Introduction to Data Analytics with R

A course designed by Cathy Trinh for HBT

This training provides a comprehensive introduction to data analytics using R, covering essential topics such as data manipulation, exploratory data analysis, statistical analysis, data visualization, machine learning, and more.

Training curriculum

Module 1: Introduction to Data Analytics and R

- Understanding the role of data analytics in decision-making
- Introduction to R and its advantages for data analytics
- Setting up the R environment (RStudio, packages, etc.)
- Basic R syntax and data structures

Module 2: Data Manipulation with R

- Importing data into R (CSV, Excel, databases)
- Exploring and summarizing data using R functions
- Data cleaning and preprocessing techniques
- Filtering, sorting, and transforming data in R

Module 3: Exploratory Data Analysis (EDA)

- Principles and goals of EDA
- Visualizing data using R (ggplot2, base plots)
- Descriptive statistics and data distributions
- Correlation analysis and data relationships

Module 4: Data Wrangling and Manipulation in R

- Advanced data manipulation techniques (dplyr, tidyr)
- Reshaping data (melting, casting, pivot tables)
- Handling missing values and outliers
- Combining and merging datasets

Module 5: Data Visualization and Reporting

- Advanced data visualization using R (ggplot2, plotly)
- Creating interactive visualizations and dashboards
- Exporting plots and reports from R
- Presenting findings and insights effectively

Module 6: Statistical Analysis with R

- Introduction to statistical concepts in data analysis
- Hypothesis testing using R (t-tests, chi-square tests, etc.)
- Analysis of variance (ANOVA) and post-hoc tests
- Linear regression and model building

Module 7: Introduction to Machine Learning with R

- Overview of machine learning concepts
- Supervised and unsupervised learning algorithms
- Model training and evaluation in R
- Introduction to popular R packages (caret, randomForest, etc.)

Module 8: Text Mining and Natural Language Processing (NLP) in R

- Basics of text mining and NLP
- Text preprocessing and feature extraction
- Sentiment analysis and text classification
- Topic modeling and text visualization

Module 9: Time Series Analysis with R

- Understanding time series data
- Time series decomposition and visualization
- Forecasting techniques (ARIMA, exponential smoothing)
- Evaluating time series models in R

Module 10: Case Studies and Real-World Applications

- Applying data analytics techniques to real-world datasets
- Case studies and examples from various industries
- Best practices and tips for effective data analytics with R
- Final project and presentation

Homework

Module 1: Introduction to Data Analytics and R

- Install R and RStudio on your computer.
- Create a new R script and write a simple program to print "Hello, World!" on the console.
- Explore the basic data structures in R (vectors, matrices, data frames) and perform basic operations on them.

Module 2: Data Manipulation with R

- Import a CSV file into R and perform data cleaning tasks, such as removing missing values or outliers.
- Use R functions to calculate summary statistics (mean, median, standard deviation) for a given dataset.
- Practice filtering and sorting data based on specific criteria using R.

Module 3: Exploratory Data Analysis (EDA)

- Choose a dataset of your choice and perform exploratory data analysis using R. Create visualizations (bar plots, scatter plots, histograms) to gain insights into the data.
- Calculate measures of central tendency and dispersion for a specific variable in a dataset.
- Conduct correlation analysis between two variables in a dataset and interpret the results.

Module 4: Statistical Analysis with R

- Choose a dataset and formulate a hypothesis. Perform a suitable hypothesis test in R (e.g., t-test, chi-square test) to validate or reject the hypothesis.
- Fit a linear regression model to a dataset and interpret the coefficients and p-values.
- Explore ANOVA analysis by comparing means across multiple groups in a dataset and conducting post-hoc tests.

Module 5: Data Visualization and Reporting

- Create a complex visualization (e.g., a stacked bar plot, a grouped scatter plot) using R's ggplot2 package.
- Generate an interactive plot or dashboard using R (e.g., using the plotly package).
- Export a plot or report generated in R to a PDF or HTML file.

Module 6: Data Wrangling and Manipulation in R

- Practice data manipulation techniques using the dplyr package, such as selecting specific columns, filtering rows, and summarizing data.
- Reshape a dataset from wide to long format or vice versa using the tidyr package.
- Merge two datasets based on a common variable using R.

Module 7: Introduction to Machine Learning with R

- Apply a machine learning algorithm (e.g., decision tree, logistic regression) to a dataset and evaluate its performance using appropriate metrics (accuracy, precision, recall, etc.).
- Use cross-validation techniques (e.g., k-fold cross-validation) to assess the generalization ability of a machine learning model.
- Experiment with different hyperparameters of a machine learning algorithm and compare the results.

Module 8: Text Mining and Natural Language Processing (NLP) in R

- Preprocess a text dataset by removing stopwords, stemming, and converting text into a matrix or term-document matrix.
- Perform sentiment analysis on a collection of text documents and interpret the sentiment scores.
- Build a text classification model (e.g., Naive Bayes, Support Vector Machines) to classify text documents into different categories.

Module 9: Time Series Analysis with R

- Choose a time series dataset and perform time series decomposition using R to identify trend, seasonality, and residual components.
- Use an appropriate time series forecasting technique (e.g., ARIMA, exponential smoothing) to forecast future values.
- Evaluate the performance of a time series forecasting model using metrics such as mean absolute error (MAE) or root mean squared error (RMSE).

Module 10: Case Studies and Real-World Applications

- Select a real-world dataset from your domain of interest and apply various data analytics techniques learned throughout the course to analyze and visualize the data.
- Present your findings and insights from the case study in a clear and concise manner, either as a report or a presentation.