Predicting Box Office Success: A Data-Driven Approach

Final Project, 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), I 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.

Building on this exploratory phase, I implemented two predictive models—Random Forest and XGBoost—to assess how well movie features can explain and predict revenue. These models complement the earlier statistical analysis by capturing nonlinear relationships and interactions that may not be evident through traditional regression alone.

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. I 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. Although TMDB is a user-contributed platform, it maintains a high standard of data reliability through strict editorial guidelines, active community moderation, and a formal reporting process for incorrect entries. Data is typically sourced from official press kits, on-screen credits, and production studio websites, with user edits subject to review. These safeguards, documented in TMDB's Contribution Bible, help ensure the accuracy and reliability of the movie metadata used in this project.

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.
- Adjusted movie budgets for inflation using monthly CPI data from Consumer Price Index for All Urban Consumers: All Items in U.S. City Average. Each movie's budget was scaled by the ratio between the CPI at the latest release month and the CPI at the movie's original release month: `budget_adj = budget * (CPI_latest / CPI_release_month)`. Revenue,

however, was left unadjusted. High-revenue films often undergo re-releases—sometimes decades after the original launch—making a clean, single-year adjustment inaccurate and potentially misleading. In many of these cases, the impact of inflation is balanced out by the extra earnings from these re-releases, so I chose to leave revenue as-is.

- 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.

Predictive Modeling

To improve upon the initial linear models and explore nonlinear relationships between movie features and revenue, I implemented two predictive algorithms: **Random Forest** and **XGBoost**. These models were used to predict `log10(revenue)` using cleaned and engineered features from the dataset.

For Random Forest, I used the `randomForest` package in R with 500 trees and default hyperparameters. This model handles nonlinearity well and provides built-in variable importance metrics, which made it useful for both prediction and interpretation. Prior to training, I excluded character-based and identifier variables (e.g., title, movie_id) and ensured `season` was treated as a categorical variable.

XGBoost, a gradient boosting model known for its strong predictive performance, was implemented using the `xgboost` package. Since XGBoost requires numeric input, I created a model matrix with one-hot encoded categorical variables and log-transformed budget (`log10(budget)`) as a predictor. Like with Random Forest, the model target was `log10(revenue)`. The XGBoost model was trained with the following hyperparameters:

- `nrounds = 100`: This specifies the number of boosting rounds (i.e., how many trees to build sequentially). Each tree attempts to correct the errors of the previous one, so more rounds generally improve learning—up to a point.
- `max_depth = 6`: This sets the maximum depth of each decision tree. Deeper trees can capture more complex patterns, but also increase the risk of overfitting. A value of 6 is a common balance between complexity and generalization.
- `eta = 0.1`: Also called the "learning rate," this controls how much each new tree contributes to the overall model. Smaller values make the model learn more cautiously (and often require more trees), helping to prevent overfitting.
- `subsample = 0.8`: This means that for each boosting round, only 80% of the training data is randomly sampled. This adds randomness and helps prevent overfitting.
- `colsample_bytree = 0.8`: This means that each tree is trained using only 80% of the available features (columns), chosen randomly. This reduces correlation among trees and improves model robustness.

For both models, I used an 80/20 train-test split and evaluated performance on the test set using R^2 and RMSE metrics. I visualized predicted versus actual log-revenue for the top 100 grossing movies to examine model accuracy across a diverse revenue range. These predictive models complement the linear regression by capturing nonlinear effects and interactions that simpler models may have missed.

Results

Exploratory Data Analysis (EDA)

Table 1 provides a quick overview of the key numeric and date variables in the dataset. It sets the stage for the analysis by showing the scale and variability across movies.

Table 1: Summary Statistics for Numeric and Date Variables

Variable	Min	Median	Mean	Max
Revenue ($)	191.5M	339.0M	445.3M	2.92B
Budget ($)	114.3K	132.7M	145.8M	535.4M
TMDB Popularity	0.38	11.34	14.83	147.03
Runtime (min)	72	116.5	119	224
Release Date	1950-02-22	2011-04-11	2009-02-20	2024-12-19
Avg. Actor Popularity	0.09	2.91	3.33	12.09
Avg. Crew Popularity	0.01	0.32	0.49	4.66

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 1950 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.

The exploratory analysis began with a look at the distribution of movie **revenue**. As shown in Figure 1, the raw revenue distribution is heavily skewed due to a handful of massive hits, while the log-transformed version offers a clearer view of underlying patterns.

Figure 1: Movie Revenue Distributions

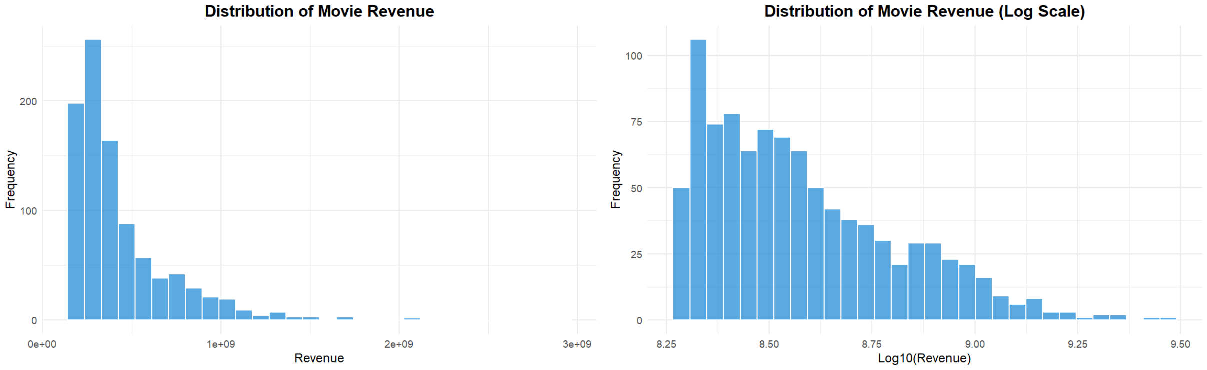

Figure 1: Histograms of movie revenue on both the raw scale (left) and a $\log_{10}$ scale (right). The raw distribution is heavily right-skewed with a few extremely high earners, while the log transformation provides a more balanced view of the data.

