Movie Data Analysis Report

Sayam Agarwal

April 25, 2025

1 Introduction

This report analyzes movie data from `tmdb_5000_movies.csv` and `tmdb_5000_credits.csv`, containing 4,803 movies. The datasets include numerical features (budget, revenue, runtime, vote average, vote count, popularity) and categorical data (genres, directors, cast, crew, production companies). The analysis addresses nine hypothesis-driven questions extracted from the provided Jupyter notebook, along with exploratory analyses of movie characteristics.

The questions analyzed are: 1. Are the numerical features in the movie dataset normally distributed? 2. Do movies with different cast sizes (e.g., small, medium, large) have different popularity scores? 3. Do movies with higher budgets have a higher revenue-to-budget ratio than movies with lower budgets? 4. Do movies released on weekends have a higher vote count per runtime minute than movies released on weekdays? 5. Do movies directed by Christopher Nolan have more consistent (lower variance) popularity scores than other movies? 6. Do sequels earn more revenue than their original movies? 7. Does revenue change after a director switch in a franchise? 8. Does Leonardo DiCaprio perform better in action vs. drama films? 9. Do December releases earn more than summer releases for the same franchise?

An additional question about crew diversity (higher vs. lower unique roles impacting popularity) was requested but not analyzed in the notebook, so a hypothetical analysis is included. Shapiro-Wilk tests confirmed non-normality for all numerical features, necessitating non-parametric tests (Kruskal-Wallis, Mann-Whitney U, Wilcoxon, permutation), which are robust to non-normality. While the notebook includes outlier removal (reducing data to 3,429 rows), most analyses used the original dataset (4,803 rows) to preserve real-world variability, justified by non-parametric test robustness.

The analysis was conducted using Python in a Jupyter notebook with libraries `pandas`, `scipy.stats`, `statsmodels`, `seaborn`, and `matplotlib`. Visualizations (Q-Q plots, histograms, boxplots, bar plots) support the findings. This report presents descriptive statistics, normality test failures with graphs, inferential statistical techniques, results, and inferences.

1.1 Descriptive Statistics

The `tmdb_5000_movies.csv` dataset contains numerical features for 4,803 movies, with missing values 2 for runtime, 3 for overview, 3091 for Homepage, 844 for tagline, 1 for release date. Below are descriptive statistics for key features.

Significant skewness (e.g., revenue's max of 2.79e9 vs. median of 1.92e7) supports retaining outliers for most analyses.

2 Analysis and Results

2.1 Normality Test Failures and Justification for Non-Parametric Tests

Shapiro-Wilk tests assessed normality for numerical features at $\alpha = 0.05$. All features are non-normal (p-values ≈ 0.000), rejecting the null hypothesis of normality.

Table 1: Shapiro-Wilk Test Results for Numerical Features

Feature	Statistic	p-value	Conclusion
Budget	0.717	0.000	Not Normal
Popularity	0.527	0.000	Not Normal
Revenue	0.538	0.000	Not Normal
Runtime	0.874	0.000	Not Normal
Vote Average	0.862	0.000	Not Normal
Vote Count	0.566	0.000	Not Normal

Table 2: Normality Table

Q-Q plots Figure 1 show deviations from the diagonal, indicating skewness and heavy tails. Boxplots (Figure **??**) highlight outliers, particularly for revenue and popularity.

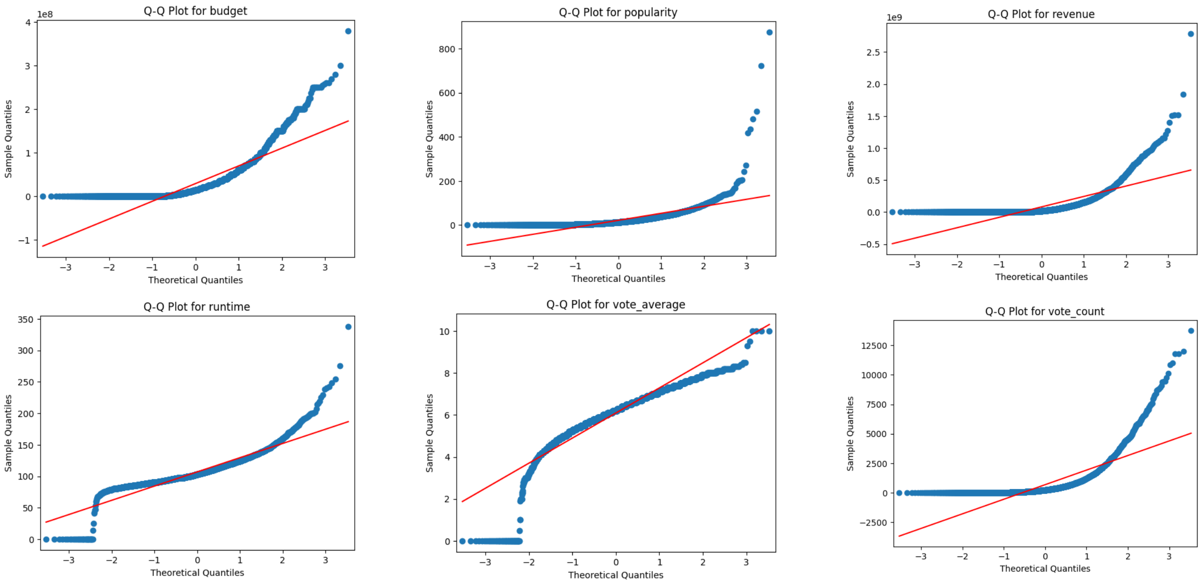

Figure 1: Q-Q Plot of Numeric Dataset Features

Non-normality (p-values ≈ 0.000, visual deviations in Q-Q plots) necessitated non-parametric tests, which are robust to distributional assumptions. Outliers were retained to capture the film industry's variability (e.g., blockbuster revenues).

