YouTube Data Collection and Analysis

To collect data from YouTube, we need to be clear about what data we need. Let’s collect data about the trending videos on YouTube to analyze and find what makes a video trend on YouTube.

So, let’s start with data collection first. To collect data from YouTube, you need to set up an API. Here are the steps you can follow:

1. Go to Google Cloud Console.
2. Click on the project drop-down at the top, then “New Project”.
3. Enter a project name and click “Create”.
4. In the Google Cloud Console, navigate to “APIs & Services” > “Library”.
5. Search for “YouTube Data API v3” and click on it.
6. Click “Enable”.
7. Go to “APIs & Services” > “Credentials”.
8. Click “+ CREATE CREDENTIALS” and select “API key”.
9. Copy the generated API key.

YouTube Data Collection and Analysis with Python

Step-1.

Code Implementation:
I have collected data about the top 200 trending videos on Youtube.
Output Observations:
We are using the YouTube Data API to fetch details of the top 200 trending videos in the US, iterating through the API’s paginated responses to collect video details such as title, description, published date, channel information, tags, duration, definition, captions, and various engagement metrics like views, likes, and comments. The script compiles

this information into a list, converts it into a pandas DataFrame, and saves the data to a CSV file named trending_videos.csv, allowing us to analyze trends and patterns in the collected video data.

Data Description using Pandas:

Step-2.

I have used pandas for description of data collected though api i.e., “trending_videos.csv”.

Code Implementation:
Now, checked for missing values and data types of columns.
Output observations:
Found that The description column has 4 missing values. This is minor and can be handled as needed. The data types seem appropriate for most columns, but we may need to convert the published_at column to a datetime format and tags might need further processing.

Descriptive Statistics of data:

Step-3.

Code implementation:
Parameters:
‘view_count', 'like_count', 'dislike_count', 'comment_count'

Distribution of viiews,likes and comments of all videos in data using matplotlib and Seaborn libraries.
Output Observations:

The histograms(fig.1) show that the distributions of view counts, like counts, and comment counts are right-skewed, with most videos having lower counts and a few videos having very high counts.

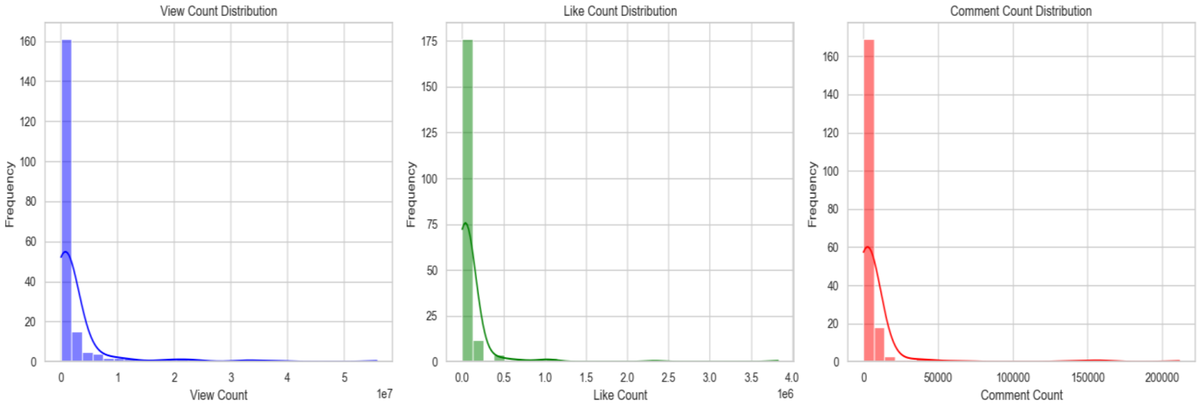


Figure 1 Histogram distributions of views,likes and comments

Correlation Matrix:

Step-4.

Code Implementation:
Have a look at the correlation between likes, views, and comments:

Output Observations:
The heatmap confirms strong positive correlations betwee views, likes, and comments.

