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: This dataset contains 7 CSV files:
    - credits.csv: 3 columns with 45504 pieces of data, including information on the actors and staff of the movie.
    - keywords.csv: 46419 pieces of data in 2 columns, including movie keywords, stored in JSON format.
    - links.csv: 3 columns of 45843 data, including the correspondence between movie IDs and imdbIDs, tmdbIDs.
    - links_small.csv: 3 columns of 9125 data, subset of links.csv.
    - movies_metadata.csv: 24 columns with 45455 pieces of data, involving the content, production, and evaluation of movies.
    - ratings.csv: 4 columns with 26024289 records, including user rating data.
    - ratings_small.csv: 4 columns with 100004 pieces of data, including some user rating data, subset of ratings.csv.
- Top 1000 Highest Grossing Movies: This dataset contains a Highest Holywood Grossing Movies.csv file with 14 columns and 1000 entries, including movie content, budget, box office, and ratings.
- Movie Dataset: This dataset contains a movie_dataset.csv file with 24 columns and 4803 entries, including movie content, revenue, and reviews
- IMDB 5000 Movie Dataset: This dataset contains a movie_stetata.csv file with 28 columns and 5043 entries. In addition to various movie details, it also includes the number of likes given to actors

3. Data preprocessing

Due to the fact that the above datasets provide rich information from different dimensions, for the convenience of research, we need to clean and merge these data to create a comprehensive dataset covering multiple dimensions such as basic information, ratings, box office, genres, production companies, etc., to support a wide range of analysis needs.

I am mainly responsible for cleaning the data 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, in order to quickly split them in Tableau, I need to parse these fields in JSON first. Considering that Tableau can also segment strings, the processing of these nested JSON contents is relatively rigid and inefficient. Therefore, in this process, I mainly use Python loops to process these nested JSON contents and convert them into comma separated strings for visualization using Tableau in the future. At the same time, it is necessary to identify duplicate values and merge them, while retaining most of the missing values 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, due to the fact that the first dataset actually had 7 CSV files, all files needed to be merged for ease of analysis. In the experiment, I merged these 7 files in Tableau with movie IDs as keywords 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 the final step in our group's visualization process and a key focus in the film industry. My goal is to demonstrate whether there is a correlation between movie ratings and box office revenue by examining the data, and which factors affect this correlation. Understanding these can provide valuable insights for filmmakers and marketers seeking to optimize film art and commercial success. Therefore, my analysis mainly consists of the following five steps.

4.2 Visualization and Analysis

4.2.1 Gross Distribution

Before studying the relationship between high box office and high ratings, we need to first define what kind of box office is high and what kind of ratings are high. So the first step is to study the distribution of box office revenue.

The distribution of box office revenue for movies is shown in the following bar chart, with 50M as a data bucket, and the visualization also excludes movies with zero revenue and zero vote counts. It can be seen that its distribution follows a long tail pattern, with most films having extremely low box office revenues and a very small number of films contributing extremely high box office revenues. As shown in the figure, about three-quarters of movies have a box office of less than 50M, about 85% of movies have a box office of less than 100M, and the highest grossing movie can exceed 2700M. This indicates that although some high box office movies have attracted people's attention, the overall performance of the industry is mainly driven by a large number of low box office movies. Therefore, understanding and optimizing the success of these low-income

films is crucial for maintaining the overall profitability of the film industry. In later high box office research, I will study movies that have a box office of over 100M.

