Ella Zaputil

Professor Michael Redmond

Data Wrangling

2 April 2025

Final Project Proposal – Analyzing Factors Contributing to Box Office Success

**Introduction**

The film industry is a complex environment that generates substantial annual revenues, making it crucial to understand the factors that contribute to a movie’s financial success. In 2022, global theater box office revenues reached approximately $25.9 billion (IMDB). ​Given the film industry's substantial growth, it is increasingly important to understand how factors such as production budget, directorial influence, star power, genre, and release timing affect a movie's financial performance. A well-allocated production budget can enhance a film’s quality and marketing reach, potentially boosting its visibility. The director’s vision and reputation often shape audience expectations, while casting a well-known actors will attract larger audiences due to their established fan bases. Certain genres typically draw substantial crowds, and strategic release timing, such as debuting during holidays or summer months, can significantly impact ticket sales. This analysis aims to explore how these factors interact and contribute to the film’s overall revenue and box office performance.

**Data**

For this project, I will use data from the “IMDB Movies Dataset” from Kaggle, which provides detailed information about the top 1000 movies on IMDB. This dataset includes fields such as movie title, release date, certificate, runtime, genre, IMDB rating, film overview, Metascore, director, top 4 stars, number of IMDB votes, and gross revenue. The original dataset can be viewed [here](https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows). Because the certificate, overview, Metascore, and IMBD votes will not contribute to the analysis, I will be dropping them from the final dataset. Additionally, I will be scraping data from “The Numbers” website, which provides financial information about films, including film title, ranking, production budget, domestic gross revenue, and worldwide gross revenue. The website I will be scraping data from can be accessed [here](https://www.the-numbers.com/movie/budgets/all).

I plan to download the “IMDB Movies Dataset” from Kaggle and scrape additional data from The Numbers website using Python and pandas. After obtaining both datasets, I will horizontally integrate them into a single, unified data frame, ensuring a more comprehensive overview of each film. This integration will allow for in-depth analysis of the financial factors contributing to a movie’s success.

*Figure 1 Data Dictionary*

|  |  |  |
| --- | --- | --- |
| **Field** | **Data Type** | **Description** |
| Movie Title | Text | Name of film |
| Release Year | Integer | Year the film was released |
| Runtime | Integer | Length of film in minutes |
| Genre | Text | Genre(s) of film |
| IMDB\_Rating | Float | Rating of film on IMDB website |
| Director | Text | Name of film’s director |
| Star1, Star2,  Star3, Star4 | Text | Name of film’s top 4 stars |
| Ranking | Integer | Ranking of film |
| Release Date | Date | Date film was released including month, day, year |
| Production Budget | Integer | Money budgeted to produce film |
| Domestic Gross | Integer | Total amount of money movie earned from ticket sales in America |
| Worldwide Gross | Integer | Total amount of money movie earned from ticket sales internationally |

**Proposed Analysis**

This analysis aims to explore several key research questions related to the financial success of movies. First, how does a film’s production budget impact its gross revenue? A higher budget could lead to better production quality and marketing reach, potentially resulting in greater earnings. To test this, I will use correlation analysis to examine the relationship between production budget and revenue, and create a linear regression model to determine the strength and significance of this relationship.

Second, I want to determine whether certain directors consistently produce higher-grossing films. Examining the relationship between a director’s past success and a film’s revenue may reveal trends in audience following and directorial influence. To analyze this, I will use grouped summary statistics to compare the average and median earnings of films by various directors. This will allow me to see a correlation between the number of films directed and gross earnings.

Finally, I want to find how release timing and genre impact gross revenue performance. Some genres may perform better during specific seasons, such as action films during the summer or family movies during the holidays. I will use pivot tables and bar charts to compare the revenues across different genres and release months. By applying multiple statistical methods, including summary statistics, regression models, and hypothesis testing, this project will provide a comprehensive analysis of the drivers behind a movie’s financial success.

References

*Global box office notched 27 gain in 2022 to hit 26 billion total, research shows - IMDB*. (n.d.). IMDb. <https://www.imdb.com/news/ni63899899>

*Full IMDB dataset (1M+)*. (2025, April 2). <https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows>

*The numbers - movie budgets*. (n.d.). The Numbers. <https://www.the-numbers.com/movie/budgets/all>