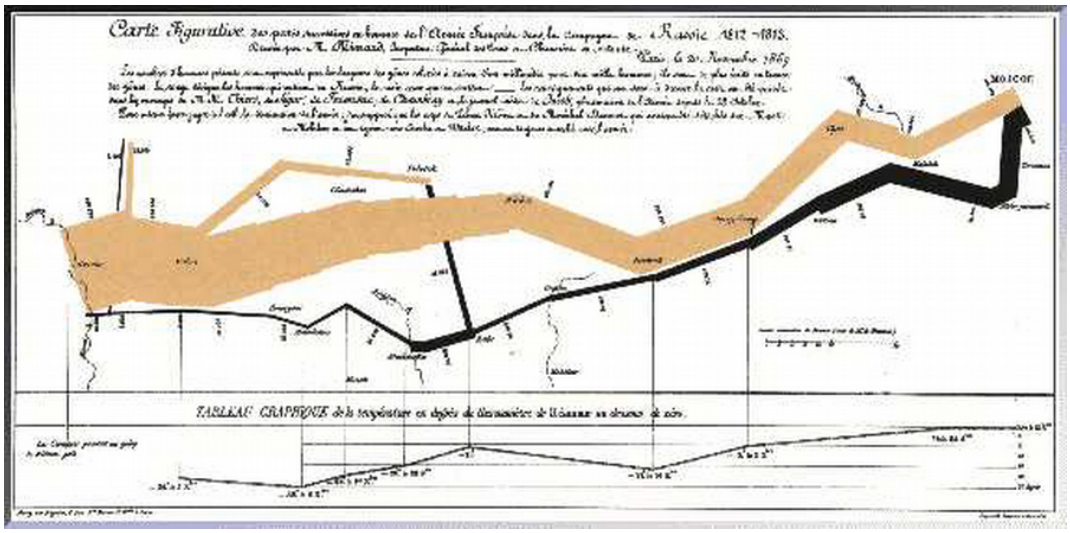
Datensätze für Lehrveranstaltungen

Manfred Brill

Stand: Sommersemester 2022



Inhaltsverzeichnis

1 Verzeichnis der Datensätze

Dieses Dokument enthält die Datensätze, die in Lehrveranstaltungen der Stochastik und Datenanalyse eingesetzt werden.

Ein Datensatz hat immer eine *Kurzbezeichnung*, die auch in Aufgaben, Folien und anderen Texten verwendet wird. Das Format der Dokumentation lehnt sich an die Dokumentation von Datensätzen für die Statistik-Software `R` an. Die Datensätze sind in verschiedene Kategorien eingeteilt. Auf Seite 59 finden Sie ein alphabetisches Verzeichnis aller Datensätze.

Dateiformate

Als Dateiformate werden zur Zeit `csv` und `json` verwendet. In den `csv`-Dateien wird das Semikolon „;“ als Trennsymbol verwendet. Dezimalzahlen sind mit der Einstellung „Deutschland“ gespeichert, also mit einem *Dezimalkomma*.

2 Datensätze mit einem Merkmal

Die hier aufgeführten Datensätze enthalten ein Merkmal und sind meist der Statistik-Literatur entnommen. Die Aufzählung ist alphabetisch geordnet.

Bedienungszeiten **Bedienungszeiten an einer Scannerkasse**

Beschreibung

An der Scanner-Kasse eines Supermarkts wurden für 50 aufeinander folgende Kunden die Bedienungszeiten in Sekunden registriert:

40	20	22	15	18	51	37	42	31	58
33	39	49	22	23	62	42	53	43	44
19	49	39	36	37	38	22	24	32	29
41	40	39	38	27	51	52	54	28	22
64	19	50	40	18	68	51	41	48	57

Quelle ist [BBK08].

Format

50 ganze Zahlen mit den Zeitangaben in Sekunden.

Dateien

bedienungszeiten.csv

Die Datei enthält eine Zeile mit den 50 Einträgen der Urliste.

bedienungszeiten.json

Die Werte der Urliste sind in der Liste `Daten` enthalten.

Bruttosozialprodukt **Bruttosozialprodukt in den Jahren** 1950 **bis** 1965

Beschreibung

Der Datensatz beschreibt die Entwicklung des realen Bruttosozialprodukts der Bundesrepublik Deutschland im Zeitraum 1950 bis 1965 in Preisen von 1980 in Milliarden DM. Quelle ist [Ass96].

Format

Der Datensatz enthält Jahre und Bruttosozialprodukt in Milliarden DM wie in Tabelle 2.1.

Tabelle 2.1: Die Werte für den Datensatz Bruttosozialprodukt

Jahr	Bruttosozialprodukt in Milliarden DM
1950	338,800
1951	370,986
1952	404,004
1953	437,132
1954	469,480
1955	525,817
1956	564,202
1957	596,362
1958	618,427
1959	663,572
1960	724,621
1961	756,504
1962	792,060
1963	814,237
1964	867,977
1965	914,848

Dateien

bsp.csv

Die Datei enthält eine Zeile mit den Überschriften, gefolgt von den Werten. In der ersten Spalte ist die Jahreszahl als ganze Zahl, in der zweiten Spalte das Bruttosozialprodukt in Milliarden DM als Dezimalzahl angegeben.

bsp.json

Die Jahre sind in der Liste `Kategorie`, das Bruttosozialprodukt in der Liste `Daten` zu finden. Die Reihenfolge der Werte entspricht dabei Tabelle 2.1.

Buffalo **Schneehöhen aus 63 Wintern in Buffalo/N.Y.**

Beschreibung

Der Datensatz enthält die in Inches gemessenen Schneefall-Höhen aus 63 aufeinander folgenden Wintern von 1910/11 bis 1972/73 aus Buffalo im Bundesstaat New York. Gegeben ist die folgende Rangliste:

25,0	39,8	39,9	40,1	46,7	49,1	49,6	51,1	51,6
53,3	54,7	55,5	55,9	58,0	60,3	63,6	65,4	66,1
69,3	70,9	71,4	71,5	71,8	72,9	74,4	76,2	77,8
78,1	78,4	79,0	79,3	79,6	80,7	82,4	82,4	83,0
83,6	83,6	84,8	85,5	87,4	88,7	89,6	89,8	89,9
90,9	97,0	98,3	101, 4	102,4	103,9	104,5	105,2	110,0
110,5	110,5	113,7	114,5	115,6	120,5	120,7	124,7	126,4

Quelle ist [Koc12].

Format

63 Dezimalzahlen mit den Schneehöhen in Inch.

Dateien

buffalo.csv

Die Datei enthält eine Zeile mit den 63 Einträgen der Urliste.

buffalo.json

Die Werte der Urliste sind in der Liste `Daten` enthalten.

Dieselpreise **Dieselpreise in einer Kleinstadt**

Beschreibung

Die Verbraucher-Beratungsstelle einer Kleinstadt stellt an einem Stichtag bei 20 Tankstellen die Preise in Cent für Diesel fest. Es ergibt sich die folgende Rangliste:

91,4 91,4 91,9 91,9 91,9 91,9 91,9 92,9 93,9 93,9

95,9 95,9 95,9 96,9 97,9 97,9 97,9 98,9 98,9 98,9

Quelle ist [BB93].

Format

20 Dezimalzahlen mit den Preisen in Cent.

Dateien

dieselpreise.csv

Die Datei enthält eine Spalte mit den 20 Einträgen der Urliste.

dieselpreise.json

Die Werte der Urliste sind in der Liste `Daten` enthalten.

Druckfestigkeit **Druckfestigkeit von Betonwürfeln**

Beschreibung

Bei der Untersuchung der Druckfestigkeit in $0{,}1 N/mm^2$ an 30 Betonwürfeln wurden die folgenden Ergebnisse ermittelt:

$$374, 358, 341, 355, 342, 334, 353, 346, 355, 344, 349, 330, 352, 328, 336,$$

$$359, 361, 345, 324, 386, 335, 371, 358, 328, 353, 352, 366, 354, 378, 324.$$

Quelle ist [Rie78].

Format

30 ganze Zahlen mit den Druckfestigkeiten der Betonwürfel.

Dateien

druckfestigkeit.csv

Die Datei enthält eine Zeile mit den 30 Einträgen der Urliste.

druckfestigkeit.json

Die Werte der Urliste sind in der Liste `Daten` enthalten.

F & E Investitionen in Forschung und Entwicklung

Beschreibung

In einer Stichprobe wurden in 50 Software-Firmen die Anteile des Jahresumsatzes erhoben, die in Forschung und Entwicklung investiert werden:

13,5	9,5	8,2	6,5	8, 4	8,1	6,9	7,5	10,5	13,5
7,2	7,1	9,0	9,9	8,2	13,2	9,2	6,9	9,6	7,7
9,7	7,5	7,2	5,9	6,6	11,1	8,8	5,2	10,6	8,2
11,3	5,6	10,1	8,0	8,5	11,7	7,1	7,7	9,4	6,0
8,0	7,4	10,5	7,8	7,9	6,5	6,9	6,5	6,8	9,5

Quelle ist [MB91].

Format

50 Dezimalzahlen mit Prozentangaben.

Dateien

R_D.csv

Die Datei enthält eine Spalte mit den 50 Einträgen der Urliste.

