Kapitel 1
Über dieses Buch

In diesem Einleitungskapitel stelle ich Ihnen die Ziele und die Struktur des Buches vor. Ich sage Ihnen auch, was Sie hier lernen und was Sie hier nicht lernen werden.

1.1 Für wen ist dieses Buch? Für Sie?

Dieses Buch richtet sich an all die, die sich bisher davor gedrückt haben, statistische Analysen zu realisieren, und an die, die ein bisschen Angst haben. Es richtet sich an die, denen in der Schule gesagt wurde, dass sie kein Mathe könnten (von Lehrern, die es nicht unbedingt gut erklärt haben). Es ist auch für die, die denken, sie sind schon zu alt, um noch etwas wie programmieren zu lernen. Mit diesem Buch gelingt es jedem. Es ist an alle gerichtet, die einfach nur mal schauen wollen, wie statistische Analysen funktionieren und warum sie überhaupt gemacht werden. So wie ich vor einigen Jahren.

Es ist aber auch für die Journalisten, die sicherer schreiben wollen, und für alle Geisteswissenschaftler, die ihre Aussagen auf noch belastbarere Füße stellen möchten. Es ist auch für die, die Grundkenntnisse in R für Bewerbungen benötigen oder eigene Analysen zu Hause erstellen wollen. Auch die, die verstehen möchten, wie 80 % aller *Graphen* und *Diagramme*, die sie in den Print- und Onlinemedien sehen, erstellt werden, sollten dieses Buch lesen. Und schließlich ist es auch für die, die einfach nur erst mal schauen wollen, wie eine Datenanalyse aussehen kann, bevor sie sich der Statistik und R eventuell weiter annähern.

1.2 Was sind die Ziele, was können Sie hier lernen?

Ziel dieses Buches ist, Ihnen eine erste Idee von statistischen Analysen und deren Nutzen zu vermitteln – nur so theoretisch wie nötig, aber so praktisch wie möglich. Das heißt, dass Sie nach dem Lesen dieses Buches fähig sein sollen, bestimmte erste grundlegende statistische Analysen selbst ausführen zu können. Diese Analysen sollen sich in diesem Buch deshalb ebenfalls auf tägliche, praktische Anwendungsmöglichkeiten beschränken. Dazu wähle ich die drei häufigsten in öffentlichen Medien verwendeten sta-

tistischen Darstellungen. Ich möchte, dass Sie mit relativ wenig Lernaufwand die gängigsten dieser Statistiken darstellen und die Ergebnisse solcher Darstellungen analysieren und interpretieren können. Ich möchte, dass Sie mit wenigen Codezeilen *Balkendiagramme* (*Barplots*) wie in Abbildung 1.1 und übliche *Streu- und Liniendiagramme* (*Scatterplots* und *Linecharts*) mit *Trendbestimmung* (*Trend*) wie z. B. die in Abbildung 1.2 und Abbildung 1.3 erstellen können.

