



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



Bibliografische Information der Deutschen Nationalbibliothek:

Die Deutsche Nationalbibliothek verzeichnetet 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

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutzgesetzgebung als frei zu betrachten wären und daher von jedem benutzt werden dürften.

Alle in diesem Buch enthaltenen Verfahren bzw. Daten wurden nach bestem Wissen dargestellt. Dennoch sind Fehler nicht ganz auszuschließen.

Aus diesem Grund sind die in diesem Buch enthaltenen Darstellungen und Daten mit keiner Verpflichtung oder Garantie irgendeiner Art verbunden. Autoren und Verlag übernehmen infolgedessen keine Verantwortung und werden keine daraus folgende oder sonstige Haftung übernehmen, die auf irgend eine Art aus der Benutzung dieser Darstellungen oder Daten oder Teilen davon entsteht.

Dieses Werk ist urheberrechtlich geschützt.

Alle Rechte, auch die der Übersetzung, des Nachdruckes und der Vervielfältigung des Buches oder Teilen daraus, vorbehalten. Kein Teil des Werkes darf ohne schriftliche Einwilligung des Verlages in irgendeiner Form (Fotokopie, Mikrofilm oder einem anderen Verfahren), auch nicht für Zwecke der Unterrichtsgestaltung – mit Ausnahme der in den §§ 53, 54 URG genannten Sonderfälle –, reproduziert oder unter Verwendung elektronischer Systeme verarbeitet, vervielfältigt oder verbreitet werden.

Die Rechte aller Grafiken und Bilder liegen bei den Autoren.

© 2021 Carl Hanser Verlag GmbH & Co.KG, München

www.hanser-fachbuch.de

Lektorat: Lisa Hoffmann-Bäuml

Herstellung: Carolin Benedix

Satz: Eberl & Koesel 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 Datensätzen 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

Vorwort	VII
Entfesseln Sie das Potenzial Ihrer Daten!	IX
1 Einführung	1
1.1 Warum Datenanalytik wichtig ist	1
1.2 Warum dieses Buch geschrieben wurde	3
1.3 Wie dieses Buch strukturiert ist	6
1.3.1 Geschäftsrelevante Frage formulieren	7
1.3.2 Daten erfassen	7
1.3.3 Daten vorbereiten	8
1.3.4 Daten analysieren	8
1.3.5 Geschäftsentscheidung vorbereiten	10
1.4 Welche Werkzeuge werden verwendet?	10
1.5 Aktivieren und Verwenden der erforderlichen Software	12
1.6 Was wird bereitgestellt	24
1.7 Welche Fallbeispiele sollte ich studieren?	24
2 Data Science und Datenanalytik	29
2.1 Komponenten der Datenanalytik	29
2.2 Big Data und ihre Beziehung zur Datenanalytik	30
2.3 Voraussetzung für Data Science und künstliche Intelligenz	33
3 Phasen von Data Science und Datenanalytik	35
3.1 Geschäftsrelevante Frage formulieren	36
3.2 Daten erfassen	37

3.3	Daten vorbereiten	39
3.4	Daten analysieren	42
3.4.1	Deskriptive Statistik	42
3.4.2	Normalverteilung	43
3.4.3	Arten von Daten	48
3.4.4	Werkzeuge für die Datenanalyse	50
3.5	Geschäftsentscheidung vorbereiten	55
3.6	Geschäftsentscheidung kommunizieren – Storytelling	56
3.6.1	Wer ist das Publikum?	56
3.6.2	Wie werden die Daten angezeigt?	57
3.6.3	Was ist der Zweck der Präsentation?	58
3.6.4	Wie kann die Präsentation vereinfacht werden?	59
3.7	Überlegungen zu den wichtigsten verwendeten Analysewerkzeugen ..	60
4	Kompetenzen eines Datenanalytikers	63
4.1	Benötigte Kompetenzen in den Phasen der Datenanalytik	63
4.2	Schlüsselrollen der heutigen Manager und Führungskräfte	67
5	Die Stimme des Kunden	73
5.1	Warum Kundenanalytik?	73
5.1.1	Hören Sie auf die Stimme Ihrer bestehenden Kunden	74
5.1.2	Kundenerwartungen verstehen	76
5.1.3	Untersuchen der Kundenerfahrung	77
5.2	Entwerfen von Kundenumfragen	79
5.2.1	Entwicklung und Durchführung einer eigenen Umfrage	80
5.2.2	Schlussfolgerung	83
6	Fall: Toll, wir haben uns verbessert ... oder nicht?	85
6.1	Das Problem der Stichprobe	86
6.2	Geschäftsrelevante Frage formulieren	88
6.3	Daten erfassen	88
6.4	Daten vorbereiten	89
6.5	Daten analysieren	89
6.6	Geschäftsentscheidung vorbereiten	92

