




# Python数据分析 (影印版)

Python Data Analysis

Ivan Idris 著

[PACKT]  
PUBLISHING

学习如何运用流行的开源Python模块实现强大的数据分析技术

 东南大学出版社  
SOUTHEAST UNIVERSITY PRESS

# Python 数据分析(影印版)

*Ivan Idris* 著

南京 东南大学出版社

## 图书在版编目(CIP)数据

Python 数据分析:英文/(印尼)伊德里斯(Idris, I.)  
著. —影印本. —南京:东南大学出版社, 2016.1

书名原文:Python Data Analysis

ISBN 978-7-5641-6064-7

I. ①P… II. ①伊… III. ①软件工具—程序设计—英文 IV. ①TP311.56

中国版本图书馆 CIP 数据核字(2015)第 243362 号

© 2014 by PACKT Publishing Ltd

Reprint of the English Edition, jointly published by PACKT Publishing Ltd and Southeast University Press, 2016.  
Authorized reprint of the original English edition, 2015 PACKT Publishing Ltd, the owner of all rights to publish and sell the same.

All rights reserved including the rights of reproduction in whole or in part in any form.

英文原版由 PACKT Publishing Ltd 出版 2014。

英文影印版由东南大学出版社出版 2016。此影印版的出版和销售得到出版权和销售权的所有者——PACKT Publishing Ltd 的许可。

版权所有,未得书面许可,本书的任何部分和全部不得以任何形式重制。

## Python 数据分析(影印版)

---

出版发行:东南大学出版社

地 址:南京四牌楼 2 号 邮编:210096

出 版 人:江建中

网 址:<http://www.seupress.com>

电子邮件:[press@seupress.com](mailto:press@seupress.com)

印 刷:常州市武进第三印刷有限公司

开 本:787 毫米×980 毫米 16 开本

印 张:21.75

字 数:426 千字

版 次:2016 年 1 月第 1 版

印 次:2016 年 1 月第 1 次印刷

书 号:ISBN 978-7-5641-6064-7

定 价:68.00 元

---

本社图书若有印装质量问题,请直接与营销部联系。电话(传真):025-83791830

# Credits

**Author**

Ivan Idris

**Reviewers**

Amanda Casari

Thomas A. Dyar

Dr. Hari Shanker Gupta

Puneet Narula

Alan J. Salmoni

**Commissioning Editor**

Akram Hussain

**Acquisition Editor**

Owen Roberts

**Content Development Editor**

Prachi Bisht

**Technical Editor**

Pankaj Kadam

**Copy Editors**

Roshni Banerjee

Sarang Chari

Adithi Shetty

**Project Coordinator**

Shipra Chawhan

**Proofreaders**

Simran Bhogal

Maria Gould

Ameesha Green

**Indexers**

Hemangini Bari

Mariammal Chettiyar

Rekha Nair

Tejal Soni

**Graphics**

Sheetal Aute

**Production Coordinators**

Adonia Jones

Manu Joseph

Komal Ramchandani

**Cover Work**

Manu Joseph

# About the Author

**Ivan Idris** has an MSc degree in Experimental Physics. His graduation thesis had a strong emphasis on Applied Computer Science. After graduating, he worked for several companies as Java developer, data warehouse developer, and QA analyst. His main professional interests are Business Intelligence, Big Data, and Cloud Computing.

Ivan Idris enjoys writing clean, testable code and interesting technical articles. He is the author of *NumPy Beginner's Guide - Second Edition*, *NumPy Cookbook*, and *Learning NumPy Array*, all by Packt Publishing. You can find more information and a blog with a few NumPy examples at [ivanidris.net](http://ivanidris.net).

---

I would like to take this opportunity to thank the reviewers and the team at Packt Publishing for making this book possible. Also, my thanks go to my teachers, professors, and colleagues, who taught me about science and programming. Last but not least, I would like to acknowledge my parents, family, and friends for their support.

---

# About the Reviewers

**Amanda Casari** is currently a data scientist and engineer in the Seattle area. Amanda received her MSEE degree and Certificate of Study in Complex Systems from the University of Vermont and a BS degree in Systems Engineering from the United States Naval Academy. She has more than 10 years of professional experience, ranging from naval officer, analyst, conservation trip leader to integration engineer. Her research interests focus on discovering attributes of natural systems to update and optimize man-made complex networks. Amanda is passionate about making Mathematics and Science approachable to everyone.

