US Trending YouTube Video Statistics

Text Mining & Exploratory Data Analysis

Lixin Xiong Jingjing Huang

Yajing Zhou Kanchan Gyani

Introduction

The project ‘Text Mining and EDA of US YouTube Trending Video Statistics’ is designed to help us uncover the insights behind these trending videos. The data is scraped from YouTube API with the most relevant information on the trending videos. According to Variety magazine, “To determine the year’s top-trending videos, YouTube uses a combination of factors including measuring users interactions (number of views, shares, comments, and likes).”

Questions to Be Explored

With this the U.S. Youtube video dataset, we are going to use analytical techniques to figure out the following questions:

a. Are there correlations between the number of views, number of likes, number of dislikes, and number of comments?

b. What factors affect how popular a YouTube video will be? (what is the trending of the YouTube video?)

c. Does publish time affect the performance of a video?

d. What are other insights from the text analysis of string variables like titles, descriptions and tags?

Data Cleaning

- As we read data using read.csv, we added na.string argument to replace the missing values ("","NA","N/A"," ","[none]") with 'NA' for future convenience.
- Then we checked the type of each column and if any column has ‘NA’ in it.
- With the data type of raw data columns, we fix the data types. We converted the columns of ‘title’, ‘channel_title’, ‘description’ and ‘tags’ from Factor type to a Character type.
- For all ‘NA’ in the ‘tags’ and ‘description’ columns, we replaced them with ‘ ’.
- The next step is separating publish_time to publish_date and publish_hour.
- Finally, we examined the data again to make sure everything was ready for our future analysis, and then exported the cleaned dataset and saved it as "USvideos_cleaned.csv".

	DataType	isNA	NumOfNAs	factorTF	numlevels
video_id	factor	FALSE	0	TRUE	6351
trending_date	factor	FALSE	0	TRUE	205
title	character	FALSE	0	FALSE	0
channel_title	character	FALSE	0	FALSE	0
category_id	integer	FALSE	0	FALSE	0
tags	character	FALSE	0	FALSE	0
views	integer	FALSE	0	FALSE	0
likes	integer	FALSE	0	FALSE	0
dislikes	integer	FALSE	0	FALSE	0
comment_count	integer	FALSE	0	FALSE	0
thumbnail_link	factor	FALSE	0	TRUE	6352
comments_disabled	factor	FALSE	0	TRUE	2
ratings_disabled	factor	FALSE	0	TRUE	2
video_error_or_removed	factor	FALSE	0	TRUE	2
description	character	FALSE	0	FALSE	0
publish_date	character	FALSE	0	FALSE	0
publish_hour	character	FALSE	0	FALSE	0

Analytical Techniques

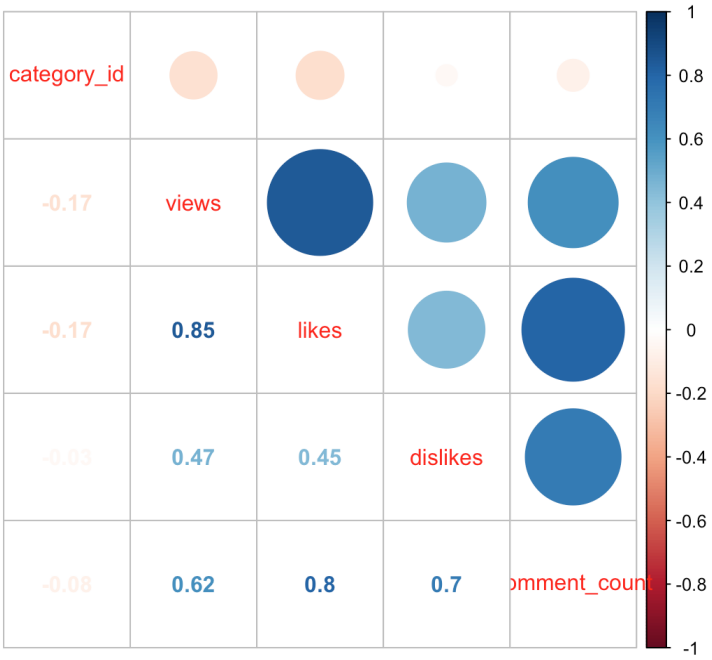
To solve the question above, we have used exploratory data analysis, text mining, and clustering analysis to process the dataset and tried to obtain insights.

Firstly, we used exploratory data analysis (EDA) for all questions. We used EDA with visualization because it helps us to understand the relationship between key variables and visually expose the video trends. For example, with a histogram, you could easily see which video's category has the most average view for a video.

Then we have used text mining, including sentiment analysis, to treat our descriptive variables like title and description. For each video, there are many text contents describing the video, which has substantial meaning for YouTubers. For example, an accurate video description helps a video to gain more views because the YouTube algorithm will recommend the video to an audience when the audience is searching or watching the content relating to the video's description or tags. In addition, Since the descriptive data are unstructured, text mining makes it accessible by extracting useful information from content and showing the patterns, trends, and insights of the words used.

