"From an Elevator Pitch to Box Office Success: Unlocking Cinema's Predictive Power"

Report By: -

Yashkumar Kalariya

Bassam Shazli

Mujtaba Nadeem

Faculty Advisor: -

Prof. Astrid Ayala

Fulfillment Statement:
A capstone project for partial fulfillment of the requirements of Master of Science in the Business Analytics program offered by Stetson-Hatcher School of Business at Mercer University.

Table Of Content

Business Question: Can film producers leverage analytics to confidently predict a movie’s success by analyzing controllable factors such as genres, budget, runtime, MPAA ratings, and release timing to forecast audience reception, critical acclaim, and blockbuster potential?

1. Introduction

The film industry, a cornerstone of modern entertainment, generates billions in revenue annually while facing significant financial risks. With average production costs soaring to $65 million per movie (Mueller, 2011) and only one-third of U.S. productions turning a profit (Moore, 2019), the need for data-driven decision-making in film investment has never been more critical.

This capstone project aims to predict film outcomes of revenue, audience reception and critics recognition to provide better investment strategies by analyzing the intricate relationships between genres, budgets, MPAA rating and release timing using a comprehensive dataset of movies from 2010 to 2018. Our research is motivated by the high-stakes nature of film production, where a single project can result in either substantial financial gains or devastating losses. By leveraging advanced data analysis techniques and predictive modeling, we seek to:

1. Identify key factors that contribute to a film's financial success and audience reception.
2. Develop actionable insights to guide producers and investors in making informed decisions.
3. Minimize investment risks while maximizing potential returns in an increasingly competitive market.

Our study builds upon existing research while incorporating novel approaches to data analysis, aiming to provide a nuanced understanding of the dynamics that drive box office performance. By doing so, we intend to equip stakeholders with the tools necessary to navigate the uncertainties inherent in film production and distribution.

2. Literature Review

The film industry's high-risk investment landscape has prompted extensive research into optimizing resource allocation and predicting movie success. Our literature review synthesizes findings from various studies that employ machine learning (ML) techniques and statistical analyses to identify critical factors influencing a film's performance.

Ensemble Machine Learning Approaches:
Gupta et al. (2022) utilized max-voting and boosting-based algorithms to analyze movie success. Their study highlighted the significance of:

- Movie ratings as a critical attribute
- Release month as a key factor in audience engagement
- The impact of multi-genre films on viewer interest
- Movie duration's role in generating excitement
- The complex relationship between popularity and audience opinions

Regression Models and Key Predictors:
Yoo, Kanter, and Cummings (2011) employed linear and logistic regression models, identifying budget as the strongest success indicator. Their findings contrasted with Gupta et al. (2022) by emphasizing audience size over ratings as a primary driver of gross profit.Comparative Model Performance:
San (2020) compared various ML models, including linear regression, polynomial regression, decision trees, and random forests. Multiple linear regression emerged as the most effective (R-squared = 0.451), closely followed by second-degree polynomial regression (R-squared = 0.443).

Advanced Techniques and Feature Selection:
Paul and Das (2022) combined random forests, support vector regression, gradient boosting, and artificial neural networks with stepwise regression, Lasso, and Ridge regression to mitigate collinearity and identify significant predictors. Their study underscored the importance of star power and film-specific characteristics in revenue generation.

Influential Studies Shaping Our Approach:

1. Dalton and Leung(2019) emphasized the role of critic reviews in box office performance, highlighting the importance of qualitative assessments alongside quantitative metrics. The authors used machine learning techniques, specifically random forests, to predict box office success. Their model achieved an R-squared of 0.83 for predicting opening weekend box office revenue.
2. Sharma et al.(2020) demonstrated the effectiveness of Random Forest models in predicting box office revenues and capturing nonlinear relationships between variables like budget and ROI. The project compared multiple algorithms, including Linear Regression, Decision Trees, Random Forest, and XGBoost.

Random Forest and XGBoost performed best, with R-squared values of 0.85 and 0.87 respectively.

3. Hagen and Jøsok(2021) underlined the significance of release timing and seasonality in determining financial performance, providing a foundation for integrating temporal analysis into our research. The project primarily used correlation analysis rather than machine learning models. They focused on examining the relationship between critical reviews and box office performance.

Methodology Selection:

Based on the literature review, we have chosen to employ multinomial logistic regression and polynomial regression, Random Forest, XGBoost and Decision Tree models as our primary analytical tools. This approach allows us to:

- Evaluate model efficiency
- Enhance result accuracy
- Avoid overfitting or underfitting issues observed in simpler or more complex models

Key Determinants of Success:

Our review consistently highlights several factors as significant predictors of box office success:

- Production budget
- Release timing and seasonality
- Movie duration
- Genre (particularly multi-genre films)
- MPAA ratings (Motion Picture Association of America)
- Critic and audience reviews

However, it's important to note that no single factor guarantees success, emphasizing the need for a multifaceted approach to film investment strategy.

Our literature review underscores the potential of data-driven forecasting models in aiding studios to allocate resources efficiently amidst the inherent risks of filmmaking. Our research aims to build upon these findings, focusing on the interplay between genre, budget, and release timing to provide actionable insights for optimizing film investments in the contemporary market landscape.

3. Data

Dataset Source and Description

The dataset used in this research was obtained from a GitHub repository (https://github.com/ntdoris/movie-revenue-analysis/tree/main). With 1765 total number of observations, it combines data from IMDb, TMDb, and Box Office Mojo, covering films released between 2010 and 2018. The dataset includes financial, audience, and productionlevel details, adjusted for inflation for consistency.

Data Dictionary:

Here's a table describing the key variables in our dataset:

Variable Name	**Description**
movie	Title of the movie
year	Release year of the movie
production_budget	Original budget for the movie production
domestic_gross	Gross revenue from domestic (U.S.) market
foreign_gross	Gross revenue from foreign markets
worldwide_gross	Total worldwide gross revenue
month	Release month of the movie
profit	Original profit (worldwide_gross - production_budget)
roi	Return on Investment
popularity	Popularity score of the movie
vote_average	Average vote rating
vote_count	Number of votes received
genres	List of genres associated with the movie
IMDb_Rating	Rating from IMDB

Variable Name	**Description**
Distributor	Distribution company
Director	Director of the movie
MPAA_Rating	Motion Picture Association of America rating
Runtime	Duration of the movie in minutes
Critic_score	Critics' rating score taken from Rotten Tomato
cpi	Consumer Price Index for the release year
production_budget_adj	Production budget adjusted for inflation
worldwide_gross_adj	Worldwide gross adjusted for inflation
profit_adj	Profit adjusted for inflation
roi_adj	Return on Investment adjusted for inflation

Data Cleaning Process:

The data cleaning process was comprehensive and aimed at ensuring the quality and relevance of our dataset for analysis. We began by focusing on the time period of interest (2010-2018) and adjusting financial data for inflation to ensure comparability across years.

A significant part of our cleaning process involved removing entries that lacked crucial information or could potentially skew our analysis. This included removing duplicate entries, movies with no reported revenue (either domestic or foreign), and those without specified genres or MPAA ratings. These steps were crucial in ensuring that our analysis would be based on movies that had a genuine commercial release and reception.

We also refined our dataset by removing variables that were no longer relevant after our cleaning steps, such as the 'TV Movie' category. This streamlining of the dataset helps to focus our analysis on the most pertinent factors affecting theatrical movie performance.

The result of this cleaning process is a more robust and reliable dataset, focused on commercially released and rated movies with reported financial performance in both domestic and international markets. This cleaned dataset provides a solid foundation for our subsequent analyses, ensuring that our insights are based on complete and relevant data points.

By removing potentially misleading or incomplete entries, we've increased the reliability of our dataset, allowing for more accurate analyses of the relationships between movie characteristics (such as genre, budget, and release timing) and their financial performance.

Feature Engineering

Our feature engineering process involved creating new variables and categorizing existing ones to enhance our analysis. Here's a detailed breakdown of each step:

1. Seasonal Categorization:-

   We categorized release months into seasons:

   a) Winter - December, January, February

   b) Spring - March, April, May

   c) Summer - June, July, August

   d) Fall - September, October, November

2. Profit Adjusted Margin

   We calculated an adjusted profit margin based on inflation-adjusted figures:

   Code: profit_adj_margin = profit_adj / worldwide_gross_adj

3. Genre Count

   We created a count of genres associated with each movie based on 18 genre categories.

4. Runtime Categories

   We categorized movies based on their runtime:

   a) Less than 90 - ≤ 90

   b) 90 to 135 - > 90 to ≤ 135

   c) Greater than 135- > 135

5. Profit Categories

   We categorized movies based on their profitability:

   a) Loss - profit_adj ≤ 0

   b) Break-even - 0 < profit_adj ≤ production_budget_adj

   c) Profitable - production_budget_adj < profit_adj ≤ 2 * production_budget_adj

   d) Successful - profit_adj > 2 * production_budget_adj

This feature engineering process created a rich set of variables for our analysis, allowing us to explore various aspects of movie characteristics and their potential impact on financial performance. These engineered features provide a multifaceted view of each movie in our dataset, enabling a more nuanced analysis of the factors influencing movie success.

Exploratory Analysis

In this section, we analyze key relationships and trends in the dataset using visualizations. Each visualization provides insights into factors influencing movie performance, including production budgets, worldwide gross revenue, and year-over-year profit trends.

Facet Grid: Profitability Trends Across Seasons: (Fig. 1)

Key Observations:

1. **Seasonal Profitability Variations**:
    - Summer consistently shows the highest peaks in profitability across the years, aligning with the release of blockbuster films during this season.
    - Winter and Fall also display notable profitability spikes, though less frequent than summer.
    - Spring tends to have lower profitability peaks compared to other seasons.
2. **Annual Trends Within Seasons**:
    - In all seasons, profitability fluctuates year-to-year, with some standout years showing much higher profits.
    - Peaks within each season indicate the release of highly profitable movies during specific years.
3. **Summer Dominance**:
    - The highest individual profits are recorded in the summer season, underscoring its importance for the movie industry.
4. **Lower Consistency in Spring and Fall**:
    - While some Spring and Fall releases achieve high profitability, these seasons exhibit more variability and lower average profits.

Potential Insights:

1. **Seasonal Blockbuster Strategy**:
    - Studios heavily rely on the summer season for high-grossing releases, possibly leveraging school vacations and favorable weather for increased theater attendance.
    - Winter also shows significant profits, likely driven by holiday releases.
2. **Opportunities for Spring and Fall**:
    - Lower average profitability in Spring and Fall could indicate opportunities for counter-programming strategies or releasing niche films with less competition.
3. **Impact of Specific Movies**:
    - The spikes in profitability during certain years suggest that individual movies (likely blockbusters) drive seasonal trends.
    - Analyzing these movies could provide further insights into genre or budget impacts on seasonal performance.

Interpretation of Box Plot: IMDb Rating by MPAA Rating: (Fig. 2)

The box plot comparing IMDb ratings across MPAA rating categories reveals several interesting patterns:

Similar Median Ratings:
PG, PG-13, and R movies have comparable median IMDb ratings, suggesting that MPAA ratings don't significantly influence audience perception of quality. This aligns with findings that IMDb scores don't strongly correlate with specific content ratings.

Rating Distribution Variability:
R and PG-13 movies show a wider range of IMDb ratings, including lower outliers. This broader distribution likely reflects the diverse quality and audience reception within these popular categories.

NC-17 and G Ratings:
NC-17 movies have a narrow range of IMDb ratings, possibly due to fewer films or niche audiences. G-rated movies show a relatively high median but also exhibit lower ratings, indicating variability in audience reception for family-friendly content.

Potential insights include:

1. Content rating alone doesn't determine perceived quality, as measured by IMDb scores.

2. R and PG-13 categories encompass a wide range of movie qualities, likely due to the volume of productions in these categories.
3. G and NC-17 ratings may cater to more specific audience expectations or niche markets.

This analysis can inform marketing strategies, quality improvement efforts for lower-rated films within each category, and potentially contribute to predicting IMDb ratings based on MPAA categories, though additional factors would be necessary for accurate predictions.

Sankey Diagram: Flow of Movies by Main Genre to Profit Category: (Fig. 3)

The Sankey diagram visualizing the flow of movies from main genres to profit categories reveals several key insights about genre profitability:

Highly Profitable Genres:
Animation, Fantasy, and Action movies show a higher proportion of Successful outcomes, likely due to their wide appeal and global revenue potential. Family and Adventure genres also have significant flows towards Profitable and Successful categories, indicating their ability to attract large audiences.

Loss-Prone Genres:
Documentary, Crime, History, and Music genres appear to have a higher proportion of movies in the Loss category. This suggests these genres may have smaller target audiences or limited commercial appeal.
Mixed Success:
Drama, being a diverse and widely produced genre, shows connections across all profit categories. This indicates variable market performance depending on factors like storyline and production quality.

Break-even Genres:
Thriller, Science Fiction, and Mystery genres often land in the Break-even category, reflecting moderate financial performance.

This analysis can be applied in several ways:

1. Strategic Investments: Investors might focus on genres like Animation and Fantasy that demonstrate higher profitability, while being cautious with genres like Documentary and Music.
2. Target Audience Optimization: For loss-prone genres, producers could experiment with niche marketing strategies to effectively target specific audience segments.
3. Content Diversification: Production houses may use this data to diversify

their portfolios by investing in a mix of highly profitable and experimental genres.

These insights can guide decision-making in film production and investment, helping stakeholders balance potential profitability with creative risk-taking across different genres.

