Trending YouTube Video Analysis

MD SAKIBUR HASAN (shstat10@gmail.com)
Grand Valley State University
Bishal Sarker
University of Dhaka
Diksha Shrestha
Grand Valley State University
Roshan Shrestha
Grand Valley State University
Sajal N. Shrestha
Grand Valley State University

Research Article

Keywords: You Tube, trend, likes, views, title

Posted Date: February 7th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2548456/v1

Abstract

YouTube is an online platform where users may view, share, like, comment on, and subscribe to videos that are created by content creators. The quantity of views, likes, and the age of the video are used to select which videos are included in the trending video category on YouTube. But that is the main challenge faced by the content creators. This main goal of this study is to analyze the factors that influence popular YouTube videos. There are five tasks that were determined to complete: determining the relationship between likes and views; comparing the average time it takes for different categories to trend; identifying the most popular tags; determining the ideal title length; and determining the best day for the video to be popular. YouTube video trending across different countries which will help the content creators to get a better understanding. A dataset from September 2020 to January 2022, which is updated daily, helps determine the correlation between likes and views, the average hours taken to trend across different categories, the most used tags across countries, the optimal title length's range, the trend over the day of the week, and the optimal day to post the videos. A dataset from September 2020 to January 2022, which is updated daily, helps content creators and users across countries determine the correlation between likes and views, the average hours taken to trend across different categories, the most used tags across countries, the optimal title length's range, the trend over the day of the week, and the optimal day to post the videos, and helps them make proper content.

1 Introduction

YouTube is an online platform where the videos are published by the content creator and the users can view, share, like, comment and subscribe. YouTube is a growing video platform which is used by millions of people worldwide and the user engagement with YouTube has escalated within the past years. Similarly, nowadays, it is widely used by the business company's as well to advertise and promote their products, hence, contributing to their profit growth. YouTube also has a trending video category which is determined by the number of views, likes, and the age of the video. Additionally, the trending category might differ across different countries with people's choice of engagement towards the content of the video.

Although, there has been a study [3] in regards to the trending YouTube video in general, there hasn't been an analysis done across different countries in regards to the reason determining the trending YouTube video analysis. Our main goal is to analyze different character- istics that help in making the YouTube video trending across different countries which will help the content creators to get a better understand- ing. We want to analyze how the viewership influences the number of likes for the trending videos. Additionally, we aim to compare the aver- age time taken for the video to be trending across different categories. Identifying the mostly used tags and video's title character count are our other objectives we plan to achieve. Lastly, we want to pinpoint the day of the week that has the maximum number of trending videos.

Our motive is to help the content creators by providing necessary information to reach the maximum number of viewers and help them grow their YouTube channel. As YouTube is the second most popular website based on the total number of users [v7]. YouTube is paying their users for their content based on the view counts. Since YouTube is paying the youtuber for their videos nowadays many people dream of starting their career as youtuber and want to be featured in a trending page which is easy to gain views. This has made the platform more lu- crative for the existing and new users throughout the world. We are also curious in understanding how the current video contents uploaded by the channels are trending and explore what various attributes contribute to reach the trending page. We also want to gain information if these attributes have the same effect across various countries and categories.

2 Literature Review

The research [3] stated by Cheng, Dale and Liu focused on the system- atic measurement study on the characteristics of the YouTube videos. The data were collected from YouTube API and YouTube video pages for the 3 month period containing 27 datasets. They identified the growth trend, figured the pattern with the life span of the videos and length distribution in YouTube using 20 related videos. The research used the YouTube crawler to figure out the most viewed and top rated videos which totalled 189 unique videos and it was performed on a weekly basis providing seven datasets. According to their research, their dataset showed skewed distribution with music being the popular category comprising 22.9%, followed by entertainment with 17.8%. The lowest category were Howto and DIY, followed by Pets and ani- mals. The other characteristics were based on the video length which showed 97.8% of the popular video lengths were under 600 seconds because the music category contributed more to it as the music videos are often within that range. The characteristics also listed the file size of the videos which were mostly below 30 MB. Similarly, the article also included the date added characteristic to study the growth trend which showed a decreasing graph and the reason was because the uploaded video was not so popular. The views and rating characteristics was considered as one of the important characteristics by this research as it helped in identifying the popularity and patterns of the videos. The last characteristic that was outlined in the research was the growth trend of the number of views with their life span where they used a power law. The visualization that has been used in the research were bar chart, histograms, line graph and scatter plot. It described how each characteristic plays a significant role in making the videos popular. The research doesn't cover the popularity of the video trends across the countries and the study is based on the dataset of 2007. We will be analyzing the recent dataset from September 2020 to January 2022 across different countries.

The research [2] by Barjasteh, Liu, and Radha also focuses on the trending video's measurement and analysis. It analyzes the trending video's time-series over the 9 months period and the data were collected from the YouTube API. The article focuses on analyzing the viewership lifecycle, comparative analysis between the trending and non-trending videos, analyzing the trending video uploader's profile, and showing the directional relationship analysis with the view count on the trending video with the trending video categories. The dataset used was 4000 trending videos and 4000 non-trending videos, and once the trending videos were generated from the YouTube API, the statistics for the videos were then extracted. The analysis was made in 4 subsets where the data were collected in 15, 30, 45, and 60 days. The histogram distribution was used to show the aggregated view of the videos across the different time periods and the cumulative line graph was used to show the percentage of views with the number of days. There has been a comparative analysis between the trending and non-trending videos according to different categories using the bar chart. As per the analysis, it states, the time duration mean for the trending video is more than that of the non-trending videos. There has been a comparative analysis made on the number of views with the trending and non-trending videos. A line graph has been used which demonstrates that the number of views for the trending videos increases over the time, whereas, for the non- trending videos, the number of view count increases for certain days once it has been uploaded then the number of views gets saturated. There has been an analysis made on the profile of the uploader's where the 86% of the uploaders are male and14% uploaders are female for the trending videos. It has also taken into account how the subscriber counts influences the videos to be categorized as trending because 84% of uploaders have more than 100 subscribers and 6% have more than one million subscribers increasing the reach of the videos to the end users. Also, the directional relationship analysis is performed between the number of views with the trending videos as per the time-series using the Granger Causality. The article's main aim was to identify, measure and analyze the statistics of the trending videos.

