Practical Data Science with R

Instructor

Matthew Renze
Twitter: @matthewrenze
Email: matthew@matthewrenze.com
Web: http://www.matthewrenze.com

Course Description

Data science is the practice of transforming data into knowledge. R is the most popular programming language used by data scientists. In our data-driven economy, this combination of skills is in extremely high demand, commanding significant increases in salary, as it is revolutionizing the world around us.

In this workshop, we'll learn about the practice of data science, the R programming language, and how they can be used to transform data into actionable insight. 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 handle Big Data, make predictions using machine learning algorithms, and deploy R to production.

Prerequisites

Please bring your own Windows laptop and complete Lab 0 to install all of the necessary software before the workshop begins.

Module Descriptions

1. **Introduction** – introduce the practice of data science and the R programming language
2. **Working with 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 Gaussian models and simple linear regression models
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

Working with Data

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

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 and generate data with a Gaussian distribution model
- How to create and predict with a simple linear regression model

Handling Big Data

- What Big Data is and what are the limitations of R
- How to work around these limitations with sampling and 3rd-party tools

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
- How to predict with k-Mean cluster analysis, decision trees, naïve Bayes, and neural networks

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

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

Introduction to R

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

Lab

- Installation and setup
- Hello World
- Working with data types
- Working with data structures
- Working with data frames
- Miscellaneous topics

Working with Data

Lecture

- What is data munging?
- What are clean data?
- The data munging process
- Data munging tools

Lab

- Importing data
- Transforming data
- Cleaning data
- Exporting data
- Using dplyr

Descriptive Statistics

Lecture

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

Lab

- Creating univariate descriptive statistics
- Creating bivariate descriptive statistics

Data Visualization

Lecture

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

Lab

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

Statistical Modeling

Lecture

- What are statistical models?
- Gaussian distribution models
- Linear regression models

Lab

- Creating Gaussian distribution models
- Creating linear regression models

Handling Big Data

Lecture

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

Lab

- Using ff to work with large data sets
- Creating linear regression models with biglm

Machine Learning

Lecture

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

Lab

- Predicting with k-means cluster analysis
- Creating training and test data sets
- Predicting with decision trees
- Predicting with naïve Bayes classifiers
- Predicting with neural networks

R in Practice

Lecture

- Using R in production
- Best practices
- Reproducible research

Lab

- Exporting charts
- Using shiny
- Creating R markdown