数据分析与决策技术丛书

HZ BOOKS 华章IT

[PACKT] PUBLISHING

R Data Analysis Cookbook

R数据分析秘笈

[美] 维西瓦·维斯瓦纳坦（Viswa Viswanathan）
珊蒂·维斯瓦纳坦（Shanthi Viswanathan） 著

鱼翔 译

资深技术专家多年经验结晶，深入剖析R数据分析的实用方法、技巧和最佳实践

通过80多个基于任务的实际应用案例，帮助你快速提升R数据分析能力，轻松搞定数据分析项目

机械工业出版社
China Machine Press

数据分析与决策
技术丛书

R Data Analysis Cookbook

R数据分析秘笈

[美] 维西瓦·维斯瓦纳坦（Viswa Viswanathan）
珊蒂·维斯瓦纳坦（Shanthi Viswanathan） 著

鱼翔 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

R 数据分析秘笈 /（美）维斯瓦纳坦（Viswanathan, V.），（美）维斯瓦纳坦（Viswanathan, S.）著；鱼翔译 . —北京：机械工业出版社，2016.3
（数据分析与决策技术丛书）
书名原文：R Data Analysis Cookbook

ISBN 978-7-111-53173-9

I. R… II. ① 维… ② 维… ③ 鱼… III. 程序语言 - 程序设计 IV. TP312

中国版本图书馆 CIP 数据核字（2016）第 045174 号

本书版权登记号：图字：01-2015-5216

R 数据分析秘笈

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：李 艺
责任校对：董纪丽
印 刷：北京诚信伟业印刷有限公司
版 次：2016 年 4 月第 1 版第 1 次印刷
开 本：186mm × 240mm 1/16
印 张：17.25
书 号：ISBN 978-7-111-53173-9
定 价：59.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426 88361066
投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259
读者信箱：hzit@hzbook.com

The Translator's Words 译者序

尽管仍然受到战争、饥饿、环境等问题的困扰，但无法否认的是，人类正处于历史上最好的时代。得益于计算机和数据处理技术的飞速发展，以自动驾驶汽车、Siri 语音助手、随时随地的手机支付等为代表的现代技术应用正将人类照顾得无微不至。对于数学、统计、计算机专业的人来说，这更是一个最好的时代，因为我们有幸见证了机器学习、自然语言处理、高速计算机集群的大规模应用，并实实在在地改变了我们的世界。

数据分析对于业界有没有用？有多大用处？从 20 世纪末的冷遇到现在的如日中天，我相信很多人都感受到了整个世界对于数据价值理解的巨大变化。在这十多年中，R 作为数据科学最为青睐的语言之一，迅速地从学界渗透到业界，发展壮大。

一直以来，R 最大的优势就是全球统计界（现在应该还要加上数据科学界）的强力支持（截至我写这篇序的这一刻，CRAN 上已经有 7514 个包）。这一点是任何其他数据分析工具所不可比拟的（比如 SAS、Python、SPSS 等）。除此之外，R 的灵活和开放性也使它能够很好地与其他语言和数据库沟通，以及处理非结构化的数据。

自从 2008 年我在美留学期间接触到 R 语言以来，不知不觉已经是第 8 个年头，而中国的 R 语言大会也已经如火如荼地开到了第八届。这几年中，我有幸目睹了 R 在学术界和业界的迅速发展，看到了一批又一批的优秀人才涌入到数据科学的浪潮中，而我自己也从 R 语言的学习者逐渐转向了它的传播者。这几年来，我在大学教授统计 / 数据分析课程，并为业界解决一些实际问题。以我浅薄的经验来看，一方面，就业市场对于统计类人才的渴望越来越强烈；然而另一方面，统计系毕业的学生又很少能在毕业时拥有在实际环境中处理数据的能力。原因是多方面的，比较重要的一点是，很多学校在教授统计 / 数据分析类课程的时候，缺少真实环境下的分析能力培养，教材也多偏重于统计理论或者 R 语言的基础，合适的教材比较匮乏。我也曾考虑将这几年教学和实践中对于数据处理的一些流程技巧整理编成一个小册子，但未能完成。

