Crisp DM Non-Technical Report

Business Understanding.

Overview:

The global film industry is undergoing constant transformation, driven by changing audience preferences, streaming disruptions, and fierce competition. A new movie studio aims to enter this dynamic space and seeks data-driven guidance to make smart, profitable decisions. Understanding the factors behind successful films is critical for positioning the studio competitively.

Challenges:

Despite the abundance of data on movies, key challenges include identifying the right metrics for "success," accounting for changes in audience tastes over time, and combining data from diverse sources (e.g., IMDb, Rotten Tomatoes, The Numbers). The complexity increases when isolating the influence of individual contributors like directors and actors.

Proposed Solution:

We propose analyzing a comprehensive dataset of films over the last 15 years, focusing on genres, directors, and actors. By evaluating box office performance, ratings, and popularity, we aim to uncover patterns and build predictive models. These insights will help the studio invest in the most promising genres and talent.

Brief Conclusion:

With targeted data mining and analysis, the studio can significantly reduce guesswork in creative and financial decision-making. This project lays the foundation for a data-driven strategy to maximize box office success.

Problem Statement

The film industry lacks a reliable, data-informed approach to determine which genres, directors, and actors are most likely to succeed at the box office. This uncertainty hinders new studios from making confident production investments.

Objectives

i. Mine and merge film data from IMDb, Rotten Tomatoes, and The Numbers.

ii. Analyze customer ratings, popularity, and box office performance.

iii. Build predictive models to estimate film success and retention value.

iv. Provide actionable recommendations to guide studio investments.

Key Research Questions.

What are the top 5 highest-grossing movie genres in the past 15 years?

Which director appears most in the top 10 grossing films since 2008?

Which region has had the strongest box office performance?

Conclusion.

By leveraging historical data, our team can offer data-driven recommendations that minimize risk and maximize profitability for the new studio. These include guidance on genre focus, director partnerships, and target markets.

Data Understanding

Several datasets have been provided for analysis ie IMDB,Rotten tomato,Numbers etc.Which are in different formats csv,Tsv,db and have data related to movies

Data Sources

(I). tn.movie_budgets.csv: Contains financial data (production budget, domestic/worldwide gross, profit).

(II). IMDb SQL Database (im.db):

a) movie_basics: Movie details (title, year, runtime, genres).

b) directors: Director information.

c) persons: Personal details of industry professionals.

d) movie_ratings: Audience ratings and votes.

e) movie_akas: Alternate titles and regional information.

Prior to each statistical test, we will need to perform some data preparation, which could include:

