Broadview
www.broadview.com.cn

行业畅销书升级版

告诉你职场图表背后的故事

谁说菜鸟不会数据分析

(Python篇)

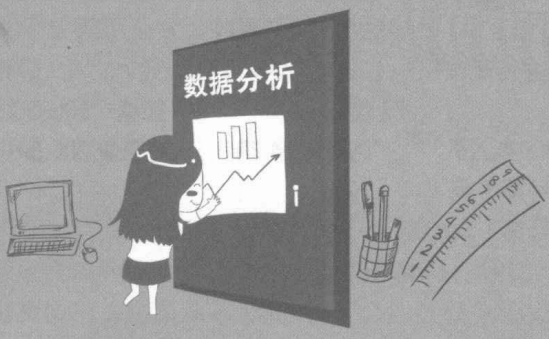
方小敏 张文霖 著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

谁说菜鸟不会数据分析

(Python篇)

方小敏 张文霖 著



電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内容简介

本书从解决工作实际问题出发，提炼总结工作中 Python 常用的数据处理、数据分析实战方法与技巧。本书力求通俗易懂地介绍相关知识，在不影响学习理解的前提下，尽可能地避免使用晦涩难懂的 Python 编程、统计术语或模型公式。

本书定位是带领 Python 数据分析初学者入门，并能解决学习、工作中大部分的问题或需求。入门后如还需要进一步进阶学习，可自行扩展阅读相关书籍或资料，学习是永无止境的，正所谓“师傅领进门，修行在个人”。

图书在版编目（CIP）数据

谁说菜鸟不会数据分析．Python 篇 / 方小敏，张文霖著．—北京：电子工业出版社，2019.6
ISBN 978-7-121-36458-7

Ⅰ．①谁… Ⅱ．①方… ②张… Ⅲ．①表处理软件②软件工具—程序设计 Ⅳ．① TP391.13 ② TP311.561

中国版本图书馆 CIP 数据核字（2019）第 085523 号

责任编辑：张月萍
印　　刷：中国电影出版社印刷厂
装　　订：中国电影出版社印刷厂
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：720×1000　1/16　　印张：14.75　　字数：306.8 千字　　彩插：1
版　　次：2019 年 6 月第 1 版
印　　次：2019 年 6 月第 1 次印刷
印　　数：8000 册　　定价：69.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：（010）51260888-819，faq@phei.com.cn。

前　言

《谁说菜鸟不会数据分析》系列图书自上市以来，已拥有数十万读者与粉丝，口口相传，成为职场人士案头必备的参考用书。同时非常荣幸地获得书刊发行业协会授予的“全行业优秀畅销品种”称号，这离不开广大读者的厚爱与支持。有读者告诉我们，每次阅读都会有新的体会与收获，这让我们很开心。

随着云计算、互联网、电子商务和物联网的飞速发展，世界已经逐步迈入大数据时代。数据分析、机器学习等数据科学技术也相应流行起来，主流的数据科学技术，都将 Python 作为主要的计算工具。Python 越来越被大家熟悉和认可，成为数据分析师的新宠儿，特别是在互联网行业。

市面上 Python 数据分析的相关书籍基本上多数由 IT 人员编写，写作角度相对侧重技术层面，很多基础知识点和编写的代码并无详细介绍，并且在数据分析思维体系方面相对薄弱，学习门槛非常高，让非 IT 专业朋友学起来较为痛苦。

鉴于此，本书作者于 2015 年开始提炼总结工作中 Python 常用的数据处理、数据分析实战方法与技巧，并录制成了视频课程《Python 数据分析实战》，发布于网易云课堂。课程上线后，受到了大量学员的支持与肯定。同时，课程上线后，根据热心学员的宝贵反馈意见，对课程不断进行升级更新。

通过《Python 数据分析实战》视频课程的录制、升级过程中，沉淀了大量的 Python 数据分析实战教学经验。同时大量的学员与读者不断来信咨询希望早日出版《谁说菜鸟不会数据分析（Python 篇）》。经过两年时间的打磨，这本书终于与读者见面了。整个写作过程是艰辛的，但是也很有成就感。

