Data-Driven Insights for a Modern Film Studio

Overview

The global movie industry has undergone major shifts in recent years, with changes in consumer behavior, streaming trends, and content production strategies. Streaming platforms such as Netflix and Prime are reshaping how content is consumed and evaluated. While theatrical revenue remains an important success metric, online ratings, platform distribution and viewer engagement have become just as critical in predicting and evaluating a movie's performance.

As a new entrant in the movie production industry, understanding what drives a movie’s commercial success is crucial for making data-driven decisions in production, marketing, and distribution.

This project seeks to uncover actionable insights from historical movie data to uncover what drives both financial and audience success, and how this knowledge can inform business strategy for a potential new film studio.

Business Understanding

Background

With increased competition from both traditional cinema and digital streaming platforms, studios need to optimize decisions around:

- Genre selection
- Budget allocation
- Casting
- Timing of releases
- Marketing focus
- Platform strategy - Balancing between the box office and streaming platforms

By analyzing both theatrical and streaming success, studios can better navigate this complex, hybrid distribution landscape. Data from past movies including box office revenue, ratings, genres, and production details can reveal patterns and predictors of success.

This analysis will serve as a proof of concept for how a data-driven approach can enhance Return on Investment(ROI) and reduce risk in movie production.

Problem Statement

A new movie production company is seeking to make informed decisions about:

- What types of movies to produce (genre, language, duration)
- How much budget to allocate
- When to release their movies
- Which actors or directors are most associated with successful projects
- Decide whether to prioritize theatrical releases, streaming or a hybrid model

The challenge is to analyze historical movie data to find patterns that can help predict which factors lead to higher box office performance or audience engagement.

Objectives

The primary objectives of this project are:

1. To identify the key factors that contribute to a movie's success (e.g. revenue, high ratings, box office success, streaming performance)
2. To provide data-backed recommendations for genre selection, ideal budgets, release strategy and cast decisions
3. To build visualizations and models that support strategic decisions for a new movie production company

Metrics of Success

- Identification of top 5 features most correlated with success metrics
- Creation of visual dashboards to communicate findings clearly
- Development of a simple predictive model for revenue or rating to estimate movie success
- Strategic, business-friendly recommendations based on findings
- Well-documented collaboration and communication via GitHub, Trello, and reporting tools

Tools and Technologies

- Python (Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn)
- Jupyter Notebooks for analysis
- Git & GitHub for version control and collaboration

- Trello for project management and workflow tracking
- Google Docs for the final data report
- Google Slides for the Presentation
- Tableau for Interactive Dashboard

Stakeholders

The primary stakeholder is the founding team of the new movie production studio.

Secondary stakeholders include potential investors, marketing consultants, streaming partners and creative directors.

Data Understanding

In order to uncover what drives a movie's success, both financially and in terms of audience reception, we must first develop a thorough understanding of the data at our disposal. This section explores the structure, scope, and quality of the datasets used for analysis.

Our data sources include publicly available movie datasets with information on:

- **Movie metadata**: title, genre, release date, language, runtime, production companies
- **Financial data**: production budget, box office revenue (domestic & worldwide)
- **Ratings**: IMDb scores, Rotten Tomatoes critic/audience ratings
- **Streaming availability**: whether the movie was released theatrically, via DVD or streaming, or both
- **Cast and crew**: actors, directors, and producers

Data Source

The movie datasets are drawn from:

- Box Office Mojo
- IMDb
- Rotten Tomatoes
- The Movie DB

- The Numbers

The datasets are in the following formats:

- bom.movie_gross.csv (CSV File)
- im.db (sqLite Database)
- rt.movie_info.tsv (TSV File)
- rt.reviews.tsv (TSV File)
- tmdb.movies.csv (CSV File)
- tn.movie_budgets.csv (CSV File)

By reviewing the attributes of these datasets and exploring initial patterns, we aim to:

- *Identify which variables are relevant to our business goals and their data types*
- *Detect any missing, duplicated, or inconsistent records*
- *Gain early insights into data trends that may inform modeling later on*

This understanding will serve as the foundation for cleaning, feature engineering, and deeper analysis in the next stages of the project.

Data Preparation

Following our Data Understanding phase, we now transition into the Data Preparation stage of the CRISP-DM methodology. This phase is crucial in transforming raw data into a clean and structured format that can be used effectively in analysis and modeling.

The tasks in this section include:

- Selecting relevant tables and columns.
- Handling missing or inconsistent data.
- Converting data types (e.g., dates, floats, integers).
- Creating derived attributes where necessary (e.g., year of release from a full date).
- Merging datasets appropriately based on relationships (e.g., movie_id, person_id).
- Filtering and sampling records for exploratory and predictive tasks.

Goals

- Ensure the dataset is clean, consistent, and analysis-ready.
- Retain only data relevant to our business objectives.
- Prepare a consolidated and structured dataset for feature engineering and modeling in subsequent phases.

