

Course Code	Course Title	L	T	P	C
BACSE301	Exploratory Data Analysis	3	0	2	4
Pre-requisite	NIL	Syllabus Version			
		1			
Course Objectives					
1. To explain the methods for data preparation and data understanding to build solid foundation for data analysis. 2. To apply essential techniques for analyzing multivariate data by summarizing through statistical and graphical methods. 3. To summarize the role of predictive analytics, data science and data visualization in extracting insights from the data.					
Course Outcomes					
1. Analyze real world data sets by selecting and applying appropriate data analysis methods. 2. Apply data preprocessing techniques and detect outliers to prepare data for analysis. 3. Develop comprehensive statistical summaries and visual dashboards to communicate data insights. 4. Implement dimensionality reduction techniques and perform time series analysis techniques. 5. Create advanced visualizations and Develop Comprehensive Exploratory Data Analysis report for the real-world problems.					
Module:1	Introduction to Exploratory Data Analysis	6 hours			
Importance of EDA in Data Science- Overview of the EDA process- Data Collection: Sources (Databases - CSV- APIs - web scraping) - Understanding dataset structure- Metadata and feature selection. Data collection and handling - Biased datasets.					
Module:2	Data Preprocessing	11 hours			
Data Cleaning: Removing duplicates - Correcting inconsistencies. Handling Missing Values: Missing Completely at Random (MCAR) - Missing at Random (MAR) - Missing Not at Random (MNAR) - Imputation techniques - mean- median- mode- MICE. Data Type Conversion: Handling categorical and date time variables. Feature Engineering: Creating new features - binning - encoding categorical variables. Feature scaling - Min-Max-Standardization. Normalization techniques: Log-square root-power transformations. Handling imbalanced datasets”dealing with skewed distributions. Outlier Detection: Z-score- IQR method - Isolation Forest.					
Module:3	Descriptive Statistics and Visualization	9 hours			
Descriptive Statistics: Measures of central tendency and dispersion (mean- median- mode- variance- IQR - skewness- kurtosis).					

Visualization Techniques: Univariate: Histograms- Boxplots - Bar charts. Bivariate: Scatter plots- Correlation Heatmap. Multivariate: Pair plots - Parallel Coordinate Plots - Andrews Curves - Heatmap. Effective data storytelling using visualizations.		
Module:4	Dimensionality Reduction and Time Series Analysis	9 hours
Dimensionality Reduction: Filter - Wrapper - Embedded and Hybrid methods - PCA - SVD- t-SNE- Local Linear Embedding. Time Series Analysis: Time based indexing- Linear regression - Identifying trends - Seasonality and cyclic patterns. Case studies in Finance- Healthcare and Retail industries.		
Module:5	Advance Visualization Techniques	8 hours
Automated EDA Tools: Introduction to tools like Pandas Profiling- Sweetviz and D-Tale. Generating automated reports and quick insights. Advanced Visualization Techniques: Sankey diagrams - Violin plots - Radial plots. Network graphs for data relationships. EDA for Unstructured Data: Text Data - Image Data.		
Module:6	Contemporary Topics	2 hours
Total Lecture Hours:		45 hours
Text Book(s)		
1. Suresh Kumar Mukhiya and Usman Ahmed, " Hands-On Exploratory Data Analysis with Python ", Packt Publishing, 1 st Edition, 2020		
Reference Books		
1. Michael Jambu, " Exploratory and Multivariate Data Analysis ", Academic Press Inc, 1 st Edition, 1991		
2. Craig K. Enders, " Applied Missing Data Analysis ", The Guilford Press, 1 st Edition, 2010		
Indicative Experiments		
1. The Titanic dataset contains passenger demographics and voyage information; performs data cleaning and handle missing values to produce a cleaned dataset suitable for a survival prediction model. Dataset can be referred from Kaggle/UCI like public repositories.		4 hours
2. A retail sales dataset contains historical transactions and product attributes; apply data transformation and feature-engineering techniques to prepare the data for a sales prediction model. Dataset can be referred from Kaggle/UCI like public repositories.		4 hours
3. The Wine Quality dataset contains physicochemical measurements and quality scores; perform multivariate analysis to explore relationships among		4 hours

features and quality. Dataset can be referred from Kaggle/UCI like public repositories.	
4. A credit card transactions dataset contains labeled fraud and legitimate transactions; develop a suitable model to detect outliers for identifying fraudulent transactions. Dataset can be referred from Kaggle/UCI like public repositories.	4 hours
5. Modified National Institute of Standards and Technology (MNIST) is handwritten digit data represented as images; reduce the dimensionality of the images using PCA and t-SNE methods for digit recognition. Dataset can be referred from Kaggle/UCI like public repositories.	4 hours
6. The Iris dataset contains measurements for different iris species; perform exploratory data analysis and present the results using visualizations. Dataset can be referred from Kaggle/UCI like public repositories.	4 hours
7. An airline passenger dataset records monthly passenger counts; apply time series analysis to capture shifts and patterns in the data. Dataset can be referred from Kaggle/UCI like public repositories.	4 hours
8. A product review dataset contains customer reviews and metadata; perform exploratory data analysis on the review text to prepare for sentiment prediction. Dataset can be referred from Kaggle/Hugging Face/UCI like public repositories.	2 hours
Total Laboratory Hours:	
	30 hours
Mode of Evaluation: Continuous Assessment Test, Quiz, Final Assessment Test, Lab Continuous Assessment, Lab Final Assessment	
Recommended by Board of Studies :	16-10-2025
Approved by Academic Council : No. 80	20-11-2025