COURSE MODULE OF THE SUBJECT TAUGHT FOR THE SESSION 2024-25 (ODD SEM)

Course Syllabi with CO's

Faculty Name:Dr Anitha D B				Academic Year: 2024 – 2025				
Department: CSE- Data Science								
Course Code	**Course Title**	**Core / Elective**	**Prerequisite**	**Contact Hours**			**Total Hrs/ Sessions**
				L	**T**	**P**	
BDS306C	Data Analytics with R	Core	Basics of Programming	2	-	2	28L+20P
Course 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							

Topics Covered as per Syllabus

Module-1

Basics of R: Introducing R, Initiating R, Packages in R, Environments and Functions, Flow Controls, Loops, Basic Data Types in R, Vectors (**5 hrs)**

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

Module-2

Basics of R Continued: Matrices and Arrays, Lists, Data Frames, Factors, Strings, Dates and Times (**5 hrs)**

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

Module-3

Data Preparation: Datasets, Importing and Exporting files, Accessing Databases, Data Cleaning and Transformation (**6 hrs)**

Chapter 3: 3.1,3.2,3.3,3.4

Module-4

Graphics using R: Exploratory Data Analysis, Main Graphical Packages, Pie Charts, Scatter Plots, Line Plots, Histograms, Box Plots, Bar Plots, Other Graphical packages (**6 hrs)**

Chapter 4: 4.1 to 4.9

Module-5

Statistical Analysis using R: Basic Statistical Measures, Normal distribution, Binomial distribution, Correlation Analysis, Regression Analysis-Linear Regression Analysis of Variance (**6 hrs)**

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

LAB COMPONENT 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.
5. Develop an R Program using functions to find all the prime numbers up to a specified number by the method of Sieve of Eratosthenes.
6. 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
7. Develop R program to create a Data Frame with the given details and do the following operations

itemCode	itemCategory	itemPrice
1001	Electronics	700
1002	Desktop Supplies	300
1003	Office Supplies	350
1004	USB	400
1005	CD Drive	800

   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
8. 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.

9. 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".
10. 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?
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.

List of Textbooks

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

List of 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

Course Outcomes	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.

Internal Assessment Marks: 50 (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 internal Assessment Tests and 10 marks for other Assessment. 25 marks for the practical component are split into 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.).

The Correlation of Course Outcomes (CO's) and Program Outcomes (PO's)

Subject Code	BDS306C	TITLE: Data Analytics with R						Faculty Name	Dr Anitha D B			
List of Course Outcomes	Program Outcomes											Total
	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12	
CO-1	2	-	-	-	3	-	-	-	2	-	-	-	7
CO-2	2	2	2	-	3	-	-	-	2	-	-	-	11
CO-3	2	2	2	-	3	-	-	-	2	-	-	-	11
CO-4	2	2	2	-	3	-	-	-	2	-	-	-	11
Total	**08**	**06**	**06**	**-**	**12**	**-**	**-**	**-**	**06**	**-**	**-**	**-**	**40**

Note: 3 = Strong Contribution, 2 = Average Contribution, 1 = Weak Contribution, - = No Contribution

The Correlation of Course Outcomes (CO's) and Program Specific Outcomes (PSO's)

Subject Code	BDS306C	TITLE: Data Analytics with R	Faculty Name	Dr Anitha D B	
List of Course Outcomes	Program Specific Outcomes				Total
	PSO-1	PSO-2	PSO-3		
CO-1	3	-	-		3
CO-2	3	1	-		4
CO-3	3	2	-		5
CO-4	3	2	-		5
Total	**12**	5	**-**		**17**