Datenanalyse mit R

SoSe 2020

Christina Bogner

Version vom 12. Mai 2020

Contents

Chapter 1

Vorwort

> "And honey, we're gonna do it in style"
>
> — Fools Garden

1.1 Organisatorisches

Die Coronaviruspandemie verändert unser Leben und unser Lernen. Die UzK bittet Lehrende, zumindest zu Beginn des SoSe 2020 auf digitale Lernformen umzusteigen. Daher wird dieser Kurs als ein Onlinekurs beginnen. Abhängig von der (sehr dynamischen) Lage werden wir im weiteren Kursverlauf das Format anpassen. Bitte seien Sie nachsichtig, wenn nicht alles so klappt, wie in Präsenzveranstaltungen. Wir müssen aktuell alle sehr viel dazu lernen in Sachen digitale Lehre. Sie können sicher sein, dass das Geographische Institut bemüht ist, die Lehre so effizient wie möglich weiter laufen zu lassen, damit Sie in Ihrem Studium fortfahren können.

In dieser Veranstaltung werden wir folgende Werkzeuge verwenden:

1. **ILIAS**: die Online-Lernplattform der UzK. Entweder sind Sie bereits automatisch in dem Kurs registriert oder werden von mir per Hand angemeldet.
2. **Campuswire**: die Live-Chatplattform dient der allgemeinen Kommunikation und der Selbstorganisation des Lernens. Verwenden Sie diese, um Fragen mit Ihren Kommilitonen und mir zu diskutieren. Sie sollten eine Einladungsmail zu Campuswire erhalten haben.
3. **Zoom**: die Videokonferenz-Software werden wir für live Einführungen nutzen. Die Anmeldemodalitäten sind auf den Kursseiten in ILIAS erklärt.

1.2 Sinn und Unsinn dieses Skripts

Dieses Skript ist ein lebendiges Begleitdokument des Kurses. Es wird laufend angepasst und aktualisiert.

Ich nutze verschiedene Farbkästen, um wichtige Stellen hervorzuheben:

Infoblock

Achtung, wichtig!

Beispielblock

Lernziele

Zusammenfassung

Chapter 2

Der Kurs

2.1 Zuordnung zum Modul und Leistungsnachweis

Dieser Kurs gehört zum Modul *Fachmethodik I* oder *Fachmethodik II* und ist aus 4 SWS Praktikum und 2 SWS Seminar aufgebaut. Das wichtigste Ziel besteht darin, Ihnen einen sicheren Umgang mit `R` beizubringen.

Den Leistungsnachweis bildet ein benoteter Praktikumsbericht.

2.2 Lernziele des Kurses

- Daten für Analysen vorbereiten
- eigene wiederverwendbare Skripte schreiben
- eigene Funktionen schreiben
- einfache Datenanalysen durchführen
- Daten visualisieren
- Ergebnisse reproduzierbar im Praktikumsbericht darstellen

2.3 Was mir im Umgang miteinander wichtig ist

- Pünktlichkeit bei live und Präsenzsitzungen
- Gute Vorbereitung durch erledigen der blenden learning Einheiten und Hausaufgaben
- Respektieren anderer Meinungen
- Offenheit gegenüber neuen Sichtweisen, Themen und Methoden
- Geduld mit sich selbst und den anderen

Chapter 3

Erste Schritte in

- Layout und Bedeutung einzelner Fenster in RStudio kennen
- Anweisungen aus dem Skript an die Konsole schicken
- R als Taschenrechner benutzen
- erste Funktionen aufrufen
- Objekte mit eckigen Klammern [ ] ansprechen
- R-Hilfeseiten aufrufen

3.1 Was ist R?

R ist eine Programmiersprache für Datenanalyse und statistische Modellierung. Es ist frei verfügbar (*open source software*) und neben Python einer der am meisten benutzten Programmiersprachen zur Datenanalyse und -visualisierung. R wurde von Ross Ihaka und Robert Gentleman 1996 veröffentlicht Ihaka and Gentleman (1996). Es gibt für R eine Vielzahl von Zusatzpaketen, die die Funktionalität und die Einsatzmöglichkeiten enorm erweitern.

Sie können R für Ihren Computer auf der offiziellen R-Seite https://www.r-project.org/ herunter laden und installieren. Auch die Pakete finden Sie dort unter CRAN (*The Comprehensive R Archive Network*). Auf den CRAN-Seiten finden Sie sogen. CRAN Task Views, eine Übersicht über Pakete in verschiedenen Themenbereichen. Für den Umweltbereich sind folgende Paketsammlungen besonders relevant:

- Environmetrics: Analyse von Umweltdaten
- Multivariate: Multivariate Statistik
- Spatial: Analyse von räumlichen Daten
- TimeSeries: Zeitreihenanalyse

Zu Beginn des Kurses, werden wir jedoch nicht auf Ihren lokalen Rechnern arbeiten, sondern in einer Cloud (s.u.). Das ermöglicht einen schnelleren Einstieg in R und bietet eine live Unterstützung durch den Dozenten beim Pro-

grammieren. Daher biete ich zu diesem frühen Zeitpunkt im Kurs keine Unterstützung bei der Installation. Für die ganz Ungeduldigen, gibt es hier eine kurze *Einleitung zur Installation*

3.2 Was ist RStudio?

RStudio Desktop ist eine Entwicklungsumgebung für R. Sie können die *open source* Version kostenlos für Ihren Rechner hier herunterladen.

Es gibt eine live Einführung in RStudio im Kurs. Zusätzlich können Sie hier ein Video dazu ansehen.

3.3 RStudio Cloud

Zu Beginn des Kurses werden wir in der RStudio Cloud arbeiten. Sie sollten eine Einladungsmail zu unserem Kurs in der Cloud bekommen haben. Ich werde in der Cloud Projekte für Sie anlegen (*assignment*), die Skripte, Arbeitsanweisungen etc. beinhalten. Wenn Sie auf so ein Assignment klicken, wird für Sie automatische ein Kopie des Projekts erstellt, in der Sie dann arbeiten können.

Der große Vorteil der Cloud ist, dass ich direkt in Ihre Projekte eingreifen kann, wenn es mal zu Fehlern kommt. Während ich in Ihrem Projekt arbeite, werden Sie kurz aus der R-Sitzung ausgeloggt, da die Cloud kein gleichzeitiges Arbeiten unterstützt. Nehmen Sie sich etwas Zeit, um die Cloud und die darin enthaltenen Tutorials kennen zu lernen.

Sowohl in der RStudio Cloud als auch in einer lokalen Installation, ist Ihr RStudio so aufgebaut wie in Abbildung 3.1.

