

# Inhaltsverzeichnis

## Teil I Rahmen

<b>1</b>	<b>Statistik heute</b> . . . . .	3
1.1	Datenanalyse, Statistik, Data Science und Co. . . . .	4
1.2	Wissensgebiete der Datenanalyse . . . . .	6
1.3	Einige Grundbegriffe . . . . .	8
1.4	Signal und Rauschen . . . . .	9
<b>2</b>	<b>Hallo, R</b> . . . . .	13
2.1	Eine kurze Geschichte von R . . . . .	13
2.2	Warum R? Warum, R? . . . . .	15
2.2.1	Warum R? . . . . .	15
2.2.2	Warum, R? . . . . .	17
<b>3</b>	<b>R starten</b> . . . . .	21
3.1	R und RStudio installieren . . . . .	21
3.2	Pakete . . . . .	23
3.2.1	Pakete von CRAN installieren . . . . .	23
3.2.2	Pakete installieren vs. Pakete starten (laden) . . . . .	24
3.2.3	Pakete wie <code>pradadata</code> von Github installieren . . . . .	25
3.3	Hilfe! R startet nicht! . . . . .	25
3.4	Zuordnung von Paketen zu Befehlen . . . . .	27
3.5	R-Skript-Dateien . . . . .	29
3.6	Daten . . . . .	29
3.6.1	Datensätze aus verschiedenen R-Paketen . . . . .	29
3.6.2	Datensätze aus dem R-Paket <code>pradadata</code> . . . . .	30
3.7	Grundlagen der Arbeit mit RStudio . . . . .	30
3.7.1	Das Arbeitsverzeichnis . . . . .	31
3.7.2	RStudio-Projekte . . . . .	32
3.8	Hier werden Sie geholfen . . . . .	33
3.8.1	Wo finde ich Hilfe? . . . . .	33
3.8.2	Einfache, reproduzierbare Beispiele (ERBies) . . . . .	33

<b>4</b>	<b>Erstkontakt</b>	37
4.1	R ist pingelig	37
4.2	Variablen zuweisen und auslesen	38
4.3	Funktionen aufrufen	39
4.4	Logische Prüfungen	40
4.5	Vektorielle Funktionen	42
4.6	Literaturempfehlungen	43

## Teil II Daten einlesen

<b>5</b>	<b>Datenstrukturen</b>	47
5.1	Überblick über die wichtigsten Objekttypen	47
5.2	Objekttypen in R	49
5.2.1	Vektoren	49
5.2.2	Faktoren	51
5.2.3	Listen	53
5.2.4	Matrizen und Arrays	53
5.2.5	Dataframes	53
5.3	Daten auslesen und indizieren	55
5.3.1	Reine Vektoren	55
5.3.2	Matrizen und Arrays	57
5.3.3	Listen	57
5.3.4	Dataframes	59
5.4	Namen geben	60
<b>6</b>	<b>Datenimport und -export</b>	63
6.1	Daten in R importieren	63
6.1.1	Excel-Dateien importieren	64
6.1.2	Daten aus R-Paketen importieren	64
6.1.3	Daten im R-Format laden	65
6.1.4	CSV-Dateien importieren	65
6.2	Textkodierung	68
6.3	Daten exportieren	69

## Teil III Daten aufbereiten

<b>7</b>	<b>Datenjudo</b>	75
7.1	Daten aufbereiten mit <code>dplyr</code>	77
7.2	Zentrale Bausteine von <code>dplyr</code>	78
7.2.1	Zeilen filtern mit <code>filter()</code>	78
7.2.2	Fortgeschrittene Beispiele für <code>filter()</code>	80

