kaggleX

KaggleX- Showcase

YouTube Engagement Analyzer:

Enhancing Content Strategy Through Data

Zoey Espinoza

Background

- My background is in Media and Broadcasting and I have an educational background with a Bachelors in Architecture
- I am currently a Creative Technologist and Video Technician making a career change into Data Science with a bootcamp and self learning tools

Project Definition

- In this data science and machine learning project, I delve into the vast world of YouTube content to analyze user engagement, and recommend personalized video suggestions. Leveraging a comprehensive YouTube statistics dataset, my goal is to provide content creators and marketers with actionable insights to optimize their strategies and boost their channel's performance.
- EDA, Statistical Data Analysis, Data Visualization, Machine Learning, Predictive Modeling
- I learned how to do an in-depth analysis using a dataset from Kaggle. Step by step process of completing my first personal project and not using an example

Project Links

My project is available to view here:

- Github project link
- Kaggle project link

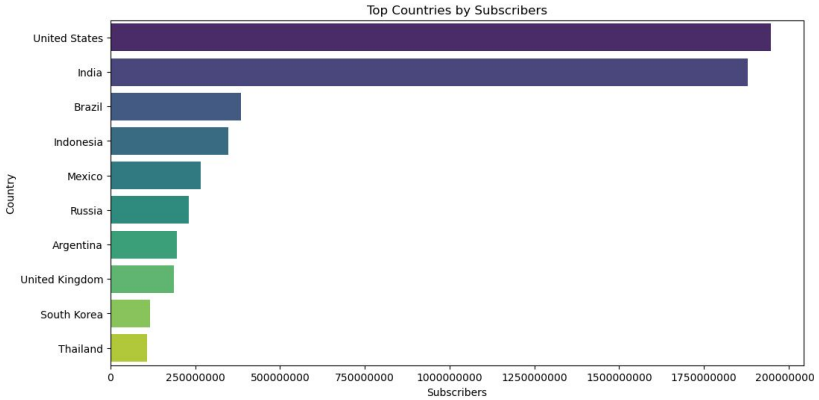
Social Links:

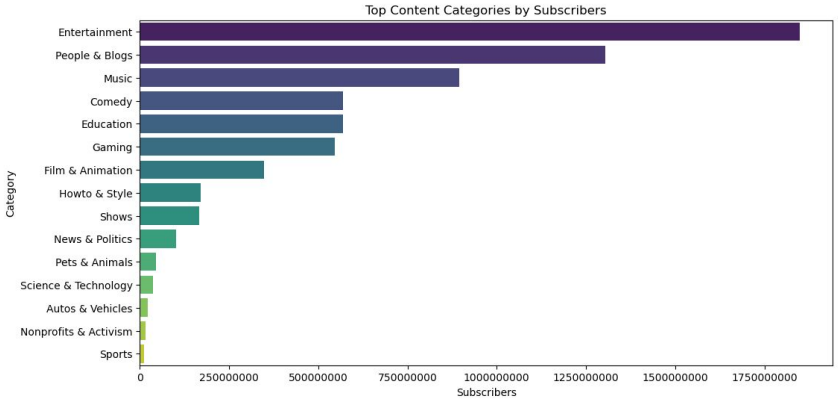
- Linkedin profile

Steps

1. **Content Creator Profiling:**
   - Use the 'subscribers' and 'video views' columns to identify top content creators.
   - Group creators by category and country to understand the diversity of content on YouTube.
2. **Performance Metrics Analysis:**
   - Analyze 'subscribers,' 'video views,' and 'uploads' to identify trends and patterns in channel growth.
   - Create visualizations to represent changes over time.
3. **Monetization Insights:**
   - Utilize 'lowest_monthly_earnings' and 'highest_monthly_earnings' to analyze the earnings potential of content creators.
   - Identify factors that contribute to higher earnings.
4. **Audience Engagement and Demographics:**
   - Explore 'Country' and 'Urban_population' to understand the geographic distribution of viewers.
   - Analyze 'Gross tertiary education enrollment (%)' to assess the education level of the audience.
5. **Predictive Modeling:**
   - Build predictive models using historical data to forecast metrics like 'subscribers' and 'video views.'
   - Evaluate the accuracy of the models and make recommendations based on predictions.

