Movie Profitability Prediction

Fatemeh Chajaei

1. Introduction

The Movie Profitability Prediction project aims to analyze a dataset containing information about various movies to predict their profitability. The dataset, sourced from TMDB (The Movie Database) via Kaggle, includes a wide range of features such as budget, revenue, genres, keywords, cast, crew, production companies, and more. The primary objective is to develop machine learning models that can predict whether a movie will be profitable or not based on its features.

2. Data Overview

The Movies Dataset and Credits Dataset provide comprehensive information about various movies, including their budget, genres, production details, release date, revenue, runtime, and credits. Below is an overview of the datasets:

Table 1. Data Description for Movie Profitability

Dataset	Attribute	Description	Data Type
Movies	budget	The budget allocated for movie production	int64
	genres	The genres of the movie (e.g., action, comedy, drama)	object
	homepage	The URL of the movie's homepage	object
	id	Unique identifier for the movie	int64
	keywords	Keywords or tags related to the movie	object
	original_language	The original language of the movie	object
	original_title	The original title of the movie	object
	overview	A brief description or summary of the movie	object

	popularity	Numeric value specifying the movie's popularity	float64
	production_companies	Companies involved in making the movie	object
	production_countries	Countries where the movie was produced	object
	release_date	The release date of the movie	object
	revenue	Worldwide revenue generated by the movie	int64
	runtime	Duration of the movie in minutes	float64
	spoken_languages	Languages spoken in the movie	object
	status	The status of the movie (e.g., Released, Rumored)	object
	tagline	A tagline or slogan associated with the movie	object
	title	The title of the movie	object
	vote_average	The average rating given to the movie	float64
	vote_count	The number of votes cast for the movie	int64
Credits	movie_id	Unique identifier for the movie.	int64
	title	The title of the movie.	object
	cast	Information about the cast members (e.g., actors, actresses) involved in the movie	object
	crew	Information about the crew members (e.g., directors, producers, writers) involved in the movie	object

These datasets provide a wealth of information for analyzing movie characteristics, production details, and the people associated with each film. By combining information from both datasets, comprehensive analyses can be conducted to understand the factors contributing to movie success and popularity.

3. Data Cleaning and Preprocessing

After obtaining the raw datasets, several data cleaning and preprocessing steps were performed to ensure the data's quality and suitability for analysis.

This involved handling columns in JSON format (such as genres, keywords, production countries, production companies, spoken languages, cast, and crew) by parsing them to extract relevant information and converting them into Python objects for ease of manipulation. Additionally, the release_date column, containing dates in string format, was converted to the appropriate datetime format to facilitate date-based operations and analysis. After converting the release_date column, further preprocessing steps were conducted to extract the month and year components. This extraction enabled easier analysis based on the release date of the movies, allowing for temporal insights and trends to be captured more effectively. Columns deemed irrelevant or redundant for the analysis were dropped from the datasets to reduce noise and focus on the most relevant features. Missing values in the datasets were addressed through various strategies, including imputation, dropping rows or columns with missing values, or utilizing domain knowledge to infer appropriate values.

By performing these data cleaning and preprocessing steps, the datasets were prepared for exploratory data analysis (EDA) and subsequent modeling tasks. The cleaned data ensures accurate and reliable results in the analysis and modeling phases of the project.

4. Exploratory Data Analysis (EDA)

The exploratory data analysis (EDA) phase of the project involved comprehensive analysis of the movie dataset to gain insights into various aspects such as budget, revenue, popularity, genres, production companies, and more. The EDA process included the following key visualizations and analyses:

Correlation Matrix

The correlation matrix was computed for numerical columns to identify relationships between variables. The heatmap visualization revealed the strength and direction of correlations among features.

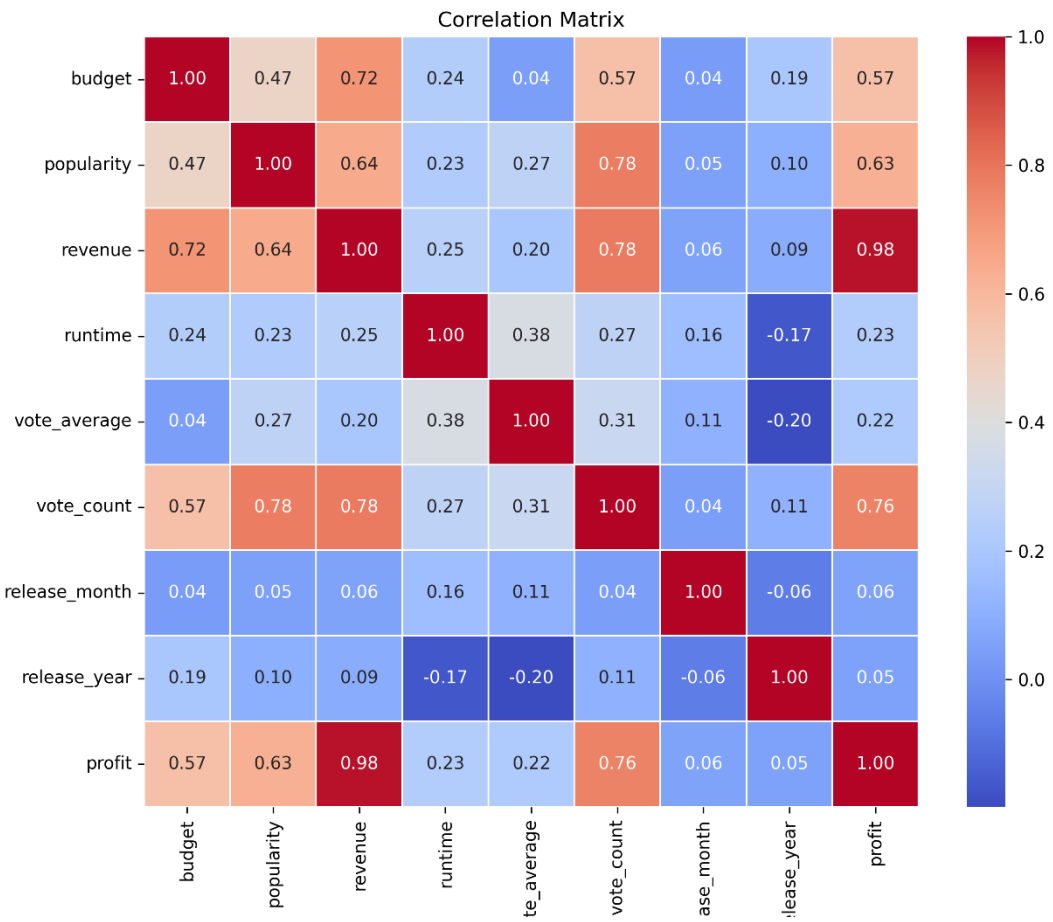

Fig 1. Correlations Between Various Attributes

Each cell in the matrix represents the correlation coefficient between two features, ranging from -1 to 1. A correlation coefficient closer to 1 indicates a strong positive correlation, while a coefficient closer to -1 indicates a strong negative correlation. Here's the description of the correlation matrix:

Budget vs. Other Features: The budget shows strong positive correlations with revenue (0.72) and vote count (0.57), indicating that higher-budget movies tend to generate more revenue and receive more votes. There's also a moderate positive correlation with popularity (0.47) and profit (0.57), suggesting that higher-budget movies are more likely to be popular and profitable.

Popularity vs. Other Features: Popularity exhibits strong positive correlations with revenue (0.64) and vote count (0.78), indicating that more popular movies tend to generate higher revenue and receive more votes. There's also a moderate positive correlation with profit (0.63), suggesting that popular movies are more likely to be profitable.

Revenue vs. Other Features: Revenue shows a very strong positive correlation with budget (0.72) and vote count (0.78), indicating that higher-budget movies tend to generate more revenue and receive more votes. It also has a very strong positive correlation with profit (0.98), implying that movies with higher revenue are highly likely to be profitable.

Runtime vs. Other Features: Runtime exhibits weak correlations with other features, suggesting minimal association with budget, popularity, revenue, and profit.

Vote Average vs. Other Features: Vote average has weak correlations with other features, indicating minimal association with budget, popularity, revenue, and profit.

Vote Count vs. Other Features: Vote count displays strong positive correlations with popularity (0.78) and revenue (0.78), indicating that movies with more votes tend to be more popular and generate higher revenue. It also has a strong positive correlation with profit (0.76), suggesting that movies with more votes are more likely to be profitable.

