# CPSC 4030/6030 UG3 Project Report: IMDb Movie Insights Dashboard

## Overview and Motivation

Our project, IMDb Movie Insights, seeks to explore the impact of cinematic productions on audiences and box offices worldwide. Fueled by a deep interest in the dynamics of the film industry, our dashboard aims to unravel the complexities of movie economics. By examining the interplay between a film's budget, its global gross income, and the shifting sands of genre trends, we aspire to provide a comprehensive tool for industry analysis and enthusiast exploration.

## Related Work

Our dashboard draws inspiration from a variety of sources. We referenced academic papers on data visualization, and integrated insights from class discussions on state-of-the-art visualization methodologies. The backbone of our visual analytics approach is influenced by studies that delve into the economic trajectories within the film industry and the evolution of genre preferences over the years.

## Questions

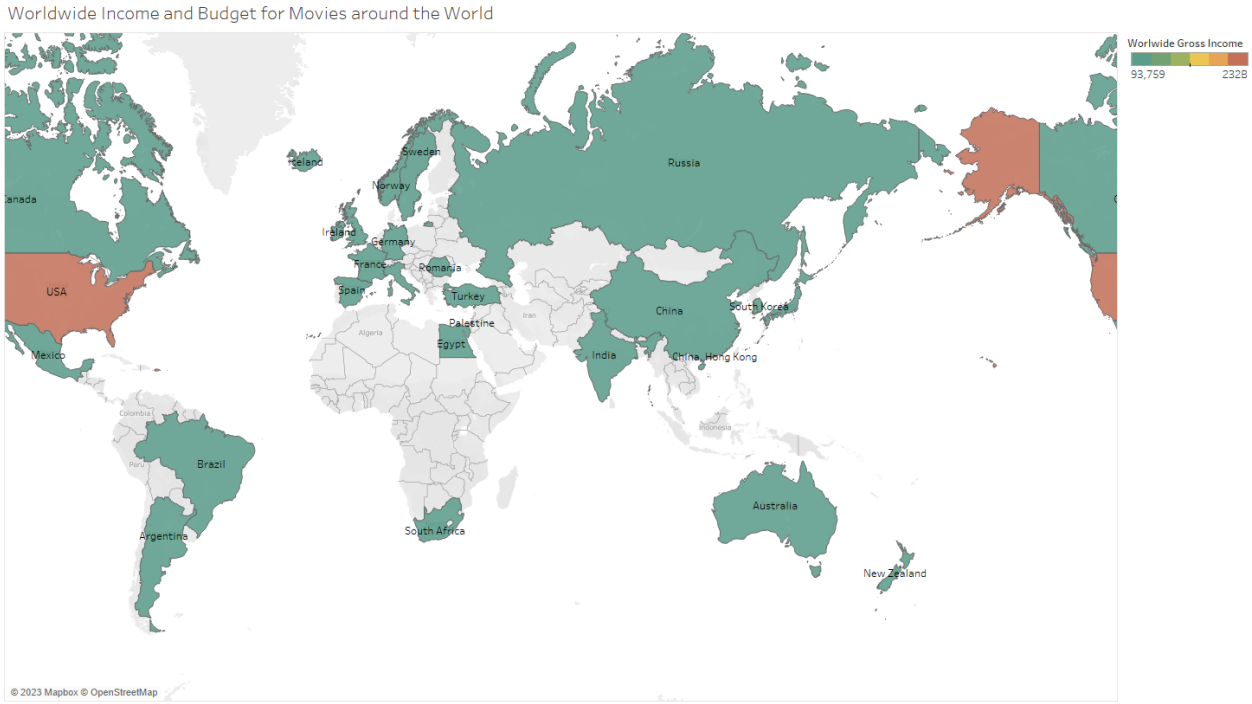
The project was initiated with the intent to dissect the correlation between movie budgets and their worldwide gross income. However, as we delved deeper into the data, our inquiries broadened to encompass genre popularity and its temporal fluctuations. This progression of our investigative focus was a natural response to the data's complexity and the technological challenges we faced.

