Einführung in die Datenanalyse mit R: Objekte, Konventionen und Datenimport

Marco Wähner


Gebäude 37.03.03.14


marco.waehner@hhu.de


Sprechstunde nach Vereinbarung

Wiederholung

- In R wird alles in Objekten gespeichert (Variablen, Daten, Funktionen etc.)
- Objekte haben verschiedene Klassen (und Typen)
- Objekte erhalten einen Wert durch Zuweisung
  - Dafür wird das „assignment“-Zeichen verwendet <-
  - Kleiner-als + minus
  - Shortcut: alt + minus

Wiederholung

- Kleine Aufgabe:

```
d <- 0
a <- b <- c <- d
print(a)


z <- 5
x <- y <- z
print(x)
```

Objekte

Objekte

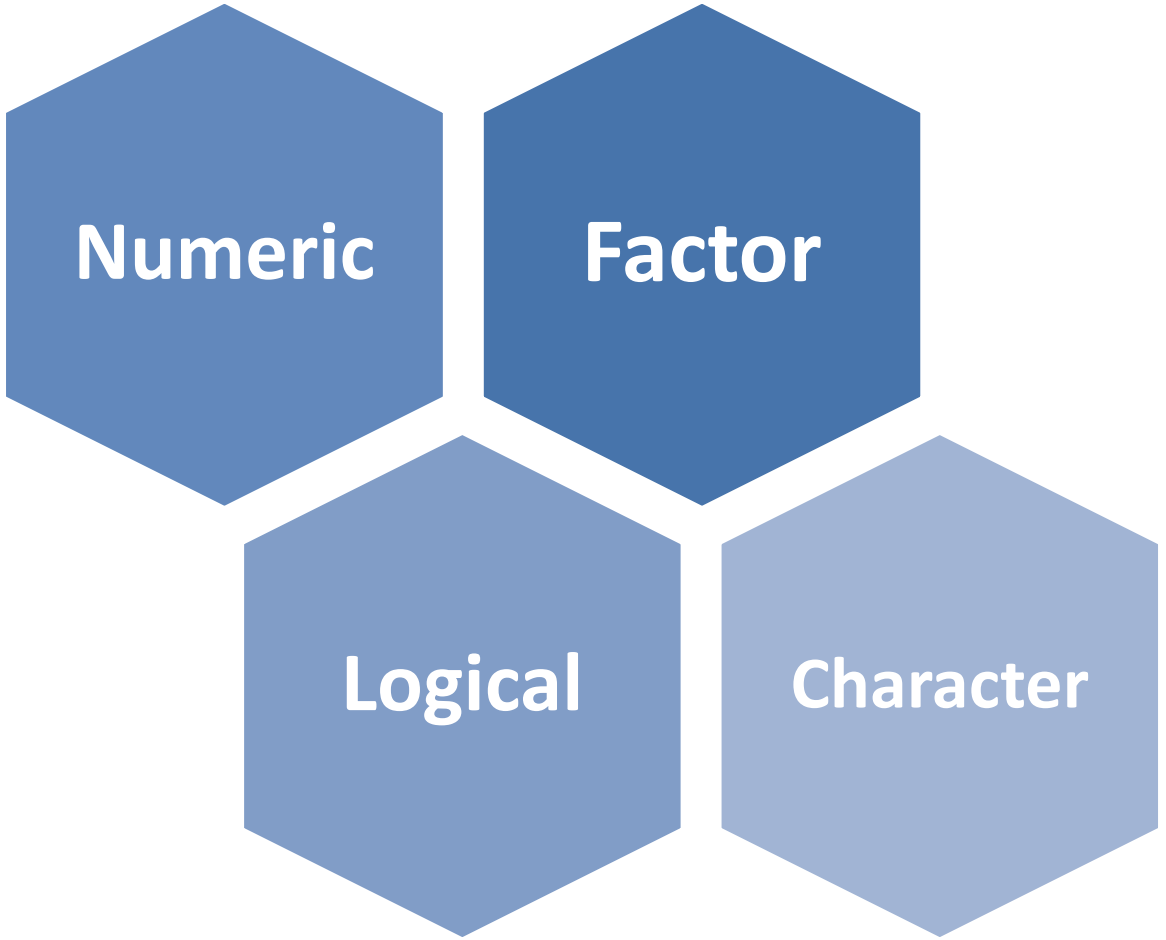
- Objekte haben unterschiedliche Klassen

```
#Objekte und Klassen
number <- 1 + 1
class(number)

## [1] "numeric"
```

```
#Objekte und Klassen
text <- "hello world!"
class(text)

## [1] "character"
```

