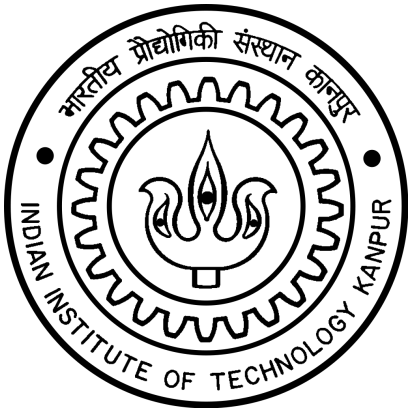
Global Movie Trend Analysis via TMDB

Project Report for CS661:Big Data Visual Analytics
2024-2025 Summer Term

Members:	**Email ID:**
Om Bhartiya	ombhartiya23@iitk.ac.in
Aryamann Srivastava	aryamanns23@iitk.ac.in
Saaumitra Raaj	saaumitra22@iitk.ac.in
Kshitiz Tyagi	ktyagi23@iitk.ac.in
Tejas Shrivastava	tejass23@iitk.ac.in
Aryan Deo	aryandeo23@iitk.ac.in
Swarnim Verma	swarnimve23@iitk.ac.in
Harshita Awasthi	harshitaa23@iitk.ac.in



Department of Computer Science & Engineering
Indian Institute of Technology, Kanpur

Contents

Chapter 1

Executive Summary

This report details "Global Movie Trends Visual Analytics Dashboard", an interactive visual analytics project developed using Python, Pandas, and Plotly within Jupyter. The project provides a comprehensive interface for exploring diverse facets of the global movie industry using the TMDB Movies Dataset sourced from Kaggle, which contains detailed information on approximately 930,000 movies spanning more than a century of film production. The visual analytics interface integrates multiple dimensions of the dataset, covering production budgets, box office revenues, genres, countries of production, production companies, runtimes, audience ratings, and popularity scores. Through a carefully designed series of interactive visualizations, users can analyze movie data temporally (year-wise production and popularity trends), comparatively (across genres, countries, and production companies), and relationally (exploring correlations between budget, revenue, runtime, and ratings).

Key features include interactive maps and dynamic charts such as bar, line, pie, sunburst, scatter, Sankey, treemap, heatmap, along with data filtering controls and integrated machine learning techniques such as linear regression and confidence intervals. This report outlines the project's objectives, the methodologies used for data preprocessing and feature selection, the design rationale behind each visualization, and a detailed explanation of the insights derived from them. Each plot is justified not only by the nature of the data but also by the story it conveys about the evolution of the global film industry. The complete source code, notebook, and interactive plots are hosted on Github for easy reproducibility and demonstration.

Please note that this report contains technical details, including CSV column names and code snippets, to help readers connect the visual output with the underlying implementation logic.

The GitHub Repository Link for the project is - Click Here

Chapter 2

Introduction

2.1 Project Overview & Goals

The global film industry is a powerful cultural and economic force that influences societies worldwide. Understanding its production trends, financial dynamics, audience reception, and evolution over time is valuable for filmmakers, producers, media analysts, and researchers alike. However, despite the availability of vast movie datasets, meaningful exploration often remains limited by the challenges of processing, cleaning, and visualizing such large and diverse data. The "Global Movie Trends Visual Analytics Portal" project was undertaken to address this challenge. The primary goal is to develop an intuitive, interactive visual analytics interface that allows users to explore detailed movie industry data in a clear and meaningful way.

The project aims to:

- Provide a centralized platform for accessing comprehensive movie statistics spanning over a century.
- Enable exploration of movie data across multiple dimensions — including time (year-wise trends), genres, production companies, production countries, audience ratings, budgets, revenues, and popularity scores.
- Facilitate comparative analysis between genres, countries, studios, and movie characteristics to reveal differences and similarities.
- Employ effective data visualization techniques to highlight patterns, correlations, and anomalies that may be hidden in raw tabular data.
- Empower users to draw insights into questions like how production budgets relate to revenue, which genres perform best, how audience tastes change over time, and which companies dominate the global box office.

2.2 Scope

The scope of this project encompasses the following key aspects:

- **Data Integration**: Loading, cleaning, and standardizing a large, single-source movie dataset (TMDB Movies Dataset) containing approximately 930,000 records. This includes handling missing values, correcting date formats, filtering unrealistic entries (such as future release years), and deriving additional features such as release year and main genre for deeper analysis.
- **Visualization Development**: Designing and implementing a range of interactive visualizations using Python, Pandas, and Plotly within a Jupyter environment. This includes area charts, line charts, scatter plots, box plots, bar charts, heatmaps, and sankey plots; each carefully chosen to reveal trends, distributions, and relationships in the data.
- **Interactive Web Interface**: Building a clear, structured and multi-tabbed Dash web interface that hosts the visualizations and provides preprocessing, plotting, and interactivity. Users can work through tabs and modify parameters (e.g., year filters) to dynamically adjust visual outputs.
- **Interactivity**: Enabling users to interactively explore the data by applying year filters, genre filters, and company filters within the code, allowing for flexible insights without needing to modify raw data manually.
- **Dashboard**: Preparing the core visualizations for easy integration into a web-based dashboard using Dash. This scope would support dropdowns, sliders, and user inputs to dynamically update plots.

2.3 Frameworks & Utilities Used

The portal and its underlying visual analytics pipeline are built entirely using the Python ecosystem.