## Data: Source, Scraping Method, Cleanup, etc.

The cornerstone of our visualizations was a comprehensive dataset titled "IMDB Movies 2000 - 2020.csv". This dataset required an extensive cleaning process to ensure the data's integrity for our visualizations. We used Python scripts, notably csv2json.py, to convert the CSV into a JSON format, which facilitated effective data manipulation using JavaScript and D3.js for our web-based interactive visualizations.

The movies.json and movies_to_actors.json files were critical for the Actors Network Graph, necessitating a meticulous cleaning process to ensure accurate actor-movie associations. This included normalizing actor names and creating accurate links between actors for shared movie roles. Our iterative and exploratory data cleaning process also involved visual tools to detect outliers and ensure data consistency.

An initial idea for a global heatmap to visualize movie GDP was eventually set aside due to its complexity and the realization that it did not add clarity to our objectives. The focus was then shifted to refining the dataset for the visualizations we deemed most effective.

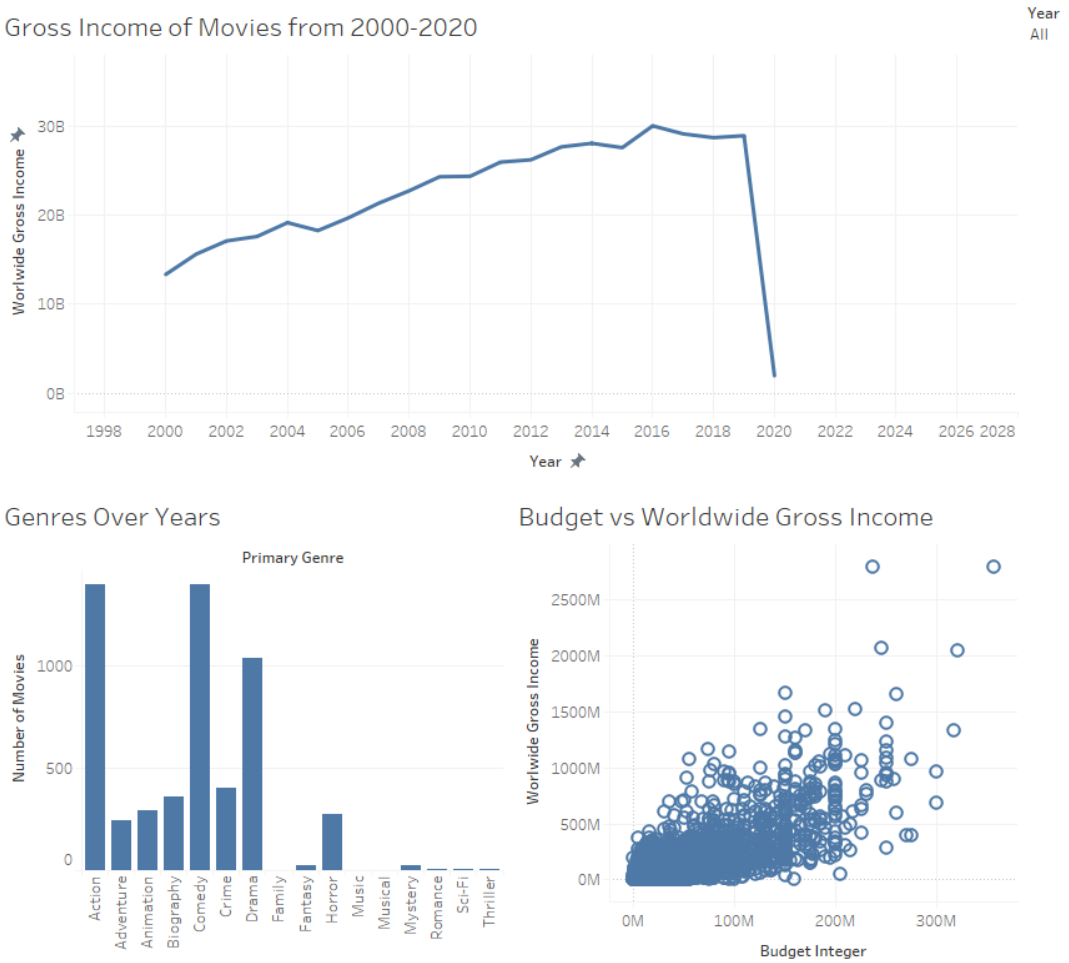



This heat map shows the Worldwide Gross Income of all the films in the database. However, the heat map is not able to show much information because of many movies being attached to several countries. The data scraping for this will be difficult to decide which movie should be mainly attached to one country.

