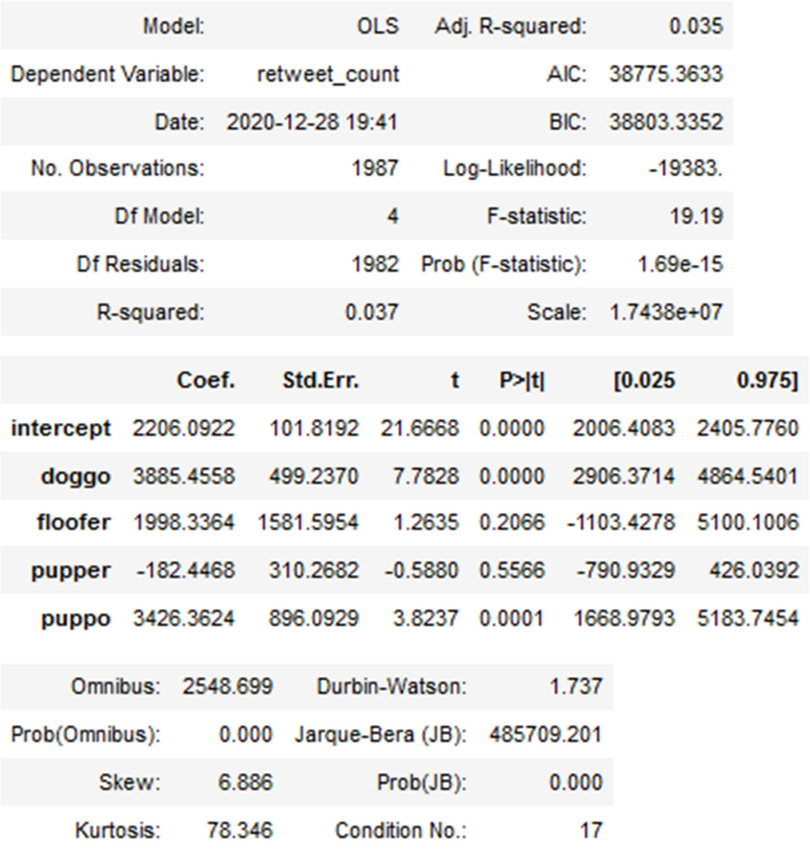
Omnibus:	1816.736	Durbin-Watson:	1.263
Prob(Omnibus):	0.000	Jarque-Bera (JB):	77144.171
Skew:	4.251	Prob(JB):	0.000
Kurtosis:	32.317	Condition No.:	17

OLS measurement on this data, placed in dfa7, shows an approximately 4% contribution to retweet counts for dogs being labelled as either ‘doggo’ or ‘puppo’ as compared to ‘None’, with high significance, but not for the other two categories of dog ‘type’.

Does type of dog influence retweet?

The null in this case is that the type of dog, using our predictive data, had no impact on retweets

$$N_0 = N_1$$



Model:	OLS	Adj. R-squared:	0.035
Dependent Variable:	retweet_count	AIC:	38775.3633
Date:	2020-12-28 19:41	BIC:	38803.3352
No. Observations:	1987	Log-Likelihood:	-19383.
Df Model:	4	F-statistic:	19.19
Df Residuals:	1982	Prob (F-statistic):	1.69e-15
R-squared:	0.037	Scale:	1.7438e+07

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
intercept	2206.0922	101.8192	21.6668	0.0000	2006.4083	2405.7760
doggo	3885.4558	499.2370	7.7828	0.0000	2906.3714	4864.5401
floofer	1998.3364	1581.5954	1.2635	0.2066	-1103.4278	5100.1006
pupper	-182.4468	310.2682	-0.5880	0.5566	-790.9329	426.0392
puppo	3426.3624	896.0929	3.8237	0.0001	1668.9793	5183.7454

Omnibus:	2548.699	Durbin-Watson:	1.737
Prob(Omnibus):	0.000	Jarque-Bera (JB):	485709.201
Skew:	6.886	Prob(JB):	0.000
Kurtosis:	78.346	Condition No.:	17

In this case a linear regression was run on data placed into dfa3 for analysis, and a low p-value for two of our values was determined, but the fit wasn't very good $R^2 = 0.035$. This means our model doesn't fit very well but, we had two categories, 'doggo' and 'puppo' that did show statistical significance as compared to 'None'. With p_values of 0 and 0.0001 respectively, this indicates some probable correlation, though 'floofer' and 'pupper' did not have the same level of apparent connection. We might be able to reject the null for 'doggo' and 'puppo' cases, but not for 'floofer' and 'pupper' cases.

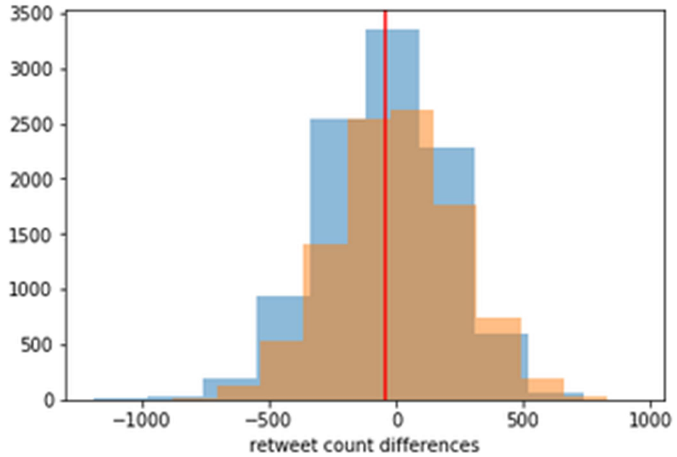
Does having a name improve retweet?

