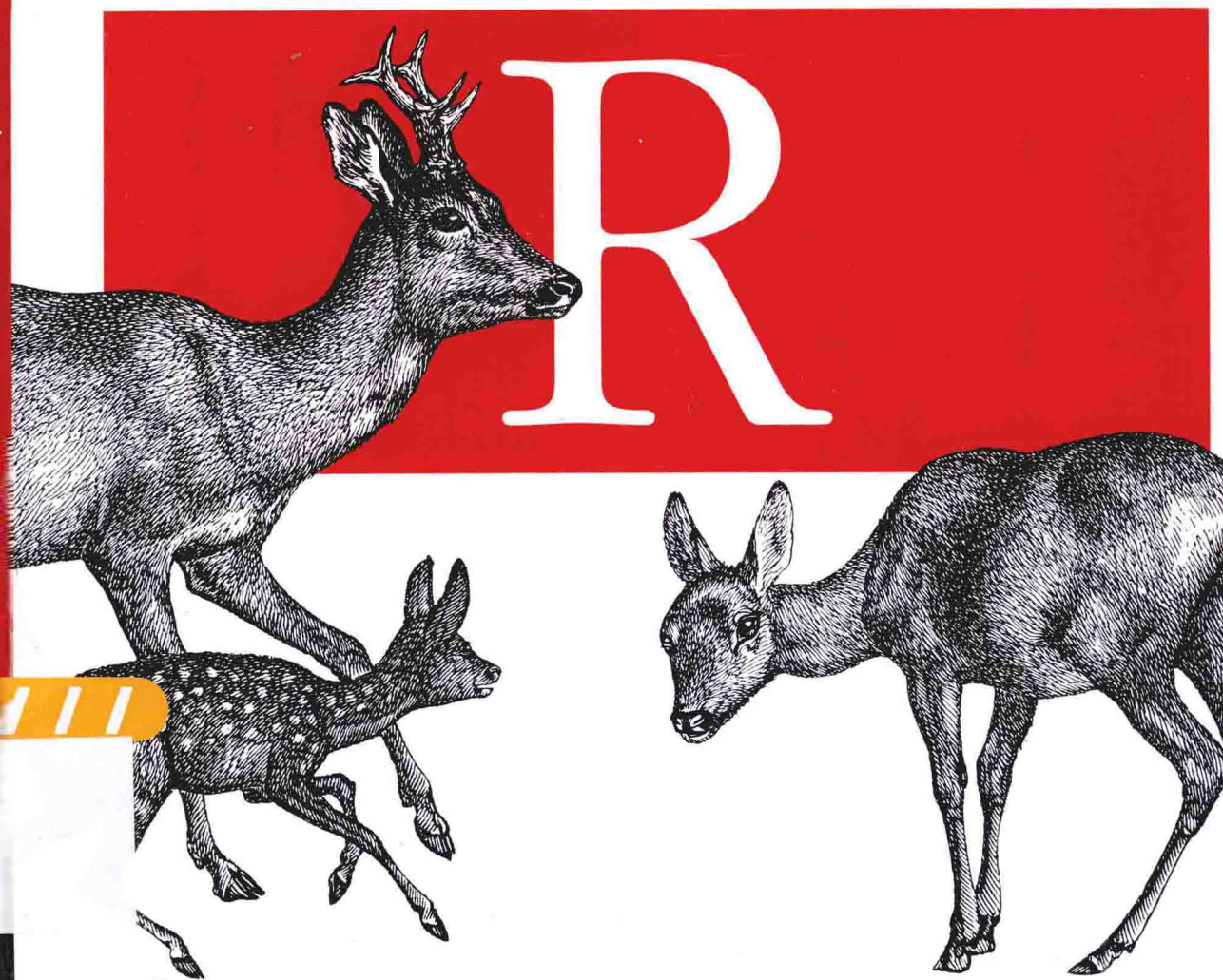


学习



[美] Richard Cotton 著

刘军 译

O'REILLY®

人民邮电出版社
POSTS & TELECOM PRESS



图灵程序设计丛书

学习R

Learning R

[美] Richard Cotton 著
刘军 译

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

O'Reilly Media, Inc. 授权人民邮电出版社出版

人民邮电出版社

北 京

图书在版编目 (C I P) 数据

学习R / (美) 柯顿 (Cotton, R.) 著 ; 刘军译. --
北京 : 人民邮电出版社, 2014. 6
(图灵程序设计丛书)
ISBN 978-7-115-35170-8

