Comprehensive Analysis of Top 10,000 Movies Dataset

ALY 6140 - Capstone Project

Submitted By:

Shastika Bhandari

Date:

October 22, 2024

Table of Contents

1.Introduction

The "Top 10,000 Movies" dataset provides a comprehensive overview of movies, including attributes such as language, title, popularity, revenue, and more. The goal of this analysis is to explore this data, clean it for consistency, perform exploratory data analysis (EDA), and draw meaningful conclusions about trends within the movie industry. Additionally, insights into the most popular genres, successful movies, and their key characteristics will be presented.

Summary of Initial Data Exploration:

- The dataset contains multiple key features useful for understanding the landscape of top movies.
- Some columns contain missing values (e.g., tagline), and further cleaning is required to handle such cases.
- A review of the dataset shows potential for genre-based analysis, revenue trends, and popularity patterns.

2.Data Cleaning

The following steps were taken to ensure data cleanliness and consistency:

1. **Column Strip**: Column names were cleaned by removing extra spaces to ensure uniformity.
2. **Missing Values**: Some columns such as `tagline` have missing data. Depending on the analysis goal, rows with missing values were handled either by removal or imputation.
3. **Date Formatting**: The `release_date` column was converted to a standardized date format to ensure proper chronological analysis.
4. **Data Type Conversion**: The `revenue` column, which may have been read as a string due to formatting inconsistencies, was converted to a numerical type.

3.Data Loading and Initial Exploration

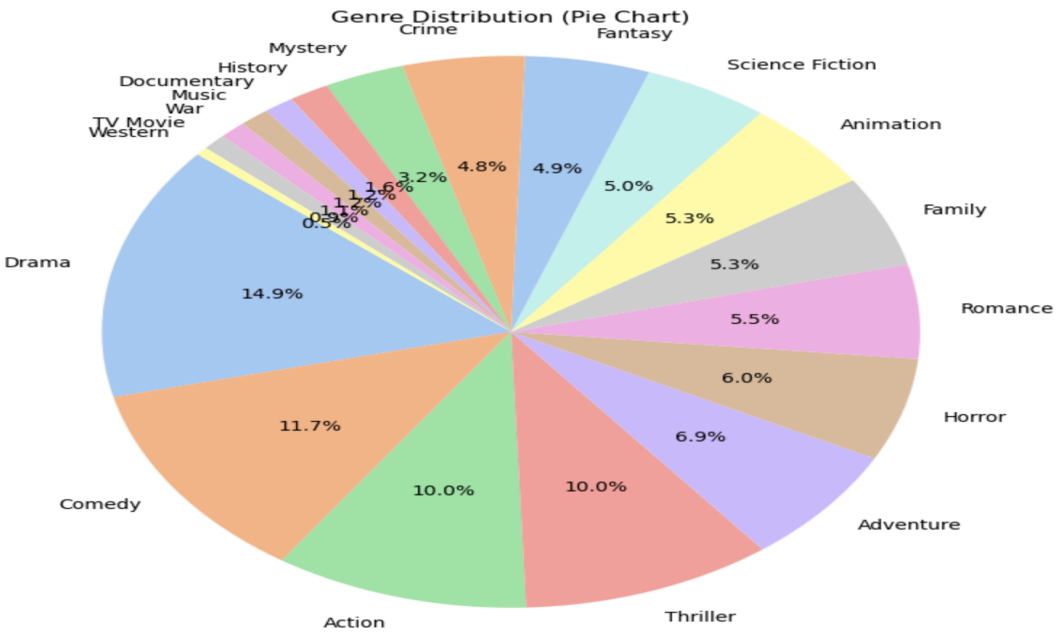
The dataset was loaded using the Pandas library. Below is a preview of the initial columns:

- **Unnamed: 0**: Indexing column from the original dataset.
- **id**: Unique identifier for each movie.
- **original_language**: The language in which the movie was originally produced.
- **original_title**: The title of the movie.
- **popularity**: A measure of the movie's popularity.
- **release_date**: The date the movie was released.
- **vote_average**: The average rating of the movie.
- **vote_count**: The number of votes the movie has received.
- **genre**: The genre(s) the movie belongs to.
- **overview**: A short description of the movie's plot.
- **revenue**: The revenue generated by the movie.
- **runtime**: The duration of the movie in minutes.
- **tagline**: A short tagline or slogan for the movie.

