Yuchen Wu, email: ywu55711@usc.edu, USC ID: 9788517696, GitHub: ywu55711
Yang Zhao, email: yzhao657@usc.edu, USC ID: 7748154612, GitHub: yzhao657-lyz
Github Repository Link: https://github.com/yzhao657-lyz/DSCI-510-Final-Project.git

Analysis of Key Factors for Movie Success Using TMDb Data

Project Description

This project utilizes data from the Movie Database (TMDb) API to analyze factors influencing the commercial success and audience response of films. We examined the relationships between production budget, box office revenue, popularity, user ratings, genre, release date, and creative personnel such as actors and directors. Our overall hypothesis is that films with higher investment, greater exposure, favorable genre characteristics, and higher creative personnel involvement tend to achieve better box office results and audience response. Through data cleaning, visualization, and exploratory analysis, we identified patterns that distinguish box office successes from failures. Furthermore, we constructed a simple predictive model to predict film box office revenue based on key features.

Data Collection

The dataset for this project was collected using the TMDb (The Movie Database) API, which provides structured and publicly available information about movies, including release details, financial performance, genres, popularity metrics, and production credits. These categories were later expanded during data cleaning into specific features such as release year and season, ROI, and summarized genre and cast attributes.

All data was retrieved through HTTP GET requests using the Python requests library, and responses were returned in structured JSON format. Relevant information such as release details, financial metrics, genres, and credits was extracted directly from these JSON responses rather than scraped from web page content.

The process began by retrieving the official list of movie genres using the /genre/movie/list endpoint. This genre reference was saved as a raw JSON file and later used to interpret genre information associated with each movie.

Movies were then collected through the /discover/movie endpoint. The query was configured to include non-adult movies released from 1990 onward, sorted by popularity. Up to 50 pages of results were retrieved, yielding 1,000 unique movie IDs. Each page of results was saved as a separate JSON file in the raw data folder to preserve the original API responses and allow traceability.

For each movie ID, two additional API calls were made: one to retrieve detailed movie metadata (/movie/{id}) and another to retrieve credit information (/movie/{id}/credits). These responses were stored in line-delimited JSON files, where each line corresponds to a single movie. This format allows the data collection process to resume safely if interrupted. To respect API rate limits and improve stability, short delays were added between requests.

Finally, the movie details file was loaded into a Pandas DataFrame, producing an initial structured dataset that retained nested fields such as genre lists and cast information for further processing.

Data Cleaning

After data collection, the raw movie data retrieved from the TMDb API was cleaned and preprocessed to create a structured dataset. These responses were stored in JSON Lines files, where each line corresponds to a single movie. These files were first loaded into Pandas DataFrames, with each row representing a unique movie identified by its TMDb movie ID. Movie IDs served as unique identifiers throughout the process, ensuring that each film appeared only once in the dataset and preventing duplicate records.

Missing and invalid values were handled carefully. Numeric fields such as budget and revenue were converted to numeric types using coercion to handle invalid entries. Movies with

missing or non-positive budget values were treated as invalid for return-on-investment (ROI) analysis and were removed from the final dataset using conditional filtering. Other non-critical fields, such as genre or cast information, were allowed to remain missing where appropriate to avoid unnecessary data loss.

Several preprocessing steps were applied to transform nested and complex data structures into a structured tabular format. Genre information, originally stored as lists of dictionaries, was converted into a pipe-separated string of genre names, along with an extracted primary genre for simplified analysis. Credit data was processed to extract the director's name and the top three billed cast members, which were summarized into concise string features. Release dates were converted to datetime format and further decomposed into release year, release month, release season, and movie age to support temporal analysis. Financial metrics were also enhanced by computing ROI as the ratio of revenue to budget.

Finally, a curated subset of relevant variables was selected, including identifiers, release timing features, financial metrics, popularity and voting statistics, genre summaries, and key credit attributes. The cleaned dataset was saved as a CSV file in the processed data directory, resulting in a final dataset of 499 movies with 18 well-defined features after filtering out records with missing or invalid financial information.

Data Analysis

Hypothesis

First, this project initially proposes four hypotheses related to movie financial performance and audience reception.

Hypothesis 1 (Budget vs. Revenue and ROI): Movies with higher production budgets tend to generate higher box office revenue.

Hypothesis 2 (Popularity, Ratings, and Success): Audience attention indicators, such as popularity and vote count, are more strongly associated with box office revenue.

Hypothesis 3 (Genre-Level and Talent-Level Performance): Movie genres differ in their typical budget levels, box office revenue, audience ratings, and ROI. In addition, movies associated with well-known directors or leading actors tend to exhibit stronger financial performance.

