Analysis of Factors Influencing Movie Box Office Performance

Yahia Gaber, Ahmed AbdelMaboud, Omar Desouky, Manar Maher, Nardin Ashraf, Youssef Khaled

***Abstract*—This paper presents a data-driven analysis of factors influencing movie box office success using a dataset collected through automated data scraping. The study follows a complete data analysis pipeline including preprocessing, exploratory data analysis, visualization, linear regression, and clustering. The primary objective is to evaluate the relative influence of budget, cast popularity, and IMDb ratings on financial performance, analyze seasonal release effects, and compare critics' and audience ratings across genres.**

***Index Terms*—Data Analysis, Exploratory Data Analysis, Linear Regression, Clustering, Movie Analytics**

I. Introduction

The film industry represents a high-risk, high-reward environment where financial success depends on multiple interacting factors. Understanding which variables most strongly influence box office performance can provide insights for producers, studios, and analysts. This study investigates financial, temporal, and perceptual factors using structured movie data and statistical analysis techniques.

The research questions addressed are:

1) What factors most strongly predict box office success?
2) Do seasonal release patterns influence financial performance?
3) Is there a significant difference between critics' and audience ratings across genres?

II. Data Collection

Movie data was collected using an external movie database API (TMDb + OMDb) through automated data scraping. The dataset includes financial metrics, release dates, ratings, popularity indicators, and genre classifications. Due to incomplete reporting, some records contained missing financial values, which were handled during preprocessing.

III. Data Preprocessing

Preprocessing steps included:

- Removal of records with missing critical variables (e.g., revenue for regression analysis)
- Data type normalization for numerical and date fields
- Feature normalization for regression and clustering analysis

Financial variables such as budget and revenue were not imputed to avoid introducing bias.

IV. Exploratory Data Analysis

Exploratory analysis was conducted to examine distributions, correlations, and anomalies.

TABLE I: Correlation of Selected Features with Worldwide Revenue

Feature	Correlation with Revenue
Worldwide Revenue	1.000
Budget	0.785
Mean Cast Popularity	0.371
IMDb Rating	0.244

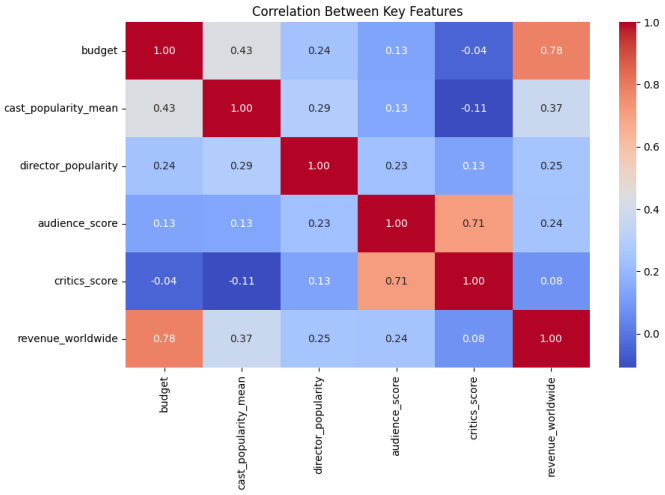

Fig. 1: Correlation Heatmap

Insight 1: Budget is the strongest predictor of box office success with the highest correlation with worldwide revenue between the other factors.

