INDIAN INSTITUTE OF TECHNOLOGY KANPUR

SOC471A Project Report

Title: Youtube Algorithm analysis

Submitted By:

Group No. : 10

Group Members:
Pulkit Dhamija (200738)
Rakesh Sankhla (200767)
Ritick Gupta (200801)
Ritik Raj (200803)

Under the Guidance of
Prof. Pradip Swarnakar

Table of Contents

Proposal 3
Introduction 4
- Overview 4
- Objective 4

Details 5
- What are we trying to do? 5
- How is it done today and what are the limits of current practice? 5
- What is new in our approach and why would it be successful? 6
- Whom would it affect and what difference would it make? 6
- What are the risks? 7
- How much will it cost? 7
- How long will it take? 8

Methodology 8
- Data Preprocessing and Exploratory Data Analysis (EDA) 8

Results 10
- 1. Top 20 Most viewed channels 10
- 2. Most viewed channel by year 10
- 3. Occurrence of Words in Video Titles 11
- 4. Most frequent tags used 12
- 5. Videos of which category are dominating in top trending videos? 13
- 6. Which channel has the most trending videos? 14
- 7. Exploration of views 14
- 8. Exploration of likes 16
- 9. Which day of the week has the most top trending videos published? 18
- 10. Channel-wise comparison between number of videos, views, ikes, and comments 18

Conclusion 21

Proposal

Title : Youtube algorithm analysis

Objective: Analysing the trend of a particular video based on views, likes, dislikes, comments, upload date

Research questions: Analytical and Statistical questions

Statistical

1. What is the correlation between video duration and the number of likes received?
2. What is the distribution of comments across different video categories?
3. How does the number of subscribers of a channel affect the likelihood of a video trending?

Analytical

1. Can we identify a pattern between the timing of video uploads and the likelihood of trending, considering different time zones?
2. How does the extremely positive, negative or catchy sentiment in the video correlate with the video's content?
3. Are there specific keywords or phrases that frequently appear in titles of trending videos, and do they impact viewership?

Data sources :
https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset

Significance: To understand the trend of the YouTube algorithm, which would help new YouTubers to grow faster by using the right amount of efforts

Scope & limitations :

It does not take into account the design and creative aspect of the video, such as designing the thumbnail of a video using the appropriate set of colours to match with the YouTube algorithm.
Eg: Various researchers have found that thumbnails which depict the scale of the video with appropriate figures and colours trend faster. Use of red colour to draw attention works as a click bait for most trending videos.
The story telling aspect of the youtuber is also not taken into account in this study.

Introduction

Overview

In the digital era, the popularity of online video content has soared, with platforms like YouTube becoming key players in delivering diverse content to global audiences. This project delves into the trending videos on YouTube in India to uncover patterns, preferences, and insights within the dataset. The analysis encompasses various aspects, including channel popularity, content categories, viewer engagement metrics, and linguistic patterns in video titles.

Objective

- Exploratory Data Analysis (EDA): Uncover patterns and trends within the dataset to gain a holistic understanding of the YouTube trending landscape in India.

- Content Category Analysis: Examine the distribution of video categories to identify the most and least common types of trending content.

- Channel Metrics: Analyse the frequency and popularity of channels, focusing on the top-performing channels and their content categories.

- Engagement Analysis: Explore viewer engagement metrics, including views, likes, and comments, to understand audience interaction with trending videos.

- Title Analysis: Investigate linguistic patterns in video titles, identifying frequently used words and themes.

- Time Analysis: Examine trends based on the time of video publication, particularly focusing on the days when videos are most likely to trend.

- Word Cloud Generation: Create word clouds to visually represent the most common words used in video titles, providing a quick and intuitive overview.

- Future Work Suggestions: Propose potential avenues for future analysis or enhancements to understand the dynamics of trending videos on YouTube.

Through these objectives, this project aims to provide valuable insights for content creators, marketers, and YouTube enthusiasts interested in the trends and patterns of trending videos in the Indian YouTube space. Additionally, this analysis seeks to contribute to the broader discourse on digital content consumption, shedding light on the preferences of the YouTube audience in India.

Details

What are we trying to do?

Analysing the trend of top 200 videos region wise based on video title, channel title, publish time, tags, views, likes , description, and comment count.

How is it done today and what are the limits of current practice?

Nowadays, consultants are hired or some YouTubers launch their courses which provide information on how to improve the reach and subscribers of a YouTube channel. Such ways are mostly non algorithmic and subject to the experience of the person hired. Such methods don't make much use of big data and is limited to people who can afford a consultant. Some creators make use of trending topics from social media or latest news trends which help them increase their reach in the short run but it is not helpful in the long run. This method is also genre specific. They need some concrete model which would be helpful in the long run and could be applied to any content creator from any genre.

What is new in our approach and why would it be successful?

Our method makes use of big data from the latest YouTube analytical data. It analyses the data through statistical approach and gives a relevant conclusion to the content creator. We take advantage of the major characteristics of big data and implement the same in our model. This method is efficient and would be accessible to all the content creators from different genres, be it a famous or a small content creator. This method is easily implementable, affordable and one does not need to hire an expensive consultant or subscribe to expensive courses. Our method would give efficient results in less time and would help increase the reach of the YouTube channel at a faster rate.

Whom would it affect and what difference would it make?

YouTube algorithm analysis" can benefit various individuals and groups, including:

- **Content Creators:** Creators on YouTube can analyse the algorithm to understand how it ranks and recommends videos. This knowledge can help them optimise their content strategy to reach a larger audience.
- **Marketers and Advertisers:** Marketers can gain insights into how YouTube's algorithm affects video visibility and engagement. This information can be used to create more effective advertising campaigns.
- **Researchers and Analysts:** Researchers can study YouTube's algorithm to understand its impact on user behaviour, content trends, and the spread of information.
- **Educators:** Educators can use algorithm analysis to understand how educational content is ranked and recommended on YouTube, helping them create more accessible and engaging lessons.
- **Viewers and Consumers:** Understanding the algorithm can help viewers discover content that aligns with their interests and preferences, enhancing their YouTube experience.

In summary, YouTube algorithm analysis can be valuable to a wide range of stakeholders, from content creators and marketers to researchers and everyday viewers. It helps enhance content quality, user engagement, and the overall YouTube ecosystem.

What are the risks?

- **Misinterpretation:** Analysing YouTube algorithms requires a deep understanding of data analysis techniques and algorithms. Misinterpretation of data can lead to incorrect conclusions.
- **Misinformation:** Findings from YouTube algorithm analysis can sometimes be misused or misunderstood, potentially contributing to the spread of misinformation.
- **Algorithm Changes:** YouTube frequently updates its algorithms, making it challenging to keep analyses up to date. Findings may become irrelevant as algorithms change. To mitigate these risks, researchers should follow ethical research practices, prioritise user privacy, and stay informed about YouTube's policies and algorithm changes

How much will it cost?

The cost of YouTube algorithm analysis can vary significantly depending on several factors, including the complexity of the analysis, data collection methods, and the scale of the project. Here's a breakup of potential costs:

- **Data Collection:** Data collection costs may involve web scraping tools, APIs, or data acquisition services. - Cost range: Free to several thousand dollars, depending on the data source and size.
- **Hardware and Software:** High-performance computers or cloud computing services are required for data processing. - Cost range: Varies based on computing needs, but cloud services may cost from a few dollars to several hundred dollars per month.
- **Data Preprocessing and Cleaning:** Data preprocessing tools and software for cleaning and structuring data. - Cost range: Varies based on the tools used, but free and paid options are available.
- **Analysis Tools and Software:** Data analysis tools, machine learning libraries, and visualisation software. - Cost range: Free and open-source options are available, but some specialised software may require licensing fees.

How long will it take?

The duration required to complete a YouTube algorithm analysis project can vary significantly based on various factors, including the complexity of the analysis, the volume of data, the specific research objectives, the availability of resources, and the expertise of the team involved. Therefore, providing a precise timeline without detailed project information is challenging. Here are some considerations to help estimate the project duration:

- **Data Preprocessing:** Cleaning and preparing the data for analysis can be time-consuming, depending on data quality and complexity.
- **Analysis Methods:** The complexity of the analysis methods, such as machine learning, natural language processing, or deep learning, can impact the timeline.
- **Project Scope:** The scope of the analysis, including the number of variables or features considered, can affect the duration.
- **Testing and Validation:** This can extend the timeline if the analysis requires rigorous testing and validation. In practice, a YouTube algorithm analysis project could take anywhere from several weeks for a basic analysis to several months for a more comprehensive research-oriented project. Smaller-scale analyses have shorter durations, while larger, research-focused projects could extend over several months or more. Proper project planning is essential to manage and optimise the timeline effectively.

