Movie Success Analysis Project – CRISP DM Data Report

Project title: Movie Success Analysis - Insights for a New Studio
Team: Group 5 - Stella Kiarie, Kumati Dapash, Doris Mutie, Morvine Otieno
Date: 6th November 2025

Business Understanding

The movie industry is highly competitive, especially as streaming platforms and major studios continue producing large volumes of films. For a new movie studio entering the market, understanding what types of films perform well is essential.

The goal of the business is to launch a new movie studio to produce movies that the box office. The objective is to conduct comprehensive research and analysis based on historical data from the Box Office Mojo and IMDb database.

The analysis should help the business understand:

The trends and patterns existing within the movie production industry

To identify the key factors that contribute to the success of a movie box office.

The existing relationship between different movie characteristics and box office movie success.

To identify the studio with the highest revenue for better investment decisions.

Introduction

Overview/ Background

The entertainment industry is currently undergoing a rapid transformation, driven by key players producing original content for competitive streaming platforms. In this volatile market, a company planning to launch a new movie studio requires competitive, data-driven intelligence to ensure profitable investment decisions. This project addresses that need by conducting a comprehensive exploratory data analysis of historical movie performance and characteristics. The analysis specifically leverages movie revenue and rating data sourced from Box Office Mojo and IMDB to scrutinize film success metrics. The primary goal is to identify patterns, relationships, and the core factors that contribute most significantly to a film's box office success. Ultimately, these findings will translate into actionable recommendations to guide the new venture on which studios to invest in and the most profitable types of films to produce.

Challenges

The company currently lacks insight into which movie attributes lead to strong box office performance. Success depends on factors such as genre, audience appeal, runtime, and studio reputation. However, these patterns are not immediately clear. Additionally, revenue data is unevenly distributed, with a few blockbuster films dominating the market, making it difficult to identify reliable trends without structured analysis.

Proposed Solution

This project proposes conducting a comprehensive exploratory data analysis using data from IMDb and Box Office Mojo. The goal is to explore genres, movie characteristics, and audience rating patterns that contribute to financial success. The insights gained will guide the studio in shaping its production strategy, selecting film concepts, and allocating resources effectively.

Brief Conclusion

By analyzing real-world movie performance data, the studio will gain actionable insights that support evidence-based decision-making. The expected outcome is a clear set of recommendations, such as high-performing genres, optimal runtime ranges, and the influence of audience ratings, that can help the new studio reduce risk and maximize box office returns.

Problem Statement

The company lacks data-driven insights into what types of movies succeed at the box office. Without understanding which genres, characteristics, or audience factors drive revenue, the new movie studio risks investing in projects that may not perform well financially.

Objectives

1. To identify high-performing movie genres.
2. To evaluate the impact of audience ratings on box office performance.
3. To identify key factors that predict movie success (e.g., genre, runtime, release timing).
4. To examine how box office performance changes over time.
5. Make conclusions and recommendations

Success criteria:

Success means delivering at least three clear, data-driven recommendations supported by visuals and summary statistics – specifically on which genres perform best, how runtime affects revenue, and how audience ratings relate to total box office performance.

Data Understanding

Datasets Used

- *im.db* (IMDb SQLite) – movie metadata such as titles, year, runtime, genres, ratings. Primary tables used: movie_basics, movie_ratings.
- *bom.movie_gross.csv.gz* (Box Office Mojo) – domestic_gross, foreign_gross, studio, and year.

Key Variables

- Target: total_gross (created as domestic_gross + foreign_gross)
- Predictors: genre, runtime, release year, IMDb rating, number of votes, studio

Initial Observations

- Revenue distribution is highly skewed – a small number of blockbusters account for a large share of total revenue.
- Missing or inconsistent entries across datasets, especially for foreign gross, runtime, and ratings.
- Merging required careful standardization of titles and years due to inconsistent formatting.

Data Preparation

Cleaning steps

- Unzipped the dataset and connected to *im.db.*
- Loaded the *movie_basics* and *movie_ratings* tables from the database.
- Imported the Box Office Mojo dataset (*bom.movie_gross.csv.gz*) using *pd.read_csv,* which handled compression automatically.