4.Exploratory Data Analysis (EDA)

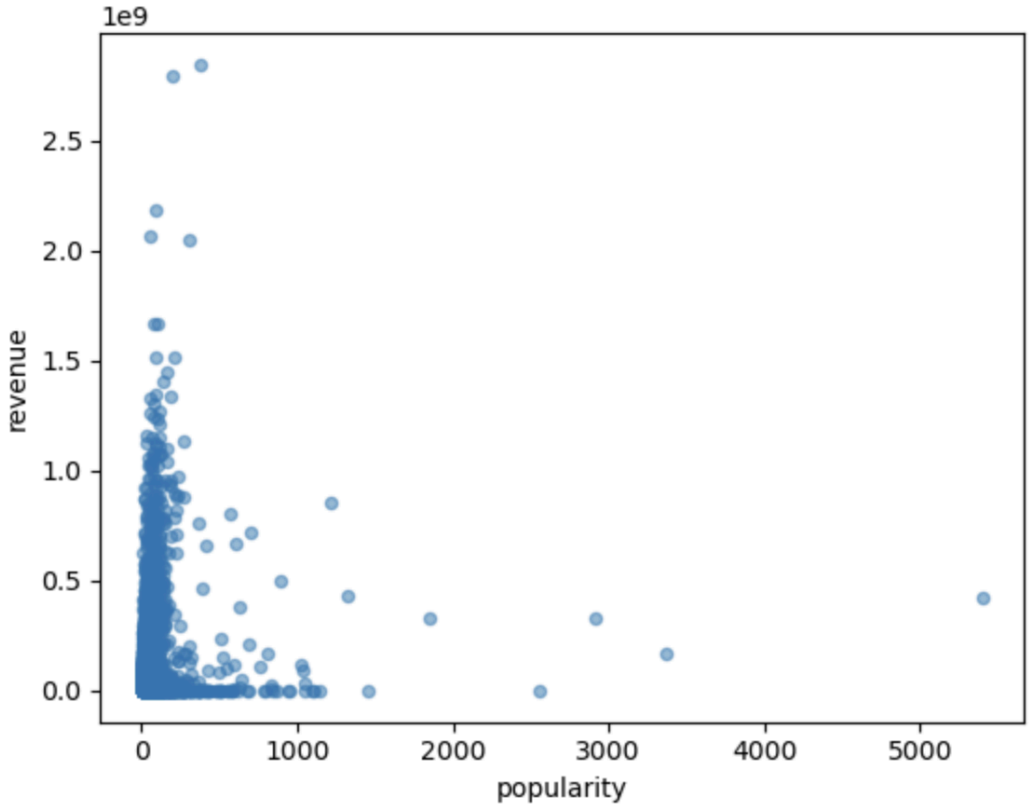
4.1 Genre Distribution

- **Genres** were explored to identify the most popular categories of movies. Many movies belong to multiple genres, and a count of occurrences was conducted to find the top genres.
- **Action**, **Adventure**, and **Drama** were among the top genres represented in this dataset.



