

III B.Tech I Semester

23A30503c	EXPLORATORY DATA ANALYSIS WITH PYTHON (Professional Elective-I)	L	T	P	C
		3	0	0	3

Course Objectives:

- To introduce the principles and practices of Exploratory Data Analysis (EDA) using Python.
- To teach techniques for data cleaning, preprocessing, transformation, and visualization.
- To apply statistical techniques and visual methods to discover patterns and relationships.
- To gain experience using popular Python libraries such as NumPy, Pandas, Matplotlib, and Seaborn.
- To prepare datasets for further machine learning and predictive modeling.

Course Outcomes:

After completion of the course, students will be able to:

- Understand and apply key concepts of EDA and data preprocessing.
- Perform exploratory analysis using Python libraries and interpret results.
- Handle missing data, outliers, and categorical features effectively.
- Create meaningful visualizations to support data-driven insights.
- Use EDA as a foundation for data science workflows.

UNIT I – Introduction to EDA and Python Environment

Introduction to Data Science and EDA, Importance of EDA in Data Science Life Cycle, Setting up Python Environment: Jupyter, Anaconda, VS Code, Introduction to NumPy and Pandas: Arrays, Series, DataFrames, Data loading, viewing, basic operations (info, describe, shape)

UNIT II – Data Wrangling and Preprocessing

Handling Missing Data (mean, median, drop, interpolation), Dealing with Duplicates, Outliers, and Anomalies, Encoding Categorical Variables (Label, One-hot), Data Transformation: Scaling, Normalization, Binning, Data Types Conversion and Data Type Casting.

UNIT III – Univariate and Bivariate Analysis

Measures of Central Tendency and Dispersion, Distribution Plots: Histograms, Boxplots, KDE, Bar Charts, Count Plots, Pie Charts, Bivariate Analysis: Scatter Plots, Pair Plots, Heatmaps, Correlation and Covariance Analysis.

UNIT IV – Data Visualization Techniques

Visualization with Matplotlib and Seaborn, Customizing Plots: Titles, Legends, Labels, Themes, Advanced Visuals: Violin Plots, Strip Plots, Swarm Plots, Multivariate Visualization and Subplots, Plotly and Interactive Visualizations (basic overview)

UNIT V – EDA Case Studies and Real-Time Datasets

Step-by-step EDA on Sample Datasets (Titanic, Iris, Sales, etc.), Outlier Detection Techniques, Feature Engineering Techniques in EDA, EDA Report Generation using Python Notebooks, Preparing Data for Machine Learning Models

Textbooks:

1. Jake VanderPlas, Python Data Science Handbook: Essential Tools for Working with Data, O'Reilly, 2016.
2. Wes McKinney, Python for Data Analysis, 2nd Edition, O'Reilly, 2018.

Reference Books:

1. Joel Grus, Data Science from Scratch, O'Reilly, 2019.
2. Aurelien Geron, Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow, 2nd Edition, O'Reilly, 2019.
3. Allen B. Downey, Think Stats: Probability and Statistics for Programmers, O'Reilly, 2014.

Online Learning Resources:

1. NPTEL Course – Data Science for Engineers
2. Coursera – *Applied Data Science with Python Specialization (University of Michigan)*