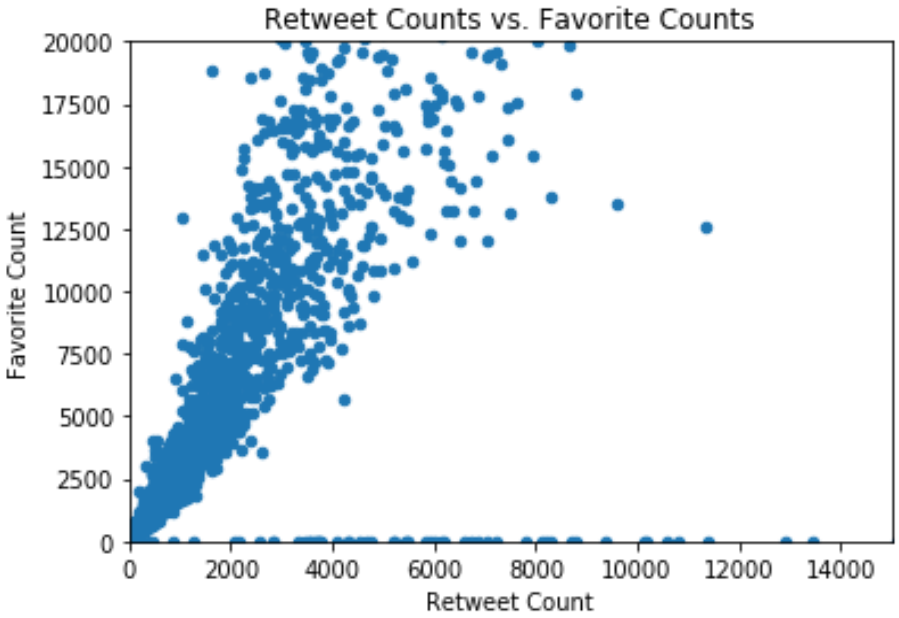
Analyzing and Visualizing

In this part, I focus on the relations between the following variables:

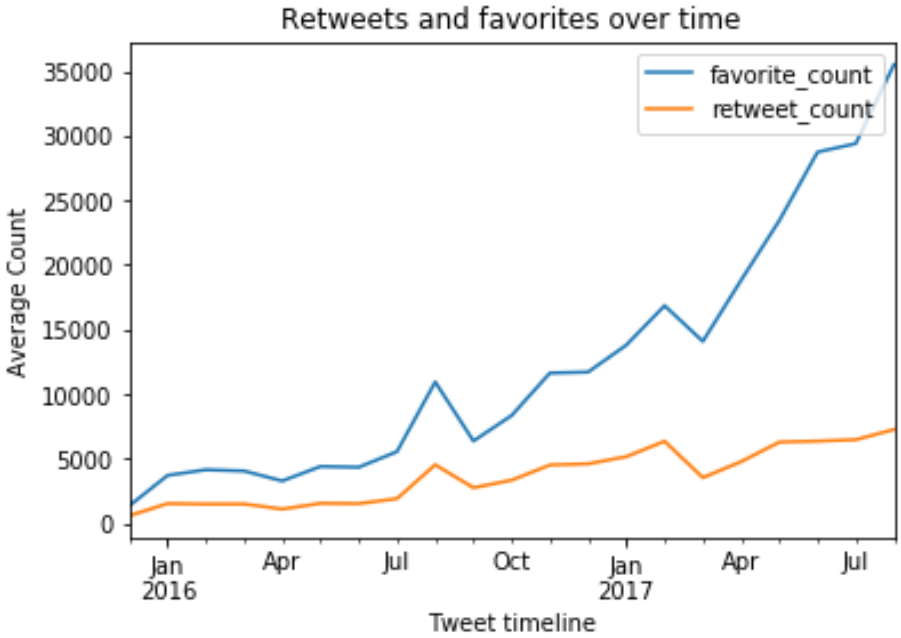
- Retweet and Favorite Counts
- Rating Ratio
- Month and Year

