Unveiling Movie Magic: Insights into Industry Trends, Audience Preferences, and Box Office Success

Group 14

ZHANG Yanfeng
u3641183@connect.hku.hk
3036411839

WAN Dingkang
u3641153@connect.hku.hk
3036411530

LI Yinghua
u3641128@connect.hku.hk
3036411281

ZHANG Hongyi
h1kari@connect.hku.hk
3036408959

HAO Xinyu
haoxinyu@connect.hku.hk
3036382002

Contents

1 Introduction

1.1 Foreword

The movie industry, a vibrant and ever-evolving domain, offers valuable insights into audience preferences, market dynamics, and creative trends. With an abundance of data available, analyzing movies from multiple dimensions provides opportunities to uncover factors that drive both critical acclaim and commercial success. Our project investigates the intricate relationships between industry trends, movie ratings, and box office performance, providing a comprehensive view of the film landscape.

Our analysis is structured around five core themes: a general overview of movies to establish foundational patterns; industry trends to identify shifts over time; factors influencing box office revenues; determinants of IMDb ratings; and the interplay between high ratings and high box office performance. The findings aim to benefit stakeholders across the entertainment ecosystem, from producers seeking to align their projects with audience expectations to platforms refining their recommendation algorithms.

1.2 Highlights

- **Comprehensive Multi-Dimensional Analysis of the Movie Industry**: The project studies the movie industry from many angles, including trends over time, what makes movies successful at the box office, and what affects IMDb ratings. It shows how these factors have shaped the industry. This analysis is helpful for producers, filmmakers, and platforms to understand what works in movies.
- **Intuitive and Interactive Visualization Techniques**: The project uses simple yet powerful charts and maps to show the data. Tools like line charts, bubble charts, and maps make it easy to see trends like how budgets and box office earnings have grown over time or which countries make the most successful movies. Interactive features let users explore specific details, like checking how different genres or years performed.
- **Giving Insights on the Global Movie Industry**: The project provides several insights on the global movie industry, such as Hollywood's massive influence on the global film market, with U.S. movies dominating both box office revenue and critical acclaim. The report also provides insights into making successful films, including rating and revenue. It emphasizes the importance of balancing factors like budget, strong storytelling, and audience appeal to create movies that perform well both critically and financially.

1.3 The Data

1.3.1 Datasets

The four raw datasets selected for the study, all from Kaggle, were used to analyze movie data and provide personalized recommendations and industry trend predictions from macro movie data, user ratings, and movie-specific information, respectively. After organizing and cleaning the original datasets, they are merged into two .csv datasets for subsequent visual analysis of the movie industry.

- **The Movie Dataset**: The dataset comprises metadata for all 45,000 movies included in the Full MovieLens Dataset, which encompasses movies released on or before July 2017. The dataset includes information on the cast, crew, plot, keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages. Additionally, it contains files with 26 million ratings from 270,000 users for all 45,000 movies, with ratings on a scale of 1-5 obtained from the official GroupLens website.

- **Top 1000 Highest Grossing Movies:** The dataset comprises information about the 1,000 highest-grossing films produced by Hollywood studios. It has been updated to reflect the most recent data as of 25 September 2023. The data has been collated from a range of sources, including the Internet Movie Database (IMDb), Rotten Tomatoes and other similar platforms, and has been aggregated for the purpose of performing various data operations.
- **Movie Dataset: Budgets, Genres, Insights:** The movies dataset is a comprehensive collection of information about 4,803 movies. It provides a wide range of details, sourced from github.com, about each film, including budget, genres, production companies, release date, revenue, runtime, language, popularity, and more.
- **IMDB 5000 Movie Dataset:** The dataset comprises detailed information about over 5,000 films sourced from the Internet Movie Database (IMDb). It encompasses a range of data points, including the cast, keywords, reviews, budgets, and other pertinent information. Of particular note is the inclusion of data from the cast's Facebook pages and associated data.

1.3.2 Data Integration Strategy

Since the above datasets provide rich information in different dimensions respectively, this study needs to organize and merge these datasets to create a comprehensive data framework covering multiple dimensions such as basic information, ratings, box office, genres, production companies, and so on, to support a wide range of analytical needs.

For the specific steps of integration, data cleansing is first required to remove duplicates, fill in missing values, and standardize the content format, followed by merging and de-duplicating the data tables based on the movie title and IMDb ID fields as key fields. Integrating multiple datasets allows for in-depth analysis across multiple dimensions of movie industry-specific information, and is more conducive to exploring the relationship between production budgets and box office revenues.

The processed dataset is divided into the following two, focusing on macro-level information and movie-specific details, respectively. The larger dataset mainly contains basic information such as movie title, release date, duration and other basic information and production information such as movie ratings and box office budget, as well as its social media situation outlining the popularity and specific performance of the movie. This data is mainly used to analyze the overall trend of the movie industry from a macro point of view, to explore the specific performance of movies in different periods and genres, and to better understand the industry development trend and user preferences. The other data is smaller and expands on the previous dataset with information on specific box office situations, distribution companies and main actors, to more deeply analyze the impact of specific characteristics of a movie on its box office and ratings.

2 Task A: General Overview of Movies

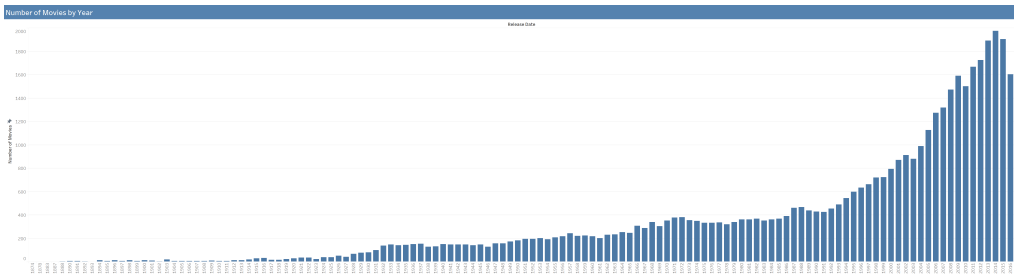
2.1 Introduction

The global film industry is vast and diverse, encompassing movies from different genres, production scales, and cultural backgrounds. To better understand the overall structure and patterns of this industry, it is crucial to first gain a comprehensive overview of the dataset. This section aims to provide insights into the foundational aspects of the movie dataset, such as the total number of films, their distribution over time, and the diversity in genres and production origins through different visualizations.

2.2 Number of Movies

We use a bar chart to illustrate the number of movies each year and its trend. The x-axis represents the year and the y-axis represents the number of movies. It can be seen that the number of movies increased steadily before the 21st century, and that the increase became significantly larger after the 21st century.

(a) (b)



Number of Movies by Region

Region	
United States of America	18,429
United Kingdom	3,072
France	2,711
Japan	1,507
Canada	1,497
Italy	1,472
Germany	1,418
Russia	800
India	780
Spain	601
Australia	506
Hong Kong	468
South Korea	457
Sweden	398
Finland	324
China	300
Belgium	299
Denmark	297
Brazil	262
Poland	245

Figure 1: Number of Movies by Year and Region

Before 21st century, the movie production is quite expensive and complicated. These limitations make it hard to produce movies, resulting the slow increase. However, when it comes to the 21st century, advances in digital filmmaking, globalization of cinema, the rise of streaming platforms, and increased accessibility significantly boosted production. Emerging markets and government incentives also played a role.

Additionally, we have designed this chart to be interactive. Users can select a specific year to view the details of the overall movie statistics for that year in the following charts.