Rotten Tomatoes (rt.movie_info.tsv / rt.reviews.tsv)

Since both datasets have a common 'id' column, we used that to merge the two using the left join so that we don't lose a lot of data. The dataset contained columns under the name;

Id: Shows the unique identification number for each movie

Rating: This gives the movies suitability for certain audiences based on its content

Genre: Stylistic categories that organize films based on criteria such as the setting, characters, plot, mood, tone, and theme

Director: The creative leader of a film production, responsible for bringing the script to life on screen

Writer: A professional who creates the scripts for films

Theater_date: The date when the movie was released on theaters

Runtime: How long the movie goes for

Rating_10_point: This is a review rating by critics on a scale of 10

Fresh: A boolean column that shows if the movie was fresh[1] (Higher or equal to 7 rating) or rotten[0] (less than 7 rating)

Movie Budgets (tn.movie_budgets.csv)

A lot of excellent feature engineering was done to this file to extact useful information that we are going to use to carry out exploratory data analysis and visualisations. The final cleaned dataset contained the following columns:

Id: The unique identification number for a specific movie

Release_date: Marks the date which the film was released in theaters

Primary_title: Title of the movie

Production_budget: States the cost of producing the entire movie

Domestic_gross: This is the measure of revenue generated from the country of origin of the film

Worldwide_gross: This is the combined revenue the movie generated from all the countries

Foreign_gross: This is the revenue the movie generated outside its country of origin

Total_profit/loss: This is the difference of worldwide gross and production budget

Domestic_profit/loss: This is the difference between the domestic gross and production budget

Foreign_profit/loss: This is the difference between the foreign gross and production budget

ROI(return on investments): performance measure used to evaluate the efficiency or profitability of an investment, in this case total profit.

Release_year: This is the year the movie was released

Release_month: This is the month the movie was released

Domestic_share%: This is the percentage of worldwide revenue that comes from domestic gross

Foreign_share%: This is the percentage of worldwide revenue that comes from foreign gross

Budget_category: This is categorical distribution of a movies production budget ranging from blockbuster, high budget, mid budget and low budget

Is_profiatble: This tells us whether the movie made a profit comparing its productionn budget and worldwide revenue.

IMDB (im.db)

This was a data base that contained various tables of which the following were used;

Movie_basics

Movie_id: The unique identification number for a specific movie

Primary_title: Title of the movie

Start_year: This is the year the movie was released

Runtime: How long the movie goes for

Genre: Stylistic categories that organize films based on criteria such as the setting, characters, plot, mood, tone, and theme

Directors

Movie_id: The unique identification number for a specific movie

Person_id: This is the unique identification for a specific person whether director or writer

Movie_ratings

Movie_id: The unique identification number for a specific movie

Average_rating: This is a review rating by critics

Persons

Person_id: This is the unique identification for a specific person whether director or writer

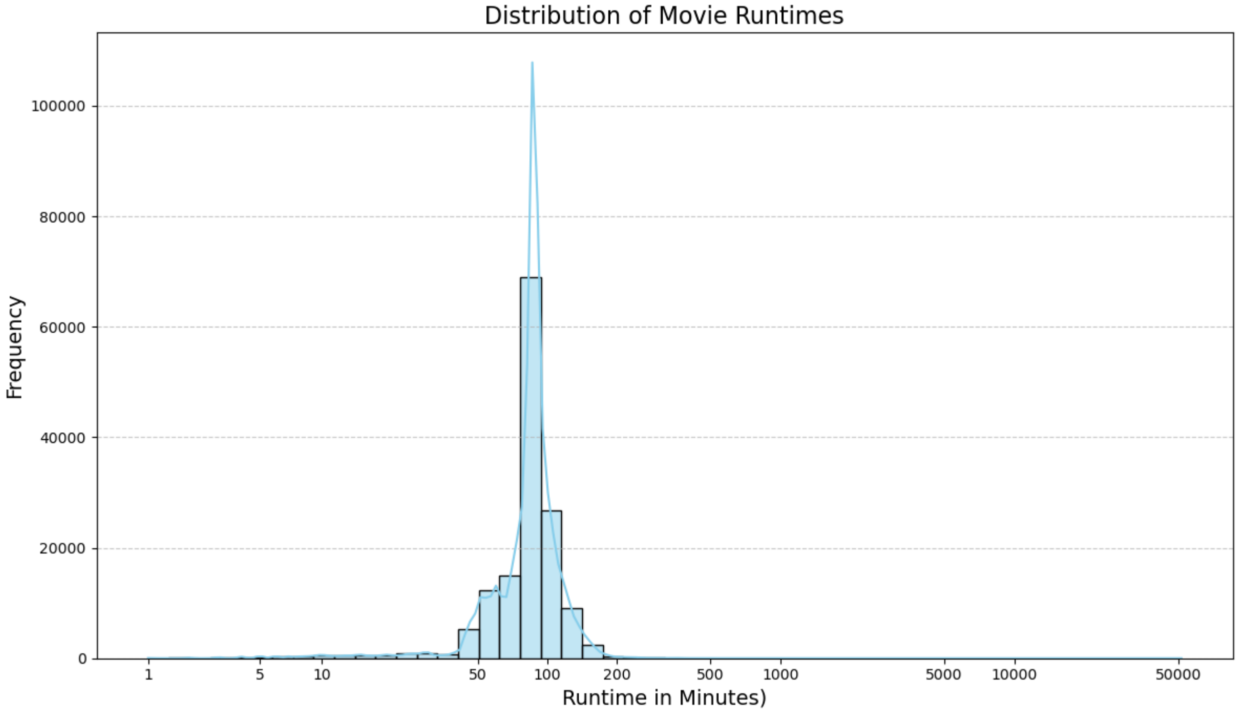
Primary_name: Name tied to the specific person_id

