Trending YouTube Videos Analysis

Saba Madadi

April 2025

Abstract

To uncover secrets of YouTube trending videos and causes of viral content, this project investigates popular YouTube videos. Examining trends, engagement patterns, and content attributes through EDA and data visualization helps to achieve the goal of this research, which is to provide answers to important questions about the effects of content features, upload times, and video categories on video performance. In addition, creating seven more insightful questions delve deeper into the relationship between engagement levels and video attributes such as titles, tags, and sentiment. In order to investigate variations in engagement across categories, the analysis also addresses statistical tests such as Kruskal-Wallis H test. This report offers insights that are useful not only for comprehending current trends, but also for forecasting what causes a YouTube video to become popular by constructing a coherent narrative around the data.

I. Introduction

YouTube is world's most famous video sharing website which maintains list of top trending videos on platform. According to Variety magazine, "To determine the year's top-trending videos, YouTube uses a combination of factors including measuring user interactions (number of views, shares, comments, and likes). Note that they're not the most-viewed videos overall for the calendar year." Top performers on YouTube trending list include music videos (such as famously viral "Gangnam Style"), celebrity and reality TV performances, and random viral videos that YouTube is well-known for. YouTube is a leading platform for video content in current digital era. However, only a few videos are able to attract attention and trend internationally. This project explores characteristics that set these videos apart by identifying trends and traits that contribute to success of popular YouTube videos. The analysis aims to provide answers to important questions regarding video engagement, content features, and trends using statistical tests, data visualizations, and EDA. The objective is to offer insights that forecast future viral content while explaining existing trends. Main goal of this research is to analyze the key factors that contribute to YouTube videos going viral by examining trends, engagement metrics, and content attributes

from a dataset of trending videos. In the second part, the preprocessing of the dataset and its analysis is discussed to gain a deeper understanding of the data. The third section focuses on the statistical tests and their analysis, with results presented in the fourth section. Finally, the conclusion summarizes the effective factors influencing the analysis of trending YouTube videos.

II. Data Exploration

a. Dataset

- Before Extra Data

(Note: Till end of the main project, this dataset have been used.) The dataset has information related to trending videos in 10 countries. It includes several months of data on daily trending YouTube videos with up to 200 listed trending videos per day. In table 1, ten countries are listed along with number of data entries available from each in the dataset (Each region's data is in a separate file.):

Table 1: Countries and Data Entries

Country Name	**Abbreviation**	**Number of Entries**
Canada	CA	40881
Germany	DE	40840
France	FR	40724
United Kingdom	GB	38916
India	IN	37352
Japan	JP	20523
South Korea	KR	34567
Mexico	MX	40451
Russia	RU	40739
United States	US	40949

Also, features of the dataset and their description has shown on Table 2:

Table 2: Features and Descriptions

Feature Name	Description
video_id	Unique identifier for the video
trending_date	Date when the video became trending
title	Title of the video
channel_title	Name of the channel that uploaded the video
category_id	Identifier for the video category
publish_time	Date and time the video was published
tags	Tags associated with the video
views	Total number of views
likes	Total number of likes
dislikes	Total number of dislikes
comment_count	Number of comments on the video
thumbnail_link	Link to the video thumbnail
comments_disabled	Indicates if comments are disabled
ratings_disabled	Indicates if ratings are disabled
video_error_or_removed	Indicates if the video is removed or has an error
description	Description of the video

This dataset was collected using the YouTube API. Additionally, there is a JSON file available for each country. In the EDA section, two important pieces of information will be extracted from this file: one related to the category of each video and the other concerning whether the category is assignable or not, which will be explained further. This section focuses solely on the details of the JSON file, as shown in Table 3:

Table 3: JSON Data Table

Field	Description
`id`	The unique identifier of the video category
`title`	The title of the video category
`assignable`	Indicates whether this category can be assigned to channels
`channelId`	The identifier of the channel associated with this category

- After Adding Extra Data

(Note: Following at marked section this dataset have been used.)

b. EDA

Main goal of this research is to analyze key factors that contribute to virality of YouTube videos by examining trends, engagement metrics, and content

features. This analysis aims to provide insights into elements that influence video trending, helping video creators understand how to prepare their videos for increased views in future. Initially, as shown in Tables 2 and 3, the ID serves as link between CSV and JSON files in dataset. Therefore, based on ID present in category column of CSV file, corresponding category from the JSON file is extracted and added to CSV file as category_title column. A DataFrame named df was created to hold all data globally, containing 375,942 entries. The DataFrame has a 'country' column, which indicates country of each data entry. To check if all data entries have assigned categories, looked for null values in 'category_title' column. Was found that there are 2,795 null values. Next, examined which 'category_id' values corresponded to null 'category_title'. It turned out that 'category_id' 29 did not have a category title. After researching online, using information from this link, discovered that 'category_id' 29 is associated with category "Nonprofits and Activism." It was found based on surfing on internet that category "Nonprofits and Activism" is assignable. A review of JSON file showed that categories are divided into two types: assignable and non-assignable. Assignable categories can be chosen by content creators when uploading videos, allowing to help viewers understand content better. Non-assignable categories cannot selected by creators during upload process; these categories are generally used for specific types of content that YouTube automatically categorizes. Tables 4 and 5 provide detailed information about these categories:

category id	**category title**
1	Film & Animation
2	Autos & Vehicles
10	Music
15	Pets & Animals
17	Sports
19	Travel & Events
20	Gaming
22	People & Blogs
23	Comedy
24	Entertainment
25	News & Politics
26	Howto & Style
27	Education
28	Science & Technology
29	Nonprofits & Activism

Table 4: Assignable Categories

category id	category title
18	Short Movies
21	Videoblogging
30	Movies
31	Anime/Animation
32	Action/Adventure
33	Classics
34	Comedy
35	Documentary
36	Drama
37	Family
38	Foreign
39	Horror
40	Sci-Fi/Fantasy
41	Thriller
42	Shorts
43	Shows
44	Trailers