To better understand how specific features relate to revenue, I created scatter plots with fitted trends for **budget**, **popularity**, and **runtime**. Figure 2 illustrates that movies with higher budgets generally earn more, and this relationship is clearest on a log-log scale. Popularity (Figure 3) shows a weaker trend, which was somewhat surprising given its strong presence on

TMDB. Runtime (Figure 4) has a moderate positive association, suggesting longer films may be perceived as more cinematic or event-like.

To better understand how specific features relate to revenue, I created scatter plots with fitted trends for **budget**, **popularity**, and **runtime**. Figure 2 shows that movies with higher budgets tend to earn more, and this relationship is especially clear on a log-log scale. Popularity (Figure 3) also shows a fairly strong upward trend, indicating that TMDB's popularity metric does capture some meaningful signal about box office performance. Runtime (Figure 4) has a moderate positive association, suggesting that longer films might feel more like big productions or special events, which could help draw larger audiences.

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

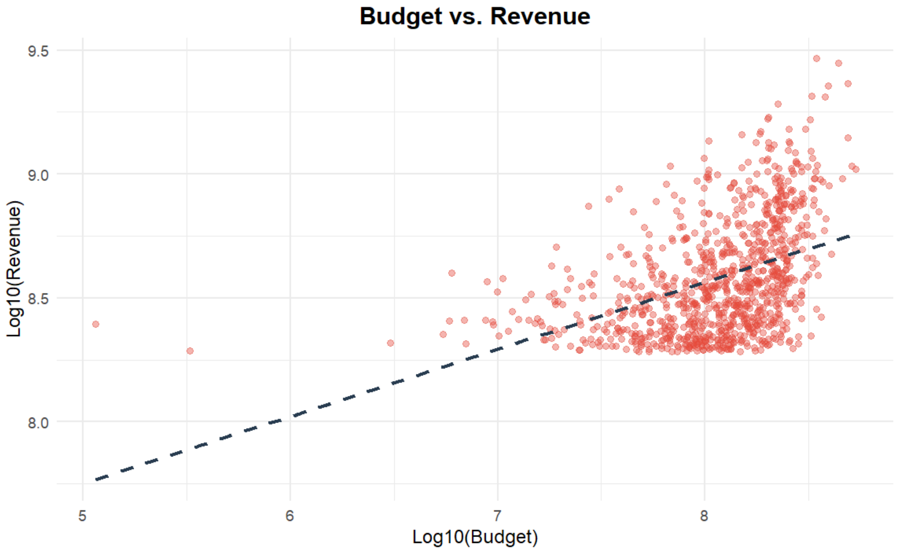

Figure 2: Scatter plot of budget vs. revenue, both on $\log_{10}$ scales. The dashed line represents the fitted linear trend, indicating a generally positive relationship between higher budgets and higher revenues.

Figure 3: Popularity vs. Revenue

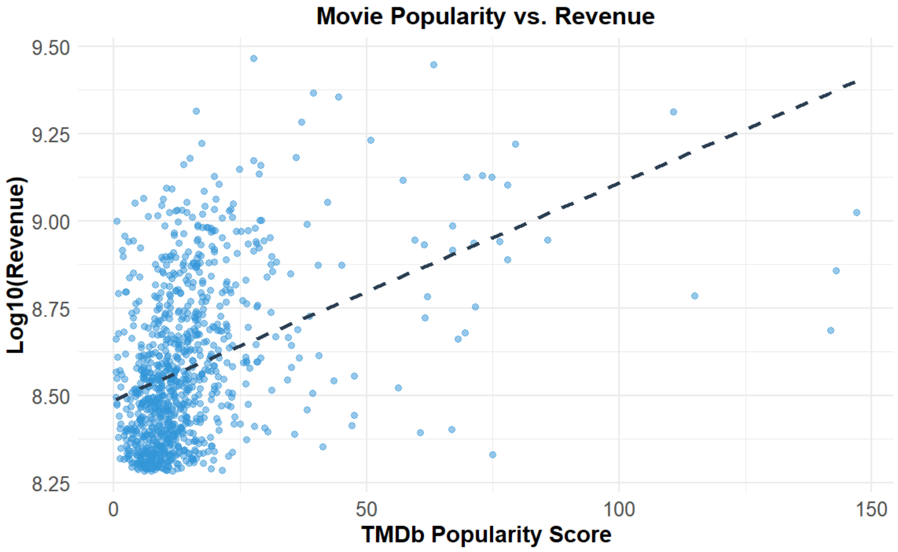

Figure 3: Scatter plot of TMDB popularity vs. $\log_{10}$(revenue). The dashed trend line suggests a modestly positive relationship, though the effect appears weaker than that of budget.

