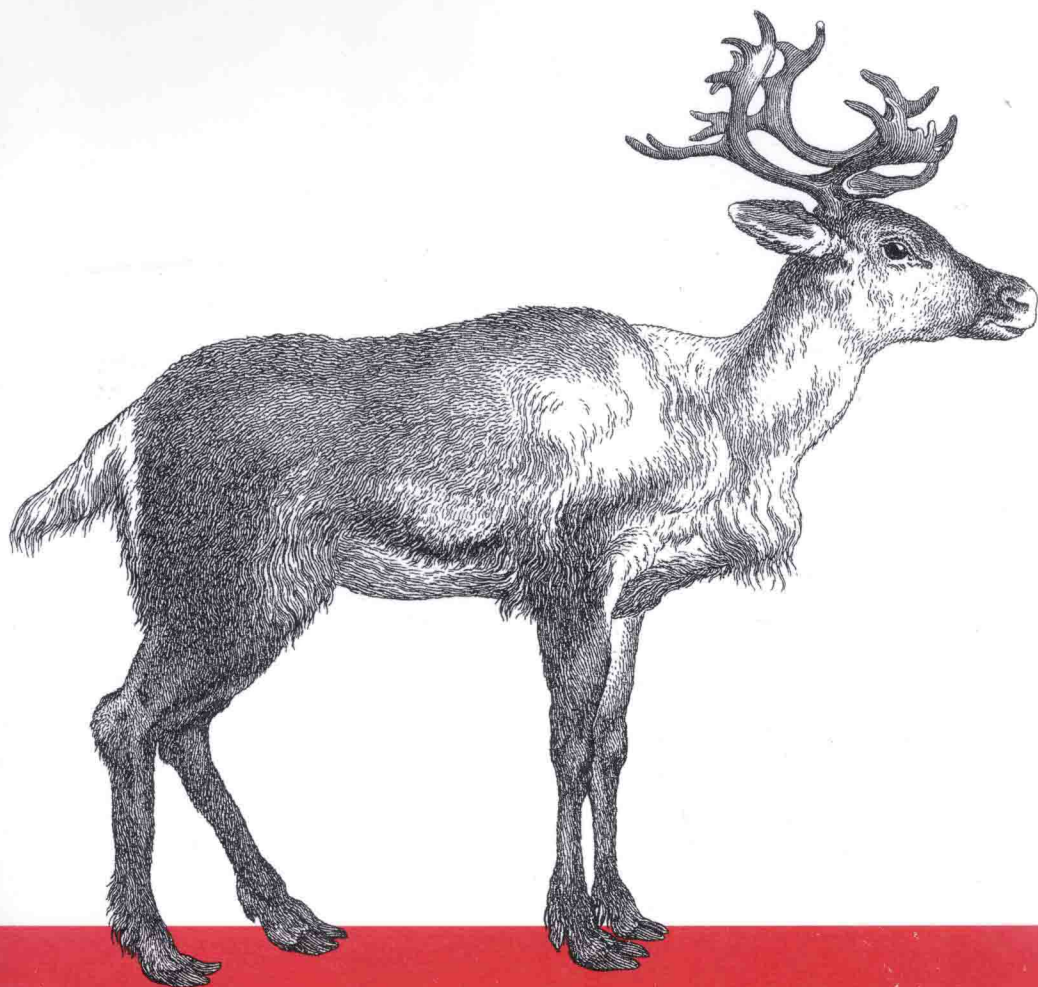


R Graphics Cookbook



R数据可视化手册

[美] *Winston Chang* 著

肖楠 邓一硕 魏太云 译

邱怡轩 审校

O'REILLY®



人民邮电出版社

POSTS & TELECOM PRESS

R 数据可视化手册

[美] *Winston Chang* 著

肖楠 邓一硕 魏太云 译

邱怡轩 审校

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 (CIP) 数据

R数据可视化手册 / (美) 常 (Chang, W.) 著 ; 肖楠, 邓一硕, 魏太云译. — 北京 : 人民邮电出版社, 2014. 5
ISBN 978-7-115-34227-0

I. ①R… II. ①常… ②肖… ③邓… ④魏… III. ①
可视化软件—手册 IV. ①TP31-62

中国版本图书馆CIP数据核字(2013)第307384号

版权声明

Copyright ©2013 by O'Reilly Media, Inc.

