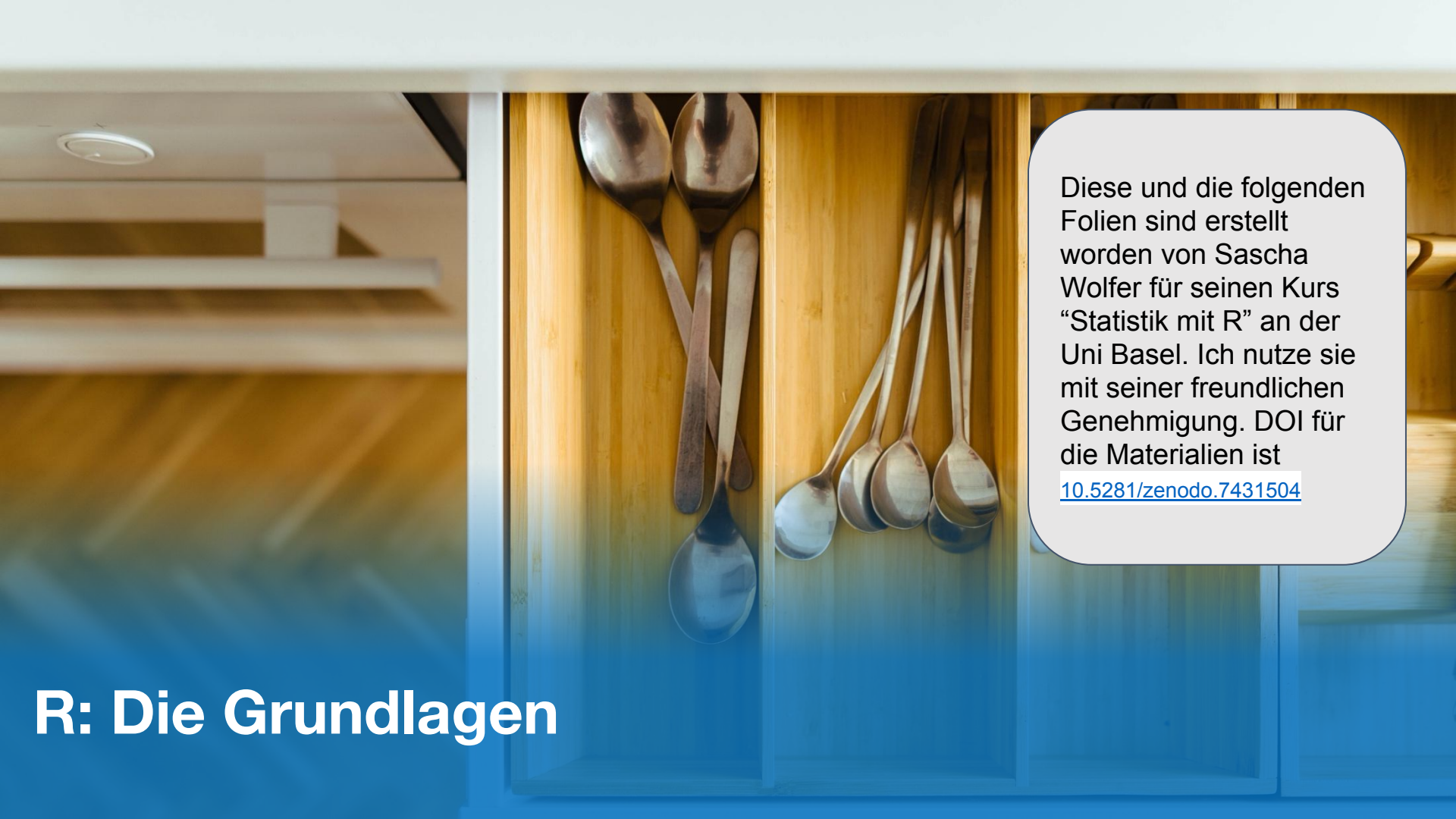
Einführung in die Statistik

Praktische Übung – Jürgen Hermes – IDH – SoSe 2023



R: Die Grundlagen

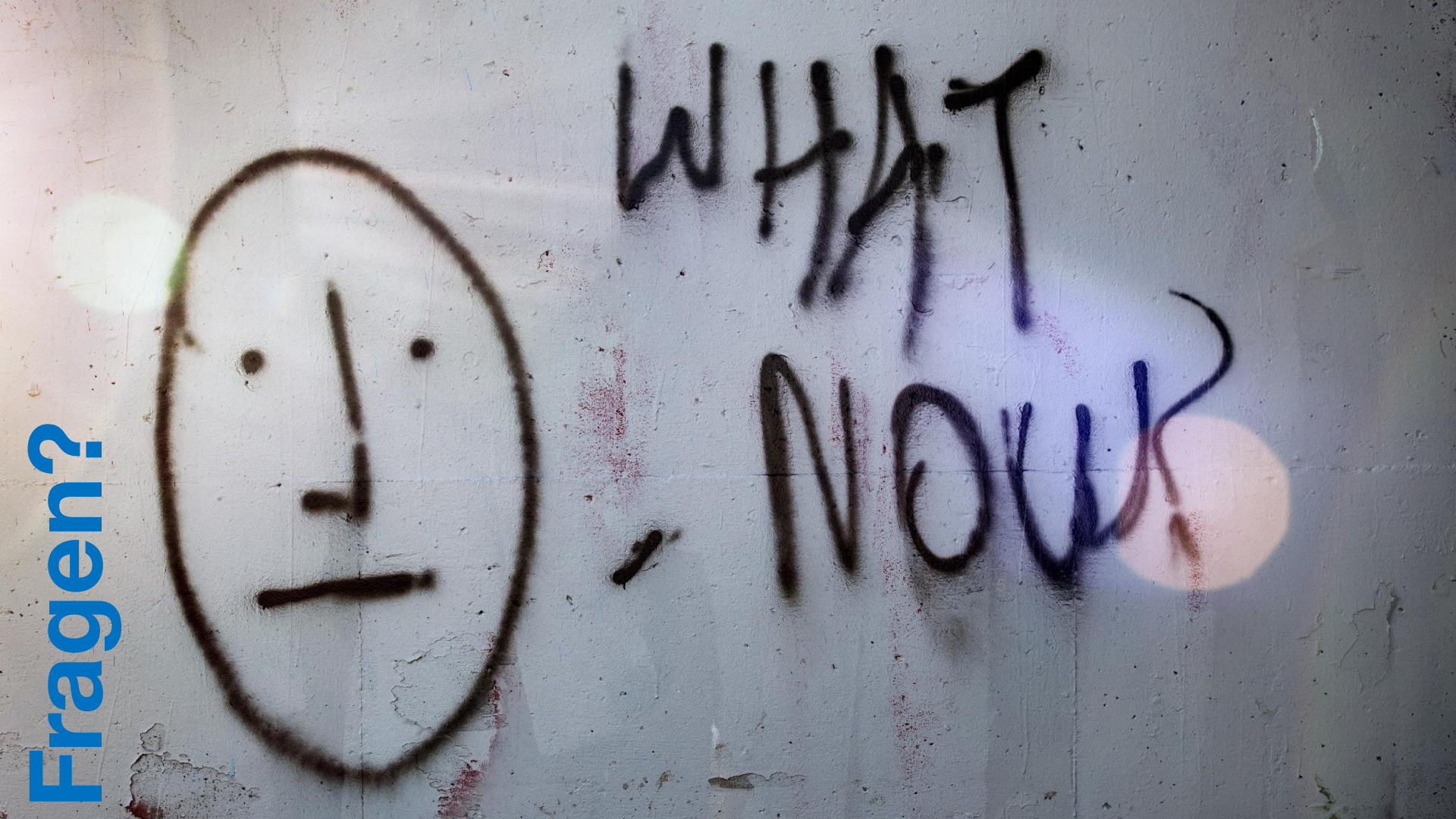
Diese und die folgenden Folien sind erstellt worden von Sascha Wolfer für seinen Kurs “Statistik mit R” an der Uni Basel. Ich nutze sie mit seiner freundlichen Genehmigung. DOI für die Materialien ist 10.5281/zenodo.7431504

