Analysis Report

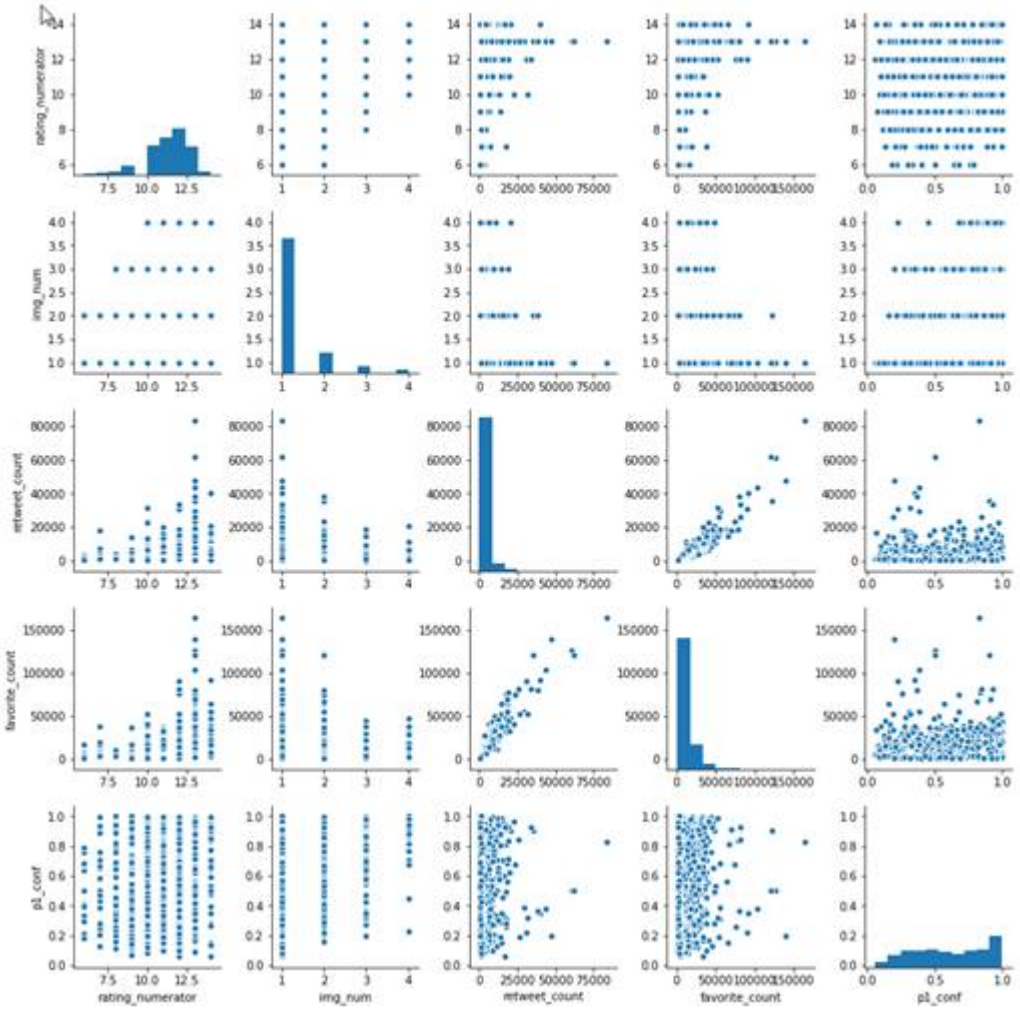
The goal of this article is to use Python and its libraries to get some insights from data provided by the twitter account "WeRate Dogs" and tweet API query. We analyzed 2000+ records from 2015/12 to 2017/8. This article starts with observations from selected variables followed by visualizations and concludes with a regression model to predict ratings for dogs.

1. Mistakes in rating – numerator and denominator

In the data provided, we have observed that there are some extremely high rating numerators, such as 1776, 420… To find out if those scores are valid, we investigated some of those suspicious tweets and observed that the numerator ratings were mistakenly extracted from the text content. Examples (images below): tweet on left says score is 9.75/10 while the input is 75 in the rating numerator column; tweet on right: "24/7" in the text is a digital abbreviation, referring to "24 hours and 7 days", which was extracted as 24 in the numerator and 7 in denominator. We believe such invalid data can be detected by comparing against a threshold or fixed by improving on the text-to-rating extraction algorithm.

