

Python数据分析与算法指南（套装共8册）

作者：江雪松

总目录

[Python数据分析](#)
[Python大数据分析算法与实例](#)
[Python机器学习：数据分析与评分卡建模\(微课版\)](#)
[Python数据可视化之Matplotlib与Pyecharts](#)
[Python数据预处理技术与实践](#)
[Python科学计算（第2版）](#)
[Python量化交易实战](#)
[Python王者归来](#)

本书封面贴有清华大学出版社防伪标签，无标签者不得销售。

版权所有，侵权必究。侵权举报电话：010-62782989 13701121933

图书在版编目（CIP）数据

Python数据分析 / 江雪松，邹静编著. —北京：清华大学出版社，2020.6

（人工智能科学与技术丛书）

ISBN 978-7-302-55517-9

I. ①P... II. ①江...②邹... III. ①软件工具-程序设计 IV. ①TP311.561

中国版本图书馆CIP数据核字（2020）第085999号

责任编辑：盛东亮 吴彤云

封面设计：李召霞

责任校对：李建庄

责任印制：杨 艳

出版发行：清华大学出版社

网 址：http://www.tup.com.cn, http://www.wqbook.com

地 址：北京清华大学学研大厦A座

邮 编：100084

社总机：010-62770175

邮 购：010-62786544

投稿与读者服务：010-62776969, c-service@tup.tsinghua.edu.cn

质量反馈：010-62772015, zhiliang@tup.tsinghua.edu.cn

课件下载：http://www.tup.com.cn, 010-83470236

印刷者：清华大学印刷厂

经 销：全国新华书店

开 本：185mm×260mm

印 张：20

字 数：487千字

版 次：2020年7月第1版

印 次：2020年7月第1次印刷

印 数：1~2500

产品编号：084953-01

作者简介

江雪松 2001年获工学硕士学位，曾服务于华为、诺基亚等公司，有丰富的产品团队与研发团队管理经验。先后担任项目经理、研发经理、产品研发负责人，管理多个产品全球支持与交付团队、软件维护业务。

邹静 2006年获理学硕士学位，现服务于四川省人民医院，担任副主任医师。拥有十多年的临床药物实验设计、数据统计与分析经验。先后参加了多个国家自然科学基金项目及四川省卫生厅科研项目。

内容简介

本书是一本系统讨论Python数据分析基础与案例实战的教程。全书共分为上下两篇：上篇为Pandas数据分析基础（第1章～第10章），首先介绍了数据分析及其基本流程、如何构建基于Python的数据科学开发环境，然后深入讨论了如何利用Python中的Pandas库进行基本的数据操作、数据清洗、数据整理，以及如何对数据进行可视化，最后用一个电商销售数据的分析案例对上篇的知识进行了总结；下篇为Python数据分析实战（第11章～第21章），尽可能多地为读者展示各种数据分析应用，目的是让读者体会到数据分析的作用。此外，下篇也对时间序列数据和大规模数据分析等进行了讨论。

为便于读者高效学习，快速掌握Python数据分析，本书提供全部的配套源代码以及数据。此外，作者单独录制了配套视频课程，针对该课程提供答疑。本书也额外提供了如何利用Basemap库和Folium库完成地理信息数据的可视化教程，供读者单独下载。

本书适合作为高校数据分析相关专业的课程教材，也可以作为从事数据分析的广大工作者的自学参考用书。

前言

根据Cumulus Media 2018年的数据，全球每分钟向谷歌发起370万次搜索，发送1800万条短信，YouTube上有430万条视频被观看，Facebook登录超过97万次，等等。人们刷微博、聊微信、用滴滴打车、用淘宝购物、用美团点外卖，每一条信息，每一次出行，每一次消费，每一次互动都成为其数字足迹！未来已来，我们进入了数据时代！

本人与数据正式结缘还要追溯到2003年担任软件项目经理时，当时分析项目团队的软件问题成为我的第一个数据分析任务。2009年，我开始管理多个产品的全球支持团队，与数据的关系更进了一步。作为通信设备商，我们的客户支持有本地技术支持、欧洲及美洲等区域技术支持以及全球技术支持。每层的技术人员解决的问题类型、难度、时间以及成本都不同，电信运营商的设备故障每分每秒都将给客户带来巨大影响。面对成千上万的客户问题，如何快速、高效、低成本地处理各种问题，需要用数据来帮助我们做出决策。2013年后，我开始转向负责软件维护业务，此时我关心的是：如何从软件维护中获取更多收入；如何更好地提高客户服务质量；针对全球的客户，如何对客户进行细分；哪些客户的维护业务对公司更有价值；面对客户时如何说服他们维护业务带来的价值远远超过了成本。这一切更需要数据来支持！

