



达人迷

Python 数据科学入门

Drone ^{for} dummies[®]
A Wiley Brand

本书将教会你

- 掌握Python数据分析来编程
- 使用Python对象、函数、模块和库
- 应用概率和随机分布等统计学概念
- 使用NumPy、SciPy、Scikit-learn和Pandas库

[美] 约翰·保罗·穆勒 (John Paul Mueller)

[意] 卢卡·马萨罗 (Luca Massaron) 著

徐旭彬 译



中国工信出版集团



人民邮电出版社
POSTS & TELECOM PRESS



达人迷

Python 数据科学入门

Drone dummies[®]

A Wiley Brand

[美] 约翰·保罗·穆勒 (John Paul Mueller)

[意] 卢卡·马萨罗 (Luca Massaron) 著

徐旭彬 译

人民邮电出版社

北京

图书在版编目 (C I P) 数据

Python数据科学入门 / (美) 约翰·保罗·穆勒
(John Paul Mueller), (意) 卢卡·马萨罗
(Luca Massaron) 著; 徐旭彬 译. — 北京: 人民邮
电出版社, 2018.5
(达人迷)
ISBN 978-7-115-47962-4

I. ①P… II. ①约… ②卢… ③徐… 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(2018)第073392号

版 权 声 明

Luca Massaron, John Paul Mueller

Python for Data Science For Dummies

Copyright © 2015 by John Wiley & Sons, Inc.

All right reserved. This translation published under license.

Authorized translation from the English language edition published by John Wiley & Sons, Inc..

本书中文简体字版由 John Wiley & Sons 公司授权人民邮电出版社出版, 专有出版权属于人民邮电出版社。
版权所有, 侵权必究。

◆ 著 [美] 约翰·保罗·穆勒 (John Paul Mueller)
[意] 卢卡·马萨罗 (Luca Massaron)

译 徐旭彬

责任编辑 陈冀康

责任印制 焦志炜

◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路 11 号

邮编 100164 电子邮件 315@ptpress.com.cn

网址 <http://www.ptpress.com.cn>

三河市君旺印务有限公司印刷

◆ 开本: 800×1000 1/16

印张: 24.5

字数: 556 千字

2018 年 5 月第 1 版

印数: 1—3 000 册

2018 年 5 月河北第 1 次印刷

著作权合同登记号 图字: 01-2016-8616 号

定价: 69.00 元

读者服务热线: (010)81055410 印装质量热线: (010)81055316

反盗版热线: (010)81055315

广告经营许可证: 京东工商广登字 20170147 号

内容提要

本书的目标是介绍如何使用 Python 语言及其工具，解决和数据科学所关联的复杂任务。

全书共 6 个部分，分 22 章，涵盖了 Python 数据科学基础知识，数据的采集、整理、整形、应用，数据的可视化，数据分析和处理，数据学习，以及和数据科学相关的 10 个话题等。本书将重点放在使用正确的工具上，教读者如何使用 Anaconda、Matplotlib、NumPy、pandas、Scikit-learn 等常用的工具来解决数据科学的相关问题。

本书适合对数据科学的知识和应用方法感兴趣的读者阅读，特别适合有志于学习 Python 数据分析和处理的读者学习参考。

作者简介

Luca Massaron 是一名数据科学家和市场研究主管，他专攻多变量统计分析、机器学习和客户洞察，在解决真实世界的问题以及应用推理、统计学、数学挖掘和算法来为利益相关者创造价值方面具有超过 10 年的经验。他始终热衷于数据和分析相关的所有一切以及向专家和非专家演示由数据驱动的知识探索的潜力。相比不必要的复杂性，他更倾向于简单，他相信通过理解并实践不可或缺的关键部分就可以取得数据科学的很多成就。