Key technologies and libraries used include:

- **Google Colab**: A cloud-based Jupyter notebook environment that allows interactive coding, data cleaning, and visualization without local setup. It serves as the primary development and demonstration platform for the project.
- **Pandas**: The core Python library for data manipulation and analysis. Used extensively for loading the TMDB dataset, handling missing values, cleaning inconsistent entries, extracting new features (e.g., release year), and preparing data for visualization.

- **Numpy**: The fundamental package for numerical computations in Python. Utilized for array operations, type conversions, and efficient numerical handling within the dataset.
- **Plotly**: Interactive visualization libraries for generating high-quality, publication-ready plots directly in notebooks. Plotly Express is used for quick generation of area charts, line charts, scatter plots, box plots, and bar charts, while Graph Objects can be leveraged for more customized visualizations if needed.
- **Dash**: A framework by Plotly for building interactive analytical web applications, is planned as an optional next step to deploy the notebook visualizations as a standalone multi-tabbed web dashboard. This would enable broader accessibility, advanced interactivity, and a more polished user interface.

Chapter 3

Tasks

The web app is designed to facilitate the exploration and analysis of multiple aspects of the global movie industry. The core tasks enabled through the interactive visualizations include:

- **Production Volume Analysis:** This plot provides a clear overview of how the number of movies produced each year has changed over time, helping users identify major growth periods and historical shifts in production trends.

  A **histogram** is used instead of other plot types because it effectively shows the frequency distribution of movie releases across years, making it easy to spot peaks, declines, and overall production patterns at a glance.

- **Budget vs Revenue Analysis:** This plot enables users to explore the relationship between movie budgets and their box office revenues, highlighting typical returns on investment, blockbuster outliers, and less profitable ventures.

  A **scatter plot** is used because it clearly displays the distribution and spread of individual movies along both axes, making it easy to detect trends, clusters, and anomalies. The fitted linear regression line with a confidence interval helps illustrate the general correlation and expected range, providing deeper insight than simple aggregated statistics would.

- **Genre-Based Revenue Comparison:** This plot allows users to compare how different genres perform financially by visualizing revenue distributions and identifying consistently high-grossing genres versus more niche categories.

  A **treemap** is used because it intuitively represents hierarchical data and relative magnitudes in a compact space, making it easy to see which genres contribute most to overall revenue at a glance. Unlike bar charts or pie charts, the treemap efficiently shows both proportion and structure when many categories are involved.

- **Country-Level Genre Analysis:** Offer insights into how different genres are distributed among production countries, emphasizing the dominance of key film-producing nations and emerging markets.

  This is visualized using a **choropleth map** and a supporting **bar chart**, as a map clearly shows spatial patterns which a simple table or pie chart cannot highlight geographically.

- **Runtime vs Audience Rating Analysis:** Visualize the correlation between movie runtime and average user ratings, revealing audience preferences for narrative depth and film length.

  This uses a **scatter plot** with a fitted regression line, since scatter plots are ideal for showing relationships between two continuous variables, unlike bar charts which suit categorical data.

- **Top Production Companies Overview:** Highlight the leading production companies based on cumulative revenue, helping users understand the level of industry concentration and the market power of major studios.

  This is visualized using an interactive **donut chart** that displays each company's share at a glance. When a user clicks on a segment of the donut, it expands to reveal a detailed **bar chart** showing key metrics such as the company's top movies ranked by budget, revenue, or ROI, enabling deeper analysis of indivual studio performance.

- **Top-Rated Movies Trend Analysis:** Track how the top movies released each year change over time with reference to their budget, revenue, ROI, and popularity, indicating shifts in audience and critic tastes.

  A **bar chart** works best here to show trends over continuous time, which other charts would make too segmented.

- **Genre Evolution and Recent Trends Analysis:** Examine how the prevalence and financial performance of different genres have evolved over time, highlighting emerging trends and genre cycles.

  This is visualized using a combination of a **streamgraph** and a **sunburst chart**. The streamgraph clearly shows smooth temporal shifts and overlaps in genre popularity across decades, which would be difficult to interpret with static bar charts or line charts alone. The sunburst adds a multi-level breakdown, illustrating how main genres branch into sub-genres or distribute across periods, which a simple pie or bar chart cannot convey effectively.

- **Genre Dominance over time:** Assess which production companies lead in each genre by cumulative output and revenue, revealing genre-specific market leaders.

A **Sankey diagram** is best to show how companies flow into genres — unlike bar charts, Sankey charts make split and flow connections visually clear.

- **Evolution of Production for Companies/Countries within a Genre:** Track how individual studios or countries have increased or diversified their output in a given genre over time.

  A **heatmap** is used here because they reveal variations across time and categories together, which a line chart alone might not capture for multiple groups.

- **Movie Recommendations Based on Top Production Companies:** Suggest films to users by leveraging metrics such as budget, revenue, ROI, and popularity from leading studios, tailored to viewing preferences.

  A **ranked bar chart** makes it easy to compare exact figures, which would be tedious or unreadable in more abstract visuals.

Chapter 4

Data Sources and Preprocessing

4.1 Overview of Dataset

This robust visual analytics project relies heavily on clean, well-structured, and context-rich data. The dataset used in this project was sourced from a version of the **TMDB** open dataset from kaggle, comprising approximately ~930,000 movie entries. Each record provides extensive metadata for individual films, enabling detailed temporal and categorical analyses across multiple dimensions of the global film industry.

