数据科学与大数据技术

R 4 编程入门与数据科学实战

[美]马塔·威利(Matt Wiley)
[澳]乔舒亚·威利(Joshua F. Wiley) 著
孙云华　郭涛 译

清華大學出版社
北　京

北京市版权局著作权合同登记号　图字：01-2021-6298

图书在版编目(CIP)数据

R 4 编程入门与数据科学实战 / (美) 马塔 • 威利(Matt Wiley)，(澳) 乔舒亚 • 威利 (Joshua F. Wiley) 著；孙云华，郭涛译. —北京：清华大学出版社，2023.4
(数据科学与大数据技术)
书名原文：Beginning R4: From Beginner to Pro
ISBN 978-7-302-62938-2

Ⅰ. ①R… II. ①马… ②乔… ③孙…④郭… III. ①程序语言—程序设计 Ⅳ. ①TP312

中国国家版本馆 CIP 数据核字(2023)第 038505 号

责任编辑：王　军
装帧设计：孔祥峰
责任校对：成凤进
责任印制：丛怀宇

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-83470000　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：天津鑫丰华印务有限公司
经　　销：全国新华书店
开　　本：170mm×240mm　　**印　　张**：22.75　　**字　　数**：471 千字
版　　次：2023 年 6 月第 1 版　　**印　　次**：2023 年 6 月第 1 次印刷
定　　价：99.80 元

产品编号：093136-01

译 者 序

随着时代的发展，数字化已经渗入人们生活的各个角落，当今社会的发展验证了“科学技术是第一生产力”的论断。编程语言在整个数字化发展进程中是必不可少的工具。只有通过这个工具，才能对杂乱的数据进行有效的分析处理。

现在，越来越多的人想要走进数字化这个领域，不仅是成年人，甚至有很多儿童都逐步开始了编程语言的学习。正是在数以万计程序员的共同努力下，才有了现在各种便捷的数字化服务。编程语言有很多种，本书将专门讲解一种难度较低的解释性语言——R 语言。

本书的两名作者均为从事编程以及教育方面的专家，他们用详尽的语言，以初学者的角度进行知识点的讲解，每个细节都手把手教学，以让读者悉数掌握所有知识点，在每章的结尾都安排理论与实操相结合的习题。与同类书籍相比，除了内容详细，本书最用心的一点是尽可能地避免运用生僻的专业术语，如果无法避免则会追加详细的解释，以便读者理解领会。

本书的开头部分对主程序、运行环境以及相关程序包的安装做详细介绍，除了介绍本书学习相关的程序包，还延伸介绍一些其他的功能极为强大的程序包，以供读者了解。此外，可圈可点的是，本书针对不同的操作系统，详细讲解了每一步的操作过程。后续则从简单的数据导入导出开始构建使用基础，并形成相应的知识框架。

R 语言在此处主要用来进行数据的处理分析，也就是通过 R 语言来运用统计学的相关知识。本书的绝大部分内容讲述如何将 R 语言与统计学结合并应用。首先从如何运用 R 语言抽取完美样本开始，到如何运用 R 语言分析样本以达到分析总体的目的，例如，概率分布、相关性、回归关系、假设检验、方差分析等方面的知识，其中尤为重要的一点是，如何运用 R 语言实现数据的可视化。可视化不仅可以让不易于观察的数字呈现更明显的规律，还便于验证数据的准确性。

读者通过本书的学习，不仅能掌握 R 语言的用法，更能将 R 语言与统计学知识相结合。在数字化高速发展的今天，掌握数据处理的必要技能，将让生活与工作更加便捷。

由衷希望各位读者可以如书名一样，通读完本书即可从初学者成长为专家。

作 者 简 介

Matt Wiley 领导维多利亚大学的机构有效性、研究和评估部门，同时负责促进战略和单位规划、数据知情决策、州/地区/联邦问责制度的发展。作为一名终身数学副教授，他曾在数学教育(加利福尼亚州)和学生参与(得克萨斯州)活动中获奖。此外，Matt 拥有加利福尼亚大学和得克萨斯 A&M 系统的计算机科学、商业和纯数学学位。

除了学术成就，他还参与合著了三本关于流行的 R 编程语言的图书，并担任一家统计咨询公司的管理合伙人近 10 年。他还拥有使用 R、SQL、C++、Ruby、FORTRAN、JavaScript 语言的编程经验。

作为一名程序员、出版作家、数学家和变革型领导者，Matt 总是将写作的热情同解决逻辑和数据科学问题的乐趣融为一体。无论是在会议室，还是在教室，他都喜欢采用动态的方式与跨学科和多元化的团队进行合作，使复杂的想法和项目变得易于解决。

Joshua F. Wiley 是莫纳什大学脑与健康特纳研究所以及心理科学学院的讲师。他在加利福尼亚州大学洛杉矶分校获得博士学位，并完成了初级保健和预防方面的博士后培训。他的研究采用先进的定量方法来解释心理社会因素、睡眠和其他与心理和身体健康有关的健康行为之间的动态关系。他独立开发或与他人共同开发了许多 R 语言程序包，包括用于运行贝叶斯尺度-位置结构方程模型的 varian 程序包；将 R 语言链接到商业 Mplus 软件的 MplusAutomation 程序包；用于更快逻辑运算的额外运算符；用于诊断、效应大小和轻松显示多级/混合效应模型结果的多级工具；辅助 JWileymisc 进行数据探索或者加速分析的函数。