3.4 Inhalt der live Einführung

- Überblick über RStudio
- R als Taschenrechner
- einfache Funktionen aufrufen
- Zuordnungen (*assignments*)
- Notation mit eckigen Klammern [ ] (*array*-Notation)
- Hilfeseiten aufrufen

Funktionen, die wir in der Session nutzen werden:

Funktion	Bedeutung	Beispielaufruf
`pi`	Zahl pi	`pi`
`sin`	Sinus	`sin(2)`
`cos`	Cosinus	`cos(2)`
`sqrt`	Quadratwurzel	`sqrt(2)`

Funktion	Bedeutung	Beispielaufruf
`c`	(*concatenate*) Fügt Daten zu einem Vektor zusammen	`c(1,2,3,4)`
`help.start`	Öffnet ein Browser-Fenster mit diversen Handbüchern	`help.start()`
`help.search`	Sucht nach einem Begriff in Hilfe-Dateien	`help.search('time')`
`??`	alias `help.search`	`??time`
`help`	Sucht nach einer Funktion	`?mean`
`?`	alias `help()`	`?mean`
`mean`	Mittelwert	`mean(c(1,2,3,4))`
`var`	Varianz	`var(c(1,2,3,4))`
`sd`	Standardabweichung	`sd(c(1,2,3,4))`
`sum`	Summe	`sum(c(1,2,3,4))`
`vector`	Generiert einen Vektor	`vector(length=3, mode='numeric')`

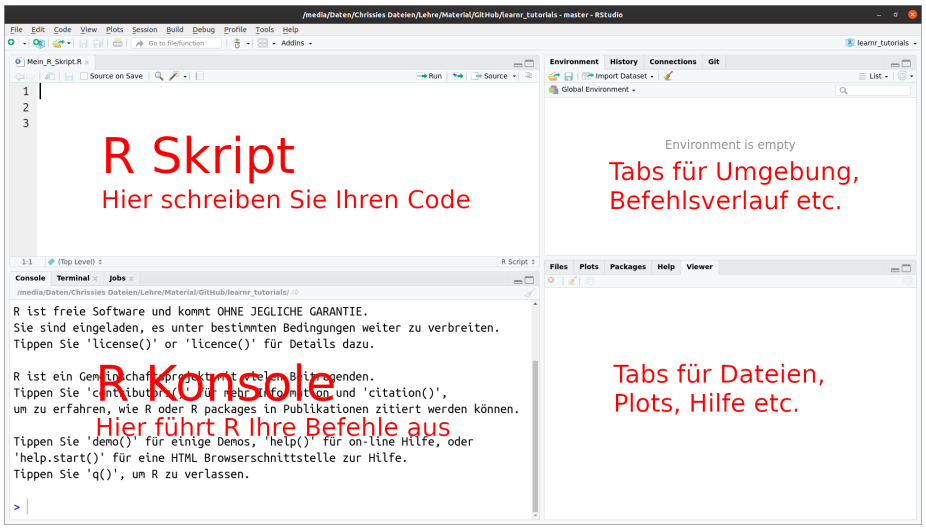


Figure 3.1: Aufbau von RStudio

Chapter 4

Daten in R

- Daten einlesen mit `read.table`
- Datenstrukturen erstellen
- Typen von Daten in R abfragen
- Daten speichern mit `write.table`

4.1 Datenstrukturen erzeugen

In R gibt es unterschiedliche Datenobjekte. Es ist wichtig, sich über die Struktur (oder Typ) des Datenobjekts Gedanken zu machen. Denn diese bestimmt, was mit einem Objekt gemacht werden kann und ob Funktionen damit richtig umgehen können. Schließlich ist es nicht egal, ob es sich bei einem Objekt um ein numerisches Objekt oder einfach Text (*character*) handelt.

Die wichtigsten Datentypen sind

- **Vektoren**: hier gruppiert man gleichartige Elemente, z.B. Zahlen. Auch eine einzelne Zahl (ein Skalar) wird von R wie ein Vektor behandelt.
- **Matrizen**: zweidimensionale (Zeilen und Spalten) Datentabellen mit gleichartigen Elementen.
- **Listen**: können beliebige Elemente beliebiger Länge enthalten.
- **Dataframes**: zweidimensionale Datentabellen, die beliebige Elemente enthalten können. Die Spalten der Dataframes müssen allerdings gleichartige Elemente enthalten. Dataframes sind eine Unterart von Listen.

Neben diesen Hauptstrukturen gibt es

- **Factor**: ein besonderer Vektor für kategorielle Variablen

Um diese Datenstrukturen zu erzeugen, gibt es jeweils eine Funktion mit gleichlautendem Namen.

```r
# Vektor erzeugen
my_vect = vector(length = 3, mode = 'numeric')
my_vect
```

```
## [1] 0 0 0
```

```r
# Matrix erzeugen
my_matrix = matrix(data = c(1:(3*4)), nrow = 3, ncol = 4)
my_matrix
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    4    7   10
## [2,]    2    5    8   11
## [3,]    3    6    9   12
```

```r
# Dataframe erzeugen
my_dataframe = data.frame('Spalte_1' = rep('Text', 10),
                          'Spalte_2' = 1:10)
my_dataframe
```

```
##    Spalte_1 Spalte_2
## 1      Text        1
## 2      Text        2
## 3      Text        3
## 4      Text        4
## 5      Text        5
## 6      Text        6
## 7      Text        7
## 8      Text        8
## 9      Text        9
## 10     Text       10
```

```r
# Liste erzeugen
my_list = list('Schachtel_1' = 3, 'Schachtel_2' = my_dataframe,
               'Schachtel_3' = 'Noch mehr Text')
my_list
```

```
## $Schachtel_1
## [1] 3
##
## $Schachtel_2
##    Spalte_1 Spalte_2
## 1      Text        1
## 2      Text        2
```

```
## 3      Text        3
## 4      Text        4
## 5      Text        5
## 6      Text        6
## 7      Text        7
## 8      Text        8
## 9      Text        9
## 10     Text       10
##
## $Schachtel_3
## [1] "Noch mehr Text"
```

```
# Factor erzeugen
my_factor = factor(c('R', 'RStudio', 'Cloud', 'Cloud', 'R', 'R'))
my_factor
```

```
## [1] R       RStudio Cloud   Cloud   R       R
## Levels: Cloud R RStudio
```

4.2 Arten von Daten in R

Die Datenstrukturen `vector`, `data.frame` usw. können unterschiedliche Arten von Daten enthalten.

Name	Beispiele
raw	3A, FE
logical	TRUE, FALSE
integer	1, 42, -3
numeric/double	3, 2.81, 6.032e23
complex	1.2+2.2i
character	“foo”

4.3 Objekt, sag mir wer du bist

Um die Struktur und/oder Datenart abzufragen, verwendet man `class`, `typeof`, `mode` und `storage.mode`.

```
class(my_vect)
```

```
## [1] "numeric"
```

```
typeof(my_vect)
```

```
## [1] "double"
```

```
class(my_dataframe)
```

```
## [1] "data.frame"
```

```
typeof(my_dataframe)
```

```
## [1] "list"
```

Mit `str` kann man das Innenleben eines Objekts anzeigen. Das ist besonders wichtig nach dem Einlesen von Daten, um das Ergebnis des Einlesens zu kontrollieren. Dabei kontrolliert man, dass z.B. alle numerischen Spalten auch als Zahlen eingelesen wurden und nichts schief gegangen ist.

```
str(my_dataframe)
```

```
## 'data.frame':    10 obs. of  2 variables:
##  $ Spalte_1: Factor w/ 1 level "Text": 1 1 1 1 1 1 1 1 1 1
##  $ Spalte_2: int  1 2 3 4 5 6 7 8 9 10
```

Weiter Funktionen, die Auskunft über Objekte geben sind `length`, sinnvoll auf nur Vektoren und Listen, und `dim`, sinnvoll auf zweidimensionalen Datenobjekten. Wenn Sie versuchen, `dim` auf einem Vektor aufzurufen, gibt es `NULL` (s.u.), weil Vektoren keine Dimensionen haben. Wenn Sie `length` auf einem `data.frame` aufrufen, bekommen Sie die Anzahl der Dimensionen, nämlich 2. Das sind keine besonders spannenden Informationen .

```
length(my_vect)
```

```
## [1] 3
```

```
dim(my_vect)
```

```
## NULL
```

```
length(my_dataframe)
```

```
## [1] 2
```

```
dim(my_dataframe)
```

```
## [1] 10  2
```

4.4 Datenlücken, Fehlschläge etc.

Datenlücken werden in R mit `NA` kodiert, Fehlschläge bei Berechnungen mit `NaN` (not a number) und Vektoren der Länge 0 mit `NULL`. Letzteres wird häufig beim Aufruf von Funktionen benutzt, wenn man bestimmte Parameter ausschalten möchte. Die Benutzung muss aber immer in der Hilfe zur jeweiligen Funktion nachgeschlagen werden.

