Part 1: Context and EDA

Context and Source Description

The datasets I have chosen for my assignment contain the YouTube statistics data which reveals parameters like views, subscribers, and video counts of about 1000 YouTube channels from all over the globe. Each column is a unique player, adding flavour to the dataset.

The two open-source datasets I have selected for this project are from Kaggle.

Dataset 1: Most Subscribed YouTube Channels
Link to the dataset: https://www.kaggle.com/datasets/surajjha101/top-youtube-channels-data

Dataset 2: Global YouTube Statistics 2023
Link to the dataset: https://www.kaggle.com/datasets/nelgiriyewithana/global-youtube-statistics-2023

Figure 1: YouTube
https://www.analyticssteps.com/blogs/how-extract-analyze-youtube-data-using-youtube-api

YouTube is a global stage for digital content that is widely accessible to people all over the world. With the help of this dataset, we can explore the world of YouTube by learning about channel growth, viewing trends, and a lot more.

Dataset Description and Summary Statistics

Dataset Description:
To get the final dataset for the visualization, I have combined two datasets. Dataset 1 contains columns 1-7, and Dataset 2 contains columns 8-10. Dataset 1 provides all of the channel details, while Dataset 2 includes information on the channel's country of origin.

Table 1: Dataset description

Column No.	Column Name and Datatype	Description of the column
1	Channel Name (chr)	Name of the YouTube channel
2	Rank (int)	Rank as per the subscribers
3	Subscribers (num)	Total subscribers of the YouTube channel
4	Video Views (num)	Total video views on the channel
5	Video Count (num)	Total videos uploaded
6	Category (chr)	Type of the YouTube channel
7	Start Year (int)	The year in which the channel started
8	Country (chr)	Origin country of the channel
9	Population(num)	Population of the country in which the channel was started
10	Unemployment rate(num)	Population of the country in which the channel was started

The final dataset contains the details of 1000 YouTube channels from 2005 to 2021, with 1000 rows and 10 columns. The size of the CSV file of the merged dataset which is used for visualization is 97.5 KB.

The dataset contains 10 rows of deleted YouTube channels. The ‘video views’ and ‘video count’ columns of these rows have a value of zero. The deleted YouTube channels are Gaming, Live, Machinima, Minecraft - Topic, Music, News, Popular on YouTube, Sports, TV Shows, and YouTube Movies.

There is an outlier in the dataset, which is ranked number 100 when the dataset is sorted as per rank. The channel name of the outlier is ‘YouTube’. This outlier row is eliminated from the dataset to get accurate plotting.

Summary of data statistics:

Subscribers and Video Views: With a mean of 20.5 million, the distribution of subscribers is extremely uneven, ranging from roughly 11 million to a maximum of 222 million. This suggests that the top channels have a wide range of audience sizes. With an average of over 9 billion views, video views follow a similar pattern as subscribers.

Video Count: Throughout the dataset, there have been roughly 8000 mean average video uploads.

Category, Country, Population, and Unemployment Rate: The dataset includes YouTube channels of 18 categories from 44 countries. The population and unemployment rates of the nation where the channel originated are included in the dataset.

Start Year: Most of the channels appeared around 2012, with the rest ranging between 2005 and 2021.

Missing Data:

Figure 2: Missing data

The combined dataset has overall missing values of 9.9%. In particular, the ‘Category’ column has a 3% missing value, whereas the ‘Country’, ‘Population’, and ‘Unemployment Rate’ columns have higher missingness, with 32% of their respective values missing.

Part 2: Design

Static Plot:
The static plot of Category against total video views per category is being plotted. The purpose of this visualization is to identify the most viewed category. And also, the relation between the views and the total video uploads in each category. The prospective audience for this is content creators, who can potentially gain more views by modifying their content creation categories. It will also help marketing companies and advertisers target their advertisements and tailor promotional campaigns more effectively by aligning with popular content categories.

