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

Name: WAN Dingkang Uid: 3036411530 Group: 14

1. Completed Tasks

task 1: Data collection

task 2: Data preprocessing

task 3: Visualization & Analysis of the Relationship between High Ratings and High Box Office

2. Data collection

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.

- 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.

3. Data preprocessing

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

I am mainly responsible for the data cleaning work of the first dataset: "The Movie Dataset" and the second dataset: "Top 1000 Highest Grossing Movies". Because many fields in these two datasets are stored in JSON format, such as Genre and cast. Therefore, in this process, I mainly use Python to parse these JSON contents and convert them into comma separated strings for subsequent visualization using Tableau. At the same time, it is necessary to identify duplicate values and merge them, while most of the missing values are retained to prevent excessive deletion from affecting the analysis. Except for the revenue gap, it will be deleted because this field is the core data in our analysis process.

Afterwards, because the first dataset actually had 7 CSV files, for the convenience of analysis, all files need to be merged. In the experiment, I merged these 7 files in Tableau with movie ID as the keyword to form a dataset containing all the data and fields.

4. Analysis of the Relationship between High Ratings and High Box Office

4.1 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.

4.2 Visualization and Analysis

4.2.1 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 below. Beyond this range, revenue contributions drop sharply, with only 9.86% coming from the next group of films. 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. 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.

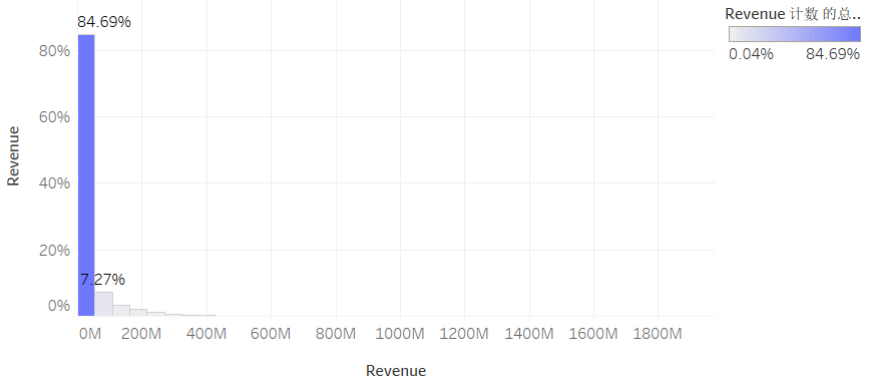

Figure 1: Gross Distribution

4.2.2 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, which follows a normal distribution, 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.

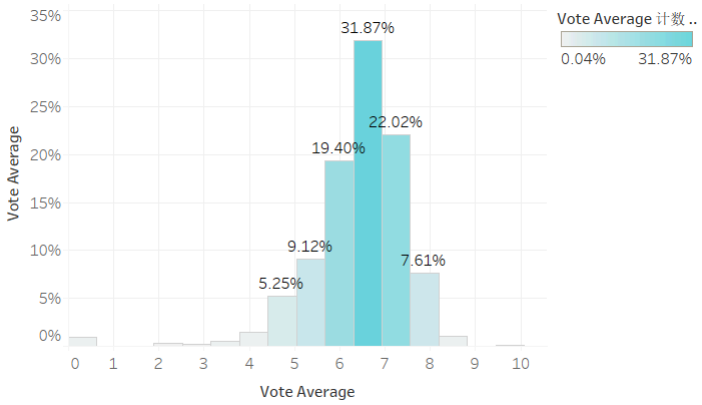

Figure 2: Rating Distribution

4.2.3 The relationship between Gross and Rating

This scatter plot 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.

