

70多个常用的数据可视化实例，让您掌握流行的Python数据可视化库

# Python 数据可视化 编程实战（第2版）

**Python Data Visualization Cookbook**  
Second Edition

[爱尔兰] 伊戈尔·米洛瓦诺维奇 (Igor Milovanović)

[法] 迪米特里·富雷斯 (Dimitry Foures) 著

[意大利] 朱塞佩·韦蒂格利 (Giuseppe Vettigli)

颀清山 译



中国工信出版集团



人民邮电出版社  
POSTS & TELECOM PRESS



# Python 数据可视化 编程实战 (第2版)

[爱尔兰] 伊戈尔·米洛瓦诺维奇 (Igor Milovanović)

[法] 迪米特里·富雷斯 (Dimitry Foures) 著

[意大利] 朱塞佩·韦蒂格利 (Giuseppe Vettigli)

颢清山 译

人民邮电出版社

北京

## 图书在版编目 (CIP) 数据

Python数据可视化编程实战：第2版 / (爱尔兰) 伊戈尔·米洛瓦诺维奇, (法) 迪米特里·富雷斯 (Dimitry Foures), (意) 朱塞佩·韦蒂格利 (Giuseppe Vettigli) 著; 颀清山译. — 北京: 人民邮电出版社, 2018.9  
ISBN 978-7-115-48842-8

I. ①P… II. ①伊… ②迪… ③朱… ④颀… 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148596号

## 版权声明

Copyright ©2015 Packt Publishing. First published in the English language under the title Python Data Visualization Cookbook, Second Edition.

All rights reserved.

本书由英国 Packt Publishing 公司授权人民邮电出版社出版。未经出版者书面许可, 对本书的任何部分不得以任何方式或任何手段复制和传播。

版权所有, 侵权必究。

- 
- ◆ 著 [爱尔兰] 伊戈尔·米洛瓦诺维奇 (Igor Milovanović)  
[法] 迪米特里·富雷斯 (Dimitry Foures)  
[意大利] 朱塞佩·韦蒂格利 (Giuseppe Vettigli)
  - 译 颀清山
  - 责任编辑 武晓燕
  - 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: 17.25  
字数: 329千字 2018年9月第1版  
印数: 1-2400册 2018年9月河北第1次印刷
- 著作权合同登记号 图字: 01-2017-7882 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

本书是一本使用 Python 实现数据可视化编程的实战指南，介绍了如何使用 Python 最流行的库、通过 70 余种方法创建美观的数据可视化效果。

全书共 9 章，分别介绍了准备工作环境、了解数据、绘制并定制化图表、学习更多图表和定制化、创建 3D 可视化图表、用图像和地图绘制图表、使用正确的图表理解数据、matplotlib 相关知识以及使用 Plot.ly 进行云端可视化。

本书适合那些对 Python 编程有一定基础的开发人员阅读，它可以帮助读者从头开始了解数据、数据格式、数据可视化，并学会使用 Python 可视化数据。

# 作者简介

**Igor Milovanović** 是一名在 Linux 系统和软件工程领域有深厚背景的、经验丰富的开发人员，熟悉创建可扩展数据驱动分布式富软件系统的技术。

身为一个高性能系统设计的布道者，他对软件架构和软件开发方法论有着浓厚的兴趣。他一直坚持倡导促进高质量软件的方法论，如测试驱动开发、一键部署和持续集成。

他也拥有丰富的产品开发知识。他拥有领域经验知识，并参加过官方培训，他能够在业务和开发人员之间很好地传递业务知识和业务流程。

Igor 非常感谢他的女友允许他把大量的时间花费在本书的写作上而没有陪伴她，并在他无休止地谈论本书时甘愿做一个热心的听众。他感谢他的哥哥一直以来做他最坚强的后盾。他还要感谢他的父母，给予他自由的发展空间，让他成为今天的自己。

**Dimitry Foures** 是一名拥有应用数学和理论物理背景的数据科学家。在里昂高等师范学校（法国）完成他的物理专业的本科学业后，他继续在巴黎综合理工学院学习流体力学，并获得了一等硕士学位。他拥有剑桥大学应用数学专业博士学位。他目前在剑桥的一家智能能源初创公司担任数据科学家一职，与剑桥大学有着非常密切的合作。

