Exploratory Data Analysis

on

Trending YouTube Videos

By

Venugopal Katragadda(A20544776)

Sannihitha Gudimalla (A20560248)

Sruthi Kondapalli (A20554780)

Bezawada Sai Sravanya (A20561552)

Executive Summary

Key research issue

This research aims to explore the dynamics behind YouTube trending videos and pinpoint the critical factors that drive their success. Through a comprehensive analysis, we investigated various aspects of trending content to gain meaningful insights into what contributes to a video's popularity on the platform.

Findings

Our analysis provided important insights into several factors influencing trending videos, including optimal category choices, ideal title lengths, and the influence of sentiment in descriptions on viewership. We also compared our results with existing trends to offer a broader perspective on the data. However, our efforts to apply cluster analysis for deeper insights were less successful than anticipated.

Future Work

Future work could involve a deeper examination of the identified clusters and the development of predictive models to forecast a video's likelihood of trending. Such models would leverage key input features to not only predict trending potential but also provide strategic recommendations for optimizing content to enhance its chances of achieving trending status.

Overview of project

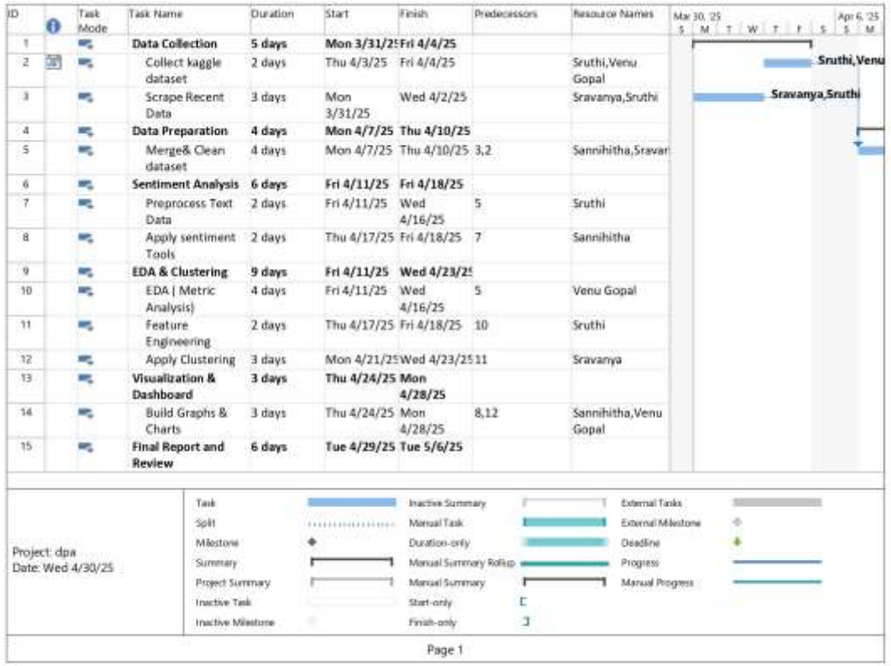
Methodology

DATA COLLECTION

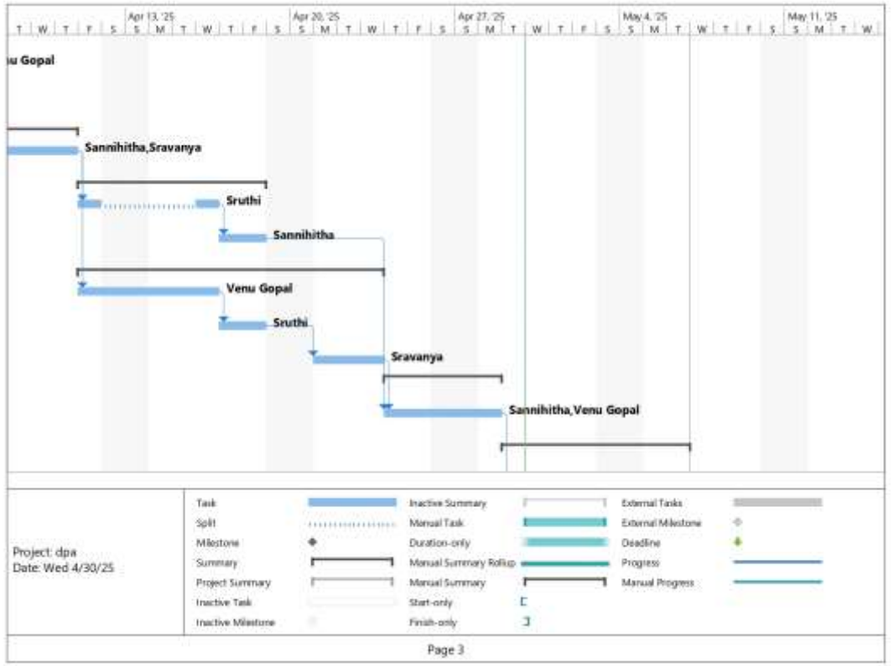
DATA PREPROCESSING

EXPLORATORY DATA ANALYSIS (EDA)

CLUSTER ANALYSIS



ID	Task Mode	Task Name	Duration	Start	Finish	Predecessors	Resource Names
1		**Data Collection**	**5 days**	**Mon 3/31/25**	**Fri 4/4/25**		
2		Collect kaggle dataset	2 days	Thu 4/3/25	Fri 4/4/25		Sruthi,Venu Gopal
3		Scrape Recent Data	3 days	Mon 3/31/25	Wed 4/2/25		Sravanya,Sruthi
4		**Data Preparation**	**4 days**	**Mon 4/7/25**	**Thu 4/10/25**		
5		Merge& Clean dataset	4 days	Mon 4/7/25	Thu 4/10/25	3,2	Sannihitha,Sravar
6		**Sentiment Analysis**	**6 days**	**Fri 4/11/25**	**Fri 4/18/25**		
7		Preprocess Text Data	2 days	Fri 4/11/25	Wed 4/16/25	5	Sruthi
8		Apply sentiment Tools	2 days	Thu 4/17/25	Fri 4/18/25	7	Sannihitha
9		**EDA & Clustering**	**9 days**	**Fri 4/11/25**	**Wed 4/23/25**		
10		EDA (Metric Analysis)	4 days	Fri 4/11/25	Wed 4/16/25	5	Venu Gopal
11		Feature Engineering	2 days	Thu 4/17/25	Fri 4/18/25	10	Sruthi
12		Apply Clustering	3 days	Mon 4/21/25	Wed 4/23/25	11	Sravanya
13		**Visualization & Dashboard**	**3 days**	**Thu 4/24/25**	**Mon 4/28/25**		
14		Build Graphs & Charts	3 days	Thu 4/24/25	Mon 4/28/25	8,12	Sannihitha,Venu Gopal
15		**Final Report and Review**	**6 days**	**Tue 4/29/25**	**Tue 5/6/25**		



Project Plan

Updated Plan:

1. We made a lot more graphs and charts than we first thought. These visuals helped us see trends in the YouTube data more clearly.

2. We faced unexpected problems with the YouTube data scrapping. It took us longer to get the data we needed, but we found solutions and kept going with our analysis.

Data Sources

Kaggle Dataset

For the analysis we have chosen a data set which has the following characteristics.

Dataset Type: Multivariate

Dataset Size: 10 documents categorized by country, each having 16 columns and around 20000 records.