- Standardized movie titles and years (e.g., converted to lowercase, removed extra spaces) to enable proper merging across datasets.
- Merged IMDb and Box Office Mojo data based on title and year.
- Filtered out movies with missing or zero box office gross values.
- Cleaned data types: converted year columns to integers and ensured numeric columns (like gross and ratings) were correctly formatted.
- Handled missing values where necessary – Filled missing runtime_minutes with the median and filled missing genres with “Unknown”
- required complete data.

Feature selection

- Focused on a small but meaningful set of predictors for EDA – including genre, runtime range, and audience rating (*IMDb rating bucket*).

Explore Key Relationships (Based on Objectives)

The analysis used summary statistics to explore key factors influencing movie success.

- **Top Genres:** Most movies earned higher foreign gross than domestic, showing strong international appeal. Adventure, Fantasy, and Sci-Fi genres performed best in both markets, while Documentary and Drama showed balanced results. Overall, foreign markets contributed more revenue, with Adventure, Drama, and Sport leading in both domestic and international earnings.
- **Audience Ratings:** The correlation matrix shows a strong positive relationship (0.81) between domestic gross and foreign gross, meaning movies that earn more locally also tend to perform well internationally. The correlations between average rating and both gross values are weak (around 0.12), suggesting that higher ratings have little impact on box office revenue.
- **Runtime:** There is a weak positive correlation (0.13) between runtime_minutes and

domestic_gross, suggesting that longer movies tend to earn slightly higher domestic revenue, but the relationship is very weak and not significant.

- **Studios:** The data shows that BV has the highest average domestic gross, followed closely by P/DW, indicating these studios lead in local revenue performance. WB, Universal, and WB (NL) follow with moderate earnings, while MGM, Paramount, Sony, and Fox show slightly lower averages. Summit (Sum.) records the lowest domestic gross among the listed studios.

Overall: High-performing genres, reputable studios, and balanced runtimes contribute most to box office success. Audience ratings support success but are not the main driver.

Hypothesis Testing

In this section, we perform a Chi-Square Test of Independence to determine whether there is a significant relationship between a movie's genre and its box office success level (measured using domestic gross revenue).

This helps us understand whether certain genres are statistically more likely to perform well than others – a key insight for our movie studio's strategy.

Chi-Square Test of Independence

We perform a Chi-Square test to determine whether there is a significant relationship between movie genres and box office success levels. This statistical test supports our earlier exploratory analysis by verifying whether observed differences are due to chance or real relationships in the data.

Hypotheses

Null Hypothesis (H_0): There is no relationship between movie genre and box office success.

Alternative Hypothesis (H_1): There is a relationship between movie genre and box office success.

