Tutorium Datenauswertung

Beschreibung kategorialer Daten

Agenda

- Häufigkeitstabellen
  - Basics
  - per Hand berechnen
  - Erstellen mit R
- Diagramme
  - Basics
  - per Hand zeichnen
  - Täuschungen mit Diagrammen
  - Erstellen mit R
- Häufigkeitsübersichten

Häufigkeitsverteilungen

Häufigkeitsverteilungen

- beschreiben, wie sich die erhobenen Daten auf die Ausprägungen einer Variablen verteilen → Überblick
- Können in Tabellen oder Diagrammen dargestellt werden
- Werden für *kategoriale Daten* (in R: Faktoren!) benutzt

Häufigkeitstabellen – Basics

Häufigkeitstabellen

- Bestehen aus vier Spalten:

Spalte	Beantwortung der Frage…
Ausprägungen	Welche Ausprägungen kann die Variable annehmen?
Anzahl	Wie oft wurde die Ausprägung realisiert (angekreuzt)?
Prozent	Wieviel Prozent der Befragten haben die Ausprägung realisiert?
Kumulierte Prozent	Wieviel Prozent der Befragten haben diese Ausprägung oder eine von den Ausprägungen in den Zeilen davor realisiert?

- Können fehlende Angaben (NAs) angeben oder nicht

Häufigkeitstabellen per Hand: Beispiel

Fall	1	2	3	4	5	6	7	8	9	10	11
Altersgruppe	15-24 Jahre	25-34 Jahre	25-34 Jahre	15-24 Jahre	15-24 Jahre	NA	35-44 Jahre	15-24 Jahre	15-24 Jahre	25-34 Jahre	15-24 Jahre

Berechnung der Prozentspalte: $\frac{\text{Anzahl der Realisationen der Ausprägung}}{\text{Gesamtanzahl der Realisationen}} * 100$

Ausprägung	Anzahl	Prozent	Kumulierte Prozent
15-24 Jahre	6	6/(6+3+1)*100 = 60	60
25-34 Jahre	3	30	60+30 = 90
35-44 Jahre	1	10	100

Häufigkeitstabellen mit R

```
#Laden von Daten und Paketen
load("daten_x.RData")

#Erstellung der Tabelle
Anzahl <- table(datensatz$variable)
Prozent <- prop.table(Anzahl) * 100
kum.Prozent <- cumsum(Prozent)
round(cbind(Anzahl, Prozent, kum.Prozent), 1)
```

Häufigkeitstabellen mit R – Beispiel

```
#Laden von Daten und Paketen
load("Beispieldatensatz.RData")

#Erstellung der Tabelle
Anzahl <- table(beispiel$Altersgruppe)
Prozent <- prop.table(Anzahl) * 100
kum.Prozent <- cumsum(Prozent)
round(cbind(Anzahl, Prozent, kum.Prozent), 1)
```

Häufigkeitstabellen – Aufgabe

Aufgabe: Erstellt eine Häufigkeitstabelle mit den Spalten „Ausprägung“, „Anzahl“, „Prozent“ und „kumulierte Prozent“ für die unten stehende Variable Haushaltsgröße aus dem Beispieldatensatz. Benutzt R als Taschenrechner und rundet auf eine Nachkommastelle. Überprüft anschließend euer Ergebnis, indem ihr mithilfe des Markdown-Scripts eine Häufigkeitstabelle erstellt (ersetzt die Subset-Zeile durch „datensatz <- beispiel“).

Fall	1	2	3	4	5	6	7	8	9	10	11
Haushaltsgröße	1	4	2	1	1	3	3	2	1	1	4

Häufigkeitstabellen – Lösung I

Fall	1	2	3	4	5	6	7	8	9	10	11
Haushaltsgröße	1	4	2	1	1	3	3	2	1	1	4

Ausprägung	Anzahl	Prozent	Kumulierte Prozent
1 Mitbewohner	5	5/(5+2+2+2)*100 = 45,5	45,5
2 Mitbewohner	2	18,2	63,6
3 Mitbewohner	2	18,2	81,8
4 Mitbewohner	2	18,2	100

Den Wert, der am häufigsten in einer Tabelle vorkommt, nennen wir „Modalwert“.

Häufigkeitstabellen – Lösung II

```
#Laden von Daten und Paketen
library(knitr)
load("Beispieldatensatz.RData")
datensatz <- beispiel

#Beschriftung ändern (optional)
datensatz$Haushaltsgröße <- factor(datensatz$Haushaltsgröße, levels=c("1", "2","3","4"), labels=c("1 Mitbewohner", "2 Mitbewohner","3 Mitbewohner","4 Mitbewohner"))

#Erstellung der Tabelle
Anzahl <- table(datensatz$Haushaltsgröße)
Prozent <- prop.table(Anzahl) * 100
kum.Prozent <- cumsum(Prozent)
kable(round(cbind(Anzahl, Prozent, kum.Prozent), 1))
```

