Analyzing and monitoring YouTube video trends before and after COVID

Manasi Choughule

· **Abstract**

YouTube is the world's most famous video streaming website, and it maintains a list of trending videos according to genre. To determine the year's top-trending videos, YouTube uses a combination of factors including measuring user's interactions (number of views, shares, comments and likes). Note that they're not the most-viewed videos overall for the calendar year". It has a list of trending videos which are updated frequently. Here we will use Python with some packages like Pandas and Matplotlib to analyze a dataset that was collected during pandemic. We will analyze this dataset to get insights into YouTube trending videos, to see what is common between these videos and what people can use to make the video popular and how the pandemic affected the viewing patterns of users.

The dataset we selected was collected using YouTube API containing data of several months on daily trending YouTube videos. Data is included for India, USA, Great Britain, Germany, Canada, France, Russia, Brazil, Mexico, South Korea, and Japan respectively, where we are using only USA's data to work for our project with up to 200 listed trending videos per day. Data includes the video title, channel title, publish time, tags, views, likes and dislikes, description, category_id, and comment count.

We came across the YouTube culture and trend articles where we found global sourdough view where at peak during the pandemic people in US were not having views on cooking related videos before pandemic and during the global shutdown people had changed viewing patterns and more will be discovered further in the report. [1]
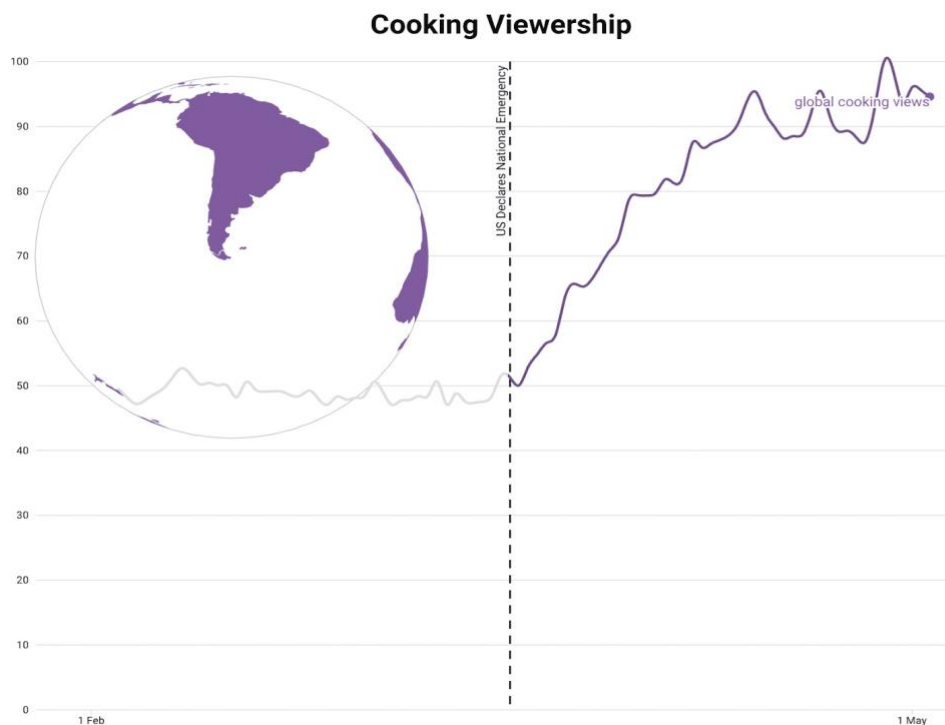
Fig 1. Cooking View ship plot after pandemic.

• Introduction

We have picked a dataset of daily records of top trending YouTube videos for our project and will analyze what factors affect in making the video a trending video. Also will find out the difference in the trends of records for before and after COVID-19 trends. Each user has various areas of interests and whenever they open the application, they will find trending suggestions according to the videos they viewed before. That is how the monitoring of the application helps the users to have better experiences and the brands improves their strategies to make customer experiences more relevant and efficient.

While we were searching for our dataset, we did some extra research to find out how and what parameters make video trending. We discovered that comment counts on a video results in high ranking, video length has a huge impact, views on the video, subscriber size of a channel, likes for a video, if a video drives a subscriber to the channel, keyword match for the video titles. All these parameters have an impact on making a video trending.

• Objectives

- The objective of the project is to analyze the dataset and monitor it to find the conclusion about how the COVID-19 impacted the YouTube business.
- What all parameters affect the dataset in making the videos tending or vice versa.
- Retrieving the video which is most viewed.
- Finding the video which has the highest count of comments.
- Finding the most disliked video.
- Finding the most liked video.
- Finding the major difference in categories viewed by users during and after pandemic.

• The Dataset

The YouTube API was used to get the data for this project. The data from 2020-08-21 to 2021-04-21 of trending YouTube is used. A script was set to run every day, fetching data from the YouTube API about the day's trending images. The data obtained from YouTube will then be stored in csv files. [2]

In our project, we have analyzed the data for the US region only, which consisted of 51991 videos. YouTube generally puts 200 videos on its trending list. Hence for the given time period it has generated almost 200 entries for each day.

The dataset has unique 8985 videos. The videos can be trending on multiple days too, hence it would have multiple entries for the same video. All videos in our data has following attributes:

video_id	3C66w5Z0ixs
title	I ASKED HER TO BE MY GIRLFRIEND...
publishedAt	2020-08-11T19:20:14Z
channelId	UCvtRTOMP2TqYqu51xNrqAzg
channelTitle	Brawadis
categoryId	22
trending_date	2020-08-12T00:00:00Z
tags	brawadis|prank|basketball|skits|ghost|funny videos|vlog|vlogging|NBA|browadis|challenges|bmw i8|faze rug|faze rug brother|mama rug and papa rug
view_count	1514614
likes	156908
dislikes	5855
comment_count	35313
thumbnail_link	https://i.ytimg.com/vi/3C66w5Z0ixs/default.jpg
comments_disabled	False
ratings_disabled	False
description	SUBSCRIBE to BRAWADIS ▶ http://bit.ly/SubscribeToBrawadis\r\rFOLLOW ME ON SOCIAL\r▶ Twitter: https://twitter.com/Brawadis\r▶ Instagram: https://www.instagram.com/brawadis/\r▶ Snapchat: brawadis\r\rHi! I'm Brandon Awadis and I like to make dope vlogs, pranks, reactions, challenges and basketball videos. Don't forget to subscribe and come be a part of the BrawadSquad!