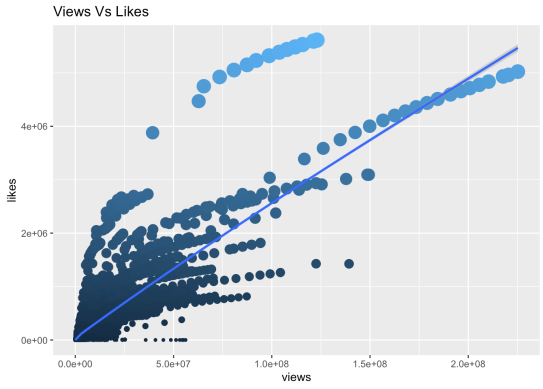
We also tried to use clustering to find some insights from the dataset. Grouping videos may help us to better understand the performance of videos.

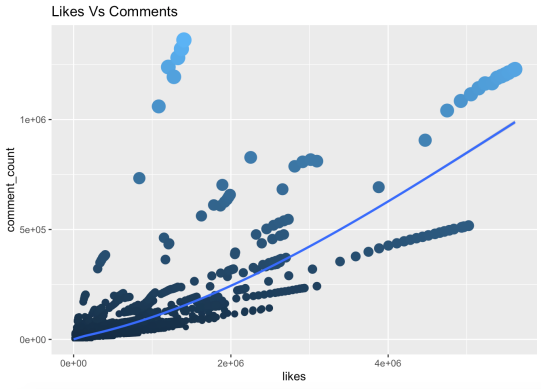
Analysis

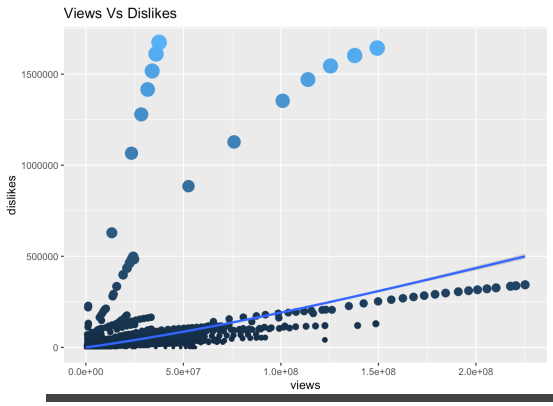


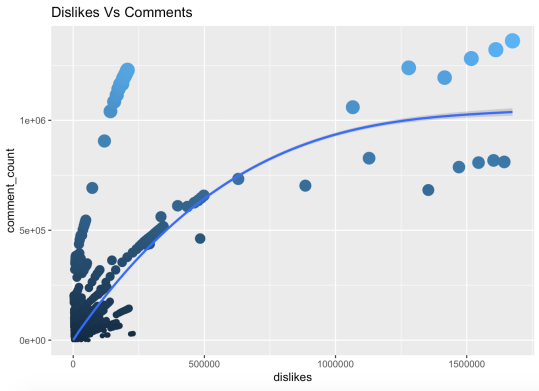
Correlation

The correlation map shows the relations among variables ‘views’, ‘likes’, ‘dislike’, and ‘comment_count’. We can see that there is a high correlation between views and likes, likes and comment_count. There is also a high correlation between dislikes and comment_count. So we decide to move forward and explore their relationships.



