Title: TMDB 5000 Movie Dataset: EDA, Modeling and Recommendation System

Abstract:

The analysis of the TMDB (The Movie Database) dataset offers valuable insights into various aspects of the movie industry. Through exploratory data analysis (EDA), modeling, and the development of a recommendation system, this report aims to uncover trends, correlations, and patterns within the dataset. By leveraging advanced analytics techniques, the report provides actionable insights and recommendations for stakeholders in the movie industry.

Leatreture Review:

Prior research in the field of movie industry analysis has emphasized the importance of understanding audience preferences, predicting movie profitability, and optimizing marketing strategies. Various studies have utilized machine learning algorithms, exploratory data analysis techniques, and collaborative filtering methods to extract valuable insights from movie datasets. This report contributes to the growing body of knowledge in movie industry analysis by building upon existing literature.

Description of analysis methods:

In this project, the analysis methodology involved several key steps aimed at preparing and analyzing the TMDB (The Movie Database) dataset. The primary focus was understanding the movie industry, predicting movie profitability, and developing a recommendation system. Below is a description of the analysis methods employed:

- Data Merging:
  - The dataset consisted of two main parts: credits and movies. The initial step involved merging these two parts to create a comprehensive dataset for analysis. By combining the information from both parts, we aimed to leverage the full spectrum of attributes available in the dataset.
- Data Cleaning:
  - Handling Null and Duplicate Rows:
    - One of the crucial tasks in data cleaning was addressing null and duplicate rows. Specifically, two features, namely 'homepage' and 'tagline,' exhibited a significant number of null values compared to the total number of rows. Given the importance of these features and the inability to simply delete rows with null values, we adopted the "Imputation with a Placeholder" method.

        - Considering the project's objective of determining movie profitability, and given the presence of essential columns such as 'budget' and 'revenue,' we opted to retain these columns and handle the missing values accordingly.
        - As a result, null values in the 'homepage' and 'tagline' columns were replaced with placeholders, namely "No Homepage" and "No Tagline," respectively.
    - Handling Missing and Outliers Values:
        - Further data cleaning addressed missing values in other features such as 'genres,' 'cast,' 'crew,' 'spoken languages,' 'keywords,' 'production_companies,' and 'production_countries.'
        - Initially, missing values in these features were converted to null values for easier handling. Subsequently, rows with null values were removed from the dataset.
        - This approach ensured that the dataset remained robust and suitable for subsequent analysis and modeling tasks.

Exploratory Data Analysis (EDA):

The next step in the analysis process involved conducting exploratory data analysis (EDA) on the numeric columns of the dataset. This step was divided into two sections:

Analysis of Numeric Columns: Various queries were prepared to analyze the distribution and characteristics of numeric columns.

For instance:

Q1: Check the distribution of the budget.

Resul1t 1: Most movies were built with almost 1 million dollars.

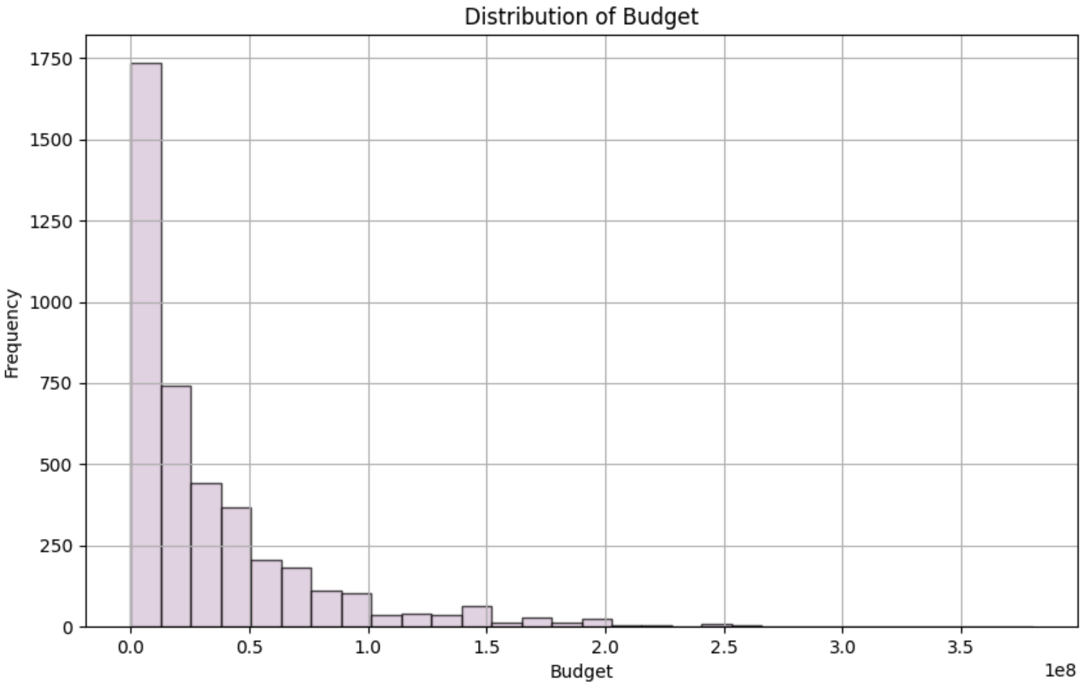

Figure1. Distribution of Budget

Q2: Plot the chart of the budget for the top 10 movies based on popularity.

Result 2: The Minions was the most popular movie with a budget of less than 80 million dollars.