R_D.json

Die Werte der Urliste sind in der Liste `Daten` enthalten.

Familienstand **Familienstand von 97 Personen**

Beschreibung

In einer Studie wird der Familienstand abgefragt hat. Quelle ist [Ben13].

Format

Der Datensatz enthält die Kategorie Familienstand mit den Werten „ledig“, „verheiratet“, „verwitwet“ und „geschieden“ und die jeweiligen absoluten Häufigkeiten wie in Tabelle 2.2.

Tabelle 2.2: Die Häufigkeiten für den Datensatz Familienstand

Familienstand	Absolut
ledig	28
verheiratet	43
verwitwet	11
geschieden	15

Dateien

familienstand.csv

Die Datei enthält eine Zeile mit den Überschriften wie in Tabelle 2.2, gefolgt von den Werten. In der ersten Spalte ist das Merkmal „Familienstand“ enthalten, in der zweiten Spalte die zugehörigen absoluten Häufigkeiten.

familienstand.json

Die Ausprägungen des Merkmals „Familienstand“ sind in der Liste `Kategorie`, die absoluten Häufigkeiten in der Liste `Daten` zu finden. Die Reihenfolge der Werte entspricht dabei Tabelle 2.2.

Jahreseinkommen **Jahreseinkommen für 200 Mitarbeiter in €**

Beschreibung

Die Personalabteilung der *Statistik KG* liefert für 200 Mitarbeiter die Angaben über das Einkommen eines bestimmen Jahres. Quelle ist [Sch94].

Format

Der Datensatz enthält Einkommensklassen und die jeweiligen absoluten Klassenhäufigkeiten wie in Tabelle 2.3.

Tabelle 2.3: Das Ergebnis der Erhebung für den Datensatz Jahreseinkommen

Jahreseinkommen in €	Absolute Häufigkeiten
(10 000, 20 000]	5
(20 000, 30 000]	15
(30 000, 40 000]	54
(40 000, 50 000]	46
(50 000, 60 000]	38
(60 000, 70 000]	18
(70 000, 80 000]	12
(80 000, 90 000]	8
(90 000, 100 000]	4

Dateien

jahreseinkommen.csv

Die Datei enthält eine Zeile mit den Überschriften wie in Tabelle 2.3, gefolgt von den Werten. In der ersten Spalte sind die Einkommensklassen enthalten, in der zweiten Spalte die zugehörigen absoluten Häufigkeiten.

jahreseinkommen.json

Die Klassen sind in der Liste `Kategorie`, die absoluten Häufigkeiten in der Liste `Daten` zu finden. Die Reihenfolge der Werte entspricht dabei Tabelle 2.3.

Körpergröße **Körpergrößen von Frauen**

Beschreibung

Bei einer Erhebung der Körpergröße in cm in einer Stichprobe von Frauen ist die folgende Urliste entstanden:

$$167, 170, 178, 154, 176, 162, 182, 166, 153, 165, 161, 175,$$

$$159, 168, 159, 158, 179, 174, 181, 174, 163, 160, 161, 178.$$

Quelle ist [Ben13].

Format

24 ganze Zahlen mit den Körpergrößen in cm.

Dateien

koerpergroesseFrauen.csv

Die Datei enthält eine Zeile mit den 24 Einträgen der Urliste.

koerpergroesseFrauen.json

Die Werte der Urliste sind in der Liste `Daten` enthalten.

Lebensdauer **Lebensdauer eines Ersatzteils in Stunden**

Beschreibung

Ein Unternehmen protokolliert die Lebensdauer eines Ersatzteils. Dabei entsteht die folgende Stichprobe, die die Lebensdauer der Teile in Stunden enthält:

$$110, 520, 490, 30, 120, 290, 370, 305, 415, 170, 280, 70, 540, 460, 260,$$

$$345, 150, 220, 435, 425, 470, 350, 130, 380, 230, 320, 360, 240, 330, 580.$$

Quelle ist [BB93].

Format

30 ganze Zahlen mit den Zeitangaben in Stunden.

Dateien

lebensdauer.csv

Die Datei enthält eine Zeile mit den 30 Einträgen der Urliste.

lebensdauer.json

Die Werte der Urliste sind in der Liste `Daten` enthalten.

Lieferzeiten **Lieferzeiten in Tagen**

Beschreibung

Die letzten 50 Lieferungen eines Lieferanten in Tagen liefert die folgende Urliste:

$$4, 5, 4, 1, 5, 4, 3, 4, 5, 6, 6, 5, 5, 4, 7, 4, 6, 5, 6, 4, 5, 4, 7, 5, 5,$$

$$6, 7, 3, 7, 6, 6, 7, 4, 5, 4, 7, 7, 5, 5, 5, 5, 6, 6, 4, 5, 2, 5, 4, 7, 5$$

Quelle ist [BB93].

Format

50 ganze Zahlen mit den Zeitangaben in Tagen.

Dateien

lieferzeiten.csv

Die Datei enthält eine Zeile mit den 50 Einträgen der Urliste.

lieferzeiten.json

Die Werte der Urliste sind in der Liste `Daten` enthalten.

Niederschlag **Monatliche Niederschläge in Dakar und Wien**

Beschreibung

Der Datensatz enthält die durchschnittlichen monatlichen Niederschläge in Millimetern für Dakar und Wien. Quelle ist [Ben13].

Format

Der Datensatz enthält Monatsangaben und Niederschlagsmengen in Millimetern wie in Tabelle 2.4.

Tabelle 2.4: Die durchschnittlichen monatlichen Niederschläge in mm im Datensatz Niederschlag

	Januar	Februar	März	April	Mai	Juni
Wien	37,5	36,8	45,0	51,7	68,0	70,8
Dakar	1	1,5	0,1	0,1	0,8	13,4
	Juli	August	September	Oktober	November	Dezember
Wien	75,6	66,6	48,7	49,4	48,4	46,2
Dakar	75,2	215,3	145,2	42,3	2,5	4

niederschlag.csv

Die Datei enthält eine Zeile mit den Überschriften wie in Tabelle 2.4. Die erste Spalte wird verwendet, um den Ort anzugeben. In der zweiten und dritten Zeile befinden sich die Niederschlagsmengen als Dezimalzahlen.

niederschlag_dakar.json

Die Monate sind in der Liste `Kategorie`, die durchschnittlichen monatlichen Niederschlagsmengen für Dakar in der Liste `Daten` zu finden.

niederschlag_wien.json

Die Monate sind in der Liste `Kategorie`, die durchschnittlichen monatlichen Niederschlagsmengen für Dakar in der Liste `Daten` zu finden.

Old Faithful **Eruptionen des Geysirs Old Faithful**

Beschreibung

Der Datensatz enthält die Dauer von 107 aufeinander folgenden Eruptionen des Geysirs Old Faithful in Minuten:

4,37	3,87	: 4	4,03	3,5	4,08	2,25	4,7	1,73	4,93
1,73	4,62	3,43	4,25	1,68	3,92	3,68	3,1	4,03	1,77
4,08	1,75	3,2	1,85	4,62	1,97	4,5	3,92	4,35	2,33
3,83	1,88	4,6	1,8	4,73	1,77	4,57	1,85	3,52	4
3,7	3,72	4,25	3,58	3,8	3,77	3,75	2,5	4,5	4,1
3,7	3,8	3,43	4	2,27	4,4	4,05	4,25	3,33	2
4,33	2,93	4,58	1,9	3,58	3,73	3,73	1,82	4,63	3,5
4	3,67	1,67	4,6	1,67	4	1,8	4,42	1,9	4,63
2,93	3,5	1,97	4,28	1,83	4,13	1,83	4,65	4,2	3,93
4,33	1,83	4,53	2,03	4,18	4,43	4,07	4,13	3,95	4,1
2,27	4,58	1,9	4,5	1,95	4,83	4,12			

Quelle ist [Tri14].

> **Tipp**:
> Dieser Datensatz steht in `R` als Default mit der Bezeichnung `faithful` zur Verfügung und enthält dort neben den Dauern von Eruptionen auch die Zeitspanne zwischen den Ausbrüchen!

Format

107 Dezimalzahlen mit den Zeitangaben in Minuten.

Dateien

oldFaithful.csv

Die Datei enthält eine Spalte mit den 107 Einträgen der Urliste.

oldFaithful.json

Die Werte der Urliste sind in der Liste `Daten` enthalten.

Rhythmik **Untersuchungen zum Rhythmus-Gefühl**

Beschreibung

Bei Rhythmik-Untersuchungen sollen drei Testpersonen einen vorgegebenen Takt halten. Die Messreihen ergeben die Differenzen zwischen dem gegebenen Takt und dem jeweils geschlagenen Takt in Millisekunden:

x	-30	-29	-28	-31	-30	-31	-27	-29	-30	-32
y	-20	25	15	-18	0	-12	8	-14	-3	19
z	2	-1	0	0	-1	0	-1	1	15	22

Quelle ist [GT96].

Format

Für jede der drei Testpersonen gibt es 10 Abweichung vom Takt, gemessen in Millisekunden. Die Daten sind als ganze Zahlen angegeben.