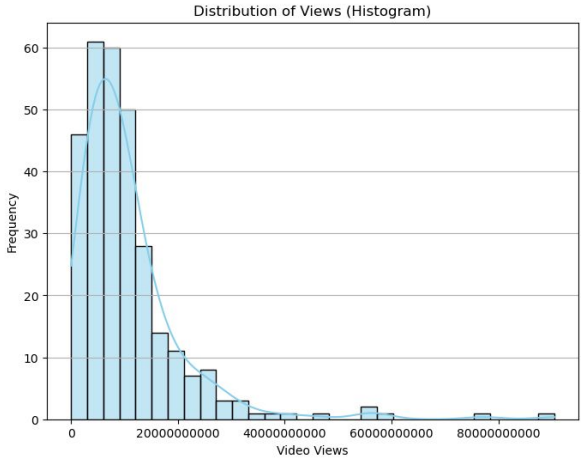
Creator Profiling





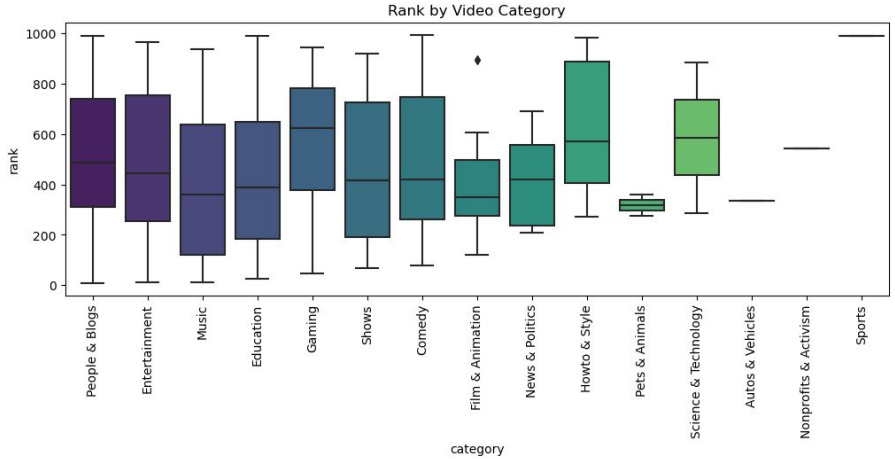
Top content categories are Entertainment, Music, and People & Blogs. The top countries are the United States, India, and Brazil. This gives us a good idea of the type of content that is popular on YouTube and where the content creators are located.

