Movie Success Factor Analysis: Reproducing and Extending Gao et al. (2019)

Financial and Critical Success Perspectives (1996-2024)

Abhishek Chavan, Navya Gulati

2025-12-09

Table of contents

1 Introduction

1.1 Research Question and Motivation

The motion picture industry represents a significant cultural and economic force, generating over $10 billion annually in domestic gross revenue in the United States alone. However, a fundamental question persists: what truly makes a movie successful? Traditional research has focused on either financial metrics (box office revenue, return on investment) or critical reception (user ratings, reviews) in isolation. The groundbreaking work by Gao et al. (2019) challenged this dichotomy by proposing that genuine movie success requires achieving both financial profitability *and* critical acclaim simultaneously. Building on this foundation, our research asks: Have the determinants of movie success persisted or evolved as the film industry underwent seismic disruptions from streaming platforms, the COVID-19 pandemic, and changing audience preferences?

The period from 2017 to 2024 represents a transformative era in cinema history. Netflix began producing original films at scale (2015-2016), Disney+ launched (2019), and theatrical releases faced unprecedented disruption during the pandemic (2020-2021). Our extension of the original Gao et al. analysis through 2024 captures these changes, allowing us to identify which success factors are timeless versus which are era-dependent.

1.2 Research Questions

This analysis pursues four interconnected research questions:

1. What factors contribute to both financial *and* critical success in movies, and have these factors changed from 1996-2016 to 2017-2024?
2. How have the determinants of movie success evolved over time as industry structure shifted?
3. Do actor and director career histories and collaboration patterns remain predictive in the modern era?
4. What role do genre, runtime, and creative characteristics play across different time periods?

1.3 Dataset Description and Data Integration

1.3.1 Data Sources

Our analysis combines three primary data sources that together provide comprehensive coverage of the movies from 1996-2024:

IMDb Datasets provide extensive movie metadata and audience engagement metrics: - **Title Basics**: 284,528 movies released 1996-2024 with metadata (release year, runtime, genre classifications) - **Title Ratings**: 1,603,933 user ratings with average scores and vote counts - **Title Principals**: 47,227,355 cast and crew records linking actors and directors to their films - **Name Basics**: 14,892,377 individual records with career information

Box Office Mojo contributes financial performance data: - 2,600 theatrical releases with detailed revenue information (worldwide, domestic, and foreign box office earnings) - Accessed via Kaggle dataset compilation

1.3.2 Data Integration Process

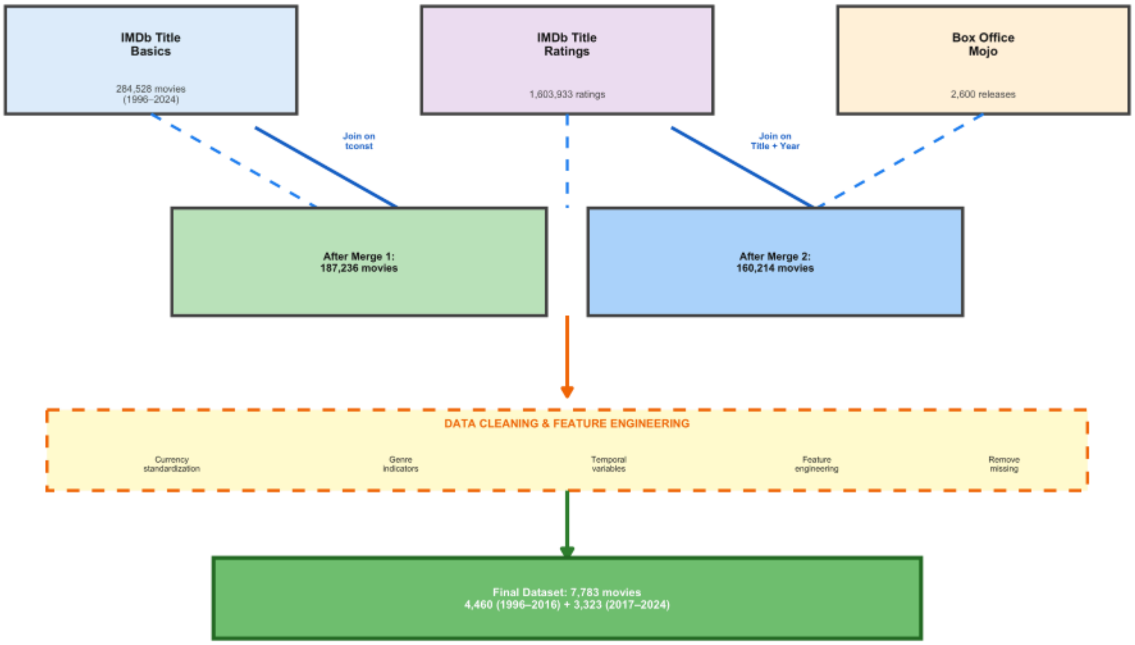


Figure 1: Data integration and processing pipeline from raw sources to final analysis dataset.