Sortieren = Indizierung

- `order(<Vektor>)` gibt die Rangfolge der Elemente in einem Vektor zurück.
  - Bei Zeichenketten: alphabetisch, bei Zahlen: von klein nach groß
  - `order(c("D", "B", "C", "A"))` → `[1] 4 2 3 1`
  - `order(c("B", "C", "A"))` → `[1] 3 1 2`
  - Zurückgegeben wird ein Vector, der die Reihenfolge der Indizes enthält!
- Deshalb können wir Dataframes sortiert ausgeben, wenn wir `order()` verwenden, um auf Zeilen zuzugreifen.
  - `df` aufsteigend nach Spalte `freq` sortiert ausgeben: `df[order(df$freq),]`
  - Absteigend sortieren: `order(..., decreasing = T)`

Zusammenfassung

- **Vektoren**, **Dataframes**, Listen und Matrizen sind die wichtigsten komplexen Datentypen.
  - Dataframes kann man sich vorstellen als nebeneinander gestellte Vektoren.
  - Matrizen sind mehrdimensionale Vektoren.
  - Listen können alle anderen Datentypen enthalten, auch Listen selbst.
- Wir können auf diese zugreifen (sie indizieren) über
  - Zahlen (= Stellen/Indizes): `vec[3]` oder `df[3,]`
  - Namen: `df$wort` oder `df[5, "wort"]`
  - Wahrheitswerte: `vec[c(T, F)]`



Fragen?

Übung

- Erstellen Sie fünf Vektoren:
  - Vektor `user` mit den Werten `["km", "smv", "sw", "al"]`
  - Vektor `tweets` `[18948, 11314, 2440, 14610]`
  - Vektor `followers` `[3584, 3609, 719, 2543]`
  - Vektor `follows` `[1374, 548, 877, 1059]`
  - Vektor `face.in.profile` `[T, F, T, T]`
- Kombinieren Sie die Vektoren in einem Dataframe `twitterData`.
- Extrahieren Sie die erste Zeile.
- Extrahieren Sie die zweite und dritte Spalte.
- Extrahieren Sie den Wert `877`.
- Extrahieren Sie die User, die ein Bild von sich im Profil anzeigen.
- Sortieren Sie den Dataframe nach der Spalte `followers`.

Funktionen

- Befehle an R
- **Argumente**:
  - Mit **was** soll etwas gemacht werden?
  - **Was** soll genau getan werden?
  - Manche Argumente haben "Default"-Werte
- Rückgabe(-Wert): Ergebnis der Funktion
  - Der Rückgabewert wird im Terminal ausgegeben.
  - Manche Funktionen (z.B. Plots) haben keinen Rückgabewert, sondern werden wegen ihres **Nebeneffekts** benutzt.

Funktionen

- Funktionen können ineinander **verschachtelt** werden. Sie werden dann **von innen nach außen** evaluiert/ausgewertet.
  - `sqrt(sum(c(5, 3, 1)))`
  - `max(nchar(c("HGW", "MA", "GSWG")))`
  - `as.character(min(c(9, 8, -1)) + max(1:5))`
  - `c(9, 3, 42)[sqrt(mean(c(7, 9, 11)))]`
  - `which()`

Paket

Data Frames

Description

The function `data.frame()` creates data frames, tightly coupled collections of variables which share many of the properties of matrices and of lists, used as the fundamental data structure by most of R's modeling software.