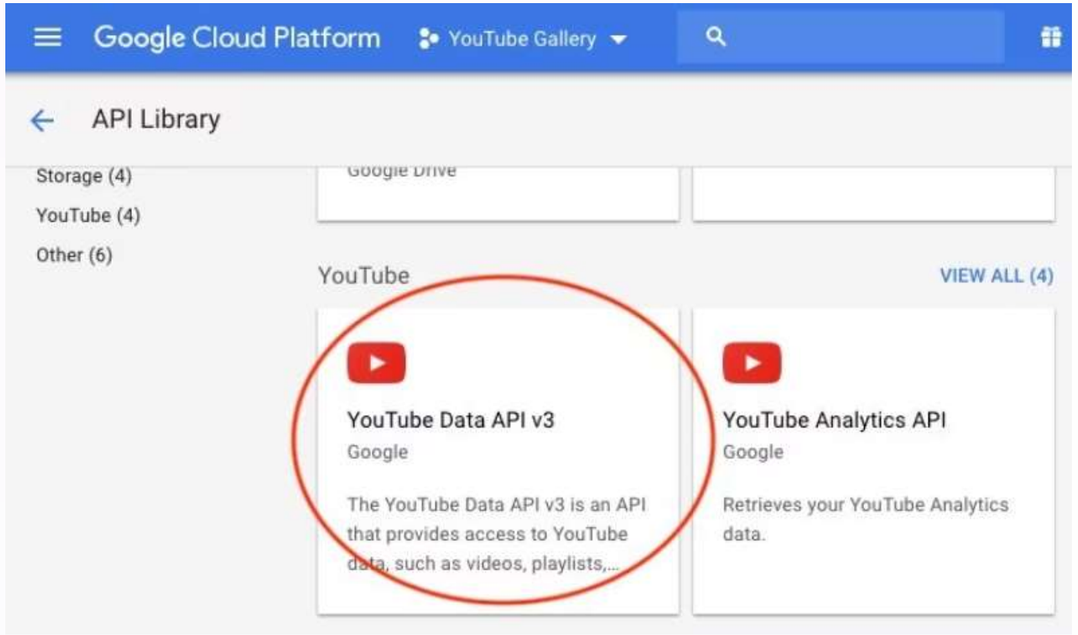
Countries chosen: Canada, USA, Great Britain, France, Germany, Russia, Mexico, South Korea, India, Japan

Missing Values: Yes

Feature Descriptions:

1. **video_id:** A unique identifier for each video on YouTube. (String)
2. **title**: The title of the video. (String)
3. **trending_date:** Date when the video appeared on the trending list. (String)
4. **channel_title:** Name of the YouTube channel that uploaded the video. (String)
5. **category_id:** Category to which the video belongs. (Integer)
6. **published_time:** Date and time when the video was published on YouTube. (DateTime)
7. **tags**: Keywords or phrases that describe the content of the video. (String)
8. **views:** Number of views. (Integer)
9. **likes**: Number of likes. (Integer)
10. **dislikes:** Number of dislikes. (Integer)
11. **comment_count:** Number of comments. (Integer)
12. **thumbnail_link:** Link to the thumbnail image that represents the video. (String)
13. **comments_disable:** Whether comments have been disabled for the video. (Boolean)
14. **ratings_disable:** Whether ratings have been disabled for the video. (Boolean)
15. **video_error_or_removed:** Whether there was an error with the video or if the video has been removed from YouTube. (Boolean)
16. **description:** Description of the video. (String)

YouTube API



Data Preprocessing

1

Converting the categorical variable `category_ID` and incorporating a new `country` variable.

2

Handling missing values and Duplicates

3

Convert the 'trending_date' and 'published_time' columns to datetime format

4

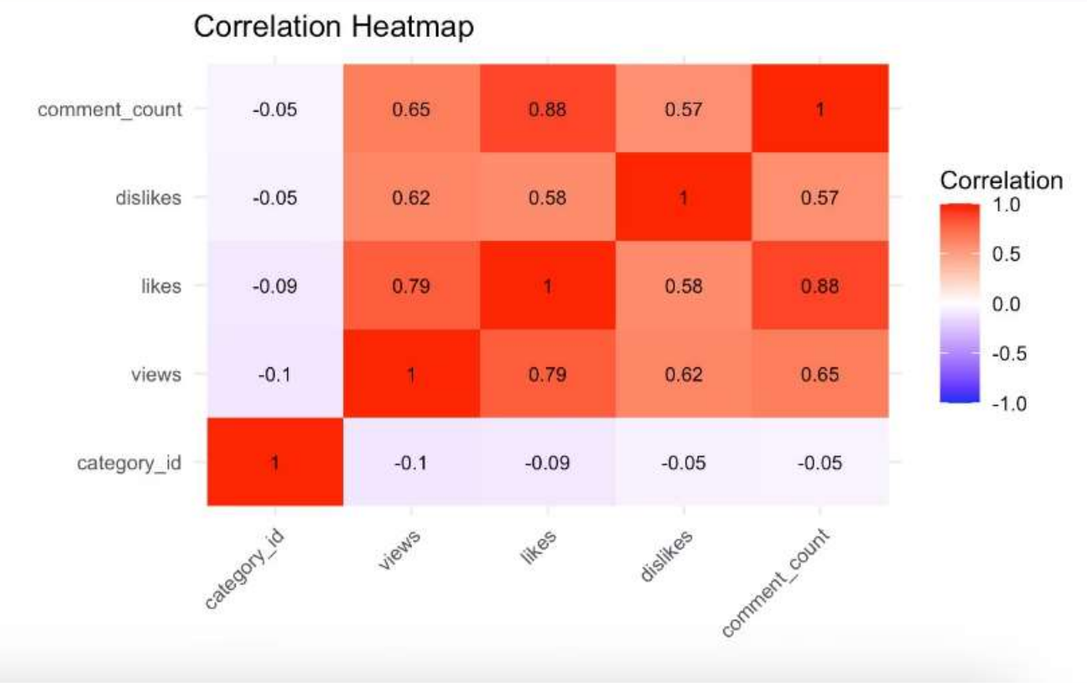
Extracting from 'published_time' such as the day of the week or hour of the day the video was published.

5

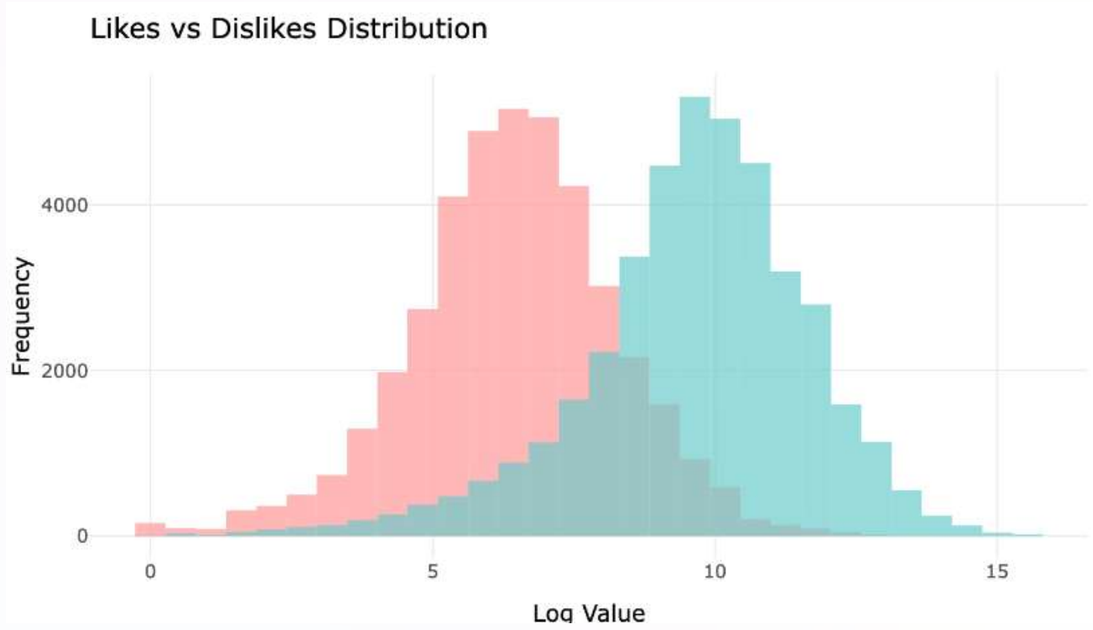
Text pre - processing for Sentiment analysis