According to the research [4] by Gajanayake and Sandanayake (2020), they aim to identify the trending pattern of YouTube gam- ing channels using sentiment analysis. The research also used machine learning methods to figure out the trending feature. However, the re- search is confined only to gaming videos posted on YouTube channels. The research collected around 1000 trending gaming videos through YouTube API. Similarly, the researchers analyzed the YouTube data upon preprocessing and converting it into meta-data, trending patterns of gaming videos on YouTube, and videos based on the sentiment analysis on comments from viewers. The authors have used Support Vector Machine (SVM), Naive Bayes, and Logistic Regression for the classification analysis. The research used a line graph to determine the most used words in the comment section of the videos. Additionally, the authors used a bar chart and histogram chart to show the publishing time and title length, respectively, to determine the trending pattern for the YouTube videos. This task is related to our project as we are also finding a perfect day to publish a video with a maximum number of views. However, this research focuses on the time and the number of videos, whereas our project focuses on the day with the view counts, and we will be using a line graph as our visualization tool. Simulta- neously, this research focuses on the title length with the number of videos to determine the title length. Our project focuses more on find- ing the title length of the trending YouTube video across countries. The research has shown how video publishing time, video title length, views and likes, and quality of the audio play an essential role in making a video trending on YouTube.

The research [5] by Hoiled, Aprem, and Krishnamurthy focused on the engagement and popularity dynamics of YouTube videos and their sensitivity to meta-data. The dataset consisted of 6 million videos from 26,000 channels. The research focused on the meta-level features such as title, tag, thumbnail, and description of the videos. Similarly, to ana- lyze the meta-level, the first-day view count, the number of subscribers, contrast of the video thumbnail, and title length are considered to find the view's popularity. For this, the researcher used machine learning methods to find out the sensitivity of the video with respect to the view counts of the video. Additionally, the Granger causality test has been used in the study to present how a view count has a causal effect on the subscriber's number. For YouTube channels, the subscriber growth in- creased with the increase in the number of view counts. The research's other task was to predict the view count of the YouTube videos, and it used a machine learning method (Extreme Learning Machine) for accomplishing the task. A line graph visualization is implemented to predict the view count using the ELM with the actual view count data. Similarly, a time series analysis method was used to identify the social interaction with view count. To measure the social interaction, it used casualty between the subscriber and the view count. Also, the dynamics of the YouTube video upload schedule have been studied as well. The research implies that the dynamics of the schedule do affect the number of views and the comments the video receives. As with our task, we are also finding out the optimal day schedule by looking at the statistics on which the YouTuber can upload their video to become trending. This research focuses more on its task concerning the average view count. Similarly, our project also has more of its task based on average view counts as it is one of the important factors influencing the video towards trending.

The research [1] by Andry, Reynaldo, Lee, Christianto, Loisa, and Manduri(2021) on the Algorithm of Trending Videos on YouTube Analysis using classification, association, and clustering. The main attributes used in this research were views, likes, dislikes, and com- ments. The data mining techniques such as classification, association, and clustering were used to find the YouTube algorithm. The research finds how the YouTube algorithm works and how the videos stay on the trending list. Through the use of classification methods, the researcher found that likes and views attributes played a significant role in the YouTube algorithm to find trending videos. The researchers then used the association technique and found out the views, likes, dislikes, and comments showed a relationship, and these attributes too played a role in the YouTube algorithm. As we found out how views and likes play a significant role in making a video trending, with our task, we focused on finding how the likes and views were correlated to know the association between them. Furthermore, the clustering technique was used where it grouped the attributes from 0 to 4 to determine which had the most effect on making the video trending, and the result concluded that the number of views played a vital role then, followed by likes and clicks, title count, and keywords. Hence the researchers concluded that the two factors contributing to the algorithm of YouTube are engagement which is through likes, views, count, and the other factor is metadata, that are title length and keywords.

3 Tasks

Our main aim is to analyze the attributes contributing towards the trending YouTube video and there are five tasks we are looking to accomplish that is finding the correlation between likes and views, comparing the average time taken for the categories to be trending, identifying the most used tags, finding the optimal title length and identifying the optimal day for the video to be trending.

Correlation between likes and views

To figure out the correlation between likes and number of views of the trending videos between September 2020 to January 2022 by scrutiniz- ing the like counts and view counts. Our first task is to discover the relationship between the likes and number of views for the trending videos. This will help us in identifying if there is any correlation be- tween them. Do all the videos with the highest number of views also have a similar number of likes? For this task, we'll be using the likes attributes and view count attributes from the dataset.

Compare average hours of trending video across mul- tiple categories

To compare the average hours taken for a video to become trending across categories between September 2020 to January 2022 by analyz- ing the published date, trending date, category, and trending duration. In this task, we want to find out how long a video category is taking to be featured on a trending page. There are 31 categories in the dataset from which we will get the information on the trending hours from the published date and trending date. We want to understand if the average time taken is consistent across multiple categories? We also want to add interactivity in the task by letting the user switch between different countries. Similarly, we can also allow the user to change the time range between hours, days, and weeks.

Identify most popular tags across countries

In this task, we want to identify the most used tags of the trending videos to help the end-users discover the video quickly between September 2020 to January 2022 across different countries by analyzing the tags used, country name, category name and category ID. The idea of finding the most used tags will allow the creators to use similar tags to reach the trending stage. What are the most used tags in the trending videos? Are the most used tags similar across countries? We will be extracting the tags from all the trending videos and it will be aggregated based on the number of times the tags were used. We can further aggregate the tags based on the countries and categories which will help us further analyze most used tags to help video reach the trending page.

Analyze the frequency of title length for the trending video

We want to analyze the overall frequency distribution of the title length for the trending Youtube videos to help the creators find the optimal title range for their content by analyzing the data from September 2020 to January 2022. For this task, we are using attributes title, frequency of the title's character and country name. Does the video title length play any role in making the video trending? Should the content creators focus on creating short titles or descriptive titles? For this, we will be extracting the length of the title for each video and aggregate them based on the countries.

Identify the trend over the day of a week with the total number of trending videos

We aim to identify the day of the week when the video is trending by analyzing the attributes trending date, view count, number of likes, comment and day of the week from September 2020 to January 2022. We focus on analyzing which day has the most number of videos trending? Is it similar across different countries? We will analyze the dataset of the video for each day and aggregate it to the day of the week. This is done by calculating the day from the date-time stamp of the trending date attribute. Additionally, in the interactive, we can further aggregate the data on a monthly basis and also allow the user to switch between different countries to see the trend.

4 Dataset

The trending video dataset [6] of YouTube is available in Kaggle from September 2020 to January 2022 which is updated daily by the user Rishav Sharma. The dataset is approximately 2.0 GB in size as of February 25, 2022. The author has extracted the data from YouTube API using a crawler. The dataset incorporates the information for the daily trending video for the USA, Canada, Britain, Japan, Germany, France, Russia, Brazil, Mexico, Japan, South Korea. Each country's data is stored in individual csv files that contains information about upload date, channel or user, trending data, and other metadata like tags, view count, likes, and dislikes. Likewise, the category information is also stored separately by country as they differ from country to country in json format. The attributes used in the dataset are:

- title - name of the trending video.
- publishedAt - datetime stamp when the video was published.