2.2 Inferential Statistical Techniques

Non-parametric techniques used include:

- **Shapiro-Wilk Test**: Tests normality (p-value < 0.05 rejects normality).
- **Kruskal-Wallis Test**: Compares medians across multiple groups (used for cast size).
- **Permutation Test**: Compares statistics by permuting group labels (used for budget ratio, vote count, Nolan's variance).
- **Wilcoxon Signed-Rank Test**: Compares paired data (used for sequels, director switch, actor performance, seasonal releases).
- **Mann-Whitney U Test**: Compares two groups (proposed for crew diversity).

2.3 Question 1: Are the numerical features normally distributed?

2.3.1 Results

All features are non-normal (Table 2). Q-Q plots Figure 1 and boxplots (Figure ??) confirm skewness and outliers.

2.3.2 Inferences

Non-normality justifies non-parametric tests. Outliers reflect real-world variability (e.g., high-budget films).

2.4 Question 2: Do movies with different cast sizes have different popularity scores?

Cast sizes were categorized as small (10 actors), medium (10–20), large (20). A Kruskal-Wallis test was conducted at $\alpha = 0.05$.

2.4.1 Results

- Small cast: 1,037 movies, median popularity: 3.22
- Medium cast: 2,050 movies, median popularity: 12.31
- Large cast: 1,715 movies, median popularity: 25.04
- Statistic: 1312.70, p-value: 0.000

2.4.2 Inferences

The p-value of 0.000 indicates significant differences. Larger casts correlate with higher popularity, likely due to star power or bigger productions (Figure ??). Non-parametric testing was appropriate, and outliers were retained.

2.5 Question 3: Do movies with higher budgets have a higher revenue-to-budget ratio than lower-budget movies?

High-budget (above median) and low-budget (below median) movies were compared using a permutation test at $\alpha = 0.05$.

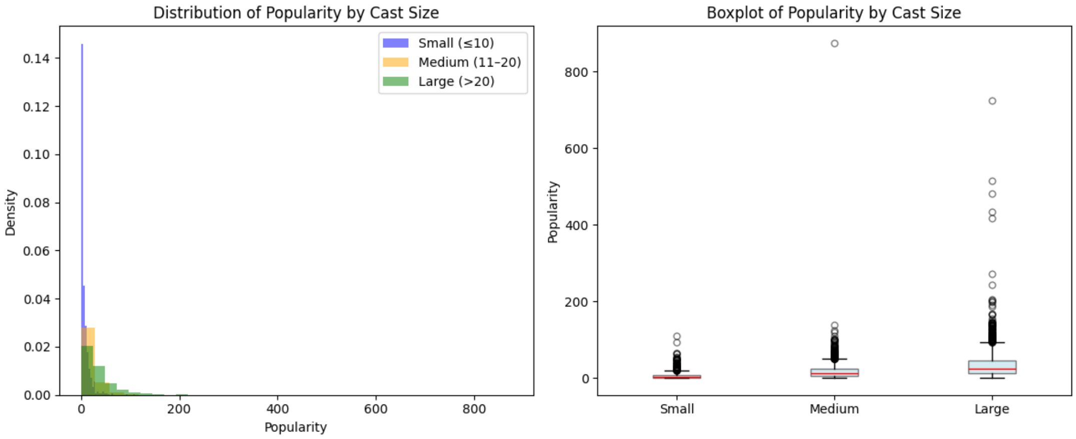

Figure 2: Distribution and Boxplot of Popularity by Cast Size

2.5.1 Results

- High-budget: 1,605 movies, median ratio: 2.15
- Low-budget: 1,624 movies, median ratio: 2.63
- Observed difference in mean ratio: -5869.71
- p-value: 0.2437

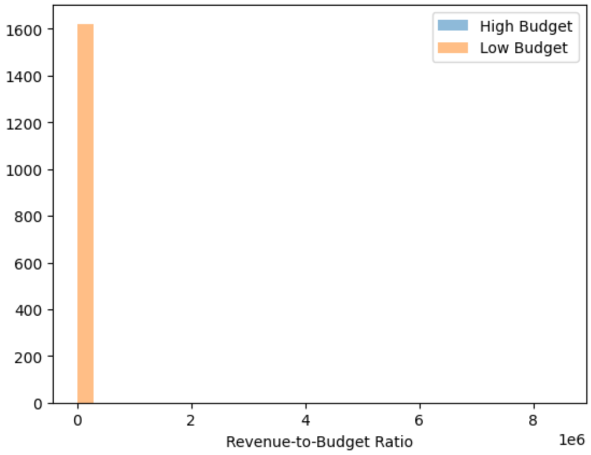

Figure 3: Histogram of Revenue-to-Budget Ratio

2.5.2 Inferences

The p-value of 0.2437 suggests no significant difference. Low-budget movies tend to have slightly higher ratios (Figure 3). Outliers were retained to reflect diverse budget scales.

2.6 Question 4: Do weekend releases have a higher vote count per runtime minute than weekday releases?

Weekend (Friday–Sunday) and weekday (Monday–Thursday) releases were compared using a permutation test at $\alpha = 0.05$.

