Python
数据分析基础

阮敬 编著

- 实用性强，强调对统计分析和数据分析知识的领悟和应用
- 语言通俗，避免用复杂数学公式推导分析所用的基本原理
- 通过python来实现数据分析和得到结论的全部过程
- 配有配套案例数据，下载地址：www.zgtjcbs.com

中国统计出版社
China Statistics Press

Python
数据分析基础

阮敬 编著

- 实用性强，强调对统计分析和数据分析知识的领悟和应用
- 语言通俗，避免用复杂数学公式推导分析所用的基本原理
- 通过python来实现数据分析和得到结论的全部过程
- 配有配套案例数据，下载地址：www.zgtjcbs.com

中国统计出版社
China Statistics Press

内容简介

本书通过真实案例，全面介绍 python 编程基础和数据分析工具的应用，并培养读者通过数据分析问题、解决问题以及对结果评价的能力。全书内容包括：python 基本配置和编程基础、数据预处理、数据描述与可视化、统计推断、相关分析、关联分析、回归分析、主成分和因子分析、聚类、判别与分类、列联分析、对应分析、定性数据分析、时间序列分析等，将读者关注的数据分析与数据挖掘技术进行剖析。

图书在版编目（CIP）数据

Python 数据分析基础 / 阮敬编著. -- 北京 : 中国统计出版社, 2017.9

ISBN 978-7-5037-8320-3

Ⅰ. ①P… Ⅱ. ①阮… 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(2017)第 217456 号

Python 数据分析基础

作　　者/阮　敬
责任编辑/姜　洋　王立群
封面设计/李雪燕
出版发行/中国统计出版社
通信地址/北京市丰台区西三环南路甲 6 号　邮政编码/100073
电　　话/邮购（010）63376909　书店（010）68783171
网　　址/http://www.zgtjcbs.com/
印　　刷/河北鑫兆源印刷有限公司
经　　销/新华书店
开　　本/787mm×1092mm　1/16
字　　数/510 千字
印　　张/27
版　　别/2017 年 9 月第 1 版
版　　次/2017 年 9 月第 1 次印刷
定　　价/89.00 元

前　言

数据分析是科学研究中的重要环节，随着大数据时代的迅猛发展，其越来越受社会和市场的重视，是科学研究、经营管理、预测与决策等过程中必不可少的基础工作。python 是当今大数据时代下最为流行的编程工具之一，在大数据领域有着十分广泛的应用，可以实现从数据收集和数据管理到数据分析和挖掘的完整过程，其高效的编程和程序执行过程，能够完全胜任日常数据分析工作的需求。

随着数据分析作用的日益凸显，如何对现有数据进行整理、加工、处理和分析，以期得到结论，作为人们进行决策的依据进而实现数据的价值？如何利用现有数据对将来可能出现的数据结果或结论进行判断或预测？不管是针对企事业单位的管理者或决策者还是从事具体数据分析的工作人员而言，都需要进行合理数据分析流程的规划，区分数据类型，利用适合的数据分析方法，使用方便、快捷、可靠的统计软件作为工具，对特定数据进行分析与预测，从而洞察市场动向，观测人心所在，把握商机，提升竞争力。

而具有深厚数学背景的统计分析和数据分析方法往往会成为相关人员继续深入学习的门槛，甚至成为枯燥乏味的代名词，无法体验到数据分析成果带来的成效。本书就是要力求降低学习难度，通过编者积累的大量真实案例和数据，主要以文字阐述替代复杂公式推导，深入浅出剖析数据分析方法的基本原理和步骤，重点在于厘清数据分析的基本思路，合理得到恰当的分析结

果。在分析过程中，本书基于python 2.7，从基础编程入手，主要通过调用python基本库和常用工具库的方式，用大量的实例来展示数据分析每一步骤的细节，带领读者走入数据分析的奇妙世界。

本书的第1章和第2章主要介绍python的基本环境、编程基础和数据预处理方面的内容，具体内容包括python数据类型及数据结构、语句与控制流、基本库、函数和面向对象编程的基础，以及数据分析最为常用的基本分析工具库numpy和pandas基础等；