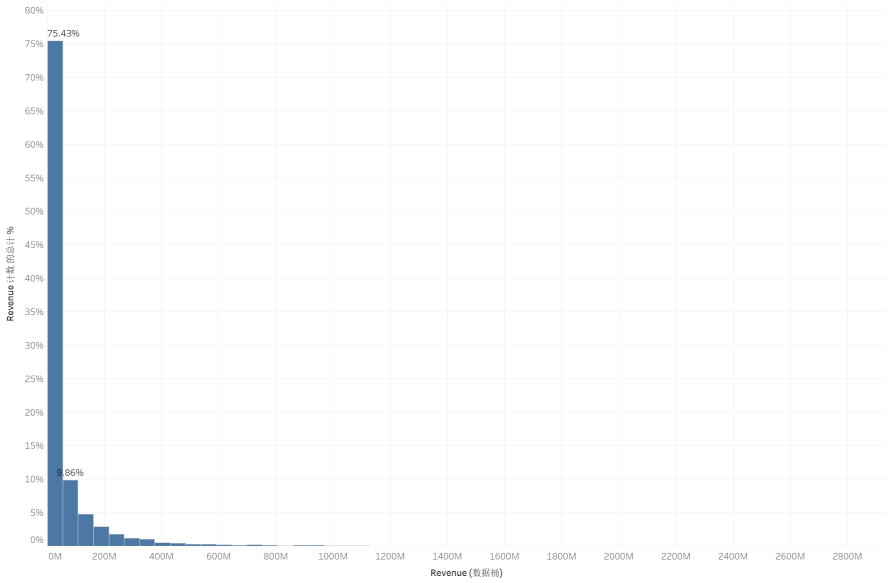

Figure 1: Gross Distribution

4.2.2 Rating Distribution

Next, we will study the distribution of ratings to determine which movies are high rated.

The following bar chart shows the overall distribution of movie ratings. It can be seen that the distribution follows a normal distribution, with more movies rated in the middle and fewer movies rated on both sides. Data shows that over 90% of movies have ratings between 4.5 and 8.0, indicating that although there are relatively few movies with extremely low or high ratings, most movies receive moderate ratings from audiences. And approximately 6% of movies have a rating greater than 8.0, which is also the high rated movie I would study later.

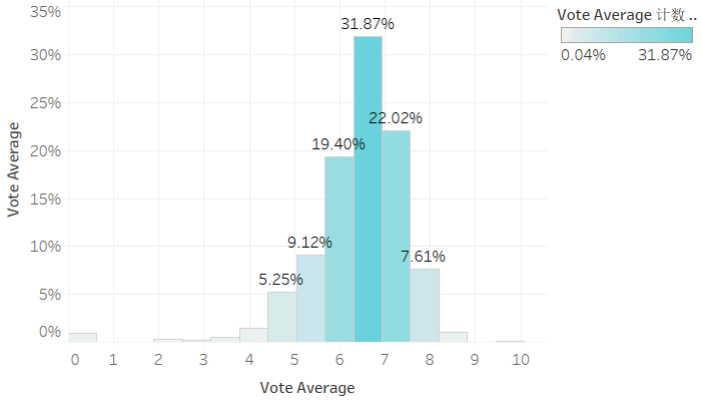

Figure 2: Rating Distribution

4.2.3 The relationship between Gross and Rating

This step begins to reveal the relationship between box office and ratings, and aims to derive the correlation between high box office and high ratings.

The following scatter plot explores the relationship between movie box office revenue and ratings. The chart contains a trend line that indicates a generally positive correlation between two variables. This indicates that movies with higher ratings often have higher box office revenues. At the same time, it can be seen that the slope of this trend line is not large, which is consistent with the distribution of box office and ratings in the previous two steps of analysis, that is, a small number of high-quality movies contribute more box office revenue, while audience ratings are

relatively conservative. This indicates that although box office revenue is generally positively correlated with ratings, high box office revenue is also influenced by many other factors.

From the graph, it can be seen that there are two obvious outliers: Avatar and Star Wars: The Force Awakens. Although the ratings are not the highest, compared to other movies, the box office of these two movies is very high. These two outliers are worth analyzing separately, as their success may be due to factors beyond the rating, and the relevant factors will be analyzed next.

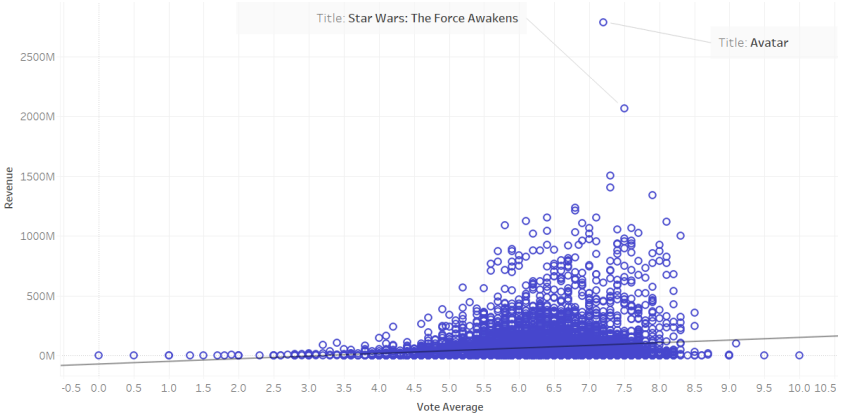

Figure 3: The Relationship between Revenue and Rating

4.2.4 Genres of highly rated and high-grossing films

The next step is to investigate how high box office and high rated movies are influenced by other factors. I have observed that other team members frequently analyzed the impact of film genres and production country in their separate studies on box office and ratings, so I will also conduct visual analysis of these two factors in the following two steps.

