

Introduction to R Programming

Course Summary

Description

A comprehensive exploration for programmers of the features and functionality of the R statistical programming language held over three days. There is an optional half-day programming basic for R that precedes the course for those taking the course that many not have a programming background. As well, an optional half-day statistics refresher course is available.

The course introduces students to the R ecosystem which includes the R programming languages, the various R libraries, the programming language itself and the associated tools like R Studio and R markdown.

Topics

- R Ecosystem Overview
- R Programming Basics
- Data Operations and TidyVerse
- Data Analysis and Visualization
- Functional Programming in R
- Data Cleaning and Transformation with R
- Building Statistical Models with R
- Presenting Reports
- High Performance R Tuning
- Big Data R (optional)
- Machine Learning with R (optional)

Audience

The course is designed for both programmers and data analysts.

Prerequisites

A basic programming knowledge in a high level language is assumed as well as a basic understanding of statistics. Optional half-day primers in both of these areas can be run before the class to ensure that students have the required prerequisites before class begins.

Duration

Three days

Introduction to R Programming

Course Outline

I. R Ecosystem Overview

- A. The first module is a tour of the R ecosystem starting with an introduction R Studio and the other tools that are used by R programmers and data scientists. Emphasis is on how to integrate R into the types of projects and work flows characteristic of those that use R in professional settings. Basic management of the R environment, including finding and installing packages from CRAN (Comprehensive R Archive Networks) is covered along with how to share R notebooks and projects.

II. R Programming Basics

- A. Introduction to basic data types, variables, operators and control structures in R. Writing and running R scripts and programs and well as writing user defined functions in R are covered. Also covered are the basics of how R workspaces are managed and used by developers, adding packages, saving workspaces, importing scripts, executing scripts at the command line and using the documentation system as well as an introduction to user defined functions.

III. Data Operations and TidyVerse

- A. The different types of data structures are examined in detail including vectors, dataframes, matrices and lists. Reading and writing data in different formats is introduced as well as manipulating the data structures in various ways in order to produce the desired data structure and format. The Tidyverse package data structures, such as tibbles, and operations are introduced as well.

IV. Data Analysis and Visualization

- A. Building on the student's ability to create and manipulate data, the statistical and graphical features of R are explored using both the core R statistical functions and plotting as well as commonly used packages like ggplot2, plotly and others. This module will also cover the R markdown and other tools used to report analytic results.

V. Functional Programming in R

- A. Much of the power of R is the ability to write functional code, an important capability of any data processing language. This module introduces the basics of functional programming, including functions and first class objects, how to write and use functional code and how to apply functional programming to working with data.

VI. Data Cleaning and Transformation with R

- A. This module is a deep dive into using R to move data through the complete data preparation process from inputting raw data, data profiling and quality analysis, applying standard cleaning techniques like missing value interpolation for example, and transforming the data to the desired form for analysis. This module integrates all of the previously introduced features of R into a workflow so that students can see R is effectively applied in real world data work.

VII. Building Statistical Models with R

- A. This module builds on the previous module and is a deep dive into the statistical features of R to show students how to use R to plan, run and analyze the data. This includes using core R tools and other common packages to fit data to standard statistical models like regression modules, correlation analyses and others. Note: This module presupposes that students have a background in statistics since standard statistical tools and methods are referenced. There is an optional half-day class available on the required statistical background that can be done before the R course.

VIII. Presenting Reports

- A. This is a continuation of the previous module and focuses on using R markdown and graphical tools to produce update-able and reproducible reports and results can be published and shared.

Introduction to R Programming

Course Outline (cont'd)

IX. High Performance R Tuning

- A. This module focuses on the internals of R and how to optimize R code for more efficient processing of large amounts of data. The performance characteristics of the various R features are examined along with tools like Rprof() for profiling R code. The module provides a basic introduction into how R code is actually compiled or interpreted and executed and how this impacts performance.

X. Big Data R (optional)

- A. This module is a very high level overview of how R is used with standard big data tools like Spark and Kafka, as well as some the techniques used to process large datasets with R.

XI. Machine Learning with R (optional)

- A. R is one of the main programming languages used to build ML models. This is a very high level overview of how ML models are build using R