HZ BOOKS
华章IT

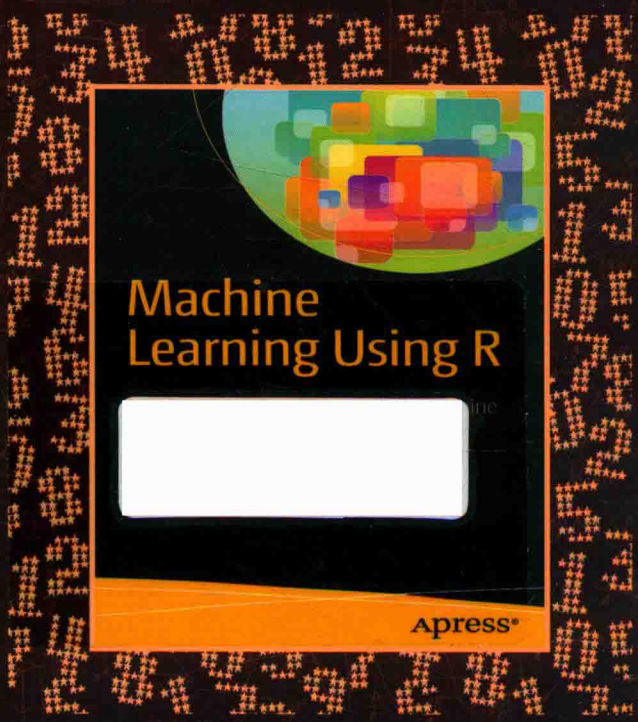
Apress®

数据科学与工程技术丛书

R语言机器学习

[印度] 卡西克·拉玛苏布兰马尼安（Karthik Ramasubramanian）
阿布舍克·辛格（Abhishek Singh） 著

吴今朝 译



MACHINE
LEARNING USING R

机械工业出版社
China Machine Press

数据科学与工程技术丛书

MACHINE LEARNING USING R

R语言机器学习

[印度] 卡西克·拉玛苏布兰马尼安（Karthik Ramasubramanian）
阿布舍克·辛格（Abhishek Singh） 著

吴今朝 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

R 语言机器学习 /（印）卡西克·拉玛苏布兰马尼安，（印）阿布舍克·辛格著；吴今朝译．
—北京：机械工业出版社，2018.4
（数据科学与工程技术丛书）
书名原文：Machine Learning Using R

ISBN 978-7-111-59591-5

I. R… II. ①卡… ②阿… ③吴… III. 程序语言—程序设计 IV. TP312

中国版本图书馆 CIP 数据核字（2018）第 064956 号

本书版权登记号：图字 01-2017-7339

R 语言机器学习

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：殷 虹
印 刷：北京市兆成印刷有限责任公司 版 次：2018 年 6 月第 1 版第 1 次印刷
开 本：185mm×260mm 1/16 印 张：27（含彩插 0.5 印张）
书 号：ISBN 978-7-111-59591-5 定 价：99.00 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

