Uwe H. Kaufmann ~ Amy B. C. Tan

Data Science

für Einsteiger

Daten analysieren, interpretieren und richtige Entscheidungen treffen

HANSER

Kaufmann

Data Science für Einsteiger

Ihr Plus – digitale Zusatzinhalte!

Auf unserem Download-Portal finden Sie zu diesem Titel kostenloses Zusatzmaterial. Geben Sie dazu einfach diesen Code ein:

plus-k8jd0-d34sf

PLUS.HANSER-FACHBUCH.DE

Uwe H Kaufmann
Amy BC Tan

Data Science für Einsteiger

Daten analysieren, interpretieren und
richtige Entscheidungen treffen

HANSER

MIX
Papier aus verantwortungsvollen Quellen
FSC® C083411

Bibliografische Information der Deutschen Nationalbibliothek:

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

Print-ISBN 978-3-446-46348-6
E-Book-ISBN 978-3-446-46677-7

www.hanser-fachbuch.de
Lektorat: Lisa Hoffmann-Bäuml
Herstellung: Carolin Benedix
Satz: Eberl & Kœsel Studio GmbH, Krugzell
Coverrealisation: Max Kostopoulos
Titelmotiv: © gettyimages.de/sesame
Druck und Bindung: CPI books GmbH, Leck
Printed in Germany

Für Nicole und Christopher, Priscilla und Pearl

Vorwort

Ein neues Buch zur Nutzung von Daten – ist das eher etwas für Manager oder für Spezialisten in der Datenverarbeitung? Oder ist heute nicht fast jeder damit konfrontiert, sich mit einer größeren Datenfülle auseinanderzusetzen? In dem vorliegenden Werk von Amy BC Tan und Dr. Uwe Kaufmann ist der Anwendungsbereich daher deutlich weiter gespannt, als wir es in der bisherigen Literatur zum Thema gewohnt sind. Das liegt u. a. an der beruflichen Entwicklung der beiden Autoren. Amy BC Tan hat ihren Erfahrungshintergrund in der Organisationsentwicklung, insbesondere im Studium des Verhaltens bei Veränderungsdruck, und Uwe Kaufmann als messtechnisch orientierter Ingenieur und Unternehmensberater fokussiert auf datengestützte Prozessverbesserungen.

Anfang 1991 lernte ich Uwe Kaufmann als frisch Promovierten der Uni Jena kennen und konnte ihn für die neu gegründete Consulting Tochter des TÜV Rheinland, Köln, und der EBASCO, USA, gewinnen. Durch seine empathische Art und seine analytischen Fähigkeiten konnte er schnell im Team überzeugen und wurde deren Leiter. Sein nächster Karriereschritt in der Inhouse-Beratung des General-Electric-Konzerns war geprägt von der Weiterentwicklung und Nutzung des gesamten Lean-Six-Sigma-Werkzeugkastens. Auch nach der Verlegung seiner Aktivitäten nach Singapur haben wir u. a. durch Buchprojekte im TÜV Verlag immer Kontakt gehalten. Es ist eine Freude, mit ihm zusammenzuarbeiten.

Was mir an diesem Buch besonders gefällt, ist die klare Strukturierung mit datenbezogenen Fallbeispielen aus unterschiedlichsten Bereichen, wobei immer der Nutzen für die Organisation im Vordergrund steht. Woran erkennt man das? Bevor Daten erfasst, vorbereitet und analysiert werden, steht die Frage der Geschäftsrelevanz. Profitieren werden alle Leser, die Ursache-Wirkungs-Zusammenhänge bei kleinen und großen Datenmengen analytisch erfassen wollen. Dies hilft dem Management, richtungsweisende Entscheidungen für die gesamte Organisation zu treffen; und es dient auch der Fachebene bei Organisationsverbesserungen in allen Bereichen.

Prof. Dr. Hermann J. Thomann

Ehemals Senior Vice President Global Consulting TÜV Rheinland

Entfesseln Sie das Potenzial Ihrer Daten!

Aus Daten aussagekräftige Informationen zu gewinnen wird immer mehr zum entscheidenden Wettbewerbsfaktor. Dabei reicht es nicht aus, einfach nur Zahlen ohne wissenschaftliche Grundlagen zu verwenden, sondern es werden Informationen benötigt, die auf Daten basieren und statistisch fundiert sind! Dazu gehört, so nah wie möglich am Kunden zu sein. Doch auch jegliche Organisationsentwicklungsmaßnahme oder Prozessverbesserung muss auf Daten aufbauen.

Es stehen uns eine Unmenge Daten zur Verfügung, mehr als jemals zuvor. Und wir haben leistungsfähige Soft- und Hardware zur Hand, die jeder – auch ohne Informatikstudium – anwenden kann.

Statistik muss nicht abschrecken und das Gewinnen von aussagekräftigen Informationen muss nicht mit einer großen Investition verbunden sein. Dieses Werk zeigt, dass auch mit geringen Mitteln Großes erreicht werden kann. Sie erhalten eine Anleitung, wie Sie mit leicht zugänglichen Programmen (MS Excel, MS Power BI oder R) die für Ihren Geschäftserfolg zentralen Daten erfassen, umwandeln und daraus aussagekräftige Informationen ableiten. Mit den dargestellten Beispielen können typische Datenanalysefälle Schritt für Schritt nachvollzogen werden. Diese Schritte können Sie einfach für Ihre eigene Datenanalyse anpassen und für Ihre individuellen Belange gewinnbringend einsetzen. Erleichtert wird Ihnen die Umsetzung durch umfassende Arbeitshilfen, die kostenlos zum Download zur Verfügung stehen (siehe vorne eingedruckten Code oder entsprechenden Downloadhinweis am Ende des Werks).

Eine gewisse Affinität zur Statistik und zum Einsatz entsprechender Software erleichtert die Lektüre. Aber das Werk ist auch ohne Vorkenntnisse verständlich.

Dieses Buch ist für alle, die das Potenzial ihrer Daten freisetzen und der Konkurrenz einen Schritt voraus sein wollen.

Wir wünschen Ihnen bei der Umsetzung viel Erfolg!

Singapur, Januar 2021

Uwe H Kaufmann

Amy BC Tan

Inhalt

1 Einführung

1.1 Warum Datenanalytik wichtig ist

„Auf Gott vertrauen wir, alle anderen bringen Daten."

W Edwards Deming

Jeder erinnert sich an den mühsamen Prozess, bei der IT-Abteilung oder einem IT-Unternehmen eine Anfrage für eine Datenanalyseaufgabe zu stellen und Tage oder sogar Wochen zu warten, bis das Ergebnis vorliegt. In den meisten Fällen wurde das Ergebnis nicht auf die nützlichste Weise präsentiert oder hat eine Folgefrage aufgeworfen, für deren Beantwortung neue Daten, neue Anfragen an die IT-Abteilung oder den IT-Consultant erforderlich waren. Jahrzehntelang haben sich Manager auf diese Art und Weise der Datenanalyse verlassen, weil sie keine Wahl hatten. Dieses Verfahren hat einen fundamentalen Fehler: Wenn man zeitnahe Entscheidungen treffen will, können verzögerte und veraltete Informationen nicht verwendet werden. Daher mussten kurzfristige Entscheidungen ohne die Grundlage von Echtzeitdaten und manchmal auf der Basis des Bauchgefühls getroffen werden.

Die Zeit hat sich geändert. Die Menge der verfügbaren Daten in allen Funktionen jeder Organisation wächst täglich. Und der Zugang zu diesen Daten wird immer einfacher. Nahezu jeder kann sich die Daten beschaffen, die er für seine eigenen Analysen benötigt. Und fast jeder hat einen modernen Computer mit leistungsstarken Analysewerkzeugen direkt auf seinem Schreibtisch. Die Frage ist nun, wie die Daten in geschäftsrelevante Informationen umgewandelt werden können, um im Bedarfsfall die richtigen Schlussfolgerungen zu ziehen.

