资深数据科学家多维度详解R语言数据可视化技术的精髓
通过500多个可视化示例，全面展现30多类统计图形的绘制

R语言数据可视化实战

米霖◎编著

机械工业出版社
China Machine Press

R语言

数据可视化实战

米霖◎编著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

R语言数据可视化实战 / 米霖编著. —北京：机械工业出版社，2020.11

ISBN 978-7-111-66791-9

Ⅰ. R… Ⅱ. 米… Ⅲ. 统计分析－应用软件 Ⅳ. C819

中国版本图书馆CIP数据核字（2020）第199849号

R 语言数据可视化实战

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）

责任编辑：陈佳媛　　责任校对：姚志娟

印　　刷：中国电影出版社印刷厂　　版　　次：2020 年 11 月第 1 版第 1 次印刷

开　　本：186mm × 240mm　1/16　　印　　张：35.75

书　　号：ISBN 978-7-111-66791-9　　定　　价：169.00 元

客服电话：（010）88361066　88379833　68326294　　投稿热线：（010）88379604

华章网站：www.hzbook.com　　读者信箱：hzit@hzbook.com

前言

数据可视化是一种非常重要的技能，能够帮助人们快速理解信息，分析数据中存在的趋势，识别关系中的模式。当人们有了新的见解后，往往需要将这些见解传达给其他人，在传达的过程中，使用图表、图形或其他对视觉有影响的表现形式非常重要，因为这些表现形式吸引人并可以迅速传达信息。

R 是由统计学家设计的专门用于统计计算的语言，它也是一种非常好的数据可视化工具。随着技术的进步，数据公司或研究机构对数据的收集变得越来越复杂，许多人已经采用 R 语言作为分析数据的首选语言。R 语言适用于机器学习、数据分析、数据可视化及科学计算等领域。它有多个主题的软件包，如计量经济学、金融学和时间序列等，也拥有一流的可视化、报告和交互性工具。这些工具对于企业和科学研究都非常重要，被科学家、工程师和商业人士广泛使用。

笔者多年来一直以 R 语言为工具从事数据可视化、数据分析、统计建模和机器学习等数据科学工作，在工作中积累了大量的经验，对 R 语言的原理和应用有比较全面和深入的认识，尤其在数据可视化方面更是有独到的见解。R 语言提供了强大的数据可视化功能，可以生成高质量的图形，如条形图、直方图、散点图、动态图和数学符号，甚至可以用很少的代码来实现自己设计的全新图形。R 语言有很多数据可视化包，如 ggplot2、ggivs 和 plotly 等，使用这些包可以非常轻松地绘制出令人印象深刻的数据可视化图表。其中，plotly 包提供了一套绘制交互式图形的工具，所绘制的图形非常有表现力。另外，R 语言的文档资源很丰富，而且质量很高，这对学习 R 语言有很大的帮助。

为了帮助广大相关从业人员和数据技术爱好者快速掌握 R 语言数据可视化技术，笔者编写了本书。本书主要介绍如何使用 R 语言绘制常见的统计图形，如直方图、散点图和箱线图等，另外也介绍了如何绘制一些不常见但很实用的统计图形，如桑基图、和弦图和时间序列图等。

本书不但介绍了普通的统计图形的绘制，而且介绍了交互式图形及动画图形的绘制，另外还介绍了如何使用 Shiny 工具包进行数据探索与可视化。相信通过阅读本书，读者可以在较短时间内比较系统地掌握 R 语言数据可视化技术。

本书特色

1. 内容全面，理论结合实践

本书全面介绍如何利用 R 语言绘制各种统计图形，涵盖普通的统计图形绘制、高级统

计图形绘制、交互式图形绘制、动画图形绘制，以及统计图形的细节调整。另外，本书还介绍了如何利用 Shiny 进行数据探索与可视化，是一部理论与实践紧密结合的数据可视化学习宝典。

2. 零基础入门，学习门槛低

阅读本书不需要读者具备太多的预备知识，只要有基本的数学与统计学知识即可，学习门槛很低，零基础即可入门。

3. 实例丰富，实用性强

本书在讲解的过程中给出了大量的 R 语言绘图实例，这些实例涵盖单变量图形绘制、两个同类型变量的图形绘制、分类变量和连续变量的图形绘制、高维图形绘制等。这些例子都非常实用，可以比较容易地迁移到实际工作中。