Häufigkeitstabellen – Lösung III

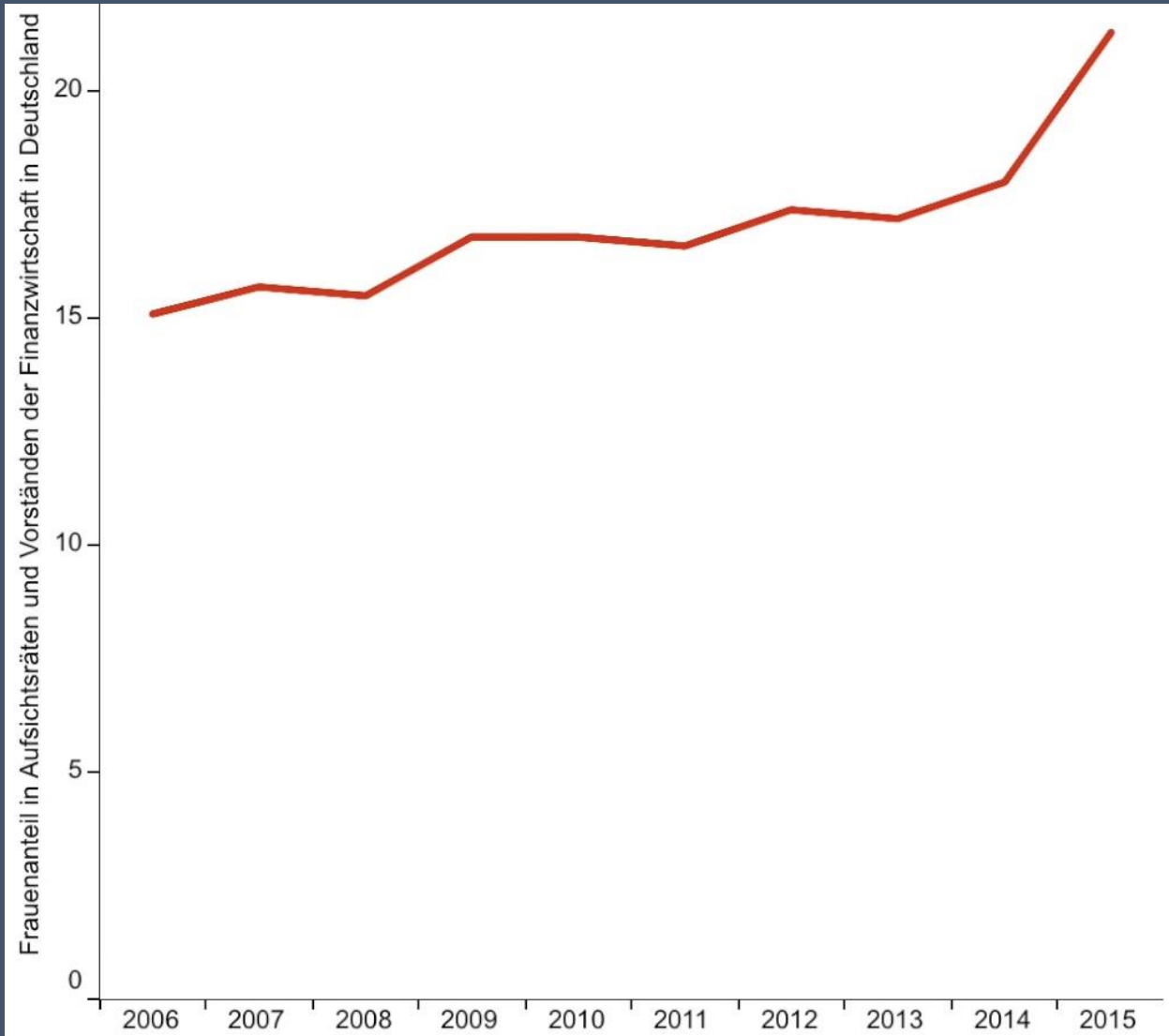
	Anzahl	Prozent	kum.Prozent
1 Mitbewohner	5	45.5	45.5
2 Mitbewohner	2	18.2	63.6
3 Mitbewohner	2	18.2	81.8
4 Mitbewohner	2	18.2	100.0

Kategoriale Daten - Diagramme

Diagramme

- Bieten einen graphischen Überblick über die Verteilung
- Sind besonders sinnvoll zum Vergleich von Ausprägungen
- Klassische Diagrammtypen: Säulendiagramm, Kreisdiagramm
- Per Hand berechnen:
  - x-Achse: Ausprägungen (Platz lassen!)
  - y-Achse: Häufigkeit (sinnvolle Skalierung!)

!!! Traut keiner Grafik, die von einer Person erstellt wurde, die in einem Interessenskonflikt mit den Ergebnissen stehen könnte !!!