- channelTitle - name of the YouTube channel.
- categoryTitle - name of the video category.
- trendingDate - datetime stamp at which time it features on the trending page of the specific country.
- tags: labels used by channel to be easily recognized by the YouTube algorithm.
- views: count of views from the user
- likes: count of likes in the video
- dislikes: count of dislike in the video
- commentCount: count of comments in the video.

The date-time stamp used in publishedAt and trendingData are in UTC time format. Tags attribute in the dataset are in text format where each word is separated by ”—” similar to csv file format. For our analysis we will be extracting each tags for the individual video.

5 Project Design

Correlation between likes and views

This visualization aims to discover the correlation between the likes and number of views for the Youtube trending video from September 2020 to January 2022. Initially, we considered using a bubble chart as our visualization. This is because they are helpful in analyzing the relationship between numerical or categorical variables. It is very similar to a scatterplot and can also be considered an extension of a scatter plot. However, our tasks only consider two numerical variables to represent the tasks, and a bubble chart can be pretty excessive to achieve it. Also, our dataset is very large, so the area of the bubble can make the chart look chaotic. Furthermore, area encoding is not an effective visual encoding as per Cleveland’s rules.

The following visualization we came across was a heatmap. Heatmaps can be used to show relationships between two variables which can be ideal for our task. It will allow us to discover any patterns between the likes and number of views in the dataset. However, as we have a large dataset, it can be overwhelming for the users to draw any conclusions, and they might miss out on details. Additionally, as per Cleveland’s rule, we are not good at decoding color encoding, which is a primary visual encoding for heatmaps.

We are inclined to use the scatterplot visualizations due to the above reasons. Scatterplot graphs can help us to detect relationships between two variables. We felt that this is perfect for our use case, i.e., to show the correlation between likes and views for the trending Youtube video. Likewise, as our dataset is large, this visualization scales very well and can help the users identify the correlation between them [1]. In addition, we also found that scatterplots can help us in discovering patterns and detecting outliers at the same time as well. Scatter plots are very simple, and most people are used to analyzing them.

The visual encodings applied in this task are position and color. We use position encoding to plot the data for like on the x-axis and the view count on the y-axis. Cleveland’s rules rank position encoding to be highly effective as humans are better at recognizing positions in a visualization. This allows the user to quickly locate the instance of the dataset. In regards to Gestalt’s principles, the law of proximity can be applied as the data that are placed together seem to be more related.

Color is another visual encoding we have used in this graph. We chose the blue color for this task as it signifies calmness. As a result, it makes the visualization easy on the eyes of end-users and makes our graph easier to comprehend. As per Cleveland’s rules, color encoding is not so effective in terms of effectiveness ranking. However, since we are using a single color and the position is set as our primary encoding, it should not affect our visualization. Furthermore, Gestalt’s Law of similarity shows the relationship between likes and views using the same color.

Compare average hours of trending video across mul- tiple categories

We aim to compare the average hours taken for a Youtube video to reach trending across categories between September 2020 to January 2022. For this task, we analyzed different visualizations, such as donut charts and bubble charts. Initially, we planned to use a donut chart for visualizing the task of comparing the average hours to trends between the categories. Nevertheless, after analyzing the graph, we found that the donut chart is not as effective in ranking as compared to the Bar chart. The donut chart uses visual encoding such as area and angle, which can be more challenging to read and analyze then length according to Cleveland's rule. Furthermore, we have around 14 categories of youtube videos, and the donut chart might not be ideal for this task, as it will be difficult to read and understand when there are many categories since that will reduce the area each category can use to represent its data.

The following visualization that we came across for our task is a bubble chart. The bubble chart is helpful in showing the relationship between the categorical variables. It is also reader-friendly and easy to understand. However, we stumbled upon a couple of issues when using this visualization. We found that this graph is ideal when there are few variables. Additionally, according to Cleveland's rule, humans are worse at comparing areas than they are at comparing the lengths. So, it will be difficult for us to discriminate between two or more categories if they have similar values.

Hence, we finally selected the bar chart as our visualization choice for the task. This graph can be used to emphasize different values, more importantly, the dissimilarity between them. Each bar represents the youtube category, and its length represents its value, i.e. average hours. These bars are then presented over a common scale. For this task, we found the bar chart helps the end-user easily perceive the trend of average hours taken to come at trending within different categories. The bar chart also helps to compare the data across different categories and helps in sorting to make the graph easy to read and understand. Furthermore, we can use the bar chart to show the frequency distribution of average hours taken to be trending for each category.

For the task, the visual encoding used is position, length, and color. We use position encoding to place all the categories next to each other along the x-axis in the visualization. As per Cleveland's rule, humans are best at identifying the position first as it helps the end-user iden- tify where each category is placed along a common scale. Similarly, Gestalt's Principles Law of Closure is reflected by the white space between the graphs as a separator between the categories.

We are using length visual encoding in the visualization to represent the average hours taken for a Youtube category to trend. Additionally, we are sorting the position of the category based on their length. This helps us to recognize the category which trends quickly as well the categories that take the longest amount of time to trend. Applying Cleveland's rule, the length of visual encoding falls into the higher spectrum of the visualization's effectiveness ranking. This means that by using length to represent the average hours, we can be assured that the end result can be conveyed without a doubt. In terms of Gestalt principles, we can apply the law of Pragnanz for our visualization tasks. This law states that our mind loves simplicity and is gravitated towards finding simple patterns that are regular, even, and orderly. Using this principle, we can quickly identify the categories that trend much faster at a glance or vice versa.

Furthermore, the next visual encoding we thought of was color encoding. Initially, we had the idea of using different colors for each visualization. But this made the visualization quite confusing, and different color scales did not bring additional meaning. For this reason, we chose to visualize the data using a single color. Additionally, it makes it much easier to compare, understand, and reduce the cognitive load of the end-user when using multiple colors. Applying Cleveland's rule, the color encoding is much worse in terms of the effectiveness ranking for the visualization. Speaking in terms of Gestalt principles, we can apply the law of similarity as we are using the same color. This makes the end-users read and understands the data effortlessly. Likewise, we can also apply the law of figure and ground. This law states that objects are either perceived as a figure or ground. In the context of our task, the bars in the chart are identified and seen first as it is the foreground object as compared to the plain white background. The contrast between the figure and ground helps the user identify between two or more objects. Lastly, we chose the color blue to represent a calm and neutral color.

