SNS COLLEGE OF TECHNOLOGY

An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A' Grade

Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

Department of Computer Science and Engineering

Course Code & Title : 23ADO201 - Data Science Fundamentals

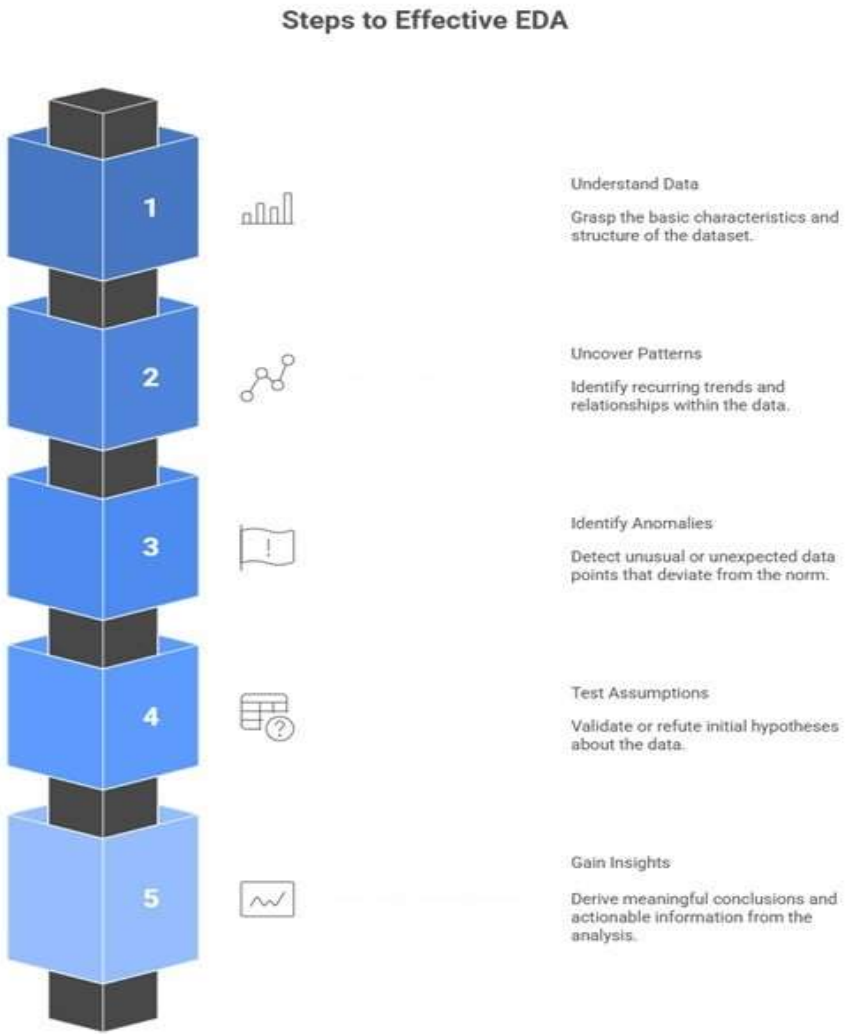
III YEAR / VI SEMESTER - EEE

Unit 1 - INTRODUCTION TO DATA SCIENCE

