Analysis of Popular Film Data via The Movie Database (TMDb) API

Effy Tang & Weiqi Yang

2025-12-01

Contents

1 Abstract

The performance of the film industry has become increasingly important in recent years. The aim of this study is to examine how a film's production budget, primary genre, and release year influence its box office performance and audience reception. A dataset of 164 films was compiled, cleaned, and analyzed using statistical and visual methods based on data obtained from the TMDb API. The data showed that, in terms of box-office revenue, films with high budgets tend to correlate with high revenues ($r = 0.604$). Additionally, the average audience ratings varied significantly by film genre, and there has been a strong trend toward increased popularity of films released in recent years. A Shiny interactive application was developed to provide an interactive tool for further exploring these relationships. The findings of this study highlight both the practical implications of API-based movie data and some of the structural trends and patterns in film performance observed today.

2 Introduction

Film production decisions, such as budget allocation, genre selection, and release timing, are made under uncertainty. However, they still have a huge influence on the movie's commercial and critical success. Therefore, it becomes important for both researchers and practitioners in the industry to understand which measurable characteristics are related with high revenue or strong audience reception.

This study addresses three main questions: the extent to which production budget is associated with box-office revenue, whether audience ratings differ systematically across genres, and what temporal patterns characterize the release and popularity of contemporary films. To answer these questions, data were gathered from the TMDb API which provided current metadata on films that are widely watched. The study presents insights integrating data management, statistical visualizations, and interactive tools that reveal the underlying dynamics of popular cinema today.

2.1 Research Question

We investigate **which movie characteristics (production budget, primary genre, and release year) are most strongly associated with a film's commercial success (measured by revenue) and audience reception (measured by vote average)**.

Understanding these relationships can help film studios make more informed decisions about resource allocation, genre selection, and release timing. This analysis uses real-world data from The Movie Database (TMDb), one of the most comprehensive film databases available.

2.2 Data Source

The data for this project comes from **The Movie Database (TMDb) API**, which provides access to information about thousands of movies, including:

- Production budgets and box office revenue
- User ratings and vote counts

- Release dates and genres
- Popularity metrics

We access this data using the TMDb API endpoint at https://developer.themoviedb.org/docs/getting-started.

2.3 GitHub Repository

The complete code, data, and documentation for this project are available at: https://github.com/effytang/Tang-Yang-Final_project

3 Data Collection

3.1 Setup

We begin by loading the necessary R packages and setting up our API key.

3.2 API Functions

We created two main functions to collect data from the TMDb API:

1. `get_popular_movies()` - Retrieves basic information about popular movies
2. `get_movie_details()` - Gets detailed information (budget, revenue) for specific movies

3.3 Collecting Basic Movie Data

We collect basic information from 10 pages of the API, with each page containing approximately 20 movies, giving us around 200 movies total.

```
## Starting to collect movie data...
## Fetching page 1 of 10
## Fetching page 2 of 10
## Fetching page 3 of 10
## Fetching page 4 of 10
## Fetching page 5 of 10
## Fetching page 6 of 10
## Fetching page 7 of 10
## Fetching page 8 of 10
## Fetching page 9 of 10
## Fetching page 10 of 10
##
##  Successfully collected 200 movies
```

The basic endpoint provides general information like title, release date, and vote average, but **does not include budget and revenue data**. Therefore, we need to make individual API calls for each movie.

3.4 Collecting Detailed Information

To get budget and revenue data, we query each movie individually. This is more time-intensive but necessary for our analysis.

```
## Collecting detailed info for 200 movies...
## This may take 1-2 minutes.
##
## Progress: 20 / 200
```

```
## Progress: 40 / 200
## Progress: 60 / 200
## Progress: 80 / 200
## Progress: 100 / 200
## Progress: 120 / 200
## Progress: 140 / 200
## Progress: 160 / 200
## Progress: 180 / 200
## Progress: 200 / 200
##
##  Successfully collected details for 200 movies
```

