Group-2

CS661 Course project

CineScope

Global Movie Trend Analysis via TMDB

DR. SOUMYA DUTTA

DEPT. OF CSE, IIT KANPUR

TEAM MEMBERS

Om Bhartiya	230714	ombhartiya23@iitk.ac.in
Kshitiz Tyagi	230585	ktyagi23@iitk.ac.in
Aryamann Srivastava	230211	aryamanns23@iitk.ac.in
Swarnim Verma	231071	swarnimve23@iitk.ac.in
Aryan Deo	230213	aryandeo23@iitk.ac.in
Saaumitra Raaj	220928	saaumitra22@iitk.ac.in
Tejas Shrivastava	231091	tejass23@iitk.ac.in
Harshita Awasthi	230463	harshitaa23@iitk.ac.in

CONTENTS

EXECUTIVE SUMMARY

Agenda- In this project we chose a very large TMDB dataset from Kaggle containing approximately 930,000 records of movies spanning more than 100 years with features such as release year, genre, budget, revenue, ratings, production companies, country, popularity.

Reason for this Dataset-

- Very Large Dataset (Around a Million Records!!!)
- Very Diverse (So many Categories to choose from)
- Very Rich Meta Data, allowing exploration from multiple angles

Our Plan: Turn Millions of rows into interactive visuals that answer real questions about the film industry, which is why we tried to build a visual interface where a user can pick filters, see live plots, and find hidden insights without writing any code

Output: An interactive multi-tabbed dashboard offering comprehensive insights into global movie trends using TMDB data.

INTRODUCTION

Project focus -This project aims to create an interactive visual analytics dashboard that allows users to explore and understand global movie industry trends using a large-scale dataset from TMDB.

Objective -

- Develop a centralized platform for analyzing century-wide movie statistics.
- Enable multi-dimensional exploration (by genre, time, geography, company, etc.).
- Reveal patterns, correlations, and anomalies hidden in raw data.
- Provide insights into industry questions like:
  - Do high budgets yield high returns?
  - How do genre preferences change over time?
  - Which studios dominate the market?

Techniques Applied - Data preprocessing, exploratory data analysis, interactive visualizations with Plotly and Dash, and basic regression modeling.

Using these Frameworks and Utilities, we created interactive charts and interfaces for users along with deploying the dashboard.

LIBRARIES AND THEIR USAGE

Library	Purpose
Pandas	Data loading, cleaning, transformation, feature engineering
NumPy	Efficient numerical operations, type conversions
Plotly	Interactive plots: scatter, sunburst, treemap, heatmap, streamplot, etc.
Dash	Web dashboard framework for multi-tab layout and user interaction
Google Colab	Development environment with cloud-based execution and sharing

DATASET DESCRIPTION

Dataset Overview

- Source: TMDB dataset from Kaggle
- Entries: ~930k movies
- Columns: 24
- Time Span: 100+ years of global film data
- Format: CSV, processed using Pandas in Google Colab

Key Features

title, release_date, runtime, genres, vote_average, vote_count,budget, revenue, popularity,production_companies,production _countries ,overview, keywords, original_language, adult flag

Data Challenges

- Missing values in genres, release_date, companies
- Zero or unrealistic budget and revenue
- Label inconsistencies in countries & companies

Use in Project

The dataset enabled visual exploration across:

- Time (release_year, release_decade)
- Genre (extracted from exploded genres)
- Geography (production_countries)
- Studios (production_companies)
- Finance & Ratings (budget, revenue, popularity, vote_average)

DATA PREPROCESSING

Data Loading & Initial Cleaning

- Parsed Dates: Extracted release_year and decade from release_date.
- Removed Future Releases: Dropped entries with release dates beyond the current year.
- Dropped Nulls: Removed rows missing critical values (e.g., release date).

Feature Engineering and Standardization

- Derived Features:
  - main_genre
- Numerical Conversions:
  - Casted budget, revenue and runtime to numeric types.
  - Coerced invalid entries to NaN.
- Zero-Value Handling:
  - Filtered out movies with zero budget or revenue for financial plots.
- Text Label Cleanup:
  - Standardized labels in production_company, country, language for consistent groupings.

Preparation for Visualization & Modeling

- Genre Exploding: Split comma-separated genres for proper genre-wise analysis.
- Filtering Unrealistic Entries: Removed extreme outliers or entries with conflicting values.
- Data Subsetting: Created smaller DataFrames per tab (e.g., for genre, country, company views).

DASHBOARD OVERVIEW

The dashboard consists of four interactive tabs — Overview, Genre, Country, and Company — each offering a unique lens into global movie trends using TMDB data.

Overview Tab

Purpose: High-level visualization of global movie trends over time.

Insights: Reveals how global movie production and genre popularity have evolved over time.

Genre Tab

Purpose: Deep dive into performance and trends of individual genres.

Insights: Shows how financial performance, audience ratings, and trends vary across individual genres.

Country Tab

Purpose: Analysis of movies by production country.

Insights: Highlights regional strengths, production volumes, and revenue contributions of different countries.

Company Tab

Purpose: Exploration of trends and outputs from specific production companies.

Insights: Analyzes top production companies' output, genre focus, and market dominance over the years.

KEY TASKS ENABLED IN THE DASHBOARD

- **Production volume over time:**

We used a **histogram** to analyze year-wise movie production trends, as it effectively displays the frequency of releases over time, helping us identify major growth periods and historical shifts in production.

- **Budget vs revenue analysis:**

We used a **scatter plot** with a fitted linear regression line to analyze the relationship between movie budgets and revenues, as it clearly reveals ROI trends, blockbuster outliers, and less profitable ventures through the spread and clustering of data points.

- **Genre-based revenue comparison:**

We used a **treemap** to compare genre-wise revenue performance, as it effectively visualizes both the relative contribution and hierarchical structure of genres, helping identify consistently high-grossing versus niche categories in a compact and intuitive format. Country-level trends analyze production and revenue distribution globally

KEY TASKS ENABLED IN THE DASHBOARD

- **Country-Level Genre Analysis:**

We used a **choropleth map** alongside a bar chart to analyze genre distribution across production countries, as the map highlights geographic patterns and dominant film-producing nations, while the bar chart provides clear comparative support for emerging markets.

- **Runtime vs Audience Rating Analysis:**

We used a **scatter plot** with a fitted regression line to analyze the relationship between movie runtime and audience ratings, as it effectively captures trends and correlations between two continuous variables, highlighting viewer preferences related to film length.

- **Top Production Companies Overview:**

We used an interactive **donut chart** to highlight leading production companies by cumulative revenue, offering a clear view of industry concentration; clicking a segment reveals a detailed bar chart of the company's top movies by budget, revenue, or ROI for deeper performance insights.

- **Top-Rated Movies Trend Analysis:**

We used a **bar chart** to track how top movies each year vary in terms of budget, revenue, ROI, and popularity, effectively revealing evolving audience and critic preferences over time through a clear, continuous visual format.

KEY TASKS ENABLED IN THE DASHBOARD

- **Genre Dominance and Recent Trends Analysis:**

We used a **streamgraph and sunburst chart** to analyze the evolution of genre prevalence and performance over time—the streamgraph captures smooth temporal shifts and overlaps in popularity, while the sunburst provides a multi-level breakdown of main genres into sub-genres or periods, offering deeper insights than static bar or pie charts.

- **Genre Evolution over time:**

We used a **Sankey diagram** to assess genre dominance over time by showing how production companies flow into different genres based on cumulative output and revenue, as it clearly visualizes split and flow connections that bar charts cannot effectively convey.

- **Evolution of Production for Companies/Countries within a Genre:**

We used a **heatmap** to track how individual studios or countries have increased or diversified their output within a genre over time, as it effectively captures variations across both time and categories, which line charts may miss when comparing 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.

OVERVIEW TAB

Visualizations:

- Histogram: Number of movies produced per year.
- Streamplot: Evolution of genre popularity over decades.
- Sunburst Charts: One shows the overall genre hierarchy other highlights year-specific genre distribution.

