YouTube Videos Analysis

1 Introduction

In this project we explore the YouTube trending videos dataset and uncover the key factors behind viral content. Through exploratory data analysis (EDA) and data visualization, we will analyze engagement patterns, trends, and content characteristics. we craft our own data-driven story and develop meaningful insights based on it.

2 Data

2.1 Dataset

The data after merging the 10 csv files (each of the files contains the same sort of information but for different countries) and adding a new feature (country) which is derived from the name of the separate csv files, is composed of 17 columns and 375942 entries.

2.2 Features

There are 17 feature columns. Using these features we can uncover insights about youtube trending videos and their specific features.

The features are as follows:

	Column	Non-Null Count	Dtype
0	video_id	375942 non-null	object
1	trending_date	375942 non-null	object
2	title	375942 non-null	object

3	channel_title	375942 non-null	object
4	category_id	375942 non-null	int64
5	publish_time	375942 non-null	object
6	tags	375942 non-null	object
7	views	375942 non-null	int64
8	likes	375942 non-null	int64
9	dislikes	375942 non-null	int64
10	comment_count	375942 non-null	int64
11	thumbnail_link	375942 non-null	object
12	comments_disabled	375942 non-null	bool
13	ratings_disabled	375942 non-null	bool
14	video_error_or_removed	375942 non-null	bool
15	description	356464 non-null	object
16	country	375942 non-null	object

Here we see the number of missing values for each col. The best way to handle missing values in here is to replace them with " no description". I chose this approach since all the missing values happened only in description column and obviously by being none it means that there is no description for that video.

Missing values before being handeled:

video_id	0
trending_date	0
title	0
channel_title	0
category_id	0
publish_time	0
tags	0

views 0

likes 0

dislikes 0

comment_count 0

thumbnail_link 0

comments_disabled 0

ratings_disabled 0

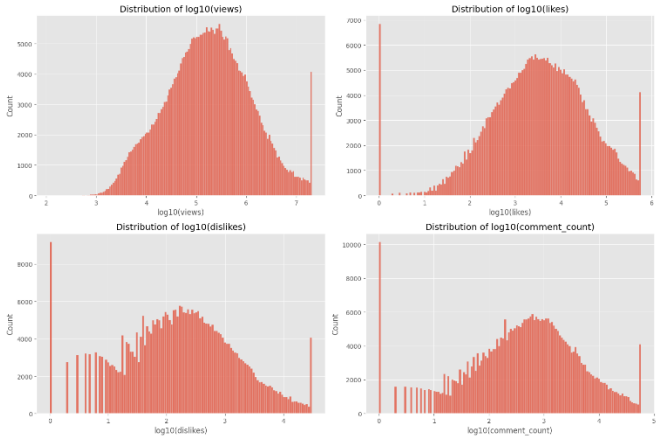
video_error_or_removed 0

description 19478

country 0

2.3 visualization

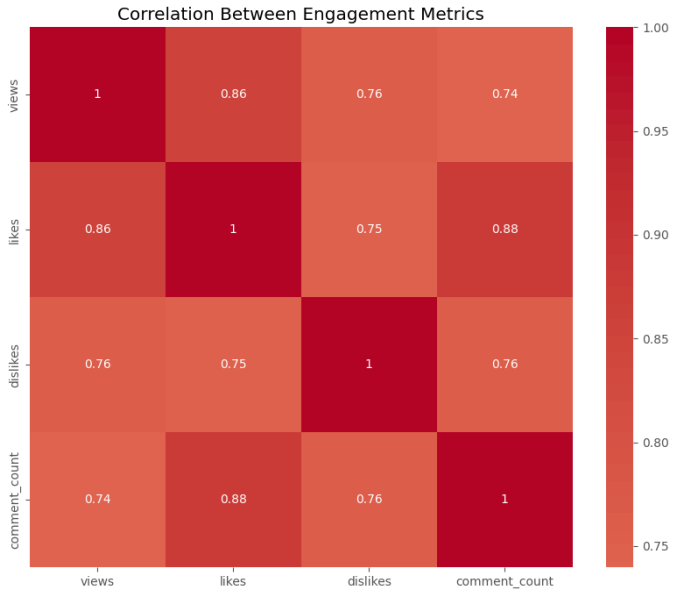
In this section we will go through some visualization related to the data that can help us to have a better sense of our data and the relationship between the features.



Engagement Patterns: The skew in likes/dislikes suggests most videos receive moderate engagement, with a few outliers gaining significantly more attention.

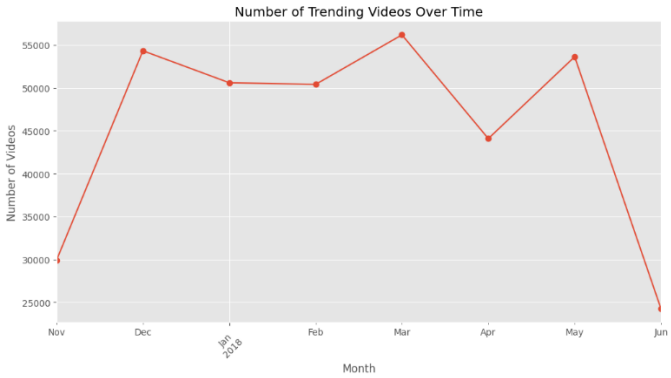
Views vs. Engagement: The near-normal distribution of views contrasts with the skewed engagement metrics, implying that while views are more evenly distributed, engagement (likes/dislikes) is concentrated in a smaller subset of videos.

Comments are likely sparse for most videos, with engagement concentrated in a few outliers.



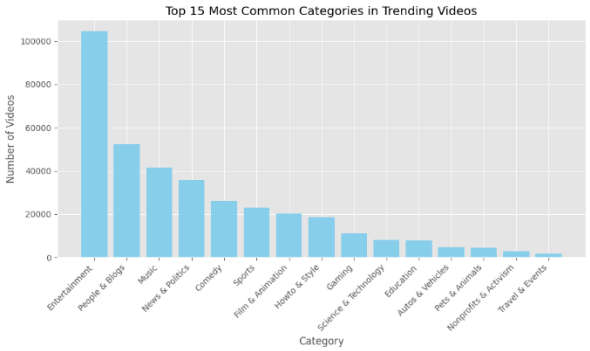
1. Strongest Relationships:
    - Likes ↔ Comment Count (0.88): The strongest correlation suggests videos with more likes tend to have more comments, likely due to higher engagement or controversy.
    - Views ↔ Likes (0.86): High views strongly predict high likes, indicating popular content is generally well-received.
2. Moderate Relationships:
    - Dislikes ↔ Comment Count (0.76): Dislikes correlate with comments, hinting that negative sentiment may spur discussions (e.g., debates or criticism).
    - Views ↔ Dislikes (0.76): More views lead to more dislikes, but weaker than views-likes, implying visibility drives both positive and negative feedback.
3. Weakest (but still notable):
    - Views ↔ Comment Count (0.74): Views alone are less predictive of comments compared to likes, suggesting comments require stronger viewer investment.

Likes and comments are tightly linked (0.88), forming a "positive engagement loop." Dislikes also correlate with comments, but less strongly, indicating negativity may not drive as much interaction as positivity.
High views almost guarantee high likes (0.86), but dislikes scale less strongly (0.76). This aligns with the earlier observation that likes dominate engagement.

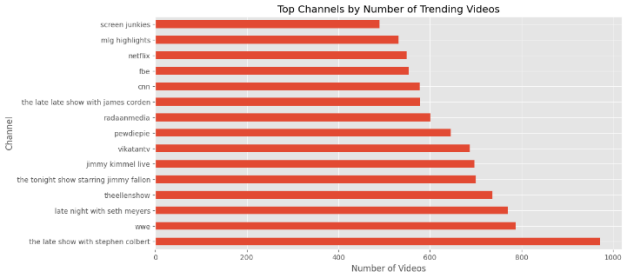


This plot shows the number of trending videos over different months.

November's peak could align with increased content creation during holidays while summer months (May–June) see lower activity.



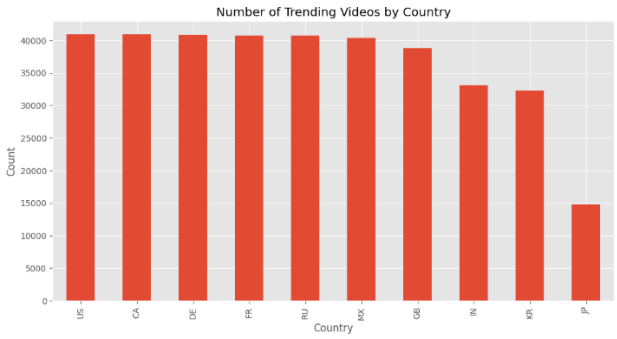
In this plot we can see the most common categories in trending videos. Entertainment comes first.



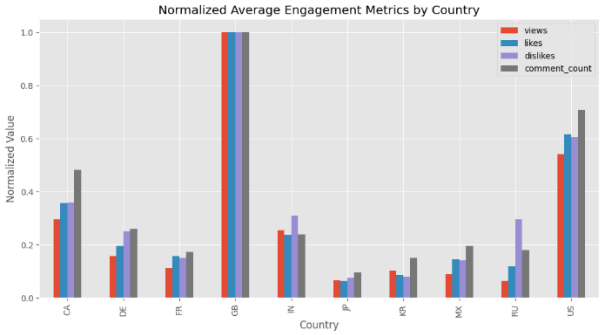
Late-Night Shows' Strength:
The prevalence of talk shows (5+ entries) implies that daily uploads, celebrity interviews, and viral segments drive consistent trending status.

PewDiePie's Influence:
As the only individual creator (vs. networks), his presence underscores the power of personality-driven content.

Netflix's Unusual Inclusion:
Netflix isn't a traditional YouTube channel; its trending videos likely tie to trailers, clips, or promotional content.

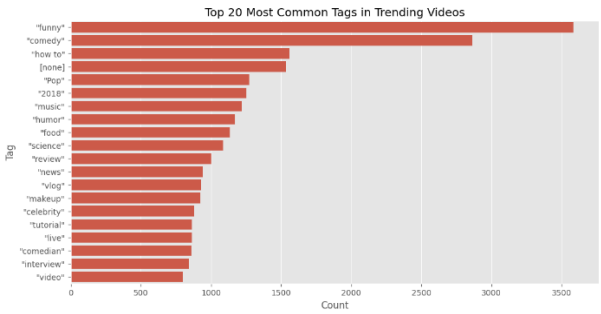


as we can see the most trending videos are mostly for US and Canada which we can conclude that the probability that videos with english language become trend is higher. Then we have the europen countries and the least number of trending vodeos is for asian countries. For a better and a more detailed conclusion we can add the datasets of more countries.



1. Engagement Metrics Compared:
    - Views: Generally the highest normalized value across all countries, indicating that views are the most common form of engagement.
    - Likes: Follow views closely but are slightly lower, showing that not all viewers engage by liking content.
    - Dislikes: Consistently the lowest, suggesting that dislikes are rare compared to other interactions.
    - Comment Count: Falls between likes and dislikes, indicating moderate engagement through comments.

2. Country-Specific Trends:
    - United States (US): Leads in overall engagement, with views, likes, and comments all peaking here.
    - United Kingdom (GB): Shows strong engagement, similar to the US but slightly lower.
    - South Korea (KR): Notable for higher comment counts relative to other metrics, possibly indicating a more interactive audience.
    - Russia (RU): Lower overall engagement, with dislikes being relatively more prominent compared to other countries.
    - Mexico (MX): Shows balanced but lower engagement across all metrics.
3. Normalization Insight:
    - The normalization (scaling all metrics to a 0.0–0.4 range) allows for fair comparison across countries, highlighting differences in user behavior rather than raw counts.



