Contents

Preface

The creation of this book has been for me, Mike McGrath, an exciting personal journey in discovering how the R programming language can be used today for data analysis and the production of beautiful data visualizations. Example code listed in this book describes how to produce R Scripts in easy steps – and the screenshots illustrate the actual results. I sincerely hope you enjoy discovering the exciting possibilities of R programming and have as much fun with it as I did in writing this book.

In order to clarify the code listed in the steps given in each example I have adopted certain colorization conventions. Components and keywords of the R programming language are colored blue, programmer-specified names are colored red, literal numeric values and literal character string values are colored black, and comments are colored green, like this:

```
# Write the traditional greeting.
greeting = "Hello World!"
print( greeting )
```

Additionally, non-literal values are colored gray like this: `color="Red"`

In order to readily identify each source code file described in the steps a file icon and file name appears in the margin alongside the steps:

Script.R

For convenience I have placed source code files from the examples featured in this book into a single ZIP archive. You can obtain the complete archive by following these easy steps:

1. Browse to **www.ineasysteps.com** then navigate to Free Resources and choose the Downloads section

2. Find R for Data Analysis in easy steps in the list, then click on the hyperlink entitled All Code Examples to download the archive

3. Next, extract the "MyRScripts" folder to a convenient location on your system

4. Now, follow the steps to call upon the R program interpreter and see the output

1 Getting started

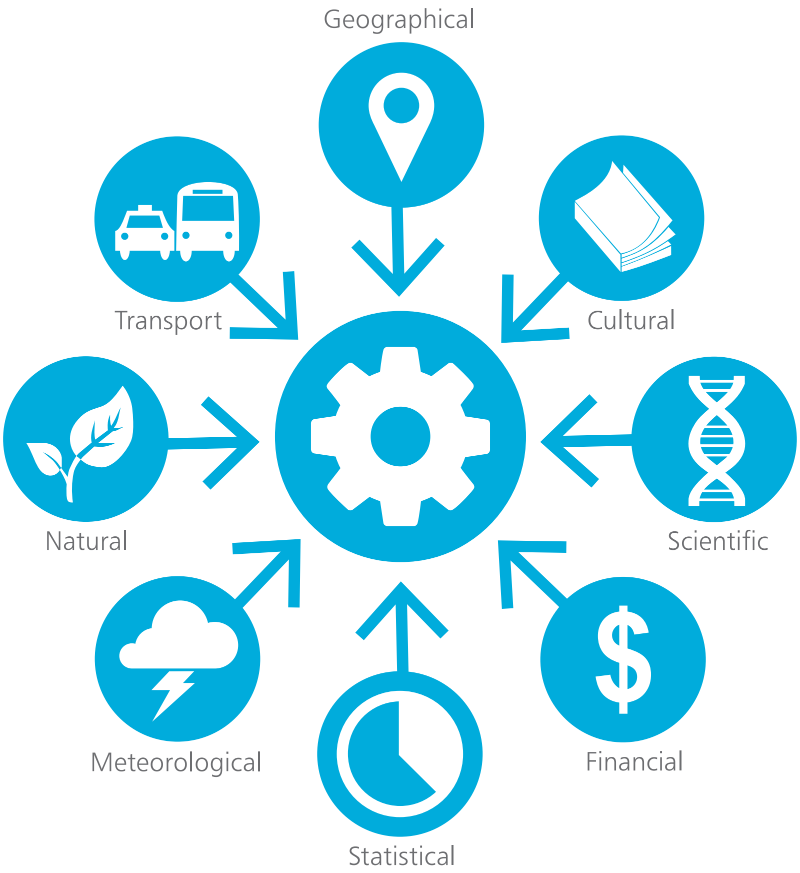
Welcome to the exciting world of R programming. This chapter describes how to set up an R environment and demonstrates how to create a first R program.

Understanding data

The term "data" refers to items of information that describe a (qualitative) status or a (quantitative) measure of magnitude. Various types of data is collected from a huge range of sources and reported for analysis to reveal pattern and trend insights:

Don't forget

This illustration depicts only some of the many data types that can be reported for analysis.



Hot tip

Around 13 billion devices are connected to the internet today. This is predicted to grow to 50 billion by 2020.

Data is increasingly being collected by devices that are able to report measurements for analysis via the internet ("The Cloud"). For example, devices that have temperature and humidity sensors can report measurements for instant analysis of climate conditions. The recent rapid decline in the cost of device sensors has given rise to the "Internet of Things" (IoT) that can easily and cheaply report vast amounts of data – this is often referred to as "big data". Big data consists of extremely large data sets that can best be analyzed by computer to reveal pattern and trend insights.

Data analysis (a.k.a. "data analytics") is the practice of converting collected data into information that is useful for decision-making. The collected "raw" data will, however, typically undergo two initial procedures before it can be explored for insights:

- **Data processing** – the raw data must be organized into a structured format. For example, it may be arranged into rows and columns in a table format for use in a spreadsheet.
- **Data cleaning** – the organized data must be stripped of incomplete, duplicated, and erroneous items. For, example, by the removal of duplicated rows in a spreadsheet.

