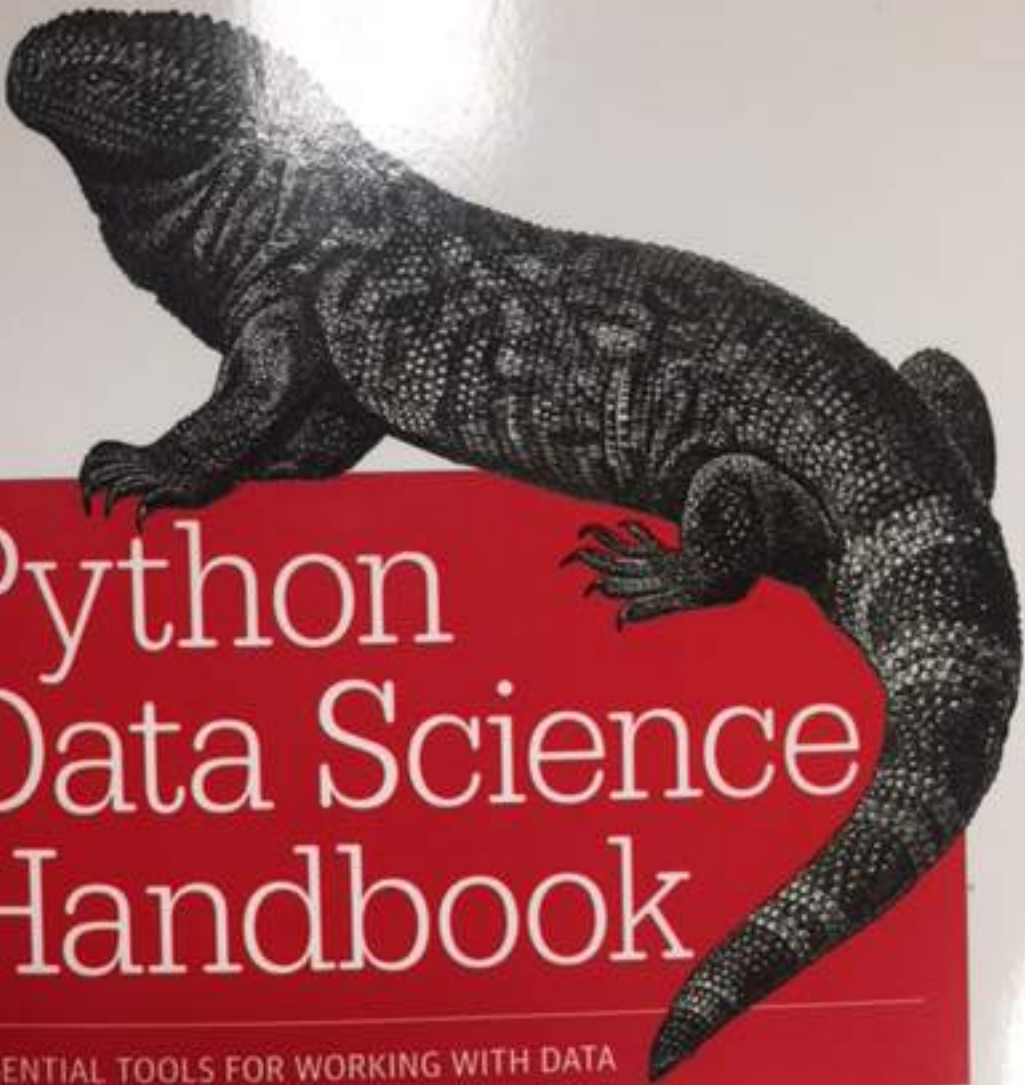


O'REILLY



Python Data Science Handbook

ESSENTIAL TOOLS FOR WORKING WITH DATA

powered by



Jake VanderPlas

Table of Contents

Preface.....	xi
1. IPython: Beyond Normal Python.....	1
Shell or Notebook?	2
Launching the IPython Shell	2
Launching the Jupyter Notebook	2
Help and Documentation in IPython	3
Accessing Documentation with ?	3
Accessing Source Code with ??	5
Exploring Modules with Tab Completion	6
Keyboard Shortcuts in the IPython Shell	8
Navigation Shortcuts	8
Text Entry Shortcuts	9
Command History Shortcuts	9
Miscellaneous Shortcuts	10
IPython Magic Commands	10
Pasting Code Blocks: %paste and %cpaste	11
Running External Code: %run	12
Timing Code Execution: %timeit	12
Help on Magic Functions: ?, %magic, and %lsmagic	13
Input and Output History	13
IPython's In and Out Objects	13
Underscore Shortcuts and Previous Outputs	15
Suppressing Output	15
Related Magic Commands	16
IPython and Shell Commands	16
Quick Introduction to the Shell	16
Shell Commands in IPython	18

