INFO-250 Final Project

Kevin Shi & Rithvik Sukumaran

September 6, 2023

shinyapps.io: https://kevinshi4.shinyapps.io/info250_project2B/

Introduction

In this project, we aim to visualize, analyze, and discover key findings and intricacies of global YouTube statistics in the year of 2023. Our dataset specifically focuses on statistics of the most subscribed YouTube channels around the globe. By using visualizations such as geographic maps, correlation matrices, scatter plots, bar graphs, and line plots, we hope to visualize and analyze the valuable insights to be gained from contemporary YouTube data. Our project also contains a special statistical analysis as an additional component (See “Statistical Analysis” tab in our Shiny App).

The main target audience of our work is aspiring YouTube content creators, individuals who create their own YouTube channels and publish content on their channels. These aspiring content creators, who are passionate about YouTube, possess a keen understanding and deep interest in the modern world of YouTube. While many YouTubers may be well versed in the intricacies of their own channel’s performance, they may not be familiar with the performance of other YouTube channels in the ever-changing landscape of contemporary YouTube data. Particularly, YouTubers may be interested in the performance of their competitors or the performance of channels in their subject field, but may not have the means or ability to perform any type of analysis. To this end, our project aims to appeal to YouTubers who wish to learn and understand more about the valuable insights of modern YouTube data.

There are many aspects of our data that may be interesting to content creators. For example, visualizing the relationships between key variables such as channel rank, channel category, subscriber count, video views, number of uploads, monthly and yearly earnings, creation date, etc, we are able to reveal key relationships between these variables that offer crucial insights to interested content creators.

Moreover, since our dataset contains the statistics of the top 995 YouTube channels in the world, our project may particularly appeal to the top-most YouTube channels in the world. As the popularity of YouTube channels grows, competition grows as well. These top channels may want to see how they are compared to other top channels throughout the world. Aspiring content creators can also draw inspiration and motivation from these top channels to pave their way to become more popular as well.

Overall, since billions of people use YouTube around the world, YouTube is undoubtedly a significant platform. Our project's objective is to visualize and analyze the valuable information derived from current YouTube data.

Research Questions

- What is the geographic distribution of the top 995 YouTube channels around the world?
- Which countries have the most successful YouTube channels?
- Is there a strong correlation between the number of subscribers and number of video views?
- Is there a strong correlation between number of uploads to number of subscribers, video views, and earnings?
- What is the correlation between earnings and number of subscribers?
- What is the correlation between earnings and number of video views?
- What are the top YouTube channels based on subscribers and views?
- What is the relationship between a channel's subscriber count and its video views?
- What are the most popular content categories on YouTube?
- How do estimated earnings vary across different YouTube channels?
- Is there a relationship between a channel's creation date and its performance?

Description of Data and Methods

Our Kaggle dataset is about global YouTube statistics in the year of 2023. Our dataset specifically focuses on statistics of the most subscribed YouTube channels around the globe. There are 995 records and 28 different variables. The key variables of this dataset are the name of the YouTube Channel, its rank, number of subscribers, number of video views, and the geographic location of the channel.

There are some null values for some of the records in columns such as channel category, country_rank, channel_type_rank, video_views_for_the_last_30_days, subscribers_for_last_30_days, Gross tertiary education enrollment (%), population, unemployment rate, and urban_population. We plan on focusing on other variables (which do not have any null values) to build our visualizations. We may still reference the variables mentioned above, but they will not be the main basis of our visualizations. It is important to note, however, that there are some records that have null values for country, country abbreviation, and latitude and longitude; these records have these null values because the particular YouTube channel is not tied to any specific geographic location. Other variables such as country_rank, Gross tertiary education enrollment (%), population, unemployment rate, and urban_population are also null because there is not a country tied to the channel. As a result, we have accounted for these factors in our visualizations.

On Kaggle, the creator of the dataset, named Nidula Elgiriyewithana, noted that the "dataset was meticulously compiled from various reputable sources, ensuring accuracy and reliability of the information presented." Nidula Elgiriyewithana is a datasets expert on Kaggle, and he collected the data from multiple reputable sources using Python. We will have to trust that the data indeed was collected from reputable sources and that the data in the dataset is valid data. The dataset is updated annually, and as of 8/13/2023, the dataset was updated sixteen days ago.

