



# 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

Copyright © 2018 by Dr. Ossama Embarak

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

Trademarked names, logos, and images may appear in this book. Rather than use a trademark symbol with every occurrence of a trademarked name, logo, or image we use the names, logos, and images only in an editorial fashion and to the benefit of the trademark owner, with no intention of infringement of the trademark.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

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](http://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](mailto:orders-ny@springer-sbm.com), or visit [www.springeronline.com](http://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](mailto:rights@apress.com), or visit [www.apress.com/rights-permissions](http://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](http://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](http://www.apress.com/978-1-4842-4108-0). For more detailed information, please visit [www.apress.com/source-code](http://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

<b>About the Author .....</b>	<b>xiii</b>
<b>About the Technical Reviewers .....</b>	<b>xv</b>
<b>Introduction .....</b>	<b>xvii</b>
<b>Chapter 1: Introduction to Data Science with Python .....</b>	<b>1</b>
The Stages of Data Science .....	1
Why Python? .....	2
Basic Features of Python.....	3
Python Learning Resources.....	4
Python Environment and Editors.....	6
Portable Python Editors (No Installation Required).....	6
Azure Notebooks .....	8
Offline and Desktop Python Editors.....	13
The Basics of Python Programming.....	13
Basic Syntax.....	14
Declaring Variables and Assigning Values .....	19
Basic Operators in Python .....	22
Python Comments .....	25
Formatting Strings.....	25
Conversion Types.....	26
The Replacement Field, {}.....	27
The Date and Time Module.....	28

## TABLE OF CONTENTS

Time Module Methods .....	29
Python Calendar Module .....	30
Fundamental Python Programming Techniques.....	32
Selection Statements .....	32
Iteration Statements .....	35
The Use of Break, Continues, and Pass Statements.....	39
try and except.....	41
String Processing .....	42
Tabular Data and Data Formats .....	54
Python Pandas Data Science Library.....	55
Python Lambdas and the Numpy Library .....	60
Data Cleaning and Manipulation Techniques .....	64
Abstraction of the Series and Data Frame .....	64
Running Basic Inferential Analyses.....	69
Summary.....	74
Exercises and Answers .....	74
<b>Chapter 2: The Importance of Data Visualization in Business Intelligence .....</b>	<b>85</b>
Shifting from Input to Output .....	86
Why Is Data Visualization Important?.....	86
Why Do Modern Businesses Need Data Visualization? .....	87
The Future of Data Visualization .....	88
How Data Visualization Is Used for Business Decision-Making .....	89
Faster Responses .....	89
Simplicity.....	90
Easier Pattern Visualization .....	90
Team Involvement .....	90

## TABLE OF CONTENTS

Unify Interpretation.....	90
Introducing Data Visualization Techniques.....	92
Loading Libraries.....	93
Popular Libraries for Data Visualization in Python.....	94
Introducing Plots in Python.....	109
Summary.....	116
Exercises and Answers .....	117
<b>Chapter 3: Data Collection Structures .....</b>	<b>125</b>
Lists .....	125
Creating Lists.....	126
Accessing Values in Lists .....	126
Adding and Updating Lists.....	127
Deleting List Elements.....	128
Basic List Operations.....	129
Indexing, Slicing, and Matrices.....	130
Built-in List Functions and Methods.....	130
List Sorting and Traversing .....	133
Lists and Strings.....	134
Parsing Lines .....	135
Aliasing.....	136
Dictionaries .....	137
Creating Dictionaries .....	138
Updating and Accessing Values in Dictionaries.....	139
Deleting Dictionary Elements .....	141
Built-in Dictionary Functions .....	141
Built-in Dictionary Methods.....	143