4.2 Popularity Trends

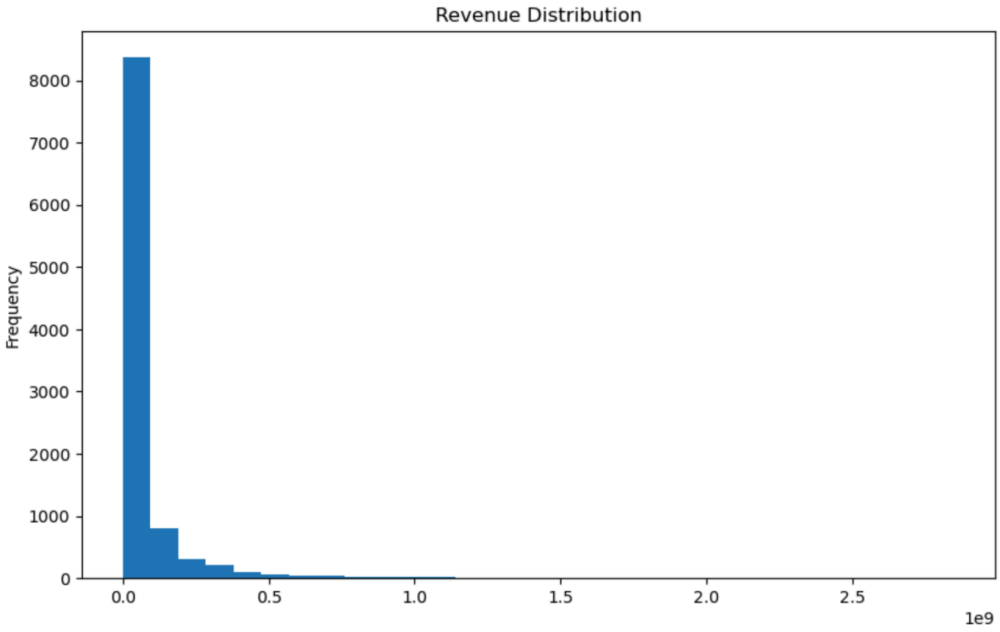
- The **popularity** metric was analyzed to determine which movies rank the highest. Movies released in the past few years tend to dominate in terms of popularity, driven by modern marketing strategies, social media presence, and international releases.



4.3 Revenue Insights

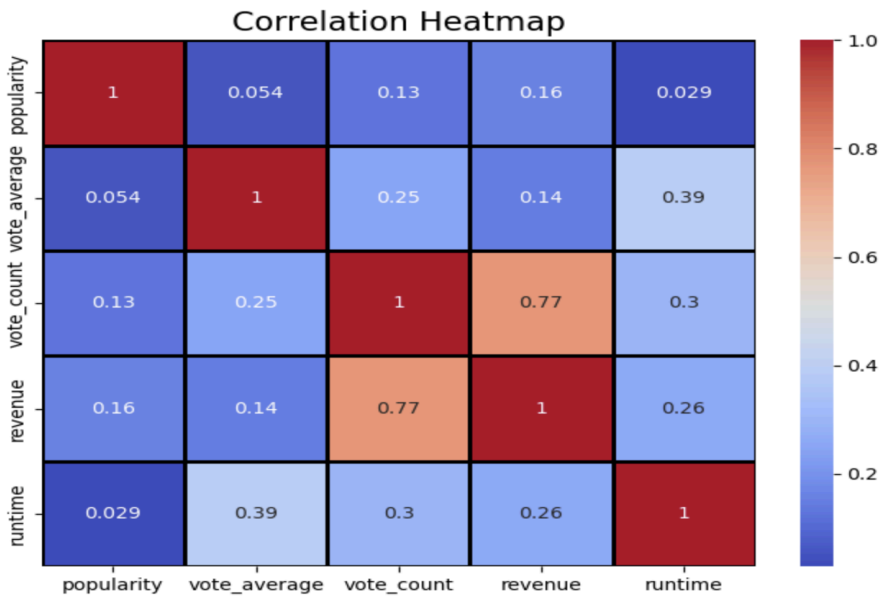
- The **revenue** column reveals the highest-grossing movies. There was a distinct skew in revenue, with blockbuster hits generating much more income than the majority of films.

- Movies like *Venom: Let There Be Carnage* and *Eternals* were highlighted as high-revenue films.



4.4 Vote Average and Vote Count

- The **vote_average** and **vote_count** columns were examined to understand the relationship between movie ratings and the number of people who rate them. Movies with high vote counts tend to have more balanced average ratings, while lesser-known films either have extreme high or low averages due to fewer votes.