**Giuseppe Vettigli** 是一名数据科学家，他在产业界和学术界从事过多年的研究工作。他的工作关注从结构化及非结构化数据中提取信息进行机器学习模型的开发和应用。他也经常发表利用 Python 进行科学计算和数据可视化的文章。

# 评阅者简介

**Kostiantyn Kucher** 出生在乌克兰敖德萨。他于 2012 年在敖德萨国立理工大学获得计算机科学专业的硕士学位。他使用 Python、Matplotlib 和 PIL 从事机器学习和图像识别的工作。

Kostiantyn 从 2013 年开始成为一名计算机专业信息可视化方向的博士研究生。在 Andreas Kerren 博导的指导下，他在瑞典林奈大学计算机科学系的 ISOVIS 小组进行研究。

Kostiantyn 也是本书第 1 版的技术评阅者。

# 译者序

图形可视化是展示数据的一个非常好的手段，好的图表自己会说话。毋庸多言，在 Python 的世界里，matplotlib 是最著名的绘图库，它支持几乎所有的 2D 绘图和部分 3D 绘图，被广泛地应用在科学计算和数据可视化领域。但是介绍 matplotlib 的中文书籍很少，大部分书籍只是在部分章节中提到了 matplotlib 的基本用法，因此在内容和深度上都力有不逮。本书则是一本专门介绍 matplotlib 的译著。

matplotlib 是一个开源项目，由 John Hunter 发起。关于 matplotlib 的由来，有一个小故事。John Hunter 和他研究癫痫症的同事借助一个专有软件做脑皮层电图分析，但是他所在的实验室只有一份该电图分析软件的许可。他和许多一起工作的同事不得不轮流使用该软件的硬件加密狗。于是，John Hunter 便有了开发一个工具来替代当前所使用的软件的想法。当时 MATLAB 被广泛应用在生物医学界中，John Hunter 最初是想开发一个基于 MATLAB 的版本，但是由于 MATLAB 的一些限制和不足，加上他本身对 Python 非常熟悉，于是就有了 matplotlib 的诞生。

所以，无论从名字上，还是从所提供的函数名称、参数及使用方法，matplotlib 都与 MATLAB 非常相似。对于一个 MATLAB 开发人员，使用 matplotlib 会相当得心应手。即使对不熟悉 MATLAB 的开发人员（譬如我），对其函数的使用也能够一目了然，而且 matplotlib 有着非常丰富的文档和实例，再加上本书的介绍，学习起来将会非常轻松。

matplotlib 命令提供了交互绘图的方式，在 Python 的交互式 shell 中，我们可以执行 matplotlib 命令来实时地绘制图形并对其进行修改。生成的图像可以保存成许多格式，这取决于其所使用的后端，但绝大多数后端都支持如 png、pdf、ps、eps 和 svg 等格式。

在本书中，作者对内容进行了整理，剔除了一些与章节联系不是很大的延伸内容，对排版也进行了修改，使得本书在内容编排上更加简洁紧凑。在本书中，作者还引入了 pandas 和 Plot.ly。其中 pandas 是一个功能强大且高性能的数据分析工具。Plot.ly 是一款非常优秀

的在线图表工具，它非常注重图表的可操作性及分享。在内容上，第 2 版更加完整地涵盖了 Python 数据可视化领域用到的主流工具。限于篇幅，本书不可能对所有工具做完整详细的介绍，但本书所讲的工具能满足读者大部分的数据可视化需要，读者可以根据自己的需要有选择地深入学习。

在这里，我要特别感谢我的妻子董秋影，在精神和专业知识上，她都给予了我莫大的帮助，没有她就没有这本译稿的完成。她从事医疗图像算法工作，对各种图形和算法以及 MATLAB 都有很深的了解，本书的每一章都经过了认真的审阅校对。最后，感谢人民邮电出版社武晓燕老师专业细心地审核，和武老师合作很轻松、很开心。

由于译者水平有限，错误和失误在所难免，如有任何意见和建议，请不吝指正，我将感激不尽。

颢清山

2018 年 1 月于 墨尔本



# 前言

最好的数据是我们能看到并理解的数据。作为开发人员和数据科学家，我们希望可以创造并构建出最全面且容易理解的可视化图形。然而这并非易事，我们需要找到数据，对它读取、清理、过滤，然后使用恰当的工具将其可视化。本书通过直接和简单（有时不那么简单）的方法解释了进行数据读取、清理以及可视化的流程。

