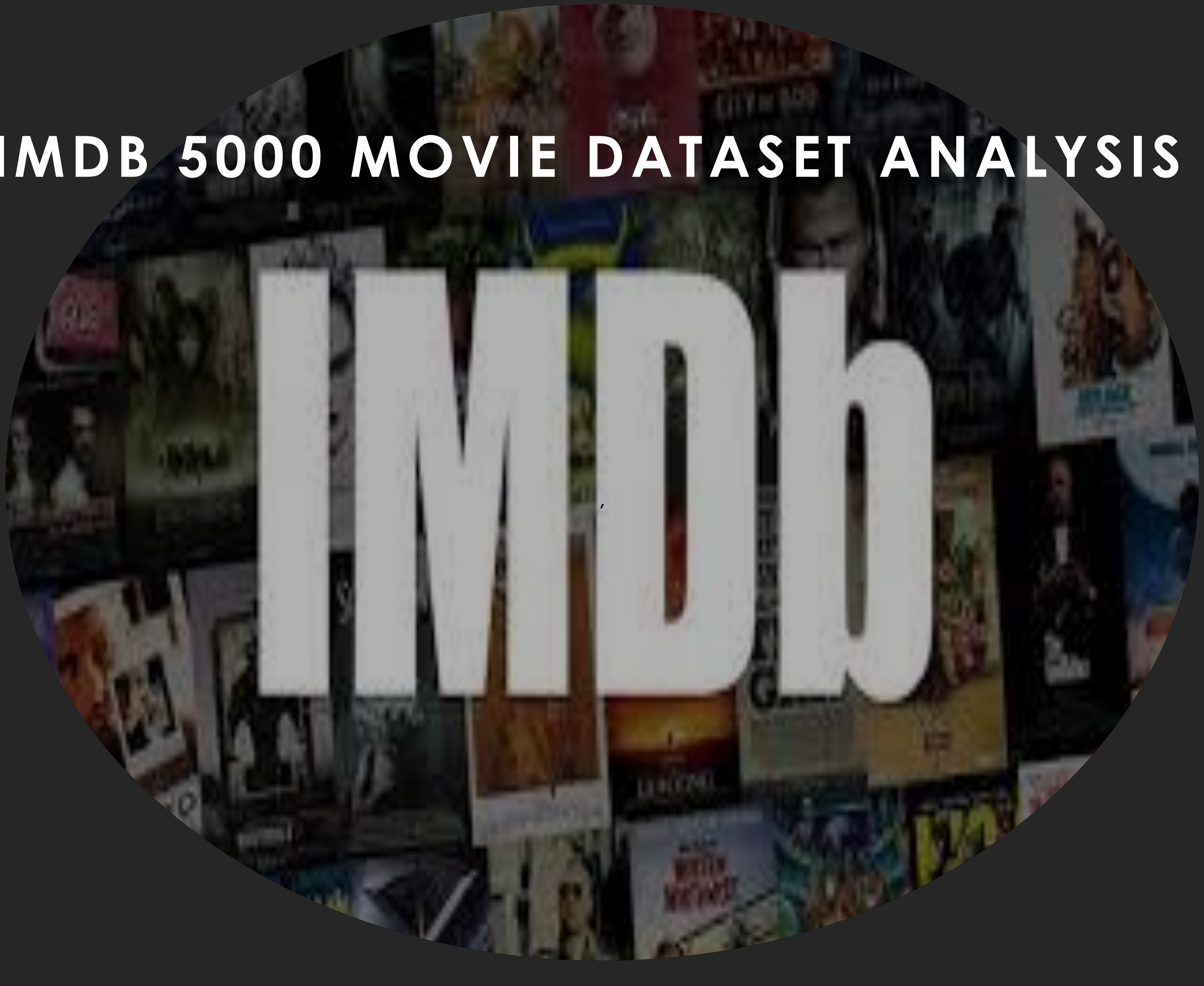
ANALYSIS REPORT



IMDB 5000 MOVIE DATASET ANALYSIS

Sama Hosseini
Git: tmdb_5000_movies

TABLE OF CONTENTS

01 - DATA CLEANING AND PREPROCESSING

ABOUT THE DATASET:

Background: The dataset contains information on thousands of movies, covering plot, cast, crew, budget, and revenue. It's valuable for predicting a movie's success before release and understanding its potential ratings or commercial performance.

Data Source Transfer: Originally from IMDB, the dataset was transferred to The Movie Database (TMDb) following a DMCA takedown request. This transition improved data consistency and added features like full credits for cast and crew.

Key Features:

- Title, language, spoken languages
- Genres, keywords, overview
- Production details, release date, status
- Financial data (budget, revenue)
- Runtime, popularity, homepage
- Cast, crew, vote count, average rating
- Profit

Dataset Overview:

The dataset comprises 4803 rows and 24 columns.

Columns include information such as movie budget, genres, homepage, keywords, popularity, production details, release date, revenue, runtime, spoken languages, tagline, and cast/crew details.

Data Quality:

Missing Values: Several columns have missing values, with 'homepage' having the highest count of 3091 missing values, followed by 'tagline' with 844 missing values.

Empty Lists: Some columns, including 'spoken_languages', 'genres', 'keywords', 'production_companies', 'production_countries', 'cast', and 'crew', contain empty lists for certain entries.

Descriptive Statistics:

- Budget: The mean budget is approximately $29 million, with a minimum of $0 and a maximum of $380 million.
- Revenue: The mean revenue is about $82 million, ranging from $0 to over $2.78 billion.
- Popularity: The average popularity score is around 21.49, with a wide range from 0 to 875.58.
- Runtime: The average runtime is approximately 106.88 minutes, with values ranging from 0 to 338 minutes.
- Vote Average: The mean vote average is 6.09, with scores varying from 0 to 10.
- Vote Count: The average vote count is about 690, with a minimum of 0 and a maximum of 13,752

Data Preprocessing:

- Custom function has_http checks for 'http' in 'homepage' column, resulting in 'has_homepage' column creation.
- 'homepage' column is removed post-application to streamline the dataset.

Feature Engineering:

- 'release_date' column converted to datetime for time-based analysis.
- Additional features ('year', 'month', 'day', 'dow') derived from 'release_date'.
- 'title_x' column converted to lowercase for consistency.
- 'popularity' column normalized using min-max scaling.

Handling Missing data:

- Missing values were identified in the dataset.
- Missing values were handled using imputation techniques.
- 'runtime', 'budget', 'revenue', 'vote_average', and 'vote_count' had missing values.
- Median imputation was applied for numerical columns.
- Modal imputation was applied for the 'vote_average' column.