Fig 2: Our data

As per Fig 1, we can see that the data has 16 attributes. YouTube assigns a unique id to each video, which is stored as "video_id". For our data, it is possible for the same video_id to be present in multiple lines, as the same video could be trending for more than one day. Video title is stored as "title" for each video. Precise first publishing time of every video is stored in "publishedAt". The publishing time consists of dates in yyyy-mm-dd format along with time in 24-hours format.

Each channel of YouTube is identified by a unique alphanumeric id, which is present in "channelId". Each YouTube content creator associates their channel with a channel title. It makes it easier for the viewers to search the channel content. It is titled as "channelTitle" in YouTube data. YouTube categorizes each video based on the type of content it has. Each such category is enumerated and added to the data based on its content, in "categoryId". Some of the categories YouTube are Music, Sports, Gaming, etc. Each entry has a record of the current "trending_date" of the video.

YouTube content creators use "tags" as a way to reach out to more audiences. They can be used to improve the searching process for a video. Generally, a video can have multiple tags too. The field "view_count" is used to store the total views of that trending video, on a given date. The number of likes and dislikes are also recorded as "likes" and "dislikes" respectively. Viewers of the video can comment on the video. Count of the comments is stored as "comment_count".

Each video has a thumbnail given by its creator. The link to such a thumbnail is stored in "thumbnail_link" by YouTube. The creator is given a choice to disable the comment on the videos.

This choice is stored as boolean value in the dataset. This choice is labeled as "comments_disabled". Each YouTube video has a rating associated with it. This is mainly for the purpose to restrict the audience with age-appropriate content. A boolean flag is also stored in the data as "ratings_disabled", to show whether the ratings are disabled for the video or not.

YouTube also allows the content creators to give a "description" of the video. The field generally contains information from the creators. It may be related to the video, links for other videos, notes from the creators, links for other social media accounts of the creator, etc.

Pre-processing

1. Import libraries:

   The different libraries we have used are:
   - Pandas: It is used for data processing and analysis.
   - NumPy: A fundamental package in Python, used for computation purposes.
   - Matplotlib: Python library which is used for visualization.
   - Seaborn: Another Python library used for visualization.
   - Scikit-learn: These libraries are used for various data preprocessing and algorithms.

   - Word cloud : A python library which is used for data visualization techniques to represent text data in which size of each word is based on the frequency or importance of the word.

2. Read data:

   The YouTube data from US_youtube_trending_data.csv file is read and stored in dataframe "data" in Python.

Reading the data file

```
data = pd.read_csv("US_youtube_trending_data.csv",header=0, na_values="?")
```

Fig 3: Reading the data

3. Checking for missing values:

   The data is then checked for the missing values.

```
In [16]: data.isnull().sum()

Out[16]: video_id            0
         title               0
         publishedAt         0
         channelTitle        0
         trending_date       0
         tags                0
         view_count          0
         likes               0
         dislikes            0
         comment_count       0
         thumbnail_link      0
         comments_disabled   0
         ratings_disabled    0
         description       394
         category            0
         dtype: int64
```

Fig 4: Checking for the null values

The data has a total 394 missing values in the "description" column. Since the number of missing values is not significant as compared to the total entries in the data, we replaced the missing values with blank string ("").

4. Checking for categorical data:

This step is performed as processing categorical data creates some issues. To avoid that the categorical data is converted to some numeric values. By analyzing our data, we could not find any categorical data.

5. Replacing category field with respective category text:

In our data, we replaced the "categoryId" field with "category" field. For this we had sent a json request to the Google API. The API sends back a json file containing the list of category id with its respective category name.

6. Splitting the data:

We have split our data into two dataframes, "data_2020" and "data_2021" for years 2020 and 2021 respectively. The video count for 2020 is 24313 and for 2021 is 27595.
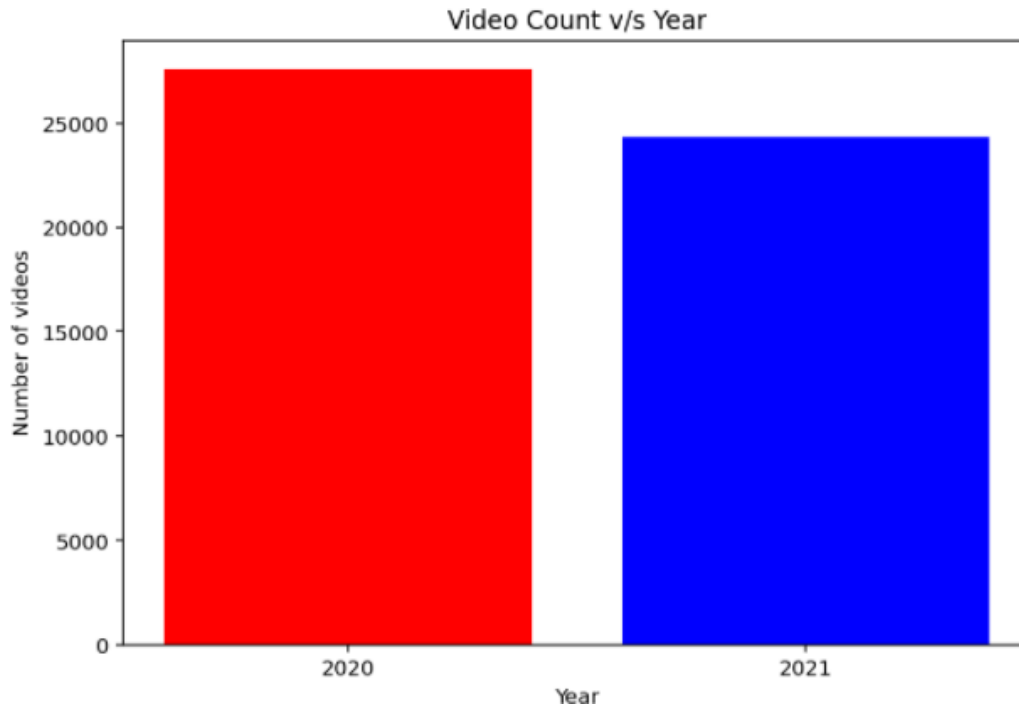
Fig 5: Video count per year