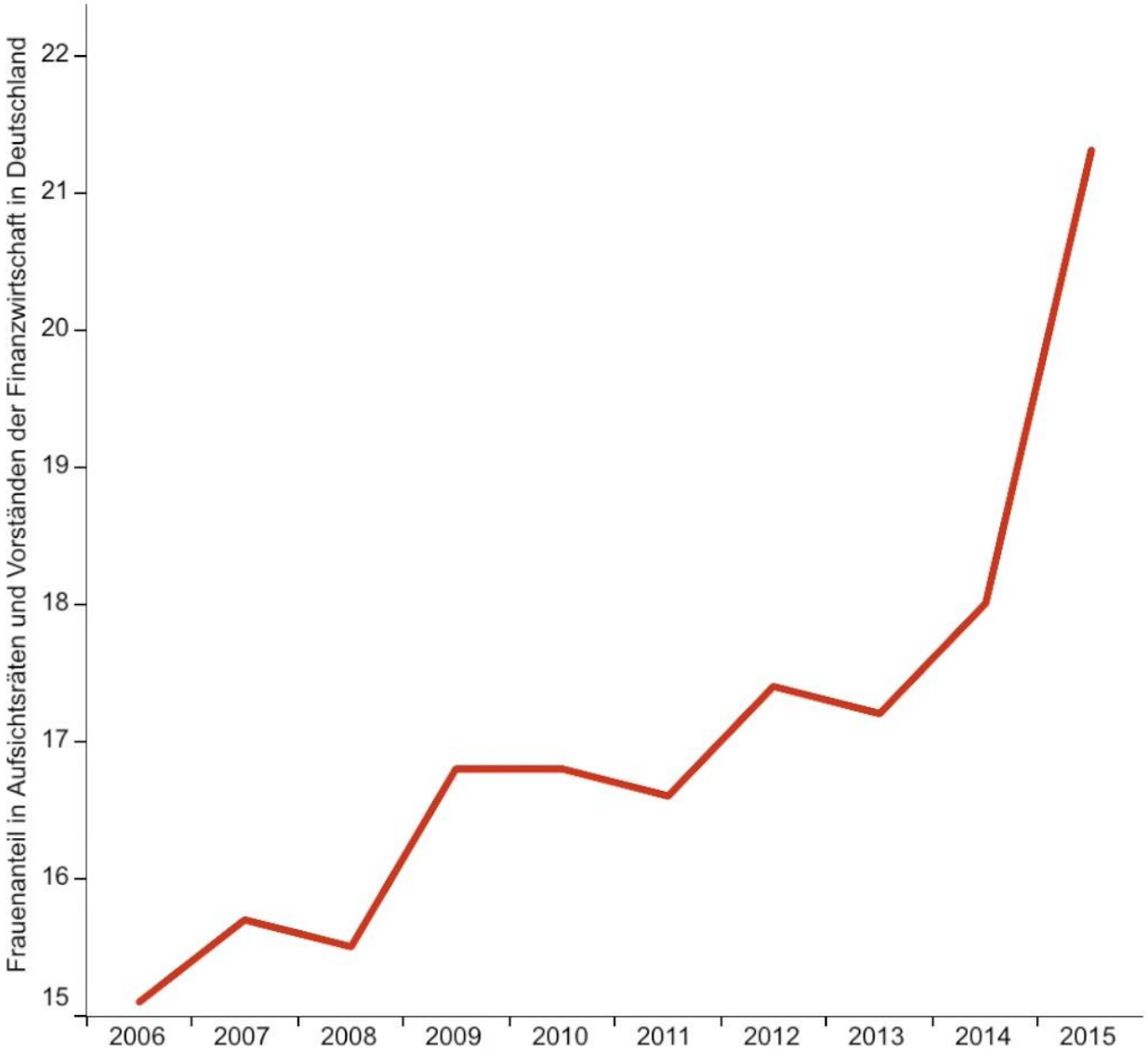


Quellen: DIW, Statistisches Bundesamt, Bundesagentur für Arbeit

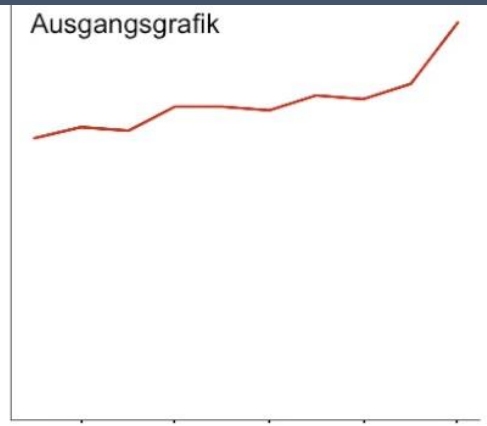
Ausgangsgrafik: Frauenanteil in Aufsichtsräten und Vorständen der Finanzwirtschaft in Deutschland

Naheliegende Lesart: Frauenanteil wächst, aber recht gemächlich.

