Datenanalyse mit Python

Auswertung von Daten mit pandas, NumPy und Jupyter

DAS VORWORT

Vorwort

Die erste (englischsprachige) Auflage dieses Buchs wurde 2012 veröffentlicht, als die Open-Source-Bibliotheken zur Datenanalyse mit Python (insbesondere pandas) ganz neu waren und sich rasant weiterentwickelten. Als es an der Zeit war, 2016/2017 die zweite Auflage zu schreiben, musste ich das Buch nicht nur an Python 3.6 anpassen (in der ersten Auflage kam noch Python 2.7 zum Einsatz), sondern auch den neuen Funktionalitäten Rechnung tragen, die sich in den fünf Jahren dazwischen entwickelt haben. Jetzt ist es 2022, und es gab weniger Änderungen an Python (wir sind aktuell bei Erscheinen dieses Buchs bei Python 3.11), aber pandas hat sich stets weiterentwickelt.

In dieser dritten Auflage ist es mein Ziel, die Inhalte an die aktuellen Versionen von Python, NumPy, pandas und anderen Projekten anzupassen, dabei aber in Bezug auf neuere Python-Projekte aus den letzten paar Jahren eher zurückhaltend vorzugehen. Da dieses Buch für viele Vorlesungen an Universitäten und für Experten in ihrem beruflichen Alltag zu einer wichtigen Quelle geworden ist, möchte ich Themen vermeiden, die eventuell in ein oder zwei Jahren schon wieder unwichtig geworden sind. So sollte sich das Buch auch noch 2023 oder 2024 gut nutzen lassen.

Ein neues Feature der dritten Auflage ist die (englischsprachige) Open-Access-Onlineversion auf meiner Website unter *https://wesmckinney.com/book*, die als Ressource und praktischer Rückgriff für Besitzer der Papier- oder Digitalversion dieses Buchs dient. Ich plane, den Inhalt dort möglichst aktuell zu halten – wenn Sie also die gedruckte Version dieses Buchs besitzen und über etwas stolpern, das nicht richtig funktioniert, sollten Sie dort nachschauen, ob sich etwas geändert hat.

Konventionen in diesem Buch

Folgende typografische Konventionen gelten in diesem Buch:

Kursiv
: Kennzeichnet neue Begriffe, URLs, E-Mail-Adressen, Dateinamen und Dateierweiterungen.

`Nichtproportionalschrift`
: Kennzeichnet Programmlistings sowie Programmelemente in Absätzen, wie etwa Variablen- oder Funktionsnamen, Datenbanken, Datentypen, Umgebungsvariablen, Anweisungen und Schlüsselwörter.

`Nichtproportionalschrift fett`
: Stellt Befehle oder anderen Text dar, der wortwörtlich vom Benutzer eingetippt werden sollte.

`Nichtproportionalschrift kursiv`
: Zeigt Text, der durch Werte ersetzt werden soll, die der Benutzer vorgibt oder die sich aus dem Kontext ergeben.

Dieses Symbol kennzeichnet einen Tipp oder Vorschlag.

Hinter diesem Symbol verbirgt sich eine allgemeine Bemerkung.

Dieses Element symbolisiert einen Warnhinweis.

Benutzung von Codebeispielen

Sie finden die Daten und dazugehöriges Material für jedes Kapitel im GitHub-Repository dieses Buchs unter *http://github.com/wesm/pydata-book*, auch gespiegelt nach *https://gitee.com/wesmckinn/pydata-book*, falls Sie keinen Zugriff auf GitHub haben.

Das Buch soll Ihnen bei Ihrer Arbeit helfen. Ganz allgemein gilt: Wenn in diesem Buch Beispielcode angeboten wird, können Sie ihn in Ihren Programmen und Dokumentationen verwenden. Sie müssen sich dafür nicht unsere Erlaubnis einholen, es sei denn, Sie reproduzieren einen großen Teil des Codes. Schreiben Sie zum

Beispiel ein Programm, das mehrere Teile des Codes aus diesem Buch benutzt, brauchen Sie keine Erlaubnis. Verkaufen oder vertreiben Sie Beispiele aus O'Reilly-Büchern, brauchen Sie eine Erlaubnis. Beantworten Sie eine Frage, indem Sie dieses Buch und Beispielcode daraus zitieren, brauchen Sie keine Erlaubnis. Binden Sie einen großen Anteil des Beispielcodes aus diesem Buch in die Dokumentation Ihres Produkts ein, brauchen Sie eine Erlaubnis.

Wir freuen uns über eine Erwähnung, verlangen sie aber nicht. Eine Erwähnung enthält üblicherweise Titel, Autor, Verlag und ISBN, zum Beispiel: »*Datenanalyse mit Python* von Wes McKinney, O'Reilly 2023, ISBN 978-3-96009-211-7.«

Falls Sie befürchten, zu viele Codebeispiele zu verwenden oder die oben genannten Befugnisse zu überschreiten, kontaktieren Sie uns unter *kommentar@oreilly.de*.

