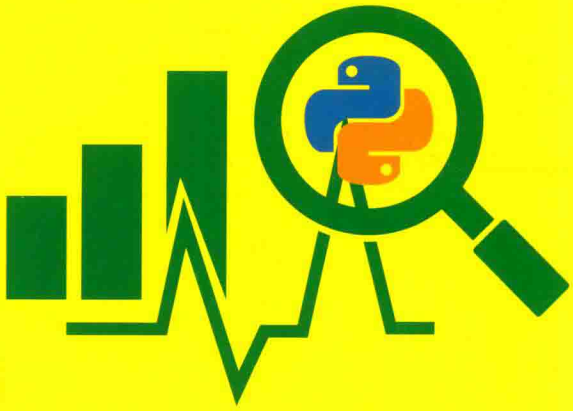
**手把手带领“小白”轻松入门Python数据分析**

全面涵盖数据分析的流程、工具、框架和方法，内容新，实战案例多

详细介绍从数据读取到数据清洗，以及从数据处理到数据可视化等实用技术



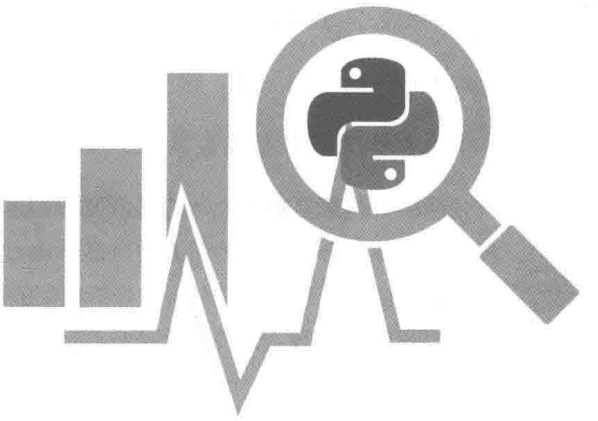
# 从零开始学 Python数据分析

（视频教学版）

罗攀◎编著

- **赠送400分钟配套教学视频、补充学习文档、素材文件、源文件和PPT等超值资源**
- 详解9个有较高应用价值的Python数据分析项目实战案例
- 详解数据分析的3大模块：NumPy、pandas库和matplotlib库
- 详解Python数据分析集成环境Anaconda的安装和使用，降低学习门槛

机械工业出版社
China Machine Press



# 从零开始学 Python数据分析

（视频教学版）

罗攀◎编著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

从零开始学Python数据分析：视频教学版/罗攀编著. —北京：机械工业出版社，2018.7

ISBN 978-7-111-60646-8

Ⅰ. 从… Ⅱ. 罗… Ⅲ. 软件工具－程序设计 Ⅳ. TP311.561

中国版本图书馆CIP数据核字（2018）第182809号

网络中的信息是很庞大的。如何提取这些信息？如何分析这些信息？这都需要用到数据分析技术。而数据分析技术的首选语言是Python，而本书便是一本适合“小白”学习Python数据分析的入门图书，书中不仅有各种分析框架的使用技巧，而且也有各类数据图表的绘制方法。本书通过讲解多个案例，让读者体验数据背后的乐趣。

本书共11章，核心内容包括Python数据分析环境安装、NumPy基础、pandas基础、外部数据读取与存储、数据清洗与整理、数据分组与聚合、matplotlib可视化、seaborn可视化、pyecharts可视化、时间序列、网站日志分析综合案例等。

本书适合Python数据分析的初学者和爱好者阅读，也适合作为各类院校相关专业的教学用书，同时还适合相关社会培训机构作为Python数据分析的培训教材或者参考书。

从零开始学 Python 数据分析（视频教学版）

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）

责任编辑：欧振旭 李华君　　责任校对：姚志娟

印　　刷：中国电影出版社印刷厂　　版　　次：2018 年 8 月第 1 版第 1 次印刷

开　　本：186mm×240mm 1/16　　印　　张：17

书　　号：ISBN 978-7-111-60646-8　　定　　价：69.00 元

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　　投稿热线：（010）88379604

购书热线：（010）68326294 88379649 68995259　　读者信箱：hzit@hzbook.com

# 前言

互联网的飞速发展伴随着海量信息的产生，而海量信息的背后对应的则是海量数据。如何从这些海量数据中获取有价值的信息来供人们学习和工作使用，这就不得不用到大数据挖掘和分析技术。数据分析作为大数据技术的核心一环，其重要性不言而喻。