Wirtschafts Woche



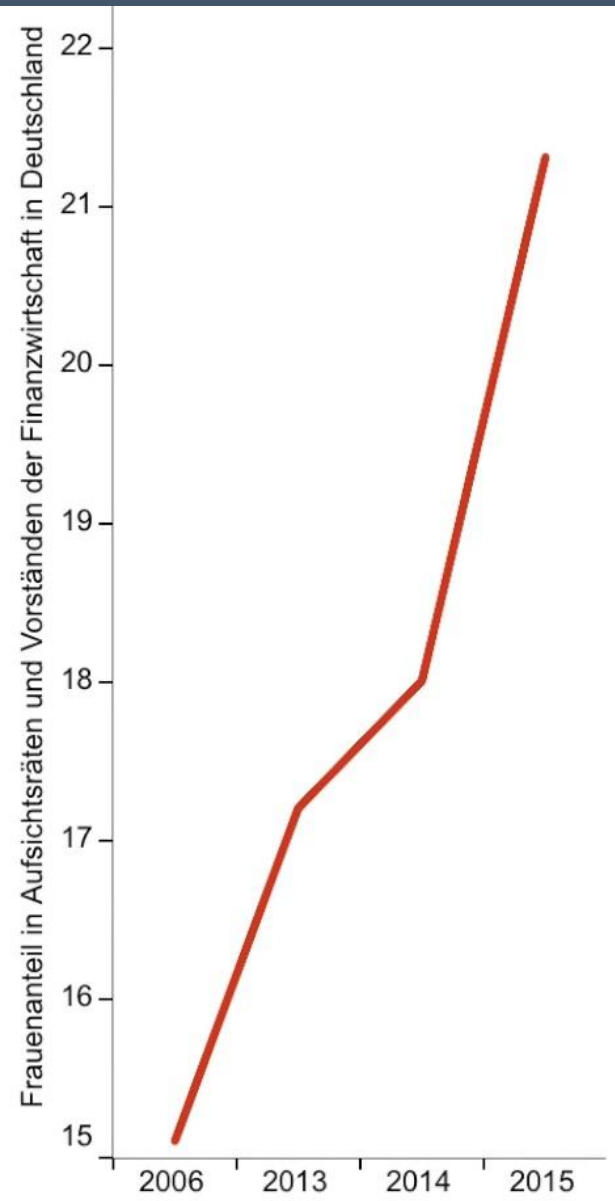
Quellen: DIW, Statistisches Bundesamt, Bundesagentur für Arbeit



Effekt 1: Als Nulllinie wird 15 Prozent gesetzt.

Naheliegende Lesart: Der Frauenanteil steigt dynamisch.

Wirtschafts Woche



Quellen: DIW, Statistisches Bundesamt, Bundesagentur für Arbeit

Effekt 2: Die Werte nur die Jahre 2006, 2013, 2014, 2015 gezeigt

Naheliegende Lesart: Der Frauenteil geht ab wie eine Rakete.

Wirtschafts Woche

Installation: ggplot2

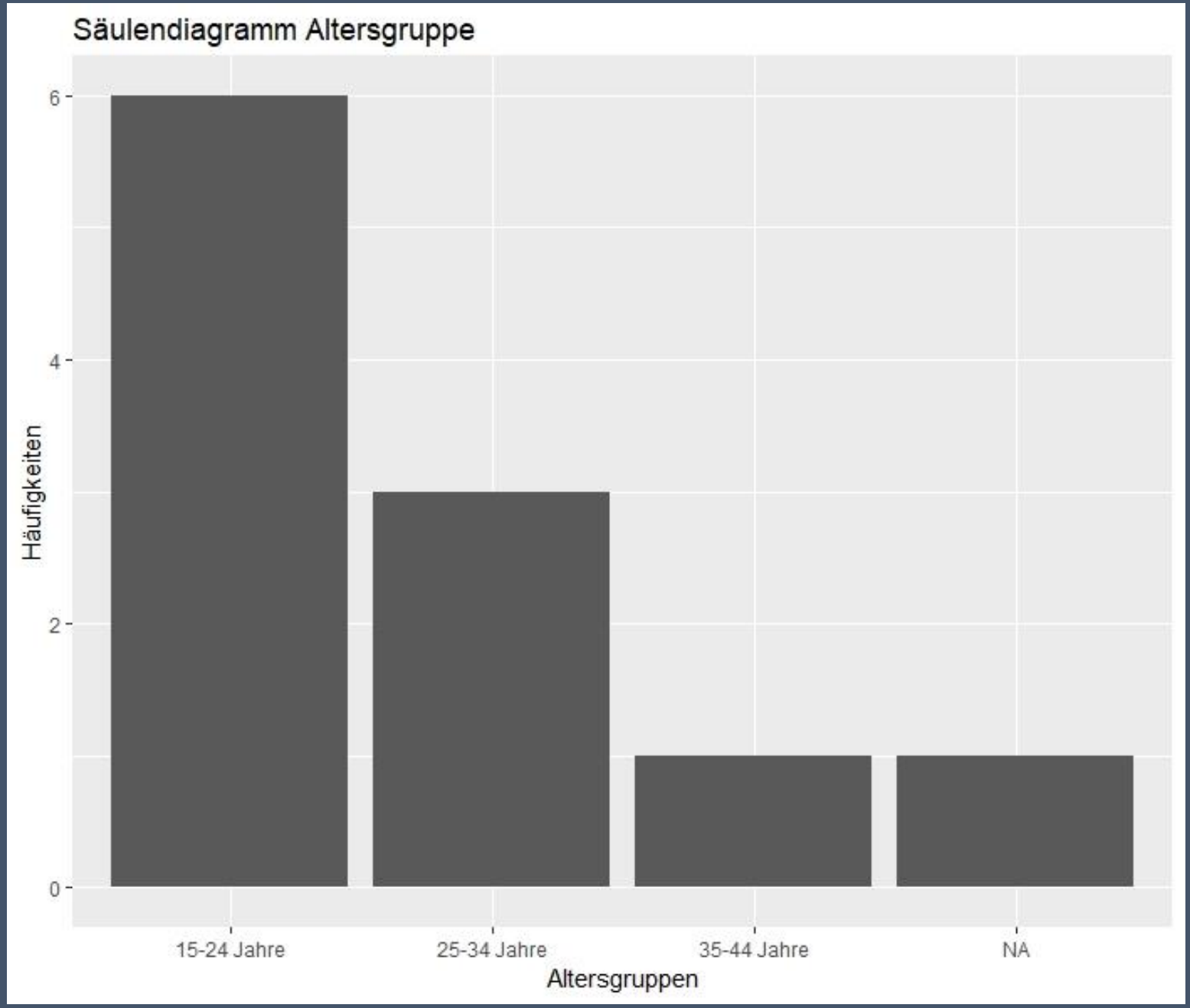
```
> install.packages("ggplot2")
Installing package into 'C:/Users/vitus/Documents/R/win-library/3.5'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/ggplot2_3.1.1.zip'
Content type 'application/zip' length 3623171 bytes (3.5 MB)
downloaded 3.5 MB

package 'ggplot2' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\vitus\AppData\Local\Temp\Rtmpgh3m3A\downloaded_packages
> library(ggplot2)
Warning message:
Paket 'ggplot2' wurde unter R Version 3.5.3 erstellt
>
```