Exploratory Data Analysis

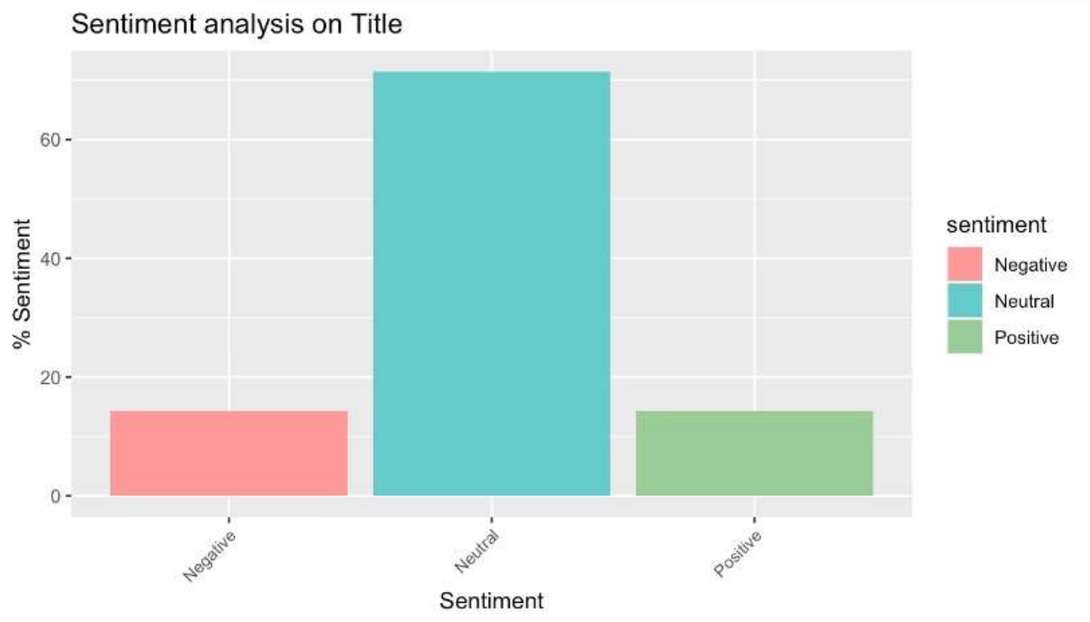
CORRELATION MATRIX

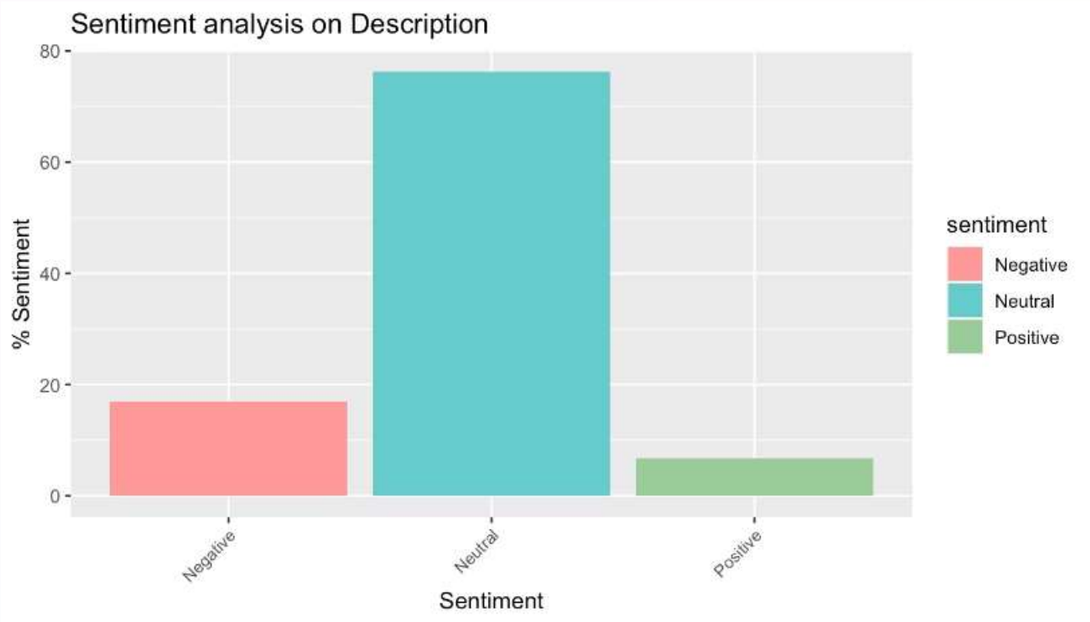


DISTRIBUTION PLOT

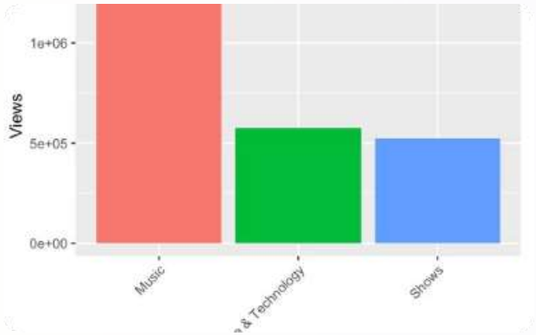


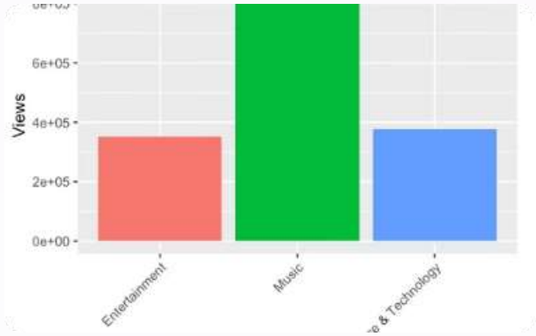
Sentiment Analysis

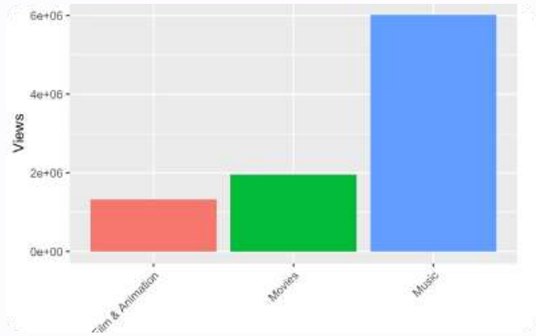


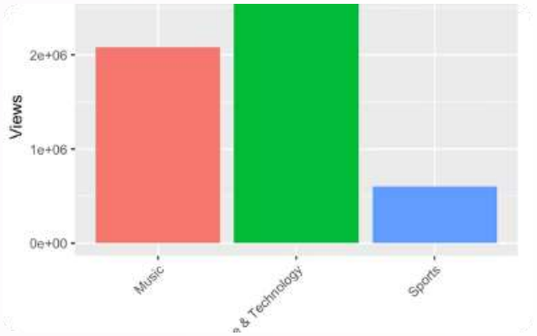


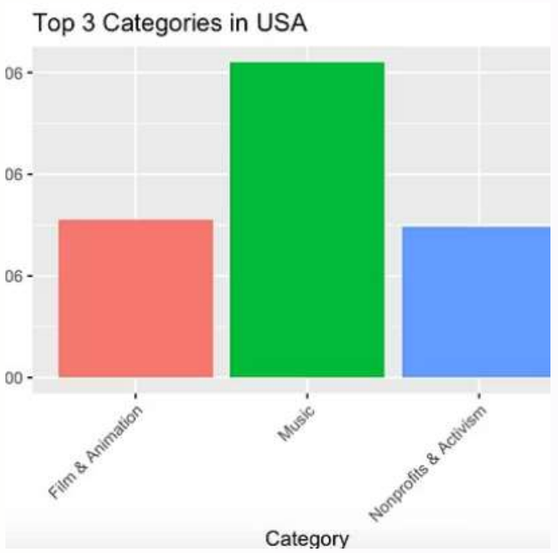
WHAT ARE THE TOP 3 CATEGORIES BASED ON THE NUMBER OF VIEWS, FOR EACH OF THE COUNTRIES?