Interactivity:

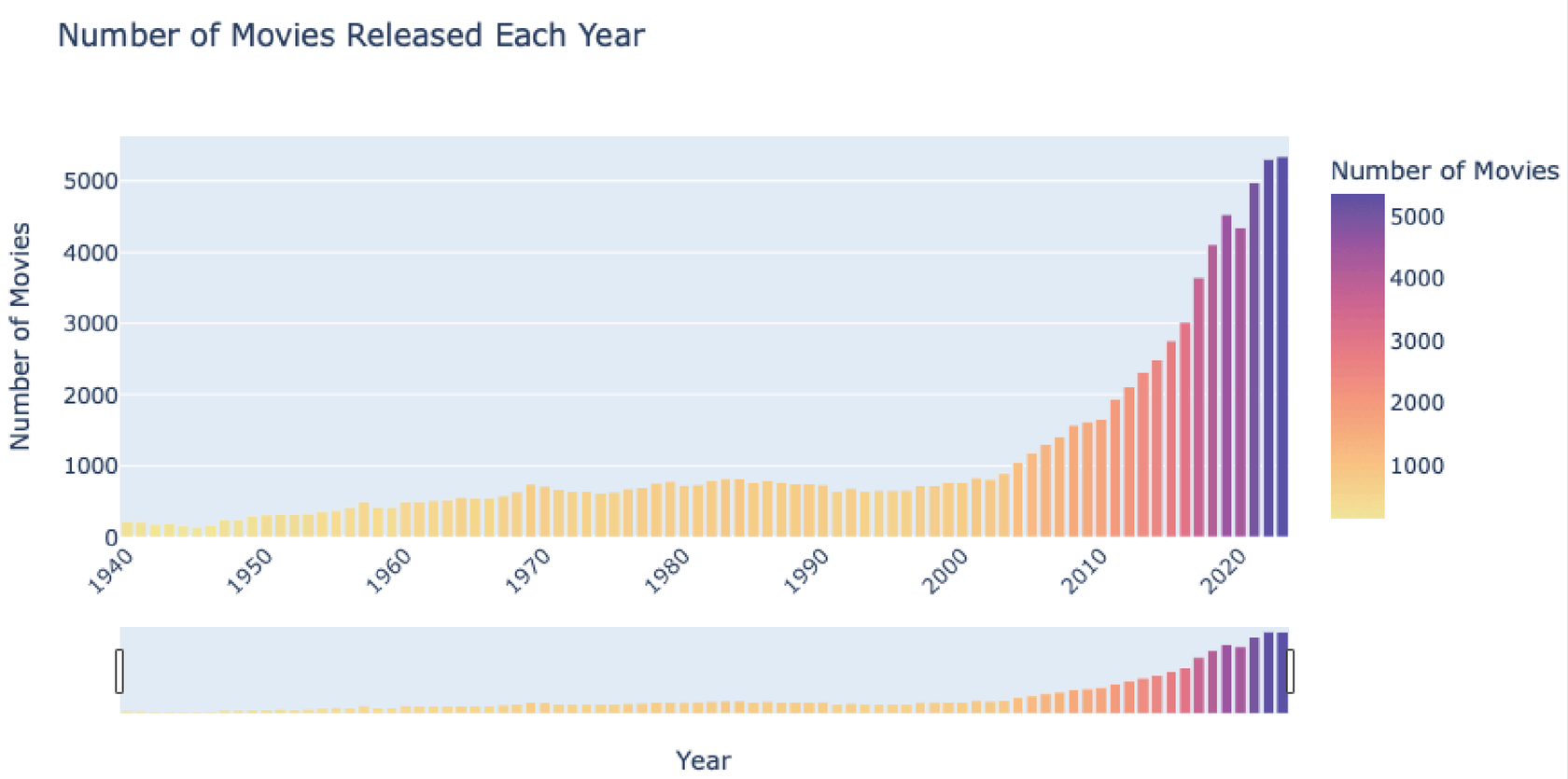
- Genre Dropdown: Filter streamplot and sunburst charts by selected genres.
- Year Range Slider: Adjusts all time-based visualizations.
- Hover Tooltips & Click Events: Show additional movie metadata interactively.

Observations:

- Spot production spikes (e.g., growth after 2000).
- See rising/falling popularity of genres like Action, Comedy, or Drama.

NUMBER OF MOVIES RELEASED EACH YEAR

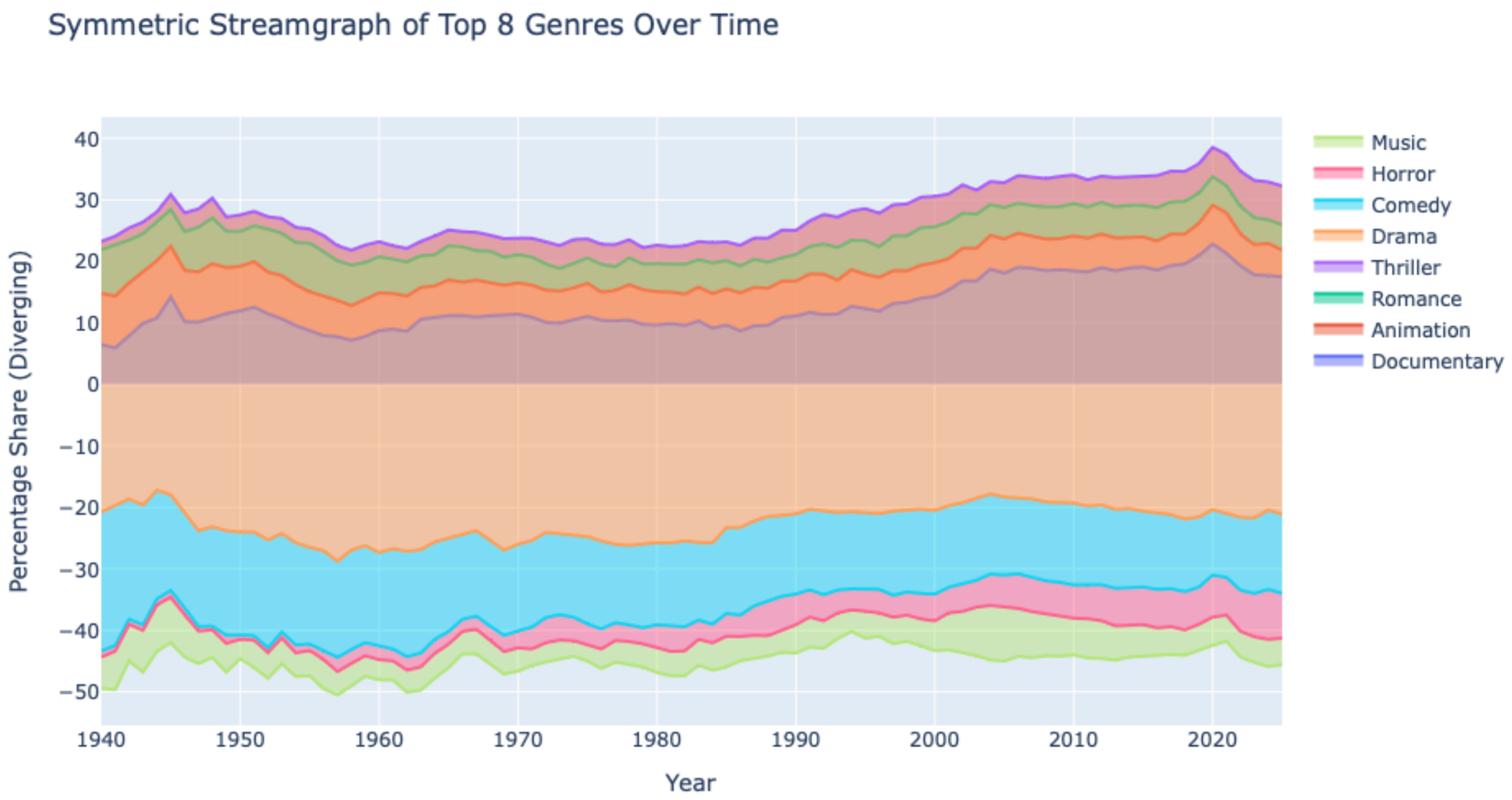
Histogram



The histogram clearly shows how the total number of movies produced each year has evolved over time, revealing periods of growth in film production as expected due to increase in audience and cinema popularity.