技术评审员简介

Rachel Winkenwerder 是维多利亚大学的数学系副教授，并担任机构有效性、研究和评估部门的助理主任。她曾任教于中国和美国得克萨斯州的中高等教育机构，具有丰富的数学教学经验。Rachel 对情景化课程的实际开发有着深刻的理解。她最近在高等教育方面的工作包括共同主持她所在机构的课程和教学委员会、审查区域认证的叙述以及主导学术评估。她将统计学方法的课程与 R 编程语言相结合，且拥有计算机科学、数学和教育学学位。

致　　谢

真诚地感谢我们所有的学生，多年来，他们给予了我们许多重要且丰富的经验。

前　言

本书内容在某种程度上严格符合本科生的数学课程要求，也可以将此称为本书的“核心”。在这方面，本书可能适合添加到初级或中级统计方法课程中)。此外，根据最后几章的内容，它也可能适用于社会科学(例如心理学或社会学)的高年级本科课程。

除了以上这些目标，本书旨在让读者通过 R 编程语言的实际应用，亲自体验统计思维。理论是难以被忽视的，技术和理论将通过模型、视觉效果和其他直观的方法来呈现。本书真正的目的是以容易理解的方式分享复杂的数字构造语言。这种实际应用——有时也可称为经验和定量分析能力——旨在使读者能够批判性地思考和探索日益复杂的数据集，更准确地描述和总结大量信息，然后进行分析、建模，最后将结果清楚地传达给专业和非专业人士。

本书由两部分构成。第一部分旨在有效地逐步引导读者安装 R 编程语言，并了解其所需的基本计算机环境。本书将尽量避免“技术用语”，坚持使用日常用语，在最初阶段尽可能快地引导读者并灵活地转向研究实际的统计数据。第二部分介绍统计学，依次为总体、样本、描述性统计、概率、分布、相关性、回归、置信区间、假设检验和方差分析(ANOVA)。本书在不回避技术数学理论的同时，首先引入统计学思想的概念。本书的内容将从读者阅读阶段到帮助读者亲自参与学习中的实践操作，目标是建立一个坚实的、真实的、连贯的上下文关系，从而使理论更具有相关性。作者的目的是写一本使读者“能在日常生活中使用统计学”的书，而不是写一本完成考试之后很快就被遗忘的书。

最后，虽然本书适用于本科课程，但也将提高读者使用 R 编程语言的能力，并为使用 R 语言进行研究、数据科学、机器学习、动态报告和可视化定制奠定基础，因此，本书也非常适合希望掌握 R 语言的有一定经验的数据分析师、希望获得更强的技能和掌握统计学知识的研究生，以及任何喜欢学习数据和统计相关知识的人。

感谢你花时间和精力阅读我们的书。请务必通过本书封底的二维码下载源代码及获取书中网址，并参与本次学习实践。如果有任何问题，请随时与我们联系。

目　录

第 1 章

R 语言的安装

学习统计学的第一步就是安装程序，整个过程就像是观看魔术师的表演。首先，需要了解一些应用统计学的知识，并学习如何编程。虽然这些内容看起来很困难，但请不要放弃。本章将逐步介绍 R 语言的安装，无论是使用个人计算机，还是使用公司/机构的计算机进行学习，本书都将用日常用语提供每一步的指导。如果需要，本书也将给 IT 部门提供足够的完整材料(适用于企业/机构学习者)。

值得一提的是，通过学习使用 R[16]语言进行统计性思考，不仅可以学习免费的开源软件，还可以从第一天开始就像在“现实”生活中一样使用统计学知识。另外，还可以学习一些其他非常有用的应用技能。

本章涵盖以下内容：

- 从 CRAN [2]下载最新版本的 R 语言，并安装在 Windows 或 mac OS 操作系统的计算机上。
- 下载 RStudio 软件并安装在 Windows 或 mac OS 操作系统的计算机上。
- 了解 RStudio 软件的项目环境。
- 通过应用一些基本的 R 语言代码来检测安装技能的掌握情况和对相关知识的理解。

如果你对 R 语言已经有所了解，并希望直接学习统计学的相关知识，可以先在表 1-1 中查阅本书所需要的软件。如果想要按照本书的规划学习，请继续阅读本章内容。

表 1-1　R 语言技术栈

软件名称	网址
R 4.0.2	网址 1-1
RStudio 1.2.x	网址 1-2
Windows 10	网址 1-3
macOS	网址 1-4

1.1 技术栈

表 1-1 中的技术软件往往以一定的顺序相互作用，这种项目运行所需的软件列表被称为技术栈。如果安装 R 语言时遇到问题，首先要做的就是将技术栈分享给帮助解决问题的人，因为计算机的操作系统可能与正在运行的 R 语言版本不一致。本书是采用 R.Version()[["version.string"]]编写的，未来的 R 语言版本应该不会出现太大变化，因此本书中的内容可以持续实践操作。

1.2 操作系统升级

R 语言对计算机操作系统(如 Windows 或 macOS)的版本不做强制要求，但最好是最新的版本。最新的版本有两个好处：首先，需要安装的 R 语言和 RStudio 软件均为最新版，它们都是在最新的计算机操作系统上进行测试的；其次，大多数技术类型的学习与工作更倾向于保持最新的软件版本，而且有一个目前被大众熟悉的系统版本将帮助你更好地克服学习中遇到的困难。

