

## Datenanalyse mit Python

Auswertung von Daten mit pandas,  
NumPy und Jupyter

# DAS INHALTS- VERZEICHNIS

» Hier geht's  
direkt  
zum Buch

<b>Vorwort</b> .....	<b>13</b>
<b>1 Einleitung</b> .....	<b>19</b>
1.1 Worum geht es in diesem Buch? .....	19
Welche Arten von Daten? .....	19
1.2 Warum Python für die Datenanalyse? .....	20
Python als Kleister .....	21
Das »Zwei-Sprachen-Problem« lösen .....	21
Warum nicht Python? .....	22
1.3 Grundlegende Python-Bibliotheken .....	22
NumPy .....	22
pandas .....	23
matplotlib .....	24
IPython und Jupyter .....	25
SciPy .....	26
scikit-learn .....	26
statsmodels .....	27
Andere Pakete .....	27
1.4 Installation und Einrichtung .....	28
Miniconda auf Windows .....	28
GNU/Linux .....	29
Miniconda auf macOS .....	29
Python-Pakete installieren oder aktualisieren .....	30
Integrierte Entwicklungsumgebungen und Texteditoren .....	31
1.5 Community und Konferenzen .....	32
1.6 Navigation durch dieses Buch .....	33
Codebeispiele .....	33
Daten für die Beispiele .....	34
Importkonventionen .....	35

<b>2</b>	<b>Grundlagen von Python, IPython und Jupyter-Notebooks</b>	<b>37</b>
2.1	Der Python-Interpreter	38
2.2	IPython-Grundlagen	39
	Die IPython-Shell ausführen	39
	Das Jupyter-Notebook ausführen	40
	Befehlsergänzung mit Tab	43
	Introspektion	44
2.3	Grundlagen der Sprache Python	45
	Sprachsemantik	46
	Skalare Typen	54
	Kontrollfluss	61
2.4	Schlussbemerkung	65
<b>3</b>	<b>In Python integrierte Datenstrukturen, Funktionen und Dateien</b>	<b>67</b>
3.1	Datenstrukturen und Sequenzen	67
	Tupel	67
	Listen	71
	Dictionarys	75
	Set	79
	Eingebaute Funktionen von Sequenzen	81
	List, Set und Dictionary Comprehensions	83
3.2	Funktionen	85
	Namensraum, Gültigkeitsbereich und lokale Funktionen	86
	Mehrere Rückgabewerte	88
	Funktionen sind Objekte	88
	Anonyme oder Lambda-Funktionen	90
	Generatoren	91
	Fehler und die Behandlung von Ausnahmen	94
3.3	Dateien und das Betriebssystem	96
	Bytes und Unicode mit Dateien	100
3.4	Schlussbemerkung	102
<b>4</b>	<b>Grundlagen von NumPy: Arrays und vektorisierte Berechnung</b>	<b>103</b>
4.1	Das ndarray von NumPy: ein mehrdimensionales Array-Objekt	105
	ndarrays erzeugen	107
	Datentypen für ndarrays	109
	Rechnen mit NumPy-Arrays	112
	Einfaches Indizieren und Slicing	113
	Boolesches Indizieren	118
	Fancy Indexing	120
	Arrays transponieren und Achsen tauschen	122

4.2	Erzeugen von Pseudozufallszahlen . . . . .	123
4.3	Universelle Funktionen: schnelle elementweise Array-Funktionen . . . . .	125
4.4	Array-orientierte Programmierung mit Arrays . . . . .	128
	Bedingte Logik als Array-Operationen ausdrücken . . . . .	130
	Mathematische und statistische Methoden . . . . .	131
	Methoden für boolesche Arrays . . . . .	133
	Sortieren . . . . .	133
	Unique und andere Mengenlogik . . . . .	135
4.5	Dateiein- und -ausgabe bei Arrays . . . . .	136
4.6	Lineare Algebra . . . . .	136
4.7	Beispiel: Random Walks . . . . .	138
	Viele Random Walks auf einmal simulieren . . . . .	140
4.8	Schlussbemerkung . . . . .	141
<b>5</b>	<b>Erste Schritte mit pandas . . . . .</b>	<b>143</b>
5.1	Einführung in die pandas-Datenstrukturen . . . . .	144
	Series . . . . .	144
	DataFrame . . . . .	148
	Indexobjekte . . . . .	156
5.2	Wesentliche Funktionalität . . . . .	158
	Neuindizierung . . . . .	158
	Einträge von einer Achse löschen . . . . .	161
	Indizierung, Auswahl und Filterung . . . . .	162
	Fallstricke bei Integer-Indizes . . . . .	169
	Arithmetik und Datenausrichtung . . . . .	172
	Funktionsanwendung und Mapping . . . . .	178
	Sortieren und Rangbildung . . . . .	180
	Achsenindizes mit duplizierten Labels . . . . .	184
5.3	Zusammenfassen und Berechnen deskriptiver Statistiken . . . . .	185
	Korrelation und Kovarianz . . . . .	188
	Eindeutigkeit, Werteanzahl und Mitgliedschaft . . . . .	190
5.4	Schlussbemerkung . . . . .	193
<b>6</b>	<b>Laden und Speichern von Daten sowie Dateiformate . . . . .</b>	<b>195</b>
6.1	Lesen und Schreiben von Daten im Textformat . . . . .	195
	Stückweises Lesen von Textdateien . . . . .	202
	Daten in Textformaten schreiben . . . . .	204
	Arbeiten mit anderen Formaten . . . . .	205
	JSON-Daten . . . . .	207
	XML und HTML: Web-Scraping . . . . .	209

