Objective

The objective of this task is to assess data quality issues in the Online Retail dataset and prepare an analysis-ready dataset for accurate revenue analysis, product performance evaluation, and sales trend analysis.

Data Quality Issues Identified

Missing Values

- Observation: The CustomerID column contains missing values in several records.
- Impact: Missing CustomerID values limit customer-level analysis. However, they do not affect product-level sales or total revenue calculations.
- Action Taken: Rows with missing CustomerID values were retained to avoid underreporting overall sales and revenue.

Duplicate Records

- Observation: Duplicate transaction records were identified in the dataset.
- Impact: Duplicate records can inflate revenue, sales quantity, and demand trends, leading to inaccurate analysis.
- Action Taken: Duplicate rows were removed based on complete record matching to ensure data accuracy.

Negative and Zero Quantities

- Observation: The Quantity column contains negative and zero values.
- Interpretation: Negative quantities indicate product returns or order cancellations.
- Impact: Including these values in revenue calculations can distort net sales figures.
- Action Taken: Negative and zero quantities were excluded from revenue calculations but retained for return and cancellation analysis.

Cancelled Transactions

- Observation: Invoice numbers starting with the letter “C” indicate cancelled transactions.
- Impact: Cancelled transactions do not represent completed sales and can misrepresent revenue figures.
- Action Taken: Cancelled invoices were excluded from net revenue analysis.

Derived Column

TotalAmount

Definition: TotalAmount = Quantity × UnitPrice

Purpose: This column is used as the primary metric for revenue calculation, sales trend analysis, and product performance evaluation.

Data Cleaning Summary

- Removed duplicate records to avoid double counting
- Retained records with missing CustomerID for accurate revenue analysis
- Excluded cancelled invoices and negative quantities from revenue calculations
- Created a TotalAmount column for precise revenue measurement

Final Data Readiness Assessment

After data cleaning and transformation, the dataset is clean, consistent, and suitable for:

Revenue analysis

- Product performance analysis
- Time-based sales trend analysis
- Geographic sales insights

Limitations

- Customer-level analysis is partially limited due to missing CustomerID values
- Profit analysis is not possible due to the absence of product cost data

Conclusion

After performing comprehensive data quality checks and cleaning actions, the Online Retail dataset is now analysis-ready. The cleaned dataset provides reliable insights and can effectively support retail business decision-making.