1.2.1 Windows

更新到 Windows 10 系统很简单，在操作系统的搜索栏(Windows 标志右边的放大镜处)输入“check for updates”，然后选择 Check for updates System settings 选项，随后在弹出的 Windows Update 对话框中单击 Check for updates 按钮。上述步骤成功后，就会出现包含当前时间的文本“Last checked: Today”。

实际安装 R 语言之前，可以先验证计算机是不是 64 位操作系统，因为计算机也有一定可能是 32 位操作系统。同样，在 Windows 操作系统搜索栏输入“About your PC”，就会出现一个标题为“About your PC System settings”的选项，单击该选项，打开一个包含计算机操作系统信息的界面，从中可以看到该计算机系统类型为 64 位还是 32 位。无论是哪种情况，在安装 R 语言时都要记住这一项。

1.2.2 macOS

R 语言需要在最新版本的 macOS 上运行，这极为重要，因为不同版本的系统安装过程不同。在写这本书的时候，最新的系统版本为 macOS Catalina，在你阅读本书时，可能会发布 macOS Big Sur 或者更高级版本，可以通过网址 1-5 获得关于系统升级的更多帮助。

1.3　从 CRAN 下载并安装 R 语言

操作系统更新之后，即可下载最新版本 R 语言。撰写本书时，R 语言的最新版本为 4.0.2 (2020-06-22)，可于网址 1-1 中下载。但是，安装的 R 语言版本会根据计算机操作系统的不同而略有不同，所以需要选择最适合自己的版本。

1.3.1　Windows

在 Windows 10 上安装 R 语言需要使用网络浏览器(如 Chrome、Firefox 或 Edge)访问网址 1-1，网页上有一个标题为“Download and Install R”的文本框，单击 Download R for Windows 选项，跳转到链接网址 1-6。因为需要安装基础版本的 R 语言，所以接下来单击 base，跳转到链接网址 1-7，网页顶部的第一个链接就是 Download R 4.0.2 for Windows (XX megabytes, 32/64 bit)。

当阅读并学习本书的时候，R 语言很有可能更新到 4.0.2 以上的版本，兆字节也有可能略有不同。此时，只需下载最新版本，只要主版本仍旧是 4.x.x，本书的全部内容就可以实践操作。

下载完成后，打开保存 R-4.0.2-win.exe 文件的 Downloads 文件夹。大多数情况下，网络浏览器会弹出一个提示框提醒文件位置，如果没有，可以先按住键盘上的 Windows 键再按下 E 键，打开 Windows 资源管理器窗口，通常，窗口左侧有一个标题为 Downloads 的文件夹。

一旦找到 R-4.0.2-win.exe 文件，双击它就可以开始安装。选择继续，并接受所有默认选项，根据需要单击 Next 和 OK 按钮。

1.3.2　macOS

为了让 R 语言在 macOS 上更灵活地运行，建议使用几个附加工具。

在安装 R 语言之前，需要从 App Store 下载安装 Xcode 工具，安装完成后，记得打开并选择接受条款，否则该工具可能不起作用。

安装 Xcode 工具之后，还需要安装命令行工具。首先打开终端(如果找不到，请尝试使用 spotlight 搜索)，然后输入 xcode-select --install，之后按 Enter 键运行。如果遇到任何访问问题，可能需要启用 root 账户，可以在终端中输入 dsenableroot，然后按 Enter/Return 键启用 root 账户。

注意

如果终端要求输入密码，请输入登录 Mac 计算机时使用的密码，然后按 Enter/Return 键。在终端输入密码时不显示字符，但字符已经输入。

安装 XQuartz/X11 工具。访问网址 1-8 下载并运行该文件，遵循屏幕上的指示完成安装。

登录网址 1-9 并按照说明获取必需的工具和库。例如，访问网址 1-10，根据指示下载并安装 gfortran 8.2 工具。

虽然 R 语言本身不需要，但很多扩展 R 语言功能的程序包可能需要一些额外的工具。

访问网址 1-11，并按照 Install Homebrew 的步骤安装 macOS 版本的 homebrew 工具。如果遇到任何访问问题，可能需要启用 root 账户，可以在终端中输入 dsenableroot，然后按 Enter/Return 键，启用 root 账户。

打开终端(可以通过搜索终端或者在启动里面查找)，输入 brew install openssl，按 Enter 键，安装 openssl 工具，以允许 R 语言安全地从互联网上下载文件和程序包。

打开终端(可以通过搜索终端或者在启动里面查找)，输入 brew install libgit2，按 Enter 键，安装绘图包所需要的 libgit2 工具。

最后，访问网址 1-12，单击 Download R for (Mac) OS X，下载 4.0.2 版本的 R 语言。下载完成后，确保将它安装到计算机的应用程序中。

1.4 下载并安装 RStudio 软件

首先，恭喜你完成了 R 语言的安装，但还有一个软件需要安装。尽管 R 语言本身很强大，但它的工作环境很苛刻，通常都是在集成开发环境(Integrated Development Environment，IDE)中进行编程。在本书的案例中，需要安装 RStudio Desktop 软件以添加便于观察的视觉效果，辅助我们更好地“看到”R 语言反馈的结果。RStudio 软件使人更容易专注于统计学的学习。

访问网址 1-13，并选择 RStudio Desktop Free 选项。同之前一样，请自行选择与操作系统相符的版本。

1.4.1 Windows

访问网址 1-13，选择 RStudio Desktop Free 选项，之后就可以开始安装 Windows 版的 RStudio Desktop 软件。撰写本书时，最新的版本为 1.3.1056.exe，在网址 1-14 上