The project utilizes a single, consolidated CSV file named `TMDB_movie_dataset_v11.csv`, which serves as the master dataset for all visualizations and computations. This file includes a wide range of features, such as:

- **Title and Release Date:** Basic identifiers for each film, with dates parsed into a standardized datetime format for temporal grouping and filtering.
- **Genres:** A comma-separated list of genres associated with each film (e.g., `Action, Adventure, Sci-Fi`). This column was processed and exploded to enable genre-level aggregation and analysis over time.
- **Adult Content Flag:** A boolean field indicating whether the film is marked as adult content, enabling comparative analyses between general and adult film trends.
- **Spoken Languages and Original Language:** Useful for analyzing linguistic diversity and the prevalence of multilingual productions.
- **Runtime:** Duration of each film in minutes, used in descriptive statistics and for exploring relationships with genre and popularity.
- **Popularity Metrics:** Includes fields such as `popularity`, `vote_average`, and `vote_count`, offering insights into audience engagement and reception.

- **Production Companies and Countries:** Strings containing the names of production entities and countries, useful for analyzing trends by region or production house.

4.2 Data Loading and Initial Cleaning

The dataset was loaded into a Google Colab notebook using `Pandas`. Initial cleaning steps included:

- **Date Parsing and Extraction:** Converting the `release_date` column to a valid datetime type while gracefully handling missing or malformed entries. Creating a `release_year` column to enable trend and time series analysis.
- **Missing Values and Outlier Filtering:** Dropping rows with critical missing information (e.g., missing release dates). Removing unrealistic future release years (e.g., entries showing dates beyond the current year).

4.3 Feature Engineering and Standardization

Key preprocessing tasks included:

- **Derived Columns:** Extracting additional features such as the release decade or main genre (if genre lists were nested).
- **Numeric Conversions:** Ensuring fields like budget, revenue, and runtime are stored as numeric types with invalid entries coerced to NaN.
- **Handling Zero Values:** Filtering or flagging movies with zero budget or revenue for more meaningful financial plots.
- **Label Cleaning:** Ensuring text fields (e.g., production company, country) are consistently formatted for groupings and aggregations.

Together, these steps ensured the dataset was reliable, consistent, and ready for robust visual analytics.

Chapter 5

Framework's core Components

5.1 Framework: Dash

The current application is developed primarily as an interactive Jupyter notebook using **Google Colab**, with all visual analytics handled through `Pandas` and `Plotly`. Colab provides a cloud-based Python environment that requires no local setup and supports live code execution, inline visualization, and easy sharing.
Key advantages relevant to this project include:

- **Python Native:** Allows developers to use familiar Python libraries (Pandas, Scikit-learn, etc.) directly within the application logic for data processing and visualization.
- **Interactive Plots:** Plotly enables high-quality, interactive charts embedded within the dash app, allowing dynamic exploration of the data.
- **Dash Integration:** While the current version runs fully in Colab, the project is designed for easy migration to Dash, which will enable building a web application version of the notebook with richer interactivity and a multi-page layout.

When extended to Dash, the same Plotly plots and Pandas logic will form the core backend, while Dash will handle the user interface and callback-based interactivity.

Chapter 6

Overall Application Story

Imagine trying to understand the global movie industry — an industry that spans continents, decades, and hundreds of thousands of films in dozens of languages. Which countries have seen explosive growth in film production? Which genres dominate box office revenue? How have audience tastes shifted over time? How do budgets translate into box office success, and which studios consistently deliver hits? Raw spreadsheets alone make these questions hard to answer.

The **TMDB Global Movie Trends Visual Analytics Portal** was designed to cut through that complexity. It transforms a massive dataset into a dynamic exploration tool that reveals hidden patterns and connections in the world of film. Instead of static charts or simple summary tables, the portal offers an interactive experience. Users can start by tracing how the number of movies produced each year has grown, spotting the rise of emerging film industries and landmark spikes in production. They can compare budgets with revenues to understand how risk and reward play out across genres and time.

Curious about what genres audiences flock to? The genre-based visualizations break down revenue distributions, showing which categories consistently draw crowds. The country-level analysis shows how different markets contribute to global box office performance, highlighting the influence of Hollywood while also surfacing growing contributions from other regions.

Beyond the financials, the portal lets users investigate creative patterns. Does runtime affect audience ratings? Are longer movies rated more highly? Which production companies dominate the market, and how have their hits changed over time? Finally, by tracking metrics like popularity scores and average ratings for top movies year over year, the portal helps uncover how global audience tastes evolve — offering both a historical perspective and hints about where the industry might be heading.

Chapter 7

Tab-Specific Analysis and Features

7.1 Overview Tab

7.1.1 Functionality Overview

The Overview Tab provides a high-level entry point into the TMDB dataset. It highlights trends over time, geographic patterns, and general statistics to help users understand how the global film industry has evolved.

7.1.2 Core Visualisation: Streamplot and Sunburst charts

These visualizations together answer how the composition and structure of movie genres have evolved across decades. The streamplot reveals long-term trends in audience preferences by showing the rise or decline of genres like action, drama, or horror over time. Complementing this, the sunburst chart provides a multi-level breakdown of the genre hierarchy within the dataset, illustrating how main genres branch into sub-genres or how genres split across different decades.

7.1.3 Interactive Elements

- **Genre Selection Dropdown**: A dropdown menu lets the user choose specific genres to include or exclude in the plots. This updates visualizations like the Genre Streamplot and Sunburst Chart to reflect only the selected genres.
- **Year Range Sliding Window**: A dynamic range sliding window allows the user to filter movies by release year. Moving the slider updates all relevant plots, including the Number of Movies Per Year, Runtime vs Rating scatter plot, and any time series graphs.
- **Interactive Plot Hover and Click:** Many plots, such as the streamplot, provide hover tooltips that display detailed movie statistics. Clicking on a data

point can reveal additional metadata or highlight related entries.

