异步图书
www.epubit.com

数据科学实战手册

Practical Data Science Cookbook
Second Edition

第2版

[印] 普拉罕·塔塔（Prabhanjan Tattar）
[美] 托尼·奥赫达（Tony Ojeda），肖恩·帕特里克·墨菲（Sean Patrick Murphy）
本杰明·本福特（Benjamin Bengfort），阿比吉特·达斯古普塔（Abhijit Dasgupta）著
刘旭华 李晗 闫晗 译

基于R和Python的数据预处理、分析及可视化实战教程

Packt>

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

Practical Data Science Cookbook
Second Edition

第2版

[印] 普拉罕·塔塔（Prabhanjan Tattar）
[美] 托尼·奥赫达（Tony Ojeda），肖恩·帕特里克·墨菲（Sean Patrick Murphy）
本杰明·本福特（Benjamin Bengfort），阿比吉特·达斯古普塔（Abhijit Dasgupta）著
刘旭华 李晗 闫晗 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 : 第2版 / (印) 普拉罕·塔塔(Prabhanjan Tattar) 等著 ; 刘旭华, 李晗, 闫晗译. -- 北京 : 人民邮电出版社, 2019.1
ISBN 978-7-115-49925-7

Ⅰ. ①数… Ⅱ. ①普… ②刘… ③李… ④闫… Ⅲ. ①软件工具－程序设计－手册 Ⅳ. ①TP311.56-62

中国版本图书馆CIP数据核字(2018)第244894号

版权声明

◆ 著　[印] 普拉罕·塔塔（Prabhanjan Tattar）
[美] 托尼·奥赫达（Tony Ojeda）
[美] 肖恩·帕特里克·墨菲（Sean Patrick Murphy）
[美] 本杰明·本福特（Benjamin Bengfort）
[美] 阿比吉特·达斯古普塔（Abhijit Dasgupta）

译　刘旭华　李　晗　闫　晗

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

涿州市京南印刷厂印刷

◆ 开本：800×1000　1/16

印张：20.5

字数：386 千字　2019 年 1 月第 1 版

印数：1－2 400 册　2019 年 1 月河北第 1 次印刷

著作权合同登记号　图字：01-2017-3661 号

定价：69.00 元

读者服务热线：(010)81055410　印装质量热线：(010)81055316

反盗版热线：(010)81055315

广告经营许可证：京东工商广登字 20170147 号

内容提要

本书对想学习数据分析的人来说是一本非常实用的参考书，书中有多个真实的数据分析案例，几乎是以手把手的方式教你一步一步地完成从数据分析的准备到分析结果报告的整个流程。无论是数据分析工作的从业者，还是有志于未来从事数据分析工作的在校大学生，都能从本书中获取一些新知识、新思想。

同时，本书也是一本优秀的学习和提高 R 及 Python 编程的参考书。很多人有这样的感触，单纯地学习编程语言是很枯燥的过程，但利用本书学习 R 和 Python 语言可以很好地解决这个问题，生动实用的数据集以及非常有意思的分析结果会极大地激发读者学习的兴趣。

本书案例包括汽车数据分析、税收数据分析、就业数据分析、股市数据分析、社交网络分析、大规模电影推荐、Twitter 数据分析、新西兰海外游客预测分析以及德国信用数据分析等。

关于作者

Prabhanjan Tattar 有 9 年的统计分析工作经验。他的主要精力集中在通过简洁优美的程序解释统计和机器学习技术。生存分析和统计推断是他主要感兴趣和研究的领域，他已经在同行评审期刊上发表了多篇研究论文，并写作了两本关于 R 的书：*R Statistical Application Development by Example*（Packt Publishing）和 *A Course in Statistics with R*（Wiley）。他还在维护几个 R 包：gpk、RSADBE 和 ACSWR。