4.5 Inhalt der live Einführung

- Daten einlesen und `data.frame` erstellen: Aufgabe 7.2.1

Funktionen, die wir in der Session nutzen werden:

Funktion	Bedeutung	Beispielaufruf
`read.table`	Liest Daten aus einer Datei ein.	`read.table(file= 'Daten.txt', header=TRUE)`
`ls`	Zeigt den Inhalt des Workspaces.	`ls`
`head`	Zeigt den ersten Teil eines Objekts.	`head(x)`
`tail`	Zeigt den letzten Teil eines Objekts.	`tail(x)`
`str`	Zeigt die Struktur (Innenleben) eines Objekts an	`str(my_dataframe)`
`length`	Gibt die Länge eines Objekts.	`length(x)`
`dim`	Gibt die Dimension eines Objekts (Reihenfolge: Zeilen, Spalten)	`dim(x)`
`seq`	Erstellt eine regelmäßige Reihe.	`seq(from=-2, to=4, by=0.1)`
`data.frame`	Erstellt eine Datentabelle.	`data.frame(x,y,z)`
`colnames`, `rownames`	Benennt Spalten bzw. Zeilen eines Datenobjekts.	`colnames(x)`
`rm`	Löscht Objekte aus dem Workspace.	`rm(x)`
`summary`	Fasst ein Objekt zusammen.	`summary(x)`
`table`	Erstellt eine Häufigkeitstabelle.	`table(x)`

Funktion	Bedeutung	Beispielaufruf
`which`	Gibt die TRUE-Indices eines logischen Objekts.	`which(LETTERS == 'R')`
`history`	Zeigt die Liste mit ausgeführten Befehlen der Session.	`history`
`write.table`	Speichert Datenobjekte als Tabelle ab.	`write.table(x, file='Tabelle.txt')`
`save.image`	Speichert den Workspace.	`save.image(file= 'RSession.Rdata')`
`savehistory`	Speichert die History.	`savehistory(file= 'Myhistory.Rhistory')`

Chapter 5

Daten visualisieren I: Einfache Grafiken

- Einfache Grafiken erstellen
- Grafiken beschriften und speichern
- Die Arbeitsweise der Funktion `par` beschreiben
- Die grafischen Parameter für Randgröße, Farbe, Schrift- und Symbolgröße einstellen
- Unterschiede zwischen *high-level* und *low-level* Grafikfunktionen erklären
- Grafiken mit mehreren Plots erstellen

5.1 Plotten mit R-Basisfunktionen

Für Grafikverliebte und Neugierige empfehle ich die Kapitel 2 und 3 in Murrell (2006).

5.1.1 *High-level* Grafikfunktion `plot` und *low-level* Grafikfunktion `lines`

Ein Streudiagramm stellt zwei numerische Variablen gegeneinander dar. Wir betrachten Klimadaten der Station Köln-Bonn, die man beim Deutschen Wetterdienst herunterladen kann (https://www.dwd.de/DE/leistungen/klimadatendeutschland/klimadatendeutschland.html).

Sie können den Code aus den Chunks leicht herauskopieren und in RStudio laufen lassen (rechts oben in den Chunks auf das Symbol *copy to clipboard* klicken).

Wir lesen die Daten ein und sehen uns deren Struktur an.

```
meteo <- read.table('produkt_klima_monat_20181001_20200430_02667.txt',
header = T, sep = ';')
str(meteo)
```

```
## 'data.frame':    19 obs. of  17 variables:
##  $ STATIONS_ID      : int  2667 2667 2667 2667 2667 2667 2667 2667 2667 2667 ...
##  $ MESS_DATUM_BEGINN: int  20181001 20181101 20181201 20190101 20190201 20190301 201
##  $ MESS_DATUM_ENDE  : int  20181031 20181130 20181231 20190131 20190228 20190331 201
##  $ QN_4             : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ MO_N             : num  4.76 5.51 6.49 6.66 4.79 5.62 4.56 5.37 3.85 4.95 ...
##  $ MO_TT            : num  12.01 7.31 5.64 2.41 6.4 ...
##  $ MO_TX            : num  17.58 10.95 8.47 4.74 12 ...
##  $ MO_TN            : num  6.41 3.51 2.61 -0.26 1.12 ...
##  $ MO_FK            : num  2.42 2.57 2.68 2.84 2.54 3.06 2.53 2.32 2.4 2.23 ...
##  $ MX_TX            : num  26.7 19.2 15 8.8 21 20.4 25.9 24.3 36.2 40.3 ...
##  $ MX_FX            : num  15.3 16.5 18.9 22.8 18.7 28.8 19.5 16.5 26.1 14.6 ...
##  $ MX_TN            : num  0.4 -4.3 -4.9 -10.7 -3.5 -2.8 -2.3 -1.8 7.3 4.4 ...
##  $ MO_SD_S          : num  145.4 74.2 33.8 38.2 127.3 ...
##  $ QN_6             : int  9 9 9 3 3 3 3 3 3 3 ...
##  $ MO_RR            : num  26.5 25 101.9 101.8 30 ...
##  $ MX_RS            : num  6.9 9.7 15.7 22.3 12.3 18.2 21.4 10.6 8.8 11.9 ...
##  $ eor              : Factor w/ 1 level "eor": 1 1 1 1 1 1 1 1 1 1 ...
```

Uns interessieren hier nur die Spalten MO_TT, MO_TN, MO_TX und MESS_DATUM_BEGINN. Das sind jeweils die Monatsmittel der Lufttemperatur in 2 m Höhe, Monatsmittel des Minimums der Lufttemperatur, Monatsmittel des Maximums der Lufttemperatur und der Beginn der jeweiligen Messperiode (d.h. des Kalendermonats). Um die Daten als Zeitreihen darstellen zu können, wandeln wir die Spalte MESS_DATUM_BEGINN in ein richtiges Zeitobjekt (d.h. ein Objekt der Klasse *Date*). Das geht mit der Funktion `as.Date`. Der Parameter `format` beschreibt den Aufbau des Datums im Objekt `meteo`: erst steht das Jahr mit 4 Zeichen (z.B. 2018), dann folgt der Monat mit 2 Zeichen (z.B. 01) und dann der Tag mit 2 Zeichen (z.B. 01). Näheres zu Datumsformaten finden Sie mit `?strptime`.

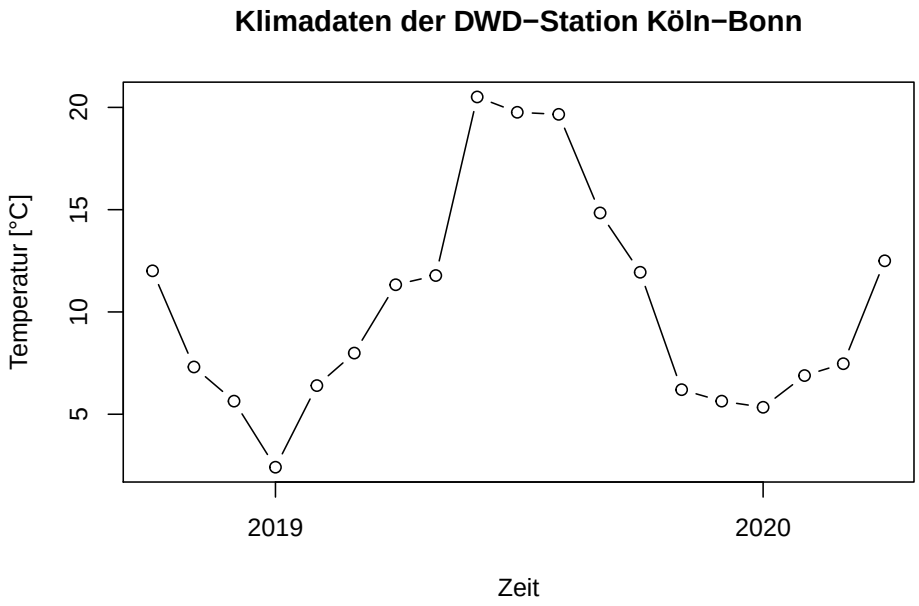
```
my_date <- as.Date(as.character(meteo$MESS_DATUM_BEGINN), format = '%Y%m%d')
my_date
```

```
##  [1] "2018-10-01" "2018-11-01" "2018-12-01" "2019-01-01" "2019-02-01"
##  [6] "2019-03-01" "2019-04-01" "2019-05-01" "2019-06-01" "2019-07-01"
## [11] "2019-08-01" "2019-09-01" "2019-10-01" "2019-11-01" "2019-12-01"
## [16] "2020-01-01" "2020-02-01" "2020-03-01" "2020-04-01"
```