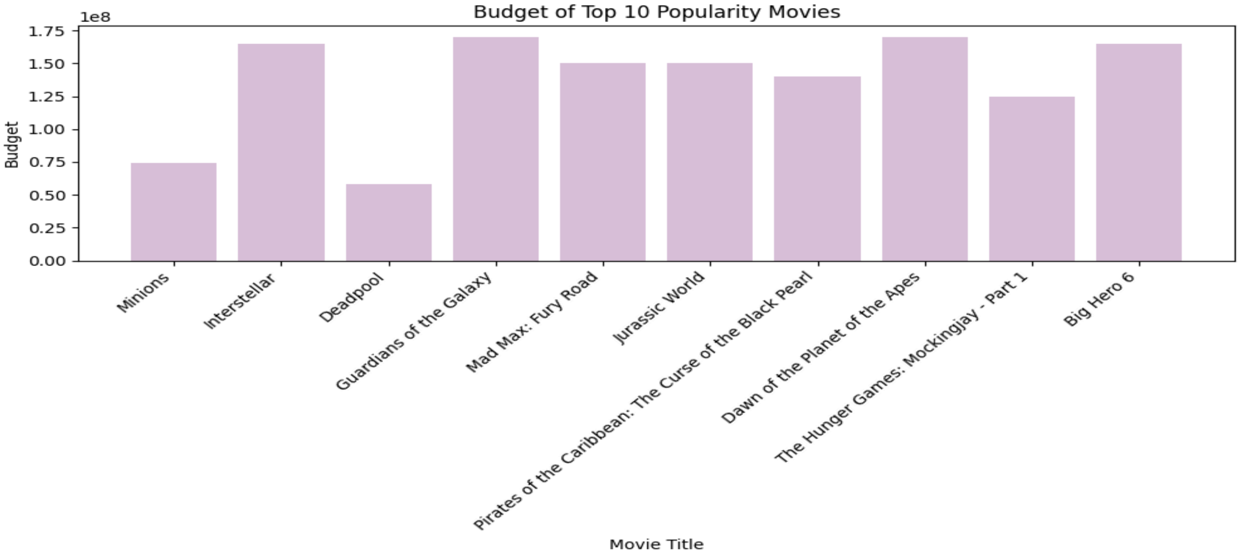

Figure 2. Budget of Top 10 Popularity Movies

Q3: Plot the movies with the maximum and minimum budgets.

Result 3: "Pirates of the Caribbean: On Stranger Tides" was the most expensive movie, while "The Modern Times" was the lowest cost.

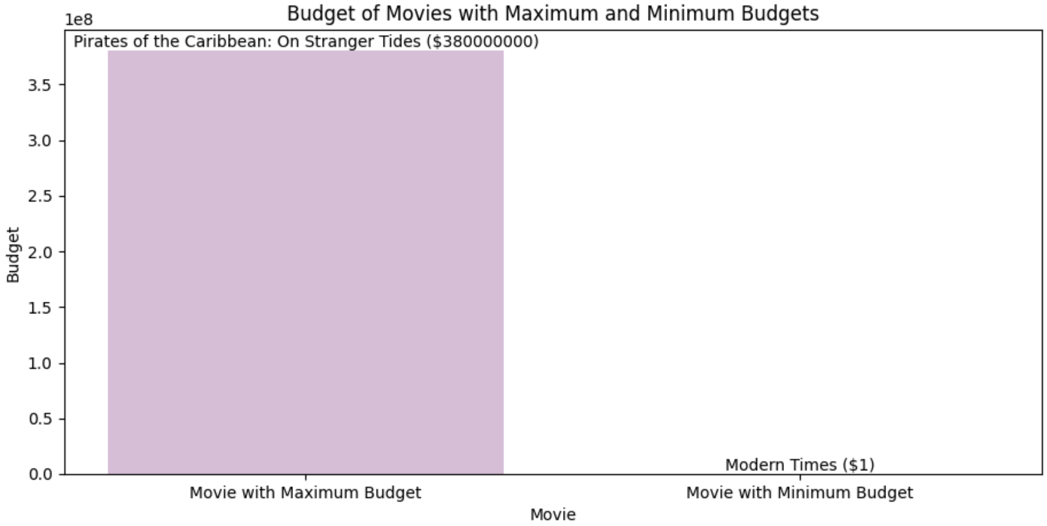

Figure 3. Budget of movies with Maximum and Minimum Budgets

Q4: What is the average rating of the 10 most expensive movies?

Result 4: "The Dark Knight Rises" has the highest average rating among the 10 most expensive movies.