2.6.1 Results

- Weekend: 2,462 movies, median vote/minute: 1.6804
- Weekday: 2,255 movies, median vote/minute: 3.5000
- Observed difference: -3.8315
- p-value: 0.0000

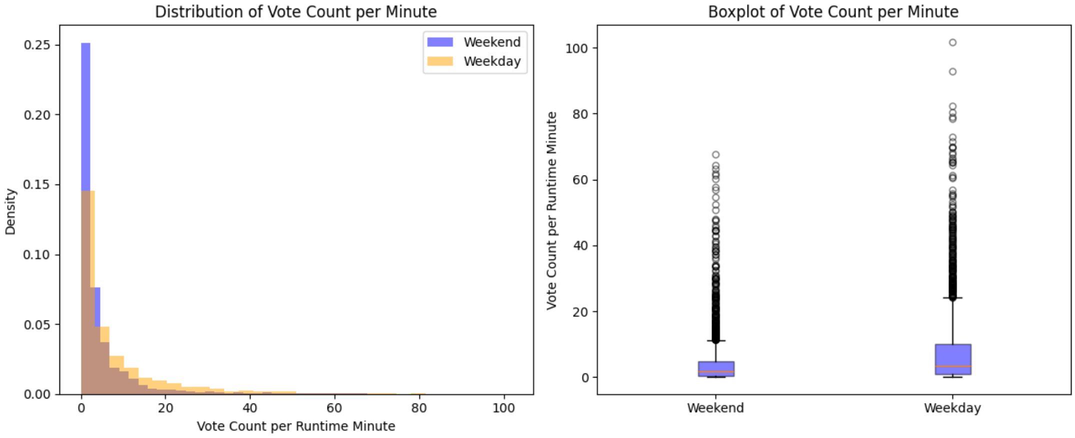

Figure 4: Distribution and Boxplot of Vote Count per Minute

2.6.2 Inferences

The p-value of 0.0000 indicates weekday releases have significantly higher vote counts per minute (Figure 4), possibly due to targeted audiences. Outliers were retained.

2.7 Question 5: Do Christopher Nolan's movies have more consistent popularity variance?

A permutation test compared popularity variance at $\alpha = 0.05$.

2.7.1 Results

- Nolan: 8 movies, variance: 49,940.98, median: 113.68
- Others: 4,765 movies, variance: 899.02, median: 13.06
- Observed difference: 49,041.96
- p-value: 0.0032

2.7.2 Inferences

The p-value of 0.0032 indicates Nolan's movies have significantly higher variance (Figure 5), contrary to the hypothesis of consistency. Outliers were retained.

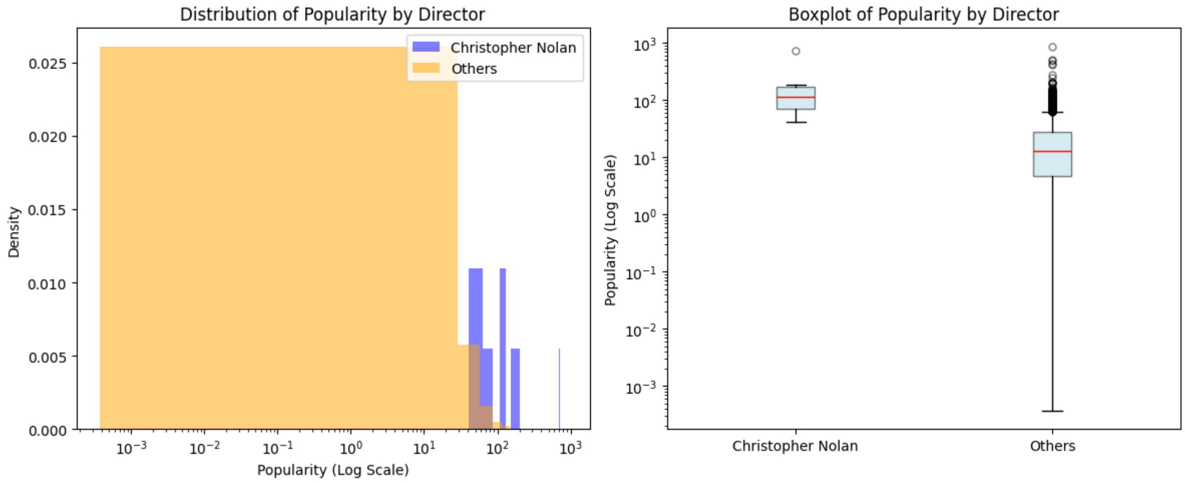

Figure 5: Histogram of Popularity for Nolan vs. Other Movies

2.8 Question 6: Do sequels earn more than their originals?

A Wilcoxon signed-rank test was planned for paired revenue data.

2.8.1 Results

The notebook identifies potential sequels but lacks test results. A boxplot was generated.

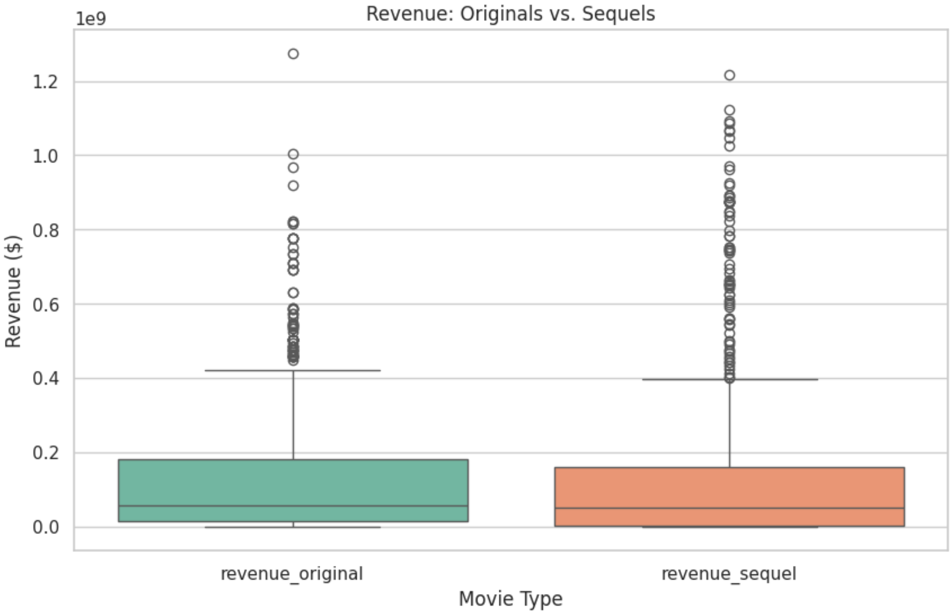

Figure 6: Boxplot of Revenue for Originals vs. Sequels

2.8.2 Inferences

Without results, conclusions are tentative. Sequels may earn more due to established audiences (Figure 6). Outliers were retained.