• The System

• Architecture

Data architecture diagrams are visual representations of how data will be managed from input to output. They usually include all the steps of your analytics architecture and show you how they connect to each other. Below architecture shows the way our project is collecting data and what all steps are taken to generate the outputs.
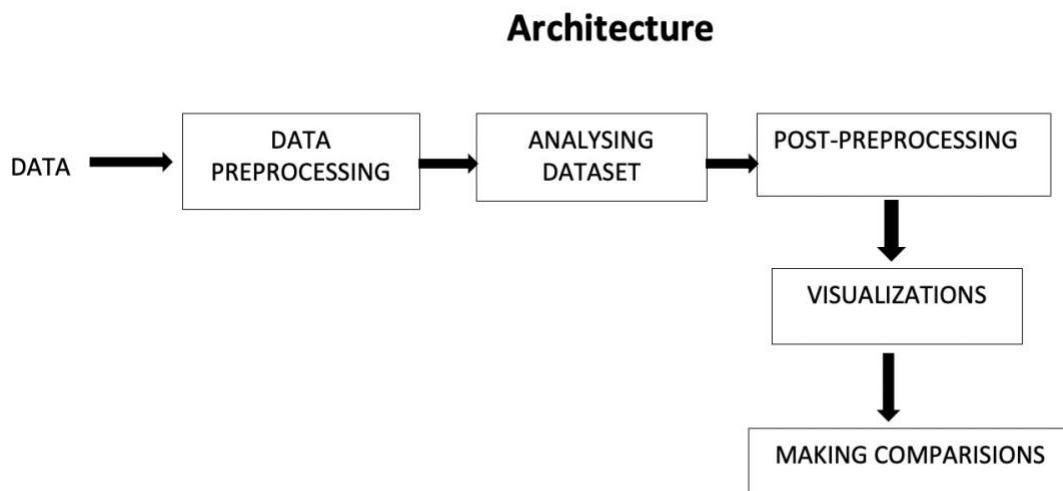
Fig 6: Project Architecture

The above architecture shows how the data is fed to the system and the output information is obtained.

• Data Processing

Data processing is done when data is collected and translated into usable format, it is important that this operation is performed correctly as not to negatively affect the end product, or output. Data processing starts with data in its raw form and converts it into a more relevant format (graphs, documents, etc.), giving it the form and context necessary to be interpreted by computers and making it readable. [2]

• SW/HW Development platforms

We have used software development platforms for data visualization and analysis.
Languages- Python
Framework- Jupyter Notebook.

• Experimental results and analysis

The analysis and results without explicitly mentioned are from all 51991 entries of YouTube videos. We have not considered only the unique videos and treating all the videos as new videos similar to how YouTube treats them. A video trending on 4 days would be considered to have more impact on trending parameters than the video trending on a single day only. That's the reason we have not removed the multiple entries of a video on multiple days.

In our project, for some analysis using unique video would give better insights. For such analysis, we have mentioned it whenever we have used only the unique videos

Trending Videos and Views

Minimum view video
Views is one of the essential parts of YouTube trending video criteria, but it is not the only parameter. Because of which we were able to even spot a few videos with 0 views in both the years.

```
In [33]: data_2020[data_2020['view_count']!=0]["view_count"].min()

Out[33]: 38510


In [34]: data_2021[data_2021['view_count']!=0]["view_count"].min()

Out[34]: 48325
```

Fig 7: Minimum view count for both the years

In 2020, a video had 38510 views and was still in the trending list. In 2021, the same number rose to 48325. We can say that the minimum view count increased from 2020 to 2021.

Maximum view video

The maximum view count a trending video had in 2020 was 232649205. The same number dropped in 2021 to 103691157.


```
In [36]: print("Maximum view count of a trending video in 2020: ", data_2020['view_count'].max())

         Maximum view count of a trending video in 2020:  232649205

In [37]: print("Maximum view count of a trending video in 2021: ", data_2021['view_count'].max())

         Maximum view count of a trending video in 2021:  103691157
```

Fig 8: Maximum view count

We can conclude that, in 2021 maximum view count of the trending video almost dropped to half of 2020.

