**ABSTRACT**

Youtube is a one-stop for all the creators, teachers, artist, or any person who wish to share his talent over the globe. Youtube has connected people like never before. From how to solve a complex maths problem to how to set up a drone, everything could be found on youtube. For a creator, the most important thing is to reach a far-reaching audience as possible. So we came up with this idea of analyzing various videos and finding trends in videos that do make it to the trending page.



With the advancement in technology, it has become feasible for us to store large databases, anlayse and make a future prediction by studying the ongoing trend. With the onset of service provided by jio, Youtube became our google, the blogs we used to read could now be viewed and understood with much more clarity. The one question that usually comes to our mind after posting a video is "will it make to the trending page", "will it get enough views", "what would have been the right time to launch it"? These questions which went unanswered could now be answered by analyzing previous trends, by remembering and analysing what we already know:- the dynamic programming approach, but now with databases that measure hundreds and thousands of GB and techniques which have a higher probability of predicting future trend accurately.

**INTRODUCTION**

YouTube maintains an inventory of the highest trending videos on the platform which is updated constantly YouTube uses a mixture of things including measuring user interactions (number of views, shares, comments and likes).

To this date, it's difficult to seek out high-level statistics on YouTube that paint a good picture of the platform in its entirety. This project attempts to provide an overall characterization of YouTube, supported a random sample of channel and video data.

Through the pictorial representation, we’ll analyse the role of views,comments, likes, dislikes in a video and using these as parameter try to predict whether a video will or will not make it to the trending page. Also we’ll compare several algorithms and their results and use them for predicting future videos growth rate.

**BACKGROUND STUDY**

* According to the study done by **Dustin J. Welbourne** on content analysis of 390 videos from 39 YouTube channels,he found that -although professionally generated content was superior in number, user-generated content was significantly more popular. Furthermore, videos that had consistent science communicators were more popular than those without a regular communicator
* In another study done by **Xu Cheng, Jiangchuan Liu and Cameron Dale,** they found that the videos have strong correlations with each other, and creates opportunities for developing novel caching and peer-to-peer distribution schemes to efficiently deliver videos to end users.

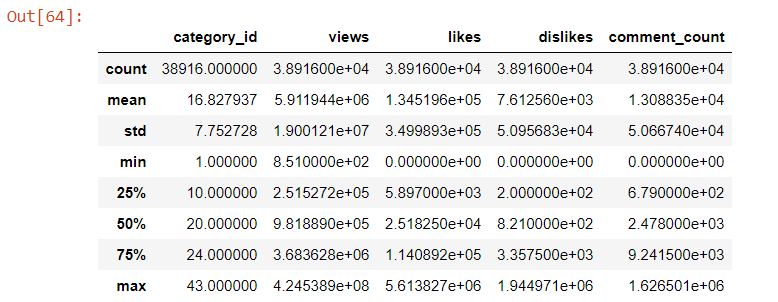
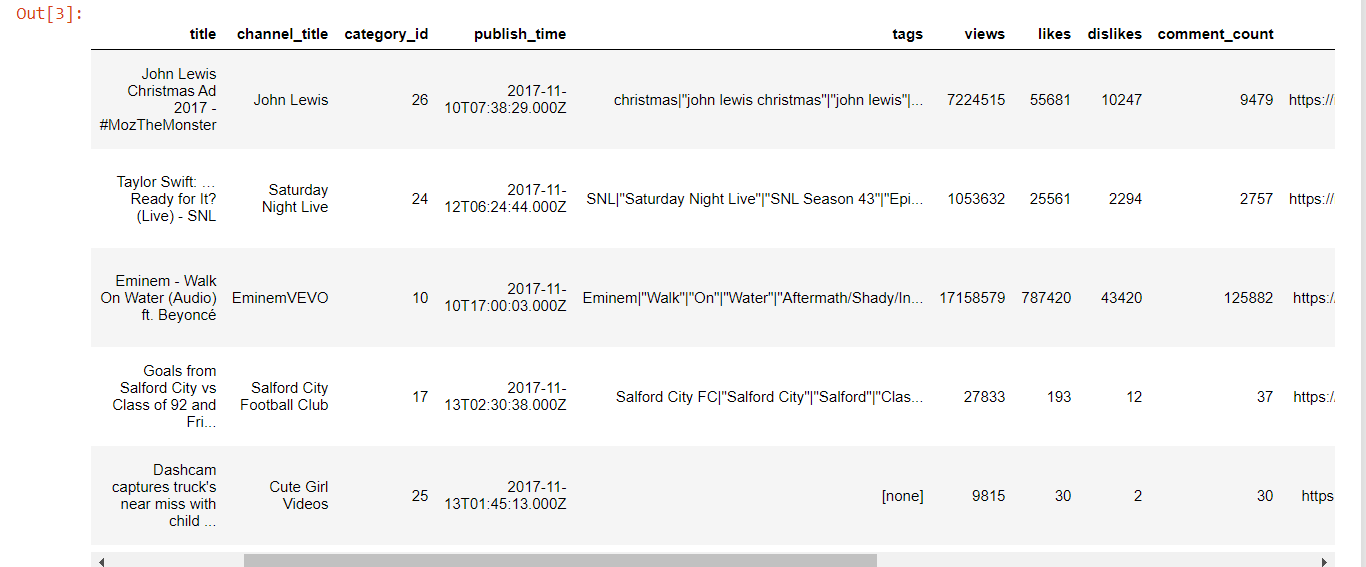
**SYSTEM ARCHITECTURE**

* **DESCRIPTION :**

In this projects, we will use **Python** with some packages like **Pandas**, **Matplotlib, seaborn, scikit learn etc.** to analyze a dataset that was fetched from Kaggle. The dataset contains data about the trending videos of a particular day. It contains data about **more than 35,**000**trending videos**. We analyzed this data to get an accurate and deep understanding of the trending videos, to see what was common between these videos.