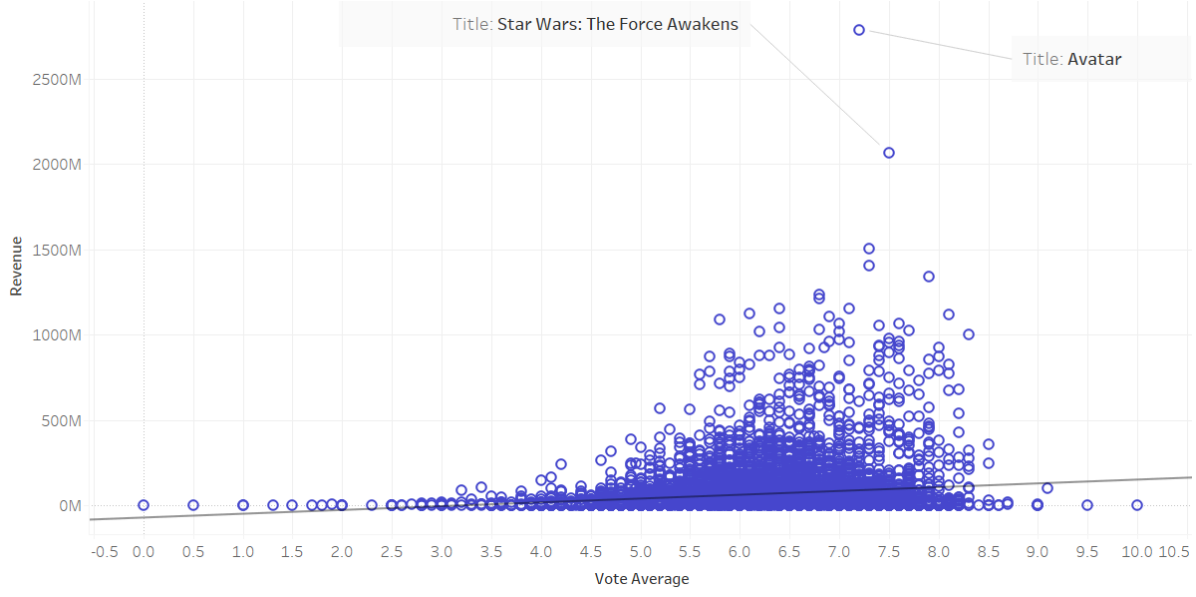

Figure 3: The Relationship between Revenue and Rating

4.2.4 Genres of highly rated and high-grossing 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. At the same time, clicking on any genre, and other images will also change according to the selected genre, showing gross distribution, rating distribution and the relationship between gross and rating under a specific genre.

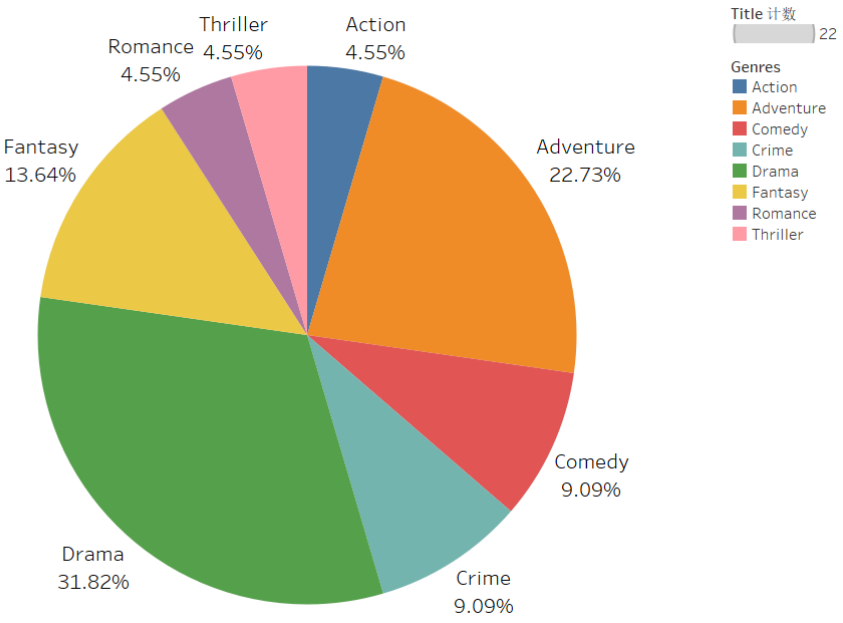

Figure 4: Genres of Highly Rated and High-revenue Movies

4.2.5 Countries of highly rated and high-grossing films

This map 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. After America and Europe, Japan has the most movies, indicating that Japan also has its own unique appeal in the cultural field. At the same time, clicking on any country will cause other images to change based on the selected country, displaying the gross distribution, rating distribution, and the relationship between gross and rating for a specific country.