Primary_profession: occupation of the specific person

Exploratory Data Analysis

Univariate Analysis

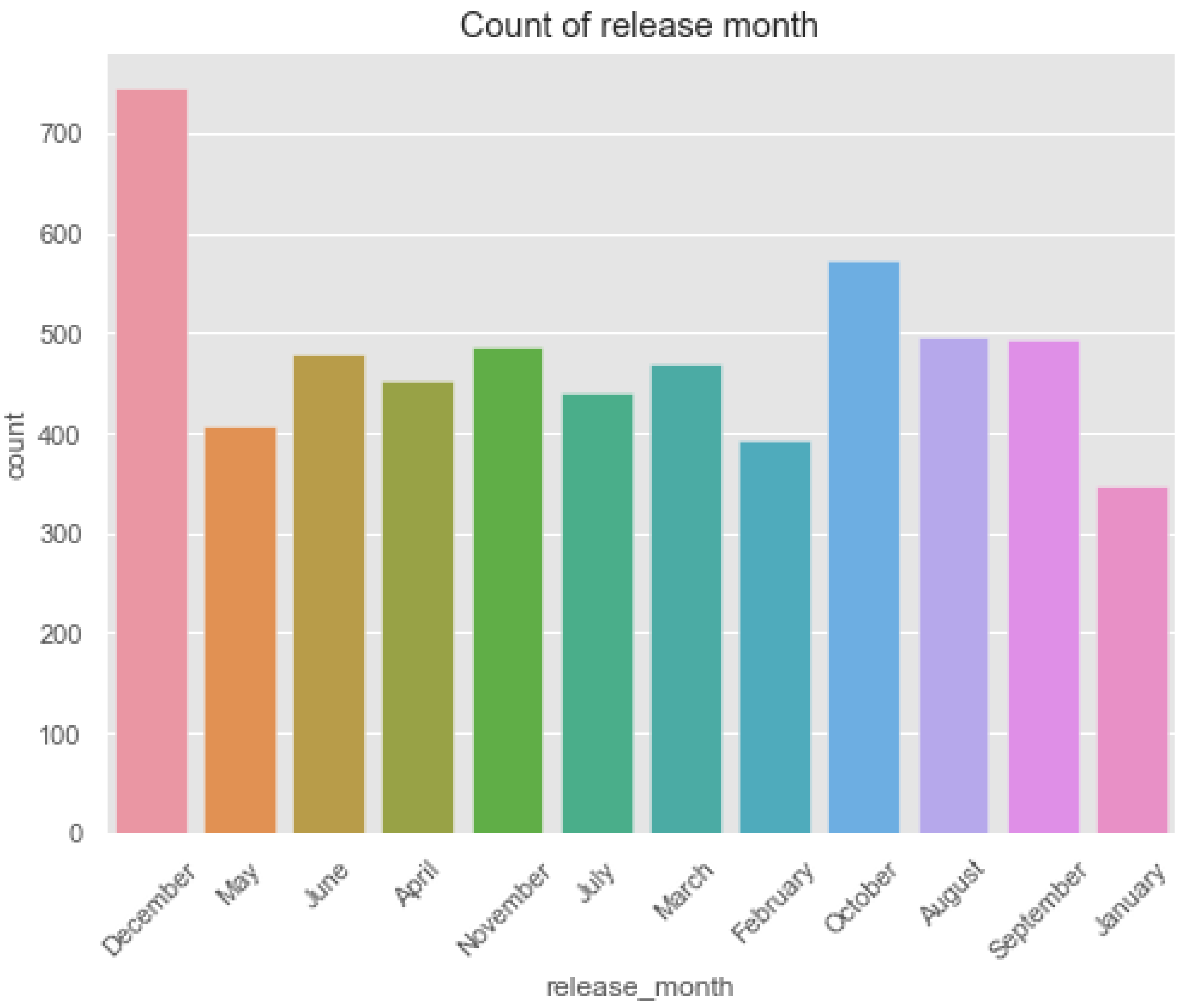
Distribution of Movie Runtimes



Key Insights and Observations

The most important takeaway is that the vast majority of our movies have a "sweet spot" runtime of about 80 to a little under 200 minutes. The tall peak you see on the graph in that range tells us that's where most of our content lies.