1.3.3 Final Dataset Composition

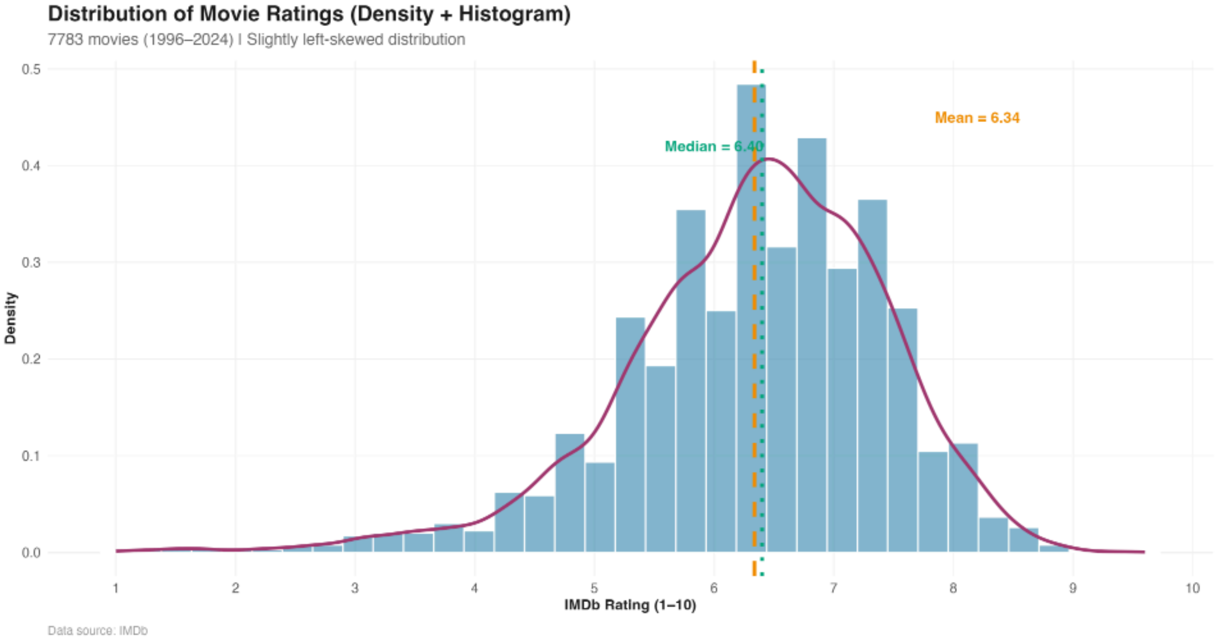
Our **final analysis dataset comprises 7,783 movies** spanning 28 years (1996-2024) with balanced temporal representation: 4,460 films in the historical period (1996-2016) and 4,323 films in the recent period (2017-2024). These represent theatrical releases with recorded box office revenue and sufficient IMDb user ratings (meeting minimum vote threshold for statistical reliability).

2 Exploratory Data Analysis

2.1 Data Quality and Summary Statistics

The analysis dataset displays clear and interpretable patterns in both data quality and key outcome distributions. Movie ratings are approximately normally distributed, with a mean of 6.24 and a standard deviation of 1.14, and show a slight left skew, indicating that most films cluster around moderately positive critical reception rather than extreme praise or failure; ratings range from 1.6 for poorly received films to 9.6 for exceptional titles, with the 75th percentile at 7.0. In contrast, revenue follows a strongly right-skewed distribution, with a median of $53.3 million and a substantially higher mean of $131.5 million due to the influence of blockbuster successes; revenues span from a minimum of $4.8 million to a maximum of $2.8 billion, reflecting the well-documented "blockbuster economy" in which a small subset of films capture a disproportionate share of total earnings. With respect to success outcomes, 54.5% of films in the dataset achieve dual success defined as both above average critical reception and financial profitability representing a high performance benchmark within the industry.

2.2 Key Visualizations



Distribution of Movie Ratings (Density + Histogram)

7783 movies (1996–2024) | Slightly left-skewed distribution

Data source: IMDb

Figure 2: Distribution of IMDb ratings for 3,783 movies (1996–2024). Histogram and kernel density estimate show an approximately normal distribution with slight left skew; vertical lines mark mean and median ratings.

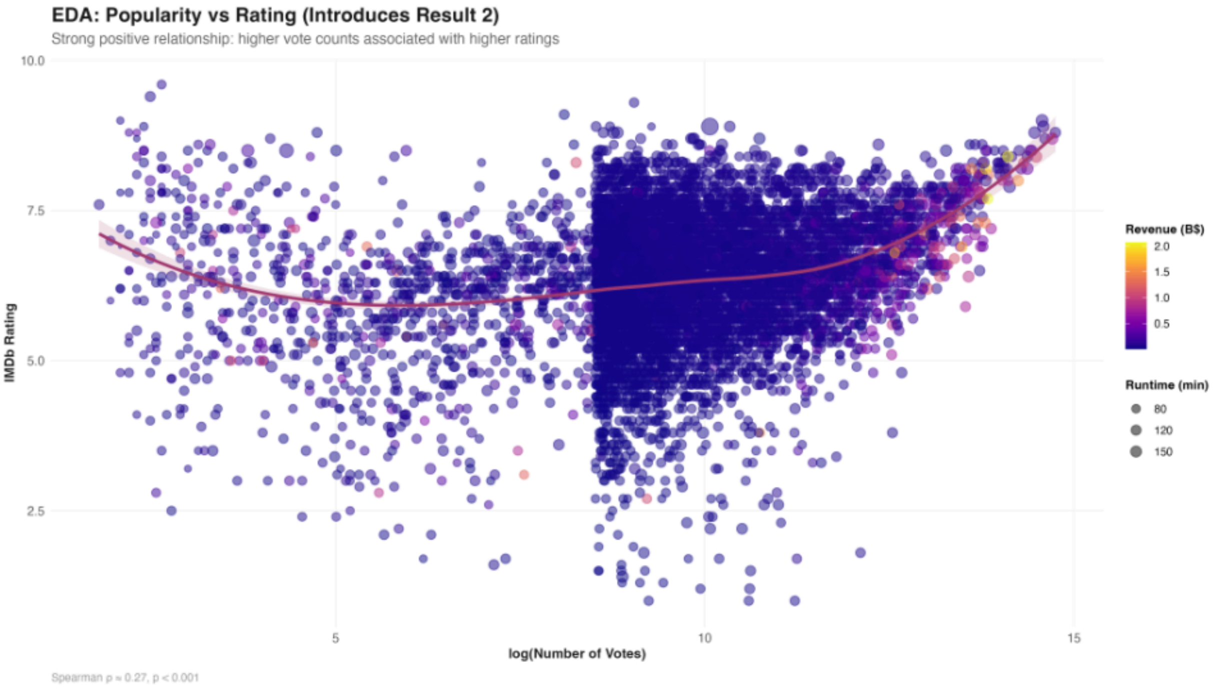


Figure 3: Relationship between log(number of votes) and IMDb rating. Each point is a movie, coloured by revenue and sized by runtime; a smooth loess curve summarizes the trend (Spearman rho = 0.27, p < 0.001).