Datenanalytik ist der Prozess des Sammelns, Verarbeitens und Analysierens von Daten mit dem Ziel, nützliche Informationen aufzuspüren, Schlüsse anzubieten und die Problemlösung sowie die Entscheidungsfindung zu unterstützen.

Datenanalytik ist eine Geschäftspraxis, mit der jeder Manager vertraut sein sollte.

Die Datenanalytik umfasst die Hauptkomponenten Deskriptive Analytik (Post-Mortem-Analyse), Prädiktive Analytik und Präskriptive Analytik.

Big Data beschreibt Datensätze, die so umfangreich und komplex sind, dass herkömmliche Datenverarbeitungswerkzeuge sie nicht bearbeiten können. Big Data wird definiert durch seine drei Vs, Volumen, Geschwindigkeit (velocity), Vielfalt (Russom, 2011). Zu Beginn der 2000er-Jahre stellten große Datenmengen für viele Organisationen ein ernstes Problem dar. Einerseits nahm die Menge der verfügbaren Daten exponentiell zu. Auf der anderen Seite konnten CPU-Geschwindigkeit und Speicherkapazität nicht mit der vorhandenen Datenmenge Schritt halten. Zu dieser Zeit war der Umgang mit großen Daten einigen wenigen Unternehmen und Organisationen vorbehalten, die auf die Analyse von Daten angewiesen waren, um im Geschäft zu bleiben.

Heutzutage stehen jedoch Computer mit riesigen Datenspeicher- und Verarbeitungskapazitäten nahezu jeder Organisation zur Verfügung, sei es durch die Installation von Hard- und Software im eigenen Haus oder durch die Anmietung externer Kapazitäten. Zwei Trends scheinen das Ergebnis dieses Wandels in der IT-Umgebung zu sein. Erstens haben immer mehr Organisationen die Mittel und sehen die Notwendigkeit, Daten über ihre Betriebsumgebung zu sammeln. Zweitens erweitern diese Organisationen daher den Umfang ihrer Datenanalysetätigkeiten mit steigender Geschwindigkeit.

Während einige Forscher früher die Auffassung vertraten, dass Datenanalyse hauptsächlich den Umgang mit Benutzerdaten beschreibt, die von CRM- und ähnlichen Systemen erzeugt und in Kundenintelligenz umgesetzt werden, öffnet sich der Anwendungsbereich der Datenanalyse heute auf alle Funktionen einer Organisation.

Es gibt nicht nur eine Bewegung von der sogenannten „Big Data“-Analyse hin zur Analyse jeglicher Art von Daten, sondern es gibt auch einen gesunden Trend zur Einbeziehung aller Managementebenen und sogar der Mitarbeiter in dieses nicht so neue Gebiet des Informationsmanagements. Fortschrittliche Manager sind mit den verfügbaren Daten und mit Trends, Verschiebungen oder anderen Mustern in ihren Daten vertraut und nutzen sie für die Entscheidungsfindung.

Das frühere Spezialgebiet der Datenanalyse gewinnt unter allen Managern einer Organisation an Popularität. Es ist daher an der Zeit, dafür zu sorgen, dass die richtigen Daten in geeigneter Weise erhoben, gesichtet, transformiert und mit gültigen Methoden so analysiert werden, dass sie geschäftsrelevante Informationen liefern, die in Erkenntnisse umgewandelt werden und geeignete Entscheidungen für den Geschäftserfolg vorbereiten.

„Die Fähigkeit, Daten aufzunehmen – sie zu verstehen, zu verarbeiten, aus ihnen Wert zu schöpfen, sie zu visualisieren und zu kommunizieren – das wird in den nächsten Jahrzehnten eine enorm wichtige Fähigkeit sein."

Hal R Varian (2009)

1.2 Warum dieses Buch geschrieben wurde

Auf falsche Daten zu vertrauen ist schlimmer als gar keine Daten zu haben.

Es ist gut und wichtig, Zahlen zu haben, aber das ist nicht hinreichend. Darüber hinaus müssen wir sicherstellen, dass die Daten ordnungsgemäß gesammelt, bereinigt und analysiert werden, bevor wir eine Entscheidung treffen.

Blutbank mit schlechten Leistungsindikatoren

Bei einem internationalen Vergleich von Leistungsindikatoren bei Blutbanken kam heraus, dass eine Blutbank deutlich mehr Blutprodukte verschwendet als die anderen. Es war die Rede von Beuteln mit Blutplättchen, die von Blutspendern entnommen, getestet und dann entsorgt wurden, weil sie nicht den Qualitätsstandards entsprachen. Eine solche Situation war für die Leiterin der Blutbank nicht zu akzeptieren.

Ein Team wurde eingesetzt, um die Ursachen für die Verschwendung der wertvollen Blutprodukte zu untersuchen. Nach der Datenerhebung und einigen grundlegenden Analysen wurde klar, dass die Blutprodukte nicht von geringerer Qualität waren als in anderen Ländern. Die Grundursache lag in der Bewertung der Qualität der Blutbeutel – der Datenerfassung.

(Auf dieses Beispiel werden wir im Buch später zurückkommen.)

Nachfolgend einige zentrale Empfehlungen:

- Erstens: Vertrauen Sie Zahlen nicht blind. Selbst Zahlen, die von einem Computer ausgespuckt werden, können falsch, verzerrt oder anderweitig unbrauchbar gemacht worden sein. Prüfen Sie, wie diese Zahlen überhaupt erst in den Computer gelangt sind.
- Zweitens: Bevor Sie Datenanalysen durchführen, stellen Sie sicher, dass die Daten nach dem richtigen Verfahren erfasst wurden. Daher beginnen wir dieses Buch nicht mit der Datenanalyse, sondern dort, wo die Erhebung der Daten konzipiert wird.

- Drittens: So wie die Leiterin der Blutbank ihren sehr aussagekräftigen Business Case hatte, sichern Sie, dass Ihre Datenanalyse einem Zweck dient, einem Bedürfnis, das die Menschen, für die Sie mit Ihrer Datenanalyse arbeiten, kennen, verstehen und teilen. Nur mit diesem Zweck, diesem Business Case, ist Ihr Datenanalyse-Fall mehr als ein Spiel mit Zahlen.

In den folgenden Kapiteln werden wir den Einsatz der Datenanalyse zur Lösung von Geschäftsproblemen, zum Treffen kritischer Entscheidungen und zur Steuerung der Unternehmensstrategie erläutern. Und wir werden einige typische Fallstricke und Abhilfemaßnahmen auf dem Weg zur Datenanalyse aufzeigen.

Gegenwärtig sind auf dem Markt zahlreiche Kurse zur Datenanalyse verfügbar. Interessanterweise sind viele von ihnen mit Datenanalyse für HR-Fachleute oder für das Kundenbeziehungsmanagement (CRM) betitelt. Dieses Buch spannt einen breiteren Rahmen und zeigt den Gebrauch der Datenanalyse in vielen organisatorischen Situationen, in denen die richtige Verwendung von Daten entscheidend ist. Daher nennen wir es „Das Potenzial von Daten freisetzen – Nutzung von Data Science für die Organisationsentwicklung“.

Jeder Manager sollte vier leistungsfähige Analysekonzepte kennen, um über seine Organisation informiert zu sein und datengestützte Entscheidungen treffen zu können (Gallo, 2018). Diese Konzepte sind keineswegs neu. Sie gewinnen jedoch mit der zunehmenden Menge an verfügbaren Daten und dem offensichtlichen Bedarf – und der Chance –, diese Daten in geschäftsrelevante Informationen umzuwandeln, an Bedeutung. Unterstützt wird dies durch die Verfügbarkeit einer Vielzahl von einfach zu handhabenden Werkzeugen zur Datenanalyse und Datenvisualisierung.

Diese Werkzeuge können von Managern nur dann genutzt werden, wenn diese Manager die Grundlagen der Datenanalyse von der Datenerfassung bis zur Entscheidung verstehen. Daher müssen Manager die grundlegendsten Konzepte kennen (Gallo, 2018). Bei diesen Konzepten handelt es sich um randomisierte kontrollierte Experimente, Hypothesentests, Regressionsanalysen und statistische Signifikanz.

