Predicting Box Office Success: A Data-Driven Approach

JSC370H1: Data Science II, Winter 2025

Janis Joplin | GitHub Repo

Introduction

The success of a movie at the box office is influenced by various factors, ranging from production-related aspects like budget and cast to external elements such as release timing and genre preferences. Understanding these relationships is valuable for both the film industry and researchers studying entertainment economics. While certain patterns, such as high-budget blockbuster films often generating significant revenue may seem intuitive, the extent to which different factors contribute to box office performance remains an open question.

This project explores how multiple factors—including budget, popularity, adult rating, release season, runtime, genres, actors, and crew—relate to a movie's box office revenue. By conducting exploratory data analysis (EDA), we aim to uncover potential correlations and trends within these variables, identifying which characteristics are most associated with financial success.

Research Question: "**How do factors such as budget, popularity, adult rating, release season, runtime, genres, actors, and crew influence a movie's box office revenue?**"

This analysis will serve as a foundation for further statistical modeling, helping to assess whether certain attributes have a measurable impact on revenue. The findings from this exploratory phase will provide insights into potential predictive relationships and guide future research in film analytics.

Methods

In this section, I describe how I collected the movie data from TMDB, cleaned and processed the raw information, and set up my exploratory and statistical analyses. This part is mainly about what I did to prepare and analyze the data, without going into the results themselves (those will come in Preliminary Results and Summary).

Data Collection

The data comes from The Movie Database (TMDB) API. Specifically, I used three different endpoints:

1. /discover/movie – to fetch batches of movie IDs and titles (20 at a time) sorted by descending revenue, ensuring we get high-revenue (and presumably more popular) movies. We iterated through enough pages to reach 1,000 total movies.
2. /movie/{movie_id} – to collect each movie's details such as revenue, budget, popularity, release date, runtime, and genre information.
3. /movie/{movie_id}/credits – to fetch information on up to five cast members (actors) and five crew members (directors, producers, etc.) for each film, which I then used to derive average actor and crew popularity scores as numeric features.

This approach gave me a comprehensive dataset with a variety of movie attributes. All calls were made with valid API keys and carefully iterated until the desired dataset size was reached.

Data Cleaning & Wrangling

Once I had the raw data, I did the following to clean and tidy it:

- Converted release dates to a proper Date format to simplify time-based analysis.
- Checked for missing values in each column. Rows with essential missing fields (e.g., no revenue, no budget, no release date) or obviously invalid values (like zero budget or zero runtime) were removed.
- Adult rating turned out to be `FALSE` for all entries in the top-revenue set, so I excluded that column from further analysis.
- Removed or corrected abnormal values (e.g., extremely short runtimes, if any).
- One-hot encoded genres. Since most movies have multiple genres, I split the genre strings into individual categories and created a binary indicator (0 or 1) for each possible genre (Action, Romance, Sci-Fi, and so forth).
- Created a "season" variable from the release date. Each movie was labeled as Winter, Spring, Summer, or Fall based on its release month. This is useful for assessing revenue patterns by time of year.
- Finally, made one final check to confirm there were no missing values or weird outliers left. For further visualization and analysis, the cleaned dataset will be used.

Exploratory Data Analysis (EDA)

The EDA focused on understanding how each predictor might relate to revenue. Some key points in my EDA workflow:

- Basic numeric summaries (minimum, maximum, quartiles, mean) for revenue, budget, popularity, release date, runtime, average actor popularity, and average crew popularity.
- Log-transformation of revenue (log base 10) to handle the wide range in revenue values. This often makes scatter plots and correlations more interpretable.
- Scatter plots of revenue versus budget, popularity, runtime, actor/crew popularity.
- Boxplots for revenue across different release seasons and genres.
- A time series of average revenue by release year to look for historical trends.
- A correlation matrix of numeric features to see which factors co-occur strongly.

Statistical Analysis

I performed two statistical procedures:

1. Linear regression to explore how revenue (in log scale) relates linearly to numeric predictors (budget, popularity, runtime, actor/crew popularity) and the presence of each genre (via our one-hot variables).
2. ANOVA (Analysis of Variance) to check if there's a significant difference in average (log) revenue across the four release seasons. This helps determine if the time of year has a meaningful effect on box office success.

Preliminary Results

To understand the factors influencing movie revenue, we first conduct an exploratory data analysis (EDA) to examine key patterns in the dataset. This includes visualizing revenue distributions, assessing relationships between revenue and numeric predictors, and analyzing categorical factors such as seasonal trends and genres. Additionally, we perform statistical modeling to quantify the impact of different predictors on revenue. The following section presents key findings from these analyses.

