The Analyzing and Visualizing Process

Introduction

In this project I have done the analyzing and visualizing process in the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

Analyzing and Visualizing Process

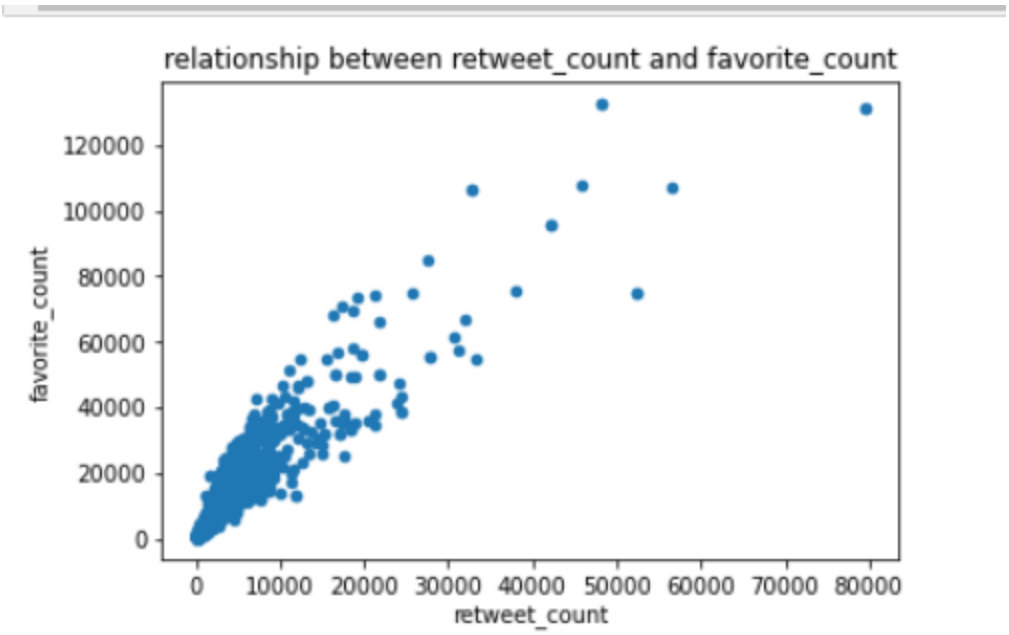
1. The relationship between retweet_count and favorite_count

Using the Regression Model and scatterplot I found these interesting conclusions about the relationship between retweet_count and favorite_count.x

Dep. Variable:	favorite_count	**R-squared:**	0.833
Model:	OLS	**Adj. R-squared:**	0.833
Method:	Least Squares	**F-statistic:**	1.203e+04
Date:	Thu, 27 May 2021	**Prob (F-statistic):**	0.00
Time:	08:14:52	**Log-Likelihood:**	-24057.
No. Observations:	2411	**AIC:**	4.812e+04
Df Residuals:	2409	**BIC:**	4.813e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
intercept	2485.5456	122.548	20.282	0.000	2245.235	2725.856
retweet_count	2.2721	0.021	109.696	0.000	2.231	2.313

Omnibus:	570.163	**Durbin-Watson:**	0.777
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	30731.724
Skew:	0.035	**Prob(JB):**	0.00
Kurtosis:	20.490	**Cond. No.**	6.83e+03



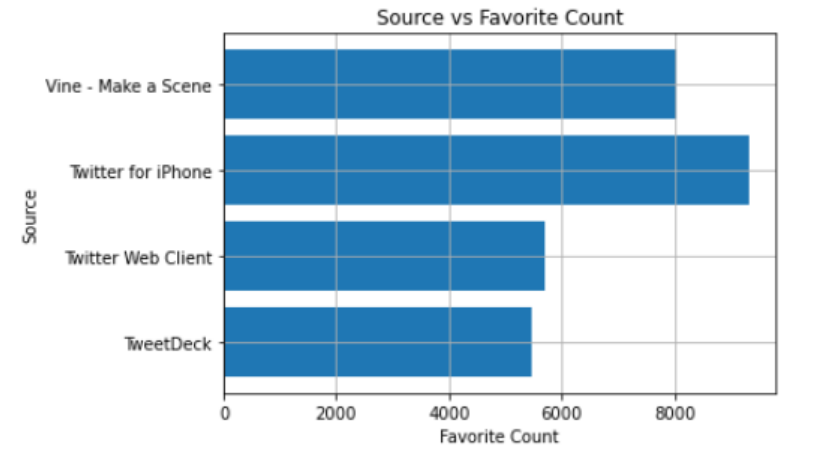
From the model summary and scatter plot we can conclude the following:

- The predicated favorite_count = 2485.5456 + 2.2721 * retweet_count -> if the retweet_count = 0, we predicate that the favorite_count will be around 2486, and for every one more retweet, we predicate that the favorite_count would be increase by 2.2721.
- The p-value for retweet_count = 0 -> retweet_count is statistically significant for predicting the favorite_count.
- The R-squared (= 0.833) is closer to 1, which mean the better our model fit, and it is suggest that there is a positive strong realtionship between favorite_count and retweet_count (as you can see in the scatterplot). Also, we can interpret the value of R-squard as that 83.3% of the variability in favorite_count is explained by the retweet_count.