Creator Video View and Frequency



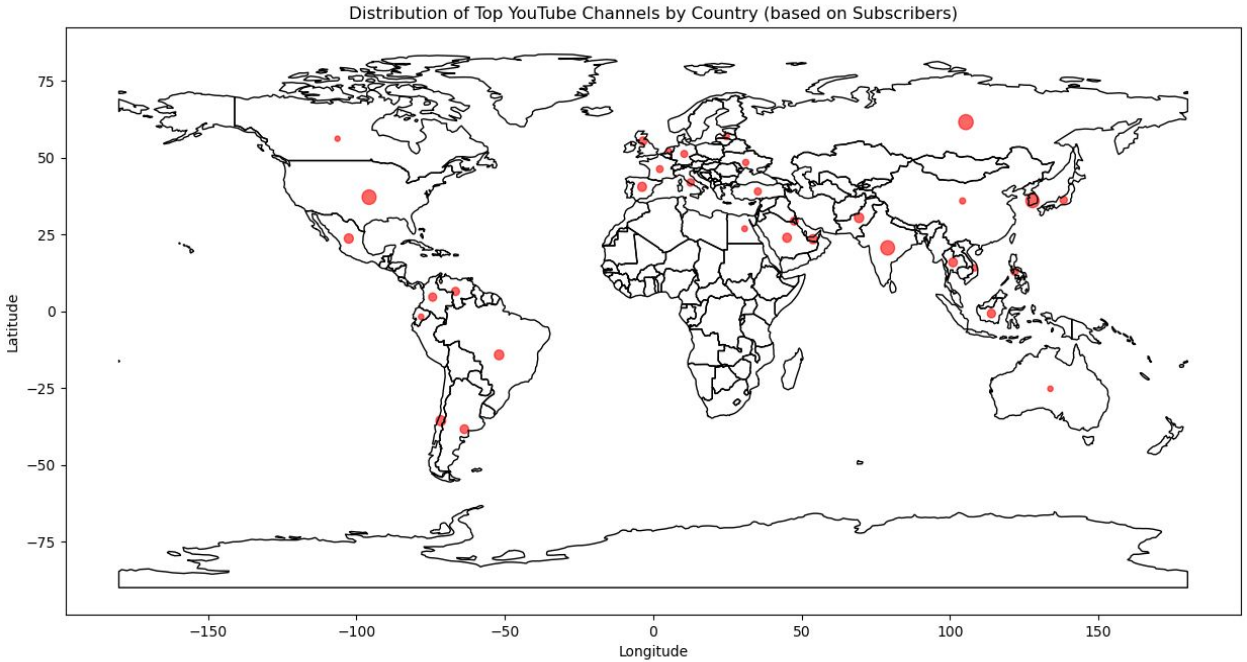
In the distribution of views, we can see that most of the videos have less than 1 million views. There are a few videos that have more than 1 million views, but they are very few in number. We can assume that most of the videos are not very popular, but there are a few videos that are very popular. The distribution is right-skewed, and the mean is greater than the median. The frequency of videos decreases as the number of views increases.

Rank by Video Category



The rank is a measure of a video's popularity within its category. In the context of this boxplot, a lower rank indicates higher relative popularity within the category. When we examine the boxplot, we observe that the median rank for 'Entertainment' videos is higher than that of other categories. This suggests that, on average, videos in the 'Entertainment' category have lower relative popularity within their category. To clarify, a higher rank corresponds to a lower number of subscribers, and the 'Entertainment' category, with its higher median rank, tends to have videos with fewer subscribers compared to other categories.

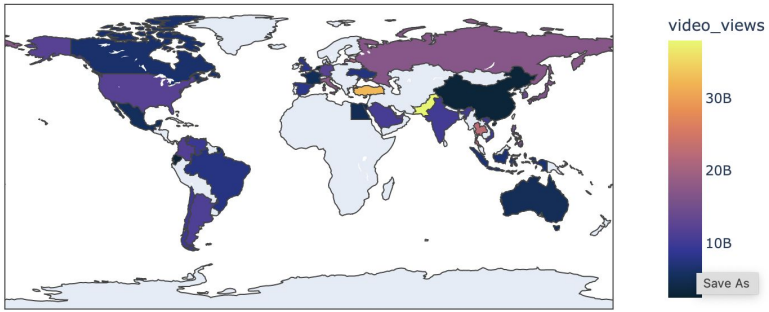
Youtube Channels by Country



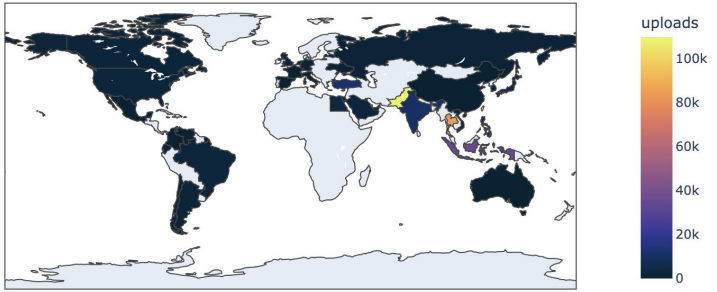
The Distribution of Top YouTube Channels by Country (based on Subscribers) can be seen in the image above as a world map, the size of the circles represent the number of subscribers.

