· 大数据时代的R语言 ·

初学者，零基础你也大胆往前走

进阶者，R的金矿大门为你打开

高阶者，开源算法任你玩国际范

实战者，精辟案例助你融会贯通

Broadview
www.broadview.com.cn

ATAGURU
数据之巅

数据分析：R语言实战

李诗羽　张飞　王正林　编著

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

· 大数据时代的R语言 ·

数据分析：R语言实战

李诗羽　张飞　王正林　编著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

大数据时代，数据成为决策最为重要的参考之一，数据分析行业迈入了一个全新的阶段。R 是一款非常优秀的统计分析软件，本书侧重于使用 R 进行数据的处理、整理和分析，重点讲述了 R 的数据分析流程、算法包的使用以及相关工具的应用，同时结合大量精选的数据分析问题对 R 软件进行科学、准确和全面的介绍，以便使读者能深刻理解 R 的精髓和灵活、高效的使用技巧。

通过本书，读者不仅能掌握使用 R 及相关的算法包来快速解决实际问题，而且能学会从实际问题分析入手，到利用 R 进行求解，以及对结果进行分析。

本书可作为计算机、互联网、机器学习、信息、数学、经济金融、管理、运筹、统计以及有关理工科专业的本科生、研究生的学习用书，也能帮助市场营销、金融、财务、人力资源管理人员及产品经理解决实际问题，还能帮助从事咨询、研究、分析行业的人士及各级管理人员提高专业水平。

图书在版编目（CIP）数据

数据分析：R 语言实战 / 李诗羽，张飞，王正林编著. —北京：电子工业出版社，2014.8
（大数据时代的 R 语言）
ISBN 978-7-121-23714-0

Ⅰ. ①数… Ⅱ. ①李… ②张… ③王… Ⅲ. ①统计数据—统计分析②程序语言—程序设计 Ⅳ. ①O212.1 ②TP312

中国版本图书馆 CIP 数据核字（2014）第 147847 号

策划编辑：张月萍
责任编辑：刘 舫
印 刷：三河市鑫金马印装有限公司
装 订：三河市鑫金马印装有限公司
出版发行：电子工业出版社
北京市海淀区万寿路 173 信箱 邮编：100036
开 本：787 × 980 1/16 印张：21.00 字数：521 千字
版 次：2014 年 8 月第 1 版
印 次：2014 年 8 月第 1 次印刷
定 价：59.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei. com.cn。

服务热线：（010）88258888。

前 言

大数据时代，数据成为决策最为重要的参考之一，数据分析随着大数据概念的普及而日益得到重视，数据分析行业迈入了一个全新的阶段。

数据分析的软件如雨后春笋般地涌现，其中 R 软件的发展备受瞩目。R 是一个免费开源软件，它提供了首屈一指的统计计算和绘图功能，尤其是大量的统计分析、数据挖掘方面的算法包，使得它成为一款优秀的、不可多得的数据分析工具软件。

本书的主要目的是向读者介绍如何用 R 进行数据分析，通过大量的精选实例，循序渐进、全面系统地讲述 R 在数据分析领域的应用。

全书分为 15 章，主要内容从数据分析的前期准备、基本分析及应用和综合实例这三篇展开。

（1）上篇　数据分析的前期准备

由第 1~3 章组成，首先简要介绍数据分析的原则、步骤和过程，常用工具及 R 在数据分析中的优势，然后介绍 R 中数据整理等数据预处理的基本函数及方法。这些内容是使用 R 进行数据分析的最基础内容。

（2）中篇　基本分析及应用

由第 4~13 章组成，主要讲述数据分析的基本算法及应用，包括数据的图形描述、描述性分析、参数估计、假设检验，以及方差分析、回归分析、主成分分析、典型相关分析、判别分析、聚类分析和时间序列分析等，这些分析方法也是数据分析中使用得最多、最普遍的算法。R 中提供了丰富的、功能强大的算法包和实现函数，数据分析的初级和中级用户务必掌握。

（3）下篇　综合实例

由第 14~15 章组成，主要结合两个大例子，综合讲述数据分析在金融数据分析和数据预测中的应用，以及如何使用 R 中的方法和工具进行应用。对于中高级的用户，可以深入学习一下。

