Analyzing Movie Success: The Impact of Credits, Revenue, and Audience Perception

https://github.com/Dennis-Ding1/movie-success-analysis/tree/main/Midterm

Haochen Ding

1. Introduction

Understanding what drives a movie's success is a crucial question in the film industry. Factors such as genre, budget, cast, and audience perception play a significant role in determining a movie's box office performance. This study using data collected from The Movie Database (TMDB) API and Kaggle. Aims to explore the relationship between movie credits, box office revenue, and audience ratings and perception, incorporating both numerical and text-based analysis. Specifically:

How do movie credits (e.g., genre, cast, production) influence box office revenue and audience perception?

To address this, I propose the following hypotheses:

- Credits and Revenue: Larger casts and larger production scale contribute to higher box office earnings. Higher budgets generally lead to greater box office revenue, but the rate of increase diminishes as budgets grow, diminishing returns.
- Audience Reviews and Perception: Sentiment expressed in audience reviews significantly differs across movie genres. And movies with larger casts or more production companies tend to receive more positively worded audience reviews.

By integrating credit-based attributes, financial performance, and text analysis, this study aims to identify key factors driving a movie's success in both revenue and audience reception.

2. Method

2.1 Data collection & Merging

For this analysis, I used four datasets: three sourced from TMDB API and one from Kaggle. Since TMDB does not provide box office revenue data, and unlimited access to the IMDb API is not available, I opted to use an existing Kaggle dataset that contains revenue, budget, IMDb ratings, and additional metadata (https://www.kaggle.com/datasets/alanvourch/tmdb-movies-daily-updates/data).

From TMDB api (https://api.themoviedb.org/), I collected:

1. The top 2000 highest-TMDB-rated movies, with related movie credits.
2. Genre IDs to genre names look up table.
3. Audience reviews (up to 10 per movie).

Datasets were merged with kaggle dataset, using movie IDs from TMDB as link to perform left join (TMDB dataset as base), ensuring proper integration of financial, genre, and audience perception data. The final dataset includes 2000 obversations, and 26 variables:

- Movie details: title, release date, original language, cast, genre etc.
- Box office performance and budget.
- Audience ratings (TMDB and IMDB).

Audience reviews are in a separate dataset, this is for faster performance. It contains 5026 observations (reviews).

2.2 Data cleaning

To ensure data usability, I performed several preprocessing steps on the dataset. This included extracting relevant numerical features, handling missing values, and filtering unrealistic financial data.

1. Feature Engineering

Some columns in the dataset contained comma-separated lists (e.g., production companies, cast, genres, and production countries). To make these attributes more analytically useful, I transformed them into numerical features representing the number of elements in each category:

- num_production_companies: The number of production companies involved in each movie.
- num_cast: The number of credited cast members.
- num_genre: The number of genres assigned to a movie.
- num_production_countries: The number of countries associated with production.
- first_genre: The most significant genre of movies. This is for data model analysis, but will not be used in text analysis. Then I use genre id and name look up dataset to cover id to actual name of the genre.

2. Handling Missing and Unrealistic Values

To ensure meaningful analysis, I filtered out movies with missing or unrealistic financial data:

- Movies with NA values in revenue or budget were removed, as these are essential for analyzing box office performance.
- Movies with revenue or budget less than 100 were excluded, as such values are likely incomplete or inaccurate records.

2. Cleaning Audience Review Text

To prepare the audience review text for text analysis, I first removed entries with missing content. Each review was then lowercased, stripped of non-alphabetic characters (while retaining apostrophes), and cleaned of common English stopwords to reduce noise. Extra whitespace was trimmed, and reviews that became empty after cleaning were removed. This preprocessing ensured that the text data was standardized and ready for tokenization and sentiment scoring.

2.3 Analyze methodology

To explore the relationship between movie credits, box office revenue, and audience perception, I employ a combination of numerical modeling and text analysis. The study is structured into two key analytical components:

1. Analyzing the Influence of Movie Credits on Box Office Revenue

- Linear Regression: A multiple linear regression model is used to estimate the individual contributions of key factors, such as budget, cast size, production scale, and genre, to box office revenue. This model provides an interpretable framework to assess whether larger productions and higher budgets are associated with increased revenue. Additionally, the regression model helps evaluate the presence of diminishing returns.

Random Forest: Because revenue-generation processes may involve complex, nonlinear relationships, a random forest model is used to capture patterns that linear models might miss. Random forests are ensemble models that build many decision trees and average their predictions to improve accuracy and reduce overfitting. This approach allows the model to detect non-linear dependencies (relationships that don't follow a straight-line pattern) and higher-order interactions (how multiple variables combine in complex ways). Additionally, random forests produce variable importance measures, which quantify how much each predictor contributes to the model's accuracy. These measures help identify which movie credit features — such as budget, cast size, or genre — have the most substantial impact on box office revenue.

2. Model Diagnostics and Evaluation

- To ensure the validity of the linear regression model, I perform standard diagnostic checks including residual plots, Q-Q plots, and leverage plots. These diagnostics evaluate key assumptions of the model such as linearity (whether the relationship between predictors and the outcome is linear), homoscedasticity (constant variance of residuals), normality of residuals, and the absence of overly influential observations. The residuals vs. fitted plot is used to identify any non-linear trends or heteroskedasticity, while the Q-Q plot assesses whether the residuals are approximately normally distributed. The leverage plot, which incorporates Cook's distance, helps identify high-leverage points that may disproportionately affect the model's estimates. These diagnostic tools are essential for verifying model assumptions and ensuring that the regression outputs are reliable and interpretable.
- The performance of the random forest model is evaluated using several metrics that capture both fit and prediction accuracy. These include the coefficient of determination (R2), which reflects the proportion of variance in revenue explained by the model, as well as mean absolute error (MAE) and root mean squared error (RMSE), which measure the average and squared differences between predicted and actual values, respectively. To interpret the contribution of individual predictors, I use two variable importance measures provided by the model: percentage increase in mean squared error (%IncMSE), which quantifies how much prediction error increases when a variable is excluded, and increase in node purity (IncNodePurity), which captures how useful a variable is for splitting the decision trees. In addition, I use partial dependence plots to visualize how each variable affects predicted revenue, holding other variables constant. These plots, generated using the pdp package in R, reveal the shape and direction of non-linear effects, offering intuitive insights into the marginal influence of each predictor on revenue outcomes.

