Ella Zaputil

Professor Redmond

BAIS:3250 Data Wrangling

9 May 2025

Behind the Screen: Exploring Movie Data for Deeper Insights

**Introduction**

As one of the most profitable sectors in the entertainment industry, the film industry generates billions in annual revenue, making it vital to understand the key factors that influence a movie’s financial success. With the immense financial stakes involved, understanding the factors that drive a film’s success is essential for studios, investors, and filmmakers alike. From the size of the production budget to the star power of the cast and the timing of a release, numerous variables can influence a film's revenue potential. As the industry continues to evolve with the rise of streaming platforms and changing audience preferences, it becomes increasingly important to identify patterns that can predict a movie’s financial outcome. This project aims to explore the key drivers of movie revenue, providing valuable insights into how various aspects of a film can contribute to its financial success and offering predictions for future industry trends.

**Data**

For this project, my first data source is the “IMDB Movies Dataset” from Kaggle, which contains detailed information about the top 1000 movies on IMDB. The dataset includes fields such as movie title, release date, runtime, genre, IMDB rating, director, top 4 stars, and gross revenue. However, since the fields "certificate," "overview," "Metascore," and "IMDB votes" are not relevant to the analysis, I removed them from the final dataset. The dataset provides a total of 1,000 data points. This data contributed 1,000 data points. The original dataset can be viewed [here](https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows).

Additionally, I scraped additional data from “The Numbers” website, which provides financial information about films, including film title, ranking, production budget, domestic gross revenue, and worldwide gross revenue. This source contributed 3,800 data points. The website I scraped data from can be accessed [here](https://www.the-numbers.com/movie/budgets/all).

After downloading and scraping data from each source, I merged the datasets using a horizontal integration based on two common fields: Movie Title and Release Year. I used an inner join to ensure that only movies present in both datasets were retained, which helped maintain consistency and data quality for the analysis. This merging process resulted in a final dataset containing 314 movies. The raw merged dataset can be accessed [here](https://iowa-my.sharepoint.com/personal/edzaputil_uiowa_edu/Documents/Desktop/Spring%20Semester%202025/Data%20Wrangling/edzaputil_IntegratedMovieData.csv).

*Figure 1: Data Dictionary*

|  |  |  |
| --- | --- | --- |
| **Feature Name** | **Data Type** | **Description** |
| Movie Title | Object | Title of film |
| Release Year | Integer | Year film was released |
| Runtime | Integer | Length of film in minutes |
| IMDB\_Rating | Float | Rating of film on IMDB website |
| Director | Object | Name of director of film |
| Star1 – Star4 | Object | Names of top 4 actors in film |
| Genres | Integers | Columns for each possible genre, 0 or 1 value for each film |
| Ranking | Float | Ranking of film amongst other films |
| Release Date | DateTime | Date film was released (YYYY-MM-DD) |
| Production Budget | Integer | Money allocated to produce film |
| Domestic Gross | Integer | Money made by film in country of filming |
| Worldwide Gross | Integer | Money made by film internationally |
| Release Month | Integer | Extracted month film was released |

**Analysis**

**Research Question 1: How does a film’s production budget impact its gross revenue?**

To investigate whether a higher production budget leads to higher worldwide gross revenue, I conducted a series of steps to quantify the relationship between these variables. First, I cleaned the data to ensure it was able to be analyzed. I dropped missing values from the 'Production Budget' and 'Worldwide Gross' columns to ensure the analysis was based on complete data. This step was necessary to prevent inaccuracies in calculations, such as misleading correlations or regression estimates caused by missing entries. Next, I examined the distribution of both production budgets and worldwide revenue using histograms. This helped visualize the spread of budget sizes and revenue outcomes across different films. The distributions provided insight into how budgets and revenues vary, showing a right-skewed pattern, indicating that while most movies have moderate budgets and earnings, a few outliers significantly exceed the norm.

A graph of a distribution of production budget

AI-generated content may be incorrect. A graph of a distribution of gross revenue

AI-generated content may be incorrect.   
 *Figure 2: Distribution of Production Budget Figure 3: Distribution of Worldwide   
 Gross Revenue*

To assess the strength and direction of the relationship between production budget and revenue, I calculated the correlation coefficient, which was 0.77. This revealed a strong positive correlation, suggesting that movies with higher budgets tend to achieve higher earnings. To visualize this relationship, I created a scatter plot that plotted production budget on the x-axis and worldwide revenue on the y-axis. The plot clearly showed the positive correlation, with higher production budgets generally corresponding to higher worldwide revenue, reinforcing the idea that a larger budget often leads to greater financial success for a movie.

