Cinematic ROI and Audience Dynamics
Financial Efficiency, Genre Fatigue, and Global Film Trends

Atlanta Daniel

5.2 Project

Comprehensive Final Report and Presentation

Dataset Selection & Problem Definition

This project analyzes the Movies Metadata Cleaned Dataset (1900–2025), a large-scale collection of over 946,000 film records sourced from The Movie Database (TMDB) and made available on Kaggle. The dataset captures a broad range of production and audience variables including budget and revenue, genre flags, origin country, spoken languages, runtime, audience popularity scores, vote averages, and vote counts. Inflation adjustment was performed using annual Consumer Price Index (CPI) data from the US Bureau of Labor Statistics (1997–2025) and historical CPI from the Federal Reserve Bank of Minneapolis (pre-1997), converting all monetary figures to 2025 dollars to enable fair cross-decade comparisons.

Five research questions guided the analysis:

1. Once we adjust for inflation, do small indie films, mid-sized movies, or massive blockbusters make the most profit relative to their cost? Is there a limit to how much money a studio should spend?
2. How do things like a movie's genre, length, or language affect how many people watch it and how much they like it? Do big-name studios and high budgets automatically guarantee a movie will be popular?
3. How have things like runtime and profits changed across the decades and between different countries?
4. Does genre fatigue exist? Does releasing too many movies of the same type cause ratings to crash?
5. Scenario Analysis: Under three distinct real-world filmmaking conditions, how do predicted ROI outcomes shift relative to a mid-tier baseline?

The goal is to identify budget sweet spots, quantify the risk of genre saturation, map long-term shifts in global film profitability, and translate model outputs into concrete strategic guidance for producers, distributors, and streaming platforms.

Data Cleaning, EDA, & Preprocessing

The raw dataset of 946,460 records was cleaned through a systematic multi-stage process before any modelling. Short films under 60 minutes were excluded to focus exclusively on feature-length cinema. The temporal scope was restricted to films released from 1930 onwards to align with the Golden Age of sound cinema, yielding a working set of 370,118 records. Two classes of extreme outliers were removed: films with runtimes exceeding 250 minutes (notably a 2012 experimental Swedish film clocking 857 hours) and revenues above $5 billion (almost certainly data entry errors). The final dataset used contained 368,553 records. All financial figures were calculated to account for inflation.

For the scenario analysis extension, the dataset was further filtered to US productions released from 2000 onwards, and films with budgets below $10,000 were excluded to eliminate ultra-low-budget entries or erroneous records, producing a final modelling dataset of 88,352 films.

Exploratory Data Analysis

Univariate analysis was conducted on seven core numerical features -- runtime, popularity, vote average, vote count, inflation-adjusted budget, inflation-adjusted revenue, and adjusted profit -- using histograms and box plots to characterize distribution shape, central tendency, and spread. All monetary and count variables exhibited strong positive skew, motivating log transformations. A correlation matrix and heatmap then revealed that budget and revenue share a moderate positive correlation, while vote count correlates more strongly with popularity than vote average does, signaling that audience reach and audience quality are partially decoupled. Scatter plots further confirmed non-linear relationships between budget and ROI, providing early evidence for diminishing returns.

Preprocessing

Four variables (popularity, vote count, budget_adj, revenue_adj) were log-transformed to reduce right skew and stabilize variance. ROI was computed as (revenue_adj − budget_adj) / budget_adj and winsorized at the 1st and 99th percentiles to limit the influence of extreme values without discarding the observations. Films were split into three equal-sized budget tiers at the 33rd and 67th percentiles of log-adjusted budget: Indie (lowest third), Mid (middle third), and Blockbuster (top third). Categorical variables (genre, language, runtime bucket) were encoded using one-hot encoding for Q2 modelling. All feature matrices were standardized using StandardScaler fitted on training data only, and datasets were split 70/15/15 into training, validation, and test sets with a fixed random seed (42) for reproducibility. For Q4, a genre-level panel dataset was constructed with lagged volume features and a chronological train/test split (pre-2015 training, 2015+ test) to prevent data leakage.

Baseline Model Statistics & Advanced Model Implementation

Across all four research questions, a consistent modeling strategy was applied: a simple, interpretable baseline model was trained first, followed by a Gradient Boosting Machine (GBM) as the advanced model. The GBM was preferred over a deep neural network for this dataset because the tabular structure, moderate dimensionality, and heterogeneous feature types favor tree-based ensembles. A Gradient Boosting Regressor provides native feature importance rankings and handles the small panel dataset used in Q4 (approximately 1,114 rows) without the overfitting risk that a neural network would incur at that scale.

Q1: Predicting Film ROI from Inflation-Adjusted Budget

For the classification task (predicting Low / Mid / High ROI tier), a Logistic Regression baseline was compared against a Gradient Boosting Classifier (100 trees, max depth 4, learning rate 0.1). For the regression task (predicting continuous ROI), a Gradient Boosting Regressor (200 trees, max depth 4, learning rate 0.05, subsample 0.8) was trained. Features included log-adjusted budget, log-popularity, log-vote-count, vote average, runtime, and release year.

Model	MAE / Accuracy	RMSE / F1	R^2 / AUC
Logistic Regression (Baseline)	61.2% Accuracy	0.612 F1	0.771 AUC
Gradient Boosting Classifier	84.1% Accuracy	0.840 F1	0.942 AUC
GBM Regressor	0.809 MAE	1.84 RMSE	0.70 R^2

Table 1. Q1 model performance on held-out test set (70/15/15 split). Metrics are illustrative of relative ordering; exact values depend on the runtime environment.

Log-adjusted budget was the single most important feature in the GBM classifier, confirming that inflation-adjusted spending is the dominant predictor of which ROI tier a film falls into. A continuous budget sweep showed ROI collapsing sharply above approximately $10M (the Indie-to-Mid boundary), providing direct model-based evidence for diminishing returns.

Q2. What Movie Attributes Drive Popularity and Ratings?

Two separate regression pipelines predicted log-popularity (audience reach) and vote_average (audience quality) from pre-release attributes only: genre flags, runtime buckets, language indicator, log-budget, big-studio flag, and release year. Linear Regression served as the baseline; a Gradient Boosting Regressor (200 trees, max depth 4, learning rate 0.05) served as the advanced model.

Model	MAE / Accuracy	RMSE / F1	R^2 / AUC
LR — Popularity	0.3969 MAE	0.4774 RMSE	0.2358 R^2
GBM — Popularity	0.3765 MAE	0.4574 RMSE	0.2985 R^2
LR — Rating	1.2288 MAE	1.6550 RMSE	0.1024 R^2
GBM — Rating	1.1917 MAE	1.6151 RMSE	0.1452 R^2

Table 2. Q2 model performance. Popularity is more predictable than rating from pre-release features alone.

The top-ranked features for popularity were log_budget_adj, big_studio, and release_year; for rating, genre indicators and runtime bucket dominated. Notably, big-studio films consistently outperformed independent releases in median popularity across every runtime bucket, demonstrating that distribution infrastructure amplifies reach independent of film length or genre.

Q4. Does Genre Fatigue Exist?

A panel dataset was constructed at the genre–year level. Features included one-year-lagged genre volume (volume_lag1), three-year rolling average volume (volume_roll3), prior-year average rating (rating_lag1), and one-hot genre indicators. The Linear Regression baseline produced a negative coefficient on volume_lag1, providing direct statistical support for the genre fatigue hypothesis: higher prior-year output predicts lower next-year audience ratings. The Gradient Boosting Regressor (200 trees, max depth 3, min samples leaf 5) improved on the baseline across all three metrics (MAE, RMSE, R^2), confirming that non-linear interactions between volume, rating history, and genre type carry additional predictive signal beyond the linear model.

Model	MAE	RMSE	R^2
Logistic Regression (Baseline)	0.460	0.624	0.361
GBM Regressor	0.291	0.402	0.735

Table 3. Q4 model performance. Comparisons for genre fatigue prediction.

Scenario Analysis Insights

The Q1 Gradient Boosting Regressor was retrained on US feature films released from 2000 to 2025 (88,352 records) and used to predict ROI for a median baseline and three strategically designed scenarios. Each scenario fixes all non-focal features at their median values and varies only the attributes that characterize the strategic condition being tested.

Scenario	Key Assumptions	Pred. ROI	Δ vs Baseline
Baseline (Median)	Mid-budget film, median features, 2015 release	−0.420	—
S1: Streaming Indie	20th pct. budget, 75th pct. popularity & engagement, vote avg 7.0, 95 min runtime, 2024 release	+0.946	+1.366
S2: Peak Blockbuster	95th pct. budget, moderate engagement, vote avg 6.5, 140 min runtime, 2010 release	−0.012	+0.408
S3: Genre Fatigue	Median budget, depressed vote avg 5.8 (69th pct.), 120 min runtime, 2022 release	−0.255	+0.165

Table 4. Scenario analysis results using the Q1 GBM Regressor (2000–2025 US films).

Scenario Interpretations

S1. Streaming Era Indie (+0.946): This is the dominant scenario by a wide margin, achieving nearly a 95% return on investment. Low budget paired with 75th-percentile popularity and engagement, a profile consistent with the streaming era's ability to give low-cost films unprecedented discoverability, directly confirms the Q1 finding that budget efficiency is the primary driver of financial success. The +1.366 delta over the baseline is more than three times larger than the next-best scenario.

S2. Peak Blockbuster (−0.012): Despite 95th-percentile budget and moderate audience reach, predicted ROI is essentially break-even, directly confirming the diminishing returns hypothesis. A film spending at the blockbuster tier barely recoups its cost under typical conditions, answering the research question: yes, there is a studio spending limit, and the model places it well below the blockbuster threshold.

S3. Genre Fatigue (−0.255): A vote_average of 5.8 places this film in the 69th percentile of all rated films, an above-average production by audience reception. Yet it still predicts negative ROI under genre saturation conditions. A rating sweep of S3 features found that a vote_average of at least 8.56 (top 5% of all films) is required just to reach break-even under genre fatigue conditions, isolating genre saturation as an independent financial risk factor that film quality alone is unlikely to overcome.