Usage

Default-Werte

```
data.frame(..., row.names = NULL, check.rows = FALSE,
           check.names = TRUE, fix.empty.names = TRUE,
           stringsAsFactors = default.stringsAsFactors())

default.stringsAsFactors()
```

genaue Beschreibung der Argumente

Arguments

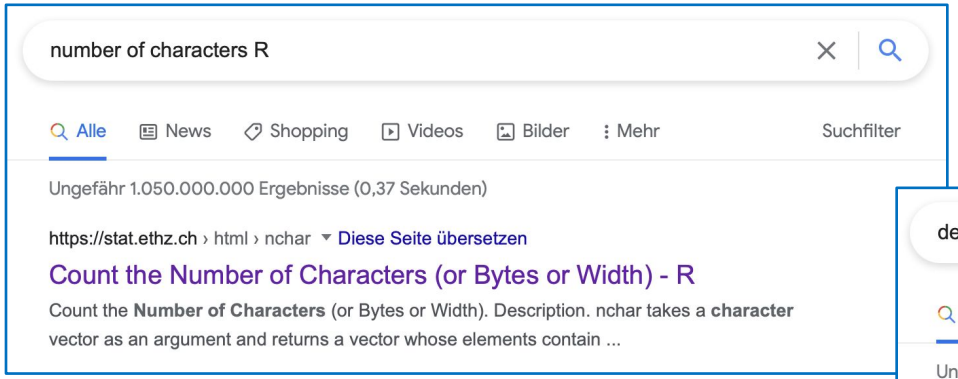
`...`	these arguments are of either the form `value` or `tag = value`. Component names are created based on the tag (if present) or the deparsed argument itself.
`row.names`	`NULL` or a single integer or character string specifying a column to be used as row names, or a character or integer vector giving the row names for the data frame.
`check.rows`	if `TRUE` then the rows are checked for consistency of length and names.
`check.names`	logical. If `TRUE` then the names of the variables in the data frame are checked to ensure that they are syntactically valid variable names and are not duplicated. If necessary they are adjusted (by `make.names`) so that they are.
`fix.empty.names`	logical indicating if arguments which are "unnamed" (in the sense of not being formally called as `someName = arg`) get an automatically constructed name or rather name `""`. Needs to be set to `FALSE` even when `check.names` is false if `""` names should be kept.
`stringsAsFactors`	logical: should character vectors be converted to factors? The 'factory-fresh' default is `TRUE`, but this can be changed by setting `options(stringsAsFactors = FALSE)`.

