Table of Contents

Preface xi

1. Preliminaries 1
1.1 What Is This Book About? 1
What Kinds of Data? 1
1.2 Why Python for Data Analysis? 2
Python as Glue 3
Solving the "Two-Language" Problem 3
Why Not Python? 3
1.3 Essential Python Libraries 4
NumPy 4
pandas 5
matplotlib 6
IPython and Jupyter 6
SciPy 7
scikit-learn 8
statsmodels 8
Other Packages 9
1.4 Installation and Setup 9
Miniconda on Windows 9
GNU/Linux 10
Miniconda on macOS 11
Installing Necessary Packages 11
Integrated Development Environments and Text Editors 12
1.5 Community and Conferences 13
1.6 Navigating This Book 14
Code Examples 15

Data for Examples
Import Conventions

2. Python Language Basics, IPython, and Jupyter Notebooks
2.1 The Python Interpreter
2.2 IPython Basics
Running the IPython Shell
Running the Jupyter Notebook
Tab Completion
Introspection
2.3 Python Language Basics
Language Semantics
Scalar Types
Control Flow
2.4 Conclusion

3. Built-In Data Structures, Functions, and Files
3.1 Data Structures and Sequences
Tuple
List
Dictionary
Set
Built-In Sequence Functions
List, Set, and Dictionary Comprehensions
3.2 Functions
Namespaces, Scope, and Local Functions
Returning Multiple Values
Functions Are Objects
Anonymous (Lambda) Functions
Generators
Errors and Exception Handling
3.3 Files and the Operating System
Bytes and Unicode with Files
3.4 Conclusion

4. NumPy Basics: Arrays and Vectorized Computation
4.1 The NumPy ndarray: A Multidimensional Array Object
Creating ndarrays
Data Types for ndarrays
Arithmetic with NumPy Arrays
Basic Indexing and Slicing

Boolean Indexing 97
Fancy Indexing 100
Transposing Arrays and Swapping Axes 102
4.2 Pseudorandom Number Generation 103
4.3 Universal Functions: Fast Element-Wise Array Functions 105
4.4 Array-Oriented Programming with Arrays 108
Expressing Conditional Logic as Array Operations 110
Mathematical and Statistical Methods 111
Methods for Boolean Arrays 113
Sorting 114
Unique and Other Set Logic 115
4.5 File Input and Output with Arrays 116
4.6 Linear Algebra 116
4.7 Example: Random Walks 118
Simulating Many Random Walks at Once 120
4.8 Conclusion 121

5. Getting Started with pandas..... 123
5.1 Introduction to pandas Data Structures 124
Series 124
DataFrame 129
Index Objects 136
5.2 Essential Functionality 138
Reindexing 138
Dropping Entries from an Axis 141
Indexing, Selection, and Filtering 142
Arithmetic and Data Alignment 152
Function Application and Mapping 158
Sorting and Ranking 160
Axis Indexes with Duplicate Labels 164
5.3 Summarizing and Computing Descriptive Statistics 165
Correlation and Covariance 168
Unique Values, Value Counts, and Membership 170
5.4 Conclusion 173

6. Data Loading, Storage, and File Formats..... 175
6.1 Reading and Writing Data in Text Format 175
Reading Text Files in Pieces 182
Writing Data to Text Format 184
Working with Other Delimited Formats 185
JSON Data 187

XML and HTML: Web Scraping
6.2 Binary Data Formats
Reading Microsoft Excel Files
Using HDF5 Format
6.3 Interacting with Web APIs
6.4 Interacting with Databases
6.5 Conclusion

7. Data Cleaning and Preparation.....
7.1 Handling Missing Data
Filtering Out Missing Data
Filling In Missing Data
7.2 Data Transformation
Removing Duplicates
Transforming Data Using a Function or Mapping
Replacing Values
Renaming Axis Indexes
Discretization and Binning
Detecting and Filtering Outliers
Permutation and Random Sampling
Computing Indicator/Dummy Variables
7.3 Extension Data Types
7.4 String Manipulation
Python Built-In String Object Methods
Regular Expressions
String Functions in pandas
7.5 Categorical Data
Background and Motivation
Categorical Extension Type in pandas
Computations with Categoricals
Categorical Methods
7.6 Conclusion

8. Data Wrangling: Join, Combine, and Reshape.....
8.1 Hierarchical Indexing
Reordering and Sorting Levels
Summary Statistics by Level
Indexing with a DataFrame's columns
8.2 Combining and Merging Datasets
Database-Style DataFrame Joins
Merging on Index

Concatenating Along an Axis 263
Combining Data with Overlap 268
8.3 Reshaping and Pivoting 270
Reshaping with Hierarchical Indexing 270
Pivoting "Long" to "Wide" Format 273
Pivoting "Wide" to "Long" Format 277
8.4 Conclusion 279

