EXPLORATORY DATA ANALYSIS

IMDB TOP 2000 MOVIES DATASET

PRIYADARSH P V

Introduction

This report presents an exploratory data analysis (EDA) of the IMDb Top 2000 Movies dataset, sourced from Kaggle. The dataset comprises a comprehensive collection of movies, featuring key attributes such as movie names, release years, running times, IMDb ratings, Metascores, vote counts, genres, directors, cast members, and gross earnings. The primary objective of this analysis is to uncover insights and patterns within the dataset that can enhance our understanding of cinematic trends and audience preferences.

The analysis begins with data cleaning, ensuring the integrity and accuracy of the dataset by addressing missing values and refining data types for optimal analysis. Subsequent feature engineering processes further enrich the dataset, enabling a deeper exploration of genres and categorization of movies based on their release years.

Utilizing various statistical techniques and visualization tools, this report aims to identify trends in movie ratings, gross earnings, and genre distributions, while also examining the relationships between different variables. Through this exploratory analysis, we seek to draw meaningful conclusions about the movie landscape, which can serve as a foundation for future predictive modeling and recommendationsystemdevelopment.

Data Overview

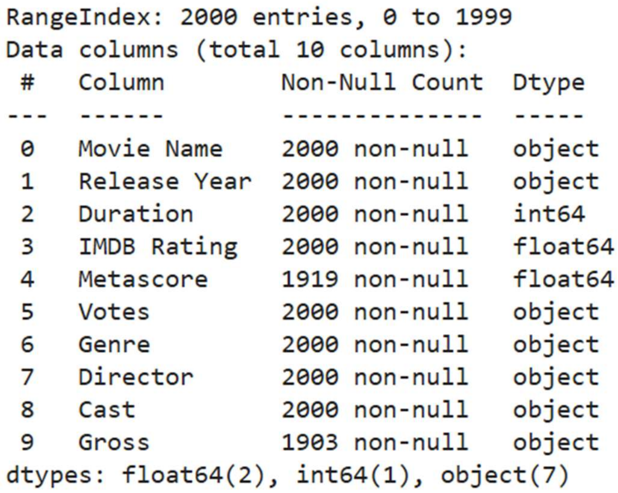
The **IMDB Top 2000 Movies** dataset provides an extensive collection of information about the most popular films on IMDb, offering valuable insights for various analytical purposes. The dataset is sourced from IMDb's official website and made available through Kaggle, encompassing 2000 entries with 10 distinct columns. Below is a detailed description of each column in the dataset:

1. **Movie Name**: The title of the film, which allows for identification and reference.
2. **Release Year**: The year the movie was released, providing context for its production and reception over time.
3. **Duration**: The running time of the movie in minutes, which can be useful for analyzing trends in movie length and its potential impact on audience reception.
4. **IMDB Rating**: A floating-point number representing the film's rating on IMDb, reflecting viewer satisfaction and critical acclaim.
5. **Metascore**: A floating-point number indicating the Metascore, which aggregates critical reviews and provides an overall assessment of the film's quality. Note that some entries may be missing this data.
6. **Votes**: The number of votes the film has received on IMDb, represented as an object type. This metric can indicate the film's popularity and level of engagement from viewers.
7. **Genre**: The genre or categories the movie falls under, allowing for classification and segmentation of films based on thematic content.
8. **Director**: The name of the film's director, providing insights into directorial style and the potential influence of specific filmmakers on the film's success.
9. **Cast**: A listing of the primary actors in the movie, which can be crucial for understanding the film's appeal and star power.

10. **Gross**: The total gross revenue generated by the film, represented as an object type. This financial metric is essential for evaluating the commercial success of the movie.

Dataset Characteristics

- **Total Entries**: 2000
- **Total Columns**: 10
- **Non-null Counts**: Most columns contain complete data, while some, such as Metascore and Gross, have missing values, indicating potential areas for further data cleaning or analysis.

