Exploratory Data Analysis of YouTube Trending Videos (2020-21)

Ramesh Chandra Vuppala, Srihari Das S G, Animesh Kumar, Ronak Dedhiya Dinesh

Email: {rameshcv , sriharidas, animeshkumar, ronakdedhiya} @iisc.ac.in

What are YouTube Trending Videos?

YouTube is the most popular and most used video platform in the world. YouTube maintains a list of trending videos. Description of trending videos as shared by google. *"Trending helps viewers see what's happening on YouTube and in the world. Trending aims to surface videos that a wide range of viewers would find interesting. Some trends are predictable, like a new song from a popular artist or a new movie trailer. The list of trending videos is updated roughly every 15 minutes".*

Goals and Motivation

We are presenting **exploratory data analysis of YouTube trending videos** of US during 2020-21. We are trying to get key insights from trending videos and find interesting facts and patterns by exploring the data and by using effective visualizations. Titles, descriptions, thumbnails, tags, views, likes/dislikes, and comments were analyzed to produce intuitive results as shown in document. All trending videos for the period of August 2020- October 2021 were analyzed (More than 90,000 videos).

We are also comparing the similarities between trending videos of multiple countries (US vs Great Britain vs India) with respect to videos and distribution across different video categories.

We are also presenting a **video recommendation** method based on current watching trending video. One may think that having many options is a good thing, as opposed to having very few, but an excess of options can lead to what is known as a "*decision paralysis*". Recommendation systems make life simpler. Therefore, recommender systems have become a crucial component in platforms, in which users have a myriad range of options available. Their success will heavily depend on their ability to narrow down the set of options available, making it easier for us to make a choice.

Technical Problem Being Solved

Above Exploratory data analysis helps people to understand which category of videos people are interested in country wise. It also helps people to estimate when video can become trending after publishing and how much advertisement needed to make it trend by showing correlation between various parameters of the video. It helps people to prepare TAGS with respect to number of words and usage of words to make video trending faster. Proposed solution also helps people to get recommendations based on current watching video, category and video TAGS.

Architecture Details

Tools Used

This analysis was performed using Python and a powerful group of Python libraries including Spark Data Frames, Spark ML, Pandas, Matplotlib, Seaborn, Word Cloud. The analysis was performed in a Google Colab Notebook.

Data Source

The data used in this analysis was retrieved from freely available data from Kaggle website. https://www.kaggle.com/rsrishav/youtube-trending-video-dataset/version/430.

Github: https://github.com/RonakDedhiya/YouTube_Trending_Video_Spark_Data_Analysis

Data Size and Description

In this Analysis, Majorly Trending Videos from US is considered. We have a total of ~**92K** video data with video trending multiple times in the duration from Aug 2020 – Nov 2021 (15 months). Amongst them we have ~**16K** unique videos. We are also using Trending Video of Great Britain and India for the same period to compare distribution across different video categories.

The following table shows an **example of the data that we have for each video**:

video_id	Q7z1Ny1kc6g
title	Super Smash Bros. Ultimate - Take Your Heart - Nintendo Switch
publishedAt	2018-12-07T04:33:52.000Z
channelTitle	Super Smash Bros.
trending_date	2019-01-01 00:00:00
tags	Nintendo|Switch|Video Game (Industry)|Smash|Bros|Smash Bros|SSBU|Game| VideoGame|sv_nintendo|ゲーム|ビデオゲーム|任天堂|ニンテンドー|スイッチ|Nintendo Direct|Direct|ダイ レクト|大乱闘|スマッシュブラザーズ|スマブラ|スペシャル|P5|ペルソナ5|PERSONA5|Jack|ジョー カー|Arseneアルセーヌ|ATLUS|アトラス|SEAG|セガ|JOKER
view_count	1,309,682
likes	64503
dislikes	1505
comment_count	0
thumbnail_link	https://i.ytimg.com/vi/Q7z1Ny1kc6g/default.jpg
comments_disabled	TRUE
ratings_disabled	FALSE
description	https://www.smashbros.com/ESRB EVERYONE 10+PEGI 12This game is classified : PG – Mild themes and violence, Online interactivity
category	Gaming

Video_id – Unique ID of Video.

Title – Name of Video.

PublishedAt – Date and time of uploading the video.

ChannelTitle – Name of Channel uploaded the video.

Trending date – Date when video became trending, there can be multiple days same video can trend. So, we can find the same video on different days with different number of views, comments, likes, etc.