After the data has been processed and cleaned it can be explored to discover its main characteristics. This may require further data cleaning to refine the data to specific areas of interest, or may require additional data to better understand its messages. Descriptive statistics, such as average values, might be calculated to understand the data. Algorithms might be used to identify associations within the data. Data visualization might also be used to produce a graphical representation of the data for examination.

After the data has been analyzed, the results can be communicated using data visualization to present tables, plots, or charts that clearly and efficiently convey the key messages within the data. Tables provide information in which the user can look up a specific number, whereas plots and charts provide information in a way that encourages the eye to make comparisons.

"R" is an interpreted programming language and software environment that is widely used for data analysis and visualization. The "RStudio" Integrated Development Environment (IDE) is often used with R, as RStudio provides a code editor, debugging features, and visualization tools that make R easier to use. The popularity of R has grown rapidly in recent years as the increase in big data has made data analysis more important than ever.

The R programming software and RStudio IDE are both available for Windows, Linux, and macOS operating systems, and both are used throughout this book to demonstrate R for data analysis.

Hot tip

"Data Science" is the study of how data can be turned into a valuable resource.

Hot tip

"Data Mining" is the process of searching large data sets to identify patterns.

Hot tip

"Data Product" is digital information that can be purchased.

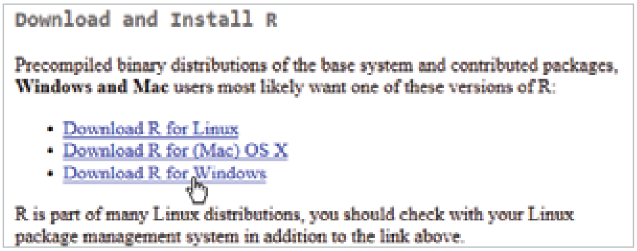
Installing R

The R programming language and software environment is freely available open source software that you can install onto your computer from the Comprehensive R Archive Network (CRAN):

If you are having difficulty downloading R click the CRAN **Mirrors** link at **cran.r-project.org** then choose a server near to your location.

1 Open a web browser and visit **cran.r-project.org**

2 Select the link appropriate for your computer operating system. For example, click **Download R for Windows**



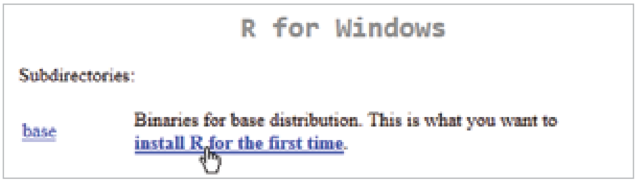
Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

3 Next, select the link for the **base** R distribution

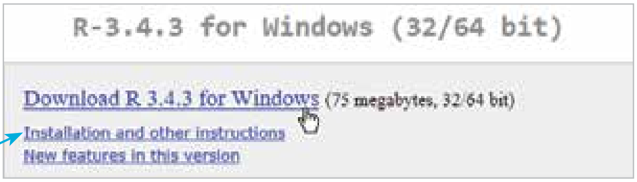


R for Windows

Subdirectories:

base Binaries for base distribution. This is what you want to install R for the first time.

4 Now, select the link to **download** the R installer



R-3.4.3 for Windows (32/64 bit)

Download R 3.4.3 for Windows (75 megabytes, 32/64 bit)
Installation and other instructions
New features in this version

Don't forget

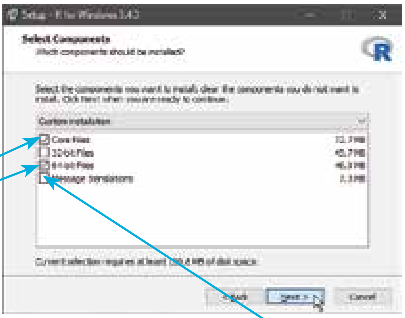
You can click the link for **Installation and other instructions** for more help with installation.

5 When the download has completed, run the installer to open the **R Setup Wizard** and click the **Next** button

6 Read the **License** information, then click the **Next** button to continue

7. Accept the suggested installation location, then click the **Next** button to continue

8. Choose to install **Core Files** and **32-bit Files** for a 32-bit machine, or choose to install **Core Files** and **64-bit Files** for a 64-bit machine, then click the **Next** button to continue



Setup - R for Windows 3.4.3

Select Components

Which components should be installed?

Select the components you want to install; clear the components you do not want to install. Click Next when you are ready to continue.

Custom installation

☑ Core Files 72.7 MB
□ 32-bit Files 45.7 MB
☑ 64-bit Files 46.3 MB
□ Message translations 7.3 MB

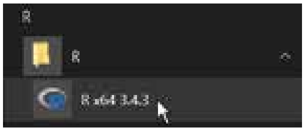
Current selection requires at least 126.4 MB of disk space.

