Code No.: **MDS 503**
Course Title:**Statistical Computing with R** Full Marks: 75
Nature: Theory +Practical (Compulsory) Credit: 3

Course Description:
This is an outcome based course to introduce basic programming in R software followed by use of R software for Statistical Computing. It focuses on the use of R software for data manipulation, data summary/data visualization, supervised and unsupervised learning sand communicate the findings.

Learning Objectives:
After completion of the course, students will be able to:

- Understand, use and apply R software for basic programming (program)
- Understand, use and apply R software for data manipulation (wrangle)
- Understand, use and apply R software for data summary and visualization (explore)
- Understand, use and apply R software for supervised learning (model)
- Understand, use and apply R software for unsupervised learning (model)
- Understand, use and apply R software to communicate findings (communicate).

Course Contents:

Unit 1: R Software for Basic Programming [8Hrs.]

R software, Statistics, Big Data and Data Science. Downloading and installing R software in Windows, Linux and Unix systems. Variables, Data types, Vectors, Lists and Matrix in R. Factors, Data Frames and Dealing with missing values in R. Logical statements, Loops, Functions and Pipes in R. Coding and naming conventions in R. Reproducible Analysis: Markdown Language, YAML Language; R Markdown/knitr document in R IDE (RStudio). Profiling and optimizing codes/scripts in R.

Unit 2: R Software for Data Manipulation [6 Hrs.]

Using R packages in R. Reading and Reviewing data in R. Manipulating and Tyding data in R. Data Wrangling in R. Data Transformation in R. Data/Text Mining in R. Big Data in R: Subsampling, Hex and 2D Density Plots.

Unit 3: R Software for Data Summary and Visualization [10Hrs.]

Basic graphics/plots in R: Bar chart and histogram, Line chart and Pie chart, Scatterplot and Boxplot, Scatterplot matrix, Social Network Analysis. The Grammar of Graphics: Data, Aesthetic mapping, Geometric objects, Statistical transformation, Scales, Coordinate system, Position adjustment and Faceting using ggplot2 package in R/RStudio. Computing measures of central tendency, dispersion, moments and relative positions in R using packages and functions/scripts.

Unit 4: R Software for Supervised Learning [10 Hrs.]

Probability Distribution Functions: Use of apply(), lapply() and sapply() fuctions in R for Breakdown Analysis. Random Sampling, Covariance and Correlation; Hypothesis Testing using common parametric and non-parametric statistical tests in R. Machine Learning and Supervised Learning. Specifying supervised models: Linear regression, Logistic Regression, Model matrices and formula. Validating models: Evaluating regression models, evaluating classification models, cross-validation, training, testing and holdouts. Supervised learning packages and its use: Decision Trees, Random Forests, Neural Networks, Support Vector Machines and Naïve Bayes.

Unit 5: R Software for Unsupervised Learning [8 Hrs.]

Dimensionality Reduction: Principle component analysis, Principle Axis Factoring, Multidimensional scaling; Clustering: k-Means clustering, Hierarchical clustering; Association rules and Monte-Carlo simulations.

Unit 6: R Software for Communication [6Hrs.]

Markdown Language, R Markdown/knitr document to produce publishable/industry level documents in HTML, PDF and Word formats. Use R Markdown to create reports, websites and dashboards. Use R Markdown to create Shiny apps for effective communication.

Practical Works:
The practical works include of class/computer lab using R/RStudiowith individual project work.

References:

1. Mailund Thomas (2017).*Beginning Data Sciences in R: Data Analysis, Visualization*, *and Modelling for the Data Scientists*.Apress: Aarhus, Denmark.
2. Goh Eric &Hui Ming (2019).*Learn R for Applied Statistics*. Apress: Singapore.
3. Wichham Hadley &GloremundGarrette (2017).*R for Data Science*. O'Reilly Media Inc: Sebastopol, Canada.

***