GUJARAT TECHNOLOGICAL UNIVERSITY (GTU)

Competency-focused Outcome-based Green Curriculum-2023 (COGC-2021)

Semester-VI

Course Title: **Introduction to Data Analysis**

(Course Code: 4360707)

Diploma programme in which this course is offered	**Semester in which offered**
Computer Engineering	Six

1. RATIONALE

The Introduction to Data Analysis course is designed to address the increasing importance of datadriven 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.

2. COMPETENCY

The course content should be taught and implemented with the aim to develop various types of related skills leading to the achievement of the following competency

- **Develop programs to build Data Analysis Applications.**

3. COURSE OUTCOMES (COs)

The practical exercises, the underpinning knowledge and the relevant soft skills associated with this competency are to be developed in the student to display the following COs:

The practical experiences and relevant soft skills associated with this course are to be taught and implemented, so that the student demonstrates the following industry-oriented COs associated with the above-mentioned competency:

a) Discuss various concepts of data analysis.
b) Utilize Python toolkits to read, manipulate, extract and analyze data.
c) Apply various Statistical analysis techniques.
d) Use various data visualization libraries for effective interpretations and insights of data.
e) Summarize fundamental concept of big data analysis.

4. TEACHING AND EXAMINATION SCHEME

Teaching Scheme			**Total Credits**	**Examination Scheme (In Hours) (CI+T/2+P/2)**				
Theory Marks			**Practical**	**Marks**	**Total**	**CI T**	**P**	**C CA**
	ESE	**CA**	**ESE Marks**					
3	0	2	4	70	30	25	25	150

Out of 30 marks under the theory CA, 10 marks are for assessment of the micro-project to facilitate integration of COs and the remaining 20 marks is the average of 2 tests to be taken during the semester for the assessing the attainment of the cognitive domain UOs required for the attainment of the COs.
Legends: CI-ClassRoom Instructions; T – Tutorial/Teacher Guided Theory Practice; P - Practical; C – Credit, CA - Continuous Assessment; ESE - End Semester Examination.

5. SUGGESTED PRACTICAL EXERCISES

The following practical outcomes (PrOs) that are the sub-components of the COs. *Some of the* ***PrOs*** *marked* ***'*'*** *are compulsory, as they are crucial for that particular CO. These PrOs need to be attained at least at the 'Precision Level' of Dave's Taxonomy related to 'Psychomotor Domain'.*

S. No.	**Practical Outcomes (PrOs)**	**Unit No.**	**Approx. Hrs. required**
1	Data Analysis Using Microsoft Excel: Predicting the number of umbrellas sold based on rainfall using Simple Linear Regression.	I	2
2	Write a Python program that scrapes the details from the given website using BeautifulSoup and Requests.	II	2
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'.	II	2
4	Write a program in Python that uses Principal Component Analysis (PCA) to reduce the dimensionality of a dataset.	II	2
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.	II	2
6	Load any multivariate dataset into a Pandas DataFrame and perform basic data analysis, including summary statistics, and correlation analysis.	III	2
7	Apply Descriptive Statistics in Python to Analyze Passenger Demographics on the Titanic, Including Mean, Median, and Mode.	III	2
8	Calculate and Interpret Pearson's Correlation Coefficient for Examining the Relationship Between Fare and Passenger Class on the Titanic dataset.	III	2
9	Explore Different Probability Distributions (Normal, Poisson, Exponential, Bernoulli) Using the Titanic Dataset to Analyze Survival Probabilities.	III	2
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.	IV	2

11	Utilize Seaborn to create a bar plot to visualize the average income across different regions in the "tips" dataset.	IV	2
12	Generate a Seaborn strip plot to visualize the distribution of total bill amounts within different days of the week in the "tips" dataset.	IV	2
13	Develop an interactive line chart with Plotly to showcase the trend in sepal lengths over time using the "iris" dataset.	IV	2
14	Explore data visualization tools like Tableau, Power BI, or QlikView. Install and explore the tool's capabilities by loading a large dataset and creating interactive visualizations.	V	2
	Total		28

Note

i. *More **Practical Exercises** can be designed and offered by the respective course teacher to develop the industry relevant skills/outcomes to match the COs. The above table is only a suggestive list**.***

ii. *The following are some **sample** 'Process' and 'Product' related skills (more may be added/deleted depending on the course) that occur in the above listed **Practical Exercises** of this course required which are embedded in the COs and ultimately the competency..*

S. No.	Sample Performance Indicators for the PrOs	Weightage in %
1	Correctness of program	30
2	Readability and documentation of the program/Quality of input and output displayed (messaging and formatting)	10
3	Code efficiency	20
4	Debugging ability	20
5	Program execution/answer to sample questions	20
Total		**100**

6. MAJOR EQUIPMENT/ INSTRUMENTS AND SOFTWARE REQUIRED

These major equipment/instruments and Software required to develop PrOs are given below with broad specifications to facilitate procurement of them by the administrators/management of the institutes. This will ensure conduction of practical in all institutions across the state in proper way so that the desired skills are developed in students.

