Prepared by group 5

Group 5 Analytics

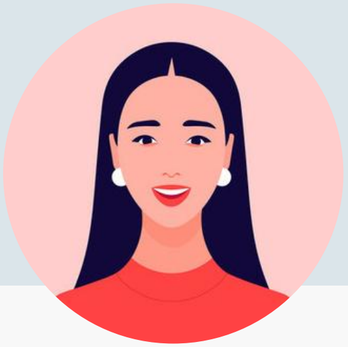
Turning Insights into Box-Office Gold

10 June, 2025

Technical Mentor: George Kamundia

Team Members

Erick Mauti
Statistical Sleuth



Marilyn Akinyi
Data Alchemist

Samwel Ongechi
Feature Forge Master

Isaack Onyango
Hypothesis Whisperer

Rose Miriti
Insight Illuminator

Rodgers Otieno
Story Teller

Introduction

This project aims to empower a new movie studio entering the film industry with no prior production experience with data-driven insights to guide its film selection and market entry strategy.

Primarily, the project aims to identify patterns in successful movies by analyzing existing industry data, focusing on key drivers such as genre performance, production budgets, audience and critic reception, and release timing.

These collectively minimize risks since the industry is very competitive and will improve profitability as a result of trategic dessicions based on trends

Business Context

Challenges for a New Studio:

- No prior production experience.
- High financial risk with film investments.
- Uncertainty about audience preferences.

Opportunities:

- Data reveals what works (genres, budgets, release timing).
- Competitive advantage by leveraging trends.

Business Context Cont'

Strategic business Questions to answer:

- Which genres have the best return on investment?
- What budget ranges balance cost and revenue most effectively?
- Do critic and audience ratings reliably indicate financial success?
- Is there an optimal release window for certain types of films?

Source of Data

- The Numbers (`tn.movie_budgets.csv`) -Financial performance metrics
- TMDb (`tmdb.movies.csv`) - Genre classification and audience sentiment
- IMDb (SQLite DB) - `movie_basics` e.g Titles, genre etc and ratings
- Rotten Tomatoes (`rt.movie_info.tsv`) - Supplementary analysis and genre validation
- Box Office Mojo (`bom.movie_gross.csv`) Backup for revenue and studio-level insights

Process Steps

1.Data Loading and Inspection

- Load all the Data
- Inspect to understand structure, dimesntions, data types and Quality

2.Data Preparation

Check shape for Each data set, column names, missing values, duplicates and summary statistics

3.Missing Data 30% Threhold Check

Check colums with high proportions of missing valies then drop or fill appropriately and drop

4.Column selection for focused Analysis

Select relevant columns that capture essential aspects of movie performance, audience engagement, and financial outcomes.

Process Steps Cont'

5.Data Cleaning

Use a function to strips whitespace from column names and string values from selected datasets rt_movies_df, tmdb_movies_df, and tn_budget_df

Standardize column names by converting letters to lower case, striping leading and trailing white spaces and replacing spaces with underscore.

6.Handling Missing Values

foreign_gross column in `bom_movie_df` and critics_consensus in `rt_movies_df` > 30% missing values threshhold

- FIll Missing categorical/text data with `"Unknown"` to retain meaningful category information
- Fill missing financial numeric columns (like `production_budget`, `worldwide_gross`, `profit`, and `roi`) with the median
- FIll missing other numeric columns (non-financial) with `0` to maintain numerical consistency and avoid errors during calculations

EDA1: Financial Performance and ROI Analysis

Financial performance summary

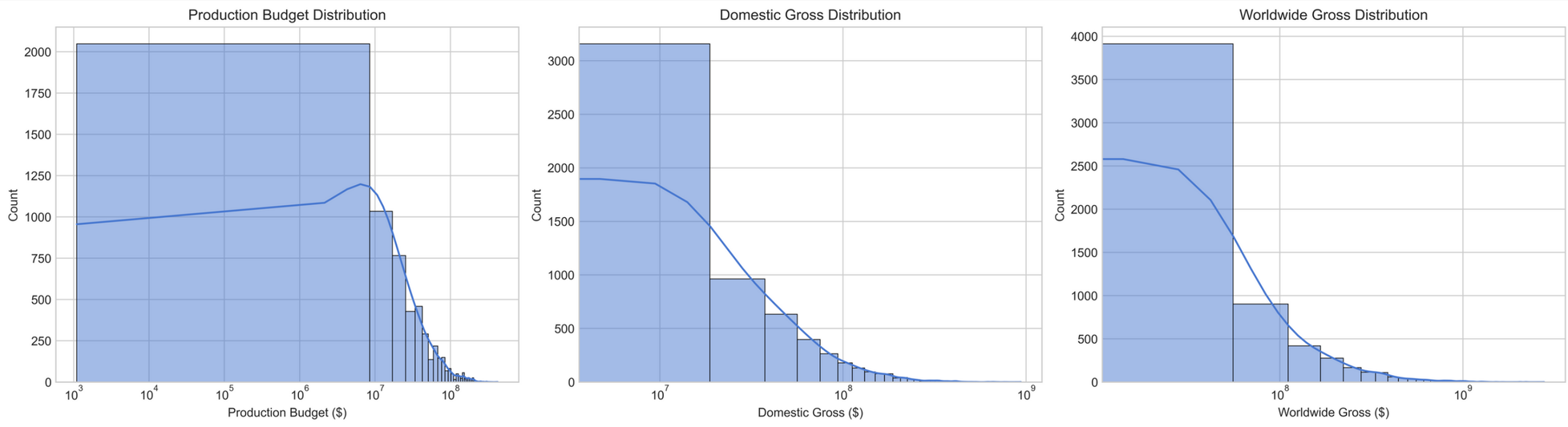
Summarized as:
ROI (Return on Investment) = (Profit ÷ Production Budget) × 100

