What Should You Know Before Becoming A YouTuber?

Yukang Zhao, Yize Liu

Summary of Questions and Results

1. What are the most popular types of content on YouTube, and how have they evolved over time?
    - This helps us understand what people liked in the past and what people like for now. We may use this research to predict what people will like in the future.
    - Through our analysis, Music, Entertainment, and Gaming have been a staple of what people watch online. In particular, the volume of music videos watched in various countries has always been high and has continued to grow in recent years.
2. What kind of Youtubers could be possibly more successful in the future?
    - This is slightly different depending on which country you are in. However, in general, whether for the sake of incomes and total views of videos, music videos dominate through all categories.
    - For the US region, there are also some other good options, for instance, Comedy, How to & Style, Blogs, and Gaming. First three categories have pretty good data on their average like-view ratio, which exceeds 4%, especially the like-view ratio of How to & Style is very close to that of Music. For Gaming, although the average like-view ratio is about 3.73%, since its total views is the highest besides Music, Gaming should also be a good option to get in.
    - In terms of what profits can YouTubers make in each category, it turns out that 'Music' has the highest profits in total but 'Shows' has the highest profits individually. We found out the total revenues for the top 500 YouTubers in 2021 are $2.8 billion and Music made $1.3 billion, nearly a half ! In order to remove the influence of quality and volume, we further found out that 'Shows' has a highest mean profit which is about $10 million and 'Music' has a second highest value of $6.2 million.
3. How do YouTube users differ by favors in different countries?
    - We mainly did the research for the USA, Canada, France, and India, but further research for other countries can be done by the same code.

- Here are top 3 categories in term of total views (in order):
  - USA: Music Film & Animation Non-Profit & Activism
  - Canada: Music Movies Film & Animation
  - France: Music Sci & Tech Shows
  - India: Gaming Movies Music
- Top 3 categories in term of like-view ratio (in order):
  - USA: Music How to & Style Comedy
  - Canada: Comedy How to & Style Gaming
  - France: Pet & Animal Education Science & Tech
  - India: Sci & Tech Comedy Pet & Animal

Motivation

- YouTube has become a staple in the world of online video sharing, and has grown to become the largest platform for video content. According to the statistics, YouTube has 2.6 billion active monthly users and around half of the internet users in the world have access to YouTube. For a world trending video platform, it has already made extraordinary contributions to connect people together. YouTubers can upload videos and they can range from individuals filming themselves with their smartphones to large media companies producing high-quality content. With the abundance of videos being shared and watched on this platform, it is important to analyze the trends and preferences of the audience as well as the YouTubers to gain insight into the evolving online video consumption landscape. By analyzing both the audience preferences and the YouTubers, we could give suggestions like what steps you should take to become a successful YouTuber.

Dataset

1. https://www.kaggle.com/datasets/themrityunjaypathak/most-subscribed-1000-youtube-channels
    - This dataset is provided by kaggle and the data can be downloaded as a csv file.
    - Record top 1000 the most subscribed YouTube channels.
    - The data contains the YouTube channel names, subscribers, video views, video counts, category and started years. We will focus on using each statistic by its category and started years.

2. https://www.kaggle.com/datasets/datasnaek/youtube-new?select=USvideos.csv
    - For instance, using the USvideos.csv, which includes several months of data on daily trending YouTube videos in the USA region. It is provided by kaggle and can be downloaded as a csv file.
    - It is important to note that each csv file also has a corresponding json file which requires us to read in a json file to merge them. The category data is inside the json file.
    - This dataset also includes similar contents as USvideos.csv but in other regions.
    - In this project, we pick four distinct countries to analyze: USA, France, Canada, and India.
3. https://us.youtubers.me/global/all/the-highest-paid-youtubers-of-2021
    - Contains a collection of income datasets of the highest-paid youtubers in 2021
    - It can be read in as a html object and then converted to a dataframe. It is important to note that this dataset includes not only individual YouTubers but also official channels. The income is collected only from YouTube not including other commercial activities.

Method

Research Question 1

Question: What are the most popular types of content on YouTube, and how have they evolved over time?

1) In the dataset 'most subscribed 1000 YouTube channels'. First, drop the rows if video views are 0 and category is invalid. Category is invalid here and represents those columns with unclear values.
    a) Converting video views to int type. This requires us to loop through each element in these columns and replace punctuations.
    b) Grouping by the total views of each category and compute the sum.
    c) Creating a plot of total views vs category.
2) Next, we used the year column and picked 5 distinctive categories based on the previous result.
    a) Masking the DataFrame to only include 'Music', 'People & Blogs', 'Gaming', 'Entertainment' and 'Education'.
    b) Calculating how many subscribers increased per video through 2009-2019. Taking a log of the number so each value can be normalized.
    c) Using a list of dictionaries to store all the results.
    d) Plotting a line graph that gives us the changes through years.

Research Question 2

Question: What kind of Youtubers could be possibly more successful in the future?

