PACKT PUBLISHING

89个R和Python秘籍，帮你解决现实世界中的数据科学问题

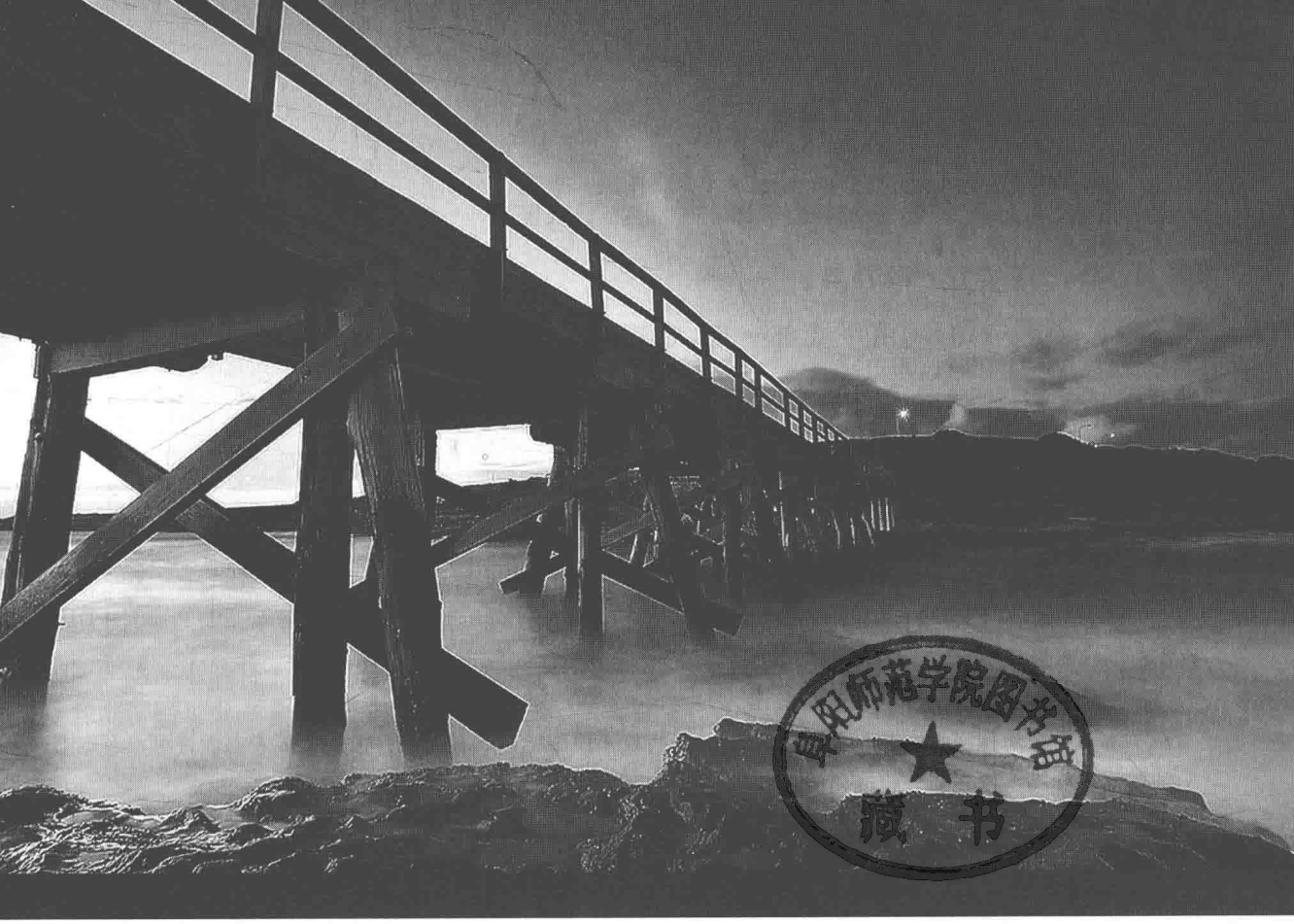
数据科学实战手册（R+Python）

Practical Data Science Cookbook

[美] Tony Ojeda Sean Patrick Murphy Benjamin Bengfort Abhijit Dasgupta 著

郝智恒 王佳玮 谢时光 刘梦馨 译

中国工信出版集团 人民邮电出版社 POSTS & TELECOM PRESS



数据科学实战手册

（R+Python）

[美] Tony Ojeda Sean Patrick Murphy
Benjamin Bengfort Abhijit Dasgupta 著

郝智恒 王佳玮 谢时光 刘梦馨 译

人民邮电出版社

北京

图书在版编目（CIP）数据

数据科学实战手册 : R+Python / (美) 托尼·奥杰德 (Tony Ojeda) 等著 ; 郝智恒等译. -- 北京 : 人民邮电出版社, 2016.8
ISBN 978-7-115-42675-8

Ⅰ. ①数… Ⅱ. ①托… ②郝… Ⅲ. ①软件工具－程序设计 Ⅳ. ①TP311.56

中国版本图书馆CIP数据核字(2016)第144243号

版权声明

◆ 著　[美] Tony Ojeda　Sean Patrick Murphy
Benjamin Bengfort　Abhijit Dasgupta
译　郝智恒　王佳玮　谢时光　刘梦馨
责任编辑　王峰松
责任印制　焦志炜

◆ 人民邮电出版社出版发行　北京市丰台区成寿寺路 11 号
邮编　100164　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京圣夫亚美印刷有限公司印刷

◆ 开本：800×1000　1/16
印张：22
字数：425 千字　2016 年 8 月第 1 版
印数：1－3 000 册　2016 年 8 月北京第 1 次印刷

著作权合同登记号　图字：01-2015-0748 号

定价：59.00 元

读者服务热线：(010)81055410　印装质量热线：(010)81055316