* **DATASET :**

The data was collected for many countries : US(United states of America), GB(Great Britain), DE(Germany), CA(Canada), and FR(France). We chose the data for GB country only.



**This is a short representation of the dataframe showing few of the columns.**

The dataset has 11 independent variables and 3 dependent variables.Our motive is to provide a systematic data pre-processing analysis working only with the dataset GB Videos. This step is important for all data mining exercises and we want to emphacize it. Before building theories from data we had to understand key data attributes, like missing values, outliers, unique counts, and time-series trends.

The description of the dataset is as follows:

• Video\_id :Single unique attribute of the entire dataset. It represents unique video id of an uploaded video on youtube database.

• Publish\_Date :It represents the date, when video was uploaded or published on Youtube site. This attribute derived from the “publish\_time” attribute of the original dataset.

• Publish\_Hour :It represents the hour, when video was uploaded or published on the Youtube site. This attribute is also derived from the “publish\_time” attribute of the earlier dataset.

• Category\_Id :It represents the distinct category id from the Youtube database, where video was uploaded/published.

• Channel\_Title :It represents the channel name of author/publisher.

• Views :It represents how many times video was viewed and reviewed by Youtube users.

• Likes :It represents the how many time’s video was liked by Youtube users.

• Dislikes :It represents the how many time’s video was disliked by other Youtube users.

• Comment\_Count :It represents the how many time’s Youtube users(including publisher) started a conversation on the video .

• Comments\_Disabled: It represents whether or not video author/owner allows others to start a conversation on the video.

• Ratings\_Disabled :It represents whether or not video author/owner allowed other users to like or dislike the video content.

• Title :It tells the title/name of the Youtube video.It was given by the author of the video.It could not be null.

• Tags :It tells the list of tag publisher attached on the video.Each tag separated by ‘|’ .Tags are functioning like keywords in Youtube.

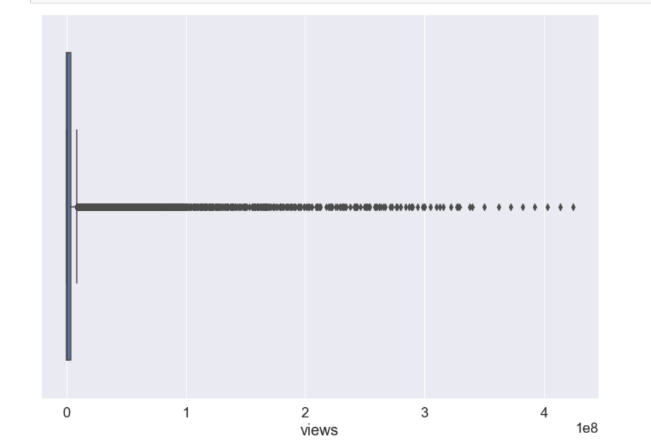
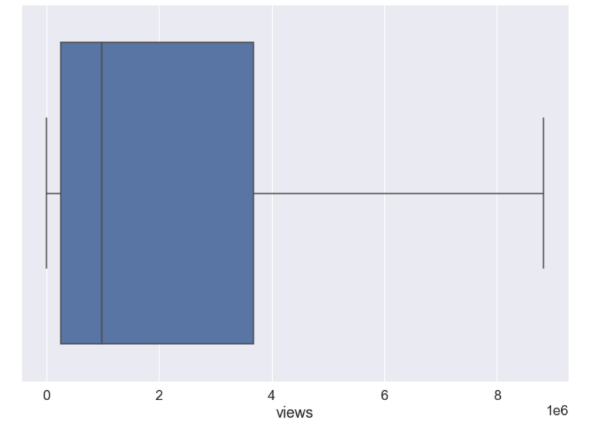
• Description : It describes the content of the Youtubevideo.It was given by the author of the video.It is optional,so it could be empty/blank.

**DATA PREPRODESSING AND FEATURE SELECTION**

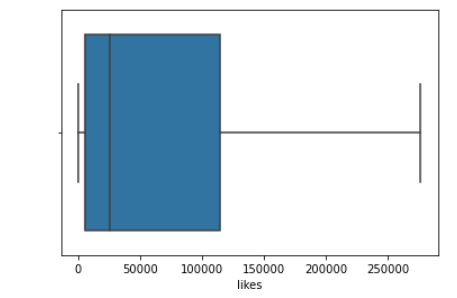
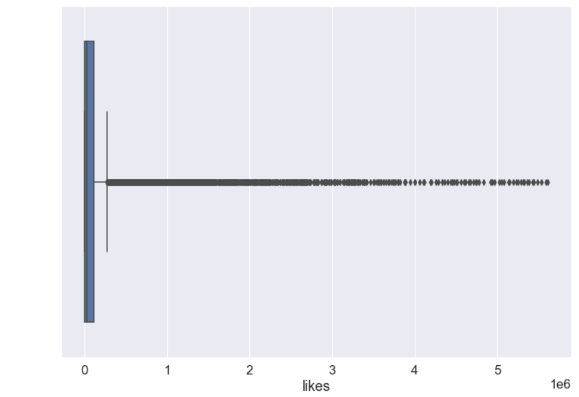
We modified the dataset to analyze some hidden information. Such as, we removed redundant video\_id’s, retrieved data using Category ID(json file)and also removed missing values, unrelated attributes and outliers. As per our requirement, w some new attributes derived through the original ones to analyse the data in a meaningful way. Below are the detailed steps that we undertook to ensure that the data is cleaned and represented properly so that it is simple to analyse it.

• First we checked if there are any empty values or not and found out that there are 612 null values in the column ‘ description’ and thus replaced each one of the null values with a blank string(“ “) as the dtype of the column is string.