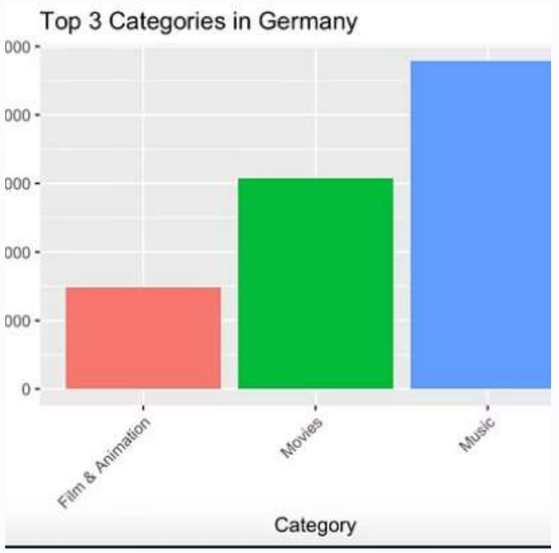


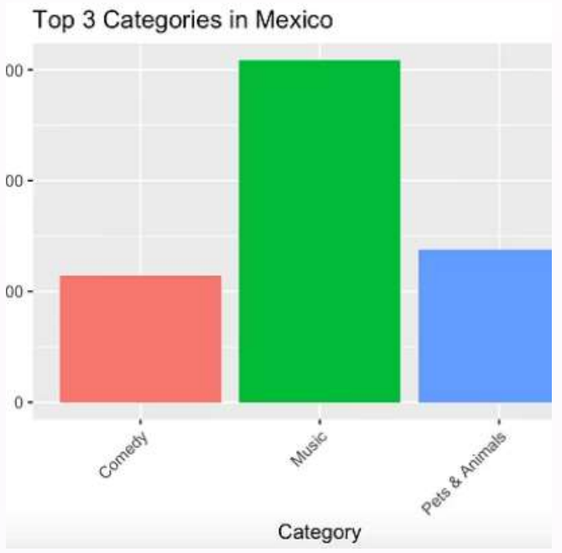


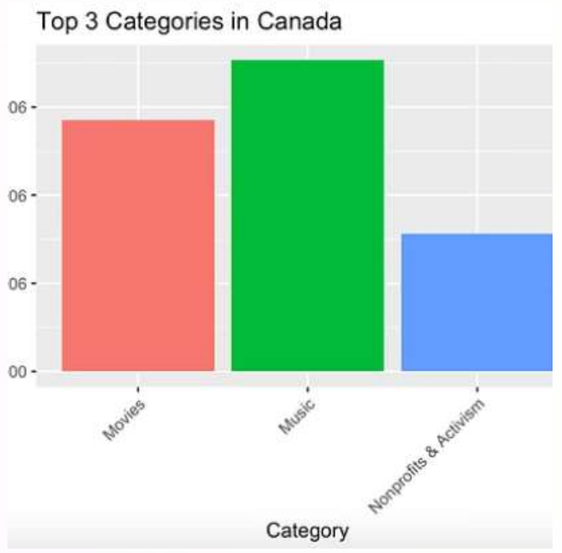


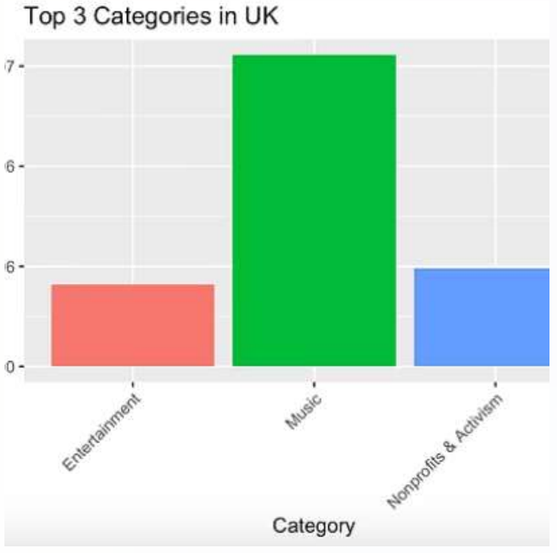


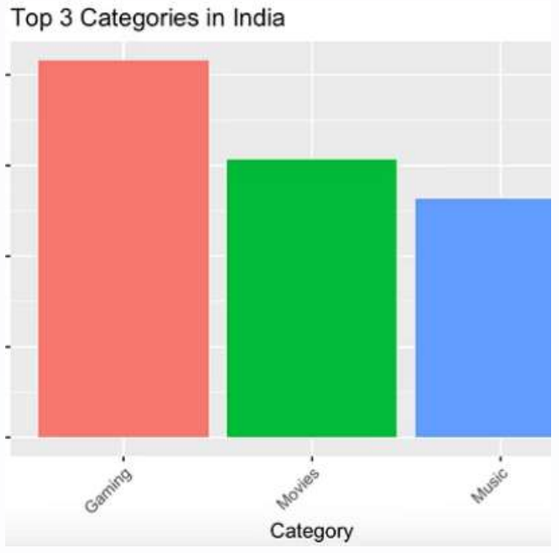

Top 3 Categories in USA
Film & Animation
Music
Nonprofits & Activism
Category


Top 3 Categories in Germany
Film & Animation
Movies
Music
Category


Top 3 Categories in Mexico
Comedy
Music
Pets & Animals
Category


Top 3 Categories in Canada
Movies
Music
Nonprofits & Activism
Category


Top 3 Categories in UK
Entertainment
Music
Nonprofits & Activism
Category


Top 3 Categories in India
Gaming
Movies
Music
Category

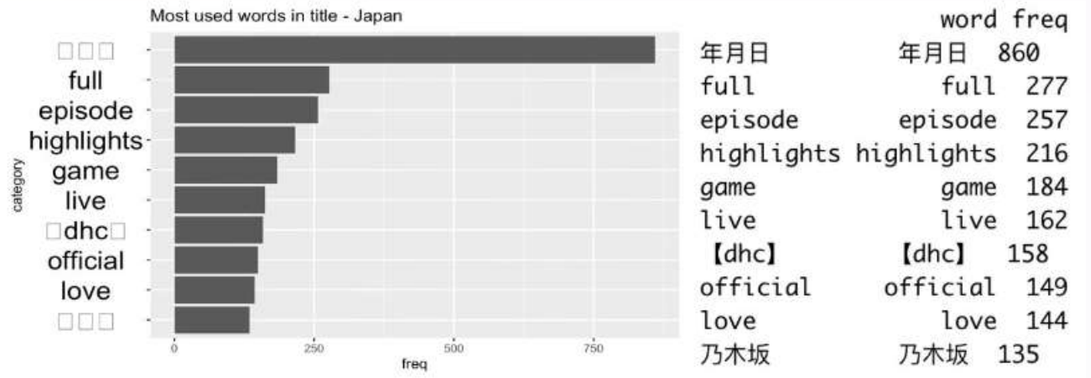
Countries with highest least views

```
> head(IN_most_category,3)
# A tibble: 3 × 3
  category_name average_views least_views
  <chr>                 <dbl>       <dbl>
1 Gaming              4162462.       72964
2 Movies              3065001.      165601
3 Music               2631116.       10971
```

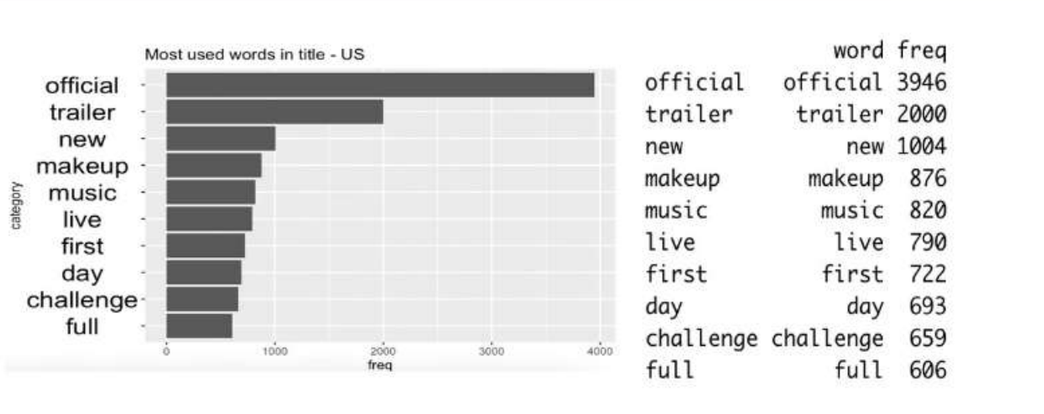
```
> head(CA_most_category,3)
# A tibble: 3 × 3
  category_name         average_views least_views
  <chr>                         <dbl>       <dbl>
1 Music                       3532525.        3201
2 Movies                      2853415       225528
3 Nonprofits & Activism       1562184.        1898
```

```
> head(GB_most_category,3)
# A tibble: 3 × 3
  category_name         average_views least_views
  <chr>                         <dbl>       <dbl>
1 Music                      12444443.        2152
2 Nonprofits & Activism       3919981.       19270
3 Entertainment               3264608.        2650
```

ARE THERE SPECIFIC KEYWORDS IN THE TITLE THAT ARE MORE LIKELY TO RESULT IN A VIDEO TRENDING?