## Exploratory Data Analysis

The exploratory data analysis (EDA) phase was a vital part of the IMDb Movie Insights project. This phase was characterized by an iterative process where visualization served as a tool for understanding the data and uncovering patterns and anomalies that could inform subsequent design decisions.

Initially, the EDA began with a high-level overview of the dataset. Using Tableau for data manipulation and D3.js for preliminary visualizations, we generated basic charts to get a sense of the data distribution. Bar charts were employed to visualize the frequency of movies across different genres and box plots to understand the spread of movie budgets and revenues. This provided an immediate sense of the data's scale and variability and highlighted potential outliers or data inconsistencies that required cleaning.




This design is our initial prototype of our dashboard. The visualization is relatively simple. As we delved deeper, we employed scatter plots to visualize the relationship between a movie's budget and its gross income. These plots were particularly revealing, displaying a positive correlation that, however, was not as strong as initially hypothesized. It became apparent that while some movies with colossal budgets did indeed garner significant revenue, there were many cases where moderate or even low-budget films achieved substantial box office success. This observation raised questions about the factors contributing to a movie's financial success beyond mere budget allocation.

Network graphs were also a significant part of our EDA, providing insights into the connections between actors based on their co-appearances in movies. These visualizations helped us understand the social network within the film industry and identify key influencers or central figures based on their connectivity within the network. For example, actors who frequently appeared in high-grossing films could be seen as central nodes, indicating their potential impact on a movie's success.

Throughout the EDA process, we were mindful of the data's temporal aspect. We created interactive line charts to track changes in genre popularity over time, revealing trends and shifts

that could correlate with cultural or economic events. For instance, the rise in popularity of superhero movies in the early 2000s could be seen within the broader context of global events and industry changes.
The EDA phase was not without its challenges. We encountered issues with data granularity and completeness, particularly when dealing with international movies where budget or revenue data was not always available or consistent. In response, we refined our data cleaning processes, such as normalizing currency values and addressing missing data through imputation or exclusion, depending on the context.
Moreover, the EDA process was crucial for determining the feasibility of certain visualizations. For instance, our initial ambition to create a global heatmap of movie GDP was reevaluated during EDA. The complexity of accurately attributing revenue to multiple filming locations and the potential for misrepresentation led us to reconsider this visualization in favor of more straightforward, yet equally informative, visualizations.

