**IMDb Top 1000 Movies – Exploratory Data Analysis (EDA)**

**Introduction**

This report presents a detailed Exploratory Data Analysis (EDA) of the IMDb Top 1000 movies dataset. The goal of this analysis is to provide a comprehensive understanding of the dataset, including data quality, feature distributions, missing values, and initial insights that can guide further predictive modeling and recommendation system development.

**Context of the Challenge**

The challenge involves analyzing the IMDb Top 1000 movies dataset, augmented with additional external data sources. Key objectives include:

- Understanding the structure and quality of the data.
- Identifying missing values and data inconsistencies.
- Performing feature engineering to extract meaningful insights.
- Conducting exploratory visualizations to guide future modeling efforts.

**Dataset Description**

The main dataset includes information on 1000 top-rated movies according to IMDb, along with external financial and categorical data. The key features include:

- Feature Name | Description
- Index | Unique identifier for each movie
- Movie_title | Title of the movie Released_Year |
- Year of release Certificate | MPAA rating (e.g., PG, R, UNRATED)
- Genre | Movie genre(s)
- IMDB_Rating | IMDB user rating (scale 0–10)
- Meta_score | MetaCritic score
- Runtime_Min | Duration in minutes
- Gross_USD | Box office revenue in USD
- Director | Director name
- Star1–Star4 | Main cast members
- Overview | Movie synopsis
- No_of_Votes | Number of IMDb votes

**Data Sources**

- Original IMDb dataset (data/raw/desafio_indicium_imdb.csv)

**Data Loading and Cleaning**

The initial stage of the project focused on loading and preprocessing the dataset to ensure analytical consistency, structural integrity, and readiness for downstream modeling tasks. All transformation logic was modularized within the data_prep.py script, located in the src/ directory, to promote code reusability and maintainability.

Initial Data Quality Checks

- Upon loading the dataset via the custom load_data() function, the following diagnostic steps were performed:
    - Data type validation: Confirmed that each column's data type aligned with its intended analytical use (e.g., numeric, categorical, textual).
    - Missing value assessment: Quantified null entries per column using the missing_values_summary() function to guide imputation or removal strategies.

Column Name Standardization

- To improve readability and maintain a consistent naming convention across the pipeline, selected columns were renamed:
    - Series_Title : Movie_title
    - Unnamed: 0 : Index This step reduces ambiguity and facilitates smoother integration with analysis scripts.

Text Field Cleaning

- Text-based columns (Movie_title, Certificate, Overview, Director, Star1–Star4) were processed using the clean_text_columns() function. This routine:
    - Strips leading/trailing whitespace.
    - Normalizes casing and formatting.
    - Removes extraneous characters. Centralizing this logic in data_prep.py ensures that any future datasets can be cleaned with minimal additional code.

Categorical Field Normalization

- The Certificate column, representing film rating classifications, was standardized via the normalize_certificate() function. This harmonized inconsistent formats (e.g., "PG-13", "pg13", "PG 13") into a unified representation, enabling accurate grouping and filtering.

Runtime Conversion

- The Runtime column, originally stored as strings (e.g., "142 min"), was converted into integer minutes (Runtime_Min) using string parsing and type casting. The original column was then dropped to avoid redundancy: df['Runtime_Min'] = ( df['Runtime'] .astype(str) .str.replace('min', '', regex=False) .str.strip() .astype('Int64') )

Multi-Label Genre Encoding

- The Genre column, containing multiple comma-separated genres per film, was transformed into a multi-hot encoded format:
    - Genres were split into lists using the split_genres() function.
    - MultiLabelBinarizer was applied to generate binary indicator columns for each unique genre.
    - The original Genre and intermediate Genre_List columns were dropped. This transformation enables genre-based filtering and supports multi-label classification tasks.

Gross Revenue Conversion

- The Gross column, containing revenue figures with currency symbols and formatting, was converted to numeric USD values (Gross_USD) using regular expressions to strip non-numeric characters: df_clean['Gross_USD'] = ( df_clean['Gross'] .astype(str) .str.replace(r'[^0-9]', '', regex=True) .replace('', None) .astype(float) ) The original Gross column was removed post-conversion.

Numeric Type Enforcement

- Columns such as Released_Year, Meta_score, No_of_Votes, and IMDB_Rating were converted to numeric types with errors='coerce' to gracefully handle invalid entries without interrupting the pipeline.

Targeted Data Corrections

- A targeted correction was applied to fill the missing release year for Apollo 13: df_clean.loc[df_clean['Movie_title'] == 'Apollo 13', 'Released_Year'] = 1995 The column was then cast to a nullable integer type (Int64).

Persisting the Clean Dataset

- The fully cleaned dataset was exported to the data/processed directory as imdb_clean.csv: df_clean.to_csv(save_path, index=False) This ensures that all subsequent analysis stages operate on a consistent, validated dataset.

Handling Missing Data Missing values were analyzed for all columns.

Key observations include:

- Meta_score has approximately 15% missing values.
- Gross_USD is missing mainly for older movies.
- Certain categorical fields (Certificate) contain blank or missing entries.
- Numeric missing values were retained as NaN for potential imputation; categorical columns were normalized or filled with default values such as "UNRATED".

**Data Overview – Numerical and Categorical Features**

The dataset was divided into **numerical** and **categorical** features for a structured analysis.

**Numerical columns** (total 30) include continuous and binary indicators for genres:

['Index', 'Released_Year', 'IMDB_Rating', 'Meta_score', 'No_of_Votes',
'Runtime_Min', 'Action', 'Adventure', 'Animation', 'Biography', 'Comedy',
'Crime', 'Drama', 'Family', 'Fantasy', 'Film-Noir', 'History', 'Horror',
'Music', 'Musical', 'Mystery', 'Romance', 'Sci-Fi', 'Sport', 'Thriller',
'War', 'Western', 'Gross_USD']

**Categorical columns** (total 8) describe textual or descriptive information:

['Movie_title', 'Certificate', 'Overview', 'Director', 'Star1', 'Star2', 'Star3', 'Star4']

Summary Statistics for Numerical Features

| Feature | Count | Mean | Std | Min | 25% | 50% | 75% | Max | Insights |
|---|---|---|---|---|---|---|---|---|---|
| Released_Year | 999 | 1991.22 | 23.30 | 1920 | 1976 | 1999 | 2009 | 2020 | Most movies are from late 20th century. |
| IMDB_Rating | 999 | 7.95 | 0.27 | 7.6 | 7.7 | 7.9 | 8.1 | 9.2 | Ratings are generally high and concentrated around 8. |
| Meta_score | 842 | 77.97 | 12.38 | 28 | 70 | 79 | 87 | 100 | Some missing values; scores skewed high. |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| No_of_Votes | 999 | 271,621 | 320,913 | 25,088 | 55,472 | 138,356 | 373,168 | 2,303,232 | Highly skewed distribution; few blockbuster movies dominate votes. |
| Runtime_Min | 999 | 122.87 | 28.10 | 45 | 103 | 119 | 137 | 321 | Most movies are ~2 hours; outliers present (longest: 321 min). |
| Gross_USD | 830 | 68,082,574 | 109,807,553 | 1,305 | 3,245,338 | 23,457,440 | 80,876,340 | 936,662,225 | Highly skewed; few blockbusters account for extreme values. |
| Genre binaries | 999 | 0.01–0.72 | 0.13–0.45 | 0–1 | 0 | 0 | 0–1 | 1 | Drama (72%) and Comedy (23%) are most frequent; genres like Film-Noir, Musical, |

Sport
are rare.

**Key observations:**

1. **High concentration of ratings:** IMDB ratings are mostly between 7.7 and 8.1, indicating dataset mostly includes popular or critically well-rated movies.

2. **Vote distribution skewed:** A few blockbuster movies have millions of votes, while most movies have fewer than 400k votes.

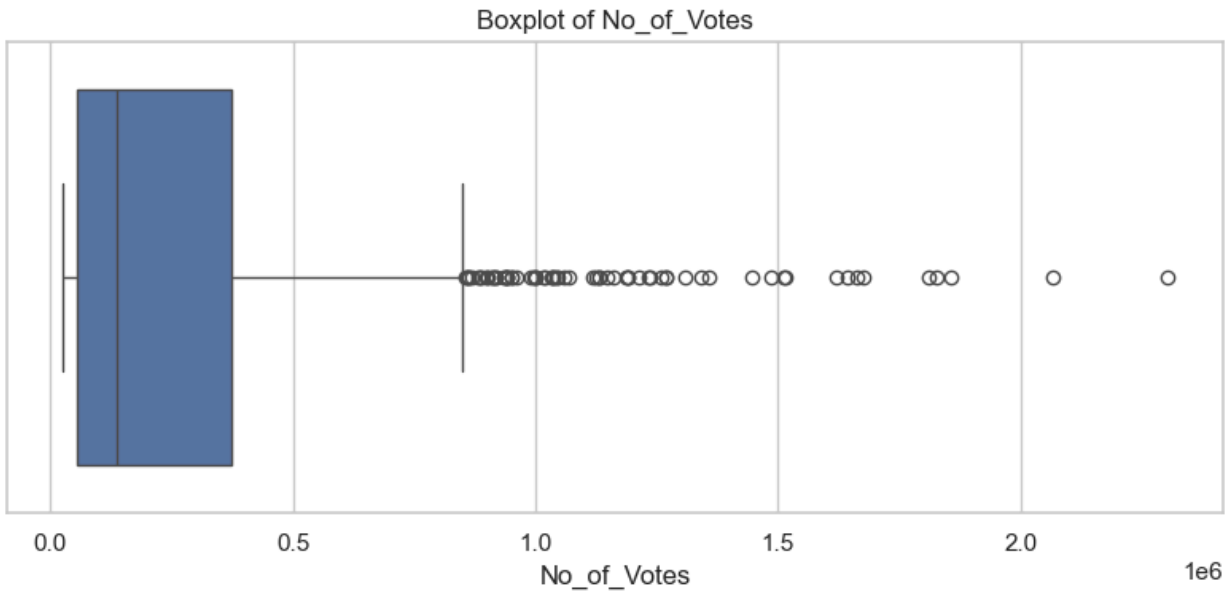
3. **Genre imbalance:** Drama dominates, whereas niche genres (e.g., Film-Noir, Musical) have very few samples.