TABLE II: Average Worldwide Revenue by Release Season

Season	Average Worldwide Revenue
Summer	0.0987
Winter	0.0800

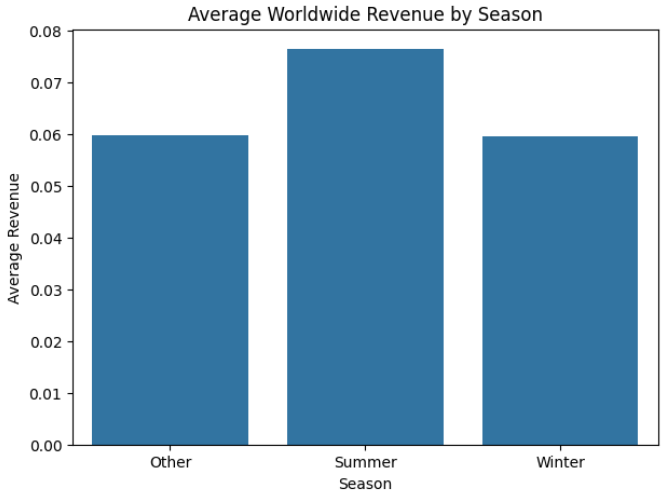

Fig. 2: Revenue according to season

Insight 2: Summer movies perform better financially than winter releases.

TABLE III: Critics and Audience Ratings by Genre

Genre	Critics Rating	Audience Rating	Difference
War	0.743	0.636	0.107
Action	0.667	0.564	0.102
Romance	0.680	0.585	0.094
Crime	0.695	0.602	0.093
Adventure	0.672	0.587	0.085
Science Fiction	0.664	0.580	0.084
Mystery	0.663	0.580	0.083
Fantasy	0.674	0.591	0.083
Thriller	0.656	0.575	0.081
Comedy	0.651	0.572	0.078
History	0.726	0.655	0.070
Western	0.632	0.578	0.055
Drama	0.717	0.667	0.050
Family	0.662	0.613	0.050
Horror	0.582	0.536	0.047
Animation	0.719	0.682	0.036
Music	0.694	0.692	0.002
Documentary	0.762	0.766	-0.004
TV Movie	0.656	0.764	-0.109

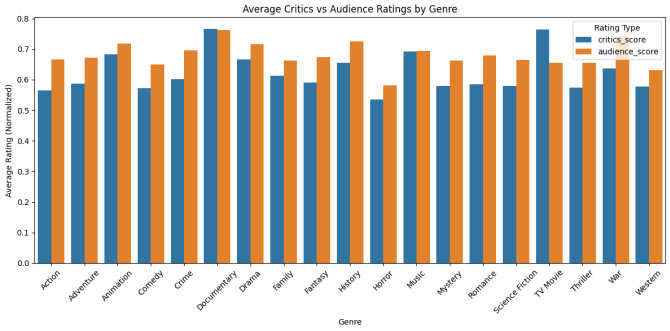

Fig. 3: Audience score vs Critics' score across genres

Insight 2: Audience score is almost always higher than critics' score.

V. Visualization Analysis

Visualizations were used to support pattern discovery and hypothesis formulation.

