

Agile Tools for Real-World Data

Python for Data Analysis



O'REILLY®

Wes McKinney

Table of Contents

Preface	xi
1. Preliminaries	1
What Is This Book About?	1
Why Python for Data Analysis?	2
Python as Glue	2
Solving the “Two-Language” Problem	2
Why Not Python?	3
Essential Python Libraries	3
NumPy	4
pandas	4
matplotlib	5
IPython	5
SciPy	6
Installation and Setup	6
Windows	7
Apple OS X	9
GNU/Linux	10
Python 2 and Python 3	11
Integrated Development Environments (IDEs)	11
Community and Conferences	12
Navigating This Book	12
Code Examples	13
Data for Examples	13
Import Conventions	13
Jargon	13
Acknowledgements	14
2. Introductory Examples	17
1.usa.gov data from bit.ly	17
Counting Time Zones in Pure Python	19

Counting Time Zones with pandas	21
MovieLens 1M Data Set	26
Measuring rating disagreement	30
US Baby Names 1880-2010	32
Analyzing Naming Trends	36
Conclusions and The Path Ahead	43
3. IPython: An Interactive Computing and Development Environment	45
IPython Basics	46
Tab Completion	47
Introspection	48
The %run Command	49
Executing Code from the Clipboard	50
Keyboard Shortcuts	52
Exceptions and Tracebacks	53
Magic Commands	54
Qt-based Rich GUI Console	55
Matplotlib Integration and Pylab Mode	56
Using the Command History	58
Searching and Reusing the Command History	58
Input and Output Variables	58
Logging the Input and Output	59
Interacting with the Operating System	60
Shell Commands and Aliases	60
Directory Bookmark System	62
Software Development Tools	62
Interactive Debugger	62
Timing Code: %time and %timeit	67
Basic Profiling: %prun and %run -p	68
Profiling a Function Line-by-Line	70
IPython HTML Notebook	72
Tips for Productive Code Development Using IPython	72
Reloading Module Dependencies	74
Code Design Tips	74
Advanced IPython Features	76
Making Your Own Classes IPython-friendly	76
Profiles and Configuration	77
Credits	78
4. NumPy Basics: Arrays and Vectorized Computation	79
The NumPy ndarray: A Multidimensional Array Object	80
Creating ndarrays	81
Data Types for ndarrays	83

Operations between Arrays and Scalars	85
Basic Indexing and Slicing	86
Boolean Indexing	89
Fancy Indexing	92
Transposing Arrays and Swapping Axes	93
Universal Functions: Fast Element-wise Array Functions	95
Data Processing Using Arrays	97
Expressing Conditional Logic as Array Operations	98
Mathematical and Statistical Methods	100
Methods for Boolean Arrays	101
Sorting	101
Unique and Other Set Logic	102
File Input and Output with Arrays	103
Storing Arrays on Disk in Binary Format	103
Saving and Loading Text Files	104
Linear Algebra	105
Random Number Generation	106
Example: Random Walks	108
Simulating Many Random Walks at Once	109
5. Getting Started with pandas	111
Introduction to pandas Data Structures	112
Series	112
DataFrame	115
Index Objects	120
Essential Functionality	122
Reindexing	122
Dropping entries from an axis	125
Indexing, selection, and filtering	125
Arithmetic and data alignment	128
Function application and mapping	132
Sorting and ranking	133
Axis indexes with duplicate values	136
Summarizing and Computing Descriptive Statistics	137
Correlation and Covariance	139
Unique Values, Value Counts, and Membership	141
Handling Missing Data	142
Filtering Out Missing Data	143
Filling in Missing Data	145
Hierarchical Indexing	147
Reordering and Sorting Levels	149
Summary Statistics by Level	150
Using a DataFrame's Columns	150