R Programming In Statistics

Prof Dr Balasubramanian Thiagarajan

ISBN: 978-93-5737-722-5

Preface

Every professional needs to perform statistical analysis in some form of the other. In order to perform this task various software tools are available. Majority of them are paid software. R programming which is an open source tool can be used to perform statistical analysis. Since it is an open source tool many front end GUI's are available to make the job easier for the user. In this book the most popular GUI RStudio is used. RStudio is a most powerful GUI front end for R programming which has been designed to use all the features of this language with ease. This book has been authored with a novice user in mind. Various steps in statistical analysis have been explained in detail using a large number of screenshots. Codes used have been clearly illustrated. The book has been structured in such a manner to ensure that basic concepts have been clearly explained with the help of screenshots before taking on challenging analytical problems.

Towards the end of the book the reader is provided with an additional resource which gives out all the codes used in this book as well as those additional ones that have not found their place in the book. Learning R coding is not difficult provided the reader spends time practicing the same. The reader is encouraged to execute all the codes provided in the R_code manual which has been provided at the end of the book. R programming can be compared to that of SPSS (the popular statistical analytical tool) as far as its ability to perform statistical analysis. One tip the author wishes to provide to the reader who is attempting to make data entry within the RStudio environment. It is always better to import data into RStudio for performing data analysis. Data can be imported from Excel , google spread sheets etc.

The reader is encouraged to download the install the software and libraries that have been described in the book and to try them out.
Advantages of R Programming :

1. It is a powerful statistical tool

2. It is open source and hence it is free

3. It is an excellent tool that can be used to perform visual analysis of a dataset. It can created different types of charts and graphs, thereby facilitating accurate analysis of data.

Being the first edition author invites comments from the readers. The same be mailed to:

Email

About the Author

Prof Dr Balasubramanian Thiagarajan is a practicing otolaryngologist
Former professor and head Department of otolaryngology
Government Stanley Medical College
Chennai
Former Registrar
The Tamilnadu Dr MGR Medical Univeristy
Guindy
Chennai.

Currently

Dean

Sri Lalithambigai Medical College
Maduravoil
Chennai

Contents

Introduction

R is a language and environment for statistical and graphics. This GNU project is similar to the "S" language and environment that was developed by Bell laboratories. Even though R can be considered as a different implementation of S, there are some important differences. Most of the code written for S runs unaltered under R.
In 1992, Ross Ihaka and Robert Gentleman created R at the University of Aukland. This was to enable the students to use this as a statistical tool. Initial version was released in 1995. Currently it is being maintained by the R Development Core Team.

R provides a variety of statistical (linear and non-linear modelling, classical statistical tests, time series analysis, classification, clustering etc). It also provides graphical techniques and is highly extensible.

One major strength of R is the ease with which well-designed publication quality plots can be produced, including mathematical symbols and formulae when needed.

1. It is a free and open source tool.

2. It has a large community of users

3. It is an independent platform and can be run without a compiler.

4. Can be considered to be the Gateway for lucrative career

5. Has a robust visualization library - R comprises libraries like ggplot2, plotly that offer aesthetic graphical plots to its users. R is recognized for its stunning visualizations which gives it an edge over Data science programming languages.

6. Used in almost every Industry

7. Distributed computing - In distributive computing, tasks are spit between multiple processing nodes to reduce processing time and to increase efficiency. R has packages lid ddr and multiDplyr that enable it to use distributed computing to process large data sets.

8. Iterfacing with Databases - R contains several packages that enable it to interact with databases like ROracle, Open database connectivity Protocol, Rmy SQL, etc.

9. Data Variety - R can handle a variety of structured as well as unstructured data. It also provides various data modeling and data operation facilities due to its interaction with databases.

10. Compatible with other programming languages - Most of the functions are written in R itself, C, C++ or Fortran can be used for computationally heavy tasks. Java, .NET, Python can also be used to manipulate objects directly.

R code can be run without any compiler. It is an interpreted language and hence compiler is not need to run the code. Calculations are done with vectors. R is actually a vector language, hence anyone can add functions to a single vector without putting in a loop. R is hence powerful and faster than other languages.

