

Einführung in die Datenanalyse mit R (700104)

Kursskript



Dr. Johannes SIGNER
Abteilung Wildtierwissenschaften
Büsgenweg 3
37077 Göttingen

jsigner@uni-goettingen.de

Dr. Kai HUSMANN
Abteilung Forstökonomie und nachhaltige Landnutzungsplanung
Büsgenweg 1
37077 Göttingen

kai.husmann@uni-goettingen.de



Fakultät für Forstwissenschaften und Waldökologie
Georg-August-Universität Göttingen



¹³ Dieses Werk ist lizenziert unter einer [Creative Commons Namensnennung - Nicht-kommerziell - Weitergabe](#)
¹⁴ [unter gleichen Bedingungen 4.0 International Lizenz](#).

¹⁵ Zitiervorschlag:
¹⁶ Signer, J. und Husmann, K. (2023) Skript zur Vorlesung Einführung in die Datenanalyse mit R, Georg-
¹⁷ August-Universität Göttingen.

¹⁸ Letzte Aktualisierung: 8. November 2023

Vorwort und Danksagung

Lernziel des Kurses ist die Einführung in die Arbeit, Visualisierung und Analyse von (forstlichen) Datensätzen mit dem Statistikprogramm *R*. Der Schwerpunkt liegt dabei auf der Datenanalyse (Data Science). Statistische Methoden werden nur an wenigen Stellen exemplarisch angewendet. Weitere Materialien als dieses Kursskript und die Übungsaufgaben (StudIP) werden nicht benötigt. Die gelöste Übungsaufgaben dieses Skriptes werden Ihnen über StudIP zugänglich gemacht. Dort werden auch ggf. Ankündigungen bekanntgegeben.

Um die Credits für den Kurs zu erhalten, müssen Sie am Ende des Kurses eine mündliche Prüfung ablegen. Für die Prüfung werden Sie zwei zufällig gezogene Prüfungsfragen aus dem Dokument "Übungen: Einführung in die Datenanalyse mit R" bearbeiten und vortellen. Nach einer 15-minütigen Vorbereitungszeit beträgt die Prüfungszeit weitere 15 Minuten. In der Prüfungszeit präsentieren Sie zunächst Ihre Lösung und beantworten anschließend vertiefende Fragen zu Ihrer Lösung und daraufhin auch zum gesamten Lehrinhalt des Kurses.

Dieses Vorlesungsskript ist ein R Markdown-Dokument, das mit R und RStudio erstellt wurde. Das Dokument besteht aus Fließtext, R Code und den entsprechenden Code-Ergebnissen. Die grau hinterlegten Codepassagen sind kurze R-Skripte. Falls das Skript eine Konsolenausgabe erzeugt, ist diese direkt mit "###" markiert (diese Begriffe werden in Kapitel 1.2 näher erläutert).

Dank für Anmerkungen gilt Markus Benesch, Sofie Biberacher und Josephine Trisl. Teile des Unterkapitels zu Schleifen und Kontrollstrukturen ist an das R Skript des Kurses Computergestützte Datenanalyse von Robert Nuske, Nikolas von Lüpke und Joachim Saborowski angelehnt.

Inhaltsverzeichnis

38	Inhaltsverzeichnis	
39	1 R und RStudio	3
40	1.1 Installation von R und RStudio	3
41	1.2 Erste Schritte in R	3
42	1.3 Kommentare	5
43	2 Variablen, Funktionen und Datentypen	6
44	2.1 Variablen beim Programmieren	6
45	2.2 Datentypen	7
46	2.3 Funktionen	8
47	2.4 Datenstrukturen	8
48	2.5 Funktionen	10
49	3 Vektoren	11
50	3.1 Funktionen zum Arbeiten mit Vektoren	12
51	3.2 Statistische Funktionen	14
52	3.3 Beispiel Fotofallen	15
53	3.4 Arbeiten mit logischen Werten	16
54	3.5 Zugreifen auf Elemente eines Vektors (=Untermengen)	18
55	3.6 Der %in%-Operator	19
56	4 Faktoren (factors)	21
57	5 Spezielle Einträge	23
58	5.1 NA	23
59	5.2 NULL	24
60	5.3 Inf	24
61	6 data.frames oder Tabellen	26
62	6.1 Wichtige Funktionen zum Arbeiten mit data.frames	27
63	6.2 Zugreifen auf Elemente eines data.frame	27
64	7 Schreiben und lesen von Daten	31
65	7.1 Textdateien	31
66	8 Erstellen von Abbildungen	33
67	8.1 Base Plot	33
68	8.1.1 Mehrere Panels	37
69	8.1.2 Speichern von Abbildungen	38
70	8.2 Histogramme	39
71	8.3 Boxplots	42
72	8.4 ggplot2: Eine Alternative für Abbildungen	44
73	8.4.1 Multipanel Abbildungen	51
74	8.4.2 Plots kombinieren	52

75	8.4.3 Speichern von plots	55
76	9 Mit Daten arbeiten	56
77	9.1 dplyr eine Einführung	56
78	9.2 Arbeiten mit gruppierten Daten	59
79	9.3 pipes oder %>%	60
80	9.4 Joins	61
81	9.5 'long' and 'wide' Datenformate	63
82	9.6 Auswählen von Variablen	64
83	9.7 Einzelne Beobachtungen abfragen (slice())	66
84	9.8 Spalten trennen	69
85	10 Arbeiten mit Text	71
86	10.1 Arbeiten mit Text	71
87	10.2 Finden von Textmustern	72
88	11 Arbeiten mit Zeit	75
89	11.1 Arbeiten mit Zeitintervallen	76
90	11.2 Formatieren von Zeit	77
91	12 Aufgaben Wiederholen (for-Schleifen)	79
92	12.1 Schleifen	79
93	12.1.1 Wiederholen von Befehlen mit for().	79
94	12.1.2 Wiederholen von Befehlen mit while()	82
95	12.2 Bedingte Ausführung von Codeblöcken	82
96	13 (R)markdown	84
97	13.1 Markdown Grundlagen	84
98	13.2 R und Markdown	85
99	14 Räumliche Daten in R	87
100	14.1 Was sind räumliche Daten	87
101	14.2 Koordinatenbezugssystem	87
102	14.3 Vektordaten in R	87
103	14.4 Arbeiten mit Vektordaten	89
104	14.5 Rasterdaten in R	91
105	15 FAQs (Oft gefragtes)	96
106	15.1 Arbeiten mit Daten	96
107	15.1.1 Einlesen von Exceldateien	96
108	16 Zusätzliche Aufgaben	97
109	16.1 Arbeiten mit Daten	99
110	17 Literatur	101

1 R und RStudio

1.1 Installation von R und RStudio

Als ersten Schritt müssen Sie R und RStudio installieren. Dabei ist wichtig zu unterscheiden, dass R und RStudio zwei unterschiedliche Programme sind. R ist die eigentliche Programmiersprache mit der wir arbeiten. RStudio hingegen ist eine sogenannte Entwicklungsumgebung¹, die das Arbeiten mit R vereinfachen soll.

Sie können also mit R arbeiten ohne RStudio (was jedoch sehr unkomfortabel ist), aber nicht mit RStudio ohne R.

Gehen Sie für die Installation von R, auf die Website <https://cloud.r-project.org/> und laden Sie die für ihren Computer passende R-Version herunter und folgen Sie den Installationsanweisungen. In Linux können Sie R über die Kommandozeile installieren.

Für die Installation von RStudio gehen Sie zu der Website <https://posit.co/download/rstudio-desktop/#download> und laden die richtige Version für Ihr Betriebssystem herunter und folgen Sie den Installationsanweisungen.

1.2 Erste Schritte in R

RStudio bietet eine Vielzahl von Funktionen, die uns das Arbeiten mit R erleichtern können. Öffnen Sie RStudio. Sie erhalten eine leere Entwicklungsumgebung. Als erstes bietet es sich an, ein neues Skript zu erstellen. Gehen Sie dafür auf das Menü: **File** » **New File** » **R Script** oder klicken Sie die Tastenkombination *Strg* + *Umschalt* + *N* (**Strg** + **⇧** + **N**).

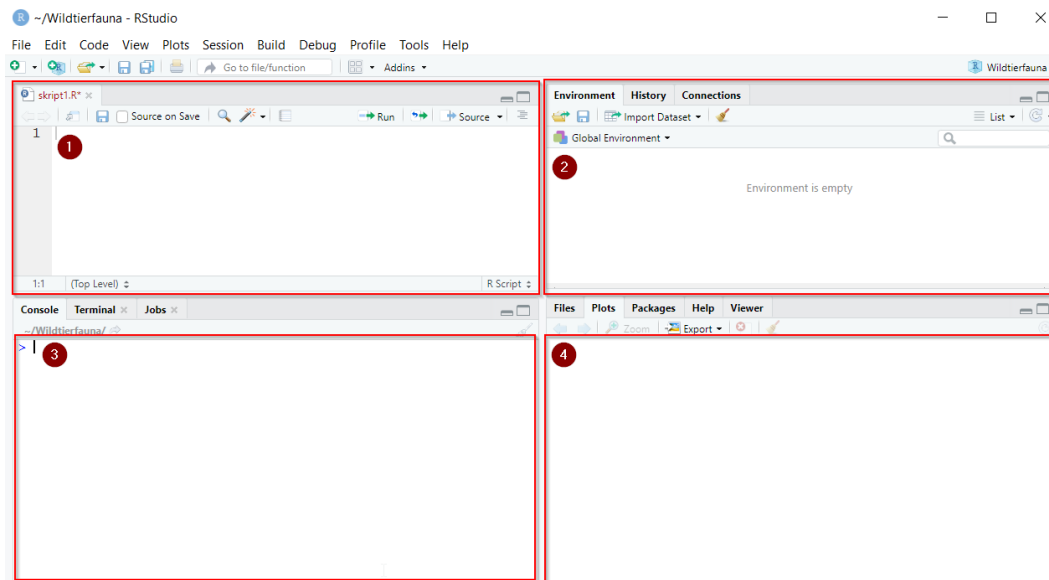


Abbildung 1: RStudio Panes.

RStudio besteht nun aus vier sogenannten Panes oder Ausschnitten (siehe auch Abbildung 1). Die Ausschnitte sind wie folgt gegliedert:

¹Oder auch IDE (=Integrated Development Environment) genannt.

1. Hier werden Skripte angezeigt, d.h., hier wird meist R Code geschrieben und dokumentiert.
2. Der zweite Ausschnitt erteilt Auskunft über den *Workspace*. Im Workspace werden alle verfügbaren Objekte angezeigt.
3. Die eigentliche R-Konsole wird in Ausschnitt 3 dargestellt. Hier wird in der Regel wenig Code eingegeben. Der normale Workflow ist vom Skript Code an die Konsole zu schicken.
4. Der vierte Ausschnitt enthält mehrere Reiter. Der Reiter *Files* zeigt den Verzeichnisbaum an. Im Reiter *Plots* werden Plots angezeigt, wenn diese im Code aufgerufen werden. Hilfeseiten zu Funktionen werden im Reiter *Help* angezeigt.

Einfache Rechenoperationen lassen sich in der R-Konsole durchführen.

```
10 + 5
```

```
## [1] 15
```

```
20 - 10
```

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

```
10 * 3
```

```
## [1] 30
```

```
100 / 19
```

```
## [1] 5.263158
```

Weitere häufig verwendete Operationen sind `^` für eine beliebige Potenz, z.B. $2^3 = 2^3 = 8$. Analog dazu gibt es die Funktion `sqrt()` zum berechnen von Wurzeln und viele weitere Funktionen.

Meist verwenden wir jedoch Skripte, um den R-Code zu schreiben und ihn dann an die Konsole “zu schicken”. Dies hat den Vorteil, dass alle Schritte nachvollziehbar bleiben und Analysen beliebig oft wiederholt werden können. Nach der Ausführung bleibt der Code erhalten. Stellen Sie sich den Code im R-Skript wie ein Kochbuch vor. Wenn wir R-Code in einem R-Skript geschrieben haben gibt es mehrere Möglichkeiten diesen Code auszuführen. Wir können entweder auf *Run* drücken (Abbildung 2) oder den Code mit der Tastenkombination *Strg + Enter* (`(Strg) + [] + [↵]`) an die Konsole schicken.

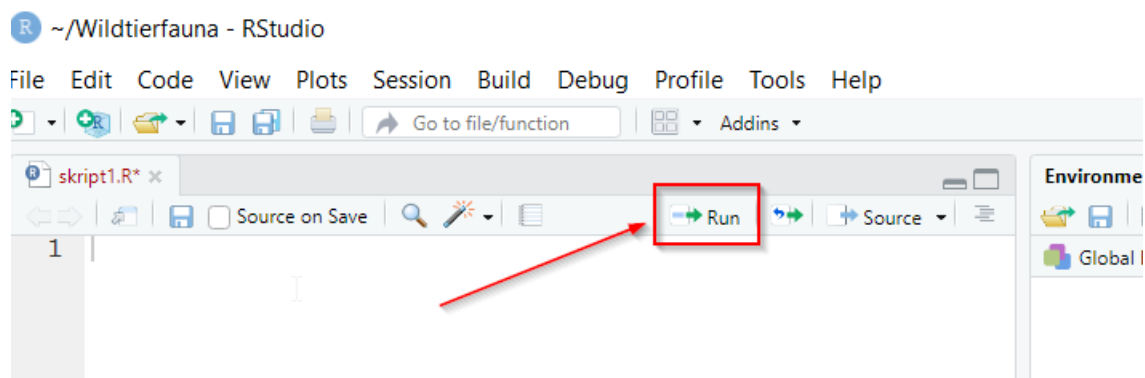


Abbildung 2: Ausführen von Code in RStudio..

1.3 Kommentare

Ein wichtiger Bestandteil für das Arbeiten mit Skripten in R sind Kommentare. Ein Kommentar ist Text in einem R-Skript, der der Dokumentation dient und von der R Konsole ignoriert wird. Sämtliche Zeilen, die mit dem Zeichen `#` beginnen, werden ignoriert.

```
# Berechnen der Quadratwurzel  
sqrt(81)
```

```
## [1] 9
```

Im vorherigen Codeblock wurde der Funktionsaufruf `sqrt(a)` auskommentiert. Die Zeile `# Berechnen der Quadratwurzel` wird bei der Ausführung ignoriert. Es empfiehlt sich, komplexere Abläufe zu kommentieren damit andere (z.B. Kommilitonen / Kommilitoninnen) verstehen, warum und wie etwas gemacht wurde. Ganz besonders gilt das jedoch für einen selbst. Oft sind in der Zukunft Zusammenhänge nicht mehr so klar, wie sie beim Schreiben des Codes waren.

Aufgabe 1: Ausführen von Quellcodes

Öffnen Sie RStudio, erstellen Sie ein neues Skript und speichern Sie dieses unter dem Namen `skript1.R` ab.

Tippen oder kopieren Sie folgenden Code in das Skript:

```
# Einfache Rechenoperationen  
1 + 3  
2^7  
  
# Einfache Funktion  
sqrt(20)
```

Führen Sie nun alle Zeilen aus.

2 Variablen, Funktionen und Datentypen

2.1 Variablen beim Programmieren

Ergebnisse aus Berechnungen (wie oben angeführt), aber auch aus deutlich komplexeren Operationen, werden in Variablen abgespeichert. Man kann sich eine Variable wie eine Hülle (oder bildlich gesprochen wie eine Schachtel) vorstellen, in die man etwas hinein legen kann und darauf zu einem späteren Zeitpunkt wieder zugreifen kann. Z.B. weist der folgende Ausdruck der Variable `alter` den Wert 102 zu.

```
alter = 102
```

Variablen können Objekte in R speichern. Ein Objekt, im einfachsten Fall ein einzelner Wert, kann mit der Anweisung `<-` einer Variablen zugewiesen werden. Der nachfolgende Code weist der Variable `a` den Wert 10 zu.

```
a <- 10
a
```

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

Man kann mit `=` oder `<-` einer Variable einen Wert zuweisen. Der Unterschied ist in den meisten Fällen vernachlässigbar, es wird aber allgemein empfohlen `<-` zu verwenden.

Wir können beliebige Variablen erstellen, z.B.

```
abc <- 10
name <- "Johannes"
```

Variablenamen dürfen nicht mit einer Zahl beginnen und müssen aus einem Wort bestehen.

- `a_123 <- 10` ist ok
- `123_a <- 10` erzeugt einen Fehler

Vorsicht: Groß- und Kleinschreibung muss beachtet werden

```
name <- "Johannes"
name
```

```
## [1] "Johannes"
```

Das Aufrufen der Variable

```
Name
```

führt zu einem Fehler.

Wir können dann mit den Werten, die in Variablen gespeichert sind, ganz normale Rechenoperationen durchführen.

```
a <- 10
b <- 5

a + b
```

189 ## [1] 15

```
b / a
```

190 ## [1] 0.5

```
a^b
```

191 ## [1] 1e+05

192 Das Ergebnis kann natürlich wieder in einer neuen Variable gespeichert werden.

```
ergebnis <- a + b
ergebnis
```

193 ## [1] 15

```
ergebnis2 <- ergebnis * 2
ergebnis2
```

194 ## [1] 30

195 Mit der Funktion `rm()` können Variablen, können nicht mehr benötigte Variablen, wieder gelöscht werden.

```
var1 <- "irgendwas"
exists("var1") # TRUE. also ja, eine Variable mit diesem Namen existiert
```

196 ## [1] TRUE

```
rm(var1)
exists("var1") # FALSE, also nein, eine Variable mit diesem Namen existiert nicht.
```

197 ## [1] FALSE

198 2.2 Datentypen

199 Es wurde bereits erwähnt, dass Variablen als Objekte gespeichert werden. Objekte sind Datenstrukturen, die
200 Objekte bzw. Messwerte aus der Realität abbilden. Wenn Sie beispielsweise eine Fotofalle in einem Wald
201 ausbringen, dann hat diese Fotofalle einen Namen (z.B. `Kamera1`) und nach einiger Zeit im Wald wurden
202 hoffentlich auch einige Fotos aufgenommen. Wir nehmen einmal an, dass nach drei Wochen 132 Fotos von
203 Rehen gemacht wurden.

204 Wir können jetzt sowohl den Namen der Fotofalle, als auch die Anzahl Fotos die aufgenommen wurden, in
205 zwei Variablen abspeichern.

```
kamera_name <- "Kamera_1"
anzahl_rehe <- 132
```

206 In den zwei vorherigen Zeilen Code haben wir zwei Objekte (im Sinne von R) erstellt. Das erste Objekt heißt
207 `kamera_name` und das zweite Objekt heißt `anzahl_rehe`. Auffällig ist, dass beide Objekte unterschiedliche
208 Datentypen haben. `kamera_name` ist vom Typ `character` (also Text). Das zweite Objekt, `anzahl_rehe`, ist
209 vom Typ `numeric` (also eine Zahl, wir unterscheiden hier nicht weiter²). Zusätzlich zu diesen zwei Typen

²Für Interessierte, man kann natürlich zwischen Ganzzahlen (`int`) und Gleitkommazahlen (`double`) unterscheiden.

(`character` und `numeric`) gibt es noch einen weiteren wichtigen Typ: nämlich das logische Wahr oder Falsch (in R: `TRUE` und `FALSE`). Zurückkommend auf das Beispiel mit den Fotofallen, könnte eine mögliche Fragestellung sein, ob auf einem der Fotos ein Fuchs gesehen wurde oder nicht. Dazu würden wir wieder eine neue Variable `fuchs_gesehen` anlegen und diese auf `TRUE` setzen, da ein Fuchs gesehen wurde.

```
fuchs_gesehen <- TRUE
```

2.3 Funktionen

Ein zweiter wesentlicher Bestandteil beim Arbeiten mit R sind Funktionen. Während eine Variable etwas speichert tut eine Funktion etwas. Beispielsweise zieht die Funktion `sqrt()` die Quadratwurzel aus einer Zahl.

```
sqrt(a)
```

```
## [1] 3.162278
```

Funktionen sind in R in der Regel daran zu erkennen, dass sie immer mit dem Funktionsnamen, gefolgt von runden Klammern (`()`), aufgerufen werden. Im vorherigen Beispiel wurde die Funktion mit dem Namen `sqrt()` aufgerufen. Das Objekt `a` haben wir bereits vorhin definiert (zur Erinnerung `a <- 10`). Die Funktion `sqrt()` arbeitet jetzt mit dem Objekt `a`, das in diesem Zusammenhang auch **Argument** genannt wird.

Argumente von Funktionen haben Namen, diese müssen jedoch nicht angegeben werden, wenn die Reihenfolge der Argumente berücksichtigt wird. Im vorherigen Beispiel, haben wir einfach die Funktion `sqrt(a)` aufgerufen und keinen Argumentnamen angegeben. Aus den Hilfeseiten für die Funktion `sqrt()` (siehe auch nachfolgender Abschnitt) ist zu entnehmen, dass die Funktion `sqrt()` ein Argument mit dem Namen `x` hat. Das heißt, der vollständige Aufruf der Funktion `x` wäre.

```
sqrt(x = a)
```

```
## [1] 3.162278
```

Um mehr über eine Funktion zu erfahren (z.B. die Bedeutung von Argumenten zu verstehen oder herauszufinden, was eine Funktion zurück gibt), können wir die Hilfeseiten konsultieren. Es gibt unterschiedliche Wege, um zu einer Hilfeseite zu gelangen.

1. In die Konsole `?<Name der Funktion>` tippen. Also, wenn wir Hilfe für die Funktion `mean()` möchten, könnten wir einfach `?mean` in die Konsole tippen.
2. Analog zu 1), können wir mit der Funktion `help` die Hilfeseite für eine andere Funktion aufrufen (z.B. wenn wir wieder die Hilfe für die Funktion `mean()` lesen möchten, dann könnten wir auch `help(mean)` in die Konsole tippen).
3. In RStudio kann man auch auf das **Help**-Panel klicken und dann einfach eine Funktion suchen (siehe Abbildung 1).
4. Wenn der Cursor in RStudio auf einer Funktion ist, kann man mit der Taste `F1` die dazugehörige Hilfeseite aufrufen.

2.4 Datenstrukturen

Verfolgen wir das Beispiel mit den Fotofallen etwas weiter. Es handelt sich um ein systematisches Monitoring. D.h. es wurde nicht nur eine Fotofalle ausgebracht, sondern insgesamt 15 Stück. Dieser Umstand erfordert

eine komplexere Datenstruktur. Nachfolgend sind die Anzahl Rehphotos für jede der 15 Fotofallen aufgeführt: 132, 79, 129, 91, 138, 144, 55, 103, 139, 105, 96, 146, 95, 118, 107.

Die Frage, die sich jetzt stellt, ist: Wie kann man diese Daten sinnvoll organisieren? Zusätzlich zur Anzahl der fotografierten Rehe soll jede Fotofalle eine eindeutige ID haben (`Kamera_1`, ..., `Kamera_15`) und wir wissen, dass jeweils 5 Fotofallen in drei unterschiedlichen Revieren aufgestellt waren (Fotofalle 1 bis 5 in Revier A, Fotofalle 6 bis 10 in Revier B und Fotofalle 11 bis 15 in Revier C).

Ein erster, wenn auch nicht wirklich sinnvoller, Ansatz könnte sein, dass wir für jede Fotofalle drei Objekte erstellen:

```
# 1. Kamera
name1 <- "Kamera_1"
anzahl_rehe1 <- 132
revier_1 <- "Revier A"

# 2. Kamera
name2 <- "Kamera_2"
anzahl_rehe2 <- 79
revier_2 <- "Revier A"

# usw.
```

Wenn wir so vorgehen würden, hätten wir 45 Objekte. Dieser Ansatz ist nicht sehr effizient und führt schnell zu einem unübersichtlichen *Workspace*³. Wir werden im Verlauf unterschiedliche Datenstrukturen kennenlernen.

Aufgabe 2: Variablen

Verwenden Sie die folgenden Daten

```
a <- 2
b <- "100"
p <- FALSE
```

und berechnen sie:

- `10 * a`
- `a / 144` und speichern Sie das Ergebnis in einer neuen Variablen `e` zwischen.
- Was ist das Ergebnis von `a + b`?
- Was ist das Ergebnis von `a + p`?

```
10 * a
e <- a / 144
a + b
a + p
```

³Als *Workspace* werden alle Objekte bezeichnet, die in einer R-Session zur Verfügung stehen.

2.5 Funktionen

Ein zweiter wesentlicher Bestandteil beim Arbeiten mit R sind Funktionen. Während eine Variable etwas *speichert* tut eine Funktion etwas. Beispielsweise zieht die Funktion `sqrt()` die Quadratwurzel aus einer Zahl.

```
sqrt(a)
```

```
## [1] 1.414214
```

Funktionen sind in R in der Regel daran zu erkennen, dass sie immer mit dem Funktionsnamen, gefolgt von runden Klammern (`()`), aufgerufen werden. Im vorherigen Beispiel wurde die Funktion mit dem Namen `sqrt()` aufgerufen. Das Objekt `a` haben wir bereits vorhin definiert (zur Erinnerung `a <- 10`). Die Funktion `sqrt()` arbeitet jetzt mit dem Objekt `a`, das in diesem Zusammenhang auch **Argument** genannt wird.

Argumente von Funktionen haben Namen, diese müssen jedoch nicht angegeben werden, wenn die Reihenfolge der Argumente berücksichtigt wird. Im vorherigen Beispiel, haben wir einfach die Funktion `sqrt(a)` aufgerufen und keinen Argumentnamen angegeben. Aus den Hilfeseiten für die Funktion `sqrt()` (siehe auch nachfolgender Abschnitt) ist zu entnehmen, dass die Funktion `sqrt()` ein Argument mit dem Namen `x` hat. Das heißt, der vollständige Aufruf der Funktion `x` wäre.

```
sqrt(x = a)
```

```
## [1] 1.414214
```

Um mehr über eine Funktion zu erfahren (z.B. die Bedeutung von Argumenten zu verstehen oder herauszufinden, was eine Funktion zurück gibt), können wir die Hilfeseiten konsultieren. Es gibt unterschiedliche Wege, um zu einer Hilfeseite zu gelangen.

1. In die Konsole `?<Name der Funktion>` tippen. Also, wenn wir Hilfe für die Funktion `mean()` möchten, könnten wir einfach `?mean` in die Konsole tippen.
2. Analog zu 1), können wir mit der Funktion `help` die Hilfeseite für eine andere Funktion aufrufen (z.B. wenn wir wieder die Hilfe für die Funktion `mean()` lesen möchten, dann könnten wir auch `help(mean)` in die Konsole tippen).
3. In RStudio kann man auch auf das **Help**-Panel klicken und dann einfach eine Funktion suchen (siehe Abbildung 1).
4. Wenn der Cursor in RStudio auf einer Funktion ist, kann man mit der Taste F1 die dazugehörige Hilfeseite aufrufen.

3 Vektoren

Die gute Nachricht zuerst, Sie haben bereits Vektoren erstellt in R (und dies wahrscheinlich nicht bewusst wahrgenommen). Wenn Sie nämlich eine Variable erstellen (z.B., `a <- 10`), wird ein Vektor der Länge eins erstellt, das heißt der Vektor enthält genau ein Element.

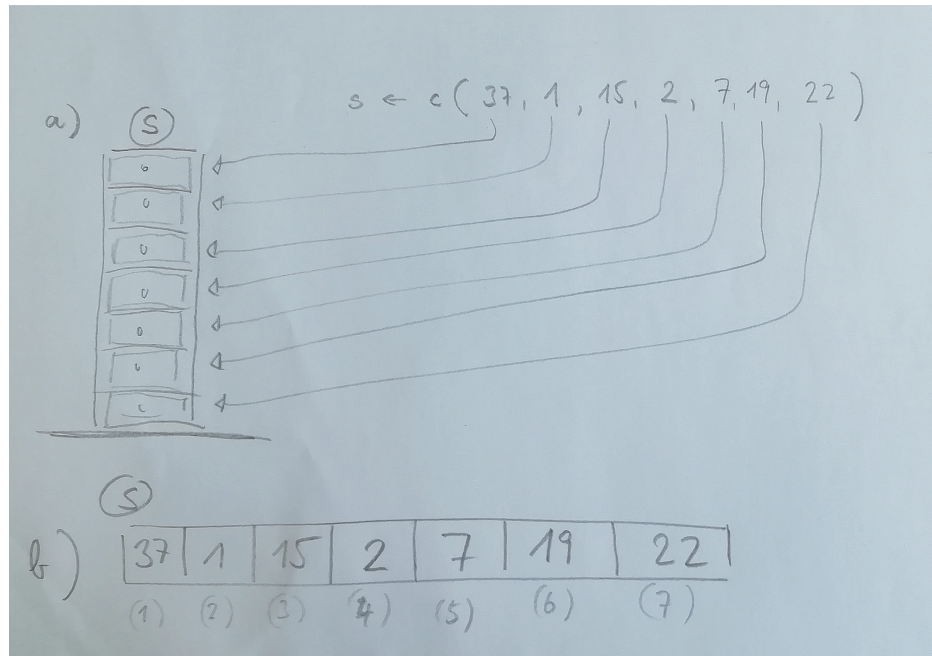


Abbildung 3: Schematische Darstellung eines Vektors in R.

Sie können sich Vektoren wie einen Schubladenschrank vorstellen (siehe auch Abbildung 3). Wichtig ist dabei, dass man in jede Schublade immer nur ein Element vom gleichen Typ verstauen kann. Etwas allgemeiner gesprochen heißt das, dass alle Elemente eines Vektors vom gleichen Datentyp sein müssen.

Es gibt zahlreiche Funktionen zum Erstellen von Vektoren (einige davon werden wir im weiteren Verlauf des Moduls kennenlernen). Die wohl wichtigste Funktion ist `c()`. Dabei steht `c` für *combine* oder *concatenate*. Die Funktion `c()` fügt einzelne Elemente in einen Vektor zusammen (und zwar genau in der Reihenfolge wie diese Elemente an `c()` übergeben werden).

Gehen wir nochmals zurück zu Abbildung 3, da wird schematisch dargestellt wie ein Vektor `s` mit 7 Elementen (in diesem Fall Zahlen) erstellt werden kann.

```
s <- c(37, 1, 15, 2, 7, 19, 22)
```

Die Funktion `c()` ordnet jetzt bildlich gesprochen die Zahl 37 der ersten Schublade zu, die Zahl 1 der zweiten Schublade und so weiter. Wenn Sie jetzt einfach `s` in die Konsole tippen, können Sie alle Elemente von `s` sehen:

```
s
```

```
## [1] 37  1 15  2  7 19 22
```

In Abbildung 3b wird der Vektor `s` nochmals systematisch dargestellt. Dabei sieht man, dass 37 an der ersten

Position des Vektors gespeichert wird und 22 an der letzten Position des Vektors gespeichert wird.