Meanwhile, the table Figure 1(b) shows the number of movies produced by different regions. The table shows the top 20 countries in terms of the number of movies produced. The United States produced the most movies (18,429), followed by the United Kingdom (3,072) and France (2,711). It is worth noting that many movies are produced by multinational joint ventures, and here we take the first field. While this allows for a simplification of data organization, it also creates some bias in our analysis.

2.3 Movie Genres

A tree map, Figure 2(a), is employed to demonstrate the distribution of movie genres. There are possibly multiple genres for a movie, and we choose the first genre tag as the main genre of the movies. The size and the color depth represent the proportion of the genre.

And Figure 2(b) illustrates the popularity of different movie genres. Drama and comedy are not only the most popular genres but also the most produced genres, reflecting their broad appeal and frequent production in the film industry. Action follows closely, driven by its high entertainment value and ability to attract a diverse audience. Genres such as horror, adventure, and thriller show moderate popularity and a relatively small production scale, appealing to more niche but dedicated audience groups. On the other hand, genres like music, western, and mystery are the least popular, likely due to their narrower audience reach or lower production volume. Overall, the charts highlight the dominance of versatile genres while showcasing the varied preferences of movie audiences.

(a) (b)

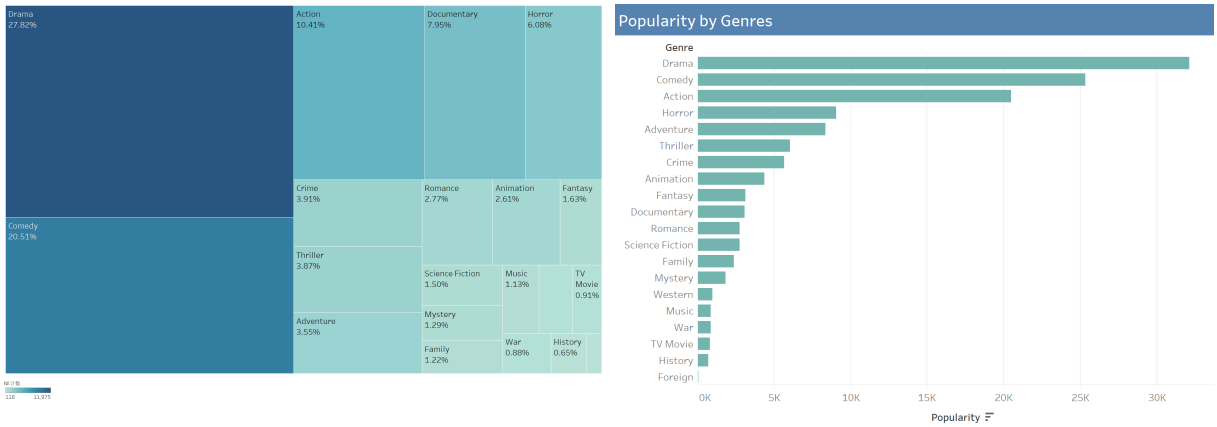


Figure 2: Number of Movies and Popularity by Genre

However, focusing on the first genre tag may overlook the impact of hybrid genres while simplifying the analysis. Therefore, it is important to conduct further data mining and in-depth visualization.

2.4 Movie Production Companies

Figure 3(a) illustrates the popularity of various film production companies through circles of differing sizes, where larger circles indicate higher popularity. Cherin Entertainment emerges as the most prominent company, significantly overshadowing its competitors and suggesting a strong influence in the industry. Newgrange Pictures, CoMix Wave Films, and 1492 Pictures are notable mid-tier players, demonstrating solid reputations without reaching Cherin's level of prominence. The diversity of companies, ranging from well-known entities like Twentieth Century Fox to smaller firms such as Pandemonium, highlights the variety within the film sector.

Figure 3(b) depicts the number of movies produced by various film companies, represented through circles of varying sizes. Each circle's size corresponds to the total number of films released by the company, providing a visual representation of their output. Companies such as Warner Bros. and Walt Disney Pictures dominate in terms of film production volume, suggesting a robust capacity for creating content. However, the average popularity of the films produced by these companies may vary significantly. For instance, while Warner Bros. has a large output, the average popularity of its films might be influenced by factors such as genre diversity and marketing strategies. Conversely, companies like Universal Pictures and Twentieth Century Fox, with fewer films, may have a higher average popularity per movie, indicating a focus on quality or blockbuster hits.

This analysis highlights the relationship between production volume and average film popularity, suggesting that both quantity and quality play crucial roles in a company's overall success in the film industry.

3 Task B: Trends of the Movie Industry

3.1 Introduction

The movie industry has experienced profound shifts in its financial landscape, characterized by changes in funding, production costs, and box office revenues. This section delves into the evolving dynamics of financial investment and returns within the industry. By analyzing the

(a) (b)

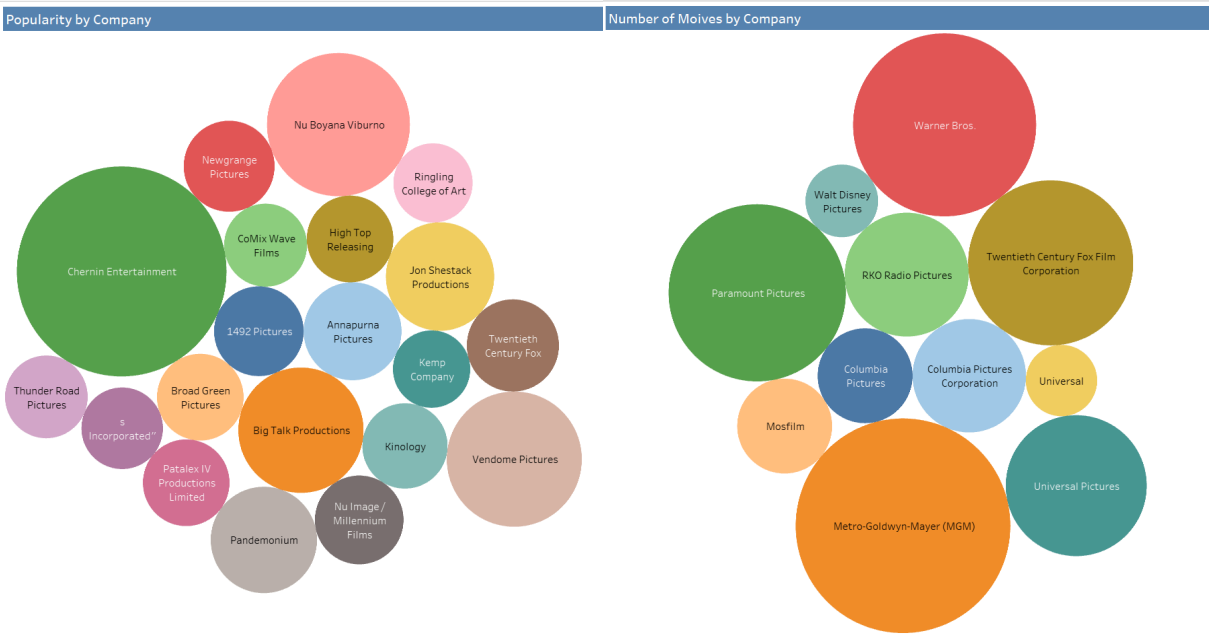


Figure 3: Movie Companies

relationship between investment and output, we aim to shed light on the flow of capital and return trends. This exploration will provide a deeper understanding of how financial strategies impact the production and success of films, offering valuable insights into the monetary mechanisms that drive the industry forward.

