DSCI 510 Final Project Report

Keys to Box Office Success: An Analysis of Movie Performance Factors and Critic-Audience Opinion Dynamics

Team Member (Solo): Chaoyang (Sunny) Yin
USC ID: 7522481537
USC Email: sunnyyin@usc.edu
GitHub repository link: https://github.com/ChaoyangYin/DSCI-510-Project-Fall-2025

Short Description

This project aims to investigate the factors that are associated with the box office success of movies, with a particular focus on the opinion dynamics between professional film critics and general audience. To gather sufficient data for this analysis, this project utilized data from two online archives, TMDB and OMDB, via their free APIs. Datasets collected from these sources contain information on movie attributes (budget, revenue, genre, runtime, release date), and ratings from different sources (Rotten Tomatoes, Metascore, IMDb rating). The analysis part of this study examines how these factors, along with derived parameters such as return of interest (ROI), drive financial success, and whether disagreement between critics and audiences affect box office results. By examining correlations and trend across genres and release time, the project uncovers insights into the movie industry.

Data

Sources
1. The Movie Database (TMDB) API:
Provided core movie attributes: title, release date, budget, revenue, genres, runtime, vote average, vote count, and IMDb ID.
2. OMDb API:
Queried using IMDb IDs obtained from TMDB. Provided ratings data: Rotten Tomatoes (RT) Tomatometer (professional score), Metascore, and IMDb rating (audience-driven user votes).
No web scraping or manual data collection was performed. All data retrieval was via HTTP requests in Python.
Note on inflation adjustment: To enable fair comparison across years, budget and revenue values were adjusted to 2025 dollars using approximate annual CPI rates (U.S. Bureau of Labor Statistics).
Source: https://www.bls.gov/cpi/tables/supplemental-files/historical-cpi-u-202412.pdf

Number of Data Samples
After combining the results of 2 separate runs of API calls to both sources (OMDB has free daliy limit of 1000) and basic filtering with request (2010-present, English laguage, minimum 50 votes and revenue>500000), the final raw dataset contains 2000 unique movies released between 2010 and 2025 in English.

Data Cleaning, Analysis & Visualization

1. Process Definition

Data Cleaning

The data cleaning pipeline in clear_data.py was designed to transform raw API response in JSON file into a structured and analysis-ready dataset in csv format. The first step of this pipeline loads multiple raw JSON files (2 in this project as result of 2 separate API runs) and combines them into one database of 2000 unique movies. In step 2, separate functions extracted TMDB attributes (title, budget, revenue, release_date, genres, runtime, vote_average) and OMDb ratings (Rotten Tomatoes Tomatometer as rt, IMDb rating as imdb, Metascore as meta). In the next step, the release_date string was converted to datetime formated and derived 'year' and 'month' columns for ease of analysis. The multi-value genres field (list of dictionaries) was converted to a list of genre names and exploded into separate rows. Finally, invalid rows were removed in order to maintain basic cleaness while preserving sample size (movies with budget or revenue equal to zero, those with missing either rt score or imdb score, those without a year, and those with ROI > 100), and the resulting structured dataset was saved to data\processed for subsequent analysis. In the scope of this study, the final cleaned data has 1811 movies, and exploded into 4812 rows with genre.

Analysis

The pipeline in run_analysis.py perfroms analysis on the cleaned dataset. First, the processed dataset is loaded into a DataFrame, and derived features are generated, including ROI (revenue/budget), adjusted budget, revenue, and ROI using CPI inflation to 2025 equivalent, critic-audience gap (RT vs IMDb), and averaged professional (RT + Metascore) vs audience (IMDb + TMDB) scores for more robust analysis. In step 2, a correlation matrix is computed for a list of key variables to identify linear relationships between budget, revenue, ROI, and ratings. In step 3, genre related statistics are aggregated (mean revenue, ROI, ratings, gap per genre, sorted by revenue) for to reveal genre-specific performance patterns. In step 4, yearly trends are examined using movie-level data (mean revenue, ROI, ratings by release year) to reveal long term temporal trends. In step 5, monthly data is analyzed (mean revenue, ROI by month) to identify optimal release periods. Additionally, top-performing movies by revenue are listed with key metrics. Finally, key insights are printed for a brief summary, highlighting patterns in commercial success and opinion dynamics. It is worth noting that in step 2, 4 and 5 the duplicate rows are dropped to maintain unique by TMDB ID to ensure one row per movie and avoid bias from the exploded genres structure.