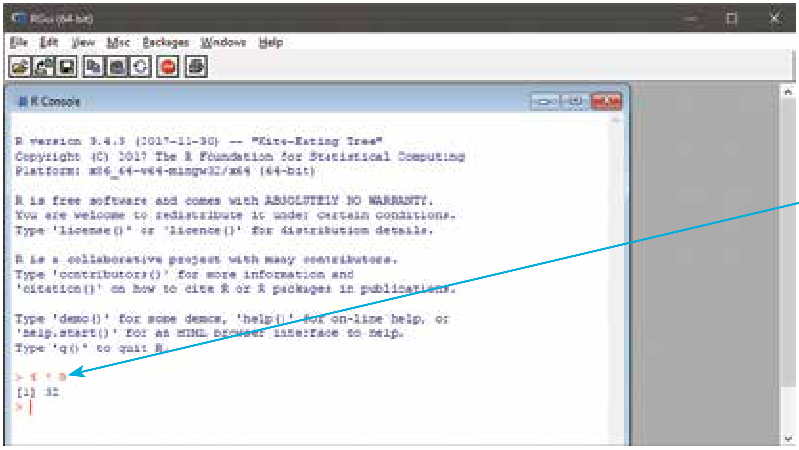
9. Choose **No (accept defaults)** to not customize startup options, then click the **Next** button to continue

10. Enter a name for a Start Menu folder (such as "R"), then click the **Next** button to continue

11. Choose additional tasks (such as **Create a desktop icon**), then click the **Next** button to begin the installation

12. When installation has completed, launch the **R** environment from the Start Menu folder you named





```
R version 3.4.3 (2017-11-30) -- "Kite-Eating Tree"
Copyright (C) 2017 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.

> 4 * 8
[1] 32
> |
```

Hot tip

You can find the System Type on Windows by pressing **WinKey** + **R** then entering **msinfo32**.

Hot tip

You can install **Message translations** for error messages, warning messages, and menu labels in languages other than English.

Hot tip

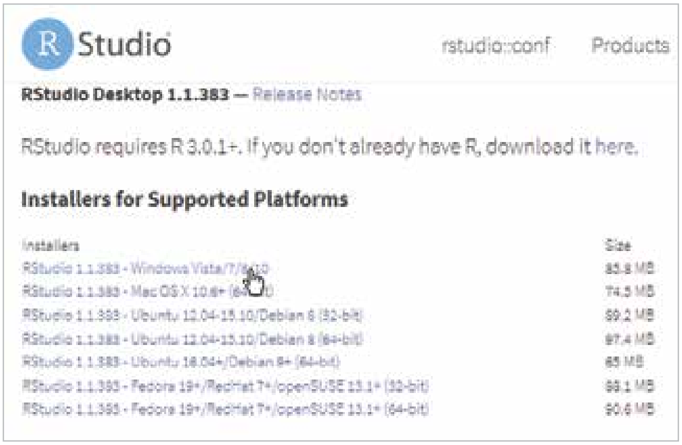
You can type expressions in the R Console to see their result – but the RStudio IDE is a much more effective programming environment.

Installing RStudio

The RStudio IDE has a freely available open source edition that you can install onto your computer from the RStudio website:

1. Open a web browser and visit the RStudio downloads page at **rstudio.com/products/rstudio/download**

2. Scroll down the page and select the **Installer** download link appropriate for your computer operating system. For example, click the edition for **Windows Vista/7/8/10**



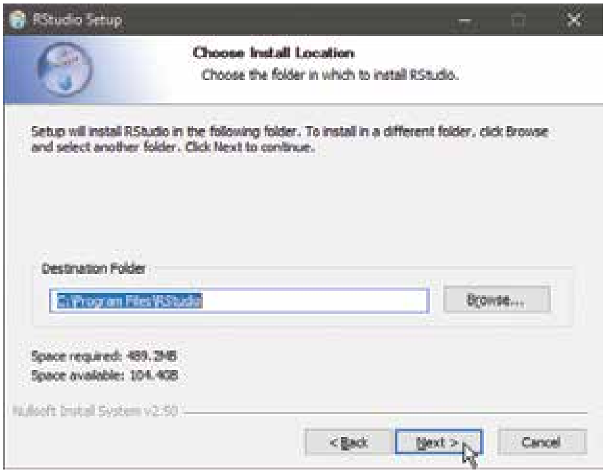
RStudio Desktop 1.1.383 — Release Notes

RStudio requires R 3.0.1+. If you don't already have R, download it here.