Techniques Used:

- Median Imputation:

  Replaced missing values in 'runtime', 'budget', and 'revenue' with the median value of each respective column.

- Mode Imputation:

  Replaced missing values in 'vote_average' with the most frequent non-zero value.

- Zero Handling:

  Values of 0 in 'budget' and 'revenue' were considered missing and imputed with medians.

Dataset Cleaning:

This keeps the data focused and ready for further analysis.

1. Removing duplicates (exact copies of rows).
2. Dropping unnecessary columns like release date, movie ID, etc.

Data Transformation:

Genres, keywords, production companies, production countries, and spoken languages columns are parsed from JSON format into lists of respective values.

Functions for Nested JSON:

- *Custom function json_convert extracts actor names from the 'cast' column, limiting to the top 10.*
- *Function fetch_crew retrieves names of directors, writers, and producers from the 'crew' column.*

Creating Profit Column:

A new 'profit' column has been added to the dataset to analyze movie financial performance. It's calculated as below. This allows to identify profitable movies, explore profitability trends, and assess budget-revenue relationships.

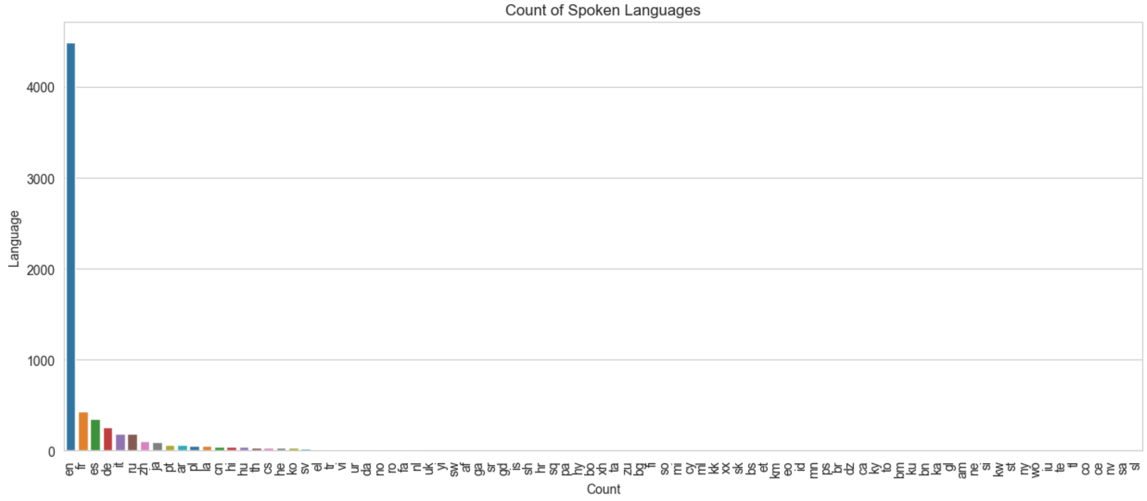
profit = revenue' - budget

02 - EXPLORATORY DATA ANALYSIS (EDA)

Movie Languages:

The bar chart illustrates the distribution of spoken languages across the dataset. Each bar represents a language, with the x-axis denoting the language and the y-axis displaying its occurrence count. The observations from the graph are as follows:

- *Language Diversity: The dataset encompasses movies spoken in various languages, indicating linguistic diversity.*
- *Dominant Language: One language, likely English, emerges prominently with the highest occurrence count, indicating its prevalence among the movies.*
- *Other Languages: Besides the dominant language, there are additional languages represented to varying extents, suggesting a diverse multilingual collection of movies.*



correlation heatmap:

The provided heatmap is a correlation heatmap, which is a graphical tool used to visualize the relationships between different numerical variables. In this case, it shows the correlation between various movie properties in the dataset.

Interpretation the heatmap:

- ***Color Strength:*** *The intensity of the color (darker or lighter) represents the strength of the correlation. Darker colors indicate stronger correlations, either positive (positive values) or negative (negative values).*
- ***Color Direction:*** *The color cast (red/orange vs. blue/purple) indicates the direction of the correlation. Red/orange hues signify a positive correlation, where higher values in one variable correspond with higher values in the other. Conversely, blue/purple hues represent a negative correlation, where higher values in one variable correspond with lower values in the other.*
- ***Positioning****: Variables are arranged on both the x-axis and y-axis, allowing us to see the correlation between any two variables by looking at the corresponding square.*

Here are some specific observations based on the heatmap:

Financial Performance:

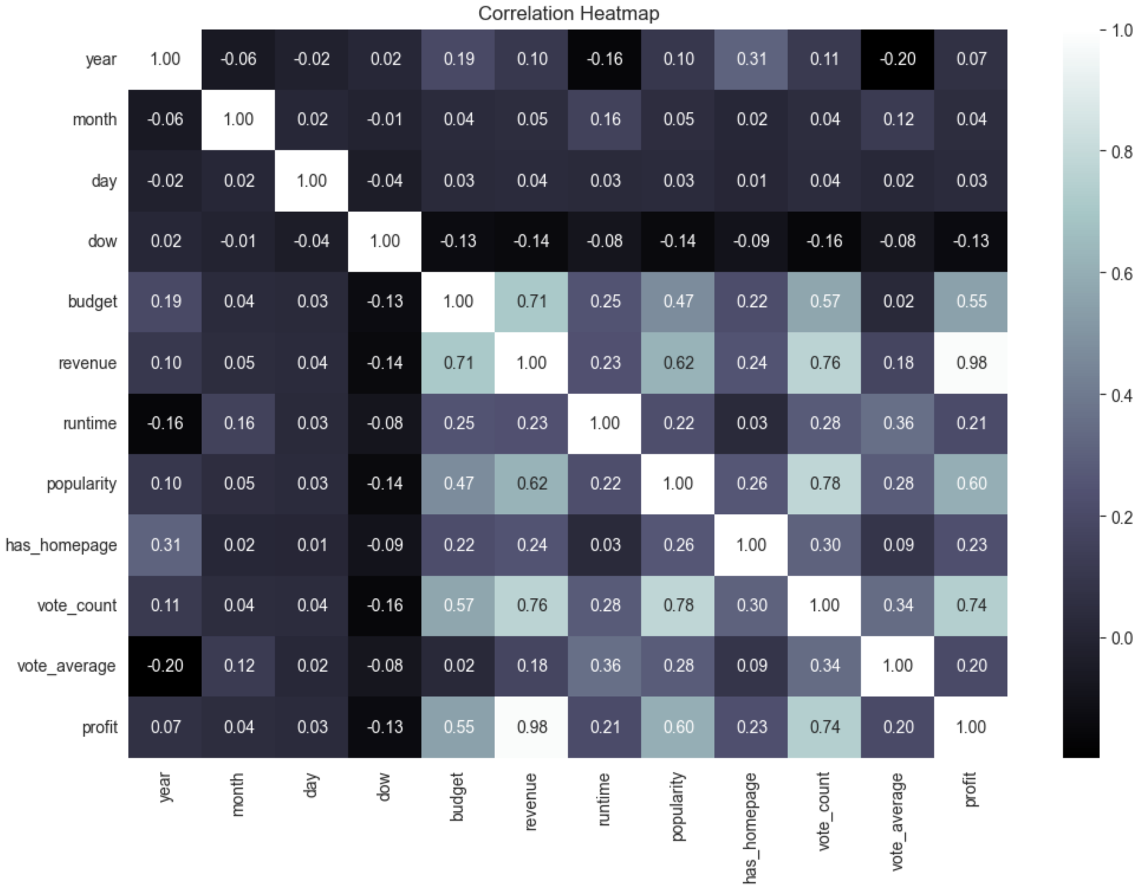
- *Budget vs. Revenue: Expect a positive correlation; higher budgets often yield higher revenues, but profitability isn't guaranteed.*
- *Budget vs. Profit: Look for a positive correlation; higher budget films may be more profitable, considering production costs.*
- *Revenue vs. Profit: Anticipate a positive correlation.*