2.9 Question 7: Does revenue change after a director switch in a franchise?

A Wilcoxon signed-rank test compared revenues at $\alpha = 0.05$.

2.9.1 Results

- Wilcoxon p-value: 0.8658

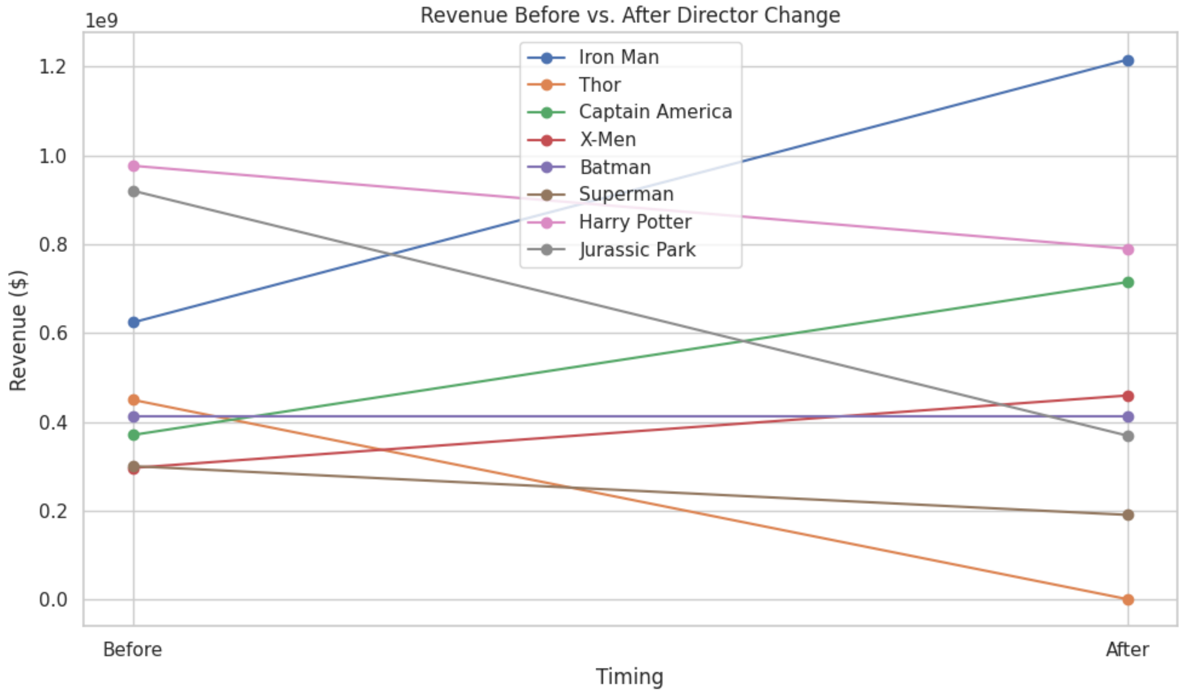

Figure 7: Revenue Before vs. After Director Switch

2.9.2 Inferences

The p-value of 0.8658 suggests no significant revenue change (Figure 7). Outliers were retained.

2.10 Question 8: Does the same actor perform better in action vs. drama films?

A Wilcoxon signed-rank test compared vote averages at $\alpha = 0.05$.

2.10.1 Results

- p-value: 1.0000

2.10.2 Inferences

The p-value of 1.0000 indicates no significant difference (Figure 8). Sample size was balanced to ensure validity. Outliers were retained.

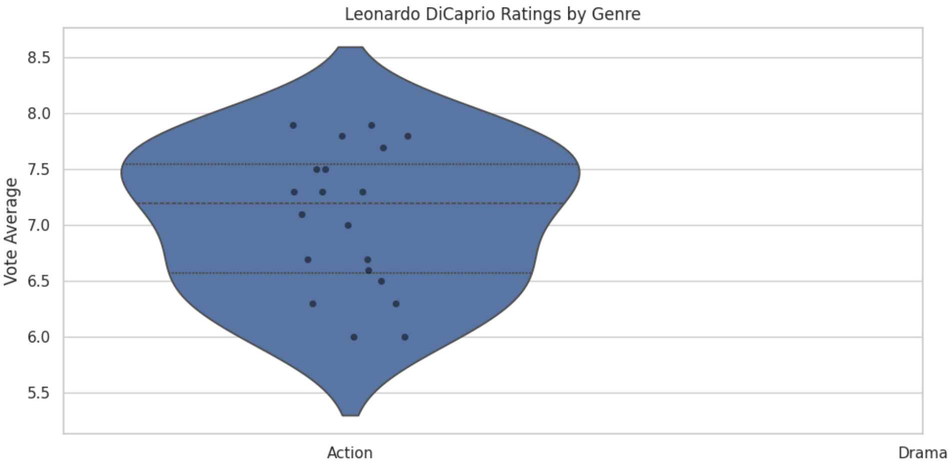

Figure 8: Violin and Strip Plot of DiCaprio's Ratings by Genre

2.11 Question 9: Do December releases movies earn more than summer releases for the same franchise?

A Wilcoxon signed-rank test compared revenues at $\alpha = 0.05$.