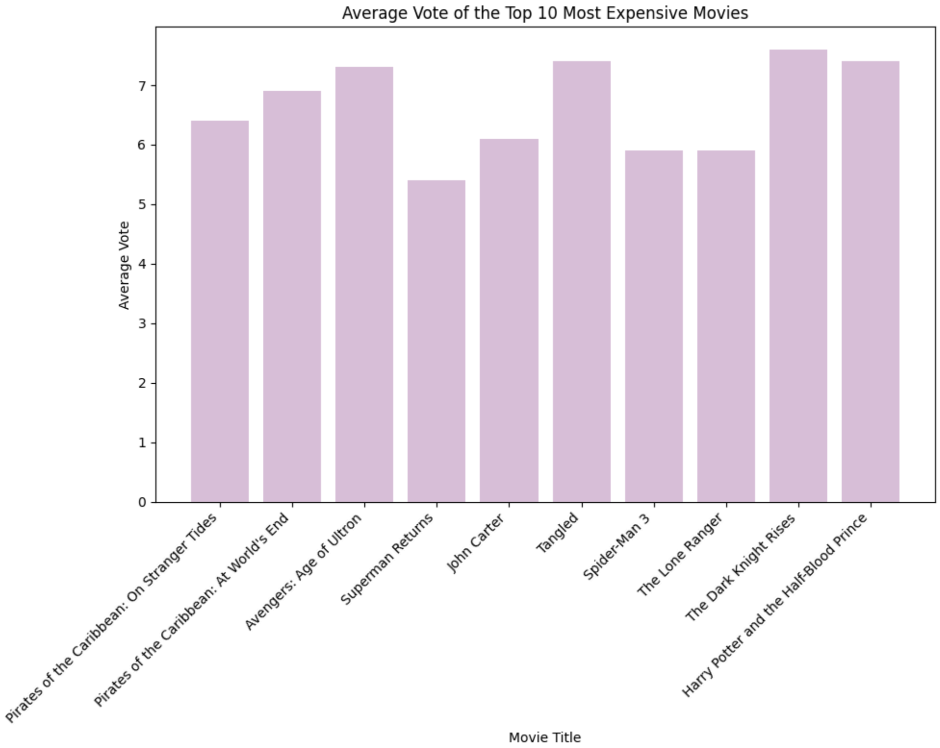

Figure 4. Average vote of the Top 10 Most Expensive Movies

Q5: How many votes did the 10 most expensive movies have?

Result 5: "The Dark Knight Rises" received more than 8000 votes.

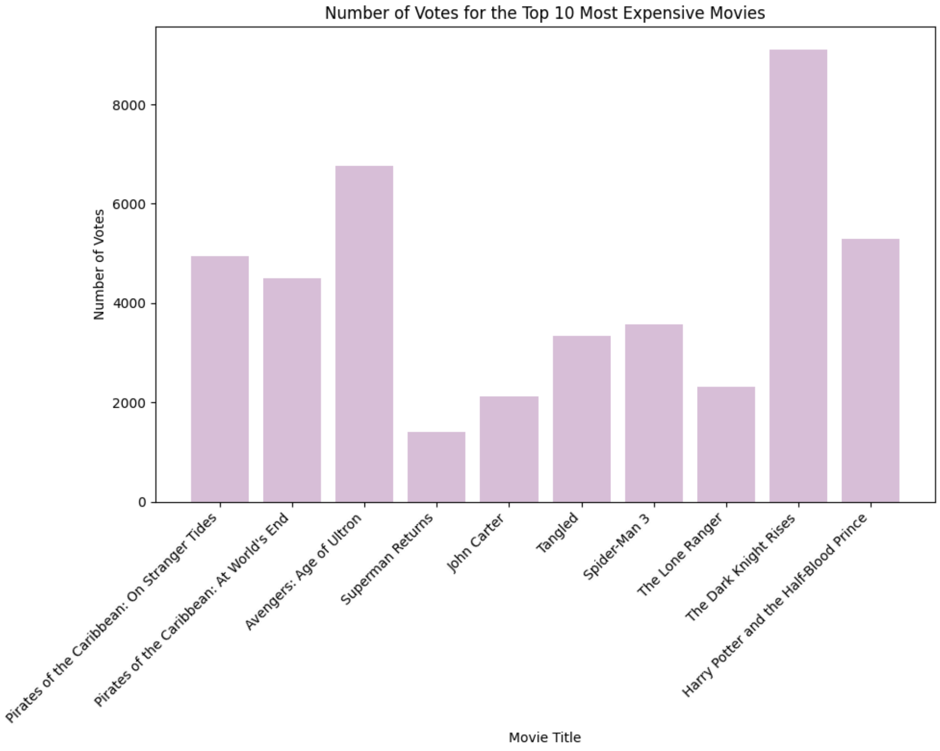

Figure 5. Number of votes for Top 10 Most Expensive Movies

Q6. Compare the budget and revenue of the movies.

Result 6: This scatter plot shows the relationship between the budget and revenue of movies. Each point on the plot represents a movie, where the x-coordinate represents the budget of the movie and the y-coordinate represents its revenue.

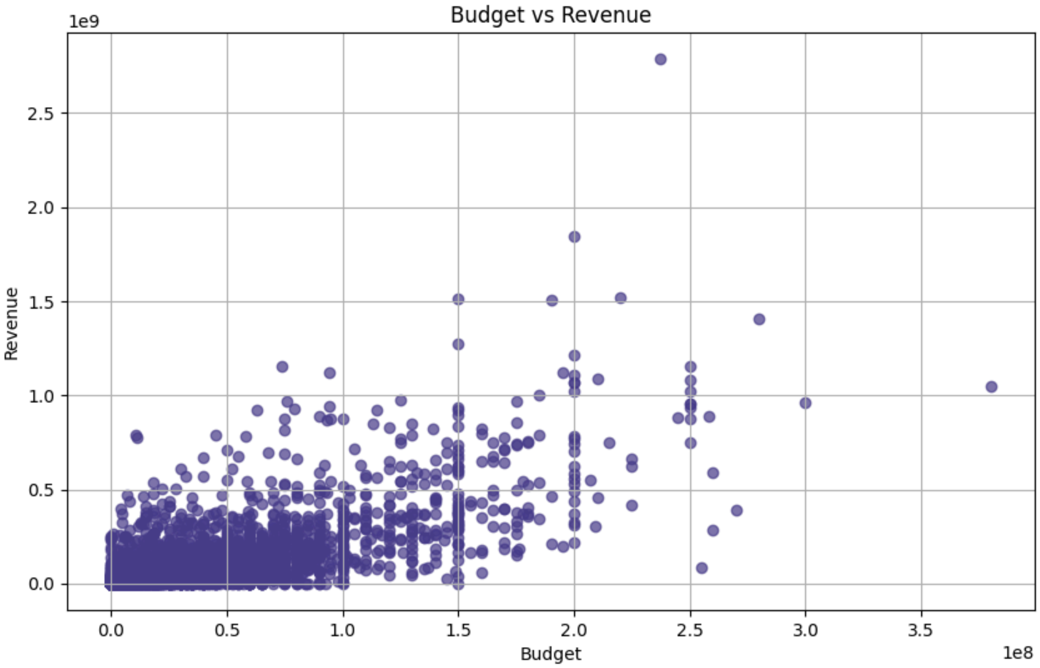

Figure 6. budget_vs_revenue

Q7. Plot the top 10 revenue movies.