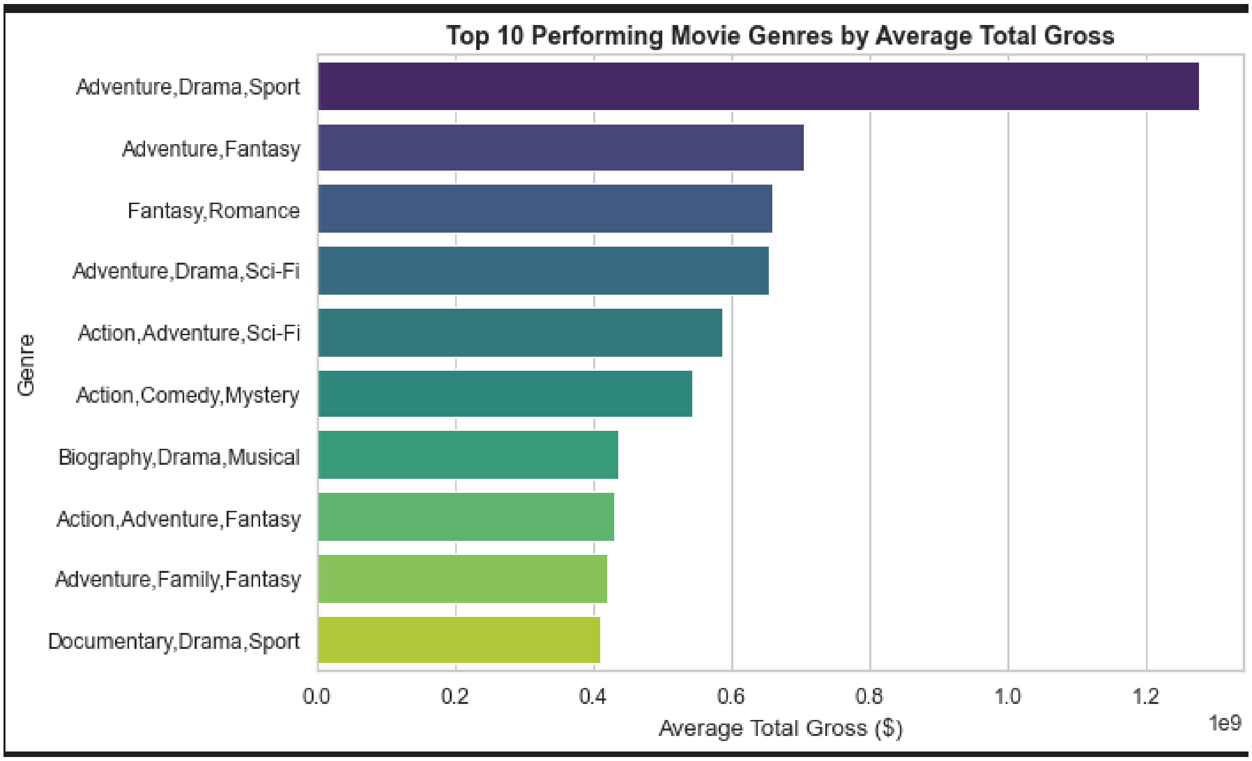
Interpretation

The Chi-Square test produced a statistic of 736.65 with a p-value of 1.06e-34, which is far below the 0.05 significance level. This means we reject the null hypothesis and conclude that there is a significant relationship between movie genre and box office success. In other words, the type of genre strongly influences how well a movie performs financially.

Key visualizations & what they show

Figure 1: Top 10 Performing Movie Genres by Average Total Gross

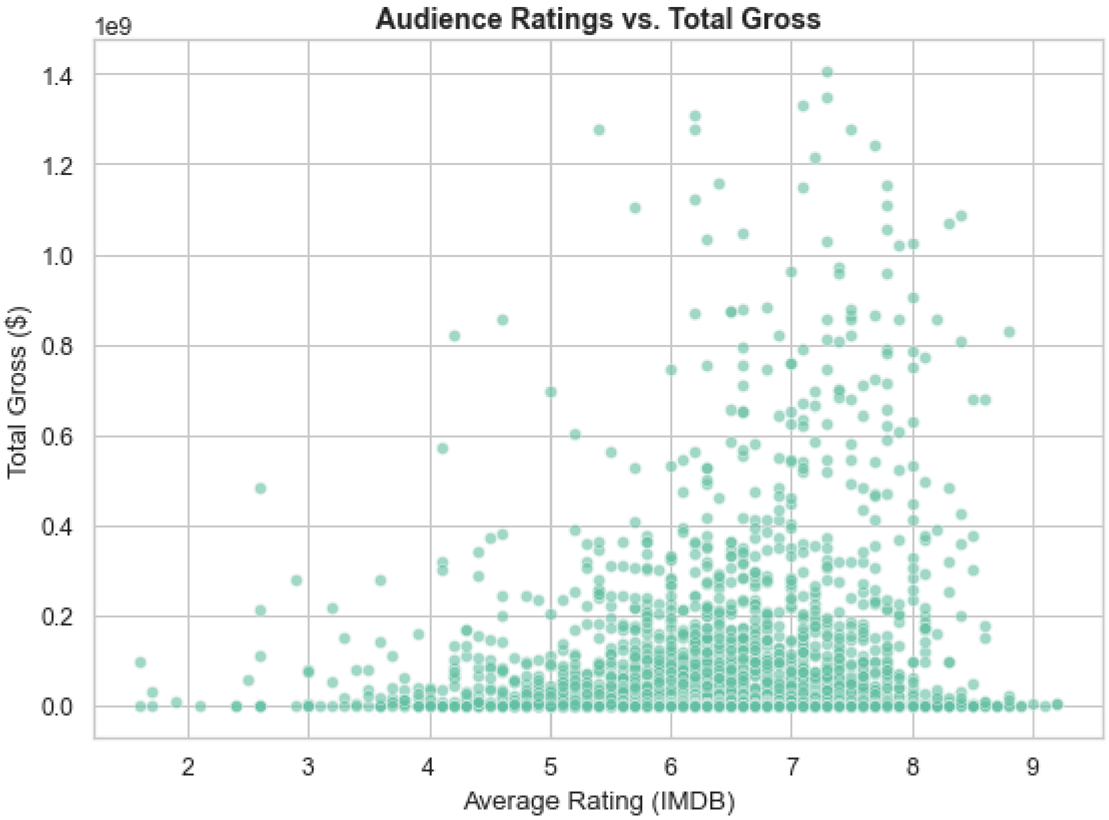
📊 **Bar chart of the highest-grossing genres (average total gross).**