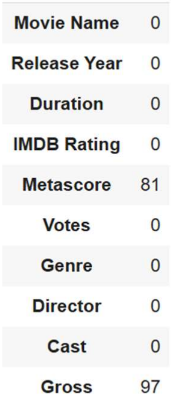


```
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   Movie Name    2000 non-null   object 
 1   Release Year  2000 non-null   object 
 2   Duration      2000 non-null   int64  
 3   IMDB Rating   2000 non-null   float64
 4   Metascore     1919 non-null   float64
 5   Votes         2000 non-null   object 
 6   Genre         2000 non-null   object 
 7   Director      2000 non-null   object 
 8   Cast          2000 non-null   object 
 9   Gross         1903 non-null   object 
dtypes: float64(2), int64(1), object(7)
```

Data Cleaning Process

In preparing the IMDB Top 2000 Movies dataset for analysis, several critical data cleaning steps were undertaken to ensure the integrity and usability of the data. The following actions were performed:

1. **Handling Missing Values**:
   - **Removal of Null Values**: To maintain the quality of the dataset, all null values were removed using the dropna() function. This step was essential as missing data can lead to inaccurate analyses and conclusions. By eliminating rows with null entries, we ensured that only complete cases were retained for further analysis.



Movie Name	0
Release Year	0
Duration	0
IMDB Rating	0
Metascore	81
Votes	0
Genre	0
Director	0
Cast	0
Gross	97

2. **Checking for Duplicates**:
    - **Duplicate Verification**: A thorough check for duplicate entries was conducted in the dataset. The process confirmed that there were zero duplicates present. This finding is crucial as duplicates can skew results and lead to misleading interpretations in any subsequent analysis.
3. **Data Type Conversion**:
    - **Changing Data Types**: The data types for the columns Release Year, Votes, and Gross were converted to float for better analytical purposes.
        - **Release Year**: Although initially represented as an object, converting this column to int enables numerical operations and analyses, such as year-based trend analyses.
        - **Votes and Gross**: Similarly, these columns were also converted to float to facilitate calculations, such as averaging, aggregating, and comparing values effectively. This transformation enhances the overall analytical capability of the dataset, allowing for more complex data manipulations and visualizations.

Feature Engineering

Feature engineering is a crucial step in the data preprocessing phase, where raw data is transformed into features that better represent the underlying problem, improving the performance of machine learning models. This process involves creating new variables or modifying existing ones to enhance the dataset's analytical capabilities. By thoughtfully engineering features, we can uncover hidden patterns and relationships that may not be immediately apparent in the original data.

In the context of the IMDB Top 2000 Movies dataset, the following features were engineered:

1. **Single Genre Column**:
    - **Creation of Genre Column**: A new column named Single Genre was created to contain only the first genre from the Genre column.
        - **Process**: This was achieved by splitting the original genre string using the split(',') method, which separates genres listed in the original format.
        - **Purpose**: By isolating the primary genre, this feature simplifies the analysis, allowing for more straightforward categorization and comparison of movies based on their most dominant genre. This change facilitates analyses such as genre-based popularity and trends without the complexity of multiple genres.
2. **Movie Type Classification**:
    - **Type Column Creation**: Another new column named Type was introduced to classify movies as either "Old Movie" or "Latest Movie" based on their release year.
        - **Threshold Year**: The year 2005 was selected as the cutoff, where movies released in 2005 and earlier are labeled as "Old Movies," while those released after 2005 are classified as "Latest Movies."

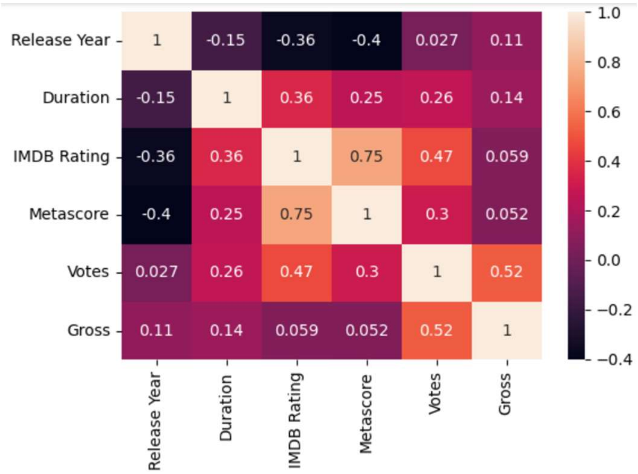
- **Purpose**: This classification aids in segmenting the dataset for analysis, enabling comparisons between older films and more recent releases. It provides a clearer perspective on trends, audience preferences, and changes in the film industry over time.