Tags – Tags used for the video.

We have likes, dislikes, views, comments, category of the video along with few more parameters.

Qualitative and Quantitative Evaluation

EDA of trending video metrics:

Views/Comments/Likes for videos when they first became trending?

- "BTS 'Butter' Official MV" hosted by "HYBE LABELS" was the most liked (7.1M), and commented (3.4M) video.
- "LISA - 'LALISA' M/V " by BLACKPINK was the most viewed (85.9M) video.
- "BLACKPINK - 'Ice Cream (with Selena Gomez)' M/V" has highest dislikes (4,05,329).
- Avg. number of Likes observed for trending first time 97666 (~100 K)
- Avg. number of Views observed for trending first time 1344952 (~1.34 M)
- Avg. number of Comments observed for trending first time 9581 (~10 K)
- "Leading the Charge | Circle K" has only 17 likes (minimum), 6 dislikes (minimum) and 10 comments when it became first trending.
- 43 Videos have 0 comments and 11 Videos have 0 views when they first became trending.
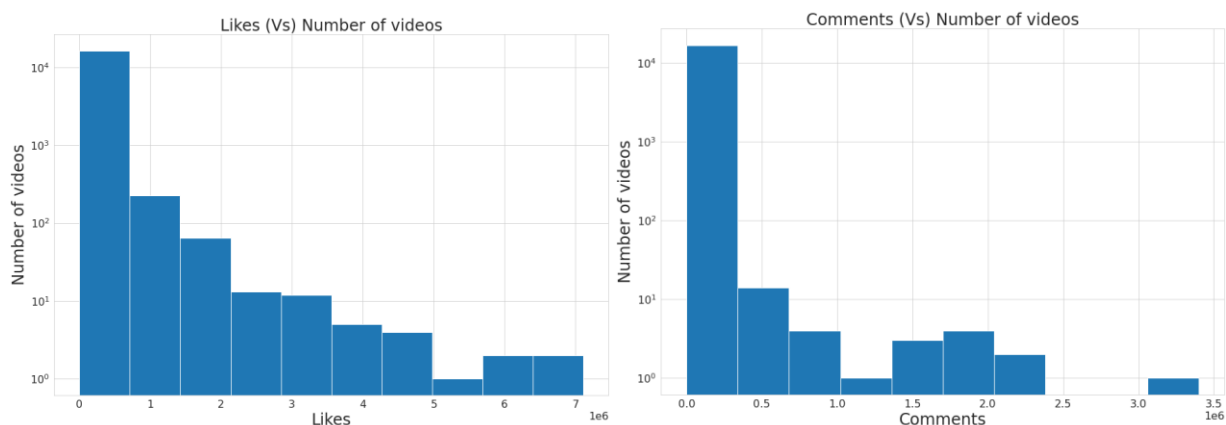
How long does it take a Video to become trending for the first time?

- Maximum it took 27 days (about 4 weeks) for a video to trend from its publishing date. (3 videos : All are from Walmart)
- 1087 videos (6.6%) Started trending on same day as Published.

Videos from which category took longer time to trend?

- Trending videos from Category ID 22 ("People & Blogs") took longer time to trend on an average compared to other categories.

Days Taken to Trend (Vs) Number of videos

Category (Vs) Avg. Days taken to trend
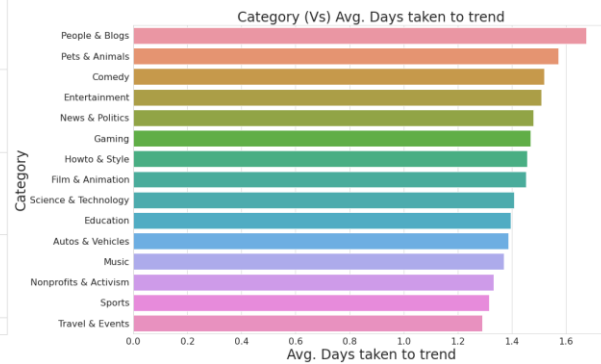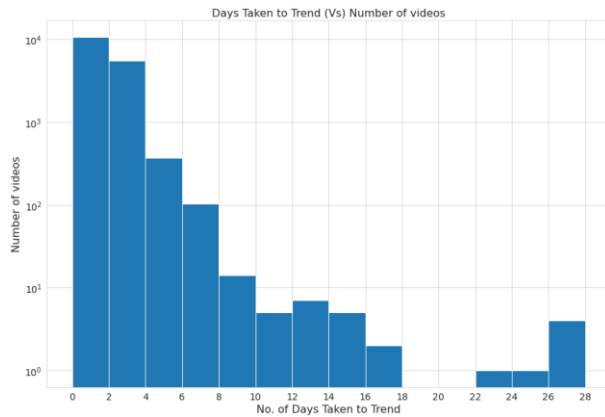
What percentage of trending Videos have more Dislikes than Likes? (Negative publicity inference)

- 59 out of 16455 I.e. Only 0.36 % Videos have more Dislikes than likes when they first became trending.
- 363 out of 91791 I.e. Only 0.395 of total Videos have more Dislikes than likes in total.

