O'REILLY®

TURING 图灵程序设计丛书

R图形化数据分析

Graphing Data with R

数据分析入门首选 // 无需编程背景

[美] John Jay Hilfiger 著

王洋洋 译

中国工信出版集团

人民邮电出版社
POSTS & TELECOM PRESS

TURING 图灵程序设计丛书

R图形化数据分析

Graphing Data with R

[美] John Jay Hilfiger 著
王洋洋 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（CIP）数据

R图形化数据分析 / (美) 约翰·杰伊·希尔菲杰 (John Jay Hilfiger) 著 ; 王洋洋译. -- 北京 : 人民邮电出版社, 2017.8
（图灵程序设计丛书）
ISBN 978-7-115-46441-5

Ⅰ. ①R… Ⅱ. ①约… ②王… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2017)第180405号

内 容 提 要

本书介绍如何使用图形化的方法来分析和理解复杂的数据，该方法突出数据中重要的关联和分布趋势，并使用尽可能简单的视觉元素来呈现尽可能丰富的信息。本书重点介绍如何理解数据分析的图形元素，以及如何使用R生成书中涉及的各种图形。附录中列有大量参考资料，以及章节练习解答、相关R函数、R包、故障排查等信息，便于读者深入学习。

本书适合任何需要数据分析和数据可视化的读者。

◆ 著 [美] John Jay Hilfiger
译 王洋洋
责任编辑 朱 巍
执行编辑 温 雪 张 憬 回 春
责任印制 彭志环

◆ 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
三河市海波印务有限公司印刷

◆ 开本：800×1000 1/16
印张：15.75 彩插：4
字数：372千字 2017年8月第1版
印数：1-3 000册 2017年8月河北第1次印刷
著作权合同登记号 图字：01-2016-4796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：京东工商广登字 20170147 号

版权声明

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*

前言

谚语说："一图胜千言。"有时，一张图也胜过很多数据。相比口头描述细微差别或者辨别成列数字间的关系，通过观察图片或图表更容易把握数据间的复杂关系。本书主要介绍如何使用图形化方法来理解复杂的数据，该方法强调重要的关系和趋势，简化数据形式，并且使大量数据一目了然。

目标读者

任何需要分析数据和可视化数据的人，都能从本书中受益。然而，我的主要目的是使更广泛的人群理解图形数据分析，特别是那些没有太多（或任何）R 相关经验，但又需要或想要创建各种类型的图表来理解重要数据的人。这些人可能来自商业、媒体、平面艺术、社会科学或者健康科学领域，真的需要分析数据，但可能并没有高等数学和计算机编程的背景。虽然本书专为自学设计，但也可作为初中级统计课程或研究的补充材料。

本书使用的工具是 R。这不是一本关于 R 的内容全面的教材。许多计算机课程和图书都试图告诉你借助一种语言或工具可能做的每一件事。对于曾经想按此方式学习的大多数人来说，这种方式令人感到十分烦恼和无聊。本书将把重点放在理解数据分析的图形元素和如何使用 R 生成本书讨论的各种图形，也将展示如何使用 R 的一些内置资源来获得帮助，很多其他内容则留给你继续探究。你应该有台可用的计算机，用它可轻松完成一些工作，如发送电子邮件、浏览互联网，或者使用文字处理软件、电子表格等应用程序。熟悉基本的统计知识有利于理解本书的一些主题，但对于大多数主题，这并不是必需的。

为什么选择R

小数据量的图表可以手工制作，但是利用计算机技术会更高效、准确地分析数据，生成有吸引力的图形。对于大批量数据来说，手工处理实际上是不可能的。而运用计算机软件，即使是针对非常大的数据量，也可以生成复杂的图形。

实际上，开源软件已经实现了该技术，只要拥有一台计算机。“开源”指的是所有人均可获取项目的源代码，可检查、使用、自由修改或增加源代码。

开源软件产品可提供免费下载给任何有需要的人。或许你会怀疑免费的东西质量不高，但我向你保证，一些自由软件遵循了最高的专业标准。

本书选用的 R 语言是一种编程语言，是一个统计、数学和绘图程序集合，已经被世界各地数百万人使用，包括科学、商业和媒体等领域的许多专业人士。在网站、主要报纸和其他出版物上，你可能见过由 R 制作的图形。你也将能够制作出这种专业的数据图表，因为 R 可运行在 Windows、Mac 或 Linux 操作系统上，而现在的 PC 和笔记本无非就这几类系统！