3. Examining Audience Perception Through Text Analysis

- Word Frequency and TF-IDF Analysis: To explore how audience perception varies across film genres, I begin by analyzing the textual content of audience reviews through word frequency and TF-IDF (term frequency–inverse document frequency) analysis. Word frequency reveals the most commonly used terms in reviews for each genre, highlighting shared audience concerns or themes. TF-IDF, on the other hand, emphasizes words that are not just frequent but also distinctive to a particular genre by downweighting terms that appear across many genres. This method helps surface genre-specific vocabulary that reflects how audiences uniquely engage with different types of films, offering insight into the narrative elements, cultural references, or emotional tones that resonate most strongly in each category.
- Sentiment Analysis and Modeling Integration: I apply sentiment analysis using the AFINN lexicon, which assigns numeric scores to words based on their emotional polarity, allowing each review to be quantified by its overall tone. These sentiment scores are aggregated at the movie level and compared across genres to examine whether certain film types elicit more positive or negative reactions. To assess whether sentiment carries predictive value, I incorporate average sentiment scores into the linear regression model and perform an ANOVA test to evaluate whether it significantly improves model fit. Sentiment is also added to the random forest model, where its impact is assessed through changes in prediction accuracy and variable importance. This modeling integration enables a quantitative assessment of whether audience emotion, as reflected in written reviews, contributes meaningfully to explaining variation in box office revenue.

3. Results

3.1 Data Summary

Table 1: Summary of Numerical Variables

Variable	Min	Max	Mean	Variance	NA_Count
TMDB Rating	7.3	8.71	7.67	0.084	0
TMDB Vote Count	300.0	37157.00	5426.49	37904068.617	0
Revenue	881.0	2923706026.00	160113771.29	83221796010854256.000	0
Budget	17300.0	460000000.00	34920075.24	2864943950035675.000	0
IMDB Rating	1.7	9.30	7.70	0.267	0
IMDB Votes	341.0	3017438.00	323867.36	158056122827.740	0
Production Companies	1.0	22.00	3.44	5.500	0
Production Countries	1.0	9.00	1.50	0.808	0
Number of Casts	1.0	308.00	49.19	1073.960	0
Number of Genres	1.0	7.00	2.58	1.000	0

Table 1 shows some statistical summary of the numerical variables used in this research. After data cleaning, they are all in reasonable range.

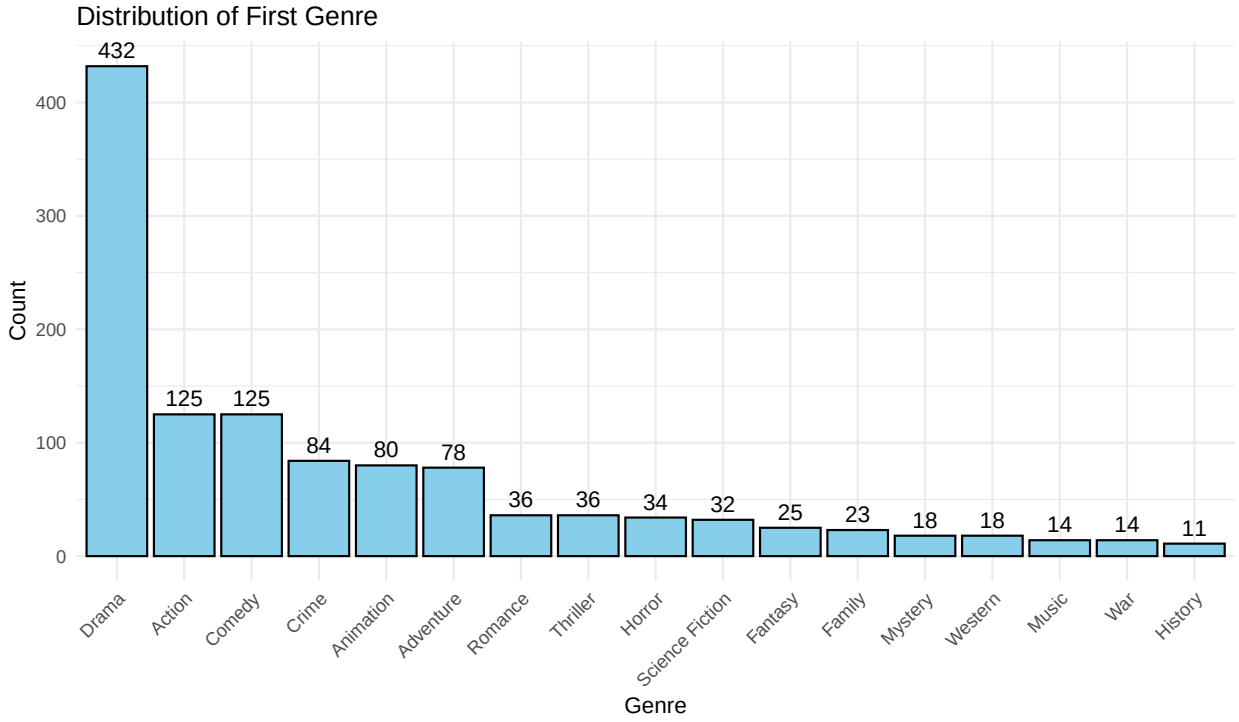

Figure 1: Distribution of First Genre

Figure 1 shows that Drama is the most common genre (432), followed by Action and Comedy (125 each). Less frequent genres include History (11), War (14), and Music (14), indicating a dominance of drama films in the dataset.