Which category of Video becomes most trending?

- OfCourse "Entertainment" Category (24) becomes more trending with 18373 trending videos count
- "Nonprofits & Activism" Category (29) stands least with 88 trending videos count

Catogories of trend videos in US

Catogories Vs Likes of trend videos in US
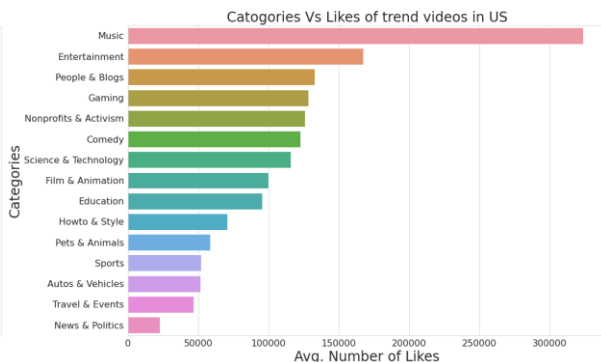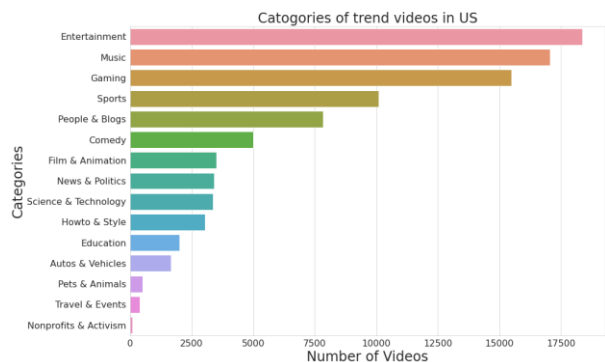
## Users like videos from which CATEGORY the most?

- "Music" Category has highest avg of likes (~324052) and "News & Politics" with least avg of likes (22934)
- "Music" Category has highest avg of dislikes also (~6068) and "Pets & Animals" with least avg of dislikes (743)

Which channels published more trending videos?

- "NBA" channel has hosted the highest number of trending videos (578).
- 41 channels have only 1 video trending and only on one day.

How many videos appeared trending most of the days?

- As we see the peak from histogram below, most of videos appeared trending for 4-8 days.
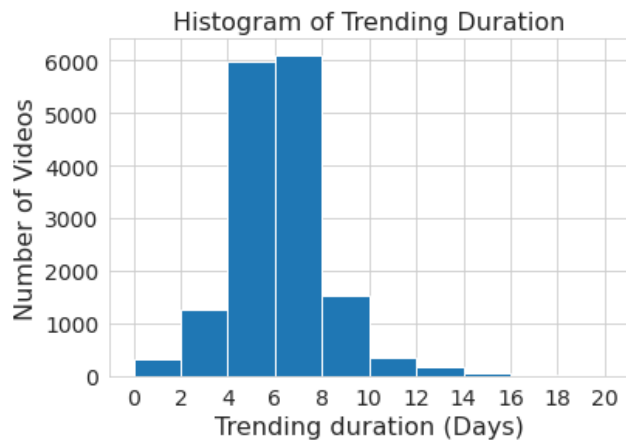
Video Recommendation based on Current Watching Video:

List of Various Categories in US Trending Video Data set:

Film & Animation	Entertainment	Documentary
Autos & Vehicles	News & Politics	Drama
Music	Howto & Style	Family
Pets & Animals	Education	Foreign
Sports	Science & Technology	Horror
Short Movies	Nonprofits & Activism	Sci-Fi/Fantasy
Travel & Events	Movies	Thriller
Gaming	Anime/Animation	Shorts
Videoblogging	Action/Adventure	Shows
People & Blogs	Classics	Trailers
Comedy	Comedy	

Steps Followed to build recommender system:

1. TAGS and CATEGORY are converted to Upper Case.
2. Space is removed in TAGS.
3. Category and First 3 TAGS for each video are selected.
4. Unique monotonically increasing ID is assigned for each Video.
5. Vector is generated for each video using TF-IDF (Term Frequency - Inverse Document Frequency) vectorizer.
6. Similarity between each Video is calculated using dot product.

