

Practical Data Science with R

Course Description

Data science is the practice of transforming data into actionable insight. R is the most popular open-source programming language currently in use by data scientists. In our data-driven economy, this combination of skills is in extremely high demand, commanding significant increases in salary, and is revolutionizing the world as we know it.

In this workshop, we'll learn about the practice of data science, the R programming language, and how they can be used to answer day-to-day questions about your business. In addition, we'll learn how to transform and clean our data, create and interpret descriptive statistics, data visualizations, and statistical models. We'll also learn how to apply machine learning algorithms and operationalize R for production.

Module Descriptions

1. **Introduction** – introduce the practice of data science and the R programming language
2. **Transforming and Cleaning Data** – learn how to import, transform, clean, and export data
3. **Descriptive Statistics** – learn how to create and interpret univariate and bivariate statistics
4. **Data Visualization** – learn how to create univariate, bivariate, and multivariate data visualizations
5. **Statistical Modeling** – learn to create linear and non-linear regression models for inference
6. **Handling Big Data** – learn about big data and how to handle it with tools in R
7. **Machine Learning** – learn about ML and how to train, test, and implement ML models
8. **R in Practice** – learn about R in production, reproducible research, and industry best practices

Learning Objectives

When students are finished with this workshop, they should understand the following:

Introduction

- What data science is, why it is important, and how the process of data science works
- What R is and why it has become so popular for data science
- How to create data types, data structures, subset data tables, and find help on R topics

Transforming and Cleaning Data

- What data munging is, what clean data are, and the steps involved in the data munging process
- How to use the dplyr package in R
- How to import, transform, clean, and export data

Descriptive Statistics

- What descriptive statistics are and how they can be used to make sense of data
- What types of variables exist and the corresponding types of data analysis we can perform
- How to create standard univariate and bivariate descriptive statistics

Data Visualization

- What data visualization is and how we can use it to identify patterns in data
- What types of data visualization we can create based on the question we are trying to answer
- How to create and interpret univariate, bivariate, and multivariate data visualizations

Statistical Modeling

- What a statistical model is and how it can be used for statistical inference
- How to create linear regression models and non-linear regression models

Handling Big Data

- What Big Data is, what are the limitations of R, and how to work around these limitations
- How to perform sampling and how to use R Open for big data processing

Machine Learning

- What machine learning is and how it can be used to make predictions
- How to train, test, and implement a machine learning algorithm

R in Practice

- How to use R in production with tools like R Server and shiny
- What industry best practices exist for using R for data science
- How to create reproducible research with R markdown

Course Outline

Introduction to Data Science and R

Introduction to Data Science (10 min)

- What is data science?
- Why is data science important?
- The data science process

Introduction to R (10 min)

- What is R?
- Why is R so popular for data science?
- R language basics

Lab (15 min)

- Installation and setup
- Hello World
- Primitive data types
- Data structures
- Working with data tables
- Miscellaneous topics

Transforming and Cleaning Data

Lecture (10 min)

- What is data munging?
- What are clean data?
- The data munging process
- Using dplyr

Lab (15 min)

- Importing data
- Transforming data
- Cleaning data
- Exporting data

Descriptive Statistics

Lecture (10 min)

- What are descriptive statistics?
- Types of data analysis
- Univariate descriptive statistics
- Bivariate descriptive statistics

Lab (15 min)

- Creating univariate descriptive statistics
- Creating bivariate descriptive statistics

Data Visualization

Lecture (10 min)

- What is data visualization?
- Univariate data visualizations
- Bivariate data visualizations
- Multivariate data visualizations

Lab (15 min)

- Creating univariate data visualizations
- Creating bivariate data visualizations
- Creating multivariate data visualizations

Break

Statistical Modeling

Lecture (10 min)

- What are statistical models?
- Linear regression models

Lab (15 min)

- Creating linear regression models

Handling Big Data

Lecture (15 min)

- What is Big Data?
- How to handle big data?

No Lab

Machine Learning

Lecture (10 min)

- What is machine learning?
- Types of machine learning
- The machine learning process

Lab (15 min)

- Cluster analysis

R in Practice

Lecture (10 min)

- Using R in production
- Reproducible research
- Best practices

Lab (15 min)

- Using shiny
- Creating R markdown