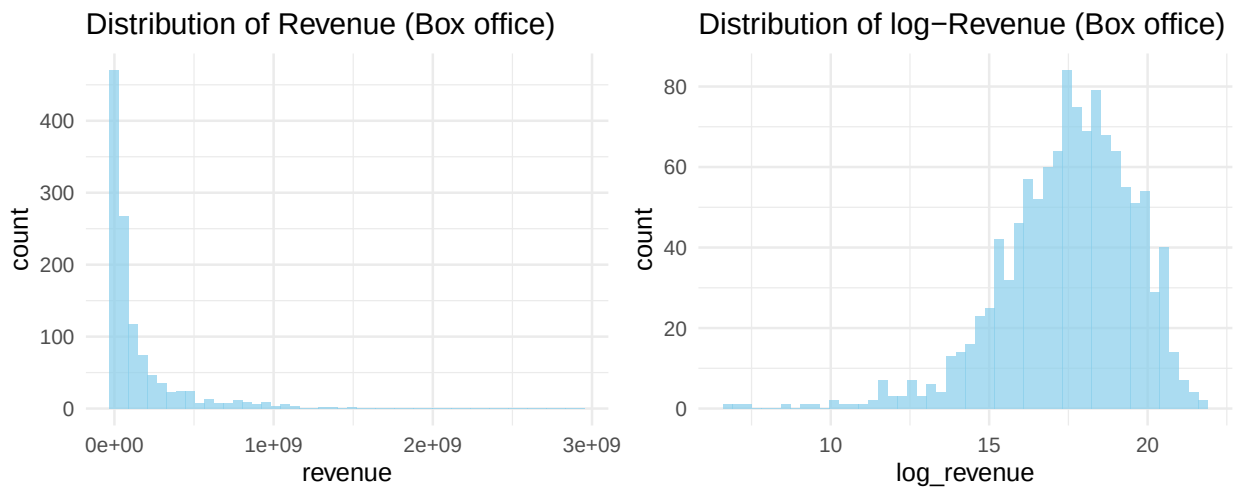

Figure 2: Distribution of raw and log-transformed revenue

Figure 2 shows the raw revenue distribution is highly skewed, with a long tail of extremely high values, making it difficult to analyze relationships effectively and create valid linear model. Taking the log transformation of revenue reduces skewness, creating a more normal-like distribution. This transformation ensures regression model is meaningful.

3.2 Linear Regression

The multiple linear regression model examines the relationship between number of production companies / countries / genre / cast, budget, and IMDB / TMDB ratings on log-transformed box office revenue. The model explains 42.5% of the variance, with an adjusted R-squared of 0.413, indicating a moderate fit.

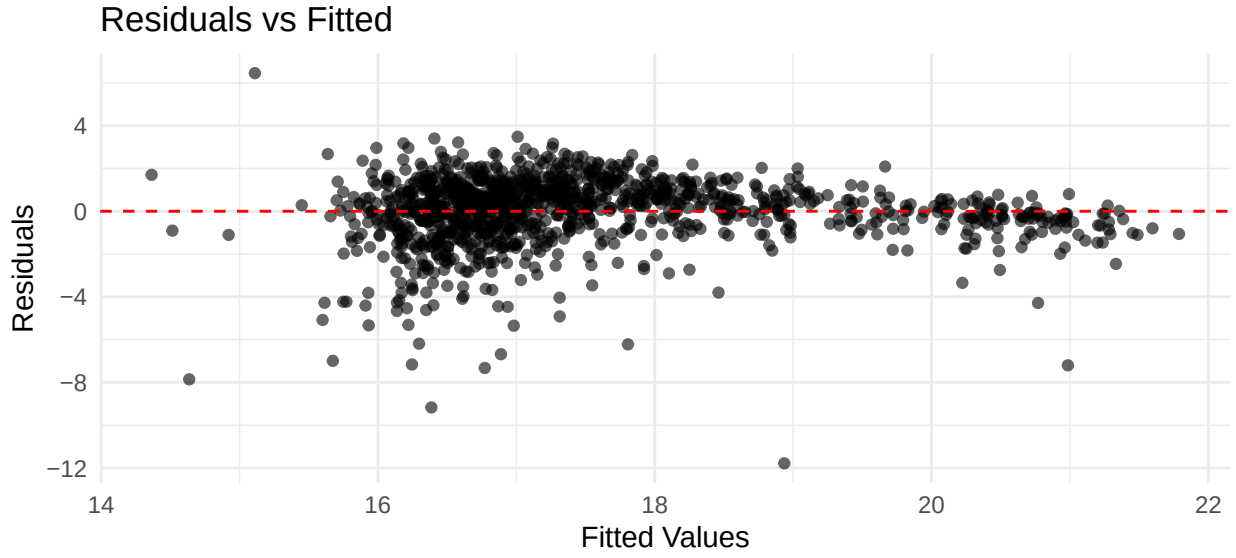

Figure 3: Residuals vs Fitted: This plot assesses linearity and homoscedasticity. The red dashed line at zero helps identify deviations from constant variance or non-linear trends.

