R语言数据高效处理指南

黄天元◎著

北京大学出版社
PEKING UNIVERSITY PRESS

R语言
数据高效处理指南

黄天元◎著

北京大学出版社
PEKING UNIVERSITY PRESS

内 容 提 要

R 语言在近 10 年来已经发生了日新月异的变化，不仅在内容上更加丰富多彩，而且在计算效率上也有了大幅的提升。它被更加广泛地用于数据可视化、统计建模、机器学习等领域，而且还能实现网络爬虫、网络应用开发等功能，成为数据科学领域的全能型工具。R 语言在学术界的地位已经不容置疑，在大数据时代中它是保证研究可重复性的重要工具。随着功能的日益完善，R 语言已经进军工业界，并在金融、保险、医疗、生物和信息计量等不同的应用场景中大放异彩，潜力不可估量。

尽管 R 语言能够实现丰富多样的实际功能和框架，但是其本质是面向数据的，因此数据处理是 R 语言核心中的核心。如果能够掌握高效的数据操作技术，就能够在各类数据分析任务中如鱼得水。本书定位即为“R 语言数据处理 101”，希望 R 语言的使用者能够在较早的阶段就习得最基本而有效的数据处理基本技术。

本书读者群体包括在校的大学生、数据分析从业人员和致力于更加高效地处理数据的所有的 R 语言使用者。尽管对数据科学、计算机编程、统计学有一定基础会帮助理解本书的内容，但这不是必需的，来自包括初学者在内的各个层次的读者群体都能从本书中有所收获。读者在本书中不仅能够学到数据处理中的实用技术，还能培养在数据分析中的探索性思维。可以作为零基础学习数据分析的教程、进阶数据分析实用技巧的参考书、常备查询的案头工具书，以及具有一定趣味性的数据分析入门启蒙书。

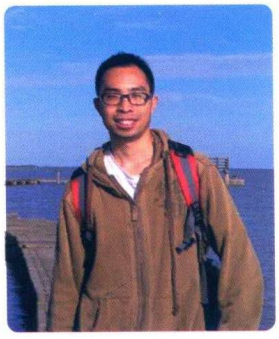
图书在版编目(CIP)数据

R语言数据高效处理指南 / 黄天元著. —北京 : 北京大学出版社, 2019.9
ISBN 978-7-301-30608-6

Ⅰ. ①R… Ⅱ. ①黄… Ⅲ. ①程序语言—程序设计—指南②数据处理—指南 Ⅳ. ①TP312-62②TP274-62

中国版本图书馆CIP数据核字(2019)第168616号

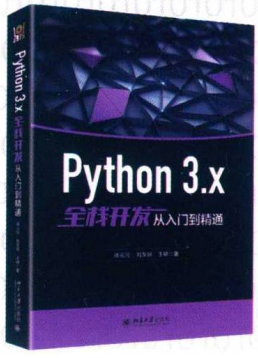
书　　名	**R语言数据高效处理指南** R YUYAN SHUJU GAOXIAO CHULI ZHINAN
著作责任者	黄天元　著
责任编辑	吴晓月　王蒙蒙
标准书号	ISBN 978-7-301-30608-6
出版发行	北京大学出版社
地　　址	北京市海淀区成府路205 号　100871
网　　址	http://www.pup.cn　　新浪微博: @ 北京大学出版社
电子信箱	pup7@ pup.cn
电　　话	邮购部 010-62752015　发行部 010-62750672　编辑部 010-62580390
印 刷 者	北京大学印刷厂
经 销 者	新华书店
	787毫米×1092毫米　16开本　13.25印张　309千字
	2019年9月第1版　2019年9月第1次印刷
印　　数	1-4000册
定　　价	59.00元

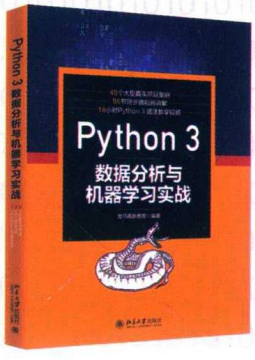


作者简介

黄天元，复旦大学博士在读，R语言忠实爱好者。热爱数据科学与开源工具，致力于利用数据科学迅速积累行业经验和科学知识，涉猎内容包括信息计量、机器学习、数据可视化、统计建模、知识图谱等。已发表SCI论文两篇，开设有知乎专栏“R语言数据挖掘”。