2. Linear relationship between retweet counts and favorite counts

After examing various pairs of numerical variables, we have found that retweet counts and favorite counts exhibit strong relationship.



The linear trend is very obvious to us, which is quite intuitive in practice. Below are the summary statistics of a linear regression run by python, a R-square of 0.86 reassured our assumption.

OLS Regression Results

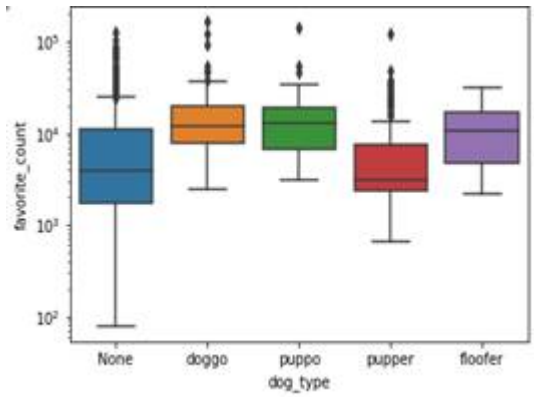
Dep. Variable:	favorite_count	**R-squared:**	0.864
Model:	OLS	**Adj. R-squared:**	0.864
Method:	Least Squares	**F-statistic:**	1.198e+04
Date:	Thu, 24 Jan 2019	**Prob (F-statistic):**	0.00
Time:	15:46:30	**Log-Likelihood:**	-18712.
No. Observations:	1892	**AIC:**	3.743e+04
Df Residuals:	1890	**BIC:**	3.744e+04
Df Model:	1		
Covariance Type:	nonrobust		

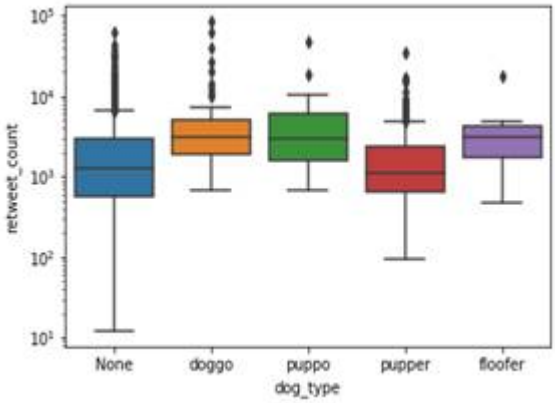
	coef	std err	t	P>|t|	[0.025	0.975]
intercept	2148.6411	125.991	17.054	0.000	1901.546	2395.736
retweet_count	2.5081	0.023	109.457	0.000	2.463	2.553

Omnibus:	492.757	**Durbin-Watson:**	0.755
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	14622.329
Skew:	0.578	**Prob(JB):**	0.00
Kurtosis:	16.570	**Cond. No.**	6.31e+03