DATA ANALYSIS

Statistical Analysis of Movie Dataset

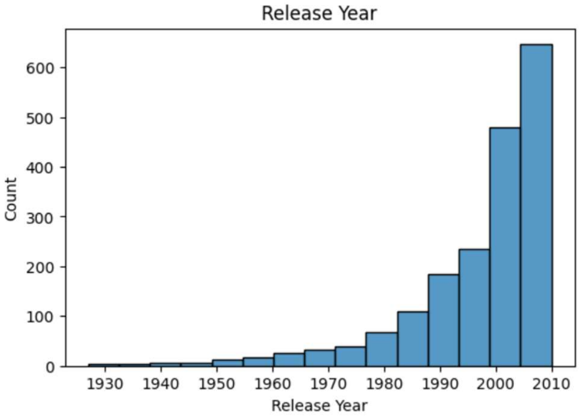
The following analysis is based on the summary statistics of the movie dataset, including **Release Year**, **Duration**, **IMDB Rating**, **Metascore**, **Votes**, and **Gross**. Each metric is examined to provide insights into the characteristics and trends of the movies in the dataset.

Correlation Heatmap



1. Release Year

- **Mean**: 1997
- **Range**: From 1927 to 2010
- **Median**: 2001

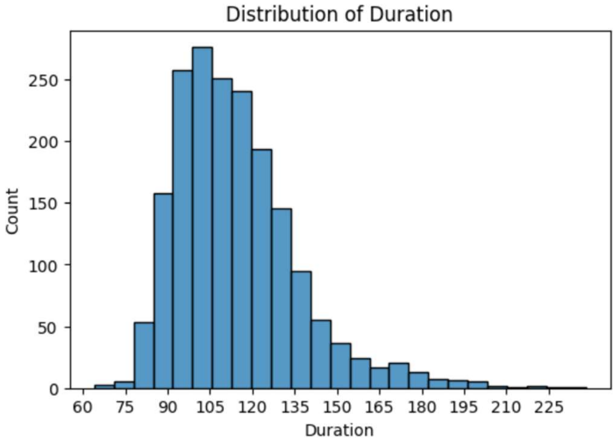


Insight:
The average release year of the movies is 1997, with a notable representation of films from the late 1990s and early 2000s. The earliest film dates back to 1927, indicating the dataset includes historical films, while the latest films extend to 2010, showing a wide temporal range in movie releases.

---

2. Duration

- **Mean**: 114.02 minutes
- **Range**: From 64 to 238 minutes
- **Median**: 110 minutes

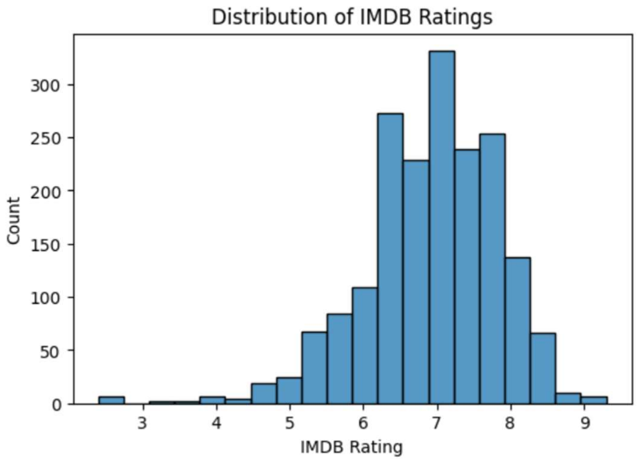


Insight:
The average duration of movies is approximately 114 minutes, suggesting that most films in the dataset are of standard length, typical for feature films. However, the longest movie runs for 238 minutes, indicating the presence of epics or comprehensive narratives, while the shortest at 64 minutes might represent TV movies or shorter formats.

