Documentation on Task – 1 Online Retail Sales Data

Project: Retail Sales Data Analysis

1. Introduction

Data analysis is only as reliable as the quality of the data used. Raw business data often contains missing values, duplicates, inconsistencies, and outliers that can distort insights. This document explains the data understanding, cleaning process, and preparation steps applied to the retail sales dataset before dashboard creation and analysis.

2. Dataset Description

The dataset represents transaction-level retail sales data containing invoice details, product information, customer identifiers, and revenue-related attributes.

Key Columns:

InvoiceNo – Transaction identifier

StockCode – Product code

Description – Product name

Quantity – Units sold

InvoiceDate – Transaction date

UnitPrice – Price per unit

CustomerID – Customer identifier

Country – Customer country

Year – Transaction year

Month – Transaction month

Revenue – Quantity × UnitPrice

3. Data Understanding & Initial Assessment

Identified issues:

- Missing CustomerID values

- Duplicate records

- Negative quantities (returns)

- Zero or negative unit prices

- Inconsistent date formats

4. Data Cleaning Process

• Missing values handled by excluding invalid customer records

• Duplicate rows removed

- We removed rows that contained negative values or duplicate empty entries. However, we deliberately kept certain rows even if they shared the same *Invoice ID*. The reason is rooted in real-world retail practices: during billing, customers sometimes add extra items after the initial entry. Instead of updating the quantity of the existing item, sales staff often record the new items as separate line entries under the same invoice. This approach reflects how transactions actually occur in practice—capturing the dynamic nature of customer behavior and the operational realities of billing systems.

• Negative and zero prices filtered

• Revenue recalculated

• Dates standardized

• Text cleaned and normalized

5. Data Transformation

Derived fields such as Revenue, Year, Month, and Average Order Value were added for analysis.

6. Data Validation

Ensured no duplicates, no invalid prices, and consistent formatting.

7. Conclusion

The cleaned dataset is analysis-ready and reliable for dashboards and business decision-making.