Zu den **randomisierten kontrollierten Experimenten** gehören Datenerfassungstechniken wie Umfragen und Erhebungen aller Art, Pilotstudien, Feldexperimente und Laborforschung. Anstatt solche Dienstleistungen an Spezialisten auszulagern und sich darauf zu verlassen, dass diese das Ergebnis analysieren und Empfehlungen entwickeln, ist es oftmals von Vorteil, die Daten und den Analyseprozess zu verstehen. Dieses Wissen würde helfen, maßgeschneiderte Schlussfolgerungen für die Organisation zu ziehen; Schlussfolgerungen, die ein Außenstehender nicht ohne Weiteres ziehen kann. Experimente umfassen auch das Testen neuer Routinen oder Produkte auf ihre Leistungsfähigkeit. Das Experimentieren mit Prozessen ist eine wirkungsvolle Möglichkeit, die Ausbeute zu verbessern und gleichzeitig andere wichtige Indikatoren kontrolliert zu verändern.

Die Gruppe der **Hypothesentests** enthält statistische Instrumente, die geschichtete geschäftsrelevante Daten vergleichen und die Frage nach dem „Besseren“ beantworten, einschließlich der Berechnung des inhärenten Risikos, dass diese Entscheidung falsch sein könnte. Hypothesentests finden ihre Anwendung in allen Einheiten jeder Organisation. Bei der Analyse von Umfrageergebnissen werden Hypothesentests zur Beantwortung von Fragen wie „Gibt es einen Unterschied zwischen dem letztjährigen und dem diesjährigen Rating?“ oder „Hat Abteilung A besser als Abteilung B abgeschnitten?“ eingesetzt. Das Ergebnis eines Hypothesentests kann viel mehr sein als nur ein „Ja“ oder „Nein“ zu solchen Fragen. Hypothesentests zeigen immer ein Risiko, das mit dem Treffen einer Entscheidung einhergeht; ein Risiko, eine falsche Schlussfolgerung zu ziehen. Viele Hypothesentests geben sogar einen Hinweis darauf, was der minimale Unterschied oder die minimal erreichbare Verbesserung ist, was zu viel besseren Entscheidungen über die Auswirkungen einer Änderung oder Verbesserung führt. „Was ist die minimale Verbesserung, wenn wir unsere Lieferungen von Lieferant B im Vergleich zu Lieferant A kaufen?“ kann mit Hypothesentests beantwortet werden.

Die Gruppe der **Regressionsanalysen** umfasst statistische Werkzeuge, die für ähnliche Aufgaben wie Hypothesentests verwendet werden. Während Hypothesentests in der Regel Fragen über die Beziehung zwischen zwei Variablen beantworten, können Regressionsmodelle eine große Anzahl von Variablen gleichzeitig umfassen. Damit kann die Interaktion zwischen mehreren Treibern (unabhängige Variablen) für dasselbe Ergebnis (abhängige Variable) analysiert werden, was bei Hypothesentests schwieriger ist. Regressionsmodelle helfen daher, komplexe Zusammenhänge zwischen vielen Variablen gleichzeitig zu erklären. Darüber hinaus werden diese Werkzeuge häufig in der prädiktiven Statistik eingesetzt, d. h. um vorhandene Daten für die Vorhersage des Verhaltens von Kunden, Maschinen, Organisationseinheiten und sogar Arbeitskräften zu nutzen.

Den genannten Methoden liegt ein wichtiges Konzept zugrunde: die **statistische Signifikanz**. Dieses oft missverstandene Konzept ist das Rückgrat aller Statistiken, das Rückgrat aller Datenanalysen. Die statistische Signifikanz informiert über das Risiko, das man eingehen muss, wenn man eine geschäftliche Entscheidung auf der Grundlage der Datenanalyse trifft.

In der Statistik gibt es „nie“ und „immer“ nicht. „0 % Wahrscheinlichkeit“ und „100 % Wahrscheinlichkeit“ sind in der Regel nicht das Ergebnis von randomisierten, kontrollierten Experimenten, Hypothesentests oder Regressionen. Höchstwahrscheinlich liegt das Ergebnis einer Analyse irgendwo dazwischen. Dann obliegt es dem Manager, eine kluge, sachkundige und datenbasierte Schlussfolgerung zu ziehen. Das Verständnis des Konzepts der Signifikanz ist der Schlüssel auf dem Weg zu einer Qualitätsentscheidung.

1.3 Wie dieses Buch strukturiert ist

Da es in diesem Buch um die Anwendung der Datenanalyse für die Organisationsentwicklung geht, decken die später besprochenen Fälle verschiedene Datenanalysesituationen in jedem Bereich der Wertschöpfungskette einer Organisation ab (Bild 1.1).

Die **Kundendomain** schließt die Erfassung, Verarbeitung und Analyse von kundenbezogenen Daten ein. Dazu gehören Umfragedaten aus verschiedenen Kundenumgebungen und Daten, die den „Moment der Wahrheit" messen, den Moment, in dem der Kunde das angebotene Produkt oder die angebotene Dienstleistung „erlebt".

Die **Prozessdomain** umfasst das Sammeln von Daten aus verschiedenen Betriebsumgebungen und die Umwandlung dieser Daten in kritische Informationen für die Entscheidungsfindung.

Die **Arbeitskräftedomain** bietet Ideen für den Umgang mit Daten aus dem Personalwesen, die verwendet werden, um Rückschlüsse auf verschiedene Aspekte im Zusammenhang mit der Belegschaft zu ziehen, wie z. B. Einstellung, Mitarbeiterfluktuation, Mitarbeiter-Engagement oder Personalplanung.

So wie im Unternehmen alle Bereiche zusammenarbeiten müssen, um Kundenforderungen unter bestmöglichen Bedingungen zu erfüllen, so ist es auch erforderlich, Daten aus unterschiedlichen Bereichen zu sammeln, zu verarbeiten und zu analysieren, um daraus unternehmensweit Schlussfolgerungen ziehen zu können. Unsere Fallbeispiele sollen dies verdeutlichen.

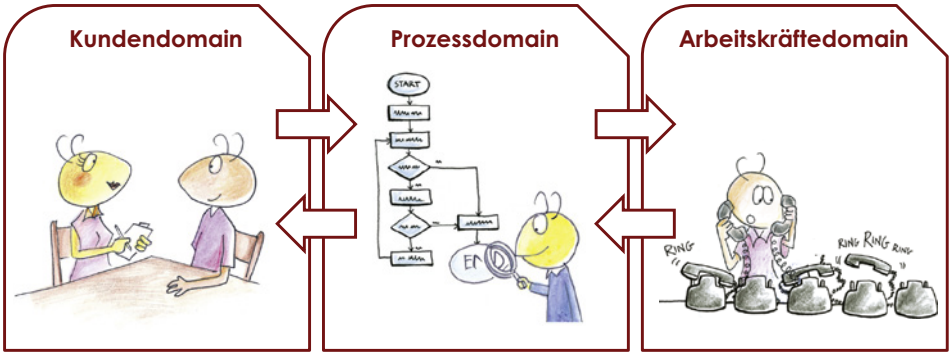


Bild 1.1 Bereiche der Wertschöpfungskette einer Organisation

Für jeden Fall verfolgen wir alle Schritte von der Aufgabenstellung, den Hypothesen oder Geschäftsfällen über die Stufen der Datenanalyse, um die richtige Auswahl zu treffen. Die in den folgenden Abschnitten erwähnten Schritte zeigt

Bild 1.2.: Formulieren einer geschäftsrelevanten Hypothese, Durchführen der Datenerfassung, der Datenvorbereitung und der Datenanalyse sowie das Ziehen von Schlussfolgerungen für das Geschäft.

Bild 1.2 Schritte eines Datenanalysefalls in Data Science

1.3.1 Geschäftsrelevante Frage formulieren

In den meisten Fällen beginnt die Datenanalyse aus der Verfügbarkeit von Daten. Dies kann zu einigen Erkenntnissen führen, die der Organisation sogar helfen können. Das kann jedoch auch in einer enormen Verschwendung von Zeit und Ressourcen aufgrund mangelnder Zweckmäßigkeit enden.

