	Data Analytics with R		Semester	3
	Course Code	**BDS306C**	CIE Marks	50
	Teaching Hours/Week (L: T:P: S)	2;0;2;0	SEE Marks	50
	Total Hours of Pedagogy	28 Hours Theory + 20 Hours Practical	Total Marks	100
	Credits	03	Exam Hours	03
	Examination type (SEE)	Theory		

Course Learning objectives:

CLO 1: To Gain the knowledge of R Programming Concepts

CLO 2: To Explain the concepts of Data Visualization

CLO 3: To Explain the concept of Statistics in R.

CLO 4: To Work with R charts and Graphs

Teaching-Learning Process (General Instructions)

1. Chalk and board, power point presentations
2. Online material (Tutorials) and video lectures.
3. Demonstration of programing examples.

Module-1 **5 hours**

Basics of R

Introducing R, Initiating R, Packages in R, Environments and Functions, Flow Controls, Loops, Basic Data Types in R, Vectors

Chapter 1: 1.1 to 1.7 Chapter 2: 2.1,2.2

Module-2 **5 hours**

Basics of R Continued

Matrices and Arrays, Lists, Data Frames, Factors, Strings, Dates and Times

Chapter 2: 2.3,2.4,2.5,2.6,2.7.2.8.1,2.8.2

Module-3 **6 Hours**

Data Preparation

Datasets, Importing and Exporting files, Accessing Databases, Data Cleaning and Transformation

Chapter 3: 3.1,3.2,3.3,3.4

Module-4 **6 Hours**

Graphics using R

Exploratory Data Analysis, Main Graphical Packages, Pie Charts, Scatter Plots, Line Plots, Histograms, Box Plots, Bar Plots, Other Graphical packages

Chapter 4: 4.1 to 4.9

Module-5 **6 Hours**

Statistical Analysis using R

Basic Statistical Measures, Normal distribution, Binomial distribution, Correlation Analysis, Regression Analysis-Linear Regression Analysis of Variance

Chapter 5: 5.1, 5.3, 5.4, 5.5, 5.6.1, 5.7

Course outcome (Course Skill Set)

At the end of the course, the student will be able to :

CO1: Describe the structures of R Programming.

CO2: Illustrate the basics of Data Preparation with real world examples.

CO3: Apply the Graphical Packages of R for visualization.

CO4: Apply various Statistical Analysis methods for data analytics.

Assessment Details (both CIE and SEE)

The weightage of Continuous Internal Evaluation (CIE) is 50% and for Semester End Exam (SEE) is 50%. The minimum passing mark for the CIE is 40% of the maximum marks (20 marks out of 50) and for the SEE minimum passing mark is 35% of the maximum marks (18 out of 50 marks). A student shall be deemed to have satisfied the academic requirements and earned the credits allotted to each subject/ course if the student secures a minimum of 40% (40 marks out of 100) in the sum total of the CIE (Continuous Internal Evaluation) and SEE (Semester End Examination) taken together.

Continuous Internal Evaluation:

- For the Assignment component of the CIE, there are 25 marks and for the Internal Assessment Test component, there are 25 marks.
- The first test will be administered after 40-50% of the syllabus has been covered, and the second test will be administered after 85-90% of the syllabus has been covered
- Any two assignment methods mentioned in the 22OB2.4, if an assignment is project-based then only one assignment for the course shall be planned. The teacher should not conduct two assignments at the end of the semester if two assignments are planned.
- For the course, CIE marks will be based on a scaled-down sum of two tests and other methods of assessment.

Internal Assessment Test question paper is designed to attain the different levels of Bloom's taxonomy as per the outcome defined for the course.

Semester-End Examination:

Theory SEE will be conducted by University as per the scheduled timetable, with common question papers for the course (**duration 03 hours).**

1. The question paper will have ten questions. Each question is set for 20 marks.
2. There will be 2 questions from each module. Each of the two questions under a module (with a maximum of 3 sub-questions), **should have a mix of topics** under that module.
3. The students have to answer 5 full questions, selecting one full question from each module.
4. Marks scored shall be proportionally reduced to 50 marks

Suggested Learning Resources:

Text Books:

R Programming: An Approach to Data Analytics, G. Sudhamathy and C. Jothi Venkateswaran, MJP Publishers, 2019

Reference Books:

1..An Introduction to R, Notes on R: A Programming Environment for Data Analysis and Graphics. W. N. Venables, D.M. Smith and the R Development Core Team. Version 3.0.1 (2013-05-16)