9. Plotting and Visualization..... 281
9.1 A Brief matplotlib API Primer 282
Figures and Subplots 283
Colors, Markers, and Line Styles 288
Ticks, Labels, and Legends 290
Annotations and Drawing on a Subplot 294
Saving Plots to File 296
matplotlib Configuration 297
9.2 Plotting with pandas and seaborn 298
Line Plots 298
Bar Plots 301
Histograms and Density Plots 309
Scatter or Point Plots 311
Facet Grids and Categorical Data 314
9.3 Other Python Visualization Tools 317
9.4 Conclusion 317

10. Data Aggregation and Group Operations..... 319
10.1 How to Think About Group Operations 320
Iterating over Groups 324
Selecting a Column or Subset of Columns 326
Grouping with Dictionaries and Series 327
Grouping with Functions 328
Grouping by Index Levels 328
10.2 Data Aggregation 329
Column-Wise and Multiple Function Application 331
Returning Aggregated Data Without Row Indexes 335
10.3 Apply: General split-apply-combine 335
Suppressing the Group Keys 338
Quantile and Bucket Analysis 338
Example: Filling Missing Values with Group-Specific Values 340
Example: Random Sampling and Permutation 343
Example: Group Weighted Average and Correlation 344

Example: Group-Wise Linear Regression
10.4 Group Transforms and "Unwrapped" GroupBys
10.5 Pivot Tables and Cross-Tabulation
Cross-Tabulations: Crosstab
10.6 Conclusion

11. Time Series
11.1 Date and Time Data Types and Tools
Converting Between String and Datetime
11.2 Time Series Basics
Indexing, Selection, Subsetting
Time Series with Duplicate Indices
11.3 Date Ranges, Frequencies, and Shifting
Generating Date Ranges
Frequencies and Date Offsets
Shifting (Leading and Lagging) Data
11.4 Time Zone Handling
Time Zone Localization and Conversion
Operations with Time Zone-Aware Timestamp Objects
Operations Between Different Time Zones
11.5 Periods and Period Arithmetic
Period Frequency Conversion
Quarterly Period Frequencies
Converting Timestamps to Periods (and Back)
Creating a PeriodIndex from Arrays
11.6 Resampling and Frequency Conversion
Downsampling
Upsampling and Interpolation
Resampling with Periods
Grouped Time Resampling
11.7 Moving Window Functions
Exponentially Weighted Functions
Binary Moving Window Functions
User-Defined Moving Window Functions
11.8 Conclusion

12. Introduction to Modeling Libraries in Python
12.1 Interfacing Between pandas and Model Code
12.2 Creating Model Descriptions with Patsy
Data Transformations in Patsy Formulas
Categorical Data and Patsy

12.3 Introduction to statsmodels 415
Estimating Linear Models 415
Estimating Time Series Processes 419
12.4 Introduction to scikit-learn 420
12.5 Conclusion 423

13. Data Analysis Examples..... 425
13.1 Bitly Data from 1.USA.gov 425
Counting Time Zones in Pure Python 426
Counting Time Zones with pandas 428
13.2 MovieLens 1M Dataset 435
Measuring Rating Disagreement 439
13.3 US Baby Names 1880–2010 443
Analyzing Naming Trends 448
13.4 USDA Food Database 457
13.5 2012 Federal Election Commission Database 463
Donation Statistics by Occupation and Employer 466
Bucketing Donation Amounts 469
Donation Statistics by State 471
13.6 Conclusion 472

A. Advanced NumPy..... 473
A.1 ndarray Object Internals 473
NumPy Data Type Hierarchy 474
A.2 Advanced Array Manipulation 476
Reshaping Arrays 476
C Versus FORTRAN Order 478
Concatenating and Splitting Arrays 479
Repeating Elements: tile and repeat 481
Fancy Indexing Equivalents: take and put 483
A.3 Broadcasting 484
Broadcasting over Other Axes 487
Setting Array Values by Broadcasting 489
A.4 Advanced ufunc Usage 490
ufunc Instance Methods 490
Writing New ufuncs in Python 493
A.5 Structured and Record Arrays 493
Nested Data Types and Multidimensional Fields 494
Why Use Structured Arrays? 495
A.6 More About Sorting 495
Indirect Sorts: argsort and lexsort 497

Alternative Sort Algorithms
Partially Sorting Arrays
numpy.searchsorted: Finding Elements in a Sorted Array
A.7 Writing Fast NumPy Functions with Numba
Creating Custom numpy.ufunc Objects with Numba
A.8 Advanced Array Input and Output
Memory-Mapped Files
HDF5 and Other Array Storage Options
A.9 Performance Tips
The Importance of Contiguous Memory

B. More on the IPython System
B.1 Terminal Keyboard Shortcuts
B.2 About Magic Commands
The %run Command
Executing Code from the Clipboard
B.3 Using the Command History
Searching and Reusing the Command History
Input and Output Variables
B.4 Interacting with the Operating System
Shell Commands and Aliases
Directory Bookmark System
B.5 Software Development Tools
Interactive Debugger
Timing Code: %time and %timeit
Basic Profiling: %prun and %run -p
Profiling a Function Line by Line
B.6 Tips for Productive Code Development Using IPython
Reloading Module Dependencies
Code Design Tips
B.7 Advanced IPython Features
Profiles and Configuration
B.8 Conclusion

Index