
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	15
Import Conventions	16
2. Python Language Basics, IPython, and Jupyter Notebooks.....	17
2.1 The Python Interpreter	18
2.2 IPython Basics	19
Running the IPython Shell	19
Running the Jupyter Notebook	20
Tab Completion	23
Introspection	25
2.3 Python Language Basics	26
Language Semantics	26
Scalar Types	34
Control Flow	42
2.4 Conclusion	45
3. Built-In Data Structures, Functions, and Files.....	47
3.1 Data Structures and Sequences	47
Tuple	47
List	51
Dictionary	55
Set	59
Built-In Sequence Functions	62
List, Set, and Dictionary Comprehensions	63
3.2 Functions	65
Namespaces, Scope, and Local Functions	67
Returning Multiple Values	68
Functions Are Objects	69
Anonymous (Lambda) Functions	70
Generators	71
Errors and Exception Handling	74
3.3 Files and the Operating System	76
Bytes and Unicode with Files	80
3.4 Conclusion	82
4. NumPy Basics: Arrays and Vectorized Computation.....	83
4.1 The NumPy ndarray: A Multidimensional Array Object	85
Creating ndarrays	86
Data Types for ndarrays	88
Arithmetic with NumPy Arrays	91
Basic Indexing and Slicing	92

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	189
6.2 Binary Data Formats	193
Reading Microsoft Excel Files	194
Using HDF5 Format	195
6.3 Interacting with Web APIs	197
6.4 Interacting with Databases	199
6.5 Conclusion	201
7. Data Cleaning and Preparation.....	203
7.1 Handling Missing Data	203
Filtering Out Missing Data	205
Filling In Missing Data	207
7.2 Data Transformation	209
Removing Duplicates	209
Transforming Data Using a Function or Mapping	211
Replacing Values	212
Renaming Axis Indexes	214
Discretization and Binning	215
Detecting and Filtering Outliers	217
Permutation and Random Sampling	219
Computing Indicator/Dummy Variables	221
7.3 Extension Data Types	224
7.4 String Manipulation	227
Python Built-In String Object Methods	227
Regular Expressions	229
String Functions in pandas	232
7.5 Categorical Data	235
Background and Motivation	236
Categorical Extension Type in pandas	237
Computations with Categoricals	240
Categorical Methods	242
7.6 Conclusion	245
8. Data Wrangling: Join, Combine, and Reshape.....	247
8.1 Hierarchical Indexing	247
Reordering and Sorting Levels	250
Summary Statistics by Level	251
Indexing with a DataFrame's columns	252
8.2 Combining and Merging Datasets	253
Database-Style DataFrame Joins	254
Merging on Index	259

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	347
10.4 Group Transforms and “Unwrapped” GroupBys	347
10.5 Pivot Tables and Cross-Tabulation	351
Cross-Tabulations: Crosstab	354
10.6 Conclusion	355
11. Time Series.....	357
11.1 Date and Time Data Types and Tools	358
Converting Between String and Datetime	359
11.2 Time Series Basics	361
Indexing, Selection, Subsetting	363
Time Series with Duplicate Indices	365
11.3 Date Ranges, Frequencies, and Shifting	366
Generating Date Ranges	367
Frequencies and Date Offsets	370
Shifting (Leading and Lagging) Data	371
11.4 Time Zone Handling	374
Time Zone Localization and Conversion	375
Operations with Time Zone-Aware Timestamp Objects	377
Operations Between Different Time Zones	378
11.5 Periods and Period Arithmetic	379
Period Frequency Conversion	380
Quarterly Period Frequencies	382
Converting Timestamps to Periods (and Back)	384
Creating a PeriodIndex from Arrays	385
11.6 Resampling and Frequency Conversion	387
Downsampling	388
Upsampling and Interpolation	391
Resampling with Periods	392
Grouped Time Resampling	394
11.7 Moving Window Functions	396
Exponentially Weighted Functions	399
Binary Moving Window Functions	401
User-Defined Moving Window Functions	402
11.8 Conclusion	403
12. Introduction to Modeling Libraries in Python.....	405
12.1 Interfacing Between pandas and Model Code	405
12.2 Creating Model Descriptions with Patsy	408
Data Transformations in Patsy Formulas	410
Categorical Data and Patsy	412

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	498
Partially Sorting Arrays	499
numpy.searchsorted: Finding Elements in a Sorted Array	500
A.7 Writing Fast NumPy Functions with Numba	501
Creating Custom numpy.ufunc Objects with Numba	502
A.8 Advanced Array Input and Output	503
Memory-Mapped Files	503
HDF5 and Other Array Storage Options	504
A.9 Performance Tips	505
The Importance of Contiguous Memory	505
B. More on the IPython System.....	509
B.1 Terminal Keyboard Shortcuts	509
B.2 About Magic Commands	510
The %run Command	512
Executing Code from the Clipboard	513
B.3 Using the Command History	514
Searching and Reusing the Command History	514
Input and Output Variables	515
B.4 Interacting with the Operating System	516
Shell Commands and Aliases	517
Directory Bookmark System	518
B.5 Software Development Tools	519
Interactive Debugger	519
Timing Code: %time and %timeit	523
Basic Profiling: %prun and %run -p	525
Profiling a Function Line by Line	527
B.6 Tips for Productive Code Development Using IPython	529
Reloading Module Dependencies	529
Code Design Tips	530
B.7 Advanced IPython Features	532
Profiles and Configuration	532
B.8 Conclusion	533
Index.....	535