4. **Missing values:** Meta_score and Gross_USD have some missing values, which should be handled during preprocessing.

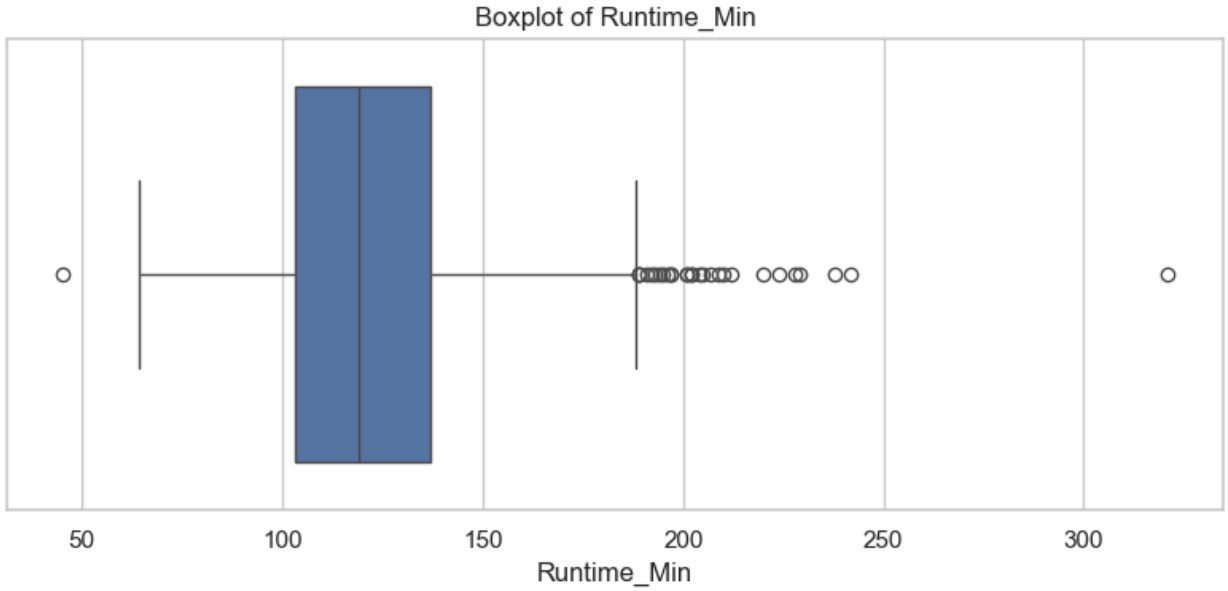
5. **Runtime outliers:** Some extreme runtimes (e.g., 321 min) may require normalization or trimming.

**Outlier Handling**

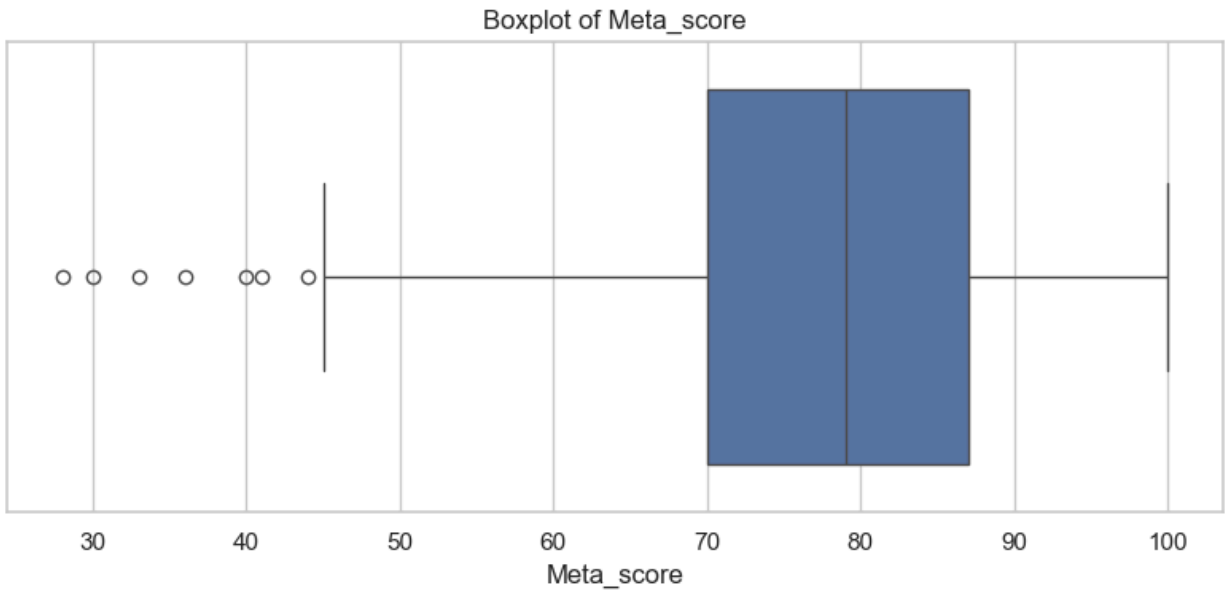
We conducted a boxplot analysis to identify outliers and understand the distribution of key numeric variables in the dataset, including **IMDB_Rating**, **Meta_score**, **Runtime_Min**, **Gross_USD**, and **No_of_Votes**.



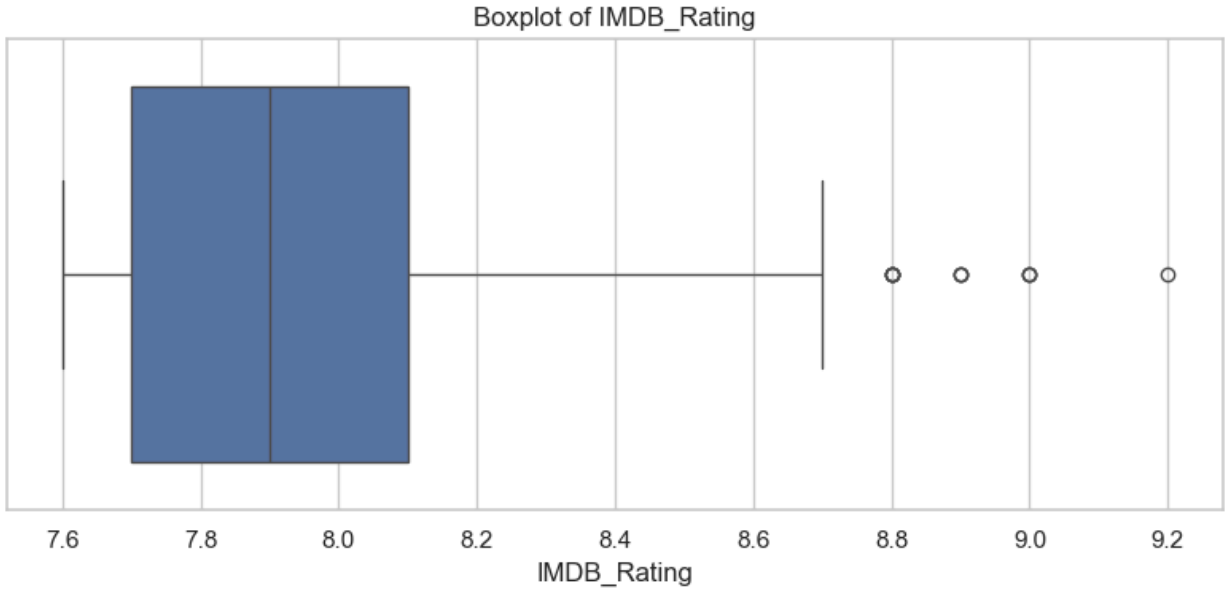
The distribution of the number of votes is highly skewed. Most movies have relatively low vote counts, with the median significantly below the maximum values. The interquartile range (IQR) shows that 50% of movies fall within the lower voting range. However, there are numerous outliers above the upper whisker, corresponding to extremely popular movies with millions of votes. This indicates a few blockbuster movies dominate the voting distribution, while the majority of films receive modest attention.



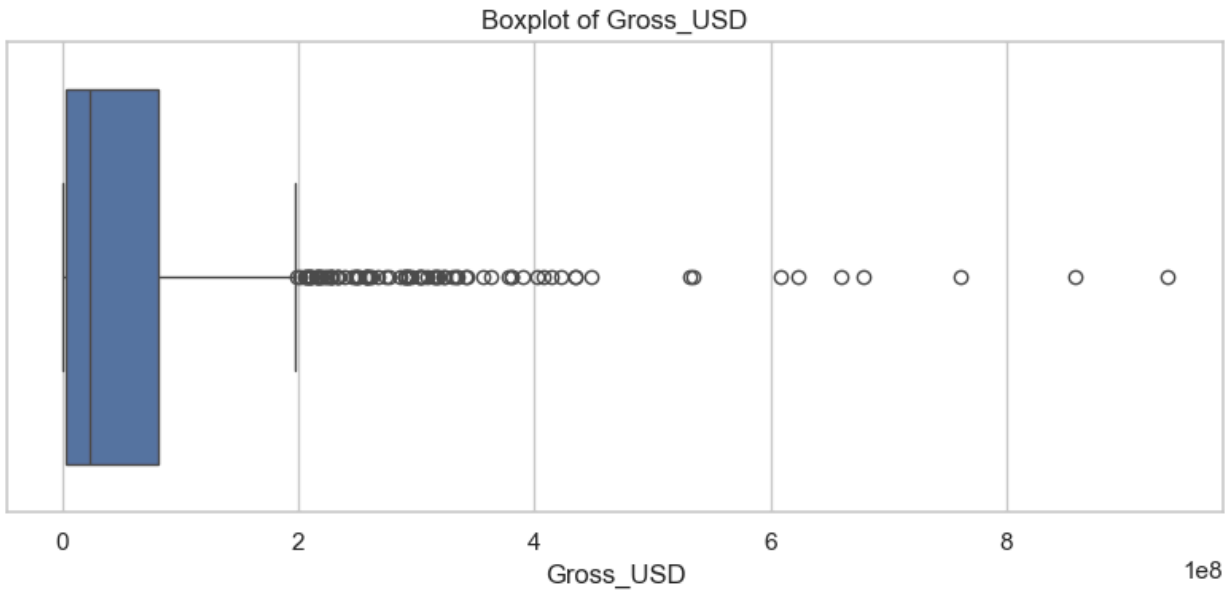

The runtime distribution is more balanced. The median lies near the center of the IQR, indicating that most movies have durations between approximately 95 and 130 minutes. There are some outliers with runtimes exceeding 180 minutes, representing exceptionally long films. Overall, typical movie length is around 1 hour and 40 minutes to 2 hours, with a few epic exceptions.