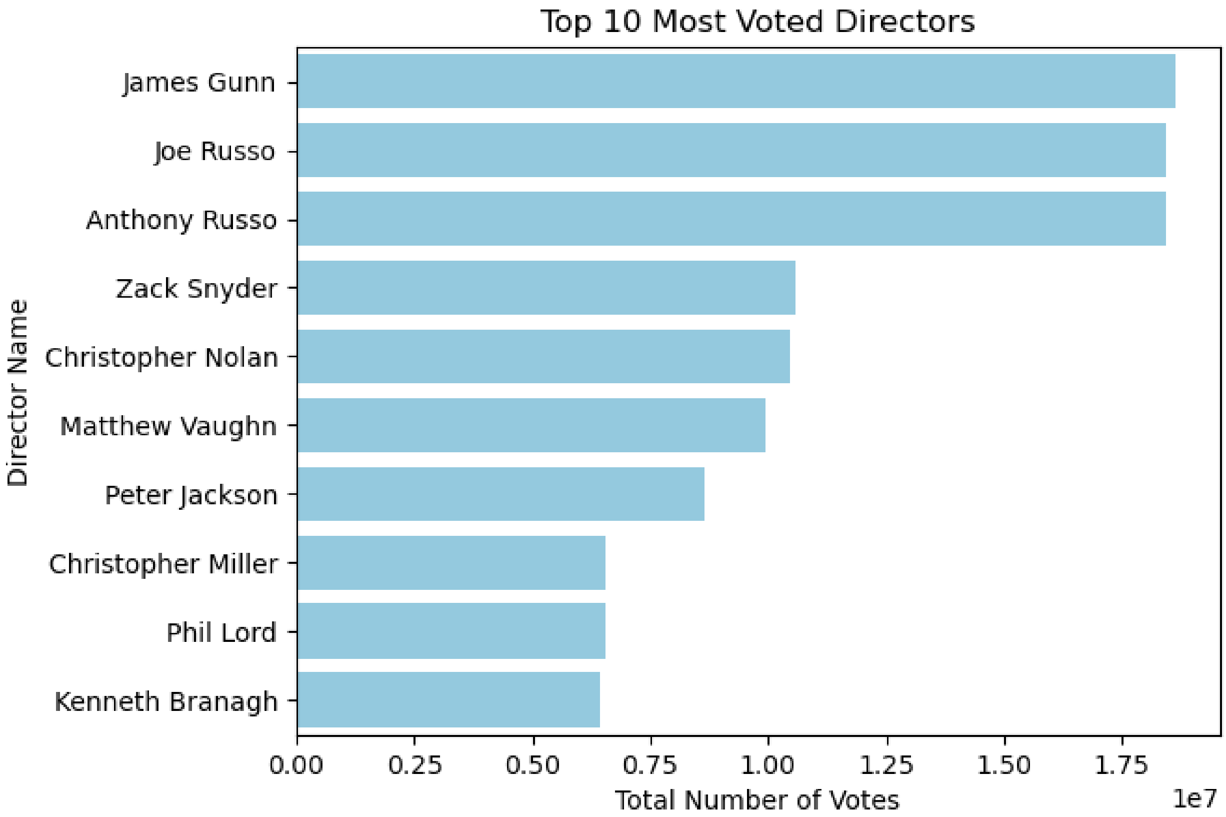
Count of Release Month



Key Insights and Observations

Most movies are released in December showing the Christmas holiday period followed by October then November, August, September, June and March which falls under the summer season and holidays.

Top 10 Most Rated Directors

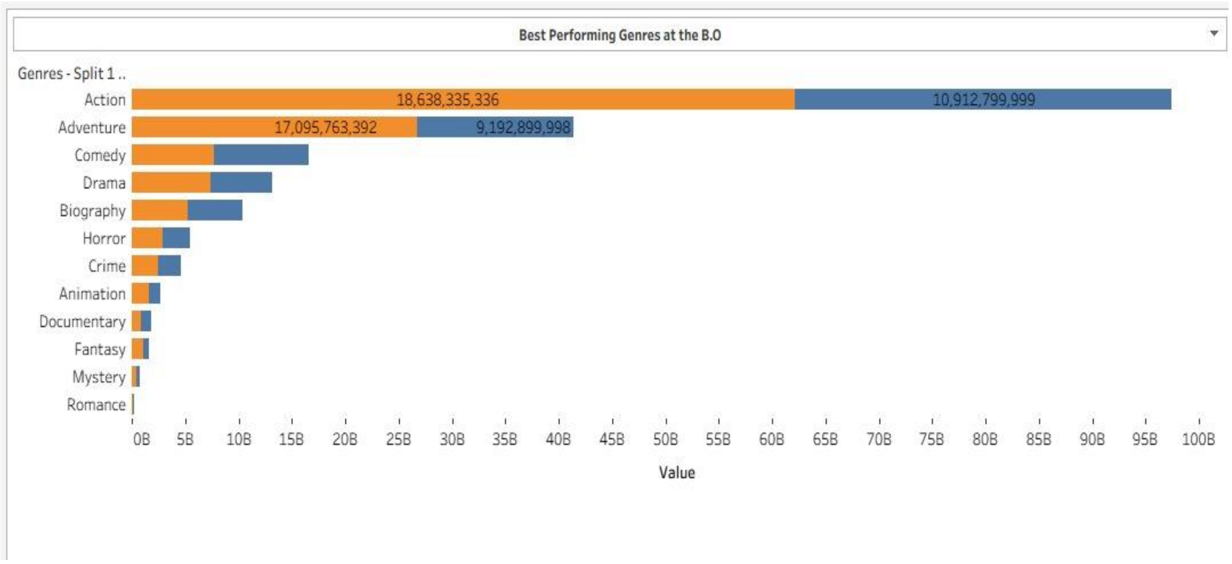


Key Insight and Observation

The list is overwhelmingly composed of directors known for helming major film franchises. The top three—James Gunn and Joe & Anthony Russo—are architects of the Marvel Cinematic Universe (MCU). This indicates that the high vote count is driven by the massive global audiences of these interconnected, big-budget films. This chart is less a measure of critical acclaim and more a reflection of mass audience participation. Directors like **Zack Snyder** (DC Extended Universe), **Peter Jackson** (*The Lord of the Rings*), and **Christopher Nolan** (known for his own brand of sci-fi blockbusters) all create movies that generate huge online discussion and rating activity.