2. Cotton, R. (2013). Learning R: A Step by Step Function Guide to Data Analysis. 1st ed. O'Reilly Media Inc
Web links and Video Lectures (e-Resources):
1. URL: https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf 2. http://www.tutorialspoint.com/r/r_tutorial.pdf 3. https://users.phhp.ufl.edu/rlp176/Courses/PHC6089/R_notes/intro.html 4. https://cran.r-project.org/web/packages/explore/vignettes/explore_mtcars.html 5. https://www.w3schools.com/r/r_stat_data_set.asp 6. https://rpubs.com/BillB/217355
Activity Based Learning (Suggested Activities in Class)/ Practical Based learning • Programming Assignment (10 Marks)

Practical Component

Sl.NO	Experiments
1	Demonstrate the steps for installation of R and R Studio. Perform the following: a) Assign different type of values to variables and display the type of variable. Assign different types such as Double, Integer, Logical, Complex and Character and understand the difference between each data type. b) Demonstrate Arithmetic and Logical Operations with simple examples. c) Demonstrate generation of sequences and creation of vectors. d) Demonstrate Creation of Matrices e) Demonstrate the Creation of Matrices from Vectors using Binding Function. f) Demonstrate element extraction from vectors, matrices and arrays
2	Assess the Financial Statement of an Organization being supplied with 2 vectors of data: Monthly Revenue and Monthly Expenses for the Financial Year. You can create your own sample data vector for this experiment) Calculate the following financial metrics: a. Profit for each month. b. Profit after tax for each month (Tax Rate is 30%). c. Profit margin for each month equals to profit after tax divided by revenue. d. Good Months – where the profit after tax was greater than the mean for the year. e. Bad Months – where the profit after tax was less than the mean for the year. f. The best month – where the profit after tax was max for the year. g. The worst month – where the profit after tax was min for the year. **Note:** a. All Results need to be presented as vectors b. Results for Dollar values need to be calculated with $0.01 precision, but need to be presented in Units of $1000 (i.e 1k) with no decimal points c. Results for the profit margin ratio need to be presented in units of % with no decimal point. d. It is okay for tax to be negative for any given month (deferred tax asset) e. Generate CSV file for the data.
3	Develop a program to create two 3 X 3 matrices A and B and perform the following operations a) Transpose of the matrix b) addition c) subtraction d) multiplication
4	Develop a program to find the factorial of given number using recursive function calls.

<table>
<tr><td>5</td><td>Develop an R Program using functions to find all the prime numbers up to a specified number by the method of Sieve of Eratosthenes.</td></tr>
<tr><td>6</td><td>The built-in data set mammals contain data on body weight versus brain weight. Develop R commands to:
a) Find the Pearson and Spearman correlation coefficients. Are they similar?
b) Plot the data using the plot command.
c) Plot the logarithm (log) of each variable and see if that makes a difference.</td></tr>
<tr><td>7</td><td>Develop R program to create a Data Frame with following details and do the following operations.
<table>
<tr><th>itemCode</th><th>itemCategory</th><th>itemPrice</th></tr>
<tr><td>1001</td><td>Electronics</td><td>700</td></tr>
<tr><td>1002</td><td>Desktop Supplies</td><td>300</td></tr>
<tr><td>1003</td><td>Office Supplies</td><td>350</td></tr>
<tr><td>1004</td><td>USB</td><td>400</td></tr>
<tr><td>1005</td><td>CD Drive</td><td>800</td></tr>
</table>
a) Subset the Data frame and display the details of only those items whose price is greater than or equal to 350.
b) Subset the Data frame and display only the items where the category is either “Office Supplies” or “Desktop Supplies”
c) Create another Data Frame called “item-details” with three different fields itemCode, ItemQtyonHand and ItemReorderLvl and merge the two frames</td></tr>
<tr><td>8</td><td>Let us use the built-in dataset air quality which has Daily air quality measurements in New York, May to September 1973. Develop R program to generate histogram by using appropriate arguments for the following statements.
a) Assigning names, using the air quality data set.
b) Change colors of the Histogram
c) Remove Axis and Add labels to Histogram
d) Change Axis limits of a Histogram
e) Add Density curve to the histogram</td></tr>
<tr><td>9</td><td>Design a data frame in R for storing about 20 employee details. Create a CSV file named “input.csv” that defines all the required information about the employee such as id, name, salary, start_date, dept. Import into R and do the following analysis.
a) Find the total number rows & columns
b) Find the maximum salary
c) Retrieve the details of the employee with maximum salary
d) Retrieve all the employees working in the IT Department.
e) Retrieve the employees in the IT Department whose salary is greater than 20000 and write these details into another file “output.csv”</td></tr>
<tr><td>10</td><td>Using the built in dataset mtcars which is a popular dataset consisting of the design and fuel consumption patterns of 32 different automobiles. The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models). Format A data frame with 32 observations on 11 variables : [1] mpg Miles/(US) gallon, [2] cyl Number of cylinders [3] disp Displacement (cu.in.), [4] hp Gross horsepower [5] drat Rear axle ratio,[6] wt Weight (lb/1000) [7] qsec 1/4 mile time, [8] vs V/S, [9] am Transmission (0 = automatic, 1 = manual), [10] gear Number of forward gears, [11] carb Number of carburetors

