Analyze WeRateDogs Data

WeRateDogs is a Twitter account that rates people’s dogs with humorous content about the dog. This analysis aims to gather insights and visualizations from the previously wrangled dataset.

Instead of using the master dataset which was cleaned, it will be assumed that the CSV file will be imported into the dataset and stored in new DataFrame called `df`. The usual steps will be followed like checking the number of columns and rows and checking the data types. The new table yielded a new irrelevant column which was subsequently deleted. The Tweet ID column was also converted from string to int.

There will be three insights and two visualizations to be communicated.

Insights

Q1: Which dog breeds are the most popular?

The five dog breeds which are featured in most of the tweets are: 1) Golden Retriever, 2) Labrador Retriever, 3) Pembroke, 4) Chihuahua, and 5) Pug. The neural network algorithm confidently predicts the breed in almost 60% of its images.

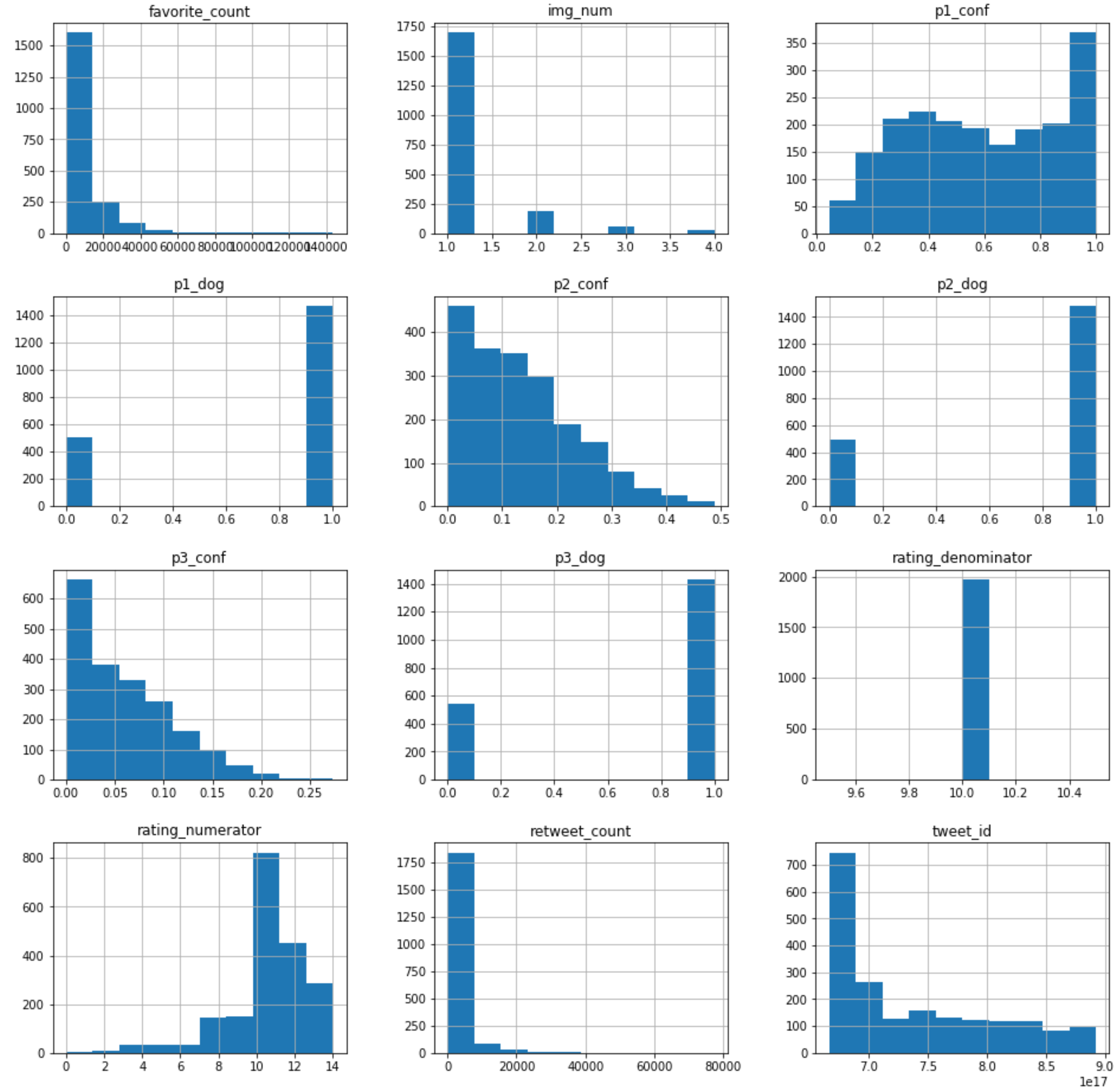
Q2: How does retweet count and favorites vary among the four different dog stages?