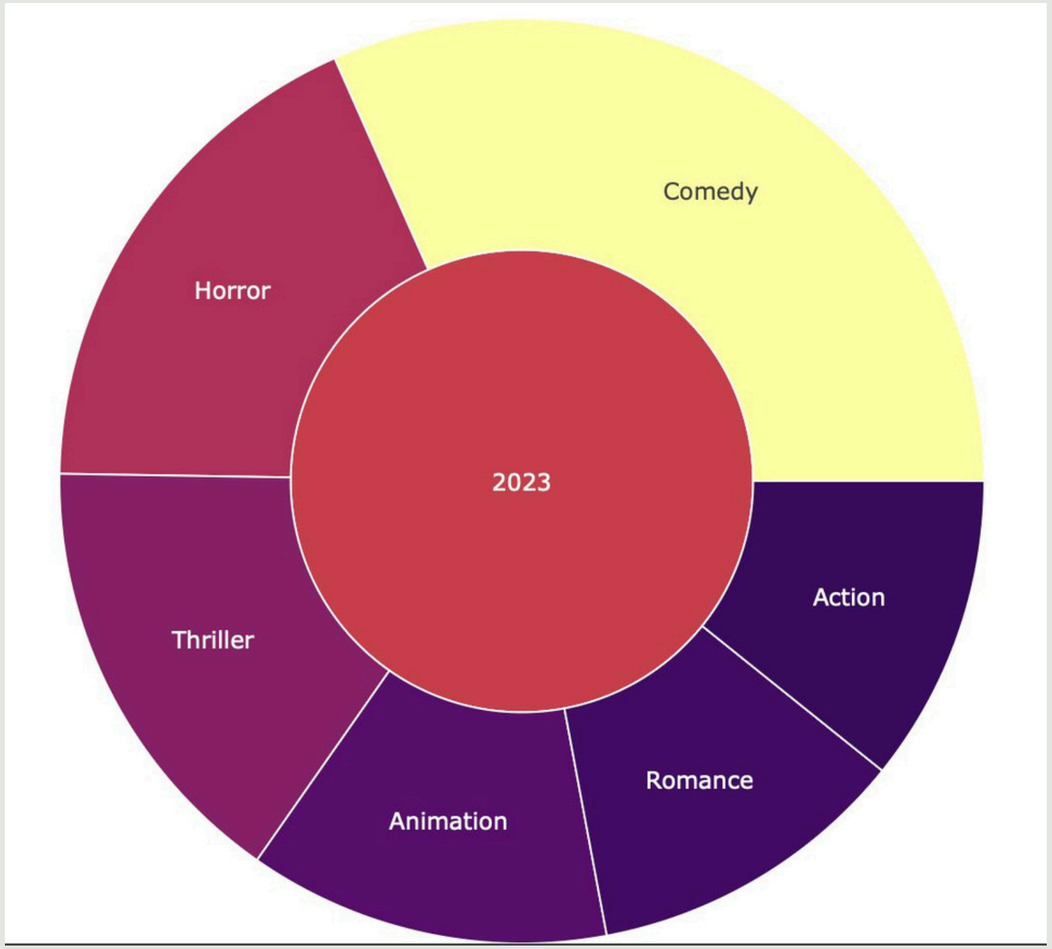
SHARE OF TOP GENRES WITH TIME

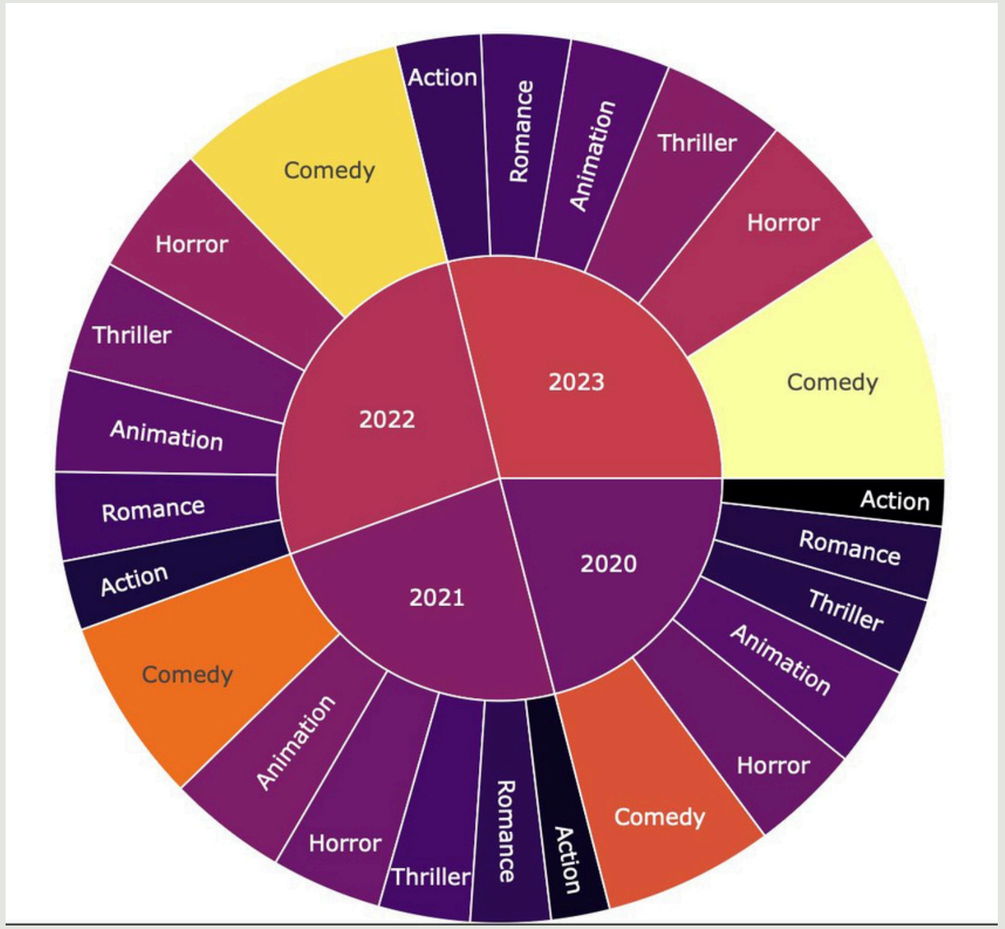
Streamplot



The streamplot highlights the changing popularity of different genres, making it easy to identify genres that have gained or lost audience share across decades.

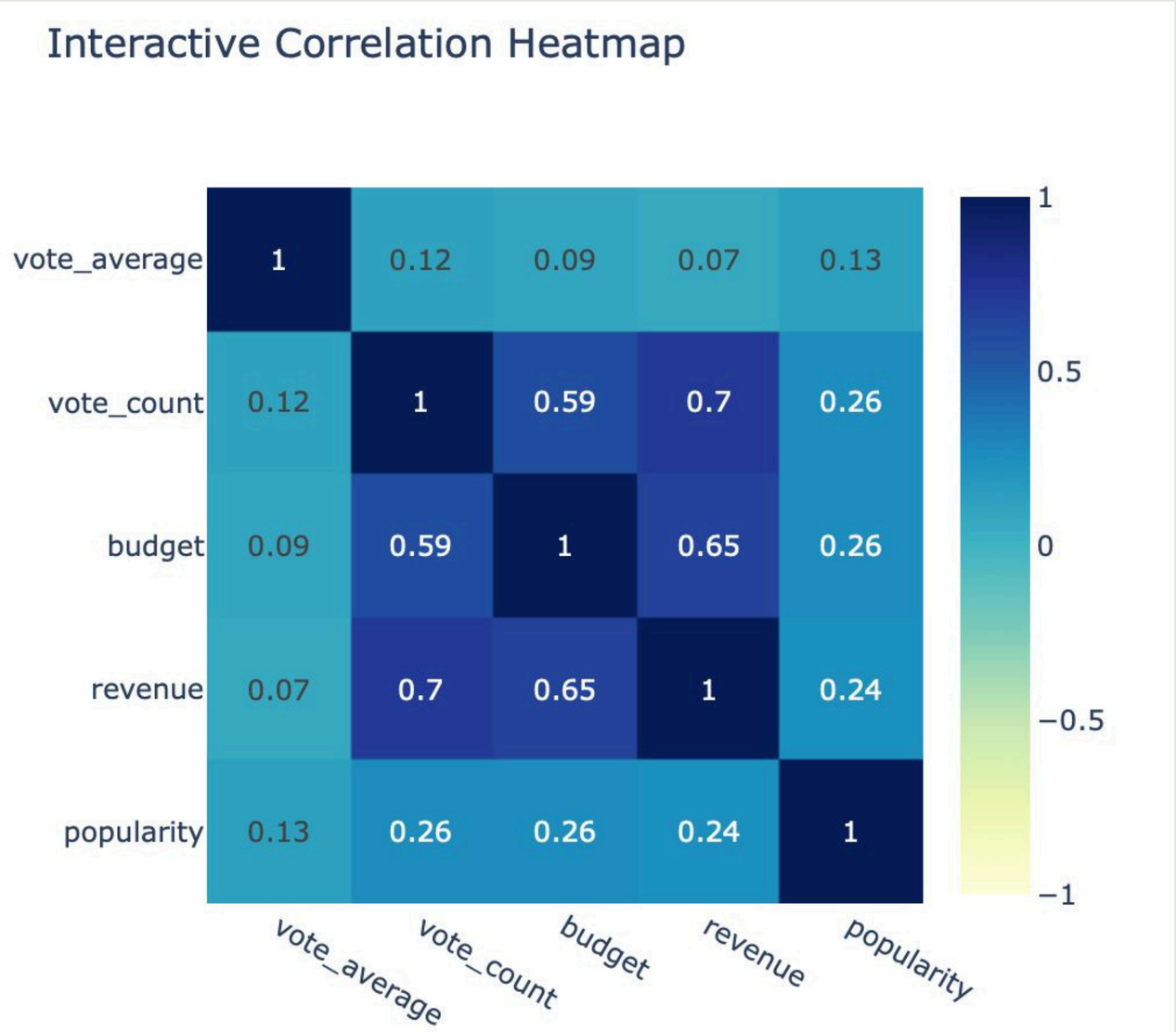
SHARE OF GENRE ACROSS YEARS - SUNBURST CHARTS





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.