Filtering out rows with irrelevant values

Transforming data from codes into human-readable values

Filtering data to transform it from numeric to categorical

Creating new columns based on queries of the values in other columns

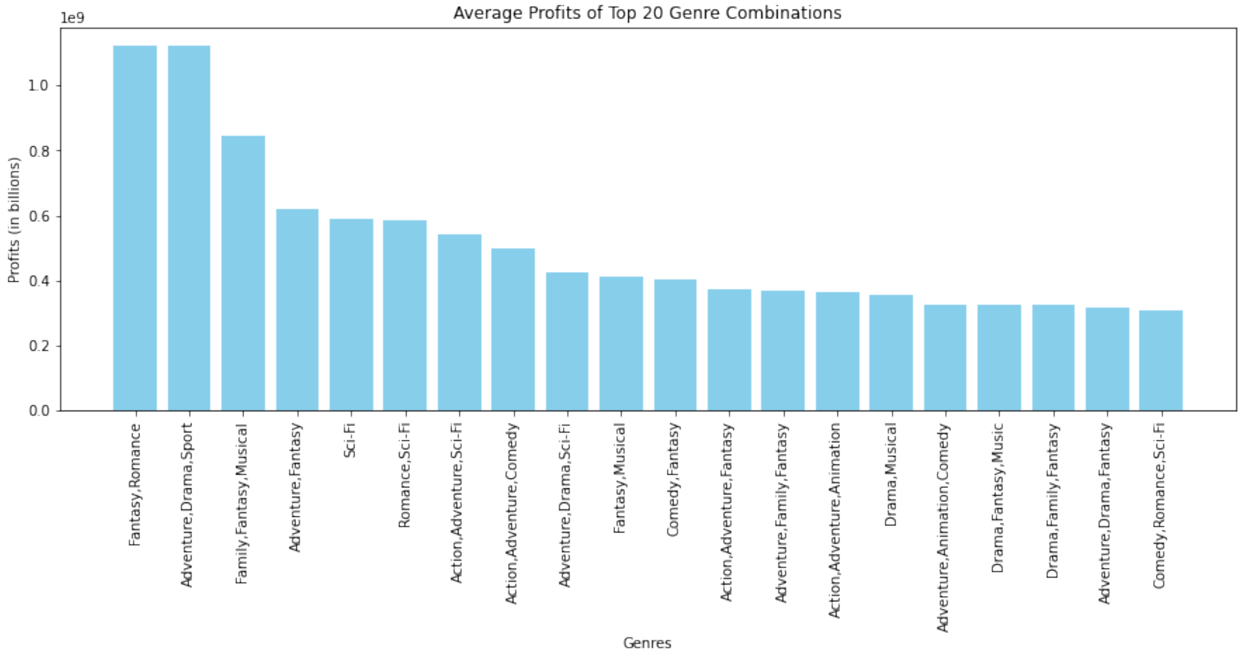
Merging datasets

KEY VISUALIZATION AND INSIGHTS

1. Genre Profitability

📊 Avg. Profit by Top 20 Genre Combos

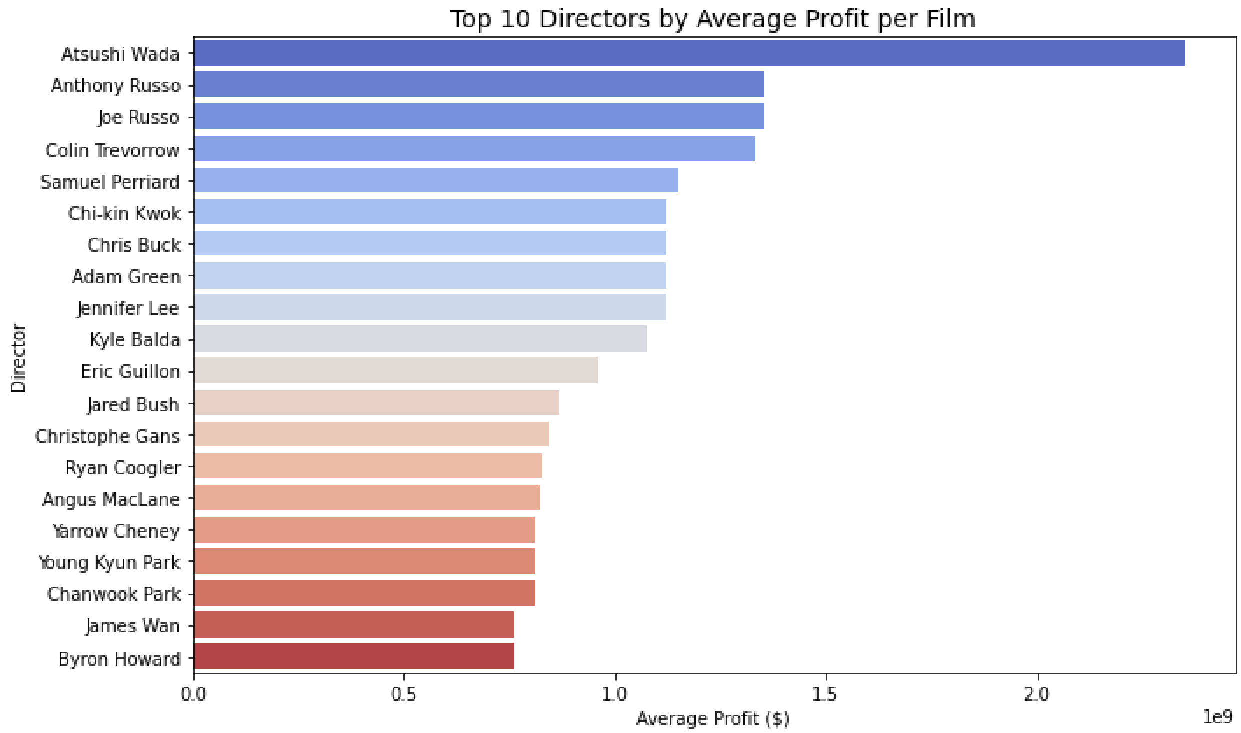
- 🎯 *Action/Adventure/Sci-Fi* lead in profitability
- 👻 *Horror* yields high returns on low budgets
- 🎭 *Drama* shows high variability in outcomes