View count distribution

As we can see in figure, View count distribution, in 2020 the view count video was densely packed under 40 million and the view count decreased exponentially till 240 millions. For 2021, the view count is densely packed till 30 million and reduced drastically till 110 million.
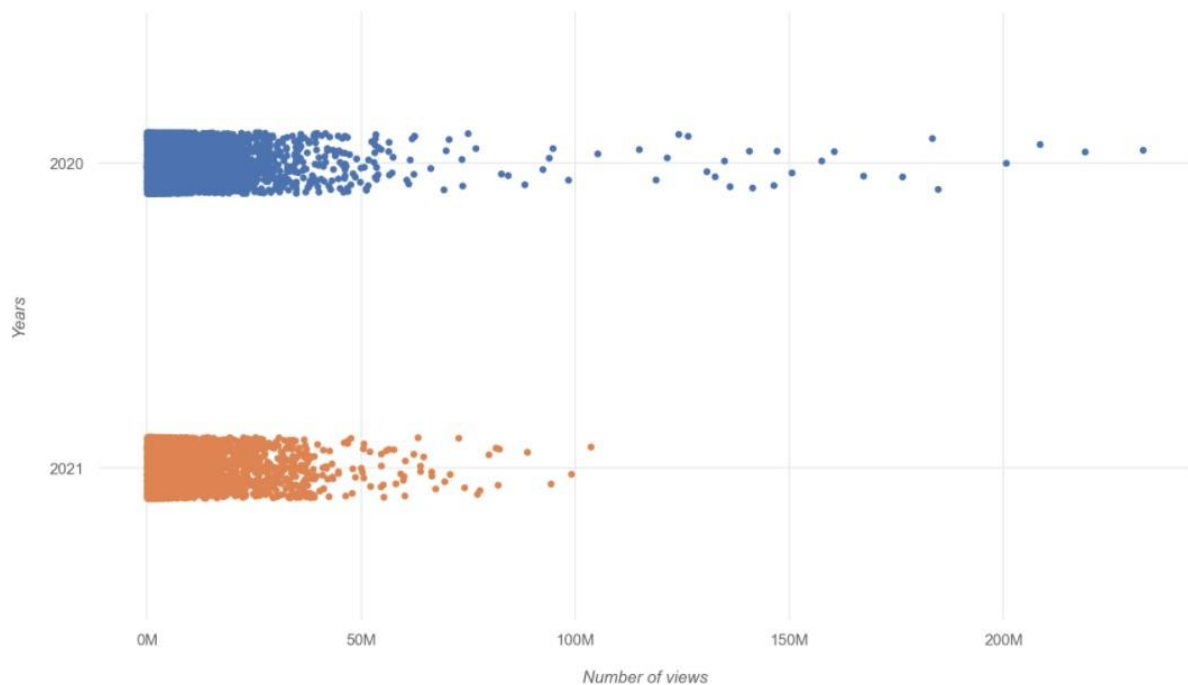
Fig 9: View count distribution

The view count distribution is more in 2020. In 2021, trending videos seem to have less video view count. The same result can be also seen in the histograms for the given years.
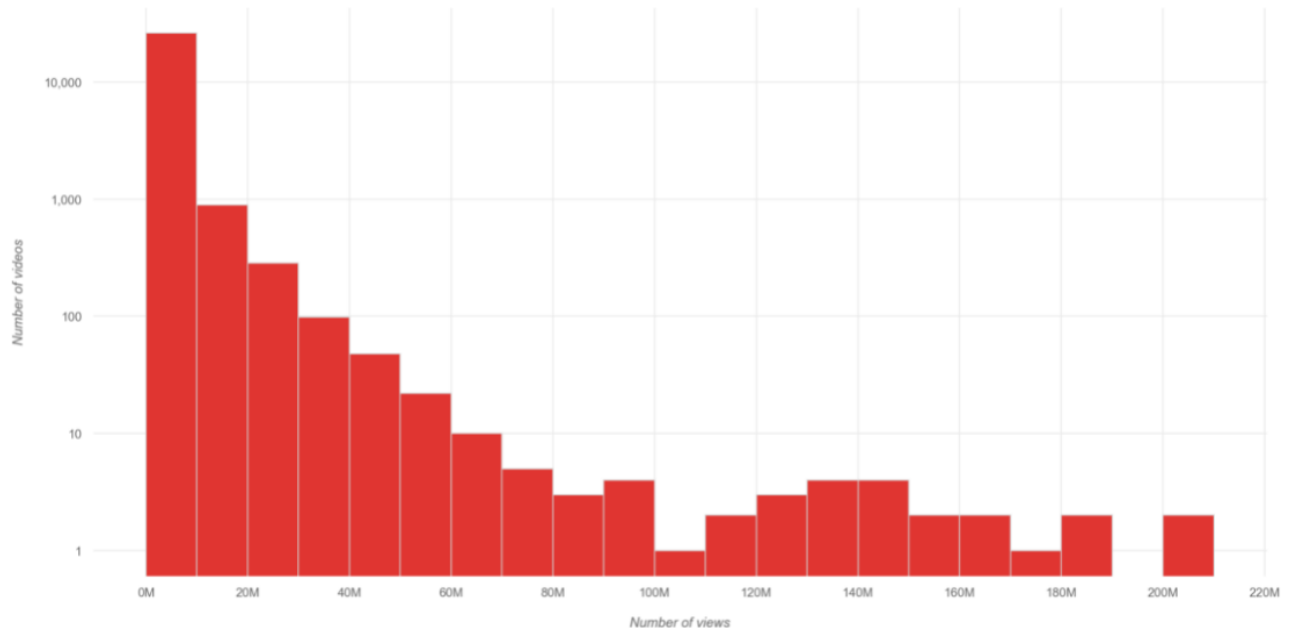
Fig 10: View count histogram for 2020
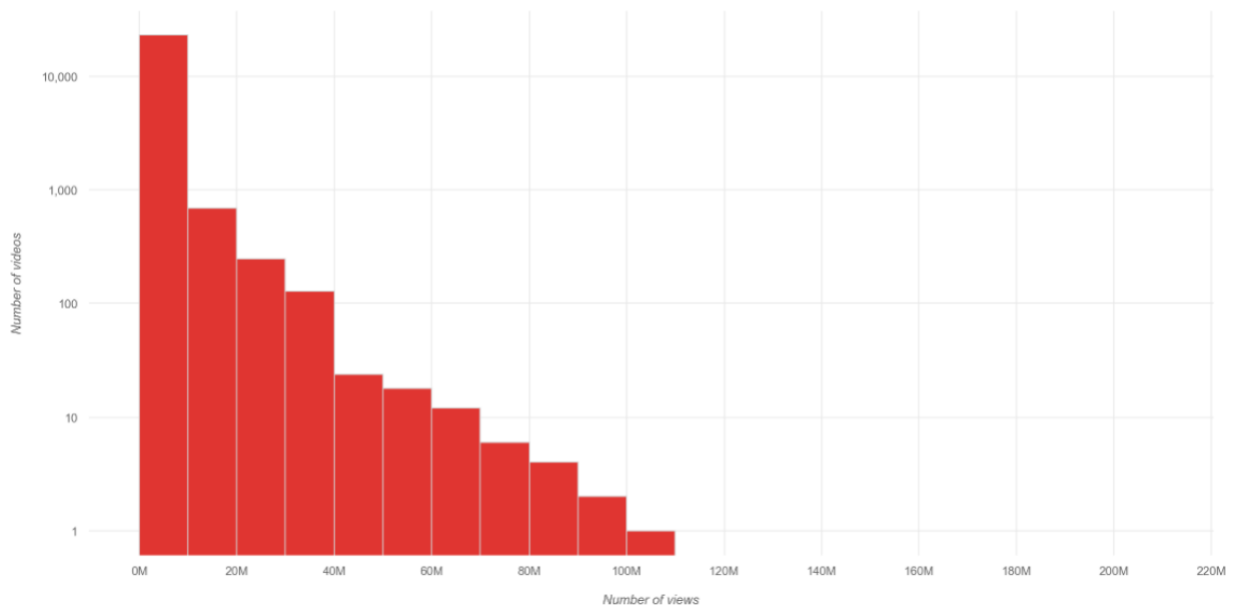
Fig 11: View count histograms for 2021