Result 7: The top revenue-generating movie is "Avatar."

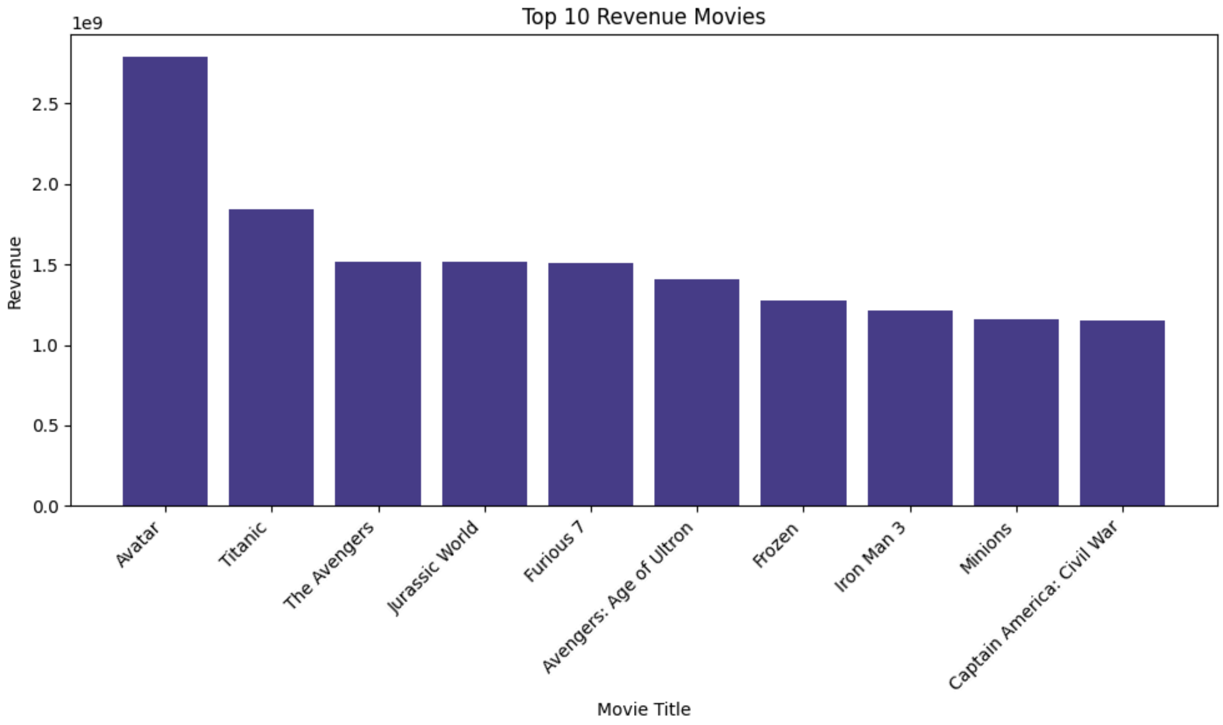

Figure 7. Top 10 revenue movies

Q8. Plot the top 10 revenue movies along with their budgets.

Result 8: The plot demonstrates that the top 10 movies were completely profitable.

