Societal and Cultural Nuances in YouTube Videos

Math 189 Group 7
Sebastian Diaz, Christine Law, Steven Liu,
Grace Murphy, Vishal Patel, Andrew Schmitz

Abstract

With social media having exponential growth in its involvement in everyone's life, it's valuable to discern patterns in trending topics across different countries. This holds significant importance in understanding societal interests, cultural nuances, and global perspectives in the digital age. By investigating the types of content that gain traction in various regions, we shed light on the prevalence of current events, beauty, vlogs, or educational content as well as unveil the underlying societal constraints and preferences shaping online consumption behaviors. More specifically, if countries push their citizens to maintain a conservative outlook on society by banning certain music videos or content that do not align with their values. Additionally, we aim to determine what's popular in a certain country and what types of content are the most likely to be able to cross language barriers. Ultimately, this project serves as a bridge between digital behavior and societal dynamics, fostering a deeper understanding of our interconnected world and differing cultures.

Statement of problem

With social media having exponential growth in its involvement in everyone's life, it's valuable to discern patterns in trending topics across different countries. This holds significant importance in understanding societal interests, cultural nuances, and global perspectives in the digital age. By investigating the types of content that gain traction in various regions, we shed light on the prevalence of current events, beauty, vlogs, or educational content as well as unveil the underlying societal constraints and preferences shaping online consumption behaviors. More specifically, if countries push their citizens to maintain a conservative outlook on society by banning certain music videos or content that do not align with their values. Additionally, we aim to determine what's popular in a certain country and what types of content are the most likely to be able to cross language barriers. Ultimately, this project serves as a bridge between digital behavior and societal dynamics, fostering a deeper understanding of our interconnected world and differing cultures.

Source of data

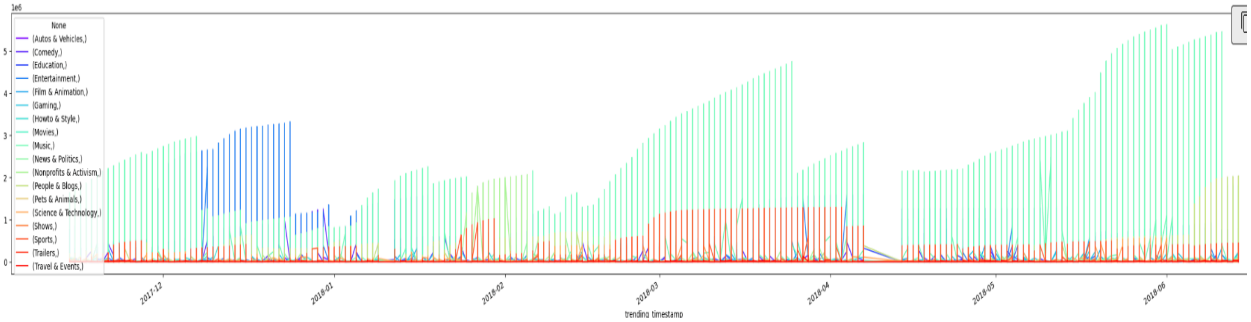
To better visualize how trends in media differ between different regions of the world, we will use the "Trending YouTube Video Statistics" dataset on Kaggle (https://www.kaggle.com/datasets/datasnaek/youtube-new/). This dataset was collected using the YouTube API, so we can trust that the metrics for the videos will be accurate at the point in time when the video was recorded into the dataset.

Description of data

More specifically, this dataset contains data on various countries such as Canada, Great Britain, Russia, the United States, and many more. Within each dataset, there is a JSON file that details the description of one of the variables: "category_id". This JSON file allowed our group to analyze the categorical variable to accurately depict the trends of popular genres of YouTube videos depending on the country. Furthermore, there are a plethora of variables such as the video ID, trending date, title, channel_title, publish time, tags, and the number of views/dislikes/likes which we were able to utilize when determining whether certain variables had any correlation with one another. Additionally, the size of this specific dataset allowed us to perform effective and accurate analysis with there being up to 200 entries per day, which provides us with a diverse pool of data expanding trends of popular YouTube videos past a single day.