Doggos had the highest median retweets and Puppos had the highest median favorites. Puppers had the least retweets and favorites.

Q3: How does the rating affect the retweets and favorites?

It is no surprise that higher rated tweets should have the higher retweets and favorites. It is evidently clear that as the rating increases, the retweets and favorites also increase. However, correlation does not necessarily imply causation and retweet and favorite count can be caused by several factors as well.

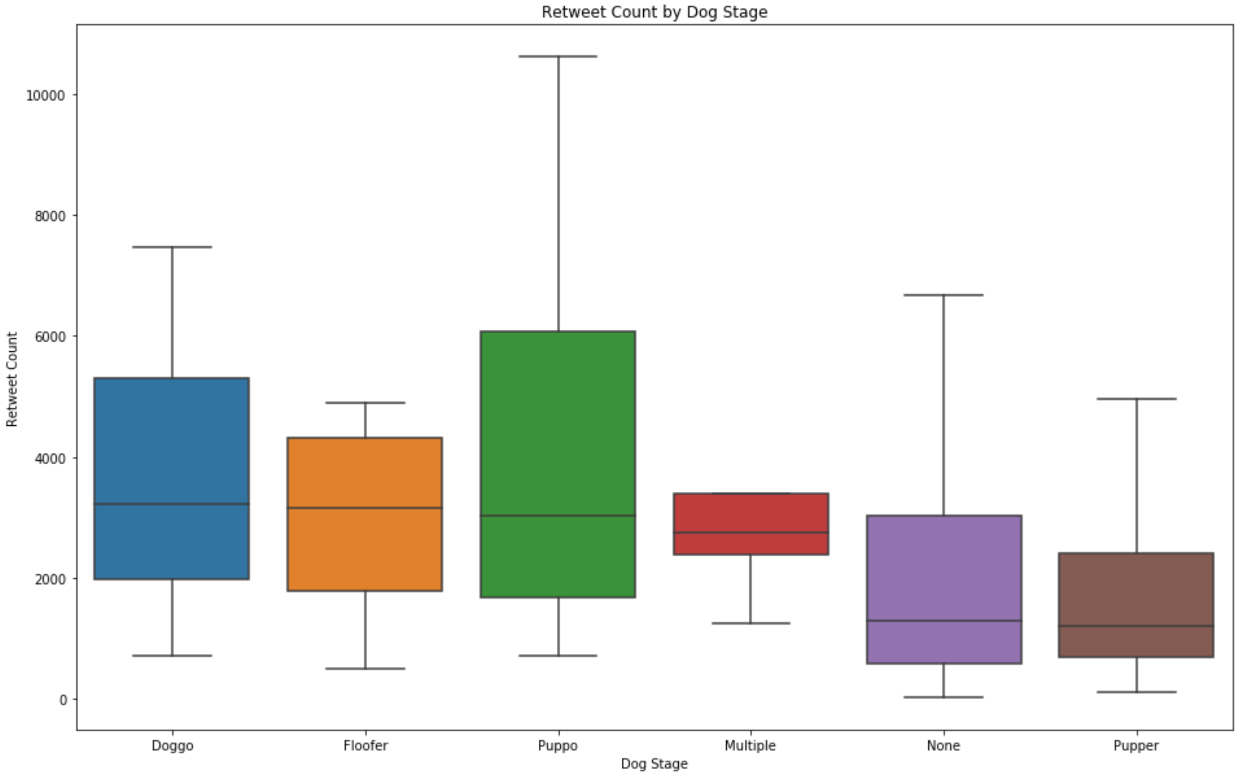
Visualizations

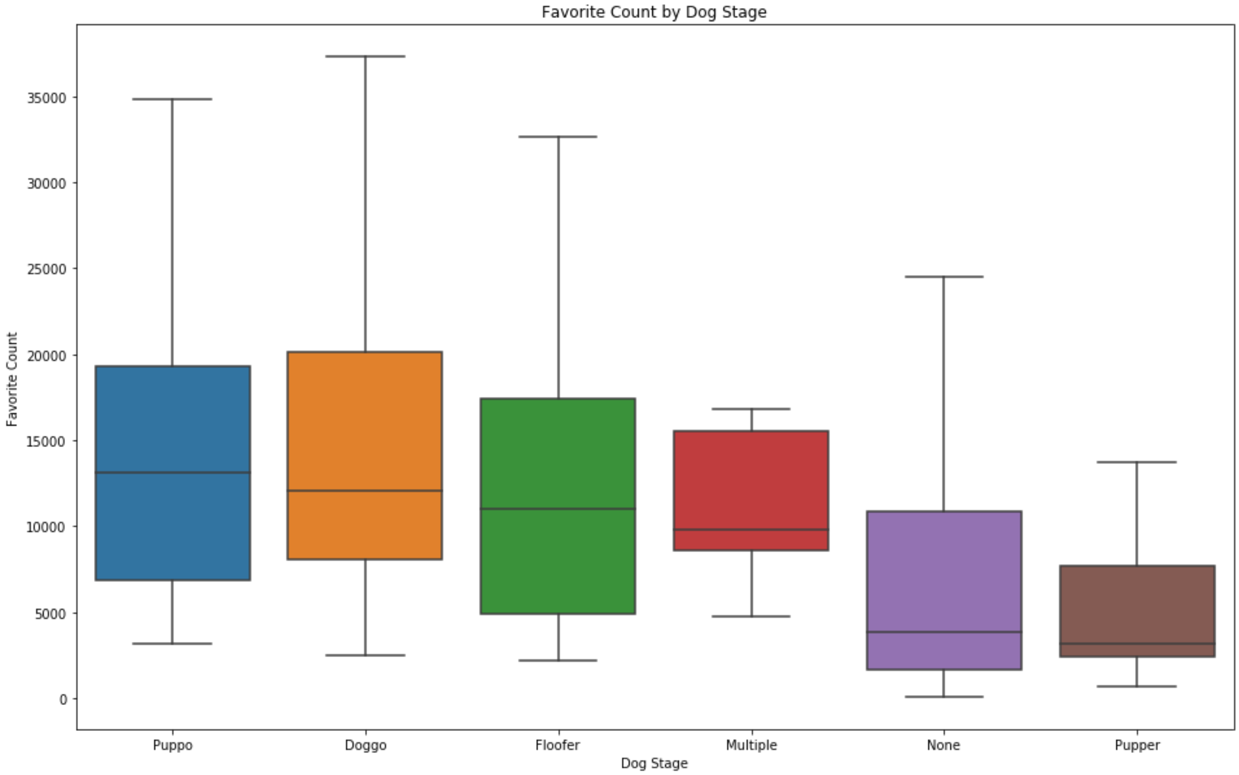


The histograms of interest are the following:

1. favorite_count: The plot is skewed to the right.
2. rating_numerator: The plot is skewed to the left.
3. retweet_count: The plot is skewed to the right.

P1: How does retweet count and favorites vary among the four different dog stages?