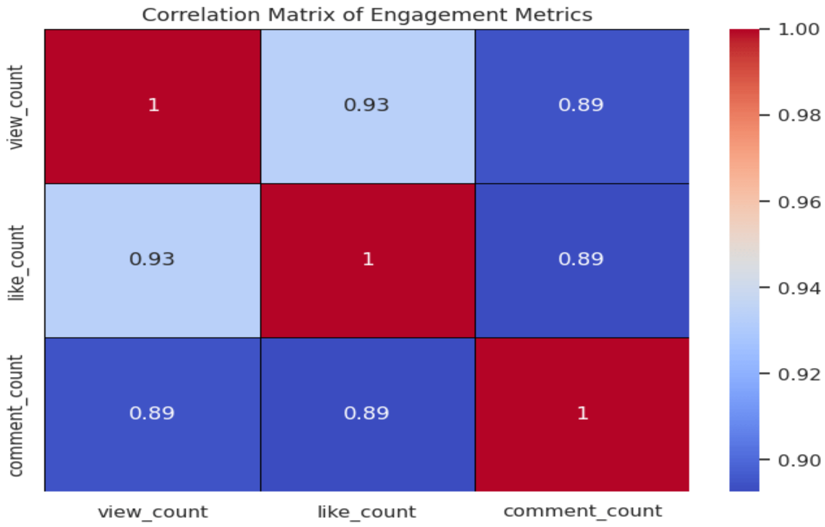


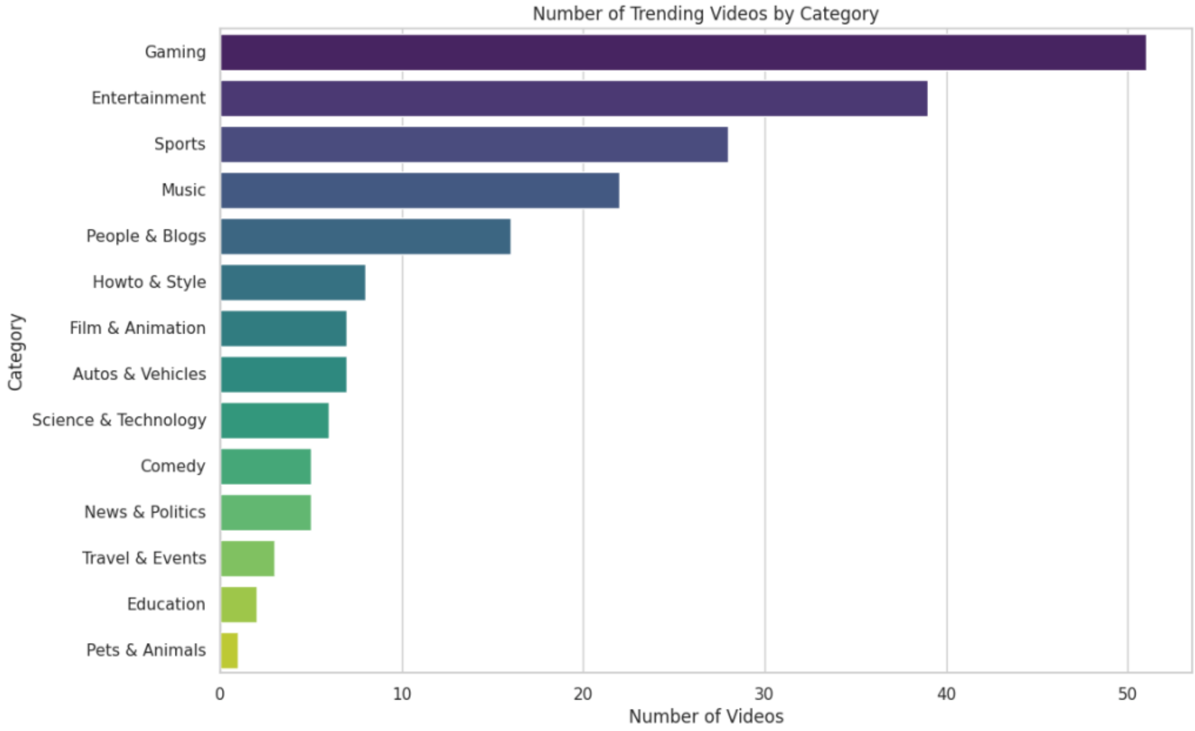
Figure 2 Heatmap

Step-5. Since we have collected only category ID, lets collect category name also from the api.

Analyze the Trending videos by Category:

Step-6.

Code implementation:
Analyze the no.of trending videos on youtube.



Output Observations:
The bar chart shows that the Gaming, Entertainment, Sports, and Music categories have the highest number of trending videos.