6.7	Was wäre, wenn wir alle Rohdaten hätten?	93
6.8	Überlegungen zu den wichtigsten verwendeten Analysewerkzeugen ..	97
7	Fall: Was beeinflusst unsere Patientenzufriedenheit?	101
7.1	Analysieren der Treiber der Kundenzufriedenheitswerte	101
7.2	Aufbau der Patientenumfrage	105
7.3	Geschäftsrelevante Frage formulieren	107
7.3.1	Hypothese 1: Es besteht ein signifikanter Unterschied zwischen den Bewertungen für die Prozessschritte – Mindestens ein Schritt wird unterschiedlich bewertet	107
7.3.2	Hypothese 2: Es besteht ein signifikanter Unterschied zwischen den Bewertungen der Indikatoren – Mindestens ein Indikator wird unterschiedlich bewertet	107
7.3.3	Hypothese 3: Es besteht eine signifikante Beziehung zwischen der Bewertung für einen Prozessschritt und der Gesamtbewertung	108
7.3.4	Hypothese 4: Es gibt ein Muster über die Zeit	108
7.4	Daten erfassen	108
7.5	Daten vorbereiten	109
7.5.1	Daten transformieren	109
7.5.2	Umgang mit nicht hilfreichen Eingaben	110
7.5.3	Umgang mit fehlenden Eingaben	111
7.6	Daten analysieren	112
7.6.1	Deskriptive Statistik und Darstellung	112
7.6.2	Hypothese 1: Es besteht ein signifikanter Unterschied zwischen den Bewertungen für Schritte (X)	115
7.6.3	Hypothese 2: Es besteht ein signifikanter Unterschied zwischen den Bewertungen der Indikatoren (X)	123
7.6.4	Hypothese 3: Es besteht eine signifikante Beziehung zwischen der Bewertung für mindestens einen der Prozessschritte (Step1 ... Step5, X) und der Gesamtbewertung (Overall, Y)	126
7.6.5	Hypothese 4: Es gibt ein Muster über die Zeit	137
7.7	Geschäftsentscheidung vorbereiten	140
7.8	Überlegungen zu den wichtigsten verwendeten Analysewerkzeugen ..	142

8 Fall: Wie erstellt man ein Dashboard zur Patientenzufriedenheit	147
8.1 Entscheidung über Metriken zur Veranschaulichung der Klinikleistungsbewertung	147
8.2 Aufbau eines Klinik-Dashboards mit MS Power BI und R	148
8.3 Verwendung von MS Power BI für analytisches Storytelling	163
8.4 Schlussfolgerung	166
9 Ohne Prozess läuft nichts	171
9.1 Warum Prozessanalytik?	171
9.2 Dimensionen der Prozessanalytik	174
9.2.1 Prozessdesign und Analytik	174
9.2.2 Definieren von Indikatoren für die Analytik	176
9.2.3 Prozessmanagement mit Analytik	178
9.2.4 Prozessverbesserung durch Analytik mit DMAIC	180
9.3 Rollen und Einsatz von Prozessanalytik	181
9.4 Schlussfolgerung	184
10 Fall: Welcher Anbieter hat die bessere Produktqualität?	187
10.1 Geschäftsrelevante Frage formulieren	187
10.2 Daten erfassen und vorbereiten	188
10.3 Daten analysieren	188
10.4 Geschäftsentscheidung treffen	198
10.5 Überlegungen zu den wichtigsten verwendeten Analysewerkzeugen ..	199
11 Fall: Warum zahlt die Finanzabteilung unsere Auftragnehmer verspätet aus?	201
11.1 Geschäftsrelevante Frage formulieren	201
11.2 Daten erfassen	202
11.3 Daten vorbereiten	203
11.4 Daten analysieren	205
11.4.1 Hypothese: Einige Geschäftseinheiten sind besser als andere ...	208
11.4.2 Hypothese: Die Finanzabteilung erhält Rechnungen, nachdem die Zahlungsfrist abgelaufen ist	208
11.4.3 Hypothese: Geschäftseinheit 1 hat sich verbessert	210

