Benjamin M. Abdel-Karim *Hrsg.*

Data Science

Best Practices mit Python

EBOOK INSIDE

Springer Vieweg

Data Science

Benjamin M. Abdel-Karim
(Hrsg.)

Data Science

Best Practices mit Python

Springer Vieweg

Hrsg.
Benjamin M. Abdel-Karim
Frankfurt am Main, Hessen, Deutschland

ISBN 978-3-658-33459-8 ISBN 978-3-658-33460-4 (eBook)
https://doi.org/10.1007/978-3-658-33460-4

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Planung: Petra Steinmueller
Springer Vieweg ist ein Imprint der eingetragenen Gesellschaft Springer Fachmedien Wiesbaden GmbH und ist ein Teil von Springer Nature.
Die Anschrift der Gesellschaft ist: Abraham-Lincoln-Str. 46, 65189 Wiesbaden, Germany

Vorwort

Im Rahmen meines Studiums der Wirtschaftsinformatik und innerhalb meiner Zeit als wissenschaftlicher Mitarbeiter und Forscher habe ich zahlreiche Probleme mithilfe der Programmierung lösen können, wie zum Beispiel die professionelle Datenanalyse oder das Implementieren von künstlichen Intelligenzen. Vor diesem Hintergrund kann ich gut nachvollziehen, welchen Herausforderungen Einsteiger[1] im Bereich der Programmierung gegenüberstehen. Ausgehend von den zahlreichen Lehrbüchern und Vorlesungen habe ich mir mit diesem Buch das persönliche Ziel gesetzt, ein Werk zu schaffen, das mit simplen Beispielen und einfachen Erklärungen den Zugang zur Programmierung, insbesondere im Bereich der Datenanalyse, ermöglicht. Hierbei möchte ich dem interessierten Leser ein Werk an die Hand geben, das meine wertvollsten Quellcodeartefakte beinhaltet, um somit einen Ansatzpunkt zu schaffen, die eigenen (alltäglichen) Datenanalyseherausforderungen zu lösen.

Ich habe dieses Buch aus der Motivation heraus geschaffen, eines der ersten deutschsprachigen Nachschlagewerke zu entwickeln, in dem relativ simple Quellcodebeispiele enthalten sind, um Lösungsansätze für die (wiederkehrenden) Programmierprobleme in der Datenanalyse weiterzugeben. Hinzu kommt, dass ich selbst wiederholt Lösungswege für ähnliche Probleme erarbeitet habe, die ich bereits in der Vergangenheit entwickelt hatte. Zweifellos gehört das Nachschlagen von Lösungsansätzen in Büchern oder im Internet zur normalen Arbeit eines Programmierers. Allerdings ist diese Suche in der Regel ein unstrukturierter und somit meist zeitaufwendiger Prozess. Entsprechend ist dieses Werk auch für mich eine Zusammenfassung wichtiger Lösungswege für die Grundprobleme des Programmierens.

Unabhängig davon, ob Sie das Buch als Studierender, Mitarbeiter, Gründer eines Start-ups oder als Unternehmer lesen, hoffe ich, dass Ihnen dieses Nachschlagewerk ein wertvoller Helfer für die ersten Anfänge sein wird. Ich gehe davon aus, dass jede Person

[1] Aus Gründen der besseren Lesbarkeit wird im Folgenden in der Regel die männliche Form verwendet. Allerdings betonen alle Autoren, dass gleichermaßen alle Geschlechterformen angesprochen sind.

die Grundlagen der Datenanalyse mithilfe moderner Programmiersprachen erlernen kann. Der Erfolg der Programmiersprache Python zeigt, dass die Programmierung kein aufwendiges Unterfangen sein muss. In den letzten Jahren ist Python in Forschung und Praxis zu einer der beliebtesten Programmiersprachen avanciert. Ein zentraler Vorteil von Python ist seine Verständlichkeit, besonderes für Anfänger. Hinzu kommt, dass Python über eine große Entwicklergemeinschaft verfügt. Dies führt zur Bereitstellung zahlreicher und kostenfreier Erweiterungen. Durch meine Erfahrungen als Wissenschaftler im Bereich der künstlichen Intelligenz, Universitätsdozent und Redner, weiß ich, dass viele Menschen eine gewisse Einstiegshürde mit Blick auf die Programmierung verspüren, die sich aber schnell nach den ersten Erfolgserlebnissen legt. Genau jenes Erfolgserlebnis kann zu einem regelrechten Motivator werden, um immer tiefer in die Programmierung einzusteigen.