Feature of R include:

1. Data inputs and data management. Data inputs such as data type, importing data and keyboard typing.

2. Data management such as data variables, operators.

Pros of R language:

1. It is the most comprehensive statistical analysis package, and new ideas often appear first in R.

2. R is an open source and can be run anywhere any time.

3. It is cross platform and runs on many operating systems.

Cons of R language:

1. The quality of some packages in R is less than perfect.

2. There is no customer support of R language.

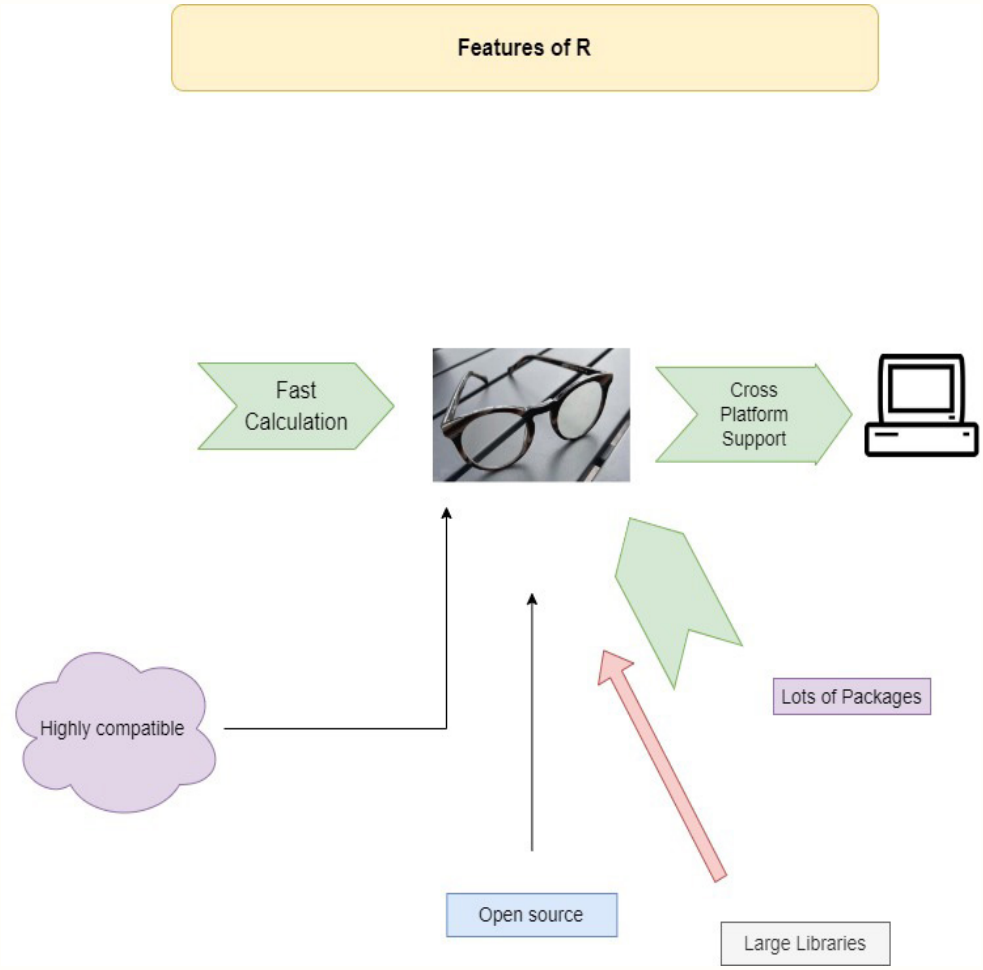
The R Environment:

This is an integrated suite of software that can be used for data manipulation, calculation and graphical display. It includes:

1. An effective data handling and storage facility
2. A suite of operators for calculations on arrays, in particular matrices
3. A large, coherent, integrated collection of intermediate tools for data analysis
4. Graphical facilities for data analysis and display either on-screen or on hard copy
5. A well developed, simple and effective programming language which includes conditions, loops, user defined recursive functions and input and output facilities.

The term environment is intended to characterize it as a fully planned and coherent system rather than an incremental accretion of very specific inflexible tools.

R has been designed around a true computer language, and it allows users to add additional functionality by defining new functions. R also has its own LaTeX like document format which is used to supply comprehensive documentation both on-line in a number of formats and in hard copy.



Prerequisites before learning R:

Before one jumps into R, it is highly recommended that they possess some basic knowledge of a few topics. These include:

1. Basic understanding of statistics, mathematics, and probability.

2. General understanding of data science and the process involved.

3. Basic understanding of various types of graphs and data representation techniques.

Unique features of R programming:

Since there are a large number of packages are available, there are many handy features in R. They include:

1. Its ability to perform directly on vectors and hence does not require too much looping.

2. It can pull data from APIs, servers, SPSS files and many other formats.

3. It is very useful for web scraping.

4. It can perform multiple complex mathematical operations with a single command.

5. It can create attractive reports combined with plain text with code and visualizations of the results if R markdown feature is used.

