CSP 571 – GROUP PROJECT REPORT

Exploratory Data Analysis on Trending YouTube Videos

Group Members:
- Venugopal Katragadda (A20544776) – Team Lead
- Sannihitha Gudimalla (A20560248)
- Sruthi Kondapalli (A20554780)
- Bezawada Sai Sravanya (A20561552)

TABLE OF CONTENTS

ABSTRACT

This study aims to understand the dynamics of popular YouTube videos and to identify key aspects of videos that lead to their popularity. In this study, we looked at YouTube trending videos to gather information on the factors associated with video popularity. Prior to our analyses, we preprocessed the data to manage missing values, re-code categorical variables, and to also reformat and recode the date and time columns. The exploratory data analysis involved observing publishing schedules, category trends, and engagement measures such as views, comments, and likes. We conducted sentiment analysis of video descriptions to measure the contribution of keywords in titles to trending. We also examined the state of the video publish dates and the time interval from publish to trending and established the top categories by views for each country. To provide a comparative analysis with the trended data, the research also includes cluster analysis.

INTRODUCTION

YouTube has changed the landscape of video content production, distribution, and consumption, with an expansive array of videos that include news, entertainment, education, and leisure. YouTube has become an important platform for influencers, brands, individuals, and content producers to engage with audiences worldwide. This research intends to understand the dynamics of viral YouTube videos and the important factors that lead, in part, to their success.

The goal of this research is to explore the patterns affecting the success of videos through exploratory data analysis (EDA) of popular videos on YouTube to explore trends in genres, timing, audience engagement including likes, views, comments, etc. With this project it is our intention to help content creators and marketers develop strategies that allow them to reach more viewers on YouTube while exploring trends in timing of posts, genres, and audience engagement. In addition, this study will explore the relationship between video length, title lengths, and tags and how these metrics along with YouTube's ranking algorithm affect performance. This type of thorough research would provide valuable information for marketers and content creators who would like to improve their presence on YouTube.

DATA PREPROCESSING

DATA SOURCES

We have two datasets for this project. The first dataset is historic YouTube video data from about 5 years ago. We got it from Kaggle. The second dataset is more recent and was scraped while using the YouTube API. We expect to perform a comprehensive analysis that will incorporate both historical and current patterns on the YouTube platform. By utilizing both datasets we will capture historic trends of video and metric data along with current trends.

1. 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)

2. SCRAPED DATASET USING YOUTUBE API

By using the YouTube Data API to get the dataset we scraped, we had the ability to collect current data directly from YouTube. As we examined the sample outputs of the scraped dataset and matched

the incompatible column in the Kaggle dataset we used, we defined consistency across the two datasets. In indicating use and consistency for the Kaggle dataset, we ensured the Kaggle dataset indicated similar data of the countries we used.

CONVERT CATEGORICAL VARIABLE CATEGORY_ID AND ADDING COUNTRY VARIABLE

The dataset's `category_id` column contains numerical IDs as opposed to descriptive text labels that indicate which category each video is in. It will be easier to understand and analyze the data according to the actual content categories by mapping these IDs to the category names.

Category Types and respective ID:
- 1: Film & Animation
- 2: Autos & Vehicle
- 10: Music
- 15: Pets & Animals
- 17: Sports
- 19: Travel & Events
- 20: Gaming
- 22: People & Blogs
- 23: Comedy
- 24: Entertainment
- 25: News & Politics
- 26: How to & Style
- 27: Education
- 28: Science & Technology
- 30: Movies
- 43: Shows
- 29: Nonprofits & Activism

HANDLE MISSING VALUES AND DUPLICATES

The missing value of the combined dataset is considered, and the below result has appeared. The missing values appear only in the column description.

```
> print(missing_sum)
              video_id          trending_date                  title          channel_title
                     0                     0                      0                      0
           category_id           publish_time                   tags                  views
                     0                     0                      0                      0
                 likes               dislikes          comment_count         thumbnail_link
                     0                     0                      0                      0
      comments_disabled      ratings_disabled video_error_or_removed            description
                     0                     0                      0                  19494
```

The sentiment analysis in the "description" column heavily deals with missing data in this column. In our first step with the data, we also removed all data records that contained missing values and carried out sentiment analysis since the analysis was based on the sentiment of the description text itself. The analysis yielded the description with the highest percentage of sentiment being neutral. Thus, the answers would not be much different if we simply replaced the missing values with a neutral value.

Having observed that a few films become the subject of continuous trending, we decided to build a dataset that eliminates these kinds of recurrences. But given that the original dataset contained vital information that we don't want to eliminate, we decided to keep it.

TRANSFORMATION OF CONVERT THE 'TRENDING_DATE' AND 'PUBLISHED_TIME' COLUMNS TO DATETIME FORMAT

The 'trending_date' and 'published_time' columns in our dataset were switched from cellular formatting to datetime formatting and this was an important change. We were capable of examining accurate trends and patterns in the video's trending and published activities in time because of this change. We increased the accuracy of our analyses and ease of flow in our data pipeline by obtaining the correct YMD (Year-Month-Day) format.

EXTRACT FROM 'PUBLISHED_TIME' SUCH AS THE DAY OF THE WEEK OR HOUR OF THE DAY THE VIDEO WAS PUBLISHED.

We extracted more data from the 'published_time' column to better quantify the patterns of publishing for YouTube videos. We were able to pull important information like the day of publishing and the time of the publishing to process the pattern within this data. We were able to identify behaviors from this extraction method like days and times to publish. This could be beneficial for content creators who want to calibrate when to publish their contents to increase engagement with their predicated audience.

TEXT PRE-PROCESSING FOR SENTIMENT ANALYSIS

We meticulously pre-processed our data in advance of the sentiment analysis. We created a corpus of titles from YouTube videos in the relevant languages, and cleaned the text data using several methods. We needed to remove special characters, punctuation, numbers, and common English stop words that did not contribute to the sentiment analysis than it would have influence the sentiment analysis. We even removed stop words from other languages in countries like France, Japan, and others.

We also removed the words "video" and "audio," which were abundant in our dataset and did not carry any sentimental value. In the final step to maintain uniformity in the entire corpus, we converted the text to lower case and eliminated additional white space. All these preprocessing

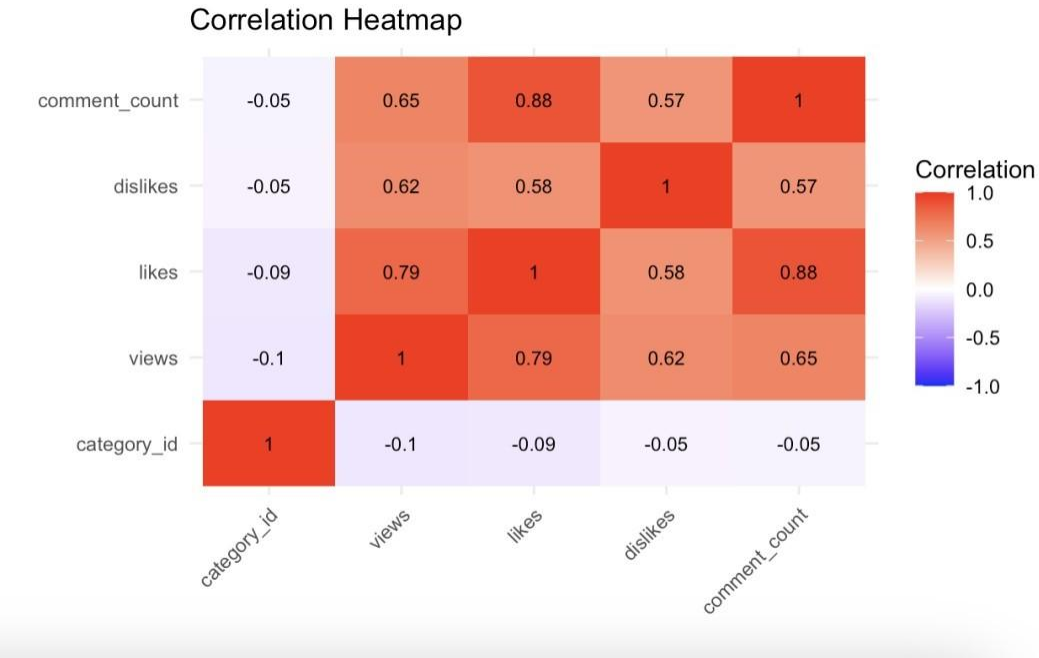
steps allowed us to diminish error and improve accuracy and continuum for only the significant portion of the video titles for sentiment analysis.

EXPLORATORY DATA ANALYSIS

To gain a thorough understanding of the overall trends and patterns, we will first do an analysis of the complete dataset. From this initial analysis, we will learn a lot about the dataset. After this, we will do a region-specific analysis, where we will look deeply within the specific region to identify region-specific trends.

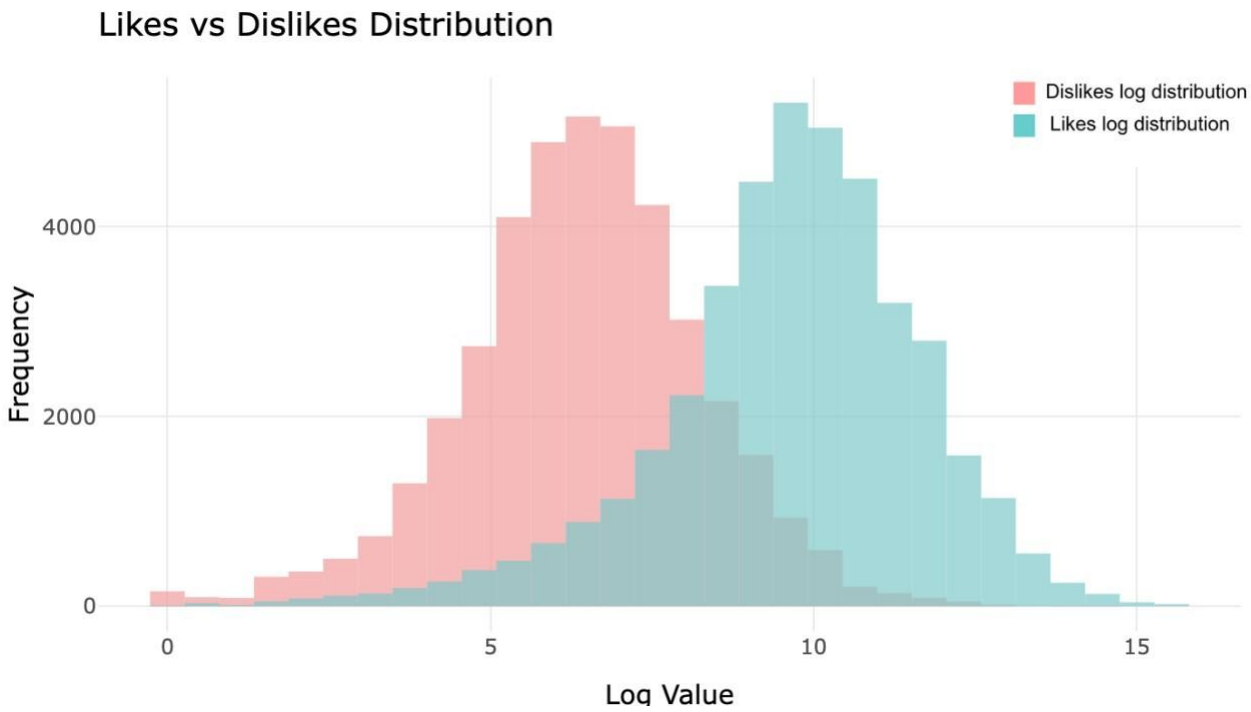
CORRELATION MATRIX

Let us first understand the correlation of comment_count, likes, dislikes views.



