

SECOND EDITION

Python for Data Analysis

*Data Wrangling with Pandas, NumPy,
and IPython*

Wes McKinney

Beijing • Boston • Farnham • Sebastopol • Tokyo

O'REILLY®

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	2
Solving the “Two-Language” Problem	3
Why Not Python?	3
1.3 Essential Python Libraries	4
NumPy	4
pandas	4
matplotlib	5
IPython and Jupyter	6
SciPy	6
scikit-learn	7
statsmodels	8
1.4 Installation and Setup	8
Windows	9
Apple (OS X, macOS)	9
GNU/Linux	9
Installing or Updating Python Packages	10
Python 2 and Python 3	11
Integrated Development Environments (IDEs) and Text Editors	11
1.5 Community and Conferences	12
1.6 Navigating This Book	12
Code Examples	13
Data for Examples	13

Import Conventions	14
Jargon	14
2. Python Language Basics, IPython, and Jupyter Notebooks.....	15
2.1 The Python Interpreter	16
2.2 IPython Basics	17
Running the IPython Shell	17
Running the Jupyter Notebook	18
Tab Completion	21
Introspection	23
The %run Command	25
Executing Code from the Clipboard	26
Terminal Keyboard Shortcuts	27
About Magic Commands	28
Matplotlib Integration	29
2.3 Python Language Basics	30
Language Semantics	30
Scalar Types	38
Control Flow	46
3. Built-in Data Structures, Functions, and Files.....	51
3.1 Data Structures and Sequences	51
Tuple	51
List	54
Built-in Sequence Functions	59
dict	61
set	65
List, Set, and Dict Comprehensions	67
3.2 Functions	69
Namespaces, Scope, and Local Functions	70
Returning Multiple Values	71
Functions Are Objects	72
Anonymous (Lambda) Functions	73
Currying: Partial Argument Application	74
Generators	75
Errors and Exception Handling	77
3.3 Files and the Operating System	80
Bytes and Unicode with Files	83
3.4 Conclusion	84
4. NumPy Basics: Arrays and Vectorized Computation.....	85
4.1 The NumPy ndarray: A Multidimensional Array Object	87

Creating ndarrays	88
Data Types for ndarrays	90
Arithmetic with NumPy Arrays	93
Basic Indexing and Slicing	94
Boolean Indexing	99
Fancy Indexing	102
Transposing Arrays and Swapping Axes	103
4.2 Universal Functions: Fast Element-Wise Array Functions	105
4.3 Array-Oriented Programming with Arrays	108
Expressing Conditional Logic as Array Operations	109
Mathematical and Statistical Methods	111
Methods for Boolean Arrays	113
Sorting	113
Unique and Other Set Logic	114
4.4 File Input and Output with Arrays	115
4.5 Linear Algebra	116
4.6 Pseudorandom Number Generation	118
4.7 Example: Random Walks	119
Simulating Many Random Walks at Once	121
4.8 Conclusion	122
5. Getting Started with pandas.....	123
5.1 Introduction to pandas Data Structures	124
Series	124
DataFrame	128
Index Objects	134
5.2 Essential Functionality	136
Reindexing	136
Dropping Entries from an Axis	138
Indexing, Selection, and Filtering	140
Integer Indexes	145
Arithmetic and Data Alignment	146
Function Application and Mapping	151
Sorting and Ranking	153
Axis Indexes with Duplicate Labels	157
5.3 Summarizing and Computing Descriptive Statistics	158
Correlation and Covariance	160
Unique Values, Value Counts, and Membership	162
5.4 Conclusion	165
6. Data Loading, Storage, and File Formats.....	167
6.1 Reading and Writing Data in Text Format	167

Reading Text Files in Pieces	173
Writing Data to Text Format	175
Working with Delimited Formats	176
JSON Data	178
XML and HTML: Web Scraping	180
6.2 Binary Data Formats	183
Using HDF5 Format	184
Reading Microsoft Excel Files	186
6.3 Interacting with Web APIs	187
6.4 Interacting with Databases	189
6.5 Conclusion	190
7. Data Cleaning and Preparation.....	191
7.1 Handling Missing Data	191
Filtering Out Missing Data	193
Filling In Missing Data	195
7.2 Data Transformation	197
Removing Duplicates	197
Transforming Data Using a Function or Mapping	198
Replacing Values	200
Renaming Axis Indexes	201
Discretization and Binning	203
Detecting and Filtering Outliers	205
Permutation and Random Sampling	206
Computing Indicator/Dummy Variables	208
7.3 String Manipulation	211
String Object Methods	211
Regular Expressions	213
Vectorized String Functions in pandas	216
7.4 Conclusion	219
8. Data Wrangling: Join, Combine, and Reshape.....	221
8.1 Hierarchical Indexing	221
Reordering and Sorting Levels	224
Summary Statistics by Level	225
Indexing with a DataFrame's columns	225
8.2 Combining and Merging Datasets	227
Database-Style DataFrame Joins	227
Merging on Index	232
Concatenating Along an Axis	236
Combining Data with Overlap	241
8.3 Reshaping and Pivoting	242

