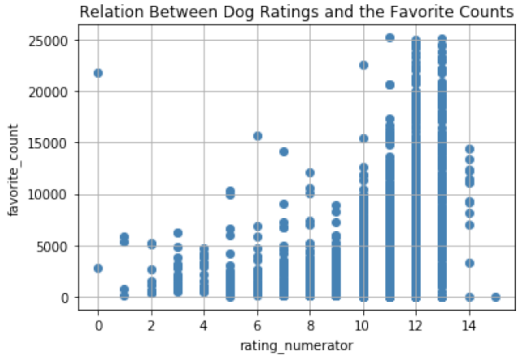
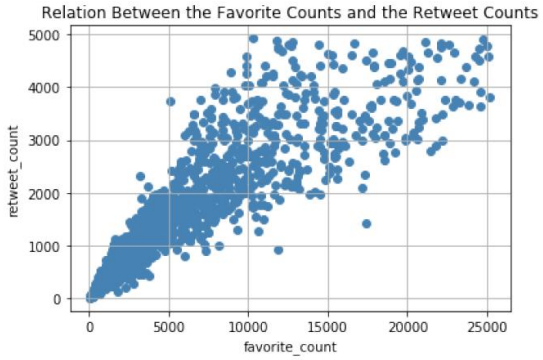
WeRateDogs is a Twitter account that rates people's dogs. I downloaded their Twitter archive. Also a data frame from the Udacity's neural network that can classify breeds of dogs was also downloaded to answer two questions: Which factor influences the number of retweet? Is it the number of favorite, or the ratings of a dog? Whether the neural network could successfully predict an image to be a dog?

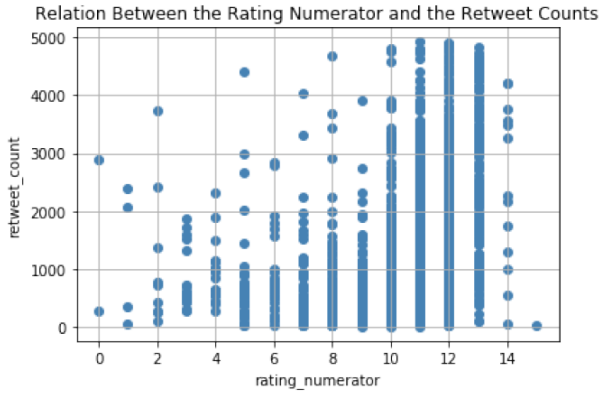
To answer the first question, I plot the rating_numerator with favorite_count. It is fair to say with a high rating, it is possible to have more favorites.



Then I plot the favorite_count with the retweet_count. From this chart we could easily see a trend with these parameters. The number of favorite has a positive relation with the number of retweet.



Also, the ratings of dogs and the number of retweet were plotted. It is also fair to say that with a higher rating, it is possible to get more retweets.



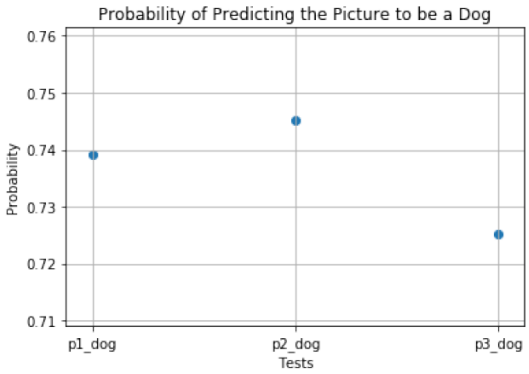
Till now, I have analyzed the relation between ratings, favorite_count and retweet_count through three plots. Although a trend could be found through these plots, it is good to have a quantitative conclusion. Hence, linear regression was used find the trend.

```
                    Results: Ordinary least squares
==================================================================
Model:              OLS              Adj. R-squared:     0.709
Dependent Variable: retweet_count    AIC:                28340.8659
Date:               2021-02-13 15:33 BIC:                28357.3492
No. Observations:   1798             Log-Likelihood:     -14167.
Df Model:           2                F-statistic:        2194.
Df Residuals:       1795             Prob (F-statistic): 0.00
R-squared:          0.710            Scale:              4.0958e+05
------------------------------------------------------------------
                   Coef.   Std.Err.    t    P>|t|   [0.025   0.975]
------------------------------------------------------------------
intercept        361.5038  55.7804  6.4808 0.0000 252.1026 470.9051
favorite_count     0.1923   0.0030 63.1994 0.0000   0.1863   0.1983

rating_numerator   6.4531   5.3484  1.2065 0.2278  -4.0366  16.9428
------------------------------------------------------------------
Omnibus:               994.817      Durbin-Watson:          1.659
Prob(Omnibus):         0.000        Jarque-Bera (JB):       10228.901
Skew:                  2.410        Prob(JB):               0.000
Kurtosis:              13.645       Condition No.:          26646
==================================================================
* The condition number is large (3e+04). This might indicate
strong multicollinearity or other numerical problems.
"""
```

From this table a conclusion can be obtained. For each additional unit increase in the favorite_count, the retweet_count is expected to increase by 0.19 as long as all the other varibales stay the same. The P-value suggests that this is statistically significant. For each additional unit increase in the rating_numerator, the retweet_count is expected to increase by 6.5 as long as all the other varibales stay the same. However, the P-value suggests that this is statistically insignificant.

Next question I want to ask is wheter the algorithm could predict the picture to be a dog?



From this plot we could see the success rate of predciting the picture to be a dog is around 74%. It is not a very high rate. The algorithm needs to be improved.

Conclusion: In this project, we analyzed which factor influence the retweet_count and whether the algorithm is good at predicting whether a image is a dog. We used linear regression to find that favorite_count has a positive influence on retweet_count. Also we found the algorithm could figure out whether a image is the picture of dog with a success rate above 70%.