Statistik zur Datenanalyse

Dr. Meike Wocken

Vorlesung 1

HS Bielefeld
Digitale Technologien (M.Sc.)
WiSe 2023/24

Meike.Wocken@codecentric.de

R4DS

https://cran.r-project.org

- Downloads
- Pakete
- Task Views

The R Project: https://www.r-project.org

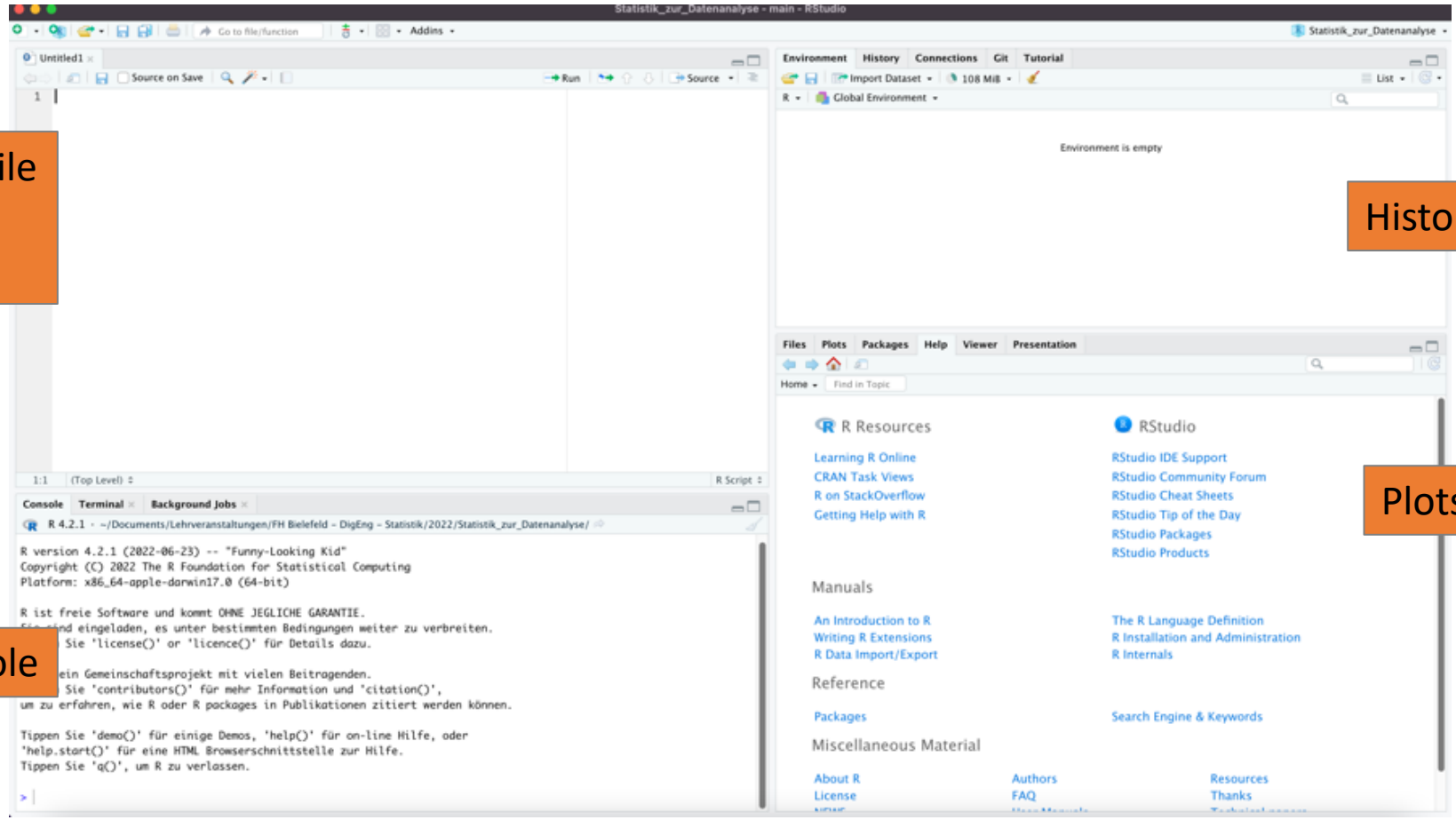
R for Data Science: ~~https://r4ds.had.co.nz~~- https://r4ds.hadley.nz

R Studio starten

File → New File

Skript

Konsole



Historie, Environment

Plots, Hilfe, Packages

Workspace

- Aktuelle R Session
- Hier befinden sich geladene Daten + Objekte
- Speichern ja/nein beim Schließen von R (bzw. R Studio)
- Pakete sind zu laden in R Session (auch wenn ich mit alter Session weiterarbeite): `library()`

R Studio: New File → New R Skript

- Object assignment

  `=` oder `<-`

  `==` ist Vergleich
- `?` Hilfsseiten immer nutzen
- CASE-Sensitive Namen
- Überschreiben erzeugt keine Fehler
- Speichern: Es ist ein .txt File, die Endung .R erleichtert das direkte öffnen mit R Studio

R Studio: New File → New R Markdown

- Wir erhalten ein html Dokument
- Output-Options:

    siehe Help → Cheat Sheets → R Markdown Cheat Sheet

Notiz: evtl. müssen beim ersten Markdown-Dokument Pakete noch installiert werden

Tidyverse

- **Datenprojekt** in Rstudio anlegen, Code Versionierung möglich
- Konzept Tidy Data:

There are three interrelated rules which make a dataset tidy:

1.Each variable must have its own column.