6.2	Binäre Datenformate . . . . .	213
	Lesen von Microsoft-Excel-Dateien . . . . .	214
	Benutzung von HDF5 . . . . .	215
6.3	Interaktion mit Web-APIs . . . . .	218
6.4	Interaktion mit Datenbanken . . . . .	220
6.5	Schlussbemerkung . . . . .	221
<b>7</b>	<b>Daten bereinigen und vorbereiten . . . . .</b>	<b>223</b>
7.1	Der Umgang mit fehlenden Daten . . . . .	223
	Fehlende Daten herausfiltern . . . . .	225
	Fehlende Daten einsetzen . . . . .	227
7.2	Datentransformation . . . . .	229
	Duplikate entfernen . . . . .	229
	Daten mithilfe einer Funktion oder eines Mappings transformieren . . . . .	231
	Werte ersetzen . . . . .	232
	Achsenindizes umbenennen . . . . .	234
	Diskretisierung und Klassifizierung . . . . .	235
	Erkennen und Filtern von Ausreißern . . . . .	237
	Permutation und zufällige Stichproben . . . . .	238
	Berechnen von Indikator-/Platzhaltervariablen . . . . .	240
7.3	Extension-Datentypen . . . . .	243
7.4	Manipulation von Strings . . . . .	247
	Methoden von String-Objekten in Python . . . . .	247
	Reguläre Ausdrücke . . . . .	249
	String-Funktionen in pandas . . . . .	252
7.5	Kategorische Daten . . . . .	255
	Hintergrund und Motivation . . . . .	255
	Der Extension-Typ Categorical in pandas . . . . .	257
	Berechnungen mit Categoricals . . . . .	260
	Kategorische Methoden . . . . .	262
7.6	Schlussbemerkung . . . . .	265
<b>8</b>	<b>Datenaufbereitung: Verknüpfen, Kombinieren und Umformen . . . . .</b>	<b>267</b>
8.1	Hierarchische Indizierung . . . . .	267
	Ebenen neu anordnen und sortieren . . . . .	270
	Zusammenfassende Statistiken nach Ebene . . . . .	271
	Indizierung mit den Spalten eines DataFrame . . . . .	272
8.2	Kombinieren und Verknüpfen von Datensätzen . . . . .	273
	Datenbankartige Verknüpfung von DataFrames . . . . .	273
	Daten über einen Index verknüpfen . . . . .	279
	Verketteten entlang einer Achse . . . . .	283
	Überlappende Daten zusammenführen . . . . .	288