3.2 Movie Industry Development

Figure 4 illustrates the trends in budget and revenue in the movie industry from 1925 to 2020. In the early years, both budget and revenue remained low and stable, reflecting modest investment and returns in the mid-20th century. During the 1980s and 1990s, there was a gradual increase, indicating expanded production scales and market growth. The 2000s and 2010s saw significant spikes, highlighting the impact of blockbuster films and technological advancements, leading to substantial financial investments and higher box office returns. However, from the 2010s to 2020, there were fluctuations despite the high levels of budget and revenue, possibly due to changing audience preferences and the rise of digital streaming platforms. Overall, the chart demonstrates a strong correlation between increasing budgets and rising revenues, underscoring the growing financial investments and returns in the movie industry over time.

3.3 Budget and Revenue

Figure 5(a) illustrates the distribution of movie budgets and revenues across different years. Each rectangle represents a year, with the larger and the darker of the rectangle indicating a higher level of budget and revenue. The larger and darker blocks, particularly in the 2000s and 2010s, reflect a period of increased spending and higher box office returns, while lighter and smaller blocks in earlier years depict lower financial activity. This visualization effectively highlights the growth in financial scale within the movie industry over time.

Figure 5(b) illustrates the relationship between movie budgets and box office revenues. Each point represents a film, with the x-axis showing the budget and the y-axis indicating the revenue. The plot reveals a general trend where higher budgets can lead to higher revenues, but there is considerable variability. Some films achieve substantial revenue with moderate budgets,

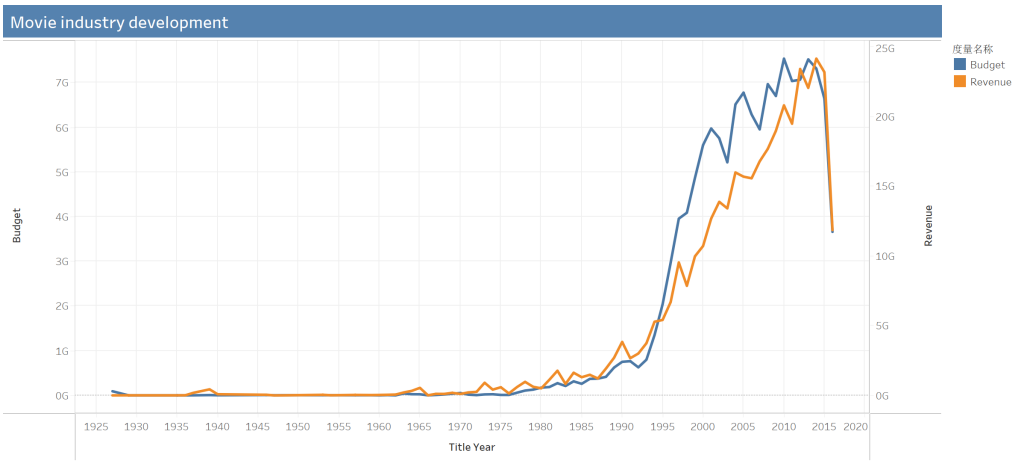


Figure 4: Movie Industry Development

while others with large budgets do not perform as well. This indicates that while budget is a factor in box office success, it's not the sole determinant.

(a) (b)

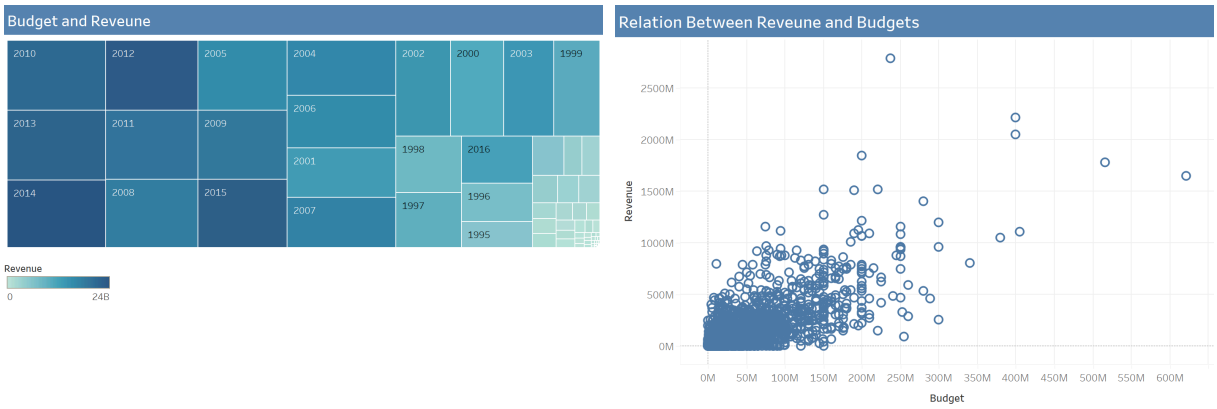


Figure 5: Budget and Revenue

Users can interact with the treemap through clicking on different years, which then links to three additional detailed charts. These charts provide an in-depth look at the geographic distribution of movie budgets and revenues for the selected year. This functionality enables a deeper exploration of how financial resources were allocated and revenue was generated across different regions, offering valuable insights into the global dynamics of the film industry for that specific year.

3.4 Company and Revenue

The pie chart displays the revenue distribution among different movie production companies, with each segment representing a company's share of total income. Different colors distinguish the contributions of each company. Major players like Walt Disney Pictures, Universal Pictures, and others are prominent, indicating their significant roles in revenue generation within the industry. This visualization effectively highlights the competitive landscape and dominance of certain studios in the movie market.

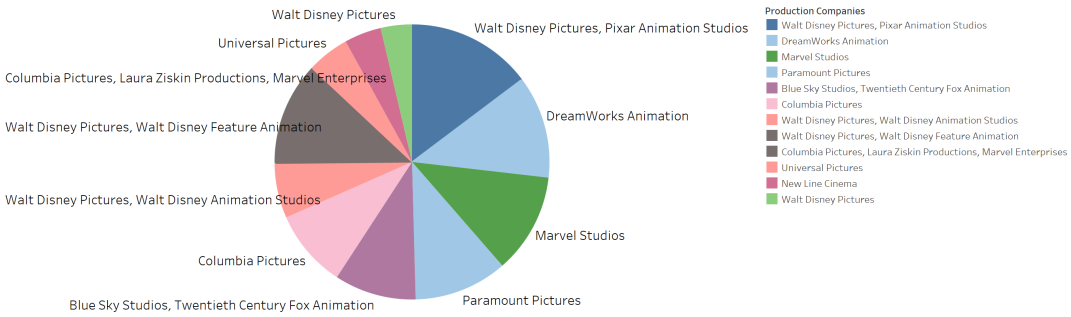


Figure 6: Companies and Revenue

3.5 Geographical Distribution of Budget and Revenue

The two maps depict the distribution of movie budgets and box office revenues across various countries and regions. Darker shades on the maps indicate higher values. The United States stands out with the deepest colors, reflecting its significant contribution to both budgets and revenues in the film industry. Other regions show varying levels of financial activity, highlighting the global nature of movie production and revenue generation. These visualizations effectively demonstrate the geographical disparities in the film industry's financial landscape.