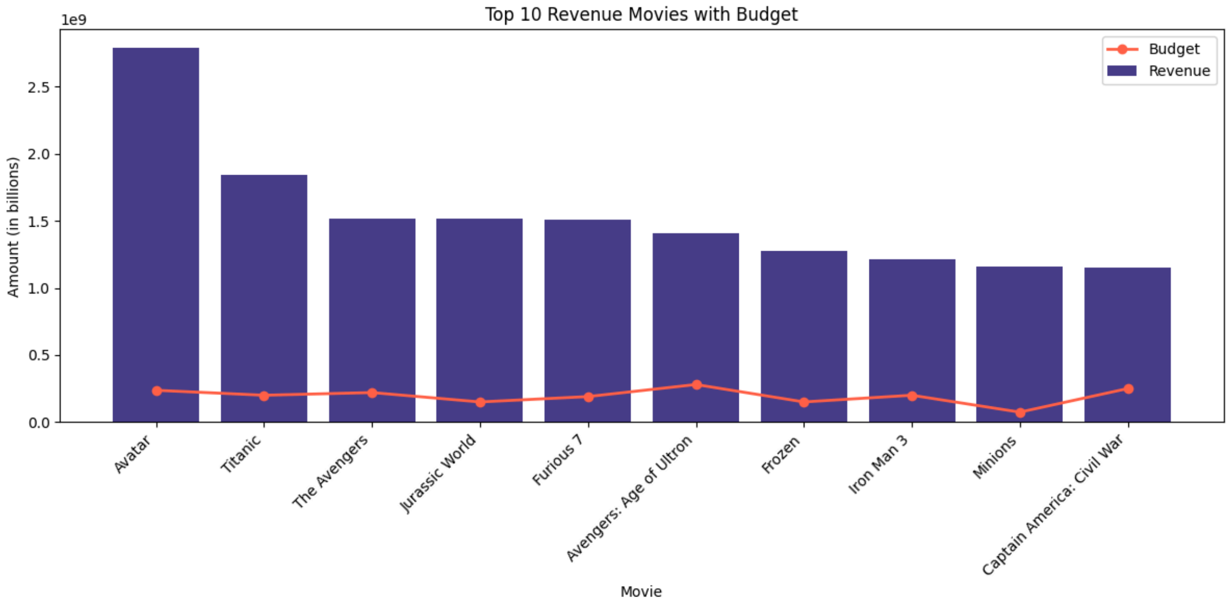

Figure 8. Top 10 Revenue Movies with Budget

Q9. Plot the revenue of the top popular movies.

Result 9: The top revenue-generating movie among the most popular ones is "Jurassic World."

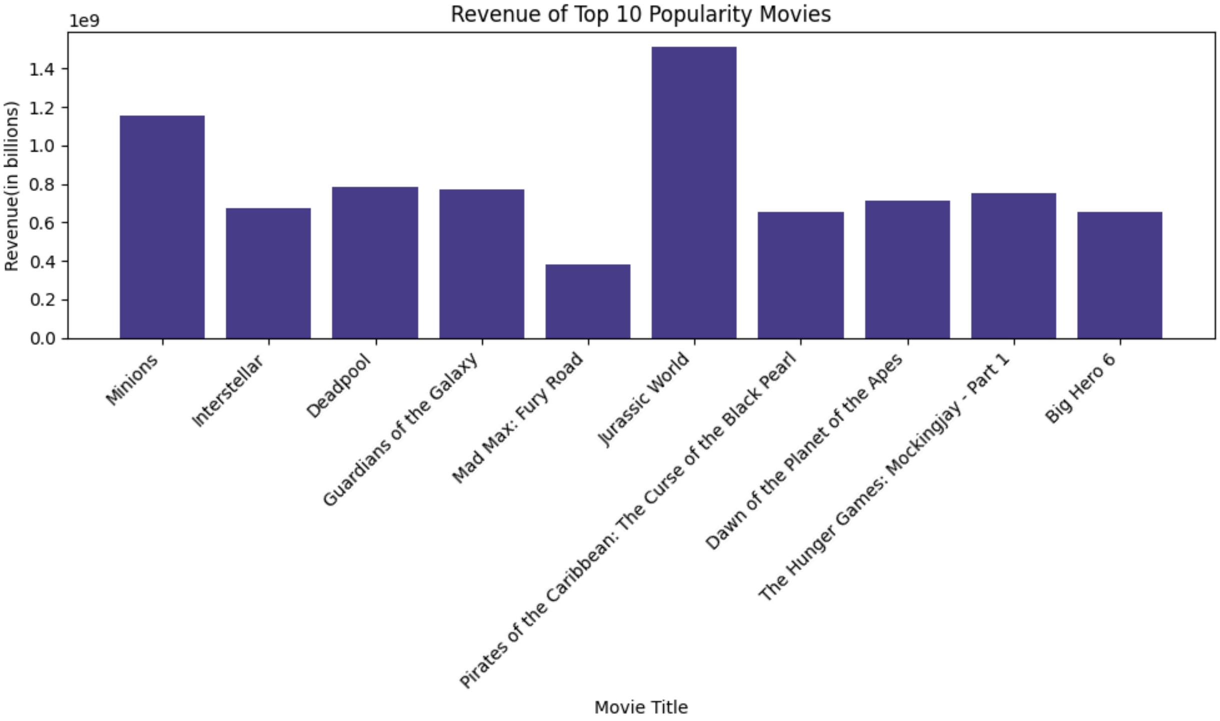

Figure 9. Revenue Of Top 10 Popularity Movies

Q10. Plot the relationship between the vote average and vote count of the movies based on their revenue.

Result 10: This plot illustrates the relationship between the vote average and the vote count of the movies based on their revenue. Each point on the plot represents a movie, with the x-coordinate representing the vote average and the y-coordinate representing the vote count. The size of each point represents the revenue of the corresponding movie.