反盗版热线：(010)81055315

内容提要

这本书是基于 R 和 Python 的数据科学项目案例集锦。业界的数据分析师、数据挖掘工程师、数据科学家都可以读一读。想要了解实际工作中如何用数据产生价值的在校学生，或者对数据科学感兴趣的人也值得一读。本书最大的优点在于其结构，每一章的每一节内容都是按照“准备工作—处理流程—工作原理”的方式组织，这种组织形式非常适合 learn-by-doing。本书的内容涵盖了基于数据科学的所有要素，包括数据采集、处理、清洗、分析、建模、可视化以及数据产品的搭建。案例包含了汽车数据分析、股票市场建模、社交网络分析、推荐系统、地理信息分析，以及 Python 代码的计算优化。通过手把手的案例解析，令读者知其然并知其所以然。

关于作者

托尼·奥杰德是一位经验丰富的数据科学家和企业家，在商业流程的最优化方面非常专业，并且对创造和执行创新型数据产品和解决方案非常有经验。他在佛罗里达国际大学获得金融硕士学位，并且在德保罗大学获得了 MBA 学位。他是社区数据实验室的创始人、华盛顿 DC 数据社区的联合创始人，致力于对数据科学家的教育提升和组织。

首先，也是最重要的，我想感谢我的合作者，正是他们不知疲倦的工作才有了这本书。我很骄傲我们一起合著了本书。我也期待在未来，我们能和你一起合作更多项目，有更多更好的产出。

我还想感谢我的审稿人，尤其是维尔·沃希思和萨拉·凯丽。他们阅读了本书的每一个章节，并且给了我们很好的反馈。正是因为他们的建议和意见，才有了本书现在的品质。

我还想要感谢我的家人和朋友，感谢他们对我工作的支持和鼓励。

最后，我要感谢我的未婚妻、我人生的伴侣——尼基。感谢她的耐心、理解和陪伴。如果我的个人生活没有现在这么稳定，这样充满爱，我是不会冒险完成各种实验的。感谢她提供的充满爱和支持的氛围。

肖恩·派特里克·莫非在约翰霍普金斯大学的应用物理实验室做了 15 年的高级科学家，他专注于机器学习、建模和模拟、信号处理以及高性能计算。现在，他是旧金山、纽约和华盛顿 DC 多家公司的数据顾问。他毕业于约翰霍普金斯大学，并在牛津大学获得 MBA 学位。他现在是华盛顿 DC 数据创新见面会的联合组织者，是 MD 数据科学见面会的联合