Die Grundrechenarten (+, -, /, *) und viele andere Funktionen funktionieren genau gleich mit Vektoren deren Länge > 1 ist. Sie werden elementweise durchgeführt. Wir können beispielsweise zu jedem Element von `s` 10 addieren

```
s + 10
```

```
## [1] 47 11 25 12 17 29 32
```

oder `s` mit sich selbst multiplizieren.

```
s * s
```

```
## [1] 1369    1  225    4   49  361  484
```

Neben der Funktion `c()` gibt es zahlreiche weitere Funktionen, um Vektoren zu erstellen. Sehr häufig braucht man Vektoren von Zahlenfolgen, solche Vektoren können mit der Funktion `seq` erstellt werden. Im einfachsten Fall benötigt `seq()` zwei Argumente: `from` und `to`⁴.

```
seq(from = 1, to = 10)
```

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

Man kann dann auch noch die Schritte angeben, mit denen erhöht wird.

```
seq(from = 1, to = 10, by = 2)
```

```
## [1] 1 3 5 7 9
```

Aufgabe 3: Vektoren erstellen

Sie haben den BHD (Brusthöhendurchmesser) in cm von vier Bäumen gemessen: 13, 15.3, 23, 9

- Erstellen Sie einen Vektor mit dem Namen `bhd` in dem Sie die Werte speichern
- Transformieren sie die BHD-Werte in mm.
- Berechnen Sie die Fläche des BHD in cm^2 (nehmen Sie dafür an, dass ein Baum kreisrund ist).

3.1 Funktionen zum Arbeiten mit Vektoren

Die Funktionen `head()` und `tail()` geben die ersten bzw. letzten `n` Elemente eines Vektors zurück. `n` hat einen voreingestellten Wert von 6, dieser kann natürlich angepasst werden.

```
head(s)
```

```
## [1] 37 1 15 2 7 19
```

```
head(s, n = 3)
```

```
## [1] 37 1 15
```

⁴Weil solche Vektoren so häufig vorkommen gibt es hier eine Abkürzung. Man kann `seq(from, to, by = 1)` mit `from:to` abkürzen. Also `1:10` würde auch alle Zahlen von 1 bis 10 zurückgeben.

```
tail(s, n = 2)
```

```
## [1] 19 22
```

Die Funktion `length()` gibt die Länge eines Vektors wieder.

```
length(s)
```

```
## [1] 7
```

Der Typ der Elemente eines Vektors kann mit der Funktion `class` abgefragt werden:

```
class(s)
```

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

Die eindeutigen Elemente eines Vektors können mit der Funktion `unique()` abgefragt werden.

```
unique(s)
```

```
## [1] 37 1 15 2 7 19 22
```

Mit der Funktion `table` kann die Häufigkeit verschiedener Elemente abgefragt werden.

```
table(s)
```

```
## s
```

```
## 1 2 7 15 19 22 37
```

```
## 1 1 1 1 1 1 1
```

Schlussendlich kann man mit der Funktion `sort()` und `rev()` die Position von Elementen in einem Vektor ändern. Die Funktion `rev` dreht die Elemente einmal um

```
rev(s)
```

```
## [1] 22 19 7 2 15 1 37
```

während `sort()` einen Vektor nach seinen Elementen sortiert⁵.

```
sort(s)
```

```
## [1] 1 2 7 15 19 22 37
```

Die Funktion `rep()` wiederholt einen Vektor.

```
rep(s, times = 2)
```

```
## [1] 37 1 15 2 7 19 22 37 1 15 2 7 19 22
```

Anstelle des Arguments `times` kann auch das Argument `each` verwendet werden. Der Unterschied liegt darin, dass `times` den gesamten Vektor `times`-Mal wiederholt und `each` jedes Element.

```
a <- 1:4
```

```
rep(a, times = 2)
```

⁵Auch für `sort()` gibt es ein zusätzliches Argument, das es ermöglicht die Elemente in absteigender Reihenfolge zu sortieren. Schauen Sie sich dazu, und auch für weitere Funktionen, die Hilfeseiten an.


```
351 ## [1] 1 2 3 4 1 2 3 4
      rep(a, each = 2)
```

```
352 ## [1] 1 1 2 2 3 3 4 4
```

```
353
```

354 *Aufgabe 4: Arbeiten mit Vektoren*

355 Es liegen jeweils zwei BHD-Messungen von vier Bäumen vor:

```
bhd <- c(32, 33, 23, 21, 21, 27, 18, 12)
```

357 Diese wurden immer abwechselnd mit zwei unterschiedlichen Messgeräten durchgeführt wurden.

358 Erstellen Sie einen Vektor von der Länge 8 mit den Einträgen die immer abwechselnd G1 und G2 sind und für
359 die zwei Geräte stehen.

360 3.2 Statistische Funktionen

361 Zahlreiche statistische Funktionen können auf Vektoren angewendet werden, hier sind nur die wichtigsten
362 aufgeführt: `mean()` berechnet den Mittelwert, `median()` berechnet den Median und `sd()` die Standardabweichung.
363

```
mean(s)
```

```
364 ## [1] 14.71429
```

```
median(s)
```

```
365 ## [1] 15
```

```
sd(s)
```

```
366 ## [1] 12.76341
```

367 Eine weitere sehr häufig verwendete Funktion ist `sample()`. Mit `sample()` werden `size` Elemente zufällig
368 aus einem Vektor, mit oder ohne Zurücklegen (mit Zurücklegen wird gezogen, wenn das Argument `replace`
369 `= TRUE` gesetzt wird), gezogen.

```
sample(s, size = 1) # 1 Element
```

```
370 ## [1] 1
```

```
sample(s, size = 3) # 2 Elemente
```

```
371 ## [1] 15 7 22
```

372 Wenn `size` weggelassen wird, dann bekommt man gleich viele Elemente zurück (wie der Vektor lang ist),
373 d.h. der Vektor wird nur permutiert.

3.3 Beispiel Fotofallen

Für den weiteren Verlauf wollen wir noch einmal zu dem Beispiel mit den Fotofallen zurückkommen. Wir können jetzt 3 Vektoren erstellen, jeweils einen für die ID, die Anzahl Rehphotos und das Revier. Dabei werden zwei weitere Funktionen eingeführt (`paste` und `rep`).

Als erstes erstellen wir einen Vektor mit den Anzahlen Rehphotos. Das geht einfach mit `c()`:

```
anzahl_rehe <- c(132, 79, 129, 91, 138, 144, 55, 103, 139,
                 105, 96, 146, 95, 118, 1007)
```

Als zweites erstellen wir einen Vektor mit den IDs. Zur Erinnerung, diese sollten die Werte `Kamera_1` bis `Kamera_15` haben. Ein erster Ansatz könnte sein, dass wir einfach 15 Fotofallen schreiben und dann die Zahlen 1 bis 15 dahinter.

```
ids <- c("Kamera_1", "Kamera_2", "Kamera_3", "Kamera_4", "Kamera_5",
         "Kamera_6", "Kamera_7", "Kamera_8", "Kamera_9", "Kamera_10",
         "Kamera_11", "Kamera_12", "Kamera_13", "Kamera_14", "Kamera_15"
)
```

Dieser Ansatz ist unbefriedigend, da wir 15 mal das Wort “Kamera” tippen müssen. Wir können das Problem in zwei kleinere Probleme zerlegen: 1) 15 mal das Wort Kamera erstellen und die Zahlen 1 bis 15 erstellen, 2) die zwei Vektoren aus 1) “zusammenkleben”.

Ein Vektor kann mit der Funktion `rep` wiederholt werden, das heißt wir können ganz einfach 15 mal das Wort “Kamera” erstellen und speichern das Zwischenergebnis in einem Vektor `v1`.

```
v1 <- rep("Kamera", 15)
```

Im nächsten Schritt müssen wir die Zahlen 1 bis 15 erstellen, auch dieses Zwischenergebnis speichern wir in einem neuen Vektor `v2`.

```
v2 <- 1:15
```

Jetzt müssen wir lediglich die Vektoren `v1` und `v2` “zusammenkleben”. Dafür gibt es die Funktion `paste`, die zwei Vektoren elementweise verbindet, dabei wird das Argument `sep` als Trennzeichen verwendet. In unserem Fall wäre das also.

```
ids <- paste(v1, v2, sep = "_")
ids
```

```
## [1] "Kamera_1" "Kamera_2" "Kamera_3" "Kamera_4" "Kamera_5" "Kamera_6"
## [7] "Kamera_7" "Kamera_8" "Kamera_9" "Kamera_10" "Kamera_11" "Kamera_12"
## [13] "Kamera_13" "Kamera_14" "Kamera_15"
```

Dann fehlt jetzt lediglich der Vektor mit den Revieren. Hier könnten wir erneut auf die Funktion `rep` zurückgreifen.

```
rep(c("Revier A", "Revier B", "Revier C"), 5)
```

```
## [1] "Revier A" "Revier B" "Revier C" "Revier A" "Revier B" "Revier C"
## [7] "Revier A" "Revier B" "Revier C" "Revier A" "Revier B" "Revier C"
```

```
399 ## [13] "Revier A" "Revier B" "Revier C"
```

400 Das Ergebnis stimmt noch nicht ganz, da wir 5 mal Revier A usw. brauchen. Mit dem zusätzlichen Argument
401 `each = 5` können wir genau zu diesem Ergebnis kommen.

```
reviere <- rep(c("Revier A", "Revier B", "Revier C"), each = 5)
reviere
```

```
402 ## [1] "Revier A" "Revier A" "Revier A" "Revier A" "Revier A" "Revier B"
```

```
403 ## [7] "Revier B" "Revier B" "Revier B" "Revier B" "Revier C" "Revier C"
```

```
404 ## [13] "Revier C" "Revier C" "Revier C"
```

```
405
```

406 *Aufgabe 5: Statistische Funktionen*

```
407
```

408 1. Berechnen Sie den Mittelwert und Median für die Anzahl Fotos.

409 2. Erstellen Sie die folgende Ausgabe:

```
410 ## [1] "Die mittlere Anzahl von Rehphotos beträgt 171.8 Rehe pro Standort."
```

411 3.4 Arbeiten mit logischen Werten

412 Weniger bekannt sind die sogenannte booleschen Rechenregeln, also das Rechnen mit wahr (TRUE) und falsch
413 (FALSE). Dabei werden die folgenden Operationen am häufigsten verwendet.

- 414 • Gleichheit (==)
- 415 • Ungleichheit (!=)
- 416 • Größer (>) und kleiner (<)
- 417 • Größer gleich (>=) und kleiner gleich (<=)

418 Das Ergebnis von logischen Operatoren ist immer TRUE oder FALSE.

419 Bei Vektoren kommt es immer zu einer elementweisen Anwendung. Wir können beispielsweise abfragen an
420 welchen Fotofallenstandorten mehr als 100 Rehe fotografiert wurden oder

```
anzahl_rehe > 100
```

```
421 ## [1] TRUE FALSE TRUE FALSE TRUE TRUE FALSE TRUE TRUE TRUE FALSE TRUE
```

```
422 ## [13] FALSE TRUE TRUE
```

423 welche Fotofallenstandorte sich in Revier B befinden.

```
reviere == "Revier B"
```

```
424 ## [1] FALSE FALSE FALSE FALSE FALSE TRUE TRUE TRUE TRUE TRUE FALSE FALSE
```

```
425 ## [13] FALSE FALSE FALSE
```

426 Des Weiteren können logische Ausdrücke miteinander verknüpft werden. Dies geschieht mit einem logischen
427 Und (&) oder einem logischen Oder (|). Für das logische Und müssen beide Ausdrücke ein TRUE zurückgeben

um ein TRUE zu erhalten. Für ein logisches Oder reicht es, wenn einer der beiden Ausdrücke TRUE zurückgibt, um ein TRUE zu erhalten.

Damit können wir nun z.B. die beiden vorherigen Abfragen verbinden. Die erste Abfrage ist: Hat eine Fotofalle mehr als 100 Rehe fotografiert und stand die Fotofalle in Revier B.

```
anzahl_rehe > 100 & reviere == "Revier B"
```

```
## [1] FALSE FALSE FALSE FALSE FALSE TRUE FALSE TRUE TRUE TRUE FALSE FALSE
```

```
## [13] FALSE FALSE FALSE
```

Das war jetzt eine Abfrage mit einem logischen Und. Würden wir ein logisches Oder verwenden, dann bekommen wir für alle Elemente ein TRUE, die entweder in Gebiet B stehen oder mehr als 100 Rehphotos aufgezeichnet haben.

```
anzahl_rehe > 100 | reviere == "Revier B"
```

```
## [1] TRUE FALSE TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE
```

```
## [13] FALSE TRUE TRUE
```

Das Arbeiten mit logischen Werten kann fürs Erste etwas abstrakt erscheinen, aber wir werden im folgenden Abschnitt (Abschnitt 3.5) zahlreiche Anwendungsbeispiele dafür sehen.

TRUE wird intern als 1 gespeichert und FALSE als 0. Es ist möglich mit TRUEs und FALSEs zu rechnen.

```
TRUE + TRUE
```

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

```
FALSE + FALSE
```

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

```
TRUE + FALSE
```

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

Aufgabe 6: Arbeiten mit logischen Werten

Überlegen Sie für jede Aufgabe erst was das richtige Ergebnis ist und Überprüfen Sie dieses dann mit R.

1. TRUE | FALSE

2. FALSE & TRUE

3. (FALSE & TRUE) | TRUE

4. (2 != 3) | FALSE

5. FALSE + 10

6. TRUE + 10

7. TRUE + 10 == FALSE + 10

8. sum(c(TRUE, TRUE, FALSE, FALSE))

3.5 Zugreifen auf Elemente eines Vektors (=Untermengen)

Sehr oft wollen wir auf bestimmte Werte in einer Datenstruktur zugreifen. Beispielsweise könnte es uns interessieren, wieviele Rehe im Mittel auf allen Fotofallen aus Revier A gesehen wurden.

Bei Vektoren kann auf die einzelnen Elemente mit eckigen Klammern (`[]`, diese werden auch Indizierungs-klammern genannt) zugegriffen werden. Der Ausdruck `anzahl_rehe[2]` gibt die Anzahl an fotografierten Rehen für die zweite Fotofalle zurück. Es gibt zwei Möglichkeiten, was in die eckigen Klammern geschrieben werden kann: 1) die Positionen der Elemente die man zurückhaben möchte. Ist es mehr als ein Element, dann muss man einfach einen Vektor mit den Positionen übergeben; 2) ein logischer Vektor von der gleichen Länge, es werden alle Elemente zurückgegeben bei denen ein `TRUE` steht.

Abfragen des zweiten Elements in dem Vektor `anzahl_rehe`:

```
anzahl_rehe[2]
```

```
## [1] 79
```

Abfragen aller Elemente aus `anzahl_rehe`, die aus dem Revier A stammen.

```
anzahl_rehe[c(1, 2, 3, 4, 5)]
```

```
## [1] 132 79 129 91 138
```

```
# oder schneller
```

```
anzahl_rehe[1:5] # da `1:5` einen Vektor mit allen Zahlen von 1 bis 5 erstellt.
```

```
## [1] 132 79 129 91 138
```

Hier ist nochmals hervorzuheben, dass innerhalb der eckigen Klammer mit dem Befehl `c(1, 2, 3, 4, 5)` bzw. `1:5` ein Vektor erstellt wird, der die Position der Elemente angibt, die zurückgegeben werden sollen.

Aufgabe 7: Zugreifen auf Vektorelemente

Erstellen Sie einen neuen Vektor `bhd`

```
bhd <- c(12, 32, 39, 41, 12, 30)
```

- Wählen Sie aus dem Vektor `bhd` nur das 2. und 3. Element aus.
- Wählen Sie aus dem Vektor `bhd` das letzte Element aus.
- Wählen Sie aus dem Vektor `bhd` das letzte Element aus, ohne die Zahl 6 zu schreiben.

Alternativ könnte das gleiche Ergebnis mit einem logischen Vektor erreicht werden. Für eine bessere Übersichtlichkeit wird erst ein Vektor `sub` erstellt, in dem die logischen Werte gespeichert werden:

```
sub <- c(TRUE, TRUE, TRUE, TRUE, TRUE, FALSE, FALSE, FALSE,
        FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE)
anzahl_rehe[sub]
```

```
## [1] 132 79 129 91 138
```

Das Erstellen des `sub`-Vektors ist mühsam und wenig zielführend. Wenn wir auf die Erkenntnisse aus dem vorherigen Kapitel zurückgreifen, kann dies leicht automatisiert werden, indem wir einfach abfragen, welche Elemente in Revier zu Revier A gehören.

```
sub <- reviere == "Revier A"
anzahl_rehe[sub]
```

```
## [1] 132 79 129 91 138
```

Das kann nochmals vereinfacht werden, indem wir den Hilfsvektor `sub` einfach weglassen und den Ausdruck direkt in die eckigen Klammern ziehen.

```
anzahl_rehe[reviere == "Revier A"]
```

```
## [1] 132 79 129 91 138
```

Wenn wir jetzt noch den Mittelwert der Anzahl fotografierten Rehe aus Revier A bilden möchten, erweitert sich der Ausdruck um einen Funktionsaufruf zur Funktion `mean`.

```
mean(anzahl_rehe[reviere == "Revier A"])
```

```
## [1] 113.8
```

Aufgabe 8: logische Werte

Verwenden Sie nochmals den Vektor mit den Anzahl Rehen die an unterschiedlichen Fotofallenstandorten fotografiert wurden.

```
anzahl_rehe <- c(132, 79, 129, 91, 138, 144, 55, 103, 139, 105, 96, 146, 95, 118, 107)
```

1. Wählen Sie alle Standorte aus für die Aussage zu $90 \leq x < 120$ zu trifft (wobei x für die Anzahl Fotos an einem Standort steht).
2. Berechnen Sie die mittlere Anzahl Fotos für alle in 1) ausgewählten Standorte.

3.6 Der %in%-Operator

Häufig wollen wir mehrere Elemente aus einem Vektor auswählen, die in einem anderen Vektor enthalten sind. Als einfaches Beispiel nehmen wir zwei Vektoren:

```
arten <- c("FI", "BU")
messungen_arten <- c("FI", "BU", "BU", "EI", "EI", "BI", "FI", "BI", "EI")
```

Wenn wir aus dem Vektor `messungen_arten` alle FI auswählen wollen, können wir dies mit einem logischen `==` machen:

```
messungen_arten[messungen_arten == "FI"]
```

```
## [1] "FI" "FI"
```

```
# oder
messungen_arten[messungen_arten == arten[1]]
```

```
## [1] "FI" "FI"
```

Etwas komplizierter wird es, wenn wir zwei oder mehr Elemente auswählen wollen. Dies geht auch mit logischen Operationen.

```
messungen_arten[messungen_arten == arten[1] | messungen_arten == arten[2]]
```

```
## [1] "FI" "BU" "BU" "FI"
```

Diese Herangehensweise wird aber für > 2 Elemente in `arten` sehr mühsam. Eine Alternative bietet der %in%-Operator. Dieser testet, ob Elemente eines Vektors in einem zweiten Vektors enthalten sind.

```
messungen_arten %in% arten
```

```
## [1] TRUE TRUE TRUE FALSE FALSE FALSE TRUE FALSE FALSE
```

```
messungen_arten[messungen_arten %in% arten]
```

```
## [1] "FI" "BU" "BU" "FI"
```

```
516
```

Aufgabe 9: Auswählen von Elementen in einem Vektor (%in%)

Der Vector `LETTERS` ist in R vorhanden und enthält die Buchstaben von A bis Z.

```
LETTERS
```

```
## [1] "A" "B" "C" "D" "E" "F" "G" "H" "I" "J" "K" "L" "M" "N" "O" "P" "Q" "R" "S"
```

```
## [20] "T" "U" "V" "W" "X" "Y" "Z"
```

Wählen Sie aus `LETTERS` nur die Vokale aus.

4 Faktoren (factors)

R besitzt einen besonderen Datentyp – *Faktoren* (engl. factors) – zum speichern von diskreten Kovariaten (z.B. Baumart, Augenfarbe oder Automarke). Faktoren erlauben es Daten vom Typ `character` effizienter abzuspeichern. Dabei wird jeder eindeutiger Wert (=Level) mit einer Zahl codiert und dann werden nur diese Zahlen zusammen mit einer Tabelle zum Nachschauen der Werte gespeichert (siehe dazu auch McNamara and Horton 2018).

Mit der Funktion `factor()` kann ein Faktor erstellt werden. Im einfachsten Fall wird nur ein Vektor übergeben.

```
a <- c("FI", "BU", "FI", "EI", "EI", "FI", "FI")
factor(a)
```

```
## [1] FI BU FI EI EI FI FI
## Levels: BU EI FI
```

Ohne weitere Spezifikation werden die Werte *Levels* alphabetisch angeordnet (das kann später z.B. beim Erstellen von Abbildungen wichtig sein), dies kann jedoch durch die Verwendung des Arguments `levels` gesteuert werden.

```
factor(a, levels = c("FI", "BU", "EI"))
```

```
## [1] FI BU FI EI EI FI FI
## Levels: FI BU EI
```

Es ist auch möglich die Beschriftung (= labels) der unterschiedlichen Levels anzugeben mit dem Argument `labels`.

```
af <- factor(a, levels = c("FI", "BU", "EI"),
             labels = c("Fichte", "Buche", "Eiche"))
af
```

```
## [1] Fichte Buche Fichte Eiche Eiche Fichte Fichte
## Levels: Fichte Buche Eiche
```

Mit der Funktion `levels()` können die unterschiedlichen Levels eines Faktors abgefragt bzw. gesetzt werden.

```
levels(af)
```

```
## [1] "Fichte" "Buche" "Eiche"
levels(af) <- c("Fi", "Bu", "Ei")
af
```

```
## [1] Fi Bu Fi Ei Ei Fi Fi
## Levels: Fi Bu Ei
```

Schlussendlich kann man mit der Funktion `relevel()` die Referenzkategorie eines Faktors (der erste Level) angepasst werden. Das ist kann für lineare Modell wichtig sein.

```
af
```

```
## [1] Fi Bu Fi Ei Ei Fi Fi
```



```
548 ## Levels: Fi Bu Ei
```

```
relevel(af, "Bu")
```

```
549 ## [1] Fi Bu Fi Ei Ei Fi Fi
```

```
550 ## Levels: Bu Fi Ei
```

551 Mit der Funktion `as.character()` kann ein Faktor wieder als Variable vom Typ `character` dargestellt werden.

```
as.character(af)
```

```
553 ## [1] "Fi" "Bu" "Fi" "Ei" "Ei" "Fi" "Fi"
```

554 Achtung mit der Funktion `as.numeric()` erhält man die interne Kodierung von Faktoren.

```
af
```

```
555 ## [1] Fi Bu Fi Ei Ei Fi Fi
```

```
556 ## Levels: Fi Bu Ei
```

```
as.numeric(af)
```

```
557 ## [1] 1 2 1 3 3 1 1
```

558 Fichte ist das erste Level des Faktors, deshalb erhalten alle Fichteneinträge den Wert 1. Bucheneinträge erhalten den Wert 2 und 3 für Eichen.

```
560
```

561 *Aufgabe 10: Faktoren*

563 Verwenden Sie den Vektor `staedte` und erstellen Sie einen Vektor mit der Anordnung der `levels` in umgekehrter alphabetischer Reihenfolge.

```
564 staedte <- c("Berlin", "Aachen", "Berlin", "Ulm", "Aachen",  
              "Berlin", "Berlin", "Aachen", "Ulm", "Ulm")
```

5 Spezielle Einträge

In vielen Fällen werden spezielle Einträge benötigt, bspw. bei

- fehlenden Einträge `NA`,
- leeren Einträgen `NULL`,
- undefinierten Einträgen `NaN` (Not a Number) oder
- unendlichen Zahlen (`Inf`).

Spezielle Einträge sind reservierte Namen. Sie können nicht überschrieben werden.

5.1 NA

R verfügt über einen speziellen Wert für fehlende Einträge. Auch wenn in Vektoren eigentlich nur ein Datentyp erlaubt ist, sind `NA` zwischen den anderen Einträgen erlaubt. Der Datentyp des Vektors wird durch `NA` Einträge nicht verändert.

```
na1 <- c("foo", NA, "foo")
str(na1)
```

```
## chr [1:3] "foo" NA "foo"
```

```
na2 <- c(3, 6, NA)
str(na2)
```

```
## num [1:3] 3 6 NA
```

Der logische Operator zum Test auf fehlende Wert ist `is.na()`. Dieser kann genauso wie die bereits bekannten logischen Operatoren bspw. zum Filtern verwendet werden. Die `na.omit()` Funktion entfernt `NA` aus dem Datensatz.

```
is.na(na1)
```

```
## [1] FALSE TRUE FALSE
```

```
na.omit(na1)
```

```
## [1] "foo" "foo"
```

```
## attr(,"na.action")
```

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

```
## attr(,"class")
```

```
## [1] "omit"
```

Die bereits bekannten logischen Operationen ergeben `NA`, wenn Sie auf Daten angewendet werden, die `NA` enthalten. Berechnungen mit `NA` ergeben ebenfalls `NA`. Bei der angewandten Programmierung müssen sie also darauf achten, dass Ihre Daten frei von `NA` sind oder sie fangen die `NA` vorher ab.

```
na2 < 3
```

```
## [1] FALSE FALSE NA
```

```
1 + NA
```

```
## [1] NA
```

Viele R Funktionen haben eingebaute Methoden zum Umgang mit NA. Die Funktion `mean()` bspw. ergibt (wie die meisten Funktionen) standardmäßig NA wenn sie auf Vektoren mit Datenlücken angewendet wird, es sei denn man stellt innerhalb der Funktion ein, dass Datenlücken entfernt werden sollen.

```
mean(na2)
```

```
## [1] NA
```

```
mean(na2, na.rm = TRUE)
```

```
## [1] 4.5
```

5.2 NULL

Im Gegensatz zu NA wird NULL für leere Einträge verwendet, und nicht für fehlende Einträge. Da in der Mathematik leere Einträge und fehlende Einträge unterschiedliche Informationen darstellen, können diese beiden Fälle unterschieden werden. Mit der Funktion `is.null()` kann man überprüfen, ob ein Element in einem Vektor NULL ist oder nicht.

5.3 Inf

Die größtmögliche Zahl in R ist $1.7976931 \cdot 10^{308}$. Größere Zahlen werden als unendlich gespeichert.

```
10^309
```

```
## [1] Inf
```

```
2 * Inf
```

```
## [1] Inf
```

```
1 + Inf
```

```
## [1] Inf
```

```
3 / 0
```

```
## [1] Inf
```

```
-3 / 0
```

```
## [1] -Inf
```

```
3 / Inf
```

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

Infinity kann mit `is.infinite` und `is.finite` getestet werden. Relationäre Operatoren funktionieren erwartungsgemäß.

```
inf1 <- c(Inf, 0, 3, -Inf, 10)
is.infinite(inf1)
```

```
## [1] TRUE FALSE FALSE TRUE FALSE
```

```
is.finite(inf1)
```

```
## [1] FALSE TRUE TRUE FALSE TRUE
```

```
inf1 < 3
```

```
## [1] FALSE TRUE FALSE TRUE FALSE
```

```
615
```

616 *Aufgabe 11: Vektoren mit speziellen Einträgen*

618 Verwenden Sie den Vektor

```
foo <- c(13563, -13156, -14319, 16981, 12921, 11979, 9568, 8833, -12968, 8133)
```

- 619
- Nehmen Sie jeden Eintrag hoch 75. Filtern Sie alle unendlichen Einträge aus dem Vektor.
- 620
- Wie viele Einträge sind unendlich negativ?

621 Verwenden Sie den Vektor

```
foo <- c(4.3, 2.2, NULL, 2.4, NaN, 3.3, 3.1, NULL, 3.4, NA, Inf)
```

622 Sind die folgenden Einträge richtig oder falsch? Überlegen Sie zunächst selbst bevor Sie die Aussagen in R
623 testen.

- 624
- Die Länge des Vektors ist 9.
- 625
- `is.na()` ergibt 2 Mal TRUE.
- 626
- `foo[9] + 4 / Inf` ergibt NA

627 Berechnen Sie den arithmetischen Mittelwert von `foo`.

6 data.frames oder Tabellen

Im vorherigen Teilabschnitt haben wir gesehen, wie mehrere Werte des gleichen Typs in einem Vektor zusammengefasst werden können. Abschließend wurde anhand des Fotofallenbeispiels gezeigt, wie Vektoren eingesetzt werden können. Wir erstellten drei Vektoren, die jeweils die Merkmalsausprägungen eines Merkmals aller Fotofallenstandorte speichern. In statistischer Sprache, sind die Fotofallen die Beobachtungen (oder auch Merkmalsträger genannt) und die Informationen zu den Fotofallen (also ID, Anzahl Rehe und das Revier) die Merkmale. Jeder beobachtete Wert (z.B. die 132 fotografierten Rehe von Kamera 1) ist dann eine Merkmalsausprägung.

Sie können sich ein `data.frame` wie eine Tabelle aus einem Tabellenkalkulationsprogramm vorstellen. Es gibt Zeilen in denen die Beobachtungen gespeichert sind und Spalten, die die Merkmale speichern. In unserem Fall gäbe es 15 Zeilen (eine Zeile für jede Fotofalle) und drei Spalten (jeweils eine Spalte für ID, Anzahl Rehe und Revier). Der Befehl zum Erstellen eines `data.frames` in R ist `data.frame()`. Für unser Beispiel wäre es:

```
monitoring <- data.frame(
  ID = ids,
  anzahl_rehe = anzahl_rehe,
  revier = reviere
)
monitoring
```

	##	ID	anzahl_rehe	revier
##	1	Kamera_1	132	Revier A
##	2	Kamera_2	79	Revier A
##	3	Kamera_3	129	Revier A
##	4	Kamera_4	91	Revier A
##	5	Kamera_5	138	Revier A
##	6	Kamera_6	144	Revier B
##	7	Kamera_7	55	Revier B
##	8	Kamera_8	103	Revier B
##	9	Kamera_9	139	Revier B
##	10	Kamera_10	105	Revier B
##	11	Kamera_11	96	Revier C
##	12	Kamera_12	146	Revier C
##	13	Kamera_13	95	Revier C
##	14	Kamera_14	118	Revier C
##	15	Kamera_15	107	Revier C

Im vorhergehenden Codebeispiel wurde ein `data.frame` erstellt und in die Variable `monitoring` gespeichert. Die Funktion `data.frame()` nimmt als Argumente beliebig viele Paare, die immer aus einem Namen und einem Vektor mit dazugehörigen Werten bestehen. D.h., dass immer eine Spalte vom selben Typ sein muss, es aber für jede Beobachtung (=Zeile) Merkmale von unterschiedlichen Typen geben kann.