🎥 Director Performance

📈 Profit vs. Director Ratings

- 📌 Highly-rated directors often drive better profits
- 📉 Some mid-rated directors deliver unexpected success
- 🔁 Positive link between reputation and financial results

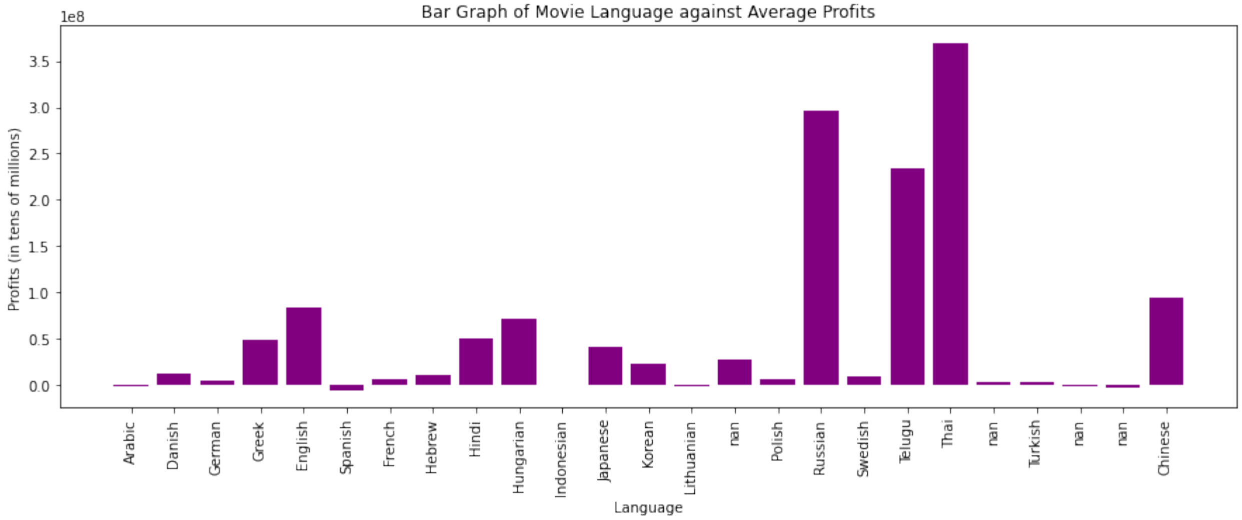


🌍 Language Profitability

🌐 Profit by Language

- us English dominates revenue share
- 🌏 Thai, Telugu & Russian show niche profit potential