7.1.4 Dashboard Screenshots

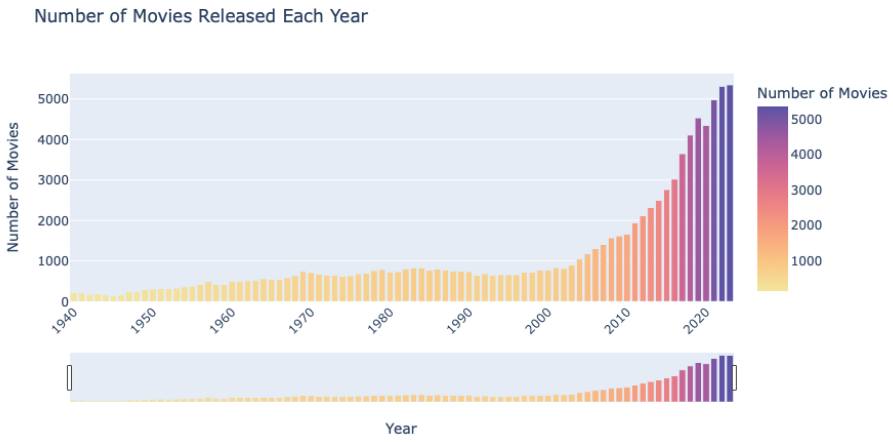


Figure 7.1: Histogram depicting total number of movies released each year over time.

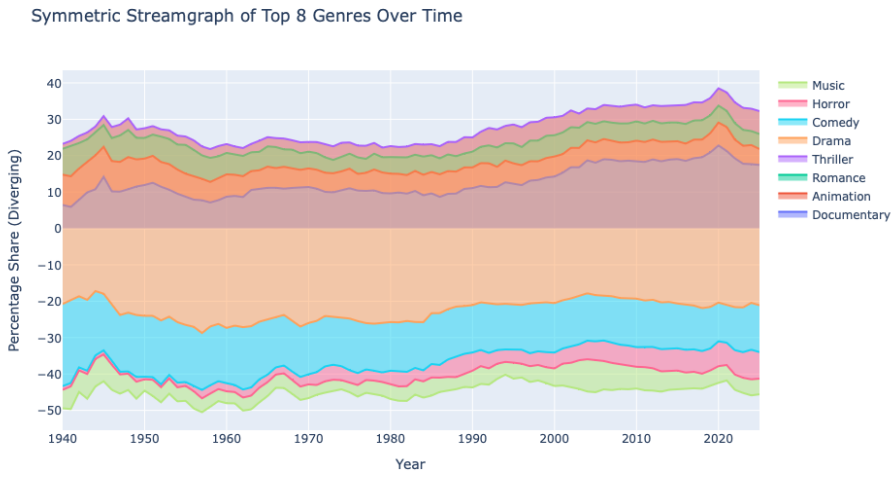


Figure 7.2: Streamplot showing percentage share of Top Genres with time

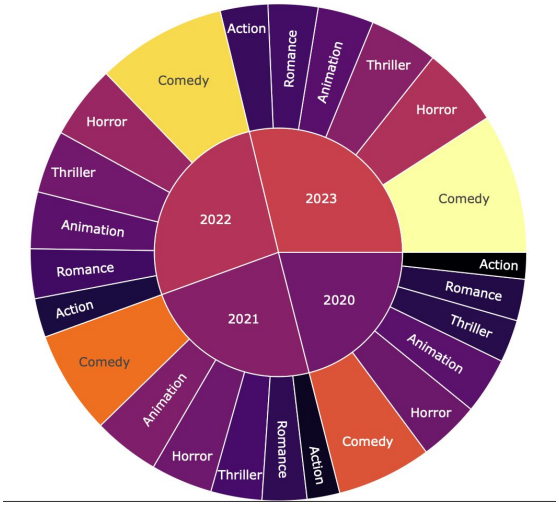


Figure 7.3: Sunburst of Genres

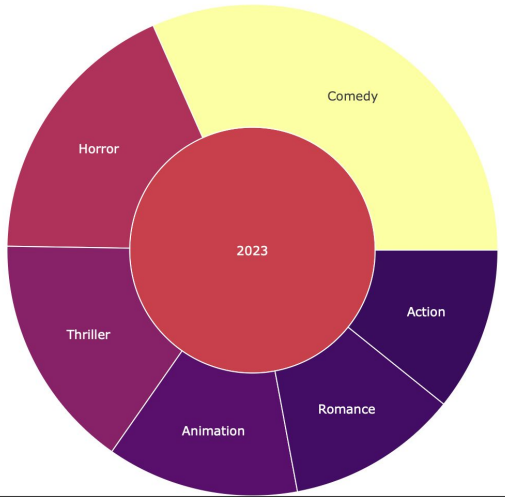


Figure 7.4: Year specific Sunburst

7.1.5 Observations

The dashboard visualizations together provide meaningful insights into global movie trends:

- The histogram clearly shows how the total number of movies produced each year has evolved over time, revealing periods of growth or decline in film production.
- The streamplot highlights the changing popularity of different genres, making it easy to identify genres that have gained or lost audience share across decades.
- The side-by-side sunburst charts give a detailed breakdown of genre distribution. The first sunburst displays the overall genre hierarchy, while the year-specific sunburst reveals how genre patterns shift in particular years.
- The word cloud visualizes the most frequent keywords or themes associated with the movies, providing an immediate sense of dominant trends, topics, or storytelling styles in the dataset.