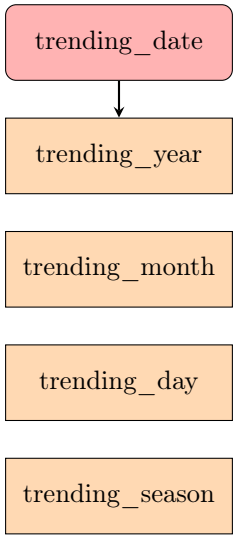
Table 5: Non-Assignable Categories

It was found that 3.34% of df was duplicated. These duplicate entries were removed. Additionally, duplicates from each country dataset were dropped because they made up a small percentage of data. Removing them resulted in cleaner data. By looking at a few rows from the beginning and the end, a general understanding of datasets was obtained. The check for missing values in df showed that only 'description' column has null values. Since this column is not analyzed in main part of research, handling this issue will be moved to specific section for 'description,' which is located at end of this section. Next, since category of each video has been determined, the 'category_id' column can be dropped. Also 'video_id' and 'thumbnail_link' columns can be dropped. DataFrame df contains a column named country. This allows for easy access to data for a specific country. Table 6 outlines features of df dataset that will be called only dataset furthermore.

Table 6: Features and Descriptions

Feature	**Description**
trending_date	The date the video was trending.
title	The title of the video.
channel_title	The name of the channel that uploaded the video.
publish_time	The date and time the video was published.
tags	Tags associated with the video.
views	The number of views the video has received.
likes	The number of likes the video has received.
dislikes	The number of dislikes the video has received.
comment_count	The number of comments on the video.
comments_disabled	Indicates whether comments are disabled.
ratings_disabled	Indicates whether ratings are disabled.
video_error_or_removed	Indicates if the video has errors or has been removed.
description	A brief description of the video's content.
category_title	The category of the video.
country	The country where the video was uploaded.

First, valuable information is extracted from trending_date column, which is in format %y.%d.%m. After this, trending_date column will be dropped. The new column trending_year is created by adding 20 to the year, since, for example, the year 2017 is written as 17 in the original format. The new columns trending_month, trending_season, and trending_day represent the month, season, and day of the week when video became trending.



Information from video title can be useful. The length of title is calcu-

lated and a new column named title_length added to dataset. Additionally, sentiment analysis using NLP is performed on title, and a new column called title_sentiment added. Finally, a boolean column added to indicate whether title contains clickbait words that might entice viewers to click. The SentimentIntensityAnalyzer used to analyze sentiment of video titles. A function created to calculate sentiment score. If score is greater than or equal to 0.05, sentiment labeled as positive. If score is less than or equal to -0.05, it is labeled as negative. Otherwise, sentiment is considered neutral.

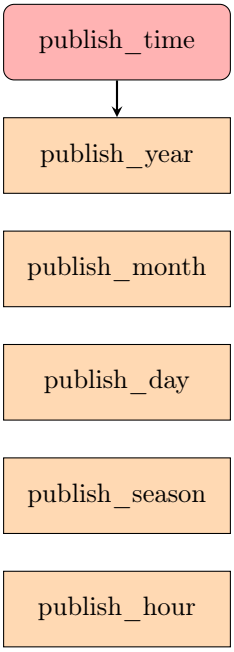
Table 7: Title Related Features and Descriptions

Feature	Description
title_length	Length of video title measured in characters.
title_sentiment	Sentiment of title categorized as positive, negative, or neutral.
clickbait	A boolean value indicating whether title contains enticing keywords to attract clicks.

Table 8: Clickbait Keywords

#	Keyword
1	shocking
2	must watch
3	amazing
4	incredible
5	unbelievable
6	you won't believe

The same process has done for 'channel_title' column. So, channel_title_length, channel_title_sentiment and channel_clickbait added to dataset. The publish_time column shows the date and time a video was published. Information is extracted regarding the year, month, season, hour, and day of the week when the video was published.



For handling the extract_tags function splits a string of tags into a list, removing any quotation marks. The extracted tags are then stored in a new column called extracted_tags in dataset.

For engagement, the total number of likes and comment_count is summed up, and for engagement_level, three categories are defined: Low, Medium, and High.

Now, all columns of dataset preprocessed and are ready to help to gain main goal of this research (analyzing important effective reasons for trending videos to help new Youtubers now what to publish to be trend!

Description As told above, description column has missed values so first handled missing ones and then using non_null_descriptions dataframe that contains description_sentiment and description_length columns.

E1: Is description sentiment related to engagement level?

In this study, the main goal is to examine engagement, as it shows how many views and likes there are. The relationship between engagement and the sentiment of the description is explored. Figure 1 shows that positive sentiment has a higher engagement level, indicating that using positive words in the description affects viewers' likes and views. The bar plot indicates a larger share for high engagement.

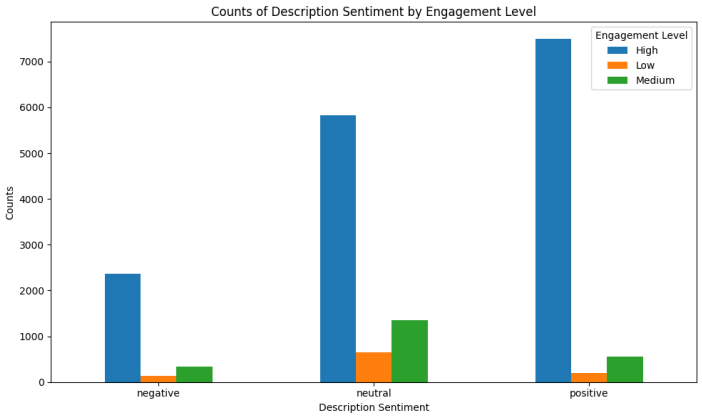


Figure 1: Description Sentiment vs. Engagement Level

Figure 2 presents a pie chart to analyze the percentage of high engagement levels. The analysis reveals that 83% is high, about 12% is medium, and around 5% is low.

