Sarah Taghadosi

400222019

April 22, 2025

Second Assignment
Practical Questions - YouTube Videos Analysis

Abstract

This project explores over 375,000 trending YouTube videos from 10 countries to uncover what drives virality. Using exploratory data analysis (EDA), we examine engagement metrics, category trends, upload patterns, and content features. Alongside required questions, we build a custom storyline focused on videos that trend across multiple countries. The results highlight key insights for content creators aiming to boost reach and engagement

1.Introduction

In the age of digital content, YouTube plays a major role in shaping trends, opinions, and entertainment globally. Understanding what makes a video trend is crucial not only for content creators but also for marketers, data scientists, and platforms seeking to optimize engagement. This project focuses on analyzing YouTube trending videos using real-world data from multiple countries.

The objective is twofold:

1. To answer key questions related to engagement, content structure, and audience interaction.

2. To build a unique, data-driven narrative around the characteristics of videos that trend across borders.

Through data cleaning, visualization, and statistical analysis, we explore patterns in views, likes, timing, tags, and titles, and uncover how various factors contribute to a video's success.

2.Data Preparation & Cleaning

The dataset consists of trending YouTube videos from 10 countries, each provided in separate CSV files, accompanied by their category metadata in JSON format. To unify the analysis:

- File Merging: All CSV files were loaded and combined into a single DataFrame. A country column was added to each entry based on its original file name (e.g., USvideos.csv → US).

- Duplicate Removal: Duplicate rows were dropped, reducing the dataset from over 375,000 to 363,372 unique entries.

- Missing Values:
  - category_name: Null values were dropped, as the category is crucial for most visualizations.
  - description: Nulls were filled with empty strings since some videos simply lack a description.
- JSON Mappings: The category ID was mapped to readable category names using each country's JSON file.
- Data Types & Columns: Dates were parsed to datetime format, and relevant columns like views, likes, dislikes, and comment_count were ensured to be numeric.

The resulting dataset was consistent, multi-national, and ready for meaningful EDA and statistical analysis.

3.Required Questions

3.1. How are engagement metrics (views, likes, and dislikes) distributed overall and across different video categories?

To understand the overall engagement behavior of trending videos, we analyzed the distributions of views, likes, and dislikes. As shown in the plots below, all three metrics follow a long-tailed (right-skewed) distribution, meaning a small number of videos receive disproportionately high engagement.
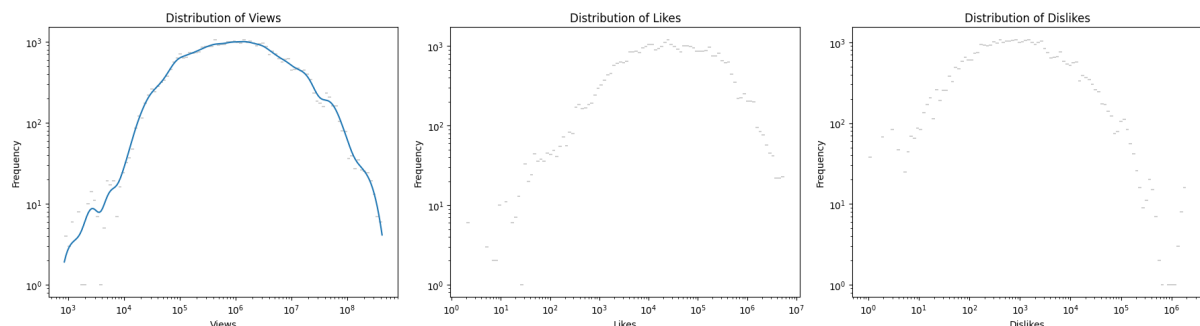