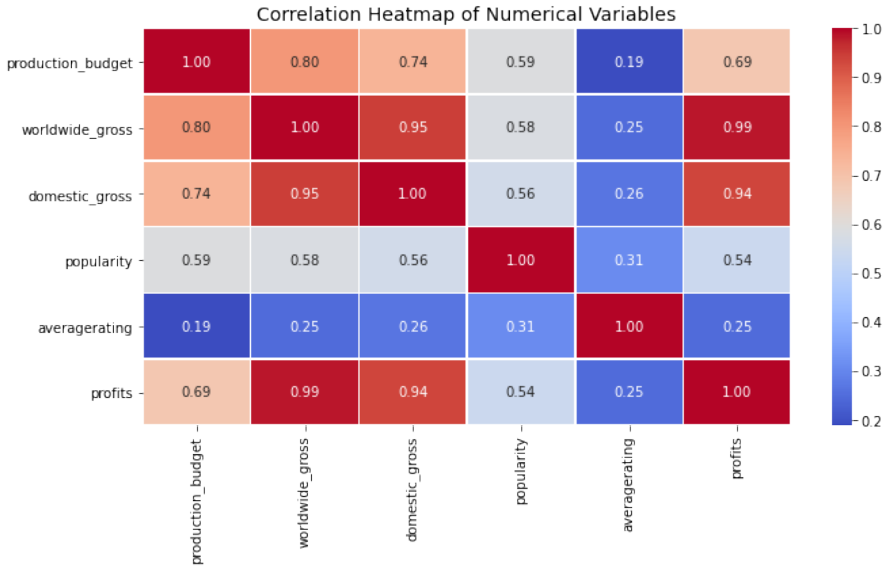
- 🧩 Dubbing/Subtitling may unlock underutilized markets



📈 Popularity vs. Profit

🔥 Heatmap: Popularity vs. Profit

- 🔗 Strong correlation *(r = 0.72)*
- ⚠ Films with *popularity < 30* rarely make a profit

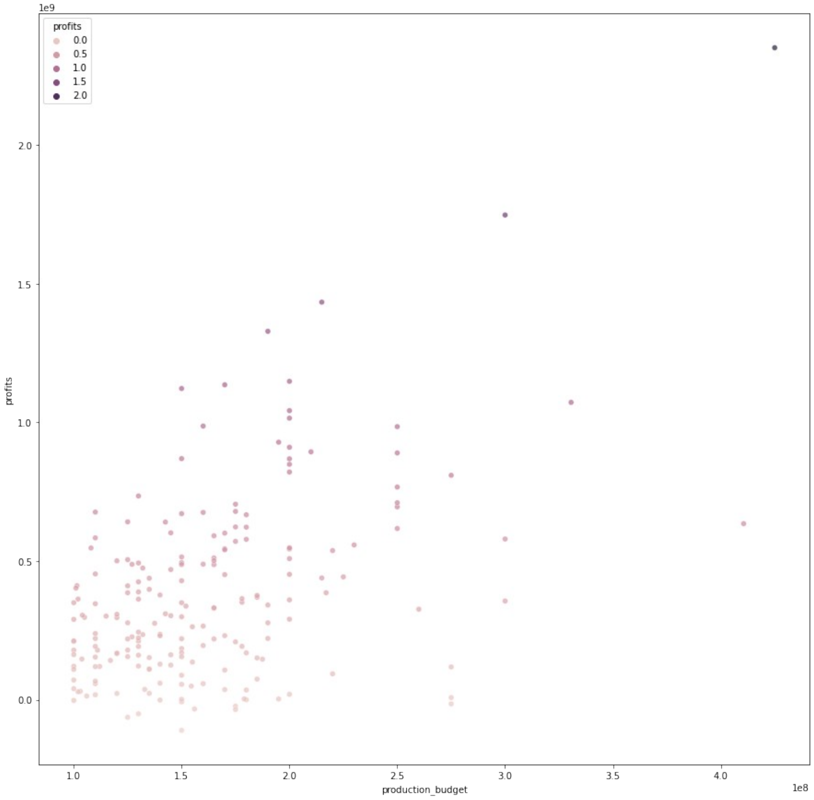


	production_budget	worldwide_gross	domestic_gross	popularity	averagerating	profits
production_budget	1.00	0.80	0.74	0.59	0.19	0.69
worldwide_gross	0.80	1.00	0.95	0.58	0.25	0.99
domestic_gross	0.74	0.95	1.00	0.56	0.26	0.94
popularity	0.59	0.58	0.56	1.00	0.31	0.54
averagerating	0.19	0.25	0.26	0.31	1.00	0.25
profits	0.69	0.99	0.94	0.54	0.25	1.00

