PLAYING WITH R

DATA ANALITICAL THINKING TO PRACTICE

R语言

从数据思维到数据实战

朱雪宁　等 著

加州大学戴维斯分校管理学院杰出教授**蔡知令**

复旦大学大数据学院创始院长**范剑青**

微家实业CDO**李舰**

北京大学光华管理学院教授**王汉生**

统计之都网站创始人**谢益辉**

联袂推荐

中国人民大学出版社

PLAYING WITH R

DATA ANALITICAL THINKING TO PRACTICE

R语言

从数据思维到数据实战

朱雪宁　等 著

中国人民大学出版社

·北 京·

图书在版编目（CIP）数据

R语言：从数据思维到数据实战/朱雪宁等著. —北京：中国人民大学出版社，2018.12
ISBN 978-7-300-26311-3

Ⅰ.①R… Ⅱ.①朱… Ⅲ.①程序语言-程序设计 Ⅳ.①TP312

中国版本图书馆CIP数据核字（2018）第232138号

R语言：从数据思维到数据实战
朱雪宁 等著
R Yuyan：Cong Shuju Siwei Dao Shuju Shizhan

出版发行	中国人民大学出版社		
社　　址	北京中关村大街31号	**邮政编码**	100080
电　　话	010－62511242（总编室）		010－62511770（质管部）
	010－82501766（邮购部）		010－62514148（门市部）
	010－62515195（发行公司）		010－62515275（盗版举报）
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com（人大教研网）		
经　　销	新华书店		
印　　刷	北京宏伟双华印刷有限公司		
规　　格	170 mm×230 mm　16开本	**版　　次**	2018年12月第1版
印　　张	22.25插页1	**印　　次**	2018年12月第1次印刷
字　　数	300 000	**定　　价**	89.00元

推荐序一

王汉生（熊大）[①]

编程语言之于数据分析是必不可少的。对于一个数据科学的新兵，应该从哪门语言开始？摆在面前的选择很多：R，SAS，Python，C，JAVA，甚至 Fortran。它们各有优势，也有不足。如果一定要选一个，我推荐 R。有两个重要原因：第一，R 是免费的，全球镜像，非常方便。第二，R 的分析建模能力很强，部分得益于基础模块的完善，部分得益于整个统计学社区的支持。很多最新的分析方法、统计模型都是用 R 首先实现，并被开发封装成为程序包的。当然，这绝不是说 R 语言是完美的。它显然不完美，还有很多缺陷。但是，这丝毫不妨碍它成为你学习数据分析的第一门语言。正因如此，狗熊会（微信公号）决定要写一本关于 R 语言的书，要写一本带有狗熊会强烈 DNA 印记的 R 语言入门教材。但是，谁来写？谁来当这个“倒霉蛋”呢？

这个“倒霉蛋”不能是我。在狗熊会的团队里，我岁数最大，有耍赖皮的特权，当然不会“压榨”自己，我更擅长“压榨”其他小伙伴。那该“压榨”谁？只能是布丁（朱雪宁）。在狗熊会的联合创始人团队里，布丁

---

① 北京大学光华管理学院商务统计与经济计量系系主任，教授，博士生导师。

的 R 编程能力公认是最强的。说来惭愧，我是布丁的博士导师，但布丁的理论功底似乎比我还好，而编程能力更比我高出不知几个量级。有时，我会有点懵圈，似乎没教过布丁什么东西，怎么就当了布丁的老师呢？她是怎么成长得如此优秀的呢？思来想去，或许我的一丢丢贡献在于点燃（或者加强了）布丁在数据分析中获得快乐。

布丁天生乐观，而且，她把数据分析的快乐完美地带入了 R 语言编程。单就汉字分词、频数统计，布丁竟然将之跟《张无忌到底爱谁》扯上了关系。这成了狗熊会第一个阅览量过万的推文。我和小伙伴们都惊呆了！说句实话，对此我很困惑。我认真看过这篇推文多遍，实在看不明白布丁在说什么。我对该作品的印象就是语无伦次，逻辑混乱，不知所云，各种差评。但是奇怪，熊粉们怎么就这么喜欢呢？也许是我老了吧。不得不承认，代沟是存在的。但是，我能感受得到，跳跃在 R 代码和《张无忌到底爱谁》文字之间的、布丁那肆无忌惮的快乐。对，这就是布丁的快乐、布丁之于数据分析的快乐。