Danksagungen

Dieses Werk ist das Produkt aus vielen Jahren der Zusammenarbeit und Hilfe sowie fruchtbarer Diskussionen mit und von Menschen auf der ganzen Welt. Ich möchte einigen von ihnen danken.

In Memoriam: John D. Hunter (1968–2012)

Unser lieber Freund und Kollege John D. Hunter verstarb am 28. August 2012 an Darmkrebs. Erst kurz zuvor hatte ich das Manuskript für die erste Auflage dieses Buchs fertiggestellt.

Man kann Johns Einfluss und Vermächtnis in der wissenschaftlichen Python-Gemeinde nicht hoch genug einschätzen. Er entwickelte nicht nur matplotlib Anfang der 2000er-Jahre (in einer Zeit, als Python nicht annähernd so beliebt war), sondern war auch an der Herausbildung der Kultur einer wichtigen Generation von Open-Source-Entwicklern beteiligt, die zu den Säulen des Python-Ökosystems gehören, das wir heute oft als so selbstverständlich hinnehmen.

Ich hatte das Glück, John zu Beginn meiner Open-Source-Karriere im Januar 2010 kennenzulernen, gerade als pandas 0.1 herausgekommen war. Seine Inspiration und seine Unterstützung halfen mir selbst in den düstersten Zeiten, meine Vision von pandas und Python als erstklassige Datenanalysesprache voranzutreiben.

John stand Fernando Pérez und Brian Granger sehr nahe, die IPython, Jupyter und vielen anderen Initiativen in der Python-Gemeinde den Weg bereiteten. Wir vier hatten gehofft, gemeinsam an einem Buch zu arbeiten, doch am Ende war ich derjenige mit der meisten freien Zeit. Ich bin mir sicher, er wäre stolz auf das gewesen, was wir einzeln und als Gemeinschaft im Laufe der letzten fünf Jahre erreicht haben.

Danksagungen für die dritte Auflage (2022)

Vor mehr als zehn Jahren habe ich mit dem Schreiben der ersten Auflage dieses Buchs begonnen, und vor mehr als 15 Jahren begann meine Reise als Python-Programmierer. Seitdem hat sich viel geändert! Python hat sich von einer relativen Nischensprache für die Datenanalyse zur beliebtesten und am weitesten verbreiteten Sprache entwickelt, die die Mehrzahl (wenn nicht sogar die Mehrheit!) der Arbeiten in den Bereichen Data Science, maschinelles Lernen und künstliche Intelligenz unterstützt.

Ich habe seit 2013 nicht mehr aktiv zum Open-Source-Projekt pandas beigetragen, aber seine weltweite Gemeinschaft ist weiter gewachsen und kann als Modell einer Community-getriebenen Open-Source-Softwareentwicklung dienen. Viele »Next Generation«-Python-Projekte, die mit Tabellendaten arbeiten, modellieren ihre Benutzeroberflächen direkt nach pandas, was zeigt, dass das Projekt einen beständigen Einfluss auf die Entwicklung des Python-Ökosystems der Data Science besitzt.

Ich hoffe, dieses Buch kann weiterhin als wertvolle Quelle für Studierende und viele andere Personen dienen, die daran interessiert sind, etwas zum Arbeiten mit Daten in Python zu lernen.

Besonders dankbar bin ich O'Reilly, dass ich eine »Open Access«-Version dieses Buchs auf meiner Website unter *https://wesmckinney.com/book* bereitstellen kann, sodass hoffentlich noch mehr Menschen erreicht werden können und ihnen dabei geholfen wird, besser in die Welt der Datenanalyse einzusteigen. J. J. Allaire war dabei unverzichtbar, er half mir, das Buch von Docbook XML nach *Quarto* (*https://quarto.org*) zu portieren – einem neuen und wunderbaren Publishing-System (Druck und Web) für Wissenschaft und Technik.

Vielen Dank auch an meine Fachkorrektoren Paul Barry, Jean-Christophe Leyder, Abdullah Karasan und William Jamir, deren umfassendes Feedback die Lesbarkeit, Klarheit und Verständlichkeit dieses Buchs deutlich verbessert hat.

Danksagungen für die zweite Auflage (2017)

Es sind fast auf den Tag genau fünf Jahre vergangen, seit ich im Juli 2012 das Manuskript für die erste Auflage dieses Buchs beendet habe. Eine Menge hat sich geändert. Die Python-Gemeinde ist unglaublich gewachsen, und das sie umgebende Ökosystem der Open-Source-Software gedeiht.

Diese neue Auflage des Buchs hätte es ohne die unablässigen Bemühungen der pandas-Entwickler nicht gegeben, die das Projekt und seine Gemeinschaft zu einem der Eckpfeiler des Python-Data-Science-Ökosystems gemacht haben. Zu ihnen gehören unter anderem Tom Augspurger, Joris Van den Bossche, Chris Bartak, Phillip Cloud, gfyoung, Andy Hayden, Masaaki Horikoshi, Stephan Hoyer, Adam

Klein, Wouter Overmeire, Jeff Reback, Chang She, Skipper Seabold, Jeff Tratner und y-p.