Bivariate Analysis

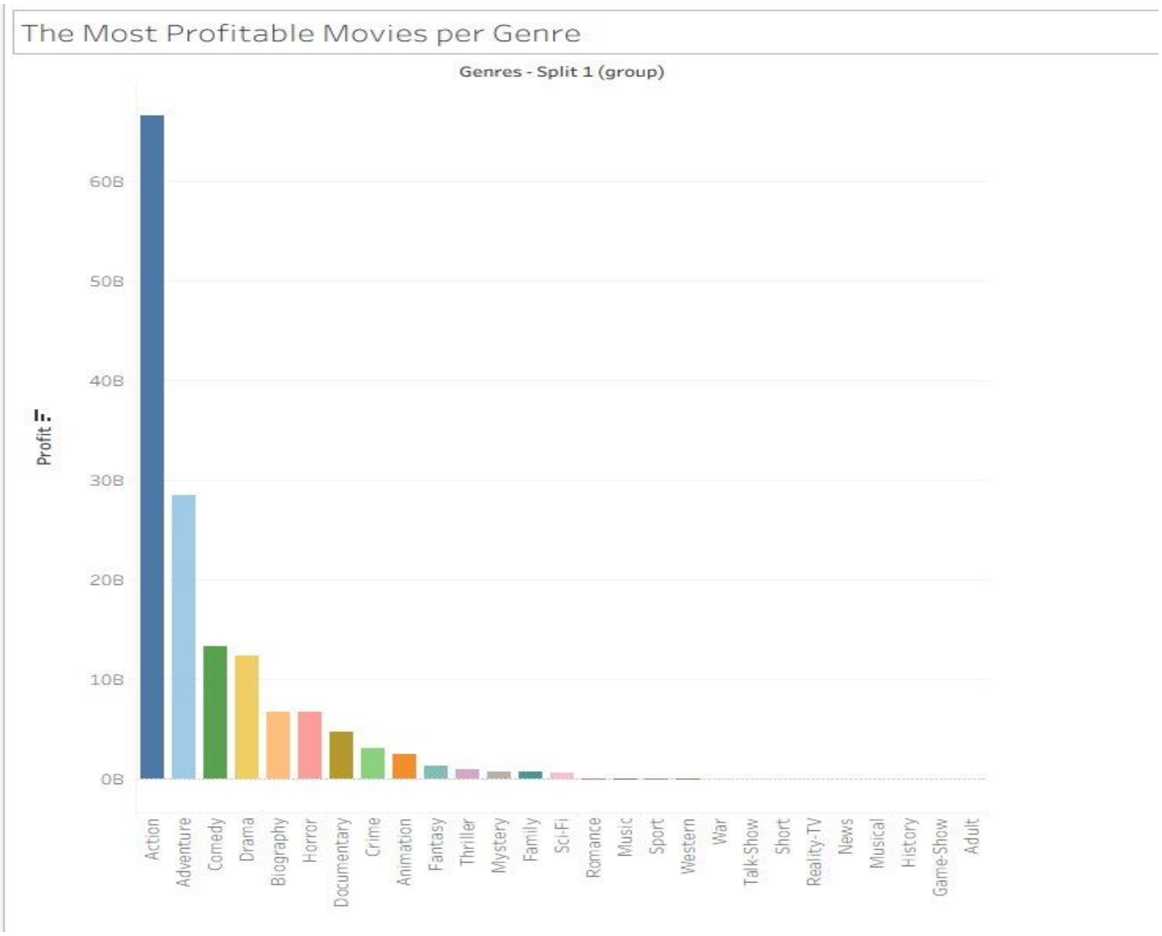
Total Box Office Revenue Per Genre



Key Insights and Observations

Market Dominance by Action & Adventure: The most immediate takeaway is the staggering dominance of the Action and Adventure genres. Together, they account for a massive portion of the total box office revenue shown, dwarfing all other categories. This highlights the "blockbuster" effect, where a few high-stakes genres drive the majority of industry profits.

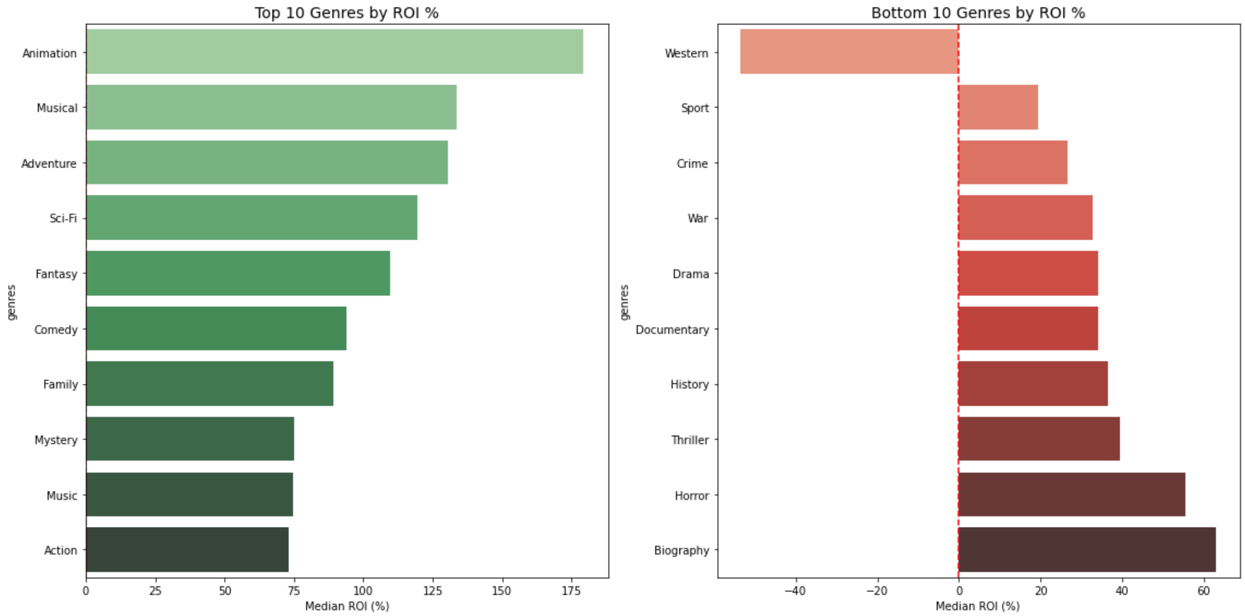
Most Profitable Movies per Genre



Key Insights and Observations

Action Genre is the Undisputed King of Profit: The most striking insight is the absolute dominance of the Action genre. Its profitability is more than double that of the next leading genre (Adventure) and surpasses the combined profit of the next five genres. This suggests that from a studio's financial perspective, investing in a successful action film offers a return on investment that is unmatched by any other genre.