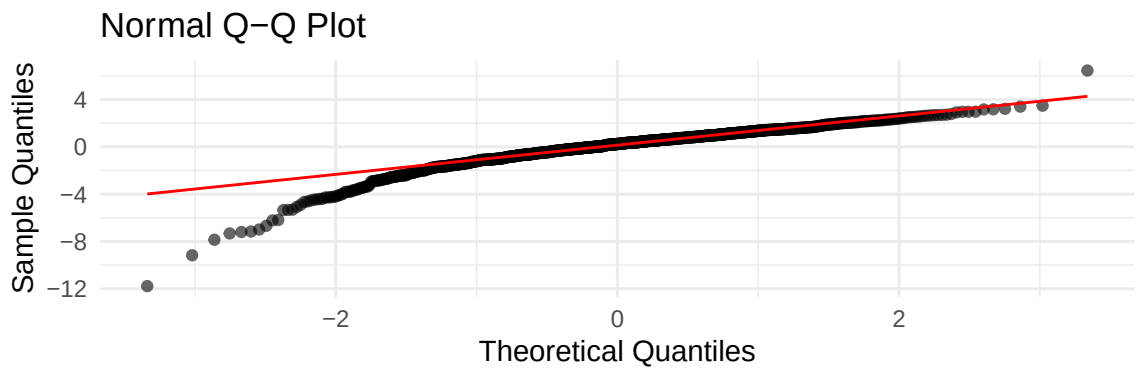

Figure 4: Normal Q-Q Plot: This plot evaluates whether the residuals follow a normal distribution. Points deviating from the red line suggest potential non-normality.

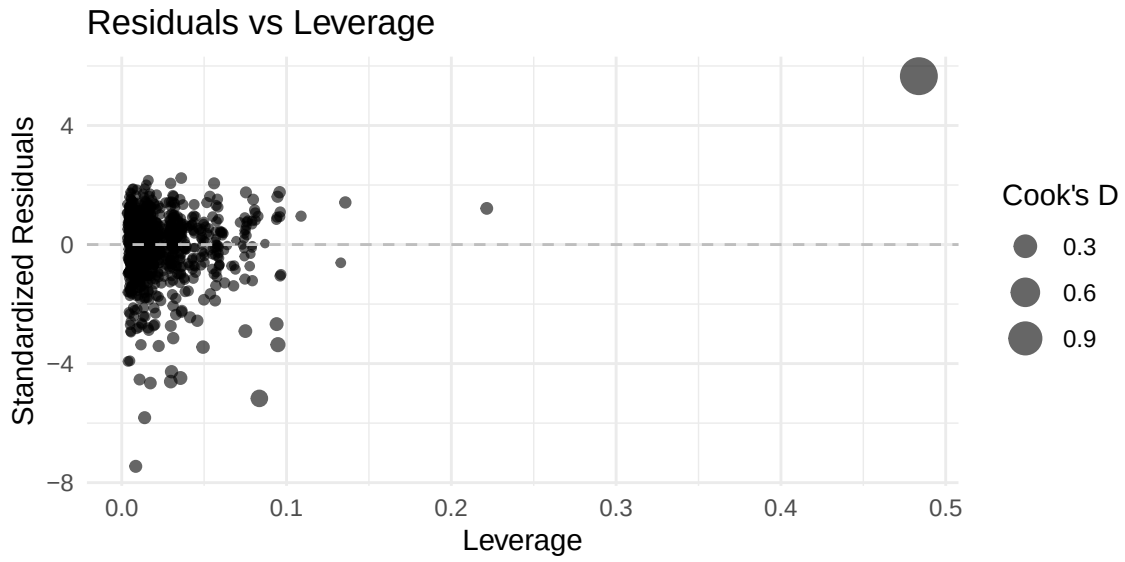

Figure 5: Residuals vs Leverage: This plot highlights influential observations using leverage and standardized residuals. Point sizes reflect Cook's distance, with larger points indicating higher influence.

Figure 3, the Residuals vs Fitted plot shows a generally random scatter around the horizontal line at zero, suggesting that the assumptions of linearity and homoscedasticity are reasonably met. However, some funneling on the left indicates potential mild heteroskedasticity in lower fitted values. Figure 4, the Q-Q plot reveals notable deviations from the theoretical line, particularly in the lower tail, indicating some departure from normality in the residuals. Figure 5, the Residuals vs Leverage plot highlights one high-leverage observation with a large Cook's distance, suggesting that a single point may exert undue influence on the model. Overall, while the model appears broadly valid, these diagnostics suggest cautious interpretation and the possible need for robustness checks or alternative modeling approaches.

Table 2 shows coeffiecent and their p values of the model:

1. *Budget and Diminishing Returns:*

- Budget has a strong positive effect on revenue, confirming that higher budgets tend to generate higher box office returns.

Table 2: Regression Results

term	estimate	std.error	statistic	p.value	significance
(Intercept)	10.812	1.276	8.474	0.000	very significant
budget	0.000	0.000	17.779	0.000	very significant
I(budget^2)	0.000	0.000	-10.533	0.000	very significant
num_production_companies	0.123	0.026	4.636	0.000	very significant
num_cast	0.004	0.002	2.642	0.008	significant
num_production_countries	-0.229	0.069	-3.339	0.001	very significant
num_genre	-0.009	0.056	-0.157	0.876	not significant
tmdb_rating	0.399	0.198	2.018	0.044	not significant
imdb_rating	0.324	0.114	2.829	0.005	significant
first_genreAdventure	0.104	0.231	0.448	0.654	not significant
first_genreAnimation	-0.249	0.238	-1.046	0.296	not significant
first_genreComedy	-0.163	0.214	-0.761	0.447	not significant
first_genreCrime	-0.386	0.230	-1.680	0.093	not significant
first_genreDrama	-0.218	0.177	-1.230	0.219	not significant
first_genreFamily	-0.346	0.367	-0.942	0.346	not significant
first_genreFantasy	-0.157	0.349	-0.450	0.653	not significant
first_genreHistory	-0.833	0.504	-1.652	0.099	not significant
first_genreHorror	0.559	0.318	1.759	0.079	not significant
first_genreMusic	0.383	0.452	0.846	0.397	not significant
first_genreMystery	-0.606	0.405	-1.497	0.135	not significant
first_genreRomance	0.155	0.308	0.505	0.614	not significant
first_genreScience Fiction	0.053	0.318	0.167	0.868	not significant
first_genreThriller	-0.651	0.305	-2.133	0.033	not significant
first_genreWar	-1.245	0.451	-2.760	0.006	significant
first_genreWestern	-0.512	0.413	-1.239	0.216	not significant

- Budget-squared is negative, indicating diminishing returns, meaning that after a certain point, increasing the budget leads to progressively smaller revenue gains.

