Title: TMDB 5k Dataset Analysis

Author: Mohsen Amiri Amjad

Github: TMDB Exploratory and Predictive Analysis

Dataset: TMDB 5000 Movie Dataset | Kaggle

INTRODUCTION

This report presents an in-depth analysis of the TMDB Movie Dataset, available on Kaggle. The dataset comprises several thousand films, providing data on the plot, cast, crew, budget, and revenues. The primary aim of this project is to gain insights about the data for further modeling to investigate the profitability of the movies.

The project follows a structured approach starting with data loading and gaining preliminary information about the dataset. This is followed by data cleaning and preprocessing to ensure the data is suitable for analysis. The next stage involves a comprehensive Exploratory Data Analysis (EDA) based on the features and different questions that may arise in the dataset. The EDA phase aims to gain more exploratory insight about the data, which will be crucial for the subsequent modeling phase.

The modeling phase involves defining a target based on the profit column to indicate whether a movie has been profitable or not. Feature engineering is performed to select a subset of useful features for training a model. The models used for prediction include Logistic Regression, Decision Tree, Random Forest, and XGBoost. Each model's performance is evaluated using four metrics: Accuracy, Precision, Recall, and F1-Score. Additionally, the ROC-AUC plot for each model is generated for a more visual comparison.

The final part of the project involves designing a simple recommendation system. Research is conducted on different recommendation system schemes, such as Collaborative filtering and content-based filtering, to design a simple recommender system.

This project aims to provide valuable insights into the factors that contribute to a movie's success and profitability. It also seeks to develop a reliable model for predicting a movie's profitability and a recommendation system that can suggest movies to users based on certain criteria. The findings from this project could be beneficial for various stakeholders in the film industry, including filmmakers, investors, and marketers, to make informed decisions and strategies.

CONTENTS

OVERVIEW

The dataset is sourced from Kaggle and contains comprehensive information and statistics about roughly 5000 movies. This dataset provides a wealth of information for analyzing and understanding trends in the movie industry.

STRUCTURE

The dataset consists of 4803 rows and 15 columns. Each row represents a movie, and each column represents a feature of said movie. The columns in the dataset are:

FEATURE	DESCRIPTION
Movie ID	The unique identifier for each movie.
Title	The title of the movie.
Cast	The cast of the movie.
Budget	The budget of the movie.
Genres	The genres of the movie.
Keywords	The keywords associated with the movie.
Original Language	The original language of the movie.
Popularity	The popularity score of the movie.
Production Companies	The production companies of the movie.
Release Date	The release date of the movie.
Revenue	The revenue of the movie.
Runtime	The runtime of the movie.
Status	The status of the movie.
Average Rating	The average vote score of the movie.
Vote Count	The vote count of the movie.

MISSING VALUES

There are very few missing values in the dataset. Specifically, the `release date` column has 1 missing value, and the `runtime` column has 2 missing values. The rest of the dataset does not seem to have any missing values. However, those columns with a list or dictionary format (like `cast`, `genres`, etc.) might contain some empty structures.

UNIQUE VALUES

The dataset, with a total of 4803 rows, exhibits a wide range of unique values across its features. The categorical features such as `original language` and `status` have varying numbers of classes. For instance, `original language` has 37 unique classes, while `status` has only 3. This indicates a high degree of diversity in the dataset's categorical features.

The numerical features also show considerable variability. For example, `budget` has 436 unique values, `runtime` has 156, `average rating` has 71, and `vote count` has 1609. This suggests that these features span a broad range of values, which could reflect the diverse nature of the movies included in the dataset.

The features `movie id` and `title` nearly have unique values for each row, indicating that they can probably serve as identifiers for the individual movies. Lastly, the `release date` and `revenue` features, with 3280 and 3297 unique values respectively, might provide insights into the temporal distribution of the movies and their financial success.

DESCRIPTIVE STATISTICS

The dataset provides a wide range of information about the movies. This section will provide a preliminary description of the numerical features of the dataset. The average `budget` of a movie is approximately 29 million, but there is a high variability in movie budgets, as indicated by a standard deviation of around 40.72 million. The `popularity scores` also vary widely, with an average score of around 21.49 and a standard deviation of 31.82. The `runtime` of movies shows moderate variability, with an average of approximately 106.88 minutes and a standard deviation of around 22.61 minutes. The `average rating` shows relatively low variability, with an average score of around 6.09 and a standard deviation of 1.19. The `revenue` of movies shows high variability, with an average of approximately 82.26 million and a standard deviation of around 162.86 million. It's important to note that the presence of 0s in the budget, runtime, and revenue columns could potentially skew these statistics and will be handled in the next phases.

PREPROCESSING

Here is a comprehensive report on the preprocessing process of the data. This process involved cleaning, reforming and adding features to the dataset and saving the result.

INITIAL DATA

The initial data consists of two dataframes: `credits df` and `info df`. These dataframes are merged into a single dataframe `movies df` based on the `movie id` column.

DATA CLEANING

In the first step, the data is cleaned using statistical, model-based, and feature-based techniques. Columns with numerical values have been cleaned separately from the non-numeric features. Here is a general overview of the cleaning process:

CLEANING NUMERICAL COLUMNS

This section is dedicated to cleaning and preprocessing the numerical columns in the dataset. This involves managing wrong values, handling missing values, and dealing with outliers.

MANAGING WRONG VALUES

In this subsection, movies with zero `runtime` are identified and removed from the dataset. This is because a runtime of zero does not make sense in the context of movies. The rest of the numerical features did not seem to have semantically or statistically wrong values and therefore did not have to change.

MANAGING MISSING VALUES

The `runtime` column has some missing values which are imputed using the average value of this column. The only other missing value is in the `release date` column which was also filled with the average. The rest of the numerical columns did not have any missing values.

MANAGING OUTLIERS

Outliers in the dataset are identified and managed in this subsection using the z-score method which is a standard-deviation-aware statistical technique. A threshold of 3.5 is used to identify the outliers of each feature. A total of 283 outliers were detected, which then were removed from the data to decrease the skewness of the distribution of features. The skewness of each column is shown in Table 1. The histograms of these columns have been shown in Figure 1 for a more intuitive visualization of the changes.

FEATURE	SKEWNESS	
	With Outliers	Without Outliers
Budget	2.437	1.828
Popularity	9.721	1.710
Revenue	4.444	2.516
Runtime	1.831	0.896
Average Rating	1.959	0.598
Vote Count	3.824	2.568

Table 1: Skewness of features before and after removing outliers from the data.

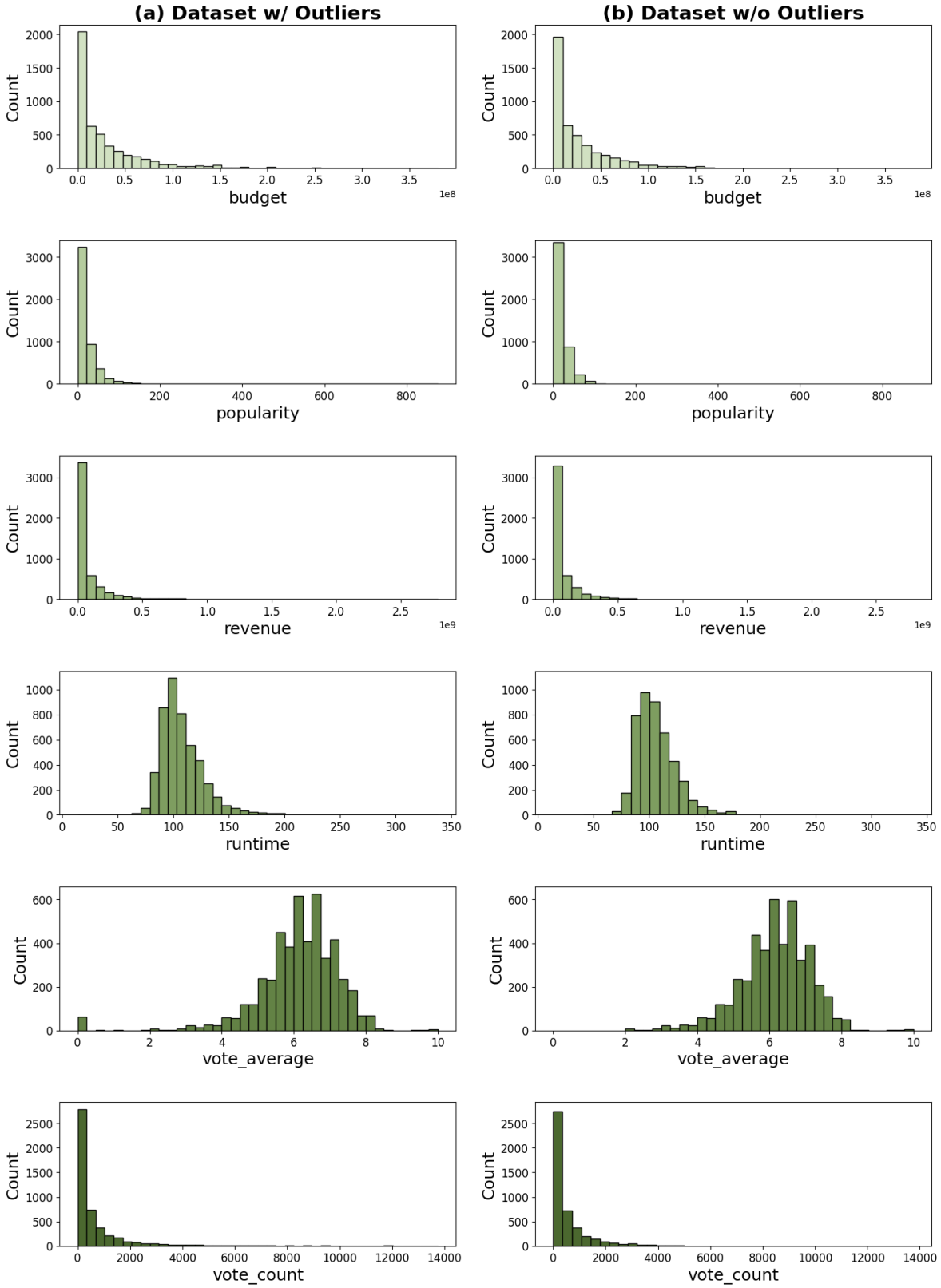


Figure 1: distributions of each numerical column. (a) the distribution of the feature before removing outliers from the dataset. (b) the distribution of the feature after removing outliers.

CLEANING NON-NUMERICAL COLUMNS

This section is dedicated to cleaning and preprocessing the non-numerical columns in the dataset. This involves generating separate and tabular columns for `casts`, `production companies`, and `genres` as new features of the dataset.

ADDING ACTORS AS A FEATURE

This subsection involves creating a separate dataframe for the `cast` of each movie. The dataframe is exploded on the `cast` column to create a row for each cast member of each movie. The optimal number of columns to have for cast members was calculated to be 5; Any higher number of columns would increase the dimensionality of the dataset and contain high numbers of missing values for movies with limited cast members. Furthermore, excess information about the cast is removed, and the dataframe is cleaned to remove any rows with missing values.

Two ranking methods are considered for ranking and choosing the top 5 actors of each movie:

1. The average revenues of their movies.
2. The provided order in the TMDB dataset.

The first method is examined and found to put too much high value on side characters of super successful movies, which is not convenient. Therefore, the provided order (second method) is used. The missing actor fields are filled with a constant token `#John Doe#` for future reference.

