UNIT-I: Introduction to Data Science

Variables and Datatypes in R, Data frames Datasets, Recasting and Joining of Data frames,Arithmetic, Logical and Matrix Operations in R, **Advanced Programming in R** : Functions, Data Visualization in R Basic Graphics.

Recap...!!

- Intro R and R Studio.
- History and Intro to R-Programming Language.
- How to download R-base and R-Studio.
- Importing Dataset into R-Studio.
- Some basics operations-GUI.

R - Base & R - Studio

- To download **R**, go to CRAN, the comprehensive R archive network. CRAN is composed of a set of mirror servers distributed around the world and is used to distribute R and R packages.
- https://cloud.r-project.org/
- **RStudio** is an IDE, for R programming. Download and install it from, https://www.rstudio.com/products/rstudio/download/.
- RStudio is updated a couple of times a year. When a new version is available, RStudio will let you know. It’s a good idea to upgrade regularly so you can take advantage of the latest and greatest features.

R Syntax

- There are **two ways** to write code in RStudio:- **first**, in the command prompt and, **second**, in the R script file.
- >print(“Hello, World!”)

  Output:

  [1] "Hello, World!"
- R Programs are usually written in R scripts and then executed in the console window.
- mystr = “Hello, World!”

  print(mystr)

  Output:

  [1] "Hello, World!"

Variables in R Programming

- A variable in programming is used to store some data which will be used by the program. Consider it as a container which holds the data.
- Variables in R programming can be used to store numbers (real and complex), words, matrices, and even tables.
- R is a **dynamically programmed language** which means that unlike other programming languages, we do not have to declare the data type of a variable before we can use it in our program.

Rules to define variables in R

- Variable names cannot contain spaces. **(eg: Bill Amount) - Invalid**
- A variable name should not start with a number. **(eg:- 2total)**- Invalid
- A variable name can contain letters, numbers, underscores and dots.**(eg:- Bill_Name1.) - Valid**
- It should not be a reserved keyword.**(eg:- for, in, repeat, if, NA, etc)**
- A variable name can start with a dot but dot should not follow the number. If starting dot is not followed by a number, then it's valid. **(eg:- .1BillAmount) - Invalid**

Reserved Keywords in R

- Following are the reserved keywords in R,
- Reserved words in R programming are a set of words that have special meaning and cannot be used as an identifier (variable name, function name etc.).

for	In	repeat	while	function
if	else	next	break	TRUE
FALSE	NULL	Inf	NaN	NA
NA_integer_	NA_real_	NA_complex_	NA_character_	

Reserved Keywords in R

- **NA:-** Not Available is used to represent missing values.
- **NULL:-** It represents a missing or an undefined value.
- **NaN:-** It is a short form for Not a Number(eg:- 0/0).
- **TRUE/FALSE: -** These are used to represent Logical values.
- **Inf :-** It denotes Infinity(eg:- 1/0).
- **NA_integer_, NA_real_, NA_complex_, and NA_character_:-** These represent missing values of other atomic types.

Data Types in R

- Data is available in various forms. In programming, data types are associated with a variable.
- A data type describes the type of data a variable can hold. Also, it is important to remember that everything in R is an object.
- The basic data types **(fundamental or atomic data types)** in R are as follows,

  1. Numeric : integer and double (real).
  2. Character.
  3. Logical.
  4. Complex.
  5. Raw.

Data Type	Example	Description
Logical	TRUE, FALSE	boolean values
Numeric	2, 45.9, 3782	Numbers of all kinds
Integer	9L, 779L	Explicitly Integers
Complex	8+9i	Real Value + Complex Value
Character	‘m’, “hello”	Characters and Strings
Raw	[68, 65, 6C, 6C,6F] is the value for string **hello**.	Any data is stored as raw bytes

Note : When data type is Raw, user has to know the format or protocol of the data.

Lets go for understanding datasets Operations.

Purpose:

For our first set of analyses, we'll use a dataset that comes pre-loaded in R. The iris data were collected by botanist **Edgar Anderson** and used in the early statistical work of R.