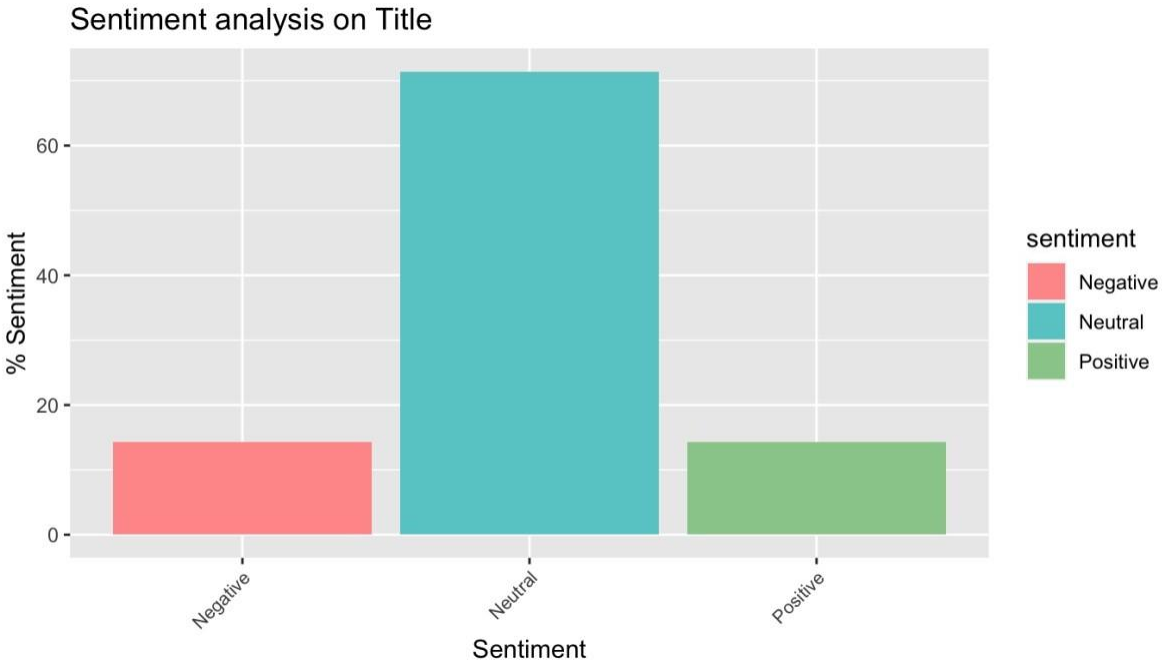
As expected, there is a strong association between views and likes. However, there is about half as much association between dislikes and views as there is for likes and views.

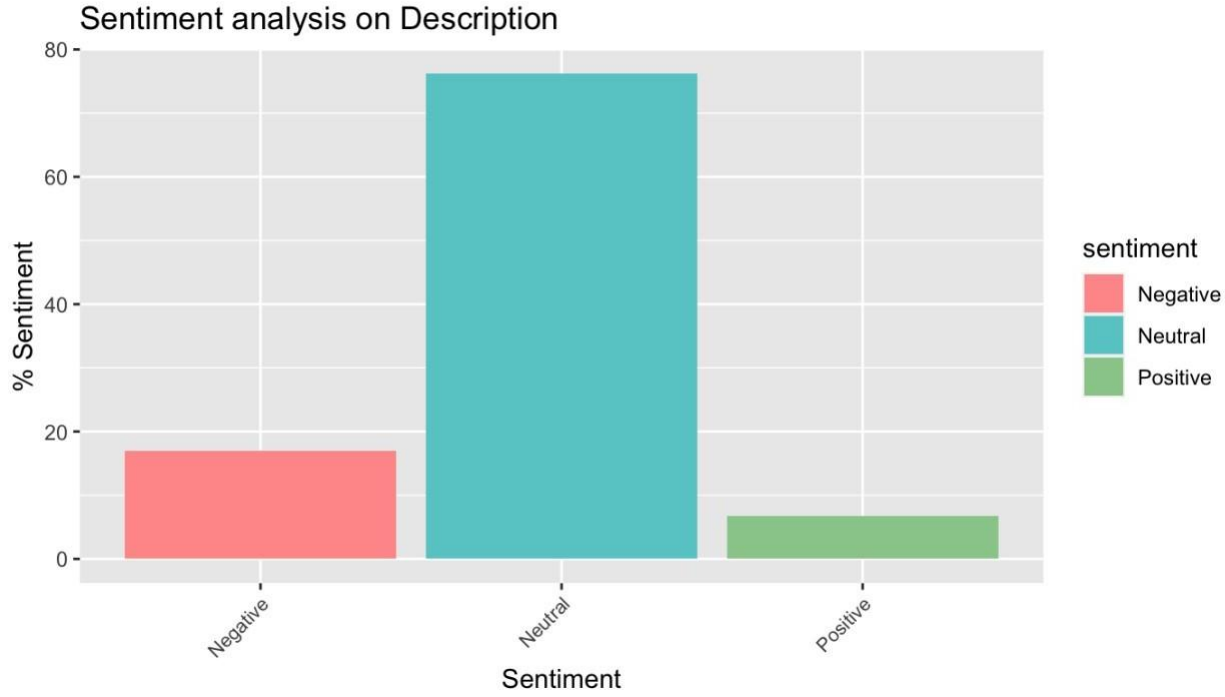
DISTRIBUTION OF THE VALUES



The distribution of likes and dislikes was normalized by taking the logarithm. After that transformation, the distributions of likes and dislikes in your dataset appear in like magnitudes. The overlap suggests that likes and dislikes appear in similar ranges, which may indicate more videos with a certain popularity (with respect to likes and dislikes) are found in the dataset.

SENTIMENT ANALYSIS

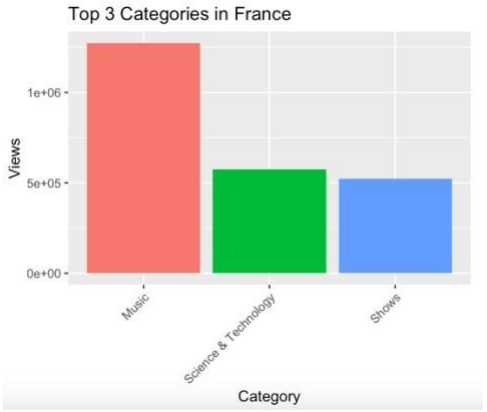


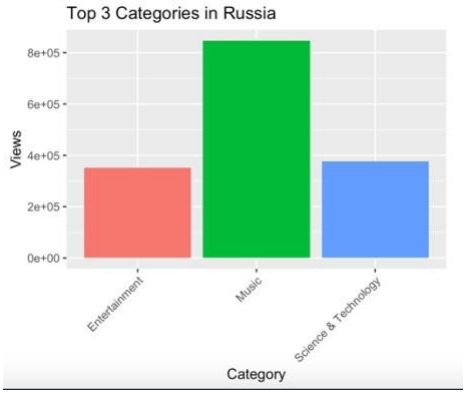


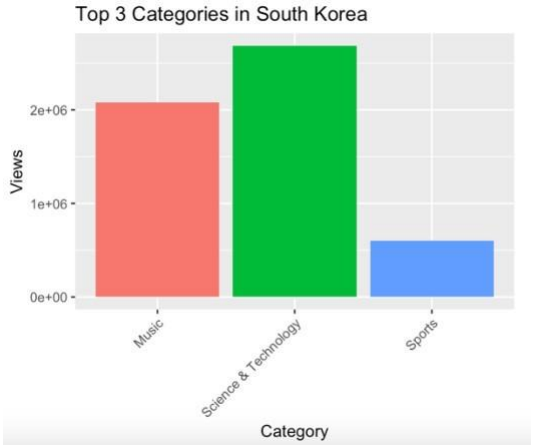
From the plots, we can observe that the sentiments in the title and description field are basically neutral.

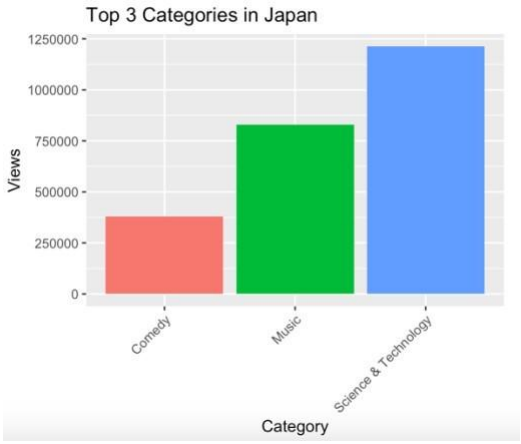
Let us try to answer some questions that can make it easier for us to understand the data.

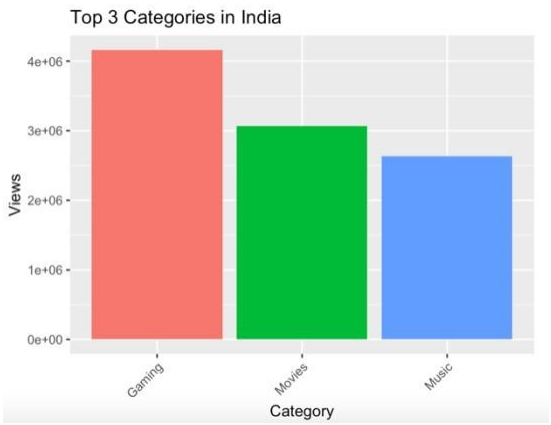
WHAT ARE THE TOP 3 CATEGORIES BASED ON THE NUMBER OF VIEWS, FOR EACH OF THE COUNTRIES?