的第二步“Download RStudio Desktop”中可下载最新版本。

与 R 语言的安装一样，仅需单击下载按钮，并记住网络浏览器保存文件的位置，然后双击运行 textttRStudio-1.3.1056.exe 文件进行安装。安装过程中需要接受所有默认选项，并根据需要单击 Next 或 OK 按钮。

以上程序安装需要计算机是 64 位操作系统，如果不是，则需要访问网址 1-15 下载并安装 RStudio 的旧版本。

1.4.2　macOS

访问网址 1-14，并选择 RStudio Desktop Free 选项，之后就可以开始安装 macOS 版的 RStudio Desktop 软件。撰写本书时，最新的版本为 1.3.1056.dmg。此外，还要确保 R.app 和 RStudio.app 能够访问所需的磁盘资源。

请按照网址 1-16 的指南要求，为程序运行提供必要的权限。

1.5　RStudio 的使用方法

现在，已经完成了 R 语言和 RStudio 软件的安装，可以第一次打开并运行 RStudio 软件。单击 RStudio 图标运行该软件，可以看到，程序窗口的一大部分由顶部的一小条图标和三个大窗格构成。左边的大窗格被称为 Console，在该窗格中有文本显示“R version 4.0.2”(或者其他刚下载安装的 R 语言版本)。在右侧，有两个较小的窗格，上方的窗格是 Environment 窗格，它应该暂时没有内容。下方的窗格是 Files/Plots/Packages/Help/Viewer 窗格，它展示的是文件目录。

第一步(仅需执行一次)需要设置一些默认选项，确保软件设置与本书相同。

在顶部的菜单功能区上找到并单击 Tools 选项，在下拉菜单中选择 Global Options，之后将会弹出 Options 菜单，此时应该已经打开了如图 1-1 所示的 General 选项卡，需要确认 General 选项卡的 Basic 栏中的以下几个选择没有被选定：

- Restore most recently opened project at startup(启动时恢复最近打开的项目)：未选中
- Restore previously open source documents at startup(启动时恢复以前的开源文档)：未选中
- Restore.RData into workspace at startup(启动时将.RData 恢复到工作区)：未选中
- Save workspace to .RData on exit(退出时将工作区保存到.Rdata)：Never

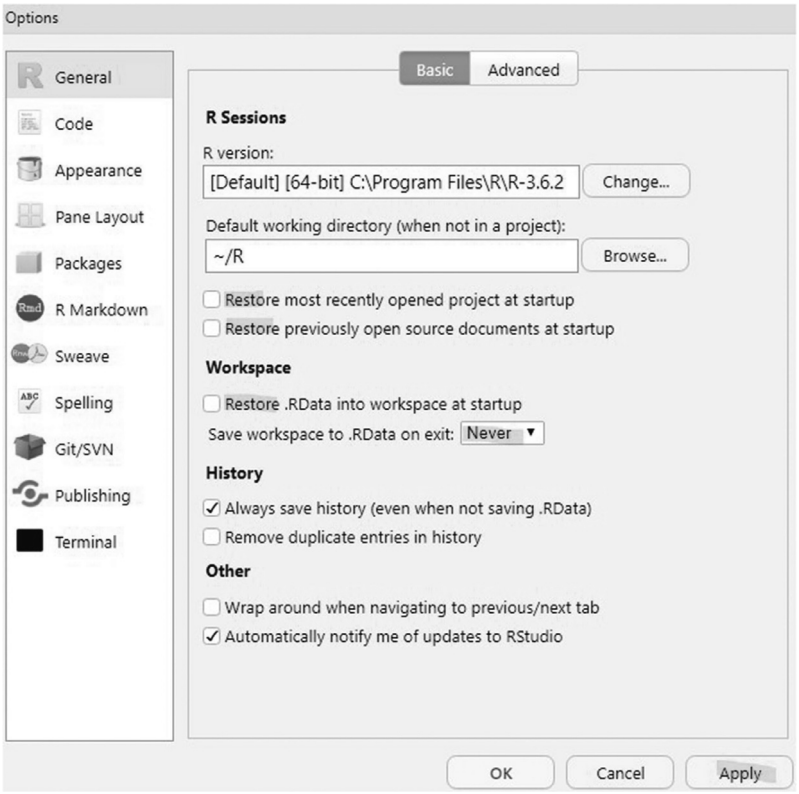

图 1-1　RStudio 软件 Options 菜单中的 General 选项卡

设置完成后，单击 Apply 按钮。

在关闭 Options 菜单之前，为了增添学习过程的趣味，请单击图 1-2 中的 Appearance 选项卡，然后选择以下选项：

- RStudio theme(RStudio 主题)：Modern
- Zoom(缩放)：根据个人喜好设置
- Editer font(编辑器字体)：根据个人喜好设置
- Editor font size(编辑器字号大小)：较大的字号让阅读更清晰，较小的字号则可以优化显示器空间
- Editor theme(编辑器主题)：Vibrant Ink(可根据个人喜好设置)