图 4-1　根据 GDP 显示前 10 名国家的折线图

图 4-2　根据农业对占 GDP 的贡献百分比显示前 10 名国家的折线图

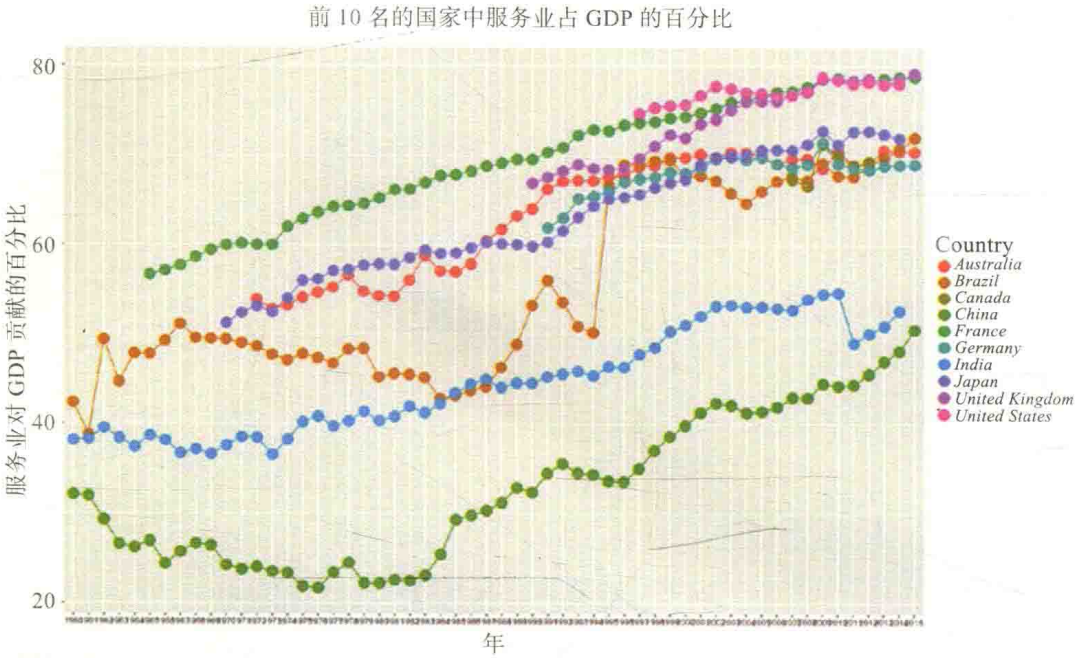


图 4-3　根据服务业占 GDP 的百分比显示前 10 名国家的折线图

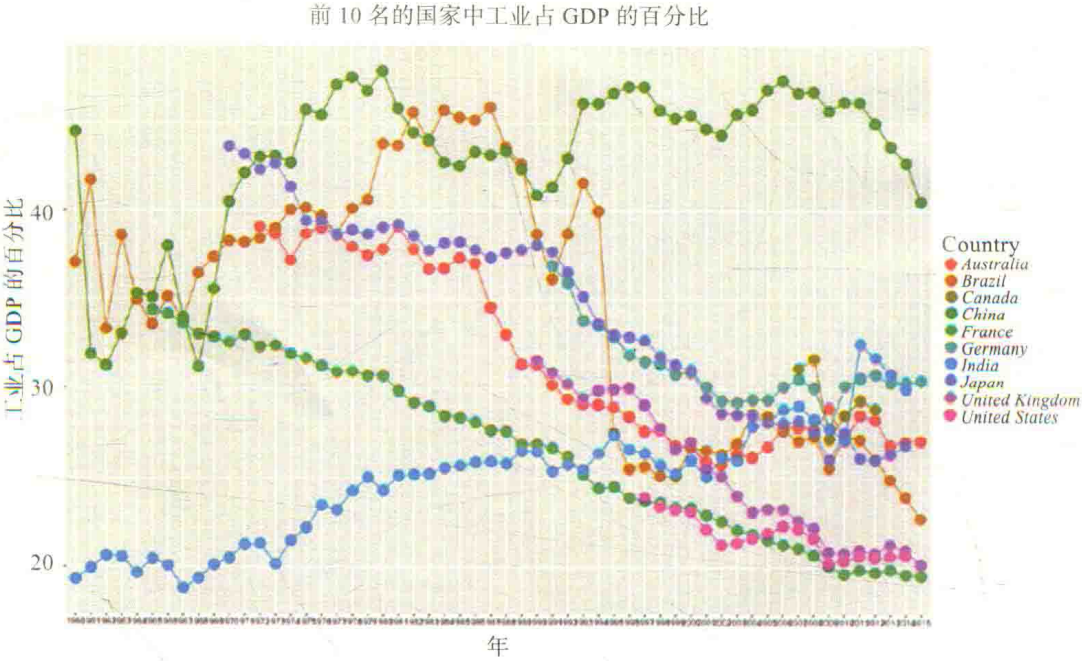


图 4-4　根据工业占 GDP 的百分比显示前 10 名国家的折线图

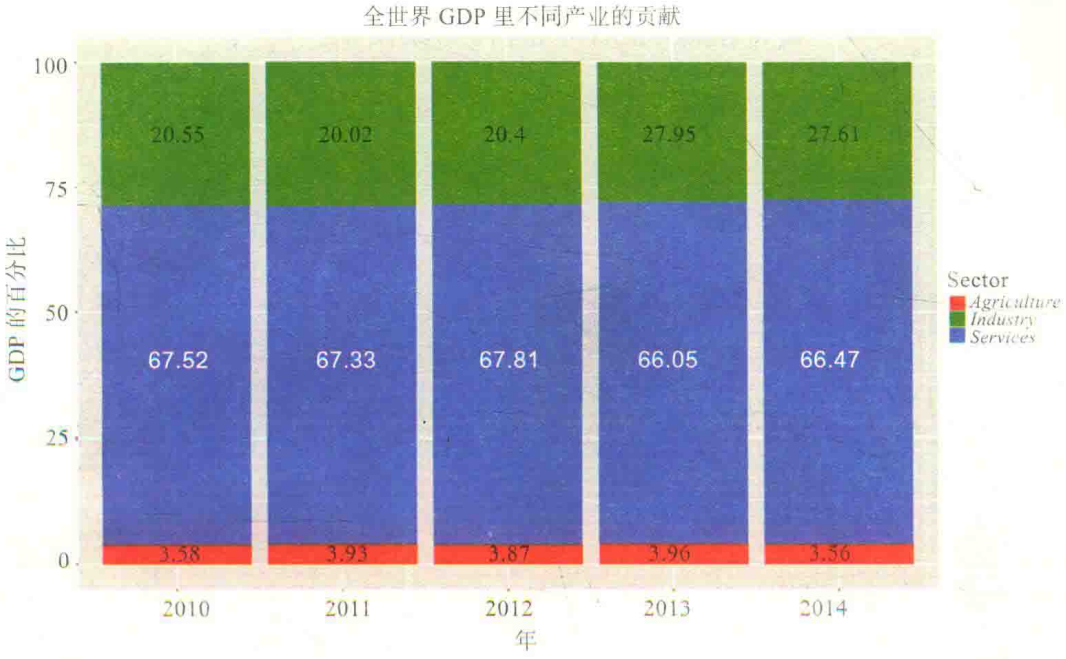


图 4-5　显示各产业对全世界 GDP 贡献情况的一个堆叠柱状图

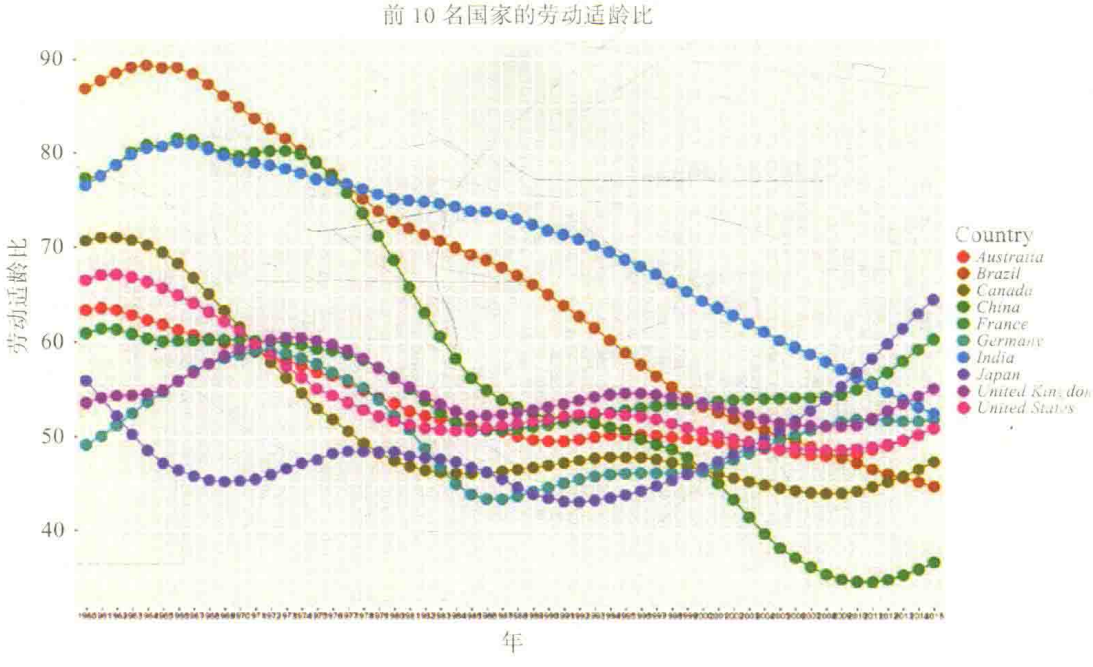


图 4-6　根据其劳动适龄比显示前 10 名国家的堆叠折线图