Dateien

rhythmik.csv

Die Datei enthält drei Zeilen mit den 10 Einträgen jedes Teilnehmers. In der ersten Spalte ist die Person angegeben; analog zur Urliste.

rhythmik_x.csv

Die Abweichungen der Testperson x ist in der Liste `Daten` enthalten.

rhythmik_y.csv

Die Abweichungen der Testperson y ist in der Liste `Daten` enthalten.

rhythmik_z.csv

Die Abweichungen der Testperson z ist in der Liste `Daten` enthalten.

Wellen **Durchmesser von Wellen**

Beschreibung

Bei der Kontrolle der Durchmesser von Wellen wird die Längein Millimeter gemessen. Quelle ist [Bau92].

Format

Die Längen sind klassifiziert. Der Datensatz enthält die Klassengrenzen in mm, die jeweiligen absoluten Klassen und die absoluten Summenhäufigkeiten wie in Tabelle 2.5.

Tabelle 2.5: Die absoluten Häufigkeiten für den Datensatz Wellen

Klassengrenzen in mm	Absolute Häufigkeiten
(125,145, 125,195]	2
(125,195, 125,245]	6
(125,245, 125,295]	18
(125,295, 125,345]	30
(125,345, 125,395]	38
(125,395, 125,445]	18
(125,445, 125,495]	9

Dateien

wellen.csv

Die Datei enthält eine Zeile mit den Überschriften wie in Tabelle 2.5, gefolgt von den Werten. In der ersten Spalte ist die linke, in der zweiten Spalte die rechte Klassengrenze enthalten. In der dritten Spalte sind die absoluten Häufigkeiten angegeben.

wellen.json

Die Klassen sind in der Liste `Kategorie`, die absoluten Häufigkeiten in der Liste `Daten` zu finden. Die Reihenfolge der Werte entspricht dabei Tabelle 2.5.

3 Zweidimensionale Daten

Die hier aufgeführten Datensätze enthalten zwei Merkmale und sind meist der Statistik-Literatur entnommen. Teilweise sind die Urlisten enthalten, teilweise enthalten die Datensätze bereits Kontingenztabellen. Die Aufzählung ist alphabetisch geordnet.

Altersverteilung **Verteilung des Alters in zwei Patientengruppen**

Beschreibung

In zwei Behandlungsgruppen wird das Alter der Patienten untersucht. Dabei besteht der Verdacht, dass sich in der Gruppe 2 vorwiegend ältere Patienten befinden. Quelle ist [JR16].

Format

Die Werte, gruppiert nach den beiden Patientenmengen, finden wir in Tabelle 3.1.

Tabelle 3.1: Die Urliste für den Datensatz „Altersverteilung“

Gruppe	Alter	Gruppe	Alter
1	40	2	63
1	42	2	64
1	43	2	66
1	50	2	66
1	52	2	67
1	54	2	67
1	55	2	67
1	55	2	70
1	56	2	70
1	60	2	76
		2	76
		2	81

Dateien

altersverteilung.csv

Die Datei enthält die Daten aus Tabelle 3.1 gruppiert nach Gruppe, so dass wir direkt ein Tibble einlesen können mit den beiden Merkmalen `Gruppe` und `Alter`.

Angebotsbearbeitung **Bearbeitungszeiten und Aufträge**

Beschreibung

Ein Hardware-Vertrieb die Bearbeitungszeit für die Erstellung des Angebots festgehalten. Zu jedem Angebot wurde auch festgehalten, ob auf Grund des Angebots ein Auftrag erteilt wurde. Quelle ist [MB91].

Format

Pro Anfrage wurde die Bearbeitungszeit in Tagen und eine Angabe, ob das Angebot zu einem Auftrag geführt hat, festgehalten. Dabei bedeutet `Y`, dass der Auftrag verloren wurde, `N` bedeutet, dass der Auftrag nicht verloren wurde, also ein Auftrag erteilt wurde.

Tabelle 3.2: Zeiten in Tagen und Angabe, ob ein Auftrag erteilt wurde im Datensatz Bearbeitungszeit

Bearbeitungszeit	Verloren?	Bearbeitungszeit	Verloren?
2,36	N	3,34	N
5,73	N	6	N
6,6	N	5,92	N
10,05	Y	7,28	Y
5,13	N	1,25	N
1,88	N	4,01	N
2,52	N	7,59	N
2	N	13,42	Y
4,69	N	3,24	N
1,91	N	3,37	N
6,75	Y	14,06	Y
3,92	N	5,1	N
3,46	N	6,44	N
2,64	N	7,76	N
3,63	N	4,4	N
3,44	N	5,48	N
9,49	Y	7,51	N
4,9	N	6,18	N
7,45	N	8,22	Y
20,23	Y	4,37	N
3,91	N	2,93	N
1,7	N	9,95	Y
16,29	Y	4,46	N
5,52	N	14,32	Y
1,44	N	9,01	N

Dateien

angebotsbearbeitung.csv

Die Datei enthält eine Zeile mit den Überschriften wie in Tabelle 3.2, gefolgt von den Werten. In der ersten Spalte ist die Bearbeitungszeit in Tagen als Dezimalzahl, in der zweiten Spalte ist festgehalten, ob der Auftrag verloren (`Y`) oder gewonnen (`N`) wurde.

angebotsbearbeitung.json

Die Bearbeitungszeit in Tagen ist in der Liste `Daten`, die Angabe, ob der Auftrag verloren oder gewonnen wurde in der Liste `Daten2` zu finden. Die Reihenfolge der Werte entspricht dabei Tabelle 3.2.

Arbeitslosenquote **Preisanstieg und Arbeitslosenquote**

Beschreibung

Um den Zusammenhang zwischen Preisanstieg und der Arbeitslosenquote in 10 Ländern zu untersuchen wurden die entsprechenden Werte aus dem OECD Main Economic Indicators Report, Dezember 1978, entnommen. Stand der Daten ist Juli 1978. Quelle ist [BB93].

Format

Die Daten sind in Tabelle 3.3 zusammengefasst.

Tabelle 3.3: Die Werte für den Datensatz„Arbeitslosenquote“

Land	Preisanstieg in %	Arbeitslosenquote in %
Belgien	4,1	10,1
BRD	2,3	4,0
Großbritannien	8,4	5,7
Irland	8,2	10,2
Italien	11,9	7,5
Japan	4,6	2,1
Kanada	9,4	8,0
Österreich	3,6	1,3
Schweden	10,6	2,2
U.S.A.	7,9	6,3

Dateien

oecd.csv

Die Datei enthält eine Matrix von Werten, wie in Tabelle 3.3 zu sehen. Die Werte für Preisanstieg und Arbeitslosenquote sind als Dezimalzahlen angegeben.

oecd_preisanstieg.json

Die Länder sind in der Liste `Daten`, der Preisanstieg in Prozent in der Liste `Daten2` zu finden. Die Reihenfolge der Werte entspricht dabei Tabelle 3.3.

oecd_arbeitslose.json

Die Länder sind in der Liste `Daten`, die Arbeitslosenquote in Prozent in der Liste `Daten2` zu finden. Die Reihenfolge der Werte entspricht dabei Tabelle 3.3.

Arztbesuch **Jährliches Einkommen und Häufigkeit von Arztbesuchen**

Beschreibung

In einer in den U.S.A. durchgeführten Studie sollte der Zusammenhang zwischen dem Einkommen und der Häufigkeit von Arztbesuchen bei Berufstätigen untersucht werden. Quelle ist [Hü03].

Format

Die Antworten (Y_i, Z_i) von $N = 2\,764$ Befragten wurden in Klassen aufgeteilt: beim jährlichen Einkommen Y_i in 5 Klassen und bei der Zeit Z_i seit dem letzten Arztbesuch in 3 Klassen. Das Ergebnis ist in Tabelle 3.4 zusammengestellt.

Tabelle 3.4: Die Urliste für den Datensatz „Arztbesuch“

	Zeit seit dem letzten Arztbesuch in Monaten			
Jährliches Einkommen in $	≤ 6	$7-12$	>12	Randhäufigkeit
unter 3 000	186	38	35	259
3 000 – 4 999	227	54	45	326
5 000 – 6 999	219	78	78	375
7 000 – 9 999	355	112	140	607
über 9 999	653	285	259	1 197
Randhäufigkeit	1 640	567	557	**2 764**

Dateien

arztbesuch.csv

Die Datei enthält die Kontingenztabelle wie in Tabelle 3.4.

arztbesuch.json

Die Werte für beide Merkmale der Kontingenztabelle sind in der Liste `Daten` abgelegt. Dabei wurde die Matrix zeilenweise abgespeichert. Die Ausprägungen des Merkmals „Jährliches Einkommen“ sind in der Liste `KategorienX` abgelegt, in der Reihenfolge wie in der ersten Spalte der Kontingenztabelle. Die Ausprägungen des Merkmals „Zeit seit dem letzten Arztbeusch“ befinden sich in der Liste `KategorienY`. Die Reihenfolge hält sich an die Angaben in der zweiten Zeile der Kontingenztabelle.

Autohersteller **Größe und Herstellerangaben für verkaufte Autos**

Beschreibung

Um die weitere Marktentwicklung besser abschätzen zu können erhebt ein Automobilhersteller eine zweidimensionale Stichprobe, in der verkaufte Autos nach Größe und Hersteller eingeordnet werden. Quelle ist [MB91].

