IMDB Top 1000 Movies – Data Analysis Report

Presented by Abdurrahmaan Tayob for ZAIO Technology Institute - Assignment 2

Presentation Overview

1

Data Collection

Sourcing and initial review of the IMDB Top 1000 dataset.

2

Data Preparation

Rigorous cleaning, transformation, and feature engineering techniques.

3

Visualizations

Key insights derived from various graphical representations.

4

Statistical Analysis

Descriptive statistics and correlation analysis of key movie metrics.

5

Conclusion & Next Steps

Summary of findings and challenges encountered.

Data Collection: The IMDB Top 1000 Dataset

The foundation of this analysis is a meticulously curated dataset of the top 1,000 movies from IMDB. This dataset provides a comprehensive view of film characteristics and performance.

- **Key Attributes:** Titles, Release Years, Runtime, Genre, Certification.
- **Personnel Data:** Director and Lead Actor information.
- **Performance Metrics:** IMDB Ratings, Meta Scores, Gross Revenue, and Number of Votes.

The data, located in the `imdb_top_1000.csv` file, was loaded using the pandas library, with a comma delimiter ensuring proper parsing of all structured columns.

Data Preparation: Ensuring Data Integrity

Missing Value Handling

Rows with null values in 'Gross' and 'IMDB Rating' were removed to maintain the analytical integrity of financial and critical assessments.

Duplicate Elimination

Identified and eliminated 24 duplicate rows within the dataset to guarantee the uniqueness and accuracy of each movie entry, preventing skewed metrics.

Feature Engineering

- **Duration:** Extracted numerical duration from the 'Runtime' string.
- **Decade:** Created a 'Decade' feature based on 'Released Year' for temporal analysis.
- **Lead Actors:** Consolidated multiple 'Star' columns into a single 'Lead Actors' field.

Challenge: Gross Revenue Parsing

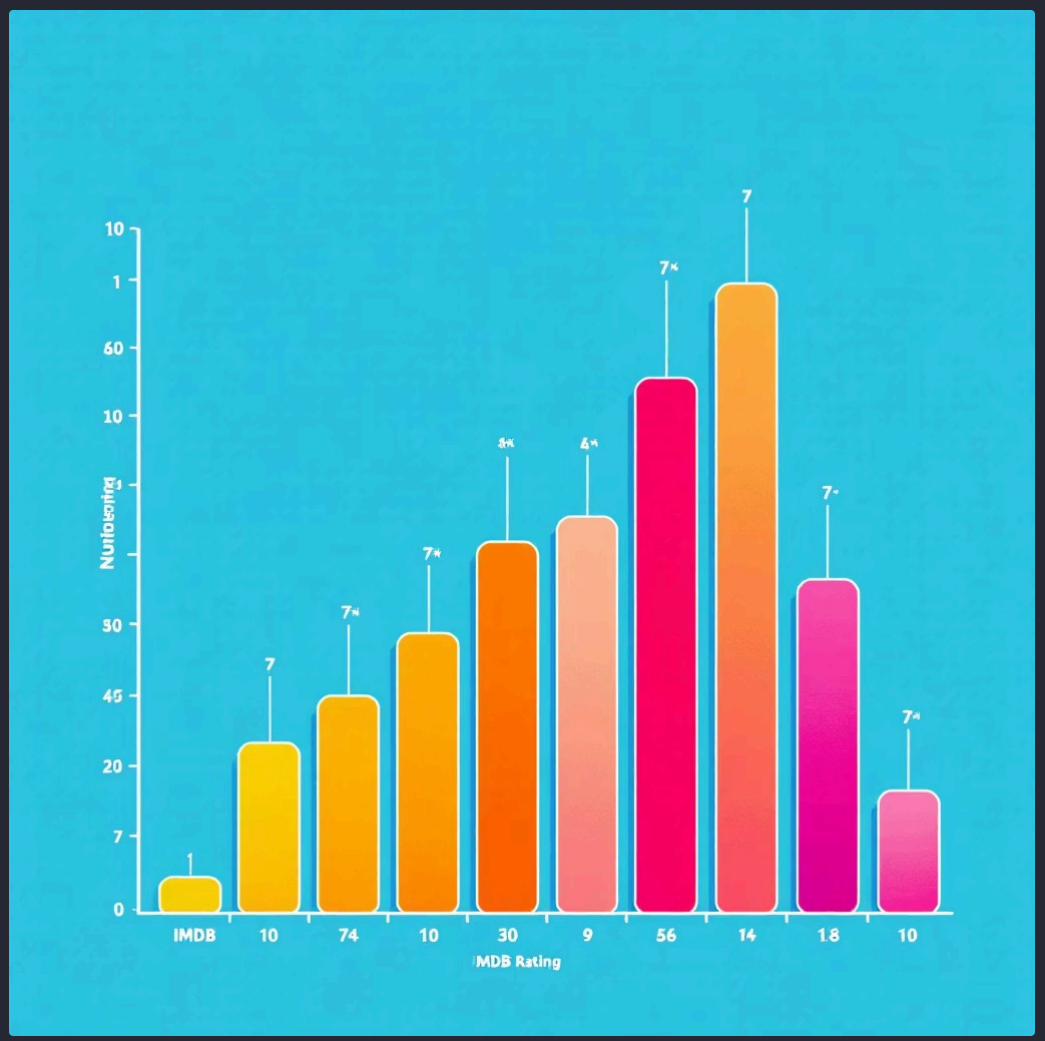
The 'Gross' revenue column, initially stored as a string with dollar signs and commas, required a complex regex cleanup and conversion to a numeric format. This was crucial for accurate financial analysis.

