

SECOND EDITION

# Python Data Science Handbook

*Essential Tools for Working with Data*

*Jake VanderPlas*

Beijing • Boston • Farnham • Sebastopol • Tokyo

**O'REILLY®**

# Table of Contents

Preface.....	xix
--------------	-----

---

## Part I. Jupyter: Beyond Normal Python

<b>1. Getting Started in IPython and Jupyter.....</b>	<b>3</b>
Launching the IPython Shell	3
Launching the Jupyter Notebook	4
Help and Documentation in IPython	4
Accessing Documentation with ?	5
Accessing Source Code with ??	6
Exploring Modules with Tab Completion	7
Keyboard Shortcuts in the IPython Shell	9
Navigation Shortcuts	10
Text Entry Shortcuts	10
Command History Shortcuts	10
Miscellaneous Shortcuts	12
<b>2. Enhanced Interactive Features.....</b>	<b>13</b>
IPython Magic Commands	13
Running External Code: %run	13
Timing Code Execution: %timeit	14
Help on Magic Functions: ?, %magic, and %lsmagic	15
Input and Output History	15
IPython's In and Out Objects	15
Underscore Shortcuts and Previous Outputs	16
Suppressing Output	17
Related Magic Commands	17

IPython and Shell Commands	18
Quick Introduction to the Shell	18
Shell Commands in IPython	19
Passing Values to and from the Shell	20
Shell-Related Magic Commands	20
<b>3. Debugging and Profiling.....</b>	<b>22</b>
Errors and Debugging	22
Controlling Exceptions: %xmode	22
Debugging: When Reading Tracebacks Is Not Enough	24
Profiling and Timing Code	26
Timing Code Snippets: %timeit and %time	27
Profiling Full Scripts: %prun	28
Line-by-Line Profiling with %lprun	29
Profiling Memory Use: %memit and %mprun	30
More IPython Resources	31
Web Resources	31
Books	32

---

## Part II. Introduction to NumPy

<b>4. Understanding Data Types in Python.....</b>	<b>35</b>
A Python Integer Is More Than Just an Integer	36
A Python List Is More Than Just a List	37
Fixed-Type Arrays in Python	39
Creating Arrays from Python Lists	39
Creating Arrays from Scratch	40
NumPy Standard Data Types	41
<b>5. The Basics of NumPy Arrays.....</b>	<b>43</b>
NumPy Array Attributes	44
Array Indexing: Accessing Single Elements	44
Array Slicing: Accessing Subarrays	45
One-Dimensional Subarrays	45
Multidimensional Subarrays	46
Subarrays as No-Copy Views	47
Creating Copies of Arrays	47
Reshaping of Arrays	48
Array Concatenation and Splitting	49
Concatenation of Arrays	49
Splitting of Arrays	50

<b>6. Computation on NumPy Arrays: Universal Functions.....</b>	<b>51</b>
The Slowness of Loops	51
Introducing Ufuncs	52
Exploring NumPy’s Ufuncs	53
Array Arithmetic	53
Absolute Value	55
Trigonometric Functions	55
Exponents and Logarithms	56
Specialized Ufuncs	56
Advanced Ufunc Features	57
Specifying Output	57
Aggregations	58
Outer Products	59
Ufuncs: Learning More	59
<b>7. Aggregations: min, max, and Everything in Between.....</b>	<b>60</b>
Summing the Values in an Array	60
Minimum and Maximum	61
Multidimensional Aggregates	61
Other Aggregation Functions	62
Example: What Is the Average Height of US Presidents?	63
<b>8. Computation on Arrays: Broadcasting.....</b>	<b>65</b>
Introducing Broadcasting	65
Rules of Broadcasting	67
Broadcasting Example 1	68
Broadcasting Example 2	68
Broadcasting Example 3	69
Broadcasting in Practice	70
Centering an Array	70
Plotting a Two-Dimensional Function	71
<b>9. Comparisons, Masks, and Boolean Logic.....</b>	<b>72</b>
Example: Counting Rainy Days	72
Comparison Operators as Ufuncs	73
Working with Boolean Arrays	75
Counting Entries	75
Boolean Operators	76
Boolean Arrays as Masks	77
Using the Keywords and/or Versus the Operators &/	78

<b>10. Fancy Indexing.....</b>	<b>80</b>
Exploring Fancy Indexing	80
Combined Indexing	81
Example: Selecting Random Points	82
Modifying Values with Fancy Indexing	84
Example: Binning Data	85
<b>11. Sorting Arrays.....</b>	<b>88</b>
Fast Sorting in NumPy: np.sort and np.argsort	89
Sorting Along Rows or Columns	89
Partial Sorts: Partitioning	90
Example: k-Nearest Neighbors	90
<b>12. Structured Data: NumPy's Structured Arrays.....</b>	<b>94</b>
Exploring Structured Array Creation	96
More Advanced Compound Types	97
Record Arrays: Structured Arrays with a Twist	97
On to Pandas	98