Simplified Chinese Edition, jointly published by O'Reilly Media, Inc. and Posts & Telecom Press, 2014. Authorized translation of the English edition, 2013 O'Reilly Media, Inc., the owner of all rights to publish and sell the same.

All rights reserved including the rights of reproduction in whole or in part in any form.

本书中文简体版由 O'Reilly Media, Inc. 授权人民邮电出版社出版。未经出版者书面许可, 对本书的任何部分不得以任何方式复制或抄袭。

版权所有, 侵权必究。

-
- ◆ 著 [美] Winston Chang
译 肖楠 邓一硕 魏太云
审校 邱怡轩
责任编辑 杨海玲
责任印制 杨林杰

- ◆ 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>
北京精彩雅恒印刷有限公司印刷

- ◆ 开本: 800×1000 1/16

印张: 21

字数: 428 千字

印数: 1—4 000 册

2014年5月第1版

2014年5月北京第1次印刷

著作权合同登记号 图字: 01-2013-3677 号



定价: 89.00 元

读者服务热线: (010)81055410 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

内容提要

R 具有强大的统计计算功能和便捷的数据可视化系统。本书重点讲解 R 的绘图系统，指导读者通过绘图系统实现数据可视化。

书中提供了快速绘制高质量图形的 150 多种技巧，每个技巧用来解决一个特定的绘图需求。读者可以通过目录快速定位到自己遇到的问题，查阅相应的解决方案。同时，作者在大部分的技巧之后会进行一些讨论和延伸，介绍一些总结出的绘图技巧。

本书侧重于解决具体问题，是 R 数据可视化的实战秘籍。本书中绝大多数的绘图案例都是以强大、灵活制图而著称的 R 包 `ggplot2` 实现的，充分展现了 `ggplot2` 生动、翔实的一面。从如何画点图、线图、柱状图，到如何添加注解、修改坐标轴和图例，再到分面的使用和颜色的选取等，本书都有清晰的讲解。虽然本书的大多数技巧使用的是 `ggplot2`，但是并不仅仅局限于 `ggplot2` 的介绍。作者的理念是用合适的工具来完成合适的绘图任务，读者也可以学到许多其他有用的绘图函数和工具，来适应各种复杂的需求。

本书是学习 R 中丰富的数据可视化方法的权威手册，非常适合对 R 语言有基本的了解的读者阅读。

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

R 官方网站的第一句话是这样介绍 R 语言的：“R 是一个用于统计计算和绘图自由软件环境。”这句话突出了 R 的两大特色：强大的统计计算功能和便捷的数据可视化系统。在很多情况下，当我们将 R 与其他同类的语言、软件进行比较时，通常都会强调其灵活的编程和计算能力，然而事实上，R 的绘图系统也是其最大的优势之一。

经过长年的开发和完善，目前 R 主要支持了四套图形系统：基础图形（base）、网格图形（grid）、lattice 图形和 ggplot2。其中前三个都内置于 R 的发行包，其功能已经非常稳定，而最为年轻的 ggplot2（最早开发于 2005 年）在历经若干次重大更新后，也逐渐成为了 R 中数据可视化的主流选择。由于 ggplot2 具有强大的语法特性和优雅的图形外观，它迅速吸引了众多的使用者和开发者，并已经开始被移植到其他语言中，如 Python、Julia 等。可以说，ggplot2 与 R 中传统的图形系统形成了良好的互补：后者适合快速的数据探索性分析，而前者则可以便捷地生成复杂的、高质量的统计图形。

本书中绝大多数的绘图案例都是基于 ggplot2 实现的，在某种意义上，本书可以认为是一部优秀的 ggplot2 入门和进阶手册。正如之前所说，ggplot2 之所以强大，是因为它不仅仅是一些函数的堆砌，而是有其内在的语法支持。因为这些特点，ggplot2 看上去更像是一门新的“语言”，从而需要使用者有一个学习和熟练的过程。

我们知道，学习一门语言（中文、英语、C++、R 等），大体都要经过一个从语法到词汇，再到句法应用的过程。ggplot2 的作者 Hadley Wickham 曾写过一本介绍 ggplot2 核心思想的书籍《ggplot2: Elegant Graphics for Data Analysis》，其中介绍了 ggplot2 的基本概念和用法，相当于为读者搭起了 ggplot2 的“语法”框架，而本书则是一本丰富、有趣的“词汇和例句书”。换句话说，本书侧重于解决具体问题，是一本实战秘籍：通过它读者可以了解到 ggplot2 生动、翔实的一面。比如，从如何画点图、线图、柱状图，到如何添加注解、修改坐标轴和图例，再到分面的使用和颜色的选取等，本书都有清晰的讲解。