4 Data Management

4.1 Merging Datasets

We merge the basic information with the detailed data and create a `release_year` variable for temporal analysis.

```
## Dataset dimensions: 204 rows × 11 columns
```

4.2 Data Cleaning

Many movies in the TMDb database have missing or zero values for budget and revenue (particularly older films or independent productions). We filter these out to ensure our analysis is based on complete financial data.

```
## After cleaning:

## Remaining movies: 164

## Removed: 40 movies with missing data

##
## Budget range: $ 70,000 to $ 400,000,000

## Revenue range: $ 184,758 to $ 2,923,706,026

## Year range: 1958 - 2025
```

4.3 SQLite Database

We store the cleaned data in a SQLite database for efficient querying using SQL. This demonstrates data management skills beyond simple data frames.

```
##
## Database created successfully!

## Tables: movies

## Rows in movies table: 164
```

4.4 Data Exploration with SQL

We use SQL queries to explore our dataset and generate summary statistics.

```
## === Top 5 Movies by Revenue ===
```

```
##                       title    budget    revenue release_year
## 1                    Avatar 237000000 2923706026         2009
## 2 Avatar: The Way of Water 350000000 2353096253         2022
## 3                   Titanic 200000000 2264162353         1997
## 4                 Ne Zha 2  80000000 2150000000         2025
## 5   Avengers: Infinity War 300000000 2052415039         2018
```

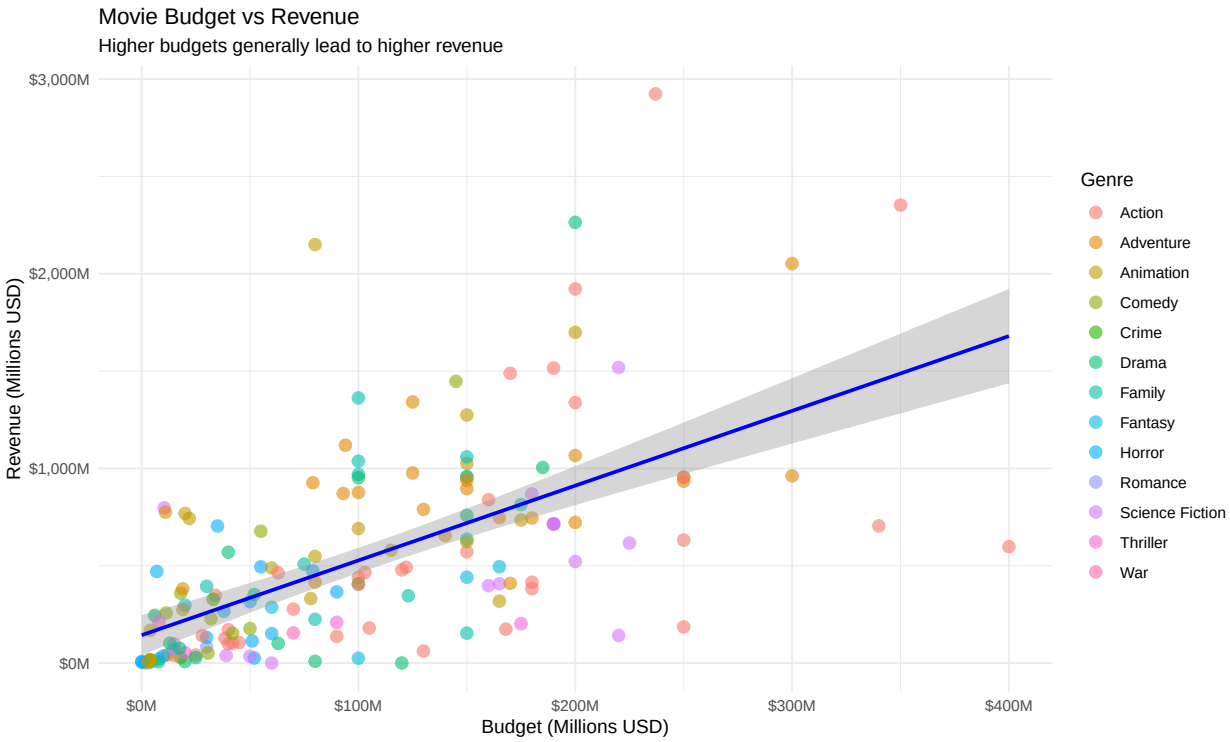
```
##
## === Average Budget & Revenue by Genre ===
```

```
##      primary_genre count avg_budget avg_revenue avg_rating
## 1           Action    35  142442857   616493090       7.30
## 2        Adventure    23  144478261   852044480       7.69
## 3        Animation    19   90778947   704984870       7.71
## 4           Horror    18   32823889   216041217       6.91
## 5  Science Fiction    16  144031250   518880419       7.46
## 6            Drama    15   71953333   431903417       7.66
## 7           Comedy    12   33330250   284248762       7.32
## 8           Family    11  112090909   711640892       7.25
## 9          Fantasy     5  125000000   376743829       7.86
## 10        Thriller     4   73250000   169864418       7.30
## 11           Crime     3   50333333   254422880       7.89
## 12         Romance     2   22300000    77107599       7.33
## 13             War     1   70000000   154984035       6.07
```

