Table of Contents

About the Author xiii

About the Technical Reviewer xv

Acknowledgments xvii

Introduction xix

Chapter 1: Getting Familiar with Python 1

Technical requirements 1

Getting started with Jupyter notebooks 2

Shortcuts and other features in Jupyter 5

Tab Completion 7

Magic commands used in Jupyter 7

Python Basics 8

Comments, print, and input 8

Variables and Constants 11

Operators 12

Data types 15

Working with Strings 20

Conditional statements 25

Loops 26

Functions 29

Syntax errors and exceptions 31

Working with files 32

Reading from a file 33

Writing to a file 34

Modules in Python 35

Python Enhancement Proposal (PEP) 8 – standards for writing code 36
Summary 38
Review Exercises 39

Chapter 2: Exploring Containers, Classes, and Objects 45

Containers 45
Lists 45
Tuples 56
Dictionaries 59
Sets 63
Object-oriented programming 65
Object-oriented programming principles 67
Summary 70
Review Exercises 71

Chapter 3: Regular Expressions and Math with Python 77

Regular expressions 77
Steps for solving problems with regular expressions 77
Python functions for regular expressions 79
Using Sympy for math problems 86
Factorization of an algebraic expression 86
Solving algebraic equations (for one variable) 87
Solving simultaneous equations (for two variables) 87
Solving expressions entered by the user 88
Solving simultaneous equations graphically 89
Creating and manipulating sets 90
Union and intersection of sets 90
Finding the probability of an event 91
Solving questions in calculus 92
Summary 94
Review Exercises 95

Chapter 4: Descriptive Data Analysis Basics 101

Descriptive data analysis - Steps 101

Structure of data 104

Classifying data into different levels 104

Visualizing various levels of data 106

Plotting mixed data 110

Summary 113

Review Exercises 113

Chapter 5: Working with NumPy Arrays 117

Getting familiar with arrays and NumPy functions 117

Creating an array 118

Reshaping an array 121

Combining arrays 125

Testing for conditions 127

Broadcasting, vectorization, and arithmetic operations 130

Obtaining the properties of an array 133

Slicing or selecting a subset of data 136

Obtaining descriptive statistics/aggregate measures 138

Matrices 140

Summary 140

Review Exercises 141

Chapter 6: Prepping Your Data with Pandas 147

Pandas at a glance 147

Technical requirements 149

Building blocks of Pandas 149

Examining the properties of a Series 152

DataFrames 156

Creating DataFrames by importing data from other formats 158

Accessing attributes in a DataFrame 160

Modifying DataFrame objects 161

Indexing 169
Type of an index object 170
Creating a custom index and using columns as indexes 171
Indexes and speed of data retrieval 173
Immutability of an index 174
Alignment of indexes 176
Set operations on indexes 177
Data types in Pandas 178
Obtaining information about data types 179
Indexers and selection of subsets of data 182
Understanding loc and iloc indexers 183
Other (less commonly used) indexers for data access 188
Boolean indexing for selecting subsets of data 192
Using the query method to retrieve data 192
Operators in Pandas 193
Representing dates and times in Pandas 194
Converting strings into Pandas Timestamp objects 195
Extracting the components of a Timestamp object 196
Grouping and aggregation 197
Examining the properties of the groupby object 199
Filtering groups 201
Transform method and groupby 202
Apply method and groupby 204
How to combine objects in Pandas 204
Append method for adding rows 205
Concat function (adding rows or columns from other objects) 207
Join method – index to index 210
Merge method – SQL type join based on common columns 211

Restructuring data and dealing with anomalies 213
Dealing with missing data 214
Data duplication 218
Tidy data and techniques for restructuring data 220
Conversion from wide to long format (tidy data) 221
Stack method (wide-to-long format conversion) 223
Melt method (wide-to-long format conversion) 226
Pivot method (long-to-wide conversion) 228
Summary 229
Review Exercises 230

Chapter 7: Data Visualization with Python Libraries 243

Technical requirements 243
External files 244
Commonly used plots 245
Matplotlib 248
Approach for plotting using Matplotlib 251
Plotting using Pandas 253
Scatter plot 254
Histogram 255
Pie charts 256
Seaborn library 257
Box plots 258
Adding arguments to any Seaborn plotting function 259
Kernel density estimate 259
Violin plot 260
Count plots 261
Heatmap 262
Facet grid 263
Regplot 265

Implot 266
Strip plot 267
Swarm plot 268
Catplot 269
Pair plot 270
Joint plot 272
Summary 273
Review Exercises 274

Chapter 8: Data Analysis Case Studies 279

Technical requirements 279
Methodology 280
Case study 8-1: Highest grossing movies in France – analyzing unstructured data 281
Case study 8-2: Use of data analysis for air quality management 288
Case study 8-3: Worldwide COVID-19 cases – an analysis 308
Summary 320
Review Exercises 321

Chapter 9: Statistics and Probability with Python 325

Permutations and combinations 325
Probability 327
Rules of probability 328
Conditional probability 330
Bayes theorem 330
Application of Bayes theorem in medical diagnostics 331
Another application of Bayes theorem: Email spam classification 333
SciPy library 334
Probability distributions 335
Binomial distribution 335
Poisson distribution 338
Continuous probability distributions 341

Normal distribution 341
Standard normal distribution 343
Measures of central tendency 347
Measures of dispersion 348
Measures of shape 349
Sampling 353
Probability sampling 353
Non-probability sampling 354
Central limit theorem 355
Estimates and confidence intervals 356
Types of errors in sampling 357
Hypothesis testing 358
Basic concepts in hypothesis testing 358
Key terminology used in hypothesis testing 359
Steps involved in hypothesis testing 361
One-sample z-test 362
Two-sample sample z-test 364
Hypothesis tests with proportions 366
Two-sample z-test for the population proportions 368
T-distribution 370
One sample t-test 372
Two-sample t-test 372
Two-sample t-test for paired samples 373
Solved examples: Conducting t-tests using Scipy functions 373
ANOVA 376
Chi-square test of association 379
Summary 383
Review Exercises 386

Index 393