Säulendiagramme mit R

```
#Laden von Daten und Pakten
library(ggplot2)
load("daten_x.RData")

#Erstellen eines Säulendiagramms
ggplot(datensatz, aes(x = factor(variable))) +
  geom_bar() + labs(title = "titel", x = "beschriftung x-Achse", y = "beschriftung y-Achse")
```

Säulendiagramme mit R – Beispiel

```
#Laden von Daten und Pakten
library(ggplot2)
load("Beispieldatensatz.RData")

#Erstellen eines Säulendiagramms
ggplot(beispiel, aes(x = factor(Altersgruppe))) +
  geom_bar() + labs(title = "Säulendiagramm
Altersgruppe", x = "Altersgruppen", y = "Häufigkeiten")
```

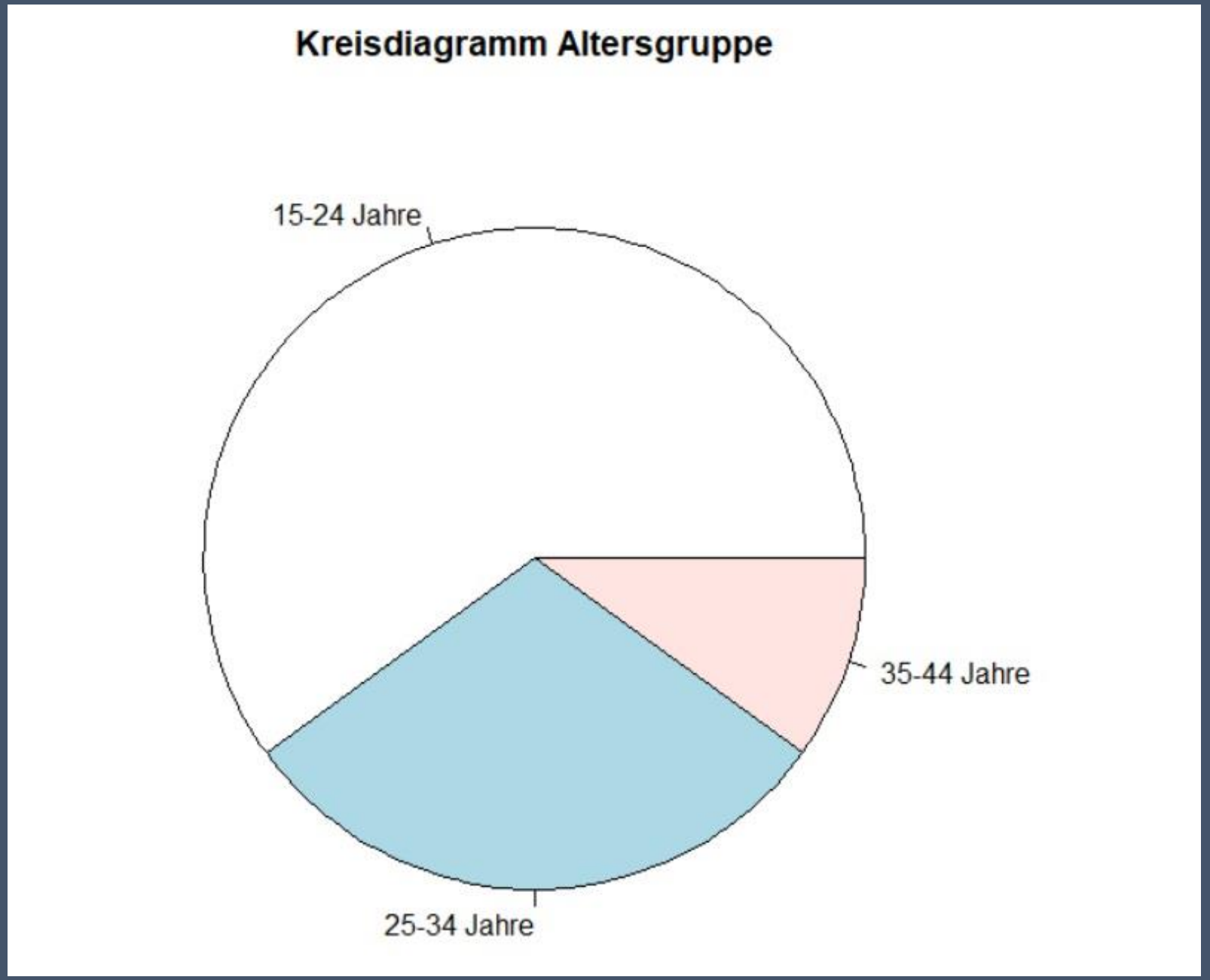

Säulendiagramm Altersgruppe
Häufigkeiten
6
4
2
0
15-24 Jahre
25-34 Jahre
35-44 Jahre
NA
Altersgruppen

Kreisdiagramme in R

```
#Laden von Daten und Paketen
load("daten_x.RData")

#Erstellen des Kreisdiagramms
pie(table(datensatz$variable), main = "überschrift")
```

Kreisdiagramme in R – Beispiel

```
#Laden von Daten und Paketen
load("Beispieldatensatz.RData")

#Erstellen des Kreisdiagramms
pie(table(beispiel$Altersgruppe), main = "Kreisdiagramm Altersgruppe")
```


Kreisdiagramm Altersgruppe
15-24 Jahre
35-44 Jahre
25-34 Jahre

Aufgabe – Diagramme

Aufgabe: Ladet den aktuellen Datensatz zu eurer Befragung (daten_2019) aus dem Learnweb herunter. Öffnet das MD Häufigkeiten und lest es euch aufmerksam durch. Erstellt für die Variable medium_GuM eine Häufigkeitstabelle, ein Säulendiagramm und ein Kreisdiagramm.