Top Genres by Return on Investment



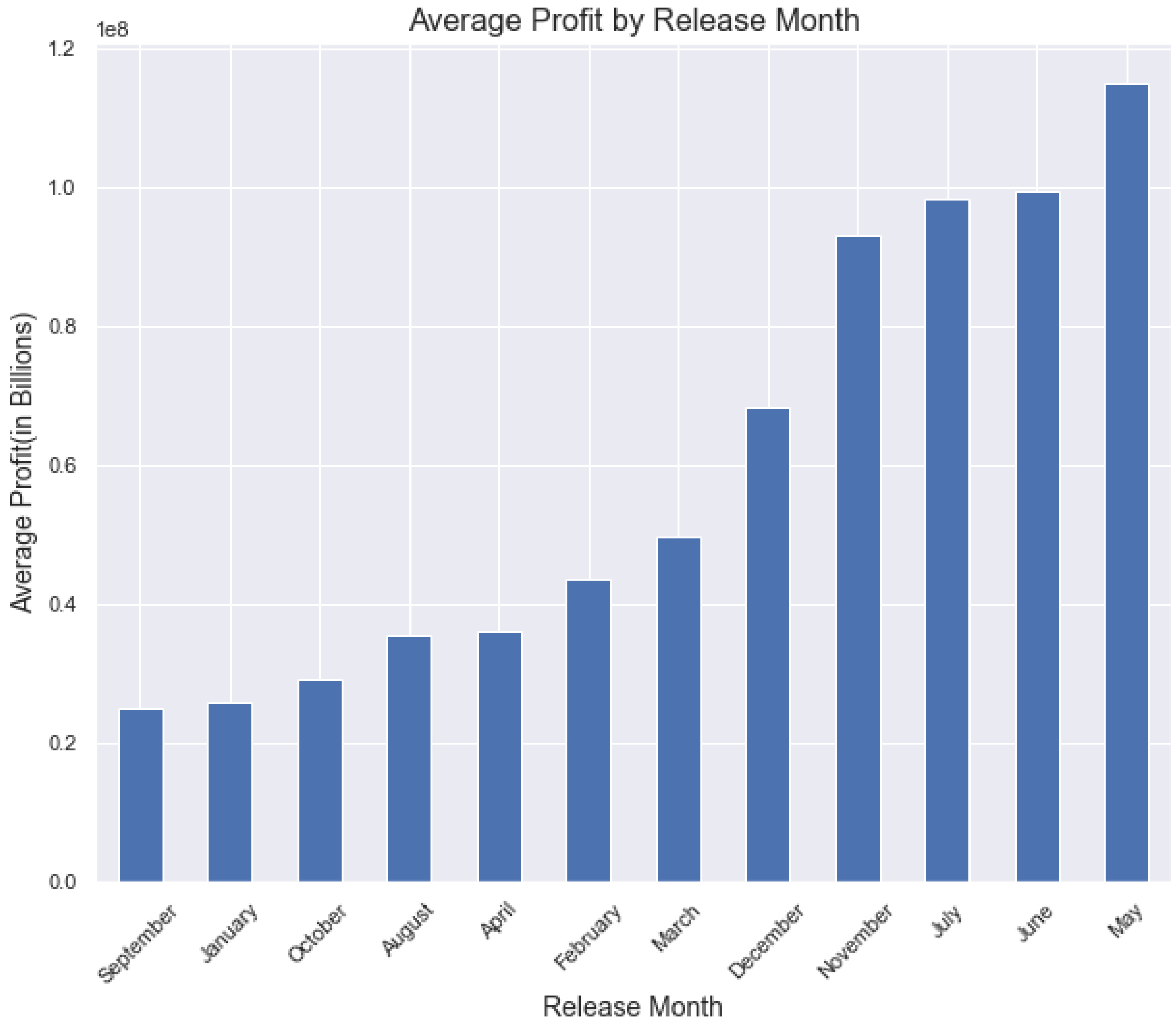
Key Observations and Insights

The top-performing genres are dominated by Animation, Musical, and Adventure. These genres, often targeted at families and broad audiences, deliver the highest percentage returns. Their success is likely driven by strong box office performance combined with lucrative merchandising and long-term replay value.

Although successful action blockbusters make enormous sums of money, the typical action film is incredibly expensive to produce.

This high cost reduces the percentage return on investment, making genres like Animation—which may have lower budgets—a more financially efficient investment on a film-by-film basis.

Average profit by Release Month



This bar chart shows the average profit of movies released in each month, helping reveal seasonal profitability trends in the film industry.