本书涉及如何读取本地数据、远程数据、CSV、JSON 以及关系型数据库中的数据。

通过 `matplotlib`，我们能用一行简单的 Python 代码绘制出一些简单的图表，但是进行更高级的绘图还需要 Python 之外的其他知识。我们需要理解信息理论和人类的审美学来生成最吸引人的可视化效果。

本书将介绍在 Python 中使用 `matplotlib` 绘图的一些实战练习，以及不同图表特性的使用情况及其用法示例。

## 本书涵盖内容

第 1 章，准备工作环境，包括一些安装方法，以及如何在不同平台上安装所需的 Python 包和库的一些建议。

第 2 章，了解数据，介绍通用的数据格式，以及如何进行读写，读写的格式如 CSV、JSON、XSL 或者关系型数据库。

第 3 章，绘制并定制化图表，着手绘制简单的图表并涉及一些图表定制化的内容。

第 4 章，学习更多图表和定制化，继续上一章内容，介绍更多的高级表格和网格定制化。

第 5 章，创建 3D 可视化图表，介绍三维数据的可视化，如 3D 柱状图、3D 直方图，以及 `matplotlib` 动画。

第6章，用图像和地图绘制图表，涵盖图像处理、在地图上投射数据以及创建 CAPTCHA 测试图像。

第7章，使用正确的图表理解数据，涵盖一些更高级的绘图技术的讲解和实战练习，如频谱图和相关性图形。

第8章，更多的 matplotlib 知识，介绍一些图表如甘特图、箱线图，并且介绍如何在 matplotlib 中使用 LaTeX 渲染文本。

第9章，使用 Plot.ly 进行云端可视化，介绍如何使用 Plot.ly 在云端环境中创建和分享可视化图形。

## 准备工作

学习本书时，你需要在自己的操作系统上安装 Python 2.7.3 或更高版本。

本书用到的另一个软件包是 IPython，它是一个交互式的 Python 环境，功能非常强大、灵活。你可以通过基于 Linux 平台的包管理工具或者用于 Windows 和 Mac OS 系统的预安装文件安装它。

一般来说，如果你对 Python 的安装和相关软件的安装不熟悉，那么强烈推荐你使用预打包的 Python 科学发行包，如 Anaconda、Enthought Python 发行包或者 Python(x,y) 进行安装。

其他所需的软件主要是一些 Python 安装包，读者可全部通过 Python 安装管理器 pip 进行安装。pip 本身通过 Python 的 easy\_install 安装工具安装。

## 本书适合的读者

本书是为那些已经了解 Python 编程，并想学习如何使用实用的方法对手头的数据进行可视化的开发人员和数据科学家编写的。如果你对数据可视化有所耳闻，但却不知道从何着手，本书将会从头开始指导你如何了解数据、数据格式、数据可视化，以及如何使用 Python 可视化数据。

你需要知道一些一般的编程概念，如果你有编程经验会非常有帮助。本书中的代码几乎是逐行讲解的。阅读本书不需要任何数学知识，书中介绍的每一个概念都有详细的讲解。此外，本书还提供了一些参考资料，以供有兴趣的人员进一步阅读。

## 章节

在本书中，你会发现几个频繁出现的标题（准备工作、操作步骤、工作原理、补充说明、另请参阅）。

为了清楚地说明如何完成一个实战练习，我们包括以下几个小节。

### 准备工作

本节告诉你练习要达到的目的，以及完成练习所需的软件或其他准备工作。

### 操作步骤

本节包括练习中需要遵循的步骤。

### 工作原理

本节通常是对上节提到的内容进行详细的讲解。

### 补充说明

本节包括了一些练习相关的附加内容，以帮助读者对该练习有更深入的了解。

### 另请参阅

本节提供了练习相关的更多有用的信息。

## 体例约定

在本书中，不同的信息由一些不同体例的文本来区分。这里有一些文本体例的例子以及它们的含义解释。

书中的代码文字、数据库表名、文件夹名称、文件名、文件扩展名、路径名、模拟 URL、用户输入和 Twitter 用户名的显示格式如下：“我们把小演示程序封装在 DemoPIL 类中，这样可以共享示例函数 `run_fixed_filters_demo` 的代码，并能很容易地对其进行扩展。”

代码块设置如下：

```
def _load_image(self, imfile):
```

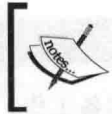
```
self.im = mplimage.imread(imfile)
```