ADDING PRODUCTION COMPANIES AS A FEATURE

This subsection involves creating a separate dataframe for the `production_companies` of each movie. The dataframe is exploded on the `production_companies` column to create a row for each production company of each movie. The optimal number of columns to have for production companies was calculated to be 1; Since higher number of columns would increase the dimensionality of the dataset and contain high numbers of missing values. Excess information about the production companies is removed, and the dataframe is cleaned to remove any rows with missing values. The missing production company fields are filled with a constant `#Jane Doe Inc#` token.

ADDING GENRES AS A FEATURE

In this subsection, all the genres are extracted and a dataframe containing movie-genre pairs is created. The dataframe is exploded on the `genres` column to create a row for each genre of each movie. The genres are then multi-hot encoded to create binary features for each genre. The one-hot encoded dataframe is merged with the main dataframe and the `genres` column is dropped from the main dataframe. Finally, any missing values in the genre columns are filled with False. Here is a list of all the new genres: `Crime`, `Comedy`, `Drama`, `Music`, `Adventure`, `Fantasy`, `Action`, `Thriller`, `Science Fiction`, `Romance`, `War`, `Western`, `Animation`, `Family`, `Mystery`, `History`, `Horror`, `Documentary`, `Foreign`, and finally, `TV Movie`

MINOR MODIFICATIONS

This section involves several cleaning tasks to prepare the dataset for further analysis or modeling. These tasks include removing unnecessary columns, adding label-encoded columns for categorical features, splitting the `release date` into separate columns, binning the `revenues`, and adding a `profit rate` to the dataset.

REMOVING THE ID COLUMN

The `movie_id` column is dropped from the dataframe since the default index can be used as a unique ID and no other reference uses the `movie_id` feature.

ADDING LABEL-ENCODINGS

Label-encoded columns are added for each of the categorical columns in the data. This includes `status`, `language`, `actors`, and `production company`. The categorical columns are converted to category data type, and then numerical codes are assigned to each category.

SPLITTING THE RELEASE DATE

The `release_date` column is split into three separate columns: `year`, `month`, and `day`. This allows for more granular analysis based on the release dates of the movies.

ADDING PROFIT RATES

A `profit_rate` column is added to the dataset. This is calculated as the ratio of `revenue` to `budget`. Any infinite values resulting from division by zero are replaced with NaN, and any profit rates greater than 100 are capped for better intuitive visualizations. Note that the capped version is not used in analyses where this capping could be problematic.

SAVING THE CLEAN DATASET

Finally, the cleaned and preprocessed dataset is saved to a CSV file for future use. This new CSV file can be accessed in the data directory with the name of "`movies_data.csv`".

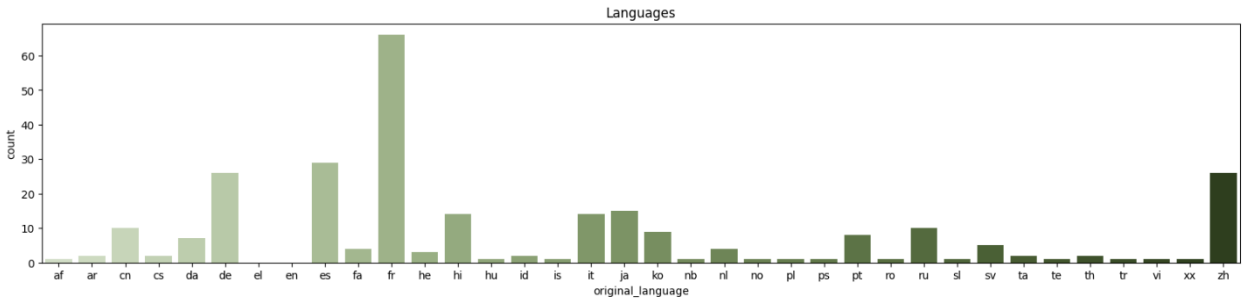


Figure 2: a bar plot showing the number of movies made in each language. (English has been removed from this list to improve the visibility of the other movies.)

DATA ANALYSIS

In this section we will explore the dataset using visualization along with statistical methods to find patterns and intuitions into how different factors affect a movie's success.

BASIC EXPLORATION

This section is dedicated to performing basic exploratory data analysis and visualizations on the dataset.

DISTRIBUTION OF CATEGORICAL FEATURES

We first plot the distributions of the categorical features to get a general understanding of the data.

GENRE VARIABILITY

The variability of genres in the dataset is visualized in Figure 3Figure 10. A pie chart is also plotted for each genre (Figure 10), showing the proportion of movies that fall into each genre. The pie charts are color-coded for better visualization. As it can be seen in the charts, Drama, Comedy, and Thriller movies are top three most common genres respectively. TV Movie, Foreign, and Western movies seem to have the least number of movies.

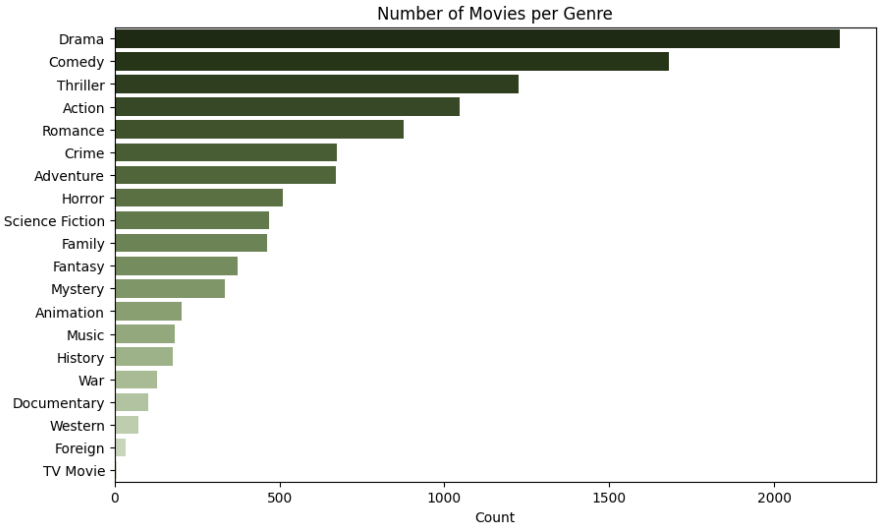


Figure 3: the number of movies in each genre.

STATUS

In this section, we visualize the `status` feature. The pie chart in Figure 4 shows that out of the 4520 movies in this dataset, only 4 have the "`rumored`", 3 have the "`post_production`" status, while the rest of the movies are all released.

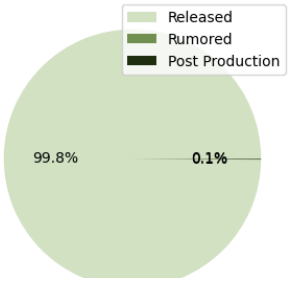


Figure 4: a pie chart depicting the almost-homogeneous distribution of statuses.

LANGUAGE

There are 36 unique languages. Unsurprisingly, the English language has the highest count of movies with 4231 movies, taking up %93.6 of the dataset. The rest of the languages are shown in Figure 2. French, Espagnole, Chinese, and German are the top 3 Languages other than English, with 69, 32, 27 and 27 movies in each of said languages respectively.

HISTOGRAMS OF NUMERICAL FEATURES

This section is dedicated to exploring the numerical features in the dataset. This involves visualizing the distribution of various numerical features such as `budget`, `popularity`, `revenue`, `runtime`, and `rating`.

BUDGET

In this subsection, this histogram is plotted to visualize the distribution of the `budget` feature in the dataset. The x-axis is formatted to display the budget in millions of dollars for easier interpretation.

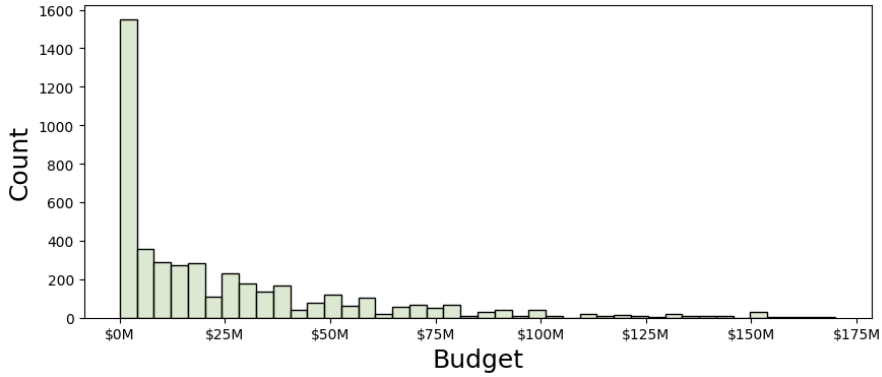


Figure 5: the distribution of movie budgets.

POPULARITY

In this subsection, a histogram is plotted to visualize the distribution of the `popularity` feature in the dataset.

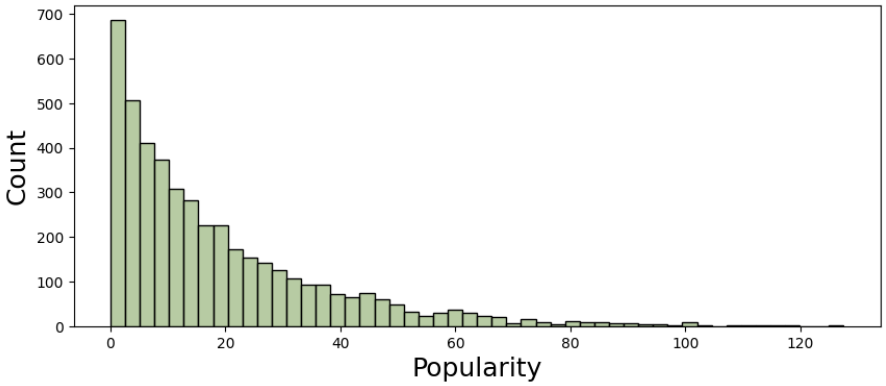


Figure 6: the distribution of movie popularities.

Revenue

In this subsection, a histogram is plotted to visualize the distribution of the `revenue` feature in the dataset. The x-axis is formatted to display the revenue in millions of dollars for easier interpretation.

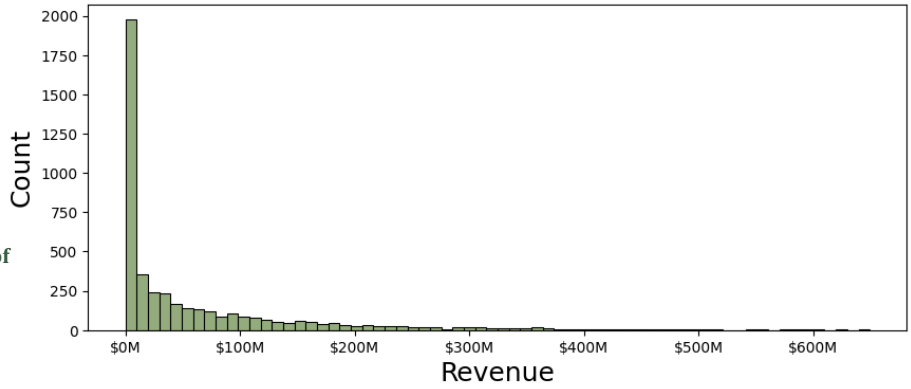


Figure 7: the distribution of movie revenues.

