Axis Bank Hackathon

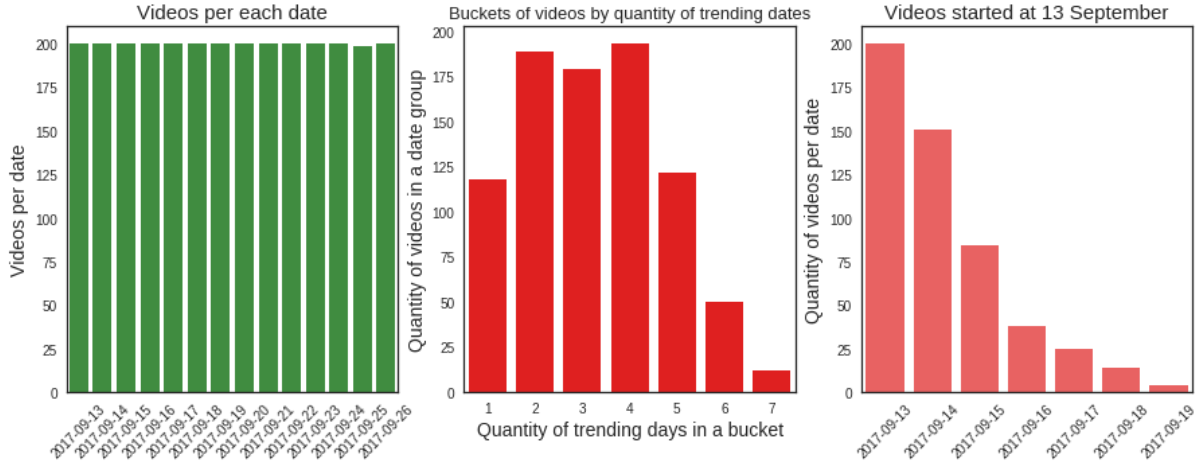
Problem Statement : The dataset includes data gathered from videos on YouTube that are contained within the trending category each day.

Tasks :

- Perform Exploratory Data Analysis on the dataset provided and analyse all the relationships between views, likes, dislikes and comments. Of course you can use other features of the data to support your analysis.
- Perform Sentiment Analysis on the comments for each video and tag them with a sentiment.