6. Since the user base is large, new ideas and technologies appear in the R community first.

Installation R base software:

Step I : R Base needs to be installed first. R is mainatined by an international team of developers and the software is available in multiple languages in their webpage "The Comprehensive R Archive Network". From here the version appropriate to the User's operating system can be downloaded. R is available for:

Windows operating system

Mac OS

Various flavors of linux

Installing R in windows is fairly simple as it comes bundled with its own installer which takes care of the entire installation process. As the user has to do is to double click on the downloaded binary file.

Step II: The windows executable file after being downloaded is double clicked to begin the installation process. All the user has got to do is keep clicking the next button till the confirmation screen appears saying that the process of installation is over. If the user is using a computer that is shared by others then Install for all users radio button needs to be selected to make the software available to all the users using the system. The first screen allows the user to choose the language of installation. R software is available in various common languages. It is preferable to allow the installation into the default folder created by the installer than customizing the process of installation. Since the user will have to install an Integrated Development Environment (IDE) software after installing R base software it will be fairly straight forward for the IDE to use R base software as it has been installed in to the default folder

Image showing CRAN webpage where the various flavors of R are available for download

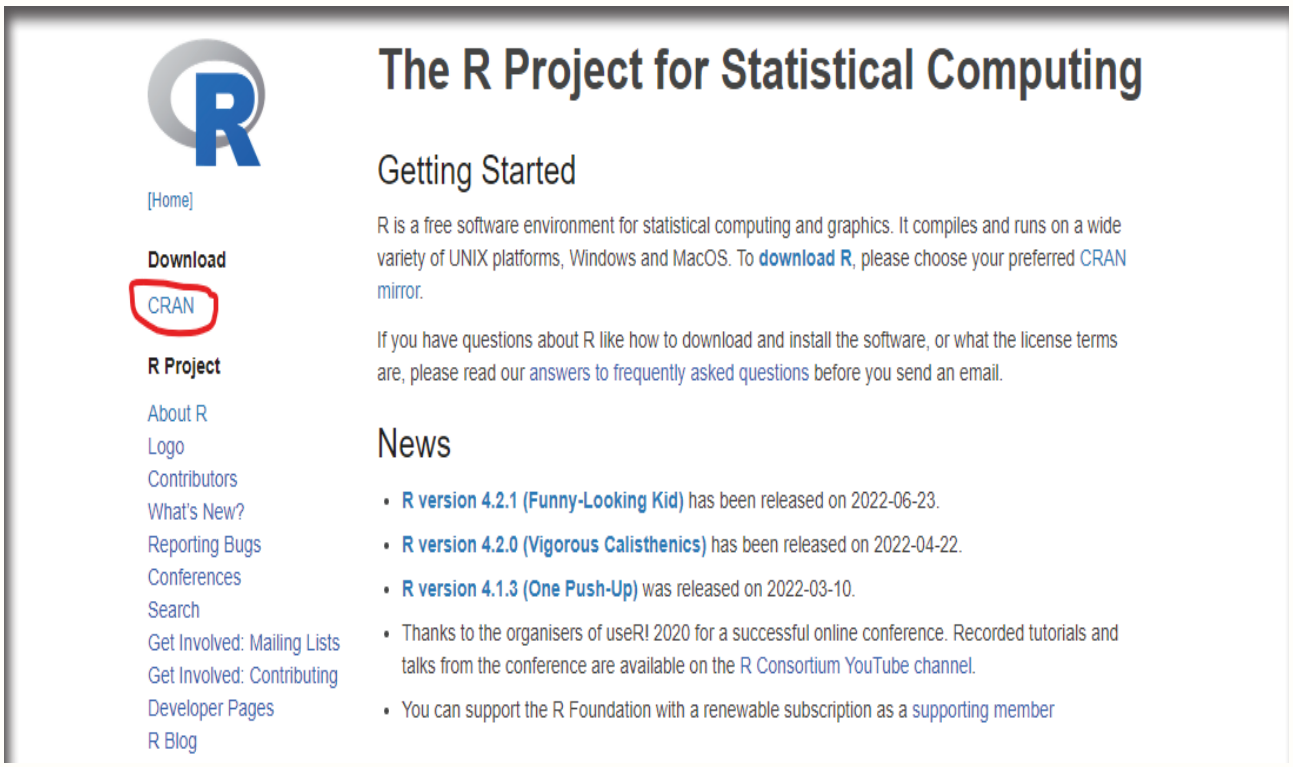
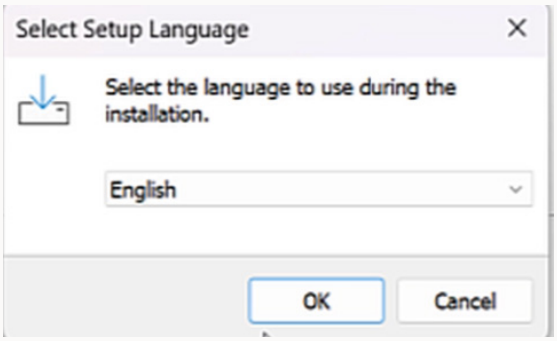