Ausgehend von meinen Erfahrungen, lernt man am besten durch einfache Beispiele. Das Lösen solcher Aufgaben ist aus meiner Perspektive besonderes wichtig, um ein Grundverständnis zu entwickeln. Vor diesem Hintergrund ist eine zentrale Prämisse des Buchs, dass alle Lösungsansätze so einfach wie möglich gehalten sind, was Sie als Leser befähigen soll, Ansätze für die Lösung Ihrer eigenen Probleme zu entwickeln. Sie benötigen für dieses Buch keine tiefgreifenden Vorkenntnisse in den Bereichen Programmierung oder Datenanalyse. Sofern Sie einen Computer bedienen können und in der Lage sind, Software zu installieren, können Sie mit dem Lesen des Buchs beginnen.

Dieses Werk grenzt sich von anderen Büchern ab, da es ein Nachschlagewerk für Quellcodeansätze im täglichen Data Science sein soll. Aus meiner Erfahrung sind viele einführende Lehrbücher in Python als universelle Bücher konzipiert, sodass sie viele Teilbereiche abdecken und ausführliche Erklärungen bieten. Damit sind diese Lehrbücher als Lernwerke im Sinn eines ausführliches Studiums konzipiert, was dazu führt, dass der interessierte Leser einem solchen Lehrbuch ausreichend Zeit widmen muss. Im Gegensatz dazu soll mein Werk praktisch anwendbar sein und beim aktiven Quellcode schreiben helfen, sodass es einfach aus dem Regal genommen werden kann, um eine schnelle Hilfestellung bei aufkommenden Fragen zu liefern. Alternativ kann das Buch als Inspirationsgeber für die Umsetzung eigenständiger Projekte dienen. Mit diesem Buch möchte ich den Leser dazu befähigen, den Data-Science-Herausforderungen zu begegnen. Meine Co-Autoren und ich haben in diesem Werk möglichst auf Fachvokabular zugunsten von Alltagssprache verzichtet, um die Verständlichkeit zu erhöhen. Notwendige Fachbegriffe werden an geeigneten Stellen ausführlich erklärt. Durch das Stichwortverzeichnis lassen sich außerdem entsprechende Begrifflichkeiten schnell finden und das Lexikon bietet eine Übersicht mit Kurzerklärungen des Data-Science-Sprachgebrauchs.

An dieser Stelle möchte ich mich bei allen Personen bedanken, die sich unmittelbar oder indirekt an diesem Buchprojekt beteiligt haben. Mein Dank gilt besonders meinen Co-Autoren, die durch ihre unterschiedlichen Ideen und ihre Mitarbeit dieses Buch maßgeblich vorangebracht haben. Ich verzichte aus Platzgründen bewusst auf

die einzelne namentliche Nennung, möchte mich aber neben den Co-Autoren auch bei meinen Kollegen, Studenten und Familienmitgliedern bedanken.

Programmierung ist für mich vergleichbar mit dem Erstellen eines eigenen Kunstwerks. Jeder eigene Programmcode trägt die Handschrift des Schöpfers. Daher freue ich mich besonders darüber, dass der interessierte Leser in diesem Buch die Gelegenheit erhält, unterschiedliche Stile der Programmierung kennenzulernen. Daher wäre es schön, wenn Sie als Leser die unterschiedlichen Ideen und Quellcodes so betrachten, als würden Sie durch eine Galerie schlendern.

