O'REILLY®

原书第2版

HZ BOOKS
华章IT

# 利用Python进行数据分析

Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython

powered by

jupyter

Wes McKinney 著

徐敬一 译

机械工业出版社
China Machine Press

原书第2版

# 利用 Python 进行数据分析

韦斯·麦金尼（Wes McKinney） 著

徐敬一 译

Beijing · Boston · Farnham · Sebastopol · Tokyo O'REILLY®

O'Reilly Media, Inc. 授权机械工业出版社出版

机械工业出版社

**图书在版编目（CIP）数据**

利用 Python 进行数据分析（原书第 2 版）/(美) 韦斯·麦金尼 (Wes McKinney) 著；徐敬一译．－北京：机械工业出版社，2018.6（2018.9 重印）

(O'Reilly 精品图书系列)

书名原文：Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython, 2nd Edition

ISBN 978-7-111-60370-2

I. 利… II. ①韦… ②徐… III. 统计分析－应用软件 IV. C819

中国版本图书馆 CIP 数据核字（2018）第 150175 号

北京市版权局著作权合同登记

图字：01-2017-8013 号



书　　名 / 利用 Python 进行数据分析（原书第 2 版）

书　　号 / ISBN 978-7-111-60370-2

责任编辑 / 冯秀泳

封面设计 / Karen Montgomery，张健

出版发行 / 机械工业出版社

地　　址 / 北京市西城区百万庄大街 22 号（邮政编码 100037）

印　　刷 / 三河市宏图印务有限公司

开　　本 / 178 毫米 ×233 毫米　16 开本　30.75 印张

版　　次 / 2018 年 7 月第 1 版　2018 年 9 月第 2 次印刷

定　　价 / 119.00 元（册）

**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

**购书热线：**(010)68326294；88379649；68995259

**投稿热线：**(010)88379604

**读者信箱：**hzit@hzbook.com

# 译者序

2014 年我刚走上工作岗位时，Python 数据分析技术开始在中国流行，越来越多的人开始学习和使用这门技术。当时，Python 的主流版本是 2.7，3.4 版本刚刚发布，但认可度和流行度并不高。很幸运，我阅读了本书的第 1 版。第 1 版基于 Python 2.7，介绍了大量的实用技术，包括 pandas、NumPy 的使用等。在实践中，书中的技术被我大量应用，帮助我更快更好地处理了各类数据。

时过境迁，从本书英文版第 1 版 2012 年出版至今，已经过去了 6 年。这 6 年中，Python 的主流版本从 2.7 升级到了 3.6。无论是否情愿，大部分 Pythoner 都不得不学会适应新版本；而 pandas 则从 0.1.0 版本迭代到如今的 0.22.0 版本。版本号的持续增加意味着新技术、新特性的不断丰富。举例来说，将带有多层索引的数据透视表写入 Excel 在 2015 年之前是无法使用 pandas 完成的，在 0.17 版本后该功能被加入。因此，本书的第一版内容已经略显落后。

2017 年 10 月下旬，本书作者 Wes McKinney 先生更新了本书的第 2 版。在第 2 版中，他将 Python 版本更新到 3.6，介绍了 pandas 的一些新接口和功能，并新增了大量现实世界的数据分析实例，以确保本书的可实践性。我在 2017 年 11 月 18 日接到出版社的翻译邀请后便开始了翻译工作。McKinney 先生的写作风格朴实、形象，因而我的翻译过程较为顺畅，但书中部分口语化的叙述也因为中英文表达方式的差异而增加了意译的难度。

在本书翻译过程中，我得到了很多帮助。首先要感谢华章公司的王春华编辑和冯秀泳编辑，他们在翻译过程中给了我耐心指导。在审稿时，来自国内 Python 圈的朋友们对本书进行了仔细而全面的审核，他们是网易数据工程师马喆诚、早稻田大学研究生梁堉、大疆工程师王波。感谢他们对本书的付梓而做出的贡献。此外，我还要感谢我的女朋友易