Für ihre Hilfe und Geduld beim Schreiben dieser zweiten Auflage möchte ich den O'Reilly-Mitarbeitern danken: Marie Beaugureau, Ben Lorica und Colleen Toporek. Ihr technisches Expertenwissen brachten Tom Augspurger, Paul Barry, Hugh Brown, Jonathan Coe und Andreas Müller ein. Danke schön.

Die erste Auflage dieses Buchs wurde in viele Sprachen übersetzt, darunter Chinesisch, Französisch, Deutsch, Japanisch, Koreanisch und Russisch. Das Übersetzen des Inhalts, der dadurch einem viel breiteren Publikum zugänglich wird, ist eine gigantische und oft undankbare Aufgabe. Ich danke den Übersetzern, dass sie Menschen auf der ganzen Welt helfen, das Programmieren und die Benutzung von Datenanalysewerkzeugen zu erlernen.

Ich hatte außerdem das Glück, dass mich Cloudera und Two Sigma Investments in den letzten Jahren bei meinen Open-Source-Entwicklungsarbeiten unterstützt haben. Oft sind Open-Source-Projekte trotz einer nicht unbeträchtlichen Benutzerbasis äußerst armselig mit Ressourcen ausgestattet. Deshalb wird es immer wichtiger – und ist auch das einzig Richtige –, dass Unternehmen die Entwicklung von wichtigen Open-Source-Projekten unterstützen.

Danksagungen für die erste Auflage (2012)

Dieses Buch hätte ich ohne die Unterstützung vieler Menschen niemals schreiben können.

Unter den O'Reilly-Mitarbeitern bin ich meinen Lektorinnen Meghan Blanchette und Julie Steele unheimlich dankbar, die mich durch den Prozess begleitet haben. Mike Loukides arbeitete mit mir während der Entwurfsphase zusammen und half mir, das Buch real werden zu lassen.

Viele Menschen haben mich als technische Gutachter unterstützt. Besonders danken möchte ich Martin Blais und Hugh Brown für ihre Hilfe bei den Beispielen für dieses Buch, bei der Übersichtlichkeit und beim Aufbau. James Long, Drew Conway, Fernando Pérez, Brian Granger, Thomas Kluyver, Adam Klein, Josh Klein, Chang She und Stéfan van der Walt haben jeweils ein oder mehrere Kapitel begutachtet, umfangreich kritisiert und von vielen verschiedenen Gesichtspunkten aus beleuchtet.

Diverse großartige Ideen für Beispiele und Datensätze kamen von Freunden und Kollegen in der Datencommunity, darunter Mike Dewar, Jeff Hammerbacher, James Johndrow, Kristian Lum, Adam Klein, Hilary Mason, Chang She und Ashley Williams.

Ich stehe natürlich in der Schuld zahlreicher Pioniere in der wissenschaftlichen Open-Source-Python-Community. Sie haben mir geholfen, das Fundament meiner Entwicklungsarbeit zu legen, und mich beim Schreiben dieses Buchs ermutigt: das

IPython-Kernteam (Fernando Pérez, Brian Granger, Min Ragan-Kelly, Thomas Kluyver und andere), John Hunter, Skipper Seabold, Travis Oliphant, Peter Wang, Eric Jones, Robert Kern, Josef Perktold, Francesc Alted, Chris Fonnesbeck und viele weitere, die hier nicht erwähnt werden können. Verschiedene Menschen gaben mir darüber hinaus ihre Unterstützung sowie Ideen und Ermutigung: Drew Conway, Sean Taylor, Giuseppe Paleologo, Jared Lander, David Epstein, John Krowas, Joshua Bloom, Den Pilsworth, John Myles-White und viele andere, die ich vergessen habe.

Ich möchte außerdem einer Reihe von Menschen aus meinen Lehrjahren danken. Zuallererst danke ich meinen früheren Kollegen bei AQR, die mich über die Jahre bei meiner Arbeit an pandas angefeuert haben: Alex Reyfman, Michael Wong, Tim Sargen, Oktay Kurbanov, Matthew Tschantz, Roni Israelov, Michael Katz, Aris Levine, Chris Uga, Prasad Ramanan, Ted Square und Hoon Kim. Und schließlich danke ich meinen akademischen Lehrmeistern Haynes Miller (MIT) und Mike West (Duke University).

Eine Menge Hilfe bekam ich im Jahr 2014 von Phillip Cloud und Joris Van den Bossche beim Aktualisieren der Codebeispiele in diesem Buch und beim Beheben einiger anderer Ungenauigkeiten, die Änderungen in pandas geschuldet waren.

Auf persönlicher Ebene danke ich Casey, die meinen tagtäglichen Schreibprozess unterstützte und meine Höhen und Tiefen tolerierte, als ich trotz eines überfüllten Terminplans den endgültigen Entwurf zusammenschrieb. Meine Eltern schließlich lehrten mich, immer meinen Träumen zu folgen und mich nie mit weniger zufriedenzugeben.