Unveiling Movie Magic: Insights into Industry Trends,

Audience Preferences, and Box Office Success

Group 14 ZHANG Yanfeng 3036411839

1. Overview

Our group project focuses on the movie data, aiming to explore the changes in the film industry over recent years and to identify a series of factors influencing the box office and rating. My work is mainly in time arrangement, data collection, data cleaning and visualization tasks. Ultimately, we completed the data visualization for the project. Through this process, I gained a deeper understanding of the entire data visualization workflow.

2. My tasks in the project

2.1 . Data collection

At the beginning of the project, I searched for datasets and analyzed every fields of them. By doing this, I explore the relationships between different fields and identified the parts that needed to be divided for out project.

2.2 . Data cleaning

During the data processing, I addressed issues such as redundancy and invalid data by performing data cleaning using Python. Finally, I merged all the movies based on their titles, resulting in several CSV files suitable for data visualization.

During the data cleaning process, I was responsible for the *'movie_dataset.csv'* and *'movie_metadata.csv'* datasets. I removed fields with missing values that could not be imputed using the mean or median, such as movie genres and directors. For the remaining missing data, I replaced them with the mean to minimize their impact on the results. Additionally, I extracted and transformed JSON-formatted data from certain fields, making it suitable for data visualization. Finally, I merged my cleaned dataset with the dataset cleaned by another team member.

```python
import pandas as pd
import numpy as np
import json

# 1. 读取数据
file_path = 'movie_dataset.csv'
data = pd.read_csv(file_path)
```

```python
merged_df = df1.merge(df2, on='title', how='inner')
```

Figure 1:python code

2.3 . Visualization Tasks

In the data visualization phase, I focused primarily on factors influencing box office performance. Throughout this process, I uses various visualizations such as scatter plots, bar charts, and tree maps for analysis.

3. Visualization and Analysis

Box office performance is one of the key indicators of a film's success. Many production companies aim to create high-grossing films, which requires understanding the key elements that make a successful commercial blockbuster. In this section, I will analyze the factors influencing box office revenue and draw conclusions based on the visualizations.

3.1. Gross of Movies Published Each Year

In this section, I used a bar chart to represent the total box office revenue for each year. The reason for choosing a bar chart is that the variation in color and height provides a more intuitive visualization of the revenue levels, making it easier to compare earnings across years.

The bar chart displays the annual total gross of films released from 1925 to 2020, revealing a clear upward trend. This growth becomes particularly pronounced from the late 1970s onward, with notable peaks during the 2000s and 2010s. The darker bars in recent years represent higher gross values, underscoring the rise of blockbuster films and the expansion of global distribution. This visualization effectively illustrates the film industry's increasing scale and financial impact over the decades.

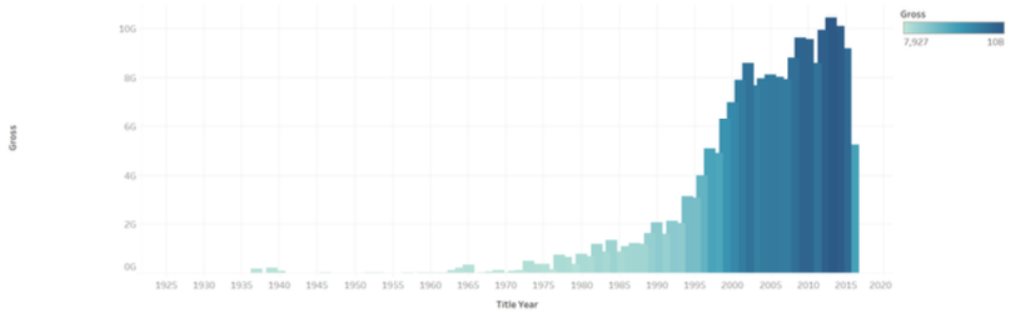

Figure 2:Gross of Movies Published Each Year

3.2. Different Genres

The bubble chart illustrates the gross revenue of films across various genres, with the size of each bubble representing the total earnings for that genre. Action films lead with the largest bubble, followed by Comedy and Adventure, indicating their strong box office performance. Although smaller, Drama also contributes significantly to revenue. Other genres, such as Crime, Horror, and Animation, show diverse yet notable earnings. This visualization highlights the financial success and popularity of different film genres within the industry.

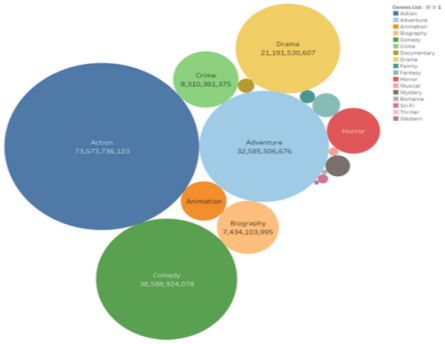

Figure 3:Gross with Different Genres

3.3. Sum of Fans, Budget and Aspect Ratio

In this section, I used three scatter plots to illustrate the relationship between various factors and box office performance. The distribution of the points and the trend lines provide a clearer, more intuitive way to identify the correlations between box office revenue and the influencing factors.