Dataset Variables we Used:

Ordinal:

- **rank:** Position of the YouTube channel based on the number of subscribers
- **video_views_rank:** Ranking of the channel based on total video views
- **country_rank:** Ranking of the channel based on the number of subscribers within its country
- **channel_type_rank:** Ranking of the channel based on its type (individual or brand)

Categorical:

- **Youtuber:** Name of the YouTube channel
- **category:** Category or niche of the channel
- **Country:** Country where the YouTube channel originates
- **Abbreviation:** Abbreviation of the country

Numerical:

- **subscribers:** Number of subscribers to the channel
- **video views:** Total views across all videos on the channel
- **uploads:** Total number of videos uploaded on the channel
- **video_views_for_the_last_30_days:** Total video views in the last 30 days
- **lowest_monthly_earnings:** Lowest estimated monthly earnings from the channel
- **highest_monthly_earnings:** Highest estimated monthly earnings from the channel
- **lowest_yearly_earnings:** Lowest estimated yearly earnings from the channel
- **highest_yearly_earnings:** Highest estimated yearly earnings from the channel
- **subscribers_for_last_30_days:** Number of new subscribers gained in the last 30 days
- **Gross tertiary education enrollment (%):** Percentage of the population enrolled in tertiary education in the country
- **Population:** Total population of the country
- **Unemployment rate:** Unemployment rate in the country
- **Urban_population:** Percentage of the population living in urban areas
- **Latitude:** Latitude coordinate of the country's location
- **Longitude:** Longitude coordinate of the country's location

  *Latitude and Longitude are of course geographical data, but in their raw form, they are decimal values.

Datetime:

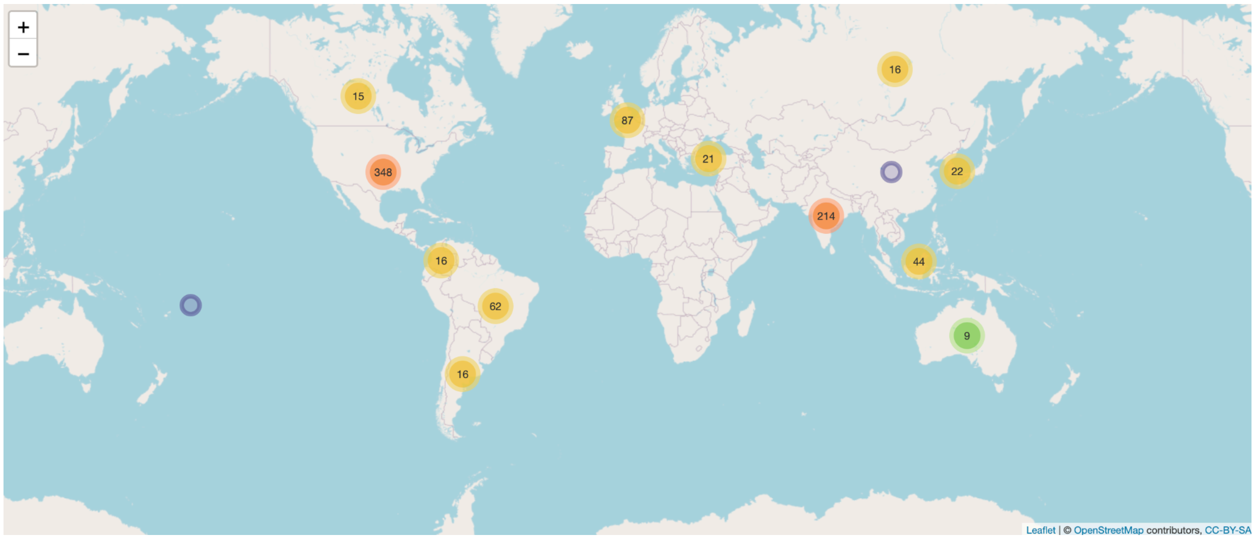
- **created_year:** Year when the YouTube channel was created

- **created_month:** Month when the YouTube channel was created
- **created_date:** Exact date of the YouTube channel's creation