Hypothesis 4 (Time Effects): Movies released during high-attention periods (such as summer, holidays, or award season) tend to achieve higher box office revenue. In addition, as films age, popularity tends to decline over time, while average user ratings remain stable or increase slightly.

Findings

Descriptive statistics and correlation analysis were performed on key variables including budget, revenue, ROI, popularity, average user rating, vote count, and movie age. The correlation heatmap shows a strong positive relationship between budget and revenue, indicating that higher production budgets are generally associated with higher box office returns (Figure 1). Vote count also exhibits a relatively strong correlation with revenue, suggesting that audience engagement and visibility contribute meaningfully to commercial performance (Figure 1). In contrast, popularity and average rating show weaker direct associations with revenue.

To test the budget–revenue relationship more formally (Hypothesis 1), a log–log regression visualization was constructed. The scatter plot reveals a clear positive trend between budget and revenue, but with substantial dispersion across films (Figure 2). This supports the hypothesis that higher budgets increase revenue potential, while also highlighting diminishing returns and variability in outcomes.

To examine whether audience attention or perceived quality better predicts financial success (Hypothesis 2), popularity, vote count, and average rating were compared. The popularity–rating plot shows only a weak relationship between popularity and user ratings (Figure 4), reinforcing the idea that visibility and engagement do not necessarily correspond to perceived quality. Combined with the

correlation results, this suggests that exposure and audience reach play a larger role in revenue generation than ratings alone.

Genre-based analysis (Hypothesis 3) reveals clear differences in financial performance. Comparison of average ROI across genres shows that some lower-budget genres achieve higher returns despite lower absolute revenue, reflecting alternative definitions of success beyond box office scale (Figure 5). These genre-level differences are further contextualized by the overall skewed distribution of ROI, where a small number of films account for disproportionately high returns (Figure 3).

Time-related effects (Hypothesis 4) were examined through both release timing and movie age. Seasonal analysis indicates that films released in certain periods, particularly spring, achieve higher average revenues, likely reflecting strategic release planning rather than direct seasonal causality (Figure 6). In addition, grouped box plots show a slight upward shift in ratings for older films, suggesting a survivorship effect in which higher-quality movies remain visible and continue to receive ratings over time (Figure 7).

Finally, talent-level analyses explored whether directors and actors are associated with financial performance. Ranking directors by average revenue highlights several consistently high-grossing individuals, although variation within this group remains substantial (Figure 8). Actor-level analysis further shows that high average ROI may be driven either by limited participation or sustained performance across multiple films, underscoring the complexity of attributing financial success to individual talent alone (Figure 9).

To incorporate a predictive perspective, we developed linear regression models to forecast box office revenue based on factors like budget, audience metrics, and release timing. For instance, a log–log model using budget to predict log-transformed revenue achieved an R^2 of approximately 0.37, indicating a moderate fit and suggesting that while higher budgets are generally associated with higher revenue, the relationship is not perfectly linear.

We extended this to a multivariate model incorporating variables such as popularity, vote average, vote count, movie age, and release month. This multivariate approach resulted in a similar level of predictive performance ($R^2 \approx 0.35$), reinforcing the idea that no single factor overwhelmingly dominates the prediction. In other words, while budget and audience engagement metrics do provide some predictive signal, the overall forecasting accuracy remains moderate due to the inherent complexity of box office outcomes.

Ultimately, these findings underscore that predicting movie revenue is challenging and influenced by a mix of factors, with no single variable serving as a definitive predictor. This aligns with the broader conclusion that movie success is multifaceted and inherently uncertain.

Visualization

To communicate the findings clearly, each figure is designed to show either relationships between key variables, distributional characteristics, or group differences.

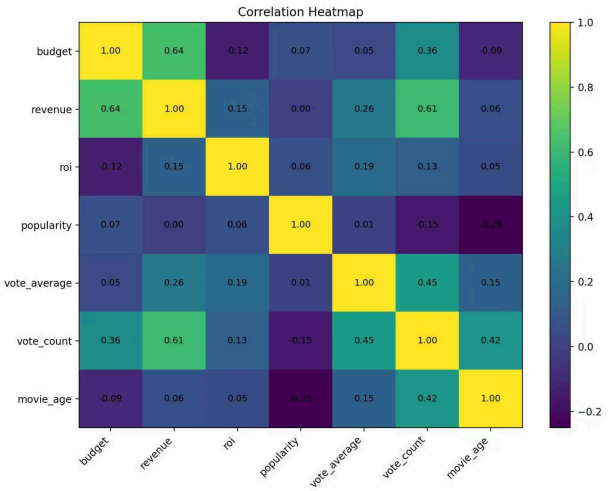
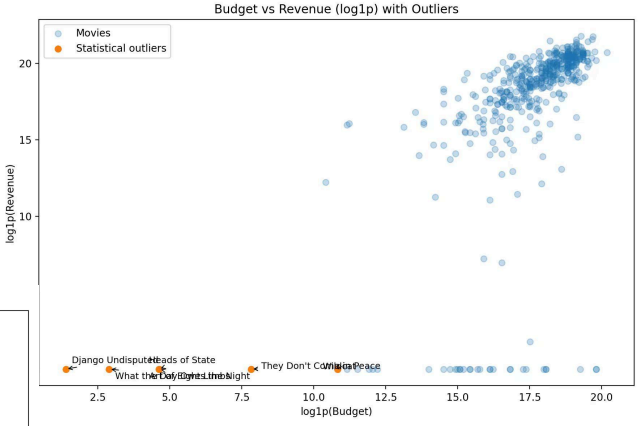