Runtime

In this subsection, a histogram is plotted to visualize the distribution of the `runtime` feature in the dataset.

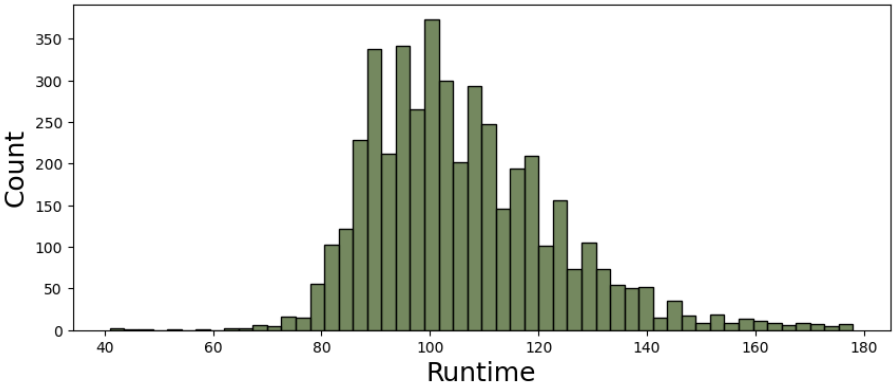


Figure 8: the distribution of movie runtimes.

Average Rating

In this subsection, a histogram is plotted to visualize the distribution of the `vote_average` feature in the dataset, which represents the average rating of the movies.

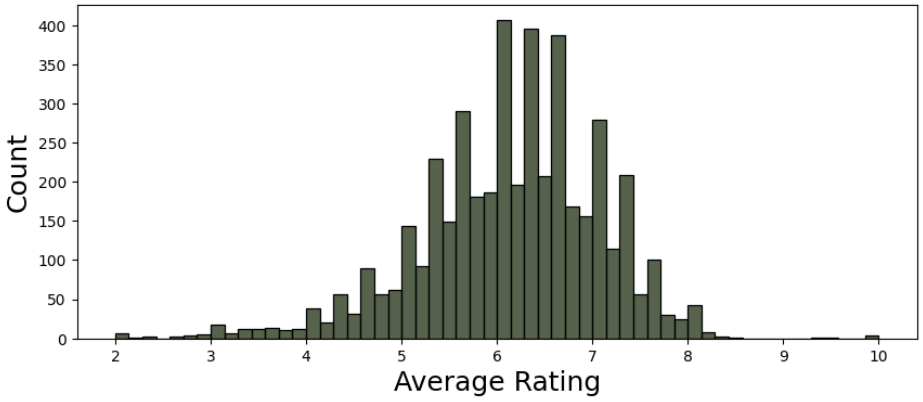


Figure 9: the distribution of movie ratings.

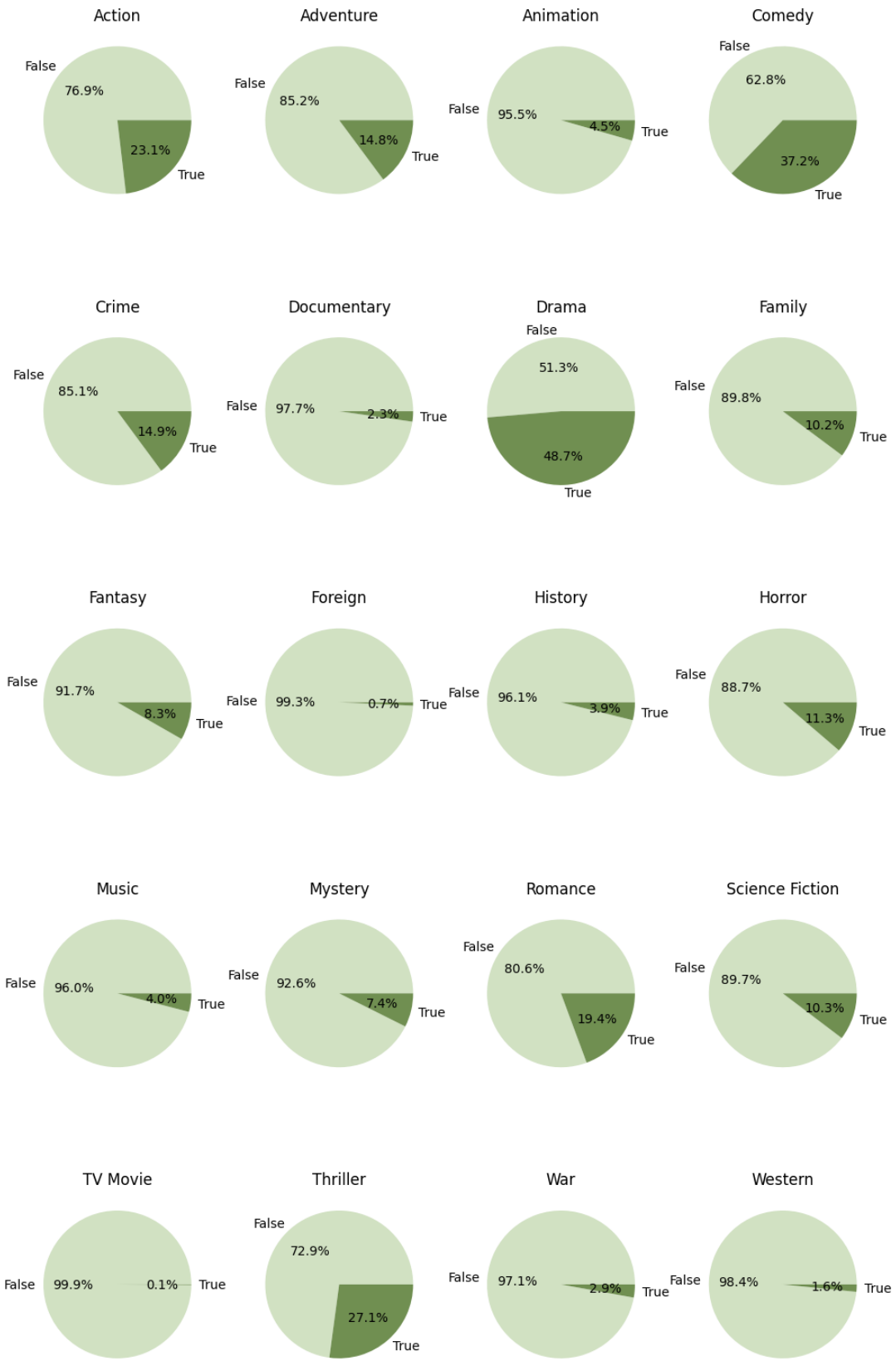


Figure 10: pie charts showing the distribution of different genres.

FEATURE IMPORTANCE AND CORRELATIONS

In this section, we focus on exploring the correlations between different features and their importance in predicting the target variable. For this analytic purpose, revenue of movies was binned into 30 separate bins and was used as the target variable.

CORRELATIONS

In this subsection, heatmaps are plotted to visualize the correlation between different groups of features and the revenue of movies. We expect the highest correlated features with profit and revenue to be among the numerical features since they have a distributed and continuous values.

GENRES

As it can be seen from Figure 11, no significant correlation exists between the profit ratio of a movie, and the genre; Meaning, there doesn't seem to be any best genre in the sense of profitability. On the other hand, Adventure and Animation genres seem to have a more significant correlation with revenue. This means that these genres have a better chance of a higher revenue. However, higher revenue does not necessarily indicate higher profits. In addition, the correlations between genres have also been plotted for further analysis.

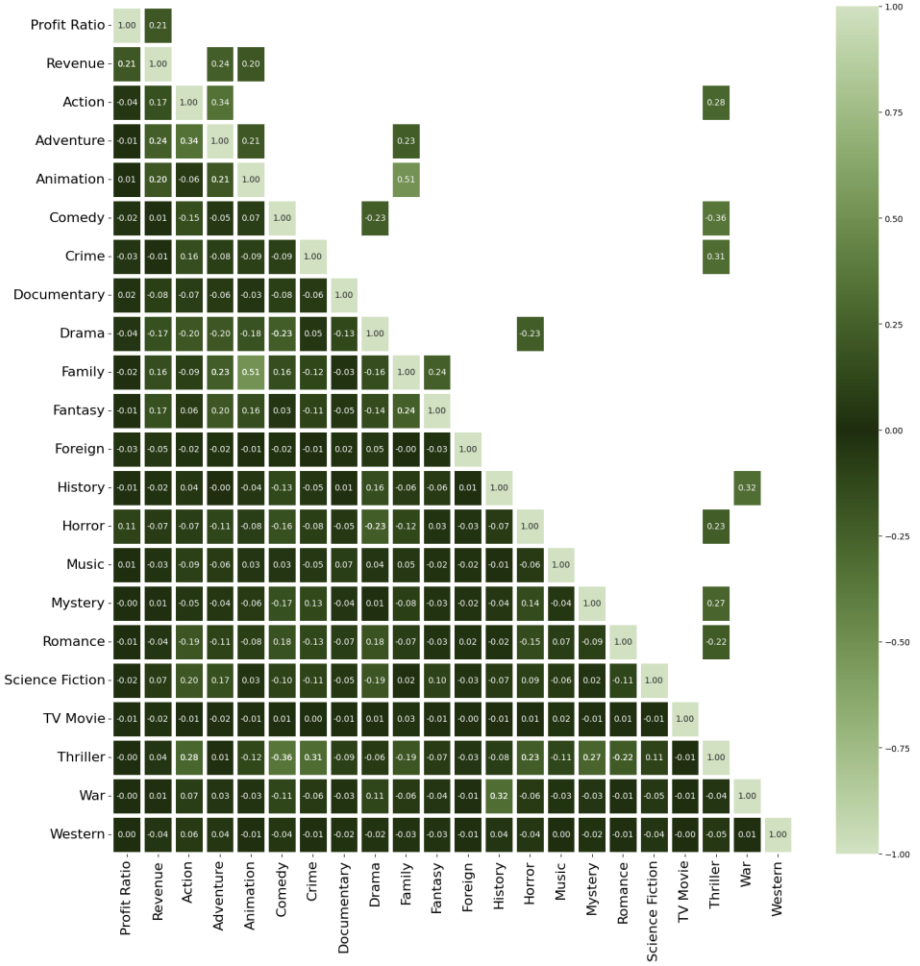


Figure 11: correlation matrix between different genres and revenue. Correlations higher than 0.2 were highlighted in the upper triangle of the matrix.

Actors and Production Company

Correlations between the cast, the production company, revenue, and profit were studied as well. Figure 12 shows a heatmap of these correlations. As can be seen, no significant correlation exists among these features. All the correlations sit inside the [-0.1, 0.1] range except for production company and revenue, which show a correlation of 0.16. This correlation is insignificant and could merely be because some companies have higher budgets and therefore have higher revenue as well. These low correlations can be due to label-encodings of these features.