A recommendation system is built for 1000 videos as finding similarity between videos takes more time.

Output for Each Search is as below:

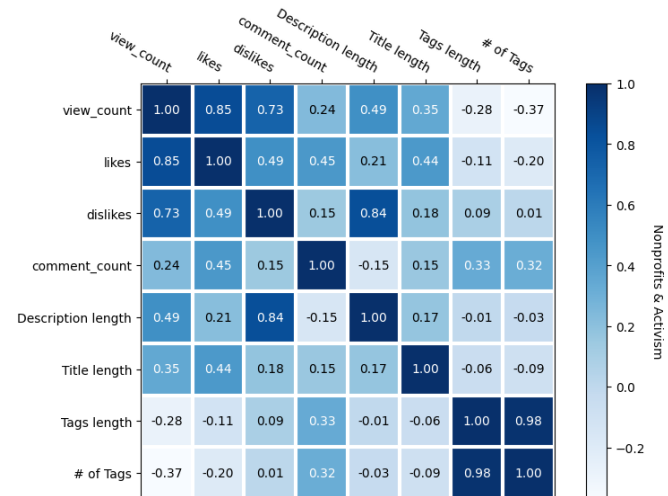|	Recommendation Title	Category_Tags
1	attempting to cook for the first time	[ENTERTAINMENT, ENTERTAINMENT, VLOG, FUNNY]
2	Sharon Osbourne Responds To Criticism Over Racism Debate	[ENTERTAINMENT, ENTERTAINMENT, NEWS, ETCANADA]
3	AGT Winner Darci Lynne Performs Baby by Justin Bieber - America's Got Talent 2020	[ENTERTAINMENT, ENTERTAINMENT, TVSERIES, HILARIOUSMUSICSONGS]
4	Friday Night Funkin' Mod Characters Reacts	Part 9	Moonlight Cactus		[ENTERTAINMENT, None, None, None]
5	Lil Baby - Behind The Scenes of "The Bigger Picture" at The 63rd GRAMMYs	[ENTERTAINMENT, None, None, None]
6	Bob Ross in a Random Outfit	[ENTERTAINMENT, None, None, None]
7	WE HAVE A NEW GROUP MEMBER!!	[ENTERTAINMENT, None, None, None]
8	*intense* room transformation	[ENTERTAINMENT, None, None, None]
9	I Gave People $1,000,000 But ONLY 1 Minute To Spend It!	[ENTERTAINMENT, None, None, None]
10	Seas #TeamSeas	[ENTERTAINMENT, None, None, None]

Correlation between various trending video metrics

What is the Correlation (Ratio) between Likes-Dislikes-Views-Comments in different categories?