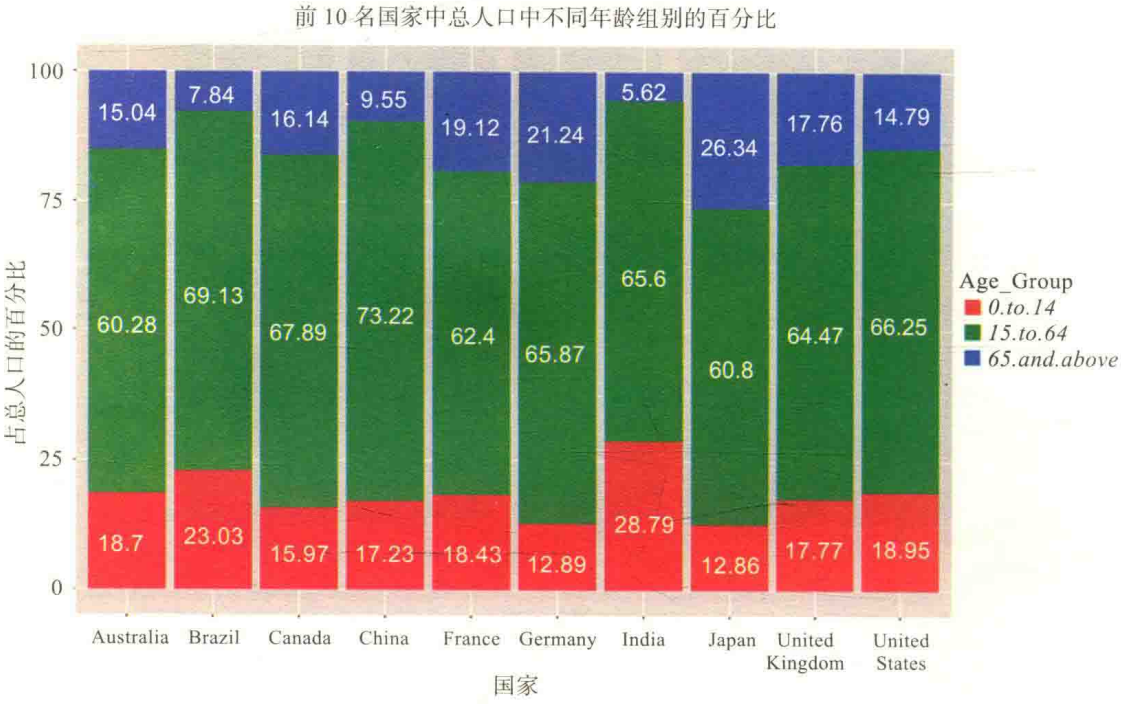


图 4-7　显示不同年龄组别占总人口百分比的堆叠条形图

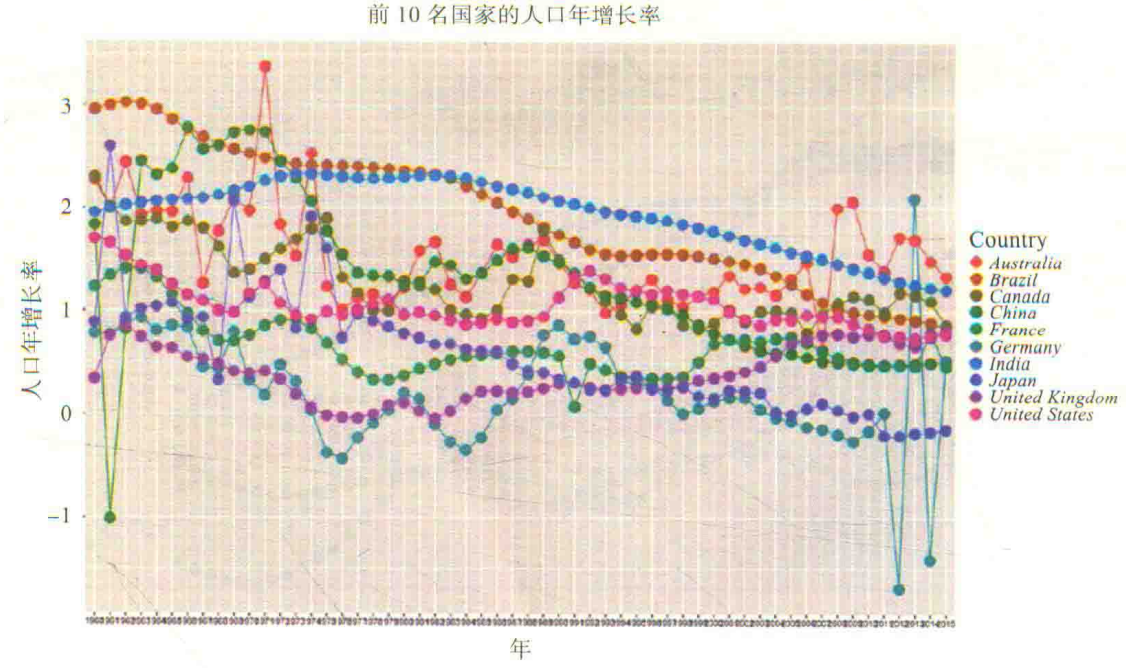


图 4-8　显示前 10 名国家人口年增长率的折线图

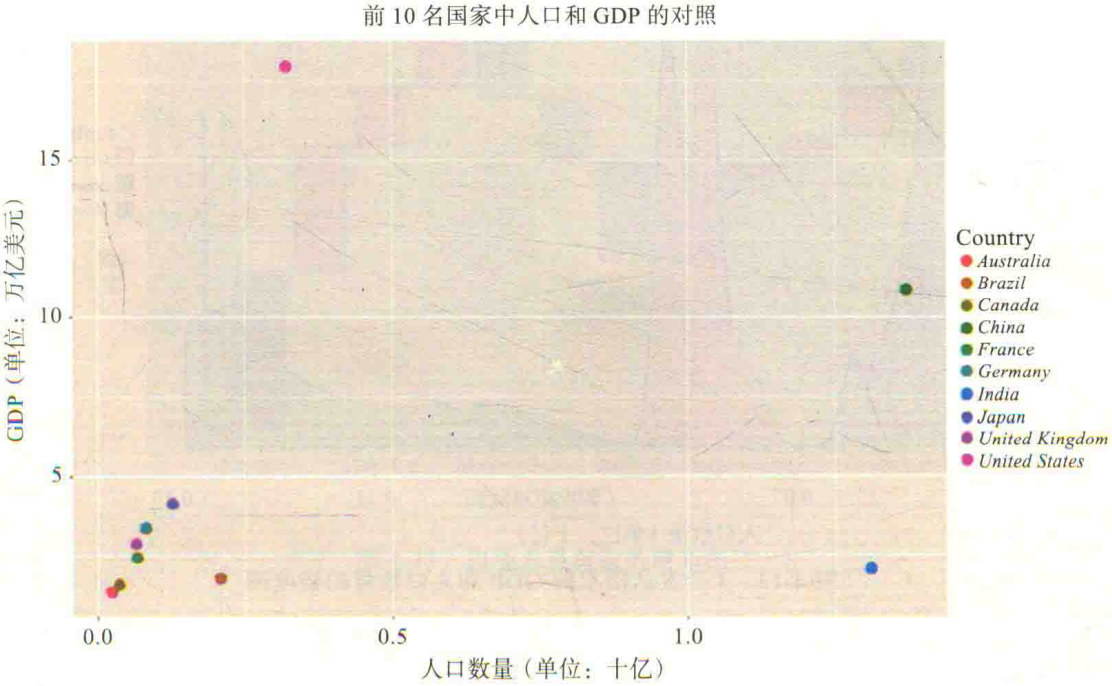


图 4-9　前 10 名国家的人口与 GDP 之间关系的散点图

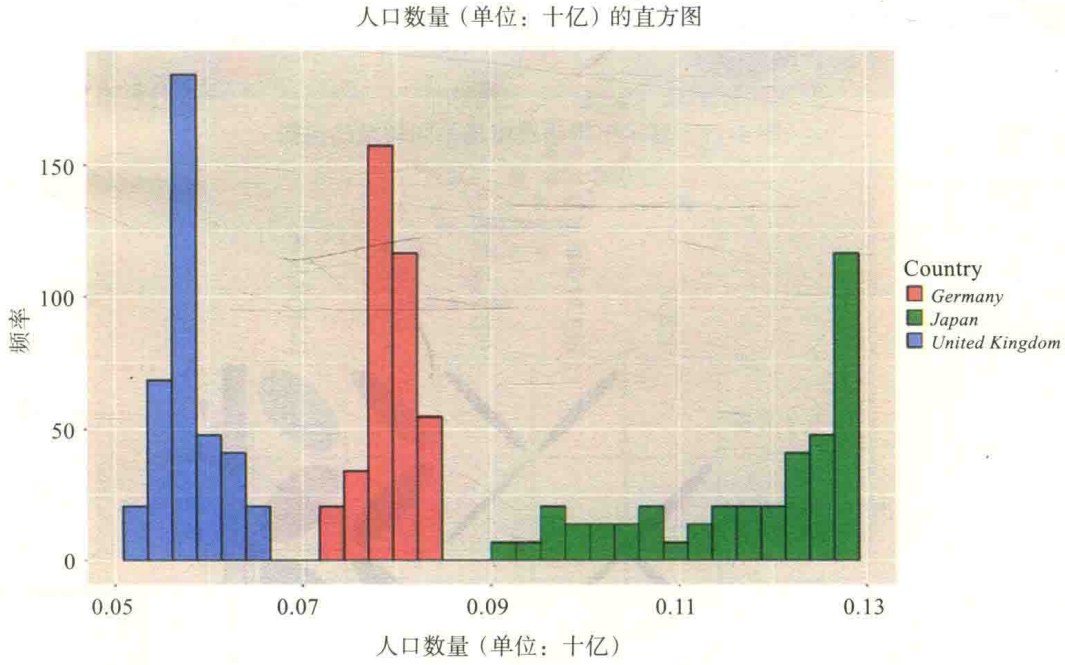


图 4-12　显示 3 个发达国家的 GDP 和人口数量的直方图

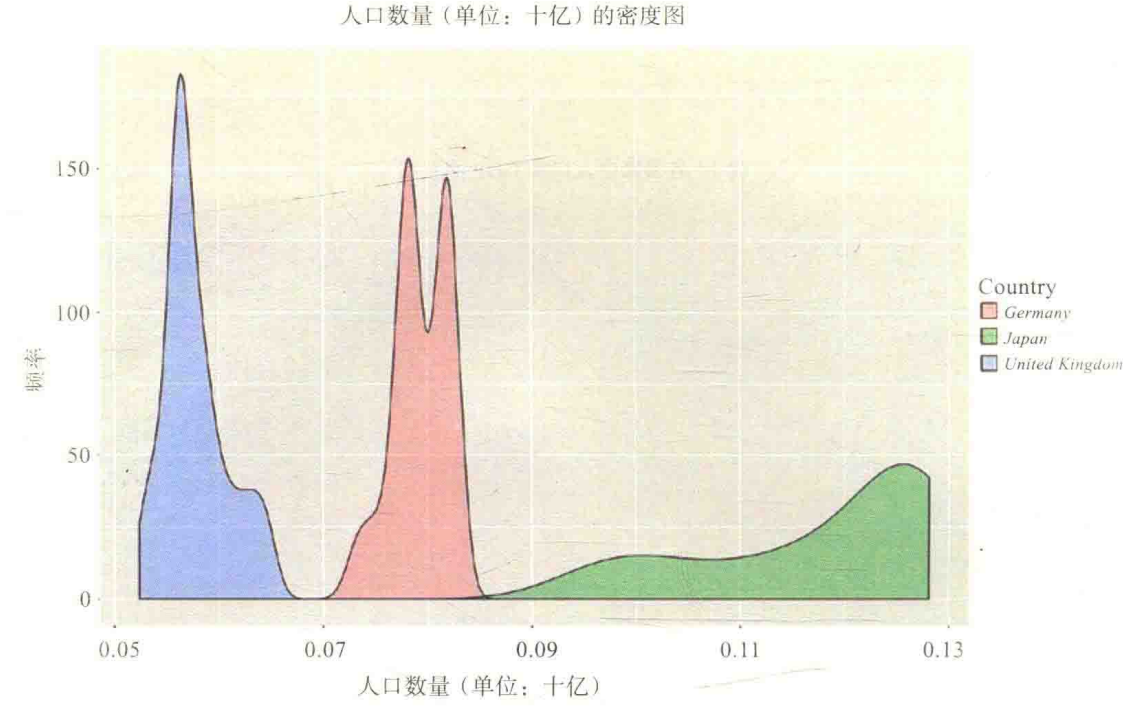


图 4-13　3 个发达国家的 GDP 和人口数量的密度图