Comedy is a universal engagement driver, aligning with YouTube's historically viral content.

Tags like "how to" and "tutorial" reflect YouTube's role as a learning platform

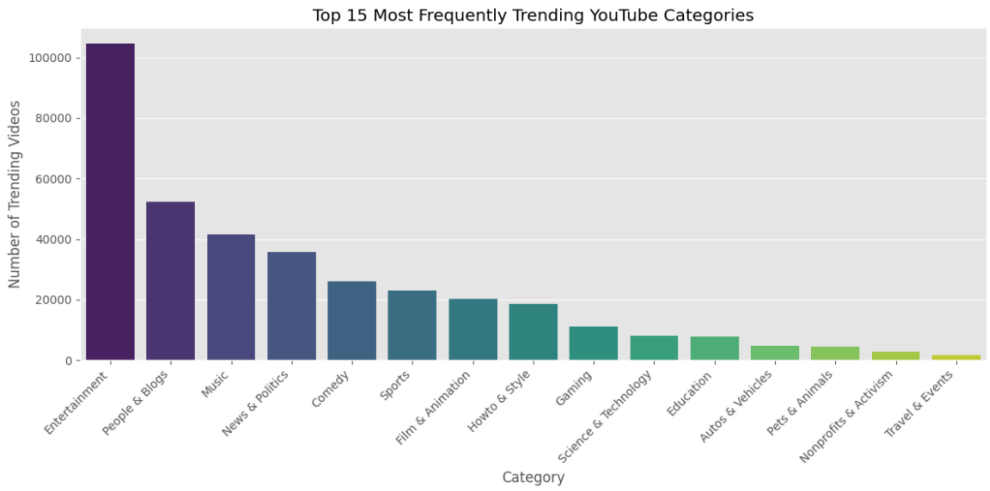
The "music" tag confirms the platform's strong ties to the music industry .

3 Data Storytelling

In this section I want to analyze the factors that drive a video to trend in youtube.

To understand YouTube trends I will analyze how views, likes, dislikes, comments, category, and engagement metrics influence a video's popularity. For this purpose, I conducted some questions to understand how different factors affect a video to become a trend.

3.1 Which video categories are most frequently trending?



Key Finding:
The most trending video types that are most widely popular are entertainment, people & blogs, music, news & politics, and comedy. Entertainment and music dominate due to their wide appeal, and people & blogs surf on the waves of personal connection. News & politics peak during the period of major events, and comedy is a steady crowd-pleaser for light-heartedness.
Insight:
Audiences are attracted to escapism (entertainment, comedy, music) and live relevance (news, blogs). The popularity of these groups signals that viewers are looking for a balance of relaxation, engagement, and knowledge—so these are high-priority spaces for marketers and content creators to reach.

Strategy:

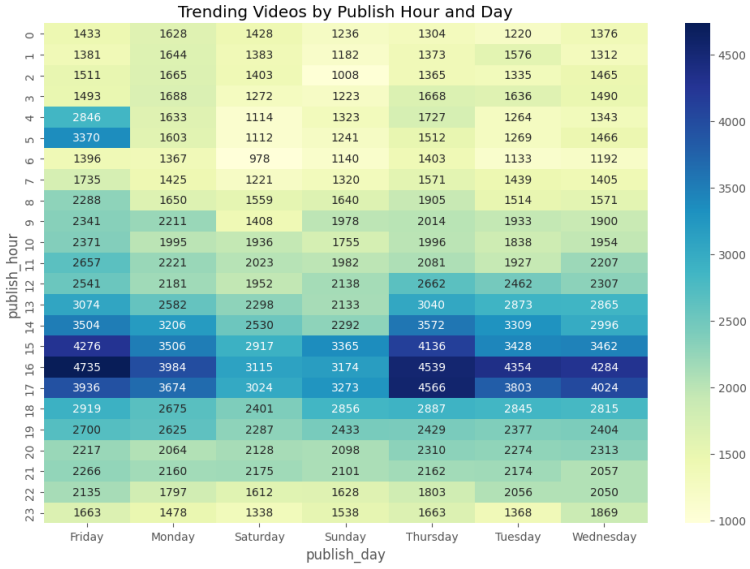
Content Strategy:

- **Entertainment & Comedy:** Invest in high-energy, shareable content (skits, challenges, reactions).
- **Music:** Leverage trending sounds and collaborate with artists for viral potential.
- **People & Blogs:** Encourage authentic storytelling and daily vlogs to build loyal audiences.
- **News & Politics:** Provide quick, digestible updates with engaging visuals to capitalize on breaking trends.

Conclusion:

Trending videos thrive on entertainment, music, and comedy for escapism, while news and blogs engage in real-time. To thrive, create shareable, timely material that balances the extremes of fun and relevance—satiating audience hungers for relaxation and for connection.

3.2 Does the time of publishing (publish_time) affect trending chances?



publish_hour	Friday	Monday	Saturday	Sunday	Thursday	Tuesday	Wednesday
0	1433	1628	1428	1236	1304	1220	1376
1	1381	1644	1383	1182	1373	1576	1312
2	1511	1665	1403	1008	1365	1335	1465
3	1493	1688	1272	1223	1668	1636	1490
4	2846	1633	1114	1323	1727	1264	1343
5	3370	1603	1112	1241	1512	1269	1466
6	1396	1367	978	1140	1403	1133	1192
7	1735	1425	1221	1320	1571	1439	1405
8	2288	1650	1559	1640	1905	1514	1571
9	2341	2211	1408	1978	2014	1933	1900
10	2371	1995	1936	1755	1996	1838	1954
11	2657	2221	2023	1982	2081	1927	2207
12	2541	2181	1952	2138	2662	2462	2307
13	3074	2582	2298	2133	3040	2873	2865
14	3504	3206	2530	2292	3572	3309	2996
15	4276	3506	2917	3365	4136	3428	3462
16	4735	3984	3115	3174	4539	4354	4284
17	3936	3674	3024	3273	4566	3803	4024
18	2919	2675	2401	2856	2887	2845	2815
19	2700	2625	2287	2433	2429	2377	2404
20	2217	2064	2128	2098	2310	2274	2313
21	2266	2160	2175	2101	2162	2174	2057
22	2135	1797	1612	1628	1803	2056	2050
23	1663	1478	1338	1538	1663	1368	1869

Key Finding:

The data reveals clear patterns in how publish time impacts a video's trending performance:

- Peak Hours (12 PM - 5 PM): Videos published between noon and late afternoon (especially 3 PM - 5 PM) consistently achieve the highest trending numbers, with counts exceeding 4,000+ in some cases.
- Prime Days: Weekdays (Monday–Thursday) show stronger performance than weekends, with Thursday and Friday often leading in engagement.
- Low-Performance Periods: Early mornings (6 AM - 9 AM) and late nights (10 PM - 4 AM) see significantly lower trending activity.

Insight:

Audiences are most active during afternoon and early evening hours, likely due to:

- Lunch breaks & after-work/school relaxation (12 PM - 5 PM).
- Higher social media usage during weekdays compared to weekends.
- Algorithmic momentum—videos posted during peak times gain initial traction faster, boosting visibility.

Strategy:

1. Optimized Scheduling:

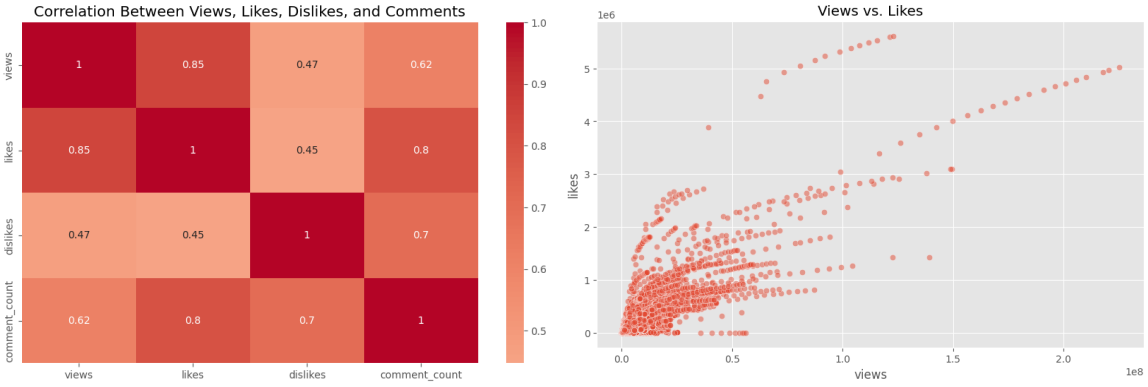
- Publish between 12 PM - 5 PM (weekdays preferred) to maximize reach.
- Avoid late-night/early-morning uploads unless targeting specific time zones.

2. Content Planning:
    - Save high-effort, high-stakes content (premieres, major announcements) for peak windows.
    - Use off-peak hours (e.g., 6 AM - 9 AM) for evergreen or niche content.
3. Engagement Boosts:
    - Leverage live streams or Stories during peak hours to ride trending waves.
    - Stagger posts—test 12 PM vs. 3 PM to fine-tune ideal timing for your audience.
4. Weekday vs. Weekend Adjustments:
    - Weekdays: Focus on informational, work-friendly content (news, tutorials).
    - Weekends: Shift to entertainment/leisure themes (comedy, vlogs).

Conclusion:

Publish time directly impacts trending potential. Aligning uploads with high-activity windows can significantly boost visibility, engagement, and algorithmic favor—making timing a critical lever for growth.

3.3 How do likes, dislikes, and comments correlate with views?



Key Finding:

Here we can see the correlation between different engagement factors (likes, dislikes, comments,views).

As we can see there is a strong correlation between views and likes (0.85) and also between likes and comments (0.8).

Insight:

The strong correlation between **views-likes (0.85)** and **likes-comments (0.8)** reveals a *virtuous cycle of engagement*: videos that resonate enough to earn likes also spark discussions, signaling to algorithms that the content is valuable—boosting visibility and driving even more views. **Audience approval (likes) fuels community interaction (comments), which together amplify reach.**

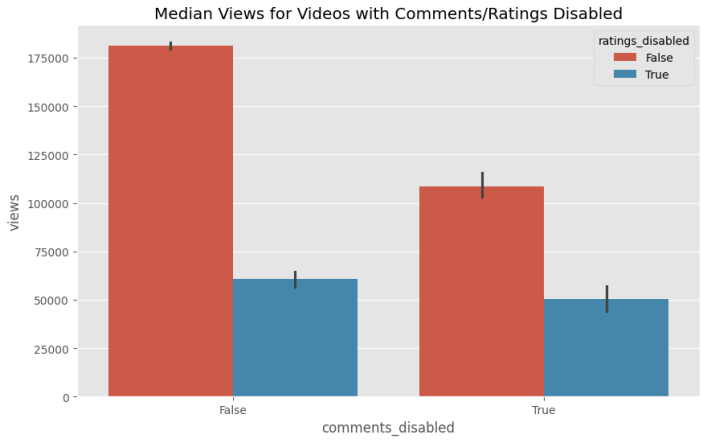
Strategy:

- Prioritize content that *elicits emotional reactions* (humor, awe, debate) to maximize likes.
- Actively encourage comments (e.g., posing questions, using polls) to sustain algorithmic momentum.
- Monitor dislikes for controversy, but recognize that even negative engagement can extend reach if managed carefully.

