Vorwort . . . . . **XIII**

1 Einleitung **1**
1.1 Worum geht es in diesem Buch? 1
Welche Arten von Daten? 1
1.2 Warum Python für die Datenanalyse? 2
Python als Kleister 2
Das »Zwei-Sprachen-Problem« lösen 3
Warum nicht Python? 3
1.3 Grundlegende Python-Bibliotheken 4
NumPy 4
pandas 5
matplotlib 6
IPython und Jupyter 6
SciPy 7
scikit-learn 8
statsmodels 8
1.4 Installation und Einrichtung 9
Windows 9
Apple (OS X, macOS) 9
GNU/Linux 10
Python-Pakete installieren oder aktualisieren 10
Python 2 und Python 3 11
Integrierte Entwicklungsumgebungen (Integrated Development Environments – IDEs) und Texteditoren 12
1.5 Community und Konferenzen 12
1.6 Navigation durch dieses Buch 13
Codebeispiele 14
Daten für die Beispiele 14
Importkonventionen 14
Jargon 15

Bibliografische Informationen
http://d-nb.info/1165334224

digitalisiert durch DEUTSCHE NATIONAL BIBLIOTHEK

2 Grundlagen von Python, IPython und Jupyter-Notebooks 17
2.1 Der Python-Interpreter 18
2.2 IPython-Grundlagen 19
Die IPython-Shell ausführen 19
Das Jupyter-Notebook ausführen 20
Befehlsergänzung mit Tab 23
Introspektion 24
Der %run-Befehl 26
Code aus der Zwischenablage ausführen 27
Terminal-Tastenkürzel 28
Über magische Befehle 29
matplotlib-Integration 31
2.3 Grundlagen der Sprache Python 32
Sprachsemantik 32
Skalare Typen 41
Kontrollfluss 48

3 In Python integrierte Datenstrukturen, Funktionen und Dateien 53
3.1 Datenstrukturen und Sequenzen 53
Tupel 53
Listen 56
Eingebaute Funktionen von Sequenzen 61
Dictionarys 63
Set 67
List, Set und Dict Comprehensions 69
3.2 Funktionen 71
Namensraum, Gültigkeitsbereich und lokale Funktionen 72
Mehrere Rückgabewerte 73
Funktionen sind Objekte 74
Anonyme oder Lambda-Funktionen 75
Currying: teilweise Anwendung von Argumenten 76
Generatoren 77
Fehler und die Behandlung von Ausnahmen 79
3.3 Dateien und das Betriebssystem 82
Bytes und Unicode mit Dateien 85
3.4 Schlussbemerkung 87

4 Grundlagen von NumPy: Arrays und vektorisierte Berechnung 89
4.1 Das ndarray von NumPy: ein mehrdimensionales Array-Objekt 91
ndarrays erzeugen 92
Datentypen für ndarrays 94
Rechnen mit NumPy-Arrays 97

Einfaches Indizieren und Slicing ... 98
Boolesches Indizieren ... 103
Fancy Indexing ... 106
Arrays transponieren und Achsen tauschen ... 107
4.2 Universelle Funktionen: schnelle elementweise Array-Funktionen ... 109
4.3 Array-orientierte Programmierung mit Arrays ... 112
Bedingte Logik als Array-Operationen ausdrücken ... 114
Mathematische und statistische Methoden ... 115
Methoden für boolesche Arrays ... 117
Sortieren ... 117
Unique und andere Mengenlogik ... 118
4.4 Dateiein- und -ausgabe bei Arrays ... 119
4.5 Lineare Algebra ... 120
4.6 Erzeugen von Pseudozufallszahlen ... 122
4.7 Beispiel: Random Walks ... 124
Viele Random Walks auf einmal simulieren ... 125
4.8 Schlussbemerkung ... 126

5 Erste Schritte mit pandas ... 127
5.1 Einführung in die pandas-Datenstrukturen ... 127
Series ... 128
DataFrame ... 132
Indexobjekte ... 138
5.2 Wesentliche Funktionalität ... 140
Neuindizierung ... 140
Einträge von einer Achse löschen ... 142
Indizierung, Auswahl und Filterung ... 144
Integer-Indizes ... 149
Arithmetik und Datenausrichtung ... 150
Funktionsanwendung und Mapping ... 155
Sortieren und Rangbildung ... 157
Achsenindizes mit duplizierten Labels ... 160
5.3 Zusammenfassen und Berechnen deskriptiver Statistiken ... 162
Korrelation und Kovarianz ... 164
Eindeutigkeit, Werteanzahl und Mitgliedschaft ... 166
5.4 Schlussbemerkung ... 169

6 Laden und Speichern von Daten sowie Dateiformate ... 171
6.1 Lesen und Schreiben von Daten im Textformat ... 171
Stückweises Lesen von Textdateien ... 177
Daten in Textformaten schreiben ... 179
Arbeiten mit separierten Formaten ... 180