...	production_budget	worldwide_gross	profit	ROI
count	6,227	6,227	6,227	6,227
mean	32,357,884	93,763,314	61,405,429	394
std	42,236,733	178,026,170	149,307,676	2,910
min	1,100	0	-200,237,650	-100
25%	5,000,000	4,569,218	-2,199,405	-49
50%	18,000,000	29,882,645	9,263,263	73
75%	40,000,000	98,800,000	62,115,664	279
max	425,000,000	2,776,345,279	2,351,345,279	179,900

Summary Explanation

- The huge variation in movie profitability and ROI.
- That while the average ROI is high, it's heavily influenced by a few extremely successful movies.
- A large number of movies operate at a loss or low profitability, especially in the lower quartiles

Data Analysis: Statistical Summary of Movie Financials



Key Insights

Blockbusters skew the data: The mean is much higher than the median, indicating a few massive productions inflate the averages.

- High-risk, high-reward industry: The wide standard deviation shows just how unpredictable returns are.

- Most films operate on mid-sized budgets: The 25th–75th percentile range shows typical budgets fall between $5M and $40M.

-Many films underperform: With a minimum gross of $0, it's clear that not all films succeed commercially.

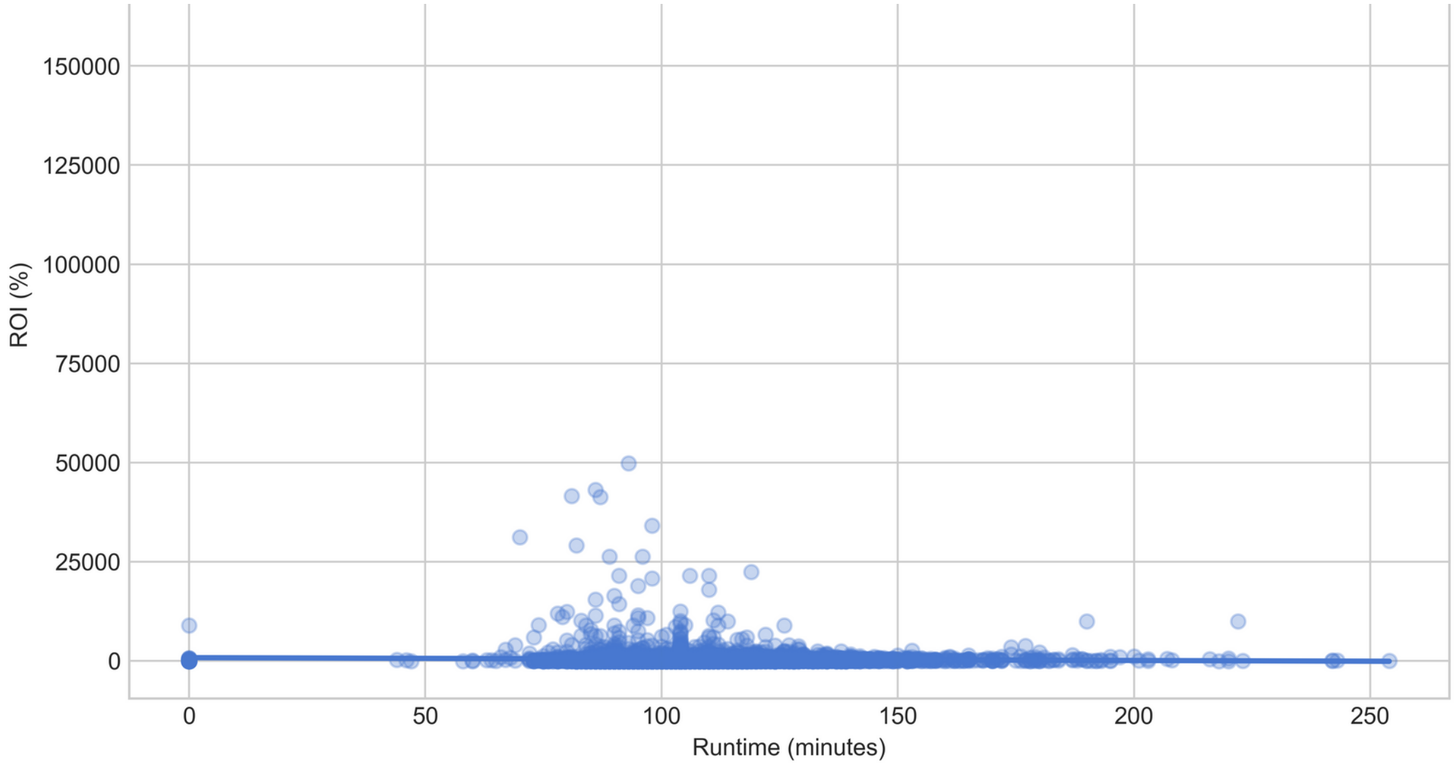
Data Analysis: ROI Vs Runtime

Correlation between runtime and ROI: -0.024

The correlation coefficient is -0.024, which is very close to zero.

