Exercise on Data Preprocessing & Visualization

# Business Scenario

A retail chain operating across multiple cities in India wants to improve its decision-making by analysing customer purchase behaviour. The company has recently expanded into Tier-2 and Tier-3 cities and is experiencing challenges in understanding customer preferences, seasonal demand, and city-level sales performance. The raw transactional data from their stores is messy, incomplete, and inconsistent, making it difficult for business analysts to directly use it for insights.  
The company’s AI team has approached you (the participants) to preprocess the data, clean it, and visualize key patterns that can help management make better marketing and inventory decisions.

# Dataset (Sample) - Retail\_Transactions\_2000.csv

The dataset contains 10,000+ transaction records (simulated sample of 2000 provided).

Each row represents a transaction. The following columns are included:  
- TransactionID – Unique ID for each transaction  
- CustomerID – Unique ID for customers  
- Gender – Male/Female/Other  
- Age – Age of the customer  
- City – City where purchase happened  
- ProductCategory – Category of product purchased (Electronics, Fashion, Groceries, Furniture, etc.)  
- Quantity – Units purchased  
- Price – Price per unit (₹)  
- TotalAmount – Derived column (Quantity × Price)  
- PurchaseDate – Date of purchase  
- PaymentMode – Cash, Card, UPI, Wallet

# Problem Statement

You are required to clean, preprocess, and visualize the retail transactional dataset to uncover hidden insights. The final objective is to create a clean dataset ready for analysis and generate visual dashboards that highlight customer and sales trends.

# Part A: Data Preprocessing Tasks

1. Data Inspection  
 - Load the dataset and check for size, structure, and column details.  
 - Identify missing values, duplicates, and inconsistent data.  
  
2. Handling Missing Data  
 - Replace missing Age with mean/median.  
 - Replace missing City with most frequent value.  
 - Drop rows where critical fields like TransactionID or ProductCategory are missing.  
  
3. Data Cleaning  
 - Remove duplicate transactions.  
 - Standardize categorical values (e.g., 'm', 'Male', 'MALE' → 'Male').  
 - Correct negative or zero values in Quantity or Price.  
  
4. Feature Engineering  
 - Derive TotalAmount if missing.  
 - Extract Month and DayOfWeek from PurchaseDate.  
 - Create an AgeGroup column (e.g., 18-25, 26-40, 41-60, 60+).  
  
5. Encoding & Transformation  
 - Encode categorical variables for analysis (e.g., Gender, City).  
 - Normalize numerical columns (Age, Price, TotalAmount) if required.  
  
6. Final Verification  
 - Ensure no missing/invalid values remain.  
 - Save the cleaned dataset as Retail\_Cleaned.csv.

# Part B: Data Visualization Tasks

1. Customer Demographics  
 - Age distribution of customers.  
 - Gender distribution.  
 - Customers by city (Top 10 cities).  
  
2. Sales Insights  
 - Total sales by product category.  
 - Monthly sales trend (line chart).  
 - Payment mode usage (pie chart).  
  
3. Advanced Insights  
 - Average spend per customer by age group.  
 - City-wise revenue contribution (bar chart).  
 - Heatmap of product category vs payment mode.  
  
4. Interactive Dashboard (Optional)  
 - Use Power BI, Tableau, or Python (Plotly/Matplotlib/Seaborn) to create a small dashboard combining key KPIs.