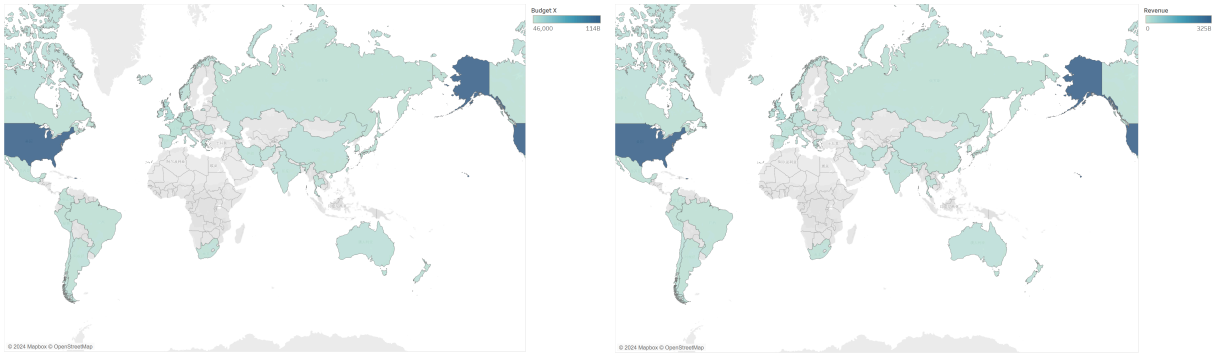


Figure 7: Geographical Distribution of Budget and Revenue

4 Task C: Factors Affecting Movie Box Office

4.1 Introduction

Understanding the factors that influence movie box office performance is crucial for stakeholders in the film industry. This section explores key elements such as marketing strategies, star power, genre preferences, release timing, and critical reviews. By analyzing these factors, we aim to provide insights into how they contribute to a film's financial success, helping filmmakers and producers make informed decisions to maximize box office returns.

4.2 Gross of Movies Published Each Year

The bar chart illustrates the total gross of films released each year from 1925 to 2020. It shows a significant upward trend, particularly from the late 1970s onwards, with noticeable peaks in the 2000s and 2010s. This increase reflects the growing scale and financial impact of the film industry over time. The darker shades in recent years indicate higher gross values, highlighting the era of blockbuster films and expanded global distribution. This visualization effectively captures the industry's growth in revenue generation across decades.

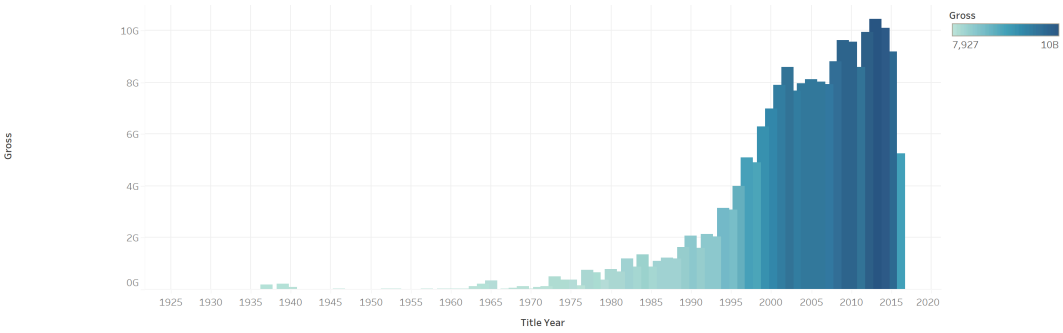


Figure 8: Gross of Movies Published Each Year

4.3 Different Genres

The bubble chart displays the gross revenue of films across different genres. Each bubble's size represents the total earnings for that genre. Action films dominate with the largest bubble, followed by Comedy and Adventure, indicating their strong box office performance. Drama, while smaller, also contributes significantly to revenue. Other genres like Crime, Horror, and Animation show varied but notable earnings. This visualization highlights the popularity and financial success of different film genres in the industry.

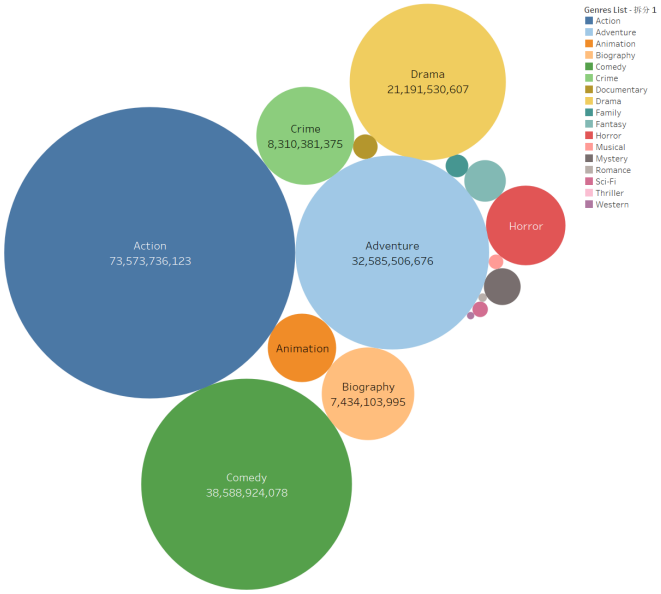


Figure 9: Gross with Different Genres

4.4 Sum of Fans, Budget and Aspect Ratio

Figure 10 (a) depicts the relationship between a movie's gross revenue and the sum of fans, measured by actor Facebook likes. Each point represents a film, with the x-axis showing the sum of fans and the y-axis indicating the gross revenue. There is a general upward trend, suggesting that films associated with actors who have larger fan bases tend to earn higher revenues. However, there is considerable variability, indicating that while fan base size can influence box office success, it is not the sole factor.

Figure 10 (b) shows the relationship between a movie's budget and its gross revenue. Each point represents a film, with the x-axis indicating the budget and the y-axis showing the gross revenue. There is a positive correlation, suggesting that higher budgets often lead to higher revenues. However, there is considerable spread, indicating that while a larger budget can contribute to box office success, it is not the sole determinant.

(a) (b) (c)

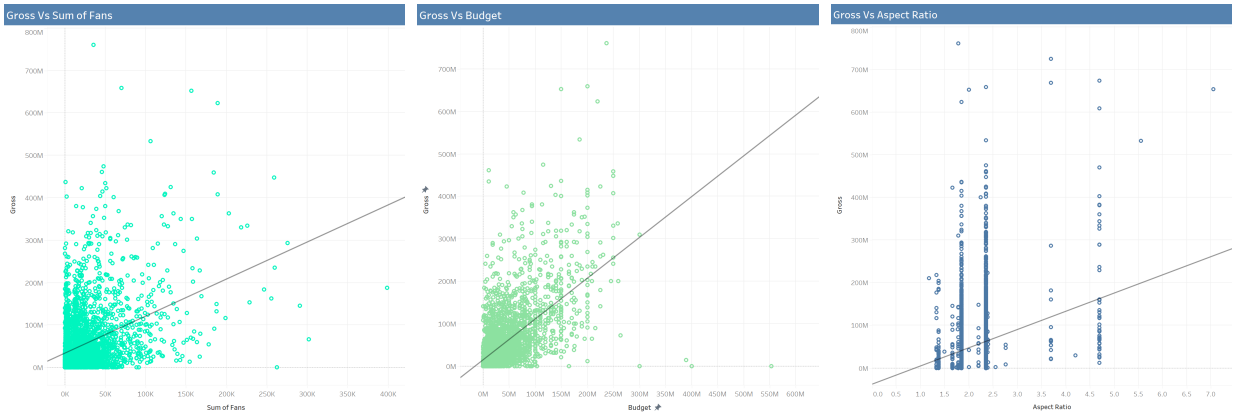


Figure 10: Gross vs. Sum of Fans, Budget and Aspect Ratio

Figure 10 (c) illustrates the relationship between a movie's aspect ratio and its gross revenue. Each point represents a film, with the x-axis showing the aspect ratio and the y-axis indicating the gross revenue. The data suggests that most films cluster around common aspect ratios, with no clear correlation between aspect ratio and box office success. While some films with typical aspect ratios achieve high revenues, this factor alone does not appear to significantly influence gross earnings.