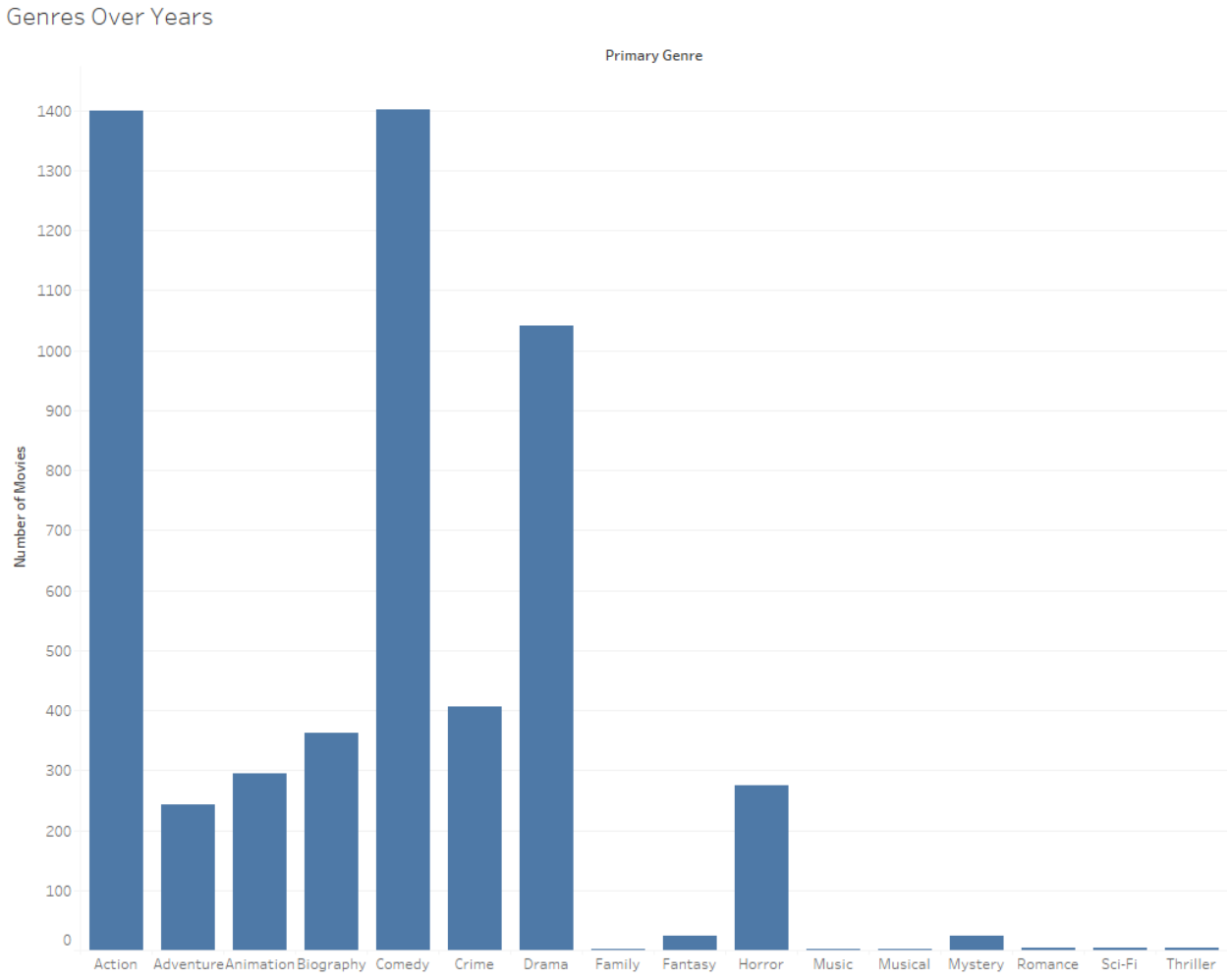
## Design Evolution

Our design process was marked by the exploration of various visualization options. Here are the designs we considered:

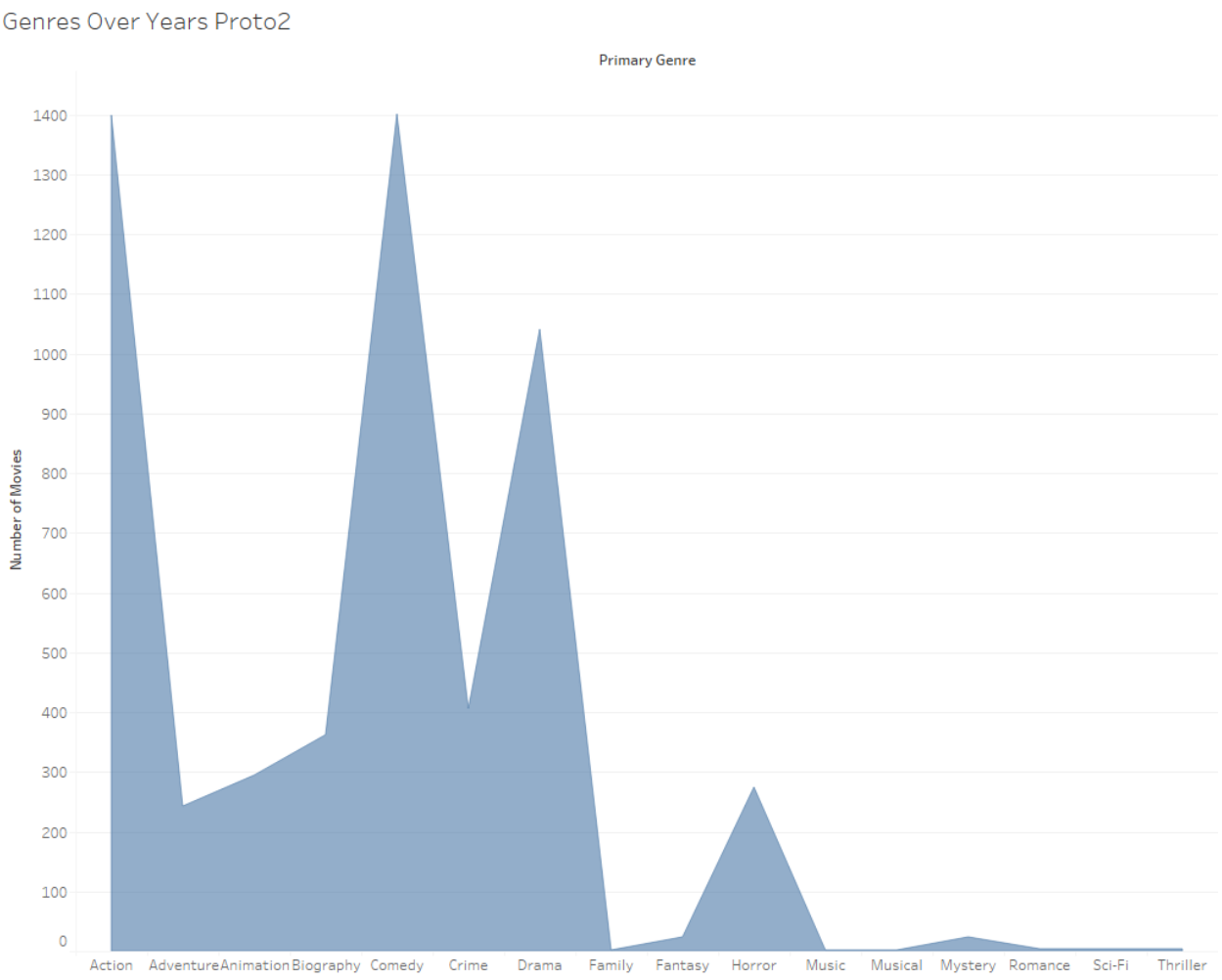
Global Heatmap of Movie GDP: Aimed to map the economic impact of movies across the globe. The complexity of attributing GDP to movies shot in multiple locations led us to discard this design.

Gross Income Trend Over Time: We'll create a line chart to show the trend of worldwide gross income over time. This will help us see if the movie industry's revenue is growing. This design was not implemented because of its simplicity and lack of intractability.

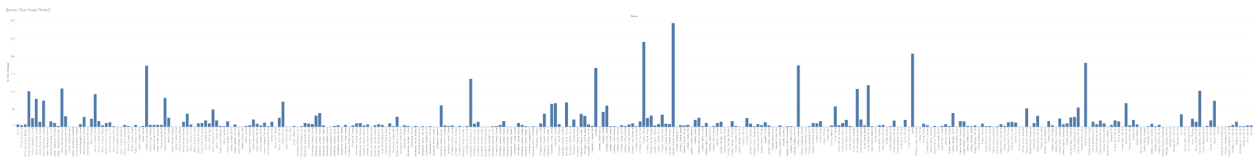
Interactive Genre Popularity Bar Chart: Initially a static representation, we decided to change this to a pie chart. The pie chart allows users to show popular genres clearer. We added an interactive feature for users to click on each pie slice to see the “Top 10 GDP movies” for the selected genre.