Figure 4: Runtime vs. Revenue

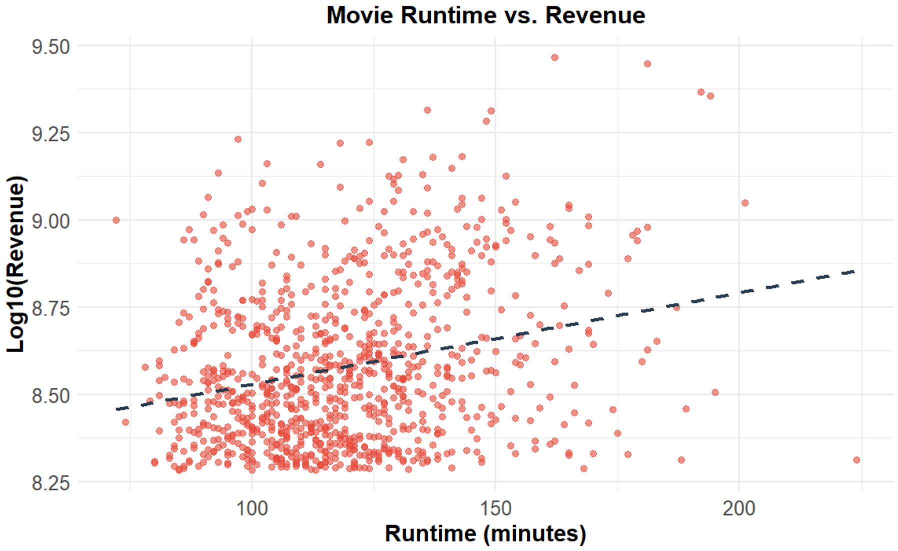

Figure 4: Scatter plot of movie runtime vs. $\log_{10}$(revenue). The dashed regression line indicates a moderate positive trend, suggesting that longer movies may tend to earn slightly higher revenues.

Beyond numeric predictors, I also explored how **seasonality** and **historical trends** relate to movie revenue. Table 2 shows that Summer has the highest number of movie releases, which aligns with the blockbuster release schedule often adopted by studios. Spring and Fall are similar, while Winter tends to have fewer high-budget films. Consistent with this, Figure 5 shows that movies released in Spring and Summer tend to earn slightly more than those released in Fall or Winter, possibly due to school breaks and holiday schedules. Figure 6 then tracks average revenue over time, revealing an overall upward trend likely driven by rising production values and international box office growth—with a noticeable dip in 2020, almost certainly due to the COVID-19 pandemic's impact on theaters.

Table 2: Count of Movies by Release Season

Season	Fall	Spring	Summer	Winter
Movie Count	215	227	302	204

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.

Figure 5: Movie Revenue by Release Season

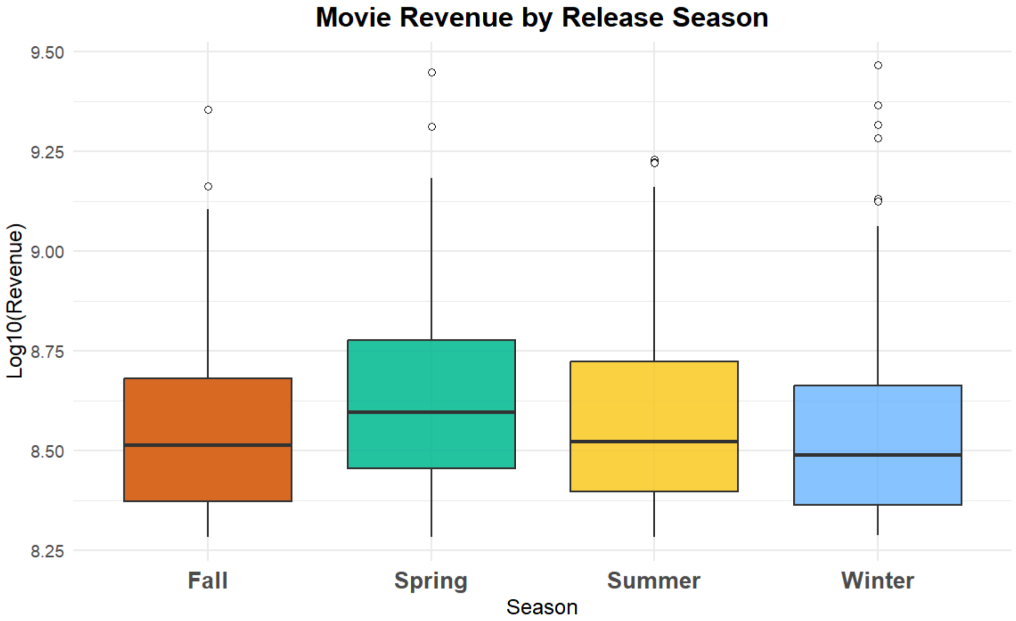

Figure 5: Boxplots of $\log_{10}$(revenue) by season (Fall, Spring, Summer, Winter). The median revenue does not vary drastically between seasons, but Summer and Spring appear to have slightly higher median values compared to Fall and Winter.

