Global YouTube Stats Dashboard

CS661: Big Data Visual Analytics

Team members: Armeet Luthra (200185), Ishaan Maheshwari (200454), Jatin Chauhan (200469), Ojsi Goel (200653), Rajarshi Dutta (200762), Ritam Jana (200798), Sourit Saha (200998), Sushmita (201027), Yash Goel (201142)

Member emails: armeet20, ishaanm20, jatinc20, ojsig20, rajarshi20, ritam20, sourits20, sushmitag20, yashgoel20

1. Introduction:

In the digital age, YouTube has emerged as a dominant platform for both creating and consuming content, exerting significant influence on public opinion, entertainment, and marketing strategies. The extensive data generated by YouTube serves as a valuable resource for understanding contemporary digital culture, user behavior, and content trends.

This analysis centers on the YouTube dataset available on Kaggle, which contains a wealth of information on trending videos. The dataset encompasses various features such as top creators' subscriber counts, video views, upload frequency, country of origin, earnings, and more. This data provides a snapshot of the content that resonates with the global YouTube audience, offering an opportunity to extract insights into viewer preferences and content performance. Our main objectives in this analysis are to:

1. Gain an understanding of the characteristics of trending YouTube videos and identify patterns and trends in viewer engagement.
2. Explore the impact of various factors such as video category, location, and content type on video popularity.
3. Utilize a range of data processing techniques, including cleaning, transformation, and exploratory data analysis, to ensure the integrity and quality of the data, facilitating a robust and meaningful analysis.

The insights derived from this analysis are valuable for content creators, marketers, sociologists, and platform developers, providing a window into the rapidly evolving landscape of online video content. By comprehending the dynamics of content preferences on YouTube, stakeholders can make informed decisions, tailor their strategies, and potentially become next top content creator.

Dataset Description

The chosen dataset is a secondary dataset sourced from Kaggle about Global YouTube Statistics. The data set has key features such as:

- **Rank**: Position of the YouTube channel based on the number of subscribers
- **YouTuber**: Name of the YouTube channel
- **Subscribers**: Number of subscribers to the channel
- **Video views**: Total views across all videos on the channel
- **Category**: Category or niche of the channel
- **Uploads**: Total number of videos uploaded on the channel
- **Country**: Country where the YouTube channel originates

- **Channel_type**: Type of the YouTube channel (e.g., individual, brand)
- **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
- **Earnings**: estimated earnings from the channel
- **video_views_for_the_last_30_days**: Total video views in the last 30 days
- **subscribers_for_last_30_days**: Number of new subscribers gained in the last 30 days

The dataset has other features but these are the main focus of our data analysis.

2. Tasks and methodology:

Here the uses of the dataset are discussed through the various visualisation techniques and analysis that we implemented:

1. **Data Pre-processing**

    - The initial steps for handling this dataset involve Data Cleaning, Data Transformations, Feature Extraction, and Plotting. We start by identifying and managing missing or null values. For instance, if there are videos lacking categories or titles, we need to decide whether to fill these gaps with default values or exclude these records entirely. Additionally, it's essential to detect and eliminate any duplicate records to ensure dataset integrity. Each column should be checked to ensure it has the correct data type (e.g., dates should be in a date/time format, and numeric values such as views and likes should be integers or floats). Entries that seem incorrect, such as negative numbers in views, likes, or comment counts, should be addressed appropriately.

2. **Channel and country based overview**

    - Proposed solution: The first experiment that was carried out on data was analysis of parameters such as subscribers, views, uploads, and earnings based on regions of worlds or channels. Bar graphs were used to highlight the absolute number of these parameters for top 10 channels in each category. For example, to compare the top 10 channels of India based on the earnings, we select India from the region and yearly/monthly earning from the parameters. This will dynamically update the bar chart to show top 10 highest earning channels in India. Pie charts were used to depict pictorial representation of split of relative number of subscribers/views/uploads from different regions of the world.
    - Nature of Plot: **Bar Chart and Pie Chart**
    - Result: Bar graphs and pie charts were successfully plotted for above parameters. We notice that USA leads in the most number of cumulative YouTube subscribers and views as there is a strong internet infrastructure in USA for a large population. Also, the dominance of English as the main content creation language gives USA edge over other regions. India leads in the most number of video uploads because of increase in number of influencers in recent years,

however the number of views are less because these creators make content in local language.