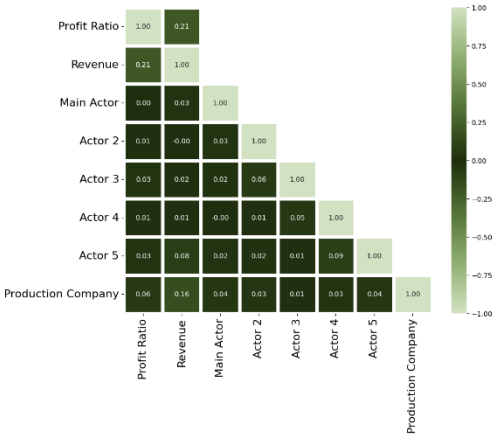


Figure 12: correlations between revenue and profit rate, and the cast members and production company of movies. Correlations higher than 0.2 were highlighted in the upper triangle of the matrix.

Numerical Features

In this subsection, the correlation between the financial success of movies, and the numerical features have been studied. As we can see in Figure 13, three features show prominent correlations with the revenue of the movie. These features are `vote count`, `budget`, and `popularity` of the movie. Additionally, the year in which a movie was released has a somewhat significant negative correlation (-0.21) with the profit rate of a movie; Meaning, the movie industry used to be more profitable in the past and is getting less profitable as the time goes by. Finally, revenues have a

Among other features, `runtime` and `rating` have a correlation of 0.35, which demands further analysis. A higher budget in a movie corresponds to a higher popularity and number of votes based on the correlation matrix. The highest correlation seems to be between popularity of a movie, and the number of votes, which could indicate a direct use of the vote count in formulation of the popularity score.

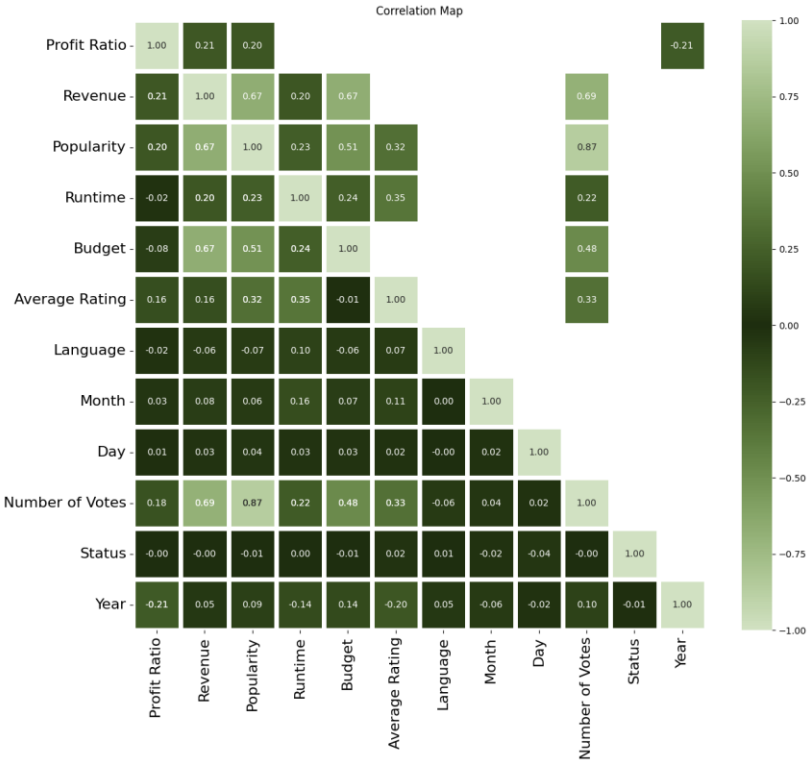


Figure 13: correlations between revenue and profit rate, and the numerical features of movies. Correlations higher than 0.2 were highlighted in the upper triangle of the matrix.

General Correlations

We also examine the correlations between all the non-text features. The figure of this correlation map has not been included in this report due to size limitations.

Based on our analyses, Adventure, Action, Animation, and Fantasy have the highest correlations with the `budget` feature, meaning these genres tend to have higher production budgets.

In Addition, Drama and Historical movies have the highest positive-valued correlation with the length of movies which means they tend to be longer. On the other hand, Animation and Comedy movies have the highest negative-valued correlations with `runtime`.

FEATURE IMPORTANCE

In this subsection, two random forest classifiers are trained on the dataset to predict the revenue and profit separately. The accuracy of each model is calculated, and a classification report is given to evaluate the performance of each model. This provides an indication of the importance of different features in predicting revenue and profit. These findings have been plotted in **Error! Reference source not found.**. The most important features for both targets are Popularity, Number of Votes and Budget Respectively, with importances of 8-12 percent. On one hand, based on the revenue of movies, many other features have importances of %4 to %6, while none of the genres were of high importance. On the other hand, the year of release has a more significant importance in predicting the profits gained by a movie. This finding is in line with our previous assumption about the drop in profits in recent years compared to years before. We will delve deeper into this subject to confirm our suspicions.

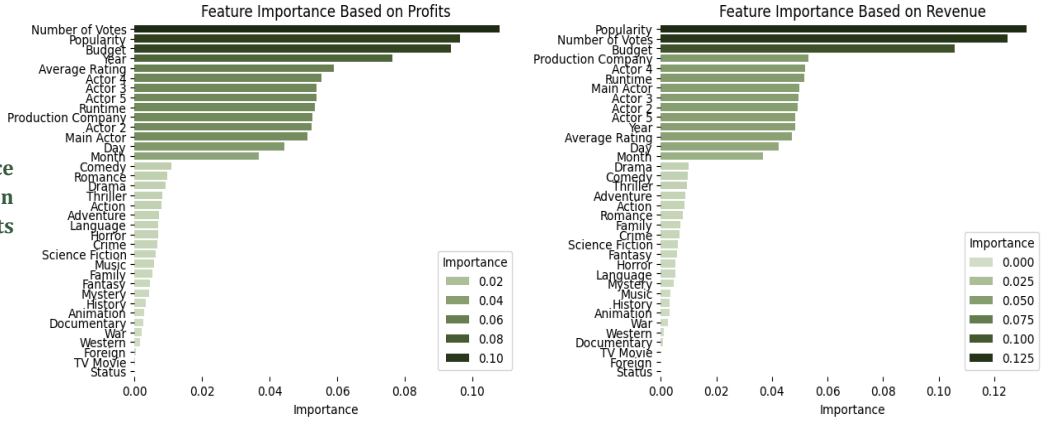


Figure 14: importance of features in prediction of revenue and profits of movies.

EXPLORATORY DATA ANALYSIS

In this section some key questions about qualitative and quantitative measures of the data are proposed and answered.

POPULARITY & NUMBER OF VOTES

In previous analyses, popularity and the number of votes showed a significant role in revenue and profitability of a movie. Figure 15 shows the high correlation between popularity and number of votes. As the red arrow indicates, there is a direct relationship between these two features.

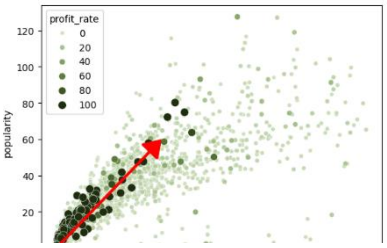


Figure 15: a scatterplot of movie popularity with respect to the number of votes.

Focusing on profits and revenues, based on Figure 16, we can see take some important notes about popularity and the number of votes:

1. Almost all movies are above the solid black line drawn in Figure 16. This shows a direct lower bound between the revenue of a movie and the number of people that vote for the said movie.
2. Almost none of the movies with revenues higher than roughly $250 Million have a popularity of more than 60x; With the four exceptions being "Jaw", "The Exorcist", "Crocodile Dundee", and "My Big Fat Greek Wedding".
3. Most of the movies with a high profit ratio sit between the two dashed lines in Figure 16.
4. We can also see a visible correlation between revenue and the plot has a smooth upward direction.

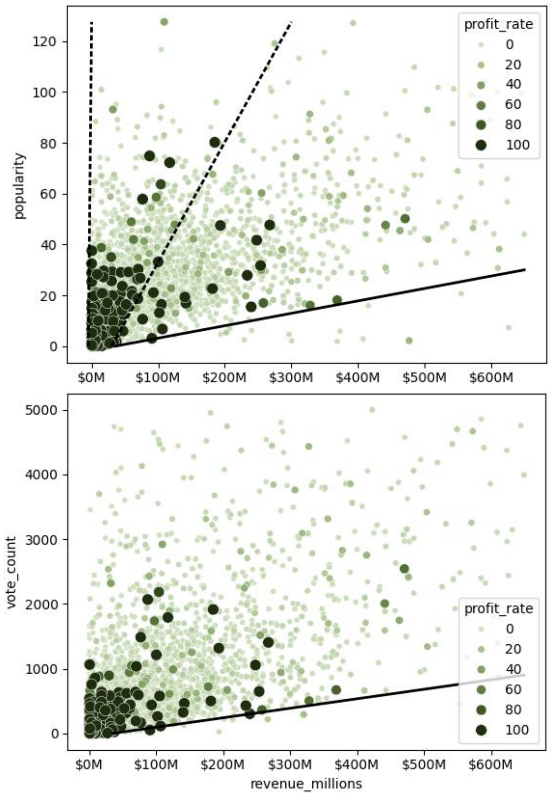


Figure 16: the scatter plot of the revenue and profit of movies based on their popularity and vote_count.

BUDGET & RUNTIME

In this section, we focus on the budget and run time of movies.

BUDGET

We first plot the relationship between the budget of movies and their revenue. We also color the points based on the length of the movie. As we can see in Figure 18, although there is a general upwards trend in the data, no other significant note can be deducted from this figure.

Next, we move on to the profitability of movies based on their budget. All the movies with profit ratios of 20x and higher have had a budget in the zero to $30 Million range. On the other hand, the highest budget of a movie with more than 50x profit ratio belongs to "The Exorcist", with an $8 Million budget. The rest of the movies that had a profit ratio of 0 to 20 are shown in the Figure 17. This shows an upward trend in this category as well.

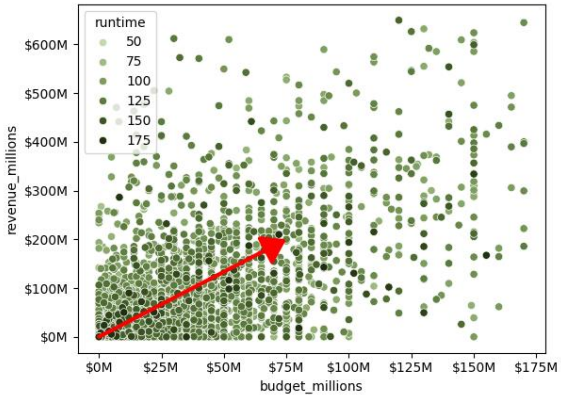


Figure 18: budget of movies with respect to their revenue. (colored by run time)

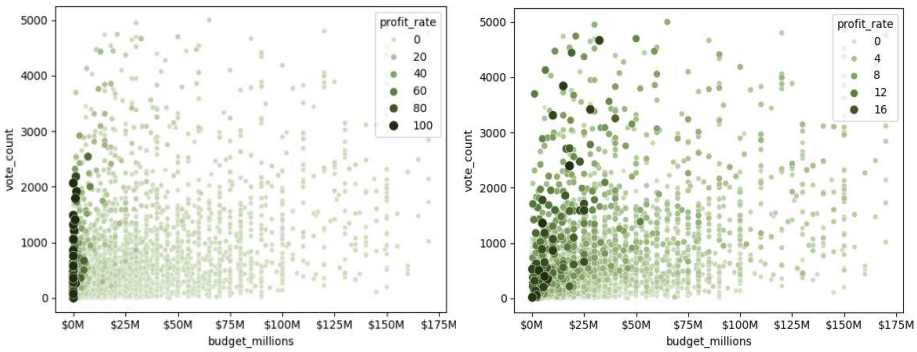


Figure 17: scatter plots of the number of votes with respect to the budget of the movie.