John Mueller 是一位自由职业作家和技术编辑。他的写作天赋与生俱来，到现在为止已经创作了 97 本书和 600 多篇文章。主题范围从网络到人工智能，从数据库管理到编程实践。他作为技术编辑已经帮助了超过 63 位作者来改善他们的手稿。John 也为《Data Based Advisor》和《Coast Compute》杂志提供技术编辑服务。在 John 为《Data Based Advisor》提供服务期间，他第一次接触到 MATLAB，从此以后他继续跟踪 MATLAB 的发展。John 在 Cubic 公司期间，他接触到了可靠性工程，并保持了在概率方面的兴趣。一定要在 <http://blog.johnmuellerbooks.com/> 读一读 John 的博客。

当 John 不在计算机旁工作时，你可以发现他在花园中、伐木，或者说就是欣赏大自然。John 也喜欢酿造葡萄酒、烘焙饼干和编织，以及制作甘油肥皂和蜡烛。你可以通过 John@JohnMuellerBook.com 联系他，其个人网站为 <http://www.johnmuellerbooks.com>。

Luca的献辞

我想要把本书献给我的父母，Renzo 和 Licia，他们都喜欢简单而且便于解释的想法，以及那些阅读我们所写的书的人。现在他们更多地理解了我在数据科学中的日常工作，并且理解了这个新领域是将会如何改变我们理解世界的方式以及其中的运转方式。

John的献辞

本书献给世上的科学家、工程师、梦想家和哲学家——那些名不见经传的群体对这个星球上的每个人的生活做出了如此之大的改变。

Luca的致谢

向我的家庭、Yukiko 和 Amelia 表达我的最大谢意，为了他们的支持和充满爱意的耐心。

我也要感谢我的全部 Kaggle 搭档，为了他们的帮助和不停歇地交换想法和观点。特别要致谢 Alberto Boschetti、Giuliano Janson、Bastiaan Sjardin 和 Zacharias Voulgaris。

John的致谢

感谢我的妻子 Rebecca。虽然她现在已经不在人世了，但她的精神在我所写的每本书中，在书页上的每个单词中出现。当没有一个人相信我时，她信任我。

应该感谢 Russ Mullen，为了他对本书的技术编辑。他大大增加了你在这里所见到的资料的准确性和深度。Russ 查找了难以发现的 URL，并也提出了许多建议，极其努力地工作来帮助本书的出版。

应该感谢我的代理 Matt Wagner，他帮助我拿到出版合同并始终会料理大多数作者无暇顾及的所有细节。我感谢他的协助。

很多人阅读了本书的全部或一部分来帮助我改善方法、测试脚本并提供了所有读者可能普遍会想到的内容。这些无偿的志愿者的帮助方式太多了，这里无法一一提及。我特别要感谢 Eva Beattie、Glenn A. Russell、Osvaldo Téllez Almirall 和 Thomas Zinckgraf 的努力，他们提供了广泛的内容，阅读了整本书并无私地致力于这个项目。

最后，我想要感谢 Kyle Looper、Susan Christophersen 和余下的编辑和产品的同事们。

出版者的致谢

策划编辑: Katie Mohr

高级编辑助理: Cherie Case

项目及文字编辑: Susan Christophersen

项目协调者: Vinitha Vikraman

技术编辑: Russ Mullen

封面图片: © iStock.com/Magnilion; © iStock.com/nadla

编辑助理: Claire Brock

前言

你每天完全依靠着数据科学来执行众多的任务或者从他人那里获得服务。事实上，你可能以自己未曾预料过的方式使用了数据科学。例如，当你在今天早晨使用你喜欢的搜索引擎来查找某些事物时，它会对可选事项做出建议。这些事项是由数据科学所提供的。当你去看医生并查明了你所发现的肉块并不是癌症时，医生可能凭借数据科学的帮助来做出诊断。事实上，你可能每天使用数据科学来工作，即使你并不知情。本书不仅让你开始着手使用数据科学来执行众多现实的任务，而且也帮助你了解有多少地方用到了数据科学。通过知晓如何来回答数据科学的问题以及从何使用数据科学，你就比其他人获得了显著的优势，增加了你升职或者获得你真心想要的新工作的机会。