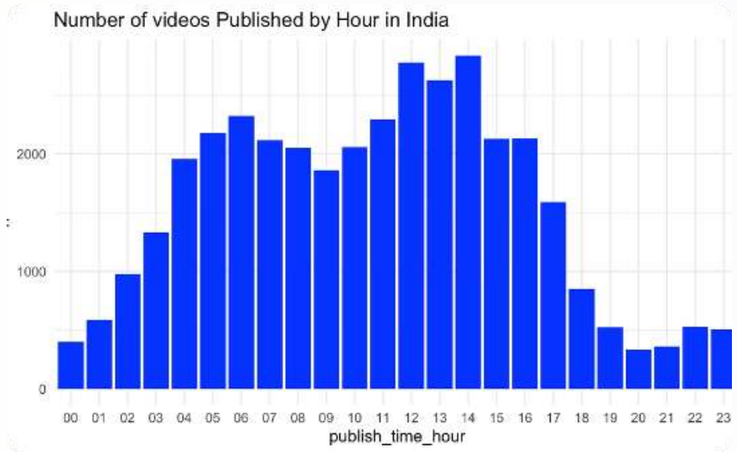


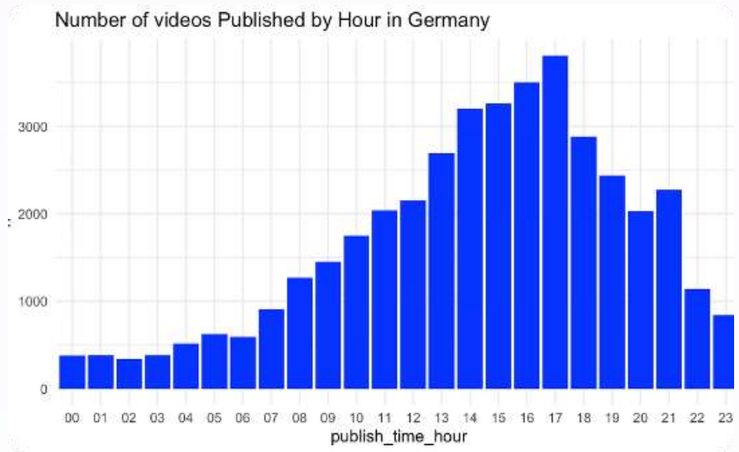
	word	freq
年月日	年月日	860
full	full	277
episode	episode	257
highlights	highlights	216
game	game	184
live	live	162
【dhc】	【dhc】	158
official	official	149
love	love	144
乃木坂	乃木坂	135

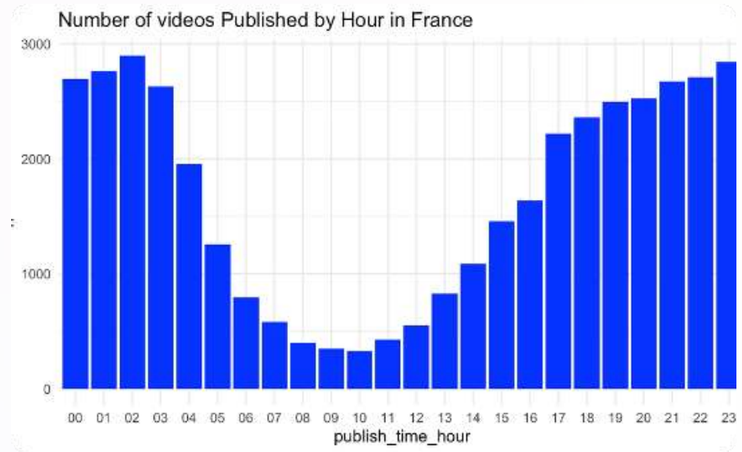


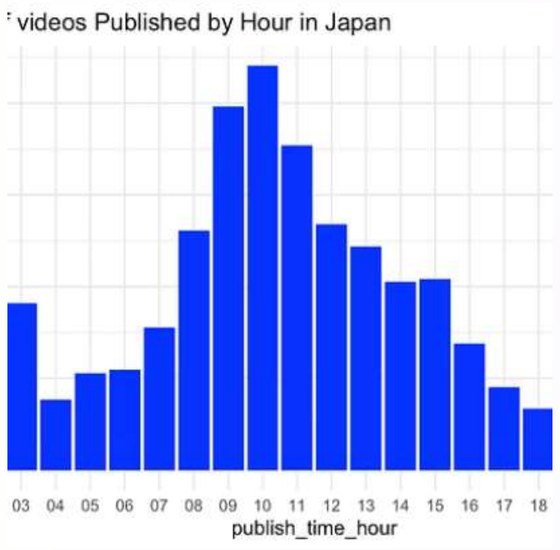
	word	freq
official	official	3946
trailer	trailer	2000
new	new	1004
makeup	makeup	876
music	music	820
live	live	790
first	first	722
day	day	693
challenge	challenge	659
full	full	606

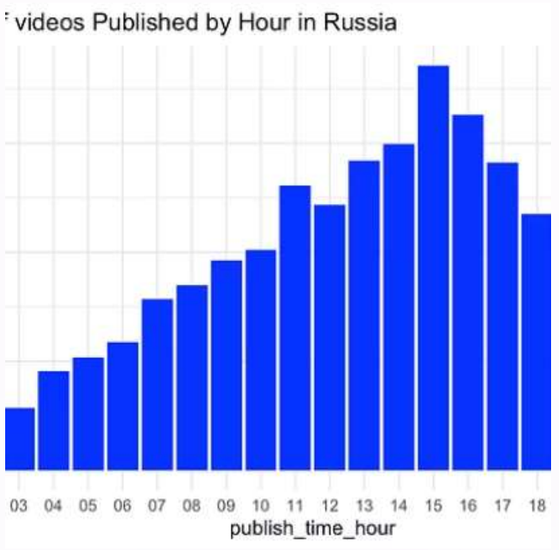
HOUR OF THE DAY WHEN TRENDING VIDEOS ARE UPLOADED

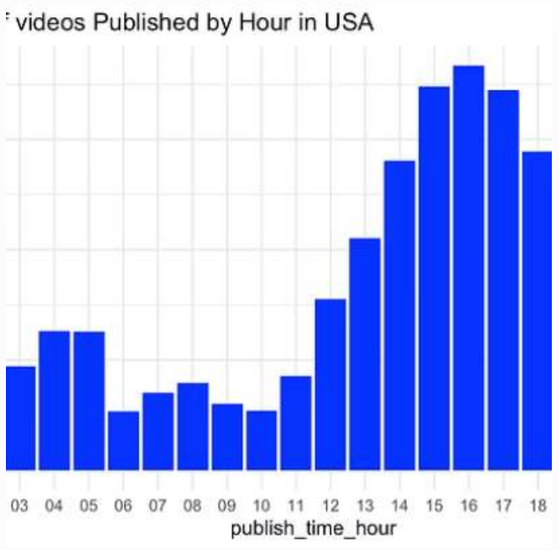


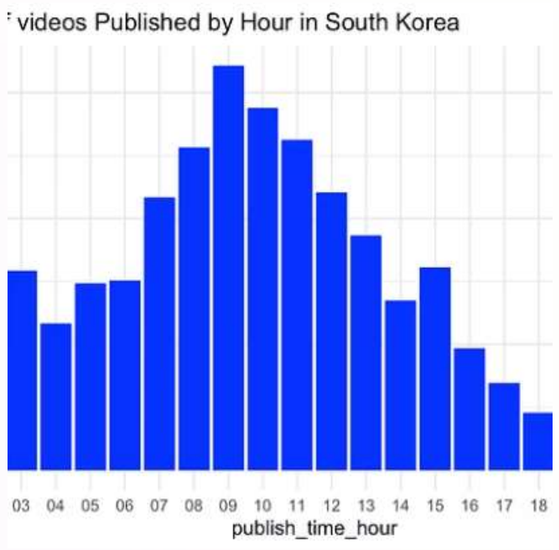


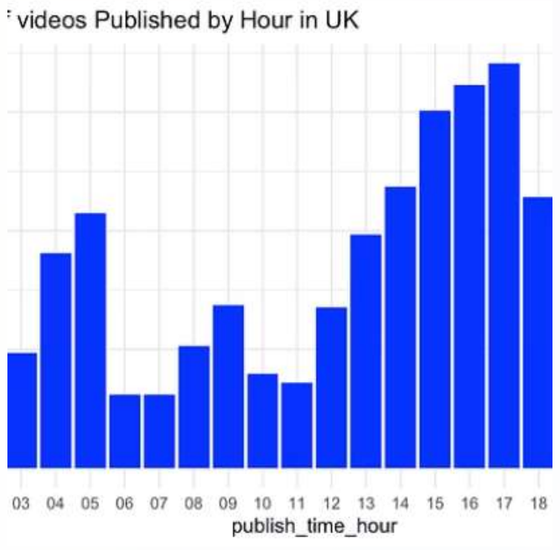


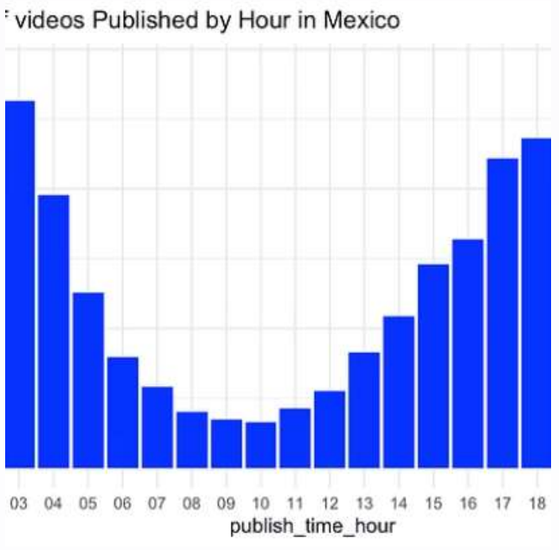

videos Published by Hour in Japan
03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18
publish_time_hour


videos Published by Hour in Russia
03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18
publish_time_hour

