YT Chronicles - Process Book

1. Basic Information

Project title : YT Chronicles

Member details

1) Kalyan Shankar Ragam u1471323@utah.edu
2) Vijay Surya Vempati u1472324@utah.edu

Github Link - https://github.com/dataviscourse2024/group-project-yt-chronicles

2. Background and Motivation

This project is driven by my deep interest in exploring the dynamics and patterns of digital media, with a particular focus on YouTube. As a major platform for global content creation and consumption, YouTube has evolved into a hub for influencers, brands, and creators to connect with audiences worldwide. Our motivation stems from a desire to understand the underlying factors that contribute to a channel's success. By visualizing how these factors interplay to influence a channel's performance, we aim to gain insights into user engagement and content virality all over the world and in a specific country.

Given YouTube's global reach, this project provides a unique opportunity to visualize data from channels across various countries and categories. By examining these, we hope to uncover broader trends in global content consumption and identify potential correlations between channel performance and regional factors. This insight can be valuable for aspiring content creators and digital marketers seeking to understand and leverage the nuances of digital media engagement.

3. Project Objectives

The primary objectives of this project are to explore and visualize various aspects of YouTube channel metrics to gain insights into content creation and consumption patterns across different countries and categories. Specifically, the project aims to:

- **Analyze Country Variations**: Identify the top 10 YouTubers in each country based on their view counts and subscriber numbers.
- **Identify Recent Trends**: Examine recent changes in subscriber counts and video views over the past 30 days for various channels to uncover current content consumption trends.

- **Determine Popular Content Categories**: Visualize the most popular YouTube channel categories (e.g., gaming, education, lifestyle) in specific countries to gain insights into regional content preferences.

4. Data

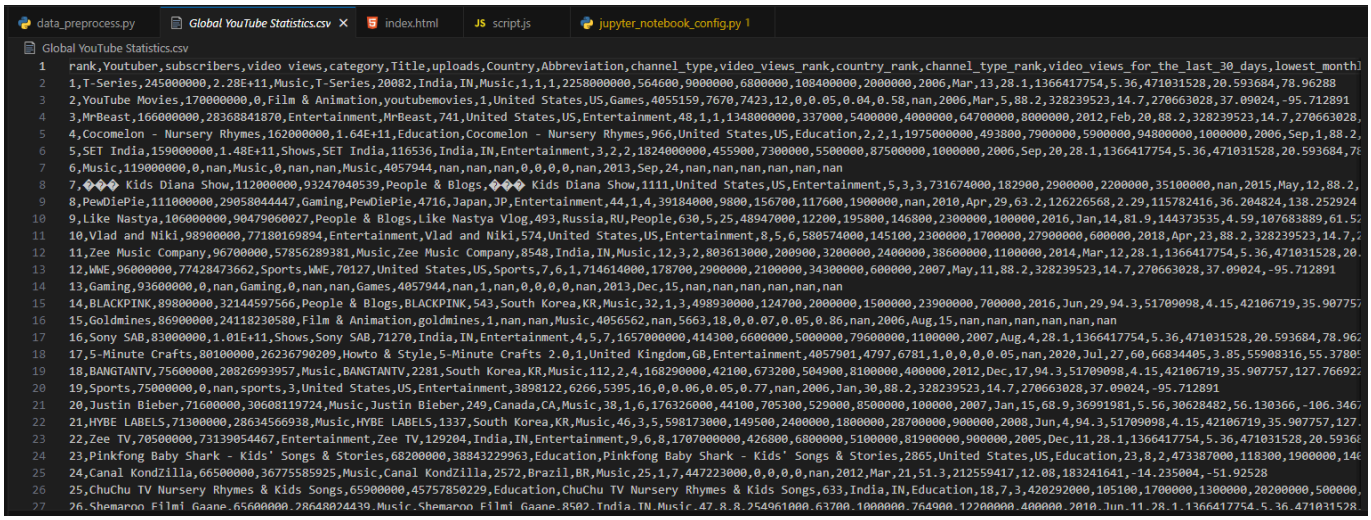
We have obtained our data from Kaggle. The youtube dataset contains information about the data of top 1000 youtubers and contains fields like rank, id, sub count, view count, category, number of uploads, country, monthly and yearly earnings, date of creation which we believe should be good for this project. The link to the dataset is https://www.kaggle.com/datasets/nelgiriyewithana/global-youtube-statistics-2023/data