6.1 Wichtige Funktionen zum Arbeiten mit `data.frames`

Wichtige Funktionen um das Arbeiten mit `data.frames` zu erleichtern sind wieder `head()` und `tail()`, um die ersten bzw. letzten `n` Zeilen eines `data.frames` anzuzeigen.

```
head(monitoring, n = 2)
```

```
##           ID anzahl_rehe   revier
## 1 Kamera_1          132 Revier A
## 2 Kamera_2           79 Revier A
```

Oder für die letzten 2 Beobachtungen.

```
tail(monitoring, 2)
```

```
##           ID anzahl_rehe   revier
## 14 Kamera_14          118 Revier C
## 15 Kamera_15          107 Revier C
```

Mit den Funktion `nrow()` und `ncol()` können die Anzahl Zeilen und die Anzahl Spalten abgefragt werden:

```
nrow(monitoring)
```

```
## [1] 15
```

```
ncol(monitoring)
```

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

Mit der Funktion `str()` (kurz für *structure*) kann schnell ein Überblick über sämtliche Variablen verschafft werden.

```
str(monitoring)
```

```
## 'data.frame':   15 obs. of  3 variables:
## $ ID           : chr  "Kamera_1" "Kamera_2" "Kamera_3" "Kamera_4" ...
## $ anzahl_rehe: num  132 79 129 91 138 144 55 103 139 105 ...
## $ revier      : chr  "Revier A" "Revier A" "Revier A" "Revier A" ...
```

Aufgabe 12: `data.frame`

Stellen Sie sich vor, Sie machen eine kleine Umfrage, in der Sie fünf Menschen nach ihrem Studienfach, Semester und Alter befragen. Erstellen Sie ein `data.frame` mit dem Namen `umfrage1` für diese Informationen und fragen Sie entweder fünf Mitstudierende oder erfinden Sie die Daten einfach.

6.2 Zugreifen auf Elemente eines `data.frame`

Für `data.frames` gilt genau das gleiche Prinzip. Nur dass wir jetzt zwei Dimensionen berücksichtigen müssen: nämlich die Zeilen und Spalten. Wir können immer noch mit eckigen Klammern (`[]`) auf Elemente innerhalb

eines `data.frames` zugreifen, müssen aber jetzt die *Zeile(n)* und die *Spalte(n)* angeben, die wir haben möchten. Die Schreibweise ist immer `[Zeile(n), Spalte(n)]`. Für Zeilen und Spalten gelten genau die gleichen Regeln wie für Vektoren. Wir können entweder einen Vektor mit den Positionen für die gewünschten Zeilen und Spalten angeben oder einen logischen Vektor, der besagt welche Zeilen und Spalten wir zurückhaben möchten. Wenn wir z.B. die Anzahl Rehphotos von der vierten Fotofalle abfragen möchten, könnte man das so machen.

```
monitoring[4, 2]
```

```
## [1] 91
```

Alternativ, kann man den Spaltennamen auch einfach Ausschreiben.

```
monitoring[4, "anzahl_rehe"]
```

```
## [1] 91
```

Wenn wir die Anzahl fotografierte Rehe von den ersten fünf Fotofallen abfragen möchten, dann müssen wir für die Zeilen einen Vektor mit den Zahlen 1 bis 5 übergeben, für die Spalten ändert sich nichts.

```
monitoring[1:5, "anzahl_rehe"]
```

```
## [1] 132 79 129 91 138
```

Wenn wir nun nicht nur die Anzahl fotografierte Rehe zurückhaben möchten, sondern auch noch das Revier für die ersten fünf Fotofallen, dann müssen wir für die Spalten lediglich das Revier hinzufügen.

```
monitoring[1:5, c("anzahl_rehe", "revier")]
```

```
##   anzahl_rehe   revier
## 1         132 Revier A
## 2          79 Revier A
## 3         129 Revier A
## 4          91 Revier A
## 5         138 Revier A
```

Wenn wir alle Spalten und/oder Zeilen eines `data.frames` abfragen möchten, dann kann man diese Position einfach frei lassen. Eine Abfrage für die ersten fünf Spalten aller Fotofallen würde so aussehen.

```
monitoring[1:5, ]
```

```
##      ID anzahl_rehe   revier
## 1 Kamera_1         132 Revier A
## 2 Kamera_2          79 Revier A
## 3 Kamera_3         129 Revier A
## 4 Kamera_4          91 Revier A
## 5 Kamera_5         138 Revier A
```

Aufgabe 13: Abfragen von Werten

Wir nehmen folgende Werte aus Übung 12 an:

```
umfrage1 <- data.frame(
  fach = c("Forst", "Bio", "Chemie", "Physik", "Forst"),
  semester = c(2, 3, 2, 1, 5),
  alter = c(21, 22, 21, 20, 23)
)
```

- Wählen Sie nur die ersten drei Zeilen aus und die erste und zweite Spalte aus.
- Wählen Sie alle Zeilen und die erste und dritte Spalte aus.
- Wählen Sie alle Spalten und die erste, dritte und vierte Zeile aus.

Mit dem `$`-Zeichen kann bei `data.frames` direkt auf Spalten zugegriffen werden. Wenn wir z.B. für alle Fotofallen die Anzahl gesehener Rehe abfragen möchten, gibt es jetzt drei Möglichkeiten:

1. über das `$`-Zeichen direkt die Spalten ansprechen.

```
monitoring$anzahl_rehe
```

```
## [1] 132 79 129 91 138 144 55 103 139 105 96 146 95 118 107
```

2. Einfach die Positionen für die Zeilen leer lassen und die Spalte abfragen.

```
monitoring[, "anzahl_rehe"]
```

```
## [1] 132 79 129 91 138 144 55 103 139 105 96 146 95 118 107
```

3. Alle Zeilen und die Spalte explizit angeben.

```
monitoring[1:nrow(monitoring), "anzahl_rehe"]
```

```
## [1] 132 79 129 91 138 144 55 103 139 105 96 146 95 118 107
```

Anmerkung zu 3), der Ausdruck `1:nrow(monitoring)` ergibt einen Vektor mit den Zahlen 1 bis 15, da `nrow(monitoring) = 15` ist. So eine Schreibweise ist zu empfehlen, wenn die Dimension des Vektors variabel ist.

Schlussendlich kann man einen `data.frame` auch mit logischen Vektoren abfragen. Ein Beispiel wäre, wenn wir alle Fotofallen abfragen möchten, die mehr als 100 Rehphotos gemacht haben. Der erste Schritt wäre abzufragen, ob eine Fotofalle mehr als 100 Rehphotos gemacht hat.

```
monitoring$anzahl_rehe > 100
```

```
## [1] TRUE FALSE TRUE FALSE TRUE TRUE FALSE TRUE TRUE TRUE FALSE TRUE
```

```
## [13] FALSE TRUE TRUE
```


739 Das Ergebnis ist ein Vektor mit 15 Elementen. Hat eine Fotofalle mehr als 100 Rehphotos gemacht ist das
740 entsprechende Element des Vektors `TRUE` ansonsten `FALSE`. In dem `data.frame` `monitoring` steht in jeder
741 Zeile eine Beobachtung (also eine Fotofalle). Nun wollen wir genau diese Fotofallen haben, die mehr als 100
742 Rehphotos gemacht gemacht haben.

```
monitoring[monitoring$anzahl_rehe > 100, ]
```

```
743 ##           ID anzahl_rehe   revier  
744 ## 1   Kamera_1         132 Revier A  
745 ## 3   Kamera_3         129 Revier A  
746 ## 5   Kamera_5         138 Revier A  
747 ## 6   Kamera_6         144 Revier B  
748 ## 8   Kamera_8         103 Revier B  
749 ## 9   Kamera_9         139 Revier B  
750 ## 10 Kamera_10         105 Revier B  
751 ## 12 Kamera_12         146 Revier C  
752 ## 14 Kamera_14         118 Revier C  
753 ## 15 Kamera_15         107 Revier C
```

754

755 *Aufgabe 14: Abfragen von Werten 2*

757 Verwenden Sie erneut den Datensatz aus Übung 13 und führen Sie folgende Abfragen durch:

- 758 • Alle Spalten für Studierende die Forstwissenschaften studieren.
- 759 • Alle Spalten für Studierende die Chemie oder Physik studieren.
- 760 • Die Spalte `fach` und `semester` für Studierende die 22 oder älter sind.

7 Schreiben und lesen von Daten

7.1 Textdateien

Bis jetzt haben wir Daten immer in R erstellt, dies ist eine eher unnatürliche Situation. In den meisten Fällen bekommen Sie Daten von Dritten, Sensoren oder sonstigen Quellen. Diese Daten müssen dann in R eingelesen werden. Daten liegen meist in einer tabellarischen Form und als Textdatei vor⁶.

Die Funktion `read.table` erlaubt es eine Textdatei in R einzulesen. Dabei sind fürs Erste drei Argumente wichtig:

- **file:** Der Pfad zur Datei die eingelesen werden soll. Dieser kann *absolut* oder *relativ* sein. Ein absoluter Pfad gibt den Ort der Datei, die gelesen werden soll, komplett an (auf einem Windows Rechner wäre das wahrscheinlich `C:/Users/...`). Im Gegensatz dazu gibt ein relativer Pfad den Ort an, an dem die Datei, die eingelesen werden soll, relativ zum aktuellen Arbeitsverzeichnis (auch `working directory`) von R an. Man kann das Arbeitsverzeichnis von R mit der Funktion `setwd()` setzen, es hat sich jedoch als sinnvoller erwiesen mit RStudio-Projekten zu arbeiten (mehr dazu im nächsten Abschnitt).
- **header:** Dieses Argument gibt an, ob die erste Zeile eine Kopfzeile mit den Spaltenüberschriften ist. Meist haben wir eine Kopfzeile, dann wäre `header = TRUE` richtig.
- **sep:** Das Trennzeichen zwischen verschiedenen Spalten. Es ist meist ein Leerzeichen (), Komma (`,`) oder Strichpunkt (`;`).

Die Datei `fotofallen.csv` finden Sie auf StudIP und kann einfach heruntergeladen werden. Die Datei kann mit dem folgenden Befehl in R eingelesen werden. Hier wurde die Datei in einem RStudio-Projekt in ein Verzeichnis `data` abgelegt.

```
dat <- read.table("data/fotofallen.csv", header = TRUE, sep = ",")
head(dat)
```

```
##          ID anzahl_rehe   revier
## 1 Kamera_1         132 Revier A
## 2 Kamera_2          79 Revier A
## 3 Kamera_3         129 Revier A
## 4 Kamera_4          91 Revier A
## 5 Kamera_5         138 Revier A
## 6 Kamera_6         144 Revier B
```

Es gibt viele Varianten der Funktion `read.table()`. Beispielsweise hat die Funktion `read.csv()` bereits die Argumente `sep = ','` und `header = TRUE` gesetzt. Die Funktion `read.csv2()` hat das Argument `sep = ';'` gesetzt. Siehe dazu auch die Hilfeseite von `read.table()`. Diese kann entweder mit `?read.table` oder `help("read.table")` aufgerufen werden.

Mit der Funktion `write.table()` kann ein `data.frame` auf die Festplatte geschrieben werden.

⁶Natürlich gibt es viele weitere Formate wie Daten vorliegen können, diese werden aber an dieser Stelle nicht weiter behandelt. Es sei lediglich auf das Paket `readxl` verwiesen, falls Sie Daten von MS Excel direkt in R einlesen möchten.

793

794 ***Aufgabe 15: Lesen und Schreiben von Dateien***
795

796 Lesen Sie die Datei `kompliziert.txt` ein. Schauen Sie die Hilfeseite an und vergewissern Sie sich, dass Sie
797 wissen was die Argumente `header`, `sep`, `dec` und `skip` bewirken. Setzen die Argumente richtig, damit die
798 Datei `kompliziert.txt` folgendes Ergebnis liefert.

8 Erstellen von Abbildungen

Abbildungen sind ein elementarer Baustein statistischer Analysen und deshalb von Beginn an Teil von R. Es gibt unterschiedliche Systeme einen Plot zu erstellen. In diesem Kurs werden wir kurz *Base Plots* vorstellen und dann das Zusatzpaket `ggplot2` vorstellen.

8.1 Base Plot

Die wichtigsten Grafiken für die einfache Datendarstellung sind schnell verfügbar. Etwas komplexere oder spezielle Grafiken erfordern mehr Programmieraufwand (folgt teilweise noch). Stellen sie sich die einfache Grafik Schnittstelle als zweidimensionale Leinwand vor, auf die Sie durch Code Ebene für Ebene Grafikelemente zeichnen:

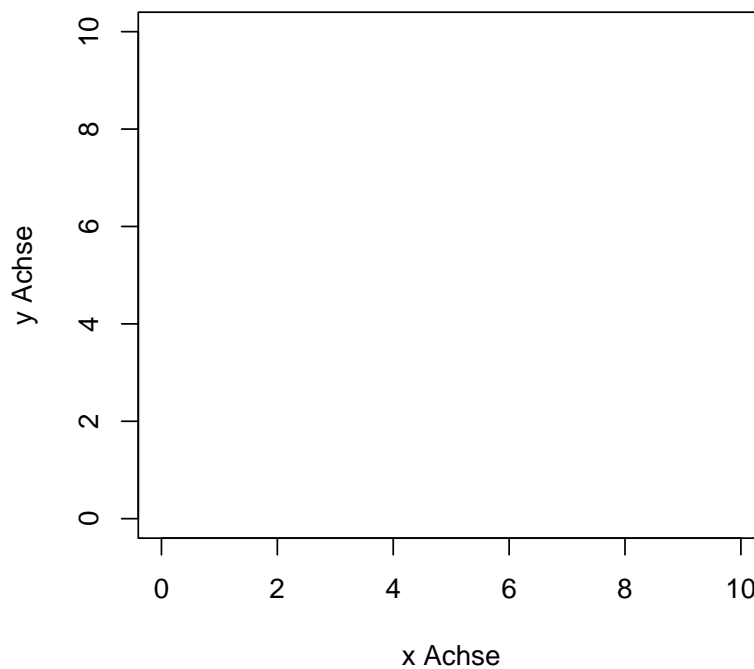
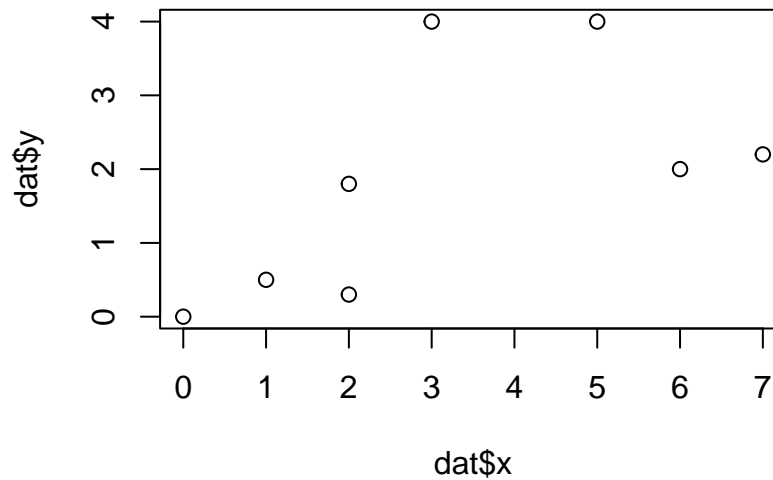


Abbildung 4: Beispiel einer leeren Grafikschnittstelle.

Hier drei einfache Beispiele für Abbildungen mit nur einer Ebene.

```
dat <- data.frame(
  x = c(0, 1, 2, 2, 3, 5, 6, 7),
  y = c(0, 0.5, 1.8, 0.3, 4, 4, 2, 2.2)
)

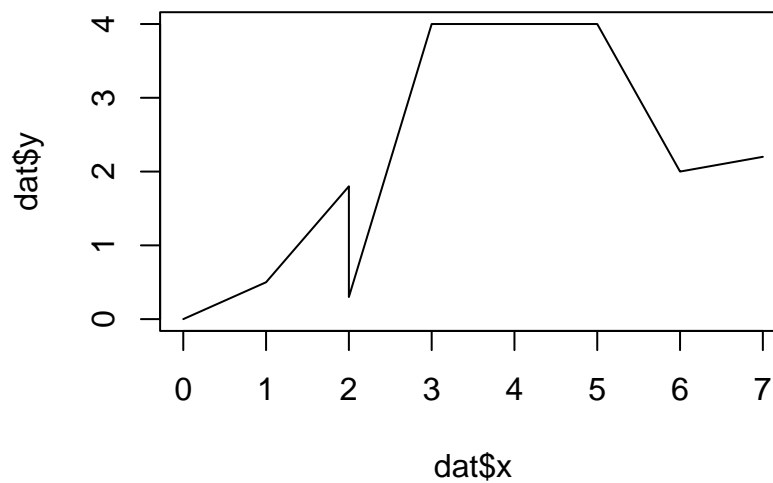
plot(dat$x, dat$y, type = "p")
```



809

810 Mit dem Argument `type` kann die Art der Darstellung gesteuert werden. Der Standardwert ist `type = "p"`
811 (für `points`). Wir können den selben Plot mit Linien (`type = "l"`)

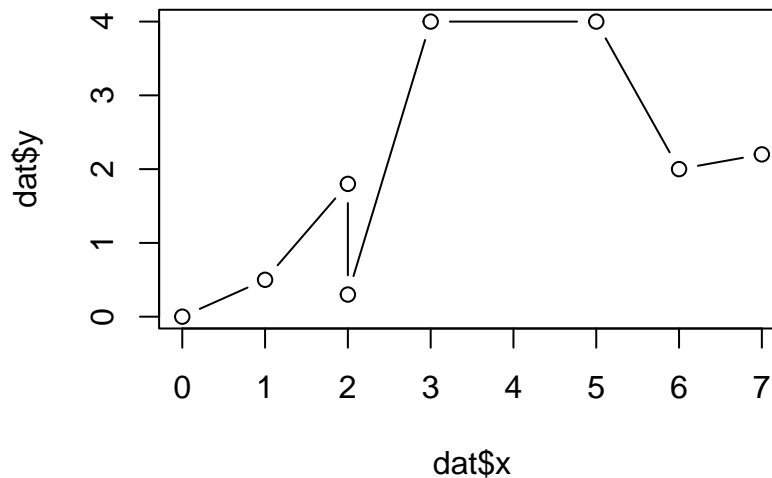
```
plot(dat$x, dat$y, type = "l")
```



812

813 oder mit Linien und Punkten (`type = "b"` für `both`)

```
plot(dat$x, dat$y, type = "b")
```



darstellen.

Aufgabe 16: Base Plot 1

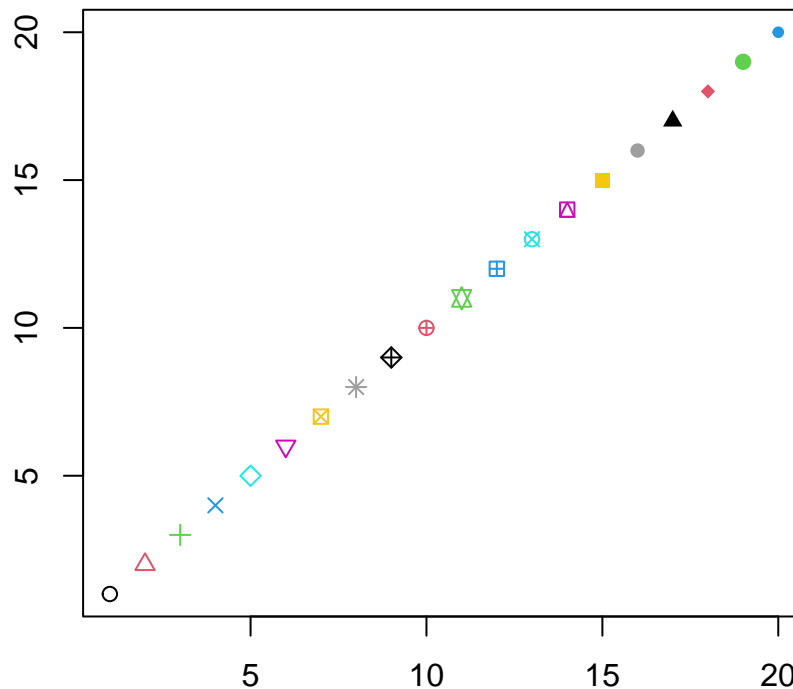
Laden Sie den Datensatz `bhd_1.txt` und erstellen Sie eine Abbildung mit dem Alter jedes Baumes auf der x-Achse und dem BHD auf der y-Achse.

Sie können entweder eine Grafik mit einem Befehl erzeugen (High-Level) oder die einzelnen Ebenen nacheinander erzeugen (Low-Level). Sie können jeder Ebenen durch zusätzliche Befehle innerhalb des Funktionsaufrufs Elemente hinzufügen und Einstellungen ändern. Die wichtigsten sind:

- `type` - Diagrammtyp
- `col` - Farbe
- `main` - Titel
- `sub` - Untertitel
- `pch` - Punktsymbol
- `lty` - Linientyp
- `lwd` - Linienstärke
- `xlab` bzw. `ylab` - Achsenbeschriftungen
- `xlim`, `ylim` - Grenzen der Achsenanschnitte
- `axes` - Sollen die Achsen eingezeichnet werden? Oder leer gelassen werden, um sie nachträglich als low-level Ebene einzuzeichnen?
- `ann` - Achsenbeschriftung kann ganz weggelassen werden.

Sehen Sie sich die Hilfeseiten `?plot.default()` oder `?par()` an für weiter Informationen. Dort finden Sie auch eine vollständige Liste der Befehle. Einige Argumente können als Vektor übergeben werden. Hier z. B. die Farben und die Punktsymbole.

```
plot(1:20, 1:20, pch = c(1 : 20), col = c(1 : 20), ann = FALSE)
```



Aufgabe 17: Anpassen von Plots

Verwenden Sie den Datensatz aus Übung 16 und passen Sie die Abbildung wie folgt an:

- Beschriften Sie die x- und y-Achse sinnvoll.
- Fügen Sie eine Überschrift hinzu.
- Wählen Sie ein anderes Symbol.
- Stellen Sie die Symbole in rot dar.

Über Low-Level Funktionen können einer Grafik Schnittstelle nacheinander Elemente hinzugefügt werden. Die wichtigsten Funktionen sind

- `points()` - Fügt Punkte ein
- `lines()` - Fügt Linien ein
- `text()` - Fügt Text ein
- `legend()` - Fügt eine Legende ein
- `abline()` - Fügt eine Gerade ein
- `curve()` - Fügt eine mathematische Funktion ein
- `arrows()` - Fügt Pfeile ein
- `grid()` - Fügt Hilfslinien ein

Dabei ist der Aufbau zunächst grundsätzlich wie in Abbildung 5 dargestellt. Der Vorteil von Low-Level Funktionen ist, dass die einzelnen Level mehr Funktionen bieten als die High-Level Funktion und, dass Sie sich die Reihenfolge der Ebenen definieren können.

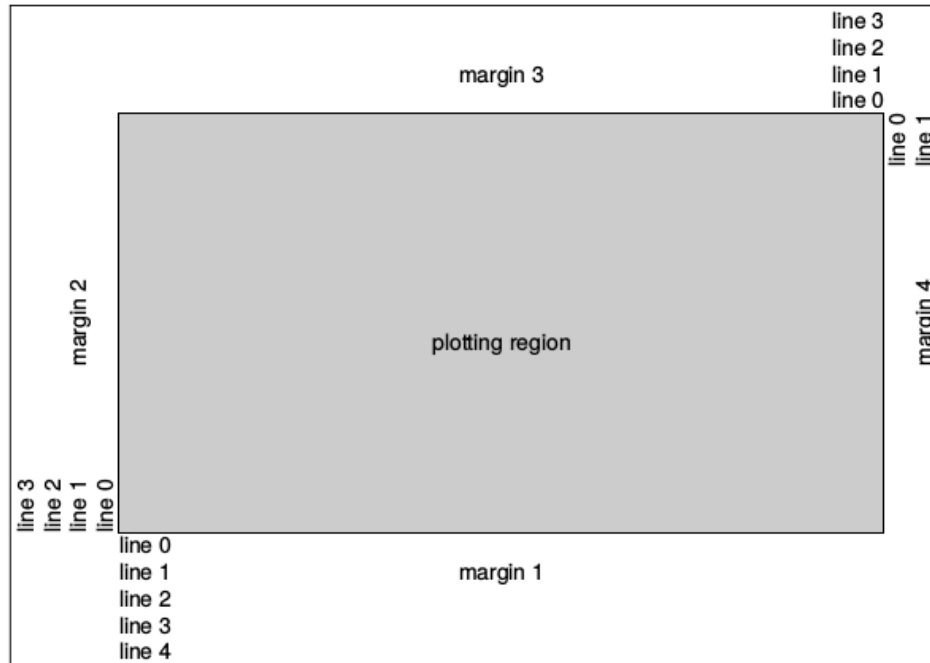


Abbildung 5: Grafikregionen eines base plots in R.

863 Mit diesem grundsätzlichen Aufbau sollten Sie bereits in der Lage sein auch komplexe Grafiken schnell zu
 864 gestalten. Wenn Sie mehrere Diagramme in einem Plot arrangieren möchten, können Sie mit dem `par()`
 865 Befehl ein Arrangement definieren. Sie haben dann zusätzlich zu den bereits bekannten Grafikregionen noch
 866 äußere Ränder (*outer margins*). Siehe Abbildung 6.

8.1.1 Mehrere Panels

868 Mit der Funktion `par()` kann auch eingestellt werden, dass ein Plot aus mehreren Subplots (= Panels)
 869 besteht. Die Argumente `mfrow` und `mfcol` können `par()` übergeben werden und kontrollieren die Anzahl
 870 Zeilen und Spalten für den Plot.

```
par(mfrow = c(1, 3))
```

871 Teilt den Plot in eine Zeile und drei Zeilen (= drei Plots nebeneinander).

```
# Erstes Panel
plot(bhd ~ alter, xlab = "Alter", ylab = "BHD", pch = 20, col = "red",
     data = dat[dat$gebiet == "A", ], main = "Gebiet A")

# Zweites Panel
plot(bhd ~ alter, xlab = "Alter", ylab = "BHD", pch = 20, col = "red",
     data = dat[dat$gebiet == "B", ], main = "Gebiet B")

# Drittes Panel
plot(bhd ~ alter, xlab = "Alter", ylab = "BHD", pch = 20, col = "red",
     data = dat[dat$gebiet == "C", ], main = "Gebiet C")
```

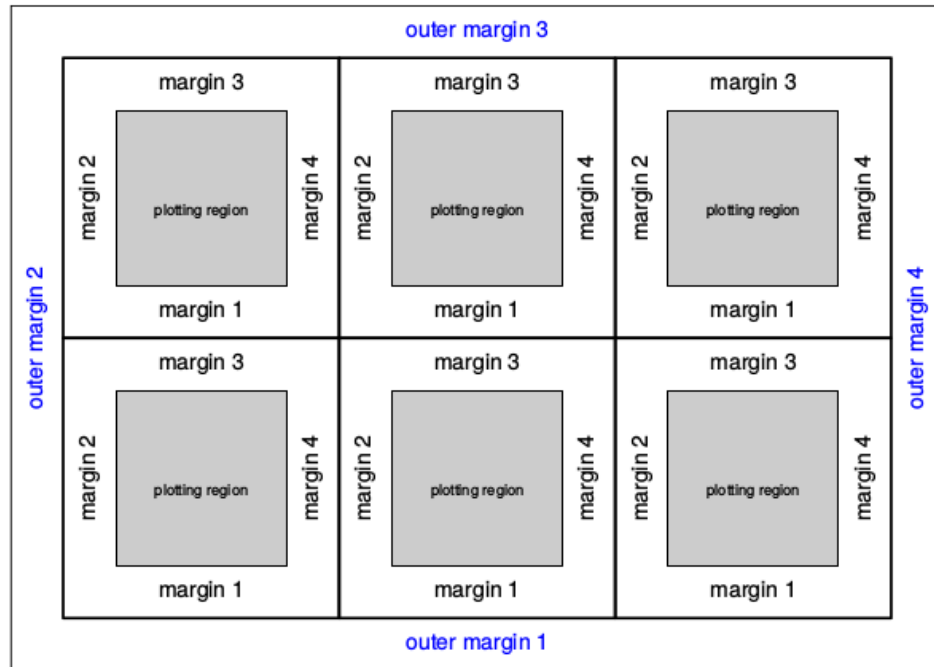
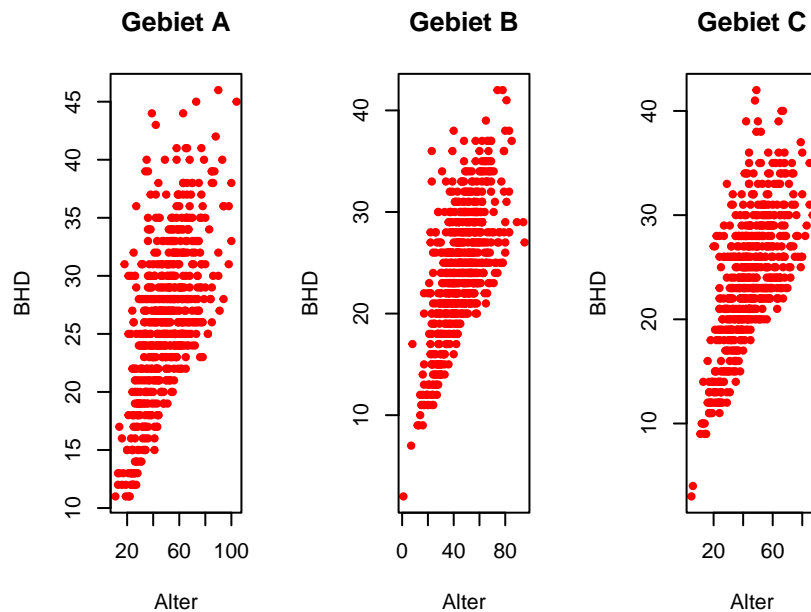



Abbildung 6: Schematischer Aufbau mehrere Diagramme in einem plot am Beispiel einer 3 x 2 Grafik.



872

873 Vergessen Sie nicht am Ende nochmals `par(mfrow = c(1, 1))` zu setzen, damit wieder nur ein Plot angezeigt
 874 wird.