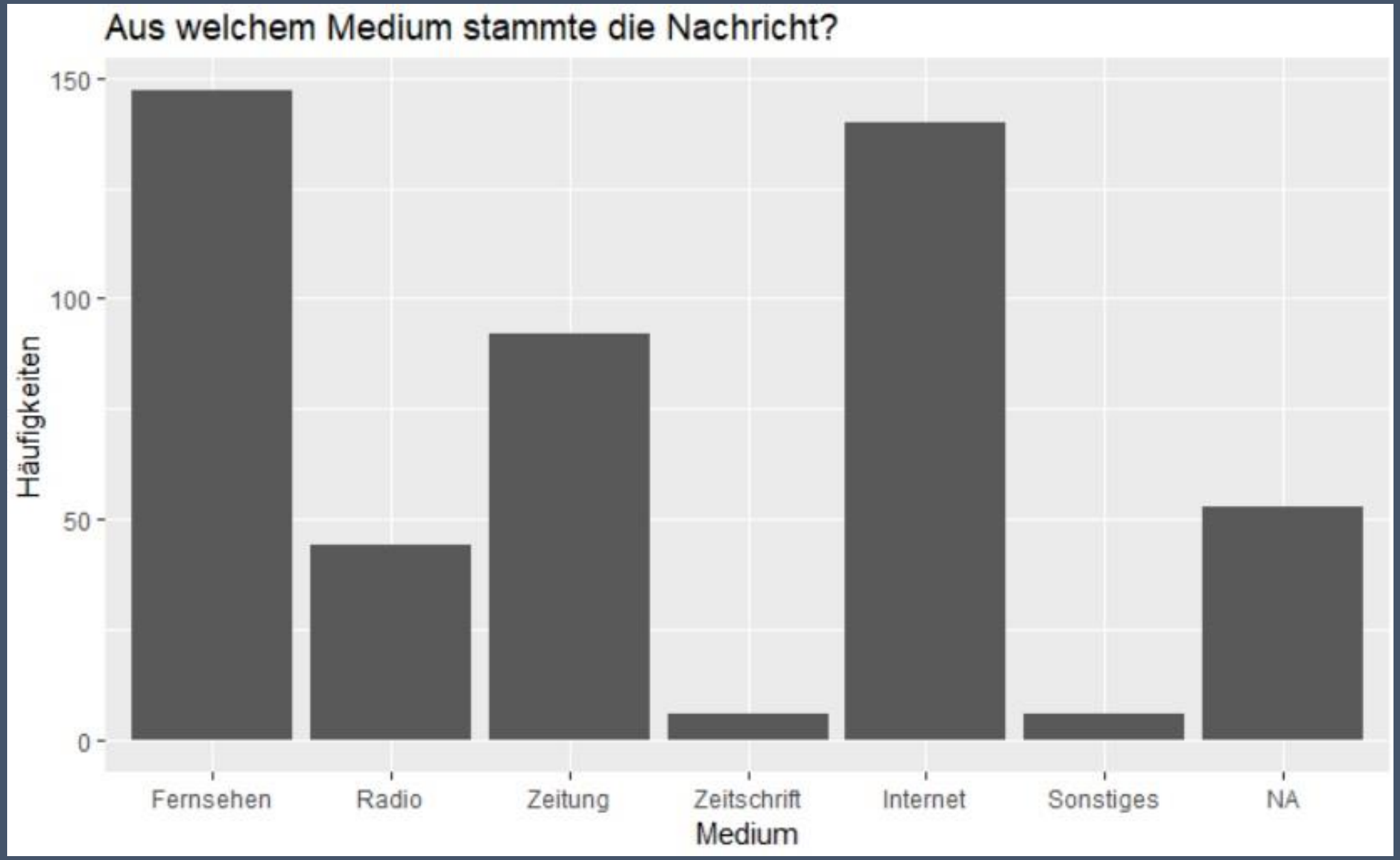
Lösung: Häufigkeitstabelle

```
#Laden von Daten und Paketen
library(knitr)
load("daten_2019.RData")
datensatz <- subset(daten_2019, jahr > 2000)

#Erstellung der Tabelle
Anzahl <- table(datensatz$medium_GuM)
Prozent <- prop.table(Anzahl) * 100
kum.Prozent <- cumsum(Prozent)
kable(round(cbind(Anzahl, Prozent, kum.Prozent), 1))
```

	Anzahl	Prozent	kum.Prozent
Fernsehen	147	33.8	33.8
Radio	44	10.1	43.9
Zeitung	92	21.1	65.1