Figure 1: Movie Revenue Distributions

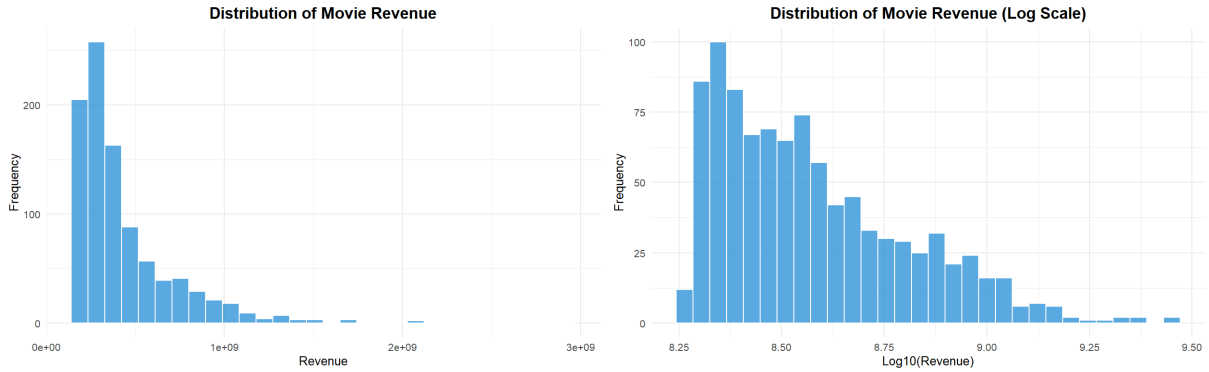

Figure 1: Histograms of movie revenue on both the raw scale (left) and a $\log_{10}$ scale (right). The raw distribution is heavily right-skewed with a few extremely high earners, while the log transformation provides a more balanced view of the data.

Figure 2: Budget vs. Revenue (Log-Log)

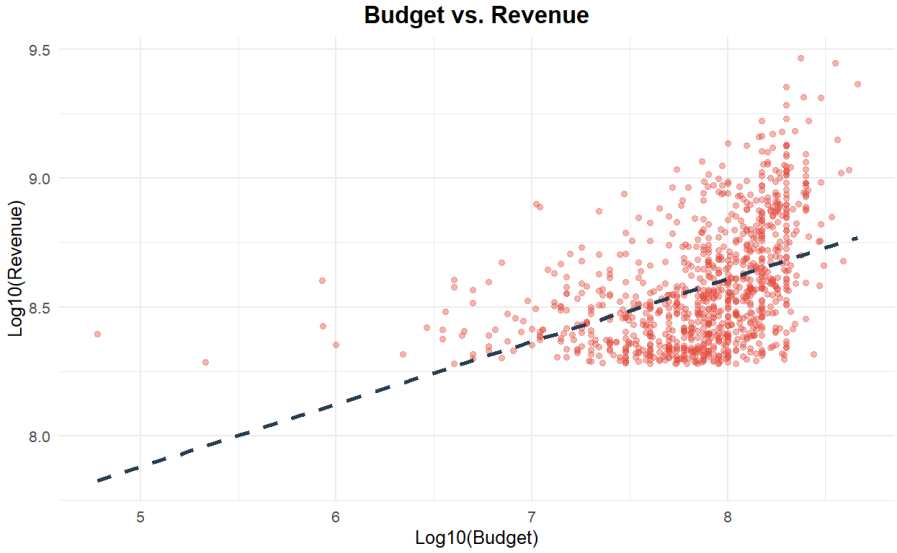

Figure 2: Scatter plot of budget vs. revenue, both on $\log_{10}$ scales. The dashed line represents the fitted linear trend, indicating a generally positive relationship between higher budgets and higher revenues.

