Data Cleaning Strategy

Datasets Used

- Books.csv – Book metadata including ISBN, title, author, year of publication, etc.
- Users.csv – User information including user ID, location, and age.
- Ratings.csv – Book ratings given by users.

1. **Missing Values Handling**

   a. **Books**

   - Converted Year-Of-Publication to numeric using pd.to_numeric(errors='coerce').
   - Removed entries with missing or invalid publication years (< 1000 or future years like > 2025).

   b. **Users**

   - Converted Age to numeric.
   - Removed rows with missing or unrealistic ages (< 5 or > 100).

   c. **Ratings**

   - Ensured Book-Rating column is numeric.
   - No missing ratings after conversion.

2. **Removing Duplicates**

   - Applied drop_duplicates() on:
     - Users
     - Books
     - Ratings

   This ensures that repeated records do not bias analysis.

3. **Standardization**

   - Trimmed whitespace from string fields like Book-Title, Book-Author.
   - Standardized formats:
     - Years are stored as integers within valid ranges.
     - All ratings are limited to the scale of 0–10.

4. **Outlier Detection using IQR**

   a. **Age (Users)**

   - Applied the Interquartile Range (IQR) method:
     - Removed users with ages beyond 1.5 × IQR from Q1 and Q3.
     - Helped eliminate improbable outliers like 0, 200, etc.

   b. **Book-Rating (Ratings)**

   - Ratings are already in a discrete scale of 0–10.
   - IQR method was applied, but no outliers were detected.

   c. **Year-Of-Publication (Books)**

   - Applied IQR filtering after removing invalid years (e.g., 0, 9999).
   - Helped in identifying extremely old or wrongly entered publication years.

5. **Visualizations**

   - Used Seaborn boxplots to visualize distributions and spot outliers for:
     - Age
     - Book-Rating
     - Year-Of-Publication

Repository structure

Data/

- Books.csv
- Users.csv
- Ratings.csv
- Book_Review.csv
- Customer_review.csv
- All_review.csv

Notebooks/

- data_cleaning.ipynb
- exploratorydataanalysis.ipynb
- recommendation.ipynb   # Collaborative & content-based models
- sentiment_analysis.ipynb# Sentiment classification model
- dashboard_dev.ipynb

Scripts/

- data_cleaning.py
- recommendation.py
- sentiment_model.py
- api_endpoints.py
- dashboard.py

Models/

- sentiment_model.pkl
- recommendation_model.pkl

Docs/

- Business_Understanding.pdf
- Project_Outline.pdf
- System_Design.pdf
- Sprint_Review.pdf

Results/

- Figures

Requirements.txt

Readme.md