第3章和第4章主要介绍利用python进行描述分析的基本过程和方法，涵盖了各种常用数据分析图形的绘制和解读以及统计量和统计表等具体内容；

第5、6、7章主要介绍利用python如何进行总体推断。在大数据时代即使数据量再大，但也离不开利用统计思想对总体特征进行推测和判断，这些具体内容包括参数估计、假设检验和非参数分析；

第8章主要介绍如何用python来分析数据之间的关系，具体涵盖了简单相关分析、非参数相关分析、偏相关分析、点二列相关分析以及数据挖掘中常用的关联分析等内容；

第9章和第10章主要介绍如何利用python来进行回归分析。回归模型可以说是大部分统计分析和数据挖掘方法的基础，本书介绍的具体内容有线性回归、非线性回归、多项式回归、分位数回归、自变量含有定性变量的回归以及因变量含有定性变量的广义线性回归分析；

第11章和第12章主要就日常数据分析中所使用的多元统计分析方法进行介绍，具体内容包括主成分分析、因子分析、列

联分析以及对应分析等；

第 13 章和第 14 章主要介绍在 python 中进行数据挖掘所使用的聚类和分类方法。内容涵盖系统聚类、*k*-means 聚类、DBSCAN 聚类、距离判别和线性判别、贝叶斯判别以及数据挖掘中的 *k*-近邻、决策树、支持向量机和随机森林等分类方法；

第 15 章主要介绍 python 中使用 ARIMA 建模进行时间序列分析的基本方法和思路。

本书以实用为主要目的，因此上述大部分的数据分析过程均会调用现有常用且公认的结果较为合理的工具库（如 numpy、pandas、matplotlib、scipy、statsmodels、scikit-learn 等）。对于本书提及的数据分析方法无法通过调用现成工具库实现的，本书在相应章节中使用 python 编制了相应的函数或类，以供读者在分析实际问题时调用和复用。读者在复用这些函数或类时，也可根据自身需要对它们进行进一步优化。

全书采用 macOS Sierra 操作系统下的 python 2.7.13 和 Anaconda 4.3.1 的 jupyter notebook 作为分析环境，希望读者参考本书的内容边做边学习。为了提高学习效果，读者应该自行把本书全部代码在 python 中一字一句的敲一遍并运行之，故本书不提供电子版程序代码。但为了提高学习效率，本书附送随书案例的全部数据（下载地址：www.zgtjcbs.com）。

本书由本人在原书《实用 SAS 统计分析教程》（中国统计出版社 2013 年版）基础上亲自编写完成。开源软件的显著特点大家都懂的。因此，读者可在阅读本书时对照原书进行实际操作，认真体会商业软件和开源软件分析流程和分析结果的异同。此外，我的研究生杨磊磊和王禹提供了部分分析程序并对全书所编

制的程序进行了运行验证。同时感谢中国统计出版社的支持。尽管作者已经投入了大量时间和精力来编写此书，但由于水平有限，如有不足之处，敬请专家与同行批评指正。同时也欢迎广大读者与作者积极联系，共同探讨数据分析方面的心得与体会。

Email: ruanjing@msn.com

读者交流群：

阮　敬

2017 年 8 月 23 日

目　录

第 1 章

Python 编程基础

Python 是一种面向对象的解释型高级编程语言，其结构简单、语法和代码定义清晰明确、易于学习和维护、可移植性和可扩展性非常强。python 提供了非常完善的基础代码库（内置库），覆盖了数据结构、语句、函数、类、网络、文件、GUI、数据库、文本处理等大量内容。用 python 进行数据分析和功能开发，许多功能直接使用现成的包（packages）或模块（modules）即可，极大的提升了效率。除了内置库外，python 还有大量的第三方库，如 numpy、scipy、matplotlib、pandas、statsmodels、sklearn 等主要用于数据分析的库，提供了向量和矩阵的操作、数据可视化、统计计算、统计推断、统计分析与建模、数据挖掘和机器学习等几乎全部数据分析的功能。