Figure 3: Popularity vs. Revenue

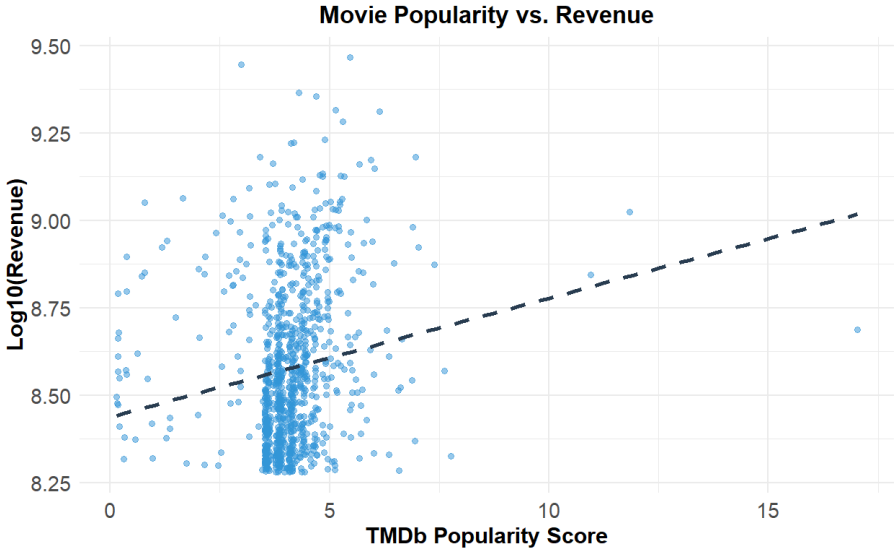

Figure 3: Scatter plot of TMDB popularity vs. $\log_{10}$(revenue). The dashed trend line suggests a modestly positive relationship, though the effect appears weaker than that of budget.

Figure 4: Runtime vs. Revenue

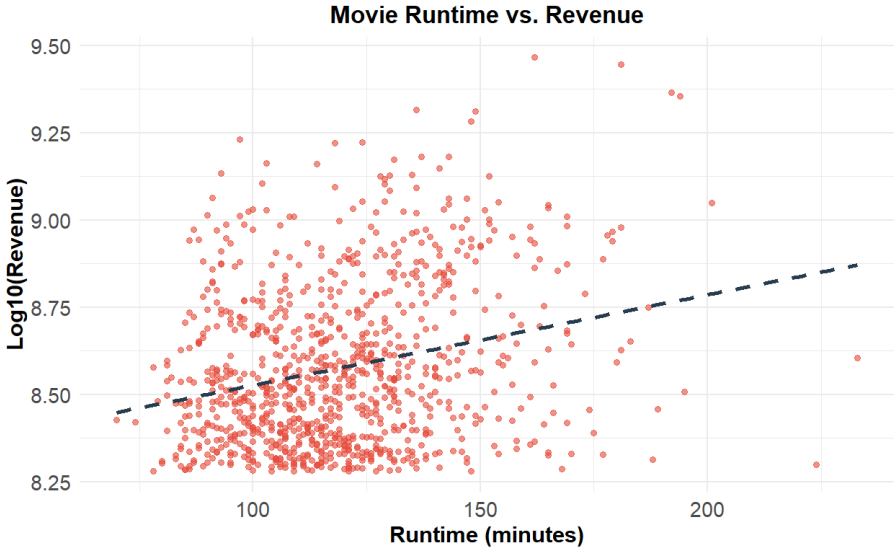

Figure 4: Scatter plot of movie runtime vs. $\log_{10}$(revenue). The dashed regression line indicates a moderate positive trend, suggesting that longer movies may tend to earn slightly higher revenues.

Figure 5: Movie Revenue by Release Season

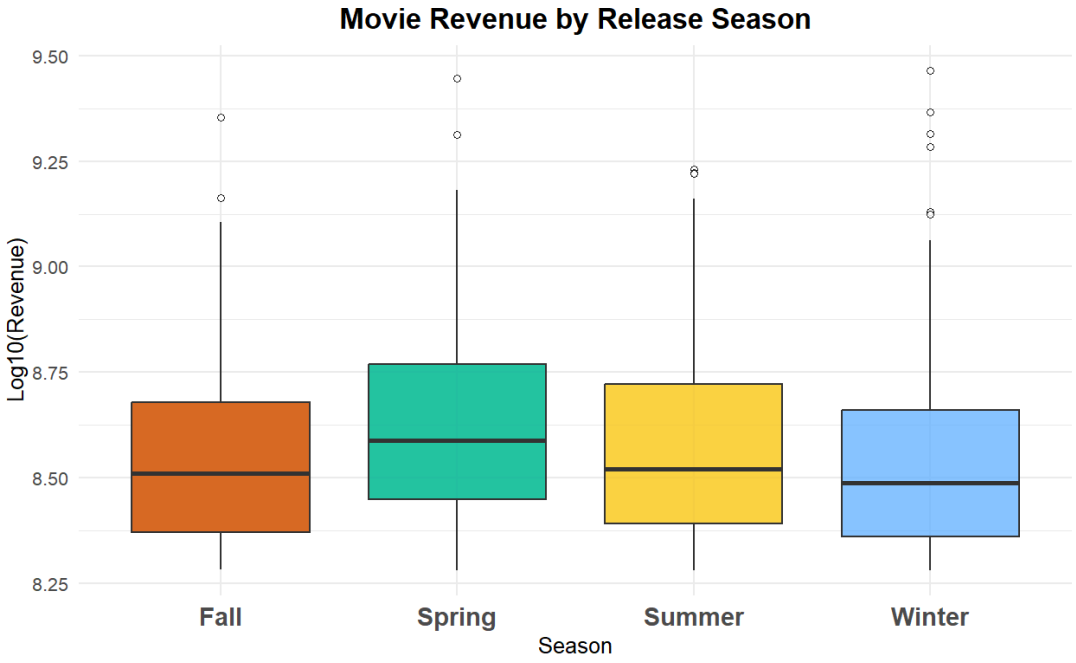

Figure 5: Boxplots of $\log_{10}$(revenue) by season (Fall, Spring, Summer, Winter). The median revenue does not vary drastically between seasons, but Summer and Spring appear to have slightly higher median values compared to Fall and Winter.