本书由 150 多个精选的“技巧”组成，每个“技巧”都用来解决一个特定的绘图需求。读者可以通过章节的目录快速定位到自己遇到的问题，然后查阅相应的解决方案。同时，作者在大部分的“技巧”之后会进行一些讨论和延伸，介绍一些总结出的绘图技巧。就我自己的经历而言，在阅读本书之前，有时用 ggplot2 作图会遇到一些非常细节的问题，例如，如何调整条形图的顺序，如何修改图例的外观，如何选择文字的大

小和字体？这些问题都会让我花费很多时间来搜索解决方案。而在我把这本书通读一遍之后，很多问题立刻迎刃而解。我相信，本书对于 ggplot2 和数据可视化的学习都是大有裨益的。

当然，本书并不仅仅局限于 ggplot2 的介绍。作者的理念是用合适的工具来完成合适的绘图任务，所以在本书的第 2 章、第 13 章和第 14 章，读者也可以学到许多其他有用的绘图函数和工具，来适应各种复杂的需求。

本书的三位译者为本书中文版的面世付出了大量的时间和精力，其中肖楠翻译了第 7~11 章和附录，邓一硕翻译了第 2~6 章，魏太云翻译了前言、第 1 章和第 12~15 章。三位译者均是统计之都 (<http://cos.name/>) 的成员，他们也是当前国内 R 语言社区的领军人物，曾翻译了许多 R 语言的经典教材和书籍，并通过组织中国 R 语言会议、开办 COS 数据分析沙龙、参与论坛问题讨论、撰写博客、整理文档、编写软件包等多种方式为 R 语言社区做出了诸多贡献。值得一提的是，本书 13.1 节中介绍的 `corrplot` 软件包，正是由译者之一的魏太云开发完成的。统计之都团队作为国内 R 语言最早的一批布道者之一，有幸见证了 R 语言在国内兴起的整个过程。从网站创始人谢益辉开始，再到如今众多 R 语言中文版书籍的面世，我相信其中的每一步都是国内统计学、以及更广义的数据科学不断向前发展的印记。我们也希望更多的有志之士加入到统计之都的团队之中，为国内统计学和数据科学贡献自己的力量。

本书的译者力求使翻译准确、生动，但疏漏之处在所难免，欢迎读者予以指正。为了让读者方便地获取和提交勘误信息，我们建立了本书的翻译项目页面：<https://github.com/cosname/gcookbook-translation>。读者也可以在统计之都的图书出版栏目 (<http://cos.name/books/>) 留言或提问。

本书的出版离不开众多人士的大力帮助，我们要郑重感谢爱荷华州立大学的王芯同学、中国人民大学的陈森同学和浙江大学的张政同学，他们在本书的翻译过程中提出了很多中肯的意见。没有他们的帮助，本书很难完成。此外，我们还要感谢人民邮电出版的杨海玲女士和编辑们，他们的专业精神让我叹服。

邱怡轩

2013 年 12 月于普渡大学

译者介绍

肖楠，中南大学数学与统计学院统计学系在读博士，统计之都论坛 R 语言版版主。合作翻译出版了《R 语言实战》、《ggplot2：数据分析与图形艺术》等图书，编写了 `protr`、`Rcpi` 等 R 软件包。关注领域为统计机器学习、化学信息学与生物信息学、定量与系统药理学。

邓一硕，毕业于中央财经大学统计与数学学院，统计之都论坛金融投资分析版版主，现效力于首钢总公司计财部。擅长的领域为时间序列分析以及数据挖掘在金融投资分析中的应用。

魏太云，毕业于中国人民大学统计学院，统计之都理事会主席。合作翻译出版了《ggplot2：数据分析与图形艺术》等图书，参与编写了 `corrplot`、`recharts`、`knitr` 以及 `fun` 等 R 软件包。感兴趣的主体包括统计建模、机器学习和数据可视化。