1) Firstly we need to clean and arrange the data of the Dataset #2:
    a) The data type of each element in columns 'publish_time' and 'trending_date' is a weird string, we use the string slicing and astype to exact the day, month, and year of both categories and turn them into integers to compute the time from publish to trending.
    b) The categories of each video in the data was not straightly stated, instead it has an id for each category, the id and corresponding category are recorded in another .json file. We convert the json file to dataframe, and then merge it with the data file by the id number.

c) We only preserve some ‘useful’ columns. Compute the like-view ratio and time from publish to trending for each video and add this as a new column to the initial DataFrame.

d) Draw the plots of total views and like-view ratio for each category, also plot of the relationship between days from publish to trending and like_view ratio by using Seaborn.

2) Using the 3rd dataset to determine the income of top subscribed YouTubers and compare them.
    a) Reading in the html file to dataframes. The column names for the second table are given as integer values so we need to give them the right string names.
    b) Combining two tables together and converting ‘income’ and ‘income / subscribers’ to float data type.
    c) Plotting top 10 income and income / subscribers YouTubers.
    d) Categorizing the income by categories and plot a pie chart to see how much portion each category takes up.
    e) Deleting the maximum and minimum values for each category so the mean value won’t be biased.
    f) Plotting a pie chart to see which category makes the most profits.

Research Question 3

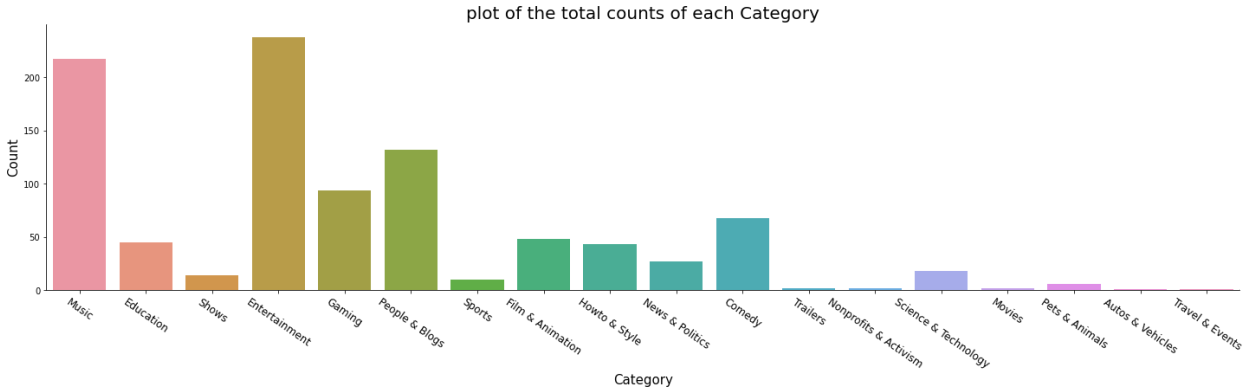
Question: How do YouTube users differ by favors in different countries?

1) We used four different countries: USA, France, Canada and India. Using the same methods above, we created 4 DataFrames of each country.
2) Combine these countries according to categories.
3) Calculate the mean of views and mean of like-view ratio by category.
4) Generate 4 bar plots for each country in 4 subplots for views; generate a comparison bar plot for like-view ratio comparing differences between countries by category.

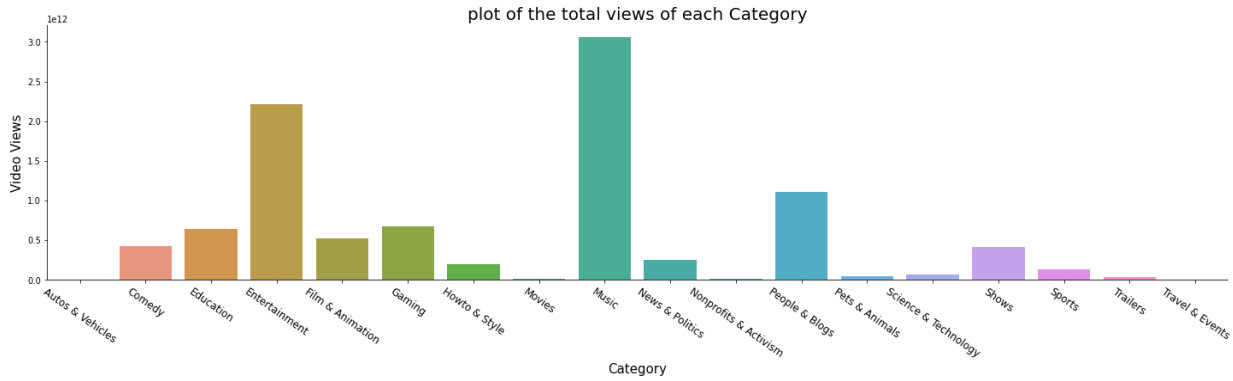
Results

Research Question 1:

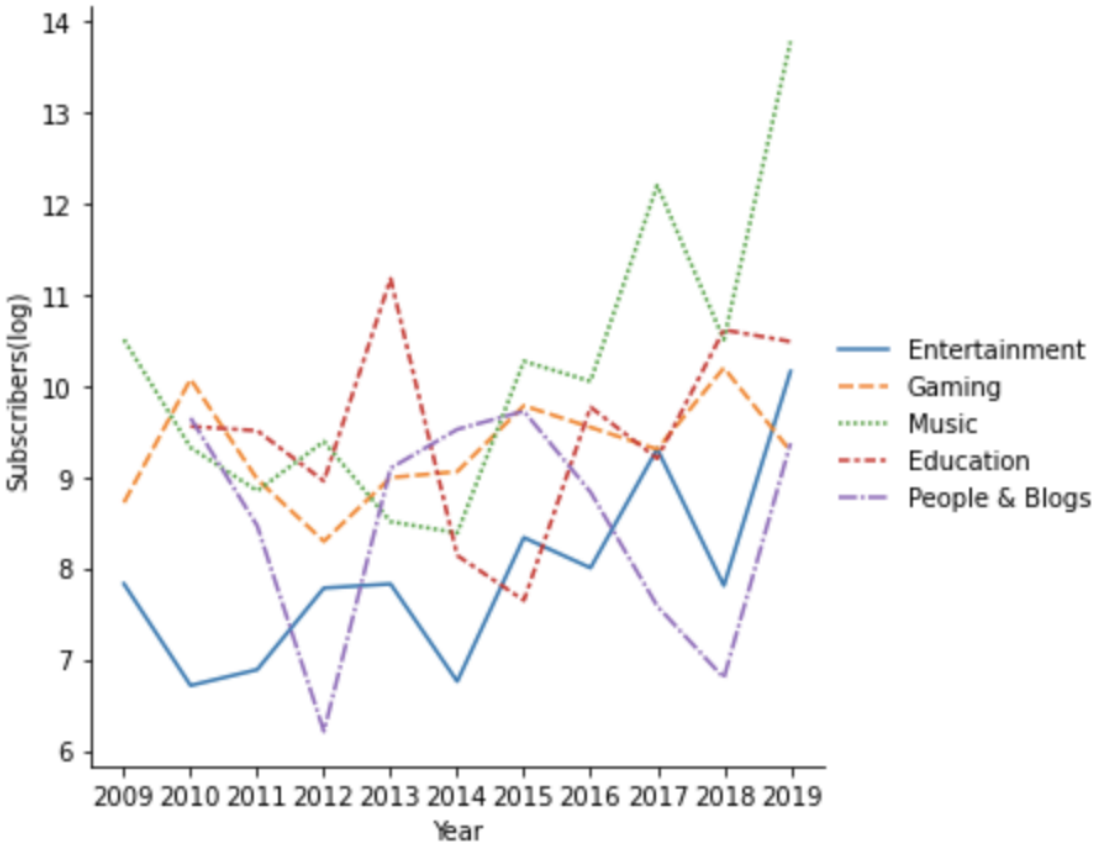
To understand what are the most popular types on YouTube, we chose the dataset ‘Most 1000 Subscribed YouTube Channels’ because it is the most representative. Furthermore, to include all YouTube videos and analyze them may take too long because the input data will be exceptionally large. We believe that the top 1000 YouTube channels can represent most viewers' favors.



Based on the above graph, most viewers tend to subscribe to channels like: Music, Entertainment, Gaming and People & Blogs. We can see Music and Entertainment took a portion of nearly 50 percent of the total top subscribed channels.



Next, we did a calculation of total views by each category. The above result justifies what we did in the previous graph.Based on our initial understanding of users' viewing preferences, we wanted to dig deeper into the relationship between subscriptions and categories.