Figure 6: Average Movie Revenue Over Time

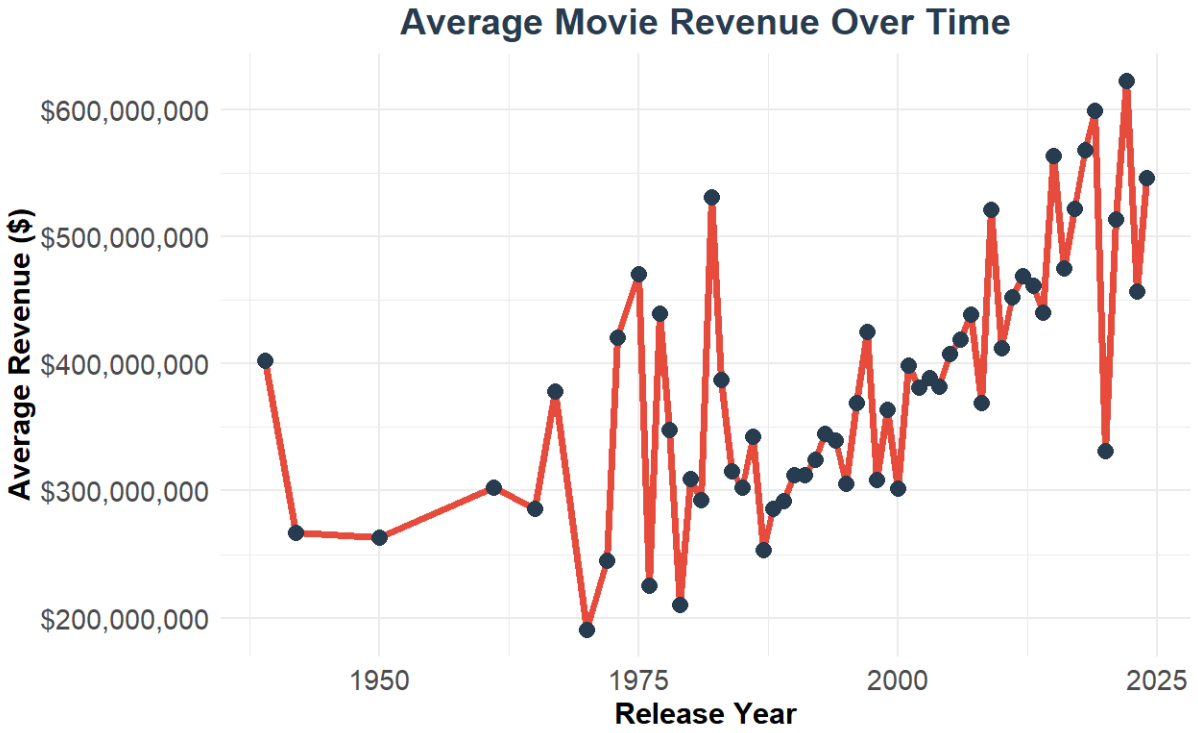

Figure 6: Time series plot of average movie revenue by release year. The general trend shows increasing revenues over time, with more volatility in recent decades. A noticeable dip around 2019-2020 could be attributed to the impact of the COVID-19 pandemic on the film industry.