Methodology

Data Preprocessing and Exploratory Data Analysis (EDA)

The initial steps involve loading the YouTube trending dataset for India and examining its structure. The dataset includes information about trending videos such as title, channel, views, likes, comments, and category. The first preprocessing steps focus on converting date columns to datetime objects and mapping category IDs to their corresponding labels for better interpretability.

Next, the tags column is cleaned by converting it into a list and handling special cases like "[none]." Natural Language Processing (NLP) techniques are applied to clean and process the video titles. This includes removing unnecessary words, special characters, and stopwords, resulting in a more streamlined representation of the titles.

The exploratory data analysis (EDA) starts with visualising the top 20 most viewed channels and identifying the most viewed channels by year. A category ID mapping exploration provides insights into the distribution of videos across different content categories.

An in-depth analysis of video titles follows, focusing on word choice and identifying the most frequently used words. The analysis extends to exploring the distribution of videos across various categories and channels, helping understand the landscape of trending content on YouTube.

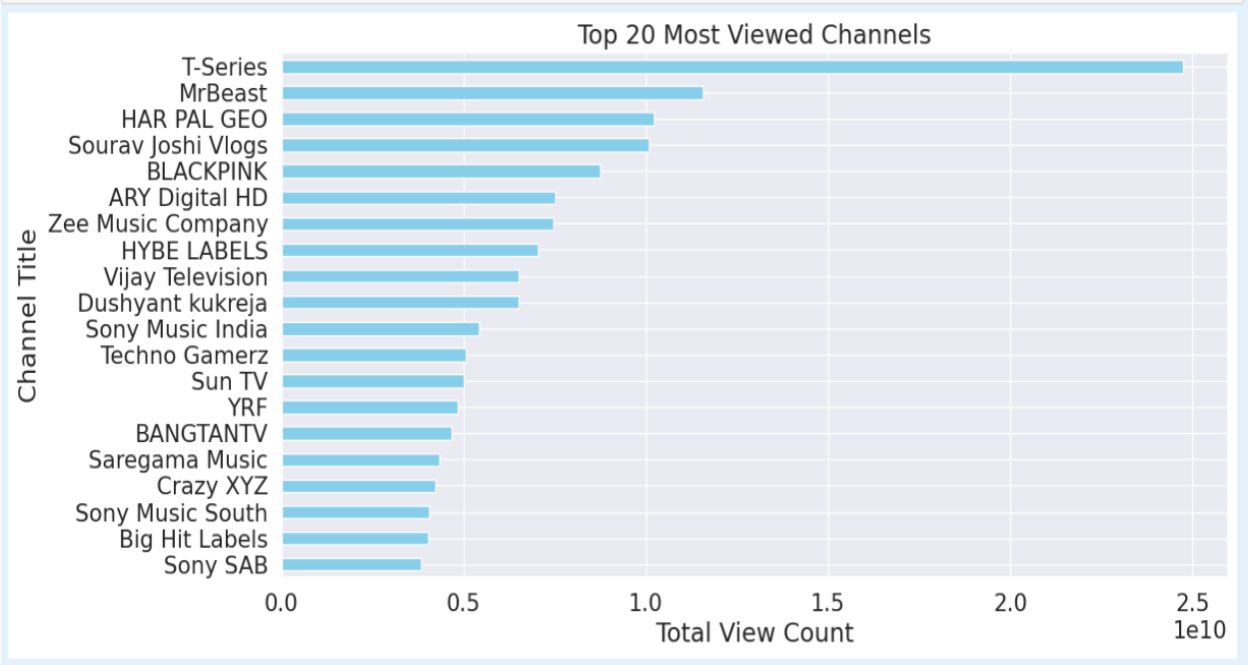
Exploring views, likes, and comments involves visualising their distributions and identifying patterns in viewer engagement. The time-based analysis looks at the publishing days to uncover any temporal trends.

Word cloud generation is employed to visually represent the most common words used in video titles, offering an intuitive overview of prevalent themes in trending content.

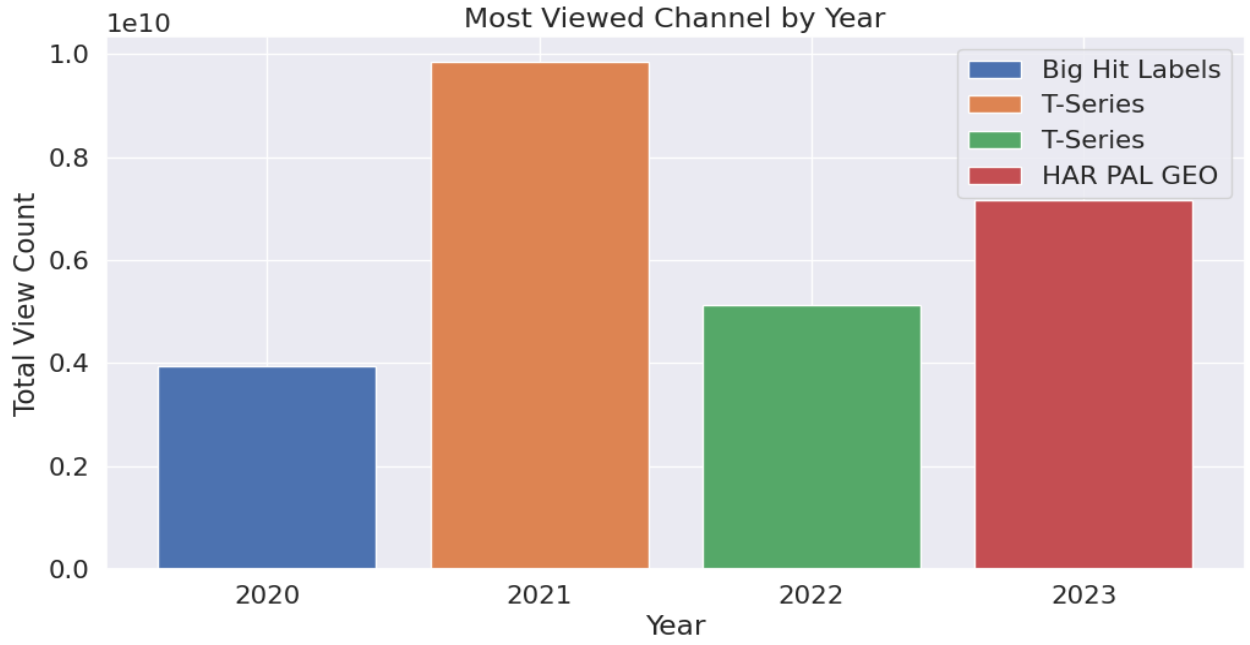
The goal of the combined data preprocessing and EDA process is to transform raw data into a more understandable and insightful format, facilitating a comprehensive exploration of the dynamics of trending videos on YouTube in India. This iterative process of cleaning, transforming, and analysing the data allows for extracting meaningful patterns and trends, providing valuable insights for subsequent analysis sections.