11.5	Geschäftsentscheidung treffen	211
11.6	Überlegungen zu den wichtigsten verwendeten Analysewerkzeugen ..	212
11.6.1	Mosaikdiagramm	213
11.6.2	Geigendiagramm	213
11.6.3	Nachweis eines signifikanten Unterschieds zwischen Gruppen ..	214
11.7	Ein Dashboard zur „Lieferantenbuchhaltung“	216
12	Fall: Warum vergeuden wir kostbare Blutprodukte?	219
12.1	Geschäftsrelevante Frage formulieren	219
12.1.1	Hypothese 1: Arbeitsstress erhöht die Verschwendungen	220
12.1.2	Hypothese 2: Unterschiedliches Material von Blutbeuteln trägt zu höherer Verschwendungen bei	221
12.1.3	Hypothese 3: Blutplättchen-Verluste hängen vom Spendenort ab	221
12.1.4	Hypothese 4: Die Zeit, die für die Blutung benötigt wird, beeinflusst die Verschwendungen von Blutplättchen	221
12.1.5	Hypothese 5: Die Ruhezeit vor der Verarbeitung von Blutplättchen beeinflusst die Verschwendungen	221
12.1.6	Hypothese 6: Der verwendete Zentrifugentyp verursacht unterschiedliche Abfallmengen	222
12.1.7	Hypothese 7: Mitarbeiter tragen zu höherer Verschwendungen bei	222
12.2	Daten erfassen	222
12.3	Daten verarbeiten	223
12.4	Daten analysieren	226
12.4.1	Hypothese 1: Arbeitsstress erhöht die Verschwendungen	226
12.4.2	Hypothese 2: Bestimmtes Material von Blutbeuteln trägt zu höherer Verschwendungen bei	230
12.4.3	Hypothese 3: Thrombozyten-Verluste hängen vom Ort der Blutspende ab	232
12.4.4	Hypothese 4: Die Zeit, die für die Blutung benötigt wird, beeinflusst die Qualität von Blutplättchen	233
12.4.5	Hypothese 5: Die Ruhezeit vor der Verarbeitung von Blutplättchen beeinflusst die Qualität	235
12.4.6	Hypothese 6: Der verwendete Zentrifugen-Typ beeinflusst die Qualität der Blutplättchen	236
12.4.7	Hypothese 7: Einige Mitarbeiter tragen zu verminderter Qualität bei	239

12.4.8 Die kostspielige Frage	241
12.4.9 Investitionen sparen	241
12.5 Geschäftsentscheidung treffen	243
12.6 Schlussfolgerung	245
12.7 Überlegungen zu den wichtigsten verwendeten Analysewerkzeugen ..	245
13 Arbeitskräfte machen den Unterschied	251
13.1 Warum Personalanalytik?	251
13.2 Warum hat sich das Thema „Arbeitskräfte“ zu einer Priorität entwickelt?	252
13.3 Die Rolle von HR in der Personalanalytik	256
13.4 Dimensionen der Personalanalytik	258
13.5 Personalplanung	258
13.5.1 Personalplanung für transaktionale Aktivitäten	259
13.5.2 Personalplanung für weniger transaktionale Aktivitäten	261
13.6 Schritte zur Personalanalytik	262
13.6.1 Beginnen Sie mit einem Problem, das die Organisation lösen will	262
13.6.2 Benötigte Informationen ermitteln und Daten sammeln	263
13.6.2.1 Schritt 1: Identifizieren der potenziellen Treiber für das Problem	263
13.6.2.2 Schritt 2: Pilotdatenerfassung und -analyse durchführen	264
13.6.2.3 Schritt 3: Vollständige Datenerhebung durchführen	265
13.6.3 Analyse der Daten	265
13.6.4 Die geschäftsrelevante Antwort formulieren – Storytelling	265
13.6.5 Change-Management ist essenziell	266
13.7 Zusammenfassung	268
14 Fall: Was macht unsere Organisation innovativ?	271
14.1 Geschäftsrelevante Frage formulieren	271
14.2 Daten erfassen	272
14.3 Daten vorbereiten	273
14.4 Daten analysieren	275
14.4.1 Vergleichen des innovativen Arbeitsverhaltens zwischen Abteilungen	276
14.4.2 Ermitteln der Treiber für innovatives Arbeitsverhalten	278