5. Data Processing

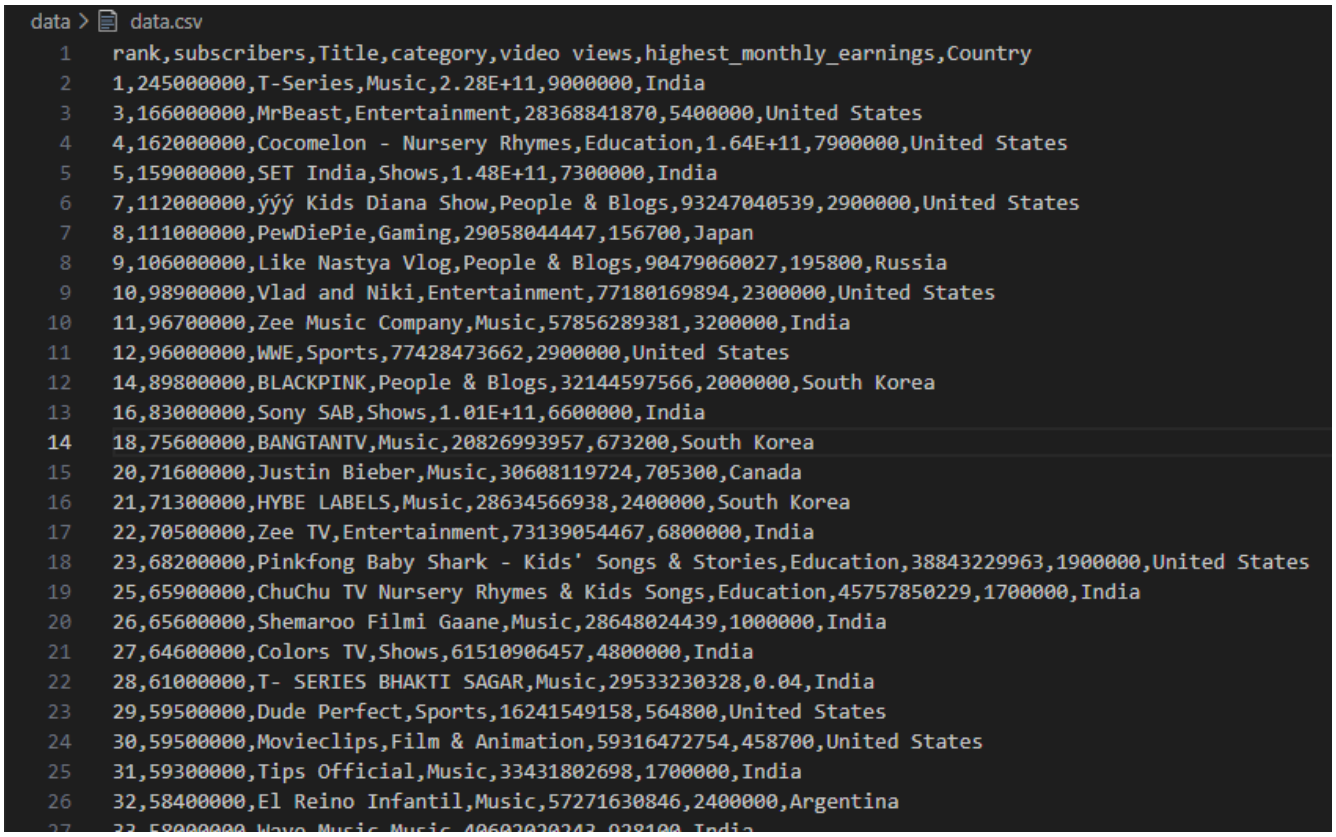
For data processing, we ran a python code which helped us get the required columns for the visualization from the original dataset we had. We had to clean and remove the rows that have invalid or garbage data in it. After writing the code accordingly and doing the data processing, we have ended up with the csv we are satisfied with and can use for our visualization. Here is the snapshot of the python code we ran for the data processing and cleaning

```python
import pandas as pd
import os

# Load the CSV file with the specified encoding
df = pd.read_csv("Global YouTube Statistics.csv", encoding='ISO-8859-1')

# Select the required columns
columns_to_keep = ['rank', 'subscribers', 'Title', 'category', 'video views', 'highest_monthly_earnings', 'Country']
processed_df = df[columns_to_keep]

# Print the number of rows before filtering
rows_before = processed_df.shape[0]
print(f"Number of rows before filtering: {rows_before}")

# Drop rows with specified conditions
filtered_df = processed_df[
    (processed_df['video views'] > 0) &
    (processed_df['highest_monthly_earnings'] > 0) &
    (processed_df['category'].notna()) & (processed_df['category'] != '') &
    (processed_df['Country'].notna()) & (processed_df['Country'] != '')
]

# Clean the 'Title' column by removing unwanted characters
filtered_df['Title'] = filtered_df['Title'].str.replace(r'\[.*?\]', '', regex=True)  # Remove anything in brackets
filtered_df['Title'] = filtered_df['Title'].str.replace(r'[Ã¯Â¿Â½]', '', regex=True)  # Remove specific unwanted characters
filtered_df['Title'] = filtered_df['Title'].str.strip()  # Strip whitespace from the ends
```

We did run into a few problems using this code which I'm gonna discuss more about in the visualization design section. But for now, this code did a great job in securing most of the data we wanted perfectly. Here is the snapshot of the original dataset we used for this visualization.