It's also worth noting that %47 of all movies in this dataset have lost money, since their revenue is lower than their budget. From Figure 19, we can see that most of these movies have a popularity below 60, while most of them are in the 0 to 30 range. Figure 19 shows the scarcity of high rated movies among these. We can also detect from both of these plots, that these movies have mostly had budgets of lower than $100 Million.

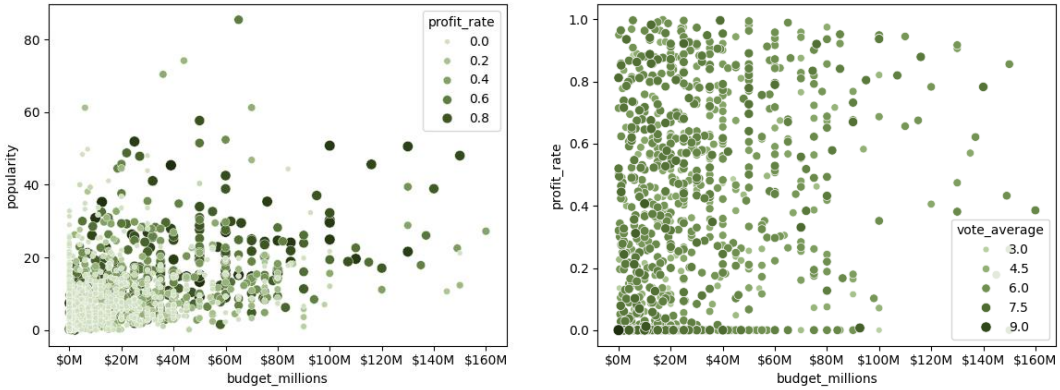


Figure 19: Left plot is the budget of money-losing movies with respect to their popularity score. Right plot is the profit ratio of these movies with respect to their budget.

RUNTIME

In this subsection we will briefly look at the popularity of movies with respect to their rating. Figure 20 shows a scatter plot of these metrics for movies longer and shorter than 90 minutes. As we can see, there aren't any movies with ratings higher than 8.3, while both groups have a higher density in the 5 to 7 rating range.

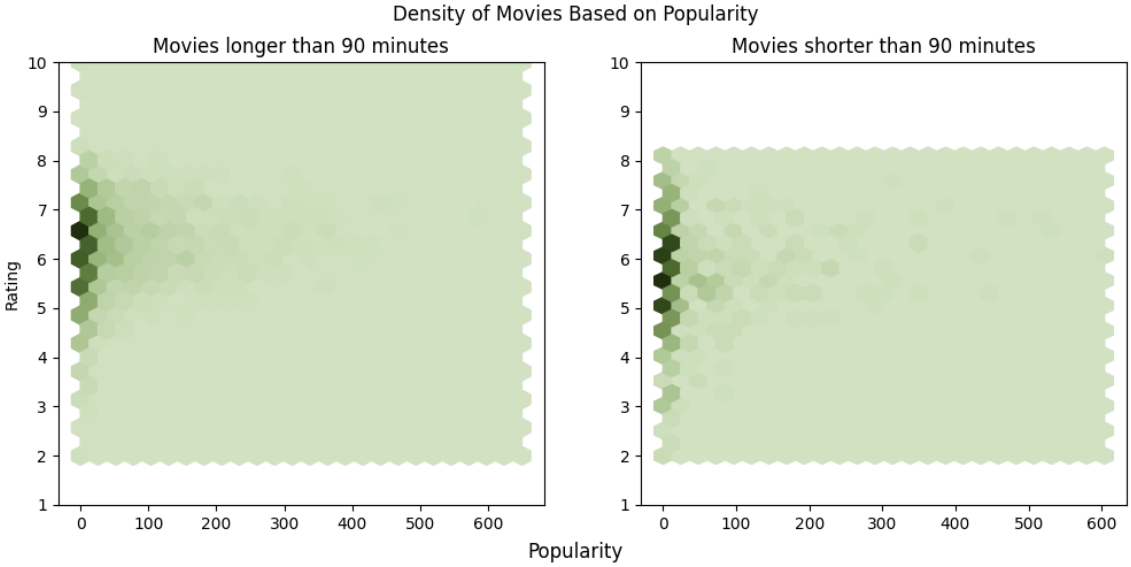


Figure 20: Plot of movie ratings and popularity based on their runtime.

RELEASE YEAR

In this section we take a look at the year of release for each movie. First, we infer from

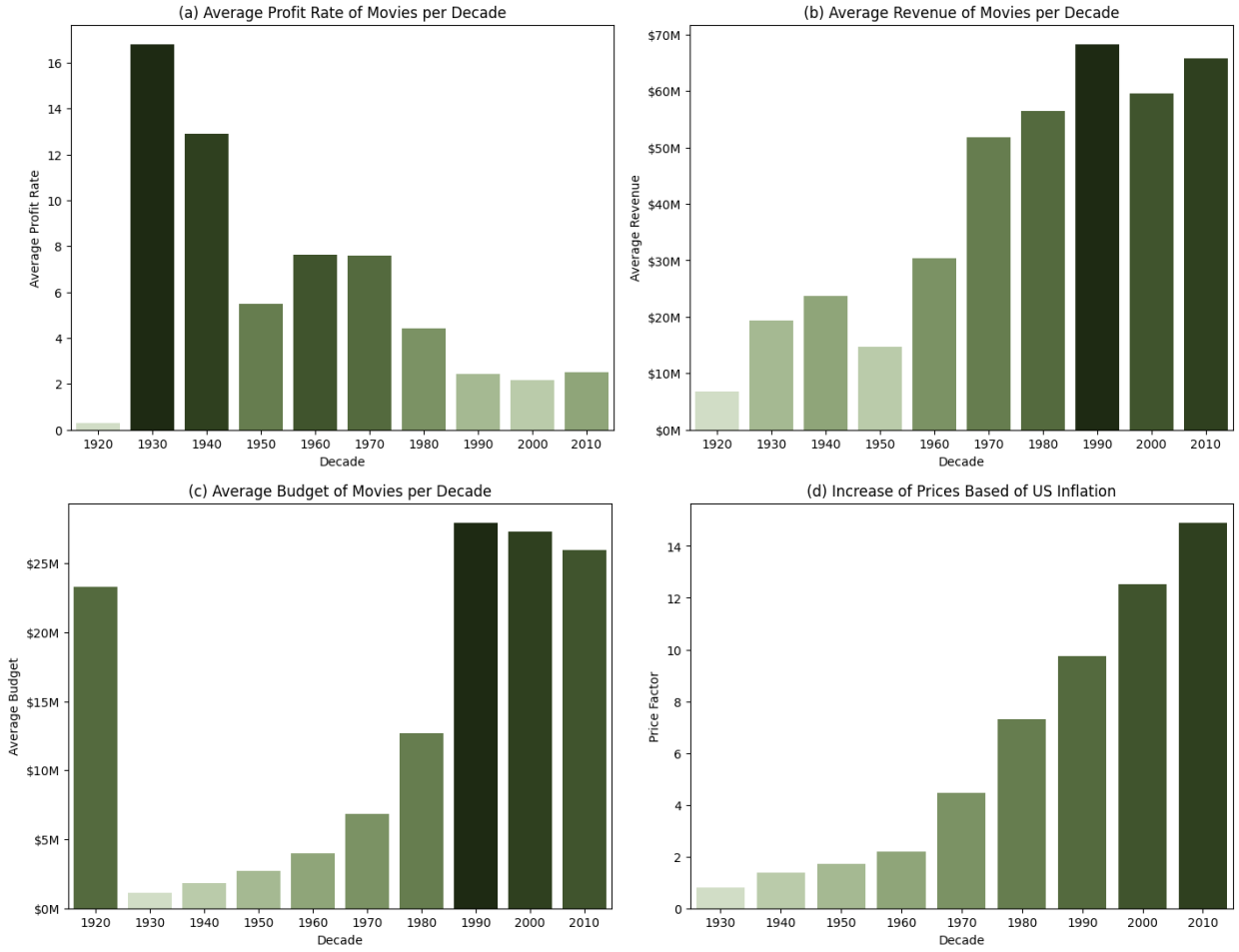


Figure 21.a that the profitability of movies has been on a downward trend since the 1930s. Although the '30s and 40's had an average profit ratio of 16.8x and 12.9x respectively, we observe that the 1920's was not so profitable. In fact, the '20s had a profit ratio of 0.29x, meaning most movies lost money on production.

To investigate this almost monotonic downward trend, we plot the budget and revenue of movies in the same time frame in