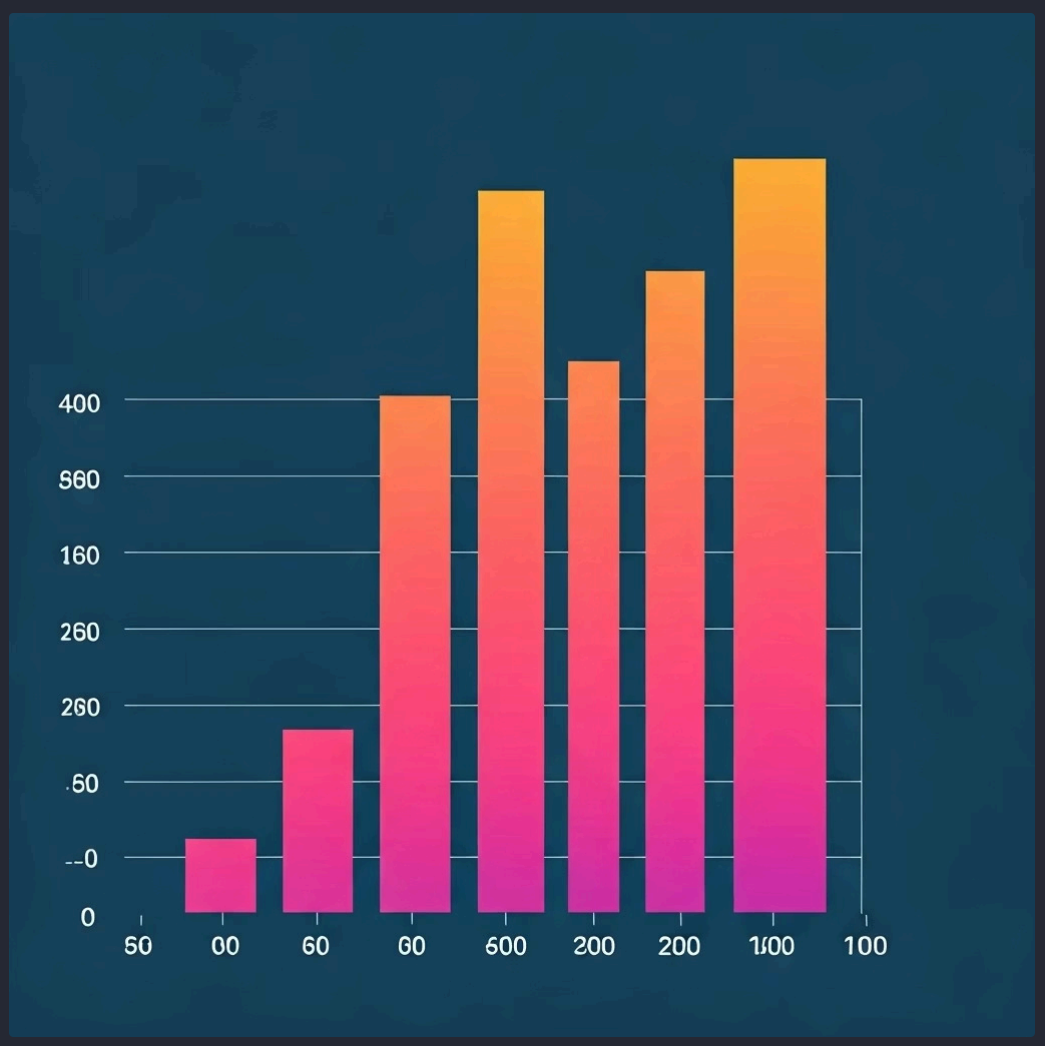
Visualizing Movie Characteristics: Ratings Distribution