JSON-Daten ... 182
XML und HTML: Web-Scraping ... 184
6.2 Binäre Datenformate ... 187
Benutzung von HDF5 ... 188
Lesen von Microsoft Excel-Dateien ... 190
6.3 Interaktion mit Web-APIs ... 191
6.4 Interaktion mit Datenbanken ... 192
6.5 Schlussbemerkung ... 194

7 Daten bereinigen und vorbereiten ... 195
7.1 Der Umgang mit fehlenden Daten ... 195
Fehlende Daten herausfiltern ... 197
Fehlende Daten einsetzen ... 199
7.2 Datentransformation ... 201
Duplikate entfernen ... 201
Daten mithilfe einer Funktion oder eines Mappings transformieren ... 203
Werte ersetzen ... 204
Achsenindizes umbenennen ... 206
Diskretisierung und Klassifizierung ... 207
Erkennen und Filtern von Ausreißern ... 209
Permutation und zufällige Stichproben ... 211
Berechnen von Indikator-/Platzhaltervariablen ... 212
7.3 Manipulation von Strings ... 215
Methoden von String-Objekten ... 215
Reguläre Ausdrücke ... 217
Vektorisierte String-Funktionen in pandas ... 220
7.4 Schlussbemerkung ... 223

8 Datenaufbereitung: Verknüpfen, Kombinieren und Umformen ... 225
8.1 Hierarchische Indizierung ... 225
Ebenen neu anordnen und sortieren ... 228
Zusammenfassende Statistiken nach Ebene ... 229
Indizierung mit den Spalten eines DataFrame ... 229
8.2 Kombinieren und Verknüpfen von Datensätzen ... 231
Datenbankartige Verknüpfung von DataFrames ... 231
Daten über einen Index verknüpfen ... 236
Verketten entlang einer Achse ... 240
Überlappende Daten zusammenführen ... 245
8.3 Umformen und Transponieren ... 246
Umformen mit hierarchischer Indizierung ... 246
Transponieren vom »langen« zum »breiten« Format ... 249

Transponieren vom »breiten« zum »langen« Format 252
8.4 Schlussbemerkung 254

9 Plotten und Visualisieren 255
9.1 Kurze Einführung in die matplotlib-API 256
Diagramme und Subplots 257
Farben, Beschriftungen und Linienformen 261
Skalenstriche, Beschriftungen und Legenden 263
Annotationen und Zeichnungen in einem Subplot 267
Diagramme in Dateien abspeichern 269
Die Konfiguration von matplotlib 270
9.2 Plotten mit pandas und seaborn 271
Liniendiagramme 271
Balkendiagramme 274
Histogramme und Dichteplots 279
Streu- oder Punktdiagramme 281
Facettenraster und kategorische Daten 283
9.3 Andere Visualisierungswerkzeuge in Python 285
9.4 Schlussbemerkung 286

10 Aggregation von Daten und Gruppenoperationen 287
10.1 GroupBy-Mechanismen 288
Iteration über Gruppen 291
Auswählen einer Spalte oder einer Teilmenge von Spalten 293
Gruppieren mit Dictionarys und Series 293
Gruppieren mit Funktionen 295
Gruppieren nach Ebenen eines Index 295
10.2 Aggregation von Daten 296
Spaltenweise und mehrfache Anwendung von Funktionen 298
Aggregierte Daten ohne Zeilenindizes zurückgeben 301
10.3 Apply: Allgemeine Operationen vom Typ split-apply-combine 302
Unterdrücken der Gruppenschlüssel 304
Analyse von Quantilen und Größenklassen 305
Beispiel: Fehlende Daten mit gruppenspezifischen Werten auffüllen 306
Beispiel: Zufällige Stichproben und Permutation 308
Beispiel: Gewichteter Mittelwert für Gruppen und Korrelation 310
Beispiel: Gruppenweise lineare Regression 312

10.4 Pivot-Tabellen und Kreuztabellierung ... 312
Kreuztabellen ... 315
10.5 Schlussbemerkung ... 316

11 Zeitreihen ... 317
11.1 Datentypen und Werkzeuge für Datum und Zeit ... 318
Konvertieren zwischen String und datetime ... 319
11.2 Grundlagen von Zeitreihen ... 322
Indizieren, auswählen und Untermengen bilden ... 323
Zeitreihen mit doppelten Indizes ... 326
11.3 Datumsbereiche, Frequenzen und Verschiebungen ... 327
Erzeugen von Datumsbereichen ... 328
Frequenzen und Offsets von Kalenderdaten ... 330
Verschieben von Datumsangaben (Vorlauf und Verzögerung) ... 332
11.4 Berücksichtigung von Zeitzonen ... 335
Lokalisieren und Konvertieren von Zeitzonen ... 335
Operationen mit Zeitstempeln bei zugeordneter Zeitzone ... 338
Operationen zwischen unterschiedlichen Zeitzonen ... 339
11.5 Perioden und Arithmetik von Perioden ... 339
Umwandlung der Frequenz von Perioden ... 340
Quartalsweise Perioden ... 342
Zeitstempel zu Perioden konvertieren (und zurück) ... 344
Erstellen eines PeriodIndex aus Arrays ... 345
11.6 Resampling und Konvertieren von Frequenzen ... 347
Downsampling ... 349
Upsampling und Interpolation ... 352
Resampling mit Perioden ... 353
11.7 Funktionen mit gleitenden Fenstern ... 354
Exponentiell gewichtete Funktionen ... 358
Binäre Funktionen mit gleitendem Fenster ... 359
Benutzerdefinierte Funktionen mit gleitenden Fenstern ... 360
11.8 Schlussbemerkung ... 361