Image showing the official R project webpage



In the first screen shown above the language of the installation needs to be chosen before clicking on the OK button

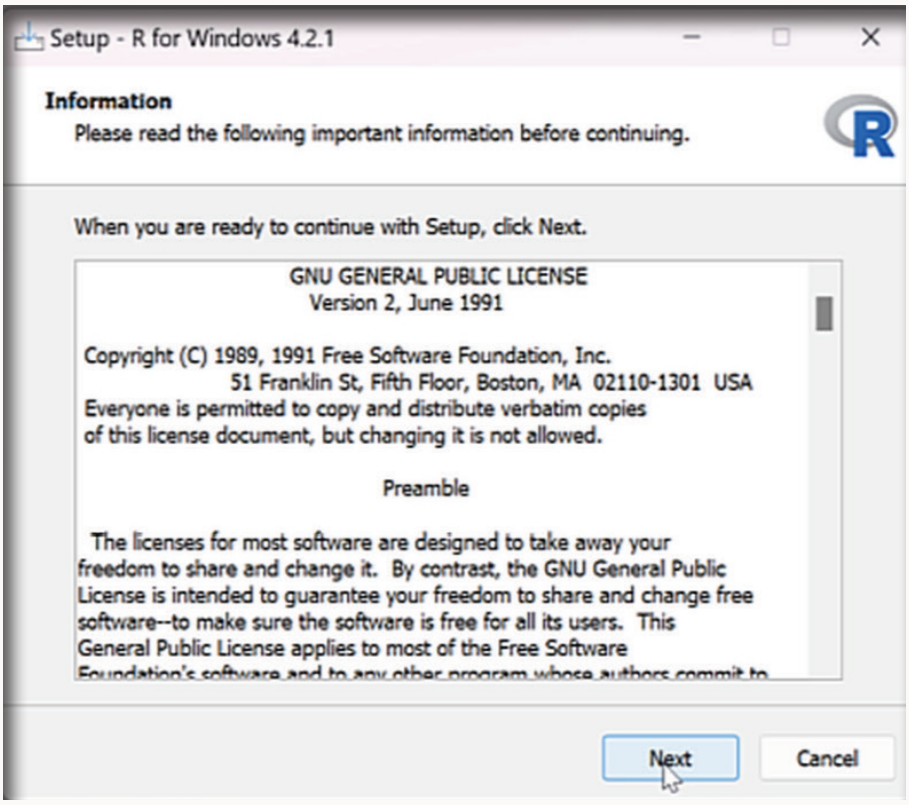


Image showing GNU licence screen which needs to be accepted by clicking the next button