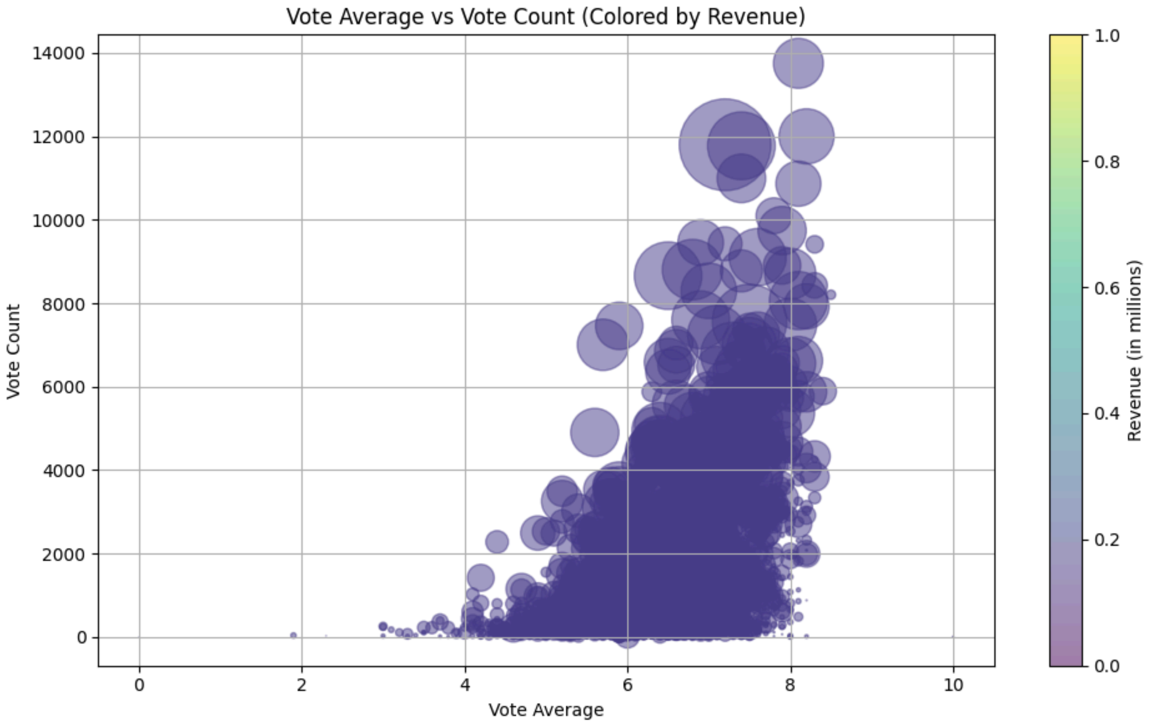

Figure 10. Vote Average vs Vote Count

Analysis of non-Numeric Columns: Various queries were prepared to analyze the distribution and characteristics of numeric columns.

For instance:

Q11. Plot the distribution of genres.

Result 11: The most frequent genres are Drama and Thriller.

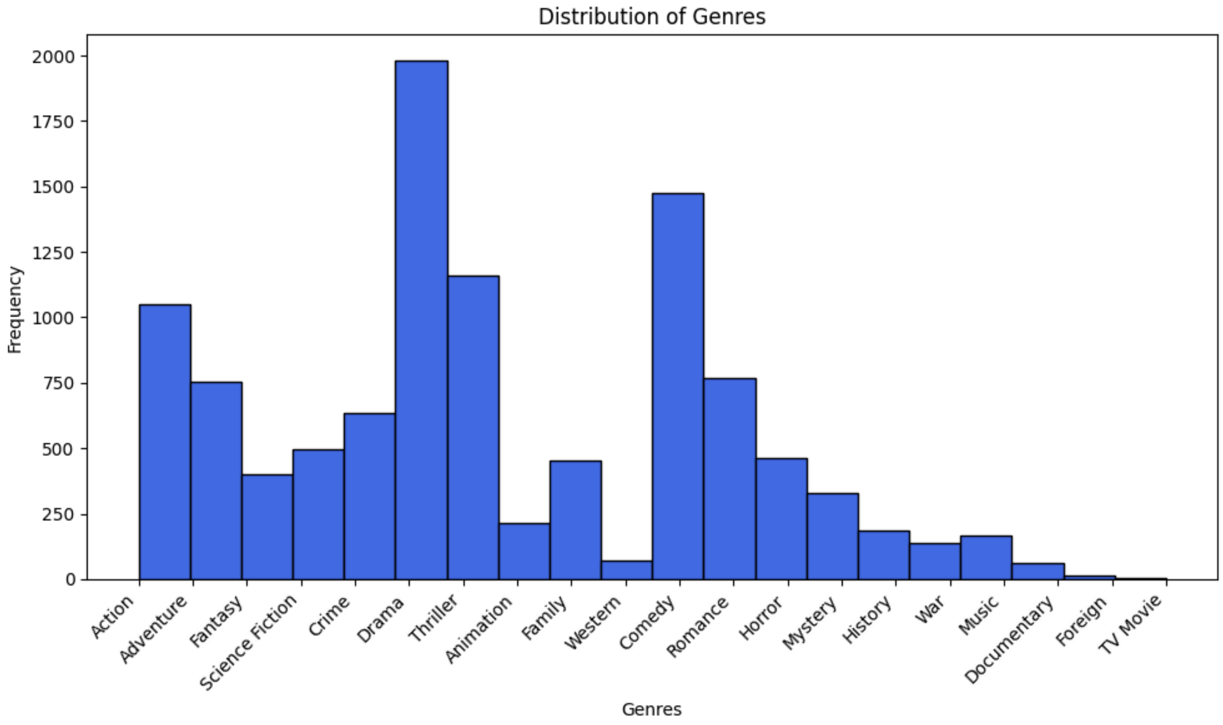

Figure 11. Distribution of Genres

Q12. Plot the top 10 genres based on popularity.

Result 12: The most popular genres are Animation and Adventure.