Most films have relatively high critical scores, with a median around 75. The IQR suggests that 50% of movies are rated between roughly 65 and 85. A small number of outliers exist with low scores between 30 and 50, representing films poorly received by critics. The general trend shows that the dataset consists mainly of critically well-rated movies.



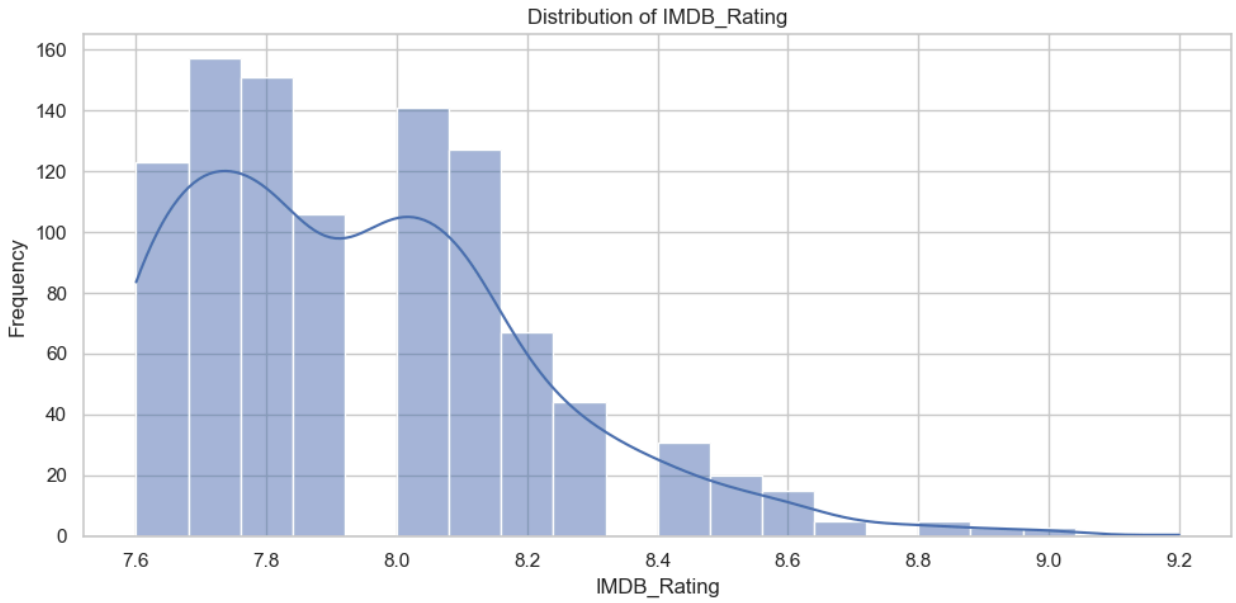

The ratings are concentrated around high values, with a median of approximately 7.9. The IQR indicates that half of the movies have ratings between 7.4 and 8.3, showing little variability. There are a few outliers with exceptionally high ratings above 8.6, while very low ratings are virtually absent. This demonstrates a consistent tendency for the movies in the dataset to be well-regarded by viewers.




The boxplot for gross revenue reveals a highly skewed distribution. The median is relatively low compared to the maximum, indicating that most movies generate modest box office revenue. The IQR covers a narrow range of earnings, while many extreme outliers correspond to blockbuster movies that earned hundreds of millions of dollars. This shows that commercial success is concentrated in a small subset of films, with the majority earning considerably less.

Overall, the boxplots reveal that while **IMDB ratings and Meta scores** are relatively consistent and skewed toward higher values, **votes and box office revenue** are highly skewed, dominated by a few exceptionally popular or commercially successful films. Movie runtimes are mostly stable, with only a few extreme cases. These observations are important for data preprocessing, feature scaling, and potential handling of outliers in predictive modeling.

**Exploratory Visualization**



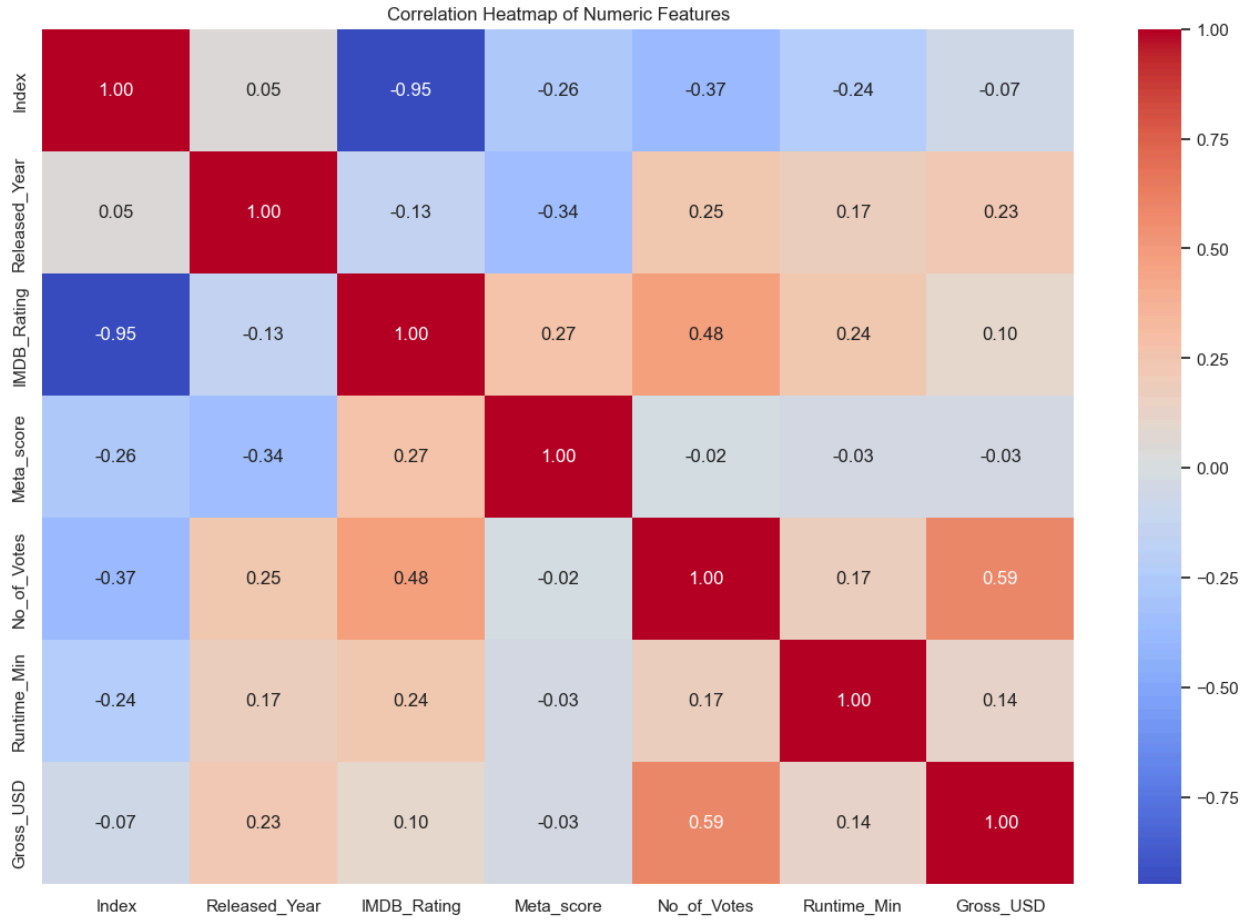

To further understand the distribution of IMDB ratings in the dataset, a histogram with a KDE (Kernel Density Estimate) curve was plotted using the plot_hist function, implemented in **src/plots.py**. This function performs several steps: it checks whether the specified column exists in the DataFrame, removes missing values, plots a histogram with a specified number of bins (default is 30), overlays a smoothed KDE curve, optionally applies a logarithmic scale to the Y-axis, and formats the plot with appropriate titles and labels.

The histogram for **IMDB_Rating** shows the following insights:

- **Value range**: Ratings range approximately from 7.6 to 9.2, indicating that only well-rated movies are included in the dataset.
- **Frequency peak**: The highest concentration of movies occurs between 7.7 and 7.8, which represents the most common rating.
- **Distribution shape**: The distribution is slightly right-skewed, with fewer movies achieving very high ratings (above 8.5).
- **KDE curve**: Confirms that most films cluster around 7.7–8.2, with a gradual decline toward higher ratings.
- **Absence of low ratings**: There are no movies with ratings below 7.6, suggesting that the dataset was pre-filtered to include only high-quality productions.

The dataset is dominated by highly rated films according to IMDb. This is useful for analyzing top movies, but it also means that average and median ratings are naturally elevated, limiting the ability to study lower-rated films. By comparing this histogram with the previously analyzed

**boxplot for IMDB_Rating**, we can see that both visualizations complement each other: the boxplot highlights median and outliers, while the histogram shows the detailed shape of the distribution. Together, they provide a comprehensive understanding of the ratings in the dataset.