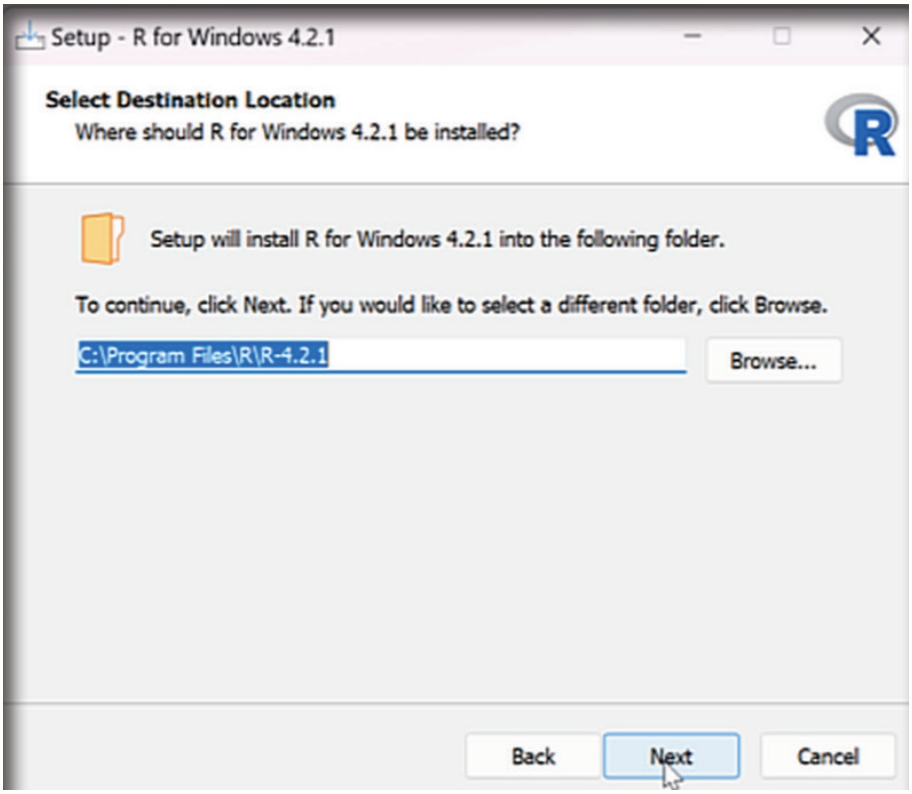


Image showing the screen that gives the choice of destination of location to the user. It is ideal for the user to allow the default settings by clicking on the next button. If the system has an SSD disk installed then installation is preferred in that disk as it would speed up the application process. If the user's system has multiple hard disks and one of them happens to be a SSD it is preferable to install it there.

R comes with both 32 bit AND 64 bit versions. The user will have a dilemma in choosing which version to use. Actually it does not matter as both versions use 32-bit integers, which indicates that they compute numbers to the same numerical precision. The difference occurs in the way each version manages the system memory. 64-bit R uses 64-bit memory pointers and 32-bit uses 32-bit memory pointers, this means that 64-bit has a larger memory space to use.

It should be pointed out that 32-bit builds of R are slightly faster than 64-bit builds. On the flip side 64-bit builds can handle larger files and data sets with fewer memory management problems. Hence if the operating system does not support 64-bit programs, or the installed RAM is less than 4 GB then it is ideal to install 32-bit R software. If the system supports 64-bit then the installer would install both versions of R.

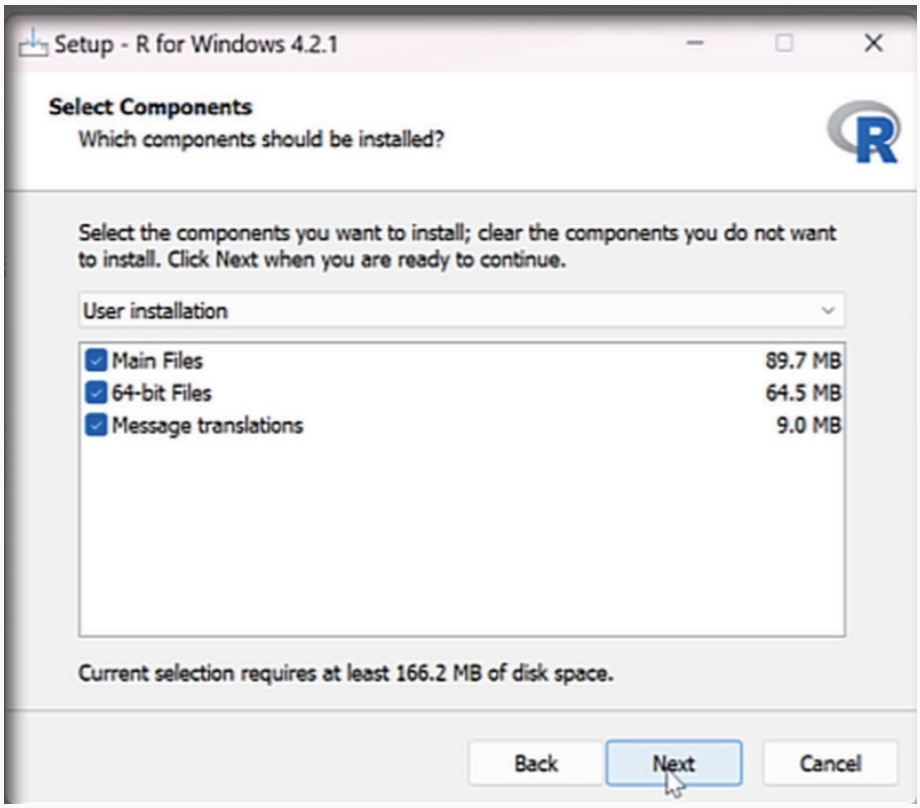


Image showing the screen that prompts user to select the desired components for installation. The user should choose the Main Files, 64-bit files if desired and Message translations if needed. The default settings is preferred and advisable. If the user wants 32 bit installation only, then 64-bit Files can be unchecked.

