Analysis of Factors Influencing Movie Box Office Performance

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

Data Analysis Project

Introduction

Research Questions:

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?

Data Sources & Preprocessing

Data Collection

- Sources: TMDb + OMDb APIs
- Automated data scraping
- Includes: financial metrics, release dates, ratings, popularity, genres

Preprocessing Steps

- Removal of records with missing critical variables (e.g., revenue)
- Data type normalization
- Feature normalization for analysis
- No imputation for budget/revenue to avoid bias

EDA Results: Feature Correlations

Table: Correlation with Worldwide Revenue

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

Key Insight

Budget has the strongest positive correlation with revenue, followed by cast popularity.
IMDb rating shows only a weak positive correlation.

EDA Results: Seasonal Analysis

Table: Average Worldwide Revenue by Release Season

Season	Average Worldwide Revenue
Summer	**0.0987**
Winter	0.0800

Key Insight

Movies released in summer generate approximately 23% higher average revenue
compared to winter releases.

EDA Results: Rating Comparison

Audience vs Critics Ratings

- **Audience scores are consistently higher** than critics' scores
- Average gap varies by genre
- Action, War, Crime: Critics rate higher than audiences
- Documentary, TV Movie: Audiences rate higher than critics

EDA Results: Visual Patterns

Insight

Higher budget movies tend to generate higher revenue, though with considerable variation.

EDA Results: Revenue Distribution

Revenue by IMDb Rating

- Movies with higher IMDb ratings generally have higher revenue
- Relationship is modest compared to budget
- Many high-revenue movies have mid-range ratings (6-8)

Main Analysis: Linear Regression

Model

- Dependent variable: Worldwide Revenue
- Predictors: Budget, Cast Popularity, IMDb Rating (standardized)

Key Results

- **Budget and popularity contribute more strongly** than IMDb ratings
- Model explains only part of variance → other factors matter
- Investment scale and visibility outweigh audience ratings in predicting revenue

Clustering Analysis: Determining Optimal Clusters

Methodology

- Applied K-means clustering
- Used Elbow Method to determine optimal cluster count
- Evaluated within-cluster sum of squares (WCSS)
- Selected K=4 as optimal

Elbow Method Interpretation

- Sharp bend at K=4 indicates optimal number
- Beyond 4 clusters, marginal improvement decreases
- Balance between complexity and explanatory power

Clustering Results: Cluster Characteristics

Table: Characteristics of Four Movie Clusters

Cluster Name	Key Characteristics
Poor Quality	• Low ratings (both audience and critics) • Low revenue performance • Typically lower budget productions
Audience Preferred	• High audience scores • Moderate critics scores • Genre-specific appeal (e.g., comedies, action)
Average	• Mid-range across all metrics • Moderate ratings, revenue, and budget

Clustering Results: Cluster Statistics

Table: Statistical Summary of Movie Clusters

Metric	Poor Quality	Audience Preferred	Average
Avg. Revenue	Low	Medium	Medium
Avg. Budget	Low	Medium	Medium
Avg. IMDb Rating	¡5.5	6.0-7.0	6.5-7.5
Avg. Cast Popularity	Low	Medium	Medium
Audience-Critic Gap	Small	Large	Medium
% of Movies	15%	25%	45%

Insight

The majority of movies fall into the "Average" cluster, with smaller proportions in extreme categories.
High quality cluster represents only 15% of movies but achieves the best financial performance.

Clustering Results: Visualization

Axes Interpretation

- X-axis: Budget/Revenue dimension
- Y-axis: Rating/Popularity dimension
- Clusters show natural separation

Cluster Separation

- Clear distinction between clusters
- Some overlap between Average and Audience Preferred
- High Quality cluster clearly separated

Clustering Analysis: Key Insights

Insight 1: Four Natural Movie Categories

- Movies naturally group into four distinct categories based on quality and popularity
- This categorization aligns with industry intuition about film types

Insight 2: Financial vs. Critical Success

- "High Quality" cluster shows alignment between critical acclaim and financial success
- "Audience Preferred" cluster shows disconnect between critics and audiences
- "Poor Quality" cluster consistently underperforms across all metrics

Insight 3: Production Strategy Implications

- Different clusters suggest different production and marketing strategies

Conclusions

Key Findings

1. **Budget** is the strongest predictor of revenue, followed by cast popularity
2. IMDb ratings show only weak relationship with financial performance
3. **Summer releases** achieve higher revenue than winter releases
4. **Critics vs. Audience:**
    - Action/War/Crime: critics rate higher
    - Documentary/TV Movie: audiences rate higher
5. **Clustering reveals** four distinct movie types with different success patterns

Implications

- Commercial success depends more on investment/marketing than ratings