---

3. IMDB Rating

- **Mean**: 6.92
- **Range**: From 2.4 to 9.3
- **Median**: 7.0



Insight:
The average IMDB rating is 6.92, which is a decent score, suggesting that the majority of films are generally well-received. The presence of the minimum rating at 2.4 points to some films that may not have been well-executed or received negatively, while the highest rating of 9.3 indicates several highly acclaimed films in the dataset.

---

4. Metascore

- **Mean**: 60.6
- **Range**: From 9 to 100
- **Median**: 61

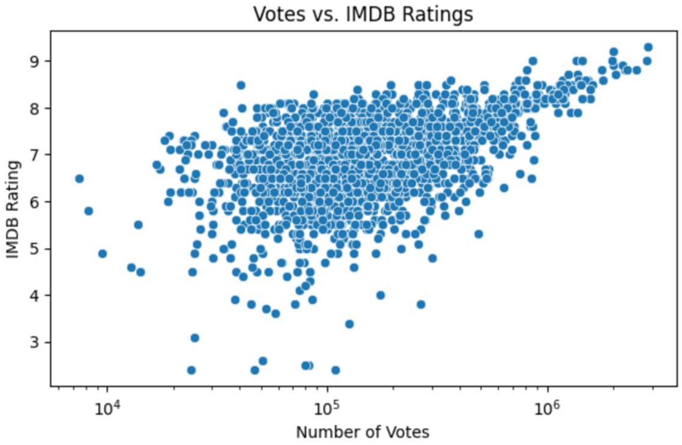
Insight:
With a mean Metascore of 60.6, it suggests that films have an average critical reception leaning towards favorable, but not outstanding. The range indicates variability, with some films being critically panned (minimum 9) while others have achieved perfection (100). This variation reflects the differing tastes of audiences and critics.

---

5. Votes

- **Mean**: 232,492
- **Range**: From 7,442 to 2,875,249
- **Median**: 140,924

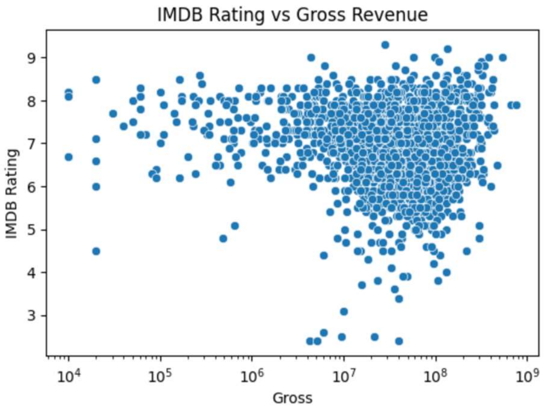
Insight:
The average number of votes indicates significant audience engagement, with the highest number of votes reaching nearly 2.9 million. This high vote count for certain movies signifies enduring popularity. However, the minimum number of votes (7,442) suggests that some movies have a much lower recognition or have been less circulated.



6. Gross Revenue

- **Mean**: $67,056,055
- **Range**: From $0 to $760,510,000
- **Median**: $45,750,000

Insight:
The mean gross revenue reflects significant box office earnings for many films, although the minimum gross of $0 raises questions about distribution or box office success. The highest gross of $760.5 million highlights blockbuster hits that resonate well with audiences. The median gross ($45.75 million) suggests that while many films do well, a sizable number may not achieve blockbuster status.



Analysis of Key Movie Metrics

1. Movies with the Lowest and Highest Votes

- **Lowest Votes:**
  - **Camelot**: 7,442 votes
- **Highest Votes:**

- **The Shawshank Redemption**: 2,875,249 votes

Insight:
The dramatic difference in the number of votes indicates that "The Shawshank Redemption" is not only a highly popular film among viewers but also stands as a classic with enduring appeal. In contrast, "Camelot" may be less well-known or appreciated, possibly due to its niche genre or older release.

---

2. Movies with the Lowest and Highest Gross Revenue

- **Lowest Gross:**
  - **The Condemned**: $0.0
- **Highest Gross:**
  - **Avatar**: $760,510,000.0

Insight:
"Avatar" leads significantly in gross revenue, reflecting its massive box office success and global appeal. In contrast, "The Condemned" earning $0 may indicate a limited release, a lack of distribution, or poor reception. Such stark differences highlight the importance of marketing and distribution in a film's financial success.

