Statistical Data Analysis and Visualization With R Programming: From Basics to Advanced Analytics

Resource Person

Sujit Kumar Roy

MSc. in Water resources Development

Bangladesh University of Engineering and Technology (**BUET**)

Phone: +8801748985407

Email: sujitroy.bejoy@gmail.com

Organizer

Advances in Geographical Research

Website: https://www.aigr.co.in/

WhatsApp: +91-94751 72399

aigeo.research@gmail.com

ADVANCES IN GEOGRAPHICAL RESEARCH

Making the Complex Simple

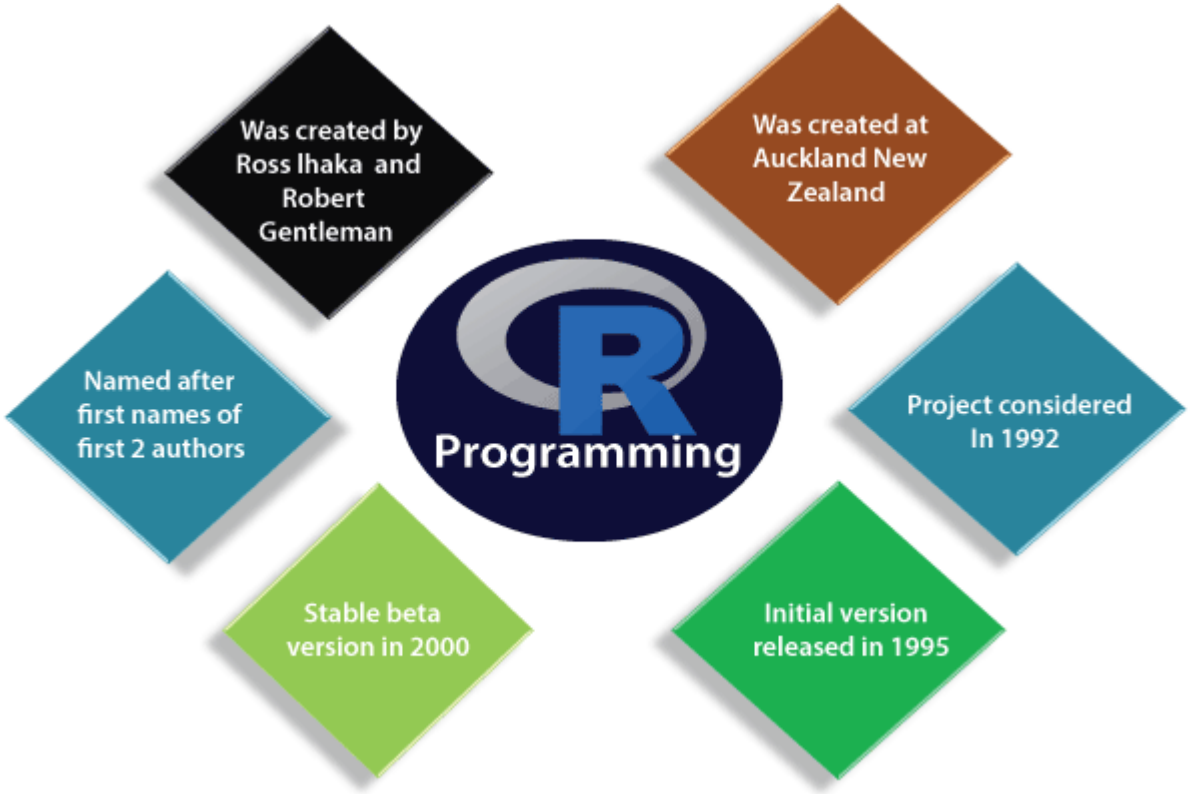
What is R Programming

- R programming is used as a leading tool for machine learning, statistics, and data analysis. Objects, functions, and packages can easily be created by R.
- "R is an interpreted computer programming language which was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand."
- The R Development Core Team currently develops R. It is also a software environment used to analyze statistical information, graphical representation, reporting, and data modeling.
- R allows integration with the procedures written in the C, C++, .Net, Python, and FORTRAN languages to improve efficiency.
- It’s a platform-independent language. This means it can be applied to all operating systems.
- It’s an open-source free language. That means anyone can install it in any organization without purchasing a license.

History of R Programming

Origins

- R goes back around 20-30 years ago
- Developed by Ross Ihaka and Robert Gentleman
- At the University of Auckland, New Zealand
- Name derived from the initials of the developers



Milestones

- First project considered in 1992
- Initial version released in 1995
- Stable beta version released in 2000
- Further development by R Development Core Team

Why Use R?

- **Statistical Analysis:** R is designed for analysis and It provides an extensive collection of graphical and statistical techniques, By making a preferred choice for statisticians and data analysts.
- **Open Source:** R is an open – source software, which means it is freely available to anyone. It can be accessble by a vibrant community of users and developers.
- **Data Visulaization :** R boasts an array of libraries like ggplot2 that enable the creation of high-quality, customizable data visualizations.
- **Data Manipulation :** R offers tools that are for data manipulation and transformation. For example: IT simplifies the process of filtering , summarizing and transforming data.
- **Integration :** R can be easily integrate with other programming languages and data sources. IT has connectors to various databases and can be used in conjunction with python, SQL and other tools.
- **Community and Packages:** R has vast ecosystem of packages that extend its functionality. There are packages that can help you accomplish needs of analytics.

R Advantages

1) Open Source: An open-source language is a language on which we can work without any need for a license or a fee. R is an open-source language. We can contribute to the development of R by optimizing our packages, developing new ones, and resolving issues.

2) Platform Independent: R is a platform-independent language or cross-platform programming language which means its code can run on all operating systems. R enables programmers to develop software for several competing platforms by writing a program only once. R can run quite easily on Windows, Linux, and Mac.

3) Machine Learning Operations: R allows us to do various machine learning operations such as classification and regression. For this purpose, R provides various packages and features for developing the artificial neural network. R is used by the best data scientists in the world.

4) Exemplary support for data wrangling: R allows us to perform data wrangling. R provides packages such as dplyr, readr which are capable of transforming messy data into a structured form.

5) Quality plotting and graphing: R simplifies quality plotting and graphing. R libraries such as ggplot2 and plotly advocates for visually appealing and aesthetic graphs which set R apart from other programming languages.

6) The array of packages: R has a rich set of packages. R has over 10,000 packages in the CRAN repository which are constantly growing. R provides packages for data science and machine learning operations.

7) Statistics: R is mainly known as the language of statistics. It is the main reason why R is predominant than other programming languages for the development of statistical tools.

8) Continuously Growing: R is a constantly evolving programming language. Constantly evolving means when something evolves, it changes or develops over time, like our taste in music and clothes, which evolve as we get older. R is a state of the art which provides updates whenever any new feature is added.

R Disadvantages

1) Data Handling: In R, objects are stored in physical memory. It is in contrast with other programming languages like Python. R utilizes more memory as compared to Python. It requires the entire data in one single place which is in the memory. It is not an ideal option when we deal with Big Data.

2) Basic Security: R lacks basic security. It is an essential part of most programming languages such as Python. Because of this, there are many restrictions with R as it cannot be embedded in a web-application.

3) Complicated Language: R is a very complicated language, and it has a steep learning curve. The people who don't have prior knowledge or programming experience may find it difficult to learn R.

4) Weak Origin: The main disadvantage of R is, it does not have support for dynamic or 3D graphics. The reason behind this is its origin. It shares its origin with a much older programming language "S."