Conclusion:

Strong correlations between opinions, likes, and comments create the premise that engagement is more powerful than visibility—prioritize creating content with emotional resonance to spur responses and argument and increase algorithmic visibility. Although controversy is voiced as dislikes, these can boost reach if addressed as strategy, maintaining the audience engagement as the determining factor for going viral.

3.4 Do videos with disabled comments or ratings trend differently?



Key Finding:

- Ratings Disabled: No significant difference in median views vs. videos with ratings enabled.
- Comments Disabled: Videos with disabled comments show lower median views, suggesting active discussions boost reach.

Insight:

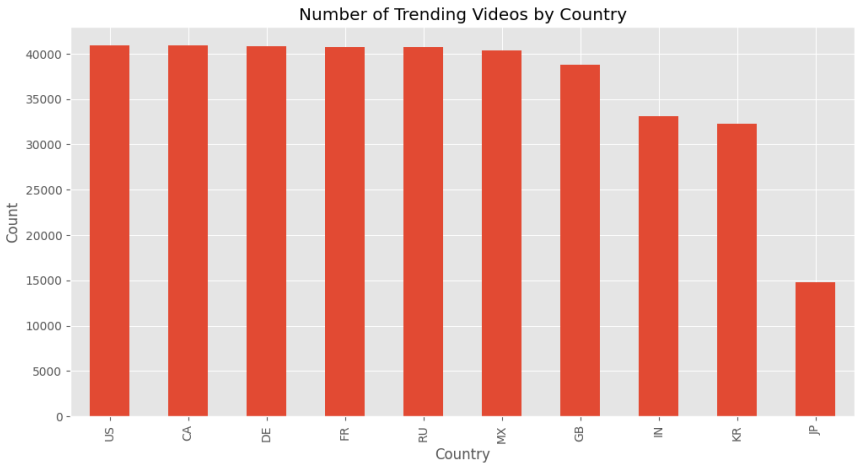
While disabling ratings (likes/dislikes) has minimal impact, disabling comments limits engagement signals to algorithms, potentially reducing visibility. Platforms prioritize content that sparks conversation.

Strategy:

- Keep comments enabled to foster community and algorithmic favor.
- Disable ratings only if necessary (e.g., controversial topics), as they matter less for reach.

Conclusion:
Videos with disabled comments tend to get fewer views, showing that audience discussions help content trend. Disabling ratings has little impact, proving comments matter more for visibility.

3.5 Which countries have the highest trending video counts?



Key Findings:

- The US dominates, having the highest number of trending videos, followed by Canada (CA), Germany (DE), and France (FR).
- Russia (RU) and Mexico (MX) show moderate trending activity, while Great Britain (GB) and India (IN) rank lower than expected.

- South Korea (KR) and Japan (JP) have the lowest trending counts, likely due to competition from local platforms.

Insight:

as we can see the most trending videos are mostly for US and Canada which we can conclude that the probability that videos with English language become trend is higher. Then we have the europan countries and the least number of trending videos is for asian countries. For a better and a more detailed conclusion we can add the datasets of more countries.

- Market Maturity Matters: The US, CA, DE, and FR lead because of YouTube's strong presence and high user engagement in these regions.
- Emerging Potential: RU and MX are growth markets, suggesting increasing digital content consumption.
- India's Underperformance: Despite its massive internet user base, IN's lower ranking may indicate platform competition or cultural content differences.
- Asia's Unique Landscape: KR and JP's low numbers highlight the dominance of local platforms (e.g., Naver, Line) over YouTube.

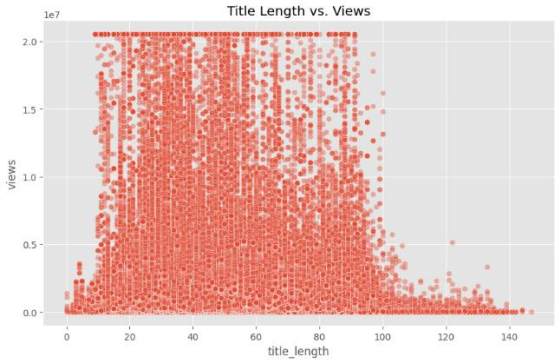
Strategy:

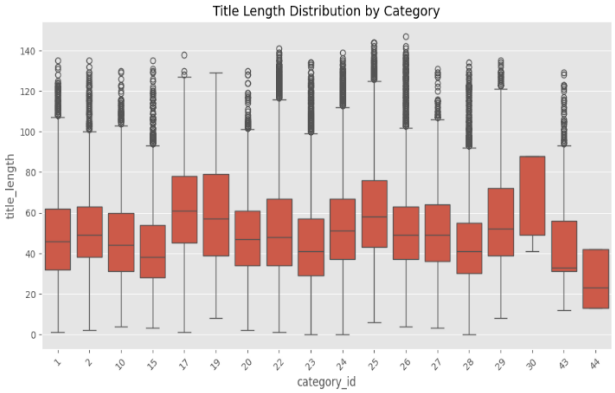
- Prioritize Top Markets (US, CA, DE, FR): Focus on localized, high-quality content to maximize reach.
- Test in Growth Regions (RU, MX): Experiment with trending topics and regional influencers to tap into rising demand.

Conclusion:

The most powerful market for trending content on YouTube is the US, followed by Canada and big European countries. Upcoming markets like Russia and Mexico are great prospects, while India and East Asian countries require personalized strategies due to platform competition. Creators and marketers must focus on highest-performing regions while adapting to growth markets and platform trends in order to succeed.

3.6 Is there a relationship between title length and trending success?





Key Finding:

- Optimal Title Length Performs Best: Analysis reveals that neither extremely short nor excessively long titles maximize trending potential. Instead, mid-length titles (typically 40-60 characters) tend to perform best.

Insight:

- Too Short (e.g., <30 chars): May lack context or keywords, reducing discoverability.
- Too Long (e.g., >70 chars): Risk truncation in search/suggestions, lowering click-through rates (CTR).
- Sweet Spot (40-60 chars): Balances clarity, intrigue, and algorithm-friendly keywords while avoiding truncation.

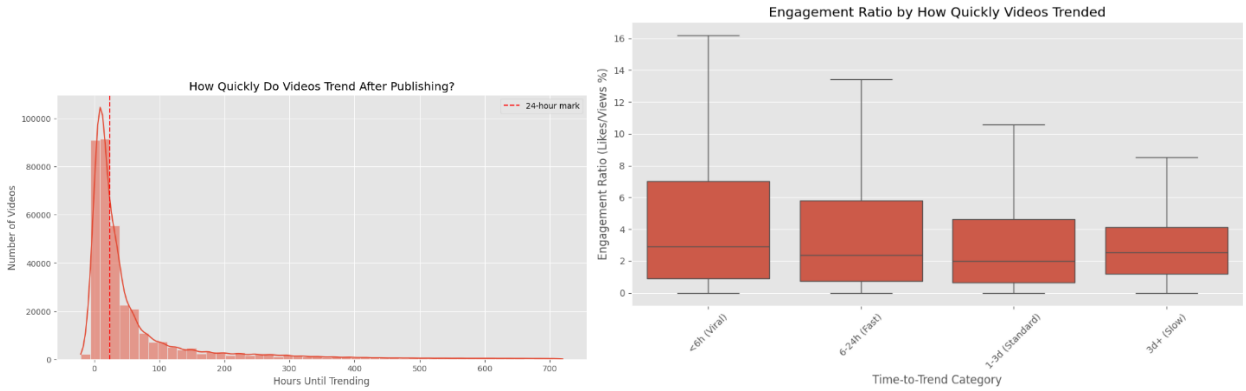
Strategy:

1. A/B Test Title Lengths:
    - Compare CTR and views for titles of varying lengths (e.g., 30 vs. 50 vs. 70 chars).
2. Prioritize Front-Loaded Keywords:
    - Place key terms in the first 40 characters to ensure visibility in truncated displays.
3. Avoid "Clickbait" Overextension:
    - Long titles may annoy users; keep them concise but compelling.

Conclusion:

Titles of moderate length (40-60 characters) strike the ideal balance—clear enough for algorithms, engaging enough for viewers. While exceptions exist (e.g., viral hooks), data suggests avoiding extremes for consistent trending success.

3.7 Does the Time-to-Trend (Gap Between Publish and Trending) Affect Video Performance?



Key Findings:

1. Most Videos Trend Quickly:
    - The majority of trending videos gain traction within 1.3 days (Standard) of publishing, with a steep drop-off after 3+ days ("Slow" trenders).
2. Engagement Ratio Differences:
    - Fast-Trending Videos (≤1.3d): Exhibit higher engagement ratios (likes, comments, shares per view) compared to slow-trending ones.
    - Slow-Trending Videos (3d+): Show lower engagement, suggesting algorithmic prioritization favors rapid audience response.

Insight:

- The "Hype Window" Matters: Videos that trend quickly likely tap into timely topics, strong hooks, or existing audience networks, triggering algorithmic promotion.
- Slow Trenders Struggle: Delayed traction may indicate weaker initial appeal or misalignment with platform trends, reducing long-term visibility.

Strategy:

1. Capitalize on Early Momentum:
    - Promote aggressively in the first 24–48 hours (e.g., shares, collaborations, ads) to trigger trending.
    - Use trending keywords/hashtags at publish time.
2. Optimize for Fast Engagement:
    - Front-load emotional hooks (surprise, urgency, humor) in titles/thumbnails.
    - Encourage early likes/comments (e.g., pinned questions).
3. Reassess Slow Performers:

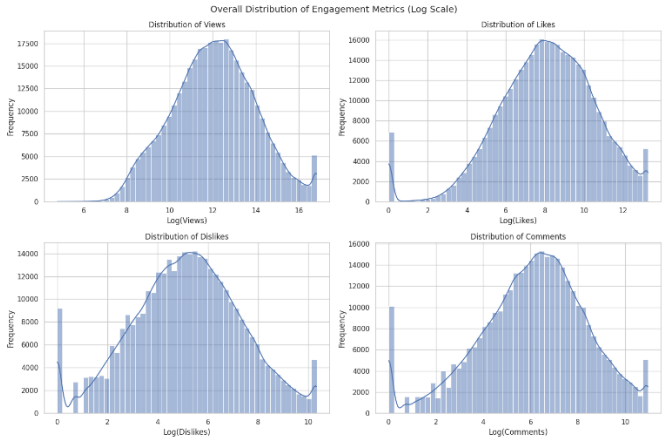
- If a video doesn't trend within 3 days, repurpose or re-promote with adjusted messaging.

Conclusion:

Speed-to-trend directly impacts performance—fast-trending videos earn higher engagement and sustained visibility. Prioritize content that resonates immediately, and leverage early signals to ride algorithmic waves.