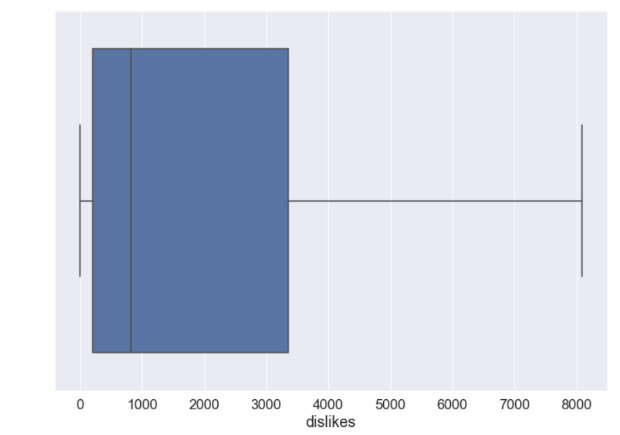
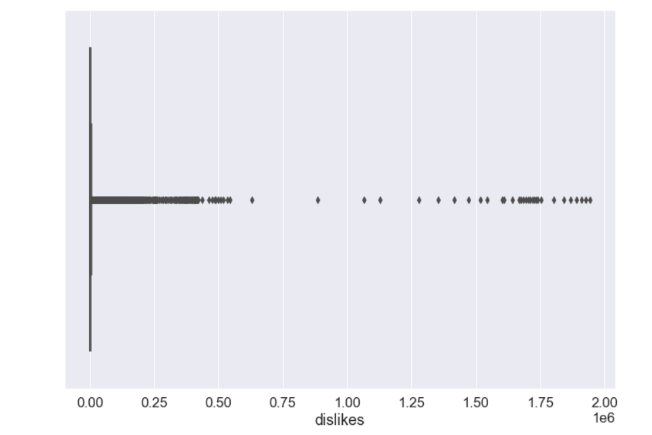
• We originally had 38916 rows in our dataset and to clean our data further we searched the data for outliers, plotted some boxplots with and without the outliers to analyse our remaining data. After removing the outliers and filling in the missing values we had 25,000 rows of data left. Below are some of the boxplots with and without the outliers as shown:



**fig1.Boxplot of ‘views’ without the outliers. Fig2.Boxplot of ‘views’ with the outliers.**

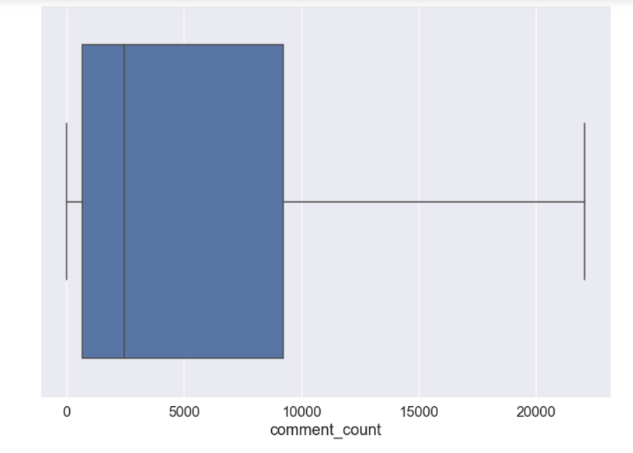
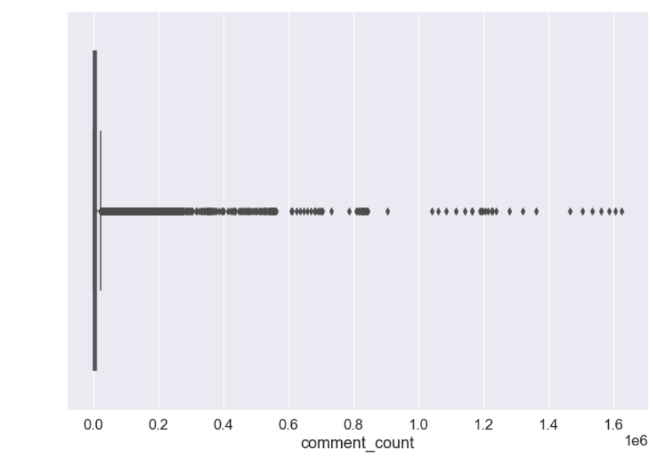


**Fig3.Boxplot of ‘likes’ without the outliers. Fig4.Boxplot of ‘likes’ with the outliers.**



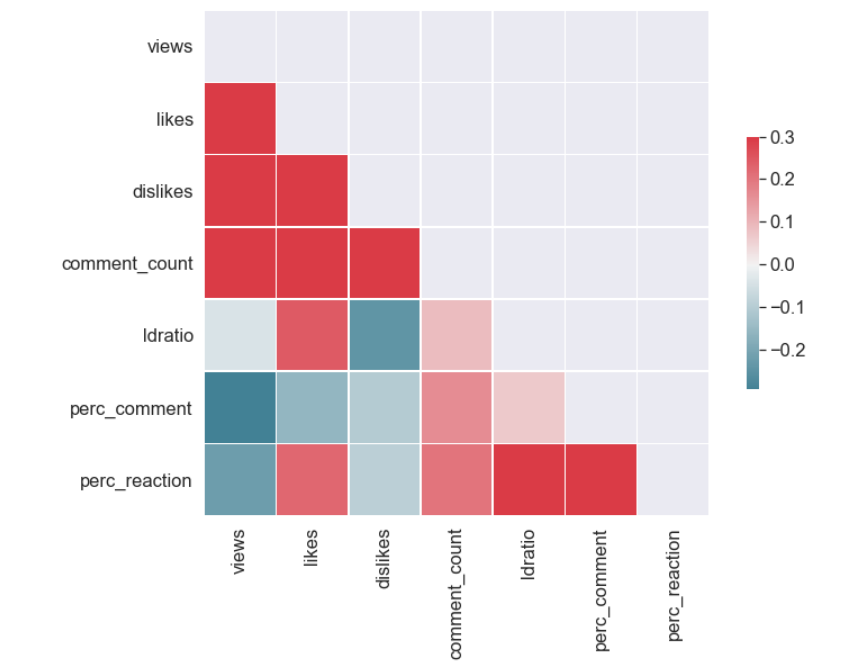
**Fig5.Boxplot of ‘dislikes’ without the outliers**