Installers for Supported Platforms

Installers	Size
RStudio 1.1.383 - Windows Vista/7/8/10	85.8 MB
RStudio 1.1.383 - Mac OS X 10.6+ (64-bit)	74.5 MB
RStudio 1.1.383 - Ubuntu 12.04-15.10/Debian 8 (32-bit)	99.2 MB
RStudio 1.1.383 - Ubuntu 12.04-15.10/Debian 8 (64-bit)	97.4 MB
RStudio 1.1.383 - Ubuntu 16.04+/Debian 9+ (64-bit)	65 MB
RStudio 1.1.383 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (32-bit)	88.1 MB
RStudio 1.1.383 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (64-bit)	90.6 MB

3. When the download has completed, run the installer to open the **RStudio Setup Wizard** – then click **Next**



Choose Install Location

Choose the folder in which to install RStudio.

Setup will install RStudio in the following folder. To install in a different folder, click Browse and select another folder. Click Next to continue.

Destination Folder: C:\Program Files\RStudio

Space required: 489.2MB
Space available: 104.4GB

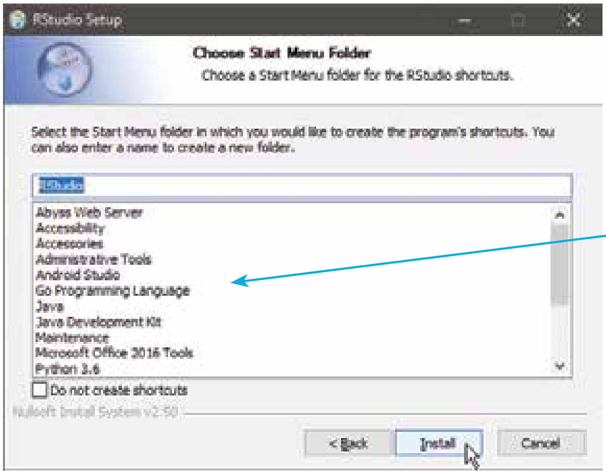
The RStudio software is available in Desktop and Server versions with Open Source Licenses and Commercial Licenses for each version – be sure to download the Desktop version with the Open Source License to try the examples in this book for free.

Beware

You must have R installed before you install RStudio. See pages 10-11 for the R software installation procedure.

4. Accept the suggested installation location and click the **Next** button to continue

5. Accept the suggested Start Menu folder name "RStudio" and click the **Install** button to continue



RStudio Setup

Choose Start Menu Folder

Choose a Start Menu folder for the RStudio shortcuts.

Select the Start Menu folder in which you would like to create the program's shortcuts. You can also enter a name to create a new folder.

RStudio

Abyss Web Server
Accessibility
Accessories
Administrative Tools
Android Studio
Go Programming Language
Java
Java Development Kit
Maintenance
Microsoft Office 2016 Tools
Python 3.6

□ Do not create shortcuts

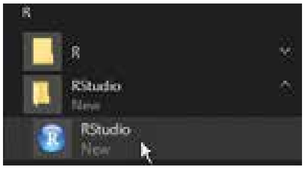
Nullsoft Install System v2.50

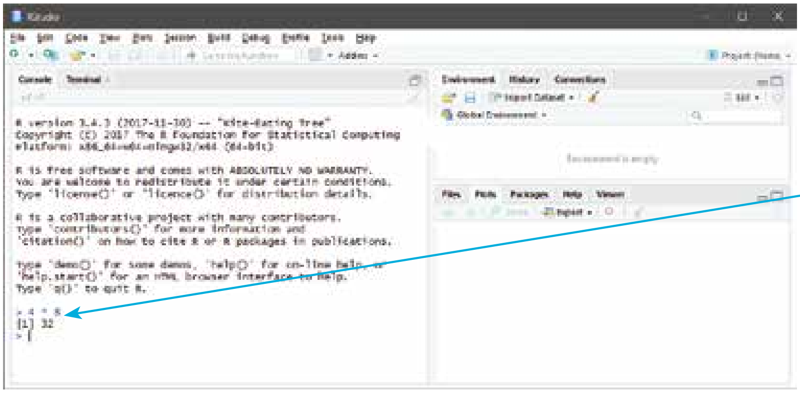
Don't forget

The items listed in this dialog box are the names of your existing Start Menu folders and will vary according to what you have installed on your computer.