---

## Part III. Data Manipulation with Pandas

<b>13. Introducing Pandas Objects.....</b>	<b>101</b>
The Pandas Series Object	101
Series as Generalized NumPy Array	102
Series as Specialized Dictionary	103
Constructing Series Objects	104
The Pandas DataFrame Object	104
DataFrame as Generalized NumPy Array	105
<i>DataFrame as Specialized Dictionary</i>	106
Constructing DataFrame Objects	106
The Pandas Index Object	108
Index as Immutable Array	108
Index as Ordered Set	108
<b>14. Data Indexing and Selection.....</b>	<b>110</b>
Data Selection in Series	110
Series as Dictionary	110
Series as One-Dimensional Array	111
Indexers: loc and iloc	112
Data Selection in DataFrames	113

DataFrame as Dictionary	113
DataFrame as Two-Dimensional Array	115
Additional Indexing Conventions	116
<b>15. Operating on Data in Pandas.....</b>	<b>118</b>
Ufuncs: Index Preservation	118
Ufuncs: Index Alignment	119
Index Alignment in Series	119
Index Alignment in DataFrames	120
Ufuncs: Operations Between DataFrames and Series	121
<b>16. Handling Missing Data.....</b>	<b>123</b>
Trade-offs in Missing Data Conventions	123
Missing Data in Pandas	124
None as a Sentinel Value	125
NaN: Missing Numerical Data	125
NaN and None in Pandas	126
Pandas Nullable Dtypes	127
Operating on Null Values	128
Detecting Null Values	128
Dropping Null Values	129
Filling Null Values	130
<b>17. Hierarchical Indexing.....</b>	<b>132</b>
A Multiply Indexed Series	132
The Bad Way	133
The Better Way: The Pandas MultiIndex	133
MultiIndex as Extra Dimension	134
Methods of MultiIndex Creation	136
Explicit MultiIndex Constructors	136
MultiIndex Level Names	137
MultiIndex for Columns	138
Indexing and Slicing a MultiIndex	138
Multiply Indexed Series	139
Multiply Indexed DataFrames	140
Rearranging Multi-Indexes	141
Sorted and Unsorted Indices	141
Stacking and Unstacking Indices	143
Index Setting and Resetting	143
<b>18. Combining Datasets: concat and append.....</b>	<b>145</b>
Recall: Concatenation of NumPy Arrays	146

Simple Concatenation with <code>pd.concat</code>	147
Duplicate Indices	148
Concatenation with Joins	149
The <code>append</code> Method	150
<b>19. Combining Datasets: <code>merge</code> and <code>join</code>.</b>	<b>151</b>
Relational Algebra	151
Categories of Joins	152
One-to-One Joins	152
Many-to-One Joins	153
Many-to-Many Joins	153
Specification of the Merge Key	154
The <code>on</code> Keyword	154
<i>The <code>left_on</code> and <code>right_on</code> Keywords</i>	155
The <code>left_index</code> and <code>right_index</code> Keywords	155
Specifying Set Arithmetic for Joins	157
Overlapping Column Names: The <code>suffixes</code> Keyword	158
Example: US States Data	159
<b>20. Aggregation and Grouping.</b>	<b>164</b>
Planets Data	165
Simple Aggregation in Pandas	165
<code>groupby</code> : Split, Apply, Combine	167
Split, Apply, Combine	167
The <code>GroupBy</code> Object	169
Aggregate, Filter, Transform, Apply	171
Specifying the Split Key	174
Grouping Example	175
<b>21. Pivot Tables.</b>	<b>176</b>
Motivating Pivot Tables	176
Pivot Tables by Hand	177
Pivot Table Syntax	178
Multilevel Pivot Tables	178
Additional Pivot Table Options	179
Example: Birthrate Data	180
<b>22. Vectorized String Operations.</b>	<b>185</b>
Introducing Pandas String Operations	185
Tables of Pandas String Methods	186
Methods Similar to Python String Methods	186
Methods Using Regular Expressions	187