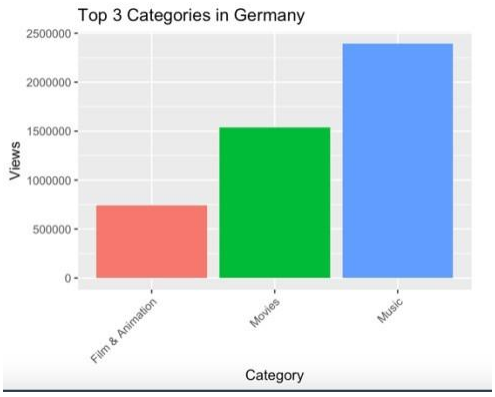


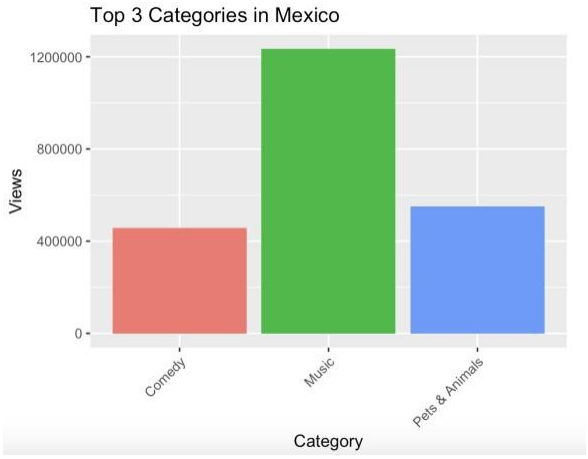


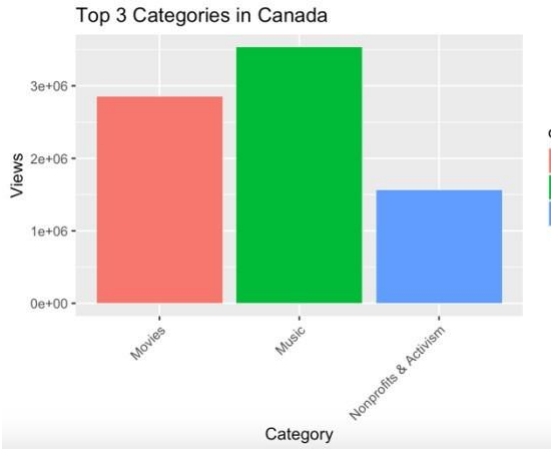


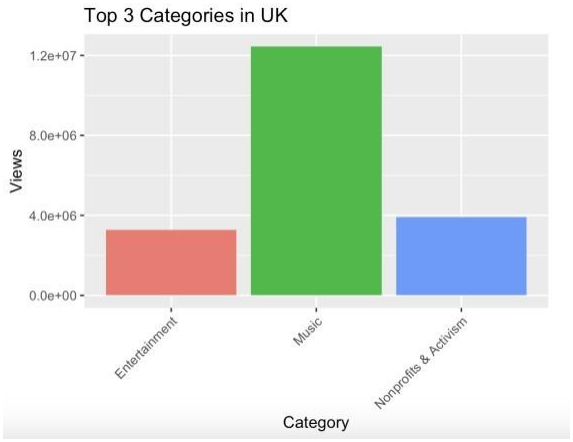


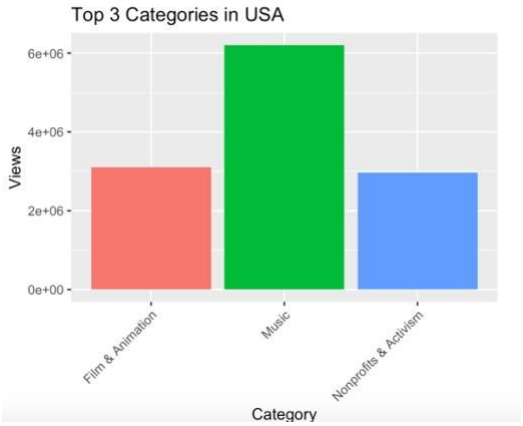


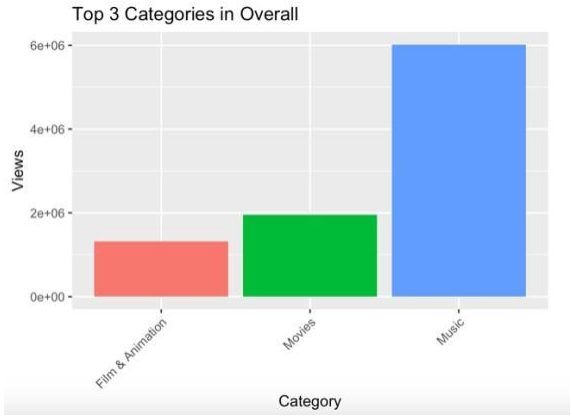












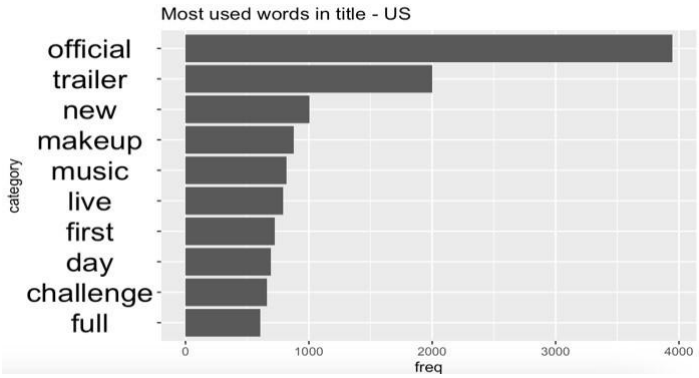
When we analyzed different video categories from different countries, we saw some interesting trends in preferences. With views typically ranging from millions to tens of millions, music is almost universally considered one of the most popular top categories in a whole set of countries including the USA, Canada and the UK. This suggests there is widespread and global interest

YouTube content related to music. Also in other places, categories such as science & technology, entertainment, film & animation had also received a significant number of views.
Importantly, the most popular categories differed by country, evidence of cultural differences, and differing audience preferences. For instance, even though science and technology were popular throughout we found that in South Korea and in Japan there was a greater consumption of this category. While we also found that gaming was the most popular category in India. Overall, this research provides useful information regarding the global consumption of video content in various forms and could help marketers and content producers tailor their approaches to better reach their target audiences.