875 8.1.2 Speichern von Abbildungen

876 Wenn nicht anders angegeben, wird die Abbildung zunächst nur in der RStudio Grafik Schnittstelle abgebildet
 877 (rechts unten). Von dort aus kann die Abbildung exportiert werden. Es bietet sich jedoch an das Speichern der
 878 Abbildung direkt im Code zu programmieren. Mögliche Formate die Abbildung als Vektorgrafik zu speichern

879 sind

- 880 • `pdf()` oder
- 881 • `postscript()`.

882 Beispiele für Rastergrafiken sind

- 883 • `png()`,
- 884 • `bmp()` oder
- 885 • `jpeg()`.

886 Die Grafikschnittstelle ist dann Ihre “Leinwand”. Mit dem Befehl `dev.off()` trennen Sie die Verbindung zur
 887 Schnittstelle wieder. Ihre “Leinwand” wird also wieder geschlossen. So lange die Schnittstelle geöffnet ist
 888 werden alle Low-Level Befehle an die Ausgabedatei gesendet. Hier am Beispiel einer PDF.

```
pdf("Grafik.pdf", height = 5)           # Öffnen der PDF Schnittstelle
plot(bhd ~ alter, type = "b", axes = FALSE, # Abbildung produzieren, Ohne Achsen
     data = dat)
axis(side = 1, line = 1)                # Achsen als High-Level Funktion hinzufügen
axis(side = 2, line = 1, las = 2)       # Sehen Sie selbst in der Hilfe was las bedeutet
dev.off()                               # Schnittstelle schließen
```

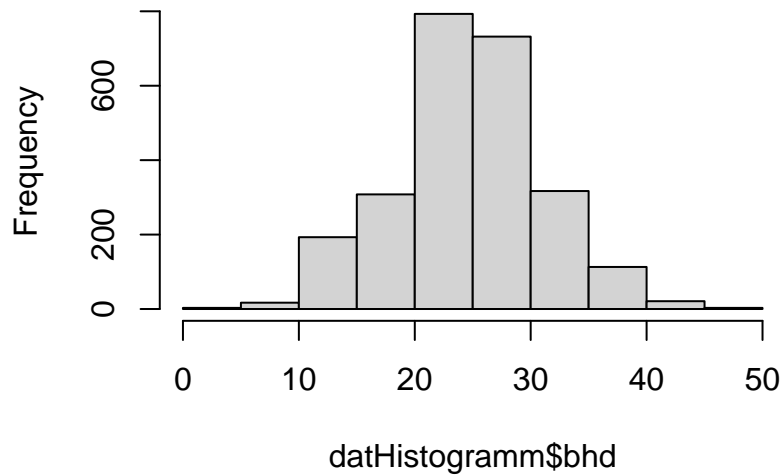
889 *Achtung*, wenn Sie die Funktion `dev.off()` nicht aufrufen, werden alle nachfolgenden Plots in die gleiche
 890 Datei geschrieben. Falls Sie nach einem Versuch einen Plot zu speichern plötzlich keine weiteren Plots mehr
 891 sehen, führen Sie einige Mal die Funktion `dev.off()` aus.

892 8.2 Histogramme

893 Neben den Streudiagrammen (*Scatterplots*, oder auch einfach x-y Diagramm) sind *Histogramme* in der
 894 angewandten Datenanalyse ein weiterer wichtiger Abbildungstyp. An Histogrammen wird die Häufigkeit
 895 von Beobachtungen nach Gruppen dargestellt. Sie sind deshalb so wichtig, weil man aus ihnen relevante
 896 Informationen über die Verteilung der Daten ablesen kann. So werden auf einen Blick der Zusammenhang
 897 von Beobachtungshäufigkeit und Streuung deutlich, sowie auch die Form der Verteilung und ihre Schiefe. Die
 898 Interpretation werden wir bei den Boxplots noch weiter vertiefen.

```
datHistogramm <- read.table("data/bhd_1.txt", header = TRUE)
# Über alle Baumarten
hist(datHistogramm$bhd)
```

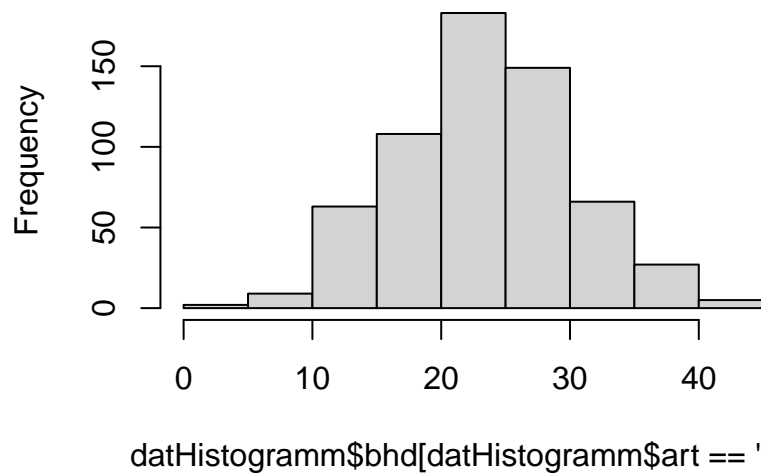
Histogram of datHistogramm\$bhd



899

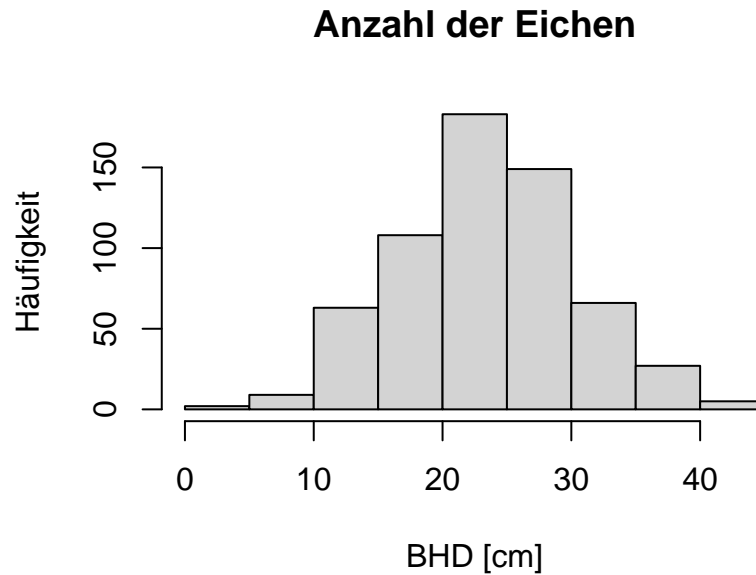
```
# Nur für Eichen, Standardeinstellungen
hist(datHistogramm$bhd[datHistogramm$art == "EI"])
```

Histogram of datHistogramm\$bhd[datHistogramm\$art == "EI"]



900

```
# Nur für Eichen, geänderte High-Level Einstellungen
hist(datHistogramm$bhd[datHistogramm$art == "EI"],
     xlab = "BHD [cm]", ylab = "Häufigkeit",
     main = "Anzahl der Eichen")
```



901

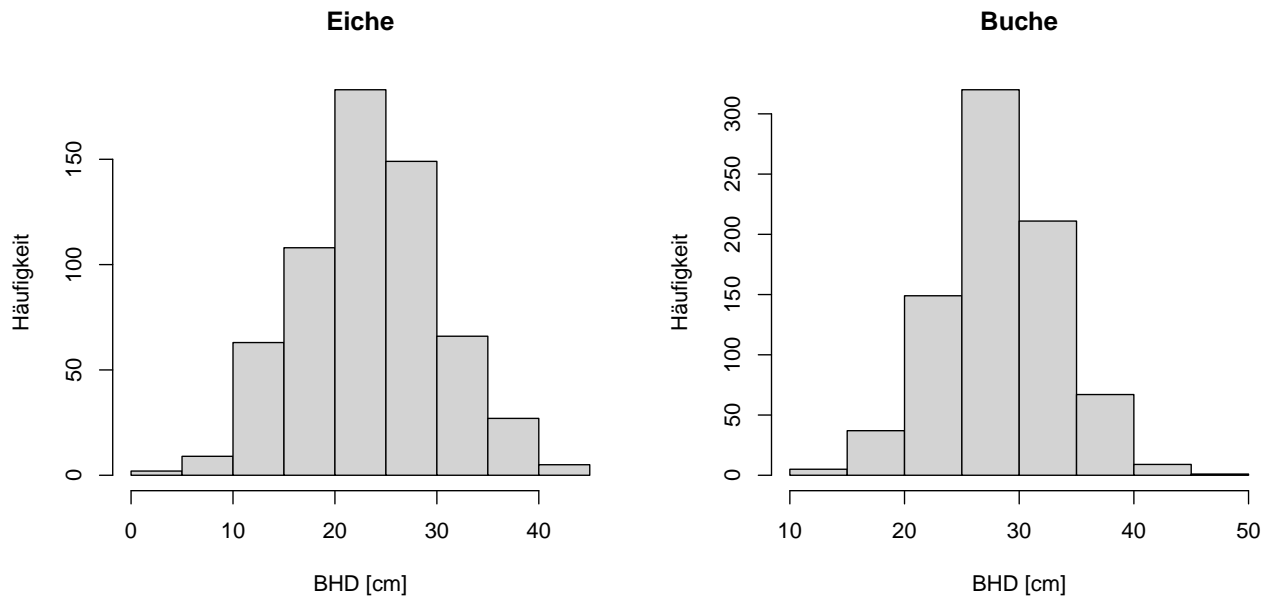
902 Eichen und Buchen im 2x1 Plot nebeneinander.

```

par(mfrow = c(1, 2))
hist(datHistogramm$bhd[datHistogramm$art == "EI"],
     xlab = "BHD [cm]", ylab = "Häufigkeit",
     main = "Eiche")
hist(datHistogramm$bhd[datHistogramm$art == "BU"],
     xlab = "BHD [cm]", ylab = "Häufigkeit",
     main = "Buche")

```

903



904

```

par(mfrow = c(1, 1)) # Alte Grafikeinstellungen wiederherstellen

```

8.3 Boxplots

Oft möchte man die Verteilung einer stetigen Variablen in Abhängigkeit einer diskreten Variable beschreiben oder Visualisieren. Ein Beispiel dafür wäre die BHD-Verteilung für unterschiedliche Baumarten. Eine häufige Darstellungsform für solche Daten sind *Boxplots*.

Boxplots bestehen aus drei Komponenten:

1. Eine *Box*, die den Bereich zwischen 0.25 und 0.75 Perzentil abdeckt, diese Distanz wird auch die IQR (Interquartile Range), bezeichnet. Zusätzlich wird die Box durch den Median (als dicke horizontale Linie) unterteilt.
2. Einzelne Punkte Ausreißer. Als Ausreißer werden Punkte bezeichnet, die $> 1.5IQR$ vom unteren oder oberen Ende der Box entfernt sind.
3. Eine senkrechte Linie von jeder Seite der Box bis zum letzten "Nicht-Ausreißer-Punkt". Diese Linie wird auch oft als *Whisker* bezeichnet.

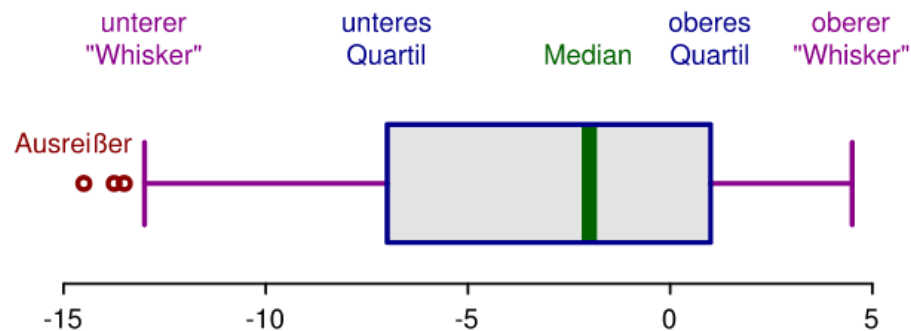
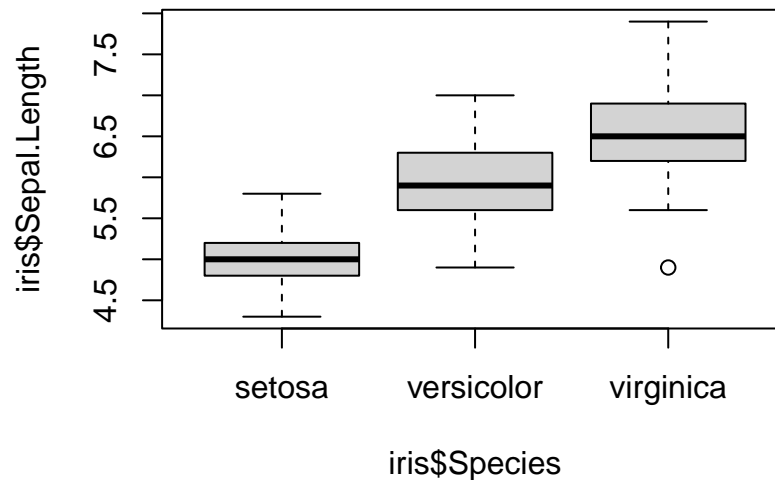


Abbildung 7: Schematische Darstellung eines Boxplots (Quelle: Von RobSeb - Eigenes Werk, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=14697172>).

Mit R kann mit der Funktion `boxplot()` ein Boxplot erstellt werden. Diese Funktion kann in zwei unterschiedlichen Ausprägungen verwendet werden.

1. `boxplot(x)` erzeugt einen Boxplot für die Variable `x`.
2. `boxplot(x ~ y)` erzeugt einen oder mehrere Boxplots für `x` aber gruppiert nach `y`, dabei sollte `y` eine kategorische Variable sein. `x` und `y` können auch die Spaltennamen eines `data.frames` sein, dann muss das `data.frame` mit dem Argument `data` zusätzlich übergeben werden.

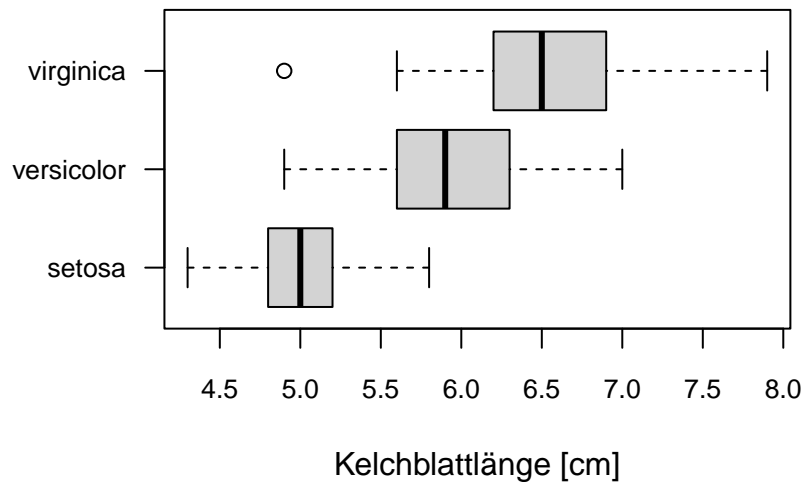
```
boxplot(iris$Sepal.Length ~ iris$Species)
```



923

924 Etwas eleganter ist es wenn wir das Argument `data` verwenden und den Plot etwas anpassen.

```
boxplot(
  Sepal.Length ~ Species, data = iris, ylab = NULL, xlab = "Kelchblattlänge [cm]",
  horizontal = TRUE, las = 1, cex.axis = 0.8
)
```



925

926

927 **Aufgabe 18: Boxplots**

928

- 929 • Lesen Sie erneut den Datensatz `daten/bhd_1.txt` ein und speichern Sie diesen in die Variable (`dat_bhd`).
- 930 • Wie viele BHD-Messungen gibt es für jedes Gebiet?
- 931 • Erstellen Sie für jedes Gebiet einen Plot

932 Erstellen Sie einen Plot mit 3 Subplots, jeweils mit einem Boxplot für die ersten drei Studiengebiete, in dem
 933 der BHD für jede Baumart dargestellt wird.

8.4 ggplot2: Eine Alternative für Abbildungen

`ggplot2` ist ein alternatives Plotting-System in *R*. Sie können mit `ggplot2` also grundsätzlich Abbildungen mit dem selben Inhalt erstellen wie mit Base Plots. Die Syntax und die optische Darstellung unterscheiden sich jedoch grundsätzlich. `ggplot2` basiert auf den *grammar of graphics* von Leland Wilkinson. Die Idee ist, alle nötigen Informationen der Abbildung miteinander zu verknüpfen. `ggplot2` ist also diametral zu Base Plots. Mit diesen gebündelten Informationen kann `ggplot2` die Abbildung automatisch verschönern. So werden bspw. die Legenden automatisch erzeugt und auch die Formatierungen automatisch an die Datenlage angepasst. `ggplot2` nimmt der*dem Entwickler*in also Arbeit ab. Dadurch sind die Abbildungen schon ohne viel Nacharbeit schick. Nachteil ist, dass der*dem Entwickler*in weniger Möglichkeiten zur Einstellung zur Verfügung stehen und nuterspezifische Sonderwünsche somit schwerer umsetzbar sind. Sehen Sie sich das *Cheatsheet* zu `ggplot2` an. Es ist in RStudio unter `Help >> Cheatsheets` zu finden.

Bei `ggplot2` sind Anweisungen zu den Daten und Anweisungen zur Darstellung voneinander getrennt. Die Daten werden in den Ästhetikbefehl übergeben und dort klassifiziert. Dann folgen die Darstellungsanweisungen. Ähnlich wie bei Base Plots werden die Grafikelemente ebenenweise nacheinander programmiert, jedoch mit einem `+` verbunden. Und hier liegt der wesentliche Unterschied zu Base Plots. Durch die `+` werden die Ebenen zu einem Befehl verbunden und damit gleichzeitig erstellt.

Die Erweiterung wird zunächst geladen⁷. Falls nicht schon geschehen, muss sie vorher installiert werden. Wir laden außerdem den Datensatz `iris`. Der Datensatz ist in *R* fest integriert. Siehe `?iris` für mehr Informationen.

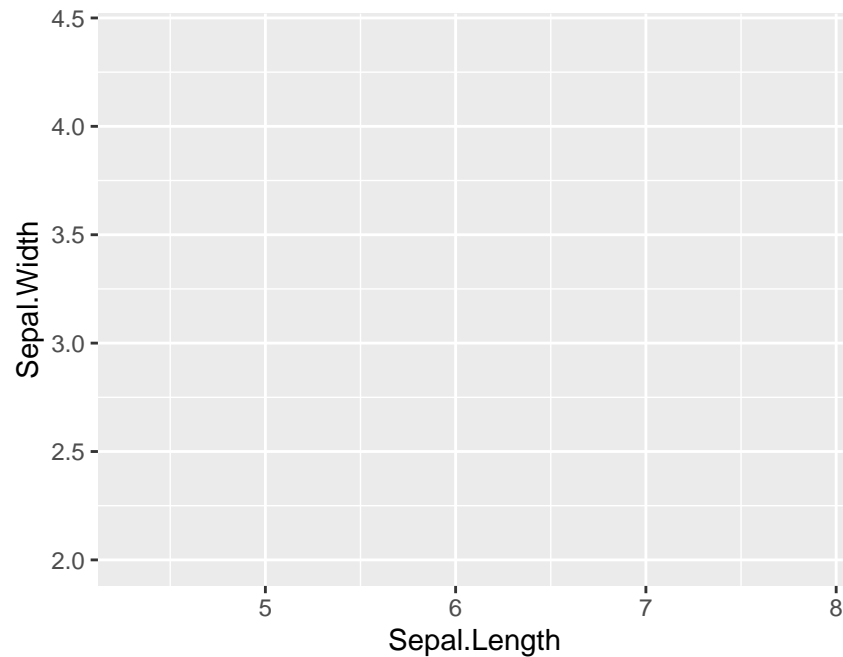
```
library(ggplot2)
head(iris)
```

```
##      Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1           5.1         3.5          1.4          0.2   setosa
## 2           4.9         3.0          1.4          0.2   setosa
## 3           4.7         3.2          1.3          0.2   setosa
## 4           4.6         3.1          1.5          0.2   setosa
## 5           5.0         3.6          1.4          0.2   setosa
## 6           5.4         3.9          1.7          0.4   setosa
```

Die Ästhetik wird bspw. folgendermaßen definiert.

```
ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width))
```

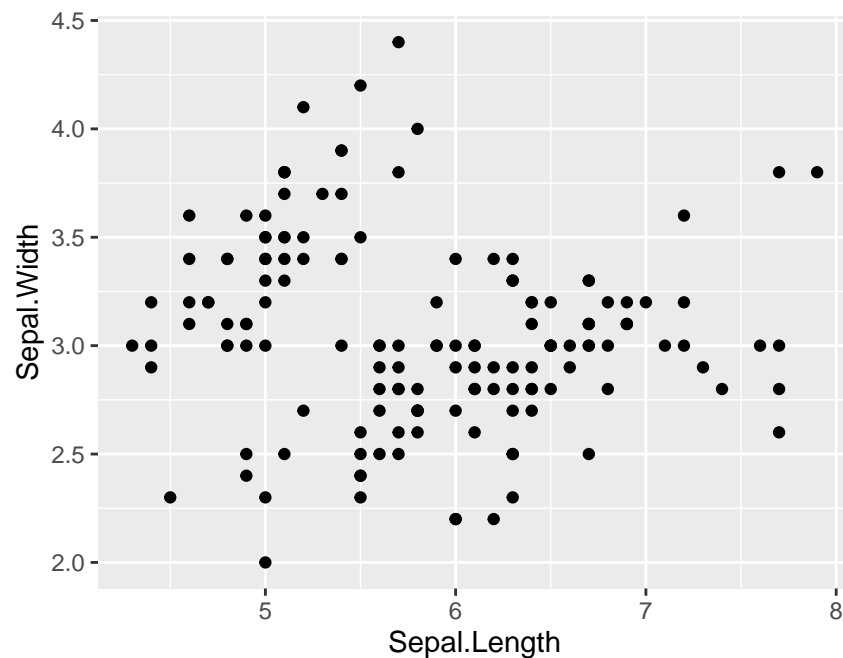
⁷Wir haben bis jetzt immer nur mit *base R* gearbeitet. D.h. wir haben nur Funktionen verwendet, die *R* bereits zur Verfügung stellt. Eine der großen Stärken von *R* sind die Erweiterungen (oder auch Pakete genannt). `ggplot2` ist so eine Erweiterung, die einmal mit `install.packages("ggplot2")` installiert werden muss. Danach muss man das Paket am Anfang jeder Session mit `library(ggplot2)` laden, damit die Funktionen aus dem Paket zur Verfügung stehen.



961

962 Dieser Befehl zeichnet noch keine Daten. Die Daten werden lediglich herangezogen, um einen leeren Plot für
963 die Daten zu erstellen. In dem Beispiel wird die Variable `Sepal.Length` aus dem `data.frame iris` als x und
964 `Sepal.Width` als y Variable definiert. Diese Informationen stehen den folgenden Layern nun zur Verfügung,
965 sodass nach den `+` nur noch x und y verwendet werden müssen. Um bspw. einen Scatterplot zu erstellen
966 wird ein `geom_point()` Layer hinzugefügt. x und y werden automatisch an `geom_point` übergeben. Weitere
967 Einstellungen sind in diesem Beispiel nicht notwendig, wären jedoch möglich. Siehe `?geom_point()`.

```
ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width)) + geom_point()
```



968

969

Aufgabe 19: Abbildungen mit ggplot2

Verwenden Sie die Daten aus Aufgabe 16 und erstellen Sie einen Scatterplot mit `ggplot2` wie in Aufgabe 16.

973

Wir haben mit der Funktion `geom_point()` dem Plot eine Punktgeometrie hinzugefügt. Es gibt noch viele weitere Geometrien. Die wichtigsten sind:

- `geom_line()` für eine Linie.
- `geom_histogram()` um ein Histogramm zu erstellen.
- `geom_boxplot()` um einen Boxplot zu erstellen.
- `geom_bar()` um ein Säulendiagramm zu erstellen.

Welche Geometrie die richtige ist, richtet nach dem Typ der darzustellenden Variablen. Beispielsweise bietet sich `geom_point()` an, wenn man zwei kontinuierliche Variable darstellen möchte. Wenn man hingegen die Verteilung von einer kontinuierlichen Variable darstellen möchte, dann bietet sich ein Histogramm (`geom_histogram()`) oder auch eine geschätzte Dichte (z.B. `geom_density()`) an.

984

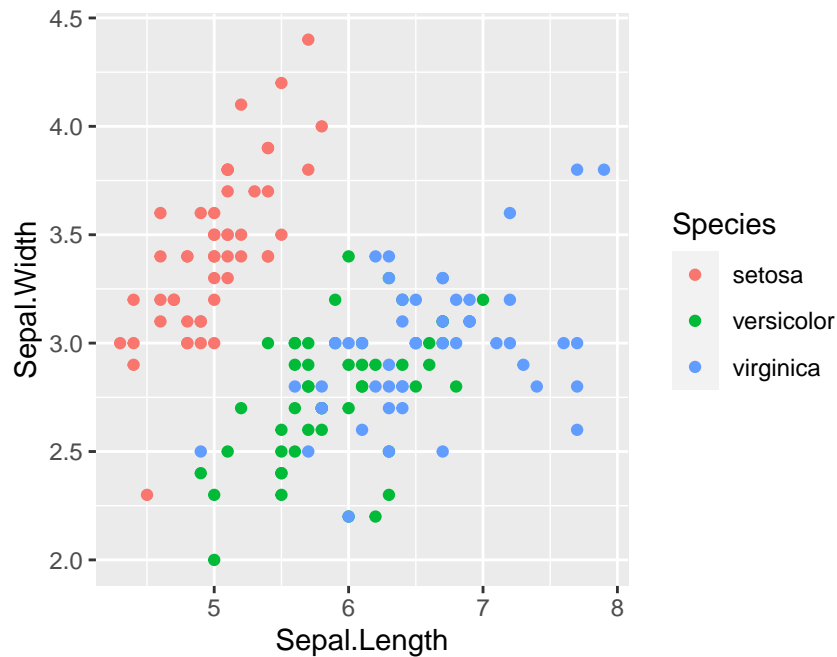
Aufgabe 20: Abbildungen mit ggplot2

Verwenden Sie die den Iris Datensatz und erstellen Sie mit `ggplot2` einen Plot der die Verteilung der Länge der Kelchblätter zeigt (Spalte `Sepal.Length`).

989

Eine der Stärken von `ggplot2` ist, dass man den Wert unterschiedlicher Variable auf unterschiedlichen Komponenten des Plots abbilden kann. Wir haben bis jetzt ein bzw. zwei Variable auf der x- und y-Achse abgebildet. Wir können aber ein weitere Variablen verwenden um das Aussehen des Plots zu beeinflussen. Beispielsweise können wir die Farbe der Punkte (für `geom_point()`) mit dem Argument `col` beeinflussen.

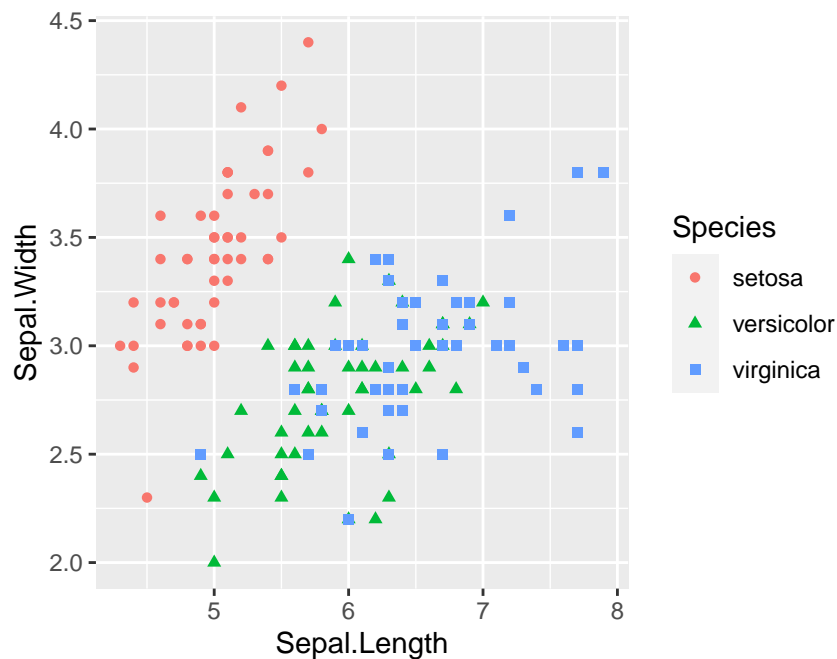
```
ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width, col = Species)) +  
  geom_point()
```



994

995 Somit bekommt jede Irisart eine eigene Farbe⁸. Gleichmaßen können wir die Punktart (`shape`), die
 996 Punktgröße (`size`) etc. anpassen.

```
ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width,
                 col = Species, shape = Species)) +
  geom_point()
```



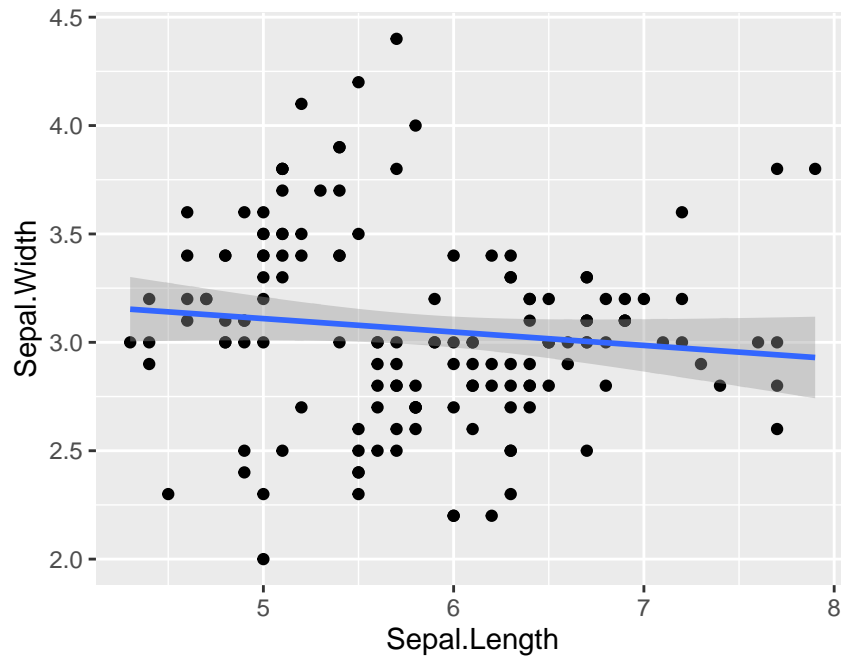
997

998 In dem Plot ist die Information zu der Art redundant (einmal als Farbe und einmal Symbolart).

999 Ein weitere sehr nützliche Geometrie ist `geom_smooth()`, die es erlaubt eine Trendlinie hinzuzufügen.

⁸Natürlich könnte man auch die Farbe anpassen.