5 Analysis and Visualizations

5.1 Budget vs Revenue Relationship

Our first analysis examines the relationship between production budget and box office revenue. Economic theory suggests that higher budgets (spent on talent, special effects, marketing) should lead to higher revenue, but we test this empirically.



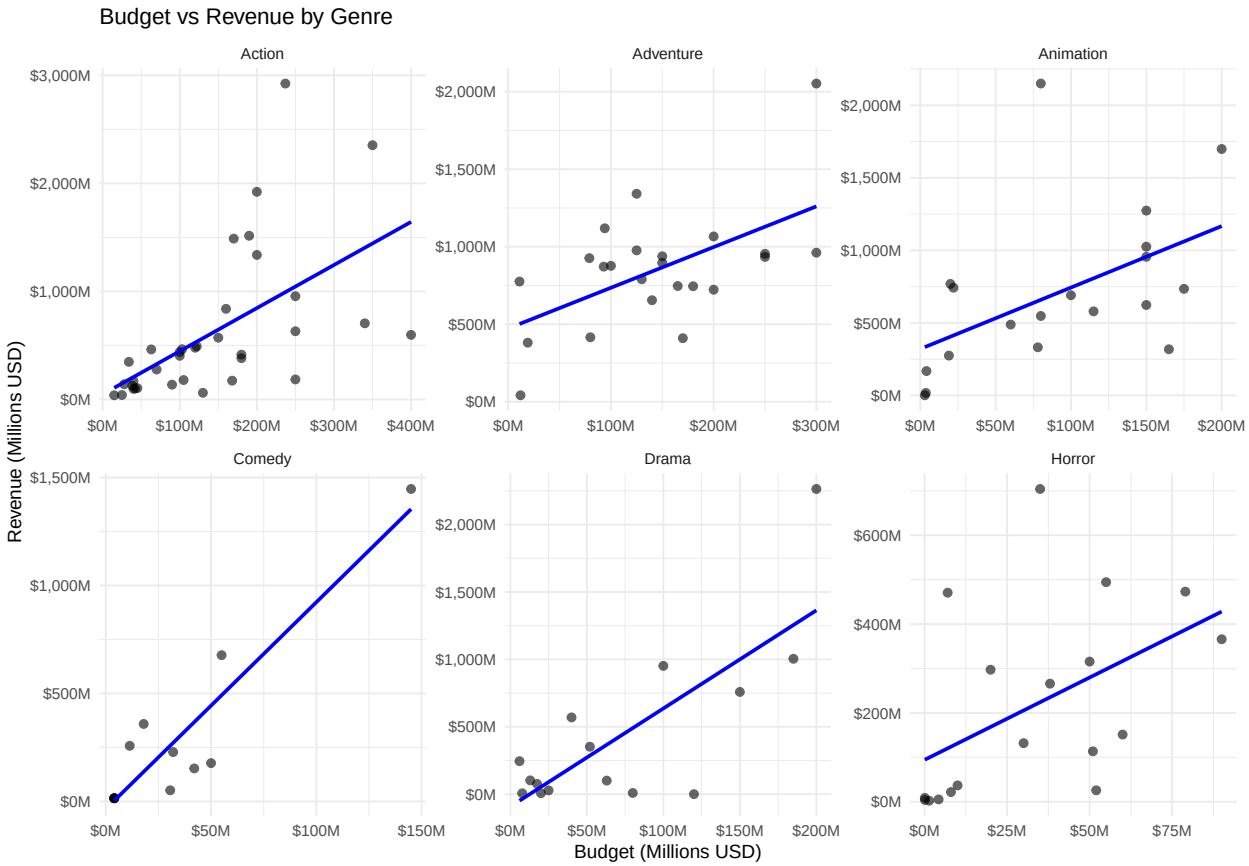

```
##
## Pearson correlation coefficient: 0.604
```