Passing Values to and from the Shell	18
Shell-Related Magic Commands	19
Errors and Debugging	20
Controlling Exceptions: %xmode	20
Debugging: When Reading Tracebacks Is Not Enough	22
Profiling and Timing Code	25
Timing Code Snippets: %timeit and %time	25
Profiling Full Scripts: %prun	27
Line-by-Line Profiling with %lprun	28
Profiling Memory Use: %memit and %mprun	29
More IPython Resources	30
Web Resources	30
Books	31
2. Introduction to NumPy	33
Understanding Data Types in Python	34
A Python Integer Is More Than Just an Integer	35
A Python List Is More Than Just a List	37
Fixed-Type Arrays in Python	38
Creating Arrays from Python Lists	39
Creating Arrays from Scratch	39
NumPy Standard Data Types	41
The Basics of NumPy Arrays	42
NumPy Array Attributes	42
Array Indexing: Accessing Single Elements	43
Array Slicing: Accessing Subarrays	44
Reshaping of Arrays	47
Array Concatenation and Splitting	48
Computation on NumPy Arrays: Universal Functions	50
The Slowness of Loops	50
Introducing UFuncs	51
Exploring NumPy's UFuncs	52
Advanced Ufunc Features	56
Ufuncs: Learning More	58
Aggregations: Min, Max, and Everything in Between	58
Summing the Values in an Array	59
Minimum and Maximum	59
Example: What Is the Average Height of US Presidents?	61
Computation on Arrays: Broadcasting	63
Introducing Broadcasting	63
Rules of Broadcasting	65
Broadcasting in Practice	68

Comparisons, Masks, and Boolean Logic	70
Example: Counting Rainy Days	70
Comparison Operators as ufuncs	71
Working with Boolean Arrays	73
Boolean Arrays as Masks	75
Fancy Indexing	78
Exploring Fancy Indexing	79
Combined Indexing	80
Example: Selecting Random Points	81
Modifying Values with Fancy Indexing	82
Example: Binning Data	83
Sorting Arrays	85
Fast Sorting in NumPy: np.sort and np.argsort	86
Partial Sorts: Partitioning	88
Example: k-Nearest Neighbors	88
Structured Data: NumPy's Structured Arrays	92
Creating Structured Arrays	94
More Advanced Compound Types	95
RecordArrays: Structured Arrays with a Twist	96
On to Pandas	96
3. Data Manipulation with Pandas	97
Installing and Using Pandas	97
Introducing Pandas Objects	98
The Pandas Series Object	99
The Pandas DataFrame Object	102
The Pandas Index Object	105
Data Indexing and Selection	107
Data Selection in Series	107
Data Selection in DataFrame	110
Operating on Data in Pandas	115
Ufuncs: Index Preservation	115
Ufuncs: Index Alignment	116
Ufuncs: Operations Between DataFrame and Series	118
Handling Missing Data	119
Trade-Offs in Missing Data Conventions	120
Missing Data in Pandas	120
Operating on Null Values	124
Hierarchical Indexing	128
A Multiply Indexed Series	128
Methods of MultiIndex Creation	131
Indexing and Slicing a MultiIndex	134

Rearranging Multi-Indices	137
Data Aggregations on Multi-Indices	140
Combining Datasets: Concat and Append	141
Recall: Concatenation of NumPy Arrays	142
Simple Concatenation with <code>pd.concat</code>	142
Combining Datasets: Merge and Join	146
Relational Algebra	146
Categories of Joins	147
Specification of the Merge Key	149
Specifying Set Arithmetic for Joins	152
Overlapping Column Names: The <code>suffixes</code> Keyword	153
Example: US States Data	154
Aggregation and Grouping	158
Planets Data	159
Simple Aggregation in Pandas	159
GroupBy: Split, Apply, Combine	161
Pivot Tables	170
Motivating Pivot Tables	170
Pivot Tables by Hand	171
Pivot Table Syntax	171
Example: Birthrate Data	174
Vectorized String Operations	178
Introducing Pandas String Operations	178
Tables of Pandas String Methods	180
Example: Recipe Database	184
Working with Time Series	188
Dates and Times in Python	188
Pandas Time Series: Indexing by Time	192
Pandas Time Series Data Structures	193
Frequencies and Offsets	195
Resampling, Shifting, and Windowing	196
Where to Learn More	202
Example: Visualizing Seattle Bicycle Counts	202
High-Performance Pandas: <code>eval()</code> and <code>query()</code>	208
Motivating <code>query()</code> and <code>eval()</code> : Compound Expressions	209
<code>pandas.eval()</code> for Efficient Operations	210
<code>DataFrame.eval()</code> for Column-Wise Operations	211
<code>DataFrame.query()</code> Method	213
Performance: When to Use These Functions	214
Further Resources	215