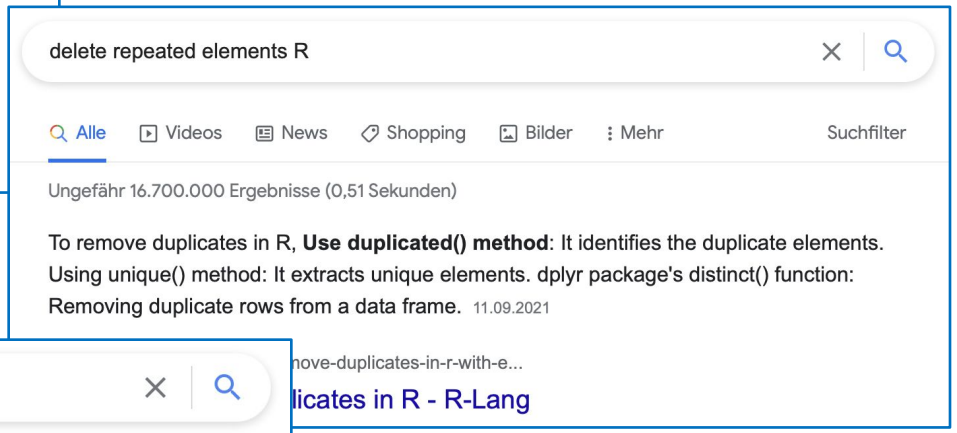
Details

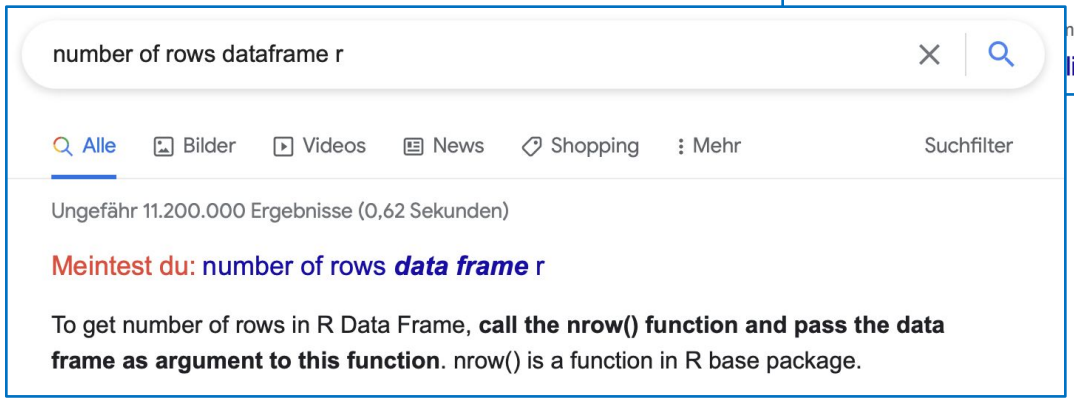
A data frame is a list of variables of the same number of rows with unique row names, given class `"data.frame"`. If no variables are included, the row names determine the number of rows.

Einige Funktionen

`min() / max()`	Minimal- / Maximalwert in Vektor
`mean()`	Mittelwert
`median()`	Median
`nchar()`	Anzahl Zeichen in Zeichenkette(n)
`length()`	Anzahl Elemente in Vektor
`ncol() / nrow()`	Anzahl Spalten / Zeilen in Dataframe
`unique()`	Gibt jedes Element / jede Zeile nur einmal zurück
`summary()`	Gibt eine Zusammenfassung des Arguments

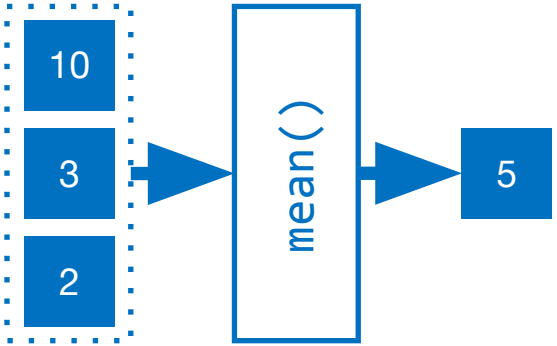

number of characters R
Alle
News
Shopping
Videos
Bilder
Mehr
Suchfilter
Ungefähr 1.050.000.000 Ergebnisse (0,37 Sekunden)
https://stat.ethz.ch › html › nchar Diese Seite übersetzen
Count the Number of Characters (or Bytes or Width) - R
Count the Number of Characters (or Bytes or Width). Description. nchar takes a character vector as an argument and returns a vector whose elements contain ...


delete repeated elements R
Alle
Videos
News
Shopping
Bilder
Mehr
Suchfilter
Ungefähr 16.700.000 Ergebnisse (0,51 Sekunden)
To remove duplicates in R, Use duplicated() method: It identifies the duplicate elements. Using unique() method: It extracts unique elements. dplyr package's distinct() function: Removing duplicate rows from a data frame. 11.09.2021
nove-duplicates-in-r-with-e...
licates in R - R-Lang


number of rows dataframe r
Alle
Bilder
Videos
News
Shopping
Mehr
Suchfilter
Ungefähr 11.200.000 Ergebnisse (0,62 Sekunden)
Meintest du: number of rows data frame r
To get number of rows in R Data Frame, call the nrow() function and pass the data frame as argument to this function. nrow() is a function in R base package.

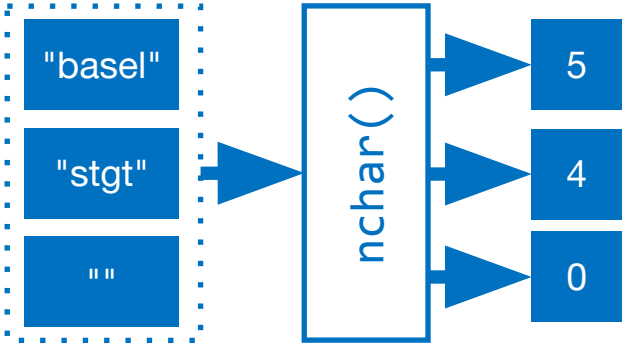
Rückgabewerte

```
mean(c(10, 3, 2))
```



Genau ein Rückgabewert

```
nchar(c("basel", "stgt", ""))
```



Ein Rückgabewert pro Element

Zwischenfazit: Funktionen

Nebenbei: Man kann sich auch eigene Funktionen schreiben. Dazu aber später mehr.