6. When the installation has completed, click the **Finish** button to close the Setup Wizard

7. Launch the **RStudio** IDE from the Start Menu folder created by the Setup Wizard





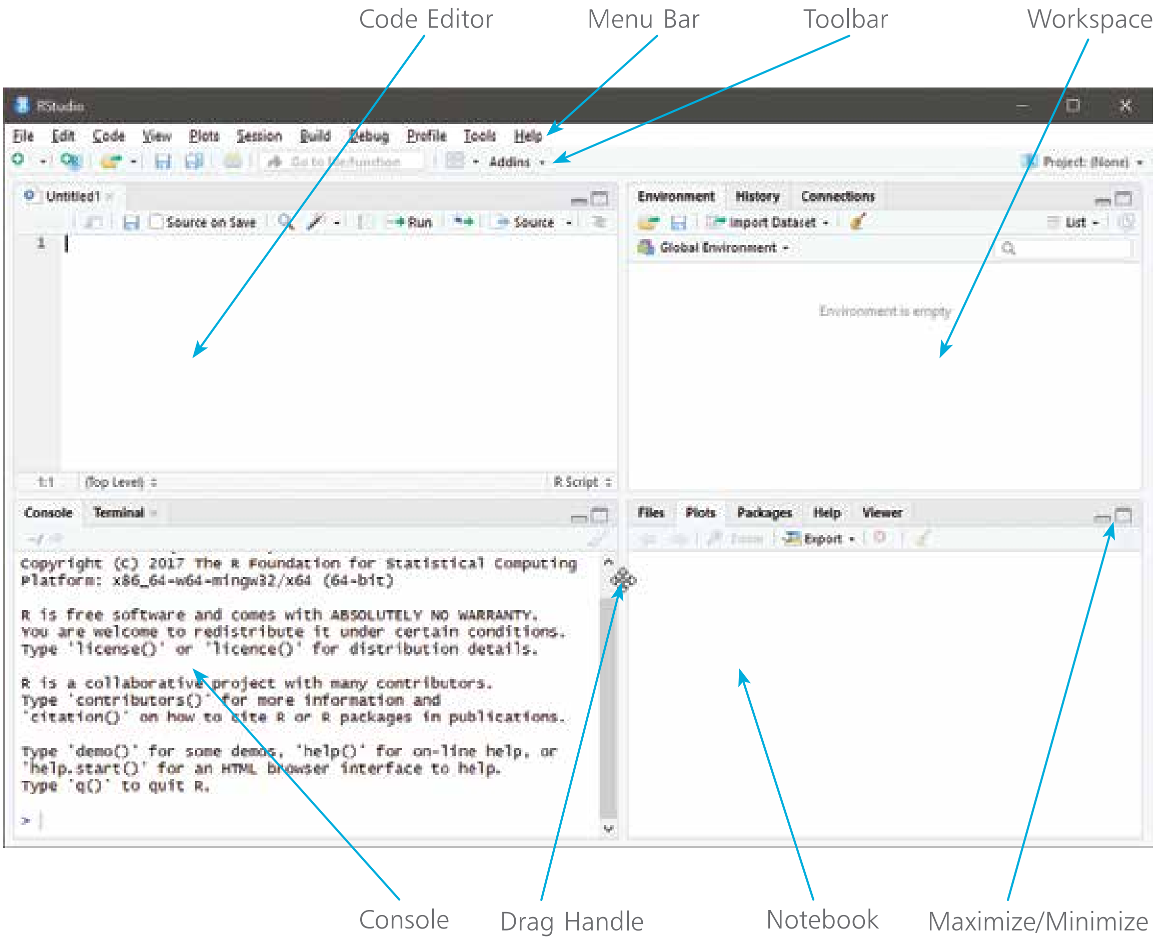
```
R version 3.4.3 (2017-11-30) -- "Kite-Eating Tree"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

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.

> 4 * 8
[1] 32
> 
```

Hot tip

You can type expressions in the RStudio Console to see their result, just as you can in the R Console – but the RStudio IDE can do so much more.

Exploring RStudio

The RStudio interface consists of a menu bar and toolbar positioned at the top of the window, and four main panes whose position can be adjusted to suit your preference. When you launch RStudio only three panes may be visible until you select **File**, **New File**, **RScript** on the menu bar to open the "Code Editor" pane. The default layout positions the four panes as shown below:



```
Copyright (C) 2017 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.

>
```

When the mouse pointer is placed on the border between any two panes, the pointer changes to a four-pointed "Drag Handle". This allows you to drag the vertical border to adjust the width of the left and right panes, and to drag the horizontal border to adjust the height of the top and bottom panes. The size of each pane can also be adjusted by clicking the Maximize and Minimize buttons.

Each RStudio pane can contain multiple tabs, and it is useful to initially explore each RStudio pane to understand its purpose:

Code Editor pane

The Code Editor is where you type or edit R Script code, and you see it automatically colored to highlight syntax – click this pane's Run button to see the script output appear in the Console pane.

R Script code can be saved as a file for later use, and multiple R Script files can be open on separate tabs in the Code Editor pane.

Console pane

- **Console tab** – This is where you can directly enter commands for immediate execution by the R interpreter.
- **Terminal tab** – This is where you can directly enter commands for execution by the operating system shell.

Workspace pane

- **Environment tab** – This is where you will see available objects such as variables and datasets.
- **History tab** – This is a list of your past commands executed by the R interpreter in the Console pane.
- **Connections tab** – This tab enables you to connect to databases to explore the objects and data inside the connection.

You can click on a data set listed in the Environment tab to open a spreadsheet of that data in the Code Editor pane.