当我们想要让你关注代码块中的某一特定部分时，相关的行或元素将设置为**粗体**：

```
for a in range(10):  
    print a
```

所有的命令行输入或者输出的写法如下：

```
$ sudo python setup.py install
```



警告或者重要的说明出现在这样的文本框中。



表示提示和技巧。

## 读者反馈

我们欢迎读者的反馈意见。如果对本书有任何的想法，喜欢或者不喜欢哪些内容，都可以告诉我们。这些反馈意见对于帮助我们创作出对大家真正有所帮助的作品至关重要。

你可以将一般的反馈以电子邮件的形式发送到 [feedback@packtpub.com](mailto:feedback@packtpub.com)，并在邮件主题中包含书名。

如果你在某一方面很有造诣，并且愿意著书或参与合著，可以参考我们的作者指南。

# 资源与支持

本书由异步社区出品，社区 (<https://www.epubit.com/>) 为您提供相关资源和后续服务。

## 配套资源

本书提供如下资源：

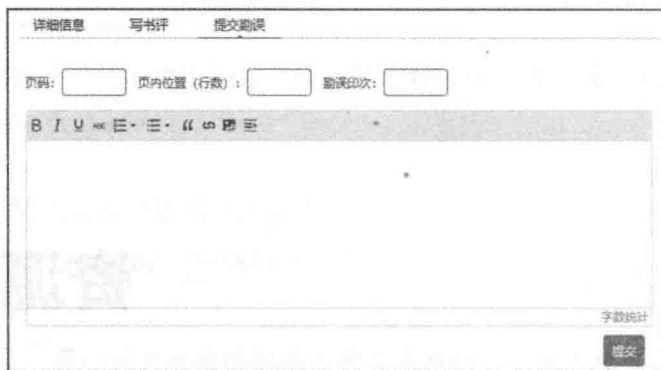
- 本书源代码；
- 书中彩图文件。

要获得以上配套资源，请在异步社区本书页面中点击 **配套资源**，跳转到下载界面，按提示进行操作即可。注意：为保证购书读者的权益，该操作会给出相关提示，要求输入提取码进行验证。

## 提交勘误

作者和编辑尽最大努力来确保书中内容的准确性，但难免会存在疏漏。欢迎您将发现的问题反馈给我们，帮助我们提升图书的质量。

当您发现错误时，请登录异步社区，按书名搜索，进入本书页面，点击“提交勘误”，输入勘误信息，单击“提交”按钮即可。本书的作者和编辑会对您提交的勘误进行审核，确认并接受后，您将获赠异步社区的 100 积分。积分可用于在异步社区兑换优惠券、样书或奖品。



## 扫码关注本书

扫描下方二维码，您将会在异步社区微信服务号中看到本书信息及相关的服务提示。



## 与我们联系

我们的联系邮箱是 [contact@epubit.com.cn](mailto:contact@epubit.com.cn)。

如果您对本书有任何疑问或建议，请您发邮件给我们，并请在邮件标题中注明本书书名，以便我们更高效地做出反馈。

如果您有兴趣出版图书、录制教学视频，或者参与图书翻译、技术审校等工作，可以发邮件给我们；有意出版图书的作者也可以到异步社区在线提交投稿（直接访问 [www.epubit.com/selfpublish/submission](http://www.epubit.com/selfpublish/submission) 即可）。

如果您是学校、培训机构或企业，想批量购买本书或异步社区出版的其他图书，也可以发邮件给我们。

如果您在网上发现有针对异步社区出品图书的各种形式的盗版行为，包括对图书全部或部分内容的非授权传播，请您将怀疑有侵权行为的链接发邮件给我们。您的这一举动是对作者权益的保护，也是我们持续为您提供有价值的内容的动力之源。

## 关于异步社区和异步图书

“异步社区”是人民邮电出版社旗下 IT 专业图书社区，致力于出版精品 IT 技术图书和相关学习产品，为作译者提供优质出版服务。异步社区创办于 2015 年 8 月，提供大量精品 IT 技术图书和电子书，以及高品质技术文章和视频课程。更多详情请访问异步社区官网 <https://www.epubit.com>。

“异步图书”是由异步社区编辑团队策划出版的精品 IT 专业图书的品牌，依托于人民邮电出版社近 30 年的计算机图书出版积累和专业编辑团队，相关图书在封面上印有异步图书的 LOGO。异步图书的出版领域包括软件开发、大数据、AI、测试、前端、网络技术 etc。