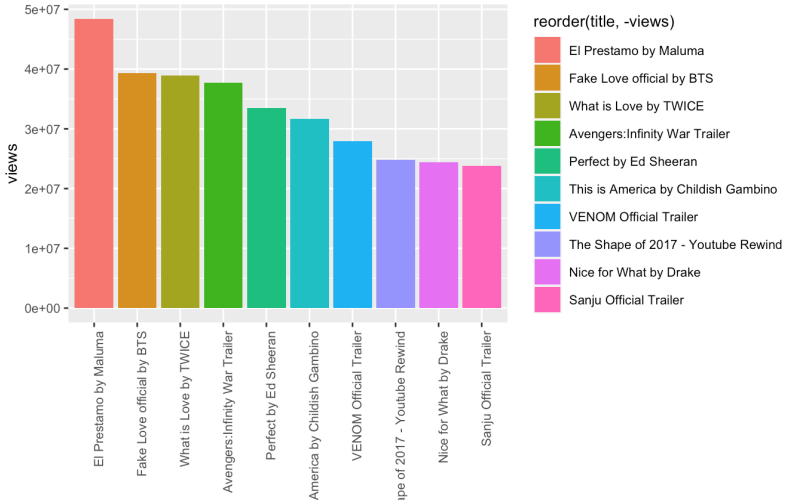


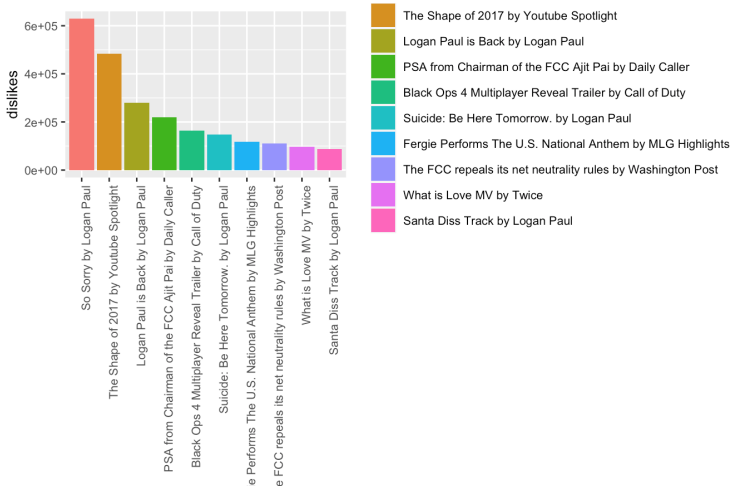


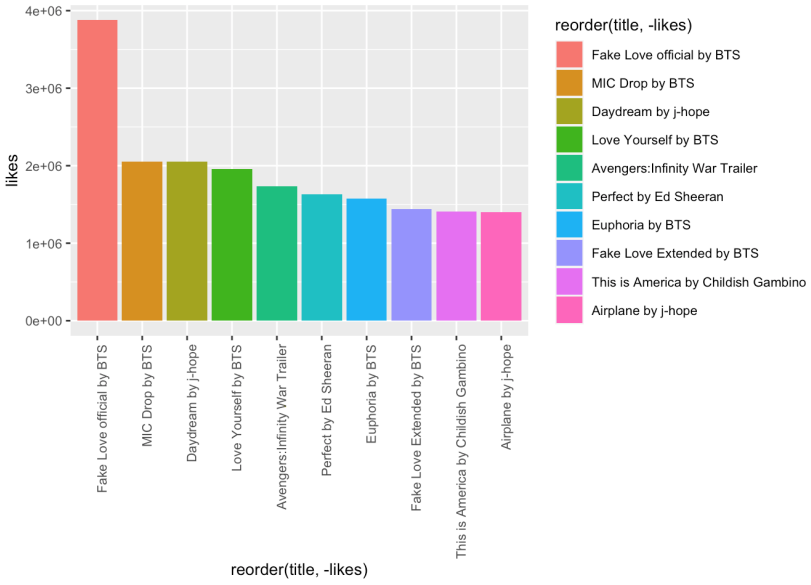


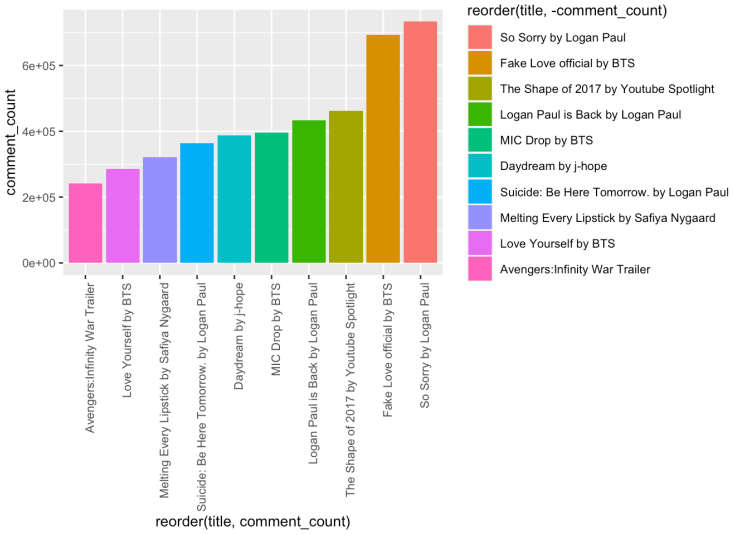
We find out that views and likes, views and dislikes, likes and comment_count, dislike and comment_count all have positive associations. This is saying that videos with higher views tend to have more likes or more dislikes, and videos with more likes or dislikes tend to have more comments. However, once the dislikes are over 1250000, the comments tend to increase slowly.

Top 10 most viewed, most liked, most disliked, and most commented videos









Top 10 most viewed videos

"El Prestamo by Maluma"
"Fake Love official by BTS"
"What is Love by TWICE"
"Avengers:Infinity War Trailer"
"Perfect by Ed Sheeran"
"This is America by Childish Gambino"
"VENOM Official Trailer"
"The Shape of 2017 - Youtube Rewind"
"Nice for What by Drake"
"Sanju Official Trailer"

Top 10 most disliked videos

"So Sorry by Logan Paul"
"The Shape of 2017 by Youtube Spotlight"
"Logan Paul is Back by Logan Paul"
"PSA from Chairman of the FCC Ajit Pai by Daily Caller"
"Black Ops 4 Multiplayer Reveal Trailer by Call of Duty"
"Suicide: Be Here Tomorrow. by Logan Paul"
"Fergie Performs The U.S. National Anthem by MLG Highlights"
"The FCC repeals its net neutrality rules by Washington Post"
"What is Love MV by Twice"
"Santa Diss Track by Logan Paul"