2.Each observation must have its own row.

3.Each value must have its own cell.

[*https://r4ds.hadley.nz/data-tidy*](https://r4ds.hadley.nz/data-tidy)

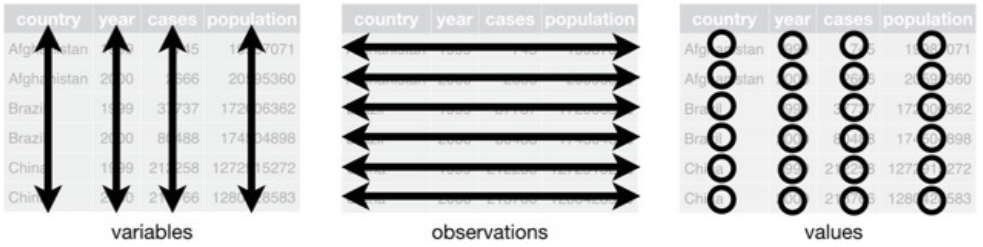
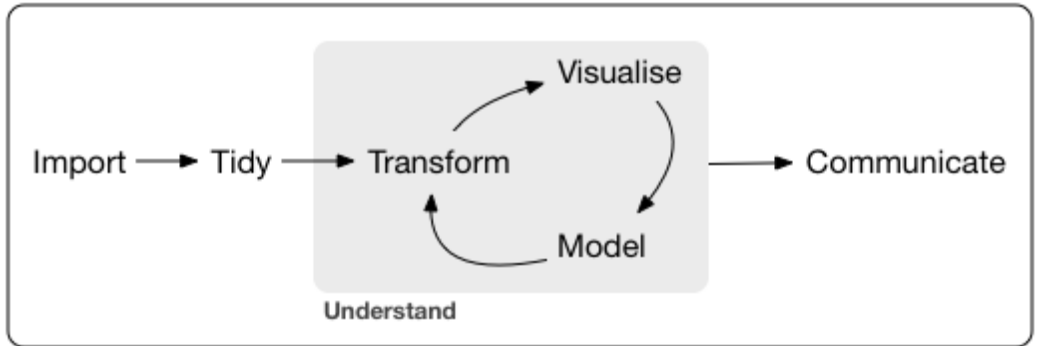


Figure 12.1: Following three rules makes a dataset tidy: variables are in columns, observations are in rows, and values are in cells.

Tidyverse



https://r4ds.hadley.nz/intro

Sammlung an Paketen, die das Tidyverse bilden:

- `readr` zum Dateneinlesen
- `tidyr` zum tidy machen
- `dplyr, purr, stringr` zum Daten aufbereiten
- `ggplot2` zum visualisieren

The Layered Grammar of Graphics

```
library(ggplot2)
ggplot(data = <DATA>) +
     <GEOM_FUNCTION>(
          mapping = aes(<MAPPINGS>), #size, color, x, y, transparency (alpha), shape
          stat = <STAT>,
          position = <POSITION>,
          )+
     <COORDINATE FUNCTION> +        # x- und y-Achse tauschen
     <FACET FUNCTION>               # unterteilen der Plot-Ausgabe in mehrere Plots (grid)
```

fahrrad_preise.csv

- Erstellen Sie einen Boxplot, der Unterschiede im Preis für unterschiedliche Farben darstellt.

Open Data - http://casas.wsu.edu/datasets/

Drei Bedingungen müssen Daten erfüllen, damit wir von Open Data sprechen:

- **Verfügbarkeit und Zugriffsmöglichkeit:**

Die Daten stehen frei zur Verfügung, bevorzugt als Download aus dem Internet. Es darf maximal eine gut zu begründende Bearbeitungsgebühr erhoben werden. Die Daten müssen in einem zweckmäßigen und bearbeitbaren Format vorliegen.

- **Verarbeitung und Weitergabe der Daten:**

Open Data werden unter Bedingungen veröffentlicht, die eine Verarbeitung, Weitergabe und Zusammenführung mit anderen Datensätzen erlauben. Die Daten müssen maschinenlesbar sein.

- **Universelle Partizipation:**

Open Data stützt sich auf die Teilnahme aller: jede und jeder muss die Daten Nutzen, Verarbeiten und Weitergeben dürfen. Es darf keine Diskriminierung geben gegenüber relevanten Anwendungsbereichen, Personen oder Gruppen. Es darf keine Einschränkungen der Verwendung der Daten geben (z.B. eingeschränkt auf nur nicht-kommerzielle Anwendungen).