Topic : Exploratory Data Analysis

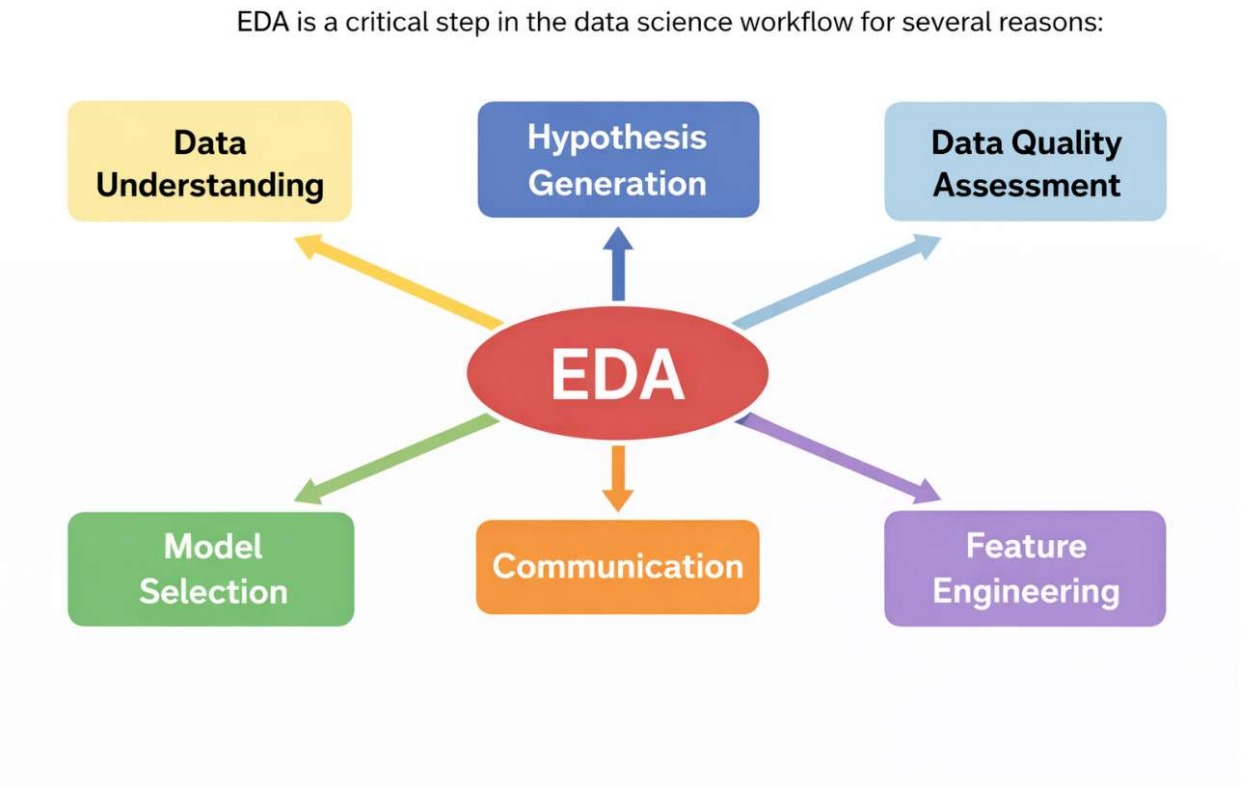
K.KARTHIKEYAN AP/CSE,SNSCT

Recall in Steps to Effective EDA



Empathy in EDA is a critical step in the data science workflow for several reasons

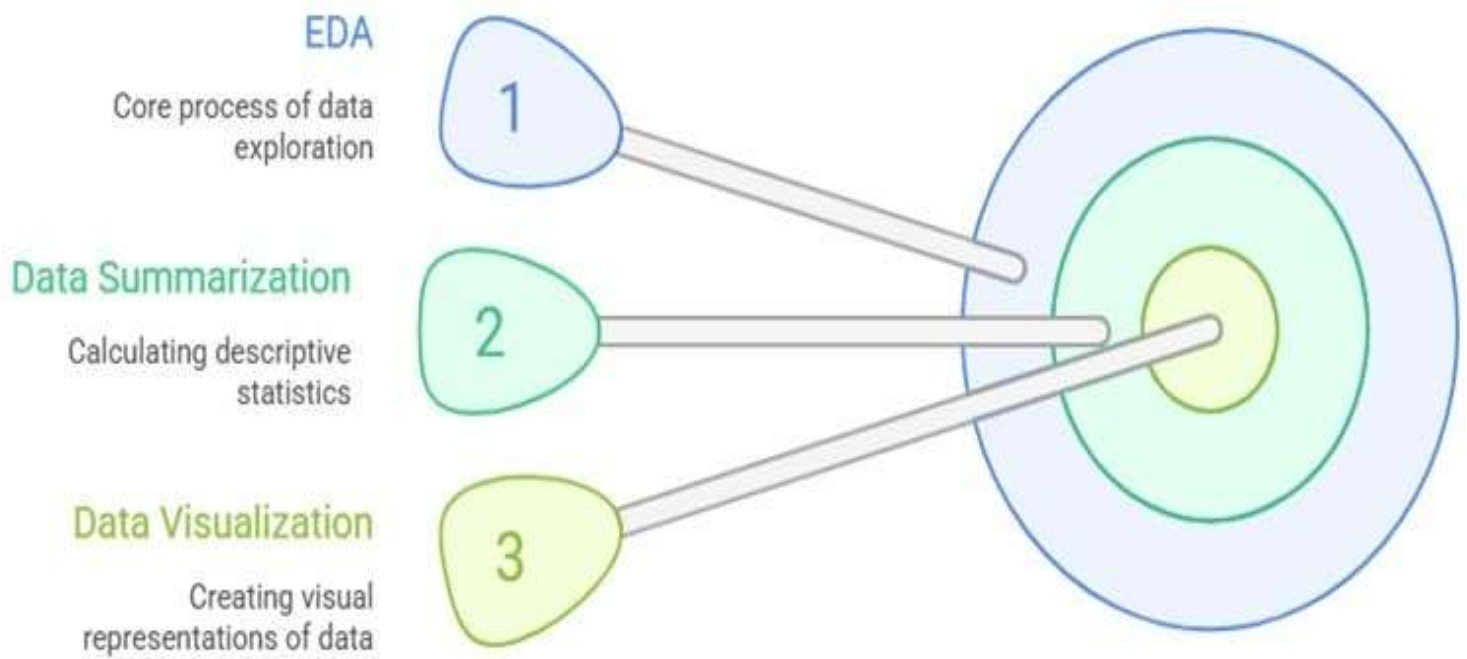
- Data Understanding
- Hypothesis Generation
- Data Quality Assessment.
- Feature Engineering
- Model Selection
- Communication