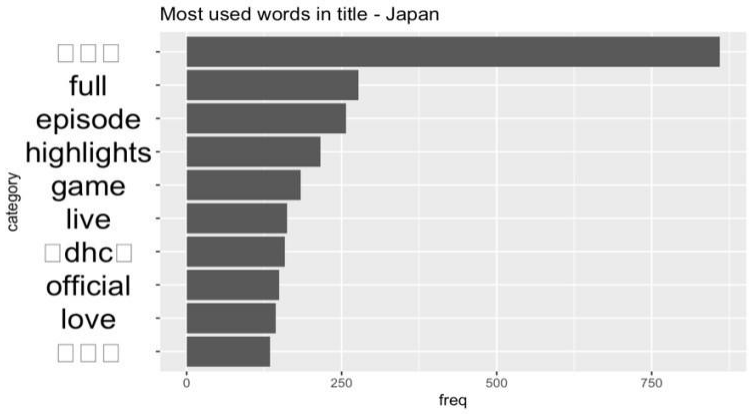
```
> head(CA_most_category,3)
# A tibble: 3 × 3
  category_name         average_views least_views
  <chr>                         <dbl>       <dbl>
1 Music                       3532525.        3201
2 Movies                      2853415       225528
3 Nonprofits & Activism       1562184.        1898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

In addition to the findings above, it is important to note that "Music" and "Movies," which are two categories with the most popular content, have less views here. They should be focused on smart targeting and promotion, while recognizing that the more users are willing to participate in a range of audience interests, the better.

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



	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

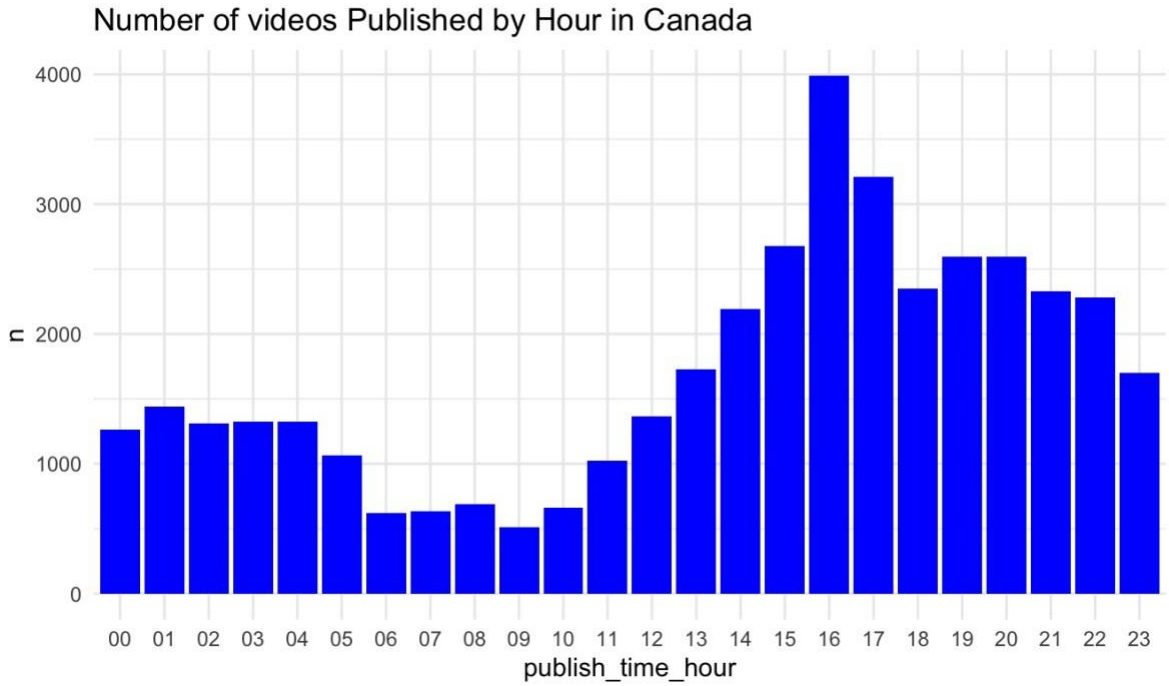


	word	freq
年月日	年月日	860
full	full	277
episode	episode	257
highlights	highlights	216
game	game	184
live	live	162
【dhc】	【dhc】	158
official	official	149
love	love	144
乃木坂	乃木坂	135

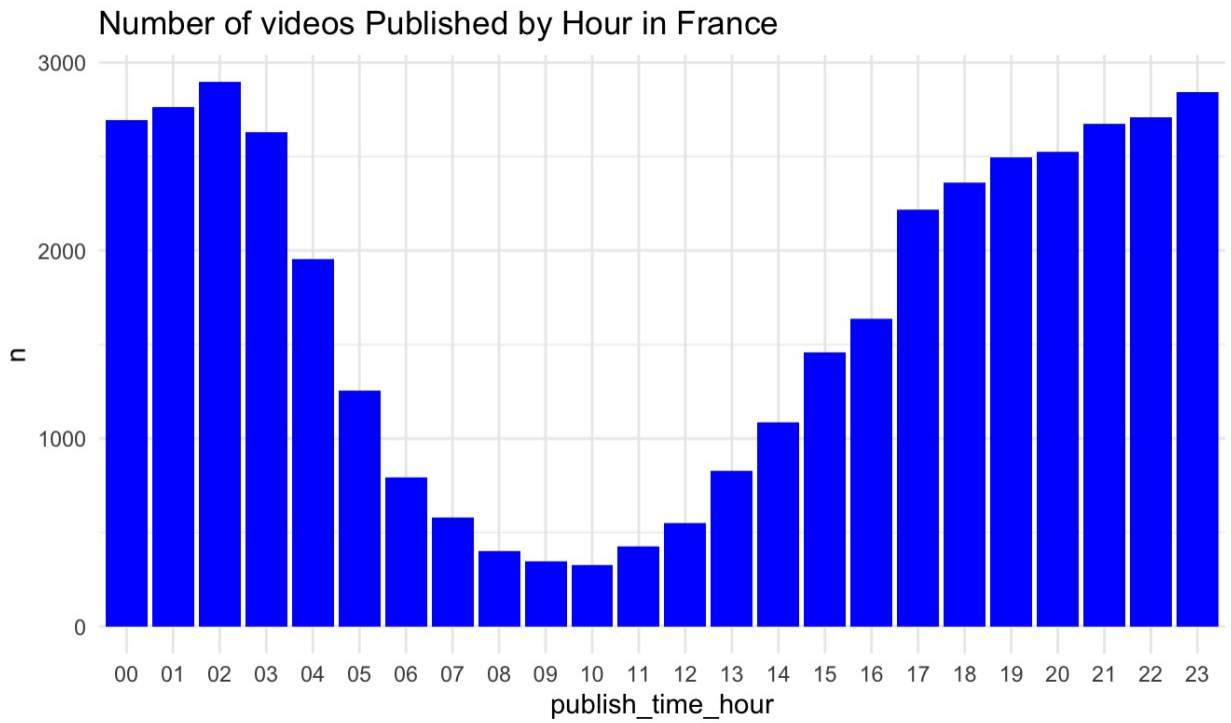
The common terms 'official,' 'trailer,' and 'new' in the US imply a lot of new content and more official releases. While words like 'makeup' and 'challenge' imply that challenge and makeup videos are in some demand. In Japan, words like "年月日" (date) and "乃木坂" (Nogizaka) show cultural differences. But both countries have words like "full" and "live" in common, indicating shared interests in certain types of content.

HOUR OF THE DAY WHEN TRENDING VIDEOS ARE UPLOADED

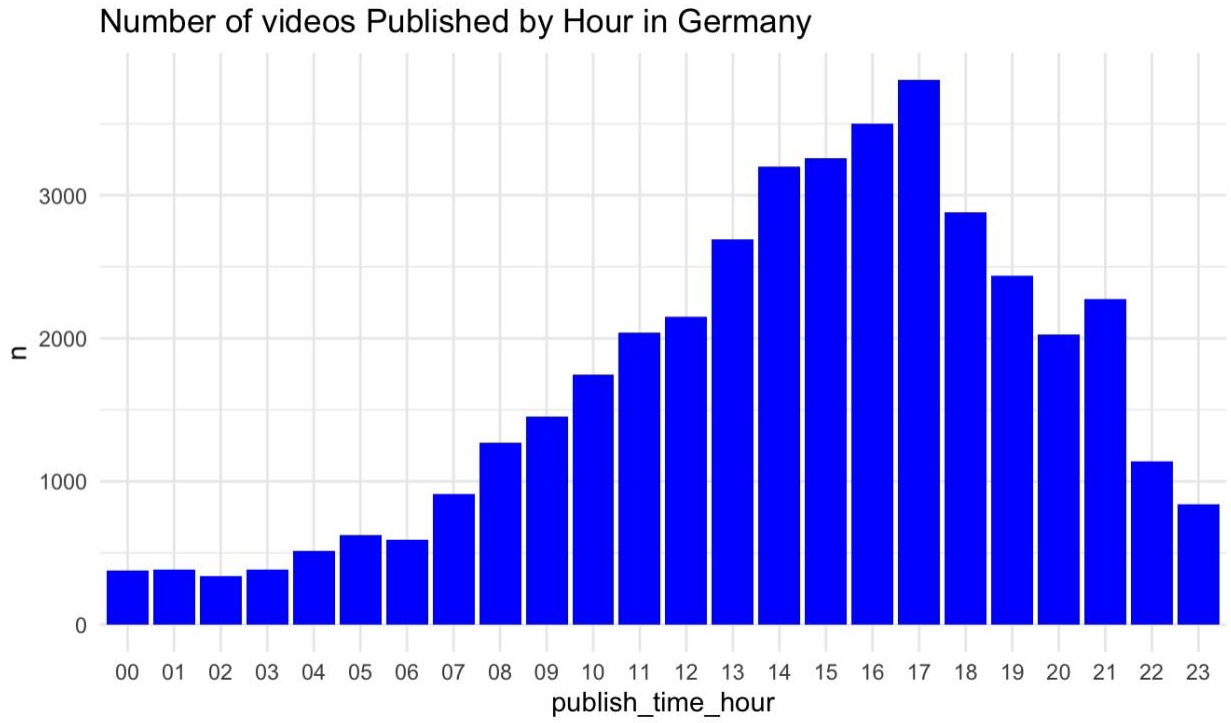
The following plots display the number of videos published in different countries, represented in the UTC time zone.



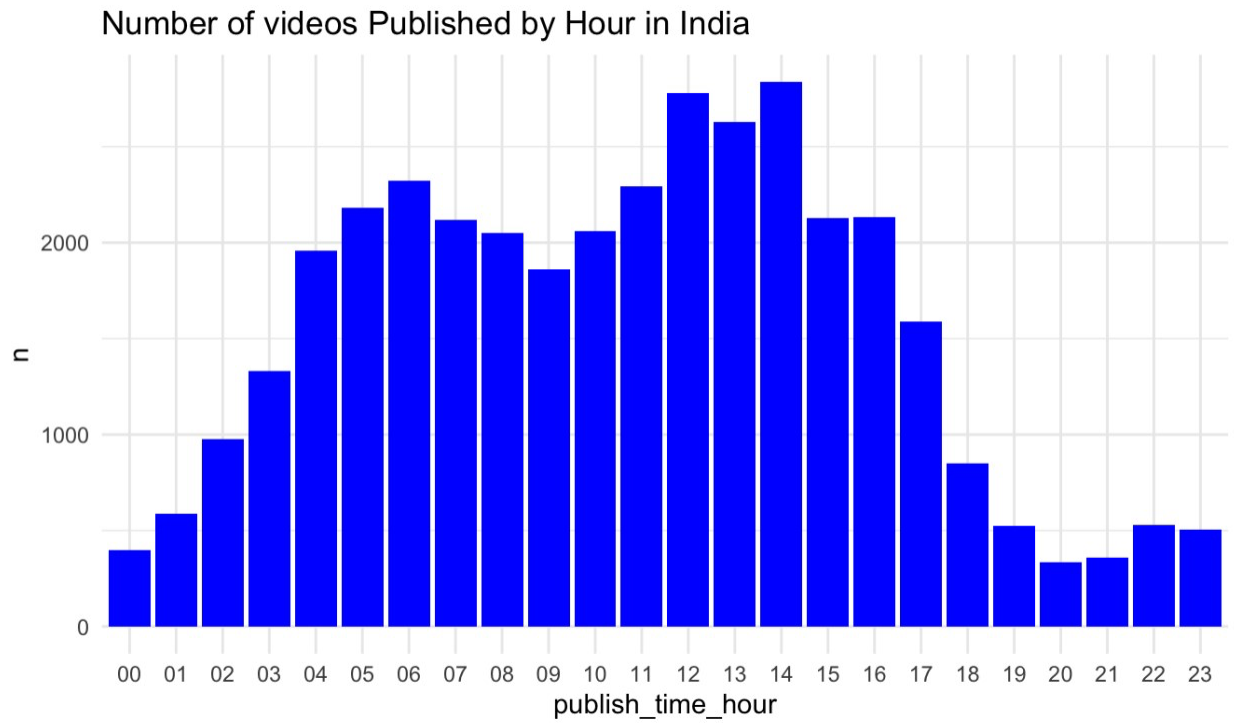
The majority of Canadian videos seem to be released at 12 PM local time, whilst less are released at 4 AM local time.



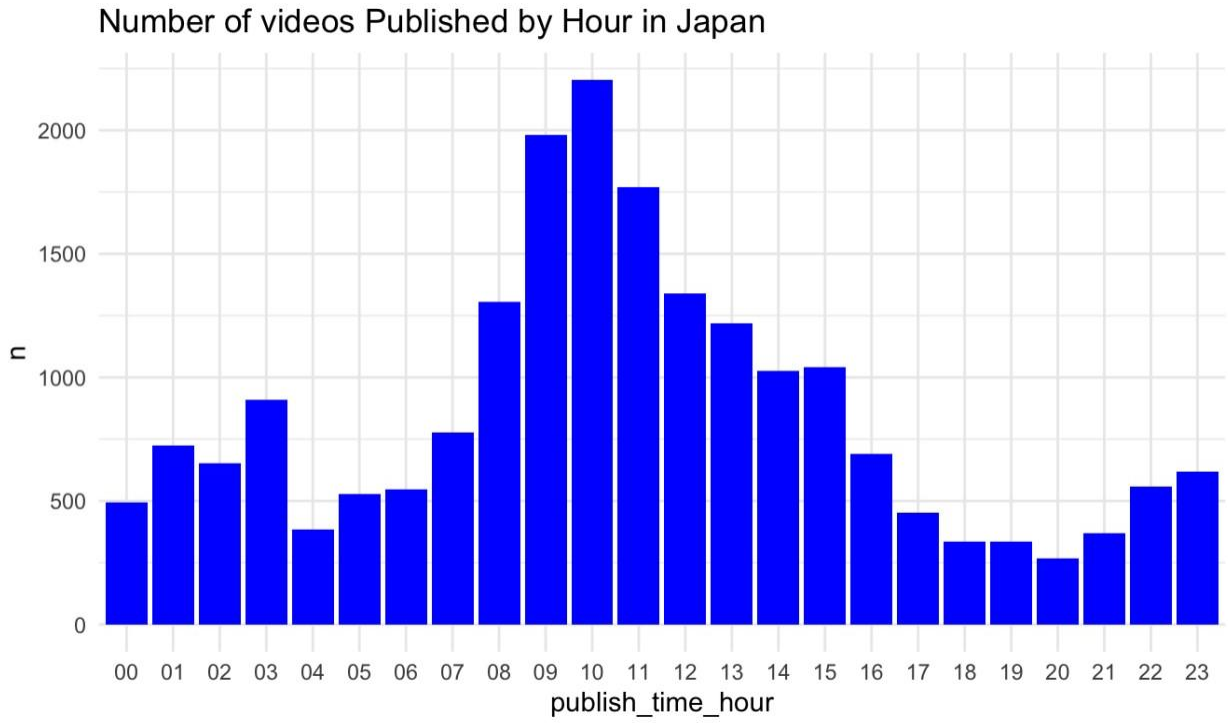
Most French videos seem to be released between 8:00 PM and 2:00 AM local time, whilst less are released between 7:00 AM and 1:00 PM local time.



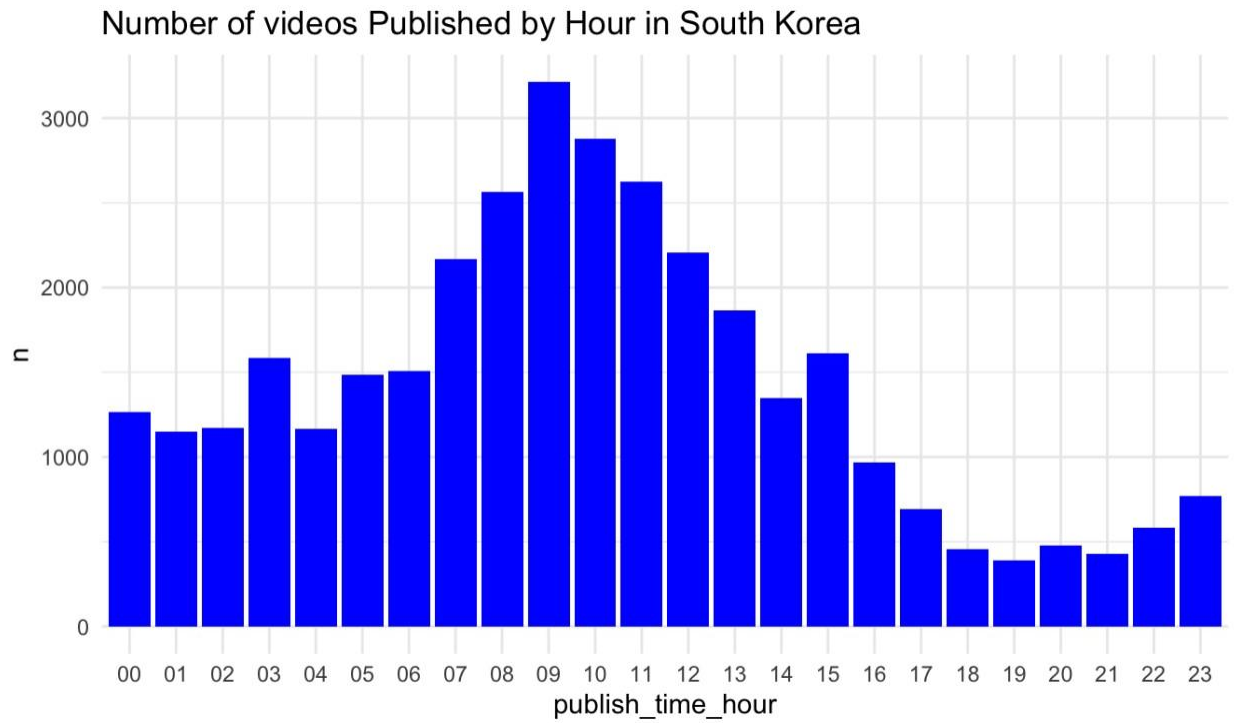
Most German videos seem to be released about 7:00 PM local time, whereas a smaller number are released at 2:00 AM local time.



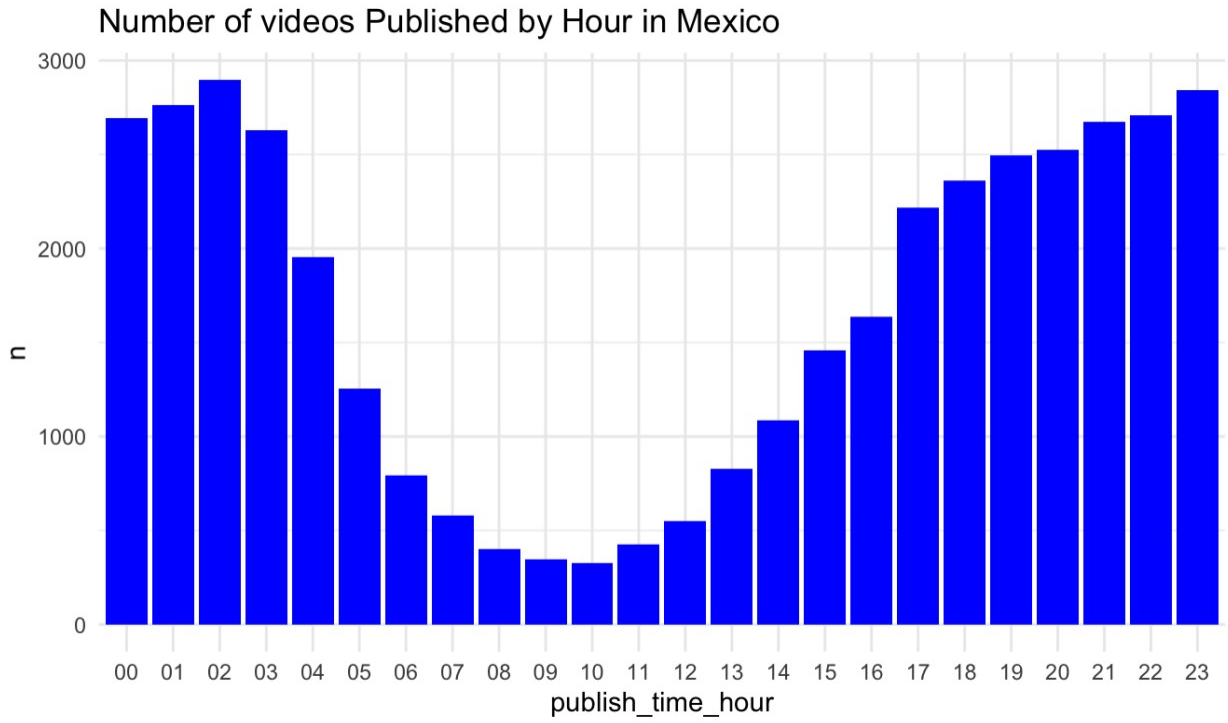
Most Indian videos seem to be released about 7:30 PM local time, whereas a smaller number are released at 1:30 AM local time.



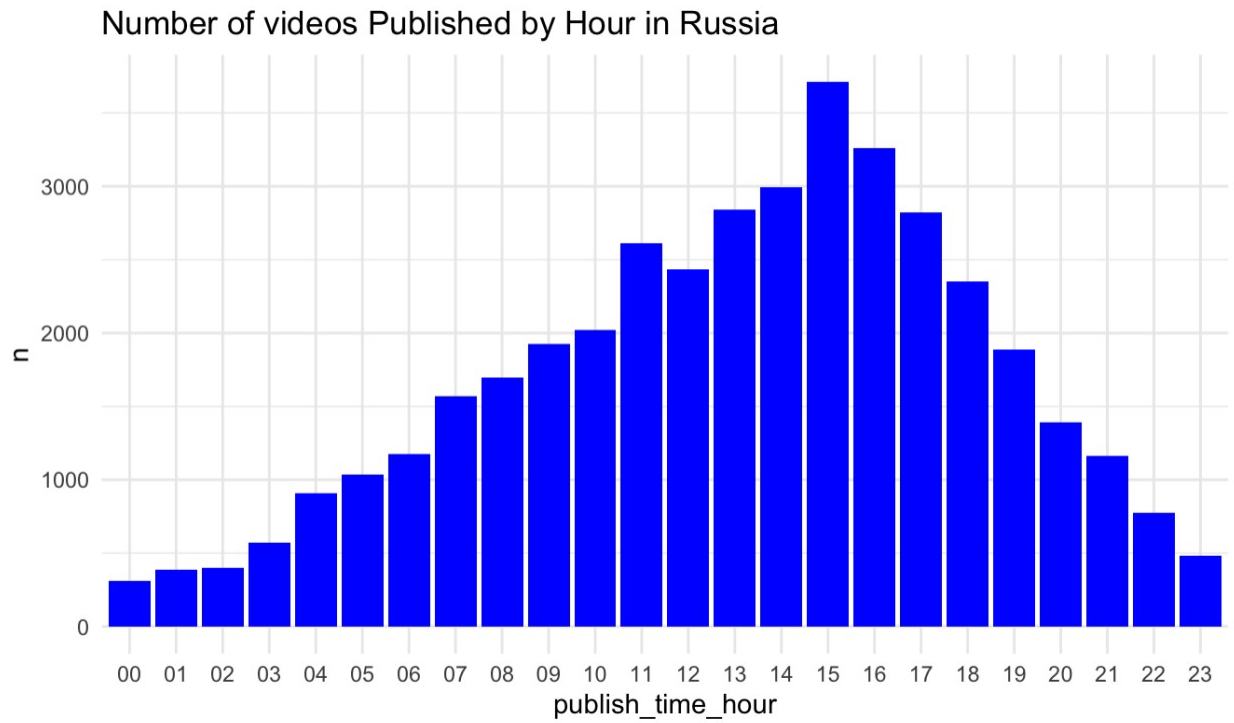
Most Japanese videos seem to be released at 7:00 PM local time, whilst less are released at 3:00 AM local time.



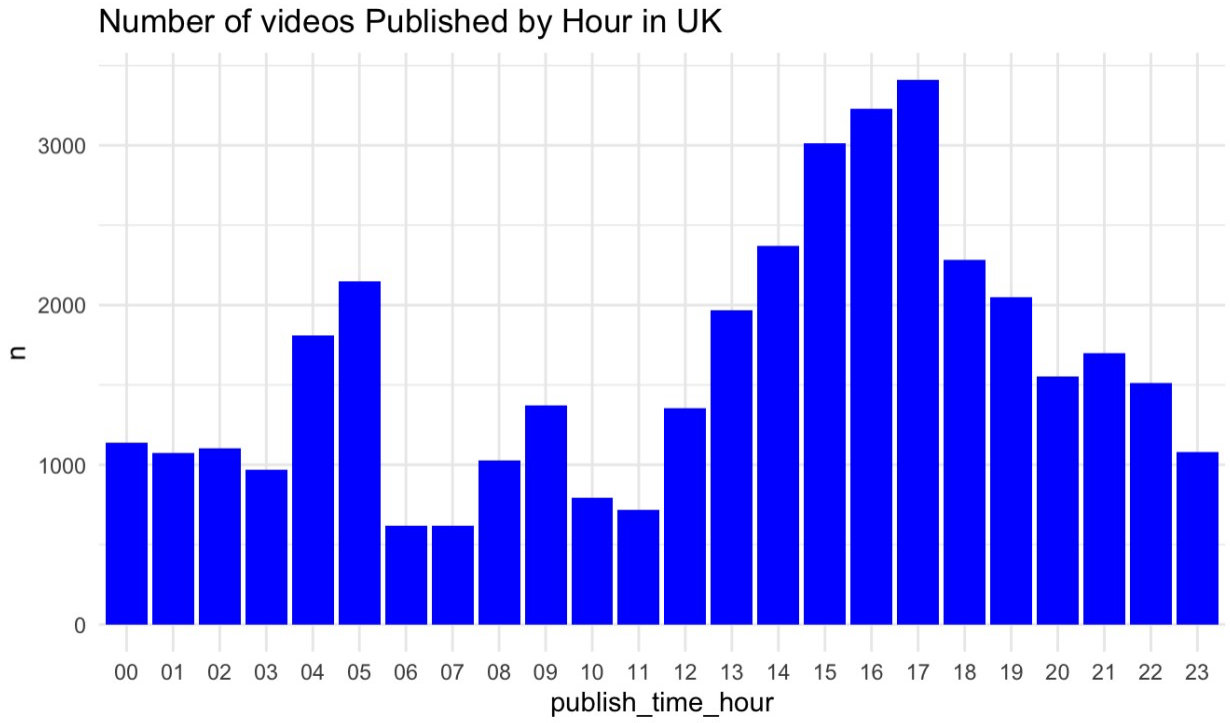