Worldwide Gross Adj by MPAA Rating: (Fig. 4)

- G-rated movies generate the highest worldwide gross, indicating strong appeal to a universal audience.
- PG-rated movies follow closely, emphasizing the financial success of family-oriented content.
- NC-17 movies show minimal financial returns, likely due to limited audience reach and distribution challenges.

Critic Score vs. Audience Reception: (Fig. 5)

Positive correlation between critic scores and audience ratings.

Higher critic approval tends to align with better audience reception.

Emphasizes the role of critics in shaping viewer perceptions and potential market reception.

Variance in Worldwide Gross Adjusted by Genre: (Fig. 6)

"Family" and "Science Fiction" genres show highest variance, indicating inconsistent revenue outcomes and higher financial risk.

"Documentary" and "Music" genres display minimal variance, suggesting more stable but lower revenue performance.

Highlights the trade-off between potential high returns and financial risk in different genres.

Boxplot of production budgets by genre: (Fig. 7)

The boxplot emphasizes clear financial disparities:

High-Budget Genres:
Action, Adventure, and Science Fiction stand out with significantly higher median budgets. These genres often rely on:

- Cutting-edge technology
- Extensive special effects
- High-profile casts
- Global marketing campaigns

Low-Budget Genres:
Documentary, Comedy, and Horror demonstrate lower budget medians. These genres can often achieve success through:

- Minimalistic setups
- Character-driven plots
- Smaller production teams

These insights underscore that genre choice is a crucial factor in determining the scale of financial investment required for a film. The data suggests a strategic approach to budgeting based on genre expectations and potential returns.

Implications for Filmmaking

1. Risk Management: High-budget genres carry greater financial risk but also potential for higher returns, while low-budget genres offer more consistent, if modest, profitability.
2. Genre Blending: Filmmakers might consider blending elements of high and low-budget genres to balance creative ambition with financial prudence.
3. Technology Impact: The significant budget differences highlight how technological requirements in genres like Science Fiction drive up costs, emphasizing the need for careful planning in these areas.
4. Market Positioning: Understanding these budget dynamics can help studios and independent filmmakers position their projects more effectively in the market, aligning production scale with audience expectations and potential revenue.

Data Reduction and further Feature Engineering

After investigating through data exploratory analysis we had to remove outliers and make certain variables into categories and do other feature engineering to make a dataset more prepped for model building.

Outlier Analysis:

- The original data showed a wide range in worldwide gross adjusted, from $28,960 to $2,048,000,000.
- 128 outliers were identified and removed using the IQR method.
- After removal, the range of worldwide gross adjusted is now from $28,956 to $448,808,221.

Implications of Outlier Removal:

1. More Representative Analysis: Removing outliers helps focus on typical movie performance, reducing

the influence of exceptional blockbusters or major flops.

2. Improved Statistical Reliability: Outlier removal often leads to more robust statistical analyses and predictions.
3. Loss of Extreme Cases: We've lost information about exceptionally successful or unsuccessful movies, which might be valuable for certain types of analysis.

Main Genres Distribution:

- Drama (315 movies) and Comedy (236 movies) are the most common main genres.
- Action (196 movies) is also well-represented.
- Some genres have very few entries, like Western (1 movie), Music (5 movies), and History (9 movies).

Log Transformation of Worldwide Gross Adjusted:

The log transformation of worldwide gross adjusted has been applied, creating a new variable 'Log_worldwide_gross_adj'. This transformation helps to:

- Reduce the range of the dependent variable, making it more manageable for analysis.
- Potentially normalize the distribution of the variable, which can be beneficial for many statistical analyses.
- Make relationships more linear, which is often an assumption in regression analyses.

The summary of Log_worldwide_gross_adj shows:

- Minimum: 10.27
- 1st Quartile: 16.72
- Median: 17.92
- Mean: 17.56
- 3rd Quartile: 18.77
- Maximum: 19.92

This transformation has indeed compressed the range of values, making the distribution more suitable for various statistical analyses.

Creation of Dummy Variables for Runtime Categories:

Dummy variables have been created for runtime categories:

- "90 to 135" minutes
- "Greater than 135" minutes

The category "Less than 90" minutes is likely being used as the reference category, which is a common practice in regression analyses.

Implications and Recommendations:

1. Improved Statistical Analysis: The log transformation of worldwide gross will likely lead to more robust statistical models, especially in regression analyses. This could provide more reliable insights into factors affecting movie performance.
2. Interpretation of Results: When using the log-transformed variable, remember that effects will be interpreted in terms of percentage changes rather than absolute dollar amounts.
3. Runtime Analysis: The creation of dummy variables for runtime categories allows for a more nuanced analysis of how movie length affects performance. This could lead to insights such as:
    - Whether movies between 90 to 135 minutes perform differently from shorter or longer films.
    - If there's a significant difference in performance for movies longer than 135 minutes.

Dummy Variables for Seasons:

- Created for Spring, Summer, and Fall (Winter is likely the reference category)

Allows for analysis of seasonal effects on movie performance

Dummy Variables for MPAA Ratings:

- Created for PG-13, R, PG, and G

Enables analysis of how different ratings impact movie performance

Dummy Variables for Genre Counts:

- Created for 1 to 6 genres

Facilitates analysis of how the number of genres affects a movie's performance

Log Transformation of Production Budget:

Similar to the log transformation of worldwide gross, this helps normalize the distribution of budget data.

Updating variable names:

These changes are important for ensuring consistency in the dataset and preventing errors in future analyses. Here's a summary of the key changes:

1. MPAA Ratings:
    - "PG-13" renamed to "PG.13"
    - "NC-17" renamed to "NC.17"

2. Runtime Categories:
    - "90 to 135" changed to "90.to.135"
    - "Greater than 135" changed to "Greater.than.135"
    - "Less than 90" changed to "Less.than.90"
3. Genre Names:
    - "Main_Science Fiction" renamed to "Main_Science_Fiction"
    - "Science Fiction" renamed to "Science_Fiction"
4. Other Changes:
    - "90 to 135" column renamed to "between_90_to_135"
    - "Greater than 135" column renamed to "Greater_than_135"

These changes will help in:

- Avoiding syntax errors in R code, especially when using these variables in formulas or function calls.
- Maintaining consistency across the dataset, which is crucial for accurate analysis and interpretation.
- Improving readability and reducing the chance of mistakes in data manipulation and modeling stages.

4 Methodology:

Our first approach was rather instinctive; that is, to evaluate the monetary value of each movie and find out the nature of Return on Investment. We built models, through the stepwise method, linear regressions, neural networks and decision trees. Here, we found very few significant variables and the genre “Horror” was the only positive significant variable when it came to ROI with all others negative. However, this is not the nature of films or the film industry.

As film producers we cannot only pitch an idea for horror films. Yes, Horror films do way better than other films, with less money involved, but filmmaking is an artform that relies on many other genres and the audience preference. Therefore we had to change our approach and see worldwide gross as the main factor. Worldwide gross in movies measures how much a movie has made in gross revenue, giving us a more diverse answers to what kind of films do good.

Our approach centred on selecting those variables that are under the control of the film producer, like, Production budget, Genre, MPAA ratings, release timing and runtime, these became our key factors (independent variables), which would help in solving the uncertain variables; Worldwide gross, IMDB rating and Critics Ratings; which define the monetary value of the film and also quantify the movie. By quantifying, we mean that, as a subjective artform, the audience reception and critics ratings define, whether a movie is worth a watch and eventually will the film remain a significant asset to film producers.

Summary for Dependent Variables and Model Selection

This section summarizes the methodology employed to analyze three dependent variables—
IMDB_Category, Critic_score_category, and Log_worldwide_gross—using various machine learning models. The goal is to classify films based on these variables and identify significant predictors that influence these classifications.

1. IMDB_Category

- **Best Model**: **Random Forest**
- **Verified Significant Variables**:
    - **Log_production_budget_adj**: A dominant predictor indicating that budget heavily influences the IMDB category.
    - **PG.13 and R**: Age ratings significantly impact categorization, reflecting audience segmentation.
    - **Genre_count**: Diversity in genres is an essential factor.
    - **Main_Drama**: The drama genre has a consistent impact on ratings.
    - **Between_90_to_135**: Movie duration in this range strongly affects categorization.
    - **Seasonal Variables (Spring, Summer, Fall)**: Timing of releases is important for success.

The Random Forest model effectively captures the relationships among these predictors, providing robust performance in categorizing films into their respective IMDB categories.

2. Critic_score_category

- **Best Model**: **XGBoost**
- **Verified Significant Variables**:
    - **Log_production_budget_adj**: The most crucial variable across all evaluation metrics, showing its significant effect on critic scores.
    - **Genre_count**: Indicates how the diversity of genres impacts critics' evaluations.

    - **Main_Drama**: The drama genre is consistently favored by critics.
    - **Between_90_to_135**: Movies in this duration range are favored.
    - **PG.13, R, and G**: Age ratings are significant predictors, reflecting audience targeting and content alignment with critics' preferences.
    - **Seasonal Variables (Spring, Fall, Summer)**: Timing strongly influences critic perceptions.

XGBoost excels in managing complex interactions between features and handling imbalanced datasets, making it suitable for predicting critic score categories.

3. Log_worldwide_gross

- **Best Model**: **XGBoost**
- **Verified Significant Variables**:
    - **Log_production_budget_adj**: Consistently the most important variable, as it directly affects revenue.
    - **Main_Action, Main_Adventure, Main_Comedy and Main_Drama**: These genres dominate revenue performance.
    - **Genre_count**: A diverse set of genres correlates with higher revenue.
    - **Between_90_to_135 and Greater_than_135**: Longer movie durations are critical for maximizing revenue.
    - **Seasonal Variables (Spring, Summer)**: Timing affects box office success.
    - **PG.13 and R**: Age ratings are key to targeting the right audience for maximum revenue.

The XGBoost model demonstrates strong performance in predicting worldwide gross revenues due to its ability to capture complex relationships among features.

Conclusion

Based on the evaluations of the models for each dependent variable:

- For IMDB_Category, the selected model is **Random Forest**, which provides robust performance and effectively captures significant predictors influencing audience ratings.

- For both Critic_score_category and Log_ worldwide_gross, the selected model is **XGBoost**, which excels in handling complex interactions and managing imbalanced datasets.

These model selections align well with the business objective of maximizing returns on film investments by providing accurate predictions based on key factors that drive audience reception and critical evaluations. Further optimization of these models may enhance their predictive capabilities, particularly for underperforming categories like "Moderate."

Model Building for Log_Worldwide_Gross_Category

Log_Worldwide_Gross_Category ~ Log_production_budget_adj + PG.13 + R + PG + G + between_90_to_135 + Greater_than_135 + Spring + Summer + Fall + genre_count + Main_Action + Main_Adventure + Main_Animation + Main_Comedy + Main_Crime + Main_Documentary + Main_Drama + Main_Family + Main_Fantasy + Main_Horror + Main_Mystery + Main_Romance + Main_Science_Fiction + Main_Thriller + Main_History

Name	Multinomial Logistic Regression	Random Forest	XGBoost	Polynomial Regression	Decision tree
Significant Variables	Log_production_budget_adj, PG.13, PG, R, Main_History	Log_production_budget_adj, Main_Action, R, PG.13, genre_count, Main_Drama, Main_Romance	Log_production_budget_adj, genre_count, Main_Horror, Fall, R, Main_Comedy, Summer, PG.13,	poly(Log_production_budget_adj, degree = 2, PG.13, R, PG, Main_History	Log_production_budget_adj, genre_count, Main_Action, PG, Main_Adventure, Main_Drama, Main_Comedy, Fall, R, PG.13

			Main_Action, Main_Drama		
Sensitivity High’s	0.6829	0.6911	0.7154	0.6423	0.6829
Sensitivity Medium	0.4622	0.4118	0.5294	0.5126	0.4790
Sensitivity Low’s	0.7105	0.6754	0.6316	0.7018	0.6491
Specificity High’s	0.8240	0.8112	0.8455	0.8584	0.8755
Specificity Medium	0.7637	0.7764	0.7257	0.7300	0.7257
Specificity Low’s	0.8388	0.8017	0.8678	0.8388	0.8058
Balanced Accuracy High’s	0.7535	0.7511	0.7805	0.7503	0.7792
Balanced Accuracy Medium	0.6129	0.5941	0.6276	0.6213	0.6024
Balanced Accuracy Low’s	0.7747	0.7385	0.7497	0.7703	0.7275
ROC / AUC High’s	0.8448306	0.8363341	0.8632018	0.8523326	0.8352176
ROC / AUC Medium	0.6920009	0.6586888	0.6993937	0.7074957	0.6504627

ROC / AUC Low's	0.8601747	0.8235283	0.8365412	0.8615884	0.8079237

Model Selection for Log_Worldwide_Gross_Category:

XGBoost stands out as the best model for predicting Log_Worldwide_Gross_Category due to its high accuracy, effective handling of complex feature interactions, and scalability for large datasets. These attributes align perfectly with the business goal of maximizing returns on film investments by enabling accurate predictions based on a comprehensive analysis of various influential factors.

