

FILM INDUSTRY PERFORMANCE AND CONTRIBUTOR IMPACT ANALYSIS

BUSINESS UNDERSTANDING

This project focuses on analyzing movie performance in the Entertainment and Media Sector to uncover patterns, trends and key drivers of commercial and critical success in the film industry. Since a single hit movie can make or break a studio's profits, the project will rely on structured data using variables such as release year, title, genre, production budget, worldwide revenue, audience rating and contributor roles (Directors and Writers) to guide smarter data driven decisions. Our insights help drive key areas like content strategy, budgeting and talent planning by revealing how different production choices influence both profits and ratings.

It is very tricky and uncertain to predict a movie's success, hence, hard for Studios to make smart decisions. Key areas they struggle with are talent evaluation, budgeting and strategic planning. Inadequate and unclear data often leads to misallocating production budgets, misjudging audience preferences and overlooking high impact talent which leads to poor performance.

Business Problem

Despite the rapid evolution of filmmaking technologies and global distribution platforms, many films still fail to achieve commercial success. Studios, distributors and investors lack a clear, data driven understanding of which factors such as genre, budget, runtime and contributors consistently influence a movie's performance. This gap in insight risks misallocating production budgets, misjudging audience preferences and overlooking high impact talent.

Project Objective

This project aims to identify trends in movie success across runtime, genre, seasonality and contributor roles. It will assess which directors and writers are most consistently associated with high ratings and strong box office returns, evaluate how budget levels correlate with profitability, assess how seasonality influence film profitability and explore how genre and runtime affect audience reception. The analysis will also highlight outliers and underperformers to inform future investment and marketing strategies.

The goal is to provide clear insights that will help studios make smarter choices especially in investing, hiring and marketing. There is need to build a data driven model to help understand what makes a movie successful by combining multiple datasets which goes through cleaning, and data analysis. The analysis will majorly focus on identifying the high impact contributors (Directors/Writers) by analyzing their track record, quantifying the relationship between budget, revenue and profitability and determining the performance influence of genre, audience ratings, season and movie length. This is aimed at coming up with a set of actionable insights and recommendations designed to improve strategic decision making across film investment, talent selection and marketing strategies.

The project will translate complex movie performance data into strategic, actionable intelligence. The findings will support improvements in Studio investment strategies, talent scouting and collaboration planning, genre specific budgeting and forecasting, audience targeting, release dates and runtime optimization

Data Understanding

The datasets used in this project are sourced from publicly available movie databases, including Box Office Mojo (BOM), The Numbers (TN), and The Movie Database (TMDb). Together, they provide structured information on thousands of films released over the past two decades. The data includes a rich mix of financial, creative, and audience-related attributes, enabling a comprehensive analysis of film performance and contributor impact. The key variables include; release year, runtime, production budget, worldwide gross, revenue, director, writer, release date, genre, average rating (vote_average), vote count, movie title. The data spans multiple formats; categorical variables such as genre, director, and writer, numerical variables such as budget, gross revenue, and ratings and temporal variables such as release year and runtime.

Data Preparation

The data will be cleaned and standardized by handling missing values, correcting inconsistent text formats, and aligning data types across key columns. This is aimed at ensuring reliable analysis of film performance and contributor impact and thorough cleaning and standardization of the movie datasets. These steps are essential for improving data quality, reducing ambiguity, and enabling accurate merging and analysis across multiple sources.