Identify most popular tags across countries

We want to identify the most used tags from the Youtube trending videos across countries from September 2020 to January 2022. We want to develop a treemap visualization to represent the top most used tags grouped by countries to implement the visualization. Treemaps data visualization is suitable for showing extensive hierarchical data and providing a high-level summary of the dataset between one or more categories. Each data value is represented by a series of nested rectan- gles proportional to the size. In the chart, the size of each rectangle box represents the count of the tags. The tags are enclosed based on their countries. The graph allows the interpreters to rapidly find the most used tags from each country presented in the visualization.

In the context of tasks, we can also implement visualizations tech- niques like word cloud and histogram. The word cloud visualization is straightforward to interpret and aids in making the frequently used tags quickly stand out from the rest. However, we found that the world cloud is not an effective tool to show all the most used tags accurately with a vast dataset. Furthermore, performing comparisons across multiple countries can also be quite challenging. Regarding histogram chart, it is very popular and mostly used to demonstrate the continuous frequency distribution. At first glance, we find it to be a good fit for the task as we can show the frequency distribution of the tags used. However, we will have too many vertical bars to represent the tags when implement- ing this chart. Moreover, similar to word cloud visualization, making cross-comparisons with different countries might not be as effective.

In this visualization, we are using area and color visual encoding. The area visual encoding is used to represent each tag inside a rect- angular box. Furthermore, the tags are nested inside another box to represent the country. The size of each rectangle is proportional to the number of times the tags are used. As per Cleveland’s rule, the area falls into the middle spectrum of the visualization effectiveness ranking. Nevertheless, it is a much more effective rank than color encoding. Similarly, we can apply Gestalt’s law of enclosure in the chart as all the tags inside a region are enclosed by a boundary around them. Again, Gestalt’s law of focal point shows the most used tag represented by the topmost used tag.

The color visual encoding is applied to visualize the tags. In this task, we are using a categorical color scale to represent the countries. We are assigning a unique color to each country in the visualization, and the tags grouped inside the country have the same color. This will allow the users to quickly identify the countries easily. Cleveland’s rules rank color encoding as the least effective in terms of effectiveness ranking. However, when color encoding is used correctly, it can result in an effective visualization depending on the task. Gestalt’s law of similarity shows that the tags with the same color belong to a specific country.

Analyze the frequency of title length for the trending video

We want to compare the video’s title length frequency across countries between September 2020 and January 2022. We planned to use a heat map to display the frequency of title length in the initial phase, where color encoding could be used to display the frequency of title length. However, it had some limitations where the character frequency was hard to differentiate because of the color encoding. As per Cleveland’s rule, humans find it hard to recognize the color difference compared to other visual encodings.

We also considered the cumulative density graph to represent the frequency and number of characters of the video. Cumulative density graphs show the distribution of frequency with the help of curve lines in the graph. However, it is pretty difficult to interpret the frequency distribution because humans find it difficult to analyze the slope of the graph as compared to position and lengths according to Cleveland’s rule. Hence, due to these reasons, we decided to implement a histogram chart. It allows us to display the frequency distribution of the length of the character in the video title. The histogram is also helpful in finding out intervals, and it is pretty easy to understand and find out the distribution of the data. Similarly, this visualization allows the user to quickly compare the height of the title length.

There are three primary visual encodings used in this task: position, length, and color. The first one is the position where the x-axis rep- resents the title length, and the y-axis represents the frequency of the title length. Also, Cleveland's rule states that humans can identify the position accurately.

The second visual encoding to consider is the length. The length encoding demonstrates the title length of the video as per its occurrence. The length also helps us find out instantly which title length has the most frequency. As the length falls under a higher rank in Cleveland's Rule, the end-user can quickly interpret the data by looking at the histogram. Similarly, Gestalt's Law of Pragnanz shows we can quickly interpret the interval of the title length for the trending video.

The final visual encoding used is the color selection. We have used the blue color to show the frequency of the title length. Additionally, we have used an orange color to grab the user's attention in identifying the most used title length for the video to become trending. Although the color falls under the lowest spectrum, as per Cleveland's rule, the user can easily identify the most used title length by the orange color. Gestalt's Law of focal point shows the title length with more frequency stands out and is noticeable by the end-user from its length and different color.

Identify the trend over the day of a week with the total number of trending videos

We want to visualize a trend that demonstrates the day of a week with the maximum number of trending videos. To accomplish this visualization, we plan on implementing a line graph. We chose this graph because it is a widespread visualization technique. It is also simple, easy to understand, and helps show the value of how something changes over time. In this task, a line graph can be a great tool to show the trend for identifying an ideal day of the week for the video to be trending across countries. We can also use different lines to show the trend of the multiple countries.This task can also be implemented using a bar chart and pie chart. We found that the bar chart helps us demonstrate the distribution of data values and compare values across multiple categories, in our case, day of the week. Bar charts are also straightforward, easy to understand, and accessible to most users as it is one of the most fundamental charts. However, bar charts are not an ideal choice when it comes to displaying trends over time. Similar to the bar chart, the Pie chart is also very easy to understand and can help us demonstrate the proportion of each day having the most trending video. However, pie charts can be poor at communicating data because it does not has any scale for reference, and human minds, in general, are not so good at comparing the size of the angles as per Cleveland's rule.

The visual encoding used in the graph is position, angle, and color. The position encoding is used in the graph to demonstrate the number of aggregated trending videos over the y-axis and the day of the week over the x-axis.
Cleveland's rules rank position to be the most effective visual encoding in visualization. This allows the user to interpret the graph effortlessly. Gestalt's law of connection shows that objects that belong to the same country are connected as a group. As each point connects to another, we can easily find the trending video over the week for different countries. Gestalt law of proximity shows that the lines that are close to each other have similar interests and characteristics.

We use angle visual encoding to connect the individual data values for the same country to show a trend. The upward slope shows the increase in the number of trending videos, and the downward slope expresses the decreasing trend. Cleveland's rules also support the slope to show the trend for the videos effectively.

The other visual encoding that we have used is color. The color encoding is used to represent the individual countries to identify them distinctly. We have implemented a categorical color scale where each color represents a country. The categorical color scale allows the user to identify the countries quickly. Additionally, we are using blind safe colors to represent the countries. Although Cleveland's rules rank color encoding to be less effective, combining this with position and angle visual encoding added more depth to the overall visualization. Gestalt's law of similarity shows that the data values with the same color belong to the same country. Gestalt's law of figure and ground shows the user will have a glance at the information on the graph and then the legends.