Popularity & Critical Reception:

- *Runtime vs. Popularity: A positive correlation suggests longer films may be more popular, influenced by genre and quality.*
- *Vote Count vs. Vote Average: A positive correlation indicates well-rated movies often attract more votes.*
- *Popularity vs. Vote Average: Expect a positive correlation, implying popular movies are often critically acclaimed.*

Release & Availability:

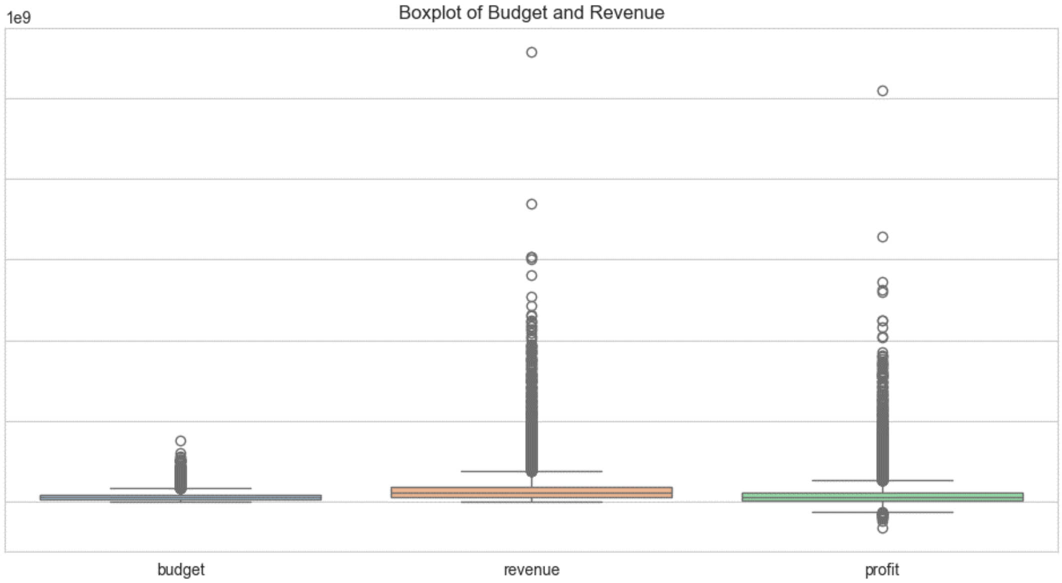
- *Year vs. Other Variables: Correlations with year can reveal trends over time, such as increasing popularity.*
- *Has Homepage vs. Other Variables: A positive correlation with popularity or vote count suggests films with a homepage have a stronger online presence.*



Boxplot of Budget, Revenue and profit:

This boxplot reveals:

- *Median Budget is Lower: Most movies cost less to produce than the revenue they generate.*
- *Budget Less Variable: Budget distribution is tighter than revenue, suggesting more consistent production costs.*
- *Revenue More Varied: Revenue distribution is wider, indicating significant differences in movie earnings.*
- *Profit Potential: The lower median budget suggests potential profitability for many movies.*
- *Outlier Extremes: High-budget and high-revenue outliers exist, representing films with either exceptional costs or earnings.*



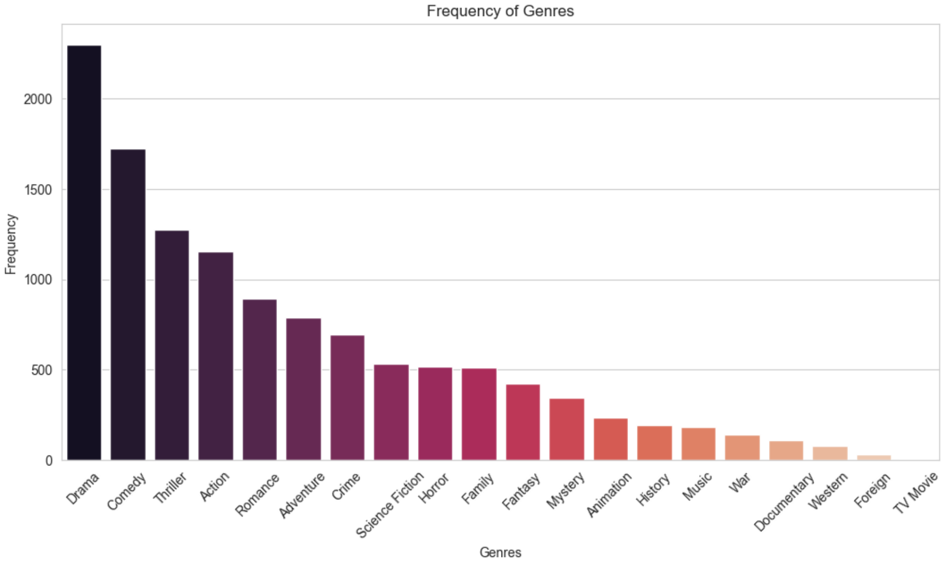
Genre Distribution:

Dominant Genres: Identify the tallest bars to uncover the most popular genres in your dataset, like "Comedy," "Action," and "Drama," indicating their frequent occurrence.