3. Dog lingo impacts to popularity of the tweet

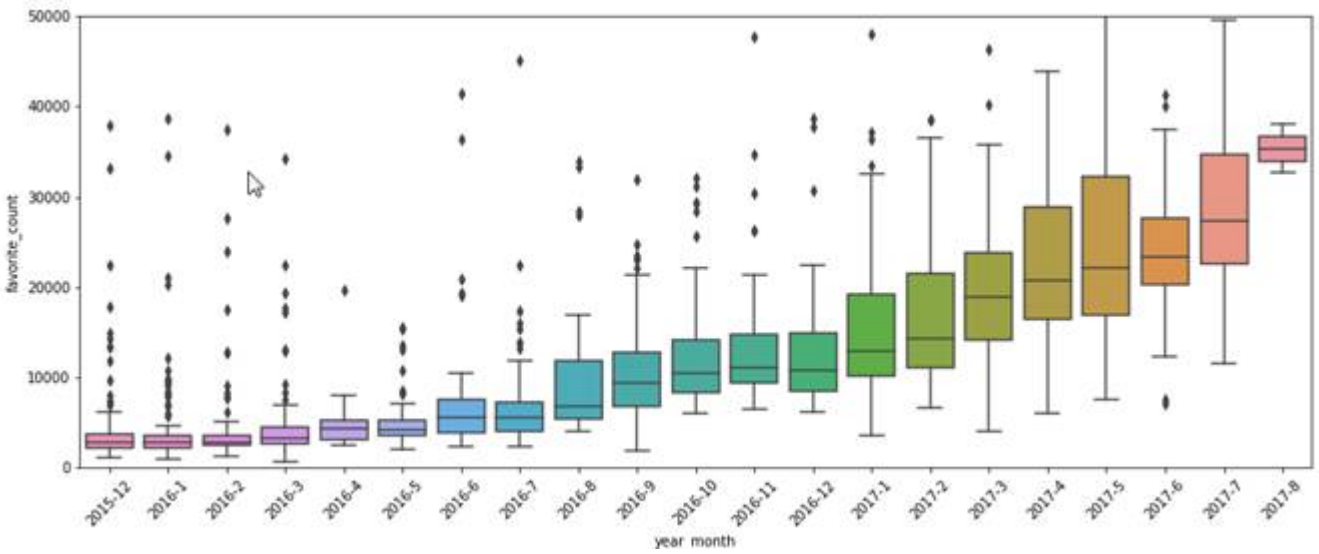
Dog lingo is a language trend that's been gaining steam on the Internet in the past few years. In our case, there are doggo, puppo, pupper and floofer and non-classified. They refer to dogs with different sizes or physical attributes.
We wonder wether the usage of dog lingo would make a difference in popularity by measuring favorite counts and retweet counts. From the chart below, we can roughly conclude that doggo and puppo receive higher favorite and retweet counts on median than other types. However, In the "favorite battle", puppo defeats doggo narrowly by median measurement. While in the "retweet battle", doggo takes the first place. Pupper bottoms in both battles. We should note that due to the different sample sizes, the result can be heavily biased. Especially when there are 200 counts for pupper, while only 7 counts for floofer in this comparison.



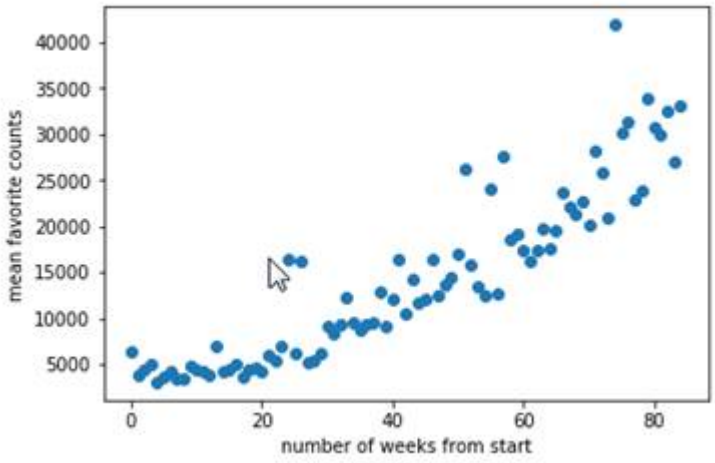


4. Time series analysis on "favorite counts" monthly median

Does WeRate Dogs get more and more web traffic over time? We can calculate the monthly median favorite counts to find out the answer. By plotting the candlesticks across 2015/11 till 2017/8, we can see that the median favorite counts climbed up steadily.