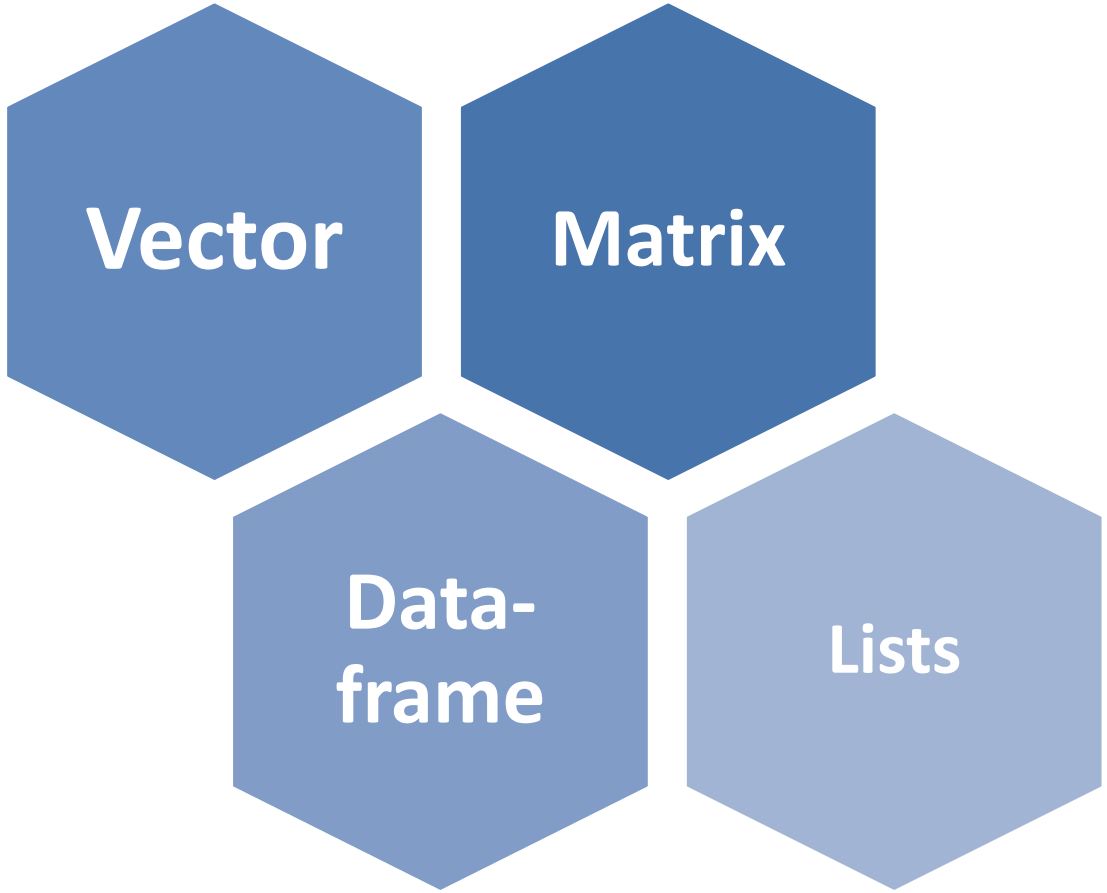
Klassen

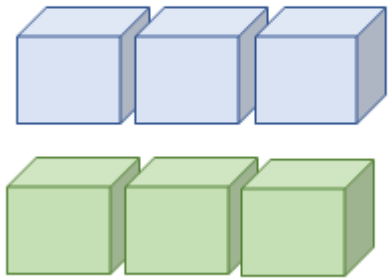
- Wesentliche Klassen sind …



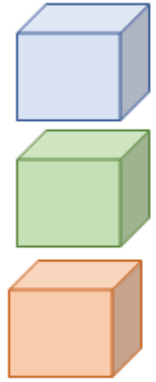
Objekte und Formate

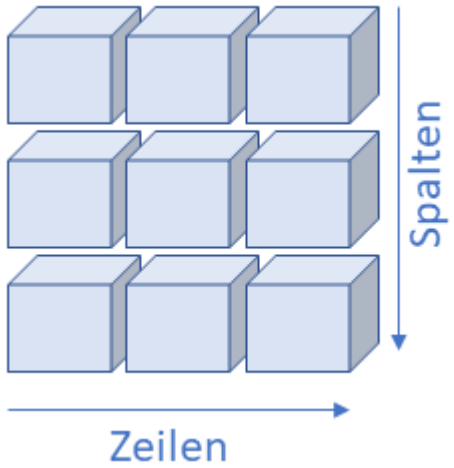
- Objekte (mit der jeweiligen Klasse) werden in unterschiedlichen Formaten gespeichert, dazu zählen …



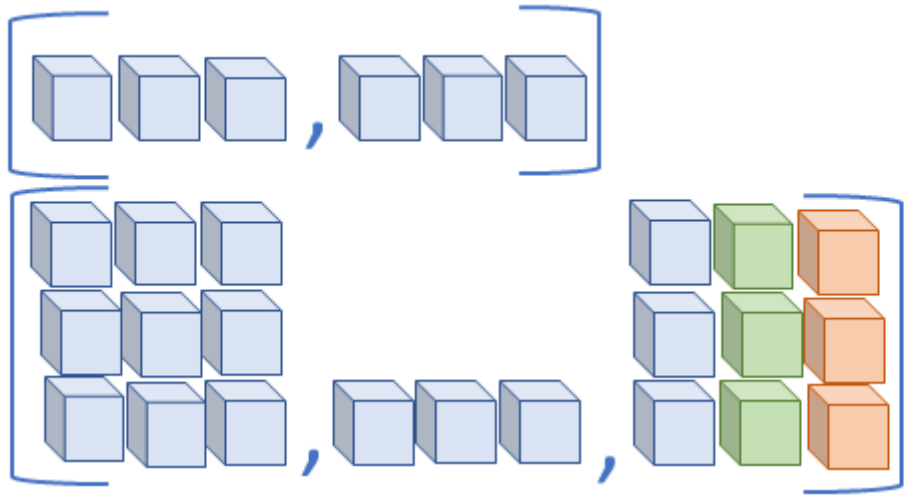


Vektor

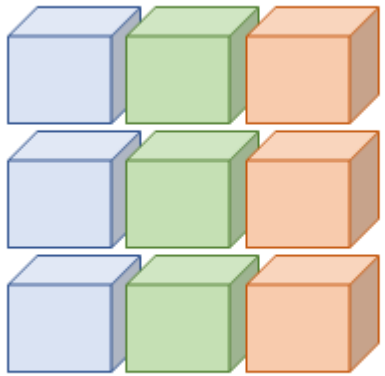




Matrix



Lists



Data Frame

Quelle: https://devopedia.org/r-data-structures

Missing Values

- Missing Values als spezifische Klasse
  - R kennt nur einen Wert (und nicht wie bei SPSS 77,777,7777) für Missing Values: NA
  - NA hat nicht die Klasse character

```
#Klasse NA
missing <- NA
class(missing)
```

```
## [1] "logical"
```

```
#Variable mit Missing Values
var_1 <- c(1:25, NA, NA, NA, NA, NA)
#test auf Missing Values
is.na(var_1)
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [25] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
```

```
#Summe NA
sum(is.na(var_1))
```