本章将就利用 python 进行统计分析和数据分析的基本功能进行详细介绍，内容包括 python 系统配置以及面向对象的编程方式，涵盖了数据结构、控制流、函数、类、模块与包等具体内容。

1.1　Python 系统配置

Python 可应用于多系统平台，如 Linux、macOS、Windows 等，用户可以直接在其官网：https://www.python.org，下载最新版本的 python。用户也可以下载其他发行版本的 python 进行使用，如 anoconda、enthought canopy 等，这些发行版本的 python 已经包含一些特定的分析库，用户可以直接在它们提供的环境中使用。此外，还可以在集成开发环境 IDE（integrated development environment）中配置和使用 python，如 PyCharm、Eclipse 以及 python 自带的 IDLE 等。本书主要介绍 python 数据分析的基础知识，对于这些系统配置问题本书不予赘述，请读者自行查阅相关资料。

本书使用 macOS 10.12.6 系统原生的 python2.7.13 和自行安装的发行版本 Anaconda4.3.1 所搭建的环境。

打开 Mac 系统的应用程序→实用工具→终端，输入命令：python，即可进入 python 的交互式环境，如图 1-1 所示：

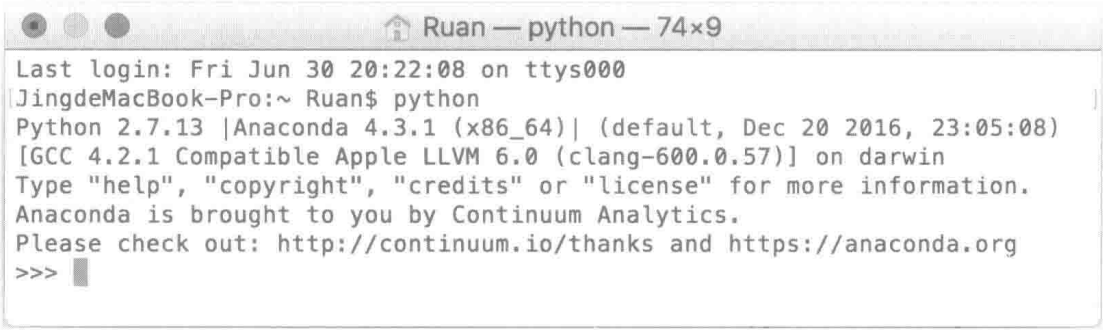

图 1-1 python 的交互式环境

用户可以在提示符“>>>”右边输入 python 命令或者语句。如：

```
>>> print 'Hello, world!'
    Hello, world!
```

在如需在 python 中安装第三方的工具库或包（packages），可以在终端中使用如下命令：

```
pip install package 的名称
```

在连接有网络的状态下，系统会自动下载对应名称的 package 并将其安装在当前系统环境当中。

macOS 原生版本 python 的交互式编程环境功能较为单一，而 Ipython 可以提供一个综合的交互式编程环境即 notebook，可以提供 tab 补全、富媒体、多客户端连接 kernel、交互式并行计算等，在科学计算和数据分析领域中发挥着极为强大的作用。在终端中可以使用 pip install 安装 ipython 和 notebook：

```
pip install ipython
pip install notebook
```

安装完毕后，在 python 环境中输入如下命令：

```
ipython notebook    #或者 jupyter notebook
#注：ipython notebook 现已升级为 jupyter notebook
```

python 中可以使用“#”作为注释，“#”右边的一切内容均不会执行。但是“#”只能对一行内容进行注释，如需对多行内容进行注释，可以在被注释的内容前后加上“'''”或“"""”，所有注释内容在程序运行过程中不会被执行：

```
ax1=fig.add_axes([0.1,0.6,0.2,0.3])
'''
这是注释内容：
指定子图在图像中的位置坐标，图像的左下角是原点(0,0)
图像横轴方向和纵轴方向总长度都为 1，(0.5, 0.5)是图像的中点
'''
line=ax1.plot([0,1],[0,1])    #绘制一条直线
ax1.set_title('Axes1')
```

运行上述的 ipython notebook 命令后，系统便会打开默认的浏览器进入 ipython notebook（现已更新为 jupyter notebook），如图 1-2 所示：