Format

Die Kontingenztabelle der Ergebnisse ist in Tabelle 3.5 zusammengefasst.

Tabelle 3.5: Die Kontingenztabelle für den Datensatz Autohersteller

	Hersteller			
	A	B	C	D
Kompaktklasse	157	65	181	10
Mittelklasse	126	82	142	46
Oberklasse	58	45	60	28

Dateien

autohersteller.csv

Die Datei enthält die Kontingenztabelle wie in Tabelle 3.5.

autoherstellerLong.csv

Die Datei enthält die Daten im Long-Format.

autohersteller.json

Die Werte für beide Merkmale der Kontingenztabelle sind in Liste `Daten` abgelegt. Dabei wurde die Matrix zeilenweise abgespeichert. Die Ausprägungen des Merkmals „Größe des Autos“ sind in der Liste `KategorienX` abgelegt, in der Reihenfolge wie in der ersten Spalte der Kontingenztabelle. Die Ausprägungen des Merkmals „Hersteller“ befinden sich in der Liste `KategorienY`. Die Reihenfolge hält sich an die Angaben in der zweiten Zeile der Kontingenztabelle.

Bankumfrage **Umfrageergebnisse einer Bank**

Beschreibung

Eine Bank plant für ihre jugendlichen Kunden spezielle Angebote einzuführen. Um die wirtschaftlichen Interessen dieser Zielgruppe besser zu verstehen wird in einer Stichprobe von 100 Menschen, die nicht älter als 30 Jahre sind, eine Umfrage durchgeführt. Jede Person kann zwei Prioritäten aus einer Liste von sechs verfügbaren Aktionen auswählen. Quelle ist [MB91].

Format

Die Ergebnisse der Umfrage sind in Tabelle 3.6 zu finden.

Tabelle 3.6: Die Ergebnisse für den Datensatz Bankumfrage

Erste Priorität	Zweite Priorität	Absolute Häufigkeiten
Autokauf	Urlaubsreise	15
Autokauf	Sparen	14
Sparen	Autokauf	22
Sparen	Urlaubsreise	23
Urlaubsreise	Autokauf	10
Urlaubsreise	Sparen	16

Dateien

bankumfrage.csv

Die Datei enthält die eine Zeile mit den Überschriften wie in Tabelle 3.6, gefolgt von den Angaben, ebenfalls wie in Tabelle 3.6.

bankumfragen.json

Die absoluten Häufigkeiten sind in der Liste `Daten` abgelegt, in der Reihenfolge wie in Tabelle 3.6. Die Antworten für die erste Priorität sind in der Liste `KategorienX`, die für die zweite Priorität in der Liste `KategorienY` abgelegt, ebenfalls in der angegebenen Reihenfolge.

Beruf und Sport **Berufsgruppe und sportliche Betätigung**

Beschreibung

In einer Befragung wurden 1000 berufstätige Personen nach der Berufsgruppe und der sportlichen Betätigung gefragt. Quelle ist [BB93].

Format

Die Kontingenztabelle der Ergebnisse ist in Tabelle 3.7 zusammengefasst.

Tabelle 3.7: Die Ergebnisse für den Datensatz „Beruf und Sport“

	Sportliche Betätigung			
Berufsgruppe	Nie	Gelegentlich	Regelmäßig	Randhäufigkeit
Arbeiter	240	120	70	430
Angestellter	160	90	90	340
Beamter	30	30	30	90
Landwirt	37	7	6	50
Freiberuflich	40	32	18	90
Randhäufigkeit	507	279	214	**1 000**

Dateien

berufsport.csv

Die Datei enthält die Kontingenztabelle wie in Tabelle 3.7.

beruf_sport_long.csv

Die Datei enthält die Werte im Long-Format.

berufsport.json

Die Werte für beide Merkmale der Kontingenztabelle sind in der Liste `Daten` abgelegt. Dabei wurde die Matrix zeilenweise abgespeichert. Die Ausprägungen des Merkmals „Berufsgruppe“ sind in der Liste `KategorienX` abgelegt, in der Reihenfolge wie in der ersten Spalte der Kontingenztabelle. Die Ausprägungen des Merkmals „Sportliche Betätigung“ befinden sich in der Liste `KategorienY`. Die Reihenfolge hält sich an die Angaben in der zweiten Zeile der Kontingenztabelle.

Bonität **Bonität von Unternehmen**

Beschreibung

Zwei unabhängige Gutachter A und B beurteilen für eine Kreditversicherungsgesellschaft die Bonität von 7 Unternehmen. Quelle ist [Sch94].

Format

Die Unternehmen werden an Hand eines Punkte-Schemas von 1 (sehr schlechte Bonität) bis 10 (sehr gute Bonität) eingestuft. Tabelle 3.8 zeigt die Ergebnisse.

Tabelle 3.8: Die Einschätzungen für den Datensatz „Bonität“

Unternehmen	1	2	3	4	5	6	7
Gutachter A	2	3	3	6	7	8	9
Gutachter B	3	2	4	6	5	8	10

Dateien

bonitaet.csv

Die Datei enthält drei Spalten, für die Unternehmen und die Einschätzungen der beiden Gutachter. Die erste Zeile ist eine Überschrift.

bonitaet.json

Die Einschätzung des Gutachters A ist in der Liste `Daten` als ganze Zahlen enthalten. Analog enthält die Liste `Daten2` die Einschätzung des Gutachters B. Die Reihenfolge der Werte entspricht dabei Tabelle 3.8.

Einstiegsgehalt **Einstiegsgehälter von College-Absolventen**

Beschreibung

Der Datensatz enthält 10 verbundene Werte über Einstiegsgehälter von weiblichen und männlichen College-Absolventen in US-Dollar. Quelle ist [MB91], Seite 425.

Format

Die Werte der Urliste finden wir in Tabelle 3.9.

Tabelle 3.9: Einstiegsgehälter von weiblichen und männlichen College-Absolventen in US-Dollar

Frauen	Männer
23800	24300
26600	26500
24800	25400
23500	23500
27600	28500
23000	22800
24200	24500
25100	26200
23200	23400
23500	24200

Dateien

einstiegsgehalt.csv

Die Datei enthält die Daten aus Tabelle 3.9 gruppiert nach dem Geschlecht, so dass wir direkt ein Tibble einlesen können mit den beiden Merkmalen `Gehalt` und `Geschlecht`.

Familienzimmer **Familiengröße und Anzahl von Zimmern**

Beschreibung

Bei einer Untersuchung der Wohnsituation wurde erhoben, wie viele Mitglieder die Familie hat und wie viele Zimmer die Wohnung der Familie besitzt. Quelle ist [BB93].

Format

Beide Merkmale werden als ganze Zahl angegeben, wie in Tabelle 3.10.

Tabelle 3.10: Die Urliste für den Datensatz „Familienzimmer“

Größe der Familie	Anzahl der Zimmer	Größe der Familie	Anzahl der Zimmer
3	3	4	2
3	4	2	1
2	1	2	4
2	3	3	4
2	4	3	3
2	3	4	3
4	3	3	4
5	4	2	4
2	3	3	2
4	4	5	4

Dateien

wohnungen.csv

Die Datei enthält eine Zeile mit den Überschriften wie in Tabelle 3.2, gefolgt von den Werten. In der ersten Spalte ist die Größe der befragten Familie, in der zweiten Spalte ist die Anzahl der Zimmer abgelegt.

wohnungen.json

Die Größe der Familie ist in der Liste `Daten`, die Anzahl der Zimmer in der Liste `Daten2` zu finden. Die Reihenfolge der Werte entspricht dabei Tabelle 3.10.

FAZ Index **Renditen von Renten und Aktienkurse**

Beschreibung

An der Börse wird oft von einem Zusammenhang zwischen der Rendite von Renten und Aktienkursen gesprochen. Deshalb wurden zu 8 Zeitpunkten die Werte für den Aktienindex der Frankfurter Allgemeinen (FAZ-Index) und die durchschnittliche Rendite öffentlicher Anleihen mit 10 Jahren Laufzeit beobachtet. Quelle der Daten ist [BBK08].

Format

Der Zeitpunkt und der FAZ Index werden als ganze Zahl angegeben; die Rendite der Renten in Prozent als Dezimalzahl. Tabelle 3.11 enthält die Urliste.

Tabelle 3.11: Die Werte den Datensatzes „FAZ Index“

Zeitpunkt	1	2	3	4	5	6	7	8
FAZ-Index	221	251	346	376	401	412	471	481
Rendite in %	9,7	7,9	8,6	7, 2	7,3	7,1	7,0	6,8

Dateien

faz.csv

Die Datei enthält drei Zeilen, wie die Daten in Tabelle 3.11.

faz.json

Der FAZ-Index ist in der Liste `Daten` als ganze Zahlen, die Rendite der Renten in Prozent in der Liste `Daten2` als Dezimalzahl zu finden. Die Reihenfolge der Werte entspricht dabei Tabelle 3.11.

faz-index.json

Die FAZ-Indices als ganze Zahl sind in der Liste `Daten`, die Zeitpunkte in der Liste `Kategorie` enthalten. Die Reihenfolge der Werte entspricht dabei Tabelle 3.11.

faz-index-rendite.json

Die Renditen der Renten in Prozent als ganze Zahl sind in der Liste `Daten`, die Zeitpunkte in der Liste `Kategorie` enthalten. Die Reihenfolge der Werte entspricht dabei Tabelle 3.11.