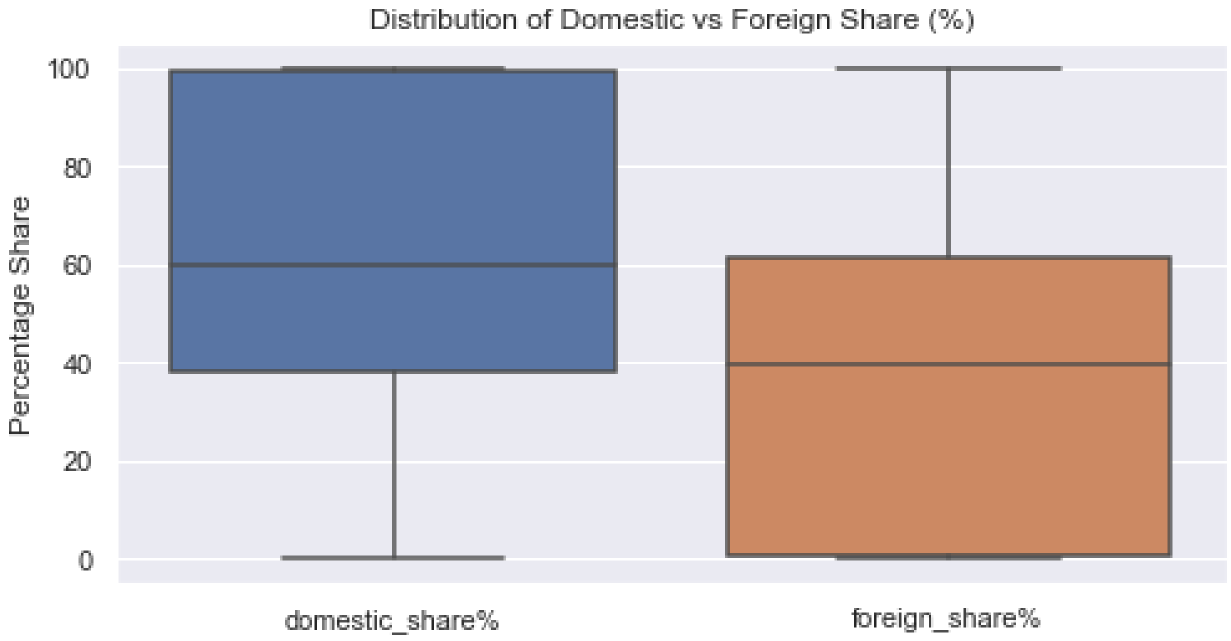
Key Insights*:*

- Top Performing Months:
  - May, June, July, and November have the highest average profits, close to or over $1 billion.
  - These months align with blockbuster seasons (summer and holidays).
- Moderate Months:
  - December and March also show strong profits.
  - February and April fall in the mid-range.

- Low Performing Months:
    - September, January, and October show the lowest average profits.
    - These months might be considered off-peak for major releases.

Distribution of Foreign Vs Domestic Revenue Share

The boxplot compares the percentage share of revenue from domestic and foreign markets.



Key Observations:

Domestic Share has a higher median than the foreign share, indicating that most movies earn a larger portion of their revenue domestically.

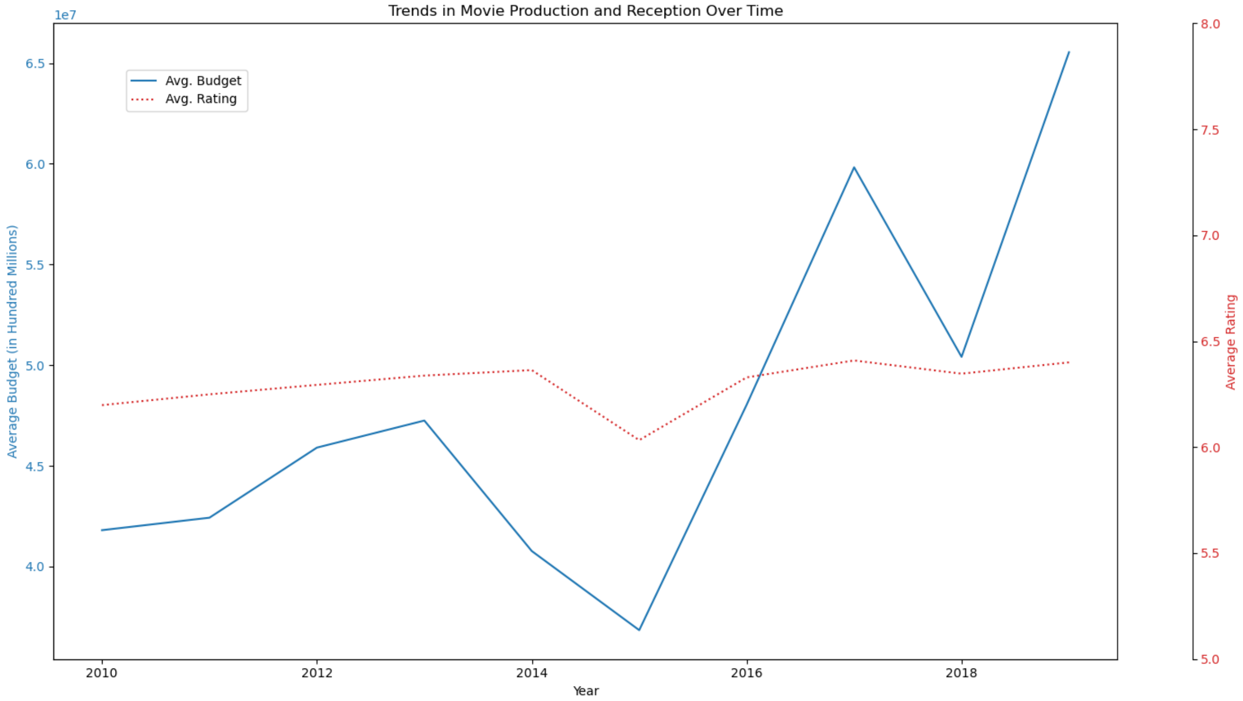
Foreign Share shows greater variability, with some movies earning almost all revenue from international markets.

The interquartile range (IQR) for domestic share is wider, but with fewer extreme outliers than the foreign share.

Insight:

While many movies tend to rely more on domestic earnings, the foreign market is highly variable and can be a significant revenue source for certain films.