Zeitschrift	6	1.4	66.4
Internet	140	32.2	98.6
Sonstiges	6	1.4	100.0

Lösung: Säulendiagramm

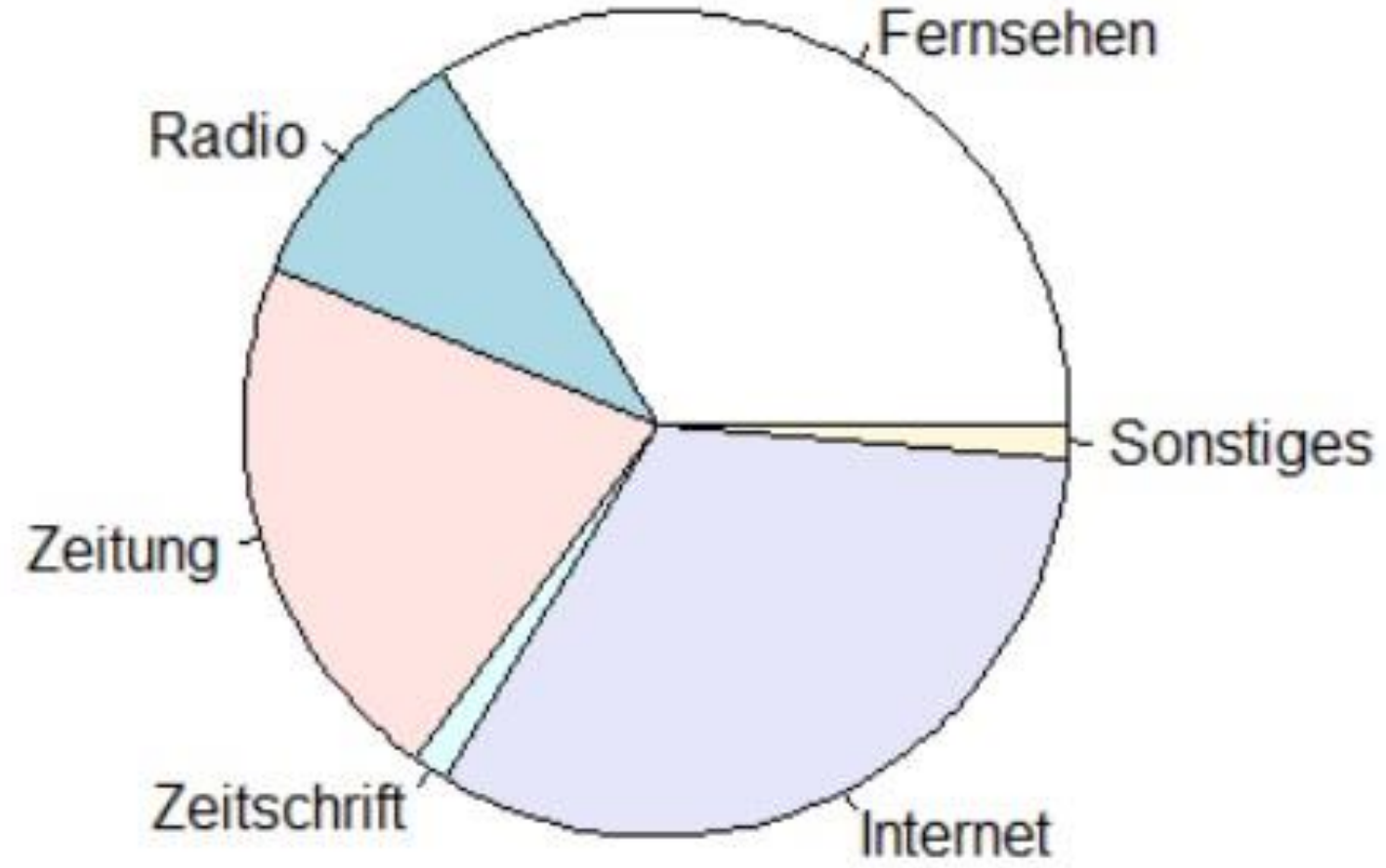
```
#Laden von Daten und Paketen
library(ggplot2)
load("daten_2019.RData")
datensatz <- subset(daten_2019, jahr > 2000)

#Erstellen eines Säulendiagramms
ggplot(datensatz, aes(x = factor(medium_GuM)))
+geom_bar() + labs(title = "Aus welchem Medium stammte
die Nachricht?", x = "Medium", y = "Häufigkeiten")
```