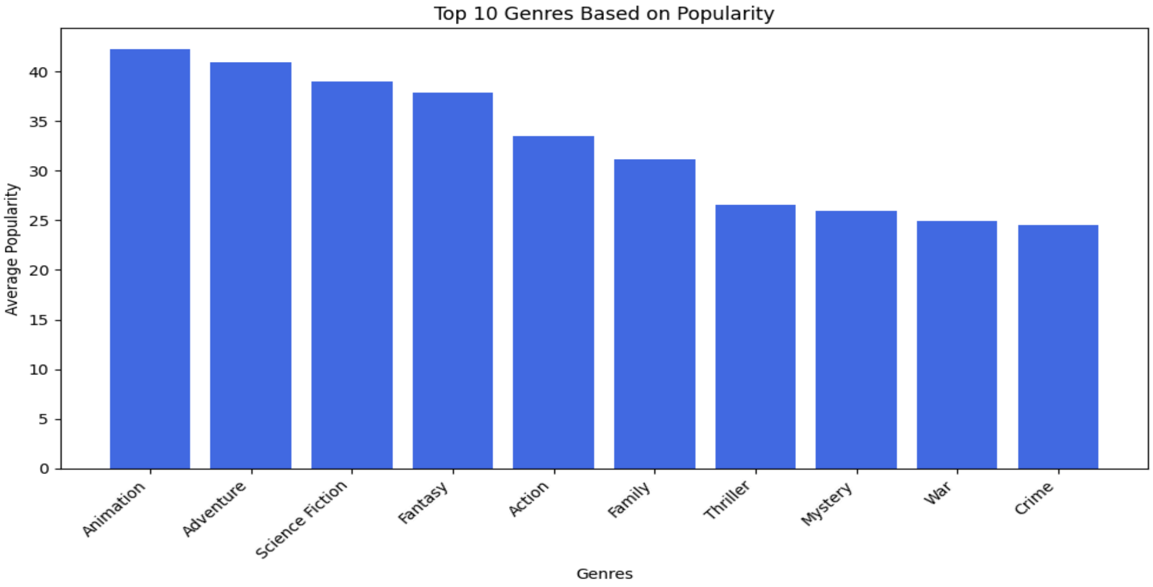

Figure 12. Top 10 Genres Based on Popularity

Q13. Plot the genres of the movie with the highest and lowest revenue.

Result 13: The genre of the movie with the highest revenue is Adventure, and the lowest revenue is TV movie.

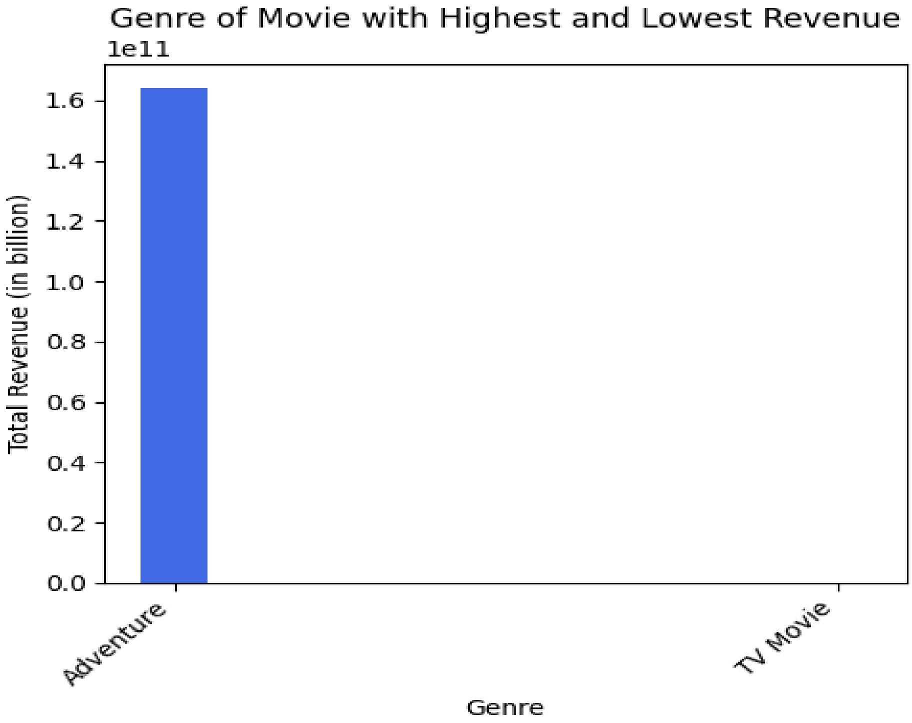

Figure 13. Genre of Movie with the Highest and Lowest Revenue

Q14. Visualize the genres of movies with the highest and lowest budgets.

Result 14: The movie with the highest budget is Pirates of the Caribbean: On Stranger Tides, and its genres are Fantasy, Adventure, and Action. The movie with the lowest budget is Modern Times, and its genres are Drama and Comedy.

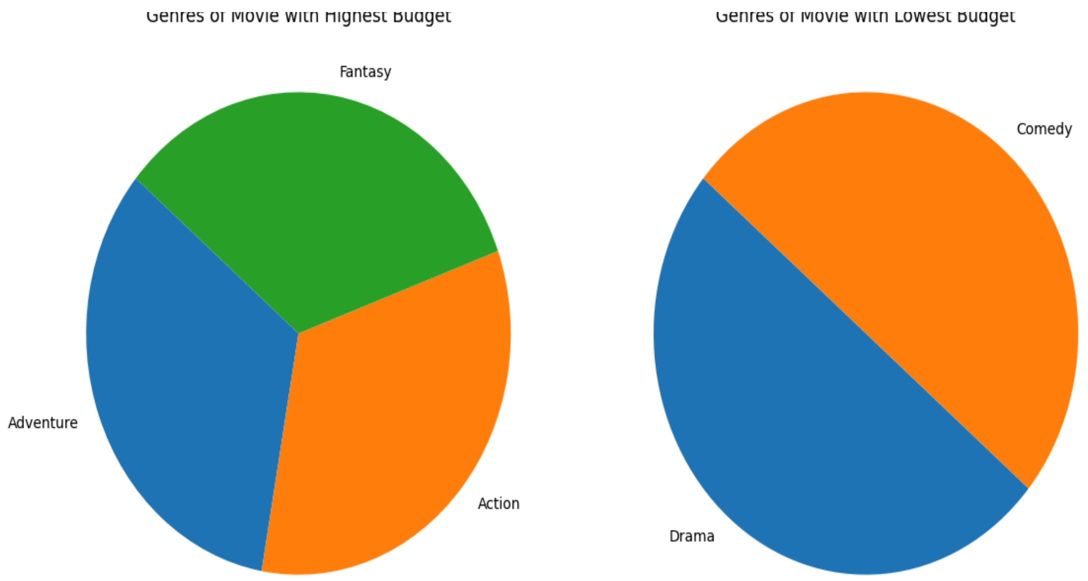

Figure 14. Genres of Movies with Highest and Lowest Budget

Modeling:

In the next step of the analysis, various machine-learning models were trained and evaluated to predict the profitability of movies based on numerical features. The following models were employed:

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

The dataset was augmented with a profitability column, which indicated whether a movie was profitable or not. This column was derived based on the difference between revenue and budget. After this preparation step, the following numerical columns were selected as features for the modeling process:

- Budget
- Popularity
- Revenue
- Runtime
- Vote Average

- Vote Count

Each model was trained on these features to predict the profitability of movies. The performance of each model was evaluated using four metrics: Accuracy, Precision, Recall, and F1-Score. Additionally, the ROC-AUC score was computed to assess the model's ability to distinguish between profitable and non-profitable movies.

The results obtained from the training and testing of each model are summarized below:

Logistic Regression:

- Accuracy: 0.9988
- Precision: 1.0000
- Recall: 0.9980
- F1-Score: 0.9990
- ROC-AUC: 0.9990

Decision Tree:

- Accuracy: 0.9844
- Precision: 0.9902
- Recall: 0.9843
- F1-Score: 0.9872
- ROC-AUC: 0.9844

Random Forest:

- Accuracy: 0.9856
- Precision: 0.9883
- Recall: 0.9883
- F1-Score: 0.9883
- ROC-AUC: 0.9848

XGBoost:

- Accuracy: 0.9820
- Precision: 0.9921
- Recall: 0.9785
- F1-Score: 0.9852
- ROC-AUC: 0.9830

These results indicate that all models achieved high performance in predicting the profitability of movies. Logistic Regression exhibited the highest accuracy and precision, indicating its effectiveness in correctly classifying profitable movies. Furthermore, the ROC-AUC scores for all models were close to 1, suggesting strong predictive capabilities in distinguishing between profitable and non-profitable movies. Overall, the modeling process utilizing numerical features proved to be successful in predicting movie profitability, providing valuable insights for decision-making in the film industry.

Recommendation System:

In the final step of the project, a recommender system was developed to provide movie suggestions based on relevant tags present in the dataset. The recommender system utilizes similarity measures between movies based on their tag vectors. Here's an overview of the process:

Feature Selection: Relevant features were selected for building the recommender system. This included 'genres', 'original_title', 'overview', 'keywords', 'cast', 'crew', and 'tagline'.

Text Pre-processing: Text pre-processing techniques were applied to standardize, normalize, and reduce the size of text data. This involved:

- Lowercasing in 'genres', 'original_title', and 'overview' to ensure uniformity.
- Handling HTML tags in 'original_title' and 'overview'.
- Converting strings to lists to facilitate merging in overview.
- Removing spaces from keywords, cast, and crew.
- Extracting the top 3 actors from the cast.
- Making tag features to capture a richer representation of each movie.
- Dropping redundant columns.
- Applying stemming in 'tags' to reduce word variation.

Vectorization: The TF-IDF vectorizer or Count vectorizer was used to map different tags to a vectorized space. This process transforms textual data into numerical vectors, making it suitable for similarity calculations.

Similarity Calculation: Cosine similarity, a commonly used similarity measure, was computed between the vectors representing different movies. Cosine similarity measures the cosine of the angle between two vectors and determines how similar they are.

Recommendation Function: Based on the input movie, a function was developed to calculate the similarity measure between the input movie's vector and all other movie vectors. The function then suggests the top 5 movies that are most similar to the input movie.

The output of the recommender system provides the top 5 recommended movies similar to the input movie.

For example, based on the input movie "The Dark Knight", the recommender system suggests the following similar movies:

"The Dark Knight Rises"
"Batman Begins"
"Batman Returns"
"Batman"
"Batman Forever"

These recommendations are generated based on the similarity of tag vectors, enabling personalized movie suggestions for users.

Conclusion:

In conclusion, this project undertook a comprehensive analysis of a movie dataset, encompassing data cleaning, exploratory data analysis (EDA), modeling, and the development of a recommender system. The project aimed to extract insights from the dataset, predict movie profitability, and provide personalized movie recommendations.

The analysis began with data cleaning, where missing values were handled, duplicates were removed, and text data underwent preprocessing to ensure consistency and standardization. EDA was then conducted to explore the distribution and relationships among various features, providing valuable insights into the dataset's characteristics.

For modeling, four different machine-learning algorithms were implemented to predict movie profitability. Each model was evaluated based on key performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. The results indicated promising performance across all models, with logistic regression achieving the highest accuracy and precision.

In the final stage, a recommender system was developed to suggest similar movies based on relevant tags present in the dataset. Text preprocessing techniques, vectorization, similarity calculation, and a recommendation function were employed to generate personalized movie recommendations.

Overall, this project demonstrates the application of data science techniques to extract insights, make predictions, and provide valuable recommendations in the domain of movie analysis. The findings and methodologies presented herein contribute to a deeper understanding of the dataset and offer practical solutions for movie enthusiasts and industry stakeholders alike.