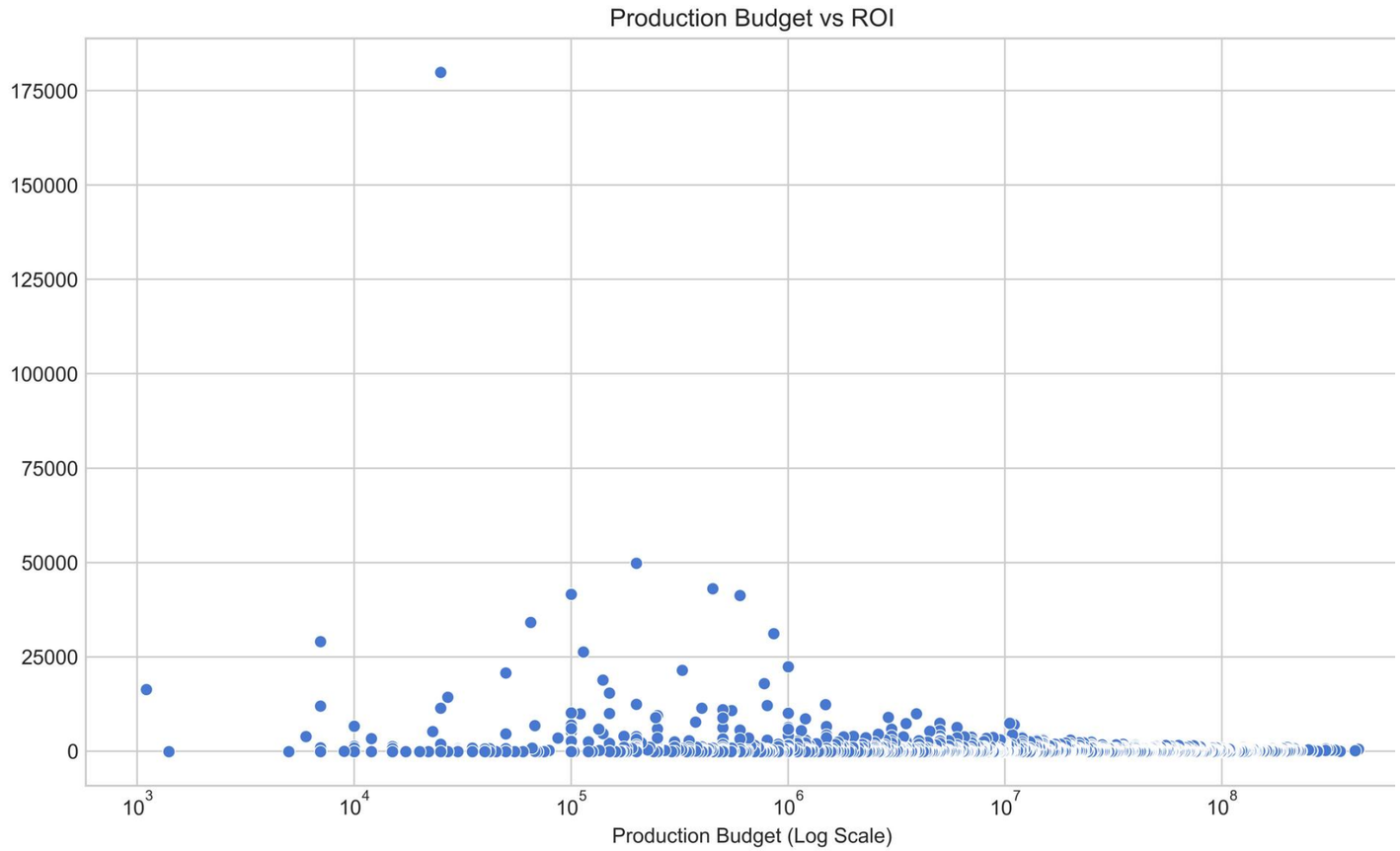
No linear relationship between the length of a movie (runtime) and its Return on Investment (ROI).

The slight negative value suggests a very weak tendency where longer runtimes might be associated with slightly lower ROI, but this effect is little and not statistically significant.