Figure 6: Average Movie Revenue Over Time

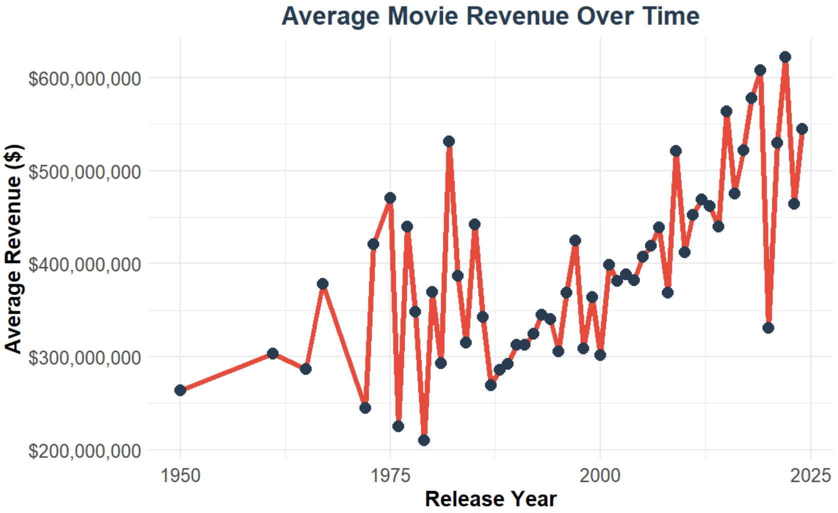

Figure 6: Time series plot of average movie revenue by release year. The general trend shows increasing revenues over time, with more volatility in recent decades. A noticeable dip around 2019-2020 could be attributed to the impact of the COVID-19 pandemic on the film industry.

Genre-wise, Figure 7 reveals that Adventure, Science Fiction, Fantasy, and Animation dominate the top revenue tier. This reinforces what we often see in practice—large-scale, visually driven blockbusters tend to perform well globally.

Figure 7: Revenue Distribution by Genre

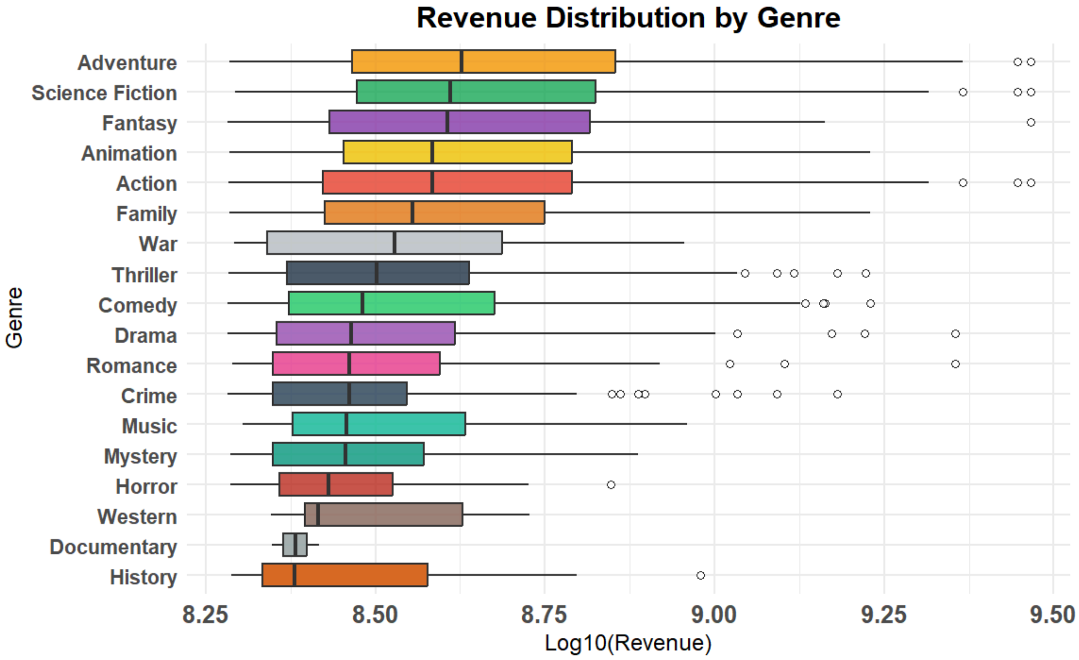

Figure 7: Boxplots of $\log_{10}$(revenue) for different genres. Adventure, Science Fiction, Fantasy, and Animation films tend to have the highest median revenues, while Documentary and History films generally have lower revenue distributions. The presence of extreme outliers suggests that a few blockbuster 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.

Actor popularity, shown in Figure 8, exhibits a moderately strong relationship with revenue, which matches expectations—big-name stars tend to attract bigger audiences. On the other hand, Figure 9 highlights that **crew popularity** has little to no correlation with revenue, reflecting how less visible these roles are to audiences.

Figure 8: Actor Popularity vs. Revenue

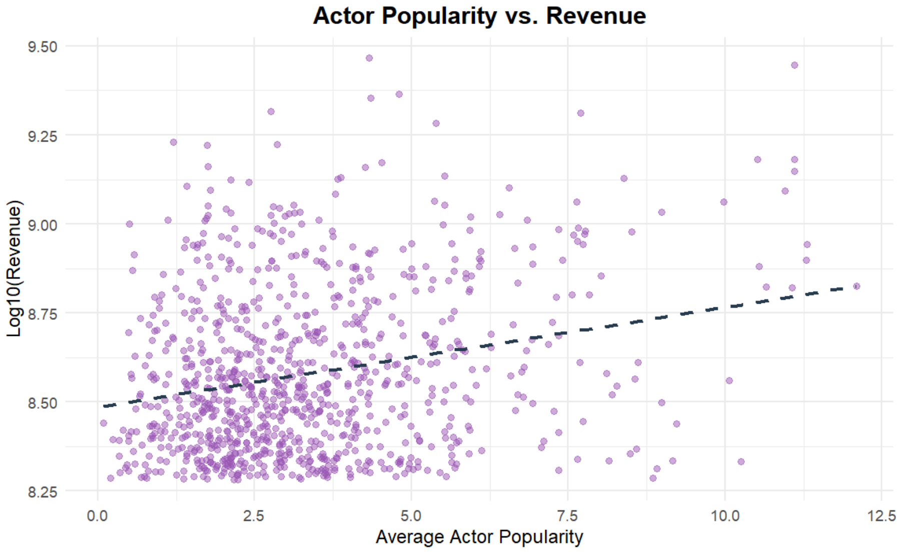

Figure 8: Scatter plot of average actor popularity vs. $\log_{10}$(revenue). There is a moderately strong positive relationship, indicating that movies featuring more popular actors tend to generate higher revenue. This aligns with expectations, as well-known actors often attract larger audiences and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