Cleaning Steps

Standardizing column names

Handling Duplicates

Handling missing values

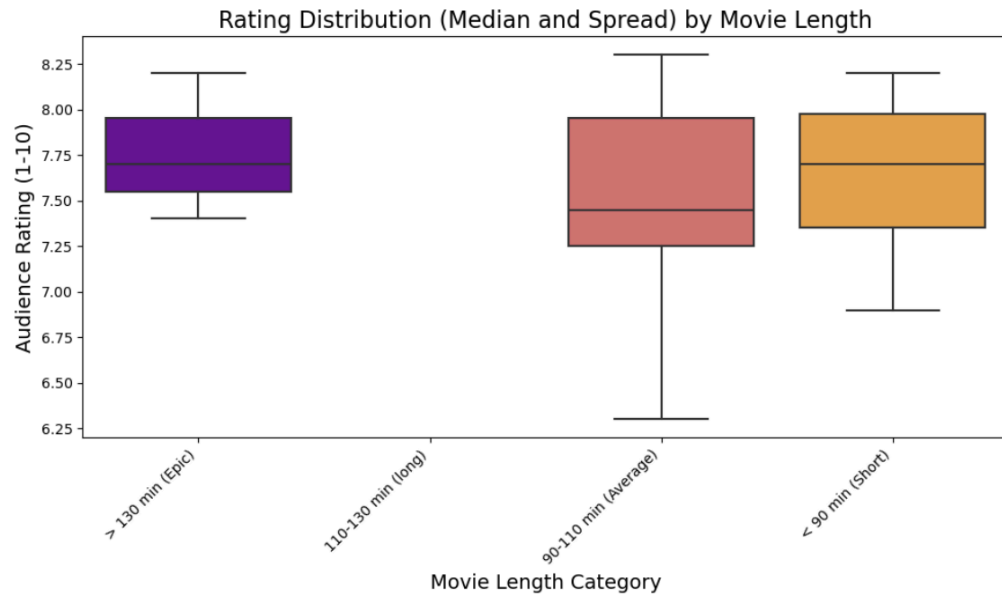
Data Analysis

Analysis of how the length of a movie affects audience ratings

The data set for movie length and ratings was merged using pandas in python based on the common movie Id column. The final combined columns were Id, runtime, vote average and vote count. The run time column was cleaned by converting string values into numeric minute values. Scatter plots showed the relationship between runtime and vote metrics. pearson correlation coefficients were calculated: runtime vs vote average and runtime vs vote count. Movies with low engagements were filtered out by requiring a minimum of 3 audience votes to ensure reliable average ratings. Distribution of vote_average and runtime was visualized and missing values in runtime identified and handled.

Runtimes were then grouped into four categories are highlighted below: short : < 90 minutes
average : 90 - 110 minutes long : 110 - 130 minutes Epic : > 130 minutes

we used boxplot to see the distribution of runtime compared to vote_average



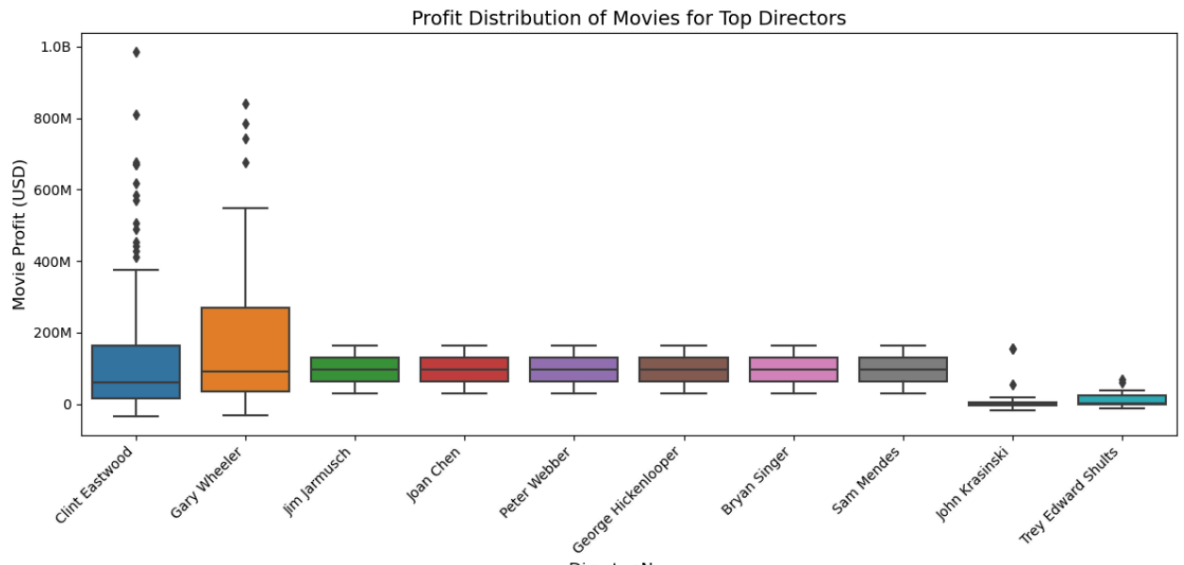
From the plot above, it is observed that epic length movies (130+ minutes) have the highest average vote rating (7.77) suggesting a positive correlation between runtime and audience ratings.