**Fig6.Boxplot of ‘dislikes’ with the outliers**

**fig7.Boxplot of ‘comment\_count’ with the outliers fig8.Boxplot of ‘commentcount’ without the outliers**

* After removing the outliers we also removed the unnecessary attribute(‘thumbnail\_link’) and substituted true, false with 1,0 respectively in the boolean columns(‘comments\_disabled’, ‘ratings\_disabled’, ‘video\_error\_or\_removed’) so that it becomes easier to predict and analyse the data.
* Using the category ID json file we successfully placed the categories at the respective indexes and also added some new columns so that we could extract more conclusions from the graph. The new columns added were derived from the the original ones like the added ones are trending date, trending, days to trending, dislike percentage, ldratio, publish time, prec\_comment, perc\_reac etc. Setting the date day time columns like publish time, days to trending to their right format was neccesary as they were jumbled up.
* We selected the target value as ‘trending’ to determine whether the video will go up on trending or not and the training columns as every column except for ‘tags’, ‘category id’, video\_id’.

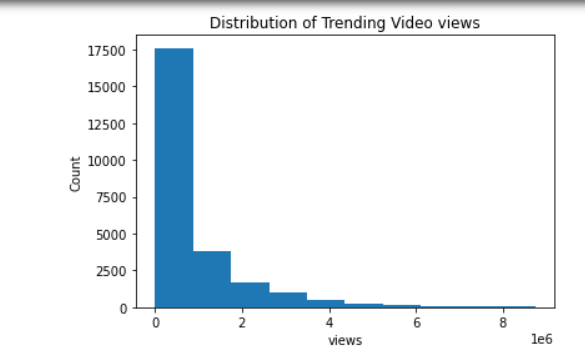
**ANALYSING GRAPHS AND CONCLUSION**



**Fig9.**

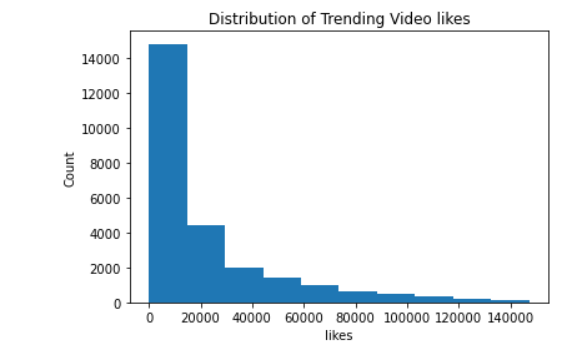
**Correlation analysis and heatmap**

* From this correlation we find that likes,dislikes and views are highly correlated with each other. We also found out that ldratio has low correlation with views and dislikes.



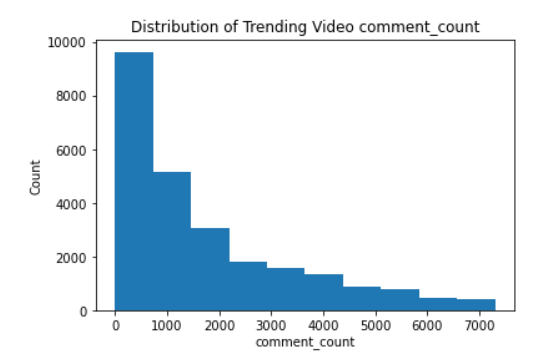
**Fig10.Distribution of video views**

* This graph showed us the count of the videos in a particular range of the no. of views, the units for ‘views’ is in lakhs. As we can observe the maximum no. of videos have views between 0 to 1 lakhs.



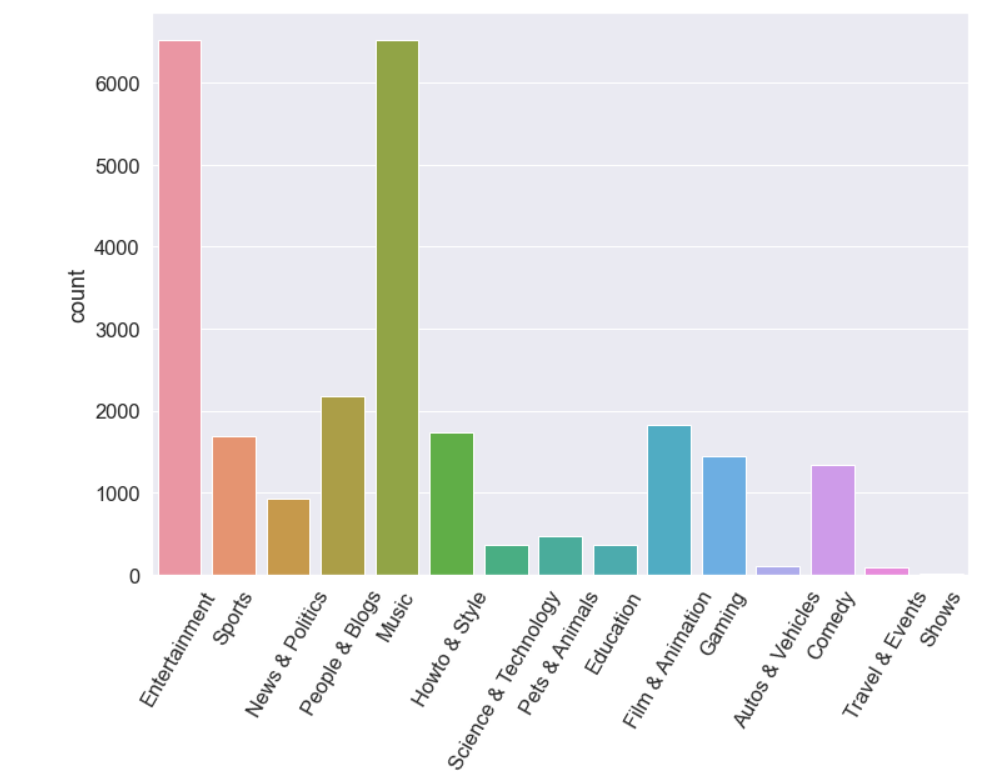
**Fig11. Distribution of video likes**

* This graph showed us the count of the videos in a particular range of the no. of likes. As we can observe the maximum no. of videos have likes between 0 to 20thousand.