如何使用本书

要想从本书获益，你需要动手制作大量图表。为此，阅读本书时，你最好坐在计算机前操作书中给出的所有命令。而且为帮助你提升水平，许多章节除示例以外还提供了练习，比如优化示例代码或将不同的数据集制成另外一张图。最好先做完这些练习再进入下一主题。

排版约定

本书使用了下列排版约定。

- 楷体
  表示新术语。
- 等宽字体（`constant width`）
  表示程序片段，以及正文中出现的变量、函数名、数据库、数据类型、环境变量、语句和关键字等。
- 加粗等宽字体（**`constant width bold`**）
  表示应该由用户输入的命令或其他文本。
- 斜体等宽字体（*`constant width italic`*）
  表示应该由用户输入的值或根据上下文确定的值替换的文本。

该图标表示一般注释。

代码示例的使用

本书会帮你完成工作。一般来说，如果本书提供了示例代码，你可以把它用在你的程序或文档中。除非你使用了很大一部分代码，否则无需联系我们获得许可。比如，用本书的几个代码片段写一个程序就无需获得许可，而销售或分发 O'Reilly 图书的示例光盘则需要获得许可；引用本书中的示例代码回答问题无需获得许可，而将书中大量的代码放到你的产品文档中则需要获得许可。

我们很希望但并不强制要求你在引用本书内容时加上引用说明。引用说明一般包括书名、作者、出版社和 ISBN。比如："*Graphing Data with R* by John Jay Hilfiger (O'Reilly). Copyright 2016 John Jay Hilfiger, 978-1-491-92261-3."

如果你觉得自己对示例代码的用法超出了上述许可的范围，欢迎你通过 permissions@oreilly.com 与我们联系。

Safari® Books Online

Safari Books Online（http://www.safaribooksonline.com）是应运而生的数字图书馆。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。技术专家、软件开发人员、Web 设计师、商务人士和创意专家等，在开展调研、解决问题、学习和认证培训时，都将 Safari Books Online 视作获取资料的首选渠道。

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

你还可以发送电子邮件到 bookquestions@oreilly.com。

勘误、示例和其他信息可到 http://www.oreilly.com/catalog/0636920038382.do 上获取。

欲了解本社图书、课程、会议和新闻等更多信息，请访问我们的网站 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

在很多人的帮助下，我完成了这本书。首先，妻子 Karen 在我整个写作过程中给予的耐心、理解和鼓励，对我完成本书至关重要。我们的儿子 Eric 和女儿 Kristen 读了第 1 章后，给出了相当直接的评价，使我感到羞愧但很有帮助。担纲技术审校的 Dr. Peter Bajorski、Sarah Boslaugh 和 Philipp K. Janert 的见解、纠正和建议是很宝贵的。本书编辑 Shannon Cutt 非常积极能干，不仅在写作上提供帮助，而且在准备手稿的所有技术和操作细节上提供帮助。我不知道竟有这么多工作需要做！最后，O'Reilly 团队做了所有你看得到和看不到的事情，这一切对于生产高质量的图书至关重要，他们是如此令人尊敬。感谢所有人。

