IMDB Top 1000 Movies – Data Analysis Report

1. Data Collection

The dataset analysed consists of **1,000 top-rated movies** curated from IMDB, containing:

* Titles, Release Years, Runtime, Genre, Certification
* Director and Lead Actor information
* IMDB Ratings, Meta Scores
* Gross Revenue and Number of Votes

The file imdb\_top\_1000.csv was loaded using pandas with a comma delimiter for parsing structured columns.

2. Data Preparation

Data was rigorously cleaned and transformed using the following steps:

* **Missing Values**:
  + Removed rows with null Gross and IMDB Rating to preserve analytical integrity.
* **Duplicates**:
  + Identified and dropped 24 duplicate rows for accuracy.
* **Feature Engineering**:
* Extracted **Duration** from the Runtime string.
* Created Decade based on the Released Year.
* Concatenated multiple Star columns into Lead Actors.

**Challenge Faced**:  
Handling Gross was tricky—it was stored as a string with dollar signs and commas. This required regex cleanup and conversion to numeric format before analysis.

3. Visualization

Multiple plots were generated to interpret movie characteristics:

* **Histograms**:
  + IMDB ratings peaked between **7 and 8**, confirming high critical quality.
  + Meta Scores showed a normal-ish distribution with center around **60–70**.
* **Bar Plot**:
  + Top genres include **Drama**, **Action**, and **Thriller**.
* **Scatter Plot**:
  + Showed a positive correlation between **Gross Revenue and Number of Votes** — movies with more votes tend to earn more.

**Box Plot**:

* + Displayed IMDB Ratings against certificate categories. **PG-13 and U/A** movies had slightly higher ratings.
* **Heat Map**:
* Visualized average IMDB Ratings by genre. **Biography**, **Crime**, and **Mystery** stood out.

4. Statistical Analysis

Descriptive Statistics

| **Feature** | **Mean** | **Median** | **Standard Deviation** |
| --- | --- | --- | --- |
| 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

| **Variables** | **Pearson Coefficient** |
| --- | --- |
| Gross vs No. of Votes | **0.67** |

This moderate positive correlation confirms that popularity (votes) influences earnings.

5. Conclusion

Summary of Key Findings:

* **Christopher Nolan** and other top directors consistently rank highest in average gross earnings.
* **Leonardo DiCaprio** and top actor pairs like *Star1 & Star2* show strong box office synergy.
* **Genres like Biography and Crime** receive the highest IMDB ratings on average.
* 💰 A clear trend exists between audience size (votes) and commercial success (gross).
* **Western** being the highest rated Genre with an average rating of 8.037500

💡 Reflection:

* **Challenges**:
* Transforming string-based Gross values into usable numeric data required careful preprocessing.
* Managing nested genres and creating meaningful visual comparisons added complexity.
* Ensuring all columns aligned for grouping and plotting needed extra attention.

This project successfully showcases your growing mastery in **data cleaning**, **visualization**, and **statistical storytelling**, Abdurrahmaan. Want help formatting this into Markdown or a polished PDF report next?