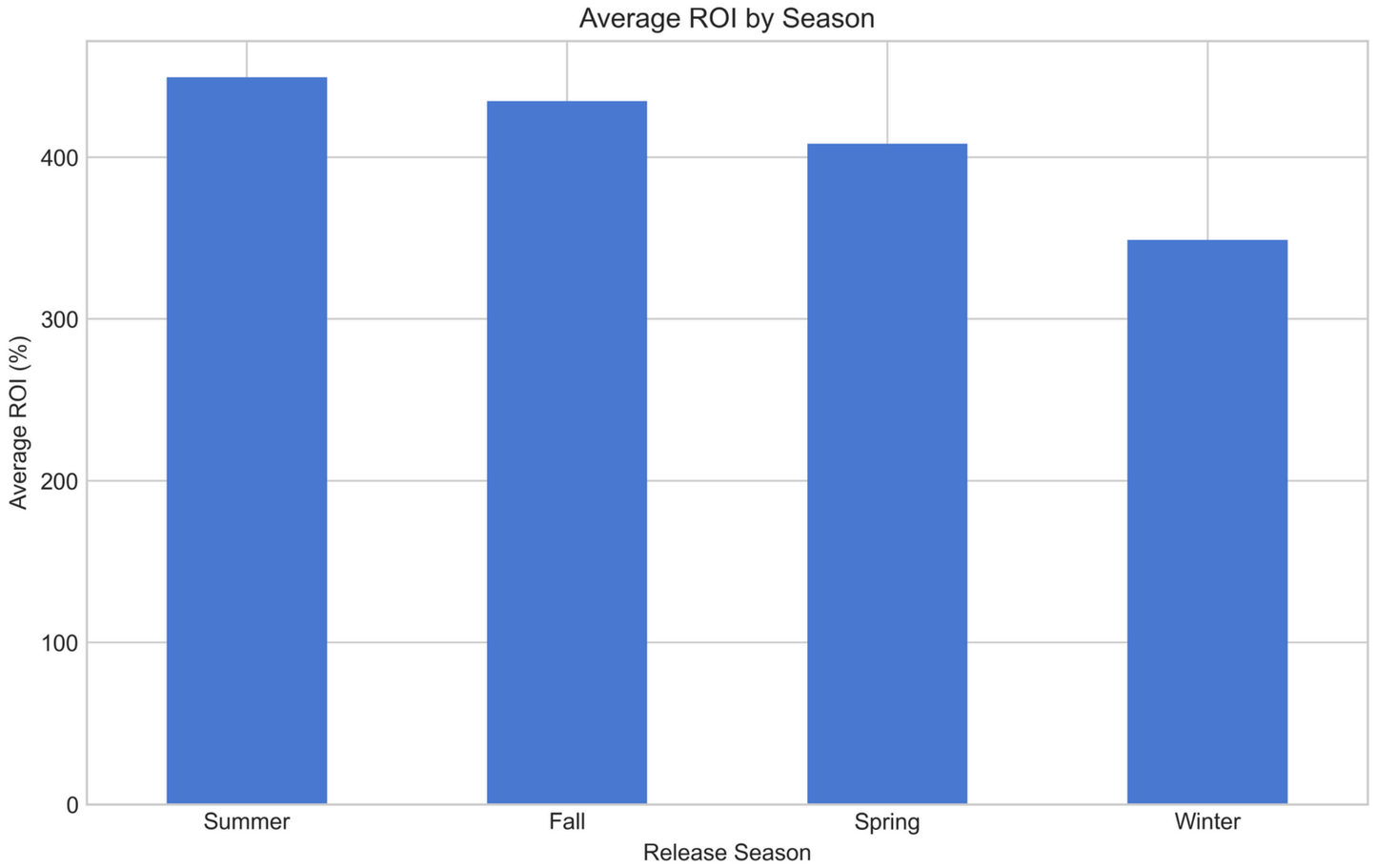
Data Analysis: ROI Vs Budget

- Most data points are tightly packed at lower ROI values, regardless of the production budget.
  - This indicates that very high returns are rare.
  - Most films tend to yield modest to moderate ROI, with only a few performing exceptionally well.