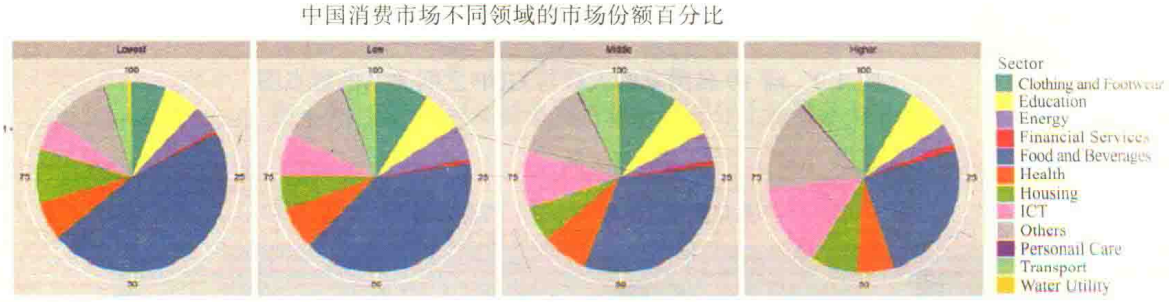


图 4-17　显示中国消费市场不同领域的饼图

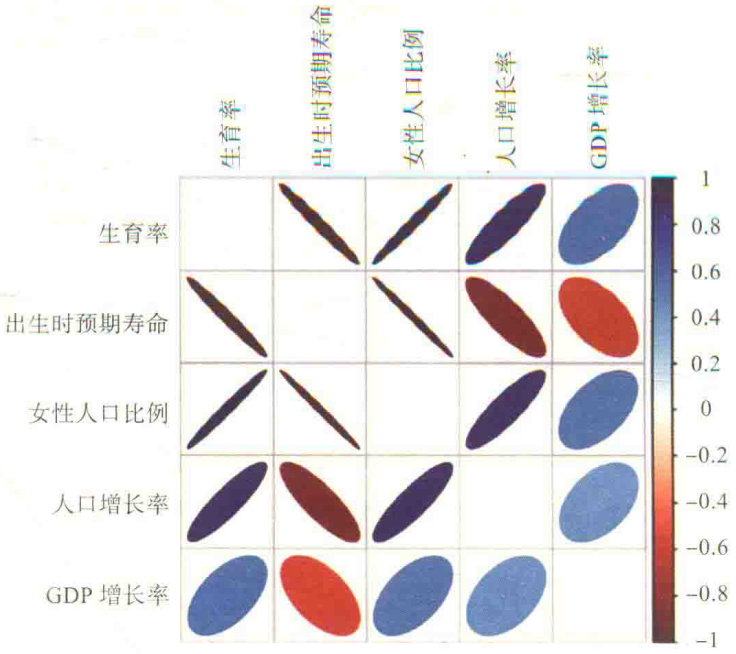


图 4-18　显示多种世界发展指标之间相关性的图

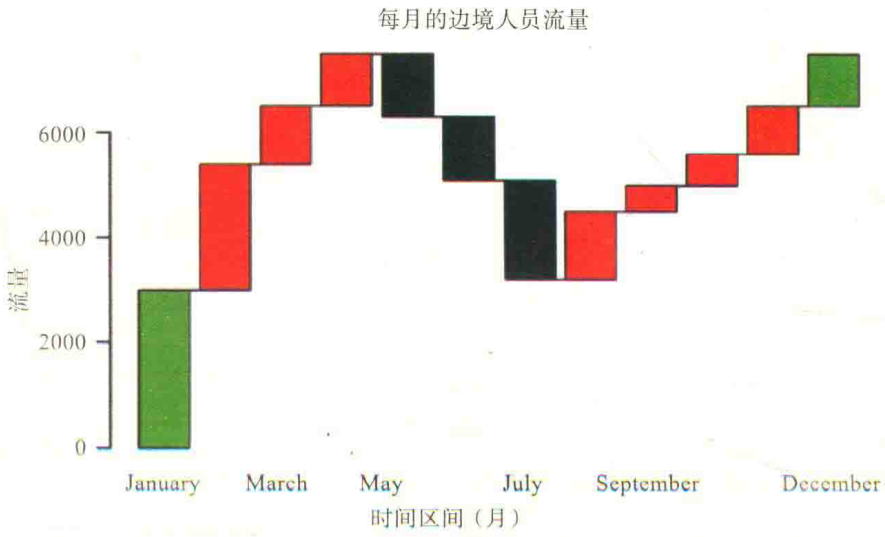


图 4-23　边境人员流量的瀑布图

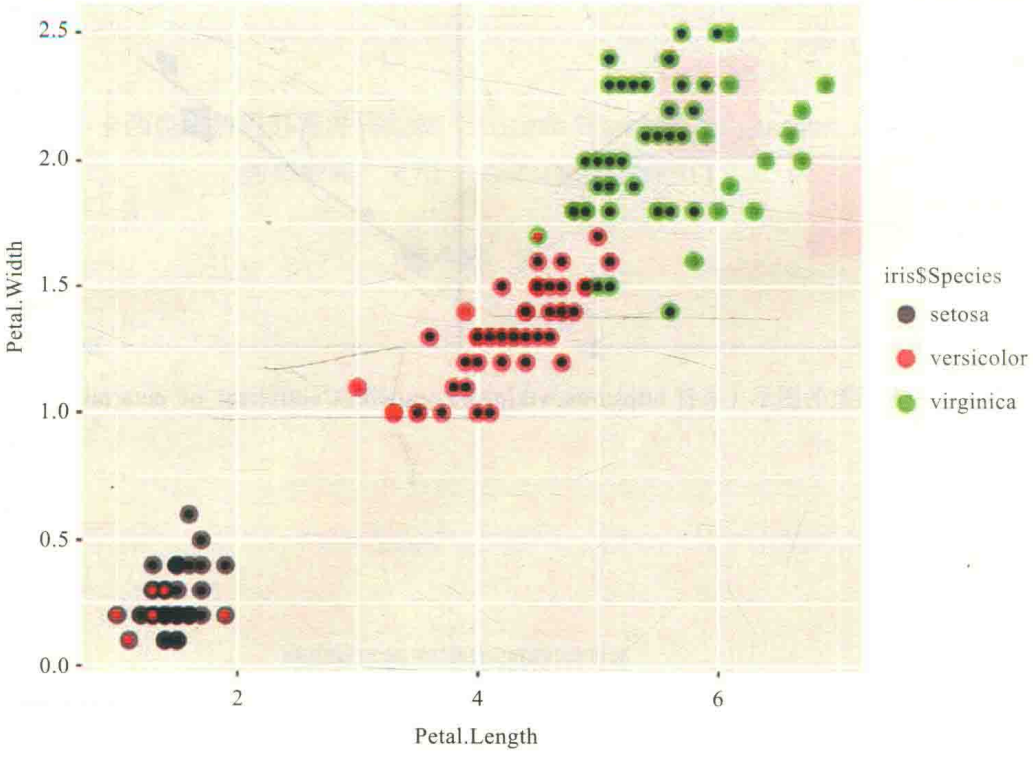


图 4-27　按鸢尾花数据中品种的实际分类形成的聚类

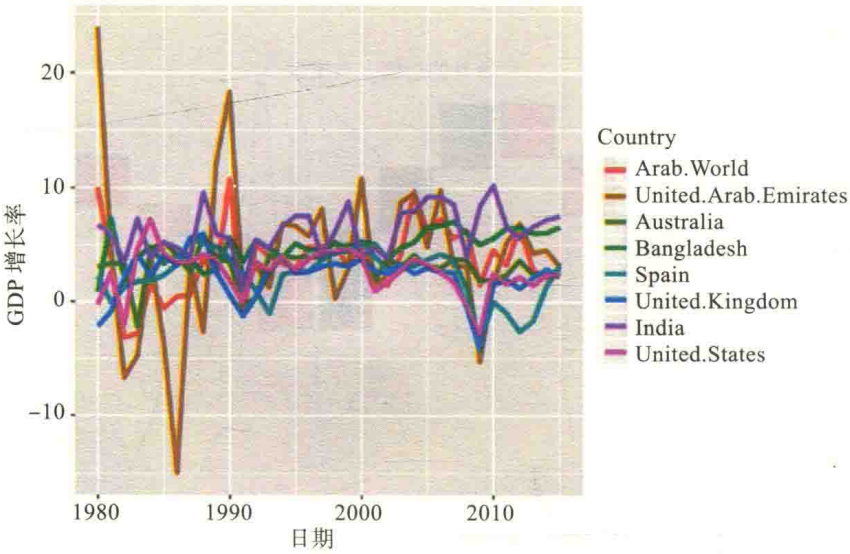


图 4-30　八个国家的国内生产总值增长

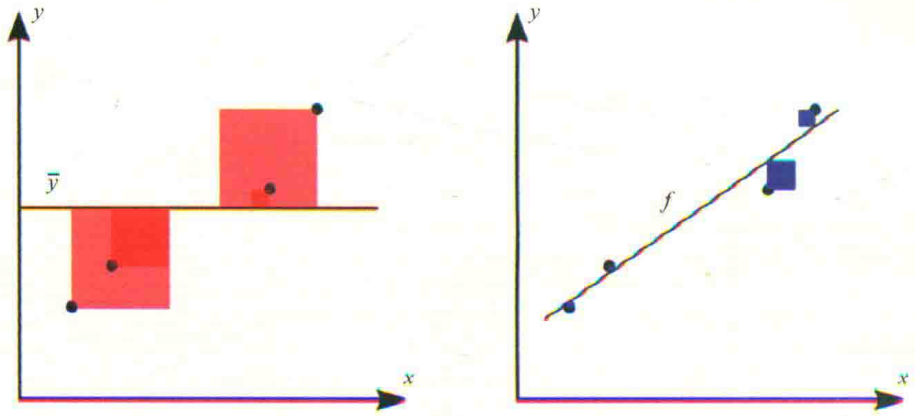


