DSE-01(a): Programming Using R

CREDIT DISTRIBUTION, ELIGIBILITY AND PRE-REQUISITES OF THE COURSE

Course title & Code	Credits	Credit distribution of the course			Eligibility criteria	Pre-requisite of the course (if any)
		Lecture	Tutorial	Practical/ Practice		
Programming using R	4	2	0	2	Class XII with Mathematics	NIL

***Learning objectives*:**

1. *Master the use of the R and RStudio interactive environment.*
2. *Expand R by installing R packages.*
3. *Explore and understand how to use the R documentation.*
4. *Read Structured Data into R from various sources.*
5. *Understand the different data types in R.*
6. *Understand the different data structures in R.*

Learning Outcomes:

1. *Develop an R script and execute it*
2. *Install, load and deploy the required packages, and build new packages for sharing and reusability*
3. *Extract data from different sources using API and use it for data analysis*
4. *Visualize and summarize the data*
5. *Design application with database connectivity for data analysis*

UNIT-I **(5 hours)**

Introduction: R interpreter, Introduction to major R data structures like vectors, matrices, arrays, list and data frames, Control Structures, vectorized if and multiple selection, functions.

UNIT-II **(10 hours)**

Installing, loading and using packages: Read/write data from/in files, extracting data from web-sites, Clean data, Transform data by sorting, adding/removing new/existing columns, centering, scaling and normalizing the data values, converting types of values, using string in-built functions, Statistical analysis of data for summarizing and understanding data, Visualizing data using scatter plot, line plot, bar chart, histogram and box plot

UNIT-III **(10 hours)**

Designing GUI: Building interactive application and connecting it with database.

UNIT-IV **(5 hours)**

Building Packages.

References:

1. Cotton, R., Learning R: a step by step function guide to data analysis. 1st edition. O'reilly Media Inc.
2. Gardener, M.(2017). Beginning R: The statistical programming language, WILEY.
3. Lawrence, M., & Verzani, J. (2016). Programming Graphical User Interfaces in R. CRC press. (ebook)

List of Practical :(60 hours)

Q1. Write an R script to do the following:

a) Simulate a sample of 100 random data points from a normal distribution with mean 100 and standard deviation 5 and store the result in a vector.
b) Visualize the vector created above using different plots.
c) Test the hypothesis that the mean equals 100.
d) Use Wilcox test to test the hypothesis that mean equals 90.

Q2. Using the Algae data set from package DMwR to complete the following tasks.

a) Create a graph that you find adequate to show the distribution of the values of algae a6.
b) Show the distribution of the values of size 3.
c) Check visually if oPO4 follows a normal distribution.
d) Produce a graph that allows you to understand how the values of NO3 are distributed across the sizes of river.
e) Using a graph check if the distribution of algae a1 varies with the speed of the river.
f) Visualize the relationship between the frequencies of algae a1 and a6. Give the appropriate graph title, x-axis and y-axis title.

Q3. Read the file Coweeta.CSV and write an R script to do the following:

a) Count the number of observations per species.
b) Take a subset of the data including only those species with at least 10 observations.
c) Make a scatter plot of biomass versus height, with the symbol color varying by species, and use filled squares for the symbols. Also add a title to the plot, in italics.
d) Log-transform biomass, and redraw the plot.

Q4. The built-in data set mammals contain data on body weight versus brain weight. Write 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.

Q5. In the library MASS is a dataset UScereal which contains information about popular breakfast cereals. Attach the data set and use different kinds of plots to investigate the following relationships:

a) relationship between manufacturer and shelf
b) relationship between fat and vitamins
c) relationship between fat and shelf
d) relationship between carbohydrates and sugars
e) relationship between fiber and manufacturer
f) relationship between sodium and sugars

Q6. Write R script to:

Do two simulations of a binomial number with n = 100 and p = .5. Do you get the same results each time? What is different? What is similar?

Do a simulation of the normal two times. Once with n = 10, μ = 10 and σ = 10, the other with n = 10, μ = 100 and σ = 100. How are they different? How are they similar? Are both approximately normal?

Q.7 Create a database medicines that contains the details about medicines such as {manufacturer, composition, price}. Create an interactive application using which the user can find an alternative to a given medicine with the same composition.
Q.8 Create a database songs that contains the fields {song_name, mood, online_link_play_song}. Create an application where the mood of the user is given as input and the list of songs corresponding to that mood appears as the output. The user can listen to any song form the list via the online link given
Q.9 Create a package in R to perform certain basic statistics functions.