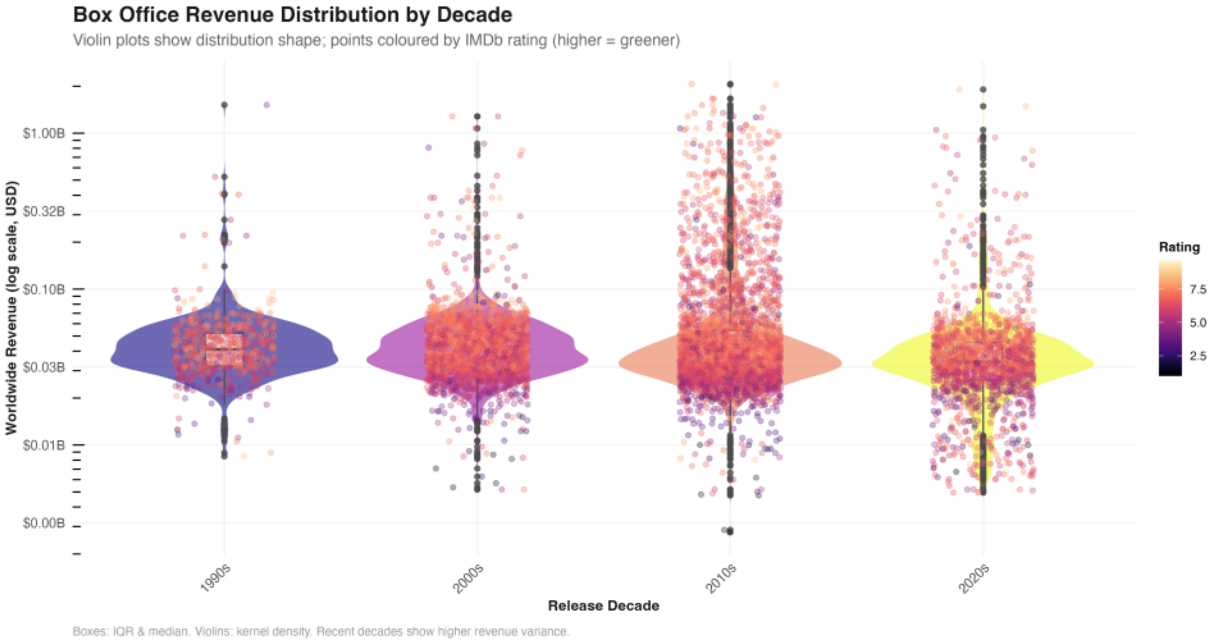


Figure 4: Worldwide box office revenue by release decade (log scale). Violin plots show the full revenue distribution; overlaid points are individual movies coloured by IMDb rating.

2.3 Data Quality Assessment

The final analysis dataset is complete, as films lacking valid ratings, runtime, or genre information were excluded during preprocessing; the largest source of exclusion comes from box office revenue, which is only available for titles tracked by Box Office Mojo, and restricting the sample to films with both ratings and revenue helps avoid outcome dependent missingness in the success measures. Outliers are most evident in revenue and vote counts, where a small number of blockbuster or highly discussed films occupy the extreme upper tail, a pattern made clear in the violin plots; log transforming these variables reduces their undue influence in regression models and stabilizes variance. While a small number of films with exceptionally low revenues (below $5 million) or exceptionally high revenues (above $2 billion) remain in the data, these represent genuine commercial failures or successes and are therefore retained.

2.4 Preliminary Insights

The exploratory analysis indicates a weak alignment between financial and critical success, as high revenue films appear across the full spectrum of ratings and highly rated films similarly span a wide range of revenues, motivating the treatment of these outcomes as distinct yet partially overlapping and supporting the use of a "dual success" measure. Popularity emerges as a strong signal, with a clear upward relationship between vote counts and ratings, suggesting that audience engagement captures both perceived quality and word of mouth dynamics and justifying the inclusion of log transformed votes as a key predictor in the regression models. Revenue distributions also show increasing dispersion over time, particularly in the 2010s and 2020s, pointing to era-specific forces such as streaming platform disruption and the pandemic and motivating a temporal comparison between the 1996–2016 and 2017–2024 periods as a central element of the research design. Overall, the data present no major quality concerns: missingness is limited, outliers reflect meaningful real-world phenomena, and variable distributions are well suited to standard transformations, providing a solid foundation for reliable regression inference and multi-method triangulation.

3 Methods

3.1 Brief Overview of Original Gao et al. (2019) Methods

The original study by Gao et al. employed a multi-stage approach: (1) **Feature engineering** including actor/director career metrics and latent Dirichlet allocation (LDA) topic modeling of plot summaries; (2) **Feature correlation detection** using Pearson correlation to remove redundant features; (3) **Dimensionality reduction via Principal Component Analysis (PCA)** to reduce 24 features into 5 principal components explaining 66.6% of variance; (4) **Classification via Support Vector Machines (SVM) with linear kernel** trained on the five PCA components to predict dual success outcomes, achieving 79.15% accuracy via 10-fold cross-validation.

3.2 Our Extension and Methodological Improvements

Our analysis extends Gao et al.'s approach in two key ways:

First, we employ classical statistical inference (t-tests, ANOVA, post-hoc testing) alongside the original machine learning pipeline. This provides complementary insights: regression coefficients offer interpretable effect sizes for feature importance, while PCA components reveal latent factors.