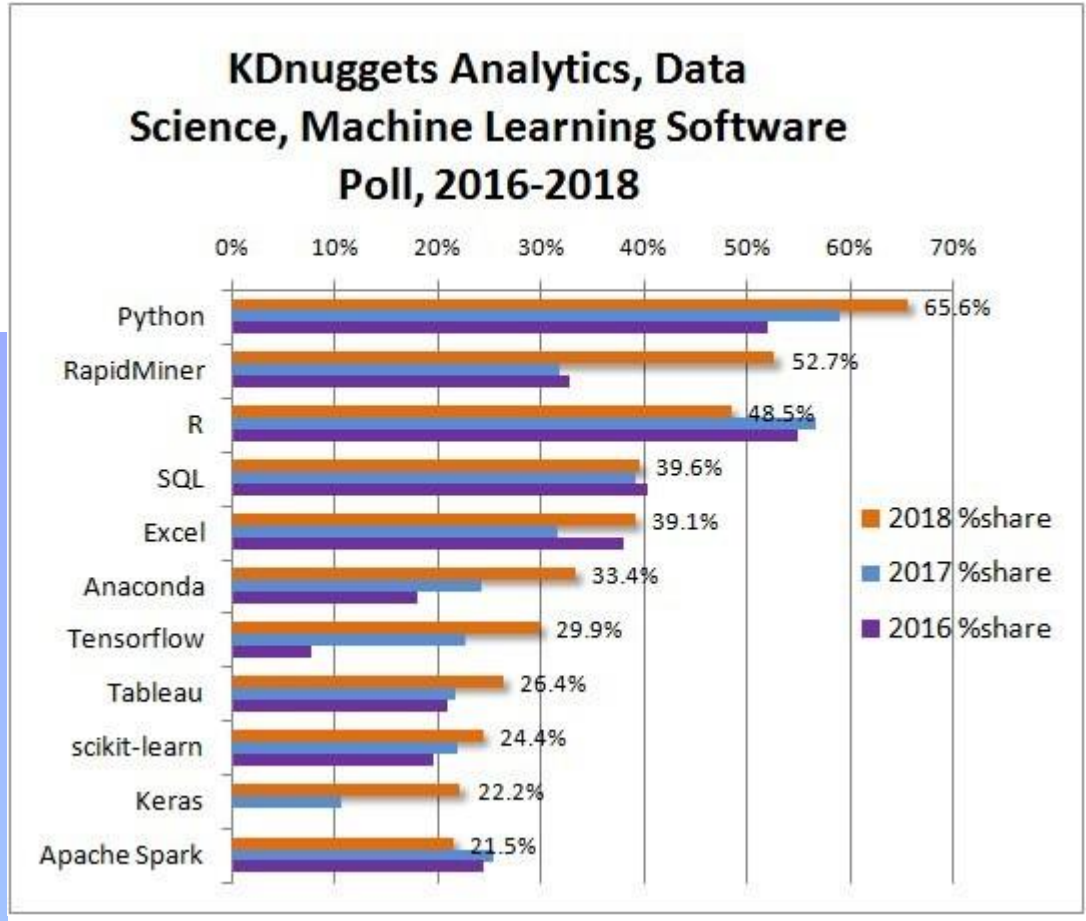
5) Lesser Speed: R programming language is much slower than other programming languages such as MATLAB and Python. In comparison to other programming language, R packages are much slower.

R VS Python

Parameter	R	Python
Objective	Data Analysis and Statistical Modeling	Data Science, Web Development, Embedded Systems
Workability	Consists of many easy to use packages	Can easily perform matrix computation as well as optimization
Integration	Locally Run Programs	Programs integrated with web-app for easy deployment
Database Handling Capacity	Poses problem for handling large dataset	Can handle large data easily without any fault
IDE	Rstudio, R GUI	Spyder, IPython, Jupyter Notebook
Essential Packages and library	ggplot2, tidyverse, caret	Numpy, pandas, scipy, scikit-learn, TensorFlow