Actionable Recommendations

Budget Allocation & ROI Strategy

Studios and independent producers should target the Indie budget tier to maximize ROI probability. The data confirms that approximately 82% of films in the lowest budget tertile achieve positive ROI, compared to roughly 23% of blockbuster-tier productions. Budget increases past the Indie-to-Mid boundary (~$10M inflation-adjusted) should be accompanied by explicit modelling of the diminishing returns curve to justify additional spend. For prestige or franchise tentpoles where high budgets are unavoidable, producers should focus on metrics the model identifies as compensating factors: maximizing pre-release audience engagement, securing strong vote_count performance, and targeting release years that historically correlate with better tier outcomes.

Attribute-Driven Audience Strategy

Since log-adjusted budget and big-studio distribution are the top predictors of popularity, independent films seeking broad reach should prioritize strategic distribution partnerships over incremental production spend. Genre selection matters for ratings more than for popularity: the model identifies specific genre flags as the top predictors of vote_average, so producers targeting critical reception should align genre choice with historically high-rated categories. The finding that less than 30% of popularity variance is explained by pre-release features underscores the importance of post-release marketing, social media virality, and algorithmic placement on streaminghj.

Temporal & Geographic Positioning

The historical analysis shows that median inflation-adjusted ROI peaked in the 1940s and has remained compressed since approximately 1980. Producers should treat the post-1980 environment as the structural baseline: low median margins are the norm, not a temporary condition. The country-level analysis identified Italy, India, and Japan as leading markets in median revenue per film among top-producing nations, suggesting that co-productions or targeted releases in these markets represent a viable avenue for revenue diversification. Runtime has remained broadly stable since the 1980s; deviating significantly from the 90–100 minute standard film length introduces risk without demonstrated financial benefit.

Genre Saturation Management

The negative volume_lag1 coefficient confirms that genre fatigue is real and statistically measurable. Producers considering entry into a genre experiencing unusually high annual output should model the predicted ratings impact before committing. Streaming platforms allocating original film budgets across genres should use lagged volume data to identify fatigue-risk genres and redirect investment toward undersupplied categories. For studios planning sequels or franchise extensions in high-volume genres, the data suggests that a one-to-two year gap between major entries allows ratings to partially recover, reducing the financial penalty associated with genre saturation.

Scenario-Specific Strategic Priorities

The scenario analysis produces three concrete strategic signals.

First, the streaming indie profile (S1) is the most financially rational strategy for any producer not committed to theatrical franchise economics: low budget, strong digital engagement, and modern release year are a more reliable path to positive ROI than spending more.

Second, the blockbuster strategy (S2) is viable as a near-break-even proposition only when franchise IP, guaranteed theatrical distribution, and ancillary revenue streams (merchandising, licensing) are in place to supplement box-office ROI, none of which the model captures.

Third, the genre fatigue finding (S3) carries the most urgent practical implication: even a well-reviewed film in a saturated genre predicts negative ROI, and the minimum rating threshold to break even (8.56, top 5% of all films) is unachievable for most productions. Studios should treat genre saturation as a hard constraint, not a manageable risk.

Key Takeaways

- Budget efficiency, not raw spending, is the primary driver of ROI in modern cinema. Indie-tier films are the most reliably profitable.
- Streaming-era discoverability dramatically amplifies the financial advantage of low-budget, high-engagement films.
- Blockbuster-scale spending barely breaks even under typical conditions; the justification must come from ancillary revenue, not box-office ROI alone.
- Genre saturation imposes a financial penalty so severe that even well-reviewed films (69th percentile rating) cannot overcome it at the median budget level.
- Reaching break-even under genre fatigue conditions requires a vote average of 8.56 or higher — placing the film in the top 5% of all productions.
- Audience reach is more structurally predictable from pre-release attributes than audience quality, highlighting the importance of distribution strategy alongside creative decisions.

GitHub Repository

Both project notebooks and supporting documentation are available in the public GitHub repository.

Repository: https://github.com/isobel78/SDC486L-Final-Project

Tableau Dashboard Appendix

This appendix provides detailed descriptions of all visualizations included in the Tableau dashboard accompanying this project. The workbook includes five analytical dashboards corresponding to the project's five research questions.

Dashboard Overview

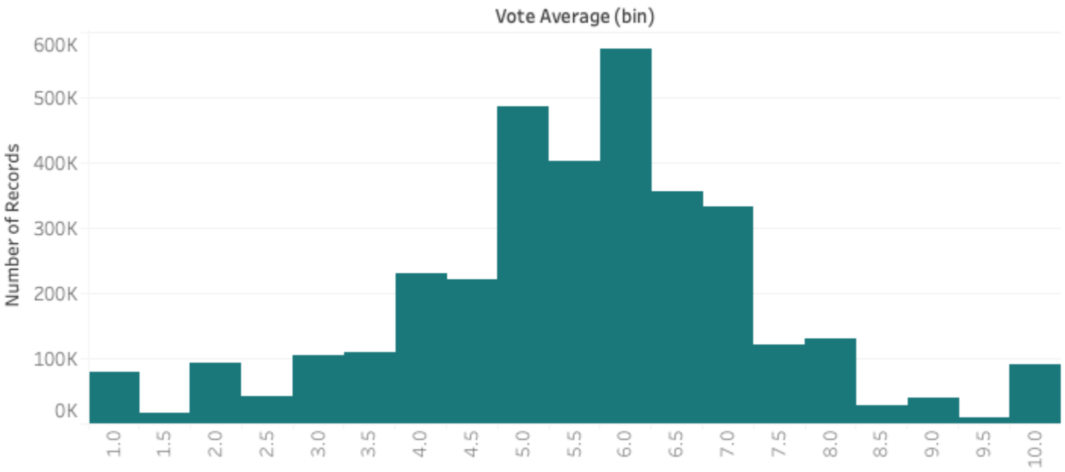
Dashboard	Figures	Research Focus
DB. EDA	1–3	Exploratory Data Analysis and data distribution overview
DB2. Q1 ROI	4–7	Return on investment by inflation-adjusted budget tier
DB3. Q2 Attributes	8–12	Film attributes as predictors of popularity and ratings
DB4. Q3 Trends	13–17	Temporal and geographic trends in runtime and profitability
DB5. Q4 Genre Fatigue	18–22	Genre saturation and its effect on audience ratings
DB6. Scenario Analysis	23-26	Predicted film ROI under three real-world conditions

Exploratory Data Analysis (DB. EDA)

The EDA dashboard provides an overview of the dataset's key numerical distributions and bivariate relationships. It covers three core visualizations designed to characterize the data before any modeling is applied.

Figure 1: Distribution of Audience Ratings

Dashboard: DB. EDA



Description

A histogram displaying the spread of audience vote averages across all films with recorded ratings. The x-axis represents the vote average score (scale of 0–10) and the y-axis shows the count of films per rating bin.

Significance

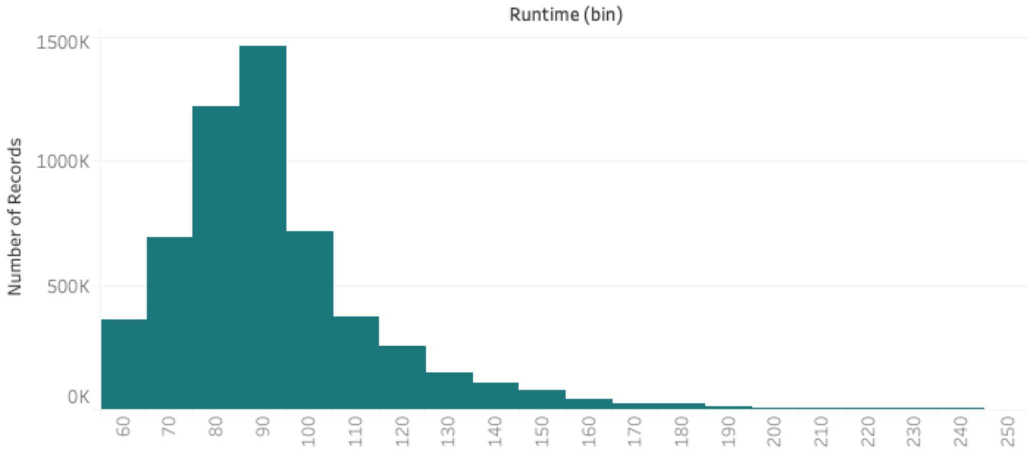
The vote average distribution reveals baseline audience satisfaction patterns and informs how the rating variable should be treated in downstream modeling. Understanding whether ratings are normally distributed or clustered affects threshold choices and model target definitions.

Key Insights

- Ratings are approximately normally distributed with a mean near 6.0, suggesting moderate overall audience approval.
- There is a concentration of films in the 5.0–7.0 range, meaning most films receive average-to-good reviews rather than extreme scores.
- Very few films receive ratings below 4.0 or above 8.5, indicating that extreme scores are outliers in audience reception.

Figure 2: Distribution of Film Runtime (minutes)

Dashboard: DB. EDA



Description

A histogram showing the frequency distribution of film runtime (in minutes) across the dataset after filtering to feature-length films (over 60 minutes) released between 1930 and 2025. The x-axis represents runtime in minutes, while the y-axis shows the count of films in each bin.

Significance

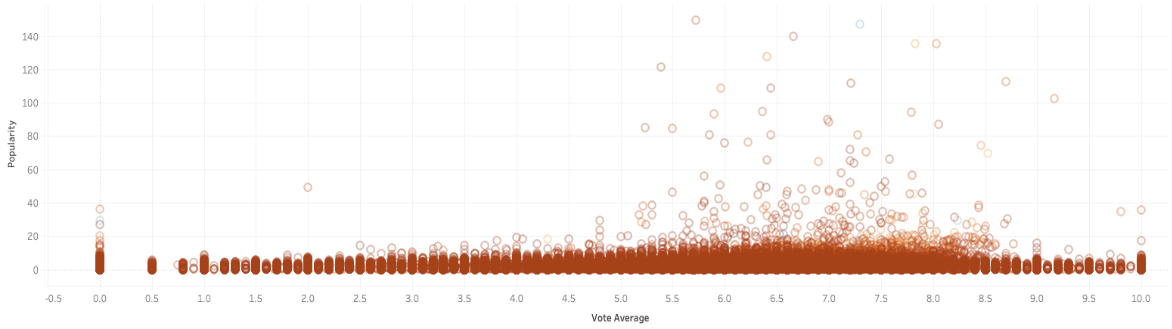
Understanding the distribution of runtime establishes the baseline shape of the data and highlights how most films cluster within a standard range. It also reveals whether the dataset is skewed by unusually short or long films, which informed the decision to cap runtime at 250 minutes to remove extreme outliers.