最后，单击 Pane Layout 选项卡，确保自己的选项与图 1-3 中的设置相匹配。

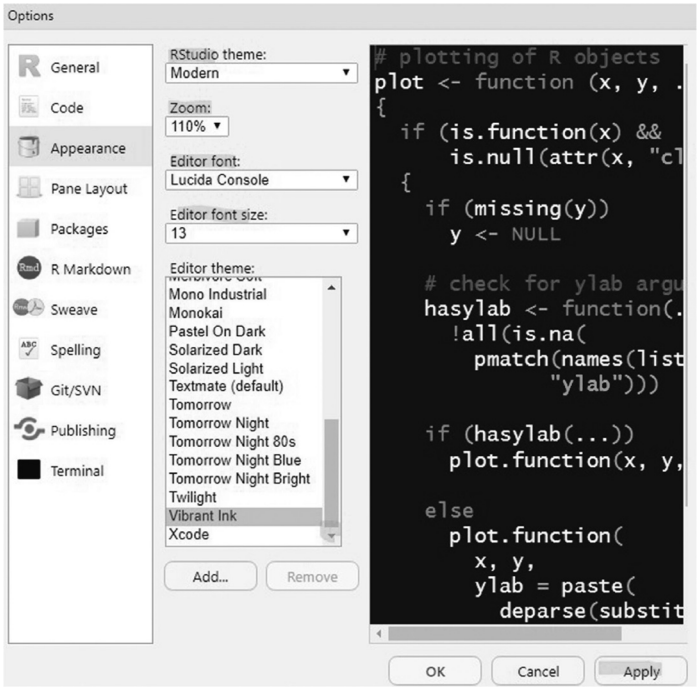

图 1-2　RStudio 软件 Options 菜单中的 Appearance 选项卡

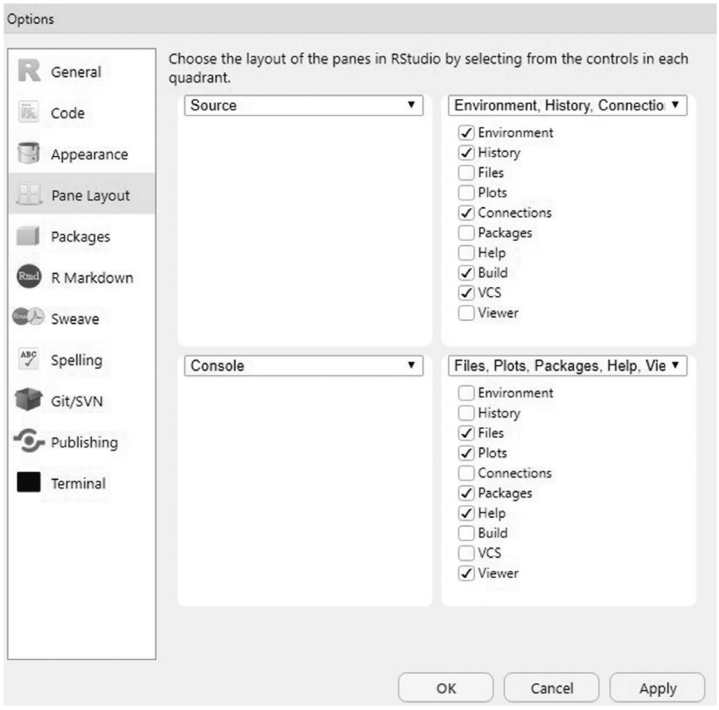

图 1-3　RStudio 软件 Options 菜单中的 Pane Layout 选项卡

到现在为止，所有的设置已经完成了。请单击 OK 按钮，为此后的学习内容做好准备。

新项目

到目前为止，RStudio 软件的默认设置已完成，可以开始构建第一个项目。为项目创建一个文件夹，用于保存已经完成的工作，文件夹会包含一些 RStudio 软件特有的文件，这些文件可以为处理错综复杂的构想提供便利。对于一个项目中应该包含多少 R 语言文件，没有明确的答案。如果把本书作为课程的一部分，那么为每一章的学习内容单独建立一个项目是很有意义的。在这种情况下，项目名称最好为章节的标题。此外，整本书中的代码是在一个名为 BeginningR_2020 的项目中编写的。

现在，需要为每章构建一个项目，以确保能够轻松地启动、关闭和打开项目。

为启动一个新项目，首先在左上角菜单栏上选择 File，选择 New Project 选项。然后使用 New Project 向导，选择 New Directory 选项卡中的 New Project 栏，填写 Directory name，此处建议使用 01Installing。一旦完成 Directory name 的填写，就可以单击 Create Project 按钮，创建新项目。

此刻你已经进入第一个项目中了！

首要任务是创建一个新的、空白的 R 语言文件。在顶部功能区的 File 栏下方有一个白纸上方带加号的小图标，单击这个图标，从新文件的列表中选择第一个名为 R Script 的文件。这个空白文件现在是可见的，在左上角的窗格中可看到一个标题为 Untitled 1 的选项卡。之后需要单击光盘形状的 save 图标，将这个文件命名为 MyFirstRScript.R 并保存。完成操作后显示的内容如图 1-4 所示。