Firstly, select movies with high box office (box office exceeding 100 million) and high ratings (ratings exceeding 8.0). Although there are about 10% of movies with high box office or high ratings when setting up separate filters, only 22 movies were selected when setting up these two filters together, which is enough to illustrate the scarcity of high-quality movies.

The following pie chart shows the genre distribution of 22 high rated (ratings over 8.0) and high box office (box office over 100 million) movies. Data shows that drama (31.82%) and adventure (22.73%) are the most likely genres to achieve both high ratings and strong box office performance simultaneously. This indicates that theatrical and adventure films often appeal to both critics and audiences, making them more likely to succeed in terms of ratings and revenue.

Afterwards, by clicking on any style in the dashboard, you can see that other images will also change according to the selected genre, displaying the gross distribution, rating distribution, and the relationship between gross and rating under a specific genre.

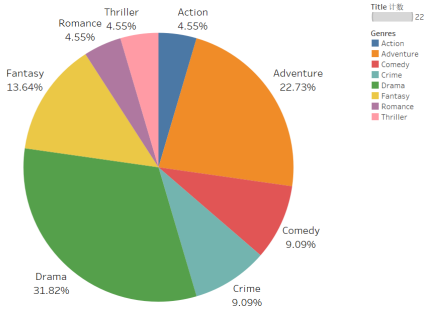

Figure 4: Genres of Highly Rated and High-revenue Movies

4.2.5 Countries of highly rated and high-grossing films

The final step is to analyze the impact of the production country on high rated and high box office films, in order to illustrate the influence of the region on culture.

The following map shows the countries that produce high rated (8.0+) and high box office (100 million+) movies. The data shows that the top three sources of these movies are the United States, Europe, and Japan, which are recognized as the dominant players in the global film industry and even culture. The figure shows that only high rated and box office movies from the United States account for 50%, indicating that Western culture is still very popular and influential globally in terms of critical acclaim and box office success. After the United States and Europe, Japan has the most movies, indicating its unique appeal in the cultural field. This indicates that popular movies nowadays are indeed deeply influenced by national culture and values.

Meanwhile, 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.

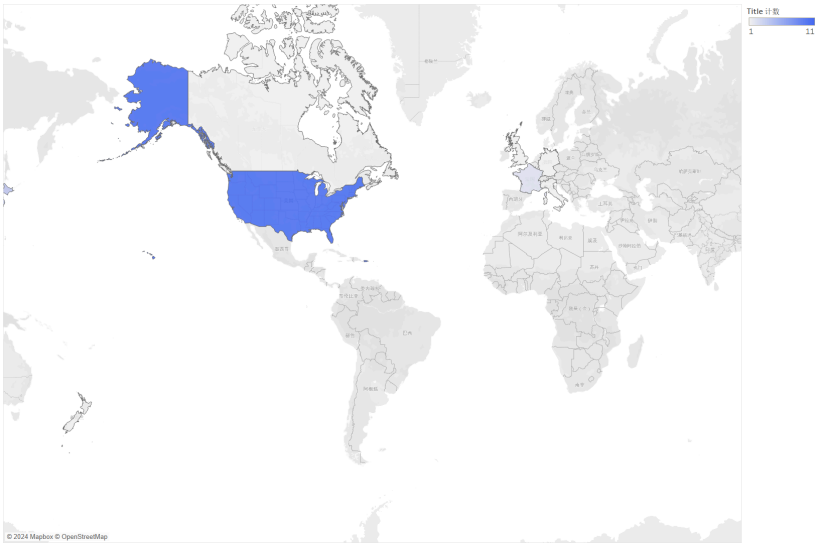

Figure 5: Countries of Highly Rated and High-revenue Movies

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.