While fewer videos are released around 4:00 AM local time, it seems that most South Korean videos are released around 6:00 PM local time.



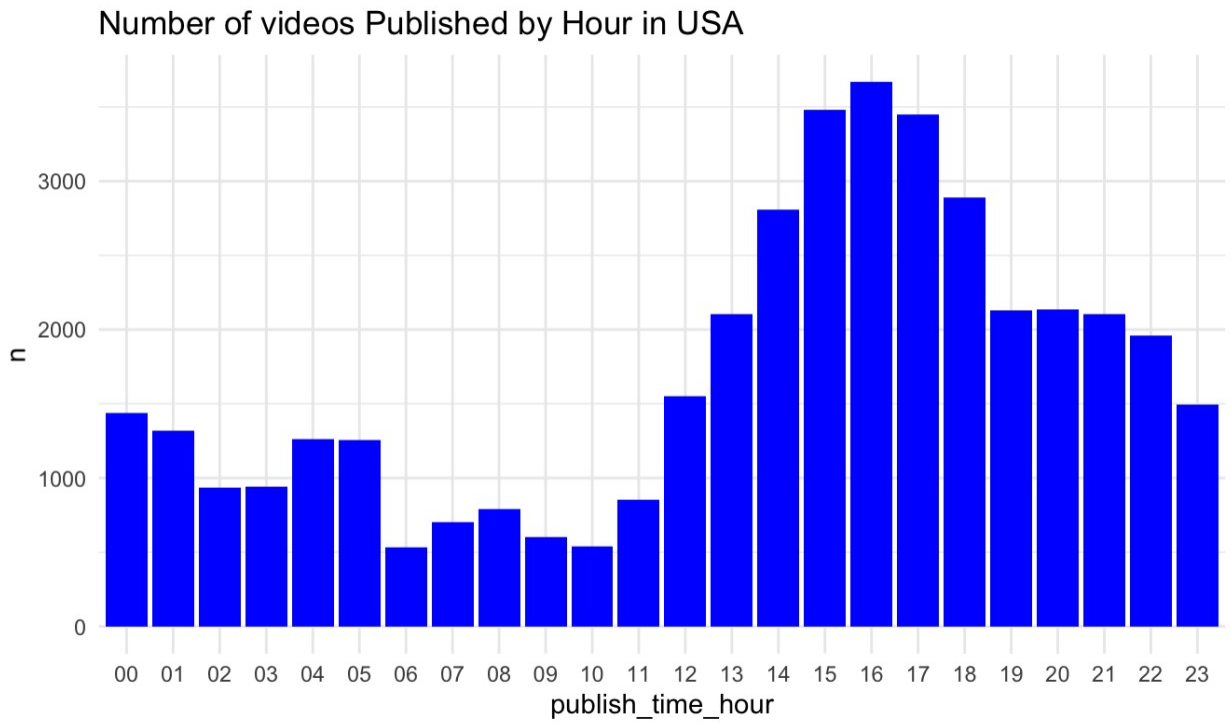
Most Mexican videos seem to be released at 8:00 PM local time, whilst less are released at 3:00 AM local time.



Most Russian videos seem to be released at 6:00 PM local time, whilst less are released at 3:00 AM local time.

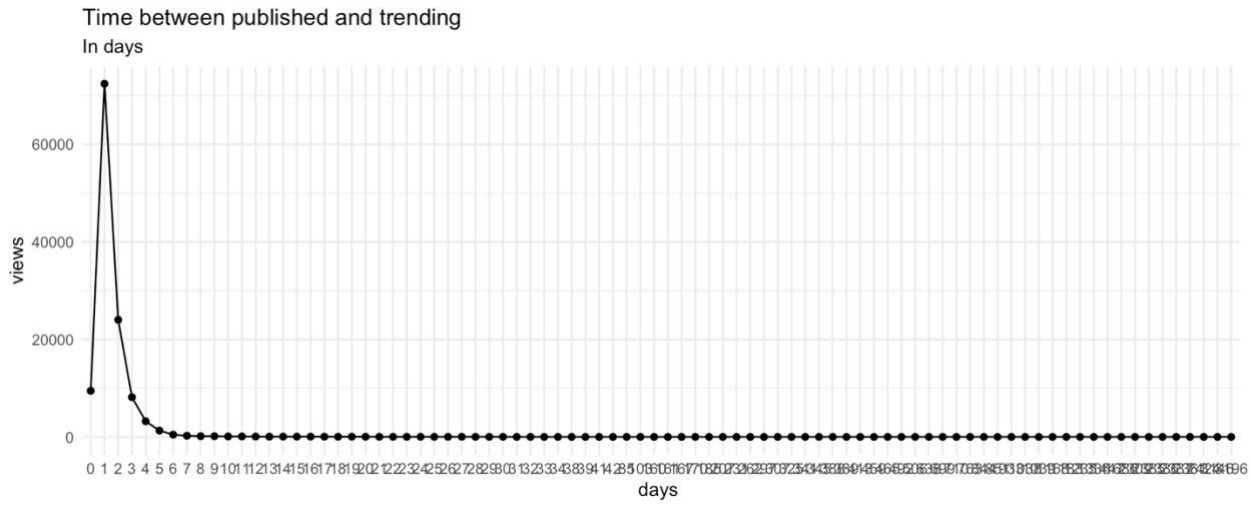


While fewer movies are released between 01:00-07:00 AM local time, it seems that most videos in the UK are released around 6:00 PM local time.



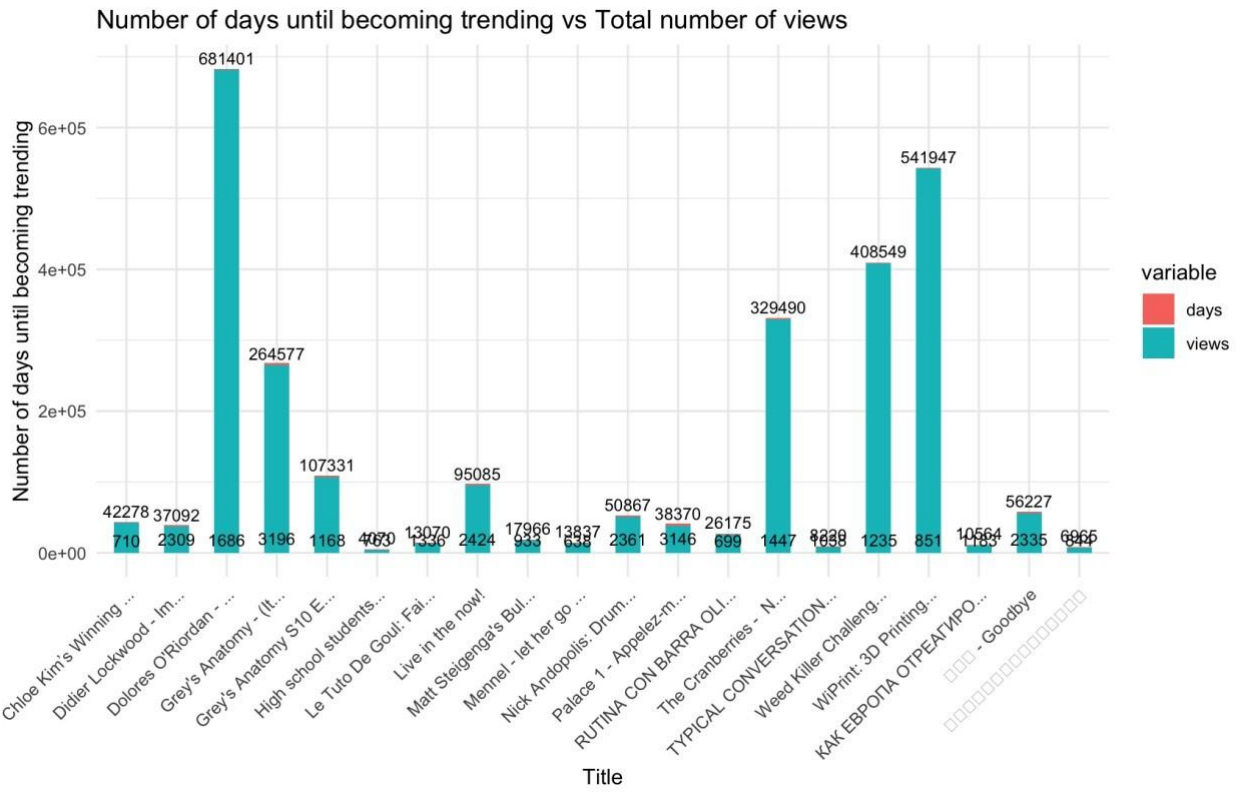
Most French videos seem to be released at 8:00 PM local time, whilst less are released at 4:00 AM local time.

TIME BETWEEN PUBLISHING AND TRENDING



YouTube videos don't seem to go viral the same day they are posted. Rather, if a video is going to become popular, it usually does so within three to four days of its release.

VIDEOS THAT WAITED LONGEST BEFORE THE STARTED TRENDING



This graph represents videos that went viral after waiting for a long time. It means long once on the YouTube platform; several videos did take a long time to go viral. The lag time from a video's posting date to its trending state is widely variable across video types. Even if some of these videos took longer to trend, they eventually achieved views to get to the YouTube homepage.

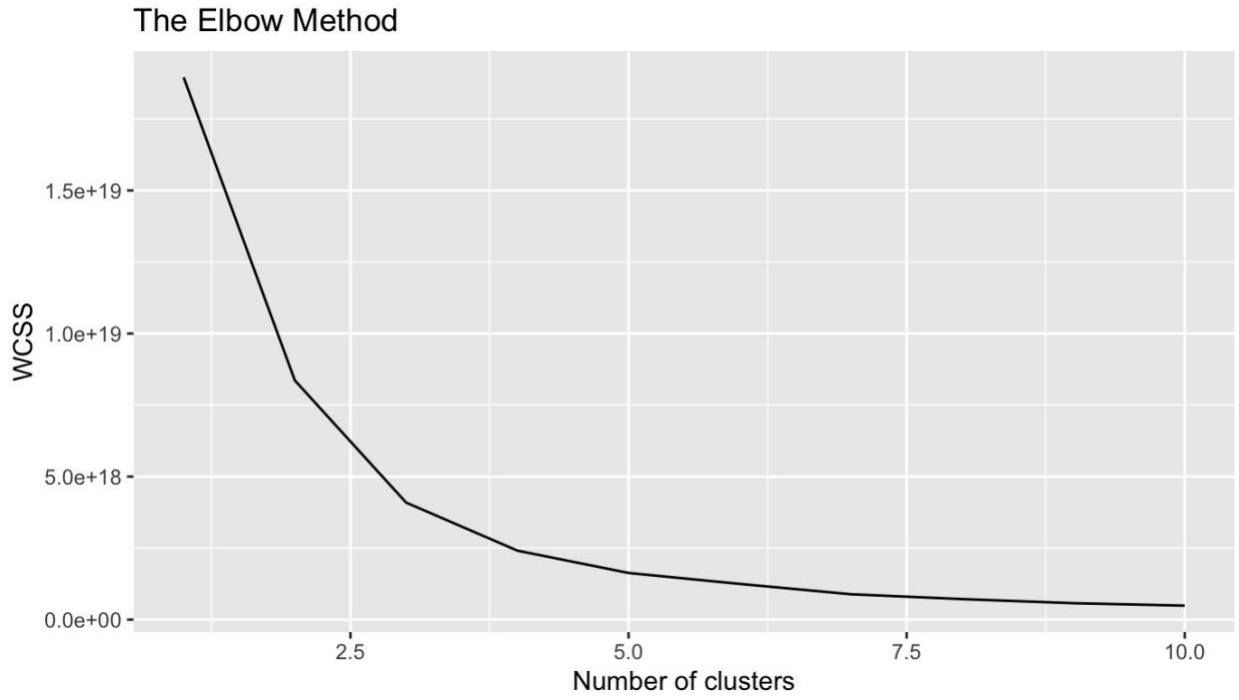
MODEL TRAINING AND VALIDATION

CLUSTER ANALYSIS

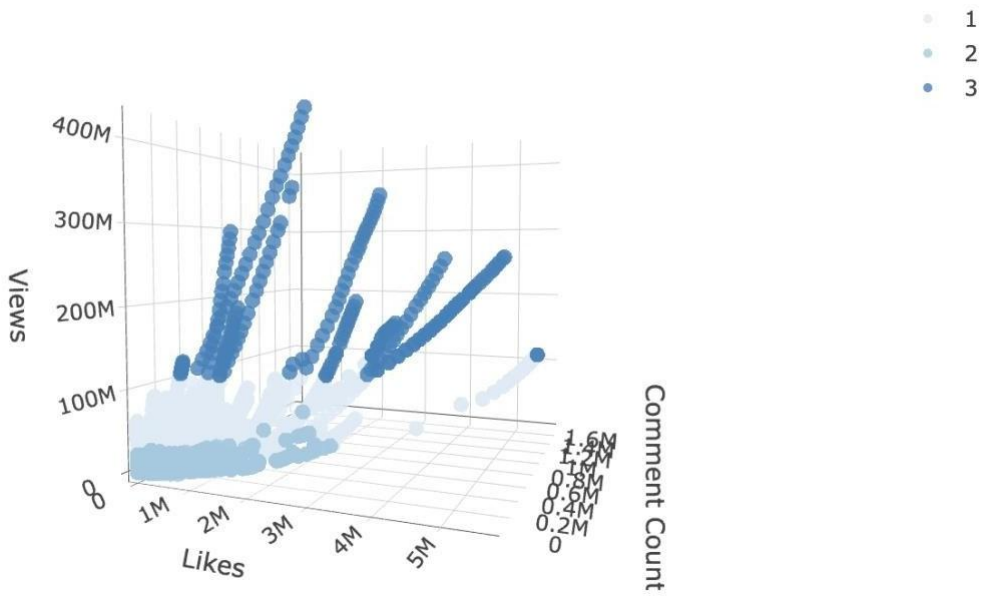
We then performed cluster analysis using the K-means algorithm to find patterns for a more accurate analysis of the data. We analyzed the variables of likes, dislikes, views, and comment_count .

WCSS was calculated for K values of 1 - 10, and we created a plot of "Number of clusters" vs "Within Cluster Sum of Squares (WCSS)" to find the best number of clusters (the K value). This

plot is known as an elbow plot, and this showed the point at which the WCSS slope begins to flatten out (see Figure 1). The point on the plot where WCSS began to flatten out was the best number of clusters to use for the K-means clustering algorithm, which will be referred to as the elbow point.



Using the elbow method in the image above, we can conclude that three clusters is the ideal number for this dataset. As we increase the number of clusters from one to three, this graph shows a large drop-off in the variance, but the drop-off declines again the more we increase clusters. This suggests that additional clusters greater than three, may not make the clustering give much better quality. We have chosen, therefore, a model with three clusters for this dataset.