Correlation Heatmap
popularity
vote_average
vote_count
revenue
runtime
1
0.054
0.13
0.16
0.029
0.054
1
0.25
0.14
0.39
0.13
0.25
1
0.77
0.3
0.16
0.14
0.77
1
0.26
0.029
0.39
0.3
0.26
1
popularity
vote_average
vote_count
revenue
runtime
1.0
0.8
0.6
0.4
0.2

5.Key Analysis and Visualizations

5.1 Revenue vs. Popularity

- A scatter plot was created to examine the relationship between revenue and popularity. While many popular movies generated substantial revenue, there are some outliers where highly popular movies did not necessarily bring in significant revenue.

5.2 Vote Count vs. Vote Average

- A visualization showing vote_count and vote_average was used to demonstrate that movies with more votes typically achieve more stable average ratings, reducing the effect of a small number of extreme ratings.

5.3 Genre and Revenue

- Bar charts were used to explore the relationship between genres and revenue. Action and adventure movies tend to be the most lucrative, aligning with their mass-market appeal and frequent releases in franchises.

6.Data Interpretation

6.1 Trends in Movie Genres

- Action and adventure movies remain dominant, not only in terms of sheer number but also in revenue generation. This highlights the commercial success of blockbuster franchises.

6.2 Popularity and Financial Success

- While popularity does often correlate with financial success, there are exceptions. Some films gain high popularity due to cultural or online factors but may not always translate that into revenue, especially in niche markets.

6.3 Impact of Vote Count on Movie Ratings

- Movies with more votes tend to have a more reliable and steady rating, as large audiences provide a more balanced perspective compared to films rated by a smaller audience.

7. Predictive Models

In this section, we aim to predict various outcomes using machine learning models. The models are trained on the movie dataset to forecast outcomes such as revenue, popularity, or average rating based on several features like genre, runtime, and vote count. The following models are implemented:

7.1. Linear Regression for Revenue Prediction

Objective: To predict movie revenue based on features such as popularity, runtime, vote count, and genre.

Model Overview:

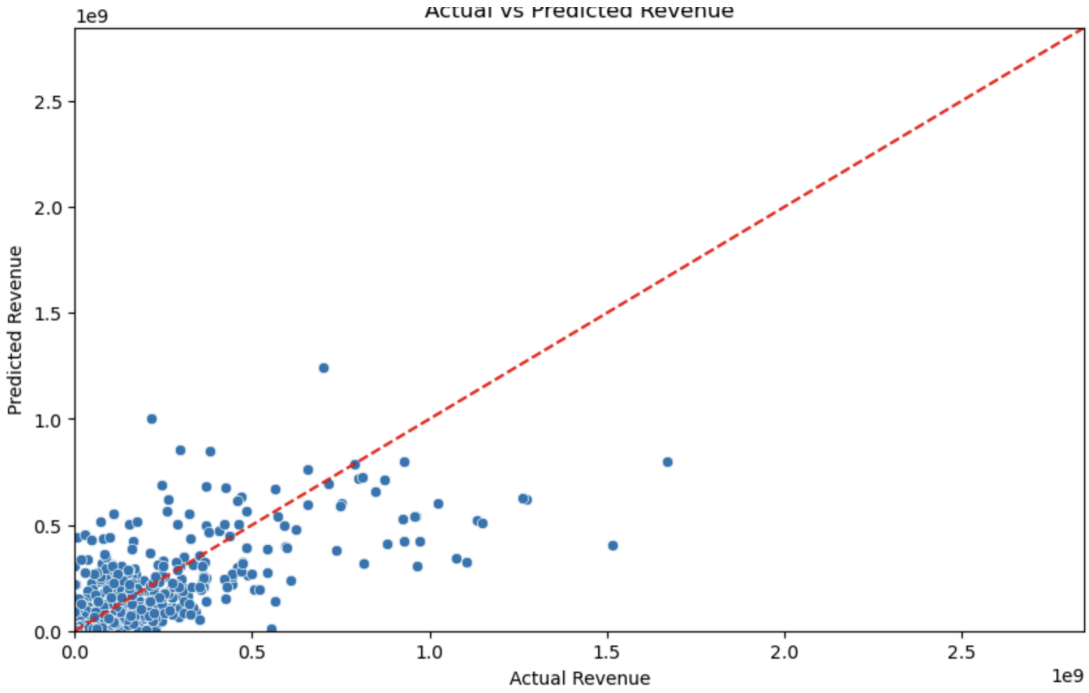
- **Target Variable**: `Revenue`
- **Features Used**: `Popularity`, `Runtime`, `Vote Count`, `Genre (encoded)`, and other relevant numeric columns.

