

水木书荟

# Python 数据分析与实战

吕云翔 李伊琳 王肇一 张雅素 编著

清华大学出版社  
北 京

## 内 容 简 介

使用 Python 进行数据分析是十分便利且高效的,因此它被认为是最优秀的数据分析工具之一。本书从理论和实战两个角度对 Python 数据分析工具进行了介绍,并采用理论分析和 Python 实践相结合的形式,按照数据分析的基本步骤对数据分析的理论知识以及相应的 Python 库进行了详细的介绍,让读者在了解数据分析的基本理论知识的同时能够快速上手实现数据分析程序。

本书适用于对数据分析有浓厚兴趣但不知从何下手的初学者,在阅读数据分析的基础理论知识的同时可以通过 Python 实现简单的数据分析程序,从而快速对数据分析的理论和实现两个层次形成一定的认知。

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

责任编辑:魏江江 王冰飞

封面设计:

责任校对:徐俊伟

责任印制:

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](mailto:c-service@tup.tsinghua.edu.cn)

质量反馈:010-62772015, [zhiliang@tup.tsinghua.edu.cn](mailto:zhiliang@tup.tsinghua.edu.cn)

课件下载: <http://www.tup.com.cn>, 010-62795954

印 刷 者:

装 订 者:

经 销:全国新华书店

开 本:185mm×260mm 印 张:12.25

字 数:211 千字

版 次:2019 年 1 月第 1 版

印 次:2019 年 1 月第 1 次印刷

定 价: .00 元

---

产品编号:077077-01

# 前言

本书是面向初学者的数据分析入门指南。按照数据分析的数据预处理、分析与知识发现和可视化 3 个主要步骤,本书逐步对数据分析涉及的理论进行讲解,并对实现这些步骤所用到的 Python 库进行详细介绍。通过理论与实践穿插的讲解方式,本书使读者能够在了解数据分析基础知识的同时快速上手实现一些简单的分析。

全书分为 10 章,第 1、3、6 章介绍数据分析理论,按照数据分析的基本流程介绍了理论知识和一些常用方法,穿插在理论章节之间的 Python 实战章节可以让读者在了解理论之后用相应的 Python 库来进行实战操作。通过阅读第 1~8 章的内容,读者已经对数据分析的各主要流程形成了一定的认识,但这些知识可能还未形成一个完整的体系,因此本书在第 9 和第 10 章引入了两个完整的数据分析实例,帮助读者建立知识点之间的联系,形成对数据分析整个知识面的清晰认知。建议读者在阅读实战章节时跟随介绍自己动手尝试一下,这样一定会发现数据的魅力所在。