I. ①学… II. ①柯… ②刘… III. ①程序语言—程
序设计 IV. ①TP312

中国版本图书馆CIP数据核字 (2014) 第062931号

内 容 提 要

本书分为上下两部分, 旨在指导你如何使用 R, 并提供练习的机会。上半部分主要介绍 R 的技术细节和使用技巧。每章都简要介绍了一组不同的数据类型 (例如第 4 章介绍向量、矩阵和数组) 或概念 (例如第 8 章介绍分支和循环)。下半部分更侧重实践, 展示了从输入数据到发布结果这一标准的数据分析流程。

即使你没有任何编程基础, 也能顺利阅读本书。

-
- ◆ 著 [美] Richard Cotton
 - 译 刘 军
 - 责任编辑 李松峰
 - 执行编辑 李 静 郎婷婷
 - 责任印制 焦志炜

- ◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路11号
- 邮编 100164 电子邮件 315@ptpress.com.cn
- 网址 <http://www.ptpress.com.cn>
- 三河市潮河印业有限公司印刷

- ◆ 开本: 800×1000 1/16
- 印张: 22.5
- 字数: 532千字 2014年6月第1版
- 印数: 1-4 000册 2014年6月河北第1次印刷
- 著作权合同登记号 图字: 01-2014-2005号



定价: 69.00元

读者服务热线: (010)51095186转600 印装质量热线: (010)81055316

反盗版热线: (010)81055315

广告经营许可证: 京崇工商广字第 0021 号

O'Reilly Media, Inc.介绍

O'Reilly Media 通过图书、杂志、在线服务、调查研究和会议等方式传播创新知识。自 1978 年开始，O'Reilly 一直都是前沿发展的见证者和推动者。超级极客们正在开创着未来，而我们关注真正重要的技术趋势——通过放大那些“细微的信号”来刺激社会对新科技的应用。作为技术社区中活跃的参与者，O'Reilly 的发展充满了对创新的倡导、创造和发扬光大。

O'Reilly 为软件开发人员带来革命性的“动物书”；创建第一个商业网站（GNN）；组织了影响深远的开放源代码峰会，以至于开源软件运动以此命名；创立了 Make 杂志，从而成为 DIY 革命的主要先锋；公司一如既往地通过多种形式缔结信息与人的纽带。O'Reilly 的会议和峰会集聚了众多超级极客和高瞻远瞩的商业领袖，共同描绘出开创创新产业的革命性思想。作为技术人士获取信息的选择，O'Reilly 现在还将先锋专家的知识传递给普通的计算机用户。无论是通过书籍出版，在线服务或者面授课程，每一项 O'Reilly 的产品都反映了公司不可动摇的理念——信息是激发创新的力量。

业界评论

“O'Reilly Radar 博客有口皆碑。”

——Wired

“O'Reilly 凭借一系列（真希望当初我也想到了）非凡想法建立了数百万美元的业务。”

——Business 2.0

“O'Reilly Conference 是聚集关键思想领袖的绝对典范。”

——CRN

“一本 O'Reilly 的书就代表一个有用、有前途、需要学习的主题。”

——Irish Times

“Tim 是位特立独行的商人，他不光放眼于最长远、最广阔的视野并且切实地按照 Yogi Berra 的建议去做了：‘如果你在路上遇到岔路口，走小路（岔路）。’回顾过去 Tim 似乎每一次都选择了小路，而且有几次都是一闪即逝的机会，尽管大路也不错。”