Trending Video Categories

Based on the content, YouTube groups the videos in various categories. The different YouTube categories are Music, Sports, Gaming, Education, etc.
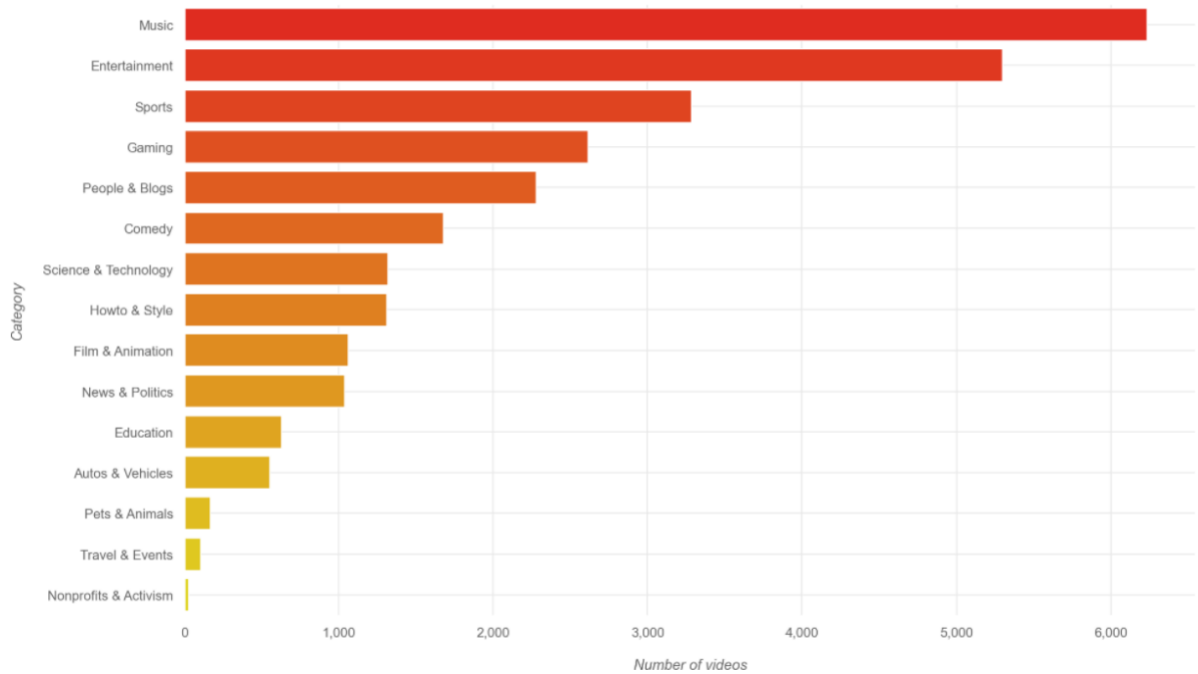
Fig 12: Trending video categories in 2020