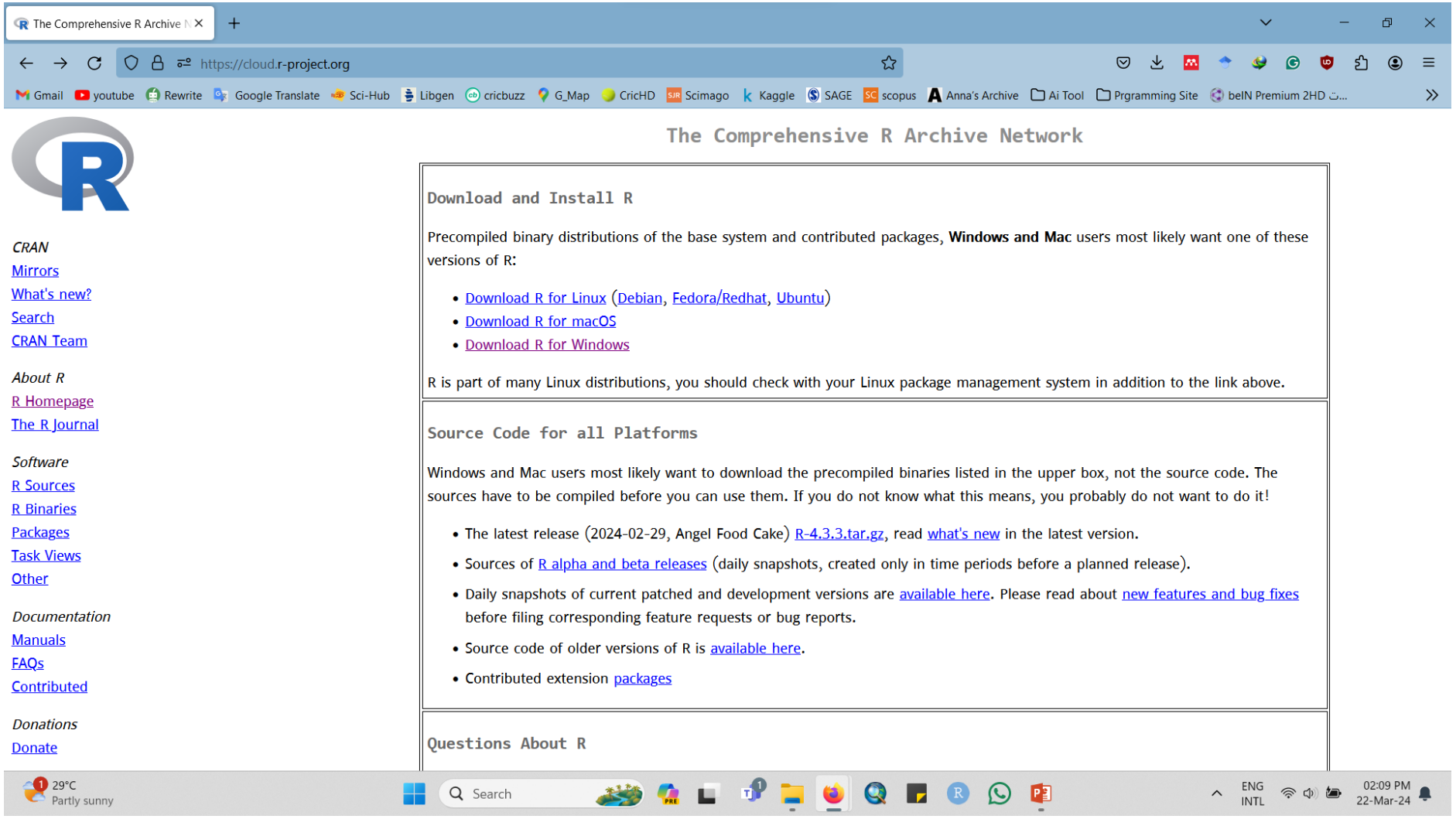
Data Flair

Comparison between R Programming and Python



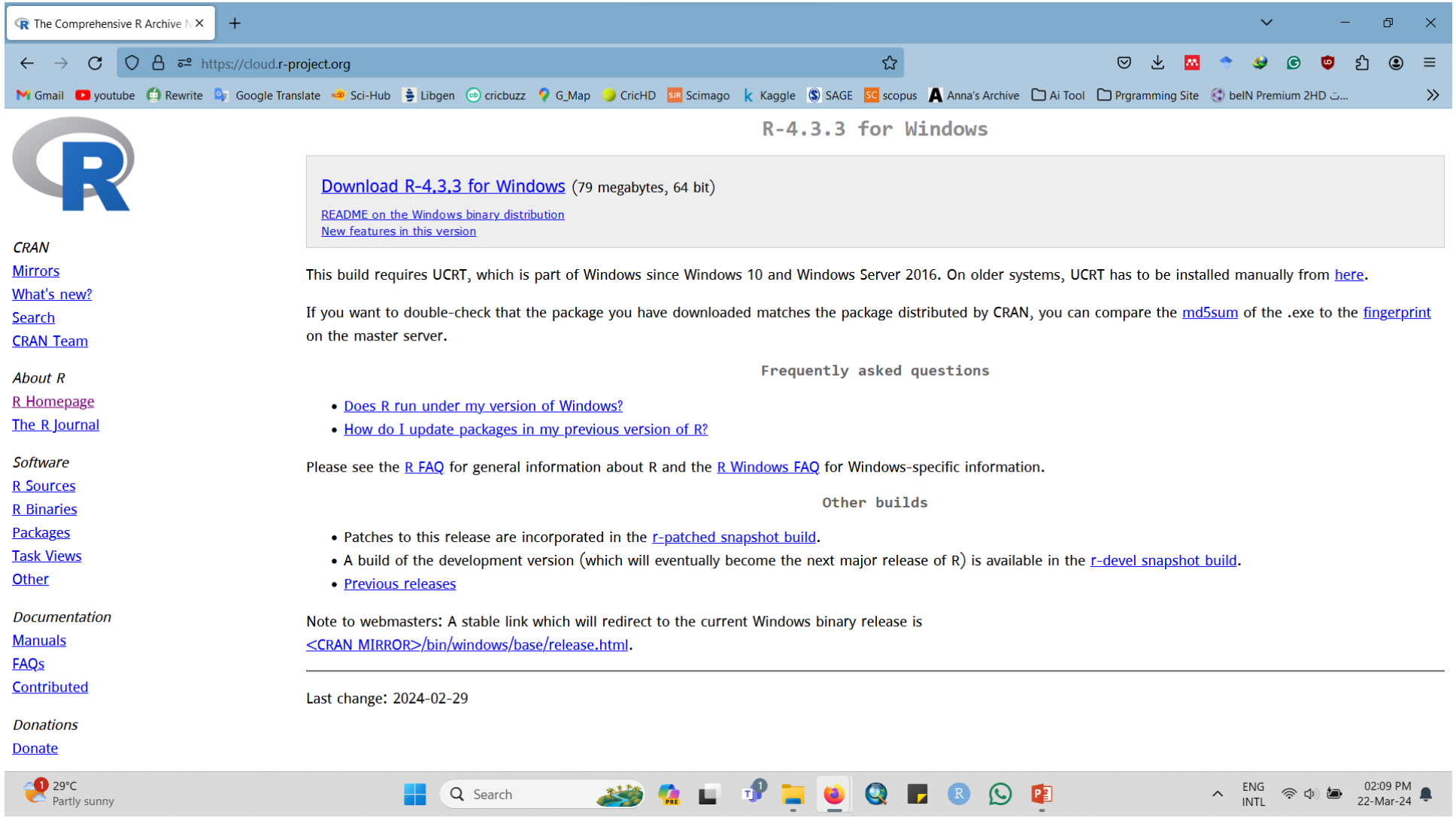
Installation of R

- To install R, either we can get it from the site https://cloud.r-project.org or can use commands from the terminal.
- **Step 1:** First, we have to download the R setup from https://cloud.r-project.org/bin/windows/base/.

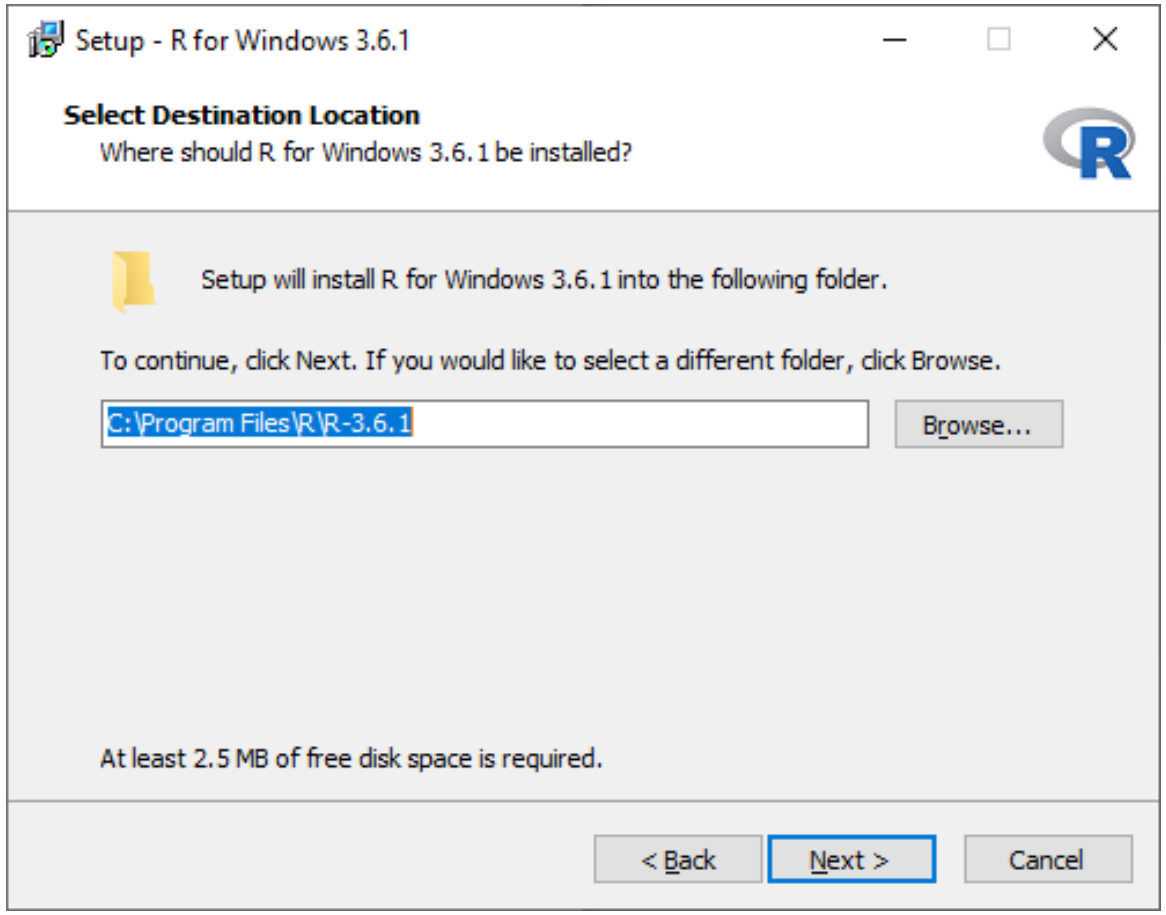


Step 2:

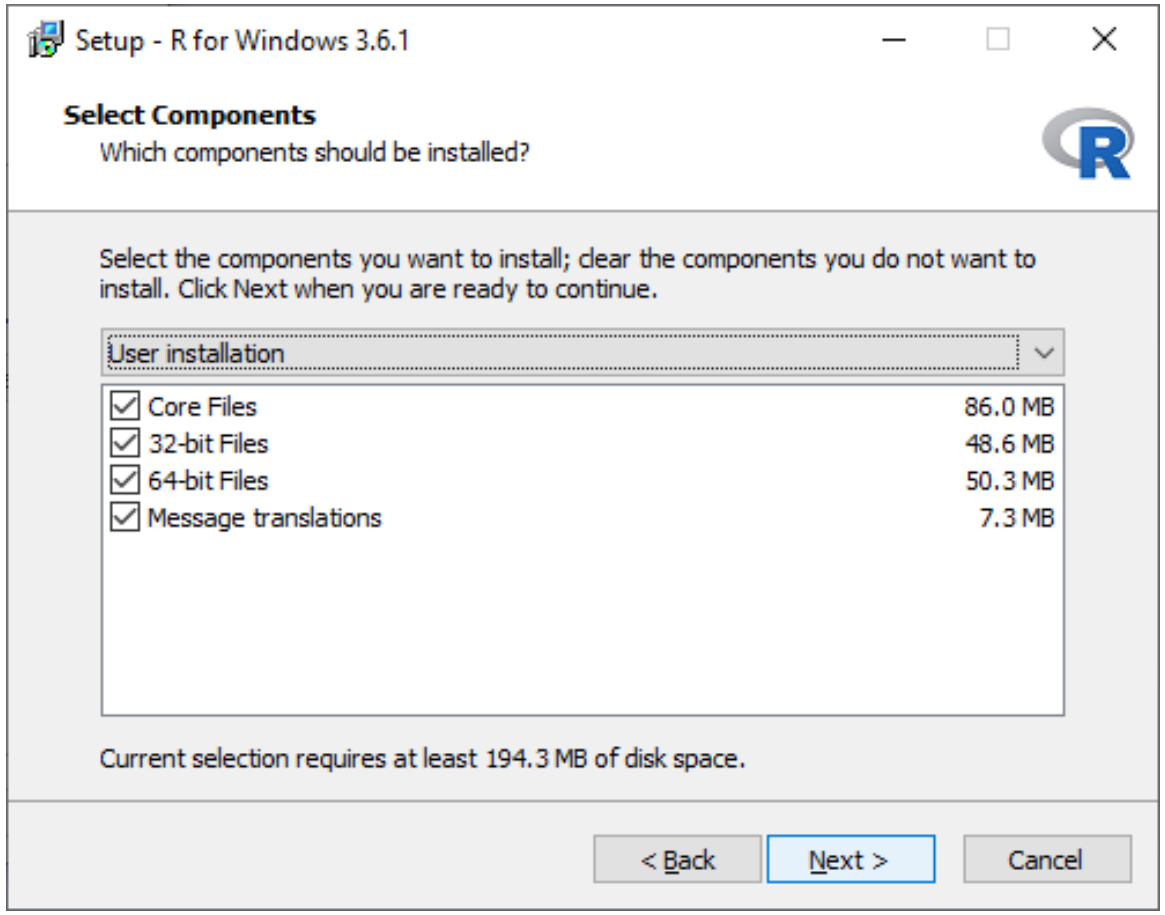
When we click on **Download R 4.3.3 for windows**, our downloading will be started of R setup. Once the downloading is finished, we have to run the setup of R in the following way: (Source: https://cloud.r-project.org/bin/windows/base/)