图 7-6　平方误差的图示（来自 https://en.wikipedia.org/wiki/Coefficient_of_determination）

译 者 序

作为译者，我觉得这本书最大的特点就是它的全面性。从基础的统计学原理和R语言编程知识，到核心的机器学习理论和算法分析，以及机器学习模型的评估和改进方法，再到机器学习技术在大数据平台上的应用，在本书中都占到了一定的篇幅。此外，对于一些比较前沿和高级的主题，作者也给出了相应的参考资料，供有兴趣的读者进一步提高。

当然，这样的写作思路也是有利有弊的。好处是书中展示了当今机器学习技术发展的全貌，有利于读者理解各种机器学习技术的适用范围及其相互联系，先从全局和整体入手，再逐步深入到每个细节中，这样会比较容易把握适合自己的方向。至于不足之处，也许是因为范围铺得太广，导致有些部分的讲解不够深入，有些代码和实例的讲解也略显简单。

因此，这本书比较适合机器学习技术的初学者，以及仅仅在某个领域有一定经验、希望从更广的角度来认识它的专业人士。如果读者希望更深入钻研某些技术，可以将本书给出的参考资料作为起点。虽然这些参考资料基本都是英文版的，不过要掌握最新的技术，阅读原版资料也是一个必要的途径。本书中还引入了很多R语言平台的组件包，它们也适用于实际环境下的分析工作，是应用领域必不可少的得力工具。限于篇幅，本书很难逐个对它们进行详细介绍，而且开源组件包的版本变化很快，在实际应用时，也需要认真阅读它们的文档。

很多看过我前几本译作的读者都知道我有在GitHub上开辟讨论区的习惯，这样可以针对书中的理论、应用、代码等问题与读者进行交流。本书也不例外，它的讨论区链接是 https://github.com/coderLMN/machineLearningUsingR/issues，欢迎读者提出问题并参与讨论。我之所以愿意花时间参与这样的讨论，是因为这样不仅能够帮助有疑问的读者，及时纠正书中可能存在的错误，同时我自己在这个过程中也经常可以开阔思路、得到启发。我不知道还有没有其他译者也会这么做，但起码我对自己的这个做法是很自豪的。

其实我还有一个更大的心愿，就是希望有读者在实际运用这些技术的过程中写出自己的理解和体会，用自己的经验来帮助其他的读者。我希望能把这样的实践经验总结放在讨论区里，其中高水平的总结还可以在书籍重印的时候作为附录添加进去，让更多的

读者能够看到。

真心希望这个心愿在这本书上能够实现。

此外，我在书中加入了一些译者注，标注了我在翻译过程中觉得有必要说明的一些问题，例如某些代码的变动、某些概念的通俗化解释、提醒读者要注意的细节，等等。由于本人水平有限，在翻译内容和译者注里难免会有不严谨或者不正确的地方，还请读者不吝指正。