This data set contained some garbage and invalid data as you can see from the snapshot. The dataset is also quite extensive and contains about 28 columns of information which we won't be using for our visualization project. We picked the columns which we feel we will use for our project but we can always go back and get more columns to work with if we want to add more visualizations apart from the planned ones. Here is the snapshot of the processed data after deciding what we want and running the code.

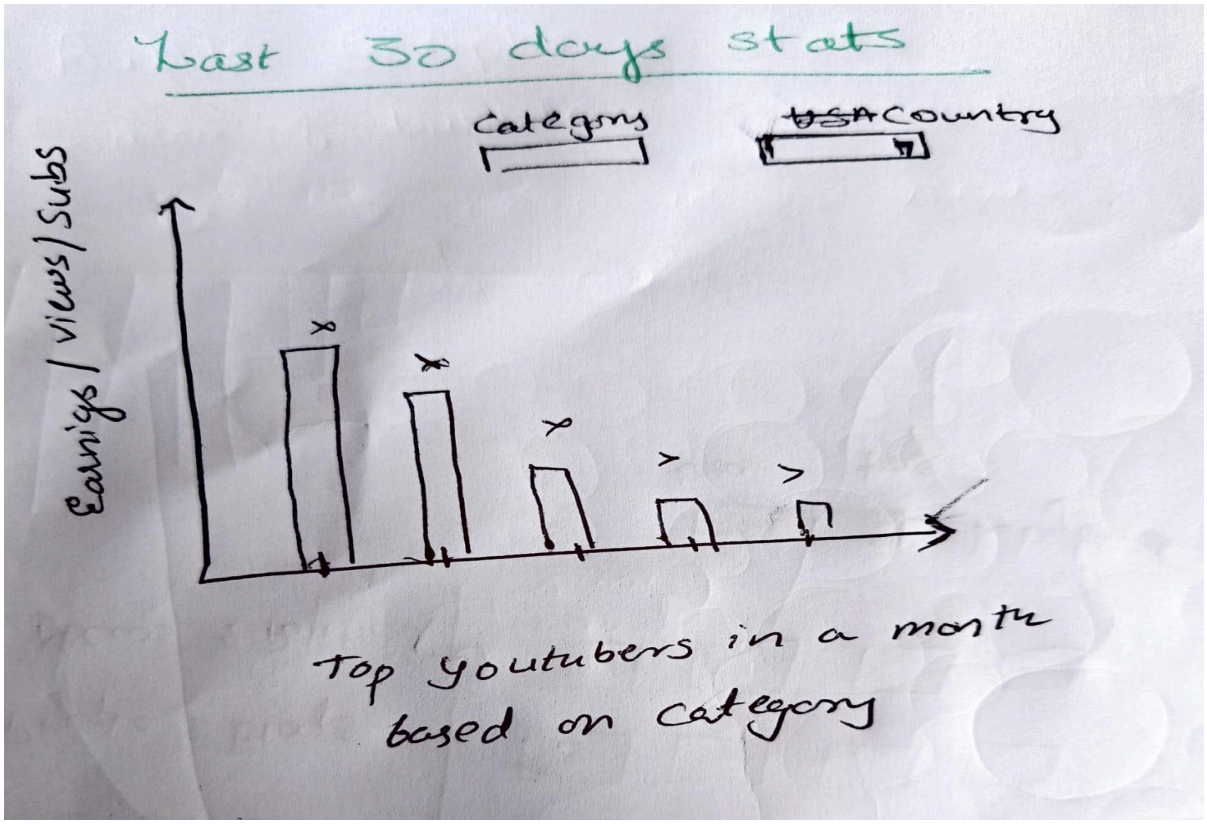


6. Proposed Visualization Design

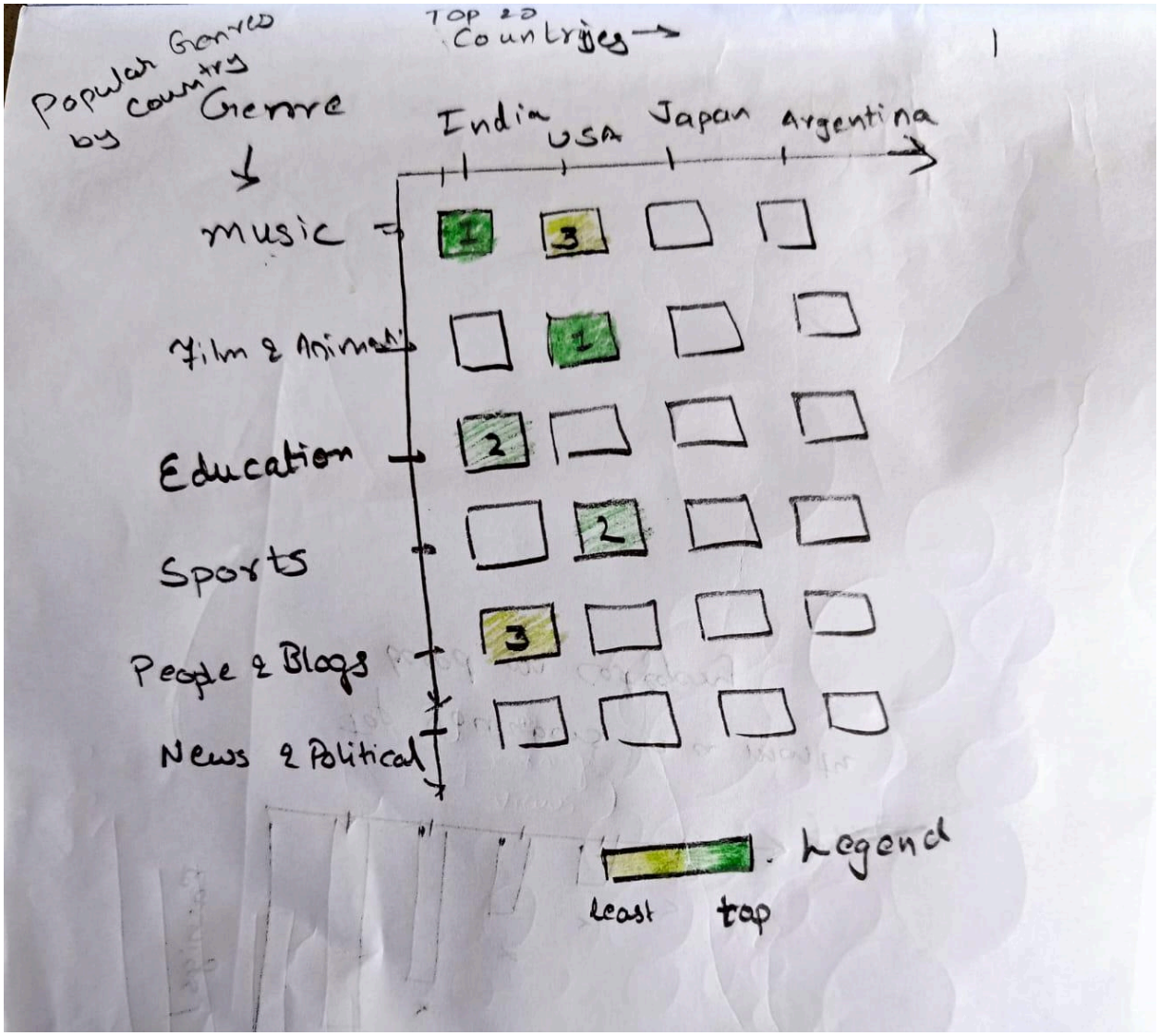
We aim to keep the page minimalistic to ensure a clean, user-friendly experience. Instead of overwhelming the user with multiple pages, our design will focus on a single, dynamic page that adapts based on user input.

1) Upon visiting the site, users will be directed to the main page, where they will be greeted with a world map and a bar graph displaying the top 10 YouTubers in the world(by default). This graph can be dynamically sorted by either views or subscriber count and by countries as well. When a user selects a specific country on the map, the graph will update to show the top 10 YouTubers from that country, with the ability to sort them by views or subscriber count, as desired.
   The world map can be used as a heatmap where the country's color intensity is proportional to the category(views/subscribers/earnings).