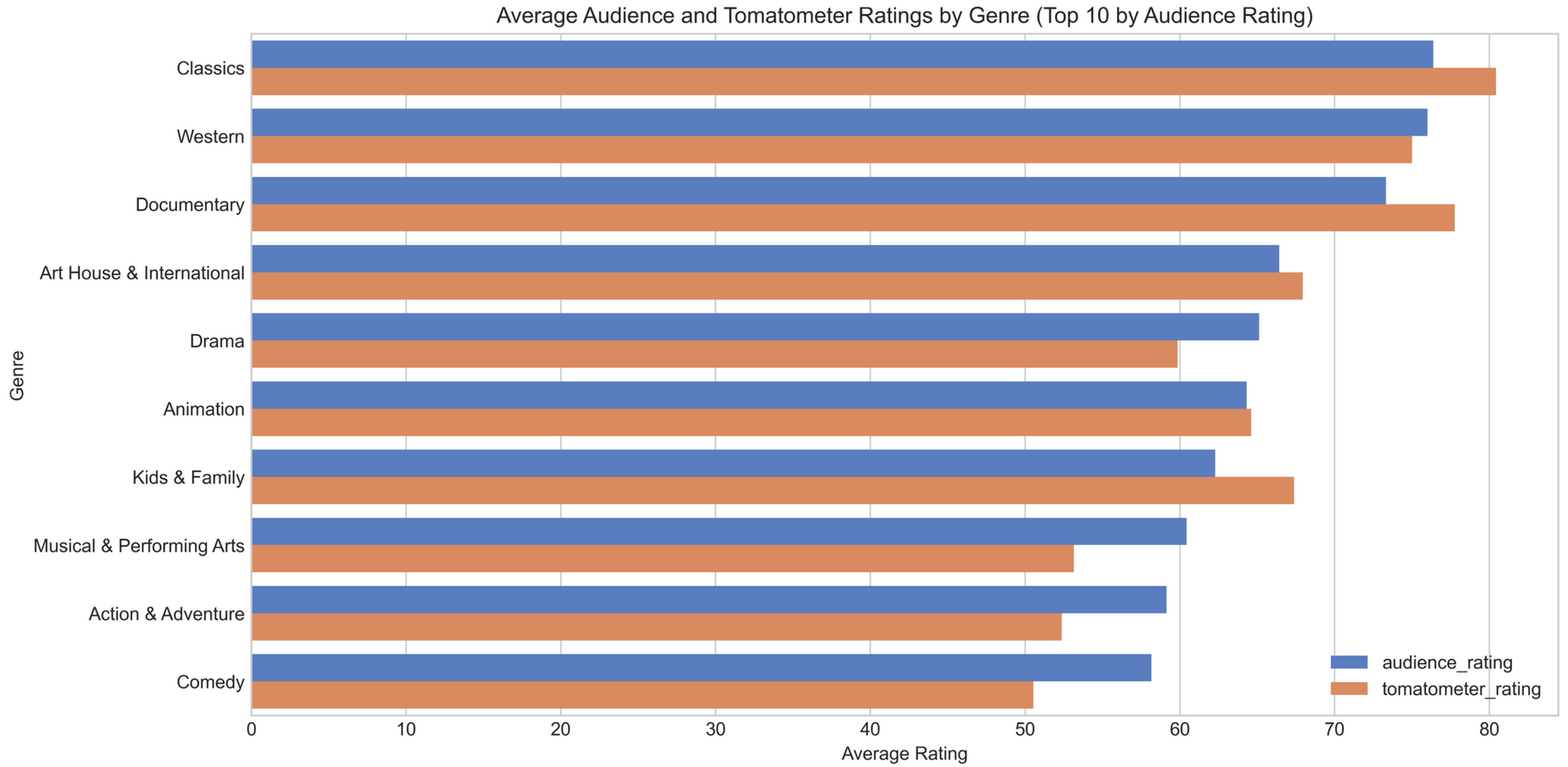
Data Analysis: ROI Vs Season

Summer and Fall yield slightly better ROI as compared to Spring and Winter Season . This information therefore informs the decison that movie releases should be Strategically released on these dates to enhance profitability.



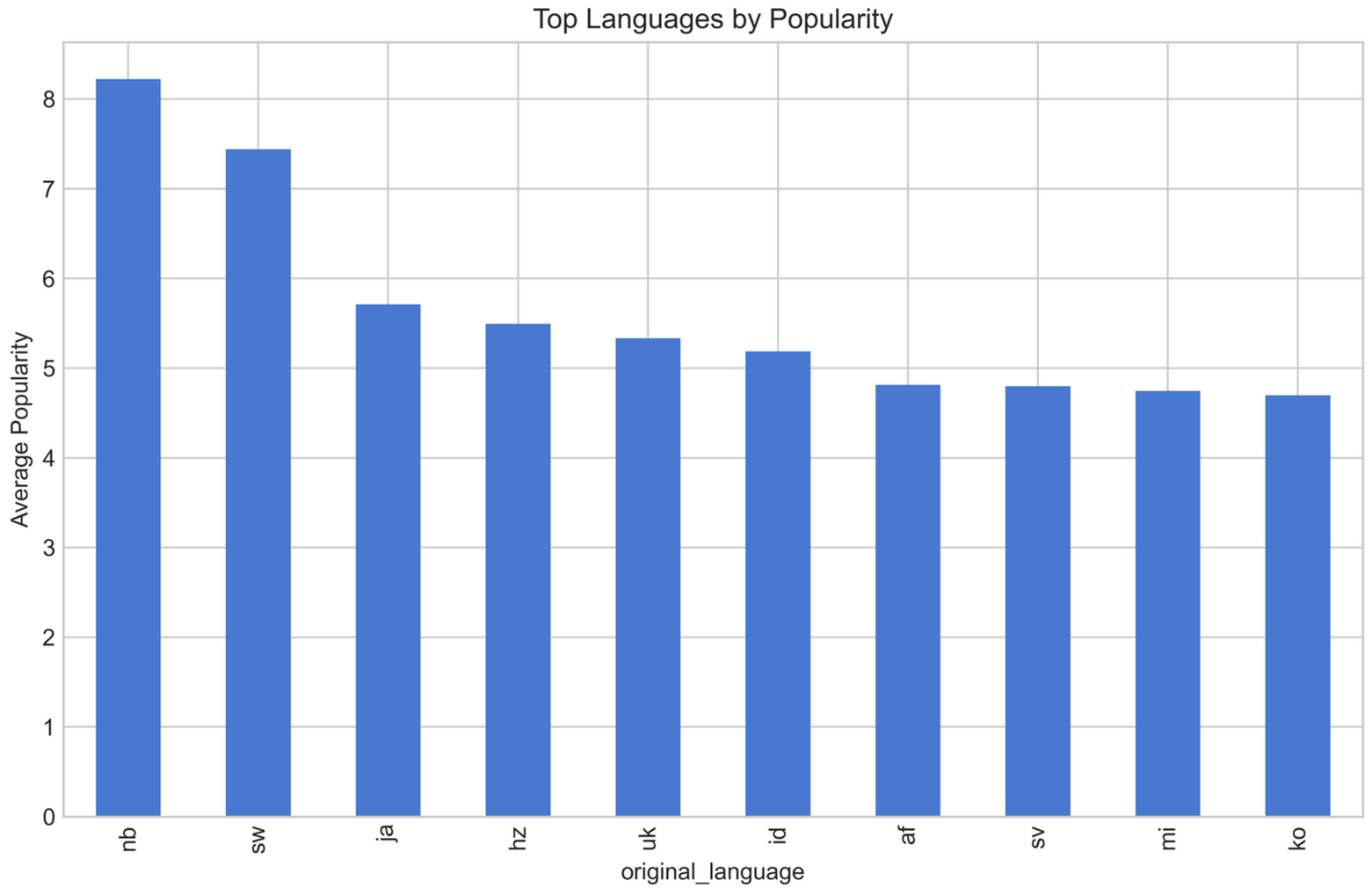
EDA2: Content Quality and Rating Impapct