In 2020, amid the pandemic music was the most trended category with more than 6000 videos. This was followed by entertainment, sports, gaming, and so on. Travel & events alongwith nonprofit & activism remain the least watched categories for this time. [3]
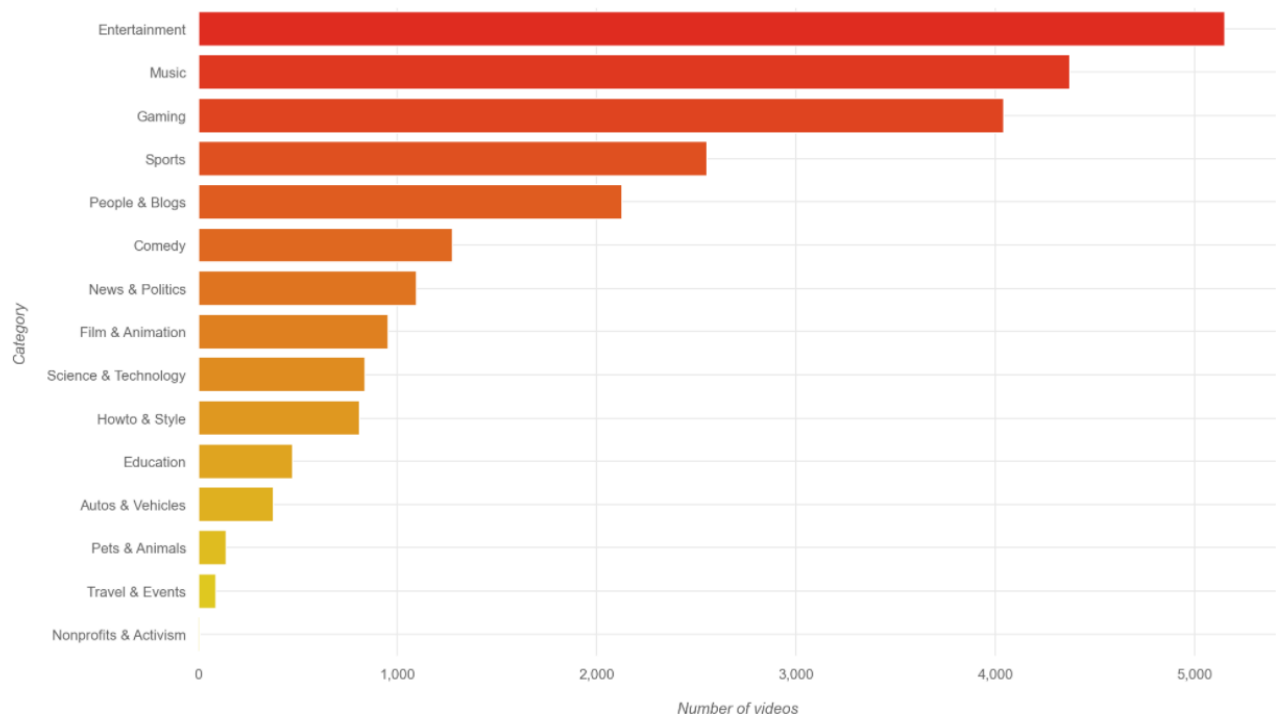
Fig 13: Trending video categories in 2021

From the start of 2021, a change has been seen in the most watched category. The most watched category shifted to entertainment. Music, which was the most trending category in 2020, became the second most watched category. News & Politics raised a few positions in the trending category list whereas science & technology moved down a few positions. [4]

Relationship between description length and views

YouTube content creators can put a description on each video. Generally, the video description contains information about video content, links to their other social media profiles, links to their other videos, etc.
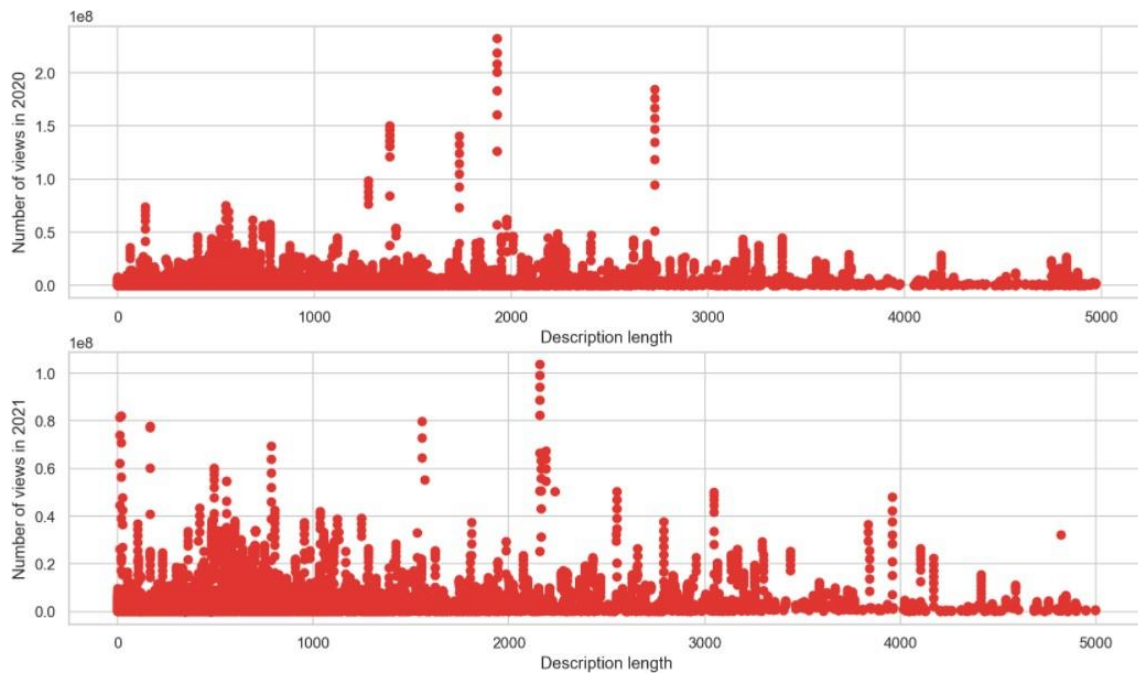
Fig 14: Description length and number of views

From the above figure we can say that for both the years the minimum description length is 0 and the maximum description length is 5000. In 2020, maximum watched videos had descriptions between 1000 to 3000. This range got shifted to 0 to 2200 in 2021.

Most Common Tags
When creators upload video they can add tags to the video. These tags help the audience to search the video in most cases. Almost all trending videos have tags.

Fig 15: Most used tags in 2020

In 2020, "none" or no tag was seen the most in trending videos with 3074 videos, funny was seen in 1533 videos, comedy was seen in 1145, and '2020' was seen in 955 videos.

Fig 16: Most used tags in 2021

No tag is still the most used tag in 2021. Funny and comedy were used again in 2021 1533 and 1077 times respectively. A new tag 'minecraft' is used the next most and is seen in 934 trending videos.