好书推荐





北京大学出版社
PEKING UNIVERSITY PRESS
第七事业部

智慧创造价值
品质铸就卓越

读者邮箱：2751801073@qq.com
投稿邮箱：pup7@pup.cn
出版咨询：010-62570390

前言

INTRODUCTION

在大数据时代，基本的数据分析知识是每一个人都必须了解和掌握的。数据分析很简单，其根本框架可以用一张思维导图来展现；数据分析很复杂，涉及数学知识、统计学知识、算法逻辑，以及硬件支持、软件实现和业务背景等内容。R 语言是实现数据分析的一大利器，作为科研工作者与开源技术爱好者，笔者用 R 语言做过数据清洗、数据批处理、数据可视化、网络爬虫、文本挖掘、社交网络分析、时空分析、机器学习和网页应用设计等各种项目。但是，万变不离其宗，R 语言的处理对象是数据，因此要想完成诸多数据任务，无一例外都要用到的最根本技术——数据操作。因此这方面的知识应该是学习 R 语言的初学者熟练掌握的内容。为了让广大的 R 语言用户能够“领先一步，领先一路”，本书对最基本的数据操作知识做了较详细的介绍。同时，也希望本书能吸引更多对数据分析感兴趣的读者，希望他们阅读之后能了解、学习并熟练掌握 R 语言，从而获得高效处理数据的基本技能。

R 语言的开源社区是一个技术交流极其活跃的地方，各种技术“大牛”都会在社区无私地分享自己辛勤劳动的成果，不仅避免了社区中的同行重复劳动，而且大大推进了数据分析技术的繁荣发展。笔者在开始学习 R 语言的时候，其实仅仅是个“调包侠”。也就是通过不断复制、粘贴高手的代码，来完成自己的任务。日积月累之后，笔者对 R 语言的整个数据分析体系越来越了解，为了对更加深入的问题进行探寻，决定系统地学习 R 语言。R 语言是一种越学越容易的语言，也就是当你学习新知识的时候，你会发现这些知识与你之前学习的知识非常相似，因此需要付出的时间相对较短。

本书的内容，最初仅是作为个人学习的记录，上传到了笔者课题组所维护的论坛中。后来笔者在开源社区学到了很多知识，就产生了把自己在数据分析中的经验和心得分享给大家的想法。那时候笔者就开始向 R 语言中文社区投稿，日积月累，写了不少帖子。但是将它们整理成书，还是遇

到了不少困难。因为笔者必须换位思考，要兼顾不同层次、不同领域读者的需求，这样才能写出一本适用性较强的书，才能够真正帮到最广泛的读者群体。本书力求成为“R 语言数据处理 101”，希望能够给初学者一个正确的引导。同时，本书也提到了很多详细的技术细节，对有一定经验的数据分析从业者也有较强的参考意义。

尽管笔者竭力将自己的研究和经验总结得全面、深入，但技术的发展日新月异，而个人的知识与能力终究有限，纰漏之处在所难免，希望各位读者不吝赐教，共同将本书打造得更完美。

本书所涉及的源代码及第 9 章的参考答案已上传到百度网盘，供读者下载。请读者关注封底“博雅读书社”微信公众号，找到“资源下载”栏目，根据提示获取。

目录

CONTENTS

第 1 部分
基础知识

万丈高楼平地起，越是希望随心所欲、灵活巧妙地处理数据，就越需要具备最扎实、最根本的基础知识。作为全书的第一部分，本章首先介绍什么是数据处理；然后对本书主要的实现工具——R 语言进行了概览式的讲解，力求让没有接触过 R 语言的读者也能够快速入门；最后会对数据基本操作的范式进行讲解，让大家对此有一个清晰的认识。

第1章

数据处理总论

“大数据”的概念在近几年被炒得很热，几乎家喻户晓。但是，其实在这个概念没有被提出或重视之前，数据处理的科学运用就已经充斥在人们生活、学习、工作的方方面面。中国古代由对农时的记录，订立的二十四节气，就是对周期性数据的记录和运用。商业兴盛之时，物质与财产的流动也会产生大量的数据，商人会雇用出纳、会计来对这些数据进行记录、整理、监督和核算。随着计算机技术的飞速发展，数据的存储和运算越来越便捷，运算方式由之前的利用算盘和账簿变成了利用计算机，但是数据处理的基本概念和核心价值是不会变的。本章首先介绍数据处理的定义，然后对其意义进行简单的探讨，最后对当代数据处理的实现工具进行介绍。