---

3. Movies with the Earliest Release Year

- **Earliest Release:**
  - **Metropolis**: 1927

Insight:
The presence of "Metropolis," a silent film from 1927, demonstrates the historical significance and evolution of cinema. Understanding how films from different eras are perceived can provide valuable context for analyzing contemporary preferences.

---

4. Top 10 Movies by Gross Revenue

Movie Name	**Director**	**Cast**	**Gross**
Avatar	James Cameron	Sam Worthington	$760,510,000.0
Titanic	James Cameron	Leonardo DiCaprio	$659,330,000.0
The Dark Knight	Christopher Nolan	Christian Bale	$534,860,000.0
Star Wars: Episode I - The Phantom Menace	George Lucas	Ewan McGregor	$474,540,000.0
Shrek 2	Andrew Adamson	Kelly Asbury	$436,470,000.0
E.T. the Extra-Terrestrial	Steven Spielberg	Henry Thomas	$435,110,000.0

Movie Name	Director	Cast	Gross
Pirates of the Caribbean: Dead Man's Chest	Gore Verbinski	Johnny Depp	$423,320,000.0
The Lion King	Roger Allers	Rob Minkoff	$422,780,000.0
Toy Story 3	Lee Unkrich	Tom Hanks	$415,000,000.0
Spider-Man	Sam Raimi	Tobey Maguire	$403,710,000.0

Insight:
The dominance of films like "Avatar" and "Titanic" shows the impact of visual effects and compelling storytelling on box office performance. Both films by James Cameron highlight his skill in creating blockbuster hits. The variety in genres among the top-grossing films illustrates the diverse tastes of moviegoers.

---

5. Top 10 Movies by Votes

Movie Name	Director	Cast	Votes
The Shawshank Redemption	Frank Darabont	Tim Robbins	2,875,249
The Dark Knight	Christopher Nolan	Christian Bale	2,857,781
Inception	Christopher Nolan	Leonardo DiCaprio	2,538,581
Fight Club	David Fincher	Brad Pitt	2,311,174
Forrest Gump	Robert Zemeckis	Tom Hanks	2,245,598
Pulp Fiction	Quentin Tarantino	John Travolta	2,209,021
The Matrix	Lana Wachowski	Keanu Reeves	2,043,258
The Godfather	Francis Ford Coppola	Marlon Brando	2,002,655
The Lord of the Rings: The Fellowship of the Ring	Peter Jackson	Elijah Wood	1,998,243
The Lord of the Rings: The Return of the King	Peter Jackson	Elijah Wood	1,970,311

Insight:
The number of votes correlates with the films' lasting popularity and cultural significance. "The Shawshank Redemption" leads by a substantial margin, showcasing its status as a favorite among audiences. The presence of Christopher Nolan’s works indicates his impact on modern filmmaking and audience engagement.

---

6. Top 10 Movies by IMDB Rating

Movie Name	**Director**	**Cast**	**IMDB Rating**
The Shawshank Redemption	Frank Darabont	Tim Robbins	9.3
The Godfather	Francis Ford Coppola	Marlon Brando	9.2
The Godfather Part II	Francis Ford Coppola	Al Pacino	9.0
12 Angry Men	Sidney Lumet	Henry Fonda	9.0
Schindler's List	Steven Spielberg	Liam Neeson	9.0
The Lord of the Rings: The Return of the King	Peter Jackson	Elijah Wood	9.0
The Dark Knight	Christopher Nolan	Christian Bale	9.0
Pulp Fiction	Quentin Tarantino	John Travolta	8.9
The Lord of the Rings: The Fellowship of the Ring	Peter Jackson	Elijah Wood	8.9
Il buono, il brutto, il cattivo	Sergio Leone	Clint Eastwood	8.8

Insight:
The top-rated films reflect critical acclaim and audience appreciation. "The Shawshank Redemption" stands out as the highest-rated film, emphasizing its storytelling and character development. The representation of various directors, including Francis Ford Coppola and Christopher Nolan, underscores their significant contributions to cinematic art.