Top 10 most liked videos

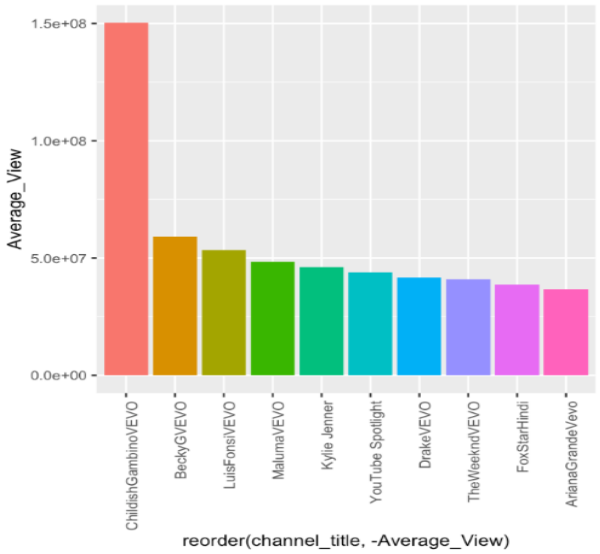
"Fake Love official by BTS"
"MIC Drop by BTS"
"Daydream by j-hope"
"Love Yourself by BTS"
"Avengers:Infinity War Trailer"
"Perfect by Ed Sheeran"
"Euphoria by BTS"
"Fake Love Extended by BTS"
"This is America by Childish Gambino"
"Airplane by j-hope"

Top 10 most commented videos

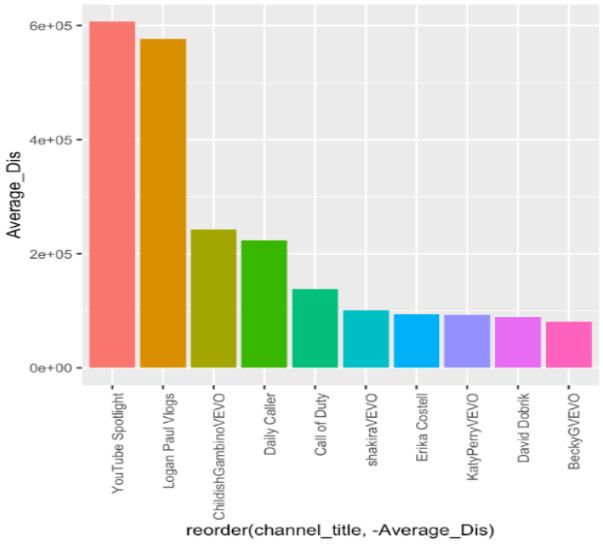
"So Sorry by Logan Paul"
"Fake Love official by BTS"
"The Shape of 2017 by Youtube Spotlight"
"Logan Paul is Back by Logan Paul"
"MIC Drop by BTS"
"Daydream by j-hope"
"Suicide: Be Here Tomorrow. by Logan Paul"
"Melting Every Lipstick From Sephora Together by Safiya Nygaard"
"Love Yourself by BTS"
"Avengers:Infinity War Trailer"

These results also proved the relationships between variables. For example, So Sorry by Logan Paul is the most disliked video and has the highest comment counts, and Fake Love by BTS is the most liked video and has the second highest comment counts.

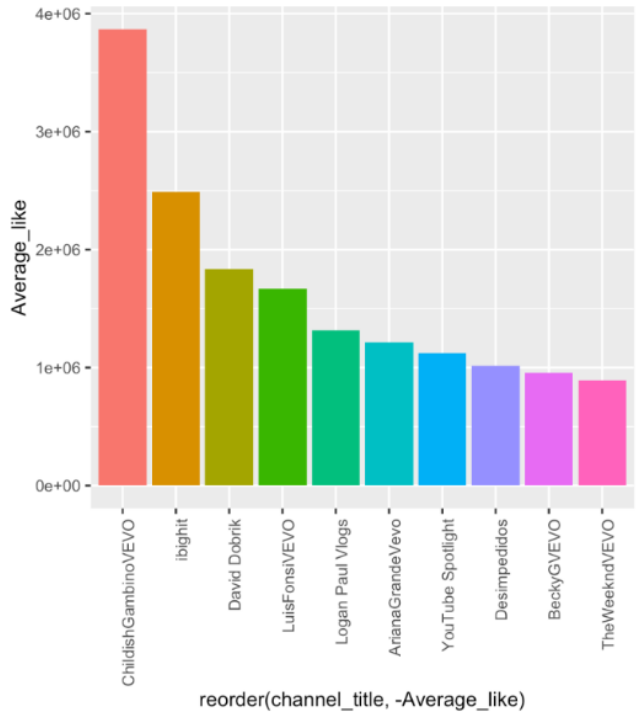
Top 10 Trending Channels by Average Number of Views, Number of Dislikes, Number of Likes and Number of Comments