Finding: The correlation coefficient of $\mathbf{r = 0.604}$ indicates a **moderate positive relationship** between budget and revenue. Higher production budgets can be inferred to have a stronger association with larger box office returns. Nevertheless, the revenue fluctuations of high-budget films imply that money invested in production is not the only factor determining sales; narrative quality, casting, marketing, and release timing are probably among the factors contributing to this phenomenon.

The plot shows a significant concentration of low- and mid-budget films, with a small number of outliers having very high revenues, including major franchise releases, highlighting the presence of a small number of exceptionally high-grossing films.

5.1.1 Budget vs Revenue by Major Genres

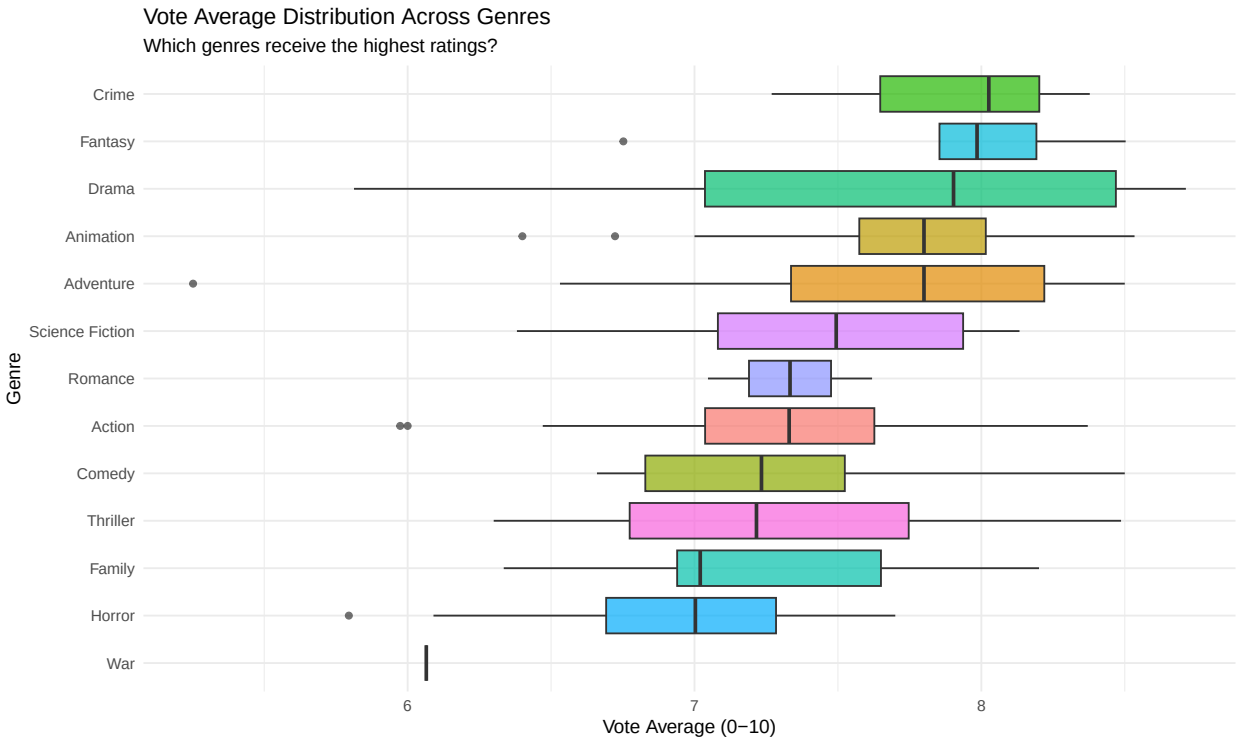
To better understand genre-specific patterns, we examine the budget-revenue relationship separately for major genres.