IMDB_Category model Building:

Name	Multinomial Logistic Regression	Random Forest	XGBoost	Polynomial Logistic Regression	Decision Tree
SignificantVariables	PG-13, R, PG, G, Main_Family, Main_Horror, Main_Mystery	Log_production_budget_adj, Main_Drama, between_90_to_135, Main_Horror,genre_count Greater_than_135, Fall, Main_Comedy, Spring	Log_production_budget_adj, genre_count,Main_Drama, Fall,between_90_to_135, R,Main_Horror,Spring,PG.13, Summer	Log_production_budget_adj, PG.13, R, PG, G, Spring, Summer, Fall, genre_count, Main_Drama, Main_Horror, Main_Family, Main_Mystery	Log_production_budget_adj,Main_Drama,genre_count,between_90_to_135, Greater_than_135,R,PG.13, Fall, Summer, Spring,Main_Horror, Main_Crime, Main_Thriller

Sensitivity Excellent	0.35135	0.33784	0.4459	0.33784	0.25676
Sensitivity Good	0.5767	0.6380	0.3129	0.5706	0.5215
Sensitivity Poor	0.3950	0.3782	0.5294	0.3697	0.4874
Specificity Excellent	0.85816	0.86525	0.6773	0.85106	0.86879
Specificity Good	0.4663	0.4404	0.7254	0.4456	0.5130
Specificity Poor	0.8059	0.8481	0.7257	0.8101	0.7342
Balanced Accuracy Excellent	0.60475	0.60154	0.5616	0.59445	0.56278
Balanced Accuracy Good	0.5215	0.5392	0.5191	0.5081	0.5172
Balanced Accuracy Poor	0.6004	0.6131	0.6276	0.5899	0.6108
ROC / AUC Excellent	0.6775925	0.7138442	0.5472733	0.6811865	0.7017203
ROC / AUC Good	0.5396548	0.5467275	0.5122223	0.5362853	0.5223148
ROC / AUC Poor	0.6835798	0.6822324	0.6884551	0.6851044	0.6566855

For the dependent variable IMDB_Category, the best model is Random Forest.

Robust Performance: Random Forest provides consistent and reliable accuracy, ensuring a balanced prediction across all IMDB categories. This minimizes risks in film investment by avoiding overfitting to specific patterns.

Feature Importance: The model naturally ranks features like Log_production_budget_adj, Main_Drama, and genre_count, which are crucial for understanding and predicting the profitability of films. This helps focus on key factors that drive returns.

Handling Complexity: Random Forest effectively handles nonlinear relationships and interactions between features like budget, genre, and seasonal trends, providing insights into complex dependencies that impact film success.

Interpretability: Unlike other models, Random Forest offers clear insights through variable importance plots and easy-to-understand predictions, enabling better decision-making for investments.

Robust to Noise: It is highly robust to outliers and noise in the data, ensuring stable predictions even with imperfect historical data, making it ideal for predicting returns in uncertain film markets.

Conclusion: Random Forest aligns with the business goal of maximizing returns by providing accurate, interpretable, and robust predictions for film investment decisions.

Model Building for Critics Score Category:

```
Log_production_budget_adj + PG.13 + R + PG + G +
    between_90_to_135 + Greater_than_135 + Spring + Summer + Fall + genre_count +
    Main_Action + Main_Adventure + Main_Animation + Main_Comedy + Main_Crime +
    Main_Documentary + Main_Drama + Main_Family + Main_Fantasy + Main_Horror +
```

Main_Mystery + Main_Romance + Main_Science_Fiction + Main_Thriller + Main_History

Name	Multinomial Logistic Regression	Random Forest	XGBoost	Polynomial Regression	Decision tree
Significant Variables	Log_production_budget_adj, PG.13, PG, R, G, Main_Documentary	Log_production_budget_adj, Main_Action, G, R, PG.13, PG, genre_count, Main_Drama, Fall	Log_production_budget_adj, genre_count, between_90_to_135, Spring, Fall, R, Main_Comedy, Summer, PG.13, Main_Drama	Log_production_budget_adj, PG.13, R, PG, G, between_90_to_135, Main_Action, Main_Adventure, Main_Drama, Main_Fantasy, Main_Science_Fiction	Log_production_budget_adj, genre_count, PG, Main_Drama, Main_Comedy, Fall, R, PG.13
Sensitivity Popular	0.7083	0.6500	0.6000	0.6667	0.5500
Sensitivity Moderate	0.17117	0.21622	0.30631	0.17117	0.17117
Sensitivity Unpopular	0.5680	0.5200	0.5200	0.5680	0.6240
Specificity Popular	0.6695	0.6398	0.6398	0.6822	0.7034
Specificity Moderate	0.90612	0.83265	0.82857	0.88980	0.86122

Specificity Unpopular	0.6537	0.7273	0.7489	0.6364	0.6147
Balanced Accuracy Popular	0.6889	0.6449	0.6199	0.6744	0.6267
Balanced Accuracy Moderate	0.53865	0.52443	0.56744	0.53048	0.51620
Balanced Accuracy Unpopular	0.6108	0.6236	0.6345	0.6022	0.6194
ROC / AUC Popular	0.7382415	0.7070445	0.6732521	0.7346751	0.6490643
ROC / AUC Moderate	0.5842618	0.5480419	0.5807685	0.5812466	0.50467
ROC / AUC Unpopular	0.6526061	0.6922771	0.6639827	0.6522251	0.6117056

The Critic_score variable was categorized into three distinct categories: “Unpopular,” “Moderate,” and “Popular,” based on specific thresholds. This transformation allows for a more straightforward analysis and prediction of how different factors contribute to the likelihood of a film falling into each category. Below is a detailed interpretation of the results, including performance metrics and insights.

Categorization of Critic Scores

The categorization was performed as follows:

- **Unpopular**: Critic scores ≤ 38
- **Moderate**: Critic scores between 38 and 67 (exclusive of both ends)
- **Popular**: Critic scores > 67

For optimizing film investments based on Critic_Score_Category, the best model is XGBoost.

High Accuracy in XGBoost offers the most robust predictive performance compared to other models, ensuring better categorization of films into Popular, Moderate, or Unpopular categories. Class Handling: It excels in managing imbalanced datasets, critical for avoiding missed predictions in key categories like Popular (which are vital for investment decisions).

Feature Interactions: XGBoost's ability to capture complex interactions between features like budget, genre, and seasonal release ensures nuanced predictions that align with film investment goals. Scalability: Its efficient computation allows handling large datasets, making it scalable for future predictions as more data becomes available. XGBoost is the most suitable model for your goal of maximizing returns on film investments.

5. Proof of Concept: Predictive Technique

Methodology Summary for Predictive Technique Development

After selecting the appropriate models for the dependent variables—IMDB_Category, Critic_score_category, and Log_worldwide_gross—the next step was to build a predictive technique that utilizes these models to estimate outcomes for future films. This approach allows film producers to make informed investment decisions based on predicted classifications and expected revenues.

Model Selection and Types

1. **IMDB_Category**
   - **Best Model**: **Random Forest**
   - This model effectively captures the relationships among predictors and provides robust performance in categorizing films into their respective IMDB categories.
2. **Critic_score_category**
   - **Best Model**: **XGBoost**
   - XGBoost excels in managing complex interactions between

features and handling imbalanced datasets, making it suitable for predicting critic score categories.

3. **Log_worldwide_gross**
   - **Best Model**: **XGBoost**
   - This model demonstrates strong performance in predicting worldwide gross revenues due to its ability to capture intricate relationships among features.

Building a Predictive Technique

To create a predictive technique, we utilized the selected models as follows:

- **Worldwide Gross Prediction**: Using the XGBoost model trained on Log_worldwide_gross.
- **IMDB Rating Prediction**: Using the Random Forest model trained on IMDB_Category.
- **Critic Score Prediction**: Using the XGBoost model trained on Critic_score_category.

The predictive technique was tested using future data inputs, simulating scenarios where film producers can control certain variables. By inputting these variables, we generated estimated results for upcoming movies.

Example Predictions

For instance, using input data for various films:

Cruella (2021):

1. **Release Year**: 2021
2. **Budget**: $200,000,000
3. **MPAA Rating**: PG-13
4. **Runtime**: 134 minutes
5. **Month**: May
6. **Season**: Spring
7. **Genres**: Family / Comedy
8. **IMDB Rating**: 7.3
9. **Revenue**: $233,503,234
10. **Critic Score**: 75
11. **Log Gross**: 19.269

Predicited Result:

Critic_score_Prediction: Moderate
Gross_Category_Prediction: High's
IMDB_Category_Prediction: Good

Proof of Concept

(https://docs.google.com/spreadsheets/d/1zi

gUHMPzjvqUY6z_37UxWtngjifVNJhJx32_Ej8cdOk/edit?usp=sharing)

Our predictive technique was validated using data from 61 random movies released between 2019 and 2024. The results demonstrate the potential of our models:

- 20% of movies had all three predictions correct
- 44% had two out of three predictions correct
- 36% had one correct prediction

Individual model performance:

- Gross Category (XGBoost): 71.99% accuracy with 78.69% predictive power
- IMDb Ratings (Random Forest): 58.46% accuracy with 40.98% predictive power
- Critics Scores (XGBoost): 60.73% accuracy with 65.57% predictive power

On average, our models achieved 63.73% accuracy and 61.75% overall predictive power.

6. Recommendations

Based on our analysis, we propose tailored strategies for films with different budget levels:

High-Budget Films ($150-300M)

- Runtime: 90-135 minutes or >135 minutes
- Genres: Action, Adventure, Sci-Fi, Animation, Fantasy
- Release Timing: Summer/Fall (ideal for blockbusters, school vacations, and holiday crowds)
- MPAA Rating: PG / PG-13 (to maximize audience reach across age groups)

Mid-Budget Films ($50-150M)

- Runtime: 90-135 minutes
- Genres: Drama, Historical, Thriller, Biographical, Mystery
- Release Timing: Fall/Winter (coincides with awards season)
- MPAA Rating: PG-13 / R (targeting older, more mature audiences)

Low-Budget Films (<$50M)

- Runtime: <90 minutes or 90-135 minutes
- Genres: Horror, Thriller, Comedy, Drama

- Release Timing: Spring/Fall (less competitive seasons, allowing niche films to thrive)
- MPAA Rating: R (targeting niche audiences expecting more mature or edgy content)

By aligning with these strategies and leveraging insights from our predictive models, producers can significantly enhance their chances of box office success and better allocate resources.

7. Project Limitations and Future Research Opportunities

Data Accessibility Challenges

One of the primary limitations of our current research was the inherent complexity of obtaining comprehensive data in the highly competitive film industry. The available datasets are often limited, curated, or require substantial financial investment. To overcome this constraint, future research would necessitate subscribing to specialized entertainment data providers such as Numbers.com or The Hollywood Reporter. These platforms offer more nuanced and detailed industry insights that could significantly enhance the depth and accuracy of our analysis.

Enhancing Predictive Variables

Talent Quantification

A critical area for improvement involves developing more sophisticated methods to quantify the subjective elements of film production. We propose a comprehensive "star power" evaluation framework that would:

- Analyze each talent's historical performance (directors, producers, screenwriters, cast)
- Develop metrics to measure monetary impact and industry influence
- Create detailed talent portfolios tracking their entire professional trajectory
- Identify optimal talent combinations that consistently produce successful films

This approach would provide a more robust predictive model for film success, moving beyond traditional quantitative metrics to capture the nuanced dynamics of creative collaboration.

Genre Optimization

Our research could be significantly expanded by:

- Conducting in-depth analysis of genre permutations

- Identifying optimal genre combinations that maximize revenue potential
- Developing predictive models that map genre intersections and audience preferences

Award and Recognition Metrics

Incorporating major awards and nominations would offer additional layers of insight:

- Quantifying the impact of critical acclaim on film success
- Mapping talent performance across different genres
- Analyzing how awards correlate with financial performance
- Identifying production companies' strategic talent acquisition patterns

Production Company Strategic Analysis

A promising avenue for future research involves a comprehensive examination of production companies' historical performance. This would include:

- Mapping resource allocation strategies
- Predicting optimal film genres for specific production companies
- Developing predictive models for seasonal film production
- Analyzing the relationship between strategic choices and financial/critical success

Conclusion

These proposed research directions represent significant opportunities to transform our understanding of the film industry. By developing more sophisticated data collection and analysis methodologies, we can create more nuanced predictive models that capture the complex interplay of creative and commercial factors in film production.

