Data Cleaning & Preprocessing Documentation

# 1. Data Import & Initial Exploration

• Multiple datasets were provided: Sales, Customers, Products, Stores, Returns.

• Datasets were imported into Python (Pandas) for cleaning and preprocessing.

• Verified data types (int, float, object, datetime) and identified missing/null values.

# 2. Handling Missing Values

• Customer Age: Some entries were missing. Used median imputation based on similar age groups.

• Gender: Missing values were marked as 'Other'.

• Product Details: Rows with critical missing identifiers (Product ID, Store ID) were dropped.

# 3. Data Type Conversions

• Converted Date columns (Order Date, Return Date) into datetime format.

• Converted categorical columns (Gender, Age Group, Category, Reason) into category datatype for optimization.

# 4. Outlier Treatment

• Detected outliers in Quantity and Profit\_per\_unit using IQR (Interquartile Range) method.

• Outliers were capped at upper whisker values to reduce distortion.

# 5. Feature Engineering

• Profit\_per\_unit = Profit / Quantity (created to analyze SKU-level profitability).

• Age Grouping: Created age bins → Young Adult (18–30), Adult (31–45), Mid-adult (46–60), Senior (60+).

• Reason\_Category: Standardized return reasons into broader categories: Defective, Size Issue, Others.

# 6. Standardization & Cleaning of Categorical Data

• Ensured uniform text formatting (e.g., 'male' → 'Male').

• Removed leading/trailing whitespaces in Product & Store names.

• Deduplicated records based on OrderID + ProductID to avoid double counting.

# 7. Validation & Consistency Checks

• Cross-checked Quantity × Unit Price = Revenue for accuracy.

• Verified that Revenue – Cost = Profit.

• Ensured each Return record matched an existing OrderID.

# 8. Final Cleaned Datasets

• Output Files: sales\_cleaned.csv, customers\_cleaned.csv, products\_cleaned.csv, stores\_cleaned.csv, returns\_cleaned.csv.

• These datasets were used for SQL queries, Power BI dashboards, and the final report.

• Outcome: After these steps, the data was consistent, outliers capped, missing values handled, and new features engineered.