💰 Budget Efficiency

📉 Budget vs. Profit Scatter

- 🚫 Diminishing returns beyond *$150M budgets*
- 🎃 *Horror/Comedy* stand out as low-cost, high-return genres



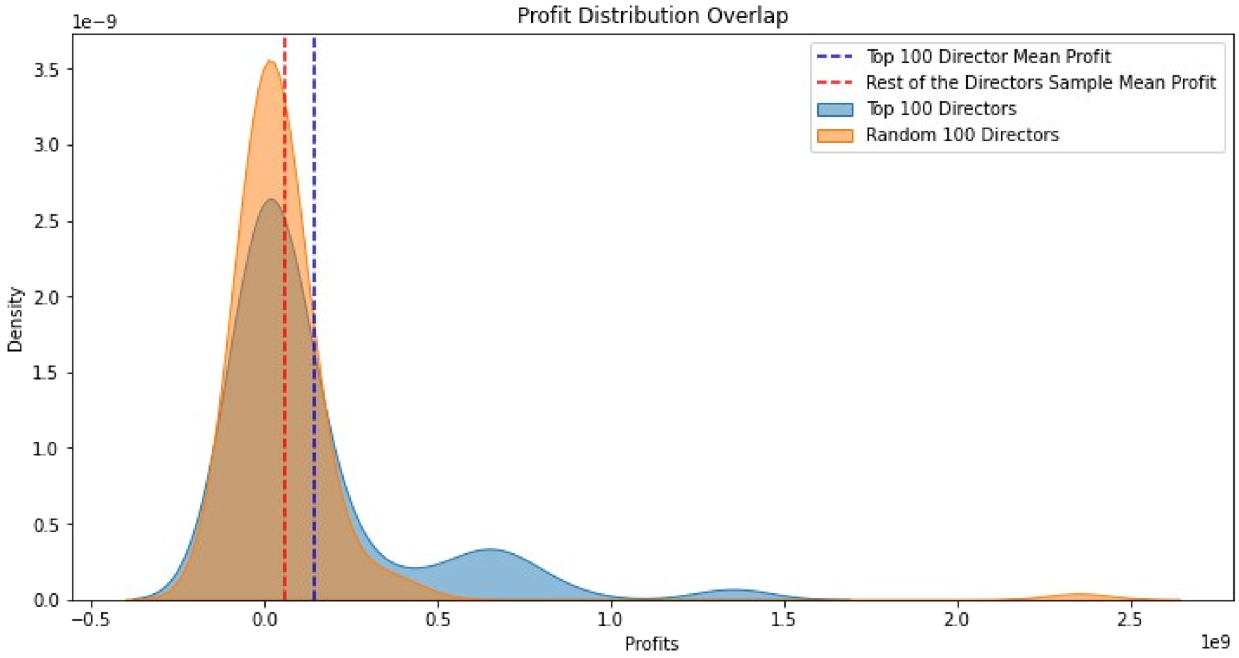
<u>Hypothesis Testing, Conclusions and Recommendations Section:</u>

Hypothesis Test 1: Top-Rated Directors vs. Lower-Rated Directors

Null Hypothesis: Movies with top-rated directors generate equal profits compared to those with lower-rated directors.

Alternative Hypothesis: Movies with top-rated directors generate more profits.

T-statistic = 2.25, p-value = 0.02, Cohen’s d = 0.32



Conclusion:

The p-value < α (0.05) indicates statistically significant evidence to reject the null hypothesis.

The positive T-statistic confirms that movies by top-rated directors have higher average profits.

Effect Size: Cohen's d = 0.32 (small-to-medium effect) suggests a meaningful but modest difference in profitability.

Implications:

Hiring top-rated directors correlates with increased profits, but the impact is not overwhelming.