Visualization

The visualization pipeline in visualize_results.py generates and saves a set of plots using the analysis result to illustrate key findings from the analysis and support the the goals of this project. First, the processed csv files saved from analysis with respect to the subdivided topics (yearly and monthly trend, correlation matrix, genre statistics). In step 2, a correlation heatmap is created from the pre-computed correlation matrix to visualize linear relationships among key variables. In step 3, a bar plot displays average ROI by genre using aggregated genre statistics. In step 4, line plots depict nominal and inflation-adjusted revenue trends over years using yearly aggregated data.

In step 5, a bar plot shows average critic-audience gap by genre to highlight opinion dynamics patterns. In step 6, a bar plot illustrates monthly seasonality in average revenue to identify optimal release periods. In step 7, a horizontal bar plot presents the top factors most strongly correlated with ROI and adjusted ROI (absolute correlation). All plots are formatted with clear titles, labels, and legends, and saved as high-resolution PNG files to the results/ folder for inclusion in the report.

2.Hypothesis / Premise & Conclusions

Initial Hypothesis/Premise

The initial premise of this project was that box office success is driven by multiple factors including budget, genre, release timing, and especially opinion dynamics between professional critics and general audiences. The hypothesis from me was a movie's commercial success might be more strongly associated with the budget than the with ratings, as majority of the paying viewers do not participate in voting. Additionally, significant disagreement between the critics and audience might potentially harming commercial performance, possibly due to audience suspicion of paid promotion or manipulated reviews.

Findings and Conclusions

The correlation analysis partially confirmed the initial hypothesis. Among factors correlated with ROI, revenue, budget, and the critic-audience gap emerged as the three strongest components, consistent with expectations that scale in finance plays a dominant role in commercial performance. However, ratings remained important, with professional critic scores showing a slightly higher correlation with ROI than audience-driven metrics by the average— a finding that was somewhat surprising given the assumption that paying viewers' preferences would carry greater weight. This suggests that critic approval may still influence perceived quality or marketing effectiveness, even in an era where audience word-of-mouth and user ratings are prominent.Within the four original rating parameters, Rotten Tomato Tomatometer exhibited the best strength in predicting financial success.

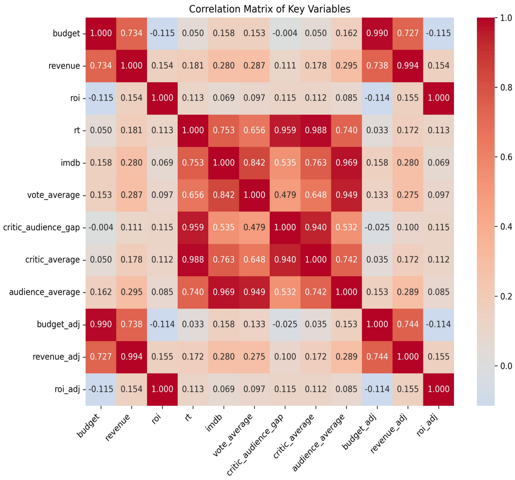


Figure 1. Correlation Matrix Heatmap

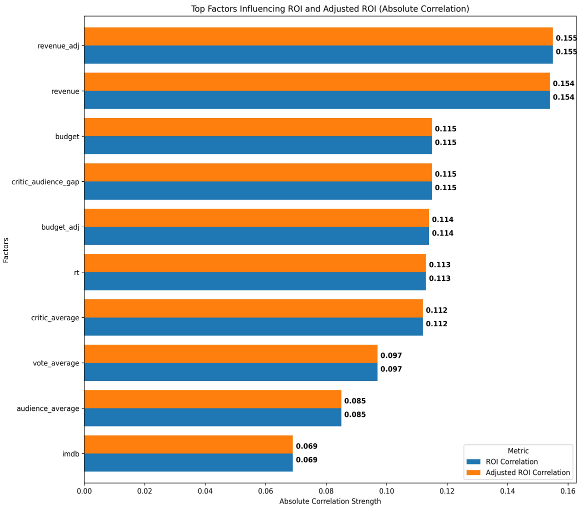


Figure 2. Top Factors influencing ROI by correlation

Another discovery is that adjusting for inflation make minimal impact on the relative strength factor on ROI. Almost all correlates remained consistent in both nominal and adjusted ROI. However, temporal analysis of revenue over years made a distinction, showing adjusted revenue higher than nominal, and converges towards 2025. A huge decrease is also discovered at 2020, likely signifying the impact of the pandemic.