还说汉字分词、两样本检验、逻辑回归，布丁将之跟《红楼梦作者之谜》扯上了关系，引得众多读者点评布丁的作品，其中既有普通熊粉，也有备受尊重的资深学者，布丁不敢怠慢，逐条答复。不得不承认，我有一点幸灾乐祸的窃喜。我想布丁的内心一定非常崩溃："我就做了一个好玩的中文数据分析，纯娱乐项目，你们怎么当真了呢？"这就是布丁的快乐、布丁之于数据分析的快乐。

布丁是一个优秀的领导者。在她的周围，团结着一帮弟弟妹妹，他们一起构成了布丁小分队（或者叫"敢死队"）。据说，布丁对弟弟妹妹们"手段凶残"，"压榨"无数。但奇怪的是，弟弟妹妹们却非常喜欢这位学姐，亲切地称她为雪姨，并且坚定不移地跟随雪姨闯荡数据江湖。为什么？我斗胆猜测，原因还是快乐。大家在一起，互相学习，互相督促，一起享受数据分析的快乐，一起享受成长的喜悦。我很喜欢这样一个团队架构。碰到极具艰难的任务，我可以通过"压榨"布丁，布丁再"压榨"她

的小分队，达到很高的团队执行力效果。这本书的出版就是一个很好的例子。这本书是我“强派”给布丁的，然后布丁把控整体设计以及很多核心内容，但是，还有很多内容是由其他小伙伴完成的，他们分别是（按姓名拼音排序）：常象宇（政委）、成慧敏、范超、李宇轩、鲁伟、潘蕊（水妈）、王健桥、王毅然、向韵桦。对此，我一并感谢，并对大家处在狗熊会“食物链”的底端深表同情。

我是不是跑题了？布丁给我的任务是给本书写序，却谈到了食物链。不，我没有跑题。我想告诉大家的是，这本书的核心不是 R 语言，是快乐，是数据分析的快乐，是跟布丁学习 R 语言的快乐。

推荐序二

谢益辉[①]

很惭愧，本人还没写几本书，序倒是写了好多篇，俨然已成为作序专业户。不过这次我很荣幸也很乐意为狗熊会摇旗呐喊一嗓子，因为我打心底认同狗熊会的朴素价值观——数据创造价值。这六个字在我看来分量相当重，尤其是在统计学术界颇为难得。如果是我，恐怕没勇气发起这个冲锋，因为我深知公式、定理、模型都是优雅的，而现实中的数据多半是混沌到让你分分钟想掀桌的程度。想用数据创造价值，需要莫大的毅力、耐心和智慧。就算作为一个跟统计沾点边的码农，我也是怯懦地选择了写代码而不是做数据分析，因为我知道后一条路不好走。

在我看来，本书最大的特色是集成了狗熊会这两年大量数据分析案例，而且这些案例都很新潮、实际。我个人最钟爱的还当属老王卖耗子药的万能例子（虽说是虚构的，但这个场景我总觉得很好笑）。我跟熊大只在 2016 年中国 R 语言大会期间某食堂餐桌上匆匆打过一次照面，也只听过他一次报告。还记得他在台上吆喝“全宇宙的中心——五道口”惹得

① 统计之都网站创始人。

我们统计之都的“萌主”（周扬，也是著名“段子手”）在后排嘿嘿一乐，深刻体现了熊大争做网红的决心。我个人完全支持统计学教授做网红，至少听众笑过之后还能留下点思考和知识。可能是受网红路线的影响，这本书也颇有网红风：热门电影、小说、事件等都在书中的案例里有所涉及。分析你关心或能吸引你注意的数据也许能让你更专心地阅读这本书。

本书的另一特点就是很细致。对我这样的读者来说可能细致得有点“令人发指”，比如我肯定没有耐心介绍如何下载安装R，或是如何在浏览器中查看HTML元素。所以写书能完全从新手的角度出发挺难得的，宁可过于细致，也不要贸然假设读者已经拥有某些基础知识。细致的好处在于你学一样就能会一样，而不必再翻别的资料补课。

就写作风格而言，本书内容比较通俗，没什么晦涩的专业术语，我觉得也很好。在模型技术方面，书中除了机器学习一章中简略提及几个稍高等的模型之外，基本以探索性分析和回归为主，这也符合我本人对简单模型的偏爱（没办法，我数学太差）。

