Agile Tools for Real-World Data

Python for Data Analysis

O'REILLY®

Wes McKinney

www.it-ebooks.info

www.it-ebooks.info

Python for Data Analysis

Wes McKinney

O'REILLY®
Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

www.it-ebooks.info

Python for Data Analysis
by Wes McKinney

Copyright © 2013 Wes McKinney. All rights reserved.
Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://my.safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editors: Julie Steele and Meghan Blanchette
Production Editor: Melanie Yarbrough
Copyeditor: Teresa Exley
Proofreader: BIM Publishing Services

Indexer: BIM Publishing Services
Cover Designer: Karen Montgomery
Interior Designer: David Futato
Illustrator: Rebecca Demarest

October 2012: First Edition.

Revision History for the First Edition:
2012-10-05 First release
See *http://oreilly.com/catalog/errata.csp?isbn=9781449319793* for release details.

Nutshell Handbook, the Nutshell Handbook logo, and the O'Reilly logo are registered trademarks of O'Reilly Media, Inc. *Python for Data Analysis*, the cover image of a golden-tailed tree shrew, and related trade dress are trademarks of O'Reilly Media, Inc.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and O'Reilly Media, Inc., was aware of a trademark claim, the designations have been printed in caps or initial caps.

While every precaution has been taken in the preparation of this book, the publisher and author assume no responsibility for errors or omissions, or for damages resulting from the use of the information contained herein.

ISBN: 978-1-449-31979-3

[LSI]

1349356084

www.it-ebooks.info

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

www.it-ebooks.info

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

www.it-ebooks.info

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

www.it-ebooks.info

Other pandas Topics 151
Integer Indexing 151
Panel Data 152

6. **Data Loading, Storage, and File Formats** **155**
Reading and Writing Data in Text Format 155
Reading Text Files in Pieces 160
Writing Data Out to Text Format 162
Manually Working with Delimited Formats 163
JSON Data 165
XML and HTML: Web Scraping 166
Binary Data Formats 171
Using HDF5 Format 171
Reading Microsoft Excel Files 172
Interacting with HTML and Web APIs 173
Interacting with Databases 174
Storing and Loading Data in MongoDB 176

7. **Data Wrangling: Clean, Transform, Merge, Reshape** **177**
Combining and Merging Data Sets 177
Database-style DataFrame Merges 178
Merging on Index 182
Concatenating Along an Axis 185
Combining Data with Overlap 188
Reshaping and Pivoting 189
Reshaping with Hierarchical Indexing 190
Pivoting “long” to “wide” Format 192
Data Transformation 194
Removing Duplicates 194
Transforming Data Using a Function or Mapping 195
Replacing Values 196
Renaming Axis Indexes 197
Discretization and Binning 199
Detecting and Filtering Outliers 201
Permutation and Random Sampling 202
Computing Indicator/Dummy Variables 203
String Manipulation 205
String Object Methods 206
Regular expressions 207
Vectorized string functions in pandas 210
Example: USDA Food Database 212

www.it-ebooks.info

8. **Plotting and Visualization** **219**
A Brief matplotlib API Primer 219
Figures and Subplots 220
Colors, Markers, and Line Styles 224
Ticks, Labels, and Legends 225
Annotations and Drawing on a Subplot 228
Saving Plots to File 231
matplotlib Configuration 231
Plotting Functions in pandas 232
Line Plots 232
Bar Plots 235
Histograms and Density Plots 238
Scatter Plots 239
Plotting Maps: Visualizing Haiti Earthquake Crisis Data 241
Python Visualization Tool Ecosystem 247
Chaco 248
mayavi 248
Other Packages 248
The Future of Visualization Tools? 249

9. **Data Aggregation and Group Operations** **251**
GroupBy Mechanics 252
Iterating Over Groups 255
Selecting a Column or Subset of Columns 256
Grouping with Dicts and Series 257
Grouping with Functions 258
Grouping by Index Levels 259
Data Aggregation 259
Column-wise and Multiple Function Application 262
Returning Aggregated Data in “unindexed” Form 264
Group-wise Operations and Transformations 264
Apply: General split-apply-combine 266
Quantile and Bucket Analysis 268
Example: Filling Missing Values with Group-specific Values 270
Example: Random Sampling and Permutation 271
Example: Group Weighted Average and Correlation 273
Example: Group-wise Linear Regression 274
Pivot Tables and Cross-Tabulation 275
Cross-Tabulations: Crosstab 277
Example: 2012 Federal Election Commission Database 278
Donation Statistics by Occupation and Employer 280
Bucketing Donation Amounts 283
Donation Statistics by State 285

www.it-ebooks.info

10. Time Series 289
Date and Time Data Types and Tools 290
Converting between string and datetime 291
Time Series Basics 293
Indexing, Selection, Subsetting 294
Time Series with Duplicate Indices 296
Date Ranges, Frequencies, and Shifting 297
Generating Date Ranges 298
Frequencies and Date Offsets 299
Shifting (Leading and Lagging) Data 301
Time Zone Handling 303
Localization and Conversion 304
Operations with Time Zone–aware Timestamp Objects 305
Operations between Different Time Zones 306
Periods and Period Arithmetic 307
Period Frequency Conversion 308
Quarterly Period Frequencies 309
Converting Timestamps to Periods (and Back) 311
Creating a PeriodIndex from Arrays 312
Resampling and Frequency Conversion 312
Downsampling 314
Upsampling and Interpolation 316
Resampling with Periods 318
Time Series Plotting 319
Moving Window Functions 320
Exponentially-weighted functions 324
Binary Moving Window Functions 324
User-Defined Moving Window Functions 326
Performance and Memory Usage Notes 327

11. Financial and Economic Data Applications 329
Data Munging Topics 329
Time Series and Cross-Section Alignment 330
Operations with Time Series of Different Frequencies 332
Time of Day and “as of” Data Selection 334
Splicing Together Data Sources 336
Return Indexes and Cumulative Returns 338
Group Transforms and Analysis 340
Group Factor Exposures 342
Decile and Quartile Analysis 343
More Example Applications 345
Signal Frontier Analysis 345
Future Contract Rolling 347

www.it-ebooks.info