Dataset Card — Hollywood Movies

|  |  |  |
| --- | --- | --- |
| Column | Type | Description |
| Film | **string** | **Movie title** |
| Genre | **string** | **Main genre** |
| Lead Studio | **string** | **Production or distribution studio** |
| Audience score % | **float** | **Audience rating percentage** |
| Profitability | **float** | **Ratio of worldwide gross to production budget** |
| Rotten Tomatoes % | **float** | **Critics rating percentage** |
| Worldwide Gross | **float** | **Worldwide box office gross in millions USD** |
| Year | **int** | **Release year** |

**1️⃣ Understand the Dataset**

* Each row represents **one movie**.
* **Columns:**
  + Film (title)
  + Genre (categorical)
  + Lead Studio (categorical)
  + Audience Score % (numeric)
  + Profitability (numeric)
  + Rotten Tomatoes % (numeric)
  + Worldwide Gross (numeric, in millions USD)
  + Year (numeric)
* **Goal:** Explore how ratings, profitability, and revenue vary by genre, studio, and over time.

**2️⃣ Initial Data Inspection**

* Shape of dataset (rows × columns).
* First 5 rows (head).
* Column data types.
* Numeric vs categorical columns.
* Missing values count per column.
* Summary statistics for numeric columns (mean, median, std, min, max).

**3️⃣ Preprocessing Steps**

**A. Handle Missing Values**

* Identify columns with missing data.
* Decide on:
  + **Categorical**: Replace with "Unknown" or mode.
  + **Numeric**: Fill with median/mean (if reasonable) or drop.
* For critical financial columns (Profitability, Gross) — investigate before filling.

**B. Handle Duplicates**

* Check for repeated Film names.
* Remove duplicates unless they represent different releases.

**C. Feature Engineering**

* **Decade** from Year (2000–2009 → "2000s").
* **Is\_Independent**: Flag if Lead Studio = "Independent".
* **Critic-Audience Gap** = Audience Score % − Rotten Tomatoes %.
* **ROI Category**: High / Medium / Low profitability groups.

**D. Encode Categorical Variables**

* Label Encode Genre if needed.
* One-Hot Encode Lead Studio for advanced analysis.

**E. Detect Outliers**

* Profitability outliers (very high values).
* Worldwide Gross blockbusters vs small releases.
* Decide whether to keep or analyze separately.

**F. Fix Data Types**

* Convert Year to integer.
* Ensure all percentages are numeric.

**4️⃣ EDA Focus Areas**

**A. Basic Distributions**

* Count of movies per genre.
* Count of movies per studio.
* Number of movies per year/decade.
* Distribution of Audience Score %, Rotten Tomatoes %, Profitability, and Worldwide Gross.

**B. Profitability Analysis**

* Average profitability by genre.
* Average profitability by studio.
* Profitability trend over years.
* Most and least profitable films.

**C. Ratings Analysis**

* Compare audience vs critic ratings.
* Genres with highest average ratings.
* Studios with highest average ratings.
* Largest critic–audience disagreements.

**D. Revenue Analysis**

* Top 10 highest-grossing movies.
* Revenue by genre.
* Revenue by studio.
* Yearly trends in Worldwide Gross.

**E. Relationship Analysis**

* Profitability vs Audience Score %.
* Profitability vs Rotten Tomatoes %.
* Correlation between all numeric variables (heatmap).
* Does high critic rating predict high profitability?

**F. Multi-Dimensional Insights**

* Profitability by genre and studio.
* Audience–critic gap by genre.
* Top profitable genres over decades.

**5️⃣ Final Insights & Storytelling**

* Which genres are both popular and profitable.
* Which studios dominate in revenue vs ratings.
* Whether critic scores matter for financial success.
* How movie performance trends changed over time.

**6️⃣ Participant Challenge Ideas**

* Find the genre with the most profitable *blockbusters*.
* Identify the studio that most consistently gets high critic ratings.
* Compare profitability between independent vs major studios.
* Spot the movie with the largest critic–audience score gap.