Analysis to identify directors and writers who make successful movies

We used the merged dataset to get the total profit per Director by grouping all movies by Director, summing the profit columns for each director and sorting the Directors from the highest to the lowest total profit.

We imported pandas and numpy for data manipulation and calculation, matplotlib.pyplot, seaborn for visualization and json, warnings for parsing and suppressing warnings.

This is aimed at identifying directors who generate the most total and average profit across their films.



From the plot above, it is observed that directors generating high revenue to should be considered for consistency and better returns e.g Clint Eastwood and Gary Wheeler.

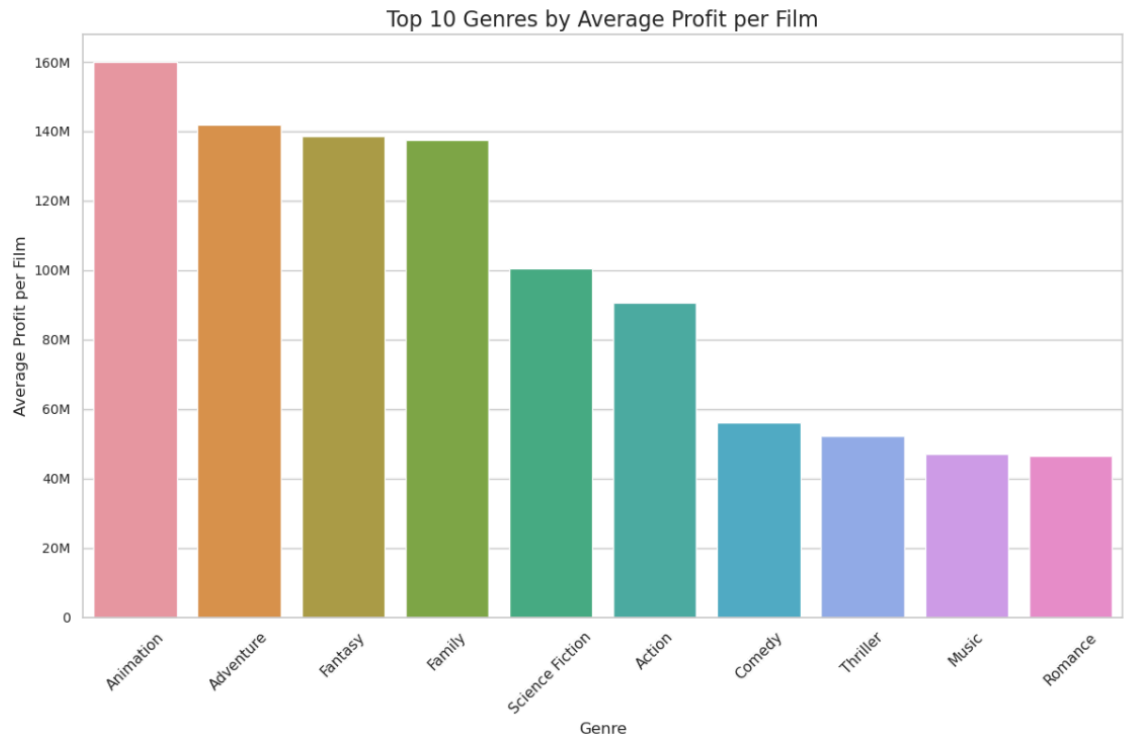
Analysis to establish which genres have the highest ratings & to assess which genres generate the highest gross revenue and profit