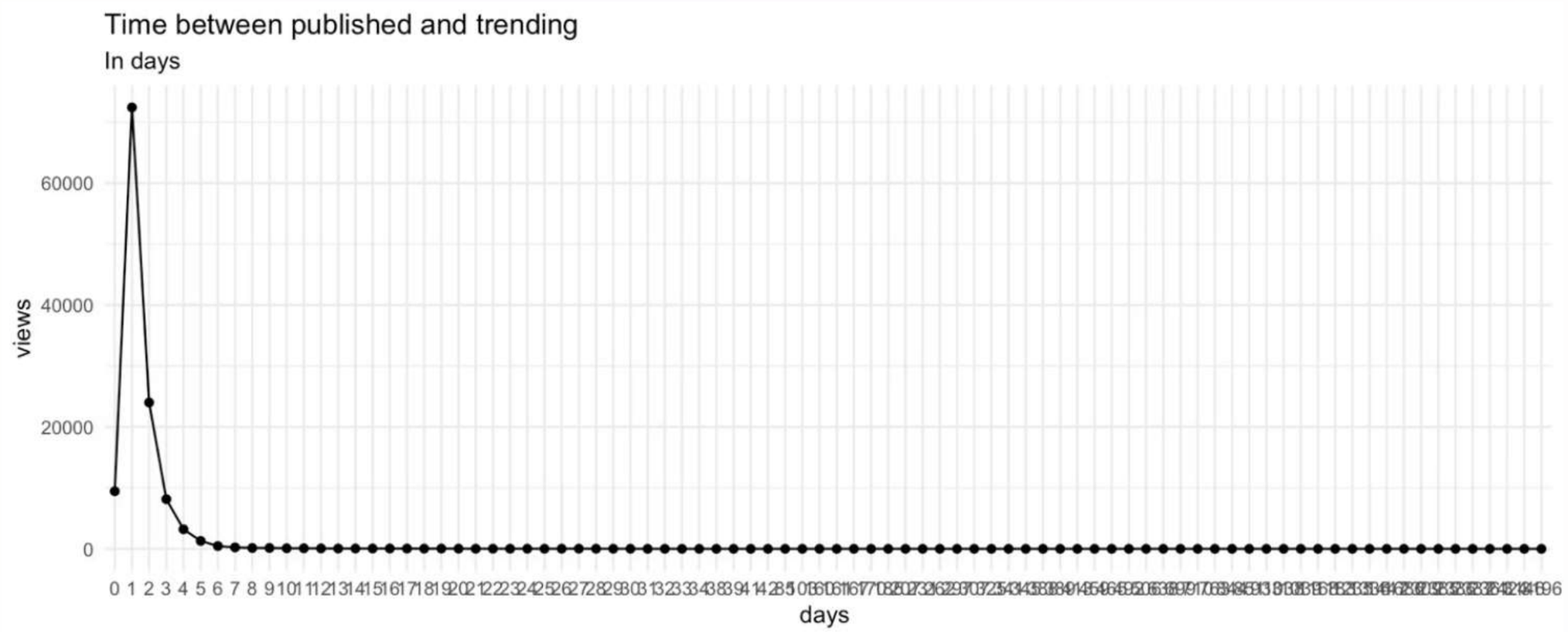

videos Published by Hour in USA
03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18
publish_time_hour


videos Published by Hour in South Korea
03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18
publish_time_hour

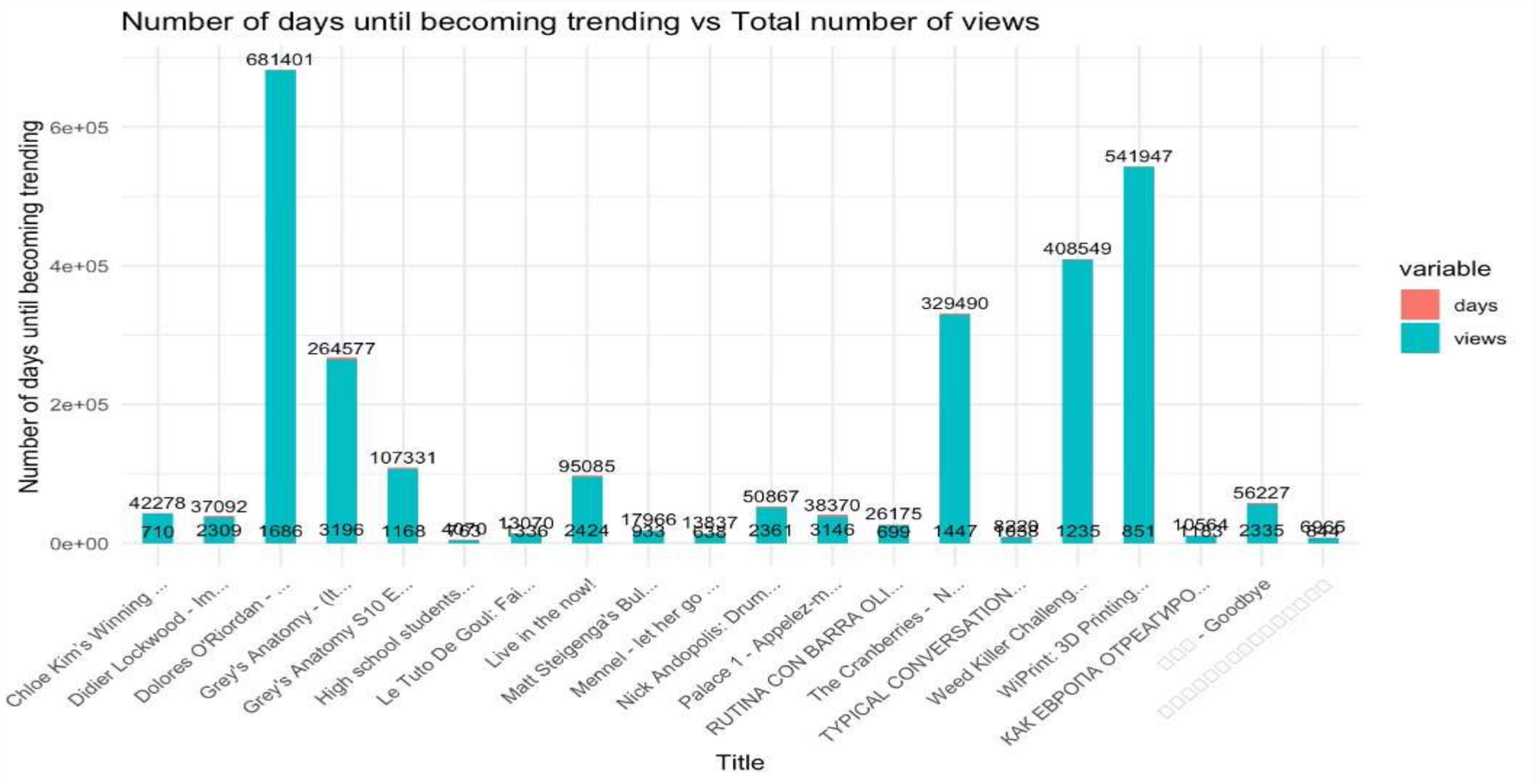

videos Published by Hour in UK
03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18
publish_time_hour


videos Published by Hour in Mexico
03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18
publish_time_hour

TIME BETWEEN PUBLISHING AND TRENDING?