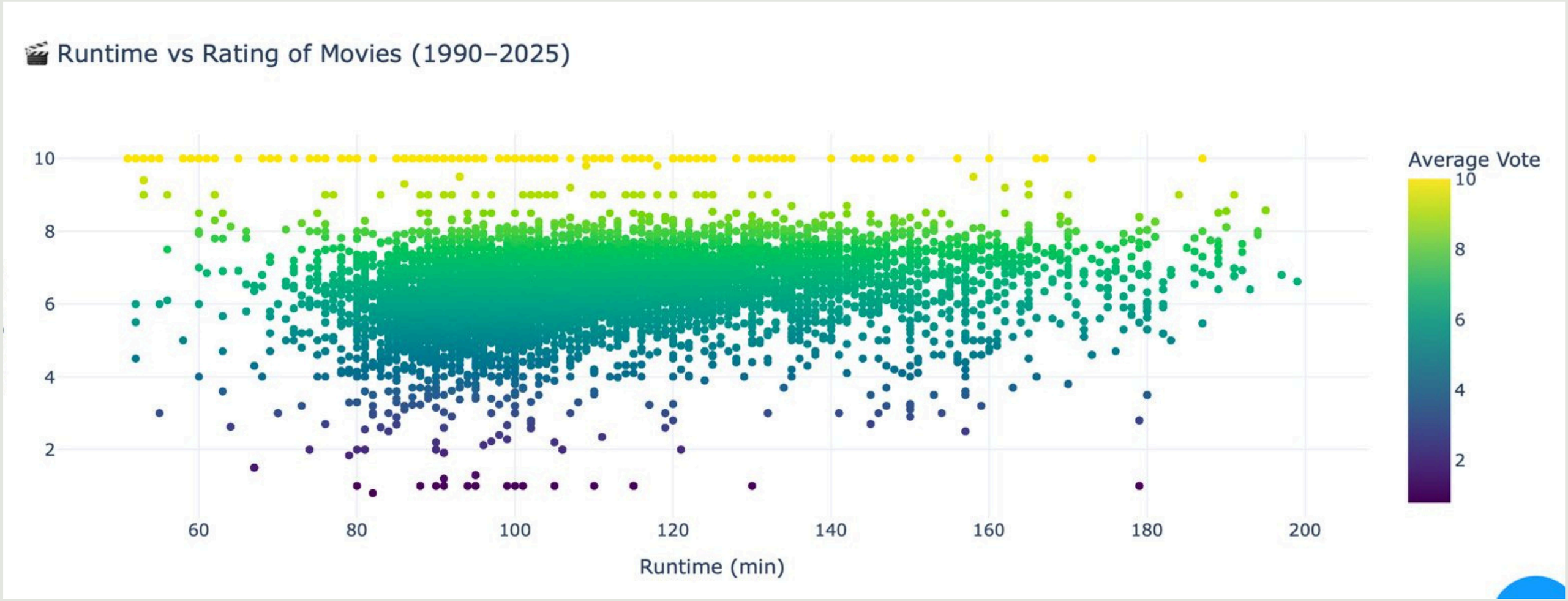
CORRELATION HEATMAP



The heatmap supports the intuitive assumption that higher budgets typically lead to better production quality, which in turn attracts a larger audience — as evidenced by the strong correlation between vote count and revenue, and consequently, higher overall revenue.

Average Vote (Rating) is not significantly correlated with any of them — suggesting critical acclaim and commercial success don't always go hand-in-hand.

SCATTER PLOT



This scatter plot depicts that density of votes decreases with increasing runtime indicating that fewer movies have higher runtimes

High ratings are achievable regardless of length, but extreme runtimes show more variance in quality.

GENRE TAB

Visualizations:

- Histogram: Yearly count of movies for each genre, with genre selection available through a dropdown menu.
- Heatmap: Movie counts by country and by studio over time, for any selected genre.
- Tree Map: Genres sized and colored by their average budget.

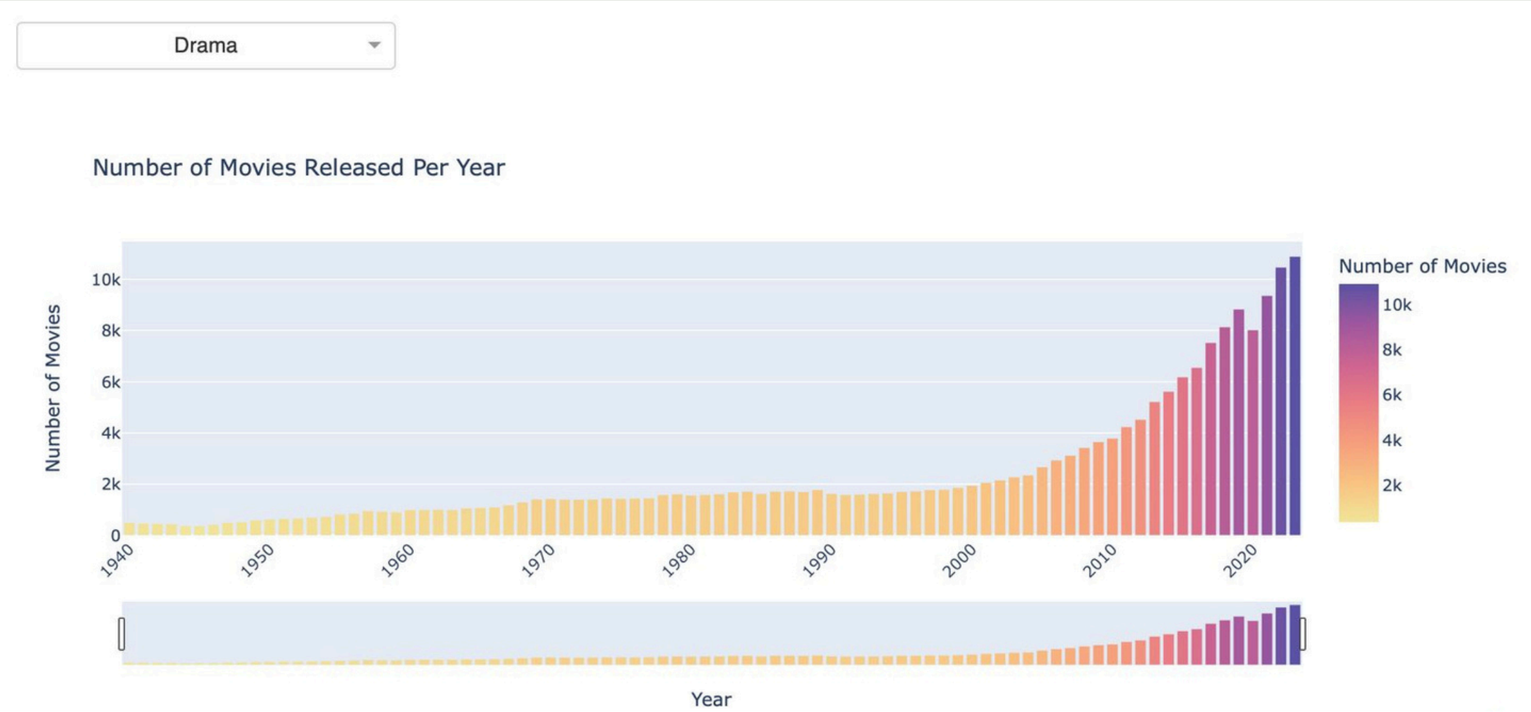
Interactivity:

- Genre Selector: Dropdown allows switching between genres in the histogram and heatmap.
- Hover: Detailed statistics appear when hovering over each data point.