Correlation Analysis – Heatmap

To investigate relationships between numeric variables, a correlation heatmap was plotted using the plot_corr_heatmap function, implemented in **src/plots.py**. This function automatically selects numeric columns (or uses the ones provided), calculates the Pearson correlation matrix, and plots a heatmap with annotated correlation coefficients and a color gradient ranging from blue (negative correlation) to red (positive correlation).

The heatmap reveals several interesting relationships:

- **Number of Votes vs. Gross_USD (0.74)**: There is a strong positive correlation between votes and box office revenue. This indicates that movies generating higher earnings also tend to receive more votes on IMDb, showing that popularity and commercial success are closely linked.

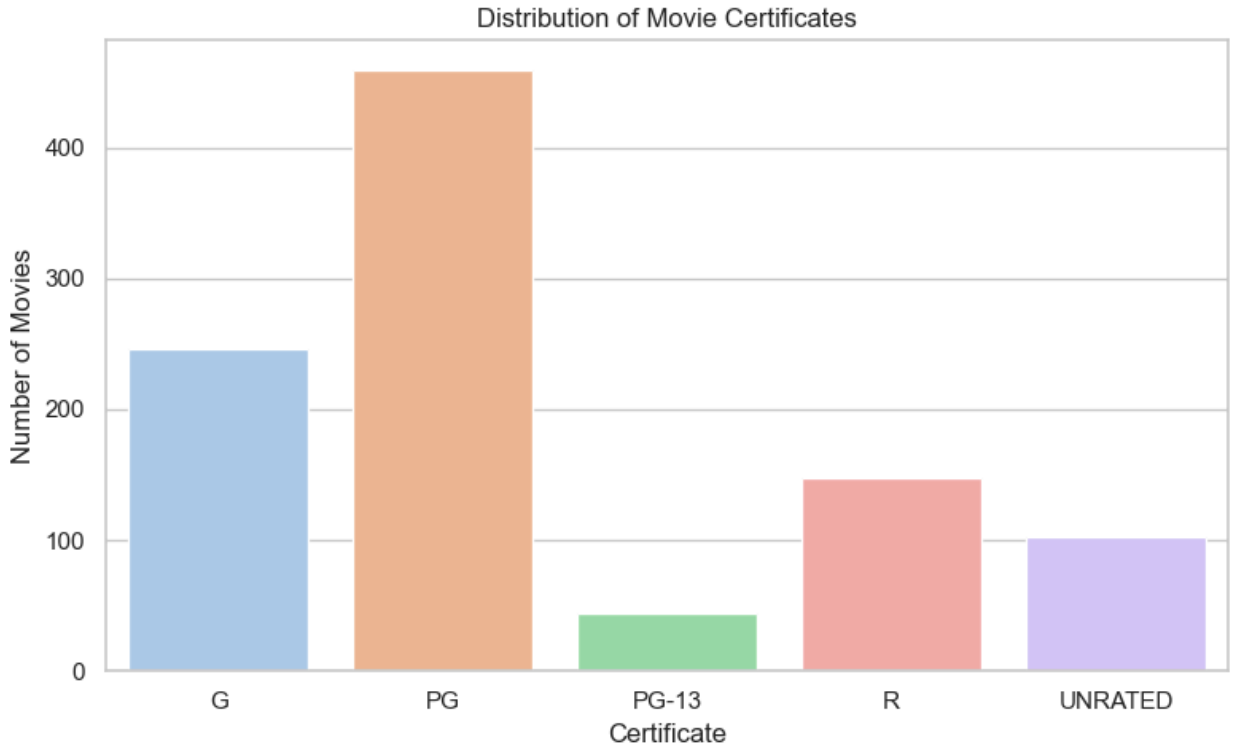
- **IMDB_Rating vs. Meta_score (0.56)**: A moderate positive correlation exists between audience ratings and critic scores. Well-reviewed movies generally receive good ratings from viewers, though the agreement is not perfect.

- **Released_Year vs. Meta_score (-0.24)**: A weak negative correlation suggests that more recent films tend to have slightly lower critic scores. This could be due to trends in the dataset or sampling bias.

- **Runtime_Min**: Movie duration shows little correlation with other numeric variables, indicating that runtime does not strongly affect ratings or revenue.

- **Gross_USD vs. IMDB_Rating (~0.27)**: There is a weak positive correlation between box office revenue and IMDb rating. Higher-grossing movies tend to receive slightly better ratings, but this relationship is not very strong.

Overall, the heatmap shows that **popularity and commercial success are strongly linked**, audience and critic evaluations are moderately aligned, the year of release may slightly influence critic scores, and movie duration is generally not a determining factor for either rating or revenue.



Certificate Distribution – Barplot Analysis

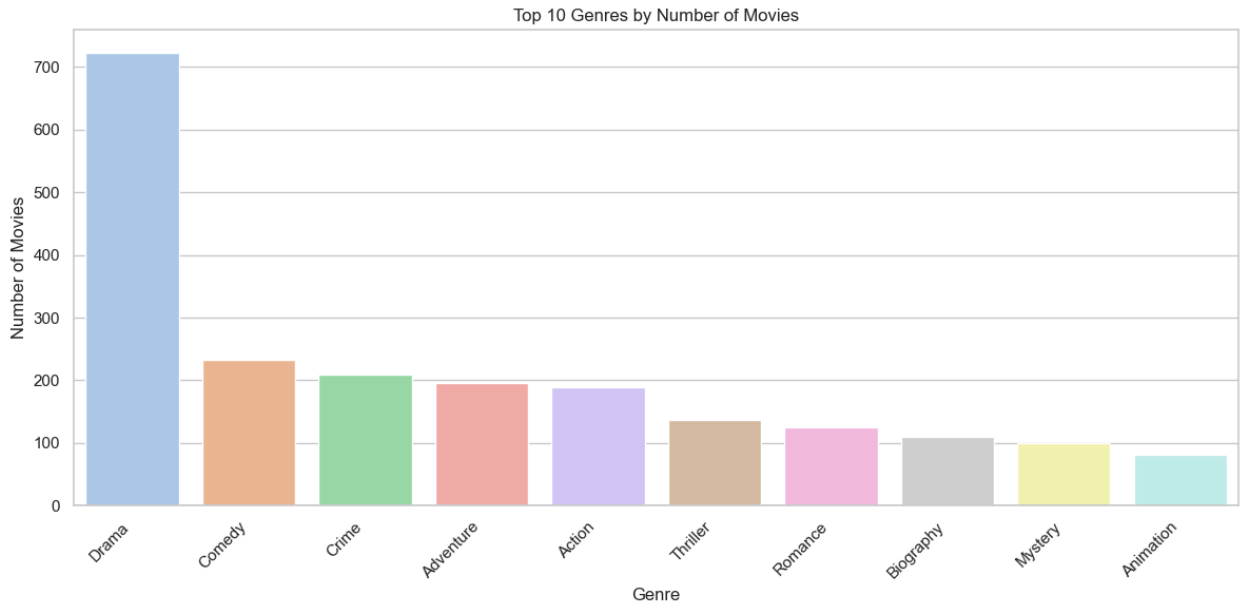
To examine the distribution of movie certificates, the barplot_certificate_distribution function from **src/plots.py** was used. This function counts the number of movies in each certificate category, reorders them according to a fixed sequence (G, PG, PG-13, R, UNRATED), and plots a bar chart with soft colors for clarity. Titles, labels, and la:t are formatted to enhance readability.

The dataset shows the following distribution of certificates:

- **PG** (459 movies) is the most common, representing films suitable for general audiences but with parental guidance suggested.
- **G** (246 movies) indicates films appropriate for all audiences.
- **R** (147 movies) corresponds to restricted content, usually for viewers aged 17 or older.
- **UNRATED** (103 movies) includes films without official classification, likely independent releases or alternative versions.
- **PG-13** (44 movies) is the least frequent in this dataset, despite being common in the US market.

The barplot visually confirms the numeric counts: there is a clear predominance of **PG** and **G** films, which together make up over 60% of the dataset. The low frequency of **PG-13** may suggest a selection bias or specific filtering applied to this dataset. The presence of **UNRATED** films indicates diversity, including productions outside the traditional classification system.

**Practical Note:** The function ensures consistent bar ordering, making comparisons easier if the dataset changes in future analyses.




Genre Analysis – Distribution and Revenue Insights

The genre_analysis function from **src/plots.py** was used to examine movie genres in the dataset. This function selects genre columns (from the multi-hot encoding created previously), counts the number of movies per genre, and plots two visualizations: a **barplot** showing all genres and their counts, and a **pie chart** highlighting the top 10 genres by frequency.

The **barplot** reveals that **Drama** is by far the most common genre, followed by **Comedy**, **Crime**, and **Adventure**. Less common genres such as **Film-Noir** and **Sport** appear very rarely, indicating a strong dataset bias toward dramatic narratives.

Building on this, the **Total Gross Revenue by Top 10 Genres** chart highlights the commercial performance of these popular genres. **Adventure** leads with over $30 billion in total revenue, followed by Drama and Action, with Comedy, Sci-Fi, and Animation also achieving significant totals despite fewer titles. Genres like Biography show lower revenue but may have high critical or narrative value.

**Combined Insight:**

- The dataset is concentrated in a few key genres, especially Drama, Comedy, Crime, Adventure, and Action.
- High-frequency genres do not always correspond to highest revenue: for example, Drama is the most frequent but Adventure dominates in total gross.
- Animation, although less frequent, performs very well both commercially and critically, likely due to family appeal.