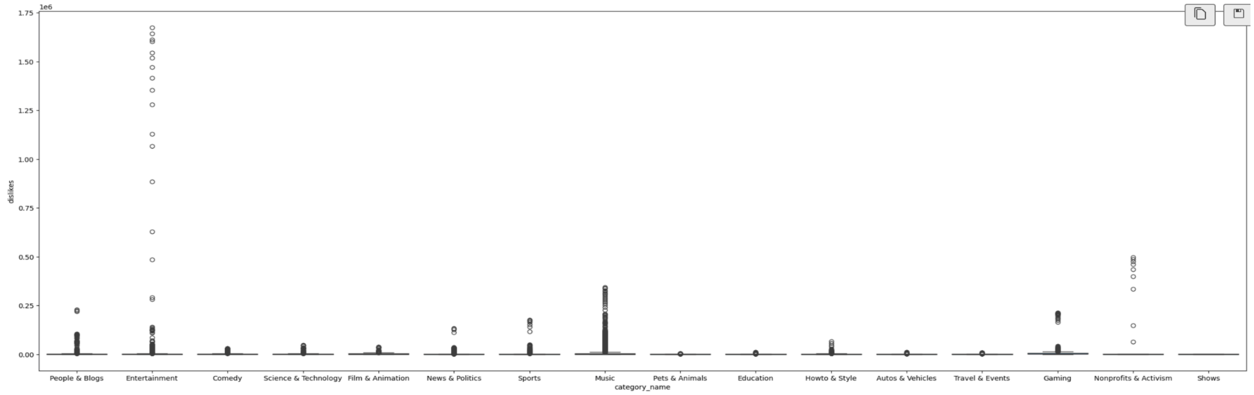
Exploratory Data Analysis

To get a baseline understanding of our data, we decided to compare the aggregate number of likes across each video category for every month between late 2017 and mid-2018. To do so, we produced the following graphic, which represents the number of likes per video (by category) on the y-axis and time (in days) on the x-axis. The "blue" and "green" data points represent likes given to videos in the "Entertainment" and "Music" categories, respectively.



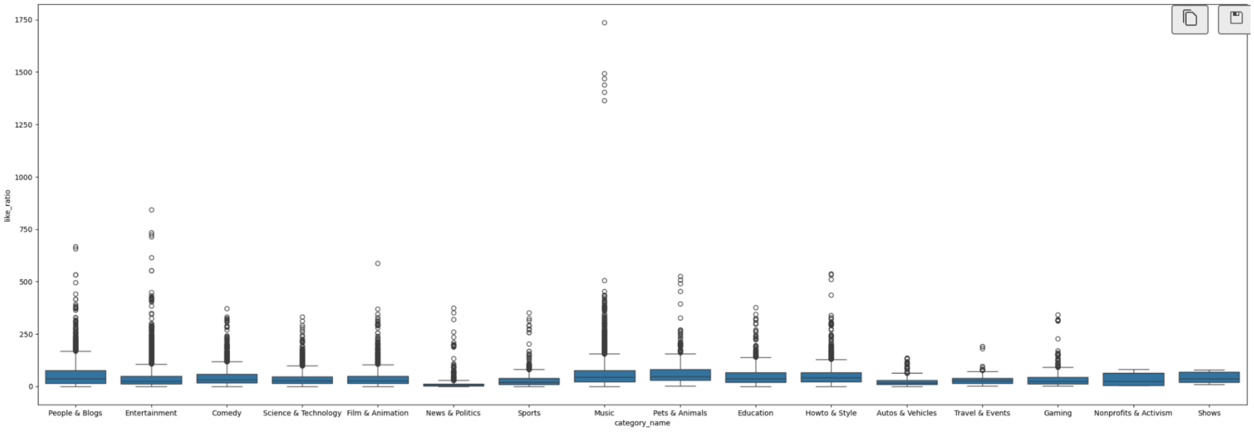
By observing the graphic, can conclude that the category with the greatest number of likes switched from Entertainment to Music sometime in early 2018, and Music would remain the most liked category for the remainder of the time horizon of the data we collected.