S. No.	Equipment Name with Broad Specifications	PrO. No.
1	Computer with latest configuration with windows or unix os	All
2	Python Anaconda	2 To 13
3	Data visualization tools like Tableau, Power BI, or QlikView	14
4	Microsoft Excel	1

7. AFFECTIVE DOMAIN OUTCOMES

The following ***sample*** Affective Domain Outcomes (ADOs) are embedded in many of the above-mentioned COs and PrOs. More could be added to fulfil the development of this competency.

a) Follow safety practices.
b) Practice good housekeeping.
c) Demonstrate working as a leader/a team member.
d) Maintain tools and equipment
e) Follow ethical practices.

The ADOs are best developed through the laboratory/field-based exercises. Moreover, the level of achievement of the ADOs according to Krathwohl's 'Affective Domain Taxonomy' should gradually increase as planned below:

i. 'Valuing Level' in 1st year ii.

'Organization Level' in 2nd year. iii.

'Characterization Level' in 3rd year.

8. UNDERPINNING THEORY

The major Underpinning Theory is formulated as given below and only higher level UOs of *Revised Bloom's taxonomy* are mentioned for development of the COs and competency in the students by the teachers. (Higher level UOs automatically includes lower level UOs in them). If required, more such higher level UOs could be included by the course teacher to focus on attainment of COs and competency.

Unit	Unit Outcomes (UOs)	Topics and Sub-topics
Unit –1: Introducti on to Data Analysis	1a. Demonstrate an understanding of diverse data sources and their nature. 1b. Differentiate between the analytic process and reporting methodologies. 1c. Apply data analysis in various real-world applications. 1d. Navigate through various phases of the data analytics lifecycle, including discovery, data preparation, model planning, model building, communicating results, and operationalization.	**Overview of Data Analysis** 1.1 Sources and nature of data, classification of data (structured, semi-structured, unstructured), 1.2 Characteristics of data, 1.3 Introduction to Big Data platform, 1.4 Need of data Analysis 1.5 Evolution of analytic scalability 1.6 Analytical process 1.7 Analysis vs. reporting 1.8 Modern data analysis tools 1.9 Applications of data analysis. 1.10 Key roles for successful analysis 1.11 Various phases of data analytics lifecycle – discovery, data preparation, model planning, model building, communicating results, and operationalization

Unit– 2: Python libraries for Data Analysis and Data extraction	2a. Demonstrate proficiency in utilizing the Python toolkit for data analysis. 2b. Develop the ability to scrape relevant data from the web for analysis purposes. 2c. Acquire the skill of cleaning and munging data to ensure data quality and accuracy. 2d. Apply dimensionality reduction methods to simplify complex datasets while preserving essential information.	**Toolkits Using Python** 2.1 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 number, Stacking - Vertical stacking, horizontal stacking, Working with RGB Image, image as a numpy array. 2.2 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** 2.3 Reading Files 2.4 Scraping the Web – Purpose, Legality and Ethical Considerations, Overview of popular libraries (Beautiful Soup, Requests, Selenium) 2.5 BeautifulSoup (Purpose and Use Cases, Parsing HTML, Working with HTML tags, Accessing tag attributes, Simple HTML Parsing Example, Extracting Data from Web Pages), 2.6 Requests (Basics of sending GET/POST requests, accessing response content, and authenticating with Requests) 2.7 Response Status Codes 2.8 Cleaning and Munging 2.9 Manipulating Data 2.10 Rescaling 2.11 Data Normalization and Transformation 2.12 Dimensionality Reduction

