Department of Computer Science and Engineering,
Anna University, Chennai- 600025
CS23005 – Business Analytics Lab
Practical
July – November 2026
Year/Sem : IV / VII

Exercise: 06

Data Preprocessing – Handling Missing Data and Normalization

14– Aug - 2026

Observation (5 Marks)

1. What is data preprocessing?
2. Define missing data.
3. Name any two methods for handling missing values.
4. What is normalization?
5. State one advantage of normalization in data analysis.

6. Consider the following sales dataset:

Customer ID	Age	Annual Income (₹)	Purchase Amount (₹)
C001	25	300000	12000
C002	32	450000	18000
C003	NULL	500000	22000
C004	45	NULL	25000
C005	28	350000	NULL
C006	40	600000	30000

1. Identify all missing values in the dataset.

2. Replace missing values using:

- Mean Imputation for Age
- Mean Imputation for Annual Income
- Mean Imputation for Purchase Amount

Example
Age Mean: (25+32+45+28+40)/5=34

Replace missing Age with 34.

3. Apply Min-Max Normalization to the "Annual Income" column.

$$Normalized\ Value = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Where

X = Original Value

Xmin = Minimum Value

Xmax = Maximum Value

Example

For Income = ₹300000

$$\frac{300000 - 300000}{600000 - 300000} = 0$$

For Income = ₹450000

$$\frac{450000 - 300000}{600000 - 300000} = 0.5$$

Customer ID	Normalized Income
C001	0.00
C002	0.50
C003	0.67
C004	0.40
C005	0.17
C006	1.00

Execution (20 Marks)

1. A retail company has collected customer purchase data. Some records contain missing values. Before performing business analytics, the dataset must be cleaned and normalized.

Write a Python program using Pandas and Scikit-learn to perform the following tasks:

Handling Missing Data

1. Display the dataset.

2. Identify missing values in each column.

3. Replace missing values in:

- Age
- Annual_Income
- Purchase_Amount

using the mean value of the respective column.

Data Normalization

1. Apply Min-Max Normalization to:

- Age
- Annual_Income
- Purchase_Amount

2. Display the normalized dataset.