

Dieses Buch richtet sich an Data Scientists, die mit den Programmiersprachen R und/oder *Python* vertraut sind und sich bereits früher (wenn auch nur punktuell oder zeitweise) mit Statistik beschäftigt haben. Zwei der Autoren entstammen der Welt der Statistik, ehe sie sich in den weiten Raum der Data Science begeben haben, und schätzen den Beitrag, den die Statistik zur Datenwissenschaft zu leisten vermag, sehr. Gleichzeitig sind wir uns der Grenzen des traditionellen Statistikkunterrichts durchaus bewusst: Statistik als Disziplin ist anderthalb Jahrhunderte alt, und die meisten Statistiklehrbücher und -kurse sind nicht gerade von Dynamik geprägt, sondern erinnern eher an die Trägheit eines Ozeanriesen. Alle Methoden in diesem Buch haben einen gewissen historischen oder methodologischen Bezug zur Disziplin der Statistik. Methoden, die sich hauptsächlich aus der Informatik entwickelt haben, wie z.B. neuronale Netze, werden nicht behandelt.

Diesem Buch liegen zwei Ziele zugrunde:

- Schlüsselbegriffe aus der Statistik, die für die Data Science relevant sind, in zugänglicher, übersichtlich gegliederter und leicht referenzierbarer Form darzulegen.
- Eine Erläuterung dazu zu geben, welche Konzepte aus datenwissenschaftlicher Sicht wichtig und nützlich sind, welche weniger wichtig sind und warum.

In diesem Buch verwendete Konventionen

Die folgenden typografischen Konventionen werden in diesem Buch verwendet:

Kursiv

Kennzeichnet neue Begriffe, URLs, E-Mail-Adressen, Dateinamen und Dateiendungen.

Konstante Zeichenbreite

Wird für Programmlistings und für Programmelemente in Textabschnitten wie Namen von Variablen und Funktionen, Datenbanken, Datentypen, Umgebungsvariablen, Anweisungen und Schlüsselwörter verwendet.

Konstante Zeichenbreite, **fett**

Kennzeichnet Befehle oder anderen Text, den der Nutzer wörtlich eingeben sollte.

Schlüsselbegriffe

Die Data Science baut auf mehreren Disziplinen auf, darunter Statistik, Informatik, Informationstechnologie und domänenspezifische Bereiche. Infolgedessen können mehrere unterschiedliche Begriffe verwendet werden, um auf ein bestimmtes Konzept zu verweisen. Schlüsselbegriffe und ihre Synonyme werden im gesamten Buch in einem Kasten wie diesem hervorgehoben.



Dieses Symbol steht für einen Tipp oder eine Empfehlung.



Dieses Symbol steht für einen allgemeinen Hinweis.



Dieses Symbol warnt oder mahnt zur Vorsicht.

Verwenden von Codebeispielen

Zu sämtlichen Beispielen zeigen wir in diesem Buch die entsprechenden Codebeispiele – zuerst immer in *R* und dann in *Python*. Um unnötige Wiederholungen zu vermeiden, zeigen wir im Allgemeinen nur Ausgaben und Diagramme, die durch den *R*-Code erzeugt wurden. Wir klammern auch den Code aus, der zum Laden der erforderlichen Pakete und Datensätze erforderlich ist. Den vollständigen Code sowie die Datensätze zum Herunterladen finden Sie unter <https://github.com/gedeck/practical-statistics-for-data-scientists>.

Dieses Buch dient dazu, Ihnen beim Erledigen Ihrer Arbeit zu helfen. Im Allgemeinen dürfen Sie die Codebeispiele aus diesem Buch in Ihren eigenen Programmen und der dazugehörigen Dokumentation verwenden. Sie müssen uns dazu nicht um Erlaubnis bitten, solange Sie nicht einen beträchtlichen Teil des Codes reproduzieren. Beispielsweise benötigen Sie keine Erlaubnis, um ein Programm zu schreiben, in dem mehrere Codefragmente aus diesem Buch vorkommen. Wollen Sie dagegen eine CD-ROM mit Beispielen aus Büchern von O'Reilly verkaufen oder verteilen,

brauchen Sie eine Erlaubnis. Eine Frage zu beantworten, indem Sie aus diesem Buch zitieren und ein Codebeispiel wiedergeben, benötigt keine Erlaubnis. Eine beträchtliche Menge Beispielcode aus diesem Buch in die Dokumentation Ihres Produkts aufzunehmen, bedarf hingegen unserer ausdrücklichen Zustimmung.

Wir freuen uns über Zitate, verlangen diese aber nicht. Ein Zitat enthält Titel, Autor, Verlag und ISBN, zum Beispiel: »*Praktische Statistik für Data Scientists* von Peter Bruce, Andrew Bruce und Peter Gedeck (O'Reilly). Copyright 2020 Peter Bruce, Andrew Bruce und Peter Gedeck, ISBN 978-3-96009-153-0.«

Wenn Sie glauben, dass Ihre Verwendung von Codebeispielen über die übliche Nutzung hinausgeht oder außerhalb der oben vorgestellten Nutzungsbedingungen liegt, kontaktieren Sie uns bitte unter kontakt@oreilly.de.

Danksagungen

Die Autoren danken den zahlreichen Menschen, die dazu beigetragen haben, dieses Buch Wirklichkeit werden zu lassen.

Gerhard Pilcher, CEO des Data-Mining-Unternehmens Elder Research, sah frühe Entwürfe des Buchs und half uns mit detaillierten und hilfreichen Korrekturen sowie Kommentaren. Ebenso gaben Anya McGuirk und Wei Xiao, Statistiker bei SAS, und Jay Hilfiger, ebenfalls Autor von O'Reilly, hilfreiches Feedback zu den ersten Entwürfen des Buchs. Toshiaki Kurokawa, der die erste Auflage ins Japanische übersetzte, leistete dabei umfassende Überarbeitungs- und Korrekturarbeit. Aaron Schumacher und Walter Paczkowski haben die zweite Auflage des Buchs gründlich überarbeitet und zahlreiche hilfreiche und wertvolle Anregungen gegeben, für die wir sehr dankbar sind. Es versteht sich von selbst, dass alle noch verbleibenden Fehler allein auf uns zurückzuführen sind.

Bei O'Reilly begleitete uns Shannon Cutt mit guter Laune und der richtigen Portion Nachdruck durch den Publikationsprozess, während Kristen Brown unser Buch reibungslos durch den Produktionsprozess geführt hat. Rachel Monaghan und Eliahu Sussman korrigierten und verbesserten unser Buch mit Sorgfalt und Geduld, während Ellen Troutman-Zaig den Index erarbeitete. Nicole Tache übernahm das Lektorat der zweiten Auflage und hat den Prozess effektiv geleitet sowie viele gute redaktionelle Vorschläge gemacht, um die Lesbarkeit des Buchs für ein breites Publikum zu verbessern. Wir danken auch Marie Beaugureau, die unser Projekt bei O'Reilly initiiert hat, sowie Ben Bengfort, Autor von O'Reilly und Ausbilder bei Statistics.com, der uns O'Reilly vorgestellt hat.

Wir und dieses Buch haben auch von den vielen Gesprächen profitiert, die Peter im Laufe der Jahre mit Galit Shmueli, Mitautorin bei anderen Buchprojekten, geführt hat.

Schließlich möchten wir besonders Elizabeth Bruce und Deborah Donnell danken, deren Geduld und Unterstützung dieses Vorhaben möglich gemacht haben.