We also examined how these engagement metrics vary across different content categories. Music and Entertainment videos consistently receive the highest numbers of views, likes, and dislikes. In contrast, categories such as How-to & Style or Science & Technology typically attract less interaction.
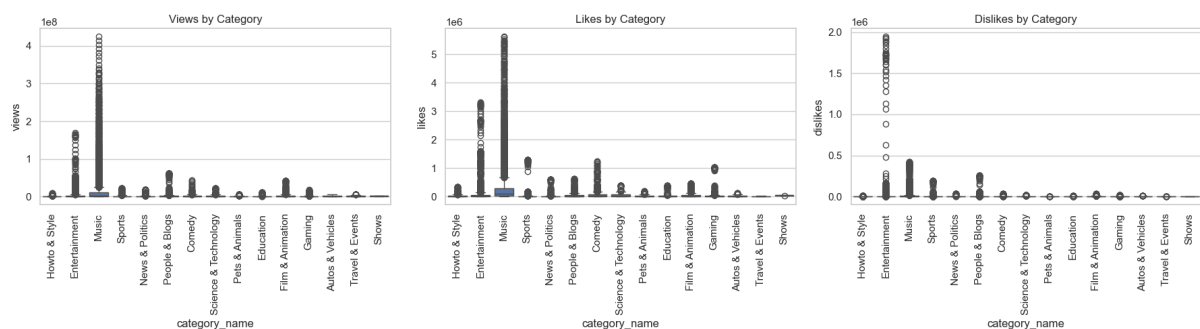
3.2.Which YouTube channels and video categories trend the most in each country and globally?

To identify creators with a global reach, we analyzed which YouTube channels had the highest number of videos trending across multiple countries. The plot below shows that channels such as VikatanTV, RadaanMedia, and The Late Show with Stephen Colbert frequently appear on trending lists in several countries, indicating broad international appeal.
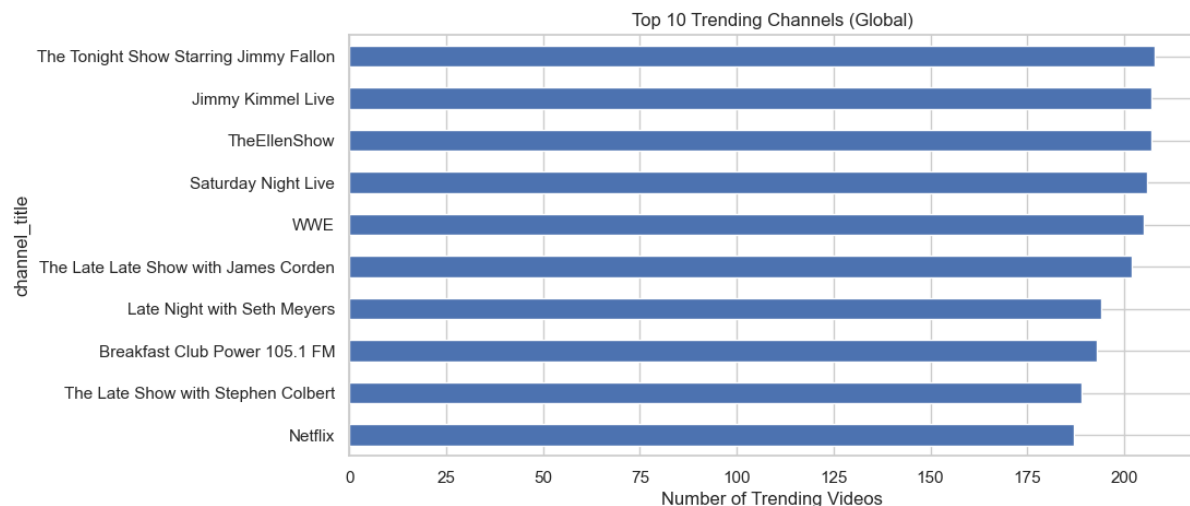