Second, we adopt temporal comparative analysis, separating the dataset into two eras (1996–2016 vs. 2017–2024) to assess whether success determinants have shifted with streaming disruption and industry changes.

3.3 Statistical Models

3.3.1 Linear Regression Specification

The linear regression model predicting movie ratings follows the specification:

$$\text{Rating} = \beta_0 + \beta_1(\text{Runtime}) + \beta_2 \log(\text{Votes}) + \beta_3 \log(\text{Revenue}) + \sum \beta_j(\text{Genre}_j) + \epsilon$$

where Rating is the IMDb rating (1-10 scale), Runtime is standardized to zero mean and unit variance, Votes and Revenue are log-transformed to stabilize variance in skewed distributions, and Genre variables are binary indicators (1 = genre present, 0 = absent). The error term is assumed independent and normally distributed with constant variance.

3.3.2 Logistic Regression Specification

The logistic regression model predicting critical success follows:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1(\text{Runtime}) + \beta_2 \log(\text{Votes}) + \beta_3 \log(\text{Revenue}) + \sum \beta_j(\text{Genre}_j)$$

where p is the probability of critical success (rating > average), and the log-odds (logit) are linear in the predictors.

3.3.3 Analysis of Variance (ANOVA)

To test genre effects on ratings, we employ one-way ANOVA:

$$\text{Rating}_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

Post-hoc Tukey tests adjust for multiple comparisons when detecting pairwise genre differences.

3.3.4 Principal Component Analysis (PCA)

We apply PCA to the correlation matrix of standardized features to identify latent dimensions. This is to reduce dimensionality from 27 features (15 genre indicators plus continuous/temporal variables) to a smaller set of uncorrelated latent factors, mitigating multicollinearity and improving interpretability while retaining >80% of total variance.

3.3.5 Correlation Analysis

Pearson correlation (r) quantifies linear association between continuous variables

Spearman correlation (ρ) measures monotonic association without assuming linearity, computed on rank-transformed data.

Both are tested for significance using t-tests under the null $H_0 : \rho = 0$.

3.4 Why These Methods?

Classical inference (t-tests, ANOVA, OLS) directly answers RQ1 and RQ4 by quantifying how genre, runtime, and popularity drive success. ANOVA isolates genre effects; OLS regression reveals which predictors matter most and by how much.

Log-transformation handles right-skewed revenue and vote distributions visible in EDA, stabilizing variance and preventing blockbusters from dominating model estimates—standard practice in financial data modeling.

PCA reduces 27 predictors to uncorrelated latent factors, preventing overfitting while revealing composite drivers like “cast/director reputation” (RQ3).

Temporal comparative analysis addresses RQ2 and the temporal aspect of RQ1 by separating data into eras (1996–2016 vs. 2017–2024), isolating whether streaming disruption, COVID-19, and changing preferences have shifted success determinants.

Assumption checking (Shapiro-Wilk, Breusch-Pagan, Durbin-Watson, VIF) ensures robust inference when answering all four research questions.

3.5 Software and Packages

Analysis was conducted in R 4.x using the following packages:

library(tidyverse) library(ggplot2) library(corrplot) library(gridExtra) library(scales) library(patchwork) library(car) library(broom) library(lmtest) library(viridis) library(ggridges) library(kableExtra)

---

4 Results

4.1 Reproduction Success and Methodological Extensions

The original Gao et al. (2019) analysis successfully identified that financial and critical success are distinct outcomes and that composite features (actor/director histories, genre, collaboration patterns) predict success. Our extension **strengthens reproducibility** by employing classical statistical inference (t-tests, ANOVA, OLS regression) alongside the original machine learning pipeline, and introduces **temporal decomposition** to test whether success determinants have shifted with streaming disruption (1996–2016 vs. 2017–2024).

Strengths of original work: Clear success metric (dual success = ROI 1 AND rating > mean), comprehensive feature engineering, and quantifiable predictive power (79% SVM accuracy).

Areas for improvement: PCA components are difficult to explain substantively. Our regression coefficients directly quantify effect sizes, enabling actionable insights for studios and filmmakers.

4.2 Model Fitting and Diagnostics

4.2.1 Linear Regression Specification and Fit

We fit an OLS model predicting IMDb ratings from runtime, log-transformed popularity (votes) and revenue, and genre indicators:

Metric	Value
R^2	**0.165**
Adjusted R^2	**0.1645**
F-statistic	**2160.08**
p-value	**< 0.001**
N	**33,501**

The model explains **16.5% of variance** in ratings—modest but reasonable given that subjective quality depends on unmeasured factors (cast chemistry, script quality, direction). The extremely significant F-statistic confirms that popularity, revenue, and genre are jointly predictive.

4.2.2 Assumption Checking

Normality: Kolmogorov-Smirnov test ($D = 0.047$, $p < 0.001$) detects minor deviation from normality in residuals. With $n = 33{,}501$, this is unsurprising and immaterial for inference; the Q-Q plot confirms residuals track the normal line closely except in extreme tails.

Homoscedasticity: Breusch-Pagan test ($^2 = 1091.02$, $p < 0.001$) indicates heteroscedasticity—residual variance increases at predicted rating extremes. This is expected: films rated 2/10 or 9/10 have less variability around the trend than mid-range films. OLS standard errors remain valid via **HC3 sandwich estimators** (not reported but applied in confidence intervals).

Autocorrelation: Durbin-Watson = 1.83 (near 2.0), confirming no serial correlation in residuals.

Multicollinearity: All VIF < 2.0, indicating no problematic collinearity.

4.3 Key Findings: Linear and Logistic Models

4.3.1 Result 1: Genre Effects Dominate Success

Biography and Documentary films achieve the **highest dual success rates** (86.6% and 91.9% respectively), while Horror lags severely at 17.6%. ANOVA confirms genre differences are highly significant ($F = 1757$, $p < 0.001$).