4.5 Countries

The map shows the gross revenue of films across different countries. The shades of blue indicate the level of gross earnings, with darker shades representing higher revenues. The United States, Canada, and several European countries display significant earnings, highlighting their substantial contributions to global box office revenues. This visualization effectively illustrates the geographic distribution of film industry success, emphasizing key markets worldwide.

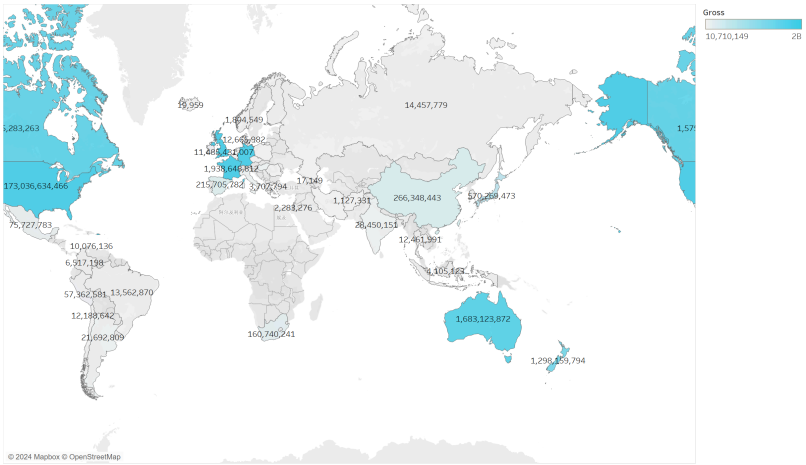


Figure 11: Gross in Different Countries

4.6 Directors and Main Actors

Figure 12(a) displays the box office gross of the top ten directors. Each rectangle represents a director, with larger areas indicating higher gross earnings. This visualization highlights which directors have achieved the most financial success in terms of box office revenue, providing a clear comparison of their impact in the industry.

Figure 12(b) shows the box office gross of films starring the top ten actors. Each rectangle represents an actor, with larger areas indicating higher gross earnings. This visualization highlights which actors have led the most financially successful films, providing a clear comparison of their impact in the industry.

(a) (b)

Figure 12: Revenue with Directors and Main Actors

4.7 Other Detailed Information

For more detailed information, we provide such things as country, movie title, language and box office receipts in a tabular format.

5 Task D: Factors Affecting Movie's IMDb Ratings

5.1 Introduction

IMDb ratings are a widely recognized metric for gauging audience sentiment and critical reception, making them a valuable indicator of a movie's quality and appeal. This section explores the factors that influence these ratings, aiming to uncover the underlying patterns and correlations that contribute to high or low audience scores. By analyzing key attributes such as release year, budget, runtime, and director popularity, we aim to identify trends and characteristics shared by highly rated films.

In this specific part, we divide the factors which influence the IMDb rating into 4 parts, including budget, content rating, region and era.

5.2 Budget

A scatter graph is employed to explore the relationship between a movie's budget and its IMDb rating. This analysis aims to identify whether higher production budgets consistently correlate with better audience reception or if low-budget films occasionally achieve critical acclaim.

From Figure 13, it is easy to see that movies with low IMDb ratings tend to have rather low budgets, while movies with relatively high budgets (e.g. over $180,000,000) can achieve at least a moderate rating (e.g. over 5.0).

We can also see that movies with quite high IMDb ratings (e.g. *The Godfather*, *Fight Club*) have quite low budgets compared to others. At the same time, movies with a super high budge, for example, *Pirates of the Caribbean: On Stranger Tides*, have an unsatisfactory rating, which may not correspond to their budget.

We have several conjectures about this phenomenon.

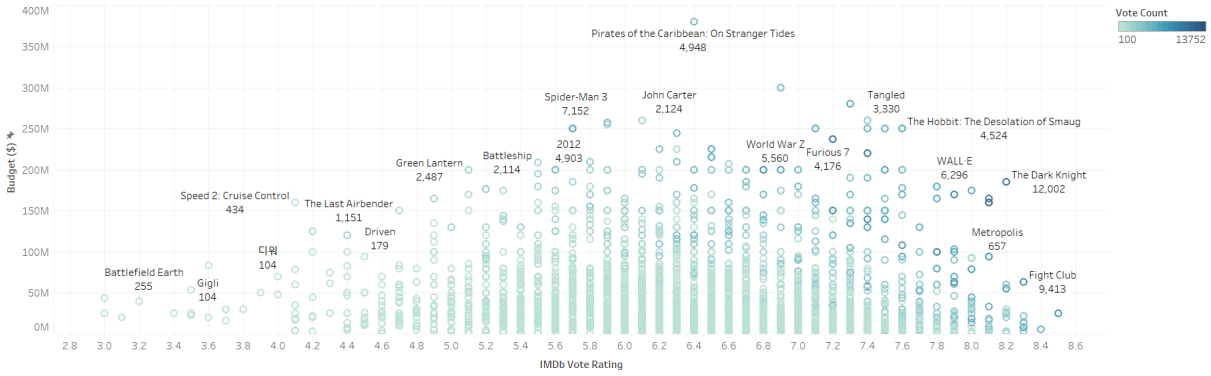


Figure 13: IMDb Rating vs. Budget

- For those high-budget movies, they often prioritize commercial success over artistic value. This strategy makes them more focus on broader audience, leads to more special effects, action sequences, and star power. These elements may not always meet audience expectations for deep storytelling or innovation.
- For those low-budget movies, their directors have more room for experimentation with themes and expression, attracting audiences who value artistic quality and compelling narratives. At the same time, these films often rely on word-of-mouth promotion. Their evaluations are more likely to reflect the quality of the story and emotional resonance rather than visual spectacle.
- And for audiences, they may have higher expectations for high-budget movies, while the "pleasant surprise" effect may cause higher ratings for low-budget films.

5.3 Regions

Utilizing a map chart, we show the geographical diversity of movie production by visualizing the average IMDb ratings of movies from different regions. Users can also check the rating distribution and proportion of movies whose ratings are above 8.0 in different regions by clicking their locations.

Figure 14 illustrates the average ratings of movies in different regions. It's noticeable that the distribution of ratings in the United States is similar to the overall distribution of global film ratings, which reflects the global influence of Hollywood. U.S. films are heavily marketed and distributed worldwide, shaping the tastes and expectations of international audiences. As a result, global rating patterns naturally align with those of U.S. films.

It is also worth mentioned that the regions that we don't think produce great films can have a fairly high average rating. That might be due to their selective international exposure and appeal to niche audiences. This creates a disparity between their rating patterns and those of U.S. films, which are subject to a broader range of criticism and viewer expectations.

5.4 Content Rating

Bar charts are used to showcase the average IMDb ratings of films categorized by their content ratings (e.g., G, PG, R). This comparison helps identify which content categories resonate most with audiences and whether content restrictions influence viewer perceptions.

