

Einführung in die Datenanalyse mit R: Datenaufbereitung

Marco Wähler



Gebäude 37.03.03.14



marco.waehner@hhu.de



Sprechstunde nach Vereinbarung



NRW-FORSCHUNGSKOLLEG
ONLINE-PARTIZIPATION

hhu Heinrich Heine
Universität
Düsseldorf



RESEARCH
FOR THE
DIGITAL AGE

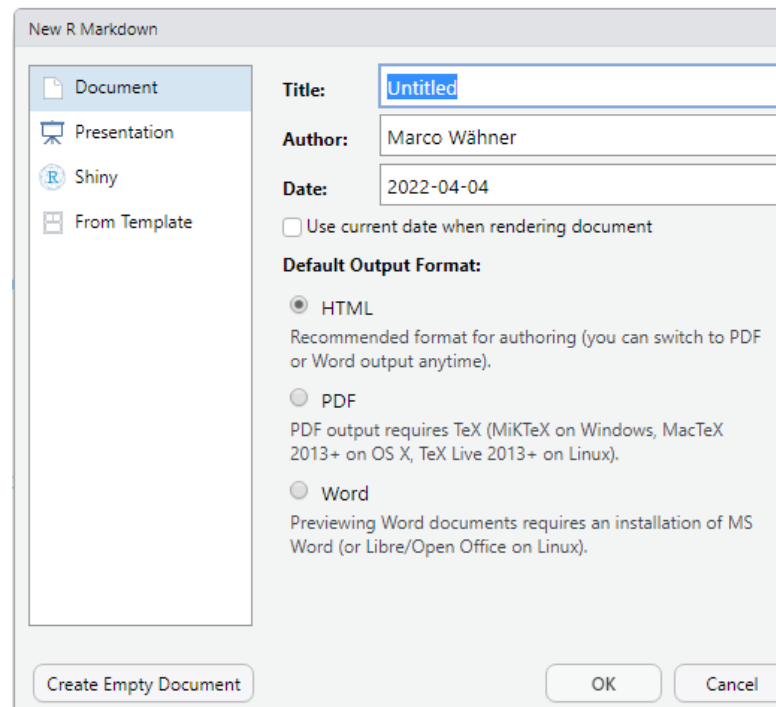
- R Markdown
 - Beispiel eines additiven Index zur politischen Online-Partizipation (Indexbildung Kapitel 11)
- Gewichtung von Survey-Daten

R Markdown


- Ermöglicht Code, Analyse und Veröffentlichung in einem Dokument
 - Tool zur reproduzierbaren Forschung
- „Rendern“ unterschiedlicher Formate
 - HTML
 - Markdown
 - PDF
 - Word
 - Open Document (u.a. Libre Office)

- Formate zur Veröffentlichung, u.a.:
 - Präsentationen ([xaringan](#))
 - Online-Books ([bookdown](#))
 - Websites ([blogdown](#))
 - Dashboards ([flexdashboard](#))
 - uvm.
- Manche Formate benötigen weitere Software
 - PDF-Dokumente benötigen z.B. eine LaTeX-Distribution
 - Am einfachsten über das [tinytex-Package](#), `install.packages(„tinytex“)` und `tinytex::install_tinytex()` zu installieren

- Über RStudio muss lediglich das R Markdown package installiert werden
- Öffnen eines R Markdown Dokuments (.Rmd) über File -> New File -> R Markdown



- Jedes File beinhaltet eine Vorlage, die direkt in ein HTML-Dokument umgewandelt werden kann



The screenshot shows an R Markdown document in a code editor. The document is divided into sections by line numbers 1 through 31. Annotations in blue text are placed next to specific parts of the code:

- YAML Header** (lines 2-5): Points to the YAML front-matter block.
- Markdown Text** (lines 12-13): Points to the section header `## R Markdown`.
- Code Chunk** (lines 18-20): Points to the first R code chunk.
- Code Chunk** (lines 26-28): Points to the second R code chunk.

```
1 ---
2 title: "first-rmd"
3 author: "Marco Wähner"
4 date: '2022-04-04'
5 output: html_document
6 ---
7
8 ```{r setup, include=FALSE}
9 knitr::opts_chunk$set(echo = TRUE)
10 ```
11
12 ## R Markdown
13
14 This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on
15 using R Markdown see <http://rmarkdown.rstudio.com>.
16
17 When you click the Knit button a document will be generated that includes both content as well as the output of any embedded R code chunks
18 within the document. You can embed an R code chunk like this:
19
20 ```{r cars}
21 summary(cars)
22 ```
23
24 ## Including Plots
25
26 You can also embed plots, for example:
27
28 ```{r pressure, echo=FALSE}
29 plot(pressure)
30 ```
31
32 Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.
```

- YAML steht für „YAML Ain't Markup Language“ (oder auch: „Yet Another Markup Language“)
- Beinhaltet Metadaten des Dokuments, die über „key values“ definiert werden
- Steht immer am Anfang eines Dokuments (---)
- Definiert bspw. das Output-Format

Source	Visual
1	---
2	title: "first-rmd"
3	author: "Marco Wähner"
4	date: '2022-04-04'
5	output: html_document
6	---
7	

Syntax

Kursiv

****Fett****

[link](https://cran.r-project.org/)

Header 1

Header 2

Header 3

* unordered list

* item 2

+ sub-item 1

+ sub-item 2

1. ordered list

2. item 2

+ sub-item 1

+ sub-item 2

Ausgabe

Kursiv

Fett

[link](#)