Es sind Daten von Oktober 2018 bis April 2020. Wir erstellen ein Streudiagramm mit der Funktion `plot`. Mit den Parametern `xlab` und `ylab` lassen sich

die beiden Achsen beschriften und `main` fügt einen Titel dazu. Der Parameter `type` bestimmt die Wahl der Symbole; hier benutzen wir `type = b` für *both*, also sowohl Punkte als auch Linien.

```
plot(my_date, meteo$MO_TT, type = 'b', xlab = 'Zeit', ylab = 'Temperatur [°C]', main = 'Klimadate
```



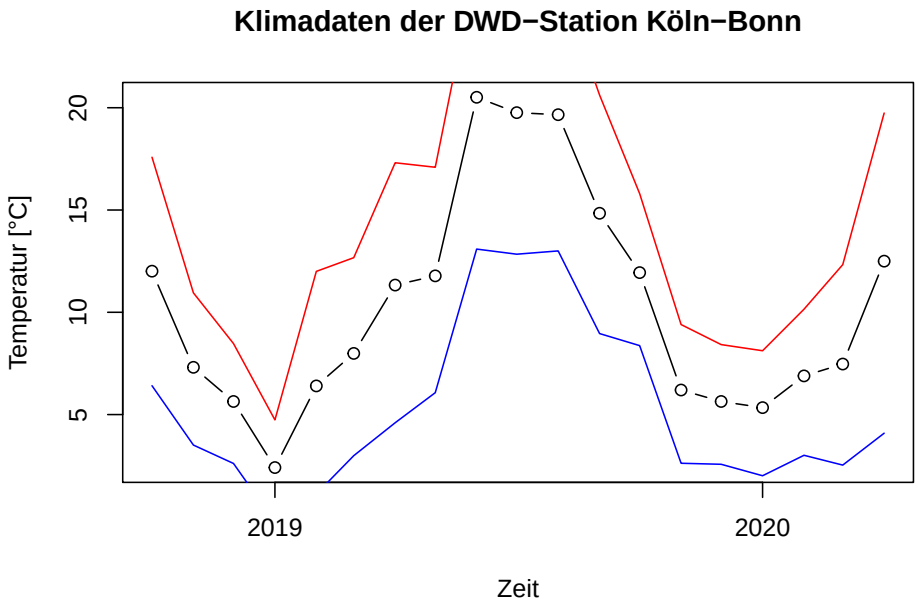
Die Funktion `plot` ist eine sogen. *high-level* Grafikfunktion. Das bedeutet, dass sie alle Schritte des Plottens übernimmt: sie öffnet ein neues Grafikfenster (ein Device), berechnet die Größe der Plotfläche und der Ränder (s. unten), berechnet die Ausdehnung der Achsen und die beste Achseneinteilung und plottet Ihre Daten.

Daneben gibt es *low-level* Grafikfunktionen, die nur in ein bestehendes Device plotten können. Wir wollen zu unserer Grafik nun die Minimum- und die Maximumtemperatur dazu plotten.

```
plot(my_date, meteo$MO_TT, type = 'b', xlab = 'Zeit', ylab = 'Temperatur [°C]', main = 'Klimadate

# Minimumtemperatur in blau
lines(my_date, meteo$MO_TN, col = 'blue')

# Maximumtemperatur in rot
lines(my_date, meteo$MO_TX, col = 'red')
```

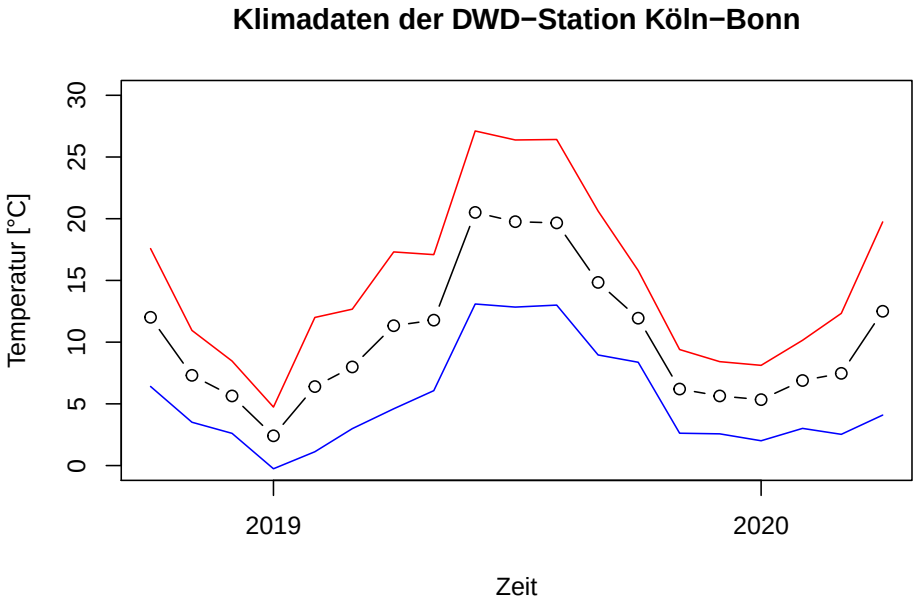


Dass `lines` nur eine *low-level* Grafikfunktion ist, erkennen Sie daran, dass sie nicht in der Lage ist, den Bereich auf der y-Achse zu vergrößern, um alle Daten sichtbar zu machen. Das kann nur `plot`. Daher muss der Bereich bereits in `plot` richtig festgelegt werden. Das macht der Parameter `ylim`.

```
plot(my_date, meteo$MO_TT, type = 'b', xlab = 'Zeit', ylab = 'Temperatur [°C]', main =

# Minimumtemperatur in rot
lines(my_date, meteo$MO_TN, col = 'blue')

# Maximumtemperatur in blau
lines(my_date, meteo$MO_TX, col = 'red')
```