Figure 7: Revenue Distribution by Genre

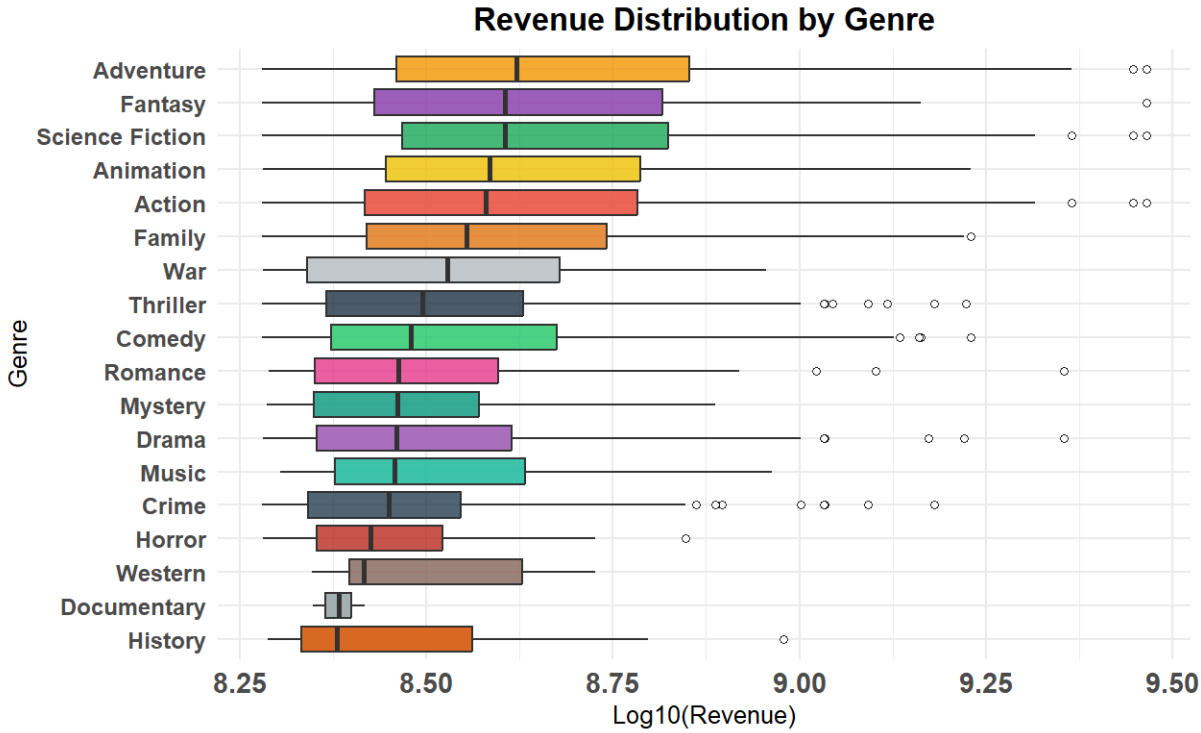

Figure 7: Boxplots of $\log_{10}$(revenue) for different genres. Adventure, Fantasy, Science Fiction, and Animation films tend to have the highest median revenues, while Documentary and History films generally have lower revenue distributions. The presence of extreme outliers suggests that a few blockbuster movies significantly impact the revenue distribution for certain genres. Upon inspection, the notable high-revenue outlier in the History genre is Oppenheimer, which is expected given its massive box office success.

Figure 8: Actor Popularity vs. Revenue

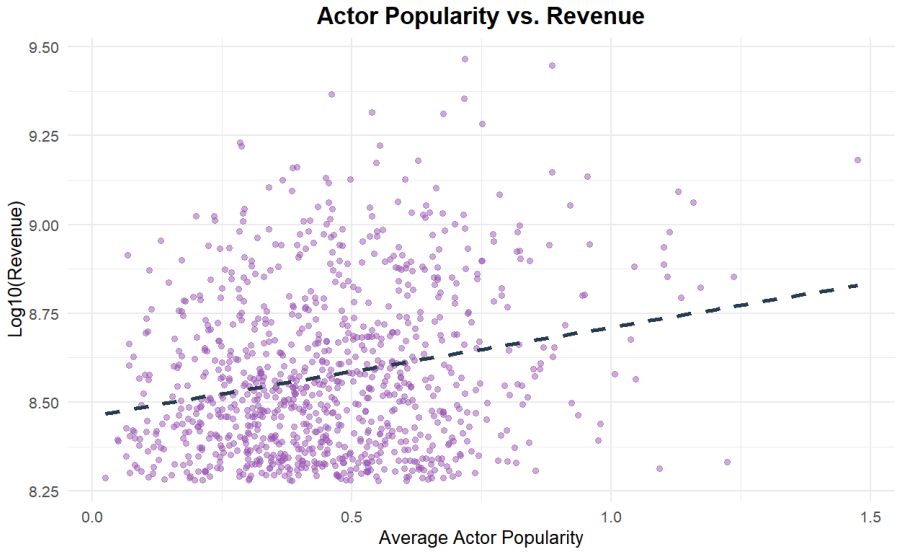

Figure 8: Scatter plot of average actor popularity vs. $\log_{10}$(revenue). There is a moderately strong positive relationship, indicating that movies featuring more popular actors tend to generate higher revenue. This aligns with expectations, as well-known actors often attract larger audiences and drive box office performance.