审校者介绍

邱怡轩，普渡大学统计系在读博士，统计之都理事会成员。合作翻译出版了《ggplot2：数据分析与图形艺术》、《R 语言编程艺术》等图书，参与编写了 `R2SWF`、`rARPACK`、`showtext`、`Layer`、`rationalfun`、`fun` 等 R 软件包。感兴趣的方向有函数型数据分析、统计计算和数据可视化等。

前言

几年前读研时我开始用 R，主要用来分析我在科研工作中收集到的数据。我使用 R 首先是想摆脱 SPSS 这样的统计软件的禁锢，即严格的环境和死板的分析。更何况，R 是免费的，所以我用不着说服别人为我购买一套这样的软件——这对一个穷研究生来说是相当的重要！此后，随着我对 R 的了解不断深入，我才发现原来 R 还可以绘制出非常优秀、动人的数据图形。

本书的每个“技巧”中，都列出了一个问题和对应的解决方法。在大多数情况下，我提供的并不是 R 中唯一的实现方法，但却是我认为的最佳方案。R 如此受欢迎的一个重要原因是它有很多附加的软件包，每一个软件包都为 R 提供了一些独特的功能。在 R 中也有很多可视化方面的软件包，但本书主要使用 ggplot2（声明：我现在工作的一部分就是开发 ggplot2；但是，在我还没意识到我可能会从事与 ggplot2 相关的工作之前，我已经完成了本书的大部分工作）。

本书并不想罗列五花八门的方法，成为 R 数据可视化的综合手册；但是我希望当你想绘制所需图形的时候，本书能够对你有所帮助。或者说，当你不知道怎么画的时候，翻一翻这本书或许就可以找到一些可行的方案。

方法

本书面向的读者需要对 R 至少有一些基本的了解。书中的技巧会让你明白如何解决一些特定的问题。在使用例子的时候，我力图简单明了，这样你就会明白它们的工作机理，并可以方便地把解决方法应用到自己的问题上。

软件和平台说明

书中的大部分“技巧”都是用 ggplot2 完成的，有些“技巧”需要 ggplot2 的最新版 0.9.3，这样也就要求有一个版本相对较新的 R——你可以在 R 的官方网站获取最新版本的 R。



如果你对 ggplot2 不熟悉，请参阅附录 A，那里对该包有一个简要的说明。

安装了 R 后，你可以再安装一些必要的包。除了 `ggplot2` 之外，你还可以选择安装 `gcookbook` 包（它包含了本书大多数例子的数据集）。要同时安装这两个包，只需运行命令：

```
install.packages("ggplot2")
install.packages("gcookbook")
```

你可能会被问到选择 CRAN（Comprehensive R Archive Network）镜像的问题。一般而言，任何一个镜像都可以正常工作，不过最好选择一个离你更近的，因为这样速度会更快。安装完包后，每次需要使用 `ggplot2` 包时在 R 会话中运行：

```
library(ggplot2)
```

本书中的技巧总是假设你已经加载了 `ggplot2`，所以不会显示这一行代码。

如果你看到这样的错误，意味着你忘记了加载 `ggplot2`。

错误：找不到函数 "ggplot"

英文版 R 的错误提示是：

```
Error: could not find function "ggplot"
```

R 的主要运行平台是 Mac OS X、Linux 和 Windows，本书中所有的“技巧”都可以在这些平台上运行。在保存位图输出的时候，会有平台的差异，详情参见第 14 章。

本书的排版约定

本书采用的体例如下：

- 等宽字体 (Constant width)：表示程序清单，以及段落中引用的程序元素，如变量或函数名、数据库、数据类型、环境变量、语句和关键字。
- 加粗的等宽字体 (**Constant width bold**)：表示需要用户手动输入的命令或其他文本。
- 斜的等宽字体 (*Constant width italic*)：表示应该用用户所提供的值或根据上下文确定的值来替换的文本。



这个图标表示一个提示、建议或者一般的注记。

代码示例的使用

本书的目标是帮助你完成工作。一般而言，你可以在自己的程序和文档中使用本书中

的代码，如果你要复制的不是很大一部分代码，则无须取得我们的许可。例如，你可以在程序中使用本书中的多个代码块，无须获取我们许可。但是，要销售或分发来源于 O'Reilly 图书中的示例的光盘则必须得到许可。通过引用本书中的示例代码来回答问题时，不需要事先获得我们的许可。但是，如果你的产品文档中融合了本书中的大量示例代码，则需要取得我们的许可。