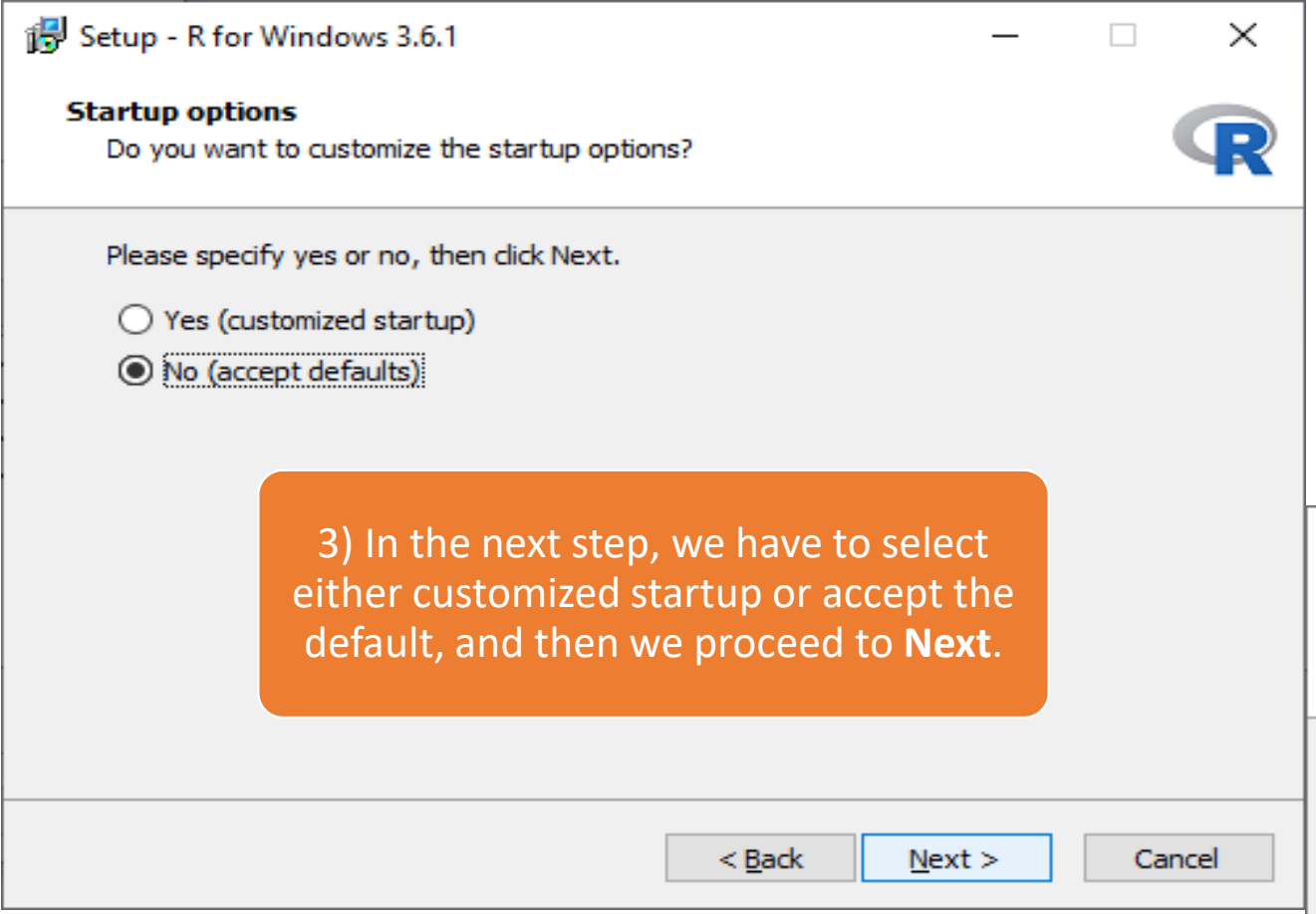


1) Select the path where we want to download the R and proceed to Next.

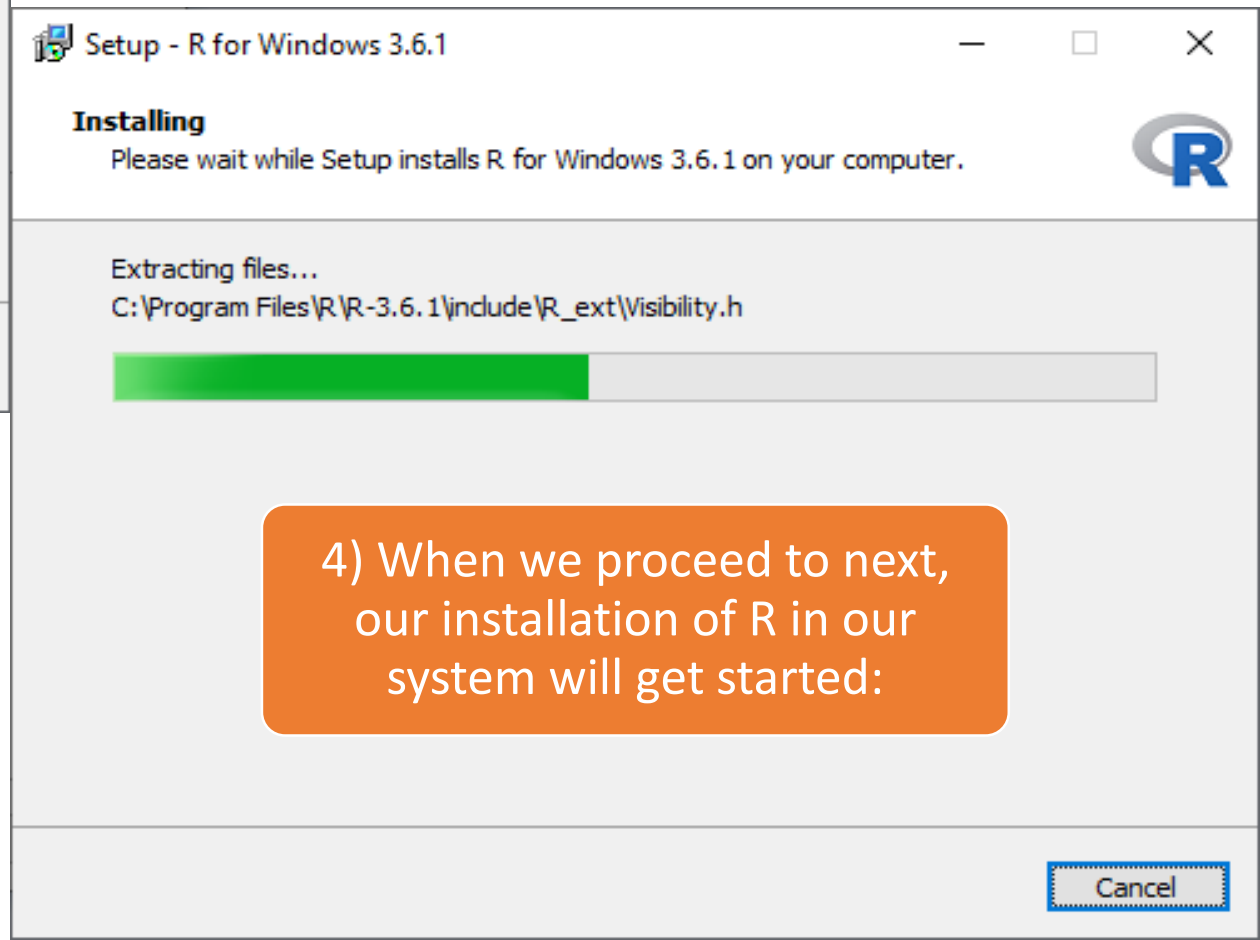


2) Select all components which we want to install, and then we will proceed to **Next**.