Results

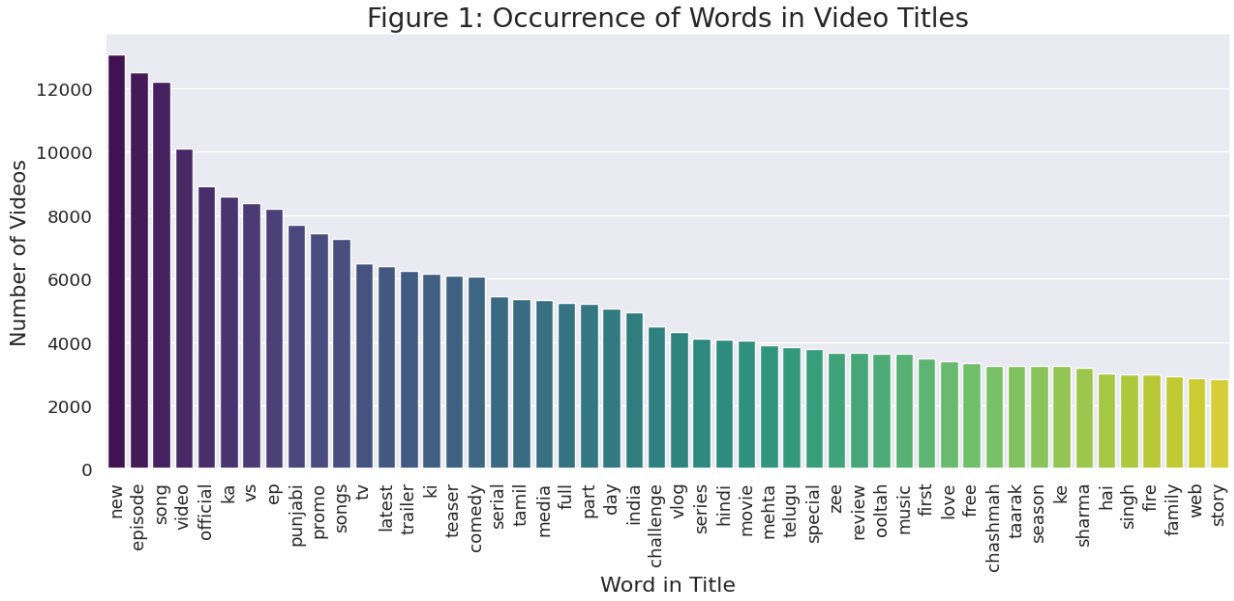
1. Top 20 Most viewed channels



2. Most viewed channel by year

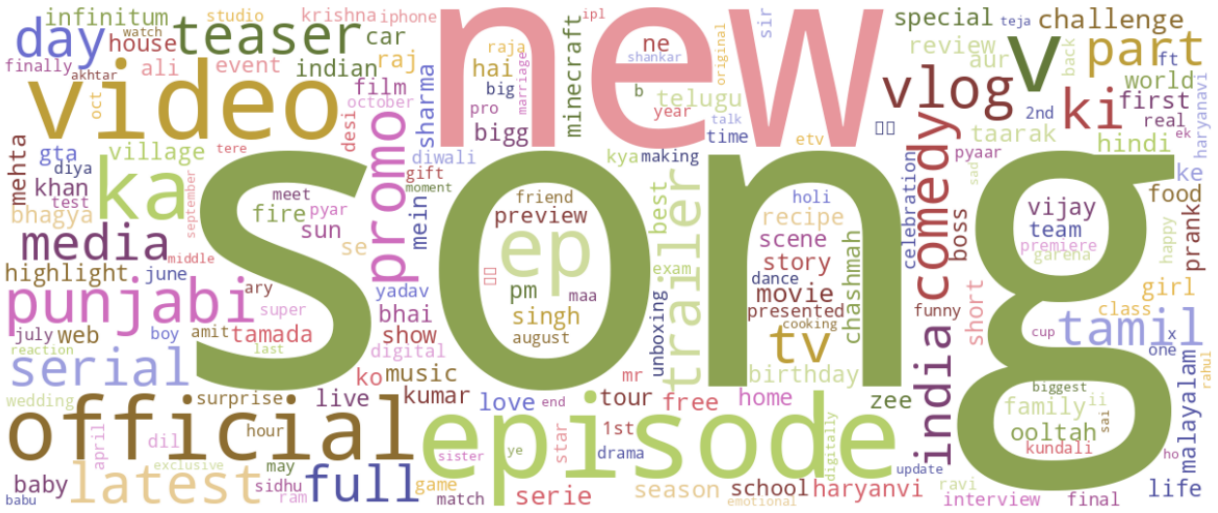


3. Occurrence of Words in Video Titles

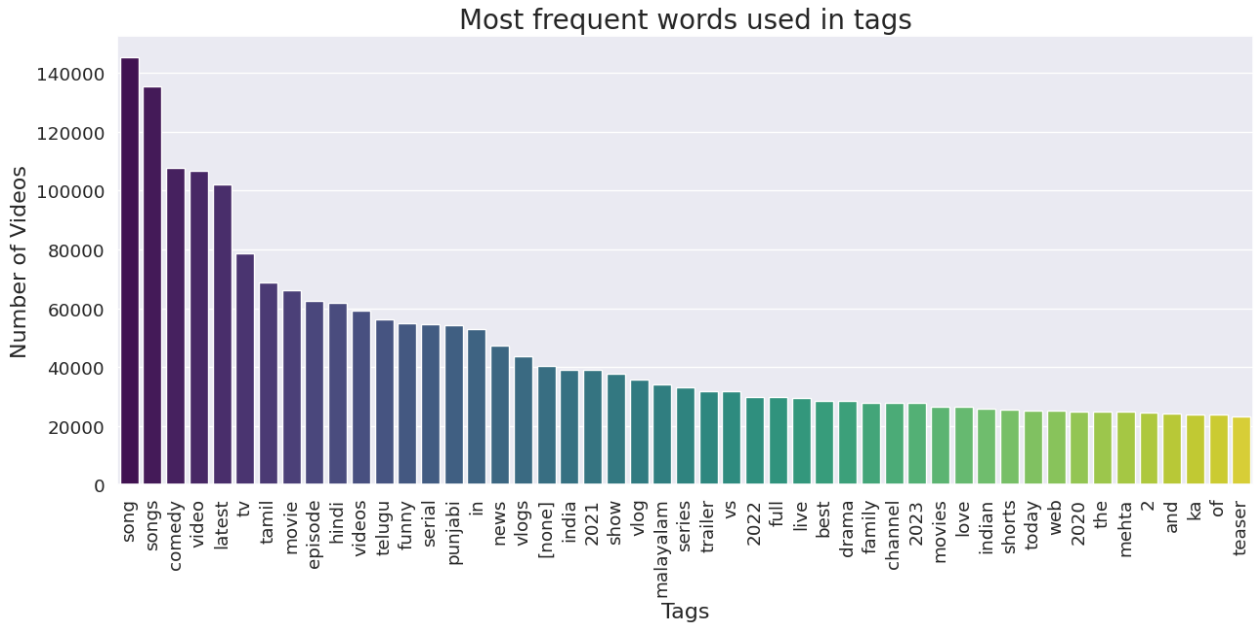


Inference: We can see that the top ten words are new, episode, song, video, official, ka, vs, ep, punjabi, and promo.
It seems that some words quite nicely go together like "new episode", "new song", "punjabi song", "official promo", "new ep", etc…

Let's have a word cloud for this



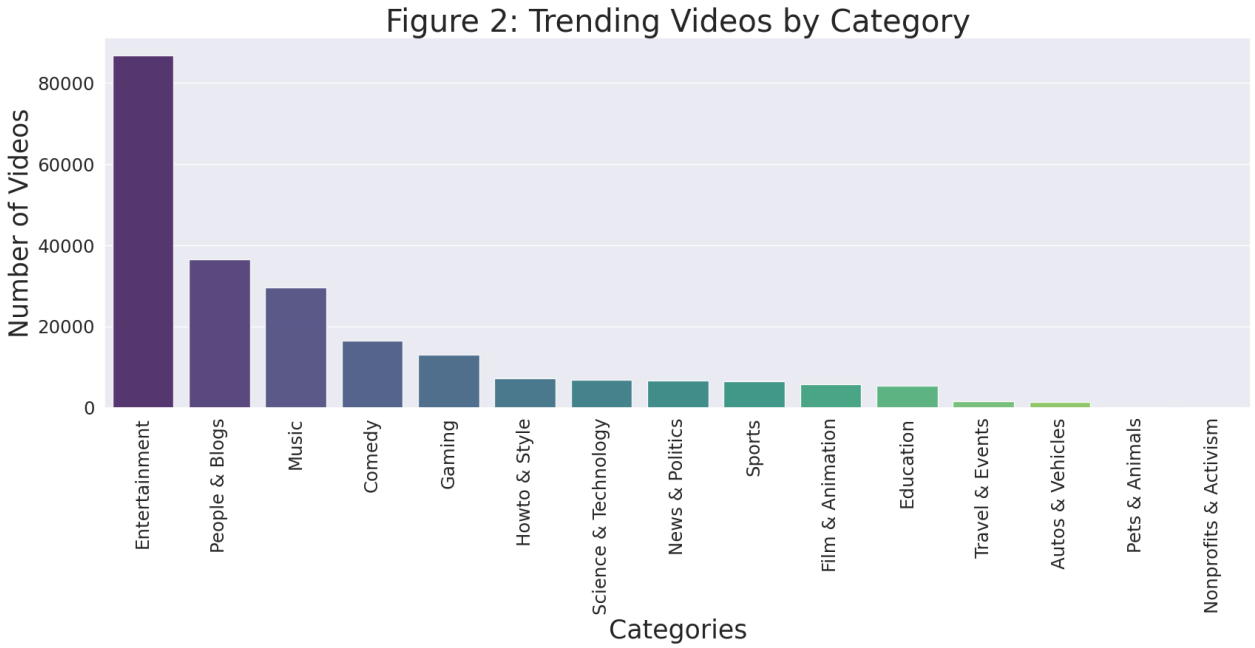
4. Most frequent tags used



So the most frequent words used in tags are song, songs, comedy, video, and latest