Figure 9: Crew Popularity vs. Revenue

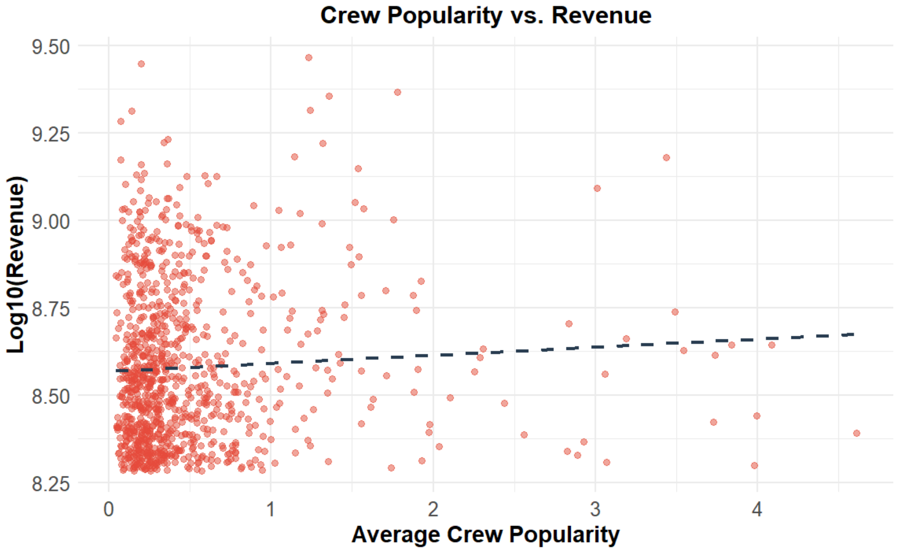

Figure 9: Scatter plot of average crew popularity vs. $\log_{10}$(revenue). Unlike actor popularity, crew popularity shows little to no correlation with revenue, suggesting that a well-known production team does not necessarily translate to higher box office earnings. This also makes sense when we think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

Figures 10 and 11 dive deeper into **actor and crew impact** by comparing **total, average, and count**-based revenue metrics. These reveal that no single metric can fully explain a person's contribution to a movie's success, highlighting the complexity of attributing box office outcomes to individuals.

Figure 10: Top 20 Actors by Revenue Metrics

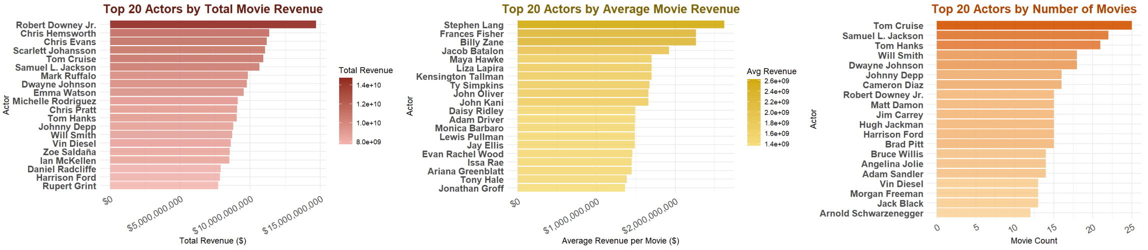

Figure 10: Bar charts showing the top 20 actors ranked by total movie revenue (left), average revenue per movie (center), and number of movies they have appeared in (right). The rankings vary significantly across the three categories, suggesting that actor influence on revenue is complex. Some actors 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

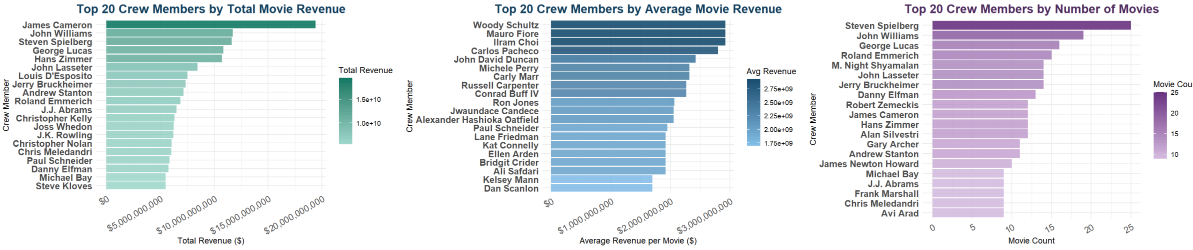

Figure 11: Bar charts showing the top 20 crew members ranked by total movie revenue (left), average revenue per movie (center), and number of movies they have worked on (right). Similar to actors, the rankings vary significantly across these metrics, suggesting that crew influence on revenue is complex. 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.

Finally, Figure 12 presents a correlation heatmap that reinforces many of the earlier observations. Budget remains the strongest numeric predictor of revenue (0.50), followed closely by popularity (0.41), which is stronger than initially expected. Actor popularity and runtime show moderate correlations (0.25 and 0.27 respectively), while crew popularity appears to have little influence (0.07).