2. Production and Cast Effects:

- Number of Production Companies has a positive impact, suggesting that films with multiple production companies tend to perform better.
- Number of Cast Members has a small but significant positive effect, meaning larger casts contribute slightly to higher revenue.
- Number of Production Countries is negatively correlated with revenue, implying that films with international co-productions may not always perform better financially.

3. Genre Influence:

The baseline genre is Action.

- Thriller and War have significantly lower revenues than Action films, suggesting they may not perform as well at the box office. However, their limited presence in the dataset suggests this trend may be influenced by sample size rather than an inherent genre effect.

- Other genres, such as Drama, Comedy, and Horror, do not show statistically significant differences from Action films.

4. Audience Ratings Impact:

- TMDB Rating and IMDB Rating both significant, and have a positive impact on revenue, indicating that higher audience ratings correlate with higher earnings.

3.3 Random Forest

A Random Forest model was trained to evaluate the factors influencing box office revenue based on budget, production details, audience ratings, and genre. The dataset was split into 80% training and 20% testing, and the model was trained with 100 trees.

Table 3: Random Forest Model Performance

Metric	Value
Mean Absolute Error (MAE)	9.128660e+07
Root Mean Squared Error (RMSE)	1.854698e+08
R-Squared (R^2)	6.180409e-01

Table 3 shows the Random Forest model explains 61.8% of revenue variance, with an MAE of $91.3M and an RMSE of $185.5M. While budget, genre, and ratings are key predictors, the high RMSE suggests missing factors like marketing or franchise status, indicating room for improvement.

Table 4: Random Forest Feature Importance

Feature	Percentage Increase in MSE	Increase in Node Purity
Budget	20.75	3.425162e+19
Main Genre	5.97	7.742918e+18
IMDB Rating	4.67	5.811665e+18
Production Countries	3.26	1.184744e+18
Production Companies	3.22	2.919030e+18
TMDB Rating	2.99	6.095374e+18
Number of Casts	1.82	8.098030e+18
Number of Genres	1.30	2.720649e+18

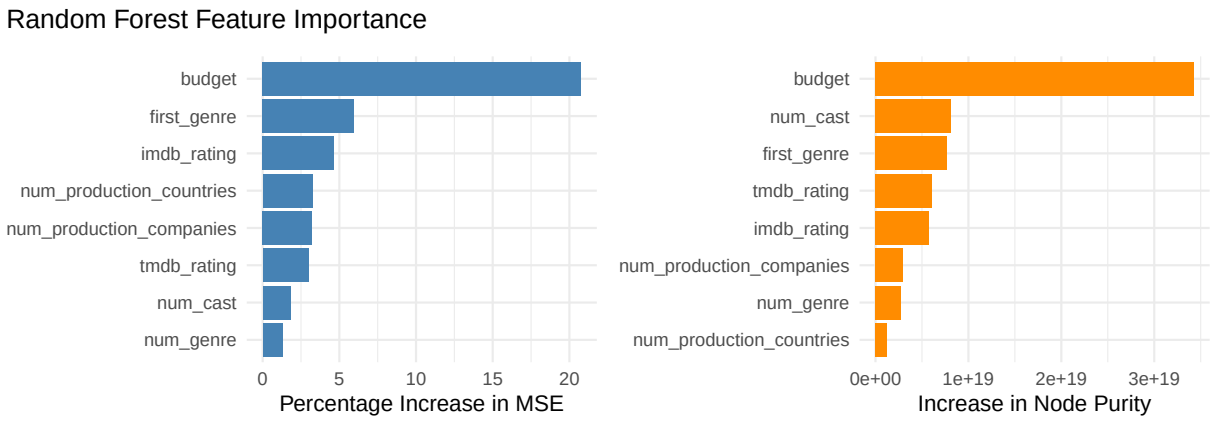

Figure 6: Feature Importance Plot

The Table 4 and Figure 6 shows the key factors affecting revenue:

- Budget is the most important predictor, with the highest Percentage Increase in MSE (20.75) and Node Purity Contribution, confirming that higher budgets lead to higher revenue.
- First Genre (5.97% Increase in MSE) also plays a significant role, suggesting that genre type impacts revenue. However, this contrasts with the regression results, where most first_genre categories were statistically insignificant. This discrepancy suggests that genre does not have a strong independent effect on revenue but rather interacts with other factors like budget and cast size. Random Forest captures these nonlinear interactions, indicating that genre's influence is context-dependent, which may not be fully reflected in a linear model.
- IMDB rating (4.67%) and TMDB rating (2.99%) influence revenue, indicating that higher audience scores correlate with better financial performance.
- Production-related variables (e.g., number of production companies, number of cast members) contribute to revenue but are less impactful than budget and ratings.