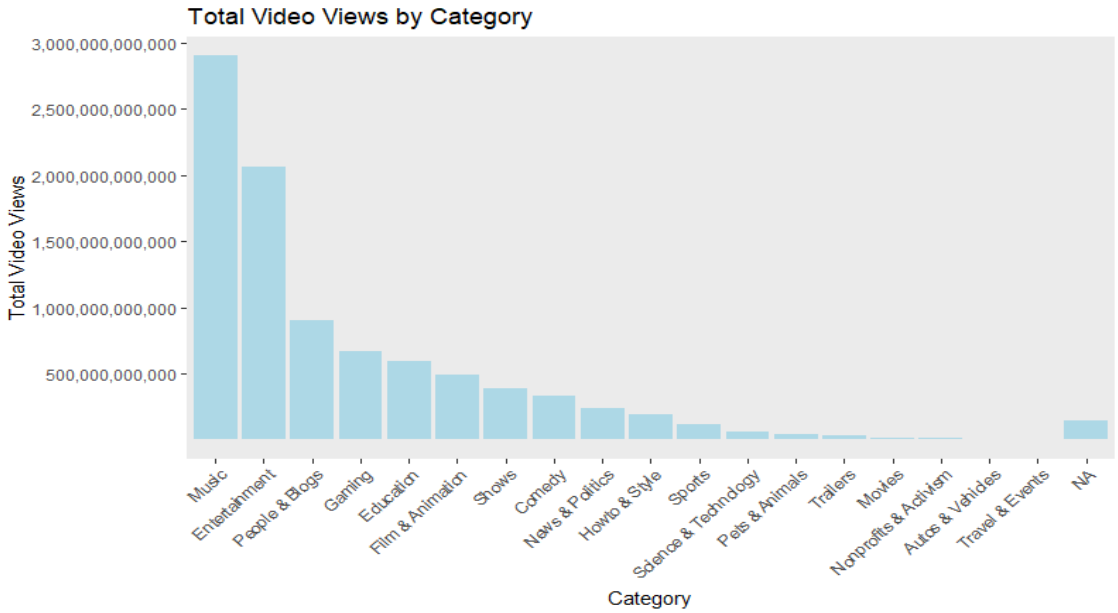

Figure 3: Total Video views vs Category Plot

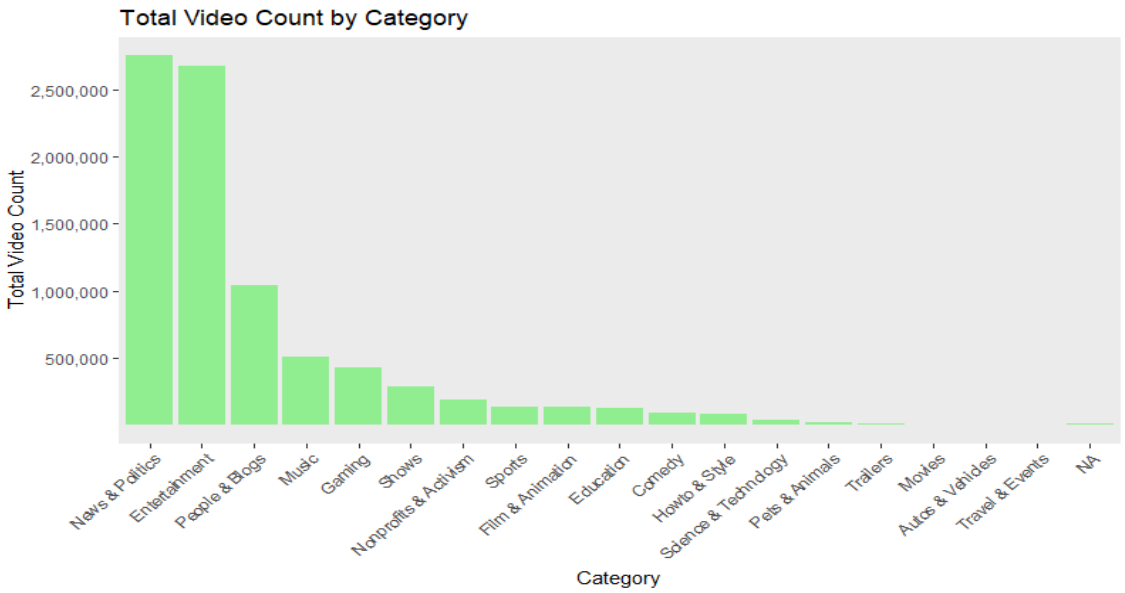

Figure 4: Total Video Count vs Category Plot

For my final submission, I'm planning to classify the data using a bar plot by having category on the x-axis and figure out how many video views each category has by placing video views on the y-axis to visually represent the insights. The categories will be ranked according to the total number of views, with the most popular category at the top. The colour of each bar will represent the total number of videos uploaded in that category. For clarity's sake, the total number of videos uploaded per category will be annotated above the bars.
The axes, title, and legend labels will be modified by changing their locations and angles to give the design a more finished appearance. These choices are intended to produce an understandable and informative visualization.

Interactive Plot:
The interactive plot would give the time series data about the total number of YouTube channels started in a year from 2005 to 2021, and also, the data of subscribers count of all the channels started in each year.
This visualization lets the audience explore the subscriber trends over time. Viewers can hover over the data points to get detailed information about the specific channels. The combination of the time series plot of the number of channels started over time and the interactive subscribers plot facilitates a comparative analysis.

Visualizations can be used by marketing firms and content creators on YouTube to comprehend the historical channel landscape. This can assist them in making well-

informed strategic decisions about the creation of content and marketing, comparing the channel's growth with that of the nation to which it belongs.

I will be starting off with a time series plot that tracks the evolution of YouTube channels over the years. The data will be grouped by channel start year, offering a summarized view of the total channels created during each year from 2005 to 2021.

Transitioning to an interactive scatter plot, I will be visualizing channels created from 2005 to 2021 along with their corresponding subscriber counts. A text box, displaying essential details like channel name, country, and video views will be shown by hovering over the scattered points.