Histograms were instrumental in understanding the distribution of movie ratings across the dataset.

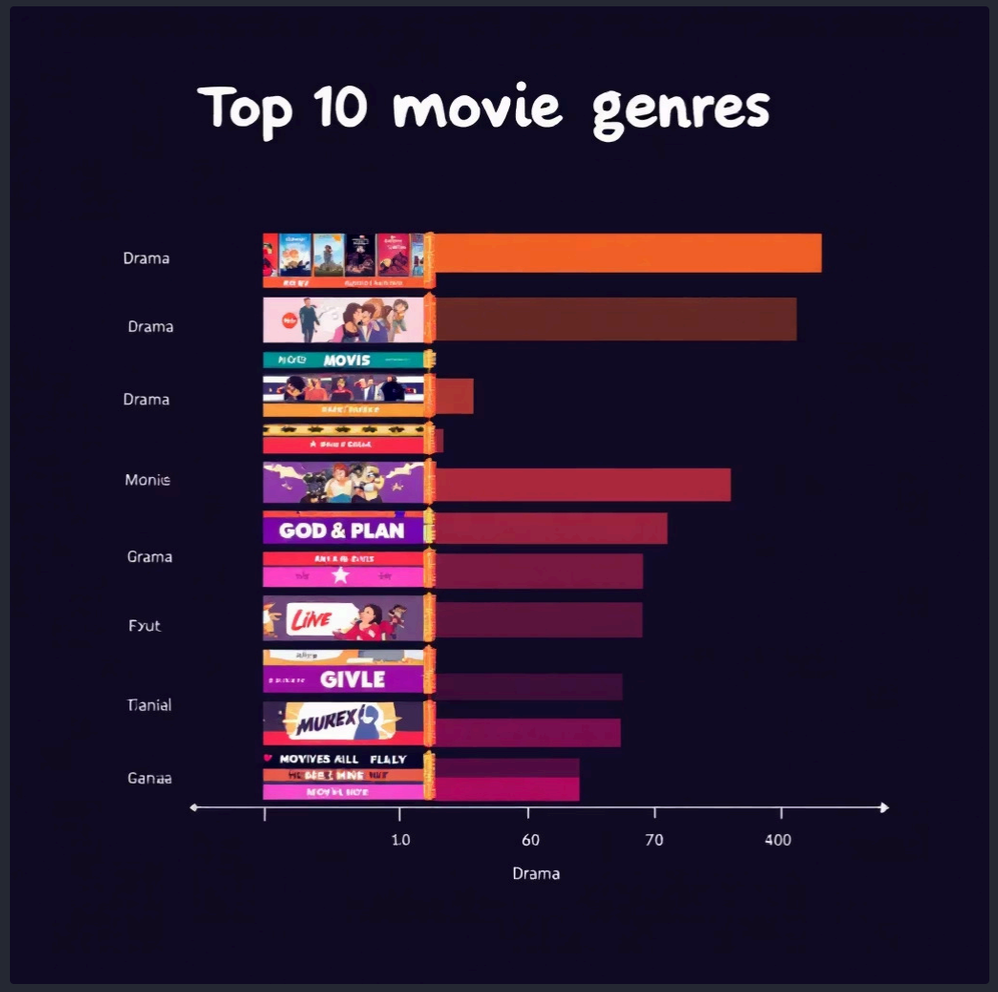
IMDB Rating Distribution: The majority of IMDB ratings are concentrated between 7 and 8, indicating a high overall critical quality for the films in the top 1000 list. This peak confirms the curated nature of the dataset.