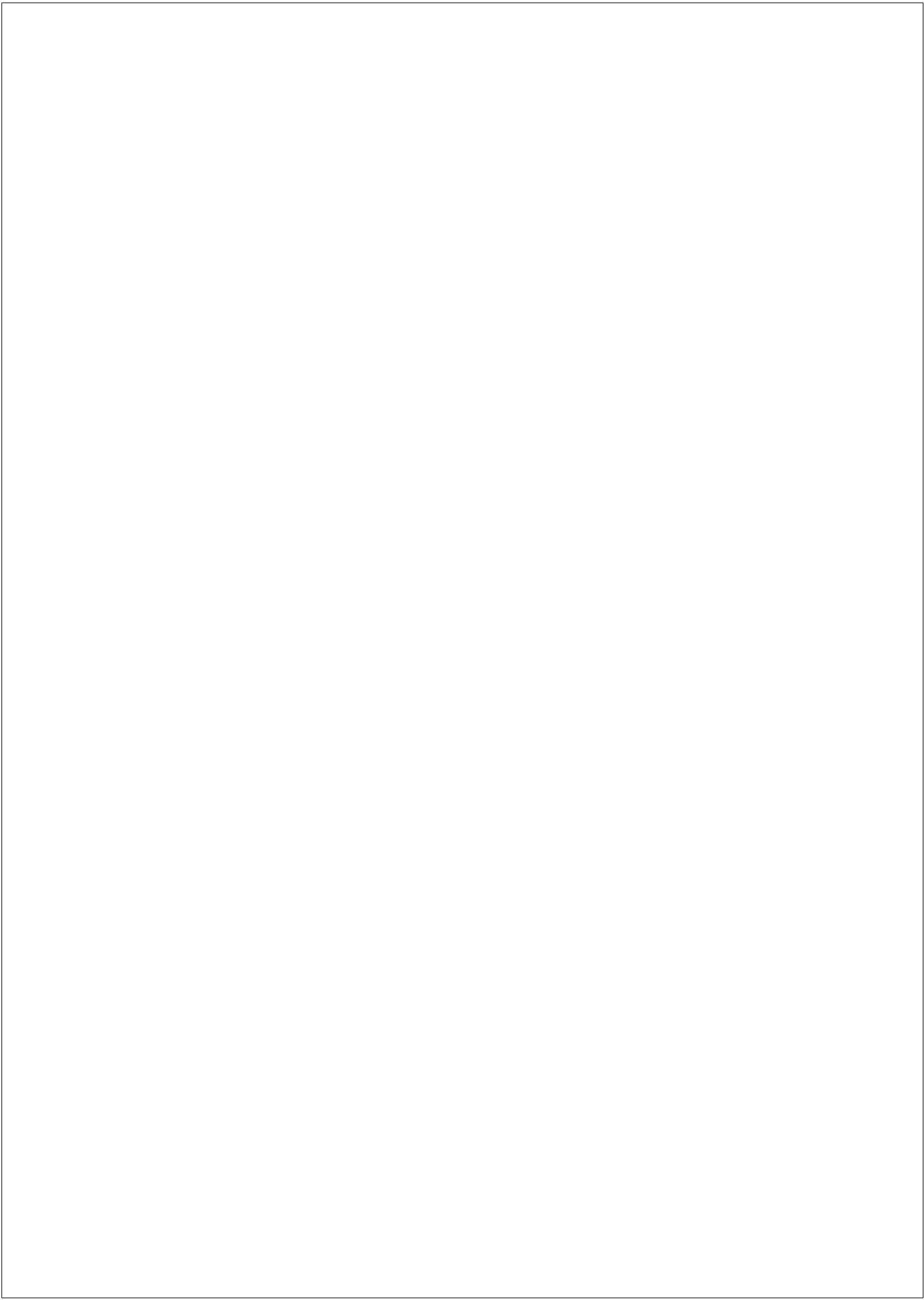
作为一本数据分析入门书籍,本书着重介绍基础知识,对前沿的内容涉及较少,这些内容留待读者在更进一步的学习中深入探索。对于 Python 语言的知识,本书仅对与数据分析相关的库进行了介绍,如果读者对 Python 语言本身有兴趣,可以参考 Python 语言工具书及官方文档等详细了解 Python 的语法和底层原理等。另外,本书所有数据分析程序的实现均在单机情况下进行,并没有对如何使用 Python 进行分布式数据分析作介绍,有兴趣的读者可以了解一下 Python 分布式数据分析的相关库,如 pyspark 等。

本书主要由吕云翔、李伊琳、王肇一、张雅素编写,曾洪立、吕彼佳、姜彦华也参与了部分内容的编写并进行了素材整理及配套资源制作等。

由于作者的水平和能力有限,本书难免有疏漏之处,恳请各位同仁和广大读者给予批评指正,也希望各位能将实践过程中的经验和心得与我们交流(yunxianglu@hotmail.com)。