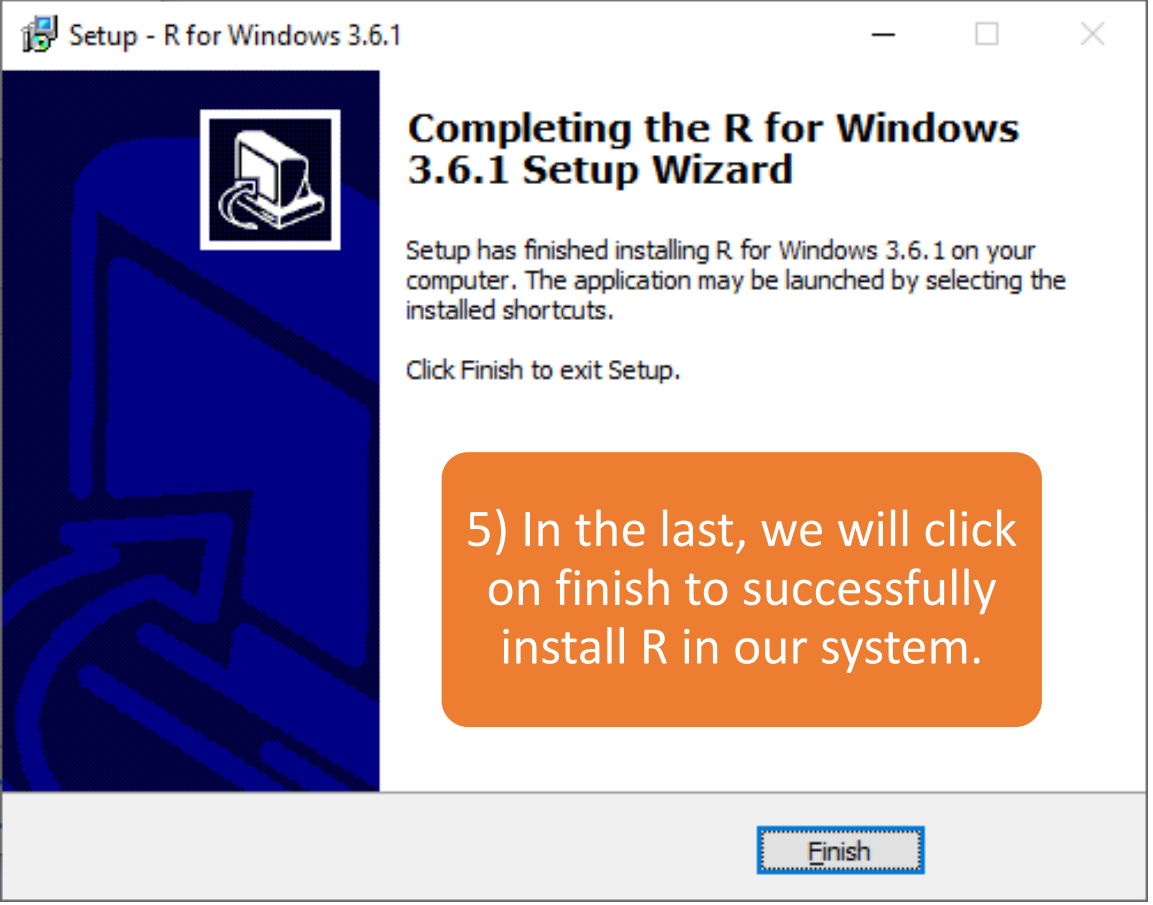




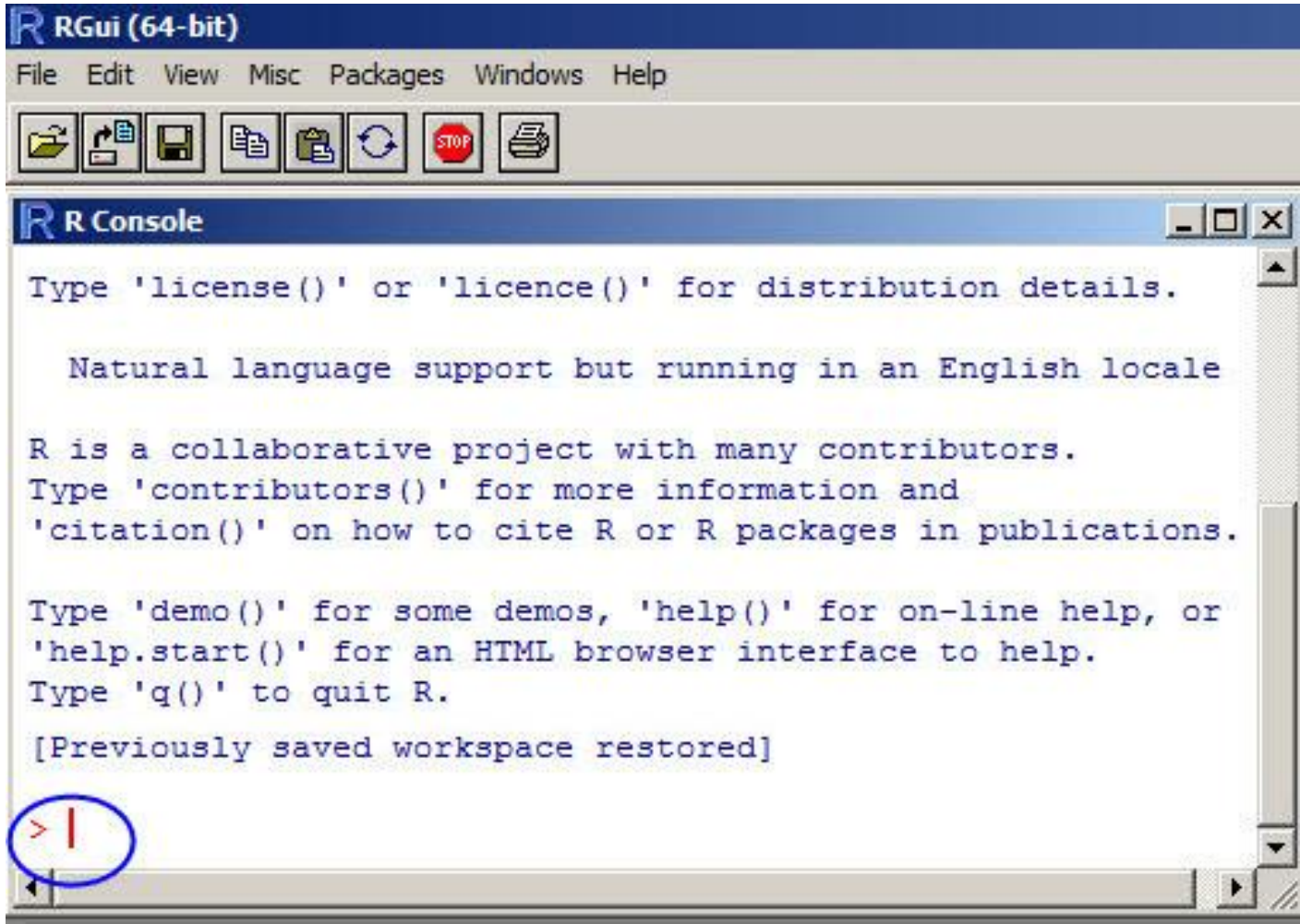
3) In the next step, we have to select either customized startup or accept the default, and then we proceed to **Next**.



4) When we proceed to next, our installation of R in our system will get started:



5) In the last, we will click on finish to successfully install R in our system.