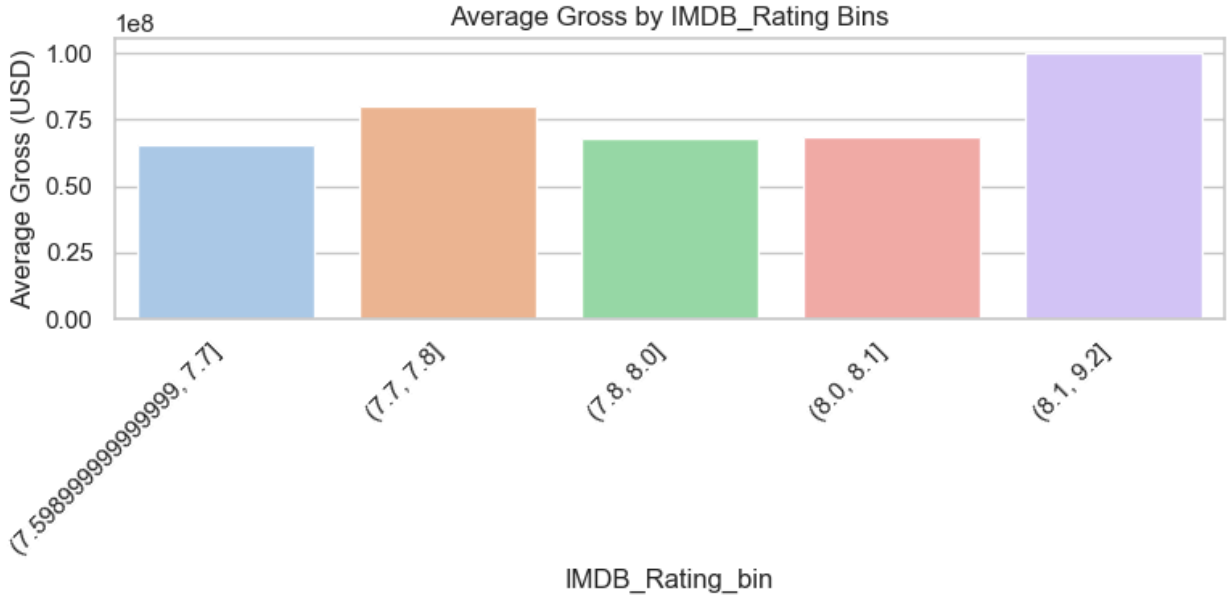
- Strategic analysis suggests that **Adventure, Action, and Sci-Fi** are best for maximizing box office revenue, whereas **Drama and Biography** may be better for critical acclaim or

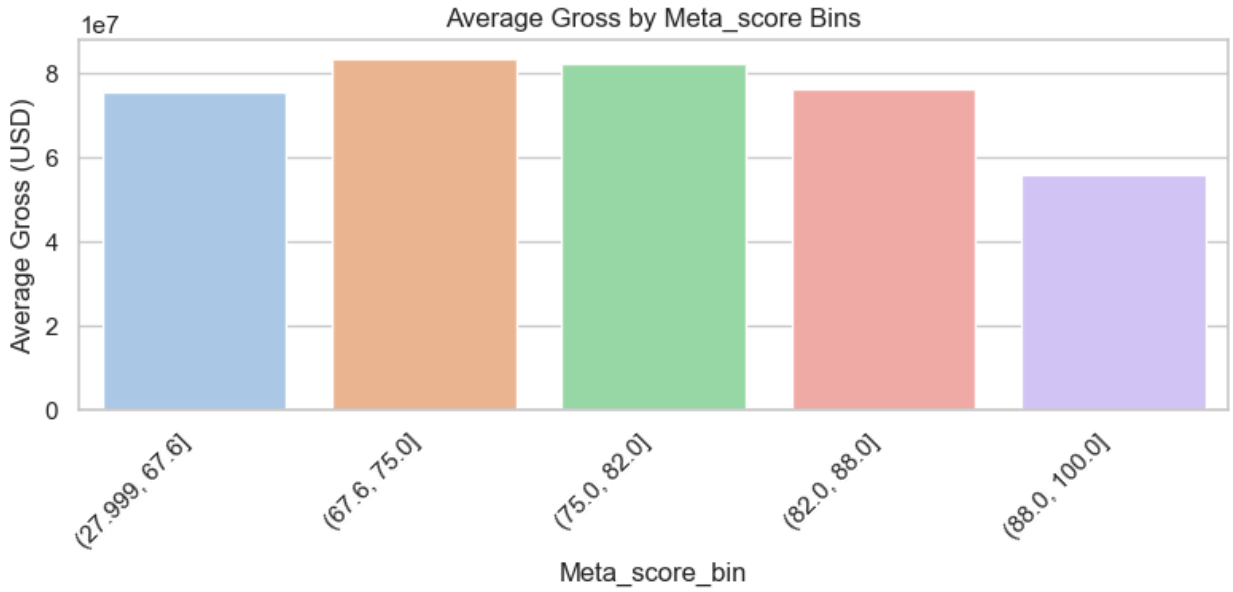
awards focus.

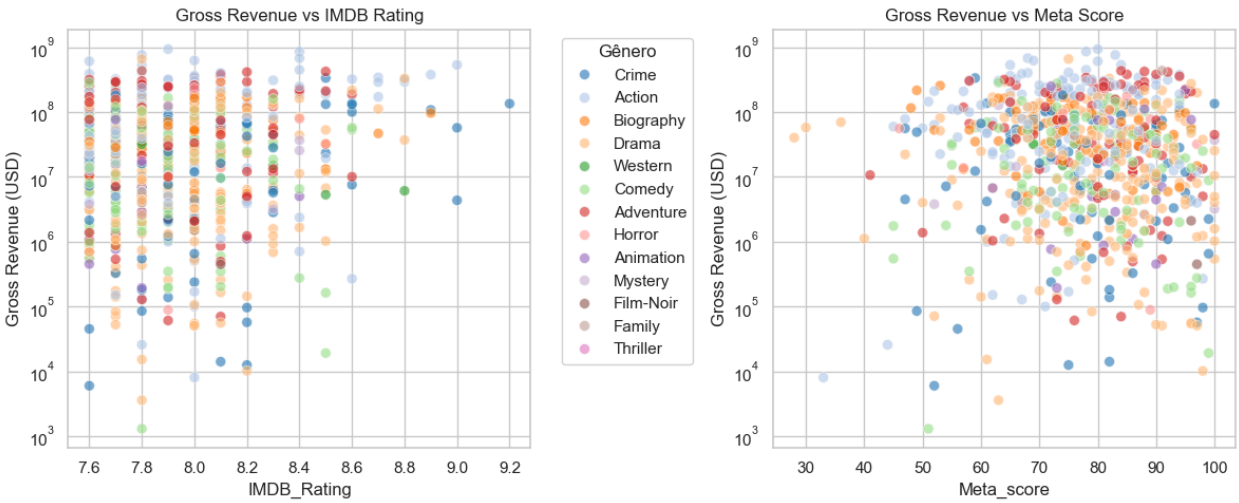
Additionally, by comparing **total revenue with IMDb ratings** per genre, we can observe patterns of **commercial success versus audience appreciation**. Expected trends indicate:

- Adventure and Action: high revenue, moderate ratings.
- Drama and Biography: lower revenue, potentially higher ratings.
- Animation: strong revenue and high ratings, appealing to broad audiences.
- Crime and Thriller: moderate revenue, potentially high ratings if well executed.

This combined analysis provides a clear overview of genre distribution, popularity, and financial and critical performance in the dataset, which can guide both production strategy and data-driven insights.






**Analysis of Revenue vs Ratings with Genre Insights**

We revisited the dataset using scatterplots colored by genre and average revenue bars per rating bins to understand how audience scores, critical scores, and genre affect box office performance.

**1. Numerical correlations**

- **Gross_USD × IMDB_Rating:** 0.13 : weak positive correlation. Higher IMDb ratings slightly relate to higher box office, but the effect is minimal.
- **Gross_USD × Meta_score:** -0.03 : essentially no correlation. Critical evaluation does not predict revenue in this dataset.

**Conclusion:** Ratings alone are not reliable predictors of revenue.

**2. Scatterplots colored by genre**

Coloring by genre reveals which genres dominate certain regions:

- **IMDB Rating × Revenue:** Adventure, Action, and Animation appear more frequently in high-revenue films, while Drama spans all revenue ranges.

- **Meta_score × Revenue:** High critical scores do not guarantee high revenue. Commercially successful genres appear across multiple score ranges.

**Conclusion:** Genre is a more visible factor than rating in explaining high box office performance.

**3. Average revenue by rating bins**

- **IMDb Rating bins:**

    - (8.1–9.2]) : highest average revenue (~$100M).
    - (7.7–7.8]) : second highest, likely driven by blockbusters.
    - Intermediate bins have lower averages, breaking the expectation that higher rating = higher revenue.

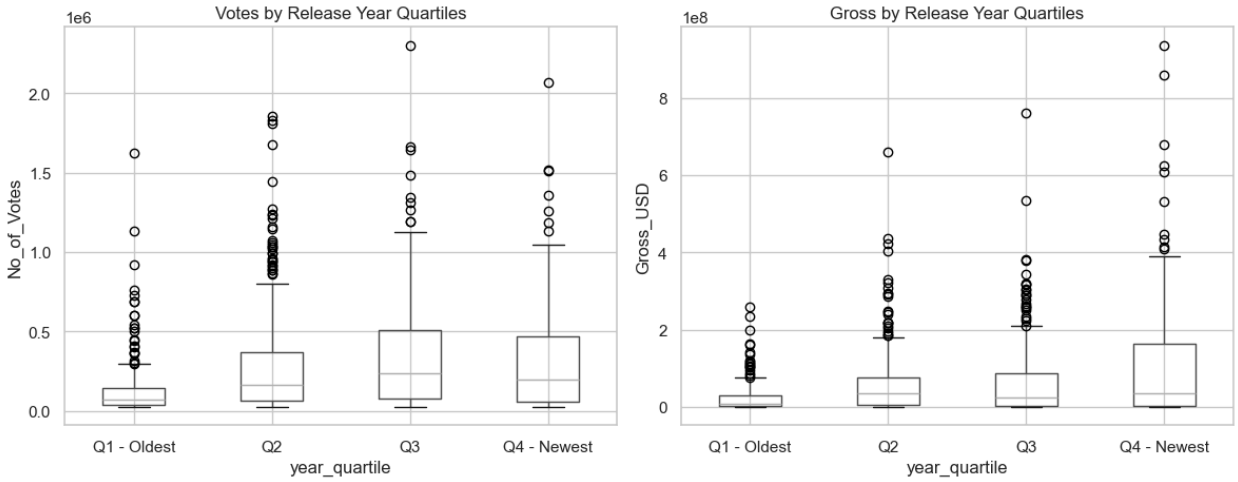
- **Meta_score bins:**

    - (67.6–75.0] and (75.0–82.0]) : highest average revenue (~$70–80M).
    - Highest range (88–100) has lower average revenue, suggesting critically acclaimed films may appeal to niche audiences.