Youtube Channels by Country

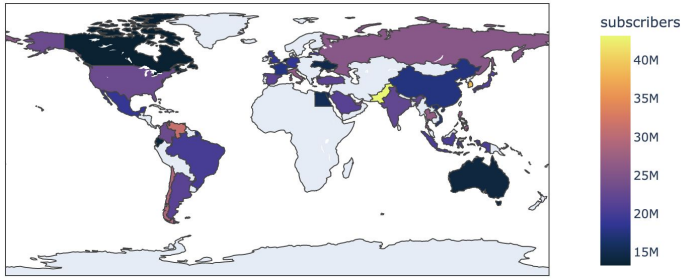
Average video_views by Country



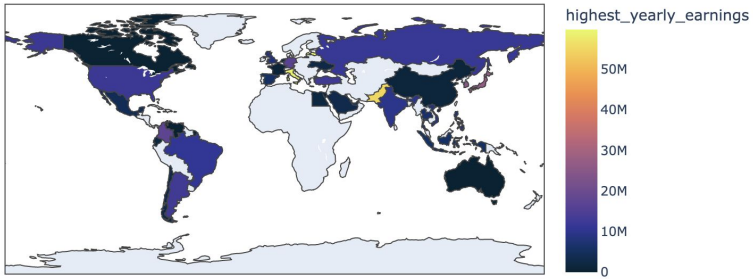
Average uploads by Country



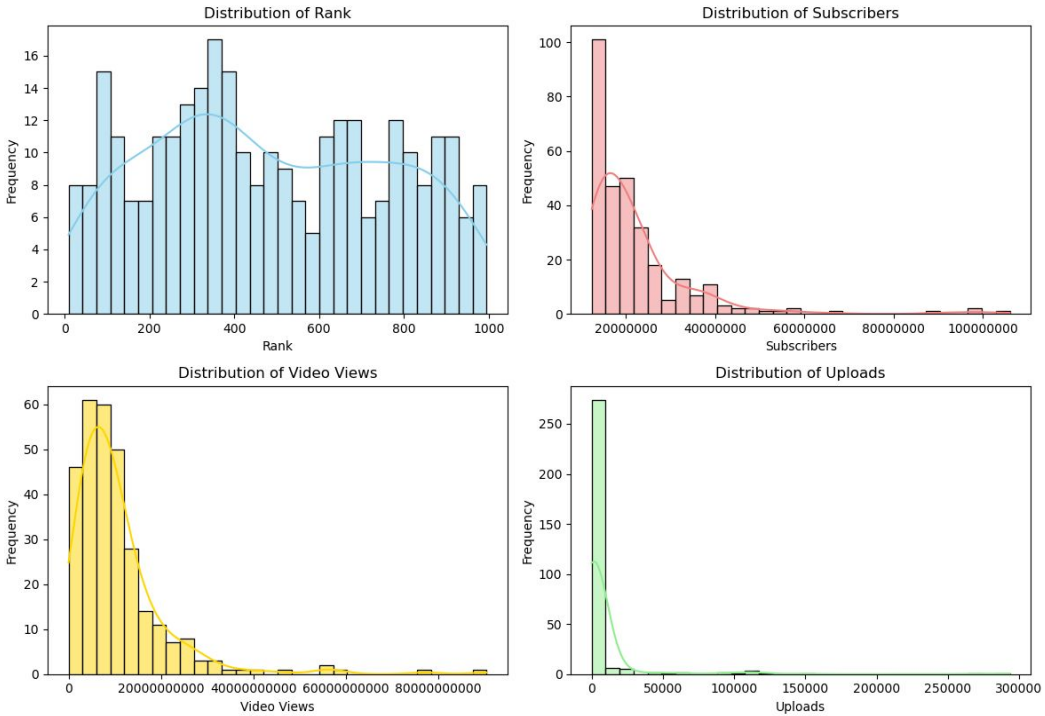
Average subscribers by Country



Average highest_yearly_earnings by Country

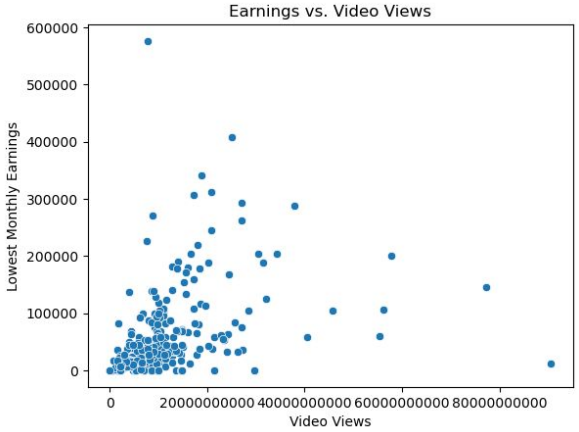


Distribution of Rank, Subscribers, Views, and Uploads

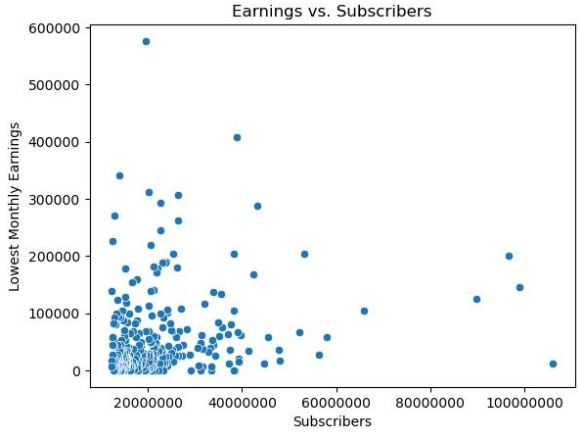


The above subplots show the distribution of the following variables: rank, subscribers, video views, and uploads. The distribution of rank is more evenly distributed, which means that most channels have a high rank. The distribution of subscribers is skewed to the right, which means that most channels have a low number of subscribers. The distribution of video views is also skewed to the right, which means that most channels have a low number of video views. The distribution of uploads is also skewed to the right, which means that most channels have a low number of uploads.