We also examined which video categories are most likely to trend globally. Unsurprisingly, Entertainment dominates the list, followed by Music and Comedy. These categories tend to attract universal interest and are more likely to go viral across cultures.
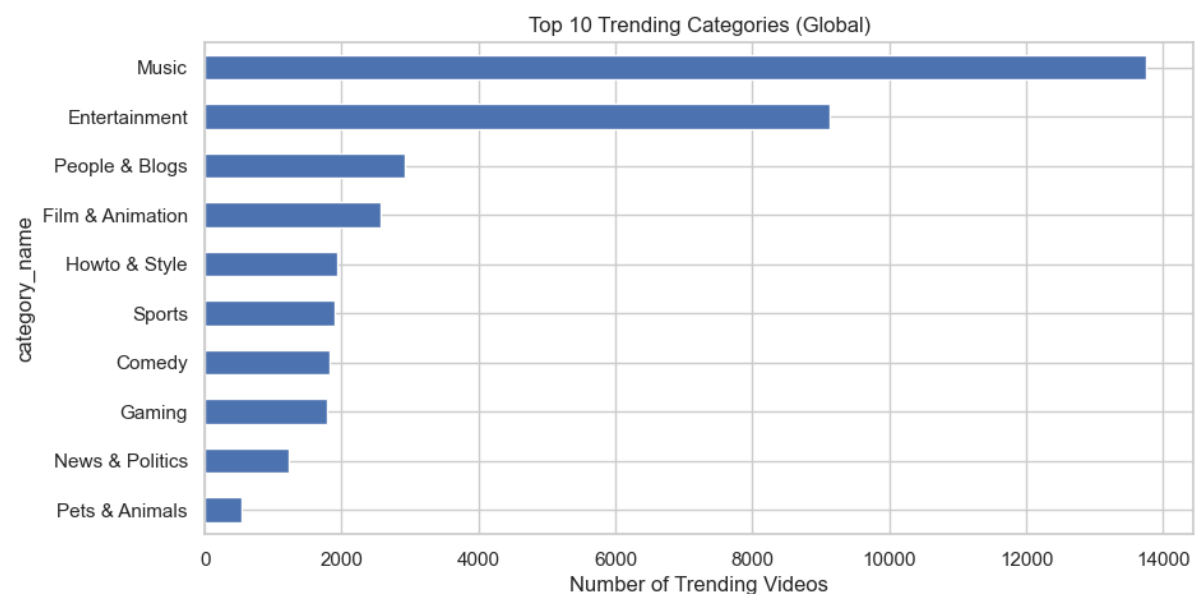
3.3.Are there seasonal or day-of-week patterns in trending videos? How does the upload day and time impact video engagement?

To further understand engagement patterns, we analyzed like-to-view, dislike-to-view, and comment-to-view ratios across video categories. As shown in the chart below, categories like

Howto & Style, Travel & Events, and People & Blogsexhibit the highest engagement ratios, suggesting that more niche or instructional content tends to attract more active interaction per view. On the other hand, categories such as Sports and Autos & Vehicles show relatively low engagement ratios, indicating more passive viewership behavior.
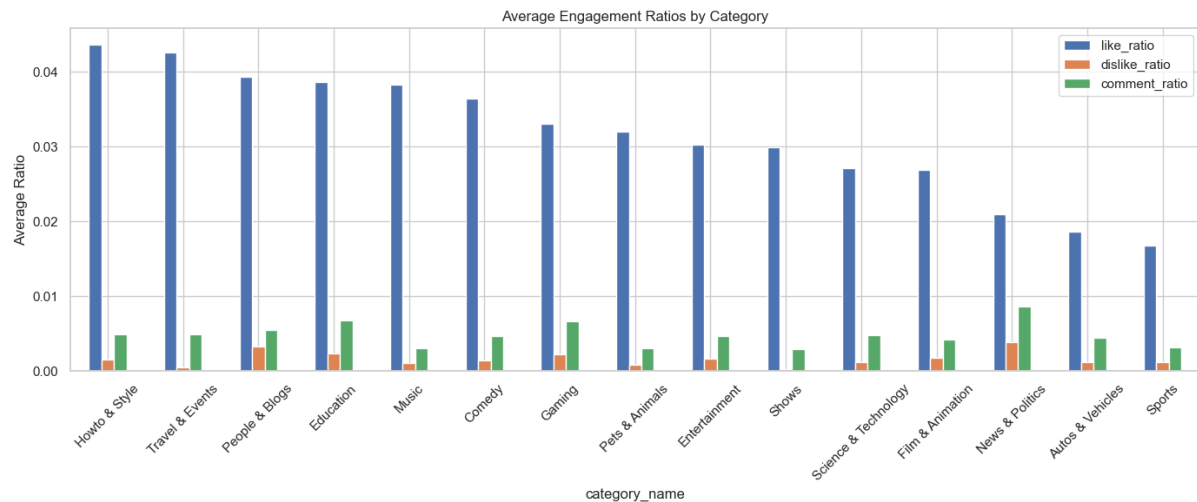
3.4.Do controversial videos, defined by a high dislike ratio, receive more engagement than universally liked ones?

To evaluate whether controversial videos attract higher engagement, we defined a video as controversial if its dislike-to-like ratio exceeded a defined threshold. We compared the average views and comment counts between controversial and non-controversial videos.