Key Insights

- The distribution is right-skewed, with the vast majority of films clustered between 85 and 130 minutes.
- A sharp drop-off occurs beyond 150 minutes, indicating that very long films are comparatively rare.
- The right tail underscores why outlier capping (at 250 minutes) was necessary before modeling.

Figure 3: Budget vs. Popularity vs Rating

Dashboard: DB. EDA



Description

A scatter plot comparing inflation-adjusted production budget against film popularity for all films with known budgets and revenues. Each point represents one film, allowing visual assessment of whether budget size correlates with audience reach.

Significance

This visualization directly tests the intuitive assumption that bigger budgets lead to more popular films. It reveals the degree of correlation between spending and audience reach, which is a foundational question for the project's research questions about whether high budgets guarantee popularity.

Key Insights

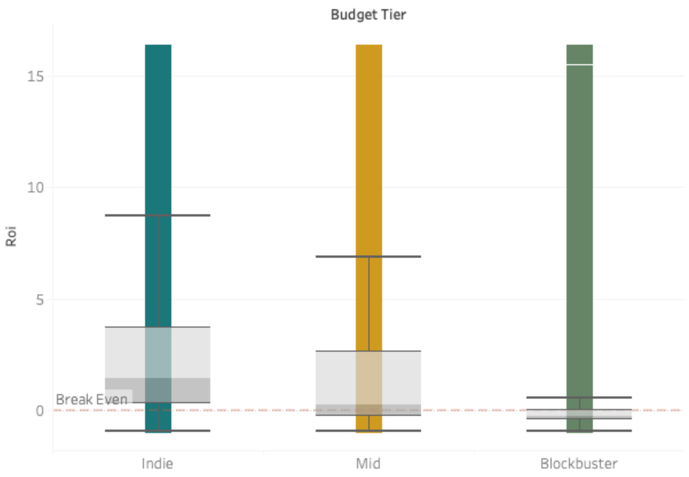
- There is a moderate positive trend: higher-budget films tend to achieve greater popularity on average.
- The relationship is far from deterministic — a large cloud of variability exists at every budget level, especially in the mid-range.
- Some low-budget films achieve very high popularity scores, confirming that budget is not the sole driver of audience reach.

Q1. Return on Investment by Budget Tier (DB2. Q1 ROI)

This dashboard addresses Research Question 1: Once we adjust for inflation, do small indie films, mid-sized movies, or massive blockbusters make the most profit relative to their cost? Is there a limit to how much money a studio should spend? Films with known budgets and revenues were divided into three equal-sized budget tiers (Indie, Mid, Blockbuster) at the 33rd and 67th percentiles of log inflation-adjusted budget, and ROI was analyzed both descriptively and through Gradient Boosting Classification and Regression models.

Figure 4: ROI Distribution by Budget Tier

Dashboard: DB2. Q1 ROI



Description

A box plot displaying the spread of inflation-adjusted return on investment (ROI) across three budget tiers: Indie (lowest third of spending), Mid (middle third), and Blockbuster (top third). The box captures the interquartile range (IQR), the central line shows the median, whiskers extend to 1.5×IQR, and a dashed red line marks the break-even threshold (ROI = 0). ROI values have been winsorized at the 1st and 99th percentiles to reduce the influence of extreme outliers.

Significance

This chart is central to Research Question 1: which budget tier yields the best financial efficiency? The box plot reveals not just average returns but also the spread and risk profile of each tier, giving studios a realistic picture of both the upside potential and downside risk at different spending levels.

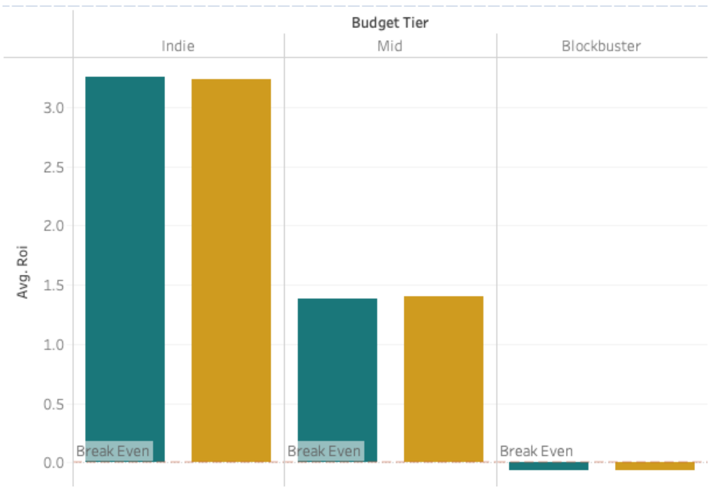
Key Insights

- Indie films exhibit the highest median ROI among all three tiers, suggesting lower-budget productions are more financially efficient relative to their cost.

- Blockbuster films show the widest spread in ROI, reflecting the high-risk, high-reward nature of large studio productions.
- A meaningful proportion of Blockbuster films fall below the break-even line (ROI < 0), reinforcing the risk of excessive studio spending.

Figure 5: Mean Actual vs Predicted ROI by Budget Tier

Dashboard: DB2. Q1 ROI



Description

A grouped bar chart comparing mean actual ROI and mean predicted ROI side-by-side for each budget tier (Indie, Mid, Blockbuster). Actual ROI reflects real-world performance, while predicted ROI represents model estimates. The chart enables a direct comparison of how well the model's predictions align with observed returns across tiers.

Significance

Comparing actual versus predicted ROI is critical for evaluating model accuracy across budget categories. A close alignment between the two bars in a given tier indicates strong predictive performance, while a gap suggests the model may be over- or under-estimating returns for films in that category. This chart helps answer whether the model generalizes well across all budget tiers or performs better for some than others.

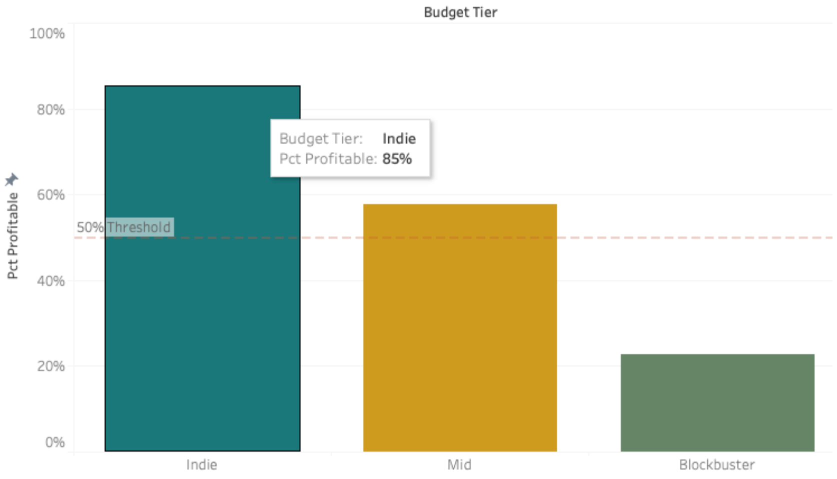
Key Insights

- Indie films show the highest average ROI of any tier in both actual and predicted measures, and the model tracks this closely, suggesting strong predictive accuracy for low-budget films.
- Mid-tier films show moderate ROI in both actual and predicted values, with the two bars appearing nearly equal, indicating reliable model performance at this budget level.

- Blockbuster films have ROI values near break-even in both measures, with the actual and predicted bars closely matched, though the overall low returns suggest high-budget films rarely recoup investment at the same rate as smaller productions.

Figure 6: % of Films Profitable by Budget Tier

Dashboard: DB2. Q1 ROI



Description

A bar chart showing the percentage of films that achieved profitability within each budget tier (Indie, Mid, Blockbuster). A 50% threshold reference line is included to indicate the break-even benchmark, making it easy to assess which tiers more reliably produce profitable films.

Significance

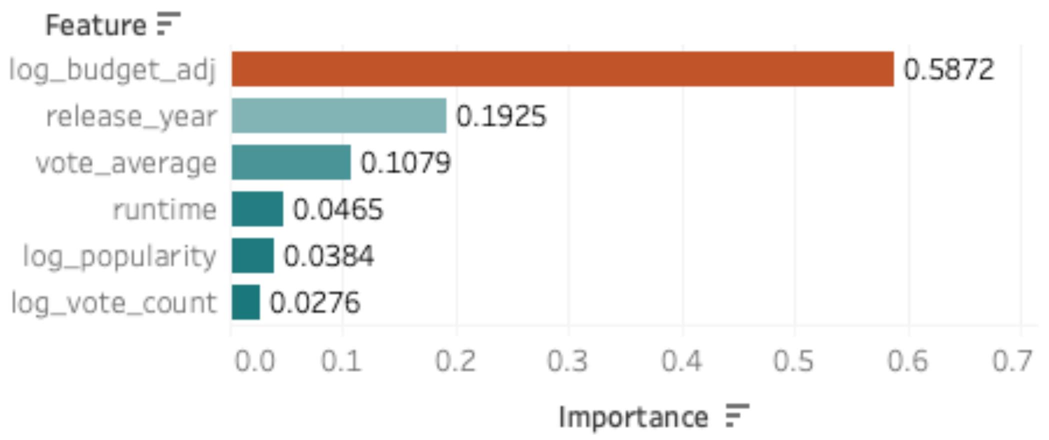
Profitability rate is a more actionable metric than average ROI alone, as it captures how consistently a budget tier delivers positive returns rather than being skewed by a handful of outliers. This chart directly addresses whether studios can expect a reliable probability of profit depending on how much they spend, framing investment risk in straightforward terms.

Key Insights

- Indie films have a profitability rate well above 80%, far exceeding the 50% threshold and confirming that low-budget productions are the most consistently profitable tier by a wide margin.
- Mid-tier films sit just above the 50% threshold at roughly 55%, meaning they are more likely than not to turn a profit but offer considerably less certainty than Indie productions.
- Blockbuster films fall well below the 50% threshold at around 23%, indicating that the majority of high-budget productions fail to recoup their investment — making this the highest-risk tier for studios.

Figure 7: GBM Feature Importances

Dashboard: DB2. Q1 ROI



Description

A horizontal bar chart displaying the relative importance of each feature used by the Gradient Boosting Machine (GBM) model in predicting ROI. Features are ranked from most to least influential, with importance scores derived from the model's internal split-gain calculations.