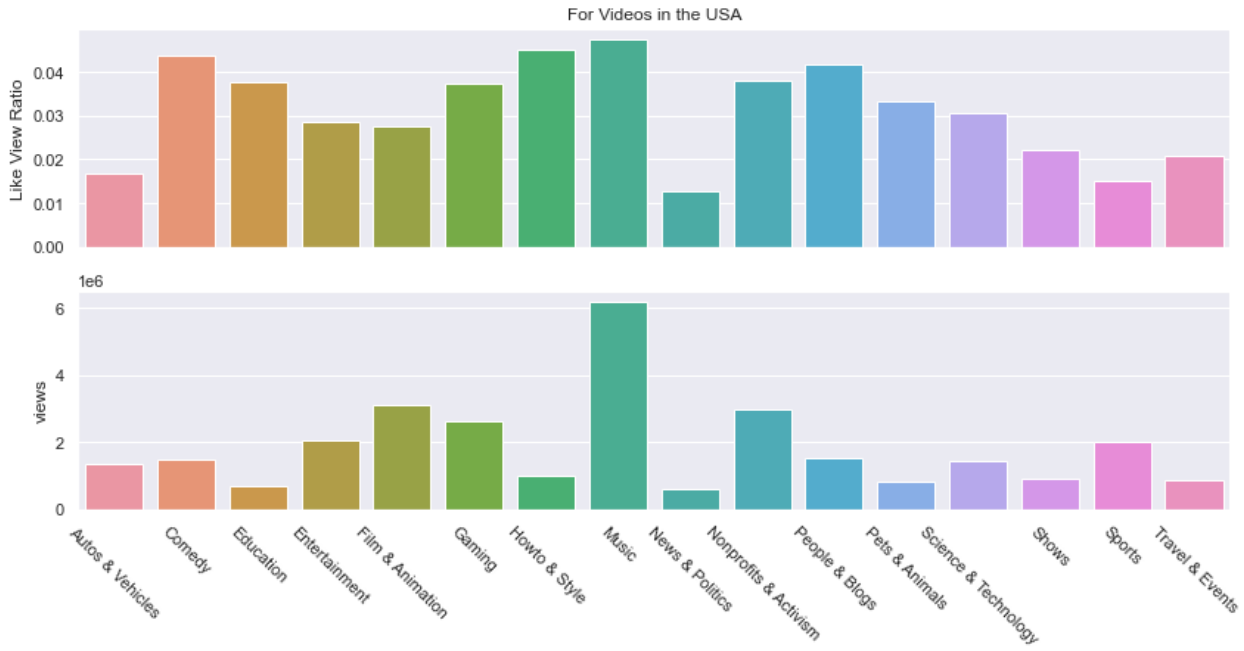


The above plot gives us how many subscriptions increased per video in each category from 2009-2019. Although the relationship is a bit ambiguous from this plot, the music category has always been the most popular both in the past and present and the most obvious increase in volume is the entertainment category. However, People&Blogs and Education which used to be popular but have declined in recent years. One thing we notice here is that the total views in 2019 of the Music category(14,400,000) is much less than categories like Entertainment(117,000,000), Education(39,200,000) and People & Blogs(85,600,000). The reason behind this is the new YouTuber:Kim Loaiza who only published 15 videos but received 14 million fans.

Research Question 2:

In order to find out what kinds of YouTube videos may have more probability to succeed, we choose to analyze the total views and like-view ratio of each category of videos for each country; which total views represents the size of population of audiences, like-view ratio represents how much people like the video, the higher the average data of those, the higher the probability a video from those categories may “succeed”.

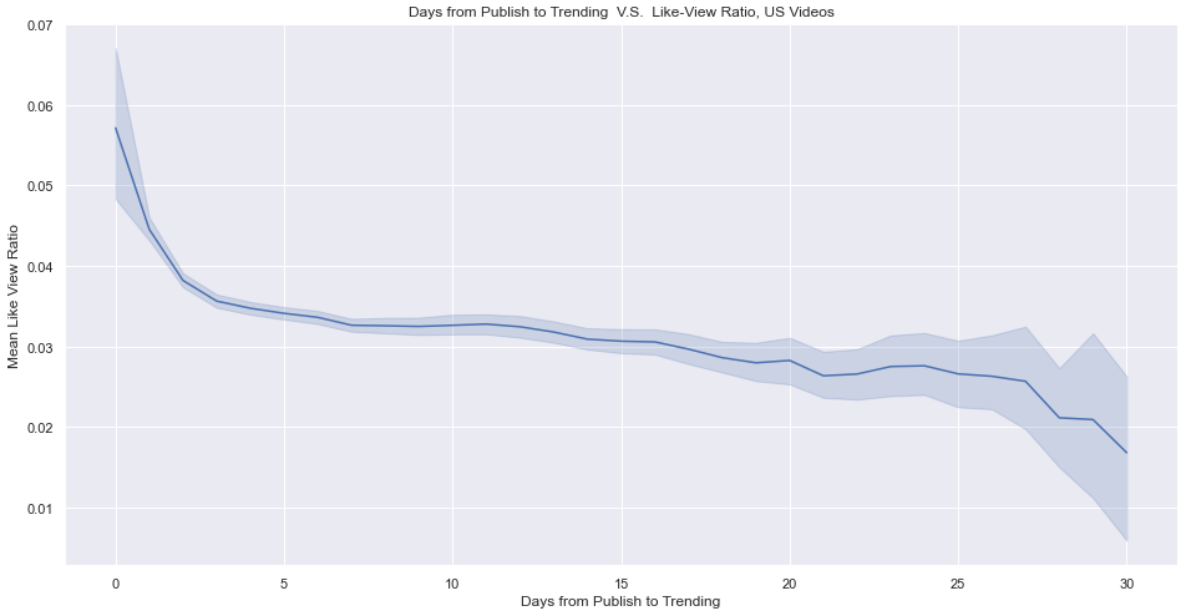
Here is a figure of total views and like-view ratio of each category in the US.



Figures of other 3 countries are in Question 3 section.