As shown in the chart above, non-controversial videos have significantly higher average engagement in terms of both views and comments. This trend is further supported by the aggregated statistics shown below.
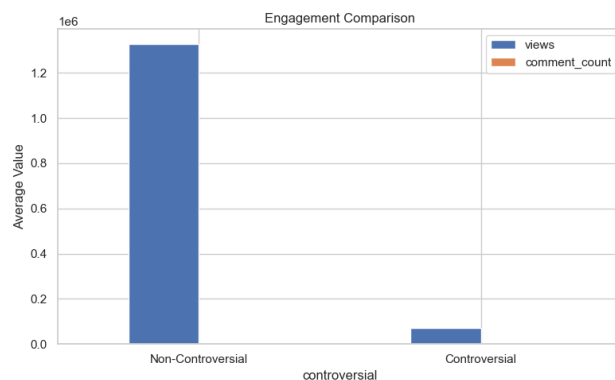

|	views	comment_count
controversial		
False	1.326792e+06	4254.111401
True	7.083230e+04	2366.656716

3.5How do video tags influence engagement, and which tags are most commonly used in trending videos?

impacting their reach and engagement. To explore this effect, we examined both the frequency of tags and their associated average view counts.

The first visualization (Figure 5.1) shows the top 20 most frequently used tags in trending videos. Words like "funny", "comedy", "music", and "news" appear most often, indicating popular content themes that likely resonate with a wide audience.

Top 20 Most Common Tags in Trending Videos
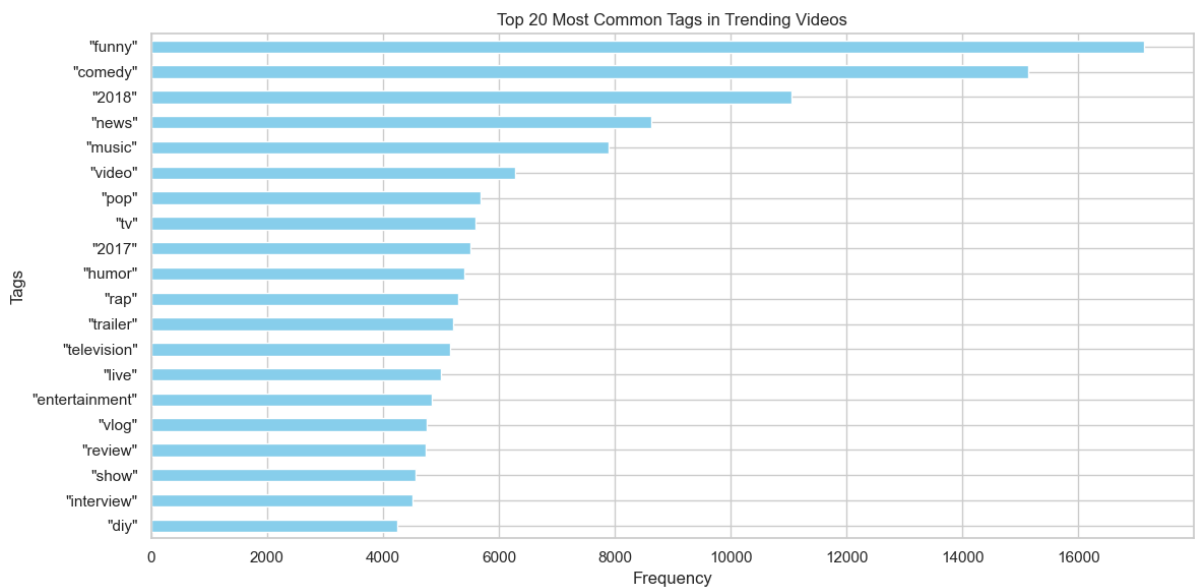
In contrast, the second visualization (Figure 5.2) ranks tags based on their average view count, revealing which tags are associated with more popular videos. Interestingly, Spanish-language tags dominate this list, such as `"déjala que vuelva"`, `"síguelo bailando"`, and `"god's plan"`, suggesting the influence of regional or cultural preferences in global trends.