Der intelligentere Auslöser für die Datenanalyse ist eine geschäftsrelevante Frage wie „Warum vernichten wir mehr von unserem kostbaren gesammelten Blut als in vielen anderen Blutbanken?“

Diese Frage leitet sich nicht nur aus der Prozesseffizienz ab. Sie vermittelt auch die Botschaft, mehr Ressourcen als wahrscheinlich notwendig aufzuwenden. Dies ist für das Management immer von Interesse.

In diesem ersten Schritt muss das geschäftsbezogene Thema klar identifiziert werden. Und es muss in einen Indikator übersetzt werden, einen KPI (Key Performance Indicator), der das Thema messbar macht. Besser noch, dieser Indikator befindet sich auf der Scorecard oder dem Dashboard von Managementmitgliedern, d. h., er ist für jemanden wichtig.

1.3.2 Daten erfassen

Es gibt eine Vielzahl von Möglichkeiten, Daten zur Beantwortung der geschäftsrelevanten Frage zu sammeln. In der Regel ist es notwendig, die Methode der Datenerhebung zu validieren, um nützliche Daten für die Analyse zu gewährleisten, d. h. Daten, die repräsentativ, reproduzierbar und genau genug sind, um ausreichende Informationen für die Beantwortung der Geschäftsfrage zu liefern. Es gibt statistische Instrumente, die dabei helfen, potenzielle Probleme innerhalb des Datenerhebungsprozesses zu identifizieren.

In unserem Beispiel der Blutbank war mit der Qualität des Blutes alles in Ordnung. Es war die Methode der Datenerhebung, die zu falschen Schlussfolgerungen führte.

1.3.3 Daten vorbereiten

Selbst wenn sich die Methode der Datenerhebung bewährt hat und das Instrument statistisch akzeptiert ist, kann es sein, dass Daten nicht nützlich sind.

Bei Umfragen zum Beispiel geben einige Umfrageteilnehmer möglicherweise keinen nützlichen Input. Das könnte daran liegen, dass sie entweder zur Teilnahme an der Umfrage gezwungen oder angeregt wurden. Im Allgemeinen können wir davon ausgehen, dass sie dann nicht daran interessiert waren. Es kann daher sein, dass sie einen gültigen Input zu einem etablierten Fragebogen geliefert haben, aber der Input ist möglicherweise nicht hilfreich. Oder schlimmer noch, der Input könnte die folgenden Analyseschritte verderben und Ergebnisse verfälschen. Solche Eingaben könnten zufällige Bewertungszahlen sein oder dieselben Bewertungszahlen für alle Fragen oder Aussagen. Derart Inputs sind unbrauchbar und manchmal schädlich.

Daher ist eine Datenaufbereitung notwendig, um solche Eingaben zu finden und zu eliminieren, um nur Daten in die Analyse einzuspeisen, die wirklich wertschöpfend sind.

Zur Datenaufbereitung gehört auch die Formatierung der Daten, sodass sie von der bevorzugten Analysesoftware verwendet werden können. In den meisten Fällen sind die von einem System heruntergeladenen Daten nicht im richtigen Format, um in die Analysesoftware, z. B. Excel, importiert zu werden. In den meisten Fällen können Daten jedoch reorganisiert, neu formatiert oder transformiert werden, sodass die Software damit umgehen kann.

Nicht immer wird die Analysesoftware wegen der falsch formatierten Daten nicht mehr funktionieren. Im schlimmsten Fall kann sie einfach funktionieren und falsche Ergebnisse ausspucken.

1.3.4 Daten analysieren

Im Allgemeinen wird die Datenanalyse auf grafische und statistische Weise durchgeführt. In der Regel ist beides notwendig, um korrekte Schlussfolgerungen zu gewährleisten. Zusätzlich kann eine grafische Analyse zur Visualisierung der Daten und zum Storytelling erforderlich sein.

Eine grafische Analyse ohne statistische Unterstützung kann jedoch zu falschen Entscheidungen führen. Ähnliches gilt für die Durchführung statistischer Analysen ohne die Verwendung grafischer Werkzeuge.

Daher sollten alle Datenanalysen in einem zweistufigen Ansatz durchgeführt werden. Zuerst sollten ein oder mehrere Diagramme zur Visualisierung der Daten erstellt werden. Allein diese Visualisierung kann die Entscheidung beeinflussen.

Zweitens ist es eine gute Praxis und oftmals eine Notwendigkeit, die grafische Analyse durch Statistiken zu validieren.

Für die Analyse von Daten steht eine Vielzahl von Werkzeugen zur Verfügung. Die Auswahl des geeigneten Werkzeugs hängt von der zu beantwortenden Geschäftsfrage, der Art der gesammelten Daten und deren Eigenschaften ab. Die eine Entscheidung beeinflussenden Faktoren werden üblicherweise als „*X*“ bezeichnet. Das daraus resultierende Ergebnis wird gewöhnlich „*Y*“ genannt.

Wenn z. B. die Ablehnungsrate eines Produkts zwischen den Monaten Januar und April verglichen wird, dann bedeutet Monat die unabhängige Variable *X*, während die Ablehnungsrate die abhängige Variable *Y* bezeichnet. Die Anwendung von Werkzeugen hängt vom Datentyp ab, der in *X* und *Y* gefunden wird.

Monat ist zum Beispiel ein diskretes *X* und die Ablehnungsrate ist ein diskretes *Y*, das durch die Zählung zufriedener Kunden und nicht zufriedener Kunden erzeugt wird. Daher wird das obere linke Feld in Bild 1.3 verwendet. Da wir nur zwei Kategorien in *X* haben, Januar und April, wäre das geeignete statistische Werkzeug ein 2-Proportionen-Test.

Diese Aufstellung in Bild 1.3 wird in den folgenden Fällen zur Auswahl des anwendbaren grafischen und statistischen Werkzeugs herangezogen.

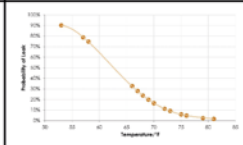
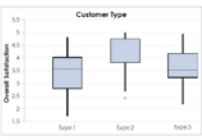
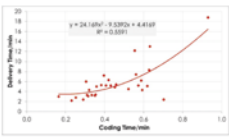
	Hypothesentests		**Regressionsanalysis**
Diskretes *Y*			
	2-Proportionen-Test	Chi²-Test	Logistische Regression Design of Experiments
Kontinuierliches *Y*			
	t-Test Tests auf Gleiche Varianzen Nichtparametrische Tests	ANOVA Tests auf Gleiche Varianzen Nichtparametrische Tests	Lineare Regression Nicht-lineare Regression Design of Experiments
	Diskretes *X*		**Kontinuierliches *X***

Bild 1.3 Analysewerkzeuge für verschiedene Datentypsituationen

Es wird eine Auswahl und Anwendung geeigneter grafischer und statistischer Instrumente demonstriert. Auf die Herleitung und detaillierte Beschreibung des statistischen Verfahrens wird verzichtet.

1.3.5 Geschäftsentscheidung vorbereiten

Sehr oft führt die Datenanalyse zu Ergebnissen, die für Mitarbeiter, die nicht in der Datenanalyse geschult sind, schwer zu verstehen sind. Ein „*p*-value“ zum Beispiel ist ein Schlüsselergebnis vieler statistischer Instrumente, das jedoch für die Mehrheit unverständlich ist.

Die Analyseausgabe wie „*p*-Wert = 0,03“ ist ein wichtiges Ergebnis, allerdings verwirrend für viele. Wenn es jedoch übersetzt wird in „Das Risiko, unser Geld zu verschwenden, wenn wir von diesem teureren Anbieter kaufen, beträgt nur 3 %“, verändert sich das Gespräch über Datenanalyse sofort.

Es ist nicht länger der Fall, sich bei der Übersetzung auf den Datenanalytiker oder Datenwissenschaftler verlassen zu müssen. Das Management sollte die Grundlagen der Datenanalyse verstehen, um Daten in Informationen umzuwandeln und entsprechende Schlussfolgerungen ziehen zu können.