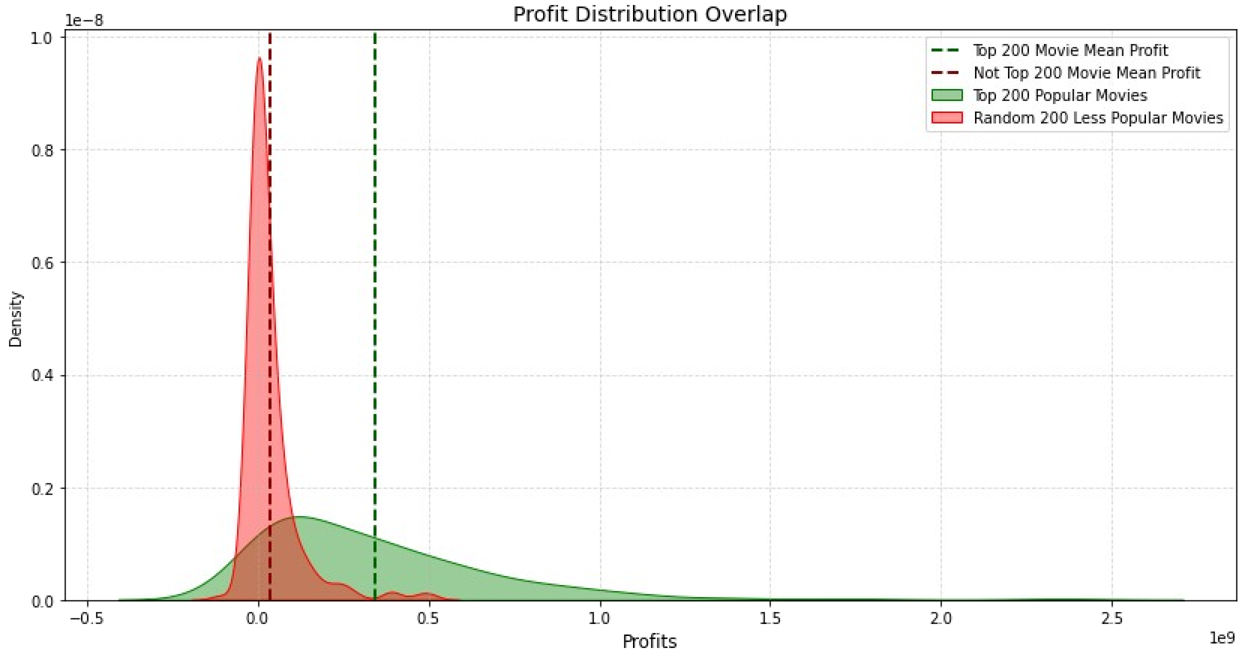
Other factors (e.g., genre, budget, marketing) may mediate the relationship between director reputation and profits.

Hypothesis Test 2: Popular Movies vs. Less Popular Movies

Null Hypothesis: More popular movies do not make more profits than less popular movies.

Alternative Hypothesis: More popular movies make more profits.

T-statistic = 12.26 , p-value = 1.6e-29 , Cohen's d = 1.23



Conclusion:

The extremely low p-value (≈0) provides evidence to reject the null hypothesis.

The large T-statistic and Cohen's d = 1.23 (large effect) indicate a strong practical significance.

Implications:

Popularity is a critical driver of profitability.