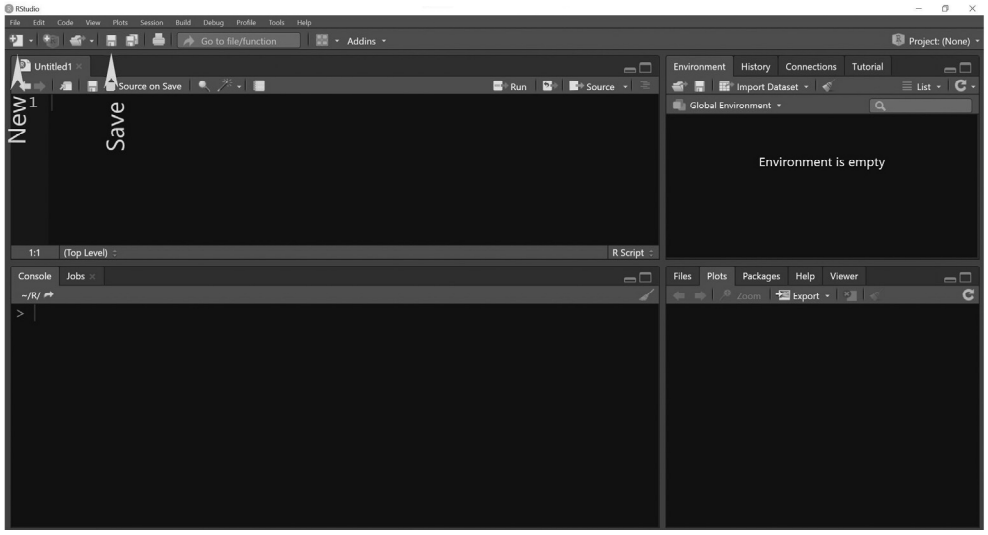

图 1-4　RStudio 软件的新建与保存示例

如果之前的步骤没有差错，那么现在显示器上应该有四个窗格。左上角的窗格名为 MyFirstRScript.R，它是一个脚本或代码窗格，将在这里面进行大部分的工作。右上角是 Environment 窗格，它仍然是空的。右下角是 Files/Plots/Packages/Help/Viewer 窗格，现在 File 栏的 project directory 选项里面应该包含三个文件，分别是.Rhistory、01Installing.Rproj 和 MyFirstRScript.R。最后，左下角是 Console 窗格，当前应该显示 R 语言的版本信息。以上是发生改变的内容，接下来将依次简要讨论每个窗格的用途。

当前保存 MyFirstRScript.R 文件的脚本窗格是输入 R 语言代码的地方，可以把它想象成 Word 文档或者 PowerPoint 幻灯片。该窗格会存储已经编写的代码，且在激活或运行代码之前，什么都不会发生。这个脚本区将成为程序区，但现在这个程序是空白的。接下来很快就会学习如何使用这个区域，这意味着你即将通过 R 语言第一次编写程序并成为一名程序员。

Environment 窗格显示程序存储在内存中的数据、对象或者变量。现在这里什么都没有显示，但本书之后的部分程序会在此窗格中显示一些数据。

Files 选项卡显示了项目的工作目录，用于此项目的任何文件都需要位于这个目录中，这一点很重要。R 语言需要知道用于获取统计分析信息的 Excel 文件或者数据文件的位置，这些文件都需要位于目录文件夹中。但是，这个窗格不仅仅展示了 Files 选项卡，Plots、Packages、Help 和 Viewer 的选项卡也存在于这个区域中。可以依次单击这些选项卡，其中，Plots、Files 和 Help 是最常用的选项卡。之后返回到 Files 选项卡。

最后是 Console 窗格。这里是运行代码的地方。RStudio 软件有两个存在代码的地方，一个是存储代码的脚本区，一个是运行代码的控制台。

现在已经结束了即将用于学习统计学的四个区域的介绍，在结束本章之前，还要确保一切都可以正常运行，因此现在需要开始第一次程序编写。

1.6　R 语言脚本的编写

对于编写的第一个程序，在确保 R 语言正常工作的同时，也希望得到一些有趣的结果。本书将展示每一行代码及输出。按照脚本操作的时候，需要将代码输入脚本区域，并用鼠标单击拖动(或者按住 Shift 键并使用方向键突出显示代码)想要运行的代码，然后在 Windows 操作系统中按 Ctrl+Enter 键，或者在 macOS 操作系统中按 Cmd+Enter 键运行代码。

需要注意，使用常规的 Ctrl+Enter 键时，需要按住 Ctrl 键，然后按下 Enter 键，最后同时松开两个键。同样，在 macOS 运行时，需要按住 Cmd 键，然后按下 Enter/Return 键，最后同时松开两个键。这是从脚本区域运行代码的最简单的方法，与每次都用鼠标单击运行相比，可以节省大量的时间。

统计学中，经常将同一类型对象的两个部分之间的联系或关系定为研究对象，例如，随着车辆重量的增加，每加仑英里数(一种燃料效率的衡量标准)会减少。为了研究怎样使用 R 语言来探索这种物理概念，首先需要简单观察第一个数据集。