7.2 Genre Tab

7.2.1 Functionality Overview

The Genre Tab allows users to conduct an in-depth, genre-specific exploration of the TMDB dataset. It enables filtering and analysis focused on a single genre at a time, helping users understand how production trends, revenue, budget, ROI, and geographic distribution vary within that genre. This view highlights how different production countries and companies contribute to the selected genre over time.

7.2.2 Core Visualizations

This tab includes visualizations specifically designed to uncover genre-level insights. **Heatmaps** display the distribution of movie counts across production countries and companies over time, helping detect dominant players and shifts in production focus. Bar charts and box plots illustrate key financial metrics such as budget, revenue, and ROI for the selected genre, highlighting high-performing segments and outliers. Together, these plots allow users to compare different facets of the genre's production and performance.

7.2.3 Interactive Elements

- **Genre Selection Dropdown:** A dropdown menu lets the user select the genre they wish to analyze. Changing the selection dynamically updates all relevant heatmaps and bar charts to reflect data for the chosen genre only.
- **Time Period Slider:** An interactive year range slider allows users to narrow the analysis to specific time windows. Adjusting the slider filters the dataset for the heatmaps and financial plots, enabling focused trend comparisons across decades.
- **Hover and Drilldown:** Plots like the heatmap support hover tooltips, displaying additional details such as exact movie counts or company names. Clicking on a heatmap cell or bar can reveal further breakdowns, such as a list of top movies for that cell's production country or company.

7.2.4 Dashboard Screenshots

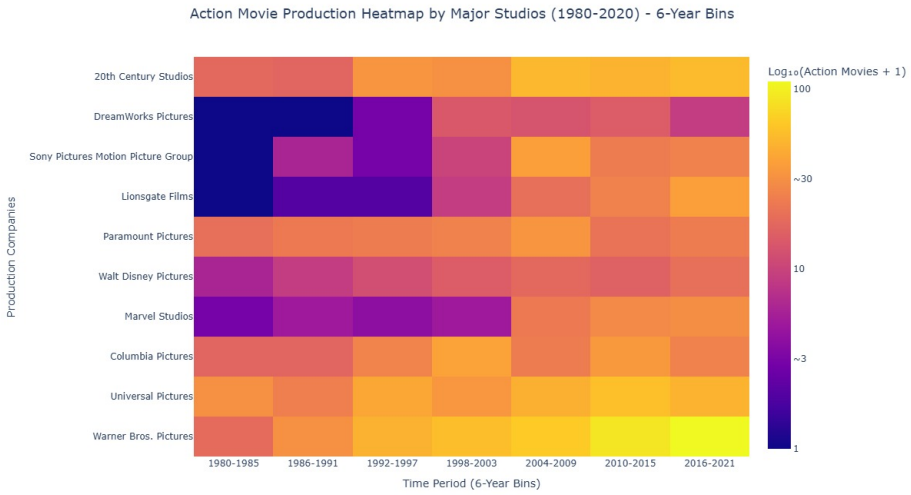


Figure 7.5: Heatmap for Movie production by Studios

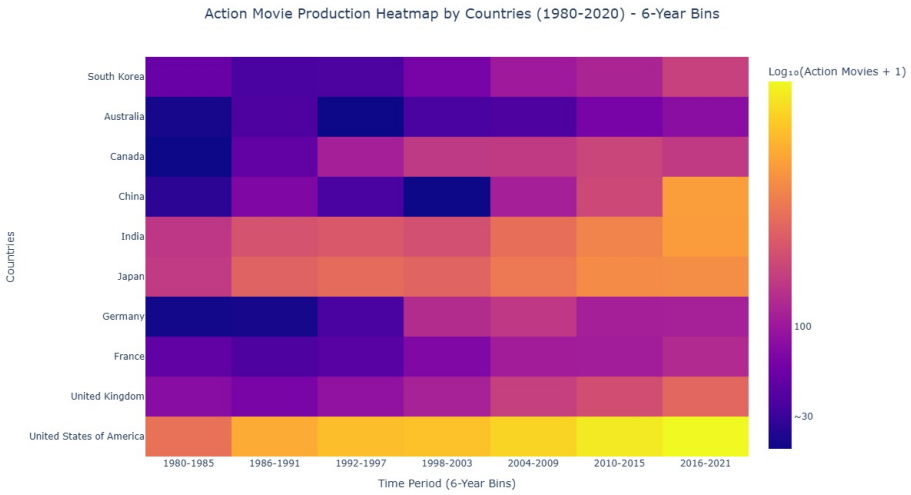


Figure 7.6: Heatmap for Movie production by Countries

Figure 7.7: Tree Plot showing revenue share by Genre

7.2.5 Observations

The heatmaps illustrate how movie production volume varies across both production studios and countries over time. The first heatmap clearly shows that certain major studios have consistently produced a high number of movies every year, while smaller studios appear intermittently, highlighting industry consolidation and the dominance of a few key players. The second heatmap shows the distribution of movie production among countries, revealing which countries are consistently major contributors to global film output and which regions have emerging or sporadic production footprints. The tree plot provides an intuitive visual of how revenue is split across different genres. It highlights genres that consistently capture the largest market share, such as action and adventure, while niche genres occupy smaller segments. This helps identify dominant revenue-driving genres and supports decisions on targeting audience preferences and investments.