---

7.2.3	Spalten wählen mit <code>select()</code> . . . . .	81
7.2.4	Zeilen sortieren mit <code>arrange()</code> . . . . .	82
7.2.5	Einen Datensatz gruppieren mit <code>group_by()</code> . . . . .	84
7.2.6	Eine Spalte zusammenfassen mit <code>summarise()</code> . . . . .	86
7.2.7	Zeilen zählen mit <code>n()</code> und <code>count()</code> . . . . .	89
7.3	Die Pfeife . . . . .	91
7.4	Spalten berechnen mit <code>mutate()</code> . . . . .	93
7.5	Bedingte Analysen mit den Suffixen von <code>dplyr</code> . . . . .	96
7.5.1	Suffix <code>_if</code> . . . . .	96
7.5.2	Suffix <code>_all</code> . . . . .	97
7.5.3	Suffix <code>_at</code> . . . . .	97
7.6	Tabellen zusammenführen ( <code>join</code> ) . . . . .	99
<b>8</b>	<b>Deskriptive Statistik</b> . . . . .	<b>103</b>
8.1	Univariate Statistik . . . . .	104
8.1.1	Deskriptive Statistik mit <code>mosaic</code> . . . . .	107
8.1.2	Deskriptive Statistik mit <code>dplyr</code> . . . . .	108
8.1.3	Relative Häufigkeiten . . . . .	109
8.2	Korrelationen berechnen . . . . .	112
<b>9</b>	<b>Praxisprobleme der Datenaufbereitung</b> . . . . .	<b>117</b>
9.1	Fehlende Werte . . . . .	118
9.1.1	Ursachen von fehlenden Werten . . . . .	118
9.1.2	Auf fehlende Werte prüfen . . . . .	119
9.1.3	Umgang mit fehlenden Werten . . . . .	119
9.1.4	Fälle mit fehlenden Werten löschen . . . . .	119
9.1.5	Fehlende Werte einer Spalte zählen . . . . .	121
9.1.6	Fehlende Werte ersetzen . . . . .	122
9.1.7	-99 in NA umwandeln . . . . .	124
9.2	Datenanomalien . . . . .	125
9.2.1	Doppelte Fälle löschen . . . . .	125
9.2.2	Nach Anomalien suchen . . . . .	126
9.2.3	Ausreißer identifizieren . . . . .	127
9.2.4	Hochkorrelierte Variablen finden . . . . .	128
9.2.5	Quasi-Konstante finden . . . . .	129
9.2.6	Auf Normalverteilung prüfen . . . . .	130
9.3	Daten umformen . . . . .	130
9.3.1	Aufgeräumte Dataframes . . . . .	130
9.3.2	Langes vs. breites Format . . . . .	131
9.3.3	z-Standardisieren . . . . .	133
9.3.4	Spaltennamen ändern . . . . .	134
9.3.5	Variablentypen ändern . . . . .	135

9.4	Werte umkodieren und partitionieren . . . . .	136
9.4.1	Umkodieren und partitionieren mit <code>car::recode()</code> . . . . .	137
9.4.2	Einfaches Umkodieren mit einer Logik-Prüfung . . . . .	138
9.4.3	Binnen mit <code>cut()</code> . . . . .	139
9.5	Vektoren zu Skalaren zusammenfassen . . . . .	141
9.5.1	Mittelwerte pro Zeile berechnen . . . . .	141
9.5.2	Beliebige Statistiken pro Zeile berechnen mit <code>rowwise()</code> . . . . .	142
<b>10</b>	<b>Fallstudie: Datenjudo</b> . . . . .	<b>145</b>
10.1	Deskriptive Statistiken zu den New Yorker Flügen . . . . .	146
10.2	Visualisierungen zu den deskriptiven Statistiken . . . . .	149
10.2.1	Maximale Verspätung . . . . .	149
10.2.2	Durchschnittliche Verspätung . . . . .	151
10.2.3	Korrelate der Verspätung . . . . .	152
<b>Teil IV Daten visualisieren</b>		
<b>11</b>	<b>Datenvisualisierung mit ggplot2</b> . . . . .	<b>157</b>
11.1	Einstieg in ggplot2 . . . . .	158
11.1.1	Ein Bild sagt mehr als 1000 Worte . . . . .	158
11.1.2	Diagramme mit ggplot2 zeichnen . . . . .	158
11.1.3	Die Anatomie eines Diagramms . . . . .	159
11.1.4	Schnell Diagramme erstellen mit <code>qplot()</code> . . . . .	162
11.1.5	ggplot-Diagramme mit <code>mosaic</code> . . . . .	164
11.2	Häufige Arten von Diagrammen (Geomen) . . . . .	166
11.2.1	Eine kontinuierliche Variable – Histogramme und Co. . . . .	166
11.2.2	Zwei kontinuierliche Variablen . . . . .	167
11.2.3	Eine oder zwei nominale Variablen . . . . .	169
11.2.4	Zusammenfassungen zeigen . . . . .	174
11.3	Die Gefühlswelt von ggplot2 . . . . .	178
11.4	<code>ggplot()</code> , der große Bruder von <code>qplot()</code> . . . . .	179
<b>12</b>	<b>Fortgeschrittene Themen der Visualisierung</b> . . . . .	<b>187</b>
12.1	Farbwahl . . . . .	187
12.1.1	Die Farben von Cynthia Brewer . . . . .	188
12.1.2	Die Farben von Wes Anderson . . . . .	190
12.1.3	Viridis . . . . .	193
12.2	ggplot2-Themen . . . . .	194
12.2.1	Schwarz-Weiß-Druck . . . . .	195
12.3	Interaktive Diagramme . . . . .	197
12.3.1	Plotly . . . . .	197
12.3.2	Weitere interaktive Diagramme . . . . .	198