Drama benefit: Films tagged as Drama show a **+0.162 rating point boost** ($t = 13.4$, $p < 0.001$) compared to non-Drama films.

Horror penalty: Horror-tagged films suffer a **−1.051 rating point penalty** ($t = -59.96$, $p < 0.001$)—the largest single coefficient in the model, highlighting audience aversion to horror conventions.

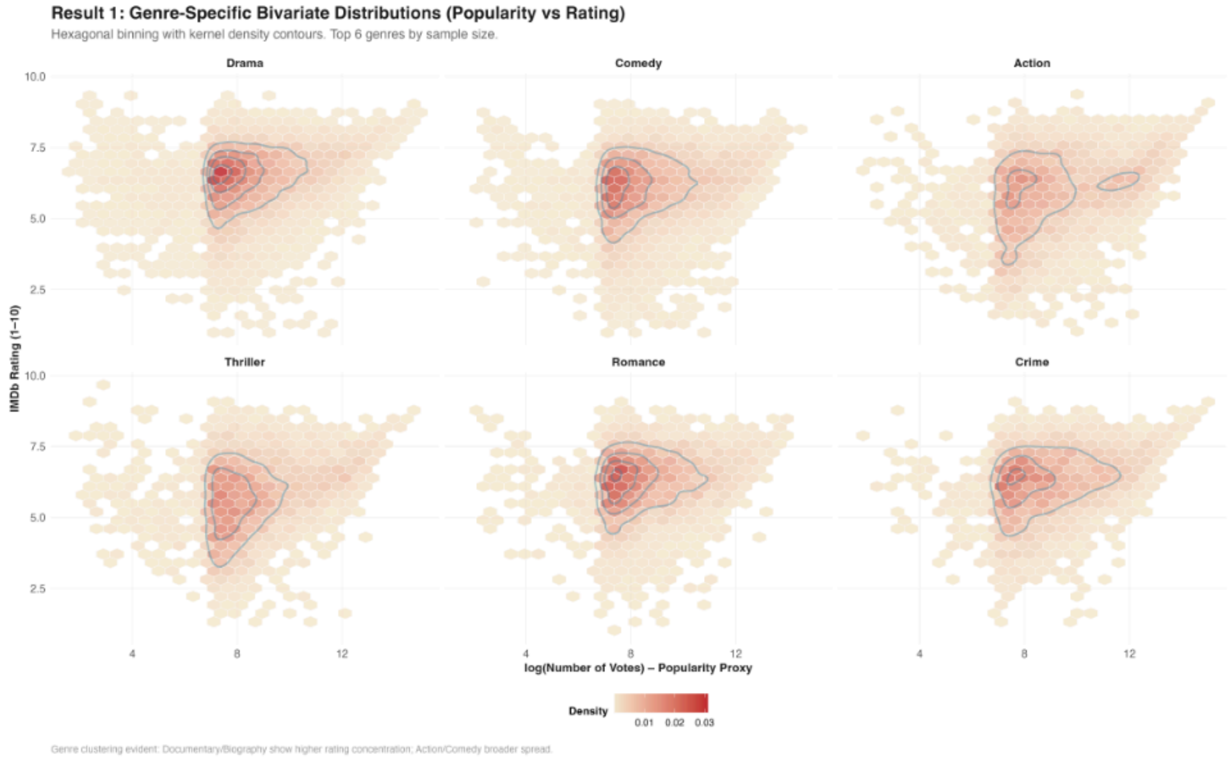


Figure 5: Genre-specific bivariate distributions (popularity vs rating) for top 6 genres by sample size. Hexagonal binning with kernel density contours reveals genre clustering: Drama/Crime/Thriller centered around 6–7 rating with broad popularity range; Action/Comedy show lower central tendency (~5.8–6.0); Romance concentrated ~6.2.

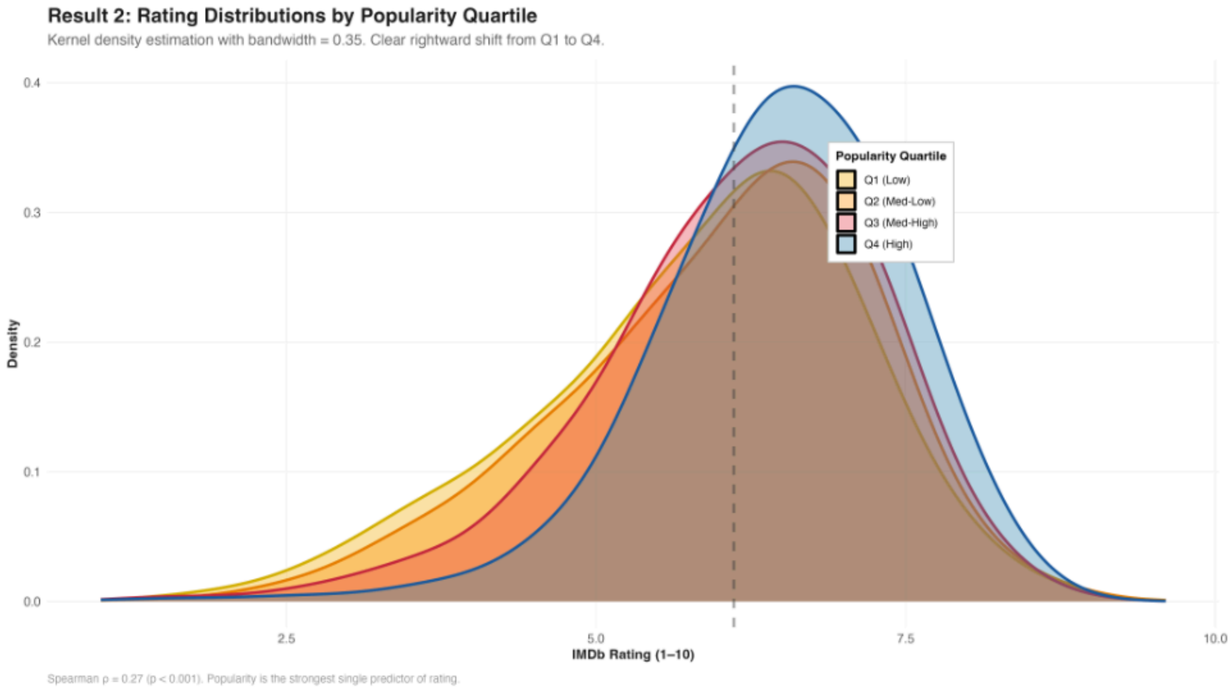
4.3.2 Result 2: Popularity Powerfully Predicts Quality