Genres Over Years
Primary Genre
Number of Movies
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
Action
Adventure
Animation
Biography
Comedy
Crime
Drama
Family
Fantasy
Horror
Music
Musical
Mystery
Romance
Sci-Fi
Thriller

(Img1)




(Img 2)



(Img 3)

These are the three visualizations made in reference to genre popularity over the years. The item is the movie. The attributes associated with the movie are year and genre in this case. Year is a quantitative attribute and genre is a categorical attribute. I believe the best visualization is the first (Img1). The problem with the third one (Img3) is there is too much data and too many categorical attributes associated with the item and it makes the graph unreadable. The problem with the second one (Img2) is that it makes it seem like there is a continuous trend with the data that in reality, each movie is different and not continuously related.

In the beginning, we envisioned a dashboard with intricate visualizations, including a global heatmap for movie GDP and an interactive timeline of movie releases. However, as we delved deeper into the data and its complexities, we recognized the need for a pivot. The initial designs, although visually appealing, proved to be overwhelming in communicating clear insights. Our first major design decision was to abandon the global heatmap. The diverse filming locations of movies introduced significant complexity in data representation, and attributing GDP to single locations was not straightforward. Moreover, it did not add substantial value to the core questions we were exploring. Similarly, the interactive timeline was reevaluated. While it provided a historical perspective, it was less effective in conveying the data's current state and trends. We decided to focus our efforts on visualizations that could offer more immediate insights.

The final dashboard design consists of three primary components:

- Top Movies by Worldwide Gross Income and Budget:
  - We chose a bar chart for its familiarity and simplicity, ensuring that even those with minimal exposure to data visualization could understand the information presented. The chart compares two key financial figures, allowing for quick assessments of a movie's financial success relative to its budget.
- Genre Distribution Pie Chart:
  - The pie chart was selected for its straightforward depiction of proportions. It provides an intuitive grasp of the popularity of various movie genres, presenting a snapshot of the market's current state without the need for historical data that could complicate the visualization.
- Actors Network Graph:
  - The network graph underwent the most significant evolution. Initially conceived as a static, complex visualization, it was transformed into a dynamic and interactive display. By allowing users to click on a movie and immediately see its actors' network, we provided a tool to explore the interconnectedness of the film industry. This network graph does not rely on the user's ability to manipulate data but rather invites them to explore and discover the relationships intuitively.

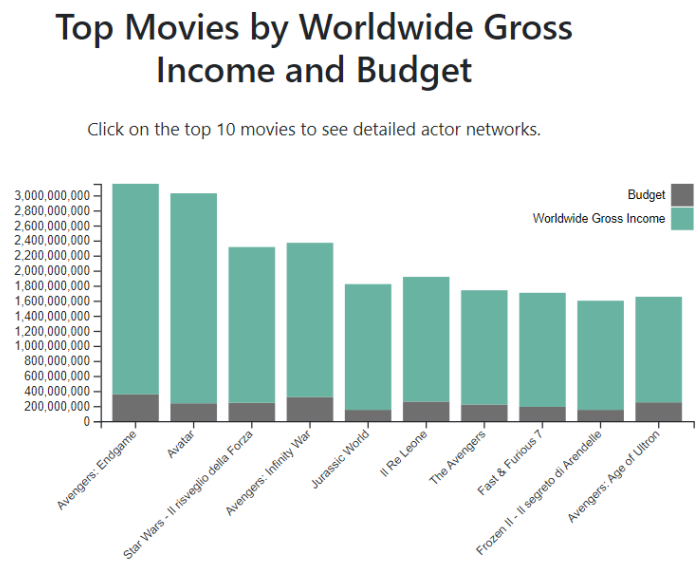
# Implementation

Our IMDb Movie Insights Dashboard is designed to be an intuitive and interactive exploration of movie data, specifically focusing on financial success, genre distribution, and the networks formed by actors across the film industry.