Jeder der im Folgenden beschriebenen Fallbeispiele basiert auf einem realen Kundenprojekt. Um unsere Mandanten zu schützen, haben wir jedoch Namen und Daten geändert.

1.4 Welche Werkzeuge werden verwendet?

Unsere Absicht war es, ein Nachschlagewerk bereitzustellen, dem die Lernenden Schritt für Schritt folgen können. Dazu wird gängige Software verwendet. Aus unserer Arbeit mit unseren Kunden kennen wir deren Anforderungen an die eingesetzte Software, die wir hier umgesetzt haben:

Software muss leicht verfügbar sein. Nahezu jeder Mensch auf der Welt hat eine Version von Microsoft Office auf dem Computer. Integraler Bestandteil davon ist MS Excel. Ein Zusatzmodul, das MS-Excel-Benutzer herunterladen können – höchstwahrscheinlich kostenlos – ist MS Power BI. MS Excel enthält viele Funktionen, die bei den meisten der in diesem Buch beschriebenen Aufgaben der Datenerfassung, Datenaufbereitung und Datenanalyse helfen. Nur wenige Excel-Benutzer kennen das jedem zugängliche Add-In „Analyse-Funktionen“, das weitere statistische Tools in die MS-Excel-Umgebung einfügt.

MS Power BI erweitert die MS-Office-Umgebung mit leistungsfähigen Visualisierungs-Werkzeugen.

R ist eine Programmiersprache für statistische Berechnungen und Grafiken. R Studio bietet eine Benutzerschnittstelle und eine Entwicklungsumgebung für R. Beide Softwarepakete sind sowohl für Windows, als auch für MacOS und Linux kostenlos erhältlich.

Software muss einfach zu bedienen sein. Zumindest MS Excel ist eine Software, die fast jeder schon einmal benutzt hat. Das bedeutet, dass Analysten mit einer vertrauten Umgebung arbeiten können, indem sie einfach einige neue Tools hinzufügen. Fast dasselbe gilt für MS Power BI. Einerseits mag es für viele Nutzer neu sein, aber andererseits hat es die Microsoft-Benutzeroberfläche und viele Funktionen, die von MS Excel übernommen wurden. Die Lernkurve für MS Power BI sollte sehr steil und kurz sein.

R ist eine Programmiersprache und freie Softwareumgebung für statistische Berechnungen und Grafiken. Die Sprache R ist unter Statistikern und Datenanalysten für die Behandlung, Analyse und Visualisierung von Daten weit verbreitet. Das Erlernen von R (über R Studio) ist für Leute mit einem leichten Programmierhintergrund einfacher. Die Lernkurve für R könnte für viele etwas länger sein. Aber die Vorteile sind ausgezeichnet. R verfügt über eine schier endlose Sammlung vorgefertigter Funktionen, die jeden Tag wächst. R kann auch in MS Power BI integriert werden, sodass spezielle Funktionen und Diagramme in R erstellt und in der vertrauten MS-Umgebung angezeigt werden können.

Software muss mit anderer allgemein verwendeter Software kompatibel sein. Die Integration von MS-Excel-Tabellen und -Diagrammen in jede MS-PowerPoint-Präsentation ist so nahtlos wie möglich. Es besteht zusätzlich die Möglichkeit, MS Excel oder MS Power BI dynamisch mit der Datenquelle auf einem beliebigen Server oder einer beliebigen Website zu verknüpfen und die Analyseausgabe in MS PowerPoint einzufügen. Dies erlaubt es dem Analysten, die gewohnte beeindruckende PowerPoint-Präsentation mit den tatsächlichen Daten zu verwenden, wann immer PowerPoint aktualisiert wird.

Jedes Mal, wenn wir andere Software wie Minitab, SigmaXL, SAS oder SPSS für Ausbildungszwecke benutzten, war die begrenzte Verfügbarkeit dieser Softwarepakete im Unternehmen ein Hindernis für die Implementierung der neu erlernten Tools in der Organisation. Und wenn wir die Analysten mit der Universion oder Testversion einer Software unterrichteten, war es vorhersehbar, dass die Software nach Ablauf der Testphase nicht mehr verwendet wurde.

Wenn Sie bei Ihren Veränderungsbemühungen erfolgreich sein wollen, sollten Sie daher die genannten Punkte berücksichtigen. Ein Manager, der für Gewinn und Verlust verantwortlich ist, muss sorgfältig abwägen, ob eine Anzahl Lizenzen mit einer jährlichen Lizenzgebühr für eine neue Software sinnvoll ist, wenn MS Excel, MS Power BI Desktop und R oder Python kostenlos zur Verfügung stehen.

Die Fallbeispiele in diesem Buch zeigen daher Analysen, die mit MS Excel, MS Power BI und R Studio durchgeführt wurden. Um die Analyse nachzuvollziehen, müssen Sie ein MS-Excel-Add-In aktivieren und Power BI und R bzw. R Studio herunterladen.

1.5 Aktivieren und Verwenden der erforderlichen Software

Vielen MS-Excel-Benutzern sind die Tools, die in dieses sehr vertraute Office-Paket geladen werden, nicht bekannt. Nicht nur, dass MS Excel mit Funktionen für fast alle möglichen Datenmanipulations- und Analyseaufgaben ausgestattet ist. Es wird auch ein Analyse-Paket mitgeliefert, das kaum benutzt wird.

Dieses muss nur aktiviert werden, damit es als eine Sammlung von Makros erscheint, die das Potenzial haben, Ihre Analysearbeit erheblich zu erleichtern.

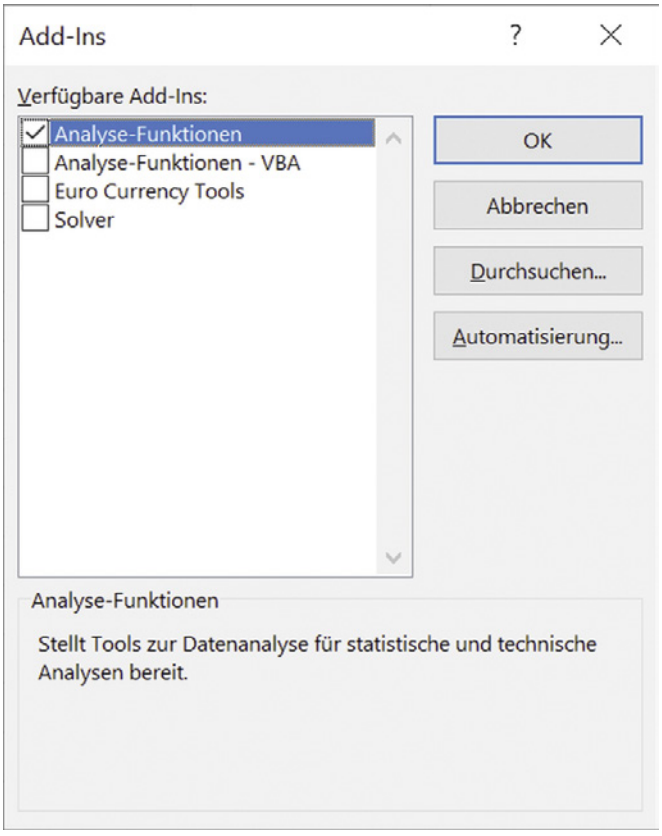


Bild 1.4
Aktivieren von MS Excels Analyse-Funktionen

Nach dem Laden von MS Excel wählen Sie Datei – Optionen – Add-Ins – Los ... und markieren das Kontrollkästchen Analyse-Funktionen (Bild 1.4). Dies ist alles, was Sie tun müssen, um eine Sammlung von häufig benötigten Analyse-Tools zu Ihrem Excel hinzuzufügen (Tabelle 1.1). Diese Tools finden Sie unter Daten – Datenanalyse (Bild 1.5).

