Data Analysis and Visualization Using Python

Analyze Data to Create Visualizations for BI Systems

Dr. Ossama Embarak

Apress®

Data Analysis and Visualization Using Python

Analyze Data to Create Visualizations for BI Systems

Dr. Ossama Embarak

Apress®

Data Analysis and Visualization Using Python

Dr. Ossama Embarak
Higher Colleges of Technology, Abu Dhabi, United Arab Emirates

ISBN-13 (pbk): 978-1-4842-4108-0 ISBN-13 (electronic): 978-1-4842-4109-7
https://doi.org/10.1007/978-1-4842-4109-7

Library of Congress Control Number: 2018964118

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Managing Director, Apress Media LLC: Welmoed Spahr
Acquisitions Editor: Nikhil Karkal
Development Editor: Matthew Moodie
Coordinating Editor: Divya Modi

Cover designed by eStudioCalamar

Cover image designed by Freepik (www.freepik.com)

Distributed to the book trade worldwide by Springer Science+Business Media New York, 233 Spring Street, 6th Floor, New York, NY 10013. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a **Delaware** corporation.

For information on translations, please e-mail rights@apress.com, or visit www.apress.com/rights-permissions.

Apress titles may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Print and eBook Bulk Sales web page at www.apress.com/bulk-sales.

Any source code or other supplementary material referenced by the author in this book is available to readers on GitHub via the book's product page, located at www.apress.com/978-1-4842-4108-0. For more detailed information, please visit www.apress.com/source-code.

Printed on acid-free paper

This book is dedicated to my family—my mother,
my father, and all my brothers—for their endless support.

Table of Contents

About the Author

Dr. Ossama Embarak holds a PhD in computer and information science from Heriot-Watt University in Scotland, UK. He has more than two decades of research and teaching experience with a number of programming languages including C++, Java, C#, R, Python, etc. He is currently the lead CIS program coordinator for Higher Colleges of Technology, UAE's largest applied higher educational institution, with more than 23,000 students attending campuses throughout the region.

Recently, he received an interdisciplinary research grant of 199,000 to implement a machine learning system for mining students' knowledge and skills.

He has participated in many scholarly activities as a reviewer and editor for journals in the fields of computer and information science including artificial intelligence, data mining, machine learning, mobile and web technologies. He supervised a large number of graduation projects, as well as he has published numerous papers about data mining, users online privacy, semantic web structure and knowledge discovery. Also he participated as a co-chair for numerous regional and international conferences.

About the Technical Reviewers

Shankar Rao Pandala is a data scientist at Cognizant. He has a bachelor's degree in computer science and a master's degree in financial markets. His work experience spans finance, healthcare, manufacturing, and consulting. His area of interest is artificial intelligence for trading.

Prashant Sahu has a bachelor's of technology from NIT Rourkela (2003) and is currently pursuing a doctorate from the Indian Institute of Technology, Bombay, in the area of instrumentation, data analytics, modeling, and simulation applied to semiconductor materials and devices. He is currently the head of training services at Tech Smart Systems in Pune, India. He is also mentoring the startup Bharati Robotic Systems (India) as an SVP of innovation. He has more than 15 years of experience in research, automation, simulation and modeling, data analytics, image processing, control systems, optimization algorithms, genetic algorithms, cryptography, and more, and he has handled many projects in these areas from academia and industry. He has conducted several faculty development training programs across India and has conducted corporate training for software companies across India. He is also an external examiner for B.E./M.E. projects and a member of the Syllabus Revision Committee at the University of Pune.

Introduction

This book looks at Python from a data science point of view and teaches the reader proven techniques of data visualization that are used to make critical business decisions. Starting with an introduction to data science using Python, the book then covers the Python environment and gets you acquainted with editors like Jupyter Notebooks and the Spyder IDE. After going through a primer on Python programming, you will grasp the fundamental Python programming techniques used in data science. Moving on to data visualization, you will learn how it caters to modern business needs and is key to decision-making. You will also take a look at some popular data visualization libraries in Python. Shifting focus to collecting data, you will learn about the various aspects of data collections from a data science perspective and also take a look at Python's data collection structures. You will then learn about file I/O processing and regular expressions in Python, followed by techniques to gather and clean data. Moving on to exploring and analyzing data, you will look at the various data structures in Python. Then, you will take a deep dive into data visualization techniques, going through a number of plotting systems in Python. In conclusion, you will go through two detailed case studies, where you'll get a chance to revisit the concepts you've grasped so far.