---

7. Top 100 Movies by Metascore

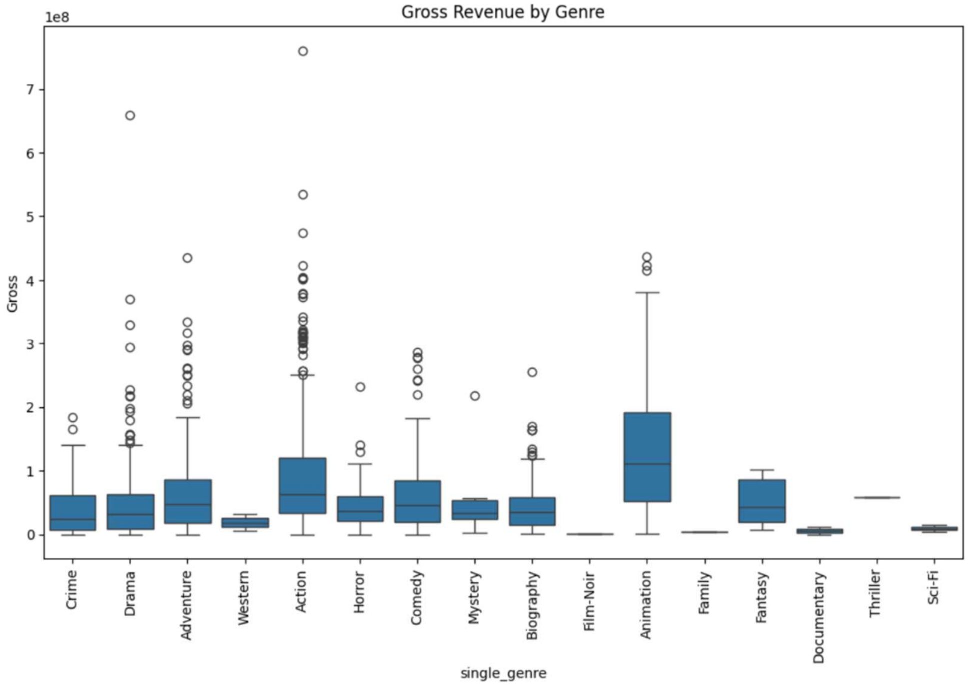
The dataset also reveals the top films according to Metascore, which highlights critical reception. Noteworthy mentions include:

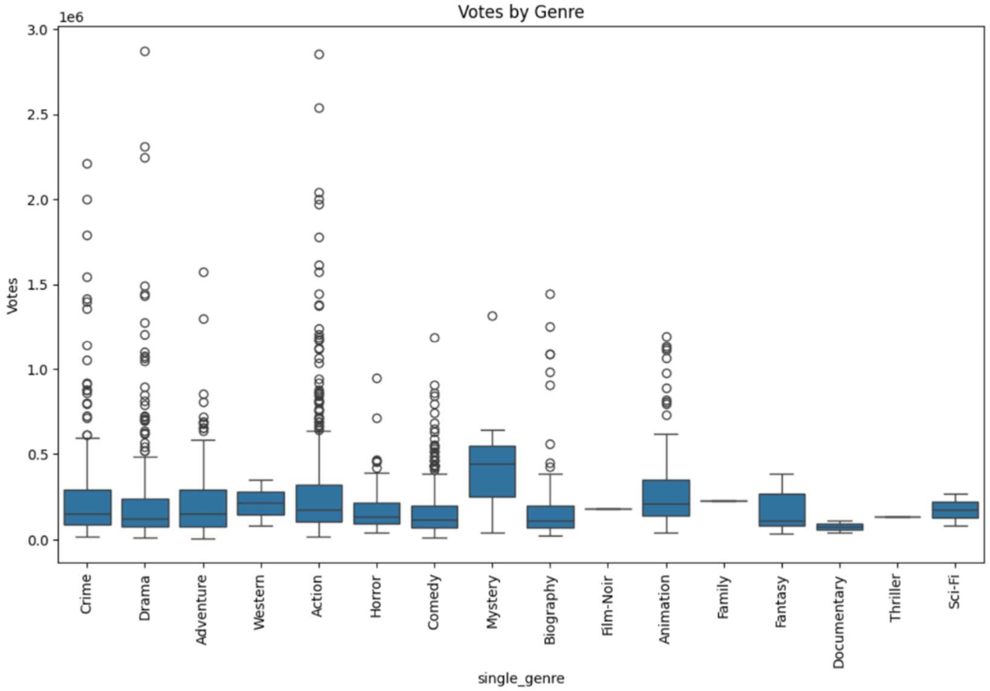
- **The Godfather**: 100
- **Lawrence of Arabia**: 100
- **Notorious**: 100
- **Vertigo**: 100