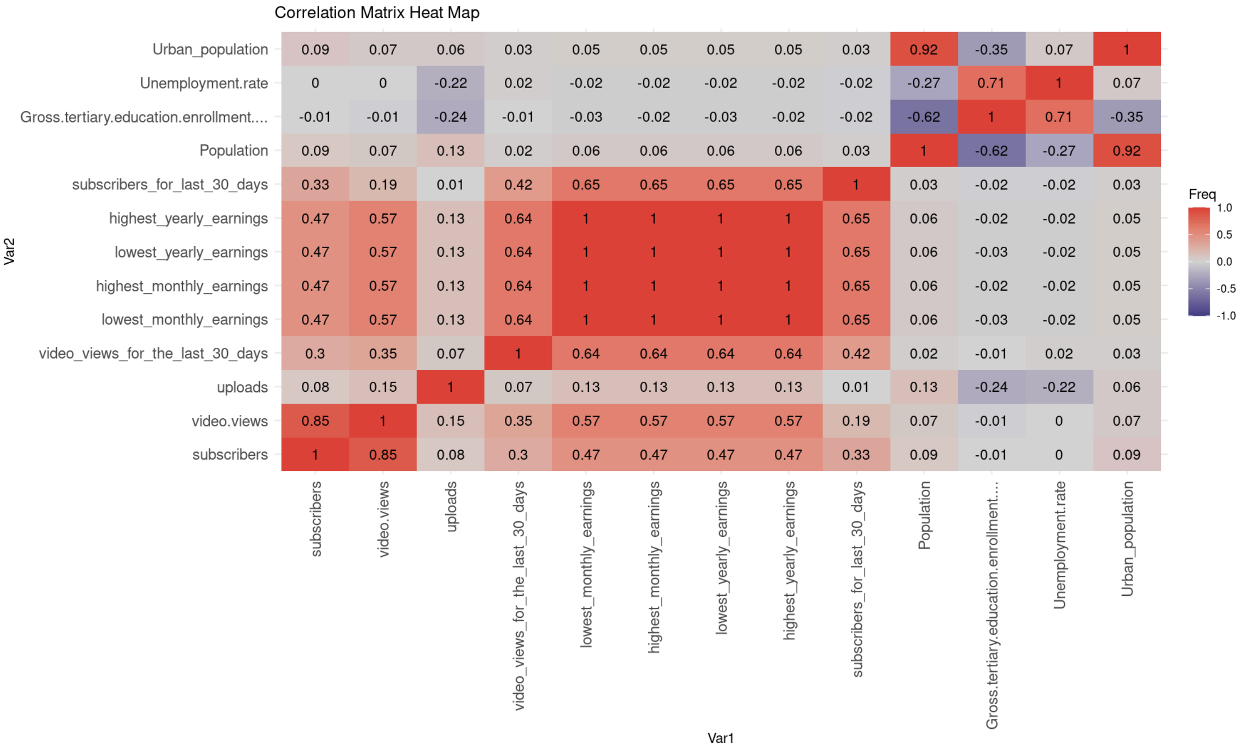
Results

To answer our research questions, we created a series of visualizations and compiled them into a single Shinyapp using R and the following R libraries: rsconnect, ggplot, ggplot2, plotly, shiny, leaflet, leaflet.extras, scales, bslib, dplyr, scales, reshape2, and ggextra. Some of these libraries, such as ggplot, plotly, and shiny, served as our main tools for creating and consolidating our visualizations. The others helped us supplement those libraries, either extending them by providing extra functionalities or making them easier to work with. Our final dashboard is a Shinyapp, consisting of a map, a correlation heatmap, a scatterplot, a line plot, and a series of box plots.