Sie finden alle Inhalte zu diesem Buch auf der Webseite des Buchs: https://github.com/BenjaminMAK/DataScienceBestPractices

Über Anregungen und Kritik zu diesem Buch, auch über Hinweise zu Entwicklungsmöglichkeiten oder Wünsche für zukünftige Auflagen, freue ich mich sehr. Gern können Sie mich über die Webseite zum Buch kontaktieren.

Nun wünsche ich allen Lesern einen interessanten Einstieg in die Programmierung und viel Freude beim Lesen und Ausprobieren.

Hochachtungsvoll im Namen aller Co-Autoren Ihr:
Benjamin M. Abdel-Karim

Herausgeber- und Autorenverzeichnis

Über den Herausgeber

Dr. Benjamin M. Abdel-Karim ist seit März 2018 als wissenschaftlicher Mitarbeiter an der Professur für Wirtschaftsinformatik und Informationsmanagement von Prof. Dr. Oliver Hinz tätig. Er hat Wirtschaftsinformatik an der Technischen Universität in Darmstadt (M.Sc.) studiert. Im Masterstudium lagen seine Schwerpunkte in den Bereichen Data Knowledge Engineering/Artificial Intelligence und Finanzierung. Die Masterarbeit befasst sich mit der Modellierung spezieller neuronaler Netze zur Abbildung komplexer Finanzmarktstrukturen zwecks der Prognose. Sein Bachelorstudium in Wirtschaftsinformatik (B.Sc.) absolvierte er an der Universität Bremen. Der Schwerpunkt seines Bachelorstudiums war Computational Finance. Praktische Erfahrungen konnte Benjamin M. Abdel-Karim unter anderem durch seine Bankausbildung und Mitarbeit in zahlreichen Banken, wie beispielsweise der Deutsche Asset & Wealth Management, sammeln. Zudem kommen zahlreiche Auftritte als Keynote Speaker im Bereich maschinelles Lernen und Data Science dazu.

Autorenverzeichnis

Dr. Kevin Bauer trat 2013 in das Doktorandenprogramm der GSEFM ein, nachdem er seinen Bachelor-Abschluss in Wirtschaftswissenschaften an der Goethe-Universität erworben hatte. Im Jahr 2017 schloss er seinen ersten Master-Abschluss in Law and Quantitative Economics ab. Im Jahr 2018 schloss er seine Promotion mit dem akademischen Titel Dr. rer. pol. mit einem summa cum laude ab. Seine Doktorarbeit Ön the Economic Significance of Social Groups: New Evidence on Self-Selection, Social Identity and Social Preferences konzentriert sich auf die empirische Analyse von Wirtschaftsbeziehungen.Während seiner Promotion begann er einen zweiten Masterstudiengang in Wirtschaftsinformatik an der Goethe-Universität mit den Schwerpunkten maschinelles Lernen und Künstliche Intelligenz. Nach seiner Promotion trat Kevin Bauer als KI-Spezialist dem TechQuartier bei, wo er weiterhin als externer Berater für KI-bezogene Themen tätig ist. Seit 2020 ist Kevin Bauer als Postdoc-Forscher am Leibniz SAFE in der Forschungsgruppe Digitalisierung in der Finanzindustrie (Leitung: Prof. Hinz) beschäftigt.

Daniel Franzmann erlangte seinen Bachelor in Wirtschaftswissenschaften mit dem Schwerpunkt in Finanzen und Accounting an der Goethe Universität Frankfurt. Nach einem Auslandssemester an der York University in Großbritannien und erster Berufserfahrung bei PwC und Senacor Technologies AG absolvierte er im Jahr 2016 seinen Master in Wirtschaftsinformatik mit dem Fokus auf Data Science an der Goethe Universität Frankfurt. Von 2016 bis 2020 war Daniel Franzmann am Lehrstuhl für Information Systems Engineering der Goethe Universität Frankfurt als Data Scientist angestellt und promovierte über das Thema Software Updates. Seit Oktober 2020 arbeitet Daniel Franzmann im Data Analytics Team der Deutschen Bank. Privat interessiert er sich fürs Laufen, Kochen und surft gerne auf r/dataisbeautiful.