本书从解决工作实际问题出发，提炼总结工作中 Python 常用的数据处理、数据分析实战方法与技巧。本书与其他《谁说菜鸟不会数据分析》系列图书一样，力求通俗易懂地介绍相关知识，在不影响学习理解的前提下，尽可能地避免使用晦涩难懂的 Python 编程、统计术语或模型公式，如需了解相关的知识，可查阅相关的书籍或资料。

本书的定位是带领 Python 数据分析初学者入门，并能解决学习、工作中大部分的问题或需求。入门后如还需进一步进阶学习，可自行扩展阅读相关书籍或资料，学习是永无止境的，正所谓“师傅领进门，修行在个人”。

本书结构

本书以数据分析主要流程为主线，介绍如何用 Python 进行数据分析。

第 1 章　数据分析概况：主要通过 2W1H 模型介绍数据分析相关知识，让读者了解与认识数据分析。

第 2 章　Python 概况：主要介绍了什么是 Python，Python 的特点，Python 的函数与模块，Python 的使用场景，以及 Anaconda 的安装与使用，让读者了解与认识 Python。

第 3 章　编程基础：主要介绍了 Python 进行数据分析所需要的编程基础，包括数

据类型、赋值和变量、数据结构、向量化运算、for 循环，让读者对 Python 在数据分析方面的使用有基本的了解与认识。

第 4 章　数据处理：主要介绍了在 Python 中如何使用 Pandas 进行数据处理操作，包括数据导入与导出、数据清洗、数据转换、数据抽取、数据合并、数据计算，让读者能够使用 Python 进行常用的数据处理操作。

第 5 章　数据分析：主要介绍了在 Python 中如何使用 Pandas、sklearn 进行相关的数据分析操作，包括描述统计分析、分组分析、结构分析、分布分析、交叉分析、RFM 分析、矩阵分析、相关分析、回归分析等常用分析方法，让读者能够使用 Python 进行常用的数据分析操作。

第 6 章　数据可视化：主要介绍了在 Python 中如何使用 matplotlib.pyplot 进行常用的数据可视化图形绘制，包括散点图、矩阵图、折线图、饼图、柱形图、条形图，让读者能够使用 Python 进行常用的数据可视化图形绘制。

本书主要基于 Python 3 进行介绍，故部分方法可能在 Python 2 中无法实现。

适合人群

★ 需要提升自身竞争力的职场新人。
★ 从事咨询、研究、分析等专业人士。
★ 在市场营销、产品运营、项目管理、开发运维等工作中需要进行数据分析的人士。

案例数据下载

本书配套案例数据下载方式：

（1）扫码关注微信订阅号：小蚊子数据分析（wzdata），回复“1”或“Python 篇”获取案例数据下载链接

（2）http://blog.sina.com.cn/xiaowenzi22

致谢

感谢广大读者与学员的支持，让笔者下定决心写这本书。在此要衷心感谢成都道然科技有限责任公司的姚新军先生，感谢他的提议和在写作过程中的支持。感谢参与本书优化的朋友：王斌、李伟、范霈璐、李萍、王晓、景小艳、余松。非常感谢本书的插画师朴提的辛苦劳动，您的作品也让本书增色不少。

感谢沈浩、张文彤、路人甲、黄成明、阿橙、许树淮、肖骁、严婷、刘志军、崔庆才、齐德胜、数据小人、郑来轶、李舰、gashero、肖凯、郑跃平等书评作者，感谢他们在百忙之中抽空阅读书稿，撰写书评，并提出宝贵意见。

最后，要感谢两位作者的家人，感谢他们默默无闻的付出，没有他们的理解与支持，同样也没有本书。

尽管我们对书稿进行了多次修改，仍然不可避免地会有疏漏和不足之处，敬请广大读者批评指正，我们会在适当的时间进行修订，以满足更多人的需要。

业内人士的推荐（排名不分先后，以姓氏拼音排序）

本书更是为非专业人士提供了应用 Python 进行数据处理的入门途径。在示例和引用的第三方库方面，本书更靠近入门者的习惯，而非专家习惯，使得入门数据分析的过程更加平滑而减少挫折，同时也避免了很多入门者常见的学了 Python 却不知道怎么用的难题。在努力降低入门门槛的同时，也没有避开一些常见的难点，比如数据清洗和多种输入输出文件类型的支持，使得本书避免成为一本纯入门的书籍。

gashero
Python 技术专家

本书将 Python 数据分析相关的模块和分析理论相结合，深入浅出地向读者阐述数据分析方法论，无论是对刚入门的业界新手，还是有经验的职场人士，都是工作学习中不可多得的一位良师益友。