Meta Score Distribution: Meta Scores, while also high, showed a more generalized distribution centered around the 60-70 mark. This suggests a broader spread of critical consensus compared to the user-driven IMDB ratings.

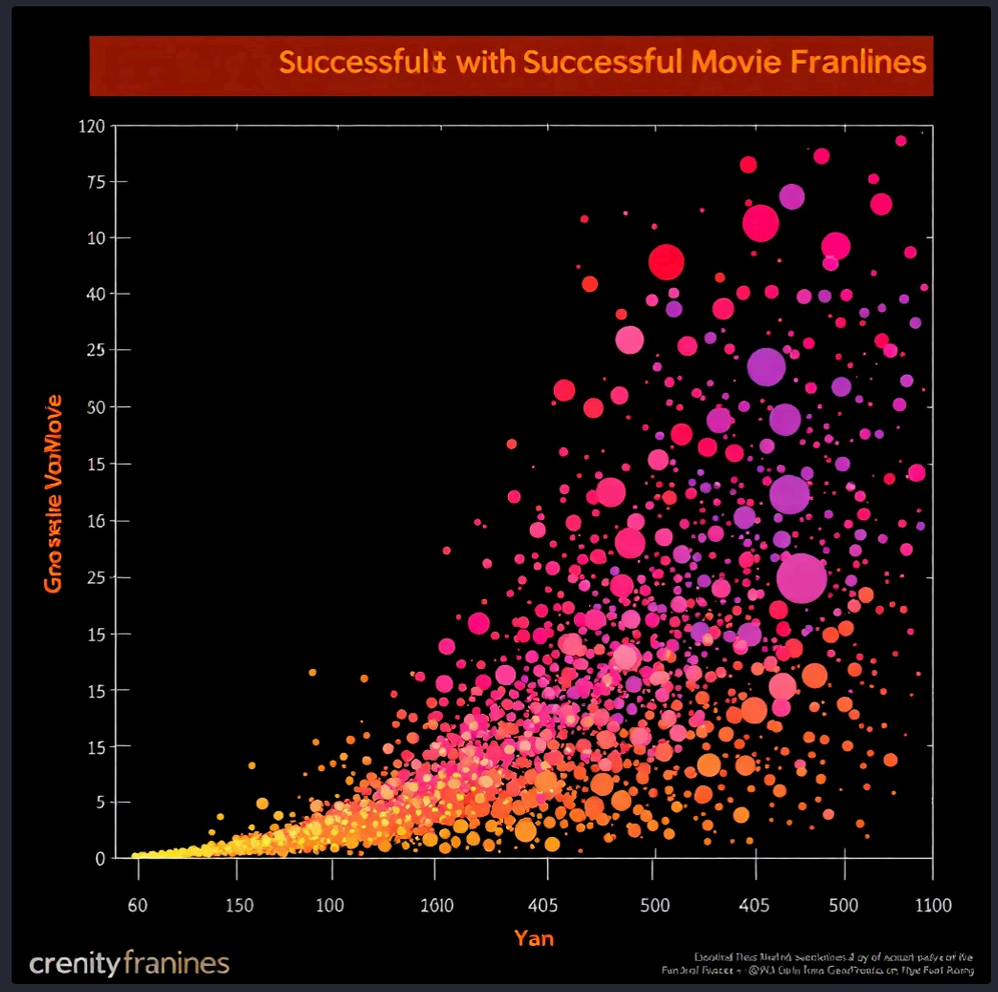




Visualizing Movie Characteristics: Genre and Revenue