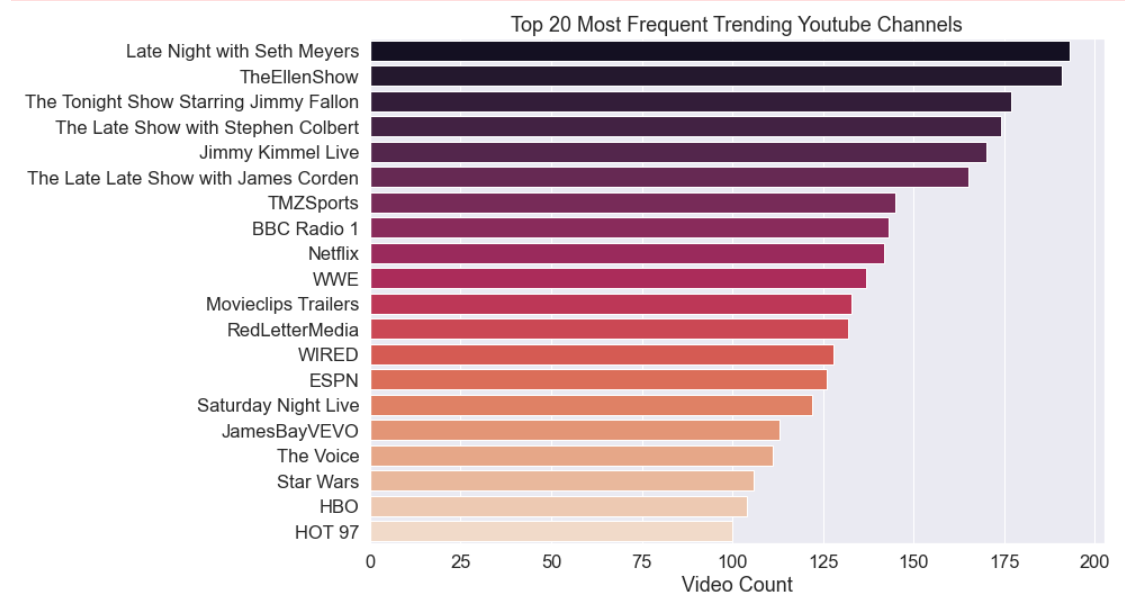
**Fig12.Distribution of comment\_count**

* This graph showed us the count of the videos in a particular range of the no. of comments on the videos. As we can observe the maximum no. of videos have comments between 0 to one thousand.



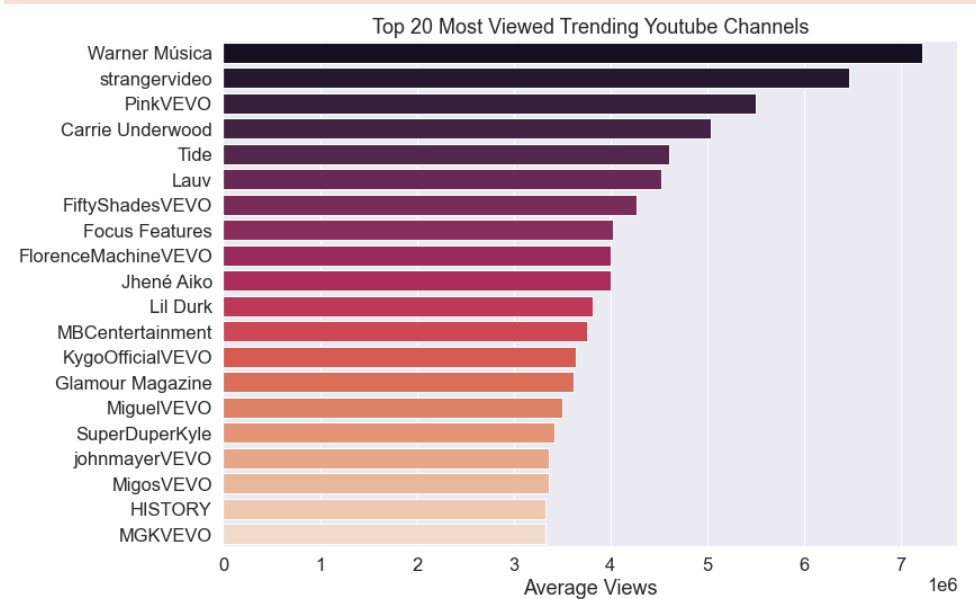
**Fig13.Video Category Distribution**

* This graph represents the count of videos in each category. As we can observe music and entertainment have the highest no. of videos and travel and events have the lowest videos posted on Youtube.