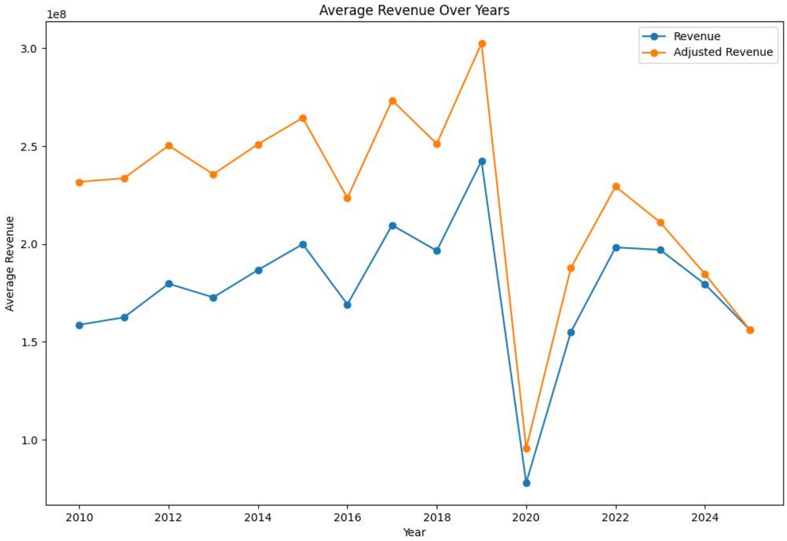
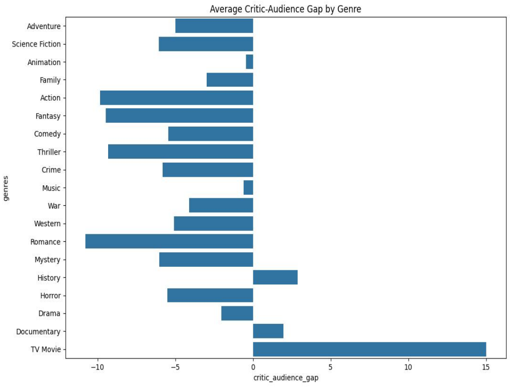


Figure 3. Average Revenue Over Years

Genre analysis revealed notable differences in critic-audience opinion dynamics. Professional critics awarded higher average ratings to TV movies, history, and documentary films, while audiences favored romance and action genres. Additionally, ROI analysis highlighted horror and mystery as the most profitable genres, achieving the highest average returns on investment.



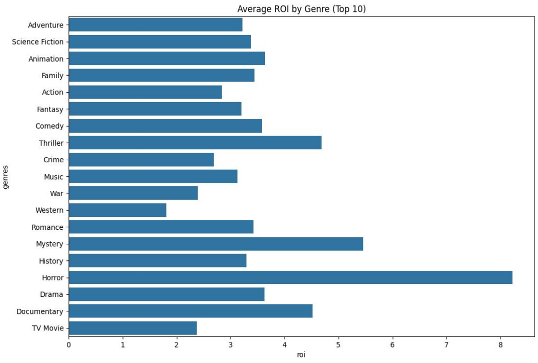


Figure 4. Average Critic-Audience Gap by Genre

Figure 5. Average ROI by Genre

Changes from Original Proposal

1. The original proposal emphasized using TMDB and OMDb APIs with a focus on Rotten Tomatoes (both critic and audience) for analysis. Due to inconsistent availability of separate Rotten Tomatoes audience scores in OMDb, IMDb rating (audience-driven user votes) was used as a reliable substitute.
2. To account for budget inflation across years, monetary values were adjusted to 2025 dollars using U.S. CPI rates, providing a more accurate view of real ROI trends over time.

3. Clarified exploratory analysis (not prediction) and specified dataset (2,000 movies, 2010–2025), and incorporated missing data handling.

Mention of Future Work

Future work could incorporate machine learning models (e.g., linear regression or random forest) to perform actual prediction on ROI based on budget, genre, ratings, and release month. Additional data sources, such as marketing spend, runtime, star influence, or social media sentiment, could be integrated to improve predictive accuracy and explore causal relationships beyond the current correlational analysis.