Monetization Insights

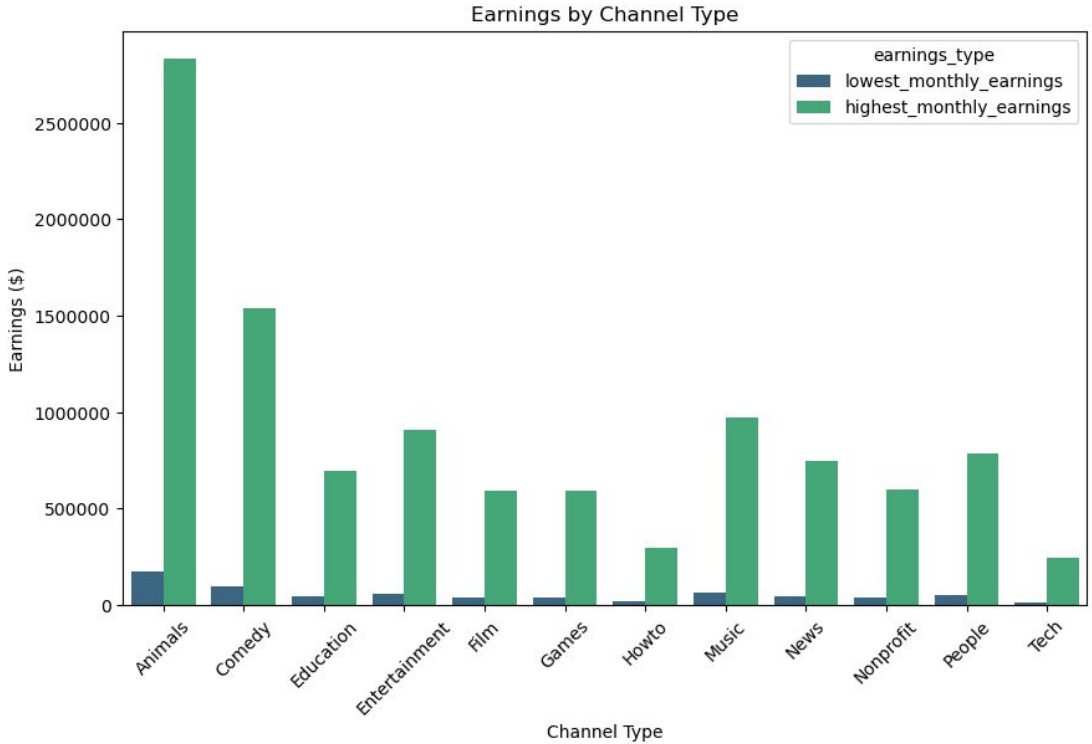


Analysis of Earnings and Video Views for top creators, find that there is a positive correlation between the two variables. As video views increase, earnings also increase. This makes sense because the more views a video gets, the more ads are shown, and the more money the content creator makes.



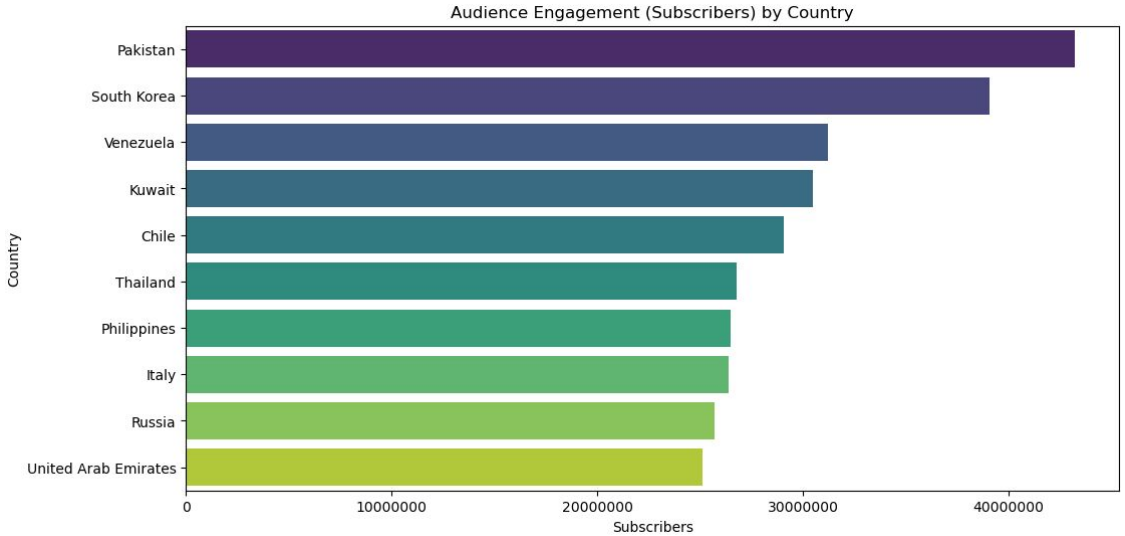
When looking at Earnings vs. Subscribers, we can see that there is a positive relationship between the two variables. As the number of subscribers increases, the earnings also increase. However, there are some outliers in the data. For example, there are some channels with a high number of subscribers but low earnings. This could be due to the fact that the channel is new and has not yet started earning money. There are also some channels with a low number of subscribers but high earnings. This could be due to the fact that the channel is old and has already earned a lot of money. Overall, there is a trend in that most subscribers are below 40 million and below 100,000 in earnings.