Hendrik Jöntgen absolvierte seinen Bachelor und Master in Wirtschaftsinformatik an der Technischen Universität Darmstadt. Im Masterstudium lag sein Schwerpunkt in dem Bereich Informationsvisualisierung und durch langjährige Arbeitserfahrung als studentische Hilfskraft an verschiedenen Lehrstühlen konnte er bereits früh Erfahrungen mit dem Erheben und Auswerten von Social Media Daten sammeln. In seiner Masterarbeit untersuchte er Vertrauensbildung innerhalb der Sharing Economy und die Effekte von unterschiedlichen Profil-Charakteristiken auf die Wahrnehmung der Plattform-Nutzer. Seit März 2019 ist Hendrik Jöntgen als wissenschaftlicher Mitarbeiter am Lehrstuhl für Wirtschaftsinformatik und Informationsmanagement von Prof. Dr. Oliver Hinz tätig und erforscht Vertrauen und soziales Kapital auf Social Media Plattformen. Seine Freizeit verbringt er gerne mit Brettspielen, am Schlagzeug oder auf Waldwegen.

Katharina Keller studierte Wirtschaftsinformatik an der Goethe-Universität Frankfurt (M.Sc.). Ihr Bachelorstudium absolvierte sie in Wirtschaftswissenschaften mit Schwerpunkt Marketing/Management. Während ihres Bachelorstudiums verbrachte sie ein Auslandssemester an der Católica Lisbon School of Business and Economics. Im Juni 2017 schloss sie ihr Studium mit dem akademischen Grad Master of Science ab. Ihre Masterthesis schrieb sie zum Thema „Technology Choice Behavior in Post-Adoption IS Usage“. Seit Juli 2017 ist Katharina Keller als wissenschaftliche Mitarbeiterin an der Professur von Prof. Dr. Oliver Hinz tätig. Zuerst am Lehrstuhl für Wirtschaftsinformatik | Electronic Markets an der TU Darmstadt und seit September 2017 an der Professur für Wirtschaftsinformatik und Informationsmanagement an der Goethe-Universität. Sie ist Teil des Sonderforschungsbereichs 1053 MAKI (Multi-Mechanismen-Adaption für das künftige Internet), der sich mit Fragen zum Thema künftiges Internet und dessen Auswirkung auf das Kommunikationsverhalten befasst. Außerdem betrachtet der SFB 1053 die Mechanismen in Kommunikationssystemen sowie die daraus resultierenden

Anforderungen an die Infrastruktur. Im Speziellen ist Katharina Keller im Teilprojekt B3 beteiligt, das die Transitionen in Kommunikationssystemen aus ökonomischer Perspektive untersucht und dezentrale Analyse- und Planungslösungen für autonome Endgeräte erstellt. Ziel ist eine effiziente und realistische Anwendbarkeit proaktiver Transitionen für koexistente Multi-Mechanismen in drahtlosen Netzen software-definierter autonomer Knoten.

Marcel Zeuch studierte Mathematik an der Technischen Universität in Darmstadt (M.Sc.). Sein erstes Mastersemester absolvierte er dabei 2016 als Erasmus-Stipendiat an der Politecnico di Milano in Mailand, Italien. Seither sammelte er Erfahrungen in einer Managementberatung, einer Top Tier Investmentbank sowie seit seinem Abschluss im Jahr 2018 in einer datengetriebenen Boutiqueberatung im Capital Markets Umfeld. In seiner aktuellen Position im Bereich Corporates & Markets bei einer deutsch-französischen Investmentbank beschäftigt er sich mit Themen der internen Prozessautomatisierung und -weiterentwicklung. In seiner Freizeit verfasste er bereits mehrfach Artikel für einen deutschen Finanzblog und interessiert sich für Themen aus den Bereichen Capital Markets, Data Science und Entrepreneurship.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Listings