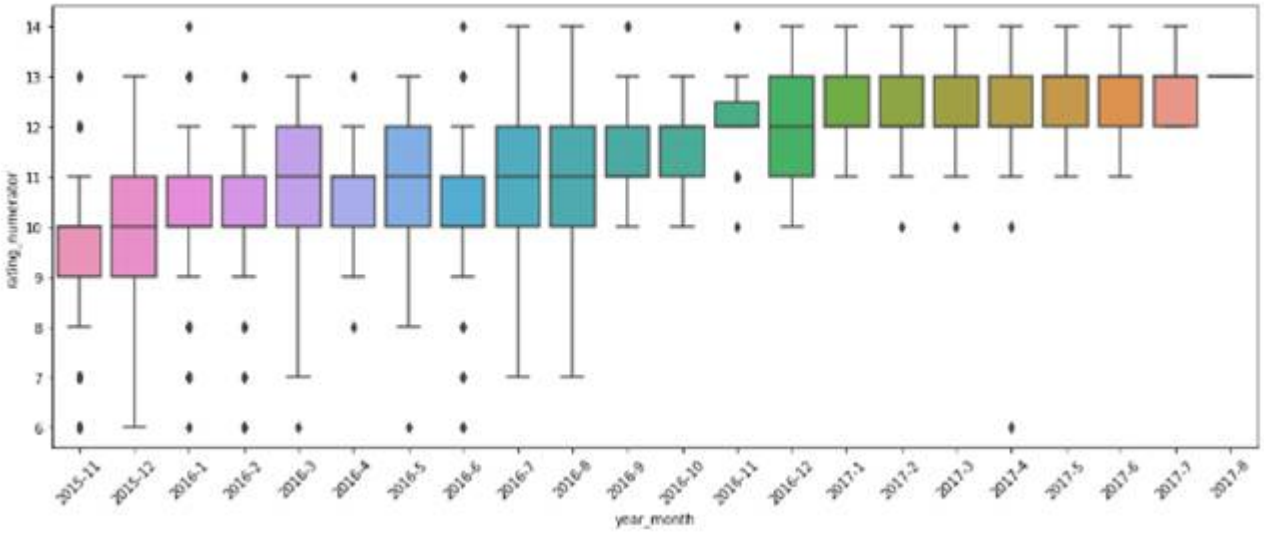


Another interesting observation is if we dive deep into the weekly mean favorite counts, a few outliers drive up the average numbers of some weeks. For example, on 2017/1 (week 55 in below chart), the mean favorite counts suddenly stand out. This change is mainly attributed to one tweet on Obama's dog. Celebrities' dogs always get a lot of attention. Even J.K. Rowling liked this post.



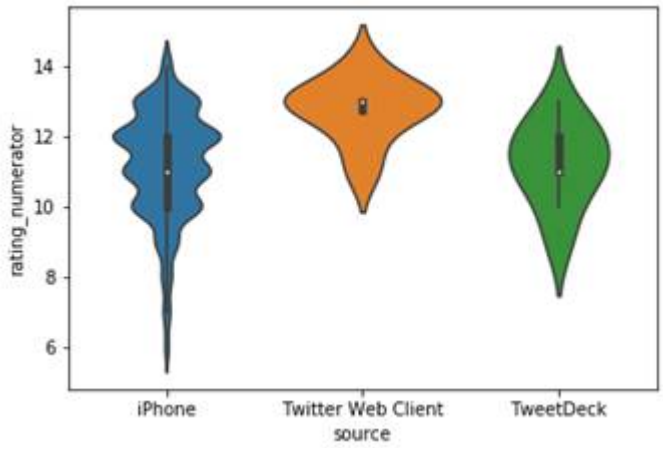
5. Time series on score (rating numerator)

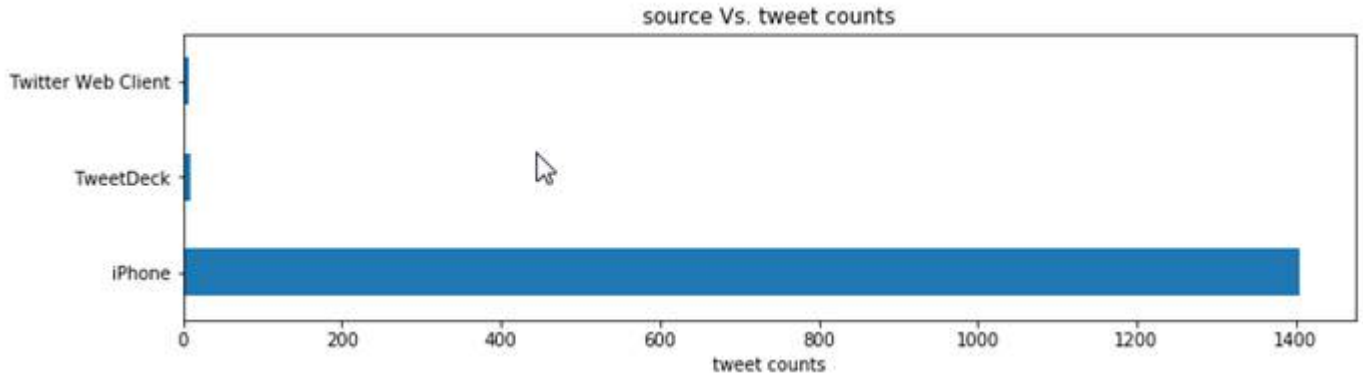
We have talked about the trend on popularity, did the rating numerator follow the same pattern? Below chart is the time series of the rating numerator, we have observed that both the score range and median score increased over time. However, the min-max range of the score seem be shrinking and there are no obvious ups and downs. This actually makes me wonder how the score is generated by WeRate Dogs.