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

Praxis

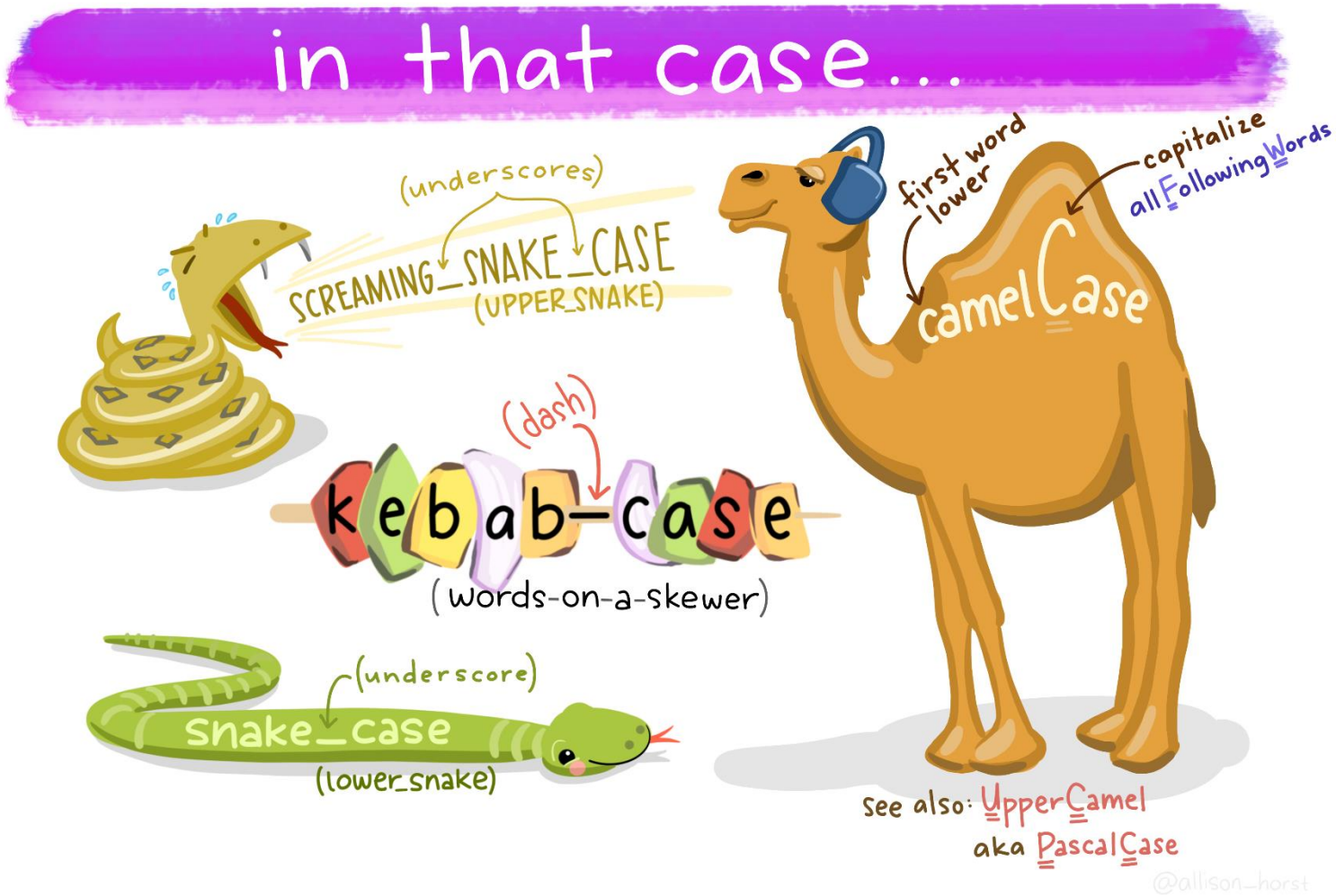
Good Practice & Konventionen

Good Practice & Konventionen

Konventionen

- Style Guides
  - Google‘s R Style Guide
  - Tidyverse Style Guide
  - Do not repeat yourself!
- Bezeichnung für Variablen, Funktionen, Daten etc.
  - R ist case-sensitive! Es macht also einen Unterschied ob die Bezeichnung mit a oder mit A anfängt
  - Darf nicht mit einer Zahl beginnen
  - Darf manche Sonderzeichen nicht beinhalten (z.B. *, #, !, $, @)
  - Besser keine Umlaute (ä,ö,ü)
  - Konvention: mit einem Unterstrich (my_variable), Bindestrich (my-variable), einem “Höcker” (“camel case”, myVariable) oder einem Punkt (my.variable)
  - Bestenfalls einheitlich! Z.B. häufig wird der Name „df“ für dataframe vergeben