Bild 1.5 Analyse-Funktionen in MS Excel

Nach der Aktivierung der Analyse-Funktionen stehen zusätzlich die in Tabelle 1.1 dargestellten Werkzeuge zur Verfügung.

Tabelle 1.1 In MS Excel Analyse-Funktionen verfügbare Tools

Deskriptive Tools	Präskriptive Tools	Regressions-analysen	Hypothesentests
Fourieranalyse Histogramm Populationskenngrößen Rang und Quantil Stichprobenziehung Zufallszahlengenerierung	Exponentielles Glätten Gleitender Durchschnitt	Kovarianz Korrelation Regression	Anova Gaußtest (*z*-Test) *t*-Test Zwei-Stichproben-*F*-Test

Machen wir uns mit dem neu entdeckten Instrumentarium besser vertraut.

Aufgabe 1.1: Erzeugen von Zufallsdaten

1. Öffnen Sie eine neue Excel-Tabelle.
2. In Reihe 1 benennen Sie Spalte A Norm und Spalte B Gruppe.
3. Wählen Sie **Daten** – **Datenanalyse** – **Zufallszahlengenerierung**.
4. Wählen Sie 1 für die **Anzahl der Variablen**, 1000 für die **Anzahl der Zufallszahlen**, Standard für die **Verteilung**, 100 für den **Mittelwert** und 5 für die **Standardabweichung** und platzieren Sie den Cursor in A2, nachdem Sie **Ausgabebereich** gewählt haben (Bild 1.6).
5. Wählen Sie **Daten** – **Datenanalyse** – **Zufallszahlengenerierung**.
6. Wählen Sie 1 für die **Anzahl der Variablen**, 1000 für die **Anzahl der Zufallszahlen**, Schematisch für die **Verteilung**, **Zwischen** 1 und 2, wobei **jede Zahl** 5-mal und **jede Folge** 100-mal wiederholt wird und der Cursor nach Auswahl von **Ausgabebereich** in B2 gesetzt wird (Bild 1.6).

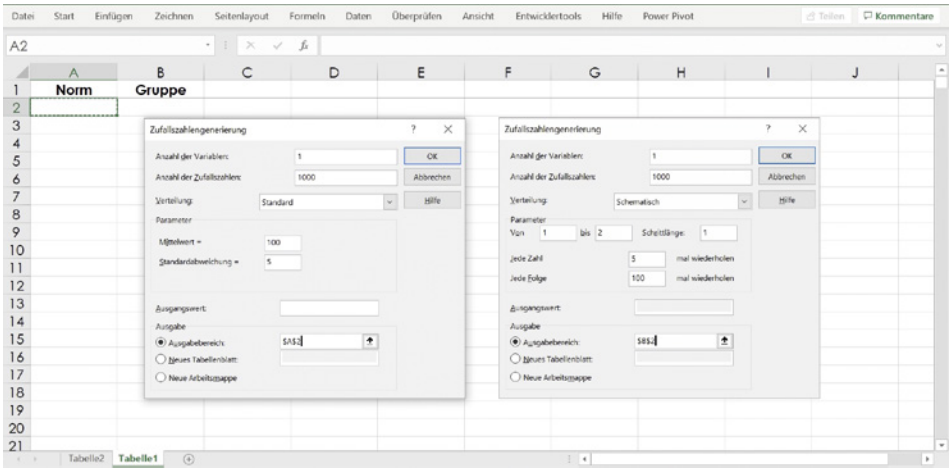
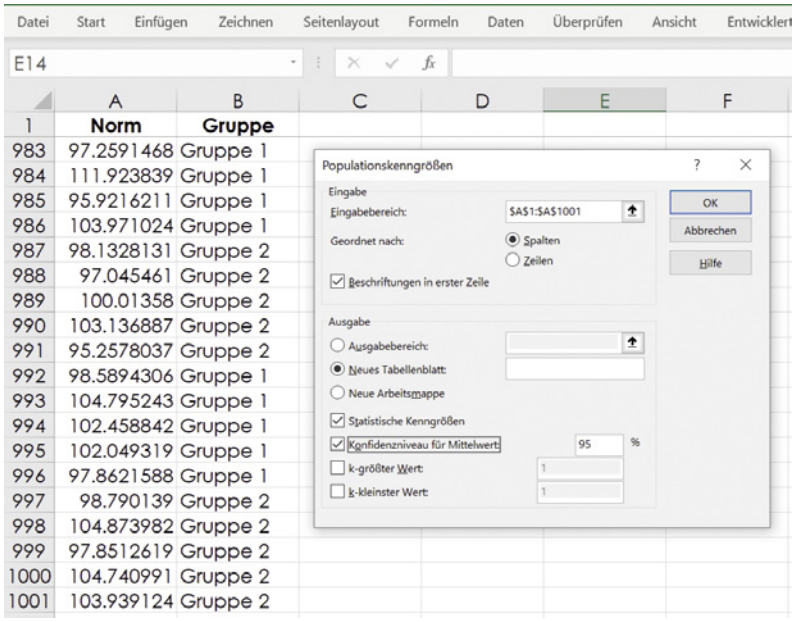


Bild 1.6 Generieren von zwei Spalten mit Zufallsdaten

Nach dem Generieren der Daten (Bild 1.7) werden Sie ein anderes Ergebnis auf Ihrem Arbeitsblatt haben.

Aufgabe 1.2: Analyse der deskriptiven Statistik von Norm

1. Wählen Spalte Gruppe, dann **Start – Suchen und Auswählen – Ersetzen** und ersetzen Sie 1 mit Gruppe 1 und 2 mit Gruppe 2.
2. Wählen Sie **Daten – Datenanalyse – Populationskenngrößen**.
3. Wählen Sie $A$1:$A$1001 als Eingabebereich. Oder wählen Sie einfach A1 mit dem Cursor und wählen Sie dann **Strg + ↑ + ↓**, um den gesamten Datenbereich zu markieren.
4. Wählen Sie **Neues Tabellenblatt** und markieren Sie **Statistische Kenngrößen** und **Konfidenzniveau für Mittelwerte bei 95 %** (Bild 1.7).



	A	B	C	D	E	F
1	Norm	Gruppe				
983	97.2591468	Gruppe 1				
984	111.923839	Gruppe 1				
985	95.9216211	Gruppe 1				
986	103.971024	Gruppe 1				
987	98.1328131	Gruppe 2				
988	97.045461	Gruppe 2				
989	100.01358	Gruppe 2				
990	103.136887	Gruppe 2				
991	95.2578037	Gruppe 2				
992	98.5894306	Gruppe 1				
993	104.795243	Gruppe 1				
994	102.458842	Gruppe 1				
995	102.049319	Gruppe 1				
996	97.8621588	Gruppe 1				
997	98.790139	Gruppe 2				
998	104.873982	Gruppe 2				
999	97.8512619	Gruppe 2				
1000	104.740991	Gruppe 2				
1001	103.939124	Gruppe 2				

Bild 1.7 Bestimmung der deskriptiven Statistik für Norm

Als Ergebnis werden die statistischen Kenngrößen für unsere Daten angezeigt (Bild 1.8), Ihre statistischen Kenngrößen werden anders aussehen.

Aufgabe 1.3: Histogramm für Norm aufzeichnen

1. Wählen Sie Norm auf Blatt1 (Markieren Sie A1 und wählen Sie **Strg** + ↑ + ↓).
2. Wählen Sie **Einfügen** – **Diagramme** – **Histogramm**.
3. Speichern Sie Ihre Arbeit unter dem Namen Norm.xlsx auf dem Desktop oder in einem Ordner Ihrer Wahl.

Norm	
Mittelwert	99.89
Standardfehler	0.16
Median	99.74
Modus	102.16
Standardabweichung	4.94
Stichprobenvarianz	24.45
Kurtosis	0.02
Schiefe	-0.03
Wertebereich	35.36
Minimum	79.95
Maximum	115.31
Summe	99891.74
Anzahl	1000.00
Konfidenzniveau (95.0%)	0.31

Bild 1.8 Deskriptive Statistik für Norm

Das Histogramm für Norm wird erstellt (Bild 1.9). Dieses Histogramm wurde verschönert, indem die *X*-Achse, in die Breite gezogen und ein Diagrammtitel hinzugefügt wurden. Dieses Histogramm könnte in eine PowerPoint-Präsentation, in ein Word-Dokument, in eine E-Mail oder einen Chat eingefügt zu werden.