Notebook pane

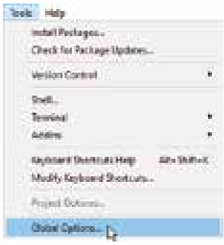
- **Files tab** – This is a file browser, which by default lists all the files in your working directory.
- **Plots tab** – This exciting tab is where your plots, graphs, and charts will appear as output from an R Script.
- **Packages tab** – This tab lists available packages that you can install to extend RStudio's functionality.
- **Help tab** – This is where you can seek assistance on the R language and RStudio IDE.
- **Viewer tab** – This is where you can see local HTML content that has been written to the session temporary directory.

You can click on an R Script file in the Files tab to open that file in the Code Editor pane.

Setting preferences

RStudio is highly customizable and it is worth setting up its features to better enjoy your R programming environment:

1. Create a new directory on your computer in which to save the R Scripts you will write. For example, on Windows you might create a directory of **C:\MyRScripts**

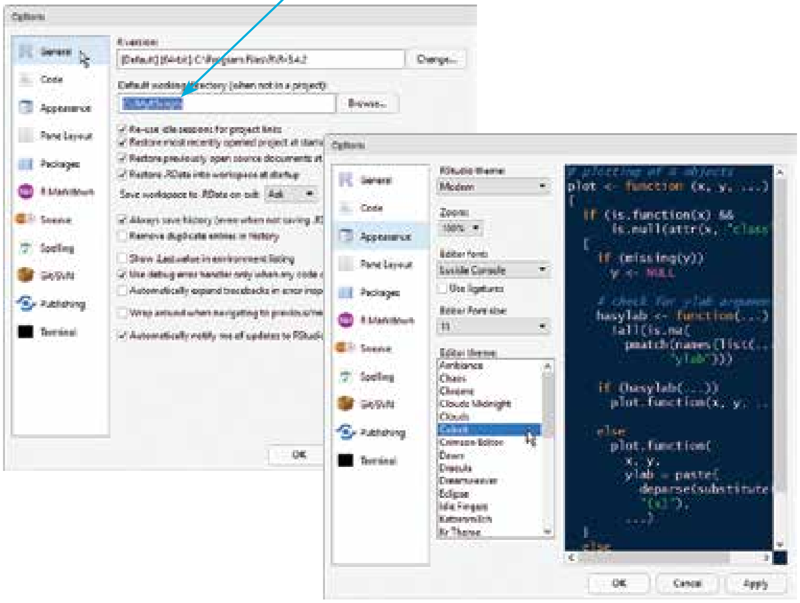


2. Launch RStudio then select **Tools**, **Global Options** on the menu bar – to open the "Options" dialog

3. Select **General** in the left panel of the "Options" dialog, then enter the path to the directory you created into the **Default working directory** box

Don't forget

Your home directory is set as the default working directory until you specify an alternative.



R version:
[Default] [64-bit] C:\Program Files\R\R-4.4.2

Default working directory (when not in a project):
C:\MyRScripts

☑ Re-use idle sessions for project links
☑ Restore most recently opened project at startup
☑ Restore previously open source documents at startup
☑ Restore .RData into workspace at startup
Save workspace to .RData on exit: Ask
☑ Always save history (even when not saving .RData)
☐ Remove duplicate entries in history
☐ Show .Last.value in environment listing
☑ Use debug error handler only when my code contains errors
☐ Automatically expand tracebacks in error inspector
☐ Wrap around when navigating to previous/next tab
☑ Automatically notify me of updates to RStudio

RStudio theme: Modern
Zoom: 100%
Editor font: Lucida Console
☐ Use ligatures
Editor Font size: 11
Editor theme:
Ambiance
Chaos
Chrome
Clouds Midnight
Clouds
Cobalt
Crimson Editor
Dawn
Dracula
Dreamweaver
Eclipse
Idle Fingers
Katzenmilch
Kr Theme

```
# plotting of R objects
plot <- function (x, y, ...)
{
  if (is.function(x) &&
    is.null(attr(x, "class"
  {
   if (missing(y))
     y <- NULL

   # check for ylab argumen
   hasylab <- function(...)
     !all(is.na(
       pmatch(names(list(...
              "ylab")))

   if (hasylab(...))
     plot.function(x, y, ...

   else
     plot.function(
       x, y,
       ylab = paste(
         deparse(substitute
         "(x)"),
       ...)
  }
  else
```

Hot tip

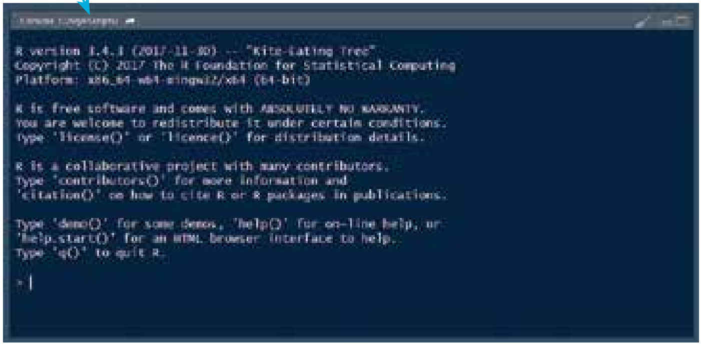
Themes with dark backgrounds, such as the "Cobalt" theme shown here, are often considered to be more restful on your eyes than those with white backgrounds.