创始人。同时他也是华盛顿 DC 数据社区的联合创始人。

本杰明·班福特是一位非常有经验的数据科学家和 Python 开发者。他曾在军方、业界和学术界工作过 8 年。他现在在马里兰大学派克学院攻读计算机博士学位，研究元识别和自然语言处理。他拥有北达科塔州立大学的计算机硕士学位，并且在那里教授过本科的计算机科学课程。他是乔治城大学的客座教授，在那里教授数据科学和分析。本杰明曾经在华盛顿 DC 参加过两次数据科学培训：大规模机器学习和多领域大数据技术应用。他非常感激这些将数据模型以及商业价值融合的课程，他正在将这些新兴组织构建为一个更成熟的组织。

我想要感谢维尔·沃希思，感谢他对我所做的一切无尽的支持。他甚至同意校对我的技术文档，并且增加了我写作部分的可读性。在我的职业生涯中能有他这样的伙伴和朋友，对我来说是至关重要的。我还要感谢我的合著者——托尼和肖恩，感谢他们的努力，才有了这本书。我还要感谢萨拉凯丽，感谢她对我们的建议和新想法。到现在为止，她已经和我们做了非常多的冒险，并且我也希望能够早日阅读她的大作。最后，我要特别感谢我的妻子——杰西，尤其是那些我需要工作到深夜的日子。没有她，我完全不会写出现在的东西。她是一个非常聪明的人，也是我们家的另一个作者。她的作品将会成为经典，每个学生都应该读一读。

阿布吉特·达斯古普塔是在大华盛顿 DC-马里兰-佛吉尼亚地区工作的数据顾问，他有着多年的生物制药行业咨询、商业分析、生物信息以及生物工程咨询方面的经验。他拥有华盛顿大学生物统计的博士学位，并且有 40 多篇被审稿人接收的论文。他对统计机器学习非常感兴趣，并且非常乐于接受有趣和有挑战性的项目。对于如何更新更好地进行数据分析，他是非常有激情参与讨论的。他是华盛顿 DC 数据社区的成员，并且是华盛顿 DC 统计编程社群的创始人和联合组织者（华盛顿 DC 地区 R 用户组的前身）。

关于译者和中文版审稿人

郝智恒：甘肃兰州人士，南开大学概率统计专业毕业。统计之都活跃会员。目前在阿里巴巴商业智能部任职，擅长数据分析和数据挖掘。喜欢用数据探索商业世界的边界。

王佳玮：黄山脚下长大，香港城大-中科大联合培养博士毕业。现于阿里云大数据孵化器团队搬砖，喜欢数据分析和挖掘在社会各领域的应用。目前正在致力于用数据和算法解决交通拥堵问题。

谢时光：2011 年博士毕业于美国弗吉尼亚理工大学工业工程系运筹学专业，毕业后从事数据分析、最优化和决策支持相关工作至今。曾先后在安飞士（Avis）、亚马逊（Amazon.com）、费埃哲（FICO）等行业领先的公司从事从供应链到风险控制等多个应用领域的数据分析科学和优化决策研究工作。

刘梦馨：灵雀云高级软件工程师，专注于容器虚拟化领域，机器学习爱好者，个人博客：http://oilbeater.com。

在本书翻译的过程中，来自阿里巴巴商业智能部的李萍、皇甫深龙，以及来自京东搜索推荐部的熊熹、车挣网络数据研发中心的徐浩、华南统计科学研究中心的张晔，来自统计之都社区的李绳、王小宁等人对本书的中文译稿进行了仔细的审阅，并提出了特别多有价值的意见和建议。没有他们，就没有本书的中文版。特别感谢统计之都，这里聚集了一群有情有义的数据人，他们来自世界各地，热心技术交流，乐于分享，通过数据让这个世界变得更美好。

关于英文版审稿人

理查德·海茵曼是 L-3 国家安全方案实验室（NSS）的技术研究员和首席数据科学家（纳斯达克代码：LLL），同时他还是拥有 EMC 认证的数据科学家，专注于空间统计、数据挖掘和大数据。理查德还是 L-3 数据策略业务小组的数据科学团队的领导者。L-3 NSS 和 L-3 数据策略小组都是初级大数据和分析服务提供者，总部在华盛顿，为全球客户服务。