在数据分析领域，Python 语言以其简单易用，并提供了优秀、好用的第三方库和数据分析的完整框架而深受数据分析人员的青睐。可以说，Python 已经当仁不让地成为了数据分析人员的一把“利器”。程序员想要进入数据分析行业，首先要掌握 Python 数据分析技术，只有这样才能在严峻的就业市场中具有较强的竞争力。

目前图书市场上关于 Python 数据分析的图书主要是几本翻译图书，其定位相对高端，而且翻译质量参差不齐，案例数据不方便下载，阅读难度系数较大，初学者不容易上手，故不适合初学者学习；而国内的几本原创 Python 数据分析图书质量也良莠不齐，不成系统，也不适合初学者阅读。可以说，图书市场上还鲜见一本通俗易懂且适合“小白”阅读的 Python 数据分析入门图书，基于此，笔者编写了本书。本书从 Python 数据分析的基础知识入手讲解，然后结合大量的数据分析案例，系统地介绍了 Python 数据分析的方法和流程，手把手带领读者掌握 Python 数据分析的相关知识，并提高读者的项目实践能力。

## 本书特色

### 1. 视频教学，高效、直观

为了便于读者高效、直观地学习，笔者专门为本书的重点内容录制了配套教学视频，读者可以一边看书，一边结合教学视频进行学习，以取得更好的学习效果。

### 2. 内容全面，讲解系统

本书不但全面介绍了从 Numpy 到 pandas，从 matplotlib 到 pyecharts 的数据分析必学技术，而且还系统地讲解了从数据读取到数据清洗，从数据处理到数据可视化的详细步骤。

### 3. 给出了数据分析环境的安装和配置步骤

本书详细介绍了 Python 数据分析集成环境 Anaconda 的安装步骤和使用方法，可以大大降低初学者学习 Python 数据分析的门槛，从而让读者快速跨进 Python 数据分析的大门。

### 4. 详细介绍了数据分析的流程

本书从一开始便对数据分析的流程进行了详细介绍，而且在讲解中结合了多个实用性很强的数据分析项目案例，带领读者掌握 Python 数据分析的相关知识，以解决实际工作中的数据分析问题。

### 5. 提供了9个有较高应用价值的项目案例，有很强的实用性

本书提供了 9 个实用性很强的数据分析项目案例，这些案例从不同的分析角度切入进行讲解，具有较高的应用价值。读者通过实际操练，可以更加透彻地理解数据分析的相关知识。

### 6. 提供教学PPT，方便教学和学习

笔者专门为本书制作了专业的教学 PPT，以方便相关院校的教学人员授课时使用；读者也可以通过教学 PPT，提纲挈领地掌握书中的内容脉络。

## 本书内容

**第 1 章　Python 环境搭建与使用**

本章介绍了如何搭建和使用 Python 数据分析环境，并介绍了如何使用 Jupyter Notebook 进行数据分析编程。

**第 2 章　NumPy 入门和实战**

本章首先介绍了 Numpy 的基本数据结构——多维数组；然后介绍了多维数组的创建和基本属性、数组的切片和索引方法，以及数组的运算与存取；最后通过综合案例，演示了如何实现图像的变换功能。

**第 3 章　pandas 入门和实战**

本章首先介绍了 pandas 中两种基础数据结构的创建和使用方法；然后详细讲解了 DataFrame 的选取和操作，同时介绍了其算术运算、函数的使用和 pandas 的可视化方法；最后结合案例，介绍了数据分析流程。