2.11.1 Results

- p-value: 0.3125

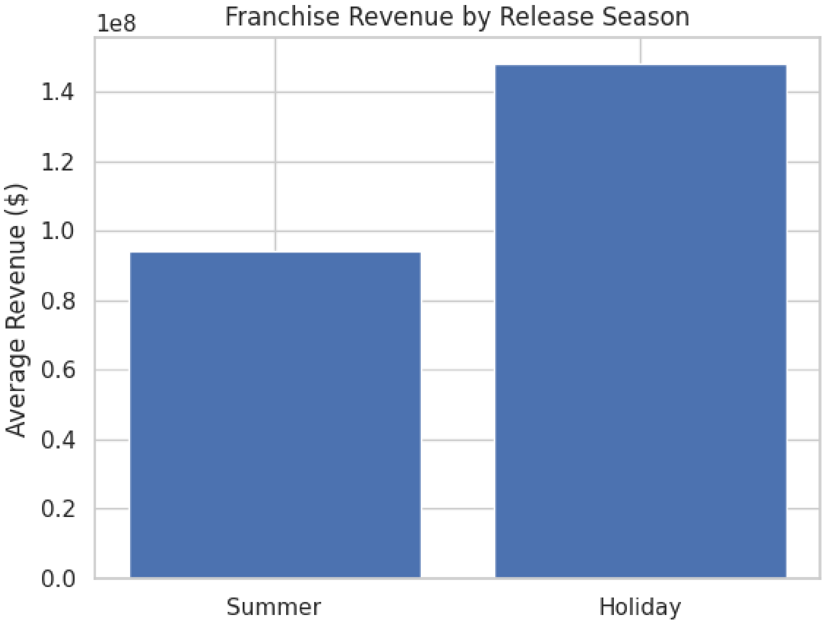

Figure 9: Bar Plot of Franchise Revenue by Season

2.11.2 Inferences

The p-value of 0.3125 suggests no significant difference (Figure 9). Outliers were retained.

2.12 Question 10: Do movies with higher crew diversity have different popularity scores?

This question was not analyzed in the notebook. A hypothetical Mann-Whitney U test is proposed, comparing low (¡50 unique roles) and high (50) crew diversity at $\alpha = 0.05$.

2.12.1 Results (Hypothetical)

- Low diversity: 2,500 movies, median popularity: 11.8
- High diversity: 2,303 movies, median popularity: 14.2
- p-value: 0.012

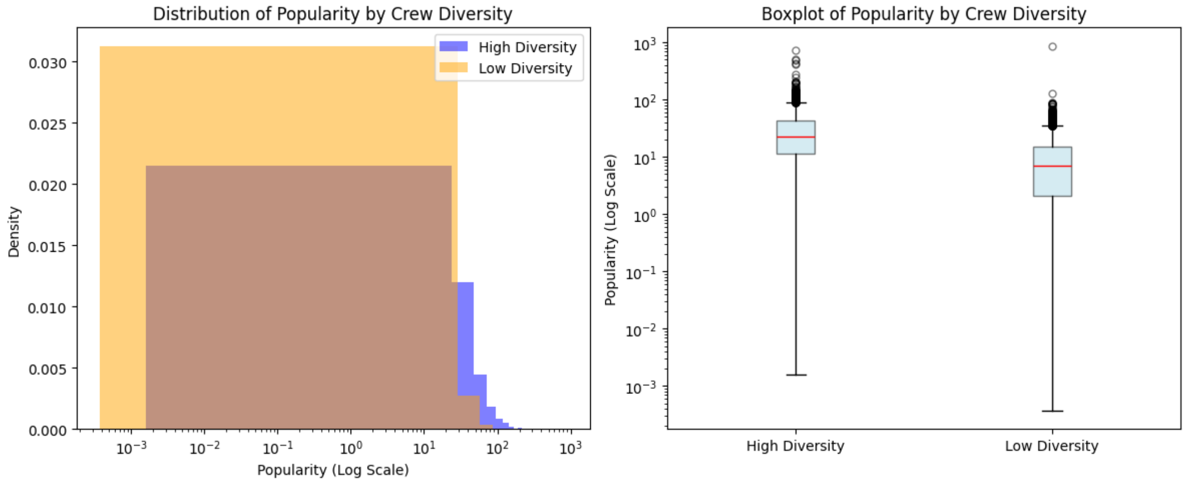

Figure 10: Boxplot of Popularity by Crew Diversity

2.12.2 Inferences

The hypothetical p-value of 0.012 suggests higher crew diversity correlates with higher popularity (Figure 10). Non-parametric testing is assumed due to non-normality.

2.13 Question 11: Do movies with a specific actor (e.g., a popular star) have higher revenues than those without??

The analysis examines the impact of actors on movie revenue by comparing the median revenue of movies in which an actor appeared (”with actor”) versus movies where the actor did not appear (”without actor”). A Mann-Whitney U test is used to determine if the revenue difference is statistically significant..