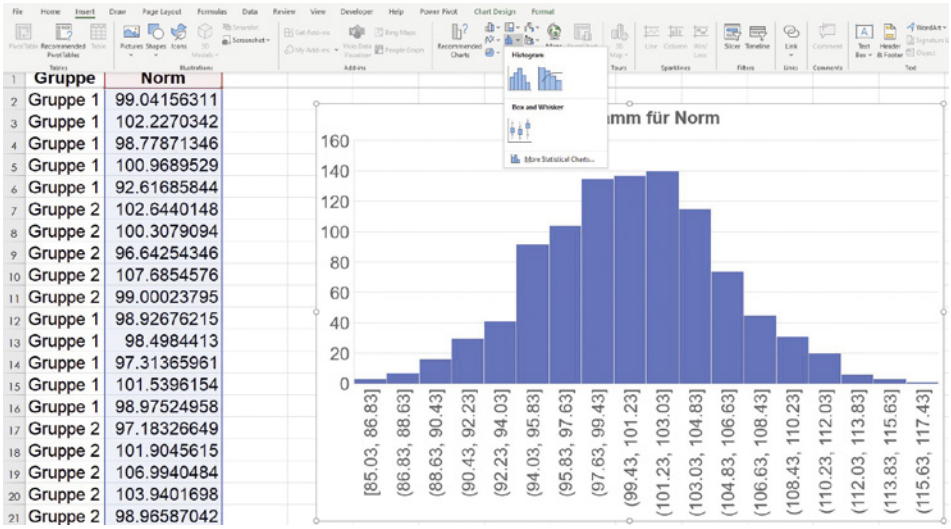


Bild 1.9 Histogramm für Norm

Aufgabe 1.4: Boxplot für Norm nach Gruppe anzeigen

1. Wählen Sie Norm auf Blatt1 (markieren Sie A1:B1 und wählen Sie **Strg** + ↑ + ↓).
2. Wählen Sie **Einfügen** – **Diagramme** – **Kastengrafik**.

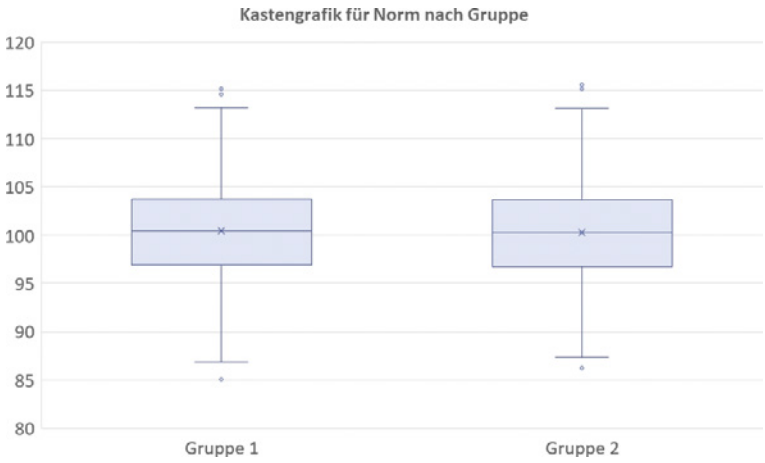


Bild 1.10 Kastendiagramm von Daten in Norm nach Gruppe

Dieser Befehl führt zu dem in Bild 1.10 gezeigten Kastendiagramm oder Box-Plot.

Herunterladen und Verwenden von MS Power BI

MS Power BI ist im Microsoft Store zum Download bereit, wenn MS Office auf Ihrem Computer läuft. Mit Microsoft Power BI stehen Ihnen Visualisierungen zur Verfügung. Mit dem optionalen Power-BI-Webdienst können Sie Ihre Visualisierungen, d. h. Ihre Dashboards, über den Power-BI-Server mit Ihrem Team gemeinsam nutzen.

Nach dem Download kann es erforderlich sein, eine kostenlose Registrierung vorzunehmen.

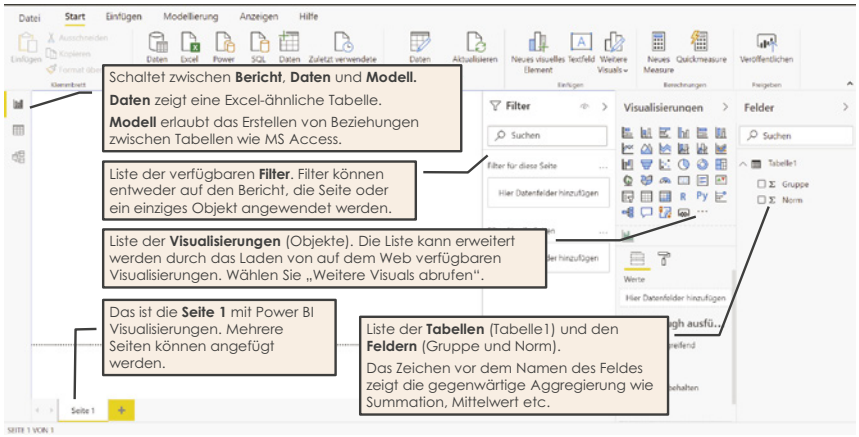


Bild 1.11 Power BI Benutzeroberfläche

Aufgabe 1.5: Laden der Daten aus Norm.xlsx in Power BI

1. Öffnen Sie MS Power BI-Desktop (Bild 1.11).
2. Wählen Sie **Daten abrufen** – **Excel** – **Verbinden**.
3. Öffnen Sie Norm.xlsx an dem Ort, an dem Sie es gespeichert haben.
4. Aktivieren Sie das Kästchen vor Tabelle 1 – **Laden** (Blatt1 trägt die Datentabelle).

Im Gegensatz zu Excel erscheinen die Daten nicht als Tabelle auf dem Bildschirm. Datentabellen und deren Spalten (Variablen) sind unter Felder auf der rechten Seite aufgeführt (Bild 1.11).

Aufgabe 1.6: Laden eines Bildmaterials für ein Histogramm aus dem Power-BI-Repository

1. Wählen Sie **...** unter **Visualisierungen** und **Weitere Visuals abrufen.**
2. Suchen Sie nach **Histogramm**.
3. **Histogram Chart hinzufügen** (oder ein anderes Histogramm Ihrer Wahl).
4. Wählen Sie das neu importierte **Histogramm**-Symbol.
5. Verändern Sie die Größe des Bereichs nach Ihren Wünschen.
6. Stellen Sie sicher, dass der Histogramm-Bereich ausgewählt ist.
7. Ziehen Sie das Feld Norm von rechts in das Feld **Werte** unter **Visualisierungen**.
8. Ändern Sie das **Format** (Malerbürste) von Titel, Achsen usw.

Power BI gibt es für Desktop- und mobile Geräte. Alle Versionen sind über den Microsoft Store erhältlich.