不仅工作中需要数据思维，我们的生活也需要数据思维，它可以帮我们找工作、买房、从投资中获利，等等。那么什么是数据思维？数据思维最核心的思想是利用数据解决问题，而利用数据解决问题则需要深度了解需求，了解真正要解决什么样的问题，解决问题背后的真实目的是什么。在解决问题的过程中，使用数据分析的方法，帮助我们 从庞杂的数据中提取有价值的信息，做出更好的决策，本书正是围绕此目标而编写。全书分上下两篇。上篇（第1章～第10章）介绍了什么是数据分析，如何利用Python中的Pandas库进行基本的数据操作、数据清洗、数据整理，如何对数据进行可视化。下篇（第11章～第21章）着重于数据分析实战，尽可能多地为读者展示各种数据分析应用，目的是让读者体会到数据分析的作用。其中既有数据分析在企业中的应用，如客户群组分析、客户细分、A/B测试等，又有数据分析在我们生活中的应用，如找工作、买房、投资等。通过这些案例，希望读者能够理解数据分析是没有边界的，只要能提出问题，就能找到它的用武之地。

随着可穿戴设备的兴起，物联网、人工智能伴随5G而来，用数据定义问题、用数据讲故事、用数据支持决策的能力也越来越重要，本书是作者对过去学习与工作的总结，希望能成为读者迈入数据世界的第一本书。

本书第7章、第14章和第15章由邹静完成，其余章节由江雪松完成。由于作者水平有限，书中难免有不足之处，还望读者不吝指正！

感谢清华大学出版社的盛东亮老师为出版本书提供的帮助！同时也要感谢本书编辑团队付出的辛勤劳动！将近一年时间的写作，意味着无法陪伴家人，在此也要感谢他们的理解与支持。

江雪松

2020年4月

说明

1. 写作风格

本书上篇为Pandas数据分析基础，因此作者对于代码的讲解相对详细。下篇为案例实战，经过上篇的学习，读者应该已经具备了自学Pandas中新函数的能力，因此在讲解时以案例为主，对部分代码不再逐行解读，希望读者能自己动手实践，加深理解。

2. 英文术语的翻译

本书中的英文术语在两种情况不进行翻译，采取保留英文原文的方式。第一种情况是专有术语，无中文翻译，如Pandas、Jupyter Notebook等。第二种情况是Pandas中的专有数据结构DataFrame和Series，采取不翻译的方式，另外对于DataFrame中的Index，如果可能在上下文产生歧义，我们会采取保留英文原文的方式。

3. 代码格式

本书中的代码均在Jupyter Notebook中编写完成，Jupyter Notebook的一大优势是提供了良好的用户交互输出。为了更好地区分代码与输出，全书中的代码采用如下格式。

□

其中，前面带有“□”的代表Jupyter Notebook中的一段代码。例如，上面代码段中的“□.retail_data.dtypes”代表的是一行代码，前面没有“□□□”的部分则代表了代码对应的输出。类似地，如下内容中“□”后面代表的是一段代码。

□

4. 全书代码与数据下载

由于本书是一本数据分析实战的书，因此书中大量使用了各种不同真实场景的数据和代码，为了方便读者学习，我们也提供了数据和代码下载。读者可以关注微信公众号“见数知理”获得代码与数据的下载方式。

读者完成代码与数据下载后，如果按照本书第2章的方式设置了Jupyter Notebook，那么请将数据放到data目录，该目录应该与代码目录同级。

5. 勘误表

全书的勘误将通过公众号发布，读者阅读本书若发现任何问题，也欢迎通过公众号与我们交流。

学习说明：请关注“人工智能科学与技术”公众号，了解本书及后续可能更新的学习资源。

目录

[前言](#)

[说明](#)

[上篇 Pandas数据分析基础](#)

[第1章 数据分析初探](#)

[1.1 “数据+”时代的到来](#)

[1.2 什么是数据分析](#)

