Introduction to R and Statistics in R

Khoi Trinh

September 20, 2024

Who is Khoi?

- PhD Student (hopefully Candidate soon)
- SecretLab with Dr. Maiti
- Data science in GenAI and Web3
- Backend and API developer at ScooterLab

Introduction to R

- R is a free software environment for statistical computing and graphics.
- Widely used for data analysis, statistics, and machine learning.
- Created by Ross Ihaka and Robert Gentleman in 1993.
- The R community is active and large, with many contributed packages available via CRAN.

Why Use R?

- Specialized for statistical analysis and visualization.
- Highly extensible with thousands of user-contributed packages.
- Integrates well with other programming languages and tools (Python, SQL, etc.).

Setting Up R and RStudio

- Download R from: `https://cran.r-project.org/`
- Download RStudio (IDE): `https://www.rstudio.com/products/rstudio/`
- Remember to install R **BEFORE** RStudio

RStudio Overview

- **Console**: Execute R code interactively.
- **Environment**: View and manage variables and datasets.
- **Source Panel**: Write and edit scripts.
- **Plots**: Visualize your data with charts.

Basic Syntax in R

Variables and Assignment

`<-` is used for assignment:

- `x <- 10` assigns 10 to the variable `x`.
- Variables store values such as numbers, strings, and vectors.

Functions and Arithmetic

- `sum(10, 20)` returns 30.
- `mean(c(1, 2, 3))` calculates the mean of a vector.
- R supports common operations: `+`, `-`, `*`, `/`, `^`.

Working with Data in R

- Data in R is often stored in data frames, a table-like structure with rows and columns.
- Common functions for inspecting data include:
  - `head()`: Shows the first 6 rows.
  - `summary()`: Provides summary statistics for each column.
  - `str()`: Displays the structure of the data.
  - `dim()`: Returns the dimensions of the data (rows, columns).

Demo