A diagram of a graph

AI-generated content may be incorrect.  
*Figure 4: Scatter plot of Production Budget vs. Worldwide Gross Revenue*

Next, I built a linear regression model to quantify this relationship further. The regression equation derived from the analysis was: Worldwide Gross Revenue = 12,176,232.57 + 4.71 × Production Budget. This equation suggests that for every $1 increase in production budget, we expect an increase of $4.71 in worldwide revenue. To assess how well the model explains variations in revenue, I examined the R-squared value, which was 0.59. This indicates a moderate fit, meaning production budget explains 59% of the variance in revenue, but other factors also influence a movie’s financial success. My analysis confirms that higher production budgets generally lead to higher revenues, as shown by the strong correlation and the positive regression coefficient. However, the R-squared value suggests that while budget is a significant factor, additional elements, such as marketing, distribution, genre, and audience reception, play crucial roles in determining a film’s total earnings. These findings reinforce the importance of strategic budgeting, demonstrating that investing more in production can substantially enhance revenue potential, but producers should also consider other influencing factors beyond budget alone.

**Research Question 2: Which variables most positively and negatively influence revenue?**

After exploring the relationship between production budget and worldwide revenue, I wanted to dive deeper and identify which variables most positively and negatively influence a movie's financial performance. To achieve this, I implemented a series of regression models: Linear Regression, Ridge Regression, Lasso Regression, and Random Forest Regressor. These models focused on identifying key predictors and understanding their relationship with revenue.

I began with data preparation, removing non-numeric and irrelevant columns such as movie title, director, cast, release date, and domestic gross revenue, as these variables were not directly relevant to the regression models I wanted to build. I also removed all rows with missing values to avoid bias or errors in model performance. Next, I identified the independent variables and the target variable. The target variable I aimed to predict was the Worldwide Gross revenue. The features included various numeric movie attributes, such as genres, runtime, IMDB rating, and production budget, excluding categorical columns like movie title and cast. To ensure that my models could generalize well to unseen data, I split the data into training and testing sets using an 80/20 split. The training set would be used to train the models, while the testing set would be used to evaluate their performance. This allowed me to assess the accuracy of the models.

I first ran a linear regression model to evaluate the simplest possible relationship between the features and worldwide revenue. This yielded a very high mean squared error (~1.19e+17) and a negative R² score (-0.56), indicating poor fit and suggesting that a strictly linear approach does not capture the complexity of the data. Certain variables, such as Animation, Fantasy, and Thriller genres, had high positive coefficients, while others, like Biography and Mystery, had negative coefficients, indicating their inverse relationship with revenue. However, this model’s inability to fit the data properly indicated that linear relationships were insufficient.

Next, I used Ridge Regression, which is a type of linear regression that adds regularization to prevent overfitting and manage multicollinearity. Ridge, with a similarly poor R² score (-0.46), helped address multicollinearity and found that variables like Production Budget, Animation, and Runtime had the most influence. While the results were still not optimal, the Ridge model helped highlight the importance of variables such as Production Budget and Animation in predicting revenue. It also provides a more stable set of coefficients compared to the Linear Regression model. I then implemented Lasso Regression, another form of regularized linear regression, which has the added benefit of performing feature selection by shrinking some coefficients to zero. This performed slightly better in terms of feature selection, shrinking some coefficients to zero and suggesting that Music, Sport, and Horror were negligible predictors. Despite these improvements, the results were still not helpful compared to the Random Forest.

Finally, I turned to a Random Forest Regressor, a more advanced machine learning algorithm capable of capturing non-linear relationships and interactions between features. This performed the best, and had a significantly lower mean squared error (~4.07e+16) and a positive R² score of 0.46. This result implies that the relationships between the predictors and worldwide revenue are more complex than linear, involving non-linear interactions. The feature importance analysis revealed that key variables like Production Budget, IMDB Rating, and Runtime were the most influential predictors of success. Through these steps, I was able to build a series of models, identify the most important predictors, and assess the effectiveness of different modeling techniques. The overall takeaway was that while simpler models like Linear Regression provided some insight, more sophisticated models like Random Forests were far more effective at capturing the complexity of the relationships between movie attributes and financial performance.

A graph with blue bars

AI-generated content may be incorrect.  
*Figure 5: Feature Importance graph of variables vs. Worldwide Gross Revenue*

**Research Question 3: Do certain actors consistently produce higher-grossing films? How much does star power affect a movie’s success?**

Actors that consistently perform in high-profile roles are more likely to generate significantly higher worldwide gross revenue on average. To investigate whether certain actors consistently produce higher grossing films and assess the impact of ‘star power’ on a movie’s success, I first transformed the dataset to enable analysis across all featured stars. Since each film in the dataset listed up to four main actors, I used the melt function to consolidate the Star1 through Star4 columns into a single actor column, paired with each film’s worldwide gross revenue. This allowed me to treat every actor appearance equally for analysis. After dropping any rows with missing actor data, I grouped the data by actor and calculated the mean, median, and count of worldwide gross revenue for each. To ensure relevance, I filtered the data set to only include those actors who have appeared in at least three films.