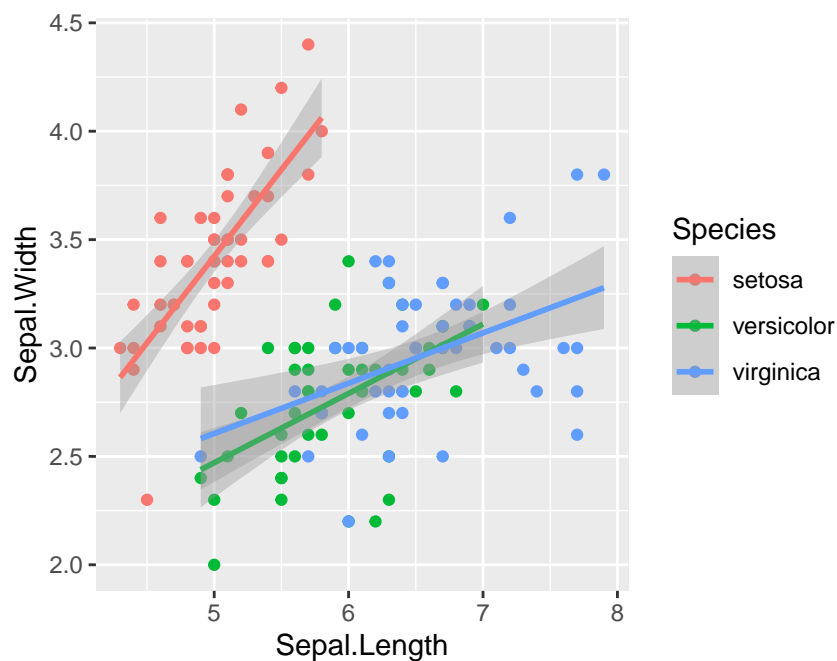
```
ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width)) +  
  geom_point() + geom_smooth(method = "lm")
```



1000

1001 Mit `method = "lm"` wird festgelegt, dass die Trendlinie gerade sein soll (es wird eine lineare Einfachregression
 1002 angepasst). Wenn wird wieder eine gruppierende Variable einführen (z.B. die Beobachtungen nach Art auf
 1003 die Farbe aufteilen), wird das von `geom_smooth()` berücksichtigt.

```
ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width, col = Species)) +  
  geom_point() + geom_smooth(method = "lm")
```



1004

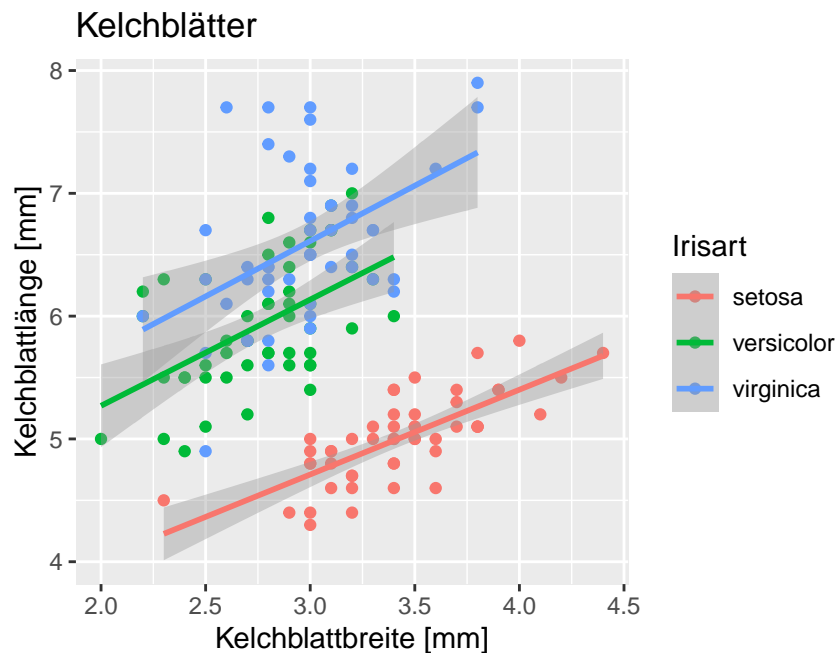
Aufgabe 21: Anpassen von Plots

Lesen Sie den Datensatz `data/bhd_1.txt` ein und erstellen Sie einen Boxplot für die Verteilung des BHDs für jede Baumart. In einem zweiten Schritt verwenden Sie erst `col = gebiet` und dann `fill = gebiet`. Welchen Unterschied stellen Sie fest?

```
dat <- read.table("data/bhd_1.txt", header = TRUE)
head(dat)
ggplot(dat, aes(art, bhd, fill = gebiet)) + geom_boxplot()
```

Mit der Funktion `labs()` werden die Beschriftungen geändert.

```
ggplot(iris, aes(x = Sepal.Width, y = Sepal.Length, color = Species)) +
  geom_point() + geom_smooth(method = "lm") +
  labs(x = "Kelchblattbreite [mm]", y = "Kelchblattlänge [mm]",
       title = "Kelchblätter", color = "Irisart")
```

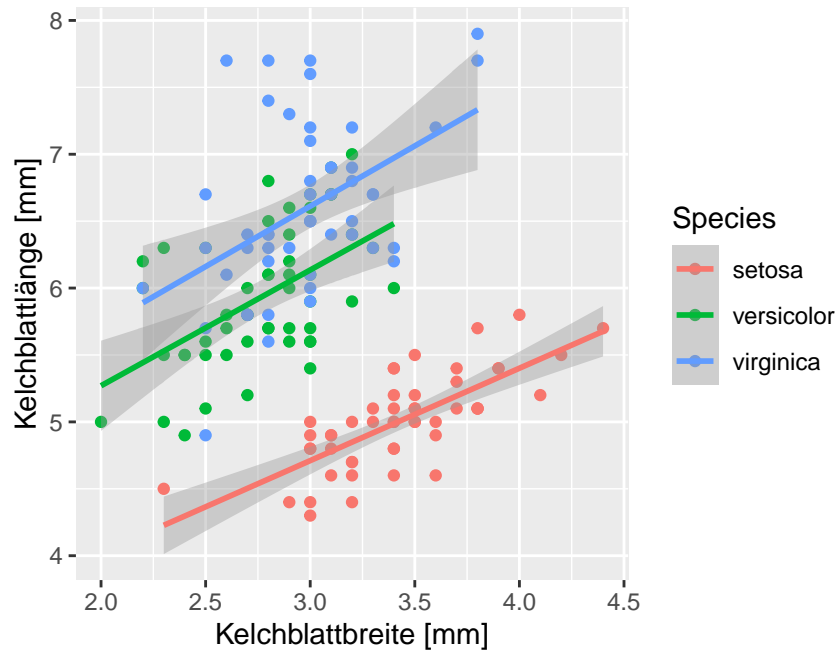


Statt einen langen Befehl zu tippen, kann ein `ggplot()` auch zwischengespeichert und wieder aufgerufen bzw. angepasst werden. Das ist vor allem sinnvoll, wenn mehrere Abbildungen auf dem selben Zwischenergebnis aufbauen sollen.

```
p1 <- ggplot(iris, aes(x = Sepal.Width, y = Sepal.Length, color = Species)) +
  geom_point() + geom_smooth(method = "lm")
```

Wir können jetzt mit `p1` weiter arbeiten und beispielsweise eine Beschriftung hinzufügen.

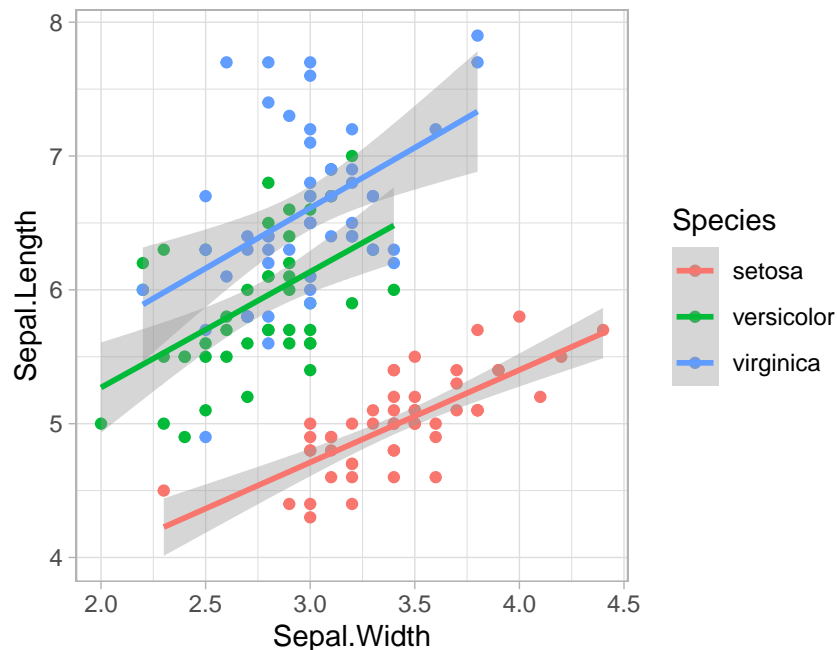
```
p1 + labs(x = "Kelchblattbreite [mm]", y = "Kelchblattlänge [mm]")
```



1018

1019 Oder auch den ganzen Plot anpassen. Dafür gibt es *themes*. Es gibt eine Reihe von vorgefertigten *themes*
 1020 oder man kann diese auch selber erstellen (das ist aber nicht Teil dieses Kurses).

```
p1 + theme_light()
```



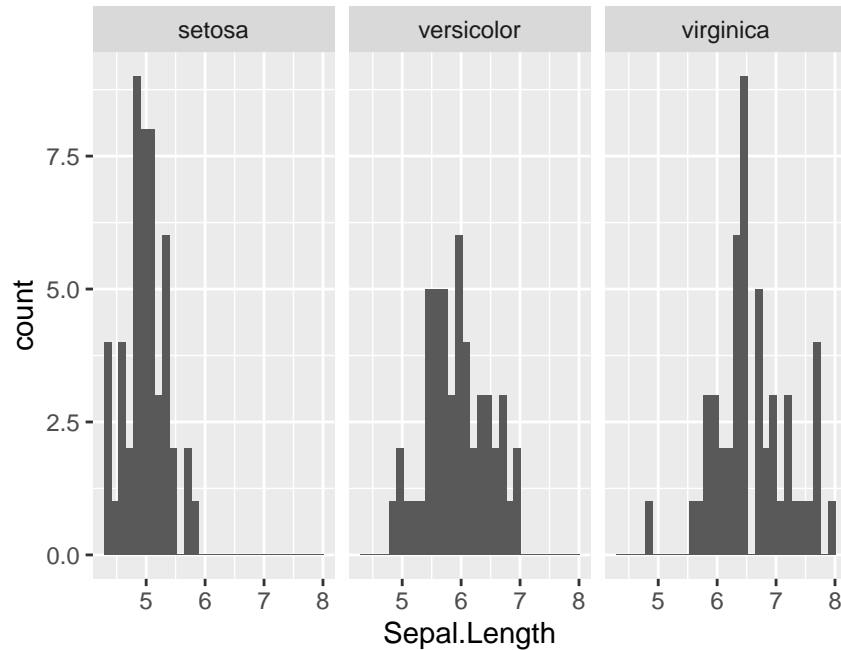
1021

1022 Weitere *themes* sind: `theme_bw()`, `theme_linedraw()` oder `theme_dark()`. Es gibt extra Pakete die viele
 1023 zusätzliche weitere *themes* an bieten. Dazu gehört z.B. das Paket [ggthemes](#).

8.4.1 Multipanel Abbildungen

Mit ggplot2 kann man einfach Abbildungen erstellen, die mehrere Panels haben. Das bedeutet, dass eine oder mehrere weitere Variablen gibt, die einen Plot in mehrere Subplots teilt. Dafür gibt es zwei Funktionen: `facet_grid()` und `facet_wrap()`.

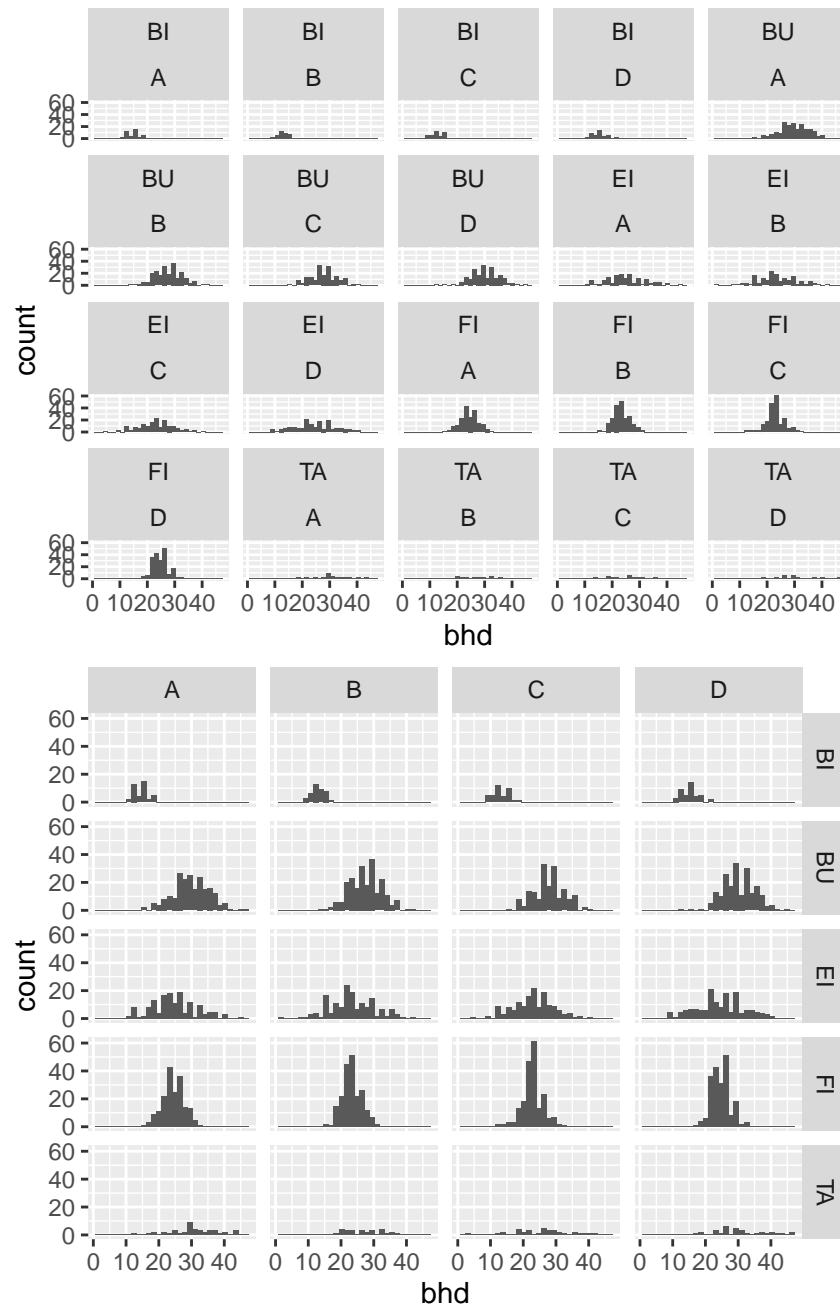
```
ggplot(iris, aes(Sepal.Length)) + geom_histogram() +  
  facet_grid(~ Species)
```



Die Funktion `facet_grid()` erzeugt eine *Grid*, während `facet_wrap()` für jedes Panel eine eigene Überschrift erzeugt.

Aufgabe 22: Multipanel Abbildungen

Lesen Sie erneut den Datensatz `daten/bhd_1.txt` ein und speichern Sie diesen in die Variable (`dat_bhd`). Erstellen Sie für jede Art und Gebiet ein Histogramm. Welche Unterschiede können Sie feststellen, wenn Sie `facet_grid()` oder `facet_wrap()` verwenden?



8.4.2 Plots kombinieren

Es gibt Situationen in denen **unterschiedliche** Plots miteinander kombiniert werden müssen. Im vorherigen Abschnitt wurde dies immer anhand einer gruppierenden Kovariate gemacht. Aber es gibt auch Situationen, in denen das nicht möglich ist. Beispielsweise wenn ein Histogramm und ein Scatterplot vom gleichen Datensatz zusammengefasst werden sollen. Dafür bietet sich das Paket **patchwork** an⁹.

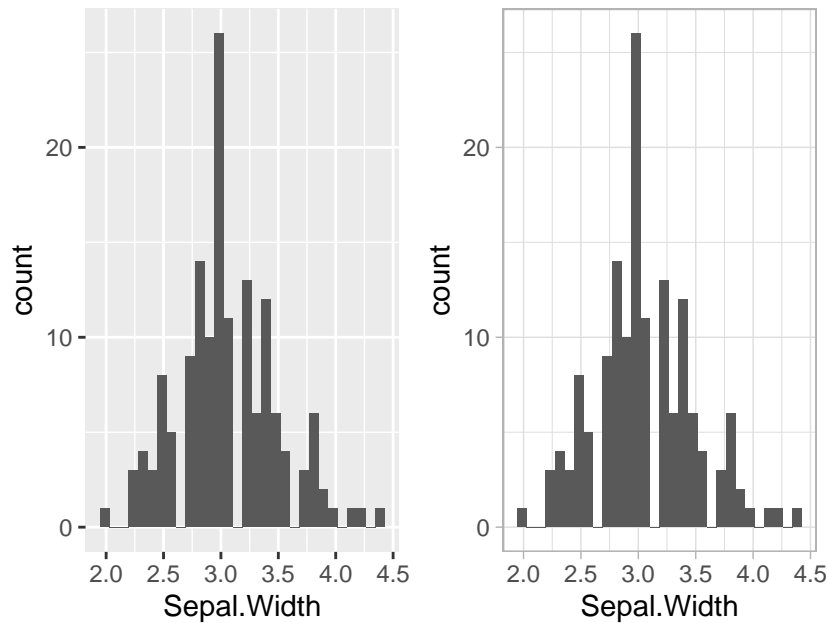
Als erstes können wir zwei (oder natürlich auch mehrere Plots) erstellen. Hier unterscheiden sich die Plots lediglich durch das Aussehen.

⁹Auch dieses Paket müssen Sie einmalig mit `install.packages("patchwork")` installieren.

```
p1 <- ggplot(iris, aes(Sepal.Width)) + geom_histogram()
p2 <- p1 + theme_light()
```

1046 Dann müssen können wir diese Plots ebenfalls mit + zusammenfügen.

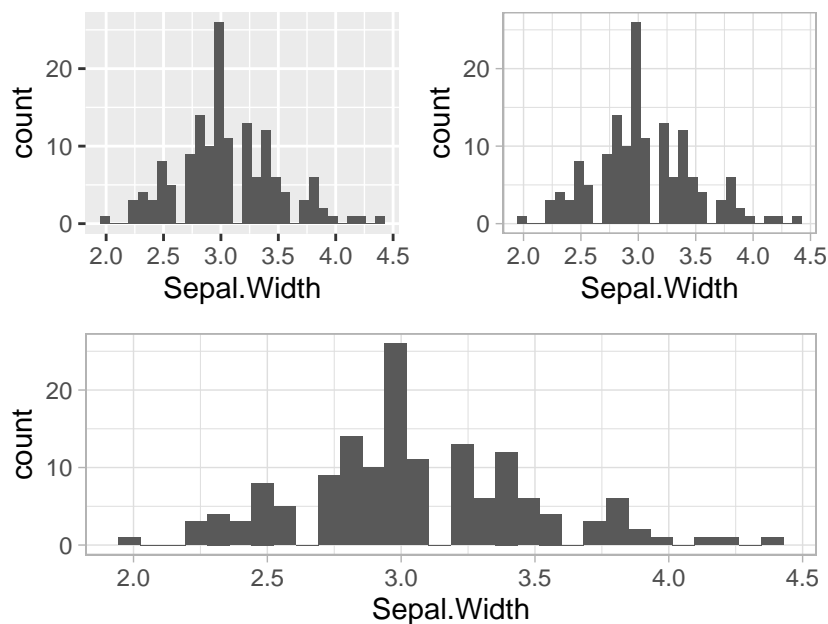
```
library(patchwork)
p1 + p2
```



1047

1048 Natürlich können auch weitere Plots hinzugefügt werden (auch in unterschiedlichen Dimensionen):

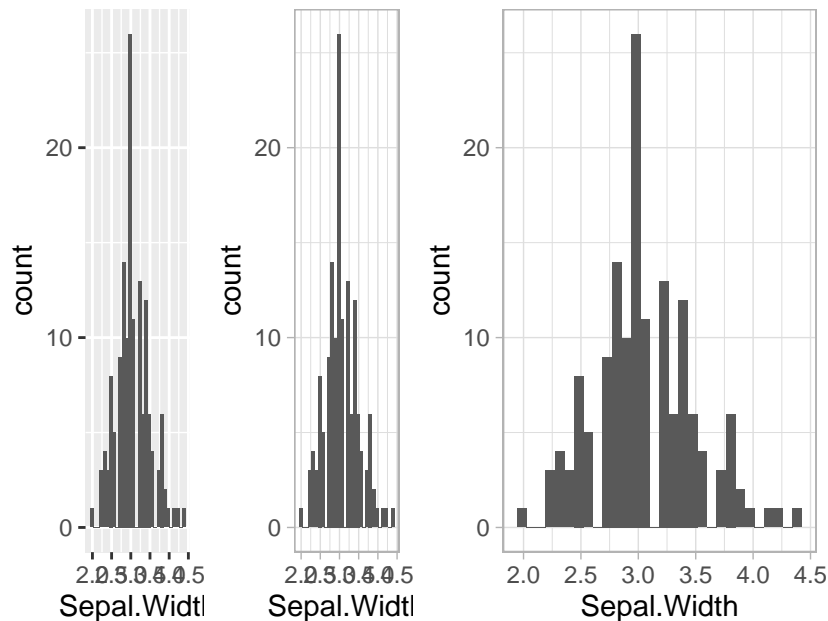
```
(p1 + p2) / p2
```



1049

1050 Des weiteren können mit | auch Plots gegenüber gestellt werden.

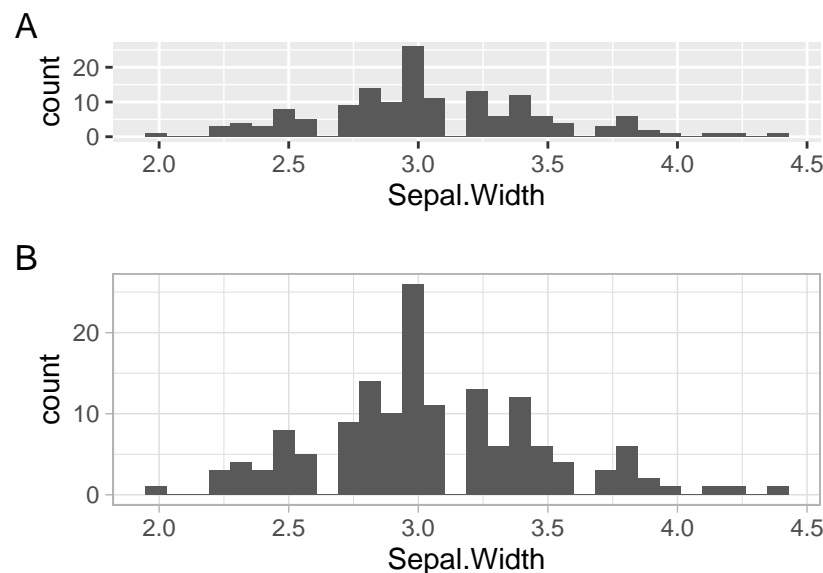

```
(p1 + p2) | p2
```



Weitere Optionen können mit `plot_layout()` und `plot_annotation()` angepasst werden. Mit `plot_layout()` kann die Anordnung der Plots bestimmt werden (z.B. über die Argument `nrow` und `ncol`), sowie deren *relative* Größe (über die Argumente `widths` und `heights`). Mit der Funktion `plot_annotation()` können zusätzliche Beschriftungen hinzugefügt werden, wie beispielsweise eine Titel (Argument `title`) oder ein Buchstabe/Zahl für jedes Element (Argument `tag_levels`).

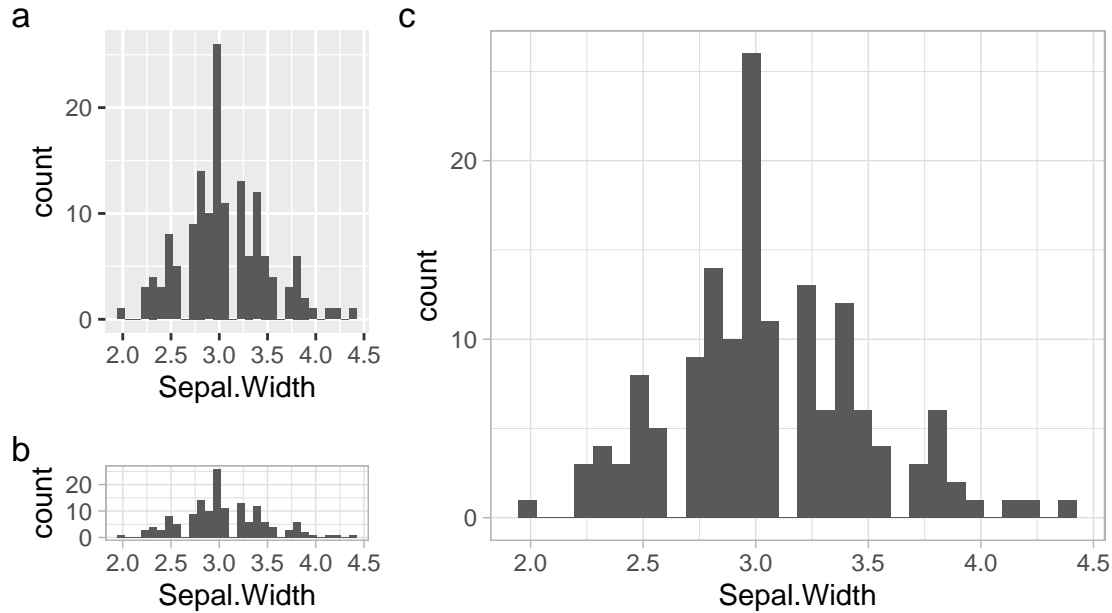
```
p1 + p2 +
  plot_layout(ncol = 1, heights = c(0.3, 0.7)) +
  plot_annotation(title = "Zwei Histogramme", tag_levels = "A")
```

Zwei Histogramme



Aufgabe 23: Mehrere Plots zusammenfügen

Versuchen Sie die folgende Zusammenstellung der Plots nachzumachen:

**8.4.3 Speichern von plots**

Sie können mit `ggsave()` eine zwischen-gespeicherte Abbildung exportieren, indem Sie den Variablennamen übergeben. Wenn Sie keine Variable übergeben, wird automatisch die letzte Abbildung gespeichert. Das Dateiformat wird aus dem Dateinamen übernommen.

```
ggsave("letzteAbb.png")
ggsave(p1, "zwischengespeicherteAbb.png")
```

9 Mit Daten arbeiten

9.1 dplyr eine Einführung

dplyr ist eine Erweiterung von R (= Paket), die das Ziel hat den Umgang mit Daten einfacher und schneller zu machen.

dplyr definiert 5 Verben, um mit Daten zu arbeiten. Diese sind:

- filter
- select
- arrange
- mutate
- summarise

```
dat <- data.frame(id = 1:5,
                  plot = c(1, 1, 2, 2, 3),
                  bhd = c(50, 29, 13, 23, 25),
                  alter = c(10, 30, 31, 24, 25))
```

Damit die Funktionen aus dplyr verwendet werden können, müssen wir als erstes das Paket dplyr laden.

```
library(dplyr)
```

Sollte dies zu einer Fehlermeldung führen, dann müssen Sie das Paket dplyr erst installieren. Dafür müssen Sie **einmalig** `install.packages("dplyr")` installieren.

dplyr stellt unterschiedliche Funktionen zum Arbeiten mit Daten zur Verfügung. Es gibt fünf Grundfunktionen für die am häufigsten vorkommenden Operationen. Mit der Funktion `filter()` können unterschiedliche Beobachtungen gefiltert werden:

```
filter(dat, bhd > 10)
```

```
##   id plot bhd alter
## 1  1    1  50    10
## 2  2    1  29    30
## 3  3    2  13    31
## 4  4    2  23    24
## 5  5    3  25    25
```

Es können auch mehrere Spalten verwendet werden.

```
filter(dat, bhd > 10, bhd < 40)
```

```
##   id plot bhd alter
## 1  2    1  29    30
## 2  3    2  13    31
## 3  4    2  23    24
## 4  5    3  25    25
```

Natürlich kann genau das gleiche Ergebnis mit dem ‘normalen’ R erreicht werden, dies wäre dann:

```
dat[dat$bhd > 10 & dat$bhd < 40, ]
```

```
1096 ##   id plot bhd alter
1097 ## 2  2   1  29   30
1098 ## 3  3   2  13   31
1099 ## 4  4   2  23   24
1100 ## 5  5   3  25   25
```

1101 Eine weitere Funktion aus dem Paket `dplyr` ist `select()`. Damit können Spalten aus einem `data.frame`
 1102 ausgewählt werden. Dabei können auch die Spaltennamen unbenannt werden.

```
select(dat, bhd)
```

```
1103 ##   bhd
1104 ## 1  50
1105 ## 2  29
1106 ## 3  13
1107 ## 4  23
1108 ## 5  25
```

```
select(dat, bhd, id)
```

```
1109 ##   bhd id
1110 ## 1  50  1
1111 ## 2  29  2
1112 ## 3  13  3
1113 ## 4  23  4
1114 ## 5  25  5
```

```
select(dat, BHD = bhd, id)
```

```
1115 ##   BHD id
1116 ## 1  50  1
1117 ## 2  29  2
1118 ## 3  13  3
1119 ## 4  23  4
1120 ## 5  25  5
```

1121 Mit der Funktion `arrange()` können die Beobachtungen in einem `data.frame` sortiert werden.

```
arrange(dat, bhd)
```

```
1122 ##   id plot bhd alter
1123 ## 1  3   2  13   31
1124 ## 2  4   2  23   24
1125 ## 3  5   3  25   25
1126 ## 4  2   1  29   30
1127 ## 5  1   1  50   10
```

Mit der Funktion `desc()` kann die Anordnung in absteigender Reihenfolge sortiert werden.

```
arrange(dat, desc(bhd))
```

```
##   id plot bhd alter
## 1  1    1   50    10
## 2  2    1   29    30
## 3  3    2   25    25
## 4  4    2   23    24
## 5  5    3   13    31
```

Mit der Funktion `mutate()` kann man eine neue Spalte hinzufügen.

```
mutate(dat, bhd_mm = bhd * 10, fl = pi * (bhd/2)^2)
```

```
##   id plot bhd alter bhd_mm      fl
## 1  1    1   50    10    500 1963.4954
## 2  2    1   29    30    290  660.5199
## 3  3    2   25    25    312 132.7323
## 4  4    2   23    24    230  415.4756
## 5  5    3   13    31    130  490.8739
```

```
mutate(dat, mean_bhd = mean(bhd))
```

```
##   id plot bhd alter mean_bhd
## 1  1    1   50    10      28
## 2  2    1   29    30      28
## 3  3    2   25    25      28
## 4  4    2   23    24      28
## 5  5    3   13    31      28
```

Mit der Funktion `summarise()` können Spalten zusammengefasst werden.

```
summarise(
  dat,
  mean_bhd = mean(bhd),
  mean_sd = sd(bhd)
)
```

```
##   mean_bhd mean_sd
## 1      28 13.63818
```

Aufgabe 24: Datenmanipulation mit dplyr

1. Laden Sie den Datensatz `data/bhd_1.txt`
2. Berechnen Sie folgende Werte für alle Einträge und speichern Sie die Ergebnisse in `erg1`

- mittlerer bhd
- maximales alter
- die Standardabweichung des BHDs
- die Anzahl Bäume mit einem BHD > 30

9.2 Arbeiten mit gruppierten Daten

Zusätzlich können `mutate` und `summarise` auch auf gruppierte Daten angewendet werden. Dafür müssen wir erst die Funktion `group_by()` aufrufen und als Argumente die Variablen übergeben, die die Gruppen definieren.

```
dat1 <- group_by(dat, plot)
mutate(dat, bhd_m = mean(bhd)) # bhd über alle Bäume
```

```
##   id plot bhd alter bhd_m
## 1  1   1  50    10    28
## 2  2   1  29    30    28
## 3  3   2  13    31    28
## 4  4   2  23    24    28
## 5  5   3  25    25    28
```

```
mutate(dat1, bhd_m = mean(bhd)) # bhd pro Plot
```

```
## # A tibble: 5 x 5
## # Groups:   plot [3]
##       id plot   bhd alter bhd_m
##   <int> <dbl> <dbl> <dbl> <dbl>
## 1     1     1     50     10  39.5
## 2     2     1     29     30  39.5
## 3     3     2     13     31   18
## 4     4     2     23     24   18
## 5     5     3     25     25   25
```

```
summarise(dat, bhd_m = mean(bhd))
```

```
##   bhd_m
## 1     28
```

```
summarise(dat1, bhd_m = mean(bhd))
```

```
## # A tibble: 3 x 2
##   plot bhd_m
##   <dbl> <dbl>
## 1     1  39.5
## 2     2   18
## 3     3   25
```

Aufgabe 25: dplyr mit gruppierten Daten

1. Laden Sie den Datensatz `data/bhd_1.txt`
2. Berechnen Sie für jede Baumart und Standort die folgenden Werte:
 - mittlerer `bhd`
 - maximales `alter`
 - die Standardabweichung des BHDs
 - die Anzahl Bäume mit einem BHD > 30
3. Ordnen Sie das Ergebnis nach Baumart und BHD.