本书简介

本书的主要目标是通过向你展示数据科学不但真的有趣而且借助 Python 相当具有可操作性，从而驱走你对数据科学的恐惧。你可能会希望成为一名计算机科学天才来执行通常和数据科学所关联的复杂任务，但这远远与事实不符。Python 具有许多有用的库来为你在后台挑起所有的重担。你甚至没有意识到有多少工作正在进行中，你不必在意。你真正需要知道的所有一切就是你想要执行的特定的任务，并且 Python 让这些任务变得唾手可得。

本书的部分重点就是在使用正确的工具上面。你从使用 Anaconda 开始，这是一个包含了 IPython 和 IPython Notebook 的产品——两个让人容易使用 Python 来工作的工具。你在一个完全交互式的环境下用 IPython 来做实验。你在 IPython Notebook 中存放的代码具有可供演示的质量，你可以在文档中恰当地混合一些演示元素。它不像完全真正地使用一个开发环境。

你也会在本书中发现一些有趣的技术。例如，你可以使用 Matplotlib 来为你的全部数据科学实验制作绘图，为此本书给你提供了所有的细节。本书也花费了相当的时间仅仅向你展示可供利用的东西，以及你该如何使用它来执行一些真正有趣的计算。很多人想要知道如何执行手写识别——如果你是他们中的一员，你能够使用本书在此过程中领先一步。

当然，你可能还是会对完整的编程环境问题感到焦虑，本书也不会把你抛

弃在黑暗中。一开始，你就可找到完整的 **Anaconda** 安装教程以及一个针对你所需运行的基础 **Python** 编程的快速入门书（和参考）。重点就是让你上手并尽快运行，让例子变得直截了当和简单，这样代码就不会成为学习的绊脚石。

甚至为了更快地吸收概念，本书使用了下面的约定。

- » 你应该键入的文字在本书中用粗体显示。除非当你正在遍历一个步骤列表的时候：因为每个步骤是粗体的，所以要键入的文字就不是粗体的。
- » 当你看到以斜体显示的作为键入序列的部分词组时，你需要把它们值替换成一些对你有用的文字。例如，如果你看到“键入你的名字并按回车键”，你就需要把你的名字替换成你实际的名字。
- » **Web 地址**和**编程源码**以 **monofont** 字体显示。如果你正在一个与因特网连接的设备上阅读本书的电子版时，注意你可以点击 **web 地址**来访问那个网站，就如这样：`http://www.dummies.com`。
- » 当你需要键入命令序列时，你会看到它们被一个特殊的箭头分割开了，就如这样：文件→新文件。在这种情况下，你首先要去文件菜单，接着在那个菜单上选择“新文件”项。结果你就会看到一个新文件被创建出来了。

对读者的要求

你可能发现我们对你所做的一切假设都是难以置信的——毕竟，我们甚至还没见过你！尽管大多数假设可能显得多余，我们还是做出了这些假设来为本书提供一个起点。

你要对想要使用的平台有所了解，这是重要的，因为本书在这方面没有提供任何指导（第3章确实提供了 **Anaconda** 的安装教程）。为了提供给你 **Python** 在数据科学方面应用的最大信息量，本书没有讨论任何平台相关的问题。在你开始使用本书工作前，你真的需要了解如何安装应用程序，如何使用应用程序，以及如何用你所选择的平台做一般性的工作。

本书不是一本数学入门读物。是的，你会看见许多复杂数学的例子，但是重点是在帮助你使用 **Python** 和数据科学来执行分析任务，而不是学习数学理

论。第 1 章和第 2 章让你更精确地理解你所需知道的内容，从而得以成功使用本书。

本书也假设你能够在因特网上获取资源。全书散布着许多可以提升你的学习经验的在线参考资料。无论如何，只有当你实际找到并使用了这些附加资源，它们才是有用的。

本书所使用的图标

当你阅读本书时，你会看见在页边上的图标，它们用来表示资料的有趣程度（或者在有些情况下没有）。本小节简单地描述了本书中的每个图标。



提示是很有用的，因为它们帮你节约时间或者运行一些无需大量额外工作的任务。在本书中的提示是节约时间的技术或者指向了你应该尝试的资源，从而可以从 Python 得到最大的收益，或者可以运行数据科学相关的任务。