4. Next, select **Appearance** in the left panel, then click items in the **Editor theme** box to preview possible color themes

5. Use the **Editor font** and **Editor font size** drop-down menus to choose your font preferences

6 Click the **Apply** button to change the RStudio settings

7 Click the **OK** button to close the "Options" dialog and see your preferences have been applied – the working directory path appears on the Console title bar



```
R version 3.4.3 (2017-11-30) -- "Kite-Eating Tree"
Copyright (C) 2017 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.

> |
```

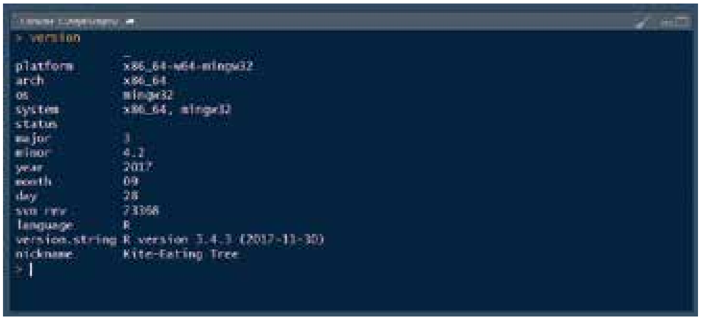
Hot tip

You can click the ➦ arrow button on the Console pane title bar to reveal the working directory's files in the Files pane.

8 You next need to select a pane to work with in RStudio – click on the Console pane to select it

9 Click the 🖌 brush button on the Console pane's title bar, or press **Ctrl** + **L** keys, to clear existing Console content

10 Now, type **version** at the Console prompt, then hit **Enter** to run the command – see the R interpreter output version details in the Console window



```
> version
               _
platform       x86_64-w64-mingw32
arch           x86_64
os             mingw32
system         x86_64, mingw32
status
major          3
minor          4.2
year           2017
month          09
day            28
svn rev        73368
language       R
version.string R version 3.4.3 (2017-11-30)
nickname       Kite-Eating Tree
> |
```

Beware

Commands typed at the Console prompt must be entered again to run the command once more – whereas commands typed in the Code Editor can be run repeatedly.

Dark background themes are great for on-screen viewing but all ensuing screenshots throughout this book use a white background theme (TextMate) for better on-page clarity.

Creating an R Script

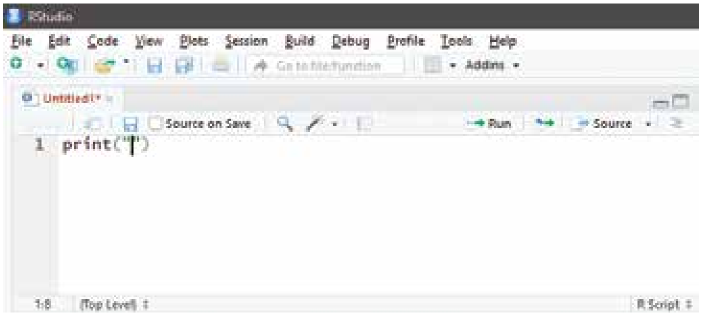
Unless you simply want to test a snippet of code directly at a Console prompt, you should always create an R Script using the Code Editor – so that your code can be run whenever required:

Hello.R

1 Launch RStudio, then click **File**, **New File**, **R Script** on the menu bar to open the Code Editor pane

2 Click on the Code Editor pane to select it and see a blinking cursor appear – here, type the command **print( )**

3 Type a **"** double-quote character between the command's parentheses and see RStudio automatically add a second character after the cursor – so you cannot forget the final double-quote that is required to enclose a text string



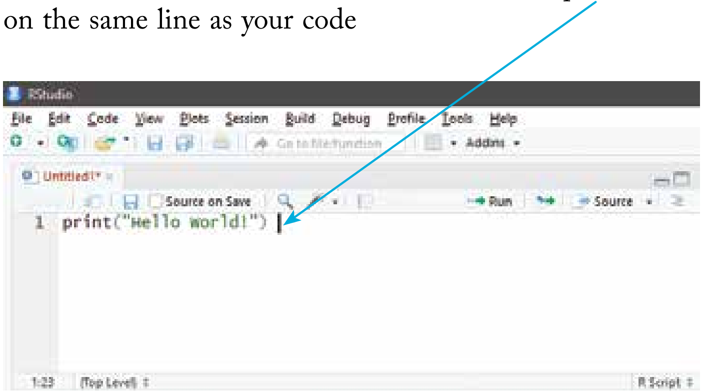
```
print("")
```

Don't forget

The command here is calling the built-in R **print( )** function. The R language is case-sensitive, so typing the command as **Print( )** or **PRINT( )** will simply produce an error.