12 pandas für Fortgeschrittene ... 363
12.1 Kategorische Daten ... 363
Hintergrund und Motivation ... 363
Der Typ Categorical in pandas ... 365
Berechnungen mit Categoricals ... 367
Kategorische Methoden ... 370
12.2 Erweiterter Einsatz von GroupBy ... 372
Gruppentransformationen und »ausgepackte« GroupBys ... 373
Gruppiertes Zeit-Resampling ... 376

12.3 Techniken für die Verkettung von Methoden 378
Die Methode pipe 380
12.4 Schlussbemerkung 380

13 Einführung in Modellierungsbibliotheken in Python 383
13.1 Die Kopplung zwischen pandas und dem Modellcode 383
13.2 Modellbeschreibungen mit Patsy herstellen 386
Datentransformationen in Patsy-Formeln 389
Kategorische Daten und Patsy 390
13.3 Einführung in statsmodels 393
Lineare Modelle schätzen 393
Zeitreihenprozesse schätzen 396
13.4 Einführung in scikit-learn 397
13.5 Ihre Ausbildung fortsetzen 401

14 Beispiele aus der Datenanalyse 403
14.1 1.USA.gov-Daten von Bitly 403
Zählen von Zeitzonen in reinem Python 404
Zeitzonen mit pandas zählen 406
14.2 MovieLens-1M-Datensatz 413
Messen von Unterschieden in der Bewertung 418
14.3 US-Babynamen von 1880–2010 419
Namenstrends analysieren 424
14.4 Die USDA-Nahrungsmitteldatenbank 433
14.5 Datenbank des US-Wahlausschusses von 2012 439
Spendenstatistik nach Beruf und Arbeitgeber 441
Spenden der Größe nach klassifizieren 444
Spendenstatistik nach Bundesstaat 446
14.6 Schlussbemerkung 447

A NumPy für Fortgeschrittene 449
A.1 Interna des ndarray-Objekts 449
Die dtype-Hierarchie in NumPy 450
A.2 Fortgeschrittene Manipulation von Arrays 451
Arrays umformen 452
Anordnung von Arrays in C und Fortran 454
Arrays verketten und aufspalten 454
Wiederholen von Elementen: tile und repeat 457
Alternativen zum Fancy Indexing: take und put 459
A.3 Broadcasting 460
Broadcasting über andere Achsen 462
Werte von Arrays durch Broadcasting setzen 465

A.4 Fortgeschrittene Nutzung von ufuncs 465
Instanzmethoden von ufunc 466
Neue ufuncs in Python schreiben 468
A.5 Strukturierte und Record-Arrays 469
Geschachtelte dtypes und mehrdimensionale Felder 469
Warum sollte man strukturierte Arrays verwenden? 470
A.6 Mehr zum Thema Sortieren 471
Indirektes Sortieren: argsort und lexsort 472
Alternative Sortieralgorithmen 474
Arrays teilweise sortieren 474
numpy.searchsorted: Elemente in einem sortierten Array finden 475
A.7 Schnelle NumPy-Funktionen mit Numba schreiben 476
Eigene numpy.ufunc-Objekte mit Numba herstellen 478
A.8 Ein- und Ausgabe von Arrays für Fortgeschrittene 478
Memory-mapped Dateien 478
HDF5 und weitere Möglichkeiten zum Speichern von Arrays 480
A.9 Tipps für eine höhere Leistung 480
Die Bedeutung des zusammenhängenden Speichers 480

B Mehr zum IPython-System 483
B.1 Die Befehlshistorie benutzen 483
Die Befehlshistorie durchsuchen und wiederverwenden 483
Eingabe- und Ausgabevariablen 484
B.2 Mit dem Betriebssystem interagieren 485
Shell-Befehle und -Aliase 486
Das Verzeichnis-Bookmark-System 487
B.3 Werkzeuge zur Softwareentwicklung 487
Interaktiver Debugger 488
Zeitmessung bei Code: %time und %timeit 492
Grundlegende Profilierung: %prun and %run -p 494
Eine Funktion Zeile für Zeile profilieren 496
B.4 Tipps für eine produktive Codeentwicklung mit IPython 498
Modulabhängigkeiten neu laden 499
Tipps für das Codedesign 499
B.5 Fortgeschrittene IPython-Funktionen 501
Ihre eigenen Klassen IPython-freundlich gestalten 501
Profile und Konfiguration 502
B.6 Schlussbemerkung 503

Index 505