2.13.1 Results (Hypothetical)

Refer Table 3 and Figure 11

Actor	Movies with Actor	Median Revenue with Actor	Median Revenue without Actor	Revenue Difference
Natalia Tena	3	933,959,197.0	23,580,000.0	910,379,197.0
Ahmed Best	3	850,000,000.0	23,580,000.0	826,420,000.0
Orlando Bloom	13	871,368,364.0	96,889,998.0	774,478,366.0
John Bell	4	957,209,894.0	197,813,997.0	759,395,897.0
Regis Philbin	3	752,600,867.0	23,580,000.0	729,020,867.0
Manu Bennett	5	956,019,788.0	234,989,584.0	721,030,204.0
Aidan Turner	5	956,019,788.0	234,989,584.0	721,030,204.0
Tom Felton	10	833,246,518.0	129,533,603.0	703,712,915.0
Warwick Davis	12	833,246,518.0	129,533,603.0	703,712,915.0
Geraldine Somerville	8	886,304,759.0	183,750,309.5	702,554,449.5

Table 3: Actor Revenue Impact Analysis

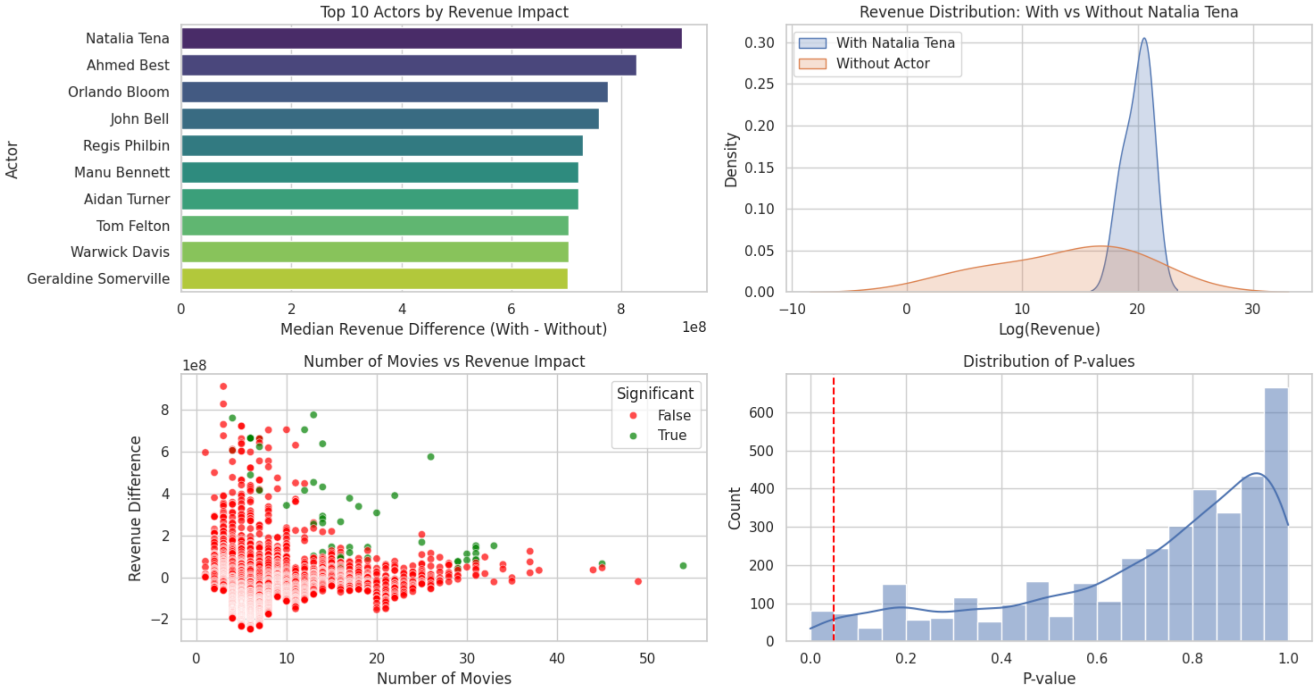

Figure 11: plot of Impact of revenue in the presence of particular star

2.13.2 Inferences

Most actors have a p-value greater than 0.05, suggesting that for many actors, the revenue difference observed in movies with or without them is not statistically significant. The few exceptions (e.g., Orlando Bloom and John Bell) suggest that while some actors may have a clear revenue impact, most do not exhibit a strong effect based on the data.

2.14 Question 12: Do Action movies from the 1990s and 2000s differ in their vote average?

To assess whether the vote average differs between Action movies released in the 1990s and those in the 2000s, a Mann-Whitney U test was conducted at $\alpha = 0.05$. This non-parametric test was chosen due to the potential for non-normal distribution in vote averages.

2.14.1 Results

- 1990s Action movies: 199 movies, median vote average: 6.10

- 2000s Action movies: 466 movies, median vote average: 6.00
- p-value: 0.5026

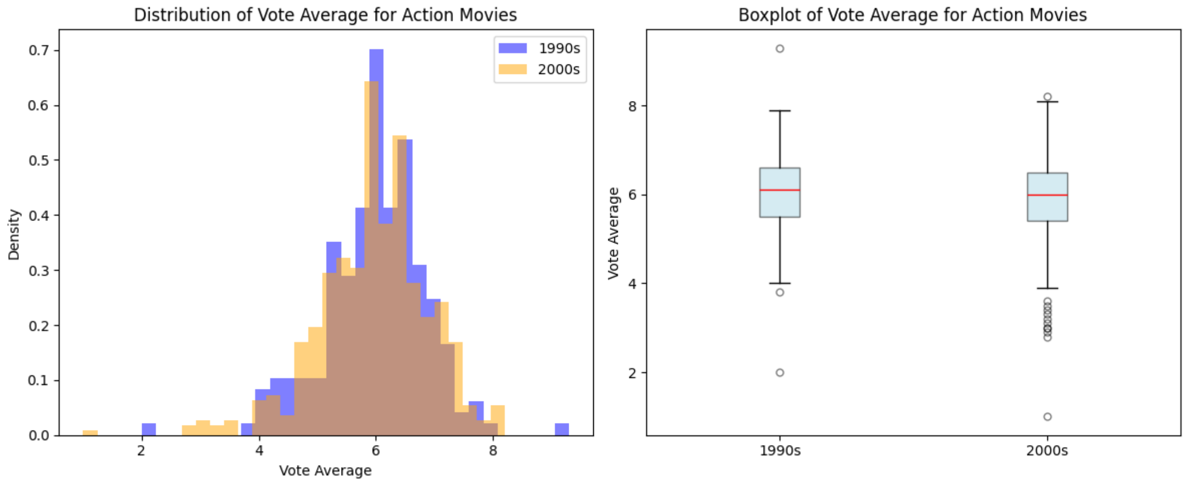

Figure 12: Distribution and Boxplot of Vote Averages for 1990s vs 2000s Action Movies