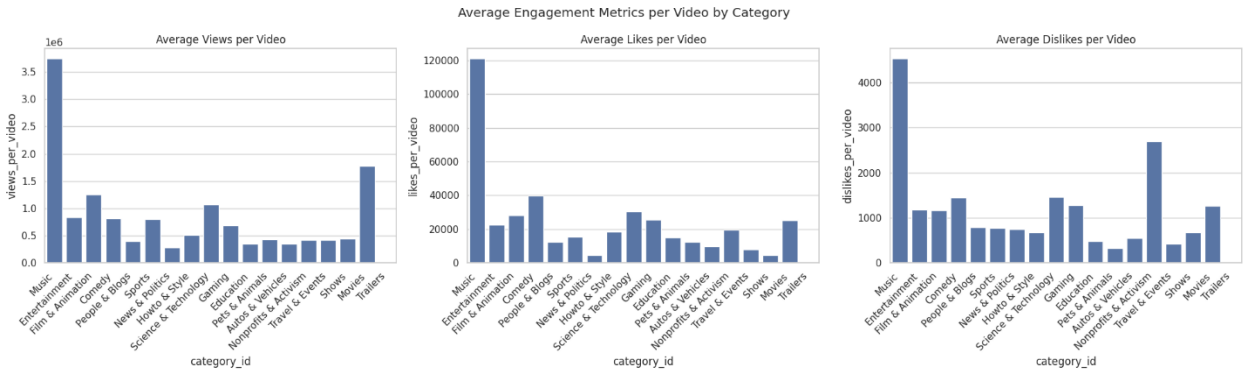
4 Additional Questions and Visualizations

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



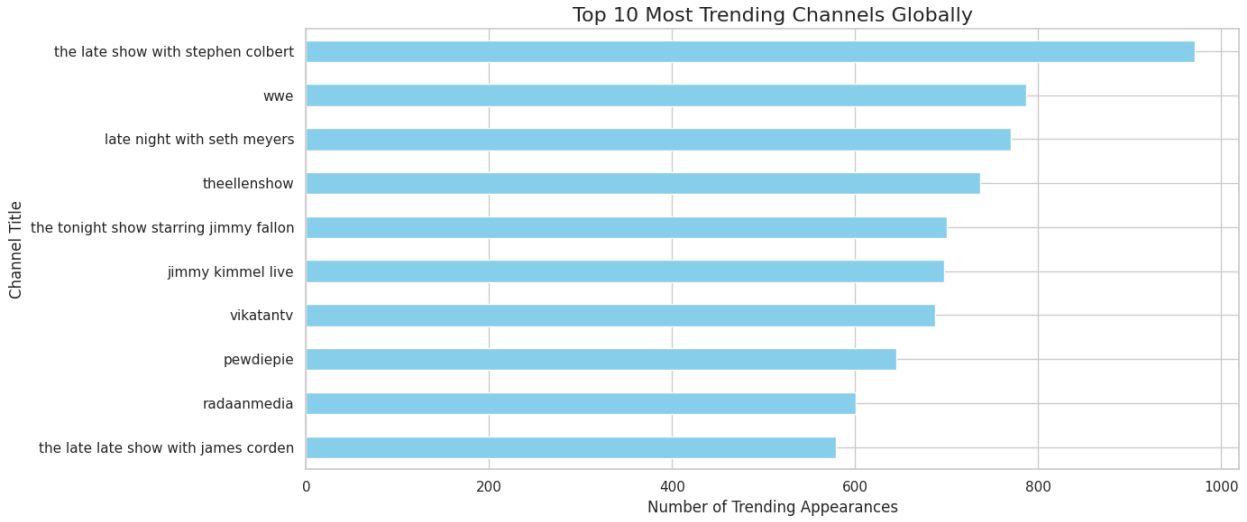
This plot effectively reveals the skewed nature of engagement metrics on YouTube, emphasizing the variability in user interaction with video content.

- Engagement Hierarchy: Views are the most frequent metric (highest range), followed by comments and dislikes, which aligns with typical YouTube behavior where not all viewers leave comments or dislikes.
- Outliers Exist: The long tails in all three distributions highlight the presence of outlier videos that achieve exceptionally high engagement (e.g., viral or controversial content).

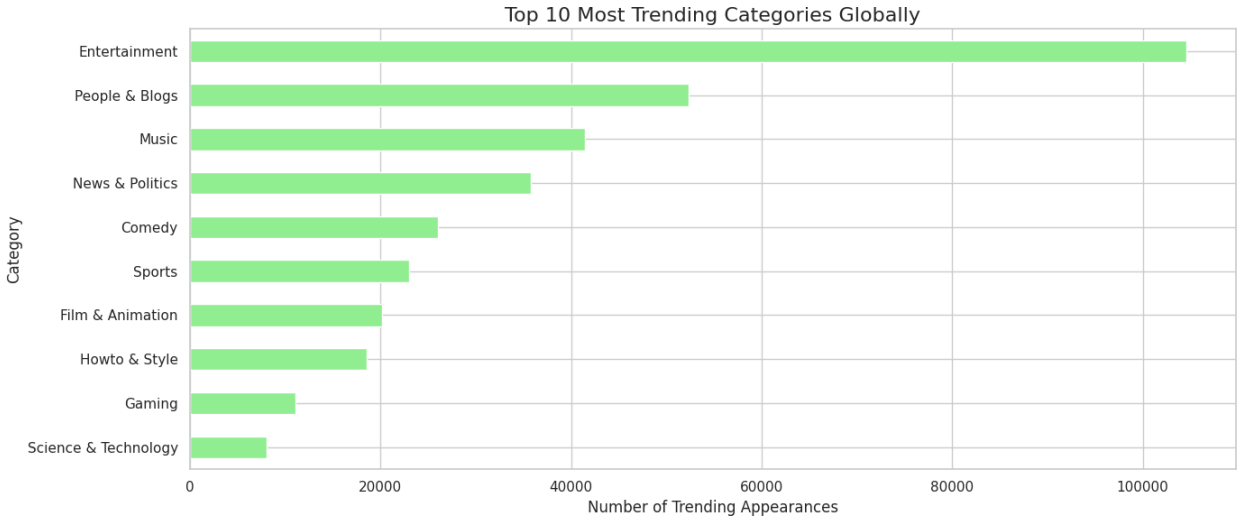


The plot likely reveals a "winner-takes-all" trend, where a few popular categories dominate in views and likes, while others lag. As we can see the Music category has the most engagement rate in comparison to other categories.

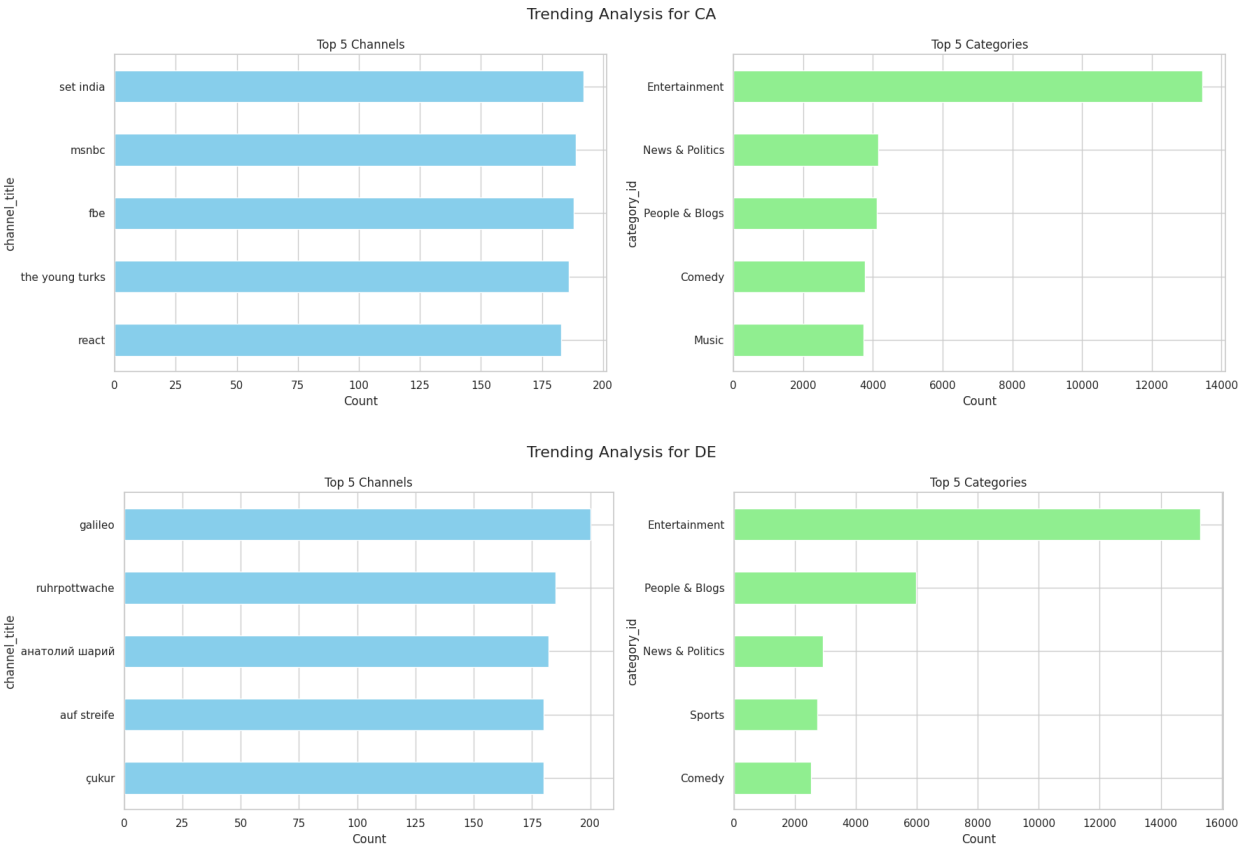
4.2 Which YouTube channels and video categories trend the most in each country and globally?



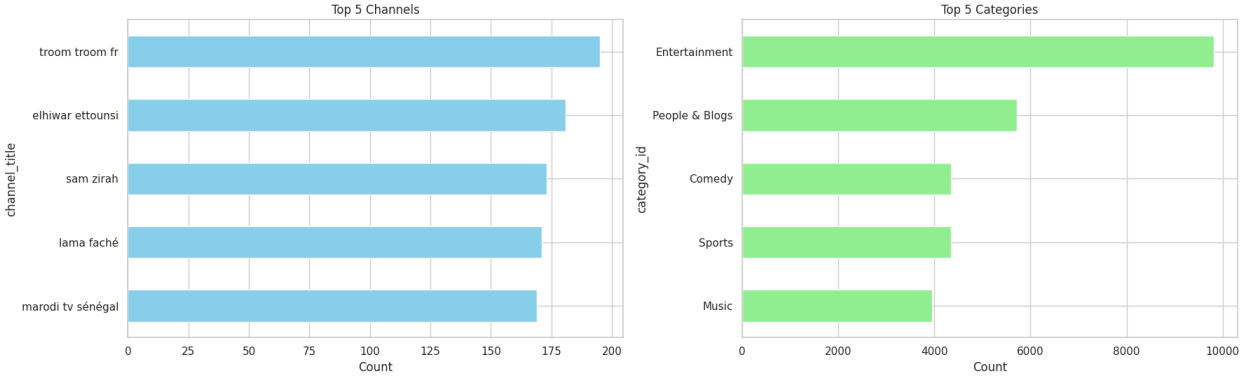
In this plot we can see the top 10 most trending channels among all the 10 countries we have in our dataset.



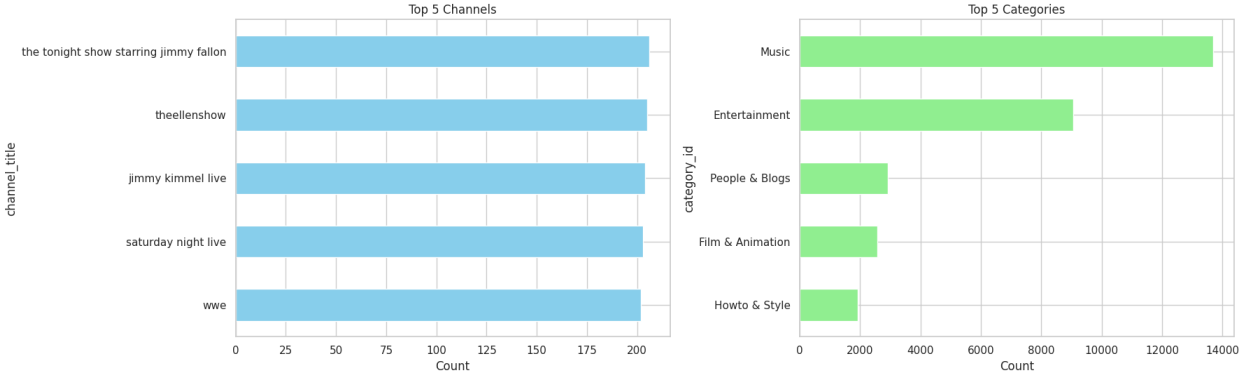
In this plot we can see the top 10 most trending categories among all the 10 countries we have in our dataset. As we can see Entertainment comes first with a meaningful difference with the other categories.