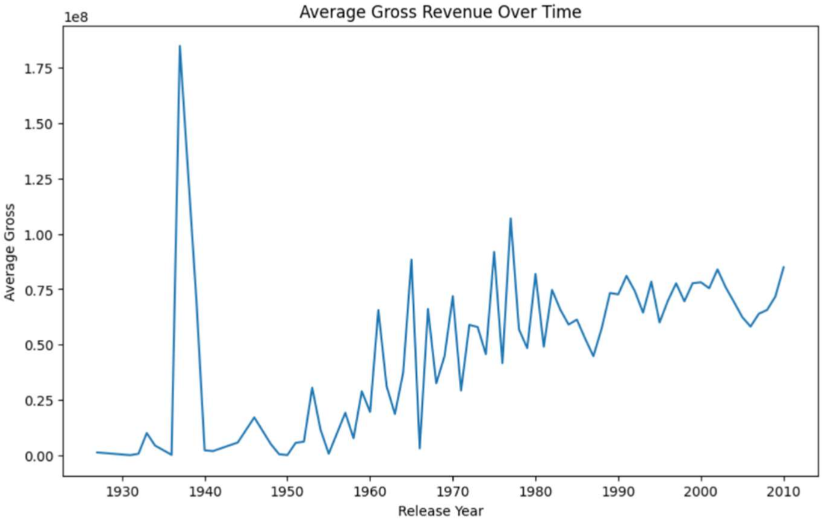
Insight:
The films with perfect Metascores signify exceptional craftsmanship and storytelling, indicating that these movies are not only commercially successful but also critically revered. This distinction can be crucial for film studies and recommendations, emphasizing the quality of content over mere box office success.

Plots Used in Analysis

In the analysis of the IMDb Top 2000 Movies dataset, several plots can be beneficial for visualizing and understanding the data more intuitively. Here’s a short explanation of the types of plots typically used and their necessity in analysis:


Gross Revenue by Genre
1e8
Gross
0
1
2
3
4
5
6
7
Crime
Drama
Adventure
Western
Action
Horror
Comedy
Mystery
Biography
Film-Noir
Animation
Family
Fantasy
Documentary
Thriller
Sci-Fi
single_genre

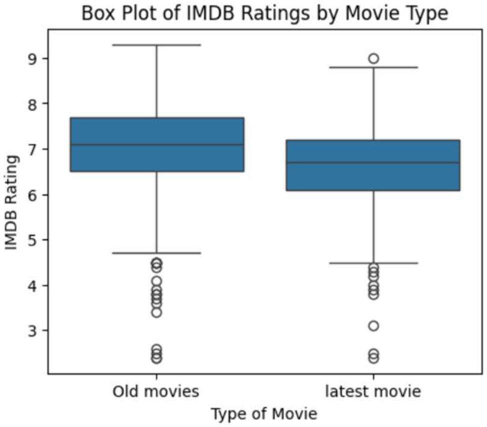

Votes by Genre
1e6
Votes
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Crime
Drama
Adventure
Western
Action
Horror
Comedy
Mystery
Biography
Film-Noir
Animation
Family
Fantasy
Documentary
Thriller
Sci-Fi
single_genre



1. Bar Plots

- **Purpose:** To show the frequency or count of categorical variables, such as the number of movies per genre, director, or cast.
- **Need:** Bar plots help in quickly identifying the most common genres, directors, or actors in the dataset, enabling comparisons between different categories.