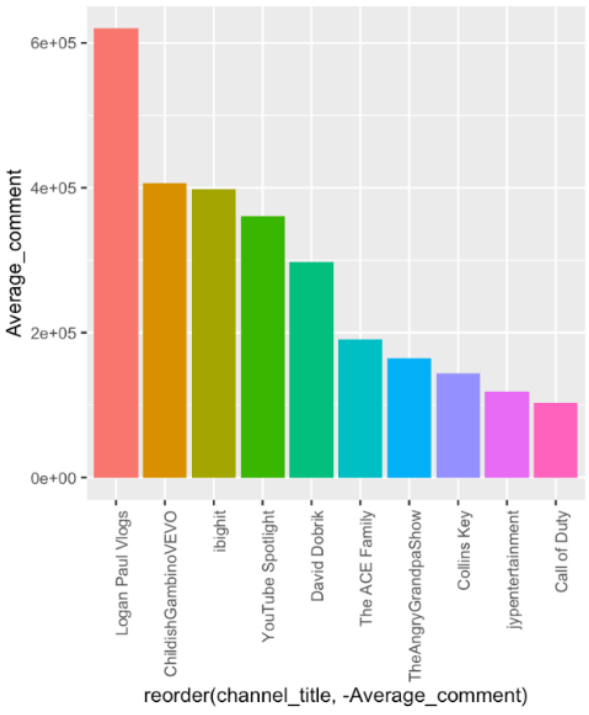


Histogram of Channels with Top 10 Average View



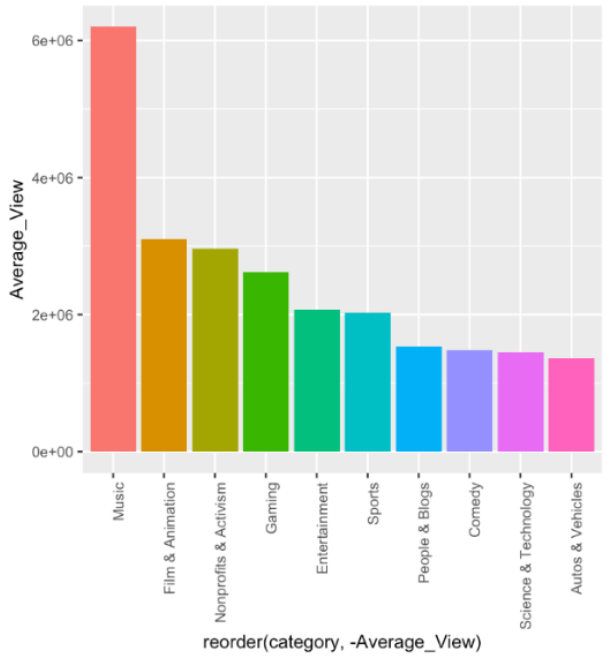
Histogram of Channels with Top 10 Average Dislike



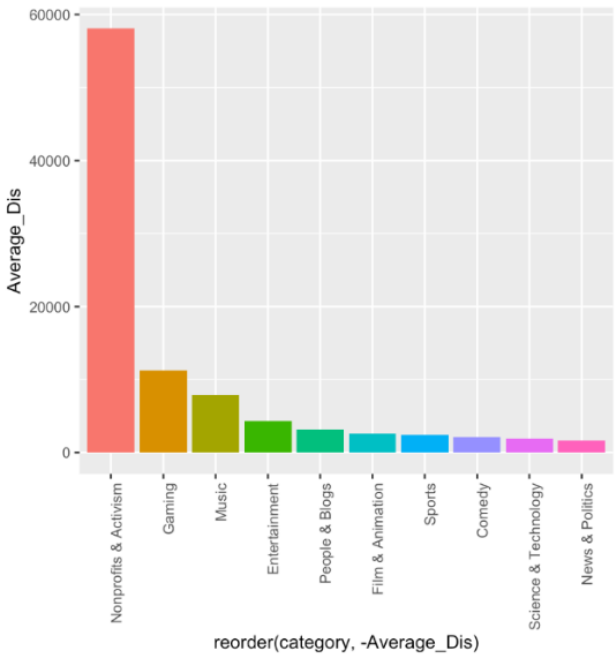


From the histograms above, we could see that whatever the select feature is, most of the top 10 trending channels are related to music.

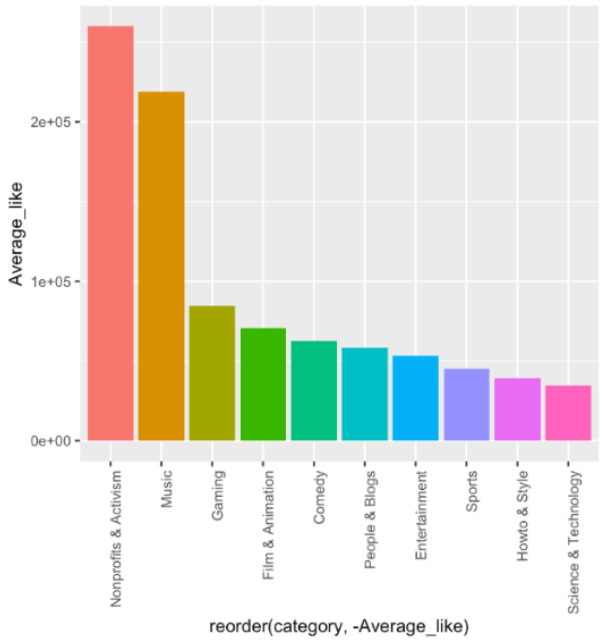
Top 10 Trending Video Categories by Average Number of Views, Number of Dislikes, Number of Likes and Number of Comments