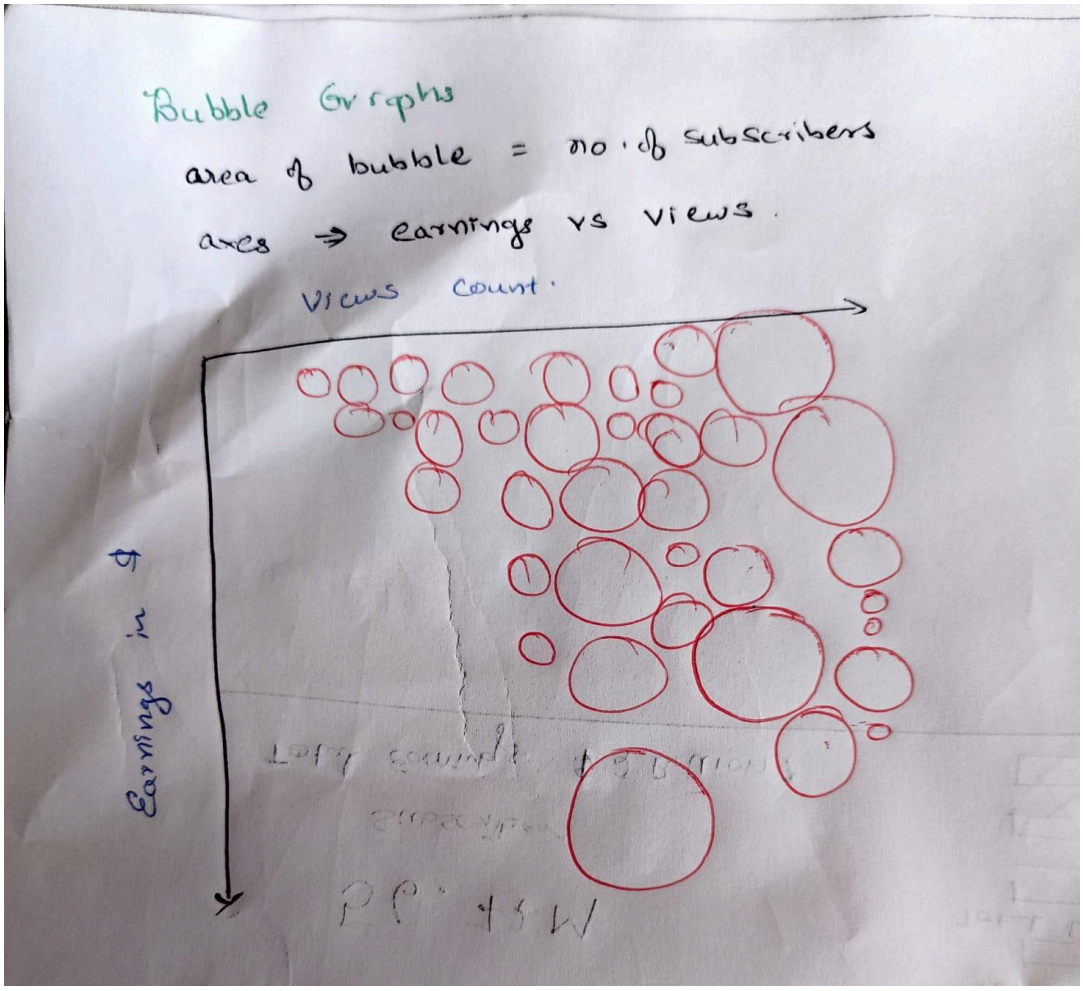
2) Below the main graph, we will display a second graph focused on recent trends, highlighting changes in subscriber counts and video views over the past 30 days. This will help uncover current content consumption trends either globally(by default) or for the selected country on the world map.



3) At the bottom of the page, we will feature a heat map that shows the popularity of different YouTube channel categories across countries. This visualization will reveal which types of content creators are most popular in specific regions.



Optional : We aim to implement a bubble graph where the size of the bubble indicates the subscriber count, axes represent Earnings vs View counts.



7. Proposed Must Have Features

- Interactive heatmaps and charts for visualizing country-specific YouTube stats.
  The world map should function as a heatmap where the color intensity reflects a selected metric (views, subscribers, or earnings).
- Filter options to select categories, countries
  By default, the bar graph displays the top 10 YouTubers globally.
  The graph should allow sorting by views or subscriber count.
  When a country is selected on the map, the graph should update to show the top 10 YouTubers from that country.
- Recent Trends Graph
  Display changes in subscriber counts and video views over the past 30 days.
  By default, show global trends, but when a country is selected, it should update to reflect trends for that specific country.
- Category Popularity Heatmap

A heatmap at the bottom of the page that visualizes the popularity of different YouTube categories (e.g., Gaming, Education, Music) across countries.

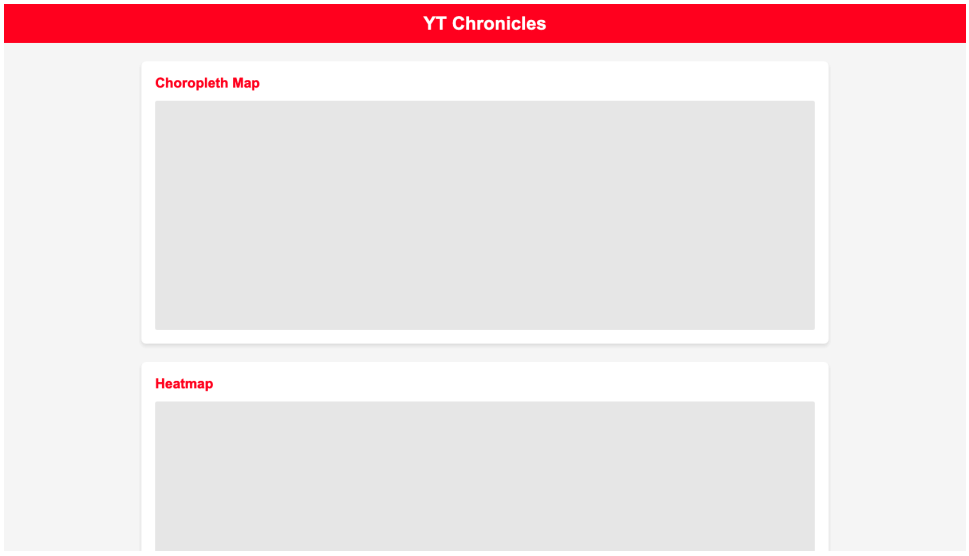
8. Proposed Optional Features

Here are some additional features we believe would enhance our project, though we recognize they may be challenging to implement within our given timeframe. Nevertheless, we will make an effort to incorporate them to further enrich the project

- We aim to create individual profiles for the top YouTubers, including their profile picture and key information displayed in a dashboard format, allowing users to quickly learn more about them.
- We want to add a feature that randomly recommends a YouTuber to watch, based on the country and category the user selects.
- We plan to leverage unused data, such as population, unemployment rate, and gross tertiary education enrollment, to find correlations with YouTube metrics and create insightful visualizations that reveal deeper patterns.
- Have a dark mode option for better user experience.