Release Month and Year vs. Other Features: Release month and year show weak correlations with other features, indicating minimal associations with budget, popularity, revenue, and profit.

Profit vs. Other Features: Profit exhibits strong positive correlations with popularity (0.62), revenue (0.98), and vote count (0.76), suggesting that profitable movies tend to be more popular, generate higher revenue, and receive more votes. It also has a moderate positive correlation with budget (0.57), indicating that movies with higher budgets are more likely to be profitable.

Top 10 Most Expensive Movies

A bar plot was created to visualize the top 10 most expensive movies based on their budget. Each bar represented a movie title, and the corresponding budget was displayed inside the bars for easy comparison.

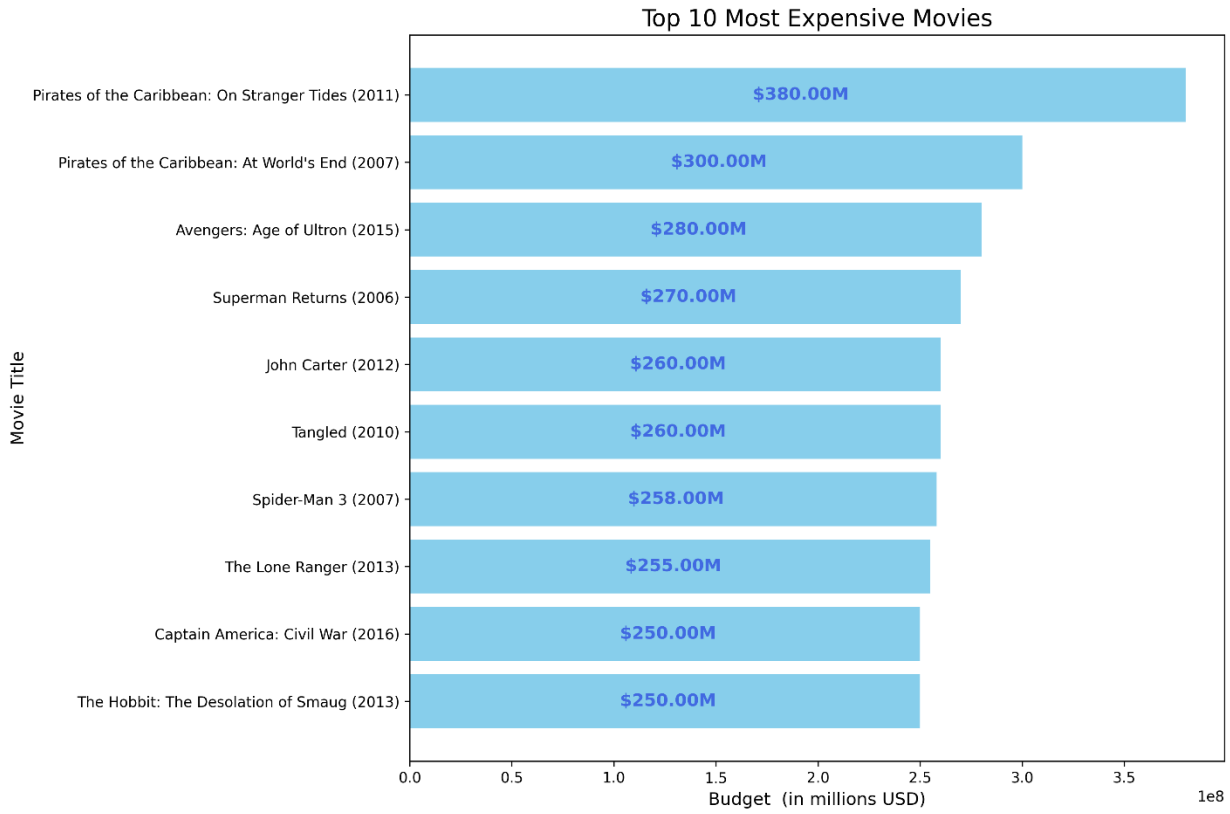

Fig 2. Top 10 Movies with Highest Budget

It appears that the "Pirates of the Caribbean" series has the highest budget among the top 10 most expensive movies, followed by the "Avengers" series. This suggests that these blockbuster franchises invested substantial amounts of money in their production, possibly due to the extensive visual effects, star-studded casts, and elaborate sets required for such high-profile productions. The significant budget allocation underscores the studio's confidence in the potential commercial success of these movies and their commitment to delivering a high-quality cinematic experience to audiences.

Top 10 Highest Grossing Movies

Another bar plot showcased the top 10 highest grossing movies, highlighting their revenue. Similar to the previous visualization, the revenue values were displayed inside the bars for clarity.

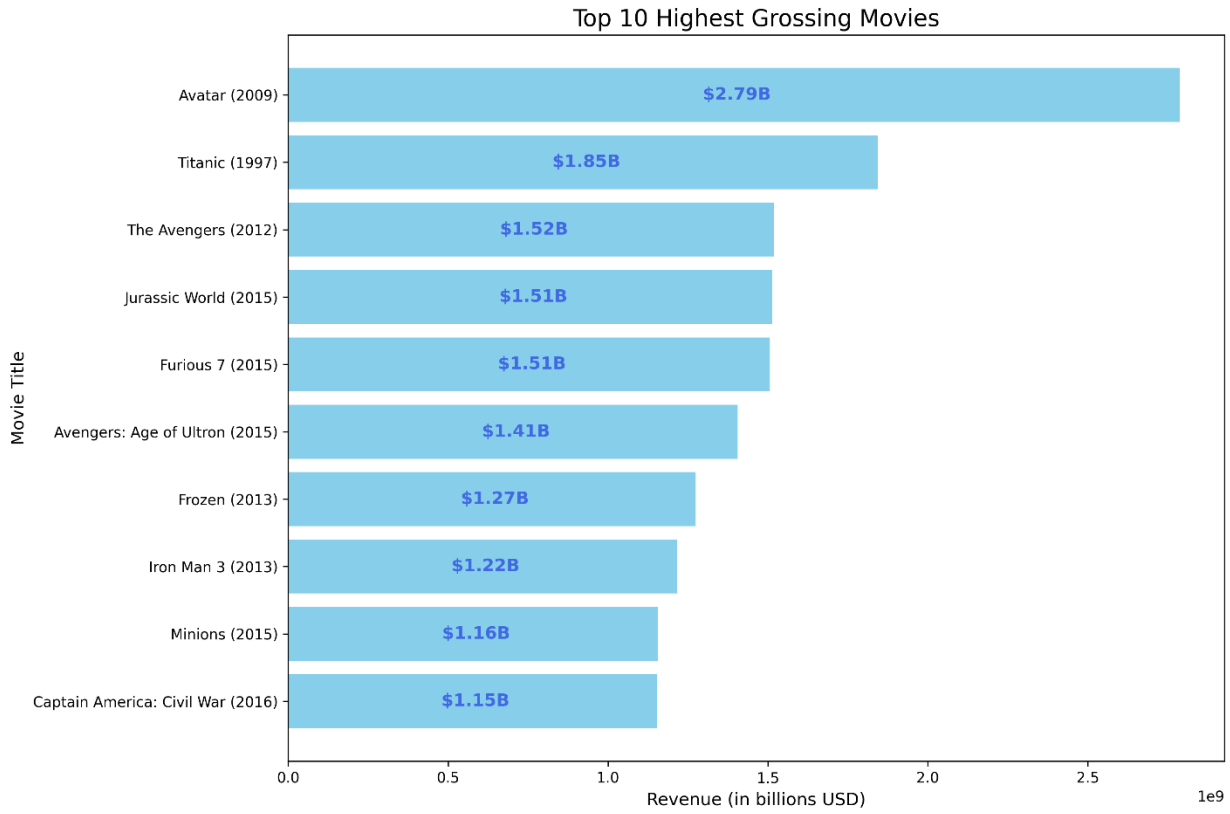

Fig 3. Top 10 Movies with Highest Revenue

The top 10 highest-grossing movies showcase some of the most iconic and commercially successful films in cinematic history. At the pinnacle sits "Avatar," directed by James Cameron, with a staggering revenue of over 2.7 billion USD, making it the highest-grossing movie of all time. Following behind is another Cameron masterpiece, "Titanic," which garnered approximately 1.85 billion USD in revenue. Interestingly, both "Avatar" and "Titanic" are directed by James Cameron, indicating his remarkable ability to craft blockbuster hits that resonate with audiences worldwide. However, there is a substantial revenue difference between "Titanic" and "Avatar." As we move down the list, the revenue gaps diminish, with movies like "Furious 7," "Jurassic World," and "The Avengers" collecting almost similar amounts of revenue.