Reshaping with Hierarchical Indexing	243
Pivoting “Long” to “Wide” Format	246
Pivoting “Wide” to “Long” Format	249
8.4 Conclusion	251
9. Plotting and Visualization.....	253
9.1 A Brief matplotlib API Primer	254
Figures and Subplots	255
Colors, Markers, and Line Styles	260
Ticks, Labels, and Legends	262
Annotations and Drawing on a Subplot	266
Saving Plots to File	268
matplotlib Configuration	269
9.2 Plotting with pandas and seaborn	269
Line Plots	270
Bar Plots	273
Histograms and Density Plots	278
Scatter or Point Plots	281
Facet Grids and Categorical Data	284
9.3 Other Python Visualization Tools	286
9.4 Conclusion	287
10. Data Aggregation and Group Operations.....	289
10.1 GroupBy Mechanics	290
Iterating Over Groups	293
Selecting a Column or Subset of Columns	295
Grouping with Dicts and Series	296
Grouping with Functions	297
Grouping by Index Levels	297
10.2 Data Aggregation	298
Column-Wise and Multiple Function Application	300
Returning Aggregated Data Without Row Indexes	303
10.3 Apply: General split-apply-combine	304
Suppressing the Group Keys	306
Quantile and Bucket Analysis	307
Example: Filling Missing Values with Group-Specific Values	308
Example: Random Sampling and Permutation	310
Example: Group Weighted Average and Correlation	312
Example: Group-Wise Linear Regression	314
10.4 Pivot Tables and Cross-Tabulation	315
Cross-Tabulations: Crosstab	317
10.5 Conclusion	318

11. Time Series.....	319
11.1 Date and Time Data Types and Tools	320
Converting Between String and Datetime	321
11.2 Time Series Basics	324
Indexing, Selection, Subsetting	325
Time Series with Duplicate Indices	328
11.3 Date Ranges, Frequencies, and Shifting	329
Generating Date Ranges	330
Frequencies and Date Offsets	332
Shifting (Leading and Lagging) Data	334
11.4 Time Zone Handling	337
Time Zone Localization and Conversion	337
Operations with Time Zone–Aware Timestamp Objects	340
Operations Between Different Time Zones	341
11.5 Periods and Period Arithmetic	341
Period Frequency Conversion	342
Quarterly Period Frequencies	344
Converting Timestamps to Periods (and Back)	346
Creating a PeriodIndex from Arrays	347
11.6 Resampling and Frequency Conversion	350
Downsampling	351
Upsampling and Interpolation	354
Resampling with Periods	355
11.7 Moving Window Functions	356
Exponentially Weighted Functions	360
Binary Moving Window Functions	361
User-Defined Moving Window Functions	363
11.8 Conclusion	364
12. Advanced pandas.....	365
12.1 Categorical Data	365
Background and Motivation	365
Categorical Type in pandas	367
Computations with Categoricals	369
Categorical Methods	372
12.2 Advanced GroupBy Use	375
Group Transforms and “Unwrapped” GroupBys	375
Grouped Time Resampling	379
12.3 Techniques for Method Chaining	380
The pipe Method	382
12.4 Conclusion	383

13. Introduction to Modeling Libraries in Python.....	385
13.1 Interfacing Between pandas and Model Code	385
13.2 Creating Model Descriptions with Patsy	388
Data Transformations in Patsy Formulas	391
Categorical Data and Patsy	392
13.3 Introduction to statsmodels	395
Estimating Linear Models	395
Estimating Time Series Processes	398
13.4 Introduction to scikit-learn	399
13.5 Continuing Your Education	403
14. Data Analysis Examples.....	405
14.1 1.USA.gov Data from Bitly	405
Counting Time Zones in Pure Python	406
Counting Time Zones with pandas	408
14.2 MovieLens 1M Dataset	415
Measuring Rating Disagreement	420
14.3 US Baby Names 1880–2010	421
Analyzing Naming Trends	427
14.4 USDA Food Database	436
14.5 2012 Federal Election Commission Database	442
Donation Statistics by Occupation and Employer	444
Bucketing Donation Amounts	447
Donation Statistics by State	449
14.6 Conclusion	450
A. Advanced NumPy.....	451
A.1 ndarray Object Internals	451
NumPy dtype Hierarchy	452
A.2 Advanced Array Manipulation	453
Reshaping Arrays	454
C Versus Fortran Order	456
Concatenating and Splitting Arrays	456
Repeating Elements: tile and repeat	459
Fancy Indexing Equivalents: take and put	461
A.3 Broadcasting	462
Broadcasting Over Other Axes	464
Setting Array Values by Broadcasting	467
A.4 Advanced ufunc Usage	468
ufunc Instance Methods	468
Writing New ufuncs in Python	470
A.5 Structured and Record Arrays	471

Nested dtypes and Multidimensional Fields	471
Why Use Structured Arrays?	472
A.6 More About Sorting	473
Indirect Sorts: argsort and lexsort	474
Alternative Sort Algorithms	476
Partially Sorting Arrays	476
numpy.searchsorted: Finding Elements in a Sorted Array	477
A.7 Writing Fast NumPy Functions with Numba	478
Creating Custom numpy.ufunc Objects with Numba	480
A.8 Advanced Array Input and Output	480
Memory-Mapped Files	480
HDF5 and Other Array Storage Options	482
A.9 Performance Tips	482
The Importance of Contiguous Memory	482
B. More on the IPython System.....	485
B.1 Using the Command History	485
Searching and Reusing the Command History	485
Input and Output Variables	486
B.2 Interacting with the Operating System	487
Shell Commands and Aliases	488
Directory Bookmark System	489
B.3 Software Development Tools	489
Interactive Debugger	490
Timing Code: %time and %timeit	494
Basic Profiling: %prun and %run -p	496
Profiling a Function Line by Line	498
B.4 Tips for Productive Code Development Using IPython	500
Reloading Module Dependencies	500
Code Design Tips	501
B.5 Advanced IPython Features	502
Making Your Own Classes IPython-Friendly	503
Profiles and Configuration	503
B.6 Conclusion	505
Index.....	507