Teil I

Grundlagen der Programmierung

Einleitung

1

Benjamin M. Abdel-Karim

Dieses Kapitel gibt einen Überblick über den Inhalt, den Aufbau und die Zielsetzung des Buchs. Mit dem wachsenden Angebot an Datenquellen wächst das Interesse, diese entsprechend zu monetarisieren. Dies gilt nicht nur für digitale Geschäftsprozesse, sondern auch für die Forschung und Entwicklung. Damit sind das Verständnis und die Fähigkeit zur Datenanalyse gewinnbringende Vermögenswerte, um die gemeinschaftliche Wohlfahrt des Staats voranzubringen. Im ersten Teil des Kapitels wird der Gegenstand dieses Buchs beschrieben. Daran knüpft der Aufbau des Buchs an und es wird eine Erklärung geboten, weshalb sich Python besonders für Einsteiger eignet.

1.1 Gegenstandsbereich dieses Buchs

Die stets zunehmende Menge an Daten und die implizit entstehende Komplexität dieser Daten führen dazu, dass neue Geschäftsbereiche und Stellenbeschreibungen entstehen, sodass die Fähigkeit, diese Daten zu verarbeiten und geeignet zu analysieren, immer gefragter wird. Das Anwendungsfeld der Datenanalyse war bis vor wenigen Jahren nur im Bereich Forschung und Entwicklung zu finden. Allerdings entdecken zunehmend Startups, der Mittelstand und die großen Technologiekonzerne diesen Bereich für sich, um neue Geschäftsmodelle zu erschließen. Vor dem Hintergrund der Wettbewerbssicherung führt dieser Umstand dazu, dass Studierende, Berufsanfänger und Unternehmer sich frühzeitig mit den Grundlagen der Datenanalyse befassen sollten. Damit richtet sich dieses Buch an eben jene interessierten Leser, die sich mit den Grundlagen der Datenanalyse befassen

B. M. Abdel-Karim (✉)
Frankfurt am Main, Hessen, Deutschland
E-mail: BenjaminM.Abdel-Karim@gmx.de

B. M. Abdel-Karim (Hrsg.), *Data Science,*
https://doi.org/10.1007/978-3-658-33460-4_1

möchten, insbesondere an Leser, die sich frühzeitig den neuen Wettbewerbsanforderungen stellen wollen.

Dieses Werk ist nicht als reines Einführungswerk zu verstehen, sondern als Nachschlagewerk zum Lösen der klassischen Problemstellungen in der Datenanalyse, die in der Regel zu Beginn der Datenanalyse auftreten. Basierend auf den Erkenntnissen zahlreicher Vorlesungen, Übungen, Mentorings und Speaker Events fasst dieses Buch einige Quellcodeausschnitte in strukturierter Art und Weise zusammen. Das Buch beginnt mit den essenziellen Grundlagen der Programmierung in Python. Diese werden durch praktische Implementierungen von Anwendungsbeispielen vertieft.

1.2 Aufbau und Zielsetzung

Als Grundlagen- und Nachschlagewerk hat dieses Buch den primären Anspruch, die Umsetzung von wesentlichen Methoden im Umgang der Datenanalyse im Kontext der Programmiersprache Python darzulegen. Der erste Teil des Buchs vermittelt die essenziellen Grundkenntnisse, um die Data-Science-Projekte im zweiten Teil des Buchs nachvollziehen zu können. Die Besonderheit in diesem Buch sind eben jene Projekte (Case Studies). Jede Case Study behandelt eine Data-Science-Herausforderung aus der Perspektive einer meiner Co-Autoren, sodass Sie als Leser in diesem Buch die Gelegenheit erhalten werden, unterschiedliche Programmierstile und -strategien kennenzulernen. Daraus ergibt sich der folgende Aufbau dieses Buchs:

- Grundlagen
  - Python Installation und Einstieg
  - Grundlegende Datentypen
  - Datenstrukturen und Indexierung
  - Kontrollstrukturen
- Data-Science-Projekte
  - Data Science und Data-Science-Prozess
  - Überblick bedeutender Bibliotheken
  - Datenvorverarbeitung
  - Datenvisualisierungen
  - Entwicklung eigener Modelle

Die Zielsetzung des Buchs ist damit, die gesammelten Erkenntnisse und Erfahrungen zu archivieren und weiterzugeben, damit der Leser aus vorangegangenen Erfahrungen der Autoren und dem entstandenen Lernprozess profitieren kann. Im Allgemeinen richtet sich das Buch mit dieser Zielsetzung an jeden interessierten und motivierten Leser, sich die Grundlagen des Programmierens in Python und von Data Science anzueignen. Außerdem ist das Buch für Studierende im Bachelor- und Masterstudium konzipiert, die ihren

Wissenshorizont erweitern möchten und durch die erworbenen Fähigkeiten imstande sind, die klassischen Aufgaben des wissenschaftlichen Arbeitens im Rahmen der Datenanalyse zu bewerkstelligen. Darüber hinaus ist das Buch für Berufspraktiker geschrieben, die im Rahmen ihrer Berufstätigkeit eine flexible und moderne Programmiersprache nutzen wollen, um die anfallenden Daten zu analysieren. In vielen Firmen werden bisher einfache Tabellenverarbeitungsprogramme eingesetzt, die allerdings schnell an ihre Verarbeitungskapazitäten stoßen. Hierzu können die vorgestellten Anwendungsfälle einen Eindruck zu den möglichen Alternativen mithilfe der Programmiersprache Python aufzeigen. Zusätzlich richtet sich dieses Buch an junge Unternehmer, die den Wert ihrer Daten erkannt haben und nun auf der Suche nach einer Inspiration sind, diese Daten geeignet auszuwerten.

Demnach möchte dieses Buch einen Beitrag zur Wissensdiffusion moderner Technologien liefern, wie beispielsweise Hardware und Programmiersprachen, und durch anschauliche und praxisnahe Beispiele zum selbstständigen Ausprobieren anregen. Vor dem Hintergrund des exponenziellen Wissenszuwachses der menschlichen Gesellschaft und der damit verbunden Entwicklung einer hochspezialisierten Wissensgesellschaft (Stehr, 2006; Bittlingmayer & Bauer, 2006), ist die private Weiterbildung ein zentraler Bestandteil der fachlichen Weiterentwicklung.

1.3 Warum Python?

In der Welt der Programmierung existieren zahlreiche Programmiersprachen. Jede einzelne von ihnen besitzt Stärken und Schwächen. Für Einsteiger bietet sich besonders die Programmiersprache Python an. An dieser Stelle möchte ich kurz auf einige gute Gründe zur Nutzung von Python eingehen, ohne den Details aus den kommenden Kapiteln vorzugreifen:

- Python ist schnell zu erlernen, da diese Programmiersprache im Vergleich zu anderen Programmiersprachen auf wenige Syntaxsymbole setzt.
- Zudem werden Variablen dynamisch deklariert, sodass das Schreiben des Codes zu Beginn vereinfacht wird.
- Python wird im Bereich Forschung und Entwicklung sowie in vielen Praxiskontexten eingesetzt, sodass ihr eine hohe praktische Relevanz beizumessen ist.
- Python wird von einem großen Nutzerkreis durch umfassende Module (Bibliotheken) unterstützt, sodass viele Programmcodeteile schon verwendet werden können. Außerdem existieren zahlreiche Onlinekurse und Literaturbeiträge.
- Python lässt sich mit anderen Programmiersprachen wie C++ oder HTML verknüpfen, sodass komplexere Programme geschrieben werden können.

Die folgenden Kapitel gehen auf die Grundlagen der Programmiersprache ein.