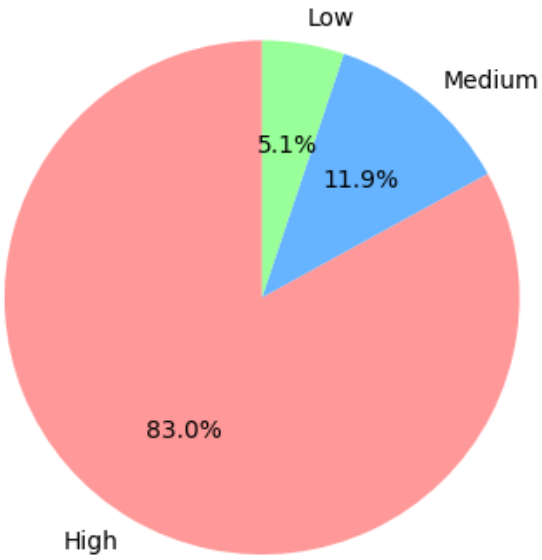


Figure 2: Engagement Levels

Additionally, the donut plot in Figure 3 represents description_sentiment, created to provide a detailed analysis of the previous question. As seen, 43.6%

is positive, 41.4% is neutral, and 15% is negative.

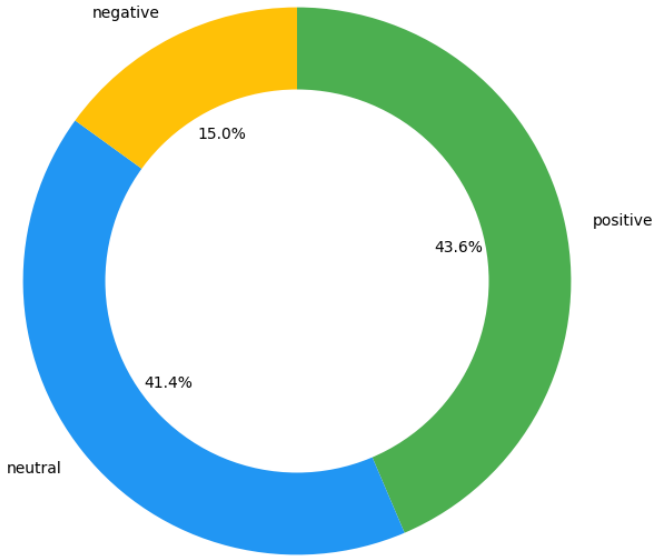


Figure 3: Description Sentiment

E2: Does description length impact engagement levels?

The average description_length is about 980 characters. According to Figure 4, the violin plot shows that there is no significant difference, indicating that the length of the description does not have much impact on views or likes.

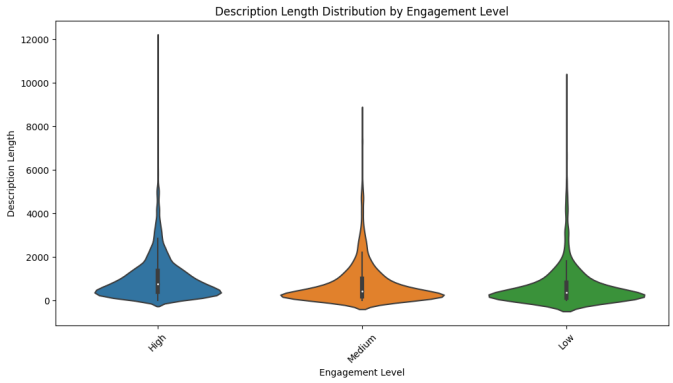


Figure 4: Description Length vs. Engagement Level

In following, 3 tables categorized by data type can be seen:

Table 9: Numerical Data

Feature
views
likes
dislikes
comment_count
trending_year
trending_month
title_length
channel_title_length
publish_year
publish_month
publish_hour
engagement

Table 10: Categorical Data

Feature
category_title
trending_day
trending_season
title_sentiment
publish_day
engagement_level

Table 11: Boolean Data

Feature
comments_disabled
ratings_disabled
video_error_or_removed
clickbait
channel_clickbait

Table 12: string Data

Feature
title
channel_title
description
country
extracted_tags

Given that it was previously concluded that description length does not significantly affect engagement level, it is interesting to investigate whether title length has any impact on engagement level. Figure 5 shows that there is not much difference in this case either, indicating that neither description length nor title length has a significant effect on attracting an audience, as viewers do not pay much attention to this aspect. (Q6)

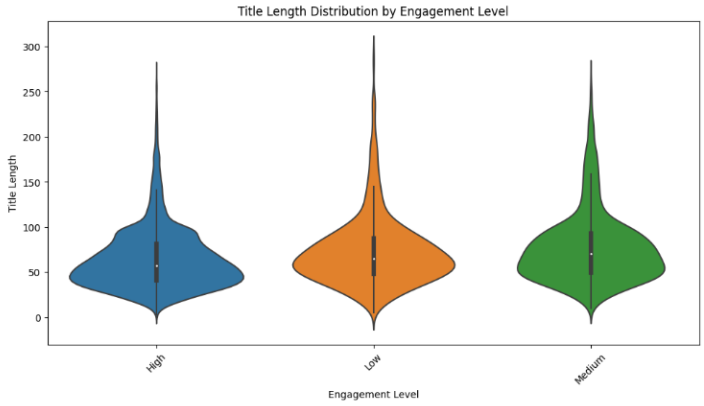


Figure 5: Title Length vs. Engagement Level

For a more detailed analysis, Figure 6 shows that the average title length is about 66 characters, which, as noted in the previous analysis, does not significantly affect the engagement level.