---

I would like to thank my family for supporting our journey and inspiring me during this effort, N. Manukyan for all of her data enthusiasm, C. Stone for creative breakfasts, the Carnation Climbing Club, and P. Nathan for kindly encouraging my myriad interests.

---

**Thomas A. Dyar** (Tom) is a senior data scientist in the Genomic Sciences group at BD Technologies ([www.bd.com](http://www.bd.com)), Research Triangle Park, North Carolina, where he develops algorithms to process genomic data in a variety of contexts—from targeted panels to whole genomes—for infectious disease and oncology diagnostics applications. His areas of expertise are scientific programming in Java, Python, and R; machine learning, including neural networks and kernel methods; and data analysis and visualization. His primary interests are in conceptualizing and developing large-scale data-driven solutions using Cloud resources.

Tom started his career in software, developing neural networks and expert systems tools for process control in the aerospace and petrochemical industries. He has also worked on distributed virtual environments for stroke rehabilitation at MIT and automated image processing for high-throughput cell biology experiments at BD.

Tom earned his BA degree in Pure & Applied Mathematics from Boston University and is a member of the ACM and IEEE associations.

**Dr. Hari Shanker Gupta** is a senior quantitative research analyst working in the area of algorithmic trading system development. Prior to this, he was a post-doctoral fellow at the Indian Institute of Science (IISc), Bangalore, India. He obtained his PhD in Applied Mathematics and Scientific Computation from IISc. He completed his MSc in Mathematics from Banaras Hindu University (BHU), Varanasi, India. During his MSc, he was awarded four gold medals for outstanding performance at BHU.

Hari has published five research papers in reputed journals in the field of Mathematics and Scientific Computation. He has experience working in the areas of Mathematics, Statistics, and Computation. His experience includes working in numerical methods, partial differential equations, mathematical finance, stochastic calculus, data analysis, finite difference, and finite element methods. He is very comfortable with the mathematics software, MATLAB; the statistics programming language, R; Python; and the programming language, C.

He has reviewed the book *Introduction to R for Quantitative Finance*, Packt Publishing.

**Puneet Narula** has over 8 years of experience in the Banking and Finance industry, but his aptitude and passion for the technology sector has brought him back into the world of data and analytics. Leaving behind a stable career in banking was a very tough decision, but following his dreams was even more important to him. He completed his MSc degree in Data Analytics from Dublin Institute of Technology in 2013 to enter the world of analytics and data science. Currently, Puneet is working with Web Reservations International as a PPC data analyst.

At Web Reservations International (WRI), Puneet works with massive clickstream data from both direct and affiliate sources. The technologies used for the analysis is a combination of RapidMiner, R, and Python.

---

I want to thank Silviu Preoteasa for all his support and motivation at all times.

---

**Alan J. Salmoni** enjoys making sense of data and is the author of Salstat (<http://www.salstat.com>). He has been using Python for data analysis since 2001 and has taught statistics to undergraduates and postgraduates. When not with his family, he spends time generating large statistical models of text for natural language processing.

Alan owns a company, Thought Into Design, which specializes in data analysis and user experience.

---

I would like to thank my wife, Jell, and my daughter, Louise,  
for their patience.

---



# www.PacktPub.com

## Support files, eBooks, discount offers, and more

You might want to visit [www.PacktPub.com](http://www.PacktPub.com) for support files and downloads related to your book.

Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at [www.PacktPub.com](http://www.PacktPub.com) and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at [service@packtpub.com](mailto:service@packtpub.com) for more details.

At [www.PacktPub.com](http://www.PacktPub.com), you can also read a collection of free technical articles, sign up for a range of free newsletters and receive exclusive discounts and offers on Packt books and eBooks.



<http://PacktLib.PacktPub.com>

Do you need instant solutions to your IT questions? PacktLib is Packt's online digital book library. Here, you can access, read and search across Packt's entire library of books.