Observations:

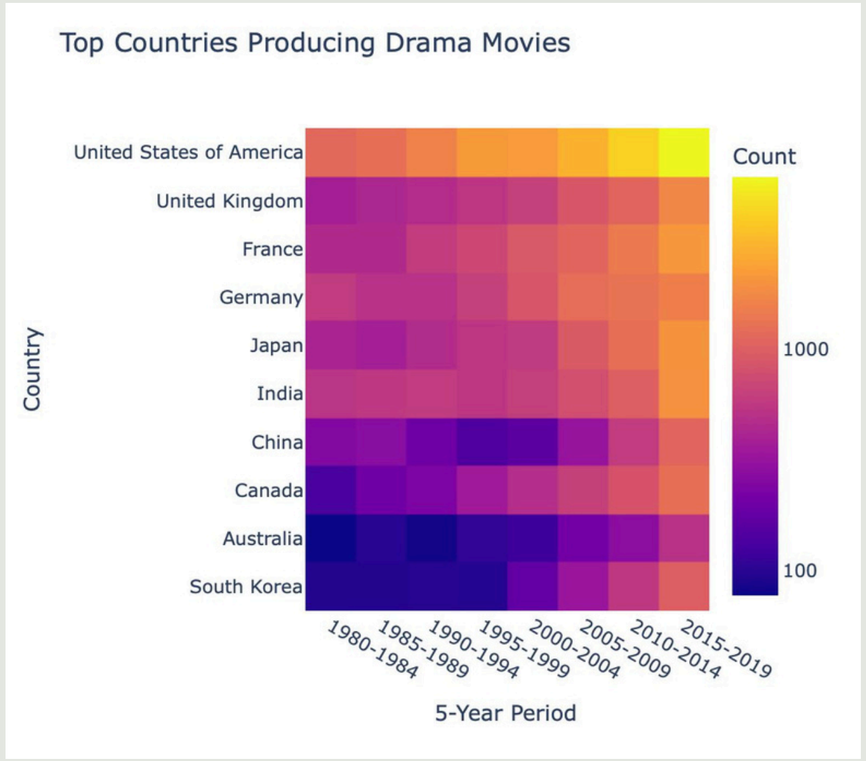
- Discover which genres are consistently high-performing.
- Understand when a genre peaked or declined.
- Treemap highlights genres with greater financial investment

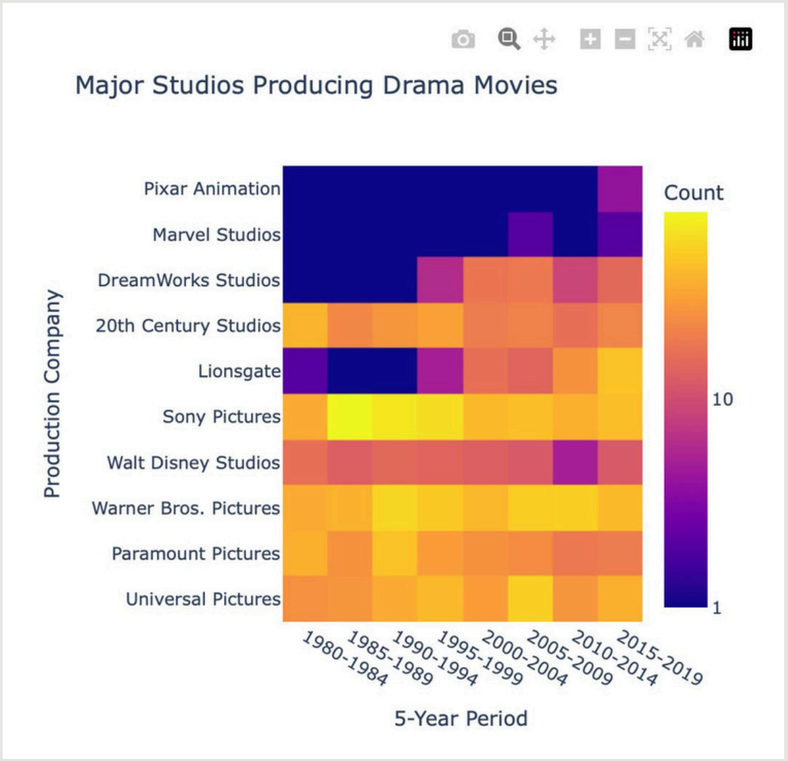
MOVIE PRODUCTION BY GENRE – HISTOGRAM



The histogram shows how the production of movies in each genre changes year by year. It makes it easy to spot trends, such as periods of rapid growth or decline in output for any selected genre, helping to quickly identify when a genre was most or least popular.

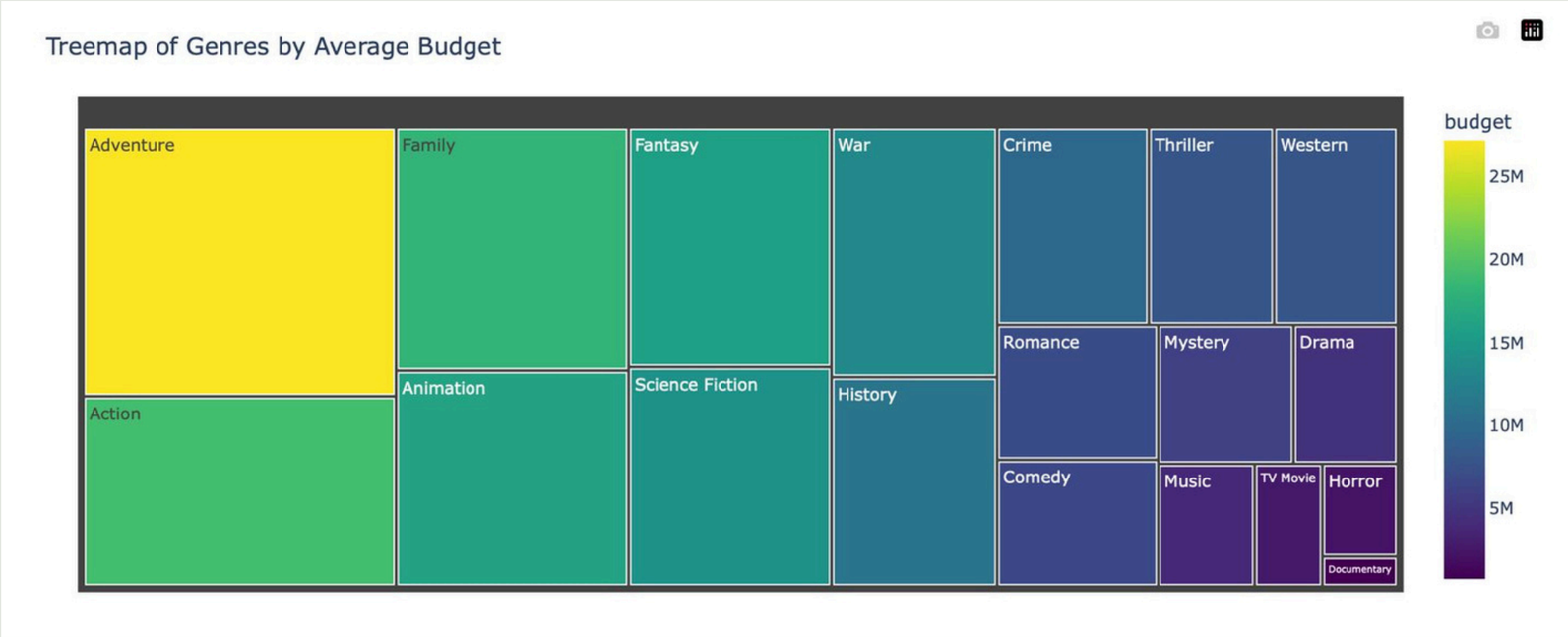
MOVIE PRODUCTION BY COUNTRY & STUDIO – HEATMAP





The heatmaps display how movie production is distributed across countries and studios over time for each genre. They reveal which countries and studios are most active in producing certain genres, and how their output shifts across different periods, highlighting key players and emerging trends.

GENRE BY AVERAGE BUDGET - TREE PLOT



The tree map visually breaks down average budgets by genre, with larger and brighter segments representing genres that attract more financial investment. This helps pinpoint which genres receive the most funding and which occupy smaller, niche spaces, guiding decisions on where to focus resources and attention.

COUNTRY TAB

Visualizations:

- Interactive Geospatial Plot: Map showing country-wise movie counts and revenue.
- Bar Charts: Top movies by rating, budget, and ROI within a selected country.
- Time Trends: Movie production or revenue evolution for the country.