4. 绘图代码详细，效果精美

本书中的绘图实例都给出了详细的实现代码，读者可以按照代码和操作步骤亲自动手实现每一个实例的效果，而且这些绘图实例的效果非常精美，让人赏心悦目。

本书内容

本书共分为 13 章，各章内容简单介绍如下：

第 1 章主要介绍 R 语言的基本概念、Rstudio 跨平台集成开发环境及常见的统计图形等。

第 2 章主要介绍一些数据处理与数据探索的方法，如数据转换和数据重塑等，因为在数据可视化之前需要先对数据进行处理。

第 3 章主要介绍在不进行更多细节调整的情况下如何快速地进行数据可视化。

第 4 章主要介绍面积图、密度图、直方图和频率图这几种单变量图形的绘制。这类图形往往只涉及数据集的一个变量。

第 5 章主要介绍两个同类型变量的图形绘制，包括散点图、抖动点图、连续二维分布图和线图。

第 6 章主要介绍分类变量和连续变量的图形绘制，包括箱线图、小提琴图、棒棒糖图、条形图、圆形条形图、饼图和甜甜圈图。

第 7 章主要介绍高维图形的绘制，包括气泡图、三维散点图、流型图、相关矩阵图、树状图、圆形包装图和树形图。

第 8 章主要介绍其他类型的统计图形绘制，包括和弦图、桑基图、网络图、旭日图、雷达图、词云、平行图、时间序列图、交互式图形及动画图。这些图形并不是很常见，但是非常有用，使用这些高级图形能够让数据可视化的效果更加引人注目。

第 9～11 章主要介绍图形的细节调整，如添加图形元素、图形的颜色调整、线条类型

调整、坐标轴范围调整、删除面板边框和网格线、合并多幅图形等。一幅优秀的统计图形往往需要许多细节上的调整，通过调整细节，可以让图形更具表现力。

第 12、13 章主要介绍一些扩展内容，包括 ggfortify 绘图包和 Shiny 工具包，它们可以实现数据可视化的一些高级功能，如交互式图形绘制和动画图形绘制等。

读者对象

- 数据可视化从业人员；
- 统计学、数学、经济学、计算机和财经等专业的本科生和研究生；
- 互联网从业人员，如产品经理；
- R 语言数据可视化初学者与进阶者；
- 对数据可视化感兴趣的人员；
- 相关培训机构的学员。

配套资源

本书的所有实例源代码文件及彩色效果图等相关资源需要读者自行下载。方法是：在华章公司官网 www.hzbook.com 上搜索到本书，然后单击“资料下载”按钮，即可在本书页面上找到“配书资源”下载链接。

售后支持

本书涉及的内容比较庞杂，R 数据可视化技术也是日新月异，加之作者水平和成书时间所限，书中难免有一些疏漏和不当之处，敬请读者指正。阅读过程中若有疑问，请发 E-mail 至 hzbook2017@163.com，以获得帮助。

目录

第 1 章　R 语言数据可视化简介

本章主要介绍 R 语言数据可视化的一些基本内容，包括什么是 R、什么是 Rstudio、什么是 R 包等，另外还会介绍 R 语言基本语法的相关内容。由于本书的重点在于数据可视化，因此不会深入讲解 R 语言的语法。本章旨在帮助读者了解 R 语言的一些核心内容，以尽可能简单的方式帮助读者了解如何使用 R 语言，最后介绍一些常见的统计图形。

1.1　R 语言介绍

R 是一种编程语言和自由软件，常用于统计计算和图形绘制。R 语言广泛应用于统计学领域和数据科学领域，用于开发统计软件和进行数据分析。R 是世界上最强大的统计计算、机器学习和图形编程语言，拥有蓬勃发展的全球用户、开发人员和贡献者社区。

R 语言是罗斯·伊哈卡（Ross Ihaka）和罗伯特·杰特曼（Robert Gentleman）在新西兰的奥克兰大学所创建的，以两个作者名字的第一个字母命名。R 项目于 1992 年构思，1995 年发布初始版本，2000 年发布稳定版本。R 可从 CRAN 网页进行下载，链接为 http://cran.r-project.org/。

R 及其库实现了各种统计和图形方法，包括线性和非线性建模、经典统计测试、时间序列分析、分类、聚类等。一般而言，最新的统计方法都有对应的 R 包，R 用户可以非常方便地学习、应用最新的统计方法。

