GUJARAT TECHNOLOGICAL UNIVERSITY

Program Name: Diploma in Engineering

Level: Diploma

Branch: Computer Engineering / Computer Science and Engineering

Subject Code : DI04000071

Subject Name: Introduction to Data Analysis

W.E. F. Academic Year:	2025-26
Semester:	4th
Category of the Course:	Professional Elective - II

Prerequisite:	Basic computer literacy, Logical Thinking and Basic Mathematics
Rationale:	The Introduction to Data Analysis course is designed to address the increasing importance of data-driven decision-making in various fields. In a world inundated with diverse data sources, this course provides students with foundational knowledge and practical skills. It covers the sources and classifications of data, introduces Big Data platforms, emphasizes the need for data analytics, and explores the evolution of analytic scalability. By incorporating modern analytic tools and the Data Analytics Lifecycle, the course ensures students are equipped to navigate real-world analytical challenges, preparing them for roles where data-driven insights are paramount.

Course Outcome:

After Completion of the Course, Student will able to:

No	Course Outcomes	RBT Level
01	Discuss various concepts of data analysis.	Understanding
02	Utilize Python toolkits to read, manipulate, extract and analyze data.	Apply
03	Apply various Statistical analysis techniques	Apply
04	Use various data visualization libraries for effective interpretations and insights of data.	Apply
05	Use Generative AI and Power BI to visualize and analyze data and create interactive dashboards.	Apply

**Revised Bloom's Taxonomy (RBT)*

Teaching and Examination Scheme:

Teaching Scheme (in Hours)			Total Credits L+T+ (PR/2)	Assessment Pattern and Marks				Total Marks
				Theory		Tutorial / Practical		
L	T	PR	C	ESE (E)	PA (M)	PA (I)	ESE (V)	
3	0	2	4	70	30	20	30	150

Course Content:

Unit No.	Content	No. of Hours	% of Weightage
1.	Introduction to Data Analysis • Sources and nature of data, classification of data (structured, semi-structured, unstructured), • Characteristics of data, • Introduction to Big Data platform, • Need of data Analysis • Evolution of analytic scalability • Analytical process • Analysis vs. reporting • Modern data analysis tools • Applications of data analysis. • Key roles for successful analysis • Various phases of data analytics lifecycle – discovery, data preparation, model planning, model building, communicating results, and operationalization	08	18
2.	Python libraries for Data Analysis and Data extraction Toolkits Using Python • NumPy - Difference between array and list, N dimension array - 1D array, 2D array, 3D array, Zeros matrix, Ones matrix, Identity matrix, Reshape, Working with random	12	26

	number, Stacking - Vertical stacking, horizontal stacking, Working with RGB Image, image as a numpy array. ● Pandas - Working with Dataframes Read csv and xlsx file, Analyze the basic dataset characteristics, Perform different merge and sort operations with multiple dataframes. Handle missing values in Dataframe, Analyze the DataFrame with loc and iloc, nlargest(), nsmallest(), add or remove an attribute from the DataFrame. Working With Data ● Reading Files ● Scraping the Web – Purpose, Legality and Ethical Considerations, Overview of popular libraries (Beautiful Soup, Requests, Selenium) ● BeautifulSoup (Purpose and Use Cases, Parsing HTML, Working with HTML tags, Accessing tag attributes, Simple HTML Parsing Example, Extracting Data from Web Pages), ● Requests (Basics of sending GET/POST requests, accessing response content, and authenticating with Requests) ● Cleaning and Munging ● Rescaling ● Data Normalization and Transformation ● Dimensionality Reduction		
3.	**Statistical Analysis** ● Regression modeling ● Multivariate analysis ● Apply basics of descriptive statistics including measures of central tendency such as mean, median, and mode Different correlation techniques: ● Pearson's Correlation Coefficient, ● Methods of Least Squares, ● scatterplots and other graphical techniques to identify the correlation between variables,	09	20

	• Different probability distributions such as Normal, Poisson, Exponential, Bernoulli Definition and Importance		
4.	**Data Visualization** • Introduction to Data Visualization • Importance of Data Visualization • Basic Data Visualization with matplotlib • Customizing matplotlib plots • Data visualization with Seaborn • Interactive visualization with Plotly • Time series data visualization • Advance plots: Yiolin plots, Box plots	08	18

5.	**Generative AI and Big Data Visualization with Power BI**
	• Recent trends in data analysis, predictive vs generative AI
	• Generative AI in Analytics: synthetic data, automated dashboards, natural language queries, storytelling with data, AI copilots in BI
	• Big Data:
	  ○ Visualizing Big Data
	  ○ Pre-attentive Attributes
	  ○ Challenges of Big Data Visualization
	• Power BI:
	  ○ Data transformation & summarization
	  ○ Dashboards: bar, line, pie charts
	  ○ AI visuals: Q&A, decomposition tree, key influencers
	  ○ Visualizing large datasets & big data considerations
	• Generative AI Challenges & Solutions: bias, fairness, privacy, security.
		08	18

	Total	**45**	**100**

Suggested Specification Table with Marks (Theory):