The chart shows that PG-13-rated movies have a high film count, indicating their dominance in the movie industry. This is likely because PG-13 films target a broad audience, including teenagers and adults, making them highly marketable. R-rated films have the highest count, reflecting their popularity among adult audiences, while PG-rated films are also significant,

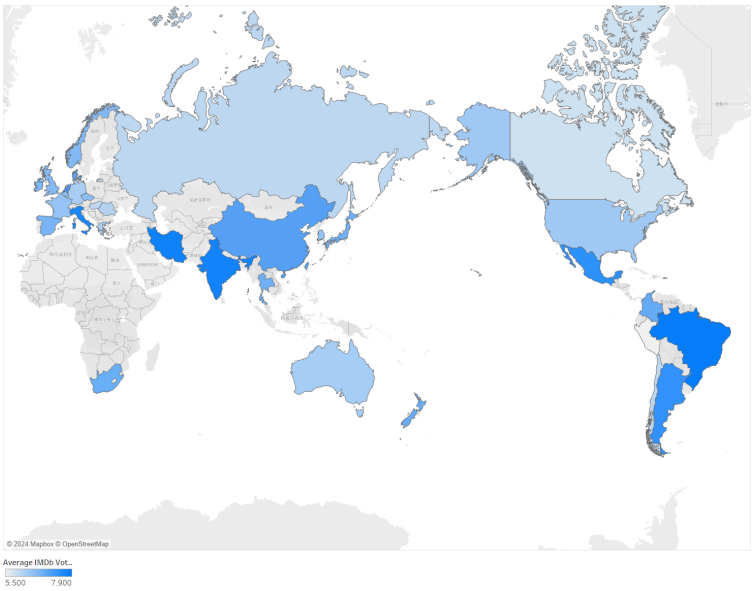


Figure 14: IMDb Ratings vs. Regions

appealing to family audiences. The count of G-rated (General Audiences) and X-rated (Adults Only) films is very low, suggesting that these categories are either less commonly produced or face limited market demand.

Despite their low count, X-rated films achieve the highest average IMDb ratings. This aligns with their niche audience appeal and selective exposure, as discussed earlier. These films are often rated by viewers who appreciate their specialized content, resulting in higher scores. Despite their high production volume, PG-13 films exhibit the lowest average IMDb ratings, likely due to their focus on mass-market appeal, which can lead to diluted quality or overexposure to criticism.

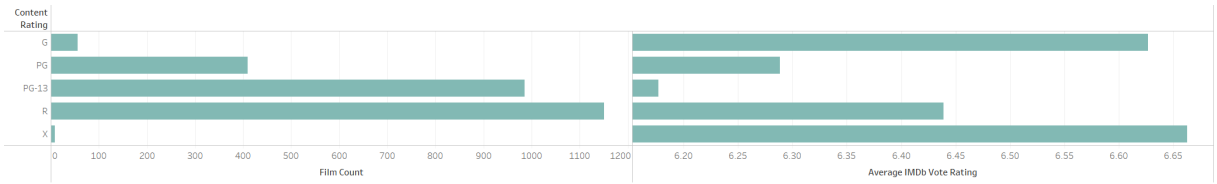


Figure 15: IMDb Ratings vs. Content Ratings

5.5 Era

By employing bar and pie charts, this analysis investigates IMDb ratings across different eras, revealing trends in audience reception over time. Additionally, the proportion of highly rated films (ratings above 8.0) in each era is examined to understand the evolution of critically acclaimed filmmaking.

From Figure 16, we can obtain that the earlier movies can get a relatively higher rating compared with recent ones. The reasons behind the phenomenon might be the nostalgia and the historical significance of old movies, especially those from 1950 or earlier. Viewers and critics often hold these films in higher regard due to their cultural and historical value.

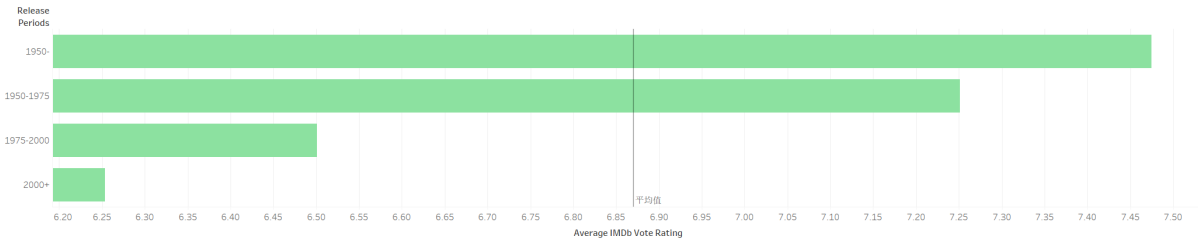


Figure 16: IMDb Rating of Movies Produced in Different Eras

The changes in production models can also be a significant factor. As the film industry grew in the 1975-2000 period, the focus shifted toward blockbuster franchises and commercial appeal, potentially leading to a dilution of quality. Furthermore, post-2000 films face more diverse and critical audiences with higher expectations, particularly due to the ease of access to films globally. Modern audiences often compare films across a broader spectrum, leading to more polarized ratings.

6 Task E: Analysis of the Relationship between High Ratings and High Box Office

6.1 Introduction

The relationship between high ratings and box office success is a key focus for the film industry. This analysis explores how critical acclaim and audience approval correlate with financial performance. By examining data on film ratings and box office earnings, we aim to uncover patterns that indicate whether high ratings consistently lead to higher revenues. Understanding this relationship can provide valuable insights for filmmakers and marketers seeking to optimize both the artistic and commercial success of their films.

6.2 Gross Distribution

The provided visualization represents the distribution of box office revenue across a range of movies. A significant 75.43% of the total box office revenue comes from films that earned less than 100 million dollars, as shown by the dominant bar on the left side of the graph. Beyond this range, revenue contributions drop sharply. The distribution follows a long-tail pattern, where a small number of blockbuster films generate extremely high revenue, but the majority of movies earn significantly less.