Figure 9: Crew Popularity vs. Revenue

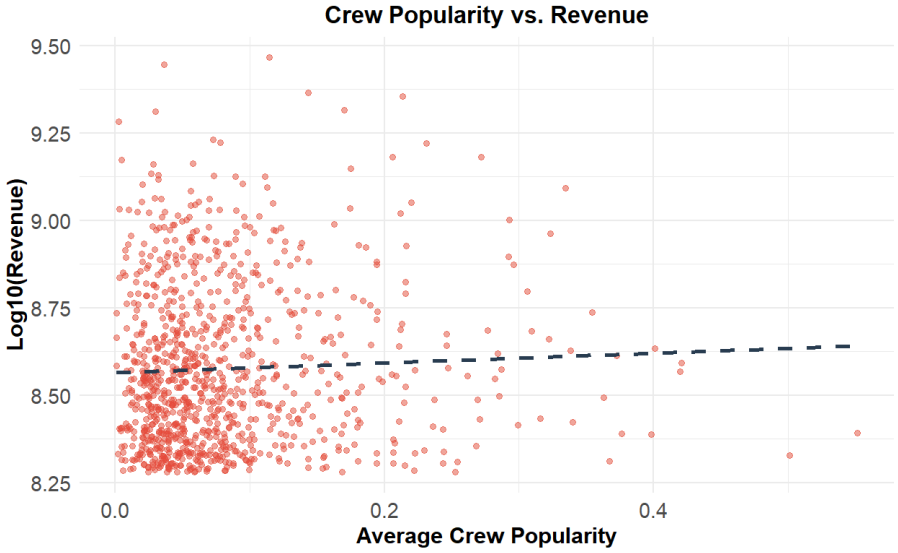

Figure 9: Scatter plot of average crew popularity vs. $\log_{10}$(revenue). Unlike actor popularity, crew popularity shows little to no correlation with revenue, suggesting that a well-known production team does not necessarily translate to higher box office earnings. This also makes sense when we think that most people might only pay attention to actors appearing on screen, rather than those behind it.

Figure 10: Top 20 Actors by Revenue Metrics

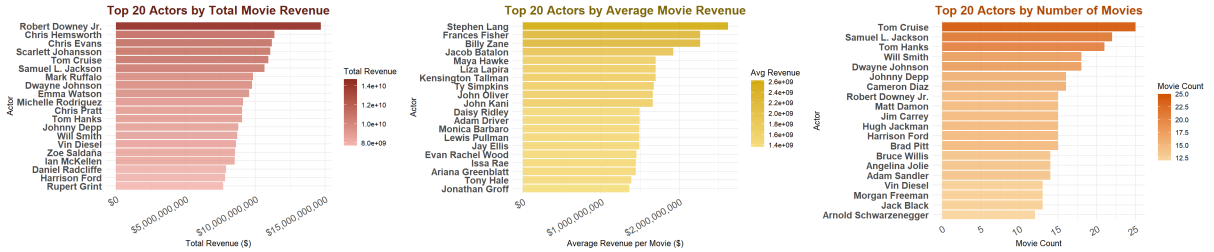

Figure 10: Bar charts showing the top 20 actors ranked by total movie revenue (left), average revenue per movie (center), and number of movies they have appeared in (right). The rankings vary significantly across the three categories, suggesting that actor influence on revenue is complex. Some actors might have high total revenue simply due to appearing in many movies, while others might have high average revenue because they starred in a few blockbuster films but were not necessarily the driving force behind the movie's success. Meanwhile, actors with high movie counts might frequently appear in lower-grossing films. Since each metric captures a different aspect of an actor's career, combining them into a single predictor of revenue is challenging, and none alone seem to provide a clear estimate of how actors influence box office performance.

Figure 11: Top 20 Crew Members by Revenue Metrics

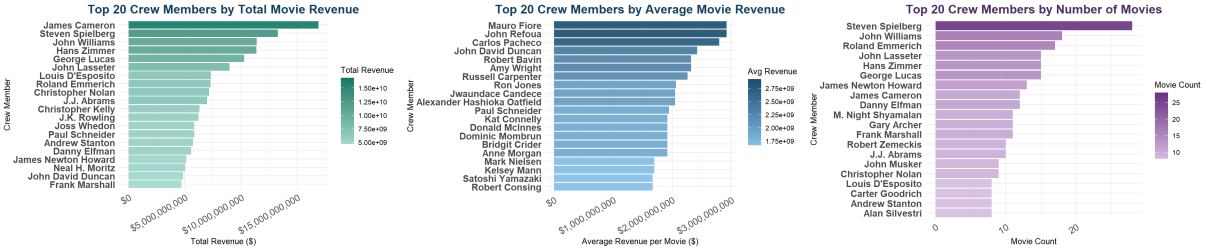

Figure 11: Bar charts showing the top 20 crew members ranked by total movie revenue (left), average revenue per movie (center), and number of movies they have worked on (right). Similar to actors, the rankings vary significantly across these metrics, suggesting that crew influence on revenue is complex. Some crew members might have high total revenue due to working on many films, while others might have high average revenue from contributing to a few blockbusters without necessarily being the key driver of success. Likewise, those with high movie counts might often work on lower-grossing films. As with actors, these factors are difficult to combine, and none alone seem to provide a reliable estimate of how crew members influence box office performance.