The genres of the top grossing movies were analyzed to identify commonalities. The most common genres among the top grossing movies were determined and presented. From the analysis, it's evident that adventure, action, and science fiction are the most common genres among the top grossing movies, with adventure being the most prevalent, followed by action and science fiction. This suggests that movies with elements of adventure, action, and fantasy tend to attract larger audiences and generate higher revenues, possibly due to their widespread appeal and ability to offer immersive cinematic experiences.

Top 10 Movies with Highest Popularity

A bar plot displayed the top 10 movies with the highest popularity ratings, showcasing their popularity scores.

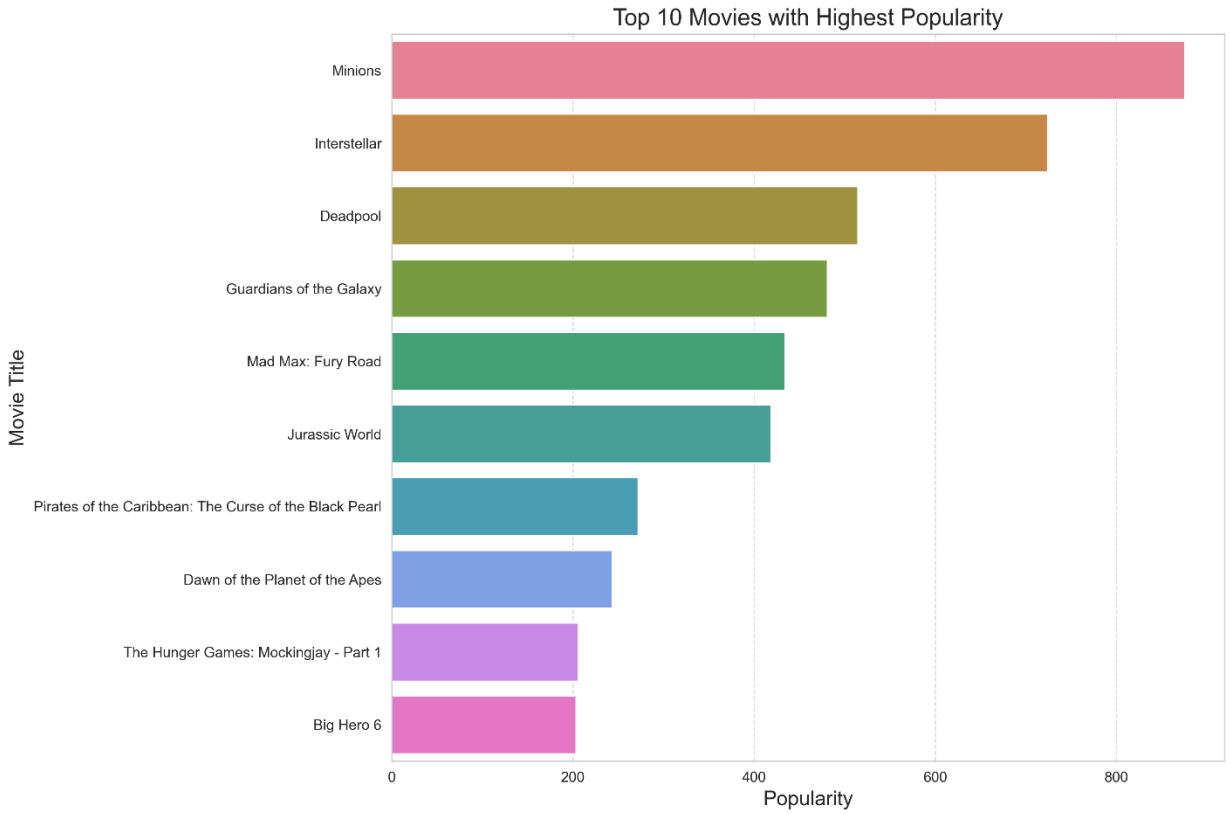

Fig 4. Top 10 Movies with Highest Popularity

These movies have garnered significant attention and engagement from audiences, leading to high popularity scores on TMDB. It's worth noting that popularity scores are influenced by various factors such as the number of votes, views, user favorites, and watchlist additions. Therefore, movies with compelling narratives, impressive visual effects, memorable characters, and engaging marketing campaigns tend to rank higher in popularity. In this list, animated films like "Minions" and "Big Hero 6" are particularly prominent, suggesting a strong appeal to audiences across different age groups. Similarly, blockbuster franchises such as "Guardians of the Galaxy," "Jurassic World," and "Pirates of the Caribbean" also feature prominently, highlighting the popularity of established cinematic universes and beloved characters.

Rating Distribution

A count plot depicted the distribution of movie ratings grouped into bins, enabling an understanding of the distribution of ratings across different groups.

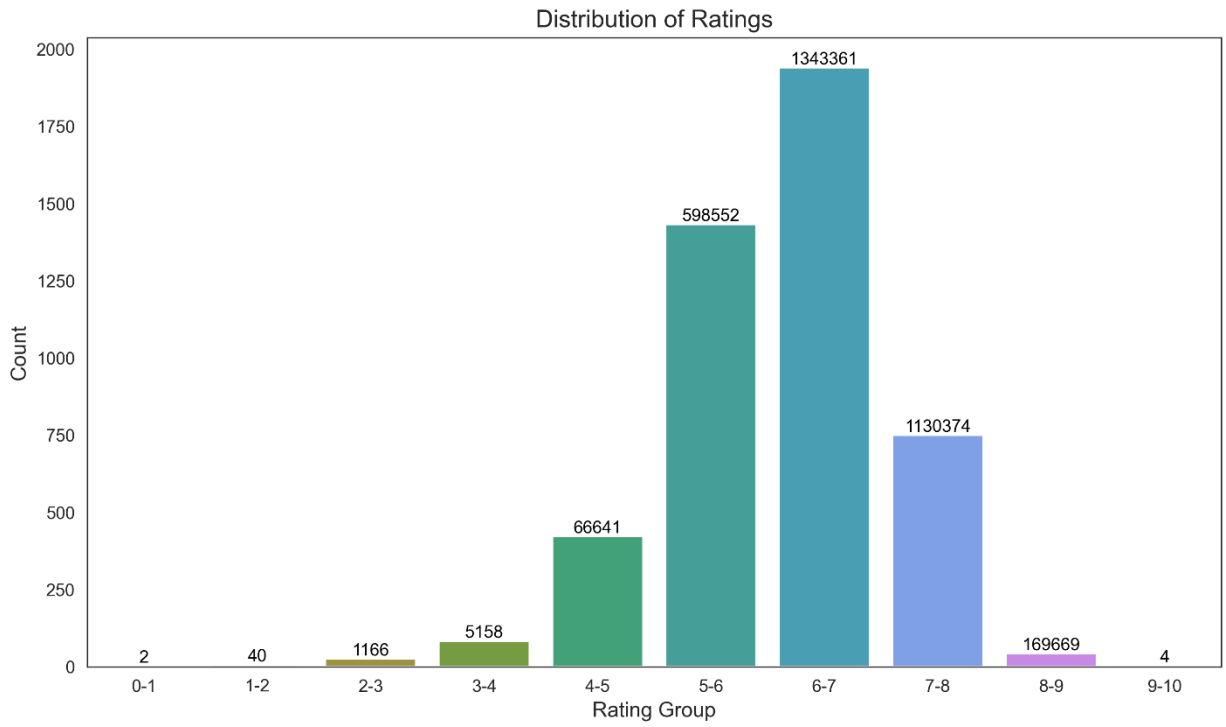

Fig 5. Distribution of Ratings (Grouped)

The rating distribution plot provides insight into the distribution of movie ratings within the dataset. The bars in the plot represent the count of movies with that rating group rather than the number of people who voted within each rating group. However, the labels on each bar indicate the number of people who voted. This means that each bar represents the number of movies falling within a specific rating range, and the label on each bar denotes the total number of votes received by all movies within that range. This distinction is important as it provides insight into both the distribution of movies across different rating ranges and the level of engagement or popularity indicated by the number of votes.

Upon analysis, it's evident that the majority of movies fall within the rating range of 6 to 7, followed by the range of 5 to 6, and then 7 to 8. This suggests that a significant portion of the movies in the dataset are rated moderately, with fewer extremes at the lower and higher ends of the rating spectrum. Interestingly, the distribution follows a roughly normal distribution, albeit slightly skewed to the right.