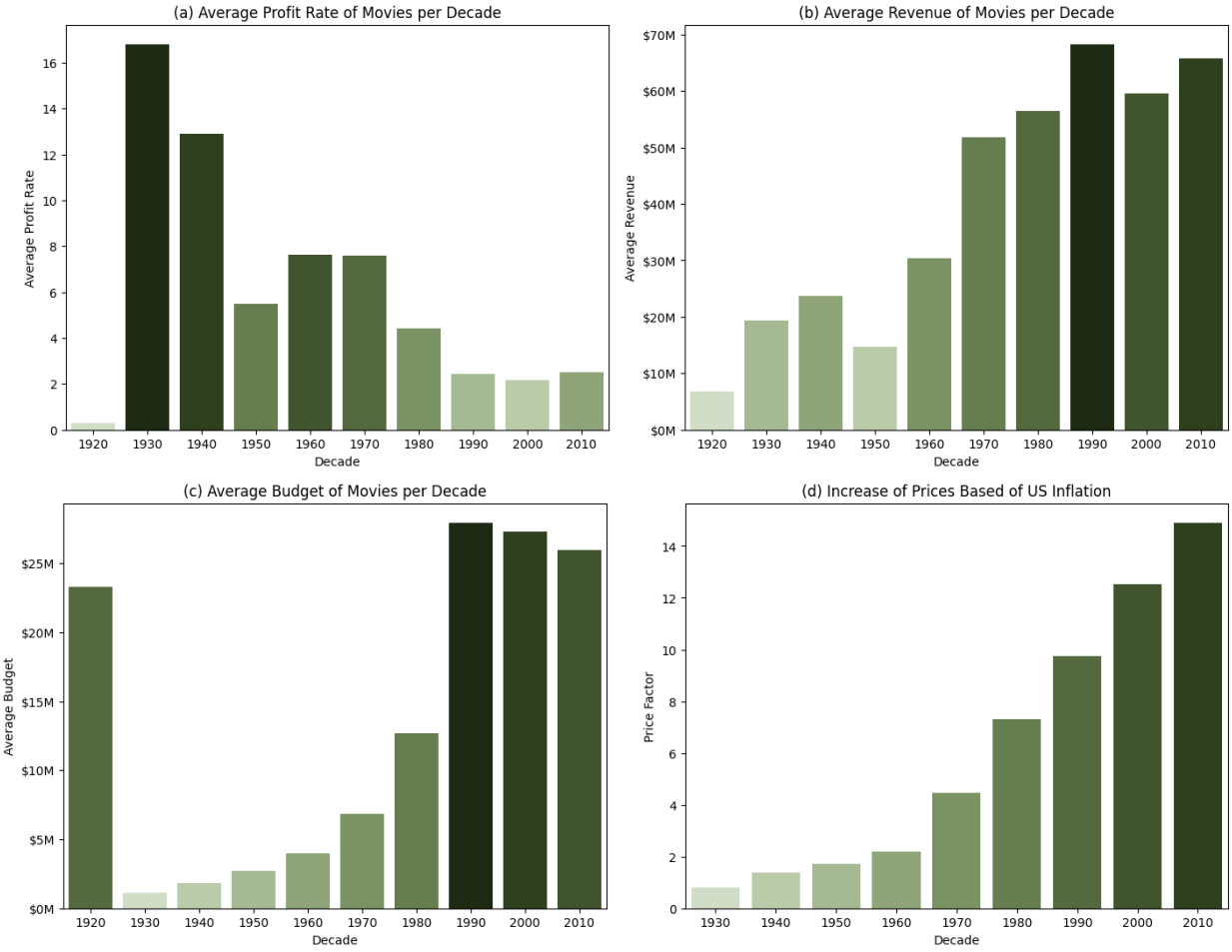


Figure 21.b and

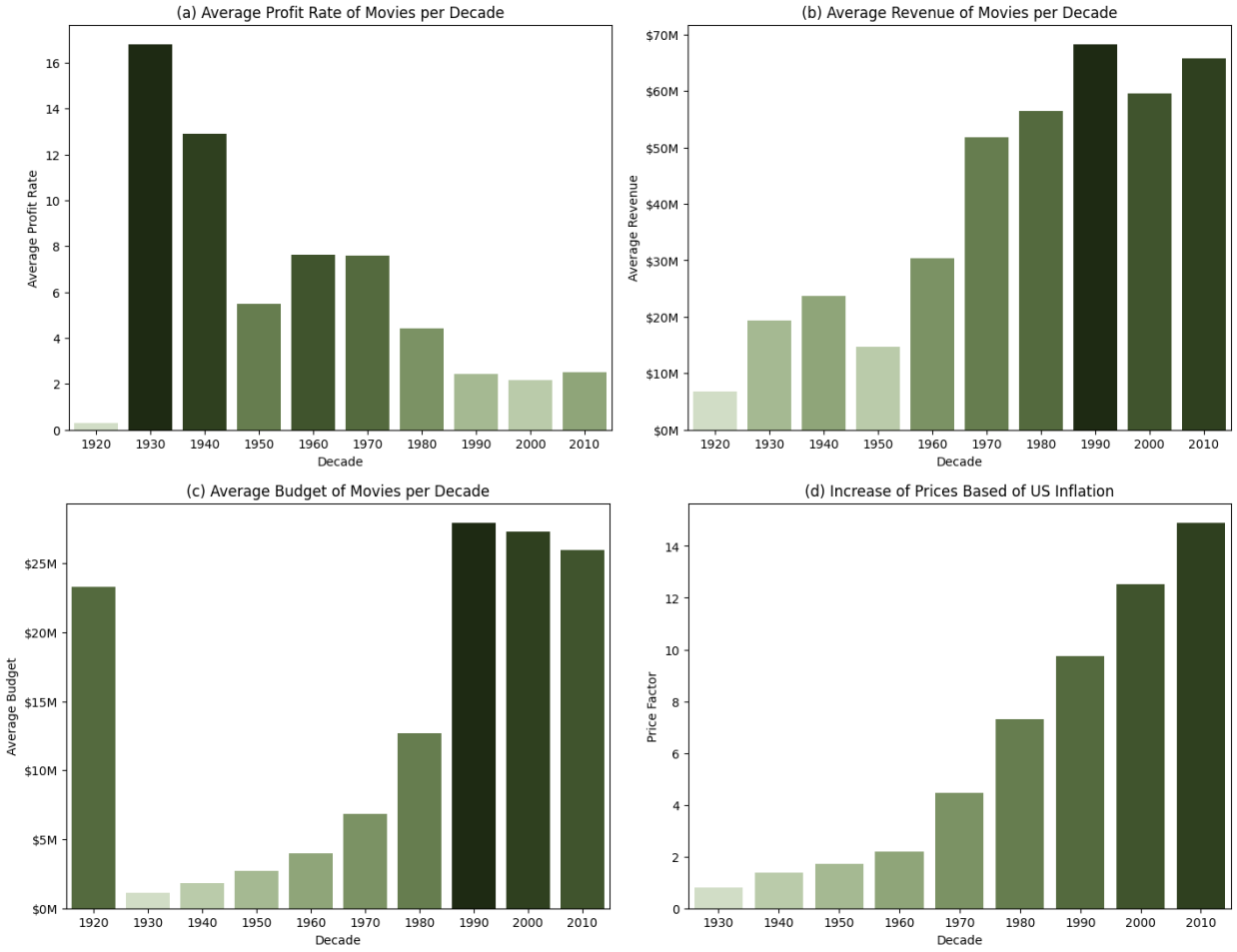


Figure 21.c. The revenue of movies has increased by a factor of 3.4x from the 1930s to the 2010s. Although this increase is significant, it is incomparable to the increase in the budgetary needs of movies. In the same span of time, the average budget of movie has increased from $1.15 Million to almost $26 Million, which indicates a 22.6x increase in the budgetary needs of movies.

For Further Analysis, we compare the revenue increase with the US cumulative inflation ratio. As we can see in

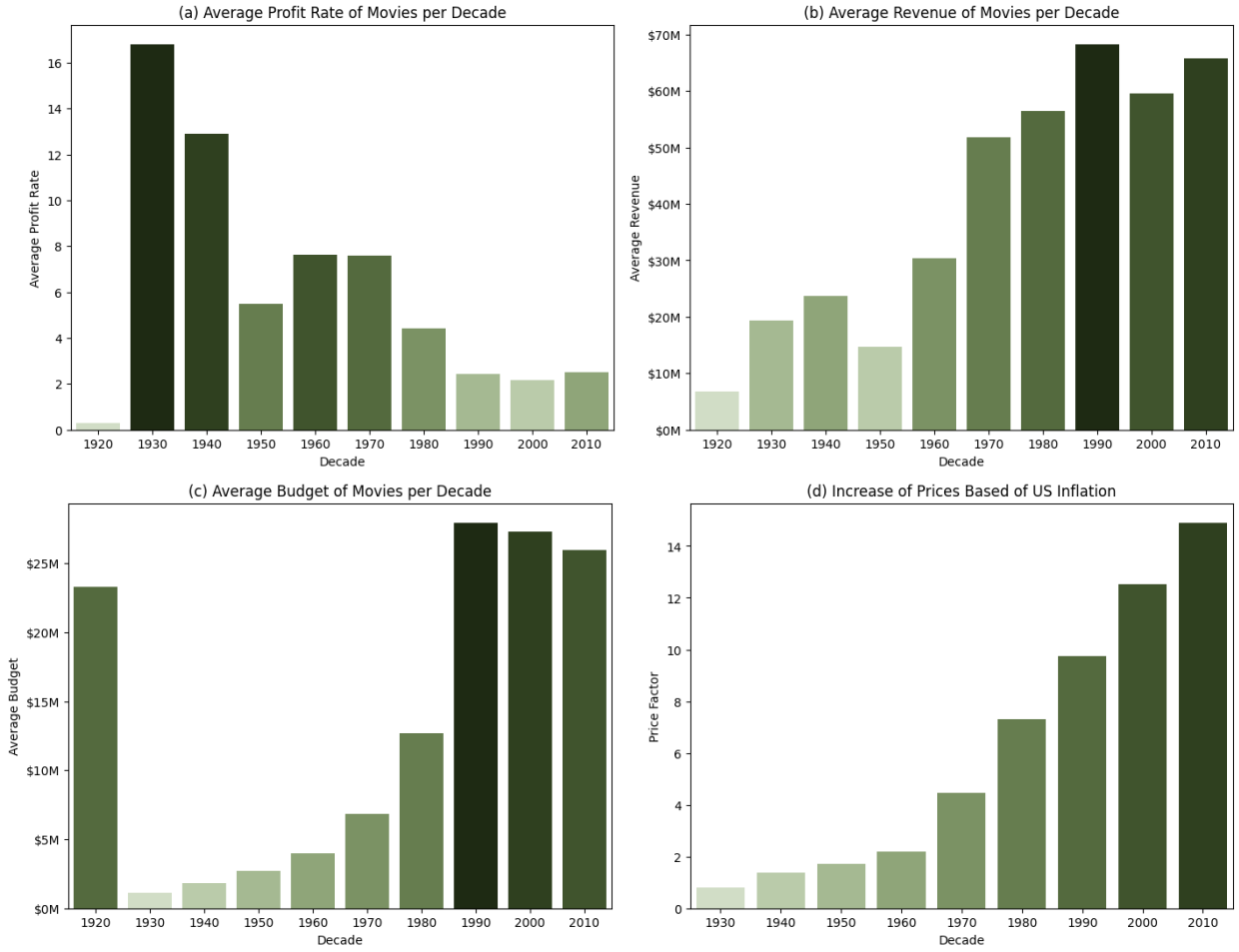


Figure 21.d, the 3.4x increase of revenue does not even keep pace with the almost 15x increase that is cause by inflation.

