INTRODUCTION TO DATA SCIENCE

B.Tech. Data Science (Minor) III Year I Sem.

L	T	P	C
3	0	0	3

Course Objectives:

- Learn concepts, techniques and tools they need to deal with various facets of data science practice, including data collection and integration
- Understand the basic types of data and basic statistics
- Identify the importance of data reduction and data visualization techniques

Course Outcomes: After completion of the course, the student should be able to

- CO-1: Understand basic terms what Statistical Inference means. Identify probability distributions commonly used as foundations for statistical modeling. Fit a model to data
- CO-2: describe the data using various statistical measures
- CO-3: utilize R elements for data handling
- CO-4: perform data reduction and apply visualization techniques.

UNIT-I: Introduction

What is Data Science? - Big Data and Data Science hype – and getting past the hype - Datafication - Current landscape of perspectives - Statistical Inference - Populations and samples - Statistical modeling, probability distributions, fitting a model – Over fitting.

Basics of R: Introduction, R-Environment Setup, Programming with R, Basic Data Types.

UNIT-II: Data Types & Statistical Description

Types of Data: Attributes and Measurement, What is an Attribute? The Type of an Attribute, The Different Types of Attributes, Describing Attributes by the Number of Values, Asymmetric Attributes, Binary Attribute, Nominal Attributes, Ordinal Attributes, Numeric Attributes, Discrete versus Continuous Attributes.

Basic Statistical Descriptions of Data: Measuring the Central Tendency: Mean, Median, and Mode, Measuring the Dispersion of Data: Range, Quartiles, Variance, Standard Deviation, and Inter-quartile Range, Graphic Displays of Basic Statistical Descriptions of Data.

UNIT-III

Vectors: Creating and Naming Vectors, Vector Arithmetic, Vector sub setting,

Matrices: Creating and Naming Matrices, Matrix Sub setting, Arrays, Class.

Factors and Data Frames: Introduction to Factors: Factor Levels, Summarizing a Factor, Ordered Factors, Comparing Ordered Factors, Introduction to Data Frame, sub setting of Data Frames, Extending Data Frames, Sorting Data Frames.

Lists: Introduction, creating a List: Creating a Named List, Accessing List Elements, Manipulating List Elements, Merging Lists, Converting Lists to Vectors

UNIT-IV

Conditionals and Control Flow: Relational Operators, Relational Operators and Vectors, Logical Operators, Logical Operators and Vectors, Conditional Statements.

Iterative Programming in R: Introduction, While Loop, For Loop, Looping Over List.

Functions in R: Introduction, writing a Function in R, Nested Functions, Function Scoping, Recursion, Loading an R Package, Mathematical Functions in R.

UNIT-V:

Data Reduction: Overview of Data Reduction Strategies, Wavelet Transforms, Principal Components Analysis, Attribute Subset Selection, Regression and Log-Linear Models: Parametric Data Reduction, Histograms, Clustering, Sampling, Data Cube Aggregation.

Data Visualization: Pixel-Oriented Visualization Techniques, Geometric Projection Visualization Techniques, Icon-Based Visualization Techniques, Hierarchical Visualization Techniques, Visualizing Complex Data and Relations.

TEXT BOOKS:

1. Doing Data Science, Straight Talk from The Frontline. Cathy O'Neil and Rachel Schutt, O'Reilly, 2014
2. **Jiawei Han**, Micheline Kamber and Jian Pei. Data Mining: Concepts and Techniques, 3rd ed. The Morgan Kaufmann Series in Data Management Systems.
3. K G Srinivas, G M Siddesh, "Statistical programming in R", Oxford Publications.

REFERENCE BOOKS:

1. Introduction to Data Mining, Pang-Ning Tan, Vipin Kumar, Michael Steinbanch, Pearson Education.
2. Brain S. Everitt, "A Handbook of Statistical Analysis Using R", Second Edition, 4 LLC, 2014.
3. Dalgaard, Peter, "Introductory statistics with R", Springer Science & Business Media, 2008.
4. Paul Teetor, "R Cookbook", O'Reilly, 2011.

R PROGRAMMING LABORATORY

B.Tech. Data Science (Minor) III Year I Sem.

L	T	P	C
0	0	3	1.5

1. R Environment setup: Installation of R and RStudio in Windows
2. Write R commands for
   - i. Variable declaration and retrieving the value of the stored variables,
   - ii. Write an R script with comments,
   - iii. Type of a variable using class () Function.
3. Write R command to
   - i. illustrate summation, subtraction, multiplication, and division operations on vectors using vectors.
   - ii. Enumerate multiplication and division operations between matrices and vectors in R console
6. Write R command to
   - i. Illustrate the usage of Vector sub setting& Matrix sub setting
   - ii. Write a program to create an array of 3×3 matrixes with 3 rows and 3 columns.
   - iii. Write a program to create a class, object, and function
7. Write a command in R console
   - i. to create a tshirt_factor, which is ordered with levels ‘S’, ‘M’, and ‘L’. Is it possible to identify from the examples discussed earlier, if blood type ‘O’ is greater or less than blood type ‘A’?
   - ii. Write the command in R console to create a new data frame containing the ‘age’ parameter from the existing data frame. Check if the result is a data frame or not. Also R commands for data frame functions *cbind*(), rbind(), sort()
8. Write R command for
   - i. Create a list containing strings, numbers, vectors and logical values
   - ii. To create a list containing a vector, a matrix, and a list. Also give names to the elements in the list and display the list also access the list elements
   - iii. To add a new element at the end of the list and delete the element from the middle display the same
   - iv. To create two lists, merge two lists. Convert the lists into vectors and perform addition on the two vectors. Display the resultant vector.
9. Write R command for
   - i. logical operators—AND (&), OR (|) and NOT (!).
   - ii. Conditional Statements
   - iii. Create four vectors namely patientid, age, diabetes, and status. Put these four vectors into a data frame patientdata and print the values using a for loop& While loop
   - iv. Create a user-defined function to compute the square of an integer in R
   - v. Create a user-defined function to compute the square of an integer in R
   - vi. Recursion function for a) factorial of a number b) find nth Fibonacci number
10. Write R code for i) Illustrate Quick Sort ii) Illustrate Binary Search Tree
11. Write R command to
    - i. illustrate Mathematical functions & I/O functions
    - ii. Illustrate Naming of functions and sapply(), lapply(), tapply() &mapply()
12. Write R command for
    - i. Pie chart& 3D Pie Chart, Bar Chart to demonstrate the percentage conveyance of various ways for traveling to office such as walking, car, bus, cycle, and train
    - ii. Using a chart legend, show the percentage conveyance of various ways for traveling to office such as walking, car, bus, cycle, and train.
      - a. Walking is assigned red color, car – blue color, bus – yellow color, cycle – green color, and train – white color; all these values are assigned through *cols and lbls variables and the legend function.*
      - b. *The fill parameter is used to assign colors to the legend.*
      - c. Legend is added to the top-right side of the chart, by assigning
    - iii. Using box plots, Histogram, Line Graph, Multiple line graphs and scatter plot to demonstrate the relation between the cars speed and the distance taken to stop, Consider the parameters data and x Display the *speed and dist parameter of Cars data set using x and data parameters*

TEXT BOOK:

1. K G Srinivas, G M Siddesh, “Statistical programming in R”, Oxford Publications.