Earnings by Channel Type



When looking at Earnings by Channel Type, we can see that the highest monthly earnings are from the Animals category. This is followed by the Comedy category. The lowest monthly earnings are from the Tech & Howto category. This is inciteful for people decided what type of content to create on YouTube, and which will yield the highest earnings.

Audience Engagement and Demographics

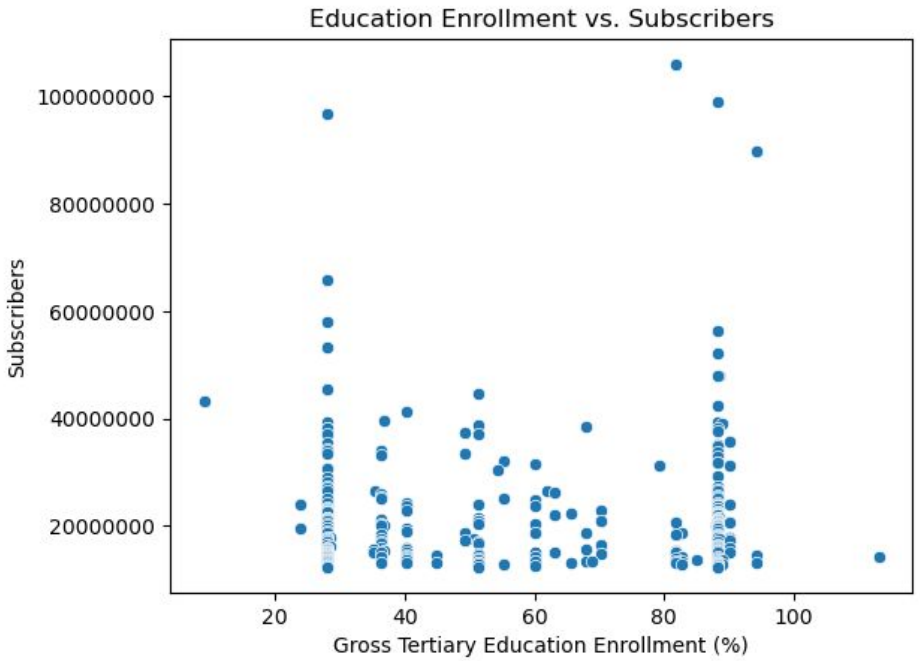


	country	subscribers
19	Pakistan	43200000
23	South Korea	39066667
31	Venezuela	31200000
15	Kuwait	30500000
4	Chile	29066667
25	Thailand	26750000
20	Philippines	26500000
13	Italy	26400000
21	Russia	25688889
28	United Arab Emirates	25100000

The list demonstrates that popular YouTube channels with a significant number of subscribers are not limited to a single country. They have a global audience, as these channels are located in various countries, including Pakistan, South Korea, Venezuela, Kuwait, Chile, Thailand, the Philippines, Italy, Russia, and the United Arab Emirates.

Some countries on the list, like Pakistan, Venezuela, and the United Arab Emirates, are considered emerging markets in terms of digital content creation. This indicates that YouTube's influence is not limited to developed nations, and creators from emerging markets can also achieve a global reach. The presence of these channels in different countries suggests that the content they produce is likely diverse and may cater to various languages, cultures, and interests.