### Why subscribe?

- Fully searchable across every book published by Packt
- Copy and paste, print and bookmark content
- On demand and accessible via web browser

### Free access for Packt account holders

If you have an account with Packt at [www.PacktPub.com](http://www.PacktPub.com), you can use this to access PacktLib today and view nine entirely free books. Simply use your login credentials for immediate access.

# Table of Contents

<b>Preface</b>	<b>1</b>
<b>Chapter 1: Getting Started with Python Libraries</b>	<b>9</b>
<b>Software used in this book</b>	<b>10</b>
Installing software and setup	10
On Windows	10
On Linux	12
On Mac OS X	13
<b>Building NumPy SciPy, matplotlib, and IPython from source</b>	<b>14</b>
<b>Installing with setuptools</b>	<b>15</b>
<b>NumPy arrays</b>	<b>16</b>
<b>A simple application</b>	<b>16</b>
<b>Using IPython as a shell</b>	<b>19</b>
<b>Reading manual pages</b>	<b>22</b>
<b>IPython notebooks</b>	<b>22</b>
<b>Where to find help and references</b>	<b>23</b>
<b>Summary</b>	<b>23</b>
<b>Chapter 2: NumPy Arrays</b>	<b>25</b>
<b>The NumPy array object</b>	<b>25</b>
The advantages of NumPy arrays	26
<b>Creating a multidimensional array</b>	<b>27</b>
<b>Selecting NumPy array elements</b>	<b>27</b>
<b>NumPy numerical types</b>	<b>28</b>
Data type objects	30
Character codes	30
The dtype constructors	31
The dtype attributes	31

<b>One-dimensional slicing and indexing</b>	<b>32</b>
<b>Manipulating array shapes</b>	<b>32</b>
Stacking arrays	35
Splitting NumPy arrays	39
NumPy array attributes	41
Converting arrays	48
<b>Creating array views and copies</b>	<b>48</b>
<b>Fancy indexing</b>	<b>50</b>
<b>Indexing with a list of locations</b>	<b>52</b>
<b>Indexing NumPy arrays with Booleans</b>	<b>53</b>
<b>Broadcasting NumPy arrays</b>	<b>55</b>
<b>Summary</b>	<b>58</b>
<b>Chapter 3: Statistics and Linear Algebra</b>	<b>59</b>
<hr/>	
<b>NumPy and SciPy modules</b>	<b>59</b>
<b>Basic descriptive statistics with NumPy</b>	<b>63</b>
<b>Linear algebra with NumPy</b>	<b>66</b>
Inverting matrices with NumPy	66
Solving linear systems with NumPy	68
<b>Finding eigenvalues and eigenvectors with NumPy</b>	<b>69</b>
<b>NumPy random numbers</b>	<b>71</b>
Gambling with the binomial distribution	72
Sampling the normal distribution	74
Performing a normality test with SciPy	75
<b>Creating a NumPy-masked array</b>	<b>78</b>
Disregarding negative and extreme values	80
<b>Summary</b>	<b>83</b>
<b>Chapter 4: pandas Primer</b>	<b>85</b>
<hr/>	
<b>Installing and exploring pandas</b>	<b>86</b>
<b>pandas DataFrames</b>	<b>87</b>
<b>pandas Series</b>	<b>90</b>
<b>Querying data in pandas</b>	<b>94</b>
<b>Statistics with pandas DataFrames</b>	<b>97</b>
<b>Data aggregation with pandas DataFrames</b>	<b>99</b>
<b>Concatenating and appending DataFrames</b>	<b>103</b>
<b>Joining DataFrames</b>	<b>105</b>
<b>Handling missing values</b>	<b>108</b>
<b>Dealing with dates</b>	<b>110</b>
<b>Pivot tables</b>	<b>113</b>
<b>Remote data access</b>	<b>114</b>
<b>Summary</b>	<b>117</b>