9.3 pipes oder %>%

Mit *Pipes* (`%>%`) kann man das Ergebnis einer Funktion einfach an eine nachfolgende Funktion weiterreichen.

```
a <- c(5, 3, 2, NA)
```

Wir kennen bis jetzt:

```
mean(na.omit(a))
```

```
## [1] 3.333333
```

Mit *Pipes*, die durch das Symbol `%>%` dargestellt werden¹⁰, können wir das etwas vereinfachen und nacheinander schreiben:

```
na.omit(a) %>% mean
```

```
## [1] 3.333333
```

Oder sogar

```
a %>% na.omit %>% mean
```

```
## [1] 3.333333
```

Aufgabe 26: Pipes %>%

Wiederholen Sie die letzte Aufgabe, aber diesmal ohne Zwischenergebnisse zu speichern:

1. Laden Sie den Datensatz `data/bhd_1.txt`.
2. Berechnen Sie für jede Baumart und Standort die folgenden Werte:
 - mittlerer `bhd`
 - maximales `alter`
 - die Standardabweichung des BHDs

¹⁰In RStudio kann `%>%` mit der Tastenkombination `Strg + Umschalt + m` (`(Strg) + ⬆ + m`) eingefügt werden.

- die Anzahl Bäume mit einem BHD > 30

3. Ordnen Sie das Ergebnis nach Baumart und BHD.

9.4 Joins

Eine weitere häufige Aufgabe beim Daten Management ist es Daten zusammenzuführen. Nehmen Sie an, dass wir folgende Aufnahmen gemacht haben

```
aufnahmen <- data.frame(
  id = 1:3,
  bhd = c(20, 31, 74)
)
```

und jeder Baum lediglich mit einer `id` versehen wurde. In einer zweiten Tabelle wurden dann weitere Daten zu Bäumen gespeichert (z.B. die Art, das Studiengebiet usw).

```
metadaten <- data.frame(
  id = 2:4,
  art = c("Ta", "Bu", "Bu"),
  gebiet = c("A", "B", "B")
)
```

Ziel ist es jetzt die Bäume aus `aufnahmen` mit den Informationen aus den `metadaten` zu verbinden. Dazu dient `id` als Bindeglied (oft auch Schlüssel genannt).

Dazu gibt es vier Möglichkeiten.

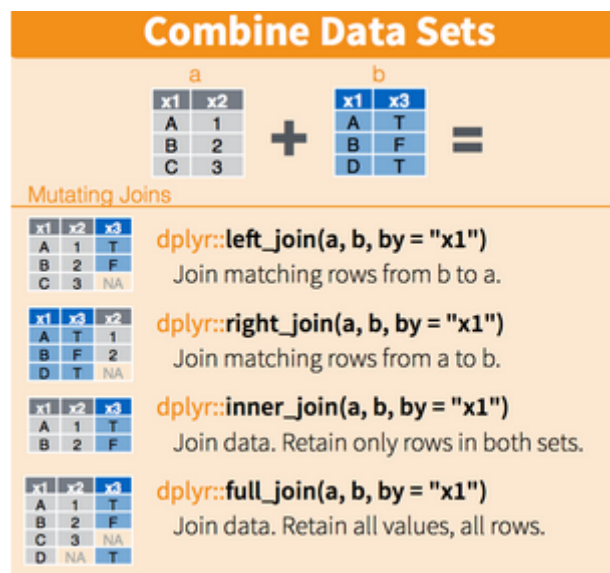


Abbildung 8: Joins (Quelle Rstudio)

Zur Durchführung gibt es in base R die Funktion `merge()`. Wir werden aber gleich die Funktionen aus dem Paket `dplyr` verwenden.


```
library(dplyr)
left_join(aufnahmen, metadaten, by = "id")
```

```
1227 ##   id bhd  art gebiet
1228 ## 1  1  20 <NA>   <NA>
1229 ## 2  2  31  Ta     A
1230 ## 3  3  74  Bu     B
```

```
right_join(aufnahmen, metadaten, by = "id")
```

```
1231 ##   id bhd art gebiet
1232 ## 1  2  31 Ta     A
1233 ## 2  3  74 Bu     B
1234 ## 3  4  NA Bu     B
```

```
inner_join(aufnahmen, metadaten, by = "id")
```

```
1235 ##   id bhd art gebiet
1236 ## 1  2  31 Ta     A
1237 ## 2  3  74 Bu     B
```

```
full_join(aufnahmen, metadaten, by = "id")
```

```
1238 ##   id bhd  art gebiet
1239 ## 1  1  20 <NA>   <NA>
1240 ## 2  2  31  Ta     A
1241 ## 3  3  74  Bu     B
1242 ## 4  4  NA  Bu     B
```

1243 by kann auch unterschiedliche Spalten in den beiden `data.frames` ansprechen:

```
metadaten <- data.frame(
  baum_id = 2:4,
  art = c("Ta", "Bu", "Bu"),
  gebiet = c("A", "B", "B")
)
```

```
left_join(aufnahmen, metadaten, by = c("id" = "baum_id"))
```

```
1244 ##   id bhd  art gebiet
1245 ## 1  1  20 <NA>   <NA>
1246 ## 2  2  31  Ta     A
1247 ## 3  3  74  Bu     B
```

1248

1249 *Aufgabe 27: Verbinden von Daten*

1250

- 1251 • Lesen Sie die Datensätze `daten/bhd_2.txt` und `daten/bhd_2_meta.txt` ein.

- Stellen Sie sicher, dass es für den `bhd` keine fehlenden Werte gibt (entfernen sie entsprechende Zeilen)
- Fügen Sie zu den Metadaten (gespeichert in `bhd_2_meta`) die Anzahl Bäume und den mittleren `bhd` hinzu pro Gebiet.

9.5 'long' and 'wide' Datenformate

Wickham (2014) propagiert das Prinzip von *tidy* Data. Nach diesem Prinzip sollten Daten wie folgt organisiert sein:

- Jede Zeile ist ein Merkmalsträger/Subjekt/Objekt (z.B. eine Person, ein Baum).
- Jede Spalte ist eine Variable (=Merkmal) die den Merkmalsträger beschreibt.
- Jede Zelle ist genau ein Wert (=Mermalausprägung), nämlich der Wert, der Variable für den Merkmals-träger.

Zum Beispiel enthalten Spaltennamen oft Informationen, die eigentlich in einer Variable gespeichert werden sollten. Folgendes Beispiel gibt die BHD Messung von 3 Bäumen in 3 Jahren wieder.

```
dat <- tibble(
  id = 1:3,
  bhd2015 = c(30, 31, 32),
  bhd2026 = c(31, 31, 33),
  bhd2017 = c(34, 32, 33)
)
```

Diese Daten sind jetzt im **wide**-Format gespeichert und nicht optimal, weil Information über die Daten (nämlich das Jahr der Aufnahme in den Spaltennamen gespeichert sind). Besser wäre eine Struktur mit nur drei Spalten: `id`, `jahr` und `bhd`. Um die Daten in so eine Struktur zu bringen, gibt es die Funktion `pivot_longer()` aus dem Paket `tidyr`.

```
library(tidyr)
dat1 <- pivot_longer(dat, cols = bhd2015:bhd2017)
dat1
```

```
## # A tibble: 9 x 3
##       id name    value
##   <int> <chr>   <dbl>
## 1     1  bhd2015    30
## 2     1  bhd2026    31
## 3     1  bhd2017    34
## 4     2  bhd2015    31
## 5     2  bhd2026    31
## 6     2  bhd2017    32
## 7     3  bhd2015    32
## 8     3  bhd2026    33
## 9     3  bhd2017    33
```

Wenn wir die Spalten für die Variable und den Wert sinnvoll benennen möchten, können wir das über die Argumente `names_to` und `value_to` machen.

```
dat1 <- pivot_longer(dat, cols = bhd2015:bhd2017, names_to = "jahr", values_to = "bhd")
dat1
```

```
1282 ## # A tibble: 9 x 3
1283 ##       id jahr      bhd
1284 ##   <int> <chr>   <dbl>
1285 ## 1     1 bhd2015    30
1286 ## 2     1 bhd2026    31
1287 ## 3     1 bhd2017    34
1288 ## 4     2 bhd2015    31
1289 ## 5     2 bhd2026    31
1290 ## 6     2 bhd2017    32
1291 ## 7     3 bhd2015    32
1292 ## 8     3 bhd2026    33
1293 ## 9     3 bhd2017    33
```

1294 Analog zu der Funktion `pivot_longer()` gibt es auch die Funktion `pivot_wider()`, um vom Daten vom
1295 long-Format ins wide-Format zu transformieren.

```
pivot_wider(dat1, names_from = jahr, values_from = bhd)
```

```
1296 ## # A tibble: 3 x 4
1297 ##       id bhd2015 bhd2026 bhd2017
1298 ##   <int>   <dbl>   <dbl>   <dbl>
1299 ## 1     1      30      31      34
1300 ## 2     2      31      31      32
1301 ## 3     3      32      33      33
```

1302

1303 *Aufgabe 28: Zeitliche Verlauf von BHDs*

1304

1305 In der Datei `bhd_3.csv` befinden sich gemessene BHDs (in cm) von unterschiedlichen Bäumen zu unter-
1306 schiedlichen Zeitpunkten. Erstellen Sie ein Liniendiagramm, das den zeitlichen (x-Achse) Verlauf der BHDs
1307 (y-Achse) für die unterschiedlichen Bäume darstellt.

1308 9.6 Auswählen von Variablen

1309 Sobald die Datensätze etwas umfangreicher werden (d.h. es gibt mehrere Spalten in einem `data.frame`),
1310 können innerhalb vieler `dplyr`-Funktionen spezielle Funktionen verwendet werden, um Variablen auszuwählen.

1311 Wenn die genaue Position der Spalten bekannt ist, kann man mit dem `:`-Operator und der Position Spalten
1312 auswählen:

```
iris %>% select(1:3) %>% head(3)
```

```
1313 ## Sepal.Length Sepal.Width Petal.Length
```

```

1314 ## 1      5.1      3.5      1.4
1315 ## 2      4.9      3.0      1.4
1316 ## 3      4.7      3.2      1.3

```

1317 Diese Vorgehensweise kann gefährlich sein, da sich manchmal Spalten verschieben und sich somit die
 1318 Positionen ändern. Eist besser Spalten immer explizit anzusprechen.

```
iris %>% select(Sepal.Length, Sepal.Width, Petal.Length) %>% head(3)
```

```

1319 ##      Sepal.Length Sepal.Width Petal.Length
1320 ## 1      5.1      3.5      1.4
1321 ## 2      4.9      3.0      1.4
1322 ## 3      4.7      3.2      1.3

```

1323 `select()` erlaubt es, auch hier den `:-`Operator zu verwenden:

```
iris %>% select(Sepal.Length:Petal.Length) %>% head(3)
```

```

1324 ##      Sepal.Length Sepal.Width Petal.Length
1325 ## 1      5.1      3.5      1.4
1326 ## 2      4.9      3.0      1.4
1327 ## 3      4.7      3.2      1.3

```

1328 Es gibt auch einige spezielle Funktionen, um Spalten innerhalb eines `select()`-Aufrufs auszuführen:

- 1329 • `starts_with()`: Hier kann man ein Muster angeben, mit dem ein Text anfangen muss.
- 1330 • `ends_with()`: Diese Funktion ist analog zu `starts_with()`, jetzt wird aber am Ende des Spaltennamens nach dem Muster gesucht.
- 1331 • `contains()`: Hier kann ein Muster übergeben werden, das irgendwo im Spaltennamen sein muss.
- 1332 • `everything()`: Mit dieser Funktion werden alle Spalten ausgewählt.
- 1333 • `last_col()`: Mit dieser Funktion wird nur die letzte Spalte ausgewählt (dass ist die Spalte, die ganz rechts ist).

1336 Sämtliche Auswahlen können mit `-` umgekehrt werden.

```
iris %>% select(starts_with("Sepal")) %>% head(3)
```

```

1337 ##      Sepal.Length Sepal.Width
1338 ## 1      5.1      3.5
1339 ## 2      4.9      3.0
1340 ## 3      4.7      3.2

```

```
iris %>% select(-starts_with("Sepal")) %>% head(3)
```

```

1341 ##      Petal.Length Petal.Width Species
1342 ## 1      1.4      0.2  setosa
1343 ## 2      1.4      0.2  setosa
1344 ## 3      1.3      0.2  setosa

```

1345 `select()` bietet auch noch die Möglichkeit, Spalten namen zu ändern.

```
iris %>% select(sep_width = Sepal.Width) %>% head(3)
```

```
1346 ##    sep_width
1347 ## 1         3.5
1348 ## 2         3.0
1349 ## 3         3.2
```

```
1350
```

1351 *Aufgabe 29: Auswählen von Spalten*

```
1352
```

1353 In der Datei `messungen_1.csv` sind Messungen von zwei Sensoren enthalten für die ersten vier Monate eines
1354 Jahres. Führen Sie folgende Abfragen durch:

- 1355 1. Wählen Sie alle Messungen für Januar aus.
- 1356 2. Wählen Sie alle Messungen für Januar und März aus.

1357 9.7 Einzelne Beobachtungen abfragen (`slice()`)

1358 Mit dem Befehl `slice()` kann man einzelne Beobachtungen (= Zeilen) aus einem `data.frame` abfragen.

```
slice(iris, c(1, 9, 18))
```

```
1359 ##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1360 ## 1         5.1         3.5         1.4         0.2  setosa
1361 ## 2         4.4         2.9         1.4         0.2  setosa
1362 ## 3         5.1         3.5         1.4         0.3  setosa
```

1363 Davon gibt es drei nützliche Varianten: 1) `slice_head()` und `slice_tail()`; 2) `slice_max()` und
1364 `slice_min()`; 3) `slice_random()`.

1365 `slice_head()` und `slice_tail()` sind analog zu `head()` und `tail()`, aber mit dem entscheidenden Unter-
1366 schied, dass Gruppierungen berücksichtigt werden. Wenn keine Gruppierung in den Daten vorhanden ist, gibt
1367 es keinen Unterschied.

```
iris %>% head(n = 2)
```

```
1368 ##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1369 ## 1         5.1         3.5         1.4         0.2  setosa
1370 ## 2         4.9         3.0         1.4         0.2  setosa
```

```
iris %>% slice_head(n = 2)
```

```
1371 ##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1372 ## 1         5.1         3.5         1.4         0.2  setosa
1373 ## 2         4.9         3.0         1.4         0.2  setosa
```

1374 Sobald jedoch eine gruppierende Variable eingeführt wird, gibt `slice_head()` die ersten `n` Beobachtungen
1375 für jede Gruppe zurück und `head()` für den gesamten Datensatz.

```
# base head
iris %>% group_by(Species) %>%
  head(n = 2)
```

```
1376 ## # A tibble: 2 x 5
1377 ## # Groups:   Species [1]
1378 ##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1379 ##         <dbl>         <dbl>         <dbl>         <dbl> <fct>
1380 ## 1         5.1         3.5         1.4         0.2 setosa
1381 ## 2         4.9         3         1.4         0.2 setosa
```

```
# dplyr slice_head
iris %>% group_by(Species) %>%
  slice_head(n = 2)
```

```
1382 ## # A tibble: 6 x 5
1383 ## # Groups:   Species [3]
1384 ##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1385 ##         <dbl>         <dbl>         <dbl>         <dbl> <fct>
1386 ## 1         5.1         3.5         1.4         0.2 setosa
1387 ## 2         4.9         3         1.4         0.2 setosa
1388 ## 3         7         3.2         4.7         1.4 versicolor
1389 ## 4         6.4         3.2         4.5         1.5 versicolor
1390 ## 5         6.3         3.3         6         2.5 virginica
1391 ## 6         5.8         2.7         5.1         1.9 virginica
```

1392 `slice_tail()` funktioniert analog zu `slice_head()` mit dem einzigen Unterschied, dass nicht die ersten `n`
 1393 Zeilen zurück gegeben werden sondern die letzten `n` Zeilen.

1394 `slice_max()` und `slice_min()` geben die Beobachtung mit dem maximalen bzw. minimalen Wert einer
 1395 Variable zurück. Auch hier werden Gruppen berücksichtigt.

```
iris %>% slice_max(Sepal.Length)
```

```
1396 ##   Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
1397 ## 1         7.9         3.8         6.4         2 virginica
```

1398 Und mit Gruppen:

```
iris %>% group_by(Species) %>%
  slice_max(Sepal.Length)
```

```
1399 ## # A tibble: 3 x 5
1400 ## # Groups:   Species [3]
1401 ##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1402 ##         <dbl>         <dbl>         <dbl>         <dbl> <fct>
1403 ## 1         5.8         4         1.2         0.2 setosa
1404 ## 2         7         3.2         4.7         1.4 versicolor
```

```
1405 ## 3          7.9          3.8          6.4          2  virginica
```

1406 `slice_min()` funktioniert genau gleich, nur dass die Beobachtung (=Zeile) mit dem kleinsten Wert einer
1407 Variable zurück gegeben wird.

1408 Die Funktion `slice_sample()` erlaubt es zufällige Beobachtungen zu ziehen. Dabei kann über das Argument `n`
1409 die Anzahl an Beobachtungen angegeben werden oder über das Argument `prop` der Anteil an Beobachtungen.

```
slice_sample(iris, n = 5)
```

```
1410 ##   Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
1411 ## 1          6.5          2.8          4.6          1.5 versicolor
1412 ## 2          6.3          3.3          4.7          1.6 versicolor
1413 ## 3          7.2          3.2          6.0          1.8  virginica
1414 ## 4          4.9          3.6          1.4          0.1    setosa
1415 ## 5          6.0          2.7          5.1          1.6 versicolor
```

1416 Das Ergebnis ist bei jedem von Ihnen anders, da es sich um eine zufällige Ziehung handelt. Wenn Sie diese
1417 Ergebnisse wiederholen möchte, können Sie über `set.seed()` die zufällige Ziehung reproduzierbar machen.

```
set.seed(123)
slice_sample(iris, n = 5)
```

```
1418 ##   Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
1419 ## 1          4.3          3.0          1.1          0.1    setosa
1420 ## 2          5.0          3.3          1.4          0.2    setosa
1421 ## 3          7.7          3.8          6.7          2.2 virginica
1422 ## 4          4.4          3.2          1.3          0.2    setosa
1423 ## 5          5.9          3.0          5.1          1.8 virginica
```

1424 Wenn beispielsweise 5% der Beobachtungen gezogen werden sollen, kann dies so gemacht werden:

```
slice_sample(iris, prop = 0.05)
```

```
1425 ##   Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
1426 ## 1          7.7          3.8          6.7          2.2  virginica
1427 ## 2          5.5          2.5          4.0          1.3 versicolor
1428 ## 3          5.5          2.6          4.4          1.2 versicolor
1429 ## 4          6.5          3.0          5.2          2.0  virginica
1430 ## 5          6.1          3.0          4.6          1.4 versicolor
1431 ## 6          6.3          3.4          5.6          2.4  virginica
1432 ## 7          5.1          2.5          3.0          1.1 versicolor
```

1433 `slice_sample()` berücksichtigt ebenfalls Gruppen. Mit den Argumenten `replace` und `weight_by` dann die
1434 Zufallsziehung genauer spezifiziert werden. `replace` sagt, ob eine gezogenen Beobachtung wieder zurück gelegt
1435 wird oder nicht. Mit dem Argument `weight_by` können optional gewichte für jede Beobachtung vergeben
1436 werden.

Aufgabe 30: Daten beschreiben

Verwenden Sie den Datensatz `bhd_1.txt` und finden Sie für jedes Gebiet und Art die Beobachtung mit kleinsten BHD.

9.8 Spalten trennen

Ein gut geplanter Datensatz besteht aus Beobachtungen (Zeilen), Variablen (Spalten) und in jeder Zelle ist immer ein *genau* ein Wert gespeichert. Leider gibt es oft Datensätze, bei denen mehr als ein Wert pro Zelle gespeichert wurde. Die Funktion `seperate()` kann helfen solche Daten zu trennen.

Wir verwenden einen erfunden Datensatz zu Beobachtungen von Tieren und einer geschätzten Distanz zu diesen Tieren.

```
dat <- tibble(
  id = 1:4,
  beobachtung = c("10m, Reh", "100m, Reh", "20m, Fuchs", "40,Reh"),
)
```

In der Spalte `beobachtung` sind zwei Informationen gespeichert: Die Distanz zur Beobachtung und die Art. Das ist ungünstig, weil wir so weder nach Tierart noch nach distanz filtern können. Mit der Funktion `seperate()`, können wir Beobachtungen einer Spalte in mehrere Spalten trennen. Dafür muss der Spaltennamen (Argument `col`), die neuen Spaltennamen (Argument `into`) und das Trennzeichen (Argument `sep`) angegeben werden.

```
seperate(dat, col = beobachtung, into = c("Distanz", "Art"), sep = ",")
```

```
## # A tibble: 4 x 3
##       id Distanz Art
##   <int> <chr>   <chr>
## 1     1    10m    " Reh"
## 2     2   100m    " Reh"
## 3     3    20m    " Fuchs"
## 4     4     40    "Reh"
```

Nach dem Aufruf von `seperate()` gibt es zwei neue Spalten (`Distanz` und `Art`), die die alte Spalte `beobachtung` ersetzen.

Aufgabe 31: Aufräumen

Verwenden Sie den folgenden Datensatz und bringen Sie ihn in eine Form, die sicherstellt dass

- jede Zelle genau einen Wert enthält.
- jede Zeile eine Beobachtung ist.
- die Spaltennamen aus einem ausschlagkräftigen Wort bestehen.


```
dat <- data.frame(  
  standort = c("a1", "a2", "b1", "b2"),  
  j2019 = c("3 x Fuchs", "4 x Reh", "1 x Fuchs", "2 x Reh"),  
  j2020 = c("2 x Fuchs", "1 x Reh", "", "2 x Fuchs")  
)
```

10 Arbeiten mit Text

Bis jetzt haben wir fast ausschließlich mit Zahlen oder Abbildungen gearbeitet. R bietet aber auch viele Werkzeuge, um mit Text zu arbeiten. Wir wollen hier ein paar Funktionen dafür vorstellen. Als erstes sollte nochmals klargestellt werden, was eigentlich ein Text ist. In R ist alles, das innerhalb von doppelten (") oder einfachen (') Anführungszeichen geschrieben ist, Text.

Anbei einige Beispiele:

```
a <- "Das ist ein kurzer Satz."
b <- "Auch das ist 'moeglich'."
z <- "30"
```

Wichtig ist hier zu sehen, dass `z` nicht als Zahl sondern, als Text interpretiert wird.

```
z + 1
```

```
## Error in z + 1: nicht-numerisches Argument für binären Operator
```

Wenn man sicher ist, dass es sich bei einem Textobjekt um eine Zahl handelt, kann man dies mit der Funktion `as.numeric()` in eine Zahl umwandeln.

```
as.numeric(z) + 1
```

```
## [1] 31
```

Aber mit `a` führt dies wieder zu einem NA-Wert, da `a` nicht in eine Zahl umgewandelt werden kann.

```
as.numeric(a) + 1
```

```
## Warning: NAs durch Umwandlung erzeugt
```

```
## [1] NA
```

10.1 Arbeiten mit Text

Wir wollen erst einmal drei Funktionen besprechen, die es erlauben mit Text zu arbeiten. Die Funktion `nchar()`¹¹ gibt an wie viele Zeichen ein Text hat. Also z.B.

```
nchar("Hallo")
```

```
## [1] 5
```

```
nchar("30")
```

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

```
nchar("Hallo und Guten Tag!")
```

```
## [1] 20
```

Die Funktion `paste()` erlaubt es verschiedene Variablen mit Text zu verbinden. Wenn wir z. B. die Variablen `vorname <- "Eva"` und `name <- "Meier"` haben und wir wollen eine neue Variable `full_name <- "Eva"`

¹¹`char` ist kurz für *character*.

Meier" erzeugen, dann kann das mit der Funktion `paste()` gemacht werden.

```
vorname <- "Eva"
name <- "Meier"
full_name <- paste(vorname, name)
full_name
```

```
## [1] "Eva Meier"
```

Die Funktion `paste()` hat das Argument `sep`, das auf ein Leerzeichen () gesetzt ist, aber auch anders sein kann und das Trennzeichen definiert.

```
full_name <- paste(vorname, name, sep = ", ")
full_name
```

```
## [1] "Eva, Meier"
```

Die Funktion `substr()` erlaubt es am Anfang oder Ende eines Wortes etwas abzuschneiden. Dabei muss immer die Anfangs- (`start`) und Endposition (`stop`) angegeben werden.

```
substr("Hallo", start = 1, stop = 3)
```

```
## [1] "Hal"
```

```
substr("Hallo", start = 2, stop = 5)
```

```
## [1] "allo"
```

Aufgabe 32: Arbeiten mit Text 1

Verwenden Sie den folgenden Vektor:

```
txt <- c("Versicherung", "Vogel", "Station", "Gutschein", "Statistik", "Stift",
        "Methoden", "Fluss", "Rudel", "Baum", "Haus", "Wahrscheinlich", "Foto",
        "Seife", "Garten", "Auto", "Handy", "Teller", "Kalender")
```

1. Aus wie vielen Buchstaben besteht jedes Wort?
2. Finden Sie das längste Wort.
3. Wie viel Prozent der Wörter fangen mit einem S an?
4. Fügen Sie jedem Wort seine Position im Vektor hinzu. Beispielsweise soll aus `Vogel` "2. Vogel" werden usw.

10.2 Finden von Textmustern

Mit der Funktion `grep()` können Muster in einem Text gefunden werden. Wenn wir beispielsweise folgenden Vektor mit Textelementen haben.

```
txt <- c("Buesgenweg", "Friedlaender Weg", "Berliner Strasse")
```

Und wir wollen alle Straßennamen die ein `weg` haben abfragen, dann können wir folgenden Befehl ausführen:

```
grep("Weg", txt)
```

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

Im zweiten Element von `txt` kommen die Zeichen `Weg` vor. Beachte, in der Standardeinstellung wird zwischen Groß- und Kleinschreibung unterschieden. Dies kann mit dem Argument `ignore.case = TRUE` angepasst werden.

```
grep("Weg", txt, ignore.case = TRUE)
```

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

Mit der Funktion `sub` können Zeichen innerhalb einer Zeichenkette ersetzt werden.

So ersetzt der folgende Ausdruck `ae` mit `ä`.

```
sub("ae", "ä", "Friedlaender Weg")
```

```
## [1] "Friedländer Weg"
```

Wenn allerdings das zu ersetzende Zeichen mehr als einmal vorkommt und beide Instanzen ersetzt werden sollen braucht man die Funktion `gsub`.

```
txt <- "Friedlaender Weg und Reinhaeuser Landstrasse sind zwei Strassen in Goettingen."
sub("ae", "ä", txt)
```

```
## [1] "Friedländer Weg und Reinhaeuser Landstrasse sind zwei Strassen in Goettingen."
```

Mit `sub()` wird nur das erste `ae` ersetzt, während `gsub()` *alle* `ae` mit einem `ä` ersetzt.

```
gsub("ae", "ä", txt)
```

```
## [1] "Friedländer Weg und Reinhäuser Landstrasse sind zwei Strassen in Goettingen."
```

Oft ist der genaue Ausdruck den man finden möchte jedoch Variable. Beispielsweise möchte man alle Wörter mit einem Umlaut oder Zahlen finden möchte, kann man das oft abkürzen. Dafür gibt es reguläre Ausdrücke. Wir werden hier nur ein paar beispielhafte Anwendungen besprechen.