Unit– 3: Statistical Analysis	3a. Apply regression modeling techniques to analyze relationships between variables. 3b. Conduct multivariate analysis for a comprehensive examination of multiple variables simultaneously. 3c. Apply basics of descriptive statistics including measures of central tendency such as mean, median, and mode. 3d. Analyse and apply various correlation techniques to uncover meaningful insights and relationships within datasets. 3d. Analyze and apply different probability distribution analysis techniques.	3.1 Regression modeling 3.2 Multivariate analysis 3.3 Apply basics of descriptive statistics including measures of central tendency such as mean, median, and mode **Different correlation techniques**: 3.4 Pearson's Correlation Coefficient, 3.5 Methods of Least Squares, 3.6 scatterplots and other graphical techniques to identify the correlation between variables, 3.7 Different probability distributions such as Normal, Poisson, Exponential, Bernoulli.
Unit–4: Data Visualization	4a. Understand the fundamentals of data visualization and its role in effective communication of insights. 4b. Create basic visualizations using the Matplotlib library for Python. 4c. Customize Matplotlib plots and Utilize Seaborn for advanced and enhanced data visualization 4d. Implement interactive data visualization techniques using Plotly. 4e. Demonstrate proficiency in creating advanced plots, including Violin plots and Box plots, for a comprehensive analysis of data distributions and outliers.	4.1 Introduction to Data Visualization 4.2 Importance of Data Visualization 4.3 Basic Data Visualization with matplotlib 4.4 Customizing matplotlib plots 4.5 Data visualization with Seaborn 4.6 Interactive data visualization with ploty 4.7 Time series data visualization 4.8 Advance plots: Yiolin plots, Box plots

Unit–5: Recent Trends in Big Data Analysis	5a. Demonstrate an understanding of recent trends in data collection and analysis techniques. 5b. Evaluate various data visualization tools for their effectiveness in handling large datasets. 5c. Apply visualization techniques specifically designed for big data to derive meaningful insights. 5d. Recognize pre-attentive attributes in data visualization for enhanced understanding. 5e. Anticipate and discuss the future progress of big data visualization, considering advancements and emerging technologies in the field.	5.1 Recent Trends in various data collection and analysis techniques 5.2 Data visualization tools – Power BI, Tableau - Data transformation, data summarization, bar charts, line charts, pie charts. 5.3 Visualizing Big Data 5.4 Pre-attentive Attributes 5.5 Challenges of Big Data Visualization 5.6 Potential Solution 5.7 Future Progress of big data visualization

***Note**: The UOs need to be formulated at the 'Application Level' and above of Revised Bloom's Taxonomy' to accelerate the attainment of the COs and the competency.*

9. SUGGESTED SPECIFICATION TABLE FOR QUESTION PAPER DESIGN

Unit No.	**Unit Title**	**Teaching Hours**	**Distribution of Theory Marks**			
			R Level	**U Level**	**A Level**	**Total Marks**
I	**Introduction to Data Analysis**	06	4	4	2	10
II	**Python libraries for Data Analysis and Data extraction**	12	2	8	8	18
III	**Statistical Analysis**	08	2	8	8	18
IV	**Data Visualization**	10	2	6	8	16
V	**Recent Trends in Big Data Analysis**	06	4	2	2	8
Total		**42**	**14**	**28**	**28**	**70**

***Legends:** R=Remember, U=Understand, A=Apply and above (Revised Bloom's taxonomy)*
***Note**: This specification table provides general guidelines to assist student for their learning and to teachers to teach and question paper designers/setters to formulate test items/questions assess the attainment of the UOs. The actual distribution of marks at different taxonomy levels (of R, U and A) in the question paper may vary from above table.*

10. SUGGESTED STUDENT ACTIVITIES

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:

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

11. SUGGESTED SPECIAL INSTRUCTIONAL STRATEGIES (if any)

These are sample strategies, which the teacher can use to accelerate the attainment of the various outcomes in this course:

a) Massive open online courses (***MOOCs***) may be used to teach various topics/sub topics.
b) Guide student(s) in undertaking micro-projects.
c) Managing Learning Environment
d) Diagnosing Essential Missed Learning concepts that will help for students.
e) Guide Students to do Personalized learning so that students can understand the course material at his or her pace.
f) Encourage students to do Group learning by sharing so that teaching can easily be enhanced.
g) ***'CI" in section No. 4***means different types of teaching methods that are to be employed by teachers to develop the outcomes.
h) About ***20% of the topics/sub-topics*** which are relatively simpler or descriptive in nature is to be given to the students for ***self-learning***, but to be assessed using different assessment methods.
i) With respect to ***section No.11***, teachers need to ensure to create opportunities and provisions for ***co-curricular activities***.
j) Guide students on how to address issues on environment and sustainability using the knowledge of this course

12. SUGGESTED MICRO-PROJECTS

Only one micro-project is planned to be undertaken by a student that needs to be assigned to him/her in the beginning of the semester. In the first four semesters, the micro-project are group-based (group of 3 to 5). However, **in the fifth and sixth semesters**, the number of students in the group should ***not exceed three.***

The micro-project could be industry application based, internet-based, workshop-based, laboratory-based or field-based. Each micro-project should encompass two or more COs which are in fact, an integration of PrOs, UOs and ADOs. Each student will have to maintain dated work diary consisting of individual contribution in the project work and give a seminar presentation of it before submission. The total work load on each student due to the micro-project should be about ***16 (sixteen) student engagement hours*** (i.e., about one hour per week) during the course. The students ought to submit micro-project by the end of the semester (so that they develop the industry-oriented COs).