[1.2.1 数据分析的目标](#)

[1.2.2 数据分析分类](#)

[1.2.3 典型的数据分析方法](#)

[1.3 数据分析的基本流程](#)

[1.3.1 问题定义](#)

[1.3.2 收集数据](#)

[1.3.3 数据处理](#)

[1.3.4 数据分析](#)

[1.3.5 结果解读与应用](#)

[1.4 硝烟中的数据分析](#)

[1.4.1 数据分析的产生](#)

[1.4.2 验证问题](#)

[1.4.3 寻找原因](#)

[1.4.4 数据怎么说](#)

[1.4.5 数据分析中应该避免的典型问题](#)

[第2章 搭建数据科学开发环境](#)

[2.1 为什么选择Python](#)

[2.1.1 人生苦短，我用Python](#)

[2.1.2 为何Python是数据科学家的最佳选择](#)

[2.2 Python数据科学开发栈](#)

[2.2.1 Cython](#)

[2.2.2 NumPy](#)

[2.2.3 IPython](#)

[2.2.4 Jupyter](#)

[2.2.5 SciPy](#)

[2.2.6 Matplotlib](#)

[2.2.7 Pandas](#)

[2.2.8 Scikit-learn](#)

[2.2.9 NetworkX](#)

[2.2.10 PyMC3](#)

[2.2.11 数据科学领域中最新的一些Python包](#)

[2.3 Anaconda的安装与使用](#)

[2.3.1 安装Anaconda](#)

[2.3.2 利用Conda管理Python环境](#)

[2.3.3 利用Conda管理Python包](#)

[2.3.4 安装本书所需的包](#)

[2.4 使用Jupyter Notebook进行可重复数据分析](#)

[2.4.1 Jupyter Notebook的配置](#)

[2.4.2 Jupyter Notebook中的单元格](#)

[2.4.3 Jupyter Notebook中的命令模式与编辑模式键](#)

[2.4.4 使用Jupyter Notebook进行数据分析](#)

[第3章 Pandas基础](#)

[3.1 什么是DataFrame](#)

[3.1.1 DataFrame的基本要素](#)

[3.1.2 数据类型](#)

[3.1.3 了解Series](#)

[3.1.4 链式方法](#)

[3.2 索引与列](#)

[3.2.1 修改索引与列](#)

[3.2.2 添加、修改或删除列](#)

[3.3 选择多列](#)

[第4章 数据筛选](#)

[4.1 使用.loc和.iloc筛选行与列数据](#)

[4.1.1 选择Series和DataFrame中的行](#)

[4.1.2 同时选择行与列](#)

[4.2 布尔选择](#)

[4.2.1 计算布尔值](#)

[4.2.2 多条件筛选数据](#)

[第5章 开始利用Pandas进行数据分析](#)

[5.1 了解元数据](#)

[5.2 数据类型转换](#)

[5.3 缺失数据与异常数据处理](#)

[5.3.1 缺失值与重复值](#)

[5.3.2 处理缺失数据](#)

[5.3.3 NumPy与Pandas对缺失数据的不同处理方式](#)

[5.3.4 填充缺失值](#)

[5.4 处理重复数据](#)

[5.5 异常值](#)

[5.6 描述性统计](#)

[第6章 数据整理](#)

[6.1 什么是数据整理](#)

[6.1.1 数据的语义](#)

[6.1.2 整齐的数据](#)

[6.2 数据整理实战](#)

[6.2.1 列标题是值，而非变量名](#)

[6.2.2 多个变量存储在一列中](#)

[6.2.3 变量既在列中存储，又在行中存储](#)

[6.2.4 多个观测单元存储在单一表中](#)

[6.2.5 一个观测单元存储在多个表中](#)

[6.2.6 思考](#)

[第7章 分组统计](#)

[7.1 分组、应用和聚合](#)

[7.2 Pandas中的GroupBy操作](#)

[7.2.1 单列数据分组统计](#)

[7.2.2 多列数据分组统计](#)

[7.2.3 使用自定义函数进行分组统计](#)

[7.2.4 数据过滤与变换](#)

[第8章 数据整合](#)

[8.1 数据读入](#)

[8.1.1 基本数据读入方法](#)

[8.1.2 文件读取进阶](#)

[8.1.3 读取其他格式文件](#)

[8.2 数据合并](#)