This book is for people who want to learn Python for the data science field in order to become data scientists. No specific programming prerequisites are required besides having basic programming knowledge.

Specifically, the following list highlights what is covered in the book:

- Chapter 1 introduces the main concepts of data science and its life cycle. It also demonstrates the importance of Python programming and its main libraries for data science processing. You will learn how different Python data structures are used in data science applications. You will learn how to implement an abstract series and a data frame as a main Python data structure. You will learn how to apply basic Python programming techniques for data cleaning and manipulation. You will learn how to run the basic inferential statistical analyses. In addition, exercises with model answers are given for practicing real-life scenarios.
- Chapter 2 demonstrates how to implement data visualization in modern business. You will learn how to recognize the role of data visualization in decision-making and how to load and use important Python libraries for data visualization. In addition, exercises with model answers are given for practicing real-life scenarios.
- Chapter 3 illustrates data collection structures in Python and their implementations. You will learn how to identify different forms of collection in Python. You will learn how to create lists and how to manipulate list content. You will learn about the purpose of creating a dictionary as a data container and its manipulations. You will learn how to maintain data in a tuple form and what the differences are between tuple structures and dictionary structures, as well as the basic tuples operations. You will learn how to create a series from

other data collection forms. You will learn how to create a data frame from different data collection structures and from another data frame. You will learn how to create a panel as a 3D data collection from a series or data frame. In addition, exercises with model answers are given for practicing real-life scenarios.

- Chapter 4 shows how to read and send data to users, read and pull data stored in historical files, and open files for reading, writing, or for both. You will learn how to access file attributes and manipulate sessions. You will learn how to read data from users and apply casting. You will learn how to apply regular expressions to extract data, use regular expression alternatives, and use anchors and repetition expressions for data extractions as well. In addition, exercises with model answers are given for practicing real-life scenarios.
- Chapter 5 covers data gathering and cleaning to have reliable data for analysis. You will learn how to apply data cleaning techniques to handle missing values. You will learn how to read CSV data format offline or pull it directly from online clouds. You will learn how to merge and integrate data from different sources. You will learn how to read and extract data from the JSON, HTML, and XML formats. In addition, exercises with model answers are given for practicing real-life scenarios.
- Chapter 6 shows how to use Python scripts to explore and analyze data in different collection structures. You will learn how to implement Python techniques to explore and analyze a series of data, create a series,

access data from a series with a position, and apply statistical methods on a series. You will learn how to explore and analyze data in a data frame, create a data frame, and update and access data in a data frame structure. You will learn how to manipulate data in a data frame such as including columns, selecting rows, adding, or deleting data, and applying statistical operations on a data frame. You will learn how to apply statistical methods on a panel data structure to explore and analyze stored data. You will learn how to statistically analyze grouped data, iterate through groups, and apply aggregations, transformations, and filtration techniques. In addition, exercises with model answers are given for practicing real-life scenarios.

- Chapter 7 shows how to visualize data from different collection structures. You will learn how to plot data from a series, a data frame, or a panel using Python plotting tools such as line plots, bar plots, pie charts, box plots, histograms, and scatter plots. You will learn how to implement the Seaborn plotting system using strip plots, box plots, swarm plots, and joint plots. You will learn how to implement Matplotlib plotting using line plots, bar charts, histograms, scatter plots, stack plots, and pie charts. In addition, exercises with model answers are given for practicing real-life scenarios.
- Chapter 8 investigates two real-life case studies, starting with data gathering and moving through cleaning, data exploring, data analysis, and visualizing. Finally, you'll learn how to discuss the study findings and provide recommendations for decision-makers.