The analysis used pandas for aggregation and seaborn for visualization to analyze profitability by genre outlining the top 10 genres on the basis of both total and average profit.

Movies were grouped by genre and sum of profits for each was calculated and top 10 genres with the highest total worldwide profit was selected. This was aimed at identifying the genres that consistently generated high revenue across all films.

Average profit per film for each genre was also calculated to reveal which genres are most profitable on a per movie basis.

Visualization was done as below:

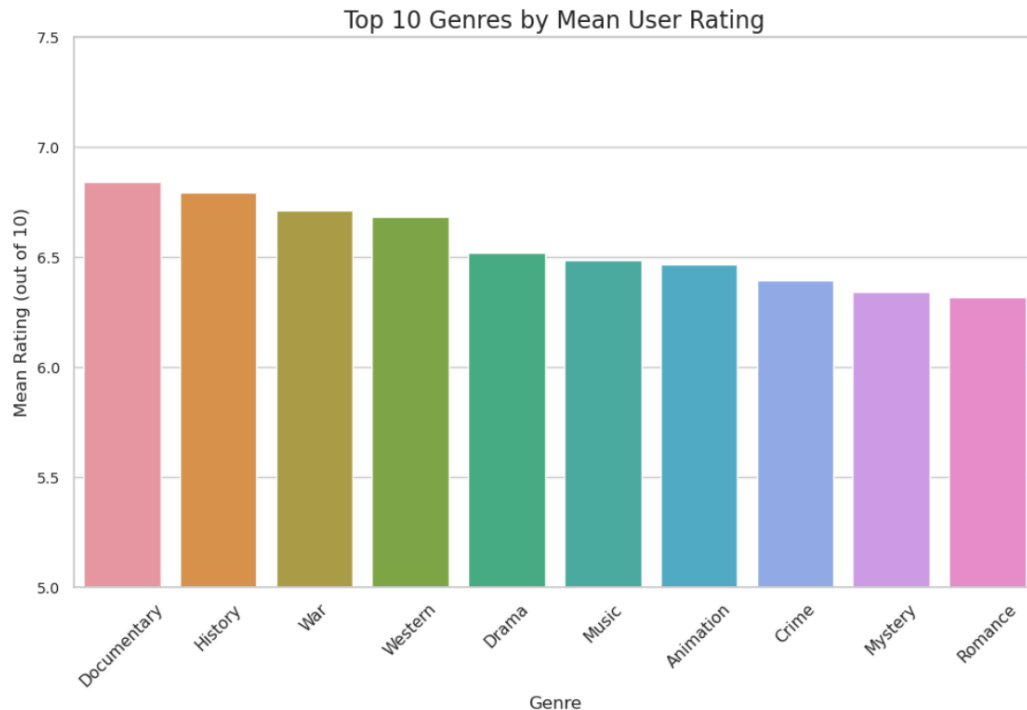


From the plot above, genre with large volume generate massive global revenue.