Größe und Gewicht **Körpergrößen und Gewicht von Erwachsenen**

Beschreibung

In einer zufälligen Stichprobe wurden für 9 Erwachsene die Körpergrößen in cm und das Körpergewicht in kg erfasst. Quelle ist [MV05].

Format

Beide Merkmale werden als ganze Zahl angegeben. Die Größe wird in cm, das Gewicht in kg angeben. Tabelle 3.12 enthält die Urliste.

Tabelle 3.12: Die Urliste für den Datensatz „Größe und Gewicht“

Größe [cm]	Gewicht [kg]
165	65
182	80
177	77
175	78
169	70
190	88
185	83
173	72
180	80

groesse_gewicht.csv

Die Datei enthält eine Zeile mit den Überschriften wie in Tabelle 3.12, gefolgt von den Werten. In der ersten Spalte ist die Körpergröße in cm als ganze Zahl angegeben. In der zweiten Spalte ist das Gewicht in kg als ganze Zahl angegeben.

groesse_gewicht..json

Die Körpergröße in cm ist in der Liste `Daten`, das Körpergewicht in der Liste `Daten2` zu finden. Die Reihenfolge der Werte entspricht dabei Tabelle 3.12.

Hausmiete **Alter von Häusern und Miete pro Quadratmeter**

Beschreibung

In einer Kleinstadt wurde eine Totalerhebung des Alters von Häusern in Jahren und der Miete pro m^2 in € durchgeführt. Quelle ist [Sch94].

Format

Beide Merkmale wurden klassifiziert. Die Kontingenztabelle der Ergebnisse ist in Tabelle 3.13 zusammengefasst.

Tabelle 3.13: Die Kontingenztabelle für den Datensatz „Hausmiete“

	Miete					
Alter	$[2,4)$	$[4,6)$	$[6,8)$	$[8,10)$	$[10,12)$	Randhäufigkeit
$[0,6)$	0	8	8	4	4	24
$[6,12)$	0	8	20	32	16	76
$[12,18)$	4	16	80	40	8	148
$[18,24)$	4	16	32	28	8	88
$[24,30)$	12	12	20	16	4	64
$[30,*)$	12	20	12	0	4	48
Randhäufigkeit	32	80	172	120	44	**448**

Dateien

alter_und_miete.csv

Die Datei enthält die Kontingenztabelle wie in Tabelle 3.13.

alter_und_miete.json

Die Werte für beide Merkmale der Kontingenztabelle sind in Liste `Daten` abgelegt. Dabei wurde die Matrix zeilenweise abgespeichert. Die Ausprägungen des Merkmals „Alter des Hauses in Jahren“ sind in der Liste `KategorienX` abgelegt, in der Reihenfolge wie in der ersten Spalte der Kontingenztabelle. Die Ausprägungen des Merkmals „Miete pro m^2 in €“ befinden sich in der Liste `KategorienY`. Die Reihenfolge hält sich an die Angaben in der zweiten Zeile der Kontingenztabelle.

Kreis **Punkte auf einem Kreis**

Beschreibung

Der Datensatz enthält die x- und y-Koordinaten von Punkten im $\mathbb{R}^2$, die annähernd auf einem Kreis liegen. Quelle ist [Old11].

Format

Die Koordinaten sind in ganzen Zahlen angegeben.

Tabelle 3.14: Punkte, die annähernd auf einem Kreis liegen

x	y
3	1
2	2
1	3
0	3
-1	-2
0	-2

Dateien

kreis.csv

Die Datei enthält eine Zeile mit den Überschriften wie in Tabelle 3.14, gefolgt von den Werten. Als Trenner für die Spalten wird das Semikolon verwendet.

Luftfeuchtigkeit **Relative Luftfeuchtigkeit und Wassergehalt von Wolle**

Beschreibung

Um den Einfluss der relativen Luftfeuchtigkeit auf den Feuchtigkeitsgehalt von Wolle zu untersuchen wurden 20 Messungen durchgeführt. Quelle ist [Bau92].

Format

Die Messwerte, als Prozentwerte angegeben, sind in Tabelle 3.15 zusammengefasst.

Tabelle 3.15: Die Urliste für den Datensatz „Luftfeuchtigkeit“ (alle Angaben in Prozent)

relative Luftfeuchtigkeit	Wassergehalt	relative Luftfeuchtigkeit	Wassergehalt
10	5	70	17
90	26	20	9
20	8	40	13
40	12	40	11
50	14	50	12
70	18	70	19
80	21	80	22
90	25	90	24
10	4	50	13
10	6	50	17

Dateien

wassergehalt.csv

Die Datei enthält eine Zeile mit den Überschriften wie in Tabelle 3.15, gefolgt von den Werten. In der ersten Spalte ist die relative Luftfeuchtigkeit, in der zweiten Spalte der zugehörigen Wassergehalt angegeben. Alle Angaben sind als ganze Zahlen abgespeichert.

wassergehalt.json

Die relative Luftfeuchtigkeit in Prozent ist in der Liste `Daten`, der Wassergehalt in Prozent in der Liste `Daten2` zu finden. Die Reihenfolge der Werte entspricht dabei Tabelle 3.15.

Restaurants **Umsätze in zwei Restaurants**

Beschreibung

Wir untersuchen die täglichen Umsätze zweier Restaurants. Insgesamt haben wir die Umsätze an 12 Tagen von zwei aufeinanderfolgenden Wochen, ohne Sonntag. Die Angaben in den Daten sind in € angegeben. Quelle ist [MB91].

Format

Die Werte der Urliste finden wir in Tabelle 3.16.

Tabelle 3.16: Umsätze zweier Restaurants in €

Restaurant 1	Restaurant 2
759	678
981	933
1005	918
1449	1302
1905	1782
2073	1971
693	639
873	825
1074	999
1338	1281
1932	1827
2106	2049

Dateien

restaurants.csv

Die Datei enthält die Daten aus Tabelle 3.16 gruppiert nach dem Restaurant, so dass wir direkt ein Tibble einlesen können mit den beiden Merkmalen `Umsatz` und `Restaurant`.

Produktionsmengen **Produktionsmengen und Gesamtkosten**

Beschreibung

Die Abteilung eines Unternehmens ist ausschließlich mit der Herstellung eines einzigen Produkts beschäftigt. Um den Zusammenhang zwischen der Produktionsmenge in Stücken und den Gesamtkosten in € zu untersuchen, wurden die entsprechenden Werte in 10 Zeitperioden festgehalten. Quelle ist [BB93].

Format

Die Ergebnisse sind in Tabelle 3.17 zusammengefasst. Der Übersichtlichkeit halber sind die Spalten des Datensatzes als Zeilen dargestellt. Die Spalte „Output“ enthält die Produktionsmenge in Stücken. Die Spalte „Kosten“ gibt die gesamte Produktionskosten in € an.

Tabelle 3.17: Die Urliste für den Datensatz Produktionsmengen

Periode	1	2	3	4	5	6	7	8	9	10
Output	9	12	14	12	12	13	10	11	12	15
Kosten	1 216	1 300	1 356	1 288	1 276	1 292	1 260	1 244	1 288	1 360

Dateien

produktionsmengen.csv

Die Datei enthält die Werten, wie in Tabelle 3.17 zu sehen. Zusätzlich sind Spaltenüberschriften vorhanden. Alle Werte sind als ganze Zahlen angegeben.

produktionsmengen.json

Die Produktionsmengen in Stück sind in der Liste `Daten`, die Gesamtkosten in € in der Liste `Daten2` zu finden. Die Reihenfolge der Werte entspricht dabei Tabelle 3.17.

Solar Stromstärke und Einfallswinkel in einer Solaranlage

Beschreibung

Der Datensatz enthält Angaben zum maximalen Strom in einer Solaranlage in Abhängigkeit vom Einfallswinkel des Sonnenlichts. Quelle ist [Old11].

Format

Über die Einheit des Stroms werden in der Quelle keine Angaben gemacht. Die Winkel sind in Grad angegeben.

Tabelle 3.18: Einfallswinkel und maximale Stromstärke in einer Solaranlage

Einfallswinkel	Strom
0	400
10	370
20	360
45	300
60	200
80	80

Dateien

solar.csv

Die Datei enthält eine Zeile mit den Überschriften wie in Tabelle 3.18, gefolgt von den Werten. Als Trenner für die Spalten wird das Semikolon verwendet.

Stärke und Stabilität **Bauteil-Stärke und Stabilität**

Beschreibung

In einer Stichprobe wird die Stärke und die Stabilität bestimmter Bauteile untersucht. Quelle ist [MV05].

Format

Beide Merkmale werden als ganze Zahl angegeben. Tabelle 3.19 enthält die Urliste.