6 Visualization And Analysis

The visualizations were created using the ReactJS, Nivo charts, and JavaScript on the front end. The data is produced by an API server im- plemented using Python and FastAPI library. The dataset was imported to the MongoDB database using a python command-line script where we pre-processed the data to extract and transform the attributes. The visualizations produce the results on showing how each characteristic helps in making the video feature on the trending page. Hence, the content creator can get a better understanding of how to make their videos trending.

Correlation between likes and views

This task analyzed the correlation between the views and likes of the trending video from September 2020 to January 2022. Our main aim was to figure out if the highest number of views also had a similar number of likes and to explore the relation between the two attributes. The scatter plot effectively visualizes the relationship between the likes and views, as depicted in Figure 1.

We extracted views and like counts for each video from the database. We used a Nivo chart scatter plot component to generate the visualiza- tion. Additional parameters such as color, labels, and scales were also provided to the component. Users can also choose to switch between different countries by interacting with the country menu. The React instance will fetch the data from the API and cause the visualization to re-render with new data.

We can observe a strong correlation between likes and views as the figure shows a linear regression line. Also, we can see the density of the views and likes are found in the lower range as it is clustered together. Gestalt's Law of Proximity is applied in the visualization, where the views and likes in the lower range are more related to the views than the points that lie farther apart. The use of a scatterplot has helped the end-user to know the relationship between likes and views by immediately looking at the figure, and we can also see that there are no outliers here. The bubble chart and the heat map would not have perfectly shown the correlation and the outliers on the likes and views. From the visualization, we can analyze that the view count of about

60 million also had a similar like count of approximately 6 million. Also, as per Cleveland's rule, the position among the standard scale is applied, making the visualization clear for the end-user. Similarly, using the same shape and size of each point in the scatterplot made the visualization pleasing to the eye. Furthermore, using the same color for each point made the visualization look simple, and the end user can easily comprehend the information on how the views and likes are correlated to each other. This also applies to Gestalt's Law of Similarity due to using the same color and size. Hence, the visualization in Figure 1 accurately shows the relation between the views and likes through the use of visual encoding further strengthening our design as previously mentioned in the design phase.

Compare average hours for trending videos across YouTube categories.

Our task of comparing the average hours taken for YouTube trending videos across categories from September 2020 to January 2022 can be answered with Figure 2. The bar chart effectively visualizes all the categories of the U.S.A, with the average hours taken for each video to be trending. We aim to determine if the average time taken was consistent over multiple categories for this visualization. Additionally, we made our visualization interactive to allow the users to switch between the countries to find out if the average time taken for different categories to be trending is similar or different.

For the visualization, the published date, the trending date, and cate- gory name attributes were extracted from the database. We generated the hours taken for a video to be trending by calculating the difference between the published date and the trending date. Lastly, we aggregated calculated hours by averaging them in each category. This data is then fetched by the React front-end to render the bar graph using the Nivo bar graph module.

The immediate insight we can get from the bar graph is that the Sports category takes the minimum average time, which is approxi- mately 115 hours, to become trending compared to the other categories in the U.S.A. It is followed by Nonprofits and Activism, taking 119 hours, and the Travels and Events category with an average of 123 hours. Similarly, the music category takes the longest to trend, which is 139 hours. The use of bar graphs helps the user quickly identify the category that takes the shortest and longest time to become trending due to its position and the length as per the Cleveland's Rule. Addi- tionally, the graph was more transparent and easily understandable due to applying a single color, blue. As discussed in our previous design phase, the donut chart and bubble chart considered area encoding, and as per the Cleveland's rule, humans are best in identifying the length than the area.

Furthermore, we can also find the categories that had similar average hours to trend. Looking at the bar graph, we can identify categories such as Education, Entertainment, and News Politics with similar average hours of 132. The other categories, such as Comedy and Pet and Animals, also had the same average hour, i.e., 133 hours. Likewise, the Auto Vehicles and gaming categories, followed by Film animation and Howto style categories, also had similar average hours of 125 hours and 134 hours, respectively, to trend. This visualization also applies Gestalt's figure and ground principle, where the user first sees the bar graph as compared to the number associated in each bar. Similarly, Gestalt's law of similarity helped us to identify the categories that are alike. The focus and context navigation strategy made it easier for the user to discover the accurate average hours taken for each category. Furthermore, the design decision taken in the previous phase for visual encoding like position, length and color helps to further strengthen the task goals that we were trying to achieve on finding out the average hours taken to be trending across the categories.

Identify most popular tags across countries

We wanted to analyze the most used tags across the countries to be trending from September 2020 to January 2022. The treemap in Figure 3 shows the top 10 most used tags across 11 different countries. Similarly, the treemap effectively visualizes each word tag that is used in different countries. From the visualization, we intend to determine if the most used tags were similar across countries. Additionally, our interactive visualizations let the end-user input the tag from top 10 to top 50, where they can get an insight into the most used tags.

For this visualization, we generated the graph using the Nivo chart treemap module. The tags for each video in a country were extracted from the database. Since the tags were in a single text form, they were further separated into individual words. In the next step, we calculated the count of tags and aggregated them by country. The front-end app will collect the processed data and initialize the graph with the Nivo chart treemap component with additional attributes such as colors for individual countries and the size of the borders.

Examining the visualization, we can observe that all the countries have empty tags with the highest frequency compared to other tags. The graph shows the area for each tag, and the size of the area is proportional to the number of times the tag is used. Almost every country had a larger area with none tag i.e., the trending video did not use any tag. Similarly, we can see that this visualization applies Gestalt's Law of Enclosure, where the border encloses each country. Furthermore, we can also get an insight that most countries had the tags such as vlog, comedy, funny, football frequently occurring.

As defined in our design phase, we found that the treemap was a perfect choice for this task because it included the most frequent tags across countries in one chart, which helps the end-user easily compare the tag across countries. Similarly, if the word cloud was used, we could not quickly identify the most used tag as the dataset was large in number. Additionally, although a histogram would have been great in showing the frequency, comparing tags across each country would have created many histograms. It would not be an easy task to compare the tags across countries from single visualization. Likewise, visual encoding helps make a clear and straightforward visualization, whereas color encoding helps separate different countries. The use of a categorical color scale is also helpful for the end-user as they can quickly identify the different countries, and each tag under each country is visible. Consequently, the visualization in Figure 3 effectively shows the most used tags across the countries.

Analyze the frequency of title length for the trending video

The histogram visualization in Figure 4 illustrates the title length of the trending video from September 2020 to January 2022 by providing us an insight into how much a video should have its title length to reach the trending. The use of color lets the end-user quickly identify the title length required for the video to be trending. Our primary motivation was to discover the distribution of the title length so that the end user and the content creator could get an overview of the optimal title range for their video.