The dashboard is structured around three core visualizations:

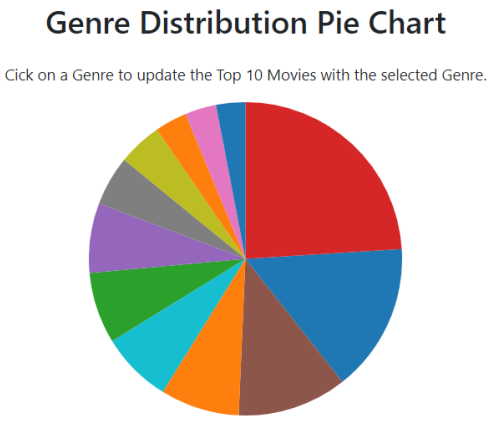
- Top Movies by Worldwide Gross Income and Budget:
  - This visualization presents a bar chart where each bar represents a movie, with the bar's length corresponding to the movie's worldwide gross income. The chart also shows a secondary measure, the movie's budget, to provide a comparison

between investment and return. Users can click on any of the top 10 movies displayed to delve deeper into the actor networks within that movie.



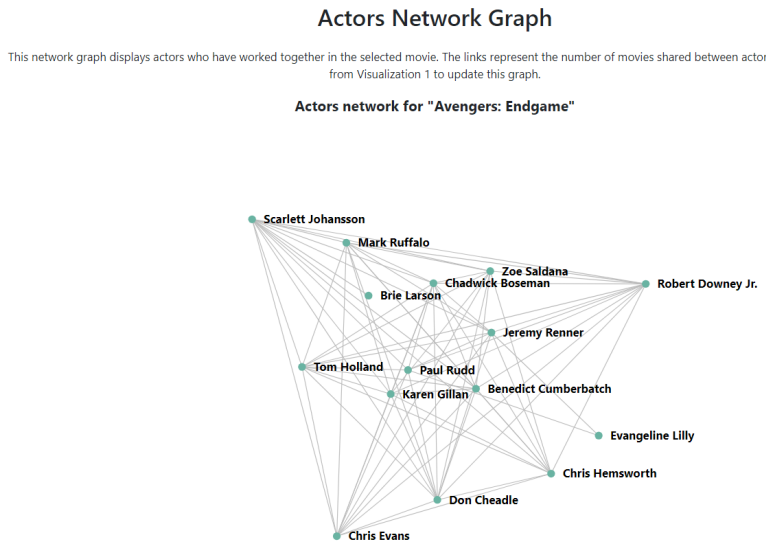

Genre Distribution Pie Chart:

A pie chart displays the distribution of movies across different genres, based on the count of movies in each genre within the dataset. When users click on a particular genre slice, the bar chart dynamically updates to show the top 10 movies within that selected genre. This interaction enables users to see which drama movies, for instance, are the highest-grossing, facilitating a genre-specific exploration of financial success.




Actors Network Graph:

Upon selecting a movie from the updated bar chart, the dashboard reveals a network graph below. This graph illustrates the connections between actors who appeared in the selected movie. Nodes represent actors, and edges denote shared movie appearances, thus providing a visual representation of the collaborative networks in the film industry.




User Interaction Flow

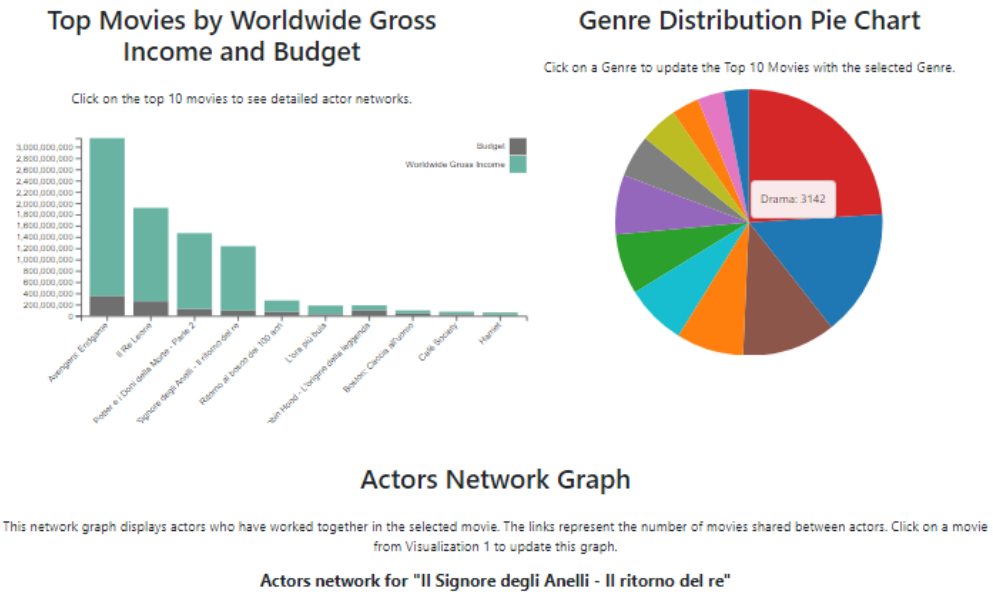
To provide an example of the dashboard's interactive capabilities, we outline a typical user interaction sequence:

A user is interested in exploring the top-performing drama movies. They click on the "Drama" slice of the genre distribution pie chart.

The bar chart responds to this action by updating to display the top 10 drama movies, ranked by worldwide gross income and budget. The visualization adjusts in real-time, offering an immediate reflection of the user's selection.

Intrigued by "The Lord of the Rings: The Return of the King," the user clicks on the corresponding bar.

The actors' network graph is then dynamically generated below the bar chart, illustrating the cast of "The Lord of the Rings: The Return of the King." The graph showcases the actors' interconnectedness, with thicker edges indicating a greater number of movies shared between actors.




# Evaluation

Throughout the development of the IMDb Movie Insights Dashboard, our evaluation process was iterative, involving both the analysis of data through visualization and the refinement of those visualizations based on insights gained.

Our visualizations led to several key insights:

Budget vs. Gross Income Discrepancy:

We discovered that a high budget does not necessarily equate to high worldwide gross income. This was particularly evident when outliers, such as low-budget films that achieved significant box office success, emerged in our visualizations. This finding challenges the common perception that financial success in movies is directly proportional to the investment made.

Genre Popularity Trends:

The Genre Distribution Pie Chart illuminated the relative popularity of movie genres. By tracking changes over time, we could identify trends, such as the rise and fall of certain genres. For instance, the resurgence of superhero movies in recent years became apparent, reflecting broader industry and cultural shifts.

Actor Networks:

Our Actors Network Graph revealed the collaborative nature of the film industry. We observed that certain actors frequently work together, suggesting a form of clustering within the industry, where certain groups of individuals tend to collaborate more closely than with the industry at large.

Answering Project Questions

The visualizations directly contributed to answering our initial questions:

- The correlation between budget and worldwide gross income was visualized through the stacked bar chart, where we could easily compare these two financial aspects for the top movies.
- Genre popularity was addressed by the pie chart, which updated the bar chart to reflect the top movies within a selected genre.
- The complexity of actor collaborations was unraveled with the network graph, which allowed us to see the connections between actors within and across movies.

Performance Evaluation

The functionality of our visualizations was robust, with real-time updates occurring smoothly upon user interaction. Users could navigate between different views without encountering any significant lag or visual glitches, which is a testament to the optimization of the D3.js code and the underlying data structures.

However, there are areas where the visualization could be improved:

- Inclusion of Time-Series Data: Incorporating a time dimension would allow users to see how movie performance and genre popularity have evolved over the years.
- Enhanced Interactivity: Adding more interactive elements, such as filtering options to include/exclude certain genres or budget ranges, would provide users with a more tailored experience.
- Mobile Responsiveness: Optimizing the dashboard for mobile devices would enhance accessibility, allowing users to interact with the data on various platforms.
- Data Expansion: Including data from streaming services would offer a more comprehensive view of the movie industry, especially as viewing habits shift towards digital platforms.

In conclusion, our evaluation demonstrates that the IMDb Movie Insights Dashboard effectively provides insights into the movie industry, allowing users to explore data on financial success, genre popularity, and actor networks. While the current implementation serves as a powerful tool for data exploration, the potential improvements identified would further enhance the user experience and provide a more complete picture of the film industry's dynamics. The dashboard

stands as a testament to the enlightening power of data visualization in uncovering and communicating complex data relationships.