It's worth mentioning that some of the average ratings in the original dataset are averaged from only a few ratings. This poses a potential issue as movies with a small number of ratings may have their averages skewed, impacting any analyses related to ratings. For instance, the group representing ratings of 9-10 contains three movies, but the total number of people who voted is only four. This scenario illustrates how a small number of high ratings can significantly influence the average rating of a movie, potentially misleading interpretations of its overall quality. Hence, filtering out movies with only a few ratings would be advisable to ensure more accurate analyses.

Popularity vs. Rating

Scatter plots were used to visualize the relationships between variables such as popularity, vote count, and vote average. These plots helped identify potential trends and patterns in the data.

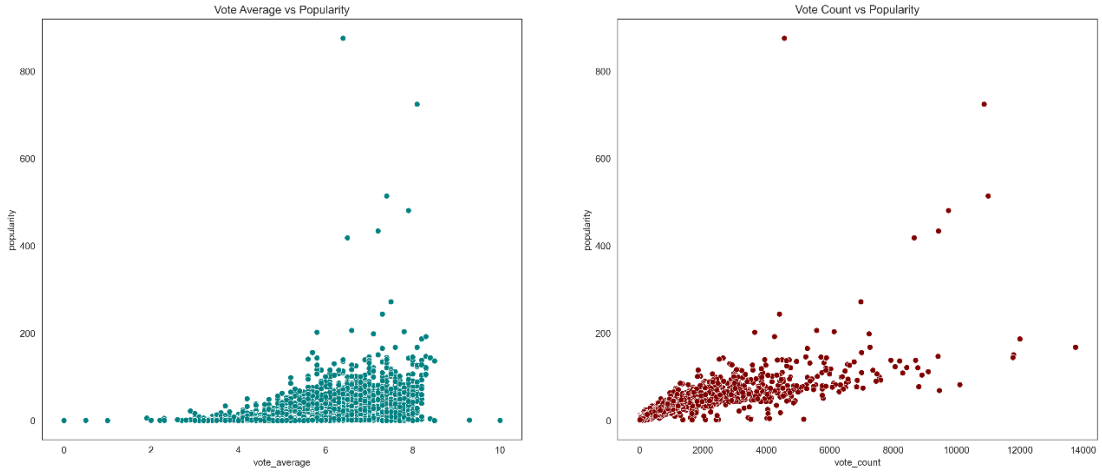

Fig 6. Relationships Between Popularity, Vote Count, and Vote Average

In the comparison between popularity and vote average, it's observed that the majority of movies with high popularity scores have an average vote score below 200. This indicates that a high popularity score does not necessarily correlate with a high average vote score. This disparity suggests that factors other than vote average contribute to a movie's popularity, such as the number of views, user engagement, and favoriting.

Additionally, examining the relationship between vote count and popularity reveals that the majority of vote counts are concentrated around the 0-4000 mark. However, as the vote count increases, there appears to be a slight uptick in popularity, implying a positive correlation between the two variables. This observation suggests that movies with higher vote counts tend to be more popular, although other factors may also influence popularity.

It's also noteworthy that there are outliers present in the scatter plots, indicated by dots far apart from the clustered data points. These outliers represent movies with exceptional values for one or more variables, which may warrant further investigation to understand their impact on the overall trends observed in the dataset.

Overall, the scatter plots provide a comprehensive visualization of the relationships between variables, offering valuable insights into the dynamics of movie popularity, vote count, and average vote scores within the dataset.

Word Clouds for Genres, Keywords, and Production Companies

Word clouds for textual data were generated to visualize the frequency of occurrence of genres, keywords, and production companies in the dataset, offering a visual representation of the most common elements.

Fig 7. Word Cloud of Keywords

The word cloud above visualizes the frequency of keywords extracted from the dataset. Some of the most prominent keywords include "woman director", "independent film", "duringcreditsstinger", "based on novel", and "murder", among others. These keywords provide insights into common themes, genres, or elements present in the movies within the dataset.

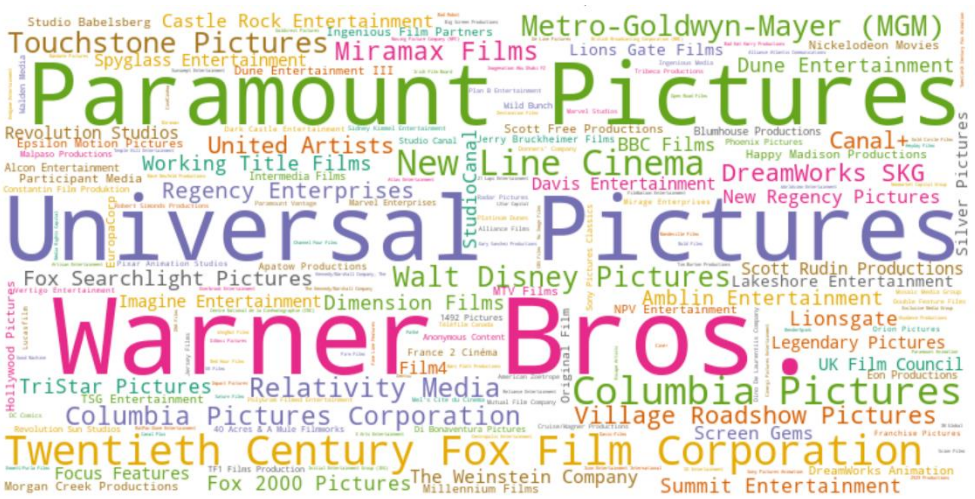

Fig 8. Word Cloud of Production Companies

The word cloud above illustrates the frequency of production companies extracted from the dataset. Some of the most prevalent production companies include "Warner Bros.", "Universal Pictures", "Paramount Pictures", "Twentieth Century Fox Film Corporation", and "Columbia Pictures", among others. These production companies are likely to have significant contributions to the movies within the dataset.

Fig 9. Word Cloud of Genres

The word cloud showcases the frequency of different genres present in the dataset. The most prevalent genres include Drama, Comedy, Thriller, and Action, indicating that these genres are highly represented among the movies in the dataset. Conversely, genres such as TV Movie and Foreign are less represented, with fewer instances in the dataset. This discrepancy in genre representation may impact future analyses, as comparisons and insights derived from genres with low representation should be interpreted with caution due to potential data skewness. Additionally, understanding the distribution of genres within the dataset is crucial for conducting accurate genre-based analyses and making informed decisions regarding movie genre classifications and recommendations.

Average Budget and Revenue by Genre

Bar plots illustrated the average budget and revenue for each movie genre, providing insights into the financial performance of different genres.

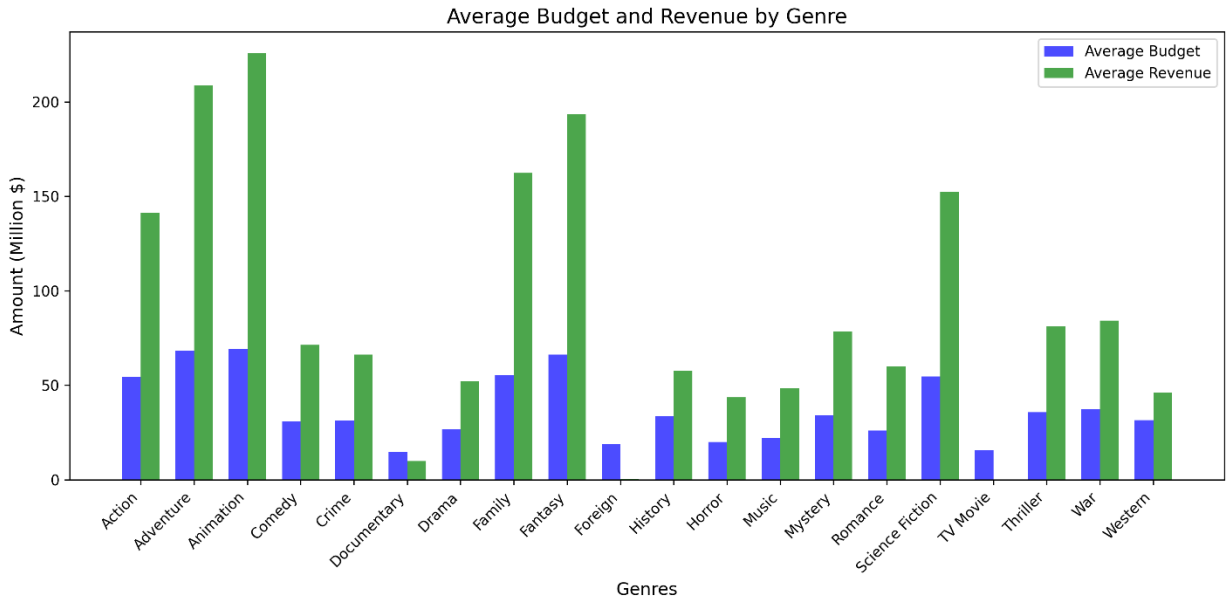

Fig 10. Average Budget and Revenue by Genre

When examining the average budget and revenue across different movie genres, several trends emerge.