你应该避免做出任何以警告标志标记的事情。否则，你可能发现你的应用程序会如预期的那样失效，你会从看起来无懈可击的方程式中得到不正确的答案，或者（在最糟糕的场景下）你会丢失数据。



无论何时当你看到这个图标时，想一想高级的提示或技术。你可能会发现这些花絮信息是有用的，只是用语言表达出来太繁琐，或者它们可能包含了让程序得以运行起来所需的解决方案。无论何时只要你喜欢，就可以略过这些信息。



如果你没有从一个特定章节或小节中吸取到任何东西，记住以这个图标所标记的资料。这段文字通常包含了一个必不可少的过程或者一些信息，它们是你用 Python 工作或者成功运行数据科学相关任务所必须要知道的。

本书之外的内容

本书不是你的 Python 或者数据科学经历的终点——它其实只是一个开端。我们提供了在线内容使得本书更加灵活而且更好地满足你的需求。那样，当我们接收到你的邮件时，我们能够解决问题并告诉你 Python 的升级或者它所关联的附属物的升级如何对本书的内容产生影响。事实上，你可以参阅这些很棒的附件。

- » **作弊单：**你记不记得在学校使用作弊条就为了在测试中取得一个更好的成绩？你这样做过吗？嗯，作弊单就是这样一种类型。它提供给你一些特殊的便条，就是关于你能够使用 Python、IPython、IPython Notebook 和数据科学来做的任务，这不是每个人都知道的。你能够在 <http://www.dummies.com/how-to/content/python-for-datascience-for-dummies-cheat-sheet.html> 上找到针对本书的作弊单。它包含了真正简洁的信息，例如在使用 Python 时使人们遭受痛苦的最常见的编程错误。
- » **Dummies.com 的在线文章：**许多读者把“*For Dummies*”系列书籍的部件页省略过去了，所以出版商决定补救。现在你有了一个真正充分的理由来阅读部件页——在线的内容。每个部件页都有一篇和它关联的文章来提供不适合放在书中的额外的有趣信息。你能够在 <http://www.dummies.com/extras/pythonfordatascience> 上找到针对本书的文章。
- » **更新：**有时候会发生变更。例如，在写本书期间，当我们预测未来时，我们可能没有看到即将浮现出来的变化。在过去，这种可能性简直意味着书籍变得过时和可用性的降低，但是现在你能够在 <http://www.dummies.com/extras/pythonfordatascience> 上找到本书的更新。
- » **配套文件：**嗨！有没有谁真的想键入书中所有的代码并手工重建所有的绘图？绝大多数读者宁愿花时间实际地用 Python 来工作，执行数据科学任务，看看它们能够去实践的有趣的事情，而不是去键入代码。对你来说，幸运的是，在书中所用的例子可供下载，所以你所需要做的一切就是阅读本书来学习 Python 的数据科学应用技术。你可以在 <http://www.dummies.com/store/product/Python-for-Data-Science-For-Dummies.productCd-1118844181,descCd-DOWNLOAD.html> 上找到这些文件。

如何阅读本书

是时候开启你的 Python 数据科学冒险之旅了！如果你对 Python 和它在数据科学方面的应用完全是一个新手，你应该从第 1 章开始并且一步一个脚印按照书本循序渐进来吸收尽可能多的资料。

如果你是一个新手，完全迫不及待地想要尽可能快地上手 Python 数据科学，你

可以跳到第 3 章，但要理解以后你可能会对一些主题产生一点困扰。如果你已经安装了 **Anaconda**（在本书中所使用的编程产品），跳到第 4 章是有可能的，但是要确保略读第 3 章，这样你就知道当我们在写此书时所做的假设。确保在 Python 2.7.9 版本已经安装的前提下安装 **Anaconda**，从而从书中的源码中获得最佳效果。