Miscellaneous Methods	188
Example: Recipe Database	190
A Simple Recipe Recommender	192
Going Further with Recipes	193
<b>23. Working with Time Series.....</b>	<b>194</b>
Dates and Times in Python	195
Native Python Dates and Times: datetime and dateutil	195
Typed Arrays of Times: NumPy's datetime64	196
Dates and Times in Pandas: The Best of Both Worlds	197
Pandas Time Series: Indexing by Time	198
Pandas Time Series Data Structures	199
Regular Sequences: pd.date_range	200
Frequencies and Offsets	201
Resampling, Shifting, and Windowing	202
Resampling and Converting Frequencies	203
Time Shifts	205
Rolling Windows	206
Example: Visualizing Seattle Bicycle Counts	208
Visualizing the Data	209
Digging into the Data	211
<b>24. High-Performance Pandas: eval and query.....</b>	<b>215</b>
Motivating query and eval: Compound Expressions	215
pandas.eval for Efficient Operations	216
DataFrame.eval for Column-Wise Operations	218
Assignment in DataFrame.eval	219
Local Variables in DataFrame.eval	219
The DataFrame.query Method	220
Performance: When to Use These Functions	220
Further Resources	221

---

## Part IV. Visualization with Matplotlib

<b>25. General Matplotlib Tips.....</b>	<b>225</b>
Importing Matplotlib	225
Setting Styles	225
show or No show? How to Display Your Plots	226
Plotting from a Script	226
Plotting from an IPython Shell	227
Plotting from a Jupyter Notebook	227

Saving Figures to File	228
Two Interfaces for the Price of One	230
<b>26. Simple Line Plots.....</b>	<b>232</b>
Adjusting the Plot: Line Colors and Styles	235
Adjusting the Plot: Axes Limits	238
Labeling Plots	240
Matplotlib Gotchas	242
<b>27. Simple Scatter Plots.....</b>	<b>244</b>
Scatter Plots with plt.plot	244
Scatter Plots with plt.scatter	247
plot Versus scatter: A Note on Efficiency	250
Visualizing Uncertainties	251
Basic Errorbars	251
Continuous Errors	253
<b>28. Density and Contour Plots.....</b>	<b>255</b>
Visualizing a Three-Dimensional Function	255
Histograms, Binnings, and Density	260
Two-Dimensional Histograms and Binnings	263
plt.hist2d: Two-Dimensional Histogram	263
plt.hexbin: Hexagonal Binnings	264
Kernel Density Estimation	264
<b>29. Customizing Plot Legends.....</b>	<b>267</b>
Choosing Elements for the Legend	270
Legend for Size of Points	272
Multiple Legends	274
<b>30. Customizing Colorbars.....</b>	<b>276</b>
Customizing Colorbars	277
Choosing the Colormap	278
Color Limits and Extensions	280
Discrete Colorbars	281
Example: Handwritten Digits	282
<b>31. Multiple Subplots.....</b>	<b>285</b>
plt.axes: Subplots by Hand	285
plt.subplot: Simple Grids of Subplots	287
plt.subplots: The Whole Grid in One Go	289
plt.GridSpec: More Complicated Arrangements	291

<b>32. Text and Annotation.....</b>	<b>294</b>
Example: Effect of Holidays on US Births	294
Transforms and Text Position	296
Arrows and Annotation	298
<b>33. Customizing Ticks.....</b>	<b>302</b>
Major and Minor Ticks	302
Hiding Ticks or Labels	304
Reducing or Increasing the Number of Ticks	306
Fancy Tick Formats	307
Summary of Formatters and Locators	310
<b>34. Customizing Matplotlib: Configurations and Stylesheets.....</b>	<b>312</b>
Plot Customization by Hand	312
Changing the Defaults: rcParams	314
Stylesheets	316
Default Style	317
FiveThirtyEight Style	317
ggplot Style	318
Bayesian Methods for Hackers Style	318
Dark Background Style	319
Grayscale Style	319
Seaborn Style	320
<b>35. Three-Dimensional Plotting in Matplotlib.....</b>	<b>321</b>
Three-Dimensional Points and Lines	322
Three-Dimensional Contour Plots	323
Wireframes and Surface Plots	325
Surface Triangulations	328
Example: Visualizing a Möbius Strip	330
<b>36. Visualization with Seaborn.....</b>	<b>332</b>
Exploring Seaborn Plots	333
Histograms, KDE, and Densities	333
Pair Plots	335
Faceted Histograms	336
Categorical Plots	338
Joint Distributions	339
Bar Plots	340
Example: Exploring Marathon Finishing Times	342
Further Resources	350
Other Python Visualization Libraries	351

