UNIT-I: Introduction to Data Science

Introduction to R and R Studio, 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...!!

- Key Components of Data Science.
- Jobs in Data Science.
- Who is Data Scientists? Role & Responsibility.
- Intro R Programming Language
- Tools:- R- base & R-Studio

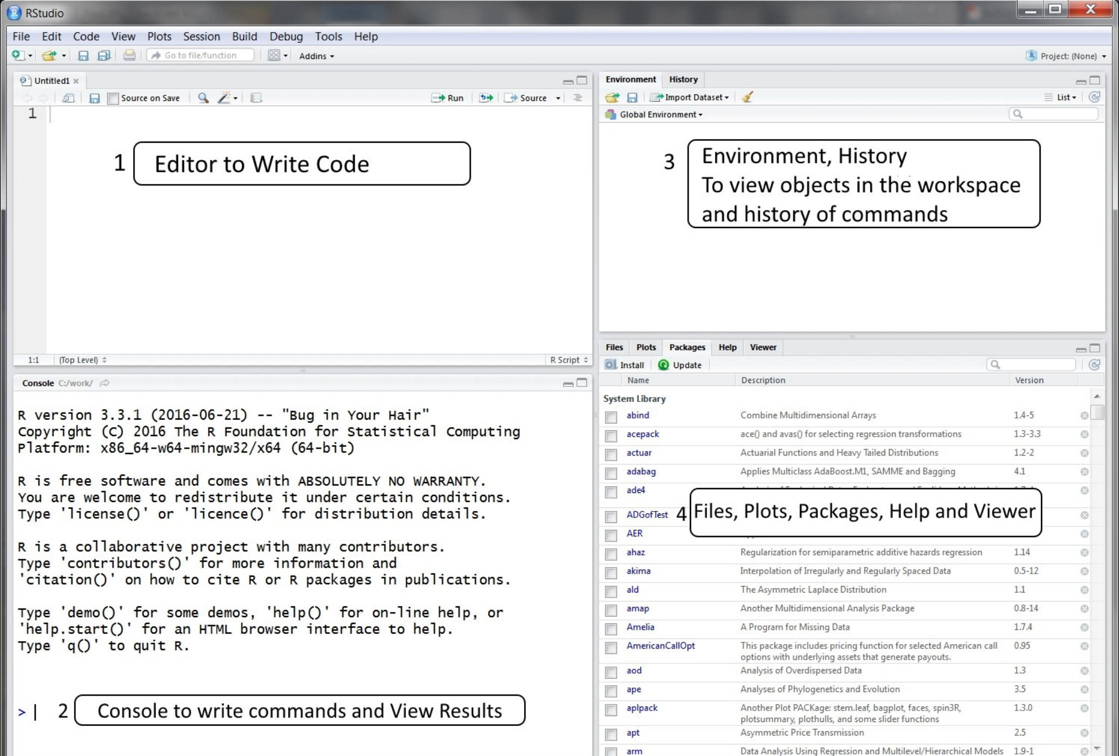
Introduction to R and R Studio

R Programming

- **R** is a language and environment for statistical computing and graphics. It is a **GNU** project which is similar to the **S** language and environment which was developed at Bell Laboratories (AT&T Corp.)
- **R** is an innovative, open-source programming language for machine learning and data science. Used for **statistical analysis** on **datasets**, it's viewed as a different implementation of the **S programming language**.
- R provides an IDE for graphics and statistical computing and has become extremely popular in the past few years. R is one of the top 20 programming languages in the https://www.tiobe.com/tiobe-index/

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.



1 Editor to Write Code

3 Environment, History
To view objects in the workspace
and history of commands

```
R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.


> |
```

2 Console to write commands and View Results

4 Files, Plots, Packages, Help and Viewer

	Name	Description	Version
	System Library		
□	abind	Combine Multidimensional Arrays	1.4-5
□	acepack	ace() and avas() for selecting regression transformations	1.3-3.3
□	actuar	Actuarial Functions and Heavy Tailed Distributions	1.2-2
□	adabag	Applies Multiclass AdaBoost.M1, SAMME and Bagging	4.1
□	ade4		
□	ADGofTest		
□	AER		
□	ahaz	Regularization for semiparametric additive hazards regression	1.14
□	akima	Interpolation of Irregularly and Regularly Spaced Data	0.5-12
□	ald	The Asymmetric Laplace Distribution	1.1
□	amap	Another Multidimensional Analysis Package	0.8-14
□	Amelia	A Program for Missing Data	1.7.4
□	AmericanCallOpt	This package includes pricing function for selected American call options with underlying assets that generate payouts.	0.95
□	aod	Analysis of Overdispersed Data	1.3
□	ape	Analyses of Phylogenetics and Evolution	3.5
□	aplpack	Another Plot PACKage: stem.leaf, bagplot, faces, spin3R, plotsummary, plothulls, and some slider functions	1.3.0
□	apt	Asymmetric Price Transmission	2.5
□	arm	Data Analysis Using Regression and Multilevel/Hierarchical Models	1.9-1

Lets go for understanding datasets.

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.

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.