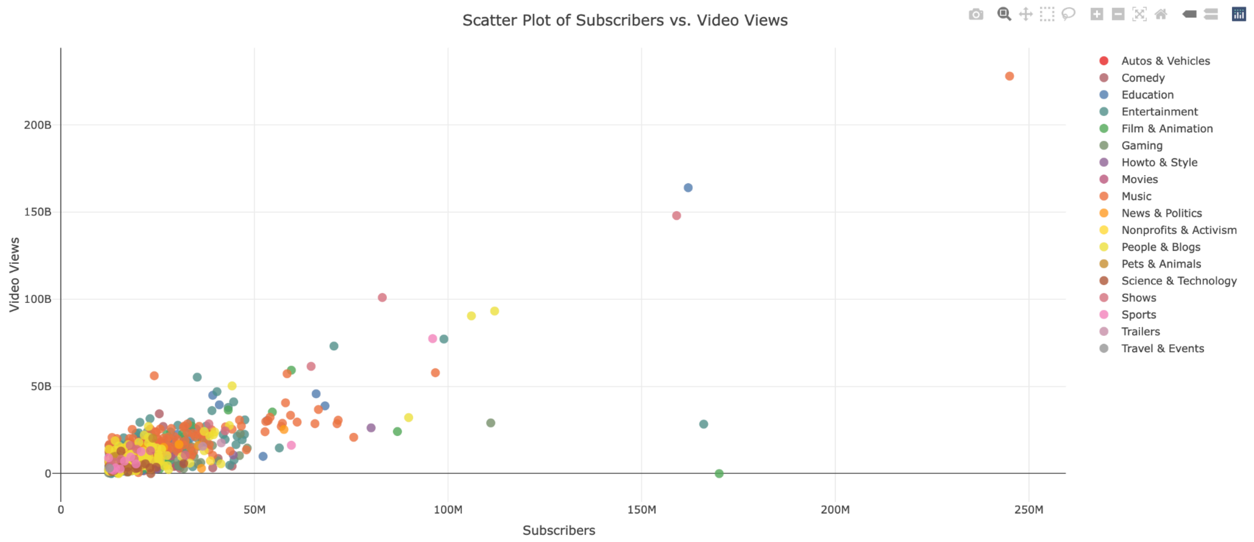
Our map takes advantage of the country column in the dataset to plot the most popular YouTubers per region. Depending on the level of zoom, it can display the number of top YouTubers at different levels of detail. These levels of detail range from top Youtubers in each hemisphere to top Youtubers in each individual country, with some intermediate levels including continents and groups of adjacent countries. Clicking on a node on the map will zoom in on that region, until you are at the level of an individual country. At this point, clicking on a node will display information regarding the top Youtubers of that country. This allows users to view the geographical distribution of the current top YouTubers. Some of our personal gleanings include the fact that the US and India are the top countries when it comes to the world's most popular YouTube channels. With 313 and 169 of the top YouTubers in the world respectively, they blow all other countries out of the water when it comes to quantities of popular YouTubers.



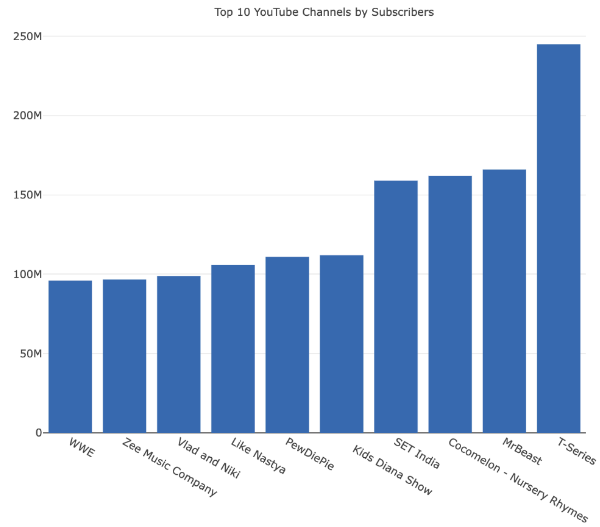
Next, our heatmap, or correlation matrix, displays Pearson correlations of each of the numerical features we used. There are many variables that are very strongly correlated with each other. For example, highest_monthly_earnings, lowest_monthly_earnings, highest_yearly_earnings, and lowest_yearly_earnings are all very strongly correlated with each other. Each of these variables have a correlation of 1 with each other, indicating an incredibly strong positive relationship between all of them. Two other variables strongly correlated with these four are video_views_for_the_last_30_days and subscribers_for_the_last_30_days, with correlation coefficients of 0.64 and 0.65 respectively. This indicates that as video views or subscribers increase, so do earnings. Overall video views and total subscribers also share a very strong positive relationship with a correlation coefficient of 0.85.