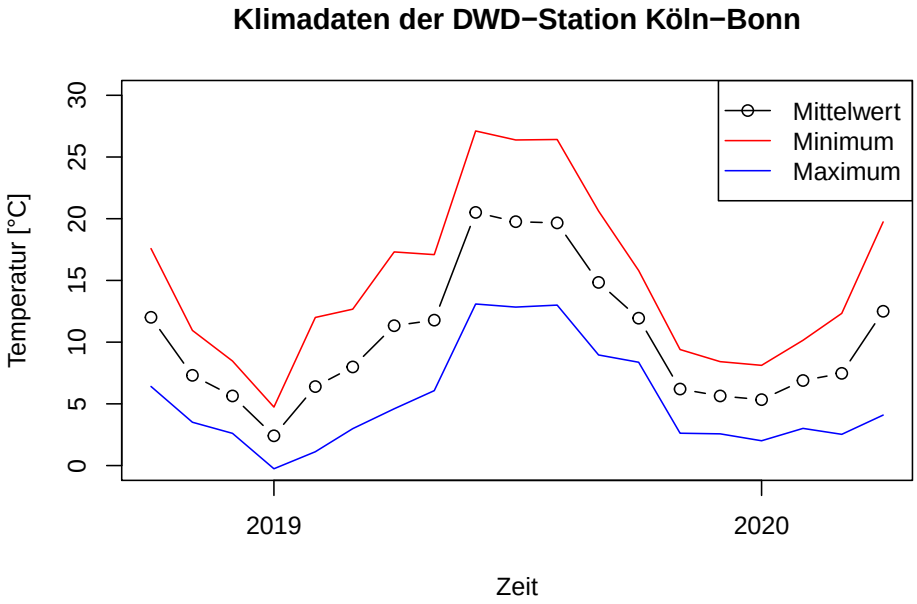
Wenn in einer Grafik mehrere Elemente dargestellt werden, benötigt man eine Legende. Das erledigt die * low-level* Grafikfunktion `legend`.

```
plot(my_date, meteo$MO_TT, type = 'b', xlab = 'Zeit', ylab = 'Temperatur [°C]', main = 'Klimadate

# Minimumtemperatur in rot
lines(my_date, meteo$MO_TN, col = 'blue')

# Maximumtemperatur in blau
lines(my_date, meteo$MO_TX, col = 'red')

legend('topright', legend = c('Mittelwert', 'Minimum', 'Maximum'),
       col = c('black', 'red', 'blue'),
       pch = c(1, NA, NA),
       lty = 1)
```



Der Parameter `lty` steht für *line type* und die 1 bedeutet durchgezogene Linie. Mit `pch` legend wir die Art des Symbols fest; hier steht die 1 für das Standardsymbol "offener Kreis". Die Funktion `legend` hat viele Möglichkeiten und es lohnt sich, in die Hilfe zu sehen `?legend`.

5.1.2 Überblick über die wichtigsten *high-level* und *low-level* Grafikfunktionen

Die wichtigsten *high-level* Grafikfunktionen nach Ligges (2008), verändert:

Funktion	Beschreibung
`plot`	kontextabhängig – generische Funktion mit vielen Methoden
`barplot`	Säulendiagramm
`boxplot`	Boxplot
`contour`	Höhenlinien-Plot
`coplot`	Conditioning-Plots: Plots zweier Variablen aufgeteilt nach Werten einer dritten
`curve`	Funktionen zeichnen
`dotchart`	Dotplots (nach Cleveland)
`hist`	Histogramm
`image`	Bilder (3. Dimension als Farbe)
`mosaicplot`	Mosaikplots (kategorielle Daten)
`pairs`	Streudiagramm-Matrix
`persp`	perspektivische Flächen
`qqnorm` und `qqplot`	QQ–Plot

Die wichtigsten *low-level* Grafikfunktionen nach Ligges (2008), verändert:

Funktion	Beschreibung
`abline`	Fügt eine Linie hinzu; diese kann horizontal, vertikal oder über Steigung und Achsenabschnitt definiert werden
`arrows`	Pfeile
`axis`	Achsen
`grid`	Gitternetz
`legend`	Legende
`lines`	Linien (schrittweise)
`mtext`	Text in den Rändern
`plot.new`	Grafik initialisieren
`plot.window`	Koordinatensystem initialisieren
`points`	Punkte
`polygon`	(ausgefüllte) Polygone
`pretty`	berechnet "hübsche" Einteilung der Achsen
`segments`	Linien (vektorwertig)
`text`	Text
`title`	Beschriftung

5.2 Tuning mit par

Zur Vertiefung dieses Kapitels, empfehle ich Ligges (2008), Kapitel 8.1.3.

Die Grafikebene in R ist aufgeteilt in drei Regionen (Abbildung 5.1) und hat innere und äußere Ränder. Die Ränder werden von unten im Gegenuhrzeigersinn durchnummeriert.

Mit der Funktion `par` lassen sich sehr viele Einstellung der Grafik verändern. Viele Einstellungen übergibt die Funktion `plot` selbständig an `par`, zu.B. `log` (Logarithmieren der Achsen), `cex` (Größe eines Punkts) oder `col` (Farbe). Andere können aber nur durch Aufrufen der Funktion `par` verändert werden. Dazu gehören die inneren Ränder `mar` und die äußeren Ränder `oma`, die Aufteilung der Grafikebene mit `mfrow` oder `mfcol`.

Richtige Benutzung von `par`:

- Parameter setzen: `op <- par( ... )`
- plotten
- Parameter auf Standard zurück setzen: `par(op)`

Die Zuweisung `op <- par( ... )` speichert die Standardeinstellungen im Objekt `par`, bevor Sie sie ändern. Der Aufruf `par(op)` setzt Ihre Änderungen zurück. Das ist sehr praktisch, wenn Sie z.B. die Aufteilung der Grafikebene nicht mehr benötigen. Wenn Sie die Parameter nicht zurücksetzen, bleiben diese bestehen, bis das Grafikfenster geschlossen wird (z.B. mit `dev.off()`).

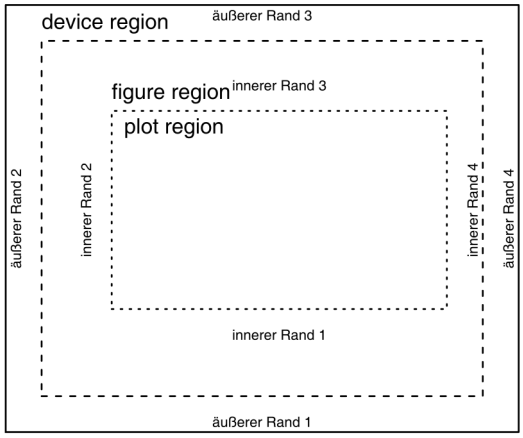


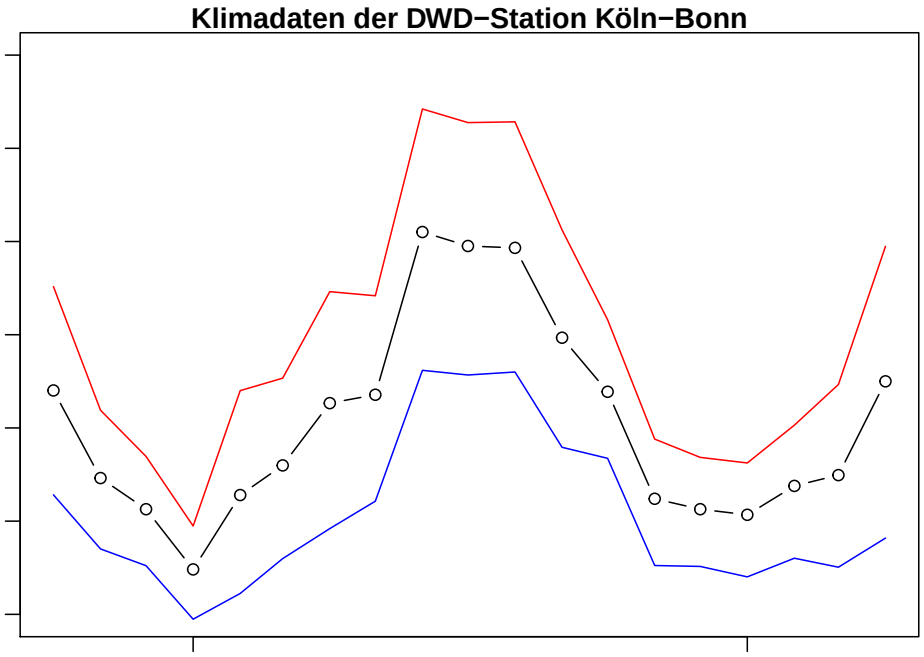
Figure 5.1: Aufteilung der Grafikfläche [@Ligges2008].