电子书

扫描如下二维码，即可购买本书电子版。

目录

第二部分 单变量图

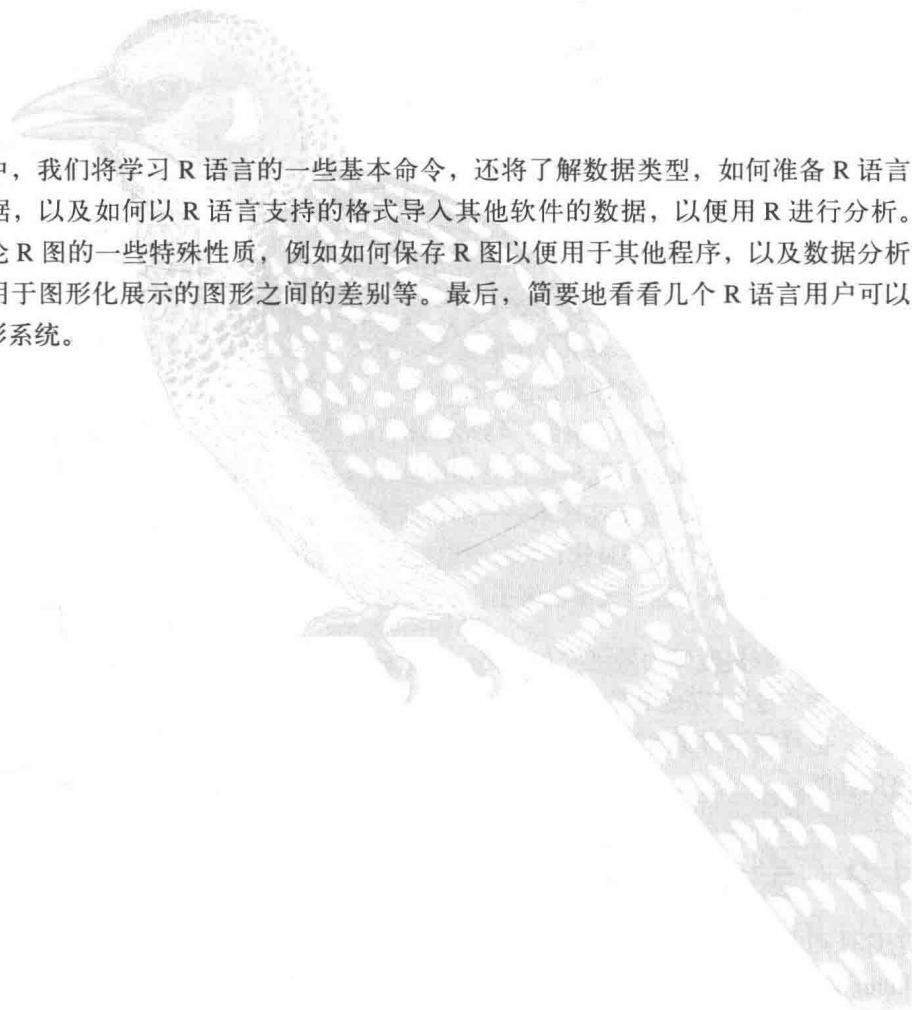
第三部分 双变量图

第四部分 多变量图

第五部分 现在该做些什么

第一部分

开始使用R



在本部分中，我们将学习R语言的一些基本命令，还将了解数据类型，如何准备R语言使用的数据，以及如何以R语言支持的格式导入其他软件的数据，以便用R进行分析。之后将讨论R图的一些特殊性质，例如如何保存R图以便用于其他程序，以及数据分析的图形和用于图形化展示的图形之间的差别等。最后，简要地看看几个R语言用户可以使用的图形系统。

第1章

R基础

1.1 下载软件

首先需要下载免费的R软件，并安装在你的计算机上。启动计算机，打开Web浏览器，通过网址http://www.r-project.org访问统计计算的R项目（R Project for Statistical Computing）。点击download R，然后选一个距离你所在地理位置较近的镜像站点。（R软件存储在世界各地的许多计算机上，而不是一台计算机上。因为它们都包含相同的文件，看起来一样，所以被称为“镜像”网站。你可以选择其中任何一台计算机。）点击网站地址，将打开一个页面，可以根据你的操作系统选择R的版本。如果你的计算机可以运行最新版本的R——3.0或更高版本——那再好不过了。然而，如果你的计算机用了几年了，不能运行最新的版本，那就选计算机可以运行的最新版本。也许会和本书中的例子有一些小的差异，但应该可运行大多数例子。

按照说明，在短时间内就可安装好R。这是*基础*R（base R），但R有数千个（这不是夸张）插件“包”。完成R的基础安装后，你可以免费下载插件来扩展R的功能。根据你的需要，也可能不需要添加任何插件，但是你可能会很惊奇地发现一些你想象不到但必须拥有的功能。

1.2 尝试一些简单的任务

若你使用的是Windows或OS X系统，单击桌面上的“R”图标即可启动R。若使用的是Linux或OS X系统，在终端窗口输入命令R可打开*控制台*（console）。在控制台窗口输入命令，可看到许多命令的结果，尽管在大多数情况下，创建图形的命令将打开一个新窗口