Sowohl in den Funktionen `grep()` als auch `(g)sub()` kann mit anstatt dem Muster (immer das erste Argument) aus ein regulärer Ausdruck angegeben werden. Mit `[1-9]` sind alle Zahlen von 1 bis 9 gemeint.

Das Ziel ist es jetzt alle Straßen zu finden, die auch einen Straßenummer haben:

```
txt <- c("Büsgenweg 1", "Berliner Strasse", "Kurze Strasse 13")
```

```
grep("[0-9]", txt)
```

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

Damit lässt sich auch das Problem mit groß- und kleingeschriebenen Wörtern lösen.

```
txt <- c("Buesgenweg", "Friedlaender Weg", "Berliner Strasse")
grep("[wW]eg", txt)
```

```
1533 ## [1] 1 2
```

```
1534
```

1535 *Aufgabe 33: Arbeiten mit Text 2*

```
1536
```

1537 Verwenden Sie den folgenden Vektor:

```
txt <- c("Versicherung", "Methoden", "Fluss", "Rudel",
        "Baum", "Haus", "Foto", "Auffahrt", "Auto", "Handy", "Teller",
        "Kalender", "Aufbau")
```

1538 1. In wie vielen Wörtern kommt der Doppellaut au vor?

1539 2. Ersetzen Sie in allen Wörtern alle au mit _ _.

```
grep("au", txt)
```

```
1540 ## [1] 5 6 13
```

```
gsub("au", "_ _", txt)
```

```
1541 ## [1] "Versicherung" "Methoden"      "Fluss"          "Rudel"          "B_ _m"
```

```
1542 ## [6] "H_ _s"         "Foto"          "Auffahrt"       "Auto"           "Handy"
```

```
1543 ## [11] "Teller"        "Kalender"      "Aufb_ _"
```

11 Arbeiten mit Zeit

Für den Computer bzw. R ist ein Datum erst einmal nichts anderes als ein Text. Für uns ist es sofort klar, dass der "13.2.2021" der 13. Februar 2021 ist, für den Computer nicht. Wir müssen R also irgendwie sagen, dass die 13 der Tag ist, die 2 der Monat und 2021 das Jahr. Dass der Computer die einzelnen Komponenten erkennt, nennt man *parsen*¹². Das Arbeiten mit Datum und Zeit kann anfangs sehr mühsam sein, aber sobald man einige Grundfertigkeiten erworben hat, kann man viele Aufgaben deutlich schneller und effizienter erledigen. Der erste Schritt ist immer ein Datum zu *parsen*. Wir verwenden dafür Funktionen aus dem Paket **lubridate**. Als erstes müssen wir wieder Paket **lubridate** laden mit:

```
library(lubridate)
```

lubridate bietet eine Vielzahl von Funktionen zum *parsen* von Datum und Zeit, die sich aus:

- **y** für Jahr,
- **m** für Monat,
- **d** für Tag,
- **h** für Stunde,
- **m** für Minute und
- **s** für Sekunde

zusammen setzten. Alle Funktionen nehmen als erstes Argument ein Textstring. Wenn wir z.B. den String 2020-01-20 parsen wollen können wir das mit der Funktion **ymd** machen.

```
ymd("2020-01-20")
```

```
## [1] "2020-01-20"
```

Dabei erkennt **lubridate** in der Regel die Trennzeichen:

```
ymd("2020.01.20")
```

```
## [1] "2020-01-20"
```

```
ymd("2020/01/20")
```

```
## [1] "2020-01-20"
```

```
ymd("2020 01 20")
```

```
## [1] "2020-01-20"
```

Wenn die die Anordnung der einzelnen Komponenten anders ist, gibt es einfach eine andere Funktion.

```
dmy("20.1.2020")
```

```
## [1] "2020-01-20"
```

Jetzt stellt sich die Frage, was der Vorteil ist, wenn R ein Datum parst.

¹²to *parse* heißt zergliedern bzw. grammatikalisch bestimmen.

```
d <- dmy("20.1.2020")
```

Wir können jetzt mit `d` arbeiten und einzelne Komponenten extrahieren.

```
day(d)
```

```
## [1] 20
```

```
month(d)
```

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

```
year(d)
```

```
## [1] 2020
```

Oder auch Zeiteinheiten hinzufügen oder abziehen.

```
d + days(10)
```

```
## [1] "2020-01-30"
```

```
d - years(20)
```

```
## [1] "2000-01-20"
```

```
d + hours(25)
```

```
## [1] "2020-01-21 01:00:00 UTC"
```

Aufgabe 34: Arbeiten mit Datum und Zeit

- Parsen Sie folgende Zeitangaben 23.1.2020, 13.2.1992 20:55:23, Mar/3/97 und 10.7.2020 19:15 und speichern Sie diese in einen Vektor `d`.
- Extrahieren Sie nun aus jedem Element aus `d` das Jahr und die Stunde.
- Fügen zu jedem Element in `d` 10 Tage hinzu.

11.1 Arbeiten mit Zeitintervallen

Mit zwei Zeitpunkten lassen sich Zeitintervalle (**Periods**) erstellen, dafür können wir die Funktion `interval()` aus dem Paket `lubridate` verwenden¹³.

```
anfang <- ymd("2020-03-18")
ende <- anfang + years(1)

int <- interval(anfang, ende)
```

Wir können jetzt mit `int` arbeiten und beispielsweise das Intervall verschieben,

¹³Alternativ zur Funktion `interval()` kann auch der `%--%`-Operator verwendet werden. Man könnte `int` auch so erstellen `int <- anfang %--% ende`.

```
int_shift(int, years(3))
```

```
## [1] 2023-03-18 UTC--2024-03-18 UTC
```

die Länge des Intervalls berechnen

```
int_length(int) # in Sekunden
```

```
## [1] 31536000
```

oder testen ob ein Datum innerhalb des Intervalls liegt.

```
ymd("2020-07-1") %within% int
```

```
## [1] TRUE
```

```
ymd("2021-07-1") %within% int
```

```
## [1] FALSE
```

%within% funktioniert genauso mit Vektoren oder mit mehreren Intervallen. Wir könnten also zwei Intervalle definieren (z.B. Ostern und Pfingsten).

```
ostern <- ymd("2021-04-02") %--% ymd("2021-04-05")
pfingsten <- ymd("2021-05-22") %--% ymd("2021-05-24")
```

Und Überprüfen welche Termine in eines der zwei Intervalle fallen.

```
termine <- ymd("2021-03-29") + weeks(0:10)
```

```
# Ostern
```

```
termine %within% ostern
```

```
## [1] FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
# Pfingsten
```

```
termine %within% pfingsten
```

```
## [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE
```

11.2 Formatieren von Zeit

Für die Ausgabe in Berichten oder Grafiken soll das Datum oft in einer speziellen Form dargestellt werden. Die Funktion `format()` bietet Möglichkeiten ein Datumsobjekt zurück in Text umzuwandeln.

Ein Beispiel wäre `ymd("2021-2-10")` als 10.2.21 auszugeben.

```
d <- ymd("2021-2-21")
format(d, "%d.%m.%y")
```

```
## [1] "21.02.21"
```

Dabei handelt sich bei `%d.%m.%y` um Abkürzungen für die unterschiedlichen Komponenten eines Datumobjekts. Siehe dazu die Hilfeseite von `strptime` (`help(strptime)`).

1606

Aufgabe 35: Arbeiten mit Intervallen

1607
1608

1609 Wie viele Einträge aus dem Vektor `v1` befinden sich in einem Intervall, das zwischen dem 1.3.2021 und dem
1610 5.3.2021 definiert ist.

```
v1 <- c(  
  "2021-03-05", "2021-03-03", "2021-03-09", "2021-03-09", "2021-03-09",  
  "2021-03-03", "2021-03-08", "2021-03-10", "2021-03-07", "2021-03-10"  
)
```

12 Aufgaben Wiederholen (for-Schleifen)

Um einfache Programme zu schreiben, müssen Sie den Ablauf der Programmcodes kontrollieren können. Kontrollieren bedeutet in diesem Zusammenhang, dass Codeabschnitte nur unter definierten Bedingungen ablaufen. Sie programmieren also zwei Sachen. 1) den Code selbst und 2) die Bedingungen die erfüllt sein müssen, damit der Code ausgeführt wird. Diese Kontrollbedingungen ermöglichen es Ihnen generisch zu programmieren. Sie schreiben Ihren Code also nicht speziell maßgeschneidert für ein Problem, sondern so generell, dass er für mehrere Auswertungen funktioniert. Um dies zu gewährleisten müssen Sie bestimmte Situationen vorhersehen und abfangen. Hierbei helfen Ihnen Kontrollstrukturen (**Control Flow**). Grundsätzlich gibt es **Control Flow** Funktionen zur Wiederholung von Codeblöcken (Schleifen) und logische Bedingungen (bedingte Anweisung).

12.1 Schleifen

Bis jetzt wurden alle Skripte einfach der Reihe nach abgearbeitet und zwischendurch bestimmte Programmteile, je nach Situation, selbstständig ausgeführt oder übersprungen. Mit einer Schleife kann man erreichen, dass eine Gruppe von Befehlen (der sog. Schleifenrumpf) mehrfach abgearbeitet wird, zum Beispiel wenn bestimmte Auswertungsschritte auf mehrere Datensätze oder Variablen angewendet oder Funktionen mit unterschiedlichen Parametern oder Startwerten aufgerufen werden sollen. Weitere Anwendungsmöglichkeiten sind iterative Algorithmen, in denen die Eingabewerte des aktuellen Iterationsschrittes von einem vorherigen abhängig sind. Besonders in Simulationen kommen Schleifen häufig zum Einsatz, da große Anzahlen von Wiederholungen benötigt werden.

Man unterscheidet zwischen zwei Arten von Schleifen: Bei den **for()**-Schleifen steht die Anzahl der Wiederholungen schon beim Eintritt in die Schleife fest, während die **while()**-Schleifen so lange ausgeführt werden, bis eine Bedingung nicht mehr wahr ist. Mit der Funktion **break** wird eine Schleife abgebrochen und die Programmausführung wird nach der Schleife fortgesetzt.

Die wesentlichen Befehle sind

- **for (i in X) {Code}**

Wiederhole den Code im “Schleifenrumpf” für jedes Element aus X.

- **while(Bedingung) {Code}**

Wiederhole den Code im “Schleifenrumpf” so lange die logische Bedingung erfüllt ist.

- **break()**

Brich die Schleife ab.

12.1.1 Wiederholen von Befehlen mit for().

Steht vor Beginn der Schleife fest wie viele Schleifendurchgänge benötigt werden, wenn zum Beispiel in einer Simulation 99 Realisierungen erzeugt oder alle Elemente eines Vektors verarbeitet werden sollen, verwendet man eine **for**-Schleife. Die allgemeine Form der **for**-Schleife ist:

```
X <- c(1 : 3) # Einträge die im Schleifenrumpf abgearbeitet werden.
for(i in X){
  # Schleifenrumpf
  print(i)
}
```

```
1645 ## [1] 1
1646 ## [1] 2
1647 ## [1] 3
```

1648 Das `i` steht in diesem Beispiel für die Schleifen-Variable. Sie muss nicht `i` heißen, sondern kann jeden
 1649 zulässigen Namen annehmen. Das `X` steht für einen existierenden Vektor oder eine existierende Liste bzw.
 1650 einen Ausdruck, der ein solches Objekt liefert (der Objektname ist ebenfalls frei wählbar). `for` und `in` sind
 1651 Schlüsselworte, sie müssen, ebenso wie die runden Klammern, vorhanden sein.

1652 Im ersten Durchgang erhält die Schleifen-Variable `i` den ersten Wert von `X` und der Schleifenrumpf wird
 1653 mit diesem Wert ausgeführt. Die Variable `i` nimmt nacheinander so lange die Werte von `X` an, bis ihr alle
 1654 Elemente zugewiesen wurden.

1655 Das folgende Beispiel wird zwar besser durch die entsprechende Vektoroperation gelöst, zeigt aber sehr
 1656 deutlich die Arbeitsweise der `for`-Schleife.

```
zahlen <- c(2, 3, 5)

for(element in zahlen){
  print(element^2)
}
```

```
1657 ## [1] 4
1658 ## [1] 9
1659 ## [1] 25
```

1660

1661 Aufgabe 36: Schleifen 1

1662

1663 Verwenden Sie den Vektor `k <- c(1, 3, 9, 12, 15)` und schreiben Sie folgende `for`-Schleifen:

- 1664 1. Eine Schleife, die jedes Element aus `k` ausgibt.
- 1665 2. Eine Schleife, die zu jedem Element aus `k` 10 addiert und den neuen Wert ausgibt.
- 1666 3. Eine Schleife wie in 2), aber der neue Wert ($k + 10$) soll jetzt nicht mehr ausgegeben werden, sondern
 1667 in `k10` gespeichert werden. Stellen Sie sicher, dass `k10` wieder von der Länge 5 ist.

1668

1669 Die Funktion `for()` ermöglicht es, einen Befehl beliebig oft zu wiederholen. Z.B. der folgende Ausdruck zieht
 1670 10-Mal eine Stichprobe der Größe 1 aus dem Vektor `v`. Beachten Sie, dass die Schleifen-Variable `i` selbst gar

1671 nicht im Schleifenrumpf vorkommt. Das Ziel dieser Schleife ist nicht die Elemente des Vektors abzuarbeiten,
 1672 sondern einfach nur den Ausdruck im Schleifenrumpf 10-Mal zu wiederholen.

```
v <- c(1, 4, 2, 3)
for (i in c(1 : 10)) {
  print(sample(v, 1))
}
```

```
1673 ## [1] 3
1674 ## [1] 4
1675 ## [1] 2
1676 ## [1] 4
1677 ## [1] 1
1678 ## [1] 4
1679 ## [1] 2
1680 ## [1] 3
1681 ## [1] 4
1682 ## [1] 1
```

1683 Auf gleiche Weise kann man auch über die Variablen eines Dataframes iterieren¹⁴. Das folgende Beispiel hat
 1684 zum Ziel die Funktionsweise von Schleifen zu verdeutlichen. Schleifen haben in R jedoch den Nachteil, dass sie
 1685 sehr langsam operieren. Wenn es geht, sollte man Alternativen verwenden. Die Funktionsweise wiederholender
 1686 Auswertungen wird jedoch mit `for`-Schleifen deutlicher. Aus diesem Grunde werden wir uns in diesem Kurs
 1687 auf Schleifen beschränken.

```
myLoopDf <- data.frame(a = c(2, 4, 7, 5),
                       b = c("Buche", "Eiche", "Eiche", "Buche"),
                       d = c(50, 60, 55, 80))

for (i in c(1 : 4)) {
  summeAd <- myLoopDf[i, "a"] + myLoopDf[i, "d"]
  print(myLoopDf$b[i])
  print(summeAd)
}
```

```
1688 ## [1] "Buche"
1689 ## [1] 52
1690 ## [1] "Eiche"
1691 ## [1] 64
1692 ## [1] "Eiche"
1693 ## [1] 62
1694 ## [1] "Buche"
1695 ## [1] 85
```

¹⁴Zur Info: Dieses Beispiel lässt sich schneller mittels der vektorwertigen Operation `apply()` lösen.

Aufgabe 37: for-Schleife

Lesen Sie den Datensatz `bhd_1.txt` ein und verwenden Sie eine `for`-Schleife.

- Ziehen Sie 500-Mal je 35 Beobachtungen für den BHD.
- Berechnen Sie jeweils den Mittelwert aus den 35 Werten.
- Speichern Sie die 500 Mittelwerte in einem neuen Vektor `mittel`.
- Wie ist die Verteilung dieser 500 Mittelwerte zu interpretieren?

12.1.2 Wiederholen von Befehlen mit `while()`

Die `while`-Schleifen finden Anwendung, wenn die Anzahl der zu durchlaufenden Wiederholungen vorher nicht bekannt ist, wie zum Beispiel bei Iterationsverfahren, die bis zu einer gewissen Genauigkeit durchlaufen werden sollen. Die `while`-Schleife besteht in R aus dem Schlüsselwort `while()` und einer Bedingung in runden Klammern.

```
while (Bedingung) {  
  # Schleifenrumpf  
}
```

Sie ist in der praktischen Programmierung nicht so relevant wie die `for`-Schleife. Sie sei deshalb hier nur kurz erwähnt. Die Abbruchbedingung wird jedes Mal geprüft bevor der Schleifenrumpf durchlaufen wird. Die Bedingung wird ausgewertet und wenn diese `TRUE` ist, wird der Schleifenrumpf ausgeführt und danach erneut die Bedingung überprüft. Ist die Bedingung nicht erfüllt, wird der Schleifenrumpf nicht durchgeführt und die Schleife beendet. Ist die Bedingung bereits vor Eintritt in die Schleife nicht erfüllt, wird die Schleife gar nicht erst durchlaufen.

Da `while`-Schleifen also so lange ausgeführt werden, bis die Bedingung nicht mehr erfüllt ist, kann eine Endlosschleife entstehen. Dies kann passieren, wenn man nicht sauber programmiert hat. Wenn innerhalb der Schleife nicht dafür gesorgt wird, dass die Bedingung irgendwann nicht mehr erfüllt wird, so läuft die Schleife immer weiter. Steckt R in einer Schleife fest und reagiert nicht mehr, kann der Befehl unter Linux mit `Strg`+`C` und unter Windows mit `Esc` abgebrochen werden. Alternativ können Sie auf das rote STOP Symbol über der Konsole klicken.

12.2 Bedingte Ausführung von Codeblöcken

Innerhalb eines Skripts ist es mitunter notwendig je nach aktueller Situation unterschiedlich fortzufahren. Die Situation wird mit einem logischen Ausdruck, einer sogenannten Bedingung, geprüft. Je nachdem, ob die Bedingung wahr (`TRUE`) oder falsch (`FALSE`) ist, werden unterschiedliche Programmteile ausgeführt, der jeweils andere Teil bleibt unberücksichtigt. Danach wird in jedem Fall die Programmausführung, mit den auf die bedingte Anweisungen folgenden Anweisung, fortgesetzt. In R kann dies mit dem `if-else`-Konstrukt realisiert werden, welches aus den Schlüsselwörtern `if()` und `else` sowie der in runde Klammern gefassten Bedingung besteht.

```
if(Bedingung){  
  # Anweisungen für Bedingung == TRUE  
} else{  
  # Anweisungen für Bedingung == FALSE  
}
```

1729 Im folgenden Beispiel sollen die bisher abstrakt beschriebenen Vorgänge praktische Anwendung finden. In
1730 dem Beispiel wird zunächst durch zufälliges Ziehen einer Zahl aus dem Bereich eins bis sechs ein Würfelwurf
1731 simuliert. Anschließend wird der Würfelwurf mit einem if-Ausdruck kommentiert. Ist die Bedingung, es wurde
1732 eine Sechs gewürfelt, erfüllt, wird der Code innerhalb der geschweiften Klammern ausgeführt, ansonsten wird
1733 der Klammerinhalt ignoriert.

```
# Würfelwurf simulieren  
x <- sample(1:6, 1)  
# if-Konstrukt zur passgenauen Gratulation  
if (x == 6) {  
  print("Glückwunsch, eine Sechs!")  
}
```

1734 In den meisten Fällen ist es für R irrelevant, ob sich zwischen den verschiedenen Elementen Leerzeichen oder
1735 Zeilenumbrüche befinden. Bei dem else-Ausdruck dagegen wird ein Fehler erzeugt, wenn vor dem **else** nicht
1736 die geschweifte Endklammer der vorausgehenden if- Bedingung steht.

```
# Würfelwurf simulieren  
x <- sample(1:6, 1)  
# if-else-Konstrukt: Gratulation oder Ermutigung  
if(x == 6){  
  print("Glückwunsch, eine Sechs!")  
} else{  
  print("Beim nächsten Wurf klappt's bestimmt.")  
}
```

1737 ## [1] "Beim nächsten Wurf klappt's bestimmt."

1738

1739 Aufgabe 38: Bedingte Programmierung

1740

- 1741 • Wenn keine 6 gewürfelt wurde, lassen Sie zusätzlich ausgeben welche Zahl gewürfelt wurde.
- 1742 • Wiederholen Sie den Würfelwurf 10 Mal.

13 (R)markdown

13.1 Markdown Grundlagen

Die Idee von Markdown ist, einfach Text strukturieren zu können und das ganze ohne umfangreiche Programme zu erstellen. Im nächsten Abschnitt sehen wir dann, wie man Markdown und R-Code verbinden kann. Hier soll es jedoch erst einmal darum gehen, die einfachsten Bausteine von Markdown vorzustellen.

Am Anfang jedes Dokuments kommt eine Präambel. Diese fängt mit `---` an und hört auch wieder mit `---` auf. Innerhalb der Präambel können dann Metainformationen über das Dokument festgelegt werden, dies beinhaltet im einfachsten Fall: Titel, Autor und Datum. Das würde dann so aussehen:

```
---
title: "Ein Titel"
author: "Der, der es geschrieben hat"
date: "März 2021"
---
```

Danach folgt strukturierter und formatierter Text. Verschiedene Hierarchieebenen von Überschriften können mit der Anzahl an `#` festgelegt werden. So ist eine Überschrift erster Ordnung `# Kapitel` eine Überschrift zweiter Ordnung `## Unterkapitel` usw.

Listen können erstellt werden, wenn man am Anfang jeder Zeile ein `-` oder `1.` schreibt.

```
- Erster Eintrag
- Zweiter Eintrag
- Dritter Eintrag
```

wird zu

- Erster Eintrag
- Zweiter Eintrag
- Dritter Eintrag

Eine zentrale Idee von Markdown ist es Text einfach zu formatieren. Werden eine oder mehrere Wörter mit zwei Sternchen (`**`) eingefasst wird dieser Text **fett** dargestellt. Also aus `**wichtig**` wird **wichtig**. Das gleiche funktioniert auch mit *kursiven* Text, jedoch muss man hier noch ein Sternchen verwenden, also aus `*kursiv*` wird *kursiv*. Soll ein text fett und kursiv sein, kann man drei Sternchen verwenden. Aus `***sehr wichtig***` wird dann ***sehr wichtig***.

Weitere Elemente wie Links oder Abbildungen können einfach eingebunden werden. Links werden mit `[Link text](url)` in den Text integriert. Beispielsweise ist eine gute Idee bei [stackoverflow](https://stackoverflow.com) bei Problemen nach einer Lösung zu suchen. Dieser Link wurde mit `[stackoverflow](www.stackoverflow.com)` erstellt.

Für Abbildungen gibt es einen ganz ähnlichen Syntax. Mit `![Das R Logo](abb/r_logo.png)` wird die Abbildung `r_logo.png` eingebunden mit der Beschriftung: "Das R Logo".



Abbildung 9: Das R Logo

Aufgabe 39: Arbeiten mit markdown

Verwenden Sie das folgende Markdownndokument:

```
---
title: "Dokument"
author: "Ihr Name"
date: "März 2021"
---
```

Einleitung

Methoden

1. Kopieren Sie die Vorlage in ein Dokument, das `test.md` heißt.
2. Fügen Sie zwei Überschriften zweiter und dritter Ordnung hinzu.
3. Fügen Sie einen *kursiven* Text hinzu.
4. Fügen Sie einen Link zu einer Website hinzu.
5. Kompilieren Sie die Datei, indem Sie in Rstudio auf **Preview** drücken (Abbildung 10).

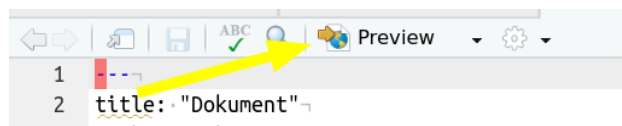


Abbildung 10: Kompilieren einer md-Datei.

13.2 R und Markdown

Markdown macht es bereits einfach Textdokumente und Dokumentationen zu verfassen, aber die wirkliche Stärke liegt in der Möglichkeit R und Markdown zu kombinieren. Man spricht dann von Rmarkdown. Ein weiteres Strukturelement, das wir noch nicht kennen gelernt haben, sind Code-Blöcke.

```
```
a <- 1:10
```



```
1801 a[1]
1802 ```
```

1803 erzeugt

```
1804 a <- 1:10
1805 a[1]
```

1806 Momentan wird noch kein Code ausgeführt, sondern lediglich der Code, als Code dargestellt. Rmarkdown  
1807 bietet nun die Möglichkeit Code beim *kompilieren*<sup>15</sup> auszuführen. Dafür müssen wir nur einen Code-Block als  
1808 R-Code-Block kennzeichnen.

```
1809 ```{R}
1810 a <- 1:10
1811 a[1]
1812 ```
```

1813 erzeugt

```
a <- 1:10
a[1]
```

```
1814 ## [1] 1
```

1815 Beachte, die Variable **a** wird beim kompilieren erzeugt und steht dann R zur Verfügung. R-Code-Blöcke  
1816 werden auch als Code Chunks bezeichnet. Diese Chunks können sehr genau angesprochen und angepasst  
1817 werden. Einige wichtige Argumente sind:

- 1818 • **echo**: Gibt an, ob der Quelltext angezeigt werden soll oder nicht.
- 1819 • **result**: Gibt an, ob die Ergebnisse gezeigt werden sollen oder nicht.
- 1820 • **eval**: Diese Option gibt an ob der Chunk ausgeführt werden soll oder nicht.

1821

#### 1822 **Aufgabe 40: Arbeiten mit Rmarkdown**

1823

1824 Erstellen Sie eine neue Rmarkdown Datei mit dem Namen **test1.Rmd**. Erstellen Sie zwei Code-Chunks. Der  
1825 erste soll nicht angezeigt werden und darin werden die Daten geladen (**bhd\_1.txt**). Im zweiten Chunk plotten  
1826 Sie das Alter der Bäume gegen den BHD. Was passiert mit dem Plot, wenn Sie die Datei kompilieren (drücken  
1827 Sie dazu auf den **Knit**-Knopf; Abbildung 11).



Abbildung 11: Kompilieren einer Rmd-Datei.

<sup>15</sup>Unter kompilieren wird hier das Übersetzen eines Markdowndokuments in ein Ausgabeformat (z.B. pdf oder html) verstanden.

## 14 Räumliche Daten in R

### 14.1 Was sind räumliche Daten

Räumliche Daten sind Beobachtungen, wie wir sie schon oft gesehen haben, mit einem räumlichen Bezug. Der Unterschied zu nicht räumlichen Daten liegt darin, dass räumliche Daten eindeutig im *Raum* verortet werden können. Häufig werden sogenannte Geoinformationssysteme (GIS) zum Arbeiten mit räumlichen verwendet. R kann in vielerlei Hinsicht wie ein GIS eingesetzt werden und hier werden einige Grundfunktionalitäten dafür besprochen. Räumliche Daten werden in zwei unterschiedliche Datentypen unterteilt: Vektor- und Rasterdaten. Vektordaten modellieren einzelne Objekte (= *Features*). Rasterdaten modellieren eine Oberfläche.

Vektordaten bestehen aus zwei Komponenten: 1) einer Geometrie, die die Form und Lage der Daten definiert und 2) Attributen, den tatsächlichen Daten. Räumliche Daten werden oft als *Features* bezeichnet. Ein Feature ist die räumliche Einheit einer Beobachtung. Je nach Art der räumlichen Daten können Features entweder Punkte (z.B. ein Baum), Linien (z.B. eine Straße) oder Polygone (z.B. ein See) sein. Auch können mehrere Geometrien zu einem Feature zusammengefasst werden. Ein Beispiel wäre eine Beobachtung für ein Land, das aber aus mehreren Polygonen bestehen kann (z.B. Festland und Inseln). Features können dann weitere Attribute (= Attributdaten) haben, z.B. eine ID, Name oder was auch immer man gemessen hat.

Rasterdaten bestehen aus einer Oberfläche von gleichgroßen Kacheln (= *Pixel*), die ein Gebiet abdecken. Meist sind Pixel viereckig, aber das ist keine Voraussetzung. Dabei hat jedes Pixel einen Wert (das kann auch ein fehlender Wert sein). Typische Beispiele für Rasterdaten sind Landnutzung oder Seehöhen.

In R kann sowohl mit Vektor- als auch mit Rasterdaten gearbeitet werden. Für Vektordaten bietet sich das Paket **sf** an und für Rasterdaten das Paket **raster**.

### 14.2 Koordinatenbezugssystem

Eine der Herausforderungen für räumliche Daten ist die eindeutige Verortung im Raum. Dazu braucht man ein *Koordinatenbezugssystem* (*KBS*). Einem KBS liegt ein mathematisches Modell der Erde zugrunde. Die Details zu KBS werden schnell relativ kompliziert und wir beschränken uns hier lediglich darauf, wie KBS verwendet werden können. Dazu müssen wir zwei Fälle unterscheiden: 1) einem Datensatz ein KBS zuweisen und 2) Transformation des KBSs eines Datensatzes in ein anderes KBS. Die technischen Details werden in den folgend Abschnitten besprochen. Für beide Aufgaben müssen wir auf ein KBS verweisen können, ein einfacher Ansatz dafür sind die sogenannten *EPSG-Codes*<sup>16</sup>.

### 14.3 Vektordaten in R

Das Paket **sf** stellt Klassen zum Abbilden von Features zur Verfügung, die dann in einem **data.frame** als Liste gespeichert werden können. In der Regel erstellen wir Features nicht individuell, sondern lesen diese aus externen Datenquellen (z.B. ESRI Shapefile) ein. Zum besseren Verständnis, erstellen wir es einmal manuell.

Wir haben die Koordinaten für drei Städte (Göttingen, Hannover und Berlin) als geografische Koordinaten vorliegen (EPSG = 4326).

<sup>16</sup> *EPSG* steht für European Petrol Survey Group

```
library(sf)
goe <- st_point(x = c(9.9158, 51.5413))
han <- st_point(x = c(9.7320, 52.3759))
ber <- st_point(x = c(13.405, 52.5200))
```

1862 Daraus könne wir jetzt eine Geometriespalte für einen `data.frame` erstellen

```
geom <- st_sfc(list(goe, han, ber), crs = 4326)
```

1863 Somit haben wir die Geometrie in der Variable `geom` gespeichert, aber noch keine dazugehörigen Attributdaten.

1864 Diese können wir jetzt in einem weiteren `data.frame` abspeichern.