When analyzing the trending videos, it is of interest to gain knowledge of the public's general opinion of certain types of videos. This led us to plot a boxplot comparing the video's categories to the total number of dislikes. It is important to note that YouTube has since removed the dislikes counter across the entire platform. However, this change took place sometime in 2020-2021. As shown in the graphic above, our data was collected from late 2017 to June 2018 which means the the dislikes of a certain video were still available at the time.



The boxplot shows an overwhelming amount of outliers in the entertainment category, meaning that certain trending videos in that category may have sparked controversy amongst viewers therefore leading them to dislike the video. Another topic that appeared to cause controversy was the Nonprofit & Activism category. The Music category also had several outliers.

When analyzing the total of dislikes, it is of interest to compare the total of dislikes and likes between categories. For instance, we can speculate whether videos with the highest number of dislikes reflect the overall public sentiment of a certain category within the dataset. This can be done by analyzing the likes:dislikes ratio across videos in our dataset, which we did using the following boxplot.



Unlike the dislikes boxplot, the outliers between categories are much more evenly dispersed except between Autos & Vehicles, Travel & Events, Nonprofit & Activism, and Shows. Note that despite the number of outliers observed in the Nonprofit & Activism regarding the total number of dislikes, it appears from the plot above that those dislikes were balanced out by the total number of likes. Thus there are no outliers observed for the category above, potentially reflecting a general split of public opinion. Music appears to have the highest likes:dislikes ratio of all the categories.

Previous Work

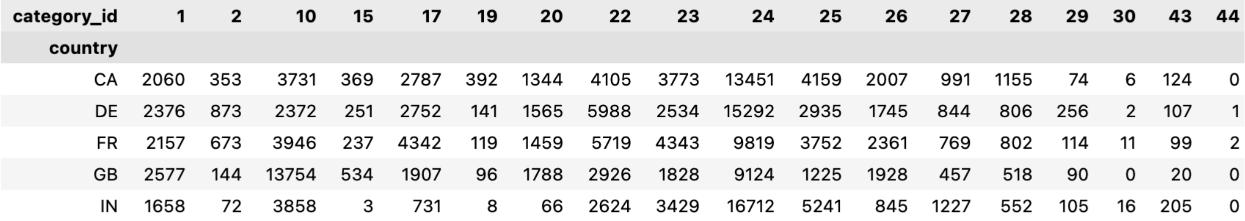
Our dataset was found on Kaggle which provides us with all the notebooks that other people have created and published to their website. Our specific dataset contained a total of about three thousand notebooks. After browsing through the notebooks we came to find that a majority of the notebooks were notebooks created for a python and data science tutorial. We searched for and found a couple notebooks with actual analyses of the dataset. The two that stood out the most were Analysis of YouTube Trending Videos of 2019 (US) and **DEEP ANALYSIS on Youtube Trending Videos - EDA, as most other notebooks contained the same or similar analyses.

The first notebook is likely the most in depth one we found. The writer of this notebook focuses on the US part of the dataset rather than all of or some of the countries. They also, more specifically, analyzed the trending videos just from the year 2019. They found the trending videos which stayed trending for the longest, which was found to be 30 days. An analysis on the titles of these trending videos was done to find patterns in the words used, capitalization, and the length of the titles, in order to find if there was any correlation between the title of the video and the video's success. They also analyzed the channels who uploaded the trending videos to find which channels succeeded in creating the most videos to reach the trending tab, when the videos were published, the tags on the videos, and the amount of comments, likes, dislikes, as well as how all these aspects of the video correlated with each other. The other notebook we looked at analyzed similar attributes of trending videos as the first notebook, with the difference being that this notebook looks at all the countries within the dataset rather than just the US.

These weren't the only two notebooks that we looked at, but just the most in depth ones. We looked through many notebooks in order to get a good understanding of what exactly has already been analyzed and how we'd be able to look at this dataset from a new perspective. So, although this dataset contained a lot of notebooks published by other users, a majority of them were for a data science tutorial and the rest analyzed very similar things to one another, which allowed us a lot of room to come up with a new way of analyzing the dataset.

