SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 107
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Information Technology

Course Name – 23ADT202 Fundamental of Data science and Analytics

II Year / IV Semester

Unit 1 – Introduction to Data science
Exploratory data analysis

Introduction to EDA

EDA is a crucial step in the data science process to understand the dataset.
It helps uncover patterns, detect anomalies, and validate assumptions.
EDA uses statistical and graphical techniques for summarizing data.
The goal is to ensure comprehensive understanding before deeper analysis or modeling.

Objectives of EDA

Understanding the Data: Gain insights into the dataset's structure and characteristics.
Identifying Patterns and Trends: Detect correlations and trends that add context to the data.
Validating Assumptions: Check for inconsistencies, such as normality in regression.
Feature Selection: Identify and retain relevant variables for the analysis.

Descriptive Statistics

Descriptive statistics summarize key metrics of the dataset.

Key metrics: Mean, median, mode (central tendency), and standard deviation (spread).

Example: Age and Salary data summary – includes measures like mean, minimum, maximum.

These metrics provide a high-level understanding of data distribution.

Univariate Analysis

Univariate analysis focuses on analyzing one variable at a time.
It helps understand the distribution and characteristics of individual variables.
Visualization techniques: Histograms, boxplots, and density plots.
Example: Analyzing "Age" and "Salary" distributions separately for deeper insights.

Bivariate Analysis

Bivariate analysis explores relationships between two variables.
Visualization techniques: Scatter plots, correlation matrices, and pair plots.
Example: Examining the relationship between "Age" and "Salary" through scatter plots.
It helps uncover linear or nonlinear correlations and patterns between variables.

Multivariate Analysis

Multivariate analysis investigates interactions among multiple variables.
Common techniques: Heatmaps, Principal Component Analysis (PCA), and pairwise correlation.
Example: Identifying patterns or clusters involving "Age," "Salary," and other variables.
It offers insights into complex relationships within the data.

Common EDA Tasks

Handling Outliers: Detect outliers using boxplots or z-scores.
Analyzing Missing Data: Use visualizations like heatmaps to identify missing values.
Feature Relationships: Use correlation coefficients to assess the strength of relationships.
Example: Analyzing "Age," "Salary," and "z-score" relationships to identify trends.

Case Study: Sales Data Analysis

Objective: Analyze a retail company's sales data for trends and correlations.

Steps:

1. Load and clean the data.
2. Use descriptive statistics to understand sales performance.
3. Visualize sales trends over time using line plots.
4. Identify correlations between marketing spend and sales volume.

Example: Visualizing how marketing spend impacts sales.

Challenges in EDA

High Dimensionality: Large datasets with many variables make visualizations complex.
Noisy Data: Robust techniques are needed to detect patterns amidst noise.
Subjectivity: EDA insights can vary depending on the analyst's perspective.
These challenges must be addressed for accurate and meaningful results in EDA.

THANK YOU