For the histogram visualization, we extracted each video's title from the database and then calculated the length of the title. We then aggre- gated the title length by frequency. Subsequently, the ReactJS javascript library will fetch the processed data from the API service and render the graph using the Data-UI histogram module, which uses the frequency distribution of title length to complete the task.

We can get an immediate insight from the histogram that the optimal title range for the video to be trending is from 40 to 50 title lengths.

Cleveland's position and length encoding help the user compare the length of the video's title quickly. It is then followed by the range of 30-40 and 50-60 title length. The least used title length is between 0 and 10. As described in our design phase, visual encoding strengthened our choice of using this visualization. Furthermore, we also got an insight on our second task where the content creators can have a descriptive title ranging from 40 to 50 than that of the short description title to be trending.

Similarly, the color encoding helps highlight the highest frequency title range, making it visually appealing to the end-user to get the insight at first sight. The heat map and the cumulative density graph
would not have been able to provide the findings of the task as the histogram perfectly visualizes it as the humans are worse at analyzing the color and slope than the length as per Cleveland's Rule. Furthermore, with the orange color highlighting the optimal range, Gestalt's law of focal point is also applied here, providing an optimal title length to be trending.

Identify the trend over the day of a week with the total number of trending videos

For the final task, our motivation to identify the optimal day of the week for the maximum number of views count of trending videos can be seen in Figure 5. We wanted to find the day in the week which had the maximum user engagement making the video on the trending list. Similarly, we wanted to compare if it was similar across the countries. The line graph helps in showing the trend over the week and also provides an insight into the day with maximum engagement.

To implement the graph, we first collected the trending data and the view count from the database. The trending date attribute was converted to a day of the week using Pandas library. In the next step, we further aggregated the view count by the day of the week. The ReactJs will load the processed data from the API and then instantiate the Nivo line chart component on the front end. The line chart component takes the data, colors, labels, and axis scales and then renders the graph.

The line graph shows Brazil has the lowest video count among 11 countries, whereas the United States has the highest number of view counts. The use of color encoding and implementation of the categorical scale helps the user differentiate between different countries. Similarly, from the line graph, we can analyze that countries like the United States, Russia, Mexico, South Korea, Japan, India, and Great Britain had the highest user engagement on Sunday. Initially, during our design phase, we thought that the line graphs of different countries would intertwine with each other. However, our final visualization was completely different than we thought, where none of the countries collided in terms of the view count. Similarly, the position encoding was used, which helped the user identify the day of the week on the x-axis and the view count on the y-axis. Furthermore, we can also see that Gestalt's law of proximity is applied where the lines that are closed together show a similar interest of higher engagement during that day of the week.

We can get insight for our second task of finding if the countries had a similar trend. The countries listed above showed a similar trend of lowest engagement being on Friday and highest being on Sunday. Furthermore, for Brazil, the maximum engagement was on Monday, and the lowest was on Friday and Wednesday. Similarly, Canada, Germany, and France also showed a similar trend, with the highest being on Sunday and Monday, followed by the lowest on Friday. With 11 different countries being portrayed in one visualization using a line graph, the user can easily compare the trends across the countries, whereas if we had used a bar chart or pie chart, there would have been 11 different figures for each country. Hence, we can see from the line graph that a perfect day for posting a video to be trending is Sunday for most countries.

7 Conclusion

Our motivation for the paper was to investigate the characteristics that determined a video to be trending across the countries. This report helped in determining the correlation between likes and views, the average hours taken to trend across different categories, the most used tags across the countries, the optimal title length's range, and the trend over the day of the week along with the optimal day to post the videos. The visualization for the last task did amazed us, with not even one country's line graph overlapping with each other, whereas the result for other tasks was interesting to know across the countries. This report will benefit the content creators and the users across the countries who want to make their video trending as they will get an overview of the characteristics of trending videos with each task described in our report, which was our main motive.

References

1. J. F. Andry, S. A. Reynaldo, K. Christianto, F. S. Lee, J. Loisa, and A. B. Manduro. Algorithm of trending videos on youtube analysis using classi- fication, association and clustering. In *2021 International Conference on Data and Software Engineering (ICoDSE)*, pp. 1–6. IEEE, 2021.
2. I. Barjasteh, Y. Liu, and H. Radha. Trending videos: Measurement and analysis. *arXiv preprint arXiv:1409.7733*, 2014.
3. X. Cheng, C. Dale, and J. Liu. Understanding the characteristics of internet short video sharing: Youtube as a case study. *arXiv preprint arXiv:0707.3670*, 2007.
4. G. Gajanayake and T. Sandanayake. Trending pattern identification of youtube gaming channels using sentiment analysis. In *2020 20th Interna- tional Conference on Advances in ICT for Emerging Regions (ICTer)*, pp. 149–154. IEEE, 2020.
5. W. Hoiles, A. Aprem, and V. Krishnamurthy. Engagement and popularity dynamics of youtube videos and sensitivity to meta-data. IEEE Transactions on Knowledge and Data Engineering, 29(7):1426–1437, 2017.
6. R. Sharma. Youtube trending video dataset, 2022. https://www.kaggle.com/rsrishav/youtube-trending-video-dataset.
7. Statista. Most popular social networks worldwide, 2022. https://www.statista.com/statistics/272014/global-social-networks- ranked-by-number-of-users/.