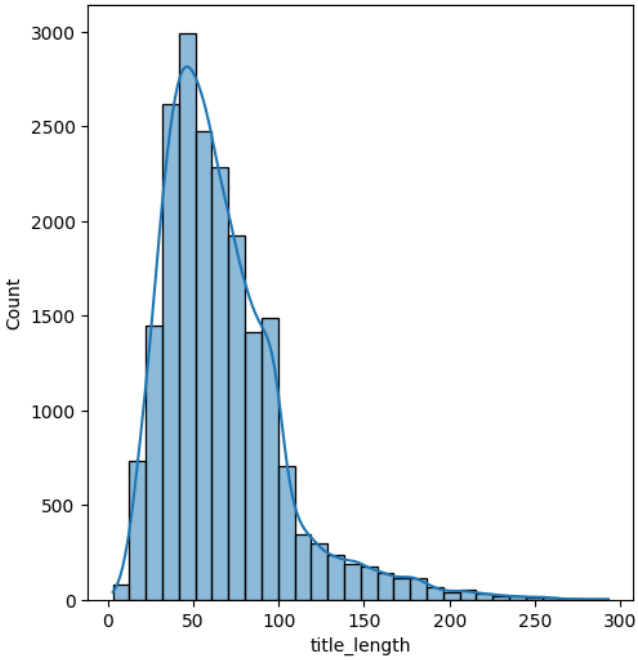


Figure 6: Title Length

Parsing tags checks if the input is a list, tuple, or string, and converts it into a list of tags. If the input is a string, it attempts to evaluate it as a literal structure or splits it by commas. After extracting the tags, calculates the number of tags for each entry. It also computes the like ratio and dislike ratio by dividing the number of likes and dislikes by the total views. This analysis helps understand audience engagement better.

As mentioned, the main goal of this research is to examine the factors influencing the trending of videos to assist new YouTubers. For this purpose, engagement metrics such as likes, views, and dislikes are displayed in the figure 7. On average, views are 1,349,209, likes are 40,150, and dislikes are 2,145. To enhance the analysis, these averages will be examined across different categories.

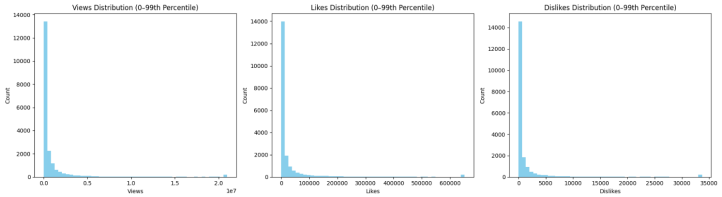


Figure 7: Engagement Metrics

Based on Figure 8, examination of the various categories reveals that music leads by a significant margin, with the majority of viewers preferring music content. Additionally, the high number of music-related videos can explain the elevated dislike counts; however, it is important to note that the scales of the charts differ. As observed in the previous figure, the number of dislikes is negligible compared to views and likes. A key point of interest is that despite the Nonprofits and Activism category having low likes and views, it ranks second in dislikes, indicating that it is not a popular category. Furthermore, analysis of all three charts shows that film, comedy, and science and technology also enjoy popularity. (Q1)

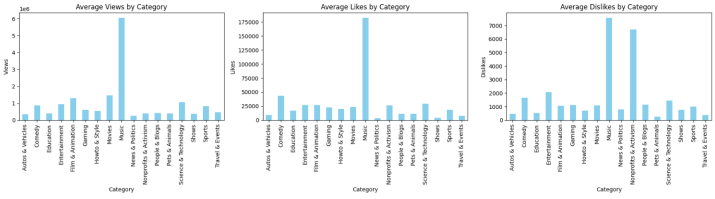


Figure 8: Engagement Metrics vs. Categories

As observed, categories have a significant correlation with engagement. Therefore, the next step is to examine which channels and categories are trending in each country, followed by an overall analysis of this issue. Initially, the top five channels for each country can be seen in below (also top 5 channels worldwide highlighted):

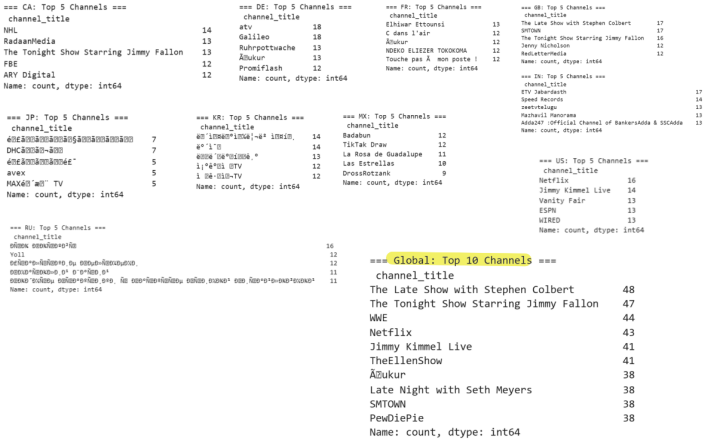


Figure 9: Top Channels

Additionally, the three trending categories in various countries are illustrated in the figures below. Finally, the plot depicting the top five trending categories

globally can be observed. (Q 2)