作者

2018 年 7 月



# 目 录

|   |    |
|---|----|
| 第 1 章 数据分析是什么 .....                                   | 1  |
| 1.1 海量数据背后蕴藏的知识 .....                                 | 1  |
| 1.2 数据分析与数据挖掘的关系 .....                                | 2  |
| 1.3 机器学习与数据分析的关系 .....                                | 2  |
| 1.4 数据分析的基本步骤 .....                                   | 2  |
| 1.5 Python 和数据分析 .....                                | 3  |
| 第 2 章 Python——从了解 Python 开始 .....                     | 5  |
| 2.1 Python 的发展史 .....                                 | 5  |
| 2.2 Python 及 Pandas、scikit-learn、Matplotlib 的安装 ..... | 6  |
| 2.2.1 Windows 环境下 Python 的安装 .....                    | 6  |
| 2.2.2 Mac 环境下 Python 的安装 .....                        | 6  |
| 2.2.3 Pandas、scikit-learn 和 Matplotlib 的安装 .....      | 7  |
| 2.2.4 使用科学计算发行版 Python 进行快速安装 .....                   | 7  |
| 2.3 Python 基础知识 .....                                 | 8  |
| 2.3.1 缩进很重要 .....                                     | 9  |
| 2.3.2 模块化的系统 .....                                    | 9  |
| 2.3.3 注释 .....  | 10 |
| 2.3.4 语法 .....  | 10 |
| 2.4 重要的 Python 库 .....                                | 11 |
| 2.4.1 Pandas .....                                    | 11 |
| 2.4.2 scikit-learn .....                              | 11 |
| 2.4.3 Matplotlib .....                                | 11 |
| 2.4.4 其他 .....  | 11 |
| 2.5 Jupyter .....                                     | 12 |



|                                 |    |
|---------------------------------|----|
| <b>第 3 章 数据预处理——不了解数据一切都是空谈</b> | 14 |
| 3.1 了解数据                        | 15 |
| 3.2 数据质量                        | 17 |
| 3.2.1 完整性                       | 18 |
| 3.2.2 一致性                       | 18 |
| 3.2.3 准确性                       | 19 |
| 3.2.4 及时性                       | 20 |
| 3.3 数据清洗                        | 20 |
| 3.4 特征工程                        | 22 |
| 3.4.1 特征选择                      | 22 |
| 3.4.2 特征构建                      | 23 |
| 3.4.3 特征提取                      | 23 |
| <b>第 4 章 NumPy——数据分析基础工具</b>    | 25 |
| 4.1 多维数组对象 ndarray              | 26 |
| 4.1.1 ndarray 的创建               | 26 |
| 4.1.2 ndarray 的数据类型             | 29 |
| 4.2 ndarray 的索引、切片和迭代           | 29 |
| 4.3 ndarray 的 shape 的操作         | 32 |
| 4.4 ndarray 的基础操作               | 32 |
| <b>第 5 章 Pandas——处理结构化数据</b>    | 35 |
| 5.1 基本数据结构                      | 36 |
| 5.1.1 Series                    | 36 |
| 5.1.2 DataFrame                 | 38 |
| 5.2 基于 Pandas 的 Index 对象的访问操作   | 45 |
| 5.2.1 Pandas 的 Index 对象         | 45 |
| 5.2.2 索引的不同访问方式                 | 48 |
| 5.3 数学统计和计算工具                   | 52 |
| 5.3.1 统计函数：协方差、相关系数、排序          | 52 |
| 5.3.2 窗口函数                      | 54 |
| 5.4 数学聚合和分组运算                   | 60 |

