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
- Data type normalization
- Feature normalization
- No imputation for budget/revenue

EDA Results: Feature Correlations

Table: Correlation with Worldwide Revenue

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

Key Insight

Budget has the strongest positive correlation with revenue.
IMDb rating shows only a weak positive correlation.

EDA Results: Seasonal Analysis

Table: Average Worldwide Revenue by Release Season

Season	**Avg. 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

Key Findings:

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

Main Analysis: Linear Regression

Model Specification

- **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

Clustering Analysis: Optimal Clusters

Methodology

- Applied K-means clustering
- Used Elbow Method to determine optimal clusters
- 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	Characteristics
Poor Quality	Low ratings, low revenue, lower budget
Audience Preferred	High audience scores, moderate critics scores
Average	Mid-range across all metrics, most common
High Quality	High ratings, high revenue, higher budget

Clustering Results: Statistics

Table: Statistical Summary of Clusters

Metric	**Poor**	**Audience**	**Average**	**High**
Avg. Revenue	Low	Medium	Medium	High
Avg. Budget	Low	Medium	Medium	High
IMDb Rating	¡5.5	6.0-7.0	6.5-7.5	¿7.5
% of Movies	15%	25%	45%	15%

Insight

Most movies are "Average" (45%). High quality cluster is only 15% but achieves best financial performance.

Clustering Analysis: Key Insights

Insight 1

Movies naturally group into four distinct categories based on quality and popularity

Insight 3

"Audience Preferred" shows disconnect between critics and audiences

Insight 2

"High Quality" shows alignment between critical acclaim and financial success

Insight 4

Different clusters suggest different production and marketing strategies

Conclusions

Key Findings

1. **Budget** strongest predictor of revenue
2. Summer releases earn 23% more
3. Critics vs. Audience gaps vary by genre
4. Four distinct movie types identified

Implications

- Success depends more on investment than ratings
- Release timing matters
- Genre affects perception
- Cluster analysis provides strategic framework