Genre Distribution: Assess the number and heights of bars to gauge the diversity of genres. Many bars of similar heights suggest a wide range of genres beyond the dominant ones.

Genre Relationships: Observe any clustering of bars to detect genres often paired together, such as "Thriller" and "Crime," indicating potential genre relationships.

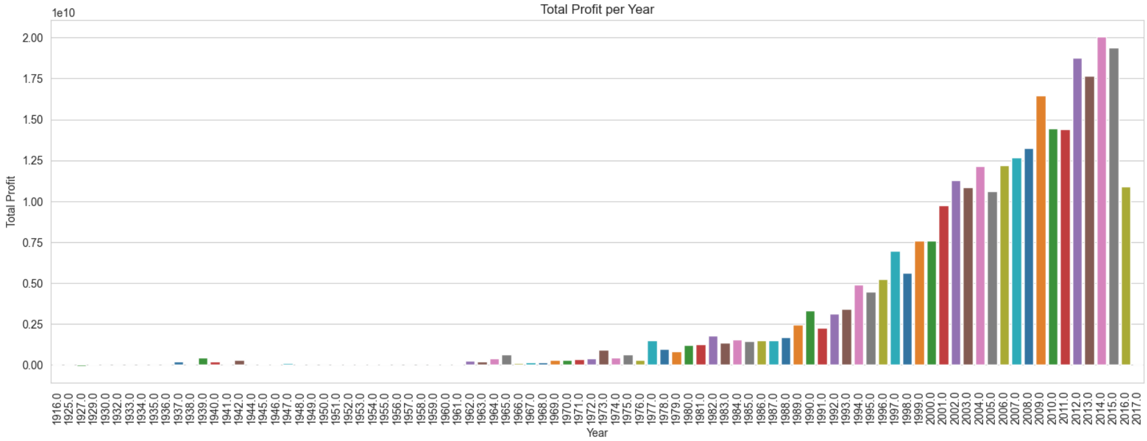
Genre Exploration: After understanding dominant and diverse genres, delve deeper by investigating specific genres' movies and comparing characteristics like budget and revenue across different genres to uncover patterns and trends.



Profit per Year:

this bar plot illustrates:

- *Yearly Profit Trend: The bar heights show how total profit fluctuates across different years. Years with taller bars represent periods with higher overall movie profits.*
- *Identifying Profitable Years: By analyzing the bars, you can pinpoint the years that yielded the highest total profits within the dataset.*
- *Profit Fluctuations: The graph reveals whether movie profits were relatively consistent over time or exhibited significant variations across different years.*



Day of Week Impact:

Each bar represents a day of the week (Sunday, Monday, etc.) and its corresponding average profit. The heights of the bars reveal variations in average profit based on release day.

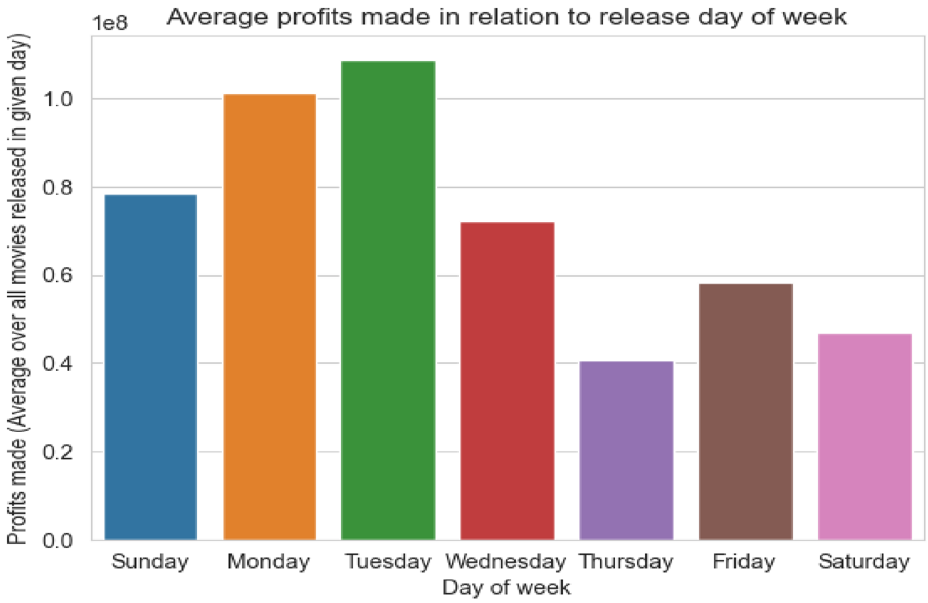
- *High-Profit Days:*

Look for days with taller bars. These days (e.g., Tuesday) are associated with movies that, on average, generate higher profits.

Movie studios or distributors might prefer releasing films on these days to potentially maximize financial returns.