Interactivity:

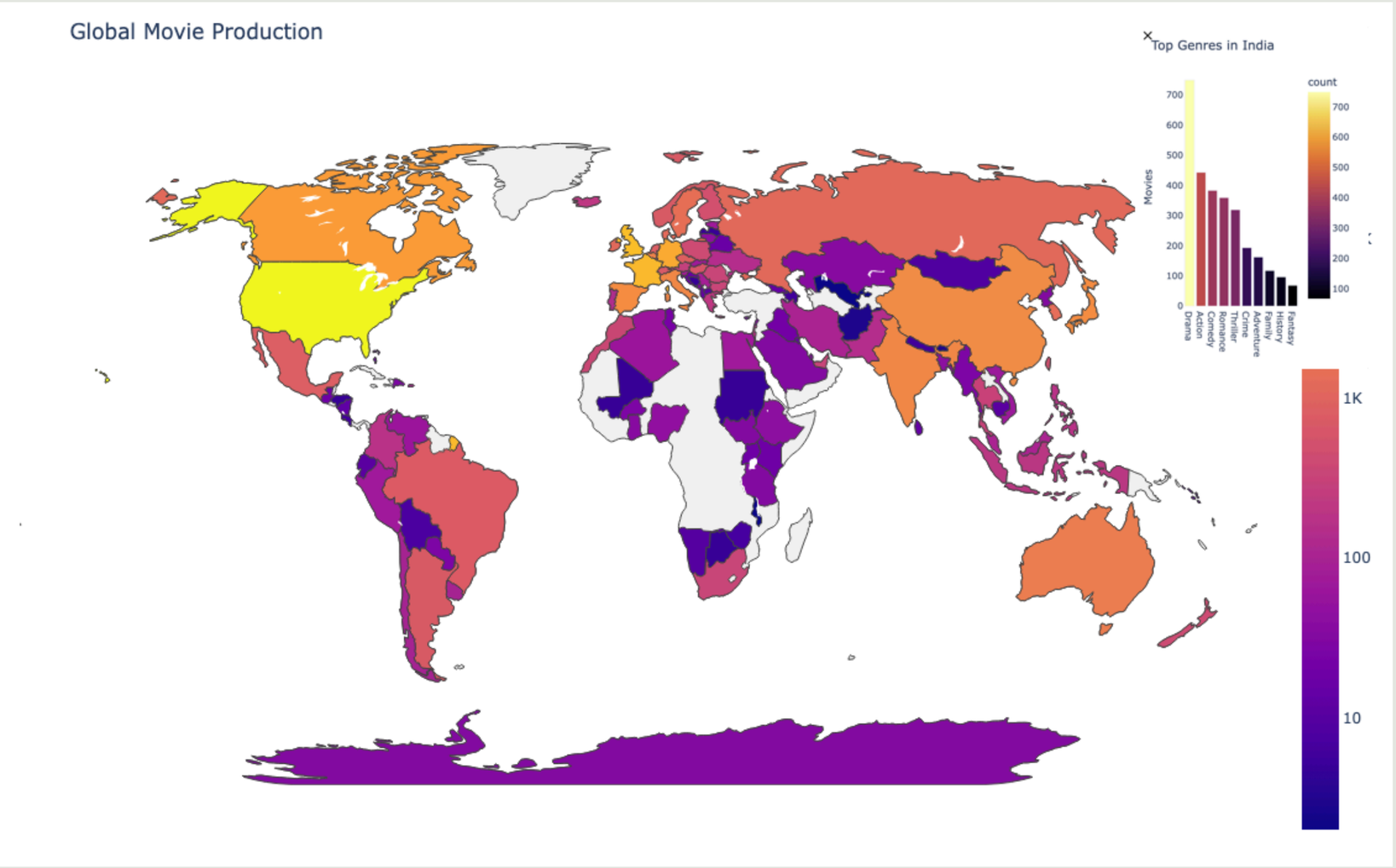
- Country Selector: Pick a specific production country.
- Map Hover & Click: Displays country stats and enables deeper drill-downs.

Observations:

- Compare dominance of Hollywood vs emerging markets (e.g., India, South Korea).
- See regional preferences in genres or production scale.
- Analyze the rise of film industries in new geographies.