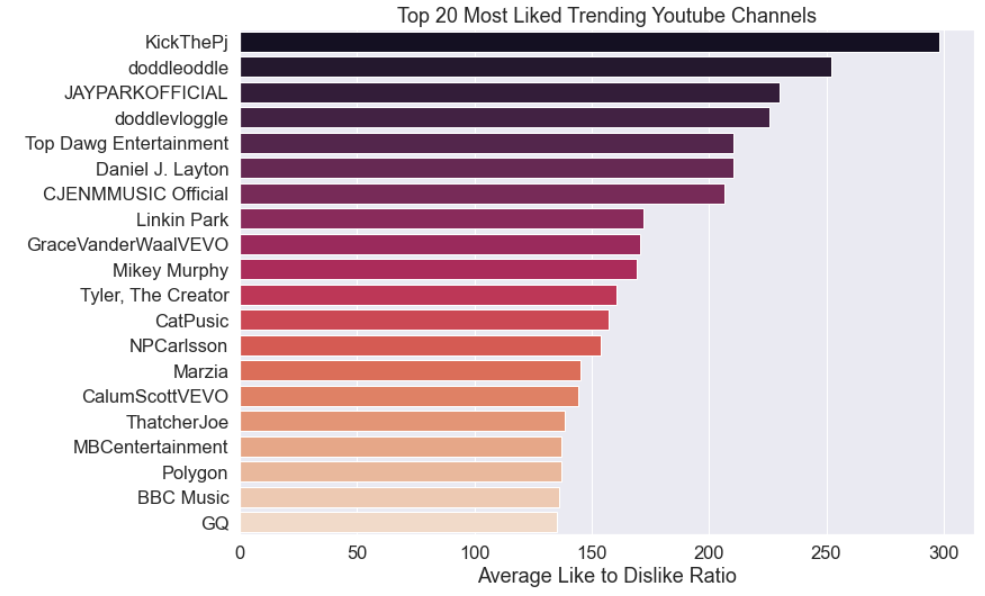
**Fig14.Most Frequent Trending channels**

* Thos graph represents the top 20 youtube channels up there on the trending chart for the longest period of time. As we can observe ‘late night with set mayers’ was the most trending youtube channel.



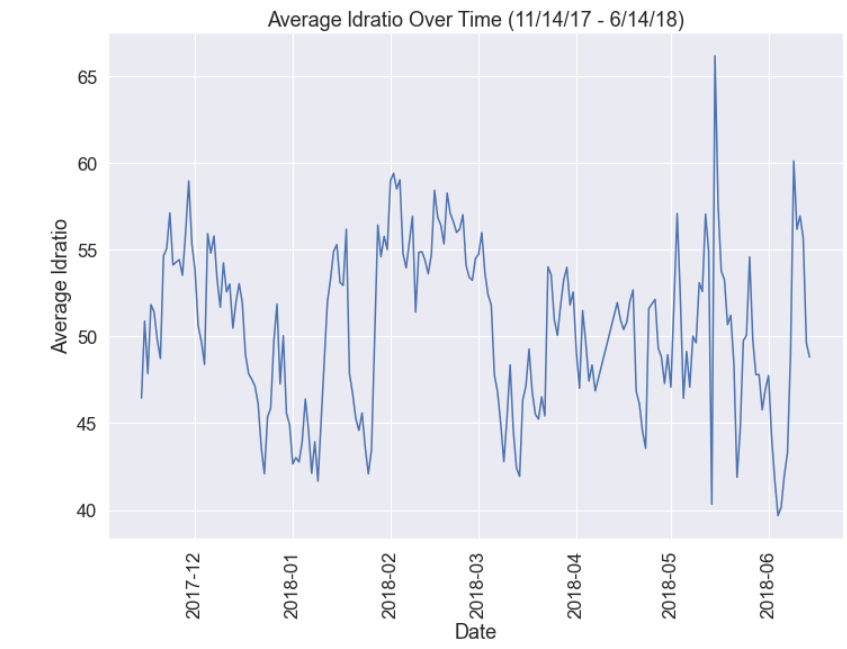
**Fig15.Most Viewed channels**

* Thos graph represents the top 20 youtubechannelswhich have the maximum no. of views. As we can observe ‘Warner Musica’ has the highest no. of views.



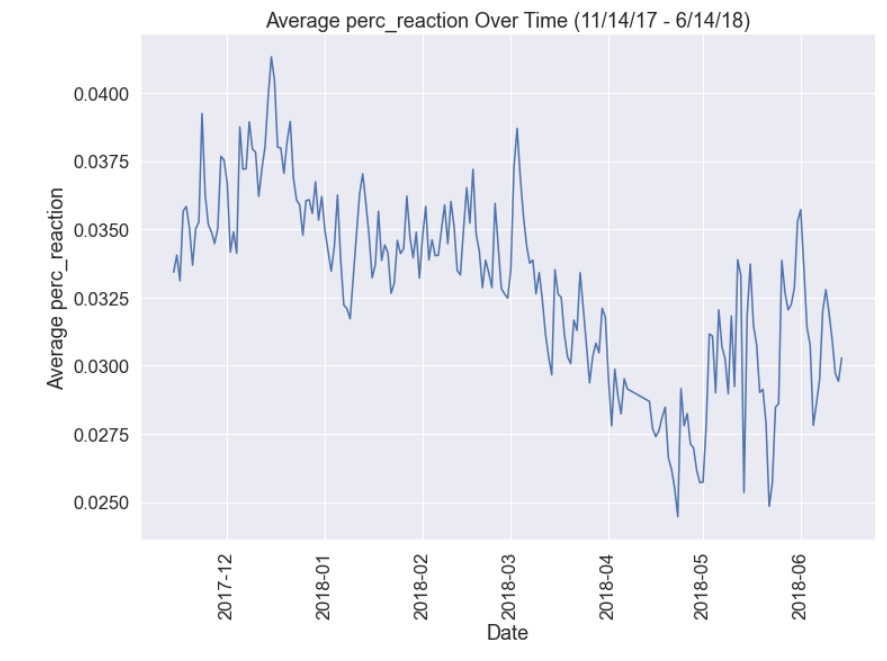
**Fig16. Most Liked channels**

* Thos graph represents the top 20 youtube channelswhich have the maximum no. of likes. As we can observe ‘Kick the pj’ Was the most liked Youtube channel.