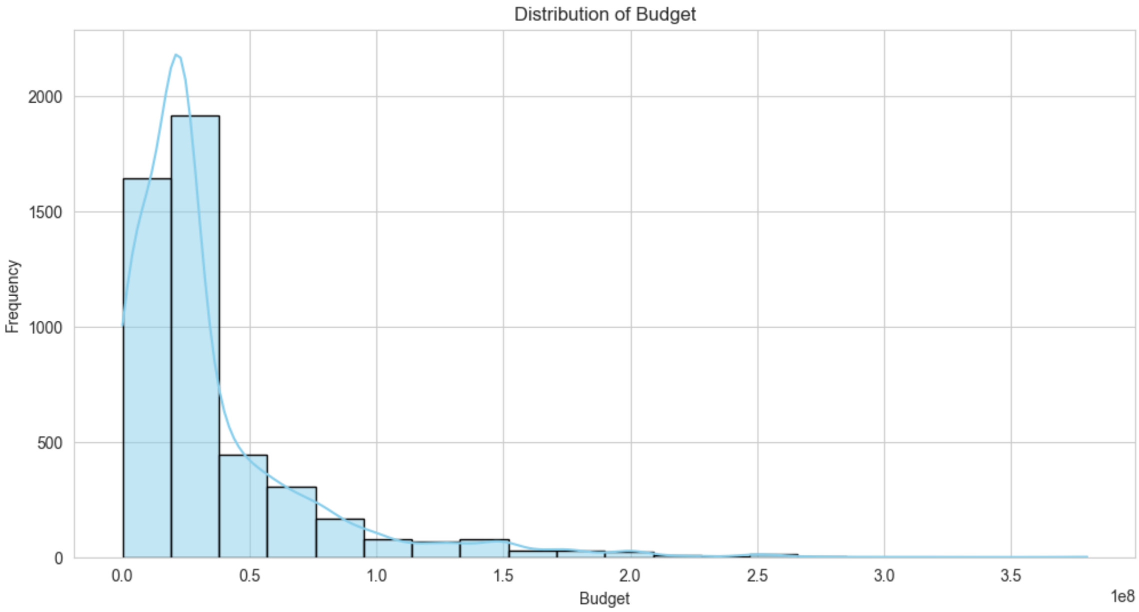
- *Low-Profit Days:*

Conversely, days with shorter bars indicate days when movies released on average tend to generate lower profits like Thursday.



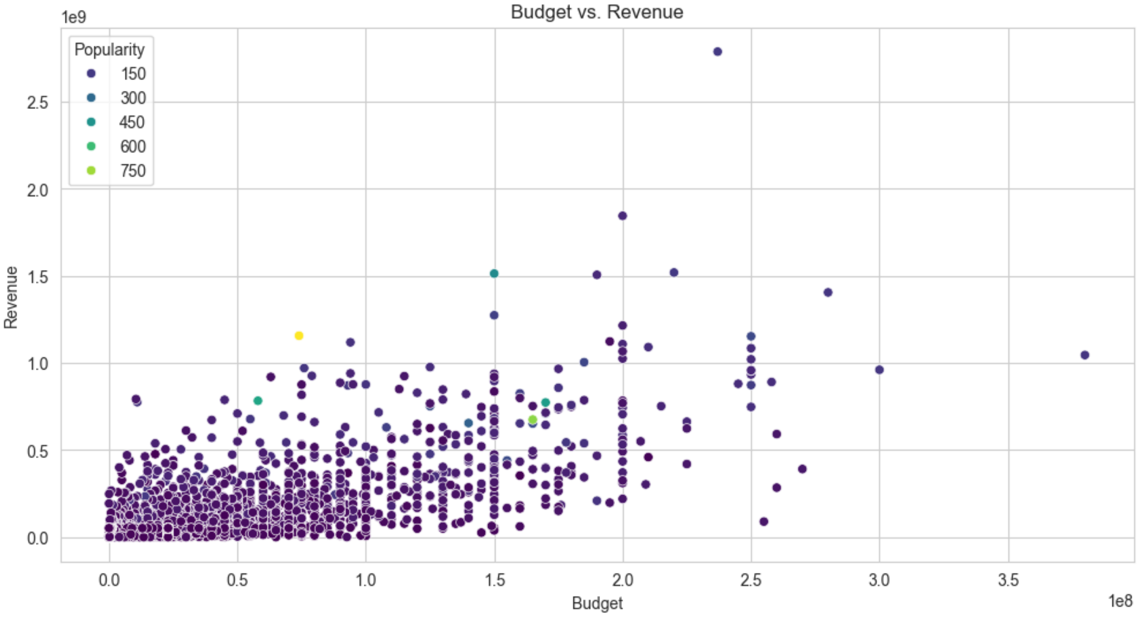
Distribution of movie budgets:

The histogram displays the distribution of movie budgets in the dataset. The x-axis represents budget ranges, likely in millions, while the y-axis indicates the frequency of movies within each range. The shape of the histogram reveals the distribution pattern: a bell-shaped curve suggests a normal distribution, with most budgets clustered around an average, while skewness to the left or right indicates predominant low budget movies, respectively.



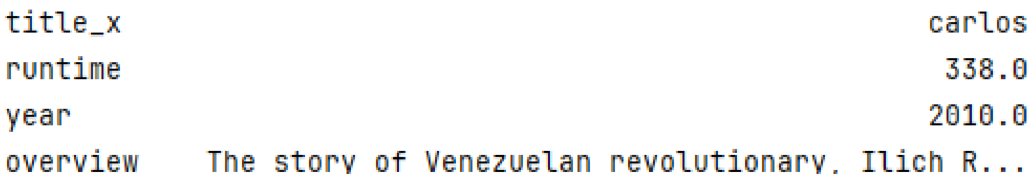
Budget vs. Revenue:

The visualization demonstrates a positive correlation between movie budgets and revenue, with higher-budget films generally yielding higher earnings, as depicted by an upward trend. Popularity, indicated by color intensity, further enhances revenue within similar budget ranges, suggesting it plays a role in financial success. Despite the trend, variations exist, implying that factors beyond budget contribute to revenue, such as genre and critical acclaim. Outliers, representing exceptional cases, are notable, potentially highlighting blockbuster successes or anomalies in the dataset.



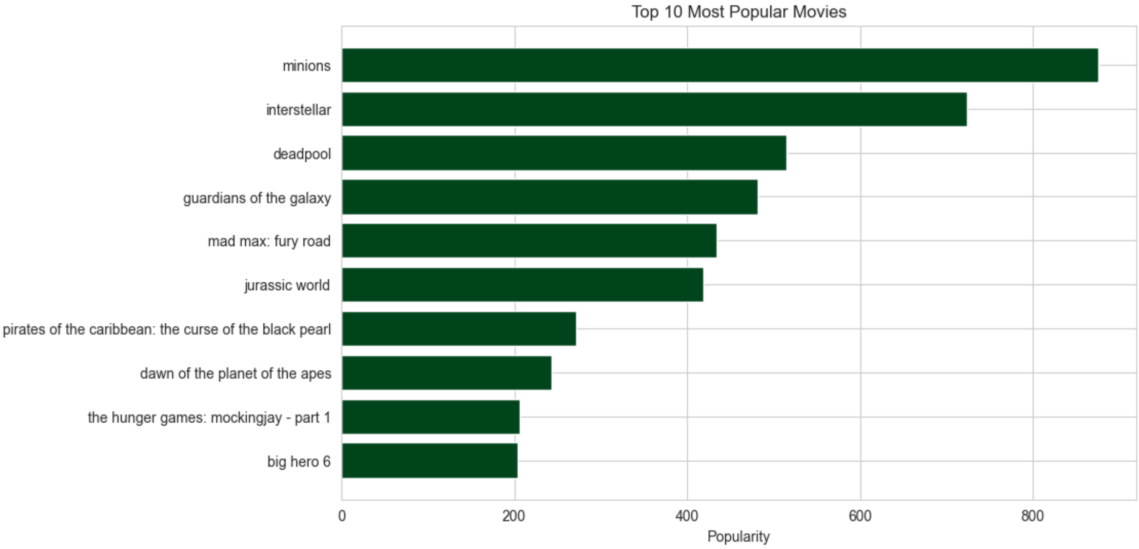
Exploring Movie Runtimes:

The average runtime of movies is approximately 107-108 minutes, with a median close to this average. However, the standard deviation and the 25%-75% ranges suggest significant variability in runtime from the mean/median. This is evident from the maximum runtime, which reaches a substantial 338 minutes. Curiosity about movies with extreme runtimes led to the utilization of the idxmax function to extract the index of the movie with a runtime exceeding 5 and a half hours.