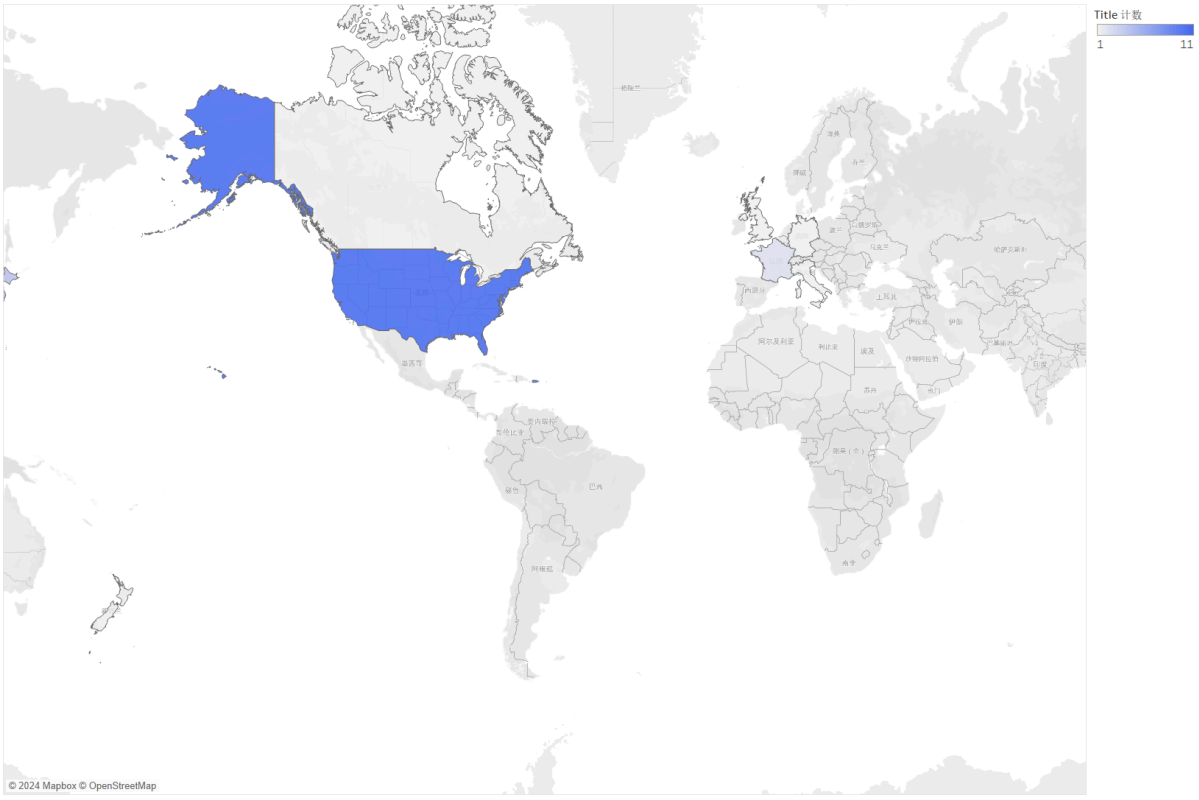

Figure 5: Countries of Highly Rated and High-revenue Movies

5 Summary

Overall, I and my team members completed this assignment together. We not only completed our assigned tasks, but also communicated progress and issues with each other every week. After this experiment, I became proficient in using Tableau for data preprocessing and visualization, and gained a more intuitive understanding of various charts. After analyzing the relationship between high ratings and high box office, it can be concluded that although the gross distribution follows a long-tail pattern and the rating distribution follows a normal distribution, the two are still roughly positively correlated and strongly influenced by factors such as genres and production countries.