**第 4 章　外部数据的读取与存储**

本章主要介绍了如何利用 pandas 库读取外部数据为 DataFrame 数据格式，并介绍了通过 Python 进行数据处理后如何将 DataFrame 类数据存储到相应的外部数据文件中。

**第 5 章　数据清洗与整理**

本章主要介绍了如何使用 pandas 进行多源数据的清洗和整理，并给出了针对多源数据的合并和连接方法，以及数据的重塑方法，最后通过一个综合案例演示了数据分析中的数据清洗过程。

**第 6 章　数据分组与聚合**

本章涵盖的主要内容有：GroupBy 的原理和使用方法；聚合函数的使用；分组运算中 transform 和 apply 方法的使用；通过 pandas 创建数据透视表；通过综合案例，巩固数据分组统计的使用。

**第 7 章　matplotlib 可视化**

本章涵盖的主要内容有：利用 matplotlib 进行图表绘制；学会使用自定义设置，个性化绘制图表；通过综合案例，巩固 matplotlib 可视化的方法和技巧。

**第 8 章　seaborn 可视化**

本章涵盖的主要内容有：使用 seaborn 绘图；学会 seaborn 样式和分布图绘制；通过综合案例泰坦尼克号的生还者数据，巩固 seaborn 的可视化方法和技巧。

**第 9 章　pyecharts 可视化**

本章涵盖的主要内容有：安装 pyecharts 库；学会使用 pyecharts 库绘制基本图表；学会绘制其他图表；通过综合案例，巩固 pyecharts 的绘制方法和技巧。

**第 10 章　时间序列**

本章涵盖的主要内容有：时间序列的构造和使用方法；时间序列的频率转换与重采样；通过综合案例，巩固时间序列数据的处理与分析方法。

**第 11 章　综合案例——网站日志分析**

本章通过一个综合案例，介绍了如何通过 Python 的第三方库解析网站日志；如何利用 pandas 对网站日志数据进行预处理；结合前面介绍的数据分析和数据可视化技术对网站日志数据进行分析。

## 本书配套资源获取方式

本书提供以下配套资源：

- 本书配套教学视频；
- 超值电子书（地图绘制技术）；
- 本书相关素材文件；
- 本书源代码文件；
- 本书教学 PPT。

这些配套资源需要读者自行下载。请登录机械工业出版社华章公司网站 www.hzbook.com，在该网站上搜索到本书，然后单击“资料下载”按钮即可找到“配书资源”下载链接。

## 适合阅读本书的读者

- 数据分析初学者；
- 数据分析爱好者；
- 数据分析从业人员；
- 数据分析培训学员；
- 高校相关专业的学生。

本书由罗攀主笔编写，蒋仟、陈瑞滕和潘丹三位小伙伴也参与了部分章节的编写工作，在此对他们表示特别的感谢！

由于作者水平所限，加之写作时间有限，书中可能还存在一些疏漏和不足之处，敬请各位读者斧正。联系我们请发电子邮件到 hzbook2017@163.com。

罗攀

# 目录

# 第 1 章　Python 环境搭建与使用

Python 语言在数据读取、数据处理、数据可视化和数据挖掘等方面都有极其广泛的应用。本章将讲解如何搭建和使用 Python 数据科学环境，并学习使用 Jupyter Notebook 进行编程。

下面给出本章涉及的知识点与学习目标。

- Anaconda 安装和使用：学会数据科学环境的搭建和使用。
- Jupyter Notebook 使用：学会 Jupyter Notebook 的基本操作和 Python 程序的编写。

## 1.1　Anaconda 的安装和使用

“工欲善其事，必先利其器”。本节将介绍 Python 数据科学环境（Anaconda）的安装和使用方法。

### 1.1.1　Anaconda 的安装

Anaconda 是一个集成的 Python 数据科学环境。简单地说，Anaconda 除了有 Python 外，还安装了 180 多个用于数据分析的第三方库，而且可以使用 conda 命令安装第三方库和创建多个环境。相对于只安装 Python 而言，安装 Anaconda 避免了安装第三方库的麻烦。