Insight: Adventure and Action-based genres dominate box office performance, especially when combined with elements of Drama, Fantasy, or Sci-Fi. These hybrid genres consistently attract large audiences and deliver the highest average total gross, exceeding $1 billion in some cases. This suggests that audiences prefer visually engaging, emotionally compelling, and high-concept stories making Adventure Drama and Action–Fantasy films the most profitable focus areas for a new movie studio.

Figure 2. Objective 2: Ratings vs. Box Office

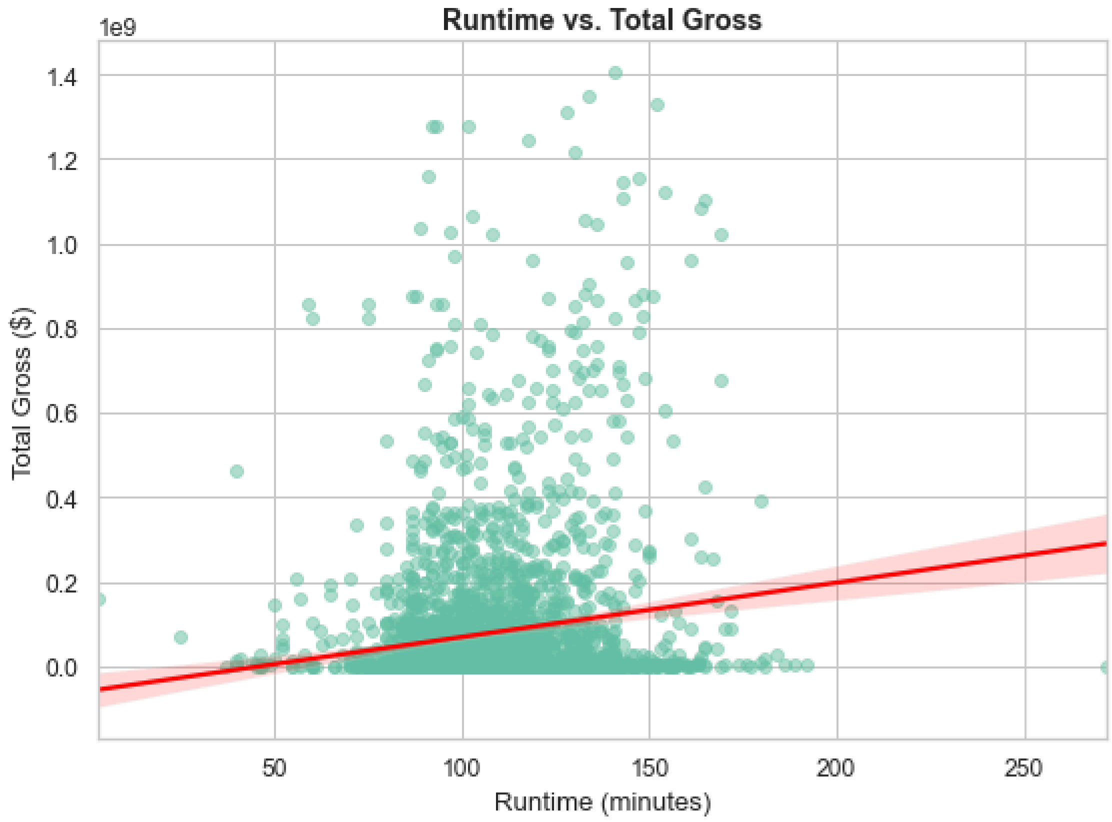
📈 Scatterplot showing the relationship between IMDb audience ratings and total gross revenue.