Tabelle 3.19: Die Urliste für den Datensatz „Stärke und Stabilität“

Stärke	11	20	15	17	21	12	19	16	20	14	11
Stabilität	5	6	12	13	4	10	11	13	9	11	8

Dateien

staerkestabilitaet.csv

Die Datei enthält eine Zeile mit den Überschriften, dabei sind die Umlaute ersetzt worden. Anschließend folgen die Daten. In der ersten Spalte ist die Stärke als ganze Zahl angegeben. In der zweiten Spalte ist die Stabilität als ganze Zahl angegeben.

staerkestabilitaet.json

Die Stärke des Bauteils ist in der Liste `Daten`, die Stabilität in der Liste `Daten2` zu finden. Die Reihenfolge der Werte entspricht dabei Tabelle 3.19.

Testarbeit **Organisatorische Geschicklichkeit und Sorgfalt**

Beschreibung

Für 10 Angestellte wurde mit einem Test sowohl ihre organisatorische Geschicklichkeit als auch ihre Sorgfalt ermittelt. Quelle ist ([BB93]).

Format

Dabei ergaben sich die Rangziffern wie in Tabelle 3.20.

Tabelle 3.20: Die Rangziffern im Datensatz Testarbeit

Angestellter	1	2	3	4	5	6	7	8	9	10
Rang für die Geschicklichkeit	7	3	9	10	1	5	4	6	2	8
Rang für die Sorgfalt	3	9	10	8	7	1	5	4	2	6

Dateien

geschicklichkeit.csv

Die Datei enthält eine Matrix von Rangziffern, wie in Tabelle 3.20 zu sehen.

geschicklichkeit.json

Die Rangliste für die Geschicklichkeit ist in der Liste `Daten`, die Rangliste für die Sorgfalt in der Liste `Daten2` zu finden. Die Reihenfolge der Werte entspricht dabei Tabelle 3.20.

Taschengeld **Taschengeld von Schülern**

Beschreibung

Der Datensatz enthält das Ergebnis einer Umfrage in drei Schulklassen nach dem Taschengeld in €. Quelle ist [Koc12].

Format

In der Schulklasse A_1 wurden 10 Schüler, in den Klassen A_2 und A_3 jeweils 5 Schüler befragt. Dabei ergeben sich die Werte aus Tabelle 3.21.

Tabelle 3.21: Die Ergebnisse der Umfrage für den Datensatz Taschengeld

Klasse	Taschengeld in €				
A_1	6	8	4	5	5
	6	7	5	6	2
A_2	5	2	5	6	6
A_3	3	2	5	6	6

Dateien

taschengeld.csv

Die Datei enthält zwei Spalten, und für jeden befragten Schüler eine Zeile. In der ersten Spalte ist die Klasse angegeben, aus der der Schüler stammt, in der zweiten Spalte das Taschengeld in €.

taschengeld_A1.json

Die Ergebnisse der Umfrage für die Schüler der Klasse A_1 sind in der Liste `Daten` zu finden.

taschengeld_A2.json

Die Ergebnisse der Umfrage für die Schüler der Klasse A_2 sind in der Liste `Daten` zu finden.

taschengeld_A3.json

Die Ergebnisse der Umfrage für die Schüler der Klasse A_3 sind in der Liste `Daten` zu finden.

4 Allgemeine Datensätze

Die hier aufgeführten Datensätze enthalten mehrere Merkmale. Die Daten stammen aus der Literatur oder dem World Wide Web. Die Aufzählung ist alphabetisch nach der Kurzbezeichnung geordnet.

Advertising **Marketing-Etats und ihre Wirkung**

Beschreibung

Ein Unternehmen hat 200 Werte über Marketing-Aktionen gespeichert. Quelle der Daten ist [JWHT13].

Format

Der Datensatz enthält Angaben über das Budgets für die drei Medien TV, Radio und Zeitung, und Angaben über die Verkaufszahlen eines Produkts. Die Budgets sind in Tausend Dollar, die Verkaufszahlen in Tausend verkauften Einheiten in 200 verschiedenen Märkten angegeben.

Dateien

Advertising.csv

Die Datei enthält 201 Zeilen, eine Überschrift und die 200 Beobachtungen in insgesamt fünf Spalten. Die erste Spalte, die keine Überschrift hat, ist eine Nummerierung. Die folgenden drei Spalten enthalten die Angaben über die Budgets in Tausend Dollar für „TV“, „Radio“ und „Newspaper“. Die letzte Spalte enthält Angaben über die Verkaufszahlen, in Tausend Einheiten.

Anscombe **Anscombe Quartett**

Beschreibung

Anscombe hat 1973 ([Ans73]) vier Urlisten aufgestellt, die im arithmetischen Mittel, der Varianz und auch der Korrelation mindestens bis auf die dritte Nachkomma-Stelle die gleichen Parameterwerte besitzen. Auch die lineare Regression ergibt die gleiche Regressionsgerade. Die Werte sind [Mun14] entnommen.

Format

Der Datensatz enthält vier zweidimensionale Urlisten mit X und Y-Werten. Die Werte sind in Tabelle 4.1 zu sehen.

Tabelle 4.1: Die Urlisten des Anscombe Quartetts

$X1$	$Y1$	$X2$	$Y2$	$X3$	$Y3$	$X4$	$Y4$
10	8,04	10	9,14	10	7,46	8	6,58
8	6,95	8	8,14	8	6,77	8	5,76
13	7,58	13	8,74	13	12,74	8	7,71
9	8,81	9	8,77	9	7,11	8	8,84
11	8,33	11	9,26	11	7,81	8	8,47
14	9,96	14	8,1	14	8,84	8	7,04
6	7,24	6	6,13	6	6,08	8	5,25
4	4,26	4	3,1	4	5,39	19	12,5
12	10,84	12	9,13	12	8,15	8	5,56
7	4,82	7	7,26	7	6,42	8	7,91
5	5,68	5	4,74	5	5,73	8	6,89

Dateien

Anscombe.csv

Der Datei enthält 8 Spalten, jeweils zwei davon sind die x- und y-Werte der vier einzelnen Urlisten. Im Header sind die vier Urlisten gekennzeichnet:

```
X1; Y1; X2; Y2; X3; Y3, X4; Y4
10,0;...
```

Jede der vier zweidimensionalen Urlisten enthält 11 Zeilen.

AutoEU **Angaben über PKW**

Beschreibung

Der Datensatz ist ursprünglich als Beispiel für [JWHT13] verfügbar und Teil des `R`-Packages `ISLR`. Da die Daten US-amerikanische Einheiten enthalten wurden die Angaben und Einheiten auf europäische und ISO-Einheiten umgerechnet.

Format

Der Datensatz enthält 9 Spalten mit den folgenden Merkmalen:

- Name des Autos,
- Region des Herstellers (U.S.A., Japan, Europa),
- Baujahr,
- Anzahl der Zylinder,
- Hubraum in Litern,
- Gewicht in Tonnen,
- Verbrauch in Litern pro 100 Kilometern,
- Anzahl der PS,
- Beschleunigung von 0 auf 100 km/h in Sekunden.

Insgesamt sind 392 Autos im Datensatz enthalten.

Dateien

AutoEU.csv

Der Datei enthält 9 Spalten und 392 Zeilen, mit Spaltenüberschriften

Golden Snowball Award **Schneehöhen in Inches**

Beschreibung

Der Datensatz enthält die in Inches gemessenen Schneefall-Höhen in fünf Städten des Bundesstaats New York in U.S.A. Quelle ist [NOA14, Wik14].

Format

Für die Städte Albany, Buffalo und Rochester wurden die Schneehöhen von der Saison 1940/41 bis 2012/2013 erfasst. Für Binghamton und Syracuse gibt es Daten für 1951/52 bis 2012/2013. Angegeben ist immer die Saison und die Schneehöhe in Inch. Hier ein Ausschnitt der Daten für Buffalo, von 1951/52 bis 1960/61:

Saison	Schneehöhe in Inch
$1951-52$	$100,5$
$1952-53$	$79,4$
$1953-54$	$91,3$
$1954-55$	$101,4$
$1955-56$	$146,8$
$1956-57$	$76,1$
$1957-58$	$143,8$
$1958-59$	$137,2$
$1959-60$	$134,8$
$1960-61$	$130,5$

Dateien

goldensnowball_albany.csv

Die Datei enthält als erste Zeile die Überschriften, gefolgt von 73 Zeilen mit den Angaben zu Saison und Schneehöhen.

goldensnowball_buffalo.csv

Die Datei enthält als erste Zeile die Überschriften, gefolgt von 73 Zeilen mit den Angaben zu Saison und Schneehöhen.

goldensnowball_rochester.csv

Die Datei enthält als erste Zeile die Überschriften, gefolgt von 73 Zeilen mit den Angaben zu Saison und Schneehöhen.

goldensnowball_binghamton.csv

Die Datei enthält als erste Zeile die Überschriften, gefolgt von 62 Zeilen mit den Angaben zu Saison und Schneehöhen.

goldensnowball_syracuse.csv

Die Datei enthält als erste Zeile die Überschriften, gefolgt von 62 Zeilen mit den Angaben zu Saison und Schneehöhen.

Jazz-Standards **Jazz-Standards, Komponisten und Aufnahmen**

Beschreibung

Der Datensatz enthält Angaben über Jazz-Standards, Komponisten, Genre und Aufnahmen von Jazz-Standards. Quelle ist [Sch09].