1.1 数据处理的定义

数据处理的基本目的是从大量杂乱无章、难以理解的数据中抽取有价值、有意义的数据，其基本内容包括对数据的采集、存储、检索、加工、变换和传输。这样的定义显然太过简单而宽泛，为了完善这个定义，我们先了解一下数据处理的子概念：数据预处理、数据清洗和 ETL。

数据预处理（data preprocessing）是指在主要处理以前对数据进行的一些处理。主要处理之前的处理就是预处理。那么，什么样的处理可以看作是主要的，什么样的处理可以看作是次要的？有数据挖掘或者建模经验的从业者应该知道，我们喜欢用数据作一些图（专业名称为数据可视化），或者做一些表格，然后揭示一定的道理。尽管如此，其实大量的时间不是花在作图和表格上，而是花在预处理上。没有高质量的数据，再华丽的模型也无济于事，业界、科研界都知道这么一个道理：垃圾进，垃圾出（garbage in, garbage out）。因此，对于实际生产、工作中不完整、不一致的杂乱数据，必须要进行预处理才能够进行数据挖掘。为了提高数据的质量，数据预处理技术应运而生，其方法包括数据清理、数据集成、数据变化、数据规约、数据审核、数据筛选和数据排序等。根据数据本身具有的特点，预处理技术种类也是丰富多样的。

数据清洗（data cleaning）是指发现并纠正数据文件中可识别错误的最后一道程序，包括检查数据一致性、处理无效值和缺失值等。输入后的数据清理一般是由计算机完成的，不需要手动操作。既然称为清洗，说明数据是“脏”的，因此才要按照一定的规则进行处理。数据清洗的任务是过滤

那些不符合要求的数据，将过滤的结果交给业务主管部门，确认是否应该过滤，还是应该由业务单位修正之后再进行抽取。不符合要求的数据主要包含不完整的数据、错误的数据、重复的数据三大类。数据出错的原因有很多，有的是因为数据采集操作不得当（例如，去做问卷调查时发现没有注意采样的性别比例，结果选的全部是女性或男性），有的是随机出现的状况（例如，想要知道客户的用电情况，结果选的日期恰好包含停电维修的日期），有的是因为人工失误（例如，手一抖，多加一个零）。无论何种情况，这些数据都是不能直接使用的，否则对最后的决策具有误导作用。

ETL，是英文 Extract Transform Load 的缩写，用来描述将数据从来源端经过抽取（Extract）、交互转换（Transform）、加载（Load）至目标端的过程。ETL 一词常用在数据仓库，但其对象并不限于数据仓库。ETL 是构建数据仓库的重要一环，用户从数据源抽取所需的数据，经过数据清洗，最终按照预先定义的数据仓库模型，将数据加载到数据仓库中。

上面提到的 3 个概念，既有重叠的部分，又有各自的特点。不过毫无疑问，三者都涵盖了数据处理的内容。本书中所讲的数据处理是指针对关系型数据模型的二维表数据结构所进行的各种读写变换。简单地说，就是针对表格数据进行的各种操作，包括筛选、排序、分组、汇总等。

1.2　数据处理的意义

为什么要进行数据处理？这个问题很好回答。因为数据都是零散的、不规整的、不符合要求的，为了把它们转化为可以直接使用的数据，必须进行数据处理。

例如，现在想要分析某高校男生的身高水平，但是拿到的表格数据中包括男生身高和女生身高，那么就需要进行数据处理，把只包含男生的数据筛选出来。听起来有点像 Excel 中的日常操作。再如，想要查看哪个商品卖得最好，但是表格中的数据没有任何规律可循，这时可以对商品成交量记录从大到小排序。这些操作，都是数据处理。

也就是说，数据处理最根本的意义在于，原始的表格数据没办法直接满足我们的需要。因此我们需要通过“魔法”般的处理技术，对数据进行变化，最终来满足应用需求。尽管目前机器学习、深度学习已经如日中天，但是基础的数据处理是永远不会被淘汰的技术，因为任何方法都必须尊重人类最原始的业务分析目标。在未来，会有自动参数搜索的算法，建模的过程可能会被弱化。但是如何对获得的原始数据进行整理，构造特征，人们还在不断地摸索。如何把这些零散的表格转化为结构化的数据，从而让它们能够产生实实在在的价值，是数据处理的根本意义所在。