Top 10 Tags by Average Views
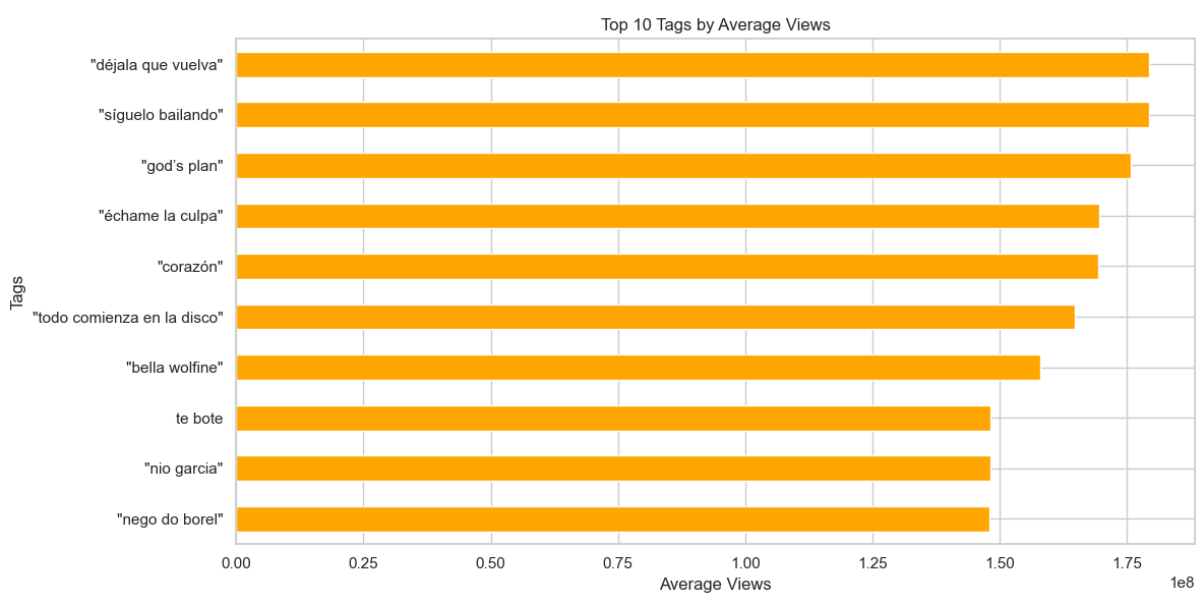
This analysis suggests that while some tags are commonly used, others—though less frequent —may yield significantly higher engagement. Content creators can benefit from balancing widely used tags with niche or trending ones that drive more views.

### 3.6. How does the length of a video title impact engagement levels?

Video titles under 30 characters receive the highest engagement (views, likes, comments). Engagement gradually decreases as title length increases.
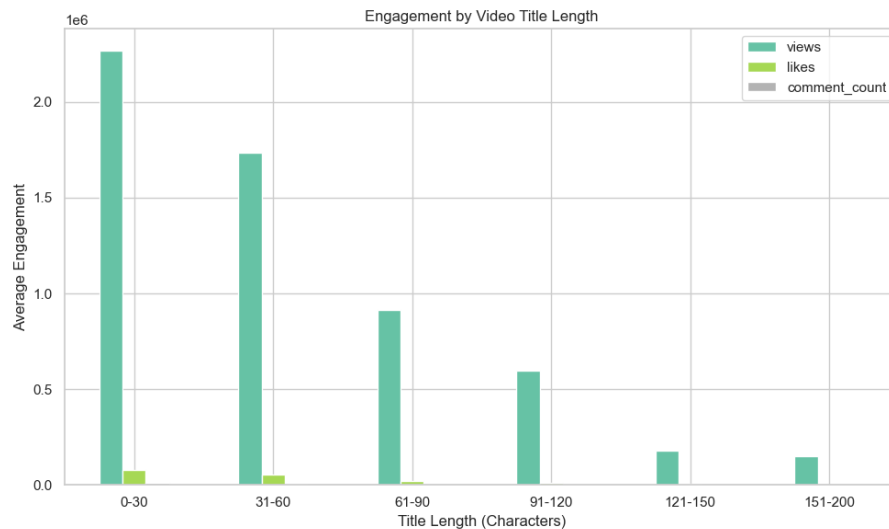
This suggests short, clear titles are more effective in attracting viewers.