Format

Der Autor gibt 30 Aufnahmen von Jazz-Standards an. Hier eine Zeile des Datensatzes:

Id	Titel	Komponisten	Musiker	Instrumente	Jahr	Art	Ort	Länge
22	Caravan	3	6	12	1990	Studio	Köln	332

Dateien

Jazzstandards.csv

Die Datei enthält als erste Zeile die Überschriften, gefolgt von 30 Zeilen mit den Angaben zu Id, Titel, Anzahl Komponisten, Anzahl Musiker in der Aufnahme, Anzahl Instrumente in der Aufnahme, Jahr der Aufnahme, Art der Aufnahme, Ort der Aufnahme und Länge der Aufnahme in Sekunden.

Minard **Daten zu Napoleons Russland-Feldzug**

Beschreibung

Minard's grafische Darstellung des Schicksals der Truppen von Napoleons im Russland-Feldzug 1813 wurde von Tufte ([Tuf83]) als „greatest statistical graphic ever drawn" bezeichnet. Friendly ([Fri02]) beschreibt, wie man diese Grafik mit modernen Werkzeugen wieder erstellen kann. Abbildung 4.1 zeigt eine Reproduktion dieser berühmten Grafik.

Quelle der Daten ist [Pro14].

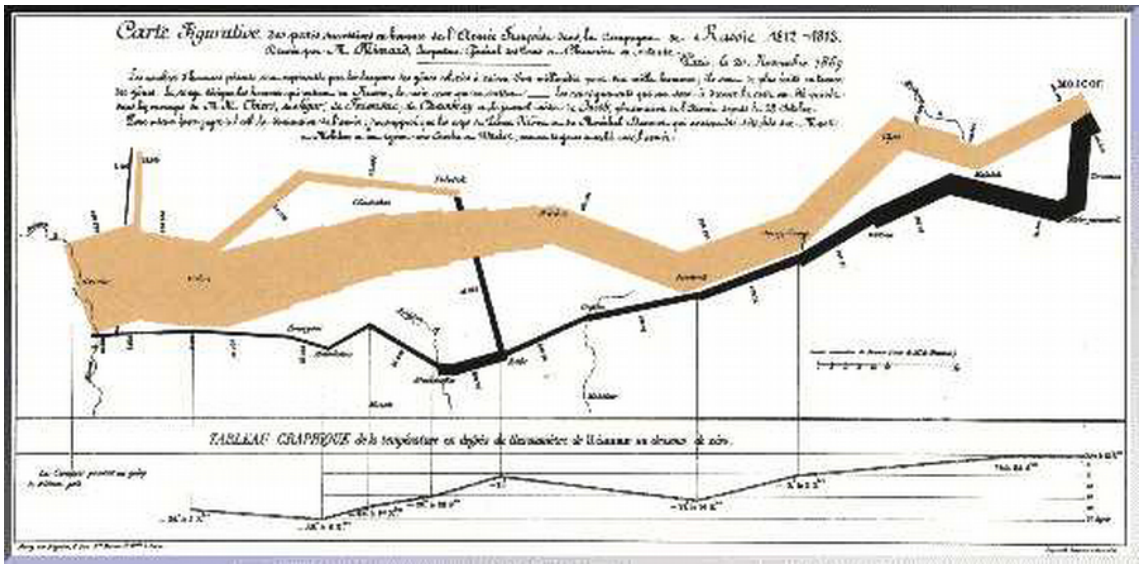


Abbildung 4.1: Minard's Darstellung des Russland-Feldzugs

Format

Der Datensatz besteht aus drei Dateien. In `Minard.troops.csv` enthält 51 Beobachtungen für die folgenden 5 Merkmale:

long Geographische Länge (longitude)

lat Geographisce Breite (latitude)

survivors Größe der Truppe als natürliche Zahl

direction Angabe, ob die Truppe auf dem Vormarsch („Advance") oder auf dem Rückzug („Retreat") war

group Angabe über den Teil der Truppe, auf den sich die Daten bezieht, als natürliche Zahl.

Die Datei `Minard.cities.csv` enthält 20 Angaben über Orte, die während des Russland-Feldzugs von Bedeutung waren:

long Geographische Länge (longitude)

lat Geographisce Breite (latitude)

city Ortsname

Die Datei `Minard.temp.csv` enthält 9 Angaben über die Temperaturen auf dem Rückzug von Moskau nach Westen:

long Geographische Länge (longitude)

lat Geographisce Breite (latitude)

days Anzahl der Tage auf dem Rückzug

date Datum, in der Form „Nov28“

Dateien

Minard.cities.csv

Die Datei enthält 21 Zeilen, eine Überschrift und die 20 Beobachtungen. In den Spalten sind die 3 Variablen abgelegt, die Spalte 1 enthält eine laufende Nummer.

Minard.temp.csv

Die Datei enthält 10 Zeilen, eine Überschrift und die 9 Beobachtungen. In den Spalten sind die 10 Variablen abgelegt, die Spalte 1 enthält eine laufende Nummer.

Minard.troops.csv

Die Datei enthält 52 Zeilen, eine Überschrift und die 51 Beobachtungen. In den Spalten sind die 5 Variablen abgelegt, die Spalte 1 enthält eine laufende Nummer.

Nightingale	Todesursache von Soldaten im Krimkrieg 1854-1856

Beschreibung

In der Geschichte der Visualisierung von Daten ist Florence Nightingale ist bekannt als Krankenschwester im Krimkrieg im 19. Jahrhundert, aber auch für die Sammlung von Daten und die Visualisierung dieser Daten. Abbildung 4.2 zeigt das von ihr erstellte Polardiagramm, auch bekannt als Nightingale Rose.

Florence Nightingale musste als Krankenschwester im Krimkrieg unsagbar schlechte sanitäre Bedingungen beobachten. Darauf hin veröffentliche Sie einige sehr einflussreiche Bücher wie [Nig58], aus dem auch die Abbildung entnommen ist. Sie wies darauf hin, dass viele der Opfer im Krimkrieg unter den britischen Soldaten durch unzulängliche hygienische Zustände und vermeidbare Ursachen, und nicht durch Verwundungen hervorgerufen wurden.

Quelle der Daten ist [PS07, Pro14].

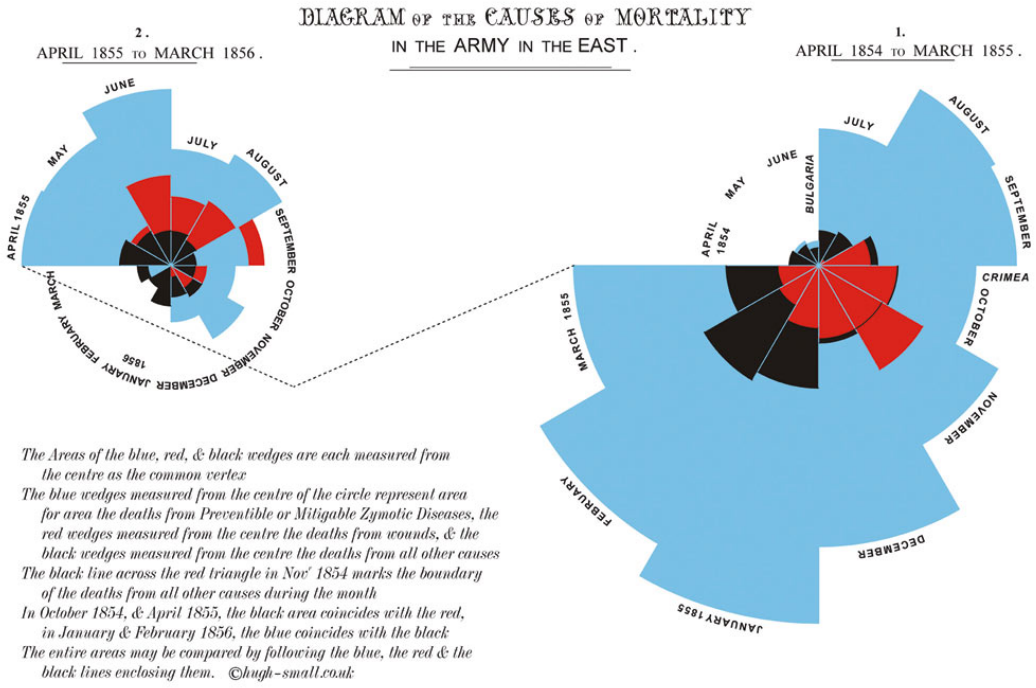


Abbildung 4.2: Das von Florence Nightingale erstellte Polardiagramm - die Nightingale Rose

Format

Der Datensatz enthält 24 Beobachtungen mit jeweils 10 Variablen:

Date Datum der Beobachtung, in der Form Jahr-Monat-Tag; Beispiel: 1854-04-01.