Figure 12: Correlation Heatmap – Movie Revenue vs. Predictors

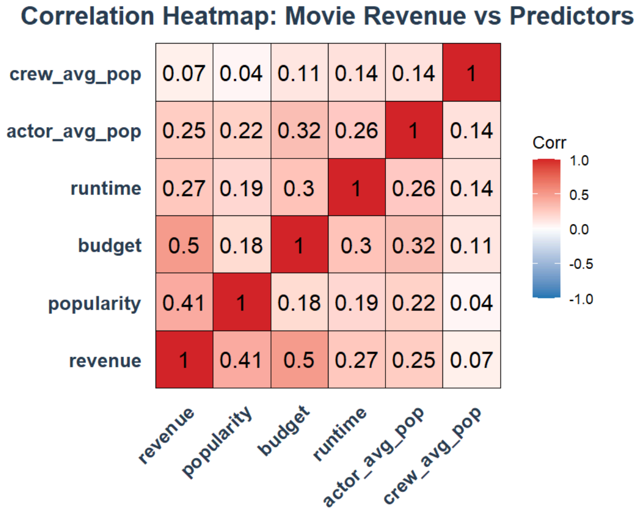

Figure 12: Correlation heat map showing relationships between movie revenue and numeric predictors. Budget has the strongest positive correlation with revenue (0.50), followed by popularity (0.41). Actor popularity (0.25) and runtime (0.27) show moderate relationships, while crew popularity (0.07) appears to have minimal effect.

Statistical Modeling

While pairwise plots offer helpful insights, they don't capture the full picture—particularly when multiple variables interact. To address this, I turn to statistical modeling, which quantifies each predictor's effect while controlling for others. Table 3 shows a linear regression using numeric predictors: log-budget, popularity, and runtime are significant, but actor and crew popularity are not—despite earlier visual trends.

Table 3: Regression Analysis for Movie Revenue by Numeric Predictors

Predictor	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	6.52	0.15	44.58	$< 2e^{-16}$
$\log_{10}$(Budget)	0.23	0.02	12.16	$< 2e^{-16}$
Popularity	< 0.01	< 0.01	9.77	$< 2e^{-16}$
Runtime	< 0.01	< 0.01	3.84	< 0.01
Avg. Actor Popularity	< 0.01	< 0.01	1.09	0.27
Avg. Crew Popularity	-0.01	0.01	-0.56	0.57
Summary Statistics	R^2 = 0.28	RMSE = 0.19		

Table 3: Regression Analysis for Movie Revenue by Numeric Predictors. This model uses log-transformed revenue as the outcome and includes five numeric predictors. Log-transformed budget, popularity, and runtime are all statistically significant ($\alpha = 0.05$), with positive relationships. Actor and crew popularity, however, are not significant after controlling for other variables. Notably, actor popularity showed a visible trend in earlier plots, but its effect weakens when adjusting for covariates, suggesting possible confounding or multicollinearity.

Table 4 builds on this by incorporating genres, revealing significant positive effects for Adventure, Science Fiction, and Animation—genres often associated with large-scale, visually spectacular productions that tend to attract wider audiences and generate higher box office returns. On the other hand, Comedy, History, and Horror show significant negative effects, which may reflect more niche appeal, cultural specificity in humor or historical context, or in the case of horror, a tendency to target narrower, genre-loyal audiences. These genres may also have shorter or more limited theatrical runs compared to blockbuster-oriented releases.

Table 4: Regression Analysis for Movie Revenue by Genre

Genre	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	8.55	0.03	340.91	$< 2e^{-16}$
Action	0.02	0.02	0.93	0.35
Adventure	0.09	0.02	5.47	< 0.01
Fantasy	0.04	0.02	1.99	0.05
Science Fiction	0.05	0.02	2.47	0.01
Drama	-0.03	0.02	-1.34	0.18
Romance	-0.01	0.02	-0.25	0.80
Animation	0.08	0.03	2.74	0.01
Comedy	-0.07	0.02	-3.94	< 0.01
Family	-0.03	0.03	-0.91	0.36
Thriller	-0.03	0.02	-1.61	0.11
Crime	-0.03	0.03	-1.21	0.23
History	-0.09	0.04	-2.18	0.03
Music	-0.01	0.05	-0.14	0.89
War	0.02	0.04	0.38	0.71
Mystery	-0.03	0.03	-1.04	0.30
Horror	-0.09	0.03	-2.70	0.01
Western	-0.11	0.08	-1.40	0.16
Documentary	-0.17	0.15	-1.11	0.27
Summary Statistics	R^2 = 0.18	Adj. R^2 = 0.16		

Table 4: Regression Analysis for Movie Revenue by Genre. This table shows the effect of movie genre on log-transformed revenue. Adventure, Science Fiction, and Animation have significant positive effects, while Comedy, History, and Horror have significant negative effects ($\alpha = 0.05$). Other genres do not show statistically significant effects on log-revenue. The model explains roughly 18% of the variance in revenue ($R^2 = 0.18$), with an adjusted R^2 of 0.16.

Table 5 shows the results of an ANOVA testing revenue differences by release season. The significant p-value supports earlier visualizations, confirming that some seasons tend to be more profitable windows for movie releases.

Table 5: ANOVA Analysis of Movie Revenue by Season

Source	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
Season	3	0.93	0.31	6.02	< 0.01
Residuals	940	48.55	0.05		

Table 5: ANOVA Analysis of Movie Revenue by Season. This table presents the ANOVA results testing whether average revenue differs across release seasons. The small p-value 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.

Together, these results provide both visual and statistical evidence on what drives movie revenue, forming the foundation for the predictive models explored below.