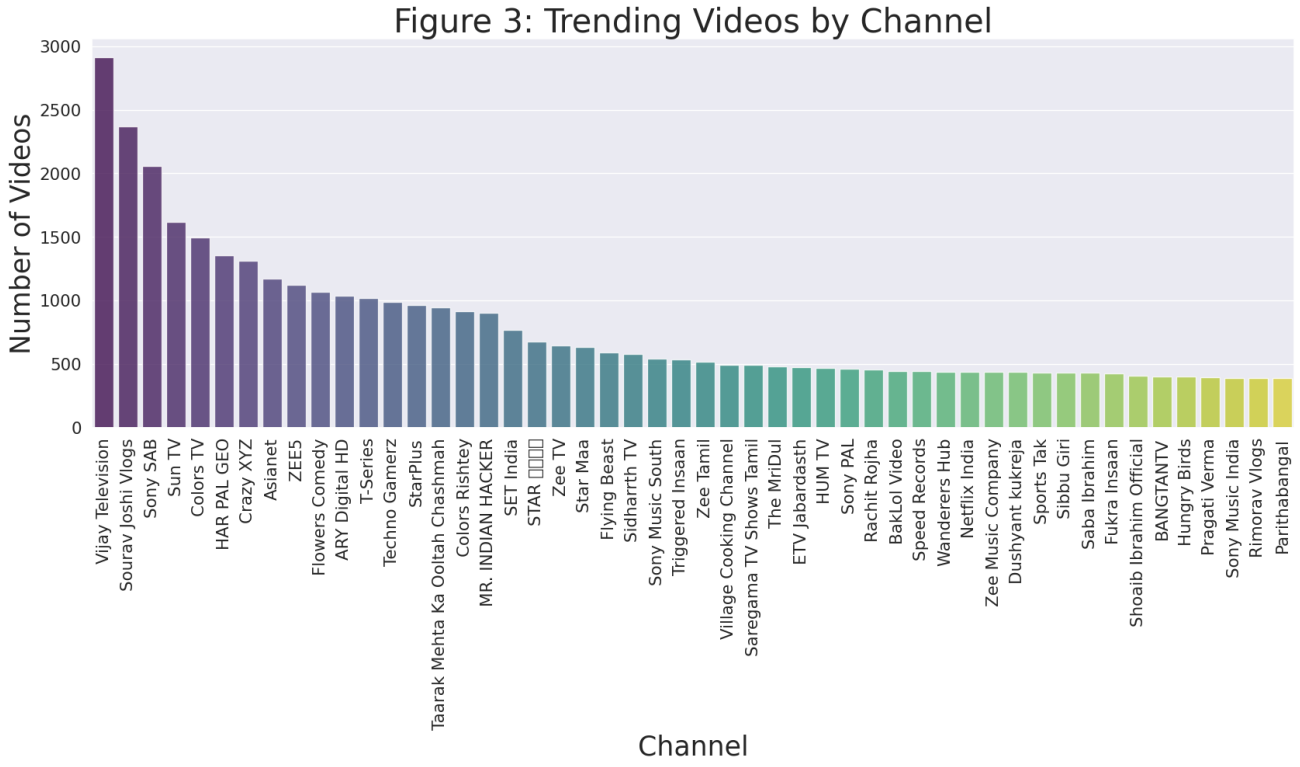
Let's have a word cloud for this as well

5. Videos of which category are dominating in top trending videos?



Videos of Entertainment category dominates the top trending videos.

6. Which channel has the most trending videos?



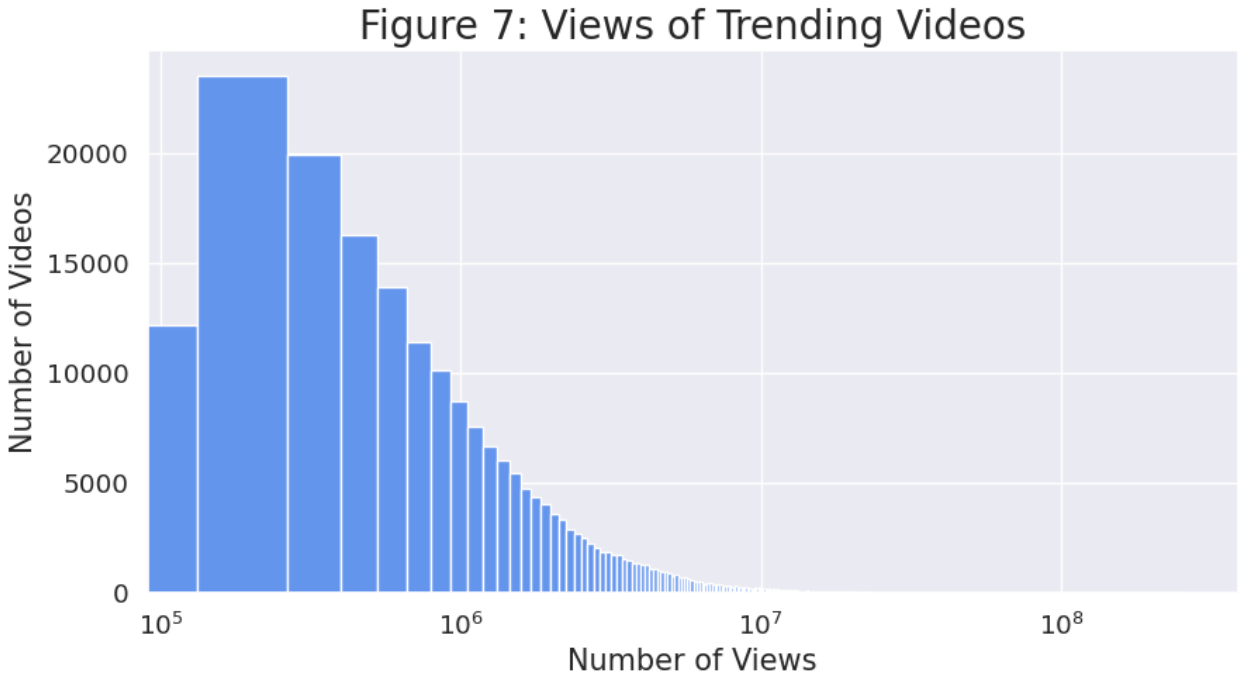
Inference: In result 1, we inferred that T-series has the most number of views but in terms of number of videos released T-series is not even in

the top 10. Releasing a large number of videos or having a large number of videos on top trending videos list does not necessarily secure a large number of views for that channel, it might be one reason but it does not explain the whole story.
But still, T-series does have a large number of videos (~1000) on the trending page, just enough to keep their subscribers' attention.

7. Exploration of views

- Average number of views: **2885734**
- Median number of views: **991707**

With the average being so much larger than the median, this tells that the distribution of views is strongly skewed to the right.