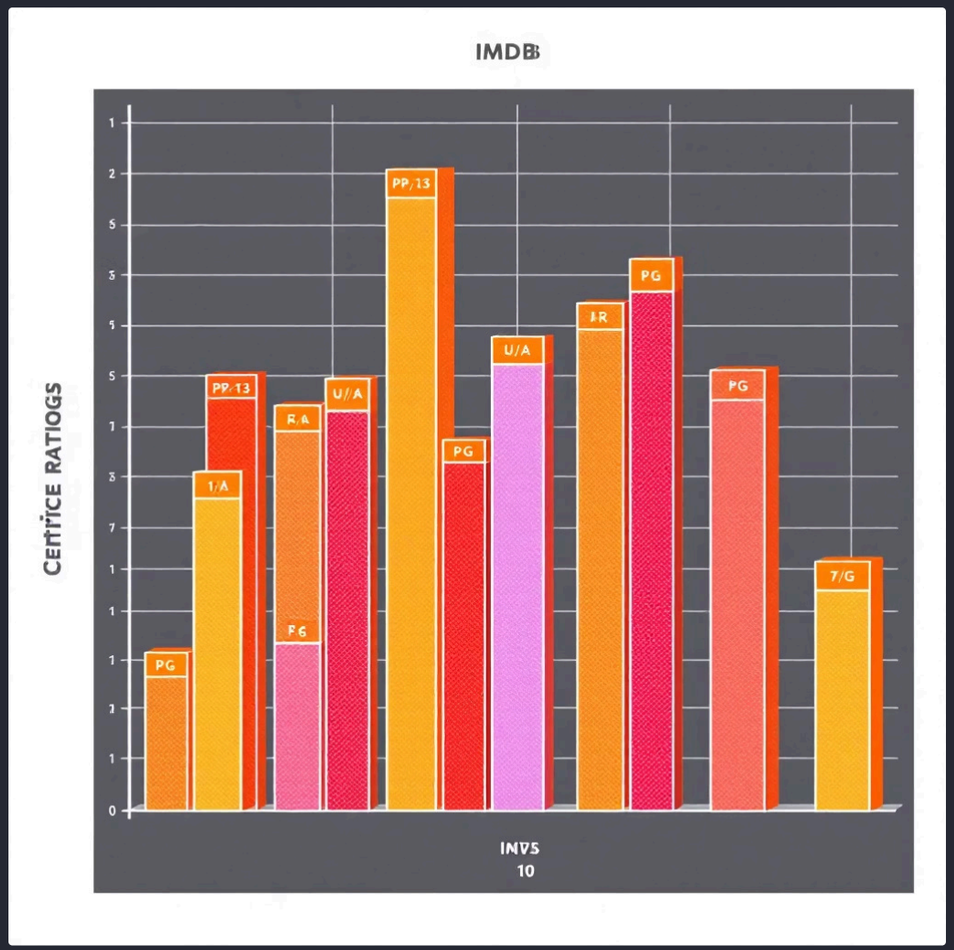


Top 10 Genres: The bar plot clearly illustrates that Drama, Action, and Thriller are the most prevalent genres within the IMDB Top 1000. Drama exhibits a considerable lead, highlighting its dominance among critically acclaimed films.