7.3 Country Tab

7.3.1 Functionality Overview

The Production Country Tab focuses on analyzing movies produced within a specific country. It allows users to filter the dataset by a selected production country and explore key statistics such as top-performing movies, revenue, budget, and ROI within that country. This tab helps users understand how different countries contribute to global film production and which genres or movies stand out.

7.3.2 Core Visualizations

The centerpiece is an interactive geospatial plot that highlights the number of movies produced by each country worldwide. When a user clicks on a country, a detailed pop-up shows country-specific statistics, including total movies produced, average budget and revenue, and notable genres. Additional visuals such as bar charts rank the top 10 movies from the selected country by vote average, budget, revenue, or ROI, while treemaps break down genre shares within that country's production output.

7.3.3 Interactive Elements

- **Country Selection on Map:** Users can click on a country directly within the geospatial plot. This action triggers a pop-up with summary stats and updates all linked visuals for that country.
- **Ranking Criteria Options:** Users can switch between ranking top movies by vote average, budget, revenue, or ROI. This helps surface both high-budget blockbusters and hidden high-ROI successes.
- **Hover Details and Drilldown:** Hovering over the map or bar charts reveals tooltips with detailed stats, while clicking a bar can expand further information about the movie, such as release year or genre.

7.3.4 Dashboard Screenshots

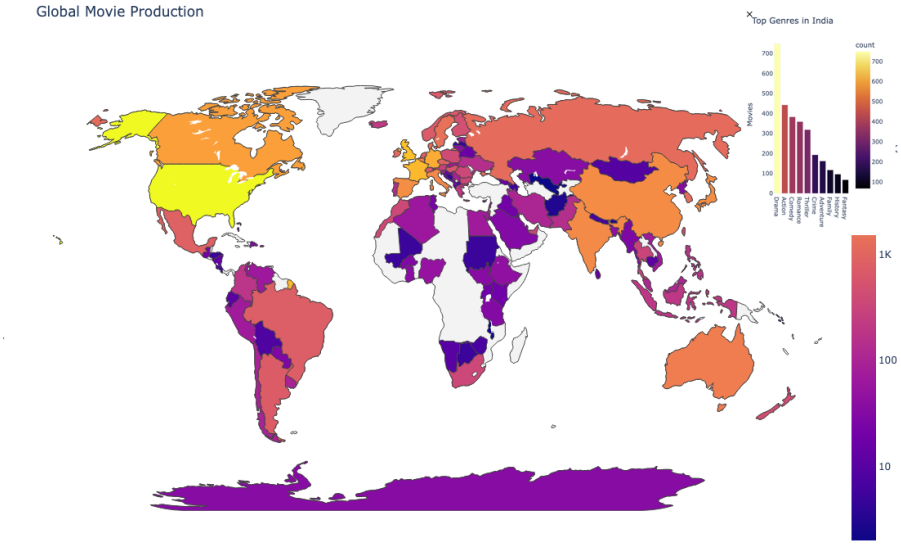


Figure 7.8: Geospatial plot with a pop-up when selected on a country (here, India).

Figure 7.9: Top Movies for a given genre(here, Action) by Revenue

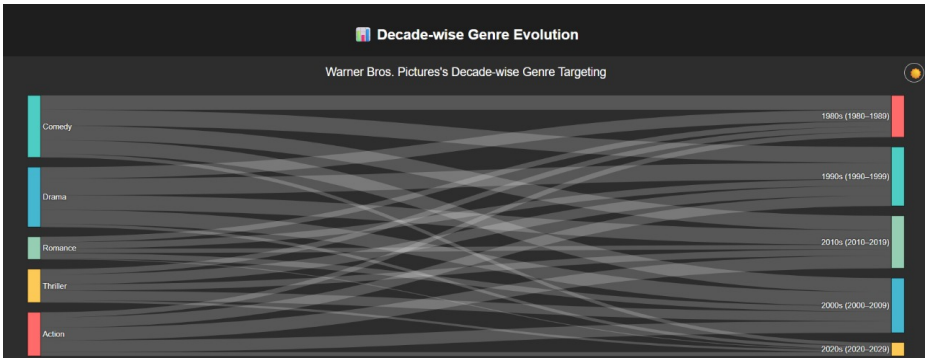


Figure 7.10: Sankey plot for Decade-wise Genre Evolution.

7.3.5 Observations

The geospatial plot provides an intuitive visual representation of movie production volume across countries, with interactive pop-ups (as shown here for India) offering quick access to key details such as total movies produced. This helps identify major film-producing hubs globally and reveals regional contributions to the worldwide film industry.

The revenue-based ranking plot for a selected genre (here, Action) highlights which individual movies have earned the highest box office revenues within that genre. This visualization makes it easy to spot blockbuster hits, compare performance across top titles, and analyze whether high-budget productions consistently translate to commercial success.

The Sankey plot illustrates how genres evolve and flow across decades, showing the connections between production eras and genre trends. This helps track the rise or decline of specific genres over time, detect shifts in audience preferences, and understand how different genres contribute to the industry's creative output across generations.

7.4 Company Tab

7.4.1 Functionality Overview

The Production Company Tab enables focused analysis of movies produced by a specific production company. Users can select a company to view its top movies ranked by rating, budget, revenue, or ROI, and see how its production trends have evolved over time. This view highlights the company's contribution to different genres and decades.