6. Source and score

We examined the three sources of the tweet (iphone, website and twitter deck), tweets from Vine has been dropped due to that fact that it is disabled already. From the violin chart below, we have found that tweets sourced from website have a higher ceiling and bottom scores than other sources. Also the frequency for scores of 13,14 are more than other sources. My guess would be photos posted through web portal are enhanced and edited, thus they are likely to have higher scores. iPhone is the biggest source for all tweets and no source for Android, which means staff of WeRate Dogs probably never use android. Once again, due to the different sample sizes (see second chart below), the result can be biased.





7. Can we predict the score?

Finally come to the question, based on the variables we mentioned above, can we build a model to predict the score for dogs? Here are independent variables I picked :

- retweet counts
- favorite counts
- number of images posted
- source of the tweet,
- time to post

From the regression result listed below, the R-square is 0.36, which represents a weak correlation relationship. Besides, favorite counts and retweet counts are highly colinear. The p-value for favorite counts, source website and source iPhone are relatively high. Therefore, it is difficult to predict numerator with this linear regression model.

OLS Regression Results

Dep. Variable:	rating_numerator	**R-squared:**	0.359
Model:	OLS	**Adj. R-squared:**	0.356
Method:	Least Squares	**F-statistic:**	128.8
Date:	Sat, 26 Jan 2019	**Prob (F-statistic):**	1.51e-129
Time:	20:40:44	**Log-Likelihood:**	-2249.8
No. Observations:	1389	**AIC:**	4514.
Df Residuals:	1382	**BIC:**	4550.
Df Model:	6		
Covariance Type:	nonrobust		

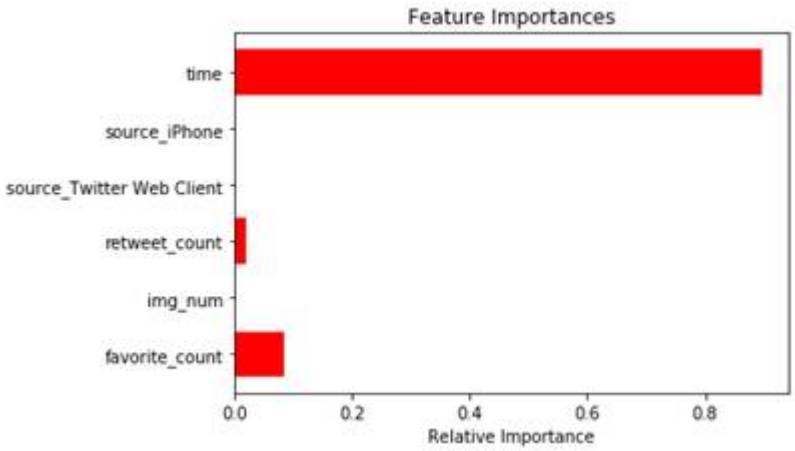
	coef	std err	t	P>|t|	[0.025	0.975]
intercept	8.6189	0.436	19.790	0.000	7.765	9.473
favorite_count	-1.925e-06	9.26e-06	-0.208	0.835	-2.01e-05	1.62e-05
retweet_count	4.501e-05	2.14e-05	2.099	0.036	2.94e-06	8.71e-05
img_num	0.2029	0.052	3.879	0.000	0.100	0.305
time	0.1368	0.009	15.340	0.000	0.119	0.154
source_Twitter Web Client	0.5819	0.597	0.975	0.330	-0.589	1.753
source_iPhone	-0.4993	0.411	-1.216	0.224	-1.305	0.306

Omnibus:	170.487	**Durbin-Watson:**	2.021
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	318.076
Skew:	-0.777	**Prob(JB):**	8.52e-70
Kurtosis:	4.756	**Cond. No.**	4.42e+05

We also try to use Random Forrest model to predict score using the same independent variables as the linear regression model does. The hyper parameters are manually picked by trial and error

- number of trees = 10
- max number of leafs in a tree = 4

After fitting the model on 1041 training record, we tested the model on 348 test records. The predict accuracy is as high as 90.74 %. The most important feature is time and favorite counts.



Both models suggest time is the most important factor in determining the rating as the scores show an increasing trend over time in my earlier observation.