Practical Problem Sheet – 1

Programme: M.Sc. Data Science
Semester: II
Module - 2: Multivariate Techniques
Topic: Exploratory Analysis of Multivariate Datasets & Data Preprocessing
Corresponding Theory Classes:

- Overview of Multivariate Data and Motivation
- Data Matrices and Factor-based Decompositions (Conceptual)

Objectives of the Practical

After completing this practical, students will be able to:

1. Understand multivariate datasets as matrices of observations and variables.
2. Perform exploratory data analysis using numerical summaries and visualizations.
3. Identify issues of scale, outliers, and correlation in multivariate data.
4. Understand the role of covariance/correlation matrices in multivariate analysis.
5. Use spectral decomposition of symmetric matrices as a foundation for PCA.
6. Use R or Python for preprocessing multivariate datasets.

1. **Dataset:** Use any one multivariate dataset with at least 4 variables and 50 observations. Clearly mention the dataset chosen. Describe the dataset.

A. Understanding the Data Matrix

i. Load the dataset into R/Python and display its dimensions.
ii. Identify:

a) Number of observations (rows)
b) Number of variables (columns)

iii. Classify the variables as:

a) Quantitative / Qualitative
b) Continuous / Discrete

iv. Represent the dataset as a data matrix and explain its structure in terms of rows and columns.

B. Exploratory Data Analysis (EDA)

i. Compute descriptive statistics for each quantitative variable. Comment on the highlights:

a) Min and max values, Range of variation
b) Mean
c) Median
d) Variance
e) Standard deviation

ii. Construct appropriate visualizations. Interpret:

Practical Problem Sheet – 1

Programme: M.Sc. Data Science
Semester: II
Module - 2: Multivariate Techniques
Topic: Exploratory Analysis of Multivariate Datasets & Data Preprocessing
Corresponding Theory Classes:

- Overview of Multivariate Data and Motivation
- Data Matrices and Factor-based Decompositions (Conceptual)

---

a) Histograms/ column diagrams for individual variables
b) Boxplots
c) Pairwise scatter plots / scatterplot matrix

iii. Compute the correlation matrix and interpret the strength and direction of associations.

C. Data Preprocessing

i. Check for missing values in the dataset. If present:

- Report the proportion of missing values
- Apply a suitable method to handle them (deletion or imputation)

ii. Detect potential outliers using boxplots or z-scores. Comment on their possible impact.
iii. Standardize the quantitative variables:

a) Explain why standardization is important in multivariate analysis
b) Compare summaries before and after standardization

D. Preparation for Principal Component Analysis

i. Covariance and Correlation Matrix

a) Compute the covariance matrix of the standardized data.
b) Comment on:

- Symmetry
- Diagonal and off-diagonal elements
- Interpretation of covariance values

ii. Spectral Decomposition

a) Perform **spectral decomposition** of the covariance or correlation matrix.
b) Report:

- Eigenvalues
- Corresponding eigenvectors

Practical Problem Sheet – 1

Programme: M.Sc. Data Science
Semester: II
Module - 2: Multivariate Techniques
Topic: Exploratory Analysis of Multivariate Datasets & Data Preprocessing
Corresponding Theory Classes:

- Overview of Multivariate Data and Motivation
- Data Matrices and Factor-based Decompositions (Conceptual)

---

iii. Interpretation

a) Explain the meaning of:

- Large vs small eigenvalues
- Orthogonality of eigenvectors

b) Explain how:

- Eigenvalues represent variability
- Eigenvectors define new uncorrelated directions

iv. Briefly explain how:

- Variance explained is obtained from eigenvalues
- Principal components are obtained from eigenvectors
- PCA reduces dimensionality while retaining maximum variability