\*\*Data Quality and Information Project Report\*\*

\*\*Introduction\*\*

This report outlines the data preparation steps undertaken as part of the Data Quality and Information project. The focus was on profiling, assessing, and cleaning the dataset to ensure consistency, accuracy, and usability. This process included detailed phases of data profiling, quality assessment, and various cleaning operations such as data transformation, error correction, and deduplication.

---

\*\*1. Data Profiling\*\*

Data profiling was conducted to understand the structure, content, and overall quality of the dataset. Key activities included:

- \*\*Schema and Attribute Analysis\*\*: Identified column names, data types, and primary attributes.

- \*\*Summary Statistics\*\*: Computed statistics for numeric fields (mean, median, standard deviation) and frequency distributions for categorical fields.

- \*\*Missing Data Overview\*\*: Quantified missing values across columns to highlight incomplete records.

- \*\*Visualization\*\*: Visual tools like histograms and bar charts were used to identify data distribution trends and potential anomalies.

Key Findings:

- Some columns contained a high proportion of missing values.

- Data formats were inconsistent in certain fields, requiring standardization.

---

\*\*2. Data Quality Assessment\*\*

The data quality assessment focused on evaluating key dimensions:

- \*\*Completeness\*\*: Quantified the extent of missing data in key columns.

- \*\*Accuracy\*\*: Compared data entries against expected patterns and rules.

- \*\*Consistency\*\*: Checked for uniformity in categorical fields and data formats.

- \*\*Uniqueness\*\*: Examined duplicate records based on critical attributes like titles and descriptions.

Key Findings:

- Several entries showed inconsistencies in categorical field spelling.

- Duplicate records were detected based on similarities in titles and plot descriptions.

---

\*\*3. Data Cleaning\*\*

To address the identified issues, multiple cleaning operations were performed:

### 3.1 Data Transformation and Standardization

- \*\*Objective\*\*: Ensure uniformity in data representation.

- \*\*Actions\*\*:

- Converted date formats to a consistent structure.

- Standardized categorical values (e.g., genre names) by correcting typos and consolidating variations.

- Normalized numeric fields by handling out-of-range values and enforcing consistent units where applicable.

### 3.2 Error Detection and Correction

- \*\*Objective\*\*: Identify and correct inaccuracies.

- \*\*Actions\*\*:

- Applied threshold-based validation for numeric fields to identify outliers.

- Checked categorical fields against predefined valid value sets.

- Corrected erroneous or missing entries using external references when possible.

### 3.3 Data Deduplication

- \*\*Objective\*\*: Remove redundant entries to ensure data uniqueness.

- \*\*Actions\*\*:

- Implemented a blocking strategy to group potential duplicates based on titles.

- Utilized similarity measures (e.g., cosine similarity for plot descriptions) to identify matches.

- Verified matches and removed duplicates from the dataset.

---

\*\*4. Outlier Analysis\*\*

Outlier detection was conducted to identify unusual data points without removing any records. Methods included:

- \*\*Statistical Analysis\*\*: Used standard deviation thresholds to highlight extreme values in numeric fields.

- \*\*Visualization\*\*: Scatter plots and boxplots were employed to visually inspect anomalies.

Key Observations:

- Outliers were identified in runtime and rating fields.

- No records were dropped as the anomalies were deemed valuable for contextual analysis.

---

\*\*5. Final Cleaned Dataset\*\*

The cleaned dataset:

- Eliminated duplicates, reducing the number of records from the initial total.

- Standardized and corrected all identified inconsistencies.

- Retained all original records, including outliers, for further analysis.

\*\*Export Details\*\*:

The final dataset was exported as a CSV file, ensuring compatibility for subsequent analysis.

---

\*\*Conclusion\*\*

The data preparation phase successfully improved the quality and reliability of the dataset through systematic profiling, assessment, and cleaning. This clean dataset provides a robust foundation for subsequent analytical tasks or reporting needs.