理查德是马里兰大学的客座教授，在那里他教授空间分析和统计推断。同时，他是乔治梅森大学的讲师，在那里他教授人口地理分析。他还是 2014—2015 年的 AAAS 大数据和分析团体的荣誉成员，是华盛顿大数据委员会成员。

理查德最近出版了《基于 R 的社会媒体挖掘》（Packt 出版社）。他现在对 DARPA、DHS、美国军方和 Pentagon 提供数据分析支持。

萨拉·凯丽是一个初级 Python 开发者，也是一个有抱负的数据科学家。她现在在马里兰班赛思达的一家初创公司工作，花费了大量时间进行数据吸收和整合。萨拉拥有西雅图大学的硕士学位。她是一个自学成才的程序员。她也鼓励她的学生们规划数学、科学和技术相关的职业生涯。

Liang Shi 2006 年毕业于佐治亚大学计算机科学专业，并于 2008 年拿到同专业的硕士学位。他博士研究的方向是机器学习和人工智能，主要是解决替代模型的优化问题。毕业后，他加入了 McAfee 的数据挖掘研究团队。他的工作是通过大数据和云计算平台开发机器学习算法来识别网络威胁。之后他加入了微软，作为一名软件工程师，继续开发应用于安全领域的机器学习算法，尤其是针对超大规模、实时的网上广告欺诈识别。2012 年，他作为高级研究员重新加入了 McAfee（英特尔），主导基于云计算平台和机器学习的网络威胁领域的研究。2014 年早些时候，他作为高级数据科学家加入了 Pivotal，主要为企业客户

进行数据科学项目。他对统计和机器学习模型以及理论都非常熟悉，并且熟悉多种编程语言和分析工具。他已经在杂志和会议上出版了多篇论文，此外还出版了一本书的一章。

维尔·沃希思是一个软件开发者，从移动应用开发到自然语言处理，再到网络安全，都非常有经验。他曾在奥地利教授英文，并且开创了一家教育技术初创公司。之后，他搬到了西海岸，加入了一个大型技术公司，现在正在网络安全领域快乐地工作，他们研发的软件被上千开发者使用。

在空闲时间，他喜欢审校技术书籍、看电影，并且安抚他的狗——是的，它是个好女孩。

前言

我们生活在数据时代。每一年，数据都在大量快速地增长，因此分析数据和从数据中创造价值的需求也比以往任何时候都更为重要。那些知道如何使用数据以及如何用好数据的公司，在后续的竞争中会比那些无法使用数据的公司更有优势。基于此，对于那些具备分析能力，能够从数据中提取有价值的洞见，并且将这些洞见用于实践产生商业价值的人才的需求会放大。

本书提供了多种令读者能够学习如何从数据创造价值的机会。书中所用的数据来自很多不同的项目，而这些项目可以体现出最新的数据科学项目的各种维度。每一章的内容都是独立的，包含了电脑屏幕截图、代码片段、必要的详细解释。我们对处理数据的过程和实际应用特别关注。这些内容都是以循序渐进的方式来安排写作的。

写作本书的目的在于，向读者介绍成为数据科学家的路径，以及向读者展示这些方法是如何应用在多种不同的数据科学项目上的。此外，我们还希望读者在今后自己做项目时，能够很方便地应用我们讲到的方法。在本书中，读者将会学到不同的分析和编程课，而所有的课程中所讲授的概念和技能，都是以实际的项目作为引导，因此读者会更好地理解它们。