**Conclusion:** Audience ratings help but do not guarantee financial success. Critical scores are less predictive; intermediate ranges seem more commercially viable.

**4. Overall insight**

- Ratings alone do not explain revenue; genre and other factors (marketing, franchise, distribution) play a larger role.
- Adventure, Action, and Animation consistently appear among the highest-grossing films.
- Drama dominates in quantity but not in average revenue.
- High critical scores do not guarantee box office success : critically acclaimed films may cater to niche audiences.




**Analysis of Movie Performance Over Time**

The dataset was divided into quartiles based on release year, from the oldest films (Q1) to the most recent (Q4), to examine trends in popularity and box office revenue over time. This approach answers two key questions:

1. How does audience popularity (number of votes) vary between older and newer films?
2. How does average box office revenue change across decades?

**1. Numerical results**

| Quartile (Year) | Average Votes | Average Revenue (USD) |
|---|---|---|
| **Q1 - Oldest** | 130,239 | 26.9 million |
| **Q2** | 304,260 | 62.5 million |
| **Q3** | 346,422 | 69.7 million |

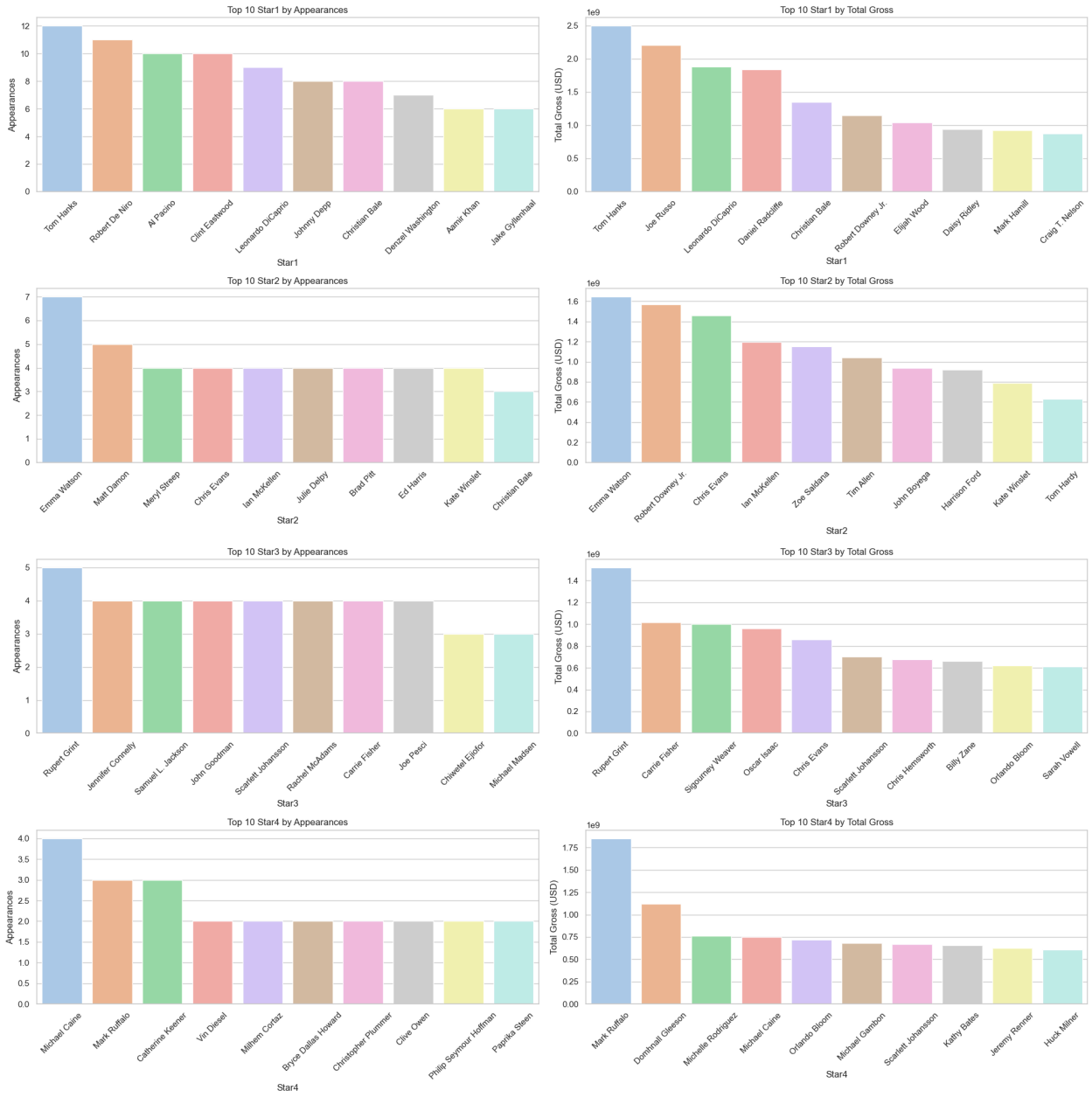
| | | |
|---|---|---|
| **Q4 - Newest** | 310,022 | **103.7 million** |

**2. Interpretation**

- **Popularity (votes):**
    - Older films (Q1) have significantly fewer votes, which is expected since fewer people watched them recently and the IMDb user base has grown over time.
    - The peak number of votes occurs in Q3, with Q4 still maintaining a high level of engagement.
- **Average revenue:**
    - Revenue consistently increases from older to newer films.
    - Q4 (most recent films) has **almost four times** the average revenue of Q1.
    - This trend likely reflects inflation, higher ticket prices, expansion into international markets, and more aggressive marketing strategies.

**3. Practical implications**

- **Understanding temporal trends:** demonstrates how the market has evolved and how recent films tend to earn more.
- **Adjusting comparisons:** when comparing revenues across eras, older films naturally show lower earnings (nominal values) and fewer votes.
- **Predictive modeling:** release year quartile can be used as a categorical variable when predicting box office performance or popularity.




**Analysis of Top Actors and Actresses – Popularity vs Revenue**

Using the functions from src/plots.py, two sets of charts were generated: the **Top 10 actors and actresses by number of appearances** and the **Top 10 by total gross revenue**. This comparison highlights the difference between popularity (measured by frequency in films) and financial impact (measured by total box office).

**1. Popularity does not equal profitability**

Some actors and actresses appear frequently across many films but do not rank among the top revenue generators. This suggests that being prolific does not guarantee financial success:genre, budget, and franchise involvement play a larger role.

**2. Female stars – frequency vs box office**

- The **Top 10 actresses by appearances** includes names absent from the **Top 10 by revenue**.
- This indicates that many female stars participate in smaller-scale productions such as dramas or independents, which contribute less to total gross.
- Conversely, those in the top revenue list are more likely tied to blockbuster franchises or high-budget projects.

**3. Male stars – higher overlap between presence and revenue**

- Among actors, there is greater convergence: those who appear frequently are often also top grossers.
- This reflects a stronger link between high visibility and participation in commercially successful genres such as action and adventure.

**4. Revenue concentration in a few names**

The **Total Gross charts** show that a small number of actors and actresses capture a disproportionate share of box office revenue. This is likely driven by recurring participation in franchises (e.g., superhero films, long-running sagas, or animated series).

**5. Genre distribution opportunities**

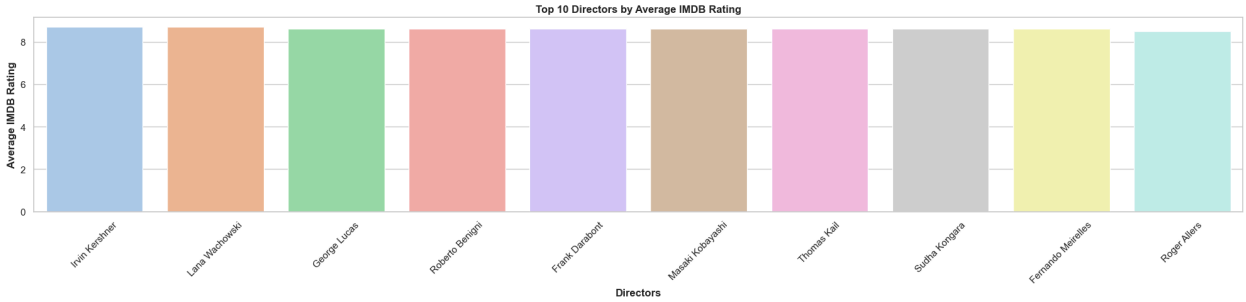
Cross-referencing these stars with the genres they appear in could reveal:

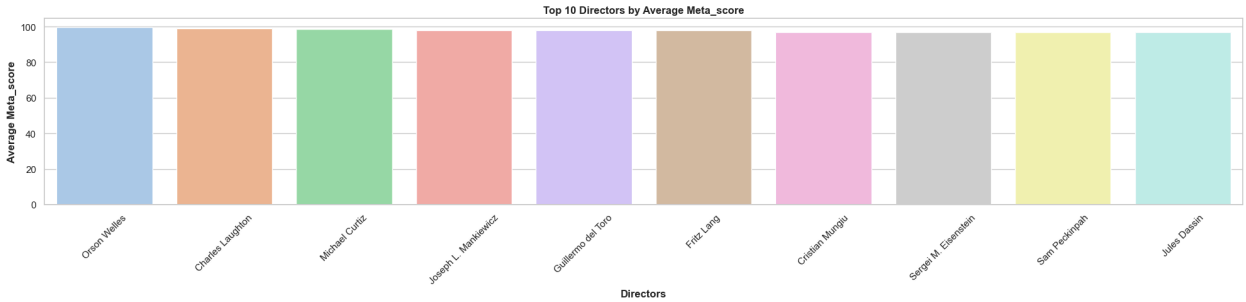
- Which genres are more financially rewarding for men versus women.