- Funktionen sind Handlungsanweisungen mit **Argumenten**
  - Mit was wird etwas gemacht?
  - Was wird genau gemacht?
- Die meisten Funktionen haben **Rückgabewerte**.
  - Die Form der Rückgabe variiert je nach Funktion.
  - Die Rückgabewerte kann man als Argument für weitere Funktionen benutzen (Verschachteln von Funktionen).

Prädikate

- Prädikate sind ganz spezielle Funktionen, denn sie geben **immer** `TRUE` oder `FALSE` zurück.
- Prädikate sind also Fragen an R, ob etwas bestimmtes gilt.

`>` `<` `==` `!=` `>=` `<=`

`%in%` `is.character()`

Übung: Prädikate

- Ist 10 größer als 10?
- Ist 10 größer gleich 10?
- Befindet sich 10 in einem Vektor, der von 2 bis 11 geht?
- Befindet sich die Zeichenkette `"a"` im Character-Vektor `["uni", "basel"]`?
- Ist `10` eine Zahl?
- Ist `"10"` eine Zahl?
- Ist *unendlich* eine Zahl?

Wozu braucht man Prädikate?

- Wir können außerdem über **Wahrheitswerte** zugreifen (das wird später nochmal wichtig!):

```
> vec <- c("das", "ist", "ein", "vektor")
> vec[c(T, T, F, T)]
[1] "das"    "ist"    "vektor"
```

Indizierung

- Wir können Fälle in Dataframes nach bestimmten Bedingungen **filtern**:

```
students[students$age >= 20,]
```

Gibt alle Fälle (Zeilen) zurück, bei denen die Spalte `laenge` größer 5 ist.

Fallunterscheidungen

- Wir können Spalten in Abhängigkeit der Werte in anderen Spalten definieren:

```
students$ageKat <- ifelse(students$age
    < 20, "Teen", "Twen")
```

Wenn in Spalte age ein Wert kleiner 20 steht, schreiben wir in Spalte `ageKat` `"Teen"`, ansonsten `"Twen"`.

Logische Operatoren

- Operatoren verknüpfen Wahrheitswerte
- Logisches UND: **&**
  - Beide Bedingungen müssen erfüllt sein.
- Logisches ODER: **|**
  - Mindestens eine Bedingung muss erfüllt sein.
- "Nicht"/Umkehren des Wahrheitswerts: **!**
  - Aus `TRUE` wird `FALSE`, aus `FALSE` wird `TRUE`.

Indizierung mit Prädikaten und Operatoren

- Welche Fälle werden ausgewählt?

```
students[students$gender == "m",]

students[students$gender %in% c("d", "w") & students$age > 20,]

students[students$age %in% 17:20 & students$secondStudyP != "History",]

students[students$secondStudyP == "Linguistics" | students$grade < 2,]

students[!(students$secondStudyP %in% c("Linguistics", "History")),]
```