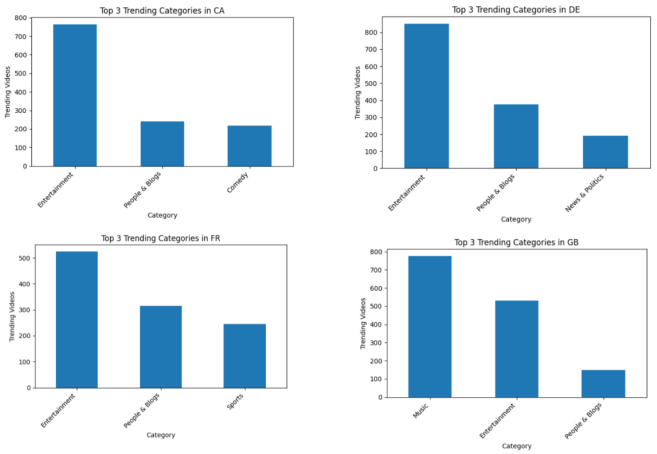


Figure 10: Trending Categories in Countries Part 1

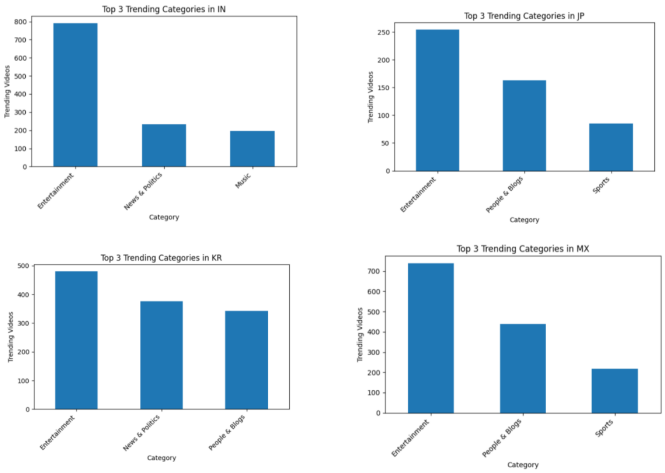


Figure 11: Trending Categories in Countries Part 2

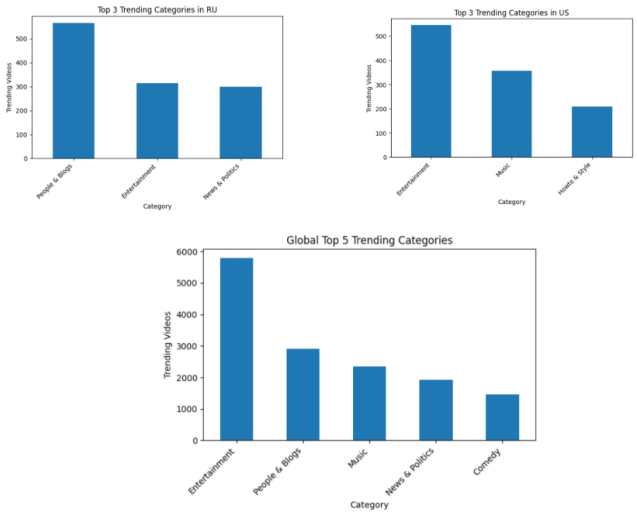


Figure 12: Trending Categories in Countries Part 3

It is interesting to examine on which day of the week and in which season the trending videos have been published. The plots below indicate that the day of the week has little impact on the trending status of a video, although there is a slight increase during the weekends. However, it is clearly observable that spring and winter have significantly more trending videos compared to autumn and summer.

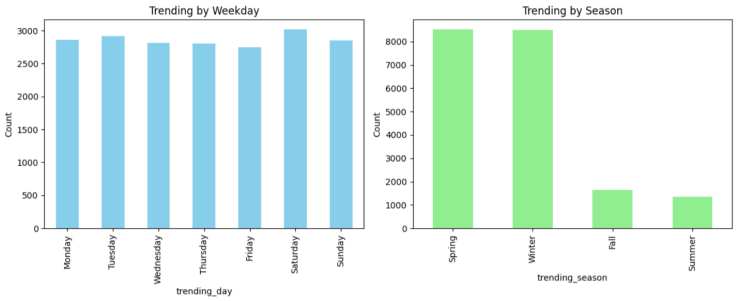


Figure 13: Trending Day and Season

Additionally, an analysis of the upload day and hour of videos shows that around 4 PM and Friday are the peak times. Knowing this can be very beneficial for new YouTubers. (Q3)

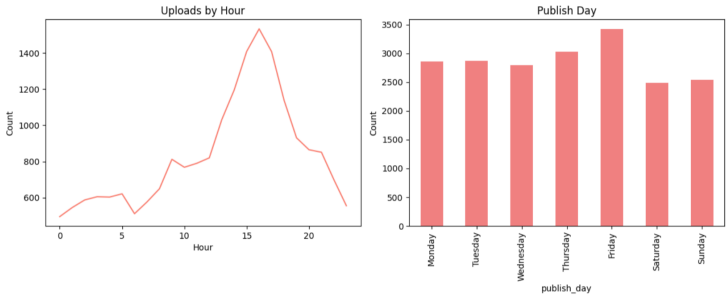


Figure 14: Upload Day and Hour

In the following analysis, a deeper examination of the engagement rate will be conducted. Do controversial videos, defined by a high dislike ratio, receive more engagement than universally liked ones? (Q4) As observed, the views for universally liked videos are higher; however, the number of likes and dislikes is also lower in this category. This suggests that while universally liked videos attract a larger audience, they may not provoke as much discussion or emotional response compared to controversial videos.

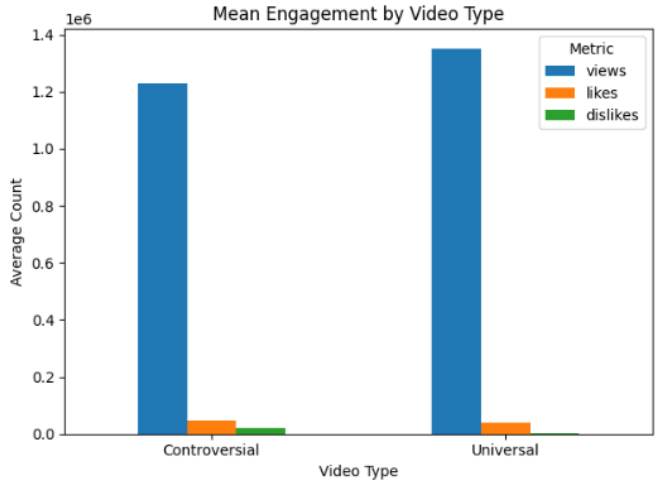


Figure 15: Universally vs. controversial

It is advisable to explore the tags used in trending videos. The table below displays the top 20 trending tags:

Tag	Count
[none]	2056
funny	975
comedy	850
2018	593
news	499
music	473
video	354
trailer	353
pop	327
tv	314
2017	305
television	304
humor	279
rap	277
review	273
vlog	269
live	268
show	247
entertainment	246
diy	244