8.3	Umformen und Transponieren . . . . .	290
	Umformen mit hierarchischer Indizierung . . . . .	290
	Transponieren vom »langen« zum »breiten« Format . . . . .	293
	Transponieren vom »breiten« zum »langen« Format . . . . .	296
8.4	Schlussbemerkung . . . . .	298
<b>9</b>	<b>Plotten und Visualisieren . . . . .</b>	<b>299</b>
9.1	Kurze Einführung in die matplotlib-API . . . . .	300
	Diagramme und Subplots . . . . .	301
	Farben, Beschriftungen und Linienformen . . . . .	305
	Skalenstriche, Beschriftungen und Legenden . . . . .	307
	Annotationen und Zeichnungen in einem Subplot . . . . .	310
	Diagramme in Dateien abspeichern . . . . .	312
	Die Konfiguration von matplotlib . . . . .	313
9.2	Plotten mit pandas und seaborn. . . . .	313
	Liniendiagramme. . . . .	314
	Balkendiagramme . . . . .	316
	Histogramme und Dichteplots . . . . .	322
	Streu- oder Punktdiagramme. . . . .	324
	Facettenraster und kategorische Daten . . . . .	326
9.3	Andere Visualisierungswerkzeuge in Python . . . . .	328
9.4	Schlussbemerkung . . . . .	329
<b>10</b>	<b>Aggregation von Daten und Gruppenoperationen . . . . .</b>	<b>331</b>
10.1	Grundlagen zu Gruppierungsoperationen . . . . .	332
	Iteration über Gruppen . . . . .	336
	Auswählen einer Spalte oder einer Teilmenge von Spalten. . . . .	337
	Gruppieren mit Dictionarys und Series . . . . .	338
	Gruppieren mit Funktionen . . . . .	339
	Gruppieren nach Ebenen eines Index . . . . .	340
10.2	Aggregation von Daten. . . . .	341
	Spaltenweise und mehrfache Anwendung von Funktionen . . . . .	343
	Aggregierte Daten ohne Zeilenindizes zurückgeben . . . . .	346
10.3	Apply: Allgemeine Operationen vom Typ split-apply-combine. . . . .	347
	Unterdrücken der Gruppenschlüssel. . . . .	349
	Analyse von Quantilen und Größenklassen . . . . .	349
	Beispiel: Fehlende Daten mit gruppenspezifischen Werten auffüllen. . . . .	352
	Beispiel: Zufällige Stichproben und Permutation . . . . .	354
	Beispiel: Gewichteter Mittelwert für Gruppen und Korrelation. . . . .	356
	Beispiel: Gruppenweise lineare Regression . . . . .	358

10.4	Gruppentransformationen und »ausgepackte« GroupBys . . . . .	358
10.5	Pivot-Tabellen und Kreuztabellierung. . . . .	362
	Kreuztabellen . . . . .	365
10.6	Schlussbemerkung. . . . .	366
<b>11</b>	<b>Zeitreihen</b> . . . . .	<b>367</b>
11.1	Datentypen und Werkzeuge für Datum und Zeit . . . . .	368
	Konvertieren zwischen String und datetime . . . . .	369
11.2	Grundlagen von Zeitreihen . . . . .	372
	Indizieren, auswählen und Untermengen bilden . . . . .	373
	Zeitreihen mit doppelten Indizes . . . . .	375
11.3	Datumsbereiche, Frequenzen und Verschiebungen . . . . .	376
	Erzeugen von Datumsbereichen . . . . .	377
	Frequenzen und Offsets von Kalenderdaten. . . . .	380
	Verschieben von Datumsangaben (Vorlauf und Verzögerung) . . . . .	381
11.4	Berücksichtigung von Zeitzonen . . . . .	384
	Lokalisieren und Konvertieren von Zeitzonen . . . . .	385
	Operationen mit Zeitstempeln bei zugeordneter Zeitzone . . . . .	387
	Operationen zwischen unterschiedlichen Zeitzonen . . . . .	388
11.5	Perioden und Arithmetik von Perioden . . . . .	389
	Umwandlung der Frequenz von Perioden . . . . .	390
	Quartalsweise Perioden . . . . .	392
	Zeitstempel zu Perioden konvertieren (und zurück) . . . . .	393
	Erstellen eines PeriodIndex aus Arrays. . . . .	395
11.6	Resampling und Konvertieren von Frequenzen. . . . .	396
	Downsampling . . . . .	398
	Upsampling und Interpolation . . . . .	400
	Resampling mit Perioden . . . . .	402
	Gruppiertes Zeit-Resampling . . . . .	403
11.7	Funktionen mit gleitenden Fenstern . . . . .	405
	Exponentiell gewichtete Funktionen . . . . .	408
	Binäre Funktionen mit gleitendem Fenster. . . . .	409
	Benutzerdefinierte Funktionen mit gleitenden Fenstern . . . . .	411
11.8	Schlussbemerkung. . . . .	412
<b>12</b>	<b>Einführung in Modellierungsbibliotheken in Python</b> . . . . .	<b>413</b>
12.1	Die Kopplung zwischen pandas und dem Modellcode . . . . .	413
12.2	Modellbeschreibungen mit Patsy herstellen . . . . .	416
	Datentransformationen in Patsy-Formeln . . . . .	419
	Kategorische Daten und Patsy . . . . .	420