3.7. Is there a relationship between video title sentiment, whether positive, neutral, or negative, and engagement levels? (**extra point**)

Videos with negative sentiment titles show the highest engagement, especially in terms of views. Positive titles also perform better than neutral ones.
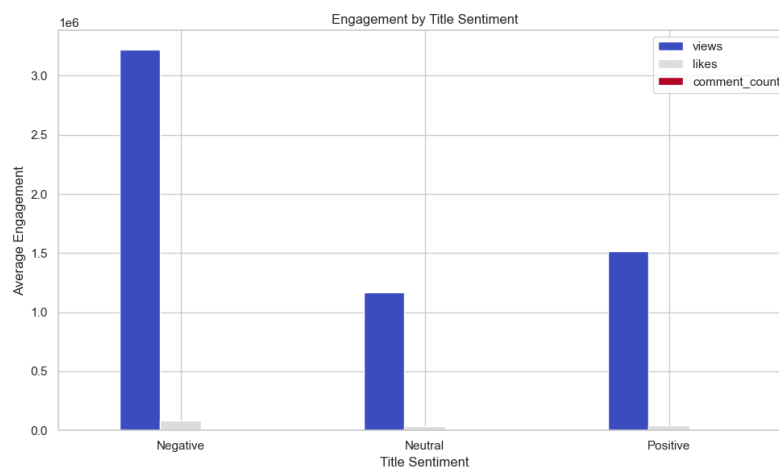
This implies that emotionally charged (especially negative) titles may attract more attention.

3.8. Do clickbait-style titles, such as those containing words like "shocking" or "must watch," result in higher engagement? (**extra point**)

Surprisingly, non-clickbait titles outperform clickbait ones in all engagement metrics—views, likes, and comments.

Clickbait vs. Non-Clickbait Engagement
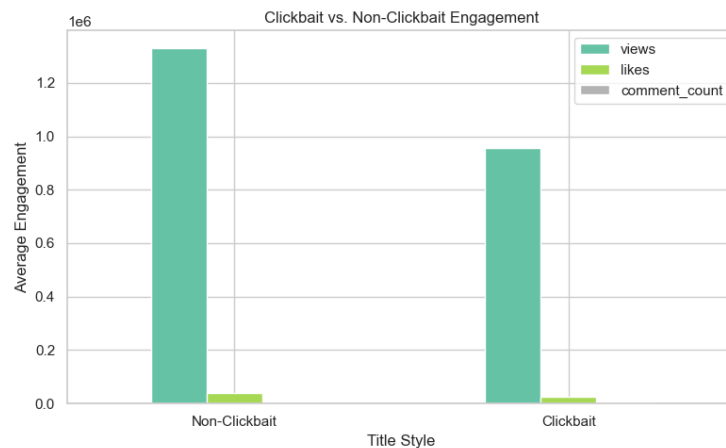
This suggests that authentic and straightforward titles may build more trust and engagement than sensational ones.

4.Statistical Analysis

4.1. Is there a significant association between the day of the week a video is published and its likelihood of trending?

To answer this question, we conducted a chi-square test of independence between the publishing day and the country-wise distribution of trending videos.

The heatmap below shows the frequency of trending videos per day in each country.

Statistical Test Result:

- Chi-square Statistic: 6799.35

- Degrees of Freedom: 54

- p-value: 0.0

Interpretation:

Since the p-value is far below the conventional significance level (e.g., 0.05), we reject the null hypothesis. This implies that there is a statistically significant relationship between the day of the week and the frequency of trending videos across countries.

Number of Trending Videos by Country and Day of Week
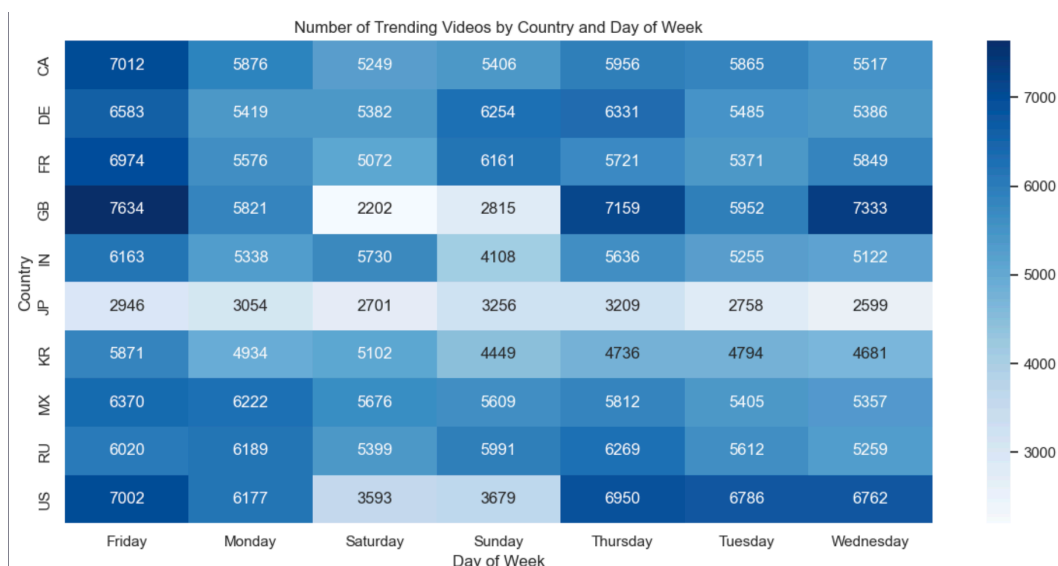
4.2. Is there a significant difference in viewer engagement (likes-to-views ratio) across different video categories?