Relationship between Average Audience and Tomatometer Ratings by Genre



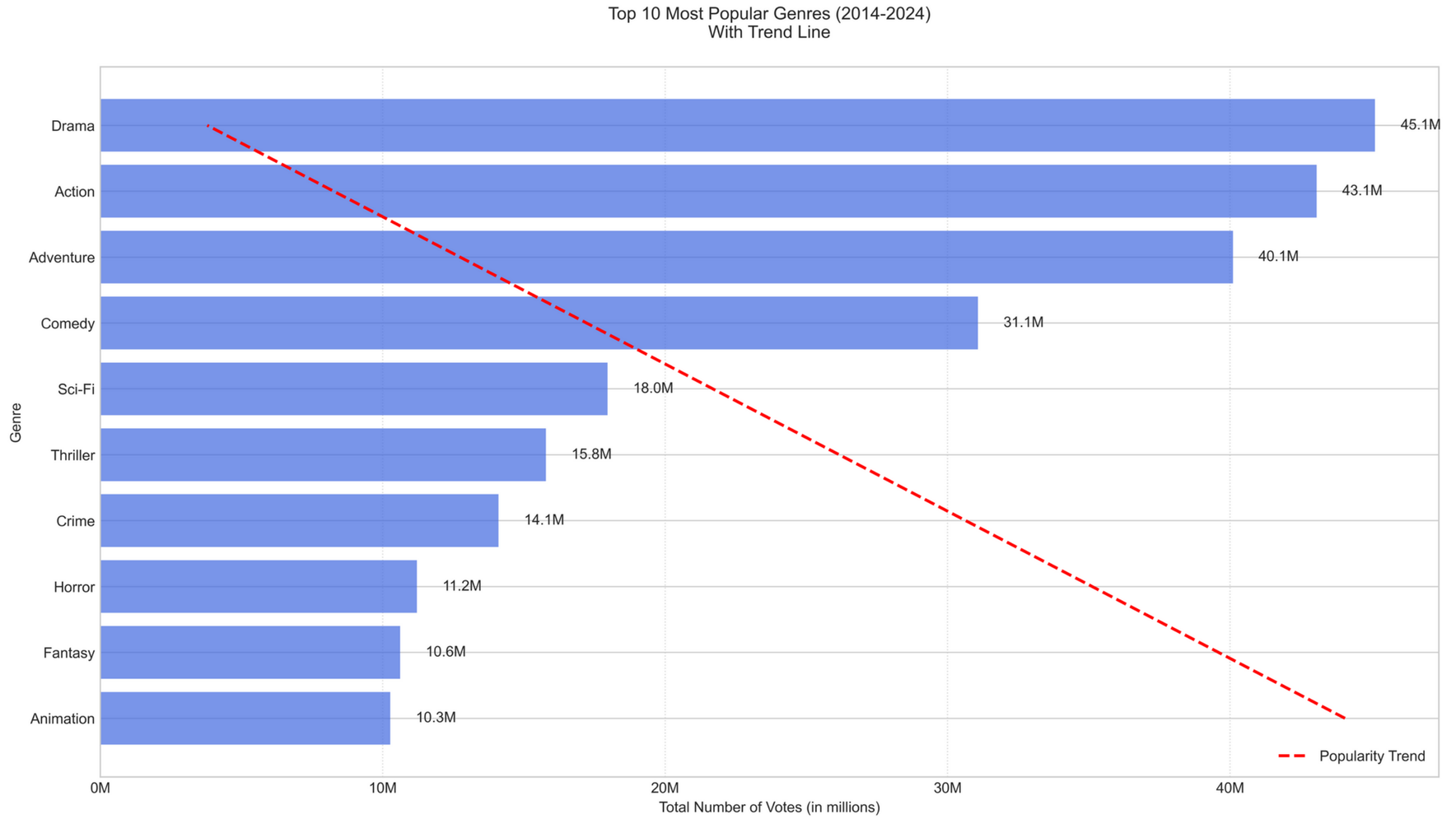
EDA3: Market Dynamics: Timing, Language, and Accessibility