对 Python 有一些认识以及安装过 **Anaconda** 的读者可以直接转到第 5 章来节省阅读时间。当你有疑问时，你总是能够根据需要回到更早的章节。无论如何，在你转移到下一章之前，你要理解每一项技术是如何工作的，这点是重要的。每一个技术点、代码例子及步骤对你都有着重要的教训，如果你开始略过太多的信息，你就可能会错失关键内容。

目录

第 1 部分 开启 Python 数据科学之门..... 1

第 1 章 探索数据科学与 Python 之间的匹配度 3

1.1 定义 21 世纪最诱人的工作 5

1.1.1 思考数据科学的出现 5

1.1.2 概述数据科学家的核心竞争力 6

1.1.3 连接数据科学和大数据 7

1.1.4 理解编程的角色 7

1.2 创建数据科学管道..... 8

1.2.1 准备数据 8

1.2.2 执行探索性的数据分析 8

1.2.3 从数据中学习 8

1.2.4 可视化 9

1.2.5 获得洞察力和数据产品 9

1.3 理解 Python 在数据科学中的角色..... 9

1.3.1 思考数据科学家的多面性 ... 9

1.3.2 使用一门多用途、简单而高效的语言来工作 10

1.4 快速学会使用 Python 11

1.4.1 加载数据 11

1.4.2 训练模型 12

1.4.3 显示结果 13

第 2 章 介绍 Python 的能力和奇迹..... 14

2.1 为什么是 Python 15

2.1.1 抓住 Python 的核心哲学 ... 16

2.1.2 探索现在和未来的开发目标 16

2.2 使用 Python 工作 17

2.2.1 品味语言 17

2.2.2 理解缩进的需求 17

2.2.3 用命令行或者 IDE 工作 18

2.3 运行快速原型和实验..... 22

2.4 考虑执行速度..... 23

2.5 可视化能力..... 24

2.6 为数据科学使用 Python 生态系统..... 26

2.6.1 使用 SciPy 来访问用于科学的工具 26

2.6.2 使用 NumPy 执行基础的科学计算 26

2.6.3 使用 pandas 来执行数据分析 26

2.6.4 使用 Scikit-learn 实现机器学习 27

2.6.5 使用 matplotlib 来标绘数据... 27

2.6.6 使用 Beautiful Soup 来解析 HTML 文档..... 27

第 3 章 为数据科学设置 Python ... 29

- 3.1 考虑现成的跨平台的用于科学的分发包 30
 - 3.1.1 获取 Continuum Analytics Anaconda 31
 - 3.1.2 获取 Enthought Canopy Express 32
 - 3.1.3 获取 pythonxy 32
 - 3.1.4 获取 WinPython 33
- 3.2 在 Windows 上安装 Anaconda 33
- 3.3 在 Linux 上安装 Anaconda 36
- 3.4 在 Mac OS X 上安装 Anaconda 37
- 3.5 下载数据集和示例代码 38
 - 3.5.1 使用 IPython Notebook 39
 - 3.5.2 定义代码仓库 40
 - 3.5.3 理解本书中所使用的数据集 45

第 4 章 复习 Python 基础 47

- 4.1 使用数字和逻辑来工作 49
 - 4.1.1 执行变量赋值 50
 - 4.1.2 做算术运算 50
 - 4.1.3 使用布尔表达式来比较数据 52
- 4.2 创建和使用字符串 54
- 4.3 与日期交互 55
- 4.4 创建并使用函数 56
 - 4.4.1 创建可复用函数 56
 - 4.4.2 以各种不同的方式调用函数 58