Table 13: Top 20 Trending Tags

Additionally, the plot below illustrates the relationship between the number of tags used and engagement. As observed, there is an initial increase in engagement with a moderate number of tags, but this trend quickly tapers off. (Q5)

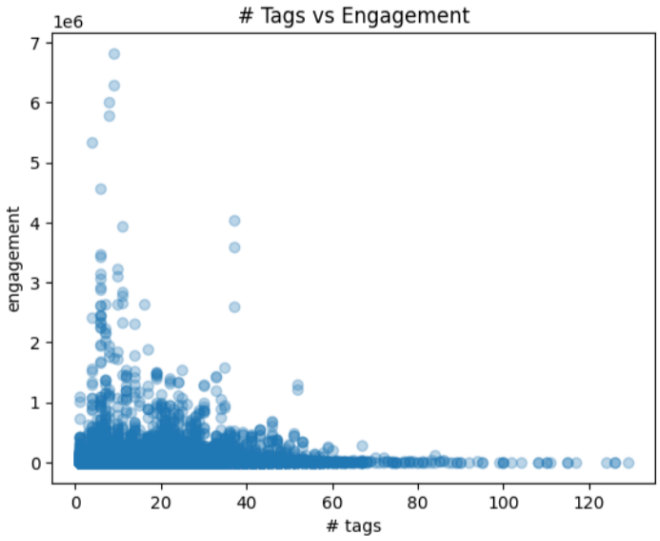


Figure 16: Tags vs. Engagement

Also, analysis reveals a clear relationship between video title sentiment and engagement levels. The plot below shows that neutral titles dominate in terms of total engagement counts, suggesting that they are more effective at attracting viewers. In contrast, both positive and negative titles show significantly lower engagement levels, indicating that while they may evoke strong reactions, they do not necessarily lead to higher viewer interaction. (Q7)

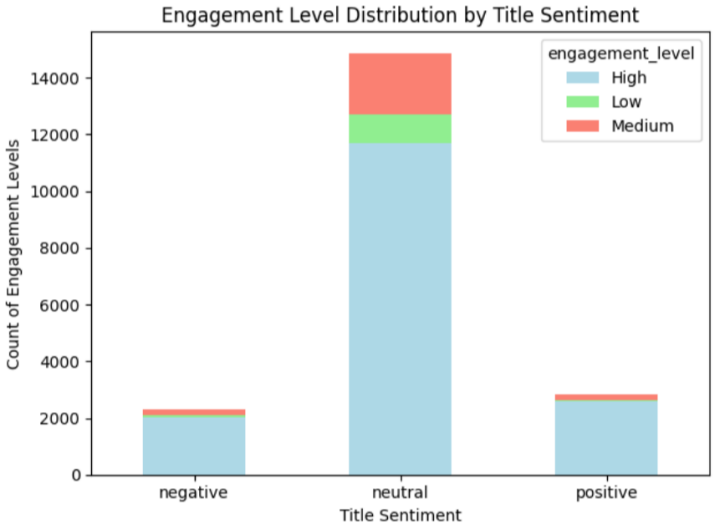


Figure 17: Sentiment vs. Engagement

Further analysis shows that using words like "shocking," "must watch," "amazing," "incredible," "unbelievable," and "you won't believe" increases engagement. These attention-grabbing phrases attract viewers and encourage them to click on videos, leading to higher interaction rates. Incorporating such compelling language in titles can effectively draw in audiences. (Q8)

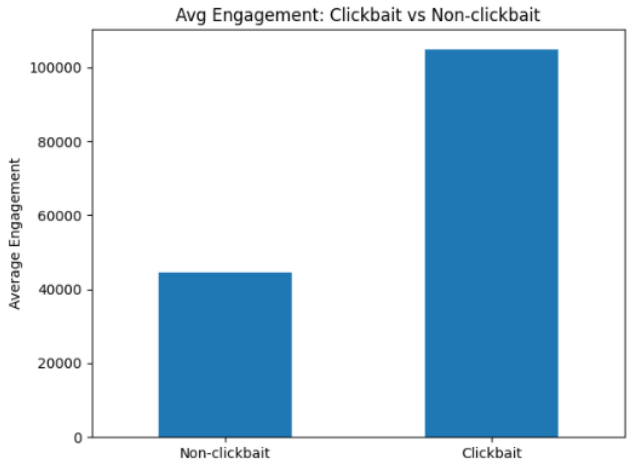


Figure 18: Clickbait vs. Engagement

In line with the goal of this research to help new YouTubers understand the factors affecting the trending of their videos, several plots are presented below,

each followed by an analysis.

E3: What is the average number of likes for videos with different publication days of the week?

As observed, Friday has the highest average likes, while Saturday has the lowest.

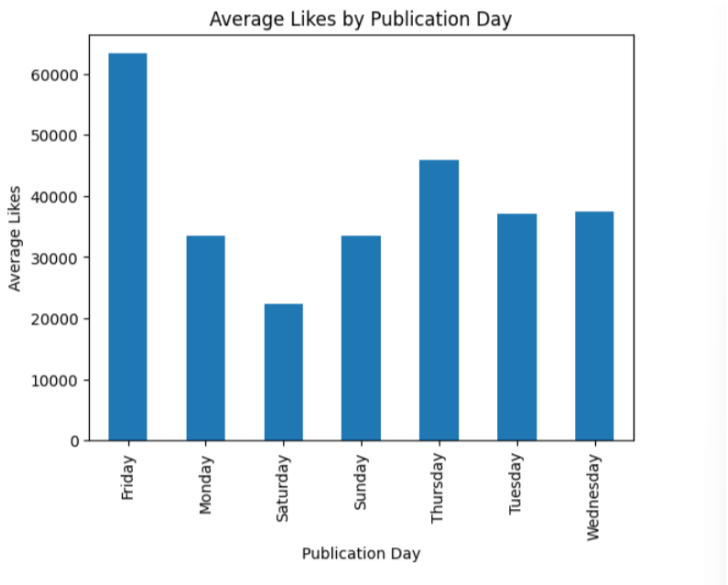


Figure 19: Like vs. Day

E4: How does the publication hour impact the views and likes of videos?