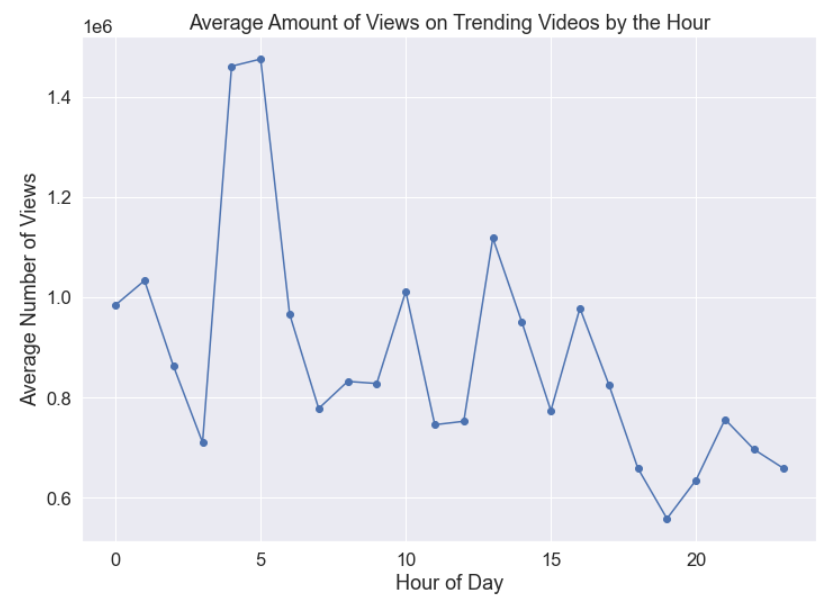
**Fig17. Time Variation of ldratio**

* This graph represents the variation of ldratio with the month-year. It can be observed that the ldratio was highest in 2018 between May and June.



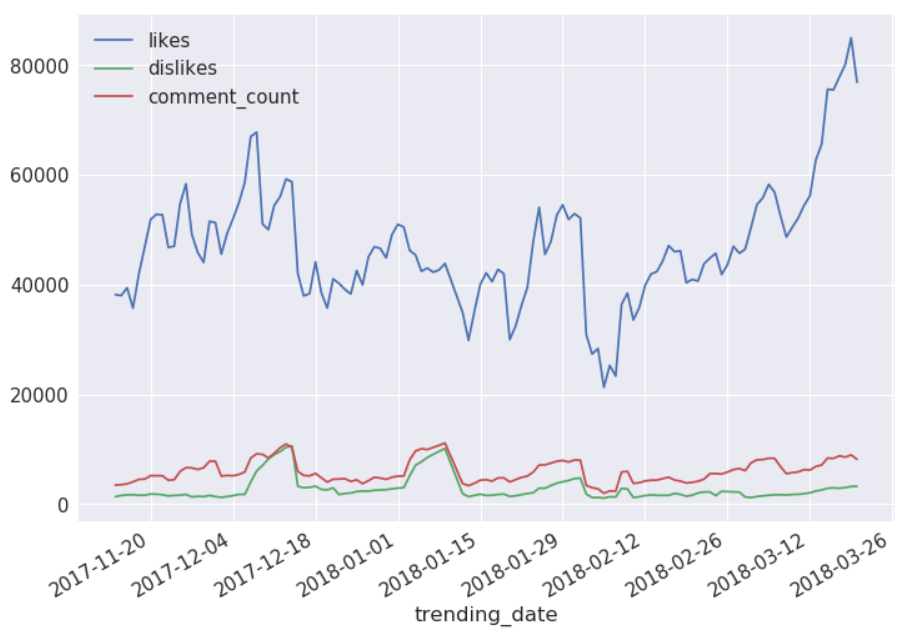
**Fig18. Time Variation of perc\_reaction**

* This graph represents the variation of perc\_reaction with the month-year. It can be observed that the ldratio was highest in 2017 between Dec and Jan.Prec\_reaction is the total no. of likes and dislikes with the total no. of views



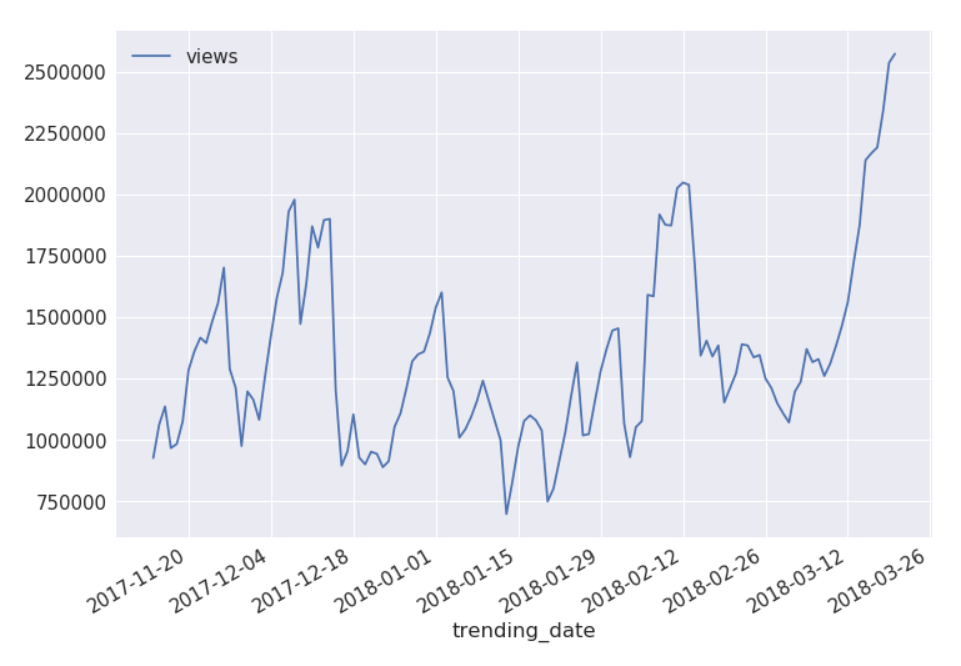
**Fig19. Time Variation of no. of views everyday**

* This graph represents the variation of average amount of views on the trending videos at each hour of the day. It can be observed that the highest no. of views came at 5 or 6 in the morning in Great Britain.