Drama: With a total count of 2290 movies, Drama is the most represented genre in the dataset. Despite its high representation, Drama movies have a comparatively lower average budget of approximately $20.74 million, and the revenue averages around $52.28 million.

Comedy: Following Drama, Comedy stands out as the second most represented genre with 1715 movies. Comedy movies have a lower average budget of approximately $25.42 million and generate moderate revenue averaging around $71.58 million.

Thriller and Action: Thriller and Action genres follow suit, with 1272 and 1152 movies, respectively. These genres tend to have slightly higher average budgets, around $32.02 million and $51.60 million, and generate considerable revenue, with Thriller averaging approximately $81.17 million and Action averaging about $141.46 million.

Adventure: Despite being less represented in the dataset with 790 movies, Adventure movies have one of the highest average budgets, approximately $66.33 million, and generate substantial revenue averaging around $208.66 million.

Documentary and TV Movie: Genres like Documentary and TV Movie have relatively lower counts of 92 and 8 movies, respectively. Documentary movies have the lowest average budget and revenue among all genres, likely due to their low representation in the dataset. Similarly, TV Movies have a minimal average budget with no recorded revenue, which could also be attributed to their limited representation.

These insights into the financial performance of different genres can guide decision-making processes for filmmakers, producers, and investors in terms of resource allocation, investment strategies, and genre-specific production planning. Additionally, understanding the relationship between budget and revenue can aid in identifying potentially lucrative genres and optimizing financial returns in the film industry.

Movies Released per Month

The number of movies released per month was plotted to identify any seasonal trends or patterns in movie releases.

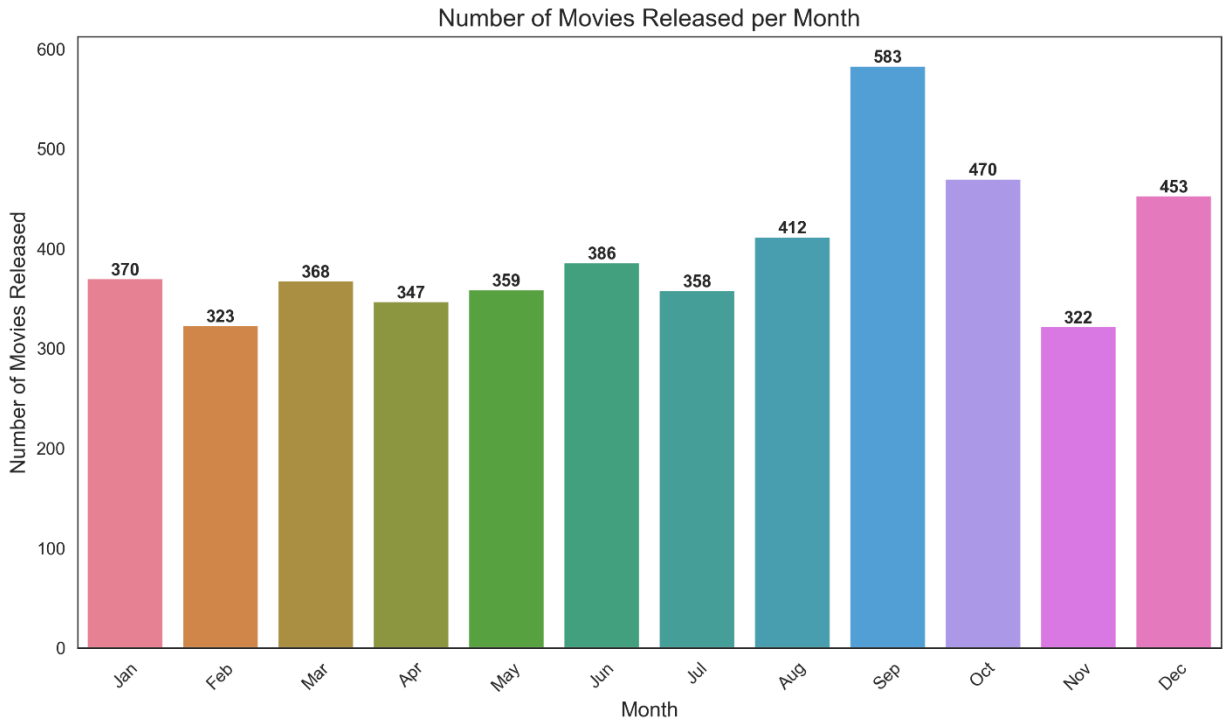

Fig 11. Number of Movies Released per Months

The distribution of movies released per month provides insights into the seasonal trends and patterns in movie releases. Surprisingly, the month of September has seen the maximum number of movie releases, followed closely by October and December. This contradicts the popular belief that August and September are "dump months" in the Hollywood film industry, characterized by lower movie attendance due to the beginning of the school year and other factors. However, the dataset reveals a different scenario, with September and October emerging as peak months for movie releases.

Boxplot of Average Votes Received by Month

The boxplot visualizes the distribution of average votes received by movies each month, offering insights into voting patterns and potentially indicating the quality of content released during different periods.

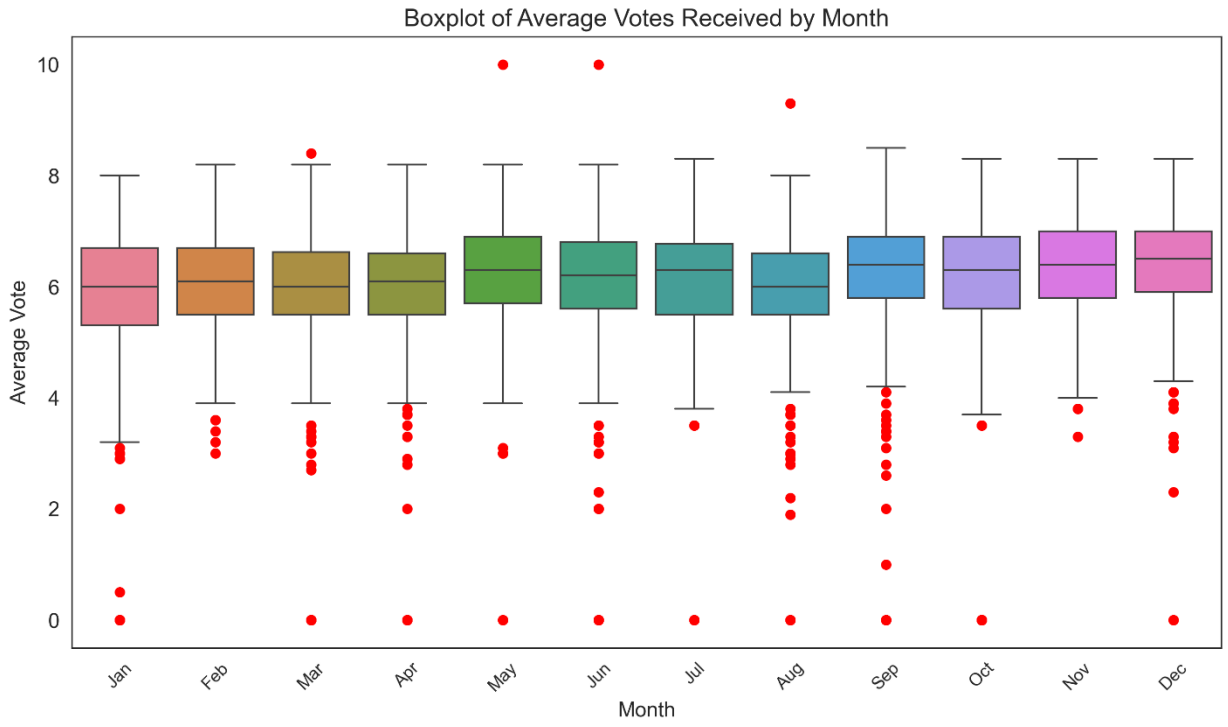

Fig 12. Boxplot of Average Votes Received by Month

The median for the month of December is the highest among all months, suggesting that movies released in December tend to receive higher average votes compared to other months. The spread of votes is more pronounced in the lower range for the month of September, indicating variability in the quality of content released during that month. An anomaly exists where movies in all months have received a rating of 0.0, which may be attributed to a data entry error. These instances should be removed to ensure accurate analysis. Upon further examination, the median vote count across all months falls within the range of 6, indicating moderate voting activity. Notably, movies released in January, May, June, July, and August have at least one instance of receiving a rating of 10.0. However, January also exhibits movies with significantly lower ratings, such as 0.5.