As observed, around 3 has the highest views, showing a significant difference compared to other hours, while the average likes remain relatively consistent throughout the day.

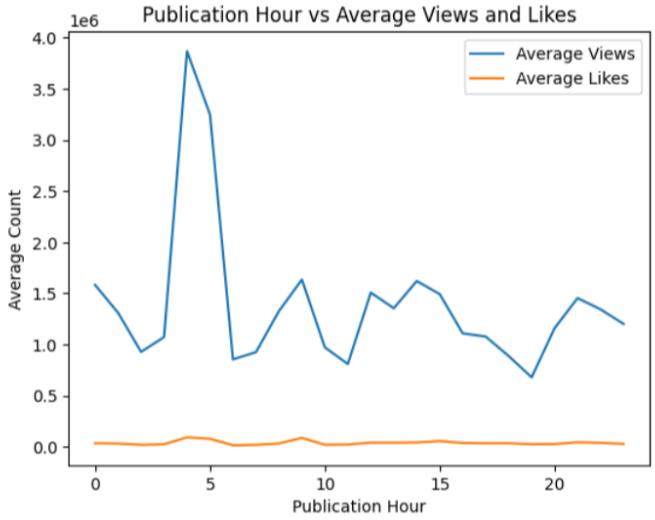


Figure 20: Like vs. Hour

E5: What is the distribution of engagement levels for videos published in different Year?

As observed, the data only relates to the years 2017 and 2018, and in both years, the highest level of engagement has been "high."

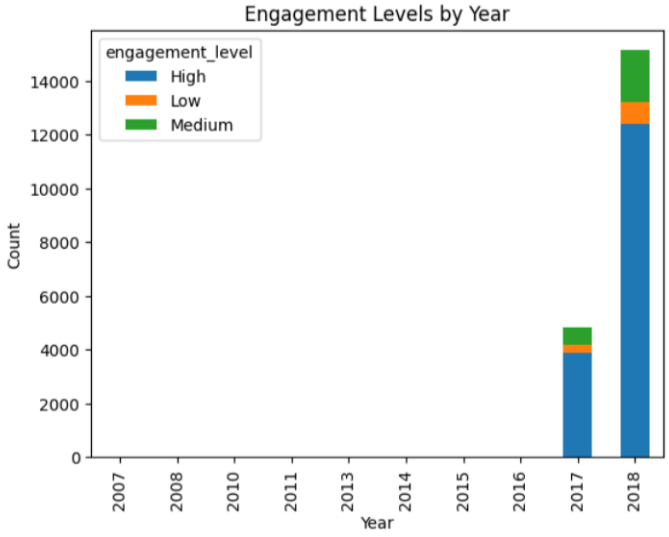


Figure 21: Engagement vs. Year

E6: How does the dislike ratio vary across different video cate-

gories?

According to the plot below, the highest number of dislikes is associated with the category "Nonprofits and Activism," followed by "News and Politics" in second place.

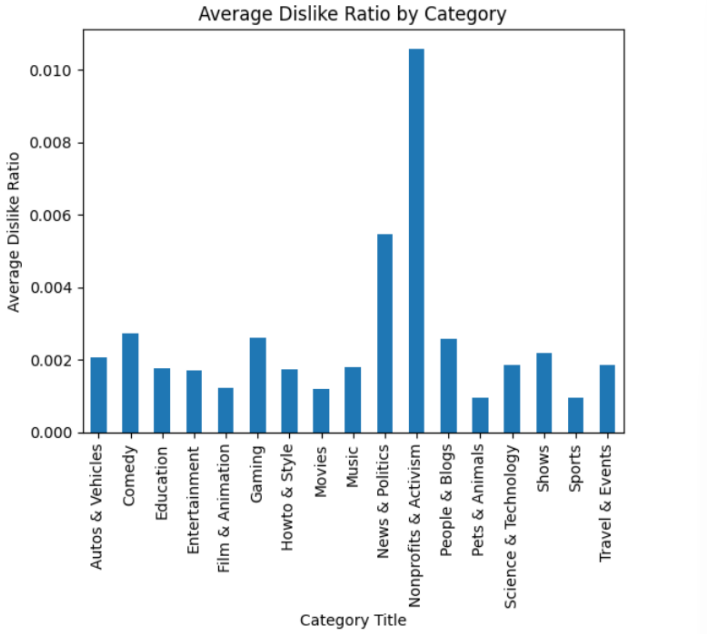


Figure 22: Dislike vs. Category

E7: How do engagement levels differ for videos that have their comments disabled compared to those that do not?

Videos with comments enabled significantly outperform those with comments disabled in terms of engagement levels. Most engagement is categorized as "High" for videos with comments allowed.

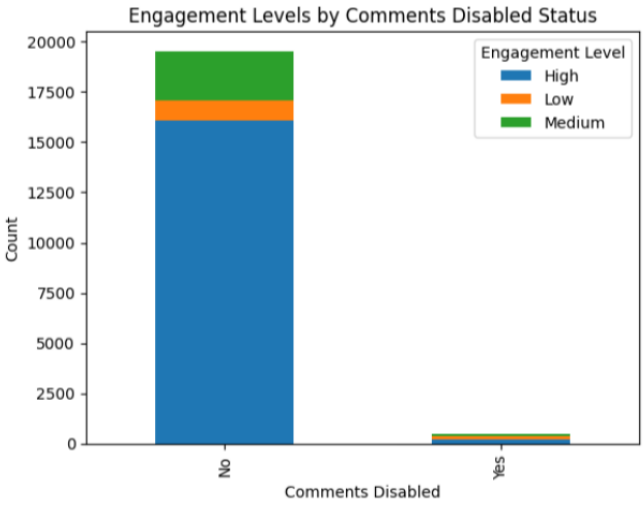


Figure 23: Engagement vs. Comments Status

III. Statistical Analysis

The analyses conducted in this section aim to understand trending behavior in order to design better senarios for Youtubers to attract visitors.