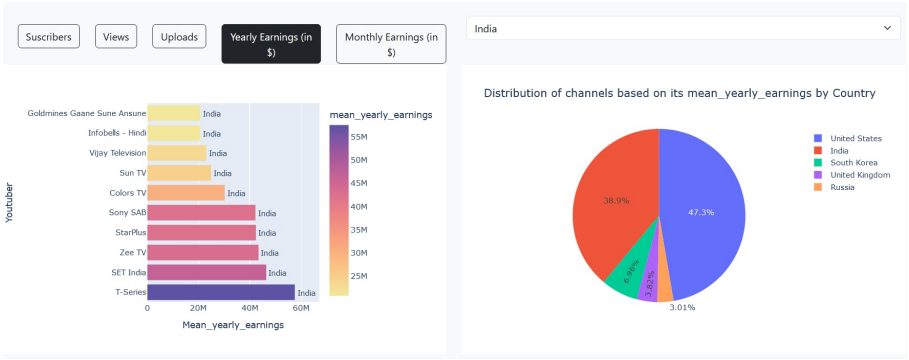

Figure 1: Channel and Country based overview

3. **Category based plots**

- Proposed solution: The "Category Based Plots" feature in the dashboard offers dynamic treemap visualizations allowing users to analyze YouTube statistics by country and category, aiding in trend identification and data-driven decision-making. It utilizes dropdown menus for user selection, dynamically generating treemaps based on chosen criteria. Data processing involves filtering, aggregation, and sorting to ensure relevance and accuracy in displayed statistics.
- Nature of Plot: **TreeMap Plot**

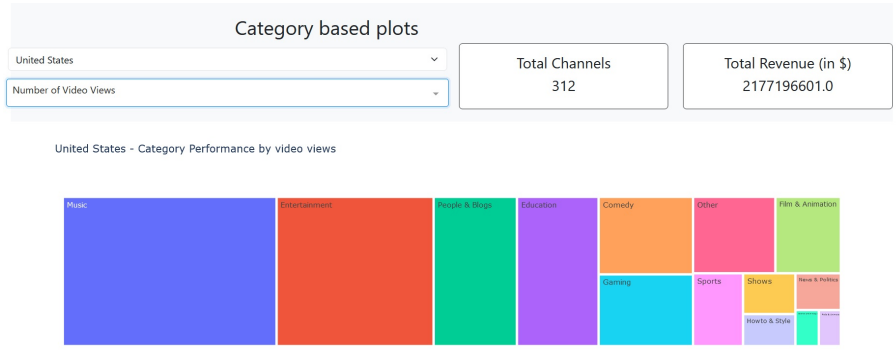

Figure 2: Category based plots

- Result: There are two dropdown boxes: first for country, second for video category. If we select one country and then one category, it will show a treemap plot. There are 14 boxes that show selected category feature in that country in ascending order. We have selected the following categories: 1. number of subscribers, 2. number of video views, and 3. monthly revenue (in $). For example, If we select India and Number of Subscribers, the treemap will show the most subscribed categories in India like Music, Entertainment, Education, Gaming etc.

4. **Distribution analysis**

- Proposed solution: The count of the number of subscribers, the number of views, earnings and other metrics for different channels that could follow different trends. So, it was proposed to plot a frequency distribution plot for such metrics to identify whether any known distribution such as normal distribution would fit the trend.
- Nature of Plot: **Histogram**
- Result: On visually analysing the plots, we can say that most YouTube channel have a particular range of subscribers, views or earnings and extreme value of these parameters are associated with only few channels. The trend is described by a truncated normal distribution.

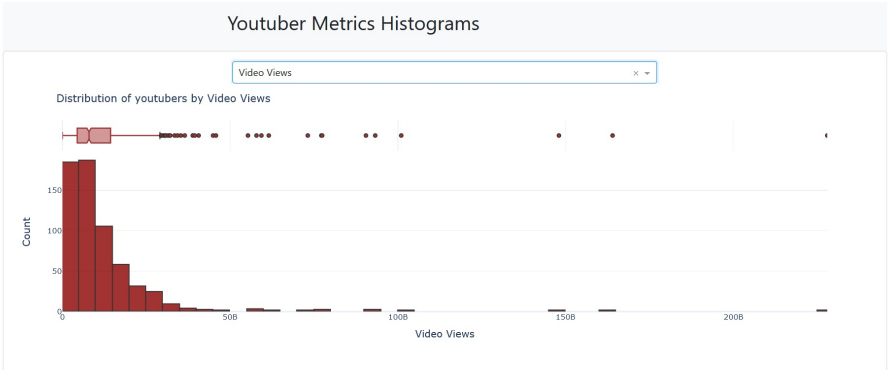

Figure 3: Distribution analysis

5. **Correlation analysis**

- Proposed solution: How the various metrics are related with each other can help understand the user patterns. For example, if the number of subscribers for a YouTube channel are more, the earnings and views should also be higher. This trend can be analysed by observing a scatter plot, calculating the correlation coefficient, and fitting data using linear regression.
- Nature of Plot: **Scatter Plot**