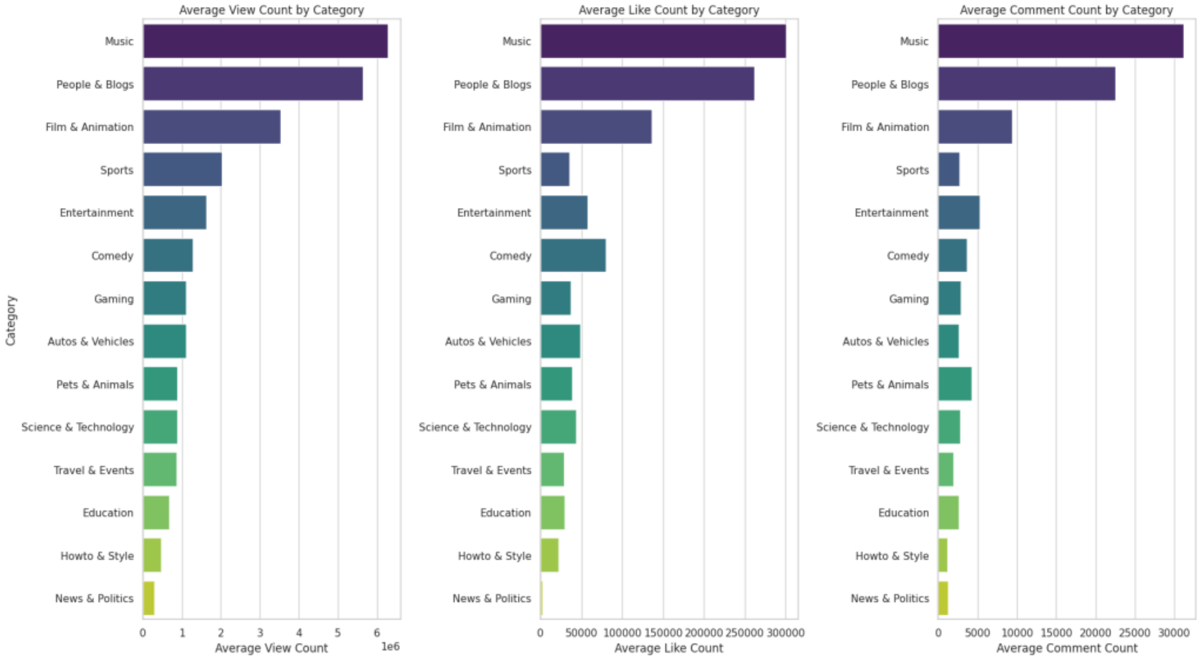
Average engagement metrics by category:

Step-7.

Code implementation:
let’s have a look at the average engagement metrics by category.

Output observations:

Music and People & Blogs categories have the highest average view counts, likes, and comments. Film & Animation also shows high engagement, especially in view counts and like counts.



Convert the duration from ISO 8601 format to seconds:

Step-8.