R 的特点是入门非常容易，使用也非常简单，因此本书也不需要读者具备 R 和数据挖掘的基础知识，不管是 R 初学者，还是熟练的 R 用户都能从书中找到对自己有用的内容，从而快速入门和提高。读者既可以把本书作为学习如何应用 R 的一本优秀教材，也可以作为数据分析的工具书。

全书以实际问题、解决方案和对解决方案的讨论为主线来组织内容，脉络清晰，并且各章自

成体系。读者可以从头至尾逐章学习，也可以根据自己的需要进行学习，找到自己实际问题的解决方案。

本书所编的源程序，都通过了反复的调试，读者可在 www.broadview.com.cn/23714 网站下载，方便读者使用。

本书主要由李诗羽、张飞、王正林编写，其他参与编写的人员有肖静、邹术来、夏路生、钟救元、郑曙霞、王成、刘亚文、肖绍英、王伟欣、朱桂莲、夏立德、王龙跃等。在此对所有参与编写的人员表示感谢！对关心、支持我们的读者表示感谢！

由于时间仓促，作者水平和经验有限，书中错漏之处在所难免，敬请读者指正，我们的电子邮箱是：wa_2003@126.com。

编著者

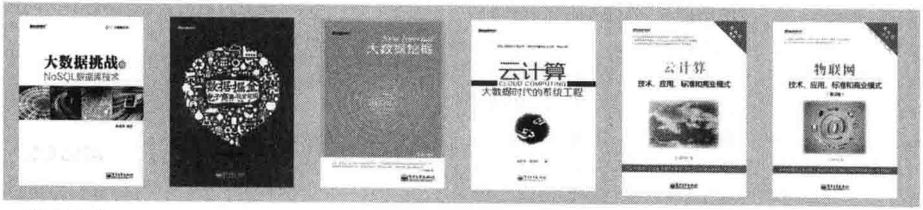
2014 年 5 月 28 日于北京

博文视点精品图书展台

专业典藏

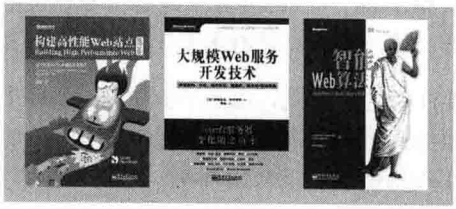
移动开发

大数据·云计算·物联网

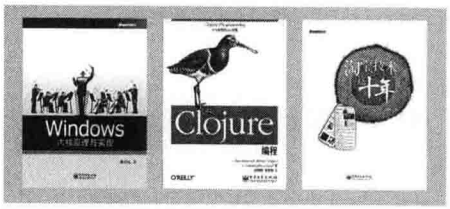


数据库

Web开发



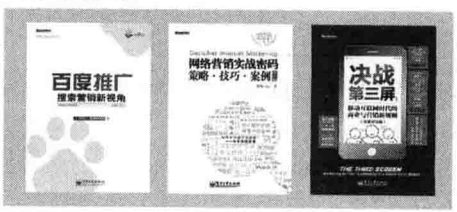
程序设计



软件工程

办公精品

网络营销



目 录

第 0 章 致敬，R！ 1

致敬，肩膀！ 1

致敬，时代！ 3

致敬，人才！ 3

致敬，R 瑟！ 5

上篇 数据分析的前期准备

第 1 章 数据分析导引 8

1.1 数据分析概述 8

1.1.1 数据分析的原则 8

1.1.2 数据分析的步骤 9

1.1.3 数据分析的过程 10

1.1.4 数据分析的对象 11

1.2 大数据分析 11

1.2.1 大数据分析的流程 11

1.2.2 大数据分析的基本方面 12

1.2.3 大数据分析的应用 13

1.3 数据分析常用工具 13

1.4 R 在数据分析中的优势 14

第 2 章 数据的读取与保存 16

2.1 数据读取 16

2.1.1 读取内置数据集 16

2.1.2 读取文本文件 17

2.1.3 读取固定宽度格式的文件 20

2.1.4 读取 Excel 数据 21

2.1.5 读取数据库文件 22

2.1.6 读取网页数据 26

2.1.7 读入 R 格式的文件 28

2.1.8 从其他统计软件读入数据 28

2.2 数据保存 31

2.2.1 使用函数 cat() 31

2.2.2 保存为文本文件 32
2.2.3 保存 R 格式文件 33
2.2.4 保存为其他类型文件 33

第 3 章 数据预处理 34

3.1 基本函数 34
3.2 数据修改 38
3.2.1 修改数据标签 38
3.2.2 行列删除 38
3.3 缺失值处理 38
3.3.1 判断缺失数据 39
3.3.2 判断缺失模式 39
3.3.3 处理缺失数据 41
3.4 数据整理 44
3.4.1 数据合并 44
3.4.2 选取数据的子集 46
3.4.3 数据排序 47
3.5 长宽格式的转换 48
3.5.1 揉数据函数 48
3.5.2 揉数据的最佳伴侣 49