## TABLE OF CONTENTS

Tuples.....	145
Creating Tuples .....	146
Concatenating Tuples .....	148
Accessing Values in Tuples .....	148
Basic Tuples Operations .....	150
Series.....	151
Creating a Series with index.....	151
Creating a Series from a Dictionary .....	154
Creating a Series from a Scalar Value .....	155
Vectorized Operations and Label Alignment with Series.....	156
Name Attribute .....	157
Data Frames.....	158
Creating Data Frames from a Dict of Series or Dicts.....	158
Creating Data Frames from a Dict of Ndarrays/Lists.....	160
Creating Data Frames from a Structured or Record Array.....	161
Creating Data Frames from a List of Dicts.....	161
Creating Data Frames from a Dict of Tuples.....	162
Selecting, Adding, and Deleting Data Frame Columns.....	163
Assigning New Columns in Method Chains .....	165
Indexing and Selecting Data Frames.....	166
Transposing a Data Frame.....	170
Data Frame Interoperability with Numpy Functions .....	171
Panels .....	172
Creating a Panel from a 3D Ndarray.....	172
Creating a Panel from a Dict of Data Frame Objects .....	173
Selecting, Adding, and Deleting Items .....	175
Summary.....	176
Exercises and Answers .....	177



<b>Chapter 4: File I/O Processing and Regular Expressions .....</b>	<b>183</b>
File I/O Processing .....	183
Data Input and Output .....	183
Opening and Closing Files .....	184
File Object Attributes .....	185
Reading and Writing to Files.....	186
Directories in Python .....	187
Regular Expressions .....	188
Regular Expression Patterns .....	188
Special Character Classes.....	197
Repetition Classes.....	198
Alternatives .....	198
Anchors .....	199
Summary.....	201
Exercises and Answer.....	202
<b>Chapter 5: Data Gathering and Cleaning.....</b>	<b>205</b>
Cleaning Data.....	206
Checking for Missing Values.....	207
Handling the Missing Values .....	209
Reading and Cleaning CSV Data .....	212
Merging and Integrating Data .....	218
Reading Data from the JSON Format.....	223
Reading Data from the HTML Format.....	226
Reading Data from the XML Format.....	233
Summary.....	235
Exercises and Answers .....	236

<b>Chapter 6: Data Exploring and Analysis .....</b>	<b>243</b>
Series Data Structures.....	243
Creating a Series .....	244
Accessing Data from a Series with a Position.....	246
Exploring and Analyzing a Series .....	248
Operations on a Series .....	251
Data Frame Data Structures .....	254
Creating a Data Frame.....	255
Updating and Accessing a Data Frame's Column Selection .....	258
Column Addition .....	259
Column Deletion .....	260
Row Selection.....	264
Row Addition .....	266
Row Deletion .....	267
Exploring and Analyzing a Data Frame.....	267
Panel Data Structures.....	273
Creating a Panel .....	273
Accessing Data from a Panel with a Position .....	274
Exploring and Analyzing a Panel.....	275
Data Analysis .....	277
Statistical Analysis .....	277
Data Grouping.....	282
Iterating Through Groups.....	283
Aggregations .....	284
Transformations.....	285
Filtration .....	286
Summary.....	287
Exercises and Answers .....	288

<b>Chapter 7: Data Visualization .....</b>	<b>293</b>
Direct Plotting .....	294
Line Plot.....	295
Bar Plot.....	298
Pie Chart.....	300
Box Plot .....	301
Histogram Plot.....	303
Scatter Plot.....	303
Seaborn Plotting System.....	304
Strip Plot.....	305
Box Plot .....	309
Swarm Plot.....	313
Joint Plot .....	315
Matplotlib Plot.....	321
Line Plot.....	321
Bar Chart .....	324
Histogram Plot.....	326
Scatter Plot.....	330
Stack Plot .....	332
Pie Chart.....	334
Summary.....	335
Exercises and Answers .....	336

TABLE OF CONTENTS

**Chapter 8: Case Studies.....343**

    Case Study 1: Cause of Deaths in the United States (1999–2015)..... 343

        Data Gathering..... 343

        Data Analysis ..... 344

        Data Visualization ..... 349

        Findings..... 353

    Case Study 2: Analyzing Gun Deaths in the United States (2012–2014)..... 354

        Data Gathering..... 355

        Data Analysis ..... 356

        Data Visualization ..... 357

        Findings..... 364

    Summary..... 366

**Index.....367**

# 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

## ABOUT THE TECHNICAL REVIEWERS

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.