Develop R program, to solve the following:
a) What is the total number of observations and variables in the dataset?
b) Find the car with the largest hp and the least hp using suitable functions
c) Plot histogram / density for each variable and determine whether continuous variables are normally distributed or not. If not, what is their skewness?
d) What is the average difference of gross horse power(hp) between automobiles with 3 and 4 number of cylinders(cyl)? Also determine the difference in their standard deviations.
e) Which pair of variables has the highest Pearson correlation?</td></tr>
</table>

11	Demonstrate the progression of salary with years of experience using a suitable data set (You can create your own dataset). Plot the graph visualizing the best fit line on the plot of the given data points. Plot a curve of Actual Values vs. Predicted values to show their correlation and performance of the model. Interpret the meaning of the slope and y-intercept of the line with respect to the given data. Implement using lm function. Save the graphs and coefficients in files. Attach the predicted values of salaries as a new column to the original data set and save the data as a new CSV file.

Assessment Details (both CIE and SEE)

The weightage of Continuous Internal Evaluation (CIE) is 50% and for Semester End Exam (SEE) is 50%. The minimum passing mark for the CIE is 40% of the maximum marks (20 marks out of 50) and for the SEE minimum passing mark is 35% of the maximum marks (18 out of 50 marks). A student shall be deemed to have satisfied the academic requirements and earned the credits allotted to each subject/ course if the student secures a minimum of 40% (40 marks out of 100) in the sum total of the CIE (Continuous Internal Evaluation) and SEE (Semester End Examination) taken together.

CIE for the theory component of the IPCC (maximum marks 50)

- IPCC means practical portion integrated with the theory of the course.
- CIE marks for the theory component are **25 marks** and that for the practical component is **25 marks.**
- 25 marks for the theory component are split into **15 marks** for two Internal Assessment Tests (Two Tests, each of 15 Marks with 01-hour duration, are to be conducted) and **10 marks** for other assessment methods mentioned in 22OB4.2. The first test at the end of 40-50% coverage of the syllabus and the second test after covering 85-90% of the syllabus.
- Scaled-down marks of the sum of two tests and other assessment methods will be CIE marks for the theory component of IPCC (that is for **25 marks).**
- The student has to secure 40% of 25 marks to qualify in the CIE of the theory component of IPCC.

CIE for the practical component of the IPCC

- **15 marks** for the conduction of the experiment and preparation of laboratory record, and **10 marks** for the test to be conducted after the completion of all the laboratory sessions.
- On completion of every experiment/program in the laboratory, the students shall be evaluated including viva-voce and marks shall be awarded on the same day.
- The CIE marks awarded in the case of the Practical component shall be based on the continuous evaluation of the laboratory report. Each experiment report can be evaluated for 10 marks. Marks of all experiments' write-ups are added and scaled down to **15 marks.**
- The laboratory test **(duration 02/03 hours)** after completion of all the experiments shall be conducted for 50 marks and scaled down to **10 marks.**
- Scaled-down marks of write-up evaluations and tests added will be CIE marks for the laboratory component of IPCC for **25 marks.**
- The student has to secure 40% of 25 marks to qualify in the CIE of the practical component of the IPCC.

SEE for IPCC

Theory SEE will be conducted by University as per the scheduled timetable, with common question papers for the course (**duration 03 hours**)

1. The question paper will have ten questions. Each question is set for 20 marks.
2. There will be 2 questions from each module. Each of the two questions under a module (with a maximum of 3 sub-questions), **should have a mix of topics** under that module.
3. The students have to answer 5 full questions, selecting one full question from each module.
4. Marks scored by the student shall be proportionally scaled down to 50 Marks

The theory portion of the IPCC shall be for both CIE and SEE, whereas the practical portion will have a CIE component only. Questions mentioned in the SEE paper may include questions from the practical component.