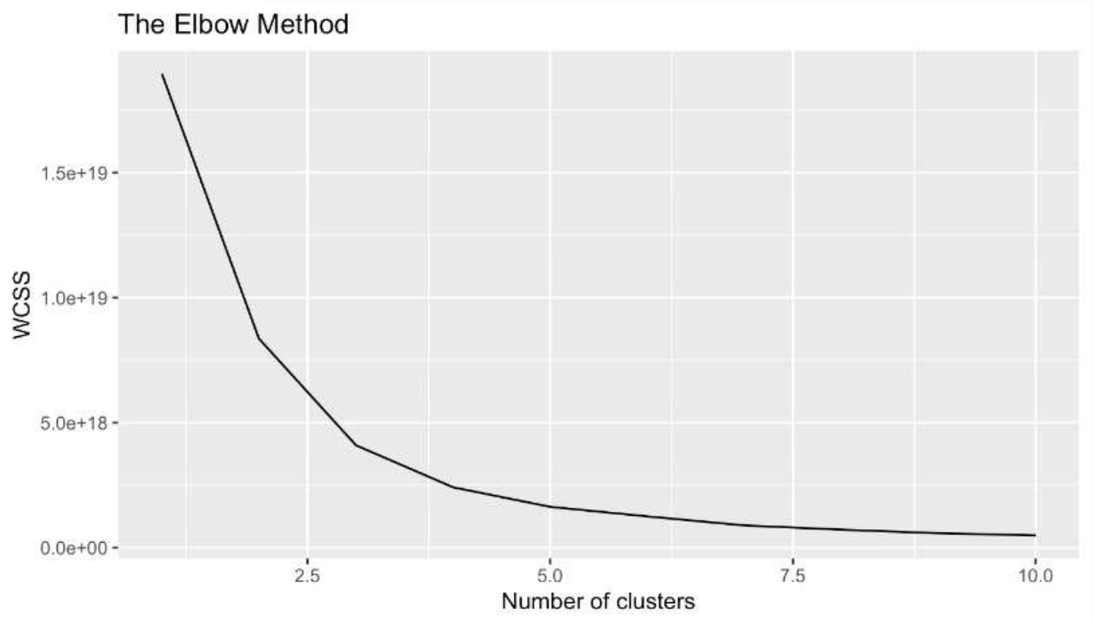


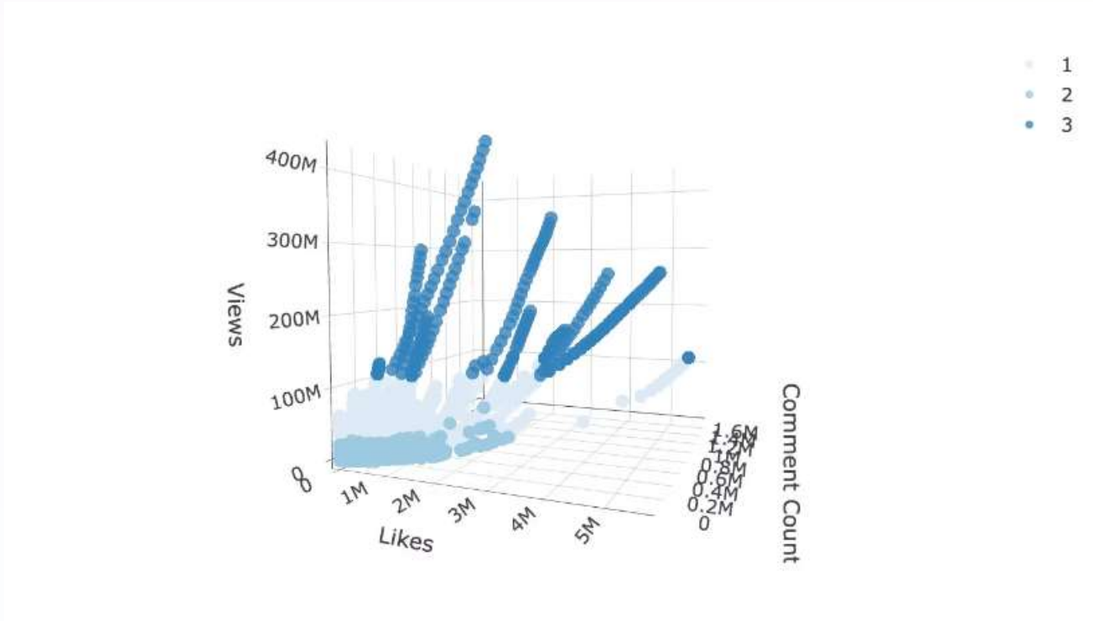
VIDEOS THAT WAITED THE LONGEST BEFORE THEY BECAME TRENDING



Model Training and Evaluation

k-means clustering





Cluster Analysis

Cluster 1 (Size: 2946):

- The mean values for `data_total.likes`, `data_total.dislikes`, `data_total.views`, and `data_total.comment_count` are notably high.
- This indicates that Cluster 1 likely corresponds to highly popular and engaging videos, characterized by substantial numbers of likes, views, dislikes, and comments.

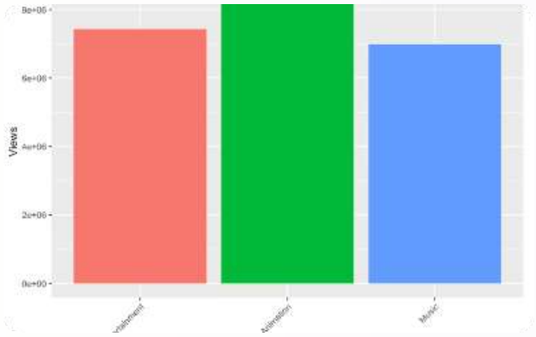
Cluster 2 (Size: 372768):

- This cluster is the largest among the three identified clusters.
- The mean values for `data_total.likes`, `data_total.dislikes`, `data_total.views`, and `data_total.comment_count` are significantly lower than those of Cluster 1 but higher than those of Cluster 3.
- This suggests that the cluster likely represents moderately popular and engaging videos, characterized by a moderate number of likes, views, dislikes, and comments.

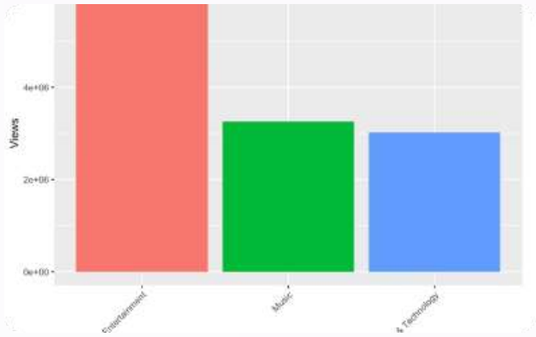
Cluster 3 (Size: 228):

- This cluster is the smallest among the three identified clusters.
- It exhibits the lowest mean values for `data_total.likes`, `data_total.dislikes`, `data_total.views`, and `data_total.comment_count` compared to the other clusters.
- This suggests that the cluster likely represents videos or content that are less popular and less engaging, characterized by comparatively fewer likes, views, dislikes, and comments.