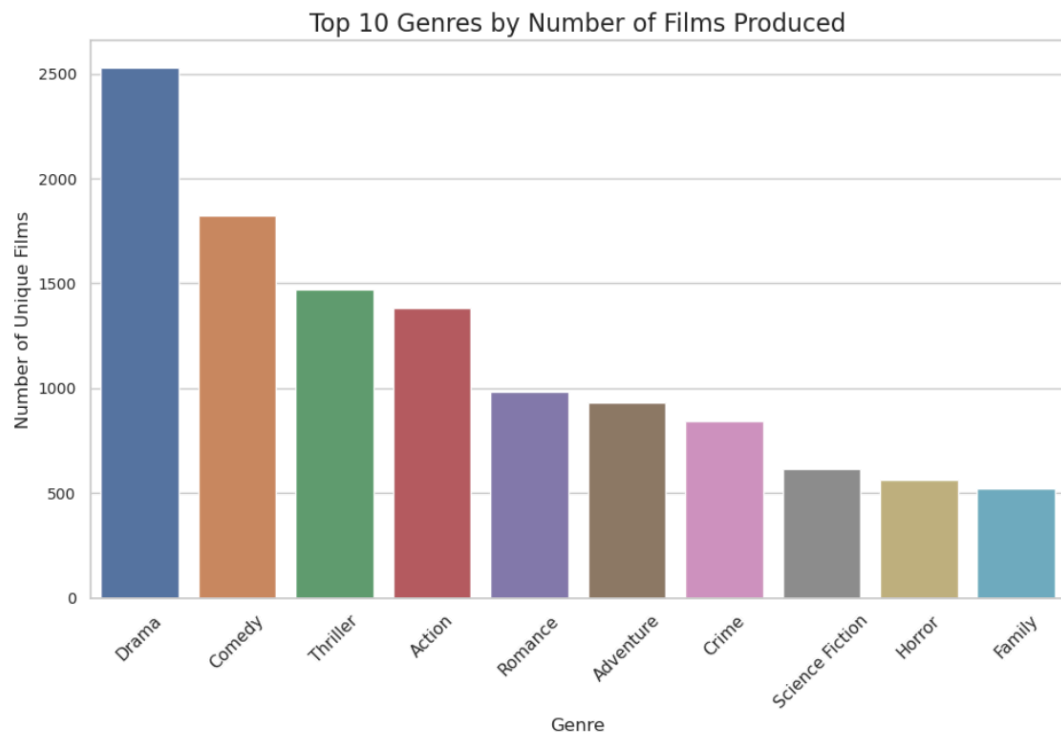
The analysis to reveal which genres are most favorably received and frequently produced by identifying top 10 movie genres based on user ratings and movie counts.

Visualization was done using Bar plot as below:

By mean user rating:



By Number of films produced:



It was observed that probably, high rated genres may reflect deeper story telling and high volume genres may indicate mainstreams popularity but with more variability in ratings.

Analyzing how release season of a movie affects its worldwide profit

The section focused on preparing data to analyze box office performance, including profit margins and release season effects. Profit was calculated by subtracting the production budget from worldwide revenue. Currency symbols were removed and data converted to numeric for both revenue and budget.

Movies that had missing values, unreliable data or near to zero values were excluded from the analysis.

Movie release dates were categorized into the following seasons:

Winter : December, January, February

CONCLUSIONS

1. **Movie length vs audience ratings:** The longer movies tend to be rated more favorably
2. **Directors and writers Vs Movie Performance:** The directors generating high revenue to be considered for better returns.
3. **Genres effects on ratings, gross revenue and profit:** Genres with high total profit are better options for studios aiming for consistent returns
4. **Movie realese season vs profit :** Studios to prioritize Summer and Spring for major release to maximize revenue

RECOMMENDATIONS

1. **Movie length vs audience ratings:** Content creators to consider longer movie run times, they probably give an allowance for deeper story telling.
2. **Directors and writers Vs Movie Performance:** The studio to consider prioritizing directors who have released movies that generated high revenues e.g Clint Eastwood and Gary Wheeler
3. **Genres effects on ratings, gross revenue and profit:** The studios to consider prioritizing genres with high total profit for consistent returns
4. **Movie realese season vs profit :** Seasonality should be considered in release planning and marketing strategies.