

R for Data Analysis: From Fundamentals to Advanced Applications

Module 1: R Foundations & Environment

- **Introduction to R and RStudio:** Why R for statistical computing? Strengths in data science, academia, and industry. A deep dive into the RStudio IDE (Console, Editor, Environment, Files, Plots, Help panes).
- **Core Syntax and Objects:** Variables, assignment operators (`<-`, `=`). Understanding R's atomic data types (character, numeric, integer, logical, complex).
- **Workspace and Scripts:** Best practices for project-oriented workflows with RStudio Projects. Managing the working directory (`getwd()`, `setwd()`). Writing, executing, and commenting .R scripts.
- **Package Management:** The role of CRAN. Finding, installing (`install.packages()`), and loading (`library()`) packages. Managing package libraries.

Module 2: Advanced Data Structures

- **Vectors, Matrices, and Arrays:** Deep dive into creating, subsetting (using indices, names, and logicals), and performing vectorized operations. Understanding data type coercion.
- **Lists:** The recursive data structure for holding heterogeneous data. Advanced subsetting (`[]`, `[[[]]`, `$`).
- **Data Frames and Tibbles:** The cornerstone of data analysis. Contrasting traditional `data.frame` with tibbles from the Tidyverse. Advanced inspection (`str()`, `glimpse()`, `summary()`).
- **Factors:** In-depth look at categorical data representation. Reordering levels (`fct_reorder`), recoding (`fct_recode`), and handling factor-based issues in modeling.

Module 3: Advanced Data Transformation with the Tidyverse

- **The Tidyverse Philosophy:** Principles of "tidy data." Mastering the pipe (`%>%`) for readable, sequential operations.
- **Core dplyr Verbs:** Review `select()`, `filter()`, `arrange()`, `mutate()`.
- **Advanced Data Manipulation:**
 - **Grouped Operations:** The power of `group_by()` combined with `summarise()` for aggregate analysis.
 - **Column-wise Operations:** Using `across()` to apply functions to multiple columns efficiently.
 - **Conditional Logic:** Implementing complex if-else logic with `case_when()`.
- **Joining Datasets:** Combining multiple tables with mutating joins (`left_join`, `right_join`, `inner_join`) and filtering joins (`semi_join`, `anti_join`).
- **Reshaping Data with tidyr:** Transforming data from wide to long format (`pivot_longer()`) and long to wide (`pivot_wider()`) for analysis and visualization.

Mini-Project 1: Data Wrangling & Cleaning Challenge

- **Objective:** Apply advanced dplyr and tidyr skills to a complex, messy, multi-table dataset.
- **Task:** You will be given several related CSV files (e.g., sales, customers, products). Your task is to import, clean, join, and reshape the data into a single, tidy analytical base table.
- **Deliverable:** A fully commented R script that documents the entire cleaning and transformation process.

Module 4: The Art of Data Visualization

- **The Grammar of Graphics with ggplot2:** Beyond the basics. A deep dive into mapping aesthetics, using statistical transformations (stat_summary), and adjusting positions (position_dodge).
- **Advanced Customization:**
 - Fine-tuning scales for axes and colors (scale_color_viridis_d, scale_x_continuous).
 - Modifying coordinate systems (coord_flip) and aspect ratios.
 - Creating faceted plots for multi-dimensional analysis (facet_wrap, facet_grid).
 - Building and applying custom themes with theme() for professional polish.
- **Data Storytelling Principles:** Choosing the right chart for your data and message. Principles of visual encoding and reducing chart junk.
- **Interactive Visualizations:** Using plotly to convert static ggplot2 objects into interactive HTML plots. Building basic interactive dashboards to allow for user-driven data exploration.

Mini-Project 2: Exploratory Data Analysis Dashboard

- **Objective:** Create a compelling narrative through a series of linked visualizations to uncover insights in a dataset.
- **Task:** Using a rich dataset, create an R Markdown document that presents an exploratory data analysis. The document must contain at least five distinct, well-labeled visualizations, including at least one interactive plot using plotly.
- **Deliverable:** An HTML report generated from R Markdown that explains your findings, supported by your visualizations.

Module 5: Statistical Modeling: From Linear to Generalized Models

- **Linear Regression (lm()):**
 - Simple, multiple, and polynomial regression.
 - Interaction terms (* vs. :).
 - In-depth model diagnostics: Checking for linearity, normality of residuals, homoscedasticity, and influential points.
- **Generalized Linear Models (glm()):**
 - **Logistic Regression:** Modeling binary outcomes. Interpreting coefficients as log-odds and odds ratios.
 - **Poisson Regression:** Modeling count data. Understanding concepts of exposure and offset.
- **Model Selection and Evaluation:** Comparing models using AIC/BIC, R-squared, and cross-validation techniques.

Module 6: Specialized & Advanced Modeling Techniques

- **For Commerce Students: Time Series Analysis:**
 - Decomposing time series data (trend, seasonality, noise) using STL.
 - Understanding autocorrelation (ACF) and partial autocorrelation (PACF) plots.
 - Building and evaluating **ARIMA** models using the forecast package.
- **For Statistics Students: Resampling & Dimensionality Reduction:**
 - **Resampling:** Cross-validation for model validation and bootstrapping for estimating uncertainty.
 - **Dimensionality Reduction:** Principal Component Analysis (PCA) for reducing feature space.
- **For Mathematics Students: Simulation & Optimization:**
 - **Monte Carlo Simulation:** Writing functions to simulate complex systems and estimate probabilities.
 - **Numerical Optimization:** Using functions like `optim()` to find minima/maxima of mathematical functions.

Module 7: Machine Learning & Real-World Data Analysis

- **Introduction to Machine Learning:** Supervised vs. Unsupervised learning. The train/test split. Introduction to the tidymodels framework for a modern, tidy approach to modeling.
- **Supervised Learning Algorithms:**
 - **Decision Trees & Random Forests:** Building intuitive, powerful predictive models for classification and regression.
 - **Support Vector Machines (SVM):** Understanding the principles of maximal margin classification.
- **Unsupervised Learning Algorithms:**
 - **Clustering:** Using K-Means to discover natural groupings and segments in your data.
- **Feature Engineering & Preprocessing:** The importance of preparing data for ML models. Techniques like scaling, normalization, and creating dummy variables using recipes.

Module 8: Reproducible Research & The Capstone Project

- **Reproducible Reporting with R Markdown:** Combining code, narrative text, and outputs into a single document. Knitting to different formats (HTML, PDF, Word). Best practices for creating dynamic reports.
- **Version Control with Git and GitHub:** Introduction to version control concepts. Using Git within RStudio to track changes and collaborate on projects.
- **Project Formulation:** How to define a clear, answerable research question from a real-world problem and dataset.
- **Project Workshop:** In-class time dedicated to starting the capstone project with instructor guidance.

Capstone Project: End-to-End Data Analysis

- **Objective:** Synthesize all course content by conducting a complete, independent data analysis project from problem formulation to final conclusion.
- **Task:** Choose a large, complex dataset relevant to your field. You will:
 1. Formulate a clear research question.

2. Perform extensive data cleaning, feature engineering, and exploratory analysis.
 3. Select, build, and validate an appropriate statistical or machine learning model (e.g., multiple regression, ARIMA, random forest).
 4. Interpret the model's results and discuss its limitations.
- **Deliverable:** A professional, 10-15 page report written in R Markdown. The report must be fully reproducible, blending narrative text, code, and output (tables, visualizations) to present a compelling analysis. The final session will be dedicated to project presentations.