——Linux Journal

译者序

2011年由布拉德·皮特主演的影片《点球成金》描述了一个运用数据运营球队的故事：一位落魄的棒球队总经理比利·比恩（Billy Beane）与他的MBA助理拍档在球队遭遇运营危机时，利用反传统的数据分析方法，开发出特有的计算机程序分析和选择球员、战术及其组合，使得这只球队从谷底反弹，而且创下美国棒球史上的连胜记录。他们不是只靠花大价钱购买高价球星，而是通过这种方法成功地使球队咸鱼翻生——“从没有一部电影将枯燥的数据转化为如此令人愉悦的娱乐体验”，《华尔街日报》如此评价。

上面只是数据分析中的一个有趣的例子，从中也反映出数据（以及大数据）分析已经逐渐成为一门越来越受关注的学问。在中国，人们同样越来越开始重视隐藏在数据背后的强大逻辑及其商业和学术价值。在我们的生活中，从“双11”的购物狂欢，身边的饭团中出现的各种“股神”，工作中的项目管理，一直到家庭账目的管理，可以说数据分析无处不在，也越来越受到个人和企业的重视，已经有大量的数据科学家在为各大、中、小企业服务。

如何从数据中得到有趣和有价值的东西，本书正是通往这一目标的桥梁之一。本书详细介绍的开源语言R正日益成熟并受到关注，其社区和影响力也越来越大。

另外，和《华尔街日报》对以上电影的评价类似，这本书也将枯燥的数据和编程理论在某种程度上转化为更令人容易接受的体验——你可以在本书中接触到奥巴马与麦凯恩的选票大战，一只无辜的螃蟹的海底之旅，麋鹿的头盖骨，古英格兰七王国等演化过来的例子。通过这些既有趣又古怪的案例来学习R以及数据分析，难道不会更加刺激你的脑细胞从而加深记忆吗？

翻译的工作是辛苦的，且需要极度的耐心和细致，非常感谢图灵公司各位老师对我工作的支持和细心的审校。最后也非常感谢我的妻子潘玮倩在翻译时对我的理解和支持。

作为译者，由于水平和视野有限，翻译不当和错误之处在所难免。欢迎大家斧正，并提出宝贵建议。希望我们所做的工作能对你有帮助。祝阅读愉快，学有所成！

刘 军

2014年2月

前言

R 是一种编程语言，也是用于数据分析和统计的软件环境。它是一个 GNU 项目，这意味着它是自由的开源软件。它正在以指数级的速度不断成长——普遍认为，它的用户人数可能超过了 100 万，它有 4000 多个由开发社区贡献的附件包，而且每年以约 25% 的速度增加。在本书创作之时，它在 Tiobe 编程社区指数 (Tiobe Programming Community Index) 的开发语言流行榜上已排至第 24 位，大致与 SAS 和 MATLAB 看齐。

R 广泛地应用在每一个需要统计或数据分析的领域，涵盖了金融、市场营销、医药、基因组学、流行病学、社会科学、教学以及许多其他较小的领域。

关于本书

因为 R 主要用于统计分析，所以很多关于 R 的书都在指导你如何计算统计或模型数据集。然而，这些书忽视了数据分析应用的实际情况。事实上，除非你做的是尖端研究，否则你所用到的统计技术往往只需用于常规任务，而且你的模型可能也不大。完整的数据分析流程更像是这样：

- (1) 取得一些数据；
- (2) 清理数据；
- (3) 探索和可视化数据；
- (4) 数据建模并做出预测；
- (5) 展示或发布你的结果。

当然，每个阶段都可能碰到一些有趣的问题，以至于你需要更多的数据，或者要以不同的方式处理现有数据，这会使你的工作倒退一步。工作流是可以迭代的，但每个步骤都不可或缺。

本书的第一部分会从头开始教你 R——你不需要任何编程语言的经验。实际上，虽然完全

没有编程经验也无妨，但有一些基本的编程知识会更好。例如，本书介绍了如何注释代码以及编写 `for` 循环，但没有作更详细的解释。因此，如果你想要找本真正的编程入门课本，那么 Jason R. Briggs 写的 *Python for Kids* 非常合适！