From this data set I plotted the top 10 actors by average worldwide gross revenue. The results highlight that actors like Joe Russo, Zoe Saldana, Chris Evans, and Robert Downey Jr. consistently appear in high-grossing films, with averages well over $1 billion. Many of these actors are associated with major franchises, such as the Marvel Cinematic Universe, Avatar, and Lord of the Rings, suggesting that franchise affiliation and casts play a significant role in box office success. This supports the idea that star power, particularly when combined with blockbuster franchises, is a strong predictor of a film’s financial performance.

A graph of actors with names

AI-generated content may be incorrect.

*Figure 6: Bar chart of top 10 actors by average Worldwide Gross Revenue*

**Research Question 4: Does the timing of a movie’s release affect its gross revenue?**

I hypothesized that movies released during peak viewing periods, such as the summer months, would generate higher gross revenue compared to those released at other times of the year.To explore whether the timing of a movie’s release affects its gross revenue, I first extracted the release month from each film’s release date by using the pd.to\_datetime function. I then visualized the distribution of movie releases by month with a count plot, which revealed that December has the highest number of movie releases, followed by November and October.

A graph of a movie release

AI-generated content may be incorrect.  
*Figure 7: histogram of movies by release month*

To assess the relationship between release timing and financial performance, I calculated the average worldwide gross revenue for each month and visualized the results using a bar plot. Interestingly, April and July had the highest average grosses, even though they are not among the most common release months.

A graph of blue bars

AI-generated content may be incorrect.  
*Figure 8: bar chart of average Worldwide Gross Revenue by Release Month*

To assess whether summer blockbusters perform significantly better, I conducted an independent t-test comparing movies released during the summer months (May-July) with those released during the rest of the year. The results, a t-statistic of 3.30 and p-value of 0.0012, indicate a statistically significant difference, suggesting that summer releases do tend to generate more revenue. These findings highlight a potential mismatch between release volume and profitability, raising a question about why studios may not more heavily target the more lucrative months like April and July. This trend may be attributed to increased audience availability during school vacations, larger marketing campaigns for summer blockbusters, and the historical success of major franchise films released during this period.

**Research Question 5: Can you predict revenue over the years based on historical data?**

To explore whether historical data can be used to predict future movie revenue, I hypothesized that trends in worldwide gross revenue over time would allow for reasonably accurate forecasting using time series modeling. I began by converting movie release dates to a datetime format and extracting the release year. I then aggregated the worldwide gross revenue by year, creating a time series from the 1950s to 2020. To visualize historical worldwide gross revenue trends, I created a line chart with years on the x-axis and revenue on the y-axis. This visualization clearly showed long-term growth patterns in the film industry, along with major disruptions, such as the sharp decline in 2020, making it easier to interpret and compare forecasted trends.

A graph showing a number of movie revenue

AI-generated content may be incorrect.  
*Figure 9: movie revenue over time by year*

Next, to evaluate the impact of unusual market disruptions, specifically the COVID-19 pandemic, I constructed two separate forecasts using Holt-Winters Exponential Smoothing, one including data from 2020 and one excluding it. The model that included 2020 predicted slower revenue growth, while the model without 2020 showed a much steeper upward trend, suggesting that the sharp decline in 2020 skewed the forecast. This indicates that global events can significantly disrupt predictable trends in revenue. Overall, the results support the hypothesis that film should be carefully considered.

A graph showing a line

AI-generated content may be incorrect.  
*Figure 10: projected revenue growth for next 10 years*

**Conclusion**

In this project, I explored the key factors that influence a film’s financial success using a comprehensive dataset of 314 films. My analysis revealed that production budget, actor influence, timing of release, and other attributes such as runtime and rating all contribute to a movie’s worldwide gross revenue. Although the analysis provides meaningful insights, certain limitations should be acknowledged. The dataset primarily contains high-ranking films, which may introduce biases favoring blockbuster successes and overlooking independent or lower-budget productions. Furthermore, categorical variables such as marketing spend and critical reviews were excluded, potentially limiting the predictive capacity of the models. Data merging also posed challenges, as discrepancies in movie title formatting may have led to the omission of some entries during integration. Future research could address these concerns by incorporating additional categorical features, expanding the dataset to include a broader range of films, and leveraging advanced natural language processing techniques to assess qualitative aspects such as audience reception and critic reviews. Ultimately, these insights can support filmmakers, studios, and investors to make more informed decisions about budget allocation, casting, and release strategies.

References

*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>