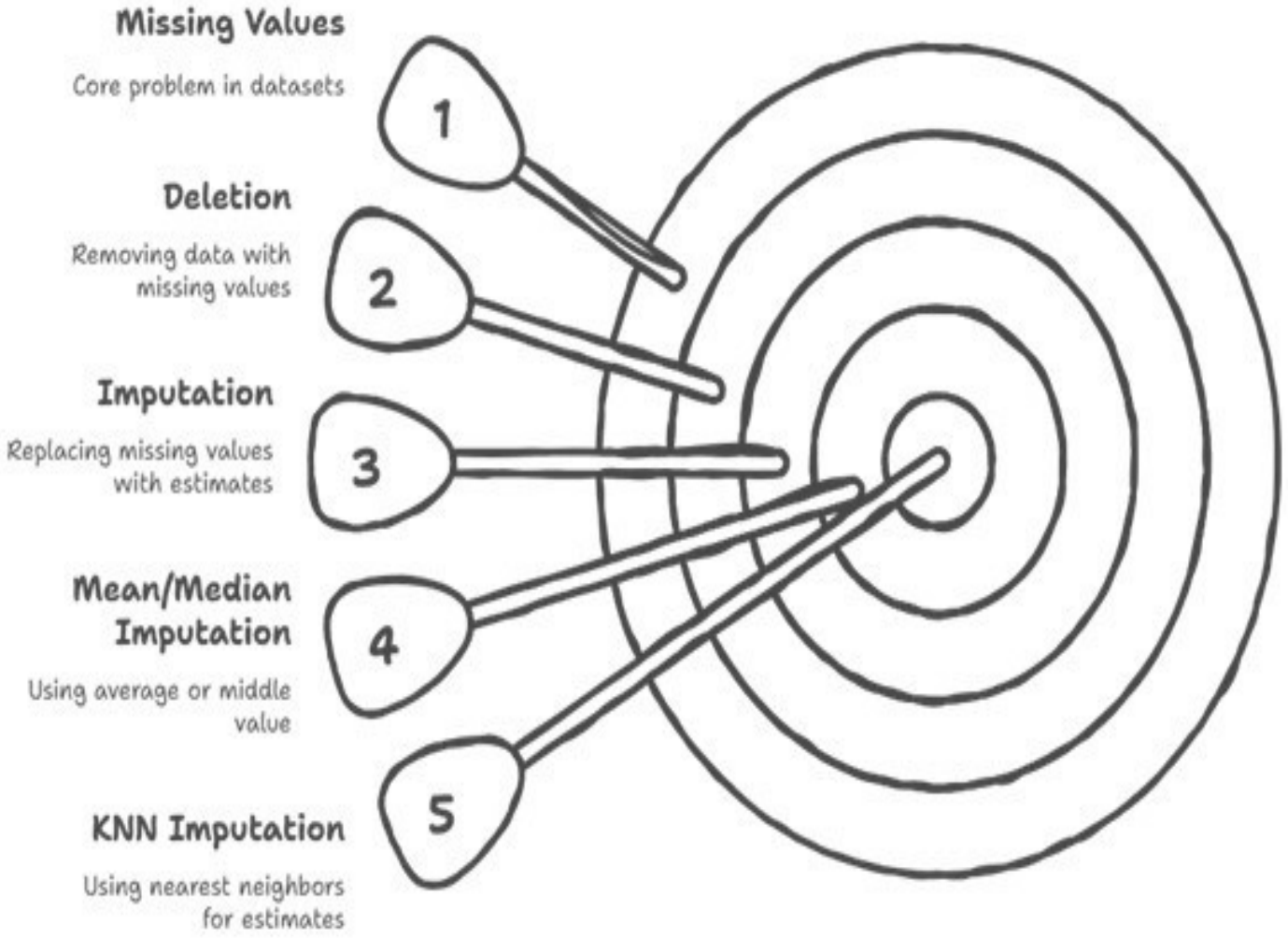
EDA Techniques

www.snsgroups.com

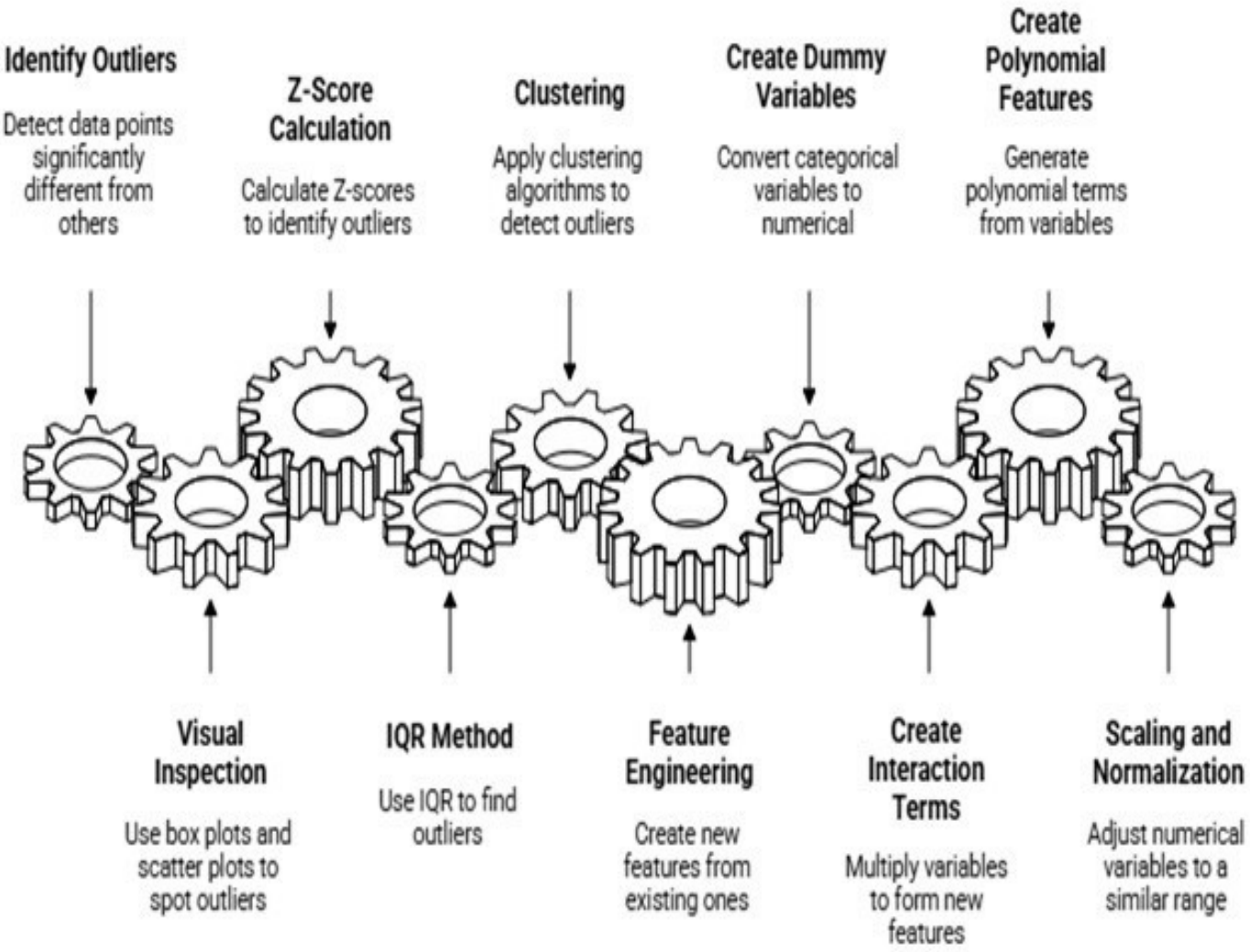
EDA Techniques



Handling Missing Values in EDA



Exploratory Data Analysis Techniques



6. Correlation Analysis

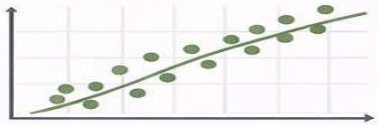
Correlation analysis involves measuring the strength and direction of the relationship between two or more variables.