本着君子和而不同的精神，以及对狗熊会求真进取精神的信任，我想坦诚地说，世上没有哪本书会是完美而全面的指南，作者和编者一定会有所取舍，比如要顾细致就不能求全面。我相信这本书会为新手打开R的大门、教给读者大量实用技能，但有雄心壮志的读者应该在此基础上继续深造。最近几年恰逢R社区比较“动荡”，这个“动荡”主要源于一个Tidyverse门派（我戏称为“极乐净土”）的异军突起。我自己作为R老用户，看到本书中的代码非常亲切和熟悉，因为我就是这样学R的，但我觉得从今往后，尝试往Tidyverse数据分析范式转型会让很多业余数据分析者受益。

本书主要作者雪宁在统计之都网站也担任主编数年，其领导风范、专业水平和敬业态度都让我深感敬佩。上可推公式，下可敲代码，办事有条有理、有始有终，可谓狗熊会中诸多英雄的突出代表。写作本书想必耗

费了主编不少心血，当然，各章节的作者也付出了大量努力（狗熊会的标准向来比较严苛）。我衷心期待更多人能通过这本轻快又实在的书了解数据分析的乐趣和技能，并进一步找到自己独特的用数据创造价值的法门。

前　言

朱雪宁（布丁）

两年前，狗熊会微信公众号刚刚投入运营，到底写点什么好呢？因为我对R语言更加熟悉，熊大（王汉生）就提议我来牵头组织关于R语言数据分析的专栏，还取了一个相当文青的名字“R语千寻”。写了几篇后，没想到竟然收到不错的读者反馈，这个专栏也就逐渐固定下来。我们意识到，R语言是一种有力的工具，在实际案例、数据分析中有无限的魅力，而“R语千寻”结合实际数据进行案例讲解的形式也受到许多朋友的喜爱。

自建立以来，“R语千寻”专栏迎来了越来越多的创作者，积累了丰富翔实的内容。于是，就有了对这些内容适时系统地梳理、总结，形成一本结合丰富的数据与案例教学的R语言数据分析书籍的想法。对“R语千寻”专栏而言，这并不是一个终点。在未来的日子里，“R语千寻”将继续为大家推出有意思的故事与有趣的分析，也希望收到更多读者朋友的反馈。

本书适合刚刚入门或者了解R语言但还没有认识到R语言在实际数据分析中强大威力的朋友。或许你是一个编程小白，渴望入门一种较为容易上手的编程语言，但又在庞大的知识体系前望而却步；或许你还在求学，

本学期刚刚学习了R语言课程，但是你想了解的不止于如何生成一个数组或者矩阵这么简单；又或许你是一个业界从业者，逐渐认识到手上开始积累越来越多的数据，它们也许能产生巨大的商业价值，而你却无所适从。本书希望能带给你一些感悟。

在这个最好的时代，我们有能力收集、积累大量的数据；数据分析、人工智能也正处在前所未有的风口上。正如狗熊会出品的第一本书——《数据思维》所强调的那样，最重要的是完成从数据到价值的转换。本书希望告诉大家，这种转换不仅需要培养严谨的数据分析思维，同时也要具备踏实的实务分析能力。如何将业务问题转变为数据可分析问题呢？对于现实中可能并不“美”的数据，如何清洗，如何描述，以及如何建模和解读呢？所有这些步骤，我们通过具体的R语言实务分析，向大家一一解读。

对于从事数据分析的人来说，这还不够，工作的需求往往不止于此。数据分析工作每天面临的是大量的细节。曾经以为数据分析就是玩转高大上的模型，然而入行后你才会发现，80%的时间你将用来理解业务、清洗数据、描述规律、大胆假设、小心求证……最后真正上模型的时间，通常也就不过剩下的20%而已。在所有的过程中，事无巨细，如果能熟练使用R语言，它将成为你得力的帮手。经常听到这样的抱怨：R语言处理实际数据太慢！我们应该去学C，Java。而实际去看看那些抱怨的人写出的代码，虽然能达到最终目的，但是效率却惨不忍睹！适当的转变编程思路，改用一两个函数或者R包，编程效率往往能数以十倍地提升。所以，那些每天喊着打语言仗的人真的不如花点时间稍微提高一下R编程的知识水平。在作者看来，急于学习多门语言不如先精通一门语言。

在内容组织方面，本书从R语言简介及优势入手，再到数据描述、建模等数据分析的各个环节，由浅入深，组成不同章节。第1章介绍R语言的背景、优势，用幽默的语言告诉你“R语言能做什么”。第2章介绍基本数据操作，包括数据基本类型、数据读写，这些组成了R语言应用的根