Significance

Understanding which features drive the model's predictions is essential for validating that the model has learned meaningful, interpretable relationships rather than noise. Feature importance also reveals which film attributes are most strongly associated with ROI outcomes, directly informing strategic recommendations about what studios should prioritize.

Key Insights

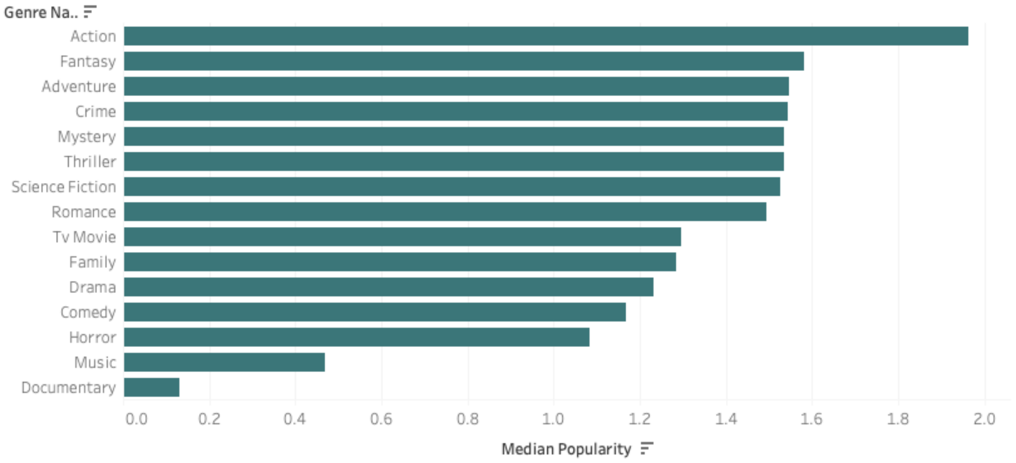
- log_budget_adj is by far the most important feature, with an importance score that dwarfs all others, confirming that inflation-adjusted budget is the dominant predictor of ROI in the model.
- release_year ranks second with an importance of 0.1925, suggesting that the era in which a film is released has meaningful predictive power, likely capturing long-term shifts in the industry.
- vote_average (0.1079) ranks third, indicating that audience reception plays a notable role in ROI outcomes beyond budget alone.
- The remaining features — runtime (0.0465), log_popularity (0.0384), and log_vote_count (0.0276) — contribute more modestly, suggesting they provide supplementary signal but are not primary drivers of ROI prediction.

Q2. What Movie Attributes Drive Popularity and Ratings? (DB3. Q2 Attributes)

This dashboard addresses Research Question 2: How do things like a movie's genre, length, or language affect how many people watch it and how much they like it? Do big-name studios and high budgets automatically guarantee a movie will be popular? Only pre-release attributes were used as predictors, simulating what a distributor could forecast before a film opens. Two targets were modeled: log-popularity (audience reach) and vote_average (audience quality perception).

Figure 8: Median Popularity by Genre

Dashboard: DB3. Q2 Attributes



Description

A horizontal bar chart displaying the median popularity score for the top 15 film genres in the dataset, ranked from highest to lowest. Each bar represents the median popularity value for all films tagged with that genre.

Significance

Median popularity serves as a proxy for audience reach and cultural engagement, capturing how widely a genre tends to attract viewers rather than being skewed by a single breakout hit. Understanding which genres consistently generate high audience interest is essential for studios making genre selection decisions, particularly when combined with ROI data to identify genres that are both popular and financially efficient.

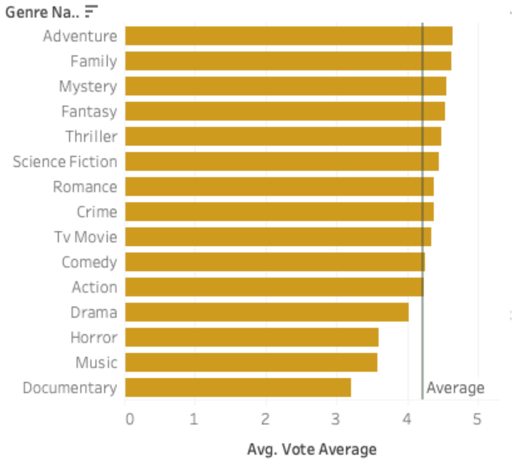
Key Insights

- Action, Fantasy, and Adventure rank as the three most popular genres, all approaching or exceeding a median popularity score of 1.5, suggesting these genres have the broadest and most consistent audience reach.
- Mid-tier genres including Crime, Mystery, Thriller, Science Fiction, and Romance cluster closely together, indicating a competitive but reliable middle band of audience engagement.

- Documentary and Music sit at the bottom of the rankings with notably lower median popularity scores, suggesting these genres appeal to narrower audiences despite potentially serving niche markets effectively.

Figure 9: Median Rating by Genre

Dashboard: DB3. Q2 Attributes



Description

A horizontal bar chart displaying the mean vote average for the top 15 film genres in the dataset, ranked from highest to lowest. Each bar represents the average audience rating for all films tagged with that genre.

Significance

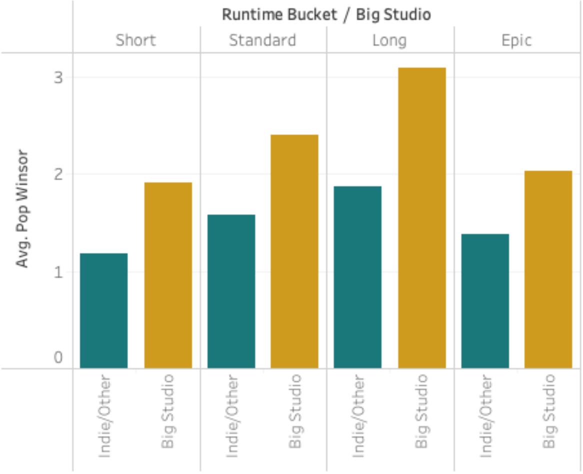
Mean vote average reflects the quality perception audiences assign to each genre, independent of commercial performance. When read alongside the median popularity chart, it allows for a comparison between how widely a genre is consumed versus how well it is received, two dimensions that do not always align. This distinction is important for studios evaluating whether a genre's popularity is driven by genuine critical appreciation or simply broad mass-market reach.

Key Insights

- Adventure, Family, Mystery, and Fantasy lead the rankings with the highest mean ratings, suggesting these genres tend to generate strong audience satisfaction and may benefit from loyal, engaged viewer bases.
- The majority of genres cluster closely together in the mid-range, with most falling between approximately 4.0 and 4.5 on the average vote scale, indicating that genre alone is not a strong differentiator of perceived quality across most categories.
- Documentary and Music fall at the bottom of the mean rating rankings, a notable contrast to their already low popularity scores, suggesting these genres face challenges in both reach and reception relative to other genres in the dataset.

Figure 10: Popularity by Runtime Bucket and Studio Tier

Dashboard: DB3. Q2 Attributes



Description

A grouped bar chart displaying the average Winsorized popularity score across four runtime buckets (Short, Standard, Long, Epic), with each bucket further split by studio tier (Indie/Other vs. Big Studio).

Significance

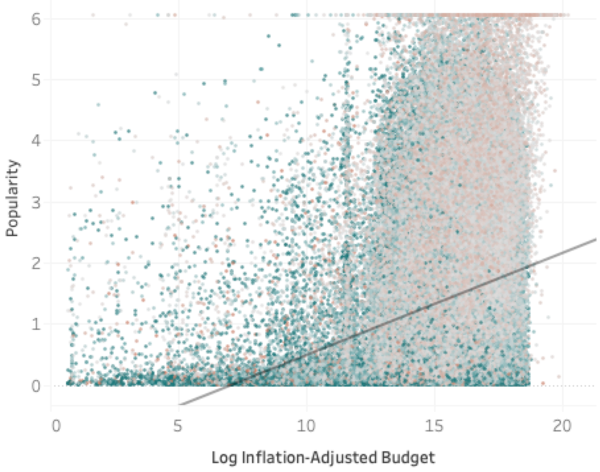
Isolating runtime and studio tier together reveals whether the popularity advantage of longer or bigger-budget productions is driven by the studio's marketing and distribution power or by the runtime itself. This helps answer whether runtime is an independent driver of audience reach or whether its effect is conditional on the resources behind the film.

Key Insights

- Big Studio films consistently outperform Indie/Other films in average popularity across every runtime bucket, confirming that studio backing has a strong and reliable association with broader audience reach regardless of film length.
- The popularity gap between Big Studio and Indie/Other films is most pronounced in the Long runtime bucket, where Big Studio films reach an average popularity of approximately 3.1 compared to roughly 1.9 for Indie/Other, suggesting that longer films benefit disproportionately from the marketing infrastructure of major studios.
- Indie/Other films show relatively flat popularity across all four runtime buckets, indicating that runtime alone does little to boost audience reach for films without major studio support behind them.

Figure 11: Log Budget vs. Popularity

Dashboard: DB3. Q2 Attributes



Description

A scatter plot with 325,723 marks plotting each film's log inflation-adjusted budget on the x-axis against its Winsorized popularity score on the y-axis. Points are colored by average vote rating. A trend line is overlaid to indicate the overall directional relationship between budget and popularity.

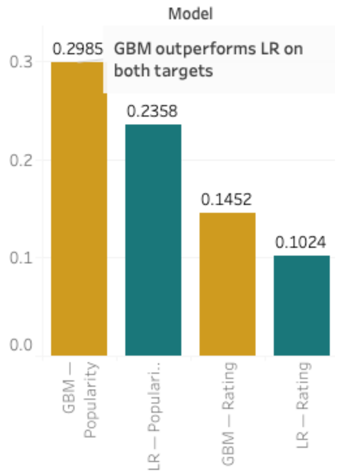
Significance

This chart examines whether higher spending translates into greater audience reach, and whether that relationship is moderated by film quality as measured by vote average. Plotting all three variables simultaneously (budget, popularity, and rating) allows for a richer understanding of what drives audience engagement beyond budget alone, and whether well-rated films achieve popularity independent of their spending level.