Like - Dislike Ratio / View - Comment Ratio / Dislike - View Ratio / Like - View Ratio
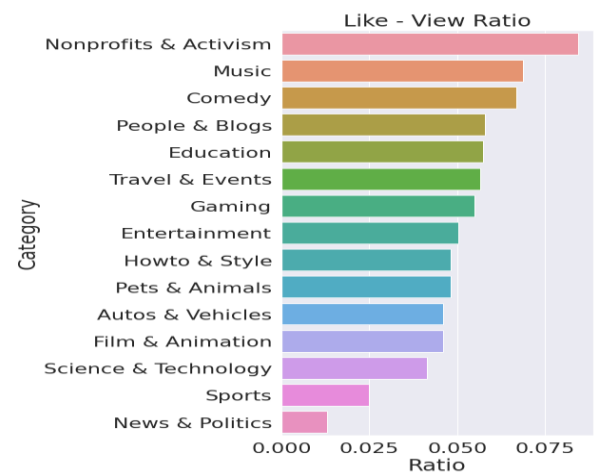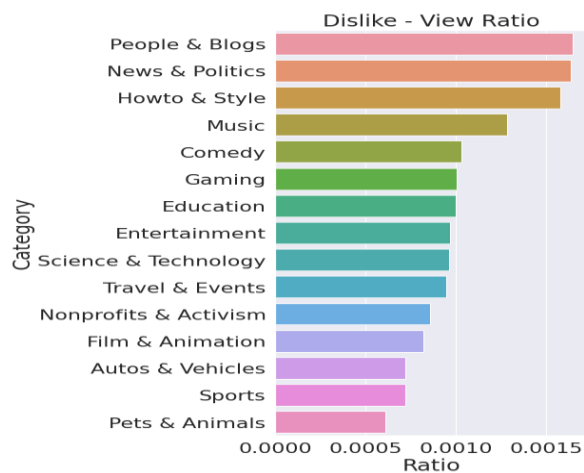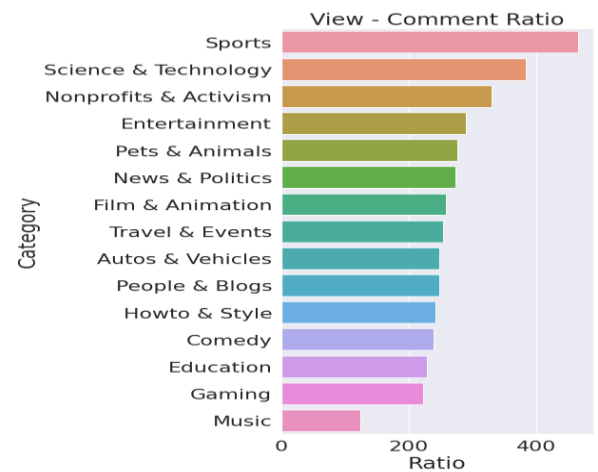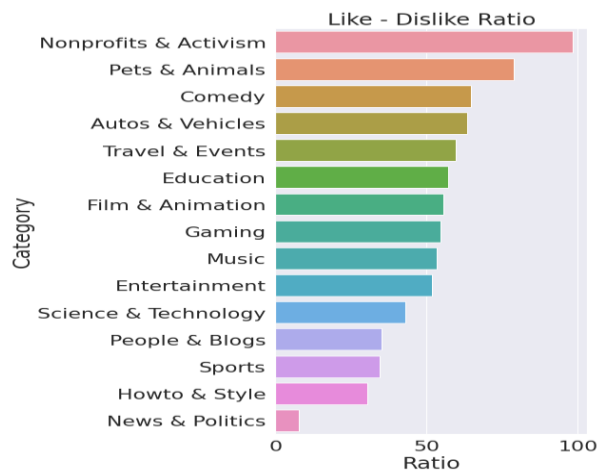
- News & Politics have lowest like-dislike ratio and view-comment ratio. People relatively dislike these videos and comment a lot. It also has the highest Dislike- view ratio.

- Pets & Animals videos & Non-profits Videos have highest likes-dislikes ratio. Not surprisingly, people find difficult to hate pets and animals and non-profit activity.

- Sports and Science Technology have largest view-comment ratio. People tend to comment less on these videos compared to music-related video

Visualization of correlation between different metrics through Heat map.

Entertainment

	view_count	likes	dislikes	comment_count	Description length	Title length	Tags length	# of Tags
view_count	1.00	0.90	0.92	0.31	-0.06	-0.10	-0.12	-0.11
likes	0.90	1.00	0.82	0.53	-0.09	-0.14	-0.16	-0.12
dislikes	0.92	0.82	1.00	0.23	-0.08	-0.10	-0.10	-0.08
comment_count	0.31	0.53	0.23	1.00	-0.02	-0.13	-0.10	-0.05
Description length	-0.06	-0.09	-0.08	-0.02	1.00	0.21	0.22	0.14
Title length	-0.10	-0.14	-0.10	-0.13	0.21	1.00	0.32	0.21
Tags length	-0.12	-0.16	-0.10	-0.10	0.22	0.32	1.00	0.88
# of Tags	-0.11	-0.12	-0.08	-0.05	0.14	0.21	0.88	1.00

Nonprofits & Activism

	view_count	likes	dislikes	comment_count	Description length	Title length	Tags length	# of Tags
view_count	1.00	0.85	0.73	0.24	0.49	0.35	-0.28	-0.37
likes	0.85	1.00	0.49	0.45	0.21	0.44	-0.11	-0.20
dislikes	0.73	0.49	1.00	0.15	0.84	0.18	0.09	0.01
comment_count	0.24	0.45	0.15	1.00	-0.15	0.15	0.33	0.32
Description length	0.49	0.21	0.84	-0.15	1.00	0.17	-0.01	-0.03
Title length	0.35	0.44	0.18	0.15	0.17	1.00	-0.06	-0.09
Tags length	-0.28	-0.11	0.09	0.33	-0.01	-0.06	1.00	0.98
# of Tags	-0.37	-0.20	0.01	0.32	-0.03	-0.09	0.98	1.00