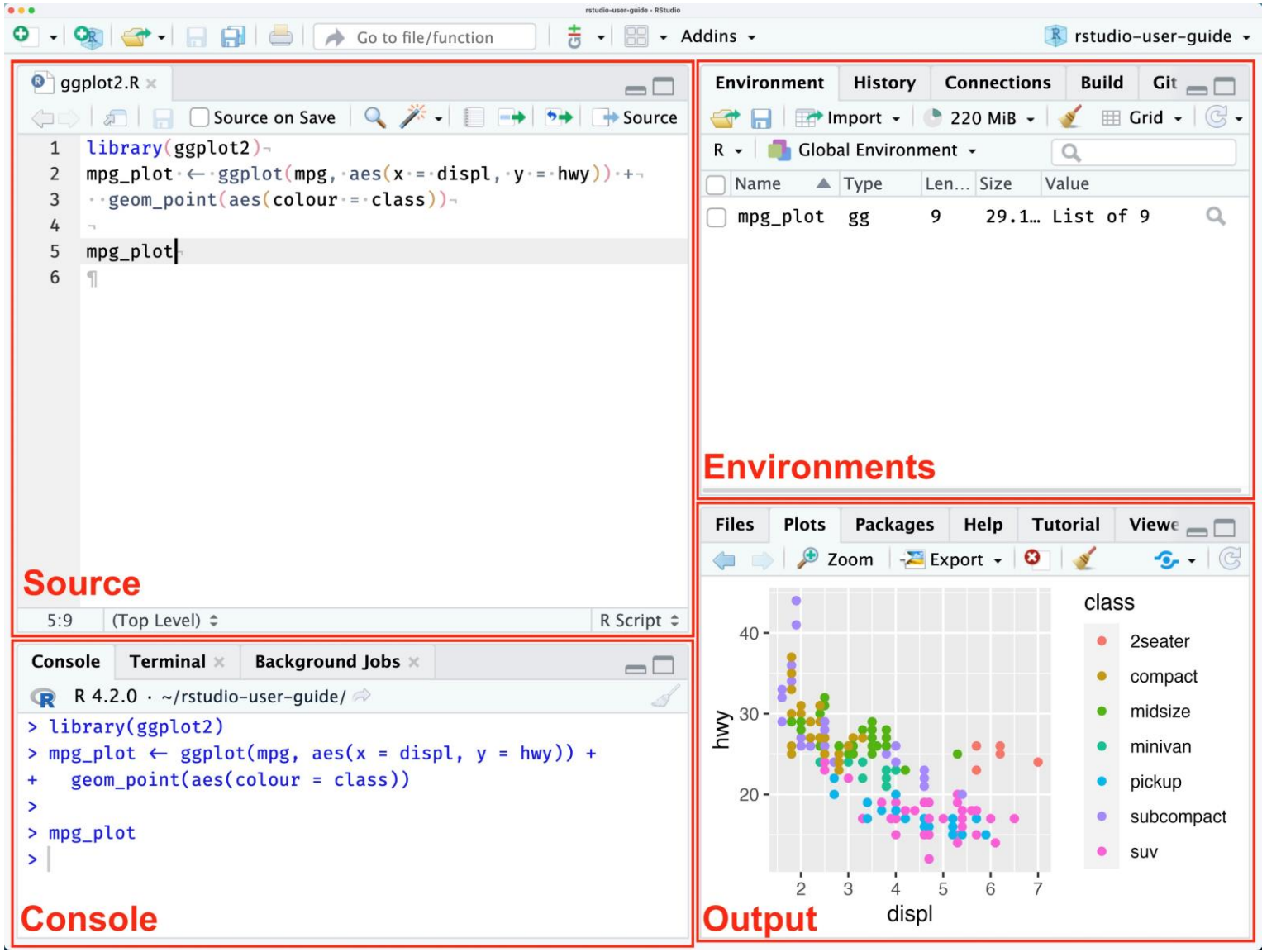
RStudio IDE

Overview

- RStudio is an integrated development environment (IDE)
- Allows for more user-friendly interaction with R
- Similar to standard RGui but more intuitive

