DATA SCIENCE USING R

Description:

The R Foundation describes R as "a language and environment for statistical computing and graphics." But, if you're familiar with R, you probably know it's a lot more than that.Created by Ross Ihaka and Robert Gentleman at the University of Auckland in New Zealand in the 1990s as a statistical platform for their students, open-source R has been extended over the decades by thousands of user-created libraries. So what is R today? It's many things:

- R is data analysis software: Data scientists, statisticians, and analysts—anyone who needs to make sense of data, really—can use R for statistical analysis, data visualization, and predictive modeling.
- R is a programming language: An object-oriented language created by statisticians, R provides objects, operators, and functions that allow users to explore, model, and visualize data.
- R is an environment for statistical analysis: Standard statistical methods are easy to implement in R, and since much of the cutting-edge research in statistics and predictive modeling is done in R, newly developed techniques are often available in R first.
- R is an open-source software project: R is free and, thanks to years of scrutiny and tinkering by users and developers, has a high standard of quality and numerical accuracy. R's open interfaces allow it to integrate with other applications and systems.
- R is a community: The R project leadership has grown to include more than 20 leading statisticians and computer scientists from around the world, and thousands of contributors have created add-on packages. With two million users, R boasts a vibrant online community.

Course Outcome:

- Analyze data find relative patterns to predict outcomes
- Analyze continuous data in varying scenarios
- Analyze and find patterns in data by applying various techniques
- Expert in Confirmatory Data analysis
- Analyze continuous data by applying various testing, regression and correlation scenarios
- Implement key components specific to text mining and analytics aided by the real world datasets and text mining
- Demonstrate expert knowledge in predicting outcomes

Session Plan :

SI.No.	Session	Topics	Content
1	Session 1	Statistics Theory	• Introduction to data Science • Scope of Data Science
2	Session 2	Statistics Theory	• Application of Data Science • Introduction to Statistics
3	Session 3	Statistics Theory	• Graphical and Tabular Descriptive Statistics • Probability
4	Session 4	Statistics Theory	• Probability distribution • Hypothesis Testing • Statistical Testing (Z-Test, Chi-Square, T-Tests, etc)
5	Session 5	R Programming, data handling and Basic Statistics	**Introduction Analytical Tool (R)** • Introduction to Data Analysis • Introduction to R programming • R environment and Basic commands
6	Session 6	R Programming, data handling and Basic Statistics	**Data Handling in R** • Importing Data • Sampling • Data Exploration • Creating calculated fields • Sorting and removing duplicates
7	Session 7	R Programming, data handling and Basic Statistics	**Basic Descriptive Statistics** • Population and sample • Measure of central tendency • Measure of dispersion
8	Session 8	R Programming, data handling and Basic Statistics	**Reporting and Data Validation** • Percentiles and Quartiles • Box plots and outlier detection • Creating graphs and reporting
9	Session 9	Project 1: Data exploration, validation and Cleaning Project	• Project on data handling • Data exploration • Data validation • Missing values identification
10	Session 10	Project 1: Data exploration, validation and Cleaning Project	• Outlier identification • Data cleaning • Basic descriptive statistics
11	Session 11	Regression Analysis	**1. Regression Analysis** • Correlation

		&Logistic Regression Model Building	• Simple regression model • R-Square • Multiple regression • Multi collinearity • Individual variable impact
12	Session 12	Regression Analysis & Logistic Regression Model Building	• Multiple regression • Multi collinearity • Individual variable impact
13	Session 13	Regression Analysis & Logistic Regression Model Building	**2. Logistic Regression** • Need of logistic regression • Logistic regression models • Validation of logistic regression models
14	Session 14	Regression Analysis & Logistic Regression Model Building	• Multiple collinearity in logistic regression • Individual impact of variables • Confusion matrix
15	Session 15	Decision Tree & Model Selection	**1. Decision Trees** • Segmentation • Entropy • Building decision trees • Validation of trees • Fine tuning and prediction using trees
16	Session 16	Decision Tree & Model Selection	**2. Model Selection and Cross validation** • How to validate a model? • What is best model? • Types of data • Types of errors • The problem of over fitting • The problem of under fitting • Bias variance of tradeoff • Cross validation • Boot strapping
17	Session 17	Project 2: Predictive Modeling Project	• Objective • Model building-1 • Model building-2 • Model validation
18	Session 18	Project 2: Predictive	• Variable selection • Model calibration

		Modeling Project	• Out of time validation
19	Session 19	Neural Network, SVM and Random Forest	**1. Neural Networks** • Neural network intuition • Neural network and vocabulary • Neural network algorithm • Math behind neural network algorithm • Building the neural networks • Validating the neural network model • Neural network applications • Image recognition using neural networks
20	Session20	Neural Network, SVM and Random Forest	**2. SVM** • Introduction • The decision boundary with largest margin • SVM-The large margin classifier • SVM-algorithm • The kernel trick • Building SVM model • Conclusion