[8.2.1 认识merge操作](#)

[8.2.2 merge进阶](#)

[8.2.3 join与concat](#)

[第9章 数据可视化](#)

[9.1 Matplotlib](#)

[9.1.1 绘制第一个散点图](#)

[9.1.2 理解figure与axes](#)

[9.1.3 Matplotlib中面向对象与类Matlab语法的区别](#)

[9.1.4 修改坐标轴属性](#)

[9.1.5 修改图形属性](#)

[9.1.6 定制图例，添加标注](#)

[9.1.7 子图](#)

[9.1.8 利用Matplotlib绘制各种图形](#)

[9.2 Pandas绘图](#)

[9.2.1 Pandas基础绘图](#)

[9.2.2 整合Pandas绘图与Matplotlib绘图](#)

[9.3 Seaborn](#)

[9.3.1 Seaborn中的样式](#)

[9.3.2 Seaborn绘制统计图形](#)

[9.4 可视化进阶](#)

[9.4.1 其他可视化工具](#)

[9.4.2 推荐读物](#)

[第10章 探索性数据分析——某电商销售数据分析](#)

[10.1 数据清洗](#)

[10.1.1 分析准备](#)

[10.1.2 了解数据](#)

[10.2 数据清洗与整理](#)

[10.2.1 数据类型转换与错误数据删除](#)

[10.2.2 添加新数据](#)

[10.3 探索性数据分析](#)

[10.3.1 客户分析](#)

[10.3.2 订单趋势分析](#)

[10.3.3 客户国家分析](#)

[10.3.4 留给读者的问题](#)

[下篇 Python数据分析实战](#)

[第11章 群组分析](#)

[11.1 群组分析概述](#)

[11.1.1 从AARRR到RARRA的转变](#)

[11.1.2 什么是群组分析](#)

[11.2 群组分析实战](#)

[11.2.1 定义群组以及周期](#)

[11.2.2 群组分析具体过程](#)

[11.2.3 思考](#)

[第12章 利用RFM分析对用户进行分类](#)

[12.1 RFM分析简介](#)

[12.1.1 RFM模型概述](#)

[12.1.2 理解RFM](#)

[12.2 RFM实战](#)

[12.2.1 R、F、M值的计算](#)

[12.2.2 利用RFM模型对客户进行细分](#)

[12.2.3 思考](#)

[第13章 购物篮分析](#)

[13.1 购物篮分析概述](#)

[13.1.1 什么是购物篮分析](#)

[13.1.2 购物篮分析在超市中的应用](#)

[13.1.3 购物篮分析实现](#)

[13.2 购物篮分析案例](#)

[13.2.1 Mxtend库中Apriori算法使用介绍](#)

[13.2.2 在线销售数据购物篮分析](#)

[13.3 留给读者的思考](#)

[第14章 概率分布](#)

[14.1 随机数](#)

[14.2 常见的概率分布](#)

[14.2.1 均匀分布](#)

[14.2.2 正态分布](#)

[14.2.3 二项分布](#)

[14.2.4 泊松分布](#)

[14.2.5 几何分布与指数分布](#)

[14.3 点估计与置信区间](#)

[14.3.1 点估计](#)

[14.3.2 抽样分布与中心极限定理](#)

[14.3.3 置信区间](#)

[14.4 留给读者的思考](#)

[第15章 假设检验](#)

[15.1 假设检验概述](#)

[15.1.1 初识假设检验](#)

[15.1.2 假设检验的步骤](#)

[15.1.3 假设检验中的I类错误与II类错误](#)

[15.2 Python中的假设检验](#)

[15.2.1 单样本 \$t\$ -test](#)

[15.2.2 双样本 \$t\$ -test](#)

[15.2.3 配对 \$t\$ -test](#)

[15.2.4 卡方检验](#)

[15.3 留给读者的思考](#)

[第16章 一名数据分析师的游戏上线之旅](#)

[16.1 游戏启动时间是否超过目标](#)

[16.1.1 启动时间是否超过3秒](#)

[16.1.2 构造启动时间监测图](#)

欢迎访问：电子书学习和下载网站 (<https://www.shgis.cn>)

文档名称：《Python数据分析与算法指南》（套装共8册）.epub

请登录 <https://shgis.cn/post/1845.html> 下载完整文档。

手机端请扫码查看：