非常感谢读者的鼓励和反馈，这使得本书（第 2 版）有了很多改进，希望读者从本书中受益。还要感谢 Tushar Gupta 把我介绍到这个项目，感谢 Cheryl Dsa 对我写作拖拉的忍耐，感谢 Karan Thakkar 鹰眼般敏锐的编辑工作以及整个 Packt 团队的大力支持。我还要感谢第 1 版的作者们，因为本书是在他们工作的基础上完成的。在个人方面，我始终感谢我的家人：可爱的 Pranathi、亲爱的妻子 Chandrika、女神般的母亲 Lakshmi 和我深爱着的父亲 Narayanachar。

Tony Ojeda 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及解决方案非常有经验。他在佛罗里达国际大学（Florida International University）获得金融硕士学位，并且在德保罗大学（DePaul University）获得了 MBA 学位。他是华盛顿特区数据实验室的创始人、华盛顿特区数据社区的联合创始人，致力于数据科学的教育事业和活动组织。

Sean Patrick Murphy 在约翰·霍普金斯大学的应用物理实验室做了 15 年的高级科研人员，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、纽约和华盛顿特区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。他还是华盛顿特区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，他也是华盛顿特区数据社区的联合创始人。

Benjamin Bengfort 是一位非常有经验的数据科学家和 Python 开发者。他曾在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（Metacognition）和自然语言处理。他拥有北达科他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特区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。

Abhijit Dasgupta 是在华盛顿特区马里兰-弗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专业的博士学位，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。他是华盛顿特区数据社区的成员，并且是华盛顿特区统计编程社群的创始人和联合组织者（华盛顿特区地区 R 用户组的前身）。

关于译者

刘旭华：现为中国农业大学理学院应用数学系副教授，北京理工大学博士，美国北卡莱罗纳大学教堂山分校（University of North Carolina at Chapel Hill）访问学者，主要从事数理统计、数据科学、数学与统计软件等领域的教学与科研工作，主持及参与过多项国家自然科学基金、北京市自然科学基金等项目。曾翻译过《R 语言统计入门》等书籍。

本书的翻译工作得到中国农业大学教务处 2016—2020 年度“概率论与数理统计”“数学实验”“数理统计”核心课程建设项目、理学院教改项目“大数据背景下的概率统计课程建设探索”的资助。他负责翻译了本书第 1～4 章的内容并对全书进行了统稿。

李晗：2015 年毕业于广州华南理工大学，硕士期间主要从事信号处理、数据分析方面的研究。目前就职于中兴通讯，主要从事数据库、数据分析、容器化微服务方面的开发与运维工作。工作之余，还对多种技术与方向怀有浓厚的兴趣，包括区块链、人工智能、互联网信息安全、技术翻译等。他负责翻译了本书第 8～11 章。

闫晗：中国人民大学统计学院硕士，“统计之都”编辑部“搬砖工”。他负责翻译了本书第 5～7 章。

在本书翻译的过程中，借鉴了由郝智恒、王佳玮、谢时光、刘梦馨等译者翻译的第 1 版译稿，本书译者对他们表示感谢。

关于英文版审稿人

Alberto Boschetti 是一位在信号处理和统计学方面有资深经历的数据科学家。他拥有电信工程博士学位，目前在伦敦生活和工作。在工作中，他每天都要面对自然语言处理、机器学习以及分布式处理领域中的一些具有挑战性的问题。他对工作很有激情，总是追踪最新的数据科学技术的发展，参加相关聚会、会议等交流。他的著作有 *Python Data Science Essentials*、*Regression Analysis with Python* 和 *Large Scale Machine Learning with Python* 等，这几本书都由 Packt 出版社出版。

非常感谢我的家人、朋友和同事们。另外，非常感谢开源社区。

Abhinav Rai 从事数据科学工作快 10 年了，目前在微软工作。他在电信、零售市场及在线广告领域都有工作经验。他感兴趣的领域包括机器学习中不断发展的方法及相关技术。他对分析大规模和海量数据集特别感兴趣，并在这种背景下有一些深刻的见解。他拥有 Deendayal Upadhyay Gorakhpur University 数学硕士学位（NBHM 奖学金）以及印度统计研究所计算机科学硕士学位。严谨和老练是他从事数据分析工作的特点。