9. Implementation Process Stages

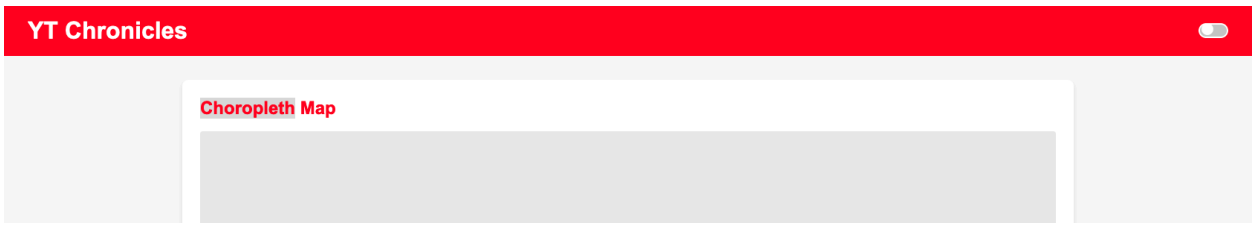
To start off, we wanted to have a skeleton website to work with and add our visualizations to which we could refine going on and rearrange our visualizations in. For the starting skeleton website design, we decided to keep it simple and go with the colors Red, White and Black primarily as those colors represent Youtube as a website. We plan to keep this color scheme going forward playing with their hues to make it visually appealing. Here is the screenshot of the initial skeleton website we made to work with



We have added names of the visualization we planned to and allocated space for them too. This is merely for our convenience for now and the website will be redesigned without those names and blocks on them in the future to make it visually appealing overall. For now, this is the canvas we are planning to go ahead and work with.

Our next goal is to add a dark mode feature to the website so we promptly did that and also made a few changes in the placement of text in the website. It looked something like this during this stage.

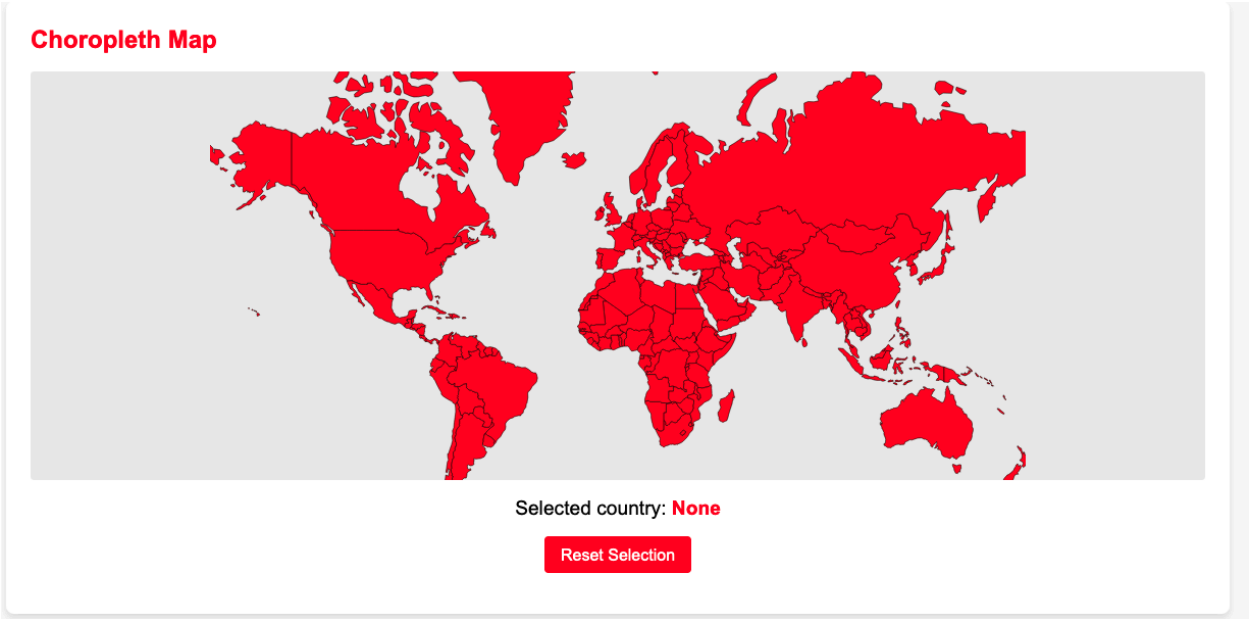
The issue we ran into here is that the dark mode button in the top right merges with the header background so we have to add an outline for the button to make it more visible like this.