阿橙
“Python 中文社区”微信公众号主理人

Python 现在已经成为数据分析的一大利器，本书从实战出发讲解了一系列使用 Python 进行数据分析的必备知识点，书中案例附有详细的案例图示和代码说明，以帮助读者更好地学习和理解。

崔庆才
《Python 3 网络爬虫开发实战》作者

读完本书，你会发现数据分析的乐趣，它并不是那么枯燥，数据背后的故事简直是太有意思了。从此你将发现：无论是新闻媒体，还是企业报表中的数字将不再孤独，因为他们在那里，在和你说着话！祝愿大家早日练就一颗数据分析的“芯”！

黄成明
《数据化管理》作者，数据化管理顾问及培训师

由浅入深、循循善诱，是一本真正站在数据分析角度的 Python 书籍。

李舰
《统计之美》作者，统计之都核心成员

这是一本对初学者非常友好的书，它将带你开启数据分析之旅。

刘志军

“Python 之禅”微信公众号主理人

两年前开始学习数据分析，因为《谁说菜鸟不会数据分析》而入门，这本书对我的影响非常大。书中的各种数据分析案例生动形象，让初学者学习起来没有丝毫的压力。《谁说菜鸟不会数据分析（Python 篇）》这本书仍然延续了系列书的风格，对于希望入门数据分析、想系统学习数据分析方法论的同学来说是一本非常值得一读的书。

路人甲

“路人甲 TM”微信公众号主理人

这是一本非常适合初学者入门的书，书中既讲解了数据分析的思路和统计学的基础知识，又提供了丰富的案例，将方法与应用紧密联系起来，还有详细的可实现的代码供读者练习。另外，这本书不仅可以作为初学者入门之选，其函数涉及之全面、参数介绍之详细，完全可以作为日常学习工作中的工具书来随时查看，是一本数据分析师的“必备宝典”！

齐德胜

中国气象局华风集团－华风象辑 研发副总监

当谈到用数据解决问题时，我经常用这样的语言去诠释：“如果你不能量化它，你就不能理解它，如果不理解它，就不能控制它，不能控制它，也就不能改变它”。数据无处不在，信息时代的主要特征就是“数据处理”，数据分析正以我们从未想象过的方式影响着日常生活。

在知识经济与信息技术时代，每个人都面临者如何有效地吸收、理解和利用信息的挑战。那些能够有效利用工具从数据中提炼信息、发现知识的人，最终往往成为各行各业的强者！

这本书向我们清晰又友好地介绍了数据分析方法、技巧与工具，欢迎来读一读这本书，或许会给你带来更大的惊喜！

沈浩教授

中国传媒大学新闻学院博士生导师，

中国传媒大学调查统计研究所所长，

大数据挖掘与社会计算实验室主任，

中国市场研究协会会长

业内人士的推荐（排名不分先后，以姓氏拼音排序）

数据分析用 Python，学分析工具 Python，用好本书就够了，基础知识、方法、流程、案例，应有尽有。

数据小人

腾讯高级数据分析师，连续创业者

市面上有很多面对初学者的 Python 书籍，大多数偏向于介绍语言本身。很多时候学完了语言却不清楚下一步应该做什么，这种情况下就需要有一本能面向具体应用场景，又不是那么厚重的书来带领大家入门。本书把数据分析的细节掰开讲透，一步步地讲清楚了各参数的含义，非常细致和有章法。对于希望从 Excel 迁移到 Python 的数据分析用户来讲，这是一本不错的入门读物。