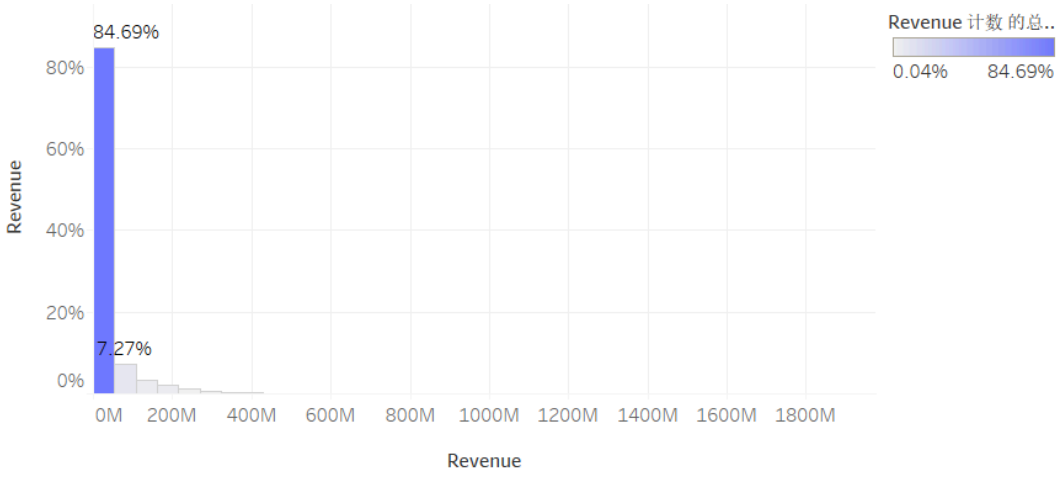


Figure 17: Gross Distribution

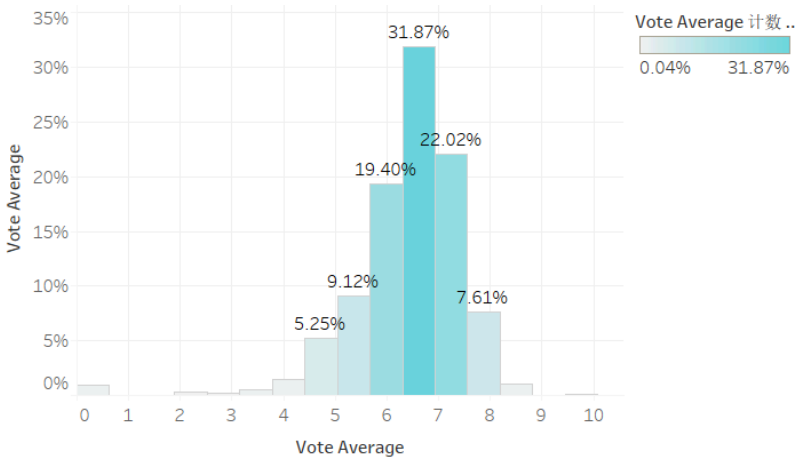


Figure 18: Rating Distribution

This suggests that while a few high-grossing films attract attention, the bulk of the industry's overall performance is driven by a large number of lower-grossing films. Therefore, understanding and optimizing the success of these lower-revenue films could be crucial for sustaining the overall profitability of the movie industry.

6.3 Rating Distribution

The bar chart displays the overall distribution of movie ratings, with the horizontal axis representing the rating scores and the vertical axis showing the percentage distribution of these ratings. The data reveals that over 90% of movies have ratings between 4.5 and 8.0, indicating that most films fall within the range of medium to relatively high scores. The peak of the distribution occurs between 6.0 and 6.5, where 27.77% of the movies are clustered. This suggests that while there are relatively few extremely low or extremely high-rated films, the majority of movies receive moderate evaluations from viewers, with a concentration around the 6 to 7 rating range.

6.4 The relationship between Gross and Rating

The scatter plot (Figure 19) explores the relationship between movie box office revenue and ratings. The horizontal axis represents the average rating, while the vertical axis represents the box office revenue. A trend line is included in the chart, showing a generally positive correlation between the two variables. This suggests that movies with higher ratings tend to have higher box office earnings.

There are, however, two significant outliers: *Avatar* and *Star Wars: The Force Awakens*, both of which had exceptionally high grosses compared to other films, despite not having the highest ratings. These two outliers warrant separate analysis, as their success may be due to factors beyond just ratings, such as franchise popularity or large fan bases. Overall, while higher ratings are generally associated with higher revenue, blockbuster films like these can achieve extraordinary box office success regardless of their specific ratings.

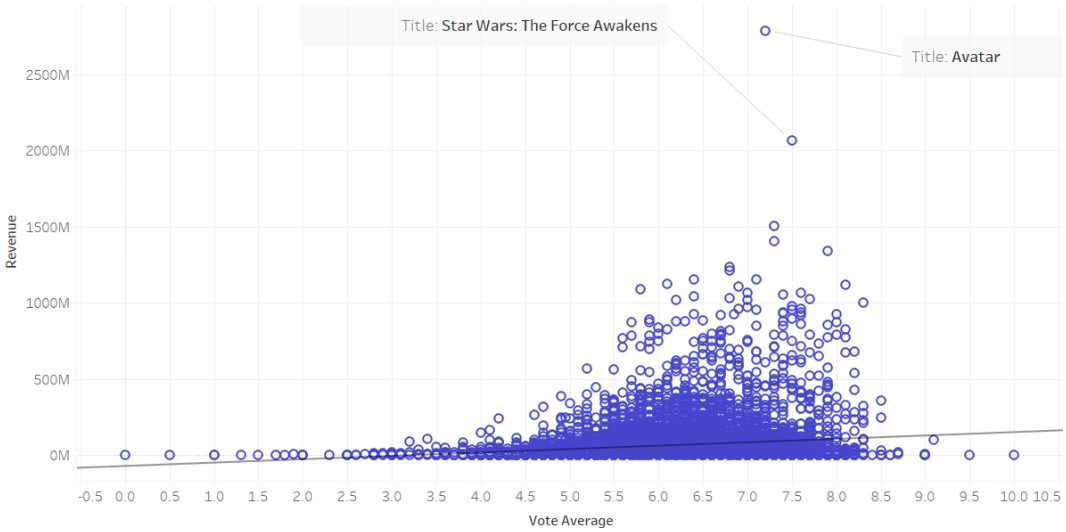


Figure 19: The Relationship between Revenue and Rating

6.5 Genres of highly rated and high-revenue films

This pie chart presents the distribution of genres among 22 highly rated (with ratings above 8.0) and high-grossing (box office above 100 million) films. The data reveals that **Drama** (31.82%) and **Adventure** (22.73%) are the genres most likely to achieve both high ratings and strong box office performance. Other notable genres include **Comedy** (9.09%) and **Romance** (4.55%), though they are less dominant. This indicates that films in the Drama and Adventure genres tend to appeal to both critics and audiences, making them more likely to achieve success in both ratings and revenue.

6.6 Countries of highly rated and high-revenue films

Figure 21 illustrates the countries where highly rated (ratings above 8.0) and high-grossing (box office above 100 million) films were produced. The data shows that the majority of these films come from the United States, which dominates the global film industry. Europe, particularly countries like the United Kingdom and France, also contributes a notable portion. This distribution highlights the strong global influence of Western cinema, particularly from the U.S. and Europe, suggesting that Western culture continues to be highly popular and influential worldwide in terms of both critical acclaim and box office success.

