Data Cleaning Report – Audible Web Scraped Dataset

Dataset Description

The dataset contains web-scraped audiobook metadata from Audible. It includes the following attributes:

- Book titles
- Authors
- Narrators
- Duration
- Release dates
- Language
- User ratings
- Prices

The dataset is suitable for audiobook market analysis, pricing strategy optimization, customer preference modeling, and business intelligence reporting.

1. name

Checks Performed:

- Text encoding consistency
- Leading and trailing whitespace
- Multiple internal spaces

Cleaning Steps Performed:

- UTF-8 enforced
- Trimmed leading and trailing spaces
- Collapsed multiple spaces into a single space

2. author

Checks Performed:

- Presence of constant prefix (Writtenby:)
- Merged names
- Multiple-author delimiters
- Capitalization consistency

Cleaning Steps Performed:

- Removed prefix
- Inserted spaces in merged names
- Standardized delimiters
- Rejoined using '; '
- Applied title case

3. narrator

Checks Performed:

- Presence of constant prefix (Narratedby:)
- Name concatenation
- Multiple narrators

- Capitalization uniformity

Cleaning Steps Performed:

- Removed prefix
- Inserted missing spaces
- Standardized delimiters
- Converted to '; ' separated text
- Applied title case

4. time

Checks Performed:

- Missing values
- Zero or negative durations
- Extreme duration values
- Format consistency

Cleaning Steps Performed:

- Parsed text-based time to minutes
- Converted to total minutes
- Validated ranges
- Removed extreme values

5. releasedate

Checks Performed:

- Two-digit year ambiguity
- Invalid day/month combinations
- Expected year range (1998–2025)
- Parsing failures

Cleaning Steps Performed:

- Converted to ISO format
- Corrected two-digit years
- Validated logical day/month
- Flagged out-of-range values

6. language

Checks Performed:

- Mixed capitalization
- Leading and trailing spaces

Cleaning Steps Performed:

- Converted to lowercase
- Capitalized first letter only
- Preserved missing values

7. stars

Checks Performed:

- Presence of 'Not rated yet'
- Rating value bounds
- Maximum rating consistency
- Rating count validity

Cleaning Steps Performed:

- Converted 'Not rated yet' to null
- Parsed rating and count
- Validated numeric bounds
- Verified constant max rating

8. price

Checks Performed:

- Currency symbols and formatting
- Decimal vs integer values
- Negative values
- Extreme outliers
- Presence of 'Free' values

Cleaning Steps Performed:

- Stripped symbols
- Converted to numeric
- Converted 'Free' to 0
- Standardized two decimals
- Flagged negatives and outliers

Overall Description of the Cleaned Dataset

The dataset has undergone comprehensive normalization, validation, numeric conversion, and standardization across all major variables. Text fields are now consistent and machine-readable, numeric fields are validated for integrity, dates follow ISO standards, and pricing is fully normalized. The dataset is now reliable for statistical analysis, business intelligence dashboards, and machine learning workflows.