Relationship between Top Languages and Popularity



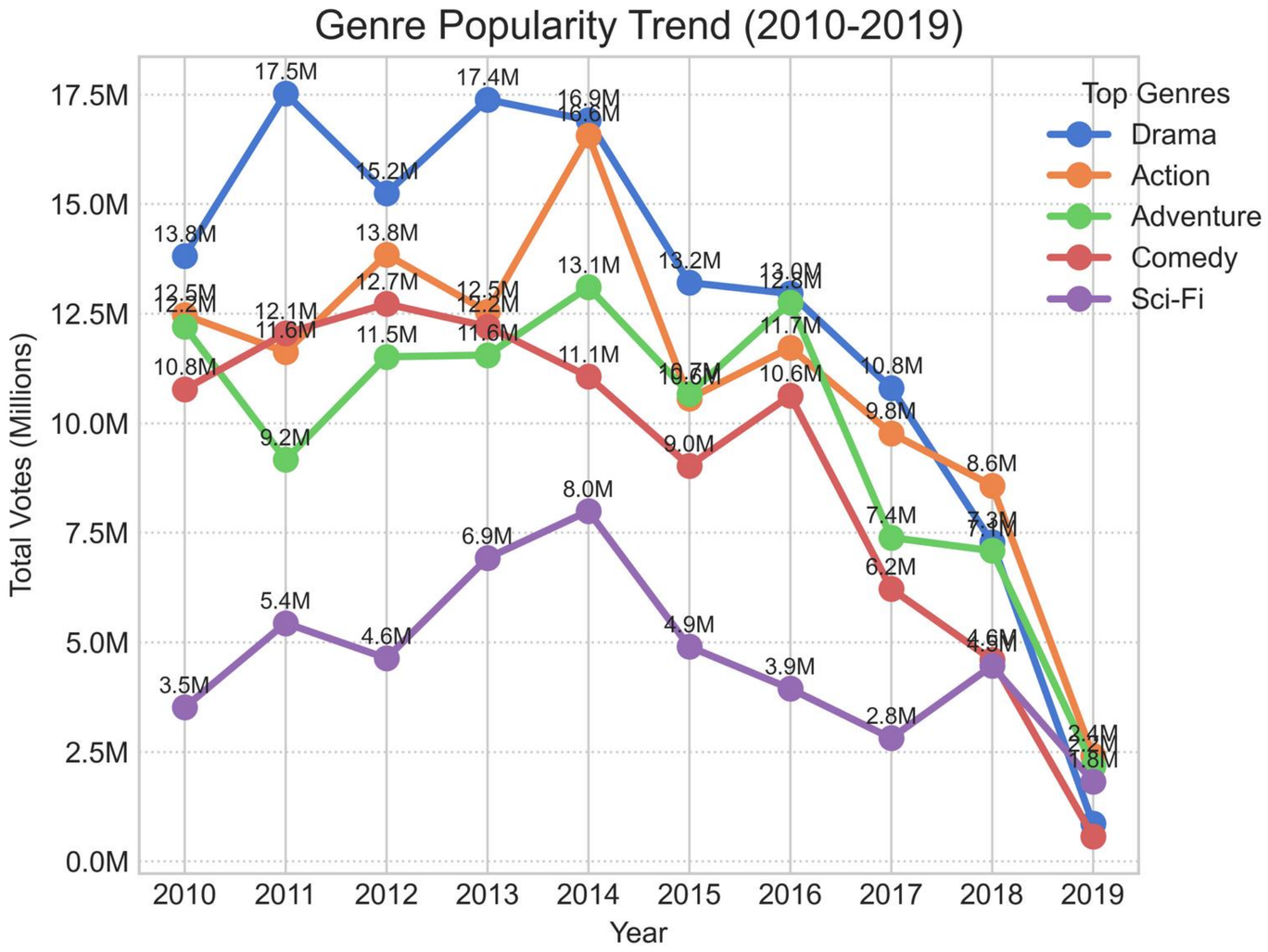
EDA3: Market Dynamics: Timing, Language, and Accessibility

Top 10 Most Popular Genres (2014–2024)



EDA3: Market Dynamics: Timing, Language, and Accessibility

Genre Popularity Trend



Hypothesis Testing Vs Strategic Validation

Hypothesis Test 1: Are there Significant Differences in ROI Across Genres?

H_0 - There is no significant difference in average ROI across movie genres.

H_1- There is a significant difference in ROI between at least one pair of genres.