Insight: The scatter plot shows a positive relationship between audience ratings and total box office gross. Movies with ratings above 6.0 on IMDB tend to earn significantly higher revenues, while poorly rated films (below 5.0) rarely achieve major financial success. This indicates that audience satisfaction and perceived quality strongly influence a movie's commercial performance. Films that resonate well with viewers are more likely to generate higher box office returns.

Figure 3: Runtime vs Total Gross

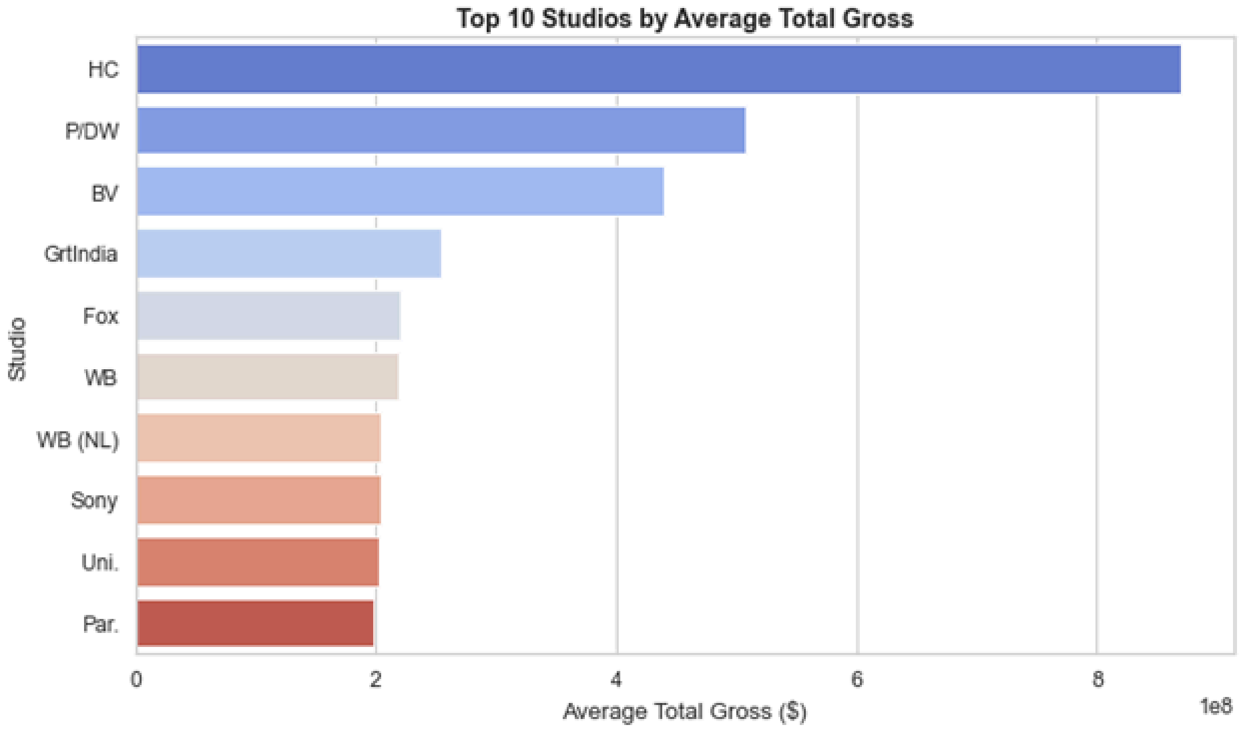
📈 **Regression plot of runtime (minutes) versus total gross.**



Insight: The visualization reveals a slight positive relationship between runtime and total box office gross. Movies with runtimes between 90 and 130 minutes generally achieve higher earnings, suggesting that audiences prefer films that are long enough to develop a story but not excessively lengthy. Extremely short or overly long films tend to earn less, indicating that finding an optimal runtime balance is key to maximizing box office performance.