- Videos belonging to entertainment category show a high correlation among views, likes, dislikes etc. It shows no correlation with description length, title length etc.
- However, for videos which are trending in category like Non-profits and activism, Description length plays an important role.
- 49% correlation is observed between views and description length. Also, 84% correlations are observed between description length and dislikes.
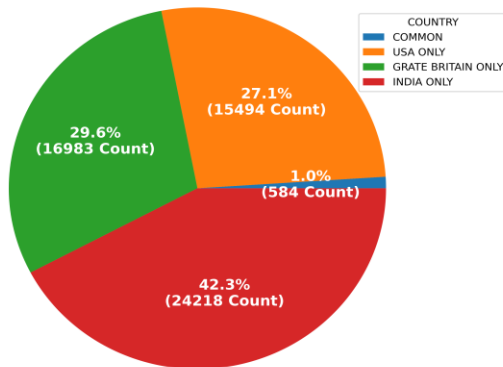
Correlation between Days of Publish to Trend (v/s) Trending Duration.

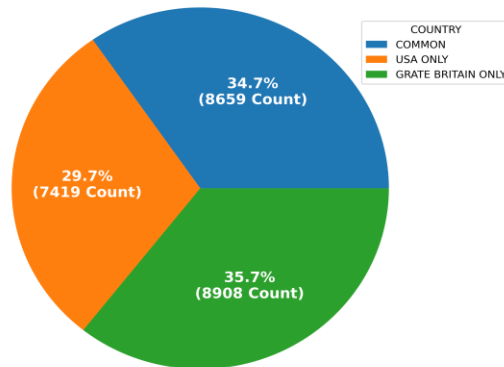Correlation between Days to Publish v/s Trend and Trending Duration
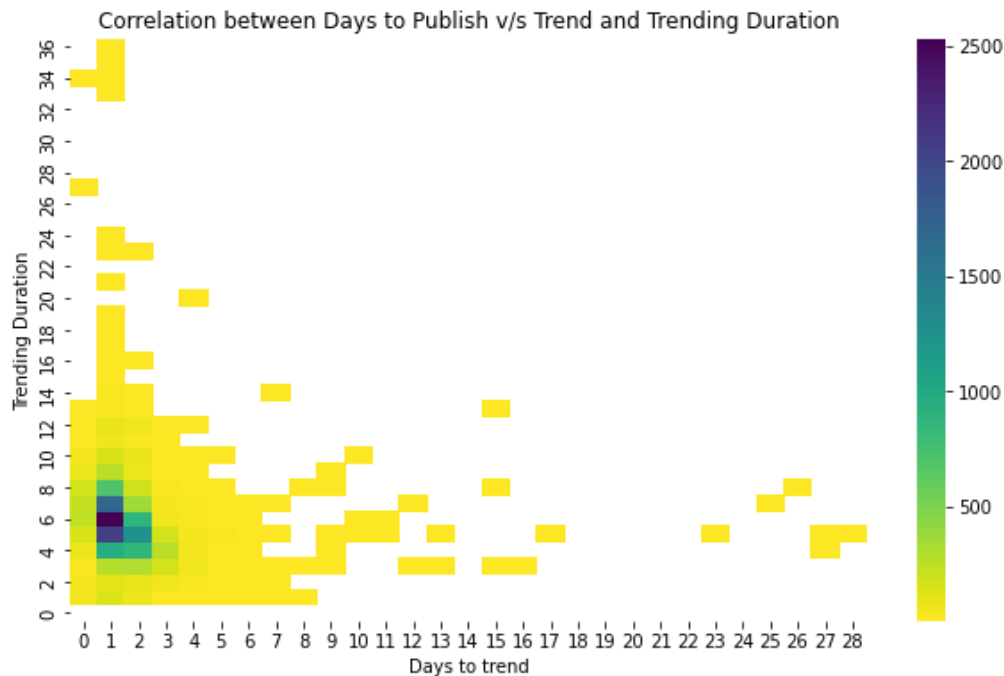
- The less days needed for a video from publishing to trend, the longer the trend duration.
- Videos that can get into trending within 3 days will have a higher probability of trending for a longer time.