Citations:

https://journals.aau.dk/index.php/NJMM/article/view/5871/5505

https://link.springer.com/article/10.1007/s10824-019-09372-1

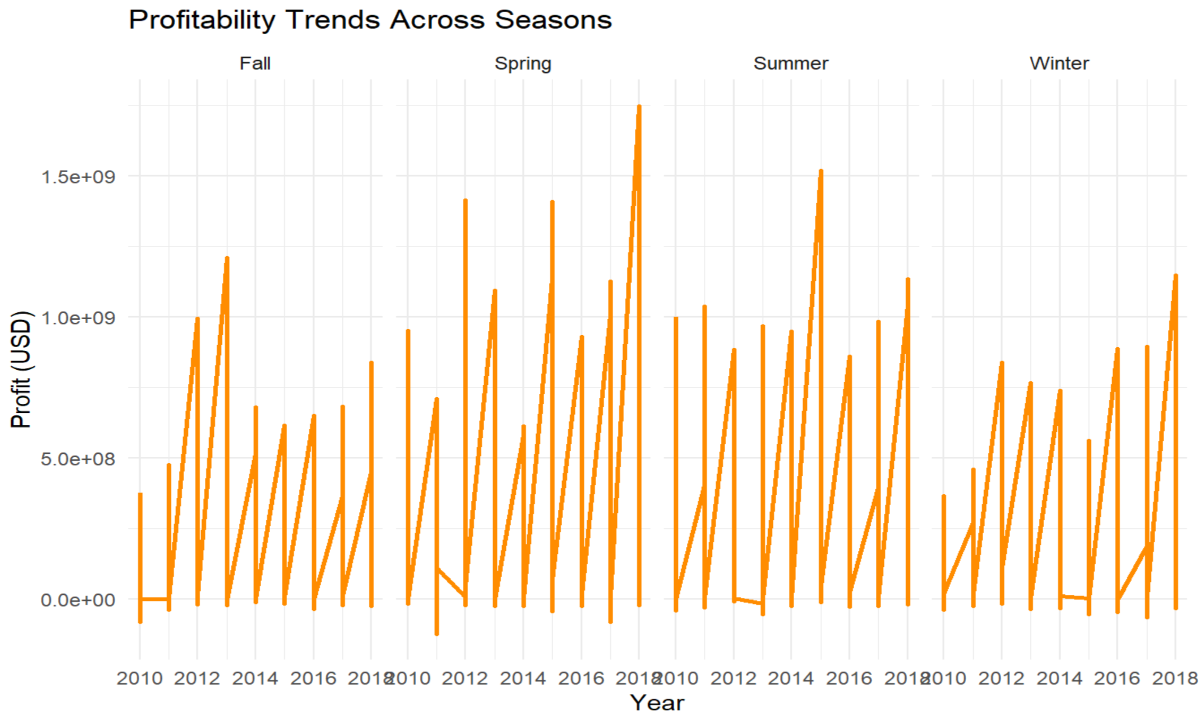
https://urfjournals.org/open-access/predictive-analytics-for-box-office-success.pdf

https://journals.aau.dk/index.php/NJMM/article/view/5871/5505

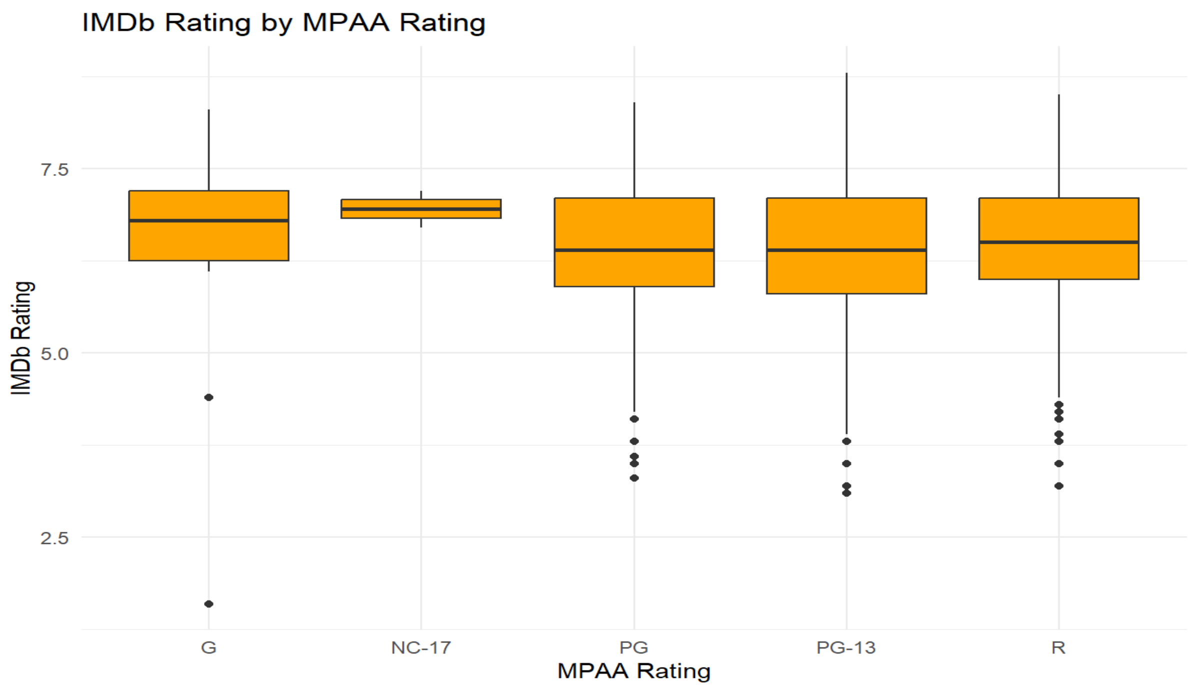
https://ieeexplore.ieee.org/abstract/document/10574956

Appendix:

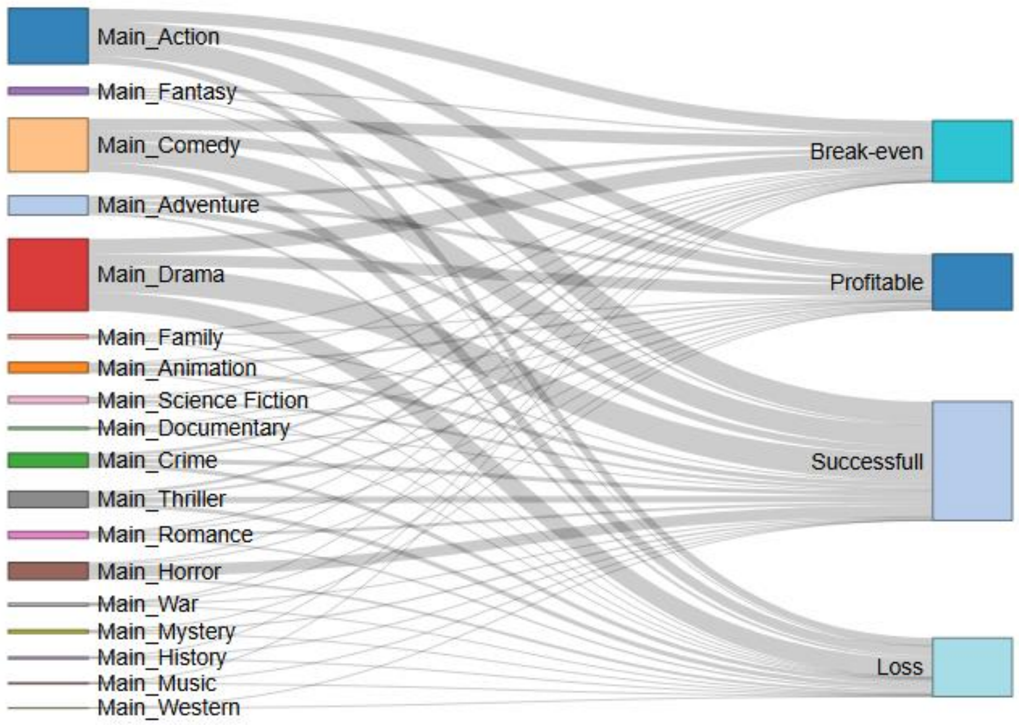
Facet Grid: Profitability Trends Across Seasons: (Fig. 1)



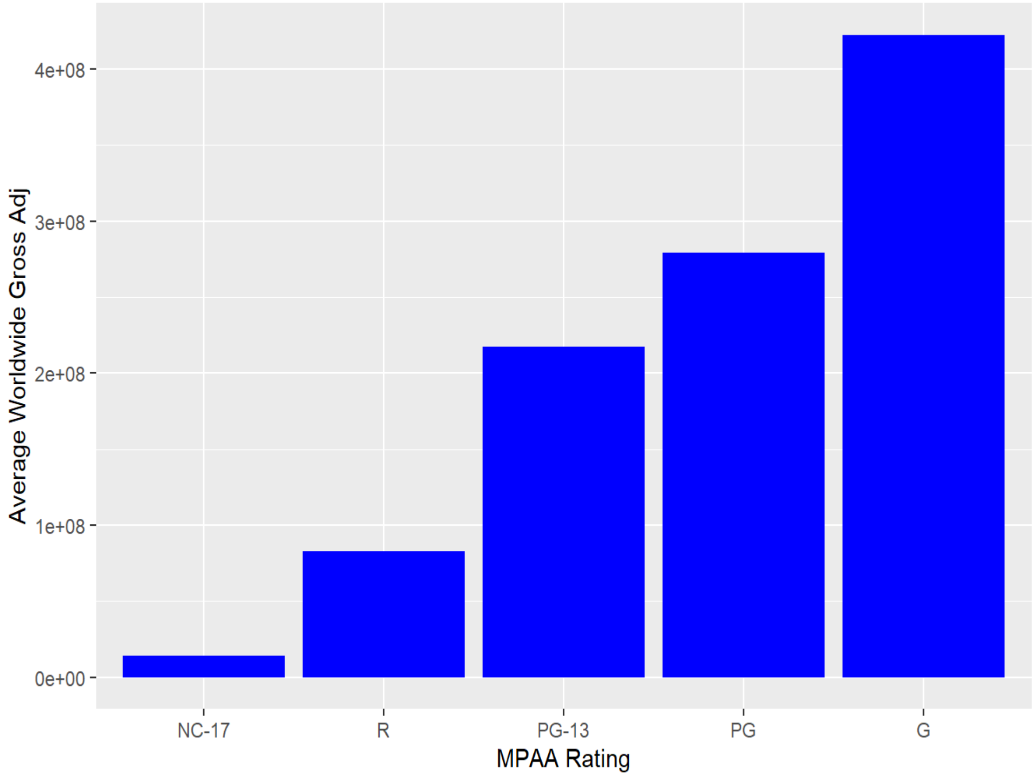
Interpretation of Box Plot: IMDb Rating by MPAA Rating: (Fig. 2)



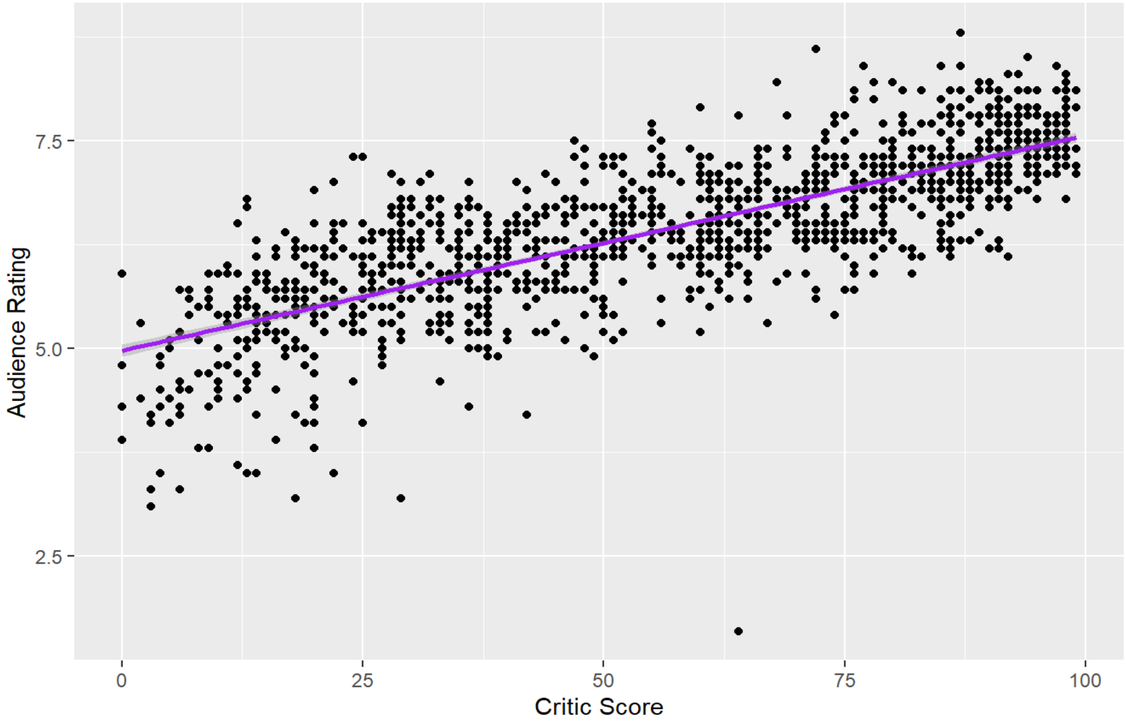
Sankey Diagram: Flow of Movies by Main Genre to Profit Category: (Fig. 3)



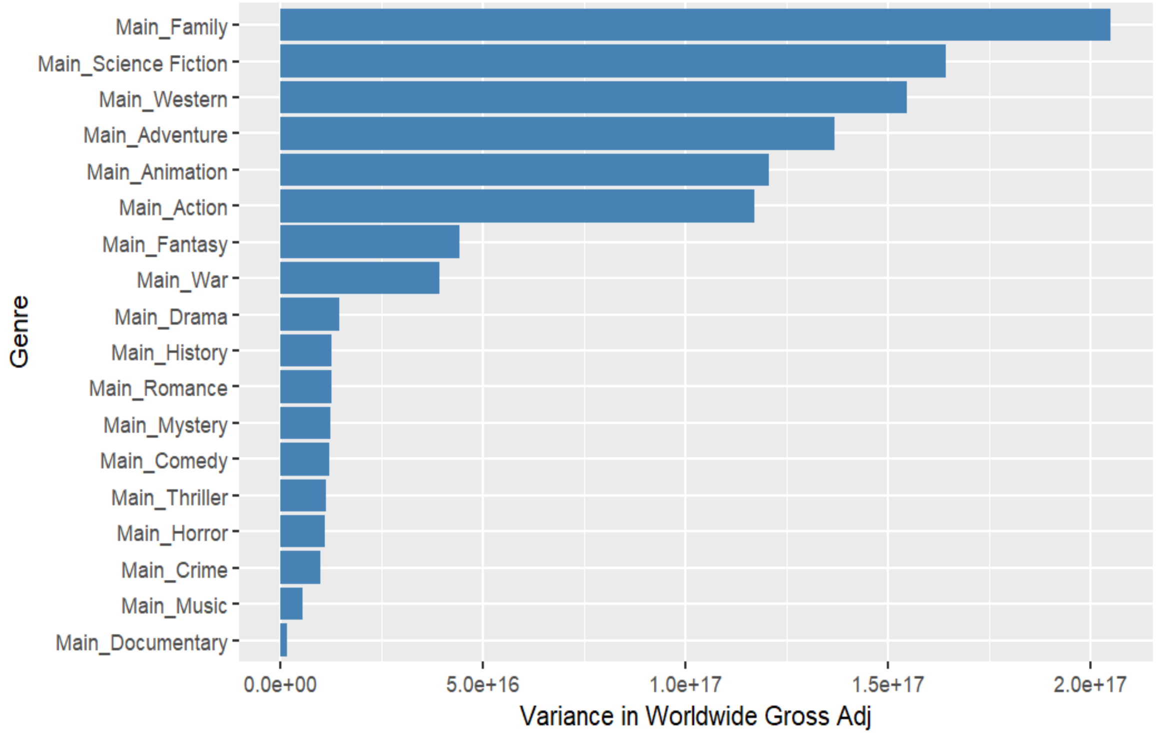
Worldwide Gross Adj by MPAA Rating: (Fig. 4)