The null in this case is that not having a name gives at least as many retweets as having a name, while our alternative is That dogs with names get more retweets than dogs without names.

$$N_0 \ \geq N_1$$

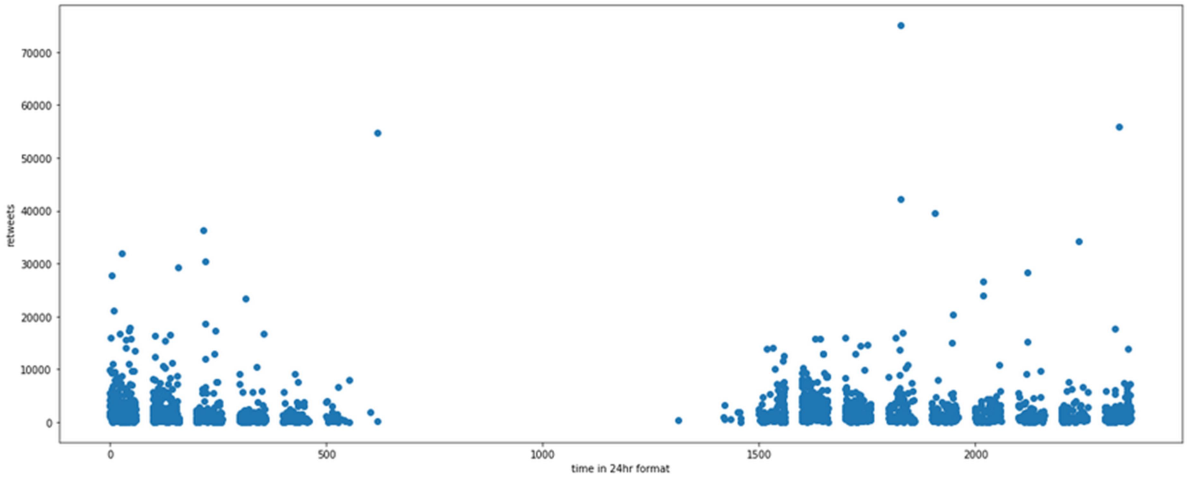
After creating an initial dfa2 for analysis, I had to return to the wrangling part of the analysis. I noticed that there were dogs named ‘a’ , ‘an’ …etc. in the list, and I needed to verify that these dogs have names or don’t and properly handle the columns. While dogs do get, from some owners, rather absurd names, such as ‘Hey You’… a simple conjunction is unlikely.

I started by creating dfa2, consisting of tweet_id and name from df1_c and retweet_count from dfa3_c. then altering the names column from strings to 0/1 int. This allowed me to run a more meaningful analysis on the subject. I then calculated the difference in means for retweet counts for these two categories. These had a mean difference of 431 retweets. I then generated a series of 10000 choices from the dfa2 dataframe in order to generate a significant sample of test values, and generated the standard deviations for named vs. unnamed dogs. In addition I generated a sample of the differences and the standard deviation for the differences. It is quite clear upon looking at the histograms that there is no effective difference between dogs with names and dogs without names. The calculated p value being .57. This is surprising due to the clear relationship between favorite and retweet, and favorite and name.



Does tweet time improve retweet?

The data from dfa1 is the scatterplot below and it shows there is really no effect of tweet time on the number of retweets provided. This is however perhaps a little surprising as well, as the retweets are very narrowly spread out over the day.. It is rather surprising that this has no noticeable impact as we saw earlier in the report time does impact favorites, and favorites and retweets are linked, but time and retweets are not.



Other questions with negative answers:

Does rating impact retweets or favorites?

Knowing that ratings are arbitrary doesn’t mean that there wasn’t some other links, this was examined and found not to be the case, that is we can not reject our null hypothesis that ratings are just that, arbitrary. These are found in dfa4 and dfa8

Insights

The single most powerful thing to take away from this is that things can be complexly related in a way that might be surprising, a number of things including retweets are directly linked to favorites with some statistical significance, but aside from favorites itself, almost none of the other variables influence retweet. It is likely that retweet influences favorites, that is a one way causation, but if that were true, we should expect to find other links back to retweet which we find missing. It is therefore my suggestion that the cause is reversed, that retweets are the result of people creating favorites, and they tend to favorite certain types and at certain times.

As mentioned in the abstract above there is significant influence on name and favorite, and there seems to be more favorites at certain times of the day. Then dog types do seem to have some impact on both favorites, ~4%, and retweets, ~3.5%.

External Resources:

https://stackoverflow.com/questions/55092403/how-to-extract-hourminute-from-a-datetime-stamp-in-python

https://www.researchgate.net/post/What_is_the_relationship_between_R-squared_and_p-value_in_a_regression