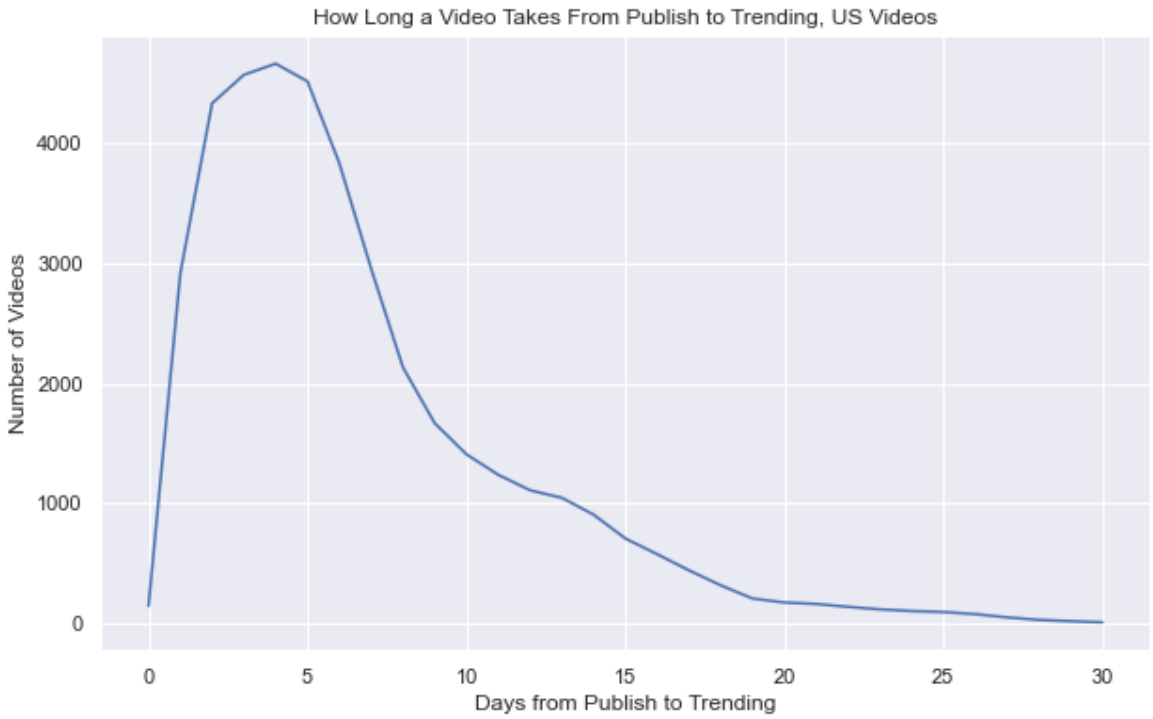
If we want to know what people like the most, or the "fan stickiness" in general, it is a good choice to look at and compare the like-view ratio of each category. The above figure of like-view ratio at the top clearly stated that people in the US favor music, How to & Style, Comedy, and Blogs a lot; Gaming is also a good choice, since even though it has not such outstanding average like-view ratio (i.e., above 4%), it has a pretty good average view. We can see Film and Nonprofits & Activism also have good statistics, but it is probably hard for normal people to get in those categories. In comparison, we can also see some categories that probably are not good choices, for instance, News & Politics and Travel & Events.

We also generated a plot of Days from Publish to Trending V.S. Like-View Ratio. The plot below clearly shows the relation between those two features, which the longer a video takes from publish to trending, the lower like-view ratio it tends to have. This suggests that we may want to make our title, cover of the video, and first tens of seconds as interesting as possible to attract the audiences to make the video trending faster, or rather the longer that process takes, the lower the probability for a video to be more outstanding.



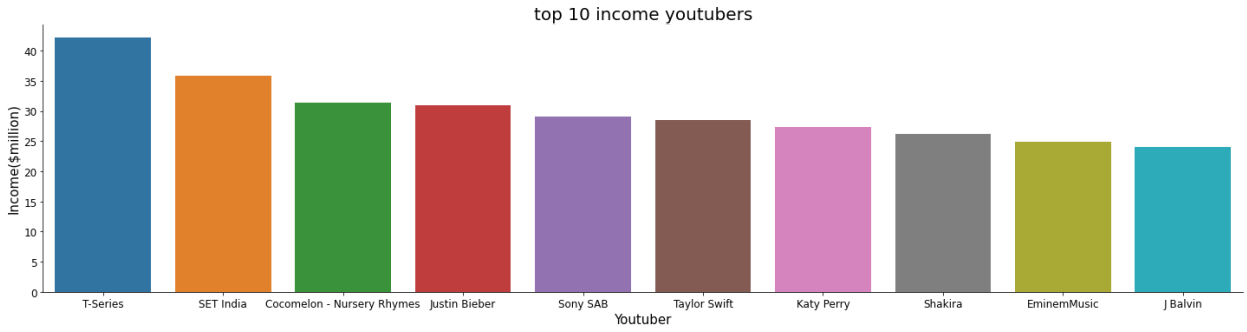
Note: There exist videos that have Days from Publish to Trending more than 30 days, but only very few of them (about 0.9%), hence we consider them as outliers and do not plot them.

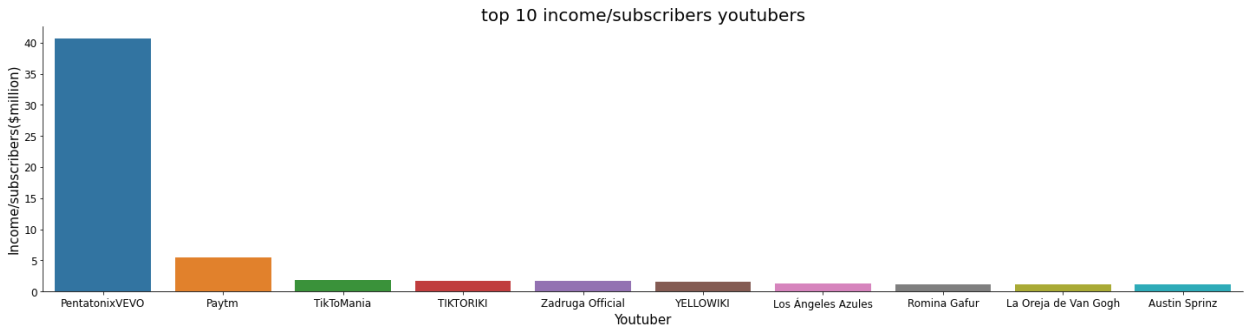
Below is a plot record number of videos sorted by Days from Publish to Trending, for videos in the US. This shows an outstanding video becomes trending very soon after publication; chances of being trending get lower and lower as time goes after a few days of publication.