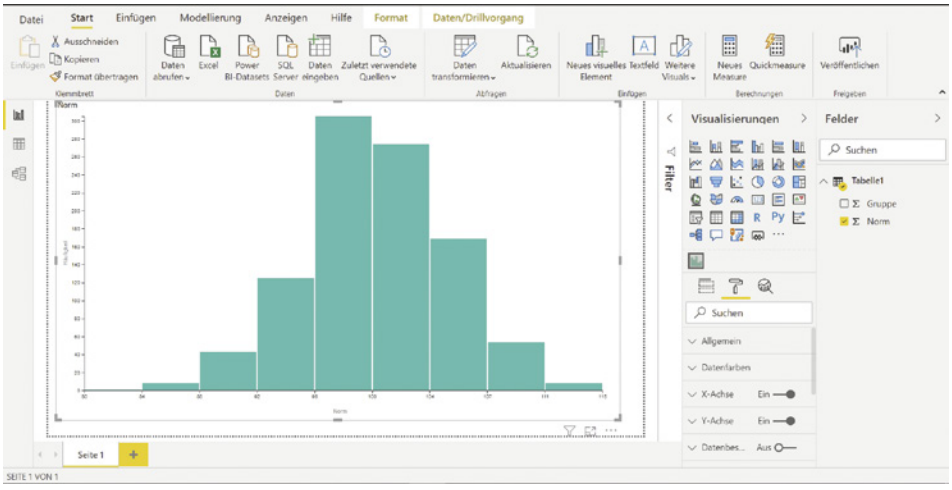


Bild 1.12 Power-BI-Histogramm von Daten in Norm

Herunterladen und Verwenden von R und R Studio

Installieren Sie R und R Studio von einem beliebigen Server, der auf den Websites bereitgestellt wird:

- Laden Sie R von *https://www.r-project.org/* herunter.
- Laden Sie R Studio von *https://www.rstudio.com/* herunter.
- Starten Sie R Studio.

Die R-Software bietet einen umfangreichen und ständig wachsenden Satz an Bibliotheken, die bei der Programmierung von Routinen, der Datenanalyse und der Visualisierung von Daten mit minimalem Code und großer Flexibilität helfen. Die meisten dieser Bibliotheken müssen geladen werden, bevor Funktionen verwendet werden können. Wir werden dies im Detail während unserer Fallanwendungen in diesem Buch erläutern.

Aufgabe 1.7: Laden der Daten aus Norm.xlsx in R Studio

1. Öffnen Sie R Studio (Bild 1.13).
2. Wählen Sie **Files**, navigieren Sie zur früher gespeicherten Datei Norm.xlsx, wählen Sie **Import Dataset** – **Sheet:** Tabelle 1 – **Range:** A1:B1001 – **Import**.
3. Die Datentabelle Norm wird im Register Norm und in Environment angezeigt.
4. Suchen Sie im Register **Packages** nach „pastecs“; wenn gefunden, wählen Sie es aus.
5. Falls nicht gefunden, wählen Sie **Install**.
6. **Install Packages** – **Packages:** pastecs – **Install**.

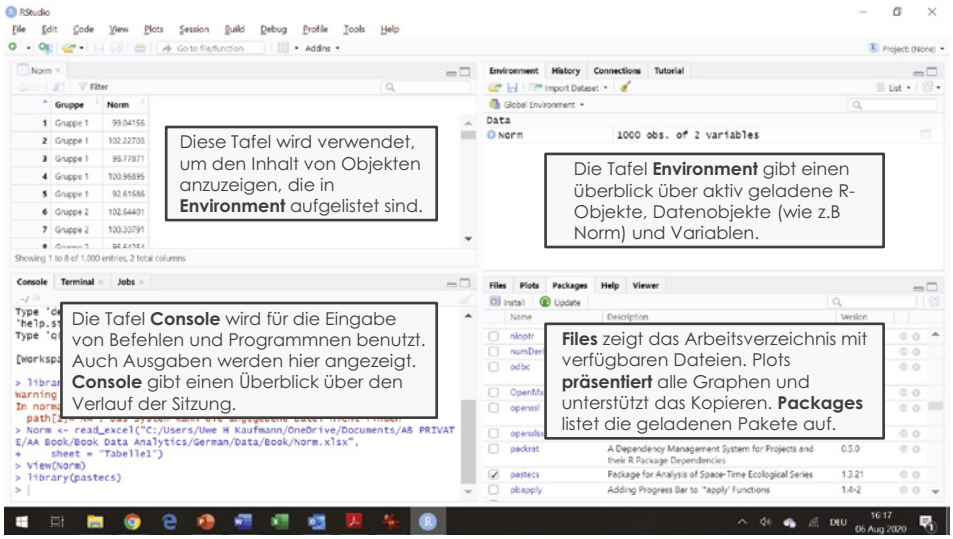


Bild 1.13 R-Studio-Benutzeroberfläche

Aufgabe 1.8: Deskriptive Statistik der Daten in Norm analysieren

Geben Sie Folgendes in die Konsole (Console) ein:

 # Installation einer Erforderlichen Bibliothek

install.packages("pastecs")

library("pastecs")

 # 1. Beschreibende Statistiken für alle Spalten im Datenrahmen (Tabelle) Norm anzeigen

stat.desc(Norm)

 # 2. Beschreibende Statistik für Spalte Norm im Datenrahmen (Tabelle) Norm anzeigen

stat.desc(Norm$Norm)

Ausgabe:

1. Alle Spalten der Tabelle Norm sind in der deskriptiven Statistik enthalten (Bild 1.14).

	Gruppe	Norm
nbr.val	NA	1000
nbr.null	NA	0
nbr.na	NA	0
min	NA	85.02608
max	NA	115.6317
range	NA	30.60559
sum	NA	100,354.1
median	NA	100.4036
mean	NA	100.3541
SE.mean	NA	0.1593795
CI.mean	NA	0.3127570
var	NA	25.40182
std.dev	NA	5.040022
coef.var	NA	0.05022239

2. Nur die Spalte Norm in der Tabelle Norm ist in der deskriptiven Statistik enthalten (Bild 1.14)

nbr.val	nbr.null	nbr.na	min	max
1.000000e+03	0.000000e+00	0.000000e+00	8.502608e+01	1.156317e+02
range	sum-	median	mean	SE.mean
3.060559e+01	1.003541e+05	1.004036e+02	1.003541e+02	1.593795e-01
CI.mean.0.95	var	std.dev	coef.var	
3.127570e-01	2.540182e+01	5.040022e+00	5.022239e-02	

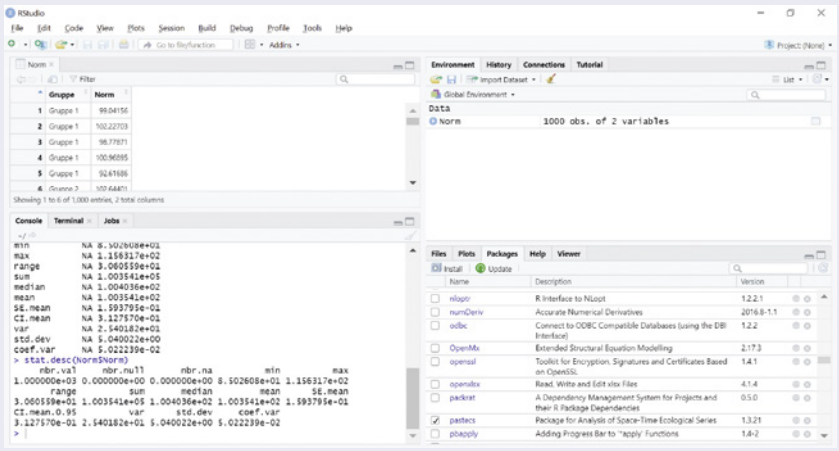


Bild 1.14 Deskriptive Statistik für Norm

Die Ausgabe der deskriptiven Statistik gibt einen Überblick über die Parameter, die den Datensatz beschreiben. Sie ist die grundlegendste Analyse, um Schlussfolgerungen über die zentrale Tendenz, Variation und Form der Verteilung des Datensatzes zu ziehen.

Aufgabe 1.9: Normalitätstest (Shapiro-Wilk-Test) für Norm durchführen

 # Durchführung eines Tests der Normalverteilung der Spalte Norm
shapiro.test(Norm$Norm)

```
shapiro.test(Norm$Norm)
```

Ausgabe:

```
Shapiro-Wilk normality test
data: Norm$Norm
W = 0.99896, p-value = 0.8525
```

Um eine Normalverteilung der Daten annehmen zu können, muss der *p*-Value des Tests größer als 0,05 sein. Daher sagen die Indikatoren des Normalverteilungstests für Norm aus, dass Norm einer Normalverteilung folgt (Bild 1.15), d. h. *p*-value > 0,05.

Aufgabe 1.10: Histogramm für Norm zeichnen

 # Plotten eines Histogramms für Norm

```
hist(Norm$Norm)
```