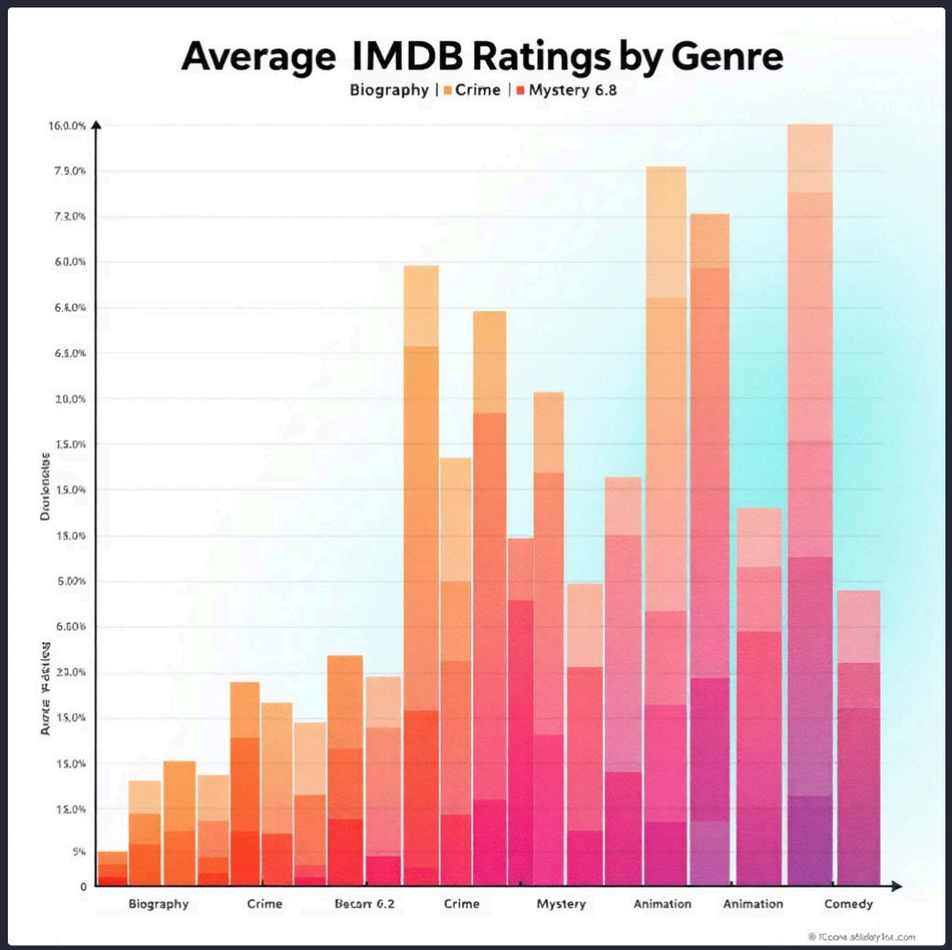


Revenue vs. Votes: The scatter plot reveals a distinct positive correlation between a movie's Gross Revenue and its Number of Votes. This suggests that films with higher audience engagement, as indicated by more votes, tend to achieve greater commercial success.

Visualizing Movie Characteristics: Ratings by Certificate & Genre



Ratings by Certificate: The box plot provides insights into IMDB Ratings across various certificate categories. While most categories show outliers, PG-13 and U/A certified movies generally exhibit slightly higher average ratings, suggesting these classifications are often associated with well-received films.



Average Ratings by Genre (Heat Map): This heat map visualizes the average IMDB Ratings for different genres. Biography, Crime, and Mystery genres consistently stand out with higher average ratings, indicating a tendency for these genres to be highly acclaimed by audiences. Westerns also emerged as a top-rated genre with an average rating of 8.03.

Statistical Analysis: Uncovering Deeper Trends

Gross (USD)	High skew	Moderate	Substantial variation
No. of Votes	Widely varied	Slightly right skewed	High dispersion
IMDB Rating	~7.5	~7.6	~0.5

Correlation Analysis: Gross vs. Number of Votes

Gross vs No. of Votes	0.67

A Pearson Coefficient of 0.67 indicates a moderate positive correlation between Gross Revenue and Number of Votes. This numerically validates the visual observation: higher audience engagement, as measured by votes, is a significant indicator of a movie's commercial success.

Conclusion: Key Findings & Reflections

Summary of Key Findings:

- **Director Impact:** Christopher Nolan and other top directors demonstrate consistent high average gross earnings.
- **Actor Synergy:** Leonardo DiCaprio and powerful actor pairings (Star1 & Star2) significantly contribute to box office success.
- **Genre Preferences:** Biography, Crime, and Mystery genres consistently achieve the highest average IMDB ratings, alongside Westerns.
- **Popularity-Revenue Link:** A strong positive correlation exists between audience size (votes) and commercial success (gross revenue).

Challenges Faced:

- **Gross Value Transformation:** Converting string-based 'Gross' values, cluttered with symbols, into usable numeric data was a significant preprocessing hurdle.
- **Genre Complexity:** Managing nested genres and ensuring meaningful visual comparisons across various movie categories added layers of complexity to the analysis.
- **Data Alignment:** Meticulous attention was required to align all columns correctly for grouping and plotting, ensuring data consistency and accurate visualization.

Thank You

Further Exploration

For a detailed look at the code, raw data, and further analysis, please visit the GitHub repository:

https://github.com/Artayob/ZAIO_assignment_ARTAYOB2.git