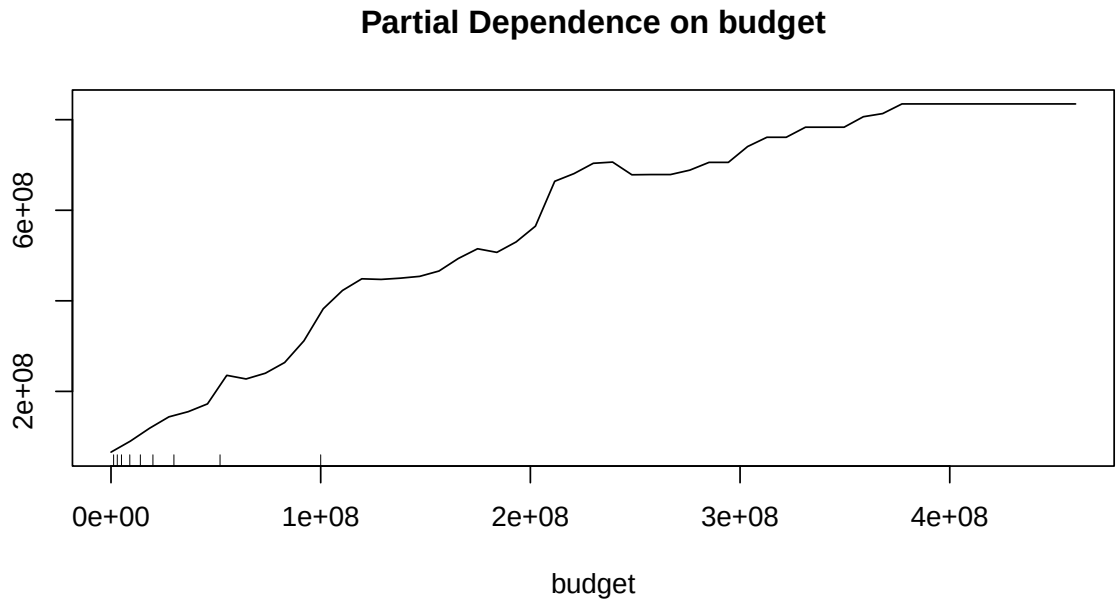

Figure 7: Partial Dependence on Budget

Figure 7 shows how predicted revenue changes with budget, holding other variables constant. The relationship is initially steep, indicating that increasing budgets significantly boost revenue. However, the curve flattens at higher budgets, suggesting diminishing returns—beyond a certain point, additional spending yields smaller increases in revenue. This aligns with our regression findings, where the quadratic budget term was negative, confirming that while budget is a key driver of revenue, its impact is nonlinear.

3.4 Text Frequency and TF-IDF anaylze

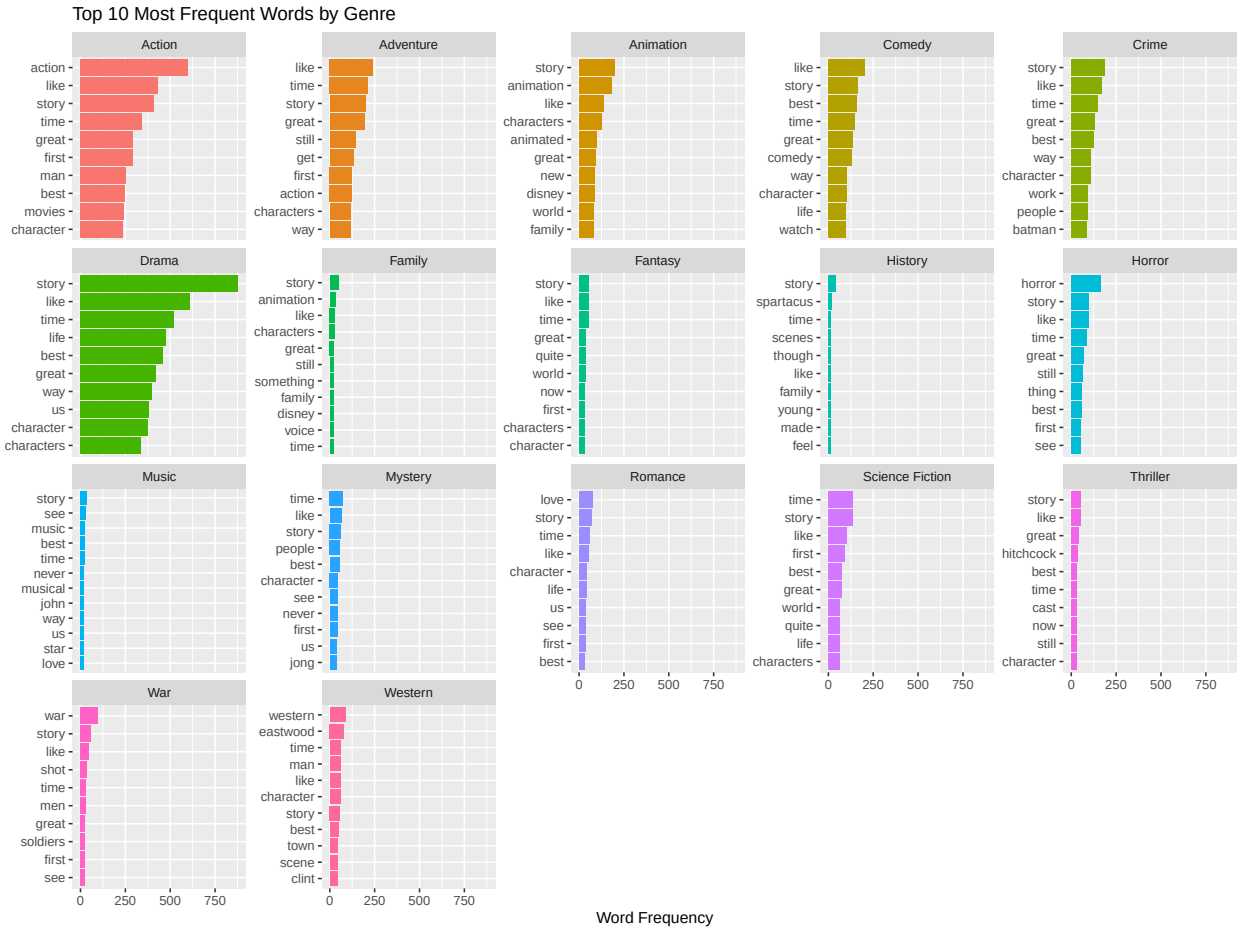

Figure 8: Top 10 Most Frequent Words by Genre

The analysis of word frequency across genres reveals that audience reviews often focus on broadly positive and emotional language, such as “story,” “like,” “great,” and “best,” regardless of genre. This consistency suggests that general storytelling quality and emotional engagement are common criteria in how audiences evaluate films. However, some genre-specific themes also emerge — for instance, words like “action” and “character” are more frequent in Action films, while “love” appears frequently in Romance. These results indicate that while audience language is often universally appreciative, frequent terms can also reflect genre-specific expectations or narrative structures.