Figure 12: Correlation Heatmap – Movie Revenue vs. Predictors

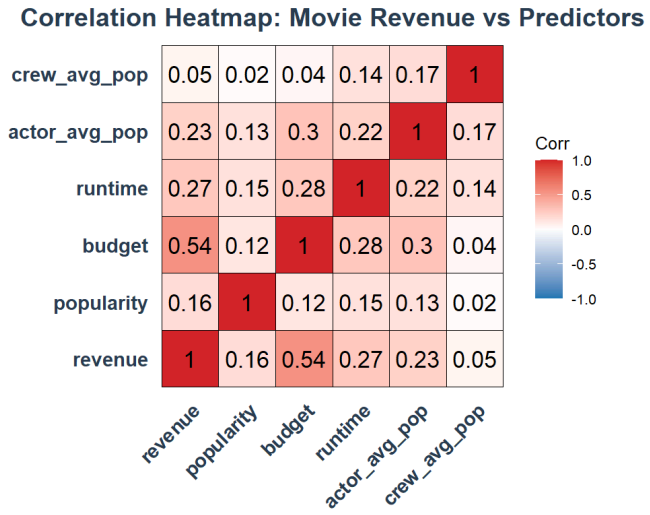

Figure 12: Correlation heat map showing relationships between movie revenue and numeric predictors. Budget has the highest positive correlation with revenue (0.54), reinforcing its importance in determining box office performance. Actor popularity (0.23) and runtime (0.27) also show moderate correlations, while crew popularity (0.05) appears to have little influence. Popularity as recorded by TMDB (0.16) has a weaker-than-expected correlation with revenue, suggesting it may not fully capture audience interest or box office potential.

Table 1: Summary Statistics for Numeric and Date Variables

Variable	Min	Median	Mean	Max
Revenue ($)	190.3M	335.8M	442.6M	2.924B
Budget ($)	60K	87.0M	100.4M	460M
TMDB Popularity	0.151	4.012	4.067	17.029
Runtime (min)	70	117	119.1	233
Release Date	1939-12-15	2011-04-03	2008-12-17	2024-12-19
Avg. Actor Popularity	0.025	0.442	0.457	1.475
Avg. Crew Popularity	0.001	0.059	0.081	0.552

Table 1: Summary statistics of numeric and date variables. Revenue and budget are in USD, while runtime is in minutes. The release date spans from 1939 to 2024, reflecting a diverse range of films. Actor has higher overall popularity than crew, reinforcing the earlier finding that actors play a more visible role in influencing box office performance.

Table 2: Count of Movies by Release Season

Season	Fall	Spring	Summer	Winter
Movie Count	214	229	305	207

Table 2: This table shows the distribution of movies across different seasons. Summer has the highest number of releases (305 movies), which aligns with the common trend of studios launching major blockbusters during peak holiday periods. Spring and Fall have similar counts, while Winter has the lowest number of releases (207 movies), possibly due to fewer big-budget films being released during film awards season.

Table 3: Regression Analysis for Movie Revenue by Numeric Predictors

Predictor	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	6.6302	0.1309	50.645	$< 2e^{-16}$
$\log_{10}$(Budget)	0.2116	0.0173	12.256	$< 2e^{-16}$
Popularity	0.0207	0.0062	3.354	0.000828
Runtime	0.0015	0.0003	4.945	$9.01e^{-07}$
Avg. Actor Popularity	0.0426	0.0347	1.227	0.220
Avg. Crew Popularity	−0.0284	0.0934	−0.304	0.761

Table 3: Regression Analysis for Movie Revenue by Numeric Predictors. This table presents the regression results for log-transformed revenue against various numeric predictors. As expected, budget, popularity, and runtime show significant positive relationships with revenue ($\alpha = 0.05$). Crew average popularity remains insignificant ($\alpha = 0.05$), aligning with earlier findings that audiences rarely consider crew members. However, actor average popularity is also ($\alpha = 0.05$), contrasting with our scatter plot, which suggested a moderately strong trend. This suggests that while popular actors correlate with higher revenue, other factors may confound this relationship in a regression setting.