A suggestive list of micro-projects is given here. This should relate highly with competency of the course and the COs. Similar micro-projects could be added by the concerned course teacher:

a) Collect data from a public API (e.g., Twitter API) or web scraping, then use Python (NumPy, Pandas) to perform basic analysis and present insights.

b) Create a regression model to predict a numerical outcome based on a dataset (e.g., housing prices based on various features)
c) Analyze and visualize time series data (e.g., stock prices) using Matplotlib and Seaborn. Explore trends, seasonality, and anomalies.
d) Develop an interactive data dashboard using Plotly to visualize and explore a dataset. Include dropdowns, sliders, and other interactive elements.
e) Explore and visualize a large dataset using Big Data visualization tools (e.g., Apache Superset, Tableau). Highlight challenges and potential solutions.
f) Perform sentiment analysis on a collection of text data (e.g., customer reviews) using NLTK. Visualize the sentiment distribution.
g) Implement a Bayesian model on a dataset with uncertainty estimates. Visualize the posterior distribution and make inferences.
h) Apply dimensionality reduction techniques (e.g., PCA) to a high-dimensional dataset and visualize the reduced features.

13. SUGGESTED LEARNING RESOURCES

S. No.	Title of Book	Author	Publication with place, year and ISBN
1	Data Science and Analytics	Jain V.K	Khanna Publishing House, Delhi
2	Big Data and Hadoop	Jain V.K	Khanna Publishing House, Delhi
3	Data Mining Concepts and Techniques	Jiawei Han and Jian Pei	Morgan Kaufmann, Third Edition2011, ISBN- 978-9380931913
4	Data Analytics	Anil Maheshwari	McGrawHil, Standard Edition-2023, ISBN- 978-9355324559
5	Introducing Data Science: Big Data, Machine Learning, and More, Using Python Tools	Davy Cielen, Arno D.B. Meysman, et al., Minning	McGrawHil, Standard Edition-2022, ISBN- 978-9355322142
6	Data Science From Scratch: First Principles with Python	Joel Grus , SPD	Shroff/O'Reilly Second Edition, 2019, ISBN-978-9352138326
7	Big Data Glossary	Pete Warden	O'Reilly
8	Data Science and Big Data Analytics	David Dietrich, Barry Heller, Beibei Yang	EMC Education Series, John Wiley

14. SUGGESTED LEARNING WEBSITES

a) https://www.anaconda.com
b) https:// www.python.org
c) https://www.w3schools.com
d) https://swayam.gov.in/nd1_noc19_cs60/preview
e) https://nptel.ac.in/courses/106106139/
f) https://www.tutorialspoint.com

15. PO-COMPETENCY-CO MAPPING

Semester II	Introduction to Data Analysis (Course Code:4360707)									
	POs and PSOs									
Competency & Course Outcomes	**PO 1 Basic & Discipline specific knowledge**	**PO 2 Problem Analysis**	**PO 3 Design/ development of solutions**	**PO 4 Engineering Tools, Experimentation &Testing**	**PO 5 Engineering practices for society, sustainability & environment**	**PO 6 Project Management**	**PO 7 Lifelong learning**	**PSO 1**	**PSO 2**	**PSO 3** (If needed)
Competency ***Use Data Analysis in various engineering applications***										
Course Outcomes CO a) Discuss various concepts of data analysis	3	2	2	2	-	-	1			
CO b) Utilize Python toolkits to read, manipulate, extract and analyze data	2	2	2	2	-	-	1			
CO c) Apply various Statistical analysis techniques	2	2	2	2	-	-	1			
CO d) Use various data visualization libraries for effective interpretations and insights of data.	2	2	2	2	-	-	1			
CO e) Summarize fundamental concept of big data analysis.	2	2	2	2	-	-	1			

Legend: **'3'** for high, **'2'** for medium, **'1'** for low or **'-'** for the relevant correlation of each competency, CO, with PO/ PSO

16. COURSE CURRICULUM DEVELOPMENT COMMITTEE

GTU Resource Persons

S. No.	Name and Designation	Institute	Contact No.	Email
1	Dr.Jigna N.Acharya, Lecturer	K.D.Polytechnic, Patan	9428752038	jignaforever@gmail.com
2	Mr. Yagnesh R. Patel, Lecturer	K.D.Polytechnic, Patan	7600501829	yagneshrpatelce@gmail.com
3	Mrs. Pravina Mehta, Lecturer	Government Polytechnic, Himmatnagar	9825446175	pravina6mehta@gmail.com