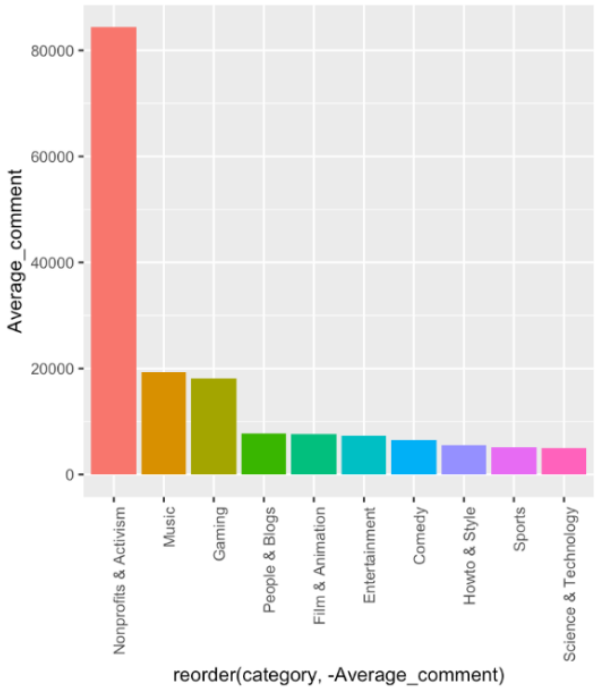
Histogram of Categories with Top 10 Average View



Histogram of Categories with Top 10 Average Dislike



Histogram of Categories with Top 10 Average Like



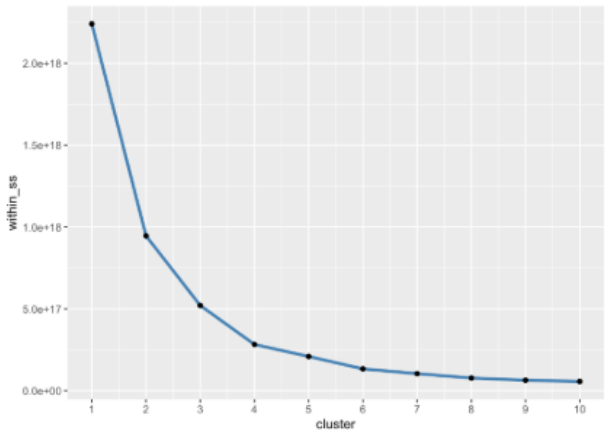
Histogram of Categories with Top 10 Average Comment Count

It seems that ‘Music’, ‘Nonprofit & Activism’, ‘Gaming’, ’Entertainment’, ‘Sports’, ‘People & Blogs’, ‘Comedy’ and ‘Science & Technology’ are shown on the all of four histograms. Therefore, we may consider these categories as top trending YouTube video categories. We noticed that most of these top trending categories are related to recreation. It may mean that the majority audience use YouTube for recreation at the most of time. Additionally, among these categories, ‘Nonprofit & Activism’ video is most disputable because this type of video has the most number of dislikes, likes and comments even though it does not have the most number of views.

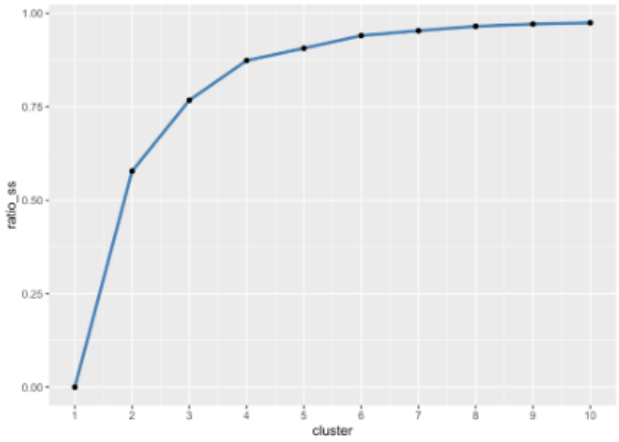
Clustering Analysis for Video Trending

We tried to use clustering analysis to obtain some insights. With this large dataset, we decided to use k-mean because we do not want a very expensive process of hierarchical clustering.

We clustered the observations based on the all numeric variables we have --- Average Number of Views, Number of Dislikes, Number of Likes and Number of Comments. And we decided to have 3 clusters because the total within sum of square and ratio plots support a two- and three- cluster solution and we think three- cluster solution makes more sense.

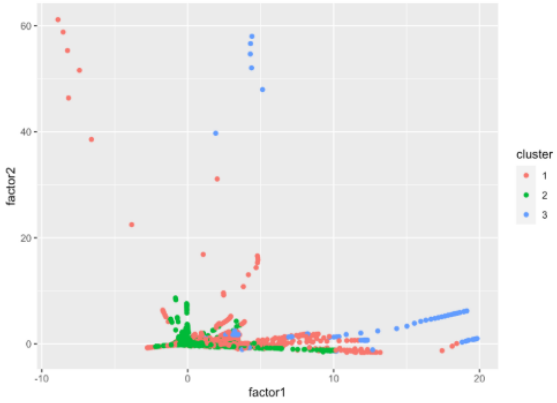


The Total within Sum of Square Plot



Ratio Plot

Using K-mean clustering, we got the following results:



2-D Cluster Plot

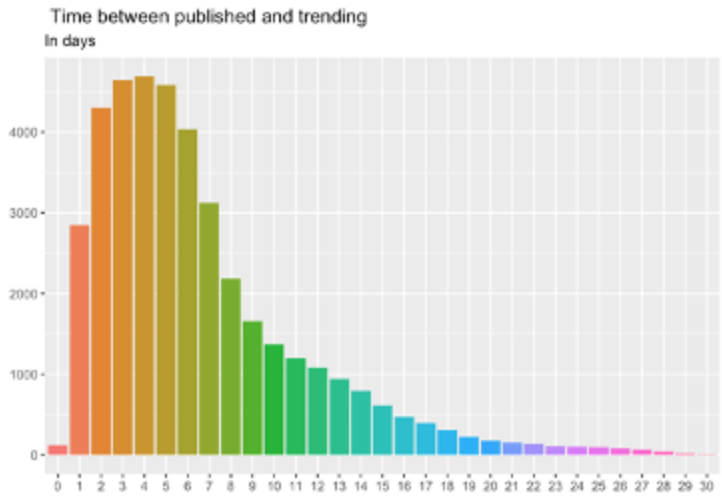
k_segments	views	likes	dislikes	comment_count
1	23960331	651853.38	36387.74	68908.13
2	1368381	47813.78	2057.70	5490.26
3	108444347	2875072.48	228910.44	381255.40

Cluster Table

We think the ratio of The number of likes v.s. The number of dislikes is low in cluster 3 and all videos in the cluster 3 belong to ‘Nonprofit & Activism’. The reason behind it might be that ‘Nonprofits & Activism’ usually launches campaigns with paid promotions, So the data is consistent within the cluster 3.

The Time Between Published and Trending

In order to examine how long it will take for a video to become trending after it was published, we first manipulate the time data to calculate the number of days it took for a



video to become trending. Then we visualize it using the histogram to have a clearer view. A video rarely becomes trending on the day it was published, 2-6 days is the golden time.

Most Popular Tags

As you can see from the word cloud, most frequently used tags for trending videos are video, funny, music etc. So for video uploaders they can leverage these tags to increase the visibility of their video.

Text Mining

The variables ‘title’ and ‘description’ from the chosen dataset were used to implement text mining along with certain other variables and a heatmap towards the end with video publish date and time to identify if there exists a relation.

The plot for ‘Top 10 words’ in the title shows that most of the ‘Official’ youtube videos are the trending videos. It also shows that 2018 has more trending videos than 2017

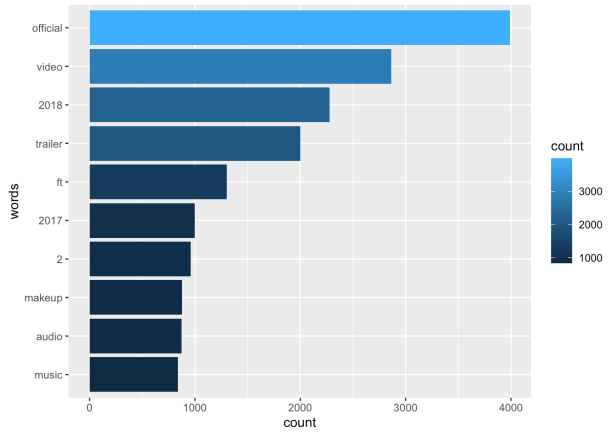


Figure 1

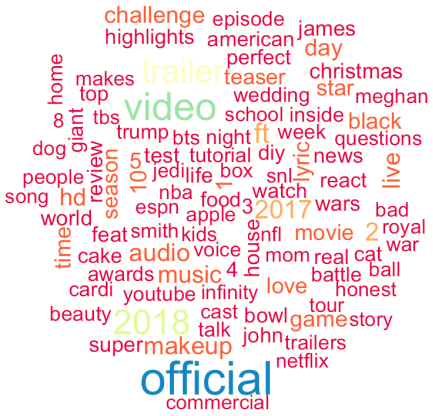


Figure 2

A ‘bing’ and NRC sentiment analysis on ‘Title’. The sentiment analysis with NRC lexicon on the video’s title shows that topmost emotions displayed by the dataset are - positive, trust, negative, joy and anticipation