Figures

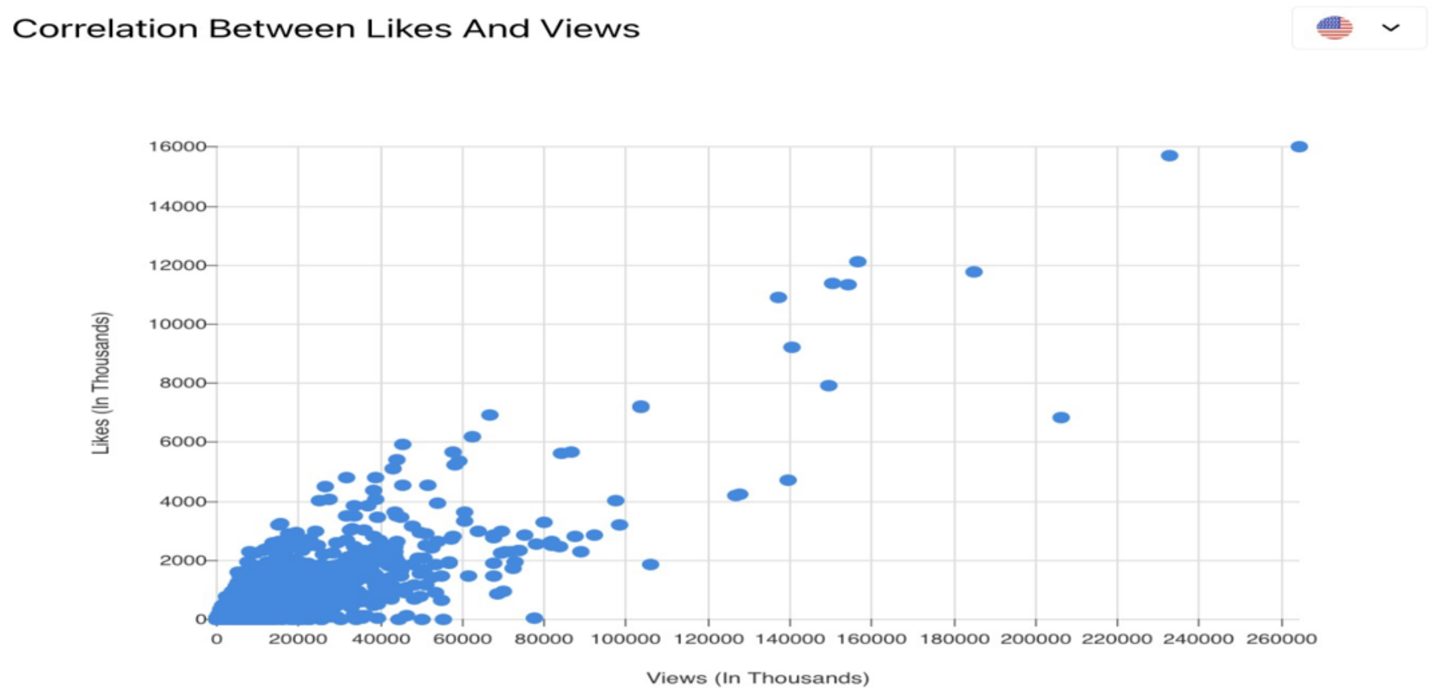

Figure 1

Correlation between likes and views.

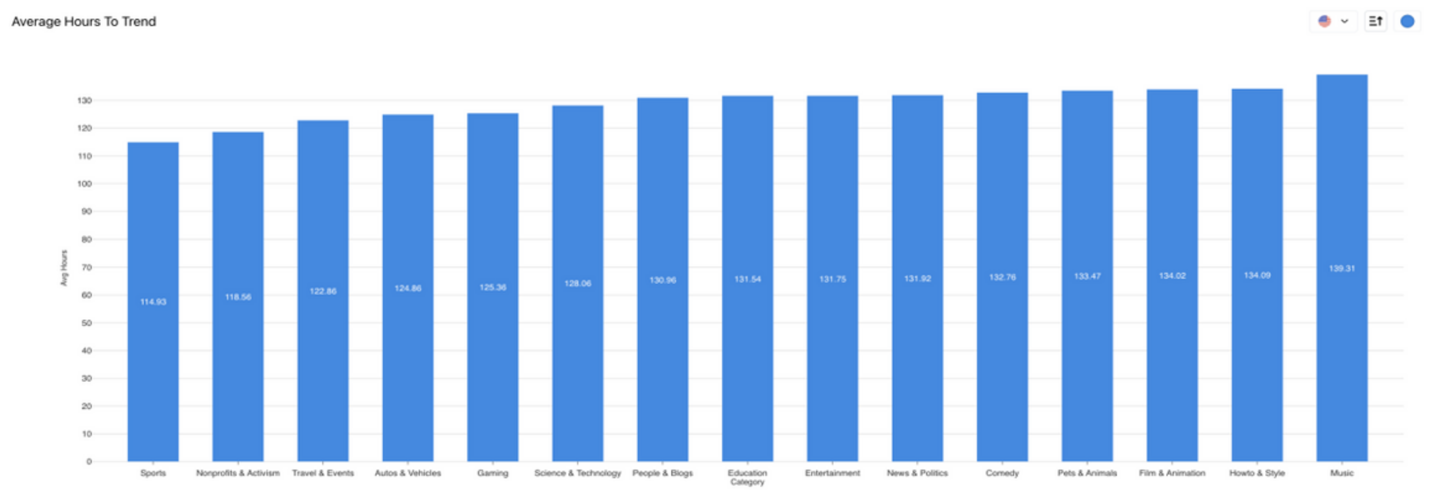

Figure 2

Bar graph comparing average hours for trending videos across YouTube categories.

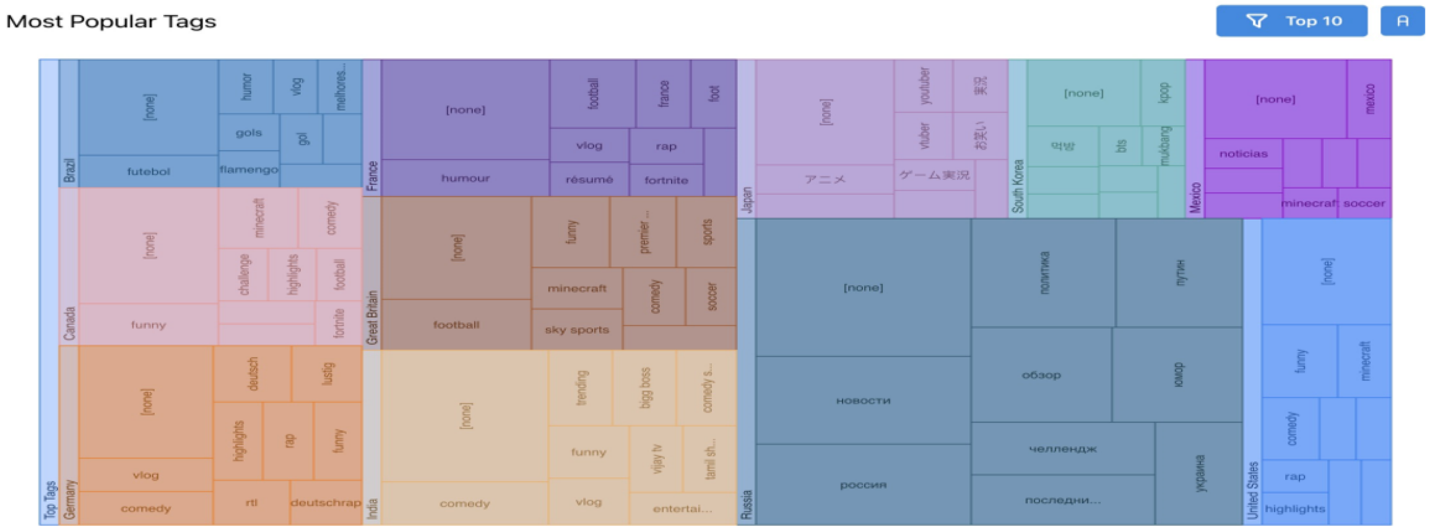

Figure 3

Treemap of most used tags across countries.