中篇 基本分析及应用

第 4 章 数据的图形描述 54

4.1 R 绘图概述 54
4.2 绘图区域分割 55
4.2.1 函数 par() 55
4.2.2 函数 layout() 56
4.2.3 函数 split.screen() 57
4.3 二维图形 58
4.3.1 高级绘图函数 58
4.3.2 多元数据绘图 61
4.3.3 低级绘图函数 63
4.3.4 图形美化 64
4.3.5 交互式绘图命令 65
4.4 三维图形 67
4.5 lattice 程序包 69
4.6 ggplot2 程序包 73
4.6.1 快速绘图 74
4.6.2 分图层绘图 76

4.7 图形保存 84
4.8 实战实例：数据地图 84

第 5 章 数据的描述性分析 88

5.1 R 内置的分布 88
5.2 集中趋势的分析 90
5.2.1 集中趋势的测度 90
5.2.2 R 语言实现 91
5.3 离散趋势的分析 93
5.3.1 离散趋势的测度 93
5.3.2 R 语言实现 94
5.4 数据的分布分析 95
5.4.1 分布情况的测度 95
5.4.2 R 语言实现 96
5.5 图形分析及 R 实现 97
5.5.1 直方图和密度函数图 97
5.5.2 QQ 图 98
5.5.3 茎叶图 100
5.5.4 箱线图 100
5.5.5 经验分布图 102
5.6 多组数据分析及 R 实现 102
5.6.1 多组数据的统计分析 102
5.6.2 多组数据的图形分析 103

第 6 章 参数估计及 R 实现 112

6.1 点估计及 R 实现 112
6.1.1 矩估计 112
6.1.2 极大似然估计 116
6.2 单正态总体的区间估计 122
6.2.1 均值 μ 的区间估计 122
6.2.2 方差 σ^2 的区间估计 125
6.3 两正态总体的区间估计 126
6.3.1 均值差 $\mu_1-\mu_2$ 的区间估计 127
6.3.2 两方差比 σ_1^2/σ_2^2 的区间估计 130
6.4 关于比率的区间估计 131

第 7 章 假设检验及 R 实现 134

7.1 假设检验概述 134
7.1.1 理论依据 135
7.1.2 检验步骤 135

7.1.3 两类错误 136
7.2 单正态总体的检验 137
7.2.1 均值 μ 的检验 138
7.2.2 方差 σ^2 的检验 141
7.3 两正态总体的检验 142
7.3.1 均值差 $\mu_1 - \mu_2$ 的检验 143
7.3.2 成对数据的 t 检验 146
7.3.3 两总体方差的检验 147
7.4 比率的检验 148
7.4.1 比率的二项分布检验 148
7.4.2 比率的近似检验 149
7.5 非参数的检验 149
7.5.1 总体分布的 χ^2 检验 150
7.5.2 Kolmogrov-Smirnov 检验 153

第 8 章 方差分析及 R 实现 157
8.1 单因素方差分析及 R 实现 157
8.1.1 基本假设的检验 157
8.1.2 单因素方差分析 160
8.1.3 多重 t 检验 164
8.1.4 Kruskal-Wallis 秩和检验 166
8.2 双因素方差分析及 R 实现 168
8.2.1 无交互作用的分析 169
8.2.2 有交互作用的分析 172
8.3 协方差分析及 R 实现 176

第 9 章 回归分析及 R 实现 180
9.1 一元线性回归 180
9.1.1 模型理论 180
9.1.2 显著性检验 181
9.1.3 R 语言实现 181
9.2 多元线性回归 187
9.2.1 模型理论 187
9.2.2 显著性检验 188
9.2.3 R 语言实现 189
9.2.4 逐步回归 192
9.3 回归诊断及 R 实现 194
9.3.1 残差诊断 195
9.3.2 影响分析 198
9.3.3 多重共线性诊断 201