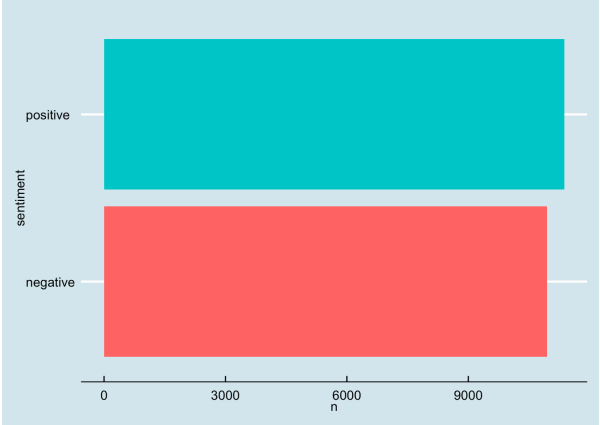


Figure 3 - ‘bing’ sentiment analysis

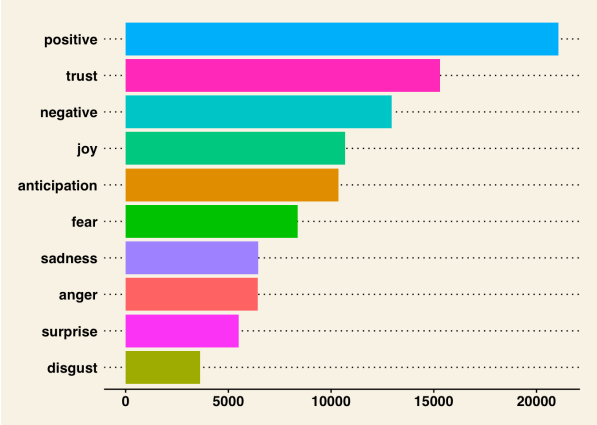


Figure 4 - NRC Lexicon

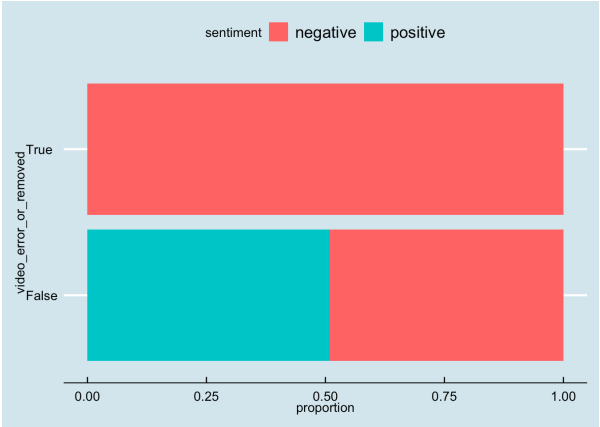


Figure 5 - ‘bing’ on title with ‘video_error_or_removed’

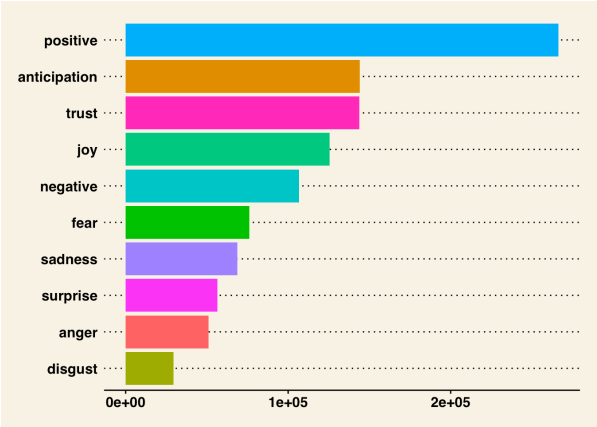


Figure 6 - NRC Lexicon on ‘description’

The videos with negative emotions were flagged to be removed or to display an error.

An NRC Lexicon use on the video’s description shows positive, anticipation, trust and joy

A ‘bing’ sentiment analysis on the video’s title shows love, perfect, beauty, awards as the top positive words. The negative words do not stand out in proportion to the positive ones (Figure 7)

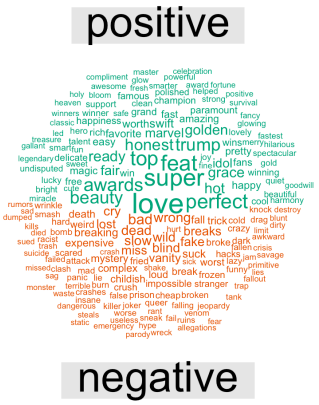


Figure 7

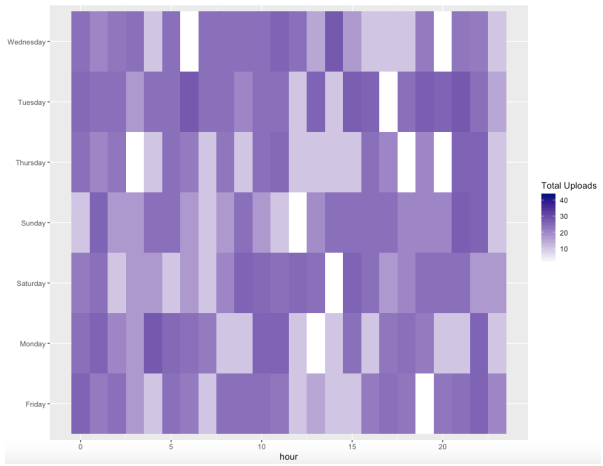


Figure 8

A heatmap was plotted to identify a relationship(if any) between the day and the hour videos were uploaded. The graph seems to spread out between 12 am - 11 pm although there are few uploads between 11 pm-12 am looking at the graph vertically (Figure 8)

Conclusion

There are positive correlations between views & likes, views & dislikes, likes & comment counts, dislikes & comment counts. This illustrated that views are the most important factor of trending videos. No matter how people react to the videos (liked or disliked), higher views would eventually bring more comments and make the video become trending.

The trending videos from 2017 and 2018 show both positive and negative emotions. So a video with negative emotion can still be trending. Most of the videos belonged to official channels or were part of the entertainment category and uploaded official trailers to Youtube. The comments section for most of the trending positive videos was disabled and those with negative sentiment were flagged as removed. The upload time is fairly spread out throughout the weekdays and hours in a day.