|              |                                    |           |
|--------------|------------------------------------|-----------|
| 5.4.1        | agg()函数的聚合操作 .....                 | 63        |
| 5.4.2        | transform()函数的转换操作 .....           | 64        |
| 5.4.3        | 使用 apply()函数实现一般的操作 .....          | 65        |
| <b>第 6 章</b> | <b>数据分析与知识发现——一些常用的方法 .....</b>    | <b>67</b> |
| 6.1          | 分类分析 .....                         | 67        |
| 6.1.1        | 逻辑回归 .....                         | 68        |
| 6.1.2        | 线性判别分析 .....                       | 68        |
| 6.1.3        | 支持向量机 .....                        | 69        |
| 6.1.4        | 决策树 .....                          | 70        |
| 6.1.5        | K 近邻 .....                         | 71        |
| 6.1.6        | 朴素贝叶斯 .....                        | 72        |
| 6.2          | 关联分析 .....                         | 72        |
| 6.2.1        | 基本概念 .....                         | 72        |
| 6.2.2        | 典型算法 .....                         | 74        |
| 6.3          | 聚类分析 .....                         | 80        |
| 6.3.1        | K 均值算法 .....                       | 80        |
| 6.3.2        | DBSCAN .....                       | 81        |
| 6.4          | 回归分析 .....                         | 82        |
| 6.4.1        | 线性回归分析 .....                       | 83        |
| 6.4.2        | 支持向量回归 .....                       | 84        |
| 6.4.3        | K 近邻回归 .....                       | 84        |
| <b>第 7 章</b> | <b>scikit-learn——实现数据的分析 .....</b> | <b>85</b> |
| 7.1          | 分类方法 .....                         | 85        |
| 7.1.1        | Logistic 回归 .....                  | 85        |
| 7.1.2        | SVM .....                          | 87        |
| 7.1.3        | Nearest neighbors .....            | 88        |
| 7.1.4        | Decision Tree .....                | 89        |
| 7.1.5        | 随机梯度下降 .....                       | 90        |
| 7.1.6        | 高斯过程分类 .....                       | 91        |
| 7.1.7        | 神经网络分类(多层感知器) .....                | 91        |



|       |                               |     |
|-------|-------------------------------|-----|
| 7.1.8 | 朴素贝叶斯示例 .....                 | 92  |
| 7.2   | 回归方法 .....                    | 93  |
| 7.2.1 | 最小二乘法 .....                   | 93  |
| 7.2.2 | 岭回归 .....                     | 94  |
| 7.2.3 | Lasso .....                   | 94  |
| 7.2.4 | 贝叶斯岭回归 .....                  | 95  |
| 7.2.5 | 决策树回归 .....                   | 96  |
| 7.2.6 | 高斯过程回归 .....                  | 96  |
| 7.2.7 | 最近邻回归 .....                   | 97  |
| 7.3   | 聚类方法 .....                    | 98  |
| 7.3.1 | K-means .....                 | 98  |
| 7.3.2 | Affinity propagation .....    | 100 |
| 7.3.3 | Mean-shift .....              | 101 |
| 7.3.4 | Spectral clustering .....     | 101 |
| 7.3.5 | Hierarchical clustering ..... | 102 |
| 7.3.6 | DBSCAN .....                  | 103 |
| 7.3.7 | Birch .....                   | 104 |
| 第8章   | Matplotlib——交互式图表绘制 .....     | 106 |
| 8.1   | 基本布局对象 .....                  | 106 |
| 8.2   | 图表样式的修改以及装饰项接口 .....          | 111 |
| 8.3   | 基础图表的绘制 .....                 | 116 |
| 8.3.1 | 直方图 .....                     | 116 |
| 8.3.2 | 散点图 .....                     | 118 |
| 8.3.3 | 饼图 .....                      | 119 |
| 8.3.4 | 柱状图 .....                     | 120 |
| 8.3.5 | 折线图 .....                     | 125 |
| 8.3.6 | 表格 .....                      | 126 |
| 8.3.7 | 不同坐标系下的图像 .....               | 127 |
| 8.4   | matplotlib3D .....            | 128 |
| 8.5   | Matplotlib 与 Jupyter 结合 ..... | 130 |



|                           |     |
|---------------------------|-----|
| 第 9 章 实例：科比职业生涯进球分析 ..... | 134 |
| 9.1 预处理 .....             | 134 |
| 9.2 分析科比的命中率 .....        | 138 |
| 9.3 分析科比的投篮习惯 .....       | 155 |
| 第 10 章 实例：世界杯 .....       | 162 |
| 10.1 数据说明 .....           | 162 |
| 10.2 世界杯观众 .....          | 164 |
| 10.3 世界杯冠军 .....          | 170 |
| 10.4 世界杯参赛队伍与比赛 .....     | 173 |
| 10.5 世界杯进球 .....          | 180 |
| 参考文献 .....                | 185 |