<b>13</b>	<b>Fallstudie: Visualisierung</b>	201
13.1	Umfragedaten visualisieren mit „likert“	202
13.2	Umfragedaten visualisieren mit ggplot	203
13.2.1	Daten aufbereiten	203
13.2.2	Daten umstellen	204
13.2.3	Diagramme für Anteile	204
13.2.4	Rotierte Balkendiagramme	207
13.2.5	Text-Labels	208
13.2.6	Diagramm beschriften	211
13.2.7	Balken mit Häufigkeitswerten	211
13.2.8	Sortieren der Balken	212
<b>14</b>	<b>Geovisualisierung</b>	215
14.1	Kartendaten	216
14.1.1	Geo-Daten der deutschen Verwaltungsgebiete	216
14.1.2	Daten der Wahlkreise	218
14.2	Unterschiede in Kartensegmenten visualisieren	219
14.2.1	Karte der Wahlkreise gefärbt nach Arbeitslosigkeit	219
14.2.2	Wahlergebnisse nach Wahlkreisen	220
14.2.3	Zusammenhang von Arbeitslosigkeit und AfD-Wahlergebnis	222
14.2.4	Ein komplexeres Modell	223
14.3	Weltkarten	224
14.3.1	rworldmap	224
14.3.2	rworldmap mit geom_sf	227
14.4	Anwendungsbeispiel: Konkordanz von Kulturwerten und Wohlbefinden	229
14.5	Interaktive Karten	234
14.5.1	Karten mit „leaflet“	234
14.5.2	Karten mit googleVis	235

## Teil V Modellieren

<b>15</b>	<b>Grundlagen des Modellierens</b>	245
15.1	Was ist ein Modell? Was ist Modellieren?	246
15.2	Abduktion als Erkenntnisfigur im Modellieren	248
15.3	Ein Beispiel zum Modellieren in der Datenanalyse	250
15.4	Taxonomie der Ziele des Modellierens	251
15.5	Die vier Schritte des statistischen Modellierens	254
15.6	Einfache vs. komplexe Modelle: Unter- vs. Überanpassung	255

15.7	Bias-Varianz-Abwägung . . . . .	256
15.8	Trainings- vs. Test-Stichprobe . . . . .	257
15.9	Resampling und Kreuzvalidierung . . . . .	259
15.10	Wann welches Modell? . . . . .	260
15.11	Modellgüte . . . . .	260
	15.11.1 Modellgüte in numerischen Vorhersagemodellen . . . . .	261
	15.11.2 Modellgüte bei Klassifikationsmodellen . . . . .	261
15.12	Der Fluch der Dimension . . . . .	262
<b>16</b>	<b>Inferenzstatistik . . . . .</b>	<b>267</b>
16.1	Wozu Inferenzstatistik? . . . . .	268
16.2	Der $p$ -Wert . . . . .	269
	16.2.1 Was sagt der $p$ -Wert? . . . . .	269
	16.2.2 Der zwielichtige Statistiker – ein einführendes Beispiel zur Inferenzstatistik . . . . .	271
	16.2.3 Von Männern und Päpsten – Was der $p$ -Wert nicht sagt . . . .	274
	16.2.4 Der $p$ -Wert ist eine Funktion der Stichprobengröße . . . . .	275
	16.2.5 Mythen zum $p$ -Wert . . . . .	276
16.3	Wann welcher Inferenztest? . . . . .	277
16.4	Beispiele für häufige Inferenztests . . . . .	278
	16.4.1 $\chi^2$ -Test . . . . .	278
	16.4.2 t-Test . . . . .	280
	16.4.3 Einfache Varianzanalyse . . . . .	281
	16.4.4 Korrelationen (nach Pearson) auf Signifikanz prüfen . . . . .	282
	16.4.5 Regression . . . . .	283
	16.4.6 Wilcoxon-Test . . . . .	284
	16.4.7 Kruskal-Wallis-Test . . . . .	284
	16.4.8 Shapiro-Test . . . . .	285
	16.4.9 Logistische Regression . . . . .	285
	16.4.10 Spearmans Korrelation . . . . .	286
16.5	Alternativen zum $p$ -Wert . . . . .	286
	16.5.1 Konfidenzintervalle . . . . .	286
	16.5.2 Effektstärke . . . . .	289
	16.5.3 Power-Analyse . . . . .	292
	16.5.4 Bayes-Statistik . . . . .	293
<b>17</b>	<b>Simulationsbasierte Inferenz . . . . .</b>	<b>301</b>
17.1	Stichproben, Statistiken und Population . . . . .	301
17.2	Die Stichprobenverteilung . . . . .	304
17.3	Der Bootstrap . . . . .	308
17.4	Nullhypothesen auf Signifikanz testen . . . . .	311

