Data Mining in R

by Deanne Larson

Module 0. About the Course (4 min)

Module 1. Introduction to RStudio (20 min)

- *Overview*
- *What is R*
- *What is RStudio*
- *Why RStudio?*
- *Navigating RStudio*
  - RStudio Interface
  - RStudio Panes
  - Source Browser
  - Navingating RStudio
  - Workspace Browser
- *R Environments*
  - R Environments
  - What is an R Package?
  - How to Install Packages
  - Exercise: Install Packages
  - Rprofile
  - Exercise: Load Packages
  - Setting up Default Environment in Rstudio

Module 2. R Basics (34 min)

- *Overview*
- *R Math*
  - R Math Overview
  - R Math Example
- *R Data Types*
  - Data Type Descriptions
  - R Data Functions
  - Vectors
  - Exercise: R Data Types
  - Exercise Review
  - Factors
  - Matrix
  - Exercise: R Data Types
  - Arrays
  - Lists
  - Data Frames
- *Working with Data Structures*
  - Indexing Vectors
  - Indexing Matrices
  - Indexing Arrays
  - Indexing Lists
  - Indexing Data Frames
  - Exercise: Indexing Data Frames
  - Search and List
  - Attach and Detach
  - Masking Errors
  - With Function

    - Assignment Operators
- *Loading Data*
    - Data Sets
    - Exercise: Loading Data
    - Loading Data Exercise Review
    - Diagnostic Functions
    - Exercise: Diagnostic Functions
- *Writing Data*
- *Summary*

Module 3. Introduction to Data Mining in R (89 min)

- *Overview*
    - What is Data Mining
    - Why Data Mining
    - The Process of Data Science
- *Exploratory Data Analysis*
    - Introduction to EDA
    - Descriptive and Inferential Statistics
    - EDA Steps
    - Summary Statistics in R
    - Exercise: Summary Statistics in R
    - The Box Plot
    - Variance and Standard Deviation Definitions
    - Variance and Standard Deviation Functions
    - Exercise: Variance and Standard Deviation
    - Covariance and Correlation Coefficient
    - Exercise: Covariance and Correlation Coefficient
    - Skewness and Kurtosis
    - Density Plot
- *Base Graphics in R*
    - Introduction to Base Graphics in R
    - Saving and Viewing Graphics
    - Dot Plot
    - Bar Charts
    - Line Graph
    - Pie Chart
    - Statistical Modeling
- *Linear Regression*
    - Introduction to Linear Regression
    - Function lm()
    - Lm() Example
    - Function predict()
    - Error Difference
    - Confidence and Prediction Intervals
    - Summary
- *Logistic Regression*
    - Introduction to Logistic Regression
    - How to Execute Logistic Regression
    - Logistic Regression Example
    - Significance Test

  - Exercise: Logistic Regression
- *Summary*

Module 4. Classification and Clustering Models in R (56 min)

- *Overview*
- *Decision Trees*
  - Decision Tree Models
  - Classification Modeling
  - How Decision Trees Work
  - Function rpart()
  - Train Data Set
  - Rpart() examples
  - Pruning
  - Advantages and Disadvantages of Decision Trees
  - Exercise: Decision Trees in R
  - Function randomForest()
  - randomForest Models
  - randomForest Example
  - Error Rates Graph
  - Exercise: randomForest
  - K Nearest Neighbor
  - Function knn()
  - Knn() Example
  - Advantages and Disadvantages of knn()
- *Clustering Models*
  - Clustering
  - Function kmeans()
  - Kmeans() Example
  - Advantages and Disadvantages of kmeans
  - Exercise: Clustering
- *Model Diagnostics*
- *Summary*