14.5 GeschäftSENTscheidung treffen	280
14.6 Überlegungen zu den wichtigsten verwendeten Analysewerkzeugen ..	284
15 Fall: Ist unsere Personalstärke angemessen?	285
15.1 Geschäftsrelevante Frage formulieren	285
15.2 Daten erfassen und vorbereiten	286
15.3 Daten analysieren	289
15.3.1 Die Nachfragestruktur verstehen	289
15.3.2 Vorhersage eines möglichen zukünftigen Problems	294
15.3.3 Verstehen des Aktivitätsmusters	295
15.4 GeschäftSENTscheidung treffen	297
15.4.1 Planung der Arbeitskräfte	297
15.4.2 „Kampf gegen die Variation“	299
15.4.2.1 „Trainieren“ der Kunden	299
15.4.2.2 Flexible Arbeitsvereinbarungen für das Personal	299
15.4.2.3 Spitzenzeiten mit Zeitarbeitskräften abdecken	300
15.4.2.4 Angestellte auf der Gehaltsliste sind nicht 100% verfügbar	300
15.4.3 Den Prozess überdenken und erneuern	301
15.5 Schlussfolgerung	302
15.6 Überlegungen zu den wichtigsten verwendeten Analysewerkzeugen ..	302
16 Fall: Was bedeutet das Ergebnis unserer Umfrage zum Engagement?	305
16.1 Geschäftsrelevante Frage formulieren	306
16.2 Daten erfassen und vorbereiten	307
16.3 Daten analysieren	307
16.4 Geschäftsrelevante Entscheidung treffen	313
16.5 Überlegungen zu den wichtigsten verwendeten Analysewerkzeugen ..	314
17 Fall: Was verursacht unsere Personalfluktuation?	319
17.1 Geschäftsrelevante Frage formulieren	319
17.2 Daten erfassen	320
17.2.1 Vergleich mit Benchmarks	323
17.2.2 Verwendung von Proxy-Messungen	324

17.2.3 Einsatz von Direktmessungen	325
17.2.4 Einbeziehen einer Kontrollgruppe	326
17.2.5 Aufnehmen von demografischen Faktoren	326
17.3 Daten vorbereiten	327
17.4 Daten analysieren	330
17.4.1 Hypothese 1: Der Umfang an Geschäftsreisen hat einen signifikanten Einfluss auf die Absicht der Mitarbeiter, das Unternehmen zu verlassen	332
17.4.2 Hypothese 2: Die Kündigungsabsicht ist für verschiedene Organisationseinheiten unterschiedlich	334
17.4.3 Hypothese 3: Die Entscheidung des Personals zur Kündigung hängt von dessen Familienstand ab	337
17.4.4 Hypothese 4: Die Entscheidung der Mitarbeiter, zu kündigen, hängt von ihrem Geschlecht ab	338
17.4.5 Hypothese 5: Die Kündigungsentscheidung des Personals hängt von den gegebenen Ausbildungsmöglichkeiten ab	338
17.4.6 Hypothesen 6 -14: Die Entscheidung des Personals, zu kündigen, hängt von den Faktoren Job Level, Alter, Amtszeit, Entfernung zum Wohnort, Bildungsniveau, Arbeitsmotivation, Führungspraxis, Teamarbeit und Gehalt ab	345
17.4.7 Berechnung der Vorhersagegenauigkeit	353
17.4.8 Priorisierung von Prädiktoren	355
17.5 Geschäftsentscheidung vorbereiten	358
17.5.1 Haupttriebkräfte zur Fluktuation von Mitarbeitern	358
17.5.2 Modell zur Vorhersage von Personal, das die Organisation verlässt	359
17.5.3 Maßnahmen zur Eindämmung der Fluktuation	359
17.6 Schlussfolgerung	360
17.7 Überlegungen zu den wichtigsten verwendeten Analysewerkzeugen ..	361
18 Bessere Entscheidungen treffen	369
18.1 Mögliche Fehler bei der Entscheidungsfindung	370
18.1.1 Fall 1: Es gibt keinen Unterschied, und wir entscheiden, dass es keinen gibt – kein Fehler	370
18.1.2 Fall 2: Es gibt keinen Unterschied und wir entscheiden, dass es einen gibt – Typ-I-Fehler	370