异步社区



微信服务号

# 目录

第 1 章 准备工作环境	1	1.6.1 操作步骤	10
1.1 介绍	1	1.6.2 安装过程说明	10
1.2 安装 matplotlib、Numpy 和 Scipy 库	2	1.6.3 补充说明	10
1.2.1 准备工作	2	1.7 安装 requests 模块	11
1.2.2 操作步骤	3	1.7.1 操作步骤	11
1.2.3 工作原理	4	1.7.2 requests 使用说明	11
1.2.4 补充说明	4	1.8 在代码中配置 matplotlib 参数	12
1.3 安装 virtualenv 和 virtualenvwrapper	4	1.8.1 准备工作	12
1.3.1 准备工作	5	1.8.2 操作步骤	12
1.3.2 操作步骤	5	1.8.3 代码解析	13
1.4 在 Mac OS X 上安装 matplotlib	7	1.9 为项目设置 matplotlib 参数	13
1.4.1 准备工作	7	1.9.1 准备工作	13
1.4.2 操作步骤	7	1.9.2 操作步骤	14
1.5 在 Windows 上安装 matplotlib	8	1.9.3 配置过程说明	14
1.5.1 准备工作	8	1.9.4 补充说明	15
1.5.2 操作步骤	9	第 2 章 了解数据	16
1.5.3 补充说明	9	2.1 简介	17
1.6 安装图像处理工具：Python 图像库 (PIL)	10	2.2 从 CSV 文件导入数据	17
		2.2.1 准备工作	17
		2.2.2 操作步骤	17
		2.2.3 工作原理	18
		2.2.4 补充说明	19



2.3 从 Microsoft Excel 文件导入数据 .....	19	2.9.1 准备工作 .....	34
2.3.1 准备工作 .....	20	2.9.2 操作步骤 .....	34
2.3.2 操作步骤 .....	20	2.9.3 工作原理 .....	37
2.3.3 工作原理 .....	21	2.9.4 补充说明 .....	37
2.3.4 补充说明 .....	21	2.10 清理异常值 .....	38
2.4 从定宽数据文件导入数据 .....	22	2.10.1 准备工作 .....	38
2.4.1 准备工作 .....	22	2.10.2 操作步骤 .....	38
2.4.2 操作步骤 .....	22	2.10.3 补充说明 .....	43
2.4.3 工作原理 .....	23	2.11 读取大块数据文件 .....	43
2.5 从制表符分隔的文件中读取数据 .....	24	2.11.1 操作步骤 .....	43
2.5.1 准备工作 .....	24	2.11.2 工作原理 .....	44
2.5.2 操作步骤 .....	24	2.11.3 补充说明 .....	45
2.5.3 工作原理 .....	25	2.12 读取流数据源 .....	45
2.5.4 补充说明 .....	25	2.12.1 操作步骤 .....	45
2.6 从 JSON 数据源导入数据 .....	26	2.12.2 工作原理 .....	46
2.6.1 准备工作 .....	26	2.12.3 补充说明 .....	46
2.6.2 操作步骤 .....	26	2.13 导入图像数据到 NumPy 数组 .....	47
2.6.3 工作原理 .....	26	2.13.1 准备工作 .....	47
2.6.4 补充说明 .....	27	2.13.2 操作步骤 .....	47
2.7 导出数据到 JSON、CSV 和 Excel .....	28	2.13.3 工作原理 .....	50
2.7.1 准备工作 .....	28	2.13.4 补充说明 .....	52
2.7.2 操作步骤 .....	28	2.14 生成可控的随机数据集合 .....	52
2.7.3 工作原理 .....	31	2.14.1 准备工作 .....	53
2.7.4 补充说明 .....	32	2.14.2 操作步骤 .....	53
2.8 用 Pandas 导入和操作数据 .....	32	2.15 真实数据的噪声平滑处理 .....	59
2.8.1 准备工作 .....	32	2.15.1 准备工作 .....	60
2.8.2 操作步骤 .....	33	2.15.2 操作步骤 .....	60
2.8.3 工作原理 .....	33	2.15.3 工作原理 .....	60
2.8.4 补充说明 .....	33	2.15.4 补充说明 .....	64
2.9 从数据库导入数据 .....	34	<b>第3章 绘制并定制化图表 .....</b>	<b>66</b>
		3.1 简介 .....	67