在本书的翻译过程中，机械工业出版社的陈佳媛和缪杰两位编辑为我提供了很多帮助，在此向他们表示感谢。

最后，还是要感谢我的家人。这本书翻译的周期比较长，工作量也不小，感谢他们的支持和激励，让我能保质保量地完成这个工作。

吴今朝

2018 年 2 月

关于作者

Karthik Ramasubramanian 就职于印度最大和发展最快的技术创业公司 Hike Messenger。他把商业分析和数据科学的最佳经验带到了 Hike Messenger 的角色中。在 7 年的研究和产业经验中，他在零售、电子商务和技术行业解决跨行业的数据科学问题，为数据驱动的解决方案进行开发和原型构建。他以前曾在印度最大的电子商务零售商之一 Snapdeal 任职，负责关于客户增长和定价分析的核心统计模型。在加入 Snapdeal 之前，作为中央数据库团队的一员，他曾负责管理 Reckitt Benckiser（RB）全球业务应用的数据仓库。他在可扩展的机器学习领域具有丰富的经验，专长包括复杂的图网络和自学习神经网络。他拥有 Anna 大学 PSG 技术学院的理论计算机科学硕士学位，是一名认证的大数据专家。通过各种在线和公众论坛，他热心于向未来的数据科学家们进行传授和指导。他喜欢在闲暇时间写诗，并热衷于旅游。

Abhishek Singh 是美国第二大的人寿保险供应商 Prudential Financial 公司的数据科学家，工作地点在爱尔兰。他在数据科学方面拥有五年的专业和学术经验，涵盖了咨询、教学和金融服务。他曾经在 Deloitte Advisory 领导了针对美国顶尖银行的监管风险、信用风险和资产负债表模型化需求的风险分析项目。目前，他正在为 Prudential 的人寿保险业务开发可扩展的机器学习算法。他拥有时间序列模型的工作经验，并曾与跨职能团队合作，在企业基础架构中实施数据科学解决方案。他一直是 Deloitte 职业大学的培训师，并领导了统计、经济学、金融风险和数据科学工具（SAS 和 R）等领域专业人士的培训和发展计划。他拥有位于 Guwahati 的 the Indian Institute of Technology 的数学和计算学士学位以及位于 Bangalore 的 the Indian Institute of Management 的 MBA 学位。他在一些数据科学的公开活动中授课，并与领先的大学合作，给研究生讲授数据科学技术。他对法律感兴趣，并拥有 NALSAR 大学网络法的博士后文凭。他在业余时间爱好烹饪和摄影。

关于技术审稿人

Jojo Moolayil 是一位数据科学家，也是“Smarter Decisions：The Intersection of Internet of Things and Decision Science”一书的作者。他在数据科学、决策科学以及物联网方面拥有超过 4 年的业界经验，曾与具有很高影响力的一些业界领导者合作，从事了覆盖多个行业的关键项目。他目前就职于工业物联网数据科学的先驱和领导者——General Electric，生活在印度的硅谷——Bengaluru。

他在印度 Pune 出生和长大，毕业于 the University of Pune，攻读了信息技术工程学专业。他在世界上最大的纯游戏分析厂商 Mu Sigma 公司开启了他的职业生涯，并与很多财富 500 强客户的领导者合作。作为物联网分析的早期先驱者，他把决策科学方面的知识融合到一套解决问题的框架里，并把数据和决策科学知识融入了物联网分析。

为了巩固他在工业物联网中利用数据科学的基础，并扩大解决问题实践的影响力，他加入了一家快速成长的物联网分析公司——Flutura，工作地点在 Bangalore，公司总部位于硅谷。在 Flutura 短暂停留之后，Jojo 继续在 Bangalore 与工业物联网领导者 General Electric 合作，在那里他专注于解决工业物联网用例的决策科学问题。作为他在 General Electric 工作的一部分，Jojo 还专注于开发数据科学和工业物联网的决策科学产品和平台。

除了撰写关于决策科学和物联网的书籍外，Jojo 还担任了 Apress 出版社的机器学习和业务分析方面书籍的技术审稿人。他是一位活跃的数据科学导师，并在 http://www.jojomoolayil.com/web/blog/ 撰写博客。

个人网站：

http://www.jojomoolayil.com/

https://www.linkedin.com/in/jojo62000

“我要感谢我的家人、朋友和导师对我一生的支持和持续的激励。”

Jojo Moolayil

致　谢

我们要感谢我们的老师、开源社区和同事用知识和信心培养了我们，让我们能写出本书的第 1 版。本书中的知识是我们多年在母校和业界获得的研究成果和专业经验的积累。我们非常感谢位于 Coimbatore 的 PSG 科技学院应用数学和计算科学系的 R. Nadarajan 教授和 R. Anitha 教授，他们不断地支持和鼓励我们在机器学习领域的工作。

在快速变化的世界中，机器学习领域的发展非常迅速，大部分最新的进展都是由开源平台驱动的。我们要感谢全球各地的开发者和贡献者，他们自由地分享了他们关于这些平台的知识。还要感谢我们在 Snapdeal、Deloitte 以及我们现在的单位 Hike 和 Prudential 的同事，他们提供了试验和创造尖端数据科学解决方案的机会。Karthik 尤其要感谢他生命中一直以来灵感的源泉，他的父亲 S Ramasubramanian 先生。他还非常感谢他的主管，Snapdeal 数据科学团队负责人 Nikhil Dwarakanath 先生为他创造机会，为他提供最好的分析专业人士，并给了他承担挑战性项目的动力。

Abhishek 要感谢他的父亲，印度气象部门的资深科学家 Charan Singh 先生，在启蒙阶段给他介绍了数据对于天气预报的作用。在个人方面，Abhishek 想感谢他的母亲 Jaya、姐姐 Asweta 和兄弟 Avilash 持续的道义支持。

我们要感谢出版商 Apress，特别是 Celestine 在这次合作机会中对我们的认可，还有 Sanchita 和 Prachi 对这个项目的管理，以及 Poonam 和 Piyush 的审稿，以及出版团队每位成员的付出。

Karthik Ramasubramanian

Abhishek Singh

目　录