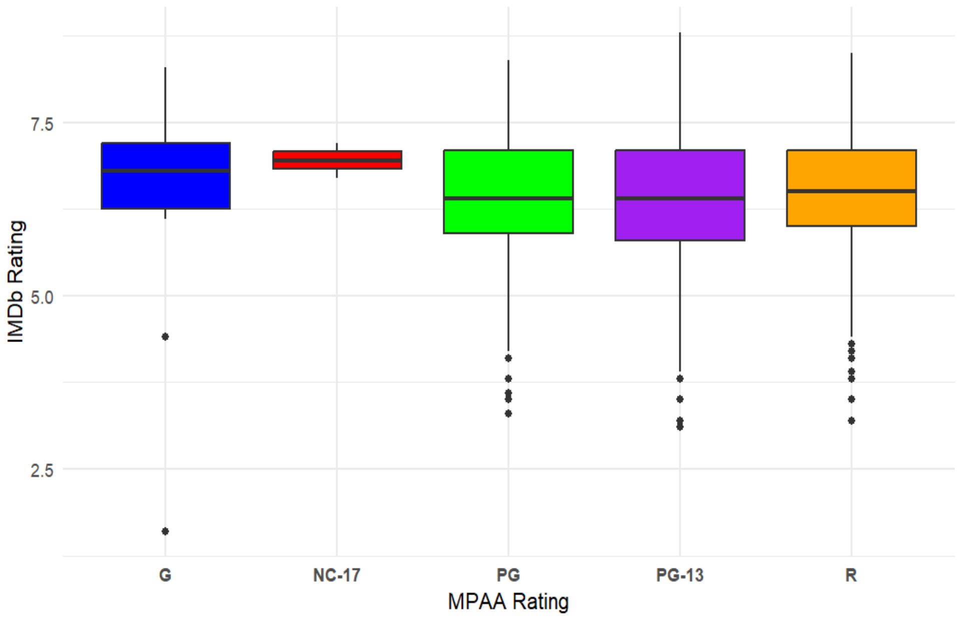
Critic Score vs. Audience Reception: (Fig. 5)



Variance in Worldwide Gross Adjusted by Genre: (Fig. 6)

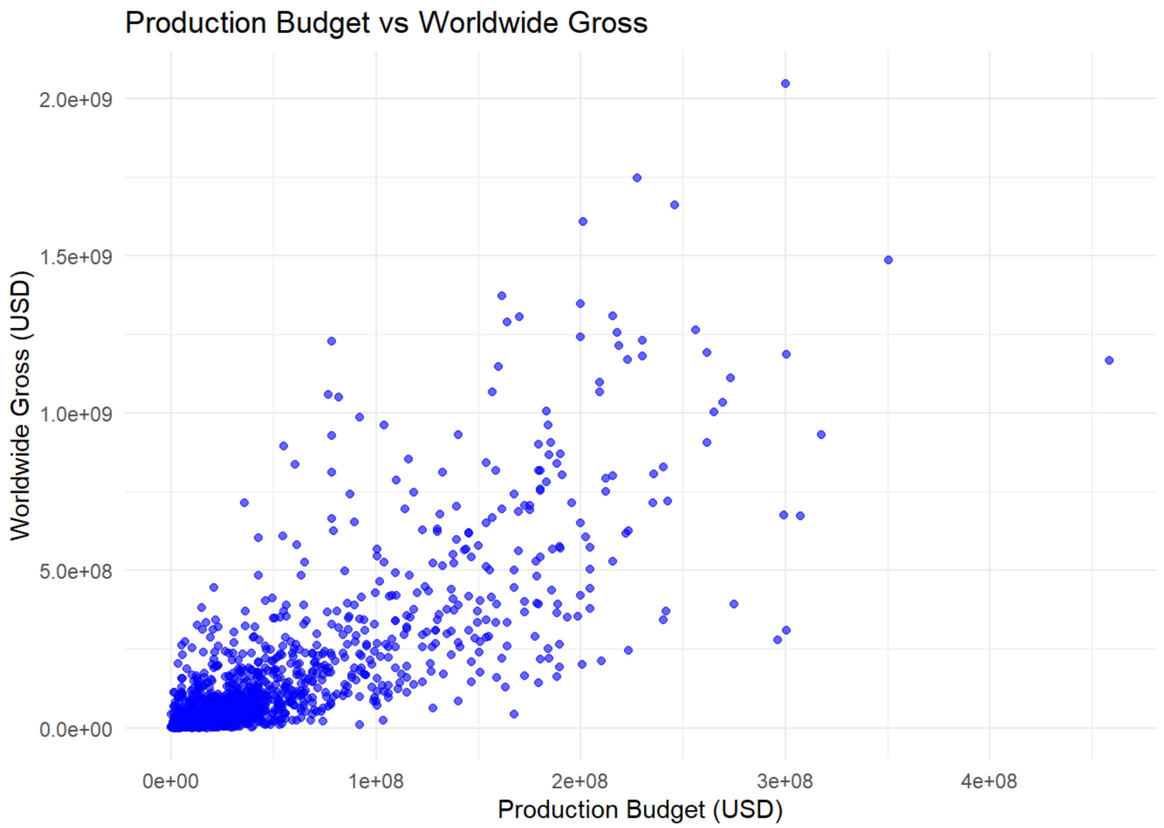


Boxplot of production budgets by genre: (Fig. 7)

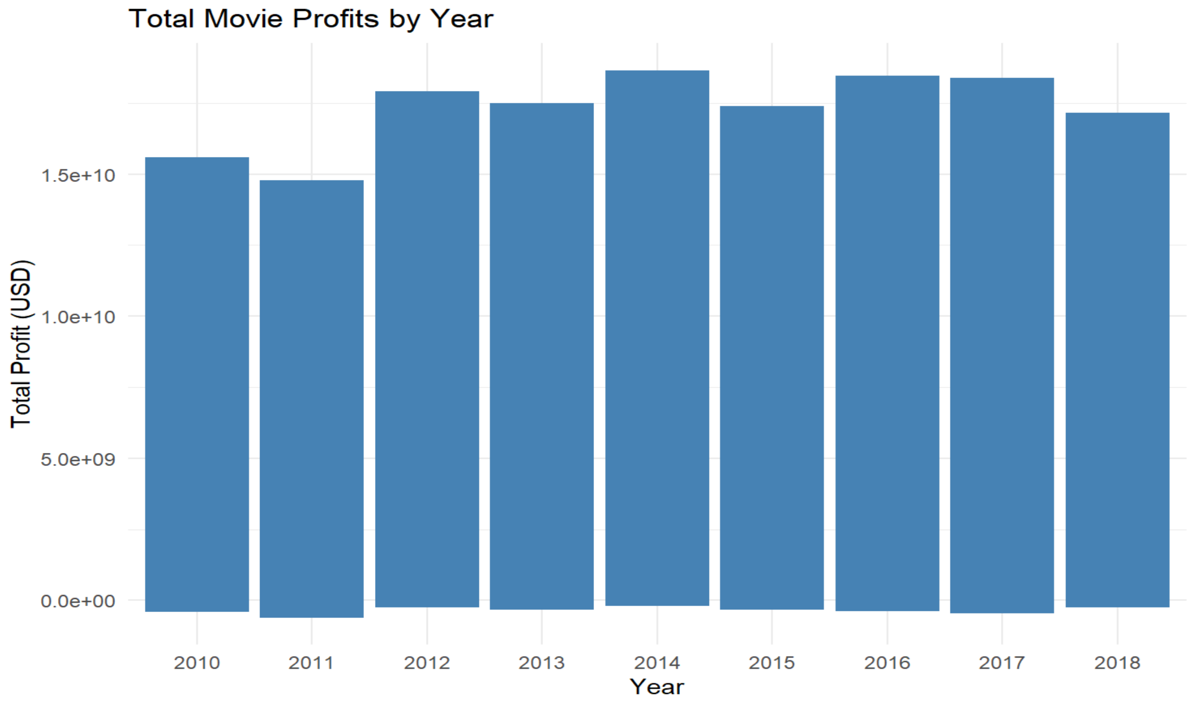


EXTRAS: -

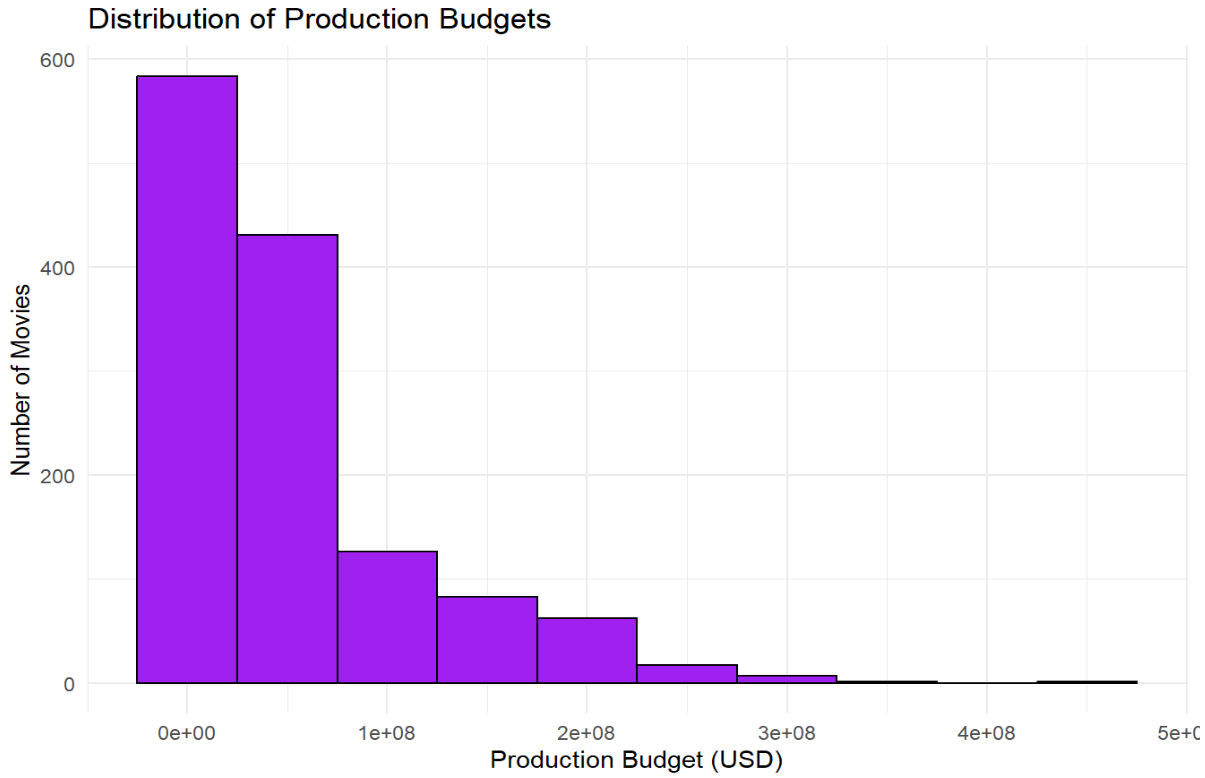
Scatter Plot: Production Budget vs. Worldwide Gross : (Extra 1)