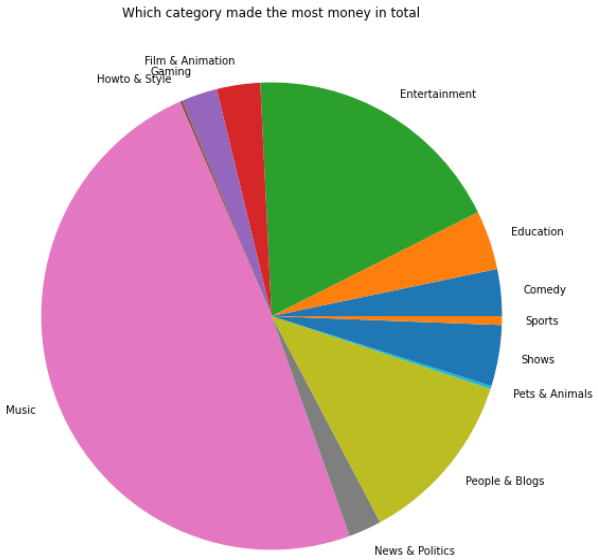
PS: This roughly looks like a Poisson Distribution with a mean of 5.

In addition, we use the 3rd dataset to determine the income of top subscribed YouTubers and compare them.

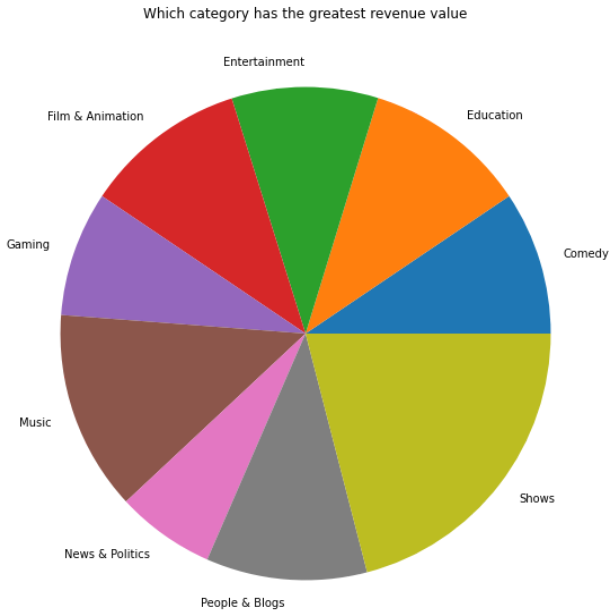




Here are two plots of the top 10 income distributions of YouTubers in 2021. We can clearly see that Music category YouTubers make up the vast majority. To further investigate how other categories look like, we did a pie chart in the following:



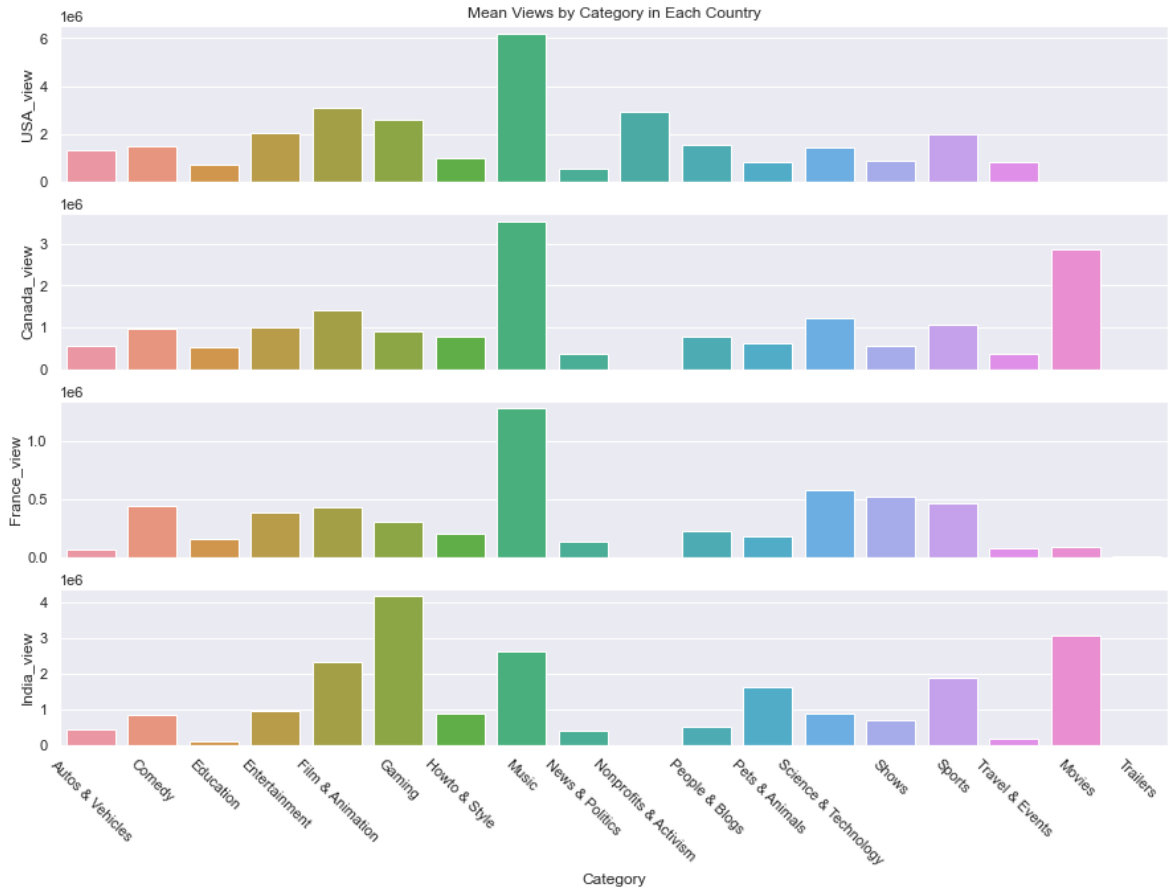
Music category total income takes up nearly a half in this pie chart, but the reason behind this may be because the total count of Music category is large. Therefore, we need to consider the mean value of the income for each category.