本书的第二部分将展示 R 语言的完整数据分析流程，这里需要一些基本的统计知识。例如，你应该了解平均值和标准差等术语，以及什么是条形图（bar chart）。

本书最后将介绍 R 的一些高级主题，例如面向对象编程和包的创建。Garrett Golem 的 *Data Analysis with R* 将会在本书的基础上深入探讨数据分析流程。

一点提醒：这不是一本参考书，许多主题叙述得并不详细。本书旨在指导你如何使用 R，并提供练习的机会。显然，我们没有那么多篇幅把所有 4000 个附件包都过一遍，但当读完此书，你将有能力找到你所需要的东西，并知道如何寻求帮助以应用它们。

本书主要内容

本书分为上下部分。上半部分旨在为你讲解 R 技术细节和使用技巧。每章都简要地介绍了一组不同的数据类型（例如第 4 章介绍向量、矩阵和数组）或概念（例如第 8 章介绍分支和循环）。

下半部分会更有意思：你能看到真实的数据分析。每章分别介绍标准数据分析流程的一部分，从输入数据到发布结果。

第一部分包括如下内容。

- 第 1 章：如何安装 R 以及在哪里得到帮助。
- 第 2 章：如何将 R 作为一个科学计算器使用。
- 第 3 章：以不同的方式检查变量。
- 第 4 章：向量、矩阵和数组。
- 第 5 章：列表和数据框（类似电子表格的数据）。
- 第 6 章：环境和函数。
- 第 7 章：字符串和因子（用于类别数据）。
- 第 8 章：分支（`if` 与 `else`）和基本的循环。
- 第 9 章：使用 `apply` 函数的高级循环应用函数及其变种。
- 第 10 章：如何安装和使用附件包。
- 第 11 章：日期和时间。

第二部分的主题如下。

- 第 12 章：如何将数据导入 R。

- 第 13 章：清理和操作数据。
- 第 14 章：通过统计计算和绘图探索数据。
- 第 15 章：如何建模。
- 第 16 章：各种高级编程技术。
- 第 17 章：如何为他人打包。

最后，第三部分的附录介绍一些有用的参考资料。

- 附录 A：比较不同类型变量的属性表。
- 附录 B：可以用 R 做的其他事。
- 附录 C：每章后测验题的答案。
- 附录 D：每章后编程练习题的答案。

你应该读哪几章

如果你从未使用过 R，那么请从第 1 章开始逐章阅读。如果你已经有了一些 R 的经验，不妨跳过第 1 章，快速浏览关于 R 语言的核心章节。

尽管章节间有一些关联，但因为每章都涉及一个不同的主题，所以你可以随机挑选感兴趣的章节阅读。

最近，我与 *R For Dummies* 的作者 Andrie de Vries 谈及这个问题，他竟然建议放弃此书，转而阅读他的著作¹！

本书排版约定

本书中使用以下排版约定。

- 楷体
表示新的术语。
- 等宽字体
用于需逐字复制的代码示例，以及段内引用的编程元素，如变量或函数名、数据类型、环境变量、语句（statement）和关键字。代码块的输出也使用等宽字体，前面加上两个井号（##）。
- 斜等宽字体
显示将由用户提供的值（或由上下文确定的值）替换的文本。

注 1：Andrie 的书所涵盖的内容和本书基本一致，并在许多方面和本书一样好。如果你也想阅读它，我对此表示欢迎。

本书中所用代码的风格指南：<http://4dpiecharts.com/r-code-style-guide>。



此图标表示小技巧、建议或一般说明。



此图标表示警告或小心。

目标、小结、小测验和练习

每章开头列出的目标会告诉你本章的内容概要，结束时的小结会重述已学的知识点，并会有个小测验以确保你是在集中精力地学习（而不是在看电视时假装看书）。测验答案可在本章内容中找到（或见本书的结尾部分，如果你想偷看的话）。每章最后还会有一些练习，大多是要求你写一些 R 的代码。每个练习题后面的方括号中有一个数字，它表示完成此题所需的大致时间。