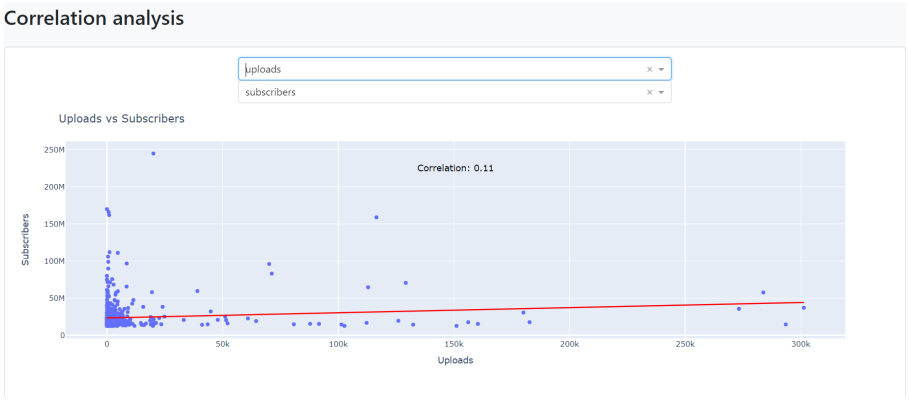

Figure 4: Correlation analysis

- Result: We can observe the strength of relation between the different pairs of metrics by the Pearson correlation coefficient. Positive and high value denotes

that if one metric increases in value the other metric also increases strongly. If it is negative, on increment of one metric, other metric shows a decrease.

6. **Annual trend analysis**

   - Proposed solution: The proposed solution for this project involves the development of an interactive dashboard tailored for YouTube data analysis. This dashboard will offer users the ability to choose specific channel categories, such as music, education, or entertainment, and conduct trend analysis based on annual data. The primary focus will be implementing user-friendly features that allow for seamless navigation and exploring YouTube trends. Additionally, the solution will integrate advanced time-series analysis techniques to uncover insightful patterns and fluctuations in video views over time within the selected channel categories.
   - Nature of Plot: **Line Chart**
   - Result: The report will showcase the interactive dashboard, emphasizing its functionalities, visualizations, and user interface to enhance user experience. Moreover, the results section will delve into insights gleaned from the time-series analysis of YouTube data, discussing notable trends, seasonal variations, and performance patterns across diverse channel categories. These insights will inform actionable recommendations aimed at optimizing content strategies and improving audience engagement on YouTube, providing stakeholders with valuable insights for informed decision-making and strategic planning.

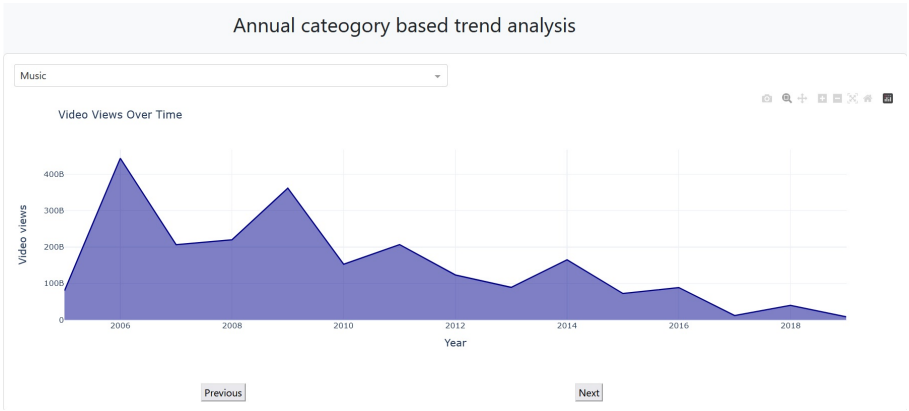

Figure 5: Annual trend analysis

7. **Region based analysis**

   - Proposed solution: The experiment encompassed the comprehensive dataset, delving into understanding of content consumption through an array of metrics such as subscriber count, video views, new subscribers within a 30-day window, new video views within the same period, and upload frequency. Geo-spatial plots have been used to visually map YouTube channels worldwide. Each point on the map was color-coded by category, with its size indicative of the magnitude of the corresponding metric.

     Hovering over a point gives information about the content creator, like channel name, subscribers, category, annual earnings, and creation date. This information, with an overlap of location, gives valuable insights into performance

trends and audience preferences to the potential content creators. Metrics like new subscribers and video views within 30 days can equip them with prevailing audience preferences and emerging trends.

Due to the non-availability of accurate location of channels, the latitudes and longitudes have been randomly sampled within the geometry enclosed by the country borders.

- Nature of Plot: **Geo-spatial Plot**
- Result: In the USA, according to video views, after the entertainment and music category, channels from education, gaming, and people and blogs category could be potential exploits. A lot of channels are observed in this category, and they have considerable video views. This reflects the strong interest of subscribers in educational, leisure, and friendly content. In India, on the company levels, entertainment, music and news, and politics dominate YouTube. On an individual level comedy, education, and people and blogs are famous. Within the country, many regional channels are famous, which indicates a strong preference for content in native languages.

