Data Cleaning Report

1. Books Dataset

Steps Taken:

- Stripped and converted text fields (Book-Title, Book-Author, Publisher) to lowercase for consistency.
- Filled missing values in Book-Author and Publisher with "Unknown".
- Converted Year-Of-Publication to numeric; invalid parsing was coerced to NaN.
- Years < 1000 or > 2025 were replaced with NaN, then filled with the median year.
- Detected and replaced outliers in Year-Of-Publication using IQR method.
- Checked for duplicate records.

Before Cleaning:

- Missing Book-Author and Publisher values were present.
- Out-of-range years were present
- No duplicates.

After Cleaning:

- All text normalized, no missing or invalid Year-Of-Publication, outliers replaced with median.

2. Ratings Dataset

Steps Taken:

- Removed ratings outside the 0–10 range.
- Used IQR method to detect outliers in Book-Rating.
- Replaced extreme values with median rating using a custom function.
- Checked for duplicate and missing records.

Before Cleaning:

- Ratings outside valid range.

After Cleaning:

- All ratings are between 0 and 10.
- Outliers replaced with the median.

3. Users Dataset

Steps Taken:

- Set age values <5 or >100 to NaN, then filled with median.
- Replaced outlier ages using IQR method.
- Checked for duplicate User-IDs.

Before Cleaning:

- Invalid and missing Age values.

After Cleaning:

- No outliers or missing values in age.