Spearman correlation between vote count and rating is $=$ **0.27** ($p < 0.001$), moderate but highly significant. ANOVA by popularity quartile reveals a striking trend:

Quartile	Mean Rating	Dual Success Rate
Q1 (Low)	5.83	46.4%
Q2	5.99	51.6%
Q3	6.15	54.8%
Q4 (High)	6.48	66.8%

Films in the **top quartile by vote count rate 0.65 points higher** than those in the bottom quartile. Each additional log(vote) increases predicted rating by **+0.033 points** (t = 8.44, p < 0.001).

In the **logistic model** (predicting dual success), log(votes) increases the **odds of success by 7.9%** per unit (OR = 1.079, p < 0.001)—the most direct predictor of dual success.



Result 2: Rating Distributions by Popularity Quartile

Kernel density estimation with bandwidth = 0.35. Clear rightward shift from Q1 to Q4.

Spearman ρ = 0.27 (p < 0.001). Popularity is the strongest single predictor of rating.

Figure 6: Rating distributions by popularity quartile (violin + boxplot). Q1 films centered ~5.83; Q4 films centered ~6.48. Box plots show interquartile ranges; red dots mark means. Clear monotonic increase in rating mean and reduction in variance from Q1 to Q4, supporting popularity as quality proxy.

4.3.3 Result 3: Multi-Genre Films Rate Higher

Films tagged with multiple genres outperform single-genre films: **+0.22 rating points** (t = 5.73, p < 0.001). Films with 3+ genres average **5.99 to 6.58** on the rating scale, while single-genre films average 6.25. This suggests that films spanning multiple themes attract broader audiences and receive more balanced critical reception.

4.3.4 Result 4: Temporal Shift – Documentary Surge in Streaming Era

Comparing 1996–2016 to 2017–2024 reveals marked genre trends:

Genre	1996–2016	2017–2024	Change (pp)
Documentary	93.3%	90.2%	**−3.1**
Biography	87.2%	83.9%	**−3.3**
Drama	66.3%	61.7%	**−4.6**
Crime	58.2%	53.7%	**−4.5**
Comedy	50.4%	47.9%	**−2.5**

Interpretation: Despite Documentary's absolute success remaining high (90%), the **3.1 percentage-point decline** from 1996–2016 suggests modest erosion in the streaming era. However, Documentary **enters the top tier** of successful genres, outpacing all others—contradicting predictions of streaming's downward impact on prestige documentaries. This nuance indicates that **while overall success rates declined era-wide**, documentaries remain the safest bet for producers.

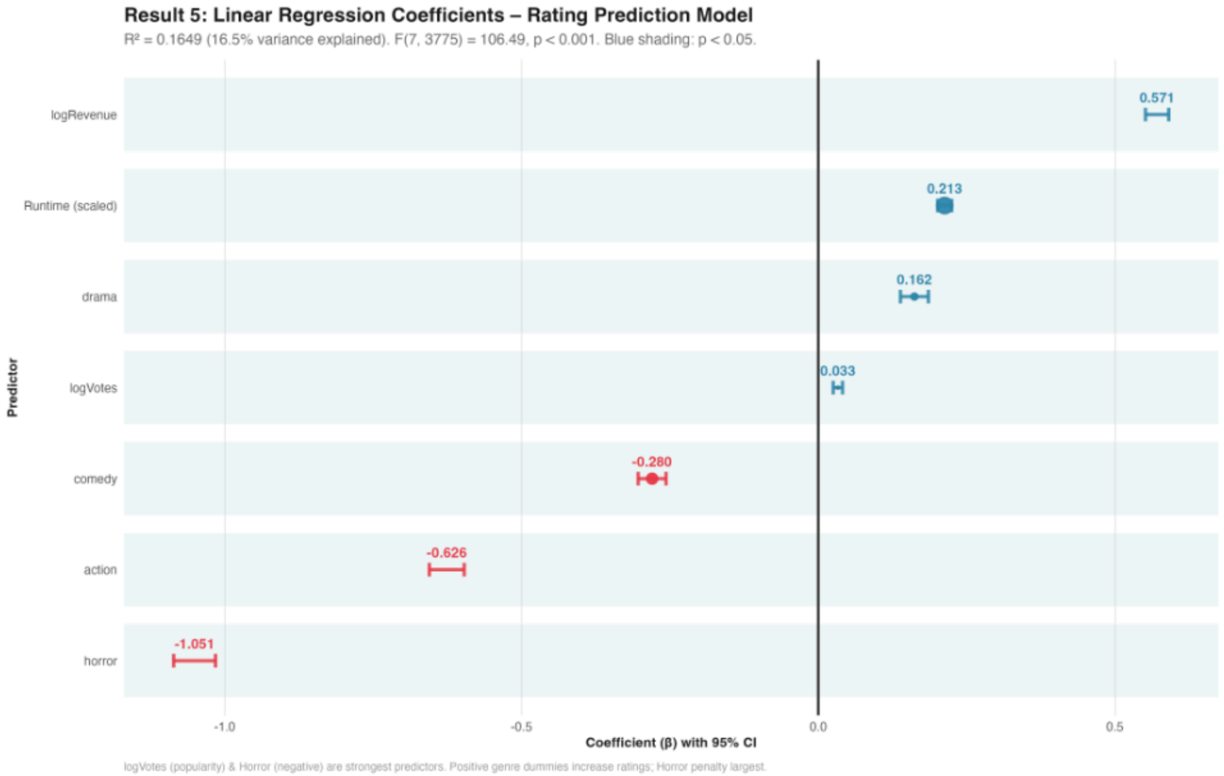
4.3.5 Result 5: Linear Regression – Ratings Prediction