Predictive Modeling

Random Forest

To move beyond linear assumptions and capture complex interactions between predictors, I trained a Random Forest regression model using 500 trees and default hyperparameters. The model used the same features as before—log-transformed budget, popularity, runtime, actor and crew popularity, season, and genre indicators.

On the test dataset, the model achieved an $\boldsymbol{R^2}$ **of 0.33** and an **RMSE of 0.17**. While not drastically outperforming the linear model, this improvement suggests that nonlinear interactions do exist in the data and are partially captured by the ensemble of decision trees.

Figure 13 shows feature importance based on two metrics: %IncMSE (how much accuracy drops when the variable is permuted) and IncNodePurity (contribution to reducing variance in splits). Budget and popularity dominate both measures, followed by runtime and actor popularity. Crew popularity again ranks near the bottom, suggesting that its predictive contribution remains weak even in nonlinear models.

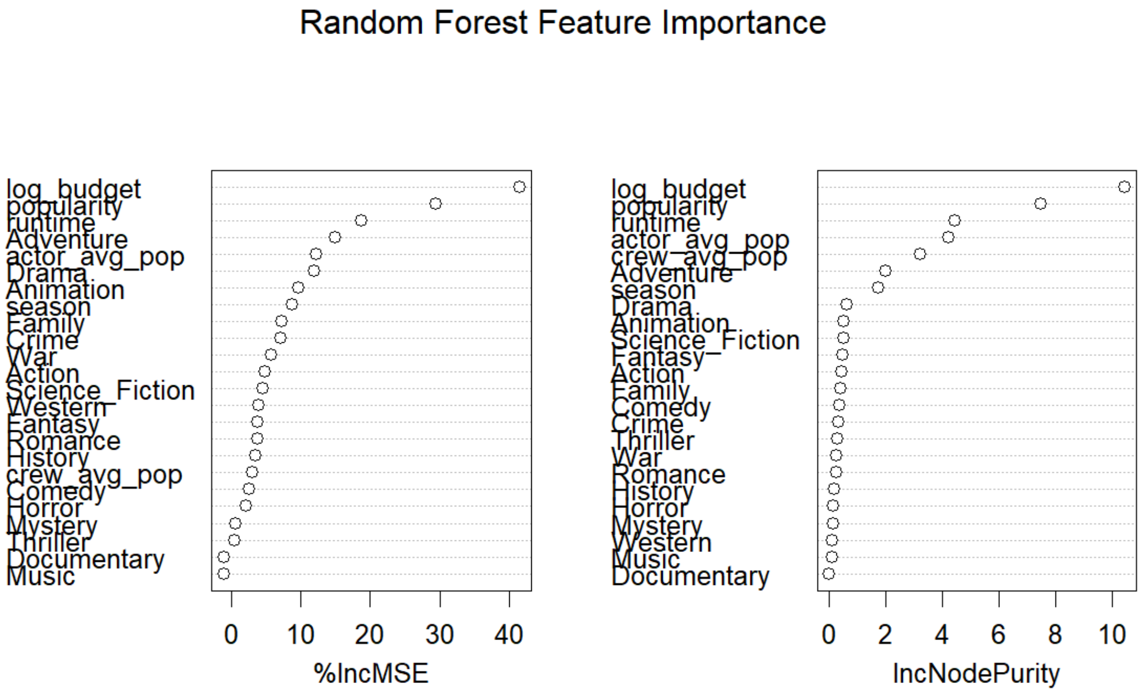

Figure 13: Random Forest feature importance plots measured by %IncMSE (left) and IncNodePurity (right). Budget and popularity stand out as the most influential predictors, followed by runtime and actor popularity. Crew popularity remains among the least important features, reaffirming earlier findings.

Overall, Random Forest provides an interpretable and moderately accurate predictive model that reinforces the importance of a few dominant features while offering some gains over simpler methods.

XGBoost

Building on the Random Forest model, I trained a gradient boosting model using XGBoost—an algorithm known for its accuracy in structured data problems. This model was tuned with reasonable default hyperparameters and trained on the same features used previously.

On the test set, XGBoost performed slightly better than Random Forest, achieving an **R^2 of 0.41** and **RMSE of 0.17**. This suggests the model captures nonlinear interactions and subtle patterns better than both linear regression and Random Forest, though gains remain modest.

Feature importance (Figure 14) once again places log-budget and popularity at the top, followed by crew and actor popularity. Interestingly, crew popularity has higher gain in this model than in Random Forest, potentially due to how XGBoost constructs additive trees. Other predictors—including runtime, drama, and seasonal effects—also play a role, though with diminishing importance.