当看到本书的目录时，我立刻感觉到非常强烈的共鸣——Viswa Viswanathan 教授已经将 R 数据分析完美呈现。从各类源数据的读入和调整，数据分析前的准备工作、清洗、转换，到面向各类需求的各种模型，再到能够显著提高效率的自动化报告系统 knitr 和交互式可视化系统 shiny，最后到与 Java、MySQL、MongoDB 和 Excel 之间的配合工作，本书为初级和中级数据分析师准备了八十多种方法，帮助他们完成真实场景中的各项任务。

同时，书中的每一个章节都相对独立，作者为其设定了非常清晰的内容结构，尽量减少读者不停翻阅前文的情况（当然，第一次从头到尾读下来的读者可能会觉得这位严谨的教授有点烦，但当你在半年后需要查询书中的某一个方法时，也许会改变这一想法）。

最后，我想在此感谢我的父母和我的妻子，在我常常翻译到深夜的日子里，是他们无微不至地照顾一岁多的小朋友。我也要感谢 Rigi，你为这个家庭带来了无数的欢乐，希望你健康快乐地成长。

——2015 年 11 月 23 日，写于第八届中国 R 语言大会之后

About the Authors 作者简介

Viswa Viswanathan 是西顿霍尔大学斯蒂尔曼商学院计算和决策科学系的一名副教授。在获得人工智能领域的博士学位之后，Viswa 先从事了十多年学术工作，接下来的十几年在软件行业高就。在这段时间中，他曾就职于 Infosys、Igate 和 Starbase 公司。他于 2011 年重新回归学术界。

Viswa 在非常广泛的领域中开展教学，包括运筹学、计算机科学、软件工程、管理信息系统，以及企业系统。除了在大学中教学之外，Viswa 还负责专业人士的培训项目。他有多篇同行评议的研究论文发表在《Operations Research》《IEEE Software》《Computers and Industrial Engineering》以及《International Journal of Artificial Intelligence in Education》等期刊上。他也编写了《Data Analytics with R: A hands-on approach》一书。

Viswa 非常享受亲自动手开发软件的过程，并且独立构思、搭建、开发、部署了几个基于网络的应用程序。

除了对数据分析、人工智能、计算机科学、软件工程等技术领域有深厚的兴趣之外，Viswa 也对教育有浓厚的兴趣，特别关注学习的根源和培养更深入学习的方法。他已经在这个领域做了不少研究并希望在未来继续研究这一学科。

Viswa 想对 Amitava Bagchi 和 Anup Sen 教授表示由衷的感激，他们在 Viswa 的早期研究生涯中鼓舞了他。同时，他也很感激几个非常聪明的同事，比如 Rajesh Venkatesh、Dan Richner 和 Sriram Bala，他们极大地影响了他的思想。他的婶婶 Analdavalli，他的姐妹 Sankari，以及他的妻子 Shanthi，在辛勤工作上教会了他很多，即便他只吸收了一点皮毛也觉得受益匪浅。他的儿子 Nitin 和 Siddarth 也在很多主题上给出了不计其数的深刻评论。

Shanthi Viswanathan 是一位经验丰富的技术专家，她为许多企业客户提供技术管理和企业结构咨询。她曾工作于 Infosys、Oracle 和 Accenture 公司。作为一名顾问，Shanthi 为一些大型机构，比如 Canon、Cisco、Celgene、Amway、Time Warner Cable 和 GE 等，在数据架构和分析，高级数据管理，面向服务的架构，商业流程管理，以及建模等方面提供