12.3	Einführung in statsmodels . . . . .	423
	Lineare Modelle schätzen . . . . .	424
	Zeitreihenprozesse schätzen . . . . .	427
12.4	Einführung in scikit-learn. . . . .	428
12.5	Schlussbemerkung . . . . .	431
<b>13</b>	<b>Beispiele aus der Datenanalyse . . . . .</b>	<b>433</b>
13.1	Bitly-Daten von 1.USA.gov . . . . .	433
	Zählen von Zeitzonen in reinem Python . . . . .	434
	Zeitzone mit pandas zählen. . . . .	436
13.2	MovieLens-1M-Datensatz . . . . .	443
	Messen von Unterschieden in der Bewertung . . . . .	447
13.3	US-Babynamen von 1880–2010 . . . . .	450
	Namenstrends analysieren. . . . .	455
13.4	Die USDA-Nahrungsmitteldatenbank . . . . .	464
13.5	Datenbank des US-Wahlausschusses von 2012 . . . . .	469
	Spendenstatistik nach Beruf und Arbeitgeber . . . . .	472
	Spenden der Größe nach klassifizieren . . . . .	475
	Spendenstatistik nach Bundesstaat . . . . .	477
13.6	Schlussbemerkung . . . . .	478
<b>Anhang A</b>	<b>NumPy für Fortgeschrittene . . . . .</b>	<b>479</b>
A.1	Interna des ndarray-Objekts. . . . .	479
	Die Datentyphierarchie in NumPy . . . . .	480
A.2	Fortgeschrittene Manipulation von Arrays . . . . .	482
	Arrays umformen. . . . .	482
	Anordnung von Arrays in C und FORTRAN . . . . .	484
	Arrays verketteten und aufspalten . . . . .	485
	Wiederholen von Elementen: tile und repeat . . . . .	487
	Alternativen zum Fancy Indexing: take und put . . . . .	489
A.3	Broadcasting. . . . .	490
	Broadcasting über andere Achsen . . . . .	492
	Werte von Arrays durch Broadcasting setzen . . . . .	494
A.4	Fortgeschrittene Nutzung von ufuncs . . . . .	495
	Instanzmethoden von ufunc . . . . .	495
	Neue ufuncs in Python schreiben . . . . .	498
A.5	Strukturierte und Record-Arrays . . . . .	499
	Geschachtelte Datentypen und mehrdimensionale Felder . . . . .	499
	Warum sollte man strukturierte Arrays verwenden? . . . . .	500
A.6	Mehr zum Thema Sortieren . . . . .	501
	Indirektes Sortieren: argsort und lexsort . . . . .	502
	Alternative Sortieralgorithmen . . . . .	503

Arrays teilweise sortieren . . . . .	504
numpy.searchsorted: Elemente in einem sortierten Array finden . . . . .	505
A.7 Schnelle NumPy-Funktionen mit Numba schreiben. . . . .	506
Eigene numpy.ufunc-Objekte mit Numba herstellen. . . . .	508
A.8 Ein- und Ausgabe von Arrays für Fortgeschrittene . . . . .	508
Memory-mapped Dateien. . . . .	509
HDF5 und weitere Möglichkeiten zum Speichern von Arrays . . . . .	510
A.9 Tipps für eine höhere Leistung . . . . .	510
Die Bedeutung des zusammenhängenden Speichers . . . . .	511
<b>Anhang B Mehr zum IPython-System . . . . .</b>	<b>513</b>
B.1 Tastenkürzel im Terminal . . . . .	513
B.2 Magische Befehle . . . . .	514
Der Befehl %run . . . . .	516
Code aus der Zwischenablage ausführen . . . . .	517
B.3 Die Befehlshistorie benutzen . . . . .	518
Die Befehlshistorie durchsuchen und wiederverwenden . . . . .	518
Eingabe- und Ausgabevariablen . . . . .	519
B.4 Mit dem Betriebssystem interagieren . . . . .	520
Shell-Befehle und -Aliase . . . . .	521
Das Verzeichnis-Bookmark-System . . . . .	522
B.5 Werkzeuge zur Softwareentwicklung . . . . .	523
Interaktiver Debugger . . . . .	523
Zeitmessung bei Code: %time und %timeit . . . . .	528
Grundlegende Profilierung: %prun and %run -p . . . . .	529
Eine Funktion Zeile für Zeile profilieren . . . . .	531
B.6 Tipps für eine produktive Codeentwicklung mit IPython. . . . .	533
Modulabhängigkeiten neu laden . . . . .	534
Tipps für das Codedesign . . . . .	535
B.7 Fortgeschrittene IPython-Funktionen . . . . .	536
Profile und Konfiguration . . . . .	536
B.8 Schlussbemerkung. . . . .	538
<b>Index . . . . .</b>	<b>539</b>