Data Cleaning & Preprocessing Report — E-Commerce Dataset

The dataset contains order-level transaction details including customer, product, region, pricing, and profitability information. Columns include: Order ID, Order Date, Customer Name, Region, City, Category, Sub-Category, Product Name, Quantity, Unit Price, Discount, Sales, Profit, and Payment Mode.

Issue Type	Column(s)	Problem Description	Action Taken
Missing Values	None	Dataset contained no null values	Verified with df.isnull().sum()
Duplicates	Entire rows	Some duplicate transactions possible	Checked & dropped duplicates
Data Type Mismatch	Order Date, numeric fields	Converted to proper datetime and numeric types	Standardized
Text Inconsistencies	Category, Sub-Category, City, Region	Variation in text casing or spacing	Standardized using .str.title()
Outliers	Quantity, Sales, Profit	Negative or unrealistic values	Removed
Derived Features	—	Needed more analytical fields	Created Total Revenue, Month, Year

New Columns Created:
- **Total Revenue** = Quantity × Unit Price × (1 – Discount)
- **Month** = Extracted from Order Date
- **Year** = Extracted from Order Date

Assumptions / Limitations:
- Negative sales or profit were treated as errors and removed.
- Discount was assumed to be between 0 and 1.
- No customer demographic data was included for segmentation.

Summary of Key Insights:
- Dataset is fully clean, consistent, and analysis-ready.
- Total Revenue enables profit and sales comparisons.
- Month/Year fields allow trend and seasonal analysis.
- Data integrity is strong, supporting dashboards and predictive analytics.