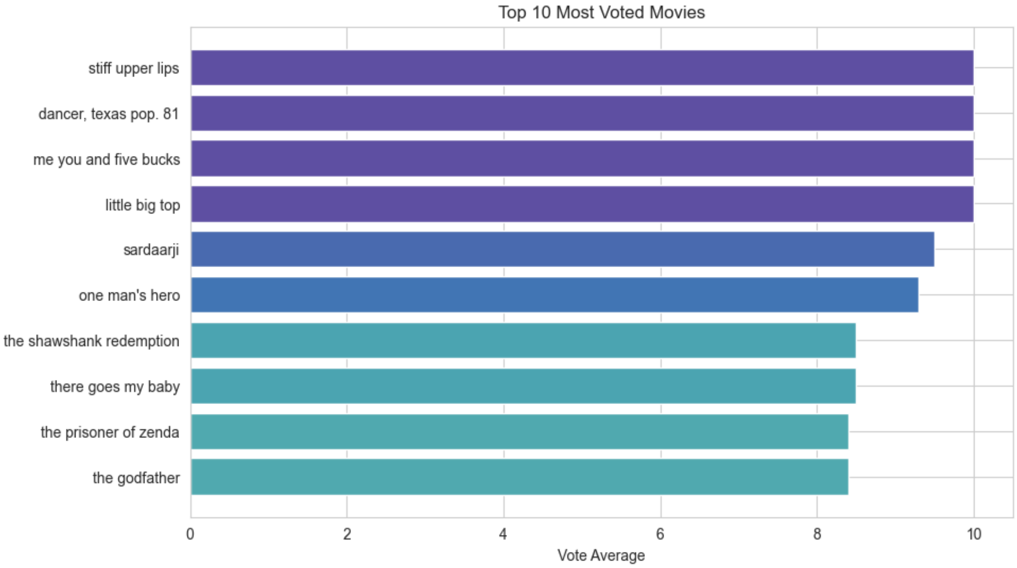


```
title_x                                                  carlos
runtime                                                   338.0
year                                                     2010.0
overview    The story of Venezuelan revolutionary, Ilich R...
```

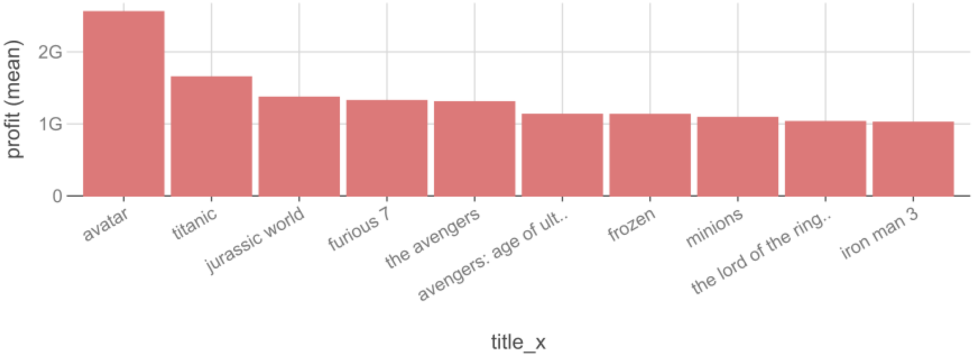
The 10 most popular movie: The most popular movie according to the popularity metric of the dataset:



Top 10 movies in regards to rating:

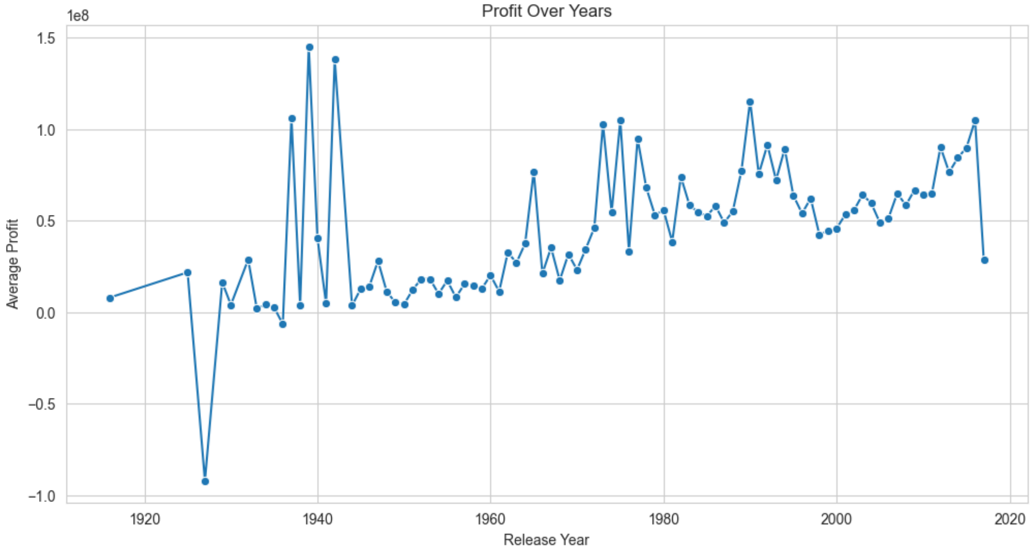


Most profitable movies:



The line graph suggests that the company's average profit has increased over the years. This could be due to a number of factors, such as the growth of the company's revenue, the increase in the number of employees, or the increase in the cost of goods and services.