9.4 岭回归及 R 实现 203
9.5 广义线性模型 206
9.5.1 模型理论 206
9.5.2 R 语言实现 207

第 10 章 主成分分析与因子分析 211

10.1 主成分分析 211
10.1.1 理论基础 211
10.1.2 R 语言实现 215
10.2 因子分析 221
10.2.1 理论模型 221
10.2.2 因子载荷矩阵的估计方法 223
10.2.3 R 语言实现 225

第 11 章 典型相关分析和对应分析 230

11.1 典型相关分析 230
11.1.1 理论基础 230
11.1.2 典型相关分析的应用 232
11.1.3 R 语言实现 233
11.2 对应分析 236
11.2.1 理论基础 236
11.2.2 对应分析的步骤 237
11.2.3 R 语言实现 238

第 12 章 判别分析和聚类分析 242

12.1 判别分析及 R 实现 242
12.1.1 距离判别法 243
12.1.2 距离判别法的 R 实现 244
12.1.3 Fisher 判别法 247
12.1.4 Fisher 判别法的 R 实现 248
12.1.5 贝叶斯判别法 251
12.1.6 贝叶斯判别法的 R 实现 252
12.2 聚类分析及 R 实现 252
12.2.1 理论概述 253
12.2.2 R 实现举例 254

第 13 章 时间序列分析及 R 实现 260

13.1 时间序列的基本分析 260
13.1.1 平稳性与非平稳性 260
13.1.2 R 实现的基本步骤 261
13.2 时间序列的分解 262

13.2.1 分解非季节性数据 263
13.2.2 分解季节性数据 265
13.3 指数平滑法预测分析 268
13.3.1 简单指数平滑法 269
13.3.2 残差的白噪声检验 272
13.3.3 Holt 指数平滑法 275
13.3.4 Winters 指数平滑法 277
13.4 ARIMA 模型分析 280
13.4.1 基本思想 280
13.4.2 平稳化处理 281
13.4.3 建模 282
13.4.4 模型的参数估计 284
13.4.5 模型预测及检验 284

下篇 综合实例

第 14 章 R 在金融数据分析中的应用 288
14.1 投资组合最优化实例 288
14.1.1 概述 288
14.1.2 均值-方差模型 289
14.1.3 模拟退火算法 292
14.2 构造投资组合的有效前沿 298
14.2.1 R 中的算法包 298
14.2.2 计算分析 298
14.3 股票聚类分析 301
14.3.1 概述 301
14.3.2 K-means 聚类分析 302
14.3.3 层次聚类分析 304

第 15 章 R 在数据预测中的应用 306
15.1 回归分析预测 306
15.1.1 概述 306
15.1.2 实战案例 306
15.2 时间序列预测 318
15.2.1 概述 318
15.2.2 实战案例 318

第0章

致敬，R！

此时，你一定想知道，书的封面上停着一只什么鸟？

那我告诉你，那是 Robin 鸟，中文名叫知更鸟，它可大有来头，是英国的国鸟，以羽毛颜色漂亮招人喜爱著称。

我把它放在封面，首先是借用其名字首字母 R，来表示 R 语言。最重要的是，我想到了股神巴菲特的一句关于知更鸟的名言，我想双关暗示一下——如果你还不学一些 R，大数据对你来说就快结束了。

如果你想等到知更鸟报春，那春天就快结束了。——巴菲特

So if you wait for the robins, spring will be over. ——*Warren Edward Buffett*

如果你想快速成功

你最好站在一个高的肩膀上

如果你想驾驭大数据时代

你最好懂点数据挖掘

如果你想玩转数据挖掘

你最好先玩转 R！

致敬，肩膀！

可能当我们还是三好小学生的时候，我们就知道，牛顿是站在巨人的肩膀上的，现如今，我们都知道，中国所有的“二代”，不是站在老爹的肩膀上，就是踩在老丈人的肩膀上的。不得不承认，脚下的肩膀有时候是很牛的。