帮助。当她空闲时，Shanthi 会在纽约州和新泽西州的郊外徒步旅行，摆弄园艺，以及教授瑜伽。

Shanthi 想要感谢她的丈夫 Viswa，在他们一起徒步旅行时关于各种主题展开的深入讨论；以及将她带入 R 和 Java 的世界。她也要感谢她的儿子 Nitin 和 Siddarth 使她进入了数据分析领域。

About the Reviewers 审校者简介

Kenneth D. Graves 相信数据科学会带给我们超能力。或者至少能让我们做出更好的决策。他在数据科学和技术上拥有超过 15 年的丰富经验，特别擅长机器学习、大数据、信号处理和营销，以及社会媒体分析。他曾就职于财务 500 强公司，比如哥伦比亚广播公司和美国无线电公司，以及金融和技术类公司，致力于设计最新的技术和数据解决方案来优化商业和企业决策流程并提高产出。他的项目包括脸部及品牌识别，自然语言处理，以及预测分析。他使用 R、Python、C++、Hadoop 和 SQL 这些语言工作，并指导他人。

Kenneth 在数据科学、商业、电影、古典语言方面都拥有学位或认证。当他没有致力于发现超能力的时候，他是一名数据科学家，并在 Soshag——一家社会媒体分析公司担任 CTO。他也参与大波士顿地区的咨询和数据科学项目。目前他居住在马萨诸塞州的韦尔斯利。

我想感谢我的妻子 Jill，她是我所做的一切的灵感来源。

Jithin S L 在 LoyoIa 理工大学获得了他的信息技术学士学位。他在分析领域开始了他的职业生涯，随后转向了多个大数据技术垂直领域。他在多家知名企业工作，例如汤姆森路透、IBM、Flytxt 等。他的工作涉及银行业、能源、保健以及电信领域，并且参与过大数据技术的全球项目。

他在国内和国际会议上发表过多篇技术和商业领域的研究论文。目前，Jithin 是 IBM 公司的系统分析师——商业分析中的大数据大视野和优化单元。

“改变能让我们思考超过人类极限的事物，恐惧改变同时也提供了以崭新的方式学到崭新事物的机会，试验、探索，以及通往成功的机遇。”

——Jithin

我要将本书献给我的父亲 N. Subbian Asari、我挚爱的母亲 M. Lekshmi，以及我可爱的

妹妹 S.L Jishma；有了他们的协助和鼓励我才能评审完这本书。最后但同样重要的是，我想感谢我所有的朋友。

Dipanjan Sarkar 是世界上最大的半导体公司 Intel 的一位 IT 工程师，他负责分析和企业程序开发的工作。作为目前业界经验的一部分，他曾在印度一家新兴的大数据分析初创公司 DataWeave 担任数据工程师，并在研究生期间在 Intel 实习。

Dipanjan 从 Bengaluru 的国际信息技术学院获得了硕士学位。他的兴趣包括研究新技术、颠覆性的初创公司，以及数据科学。他也审校了《 Learning R for Geospatial Analysis 》一书。

Hang (Harvey) Yu 毕业于伊利诺伊大学厄巴纳 – 尚佩恩分校并获得了计算生物物理学博士学位和统计学硕士学位。他在数据挖掘、机器学习和统计方面都有着丰富的经验。过去，作为学术工作的一部分，Harvey 涉及的领域包括随机过程模拟和时间序列（使用 C 和 Python）。他着迷于算法和数学建模。之后他进入了数据分析领域。

他目前是硅谷的一名数据科学家。他对数据科学满怀热情。他基于 R 中的优化方法和预测模型开发了一些统计 / 数学模型。在这之前 Harvey 在 ExxonMobil 实习。

不编程时，他会踢足球，阅读科幻书籍，或者听古典音乐。可以通过 hangyu1@illinois.edu 或者 Linkedln 网址 www.linkedin.com/in/hangyu1 来联系他。