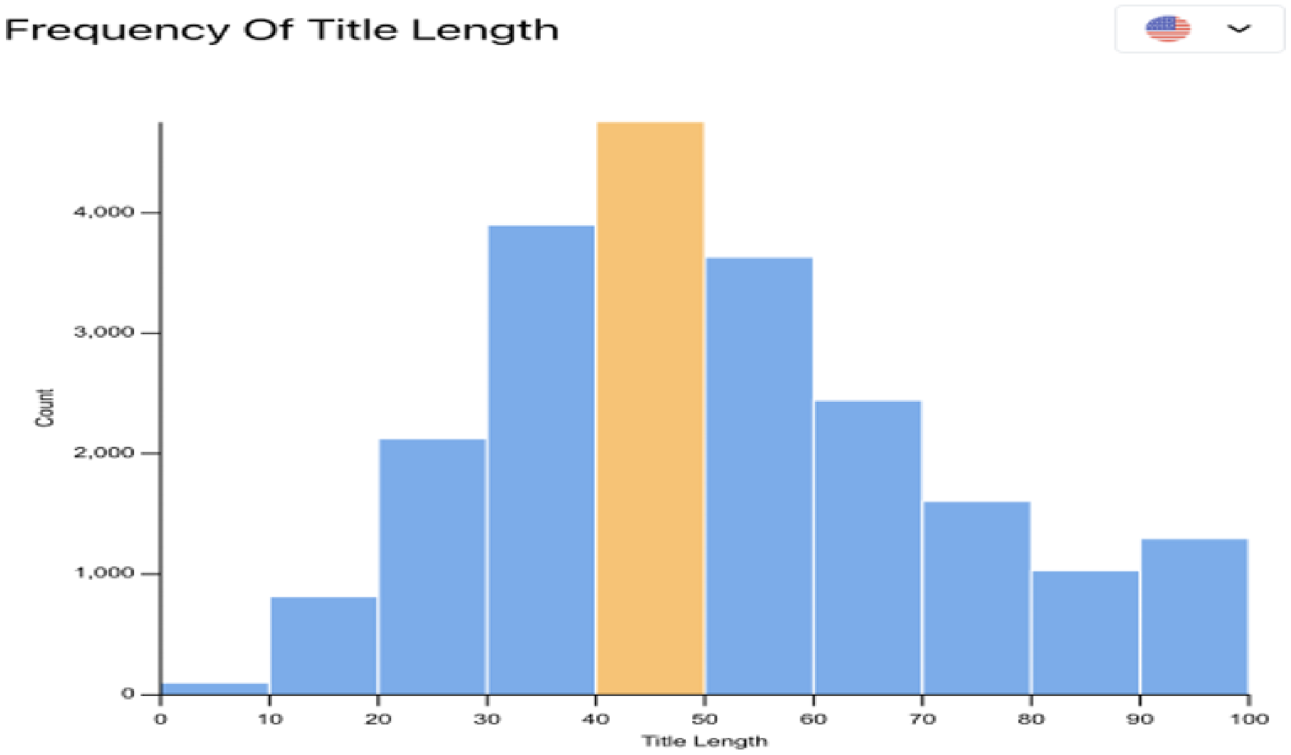

Figure 4

Histogram showing frequency distribution of title length for the trending video.

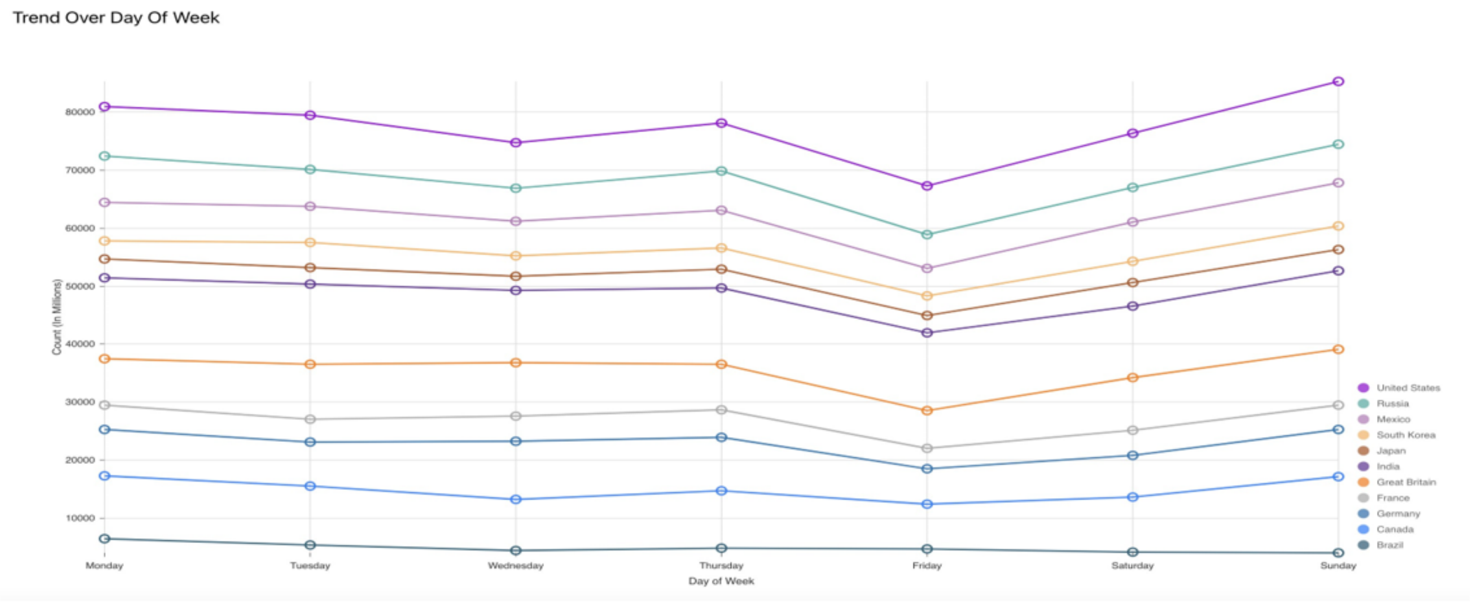

Figure 5

Line graph showing trend over the week with the total number of trending videos.