使用代码

补充材料（代码范例、练习等）可以在此下载：<http://cran.r-project.org/web/packages/learningr>。

本书就是要帮读者解决实际问题的。也许你需要在自己的程序或文档中用到本书中的代码。除非大段大段地使用，否则不必与我们联系取得授权。因此，用本书中的几段代码写成一个程序不用向我们申请许可。但是销售或者分发 O'Reilly 图书随附的代码光盘则必须事先获得授权。引用书中的代码来回答问题也无需我们授权。将大段的示例代码整合到你自己的产品文档中则必须经过许可。

使用我们的代码时，希望你能标明它的出处。出处一般要包含书名、作者、出版商和书号，例如：*Learning R* by Richard Cotton (O'Reilly). Copyright 2013 Richard Cotton, 978-1-449-35710-8。

如果还有其他使用代码的情形需要与我们沟通，可以随时与我们联系：permissions@oreilly.com。

Safari® Books Online



Safari Books Online (www.safaribooksonline.com) 是应需而变的数字图书馆。它同时以图书和视频的形式出版世界顶级技术和商务作家的专业作品。

Safari Books Online 是技术专家、软件开发人员、Web 设计师、商务人士和创意人士开展调研、解决问题、学习和认证培训的第一手资料。

对于组织团体、政府机构和个人，Safari Books Online 提供各种产品组合和灵活的定价策略。用户可通过一个功能完备的数据库检索系统访问 O'Reilly Media、Prentice Hall Professional、Addison-Wesley Professional、Microsoft Press、Sams、Que、Peachpit Press、Focal Press、Cisco Press、John Wiley & Sons、Syngress、Morgan Kaufmann、IBM Redbooks、Packt、Adobe Press、FT Press、Apress、Manning、New Riders、McGraw-Hill、Jones & Bartlett、Course Technology 以及其他几十家出版社的上千种图书、培训视频和正式出版之前的书稿。要了解 Safari Books Online 的更多信息，我们网上见。

联系我们

请把对本书的评价和问题发给出版社。

美国：

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472

中国：

北京市西城区西直门南大街 2 号成铭大厦 C 座 807 室（100035）
奥莱利技术咨询（北京）有限公司

O'Reilly 的每一本书都有专属网页，你可以在那儿找到本书的相关信息，包括勘误表、示例代码以及其他信息。本书的网站地址是：

<http://oreil.ly.learningR>

对于本书的评论和技术性问题，请发送电子邮件到：

bookquestions@oreilly.com

要了解更多 O'Reilly 图书、培训课程、会议和新闻的信息，请访问以下网站：

<http://www.oreilly.com>

我们在 Facebook 的地址如下：

<http://facebook.com/oreilly>

请关注我们的 Twitter 动态：

<http://twitter.com/oreillymedia>

我们的 YouTube 视频地址如下：

<http://www.youtube.com/oreillymedia>

致谢

本书优秀的编辑 Meghan Blanchette 给我提出了许多明智的建议。此外，很多才华横溢的人士为此书的完成提供了帮助。

数据由以下几位杰出的专业人士贡献。

- AMD 的 Bill Hogan 找到并清理了 Alpe d'Huez 循环数据集，还帮助我找到了 CDC 的淋病数据集。女士们，他特别要我强调他很健康。
- CEFAS 的 Ewan Hunter 提供北海蟹的数据集。
- 剑桥大学的 Corina Logan 整理并提供了鹿头骨的数据。
- 莱顿大学的 Edwin Thoen 整理并提供了奥巴马对战麦凯恩的数据集。
- Gwern Branwen 观看和阅读了数量惊人的日本动漫并整理了混血儿（hafu）数据集。赞一个！