Across all genres, higher budgets tend to correspond with higher revenues, but the strength of this relationship varies. Action, Adventure, and Animation show stronger positive trends where bigger budgets more reliably lead to higher box office performance. Comedy, Drama, and Horror exhibit weaker correlations with wide variability, suggesting that factors other than budget play a larger role in their revenue outcomes. Horror in particular demonstrates that relatively low budget films can still achieve strong financial success.

5.2 Vote Average by Genre

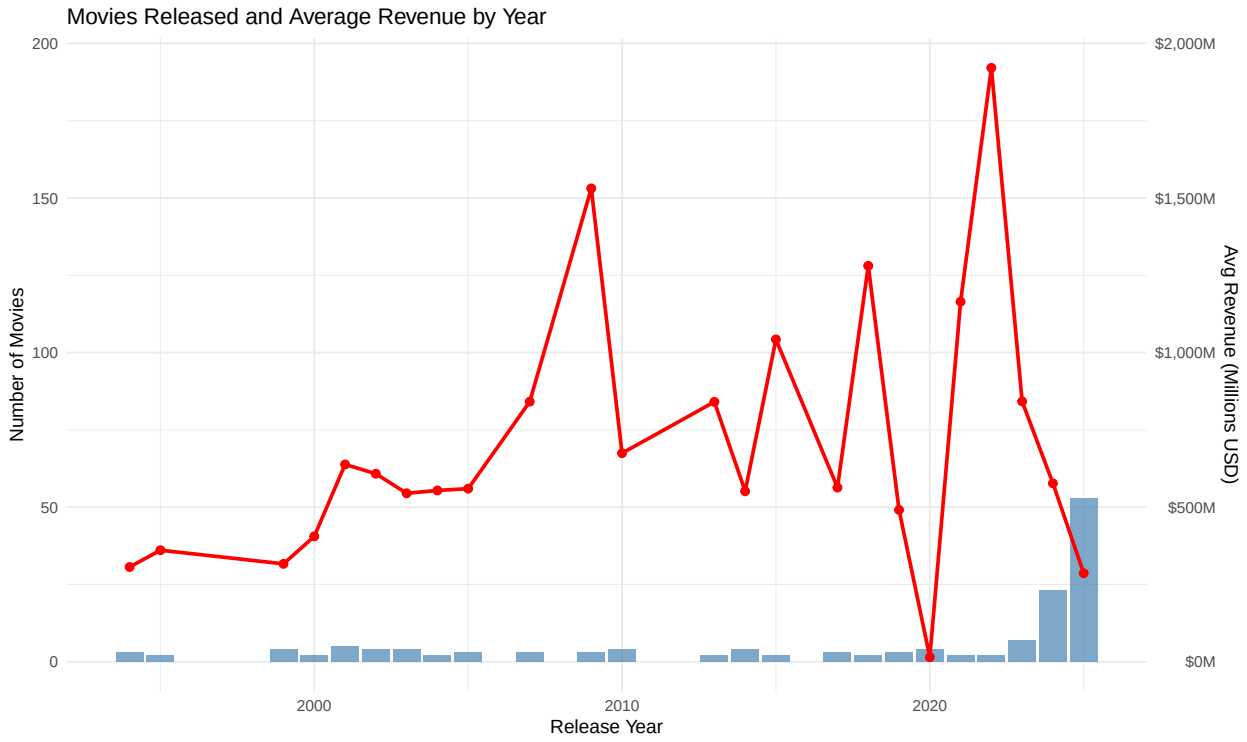
Next, we examine how different genres perform in terms of audience ratings. This helps identify which types of films tend to receive better critical and audience reception.