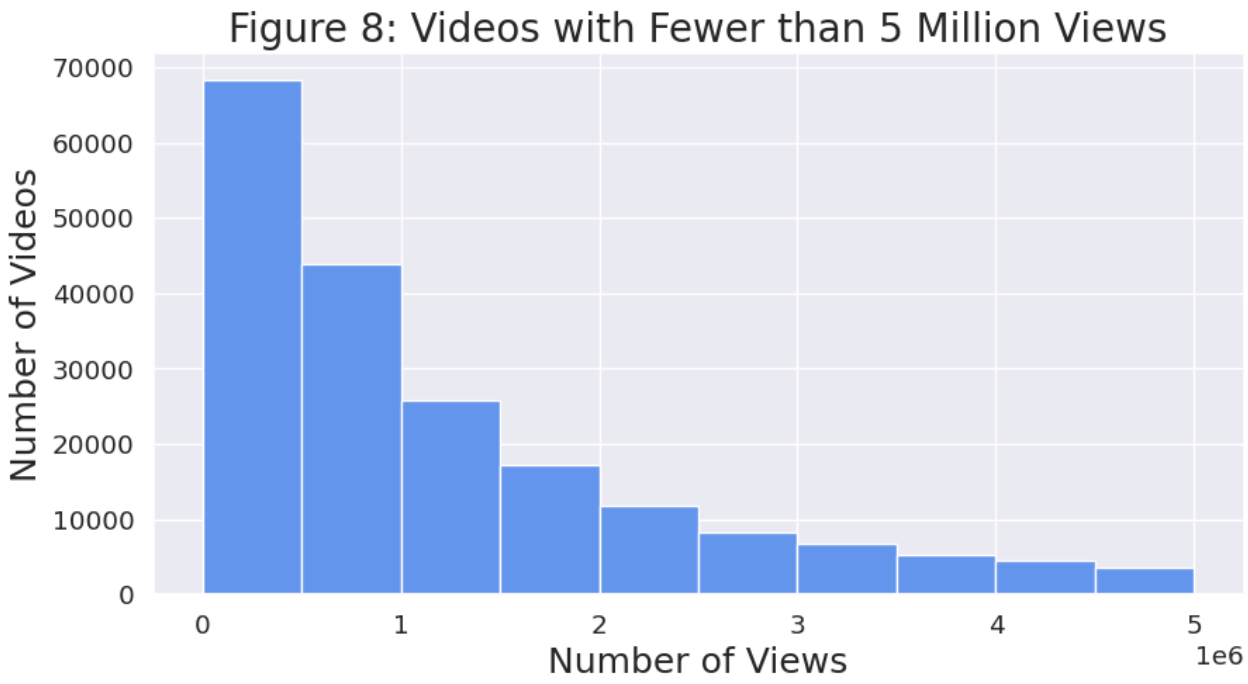
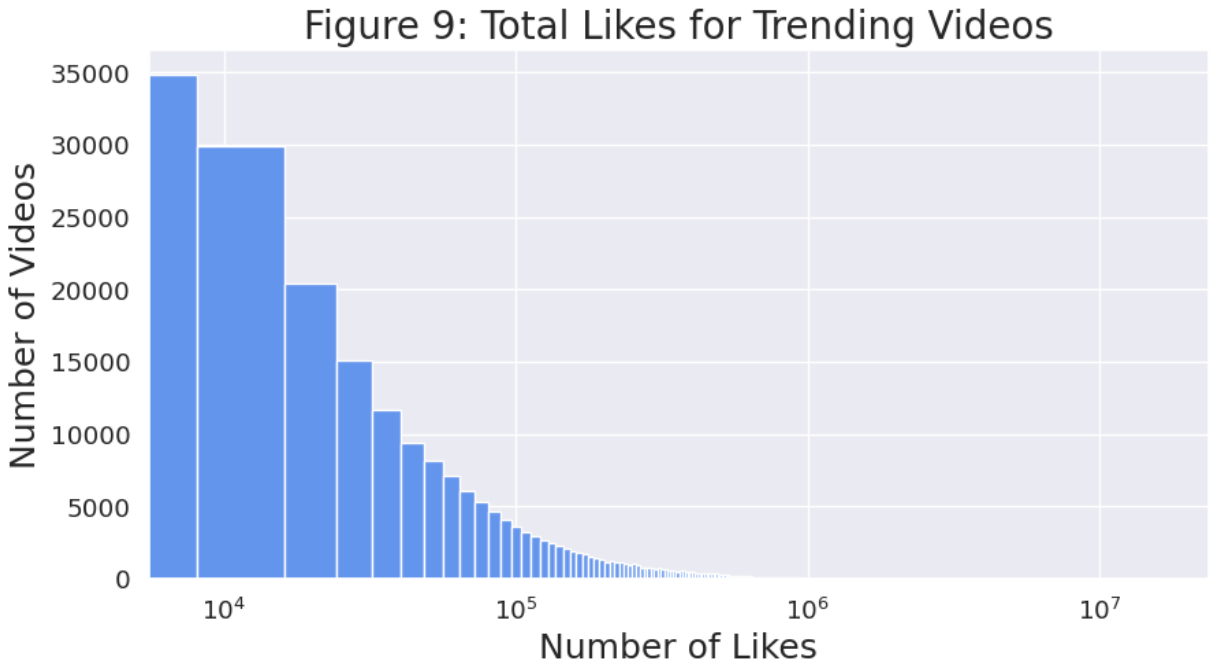


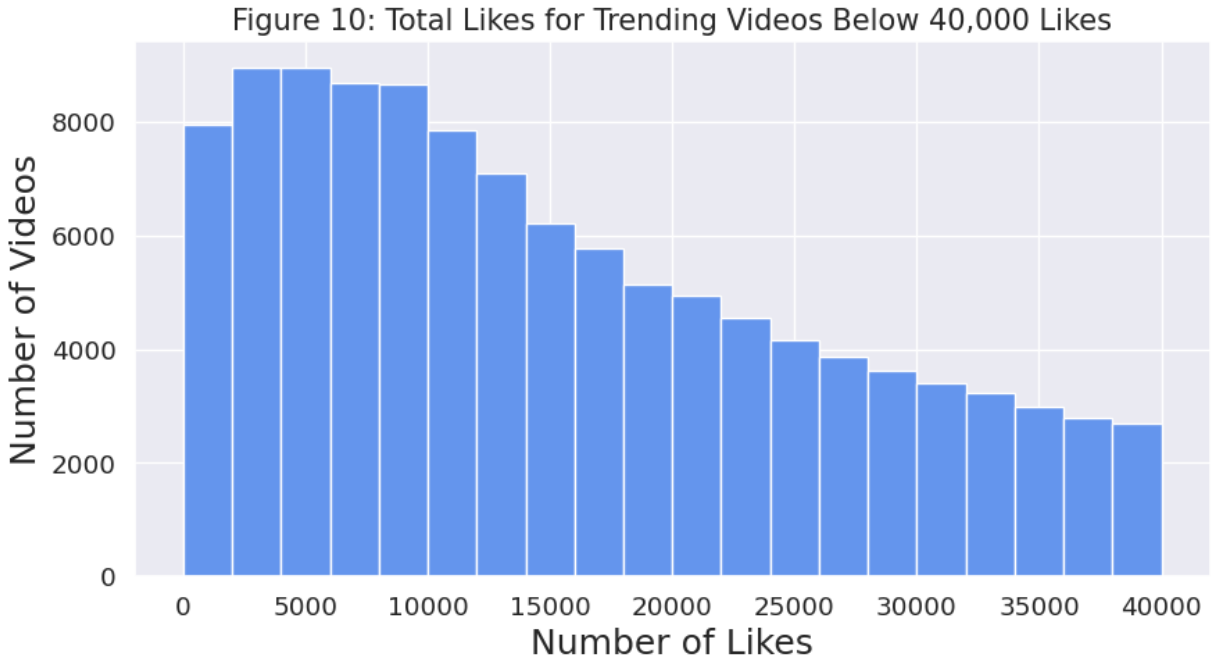
Figure 7 shows that the distribution is very right skewed and shows us that the majority of videos under million. We will plot another histogram to look at the distribution a little bit closer.

Now it is easy to see that the majority of videos have under 5 million views. We will calculate the exact percentage below.