2.Box Plots



- **Purpose:** To visualize the distribution of numerical data (e.g., Duration, IMDB Rating, Gross) and highlight outliers.
- **Need:** Box plots are essential for understanding the central tendency and variability of the data, as well as spotting any extreme values that may influence analysis.

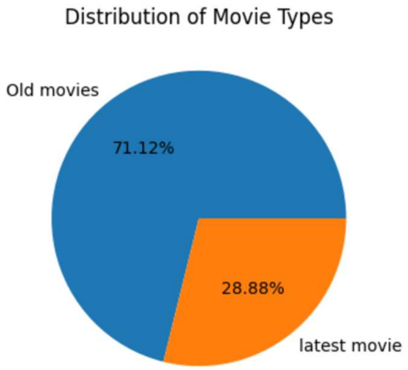
3. Histograms

- **Purpose:** To show the distribution of a single numerical variable, such as IMDB Ratings or Gross revenue.
- **Need:** Histograms allow for visual assessment of the distribution shape (normal, skewed, etc.), helping to identify trends and patterns within the dataset.

4. Scatter Plots

- **Purpose:** To explore relationships between two numerical variables (e.g., IMDB Rating vs. Gross).
- **Need:** Scatter plots help in identifying correlations between variables, providing insights into how one variable may affect another, such as whether higher-rated movies tend to have higher gross earnings.

5.Pie Charts



- **Purpose:** To display the proportions of categories within a whole, such as the percentage of movies by genre.
- **Need:** Pie charts can give a quick overview of the distribution of categories, making it easy to visualize which genres dominate the dataset.

6.Line Plots

- **Purpose:** To display trends or changes over time, such as the progression of average IMDb ratings or gross earnings by release year.
- **Need:** Line plots are essential for understanding temporal changes and patterns, showing how variables like movie ratings or box office revenue fluctuate over time.

7. Heatmaps

- **Purpose:** To visualize the correlation matrix between numerical variables, indicating the strength and direction of relationships among features like IMDb Rating, Metascore, Votes, Duration, and Gross.
- **Need:** Heatmaps highlight key interactions and dependencies between variables, making it easier to identify factors that may influence movie success, such as the correlation between votes and gross revenue.

Conclusion

The exploratory analysis of the IMDb Top 2000 Movies dataset has provided valuable insights into movie trends, genre distributions, and influential factors in cinematic success. Key findings from the analysis include:

1. **Time-Based Trends**: There is a significant difference between older and newer movies in terms of gross earnings, ratings, and votes. Older movies, especially those pre-2005, often hold high IMDb ratings and Metascores, highlighting their enduring appeal and critical acclaim.

2. **Genre Insights**: The dataset revealed that genres like Action, Comedy, and Drama are among the most prevalent, with genre-specific patterns in both popularity and audience engagement. The addition of a simplified genre column allowed for a clearer analysis of primary genres and their impacts.

3. **Top-Ranked Directors and Cast**: Directors like Steven Spielberg, Ridley Scott, and Christopher Nolan frequently appear in the dataset, indicating their consistent ability to create popular and well-regarded films. Similarly, actors such as Tom Cruise and Robert De Niro demonstrate high engagement across multiple titles, suggesting a strong fan base and impact on movie success.

4. **Highest Grossing and Most Voted Movies**: The analysis identified films such as *Avatar*, *The Dark Knight*, and *Titanic* as high-grossing, popular choices among audiences. Additionally, movies with high vote counts generally have strong IMDb ratings, reinforcing the idea that audience engagement often correlates with higher quality perception.

5. **Variable Correlations**: Notable correlations were found between IMDb Ratings and Metascores, as well as between Votes and Gross earnings. These relationships suggest that critical acclaim and audience popularity often go hand-in-hand with commercial success, highlighting the interconnected nature of these factors in determining a movie’s impact.

This analysis lays a foundational understanding of the dataset, enabling informed hypotheses for future research or machine learning applications. By leveraging these insights, data-driven strategies can be formulated to explore user preferences further, refine recommendation systems, or develop predictive models to anticipate movie success.