4 Next, type the traditional program greeting **Hello World!** text string between the double-quotes

5 IMPORTANT: Ensure that the cursor is now positioned on the same line as your code

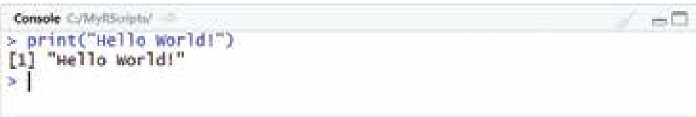


```
print("Hello world!")
```

Beware

The R interpreter will only run code on the line containing the cursor or multiple lines that you have selected (highlighted) by dragging the cursor over them.

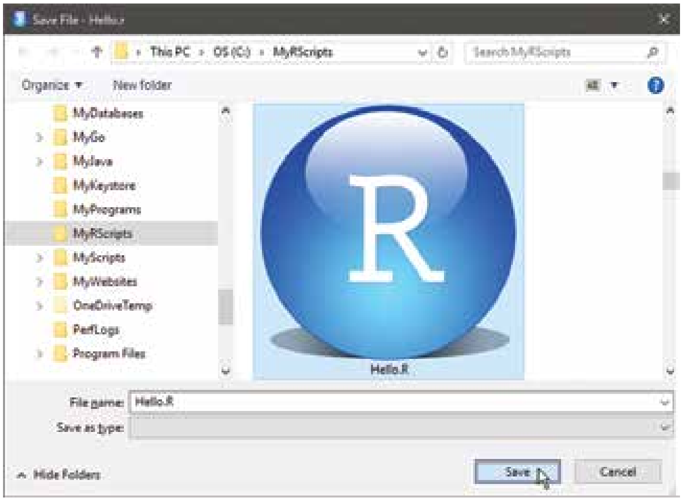
6 Click the Run button in the Code Editor, or press the **Ctrl** + **Enter** keys, to run the code – see the R interpreter repeat the code and display its output in the Console pane



```
Console C:/MyRScripts/
> print("Hello World!")
[1] "Hello World!"
> |
```

7 Click the Save button in the Code Editor, or press the **Ctrl** + **S** keys, to open the "Save File" dialog

8 Save the R Script as a file named "Hello.R" in the current working directory

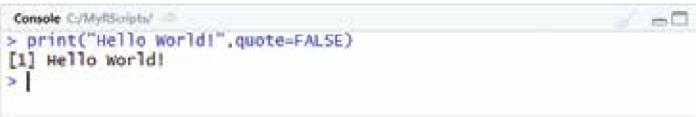


Save File - Hello.r

File name: Hello.R

9 Edit the command in the Code Editor by adding a second argument between the parentheses to become **print("Hello World!", quote=FALSE)**

10 Run the code again – see the R interpreter repeat the code and display its output with quotes now suppressed



```
Console C:/MyRScripts/
> print("Hello World!",quote=FALSE)
[1] Hello World!
> |
```

Hot tip

The bracketed number [1] that appears before the output indicates that the line begins with the first value of the result. Some results may have multiple values that fill several lines, so this indicator is occasionally useful but can generally be ignored.

Hot tip

Click the Open button in the Code Editor, or press the **Ctrl** + **O** keys to open the "Open File" dialog then choose a saved R Script file to reopen in the Code Editor. Click the arrow button beside the Open button to see a list of recently opened files that you can select to quickly reopen.

Summary

- Data is items of information that describe a qualitative status or a quantitative measure of magnitude.
- Devices that are connected to the internet are able to report sensor measurements for analysis in The Cloud.
- The decline in the cost of device sensors has given rise to the Internet of Things that can report on vast amounts of data.
- Big data consists of large data sets that can best be analyzed by computer to reveal pattern and trend insights.
- Data analysis is the practice of converting collected raw data into information that is useful for decision-making.
- Before analysis, raw data must be organized into a structured format and cleaned to remove incomplete, duplicated, and erroneous items.
- After data has been analyzed, the results can be communicated using data visualization to present tables, plots, or charts that efficiently convey the messages within the data.
- R is an interpreted programming language and software environment for data analysis and data visualization.
- RStudio is an Integrated Development Environment for R that provides a code editor, debugger, and visualization tools.
- The RStudio interface consists of a menu bar and toolbar, plus Code Editor, Console, Workspace, and Notebook panes.
- R Script code typed into the Code Editor can be run to see its output appear in the Console.
- Code snippets can be typed at the Console prompt for immediate execution by the R interpreter.
- RStudio's Global Options let you choose colorization themes, font settings, and default working directory.
- R Script in the Code Editor can be saved as a file with a **.R** file extension so the code can be re-run whenever required.