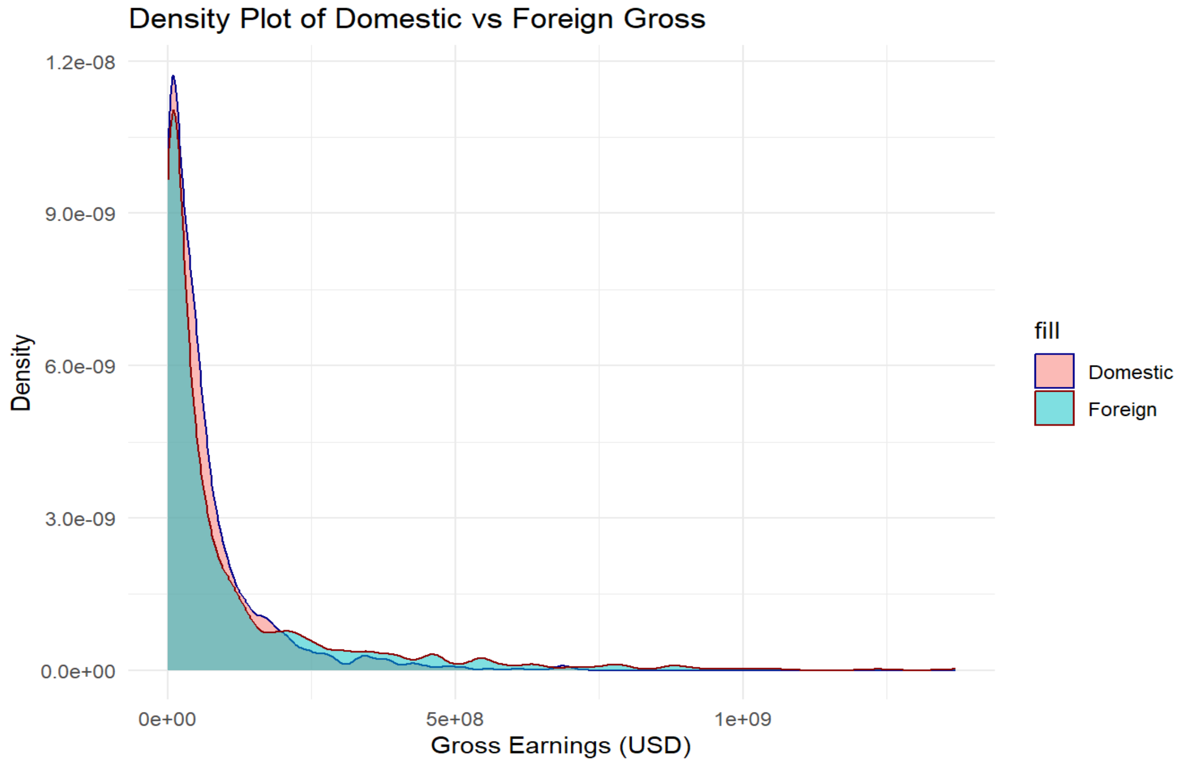
Bar Chart: Total Movie Profits by Year : (Extra 2)



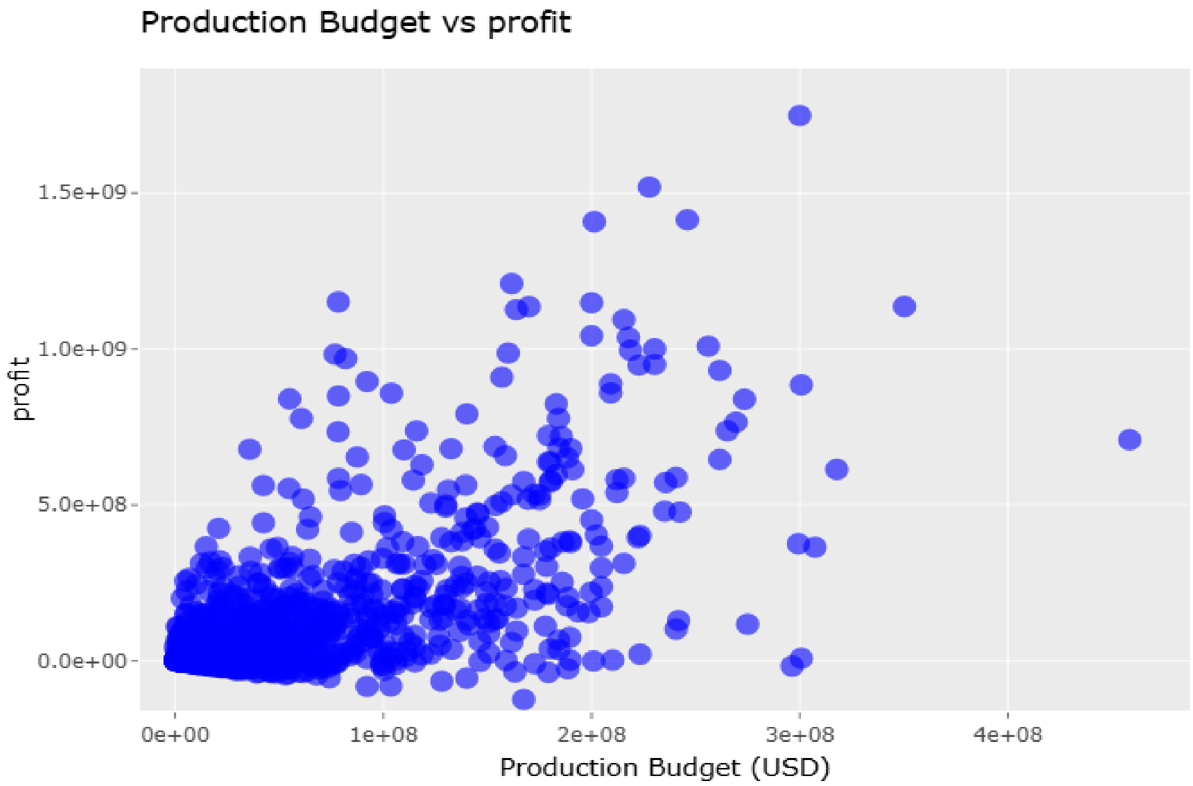
Histogram - Distribution of Production Budgets : (Extra 3)



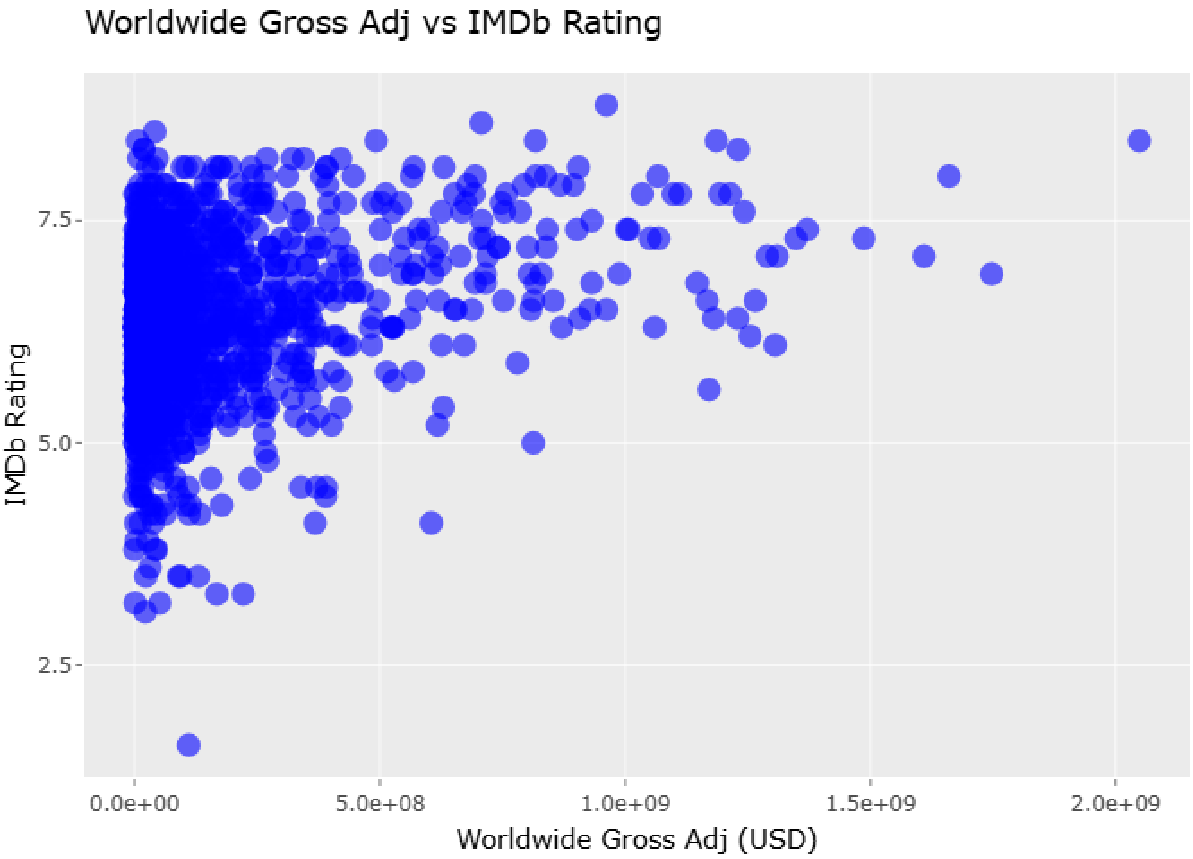
Density Plot: Domestic vs. Foreign Gross : (Extra 4)



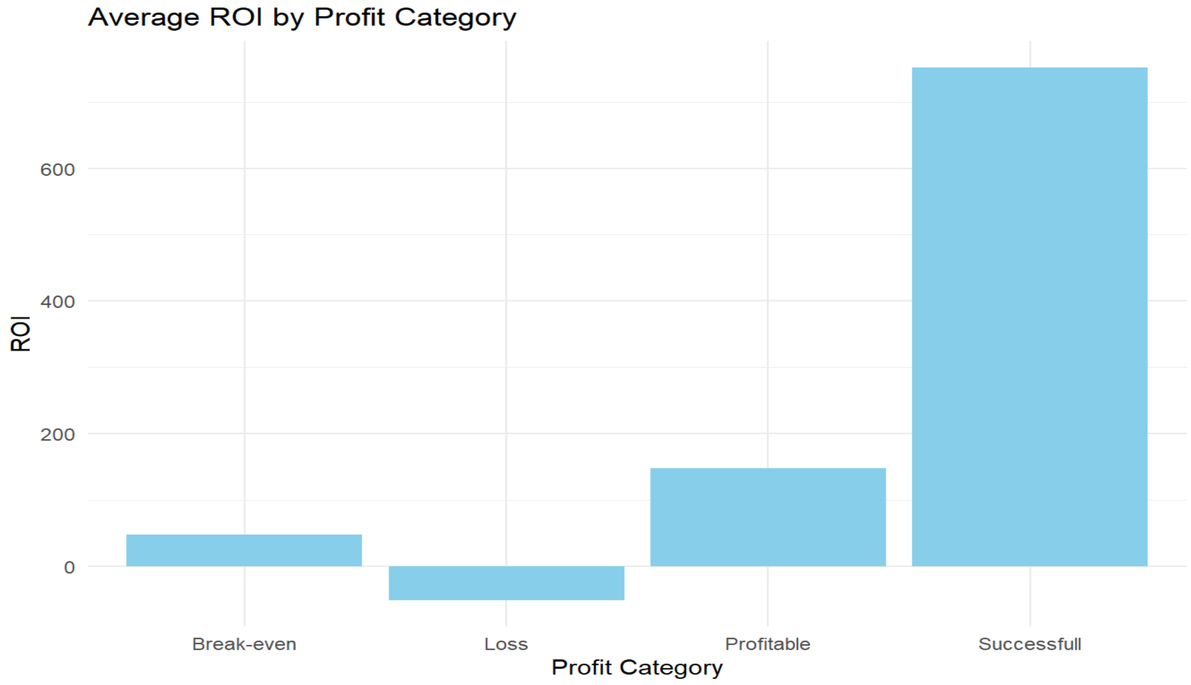
Interpretation of Interactive Plot: Production Budget vs. Profit : (Extra 5)



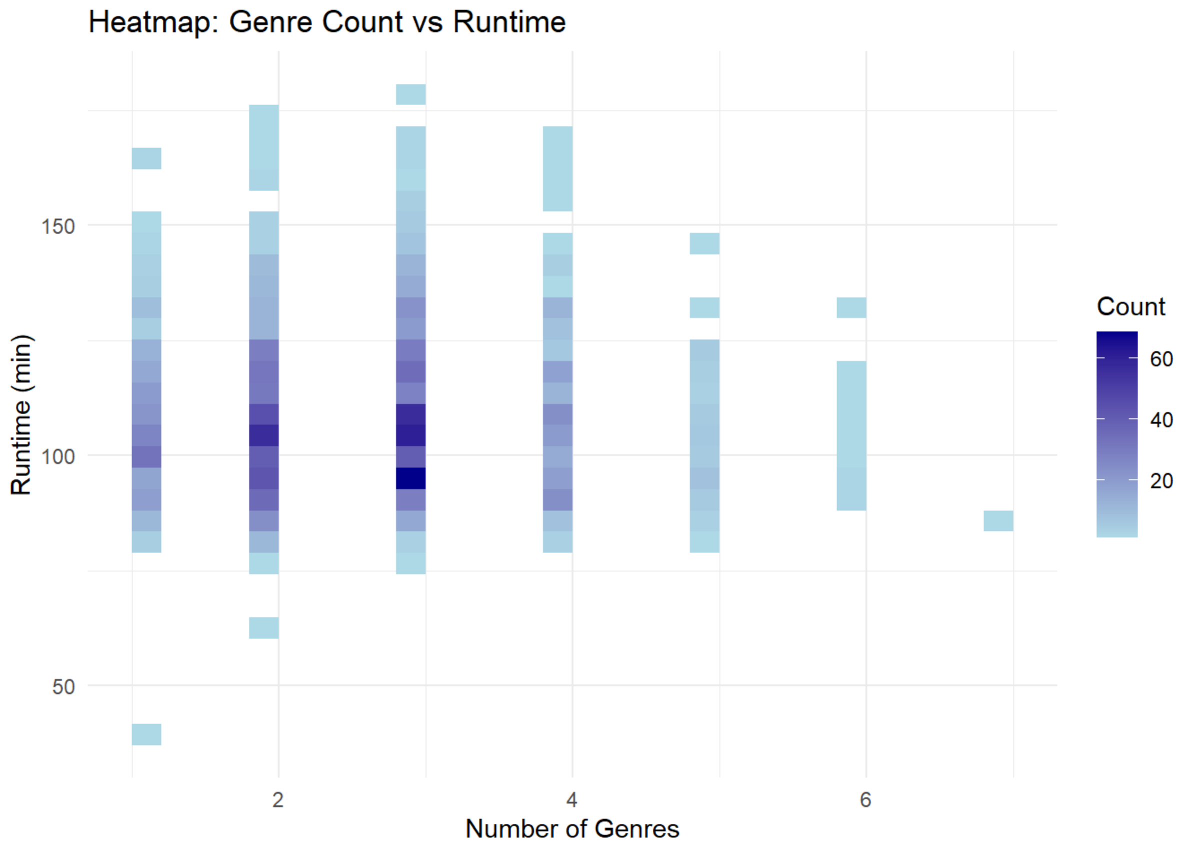
Production Budget vs IMDb Rating : (Extra 6)