Um die Ränder zu verändern, rufen wir `par` auf und beschneiden die Ränder, damit Sie den Unterschied erkennen können.

```
op <- par(mar = c(1, 1, 1, 1))
plot(my_date, meteo$MO_TT, type = 'b', xlab = 'Zeit', ylab = 'Temperatur [°C]', main =

# Minimumtemperatur in rot
lines(my_date, meteo$MO_TN, col = 'blue')

# Maximumtemperatur in blau
lines(my_date, meteo$MO_TX, col = 'red')
```



Die Achsenbeschriftungen und die Zahlen haben jetzt nicht mehr genug Platz und verschwinden. Die Größe der Ränder wird in Zeilen angegeben, ist also relativ zur Gesamtgröße. Die Standardeinstellung ist `c(5, 4, 4, 2) + 0.1`.

Einige häufig genutzte Argumente in Grafikfunktionen und in `par` (nach Ligges, 2008, verändert). Schlagen Sie die Erklärungen dazu immer in `?par` oder `?plot` nach.

Funktion	Beschreibung
`axes`	Achsen sollen (nicht) eingezeichnet werden
`bg`	Hintergrundfarbe
`cex`	Größe eines Punktes bzw. Buchstaben
`col`	Farben
`las`	Ausrichtung der Achsenbeschriftung
`log`	Logarithmierte Darstellung
`lty`, `lwd`	Linientyp (gestrichelt, …) und Linienbreite
`main`	Überschrift
`mar`	Größe der inneren Ränder für Achsenbeschriftung etc.
`mfcol`, `mfrow`	mehrere Grafiken in einem Bild
`pch`	Symbol für einen Punkt
`type`	Typ (l für Linie, p für Punkt, b für beides, n für nichts)
`usr`	Ausmaße der Achsen auslesen
`xlab`, `ylab`	x-/y-Achsenbeschriftung
`xlim`, `ylim`	zu plottender Bereich in x-/y- Richtung

Funktion	Beschreibung
`xpd`	in die Ränder hinein zeichnen

5.3 Inhalt der live Einführung

- `plot`, `barplot`, `mfrow`
- Aufgaben 7.3.1, 7.3.2 und 7.3.3.
- Speichern als pdf

Chapter 6

Reproduzierbare Forschung

- Wichtigkeit der Reproduzierbarkeit erklären
- Begriff *literate programming* definieren
- Aufbau einer RMarkdown-Datei erklären
- Einen einfachen ersten reproduzierbaren Bericht schreiben

6.1 Warum Reproduzierbarkeit in der Forschung wichtig ist

6.2 *Literate Programming* Idee von Donald Knuth

Die Idee, dass man den Code und die dazugehörige Interpretation (Text, Bericht etc.) nicht von einander trennen sollte, geht auf Knuth (1984) zurück. Mit *Literate Programming* meinte Knuth, dass Programme auch nichts anderes wie literatirsche Werke sind. Er setzte den Fokus darauf, mit Programmen menschlichen Benutzern zu erklären, was man den Computer machen lassen möchte. Also weg vom computer- hin zum menschzentrierten Zugang. So wird Programmieren und in unserem Fall die Datenanalyse verständlich und vor allem reproduzierbar.

Leider ist es in unserer modernen Forschungslandschaft immer noch nicht Standard. Das Trennen von Analyseergebnissen und Berichten (Forschungsartikeln) sorgt für viele (unendeckte und unnötige) Fehler und Frust.

6.3 Reproduzierbare Berichte mit R Markdown

R hat sein eigenes System von reproduzierbaren Berichten, genannt R Markdown (Xie et al., 2018). Es ist benutzerfreundlich und ermöglicht unterschiedliche Formate von Berichten, wie HTML-Dokumente, PDF-Dateien, Präsentationsfolien usw.

Es wird Sie vielleicht überraschen, aber das Skript, das Sie gerade lesen ist nichts anderes als ein "literarisch" programmierter Bericht in R Bookdown (Xie, 2016), einem R-Paket speziell für lange R Markdown-Dokumente.

Wir werden vor allem mit R Notebooks arbeiten, die eine gute Interaktion zwischen dem geschriebenen Text und dem R-Code ermöglichen. Das Notebook kann sowohl in ein HTML-Dokument als auch in PDF oder Word als endgültiges Berichtdokument umgewandlet werden. Diesen Prozess nennt man *knit* (der Knopf in RStudio mit dem Wollknäuel).

6.4 Wichtigste Regeln für Reproduzierbarkeit

6.5 Weiterführende Videos und Literatur

Die Playlist zu *Reporducible Research* finden Sie hier.

Report Writing for Data Science in R (Peng, 2019) (auf ILIAS)

6.6 Inhalt der live Einführung

- Erstellen eines einfachen R Notebooks
- R-Code Chunks
- Einfache Layoutelemente: Überschriften, Listen, fett und kursiv

Chapter 7

Aufgabensammlung

7.1 Erste Schritte

7.1.1 Ars Haushaltsbuch

Der angehende Datenanalyst Ar Stat möchte dem Rat seiner Mutter folgen und ein Haushaltsbuch anlegen. Als erstes möchte er sich einen Überblick über seine Ausgaben in der Uni-Mensa verschaffen und erstellt die folgende Tabelle:

1. Wie viel hat Ar insgesamt in der Woche ausgegeben?
2. Wie viel hat er im Schnitt pro Tag ausgegeben?
3. Wie stark schwanken seine Ausgaben?

Leider hat Ar sich beim übertragen der Daten vertippt. Er hat am Dienstag seine Freundin zum Essen eingeladen und 7,95 € statt 2,90 € ausgegeben.

4. Korrigieren Sie Ars Fehler.
5. Wie verändern sich die Ergebnisse aus den Teilaufgaben 1 bis 3 Warum?

Table 7.1: Ars Mensaausgaben

Wochentag	Ausgaben
Montag	2,57
Dienstag	2,90
Mittwoch	2,73
Donnerstag	3,23
Freitag	3,90

7.2 Daten in R

7.2.1 Bestandesaufnahme im Wald

Ar Stat arbeitet als HiWi in der AG Ökosystemforschung und soll im Nationalpark Eifel eine Bestandsaufnahme durchführen (d.h. Baumhöhen und -durchmesser vermessen). Er notiert den BHD (Brusthöhendurchmesser) und die Art der Bäume.

1. Lesen Sie den Datensatz `BHD.txt` ein und ordnen Sie ihn der Variable `BHD` zu.
2. Erstellen Sie einen Vektor `a` mit Baumnummern. Von welcher Art sind die Elemente des Vektors `a`?
3. Fügen Sie die Datensätze `BHD` und `a` zu einem `data.frame` zusammen und benennen Sie die Spalten sinnvoll.
4. Löschen Sie den Vektor `a`.
5. Lesen Sie den Datensatz `Art.txt` ein und ordnen Sie ihn der Variablen `art` zu.
6. Fügen Sie die Art in den `data.frame` ein.
7. Erstellen Sie eine Tabelle mit der Anzahl der jeweiligen Arten. Nutzen Sie die Funktion `table`.
8. Speichern Sie die Tabelle mit `write.table`.

7.3 Daten visualisieren, Teil I: Fokus auf R

7.3.1 Wahlbeteiligung bei der Bundestagswahl 2017

Bauen Sie die Grafiken aus der Einführung nach (Abbildung 7.1).