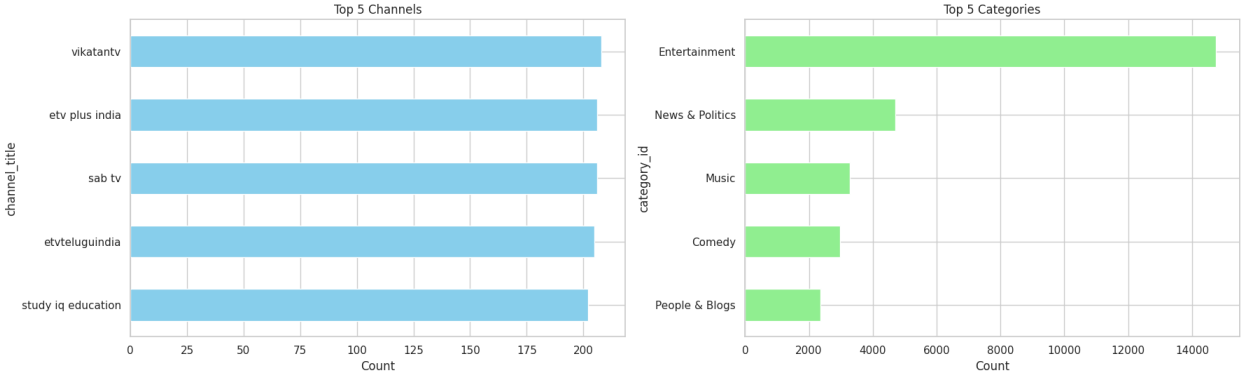
Trending Analysis for FR



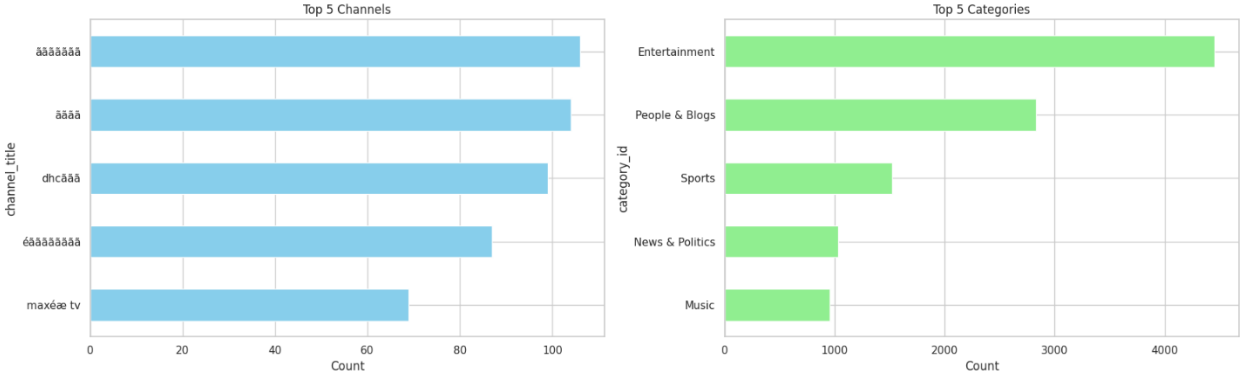
Trending Analysis for GB



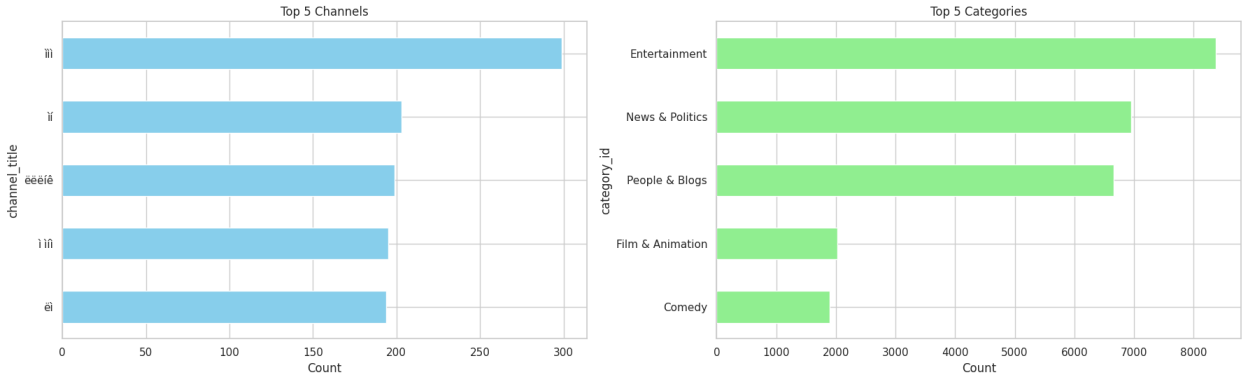
Trending Analysis for IN



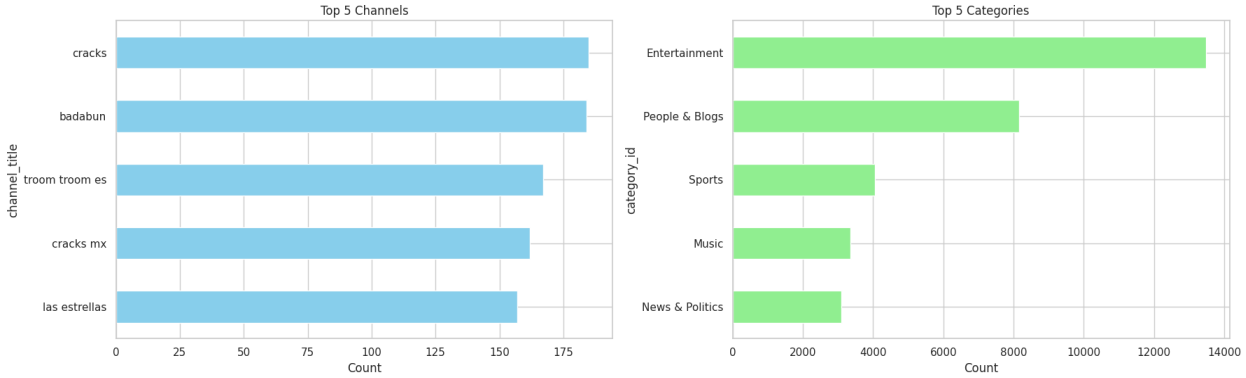
Trending Analysis for JP



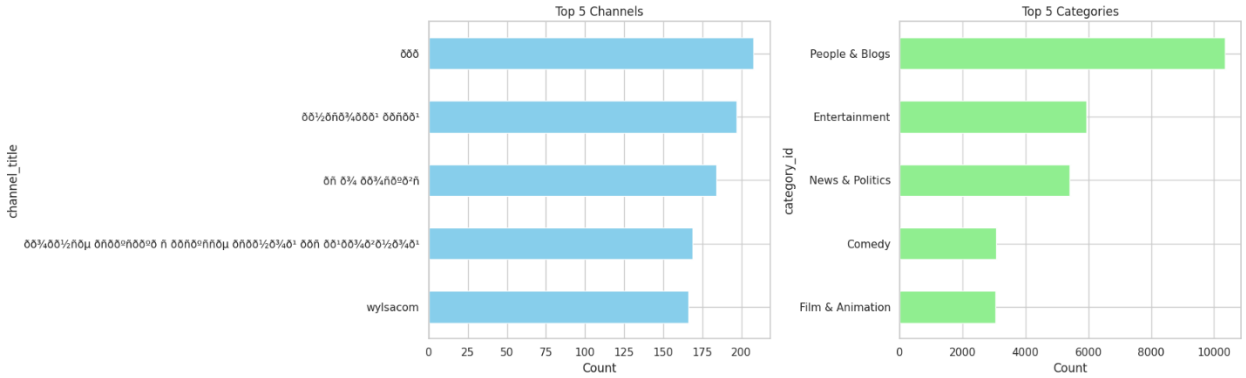
Trending Analysis for KR



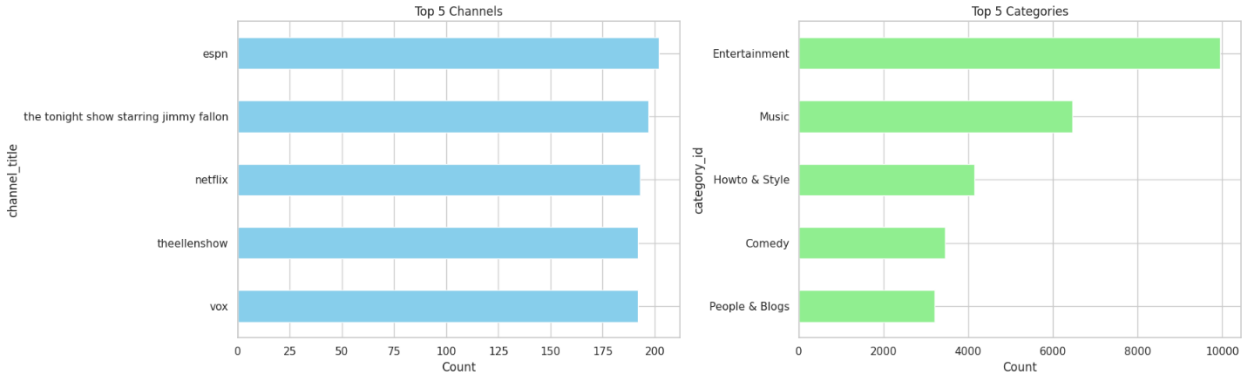
Trending Analysis for MX



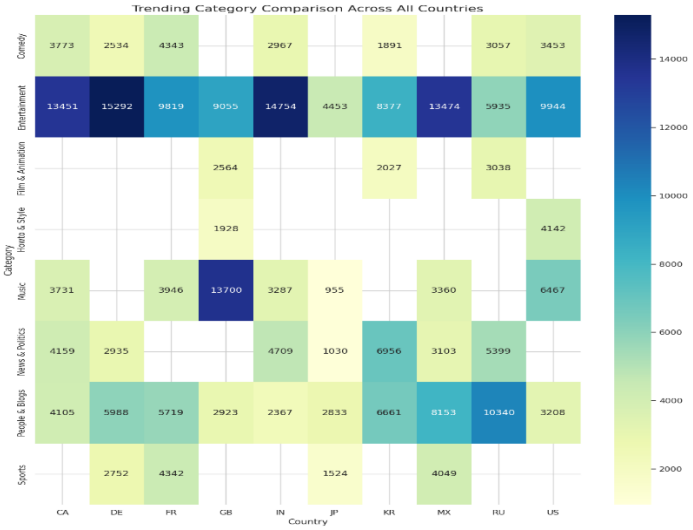
Trending Analysis for RU



Trending Analysis for US

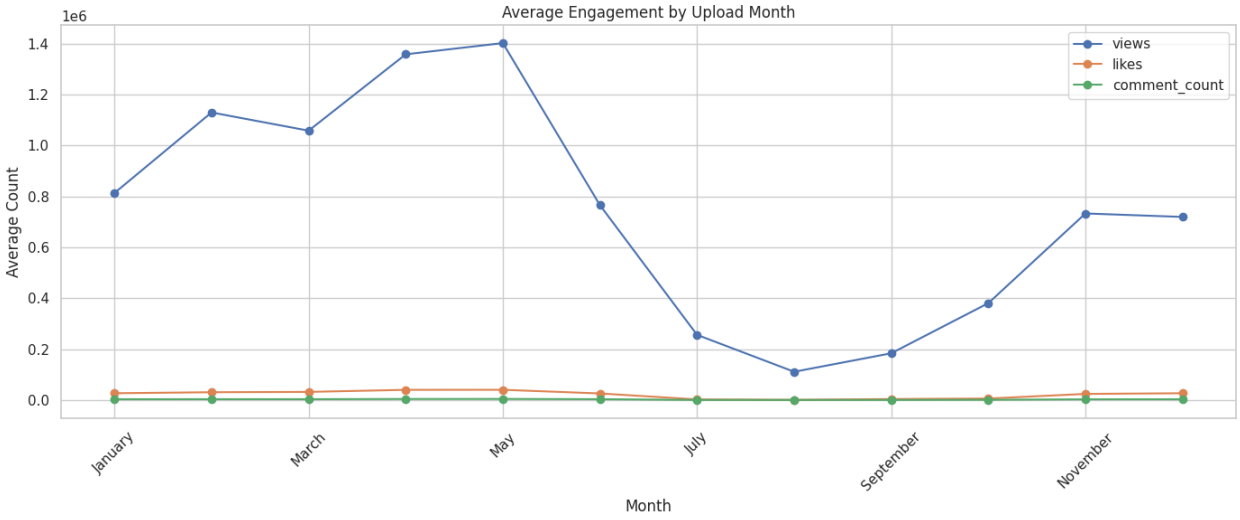


In these figures we can see the top 5 most trending categories and the channels in each country.

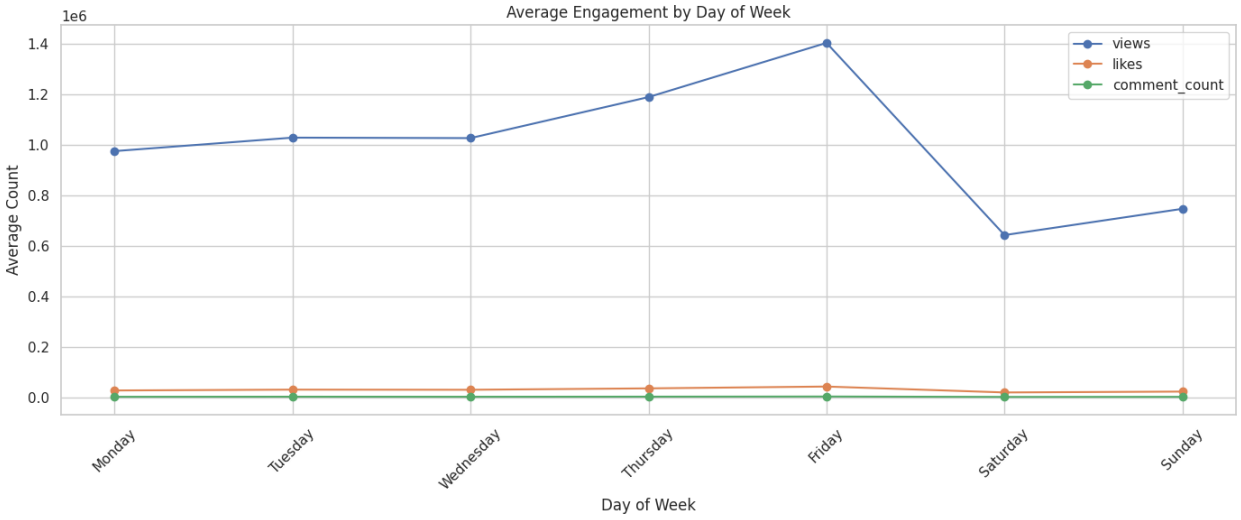


