Act Report for Data Analysis and Visualization

1. First thing that needed to be checked was if the parameter *retweet_count* is correlated with the *favorite_count* in general, since we know that normally it is true for a tweet to get widely liked (favorite) and widely spread (retweet).

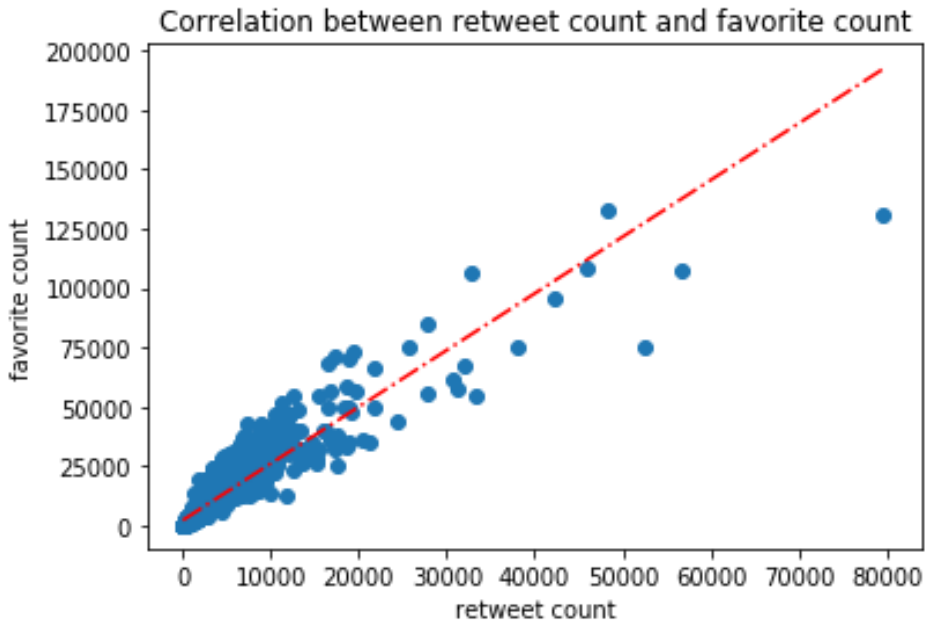

Figure 1 Correlation between retweet count and favorite count

From *Figure 1*, it is clear that the correlation of the two is quite easy to observe. After checking the linear regression model, the correlation coefficient R-squared is 0.833, which is quite high, and the slope for the linear regression line is significantly different from 0, aka the flat line.

2. Next thing checked was if the *retweet_count* or *favorite_count* correlated with the time of the tweet. This is because generally the tweet account gets more and more followers over time, so the number of retweet and favorite could simply increase over time regardless of the content of the tweets.

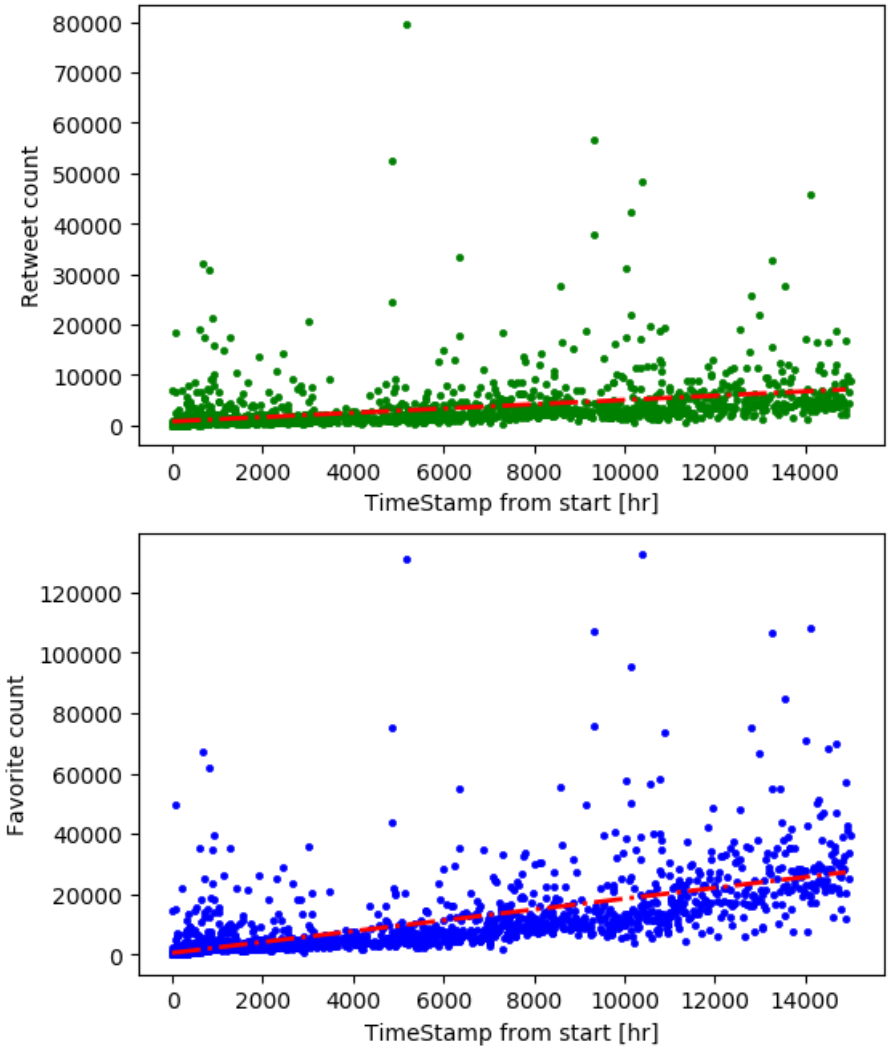

Figure 2 Retweet count and favorite count over time

Here the unit of the timestamp is *hour* instead of *second* or *minute* because the tweets were collected over 2 years and *second* or *minute* are too precise for the linear regression. From *Figure 2*, a more tilted slope of the linear regression line could be observed in favorite count's graph than retweet count's. And it was also true statistically, with coefficient of 0.42 for *retweet_count* compared with 1.79 for *favorite_count*, both had p-value less than 0.05.

On the other hand, the correlation coefficient R-squared was also smaller for *retweet_count* (0.166) compared with *favorite_count* (0.432).

In conclusion, the *favorite_count* parameter seems to have a stronger correlation with time as compared to the *retweet_count* parameter, although both are generally increasing over time.

3. Last thing needed to be check was to see if the parameters: *retweet_count, favorite_count* and *rating* were correlated with whether or not a dog was in the picture of the tweet. Note that the *rating* parameter was defined as *rating = rating_numerator/rating_denominator.*

   Since we have the image prediction results to show if there is a dog/dogs in the picture, I used these results as the label for dog existence, although there are limitations sometimes. Logistic regression was used to analyze the correlation to the dog existence because it was a categorical/Boolean value.

   As a result, only *favorite_count* showed a statistically significant coefficient to the logistic regression model with p-value of 0.0043, and the other parameters were not significant, with p-value of 0.21 for *rating*, and 0.27 for *retweet_count*.

   It seems people are more likely to “like” the tweets which contain a dog/dogs.