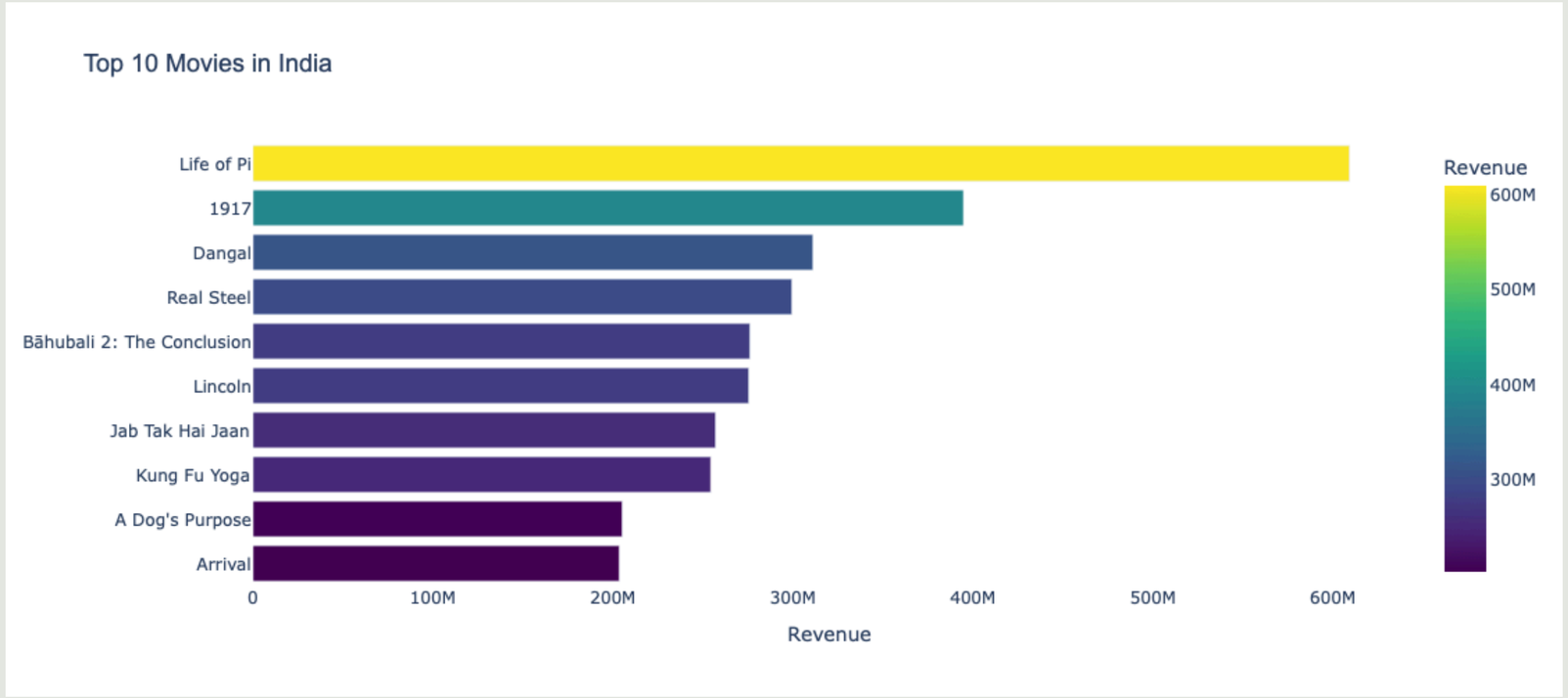
NUMBER OF MOVIES PRODUCED ACROSS COUNTRIES

Geospatial plot



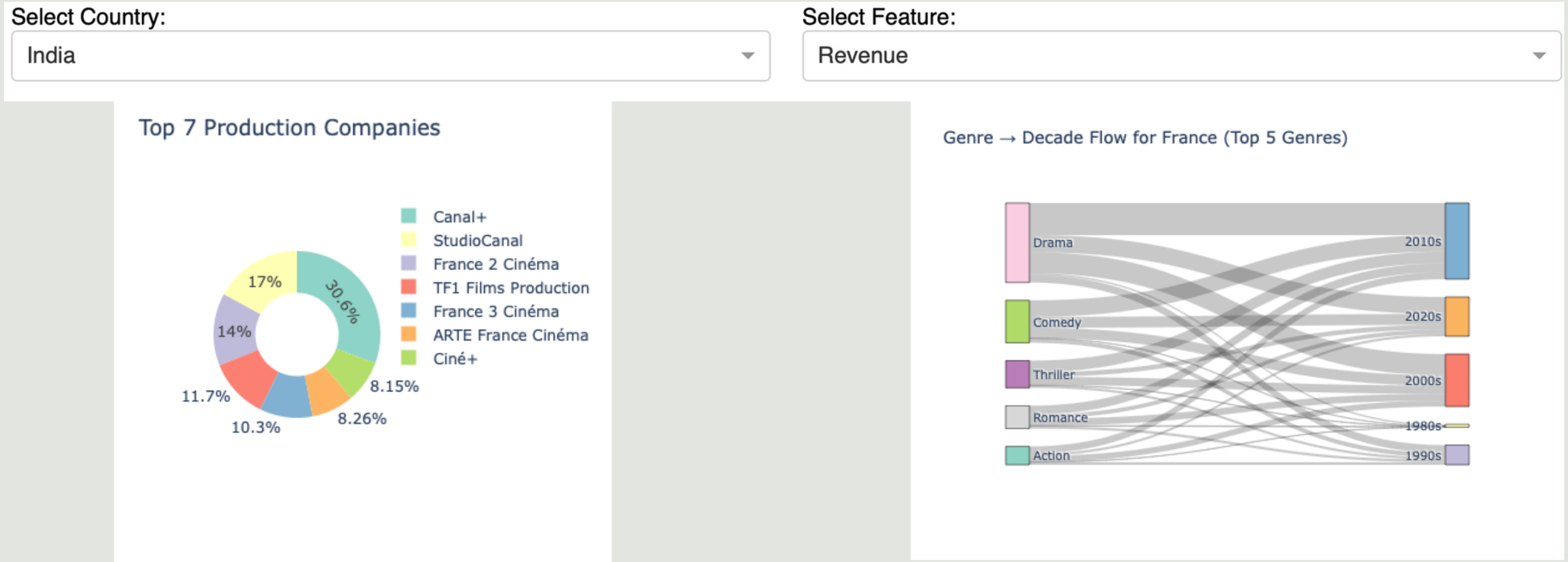
The geospatial plot provides an intuitive visual representation of movie production volume across countries, with interactive pop-ups (as shown here for India) offering quick access to key details such as total movies produced. This helps identify major film-producing hubs globally and reveals regional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

TOP REVENUES FOR GIVEN GENRE - RANKING PLOT



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.

DECADE-WISE GENRE EVOLUTION - SANKEY PLOT



Allows users to select a country and analyze its movie industry trends. Top 7 Production Companies (Donut Chart):

Displays the most dominant production companies from the selected country based on the number of movies produced.

Decade-wise Genre Flow (Sankey Diagram):

Visualizes the evolution of the top 5 movie genres across decades, highlighting shifting audience preferences and industry trends.

COMPANY TAB

Visualizations:

- Sankey Diagram: Flow of genres, and movie counts per company.
- Donut Chart: Market share distribution among major studios.
- Line/Bar Plots: Company-wise trends over time (e.g., revenue per year).

Interactivity:

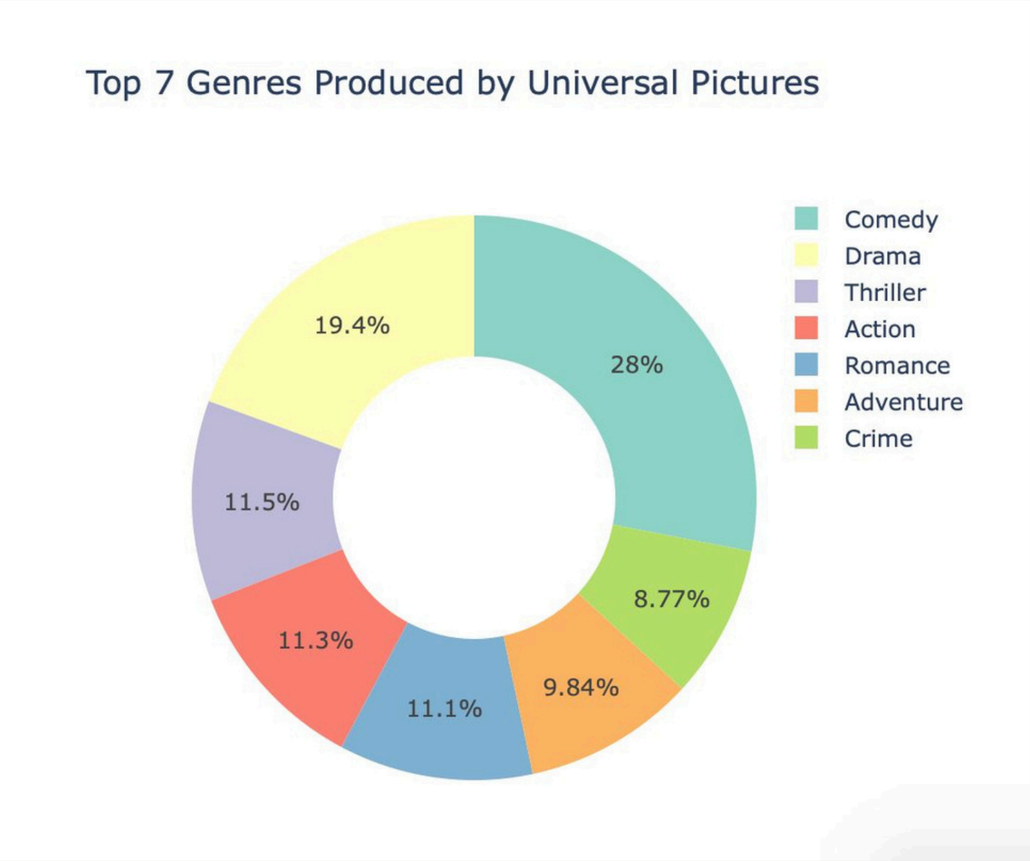
- Company Dropdown: Select a production company to explore its stats.
- Genre Filter: Analyze genre-specific contributions.
- Tooltip Details: Show movie counts, or ROI per visual element.

Observations:

- Identify market leaders (e.g., Warner Bros, Universal, etc.).
- Understand company strengths by genre.
- Track how companies adapt their strategies over decades.

NUMBER OF MOVIES PRODUCED COMPANY WISE

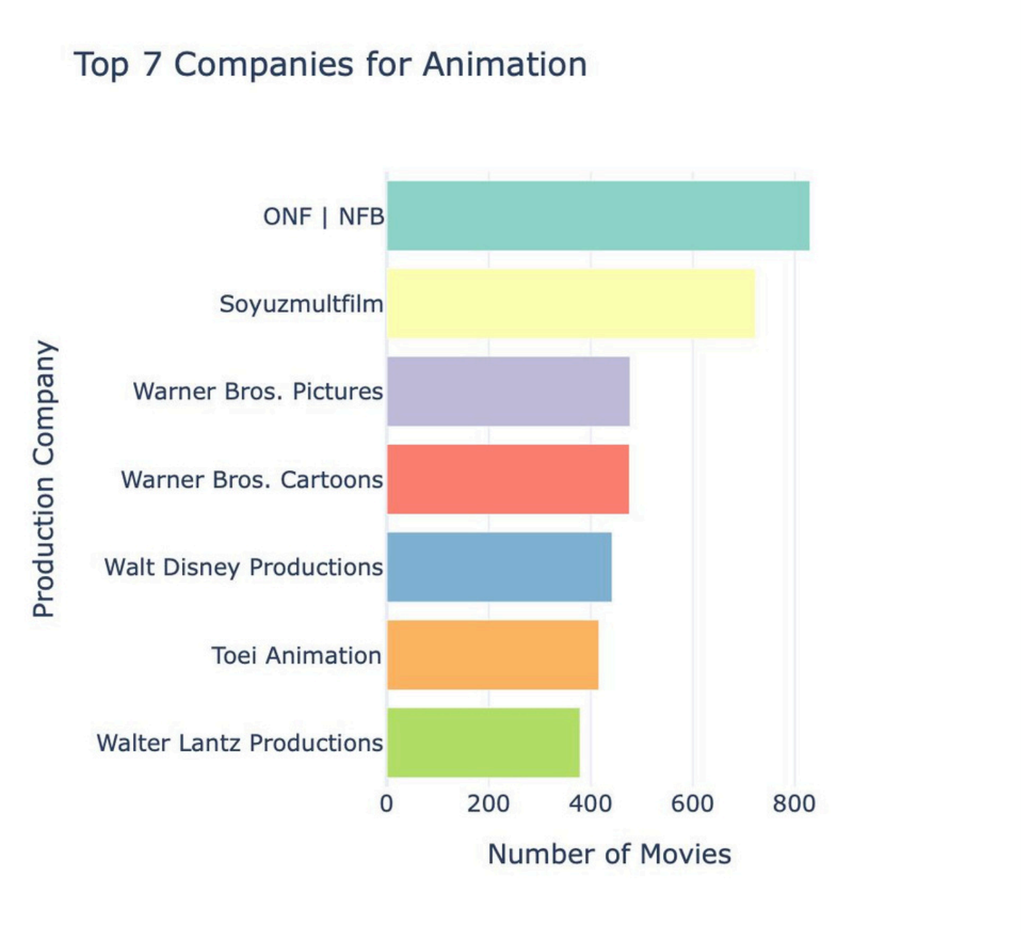
Donut chart



This pie chart illustrates the distribution of the top 7 movie genres produced by a selected production company, based on the number of movies in each category. By focusing on the most dominant genres, the visualization highlights the creative orientation and market priorities of the company. It provides a clear snapshot of where the company’s cinematic efforts have been most concentrated, offering insight into their strategic positioning within the film industry.