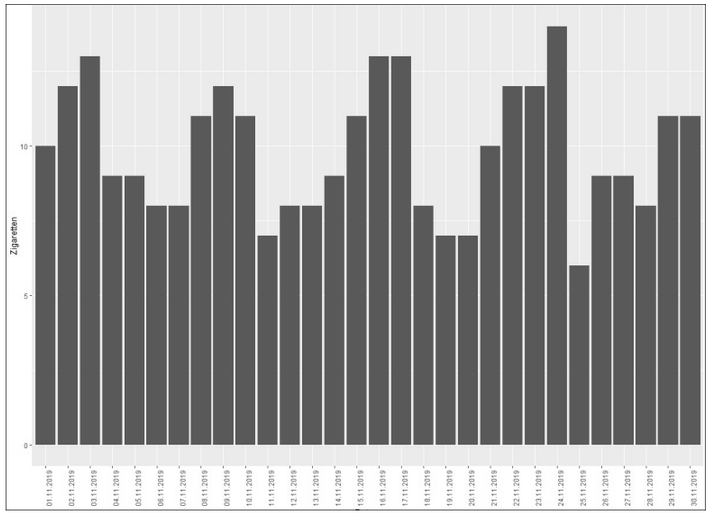

Abbildung 1.1 Privater monatlicher Zigarettenverbrauch 1

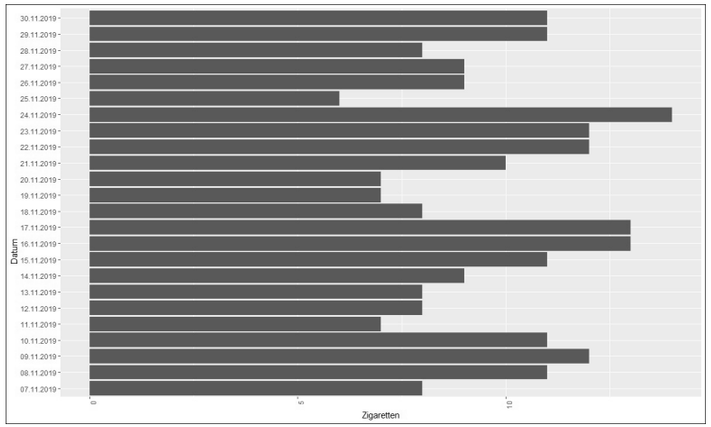

Abbildung 1.2 Privater monatlicher Zigarettenverbrauch 2

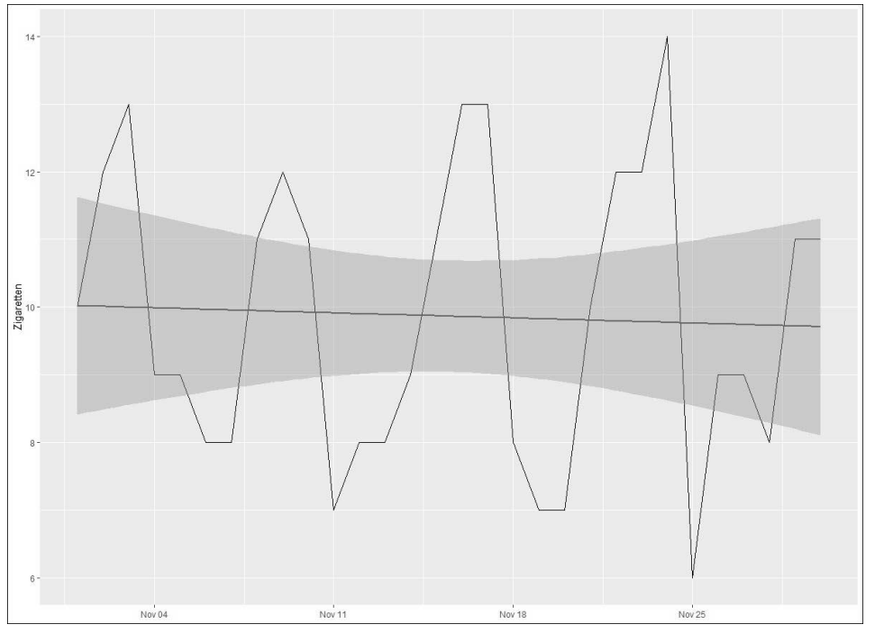

Abbildung 1.3 Liniendiagramm mit Trendbestimmung des privaten Zigarettenkonsums

1.3 Was Sie nicht lernen werden

In diesem Einstieg ist selbstverständlich nicht geplant, alle Themen der Data Science zu behandeln. Worum es hier insbesondere nicht geht, sind verwandte Programmiersprachen, Big Data, externe Datenbankabfragen nach Kriterien, andere Datenformen sowie kompliziertere Rechenverfahren.

1.3.1 Verwandte Programmiersprachen

Verwandte Programmiersprachen wie Python, Ruby, Stata oder MATLAB kommen in diesem Buch nicht vor.

1.3.2 Big Data

In diesem Buch ist ebenfalls nicht geplant, auf sehr große Datenmengen einzugehen (Big Data). Die Bearbeitung von Big Data hat noch einmal weitere Besonderheiten, die fünf Vs: Das sind ein großes Volumen (Volume), eine hohe Verschiedenartigkeit (Va-

riety), eine gewisse Schnelligkeit in der Verarbeitung (Velocity), die Validität der Daten (Validity) und der Wert, den ihre Analyse z. B. einem Unternehmen bietet (Value). Es empfiehlt sich, damit nach dem Einstieg in die Data Science zu beginnen. Nach dem Durcharbeiten dieses Buches wird es Ihnen möglich sein, kleine bis mittelgroße Datensätze zu verstehen und zu bearbeiten.