Distribution of Theory Marks (in %)					
R Level	**U Level**	**A Level**	**N Level**	**E Level**	**C Level**
30	**50**	**20**	**0**	**0**	**0**

Where R: Remember; U: Understanding; A: Application, N: Analyze and E: Evaluate C: Create (as per Revised Bloom's Taxonomy)

References/Suggested Learning Resources:

(a) Books:

a) Jain V.K, "Data Science and Analytics", Khanna Publishing House, Delhi
b) Jain V.K, "Big Data and Hadoop", Khanna Publishing House, Delhi
c) Jiawei Han and Jian Pei, "Data Mining Concepts and Techniques", Morgan Kaufmann, Third Edition-2011, ISBN- 978-9380931913
d) Anil Maheshwari, "Data Analytics", McGrawHil, Standard Edition-2023, ISBN- 978-9355324559
e) DavyCielen, Arno D.B. Meysman, et al., Minning, "Introducing Data Science: Big Data, Machine Learning, and More, Using Python Tools", McGrawHil, Standard Edition-2022, ISBN- 978-9355322142
f) Joel Grus, SPD, "Data Science From Scratch: First Principles with Python", Shroff/O'Reilly Second Edition, 2019,ISBN-978-9352138326
g) Pete Warden, "Big Data Glossary", O'Reilly
h) David Dietrich, Barry Heller, Beibei Yang, "Data Science and Big Data Analytics", EMC Education Series, John Wiley

(b) Open source software and website:

i) https://www.anaconda.com
j) https:// www.python.org
k) https://www.w3schools.com
l) https://swayam.gov.in/nd1_noc19_cs60/preview
m) https://nptel.ac.in/courses/106106139/
n) https://www.tutorialspoint.com

Suggested Course Practical List:

1. Data Analysis Using Microsoft Excel: Predicting the number of umbrellas sold based on rainfall using Simple Linear Regression.
2. Write a Python program that scrapes the details from the given website using BeautifulSoup and Requests.
3. Write a Pandas program to implement following operations:
    - Use the loc function to display rows where 'Survived' is 1.
    - Use the iloc function to display the value in the first row and second column.
    - Display the top 3 passengers with the largest 'Age'.

- Show the 3 Passengers with the Smallest 'Age'.

4. Write a program in Python that uses Principal Component Analysis (PCA) to reduce the dimensionality of a dataset.
5. Implement a Python program that takes a dataset with numerical features and applies min-max scaling to normalize the values between 0 and 1.
6. Load any multivariate dataset into a Pandas DataFrame and perform basic data analysis, including summary statistics, and correlation analysis.
7. Apply Descriptive Statistics in Python to Analyze Passenger Demographics on the Titanic, Including Mean, Median, and Mode.
8. Calculate and Interpret Pearson's Correlation Coefficient for Examining the Relationship Between Fare and Passenger Class on the Titanic dataset.
9. Explore Different Probability Distributions (Normal, Poisson, Exponential, Bernoulli) Using the Titanic Dataset to Analyze Survival Probabilities.
10. Create a Python script that uses Matplotlib to generate simple line charts, bar charts, and scatter plots from sample data. Customize the appearance of these plots, including labels, colors, and annotations.
11. Utilize Seaborn to create a bar plot to visualize the average income across different regions in the "tips" dataset.
12. Generate a Seaborn strip plot to visualize the distribution of total bill amounts within different days of the week in the "tips" dataset.
13. Develop an interactive line chart with Plotly to showcase the trend in sepal lengths over time using the "iris" dataset.
14. Explore data visualization tools like Tableau, Power BI. Install and explore the tool's capabilities by loading a large dataset and creating interactive visualizations.
15. To explore AI-driven analytics in Power BI by creating interactive dashboards using the Q&A visual and generating insights from natural language queries.

List of Laboratory/Learning Resources Required:

1. Computer with basic configuration with windows or unixos
2. Python Anaconda
3. Data visualization tools like Tableau, Power BI.
4. Microsoft Excel

Suggested Activities for Students:

Other than the classroom and laboratory learning, following are the suggested student-related co-curricular activities which can be undertaken to accelerate the attainment of the various outcomes in this course: Students should conduct following activities in group and prepare small reports (of 1 to 5

pages for each activity). For micro project report should be as per suggested format, for other activities students and teachers together can decide the format of the report. Students should also collect/record physical evidences such as photographs/videos of the activities for their (student's) portfolio which will be useful for their placement interviews:

1. Undertake micro-projects in teams.
2. Prepare charts to explain use/process of the identified topic.
3. https://www.codechef.com/ , in this website very elementary programs are available, students are expected to solve those programs
4. https://code.org/, an hour of code may be organized and students are encouraged to participate
5. Students are encouraged to register themselves in various MOOCs such as: Swayam, edx, Coursera, Udemy etc to further enhance their learning.
6. List the applications which are developed using C
7. Encourage students to participate in different coding competitions like hackathon, online competitions on codechef etc.
8. Encourage students to form a coding club at institute level and can help the slow learners

****