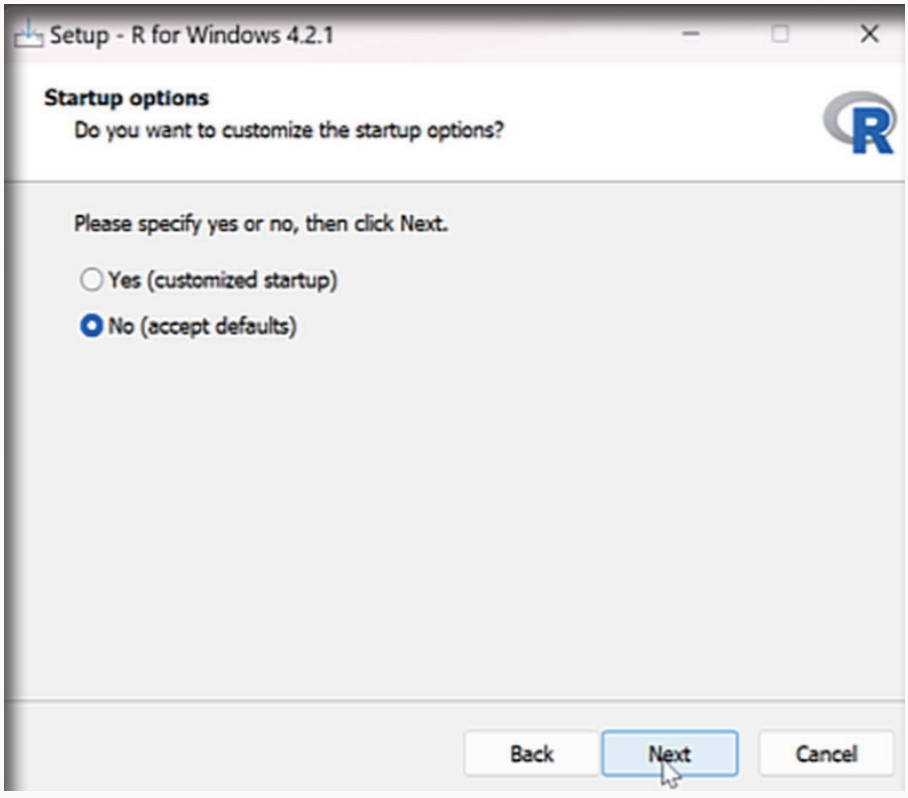


Image showing startup options window

Startup options:

When R is started, it will by default source a .Rprofile file if it exists. This allows the user to automatically tweak the R settings to meed the everyday needs. The startup package extends the default R startup process by allowing the user to put multiple startup scripts in a common "Rprofile.d" directory. If customization is needed for startup then during installation "customize startup radio button is selected" and in the ensuing window the customized file is pointed to enable customized startup. The user can have one file to configure the default CRAN repository and another one to configure their personal devtools settings. The user can also use a "Renviron.d" directory with mulitple files defining different environmental variables like language etc,. One file could contain the private GITHUB_pat key.
This customization is needed for advanced users who are well versed in R language scripting and advanced computing techniques. This step is narrated not to daunt the first time user but to illustrate the extensive customizations that are available within R environment which can be used if desired.