We may deduce the following about each of the three clusters from the information provided about the k-means clustering:

1. Cluster 1 (Size: 2946):
    - "Data_total.likes," "data_total.dislikes," "data_total.views," and "data_total.comment_count" all have comparatively high mean values.
    - This implies that videos or content with a high amount of likes, views, dislikes, and comments are probably represented by Cluster 1.
2. Cluster 2 (Size: 372768):
    - Out of the three clusters, this one is the biggest.
    - Cluster 1 has much lower mean values than Cluster 3, whereas Cluster 3 has higher mean values for `data_total.likes`, `data_total.dislikes`, `data_total.views`, and `data_total.comment_count`.
    - With a modest number of likes, views, dislikes, and comments, this cluster most likely contains videos or content that are interesting and relatively popular.
3. Cluster 3 (Size: 228):
    - Of the three clusters, this one is the smallest.
    - The three clusters with the lowest mean values are `data_total.likes`, `data_total.dislikes`, `data_total.views`, and `data_total.comment_count`.
    - With fewer likes, views, dislikes, and comments than the other clusters, this cluster probably represents videos or content that are less well-liked and interesting.

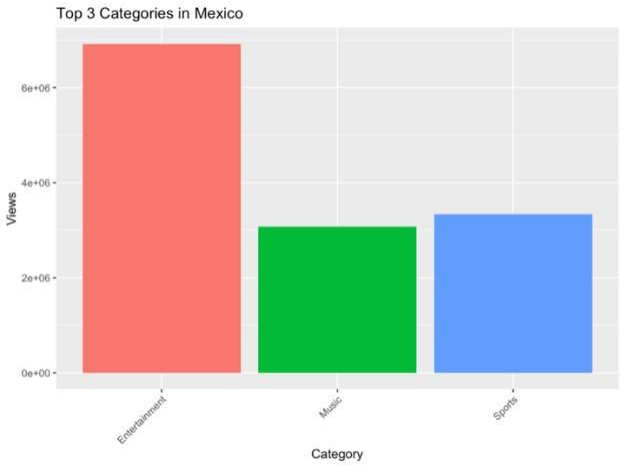
From the mean values of the characteristics, we can interpret from the data shared above that Cluster 1 is extremely popular and engaging, Cluster 2 is moderately popular and engaging, and Cluster 3 is less popular and engaging.

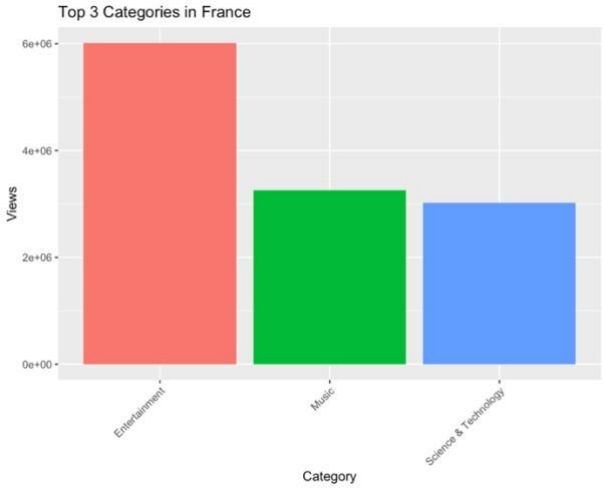
Information on the "Within cluster sum of squares by cluster" shows how well each cluster's data points are packed together. From the screenshot shown, we may deduce:

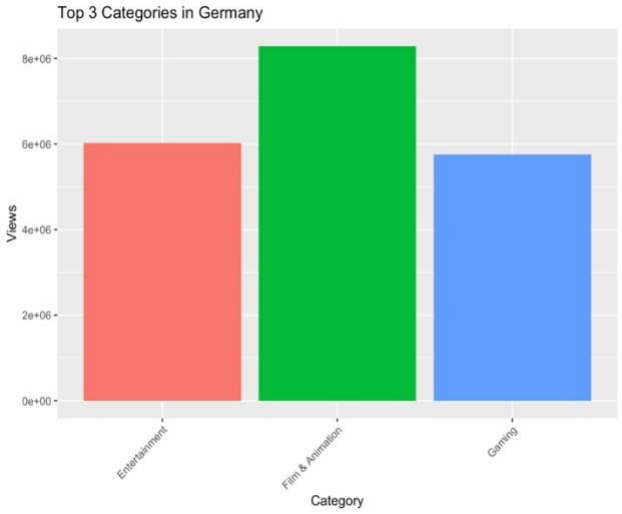
- Cluster 1: Within cluster sum of squares is approximately 1.299663e+18.
- Cluster 2: Within cluster sum of squares is approximately 1.827319e+18. - Cluster 3: Within cluster sum of squares is approximately 9.648403e+17.

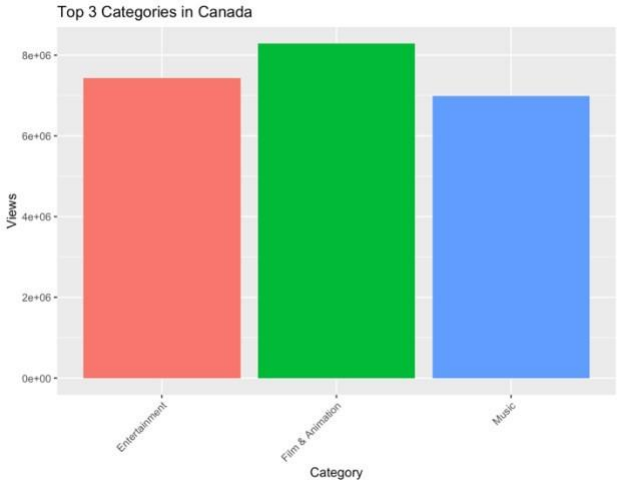
The "between_SS / total_SS" ratio also indicated that 78.4% of the overall variance the data has can be accounted for by the variance of the cluster means. That is a very large ratio indicating that the clusters capture a very significant part of the overall variance in the data, and that the clusters are clearly separated. They appear to each be distinctive.

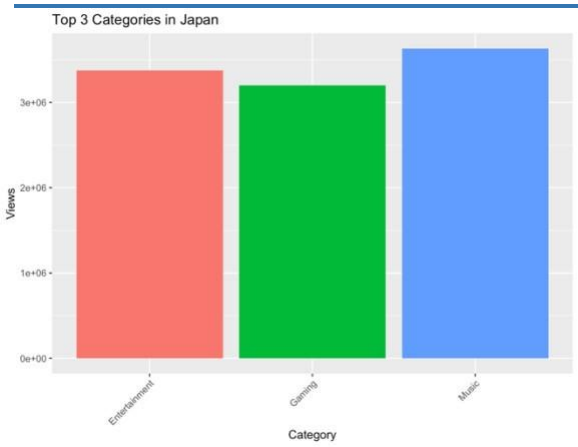
ANALYSIS ON SCRAPED DATA

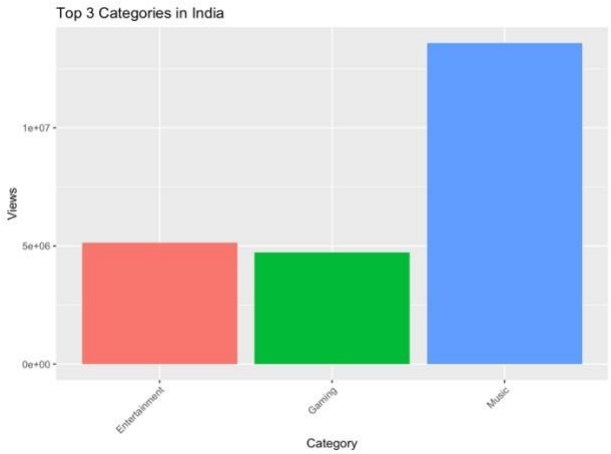


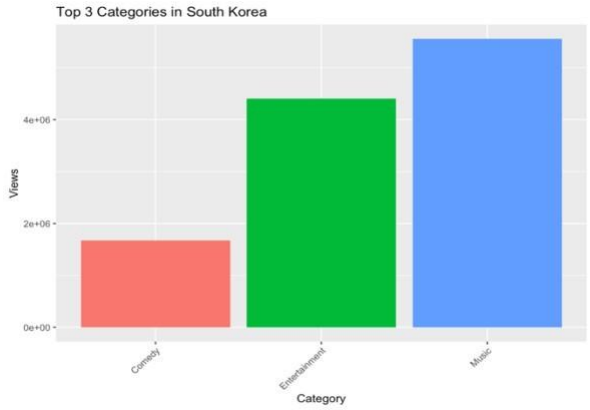


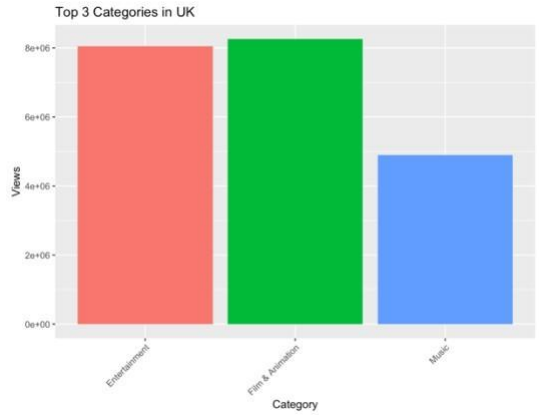


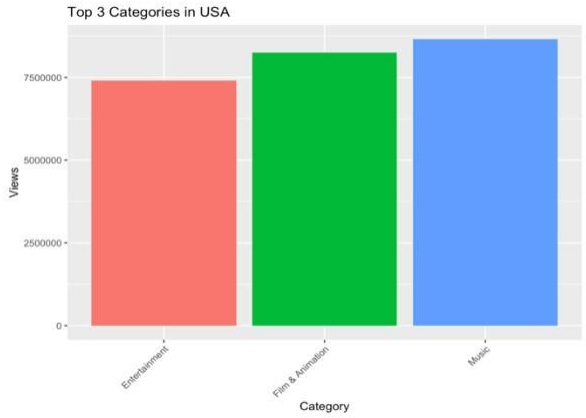


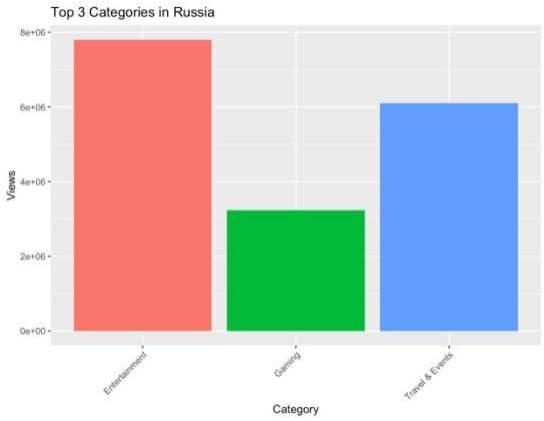