- 4.5 使用条件和循环语句 61
 - 4.5.1 使用 if 语句做决策 61
 - 4.5.2 使用嵌套决策在多个选项间做出选择 62
 - 4.5.3 使用 for 执行重复任务 63
 - 4.5.4 使用 while 语句 64
- 4.6 使用 Sets、Lists 和 Tuples 来存储数据 64
 - 4.6.1 在 set 上执行操作 65
 - 4.6.2 使用 list 来工作 66
 - 4.6.3 创建和使用 Tuple 67
- 4.7 定义有用的迭代器 69
- 4.8 使用 Dictionaries 来索引数据 70

第 2 部分 开始着手于数据 ... 71

第 5 章 使用真实数据工作 73

- 5.1 上传、流化并采样数据 74
 - 5.1.1 把少量数据上传至内存 75
 - 5.1.2 把大量数据流化放入内存 76
 - 5.1.3 采样数据 77
- 5.2 以结构化的平面文件形式来访问数据 78
 - 5.2.1 从文本文件中读取 79
 - 5.2.2 读取 CSV 定界的格式 80
 - 5.2.3 读取 Excel 和其他的微软办公文件 82
- 5.3 以非结构化文件的形式来发送数据 83
- 5.4 管理来自关系型数据库中的数据 86

5.5 与来自 NoSQL 数据库中的 数据进行交互	87	6.8 在任何层次聚合数据	115
5.6 访问来自 Web 的数据	88	第 7 章 数据整形	117
第 6 章 整理你的数据	92	7.1 使用 HTML 页面来工作	118
6.1 兼顾 NumPy 和 pandas	93	7.1.1 解析 XML 和 HTML	118
6.1.1 知道什么时候使用 NumPy	93	7.1.2 使用 XPath 来抽取 数据	119
6.1.2 知道什么时候使用 pandas	93	7.2 使用原始文本来工作	120
6.2 验证你的数据	95	7.2.1 处理 Unicode 码	120
6.2.1 了解你的数据中有什么	95	7.2.2 词干提取和停止词 移除	122
6.2.2 去重	96	7.2.3 介绍正则表达式	124
6.2.3 创建数据地图和数据规划	97	7.3 使用并超越词袋模型	126
6.3 处理分类变量	99	7.3.1 理解词袋模型	127
6.3.1 创建分类变量	100	7.3.2 用 n 元文法模型 (n -grams) 工作	128
6.3.2 重命名层级	102	7.3.3 实现 TF-IDF 变换	130
6.3.3 组合层级	102	7.4 使用图数据来工作	131
6.4 处理你数据中的日期	104	7.4.1 理解邻接矩阵	131
6.4.1 格式化日期和时间值	104	7.4.2 使用 NetworkX 基础	132
6.4.2 使用正确的时间转换	105	第 8 章 将你所知的付诸于实践	134
6.5 处理丢失值	106	8.1 将问题和数据置于上下文中 去理解	135
6.5.1 寻找丢失的数据	106	8.1.1 评估数据科学问题	136
6.5.2 为丢失项编码	107	8.1.2 研究方案	136
6.5.3 为丢失数据估值	108	8.1.3 构想出假设	137
6.6 交叉分析: 过滤并选取数据	109	8.1.4 准备数据	138
6.6.1 切分行	109	8.2 思考创建特征的艺术	138
6.6.2 切分列	110	8.2.1 定义特征创建	138
6.6.3 切块	110	8.2.2 组合变量	139
6.7 连接和变换	111		
6.7.1 增加新的实例和变量	112		
6.7.2 移除数据	113		
6.7.3 排序和搅乱	114		

8.2.3	理解分级和离散化	140
8.2.4	使用指示变量	140
8.2.5	变换分布	140
8.3	在数组上执行运算	141
8.3.1	使用向量化	141
8.3.2	在向量和矩阵上执行 简单的算法	142
8.3.3	执行矩阵向量乘法	142
8.3.4	执行矩阵乘法	143

