

LAHORE COLLEGE FOR WOMEN UNIVERSITY

**REPORT**

**COURSE:** Data Science

**TOPIC:** Cleaning on project dataset

**SUBMITTED BY:**

**NAME:** Asma Imran

**CLASS:** BSSE-VII

**DEPARTMENT:**  Software engineering

**ROLL NO:** 2225165005

**SECTION:** A (REGULAR)

**SESSION:** 2022-2026

**SUBMITTED TO: Dr.**Mohsin

**Before & After Cleaning Report –**

**E-Commerce Recommendation system**

# **Dataset Information**

# Source: Kaggle- Amazon Product Reviews

https://www.kaggle.com/datasets/bittlingmayer/amazonreviews

Merged File: **data.csv**

# **Data Cleaning Report**

|  |  |  |
| --- | --- | --- |
| Step | Before Cleaning | After Cleaning |
| Total Rows | 541,909 | ~406,829 |
| Duplicate Rows | ~5,000 | 0 |
| Negative Quantities | ~10,000 rows | 0 |
| Missing CustomerID | ~135,000 rows | 0 (dropped NA rows) |
| UnitPrice (min) | 0.0 | >0 |
| UnitPrice (max) | 38,970 | ~600 (after IQR filter) |
| Countries | 38 countries | 38 countries |
| Description Missing | Some rows blank | Still blank (not imputed) |
| Cleaned Dataset | Raw dataset with errors | 406,829 rows, clean + valid data |