From the above plots we can observe the following information.

1. Category ranks: The two datasets also rank categories differently in respective orders of average and least amount of views. For example, in the Kaggle dataset the Music category ranks second (2nd), but in the scraped dataset it has the most average views in France.

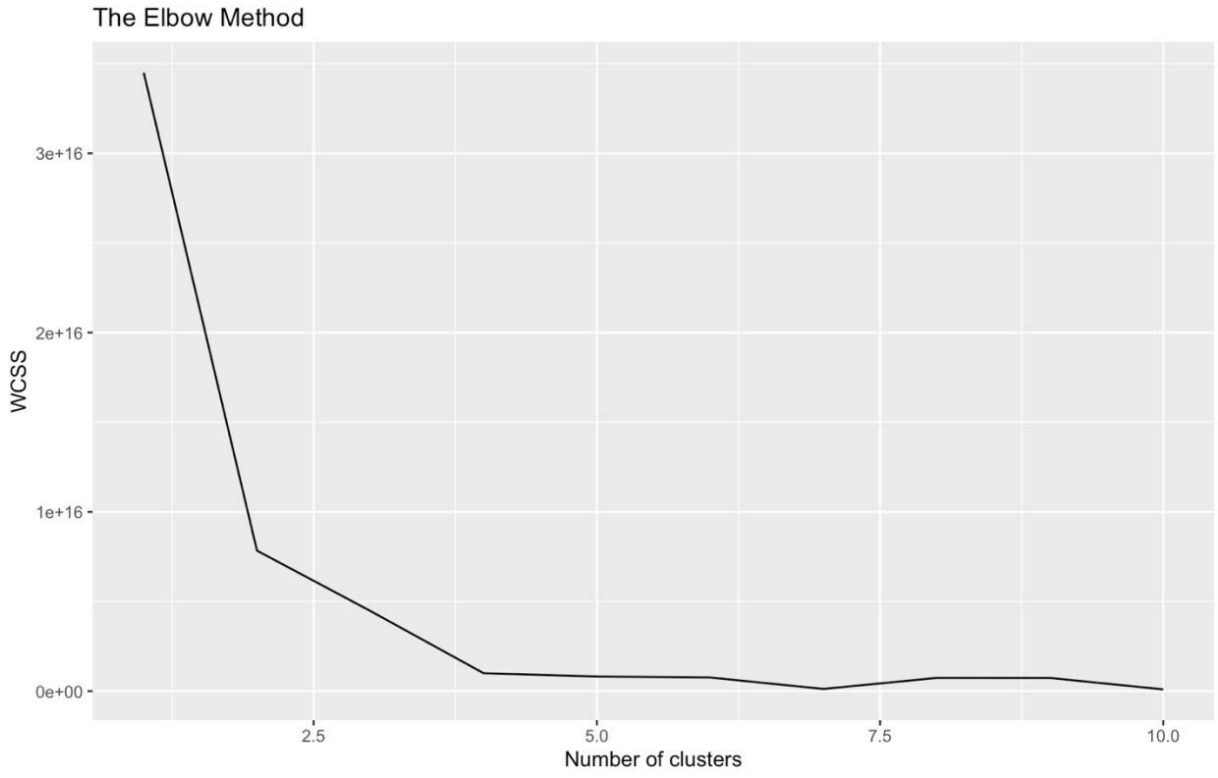
2. Category Popularity: Trends in category popularity differ among datasets and therefore can vary quite a bit. For instance, music remains overwhelmingly popular, being one of the highest ranked categories consistently – music appears to be a consistently popular category over all the datasets. Even within category popularity, popularity rankings within some of the categories among the datasets shows more fluctuation. For instance, Science & Technology and Nonprofits & Activism categories have wider variations (looking at both popularity ranking and total ridership counts).

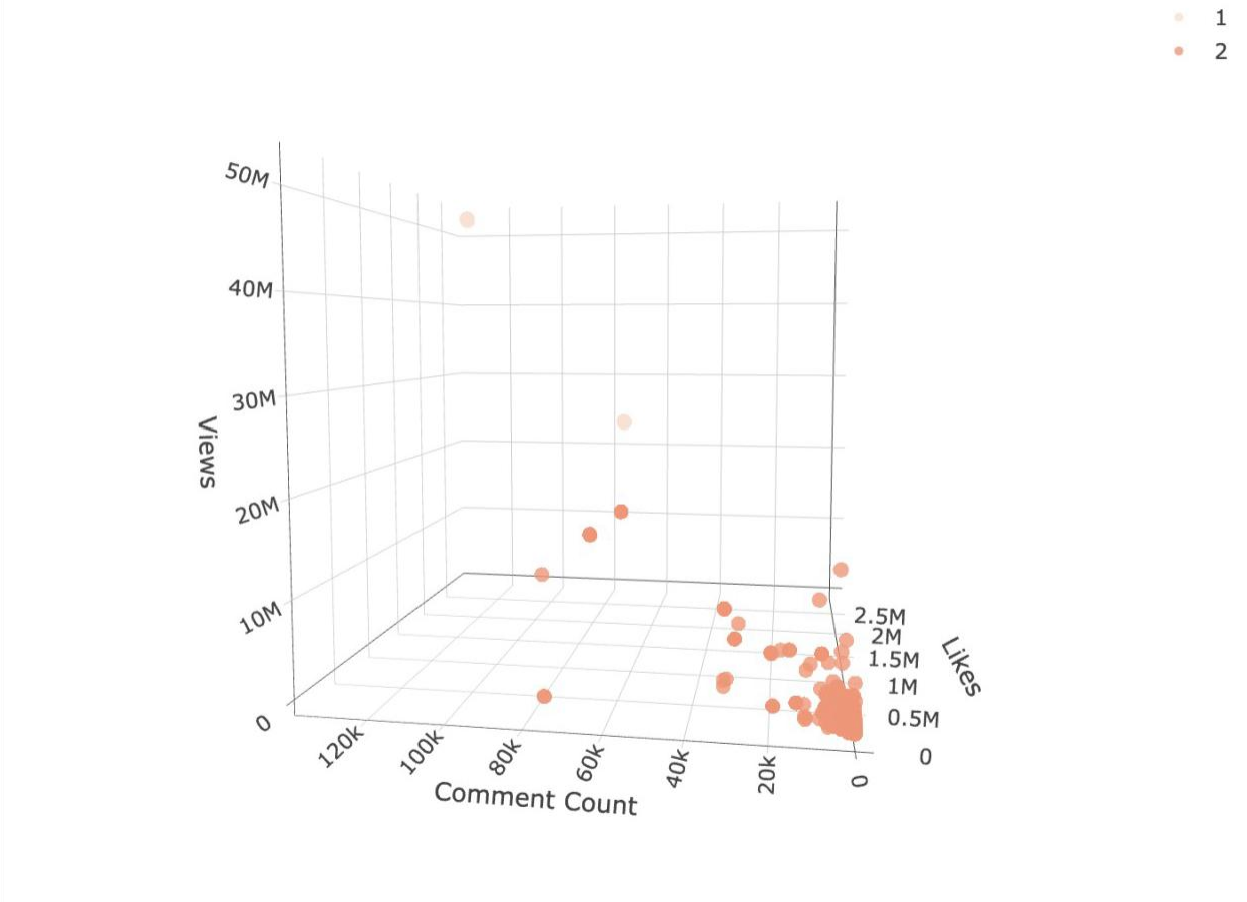
3. Global Trends: Despite these differences, there are several similarities among the trends. The rankings of entertainment, music, and film categories for example, indicate that these categories are very commonly used across a number of datasets.

By being aware of these variations, marketers and content producers can better adjust their approaches to target audiences with various datasets.

CLUSTER ANALYSIS

Following the elbow plot like we performed in the Kaggle dataset, we determined the best number of clusters. The below plot suggests the best value of k is 2. It is reasonable to have fewer clusters since the data set is small.





From the scraped dataset, we arrived at two clusters sizes 19 and 481. The cluster means for each variable were different between the clusters just like the Kaggle dataset. The WCSS said clustering explained 77.3% of the variance for this dataset.

When we look at the two datasets, we see that the Kaggle dataset had larger cluster sizes and more clearly differentiated clusters than the scraped dataset. Moreover, the within-cluster sum of squares was larger in the Kaggle datasets than the scraped dataset, which means that the clustering captures more variance in the data.

CONCLUSION

To gain greater insight into the factors influencing the popularity of particular videos, we wanted to conduct an exploratory data analysis (EDA) of popular YouTube videos. Our study provided informative knowledge related to a number of topics for popular videos, including how to choose categories, how to create titles of different lengths, and how to express sentiment in descriptions which may impact views.