- Percentage of videos with under 5 million views: **87.53%**
- Percentage of videos with under 1 million views: **50.24%**

8. Exploration of likes

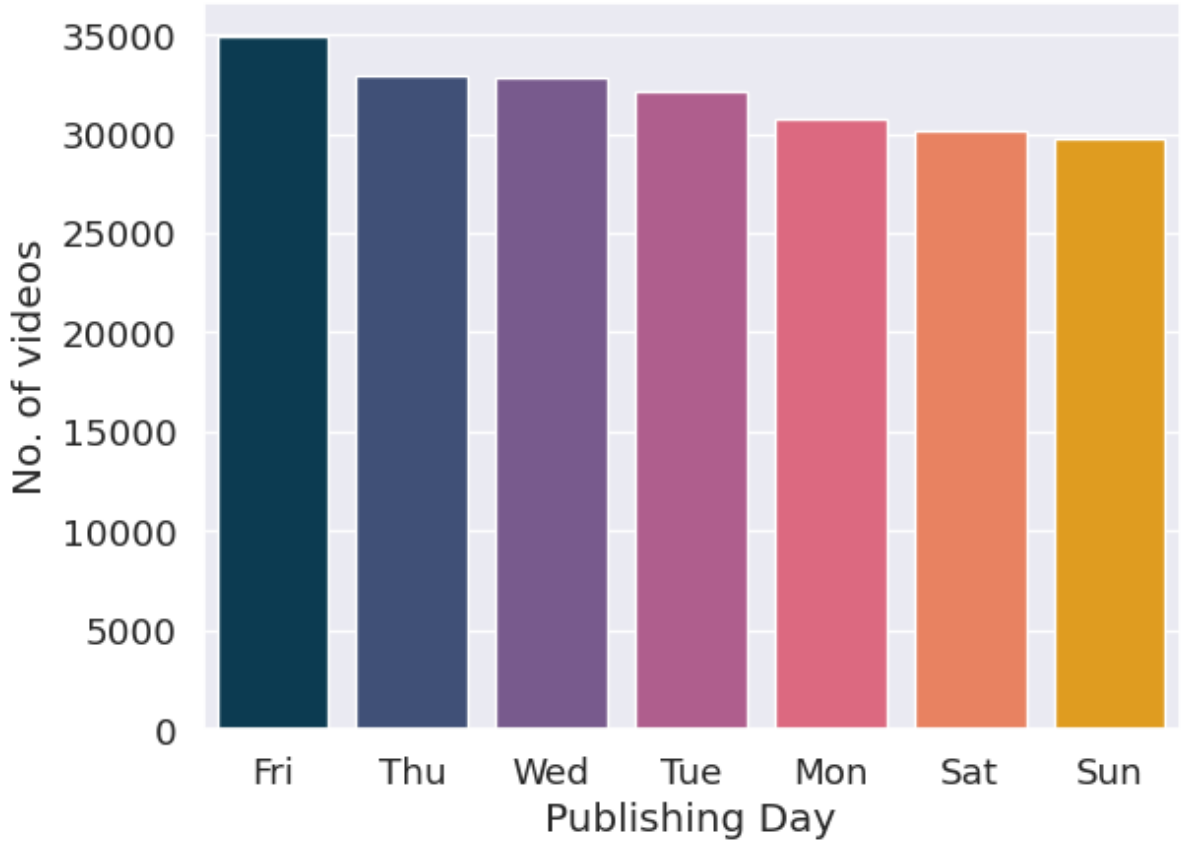




- Average number of likes: **146219**
- Median number of likes: **40163**

This tells us that the distribution of views is strongly skewed to the right, with the majority of videos having below **40,613 likes**.
To visualise the data we will create a histogram of the like data for all the videos.

9. Which day of the week has the most top trending videos published?



Inference: We can observe that most of the videos which are on trending pages are released just before the weekend and surprisingly the least number of videos are released on the weekend.