- *The y-axis does not start at zero. This means that the average profit is always positive over the years shown in the graph.*
- *It is difficult to say definitively whether the increase in profit is linear or exponential from this graph.*

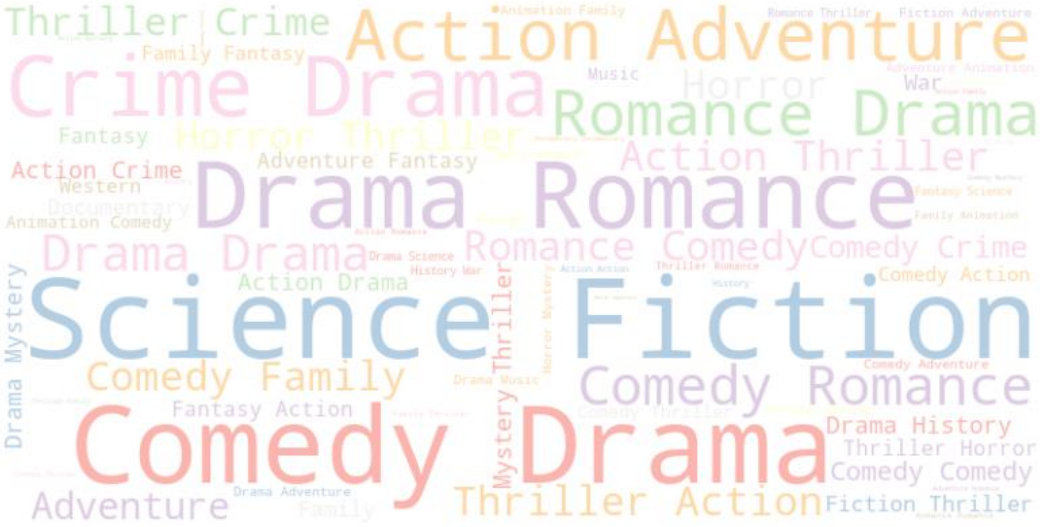


Genre Word Cloud:

Based on this word cloud, it appears to be centered around movie genres. Here's a breakdown of the prominent genres:

Central Genres: Sci-Fi, Action, Adventure, Comedy, Drama, Fantasy, Thriller

Other Genres: Romance, Crime, Western, Mystery, Family



03 – DATA MODELING

This script analyzes movie data to predict whether a movie will be profitable ("target variable").

Models Used:

- ***Logistic Regression:*** *A statistical method that predicts a binary outcome (profitable/not profitable) based on a linear relationship between features (e.g., budget, popularity).*
- ***Decision Tree:*** *A tree-like model that makes predictions by splitting data based on feature values (e.g., high-budget movies are more likely profitable).*
- ***Random Forest:*** *Combines multiple decision trees, improving accuracy and reducing overfitting.*
- ***XGBoost:*** *An advanced tree-based model known for its efficiency and accuracy in various classification tasks.*

Process:

1. ***Data Preparation:***

- *"target" variable is created to indicate profit (positive) or not (negative).*
- *Specific features are chosen, potentially including year, budget, runtime, popularity, etc.*
- *Data is split into training and testing sets for model evaluation.*
- *Features are standardized for better model performance.*

2. ***Model Training & Evaluation:***

- *Each model is trained with default hyperparameters.*
- *GridSearchCV is used to automatically tune hyperparameters for each model to improve its performance.*
- *Performance is evaluated using various metrics like accuracy, precision, recall, and F1-score.*
- *ROC-AUC curves are plotted to visualize how well each model distinguishes profitable movies from non-profitable ones.*

3. ***Feature Importance (for tree-based models):***

- *For decision tree, random forest, and XGBoost, feature importance is analyzed. This reveals which features (e.g., budget) contribute most to the model's predictions.*

Overall Performance:

- *All four models achieved relatively high accuracy scores (above 0.75), indicating they can effectively distinguish profitable movies from non-profitable ones.*
- *XGBoost appears to perform slightly better than the others, with the highest accuracy (0.82) and F1-score (0.89). However, the differences in performance between the models might be statistically insignificant depending on the dataset size.*

Model Breakdown:

- *Logistic Regression: Achieved good accuracy (0.81) but has a very high recall (1.00). This suggests it might be overly cautious and classify most movies as profitable, even if some are not.*
- *Decision Tree: Has the lowest accuracy (0.76) among the four models.*
- *Random Forest: Offers a good balance between accuracy (0.82) and precision (0.83), suggesting it can correctly identify profitable movies without making too many false positive predictions (classifying non-profitable movies as profitable).*

Feature Importance (for Tree-based models):

Most Important Features:

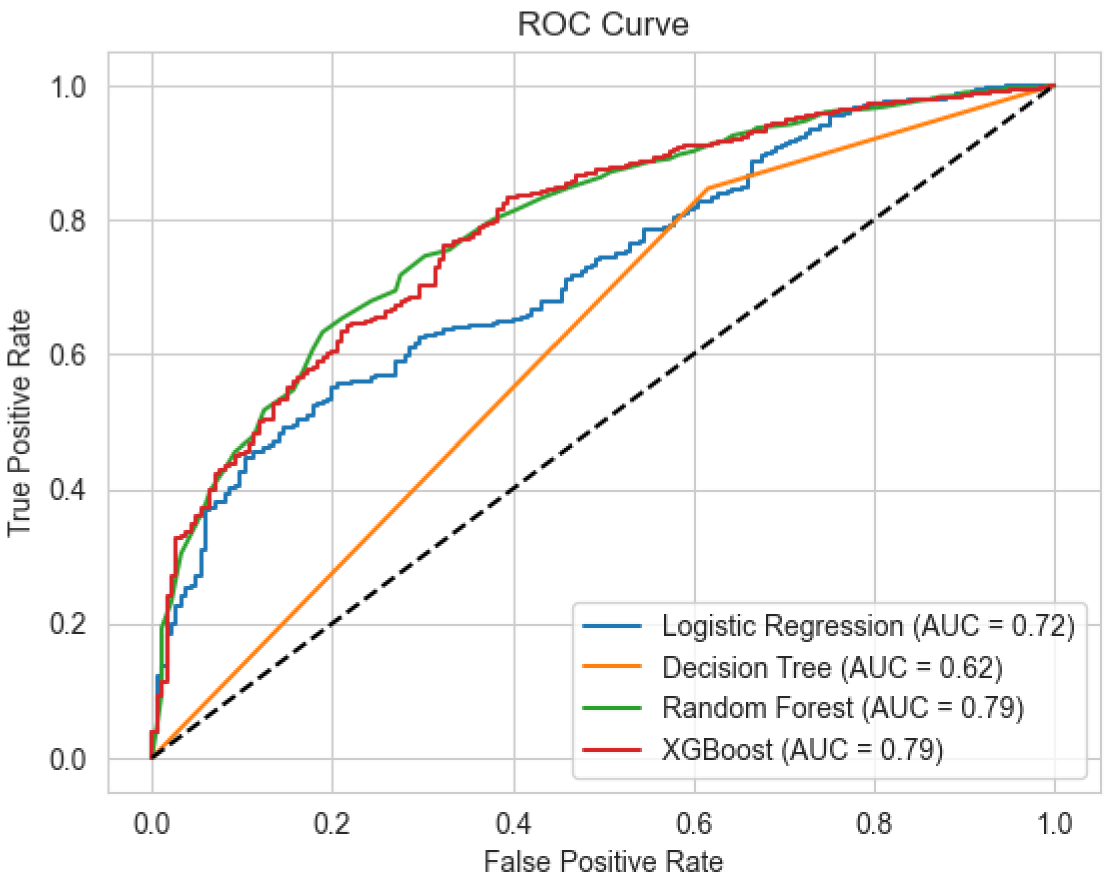
- *Day of the week (dow)*
- *Whether the movie has a homepage (has_homepage)*
- *Runtime*