```
attr <- data.frame(
 name = c("Goettingen", "Hannover", "Berlin"),
 bundesland = c("Niedersachsen", "Niedersachsen", "Berlin"),
 einwohner = c(119000, 532000, 3650000)
)
```

1865 In einem letzten Schritt möchten wir jetzt die Geometrie (`geom`) und die Attributdaten (`attr`) zusammenführen.  
1866

```
staedte <- st_sf(attr, geom = geom)
staedte
```

```
1867 ## Simple feature collection with 3 features and 3 fields
1868 ## Geometry type: POINT
1869 ## Dimension: XY
1870 ## Bounding box: xmin: 9.732 ymin: 51.5413 xmax: 13.405 ymax: 52.52
1871 ## Geodetic CRS: WGS 84
1872 ## name bundesland einwohner geom
1873 ## 1 Goettingen Niedersachsen 119000 POINT (9.9158 51.5413)
1874 ## 2 Hannover Niedersachsen 532000 POINT (9.732 52.3759)
1875 ## 3 Berlin Berlin 3650000 POINT (13.405 52.52)
```

1876 Wir können nun mit `staedte` genau so arbeiten wie mit jedem anderen `data.frame` und die Geometrien  
1877 werden immer ‘berücksichtigt’. Zusätzlich kann man eine Reihe von geometrischen Operationen durchführen.

1878 Wenn ein `data.frame` Punkten hat, kann man dies relative einfach mit der Funktion `st_as_sf()` “räumlich”  
1879 machen. Für das vorherige Beispiel würden wir zuerst einen `data.frame` mit allen Informationen (zur  
1880 Geometrie und zu den Attributen erstellen).

```
dat <- data.frame(
 name = c("Goettingen", "Hannover", "Berlin"),
 bundesland = c("Niedersachsen", "Niedersachsen", "Berlin"),
 einwohner = c(119000, 532000, 3650000),
 x = c(9.9158, 9.7320, 13.405),
 y = c(51.5413, 52.3759, 52.5200)
)
```

Dann kann man mit der Funktion `st_as_sf()` weiter arbeiten:

```
staedte1 <- st_as_sf(dat, coords = c("x", "y"), crs = 4326)
```

## 14.4 Arbeiten mit Vektordaten

Es gibt sehr viele Funktionen, um mit räumlichen Daten zu arbeiten, von denen wir hier einige vorstellen.

```
Zeigt das KBS an
st_crs(staedte)
```

```
Coordinate Reference System:
User input: EPSG:4326
wkt:
GEOGCRS["WGS 84",
ENSEMBLE["World Geodetic System 1984 ensemble",
MEMBER["World Geodetic System 1984 (Transit)"],
MEMBER["World Geodetic System 1984 (G730)"],
MEMBER["World Geodetic System 1984 (G873)"],
MEMBER["World Geodetic System 1984 (G1150)"],
MEMBER["World Geodetic System 1984 (G1674)"],
MEMBER["World Geodetic System 1984 (G1762)"],
MEMBER["World Geodetic System 1984 (G2139)"],
ELLIPSOID["WGS 84",6378137,298.257223563,
LENGTHUNIT["metre",1]],
ENSEMBLEACCURACY[2.0]],
PRIMEM["Greenwich",0,
ANGLEUNIT["degree",0.0174532925199433]],
CS[ellipsoidal,2],
AXIS["geodetic latitude (Lat)",north,
ORDER[1],
ANGLEUNIT["degree",0.0174532925199433]],
AXIS["geodetic longitude (Lon)",east,
ORDER[2],
ANGLEUNIT["degree",0.0174532925199433]],
USAGE[
SCOPE["Horizontal component of 3D system."],
AREA["World."],
BBOX[-90,-180,90,180]],
ID["EPSG",4326]]
```

Wenn wir jetzt zu einem anderen KBS (z.B. EPSG:3035, ein europäisches projiziertes KBS) umrechnen möchten, können wir das mit

```
s2 <- st_transform(staedte, 3035)
st_crs(s2)
```

```

1915 ## Coordinate Reference System:
1916 ## User input: EPSG:3035
1917 ## wkt:
1918 ## PROJCRS["ETRS89-extended / LAEA Europe",
1919 ## BASEGEOGCRS["ETRS89",
1920 ## ENSEMBLE["European Terrestrial Reference System 1989 ensemble",
1921 ## MEMBER["European Terrestrial Reference Frame 1989"],
1922 ## MEMBER["European Terrestrial Reference Frame 1990"],
1923 ## MEMBER["European Terrestrial Reference Frame 1991"],
1924 ## MEMBER["European Terrestrial Reference Frame 1992"],
1925 ## MEMBER["European Terrestrial Reference Frame 1993"],
1926 ## MEMBER["European Terrestrial Reference Frame 1994"],
1927 ## MEMBER["European Terrestrial Reference Frame 1996"],
1928 ## MEMBER["European Terrestrial Reference Frame 1997"],
1929 ## MEMBER["European Terrestrial Reference Frame 2000"],
1930 ## MEMBER["European Terrestrial Reference Frame 2005"],
1931 ## MEMBER["European Terrestrial Reference Frame 2014"],
1932 ## ELLIPSOID["GRS 1980",6378137,298.257222101,
1933 ## LENGTHUNIT["metre",1]],
1934 ## ENSEMBLEACCURACY[0.1]],
1935 ## PRIMEM["Greenwich",0,
1936 ## ANGLEUNIT["degree",0.0174532925199433]],
1937 ## ID["EPSG",4258]],
1938 ## CONVERSION["Europe Equal Area 2001",
1939 ## METHOD["Lambert Azimuthal Equal Area",
1940 ## ID["EPSG",9820]],
1941 ## PARAMETER["Latitude of natural origin",52,
1942 ## ANGLEUNIT["degree",0.0174532925199433],
1943 ## ID["EPSG",8801]],
1944 ## PARAMETER["Longitude of natural origin",10,
1945 ## ANGLEUNIT["degree",0.0174532925199433],
1946 ## ID["EPSG",8802]],
1947 ## PARAMETER["False easting",4321000,
1948 ## LENGTHUNIT["metre",1],
1949 ## ID["EPSG",8806]],
1950 ## PARAMETER["False northing",3210000,
1951 ## LENGTHUNIT["metre",1],
1952 ## ID["EPSG",8807]]],
1953 ## CS[Cartesian,2],
1954 ## AXIS["northing (Y)",north,
1955 ## ORDER[1],
1956 ## LENGTHUNIT["metre",1]],
1957 ## AXIS["easting (X)",east,

```

```

1958 ## ORDER[2],
1959 ## LENGTHUNIT["metre",1]],
1960 ## USAGE[
1961 ## SCOPE["Statistical analysis."],
1962 ## AREA["Europe - European Union (EU) countries and candidates. Europe - onshore and offshore: A"],
1963 ## BBOX[24.6,-35.58,84.73,44.83]],
1964 ## ID["EPSG",3035]]

```

1965 Die Funktion `st_buffer()` erlaubt es Features zu puffern, mit `st_distance()` kann die Distanz zwischen  
 1966 Features berechnet werden, mit `st_area()` kann die Fläche eines Features zu berechnen.

1967 Funktionen wie `st_intersection()`, `st_union()` und `st_difference()` erlauben es geometrische Opera-  
 1968 tionen zwischen unterschiedlichen Features zu berechnen. Für eine ausführliche Diskussion siehe auch hier:  
 1969 <https://geocompr.robinlovelace.net/geometric-operations.html>.

1970 Normalerweise lesen wir Daten von externen Datenquellen (z.B. ESRI Shapefiles). Das geht mit der Funktion  
 1971 `st_read()`.

## 1972 14.5 Rasterdaten in R

1973 Für Rasterdaten gibt es das R-Paket **raster**. Auch hier wollen wir uns wieder auf einige Grundfunktionalitäten  
 1974 konzentrieren. Diese umfassen das Einlesen, Zuschneiden, Rechnen und Abfragen von Rastern.

1975 Mit der Funktion `raster()` kann ein Raster in R eingelesen werden.

```

library(raster)
dem <- raster(here::here("data/dem_3035.tif"))

```

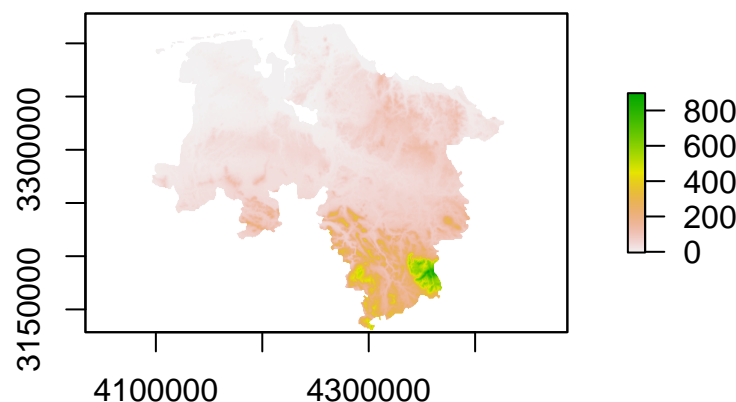
1976 `dem` steht für *Digital Elevation Model* und ist ein Raster mit den Seehöhen in Niedersachsen mit einer  
 1977 500-m-Auflösung. Wir können diese mit der Funktion `res()`<sup>17</sup> abfragen.

```
res(dem)
```

```
[1] 500 500
```

1979 Bzw. wir können den Raster auch plotten.

```
plot(dem)
```



<sup>17</sup>kurz für *resolution* also Auflösung.

Wenn wir den Raster `dem` auf ein Gebiet zuschneiden wollen (z.B. Göttingen), müssen wir drei Schritte durchführen. Als erstes müssen wir ein Shapefile für Göttingen einlesen.

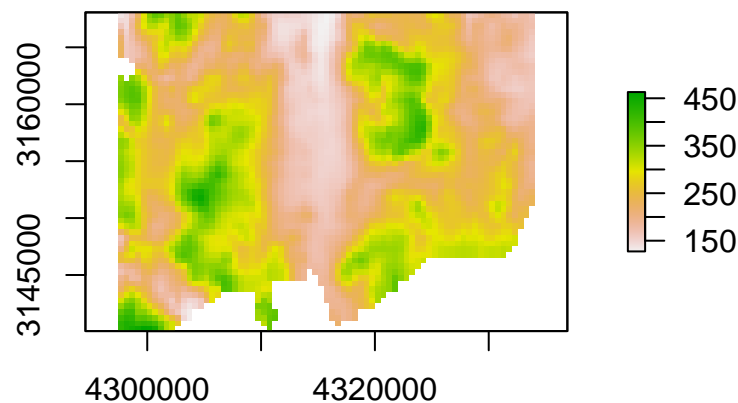
```
goe <- st_read(here::here("data/goettingen/stadt_goettingen.shp"))
```

Dann müssen wir sicher stellen, dass sowohl der Raster `dem` als auch das `sf`-Objekt `goe` im selben KBS sind. Es bietet es sich in der Regel an, das KBS des Vektors zu transformieren. Mit der Funktion `projectRaster()` kann das KBS eines Raster transformiert werden.

```
goe <- st_transform(goe, 3035)
```

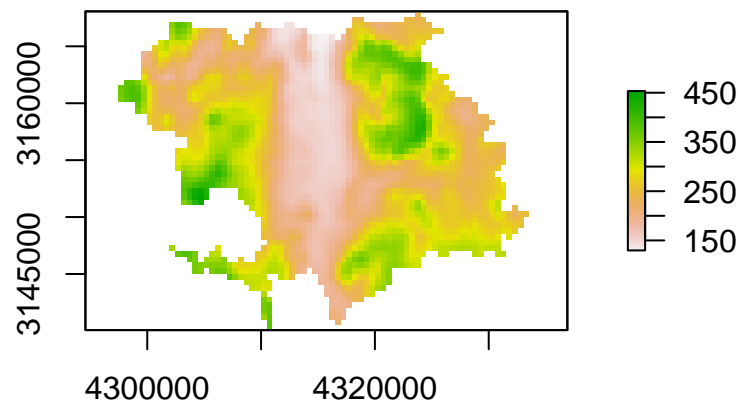
Mit der Funktion `crop()` kann der Raster `dem` auf Göttingen zugeschnitten werden.

```
dem1 <- crop(dem, goe)
plot(dem1)
```



Der Raster hat jetzt die Größe einer Bounding-Box (BBX) von Göttingen (das ist ein Rechteck, das Göttingen umfasst). Mit der Funktion `mask()` kann der Raster auf die genauen Grenzen des Vektors `goe` angepasst werden.

```
dem2 <- mask(dem1, goe)
plot(dem2)
```



Wenn wir an bestimmten Punkten den Wert des Rasters abfragen wollen (z.B. an `cities`) von vorhin, dann gibt es dafür die Funktion `extract`. Dann müssen wir erst sicherstellen, dass `staedte` und `dem` gleichen KBS zu grunde liegt. Dafür transformieren wir einfach `staedte` in das KBS von `dem`. Mit der Funktion `projection()` erhalten wir das KBS des Rasters.

```
s1 <- st_transform(staedte, 3035)
```

Wenn wir das KBS eines Objektes nicht kennen, können wir auch einfach das KBS übergeben. Der folgende Code-Block macht genau das Gleiche mit dem Vorteil, dass wir keinen EPSG-Code angeben müssen.

```
s1 <- st_transform(staedte, projection(dem))
```

Dann können wir für jede Stadt die Seehöhe abfragen:

```
raster::extract(dem, s1)
```

```
[1] 149.18181 57.21486 NA
```

Mit `raster::extract()` rufen wir *eindeutig* die Funktion `extract()` aus dem Paket `raster` auf. Wir müssen das so machen, weil es im Paket `dplyr` auch eine Funktion `extract()` gibt, die wir hier nicht anwenden möchten, da sie einen Fehler verursachen würde.

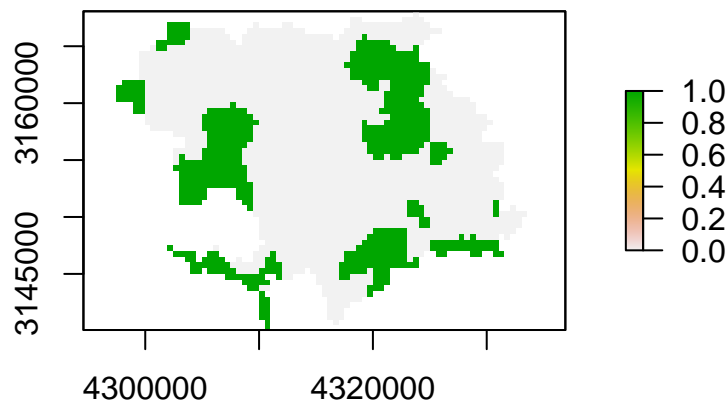
Ein analoges Vorgehen ist auch für Linien und Polygone möglich.

Mit Rastern kann auch einfach gerechnet werden. Wir können z.B. die Seehöhe in Kilometern anstatt Metern berechnen:

```
dem_km <- dem / 1e3
```

Auch logische Operationen sind möglich, wenn wir alle Rasterzellen mit einer Seehöhe von mehr als 300 m in Göttingen suchen, dann geht das so:

```
dem3 <- dem2 > 300
plot(dem3)
```



Wenn wir jetzt auf die Werter des Rasters `dem3` zugreifen wollen, geht das mit eckigen Klammern.

```
head(dem3[])
```

```
[1] NA NA NA NA NA NA
```

Das sind erst einmal viele NA-Werte für die ganzen Zellen, die außerhalb von Niedersachsen liegen. Aber wir können mit so einem Vektor ganz normal arbeiten und z.B. die Fläche des Landes Niedersachsen die eine Seehöhe von mehr als 500m Seehöhe hat ausrechnen.



```
h <- dem3[]
sum(h, na.rm = TRUE) / sum(!is.na(h))
```

```
[1] 0.265713
```

2015

### 2016 *Aufgabe 41: Arbeiten mit Rastern*

2017

2018 Verwenden Sie den Raster `wald.tif`, der auf einer 10 m Auflösung den Waldanteil jeder Rasterzelle angibt<sup>18</sup>.  
 2019 Der EPSG-Code für das KBS von `wald.tif` ist 3035. Nehmen Sie an, dass wenn der Waldanteil in einer  
 2020 Raster größer als 50 % ist, dass die Rasterzelle als Wald klassifiziert werden kann. Wie viel Prozent des  
 2021 Göttinger Stadtgebietes sind Wald? Wie ändert sich dieser Wert, wenn sie 70 % anstatt 50 % als Schwellenwert  
 2022 für Wald annehmen?

2023

### 2024 *Aufgabe 42: Studiendesign*

2025

2026 Mit der Funktion `st_sample()` können Sie innerhalb oder entlang eines Features zufällige Punkte legen. Das  
 2027 Argument `n` steuert die Anzahl Punkte und das Argument `type` wie die Punkte angeordnet werden. Für `type`  
 2028 sind für uns die Werte `type = "random"` (komplett zufällig), `type = "regular"` (regelmäßiger Grid) und  
 2029 `type = "hexagonal"` von Bedeutung (ein hexagonaler Grid, d.h. ein secheckiger Raster). Unglücklicherweise  
 2030 ist das Ergebnis von `st_sample()` erst eine Geometrie. Um daraus ein vollständiges `sf`-Objekt zu machen  
 2031 und problemlos weiter arbeiten zu können, müssen Sie noch einmal die Funktion `st_as_sf()` ausführen.

2032 Stellen Sie sich vor, dass wir die tatsächliche Waldbedeckung des Göttinger Stadtgebietes **nicht** kennen und wir  
 2033 eine Studie durchführen, um den Anteil des Göttinger Stadtgebietes, der mit Wald bedeckt ist herauszufinden.  
 2034 Erstellen Sie dafür einige unterschiedliche Stichproben (diese können in der Anzahl und Anordnung variieren).

2035 Berechnen Sie für jedes Stichprobendesign den Anteil an Wald und ein dazugehöriges Konfidenzintervall  
 2036 (dieses können Sie mit der Formel  $\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$  berechnen, wobei  $\hat{p}$  der geschätzte Waldanteil ist und  $n$   
 2037 die Stichprobengröße). Nehmen Sie an, dass eine Rasterzelle Wald ist, sobald  $> 50\%$  der Rasterzelle mit  
 2038 Wald bedeckt ist.

2039

### 2040 *Aufgabe 43: Räumliche Daten*

2041

2042 Verwenden Sie den folgenden Datensatz:

```
set.seed(123)
df1 <- data.frame(
 x = runif(100, 0, 100),
```

<sup>18</sup>Die können hier <https://land.copernicus.eu/pan-european/high-resolution-layers/> für ganz Europa bezogen werden

```

y = runif(100, 0, 100),
kronendurchmesser = runif(100, 1, 15),
art = sample(letters[1:4], 100, TRUE)
)

```

1. Erstellen Sie ein `sf`-Objekt aus `df1`.
2. Puffern Sie jeden Baum mit seinem Kronendurchmesser.
3. Berechnen Sie die Kronenfläche jedes Baumes. *Hinweis: Die Funktion `st_area()` könnte dafür hilfreich sein.*
4. Welcher Baum hat die größte Kronenfläche?
5. Finden Sie für jede Art, den Baum mit der größten Kronenfläche.

#### Aufgabe 44: Arbeiten mit räumlichen Daten

1. Lesen Sie das ESRI Shapefile `goettingen/stadt_goettingen.shp` ein.
2. Wie viele Features befinden sich in dem Shapefile?
3. Welches Koordinatenbezugssystem (KBS) hat das Shapefile?
4. Transformieren Sie das Shapefile in das KBS 3035.
5. Erstellen Sie eine neue Spalte `A` in der Sie die Fläche jeder Gemeinde/Stadt speichern.
6. Welche Gemeinde/Stadt (Spalte `GEN`) ist am größten?
7. Wählen Sie nun nur die Stadt Göttingen aus.

#### Aufgabe 45: Arbeiten mit räumlichen Daten 2

1. Lesen Sie erneut das ESRI Shapefile `goettingen/stadt_goettingen.shp` ein.
2. Lösen Sie die Gemeindegrenzen auf (die Funktion `st_union()` könnte hier nützlich sein).
3. Wie groß ist das resultierende Feature?

## 15 FAQs (Oft gefragtes)

### 15.1 Arbeiten mit Daten

#### 15.1.1 Einlesen von Exceldateien

Mit der Funktion `read_excel()` aus dem Paket `readxl` können Exceldateien direkt in R eingelesen werden.

Ein Export als csv-Datei aus Excel ist nicht notwendig.

## 16 Zusätzliche Aufgaben

### Aufgabe 46: Standardisierung

Unter Standardisierung (oder auch z-Transformation) versteht man die Transformation einer Variable, so dass sie den Mittelwert 0 und die Varianz 1 hat. Die Formel für die Standardisierung ist

$$x_s = \frac{x - \mu_x}{\sigma_x}$$

wobei  $x$  die Variable ist,  $\mu_x$  ist der Mittelwert von  $x$  und  $\sigma_x$  ist die Standardabweichung von  $x$ .

Standardisieren Sie folgenden Vektor:

```
h <- c(0, 2, 3, 1, 0, 8, 3.4, 9, 6.8, 2.1)
```

Und speichern Sie das Ergebnis in `h_s`. Vergewissern Sie sich, dass die Standardisierung geklappt hat und berechnen Sie den Mittelwert und Standardabweichung von `h_s`.

### Aufgabe 47: Arbeiten mit logischen Werten

Verwenden Sie nochmals den Vektor mit der Anzahl Rehe, die an unterschiedlichen Fotofallenstandorten fotografiert wurden.

```
anzahl_rehe <- c(132, 79, 129, 91, 138, 144, 55, 103, 139, 105, 96, 146, 95, 118, 1007)
```

Für wie viele Standort trifft die Aussage zu  $90 \leq x < 120$ , wobei  $x$  für die Anzahl Fotos an einem Standort steht.

### Aufgabe 48: Auswählen von Elementen in einem Vektor

Lesen Sie die Datei `bhd_1.txt` ein. Und bearbeiten Sie folgende Aufgaben mit dieser Datei:

- Finden Sie den mittleren BHD aller Eichen.
- Wie viele Beobachtungen haben Sie für Eichen, Fichten und Buchen?
- Finden Sie alle Bäume, die 10, 20, 21, 23, 30, 37, 78, 79, 90, 91, 92 Jahre alt sind.

2094

---

2095 **Aufgabe 49: Arbeiten mit Daten**


---

2096

2097 Wang et al. (2019) haben in einer Fotofallenstudie das Verhalten und die Habitatselektion von Ozeloten  
 2098 im brasilianischen Amazonas untersucht. Ziel dieser Übung ist es mit dem Datensatz etwas vertraut zu  
 2099 werden, wir werden noch keine ökologischen Analysen durchführen. Mehr zu dem Datensatz erfahren Sie [hier](#).  
 2100 Eine etwas angepasste Version des Datensatzes können Sie aus dem StudIP Ordner **daten** (die Datei heißt  
 2101 **ozelote.zip**) herunterladen. Speichern Sie die Datei in ihrem RStudio Projekt und entzippen Sie sie. Der  
 2102 Ordner enthält zwei Dateien, für diese Übung brauchen wir lediglich die Datei **ozelote\_standorte.csv**, die  
 2103 für jeden Fotofallen Standort einige Kovariaten angibt.

2104 Bearbeiten Sie folgende Aufgaben:

- 2105 1. Lesen Sie die Datei **ozelote\_standorte.csv** in R und speichern Sie das Ergebnis in eine Variable  
 2106 **standorte**.
- 2107 2. Wie viele Fotofallenstandorte gab es in der Studie?
- 2108 3. Welcher Standort ist am Höchsten gelegen? Die Spalte **seehoehe** enthält die mittlere Seehöhe.
- 2109 4. Finden Sie alle Standorte, die in unmittelbarer Nähe zu Flüssen sind. Eine Distanz von  $< 5$  m kann als  
 2110 Schwellenwert angenommen werden. Die Spalte **dist\_fluss** gibt die Distanz zu Flüssen an.
- 2111 5. Der Datensatz besteht aus verschiedenen Kameras, die jeweils für einen Zeitraum von 12 Tagen in einer  
 2112 Region aufgestellt wurden (Spalte **Region**). Erstellen Sie einen Plot, der den Zusammenhang zwischen  
 2113 der Region und Seehöhe darstellt.

2114

---

2115 **Aufgabe 50: Base Plots**


---

2116

2117 Erstellen Sie die folgende Beispielabbildung Schritt für Schritt selbst über Low-Level Funktionen. Die Rohdaten  
 2118 finden Sie in den Dateien **abbBeispiel.R** und **ertragstafeldaten.csv**.

- 2119 • Die Wachstumskurve der Region 1 (blau) lautet  $41.45752(1 - \exp(-0.02168x))^{1.61787}$
- 2120 • Die Wachstumskurve der Region 2 (rot) lautet  $51.11203(1 - \exp(-0.009129x))^{1.202401}$

2121 wobei  $x$  das Baumalter in Jahren angegeben ist. Die 3 schwarzen Linien sind auf der Ertragstafel abgelesen.  
 2122 Die Beschriftungen der 3 Ertragstafelkurven, sowie des Ausreißers, sind Zusatzaufgaben.

2123

---

2124 **Aufgabe 51: ggplot2 Aufgabe**


---

2125

- 2126 1. Laden Sie den Datensatz **daten/bhd\_1.txt**
- 2127 2. Erstellen Sie ein Streudiagramm. Bilden Sie dabei den BHD gegen das Alter ab, wobei dies als Subplot  
 2128 für jedes Affnahmegebiet dargestellt werden sollte.
- 2129 3. Verwenden Sie für jede Baumart eine eigene Farbe.
- 2130 4. Erstellen Sie für jede Baumart einen Boxplot des BHDs.

5. Teilen Sie die Boxplots aus 4) auf jeweils einen Subplot pro Aufnahmegebiet auf.

### Aufgabe 52: Anwendungsbeispiel kontrollierter Programmabläufe

- Öffnen Sie ein neues, leeres R Skript.
- Laden Sie die Datei "stichprobe.csv" in eine Variable.

```
stpr <- read.csv("data/stichprobe.csv", fileEncoding = "UTF-8")
```

- Filtern Sie den Data Frame so, dass er nur noch die Baumart "Eiche" enthält. Speichern Sie den gefilterten Data Frame in einer NEUEN Variable ab.
- Berechnen Sie die deskriptiven Statistiken `mean()`, `sd()`, `median()`, `min()` und `max()` des Kapitels "Deskriptive Statistik" für den BHD (des gefilterten Data Frames).
- Erstellen Sie ein Histogramm des BHD (ebenfalls mit dem gefilterten Data Frame), zeichnen Sie den arithmetischen Mittelwert als horizontale Linie in das Histogramm ein.
- Speichern Sie den R Code und kopieren Sie ihn in ein neues R Skript.
- Erstellen Sie nun eine Schleife, die alle Statistiken und auch die Abbildung für jede Baumart berechnet. Lassen Sie die Statistiken mit `print()` in die Konsole ausgeben.
- ZUSATZ: Exportieren Sie die Histogramme (bspw. als PDF). TIPP: Verwenden Sie `paste()` um sinnvolle Namen für die Dateien zu erstellen. Machen Sie sich selbst mit der Funktion vertraut.
- ZUSATZ: Sie wollen Fehlermeldungen vermeiden. Deshalb programmieren Sie eine bedingte Ausführung, um die gesamten statistischen Berechnungen und auch die Abbildung. Führen Sie Ihren gesamten Code nur unter der Bedingung aus, dass die Baumart "Ei", "Bu", "Fi", "Kie" oder "Dou" ist. TIPP: Sie können den `%in%` Operator verwenden.

## 16.1 Arbeiten mit Daten

Verwenden Sie erneut die Datensatz von Wang et al. (2019) zu Ozeloten in Brasilien für die nachfolgenden Übungen.

### Aufgabe 53: Datenzusammenfassen

1. Laden Sie die Datei `ozelote_standorte.csv` in R und speichern Sie das Ergebnis in eine Variable `standorte`.
2. Berechnen Sie die Anzahl an Fotofallen für jede Region. Welche Region weist die meisten Fotofallen auf?
3. In welcher Region ist die größte Variabilität der Seehöhe zu finden?
4. In welchen Region beträgt der Anteil an Fotofallen, die < 5m vom nächsten Fluss entfernt sind, mindestens 20%?

2165

---

2166 **Aufgabe 54: Datenmanipulation 1**

---

2167

- 2168 1. Laden Sie nun zusätzlich die Datei `ozelote_fanghistorien.csv` und speichern Sie diese in die Variable  
2169 (`fh`). In diesem `data.frame` gibt es für jede Session eine Spalte (`V1` bis `V10`). Eine 1 bedeutet, dass  
2170 mindestens ein Ozelot fotografiert wurde und eine 0 bedeutet, dass kein Ozelot in diesem Zeitraum  
2171 fotografiert wurde. `NA` heißt, dass die Kamera nicht aktiv war.
- 2172 2. Wählen Sie nur das 3. Fangereignis (das ist die Spalte `V3`).
- 2173 3. Wie viele Kameras waren beim 3. Fangereignis aktiv?
- 2174 4. Vergleichen Sie anhand einer Abbildung, ob sich die Distanz zum Fluss (Spalte `dist_fluss`) zwischen  
2175 Standorten mit Fotos (`V3 == 1`) und Standorten ohne Fotos (`V3 == 0`) unterscheidet.

2176

---

2177 **Aufgabe 55: Datenmanipulation 2 (etwas knifflig)**

---

2178

- 2179 1. Verwenden Sie erneut die Daten zu den Fotofallenstandorten und Fanghistorien der Ozelote.
- 2180 2. Finden Sie alle Fotofallenstandorte an denen  $\geq 3$  Ozelote fotografiert wurden?
- 2181 3. Gibt es einen Zusammenhang zwischen der Häufigkeit an Ozelotfotos (pro Fotofallenstandort) und der  
2182 Distanz zum nächsten Fluss (Spalte `dist_fluss`)? Eine Abbildung ist ausreichend.

## 17 Literatur

Ein guter Überblick über viele der angeschnittenen Themen gibt es in dem [R for data science](#), das online frei zugänglich ist.

McNamara, Amelia, and Nicholas J Horton. 2018. “Wrangling Categorical Data in r.” *The American Statistician* 72 (1): 97–104.

Wang, Bingxin, Daniel G. Rocha, Mark I. Abrahams, André P. Antunes, Hugo C. M. Costa, André Luis Sousa Gonçalves, Wilson Roberto Spironello, et al. 2019. “Habitat Use of the Ocelot (*Leopardus Pardalis*) in Brazilian Amazon.” *Ecology and Evolution* 9 (9): 5049–62. <https://doi.org/10.1002/ece3.5005>.

Wickham, Hadley. 2014. “Tidy Data.” *Journal of Statistical Software* 59 (10). <https://doi.org/10.18637/jss.v059.i10>.