本书内容

第 1 章，准备你的数据科学环境。本章向读者介绍了数据科学的路径，并且帮助 Mac、Windows 和 Linux 操作系统的读者恰当地搭建数据科学环境。

第 2 章，汽车数据的可视化分析（R）。本章将带领读者对汽车数据进行分析和可视化，

从中发现不同时间燃料效率的变化趋势和模式。

第 3 章，模拟美式橄榄球比赛数据（R）。这一章的项目非常有趣且具有娱乐性。本章分析橄榄球球队进攻防守强度的关系，并且模拟比赛，预测哪个球队会取得胜利。

第 4 章，建模分析股票市场数据（R）。这一章向读者展示如何搭建自己的选股系统，并且使用移动平均法分析股票历史数据。

第 5 章，就业数据的可视化探索（R）。这一章向读者展示如何从劳动统计局获取雇佣和收入数据，并且用 R 对不同水平的数据进行空间分析。

第 6 章，运用税务数据进行应用导向的数据分析（Python）。本章向读者展示如何使用 Python 将自己的分析从一次性临时的工作转变为可复用的产品化的代码。这些工作都是基于一份收入数据展开。

第 7 章，运用汽车数据进行可视化分析（Python）是第 2 章内容的复用，但是这里使用的是强大的编程语言 Python。

第 8 章，社交网络分析（Python），向读者展示如何建立、可视化和分析社交网络。本章所用数据来自于漫画书中的角色关系。

第 9 章，大规模电影推荐（Python）。本章介绍如何用 Python 搭建电影推荐系统。

第 10 章，获取和定位 Twitter 数据（Python）。这一章向读者展示如何调用 Twitter 的 API 获取 Twitter 用户数据，并绘制用户信息中包含的地理信息数据。

第 11 章，利用 NumPy 和 SciPy 优化数值计算（Python）。本章将带领读者领略如何优化 Python 代码，从而在处理大数据集时节省时间和金钱。

阅读本书，你需要什么

要阅读本书，你需要一个能够连接到互联网的电脑，并且能够安装本书项目中所需要的开源软件。本书主要用到的软件包括 R 和 Python，这两个编程语言带有大量的免费的包和库。第 1 章会介绍如何安装这些软件以及它们的包和库。

本书面向读者

本书旨在使用能够亲自实践的现实案例，启发那些想要学习数据科学以及数值编程的

数据科学家们。无论你是一个全新的数据科学家，还是具有丰富的经验，在学习了数据科学项目的结构、数据科学管道的路径以及本书中展示的示例代码之后都会有所收获。本书是按照循序渐进的方式组织内容的，因此并不需要读者具有太多的编程经验。

读者反馈

我们欢迎读者反馈。请让我们知道你对本书的想法——哪些部分你喜欢，哪些部分你不喜欢。对于我们后续开发更多的话题来说，读者反馈是非常重要的。

只需要发送电子邮件到 feedback@packtpub.com，并且在邮件标题中提及本书的书名，在邮件中告诉我们你的想法即可。

如果你对某个领域特别专长，并且你也想写本书或者和别人合作写本书，那么请访问我们的作者指南：www.packtpub.com/authors。

用户支持

作为我们 Packt 图书的荣誉用户，你可以享受以下权益。

下载示例代码

你可以通过你在 http://packtpub.com 的账户，下载所有 Packt 图书的代码。如果你是在其他地方购买的本书，你可以在 http://www.packtput.com/support 上注册，并且通过电子邮件获得代码文件。

下载彩色图片

我们也提供书中彩图的 PDF 文件。这些彩图将会帮助你更好地理解本书的内容。你可以在这里下载这些文件：http://www.packtpub.com/sites/default/files/downloads/0246OS_ColorImages.pdf.

勘误

尽管我们已经尽量细致地检查本书的内容，以免发生不确切的部分，但是书中的错误依然是难免的。如果你发现我们书中的内容有错，无论是内容的错误，还是代码的错误，请你将错误反馈给我们，我们将非常感激。当你发现任何错误，请通过如下网址来反馈：

http://www.packtpub.com/submit-errata。一旦你的勘误信息被确认，那么你提交的信息将被我们接受，并且上传到我们的网站，添加进勘误列表。你可以在 http://www.packtpub.com/support 上查看所有的勘误信息。

版权问题

互联网的版权问题一直是个问题。在 Packt，我们非常严肃地保护我们的版权。如果你在互联网上遇到我们出版图书的任何盗版版本，无论是什么形式，请将该盗版版本的地址和网站名称反馈给我们。