肖凯

蚂蚁金服数据技术专家

Python语言用途广泛，很容易让初学者迷失方向。本书是新手数据分析师的指路标，Python 数据分析入门，请从这本书设定的学习路径开始。

肖骁

58 同城数据分析师

俗话说万事开头难，入门一门新的编程语言也是一件令人头痛的事。但是这本书既简洁又全面，并且简单易懂的方式重新组织了整个知识体系，让小白的 Python 入门之路更加平坦。这应该是每一位 Python 小白入门的第一本书。

许树淮

东风本田发动机有限公司 市场数据分析师

这本书基于工作中常用的数据分析实战方法与案例，通过结合 Excel、Sql 等实现类比，通俗易懂地介绍 Python 的实现方法逻辑，大大降低了学习门槛，本书堪称 Python 数据分析入门不二之选！

严婷

猎聘网 数据分析师

统计学是一门很难，但是很有趣，更很有用的工具学科。懂得如何使用他的人总是乐在其中，而尚未入门的人则畏之如虎。国内讲述统计学理论，以及讲述统计软件操作的书籍可谓汗牛充栋，但是多数流于理论，疏于应用和实践指导。存在着明显未被满足的读者需求。

近年来随着信息技术的普及，各行各业的业务数据自动化趋势愈来愈明显，使得数据分析的需求开始从统计专业人士向各行业人员全面扩展。在此背景之下，一本能够深入浅出，从实际应用的角度介绍基本统计分析知识的书就变得很有必要。

本书在理论和实践的平衡方面做了很有价值的尝试，基于很为普及的 5W2H、PEST 等数据分析方法论为指导，深入浅出的介绍了如何满足具体工作中的常见统计分析需求，对于需要应用统计分析，但是又未接受过这方面系统培训的读者来说，本书应当是一本非常合适的数据分析入门教材。

张文彤博士

上海昊鲲企业管理咨询有限公司 合伙人

此书秉承《谁说菜鸟不会数据分析》系列图书的特点，结构有层次、内容全面、通俗易懂，一步步引导你走进数据分析的世界、手把手教你如何使用 Python 进行数据处理、数据分析和数据呈现。针对数据分析新人，是一本从了解数据分析到自己动手操作、逐步递进的好图书。

郑来轶

数据分析网创始人，JollyChic 数据分析总监

迈入大数据时代后，就理论研究和实践创新而言，Python 都已成为重要的数据分析工具。本书通过完整的结构、清晰的构思、严谨的逻辑、生动的语言，为初学者设计了一条学习和使用 Python 的有效路径。

郑跃平

中山大学政务学院助理教授

目　录

第1章

数据分析概况

21 世纪是一个数据信息爆炸性增长的时代，随着云计算、互联网、电子商务和物联网的飞速发展，世界已经逐步迈入大数据时代。数据分析在各个行业的应用越来越广泛，与之相应的是，决策也将会越来越依靠数据分析做出，而不是依靠个人直觉和经验。

> *If you can't measure it, you can't manage it.*
>
> — *Peter F. Drucker*

管理学大师彼得·德鲁克曾经说过：如果你无法衡量它，就无法管理它，这其实说的就是数据分析。那么数据分析究竟是什么呢？我们可以使用 2WIH 模型解答这个问题，也就是 What——数据分析是什么？ Why——数据分析有什么用？ How——数据分析如何做？

1.1 数据分析定义（What）

数据分析的目的是把隐藏在一大批看似杂乱无章的数据背后的信息集中和提炼出来，总结出所研究对象的内在规律。数据也称观测值，是通过实验、测量、观察、调查等方式获取的结果，常常以数量的形式展现出来。

在实际工作当中，数据分析能够帮助管理者进行判断和决策，以便采取适当的策略与行动。例如，公司管理者希望通过市场分析和研究，把握当前产品的市场动向，从而制订合理的产品研发和销售计划，这就必须依赖数据分析才能完成。