Retweet vs. Favorite Counts



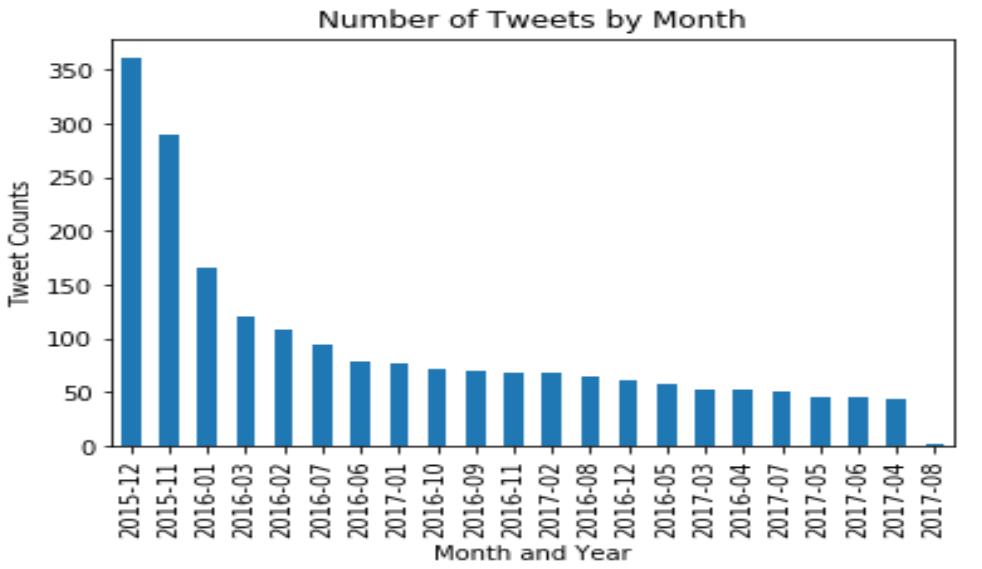
The Pearson correlation coefficient of retweet count and favorite count is at 0.8559, it is close to 1. This indicates that there is a strong positive correlation, which means a tweet with a high favorite count can expect a high retweet count and vice versa. We can see it in the generated scatter plot above. I set a limit for the plot since there are some outliers in the dataset with extreme great numbers.

Average Retweet vs. Favorite Overtime



This plot shows the average retweet counts and favorite counts by month from December 2015 till 1st August 2017. As we know from the first chart, the retweet and favorite counts are positive correlated, the correlation is also shown in the second chart. We can notice a continuous increasing trend of the favorite and retweet counts over time. Also the favorite counts are constantly greater than the retweet counts in the whole period. There is a small peak for both counts in August 2016. Several reasons may cause a higher favorite counts in the specific month. To check if there is a relation between favorite counts based on tweet activity I look closer on the tweet counts by month.