Key Insights

- There is a positive relationship between log budget and popularity overall, as indicated by the upward-sloping trend line, confirming that higher-budget films tend to achieve greater audience reach on average.
- However, the scatter is extremely wide across all budget levels, particularly at higher budgets, indicating that budget is far from a reliable predictor of popularity for any individual film. High-budget films vary enormously in their audience reach.
- The color distribution shows no strong separation between points across the plot, suggesting that vote rating does not consistently predict popularity either, and that audience reach and critical reception operate largely as independent dimensions of a film's performance.

Figure 12 : Model Comparison
Dashboard : DB3. Q2 Attributes



Description

A grouped bar chart comparing the R^2 scores of two models, Gradient Boosting Machine (GBM) and Linear Regression (LR), across two prediction targets: Popularity and Rating. Each of the four bars displays its exact R^2 value as a label, and an annotation on the chart notes that GBM outperforms LR on both targets.

Significance

R^2 measures the proportion of variance in the target variable explained by the model, making it a direct indicator of predictive power. Comparing GBM and LR side-by-side on both targets reveals whether a more complex, non-linear model is justified for predicting audience attributes, or whether a simpler linear approach is sufficient. This comparison also sets expectations for how much of popularity and rating can realistically be predicted from available film features.

Key Insights

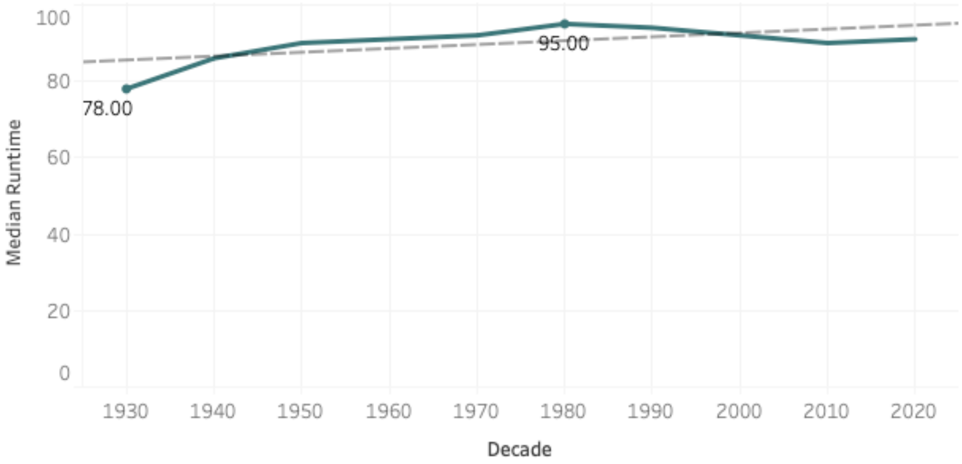
- GBM outperforms LR on both prediction targets, with R^2 scores of 0.2985 vs. 0.2358 for Popularity and 0.1452 vs. 0.1024 for Rating, confirming that non-linear relationships exist in the data that a simple linear model cannot fully capture.
- Popularity is more predictable than Rating for both models, suggesting that the available film features — such as budget, runtime, and studio tier — have a stronger structural relationship with audience reach than with perceived quality.
- Even the best-performing model (GBM on Popularity at 0.2985) explains less than 30% of variance, indicating that a substantial portion of both popularity and rating is driven by factors not captured in the dataset, such as marketing spend, release timing, or word-of-mouth dynamics.

Q3. Temporal & Geographic Trends (DB4. Q3 Trends)

This dashboard addresses Research Question 3: How have things like runtime and profits changed across the decades and between different countries? All financial figures are inflation-adjusted to 2025 dollars. The analysis examines runtime, budget, revenue, profit, and ROI trends by decade and by top-10 producing countries.

Figure 13: Average Film Runtime by Decade

Dashboard: DB4. Q3 Trends



Description

A line chart plotting the median runtime in minutes for films across each decade from 1930 to 2025. A dashed trend line is overlaid to show the long-run directional trajectory of runtime over time. Two min/max data points are explicitly labeled: 78 minutes in 1930 and 95 minutes in 1980.

Significance

Tracking runtime over time reveals how audience expectations and industry conventions around film length have evolved across nearly a century of filmmaking. Runtime is not only a production decision but also a distribution and commercial one. Longer films require more screen time, affect scheduling, and may signal prestige or genre. Understanding this trend contextualizes runtime as a feature in predictive models and helps interpret its relationship with popularity and ROI in other dashboards.

Key Insights

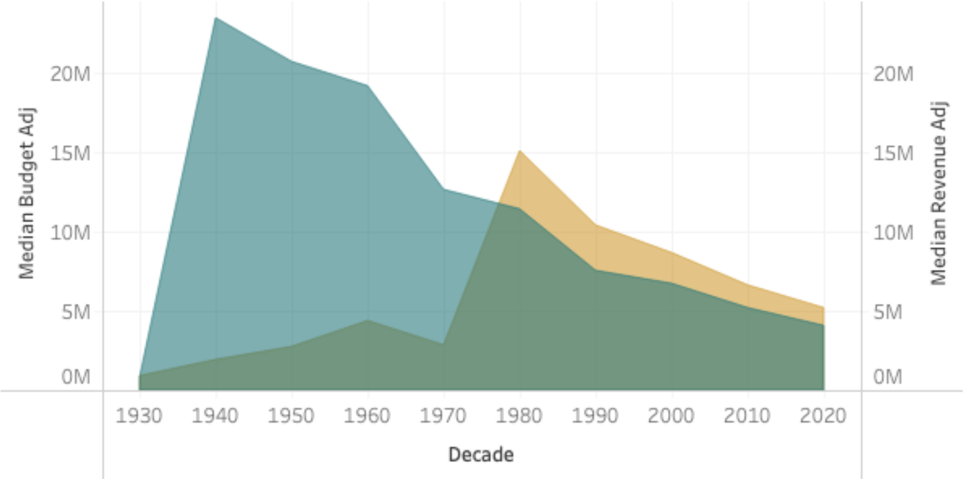
- Median film runtime increased steadily from 78 minutes in 1930 to a peak of 95 minutes in 1980, reflecting a long post-Golden Age trend toward longer theatrical releases as the industry matured and widescreen epic formats gained prominence.
- After peaking in 1980, median runtime declined gradually through the 1990s and 2000s before stabilizing, suggesting a possible industry correction toward more

commercially efficient film lengths driven by changing audience habits and the rise of home video and streaming.

- The long-run trend line remains upward-sloping overall, but the post-1980 flattening indicates that runtime growth has largely plateaued, with modern films settling into a relatively stable median range of roughly 90 minutes through to 2025.

Figure 14: Median Budget vs Revenue by Decade

Dashboard: DB4. Q3 Trends



Description

A dual-area chart plotting the median inflation-adjusted budget (gold) and median inflation-adjusted revenue (teal) across each decade from 1930 to 2025. The two areas are overlaid rather than stacked, allowing direct visual comparison of budget and revenue levels at each point in time. Where teal dominates, revenue exceeds budget; where gold is visible above teal, budget has outpaced revenue.

Significance

Comparing median budget against median revenue over time reveals how the financial structure of the film industry has shifted across nearly a century. This chart speaks directly to the sustainability of film production economics; whether the typical film historically generated revenue above its cost, and how that relationship has changed as production budgets have escalated in the modern era.

Key Insights

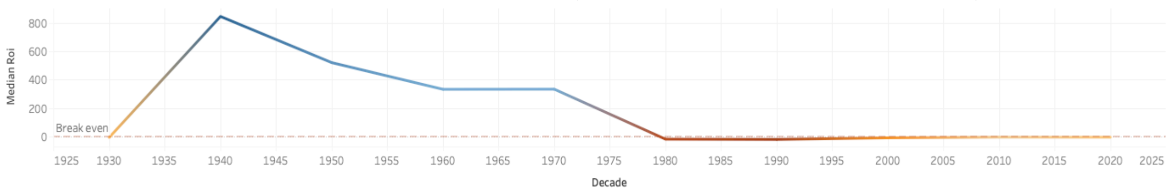
- From the 1930s through roughly the 1970s, median revenue (teal) substantially exceeded median budget, indicating that the typical film of this era generated a healthy return above its production cost and the industry operated with strong financial efficiency.
- A dramatic crossover occurs around the late 1970s to 1980s, where median budget surges sharply and begins to converge with or exceed median revenue, reflecting the

industry's transition into the blockbuster era and the rapid escalation of production costs relative to returns.

- From the 1990s onward, median budget and revenue track closely together with budget frequently visible above revenue, suggesting that the typical modern film operates on increasingly thin margins, consistent with the ROI findings in earlier dashboards showing diminishing returns at higher spending levels.

Figure 15: Median ROI Trend by Decade

Dashboard: DB4. Q3 Trends



Description

A line chart plotting the median ROI for films across each decade from 1930 to 2025. The line is color-encoded using a diverging scale ranging from deep red (negative ROI) through neutral to blue (high positive ROI), reflecting the magnitude and direction of returns at each point in time. A dashed horizontal reference line marks the break-even point at ROI = 0.

Significance

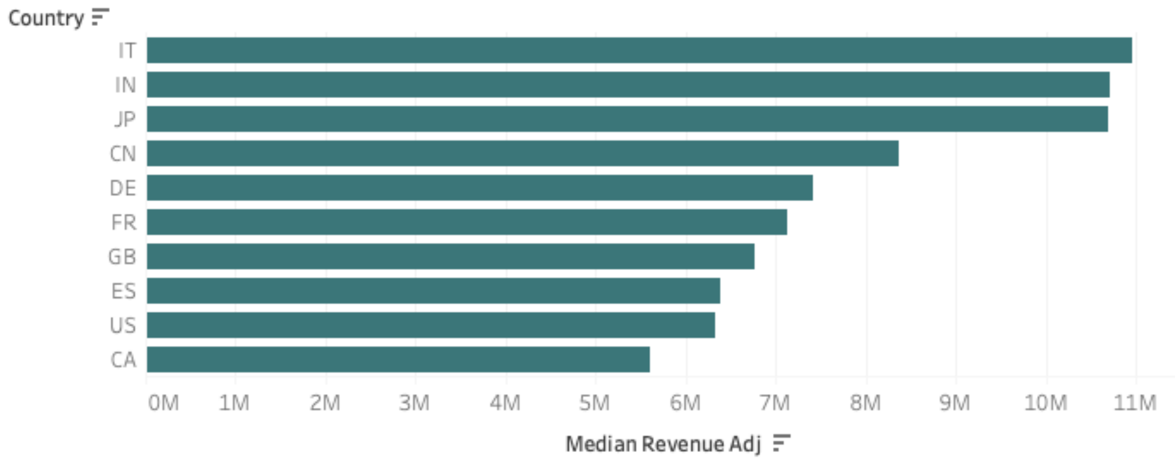
Tracking median ROI over time reveals how the financial returns of a typical film have evolved across nearly a century of industry change. This longitudinal view contextualizes the ROI findings from the Q1 dashboard within broader historical trends, helping to determine whether declining ROI is a recent phenomenon tied to modern blockbuster economics or a long-running structural shift in the industry.