R 可以通过函数和扩展包轻松扩展其功能，R 社区因在软件包方面的积极贡献而闻名。任何人都可以在 R 社区贡献出自己的包。到目前为止，R 中有超过 10 000 个包可供下载。R 开发人员社区非常活跃。R 在统计计算、数据科学及数据可视化方面有着无与伦比的优势，几乎你面临的所有问题或者你关心的问题都有人实现了 R 包。

R 是数据科学领域最流行的语言，并且 R 语言是完全面向数据的，更加注重从数据的角度去思考问题，在这一点上与其他的编程语言有很大的区别。

另外，R 是用于统计研究的主要工具。因此当新的方法被开发的时候，它不仅仅被作为论文发表，而且往往会被开发成为一个 R 包。这让 R 永远成为新算法的前沿，而 R 用户可以非常方便地使用这些新的算法。

CRAN 本身是一个非常有效的共享 R 扩展平台，具有用于包创作、构建、测试和分发的成熟系统。R 核心团队，特别是 CRAN 维护者，为 R 包创建了这样一个充满活力的生

态系统。图 1.1 所示为 R 包数量在不同年份的变化。

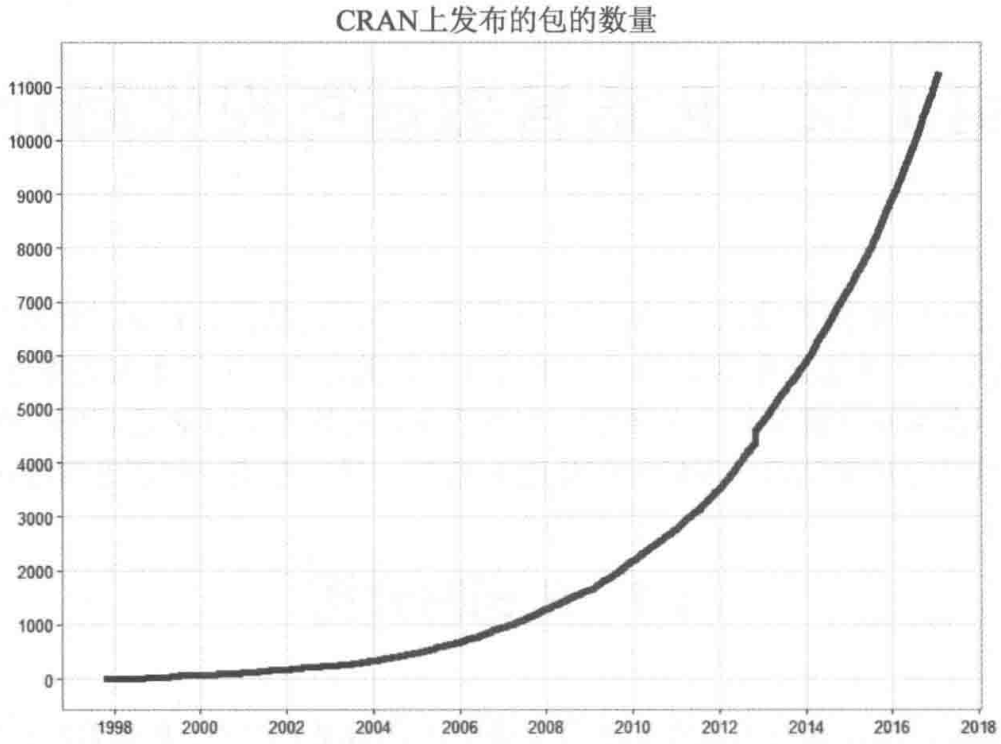

图 1.1　R 包数量的变化

从图 1.1 中可以看到，R 包的开发速度越来越快。但是过多的包导致人们在寻找自己所需要的包时困难重重，可能需要借助一些搜索工具来完成。目前有很多包的搜索与汇总工具，例如：

- CRAN 提供包任务视图（https://cran.r-project.org/web/views/），按主题区域（例如财务或临床试验）提供包目录。
- MRAN（Microsoft R 应用程序网络）为 CRAN 上的 R 软件包提供搜索工具。MRAN 的链接为 https://mran.microsoft.com/taskview。
- 为了找到最受欢迎的软件包，Rdocumentation.org 按下载次数提供了软件包的排行榜，并提供了新发布和最近更新的包的列表。RDocumentation.org（https://www.rdocumentation.org/taskviews#Bayesian）还提供了基于 CRAN 任务视图的可搜索版本。