数据分析可以分为广义的数据分析和狭义的数据分析（如图 1-1 所示），广义的数据分析包括狭义的数据分析和数据挖掘，我们常说的数据分析通常指的是狭义的数据分析。

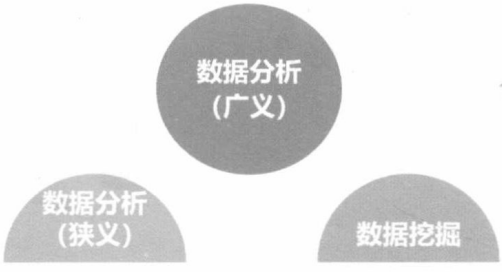


图 1-1　数据分析分类

1. 数据分析（狭义）

（1）定义：数据分析是指根据分析目的，用适当的分析方法及工具，对数据进行

处理与分析，提取有价值的信息，形成有效结论的过程（参见图 1-2）。

（2）作用：它主要实现三大作用，分别是现状分析、原因分析、预测分析，这里的预测分析主要是指数值预测分析。数据分析的目标明确，先做假设，然后通过数据分析来验证假设是否正确，从而得到相应的结论。

（3）方法：主要采用对比分析、分组分析、结构分析、分布分析、交叉分析、矩阵分析、回归分析等常用分析方法。

（4）结果：数据分析一般是得到一个指标统计量结果，如总和、平均值、计数等，这些指标数据需要与业务结合进行解读，才能发挥出数据的价值与作用。

项目	数据分析	数据挖掘
定义	指根据分析目的，用适当的分析方法及工具，对数据进行处理与分析，提取有价值的信息，形成有效结论的过程	从大量的数据中，通过统计学、机器学习、数据可视化等方法，挖掘出未知且有价值的信息和知识的过程
作用	现状分析、原因分析、预测分析	解决四类问题：分类、聚类、关联、预测
方法	对比分析、分组分析、结构分析、分布分析、交叉分析、矩阵分析、回归分析等	决策树、神经网络、关联规则、聚类分析、时间序列分析等
结果	指标统计量结果，如总和、平均值等	输出模型或规则

图 1-2　数据分析与数据挖掘

2. 数据挖掘

（1）定义：数据挖掘是指从大量的数据中，通过统计学、机器学习、数据可视化等方法，挖掘出未知且有价值的信息和知识的过程，如图 1-3 所示。

图 1-3　数据挖掘相关方法

（2）作用：数据挖掘主要侧重解决四类问题，分别是分类、聚类、关联和预测，数据挖掘的重点在于寻找未知的模式与规律。例如我们常说的数据挖掘案例：啤酒与尿布、安全套与巧克力等，这就是事先未知但又是非常有价值的信息。

（3）方法：主要采用决策树、神经网络、关联规则、聚类分析、时间序列分析等涉及统计学、机器学习等相关领域的方法进行挖掘。

（4）结果：输出模型或规则，并且可相应得到模型得分或标签，模型得分如流失概率值、综合得分、相似度、预测值等，标签如流失与非流失、高中低价值用户、信用优良中差等。

综合起来，数据分析（狭义）与数据挖掘的本质是一样的，都是从数据里面发现关于业务的知识（有价值的信息），从而帮助业务运营、改进产品以及帮助企业做更好的决策。所以数据分析（狭义）与数据挖掘构成广义的数据分析。本书后续所说的数据分析均指狭义的数据分析。

1.2 数据分析作用（Why）

那么数据分析在我们日常运营分析工作中具体有哪些作用呢？体现在哪几方面呢？

数据分析要达到帮助管理者的有效决策提供有价值信息的目的，那么我们在日常数据分析工作中该做些什么呢？比如日常通报、专题分析等，这些就是数据分析具体工作的体现。而什么时候做通报工作，什么时候开展专题分析，这都需要我们根据实际情况做出选择。很多人经常做这些工作，但不知为何而做，只是为做而做，只有当你对数据分析的目的及作用有了足够清晰、体系的正确认识后，数据分析开展才能如鱼得水，游刃有余。

数据分析在我们日常运营分析工作中主要有三大作用，如图 1-4 所示。

图 1-4　数据分析三大作用

1. 现状分析

简单来说就是告诉你过去发生了什么，具体体现在：

第一，告诉你企业现阶段的整体运营情况，通过各个经营指标完成情况来衡量，以说明企业整体运营是好了还是坏了，好的程度如何，坏的程度又到哪里。