Konventionen



Kommentare

I will remember this code without comments (And Other Hilarious Jokes You Can Tell Yourself, Volume II)

- Bitte stets den Code kommentieren!
  - Kommentare mit einen „#“ einleiten
- Chunks im Skript
  - Mit ### ---- Chunk-Name ----

Good Practice & Konventionen

- R und R-Packages zitieren
  - R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
  - Citation()-Funktion, z.B. citation(„package“)
    - Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Versionskontrolle

- Bewährter Workflow (Ironie!):
  - Analyse.r
  - Analyse_neu.r
  - Analyse_neu_1.r
  - Analyse_ganz_neu.r
  - Analyse_wirklich_ganz_neu.r
  - Analyse_wirklich_ganz_neu_final.r
  - Analyse_wirklich_ganz_neu_final_25.03.22.r

- Versionskontrolle als Lösung!
- Happy Git and GitHub for the useR – Jennifer Bryan

Hands-On:

Customer (Lösung)

RProjects

R Projects

- Organisieren den Workflow in R; z.B. Ordner, Unterordner, Skripte, Grafiken, etc.
- Erstellen einen relativen Pfad
- Ermöglichen die Reproduktion des Codes

- What They Forgot to Teach You About R – Jennifer Bryan & Jim Hester

Working Directory

- Absoluter Pfad: Working Directory ist der Ordner „R_MA_SoSe_22“
  - C:\Users\Marco\Desktop\Lehre\R_MA_SoSe_22\data
- Relativer Pfad: Working Directory ist der Ordner „R_MA_SoSe_22“ (Shortcut „.\“)
  - .\data
- Relative Pfade sind absoluten Pfaden vorzuziehen,
  - Weil Ordner verschoben werden und der Pfad nicht aktuell ist
  - Weil Projekte sonst nicht von anderen Systemen genutzt werden können
- Falsche Pfade zählen zu den häufigsten Fehlerquellen am Anfang

- RProjects organisieren

```
C:./User/Desktop/Projekt/Projekt_1
|    mein-R-Project.Rproj
|    set-up.R
|    script_1.R
|    script_2.R
+---data
|       allbus_2018_gesamt.sav
|       allbus_fb.pdf
+---export
|       data_export_1.csv
+---graphics
|       plot_1.png
|       plot_.png
```

Datenimport (Kapitel 2)

Datenimport

- Test- und Beispieldaten
  - Werden häufig von Ressourcen genutzt, um Beispiele zu zeigen
  - Überblick über data()
    - Mtcars
    - Titanic
    - Iris
- Unterschiedliche Daten-Typen können importiert werden
  - .RData
  - . sav (SPSS)
  - .dta (STATA)
  - .csv

Datenimport

- Häufigste Fehlerquelle: Arbeitsverzeichnis beachten!
  - Setwd()
  - Getwd()

- Nutzung des Rio-Packages macht das Leben einfach
  - Install.packages(„rio“) und library(rio)
  - Wahrscheinlich ist der zusätzliche Befehl install_formats() notwendig

```
library(rio)
allbus <- import("data/allbus_2018_gesamt.sav")
```

- Überblick über Datensatz

```
#Variablen
names(allbus)
#Beobachtungen
nrow(allbus)
#ersten sechs Beobachtungen (Default)
head(allbus)
```

Praxis

Kontakt

Vielen Dank!

Marco Wähner

 Gebäude 37.03.03.14

 marco.waehner@hhu.de

 Sprechstunde nach Vereinbarung