请将盗版链接发送到 copyright@packtpub.com。

我们将非常感谢你对我们的帮助，你的行为能保护我们的作者，从而让他们能够更好地为你提供更多优质的内容。

问答

如果你对我们的图书有任何问题，请通过如下电子邮件联系我们：questions@packtpub.com。我们将尽我们所能帮你解答。

目录

第 1 章　准备你的数据科学环境 …… 1

简介 …… 1

理解数据科学管道 …… 3

处理流程 …… 3

工作原理 …… 3

在 Windows、Mac OS X、Linux 上安装 R …… 5

准备工作 …… 5

处理流程 …… 5

工作原理 …… 7

参考资料 …… 7

在 R 和 RStudio 中安装扩展包 …… 7

准备工作 …… 8

处理流程 …… 8

工作原理 …… 9

更多内容 …… 10

参考资料 …… 10

在 Linux 和 Mac OS X 上安装 Python …… 10

准备工作 …… 11

处理流程 …… 11

工作原理 …… 11

更多内容 …… 11

参考资料 …… 12

在 Windows 上安装 Python ····· 12
处理流程 ····· 13
工作原理 ····· 13
参考资料 ····· 14
在 Mac OS X 和 Linux 上安装 Python 数据分析库 ····· 14
准备工作 ····· 14
处理流程 ····· 14
工作原理 ····· 15
更多内容 ····· 16
参考资料 ····· 16
安装更多 Python 包 ····· 17
准备工作 ····· 17
处理流程 ····· 17
工作原理 ····· 18
更多内容 ····· 18
参考资料 ····· 18
安装和使用 virtualenv ····· 19
准备工作 ····· 19
处理流程 ····· 19
工作原理 ····· 21
更多内容 ····· 21
参考资料 ····· 22

第 2 章　汽车数据的可视化分析（R） ····· 23

简介 ····· 23
获取汽车燃料效率数据 ····· 24
准备工作 ····· 24
处理流程 ····· 25
工作原理 ····· 25
为了你的第一个项目准备好 R ····· 26
准备工作 ····· 26
处理流程 ····· 26
工作原理 ····· 26
参考资料 ····· 26
将汽车燃料效率数据导入 R ····· 27

准备工作 …… 27
处理流程 …… 27
工作原理 …… 28
更多内容 …… 29
参考资料 …… 30
探索和描述燃料效率数据 …… 30
准备工作 …… 30
处理流程 …… 30
工作原理 …… 32
更多内容 …… 33
进一步分析汽车燃料效率数据 …… 34
准备工作 …… 34
处理流程 …… 34
工作原理 …… 43
参考资料 …… 44
研究汽车的产量以及车型 …… 44
准备工作 …… 44
处理流程 …… 44
工作原理 …… 46
更多内容 …… 47
参考资料 …… 47

第 3 章　模拟美式橄榄球比赛数据（R） …… 48

简介 …… 48
准备工作 …… 49
获取和清洗美式橄榄球比赛数据 …… 49
准备工作 …… 50
处理流程 …… 50
工作原理 …… 53
参考资料 …… 53
分析和理解美式橄榄球比赛数据 …… 53
准备工作 …… 53
处理流程 …… 53
工作原理 …… 61
更多内容 …… 61

参考资料 …… 62
构建度量攻防能力的指标 …… 62
准备工作 …… 62
处理流程 …… 62
工作原理 …… 64
参考资料 …… 65
模拟单场由程序决定胜负的比赛 …… 65
准备工作 …… 65
处理流程 …… 65
工作原理 …… 68
模拟多场由计算决定胜负的比赛 …… 68
准备工作 …… 68
处理流程 …… 69
工作原理 …… 73
更多内容 …… 74

第 4 章　建模分析股票市场数据（R） …… 75

简介 …… 75
准备工作 …… 76
获取股票市场数据 …… 76
处理流程 …… 77
描述数据 …… 78
准备工作 …… 79
工作原理 …… 80
更多内容 …… 81
清洗和研究数据 …… 82
准备工作 …… 82
处理流程 …… 82
工作原理 …… 87
参考资料 …… 87
形成相对估值法 …… 87
准备工作 …… 87
处理流程 …… 88
工作原理 …… 91
分析历史价格筛选股票 …… 92