我们很欢迎引用时给出署名，但不做要求。署名一般包括书名、作者、出版社和 ISBN。例如“R Graphics Cookbook by Winston Chang (O'Reilly). Copyright 2013 Winston Chang, 978-1-449-31695-2”。

如果你觉得你对例子代码的使用已经超出了合理程度或者上文所述的许可范围，请通过 permissions@oreilly.com 联系我们。

我们的联系方式

如果你想就本书发表评论或有任何疑问，敬请联系出版社。

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

我们还为本书建立了一个网页，其中包含了勘误表、示例和其他额外的信息。你可以通过如下地址访问该网页：

http://oreil.ly/R_Graphics_Cookbook

关于本书的技术性问题或建议，请发邮件到：

bookquestions@oreilly.com

欢迎登录我们的网站（<http://www.oreilly.com>），查看更多我们的书籍、课程、会议和最新动态等信息。

我们的其他联系方式如下：

Facebook: <http://facebook.com/oreilly>
Twitter: <http://twitter.com/oreillymedia>
YouTube: <http://www.youtube.com/oreillymedia>

致谢

没有一本书的诞生可完全归结为个人的成果。本书的完成得到了很多人直接或间接的帮助。我要感谢 R 社区创造并培育了一个积极活跃的生态系统。非常感谢 Hadley Wickham 的诸多帮助：他编写了本书所依赖的软件包 `ggplot2`，并在 O'Reilly 出版社考虑出版 R 图形书籍的时候推荐了我，他还为我打开了一扇深入了解 R 的窗户。

感谢本书的技术审稿人 Paul Teetor、Hadley Wickham、Dennis Murphy 和 Erik Iverson。他们渊博的知识和对细节的重视极大地提高了本书的质量。我还要感谢 O'Reilly 出版社积极推进此书的编辑们：Mike Loukides 在初始阶段给了我很多指导，Courtney Nash 伴我走到了最后。此外，我还要郑重感谢 O'Reilly 的 Holly Bauer 及制作团队的其他成员，他们从头到尾耐心地做了很多细致的编辑工作，并为此书增添了不少特色。

最后，我还要感谢我的妻子 Sylia，感谢她一贯的支持和理解——当然不只是在写本书的时候。

目录

第 1 章 R 基础	1
1.1 安装包	1
1.2 加载包	2
1.3 加载分隔符式的文本文件	2
1.4 从 Excel 文件中加载数据	4
1.5 从 SPSS 文件中加载数据	5
第 2 章 快速探索数据	6
2.1 绘制散点图	6
2.2 绘制折线图	8
2.3 绘制条形图	9
2.4 绘制直方图	11
2.5 绘制箱线图	13
2.6 绘制函数图像	14
第 3 章 条形图	16
3.1 绘制简单条形图	16
3.2 绘制簇状条形图	19
3.3 绘制频数条形图	21
3.4 条形图着色	23
3.5 对正负条形图分别着色	24
3.6 调整条形宽度和条形间距	26
3.7 绘制堆积条形图	28
3.8 绘制百分比堆积条形图	31
3.9 添加数据标签	33
3.10 绘制 Cleveland 点图	37
第 4 章 折线图	42
4.1 绘制简单折线图	42

4.2	向折线图添加数据标记	44
4.3	绘制多重折线图	45
4.4	修改线条样式	49
4.5	修改数据标记样式	50
4.6	绘制面积图	52
4.7	绘制堆积面积图	54
4.8	绘制百分比堆积面积图	56
4.9	添加置信域	58
第 5 章	散点图	60
5.1	绘制基本散点图	60
5.2	使用点形和颜色属性, 并基于某变量对数据进行分组	62
5.3	使用不同于默认设置的点形	64
5.4	将连续型变量映射到点的颜色或大小属性上	66
5.5	处理图形重叠	69
5.6	添加回归模型拟合线	74
5.7	根据已有模型向散点图添加拟合线	78
5.8	添加来自多个模型的拟合线	81
5.9	向散点图添加模型系数	84
5.10	向散点图添加边际地毯	87
5.11	向散点图添加标签	88
5.12	绘制气泡图	92
5.13	绘制散点图矩阵	94
第 6 章	描述数据分布	99
6.1	绘制简单直方图	99
6.2	基于分组数据绘制分组直方图	101
6.3	绘制密度曲线	104
6.4	基于分组数据绘制分组密度曲线	107
6.5	绘制频数多边形	109
6.6	绘制基本箱线图	110
6.7	向箱线图添加槽口	112
6.8	向箱线图添加均值	113
6.9	绘制小提琴图	114
6.10	绘制 Wilkinson 点图	117