10. Channel-wise comparison between number of videos, views, ikes, and comments

Figure 11: Trending Videos by Channel

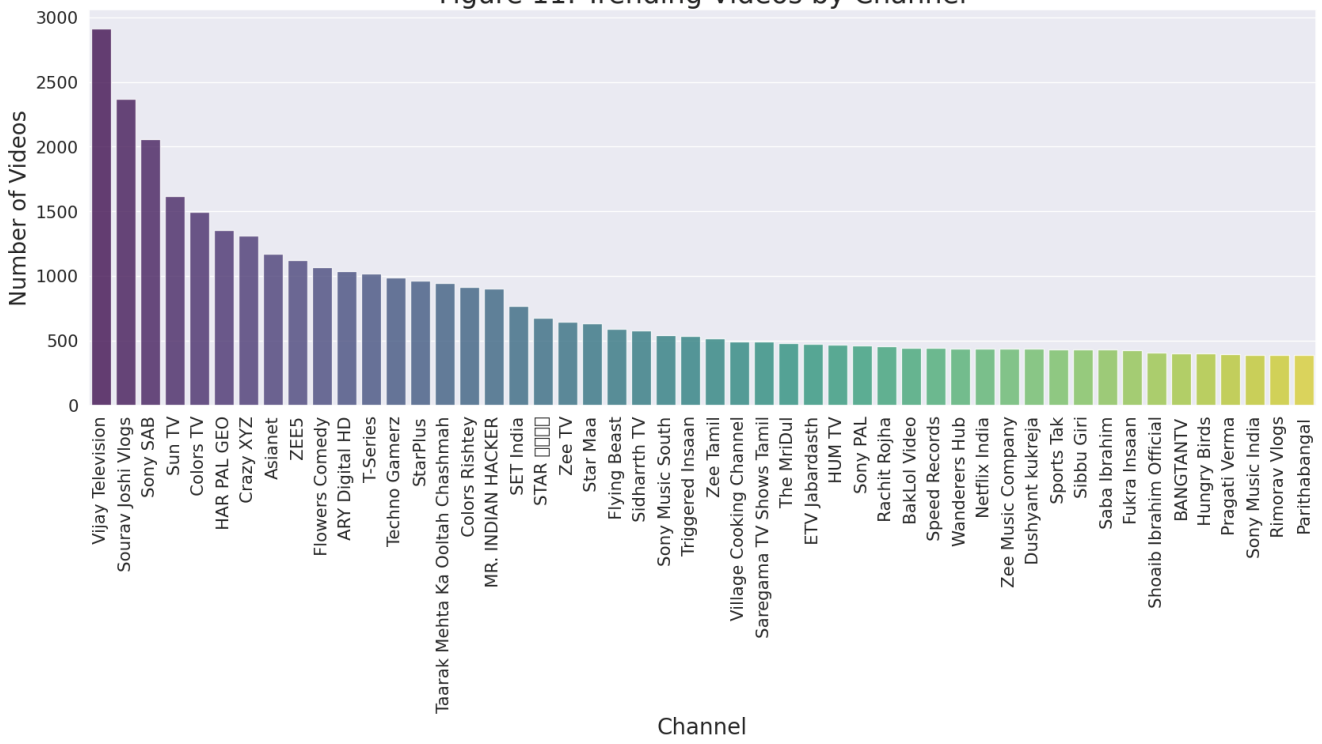


Figure 12: Number of Likes by Channel

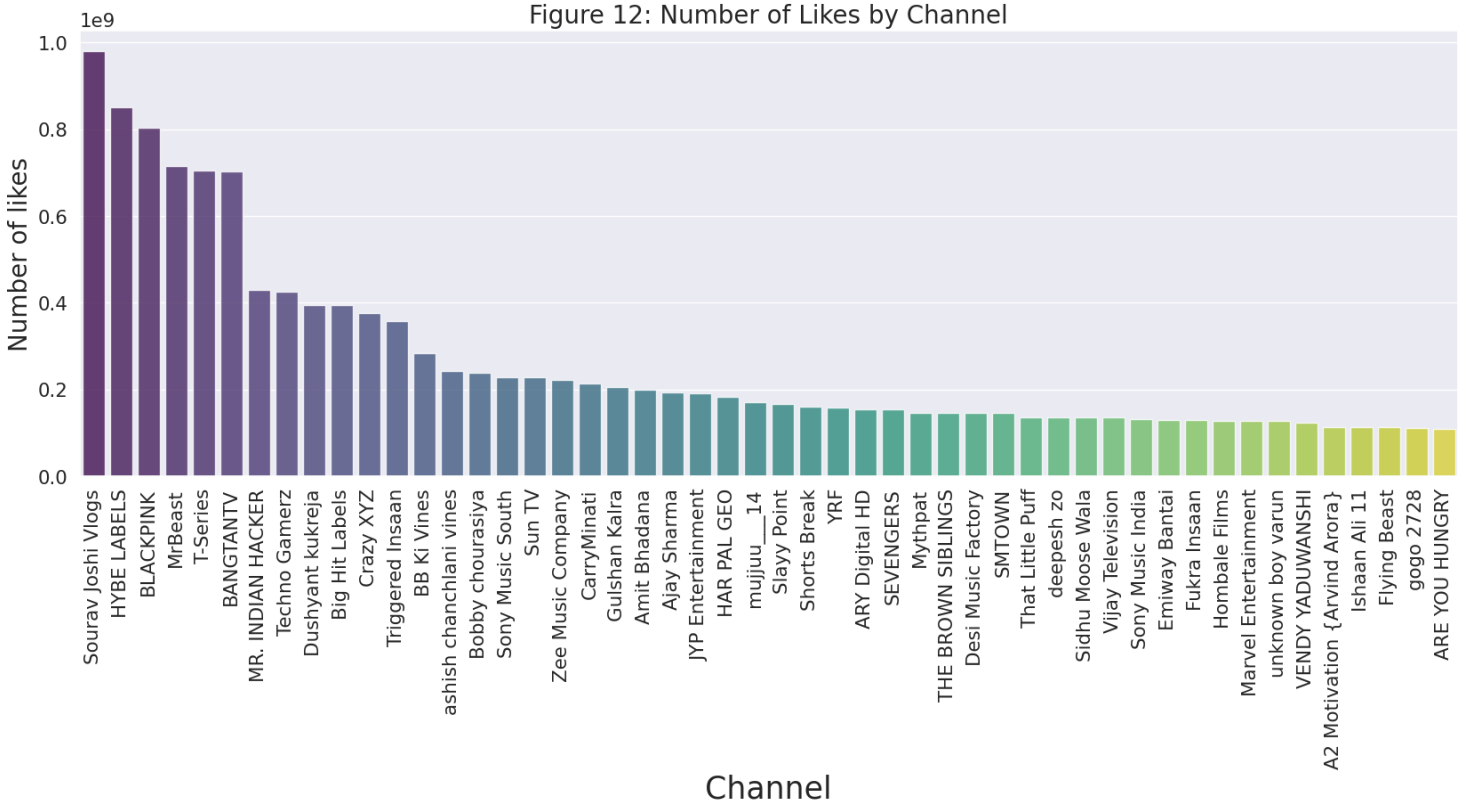


Figure 13: Number of views by Channel

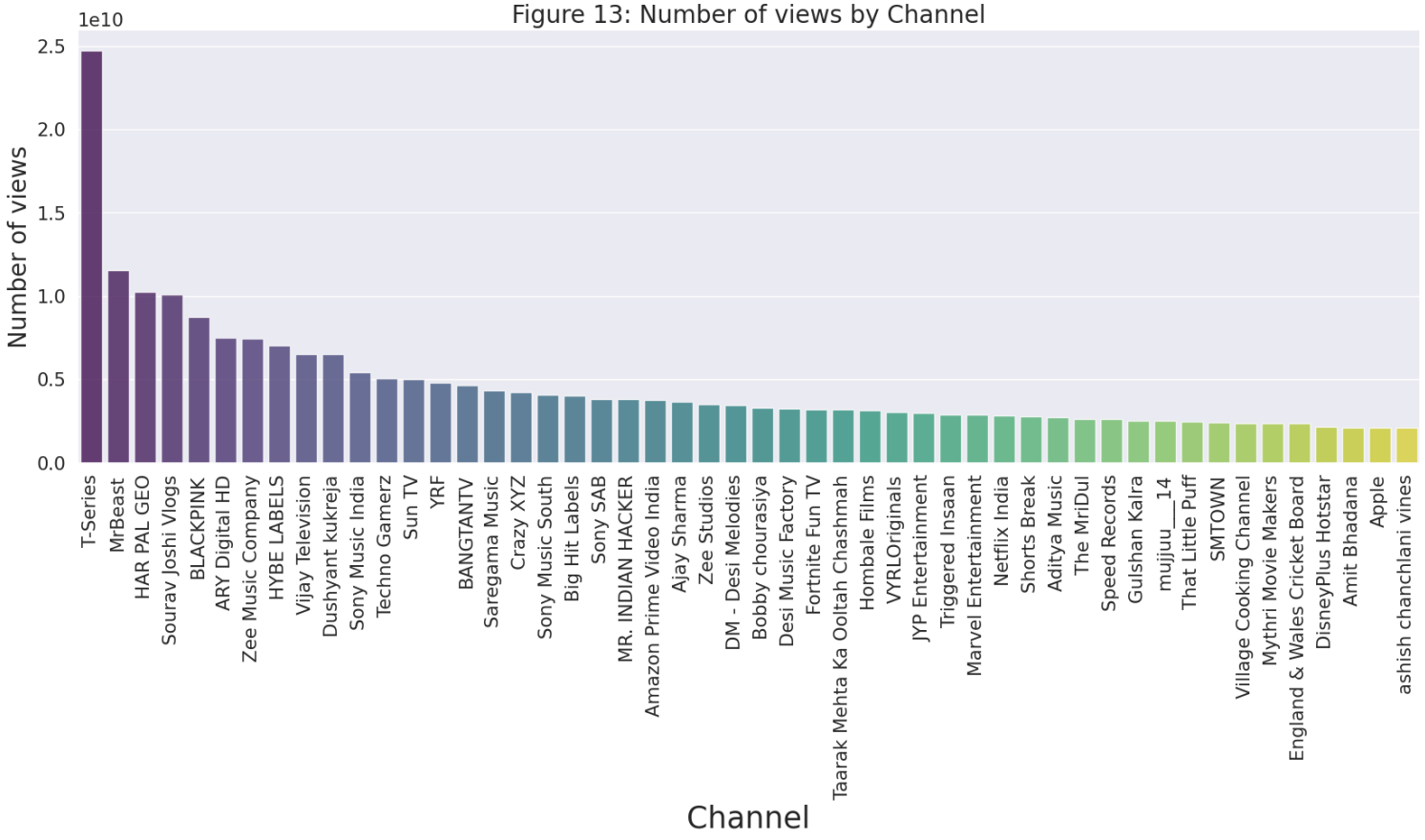
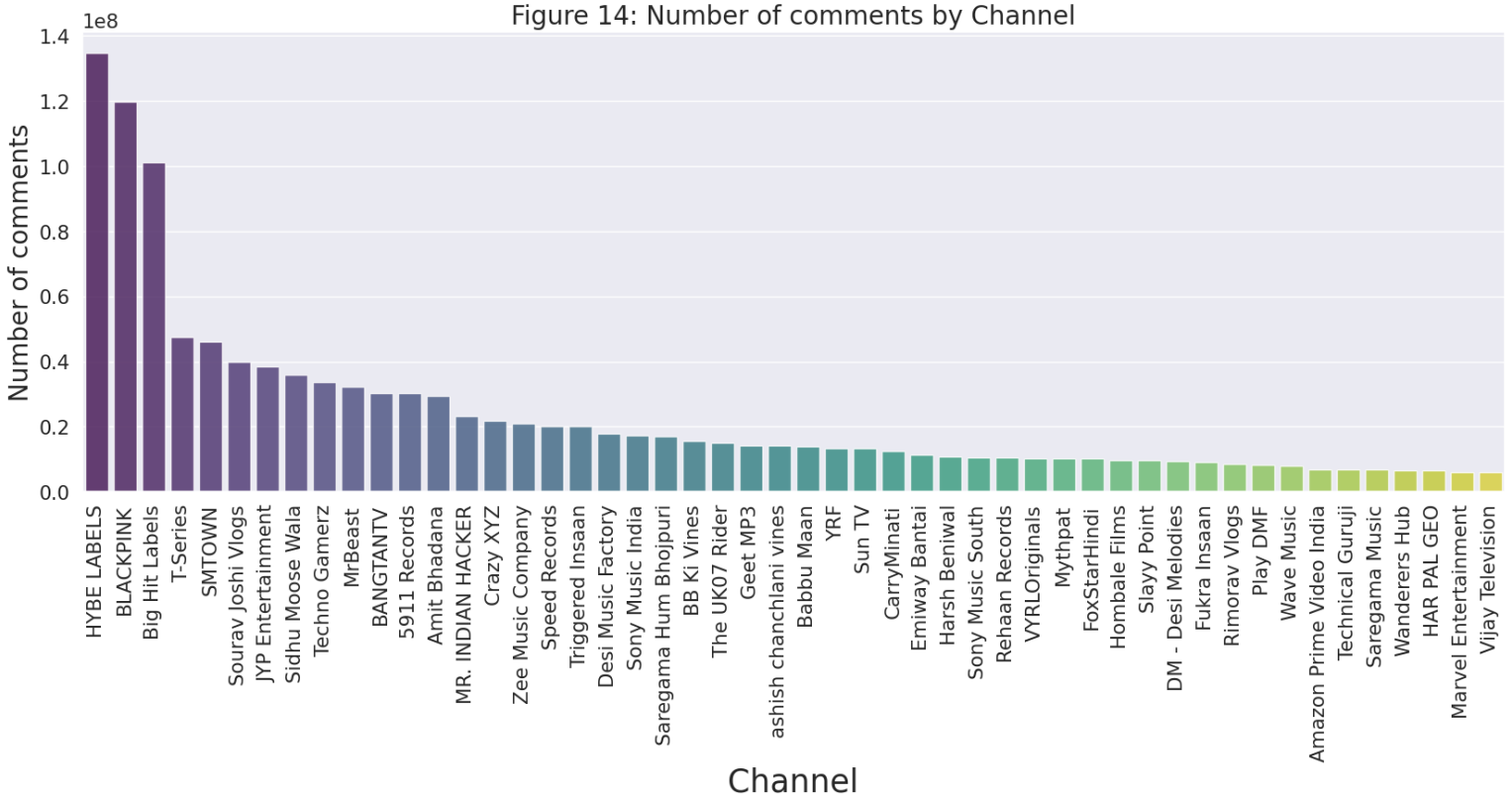
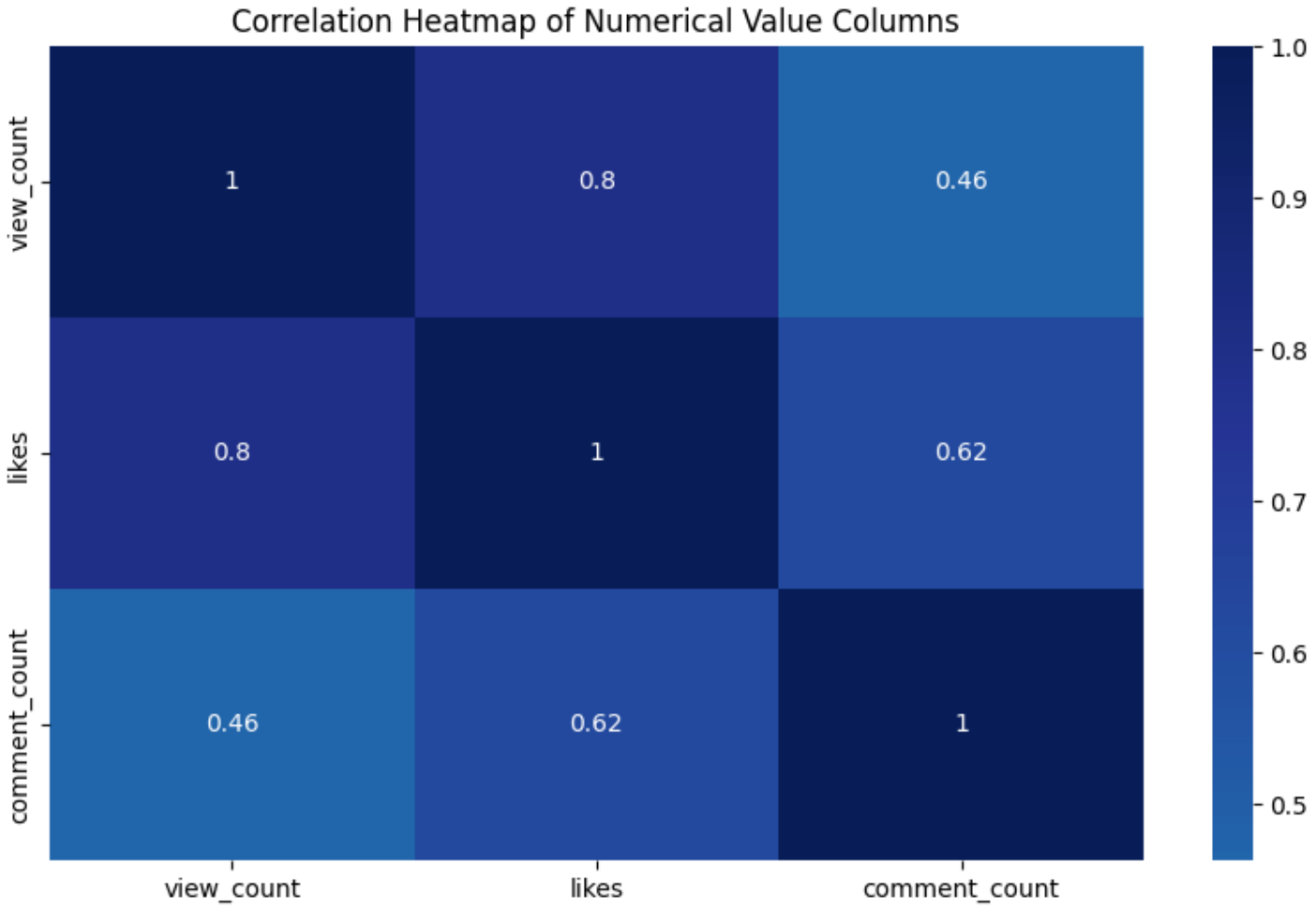


Figure 14: Number of comments by Channel



Analysis

We analyse the above plots by drawing a Correlation Heatmap of the Views Count, Likes and Comments Count.



Inference: The Heatmap derives the following about the correlation values.

- Views Count and Likes have the **max correlation** of **0.8**
- Views Count and Comments Count have **least correlation** of **0.46**
- Likes and Comments Count have correlation of **0.62**

Conclusion

The analysis of the YouTube algorithm for trending videos in India provided valuable insights into various aspects of content creation and viewer engagement. Here are some key findings and conclusions from the project:

- **Top Viewed Channels:** The analysis identified the top 20 most viewed channels, offering content creators and marketers insights into the channels that consistently attract a large audience.

- **Word Usage in Titles:** By exploring the occurrence of words in video titles, the study highlighted common themes such as "new episode," "official promo," and specific genre-related terms. This information can guide content creators in crafting titles that align with viewer preferences.

- **Tag Analysis:** The most frequent tags used in trending videos were identified, including words like "song," "comedy," "video," and "latest." This information can assist creators in selecting relevant tags to enhance video discoverability.

- **Category Dominance:** The Entertainment category emerged as the dominant genre in top trending videos, indicating the popularity of entertainment-focused content among viewers.

- **Channel Performance vs. Views:** While T-series had the most views, the analysis revealed that the number of videos released by a channel does not necessarily correlate with high viewership. This suggests that other factors, such as content quality or viewer engagement, play a crucial role.

- **Distribution of Views, Likes, and Comments:** The distribution of views, likes, and comments exhibited right-skewness, with a majority of videos having views and likes below certain thresholds. This information can guide content creators in setting realistic expectations for engagement.

- **Day of Week for Video Release:** The analysis revealed a trend where most top trending videos are released just before the weekend, indicating strategic timing for video publication.

- **Correlation Analysis:** A correlation heatmap showed that views and likes have a strong positive correlation, while views and comments have least correlation. This information can help creators understand the interplay between different engagement metrics.

In conclusion, this YouTube algorithm analysis provides a comprehensive understanding of the factors influencing video trends in India. Content creators, marketers, researchers, educators, and viewers can benefit from these insights to optimise their strategies, create engaging content, and navigate the dynamic landscape of YouTube. It is crucial to note that ongoing monitoring and adaptation to algorithm changes are necessary for sustained success in the ever-evolving digital space.