---

<b>Chapter 5: Retrieving, Processing, and Storing Data</b>	<b>119</b>
Writing CSV files with NumPy and pandas	120
Comparing the NumPy .npy binary format and pickling pandas DataFrames	122
Storing data with PyTables	124
Reading and writing pandas DataFrames to HDF5 stores	126
Reading and writing to Excel with pandas	129
Using REST web services and JSON	131
Reading and writing JSON with pandas	132
Parsing RSS and Atom feeds	134
Parsing HTML with BeautifulSoup	135
Summary	142
<b>Chapter 6: Data Visualization</b>	<b>143</b>
matplotlib subpackages	144
Basic matplotlib plots	144
Logarithmic plots	146
Scatter plots	148
Legends and annotations	150
Three-dimensional plots	153
Plotting in pandas	155
Lag plots	158
Autocorrelation plots	159
Plot.ly	160
Summary	163
<b>Chapter 7: Signal Processing and Time Series</b>	<b>165</b>
statsmodels subpackages	166
Moving averages	167
Window functions	168
Defining cointegration	170
Autocorrelation	173
Autoregressive models	176
ARMA models	179
Generating periodic signals	181
Fourier analysis	184
Spectral analysis	186
Filtering	187
Summary	189
<b>Chapter 8: Working with Databases</b>	<b>191</b>
Lightweight access with sqlite3	192
Accessing databases from pandas	194

---

<b>SQLAlchemy</b>	<b>196</b>
Installing and setting up SQLAlchemy	196
Populating a database with SQLAlchemy	198
Querying the database with SQLAlchemy	200
<b>Pony ORM</b>	<b>201</b>
<b>Dataset – databases for lazy people</b>	<b>202</b>
<b>PyMongo and MongoDB</b>	<b>204</b>
<b>Storing data in Redis</b>	<b>206</b>
<b>Apache Cassandra</b>	<b>207</b>
<b>Summary</b>	<b>210</b>
<b>Chapter 9: Analyzing Textual Data and Social Media</b>	<b>211</b>
<b>Installing NLTK</b>	<b>212</b>
<b>Filtering out stopwords, names, and numbers</b>	<b>214</b>
<b>The bag-of-words model</b>	<b>216</b>
<b>Analyzing word frequencies</b>	<b>217</b>
<b>Naive Bayes classification</b>	<b>219</b>
<b>Sentiment analysis</b>	<b>222</b>
<b>Creating word clouds</b>	<b>225</b>
<b>Social network analysis</b>	<b>230</b>
<b>Summary</b>	<b>232</b>
<b>Chapter 10: Predictive Analytics and Machine Learning</b>	<b>233</b>
<b>A tour of scikit-learn</b>	<b>235</b>
<b>Preprocessing</b>	<b>236</b>
<b>Classification with logistic regression</b>	<b>238</b>
<b>Classification with support vector machines</b>	<b>240</b>
<b>Regression with ElasticNetCV</b>	<b>242</b>
<b>Support vector regression</b>	<b>245</b>
<b>Clustering with affinity propagation</b>	<b>248</b>
<b>Mean Shift</b>	<b>250</b>
<b>Genetic algorithms</b>	<b>252</b>
<b>Neural networks</b>	<b>257</b>
<b>Decision trees</b>	<b>259</b>
<b>Summary</b>	<b>261</b>
<b>Chapter 11: Environments Outside the Python Ecosystem and Cloud Computing</b>	<b>263</b>
<b>Exchanging information with MATLAB/Octave</b>	<b>264</b>
<b>Installing rpy2</b>	<b>265</b>
<b>Interfacing with R</b>	<b>265</b>
<b>Sending NumPy arrays to Java</b>	<b>268</b>
<b>Integrating SWIG and NumPy</b>	<b>269</b>