Analysis

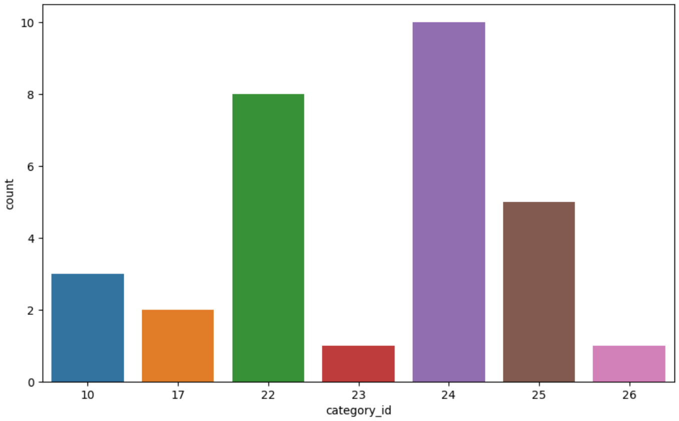
With our dataset, we performed various hypothesis tests to get a more in-depth understanding of the nature of category IDs and their regions. The first hypothesis test we did was to see if the total distribution of category IDs by region has a statistically significant difference. For this, we made the null hypothesis: There is no significant difference in the distribution of popular YouTube categories across different countries, and the alternative hypothesis: there is a significant difference in the distribution of popular YouTube categories across different countries. Both of these features, category id, and country, are both categorical variables, so we used the Pearson chi-squared test. To use this test you need first to make a contingency table. This table has the index as the countries and the columns as the category IDs. This creates a table where the i, jth value of the table corresponds to the ith country and its total video count that has the jth category. To make the table we used the pandas crosstab() method and assigned this to a variable, cont_df. Here is a snapshot of the contingency table.



category_id	1	2	10	15	17	19	20	22	23	24	25	26	27	28	29	30	43	44
country																		
CA	2060	353	3731	369	2787	392	1344	4105	3773	13451	4159	2007	991	1155	74	6	124	0
DE	2376	873	2372	251	2752	141	1565	5988	2534	15292	2935	1745	844	806	256	2	107	1
FR	2157	673	3946	237	4342	119	1459	5719	4343	9819	3752	2361	769	802	114	11	99	2
GB	2577	144	13754	534	1907	96	1788	2926	1828	9124	1225	1928	457	518	90	0	20	0
IN	1658	72	3858	3	731	8	66	2624	3429	16712	5241	845	1227	552	105	16	205	0

Now that we had the data prepared for the test, we imported scipy.stats as stats to be able to run the test. Using Scipy's chi2_contingency() method, we inputted our cont_df variable into the function which resulted in the test statistic and p-value. The test statistic was 88279.61 and it has a p-value of 0.0. Since the p-value is so low, we can reject the null hypothesis that there is no significant difference between region and category ID, at the significance level of 5%, and may conclude that there may occur a statistically significant difference between the countries and category IDs. After concluding this hypothesis test, we also decided to compute a different test to see a different type of distribution of categories and regions.

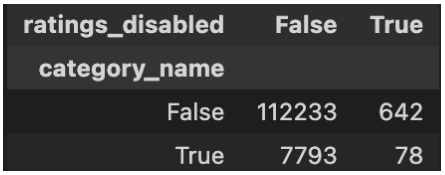
Our next hypothesis test tested to see if there is an underlying statistically significant difference between the top 3 category IDs per country. To do this we stated the null hypothesis to be there is no difference between the top 3 genres of popular YouTube videos in English-speaking countries, and the alternative to be that there is a noticeable difference between the top 3 category id of popular YouTube videos in English-speaking countries. Since this hypothesis test requires us to only include the top 3 category IDs per country, we first need to extract these categories. To do so it uses the pandas groupby() method to be able to group the data by 'country' and 'coutnry_id' with the agg function being size() to get a count for each category and country. We then used the pandas rank method to be able to assign a "rank" or a within-country value that determines the ranking of size where 1 is the largest. After assigning a rank to all the columns, we can filter out all the columns to only include rows with ranks from [1, 3] (inclusive), and create a new panda data frame only including the 'country' and 'category id'. To visualize the distribution we plotted on Seaborn's countplot, a distribution of the categories and their respective total frequency of being the top three for their country.