Trends in Movie production and Reception Over Time



Observation

Ratings are Stagnant: Conversely, the average movie rating (in orange) has remained remarkably flat, hovering around 6.0-6.5.

This tells us that despite a significant increase in investment in film production over the last two decades, the average audience reception has not changed.

Hypothesis Testing of Movie Statistical Data

Do high budget films have higher Return on Investment(ROI) than low-budget films?

State the Null and Alternate Hypothesis

H_0 (null hypothesis): → High-budget films are less profitable than low budget films or have equal profitability

H_1 (alt hypothesis): → High-budget films are more profitable than low-budget films

Specify the significance level

The significance level is 0.05.

Calculate and get the test statitistic and p-value
T-statistic: 20.4893
P-value: 1.0299031654462523e-79

Interpret the p-value

Reject the null hypothesis: High-budget films are significantly more profitable than low budget films.

Performing A Chi-Squared Test: Analyzing Independence Between Categorical Variables

The Chi-squared test of independence is ideal for determining if there is a statistically significant association between two categorical variables.

Hypothesis: Is there a relationship between a movie's profitability (is_profitable) and the month it was released?

Null Hypothesis: A movie's profitability is independent of its release month.

Alternative Hypothesis: A movie's profitability is dependent on its release month.

Calculate and get the chi-squared statitistic and p-value
Chi-squared statistic: 56.74972253928975
P-value: 3.6997308209245245e-08
Degrees of freedom: 11

Results and Interpretation

The test results show a p-value of approximately 3.70 times(10 −8) since this value is much smaller than the standard significance level of alpha=0.05, we reject the null hypothesis.

This means there is strong statistical evidence to suggest that there is a significant relationship between a movie's profitability and its release month. The month in which a movie is released is not independent of whether or not it is profitable

ANOVA for testing multiple pairwise comparisons

Since we are comparing more than two groups, which in this case is ratings from movies released in multiple different months, ANOVA is the most appropriate test, because it allows us to see if at least one month has a significantly different average rating compared to the others, without having to run multiple pairwise t-tests.

Hypothesis Test: Does the month of release significantly affect a movie's average rating? This helps the studio decide which months to release their movies for better audience reception.

Null Hypothesis: There is no significant difference in average movie ratings across different release months.
Alternative Hypothesis: At least one month has a significantly different average rating compared to others.

Calculate and get the f statitistic and p-value
F-statistic: 2.392
P-value: 0.00621

Overview and Interpretation

After running the ANOVA test to determine whether the month of release has a significant impact on a movie's average rating, we obtained a p-value of 0.00621.

This p-value is less than the commonly used significance level of 0.05, which means that the results are statistically significant. In simpler terms, there's strong evidence to reject the null hypothesis which stated that there is no difference in average movie ratings across different months.

This result suggests that the month in which a movie is released does have a significant effect on how it is rated. Therefore, timing a movie's release could influence how well it is received by audiences.

Conclusions

Based on all the statistical tests we've done and visualizations we've created and analyzed, here is a summary of the key conclusions and actionable business recommendations for a film studio.

Financial Investment Does Not Guarantee Critical Success: There is no significant correlation between a movie's production budget and its average rating. We've seen that budgets have consistently risen over the last two decades, while average audience ratings have remained flat. This indicates that spending more money does not ensure a better-received film.

There is a "Sweet Spot" for Runtime: The vast majority of highly-rated films (between 6.5 and 8.0) have a runtime between 90 and 120 minutes. Films that are significantly shorter or longer are less common and don't hit this sweet spot of audience reception as frequently.

The Film Market is Crowded and Increasingly Expensive: Movie production saw a major boom in the 2010s and the market remains saturated. Combined with the rising costs of production, this makes it harder and more expensive than ever for a film to stand out and capture an audience.

Business Recommendations

Domestic vs Foreign Gross Shares

Some films generate the majority of their revenue from foreign markets, while others rely more on domestic audiences. Invest in market research to tailor content for global audiences and consider partnerships for international distribution.

Foreign Gross is Underutilized in Some Films

A subset of movies has minimal foreign gross compared to domestic, indicating untapped market potential. Improve foreign market strategy (subtitling, cultural adaptations, international trailers) to increase reach and ROI.

Target the 90-120 Minute Runtime

For films aiming for broad commercial appeal, directors and producers should aim for a final cut between 90 and 120 minutes. Our data clearly shows this is the optimal length to achieve favorable audience ratings, making it a less risky and more commercially viable choice.

Seasonality in Releases

Films released in certain months (e.g., summer or holiday seasons) tend to perform better. Plan major releases around high-traffic months (e.g., May–July, November–December). By aligning your release schedule with these findings, you can increase a film's chances of financial success. This may involve avoiding crowded release periods or targeting months with historically higher profitability.

Optimize Your Budgeting Strategy

Given the positive correlation between budget and gross revenue, a key recommendation is to be strategic with your investment. For films with high potential, consider a larger budget to increase the likelihood of a higher worldwide gross. However, this must be balanced against the risk of loss, so a thorough financial analysis is crucial.

Adopt a Niche Audience Strategy

Given the crowded market, trying to create a film for everyone can result in a film for no one. Instead, focus on creating films for specific, underserved audiences. As seen with the success of genre films, cultivating a loyal niche fanbase can be more profitable and sustainable

Prioritize Story over Spectacle

Since a large budget doesn't guarantee a good rating, resources should be strategically allocated. Shift focus from inflating budgets with expensive CGI or A-list salaries toward investing in high-quality scriptwriting and unique storytelling. A compelling story is a more cost-effective way to achieve critical success.