This figure indicates the most trending category for most of the countries in Entertainment.

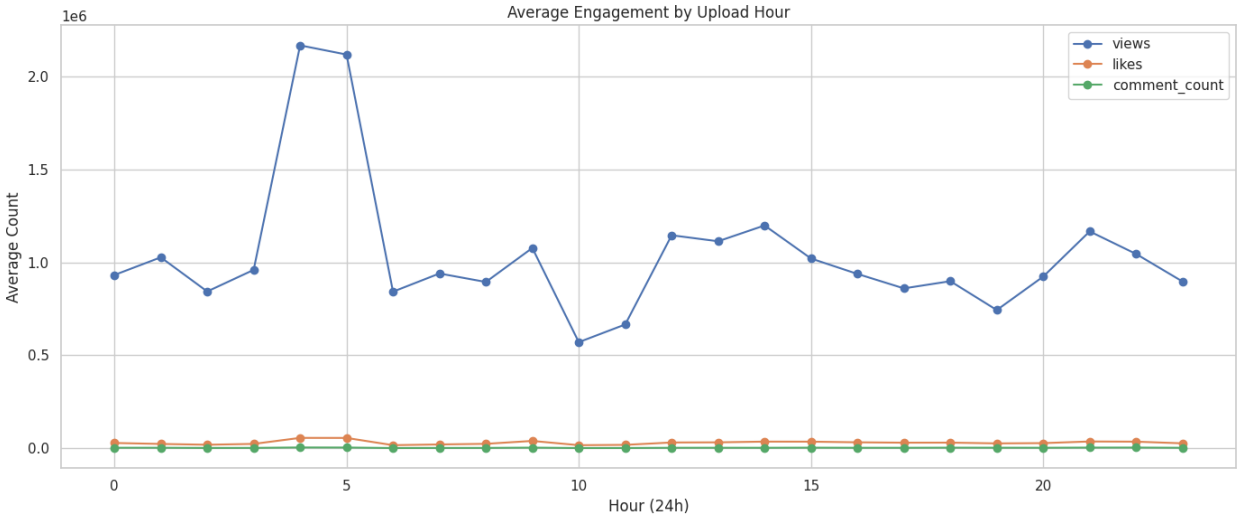
4.3 Are there seasonal or day-of-week patterns in trending videos? How does the upload day and time impact video engagement?



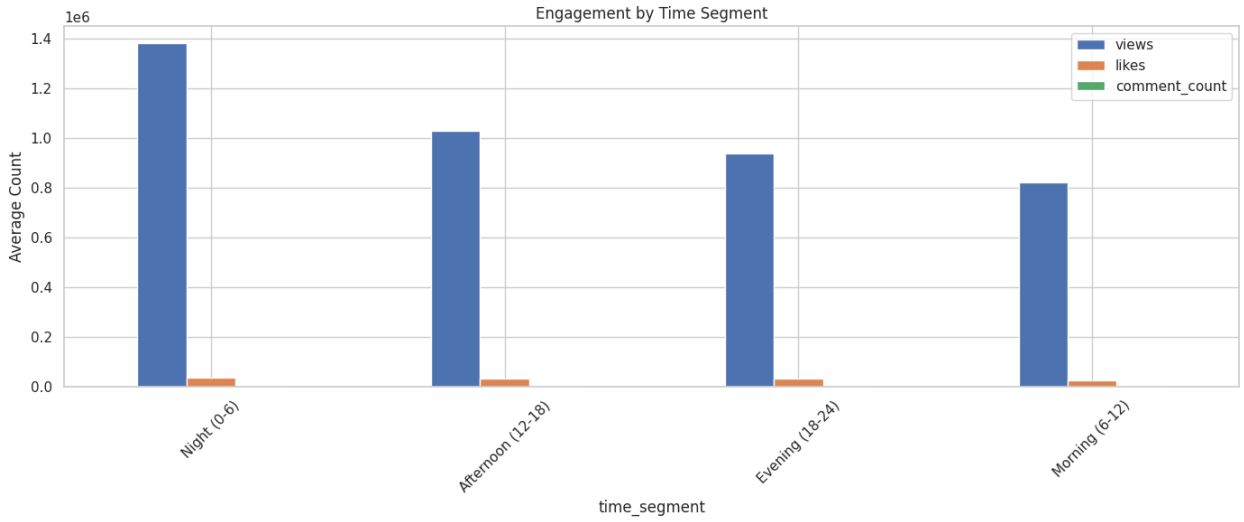
This plot clearly indicates that the engagement seems to be more on some months of year (March – May) and we can observe the least amount of engagement during July, August and September.



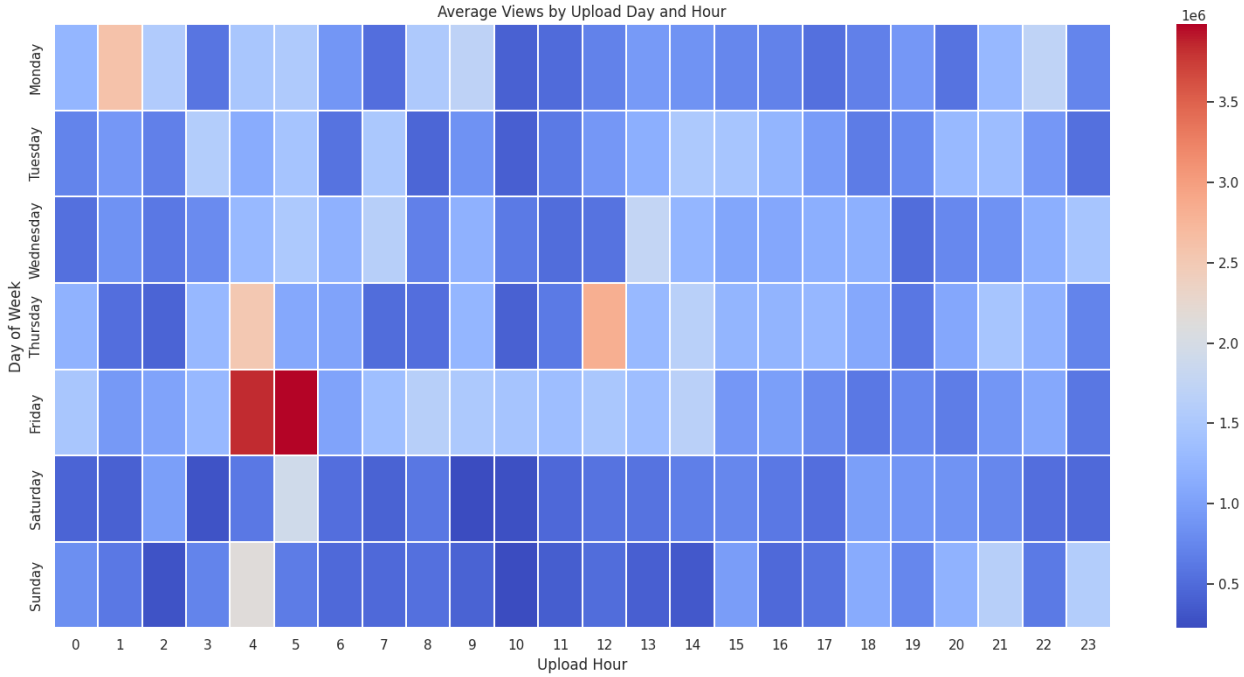
As we can observe the engagement rate is ascendant from Wednesday to Friday and on Friday it reaches its highest point.