Here you can see that category ID 24, comes up in the top three category IDs the most, while category id like 23 and 26 comes up the least frequently. After preparing all the data, we can now use the new data frame we created to make a contingency table. Since, like the previous hypothesis test, we are comparing the distribution of two categorical variables, we can use the person's chi-square test, which requires a contingency table. Just like last time, we use the crosstab() function from pandas and assign "country" to be our index and "category_id" to be our columns. We then put it into our scipy.stats function, chi2_contingency, with our contingency table as the parameter. This then gives us the statistical value and p-value of 40 and 0,92, respectively. This tells us that since the p-value is a high number, we fail to reject the null hypothesis that there is no statistically significant difference between the top 3 genres of popular YouTube videos in English-speaking countries, at a 5% significance level.

Now that we have the results of our first two hypothesis tests, we then wanted to determine whether or not there is a correlation between more controversial topics and censorship. With this idea in mind, we came up with three hypothesis tests to determine if “News and Politics” content is more likely to have their ratings disabled, comments disabled, and/or be removed from the platform entirely.

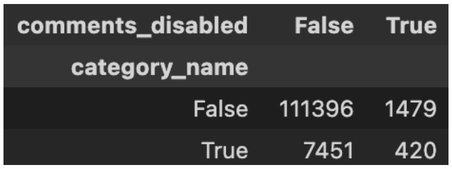
The first test conducted had the following null hypothesis: There is no correlation between a politically centric video and the likes/dislikes ratio of the video eventually being disabled in countries where the primary language is English; and the alternative hypothesis: There is a correlation between a Political video and the likes/dislikes ratio of the video eventually being disabled in countries where the primary language is English. Constructing a contingency table using pandas’ crosstab method very similar to the one above comparing videos tagged under the news and politics category and ratings disabled resulted in:



ratings_disabled	False	True
category_name		
False	112233	642
True	7793	78

Then, to run our test we used the scipy.stats method chi2_contingency and obtained a p-value of 0.00000369 which is small enough to reject the null hypothesis in favor of the alternative.

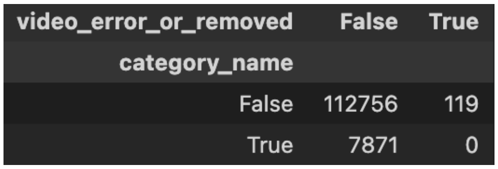
Next, we checked if Political content was more likely to have the comments section disabled. For this test, our null hypothesis was: There is no correlation between a Political video and the comments being disabled in countries where the primary language is English, and our alternative hypothesis was: There is a correlation between a politically centric video and the comments being disabled in countries where the primary language is English. As with the last test, we constructed a contingency table comparing videos in the news and politics category to having a disabled comment section:



comments_disabled	False	True
category_name		
False	111396	1479
True	7451	420

Then, we ran a chi2_continency method which gave us a p-value of 5.58x10^(-169). The p-value is small enough to reject the null hypothesis in favor of the alternative. This means that we can conclude that videos under the 'News and Politics' are more statistically likely to have disabled comments.

Finally, we checked if the videos were more likely to be removed entirely from YouTube in comparison to the other categories. The null hypothesis for this test was: There is no correlation between a politically centric video and the video eventually being taken down in countries where the primary language is English, and the alternative hypothesis was: There is a correlation between a Political video and the video eventually being taken down in countries where the primary language is English. Upon constructing the contingency table we found that no Political videos in our dataset were taken off the platform:



video_error_or_removed	False	True
category_name		
False	112756	119
True	7871	0

x

So, it is no surprise that running the chi2_contingency resulted in a p-value of ~0.007, which is low enough to reject the null hypothesis.