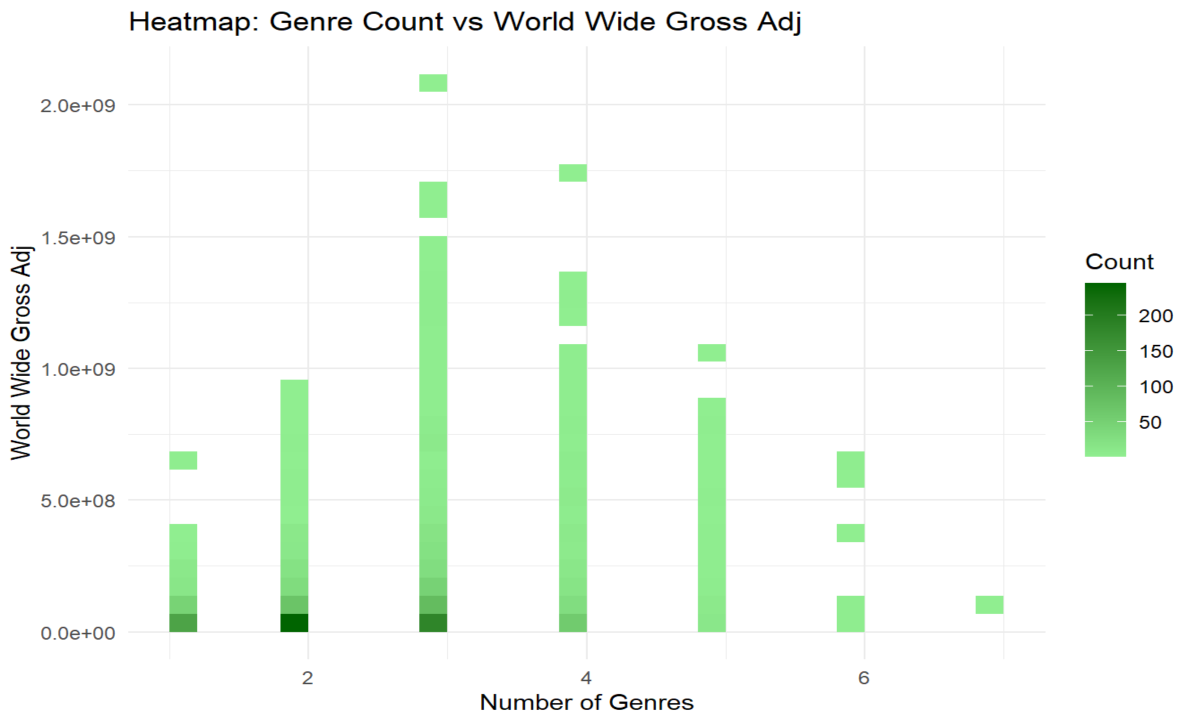
Interpretation of Bar Plot: Average ROI by Profit Category : (Extra 7)



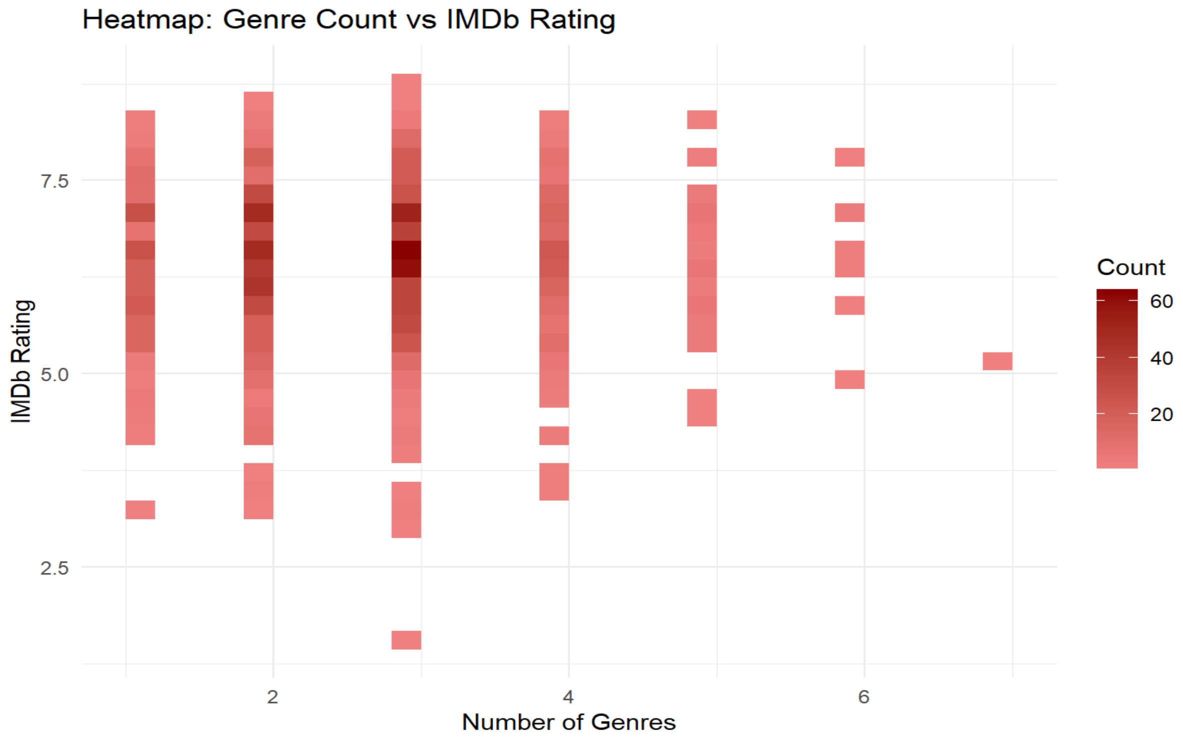
Heatmap: Genre Count vs. Runtime: (Extra 8)



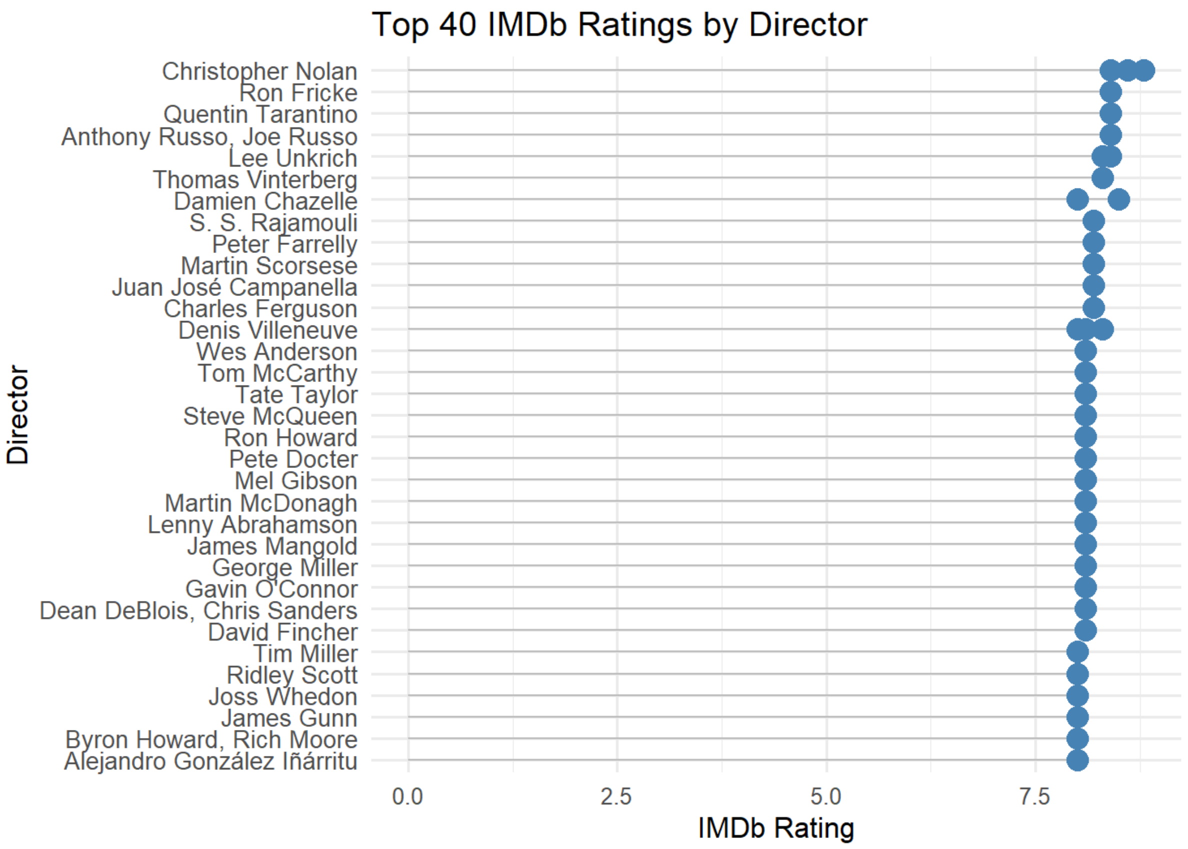
Heatmap: Genre Count vs. Worldwide Gross (Adjusted) : (Extra 9)



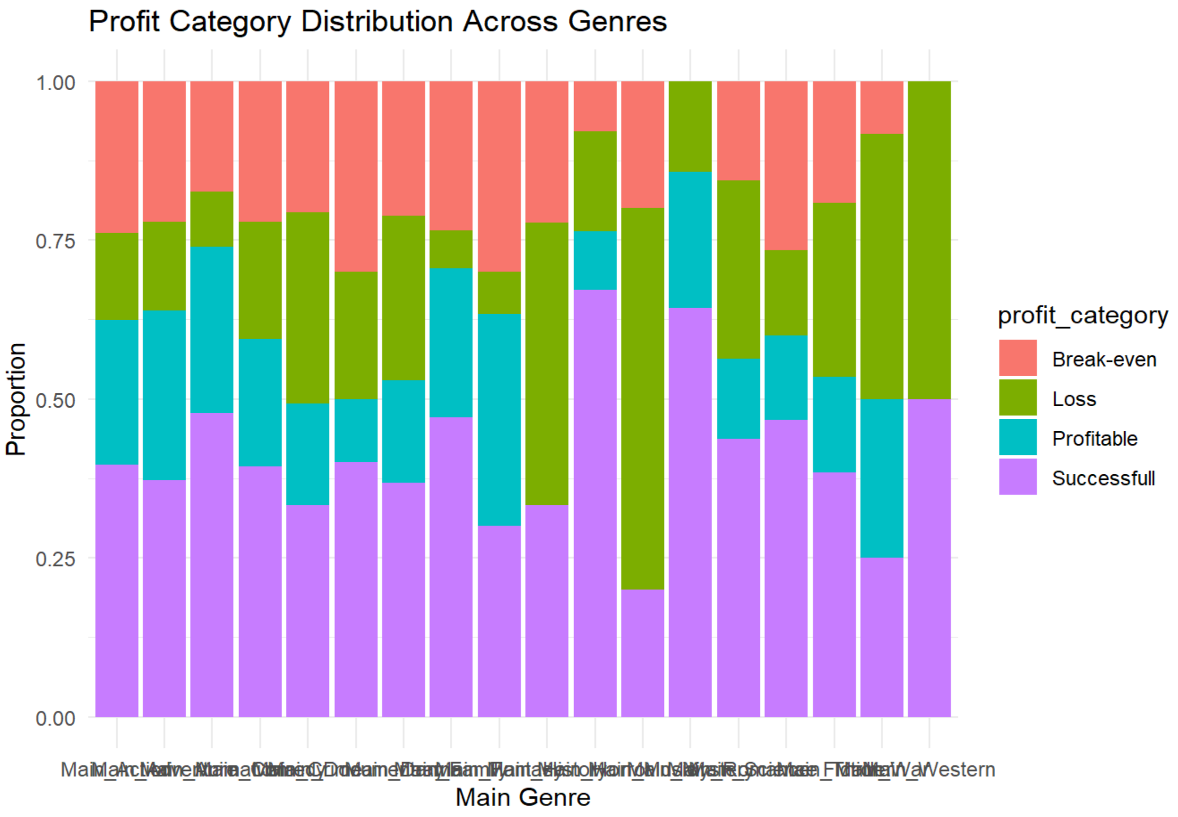
Heatmap: Genre Count vs. IMDb Rating: (Extra 10)