当你走进数据挖掘，当你走进 R 的世界，你会发现，R 的脚下也有一个肩膀，有肩膀的 R 也

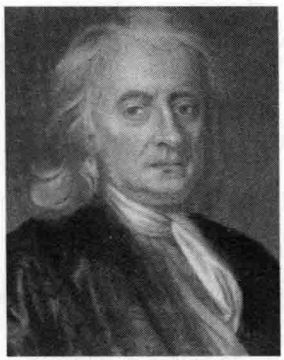
是很牛的！

R 的肩膀，是谷歌首席经济学家范里安先生发现的，先生说了好几句活，我只记住了这句“使用 R，你已经站在了巨人的肩膀上”。

在此，我只想致敬一下肩膀，与“二代”无关！

我之所以能取得现在的成就，是因为我站在巨人的肩膀上。——牛顿

If I have seen further it is by standing on the shoulders of giants. ——*Isaac Newton*

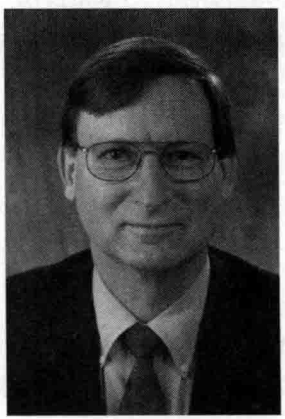


艾萨克·牛顿爵士（Isaac Newton，1643.12.25—1727.3.20），
英国数学家、物理学家、天文学家和经典力学体系奠基人。

R 的最美之处在于，你能够通过修改很多牛人预先编写好的包的代码，解决你想解决的各种问题，因此，事实上，使用 R，你已经站在了巨人的肩膀上。——哈尔·罗纳德·范里安

The great beauty of R is that you can modify it to do all sorts of things. And you have a lot of prepackaged stuff that's already available, so you're standing on the shoulders of giants.

——*Hal Ronald Varian*



哈尔·罗纳德·范里安（Hal Ronald Varian），
谷歌首席经济学家，美国著名研究微观经济学和信息经济学学者。

致敬，时代！

“大数据”一词，最早是全球知名咨询公司麦肯锡提出来的，“数据，已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。人们对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来。”

我们，已经身处大数据时代了，对于做数据挖掘、用 R 的我们来说，好时代来了！

“大数据”时代已经降临，在商业、经济及其他领域中，决策将日益基于数据和分析而做出，而并非基于经验和直觉。

——摘自《纽约时报》，2012 年 2 月的一篇专栏

摘自《纽约时报》，How Big Data Became So Big 一文

“在美国具备高度分析技能的人才（大学及研究生院中学习统计和机器学习专业的学生）供给量，2008 年为 15 万人，预计到 2018 年将翻一番，达到 30 万人。然而，预计届时对这类人才的需求将超过供给，达到 44 万～49 万人的规模，这意味着将产生 14 万～19 万的人才缺口。仅仅四五年前，对数据科学家的需求还仅限于 Google、Amazon 等互联网企业中。然而在最近，重视数据分析的企业，无论是哪个行业，都在积极招募数据科学家，这也令人手不足的状况雪上加霜。”

——摘自麦肯锡全球研究院的报告 Big data: The next frontier for innovation, competition and productivity（大数据：未来创新、竞争、生产力的指向标），2011.5

……2017 年大数据技术和服务市场将增至 324 亿美元，实现 27%的年复合增长率。……大数据不仅是新兴行业，也是市场的主要驱动力，它正在酿成一个主要的市场。

——摘自国际数据公司 IDC 的预测报告 Worldwide Big Data Technology and Services 2013–2017 Forecast，2013.12

致敬，人才！

Google 首席经济学家范里安先生，在 2008 年 10 月与麦肯锡总监 James Manyika 先生的对话

中，曾经讲过下面一段话：“我总是说，在未来 10 年里，从事最有意思的工作的人将是统计学家。人们都认为我在开玩笑。但是，过去谁能想到电脑工程师会成为 20 世纪 90 年代从事最有趣的工作的人？在未来 10 年里，获取数据——以便能理解它、处理它、从中提取价值、使其形象化、传送它——的能力将成为一种极其重要的技能，不仅在专业层面上是这样，而且在教育层面（包括对中小学生、高中生和大学生的教育）也是如此。由于如今我们已真正拥有实质上免费的和无所不在的数据，因此，与此互补的稀缺要素是理解这些数据并从中提取价值的能力。”

范里安教授在当初的对话中使用的是 statisticians（统计学家）一词，虽然当时他没有使用数据科学家这个词，但这里所指的，正是现在我们普遍所指的数据科学家。