By delving deeper into the ratings of movies released each month, a more comprehensive understanding of audience preferences and content quality can be obtained, informing strategic decisions in the film industry.

Total Revenue by Month

A bar plot displayed the total revenue generated by movies released each month, highlighting any revenue trends over time.

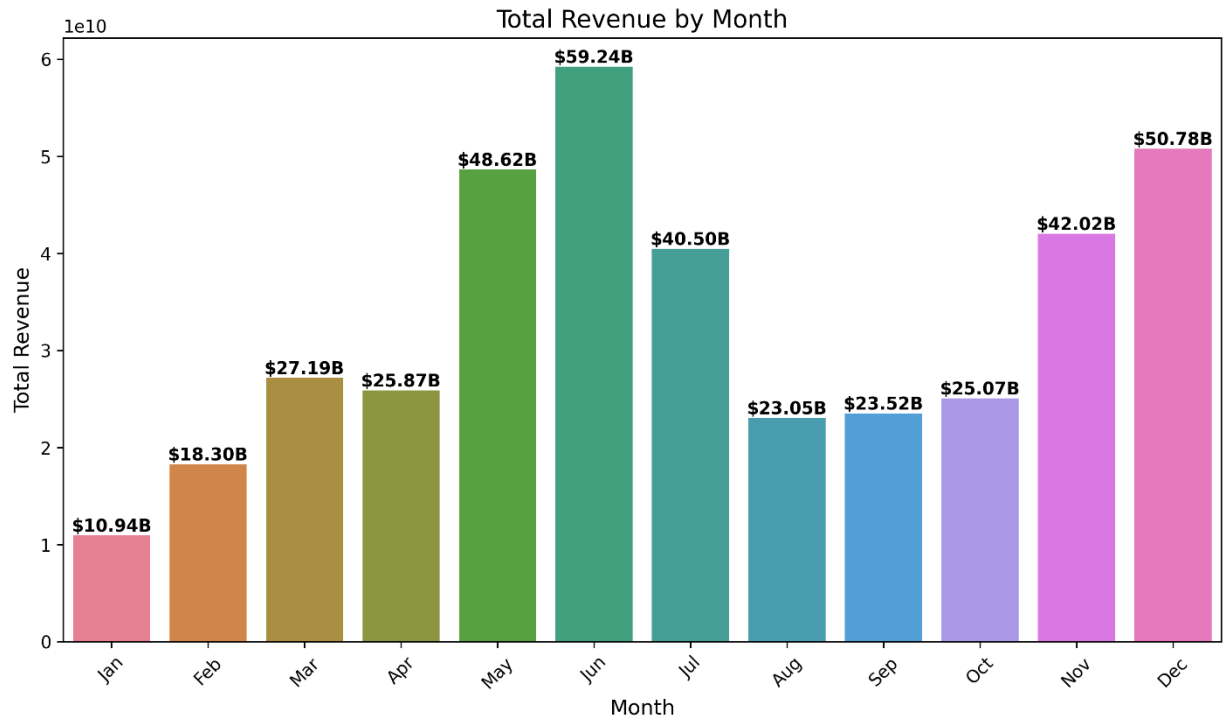

Fig 13. Total Revenue of Movies per Months

The total revenue generated by movies released each month provides insights into revenue trends over time. According to the provided data, June, May, December, and November stand out as the months with the highest total revenue. Months such as March, April, July, and August also contribute significantly to the total revenue, albeit to a lesser extent compared to the aforementioned months. January and February have comparatively lower total revenue, possibly due to factors such as post-holiday season slowdown and fewer major releases during these months, While December traditionally sees high revenue due to the holiday season. These trends suggest that certain months, particularly those coinciding with holiday seasons and peak moviegoing periods, tend to attract higher revenue. One thing to note is that the United States makes up a large proportion of the dataset, which may skew the results of both of these charts since they have their summer vacation in the months of May, June, July, and August. This may answer the question of high revenues at these months. However, other countries may have their school holidays on different times of the year.

However, the specific factors driving revenue variations across different months would require further analysis, considering factors such as blockbuster releases, competition from other forms of entertainment, and economic conditions during each period.

Profit Over Years

A time series plot illustrated the trend of average profit over the years, allowing for the analysis of profitability trends in the movie industry.

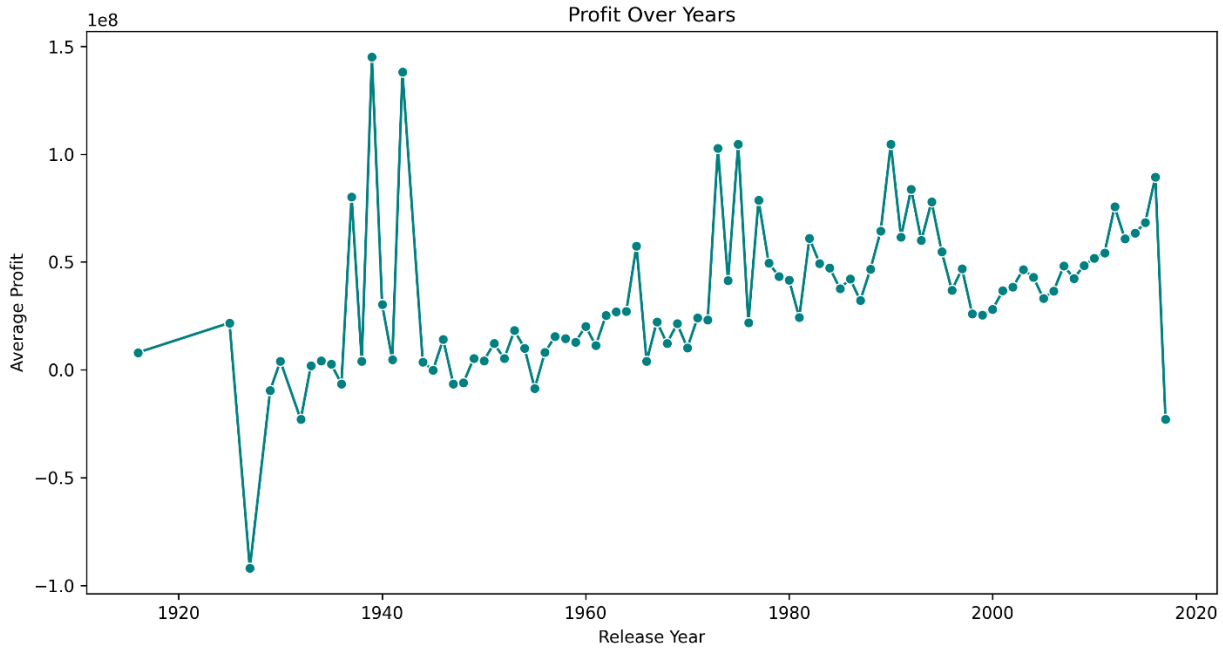

Fig 14. Profits of Movies over Years

Pairplot

A pairplot was generated to visualize the relationships between multiple numerical features simultaneously, facilitating a comprehensive analysis of feature interactions. The pairplot allows for the simultaneous visualization of relationships between multiple numerical features. In this analysis, relevant_columns including 'popularity', 'vote_average', 'vote_count', 'budget', 'revenue', and 'profit' were selected for examination. By plotting these variables against each other, we can gain insights into potential correlations and patterns among these features, aiding in a comprehensive analysis of their interactions.

