

SCSA1309	DATA FOUNDATION	L	T	P	Credits	Total Marks
		3	0	0	3	100

COURSE OBJECTIVES

- To acquire knowledge on Data science and its Foundations.
- To explore about the various data process and evaluation methods.
- To understand distinct analysis tools and practice ethical decision and actions.

UNIT 1 INTRODUCTION**9 Hrs.**

Overview of Data: Definition - Types of data – Quantitative and Qualitative (Nominal, Ordinal, Discrete and Continuous) **Big Data:** Structured, Unstructured and semi-structured - **Metadata:** Concepts of metadata – Types of metadata – Uses Data Source: **Enterprise Data Source, Social Media Data Source, Public Data Source – Web Scrapping-** Basic Concepts of **Data Warehouse and Data Mining – Distributed File System.**

UNIT 2 DATA PROCESS OVERVIEW**9 Hrs.**

Defining Goals- Data Acquisition – Sources of acquiring the data - **Data preprocessing-** Imputation of Missing values - **Data cleaning - Data Reduction, Data Transformation and Data Discretization. Exploratory Data Analysis (EDA) – Philosophy of EDA - The Data Science Process. Significance of EDA in data science - Basic tools** (plots, graphs and summary statistics) of EDA.

UNIT 3 DATA ORGANIZATION**9 Hrs.**

Data Structures: Basics – stack, Queue, Linked List, Tree, Graph - **Data Organizational Models-Centralized Model-Embedded Model- Hybrid Model-The Three-Layered structure-Centre of Excellence Model – Roles and Responsibilities- Data Governance-Data Privacy-Data Quality- Data Extraction-Extraction and ETL(Extract,Load,Transform)-Types- Physical -Logical-Data extraction with SQL.**

UNIT 4 DATA ANALYSIS & VISUALIZATION**9 Hrs.**

Spreadsheets: Data Manipulations- Sort, filter, remove duplicates-text and math functions-pivot table-lookup functions-Data visualizations for quantitative and qualitative data- charts-Excel Modelling- forecast models using advanced lookup and **data validation tools.**

Tableau: Creating Visualizations in Tableau-**Data hierarchies, filters, groups, sets, calculated fields-Map based visualizations-** Build interactive dashboards-Data Stories.

UNIT 5 ETHICS AND RECENT TRENDS**9 Hrs.**

Data and Business Insights- Data Science Engineering: - Need of Data Science - Ethics – **Doing good data science – Natural Language Processing – Machine Learning Model-** Valuing Data privacy – Getting informed consent - The Five Cs – Diversity – Inclusion – Future Trends.

Max.45 Hours.**COURSE OUTCOMES**

On completion of the course, student will be able to

- CO1:** Explore the fundamental concepts of Data science.
- CO2:** Understand Data Science Process and Tools of EDA.
- CO3:** Address how Organizational structure's influence efficiency and effectiveness.
- CO4:** Analyse and Validate data using Spreadsheets and Tableau.
- CO5:** Think through the ethics incorporating privacy, data sharing and decision-making.
- CO6:** Build interactive dashboards for Business.

SCSA2305	DATA FOUNDATION LAB	L	T	P	Credits	Total Marks
		0	0	2	1	100

COURSE OBJECTIVES

- To import and create metadata for the given data.
- To identify missing values and clean the data.
- To perform data manipulation.
- To create different views

LIST OF EXPERIMENTS**MS-Excel: Spreadsheet**

1. Load the data in excel file and apply the following operations:
 - a. Find Number of rows and columns
 - b. Add and Delete a new column
 - c. Hiding the column
 - d. Filtering the column
 - e. Find Quantitative and Qualitative values
 - f. Create new sheets and categorize the data
2. Load the data in Excel sheet and apply Pivot operation and understand the data set.
3. Apply different Arithmetic operations (sum, aggregate, average etc.,) in the data and report the result.
4. Apply statistics operations and find the central tendency (Mean, Mode, Median, Standard Deviation, and Variance) of the data set.
5. Apply different data formatting commands. (Ex. Changing date format)
6. Perform sorting and Replace operations based on selected columns (Practice Ascending and Descending order)
7. Perform different ANOVA test to find the relationship between the features (Column).
8. Perform operations to find the missing value and impute the NULL values using different imputation methods such as mean, mode, median etc.,
9. Load the file which has categorical data and encode it.
10. Perform normalization in the data.
11. Create data joins.

Tableau

12. Learn Sheet, Dashboard and story board in tableau and observe its purposes.
13. Load the data set (excel and .csv) into the tableau software and find the central tendency of the different columns.
14. Sort Top N, Bottom N
15. Apply filters

COURSE OUTCOMES

On completion of the course, the student will be able to

- CO1:** Handle data from various sources
- CO2:** Clean the data and make it suitable for processing.
- CO3:** Scale the data
- CO4:** Normalize the data and implement joins.
- CO5:** Find anomalies in the data.
- CO6:** Visualize the data by creating different views.