还有另外很多人也为我提供了数据集，由于篇幅有限不能一一列出，在此表示感谢！

Bill Hogan、Marin 软件的 Daisy Vincent 和 JD Long 审阅了本书。我不知道 JD 在哪里工作，但他住在百慕大，有可能在三角带吧。其他提供意见和反馈的人包括：James White、Ben Hanks、Beccy Smith、TDX 集团的 Guy Bourne；HSL 的 Alex Hogg 和 Adrian Kelsey；Tom Hull、Karen Vanstaen、Rachel Beckett、Georgina Rimmer、Ruth Wortham、Bernardo Garcia-Carreras 以及 CEFAS 的 Joana Silva；特拉维夫大学的 Tal Galili；RStudio 的 Garrett Grolemond；纽约市立大学的 John Verzani。CEFAS 的 David Maxwell 出色地召集了几乎所有的 CEFAS 人士审阅了我的书。

非常感谢 John Verzani 帮助酝酿本书，并提供了结构方面的意见。

当我在为手稿的字符编码抓狂时，O'Reilly 的 Sanders Kleinfeld 提供了重要的技术支持。Yihui Xie 完全义务地帮我使用 knitr 生成 AsciiDoc。Rachel Head 在审稿时单枪匹马地发现了 4000 多个 bug、排印错误和其他错误。

Garib Murshudov 是第一个教我 R 的讲师，那得追溯到 2004 年。

最后，非常感谢 Janette Bowler 在我写作过程中给予的无比耐心和支持。

目录

译者序	XII
前言	XIII

第一部分 R 语言

第 1 章 简介	2
1.1 本章目标	2
1.2 R 是什么	2
1.3 安装 R	3
1.4 选择一个 IDE	4
1.4.1 Emacs+ESS	4
1.4.2 Eclipse/Architect	4
1.4.3 RStudio	5
1.4.4 Revolution-R	5
1.4.5 Live-R	5
1.4.6 其他 IDE 和编辑器	6
1.5 你的第一个程序	6
1.6 如何从 R 中获得帮助	7
1.7 安装其他相关软件	9
1.8 小结	9

1.9 知识测试：问题	10
1.10 知识测试：练习	10
第2章 科学计算器	11
2.1 本章目标	11
2.2 数学运算符和向量	11
2.3 变量赋值	15
2.4 特殊数字	17
2.5 逻辑向量	18
2.6 小结	20
2.7 知识测试：问题	20
2.8 知识测试：练习	20
第3章 检查变量和工作区	22
3.1 本章目标	22
3.2 类	22
3.3 不同类型的数字	23
3.4 其他通用类	24
3.5 检查和更改类	26
3.6 检查变量	29
3.7 工作区	32
3.8 小结	33
3.9 知识测试：问题	33
3.10 知识测试：练习	34
第4章 向量、矩阵和数组	35
4.1 本章目标	35
4.2 向量	35
4.2.1 序列	37
4.2.2 长度	38
4.2.3 命名	39
4.2.4 索引向量	39
4.2.5 向量循环和重复	41
4.3 矩阵和数组	43
4.3.1 创建数组和矩阵	43
4.3.2 行、列和维度	45
4.3.3 行名、列名和维度名	46
4.3.4 索引数组	47
4.3.5 合并矩阵	47

4.3.6 数组算术	48
4.4 小结	50
4.5 知识测试：问题	50
4.6 知识测试：练习	51
第 5 章 列表和数据框	52
5.1 本章目标	52
5.2 列表	52
5.2.1 创建列表	52
5.2.2 原子变量和递归变量	54
5.2.3 列表的维度和算术运算	55
5.2.4 索引列表	56
5.2.5 向量和列表之间的转换	58
5.2.6 组合列表	60
5.3 NULL	61
5.4 成对列表	64
5.5 数据框	65
5.5.1 创建数据框	65
5.5.2 索引数据框	68
5.5.3 基本数据框操作	69
5.6 小结	71
5.7 知识测试：问题	71
5.8 知识测试：练习	72
第 6 章 环境和函数	73
6.1 本章目标	73
6.2 环境	73
6.3 函数	76
6.3.1 创建和调用函数	76
6.3.2 向其他函数传递和接收函数	80
6.3.3 变量的作用域	82
6.4 小结	84
6.5 知识测试：问题	84
6.6 知识测试：练习	84
第 7 章 字符串和因子	86
7.1 本章目标	86
7.2 字符串	86
7.2.1 创建和打印字符串	87