4. Visualization with Matplotlib	217
General Matplotlib Tips	218
Importing matplotlib	218
Setting Styles	218
show() or No show()? How to Display Your Plots	221
Saving Figures to File	222
Two Interfaces for the Price of One	224
Simple Line Plots	226
Adjusting the Plot: Line Colors and Styles	228
Adjusting the Plot: Axes Limits	230
Labeling Plots	233
Simple Scatter Plots	233
Scatter Plots with plt.plot	235
Scatter Plots with plt.scatter	237
plot Versus scatter: A Note on Efficiency	237
Visualizing Errors	238
Basic Errorbars	239
Continuous Errors	241
Density and Contour Plots	241
Visualizing a Three-Dimensional Function	245
Histograms, Binnings, and Density	247
Two-Dimensional Histograms and Binnings	249
Customizing Plot Legends	251
Choosing Elements for the Legend	252
Legend for Size of Points	254
Multiple Legends	255
Customizing Colorbars	256
Customizing Colorbars	261
Example: Handwritten Digits	262
Multiple Subplots	263
plt.axes: Subplots by Hand	264
plt.subplot: Simple Grids of Subplots	265
plt.subplots: The Whole Grid in One Go	266
plt.GridSpec: More Complicated Arrangements	268
Text and Annotation	269
Example: Effect of Holidays on US Births	270
Transforms and Text Position	272
Arrows and Annotation	275
Customizing Ticks	276
Major and Minor Ticks	277
Hiding Ticks or Labels	277
Reducing or Increasing the Number of Ticks	278

Fancy Tick Formats	279
Summary of Formatters and Locators	281
Customizing Matplotlib: Configurations and Stylesheets	282
Plot Customization by Hand	282
Changing the Defaults: rcParams	284
Stylesheets	285
Three-Dimensional Plotting in Matplotlib	290
Three-Dimensional Points and Lines	291
Three-Dimensional Contour Plots	292
Wireframes and Surface Plots	293
Surface Triangulations	295
Geographic Data with Basemap	298
Map Projections	300
Drawing a Map Background	304
Plotting Data on Maps	307
Example: California Cities	308
Example: Surface Temperature Data	309
Visualization with Seaborn	311
Seaborn Versus Matplotlib	312
Exploring Seaborn Plots	313
Example: Exploring Marathon Finishing Times	322
Further Resources	329
Matplotlib Resources	329
Other Python Graphics Libraries	330
5. Machine Learning	331
What Is Machine Learning?	332
Categories of Machine Learning	332
Qualitative Examples of Machine Learning Applications	333
Summary	342
Introducing Scikit-Learn	343
Data Representation in Scikit-Learn	343
Scikit-Learn's Estimator API	346
Application: Exploring Handwritten Digits	354
Summary	359
Hyperparameters and Model Validation	359
Thinking About Model Validation	359
Selecting the Best Model	363
Learning Curves	370
Validation in Practice: Grid Search	373
Summary	375
Feature Engineering	375

Categorical Features	376
Text Features	377
Image Features	378
Derived Features	378
Imputation of Missing Data	381
Feature Pipelines	381
In Depth: Naive Bayes Classification	382
Bayesian Classification	383
Gaussian Naive Bayes	383
Multinomial Naive Bayes	386
When to Use Naive Bayes	389
In Depth: Linear Regression	390
Simple Linear Regression	390
Basis Function Regression	392
Regularization	396
Example: Predicting Bicycle Traffic	400
In-Depth: Support Vector Machines	405
Motivating Support Vector Machines	405
Support Vector Machines: Maximizing the Margin	407
Example: Face Recognition	416
Support Vector Machine Summary	420
In-Depth: Decision Trees and Random Forests	421
Motivating Random Forests: Decision Trees	421
Ensembles of Estimators: Random Forests	426
Random Forest Regression	428
Example: Random Forest for Classifying Digits	430
Summary of Random Forests	432
In Depth: Principal Component Analysis	433
Introducing Principal Component Analysis	433
PCA as Noise Filtering	440
Example: Eigenfaces	442
Principal Component Analysis Summary	445
In-Depth: Manifold Learning	445
Manifold Learning: "HELLO"	446
Multidimensional Scaling (MDS)	447
MDS as Manifold Learning	450
Nonlinear Embeddings: Where MDS Fails	452
Nonlinear Manifolds: Locally Linear Embedding	453
Some Thoughts on Manifold Methods	455
Example: Isomap on Faces	456
Example: Visualizing Structure in Digits	460
In Depth: k-Means Clustering	462

Introducing k-Means	463
k-Means Algorithm: Expectation–Maximization	465
Examples	470
In Depth: Gaussian Mixture Models	476
Motivating GMM: Weaknesses of k-Means	477
Generalizing E–M: Gaussian Mixture Models	480
GMM as Density Estimation	484
Example: GMM for Generating New Data	488
In-Depth: Kernel Density Estimation	491
Motivating KDE: Histograms	491
Kernel Density Estimation in Practice	496
Example: KDE on a Sphere	498
Example: Not-So-Naive Bayes	501
Application: A Face Detection Pipeline	506
HOG Features	506
HOG in Action: A Simple Face Detector	507
Caveats and Improvements	512
Further Machine Learning Resources	514
Machine Learning in Python	514
General Machine Learning	515
Index	517