通过上面的几个方法可以很方便地找到自己所需要的包。

R 有许多标准函数，使用这些函数可以解决非常多的统计机器学习的任务。对于计算密集型任务，可以在运行时连接和调用 C、C++和 Fortran 代码。高级用户可以编写 C、C++、Java、.NET 或 Python 代码直接操作 R 对象，还可以通过使用用户提交的包来执行特定功能或特定研究领域的任务。软件包使 R 具有高度可扩展性。

R 的语法非常简单，介绍 R 的语法首先需要从数据的角度入手，在使用任何编程语言进行编程时，需要使用各种变量来存储各种信息。与其他编程语言（如 C 和 Java）不同，在 R 中使用变量不需要声明类型，并且 R 中的数据由 R 语言的对象存储。这些对象包括

向量、列表、矩阵、数组、因子、数据框。

数据是数据科学的基础，也是数据可视化的基础。下面开始介绍 R 的基本数据结构。

1.1.1　向量

这里所说的向量并不是物理学中有方向的一个量，而是一组数据。例如，1～10，这 10 个数字就可以表示为一个向量。或者 26 个字母，也可以表示为一个向量。向量是 R 中的一个基本数据结构。当你想用多个元素创建向量时，应该使用 c 函数，这意味着将元素组合成一个向量。下面的代码则是创建向量的一个例子。

```
# 创建一个向量
letter <- c('a','b',"c")
letter
## [1] "a" "b" "c"
# 查看数据类型
class(letter)
## [1] "character"
```

上面的代码使用 c 函数创建了一个向量，包含 3 个元素，分别为 a、b、c 这 3 个字母。然后 class 函数可以用于查看数据的类型。从代码的输出结果中可以看出，这个向量中元素的数据类型是字符类型（character）。上面创建的向量有 3 个元素，如果希望提取向量中的某个元素，则可以通过[]进行提取，代码如下：

```
letter[1]
## a
```

上面的代码选取了向量中的第一个元素，希望提取第几个元素只需要传入对应数字即可。

1.1.2　列表

列表（list）这个数据类型常用于存储不同数据类型的数据。列表里面可以存储向量、列表、矩阵、数组、因子、数据框等所有数据类型。另外，列表还可以嵌套多层列表，下面的代码创建了一个列表，列表中分别包含了向量、列表、数组、矩阵、数据框这几种 R 语言中的数据结构。

```
l <- list(c(1,2,3),list(1),array(c(1,2,3,3,1,2),dim = c(1,2,3)),matrix
(c(1,2,3,4),nrow = 2),factor(c(1,2,3)),data.frame(nu = c(1:3),id =c("a",
"b","c")))
l
## [[1]]
## [1] 1 2 3
##
## [[2]]
## [[2]][[1]]
## [1] 1
```

```
##
##
## [[3]]
## , , 1
##
##      [,1] [,2]
## [1,]    1    2
##
## , , 2
##
##      [,1] [,2]
## [1,]    3    3
##
## , , 3
##
##      [,1] [,2]
## [1,]    1    2
##
##
## [[4]]
##      [,1] [,2]
## [1,]    1    3
## [2,]    2    4
##
## [[5]]
## [1] 1 2 3
## Levels: 1 2 3
##
## [[6]]
##   nu id
## 1  1  a
## 2  2  b
## 3  3  c
class(l)
## [1] "list"
```

上面的代码创建了 6 种类型的数据，然后将它们全部放到一个 list 里面。如果希望提取列表中的元素，则需要使用[[]]这个符号。l[[1]]表示提取列表的第一个元素，也就是整个向量。如果希望进一步提取向量中的元素，提取方式与普通向量的提取方式相同。例如，代码 l[[1]][1]提取了列表中向量的第一个元素。

1.1.3 矩阵

矩阵可以理解为二维数组，它可以使用矩阵函数（matrix）来创建。执行下面的代码，将创建一个矩阵。

```
# Create a matrix.
M = matrix( c('a','a','b','c','b','a'), nrow = 2, ncol = 3, byrow = TRUE)
print(M)
##      [,1] [,2] [,3]
## [1,] "a"  "a"  "b"
## [2,] "c"  "b"  "a"
```