7.2.2	格式化数字	88
7.2.3	特殊字符	90
7.2.4	更改大小写	91
7.2.5	截取字符串	91
7.2.6	分割字符串	92
7.2.7	文件路径	93
7.3	因子	94
7.3.1	创建因子	94
7.3.2	更改因子水平	96
7.3.3	去掉因子水平	96
7.3.4	有序因子	97
7.3.5	将连续变量转换为类别	98
7.3.6	将类别变量转换为连续变量	99
7.3.7	生成因子水平	100
7.3.8	合并因子	100
7.4	小结	101
7.5	知识测试：问题	101
7.6	知识测试：练习	101
第 8 章	流程控制和循环	103
8.1	本章目标	103
8.2	流程控制	103
8.2.1	if 和 else	103
8.2.2	向量化的 if	105
8.2.3	多个分支	106
8.3	循环	108
8.3.1	重复循环	108
8.3.2	while 循环	110
8.3.3	for 循环	111
8.4	小结	113
8.5	知识测试：问题	113
8.6	知识测试：练习	113
第 9 章	高级循环	115
9.1	本章目标	115
9.2	replication	115
9.3	遍历列表	116
9.4	遍历数组	122

9.5 多个输入的应用函数	124
9.6 拆分 – 应用 – 合并 (Split-Apply-Combine)	126
9.7 plyr 包	127
9.8 小结	130
9.9 知识测验：问题	130
9.10 知识测试：练习	131
第 10 章 包	133
10.1 本章目标	133
10.2 加载包	133
10.2.1 搜索路径	135
10.2.2 库和已安装的包	136
10.3 安装包	137
10.4 维护包	139
10.5 小结	139
10.6 知识测试：问题	140
10.7 知识测试：练习	140
第 11 章 日期和时间	141
11.1 本章目标	141
11.2 日期和时间类	141
11.2.1 POSIX 日期和时间	142
11.2.2 Date 类	143
11.2.3 其他日期类	144
11.3 日期与字符串的相互转换	144
11.3.1 解析日期	144
11.3.2 格式化日期	145
11.4 时区	145
11.5 日期和时间的算术运算	147
11.6 lubridate	149
11.7 小结	153
11.8 知识测试：问题	153
11.9 知识测试：练习	153

第二部分 数据分析 workflow

第 12 章 获取数据	156
12.1 本章目标	156

12.2	内置的数据集	156
12.3	读取文本文件	157
12.3.1	CSV 和制表符分隔 (Tab-Delimited) 文件	157
12.3.2	非结构化文本文件	161
12.3.3	XML 和 HTML 文件	161
12.3.4	JSON 和 YAML 文件	163
12.4	读取二进制文件	165
12.4.1	读取 Excel 文件	165
12.4.2	读取 SAS、Stata、SPSS 和 MATLAB 文件	167
12.4.3	读取其他文件类型	167
12.5	Web 数据	168
12.5.1	拥有 API 的网站	168
12.5.2	抓取网页	169
12.6	访问数据库	171
12.7	小结	174
12.8	知识测试：问题	174
12.9	知识测试：练习	174
第 13 章	数据清理和转换	175
13.1	本章目标	175
13.2	清理字符串	175
13.3	操作数据框	180
13.3.1	添加和替换列	180
13.3.2	处理缺失值	181
13.3.3	在宽和长表格之间进行转换	182
13.3.4	使用 SQL	183
13.4	排序	184
13.5	函数式编程	185
13.6	小结	188
13.7	知识测试：问题	188
13.8	知识测试：练习	189
第 14 章	探索和可视化	190
14.1	本章目标	190
14.2	汇总统计	190
14.3	三种绘图系统	194
14.4	散点图	195
14.4.1	第一种方法：base 绘图法	195