This gives us a surprising result that the Music category no longer takes the lead. ‘Shows’ becomes the top first category in our visualization. However, this category is usually set up by different big companies or organizations. It is difficult for individuals without top-notch equipment and professional teams to make ‘good shows’. Therefore, individual YouTubers should choose to do People & Blogs or Education which could be more feasible.

Research Question 3:

To see the differences of tastes for those four countries, we decided to compare the mean views and mean like-view ratio by category.

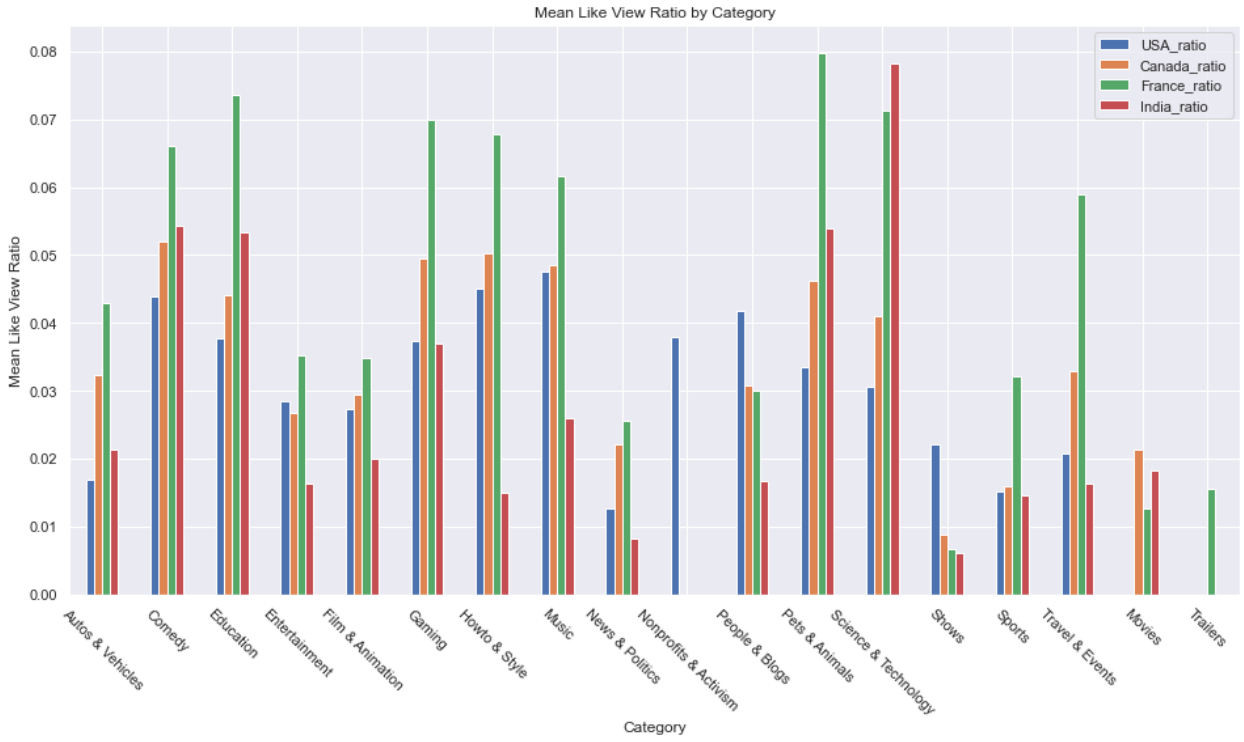


Notice the scales of y-axis for each country are different, this plot gives us a sense of the proportion of the population in each country that watches videos of each category, i.e., within a country, which categories are relatively favored more. This is reasonable, since countries vary in size and population, it is not fair to compare them by overall numbers. There are also some categories that are not included by some countries.