Trending for most Days!

	video_id	title	trending_days	view_count	likes	dislikes	comment_count	category
7401	pvPsJFRGleA	Justin Bieber - Holy ft. Chance The Rapper	27	6217404	936304	13331	91366	Music
21599	UF7hTMMeRv0	Mike Tyson and Roy Jones Jr hilarious joint interview after their exhibition	17	8273387	137286	2822	16663	Sports
17401	moOxq_8I_34	How President Trump and the White House reacted to Biden's projected presidential election victory	13	2647578	24172	5344	17599	News & Politics
15998	Q4LKWQyeBw8	Gervonta Davis KOs Leo Santa Cruz With Vicious Uppercut	SHOWTIME PPV	12	980376	20336	466	6263	Sports
16006	uioLnStATUc	Among Us Logic 6	Cartoon Animation	12	3089123	148835	2752	14631	Film & Animation
16200	1QfJya5dzI4	NLE Choppa - Bryson (Official Music Video)	12	1283217	130336	1432	16840	Music
15599	TmKh7lAwnBI	Bad Bunny x Jhay Cortez - Dákiti (Video Oficial)	11	3522783	734615	6144	36758	Music
16013	FafXBaAEowM	Among Us But PewDiePie Goes 90,000 IQ!	11	5115516	335726	2951	20231	Gaming
16398	4axpdkC2m2s	GETTING READY WITH LARRAY!!!	Addison Rae	11	2436276	142613	4226	6925	People & Blogs
16403	mW-FceZgH8k	Boxer Adrien Broner held in contempt of court, sent to jail	11	163521	2417	218	1716	News & Politics
16418	TPpmOzpHiE0	TWICE I CAN'T STOP ME Dance Practice Video	11	2339429	299498	2111	11496	Music

Fig 17: Trending on most days in 2020

In 2020, while the pandemic was still in its prime state, a video was trending for 27 days. It belongs to the category music and it was a video from singer Justine Bieber. Other videos were trending for less than 18 days.

	video_id	title	trending_days	view_count	likes	dislikes	comment_count	category
31201	wY6UyatwVTA	India claim stunning series win, end Australia's Gabba streak	Vodafone Test Series 2020-21	33	13516854	492839	8512	35888	Sports
48993	ly6yEJSDFko	Prince Philip's funeral: the key moments	15	2886289	31712	1335	0	Entertainment
48994	GdJkO3PLZrs	Prince Harry and William Walk Together as Royal Family Departs Prince Philip's Funeral	14	6640739	45805	2422	6844	News & Politics
27629	1SIX0HTH-50	Worst Dude Perfect Videos of All Time	OT 23	10	9107920	362221	7091	38877	Sports
29803	PmDx0wPG7Cc	Dax - Eminem Rap God Remix [One Take Video]	10	526402	126669	2108	14610	Music
30197	gjIGeMsgfhA	Stephen A. reacts to James Harden being traded to the Brooklyn Nets	SportsCenter	10	1643991	25372	721	5398	Sports
30198	jny0fATDVBo	How Lebron was when he found out James Harden Went to the Nets	10	1236121	154140	939	8979	People & Blogs
35995	aXzVF3XeS8M	Taylor Swift - Love Story (Taylor's Version) [Official Lyric Video]	10	2899792	636149	1619	62796	Music
39195	28RxpkcVAFw	Customizing Nintendo Switch In 10 Hours, 1 Hour, 10 Minutes, 1 Minute & 10 Seconds!	10	1722872	111691	1300	15020	Entertainment
42792	PCzkHAI9D7k	Minecraft, But I Put A T Shirt On Every Minute...	10	3120341	362105	3210	30684	Gaming
42993	tQ0yjYUFKAE	Justin Bieber - Peaches ft. Daniel Caesar, Giveon	10	4083919	633407	6805	37988	Music

Fig 18: Trending on most days in 2021

In 2021, when things are starting to improve, a sports category was trending the most which is for 33 days. It is a cricket video between India and Australia. The other most trending videos were trending for less than 16 days.

From both images we can say that in 2021 YouTube video was able to be in the trending list more, but 2020 has other videos with more trending days.

Title Length
How long are the titles of the trending videos in both years?
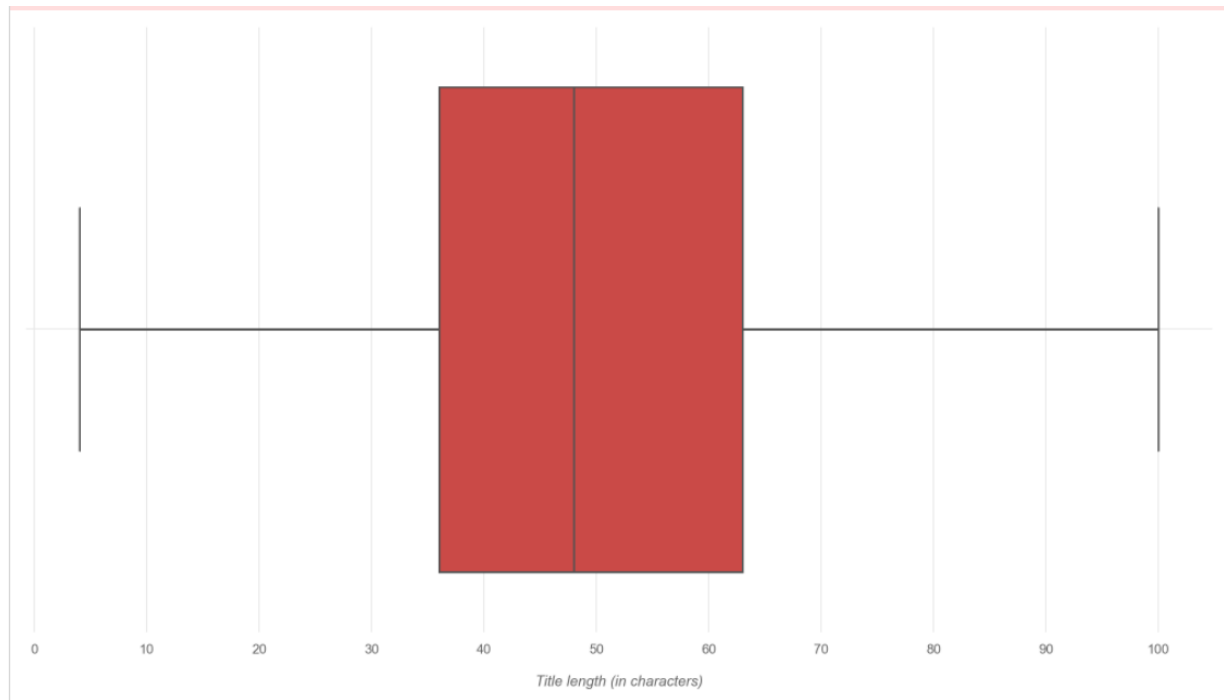
Fig 19: Boxplot for title length in 2020