Joining both plots using the subplot function, I will ensure they share a common Y-axis for better comparison. To maintain visual consistency, the same font sizes will be kept across both plots, resulting in an informative presentation.

Part 3: Final Visualisations

Visualization 1 Commentary

From visualization 1, we can notice that the Music, entertainment, and comedy categories have the highest number of video views, which suggests that these categories are the most popular among the viewers. But we can also notice that music has the highest views with fewer video uploads and the Entertainment category has lesser viewership even after having five times more uploads than the music category. This means that the number of videos viewed is not solely reliant on the number of uploads.

The use of descending order for category-wise video views helps in better understanding the popularity of the categories. The bars in the plot use Viridis color palette, which is color-blind friendly visualization. The use of text labels directly on the bars enhances accessibility to the details. The x-axis scale is angled to 45 degrees for better readability and the Y-axis scale is modified to show the views in billions for easier interpretation of large numbers.

Visualization 2 Commentary

From the line plot, we can see that 119 YouTube channels that were started in the year 2014 could make up the list of top 1000 YouTube channels, which is the highest from year 2005 to 2021. We can see a constant decrease since the year 2014.

The use of clear line and point plot with the removed background gridlines enhances focus on the data. The points along the line plot emphasize the specific data points for clarity.

The scatter plot includes a tooltip along the points with details information about the channel.
The Viridis color palette is chosen to get the uniform color along the points which are color coded by start year. Conversion of the y-axis scale to show the subscribers data in millions improves the readability.

Consistent font sizes are given across subplots to maintain a good visual experience.
Subplots allow for a side-by-side comparison of interactive visualizations with a common x-axis, providing a more comprehensive view of data.