基。第3章介绍R语言与统计分析，包括三大利器：描述分析、统计检验、回归分析，这些环节在实际的数据分析中缺一不可。第4章解读R语言与非结构化数据分析，主要包括无处不见的文本数据和图像数据。第5章介绍如何用R语言进行当下最火的机器学习建模，从数据清洗到模型集成、建模调参一网打尽。第6章介绍R语言的爬虫原理及技巧。本书对于R语言的整个知识体系框架也许不是涉及最广的，但是希望能对实际数据分析产生直接的借鉴作用。

本书由狗熊会核心创作团队齐心协力完成，希望向大家展示R语言有趣、实用、高效的一面。参与创作的成员有（按姓名拼音排序）：常象宇（政委）、成慧敏、范超、李宇轩、鲁伟、潘蕊（水妈）、王健桥、王毅然、向韵桦；参与本书整理、校对的同仁有（按姓名拼音排序）：何通、杨瀚轩，感谢所有参与成员付出的巨大心血和努力。本书还要特别感谢狗熊会CEO李广雨先生给予的鼓励和支持；感谢蔡知令教授、王汉生教授在写作过程中关于内容组织、时间安排等提出的宝贵建议；感谢狗熊会所有同事提出的宝贵建议以及细致的审查意见；感谢中国人民大学出版社李文重编辑在书稿形成、章节安排等方面付出的巨大努力。

另外，本书中引用的图片除特殊标注外均来源于网络，鉴于引用这些图片时无法获知原创作者及出处，在此对原创作者统一表示感谢。

最后，把本书献给所有培养过我们的老师、企业合作伙伴；献给我们的朋友、家人。正是因为有你们，我们才能站在更高更大的舞台上，施展抱负，勇往直前。在这里，再次想起狗熊会的理念：聚数据英才，助产业振兴。同时，也祝福狗熊会的明天会更好，愿越来越多志同道合的小伙伴加入我们，分享数据分析带给你的快乐。由于本书写作仓促，疏漏之处在所难免，请大家多多批评指正！

目 录

CONTENTS

第 1 章/*Chapter One*

初识 R 语言

1.1 初识 R 语言

R 语言可以说是一款在开源世界里集万千宠爱于一身的软件，你能想到的地方都有它的身影。

做学术，看看权威的 TIOBE 开发语言的排行榜[①]（见图 1-1）！

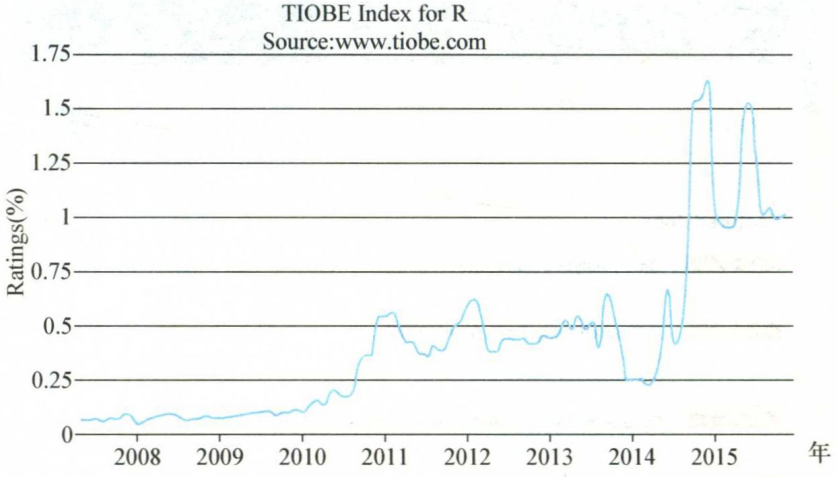

图 1-1 R 语言热门程度

① TIOBE 排行榜是根据互联网上有经验的程序员、课程和第三方厂商的数量，并使用搜索引擎（如谷歌（Google）、必应（Bing）、雅虎（Yahoo!））以及维基百科（Wikipedia）、亚马逊（Amazon）、YouTube 统计出的排名数据，用来反映某种编程语言的热门程度。