**Teil VI Geleitetes Modellieren**

<b>18</b>	<b>Lineare Modelle</b>	321
18.1	Die Idee der klassischen Regression	321
18.2	Modellgüte	324
18.2.1	Mittlere Quadratfehler	325
18.2.2	R-Quadrat ( $R^2$ )	325
18.3	Die Regression an einem Beispiel erläutert	327
18.4	Überprüfung der Annahmen der linearen Regression	329
18.5	Regression mit kategorialen Prädiktoren	331
18.6	Multiple Regression	333
18.7	Interaktionen	335
18.8	Prädiktorenrelevanz	337
18.9	Anwendungsbeispiel zur linearen Regression	339
18.9.1	Overfitting	339
18.9.2	Konfidenzintervalle der Parameter	341
<b>19</b>	<b>Klassifizierende Regression</b>	345
19.1	Normale Regression für ein binäres Kriterium	346
19.2	Die logistische Funktion	347
19.3	Interpretation des Logits	350
19.4	Kategoriale Prädiktoren	351
19.5	Multiple logistische Regression	352
19.6	Modellgüte	353
19.6.1	Vier Arten von Ergebnissen einer Klassifikation	353
19.6.2	Kennzahlen der Klassifikationsgüte	355
19.7	Vorhersagen	356
19.8	ROC-Kurven und Fläche unter der Kurve (AUC)	357
19.8.1	Cohens Kappa	360
<b>20</b>	<b>Fallstudie: Titanic</b>	365
20.1	Explorative Analyse	366
20.1.1	Univariate Häufigkeiten	366
20.1.2	Bivariate Häufigkeiten	367
20.2	Inferenzstatistik	368
20.2.1	$\chi^2$ -Test	368
20.2.2	Effektstärke	369
20.2.3	Logistische Regression	373
<b>21</b>	<b>Baumbasierte Verfahren</b>	377
21.1	Entscheidungsbäume	378
21.1.1	Einführendes Beispiel	378
21.1.2	Tuningparameter	383

21.2	Entscheidungsbäume mit <code>caret</code> . . . . .	384
21.2.1	Vorhersagegüte . . . . .	388
21.3	Der Algorithmus der Entscheidungsbäume . . . . .	391
21.4	Regressionsbäume . . . . .	391
21.5	Stärken und Schwächen von Bäumen . . . . .	391
21.6	Bagging . . . . .	393
21.7	Grundlagen von Random Forests . . . . .	394
21.7.1	Grundlagen . . . . .	394
21.8	Variablenrelevanz bei Baummodellen . . . . .	398
<b>22</b>	<b>Fallstudie: Kreditwürdigkeit mit <code>caret</code></b> . . . . .	<b>401</b>
22.1	Zwei Arten der prädiktiven Modellierung . . . . .	402
22.2	Daten aufbereiten . . . . .	403
22.2.1	Fehlende Werte . . . . .	403
22.2.2	Trainings- und Test-Sample aufteilen . . . . .	403
22.2.3	Variablen ohne Varianz . . . . .	404
22.2.4	Hochkorrelierte Variablen entfernen . . . . .	405
22.2.5	Parallele Verarbeitung . . . . .	406
22.3	Modelle anpassen . . . . .	407
22.3.1	Kreuzvalidierung . . . . .	407
22.3.2	Modell im Trainings-Sample anpassen mit <code>train()</code> . . . . .	407
22.3.3	Ein einfaches Modell . . . . .	408
22.3.4	Random Forest . . . . .	411
22.3.5	Support Vector Machines . . . . .	414
22.3.6	Penalisierte lineare Modelle . . . . .	416
22.4	Modellgüte bestimmen . . . . .	418
22.4.1	Modellgüte in der Test-Stichprobe . . . . .	418
22.4.2	Modellgüte in der Kreuzvalidierung . . . . .	421
22.5	Wichtigkeit der Prädiktoren bestimmen . . . . .	426
22.5.1	Modellunabhängige Variablenwichtigkeit für Klassifikation . . . . .	427
22.5.2	Modellunabhängige Prädiktorenrelevanz bei numerischen Vorhersagen . . . . .	429
22.5.3	Modellabhängige Variablenwichtigkeit . . . . .	430
 <b>Teil VII Ungeleitetes Modellieren</b>		
<b>23</b>	<b>Clusteranalyse</b> . . . . .	<b>437</b>
23.1	Grundlagen der Clusteranalyse . . . . .	437
23.1.1	Intuitive Darstellung der Clusteranalyse . . . . .	438
23.1.2	Euklidische Distanz . . . . .	440
23.1.3	k-Means-Clusteranalyse . . . . .	442