Header 1

Header 2

Header 3

- unordered list

- item 2

- sub-item 1

- sub-item 2

1. ordered list

2. item 2

- sub-item 1

- sub-item 2

- R Markdown am Beispiel politischer Online-Partizipation
 - pp41 - HABE AN ONLINE-PROTESTAKTION TEILGEN.
 - pp71 - HABE ONLINE-PETITION GESTARTET
 - pp72 - HABE IN SOZ. MEDIEN POL. MEINUNG GESAGT
- Additiver Index zur Online-Partizipation in Code-Chunks schreiben

Praxis

Gewichtungen

Zur Gewichtung von ALLBUS-Umfragedaten (Codebook, ab S. iii, Achtung Beispiel im Codebook von 2016)

- Zwei Merkmale des Stichprobendesigns
 1. „Befragte in Ostdeutschland werden seit 1991 zu einem größeren Anteil in die Stichprobe einbezogen als es ihrem Anteil an der Grundgesamtheit entspräche (Oversampling). Dieses Oversampling soll auch für kleinere Bevölkerungsgruppen in Ostdeutschland noch statistisch vertretbare Analysen ermöglichen.“
 2. „Die Stichproben der Umfragen in den Jahren 1980 bis 1992 sowie 1998 basierten auf Haushaltsstichproben nach dem ADM-Stichprobendesign (mit den Auswahlstufen Wahlbezirke - Haushalte - Personen, siehe vertiefend Schnell et al. 2008; von der Heyde 2009), 1994 und 1996 sowie in allen Erhebungen seit 2000 wurden dagegen Personenstichproben aus den Einwohnermelderegistern gezogen (mit den Auswahlstufen Gemeinden Personen).“

-> Beide Umstände müssen – abhängig von der FF – bei der Analyse berücksichtigt werden (wir konzentrieren uns auf die Analyse der Personenebene)

Personenstichprobe (Codebook, S. v)

- „Im Umfrageprogramm des ALLBUS werden seit der ersten Befragung Ostdeutscher im Jahr 1991 mehr Personen in den neuen Bundesländern befragt als es ihrem Anteil an der gesamtdeutschen Bevölkerung entspricht. Dieses Oversampling intendiert, auch für Ostdeutschland eine Fallzahl zu erzielen, die differenzierte Analysen für einzelne Bevölkerungsgruppen erlaubt. [...] Wenn aber beide Bereiche (Ost- und Westdeutschland) gemeinsam als Gesamtdeutschland analysiert werden sollen, muss die Überrepräsentation von ostdeutschen Befragten im ALLBUS durch eine Gewichtung aufgehoben werden.“

- Grundlage der Gewichtung ist die Verteilung in der Zielpopulation

Tabelle 2: Datengrundlage für die Ost-West-Gewichtung auf Personenebene: Mikrozensus 2015 und ALLBUS 2016

	Mikrozensus 2015 (in tausend)			ALLBUS 2016		
	West N _w	Ost N _o	Gesamt N	West n _w	Ost n _o	Gesamt n
Personen in Privathaushalten (Alter: 18 Jahre oder mehr)	55.586 82,2%	12.040 17,8%	67.626 100%	2.325 66,6%	1.165 33,4%	3.490 100%

- „Um ihrem Anteil in der gesamtdeutschen Grundgesamtheit zu entsprechen, muss den Angaben von Befragten aus Ostdeutschland bei gesamtdeutschen Analysen ein geringeres Gewicht beigemessen werden als den Befragten aus Westdeutschland. Den Angaben von Befragten aus Westdeutschland muss ein höheres Gewicht beigemessen werden.“ (Codebook, S. v)

Personenbezogenes Ost-West-Gewicht (Variable: wghtpew)

- Gewichtungswert für ostdeutsche Befragte: 0,54 (multipliziert mit 1090 Befragte)
- Gewichtungswert für westdeutsche Befragte: 1,20 (multipliziert mit 2387 Befragte)

Tabelle 3: Verteilung der Befragten auf Ost- und Westdeutschland: Vergleich des Mikrozensus 2015 mit gewichteten Daten des ALLBUS 2016

	Mikrozensus 2015 (in tausend)			ALLBUS 2016		
	West N _w	Ost N _o	Gesamt N	West n _w	Ost n _o	Gesamt n
Personen in	55.586	12.040	67.626	2.869	621	3490
Privathaushalten (Alter: 18 Jahre oder mehr)	82,2%	17,8%	100%	82,2%	17,8%	100%

- Hinweis: In der Praxis werden Gewichte zuweilen verwendet, um systematische Verzerrungen in der Stichprobe zu kaschieren (Diekmann, 2010, S. 428) z.B. Civey-Umfragen
 - Gewichtungsfaktoren, insb. In der Marktforschung sind ein gut gehütetes (Geschäfts-)Geheimnis.

- Für die Deskriptivstatistik: „Survey“-Package:
 - *install.packages("survey")*
 - *library(survey)*
- Erstellung eines Umfrage-Designs mit Gewichtung:
`df.w <- svydesign(ids =~ 1, data = df, weights =~ wghtpew)`

Praxis

Hands-On:

R Markdown (Lösung)

Gewichtung (Lösung)

Vielen Dank!

Marco Wähner



Gebäude 37.03.03.14



marco.waehner@hhu.de



Sprechstunde nach Vereinbarung