Figure 1 (Correlation Heatmap) presents pairwise correlations among budget, revenue, ROI, popularity, vote average, vote count, and movie age. Color intensity indicates correlation strength, while numeric labels show exact values. This overview identifies which variables are most strongly related and motivates deeper analysis.

Figure 2 (Budget vs. Revenue, log scale) shows the relationship between log-transformed budget and revenue, with each point representing a movie. Outliers are highlighted and labeled to illustrate atypical cases. The plot demonstrates a clear positive trend with substantial variability, supporting the budget–revenue hypothesis.



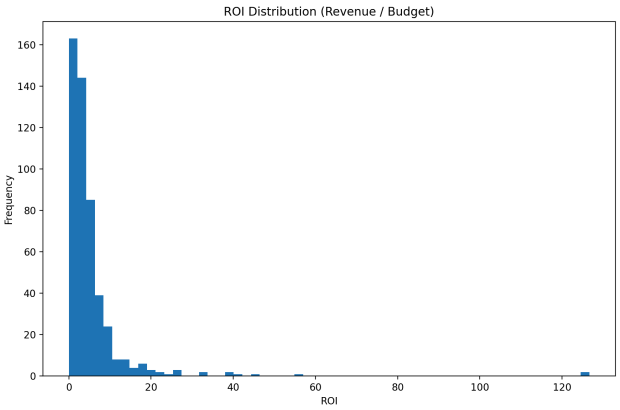


Figure 3 (ROI Distribution) displays the frequency distribution of ROI values. The histogram reveals a heavily right-skewed pattern, indicating that most movies achieve modest returns while a small number generate extremely high ROI.

Figure 4 (Popularity vs. Rating with Trend Line) plots popularity against average rating, with a fitted linear trend line. The relatively flat slope shows that higher popularity does not strongly correspond to higher ratings, distinguishing audience attention from perceived quality.

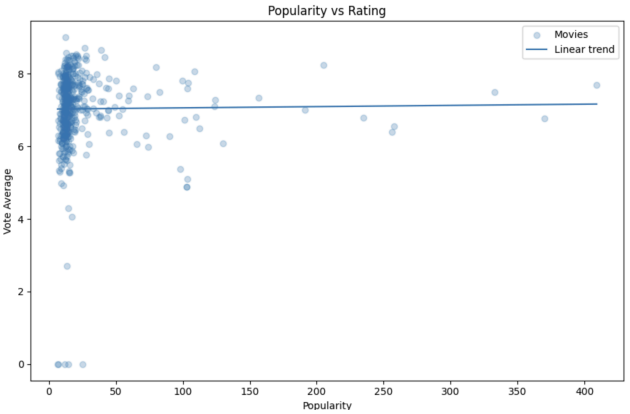
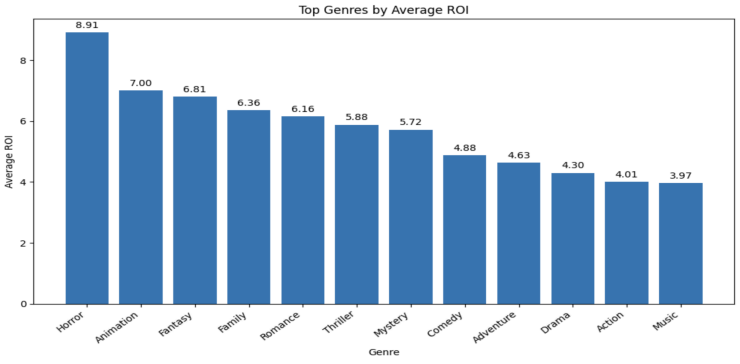


Figure 5 (Top Genres by Average ROI) compares genres based on mean ROI using a bar chart with labeled values. The figure highlights that some lower-budget genres achieve higher returns, emphasizing different definitions of success across genres.