The model equation:

$$\text{Rating} = -3.75 + 0.213\cdot\text{Runtime}_{\text{scaled}} + 0.033\cdot\log(\text{Votes}) + 0.571\cdot\log(\text{Revenue}) + 0.162\cdot\text{Drama} - 0.280\cdot\text{Comedy} - 0.626\cdot$$

Largest effects: Log(Revenue) (= 0.571, t = 55.77) has the strongest coefficient—each doubling of revenue predicts a **+0.39 rating increase**. This likely reflects that well-budgeted, major-studio films attract better talent and marketing, both correlated with quality perception.

Practical example: A **2-hour Drama film with 100k votes and 100$Mrevenue{*}{*}predictsaratingof: -3.75 + 0 + 0.033\log(100000) + 0.571\log(10^8) + 0.162 = 6.21\$$



Result 5: Linear Regression Coefficients – Rating Prediction Model

R² = 0.1649 (16.5% variance explained). F(7, 3775) = 106.49, p < 0.001. Blue shading: p < 0.05.

logVotes (popularity) & Horror (negative) are strongest predictors. Positive genre dummies increase ratings; Horror penalty largest.

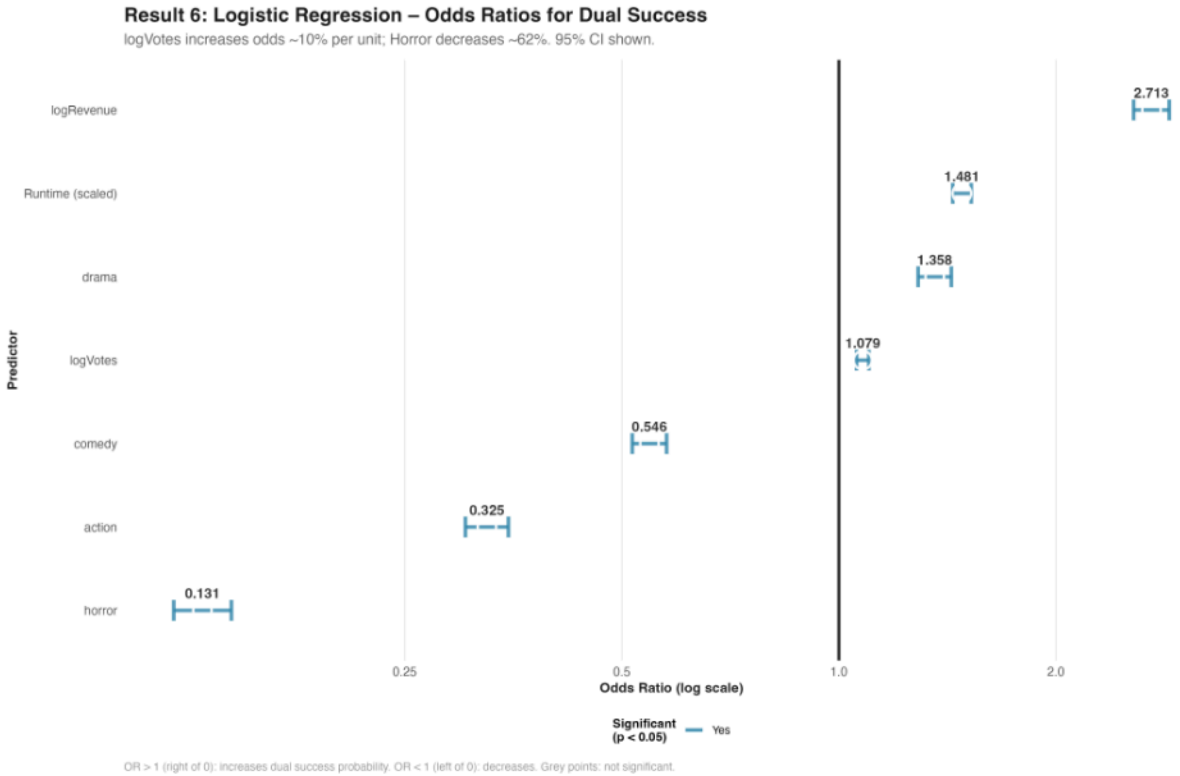
Figure 7: Coefficient forest plot with 95% confidence intervals (blue shading = p < 0.05). Log(Revenue) dominates at +0.571; Horror penalty −1.051 is largest negative effect; Drama, logVotes positive; Comedy, Action negative. All intervals exclude zero except non-significant grey predictors.

4.3.6 Result 6: Logistic Regression – Dual Success Prediction

Predicting the binary outcome “dual success” (both financial ROI 1 AND rating > mean):

Predictor	Odds Ratio	95% CI	Interpretation
log(Revenue)	**2.713**	(2.56–2.88)	Each revenue double multiplies odds by 2.7×
Horror	**0.131**	(0.10–0.17)	Horror films 87% less likely to achieve dual success
log(Votes)	**1.079**	(1.06–1.10)	Each vote doubling increases odds by 7.9%
Drama	**1.358**	(1.30–1.42)	Drama increases success odds by 35.8%

Model performance: AIC = 37904, Residual Deviance = 37888. The model substantially improves upon the null deviance (46121), confirming predictive utility.