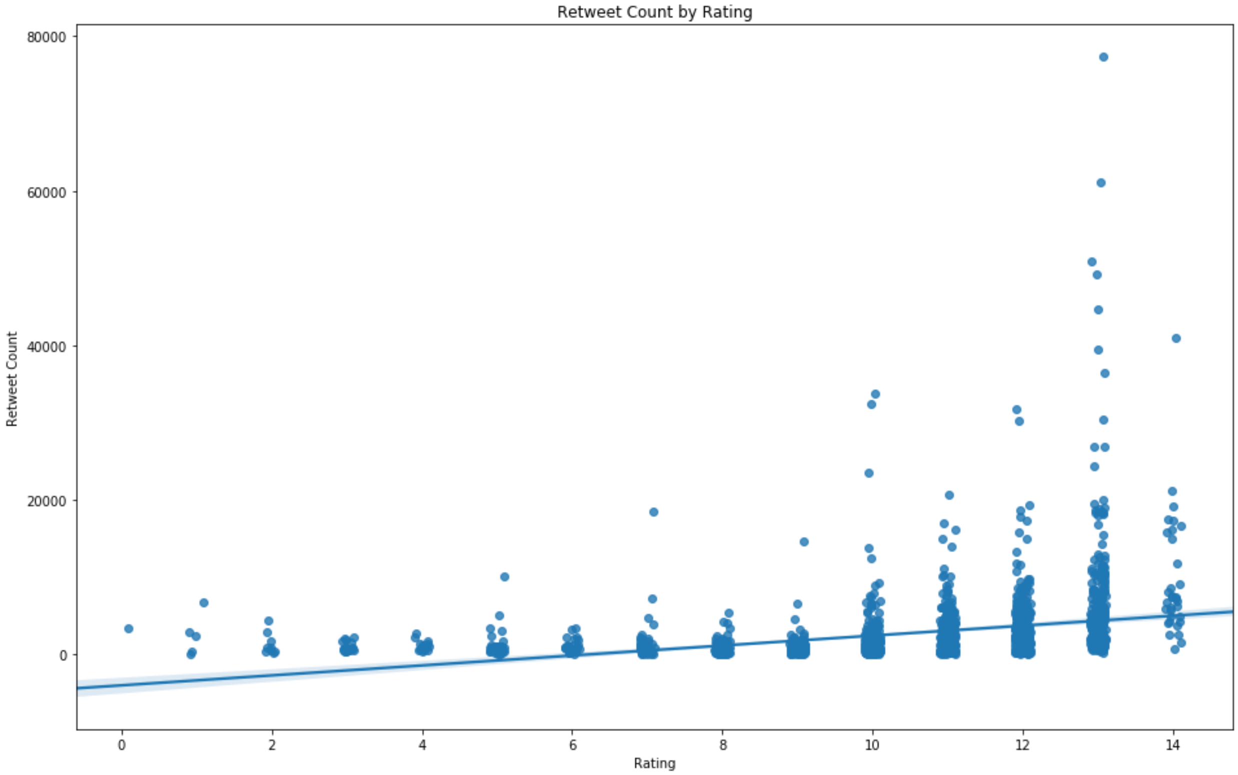


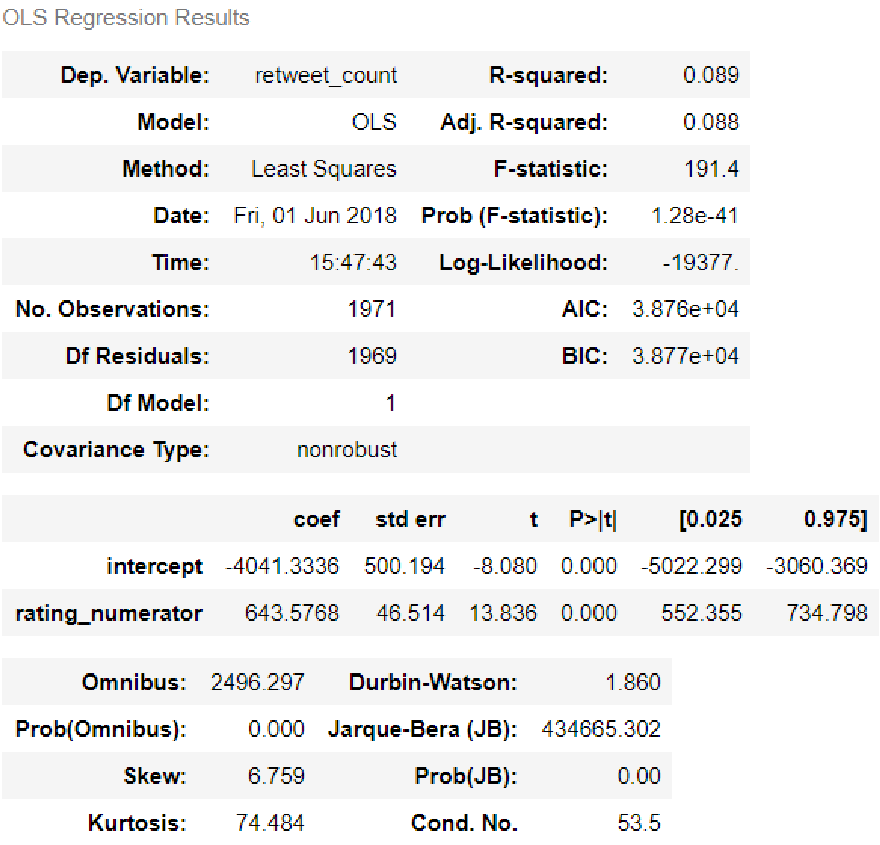


The boxplots show the median retweets and favorites decreasing across the dog stage identified earlier. Outliers are not shown to emphasize the 25% to 75% percentile ranges and the median values.