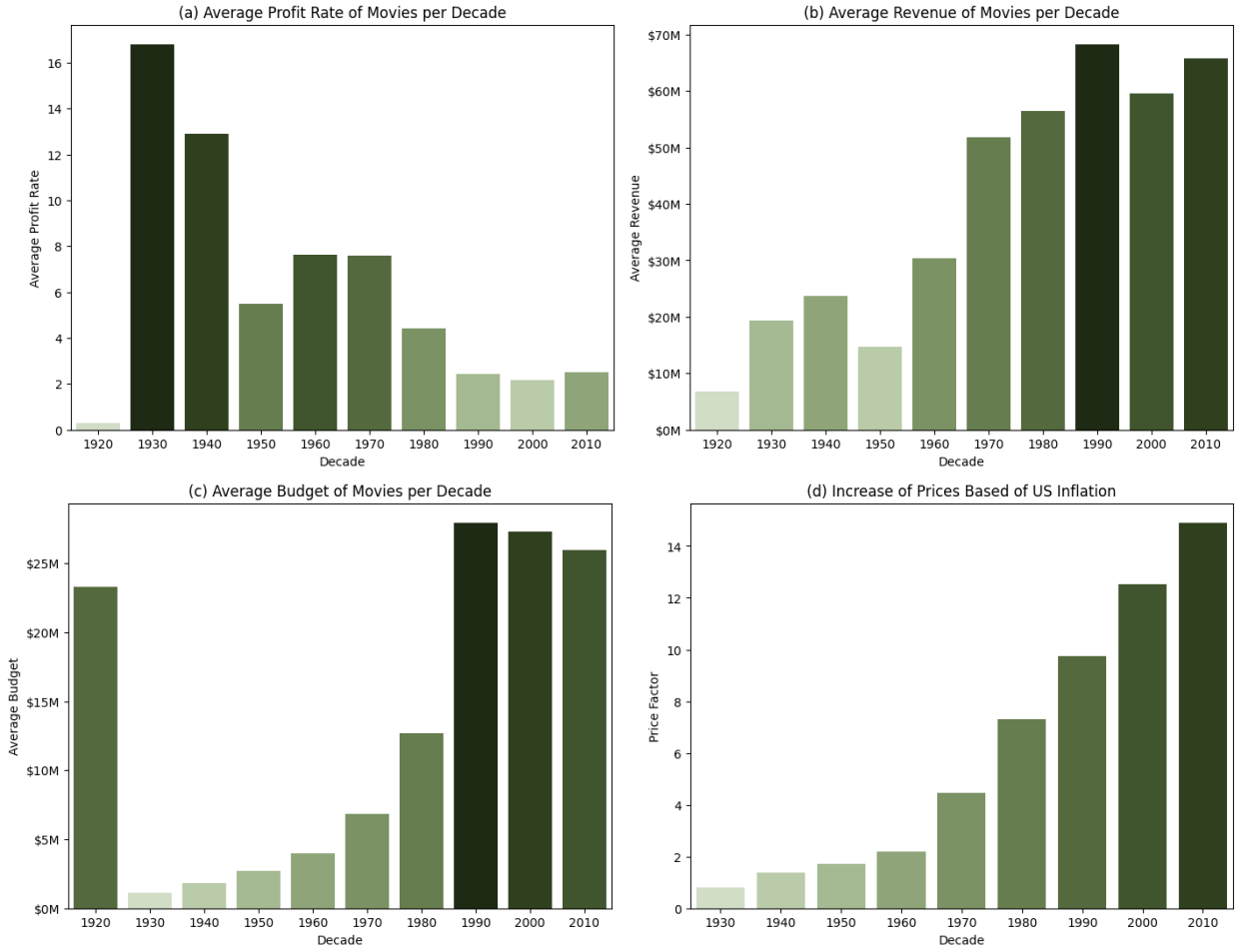


Figure 21: Bar plots of average profits, budgets, revenues, and inflation by decade.

CAST & PRODUCTION

Finally, we look at the top actors and production companies involved in making different movies.

PRODUCTION COMPANIES

The top 20 production companies, based on the accumulated revenue of the movies in which they were the top production company, have been shown in Figure 22.

CAST

The top 20 actors, based on the accumulated revenue of the movies in which they acted as one of the top 5 roles, have been shown in Figure 23.

In addition, as an interesting trivia, a plot of most paired actors has been shown in Figure 24.

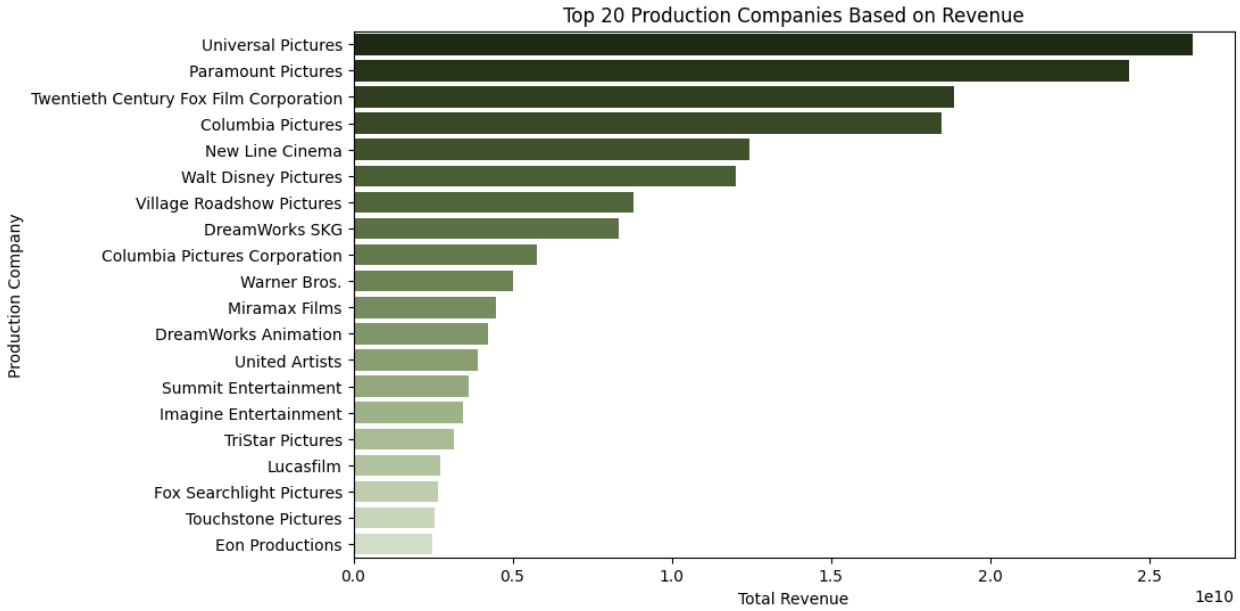


Figure 22: Bar plot of the top 20 production companies.

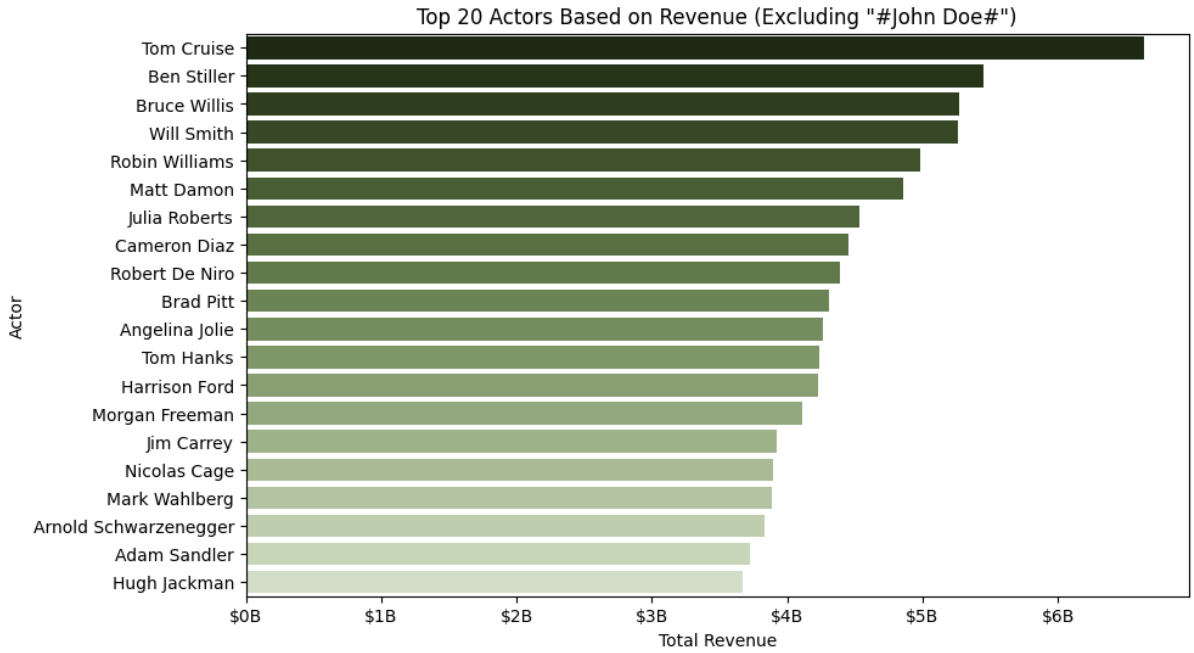


Figure 23: Bar plot of top 20 actors.

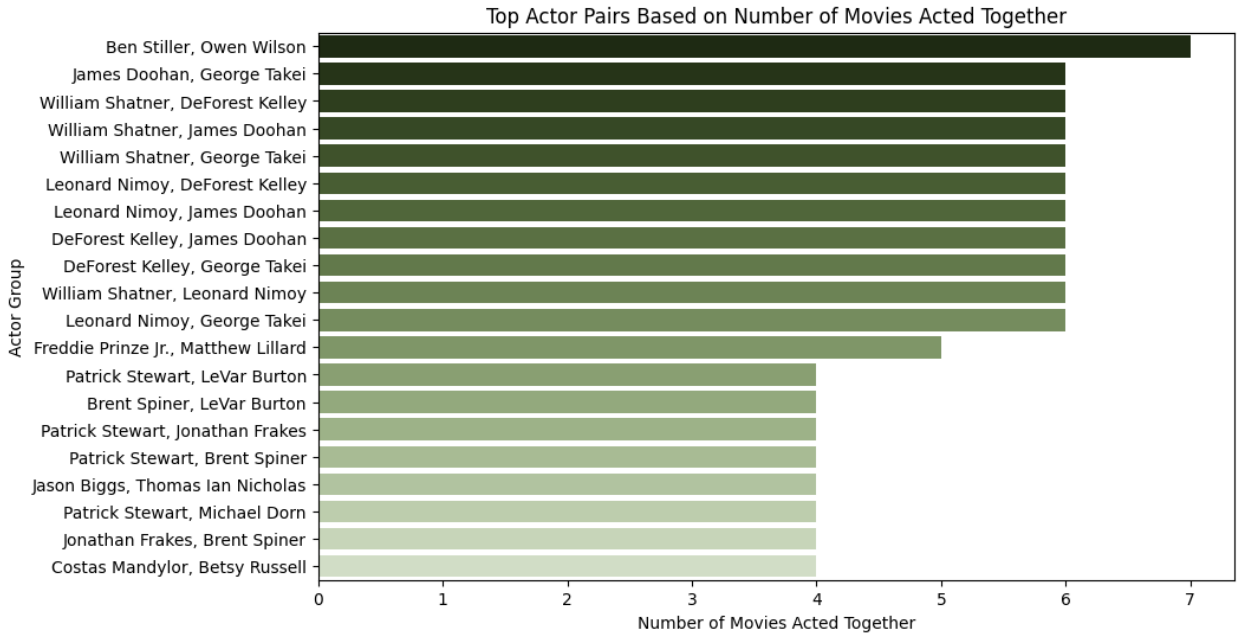


Figure 24: Bar plot of the top acting pairs.

PREDICTIVE MODELING

In this section, predictive modeling approaches are performed on the dataset, and evaluations are presented and plotted for a more detailed understanding.

PREPARING DATA

In this section, we perform feature standardization, feature extraction, and feature selection on our data. At last, we reduce the size of the dataset to 15 features. We also create a Boolean target feature representing the profitability of each movie.

NOTE: All of the scaling, normalization, feature extraction, and feature selection models are saved locally for future use. These models can be accessed through the `path/to/project/models` path where the path to the project folder should be replaced based on your project location.

DATA STANDARDIZATION

Features of the dataset are divided into two categories: Numerical Features, and Label-encoded Features. We rescale label-encoded features, so that all the classes of each feature are between 0 and 1. We then standardize the numerical features to have an average of zero, and a standard deviation of 1.

FEATURE EXTRACTION

Feature extraction is only performed on genre features. This process is done by utilizing the PCA (Principal Component Analysis) technique on these features. We opted to use the top 5 components of these features.

FEATURE SELECTION

Finally, we use feature selection techniques to pick the 10 remaining features of our analysis. For this purpose, we use the `f_classif` metric along with the KBest choosing technique to select the top 10 features of our dataset. The selected features are listed in Table 2. The f_classif method uses the ration between the variation between sample means and the variation withing the samples in different groups of features to find the most significant group of features for the prediction of the target.