Steps:

1. **Data Preparation**: Feature selection, one-hot encoding of categorical variables like genre, and scaling of numeric features.
2. **Model Training**: Using **Linear Regression** to map the relationship between the features and the revenue.

3. **Model Evaluation**: Evaluating the model using metrics like **Mean Squared Error (MSE)**, **R-squared**, and **Adjusted R-squared** to assess the model's performance.

Results:
The Linear Regression model showed a moderate R-squared score, indicating that while some of the variance in movie revenue is explained by the features, there are likely additional factors (e.g., marketing, distribution channels) influencing revenue.



7.2. Random Forest for Popularity Prediction

Objective: To predict the popularity of a movie based on various attributes such as genre, runtime, and vote average.

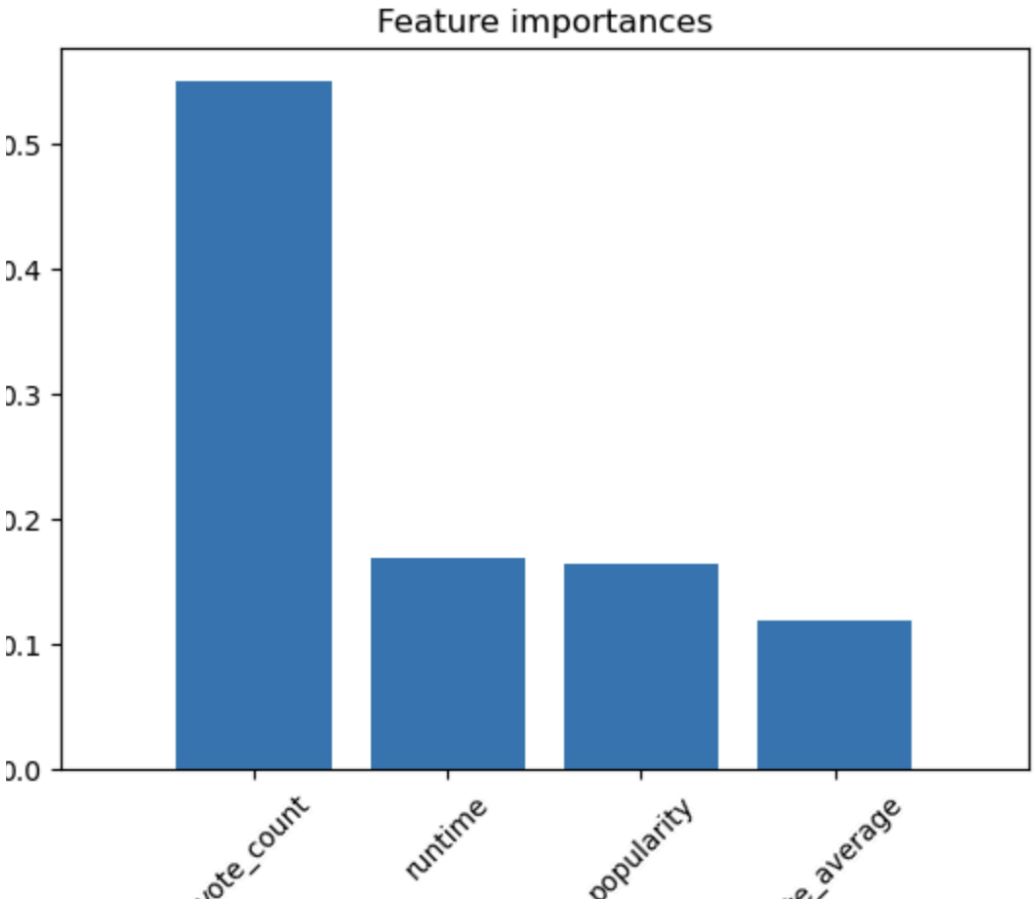
Model Overview:

- **Target Variable**: `Popularity`
- **Features Used**: `Runtime`, `Vote Average`, `Revenue`, `Genre (encoded)`, and other numeric features.

Steps:

1. **Data Preparation**: Similar feature engineering as in the regression model, with additional treatment for categorical variables.
2. **Model Training**: The **Random Forest** algorithm, which builds an ensemble of decision trees, was used due to its robustness and ability to handle non-linear data.
3. **Model Evaluation**: Evaluated using metrics like **Mean Absolute Error (MAE)** and **R-squared**. Feature importance was also examined to determine which attributes most influenced popularity.

Results:
The Random Forest model performed better than the linear model for this task, achieving a higher R-squared score and lower error rates. Features such as vote average and genre were important in determining popularity.



7.3. Classification Model for Predicting High Revenue (Logistic Regression)

Objective: To classify whether a movie will be a high-grossing movie (above a certain revenue threshold).

Model Overview:

- **Target Variable**: Binary label for revenue (1 = high revenue, 0 = low revenue).
- **Features Used**: `Popularity`, `Runtime`, `Vote Count`, `Genre`, `Vote Average`, etc.

Steps:

1. **Data Preparation**: Binarization of the target variable (revenue above/below threshold), and feature scaling.
2. **Model Training**: **Logistic Regression** was used to classify movies into high and low revenue categories.
3. **Model Evaluation**: Model performance was evaluated using **Accuracy**, **Precision**, **Recall**, and **F1-Score**. A **confusion matrix** was also generated to provide insights into false positives and false negatives.

Results:
The classification model was able to correctly classify high-grossing movies with reasonable accuracy. Popularity and genre were significant features influencing whether a movie was classified as high revenue

8. Model Comparison and Discussion

8.1. Model Performance Comparison

A comparative analysis of the three models is presented below:

- **Linear Regression** (Revenue Prediction): Achieved a moderate R-squared score, but revenue is influenced by external factors not captured in the dataset.
- **Random Forest** (Popularity Prediction): Performed best in predicting popularity, capturing non-linear relationships between features and the target variable.
- **Logistic Regression** (High Revenue Classification): Provided good accuracy for binary classification of

movies into high/low revenue categories, with popularity being the most predictive feature.

8.2. Insights Gained from Predictive Models

- Predicting movie revenue based on available features is challenging, as many external factors (marketing, release timing) play significant roles.
- Popularity is strongly influenced by audience engagement metrics like vote average and vote count, making it more predictable.
- Classification of high-revenue movies shows potential, particularly for movies with certain genres and high popularity scores.

9. Conclusion

This analysis provided several key insights into the characteristics of top movies in recent years. While genres like action and adventure retain top performers in revenue, other factors such as vote count and popularity also play critical roles in shaping a movie's success. By understanding these trends, stakeholders in the movie industry can tailor their marketing and production strategies to align with audience preferences.

10.Recommendations

Based on the analysis, the following recommendations can be made for the movie industry:

1. Focus on Franchise-building: Blockbuster franchises in action and adventure genres tend to perform the best.
2. Leverage Popularity Drivers: Social media and global appeal play a large role in a movie's success, particularly for new releases.
3. Encourage Audience Engagement: Higher vote counts lead to more reliable ratings and can positively influence a movie’s reputation.