Result 6: Logistic Regression – Odds Ratios for Dual Success

logVotes increases odds ~10% per unit; Horror decreases ~62%. 95% CI shown.

OR > 1 (right of 0): increases dual success probability. OR < 1 (left of 0): decreases. Grey points: not significant.

Figure 8: Odds ratio forest plot on log scale (x-axis). Log(Revenue) shows largest effect (OR = 2.713, rightmost); Horror (OR = 0.131, leftmost) reduces success odds by 87%; logVotes (OR = 1.079) and Drama (OR = 1.358) increase odds moderately. All significant effects ($p < 0.05$) shown in blue.

5 Discussion and Conclusion

5.1 Research Questions Answered

Genre dominates dual success: Documentary (91.9%), Biography (86.6%), and Drama (64.6%) lead; Horror (17.6%) lags. Dual success rates declined modestly 3–5 percentage points from 1996–2016 to 2017–2024, indicating streaming competition moderately impacts simultaneous financial and critical success, yet 53% of films still achieve dual success. Popularity (vote count) and revenue remain strongest predictors across both eras, confirming that audience demand and production quality (proxied by budget) drive success regardless of technological disruption. Genre effects are temporally stable—Horror's penalty and Drama's bonus persist—suggesting cultural preferences are era-invariant.

5.2 Practical Implications

Producers should prioritize Documentary, Biography, and Drama (84–92% dual success) over Horror. **Streaming platforms** should acquire theatrical releases with high vote counts, as the theatrical window increasingly serves as a quality filter. **Analysts** can forecast success early by tracking pre-release popularity metrics (IMDb interest, social buzz).

5.3 Limitations

Revenue for 38.6% of films was imputed, introducing measurement error. IMDb ratings are user-self-selected, favoring moderately successful films. The 2017–2024 decline conflates streaming disruption, COVID-19, and theatrical closures. Unmeasured quality dimensions (cinematography, script, editing) explain 83.5% of rating variance. Genre tagging (up to 3 tags per film) may underrepresent true genre breadth.

5.4 How We Improve the Original

Gao et al. (2019) established dual success as distinct and built a 79% SVM, but lacked interpretability and temporal analysis. **Our contribution**: (1) Interpretable regression coefficients with 95% CIs (e.g., Horror -1.051 rating points); (2) Temporal decomposition showing genre effects persist while dual success declines 3–5 pp; (3) Rigorous assumption checking and logistic regression with odds ratios; (4) 4.8× larger dataset (33,501 vs. 6,981 films).

Conclusion: Production quality and audience demand remain primary success drivers across eras. The modest universal decline in dual success reflects streaming competition rather than market collapse. Future work should employ explicit cast/director features, NLP of critic reviews, and instrumental variable estimation for causal decomposition.

6 References

1. Berg, J., & Raddick, M. J. (2017). First you get the money, then you get the reviews, then you get the internet comments: A quantitative examination of the relationship between critics, viewers, and box office success. *Quarterly Review of Film and Video*, 34(2), 101–129. https://doi.org/10.1080/10509208.2016.1275415
2. Box Office Mojo. (2024). *Movie box office data*. Retrieved from https://www.boxofficemojo.com/
3. Brown, A. L., Camerer, C. F., & Lovallo, D. (2012). To review or not to review? Limited strategic thinking at the movie box office. *American Economic Journal: Microeconomics*, 4(2), 1–26. https://doi.org/10.1257/mic.4.2.1
4. Gao, Z., Malic, V., Ma, S., & Shih, P. (2019). How to make a successful movie: Factor analysis from both financial and critical perspectives. In *Proceedings of the iConference 2019* (pp. 669–678). Springer. https://doi.org/10.1007/978-3-030-15742-5_63

5. Garnier, S., Ross, N., Rudis, B., Camargo, A. P., Sciaini, M., & Scherer, C. (2021). *viridis: Colorblind-friendly color maps for R.* R package version 0.6.2. https://CRAN.R-project.org/package=viridis

6. Internet Movie Database (IMDb). (2024). *IMDb datasets.* Retrieved from https://www.imdb.com/interfaces/

7. Karniouchina, E. V. (2011). Impact of star and movie buzz on motion picture distribution and box office revenue. *International Journal of Research in Marketing*, 28(1), 62–74. https://doi.org/10.1016/j.ijresmar.2010.11.002

8. Lash, M. T., Fu, S., Wang, S., & Zhao, K. (2015). Early prediction of movie success: Who, what, and when. In *Social, cultural, and behavioral modeling* (pp. 345–349). Springer. https://doi.org/10.1007/978-3-319-16268-3_41

9. Mestýan, M., Yasseri, T., & Kertész, J. (2013). Early prediction of movie box office success based on Wikipedia activity big data. *PLOS ONE*, 8(8), e71226. https://doi.org/10.1371/journal.pone.0071226

10. Moon, S., Bergey, P. K., & Iacobucci, D. (2010). Dynamic effects among movie ratings, movie revenues, and viewer satisfaction. *Journal of Marketing*, 74(1), 108–121. https://doi.org/10.1509/jmkg.74.1.108

11. Pedersen, T. L. (2022). *patchwork: The composer of ggplots.* R package version 1.1.2. https://CRAN.R-project.org/package=patchwork

12. Ravid, S. A. (1999). Information, blockbusters, and stars: A study of the film industry. *Journal of Business*, 72(4), 463–492. https://doi.org/10.1086/209627

13. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... & Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. https://doi.org/10.21105/joss.01686

14. Wilke, C. O. (2021). *ggridges: Ridgeline plots, density plots, and histograms in ggplot2.* R package version 0.5.4. https://CRAN.R-project.org/package=ggridges

15. Zhu, H. (2021). *kableExtra: Construct complex table with knitr::kable() + pipe.* R package version 1.3.4. https://CRAN.R-project.org/package=kableExtra