mtcars 数据集源于一本很久之前的汽车杂志。在脚本区域输入 mtcars 文本，并突出显示此文本，之后按下 Ctrl+Enter 键，以下 32 行 11 列的数据就会显示在屏幕上。

```
mtcars
##                   mpg cyl disp  hp drat  wt qsec vs am gear carb
## Mazda RX4          21   6  160 110  3.9 2.6   16  0  1    4    4
## Mazda RX4 Wag      21   6  160 110  3.9 2.9   17  0  1    4    4
## Datsun 710         23   4  108  93  3.8 2.3   19  1  1    4    1
## Hornet 4 Drive     21   6  258 110  3.1 3.2   19  1  0    3    1
## Hornet Sportabout  19   8  360 175  3.1 3.4   17  0  0    3    2
## Valiant            18   6  225 105  2.8 3.5   20  1  0    3    1
## Duster 360         14   8  360 245  3.2 3.6   16  0  0    3    4
## Merc 240D          24   4  147  62  3.7 3.2   20  1  0    4    2
## Merc 230           23   4  141  95  3.9 3.1   23  1  0    4    2
## Merc 280           19   6  168 123  3.9 3.4   18  1  0    4    4
## Merc 280C          18   6  168 123  3.9 3.4   19  1  0    4    4
## Merc 450SE         16   8  276 180  3.1 4.1   17  0  0    3    3
## Merc 450SL         17   8  276 180  3.1 3.7   18  0  0    3    3
## Merc 450SLC        15   8  276 180  3.1 3.8   18  0  0    3    3
## Cadillac Fleetwood 10   8  472 205  2.9 5.2   18  0  0    3    4
## Lincoln Continental 10  8  460 215  3.0 5.4   18  0  0    3    4
## Chrysler Imperial  15   8  440 230  3.2 5.3   17  0  0    3    4
## Fiat 128           32   4   79  66  4.1 2.2   19  1  1    4    1
## Honda Civic        30   4   76  52  4.9 1.6   19  1  1    4    2
## Toyota Corolla     34   4   71  65  4.2 1.8   20  1  1    4    1
## Toyota Corona      22   4  120  97  3.7 2.5   20  1  0    3    1
## Dodge Challenger   16   8  318 150  2.8 3.5   17  0  0    3    2
## AMC Javelin        15   8  304 150  3.1 3.4   17  0  0    3    2
## Camaro Z28         13   8  350 245  3.7 3.8   15  0  0    3    4
## Pontiac Firebird   19   8  400 175  3.1 3.8   17  0  0    3    2
## Fiat X1-9          27   4   79  66  4.1 1.9   19  1  1    4    1
## Porsche 914-2      26   4  120  91  4.4 2.1   17  0  1    5    2
## Lotus Europa       30   4   95 113  3.8 1.5   17  1  1    5    2
## Ford Pantera L     16   8  351 264  4.2 3.2   14  0  1    5    4
## Ferrari Dino       20   6  145 175  3.6 2.8   16  0  1    5    6
## Maserati Bora      15   8  301 335  3.5 3.6   15  0  1    5    8
## Volvo 142E         21   4  121 109  4.1 2.8   19  1  1    4    2
```

这是一个很好的原始数据示例，每辆车都有很多信息，且同类信息都位于同一列。此时，需要运用计算机的强大功能来处理数据。本章将投入大量时间来练习 R 语言，

而不是马上学习统计学知识。此刻，可以通过将这些数字输入计算器中，以检验本章的学习成果。但是，你能想象将所有这些数字输入计算器吗？

我们不需要示例中的全部内容。事实上，此案例只需要讨论车辆重量(wt)和每加仑英里数(mpg)之间的关系，不必使用所有的信息。在 R 语言中，可以使用“$”运算符进入数据列来实现一次只查看一组数字。接下来，在一行中输入 mtcars$wt，按 Enter 键移到下一行，再输入 mtcars$mpg，并将两行突出显示，最后，通过按 Ctrl+Enter 键(macOS 的 Cmd+Enter 键)运行代码。此时应得到如下结果：

```
mtcars$wt
##  [1] 2.6 2.9 2.3 3.2 3.4 3.5 3.6 3.2 3.1 3.4 3.4 4.1 3.7 3.8 5.2 5.4
## [17] 5.3 2.2 1.6 1.8 2.5 3.5 3.4 3.8 3.8 1.9 2.1 1.5 3.2 2.8 3.6 2.8
mtcars$mpg
##  [1] 21 21 23 21 19 18 14 24 23 19 18 16 17 15 10 10 15 32 30 34 22 16
## [23] 15 13 19 27 26 30 16 20 15 21
```

从现在开始，本书将不再提及脚本的输入、突出显示或者按 Ctrl+Enter 键(macOS 的 Cmd+Enter 键)等操作，而是简单地讨论构想，以及如何将这些构想转化为代码，并展示这些代码及输出。如果在 R 语言中运行代码遇到了障碍，且无法继续进行，可以在网上找到一些有用的视频来获得帮助，此外，寻求教授或者导师的帮助也是不错的选择。

得到车辆重量和每加仑英里数的数据后，若想从视觉上观察这两类数据，可以编写一个运用 plot()函数的程序，使这两个数据集可视化。此时，还需要确保所有系统都与 RStudio 软件和 R 语言兼容。

在 R 语言中可以运用 plot()函数进行图像绘制。此函数将接收一个或多个输入(在本例中为车辆重量和每加仑英里数)，并提供输出(在本例中为两个数据对的图像)。如果对代数中的 x 轴和 y 轴有一定了解，就会知道，期望输出的图像如图 1-5 所示。首先需要将看到的代码精确复制到脚本窗格内，并突出显示，然后运行代码。之后，图像将在 Plots 窗格中显示，该窗格之前切换在 Files 选项卡。切记，可以在任何时间单击返回 Files 选项卡，且可以导出图像。现在，可以向全世界分享编程得到的第一幅图像了。在编程中，这种类型的第一个程序有时被称为“Hello World”(因为最初学习编程的时候图像和可视化效果还很少见，所以第一个程序就只是将 Hello World 打印到控制台——那是一个艰难的时期！)。现在，不仅可以直观地看到 20 世纪 70 年代的旧数据集，还可以将任何可以进入 R 语言的数据集可视化，编程已然是一种强大的工具。

```
plot(x = mtcars$wt, y = mtcars$mpg)
```