1. Lesen Sie den Datensatz `Wahlbeteiligung.csv` in R ein und ordnen Sie ihn dem Objekt `bet` zu. Der Datensatz hat einen *header* und haben einen Strichpunkt als Spaltentrenner.
2. Sehen Sie sich die Struktur und die ersten und letzten 6 Zeilen des Datensatzes an.
3. Stellen Sie die Wahlbeteiligung als Funktion der Zeit in einem Streudiagramm dar. Wählen Sie die passende Darstellungsform `type`.
4. Beschriften Sie die Grafik.
5. Speichern Sie die Grafik als pdf ab.

7.3.2 Zweitstimme bei der Bundestagswahl 2017

Bauen Sie die Grafiken aus der Einführung nach (Abbildung 7.2).

1. Lesen Sie den Datensatz `Zweitstimme.csv` in R ein und ordnen Sie ihn dem Objekt `zweit` zu. Der Datensatz hat einen *header* und haben einen Strichpunkt als Spaltentrenner.

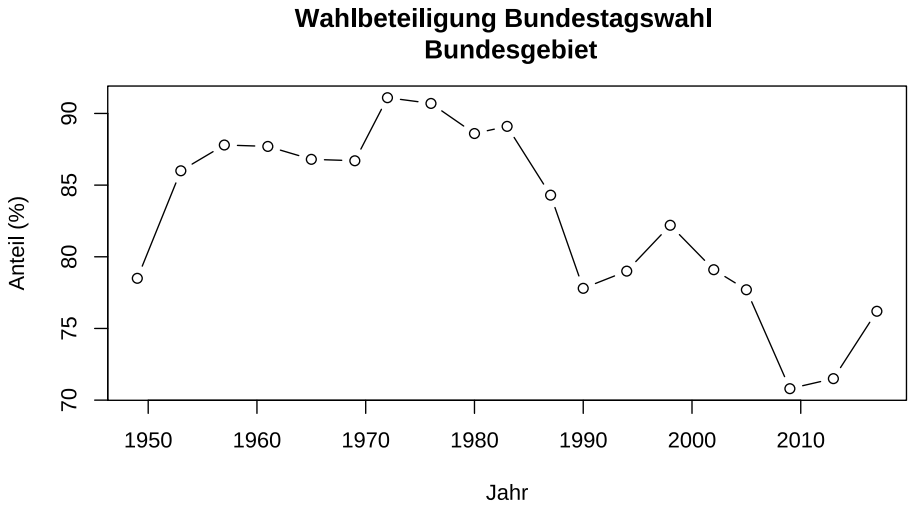


Figure 7.1: Wahlbeteiligung bei den Bundestagswahlen. Quelle: Der Bundeswahlleiter.

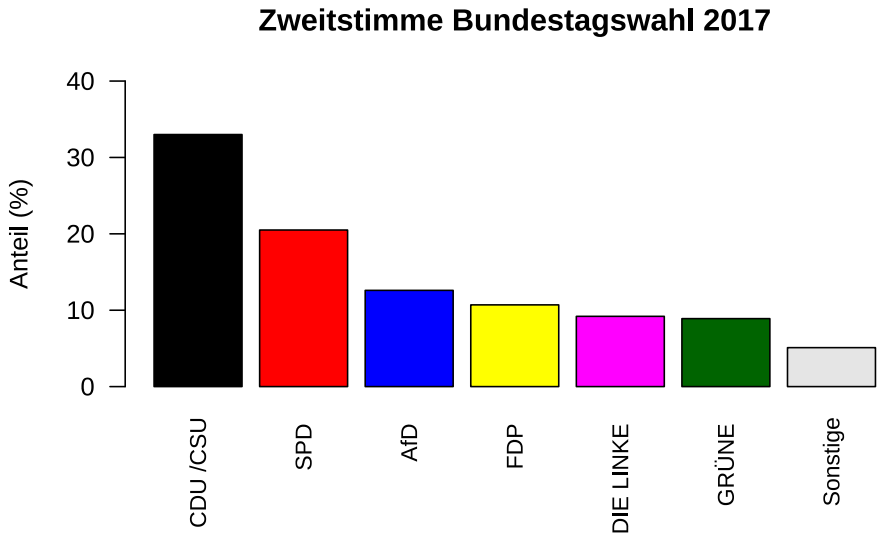


Figure 7.2: Zweitstimme bei der Bundestagswahl 2017. Quelle: Der Bundeswahlleiter.

2. Sehen Sie sich die Struktur und die ersten und letzten 6 Zeilen des Datensatzes an.
3. Stellen Sie die die Zweitstimmen pro Partei in einem Säulendiagramm dar. Sortieren Sie die Zweitstimmen in absteigender Reihenfolge.
4. Beschriften Sie die Grafik.
5. Speichern Sie die Grafik als pdf ab.

7.3.3 Ergebnisse der Bundestagswahl in einer Grafik

Stellen Sie beide Grafiken nebeneinander dar wie in Abbildung (7.3.3) gezeigt.

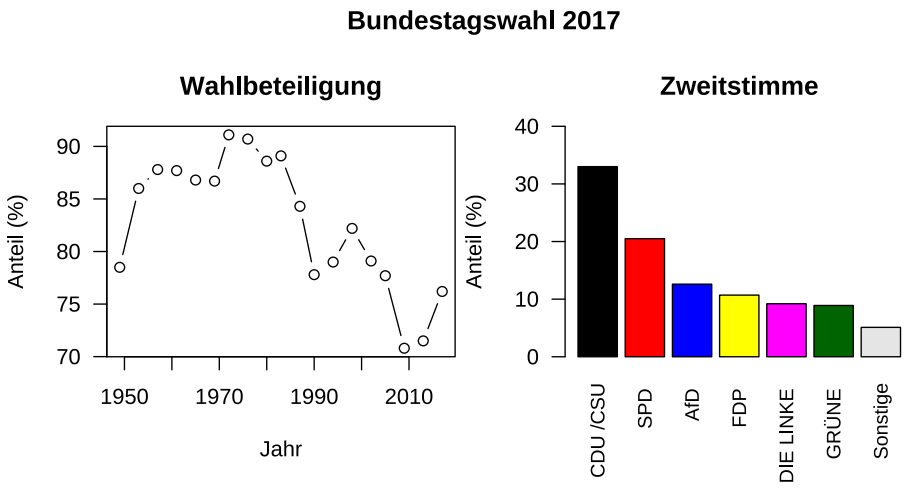


Figure 7.3: Ergebnisse der Bundestagswahl 2017. Quelle: Der Bundeswahlleiter.

7.3.4 Einen zu großen weißen Rand vermeiden

Bei Berichten haben Abbildungen meistens keine Überschrift, da alles in der Bildunterschrift erklärt wird. Wenn man die Überschrift beim plotten weglässt, die Standardeinstellungen für die Ränder aber beibehält, entsteht ein zu großer weißer Rand um die Grafik. Diesen wollen wir nun abschalten.

1. Kopieren Sie den Code zum Plotten der Temperaturen aus dem Kapitel 5.
2. Stellen Sie oben und rechts einen Rand von 0.1 Zeilen ein.
3. Speichern Sie die Grafik als pdf ab.

7.3.5 Spielen mit der Funktion `par`

Setzen Sie die Übung 7.3.4 fort. Denken Sie an den richtigen Aufruf mit der Zuweisung von `op <- par( ... )`!

1. Probieren Sie die Größeneinstellung `cex = 2` in `plot` aus. Testen Sie unterschiedliche Werte.
2. Probieren Sie die Einstellungen `cex.axis`, `cex.lab` und `cex.main` in `par` aus.
3. Probieren Sie die Einstellung `col` in `plot` aus.
4. Probieren Sie die Einstellungen `col.axis`, `col.lab` und `col.main` aus.
5. Probieren Sie die Schrifteinstellungen aus. Dazu stellen Sie den Parameter `family` in `par` auf "serif", "sans" oder "mono".
6. Probieren Sie die Parameter `font.lab = 2` und `font.axis = 2` direkt in `plot` aus. Zahlen 1 bis 5 stehen jeweils für normal, fett, kursiv, fett-kursiv und symbolisch.

