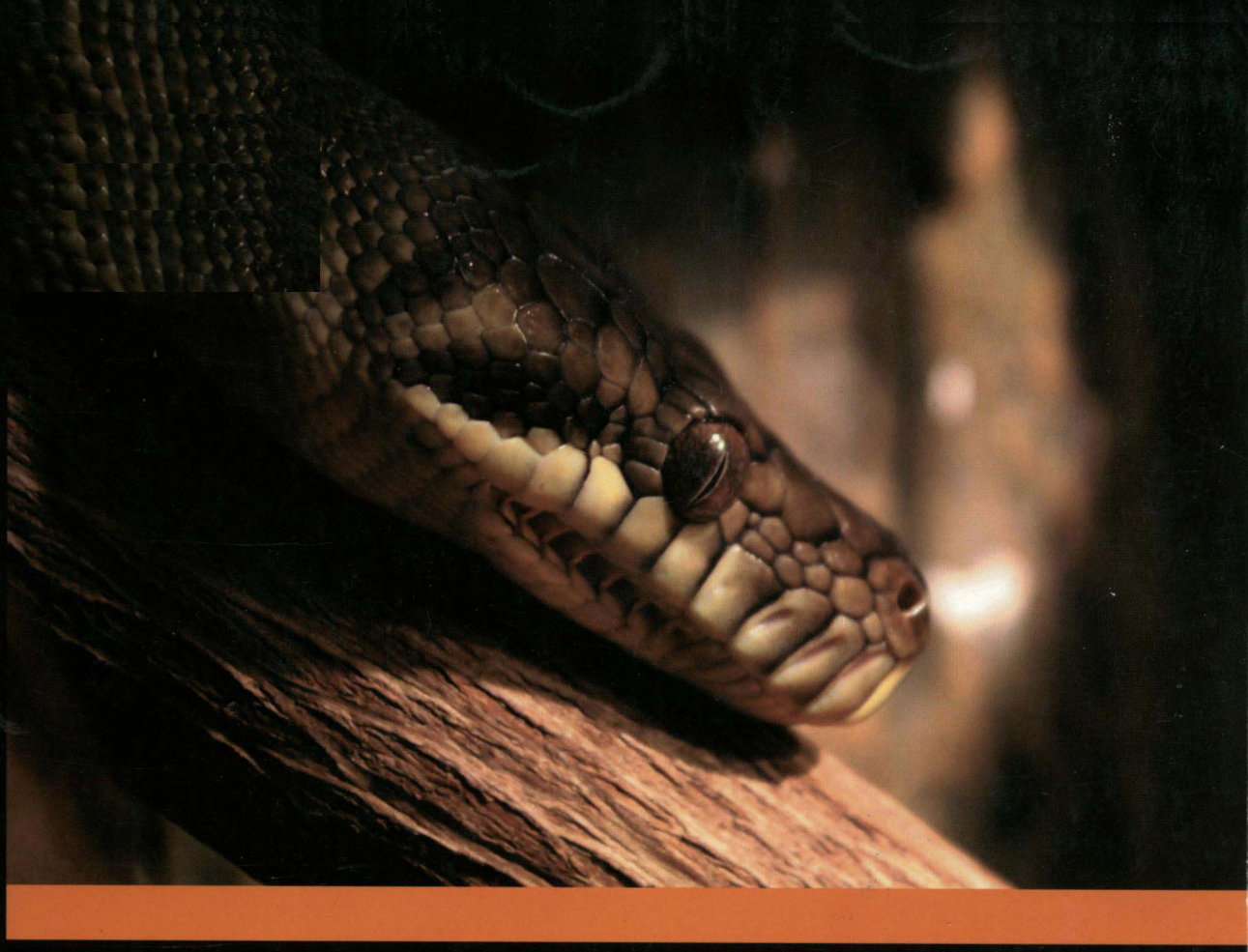

Python数据分析（影印版）

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

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
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`.

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`), 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` 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` 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` for more details.

At `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`, 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

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.