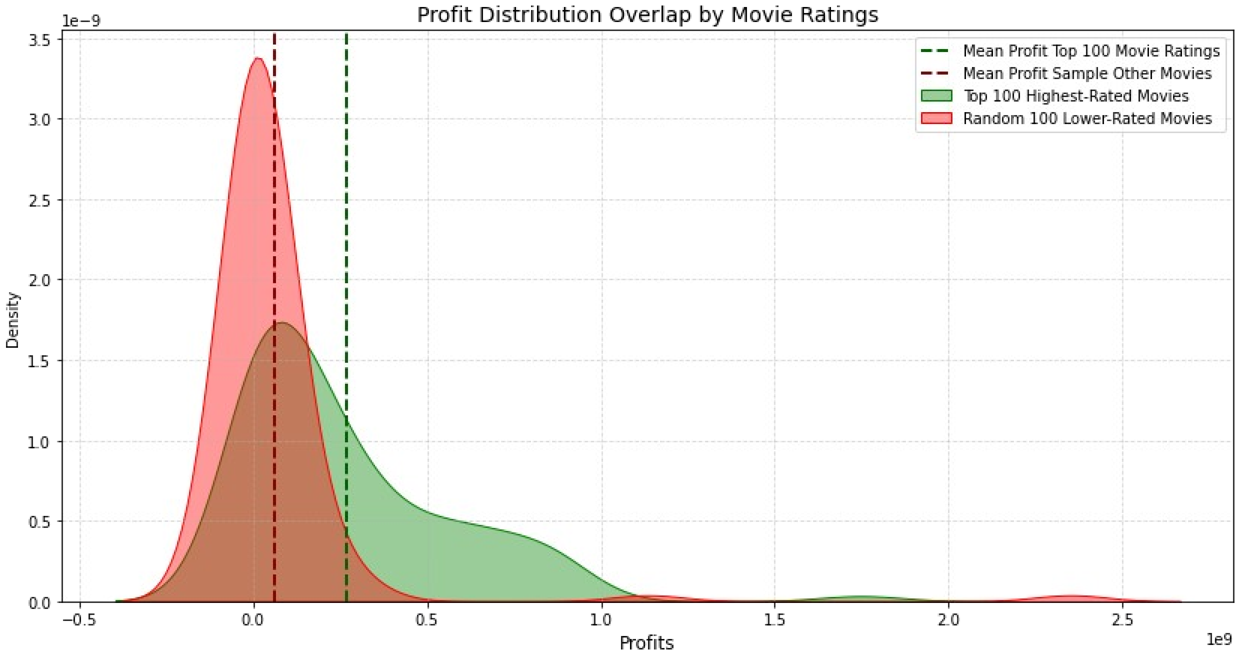
Strategies to boost popularity (e.g., marketing, star actors, viral campaigns) are likely to yield substantial returns.

Hypothesis Test 3: High-Rated Movies vs. Lower-Rated Movies

Null Hypothesis: Movies with higher ratings do not make more profits than lower-rated movies.

Alternative Hypothesis: Higher-rated movies make more profits.

T-statistic = 5.2, p-value = 4.99e-7, Cohen's d = 0.73



Conclusion:

The p-value < α confirms statistical significance, rejecting the null hypothesis.

Cohen's d = 0.73 (medium-to-large effect) highlights a substantial difference in profitability.

Implications:

Critical acclaim (high ratings) correlates with higher profits, though the effect is weaker than popularity.

Quality and audience reception matter, but they may not guarantee blockbuster success alone.

Comparative Insights

Popularity Reigns Supreme:

Popularity (Test 2) has the largest effect size (d=1.23), making it the strongest predictor of profits.

Example: A $100M budget film with high popularity could generate 2–3x more profit than a similar-budget film with low popularity.

Ratings vs. Directors:

Ratings (Test 3) have a stronger effect (d=0.73) than directors (Test 1: d=0.32).

This suggests that audience reception (ratings) is more impactful than director reputation alone.

Strategic Priorities:

Invest in Popularity: Allocate resources to marketing, social media campaigns, and casting popular actors.

Balance Quality and Appeal: High ratings enhance profitability, but prioritize projects with mass-market potential.

Director Selection: While top directors add value, their impact is secondary to popularity and ratings.

Limitations & Future Work

Confounding Variables: Budget, genre, and release timing may influence both popularity/ratings and profits.

Interaction Effects: Explore how directors/actors and ratings jointly influence popularity (e.g., do top directors/actors boost ratings?).

Final Recommendations

Maximize Popularity: Prioritize marketing and trend-aligned content.

Leverage Critical Acclaim: Use high ratings to enhance long-term revenue (e.g., awards, streaming rights).

Hire Strategically: Pair top directors with high-potential scripts to amplify their modest profit impact.