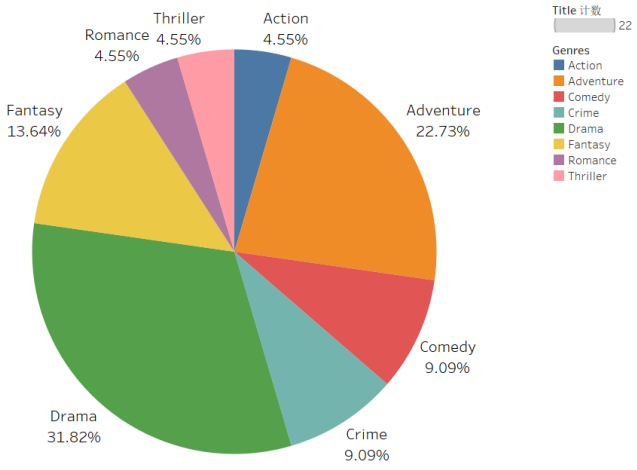


Figure 20: Genres of Highly Rated and High-revenue Movies

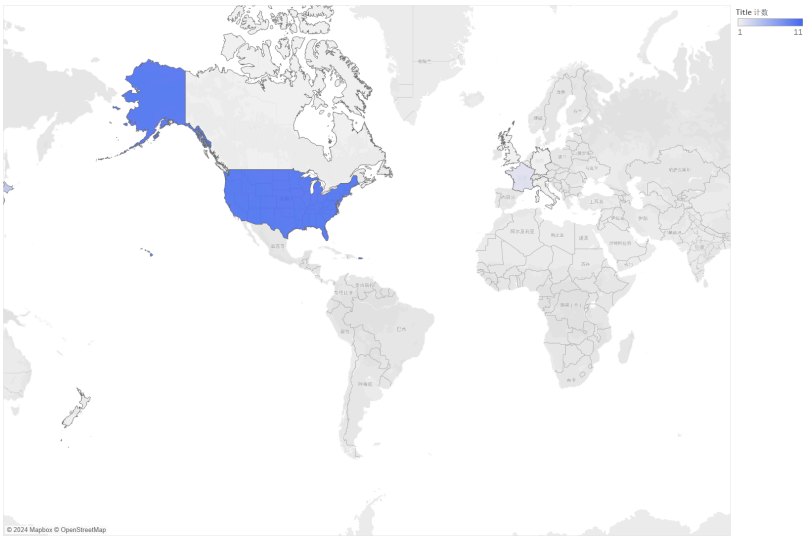


Figure 21: Countries of Highly Rated and High-revenue Movies

7 Summary and Analysis

7.1 Advantages of the visualization approach

For the analysis of the movie industry, the Tableau tool was selected for this study to perform a multi-dimensional visualization, choosing a number of different types of charts and graphs to reveal overall trends and key factors in the film industry. The following section analyzes the advantages of these specific visualization methods and explains why these chart formats are particularly effective when analyzing film data.

1. Line Chart: Used to show annual trends in movie numbers, budgets and revenues.
   - Trend analysis: line charts are well suited for presenting time-series data and can clearly reflect trends in the data. In the analysis of movie numbers and box office revenues, line charts help observe the rapid growth of movie production and investment after the 21st century.
   - Multiple Line Comparison: In the analysis of budget and box office revenue, line charts can show multiple data lines at the same time, making it easy to compare the trends of different data items and reveal the relationship between budget and box office.
2. Tree map: Analyze the distribution of movie genres, box office performance of different directors or lead actors.
   - Hierarchical Structure and Proportion Display: The tree diagram can display the hierarchical structure and proportion of the data at the same time. When used to analyze the distribution of movie genres, the proportions of different types of movies can be clearly seen, such as the largest proportion of drama and comedy movies.
   - Efficient use of space: Compared with bar charts and pie charts, tree charts utilize space more efficiently and are suitable for displaying a large number of categories of data, especially when the data has more dimensions and needs to be compared.
3. Bubble Chart: Analyze the number of production, popularity and box office revenue of different types of movies produced by movie studios.
   - Multi-dimensional display: Bubble Chart displays data in multiple dimensions at the same time through bubble size, position and color, such as the number of production and popularity of movie studios, which can visualize the comparison between different companies.

- Suitable for displaying distribution and concentration: It can help to identify concentration trends and outliers in the data, for example, certain companies have a high number of movies but not a high popularity.

4. Map Visualization: Show movie production, box office revenue and IMDb ratings for each country around the world.
   - Spatial data analysis: Map visualization visualizes geographic distribution characteristics and is ideal for analyzing the performance of the film market on a global scale. Users can visualize countries with higher box office revenues, such as the United States and China, through color shades.
   - Enhanced User Experience: The map visualization is highly interactive, allowing users to click on different countries to view detailed information and increase user engagement.
5. Interactive Visualization: Linkage analysis between multiple charts, e.g. filtering the corresponding data display according to year, type or company changes.
   - Enhance analysis efficiency: The interactive feature enables users to filter by specific dimensions (e.g., year, movie genre) to quickly locate and analyze data of interest.
   - Dynamic Exploration of Data: By clicking on a specific genre or year, users can dynamically view the changes in the relevant data and dig deeper into the trends and reasons behind the data.
   - Personalized Analysis: Interactive visualization allows users to adjust the filtering criteria according to their needs for personalized in-depth analysis, enhancing the flexibility of data exploration.

Overall, Tableau, as a powerful data visualization tool that combines a variety of chart forms. Supports the visualization of different data types, capable of displaying numerical, sub-typed, time-series and geographic data at the same time to meet the needs of multi-dimensional analysis.Through the combination of colors, sizes, shapes and other visual elements, it makes the trends and outliers of the data clear at a glance, which effectively improves the efficiency of data analysis.Supporting linkage between charts and dynamic filtering functions, users can adjust filtering conditions in real time according to their needs for personalized analysis and improve the depth of data exploration.

7.2 Other visualization methods

In this project, I chose visualization methods such as scatter plots, bubble plots, tree plots, and maps to achieve a comprehensive presentation of multi-dimensional data and trend analysis. These methods are able to balance information density and readability, while further enhancing the user experience through Tableau's interactive features. Compared with other commonly used visualization methods, these charts have obvious advantages in data expression, user-friendliness and depth of analysis, the specific comparison is as follows:

1. Box Plot: Box plots are ideal for analyzing data distribution, outliers, and interquartile range, but they are less effective for trend analysis over time. Their interpretative threshold is higher, making them less intuitive for users without a data analysis background, especially compared to line graphs, which better highlight trends such as box office revenue changes.
2. Bar Chart: Bar charts can appear crowded when presenting many categories like movie genres or production companies, as horizontal comparisons become unclear. Tree charts better utilize space by showing hierarchical structures. Additionally, bar charts struggle with time-series trend analysis, where horizontal or line charts provide a clearer visualization of growth or decline.

3. Heat map: While heat maps effectively display value magnitudes through color, they lack precision for comparisons, such as specific ratings or box office figures, where scatter or bubble charts are more intuitive using position and size. Heat maps also face limitations in visualizing multi-dimensional data, unlike scatter or bubble charts, which combine position, size, and color to analyze relationships between features like ratings, box office, and budget.

The chosen visualizations, including scatter charts, bubble charts, tree diagrams, and maps, balance multi-dimensional analysis with trend visualization. They combine high information density with clarity, effectively showing differences across categories and regions. Tableau enhances interactivity with filters and hover prompts, allowing users to explore data intuitively and comprehensively.

7.3 Shortcomings and reflections

While the project effectively utilized Tableau and cleaned datasets to deliver comprehensive visualizations, some limitations in data coverage, analytical depth, and tool capabilities constrained the scope of analysis. Addressing these gaps can further enhance the quality and insights of future projects.

1. Data Limitations: The datasets lacked completeness in key features (e.g., budgets, regional box office, and audience ratings) and omitted dynamic elements like social media trends and time-series data. Expanding data sources, incorporating APIs, and using predictive models to fill gaps could enhance analysis depth and accuracy.
2. Missed Analytical Dimensions: The absence of detailed regional box office data and audience sentiment limited the exploration of market-specific performance and emotional feedback. Integrating regional data and applying NLP for sentiment analysis could reveal nuanced insights into audience preferences and word-of-mouth effects.
3. Tool Constraints: Tableau's limitations in advanced statistical modeling restricted deeper predictive analysis. Combining programming tools like Python or R for modeling with Tableau's visualization strengths could improve both analytical rigor and presentation clarity.

Despite some limitations, the project successfully showcased multi-dimensional insights and highlighted key trends in the movie industry. Addressing data gaps, incorporating advanced analysis techniques, and leveraging complementary tools will pave the way for more robust and impactful future analyses.

8 Contribution

	ZHANG Yanfeng	WAN Dingkang	LI Yinghua	ZHANG Hongyi	HAO Xinyu
Proposal			✓	✓	
Data Preprocessing	✓	✓			
Visualization & Analysis	Task C	Task E	Task B	Task D	Task A
Final Report			✓	✓	
Presentation					✓