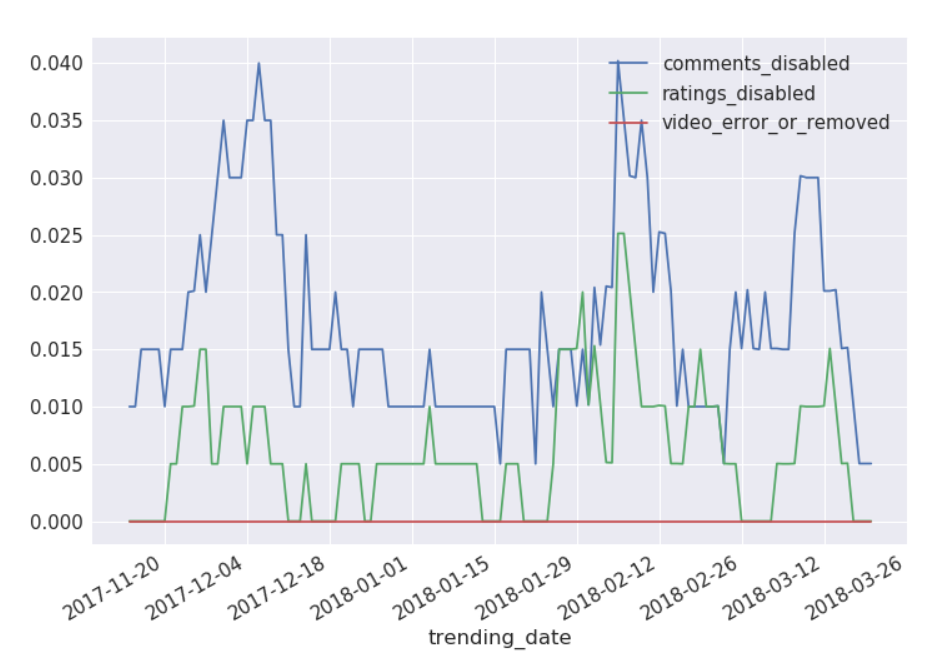
**Fig20. Time Variation of likes, comment\_count, dislikes**

* This graph represents the variation or comparison of likes, comment\_count, dislikes with each other. We can clearly observe that channels had the maximum no. of likes and comments in 2018.



**Fig21. Time Variation of Views**

* This graph represents the variation of Views with the month-year. It can be observed that the views were highest in 2018 in march. It can also be observed that the ‘views’ column follows an increasing trend .



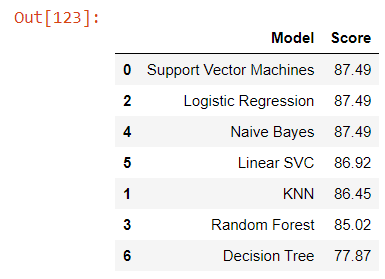
**Fig22. Time Variation of videoerror, comments, ratings**

* This graph represents the variation or comparison of comments disabled, video error, ratings disabled with each other. We can clearly observe that the maximum channels had their ratings disabled and comments disabled in 2018.

**PREDICTION**

After analysing all the graphs we came to a conclusion that it would be best to predict whether a video will come up on trending or not. For this prediction we selected columns such as 'dislike\_percentage', 'ldratio','title','trending\_date', 'video\_id','channel\_title','category\_id','publish\_time','tags','description','trending','publish\_date','days\_to\_trending'.

* First we imported all models and predicted the ‘trending’ column. The results with various models is as shown below:



* Since SVM, naïve bayes has the highest prediction accuracy so we decided to go ahead with SVM.
* Our prediction accuracy is 87.49.

**NOVELTY:**

Unlike famous videos, which would have already gained high viewership by the time they are declared popular, YouTube trending videos represent content that aims at viewers focus over a relatively short period of time, and has the potential of becoming more popular. Despite their prominence, YouTube trending videos have not been studied, analyzed or researched thoroughly in the past.

In this project, we present our key findings for measuring, analyzing, and comparing key aspects of YouTube trending videos. Our study is based on collecting and monitoring videos of great Britain and related statistics of more than 35,000 YouTube videos over an aggregate period of twenty four months.

**REFERENCES**

* ***Cha, M, Kwak, H, Rodriguez, P. (2009)****Analyzing the video popularity characteristics of large-scale user generated content systems. Transaction on Networking 17(5): 1357–1370. New York: ACM*
* ***Cheng, X (2013)****Understanding the characteristics of internet short video sharing: a youtube-based measurement study. Transactions on Multimedia 15(3): 1184–1194.*
* ***Figueiredo, F, Benevenuto, F, Almeida, JM (2011)****The Tube over Time: Characterizing Popularity Growth of YouTube Videos. In: Proceedings of the fourth ACM international conference on Web search and data mining, Hong Kong, China, 9–12 February 2011, pp.745–754. New York: ACM.*
* ***Welbourne, DJ, Grant, WJ (2016)****Science communication on YouTube: Factors that affect channel and video popularity. Public Understanding of Science 25(6): 706–718*
* ***Borghol, Y, Mitra, S, Ardon, S. (2011)****Characterizing and modelling popularity of user-generated videos. Performance Evaluation 68(11): 1037–1055.*