Programming for Data Scienec-Lab_EXP-6, Date:4-9-2025 (100 Marks)

Scenario:

A retail company wants to analyze its customer purchase data to understand sales trends, customer behavior, and product performance. Since real data is not available, you are required to generate synthetic data that mimics real-world sales records.

Tasks:

Data Generation (20 Marks)

- Generate a synthetic dataset of 1000 records with the following attributes:
- - CustomerID (unique ID)
- - Age (numeric, 18–65)
- - Gender (Male/Female/Other)
- - ProductCategory (e.g., Electronics, Clothing, Grocery, Furniture)
- - Quantity (1–10)
- - Price (random range depending on category)
- - PurchaseDate (random dates within the last 2 years)
- - PaymentMode (Cash, Credit Card, UPI, NetBanking)
- Introduce 5% missing values randomly across different columns.
- Save the dataset as retail_sales.csv.

Reading and Exploring Data (10 Marks)

- Import the dataset from the CSV file into R.
- Display the first few rows and summarize the dataset.
- Identify and report the number of missing values in each column.

Handling Missing Data (20 Marks)

- Apply at least three different missing value handling techniques, such as:
- - Removal of missing values (complete case analysis).
- - Mean/Median/Mode imputation.
- - Predictive imputation (e.g., using mice or missForest).
- Report which method gives the most reasonable results for this dataset.

Data Analysis (25 Marks)

- Perform at least three meaningful analyses, such as:
- - Total sales per product category.

- - Average spending per customer.
- - Trend of sales over time (monthly or quarterly).
- - Payment mode preference by customers.

Data Visualization (15 Marks)

- Create at least three appropriate visualizations, for example:
- - Bar chart of product category sales.
- - Line chart showing sales trend over time.
- - Pie/Donut chart for payment mode distribution.

Final Report (10 Marks)

- Prepare a single report in Word/PDF including:
- - Introduction (scenario + problem statement).
- - Data generation steps.
- - Handling missing values (with explanation).
- - Data analysis results.
- - Visualizations.
- - Conclusion.
- Include all R scripts used in the process.

Rubrics (100 Marks)

Component	Marks
Data Generation with 1000 records & 5% missing values	20
Importing data, exploration & summary	10
Handling missing values (3 techniques applied)	20
Data analysis (at least 3 insights)	25
Visualization (at least 3 appropriate plots)	15
Report writing with R scripts included	10
Total	100