5th Main Cast Member
Budget
Language
Number of Votes
Month
Popularity
Production Company
Run Time
Rating
Year of Release

Table 2: Top 10 numerical features for profitability.

MODELING

We have opted to use the following four models for this dataset:

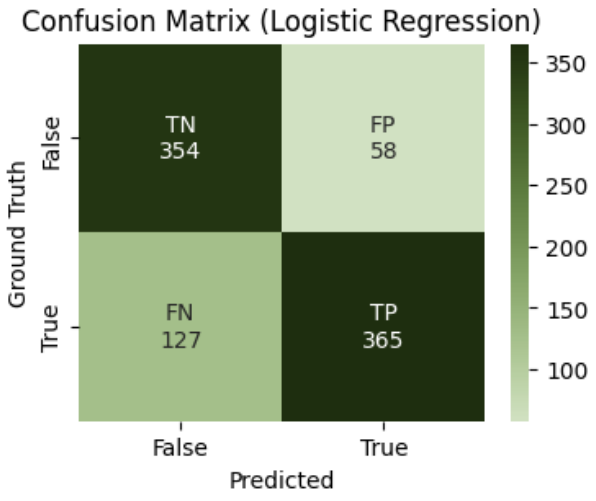
1. Logistic Regression
2. Decision Tree
3. Random Forest
4. XGBoost

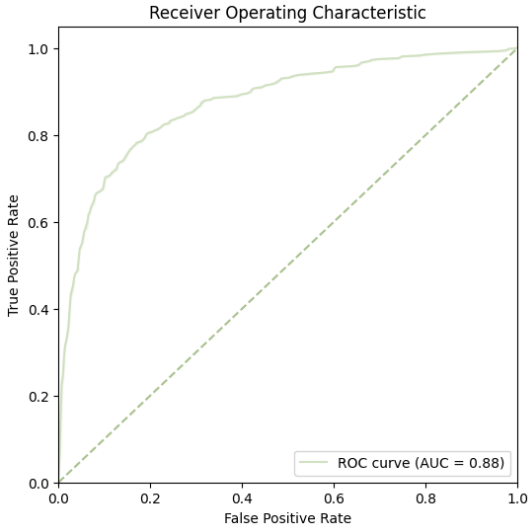
For each model, a report on the performance of the model is provided. Moreover, we plot the confusion matrix and the ROC-AUC of each model for an intuitive understanding of the performance of each model.

NOTE: All of the models are saved locally for future use. These models can be accessed through the path given below where the path to the project folder should be replaced based on your project location: `path/to/project/models`

LOGISTIC REGRESSION

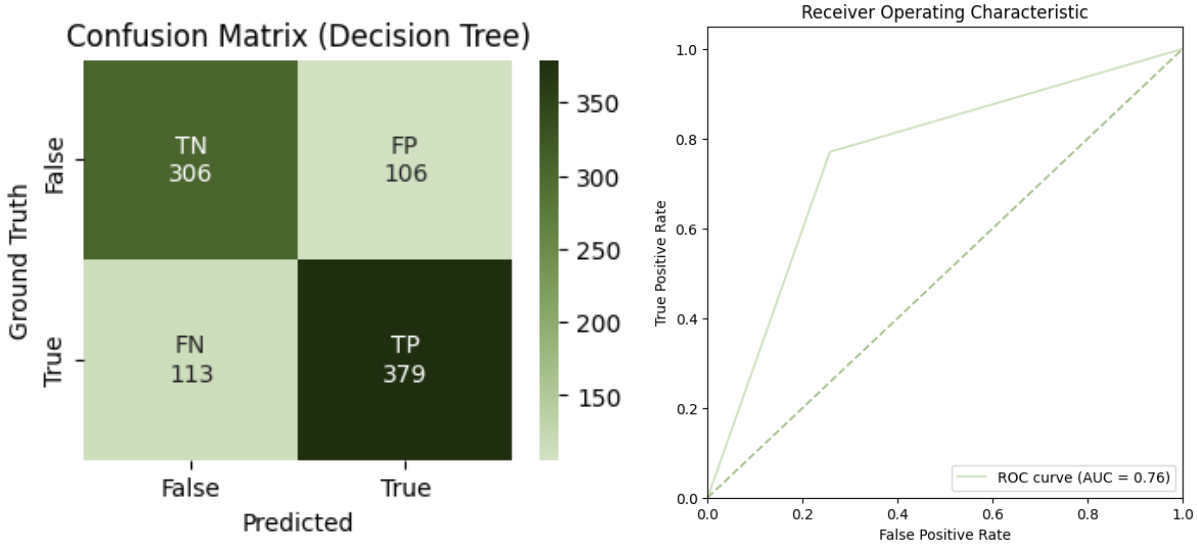
LOGISTIC REGRESSION	Precision	Recall	F1-Score	Support
Not Profitable	0.73	0.74	0.74	412
Profitable	0.78	0.77	0.78	492
Weighted Average	0.76	0.76	0,76	904





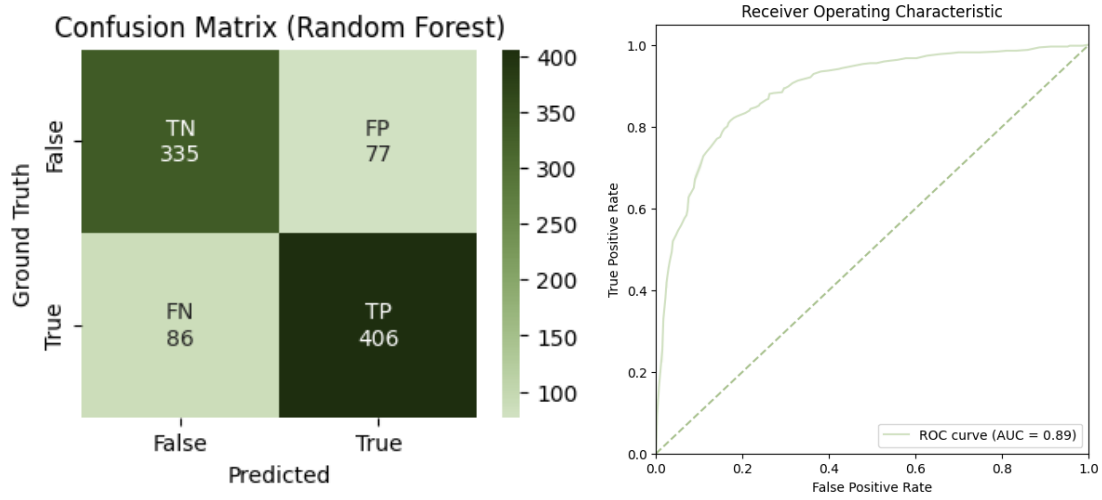
DECISION TREE

DECISION TREE	Precision	Recall	F1-Score	Support
Not Profitable	0.73	0.74	0.74	412
Profitable	0.78	0.77	0.78	492
Weighted Average	0.76	0.76	0.76	904



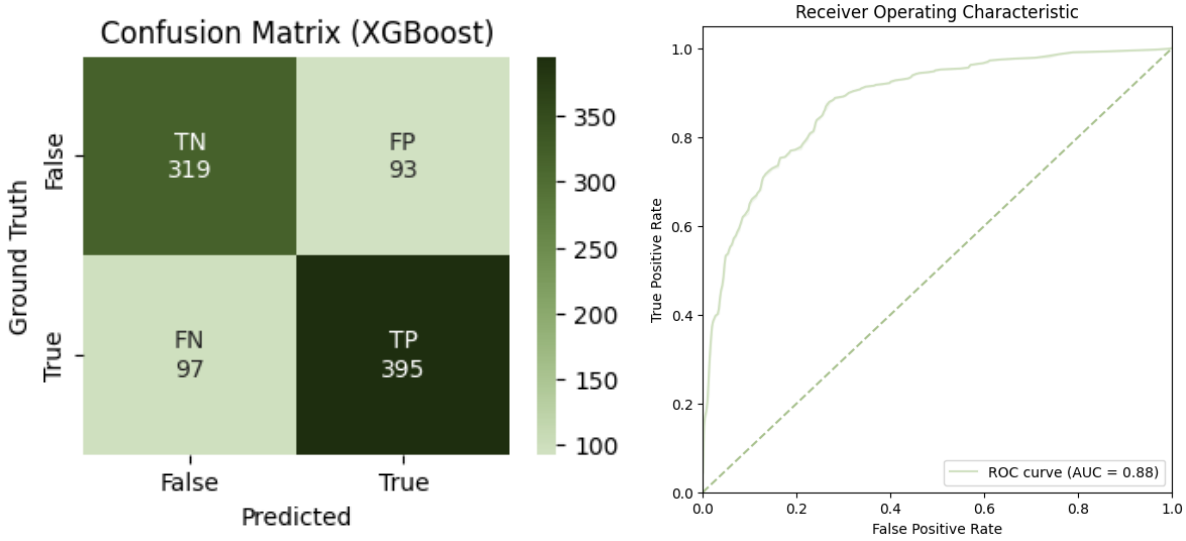
RANDOM FOREST

RANDOM FOREST	Precision	Recall	F1-Score	Support
Not Profitable	0.80	0.81	0.80	412
Profitable	0.84	0.83	0.83	492
Weighted Average	0.82	0.82	0.82	904



XGBOOST

XGBOOST	Precision	Recall	F1-Score	Support
Not Profitable	0.77	0.77	0.77	412
Profitable	0.81	0.80	0.81	492
Weighted Average	0.79	0.79	0.79	904



ANALYSIS & REPORT

Each model's performance is evaluated using various metrics such as precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC). Additionally, the confusion matrix for each model is also provided to give an intuitive understanding of the model's performance.

The first model analyzed is Logistic Regression. The performance metrics indicate that this model has a precision, recall, and F1-score of 0.76 for the weighted average. The confusion matrix reveals that the model correctly predicted 354 non-profitable and 365 profitable instances. However, it misclassified 58 profitable and 127 non-profitable instances. The AUC score is 0.88, which suggests a good balance between sensitivity and specificity.

Next, the Decision Tree model is examined. This model also has a precision, recall, and F1-score of 0.76 for the weighted average. The confusion matrix shows that the model correctly predicted 309 non-profitable and 378 profitable instances. However, it misclassified 103 profitable and 114 non-profitable instances. The AUC score is also 0.88, similar to the Logistic Regression model.

The third model, Random Forest, shows a higher precision, recall, and F1-score of 0.82 for the weighted average. The confusion matrix shows that the model correctly predicted 329 non-profitable and 406 profitable instances. However, it misclassified 83 profitable and 86 non-profitable instances. The AUC score is 0.88, similar to the previous two models.

Lastly, the XGBoost model has a precision, recall, and F1-score of 0.79 for the weighted average. The confusion matrix shows that the model correctly predicted 329 non-profitable and 396 profitable instances. However, it misclassified 83 profitable and 96 non-profitable instances. The AUC score is 0.88, similar to the other models.

In conclusion, all four models have similar AUC scores, but the Random Forest model has the highest precision, recall, and F1-score, making it the best performing model among the four according to this report. However, the choice of the best model can also depend on the specific requirements of your project. For example, if minimizing false positives is a priority, you might prefer a model with higher precision. If minimizing false negatives is more important, a model with higher recall might be preferable. It's also worth noting that these models are saved locally for future use at the given path. This allows for easy access and application of the models in future tasks. This approach not only saves time but also ensures consistency in the application of the models.