*Note: In our video submission, we misspeak by concluding that the above p-value is not small enough to reject the null hypothesis

Finally, we chose to build a set of binary logistic regression models to try and predict the country that a video was trending in. We combined sets of data from pairs of countries, with one always being the US dataset to act as a standard for comparison, and observed how well these logistic regression models could fit to the data. The independent variables we used from the dataset to act as predictors are views, trending_timestamp, comments_disabled, and ratings_disabled. We avoided including some of the other variables to prevent multicollinearity. While in some of these models, the addition of certain variables were not statistically significant, we chose to keep them in the models as a way to compare them against each other. Additionally, there were issues of producing a singular matrix during the process, so we chose to remove data whose category is among the top in the combined dataset to try and resolve this issue. Thus, we could create a set of regression models that attempt to differentiate between videos trending in the US and videos trending in other countries.

Interpretation

Our first two hypothesis tests showed conflicting results where there may be a statistically significant difference between category IDs and country, and there is no statistically significant difference between the top three category IDs and country for English-speaking countries. One interpretation that can be made from this is that since the top 3 category IDs are amongst the most popular for that country, there is a high chance that these topics are also popular for other similar countries. In our hypothesis test, the countries we grouped were English-speaking countries that may share similar cultures and mindsets compared to other non-English-speaking countries. So, since all the English-speaking countries share the same overall ideas, that may be a reason why there is no significant difference for the "popular" categories. But, when we measure the total distribution for all categories and countries we see that there is a significant difference. This may be because we are also including the nonpopular among all countries categories in the test. What we mean by this is that our first test includes categories that may only be popular for that one country and nowhere else, due to a numerous number of factors such as region, culture, country, etc. So since the test also includes these categories we saw a strong possibility that there may be a significant difference present between the countries and categories.

The next three hypothesis tests demonstrated that videos featuring political content are more likely to experience censorship through the removal of ratings and comments. Ratings and comments are a good way for viewers to determine public sentiment on the videos they watch, but interestingly, political videos are more likely to have them disabled. This could be an indication that news publications want their viewers to develop their own thoughts and opinions about certain topics and may see reactions like comments as a means of forming a bias amongst users. YouTube itself can also disable the comments section when there are too many comments that break community guidelines, so maybe comments under videos surrounding politics are more likely to violate those guidelines. This is largely speculatory because our dataset does not give us any indication of when these disablements took place or for what reason.

Regarding the results of the models, we were able to gain a range of accuracies for the different pairings of data with the US dataset. From this, we can conclude that certain countries have more similarities to the US than others in regards to the kinds of videos that trend and become popular. For example, while the model predicting on the Canada-US dataset has an accuracy of about 0.62, the model predicting on the Russia-US dataset has an accuracy of about 0.78, which is better than the first model. Similarly, the Japan-US dataset's model produces an accuracy of about 0.88, which is significantly better than the previous two models. Thus, we can identify which countries are more similar to the US in watching habits on YouTube, and which kinds of videos are more prevalent in each country by looking at the model coefficients.

Conclusion and discussion for future work

We comprehensively analyzed our dataset to make significant analyses and discoveries on how content is consumed and pushed out across the world. It was very interesting to our team to see how various regelations and cultures impact the content that is viewed across the world, especially when it came down to controversial material such as political content. By using our hypothesis tests, we were able to determine that there is a significant impact on the way that content is shown (whether or not comments are disabled) based on the category and region of the video. Our model expanded on this reasoning by determining the impact of region on the category of videos shown. Although we got conflicted results on the top categories in countries, we were able to determine that there was an impact on whether or not comments were shown depending on the region.

Our project provided an interesting view into the nature of social media and its distribution. We'd be interested in expanding this project to other platforms, especially short-form platforms such as YouTube Shorts and TikTok. Another way to extend our research is by looking at the content of the videos themselves by analyzing transcripts or frames from the videos themselves. Ultimately, we were impressed by the research we were able to complete on the distribution of content around the world.