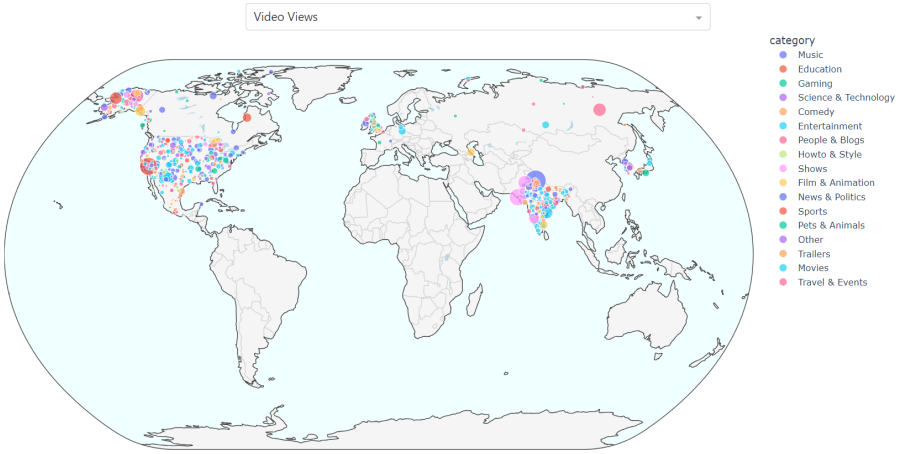

Figure 6: Region based analysis

8. **Dashboard UI components**

- Proposed solution: The primary objective of this dashboard is to provide users with an intuitive and interactive platform for visualizing and analysing Global YouTube Stats.
  - Front End: **Dash by Plotly** for creating interactive components.
  - Styling: **Dash Bootstrap Components** to utilize Bootstrap's grid system and responsive design.
  - Data Visualization: **Plotly** for dynamic and interactive charts. Display raw data that can be sorted and filtered according to user preference.
  - Charts and Graphs: Include a variety of charts (line, bar, pie, scatter, etc.) to depict trends, distributions, and relationships in the data.
- Result: After implementation, the dashboard should provide the following capabilities:

- User Interaction: Users can interact with the dashboard to select different data views, filter results, and adjust parameters without page reloads.
- Data Insights: The dashboard should present data, allowing users to derive insights quickly and efficiently. For example, changes in YouTube subscribers over time or the impact of marketing strategies on sales.
- Aesthetic and Usability: With Bootstrap, the UI should be clean and navigable, enhancing the user experience with an intuitive layout and easy-to-understand components.
- Performance: Data visualizations should load swiftly, and interactions should not cause significant delays, ensuring a smooth user experience.

9. **Web Deployment**

   - Our dashboard, which is a **flask** application, was successfully deployed by Gunicorn in **Render** which is a hosting platform. In the Render dashboard, you specify how your application should start. This is often done via a Procfile or directly in the settings where you set a command like `gunicorn app:server` (assuming `app.py` is your file and server is your **Flask** server variable).
   - The link to the dashboard is Global YouTube Stats Dashboard

5. Conclusion:

1. **Successful analysis:** The project adeptly employed statistical analysis to delve into YouTube's trending video data, unveiling crucial patterns in viewer engagement and video popularity.

2. **Interactive Dashboard:** A dynamic dashboard was crafted using Dash and Plotly, offering a robust tool for visualizing the mechanisms driving video success on YouTube.

   The link to the dashboard is Global YouTube Stats Dashboard

3. **Insightful Findings:**

   - Identified notable trends and patterns in trending YouTube videos across diverse regions and channel types.
   - Highlighted the influence of factors like video category, publish time, and content type on a video's popularity.

4. **Value to Stakeholders:**

   - Delivered valuable insights for content creators and marketers to enhance their strategic planning.
   - Provided a valuable resource for sociologists and platform developers keen on understanding digital culture and user behavior.

6. Link to source code:

All the codes and dataset have been uploaded in the following repository: Rajarshi1001/CS661_Project .

Work Distribution

Task	Member(s)
Project Ideation	All Members
Data Preprocessing	Rajarshi, Armeet, Ojsi
Channel and country-based analysis plots	Sourit, Ishaan
Category-based plots	Rajarshi, Sourit
Distribution analysis and related plots	Armeet
Annual channel-based trend analysis and plots	Yash, Jatin
Correlation analysis and related plots	Armeet, Sourit
Region-based statistics and related plots	Ojsi, Sushmita
Dashboard UI components	Ritam
Web deployment	Rajarashi

Table 1: Work Distribution

References

[1] Global YouTube Statistics 2023(link)