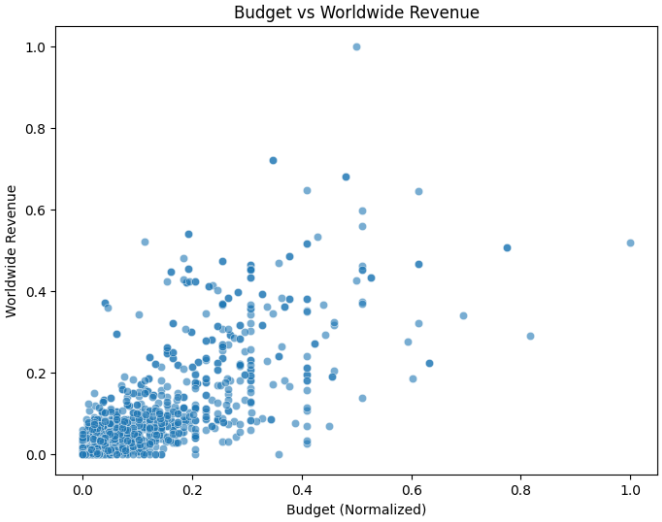

Fig. 4: Worldwide Revenue vs Budget

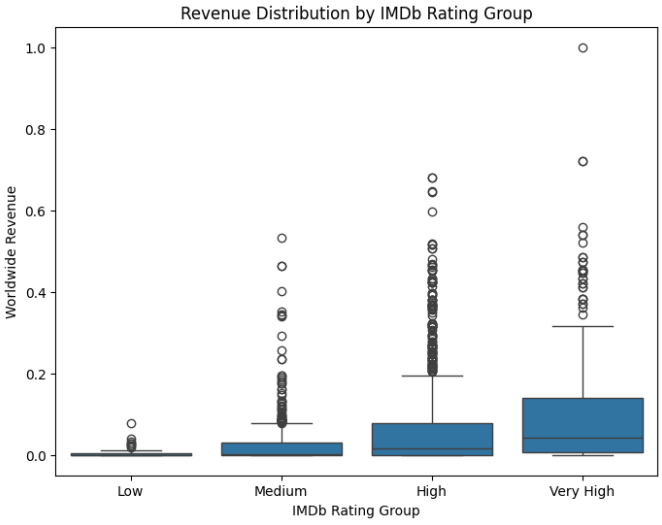

Fig. 5: Revenue Distribution by IMDb Rating Group

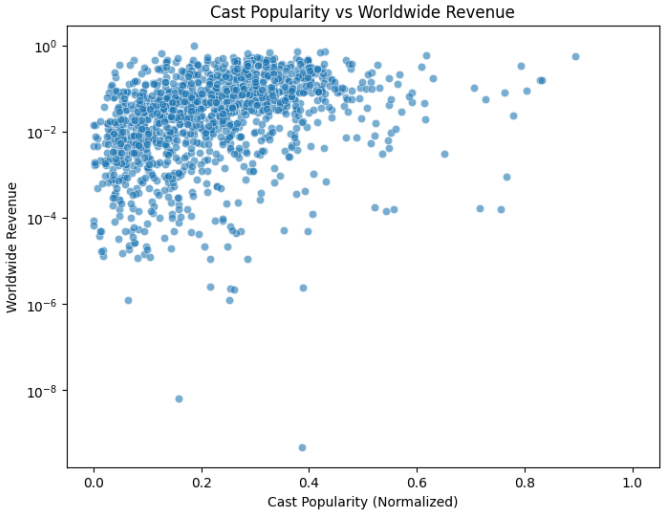

Fig. 6: Worldwide Revenue vs Cast Popularity

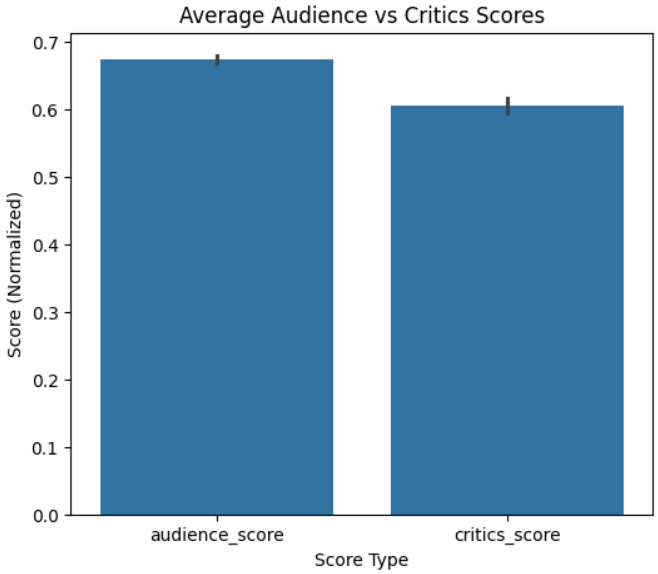

Fig. 8: Average Audience vs Critics Scores

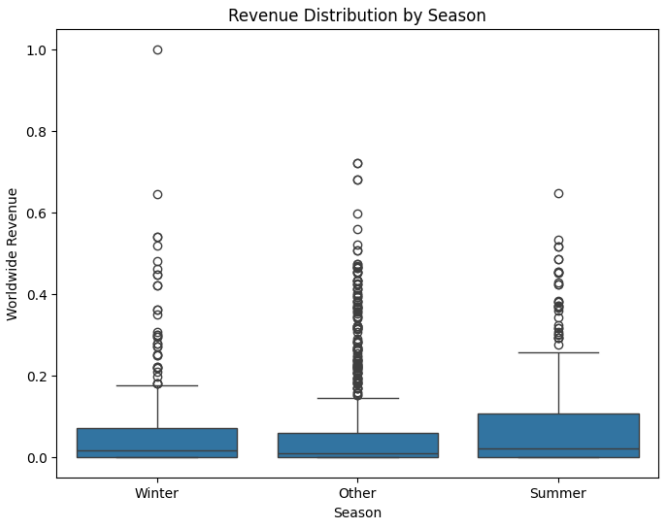

Fig. 7: Revenue Distribution by Season

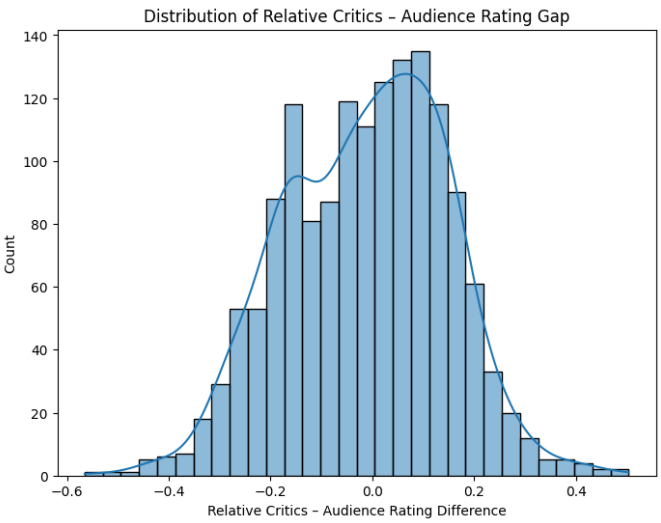

Fig. 9: Distribution of Relative Critics – Audience Rating Gap

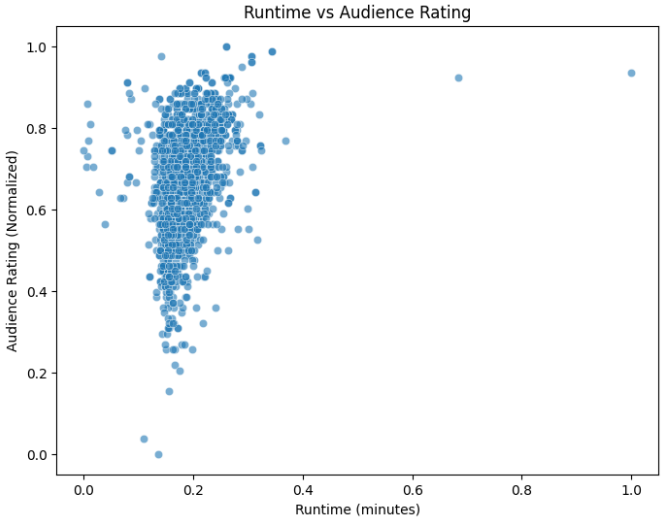

Fig. 10: Runtime vs Audience Rating

VI. Linear Regression Analysis

Linear regression was used to quantify the influence of budget, cast popularity, and IMDb ratings on revenue.

A. Model Description

Revenue was treated as the dependent variable, with standardized budget, popularity, and ratings as predictors.

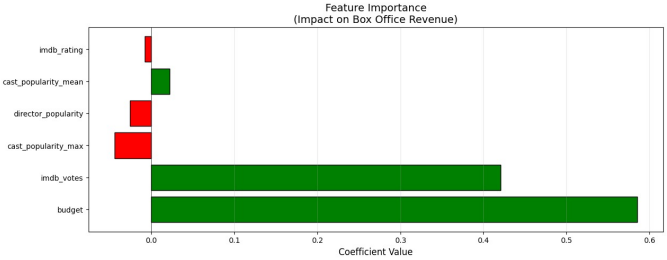

Fig. 11: Regression coefficients comparison (Placeholder Figure 6)

Insight 4: Budget and popularity contribute more strongly to revenue prediction than IMDb ratings. **Insight 5:** The model explains only part of the variance, indicating external unobserved factors.