After that is done, it is now time to work on the actual visualizations and we decided to start with the World Map visualization as we believe it is the most crucial and integral visualization in the whole webpage and making it work would mean other visualizations that are going to be integrated to it in the future are going to work fine too.

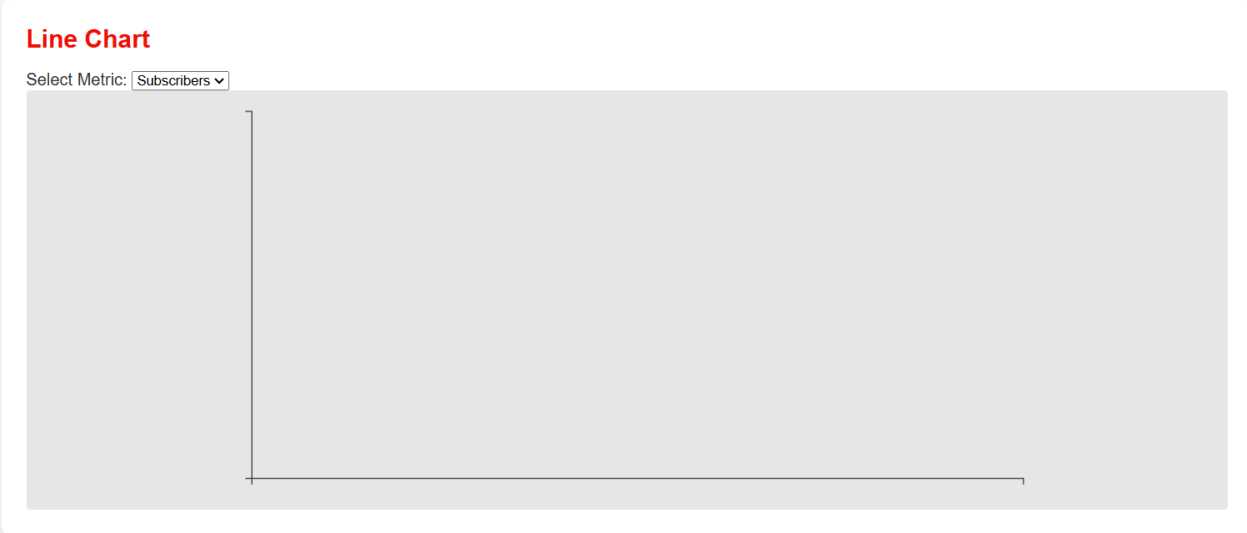
To start with our world map visualization, i first visited the website https://geojson-maps.kyd.au/ which has been really helpful to me in downloading the world map in a .geojson format. After I have downloaded it, it is time to map it properly, give it color and arrange it into our website.

After our initial steps the visualization looked something like this. Though it is not perfect as the visualization exceeded the space allocated to it, we are happy that the visualization is working perfectly. We also added a selected country option to see what country we have selected at the moment. The next step obviously is to get our visualization into the required space. After further coding and tweaking this is what we ended up with.

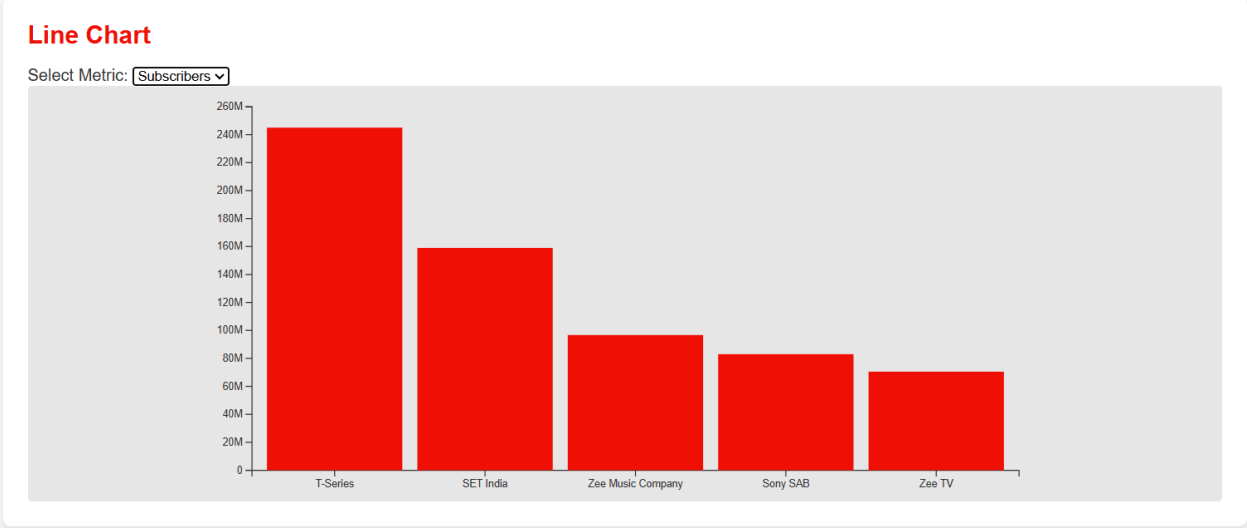


We were able to get the map properly into the block and also added a reset selection button to get back to none. This will be helpful to us when we start visualizing our line chart as setting the option to none would give us the top youtube channels in a set category in the entire world.