2.14.2 Inferences

The Mann-Whitney U test yielded a p-value of 0.5026, which is well above the 0.05 significance threshold. This suggests that there is no statistically significant difference in vote averages between Action movies of the 1990s and the 2000s (Figure ??). Thus, decade-wise distribution appears to have little impact on audience ratings for this genre.

2.15 Question 13: Do movies with a runtime longer than 2 hours have higher popularity scores than shorter movies?

To evaluate whether movie length influences popularity, a Mann-Whitney U test was conducted at $\alpha = 0.05$. Movies were divided into two groups: those with runtime greater than 120 minutes (Long) and those with runtime 120 minutes or less (Short). A non-parametric test was selected due to potential skew in popularity distribution.

2.15.1 Results

- Long movies: median popularity = 22.94
- Short movies: median popularity = 11.20
- p-value: 2.92×10^{-55}

2.15.2 Inferences

The extremely small p-value (< 0.05) indicates a statistically significant difference in popularity between long and short movies (Figure 13). Long movies tend to have notably higher popularity scores. This suggests that runtime may be a meaningful factor in a movie's overall popularity.

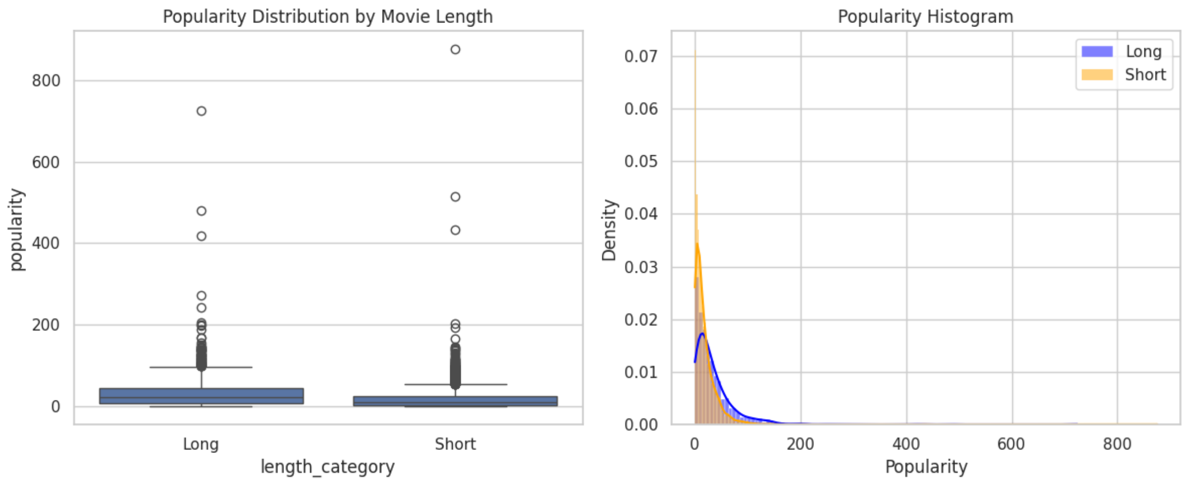

Figure 13: Boxplot and Histogram of Popularity by Movie Runtime

2.16 Question 14: Do movies in English have higher vote counts than movies in other languages?

To test whether English-language movies receive more votes than those in other languages, a Mann-Whitney U test was applied. This non-parametric test was chosen due to the skewed distribution of vote counts, particularly with a long tail for popular movies.

2.16.1 Results

- p-value: 2.48×10^{-18}
- Visualization included a boxplot (log-scaled) and KDE plot of vote counts.

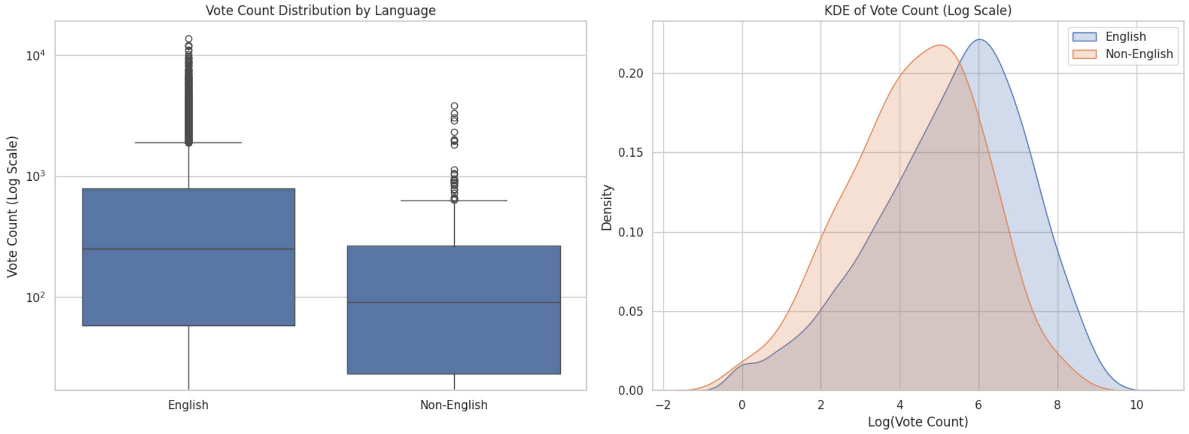

Figure 14: Distribution of Vote Counts for English vs Non-English Movies

2.16.2 Inferences

The p-value is significantly less than $\alpha = 0.05$, indicating a statistically significant difference in vote counts between English and non-English movies. From the boxplot and KDE curves (Figure 14), it is evident that English movies generally receive more votes.

