R for Data Scientists

Dr Sainath S Pawaskar – spawaskar@gmail.com

Any Programming Language

- Contains four main parts
  - Part that does something which we (mostly) intend to do
    - Calculate area of triangle
  - Decision making
  - Loops
  - Input and output
- Throws three types of errors
  - Syntax
  - Run-time (exceptions)
  - Semantic

R Programming Language

- Statistical analysis and data mining
- Graphics representation
- Reporting
- Developed by
  - **R**oss Ihaka
  - **R**obert Gentleman
  - University of Auckland, New Zealand
  - Conceived in 1992
  - Released in 1995
  - Inspired from S programming language of Bell Labs
  - R 1.0.0 released in February 2000

Why R

- Large, coherent and integrated collection of tools for data analysis.
  - Large number of statistical packages
- Graphical facilities for data analysis and display
  - Better visualisation
- Preferred by data scientists along with Python
- Supported by talented contributors
- Used in universities as well as in business critical setup

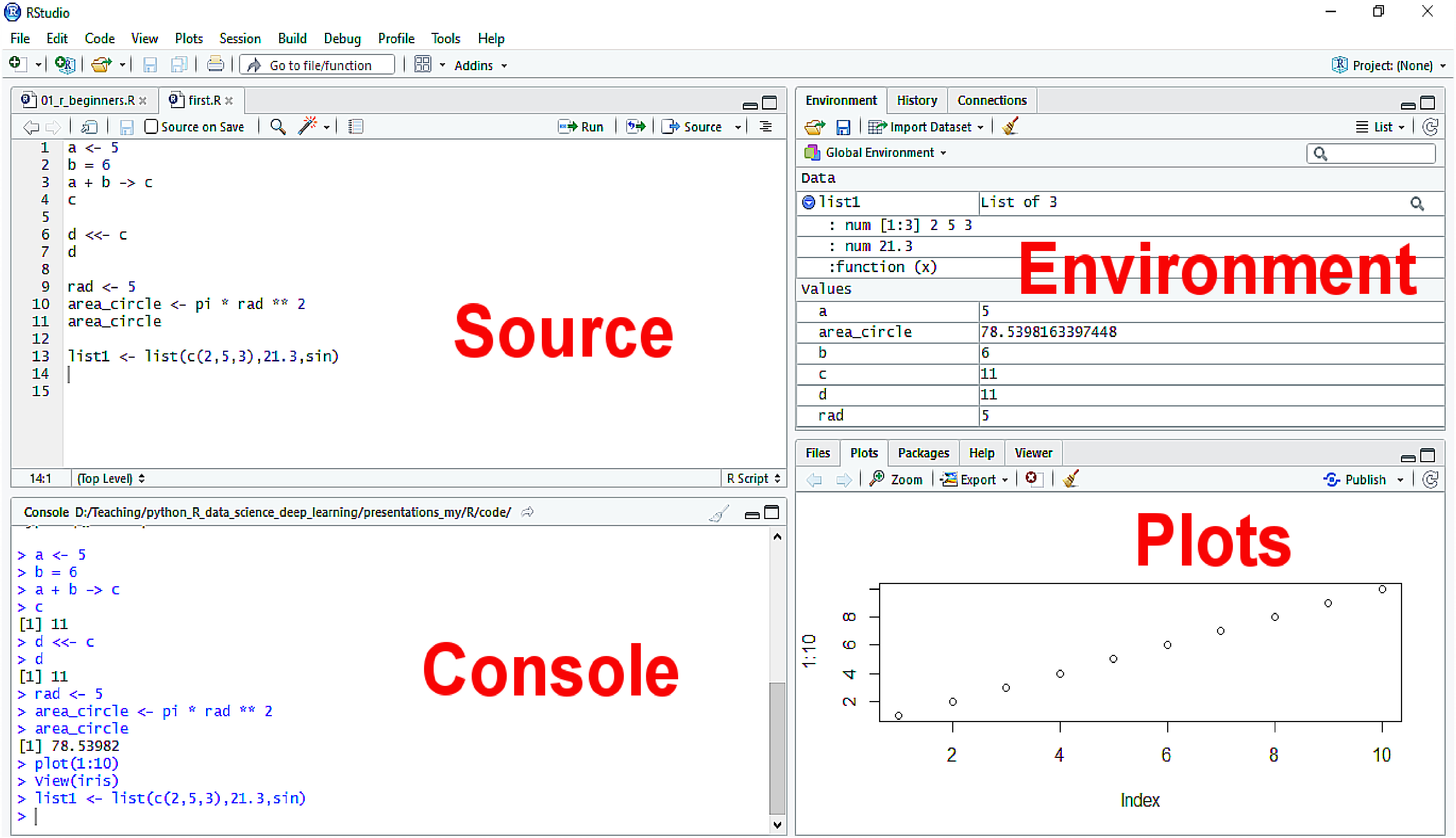
How to Execute R Programmes

- Interpreted language
  - Not compiled into object file as in C/C++
- R Console
- R script file
- Ctrl – Enter
- Rscript filename.R
  - Runs script at linux/windows command prompt

Comments

- Help programmers and others to understand the logic
- Help in debugging and troubleshooting
- Give information about programs and functions
- Will help if you come back to the script file after a long time
- Single line comment
  - Written with # at the beginning
- Multi-line comments are now supported
  - Should be enclosed with either single or double quote

R Studio



```r
a <- 5
b = 6
a + b -> c
c

d <<- c
d

rad <- 5
area_circle <- pi * rad ** 2
area_circle

list1 <- list(c(2,5,3),21.3,sin)
```

```
> a <- 5
> b = 6
> a + b -> c
> c
[1] 11
> d <<- c
> d
[1] 11
> rad <- 5
> area_circle <- pi * rad ** 2
> area_circle
[1] 78.53982
> plot(1:10)
> View(iris)
> list1 <- list(c(2,5,3),21.3,sin)
>
```

Data

list1	List of 3
: num [1:3] 2 5 3	
: num 21.3	
:function (x)	

Values

a	5
area_circle	78.5398163397448
b	6
c	11
d	11
rad	5

Basics

- Variable: named storage
- Assignment: <-, =, <<-, ->, ->>
- Operation
  - +, *, ** or ^
- Print results
  - Values can be printed/shown
    - Writing variable name
    - Print()
      - paste() or paste0() for concatenation
    - cat()
      - Concatenates multiple items

Basics (Cont ...)

-
- User input: readline()
- Read csv file
  - read.csv()
- Explore data
  - Head() and str()
- Plots
- Vectors with mixed classes
  - Classes will get typecast (coercion)