对数据科学家的关注，源于大家逐步认识到，Google、Amazon、Facebook 等公司成功的背后，存在着这样一批专业人才。这些互联网公司对于大量数据不是仅进行存储而已，而是将其变为有价值的金矿——例如，搜索结果、定向广告、准确的商品推荐、可能认识的好友列表等。

仅仅在几年前，数据科学家还不是一个正式确定的职业，然而一眨眼的工夫，这个职业就已经被誉为“今后 10 年 IT 行业最重要的人才”了。

摘自 The Emerging Role of the Analyst 一文

在国外，据统计，目前世界 500 强企业中，有 90%以上都建立了数据分析部门。IBM、微软、Intel 等公司也积极投资数据业务，建立大数据部门，培养数据分析团队。

美国的小伙伴们，在数据挖掘、数据科学等方面比我们下手早。2011 年，美国的加州大学伯克利分校开始开设《数据科学导论》课程；伊利诺伊大学香槟分校从 2011 年起举办“数据科学暑期研究班”；哥伦比亚大学从 2013 年起开设《应用数据科学》课程，并从 2013 年起开设相关培训项目，还计划从 2014 年起设立硕士学位，2015 年设立博士学位；纽约大学从 2013 年秋季起设立“数据科学”硕士学位；在英国，邓迪大学从 2013 年起设立“数据科学”硕士学位……

怎么办，那就自学吧，从 R 开始，站上那个肩膀，做今后 10 年最重要的人才吧！

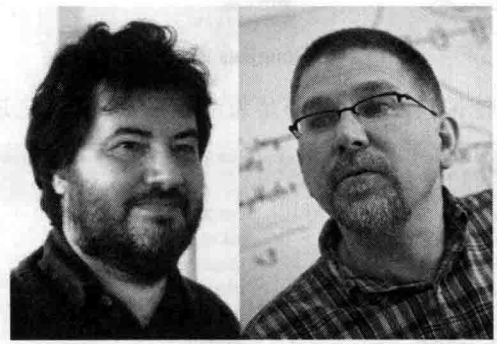
致敬，R 瑟！

1976 年，John Chambers 在贝尔实验室开发的 S 语言是为了替代昂贵的 SPSS 和 SAS 工具。如果说 S 是 VAX 和 UNIX 小型机时代的产物，那么 R 则是 PC 和 Linux 时代的产物，R 语言大量借用了 S 语言的方法。

1992 年，新西兰奥克兰大学的两位统计学教授，两位“R 姓”先生（R Sir，“R 瑟”）Ross Ihaka 和 Robert Gentleman 成为了同事，为了方便教授初等统计课程，这哥儿俩开发了一种语言，而恰巧他们名字的首字母都是 R，于是 R 便成为这门语言的名称。

这两位 R 教授也是 R 开发团队的核心成员，值得注意的是，S 语言的发明者 John Cambers 也是 R 开发团队的成员，因此不难理解 R 语言的一些数据处理路径与 S 语言相同。

R 可以看作 S 的一种实现，Insightful 公司开发的 S-PLUS 也是 S 的实现版本，2004 年 Insightful 把 S-PLUS 授权给了朗讯科技，后来又被 Tibco 软件于 2008 年收购。



R 语言的发明者 Ross Ihaka 和 Robert Gentleman

与 S 和 S-PLUS 不同的是，R 并不是象牙塔里炮制出的代码，而是一个由分析师和程序员构成的社区的产物，这个社区为处理各种数据集创建了超过 5000 个函数包和 2500 个插件。

今天，根据 Revolution Analytics 的统计，R 被全球超过 200 万个量化分析师采用。Revolution Analytics 成立于 2007 年，并开发出了 R 的并行实现，该公司采用了开放内核的方式开发 R，为开源软件包推广商业支持，同时扩展 R 环境，提升其在计算机集群上的表现，并将其与 Hadoop 集群对接。

在 2013 年中，数据挖掘专业网站 KDnuggets 做了一个关于“什么样的程序或者统计语言是你在做分析、挖掘、科学计算的时候所需要的？”的调查。

调查结果是：最受欢迎的是 R 语言（61%的调研会员在用），然后是 Python(39%)、SQL(37%)等，每个调研对象平均使用 2～3 种语言。