6.11	基于分组数据绘制分组点图	119
6.12	绘制二维数据的密度图	120
第 7 章	注解	123
7.1	添加文本注解	123
7.2	在注解中使用数学表达式	126
7.3	添加直线	127
7.4	添加线段和箭头	129
7.5	添加矩形阴影	131
7.6	高亮某一元素	132
7.7	添加误差线	133
7.8	向独立分面添加注解	136
第 8 章	坐标轴	139
8.1	交换 x 轴和 y 轴	139
8.2	设置连续型坐标轴的值域	140
8.3	反转一条连续型坐标轴	143
8.4	修改类别型坐标轴上项目的顺序	144
8.5	设置 x 轴和 y 轴的缩放比例	145
8.6	设置刻度线的位置	147
8.7	移除刻度线和标签	148
8.8	修改刻度标签的文本	149
8.9	修改刻度标签的外观	151
8.10	修改坐标轴标签的文本	153
8.11	移除坐标轴标签	154
8.12	修改坐标轴标签的外观	155
8.13	沿坐标轴显示直线	157
8.14	使用对数坐标轴	158
8.15	为对数坐标轴添加刻度	163
8.16	绘制环状图形	165
8.17	在坐标轴上使用日期	169
8.18	在坐标轴上使用相对时间	172
第 9 章	控制图形的整体外观	174
9.1	设置图形标题	174
9.2	修改文本外观	176

9.3	使用主题	178
9.4	修改主题元素的外观	180
9.5	创建自定义主题	183
9.6	隐藏网格线	184
第 10 章	图例	185
10.1	移除图例	185
10.2	修改图例的位置	187
10.3	修改图例项目的顺序	188
10.4	反转图例项目的顺序	190
10.5	修改图例标题	191
10.6	修改图例标题的外观	193
10.7	移除图例标题	194
10.8	修改图例标签	195
10.9	修改图例标签的外观	198
10.10	使用含多行文本的标签	199
第 11 章	分面	200
11.1	使用分面将数据分割绘制到子图中	200
11.2	在不同坐标轴下使用分面	202
11.3	修改分面的文本标签	204
11.4	修改分面标签和标题的外观	206
第 12 章	配色	207
12.1	设置对象的颜色	207
12.2	将变量映射到颜色上	208
12.3	对离散型变量使用不同的调色板	210
12.4	对离散型变量使用自定义调色板	214
12.5	使用色盲友好式的调色板	215
12.6	对连续型变量使用自定义调色板	217
12.7	根据数值设定阴影颜色	218
第 13 章	其他图形	221
13.1	绘制相关矩阵图	221
13.2	绘制函数曲线	224
13.3	在函数曲线下添加阴影	225

13.4	绘制网络图	227
13.5	在网络图中使用文本标签	230
13.6	如何绘制热图	232
13.7	绘制三维散点图	234
13.8	在三维图上添加预测曲面	237
13.9	保存三维图	240
13.10	三维图动画	241
13.11	绘制谱系图	241
13.12	绘制向量场	244
13.13	绘制 QQ 图	248
13.14	绘制经验累积分布函数图	249
13.15	创建马赛克图	250
13.16	绘制饼图	254
13.17	创建地图	255
13.18	绘制等值区域图	258
13.19	创建空白背景的地图	262
13.20	基于空间数据格式 (shapefile) 创建地图	263
第 14 章	输出图形用以展示	266
14.1	输出为 PDF 矢量文件	266
14.2	输出为 SVG 矢量文件	267
14.3	输出为 WMF 矢量文件	268
14.4	编辑矢量格式的输出文件	268
14.5	输出为点阵 (PNG/TIFF) 文件	270
14.6	在 PDF 文件中使用字体	272
14.7	在 Windows 的点阵或屏幕输出中使用字体	274
第 15 章	数据塑形	276
15.1	创建数据框	277
15.2	从数据框中提取信息	277
15.3	向数据框添加列	278
15.4	从数据框中删除一列	279
15.5	重命名数据框的列名	279
15.6	重排序数据框的列	280
15.7	从数据框提取子集	281