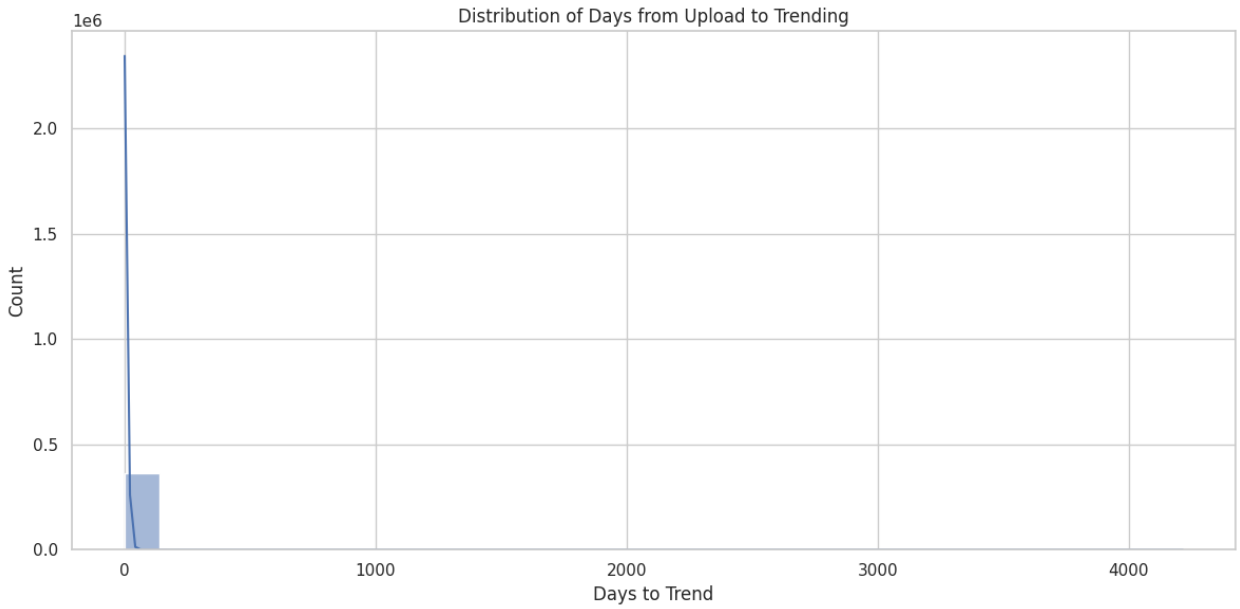
As we can observe the engagement rate peaks when the uploaded hour is during 3 to 5 AM.



In this plot we observe that the engagement seems to be more in the nights in comparison with other parts of the day.

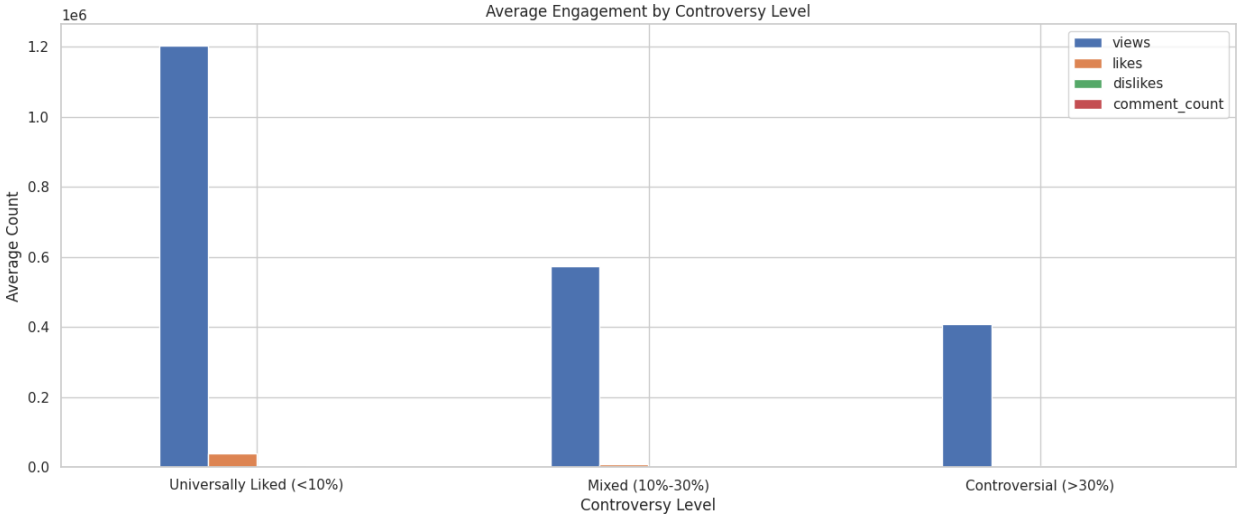


According to this plot, the Views rate is at its highest on Fridays during 3 and 4 AM.

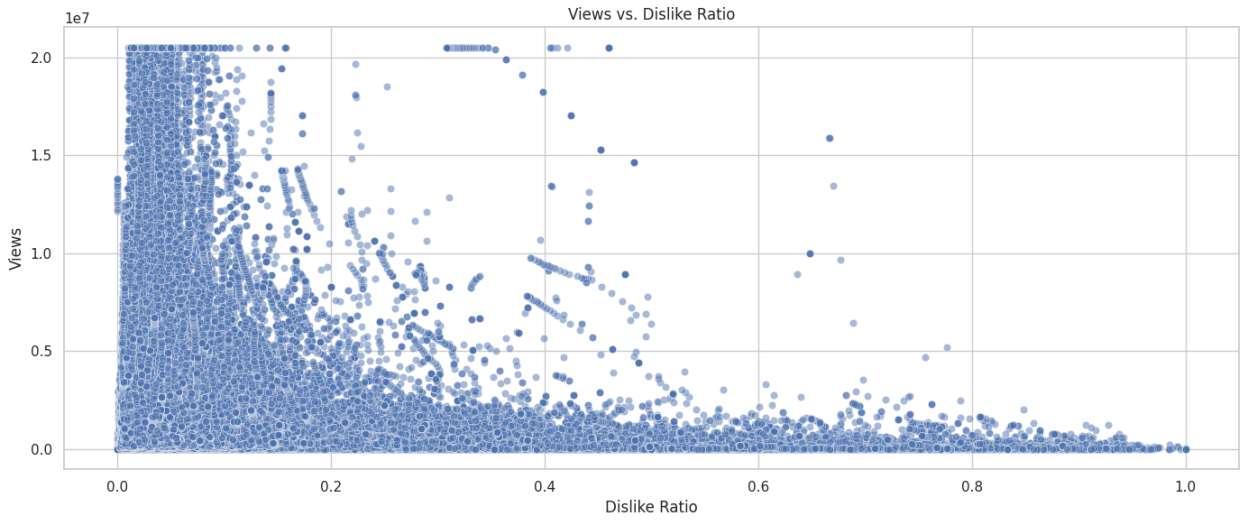


This plot indicates that almost all the trending videos became trends in a very short time after being uploaded (less than a day).

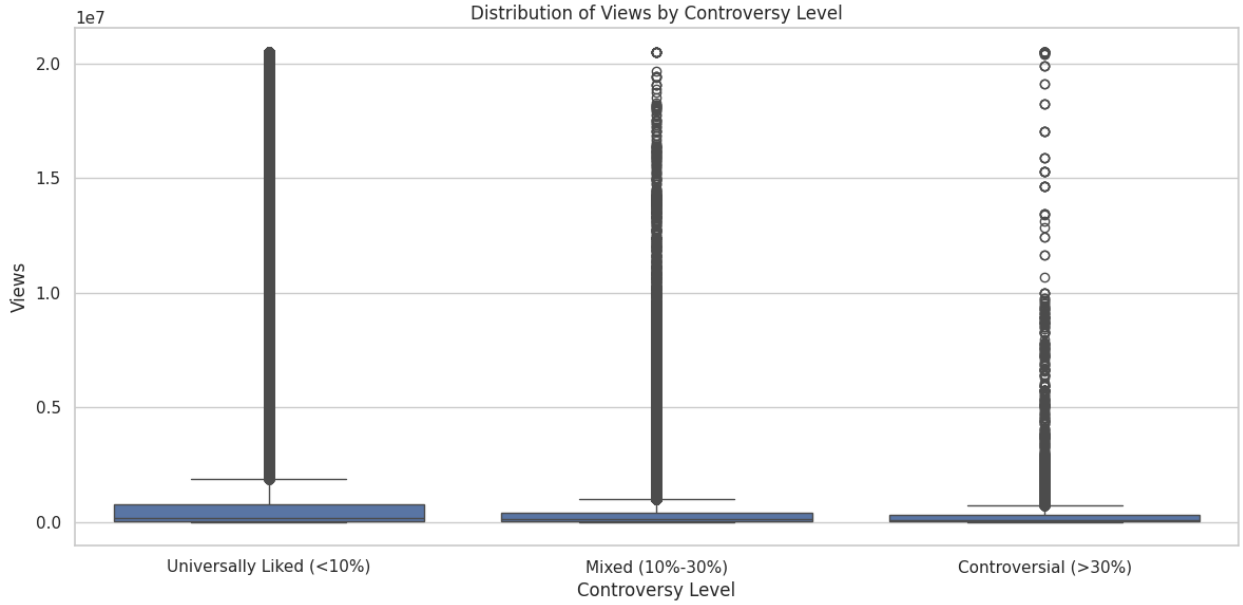
4.4 Do controversial videos, defined by a high dislike ratio, receive more engagement than universally liked ones?



This plot shows that the Engagement rate is meaningfully higher in videos with less controversy levels.

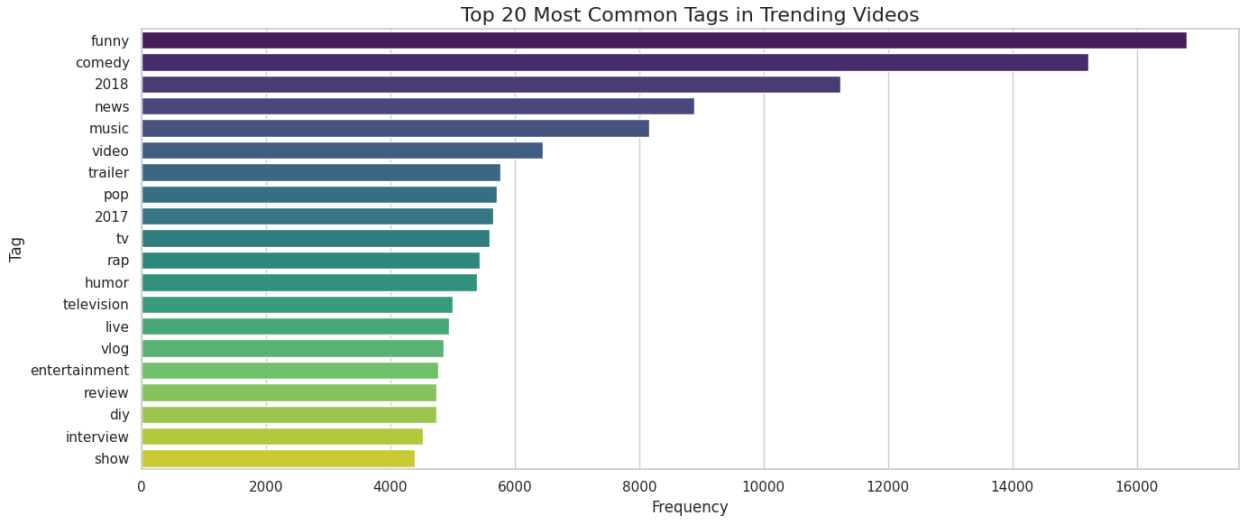


The more dislikes the videos get the less views they get.