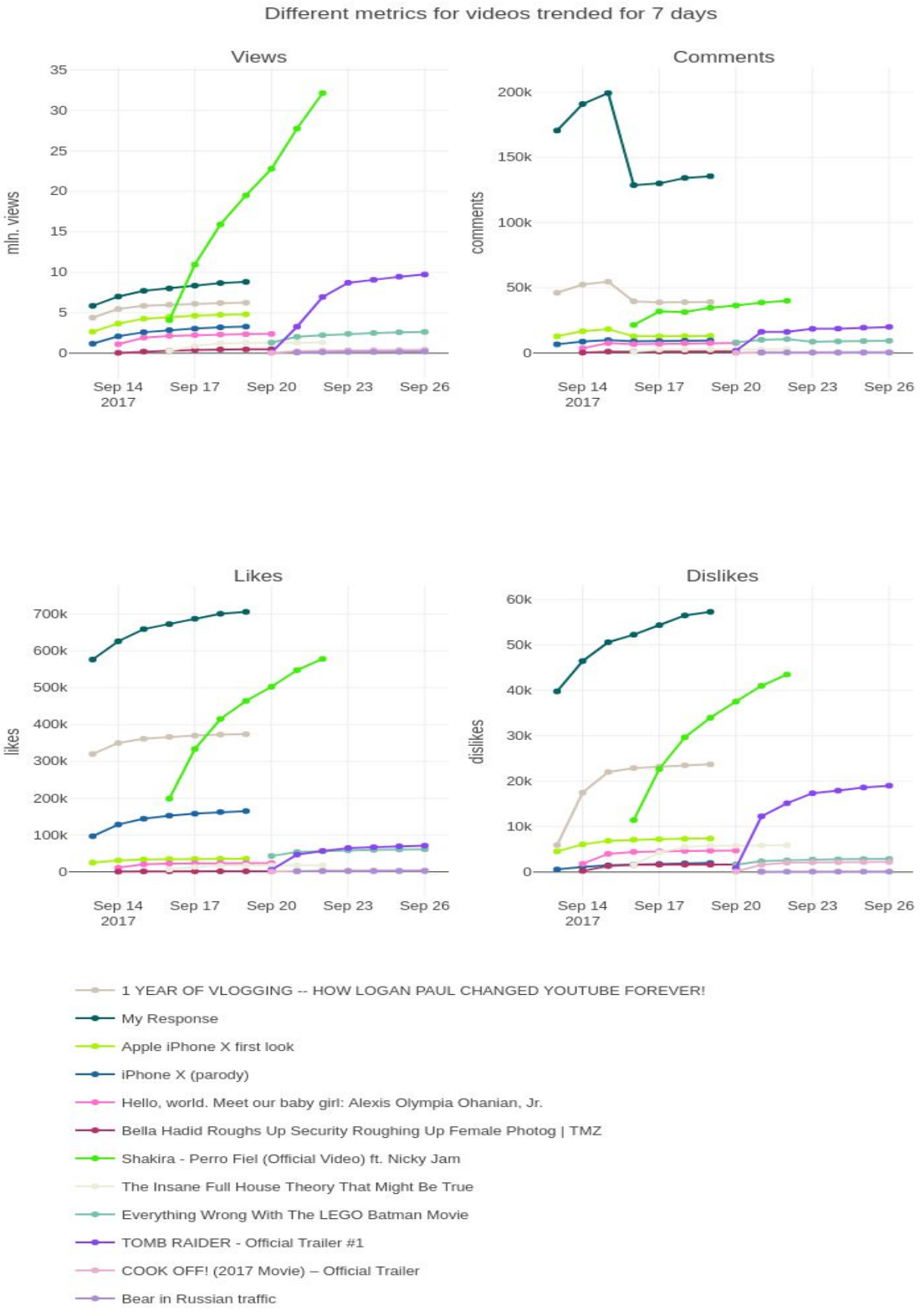
Exploratory Data Analysis

The given data set was test against if it contains any error. Using the pandas package for the data analysis we were able to come across the fact that there were repeating datas. There were also some error in the column ‘date’ of the dataset. Thus relevant code is written to remove such discrepancy.
Also the data format of the given dataset was as ‘26.09’ which resembles the 26th day of September. Thus proper module of pandas was used to make the dates in the format of ‘2017-09-26’.
Now we can our first overall analysis. We are interested in is to support my understanding of database nature (top 200 trending videos per each day; since there are 14 days of dataset and there are approx 2800 rows of data which means 200 videos per day on average), check how many days a video can be in trend and get some details on longest trending videos.
So for that I have written a function as in the code quick_insight(), which takes the dataset and outputs three graphs. These were the output from the written code.



The above plots suggests that there are in average 200 videos in trending per day. The second plot suggests the quantity of videos that can be in trending. The third video suggest the longest trending videos stats.

I then created another function which is best_survivors(). This another function is designed to give deeper information of what videos were in trend for longest period. How many and what dynamics of views, likes, dislikes and comments are required to become trending video are answered through this plot. The resulting plot is as shown below



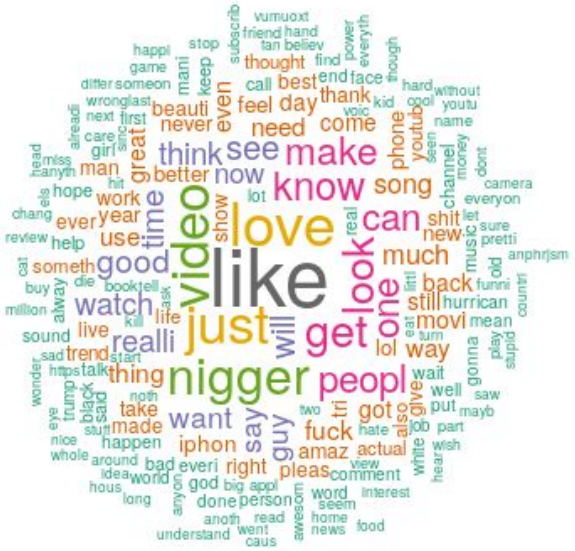
.