For a more thorough understanding of the data, we also contrasted our results with recent patterns. To find more detailed information, we tried cluster analysis, but it did not work out as well as we had planned.

We could investigate methods to forecast video trends and go farther into comprehending the clusters found in future studies. This can entail creating a model that, given specific inputs, can forecast video trends and recommend changes to improve the chances of a video trending.

DATA SOURCES

Kaggle Dataset: https://www.kaggle.com/datasets/datasnaek/youtube-new/data

YouTube API: https://developers.google.com/youtube/v3/docs/videos

SOURCE CODE

Github Link: https://github.com/SannihithaGudimalla/CSP571DPA_PROJECT

BIBLIOGRAPHY

[1] Kousha, Kayvan & Thelwall, Mike & Abdoli, Mahshid. (2012). The role of online videos in research communication: A content analysis of YouTube videos cited in academic publications.
Journal of the American Society for Information Science and Technology. 63. 1710-1727. 10.1002/asi.22717.

[2] Rui, Lau & Afif, Zehan & Saedudin, Rd & Mustapha, Aida & Razali, Nazim. (2019). A regression approach for prediction of Youtube views. Bulletin of Electrical Engineering and Informatics. 8. 10.11591/eei.v8i4.1630.

[3] Pinto, Henrique & Almeida, Jussara & Gonçalves, Marcos. (2013). Using early view patterns to predict the popularity of YouTube videos. WSDM 2013 - Proceedings of the 6th ACM International Conference on Web Search and Data Mining 10.1145/2433396.2433443.

[4] Che, X, Ip, B, Lin, L (2015) A Survey of Current YouTube Video Characteristics. IEEE Multimedia 22(2): 56–63.

[5] Bärtl M. YouTube channels, uploads and views: A statistical analysis of the past 10 years. Convergence. 2018;24(1):16-32. doi:10.1177/1354856517736979

[6] Cheng, X (2013) Understanding the characteristics of internet short video sharing: a youtubebased measurement study. Transactions on Multimedia 15(3): 1184–1194.

[7] Borghol, Y, Ardon, S, Carlsson, N. (2012) The Untold Story of the Clones: Content-Agnostic Factors that Impact Youtube Video Popularity. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012, pp

[8] OpenAI (Nov 2022) ChatGPT (3.5) [Large language model] - https://chat.openai.com/chat

[9] Eliganti, Ramalakshmi and Reddy, A Bindhu Sree and G, Sharvani, YouTube Data Analysis & Prediction of Views and Categories (April 6, 2022). Available at SSRN: https://ssrn.com/abstract=4076559 or http://dx.doi.org/10.2139/ssrn.4076559

[10] Iman Barjasteh, Ying Liu, Hayder Radha , Trending Videos: Measurement and Analysis

[11] [7] X. Cheng, C. Dale, and J. Liu,"Statistics and social network of YouTube videos", IEEE IWQos, pp. 229-238, 2008.