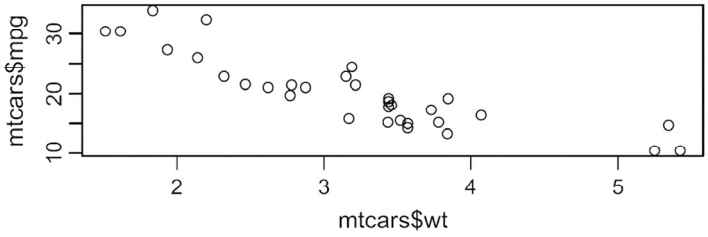

图 1-5　由 mtcars 数据库中车辆重量和每加仑英里数两组数据构成的图像

在观察图像时，有一些特性需要注意。随着 mtcars$wt 在 *x* 轴的增加，mtcars$mpg 在 *y* 轴减少。该结果符合期望：汽车重量的增加会导致燃油效率降低。在自己的系统上运行此代码得到的图像看起来会略有不同。图 1-5 中宽度大于高度，但是纵横比(图像的高度和宽度)可以更改，因此这会导致相同的数据产生看起来略有不同的图像。为了更好地对此进行理解，图 1-6 使用了相同的数据，但宽度与高度大致相等。

```
plot(x = mtcars$wt, y = mtcars$mpg)
```

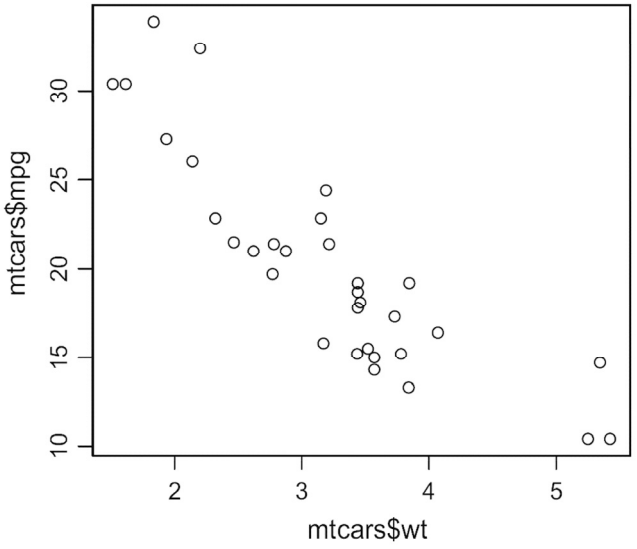

图 1-6　由 mtcars 数据库中车辆重量和每加仑英里数两组数据构成不同纵横比的图像，此时高度与宽度大致相等，而不是宽度比高度大很多

在 RStudio 软件中，把鼠标光标移到面板之间的边缘可以看到多向箭头，单击并绘制可以调整 RStudio 面板的大小和形状，此时，图像也将自动更新为相应的形状。

第 2 章将探索更多有趣的图像，本例仅是一个良好的开端。接下来，再次单击保存图标，然后选择 File 窗格，并选择位于文件菜单末尾的 Close Project 选项。关闭项

目之后，可以单击屏幕右上角的×图标关闭 RStudio 软件。

下次打开 RStudio 软件时，如果想要回到第 1 章的项目，可以先选择 File 窗格，此时在 Recent Projects 界面会有一个链接到 01Installing 项目的选项。此外，还可以使用 Open Project...导航到项目文件夹。

1.7　总结

本书将用章总结来结束每一章的学习。本章总结如表 1-2 所示，表中详细介绍了一些有助于快速参阅的条目。这些条目也可以作为指导课后练习的非常有用的工具。

表 1-2　章总结

条目	概念
R 语言	一种统计编程语言
RStudio 软件	为 R 语言提供集成开发环境(IDE)
CRAN	综合 R 档案网络
mtcars	与车辆相关的常用数据集
$	一个按名称访问列数据的 R 语言指令
plot(*x* =, *y* =)	绘制含 *x* 轴和 *y* 轴图像的函数的名称

1.8　练习与融会贯通

本节将通过做一些练习题来检查你的进步与成长。理论核查部分会提出批判性思维的问题，最好用书面方式或口头方式回答。统计学的美妙之处在于将结果成功地传达给利益相关者或者其他听众。有时这些听众非常专业，有时则不是。练习题部分则对本章探讨过的概念进行更直接的应用。

1.8.1　理论核查

1. 在 RStudio 软件的脚本窗格使用了 mtcars 数据集。如果在 RStudio 软件的 Console 窗格的提示符“¿”下面输入 mtcars，并按 Enter 键，会发生什么？将此操作与在 MyFirstRScript.R 文件中突出显示并运行 mtcars 代码的结果进行比较。

2. 在 MyFirstRScript.R 文件中可以通过 plot(*x* = mtcarswt, *y* = mtcarsmpg)代码得到图 1-6。如果运行 plot(mtcarswt, mtcarsmpg)代码，将得到什么结果？可以在 Console 窗格运行 plot(mtcarswt, mtcarsmpg)代码吗？这说明了什么？

1.8.2 练习题

1. 之前的练习中已经使用 mtcars 数据集绘制了关于车辆重量和每加仑英里数的图像。现在使用相同的 plot()函数，通过更改输入内容来创建每加仑英里数位于水平的 *x* 轴且重量位于 *y* 轴的图像。

2. R 语言中还有一个名为 iris 的数据集。运用与本章所学的处理 mtcars 数据集相同的方法，观察此数据集中 Sepal.Length 与 Petal.Length 两列数据，并创建一个 Sepal.Length 位于水平的 *x* 轴、Petal.Length 位于 *y* 轴的图像。