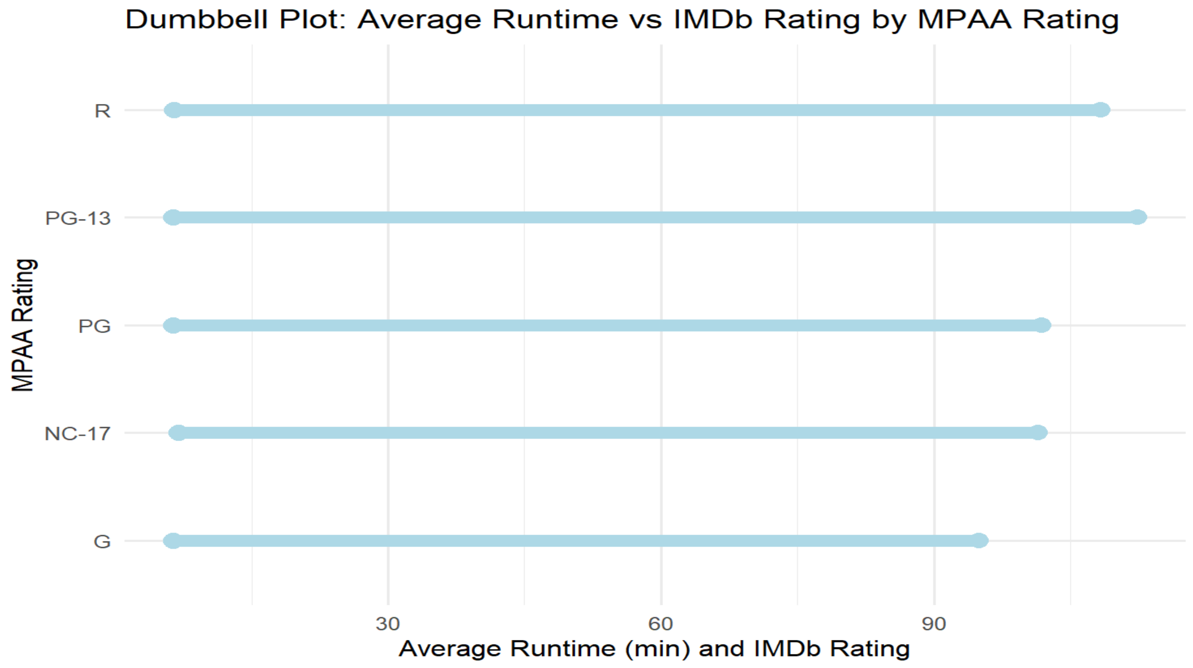
Lollipop Plot: IMDb Rating by Director (Top 40): (Extra 11)



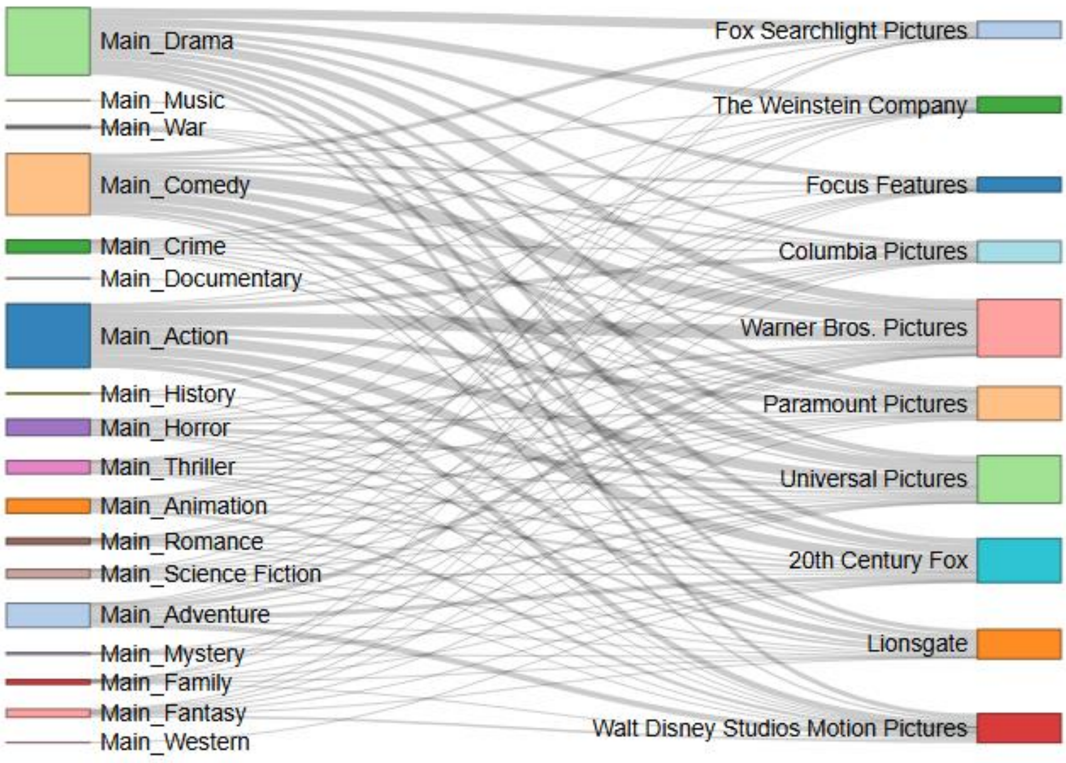
Stacked Bar Plot: Profit Category by Main Genre: (Extra 12)



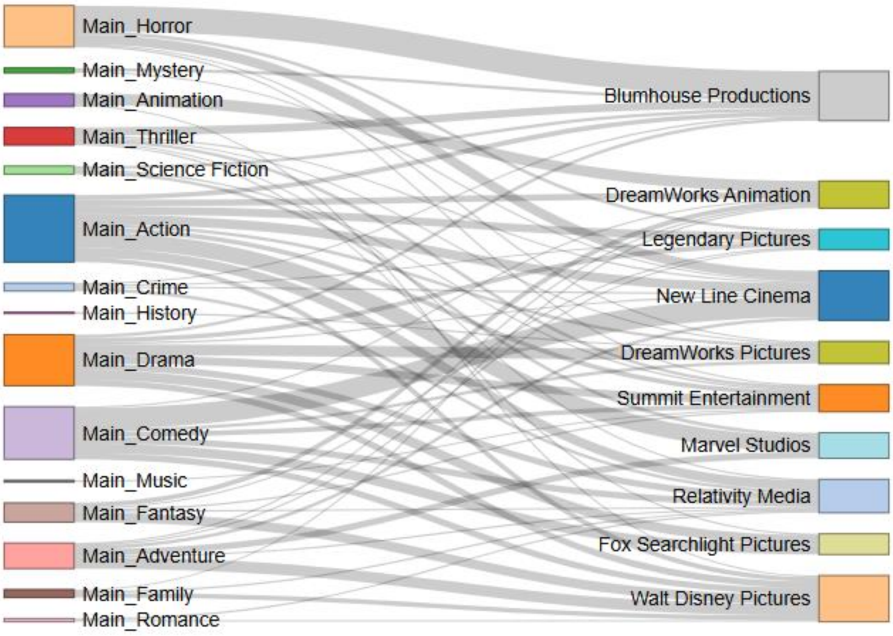
Dumbbell Plot: Runtime vs IMDb Rating by MPAA Rating: (Extra 13)



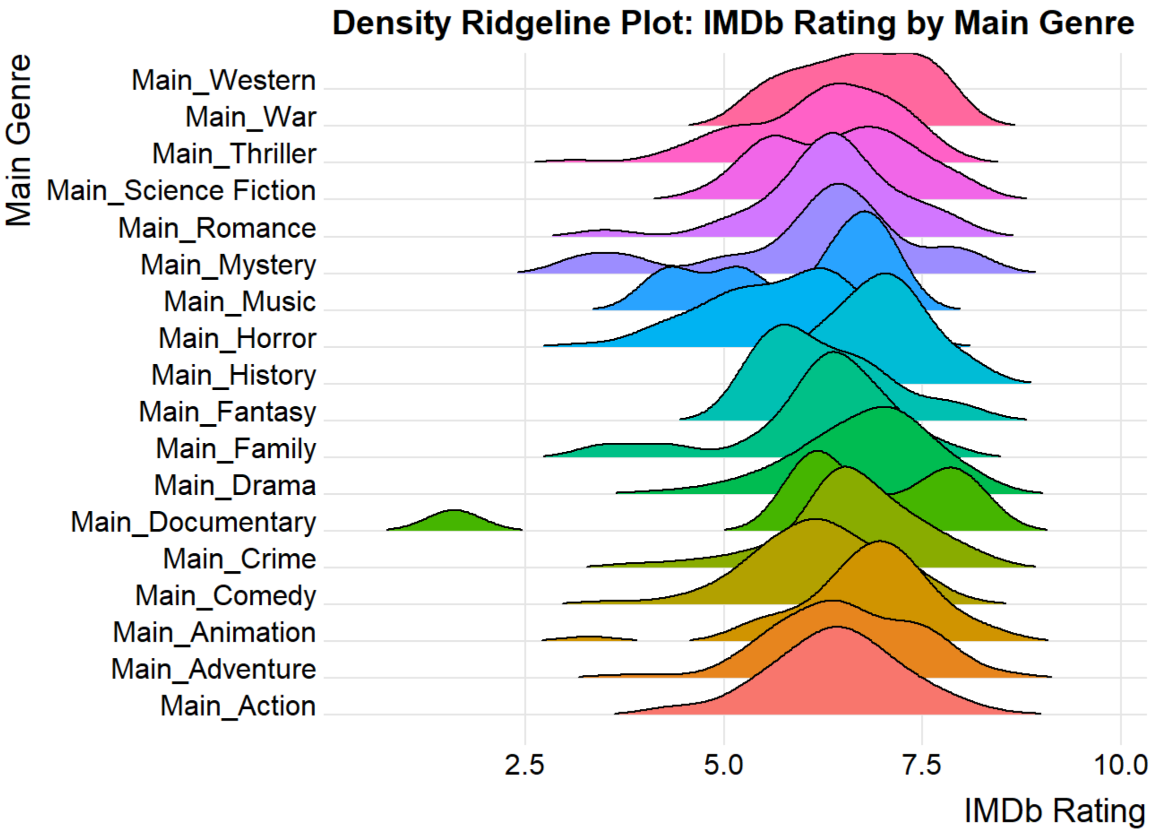
Sankey Diagram: Flow of Movies by Main Genre to Distributor: (Extra 14)



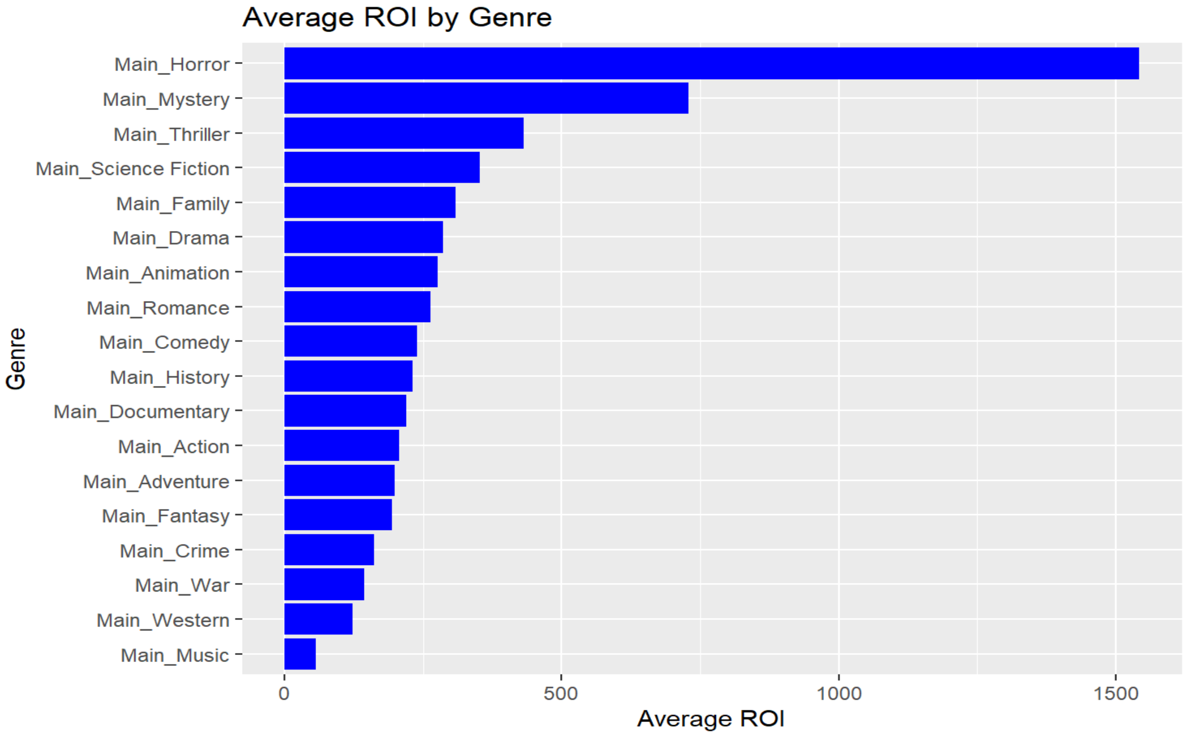
Sankey Diagram: Flow of Movies by Main Genre to Production Company: (Extra 15)



Density Ridgeline Plot: (Extra 16)



Average ROI (Return on Investment) by Genre: (Extra 17)


Average ROI by Genre
Main_Horror
Main_Mystery
Main_Thriller
Main_Science Fiction
Main_Family
Main_Drama
Main_Animation
Main_Romance
Main_Comedy
Main_History
Main_Documentary
Main_Action
Main_Adventure
Main_Fantasy
Main_Crime
Main_War
Main_Western
Main_Music
Genre
Average ROI