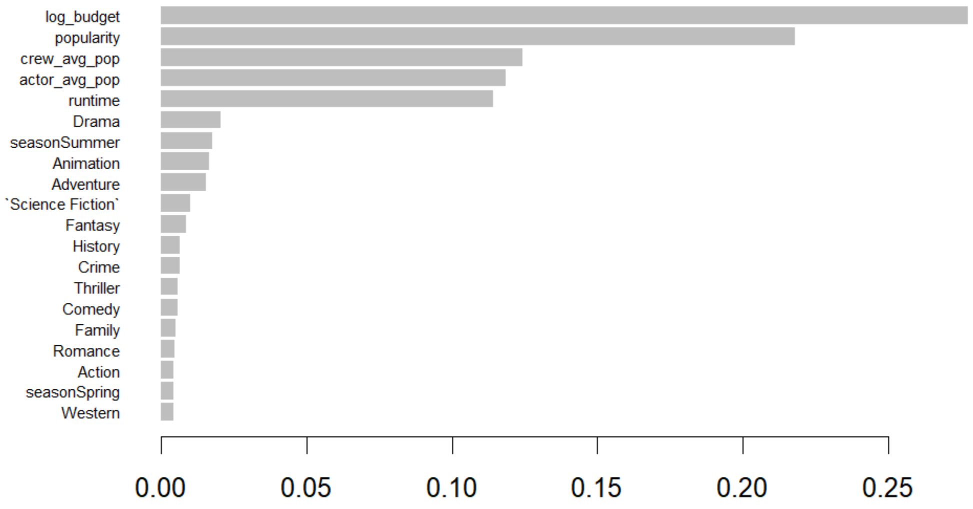

Figure 14: XGBoost feature importance plot. The top features by gain are log-budget, popularity, crew popularity, and actor popularity. Runtime, drama, season, and genre indicators follow with lesser contributions. This distribution reinforces earlier findings that budget and popularity are central to predicting box office success.

Overall, XGBoost outperformed other models in terms of R^2, confirming its strength in capturing complex feature interactions. That said, even the best model explains only about 40% of the variation in revenue, suggesting that unobserved variables (e.g. marketing budget, franchise loyalty, or timing competition) still play a substantial role.

Conclusions and Summary

This project examined the key drivers of box office revenue using data from The Movie Database (TMDB) API. By combining exploratory data analysis, statistical modeling, and machine learning, I investigated how production factors like **budget**, **popularity**, **runtime**, **seasonal release**, **actor/crew popularity**, and **genre** relate to a movie's financial success.

Key Findings

1. **Budget** consistently emerged as the strongest predictor of revenue—both visually and statistically. This confirms the intuitive idea that larger budgets often enable better production, marketing, and distribution, which in turn support box office success.
2. **Popularity** and **runtime** also showed significant positive effects, suggesting that more frequently searched and longer-form movies perform better financially.
3. While **actor popularity** displayed a positive visual trend, it was not statistically significant in the regression model. This implies that its effect may be entangled with other factors like budget or franchise brand, diluting its standalone impact.
4. **Genres** mattered: Adventure, Science Fiction, and Animation had significantly positive effects on revenue, while Comedy, History, and Horror were associated with significantly lower earnings. These results are consistent with the performance of visually grand, globally appealing blockbusters versus more niche or culturally dependent genres.
5. **Release season** plays a role: movies released during Spring and Summer tended to earn more, and this was supported by ANOVA results. These seasons often coincide with school breaks and major holidays, making them prime time for theatrical releases.

Model Comparison

To assess how well these variables can predict revenue, I compared the performance of three models: linear regression, random forest, and XGBoost. Table 6 summarizes their predictive accuracy on the test set using R^2 and RMSE.

Model	R^2	RMSE
Linear Regression	0.28	0.19
Random Forest	0.33	0.17
XGBoost	0.41	0.17

Table 6: Comparison of predictive performance across three models. XGBoost achieved the best R^2, capturing about 41% of the variance in log-revenue. Random Forest also performed well, slightly outperforming linear regression. Despite these improvements, all models leave a substantial portion of variance unexplained, pointing to the limits of prediction using metadata alone.

XGBoost outperformed both Random Forest and linear regression, likely due to its ability to model complex interactions and nonlinearities. However, even the best model explains just 41% of the variance in revenue—highlighting the role of unobserved factors like franchise loyalty, marketing, or critical reception.

Limitations

- Actor and crew-based prediction challenges
  Only the top 5 actors and crew members were used to simplify the analysis. While actor popularity had visual significance, it was not a strong standalone predictor in models. Expanding this to include more cast/crew or alternative popularity metrics may improve accuracy—but would also add complexity due to API limitations.

- Popularity is not always available pre-release
  TMDB popularity is often computed after release by aggregating user activity like upvotes, downvotes, ratings, and page views. As a result, it reflects post-release audience engagement rather than pre-release expectations. For predictive use before a movie's debut, this metric may not be available. Incorporating it in a forecasting model would require a separate model to predict popularity itself based on pre-release information (e.g., cast, trailer views, franchise history), which introduces another layer of uncertainty and complexity.

- Lack of external context and marketing data
  Important variables like marketing spend, franchise affiliation, or competitive releases are not captured in this dataset, yet likely explain much of the revenue variation. These omissions limit the models' ability to fully predict outcomes.

- Inflation adjustment limited to budget only
  While budget was adjusted for inflation using monthly CPI data, revenue was not. Many top-grossing films include re-releases years after their original debut, making it difficult to pin down a single "release year" for adjusting revenue without overcomplicating the model. I therefore left revenue as-is, assuming that inflationary loss is partially offset by re-release earnings.

Next Steps

With the foundations established, several avenues exist to strengthen this analysis:

- Model enhancements: Explore polynomial regression, splines, or ensemble methods combining multiple models.
- Feature engineering: Improve how popularity is quantified and test additional time-based variables like decade or holiday release.
- Incorporate external datasets: Adding critic/audience scores, award wins, or international box office breakdowns could improve performance.

Overall, this project demonstrates both the potential and the limits of using metadata to predict movie success. While models like XGBoost offer powerful tools for capturing complex patterns, their effectiveness is ultimately bounded by the scope and quality of available data. Many influential factors—such as marketing budget, franchise loyalty, critical reception, timing of competing releases, and social media buzz—are difficult to quantify or are simply unavailable in public datasets. This makes movie revenue prediction a uniquely challenging problem, one where data science can provide strong signals but not full certainty. Still, the ability to forecast even part of the story is valuable—and perhaps, in the world of movies, leaving room for surprises is part of the magic.