The first graph illustrates the relationship between a movie's gross revenue and the total number of fans, as measured by actor Facebook likes. Each point represents a film, with the x-axis depicting the sum of fans and the y-axis showing the gross revenue. There is a general upward trend, suggesting that films associated with actors who have larger fan bases tend to generate higher revenues. However, significant variability exists, indicating that while fan base size may influence box office success, it is not the sole factor.

The second graph presents 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. A positive correlation is evident, suggesting that higher budgets often result in higher revenues. However, the data shows considerable spread, implying that while a larger budget can contribute to box office success, it is not the only determining factor.

The third graph demonstrates 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 representing the gross revenue. The data indicates 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.

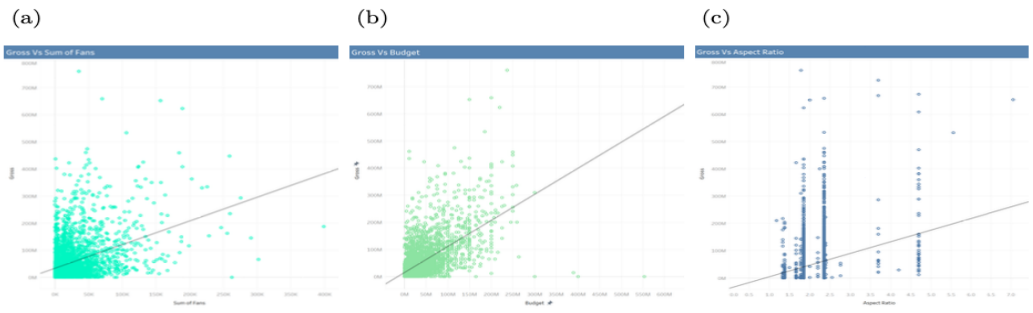

Figure 4:Gross vs. Sum of Fans, Budget and Aspect Ratio

3.4. Countries

In this section, I used a map visualization to illustrate the relationship between box office revenue and different regions and countries. The box office analysis reveals which countries have the largest film markets, and the map provides a more intuitive way to present these conclusions.

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 5:Gross in Different Countries

3.5. Directors and Main Actors

In this section, I used a tree map to visualize the top ten directors and actors based on box office performance. The size and color distribution of the tree map provide a more concise and clear representation of the data.

The first graph presents the box office gross of the top ten directors. Each rectangle represents a director, with larger areas corresponding to higher gross earnings. This visualization highlights the directors who have achieved the greatest financial success in terms of box office revenue, offering a clear comparison of their impact on the industry.

The second graph displays 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 emphasizes which actors have led the most financially successful films, providing a clear comparison of their influence within the industry.

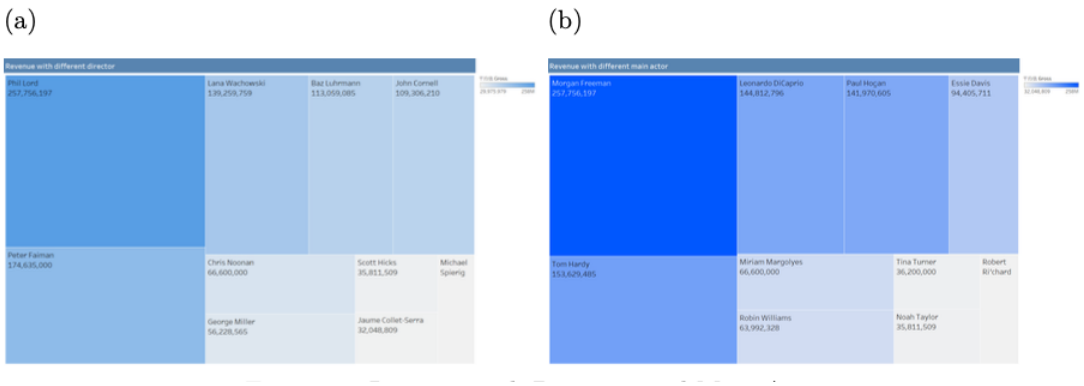

Figure 6: Revenue with Directors and Main Actors

4. Summary

After completing the movie data visualization project, I gained a deeper understanding of how data analysis and visualization can uncover key trends and patterns in the film industry. Through tasks such as data cleaning, field analysis, and the use of various visualization techniques like bar charts, scatter plots, and tree maps, I was able to explore the factors influencing box office performance. I learned how to effectively manage and visualize large datasets, and how to draw meaningful conclusions from them. This project also enhanced my skills in using tools like Python for data processing and visualization, and helped me better appreciate the complexity of factors such as director influence, budget, and genre popularity in determining a film's success.