Data analysis and Visualization

Introduction

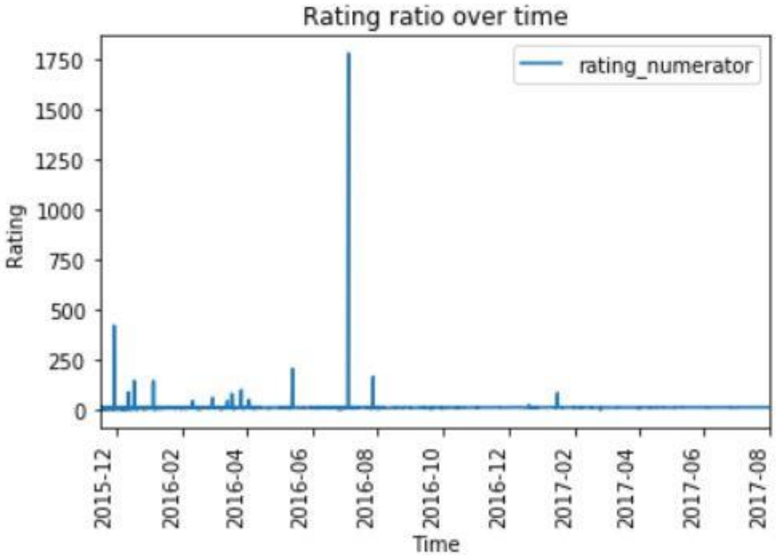
This document contains the documentation of analysis and visualization made on WeRateDogs dataset. The analysis involves finding the outliers in the data and variation of tweet with respect to the time. The visualizations are made using the ‘matplotlib’ library.

Analysis

The key interest of variables in the dataset are rating_numerator, retweet_count, favorite_count, timestamp and source. The initial analysis contains the summary and the descriptive statistics about the data. Then the analysis revealed that the maximum rating given to a dog was more than what expected and it touched 1776. These are the outliers in the data.

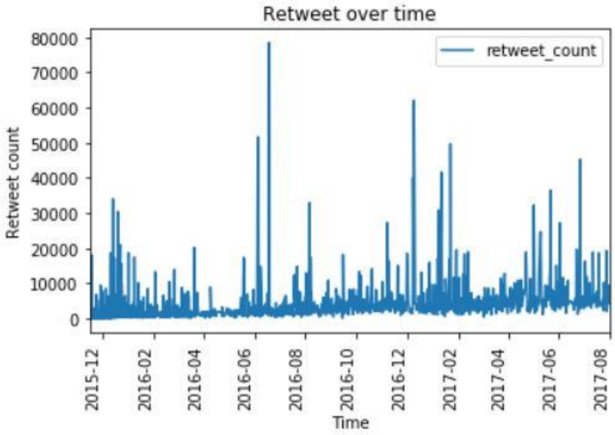
Visualization

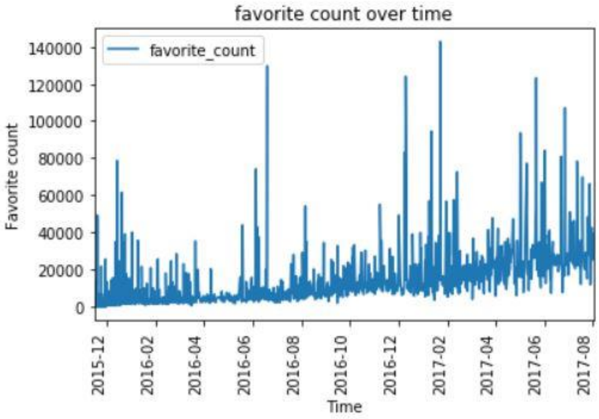
The visualization contains the plot which shows the evolution of data over time. In the below plot, the rating_numerator was visualized with respect to time.



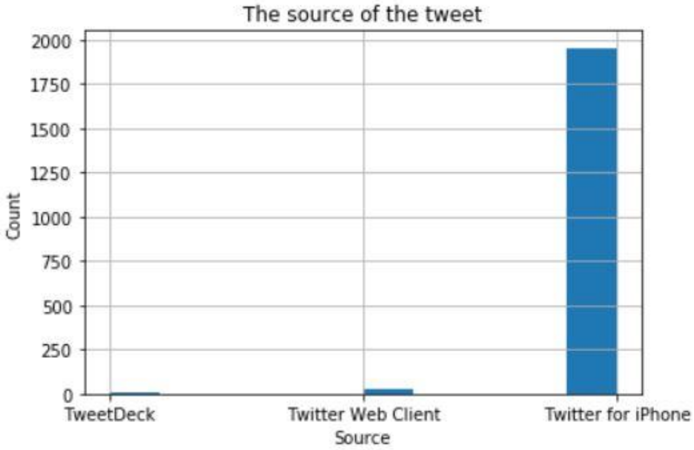
We can find that one data reaches a maximum of 1776 that’s actually a funny tweet and not a mistake.

Then, the favorite_count and retweet_count are plotted with respect to time. This shows the evolution of the data with time. We can observe that both the retweet and favorite count were minimum in the beginning. The retweet count reached a maximum in the year 2016 between june and augest. The favorite count reached a maximum between the 2016, December and early 2017.





The next plot shows the histogram of souce of the tweet. From the graph it can be inferred that the tweets from the iphone is maximum. The tweets from the Web client and TweetDeck are very loww when compared to the tweets from iphone.


The source of the tweet
Count
TweetDeck
Twitter Web Client
Twitter for iPhone
Source