Less Important Features:

- *Movie ID (id)*
- *Year*
- *Day of the month (day)*
- *Popularity*

In Conclusion:

- *XGBoost seems to be the best performing model based on accuracy and F1-score.*
- *While Logistic Regression has high overall accuracy, its high recall suggests it might be too conservative in classifying movies as profitable.*
- *Random Forest offers a balanced approach with good accuracy and precision.*
- *Decision Tree, although less accurate than the other models, highlights features like day of the week, presence of a movie homepage, and runtime as potentially important factors in predicting movie profitability***.**



```
                     Accuracy  Precision    Recall  F1-Score
Logistic Regression  0.811655   0.810867  1.000000  0.895557
Decision Tree        0.757544   0.852140  0.846649  0.849386
Random Forest        0.817898   0.833518  0.967784  0.895647
XGBoost              0.819979   0.852632  0.939433  0.893930
```

04 – RECOMMENDER SYSTEM

A Recommender System is a subclass of information filtering systems that seeks to predict the "rating" or "preference" a user would give to an item. It is commonly employed in various online platforms to suggest items that users might be interested in. Recommender systems can be categorized into several types, including collaborative filtering, content-based filtering, and hybrid methods. These systems are widely used in e-commerce platforms like Amazon, streaming services like Netflix and Spotify, social media platforms like Facebook, and many others to personalize user experiences and improve user engagement.

TF-IDF Vectorization:

- *TF-IDF (Term Frequency-Inverse Document Frequency) is a numerical statistic that reflects the importance of a word in a document relative to a collection of documents.*
- *In this code, TfidfVectorizer from scikit-learn is used to convert text data (movie titles, genres, keywords, overview, cast) into numerical vectors.*
- *The fit_transform method is used to learn the vocabulary and transform the data into a document-term matrix.*

Cosine Similarity:

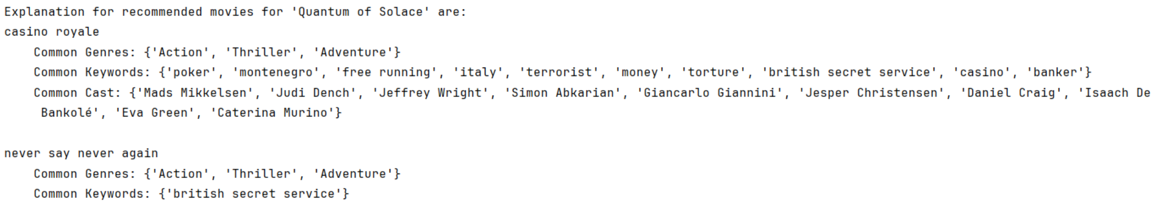
- *Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space.*
- *After TF-IDF vectorization, cosine similarity is calculated between all pairs of movies based on their feature vectors.*
- *The resulting cosine similarity matrix represents the similarity between movies.*

Recommendation Function (recommend_movies):

- *This function takes a movie title as input and returns similar movies based on cosine similarity scores.*
- *It first checks if the input movie title exists in the dataset. If not, it returns a message indicating that the movie was not found.*
- *If the movie exists, it retrieves the index of the movie in the dataset and computes similarity scores with other movies.*
- *The function sorts the movies based on similarity scores and returns the top 5 similar movies.*
- *For each recommended movie, it generates an explanation that highlights common genres, keywords, and cast with the input movie.*

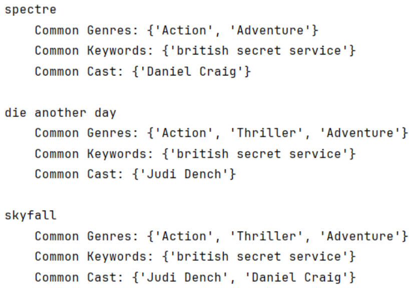
Example Usage:

- *An example is provided to demonstrate how to use the recommend_movies function with the movie title "Quantum of Solace".*
- *The function returns an explanation for recommended movies similar to "Quantum of Solace", including common genres, keywords, and cast.*



```
Explanation for recommended movies for 'Quantum of Solace' are:
casino royale
    Common Genres: {'Action', 'Thriller', 'Adventure'}
    Common Keywords: {'poker', 'montenegro', 'free running', 'italy', 'terrorist', 'money', 'torture', 'british secret service', 'casino', 'banker'}
    Common Cast: {'Mads Mikkelsen', 'Judi Dench', 'Jeffrey Wright', 'Simon Abkarian', 'Giancarlo Giannini', 'Jesper Christensen', 'Daniel Craig', 'Isaach De
     Bankolé', 'Eva Green', 'Caterina Murino'}

never say never again
    Common Genres: {'Action', 'Thriller', 'Adventure'}
    Common Keywords: {'british secret service'}
```



```
spectre
    Common Genres: {'Action', 'Adventure'}
    Common Keywords: {'british secret service'}
    Common Cast: {'Daniel Craig'}

die another day
    Common Genres: {'Action', 'Thriller', 'Adventure'}
    Common Keywords: {'british secret service'}
    Common Cast: {'Judi Dench'}

skyfall
    Common Genres: {'Action', 'Thriller', 'Adventure'}
    Common Keywords: {'british secret service'}
    Common Cast: {'Judi Dench', 'Daniel Craig'}
```