听讲座，看看每年都会举办的中国 R 语言大会的阵容（见图 1－2）！

图 1－2　中国 R 语言大会现场

找工作，看看与 R 语言相关的工作（见图 1－3）！

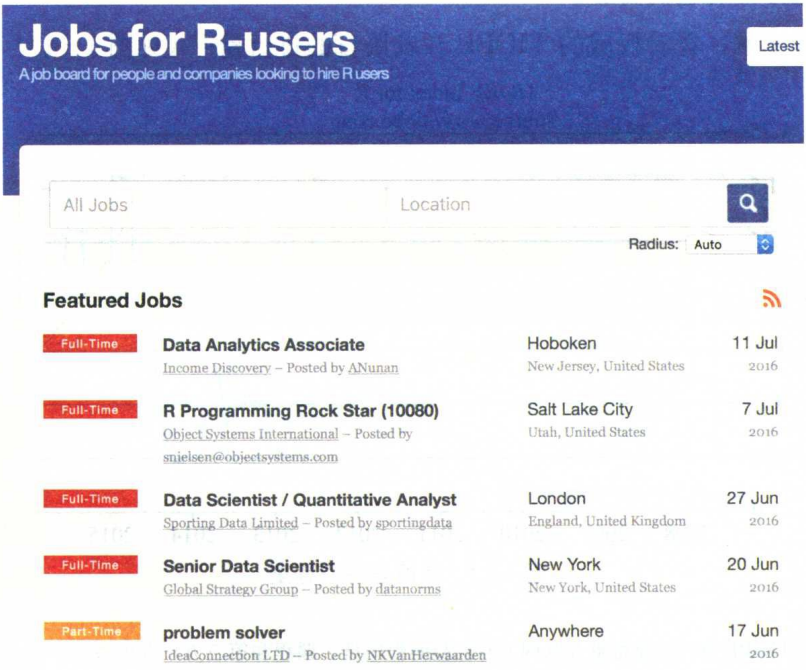

图 1－3　与 R 语言相关的工作

如果还不够，看看每年让你“剁手吃土”的它们同样在用 R（见图 1-4）！

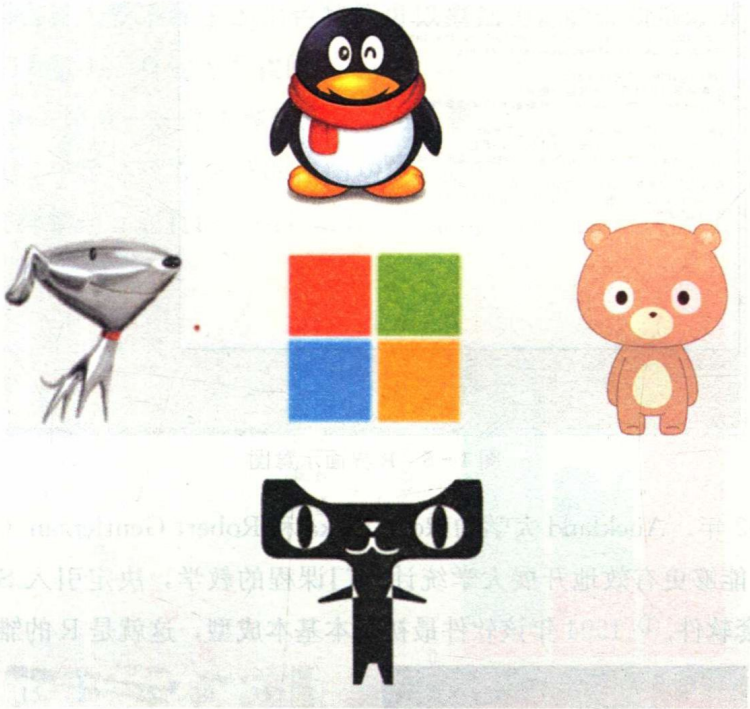

图 1-4　让你“剁手吃土”的它们

1.1.1　R 语言是什么？

R 是一个有着强大统计分析功能及作图功能的软件系统。图 1-5 是它的界面示意图。①

说到 R 语言的发展历程，还要先从另一门语言 S 谈起。S 语言是由 AT&T 贝尔实验室 John Chambers 等开发的一种用来统计编程的语言。它目前有两种实现版本：一种是由 TIBCO 经营的商业软件 S-plus；另一种就是免费开源的 R 语言。

① 别看 R 的页面很丑，但 R 可是统计、计算样样精通。不过，这是个颜值可以统领世界的年代，我们之后将介绍一款更加美观的编辑器。