Data Cleaning Approach

A systematic data cleaning process was applied to ensure data accuracy, consistency, and readiness for analysis.

**1. Remove Duplicates**

All duplicate records across columns were identified and removed to maintain data integrity and ensure unique entries for each movie.

**2. Standardize Year Format**

* Corrected inconsistent year formats and removed unwanted characters.
* Fixed encoding issues (e.g., special characters or symbols).
* Handled missing or null year values by replacing them with “N/A” or “Unknown” for clarity.

**3. Handle Missing Values — Genre**

Blank or missing entries in the Genre column were replaced with “Unknown” to maintain categorical completeness.

**4. Rating**

* Missing ratings were replaced with 0, not "N/A", to maintain numeric consistency for exploratory data analysis (EDA).
* A value of 0 represents “Not Rated / Missing”, ensuring that statistical summaries (mean, median, histograms) can be computed without data type errors.
* This approach allows seamless numeric operations while keeping the semantic meaning of missing ratings documented.

**5. Votes**

* Blank cells in the Votes column were replaced with 0.
* A 0 indicates “No Votes Recorded”, which differentiates between movies that have zero votes versus missing data.
* This helps maintain accurate aggregations and ensures valid numeric data during visualization and statistical analysis.

**6. Runtime**

* Missing or blank Runtime values were replaced with “0” to indicate unavailable duration information.
* Since runtime is a numeric field used for descriptive purposes (not computation-heavy), text substitution prevents false averages or misleading runtime summaries.

**7. Gross**

* Missing or undisclosed Gross values were replaced with 0, and documented as “Not Disclosed” in metadata.
* The numeric 0 allows consistent data typing during EDA, while the label “Not Disclosed” communicates that the actual value was not provided rather than zero revenue.

”to ensure transparency in reporting and preserve the professional tone of the dataset.

✅ Outcome

After cleaning, the dataset was standardized, complete, and analysis-ready — free from duplicates, inconsistent formats, and missing data gaps.