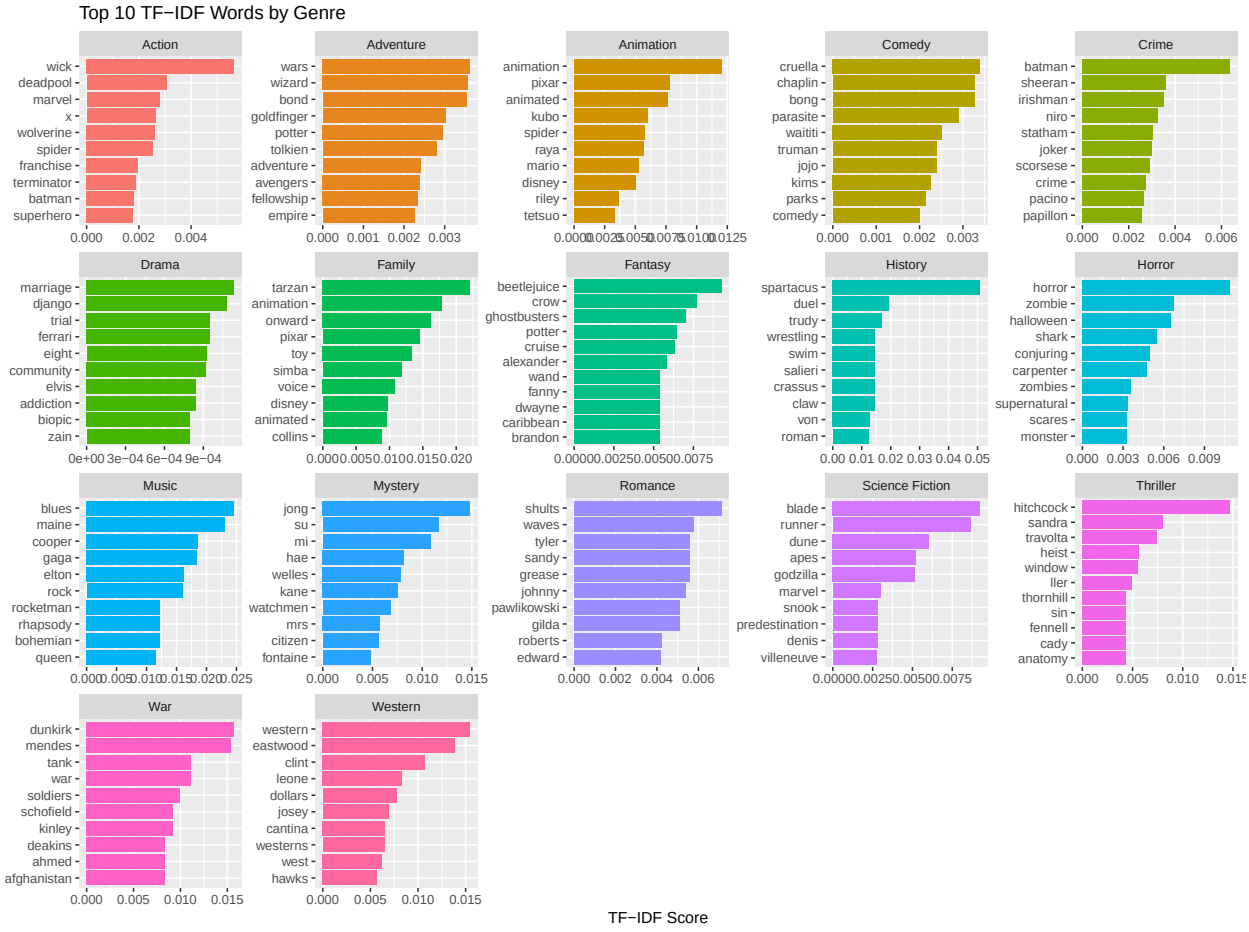

Figure 9: Top 10 TF-IDF Words by Genre

TF-IDF analysis further sharpens these distinctions by highlighting words that are particularly characteristic of each genre's reviews. Unlike word frequency, which favors common language, TF-IDF emphasizes uniqueness — uncovering terms like "zombie" and "supernatural" for Horror, "elvis" and "addiction" for Drama, and "clint" and "eastwood" for Westerns. Many of these words reference notable characters, actors, or themes, suggesting that audience perception is often tied to genre-defining films or tropes. Overall, the TF-IDF results demonstrate how linguistic patterns in reviews reflect the cultural and thematic signatures of each genre more precisely than raw frequency counts.