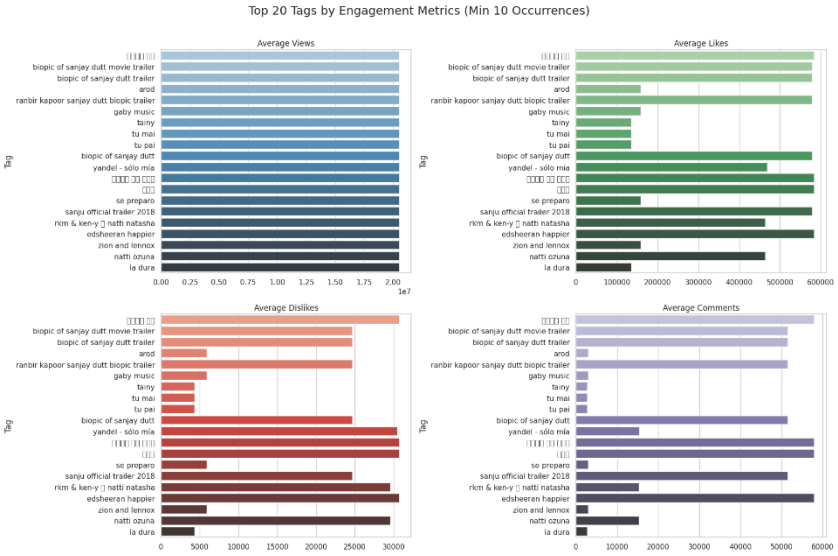


Videos with less controversy levels get more views.

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

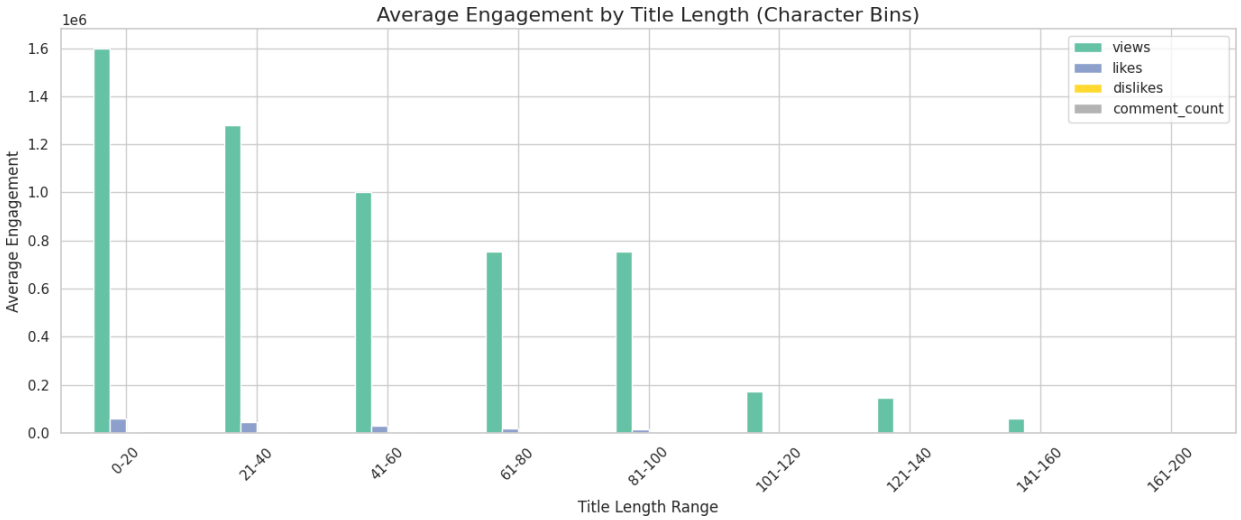


This plot indicates the top 20 common tags in trending videos. “Funny” and “Comedy” are the most repeated ones.



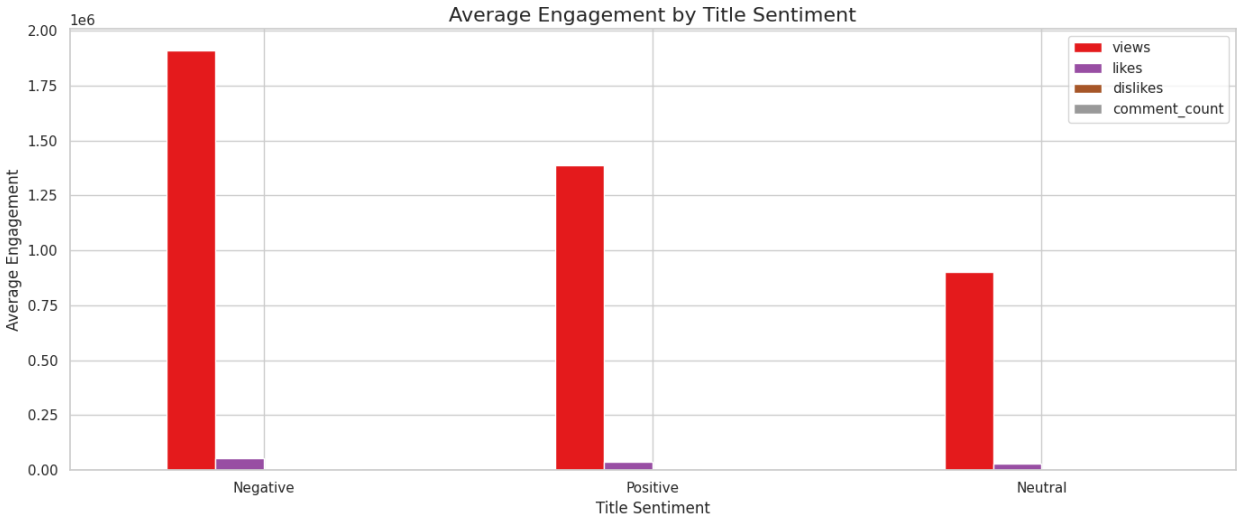
Here we see the top 20 tags by engagement metrics (views, likes, dislikes and comments) individually.

4.6 How does the length of a video title impact engagement levels?



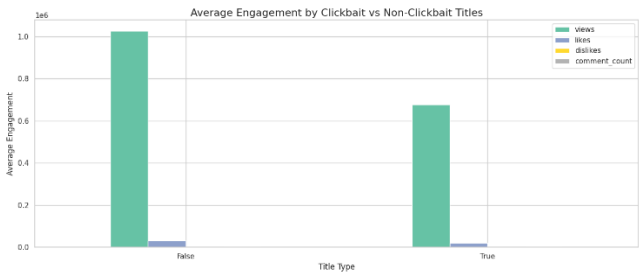
This plot shows how the title length of the video impacts the engagement level.

4.7 s there a relationship between video title sentiment, whether positive, neutral, or negative, and engagement levels? (extra point)



It seems like the videos with a negative title sentiment will attract more engagement than a positive or neutral title.

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



In this plot we can observe the effect of clickbait on Engagement.

5 Statistical Questions

5.1.1 Hypotheses:

H_o (Null Hypothesis): There is no association between the day of the week a video is published and its likelihood of trending. (They are independent.)

H_1 (Alternative Hypothesis): There is an association between the day of the week a video is published and its likelihood of trending. (They are dependent.)

5.1.2 Test:

Chi-Square Test

5.1.3 Result:

Chi2 Statistic: 0.0675

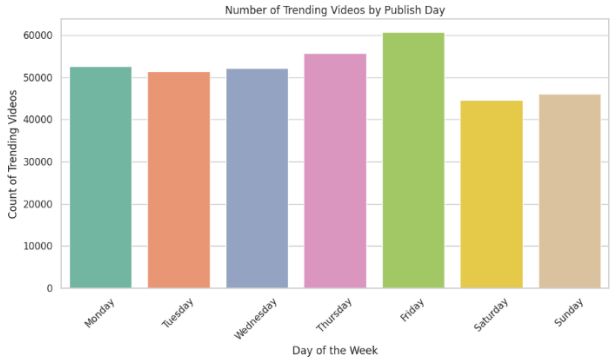
Degrees of Freedom: 6

P-value: 1.0000

Fail to reject H_o: No significant association between publish day and likelihood of trending.

5.1.4 Visualization:

Although the Chi-Square test does not find a statistically significant association, visual analysis reveals that Fridays see a noticeably higher number of trending videos, suggesting a potential trend worth exploring further.



5.2.1 Hypotheses:

H_o (null hypothesis): The distribution of engagement (likes-to-views ratio) is the same across all categories.

H_1 (alternative hypothesis): At least one category has a different distribution of engagement than the others

5.2.2 Test:

Kruskal-Wallis H-test

5.2.3 Result:

H statistic = 45434.7984

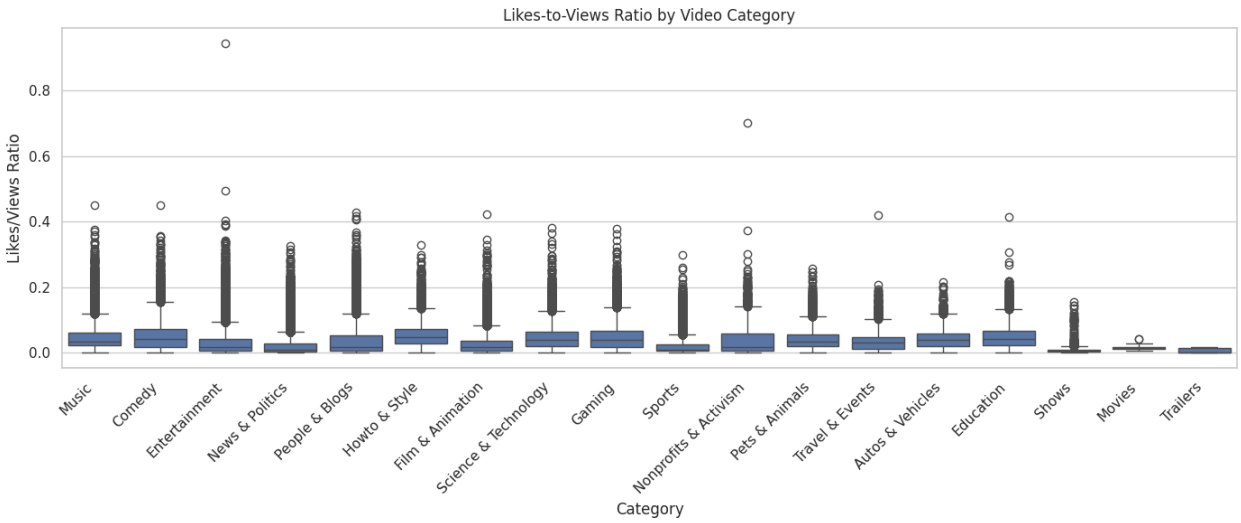
p-value = 0.0000

Significant difference in engagement between categories.

5.2.4 Visualization:

Here we can see the plot of category names v.s Likes/views ratio.

As we can see, there is a significant difference in engagement between categories.



Conclusion

In this project, we conducted a deep analysis of trending videos on YouTube to determine the most critical drivers of virality. From an analysis of engagement behaviors, content attributes, and geography trends, we formulated actionable insights that can help content creators, marketers, and platform planners better calibrate their strategy to drive maximum reach and engagement.

References

[1] https://www.kaggle.com/datasets/datasnaek/youtube-new/

Maryam Soleimani 401222075