Some conclusions from these plots are:

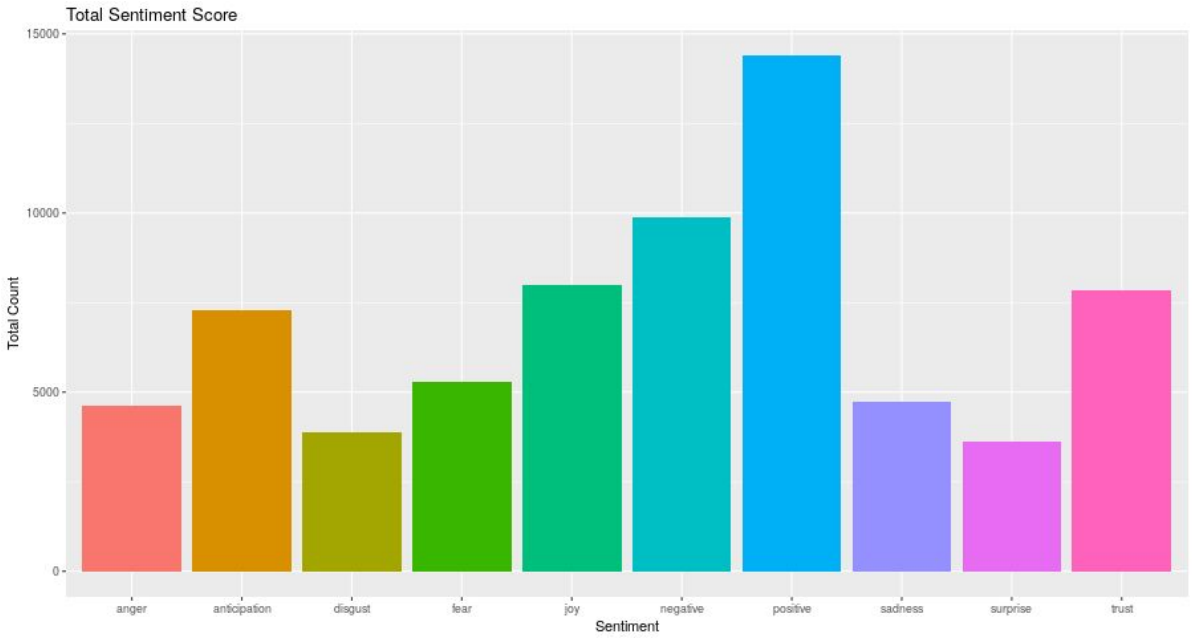
- There are 200 trending videos per each date. Database does not look for some specific videos, only for top 200 trending. Video may be in this top list for a few days or only one day.
- Maximum amount of days being in trend for audience was 7 days.
- These plots also reveal that in most cases likes and dislikes have similar shapes but different magnitudes.
- Some of the top trending videos have so low amount of views/comments. It can be because trending is more about speed of new views than the amount of views overall.

Sentiment Analysis

R was used for sentiment analysis. The given UScomments.csv file was used and its first 20,000 lines were used for the analysis. The word cloud was created and it is displayed below:



The total sentiment plot was also drawn and the plot is given below:


Total Sentiment Score
15000
10000
5000
0
Total Count
anger
anticipation
disgust
fear
joy
negative
positive
sadness
surprise
trust
Sentiment