Begriffe

Dataframes

Matrizen

Listen

Indizierung

[] [[]] [,] $

Funktionen

Argumente

Rückgabewert

Nebeneffekt

Verschachteln

Default-Werte

Prädikate

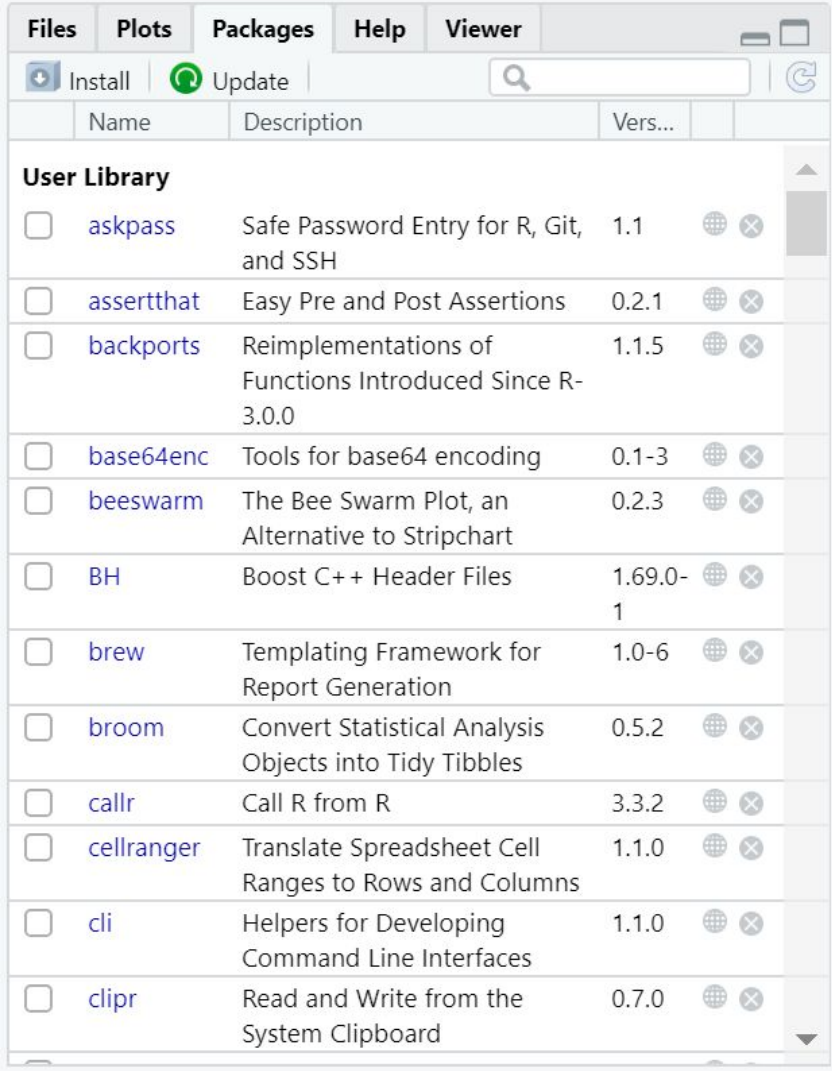
Log. Operatoren

& | !


Fragen?

Zusatzpakete

- R kann mit (sehr vielen) Zusatzpaketen erweitert werden.
- Befehl: `install.packages("<Paketname>")`
- "Packages" pane in RStudio
- Packages werden häufig mit geschweiften Klammern genannt: {beeswarm}
- `data.table::fwrite()` bedeutet: Die Funktion `fwrite()` aus dem Paket {data.table}
- Installierte Pakete laden mit `library(<Paketname>)`



Übung

- Installieren Sie das Paket `{openintro}`.
- Schauen Sie sich die ersten 6 Zeilen des Dataframes `cia_factbook` an.
- Lassen Sie sich die Zeile für die Schweiz ausgeben.
- Lassen Sie sich die Zeilen für die Schweiz, Österreich und Deutschland ausgeben.
- Welches Land hat die größte Fläche im Datensatz?
  - Hier taucht ein Problem mit nicht definierten Werten (NA) auf, das man mit is.na() behandeln kann
- Berechnen Sie den Anteil von Internetbenutzer*innen für jedes Land, speichern Sie diesen Wert in der Spalte `www_percent`.

Hausaufgabe

- Evtl. haben Sie die Hausaufgabe zur letzten Woche nicht vollständig gelöst (Abgabefrist ist verlängert worden)
- Aufgaben zu dieser Woche werden nach der Unterrichtsstunde auf Ilias gestellt (wenn klar ist, wie weit wir gekommen sind)
- Die Studienleistung wird vergeben aufgrund der Ergebnisse in den Testaten. Sollten Sie *eines* der Testate entschuldigt (Attest!) versäumen, *könnte* ggfs. die regelmäßige, rechtzeitige, vollständige und korrekte Abgabe der HA in die Bewertung mit einbezogen werden statt ein Nachtestat anzusetzen. [Hinweis: Das ist kein Automatismus!]