Preface 前 言

作为一种统计计算、数据分析和绘图环境，自从 2000 年 1.0 版本问世以来，R 的流行度获得了指数级的增长。一些电子表格用户想要完成电子表格软件无法实现的功能，或需要处理的数据量大到电子表格软件无法方便地完成，他们寄希望于 R。类似地，商业分析软件用户也被这个免费且强大的选项所吸引。于是，一大群人目前正寄希望于用 R 快速处理事务。

作为一个可扩展的系统，R 的功能分布在众多的包中，每一个包囊括了大量函数。即使是经验丰富的使用者也很难将所有的细节记在脑海中。本书旨在为已有一定基础的 R 用户提供现成的方法来实现很多重要的数据分析任务。当面对一个特定任务时，用户可以在几分钟内找到合适的方法并实施，而不必在互联网或众多书籍中苦苦搜索。

本书涵盖以下内容

第 1 章涵盖了进行真正的数据分析任务之前的准备工作。本章提供了从不同源文件格式中读取数据的方法。此外，在实际分析数据前，我们执行了几个预处理和数据清洗步骤，本章还提供了以下任务的处理方法：处理缺失值和重复值、数值的缩放或标准化、在数值型变量和分类变量之间的转换，以及创建哑变量。

第 2 章讨论了分析师在实施特定的分析手段之前常用来理解数据的几种做法。本章提供了用于汇总数据、分割数据、抽取子集和建立随机数据分块的方法，也提供了使用标准化图像来展现潜在模式的方法，还提供了使用 `lattice` 和 `ggplot2` 包绘图的方法。

第 3 章涵盖了运用分类技术的方法。本章包括分类树、随机森林、支持向量机、朴素贝叶斯、K 最近邻、神经网络、线性和二次判别分析，以及逻辑回归。

第 4 章是关于回归技术的方法。本章包括 K 最近邻、线性回归、回归树、随机森林和神经网络。

第 5 章介绍了数据简化的方法。本章提供了通过 K- 均值和系统聚类的聚类分析手段，同时也涵盖了主成分分析。

第 6 章包含了一些技巧，包括处理日期和日期 / 时间对象，创建时间序列对象并画图，时间序列的分解、滤波和平滑，以及执行 ARIMA 分析。

第 7 章讨论了社交网络。本章介绍如何通过公共 API 获取社交网络数据，创建、绘制社交网络图，并计算重要的网络度量指标。

第 8 章讨论了呈现分析结果的技术。本章包含以下方法：使用 R Markdown 和 knitR 来创建报告，通过使用 shiny 创建交互式应用使读者直接与数据进行交互，用 RPres 创建幻灯片。

第 9 章解决了面对大型数据时如何书写高效且简洁的 R 代码的问题。本章包含了通过 apply 系列函数、plyr 包和数据表来切割数据的方法。

第 10 章包含了开拓 R 在处理空间数据上的强大功能的主题。本章涵盖了以下方法：通过 RGoogleMaps 来获取 GoogleMaps 数据并且在其上添加自有数据，导入 ESRI 形状文件并绘图，从 maps 包中导入地图数据，利用 sp 包创建并绘制空间数据框对象。

第 11 章包含了 R 与其他系统的交互。本章包含了 R 与 Java、Excel、关系型数据库和非关系型数据库（分别以 MySQL 和 MongoDB 为例）之间的连接。

阅读须知

本书中的所有代码均在 R 3.0.2 (Frisbee Sailing) 版本和 3.1.0 (Spring Dance) 版本上测试通过。当安装或者载入某些包时，你也许会得到警告消息，提示你这些代码是为不同的版本编译的，不过这并不会实际影响本书中的任何代码。