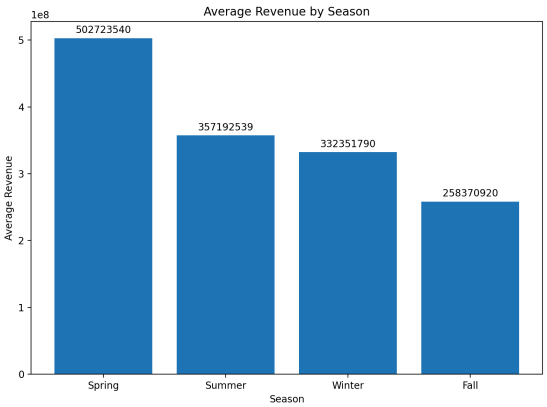
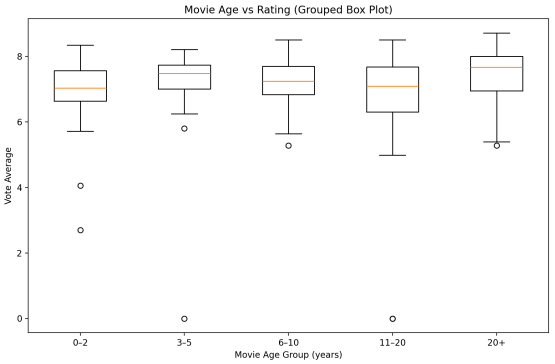


Figure 6 (Average Revenue by Season) compares average revenue across seasons using a bar chart. Value labels clarify that certain seasons, especially spring, are associated with higher average revenue.

Figure 7 (Movie Age vs. Rating) uses grouped box plots to show rating distributions across movie age categories. Medians, spreads, and outliers reveal a slight upward shift in ratings for older movies.



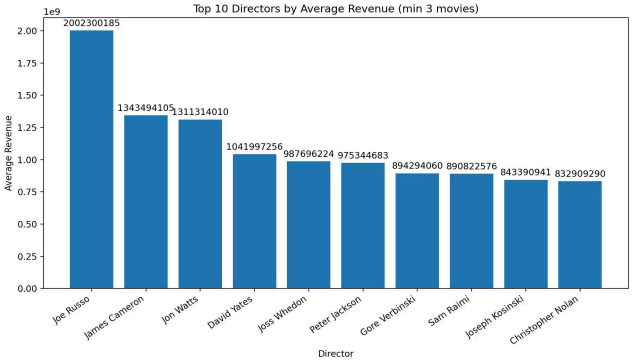
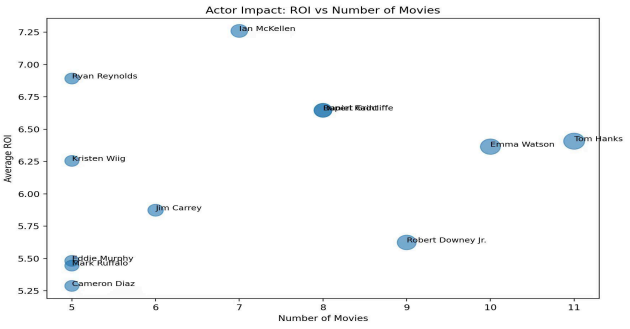


Figure 8 (Top Directors by Average Revenue) ranks directors by mean revenue, applying a minimum movie threshold to reduce noise. This figure examines whether consistently high-grossing directors emerge in the data.

Figure 9 (Actor ROI vs. Number of Movies) plots average ROI against the number of movies per actor, with point size reflecting sample size. The visualization helps interpret whether high ROI is driven by limited or sustained participation.



Challenges, Changes and Future Word

The overall goals of this project remain consistent with the initial proposal, which aimed to examine how factors such as budget, box office revenue, popularity, ratings, genre, and release timing influence film success using data from the TMDb API.

During the early stage of the project, the analytical plan was relatively simple and focused primarily on linear relationships, particularly between production budget and box office revenue. However, preliminary results revealed substantial variation in revenue among films with similar budgets, indicating that budget alone could not sufficiently explain financial outcomes. This observation highlighted the need to incorporate additional factors and motivated several refinements to the original plan.

As a result, the scope of the analysis was expanded to include return on investment (ROI), genre-level differences, release season effects, and talent-related factors such as directors and leading actors. ROI was introduced to better compare films with different budget scales, while release dates were simplified into four seasonal categories to improve interpretability and maintain sufficient sample sizes. In addition, films with missing or zero budget values were removed to ensure the validity of financial analyses. Regression-based models were also incorporated to examine the relative importance of multiple variables rather than relying solely on simple correlations.

With additional time or resources, future work could explore more advanced predictive models, incorporate richer measures of director and actor popularity, adjust financial values for inflation, and analyze audience sentiment using textual review data. These extensions could further enhance both the explanatory depth and predictive potential of the project.