- Potential disparities in commercial opportunities, where female stars may be concentrated in lower-grossing genres like drama or romance, while male stars dominate action and adventure.

**Combined Insight**

- Frequency of appearance is not a reliable indicator of financial success.
- A few key stars dominate box office returns, often tied to franchise films.
- Gender dynamics suggest that male actors have greater overlap between popularity and financial performance, while female actors' careers may be more dispersed across less commercially lucrative genres.
- Strategic follow-up analyses should include **average gross per film** and **genre breakdown by star**, to identify which performers consistently deliver the highest return per project.




The three rankings of directors analyzed: by number of films, average IMDb rating, and average Metascore, highlight distinct dimensions of influence and success in the film industry. Each perspective reveals unique patterns, showing that productivity, popular reception, and critical acclaim do not always overlap.

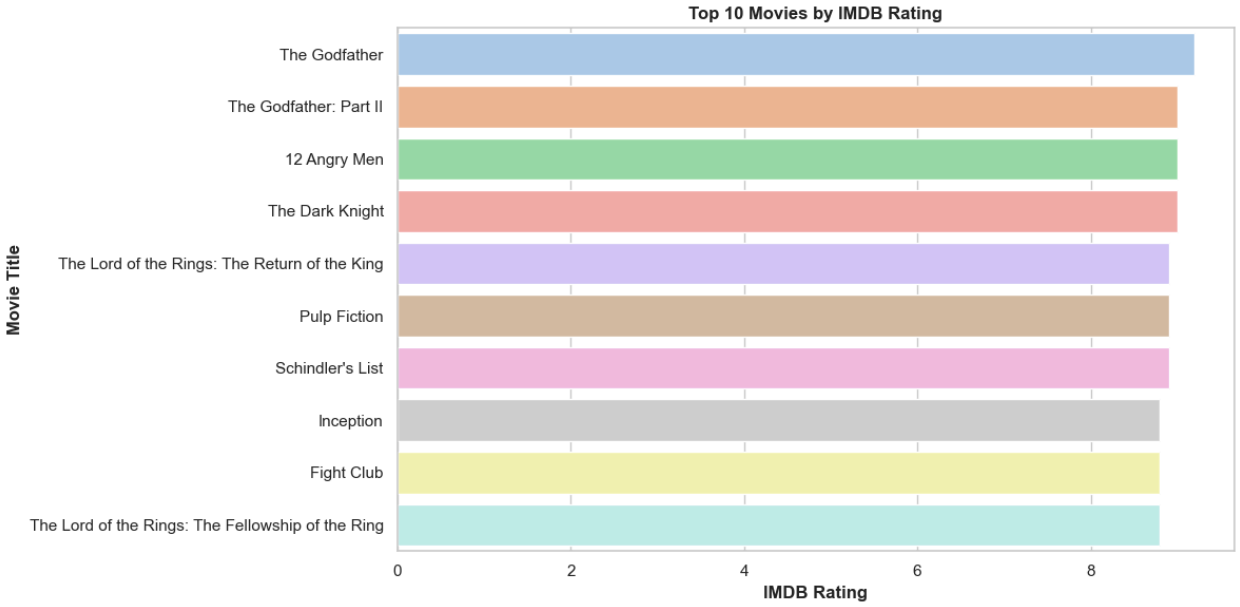
When ranking by the number of films directed, names such as Alfred Hitchcock, Steven Spielberg, Woody Allen, Martin Scorsese, and Ridley Scott stand out. These directors are highly prolific, often with long careers spanning multiple decades and genres. Their strong presence in the industry is associated with versatility and consistent output. However, this list also shows that directing many films does not automatically guarantee either high ratings or critical prestige, which suggests that quantity does not equate to quality.

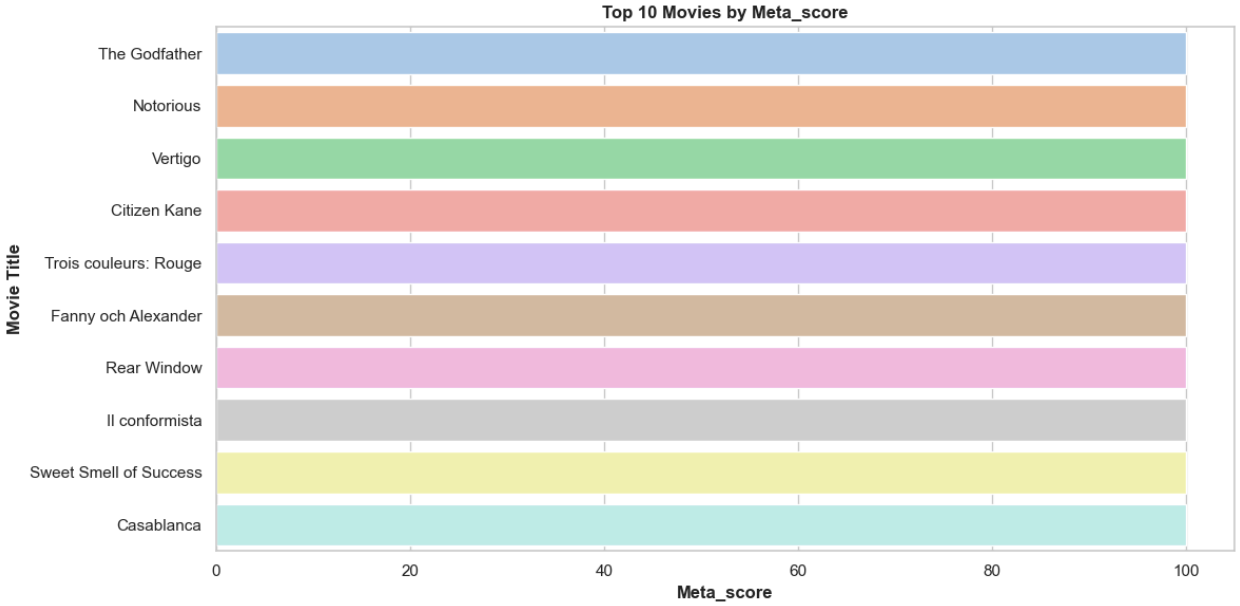
The ranking by average IMDb rating highlights a different set of directors, including Fritz Lang, Luis Buñuel, George Lucas, Stanley Kubrick, and Francis Ford Coppola. These filmmakers, often with fewer works compared to the most prolific names, achieve higher average ratings from audiences. Their films are frequently considered cult classics, groundbreaking, or narratively impactful. This suggests that selectivity in projects or strong authorial voices often resonate more deeply with audiences, resulting in higher ratings despite smaller filmographies.

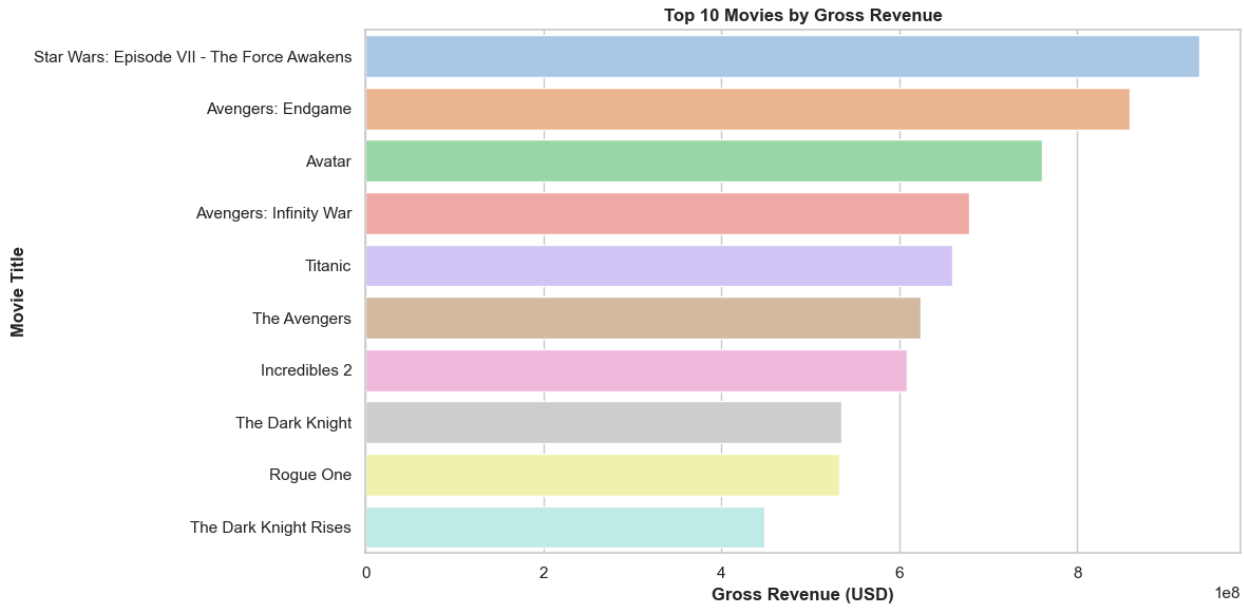
The ranking by average Metascore emphasizes directors such as Christopher Nolan, Quentin Tarantino, Jessica Hausner, Steve McQueen, Julie Ducournau, and Céline Sciamma. Many of these directors are contemporary, with fewer films, but they receive strong recognition from critics for originality, technical execution, and narrative depth. Unlike the IMDb ranking, which reflects audience preferences, this ranking captures the perspective of professional critics, who tend to value innovation and artistic vision over commercial appeal.