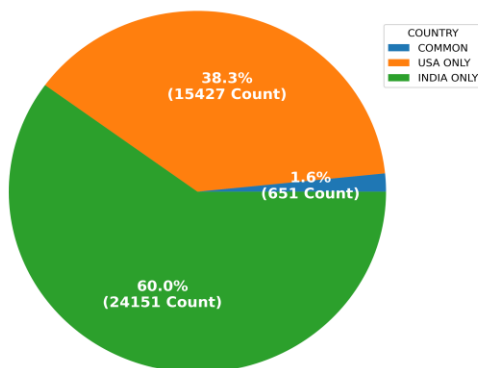## Does video trending in one country trend in other countries too?
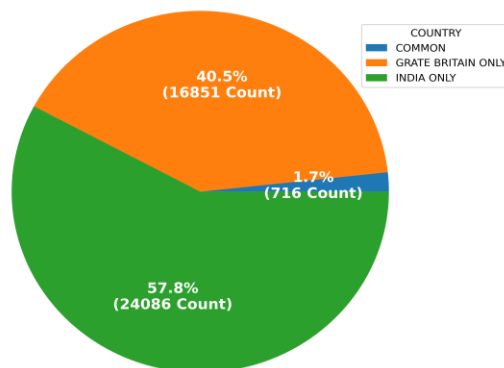
All Country Trending Video shares

COUNTRY
- COMMON
- USA ONLY
- GRATE BRITAIN ONLY
- INDIA ONLY

27.1%
(15494 Count)

29.6%
(16983 Count)

1.0%
(584 Count)

42.3%
(24218 Count)

USA & Great Britain Trending Video shares

COUNTRY
- COMMON
- USA ONLY
- GRATE BRITAIN ONLY

34.7%
(8659 Count)

29.7%
(7419 Count)

35.7%
(8908 Count)

USA & India Trending Video shares

COUNTRY
- COMMON
- USA ONLY
- INDIA ONLY

38.3%
(15427 Count)

1.6%
(651 Count)

60.0%
(24151 Count)

Great Britain & India Trending Video shares

COUNTRY
- COMMON
- GRATE BRITAIN ONLY
- INDIA ONLY

40.5%
(16851 Count)

1.7%
(716 Count)

57.8%
(24086 Count)

- 3 countries USA, Great Britain (GB) and India, trending videos are compared.
- 1% of trending videos found trending commonly in the 3 countries.
- 34.7% of trending videos found trending commonly in USA and Great Britain.
- 1.6% of trending videos found trending commonly in USA and India.
- 1.7% of trending videos found trending commonly in Great Britain and India.

Summary:
- India has highest number of trending videos, almost double from any other country.
- India have so many cultures, dialects, and language, so very few are trending in common with other 2 English speaking countries.
- 1/3rd of trending videos is in common between two English speaking Countries (USA and GB).

Comparison between the number of trending videos per category across 3 countries.
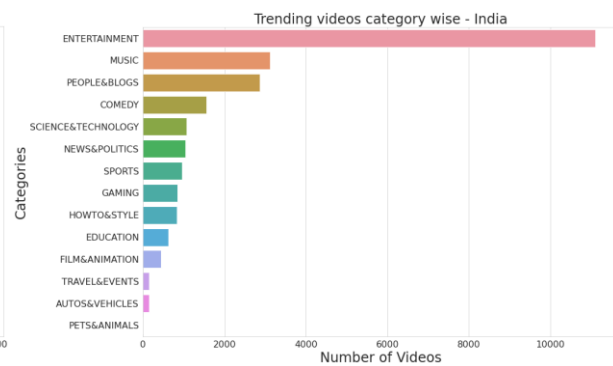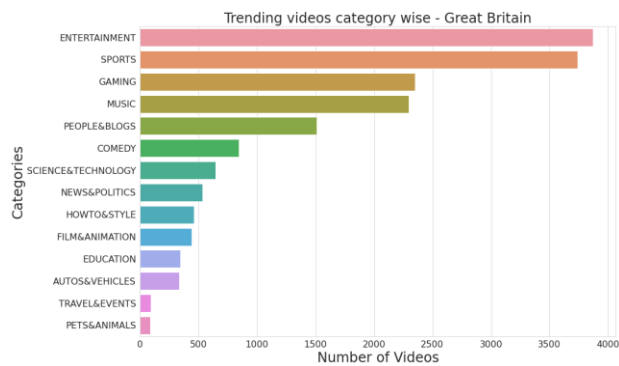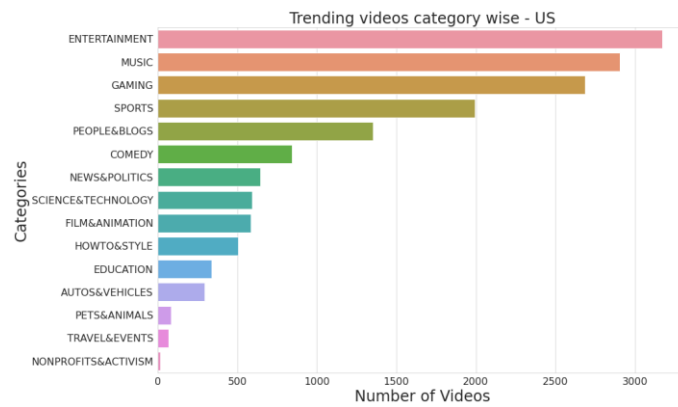
- In USA most trending categories are "Entertainment" closely followed by "Music" and "Gaming", almost 55% of total trending videos belongs to these 3 categories.
- In Great Britain most trending categories are "Entertainment" closely followed by "HowTo&Style" and then "Music" and "Gaming", almost 70% of total trending videos belongs to these 4 categories.
- In India most trending categories are "Entertainment" which is almost half of total trending video list.