SCSA4001	R PROGRAMMING	L	T	P	Credits	Total Marks
		3	0	0	3	100

COURSE OBJECTIVES

- Understand how to link data, statistical methods, and actionable questions
- Use R for statistical programming, computation, graphics, and modeling.
- Fit some basic types of statistical models.
- Be able to expand their knowledge of R on their own.
- To identify types of data analytics and data visualization tools.

UNIT 1 INTRODUCTION TO R**9 Hrs.**

History and fundamentals of R, Installation and use of R / R Studio / R Shiny, Installing R packages, R – Nuts and Bolts -Getting Data In and Out –Objects in R -Arrays,DataFrame and List -Control Structures and Functions- Loop Functions-Data Manipulation- String Operations- Matrix Operations.

UNIT 2 R DATA INTERFACES**9 Hrs.**

R Data interfaces - CSV Files, XSL files, XML files, Web Data- Data Preprocessing: Missing Values, Outliers ,Principle Component Analysis - Data Visualization – Charts & Graphs-Pie Chart, Bar Chart, Box plot, Histogram, Line graph, Scatter Plot.

UNIT 3 STATISTICAL MODELING IN R**9 Hrs.**

Statistical Modeling in R - Descriptive statistics-R Packages: Regression (MASS package) - Distribution (STATS package) - ANOVA - Time Series Analysis.

UNIT 4 MACHINE LEARNING IN R**9 Hrs.**

Machine Learning in R - Classification: Decision Trees, Random Forest, SVM – Clustering: K-Means, Fuzzy - Association Rule Mining - Outlier Detection.

UNIT 5 BUILDING R SHINY APPLICATION**9 Hrs.**

Building R shiny Application: User Interface, Control Widgets, Dynamic Output - R Hadoop :Installation of R Hadoop -rhdfs – rmr2-Data Analysis with RHadoop- Case Study .

Max. 45 Hours.**COURSE OUTCOMES**

- CO 1:** List motivation for learning R programming language
CO2: Access online resources for R and import new function packages into the R workspace
CO3: Import, review, manipulate and summarize data-sets in R
CO4: Explore data-sets to create testable hypotheses and identify appropriate statistical tests
CO5: Perform appropriate statistical tests using R
CO6: Create and edit visualizations with R

TEXT / REFERENCE BOOKS

- Hands-On Programming with R: Write Your Own Functions and Simulations By Garrett Golemund, O'Reilly Media, Inc., 2014.
- R for Data Science, Hadley Wickham, Garrett Golemund,"O'Reilly Media, Inc.2016.
- Introduction to Statistics and Data Analysis - With Exercises, Solutions and Applications in R By Christian Heumann, Michael Schomaker and Shalabh, Springer, 2016
- The R Software-Fundamentals of Programming and Statistical Analysis -Pierre Lafaye de Micheaux, Rémy Drouilhet, Benoit Lique, Springer 2013
- A Beginner's Guide to R (Use R) By Alain F. Zuur, Elena N. Ieno, Erik H.W.G. Meesters, Springer 2009

END SEMESTER EXAM QUESTION PAPER PATTERN

Max. Marks: 100

Part A: 10 question of 2 marks each – No choice

Part B: 2 questions from each unit of internal choice, each carrying 16 marks

Exam Duration:3 Hrs.

20 Marks

80 Marks

SCSA2406	R PROGRAMMING LAB	L	T	P	Credits	Total Marks
		0	0	2	1	100

COURSE OBJECTIVES

- To understand the basic R programming constructs.
- To learn the R data structures namely vectors, frames and lists.
- To plot different types of graphs for effective data visualization.
- To implement machine learning algorithms in R
- To develop R shiny applications.

SUGGESTED LIST OF EXPERIMENTS

1. Implement the built-in functions in R
2. Using vectors and matrices
3. Using data frames
4. Using lists
5. Using looping constructs.
6. Creating custom functions.
7. Plotting Data (Dot plots, Histogram, Box plots, 3D plots, Scatter plots)
8. Fit Statistical Models using R
9. Implementing machine learning algorithms in R
10. Building R shiny applications.

COURSE OUTCOMES

On completion of the course the student will be able to

- CO1:** Design and develop simple R programs.
- CO2:** Create custom functions in R.
- CO3:** Create visualizations using R.
- CO4:** Compute basic summary statistics
- CO5:** Implement machine learning algorithms in R
- CO6:** Build applications in R