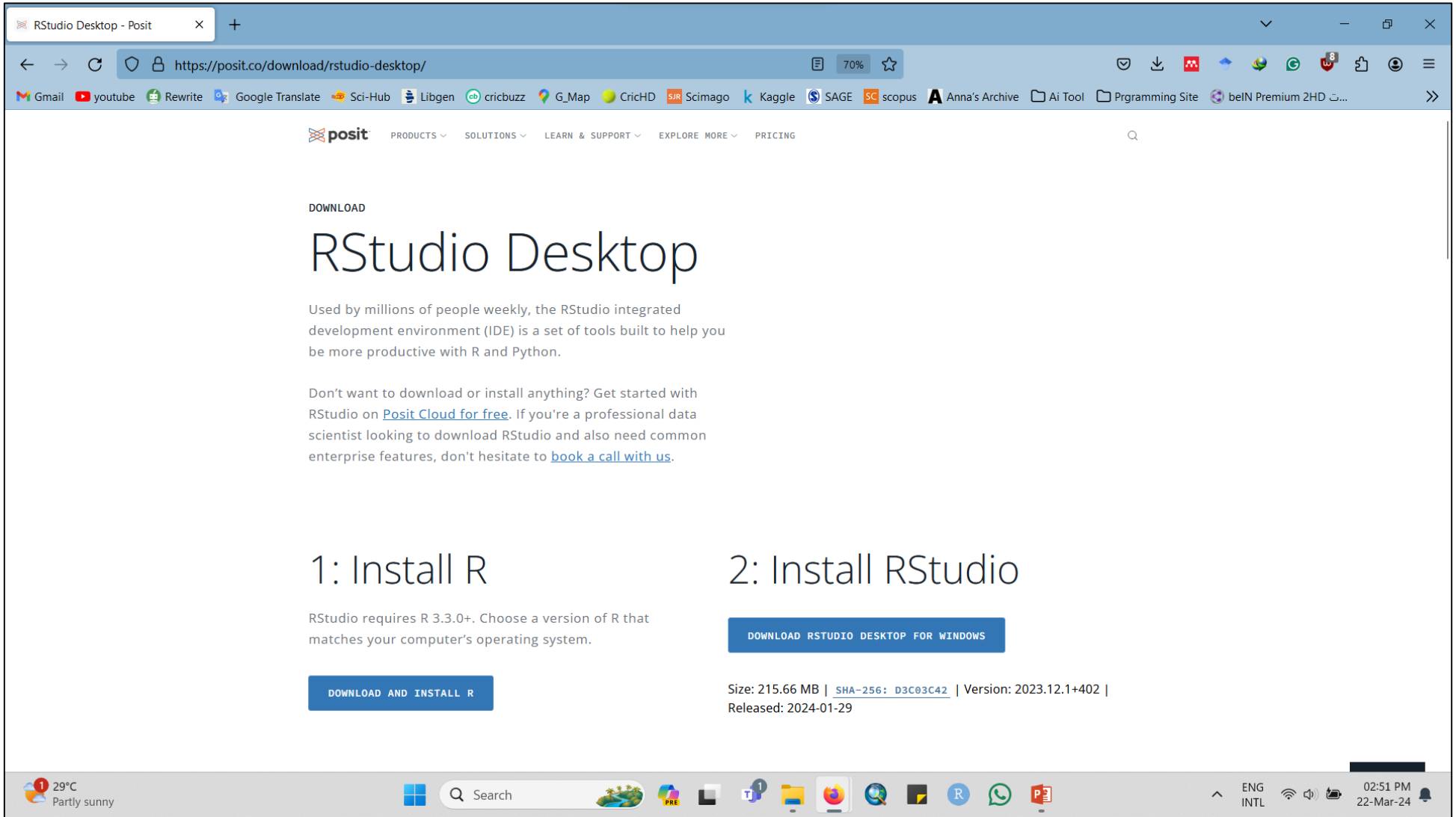
Features

- Drop-down menus
- Multiple tabbed windows
- Extensive customization options



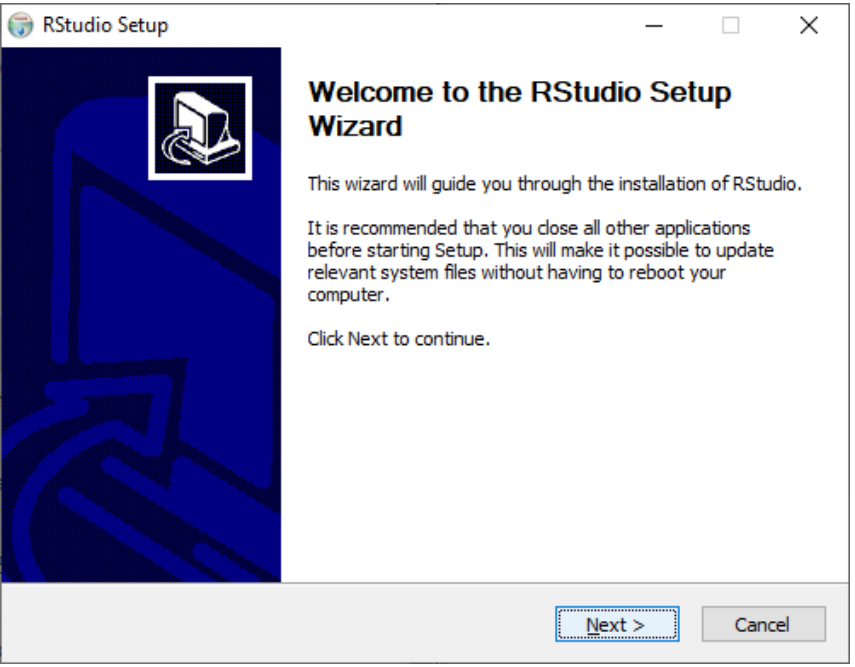
Installation of RStudio

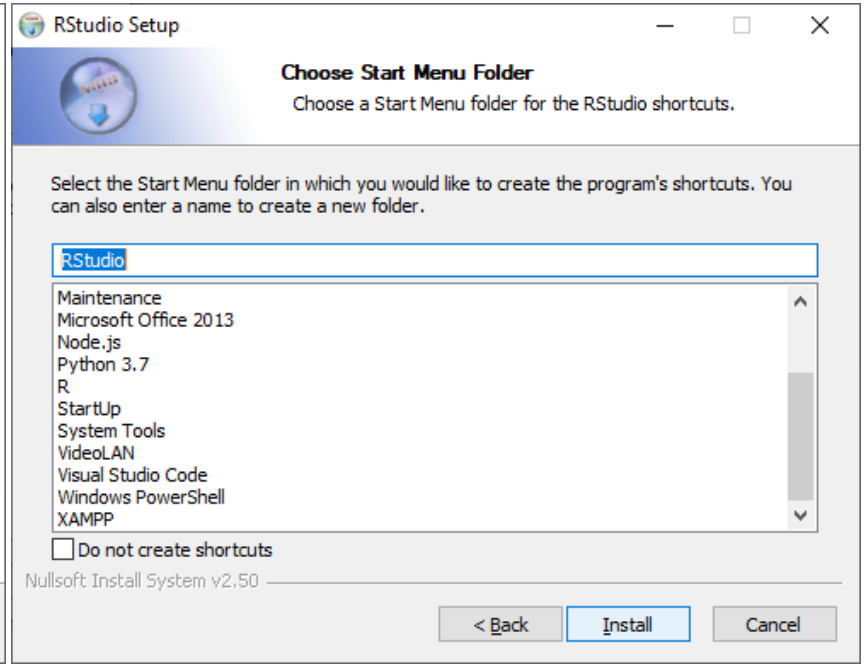
- Step 1: In the first step, we visit the RStudio official site and click on Download Rstudio (https://posit.co/download/rstudio-desktop/)

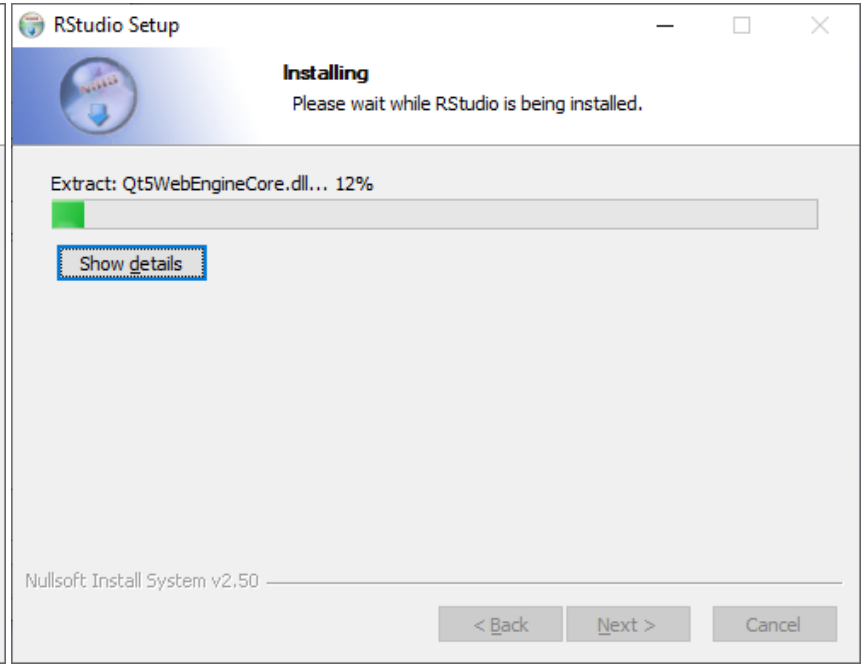


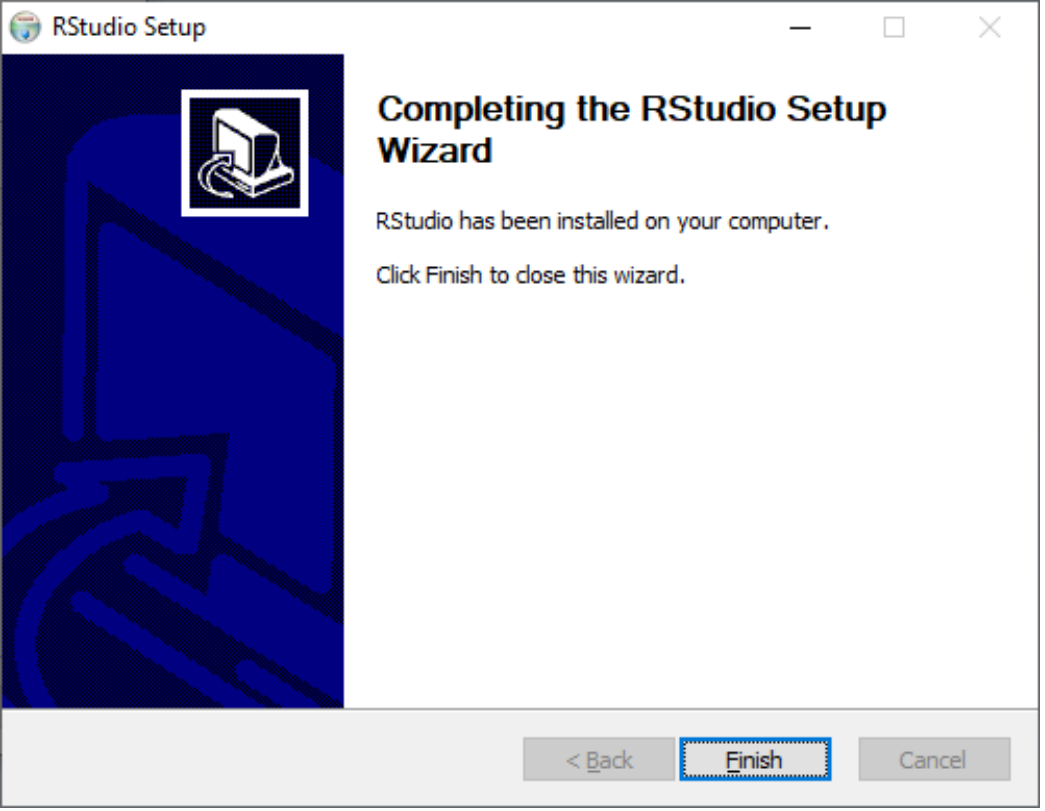
Step 2: In the next step, we will select the RStudio desktop for open-source license and click on download.