前言

欢迎阅读本书。对上一版非常正面的反馈以及读者从中的获益使得本书出版成为可能。当 Packt 出版社邀请我作为第 2 版的合作者时，我在网上事先浏览了一些评论，很快发现了本书流行的原因以及它的一些小小的不足之处。因此，这一版保留了上一版的优点并尽可能地去掉那些缺陷。为提升本书的实用性，本书新增了第 10 章和第 11 章。

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信息，并且将这些信息用于实践产生商业价值的人才的需求会继续增大。本书为读者提供了多种学习如何利用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，而且这些项目可以体现出数据科学项目的各种新进展。每一章的内容都是独立的，它们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，读者将学到不同的分析和编程方法，而所有讲授的概念和技能都是以实际的项目作为引导，因此读者可以更好地理解它们。

本书内容

第 1 章向读者介绍了数据科学管道，并且帮助使用 Mac、Windows 和 Linux 操作系统的读者恰当地搭建数据科学环境。这一章还引导读者在前述平台上安装 R 和 Python。

第 2 章带领读者对汽车数据进行分析和可视化，从中发现随时间变化的燃料效率的变化趋势和模式。在这一章中，读者将初尝获取、探索、修正、分析和沟通等流程。上述过程将在 R 中实现。

第 3 章向读者展示如何使用 Python 将自己的分析从一次性的临时的工作转变为可

复用的产品化代码。这些工作都是基于一份收入数据展开的。

第 4 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。在这一章中，读者将学习如何获取、描述、清洗数据以及如何产生相对估值。

第 5 章展示如何得到美国劳工统计局的就业和薪资数据，然后在 R 中进行不同层级的地理空间分析。相同的分析用 Python 也可以实现。本章的关注点在于数据的转换、操控和可视化。

第 6 章对应于第 2 章的内容，但是使用的是强大的编程语言 Python。这一章关注分析模型的 Python 实现。

第 7 章向读者展示了如何构建、可视化以及分析由漫画书中人物关系构成的社交网络。读者将看到 R 和 Python 两种语言的实现。

第 8 章将带领读者使用 Python 创建一个电影推荐系统。另外，读者也能够学习如何利用 R 和 Python 代码实现一个预测模型，并掌握在实现预测模型时协同过滤的使用方法。

第 9 章向读者展示如何连接到 Twitter API，以及如何画出 Twitter 用户档案中包含的地理信息。另外，读者也能学习 RESTful API 在文本挖掘中的使用方法。

第 10 章解释如何创建时间序列对象，并描述各种可视化时间序列数据的方法。另外，读者也能学到如何为数据创建一个合适的模型，并识别数据中是否包含趋势和季节性组成元素。

第 11 章利用一些基本的树方法和随机森林来展示探索性数据分析。通过本章，读者将能学习到如何对特定数据应用探索性数据分析方法。

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，并且能够安装本书中所需要的开源软件。本书用到的主要软件包括 R 和 Python，这两个编程语言带有大量免费的包和库。第 1 章会介绍如何安装这些软件以及它们的包和库。

本书面向的读者

本书旨在使用能够亲自实践的现实案例，启发那些希望学习数据科学以及数值编程的数据科学工作者。无论你是一名数据科学领域的新手，还是一名具有丰富经验的专家，在学习了数据科学项目的结构以及本书中所展示的示例代码之后都会有所收获。