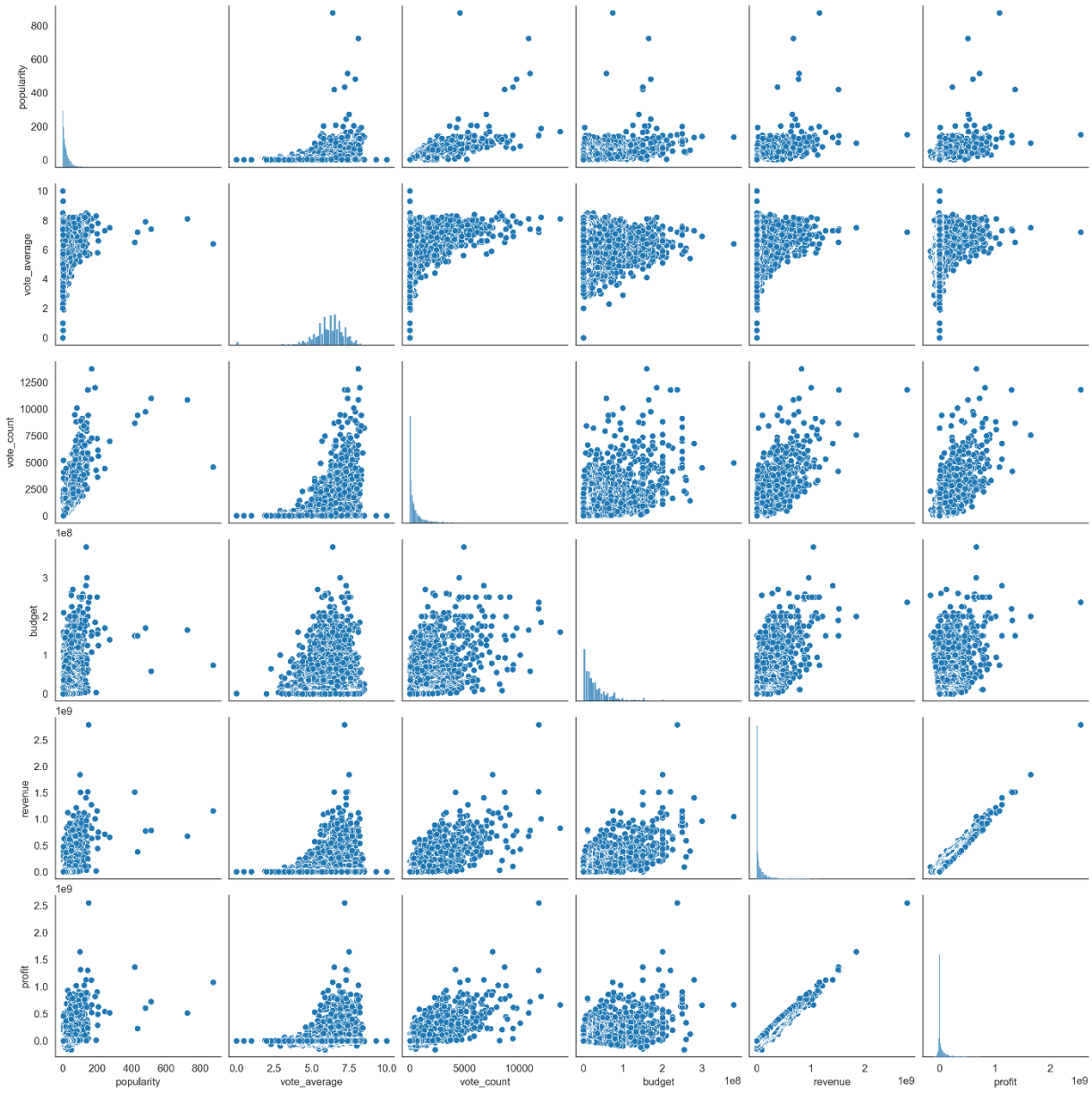

Fig 15. Relationships of Multiple Numerical Features

Overall, the exploratory data analysis phase provided valuable insights into various aspects of the movie dataset, laying the foundation for subsequent modeling and analysis tasks.

5. Data Modelling

In the data modeling phase, the profitability of movies was defined as the target variable, and feature engineering was performed to select relevant features for modeling. Machine learning models, including Logistic Regression, Decision Tree, Random Forest, and XGBoost were trained and evaluated.

For each model, metrics such as accuracy, precision, recall, and F1-score were reported, and ROC-AUC plots were generated to visualize model performance. The results were compared to determine the most effective model for predicting movie profitability.

Table 2. Models Evaluation

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.789	**0.817**	0.822	0.820	0.782
Decision Tree	0.744	0.782	0.779	0.781	0.737
Random Forest	**0.809**	**0.817**	**0.867**	**0.841**	**0.797**
XGBoost	0.799	0.806	0.865	0.834	0.786

Among the models evaluated, the Random Forest model demonstrated the highest overall performance with the highest accuracy, precision, recall, and F1-score. This suggests that Random Forest is the most effective model for predicting movie profitability based on the selected features. However, further fine-tuning and experimentation with hyperparameters could potentially improve the performance of the models even further.

The ROC curves for each model, along with their respective AUC scores, are shown below. These plots visualize the trade-off between the true positive rate and false positive rate for different classification thresholds. It's evident that Decision Trees have the weakest performance, with a lower AUC score compared to the other models. Conversely, Logistic Regression, Random Forest, and XGBoost exhibit similar performance, with Random Forest demonstrating the highest AUC score among them. This reaffirms the findings from the performance metrics presented in Table 2, where Random Forest emerges as the best-performing model in terms of accuracy, precision, recall, F1-score, and ROC-AUC.