7.4.2 Core Visualizations

This tab features visualizations designed to reveal how a production company's output is distributed across genres and time. The Sankey diagram illustrates the flow from the selected company to different decades and genres, showing how its production priorities have shifted over time. A donut chart complements this by summarizing the share of various genres within the company's total output, making it easy to spot dominant genres or niche areas the company has invested in.

7.4.3 Interactive Elements

- **Company Selection Dropdown:** A dropdown menu allows the user to choose a production company. Selecting a company updates all connected visualizations, including the Sankey diagram and donut chart.
- **Dynamic Ranking Filters:** Users can choose ranking criteria such as vote average, budget, revenue, or ROI to view the top 10 movies produced by the selected company. This filter helps users explore high-impact titles.
- **Hover and Drilldown:** Both the Sankey diagram and donut chart offer interactive tooltips. Hovering reveals details about genre shares, decade splits, and movie counts. Clicking a segment can show related top movies or deeper genre breakdowns.

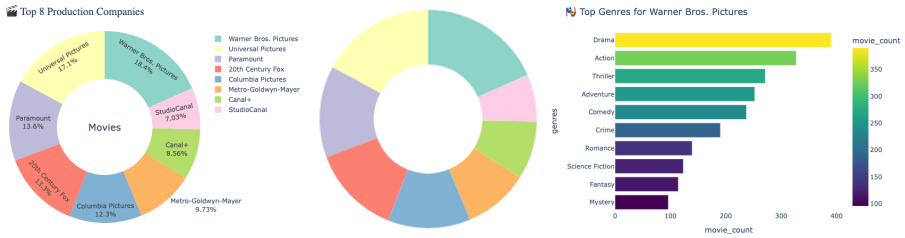


Figure 7.11: Production Companies' Donut Chart

7.4.4 Dashboard Screenshots

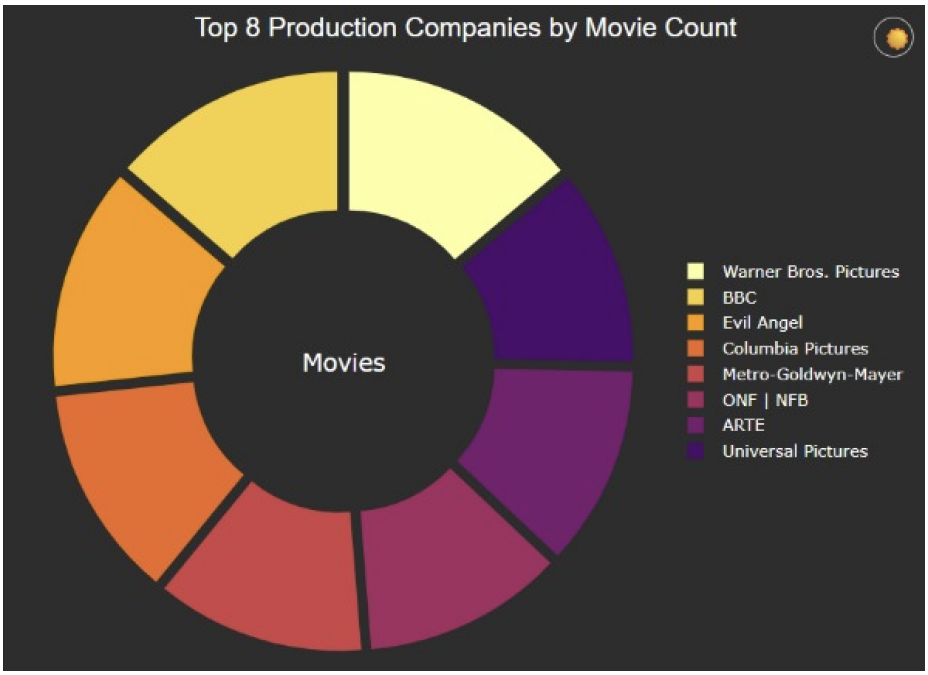


Figure 7.12: Movie Count Donut Chart

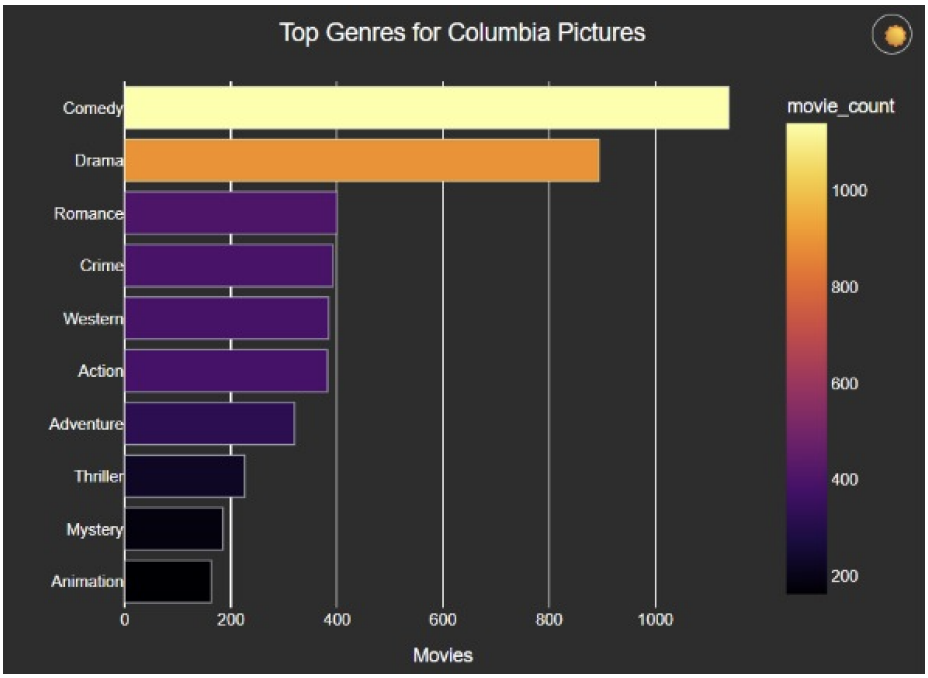


Figure 7.13: Top Genres for a Production Country

7.4.5 Observations

The Production Companies' Donut Chart provides a clear overview of how movie production is distributed among major studios, highlighting the dominant companies with the largest share of output and allowing quick comparisons of market share proportions. The Movie Count Donut Chart shows the count of movies produced across various segments or time frames, giving a snapshot of production volume distribution and indicating whether output is concentrated within a few entities or spread more evenly. The Top Genres plot for a selected production country reveals which genres are most popular or prioritized by that country's film industry. It helps identify regional strengths, creative focus areas, and genre specialization, supporting cross-country comparisons and production strategy analysis.

The donut chart of production companies provides an immediate visual summary of how a company's output is distributed across genres, highlighting dominant categories and niche areas of focus. The movie count donut chart offers a clear snapshot of the total number of movies associated with the selected company or country, allowing quick comparisons between different producers or regions. The final chart showing the top genres for a production country helps identify which genres are most popular or profitable within that country's film industry, revealing regional storytelling preferences and market trends.

Chapter 8

Contributions

Student Name	**Specific Contributions**
Om Bhartiya	Donut & Pie plot, PPT Making, Data Preprocessing, Report Format
Aryamann Srivastava	Budget vs Revenue plot, Runtime vs Ratings analysis, Latex Report
Kshitiz Tyagi	Revenue by Genre plot, Data Preprocessing, Setting up Dashboard
Tejas Shrivastava	Revenue by Country plot, Top Production Companies analysis
Saaumitra Raaj	Top Rated Movie Trend plots, Data Validation, Chloropleth
Swarnim Verma	Popularity Trend plot, and Visual Styling and Geospatial plot
Aryan Deo	Helped integrate all plots, Wrote draft conclusions, PPT Making
Harshita Awasthi	Created genre-trend plots, Sankey diagrams, Did final proofreading

Table 8.1: CONTRIBUTIONS MADE BY EACH TEAM MEMBER

* Every Team member was also responsible for applying required data preprocessing on their respective file and had complete freedom to come up with innovative representation for their data.