Das könnte auch nur ein erster Schritt sein auf dem Weg zur sicheren Analyse von größeren Datensätzen und Big Data, die im Alltag und Berufsleben immer gefragter ist. Ob in der Medizin, im Marketing, in der Forschung, im Versicherungswesen, in den Medien, der automatisierten Produktion, der Bildung, im Bankwesen und vielem anderen mehr, überall sind heute Spezialisten für die Analyse großer Datenmengen verantwortlich. Der Übergang zur Analyse von immer größeren Datensätzen gelingt Ihnen mit diesem Buch dann in einem zweiten Schritt sicher einfacher.

1.3.3 Datenbankabfragen

Datenbankabfragen, die häufig z. B. mit SQL ausgeführt werden, sollen hier ebenfalls nicht behandelt werden. Für die kleineren bis mittelgroßen Datensätze genügen eine gute Zusammenstellung der Daten und wenige Funktionen in R.

1.3.4 Andere Datenformen

Andere Datenformen haben andere Besonderheiten in der Erstellung und Verarbeitung. Geodaten z. B. erfordern eine gewisse Expertise in den Themenfeldern Geografie und Geometrie. Weit verbreitete Tools zur Analyse solcher Daten sind moderne Geographische Informationssysteme (GIS) wie z. B. QGIS oder MAP3D. Die Analyse akustischer Daten erfordert komplexes Wissen z. B. in den Bereichen Physik, Psychologie, Nachrichtentechnik und der Materialwissenschaft. Auch Erfahrungen in der mathematischen Modellierung helfen weiter. Das alles soll hier aber keine Voraussetzung sein. Es bleibt in diesem Buch bei, sagen wir fürs Erste, einfachen, »klassischen Daten«, die aus Zahlen und/oder Buchstaben bzw. Wörtern bestehen.

1.3.5 Kompliziertere Rechenverfahren

Kompliziertere Rechenverfahren wie multiple Regressionen oder Ereigniszeitanalysen werden in diesem Buch nicht besprochen. Das sind zwar auch wichtige und nützliche Verfahren, aber wie Sie feststellen werden, brauchen Sie das für einfache Datenanalysen alles nicht. Um kleinere bis mittelgroße Datensätze zu verstehen und Erkenntnisse aus ihnen zu ziehen, genügen häufig die Methoden der beschreibenden Statistik, die wir

hier um die Methode der einfachen linearen Regression erweitern. Behalten Sie sich nach dem Eindruck, den Sie durch dieses Buch hier gewinnen, komplexere Methoden für die Zukunft vor. Multiple lineare Regressionen und Ereigniszeitanalysen benötigen bspw. Analysten in der Privatwirtschaft oder der Versicherungsbranche, wenn sie etwa bestmögliche Preise oder die Höhe von Versicherungsbeiträgen bestimmen sollen.

1.4 Wie Sie mit diesem Buch arbeiten

Thematisch führe ich Sie zunächst in die wichtigsten Konzepte der Statistik und Daten ein. Es folgen Grundregeln und die wichtigsten Befehle in der R-Sprache R. Dann zeige ich Ihnen, wie ich mit nur wenigen Codezeilen ca. 80 % meiner Daten analysiere, sei es in Geschichte, Wirtschaft, Demografie, den Sozial- und Geisteswissenschaften allgemein oder ganz einfach im Alltag. Ich stelle hierbei zwischendurch immer wieder Fragen, die ich mir selbst auf meinem Weg gestellt habe. Ob Sie dabei das Buch nur lesen oder das Skript mit mir zusammen erstellen, bleibt Ihnen überlassen. Der Lernerfolg ist aber beim Selber-(Fehler-)machen noch immer der größte.

Das Buch ist hierbei rein praktisch organisiert und anwenderorientiert gestaltet. Es kommt ohne unnötigen Schnickschnack aus. Die nötigen theoretischen Grundlagen werden erklärt, aber immer nur in einem absolut nötigen Rahmen. Die Fachbegriffe neben meinen Erklärungen liefere ich Ihnen immer mit, sodass Sie auch zukünftig in anderen Büchern oder (Online-) Kursen verstehen, wovon die Rede ist. Ich zeige Ihnen in diesem Buch einen möglichen Weg der Datenanalyse – meinen Weg. Wenn Sie im Laufe dieses Buches auf Probleme stoßen, wie das beim Programmieren grundsätzlich zu erwarten ist, können Sie zusätzlich das Internet befragen (wie das treffsicher gelingen kann, zeige ich Ihnen auch). Arbeiten auch Sie sich Problem für Problem vor, knacken Sie jedes Rätsel einzeln. So kommen Sie weiter, Schritt für Schritt, jedes Mal ein Stückchen weiter als das Mal davor.