慧娟，感谢你在生活中对我的各种好，我爱你！

为了让国内读者在第一时间读到这本畅销国外技术著作，出版社和我都加快了工作进度，但时间紧、任务重，再加之本人水平有限，翻译工作中难免会出现一些失误。欢迎读者将阅读过程中发现的问题发送至我的邮箱（xujingyi46@163.com）。

本书英文版的副书名是“Data Wrangling with Pandas, NumPy, and IPython”，其中Wrangling是一个很难直译的词汇，它的原意是争执、争论，但在书中它描述的是将数据进行规整、处理的意思。希望读者读完本书后，可以使用好pandas、NumPy和IPython这些工具，更好地完成数据处理、分析的学习和工作。Enjoy it！

徐敬一

2018年5月于中国工商银行合肥后台中心

# 目录

# 前言

## 第 2 版新内容

本书第 1 版出版于 2012 年，彼时基于 Python 的开源数据分析库（例如 pandas）仍然是一个发展迅速的新事物。在本次更新、拓展的第 2 版中，我在一些章节内进行了修改，以解释过去 5 年中发生的不兼容的变更、弃用和一些新特性。此外，我还添加了新内容，用以介绍在 2012 年还不存在或者不成熟的工具。最后，我会避免把一些新兴的或者不太可能走向成熟的开源项目写入本书。我希望本版的读者能够发现本书内容在 2020 年或者 2021 年仍然几乎像在 2017 年一样适用。

第 2 版中的主要更新包括：

- 所有的代码，包括把 Python 的教程更新到了 Python 3.6 版本（第 1 版中使用的是 Python 2.7）
- 更新了 Python 第三方发布版 Anaconda 和其他所需 Python 包的安装指引
- 更新 pandas 库到 2017 年的最新版
- 新增一章，关于更多高级 pandas 工具和一些使用提示
- 新增 statsmodels 和 scikit-learn 的简明使用介绍

除了以上更新内容，我还重新组织了第 1 版的部分重要内容，使本书对新手来说更易于理解。

## 本书约定

以下印刷约定将在本书中使用：

斜体（*Italic*）

表示新的术语、URL、email 地址、文件名和文件扩展名。

等宽字体（Constant width）

用于程序清单以及段落中的程序元素，例如变量名、函数名、数据库、数据类型、环境变量、表达式和关键字等。

等宽粗体（**Constant width bold**）

表示命令或其他应当由用户键入的文本。

等宽斜体（*Constant width italic*）

表示应当由用户提供的值来替代的文本，或者其他由上下文决定的值。

本符号表示提示或建议。

本符号表示一般性说明。

本符号表示警告。

# 使用代码示例

可以通过本书的 GitHub 仓库获取本书每一章中的数据文件和相关材料。GitHub 仓库地址：*http://github.com/wesm/pydata-book*。

本书的目的在于帮助你完成工作。一般来说，本书提供的示例代码，你可以在你的程序或文档中使用而无须联系我们获取许可，除非你需要重造大量代码。举例来说，使用本书中的代码段编写程序无须授权许可，但销售或发行 O'Reilly 图书的 CD-ROM 代码示例则需要许可。引用本书代码回答问题不需要许可，但在你的产品文档中大量使用本书示例代码则需要许可。

我们鼓励注明资料来源的行为，但这并不是必需的。来源注明通常包括书名、作者、出版社及 ISBN，例如："*Python for Data Analysis* by Wes McKinney(O'Reilly). Copyright 2017 Wes McKinney, 978-1-491-95766-0"。