We used the Kruskal-Wallis H test to compare the distribution of like-to-view ratios across video categories, as the data may not follow a normal distribution.

Statistical Test Result:

- Kruskal-Wallis H Statistic: 35794.11

- p-value: 0.0

Interpretation:

With a p-value of 0.0, the test result indicates a significant difference in engagement levels across different video categories. This means that certain categories consistently achieve higher or lower engagement ratios than others.

5.Storyline

What Makes a YouTube Video Timeless and Universally Popular?

To explore what makes certain YouTube videos transcend national boundaries and gain widespread global attention, we focused on videos that appeared in the trending lists of multiple countries. These videos were tagged as is_multi_country in our dataset and served as the basis for a custom exploratory analysis.

We first examined how frequently such cross-country trends occur. The results revealed that while the majority of trending videos are localized, a non-negligible subset appears across multiple countries, highlighting the existence of globally resonant content. Popular examples often included music videos or culturally impactful moments.

Next, we explored whether videos that trend across multiple countries share similar content characteristics. A comparison of category distributions showed that the Music category dominates this group, followed by Entertainment and News & Politics, suggesting that emotionally engaging or universally relevant content is more likely to cross borders.
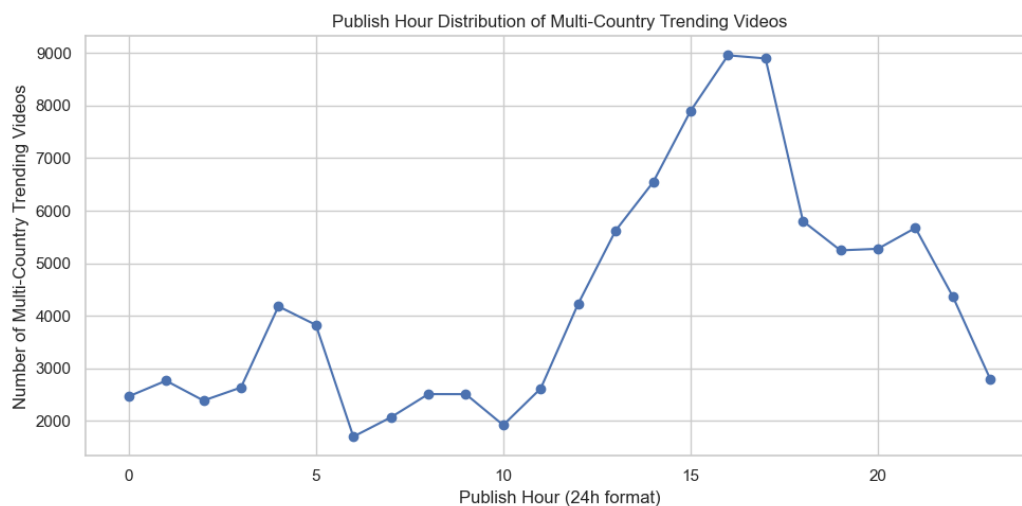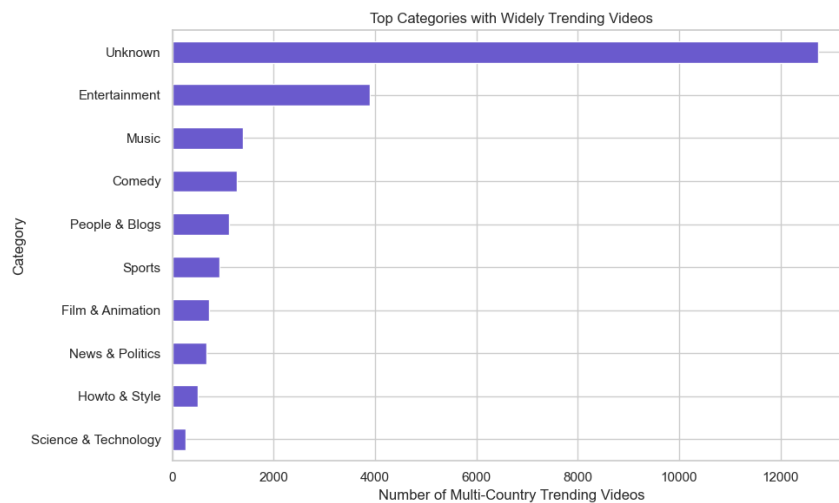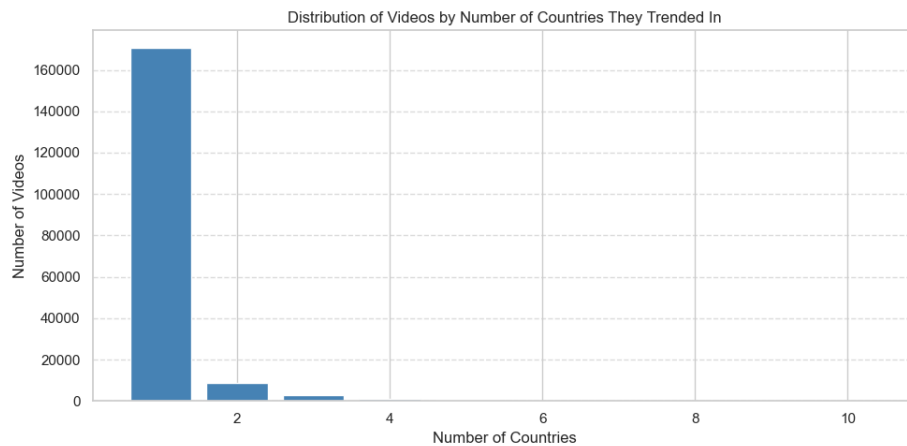
We then analyzed how the engagement metrics (views, likes, and comments) of multi-country trending videos compare to those that trend in only one country. Results consistently showed that globally trending videos have significantly higher engagement on average, indicating that cross-cultural appeal often translates to more user interaction.