VII. Clustering Analysis

K-menas clustering was applied to identify groups of movies with similar characteristics.

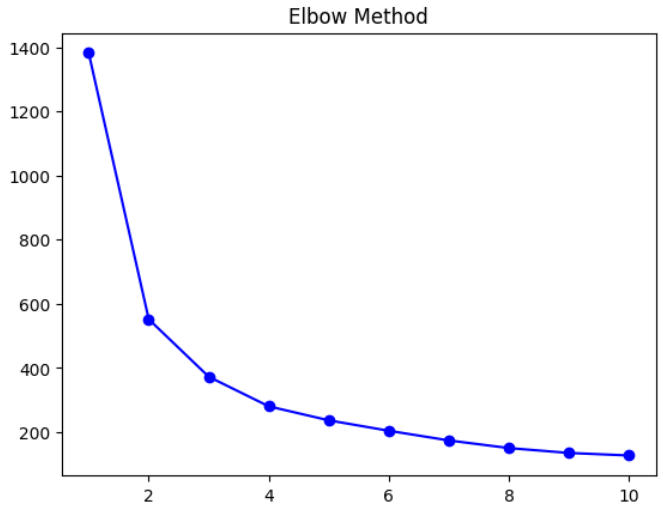

Fig. 12: Using elbow method to find the optimal number of clusters.

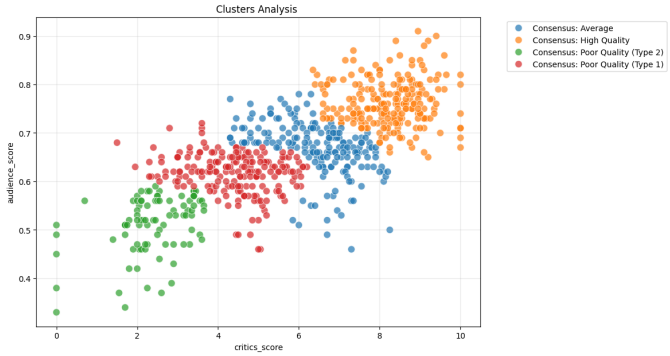

Fig. 13: Four clusters produced through K-means.

Insight 10: Clustering reveals four distinct groups; Poor quality, Audience preferred, Average, and High quality.

VIII. Conclusion

This study examined the factors influencing movie box office performance using a structured data analysis pipeline incorporating exploratory analysis, visualization, linear regression, and clustering. The results provide clear insights into the financial, seasonal, and perceptual dimensions of movie success.

First, the analysis indicates that production budget is the strongest predictor of worldwide revenue, followed by cast popularity. IMDb ratings exhibit only a weak relationship with financial performance. Linear regression results confirm that investment scale and visibility-related factors explain a substantially larger portion of revenue variability than audience ratings. This suggests that while ratings reflect audience perception, they are not reliable indicators of commercial success on their own.

Second, the seasonal analysis reveals that movies released during the summer season achieve higher average worldwide

revenue than winter releases. This trend supports the hypothesis that release timing plays a role in financial performance, likely due to increased audience availability and strategic scheduling of high-budget films. However, the observed seasonal effect is moderate, indicating that release timing must be considered in conjunction with other factors such as budget and popularity.

Finally, the comparison of critics' and audience ratings across genres demonstrates systematic genre-dependent differences in perception. Action, War, and Crime genres show larger discrepancies where critics rate films more favorably than audiences, whereas genres such as Documentary and TV Movie display higher audience ratings. These findings highlight that critical reception and audience satisfaction do not align uniformly across genres.