In 2020, the titles were 36 to 84 characters long. On average, the video title length is 48. There were videos with the shortest title of 4 characters in 2020 and videos had title length of 100 too, which is the maximum allowed title length.
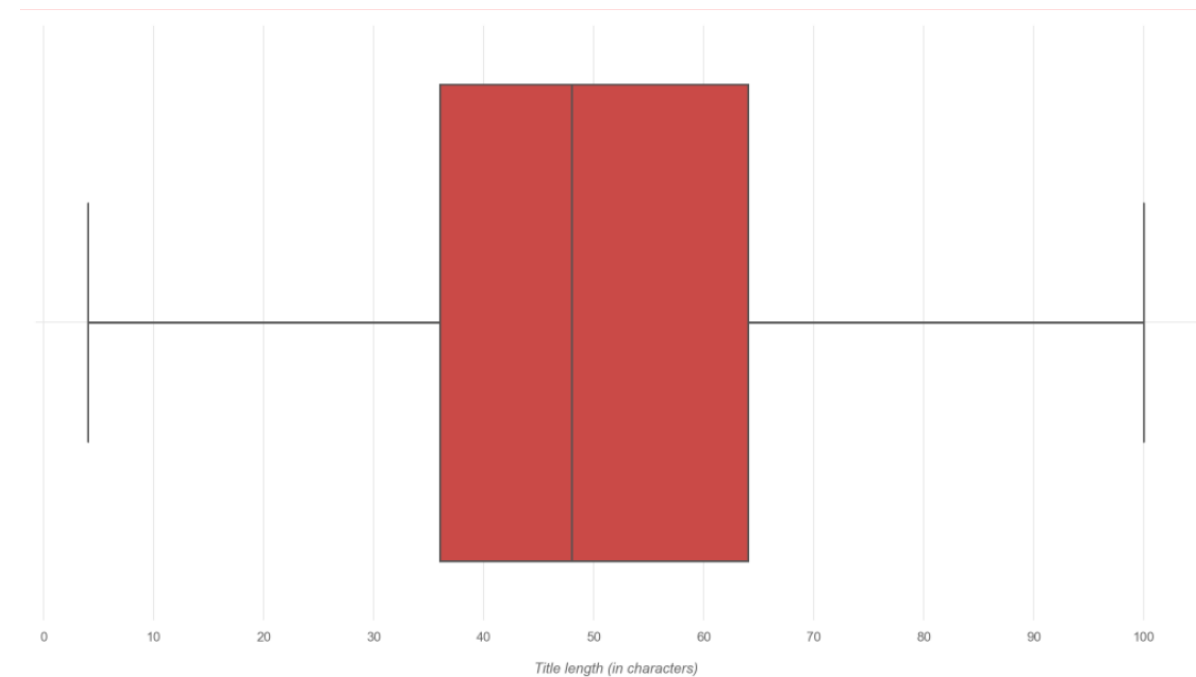
Fig 20: Boxplot for title length in 2021

For year 2021, the title were between 37 to 65 characters. With the average title length, 48, minimum video title length is 4 and maximum video title length is 2021.

For both the years, the minimum and maximum title length remained the same with a minor change in   the   average   title   length.

• Conclusions

The project's key aim was to explore the data and use powerful visualizations to discover interesting statistics and patterns. The main takeaways from the data which we found are:

1. The minimum view count of videos on the trending list increased from amid the covid to after the covid.
2. After the covid, maximum view count of the trending video almost dropped to half of the view count in the covid.
3. Covid has also impacted the view count distribution. More distribution is seen in 2020 than in 2021.
4. There is also change in most watched video categories due to covid. In 2021 the most watched category got shifted to entertainment from music in 2020. In 2021, News & Politics raised a few positions in the trending category list whereas science & technology moved down a few positions.
5. During the covid, the description would be between 1000 to 3000 of most trended videos. This range got shifted to 0 to 2200 in after covid duration.
6. Trending videos have the first three most used tags the same for both the duration. The fourth most used got changed from '2020' during covid to 'Minecraft' in post covid.
7. Post covid, the same video is trending for more days.
8. For both the duration, title length remains the same of trending videos.

• Lessons Learned

In this project, we were able to understand and perform the different procedures of data analytics projects. This helped us in enhancing the knowledge about research selection, data preprocessing, data analysis and data visualization methods. This also helped us in discovering the different effects and patterns on YouTube trending videos and derive insightful information about them.

Works Cited

[1 "https://www.youtube.com/trends/articles/covid-impact/," [Online].
]

[2 "https://www.kaggle.com/rsrishav/youtube-trending-video-dataset/metadata," [Online].
]

[3 "https://www.talend.com/resources/what-is-data-
] processing/#:~:text=Data%20processing%20occurs%20when%20data,end%20product%2C%
   20or%20data%20output.," [Online].

[4 "https://www.thinkwithgoogle.com/consumer-insights/consumer-trends/2020-youtube-
] trends/," [Online].

[5 "https://ammar-alyousfi.com/2020/youtube-trending-videos-analysis-2019-
] us?src=kgl#trending-videos-categories," [Online].

• Appendix

Snippet of YouTube data