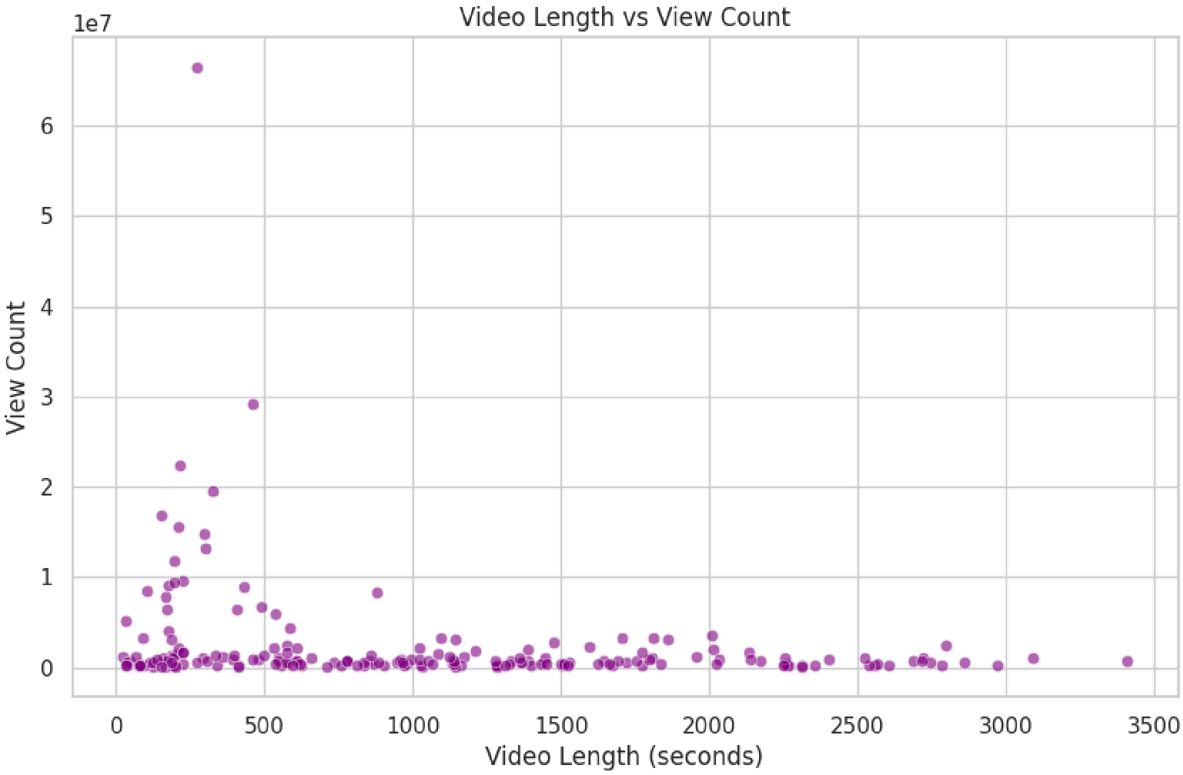
We are using the isodate library to convert the duration of each video from the ISO 8601 format to seconds, which allows for numerical analysis. After converting the durations, we are categorizing the videos into different duration ranges (0-5 minutes, 5-10 minutes, 10-20 minutes, 20-60 minutes, and 60-120 minutes) by creating a new column called duration_range. This categorization enables us to analyze and compare the engagement metrics of videos within specific length intervals, providing insights into how video length influences viewer behaviour and video performance.

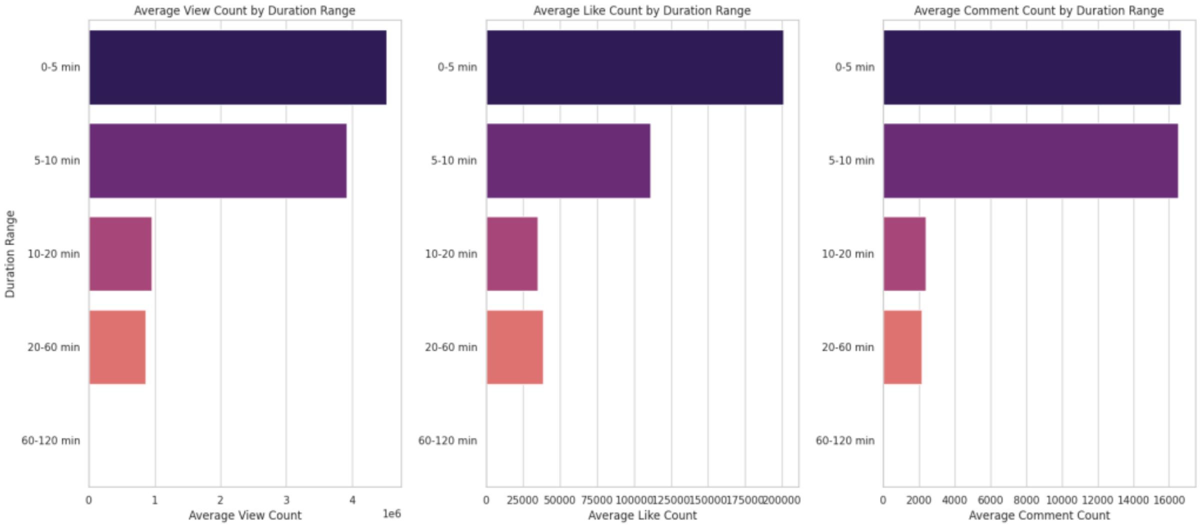
Analyze the Content and the Duration Of videos:

Step-9.

Code implementation:
Analyze the content and duration the videos.

Output Observations:
The scatter plot shows a slight negative correlation between video length and view count, indicating shorter videos tend to have higher view counts. Videos in the 0-5 minute range have the highest average view counts, likes, and comments. Engagement decreases as video length increases.


