Review Data Standardization and Merging Report

1. Overview of Review Datasets

Three distinct review datasets were utilized in this project, each originating from different platforms with varying schema and formats:

File Name	Source Platform	Sample Columns
Customer_Reviews.csv	Amazon	Review Title, Review Text, Rating, Date, User
book_review.csv	Goodreads	text, stars, posted_on, user_name
all_review.csv	Unknown/Mixed	review, score, timestamp, author

Each dataset had inconsistent naming conventions, formats for dates and ratings, and missing or incomplete data.

2. Standardization Approach

To ensure compatibility and analytical readiness, all datasets were standardized through the following steps:

Schema Normalization

All datasets were transformed to adhere to a unified schema with the following columns:

- review_id: Unique identifier generated using an MD5 hash
- platform: Source of the review
- review_text: Main content of the review
- rating: Standardized numerical rating (1 to 5 scale)
- review_date: Review submission date in YYYY-MM-DD format
- reviewer_name: Name of the reviewer

Data Cleaning and Formatting

- Removed special characters, HTML tags, and extraneous whitespace from textual data.
- Converted all text to UTF-8 encoding to ensure consistency across datasets.
- Standardized rating scales to a common numerical range.
- Normalized date formats using pandas.to_datetime() to ensure temporal consistency.

Handling Missing Values

- Textual fields such as review_text and reviewer_name were filled with placeholders like "No Review" and "Anonymous".

- Missing review_date entries were set to NaT (Not a Time) to retain data integrity.
- Missing ratings were defaulted to 0 to avoid analysis errors.
- The platform field was added where missing to maintain source traceability.

3. Dataset Merging Process

After standardization:

- All datasets were concatenated into a single DataFrame.
- A unique review_id was generated for each record to enable identification and avoid duplication.
- Duplicate reviews (based on the generated hash) were removed to ensure data cleanliness.

4. Issues Encountered and Resolution Strategies

Issue	Resolution Approach
Inconsistent column names and formats	Unified using a common schema and applied consistent data types
Missing or undefined platform data	Added a manual identifier to maintain source traceability
Encoding mismatches and special characters	Enforced UTF-8 encoding and used regex-based cleaning for text fields
Date format discrepancies	Standardized using pandas.to_datetime()
Variations in rating scales	Normalized to a uniform 1–5 rating scale
Missing values across various fields	Replaced with standardized placeholder values to ensure analytical utility

5. Final Output Summary

- The final merged dataset is saved as: merged_reviews.csv
- It includes the following columns:
    - review_id
    - platform
    - review_text

- rating
- review_date
- reviewer_name