章节安排

阅读本书时，你会发现几个标题会多次出现（准备工作、操作流程、工作原理、更多内容），下面我们简单介绍这些版块的作用。

准备工作

这个版块告诉你在项目流程中你能得到什么，描述如何安装软件或为完成项目事先需要的设置。

操作流程

这个版块包含为了完成项目所需要的步骤。

工作原理

这个版块通常由对“操作流程”版块内容的详细解释构成。

更多内容

这个版块由与项目流程有关的其他信息构成，目的是使读者对项目流程有更好的理解。

习惯约定

本书中，你将发现对不同的信息用到了许多不同的文本格式。下面有一些格式的例子以及对它们含义的解释。正文中的代码、数据库表名、文件夹名称、文件名、文件扩展名、路径名、虚拟 URL、用户输入、Twitter 用户定位等的格式展示如下：“在数据库中为 JIRA 创建一个新用户，使用如下命令授权用户链接到我们刚创建的 `jiradb` 数据库。”

代码块如下：

```
<Contextpath="/jira"docBase="${catalina.home}
/atlassian- jira" reloadable="false" useHttpOnly="true">
```

命令行输入或输出如下书写：

```
mysql -u root -p
```

新术语和**重要词语**用黑体显示。在屏幕上看到的词语，比如在菜单或对话框中，类似这样的形式：“从**管理面板**选择**系统信息**。”

这种图标表示警告或重要提示。

这种图标表示一些小技巧。

资源与支持

本书由异步社区出品，社区（https://www.epubit.com/）为您提供相关资源和后续服务。

配套资源

本书提供如下资源：

- 源代码；
- 书中彩图文件。

要获得以上配套资源，请在异步社区本书页面中单击 配套资源 ，跳转到下载界面，按提示进行操作即可。注意：为保证购书读者的权益，该操作会给出相关提示，要求输入提取码进行验证。

提交勘误

作者和编辑尽最大努力来确保书中内容的准确性，但难免会存在疏漏。欢迎您将发现的问题反馈给我们，帮助我们提升图书的质量。

当您发现错误时，请登录异步社区，按书名搜索，进入本书页面，单击“提交勘误”，输入勘误信息，单击“提交”按钮即可。本书的作者和编辑会对您提交的勘误进行审核，确认并接受后，您将获赠异步社区的 100 积分。积分可用于在异步社区兑换优惠券、样书或奖品。

扫码关注本书

扫描下方二维码，您将会在异步社区微信服务号中看到本书信息及相关的服务提示。

与我们联系

我们的联系邮箱是 contact@epubit.com.cn。

如果您对本书有任何疑问或建议，请您发邮件给我们，并请在邮件标题中注明本书书名，以便我们更高效地做出反馈。

如果您有兴趣出版图书、录制教学视频，或者参与图书翻译、技术审校等工作，可以发邮件给我们；有意出版图书的作者也可以到异步社区在线提交投稿（直接访问 www.epubit.com/selfpublish/submission 即可）。

如果您是学校、培训机构或企业，想批量购买本书或异步社区出版的其他图书，也可以发邮件给我们。

如果您在网上发现有针对异步社区出品图书的各种形式的盗版行为，包括对图书全部或部分内容的非授权传播，请您将怀疑有侵权行为的链接发邮件给我们。您的这一举动是对作者权益的保护，也是我们持续为您提供有价值的内容的动力之源。

关于异步社区和异步图书

“**异步社区**”是人民邮电出版社旗下 IT 专业图书社区，致力于出版精品 IT 技术图书和相关学习产品，为作译者提供优质出版服务。异步社区创办于 2015 年 8 月，提供大量精品 IT 技术图书和电子书，以及高品质技术文章和视频课程。更多详情请访问异步社区官网 https://www.epubit.com。

“**异步图书**”是由异步社区编辑团队策划出版的精品 IT 专业图书的品牌，依托于人民邮电出版社近 30 年的计算机图书出版积累和专业编辑团队，相关图书在封面上印有异步图书的 LOGO。异步图书的出版领域包括软件开发、大数据、人工智能、软件测试、前端、网络技术等。

异步社区

微信服务号

目录