注意：因为许多与数据分析相关的 Python 第三方库依赖性强，不容易安装，建议使用 Anaconda。

（1）打开浏览器，进入清华大学 Anaconda 开源镜像源（https://mirror.tuna.tsinghua.edu.cn/help/anaconda/），单击链接进行下载，如图 1.1 所示。

注意：由于网络限制，这里没有介绍如何在官网下载，而是选择通过镜像下载。

（2）通过下拉列表选择最新的 Anaconda 版本，读者可根据计算机操作系统进行下载，如图 1.2 所示。

注意：Anaconda 3 为 Python 3 的版本。本书以 Windows 64 位操作系统为例。


图 1.1　安装步骤 1


图 1.2　安装步骤 2

（3）双击打开下载好的程序，在欢迎界面单击 Next 按钮进入下一步，然后单击 I Agree 按钮同意进行安装，此时会要求读者选择使用的用户类型，读者可选择所有用户（All Users），然后单击 Next 按钮进入下一步，如图 1.3 所示。

（4）选择安装路径，建议安装到非系统盘，并安装到磁盘的根目录下，然后单击 Next 按钮进入下一步，如图 1.4 所示。

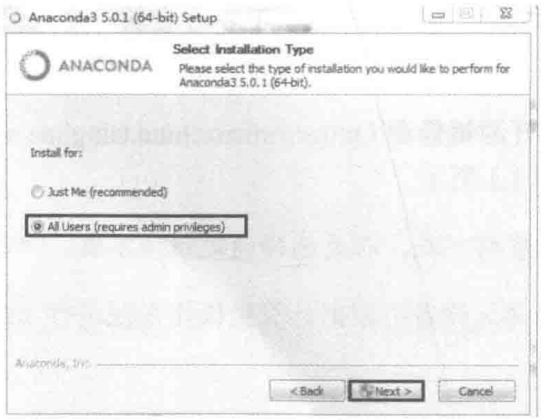



图 1.3　安装步骤 3

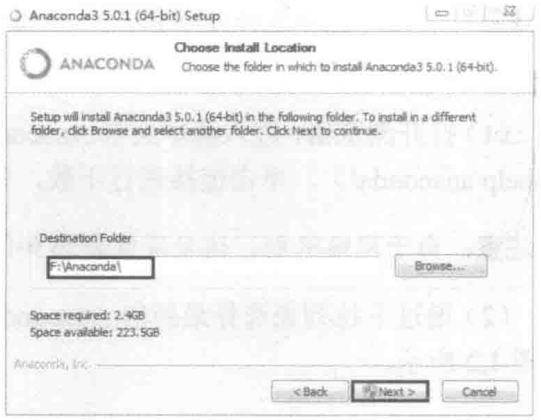



图 1.4　安装步骤 4

（5）在自定义选项中，勾选所有选项，单击 Install 按钮进行安装即可。安装完成后，单击 Next 按钮，再单击 Finish 按钮即可完成安装，如图 1.5 所示。

注意：第一个复选框选项是把 Anaconda 加入环境变量，勾选第二个复选框可以关联一些编辑器。

（6）安装完成后，在“开始”菜单栏中，选择 Anaconda Prompt 命令即可运行 Anaconda 环境，如图 1.6 所示。

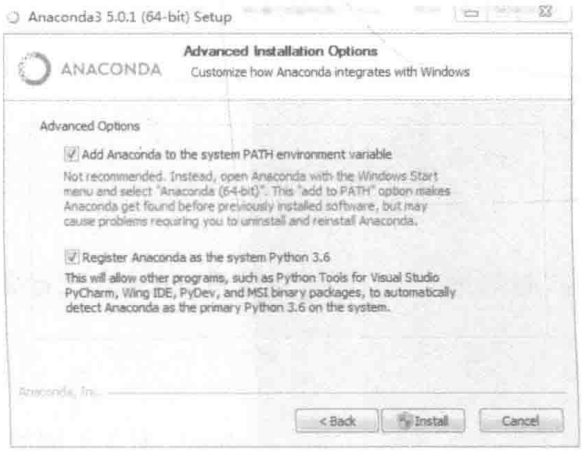



图 1.5　安装步骤 5


图 1.6　运行 Anaconda

## 1.1.2　Anaconda 的使用

Anaconda 其实是一个打包的集合，里面预装好了 conda、某个版本的 Python、众多包及科学计算工具等。对于 Anaconda 的使用，其实就是对 conda 的使用。conda 可以理解为一个工具（或者是可执行的命令），其核心功能为第三方库（包）和环境的管理。