1.3　数据处理基本工具

数据处理的工具非常多，相信读者或多或少地接触过其中的一种或几种。能够完成数据处理的软

件工具包括 Access、Excel、MATLAB、R、Python 和 Oracle 等。它们有的需要付费，有的则完全开源免费。从普及程度来说，相信使用 Windows 的用户大部分都会用 Excel，它是做数据处理的基本软件，里面有筛选、替换、排序等功能，交互式操作极其便捷。事实上，在“大数据”还没有兴起时，能够用 Excel 做透视表是求职时非常重要的加分项（也许现在还是）。但是随着科学技术的发展，各个机构、企业的数据越来越多，Excel 已无法满足多样化的需求了。

如果让笔者对这些工具进行分类，主要依据两个标准。一个标准是操作以什么为主，分鼠标流和键盘流，也就是以鼠标操作为主，还是以键盘操作为主。毫无疑问，Excel 是一个以鼠标操作为主的工具。尽管它有 VBA，能够支持使用函数进行结构化的查询，但只有少数用户能够掌握这个技能。对于轻量级的数据而言，这种能够采用鼠标拖曳来赋值的功能非常直观，而且在可视化方面，无论是做三线表还是一些简单的统计图，都相当便捷。但是鼠标的局限就在于，如果数据量庞大，只用鼠标就应付不过来了，这时键盘流就应运而生了。R 语言就属于键盘流，它的好处是在处理大型数据集的时候，能够更加灵活地处理数据。还有一点，目前 Office 套件的工具实在是太多了，光是记忆这些功能就非常困难，然后还要知道它们各自在什么位置，又要费一番工夫。R 语言就完全不同，它是完全基于命令行进行操作的。当学会了如何查找帮助文档之后，要用什么就找什么，通过加载相应的包来解决特定的问题。所以当看到 Excel 界面的时候也许会觉得工具太多了（图 1-1），而 R 语言竟然就像一张白纸一样（图 1-2）。