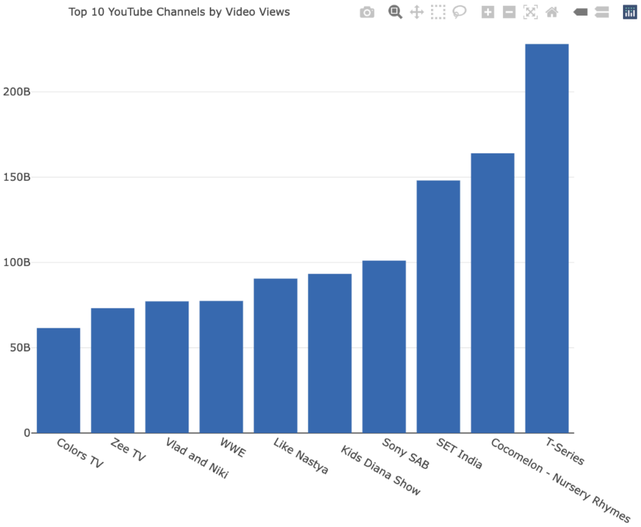

Correlation Matrix Heat Map

To model the relationship between number of subscribers and number of video views, we made an interactive scatterplot of each YouTuber. This helped us to visualize the strong positive relationship indicated by our correlation matrix (correlation coefficient: 0.85). From this scatterplot, we learned that the relationship is not completely linear but it is clear that as one variable increases, so does the other. Additionally, data points in the scatter plot are color coded by channel category, and channel categories can be hidden in the scatter plot to show the relationship of number of subscribers and number of video views for chosen categories. Mousing over the data points shows information about the channel's name, category, number of subscribers, number of video views, number of video views in the last 30 days, and country of the YouTube channel.

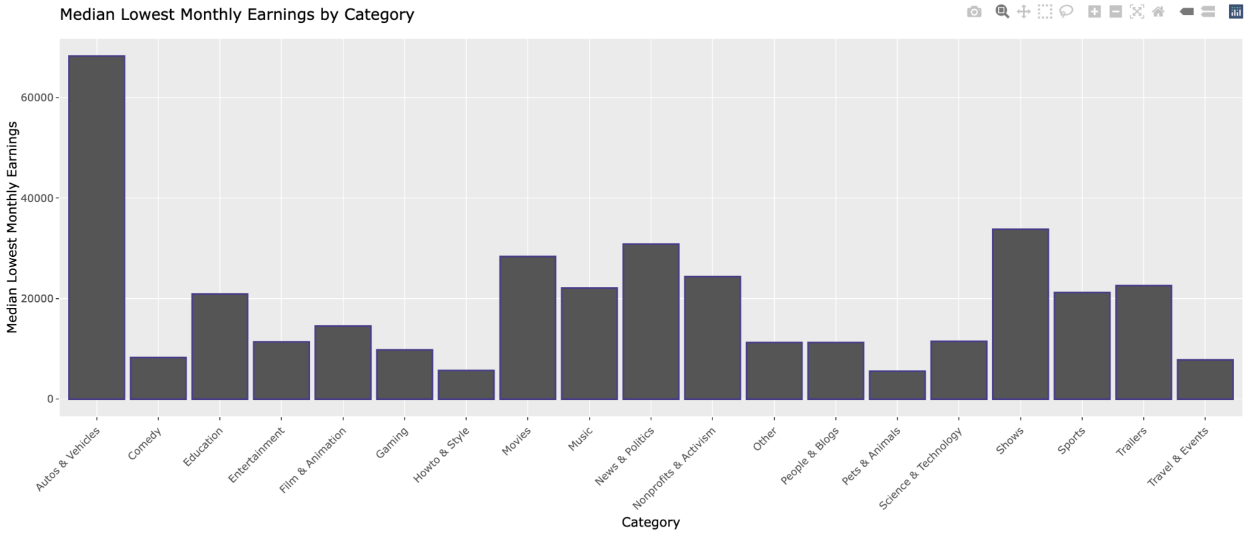


Our bar plots of top 10 channels by subscribers and top 10 channels by views show that the most successful channels are not individuals but are rather corporations. Some of the most successful individual YouTubers on the platform, such as *Like Nastya* and *Vlad and Niki* make it onto the top 10 channels by views, and Pewdiepie and Mr. Beast join them in the top 10 channels by subscribers. However, even if they are not corporate accounts, they are professionals who do (or at some point did) YouTube or content creation full-time. It is common understanding that someone doing YouTube for fun would not be able to make the ranks of the most successful on the platform, but these bar plots we have created provide concrete numerical proof of that understanding.