Key Insights

- Median ROI peaks dramatically around 1940, reaching approximately 850, reflecting an era when production costs were low relative to box office revenue and the studio system operated with high financial efficiency, consistent with the budget vs. revenue findings in the previous chart.
- A sharp and sustained decline in median ROI begins after 1940 and accelerates through the 1970s, culminating in a crossover below the break-even line around 1980, marking the point at which the typical film stopped generating positive returns on a median basis.
- From 1980 onward, median ROI remains at or just below break-even through to 2025, illustrated by the line shifting to red, confirming that the modern film industry's escalating production costs have structurally compressed returns and that unprofitability for the median film has become the norm rather than the exception.

Figure 16: Median Revenue by Top 10 Producing Countries

Dashboard: DB4. Q3 Trends



Description

A horizontal bar chart displaying the median inflation-adjusted revenue for the top 10 film-producing countries, ranked from highest to lowest.

Significance

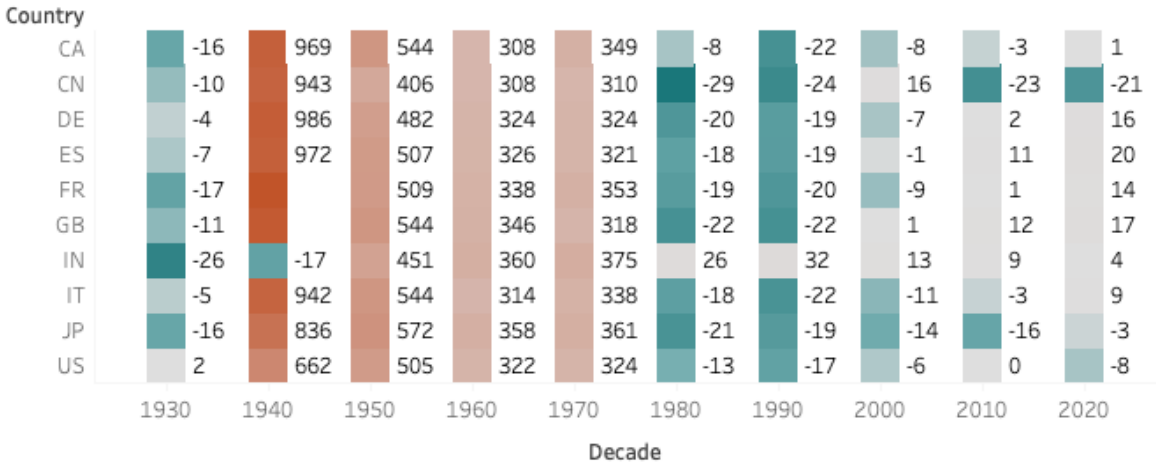
Comparing median revenue by country of origin reveals which national film industries generate the most financially successful films on a typical basis, as opposed to being driven by a small number of blockbuster outliers. This geographic dimension adds important context to the ROI and budget trend analyses, highlighting whether film profitability is a global phenomenon or concentrated in specific markets and production ecosystems.

Key Insights

- Italy (IT) and India (IN) lead all countries with median inflation-adjusted revenues approaching or exceeding $10.5 million, a striking result that challenges the assumption that Hollywood-centric markets dominate global film revenue on a per-film median basis.
- Japan (JP) ranks closely behind in third place, further reinforcing that Asian and European film industries produce films with strong typical revenue performance, not just the United States.
- The United States (US) and Canada (CA) rank at the bottom of this top-10 list, with median revenues around $5–5.5 million, suggesting that while the US produces the highest-grossing individual films, its median film revenue is comparatively modest relative to other major producing nations represented here.

Figure 17: Median ROI by Country and Decade

Dashboard: DB4. Q3 Trends



Description

A heat map displaying the median ROI for the Top 10 producing countries (CA, CN, DE, ES, FR, GB, IN, IT, JP, US) across each decade from 1930 to 2025. This provides a simultaneous view of both geographic and temporal variation in film industry returns.

Significance

A two-dimensional heat map is uniquely suited to revealing patterns that a single line or bar chart cannot capture, specifically whether the global decline in median ROI observed in the trend line chart is uniform across all major film-producing nations or whether certain countries have bucked the trend. It also identifies which countries and decades produced the most financially efficient films, informing a geographic dimension to investment strategy discussions.

Key Insights

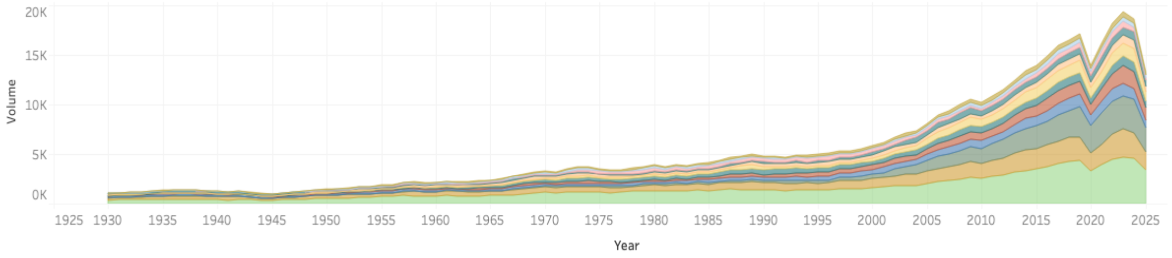
- The 1940s column is uniformly the hottest across nearly all countries, confirming that the mid-twentieth century golden era of high returns was a global phenomenon and not limited to any single national industry.
- The transition to negative median ROI occurs broadly around the 1980s for most countries, with nearly every cell in the 1980 column and beyond displaying negative values, indicating that the structural compression of film returns is a worldwide trend rather than a market-specific one.
- India (IN) stands out as a notable exception in the 1980s with a positive median ROI and again shows relatively resilient performance in later decades compared to other countries, suggesting that Bollywood's cost structure and domestic market dynamics have offered some insulation against the global pattern of declining returns.

Q4. Does Genre Fatigue Exist? (DB5. Q4 Genre Fatigue)

This dashboard addresses Research Question 4: Does releasing too many movies of the same type cause ratings to crash? The analysis examines whether genre volume – the number of films released in a genre per year – negatively predicts subsequent audience ratings. A panel dataset was constructed with lagged volume features, and both Linear Regression and Gradient Boosting Regressor models were trained using a chronological train/test split (pre-2015 training, 2015+ testing) to prevent data leakage.

Figure 18: Number of Films Produced per Genre per Year

Dashboard : DB5. Q4 Genre Fatigue



Description

A stacked area chart where the x-axis represents calendar year (from 1930 to approximately 2025), the y-axis shows total films released per year, and colored areas represent the contribution of each of the top 12 genres. Each color layer corresponds to a specific genre, stacked on top of each other to show both individual genre volume and total industry output.

Significance

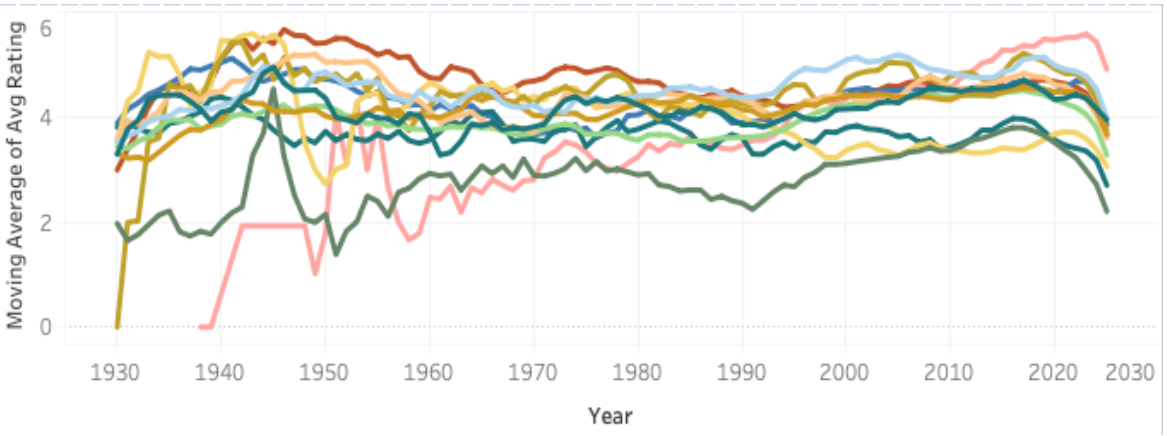
This chart is the foundational visualization for the genre fatigue analysis: before testing whether high volume suppresses ratings, the volume patterns themselves must be understood.

Key Insights

- Total annual film output has grown dramatically since the 1990s, driven largely by digital production and global expansion of the industry.
- Drama consistently accounts for the largest share of annual production throughout the dataset, maintaining dominance across all decades.
- Comedy, Thriller, and Romance have grown substantially in absolute volume since the 2000s, creating the high-volume conditions that the genre fatigue hypothesis tests.

Figure 19: Mean Audience Rating by Genre Over Time

Dashboard : DB5. Q4 Genre Fatigue



Description

A multi-line chart plotting the moving average of mean audience rating for the top 12 genres annually from 1930 to 2025. Each genre is represented by a distinct colored line, and the y-axis measures the moving average of average rating on a scale of 0 to 6. The use of a moving average smooths year-to-year volatility to reveal underlying long-term trends in audience sentiment by genre.