3.6 Sentiment Analysis

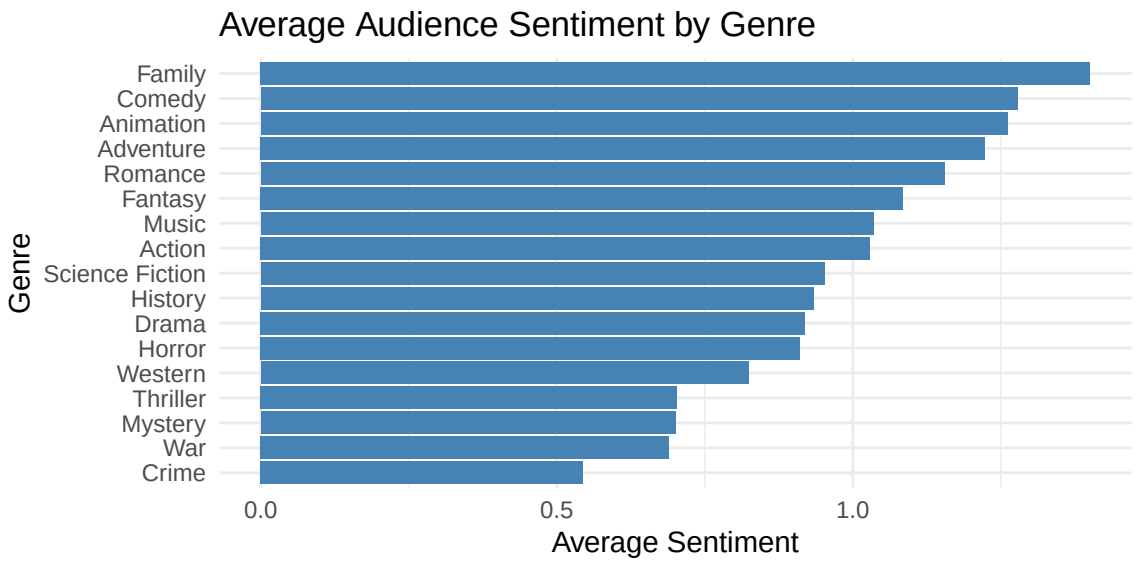

Figure 10: Average Audience Sentiment by Genre

The analysis of average sentiment scores by genre reveals clear variation in audience perception. Family, Comedy, and Animation films receive the most positive sentiment on average, suggesting a generally favorable emotional response. In contrast, genres like Crime, War, and Mystery tend to have lower sentiment scores, indicating more critical or emotionally neutral reviews. These patterns suggest that genre significantly shapes the emotional tone of audience feedback.

To assess whether these sentiment patterns have predictive value beyond descriptive insight, I integrated the average sentiment scores into both the linear regression and random forest models.

Table 5: Comparison of Linear Models With and Without Sentiment Score

Model	Residual DF	Residual RSS	DF	Sum of Squares	F Statistic	p-value
Without Sentiment	945	2387.258	NA	NA	NA	NA
With Sentiment	944	2379.288	1	7.9704	3.1623	0.0757

The ANOVA test yielded an F-statistic of 3.16 with a p-value of 0.0757. It shows that the improvement of adding sentiment score is not statistically significant at the conventional 0.05 threshold. This suggests that while sentiment may carry some explanatory value, its effect on predicting box office revenue is limited and should be interpreted with caution.

Table 6: Random Forest Model with average sentiment score Performance

Metric	Value
Mean Absolute Error (MAE)	1.004962e+08
Root Mean Squared Error (RMSE)	2.017135e+08
R-Squared (R^2)	6.178282e-01

Adding average sentiment to the random forest model resulted in a slight decline in performance, with MAE increasing from 91.29 million to 100.50 million and RMSE rising from 185.47 million to 201.71 million. The

R-squared value remained virtually unchanged at approximately 0.618, indicating that sentiment did not enhance the model's explanatory power. These results suggest that, while sentiment may reflect audience perception, it does not meaningfully improve the predictive accuracy of revenue outcomes in this setting.

4. Conclusions and Summary

4.1 Key Findings

- Budget is the most influential driver of revenue, as consistently shown in both linear regression and random forest models. However, its effect is nonlinear — the marginal returns on revenue diminish as budgets increase, emphasizing the importance of optimizing investment levels rather than simply maximizing them.
- Audience ratings (IMDB and TMDB) positively correlate with box office revenue, indicating that critically appreciated films are more likely to perform well commercially. Still, the moderate strength of this association implies that critical acclaim is only one piece of the commercial success puzzle.
- The role of genre is nuanced and context-dependent. Linear regression found few genres to be statistically significant predictors, likely due to imbalanced representation across categories. In contrast, random forest models suggest that genre exerts an indirect effect through interactions with other variables like budget and production scale. This highlights the importance of considering nonlinear relationships and interaction effects when modeling real-world phenomena.
- Production and cast size matter, but less than budget or ratings. The number of production companies — likely tied to distribution and marketing capabilities — showed a stronger association with revenue than cast size. This suggests that behind-the-scenes production decisions may be more influential than star power alone.
- Text analysis of audience reviews revealed genre-specific discourse patterns. High-frequency words and top TF-IDF terms aligned closely with genre-defining themes, character names, and iconic film elements. This supports the notion that audiences engage with movies through both emotional and thematic lenses, and these discussions reflect collective cultural associations tied to genre.
- Sentiment analysis revealed emotional differences across genres, with Family, Comedy, and Animation films receiving more positive sentiment than darker genres like Crime, War, and Horror. However, incorporating average sentiment into predictive models did not significantly improve their performance. This suggests that while sentiment reflects audience response, it may not directly translate into financial impact — possibly due to lag effects, review bias, or redundancy with other features like ratings.

4.2 Broader Implications

These findings highlight that data-driven decision-making can play a key role in film production and marketing. Budget and production scale remain the most reliable predictors of box office success, underscoring the value of strategic resource allocation. Audience ratings also matter, but their impact is moderate, suggesting that critical acclaim supports — but does not guarantee — financial performance. While genre and sentiment reflect important aspects of audience engagement, their influence appears more complex and indirect, likely interacting with other production factors. Sentiment analysis, though not strongly predictive in this context, still provides valuable qualitative insights into how different genres resonate emotionally with viewers.

4.3 Limitations

This study has several limitations. The dataset likely reflects survivorship bias, excluding low-profile or unreleased films. Genre imbalance may reduce the reliability of genre-specific results. Additionally, the

sentiment analysis was based on a simple lexicon (AFINN), which may miss contextual nuance. More advanced models or audience segmentation by region or platform could reveal deeper patterns. Lastly, the models assume fixed relationships, while real-world audience behavior may shift over time, suggesting that future work should explore temporal dynamics or social media-based signals.