来展示结果图。R 可接受命令时，会显示一个大于符号（>）作为提示符。R 最简单的应用是计算器。在提示符后，输入一个你想要获得答案的数学表达式：

```
>12/4
[1] 3
>
```

此处，我们请求“12 除以 4”。R 返回“3”，然后显示另一个提示符“>”，表明可以输入下一个命令。返回值前的 [1] 是一个*索引*（index），在本例中，它只是表明返回值从向量（vector）中的第一个数开始。本例中只有一个值，但有时会有多个值，所以了解数据集合从哪里开始会有帮助。如果你不理解索引，现在也不用担心；看到更多的例子后，你将更清楚。除法符号（/）叫作*操作符*（operator）。表 1-1 给出了标准算术运算符的符号。

表1-1：R算术运算符

运算符	运算	举例
+	加法	3 + 4 = 7 或 3+4（即无空格）
-	减法	5 - 2 = 3
*	乘法	100*2.5=250
/	除法	20/5= 4
^ 或 **	幂	3^2 = 9 或 3**2 = 9
%%	取余	5%%2 = 1（5/2=2 余 1）
%/%	除法，向下取整	5%/% = 2（5/2=2.5，向下取整等于 2）

你可以像在普通算术中那样使用括号，以表明操作的执行顺序：

```
> (4/2)+1
[1] 3
> 4/(2+1)
[1] 1.333333
```

尝试另一个例子：

```
> sqrt(57)
[1] 7.549834
```

这次通过*函数*（function）来完成运算，本例用的是 sqrt() 函数。表 1-2 列出了一些常用的算术函数。

表1-2：常用的R数学函数

函数	运算	函数	运算	函数	运算
cos()	余弦	exp()	指数函数	max()	最大值
sin()	正弦	sum()	求和	var()	方差
tan()	正切	mean()	均值	sd()	标准差
sqrt()	平方根	median()	中间值		
log()	对数	min()	最小值		

调用函数时可带参数（argument）。参数是一种修饰符，与函数一起使用时可以用更多特殊的方式请求R。因此，可能会请求计算特定数字的和，而不是单单请求sum函数；或者，你可以使用参数来指定线的颜色或宽度，而不是简单地在图上画一条线。单个或多个参数，必须在函数名后用括号括起来。当使用函数或任何R命令时，如果需要帮助，可使用如下方式寻求帮助：

```
> help(sum)
```

R 将打开一个新窗口，显示要查找的指定函数及其参数的信息。如下命令是一个快捷方式，可得到完全相同的响应：

```
> ?sum
```

请注意，R 是区分大小写的，所以“help”和“Help”是不同的！然而，空格无关紧要，因此上面的命令也可写为：

```
> ? sum
```

有时，函数中只有一个参数，比如 sqrt() 的示例。其他情况下，一个函数作用于一组数字，称为向量，如下所示：

```
> sum(3,2,1,4)
[1] 10
```

本例使用 sum() 函数计算 3、2、1、4 的和。但不可能总是像这个例子一样，把所有的值写入函数表达式。因此，通常需要先创建向量，如下所示：

```
> x1 <- c(1,2,3,4)
```

输入这个命令后，什么也没发生！你确实看到什么也没有发生。一旦出现由“<”和“-”两个符号组成的特殊操作符，操作符右边的值就会赋给左边的变量。（R 的较新版本允许使用“=”号来完成赋值。第 1 章以后，我们也将使用这种较简单的形式。）在这种情况下，创建一个新的向量，用户称其为 x1。R 是一种面向对象语言（object-oriented language），向量 x1 是工作区中的一个对象（object）。

什么是“对象”

将对象想象为一个盒子，其中装满了彼此关联的项。这些项可以是简单的数字、名称、统计分析的结果、这些项或其他项的组合。对象有助于以结构化的方式组织事物，将彼此相关的东西封装在同一个盒子里，无关的东西封装在其他盒子里；对象可以告诉 R 在其中有哪些项，以便 R 合理地操作特定对象中的项。向量是一类包含相同类型数据（都是数字或者都是字母）的对象。对象可以包含其他对象。毕竟，你可以把一个盒子放入一个更大的盒子里。因此，可以把一个或多个向量放入一个**数据框**（data frame），数据框是另一种类型的对象。输入命令 **ls()**，可以查看当前工作区有哪些对象。