图 1-1　Excel 交互式界面

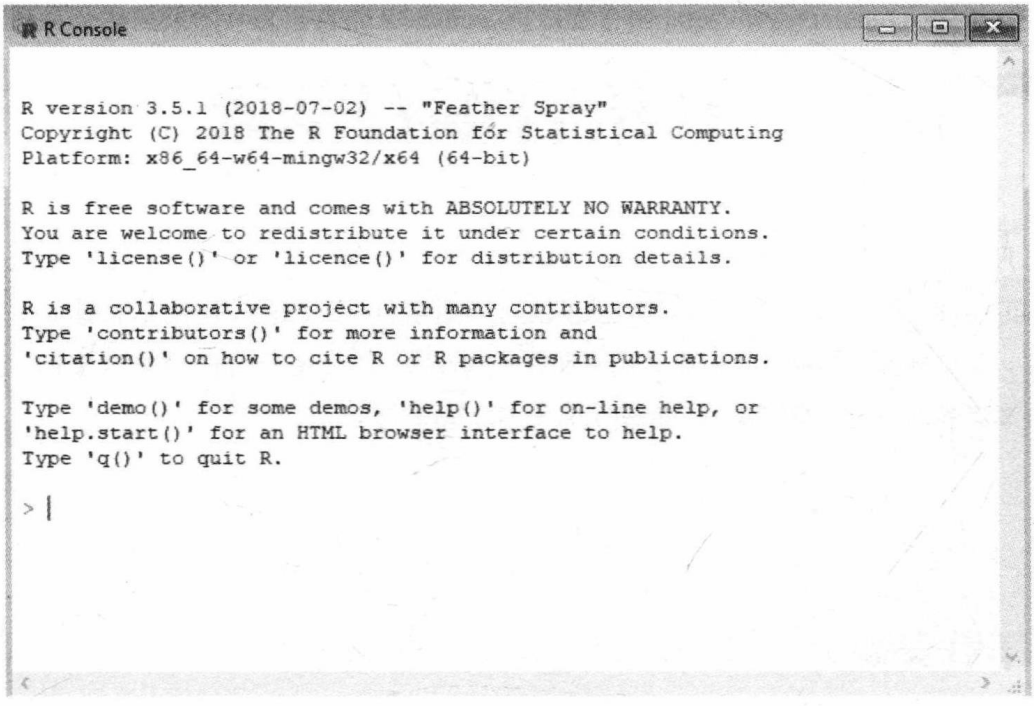


图 1-2　R 语言交互式界面

另一个分类标准，就是这些工具是否开源。开源的定义是用户能够利用源代码，并在此基础上修改、学习，不过开源是有版权的，同样受到法律保护。一个重要的区别是，开源软件是可以免费试用的，而闭源软件则需要付费使用。尽管家喻户晓的 Excel 在开机的时候已经装在了一部分计算机中，但是其实人们在买计算机时，就已经包含了购买 Windows 系统和 Office 工具的费用。也就是说，Excel 其实是需要付费使用的。付费的软件是由微软维护的，微软公司的工程师会不断开发软件的新功能。而开源的 R 语言则是开放社区自发维护的，也就是说如果开发者做了一个包感觉不错，并愿意分享给大家，那么他就可以把包发布到网上，所有人都可以使用。如果有 bug，大家可以及时指出来，开发者会不断优化这个包，或者有“大咖”直接在该包的基础上进行修改，然后发布到社区。

本书将要介绍的数据处理工具就是 R 语言，不过它的功能不仅限于数据处理。R 语言其实是 S 语言的衍生品。AT&T 贝尔实验室在 1980 年开发了 S 语言，当时的目标就是开发一款进行数据探索、统计分析和作图的解释型语言。语言不能脱离软件而存在，当时实现 S 语言的软件是 S-PLUS，它是一款商业软件。随后，新西兰奥克兰大学的志愿者开发了 S 语言的另一种实现，带头人是 Robert Gentleman 和 Ross Ihaka，后来把他们名字的首字母作为该语言的名称，R 语言就此诞生。R 语言最初的定位就是统计分析与可视化，语法通俗易懂。而且只要学会之后，能够充分利用前人积累的代码和软件包，迅速实现个性化的需求，从而达到“站在巨人的肩膀上”的效果。

第2章

R语言编程基础

凡是语言，皆为工具。C、Java、Python乃至英语，无一例外。工具是表现人类思维的手段，而不是目的。R语言是人们为了实现统计计算和数据可视化愿望而开发的工具，简单易用，能够大大提高人们实现算法的效率。本章将介绍R语言的基本概念，希望读者能够通过本章的学习掌握使用R语言的基本技巧。

2.1 下载安装

R软件是一款免费开源的软件，能够在包括Windows、Linux和macOS X在内的各种操作系统上运行。要安装R软件，首先要登录官网https://www.r-project.org/。首页就会显示下载的链接（图2-1的方框中显示的是通向下载页的下载链接），进入后会让我们选择镜像，我们选择中国（China）地区的镜像即可。下载后单击运行，整个安装过程基本是向导式的操作，进行相应选择设置，然后单击“下一步”按钮即可。用户可以自定义安装在任意的路径，可以选择安装32位或64位，也可以两者都安装。

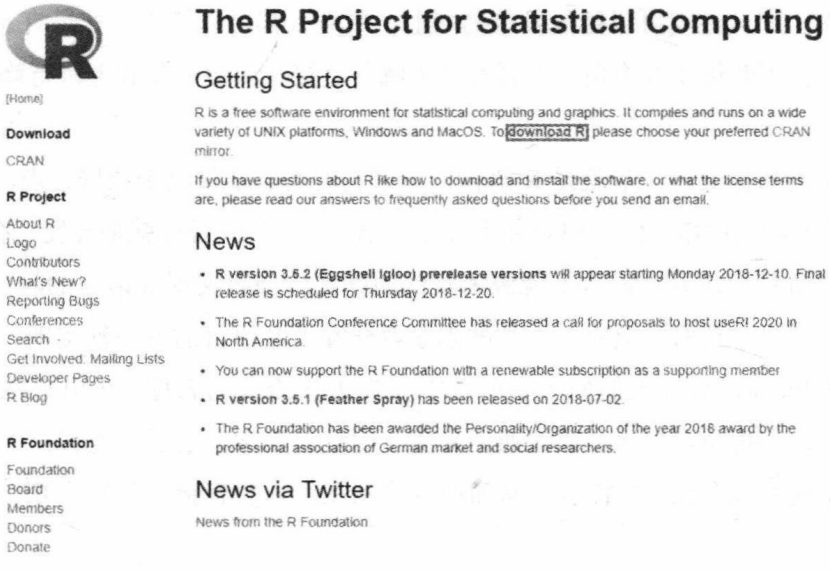


图2-1 R官网首页