Common correlation measures include:

Pearson Correlation

Measures the linear relationship between two numerical variables.

Spearman Correlation



Measures the monotonic relationship between two numerical variables.

Kendall Correlation

Measures the ordinal relationship between two numerical variables.

7. Grouping and Aggregation

Grouping and aggregation involve grouping data by one or more variables and calculating summary statistics for each group. This can help identify patterns and trends in the data. Common aggregation functions include:

Employee Productivity

Department	Hours	Tasks
Marketing	22	18
Engineering	34	25
Sales	27	20

Sum: Calculate the sum of a variable for each group.

Mean: Calculate the mean of variable for each group.

Median: Calculate the **median** of variable for each group.

Count: Count the number of of observations in each group.

Activity: Exploratory Data Analysis (EDA)

Activity Description

Students are given a real-world dataset (e.g., student performance, sales data, healthcare data) and asked to explore, summarize, visualize, and interpret the data using EDA techniques to uncover patterns, trends, and insights.

Objectives of the Activity

After completing this EDA activity, learners will be able to:

1.Understand the structure and characteristics of a dataset.

2.Identify missing values, outliers, and data quality issues.

3.Apply statistical summaries to describe data.

4.Analyze relationships between variables.

5.Use visualizations to communicate insights.

6.Formulate hypotheses based on observed patterns.

7.Prepare data for further modeling or decision-making.

Best Practices for EDA

- **Start with a clear objective**
  Define the goals of the EDA before starting.

Understand the data
Familiarize yourself with the data's source, structure, and content.

Use a variety of techniques
Combine different EDA techniques to gain a comprehensive understanding of the data.

Document your findings
Keep a record of your EDA process and findings.

Communicate your results
Share your findings with stakeholders using visualizations and summaries.

Be iterative
EDA is an iterative process. Be prepared to revisit your analysis as you gain new insights.

Use appropriate tools
Choose the right tools for the job. Common EDA tools include Python with ilibraries like Pandas, Matplotlib, and Seaborn, as well as R.

Use appropriate tools
Choose the right tools for the job. Common EDA tools include Python with ilibraries like Pandas, Matplotlib, and Seaborn, as well as R.

MIND MAP


Exploratory Data Analysis (EDA)
Data Exploration
Understand the dataset
Inspect data types
s & Outliers
Identify missing values
Numerical or Categorical?
Descriptive Statistics
Calculate summaries
Mean, Median, Mode, Std Dev
Histograms, Frequencies
Data Visualization
Histograms, Scatter plots
Bar charts, Line plots
Correlation Analysis
Pearson, Spearman, Kendall
Pie charts, Line plots
Grouping & Aggregation
Group By criteria
Sum, Mean, Median, Count
Missing Values & Outliers
Null and duplicate values
Box plots to find outliers
Descriptive Statistics
Calculate summaries
Mean, Median, Mode, Sd Dev
Data Visualization
Histograms, Scatter plots
Bar charts, Line plots
Correlation Analysis
Measure relationships
Heat map
Insight Generation
Discover patterns
Generate hypotheses

ASSESSMENT

Exploratory Data Analysis (EDA) – MCQs

1. What is the primary goal of Exploratory Data Analysis (EDA)?

A. To build predictive models

B. To clean data automatically

C. To understand data patterns and relationships

D. To deploy machine learning models

✅ **Answer:** C

2. Which of the following is usually the first step in EDA?

A. Model evaluation

B. Data visualization

C. Understanding the dataset structure

D. Feature selection

✅ **Answer:** C

REFERENCE BOOKS

1.Allen B. Downey, “Think Stats: Exploratory Data Analysis in Python”, Green Tea Press, 2014.
2.Sanjeev J. Wagh, Manisha S. Bhende, Anuradha D. Thakare, “Fundamentals of Data Science”, CRC Press, 2022.
3.Chirag Shah, “A Hands-On Introduction to Data Science”, Cambridge University Press, 2020.
4.Vineet Raina, Srinath Krishnamurthy, “Building an Effective Data Science Practice: A Framework to Bootstrap and Manage a Successful Data Science Practice”, A press, 2021.


Thank
You