如果你认为你对本书示例代码的使用超过了正常使用范围或者需要以上介绍的授权许可，请联系 *permissions@oreilly.com*。

## O'Reilly Safari

Safari（前身为 Safari Books Online）是一个会员制的培训、参考网站，服务于企业、政府、教育者和个人。

会员可以访问数千书籍、培训视频、学习路径、交互教程和超过 250 家出版商的企划列表，包括 O'Reilly Media、Harvard Business Review、Prentice Hall Professional、Addison-Wesley Professional、Microsoft Press、Sams、Que、Peachpit Press、Adobe、Focal Press、Cisco Press、John Wiley & Sons、Syngress、Morgan Kaufmann、IBM Redbooks、Packt、Adobe Press、FT Press、Apress、Manning、New Riders、McGraw-Hill、Jones & Bartlett、Course Technology 等。

更多信息，请访问 *http://oreilly.com/safari*。

## 如何联系我们

对于本书如果有任何意见或疑问，请按照以下地址联系本书出版商。

美国：

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472

中国：

北京市西城区西直门南大街 2 号成铭大厦 C 座 807 室（100035）
奥莱利技术咨询（北京）有限公司

我们为本书准备了一个网页，用于陈列勘误、示例和其他附加信息。访问地址是：*http://bit.ly/python_data_analysis_2e*。

针对本书评论或提出技术问题，请发送邮件至：*bookquestions@oreilly.com*。

关于本书的更多信息、课程、会议及新闻，请访问我们的网站：*http://www.oreilly.com*。

Facebook 联系我们：*http://facebook.com/oreilly*
Twitter 联系我们：*http://twitter.com/oreillymedia*
YouTube 观看我们的视频：*http://www.youtube.com/oreillymedia*

# 致谢

本书是全世界很多人多年来富有成效的讨论、协作和支持的成果。我想对他们中的一些代表致以谢意。

## 怀念：John D. Hunter（1968—2012）

我们亲爱的朋友和同行 John D. Hunter 在经历了一场与结肠癌的战斗后，于 2012 年 8 月 28 日离开了世界。那时正是我完成本书第 1 版最终手稿后不久。

John 对 Python 科学计算和数据社区的影响之大难以估量，他给我们留下的遗产价值非凡。除了在 2000 年初期开发 matplotlib 之外（那时 Python 还没有当下如此流行），他还帮助塑造了一代核心开源开发者的文化，如今这些开发者已经成为 Python 生态系统的顶梁柱，而 Python 生态系统对于现如今的我们来说似乎是理所当然的。

在 2010 年 1 月，我开源生涯的早期，那时候 pandas 刚刚发布了 0.1 版本，我便有幸结识了 John。即便在最黑暗的时期，他的才华和指导仍在帮助我推动 pandas 前进，实现 Python 成为数据分析第一语言的愿景。

John 与 IPython、Jupyter 项目的先锋 Fernando Pérez、Brian Granger 及其他很多 Python 社区的倡议人联系紧密。我们四人曾经希望共同写作一本书，但只有我个人时间最为自由，所以这个想法被搁置了。我非常确信他会为过去 5 年中我们个人及我们社区所取得的成就感到骄傲。

## 第 2 版致谢（2017）

距离我在 2012 年 7 月完成第 1 版手稿已经 5 年了。很多事情都发生了变化。Python 社区获得了极大的成长，围绕 Python 的开源软件生态系统也十分繁荣。pandas 核心开发者孜孜不倦的付出，使得 pandas 项目高速成长，也使得 pandas 的用户群体遍布 Python 数据科学生态系统的各个角落，没有他们本书将不会存在。pandas 的核心开发者包括但不限于：Tom Augspurger、Joris van den Bossche、Chris Bartak、Phillip Cloud、gfyoung、Andy Hayden、Masaaki Horikoshi、Stephan Hoyer、Adam Klein、Wouter、Overmeire、Jeff Reback、Chang She、Skipper Seabold、Jeff Tratner 和 y-p。

在第 2 版的实际写作过程中，非常感谢 O'Reilly 的工作人员在写作进程中给予的耐心帮助。他们是 Marie Beaugureau、Ben Lorica 和 Colleen Toporek。我再次得到了优秀技术审阅人的支持，他们是 Tom Augpurger、Paul Barry、Hugh Brown、Jonathan Coe 和 Andreas Müller。感谢你们。