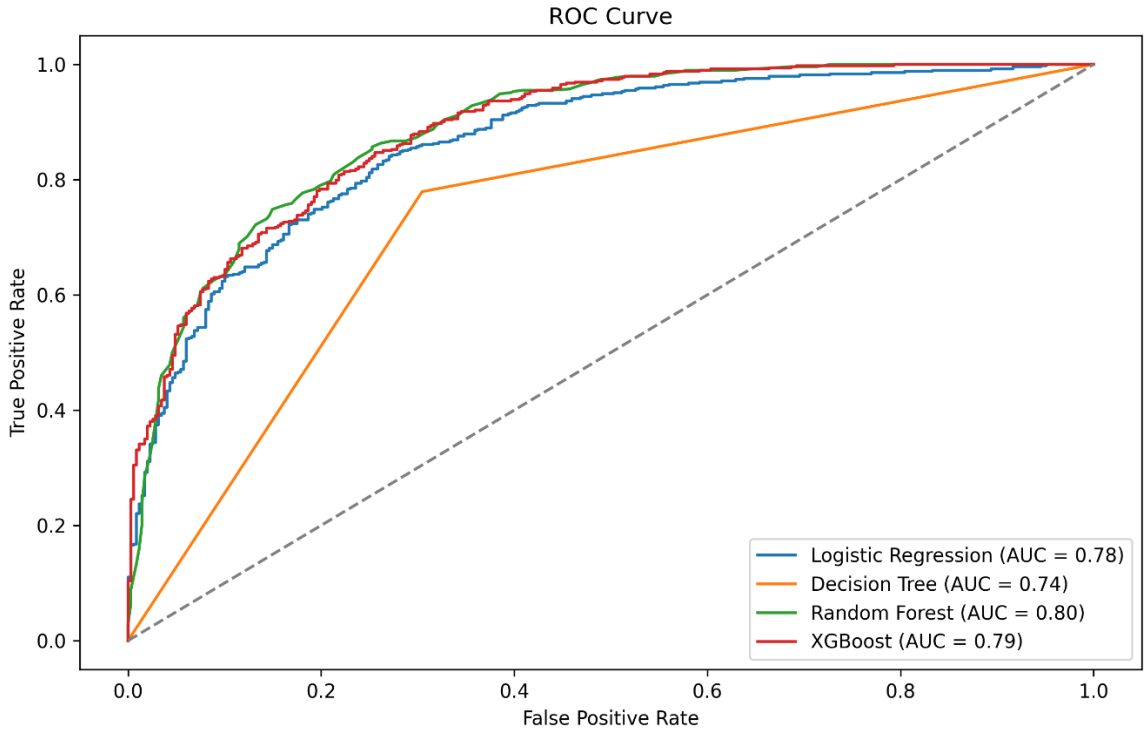

Fig 16. ROC Curves of Different Machine Learning Models

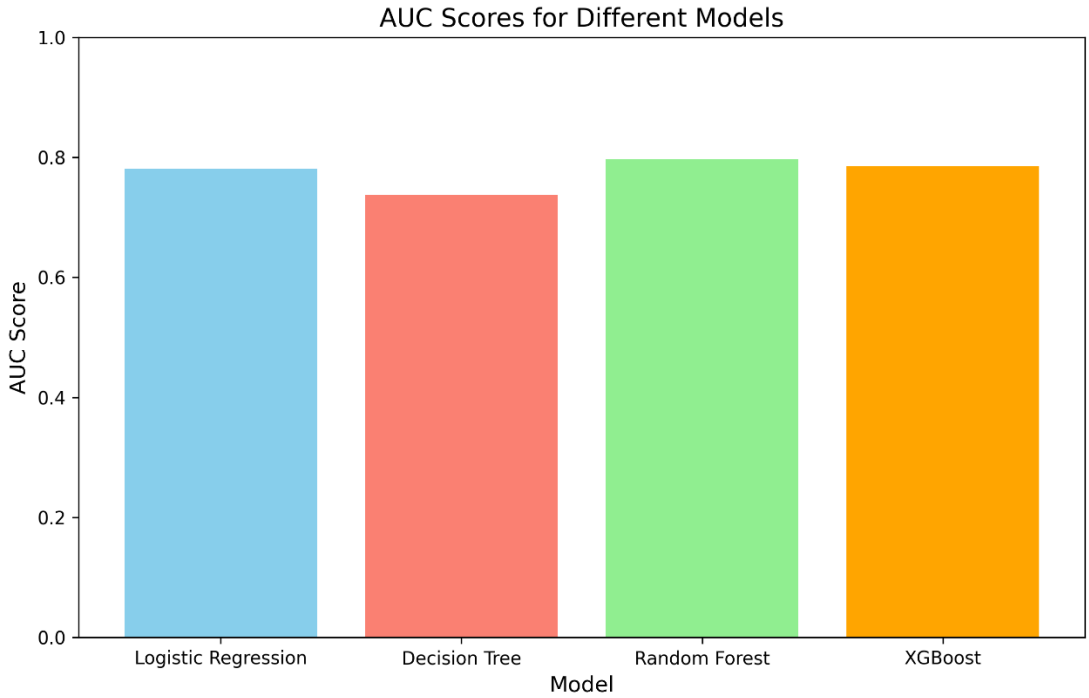

Fig 17. AUC Scores

The confusion matrices for each model provide insights into the performance of the classifiers by showing the counts of true positive, true negative, false positive, and false negative predictions.

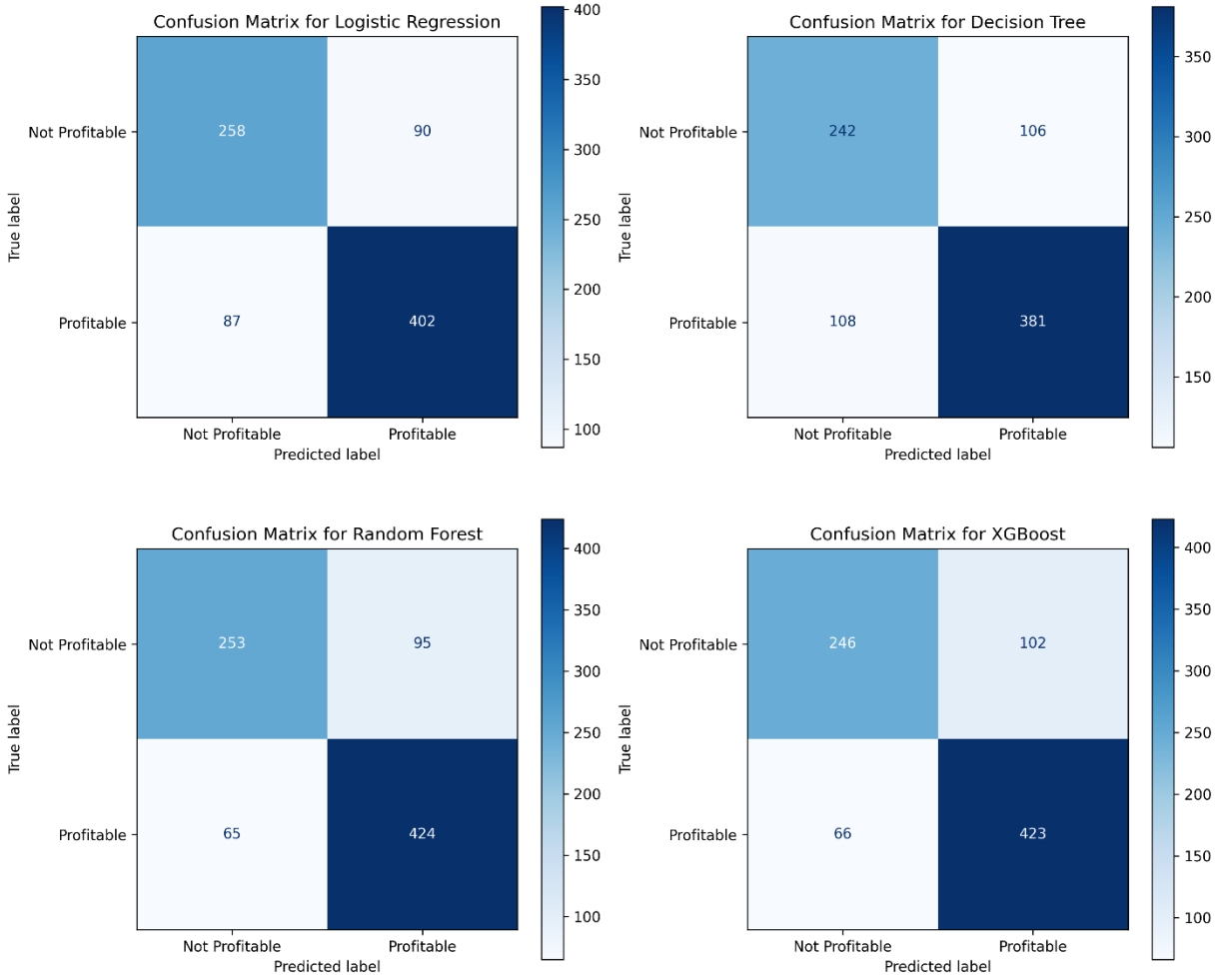

Fig 18. Confusion Matrix of Different Machine Learning Models

Based on the confusion matrices for each model:

XGBoost: This model correctly classified 246 instances as not profitable (True Negatives) and 423 instances as profitable (True Positives). However, it misclassified 102 instances as profitable when they were not (False Positives) and 66 instances as not profitable when they were (False Negatives).

Random Forest: Similar to XGBoost, Random Forest correctly classified 253 instances as not profitable and 424 instances as profitable. It misclassified 95 instances as profitable when they were not and 65 instances as not profitable when they were.

Logistic Regression: Logistic Regression correctly classified 258 instances as not profitable and 402 instances as profitable. However, it misclassified 90 instances as profitable when they were not and 87 instances as not profitable when they were.

Decision Trees: Decision Trees correctly classified 242 instances as not profitable and 381 instances as profitable. It misclassified 106 instances as profitable when they were not and 108 instances as not profitable when they were.

Overall, while all models achieved reasonable accuracy, Random Forest performed the best in terms of minimizing misclassifications, followed closely by XGBoost. Logistic Regression and Decision Trees exhibited slightly weaker performance, with a higher number of misclassifications compared to Random Forest and XGBoost.

6. Recommender System

In addition to predicting movie profitability, a recommender system was developed to suggest similar movies based on shared characteristics such as genres, keywords, and cast members. This system utilizes the TF-IDF vectorization technique to represent these features in a vectorized space, allowing for the calculation of similarity scores between movies.

Vectorization of Features: The cast, genres, and keywords of each movie were vectorized using TF-IDF vectorization. This process involves transforming textual data into numerical representations suitable for machine learning algorithms.

Combining Vector Spaces: The vectorized features were combined into a single feature space to capture the overall similarity between movies based on their cast, genres, and keywords.

Calculating Similarity: The cosine similarity metric was employed to calculate the similarity scores between movies in the combined vector space. This metric measures the cosine of the angle between two vectors, indicating the degree of similarity between them.

Recommendation Generation: Using the similarity scores, the system generates a list of top movie recommendations for a given input movie. These recommendations are selected based on their highest similarity scores with the input movie.

The recommender system successfully generated accurate movie recommendations based on shared characteristics. Each recommendation is accompanied by an explanation highlighting the common genres, keywords, and cast members between the input movie and the recommended movies.