Above is the figure of mean views by category for four countries. From the figure, we can see that Music is loved "universally", which dominates in three out of four, being the second highest in India.

- Here are top 3 categories in term of total views (in order):
  - USA: Music Film & Animation Non-Profit & Activism
  - Canada: Music Movies Film & Animation
  - France: Music Sci & Tech Shows
  - India: Gaming Movies Music

Something interesting is that Movies have only a few views in France.



Above is the plot of comparison of like-view ratio by category between countries.

- Top 3 categories in term of like-view ratio (in order):
  - USA: Music How to & Style Comedy
  - Canada: Comedy How to & Style Gaming
  - France: Pet & Animal Education Science & Tech
  - India: Sci & Tech Comedy Pet & Animal

We can see that even though mean views of some categories, e.g. Pet & Animal in France, are not outstanding, they may have an impressive like-view ratio, which can represent the fan stickiness. Also an interesting fact is that it seems French people are more willing to hit the "like" button overall, but for the third most viewed category, Shows, the like-view ratio is very low.

Impact and Limitations

Our result may serve as a reference and inspiration for people who are considering becoming a YouTuber. Also, people may use our results to learn more about what is trending and make implications about what will be trending in the future. However, due to the recommendation algorithm used by YouTube, the data may be influenced by the popularity bias, making the "popular" videos more popular. Another limitation is that as the time goes, the data will be outdated and have less and less reference value. For Top 1000 Subscribed data, since it only contains those top video channels, the data may not be so suitable for beginners.

One potential misunderstanding is that our result is the analysis for the entire population of a country as a whole, rather than for a certain person. For instance, people should not look at the graph and claim that an Indian person must like Gaming more than Education.

Challenge Goals

1. **Multiple Datasets:** Based on the complexity of our project, we have selected multiple datasets for our analysis. Each dataset has different columns and rows. In order to enrich our research goals we have to combine tables together. This involves creating new columns, and using methods like groupby and join. For exploring the factors affecting watching, we need to use the category column. However, the US videos csv file doesn't explicitly label the category for each video and instead gives an id for it. That requires us to combine the US_category_id.csv with the original csv file.

2. **Messy Data:** We encountered many obstacles while doing this project and one of these is that many files are not presented as a csv file. In our second dataset, it has category columns in a json file and we need to import the json library to handle this data. Fortunately, pandas has functions that can tell us how to normalize json files so that we can get a nicer dataframe after processing them. For the third dataset, it is a table of ranking on a website that does not provide any file for people to download. However, pandas has ways to read a html to a DataFrame. The complexity of doing that is just based on how messy the webpage is. We had a few advertisements on that webpage so after reading the html we still need to do some extra work to get it done.

Work Plan Evaluation

- In general, we worked together throughout the whole project. Yize processed the dataset 1 and 3, did the research question 1; Yukang processed the dataset 2, did the research question 3; we combined our works to answer research question 2.
- Get the data set up
    - About 3 hours.
    - Find the dataset we will be using, filter out the columns that we will not be using, and turn the data to usable format.
    - This takes slightly longer than we expected, since we find cleaning data is much more difficult than what we thought.
- Categorize the data
    - About 2 hours.
    - Use methods like groupby to group the videos by their categories, published time, countries, and etc.. After the above categorization, sort the data into different classifications of comparing, i.e., comparison of popular videos between countries, popular videos between years, popularity between categories.
    - This also takes a little longer to think of how to choose the ways of comparing.
- Get the plots
    - About 3 hours.
    - Use libraries matplotlib.pyplot and seaborn to plot the graphs. Draw the plots of views, like-view ratio, days from publish to trending, popular categories over years, and the comparison of four countries and so on.
    - Time spending is approximately as expected, since getting the plots after having the data became much easier.
- Analyze the result and write the report
    - About 7 hours.
    - Look at the final result generated by the codes and figures, convert the numbers and figures into analysis and answers correspond to the research questions. Explain and generalize what the data imply.

- We collaborate to finish the report. It is also time consuming to integrate the results and phrase them into words.

Testing

The accuracy and credibility of our data is undoubtable because all of our data is based on real YouTube statistics. One way to test our results is to compare our conclusions with other influential website's. https://www.statista.com/statistics/373753/most-viewed-youtubers-all-time/ This link provides the most viewed channels in 2023 which are similar to what we did in 2021 analysis(some of these channels changed but the category distribution didn't change too much).

To test our code works, we also created small dataframes(whether by picking rows directly or randomly) and used the assert equals function to test our codes.

Collaboration

Since the dataset 2 contains json files, and we were interested in the data listed by a website (dataset 3), which does not provide any file to download, we used websites like pandas documentation and stackoverflow to figure out the syntax and ways to convert them into CSV files.

Website we got the dataset 3:
https://us.youtubers.me/global/all/the-highest-paid-youtubers-of-2021