7.4 Reproduzierbare Berichte mir R Markdown

7.4.1 Erster eigener Bericht

Erstellen Sie ein R Notebook aus den Notizen der ersten 3 R-Sessions.

7.5 Eigene Funktionen schreiben

7.5.1 R-Hausaufgaben

An dem Kurs "Einführung in R" nehmen 49 Studierende teil. Der Leistungsnachweis besteht aus Hausaufgaben, die insgesamt mit 100 Punkten bewertet werden. Ab 50 Punkten gilt der Kurs als bestanden.

1. Lesen Sie den Datensatz `R-HAs.txt`, der die Endpunkte enthält, ein.
2. Ermitteln Sie, wie viele Teilnehmer bestanden und wie viele nicht bestanden haben.

7.5.2 Fledermäuse, die Zweite

Wir beschäftigen uns erneut mit den Fledermäusen.

1. Lesen Sie den korrigierten(!) Datensatz \texttt{Fledermaus_cor.txt} ein.
2. Schreiben Sie eine Funktion, die den Entwicklungsstand der Tiere klassifiziert. Nutzen Sie dazu die ad hoc Regel: Individuum < 5 cm ist ein Jungtier, sonst erwachsen.
3. Erstellen Sie eine ordinal-skalierte Variable `alter` mit dem Entwicklungsstand der Tiere.
4. Schreiben Sie eine Funktion, die die Mittelwerte der Größe für weibliche und männliche Individuen berechnet.
5. Berechnen Sie die Mittelwerte der Größe und runden Sie auf 2 Nachkommastellen.

7.5.3 Unfaire Klausur?

Ar belegt im 4. Semester die Veranstaltung "Spaß mit R". Bei der Klausur gibt es 2 Aufgabengruppen mit jeweils 60 Punkten. Aufgabengruppe 1 wird an Studierende auf ungeraden Sitzplätzen und Aufgabengruppe 2 an Studierende auf geraden Sitzplätzen ausgegeben.

1. Lesen Sie den Datensatz `Klausurpunkte.txt` ein.
2. Überprüfen Sie Ars Vermutung, dass die Aufgabengruppe 1 im Schnitt leichter war als Aufgabengruppe 2 (d.h. in der Gruppe 1 im Schnitt mehr Punkte erzielt wurden).

7.6 Daten visualisieren, Teil II: Fokus auf Daten

7.6.1 Zeitreihen aus der Langen Bramke (Harz)

Im Harz wurden über eine längere Zeit Niederschlag, Abfluss und Temperatur gemessen.

1. Laden Sie den Datensatz `Data.dat`.
2. Stellen Sie die Temperatur in einem Streudiagramm dar. Welche Darstellungsart (Argument `type` in `plot`) erscheint Ihnen am sinnvollsten?
3. Beschriften Sie die Graphik und fügen Sie einen Titel hinzu.
4. Speichern Sie die Graphik als pdf ab.
5. Stellen Sie die Niederschläge in einem Diagramm dar. Wählen Sie einen geeigneten Darstellungstyp mit `type` (Tipp: geben Sie für die Hilfe `?plot` in die Console ein).

7.6.2 Temperatur-Datensatz

1. Laden Sie den Temperatur-Datensatz aus Zuur et al. (2009).
2. Berechnen Sie die Monatsmittelwerte für alle Stationen, sowie die Standardabweichungen.
3. Stellen Sie die Monatsmittel der Temperatur in einem Säulendiagramm dar.
4. Beschriften Sie die Graphik sinnvoll.
5. Fügen Sie die Standardabweichungen zu den einzelnen Balken hinzu.

7.6.3 Artenvielfalt in Grasländern

Sie erhalten Daten aus dem Grasland-Monitoring im Yellowstone Nationalpark und dem National Bison Range (USA). Das Ziel des Monitorings ist die Untersuchung möglicher Änderungen der Biodiversität und des Zusammenhang mit Umweltfaktoren. Biodiversität wurde durch die Anzahl unterschiedlicher Arten quantifiziert. Insgesamt haben die Forscher ca. 90 Arten in 8 Transekten kartiert. Die Aufnahmen wurden alle 4 bis 10 Jahre wiederholt. Insgesamt

liegen 58 Beobachtungen vor. Die Daten sind in der Datei `Vegetation2.xls` gespeichert.

1. Laden Sie den Datensatz in R und sehen Sie sich das Ergebnis genau mit `str`, `head` und `tail` an. Diese Aufgabe dient dazu, das Einlesen von Excel-Dateien zu erarbeiten. Tipp: eine mögliche Bibliothek, die dabei helfen kann, wäre `xlsx`.
2. Berechnen Sie den Mittelwert und die Standardabweichung der Artenzahl (Variable `R`) pro Transekt.
3. Plotten Sie die Artenzahl gegen die Variable `BARESOIL` (Anteil von unbewachsenem Boden).
4. Benutzen Sie unterschiedliche Symbole pro Transekt, erstellen Sie eine Legende.
5. Beschriften Sie die Graphik sinnvoll und speichern Sie sie als pdf ab, ohne die Maus zu benutzen.

7.6.4 Tracerversuche

Im Waldstein wurden Tracerversuche mit dem Farbstoff Brilliant Blue durchgeführt und die gefärbten Bodenprofile *binärisiert* (d.h. in ein schwarz-weiß Bild umgewandelt). Schwarze Pixels stellen gefärbten Boden und weiße ungefärbten dar. Aus diesen Binärbildern wurde anschließend eine Reihe von Kenngrößen berechnet.

1. Lesen Sie die Datei \texttt{Waldstein2005_ind.txt} ein. Die Tiefe eines Profils ist 579 Pixel und es liegen 6 Profile untereinander in der Spalte d.
2. Berechnen Sie die 5%, 50% und 95% Quantile des Färbeanteils (Index d) der 6 Profile.
3. Stellen Sie den Median des Anteils der Färbung mit der Tiefe dar und fügen Sie die Quantile als transparente Fläche hinzu (Tipp: `polygon`).

7.7 Effizientes Programmieren

7.7.1 Lagerungsdichten

Auf 10 verschiedenen landwirtschaftlichen Feldern wurden im Oberboden je 25 Stechzylinder entnommen.

1. Lesen Sie den Datensatz `Bodendaten.txt` ein.
2. Bestimmen Sie die mittlere Lagerungsdichte pro Feld.

7.7.2 Temperatur-Datensatz, revisited

1. Laden Sie den Temperatur-Datensatz aus Zuur et al. (2009), Datei `Temperatur.csv`.

2. Berechnen Sie die Jahresmittelwerte je Station. (Tipp: Hilfe von `tapply` genau lesen!)

Bibliography

Ihaka, R. and Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314.

Knuth, D. E. (1984). Literate Programming. *The Computer Journal*, 27(2):97–111.

Ligges, U. (2008). *Programmieren mit R*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Murrell, P. (2006). *R Graphics*. Computer Science and Data Analysis Series. Chapman & Hall/CRC, Boca Raton, Fla. OCLC: 255097201.

Peng, R. D. (2019). *Report Writing for Data Science in R*.

Xie, Y. (2016). *bookdown: Authoring Books and Technical Documents with R Markdown*. Chapman and Hall/CRC, Boca Raton, Florida. ISBN 978-1138700109.

Xie, Y., Allaire, J., and Grolemund, G. (2018). *R Markdown: The Definitive Guide*. Chapman and Hall/CRC, Boca Raton, Florida. ISBN 9781138359338.

Zuur, A. F., Ieno, E., and Meesters, E. (2009). *A Beginner's Guide to R*. Springer.