### 1. 包管理

首先运行 Anaconda，查看 Python 版本，如图 1.7 所示。由于这里安装的是 Anaconda 3，对应的是 Python 3 版本。

通过 conda list 命令可以查看安装的包，部分结果如图 1.8 所示，可以看出，Anaconda 集成了大量的包。

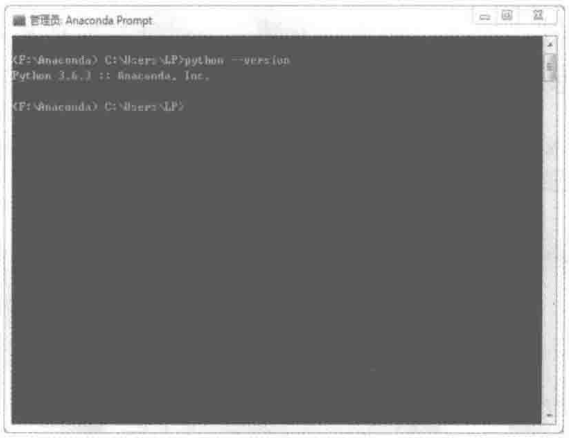



图 1.7　查看 Python 版本

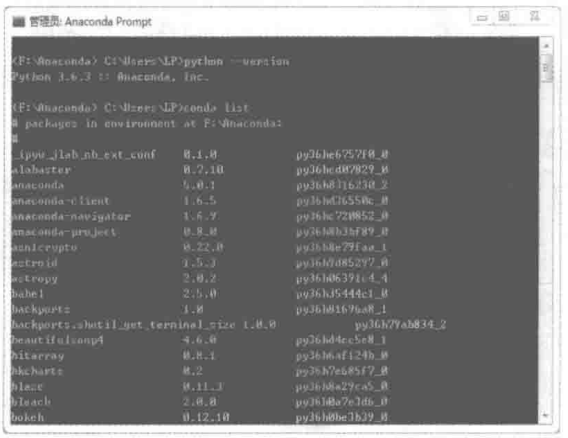



图 1.8　查看包

通过 conda 命令可以进行包的安装和卸载，也可以通过 pip 命令进行包的安装和下载。具体使用方法如下：

```
conda install xxx                    #xxx 为包名称
conda remove xxx                     #卸载包
pip install xxx
pip uninstall xxx
```

注意：建议使用 conda 命令安装包，在出错的情况下，再考虑使用 pip 命令进行安装。

### 2. 环境管理

conda 的环境管理功能允许开发者同时安装若干不同版本的 Python。对于不同的项目而言，使用独立稳定的 Python 环境很重要，以下就是安装 Python 环境的 conda 命令。

```
conda create --name xxx python=2     #xxx 为环境名称，创建了 python 版本为 2 的环境
conda create --name xxx python=3     #创建了 python 版本为 3 的环境
conda create --name xxx python=3 anaconda
                          #创建了 python 版本为 3 的环境，并具有 Anaconda 的所有包
```

这里创建一个名为 data-analysis 的 Python 环境，其是用于本书讲解 Python 数据分析的环境。由于数据分析需要 Anaconda 的原生包，这里需在 conda 命令末尾加上 anaconda，如图 1.9 所示。

创建环境成功后，可通过 activate data-analysis 命令进入该环境中，通过 deactivatedata-analysis 命令退出该环境，如图 1.10 所示。

```
activate xxx                         #激活环境
deactivate xxx                       #退出环境
```

通过以下命令可以查看 Anaconda 的环境，方便用户进行 Python 环境的管理和使用，如图 1.11 所示。

```
conda info -envs
```