## Part V. Machine Learning

<b>37. What Is Machine Learning?.....</b>	<b>355</b>
Categories of Machine Learning	355
Qualitative Examples of Machine Learning Applications	356
Classification: Predicting Discrete Labels	356
Regression: Predicting Continuous Labels	359
Clustering: Inferring Labels on Unlabeled Data	363
Dimensionality Reduction: Inferring Structure of Unlabeled Data	364
Summary	366
<b>38. Introducing Scikit-Learn.....</b>	<b>367</b>
Data Representation in Scikit-Learn	367
The Features Matrix	368
The Target Array	368
The Estimator API	370
Basics of the API	371
Supervised Learning Example: Simple Linear Regression	372
Supervised Learning Example: Iris Classification	375
Unsupervised Learning Example: Iris Dimensionality	376
Unsupervised Learning Example: Iris Clustering	377
Application: Exploring Handwritten Digits	378
Loading and Visualizing the Digits Data	378
Unsupervised Learning Example: Dimensionality Reduction	380
Classification on Digits	381
Summary	383
<b>39. Hyperparameters and Model Validation.....</b>	<b>384</b>
Thinking About Model Validation	384
Model Validation the Wrong Way	385
Model Validation the Right Way: Holdout Sets	385
Model Validation via Cross-Validation	386
Selecting the Best Model	388
The Bias-Variance Trade-off	389
Validation Curves in Scikit-Learn	391
Learning Curves	395
Validation in Practice: Grid Search	400
Summary	401
<b>40. Feature Engineering.....</b>	<b>402</b>
Categorical Features	402

Text Features	404
Image Features	405
Derived Features	405
Imputation of Missing Data	408
Feature Pipelines	409
<b>41. In Depth: Naive Bayes Classification.....</b>	<b>410</b>
Bayesian Classification	410
Gaussian Naive Bayes	411
Multinomial Naive Bayes	414
Example: Classifying Text	414
When to Use Naive Bayes	417
<b>42. In Depth: Linear Regression.....</b>	<b>419</b>
Simple Linear Regression	419
Basis Function Regression	422
Polynomial Basis Functions	422
Gaussian Basis Functions	424
Regularization	425
Ridge Regression ( $L_2$ Regularization)	427
Lasso Regression ( $L_1$ Regularization)	428
Example: Predicting Bicycle Traffic	429
<b>43. In Depth: Support Vector Machines.....</b>	<b>435</b>
Motivating Support Vector Machines	435
Support Vector Machines: Maximizing the Margin	437
Fitting a Support Vector Machine	438
Beyond Linear Boundaries: Kernel SVM	441
Tuning the SVM: Softening Margins	444
Example: Face Recognition	445
Summary	450
<b>44. In Depth: Decision Trees and Random Forests.....</b>	<b>451</b>
Motivating Random Forests: Decision Trees	451
Creating a Decision Tree	452
Decision Trees and Overfitting	455
Ensembles of Estimators: Random Forests	456
Random Forest Regression	458
Example: Random Forest for Classifying Digits	459
Summary	462

<b>45. In Depth: Principal Component Analysis.....</b>	<b>463</b>
Introducing Principal Component Analysis	463
PCA as Dimensionality Reduction	466
PCA for Visualization: Handwritten Digits	467
What Do the Components Mean?	469
Choosing the Number of Components	470
PCA as Noise Filtering	471
Example: Eigenfaces	473
Summary	476
<b>46. In Depth: Manifold Learning.....</b>	<b>477</b>
Manifold Learning: “HELLO”	478
Multidimensional Scaling	479
MDS as Manifold Learning	482
Nonlinear Embeddings: Where MDS Fails	484
Nonlinear Manifolds: Locally Linear Embedding	486
Some Thoughts on Manifold Methods	488
Example: Isomap on Faces	489
Example: Visualizing Structure in Digits	493
<b>47. In Depth: k-Means Clustering.....</b>	<b>496</b>
Introducing k-Means	496
Expectation–Maximization	498
Examples	504
Example 1: k-Means on Digits	504
Example 2: k-Means for Color Compression	507
<b>48. In Depth: Gaussian Mixture Models.....</b>	<b>512</b>
Motivating Gaussian Mixtures: Weaknesses of k-Means	512
Generalizing E–M: Gaussian Mixture Models	516
Choosing the Covariance Type	520
Gaussian Mixture Models as Density Estimation	520
Example: GMMs for Generating New Data	524
<b>49. In Depth: Kernel Density Estimation.....</b>	<b>528</b>
Motivating Kernel Density Estimation: Histograms	528
Kernel Density Estimation in Practice	533
Selecting the Bandwidth via Cross-Validation	535
Example: Not-so-Naive Bayes	535
Anatomy of a Custom Estimator	537
Using Our Custom Estimator	539

<b>50. Application: A Face Detection Pipeline.....</b>	<b>541</b>
HOG Features	542
HOG in Action: A Simple Face Detector	543
1. Obtain a Set of Positive Training Samples	543
2. Obtain a Set of Negative Training Samples	543
3. Combine Sets and Extract HOG Features	545
4. Train a Support Vector Machine	546
5. Find Faces in a New Image	546
Caveats and Improvements	548
Further Machine Learning Resources	550
<b>Index.....</b>	<b>551</b>