第 3 部分 把不可见的东西 可视化..... 145

第 9 章 获得 Matplotlib 的速成 课程 147

9.1	开始使用图表	148
9.1.1	定义标图	148
9.1.2	画多条和多标图	149
9.1.3	保存你的工作	149
9.2	设置轴、刻度和网格	150
9.2.1	得到轴	151
9.2.2	格式化轴	151
9.2.3	添加网格	152
9.3	定义线条外观	153
9.3.1	使用线条样式工作	153
9.3.2	使用颜色	155
9.3.3	添加标记	155
9.4	使用标签、注释和图例	157
9.4.1	添加标签	158
9.4.2	注释图表	158

9.4.3	创建图例	159
-------	------------	-----

第 10 章 将数据可视化 161

10.1	选择合适的图表	162
10.1.1	用饼图展示整体的局部 组成	162
10.1.2	用柱状图来创建比较 ...	163
10.1.3	用直方图来展示分布 ...	164
10.1.4	使用箱线图来描绘组 ...	166
10.1.5	使用散点图看数据 模式	167
10.2	创建高级的散点图	168
10.2.1	描绘组群	168
10.2.2	展示关联	169
10.3	标绘时间序列	171
10.3.1	在轴上表示时间	171
10.3.2	标绘随时间的趋势	172
10.4	标绘地理数据	174
10.5	把图做可视化	176
10.5.1	开发无向图	176
10.5.2	开发有向图	177

第 11 章 理解工具 180

11.1	使用 IPython 控制台	181
11.1.1	与屏幕文本交互	181
11.1.2	改变窗口外观	182
11.1.3	获取 Python 帮助	184
11.1.4	获取 IPython 帮助	185
11.1.5	使用魔法函数	186
11.1.6	探索对象	187

11.2 使用 IPython Notebook	188	13.2.1 度量集中化趋势	217
11.2.1 使用样式来工作	189	13.2.2 测量方差和区间	217
11.2.2 重启内核	190	13.2.3 使用分位数来工作	218
11.2.3 恢复检查点	191	13.2.4 定义正态化度量	219
11.3 执行多媒体和图像整合	192	13.3 为分类型数据计数	220
11.3.1 嵌入标图和其他图片 ...	192	13.3.1 理解频率	220
11.3.2 从在线网站上加载例子...	193	13.3.2 创建列联表	221
11.3.3 获取在线图像和多媒体...	193	13.4 为 EDA 创建应用可视化.....	222
第 4 部分 处理数据	195	13.4.1 检查箱线图	222
第 12 章 拓展 Python 的能力 ...	197	13.4.2 在箱线图之后执行 t 检验	223
12.1 玩转 Scikit-learn.....	198	13.4.3 观察平行坐标	224
12.1.1 理解 Scikit-learn 中的类 ...	198	13.4.4 为分布作图	225
12.1.2 为数据科学定义应用 ...	199	13.4.5 标绘散点图	226
12.2 执行散列法	202	13.5 理解相关性	228
12.2.1 使用散列函数	202	13.5.1 使用协方差和关联性 ...	228
12.2.2 演示散列法	203	13.5.2 使用非参数相关性	230
12.2.3 使用确定性选择来工作...	205	13.5.3 考虑表格的卡方检验 ...	230
12.3 考虑计时和性能	206	13.6 修改数据分布	231
12.3.1 用 timeit 来做基线检测 ...	207	13.6.1 使用正态分布	232
12.3.2 使用内存剖析器来工作...	209	13.6.2 创建 Z 评分标准化	232
12.4 并行运行	210	13.6.3 转换其他的著名分布 ...	232
12.4.1 执行多核并行化	211	第 14 章 降维	234
12.4.2 演示多核处理	212	14.1 理解 SVD	235
第 13 章 探索数据分析	214	14.1.1 寻求降维	236
13.1 EDA 方法	215	14.1.2 使用 SVD 来测量不可见 的信息	237
13.2 为 Numeric 数据定义描述 性的统计量	216	14.2 执行因子和主成分分析	238
		14.2.1 考虑心理测量模型	239