# **R-Programming for Data Science**

Learning to Program is important if you are serious about understanding data and intend to make a career in data science. Data Science is performed on computers and you need to communicate with the computers. R-Programming is the chosen language for Data Science. R is a computer language that allows the user to program algorithms and use tools that have been programmed by others. R can do anything you can imagine, with R you can write functions, do calculations, apply most available statistical and machine learning techniques, create simple or complicated graphs and even write your own library functions.

There are many courses on doing statistics with R. But this course does not deal with statistics, as, in our experience teaching statistics and R at the same time means two steep learning curves, one on the statistical methodology and one for R code. This more than what, many students are prepared to handle. Hence, we follow the divide and conquer method where we first deal with Programming in R in this course and Subsequently Statistics and Machine Learning in other courses.

### **Detailed Syllabus.**

### Introduction to R- Programming

- What is R
  - Need for Programming in Data Science.
  - GUI vs Programming
  - Short Historical Overview of R
- The First look at R
  - Small talk on the Art of Programming.
  - Exploring the Interface.
  - Appreciating the need for an Integrated development environment (IDE).
  - Installing R Studio IDE of choice.
  - Let's use it as a glorified calculator.
- Setting up Working Directory
- GITHUB Integration
  - What is GITHUB?
  - GITHUB & R Integration.

#### **Getting Data Into R**

- First Steps
  - Typing small Datasets.
  - Concatenating data with *c* function.
  - Combining variables with *c, cbind* and *rbind* function.
- Importing data
  - o Importing Excel Data.
  - Accessing data from other data sources.

# **R** Objects

• Atomic Vectors

- What are Vectors?
- Discussion on Dimensions.
- o Doubles
- o Integers
- Characters
- Logical.
- Attributes
  - o Names
  - o DIM
- Matrices
- Arrays
- Class
  - o Dates & Times
  - o Factor
- Coercion
- Lists
- Data Frame

# Accessing Variables & Managing Subsets of Data.

- Accessing variables from a Dataframe.
  - The *str* function
  - The data argument in a function
  - The \$ Sign
  - The attach function.
  - Accessing subsets of a Data.
    - R Index Notation system
    - Sorting data.
- Combining 2 datasets with a common identifier.
- Exporting data.

# Functions

•

- What is a Function.
- The *tapply* function.
- The *sapply* and *lapply* function.
- The *summary* function.
- The *table* function.
- When should you write your own function?
- Conditional execution
- Function Arguments
- Return values
- Environment

# Plotting Graph (Visualization)

- An overview of R Graphics
- An introduction to plotting
- Histogram

#### Deccansoft Software Services

### R-Programming Syllabus

- Stem & Leaf plots
- Boxplots
- Scatterplots
- Bar Graphs
- Saving Image files

### **Programs & Loops**

- Strategy
  - Sequential Steps
  - o Parallel Case
- If statements
- Else statements
- Lookup tables
- Code comments
- Loops
  - o Expected values
  - $\circ$  Expand.grid
  - $\circ \quad \text{For loops}$
  - o While loops
  - Repeat loops

# **Application – Exploratory Data Analysis**

- Introduction to EDA
- Questions
- Variation
- Missing values
- Covariation
- Patterns & Models
- Ggplot2 calls