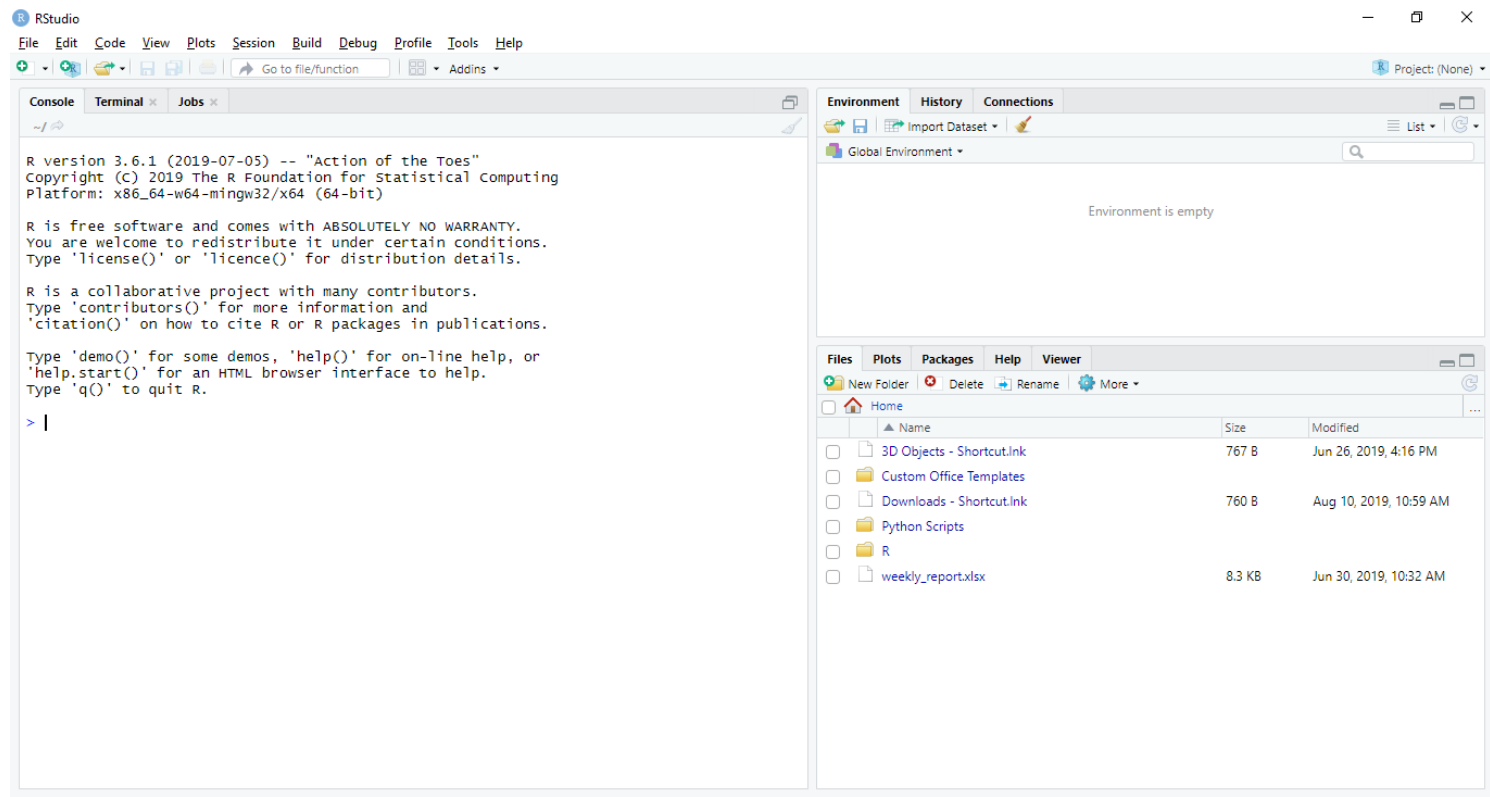
Step 3: In the next step, we will run our setup in the following way: 1) Click on Next. 2) Click on Install. 3) Click on finish.








RStudio Setup
Completing the RStudio Setup Wizard
RStudio has been installed on your computer.
Click Finish to close this wizard.
< Back
Finish
Cancel


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Console Terminal Jobs
R version 3.6.1 (2019-07-05) -- "Action of the Toes"
Copyright (C) 2019 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.
>
Environment History Connections
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
New Folder Delete Rename More
Home
Name Size Modified
3D Objects - Shortcut.lnk 767 B Jun 26, 2019, 4:16 PM
Custom Office Templates
Downloads - Shortcut.lnk 760 B Aug 10, 2019, 10:59 AM
Python Scripts
R
weekly_report.xlsx 8.3 KB Jun 30, 2019, 10:32 AM

R - Basic Syntax

- Once you have R environment setup, then it’s easy to start your R command prompt by just typing the following command at your command prompt −
- $ R
- This will launch R interpreter and you will get a prompt > where you can start typing your program as follows −
- > myString <- "Hello, World!"
- > print (myString)
- [1] "Hello, World!"

Comments in R

Comments are generally used for the following purposes:

- Code Readability
- Explanation of the code or Metadata of the project
- Prevent execution of code
- To include resources

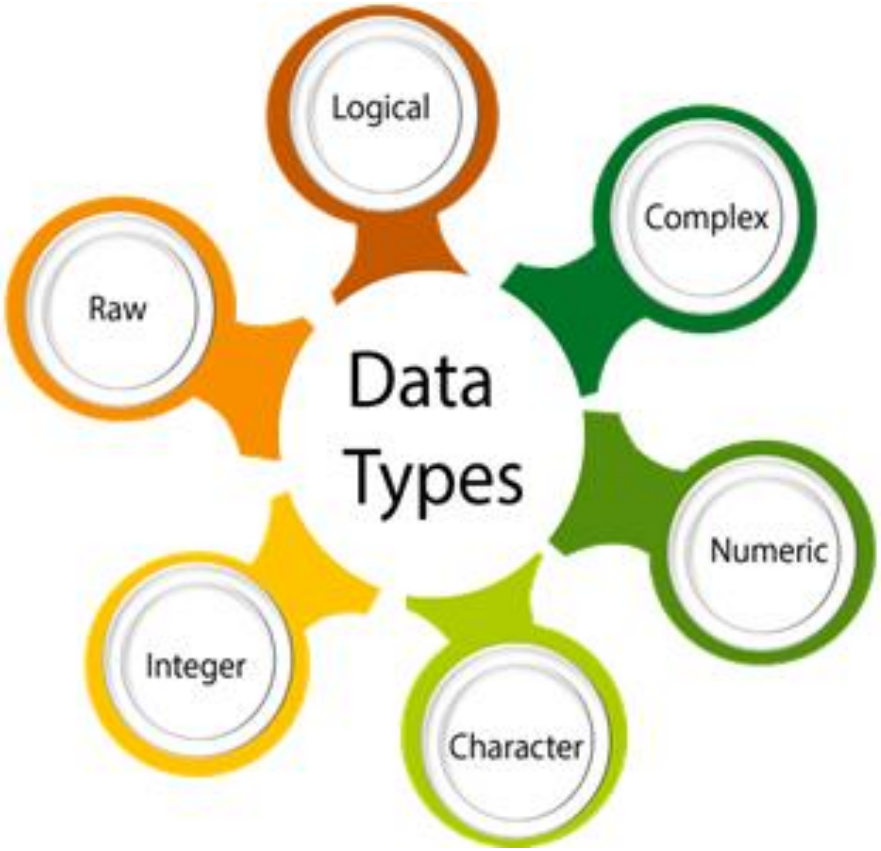
Types of Comments

- There are generally three types of comments supported by languages, namely-
- **Single-line Comments-** Comment that only needs one line
- **Multi-line Comments-** Comment that requires more than one line.

Data Types in R Programming Language

Data Types in R are:

- numeric – (3,6.7,121)
- Integer – (2L, 42L; where ‘L’ declares this as an integer)
- logical – (‘True’)
- complex – (7 + 5i; where ‘i’ is imaginary number)
- character – (“a”, “B”, “c is third”, “69”)
- raw – (as.raw(55); raw creates a raw vector of the specified length)



Data Types in R Programming Language

Character: The character data type includes text. Combinations of letters or numbers can become the character data type if quotes wrap around the text string.

Numeric: As the default data type in R, numeric data comes in decimal form. For example, R stores the number 13.2 as numeric.

Integer: All integers represent numbers created through the integer function in R and do not contain decimals. You must use the specific integer function as.integer to create an integer. For example, an input of as.integer(13.2) returns just 13 when printed.

Logical: The logical data type appears when you compare multiple variables. The variable returned in these cases is either True or False. If you write 2 > 3, R returns False. This response from R is Boolean and represents the logical data type.

Complex: The complex data type combines real and imaginary parts. An easy way to think of this is a mathematical equation with an undefined variable. R utilizes i to represent the imaginary or undefined portion. The data type returned is complex if you entered a = 12 - 4i.

Thank You

We Hope You Enjoy The Presentation