本书面向的读者对象

本书非常适合于那些已经有一定的 R 基础，但尚无将 R 广泛用于各种数据分析的经验，同时希望快速入门分析任务的读者。本书有助于在下列几个方面提高分析技巧的人士：

- 实现高级分析并创建信息充实的专业图表。
- 熟练地从各种来源获取数据。
- 应用监督型和无监督型的数据挖掘技术。
- 使用 R 的功能来呈现专业的分析报告。

每章的内容安排

在本书中，你会发现有几个标题是频繁出现的（准备就绪、要怎么做、工作原理、更多细节、参考内容）。

为了让读者在完成一个方法时获得清晰的指导，我们采用了以下内容编排方式：

准备就绪

这一节会给出内容概述，并且会描述如何准备好本节所需的软件以及任何其他前期准备工作。

要怎么做

这一节包含了完成方法所需的步骤。

工作原理

这一节通常包含了前一节中每一步的具体解释。

更多细节

这一节包含了关于所用方法的额外信息，以便让读者获得一个更加全面的认识。

参考内容

这一节提供了其他有用信息的链接。

本书约定

在本书中，你会发现我们使用不同类型的字体来区分不同类型的信息。下面有一些例子和解释。

文字形式的代码、数据库表名、文件夹名、文件名、文件扩展名、路径名、URL、用户输入和 Twitter 账户展示如下：

“函数 `read.csv()` 从 `.csv` 文件的数据中创建了一个数据框。”

代码块写成如下形式：

```
> names(auto)

[1] "No"           "mpg"          "cylinders"
[4] "displacement" "horsepower"   "weight"
[7] "acceleration" "model_year"   "car_name"
```

命名行的输入和输出写成如下格式：

```
export LD_LIBRARY_PATH=$JAVA_HOME/jre/lib/server
export MAKEFLAGS="LDFLAGS=-Wl,-rpath $JAVA_HOME/lib/server"
```

提示　小提示和小技巧出现在这里。

本书相关资源下载

下载代码范例和数据

本书提供代码范例和数据下载，读者可登录华章网站（www.hzbook.com）关于本书的页面获取相关资源。

关于本书中用到的数据

我们已经生成了本书中用到的很多数据文件。我们也使用了一些公开可获取的数据集。下表列出了这些公开的数据集。大部分公开数据集来自于加州大学欧文分校的机器学习库http://archive.ics.uci.edu/ml/。表中我们用“下载自 UCI-MLR”来标志这些数据集。

数据文件名	来　源
auto-mpg.csv	Quinlan, R. Combining Instance-Based and Model-Based Learning, Machine Learning Proceedings on the Tenth International Conference 1993, 236-243, held at University of Massachusetts, Amherst published by Morgan Kaufmann. (下载自 UCI-MLR)
BostonHousing. csv	D. Harrison and D.L. Rubinfeld, Hedonic prices and the demand for clean air, Journal for Environmental Economics a Management, pages 81–102, 1978. (下载自 UCI-MLR)
daily-bike- rentals.csv	Fanaee-T, Hadi, and Gama, Joao, Event labeling combining ensemble detectors and background knowledge, Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg. (下载自 UCI-MLR)
banknote- authentication. csv	• 数据库来源 : Volker Lohweg, University of Applied Sciences, Ostwestfalen-Lippe • 数据库捐赠 : Helene Darksen, University of Applied Sciences, Ostwestfalen-Lippe (下载自 UCI-MLR)
education.csv	Robust Regression and Outlier Detection, P. J. Rouseeuw and A. M. Leroy, Wiley, 1987. (下载自 UCI-MLR)
walmart.csv walmart- monthly.csv	下载自 Yahoo! 金融
prices.csv	下载自美国劳工统计局
infy.csv, infy- monthly.csv	下载自 Yahoo! 金融
nj-wages.csv	下载自新泽西州教育部网站以及 http://federalgovernmentzipcodes.us.
nj-county-data. csv	改编自维基百科：http://en.wikipedia.org/wiki/List_of_counties_in_New_Jersey

下载本书中的彩色图片

我们为你提供了本书中用到的截图和图表的彩色 PDF 文件。这些彩图有助于你更好的理解输出中的变化。可以从 https://www.packtpub.com/sites/default/files/ downloads/9065OS_ColorImages.pdf 下载这个文件，也可以登录华章网站获取相关内容。

目　录 *Contents*