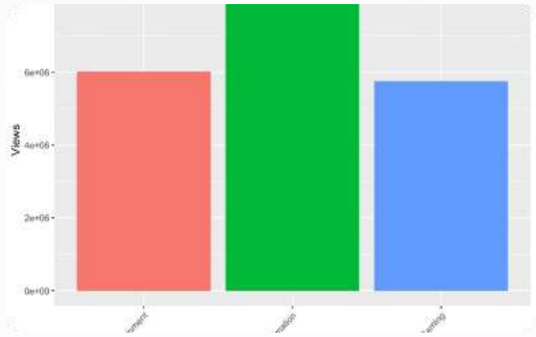
Analysis on scraped data



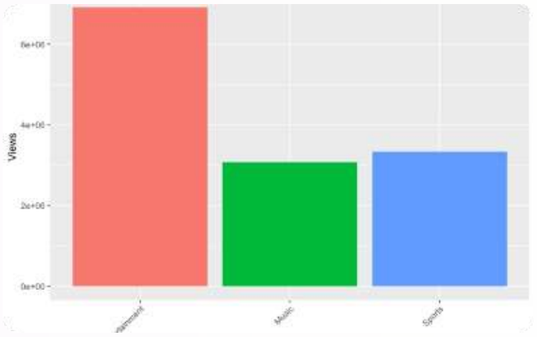
Top 3 Categories in Canada



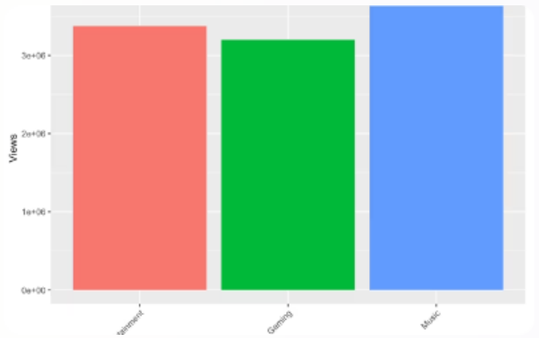
Top 3 Categories in France

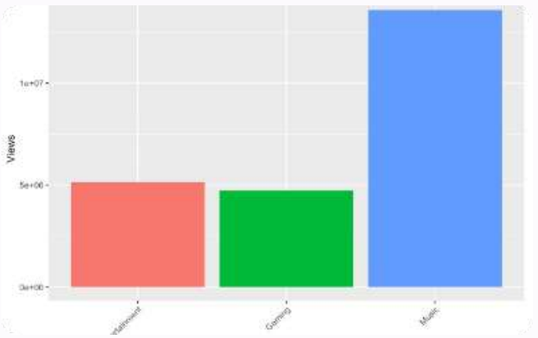


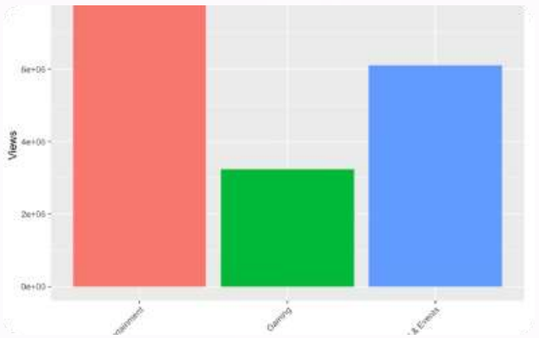
Top 3 Categories in Germany

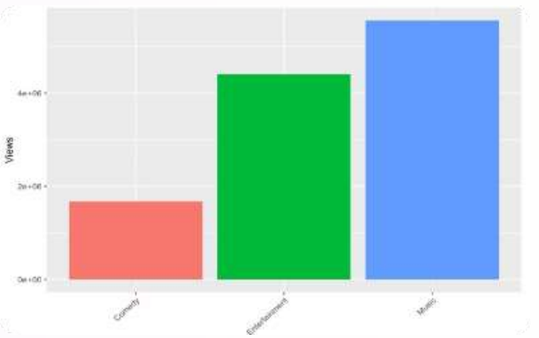


Top 3 Categories in Mexico

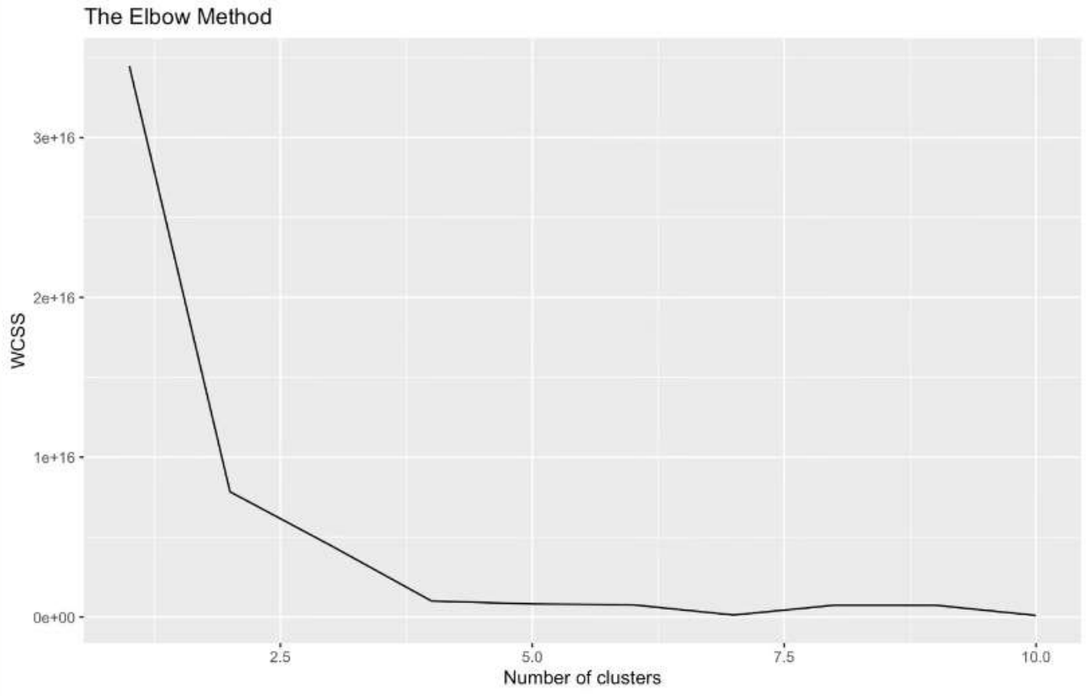

Views
3e+06
2e+06
1e+06
0e+00
Gaming
Music

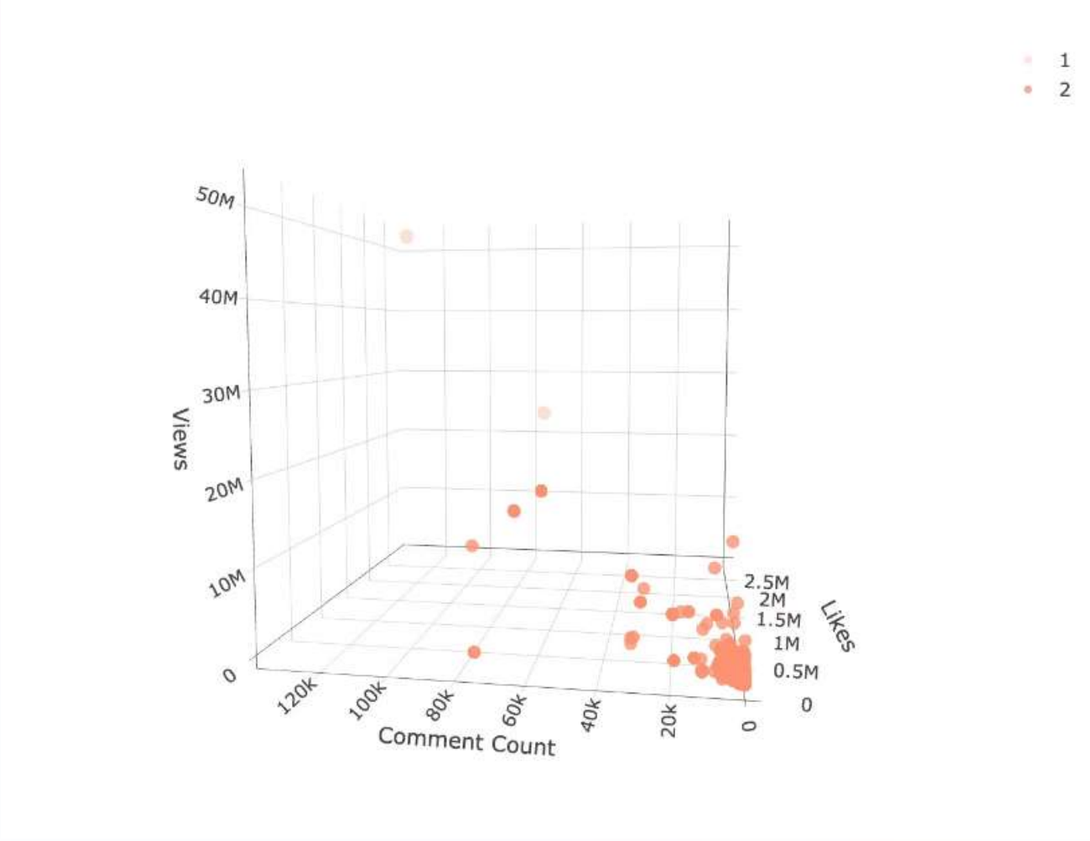

Views
1e+07
5e+06
0e+00
Gaming
Music


Views
6e+06
4e+06
2e+06
0e+00
Gaming


Views
4e+06
2e+06
0e+00
Comedy
Entertainment
Music

k-means clustering on scraped data





Conclusions

Gained insights into the optimal choice of video category, the ideal title length, and the potential influence of description sentiment on viewership.

Compared the findings with current trends to obtain a comprehensive view of the data. Overall, the trends observed in the current data closely align with those identified in past data.

An attempt was made to perform cluster analysis to uncover more nuanced insights; however, the results were not as successful as anticipated.

Future work involves a deeper exploration of the identified clusters and the development of methods to predict video trends.

Thank You!