Video Length vs View Count
1e7
View Count
0
1
2
3
4
5
6
0
500
1000
1500
2000
2500
3000
3500
Video Length (seconds)


Average View Count by Duration Range
Duration Range
0-5 min
5-10 min
10-20 min
20-60 min
60-120 min
0
1
2
3
4
1e6
Average View Count
Average Like Count by Duration Range
0-5 min
5-10 min
10-20 min
20-60 min
60-120 min
0
25000
50000
75000
100000
125000
150000
175000
200000
Average Like Count
Average Comment Count by Duration Range
0-5 min
5-10 min
10-20 min
20-60 min
60-120 min
0
2000
4000
6000
8000
10000
12000
14000
16000
Average Comment Count

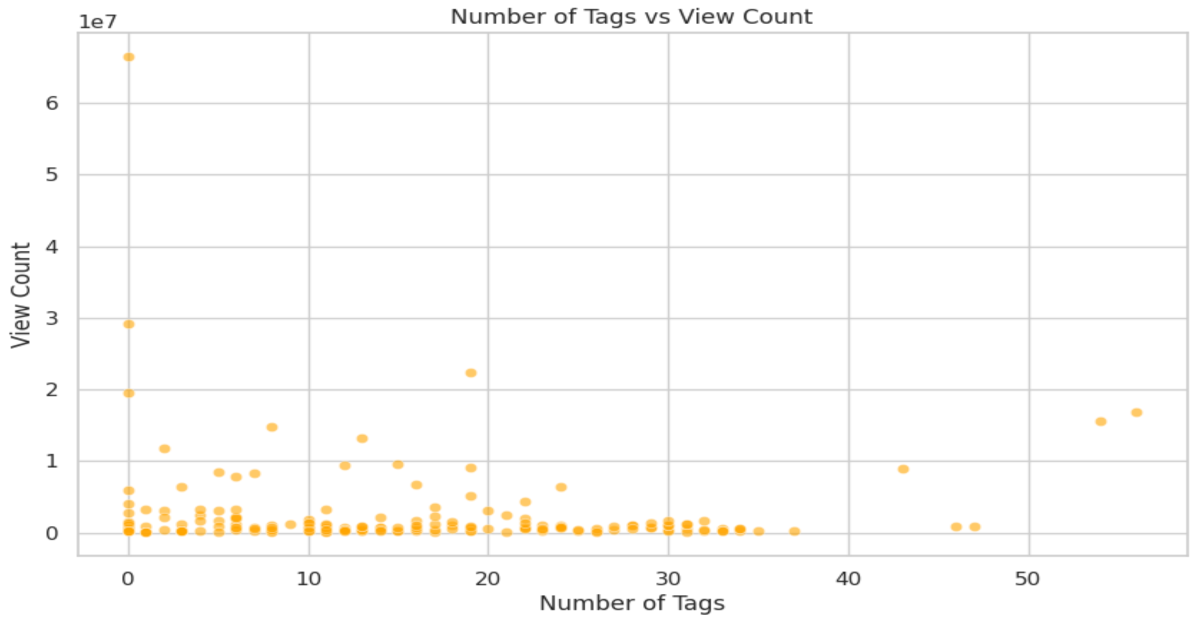
Analyze the relationship b/w views and no.of Tags:

Step-10.