2. The relationship between soure and favorite_count

	coef	std err	t	P>|t|	[0.025	0.975]
intercept	8034.0783	1190.238	6.750	0.000	5700.080	1.04e+04
iphone	1286.0773	1220.277	1.054	0.292	-1106.824	3678.979
web_client	-2332.1995	2520.623	-0.925	0.355	-7275.015	2610.616
tweetdeck	-2558.0783	3734.800	-0.685	0.493	-9881.834	4765.677

Omnibus:	2049.458	**Durbin-Watson:**	1.208
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	65343.942
Skew:	3.917	**Prob(JB):**	0.00
Kurtosis:	27.271	**Cond. No.**	20.3



We can get the following conclusions from the multiple linear regression summary and the plot

- Based on Multiple Linear Regression : If our tweets are from *vine* we predict its favorite_count to be 8,034, the tweets from *iPhone* 1,286 greater than *vine*, a *web_client* 2,332 less than *vine*, a *tweetdeck* 2,558 than *vine*.

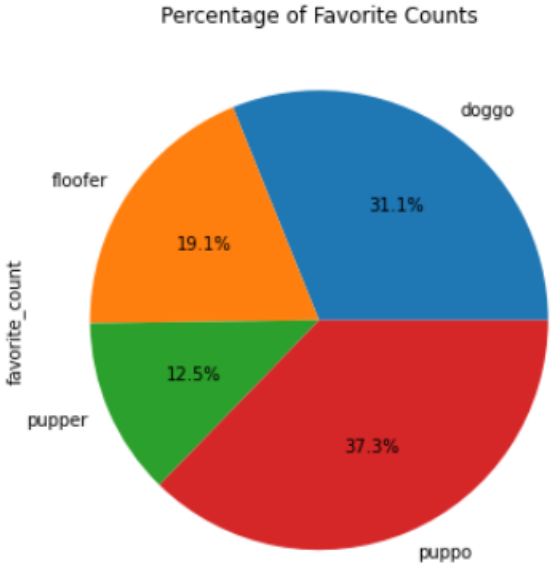
- Based on plot: we can notice that if we tweets from iPhone the means of favorite_count will be higher.

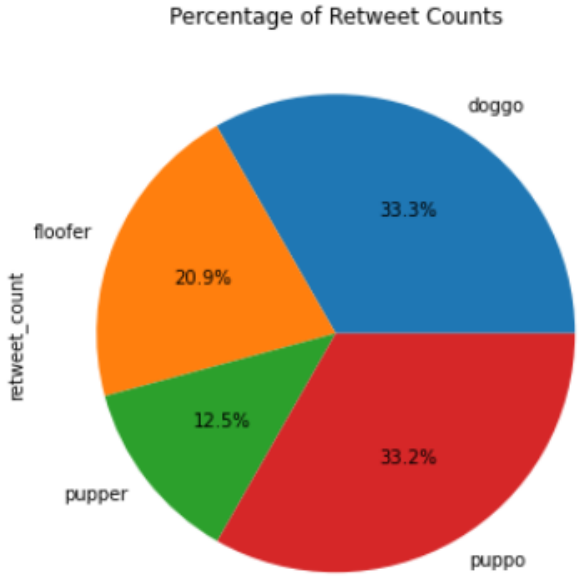
3. The most popular breed of dogs

- Based on prediction1 for the algorithm, the most favorite breed is **Golden Retriever** with sum of **1,977,583** favorite_count.
- Based on prediction2 for the algorithm, the most favorite breed is **Labrador Retriever** with sum of **1,670,195** favorite_count.

- Based on prediction3 for the algorithm, the most favorite breed is **Labrador Retriever** with sum of **776,200** favorite_count.

4. The heighest favorites and retweets means by dog stages





From the above two pie plot we can conclude the following:

1. The highest favorite_count for **puppo**
2. The highest retweet_count for **doggo**

5. the relationship between favorite_count and rating

I found these conclusions about the relationship between favorite_count and rating:

1. When the rating was > (median = 11) the mean of favorite_count was = 15,936, and the retweet_count = 4,934

2. When the rating was <= (median = 11) the mean of favorite_count was = 4,685, and the retweet_count = 1,625
3. The Higher rating go along with higher favorite_count and retweet_count