Bar Chart

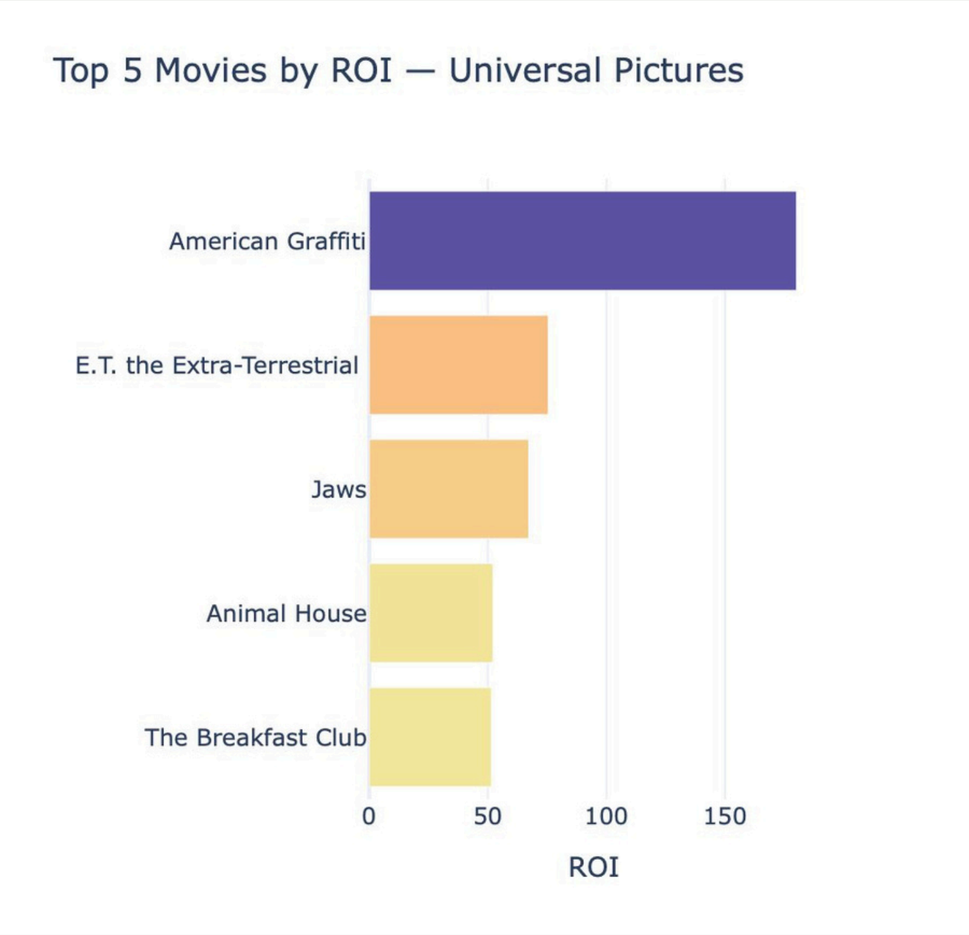
TOP COMPANIES FOR A GENRE



This bar chart visualizes the top 7 production companies for a selected genre based on the number of movies produced. It offers a clear comparison of dominance within a genre, highlighting the key players driving content creation in that thematic space. By filtering the data by genre, we gain insights into how different studios align with specific audience preferences and content niches. This representation helps in understanding market concentration and company-level genre specialization.

MOST SUCCESSFUL FILMS PER COMPANY

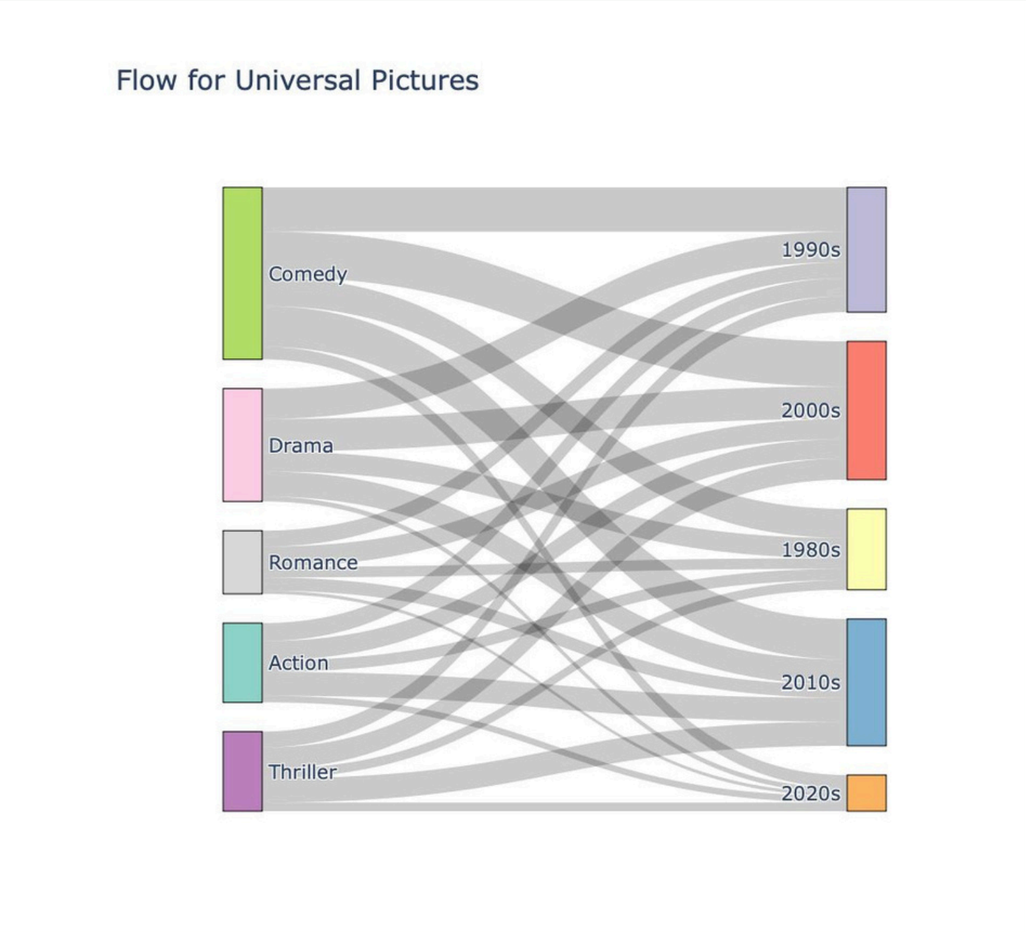
Bar Chart



This bar chart presents the top 5 movies produced by a selected company, ranked by Return on Investment (ROI). ROI is calculated as the ratio of revenue to budget, offering a standardized metric to evaluate financial performance. The visualization highlights the company's most profitable projects, regardless of scale, and offers insights into strategic successes. By focusing on ROI, this chart helps identify high-efficiency films that delivered exceptional returns relative to their production cost.

EVOLUTION OF FAVOURABLE GENRES

Sankey Plot



This Sankey diagram visualizes the flow of movie production from genres to decades for a selected company. Each link represents the number of movies produced in a particular genre during a specific decade. The width of each flow corresponds to the volume of movies, allowing us to observe how the company's focus evolved over time. This format reveals patterns such as dominant genres in certain eras, diversification trends, and shifts in production strategy across decades.

CONCLUSION

Centralized Visualization Platform

Delivered a multi-tabbed dashboard covering production, genre, country, & company-level insights.

Temporal & Comparative Analysis

Enabled users to analyze trends across time, genres, regions, and studios with interactive controls.

Pattern & Correlation Discovery

Revealed relationships (budget vs revenue, runtime vs rating) using scatter plots & regression lines.

Rich Multi-Dimensional Insights

Combined genre, geography, and financial data to uncover hidden trends and outliers.

Future Improvements

1. Use ‘tagline’ feature to create embeddings to identify potential clusters for given genres and other relevant insights.
2. Use posters of the images to draw other relevant insights.

Thank You