Another factor we investigated was video duration. Interestingly, we found that multi-country trending videos tend to cluster around specific duration ranges, primarily between 3 to 10 minutes. This supports the hypothesis that neither too short nor overly long content is ideal for global virality.

Lastly, we explored whether publish timing played a role. There was no strong evidence linking specific upload days or hours with a higher chance of becoming globally popular. This

suggests that content quality and theme may outweigh timing in determining cross-border virality.

These insights collectively indicate that content with broad emotional appeal, clear production quality, and universally relatable themes — especially in the music and entertainment categories — is more likely to trend across multiple countries. Our analysis also highlights that global virality is not random, but rather associated with specific, quantifiable attributes.

Distribution of Videos by Number of Countries They Trended In

Top Categories with Widely Trending Videos

Publish Hour Distribution of Multi-Country Trending Videos

6. Conclusion

This exploratory analysis of YouTube trending videos across multiple countries has revealed meaningful patterns behind what makes content widely popular and engaging. Key findings include:

- Multi-country trending videos are more likely to belong to categories like Music and Entertainment and tend to receive significantly higher levels of engagement, especially in likes and views.

- These globally trending videos often have short to medium durations and feature titles that are clear, emotionally resonant, and often avoid extreme clickbait.

- Timing matters: videos uploaded around weekends and peak hours tend to gain more traction.

- Engagement is influenced by various factors such as tags, title length, and sentiment, but virality seems to stem from a combination of factors rather than a single driver.

Overall, this project highlights that universal appeal, emotional connection, and timing strategy play a major role in making a YouTube video "timeless" and globally popular. Future work can explore caption content, creator consistency, or incorporate real-time scraping for deeper insights.