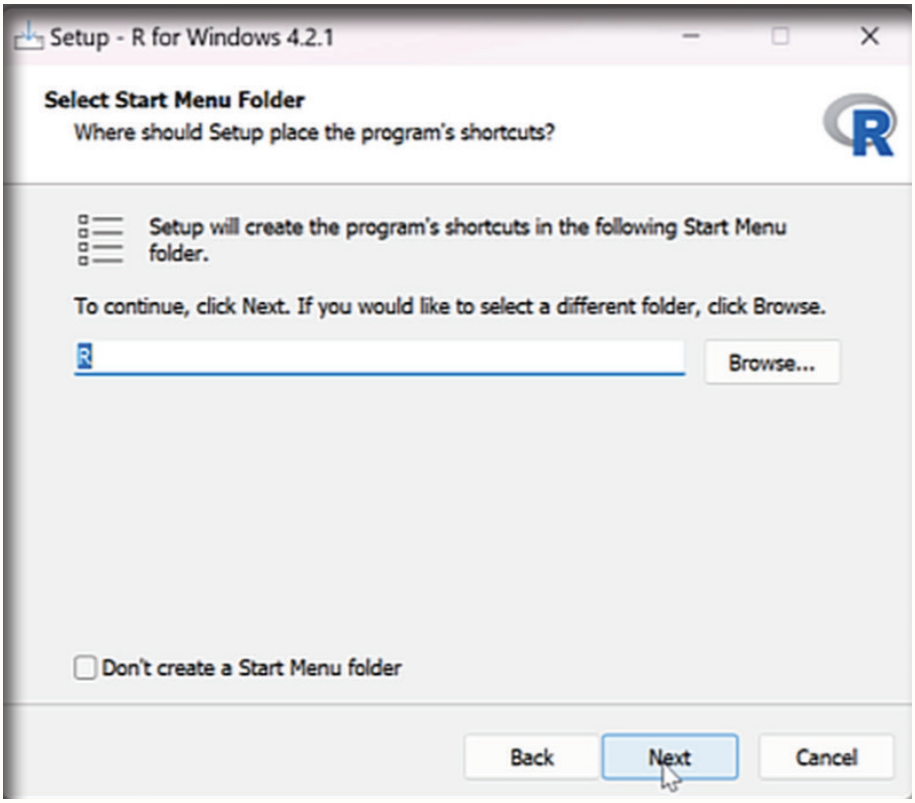


Image showing the prompt screen that allows the user to select the start menu folder where R short-cut is going to be stored. Here if the next button is clicked the defualt folder named R will be created in startup menu folder.

A small tip regarding the choice of installation folder in R programing installation:

If the user desires to install this software in a company owned computer where usually C drive access is not provided to the user as part of the company policy it is important to change the installation drive to where the user has access to. Installation will not progress if the user does not have access to the drive where installation folder is being created.

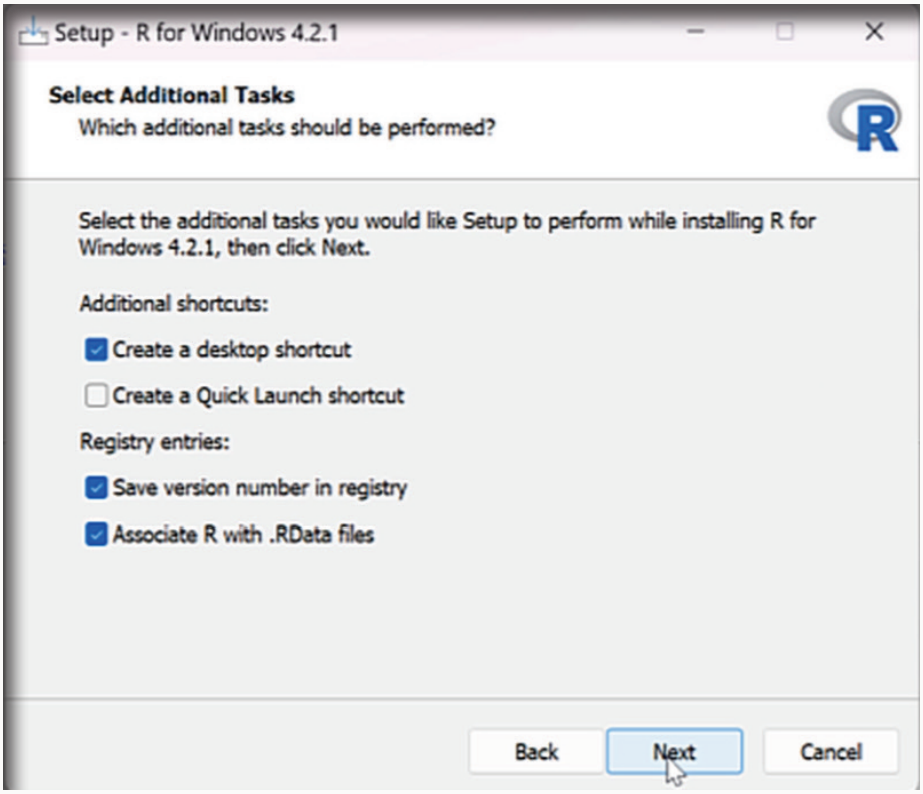


Image showing the installation screen where additional tasks can be selected during installation process.

In the image shown above the additional tasks that needs to be performed has been selected by default. The additional tasks already selected by default is sufficient for the installation to proceed. If the user desires to create a quick launch short cut then that box needs to be checked. Save version number in the registry helps in the process of identification of updates released if any. Another setting that has been chosen by default is Associate R with .RData files. This setting which is chosen by default will ensure that R files are associated with this software.