When comparing the three rankings side by side, the differences are clear. Productivity is dominated by classical and established directors with long careers. Audience ratings favor

impactful storytellers and cult filmmakers. Critical acclaim is concentrated among contemporary auteurs who push the boundaries of narrative and style. These contrasts demonstrate that popularity, critical prestige, and productivity are distinct measures of success in cinema.


Top 10 Movies by IMDB Rating
The Godfather
The Godfather: Part II
12 Angry Men
The Dark Knight
The Lord of the Rings: The Return of the King
Pulp Fiction
Schindler's List
Inception
Fight Club
The Lord of the Rings: The Fellowship of the Ring
Movie Title
0
2
4
6
8
IMDB Rating


Top 10 Movies by Meta_score
The Godfather
Notorious
Vertigo
Citizen Kane
Trois couleurs: Rouge
Fanny och Alexander
Rear Window
Il conformista
Sweet Smell of Success
Casablanca
Movie Title
0
20
40
60
80
100
Meta_score




Let's analyze the three rankings generated, revealing different dimensions of cinematic success: box office revenue, audience rating (IMDb), and critical evaluation (Meta_score). We will interpret each and then cross insights.

Top 10 by Box Office (Gross_USD)

Movie | Revenue (USD)
Star Wars: Episode VII | 936M
Avengers: Endgame | 858M
Avatar | 760M
Avengers: Infinity War | 678M
Titanic | 659M
The Avengers | 623M
Incredibles 2 | 608M
The Dark Knight | 535M
Rogue One | 532M
The Dark Knight Rises | 448M

Observations:

- Franchises dominate: Marvel, Star Wars, DC, Pixar.
- Global blockbusters rely on visual effects, action, and adventure.
- Few auteur or independent films appear.
- The Dark Knight is the only movie also in the IMDb rating ranking.

Insight: Financial success is strongly tied to franchises and major studios, not necessarily to perceived quality.

Top 10 by IMDb Rating

Movie | Rating
The Godfather | 9.2
The Godfather II | 9.0
12 Angry Men | 9.0
The Dark Knight | 9.0
LOTR: Return of the King | 8.9
Pulp Fiction | 8.9
Schindler's List | 8.9
Inception | 8.8
Fight Club | 8.8
LOTR: Fellowship of the Ring | 8.8

Observations:

- Classics and cult films dominate: drama, crime, fantasy, and psychological genres.
- Greater diversity of genres and styles.
- The Dark Knight appears here and in the box office ranking, showing rare convergence.

Insight: Audiences value narrative depth, memorable characters, and impactful scripts, not only visual spectacle.

Top 10 by Meta_score

All films have a perfect Meta_score of 100, representing maximum critical evaluation.

Movies included:
The Godfather, Notorious, Vertigo, Citizen Kane, Trois couleurs: Rouge, Fanny och Alexander, Rear Window, Il conformista, Sweet Smell of Success, Casablanca

Observations:

- Absolute classics in cinema history.
- Many are older, European, or auteur films.
- None appear in the box office ranking, and few in IMDb.

Insight: Critics value innovation, aesthetics, depth, and legacy, not necessarily popularity or financial return.

Crossing the Rankings

Criterion | Dominant Profile | Convergence
Box Office | Blockbusters, franchises | Low
IMDb Rating | Classics and cults | Medium
Meta_score | Auteur and historical films | Very low

- The Dark Knight is the only movie appearing in both box office and IMDb rankings, showing commercial success and audience acclaim.
- The Godfather appears in both IMDb and Meta_score, showing that critics and audiences can sometimes agree.

For profit, invest in franchises and visually appealing films, for critical prestige, focus on auteur direction and deep storytelling and for balance between audience and critics, study cases like The Dark Knight, Inception, and LOTR, which combine commercial impact and perceived quality.

Answering questions:

1. **Which movie would you recommend to someone you do not know?**

To best answer the question, "Which movie would you recommend to someone you do not know?", a code-based approach can be used. The logic behind such a recommendation system would involve **combining multiple general metrics** from the dataset since no personalized preferences are available. For example, the code could rank movies based on **box office revenue** to capture broad appeal, **IMDb ratings** to reflect audience satisfaction, and **Meta_score** to account for critical acclaim. By normalizing and weighting these metrics, the code can calculate a composite score for each movie, allowing it to identify titles that are both widely popular and well-regarded. The movie with the highest combined score would then be recommended, ensuring that the suggestion balances **entertainment value, popularity, and quality**, which is the most reliable strategy when the viewer's personal tastes are unknown.

**Code:**

```
df_clean['IMDB_Rating'] = pd.to_numeric(df_clean['IMDB_Rating'], errors='coerce')
df_clean['Meta_score'] = pd.to_numeric(df_clean['Meta_score'], errors='coerce')

df_ratings = df_clean.dropna(subset=['IMDB_Rating', 'Meta_score'])

top_imdb = df_ratings.sort_values(by=['IMDB_Rating', 'Meta_score'], ascending=False).iloc[0]

print("Top-rated movie recommendation:")
print(f"Title: {top_imdb['Movie_title']}")
print(f"IMDB Rating: {top_imdb['IMDB_Rating']}")
print(f"Meta Score: {top_imdb['Meta_score']}")
```

**The output:**

Top-rated movie recommendation:

Title: The Godfather

IMDB Rating: 9.2

Meta Score: 100.0

2. **What are the main factors related to a movie's high box office expectations?**

To identify the main factors associated with high box office expectations for a movie, a combination of numeric correlations and star power analysis can be performed using the dataset.

First, numeric features such as IMDb rating, Meta_score, and runtime are converted to numeric types to ensure proper calculations. The correlation of these features with Gross_USD is then computed. This step reveals how strongly each variable is associated with box office performance. For example, a positive correlation with IMDb rating indicates that films with higher audience scores tend to earn more, while runtime might show whether longer or shorter films have an impact on revenue.

Second, the star power of the cast is analyzed by calculating the total gross revenue associated with the top actors in each main star position (Star1, Star2, Star3, Star4). By summing Gross_USD per actor, it is possible to identify which performers are linked to the highest-grossing films. This approach captures the commercial influence of well-known actors, which often drives audience interest and contributes significantly to a film's financial success.

Combined, these analyses allow us to determine both quantitative factors (ratings, runtime) and qualitative factors (actors' commercial appeal) that are most strongly related to high expected box office revenue. Such insights can guide predictive models or production strategies for maximizing financial outcomes.

3. **What insights can be drawn from the Overview column? Is it possible to infer a movie's genre from this column?**

Insights from the Overview Column and Genre Prediction

The Overview column, which contains textual summaries of movies, provides useful information about narrative tone, thematic content, and, to some extent, genre. To analyze this, two complementary approaches were applied: sentiment analysis and multilabel genre classification.

1. Sentiment Analysis of Overviews

The sentiment of overviews was categorized into negative, neutral, and positive, revealing the following distribution:

- Neutral: 485 movies. Most overviews are descriptive and objective, focusing on plot rather than emotional tone.
- Positive: 282 movies. These overviews contain optimistic or hopeful language.
- Negative: 232 movies. Summaries include darker themes, conflict, or tragedy.

When comparing sentiment to audience ratings (IMDB_Rating), the differences were minimal: films with negative overviews had slightly higher average ratings (7.96) than neutral (7.95) or positive (7.93) overviews. This suggests that overviews with darker or more dramatic content are slightly more appreciated by audiences.

Wordcloud analysis further highlighted genre tendencies:

- Negative overviews: Words like *violent, criminal, war, brutal, murder, escape*, commonly associated with thrillers, dramas, and war films.
- Neutral overviews: Words like *man, woman, story, family, find, young*, indicative of descriptive narratives with a focus on characters and relationships.
- Positive overviews: Words like *love, young, life, new, discover, school*, pointing toward romance, coming-of-age, or family-oriented films.

Interpretation: The tone of the overview reflects the type of content and narrative style, which can indirectly suggest genre tendencies. However, sentiment alone is not a reliable predictor of audience rating or definitive genre classification.

2. Multilabel Genre Classification Using Overviews

A multilabel classification model was built using TF-IDF vectorization and a Multinomial Naive Bayes classifier within a MultiOutputClassifier framework to predict multiple genres from the text. Evaluation per genre revealed:

- Best-performing genre: Drama showed high positive-class recall (1.00) and reasonable F1-score (0.82), likely because it is the most frequent genre in the dataset.
- Other genres: Action, Adventure, Thriller, and Romance showed overall accuracy between 0.80–0.88, but these values are misleading due to predicting predominantly the negative class (“not this genre”).

- Poorly predicted genres: Music, Sport, Film-Noir, Western, War, and History had very few positive examples. The model almost always predicted “0” (not the genre), resulting in high accuracy but zero utility for genre identification.

Limitations Identified:

- Extreme class imbalance leads to poor recognition of rare genres.
- Low recall for positive classes demonstrates the model’s limited ability to detect true genre membership.
- TF-IDF + Naive Bayes limitations: ignoring semantic context and assuming word independence restricts model effectiveness for rich narrative text.

Insights and Strategic Conclusion:

- Movie overviews contain textual signals correlated with genre, particularly for frequent genres like Drama.
- Rare genres remain challenging to predict, highlighting the need for techniques that handle class imbalance and capture semantic meaning, such as contextual embeddings (e.g., BERT) or more advanced classifiers.
- Sentiment and textual content together can inform genre tendencies: for example, negative overviews align with thrillers or dramas, while positive overviews suggest romance or family films.

Answering the Question:
Yes, insights from the Overview column can provide indirect clues about a movie’s genre. Sentiment analysis highlights thematic tendencies, and text-based classification captures patterns for frequent genres. However, the predictive power is limited for rare genres, and sentiment alone is insufficient to determine genre definitively. Therefore, overviews are complementary features for genre prediction rather than standalone indicators.