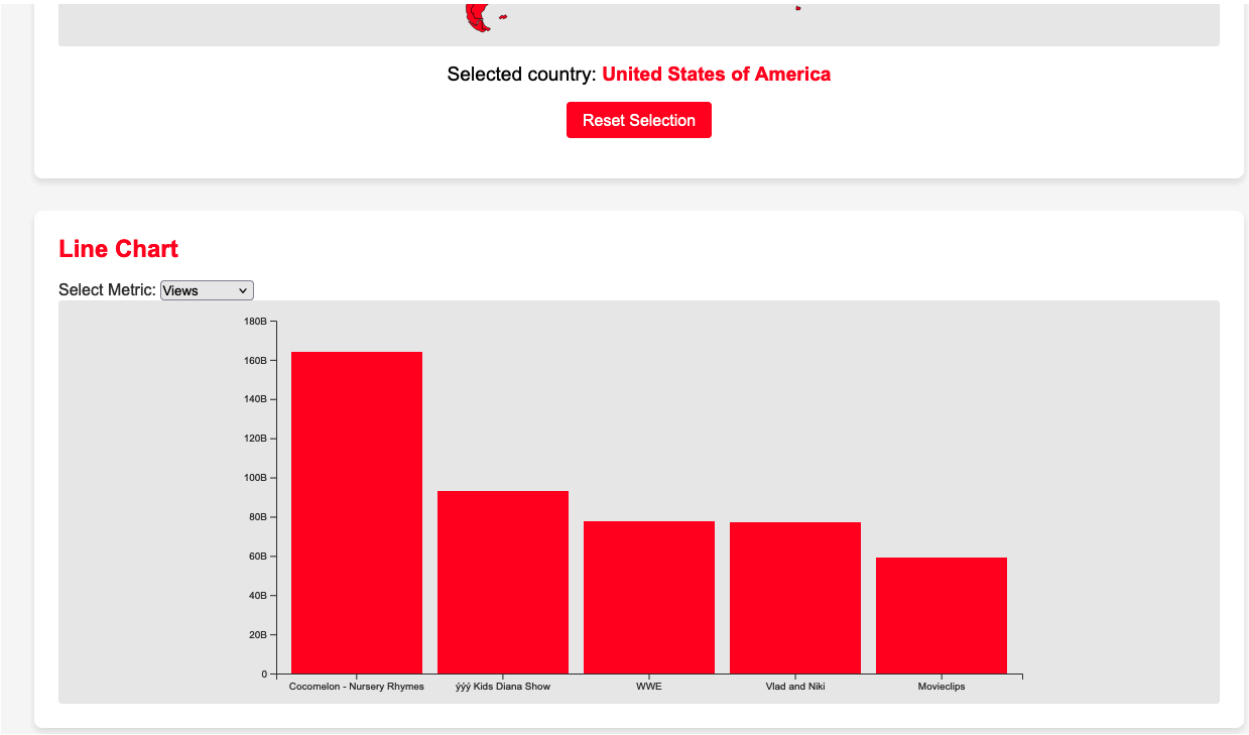
Now, the next step is to work on the line chart and this is where we ran into our first hurdle. The map we made has the location of USA as United States of America while our data set has it has just United States making it not being able to pull the data properly from our dataset showing something like this



But the visualization is coded and implemented perfectly in other cases where the country matches the dataset entry perfectly. So when you select India, it works properly showing something like this



We have worked on our dataset as well as the code to iron out these errors so that the data shows perfectly for all the countries as expected. After making a few changes we are able to output the desired output we need just like this



The line chart is now working perfectly. We have included three metrics to sort and select the youtube channels which are Views, Subscribers and Earnings. The plan was to include top 10 youtube channel per metric on the line chart but on further discussion we have decided to bring the count down to top 5 as there is just not enough data in the dataset we decided to show top 5 channels in most countries other than USA and India.

9. Planned Features

As of this milestone, we are able to get the World Map and Line chart visualization working as we intended. The project will continuously be updated in the upcoming milestones with more planned visualizations and other features. Here is a list of visualizations and features that are planned for the final release.

- A heatmap showing the top countries and the popular categories just as shown in our planned visualization design section (Must have)
- Adding a bubble graph comparing earnings to views (optional)
- Redesigning the website and structuring the visualization properly to make it more visually appealing (Must do)
- Make the visualizations be able to convey much more information like making the color of the country red go darker to lighter depending on the subscriber count in a World Map (Must do)
- Adding tasteful transitions and effects to our visualizations (optional)

This section and the entire process book will keep getting updated as we add more visualizations and features to the webpage. Thank you!