Figure 4: Top 10 Studios by Average Total Gross

🏢 **Bar chart of studios ranked by average total gross across their films.**



Insight: The analysis shows that a few major studios dominate the box office market. Studios such as HC, P/DW, and BV achieve the highest average total gross, far outperforming others. This suggests that established studios benefit from larger production budgets, strong marketing strategies, and brand loyalty. Meanwhile, smaller studios like Sony, Universal, and Paramount still maintain consistent performance but on a smaller scale. Overall, studio reputation and financial capacity play a significant role in driving box office success.

Evaluation – Findings & Insight

1. Genre performance

Action, Adventure, Fantasy, and Sci-Fi films show the highest average total gross, while Drama and Documentary films earn much less.

2. Audience rating & runtime

Ratings and runtime have weak positive relationships with revenue. Movies rated higher and those running **90–130 minutes** tend to earn slightly more.

3. Studio performance

BV, P/DW, and WB record the highest average grosses, indicating stronger performance from established studios.

4. Data distribution

Revenue is highly skewed, with a few blockbusters dominating totals. Median values provide a more realistic picture of typical earnings.

Business recommendations

Based on the analysis of movie performance data from Box Office Mojo and IMDB, the following strategic recommendations are proposed for the company's new movie studio:

1. **Focus on High-Performing Genres**

   Prioritize producing Adventure, Action, and Fantasy films, especially hybrid genres that combine emotional storytelling elements such as Drama or Romance. These genres consistently deliver the highest box office returns.

2. **Prioritize Quality and Audience Satisfaction**

   Invest in strong storytelling, character development, and production quality. The analysis shows that movies with IMDB ratings above 6.0 tend to earn significantly more revenue, proving that audience approval directly drives profitability.

3. **Optimize Movie Runtime**

   Aim for runtimes between 90 and 130 minutes. This range achieves the best balance between story depth and audience attention, leading to better commercial success.

4. **Collaborate with Established Studios and Talent**

   Partnering with or hiring talent from top-performing studios (like HC, P/DW, and BV) can provide valuable expertise in production, marketing, and distribution, increasing the studio's chances of success in the competitive film industry.

Deployment / Practical Use

How the studio could use these insights.

- **Greenlighting decisions:**
  Use a simple checklist when approving movie projects: strong genre fit (Action, Adventure, Fantasy, Sci-Fi), compelling story concept, expected runtime between **90–130 minutes**, and projected audience appeal based on rating patterns.
- **Marketing allocation:**
  Focus marketing resources on films in high-performing genres and titles showing strong early audience interest or strong IMDb ratings.
- **Content pipeline planning:**
  Include at least one major film each year supported by several mid-budget genre films to maintain steady revenue and diversify risk.

Limitations

- Some movies lacked complete financial or rating information.
- Matching titles across IMDb and BOM may introduce small merge errors.

- Important factors such as cast popularity, marketing spend, or franchise strength were not included.
- A few blockbuster hits can skew average revenue results.

Executive Summary

- Action, Adventure, Top-performing genres are Action, Adventure, Fantasy, and Sci-Fi, making them the strongest candidates for future investment.
- Movies with higher IMDb ratings and runtimes between 90–130 minutes earn slightly more, though the effects are weak.
- Established studios like BV, P/DW, and WB consistently produce higher-grossing films; a balanced slate of one major film plus several mid-budget releases helps manage risk and maximize returns.

Next Steps

1. Deepen the Analysis:

Explore how release year and timing (seasonal trends) affect movie performance.Analyze budget and marketing expenditure data (if available) to strengthen profitability insights.

2. Market Testing:

Conduct audience surveys or focus groups to validate genre preferences and refine story concepts before production.

3. Predictive Modeling (Future Work):

Build a machine learning model to predict expected box office performance based on movie characteristics (genre, runtime, ratings, etc.).

4. Strategic Implementation:

Use these insights to guide content investment decisions, set realistic revenue targets, and design an effective marketing strategy for upcoming productions.