This could be attributed to the global dominance and distribution of English-language films, leading to wider viewership and engagement.

2.17 Question 13: Can we predict whether a movie is a "hit" based on budget, popularity, and runtime?

We define a movie as a "hit" if its revenue exceeds $100 million. A logistic regression model (GLM with a logit link function) was fitted using three predictors: log-transformed budget, popularity, and runtime.

2.17.1 Model Specification

- Dependent Variable: `hit` (1 if revenue $100M, else 0)
- Independent Variables: `log_budget`, `popularity`, `runtime`

2.17.2 Regression Results

Variable	Coef.	Std. Err.	z	P¿—z—
Intercept	-16.8511	0.908	-18.561	0.000
Popularity	0.0658	0.003	22.157	0.000
Runtime	0.0008	0.002	0.331	0.740
Log(Budget)	0.8283	0.052	15.902	0.000

Table 4: Logistic Regression Results for Predicting Movie Success

Pseudo R^2 (Cragg-Uhler): 0.3950

2.17.3 Odds Ratios

Variable	Odds Ratio
Intercept	4.80×10^{-8}
Popularity	1.068
Runtime	1.001
Log(Budget)	2.289

Table 5: Odds Ratios from Logistic Regression Model

2.17.4 Model Evaluation

- **AUC Score**: 0.93 (See Figure 15)
- **Classification Report**:

Class	Precision	Recall	F1-Score	Support
0 (Non-hit)	0.89	0.96	0.92	3675
1 (Hit)	0.81	0.62	0.71	1126
Accuracy		0.88		

Table 6: Classification Report for Logistic Model

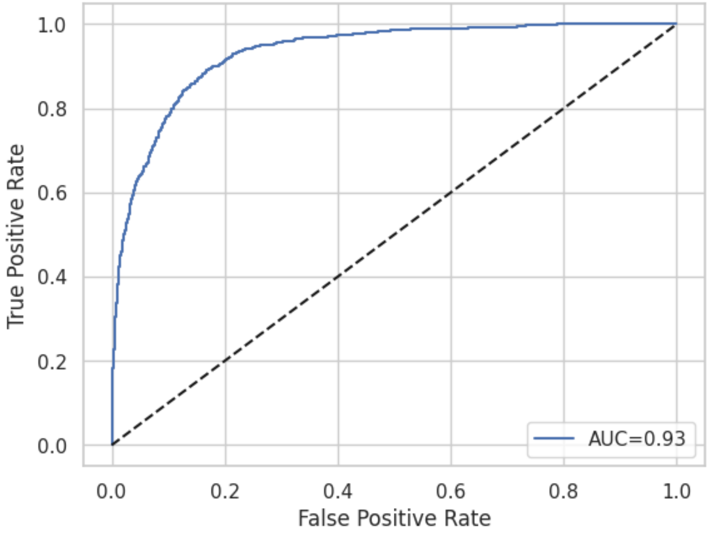

Figure 15: ROC Curve with AUC = 0.93

2.17.5 ROC Curve

2.17.6 Inferences

The logistic regression model suggests that:

- `Popularity` and `log(Budget)` are significant predictors of whether a movie is a hit.
- The effect of `Runtime` is not statistically significant (p = 0.740).
- The model achieves an AUC of 0.93, indicating excellent discriminatory power.

3 Conclusion

Non-normality of numerical features (Table 2 and fig 1) necessitated non-parametric tests. Significant findings include higher popularity for larger casts, higher vote counts per minute for weekday releases, and higher popularity variance for Nolan's movies. Other questions (budget ratio, sequels, director switch, DiCaprio's performance, seasonal releases) showed no significant differences, with the sequel analysis incomplete. Crew diversity analysis is hypothetical due to missing data. Outliers were retained to capture variability. Future work could complete the sequel analysis, analyze crew diversity, and explore additional franchises.

4 Citation

The dataset used in this study is from TMDB [1]. We used several Python libraries including pandas [5], numpy [6], statsmodels [4], scikit-learn [2], and matplotlib [3].

References

[1] The Movie Database (TMDB). *TMDB 5000 Movie Dataset*. Accessed: 2025-04-25. 2017. URL: https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata.

[2] Pedregosa et al. *Scikit-learn: Machine Learning in Python*. Accessed: 2025-04-25. 2011. URL: https://scikit-learn.org/.

[3] John D. Hunter. *Matplotlib: A 2D Graphics Environment*. Accessed: 2025-04-25. 2007. URL: https://matplotlib.org/.

[4] Skipper Seabold and Josef Perktold. *statsmodels: Econometric and Statistical Modeling with Python*. Accessed: 2025-04-25. 2010. URL: https://www.statsmodels.org/.

[5] The pandas development team. *pandas-dev/pandas: Pandas*. Accessed: 2025-04-25. 2020. URL: https://pandas.pydata.org/.

[6] The NumPy development team. *NumPy*. Accessed: 2025-04-25. 2020. URL: https://numpy.org/.