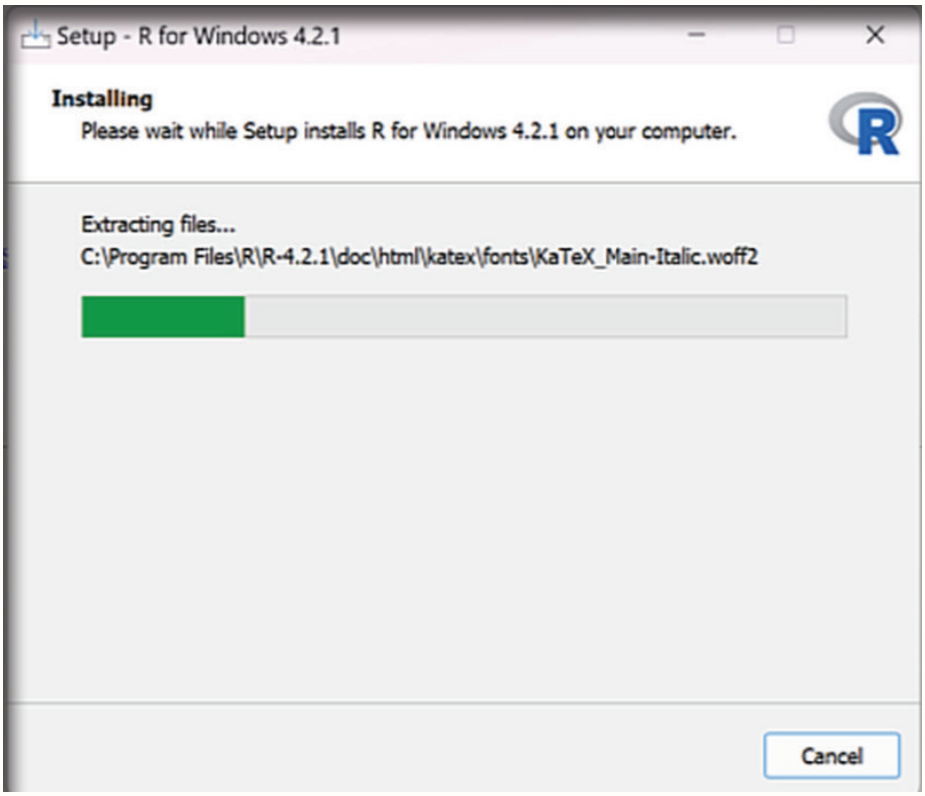


Image showing the file extraction process progressing

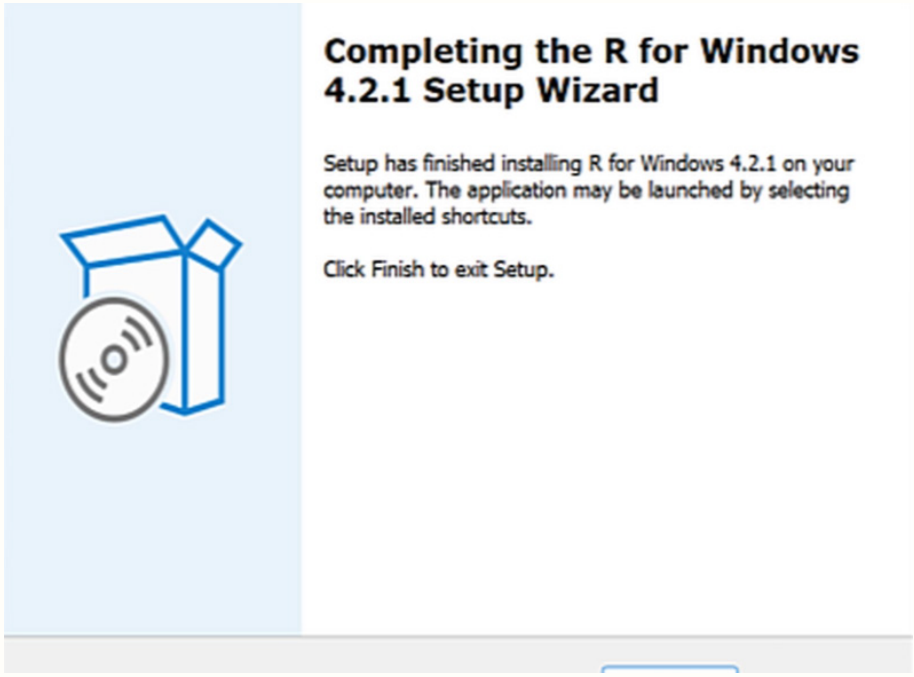


Image showing confirmation screen showing installation has been compteted successfully