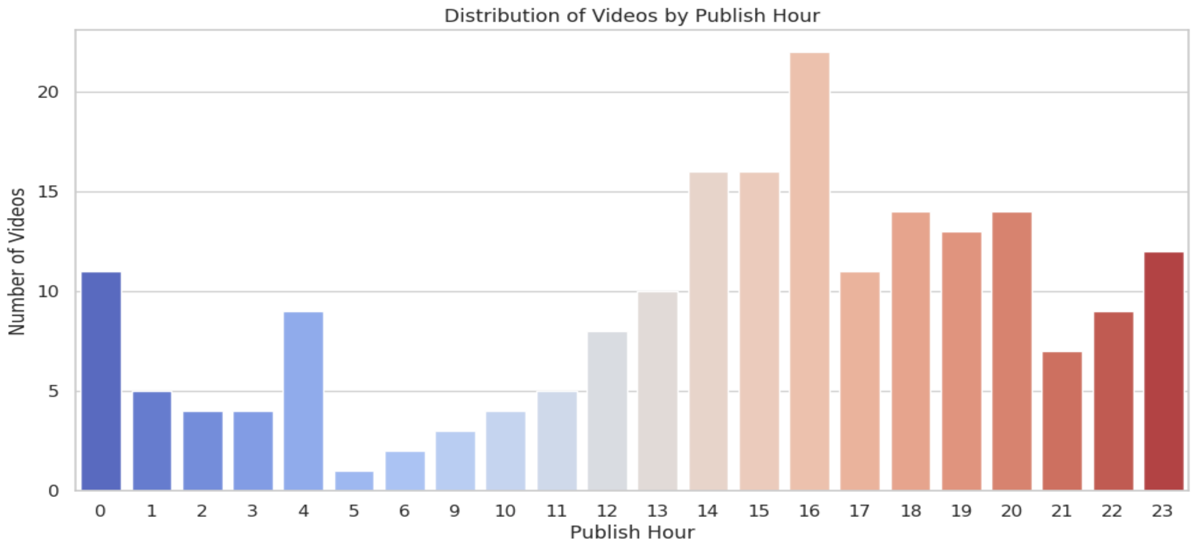
Code implementation:
Analyze the relationship b/w views and no.of Tags

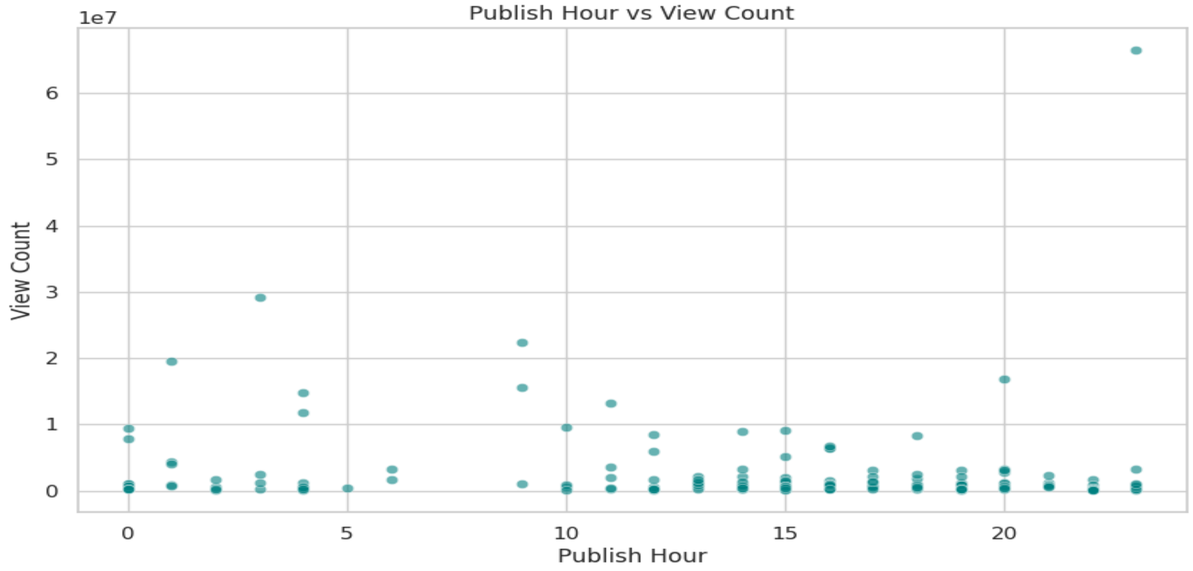
Output Observations:

The scatter plot shows a very weak relationship between the number of tags and view count, suggesting that the number of tags has minimal impact on a video's view count.



Impact of time a Video is posted on it’s views:





The distribution shows that most videos are published between 14:00 and 20:00 hours (2 PM – 8 PM), indicating this may be an optimal time for uploading videos. There is a very weak negative relationship between publish hour and view count, suggesting that the hour of publication has minimal impact on engagement metrics.

Conclusion:

So, here’s my conclusion on what makes a video trend on YouTube:

1. Encourage viewers to like and comment on videos to boost engagement metrics.
2. Aim to create shorter videos (under 5 minutes) for higher engagement, especially for categories like Music and Entertainment.
3. Schedule video uploads around peak times (2 PM – 8 PM) to maximize initial views and engagement.