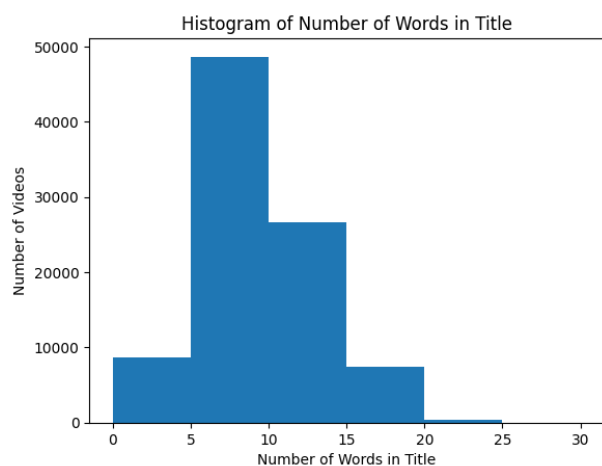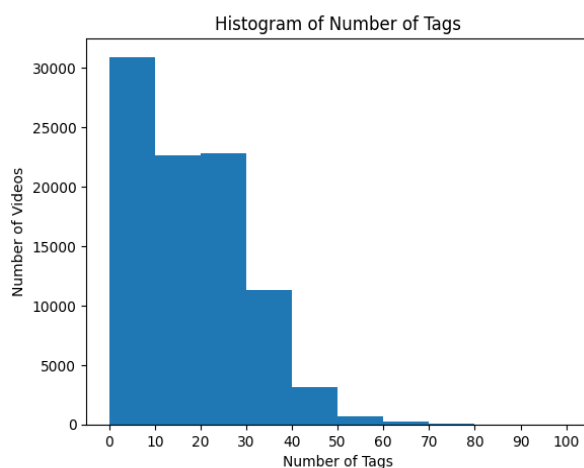## Recommendations to make Video trending:

Trending Videos TAGS and Titles

TAGS and Titles constitute an important part of each video. They describe the video for people before deciding to click on the video or not. And because of that, video TAGS and Titles are one of the crucial factors in video success; Here are some interesting facts about trending-videos TAGS and Titles.

What should be average number of TAGS? how many TAGS are there in most trending videos? What should be average number of words in video titles? how many words are there in the title of most trending videos?

Average Number of TAGS per video in data set is 16. Average Number of Words in title per video in data set is 8.

Distribution of Number of TAGS per video is as below.

What are the 50 most common Words used in TAGS with Videos greater than 1M views?

Are there some words that occur in trending video TAGS more than others? To get the answer, we analyzed the TAGS of all trending videos and counted the occurrences of each word in those TAGS. Before that, STOPWORDS were removed from the words. Here is a word cloud of the most common 50 words in the trending TAGS. The size of the word reflects how common it is:

Below facts can be considered when uploading video to make the video trending.

Total time taken to run exploratory and recommendation model

Approximately 20minutes are taken to completely run the code including graphical analysis.

Challenges faced and Gaps from Proposal

Calculating Similarity between each video for entire data set of 16009 was consuming lot of time. So, we are going ahead with Video Search recommendations for 1000 videos.