Aus welchem Medium stammte die Nachricht?
150
100
50
0
Häufigkeiten
Fernsehen
Radio
Zeitung
Zeitschrift
Internet
Sonstiges
NA
Medium

Lösung: Kreisdiagramm

```
#Laden von Daten und  Paketen
load("daten_2019.RData")
datensatz <- subset(daten_2019, jahr > 2000)

#Erstellen des Kreisdiagramms
pie(table(datensatz$medium_GuM), main = "Aus welchem Medium stammte die Nachricht?")
```

Aus welchem Medium stammte die Nachricht?



Häufigkeitsübersichten

Häufigkeitsübersichten

- Sind Tabellen, die die Häufigkeiten *mehrerer Variablen* anzeigen
- Machen nur Sinn, wenn die Variablen *die gleichen Ausprägungen* haben
- Beispiele:
  - GuMvergl_a – GuMvergl_f (Ausprägungen: mehr/weniger/gleich)
  - gem-Variablen (Ausprägungen: ja/nein)

Häufigkeitsübersichten in R

```
#Laden von Daten und Paketen
library(knitr)
load("daten_x.RData")
datensatz <- subset(datensatz, Bedingung)

#Umcodierung in Prozentwerte
var1neu <- c(100*prop.table(table(datensatz$variable_1)))
                                                    ...
var3neu <- c(100*prop.table(table(datensatz$variable_n)))

#Erstellung der Übersicht
round(cbind(datensatz$variable_n, ..., variable_nneu), 1)
```

Häufigkeitsübersichten in R – Beispiel

```
#Laden von Daten und Paketen
library(knitr)
load("daten_2019.RData")
datensatz <- subset(daten_2019, jahr > 2000)
#Level umbenennen (optional)
datensatz$GuMvergl_a <- factor(datensatz$GuMvergl_a, labels=c("weniger", "gleich", "mehr"))
[…]
#Umcodierung in Prozentzahlen
Intensivität <- c(100*prop.table(table(datensatz$GuMvergl_a)))
[…]
#Erstellung der Übersicht
round(cbind(Intensivität, Interesse, Emotionalität, Lebhaftigkeit, Anstrengung, Unterhaltung), 1)
```

	Intensivität	Interesse	Emotionalität	Lebhaftigkeit	Anstrengung	Unterhaltung
weniger	19.3	15.8	30.2	22.5	35.9	24.8
gleich	51.1	47.7	39.1	48.2	52.3	48.6
mehr	29.5	36.4	30.7	29.3	11.8	26.6

Selbststudium

- Wiederholung: Darstellung kategorialer Variablen
  - Häufigkeitstabellen (mit R, per Hand)
  - Diagramme (mit R, Säulendiagramme per Hand)
  - Häufigkeitsübersichten (mit R)
- Forschungsbericht (EBF und ZFB):
  - Häufigkeitstabelle, Säulen- und Kreisdiagramm für die Variable daten_2019$geschlecht
  - Häufigkeitsübersicht für die Variablen daten2019$gem_wwwp bis daten_2019$gem_wwwb
  - Interpretation der Ergebnisse (für beide Variablen)

Bis nächste Woche!