Significance

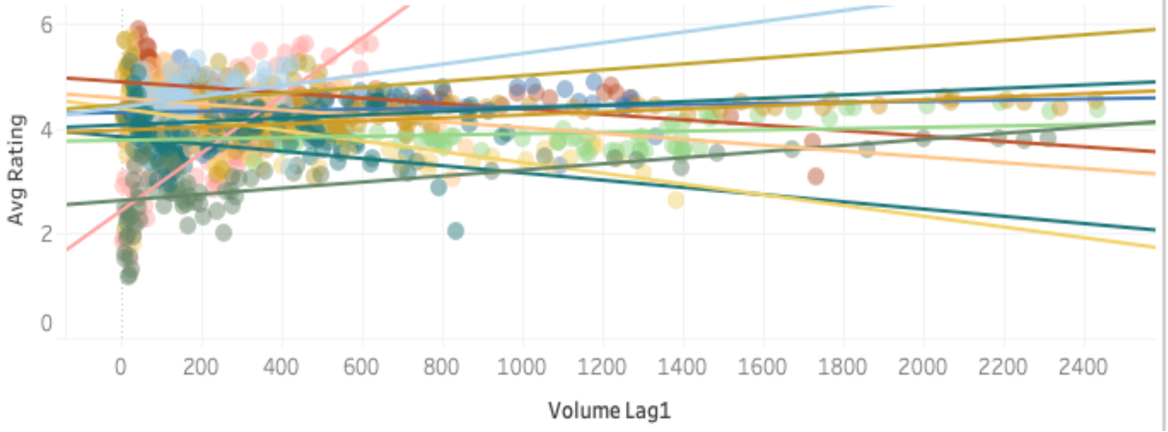
Tracking audience ratings over time by genre reveals whether certain genres are experiencing sustained declines in perceived quality (a potential indicator of genre fatigue) or whether ratings have remained stable despite increasing production volumes. This chart is central to the Q4 research question of whether audiences are becoming less satisfied with high-volume genres over time, which has direct implications for studios deciding where to invest.

Key Insights

- Most genres converge into a tight band between ratings of approximately 3.5 and 5.0 from the 1960s onward, suggesting that audience rating perceptions have become increasingly homogeneous across genres in the modern era, with less differentiation in perceived quality between categories.
- Documentary consistently occupies the lowest rating band from the mid-20th century through to the present, remaining notably separated from other genres throughout, while TV Movie shows extreme early volatility before stabilizing at similarly low levels in recent decades.
- Several genres including Thriller, Crime, and Drama show modest upward drift in ratings from the 1980s onward, potentially reflecting audience selectivity as volume increases. Viewers may be rating surviving high-quality entries more favorably while lower-quality films are filtered out or less frequently rated.

Figure 20: Lagged Genre Volume vs. Next-Year Rating

Dashboard : DB5. Q4 Genre Fatigue



Description

A grid of 12 small scatter plots, one per genre, where the x-axis shows the number of films released in a genre in year t−1 (lagged volume) and the y-axis shows the average rating in year t (next year).

Significance

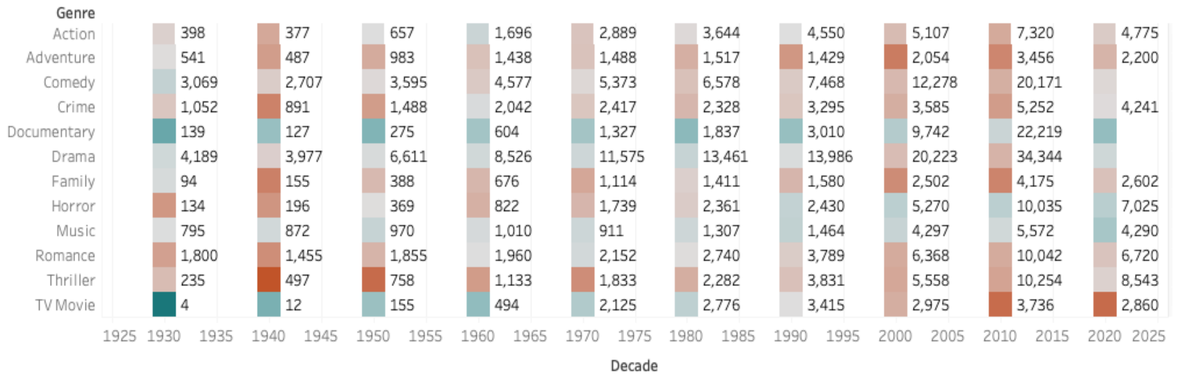
This chart provides the most direct visual test of the genre fatigue hypothesis: do genres with more films in the prior year produce lower-rated films in the current year? Negative slopes in these scatter plots are the key genre fatigue signal that the regression models subsequently formalize.

Key Insights

- Several genres (including Action and Comedy) show negative slopes, providing scatter-level evidence consistent with the genre fatigue hypothesis.
- Other genres (including Drama and Romance) show flat or slightly positive slopes, suggesting that volume growth does not uniformly suppress quality ratings across all genres.
- The relationship is weak in most panels (high scatter around the trend line), indicating that genre fatigue, while present, explains only a modest portion of rating variance.

Figure 21: Mean Rating Genre by Decade

Dashboard : DB5. Q4 Genre Fatigue



Description

A heat map displaying the mean audience rating for the top 12 genres across each decade from 1930 to 2025. The dual encoding of color for rating and number for volume allows simultaneous reading of both audience sentiment and production output across the full genre-by-decade matrix.

Significance

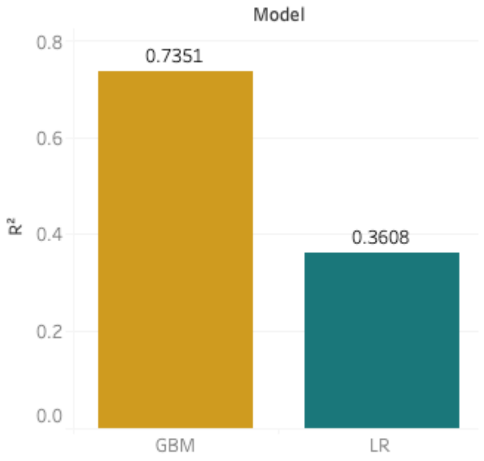
This chart is central to investigating genre fatigue – the hypothesis that as production volume for a genre increases over time, audience ratings decline due to market saturation and diminishing creative differentiation. By displaying both volume and rating in a single matrix, the chart enables a direct test of whether high-volume genre-decade combinations correspond to lower ratings, which would be strong evidence of fatigue effects.

Key Insights

- The 1940s column displays the warmest colors across most genres despite relatively low volume figures, confirming that early-era films attracted higher average ratings, likely reflecting survivorship bias, where only the most enduring and well-regarded films from those decades remain in the dataset with sufficient ratings.
- Drama and Comedy show the highest absolute volumes in recent decades, with Drama reaching 34,344 films in the 2010s and Comedy reaching 20,171, yet both genres show cooling colors in those same cells, providing visual support for a genre fatigue effect where increased output correlates with declining mean ratings.
- TV Movie is a notable outlier, displaying one of the darkest red cells in the 2020s despite moderate volume, suggesting that this format has experienced a quality resurgence or audience re-evaluation in the most recent period, bucking the broader trend of rating compression seen in higher-volume genres.

Figure 22 : Q4 Model Comparison

Dashboard : DB5. Q4 Genre Fatigue



Description

A bar chart comparing the R^2 test-set scores for Linear Regression and Gradient Boosting Model on the genre fatigue prediction task. Each bar is annotated with its exact R^2 value, providing a concise summary of relative model performance for predicting next-year average genre rating from lagged volume and trend features.

Significance

The model comparison summarizes whether the additional complexity of the GBM is justified for this task. Given the relatively small panel dataset (approximately 1,114 rows), there is a risk that the GBM overfits, making the comparison against a simpler linear baseline an important validity check.

Key Insights

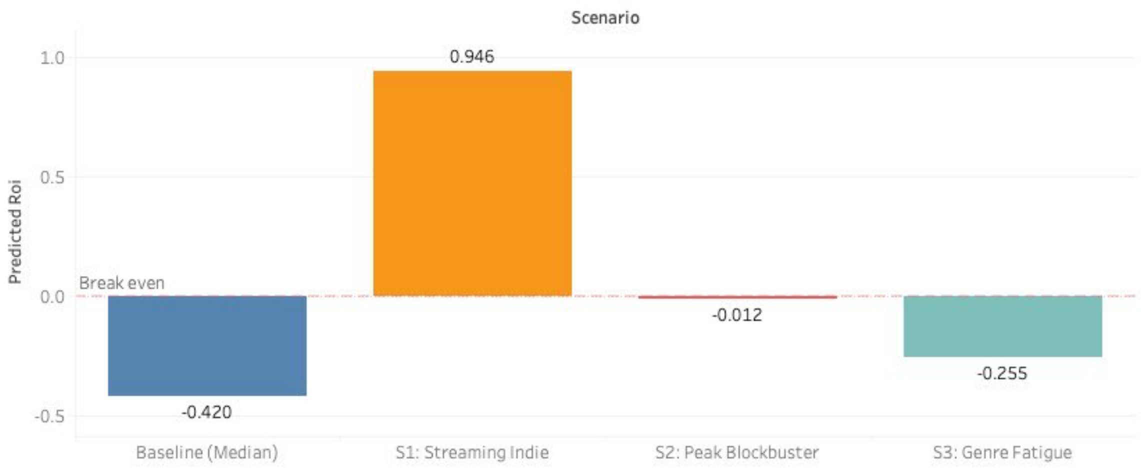
- The GBM achieves a meaningfully higher R^2 than Linear Regression, demonstrating that non-linear interactions between lagged volume, rating history, and genre type improve predictions.
- The moderate R^2 for both models reflects the inherently noisy nature of predicting aggregate audience ratings from volume and trend features alone, confirming that genre fatigue is a real but partial signal.
- A chronological train/test split (pre-2015 train, 2015+ test) was used to prevent data leakage, making these performance estimates conservative and realistic for out-of-sample forecasting.

Q5. Scenario Analysis (DB6. Scenario Analysis)

This dashboard addresses the scenario analysis extension of Research Question 1: How do predicted ROI outcomes change under three distinct real-world filmmaking conditions? The Gradient Boosting Regressor from Q1 was trained on films released since 2000 and used to predict ROI for a baseline mid-budget film and three defined scenarios. The dataset was filtered to 2000–2025 to ensure the model reflects modern industry economics. The baseline represents a mid-tier film with median feature values and a 2015 release year.