First, checking Is there a significant association between the day of the week a video is published and its likelihood of trending?

Chi-Squared Test:

$$\chi^2 = \sum \frac{(O - E)^2}{E} \tag{1}$$

Where:

- χ^2 is the Chi-Squared statistic,
- O is the observed frequency,
- E is the expected frequency.

There is no significant association between the day of the week a video is published and its likelihood of trending. The chi-square test statistic is 7.78 with a p-value of 0.2547, indicating that trending counts are uniform across all weekdays. Therefore, we fail to reject the null hypothesis.

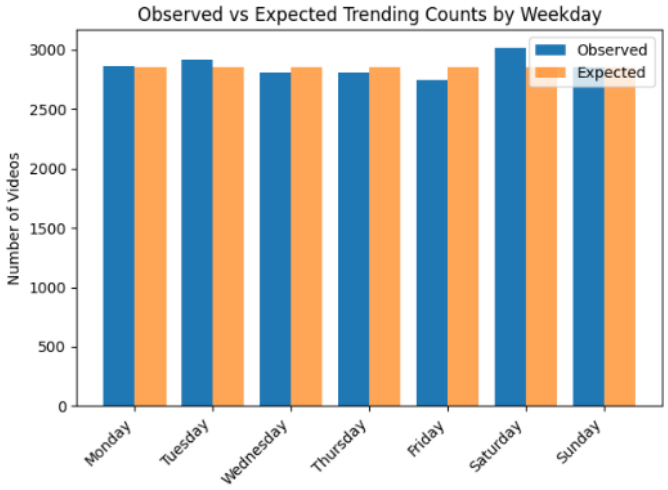


Figure 24: Based on Weekday

Then, checking is there a significant difference in viewer engagement (likes-to-views ratio) across different video categories?

Kruskal-Wallis H statistic:

$$H = \frac{(N-1)}{\sum_{i=1}^{g} n_i} \left(\sum_{i=1}^{g} n_i (\bar{r}_i - \bar{r})^2 \right) \Big/ \sum_{i=1}^{g} \sum_{j=1}^{n_i} (r_{ij} - \bar{r})^2$$

- g = number of categories
- n_i = number of observations in category i
- r_{ij} = rank of observation j from category i
- N = total number of observations
- $\bar{r}_i$ = mean rank of category i
- $\bar{r}$ = mean of all ranks

There is a significant difference in viewer engagement (likes-to-views ratio) across different video categories. The Kruskal-Wallis test statistic is 2567.99 with a p-value of 0.0000, leading to the rejection of the null hypothesis.

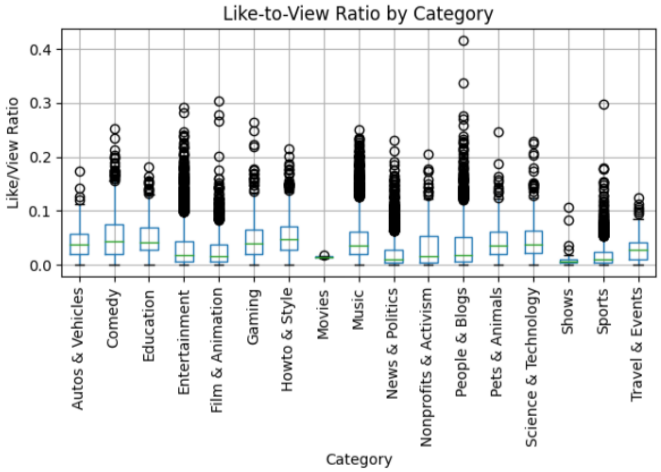


Figure 25: Engagement vs. Category

IV. Results

As mentioned, the main goal of this research was to examine the factors influencing a video's trending status on YouTube. Key insights for new YouTubers include a preference for content in the categories of entertainment, music, and film, with higher engagement in spring and winter. Title and description length, however, have minimal impact on engagement. Sentiment analysis revealed that positive sentiment in both titles and descriptions increases engagement, while words that evoke excitement can boost video views. Next, we will discuss the results related to the two hypotheses tested.

Table 14: Chi-Square Test Results

Test Statistic	7.78
P-value	0.2547
Decision	Fail to reject H0
Conclusion	No evidence of difference in trending videos by weekday.

The Chi-Square test statistic is 7.78 with a p-value of 0.2547, indicating that we fail to reject the null hypothesis. This suggests there is no significant difference in the trending status of videos based on the day of the week.

Table 15: Kruskal-Wallis Test Results

Test Statistic	2567.99
P-value	0.0000
Decision	Reject H0
Conclusion	Distributions differ by category.

Distributions differ by category.

V. Conclusion

The main goal of this research was to analyze customer personality in order to assess the acceptance of campaigns by customers and to identify various factors influencing this acceptance. The findings revealed that only 14.96% of customers responded positively to these campaigns, which is a very low figure. Therefore, an analysis of different sections of the dataset was conducted to gather sufficient reasons for the business to improve its sales policies or consider replacing them with new strategies. The analysis showed that the best-selling products were wine, meat, and gold, with significant impacts from both in-store and online sales. This suggests that future campaigns should focus on these top-selling products. Additionally, no significant relationship was found between age and campaign acceptance among customers, indicating that age is not a significant factor. However, marital status and having children are important factors influencing campaign acceptance. In particular, individuals with children, especially those with teenagers, showed very low interest in the campaigns. Therefore, it is recommended to design future campaigns specifically targeting this group. Since online sales emerged as one of the top sales methods, and customer visits to the website have favorable statistics, it is advisable to display campaigns through the website to reach customers effectively. For future work, it is suggested to conduct a more in-depth analysis of the 24 complaints made by customers. Additionally, a closer examination of the second campaign, which had the lowest acceptance rate, should be carried out. More tests can be implemented for further analysis of the dataset, including a detailed review of five campaigns based on spending behaviors.