The median retweets for Doggo, Floofer, Puppo and Multiple are almost similar but their distributions are different. Puppos have very large retweets and very small retweets while Floofers and Multiple have smaller distributions. Similarly, for median favorites, Puppo, Doggo, Floofer and Multiple have almost similar values but this time, their distributions are more similar except for Multiple. The None dog stage have the same height for retweets and favorites considering their value counts.

P2: How does the rating affect the retweets and favorites?

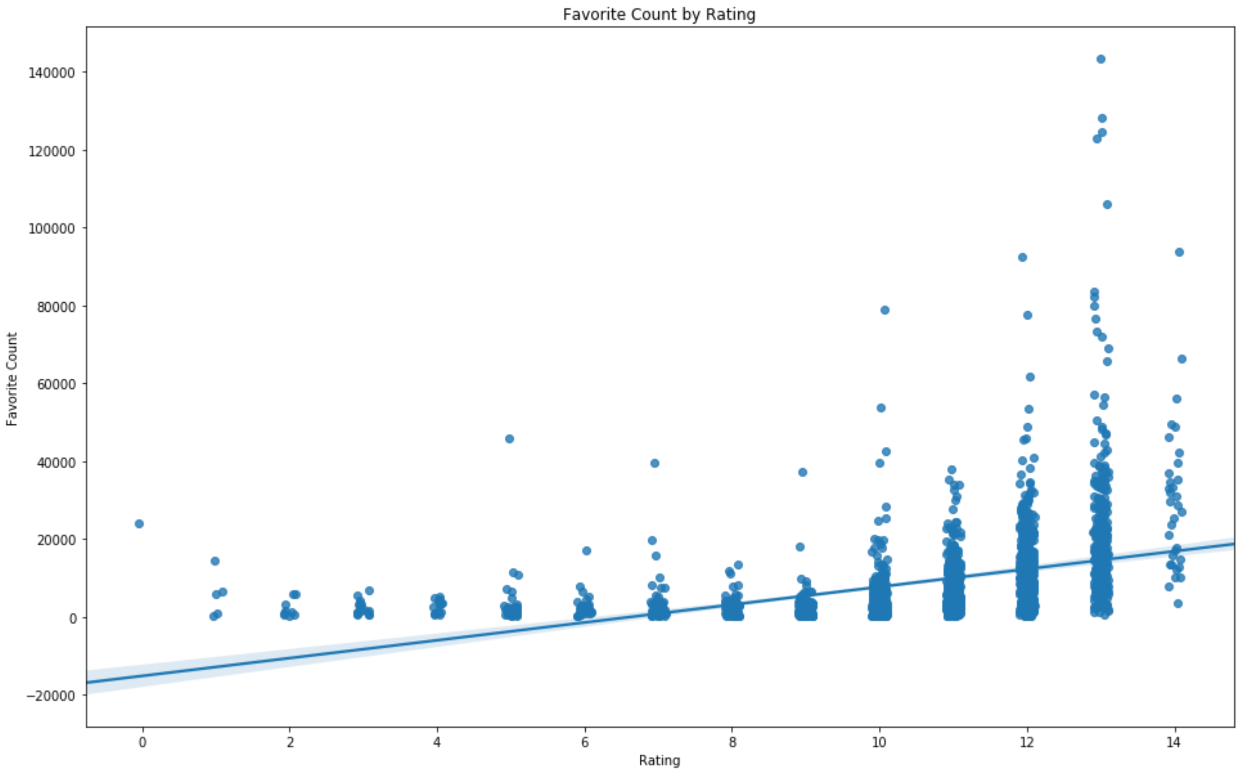


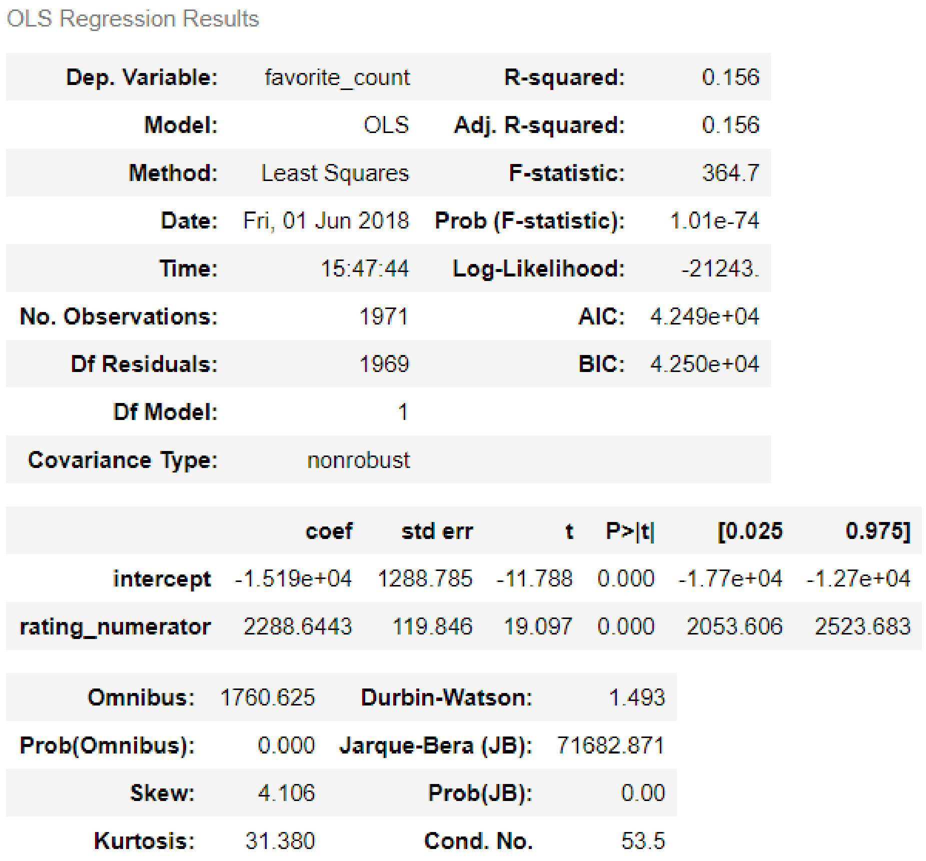

OLS Regression Results

Dep. Variable:	retweet_count	**R-squared:**	0.089
Model:	OLS	**Adj. R-squared:**	0.088
Method:	Least Squares	**F-statistic:**	191.4
Date:	Fri, 01 Jun 2018	**Prob (F-statistic):**	1.28e-41
Time:	15:47:43	**Log-Likelihood:**	-19377.
No. Observations:	1971	**AIC:**	3.876e+04
Df Residuals:	1969	**BIC:**	3.877e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
intercept	-4041.3336	500.194	-8.080	0.000	-5022.299	-3060.369
rating_numerator	643.5768	46.514	13.836	0.000	552.355	734.798

Omnibus:	2496.297	**Durbin-Watson:**	1.860
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	434665.302