Education Enrollment and Subscribers

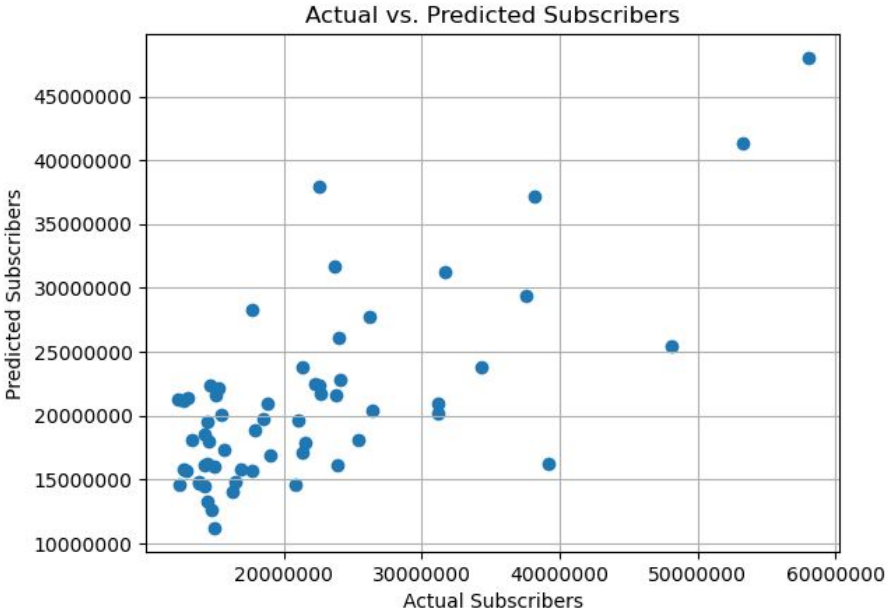


The scatterplot visualizes the relationship between the percentage of gross tertiary education enrollment and the number of subscribers for YouTube content creators. This analysis aims to understand if there is any observable connection between the level of education enrollment in a country and the popularity of YouTube channels within that country.

The data points appear to cluster in various regions of the plot. While some channels from countries with high education enrollment have a substantial number of subscribers.

There are noticeable outliers in the plot, representing channels with exceptionally high subscriber counts regardless of the education enrollment rate in their respective countries. These outliers may be driven by unique content, effective marketing strategies, or other external factors.

Predictive Modeling



Mean Squared Error (MSE): 6.79×10136.79×1013
Represents the average of the squares of the differences between observed and predicted values

Root Mean Squared Error (RMSE): 8,239,489
Represents the square root of MSE and gives an idea of the magnitude of the error in the same unit as the target variable.

Mean Absolute Error (MAE): 4,149,722.81
Represents the average absolute difference between observed and predicted values.

R-squared (R2R2): 0.3133
Represents the proportion of the variance in the target variable that is predictable from the features. A value of 1 indicates perfect prediction, while a value of 0 indicates that the model is no better than simply predicting the mean of the target variable for all observations.

Conclusion

In this project, we conducted an exploratory data analysis (EDA) and predictive modeling on a dataset containing information about various YouTube channels. The dataset covered a wide range of variables, including channel rankings, subscribers, video views, uploads, and demographic details. Our primary objectives were to understand the data, perform necessary preprocessing, conduct an exploratory analysis, and create a predictive model for subscribers based on other channel attributes.

Key Steps and Findings:

Data Cleaning and Preprocessing: We began by cleaning the dataset, addressing issues like missing values, data types, and scaling. Additionally, we converted columns to lowercase and snake case for consistency.

Exploratory Data Analysis (EDA): We conducted EDA to gain insights into the data. Visualized the distribution of key features, such as subscribers and video views, and observed relationships between variables. Notably, we examined demographic factors, such as education enrollment and unemployment rates, and their correlation with subscribers.

Challenges:

Predictive modeling presented challenges, as the Mean Squared Error (MSE) values were initially high. Further feature selection and engineering might improve model performance.

In summary, this analysis provides valuable insights into the factors that influence the number of subscribers for YouTube channels. It serves as a foundation for further exploration and model refinement to better understand and predict channel growth on this popular platform.


kaggle