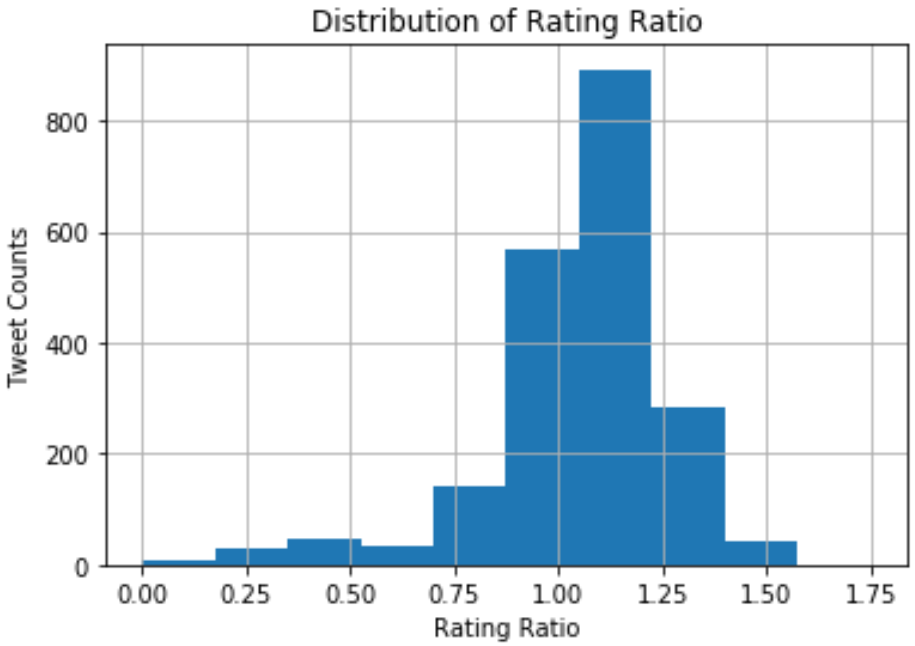
Monthly Tweet Counts



The average tweet activity per month is about 68 tweet, I choose the median as the center of measurement in our case, because median is less sensitive towards outliers. The tweet count for August 2017 is an outlier in our data set, this is due to incomplete data for August, we only have tweet data from 1st August, so the tweet count is not representative for the whole month.

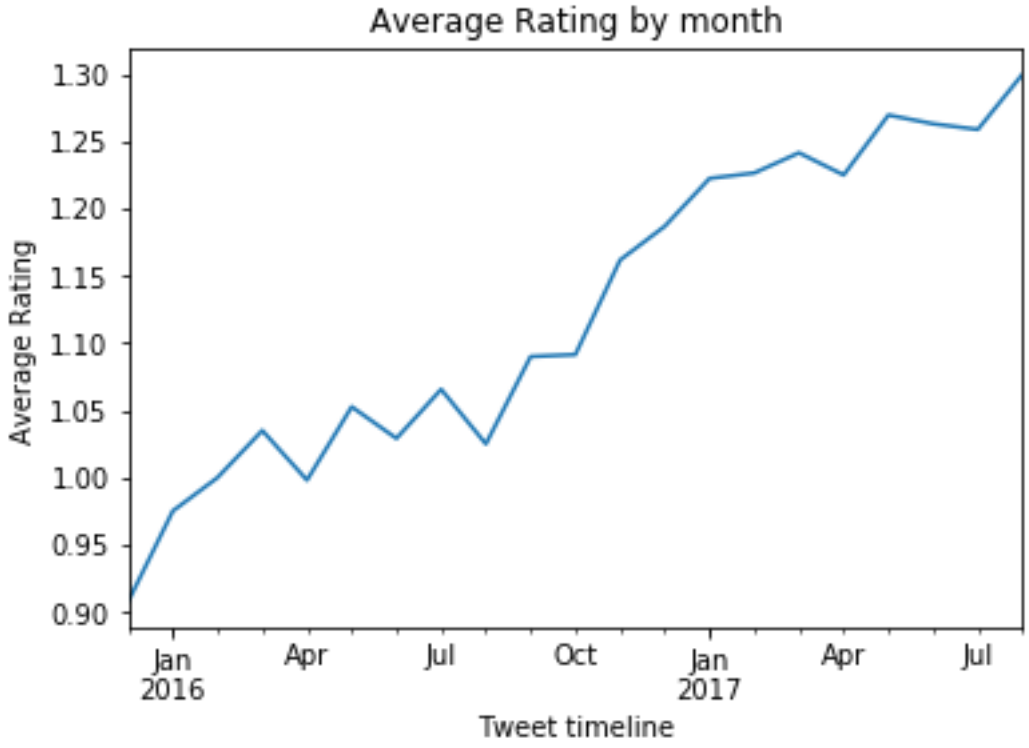
As we can see in the plot, December 2015 is the month with the highest tweet activity with 361 tweets, however, the average retweet and favorite counts are the smallest throughout the period. The tweet counts of August 2016 is less than the tweet counts of July 2016, so we are not able to make conclusion about a relationship between the tweet activity and the favorite counts in our data set.

Rating Distribution



The mean of the rating ratio is 1.065. More than 800 tweets have a rating ratio between 1.1 and 1.2. On average the rating numerator given in the tweets are greater but closer to the rating denominator. The standard deviation of rating ratio is at 0.228, so the data tend to be close to the expected value. The rating numerator given tend to be greater but close to the rating denominator.

However, when we look at the trend of the average rating by month, the rating ratio is growing continuously since December 2015. The average rating ratio is in December 2015 at 0.91 and in July 2017 at 1.26. The Difference between the rating numerator and denominator increase over time.



The lines in second chart and in the last chart have the same increasing tendency over time. According to this I conclude that tweets with higher average ratings have higher favorite and retweet counts. A high Tweet activity do not imply higher favorite and retweet counts.