Month Der Monat des Krim-Kriegs

Jahr Das Jahr des Krim-Kriegs

Army Geschätzte monatliche Stärke der britischen Armee im Krimkrieg

Disease Anzahl der Todesfälle auf Grund vermeidbarer Infektionskrankheiten

Wounds Anzahl der Todesfälle auf Grund von Verwundungen

Other Andere Todesursachen

Disease.rate Jährliche Rate auf Grund vermeidbarer Infektionskrankheiten, pro $1\,000$

Wounds.rate Jährliche Rate auf Grund von Verwundungen, pro $1\,000$

Other.rate Jährliche Rate auf Grund anderer Ursachen, pro $1\,000$

Für eine Todesursache D ist die jährliche Rate berechnet als

$$\frac{12 \cdot 1\,000 \cdot D}{\text{Army}},$$

gerundet auf eine Dezimalstelle.

Die grafischen Darstellungen in Abbildung 4.2 beziehen sich auf die Daten vor und nach März 1855.

Dateien

nightingale.csv

Die Datei enthält 25 Zeilen, eine Überschrift und die 24 Beobachtungen. In den Spalten sind die 10 Variablen abgelegt.

Tips Trinkgelder in einem amerikanischen Restaurant

Beschreibung

In U.S.A. ist die Höhe des Trinkgelds sehr wichtig für das Bedienungspersonal. In einem amerikanischen Restaurant hat eine Bedienung Daten über alle Kunden festgehalten. Die Daten wurden in einem Zeitraum von 2, Monaten erhoben, im Jahr 1990. Das Restaurant befand sich in einer Shopping Mall. Die Gesetzgebung des Bundesstaates, in dem sich das Restaurant befand verlangte, dass Gäste auf Wunsch einen Tisch in einem Nichtraucher-Teil des Restaurants erhalten konnten.

Quelle des Datensatzes ist [BS95]. Der Datensatz wird in vielen Veröffentlichungen diskutiert, insbesondere in [CS07] und [TU09].

Format

Tabelle 4.2: Die Merkmale des Datensatzes Tips

Merkmal	Erläuterung
`obs`	Zeilennummer
`totbill`	Höhe der Rechnung, incl. Steuern, in US Dollar
`tip`	Trinkgeld, in US Dollar
`sex`	Geschlecht der Person, die die Rechnung bezahlt hat
`smoker`	Gab es Raucher am Tisch?
`day`	Wochentag
`time`	Nachmittags oder Abends?
`size`	Anzahl der der Gäste am Tisch

Dateien

tips.csv

Die Datei enthält als erste Zeile die Überschriften, gefolgt von 244 Zeilen mit den Angaben wie in Tabelle 4.2. Als Feldtrenner wir das Komma und als Dezimaltrenner der Punkt verwendet!

Titanic **Statistik der überlebenden Passagiere der Titanic**

Beschreibung

Der Datensatz enthält Angaben zu den überlebenden Passagieren der Titanic, aufgeschlüsselt nach Alter, Geschlecht, Status und weiteren Merkmalen. Quelle ist [Sta14, Cor17].

Format

Tabelle 4.3: Die Angaben in der Quelle für den Datensatz Titanic

Class	Age Group	Sex	Number Survived	Number Aboard	Percent Survived
1st	child	male	5	5	100,0
2nd	child	male	11	11	100,0
3rd	child	male	13	48	27,1
1st	adult	male	57	175	32,6
2nd	adult	male	14	168	8,3
3rd	adult	male	75	462	16,2
Crew	adult	male	192	862	22,3
1st	child	female	1	1	100,0
2nd	child	female	13	13	100,0
3rd	child	female	14	31	45,2
1st	adult	female	140	144	97,2
2nd	adult	female	80	93	86,0
3rd	adult	female	76	165	46,1
Crew	adult	female	20	23	87,0

Format

Dateien

titanic.csv

Die Datei enthält als erste Zeile die Überschriften, gefolgt von 14 Zeilen mit den Angaben wie in Tabelle 4.3.

Literaturverzeichnis

[Ans73] ANSCOMBE, F. J.: *Graphs in Statistical Analysis*. American Statistician, 27:17–21, 1973.

[Ass96] ASSENMACHER, WALTER: *Deskriptive Statistik*. Springer, 1996.

[Bau92] BAUMANN, HORST: *Lehr- und Übungsbuch Mathematik IV*. Fachbuchverlag Leizig-Köln, 1992.

[BB93] BAMBERG, GÜNTER und BAUR, FRANZ: *Statistik*. Oldenbourg, 8. Auflage, 1993.

[BBK08] BAMBERG, GÜNTER, BAUR, FRANZ und KRAPP, MICHAEL: *Statistik-Arbeitsbuch – Übungsaufgaben, Fallstudien, Lösungen*. Oldenbourg, 2008.

[Ben13] BENESCH, THOMAS: *Schlüsselkonzepte zur Statistik*. Springer Spektrum, 2013.

[BS95] BRYANT, PETER und SMITH, MARLENE: *Practical Data Analysis: Case Studies in Business Statistics*. Irwin Publishing, 1995.

[Cor17] CORETEAM: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[CS07] COOK, DIANNE und SWAYNE, DEBORAH: *Interactive and Dynamic Graphics for Data Analysis With R and GGobi*. Springer, 2007.

[Fri02] FRIENDLY, MICHAEL: *Re-Visions of Minard*. Journal of Educational and Behavioral Statistics, 27(1):31–51, 2002.

[GT96] GREINER, MICHAEL und TINHOFER, GOTTFRIED: *Stochastik fr Studienanfänger der Informatik*. Hanser, 1996.

[Hü03] HÜBNER, GERHARD: *Stochastik – Eine anwendungsorientierte Einführung fÃ¼r Informatiker, Ingenieure und Mathematiker*. Mathematische Grundlagen der Informatik. Vieweg, 4. Auflage, 2003.

[JR16] JÄGER, BERND PAUL und RUDOLPH, PAUL EBERHARD: *Statistik - Verstehen durch Experimente mit SAS*. De Gruyter Studium. De Gruyter, 2016.

[JWHT13] JAMES, GARETH, WITTEN, DANIELA, HASTIE, TREVOR und TIBSHIRANI, ROBERT: *An Introduction to Statistical Learning with Applications in R*. Springer Texts in Statistics. Springer, 2013.

[Koc12] KOCKELKORN, ULRICH: *Statistik für Anwender*. Springer Spektrum, 2012.

[MB91] MCCLAVE, JAMES und BENSON, GEORGE: *Statistics for Business and Economics*. Prentice Hall, 1991.

[Mun14] MUNZNER, TAMARA: *Visualization Analysis and Design*. A K Peters Visualization Series. CRC Press, 2014.

[MV05] MONKA, MICHAEL und VOSS, WERNER: *Statistik am PC – Lösungen mit Excel 97, 2000, 2002 und 2003*. Hanser, 4. Auflage, 2005.

[Nig58] NIGHTINGALE, FLORENCE: *Notes on Matters Affecting the Health, Efficiency, and Hospital Adminstration of the British Army*. Harrison and Sons, 1858.

[NOA14] NOAA: *Unique Local Climate Data*. `http://www.nws.noaa.gov/climate/local_data.php?wfo=buf`, 2014. Zuletzt gesehen 28. 3. 2014.

[Old11] OLDENBURG, REINHARD: *Mathematische Algorithmen im Unterricht - Mathematik aktiv erleben durch Programmieren*. Vieweg+Teubner, 2011.

[Pro14] PROJEKT R: *HistData: Data sets from the history of statistics and data visualization*. `http://cran.r-project.org/web/packages/HistData/`, 2014. Zuletzt gesehen: 17. 4. 2014.

[PS07] PEARSON, M. und I. SHORT: *Understanding Uncertainty: Mathematics of the Coxcomb*. `http://understandinguncertainty.org/node/214`, 2007. Zuletzt gesehen: 17. 4. 2014.

[Rie78] RIEDWYL, HANS: *Angewandte mathematische Statistik in Wissenschaft, Administration und Technik*. Verlag Paul Haupt, 1978.

[Sch94] SCHULZE, PETER: *Beschreibende Statistik*. Oldenbourg, 2. Auflage, 1994.

[Sch09] SCHUBERT, MATTHIAS: *Mathematik für Informatiker*. Vieweg+Teubner, 2009.

[Sta14] STATISTICAL CONSULTANTS LTD: *Titanic Survival Data Summarised by Class, Age Group and Sex*. `http://www.statisticalconsultants.co.nz/blog/B47.html`, 2014. Zuletzt gesehen: 4. 4. 2014.

[Tri14] TRIOLA, MARIO: *Internet Project Old Faithful Data*. `http://wps.aw.com/wps/media/objects/15/15719/projects/ch3_faithful/duration.html`, 2014. Zuletzt gesehen: 27. MÃ¤rz 2014.

[TU09] THEUS, MARTIN und URBANEK, SIMON: *Interactive Graphics for Data Analysis*. CRC Press, 2009.

[Tuf83] TUFTE, EDWARD: *The Visual Display of Quantitive Information*. Graphics Press, 1983.

[Wik14] WIKIPEDIA: *Golden Snowball Award*. `http://en.wikipedia.org/wiki/Golden_Snowball_Award`, 2014. Zuletzt gesehen: 28. 3. 2014.

Alphabetisches Verzeichnis der Kurzbezeichnungen

Hier finden Sie ein alphabetisches Verzeichnis der Kurzbezeichnungen der Datensätze.