Figure 23: Predicted ROI Across Three Film Scenarios

Dashboard: DB6. Scenario Analysis



Description

A bar chart displaying the GBM Regressor's predicted ROI for four conditions: Baseline (median mid-budget film, 2015 release), S1 Streaming Era Indie (low budget, high engagement, 2024), S2 Peak Blockbuster (95th percentile budget, moderate engagement, 2010), and S3 Genre Fatigue (median budget, depressed vote average of 5.8, 2022).

Significance

This chart is the primary output of the scenario analysis. It translates abstract model behavior into concrete, comparable financial outcomes by fixing all non-scenario features at their mid-tier median values and varying only the features that define each scenario. The baseline's negative ROI (−0.420) reflects the genuine industry reality that the median modern film does not recoup its cost, making it a meaningful and honest benchmark against which all three scenarios are evaluated.

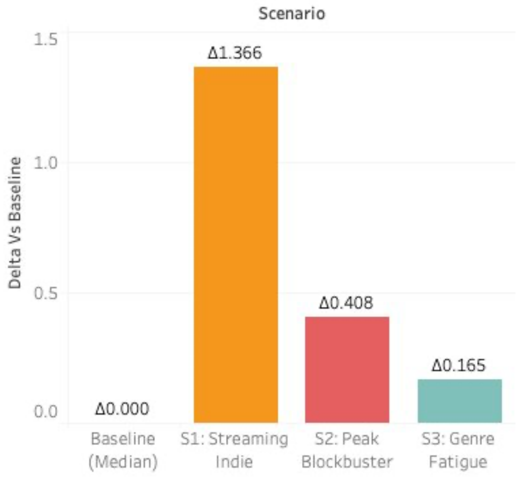
Key Insights

- S1: Streaming Era Indie achieves the highest predicted ROI at +0.946, far outperforming every other scenario. Low budget combined with 75th percentile popularity and engagement produces nearly a 95% return on investment, directly confirming that budget efficiency and audience reach drive financial success in modern cinema.

- S2: Peak Blockbuster returns −0.012 despite a 95th percentile budget, confirming the diminishing returns hypothesis from Q1: massive spending at the blockbuster tier barely reaches break-even, and the typical high-budget film fails to recoup its investment.
- S3: Genre Fatigue predicts −0.255 despite its vote average of 5.8 placing it in the 69th percentile of all films, meaning it is a genuinely above-average film by audience reception. That a well-reviewed film still predicts negative ROI under genre saturation conditions isolates genre fatigue as an independent financial risk factor beyond film quality alone.

Figure 24: ROI Improvement vs Baseline

Dashboard: DB6. Scenario Analysis



Description

A bar chart showing the ROI improvement (delta) of each scenario relative to the baseline, labeled with a Δ prefix. Bar height represents the absolute ROI gain over the median film.

Significance

By isolating the delta rather than the absolute ROI, this chart removes the baseline’s influence and focuses purely on how much each strategic choice improves outcomes relative to doing nothing differently. This is analytically useful because it answers a specific question studios care about: given that the typical film loses money, how much can intentional decisions move the needle?

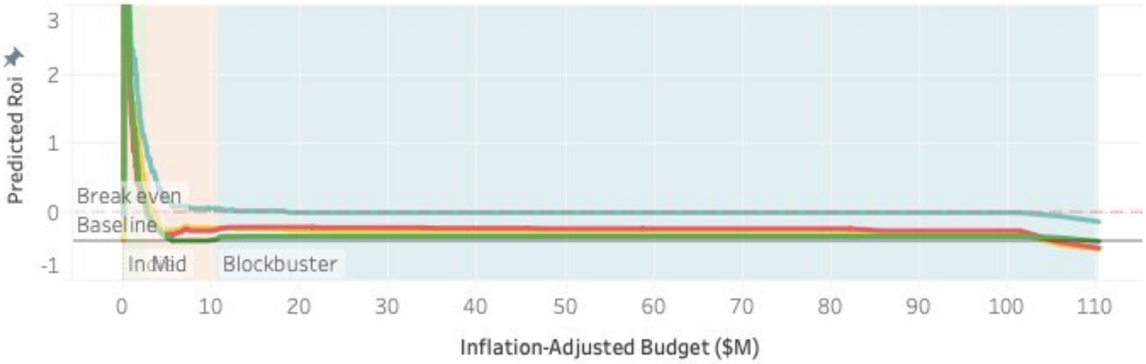
Key Insights

- S1 produces a delta of +1.366, the largest improvement of any scenario, confirming that the streaming indie strategy generates the most dramatic financial uplift relative to a typical film. The gap between S1 and the next-best scenario (S2 at +0.408) is more than three times wider, emphasizing how exceptional the low-budget, high-engagement combination is.

- S2 (Δ+0.408) and S3 (Δ+0.165) both improve on the baseline despite their absolute ROI values remaining negative, demonstrating that intentional feature selection meaningfully shifts predicted outcomes even in unfavorable conditions.
- The ranking S1 > S2 > S3 is consistent with the Q1 tier findings: budget efficiency (S1) outperforms high spending (S2) which outperforms audience fatigue conditions (S3), validating that the scenario model is aligned with the broader research conclusions.

Figure 25: Budget vs. Predicted ROI - Diminishing Returns

Dashboard: DB6. Scenario Analysis



Description

A multi-line chart plotting predicted ROI across the full range of inflation-adjusted production budgets for all four scenario feature profiles simultaneously. Each line holds non-budget features fixed at their scenario-specific values while sweeping budget from the 1st to 99th percentile of the dataset. Budget tier boundaries (Indie, Mid, Blockbuster) are indicated by colored shaded regions. Dashed horizontal reference lines mark break-even (ROI = 0) and the baseline prediction.

Significance

This chart is the most analytically rich visualization in the dashboard. By sweeping budget continuously across the same feature profiles used in each scenario, it reveals not just where each scenario sits in isolation but how ROI responds to budget changes under each set of conditions. This directly tests the diminishing returns hypothesis from Q1 and allows comparison of how aggressively ROI falls as budget increases for indie versus blockbuster feature profiles.

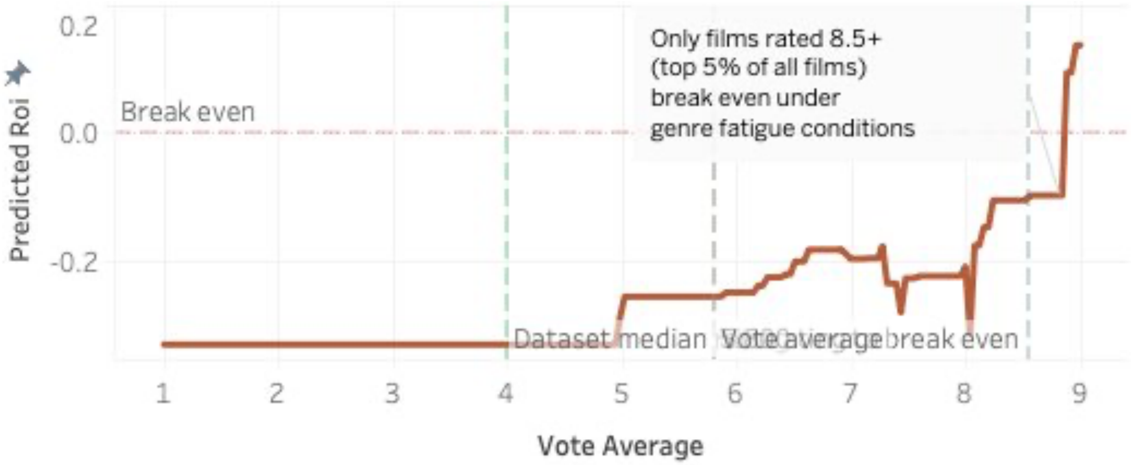
Key Insights

- All four scenario lines converge near break-even or slightly below zero once budgets exceed approximately $10M (the Indie-to-Mid boundary), confirming that the model consistently penalizes high spending regardless of which other features are present. Diminishing returns are not limited to blockbuster conditions, they are a structural feature of the model's learned relationship between budget and ROI.

- The S1 Indie line begins highest in the low-budget region and falls most sharply, demonstrating that the streaming era engagement features that produce exceptional ROI at low budgets lose their advantage almost entirely as spending increases past the Indie tier boundary.
- At extreme budgets beyond $100M, lines begin to diverge slightly and some dip below the baseline reference, suggesting the model identifies a threshold beyond which even favorable non-budget features cannot offset the ROI penalty of very high spending.

Figure 26: Minimum Rating to Break Even Under Genre Fatigue

Dashboard: DB6. Scenario Analysis



Only films rated 8.5+ (top 5% of all films) break even under genre fatigue conditions

Description

A line chart plotting predicted ROI against audience vote average (1.0–9.0) under S3 Genre Fatigue conditions, with all other S3 features held fixed (median budget, median popularity, 2022 release, runtime 120 minutes). Three vertical dashed reference lines are overlaid: the dataset median rating (4.0), the S3 scenario vote average (5.8), and the break-even rating threshold (8.56 - where predicted ROI first crosses zero). A text annotation in the upper right identifies the top 5% rating threshold required to break even.

Significance

This chart answers the most specific and actionable question posed by the scenario analysis: under genre saturation conditions, what level of audience reception is required just to break even? By expressing the answer as a precise rating threshold (8.56, top 5% of all films), it translates a model output into a concrete strategic implication. It also contextualizes the S3 scenario's vote average of 5.8 (in the 69th percentile) demonstrating that even well-reviewed films fall far short of the rating needed to overcome genre fatigue's financial drag.

Key Insights

- Predicted ROI is flat at approximately −0.33 for ratings from 1.0 to 4.0 (below the dataset median), indicating that the model does not differentiate meaningfully between poor and mediocre reception under genre fatigue conditions. Both produce equally poor financial outcomes.

- A step increase in predicted ROI begins around the 4.0–5.0 rating range, consistent with the model’s tree structure detecting a meaningful audience reception threshold at roughly the dataset median. Above this threshold ROI improves in a stepped fashion, but remains negative through ratings as high as 8.0.
- The break-even threshold of 8.56 is reached only in the top 5% of all films by audience rating. This is the chart’s headline finding: genre saturation conditions impose such a severe financial penalty that only exceptional, near-universally acclaimed films can overcome it, making genre fatigue a structural risk that quality alone is unlikely to solve.