ANOVA F-statistic: 1.91

p-value is less than 0.05, **reject the null hypothesis**.

There is a **statistically significant difference** in average ROI across movie genres.

At least one genre tends to perform differently better or worse in terms of ROI.

It is recommended to **invest in movies from genres such as Cult Movies, Horror , Special Interest and Doucumntary**, which tend to show **higher returns on investment** on average.

Hypothesis Testing Vs Strategic Validation

Hypothesis 2: Is There a Significant Difference in ROI Across Budget Categories?

H_0 - There is no significant difference in average ROI across budget groups.

H_1 - There is a significant difference in average ROI between at least one pair of budget groups.

ANOVA F-statistic: 15.63, p-value of 0.0000

Below the significance threshold of 0.05. This indicates:

- There is strong statistical evidence to reject the null hypothesis that the average ROI is the same across all budget groups.
- Budget size has a significant effect on ROI.
- Different budget categories (Low, Mid, High) show different average returns on investment.

This confirms that production budget meaningfully influences movie profitability.

Recommendation

-**Focus investments on movies with production budgets in the $20M–$60M range (Mid group).**

This mid-budget range tends to offer the best balance between cost and ROI, avoid very ;ow budget

Hypothesis Testing & Strategic Validation

Hypothesis 3: Is There a Relationship Between Ratings and Worldwide Gross Revenue?

H_0 - There is no linear relationship between movie ratings and worldwide box office revenue.

H_1 - There is a linear relationship between movie ratings and worldwide gross revenue.

Audience Rating vs Worldwide Gross: Correlation = 0.180, P-value = 0.0000 Critic Rating vs Worldwide Gross: Correlation = 0.164, P-value = 0.0000

Audience Rating vs Worldwide Gross Correlation = 0.180 indicates a weak positive correlation between audience ratings and worldwide gross revenue. As audience ratings increase, worldwide gross tends to increase slightly, but the relationship is not very strong.

- P-value = 0.0000 The very small p-value (less than 0.05) means this correlation is statistically significant, and the chance of this result occurring randomly is extremely low.

*

Hypothesis Testing Vs Strategic Validation

Hypothesis 3: Is There a Relationship Between Ratings and Worldwide Gross Revenue?

*__Critic Rating vs Worldwide Gross:__ Correlation = 0.164

There is also a weak positive correlation between critic (tomatometer) ratings and worldwide gross revenue, slightly less than audience ratings but still positive.

- P-value = 0.0000

This correlation is also statistically significant, showing that this weak relationship is unlikely to be due to chance.

Both audience and critic ratings are positively associated with how much money a movie makes worldwide. Audience ratings have a slightly stronger relationship with financial success than critic ratings in your data. However, since the correlations are weak (<0.2), ratings explain only a small part of the variation in worldwide gross.

Hypothesis Testing Vs Strategic Validation

Hypothesis 4: Do Summer Releases Earn More Than Non-Summer Releases?

H_0 - There is **no significant difference** in average worldwide gross between summer and non-summer movie releases.
H_1 - Movies released in summer earn significantly higher worldwide gross revenue.

Business Relevance:
This test informs strategic **release timing** decisions, helping producers and distributors maximize box office potential by targeting high-earning windows in the calendar
T-statistic: 4.34 P-value: 0.0000
Result: Statistically significant Interpretation: There is sufficient evidence to suggest that movies released in summer have a significantly different worldwide gross compared to those released in other seasons

Hypothesis Testing Vs Strategic Validation

Hypothesis 5a: Is There a Relationship Between Audience Ratings and Worldwide Gross?

H_0- There is **no correlation** between audience ratings and worldwide gross revenue.

H_1 - There **is a statistically significant correlation** between audience ratings and worldwide gross revenue.

Movies with higher audience ratings tend to perform better at the global box office.

Correlation Coefficient: 0.18 - This indicates a weak but positive correlation, meaning that higher audience ratings tend to be associated with higher worldwide gross revenue, although this relationship is not strong.

P-value: 0.0000- Since the p-value is well below the common significance threshold of 0.05, we can conclude that the correlation is statistically significant.

Overall Interpretation

There is strong statistical evidence to support a positive association between audience ratings and worldwide gross revenue. However, the weak correlation suggests that audience ratings alone account for only a small portion of the variation in box office revenue. This implies that while audience perception influences financial success, many other factors also contribute significantly to a movie’s worldwide earnings.

Business Recommendations

1. Recommendation on ROI & Financials

It is recommended to Invest in high-ROI genres that is Cult Movies, Horror, Special Interest , Documentary and Classics, within a moderate budget range ($20M–$60M) and aim for runtimes between 100–120 minutes.
Cult Movies, Horror, Special Interest, Documentary and Classics

2. Recommendation on Ratings & Content Quality

Prioritize well-written scripts that are likely to resonate with both audiences and critics to boost revenue potential in the genres that resonate most with the audience that is Classics, Western, Documentary, Art House & International and Drama.

3. Recommendation on Market Dynamics & Timing

Release films during summer/holiday windows and explore multilingual productions to increase global box office reach.

Thank you

Q & A?