---

<b>Integrating Boost and Python</b>	<b>272</b>
<b>Using Fortran code through f2py</b>	<b>274</b>
<b>Setting up Google App Engine</b>	<b>275</b>
<b>Running programs on PythonAnywhere</b>	<b>276</b>
<b>Working with Wakari</b>	<b>277</b>
<b>Summary</b>	<b>278</b>
<b>Chapter 12: Performance Tuning, Profiling, and Concurrency</b>	<b>279</b>
<b>Profiling the code</b>	<b>280</b>
<b>Installing Cython</b>	<b>284</b>
<b>Calling C code</b>	<b>288</b>
<b>Creating a process pool with multiprocessing</b>	<b>290</b>
<b>Speeding up embarrassingly parallel for loops with Joblib</b>	<b>293</b>
<b>Comparing Bottleneck to NumPy functions</b>	<b>294</b>
<b>Performing MapReduce with Jug</b>	<b>296</b>
<b>Installing MPI for Python</b>	<b>298</b>
<b>IPython Parallel</b>	<b>299</b>
<b>Summary</b>	<b>303</b>
<b>Appendix A: Key Concepts</b>	<b>305</b>
<b>Appendix B: Useful Functions</b>	<b>311</b>
<b>matplotlib</b>	<b>311</b>
<b>NumPy</b>	<b>312</b>
<b>pandas</b>	<b>313</b>
<b>Scikit-learn</b>	<b>314</b>
<b>SciPy</b>	<b>315</b>
<b>scipy.fftpack</b>	<b>315</b>
<b>scipy.signal</b>	<b>315</b>
<b>scipy.stats</b>	<b>315</b>
<b>Appendix C: Online Resources</b>	<b>317</b>
<b>Index</b>	<b>319</b>

---



# Preface

"Data analysis is Python's killer app."

- Unknown

Data analysis has a rich history in the natural, biomedical, and social sciences. You may have heard of *Big Data*. Although, it's hard to give a precise definition of Big Data, we should be aware of its impact on data analysis efforts. Currently, we have the following trends associated with Big Data:

- The world's population continues to grow
- More and more data is collected and stored
- The number of transistors that can be put on a computer chip cannot grow indefinitely
- Governments, scientists, industry, and individuals have a growing need to learn from data

Data analysis has gained popularity lately due to the hype around *Data Science*. Data analysis and Data Science attempt to extract information from data. For that purpose, we use techniques from statistics, machine learning, signal processing, natural language processing, and computer science.

A mind map visualizing Python software that can be used for data analysis can be found at <http://www.xmind.net/m/WvfC/>. The first thing that we should notice is that the Python ecosystem is very mature. It includes famous packages such as NumPy, SciPy, and matplotlib. This should not come as a surprise since Python has been around since 1989. Python is easy to learn and use, less verbose than other programming languages, and very readable. Even if you don't know Python, you can pick up the basics within days, especially if you have experience in another programming language. To enjoy this book, you don't need more than the basics. There are plenty of books, courses, and online tutorials that teach Python.



## What this book covers

This book starts as a tutorial on NumPy, SciPy, matplotlib, and pandas. These are open source Python packages useful for numerical work, data wrangling, and visualization. Combined, they can compete with MATLAB, Mathematica, and R. The second half of the book teaches more advanced topics such as signal processing, databases, text analysis, machine learning, interoperability, and performance tuning.

*Chapter 1, Getting Started with Python Libraries*, guides us to achieve a successful installation of the numerical Python software and set it up step by step. Also, we will create a small application.

*Chapter 2, NumPy Arrays*, introduces us to NumPy fundamentals and arrays. By the end of this chapter, we will have basic understanding of NumPy arrays and the associated functions.

*Chapter 3, Statistics and Linear Algebra*, gives a quick overview of linear algebra and statistical functions.

*Chapter 4, pandas Primer*, provides a tutorial on basic pandas functionality where we learn about pandas data structures and operations.

*Chapter 5, Retrieving, Processing, and Storing Data*, explains how to acquire data in various formats and how to clean raw data and store it.

*Chapter 6, Data Visualization*, teaches how to plot data with matplotlib.

*Chapter 7, Signal Processing and Time Series*, contains time series and signal processing examples using sunspot cycles data. The examples mostly use NumPy/SciPy, along with statsmodels in at least one example.

*Chapter 8, Working with Databases*, provides information about various databases (relational and NoSQL) and related APIs.

*Chapter 9, Analyzing Textual Data and Social Media*, analyzes texts for sentiment analysis and topics extraction. A small example is also given of network analysis.

*Chapter 10, Predictive Analytics and Machine Learning*, explains artificial intelligence with weather prediction as a running example and mostly uses scikit-learn. However, some machine learning algorithms are not covered by scikit-learn, so for those, we use other APIs.

*Chapter 11, Environments Outside the Python Ecosystem and Cloud Computing*, gives various examples on how to integrate existing code not written in Python. Also, setup in the Cloud will be demonstrated.