23.2	Beispiel für eine einfache Clusteranalyse	443
23.2.1	Distanzmaße berechnen	443
23.2.2	Fallstudie: kmeans für den Extraversionsdatensatz	444
<b>24</b>	<b>Textmining</b>	449
24.1	Grundlegende Analyse	450
24.1.1	Tidytext-Dat.frames	450
24.1.2	Regulärausdrücke	453
24.1.3	Textdaten einlesen	455
24.1.4	Worthäufigkeiten auszählen	456
24.1.5	Visualisierung	457
24.2	Sentimentanalyse	459
<b>25</b>	<b>Fallstudie: Twitter-Mining</b>	463
25.1	Zum Einstieg: Moderne Methoden der Sentimentanalyse	464
25.2	Grundlagen des Twitter-Minings	465
25.2.1	Authentifizierung bei der Twitter-API	466
25.2.2	Hashtags und Nutzer suchen	467
25.2.3	Tweets einer Nutzermenge auslesen	469
25.2.4	Aufbau einer Tweets-Datenbank	471
 <b>Teil VIII Kommunizieren</b>		
<b>26</b>	<b>RMarkdown</b>	475
26.1	Forderungen an Werkzeuge zur Berichterstellung	476
26.2	Start mit RMarkdown	478
26.3	RMarkdown in Action	480
26.4	Aufbau einer Markdown-Datei	482
26.5	Syntax-Grundlagen von Markdown	483
26.6	Tabellen	484
26.7	Zitieren	487
26.8	Format-Vorlagen für RMarkdown	489
 <b>Teil IX Rahmen 2</b>		
<b>27</b>	<b>Projektmanagement am Beispiel einer Fallstudie</b>	495
27.1	Was ist Populismus?	496
27.2	Forschungsfrage und Operationalisierung	497
27.3	Emotionslexikon	498
27.4	Daten, Stichprobe und Analysekontext	499

27.5	Prozess der Datenanalyse . . . . .	499
27.6	Zentrale Ergebnisse . . . . .	501
27.7	Projektmanagement . . . . .	504
27.7.1	Gliederung eines Projektverzeichnisses . . . . .	504
27.7.2	Faustregeln zur Struktur eines Projekts . . . . .	505
27.7.3	Versionierung mit Git . . . . .	508
<b>28</b>	<b>Programmieren mit R . . . . .</b>	<b>511</b>
28.1	Funktionen schreiben . . . . .	511
28.2	Wiederholungen . . . . .	514
28.2.1	Wiederholungen für Elemente eines Vektors . . . . .	515
28.2.2	Wiederholungen für Spalten eines Dataframes . . . . .	516
28.2.3	Dateien wiederholt einlesen . . . . .	518
28.2.4	Anwendungsbeispiele für map . . . . .	520
28.2.5	Einige Rechtschreibregeln für map ( ) . . . . .	523
28.3	Defensives Programmieren . . . . .	523
<b>29</b>	<b>Programmieren mit dplyr . . . . .</b>	<b>527</b>
29.1	Wie man mit dplyr nicht sprechen darf . . . . .	527
29.2	Standard-Evaluation vs. Non-Standard-Evaluation . . . . .	528
29.3	NSE als Backen . . . . .	530
29.4	Wie man Funktionen mit dplyr-Verben schreibt . . . . .	534
29.5	Beispiele für NSE-Funktionen . . . . .	537
29.5.1	Funktionen für ggplot . . . . .	539
	<b>Anhang A . . . . .</b>	<b>541</b>
	<b>Literatur . . . . .</b>	<b>547</b>
	<b>Sachverzeichnis . . . . .</b>	<b>559</b>