Table 4: Regression Analysis for Movie Revenue by Genre

Genre	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	8.55590	0.0242	353.551	$< 2e^{-16}$
Action	0.01398	0.01841	0.759	0.448
Adventure	0.09573	0.01711	5.594	$2.92e^{-08}$
Fantasy	0.04043	0.01892	2.137	0.03287
Science Fiction	0.04676	0.01941	2.410	0.01616
Drama	−0.02894	0.01941	−1.491	0.136
Romance	−0.00432	0.02411	−0.179	0.858
Animation	0.09180	0.02896	3.170	0.00157
Comedy	−0.07384	0.01839	−4.016	$6.39e^{-05}$
Family	−0.04474	0.02793	−1.602	0.110
Thriller	−0.03759	0.01990	−1.889	0.05922
Crime	−0.03987	0.02437	−1.636	0.102
History	−0.08651	0.04049	−2.137	0.03289
Music	−0.00389	0.04673	−0.083	0.934
War	0.00932	0.04048	0.230	0.818
Mystery	−0.02485	0.03120	−0.797	0.426
Horror	−0.09746	0.03395	−2.870	0.00419
Western	−0.11233	0.08075	−1.391	0.165
Documentary	−0.17190	0.15171	−1.133	0.257

Table 4: Regression Analysis for Movie Revenue by Genre. This table shows the effect of movie genre on log-transformed revenue. Adventure, Fantasy, Science Fiction, and Animation have significant positive effects, while Comedy, History, and Horror have significant negative effects ($\alpha = 0.05$). Other genres (Action, Drama, Romance, Family, Thriller, Crime, Music, War, Mystery, Western, and Documentary) show no strong statistical relationship with revenue.

Table 5: ANOVA Analysis of Movie Revenue by Season

Source	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
Season	3	0.84	0.28086	5.399	0.0011
Residuals	951	49.47	0.05202		

Table 5: ANOVA Analysis of Movie Revenue by Season. This table presents the ANOVA results testing whether average revenue differs across release seasons. The p-value of 0.0011 suggests a statistically significant difference in revenue among seasons ($\alpha = 0.05$). This supports our earlier visualization showing variations in seasonal revenue distributions, particularly with Spring and Summer tending to host more high-grossing films.

Summary

This study explored factors influencing movie revenue using data from The Movie Database (TMDB) API. The analysis included key numeric predictors **budget, popularity, runtime, seasonal release timing, actor popularity,** and **crew popularity**, along with categorical predictors representing movie **genres**.

Key Findings

1. Budget is the strongest predictor of revenue, showing a significant positive correlation. This aligns with expectations, as higher budgets typically fund better production, marketing, and distribution.
2. Popularity and runtime also exhibit significant positive effects, suggesting that highly searched movies and longer films tend to generate more revenue.
3. Actor popularity was expected to be a strong predictor based on scatter plot analysis, but in the regression model, it was not statistically significant. This indicates that while popular actors may correlate with higher revenue, their impact is likely confounded by other factors.
4. Genres influence revenue: Adventure, Fantasy, Science Fiction, and Animation genres significantly boost revenue, whereas Comedy, History, and Horror have negative effects.
5. Seasonal release timing matters, with ANOVA analysis confirming revenue differences across seasons, supporting the trend of high-revenue films often being released in Spring and Summer. This aligns with industry practices, as these seasons coincide with major holiday periods, leading to higher audience turnout and increased box office potential.

Limitations

While this analysis provides valuable insights, several limitations exist:

- Actor and crew-based prediction challenges
  Only the top 5 actors and top 5 crew members (directors, producers, etc.) were considered to simplify the analysis. While individual popularity alone was not a strong predictor, it makes sense that well-known figures in the industry can influence a movie's success. A more advanced approach could explore alternative popularity metrics or include a larger set of actor and crew members. However, this adds complexity, especially since the TMDB API may not return the full list, requiring additional handling of missing or incomplete data.
- Lack of marketing and external factors
  This analysis does not account for marketing budget, critical reception (e.g., reviews and awards), or competition from other films released at the same time. These factors can significantly impact a movie's revenue but are difficult to quantify and incorporate without additional data sources.

Next Steps

This midterm project sets the foundation for further analysis, which will be expanded in the final project. The next steps will focus on statistical modeling and incorporating more advanced techniques to improve prediction accuracy.

- Refining regression models: Expanding the linear model by considering interaction terms and applying polynomial basis functions or splines to capture potential nonlinear relationships.
- Exploring tree-based models: Implementing decision trees, random forests, and XGBoost to compare their predictive performance against regression models.
- Feature selection and refinement: Investigating alternative popularity metrics for actors and crew, as well as adjusting for inflation in budget and revenue to improve model consistency.
- Deeper analysis of genres and trends: Examining multi-genre effects and identifying long-term revenue trends over different decades.

These improvements will enhance the analysis and contribute to a more comprehensive final project. Building on these findings, the final report will refine the models, incorporate new insights, and provide a more complete understanding of what drives movie revenue.