Finding: Crime and Fantasy genres show the highest median ratings, while Drama and Thriller films display notable variability in audience reception. The results indicate that some genres have the capability to receive consistently positive audience responses. Crime, Fantasy, and Drama films tend to achieve higher ratings overall, while there is more fluctuation in ratings for categories like Action and Horror, indicating that quality or audience expectations differ significantly among films in these categories. This variability suggests that certain genres like Action and Horror may be more "hit-or-miss" in terms of audience satisfaction.

5.3 Temporal Trends

Finally, we examine how the number of movies and their average revenue has changed over time.



Finding: This visualization shows the number of movies (blue bars, left axis) and average revenue (red line, right axis) by release year. Because our dataset is drawn from the "popular movies" endpoint, it is dominated by recent and upcoming releases, but it also includes a smaller number of classic older films that have remained widely viewed. This leads to a strong clustering in the most recent years, particularly 2024 and 2025, while earlier years appear only sparsely. The peak in average revenue, nearing $2,000M, occurs in an earlier year rather than in the most recent period, and is driven by a few exceptionally high-grossing titles. The uneven distribution of films across years limits our ability to infer long-term trends, meaning that the visualization primarily reflects the behavior of films that are currently popular rather than the entire industry over time.

6 Interactive Shiny Application

To enable interactive exploration of the data, we created a Shiny web application that allows users to filter movies by genre and release year, with real-time visualization updates.

The Shiny app provides dynamic filtering and instant visualization updates, allowing users to explore patterns within specific genres or time periods. The application provides an additional tool for exploring the dataset interactively, allowing a detailed examination that is not possible with the static results shown in this paper.

7 Conclusion

7.1 Summary of Findings

The results of this research offer multiple insights into the performance patterns of contemporary popular films. Our analysis of 164 popular films from The Movie Database reveals several key insights:

1. **Budget-Revenue Relationship**: Production budget has a moderate positive correlation ($r = 0.604$) with box office revenue, suggesting that greater financial investment is generally associated with

stronger commercial outcomes. However, high-budget films show vast divergence in revenue, making it clear that budget alone is not a good predictor of financial success; narrative quality, marketing strategies, and external market conditions are probably major factors alongside budget.

2. **Genre Performance**: Differences in audience reception based on genre are quite remarkable. Drama, Fantasy, and Animation films are usually rated higher, while there is more fluctuation in ratings for categories like Action and Horror, indicating that quality or audience expectations differ significantly among films in these categories. Action dominates in terms of quantity, while Horror films show high variability in both ratings and revenue.

3. **Temporal Patterns**: The dataset points out that the concentration of popular films in the past few years was very strong, which influenced the annual release count and average revenue distribution. Our dataset reflects the current popularity landscape, with most films being recent or upcoming releases from 2024-2025.

7.2 Limitations

Despite these insights, several limitations should be acknowledged:

- **Selection Bias**: The dataset only comes from popular films in TMDb, which may result in selection bias that makes the sample less representative of the whole film industry, excluding independent or niche productions that might be financially or aesthetically different.
- **Missing Data**: The sample is narrowed down as 40 films were excluded due to missing financial data, potentially biasing results towards major studio releases.
- **Genre Simplification**: Using only primary genre classifications simplifies the multi-genre nature of many films, which may hide the effects of cross-genre influences.
- **Temporal Imbalance**: The strong focus on recent releases (2024-2025) limits our ability to draw conclusions about long-term trends in the film industry.

Despite these limitations, this analysis demonstrates how web APIs can provide valuable data for understanding the film industry's commercial and artistic landscape.