Skew:	6.759	**Prob(JB):**	0.00
Kurtosis:	74.484	**Cond. No.**	53.5




OLS Regression Results

Dep. Variable:	favorite_count	**R-squared:**	0.156
Model:	OLS	**Adj. R-squared:**	0.156
Method:	Least Squares	**F-statistic:**	364.7
Date:	Fri, 01 Jun 2018	**Prob (F-statistic):**	1.01e-74
Time:	15:47:44	**Log-Likelihood:**	-21243.
No. Observations:	1971	**AIC:**	4.249e+04
Df Residuals:	1969	**BIC:**	4.250e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
intercept	-1.519e+04	1288.785	-11.788	0.000	-1.77e+04	-1.27e+04
rating_numerator	2288.6443	119.846	19.097	0.000	2053.606	2523.683

Omnibus:	1760.625	**Durbin-Watson:**	1.493
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	71682.871
Skew:	4.106	**Prob(JB):**	0.00
Kurtosis:	31.380	**Cond. No.**	53.5

The scatterplots show the earlier findings on how the rating affect the retweets and favorites. The plots show an increasing count when the rating is also increased. The low P-value suggests we have evidence that the ratings have a statistically significant linear relationship with either the retweets or favorites. However, their R^2 values indicate that ratings only account for 8.9% and 15.6% of the variance in retweets and favorites, respectively.