18.1.3 Fall 3: Es gibt einen Unterschied, und wir entscheiden, dass es einen gibt – kein Fehler	371
18.1.4 Fall 4: Es gibt einen Unterschied, aber wir entscheiden, dass es keinen gibt – Typ-II-Fehler	371
18.2 Bessere Entscheidungen treffen – der Statistik nicht blind vertrauen ..	373
18.2.1 Signifikanter Unterschied bedeutet nicht wichtiger Unterschied ..	374
18.2.2 Ein nichtsignifikanter Unterschied könnte für die Organisation wichtig sein	374
18.3 Schlussfolgerung	375
19 Sicherstellung des Erfolgs	377
19.1 Grundsätzliches	377
19.2 Schritte zur Implementierung der Datenanalytik	379
19.3 Managementunterstützung sicherstellen	381
19.4 Begeisterung für die Datenanalytik und deren Vorteile erzeugen	383
19.5 Wissen aufbauen – fangen Sie klein an	386
19.6 Analysen zum Aufbrechen von Silos verwenden	387
19.7 Kreislauf schließen	388
19.8 Datenanalytik-Implementierung überprüfen	389
20 Literatur und Links	393
Literatur	393
Links	395
21 Stichwortverzeichnis	397
22 Zusatzmaterial zum Download	399
23 Die Autoren	401

■ 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 Verschwendungen 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 geschickte 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.

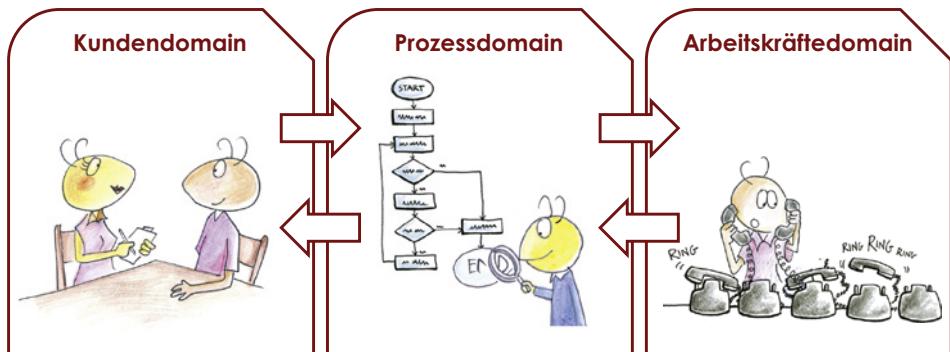


Bild 1.1 Bereiche der Wertschöpfungskette einer Organisation

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 zeigen

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 Verschwendungen 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, so dass 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.

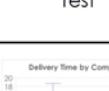
	Hypothesentests		Regressionsanalyse
Diskretes Y	 January = Good ■ Rejected	 Step 1 ■ Step 2 ■ Step 3 ■ Step 4 ■ Step 5	 Sales % ■ Temperature °F
	2-Proportionen-Test	Chi²-Test	Logistische Regression Design of Experiments
Kontinuierliches Y	 CHL Red ■ CHL Green	 Overall Satisfaction ■ Customer Type	 $y = 24.18h^2 - 9.339h + 4.418f$ $R^2 = 0.5591$ Coding Time min ■
	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