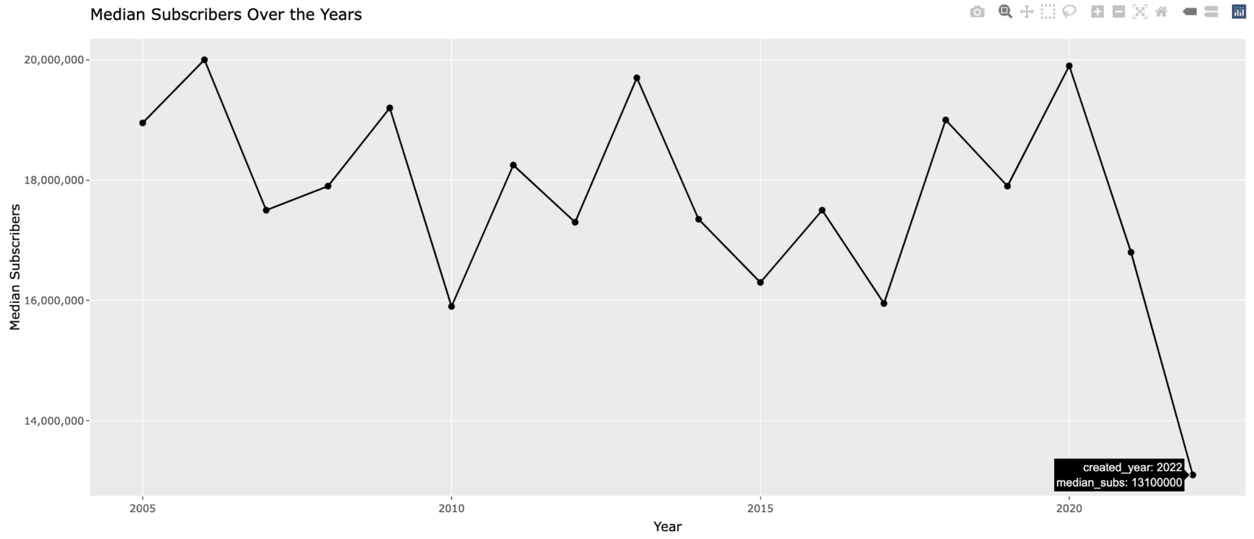




Our final bar chart models the median lowest monthly earnings for each channel category. This bar chart shows which categories tend to make the most money per month, even at their lowest performance. From this, it is apparent that Shows, New & Politics, and Movies are three of the most popular categories. However, Autos & Vehicles significantly outperforms any of the other categories, with a bar roughly twice as high as the next highest, which is Shows.



To understand if the year a channel was created had any impact on its long term success, we grouped the channels by year, and then we took the median number of subscribers for each year and plotted them against the years of creation. In other words, this is a time series plot of the median number of subscribers of each channel, with the x-axis being each year and the y-axis being the median number of subscribers of every channel created that year. From this line plot, we can see that the year of creation does not seem to have any significant impact on the number of subscribers. There is a wave-like pattern where the median number of subscribers rises and falls, but there is no one year that stands out from the others, aside from channels made in 2022 having a median subscriber count of roughly 13 million, which is the lowest median subscriber count; due to the axis scaling, this count appears to be an outlier but is actually not an outlier. There is also no conclusive proof in this plot that older channels tend to have more subscribers than newer channels, which is an interesting finding.



Conclusion

In this project, our primary goal was to visualize, analyze, and discover key findings and intricacies of global YouTube statistics in the year of 2023. Through the utilization of R and various libraries such as ggplot, plotly, shiny, leaflet, and more, we were able to create a comprehensive and interactive R Shinyapp, allowing users to explore our visualizations about YouTube data.

Our dataset encompassed a diverse range of variables, some of the most important including channel rank, category, subscriber count, number of video views, number of uploads, monthly and yearly earnings, creation date, geographic longitude and latitude, and more. These variables served as the bases of our visualizations: geographic map, correlation matrix, scatter plot, bar charts, and line graph. These visualizations in turn allowed us to visualize, analyze, and understand the factors that contribute to YouTube channel performance.

Summary of Key Findings:

- **Geographical Popularity:** The United States and India have the most popular YouTube channels, showing the highest numbers of top YouTubers globally. This insight sheds light on the global distribution of YouTube content creators.
- **Correlation:** Earnings, video views, and subscriber counts have strong correlations, which highlights strong relationships between certain variables and indicates insights to content creators looking to optimize their channels.
- **Top Channels:** While corporations dominated the top channels by subscribers and views, successful individual YouTubers like Like Nastya, PewDiePie, and MrBeast showcased that engaging content and channel advertisement can also lead to exceptional popularity.
- **Category Analysis:** We found that certain channel categories, namely Shows, News & Politics, Movies, and Autos & Vehicles, tend to outperform others in terms of earnings, even at their lowest performance levels.
- **Creation Year:** We found no conclusive proof that older channels tend to have more subscribers than newer channels. Our analysis also did not reveal a linear relationship between channel creation year and median subscriber count, which indicates that factors beyond the creation date greatly influence channel popularity.

To conclude, YouTube is undoubtedly one of the most significant platforms in the world and by creating comprehensive visualizations of YouTube data in 2023, our project serves as a valuable resource for aspiring YouTube content creators, presenting them with key insights to help them achieve greater success in the contemporary YouTube platform.