* Also all team members have combined effort for integrating their part of code and mentioning their part and observation in the final report.

Chapter 9

Challenges and Limitations

While our TMDB Movies Visual Analytics project offers valuable insights into global movie industry trends, it is important to acknowledge a few key challenges and limitations:

- **Data Completeness:** Although the TMDB dataset is extensive, some movies have missing or incomplete data for critical fields like budget, revenue, or genre. We handled missing data using cleaning and filtering, but some noise may remain.
- **Accuracy of Self-Reported Data:** Many budget and revenue figures in TMDB are crowd-sourced or self-reported by studios. These may not always match actual financials, especially for older or independent films.
- **Currency and Inflation Effects:** The dataset contains budgets and revenues in nominal terms across decades. We did not adjust figures for inflation or exchange rates. As a result, older movies' revenues may not be directly comparable to recent releases in real monetary terms.
- **Aggregation Level:** Some of our plots aggregate movies by country or company. However, many films are co-productions involving multiple studios or countries. Assigning a single country or company might oversimplify the real production context.
- **Machine Learning Limitations:** For our linaer regression and confidence interval extension, results depend heavily on the selected features and assumptions. These exploratory models provide rough groupings and forecasts, but should not be interpreted as precise predictions.
- **Computational Load:** Processing a dataset with nearly a million entries requires significant memory and compute resources. For larger feature combinations or future enhancements, more efficient pipelines and cloud-based deployment could help.

Chapter 10

Conclusion

The TMDB Movies Visual Analytics project, developed using Python, Pandas, and Plotly, demonstrates how large-scale entertainment datasets can be transformed into meaningful, interactive insights. By combining detailed data cleaning, thoughtful feature selection, and a carefully designed set of visualizations, our project enables users to explore global movie trends spanning more than a century.

Through dynamic charts, clear plots, and an optional layer of simple machine learning, the interface allows stakeholders to examine movie production growth, genre popularity, revenue patterns, and relationships between factors like budget, revenue, and runtime. Each visualization tells a part of the larger story of how the global film industry has evolved and which factors drive its financial and cultural impact. While acknowledging limitations such as missing data, currency effects, this project shows the power of visual analytics in uncovering hidden trends and patterns. By transforming raw movie data into an accessible, interactive narrative, our work provides a robust foundation for future improvements, such as adding more sophisticated predictive models or deploying the interface as a standalone web app.

Overall, this project illustrates how big data visual analytics can make complex datasets interpretable and actionable, paving the way for deeper understanding and future exploration in the entertainment industry.

Bibliography

[1] Rajarshi0001, *CS661_Project Repository*, Github.

[2] Sample Project Report, *Indian Crime Data Visualization Portal: An Interactive Exploration Platform.*

[3] Python Documentation, *Dash by Plotly.*

[4] Dataset, *TMDB movie dataset v11.csv, from Kaggle.*

[5] Gen AI tools, *Note: They were used only for formatting, the codes are our own.*

[6] Course Notes of *CS661*, by Prof. Soumya Dutta.