Blogbeitrag: https://foundersfoundation.de/content-library/potenziale-open-data/

Dataset 33 http://casas.wsu.edu/datasets/hh111.zip

- Datenset runterladen, Readme lesen
- Einlesen der annotierten Daten hh111.ann.txt
- Welche Sensor-Typen liegen vor?
- Wie viele Beobachtungen haben Sie von welchen Einzel-Sensoren?
- Gibt es Fehlwerte in dem Datensatz?
- Grafische Betrachtung der Temperatursensoren. Was fällt auf?

Wahrscheinlichkeitsbegriff

Lostopf

E := Gewinn

$P(E)$ = Anzahl günstige Ergebnisse / Anzahl mögliche Ergebnisse

$= g/m$

$= 10/(10 + 20) = 1/3$

Ereignisse

- Losziehen ist **Zufallsvorgang**, der zu einem von mehreren, **sich gegenseitig ausschließenden** Ergebnissen (Niete, Gewinn) führt.
- Mögliche Ergebnisse sind **Elementarereignisse** $\omega \in \Omega$ (**Ergebnismenge**)
- Elementarereignisse sind **nicht weiter zerlegbar**, können aber zu Ereignissen zusammen geführt werden (A, B, C, …)
- Extremfälle: $g = 0\ P(E = Gewinn) = 0$

  $g = m\ P(E = Gewinn) = 1$, sicheres Ereignis

Begrenzung Wahrscheinlichkeit $0 \leq p \leq 1$

Gegenereignisse

$$p(-E) = \frac{m-g}{m} = 1 - \frac{g}{m} = 1 - p(E)$$

- Summe der Wahrscheinlichkeiten aller Elementarereignisse ist 1
- Wahrscheinlichkeit ist 1, dass mind. Ein Ereignis eintritt
- Wenn alle Ereignisse die gleiche Wahrscheinlichkeit haben, nennt man es ein **Laplace-Experiment**

Bedingte Wahrscheinlichkeit

- $E1$ = erstes gezogenes Los gewinnt
- $E2$ = zweites gezogenes Los gewinnt

Die Wahrscheinlichkeiten verändern sich, nachdem ein erstes Los gezogen wurde.

$$P(E2|E1) = \frac{\#(E1 \cdot E2) günstige\ Fälle}{\#\ E1\ günstige\ Fälle} = \frac{P(E1 \cdot E2)}{P(E1)}$$

UND Bedingung

- $E1$ = erstes gezogenes Los gewinnt
- $E2$ = zweites gezogenes Los gewinnt

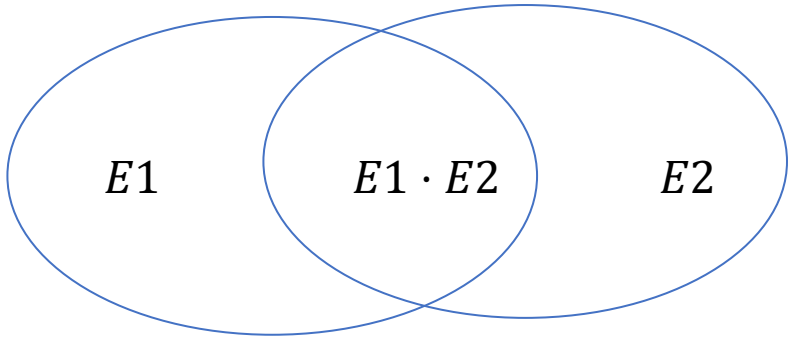
$$P(E1 \cdot E2) = P(E1)P(E2|E1) = \frac{1}{3} \cdot \frac{9}{29}$$

ODER Bedingung

- $E1$ = erstes gezogenes Los gewinnt
- $E2$ = zweites gezogenes Los gewinnt

$$P(E1 + E2) = P(E1)P(E2|E1) + P(-E1)P(E2|-E1) + P(E1)P(-E2|E1)$$
$$= p(E1)[p(E2|E1) + p(-E2|E1)] + p(-E1)\frac{p(E2 \cdot -E1)}{p(-E1)}$$
$$= p(E1) + p(-E1 \cdot E2) = p(E1) + p(E2)p(-E1|E2)$$
$$= p(E1) + p(E2) - p(E1 \cdot E2)$$

$$P(E1 + E2) = \frac{1}{3} + \frac{1}{3} - \frac{1}{3} \cdot \frac{9}{29}$$



ODER Bedingung: Gegenereignis

- $E1$ = erstes gezogenes Los gewinnt
- $E2$ = zweites gezogenes Los gewinnt

$$P(E1 + E2) = 1 - P(-E1)P(-E2| - E1)$$

$$P(E1 + E2) = 1 - \frac{2}{3} \cdot \frac{19}{29}$$

Unabhängigkeit von Ereignisses

Bedingung hat keinen Einfluss, es gilt:

$$P(E1) = P(E1|E2)$$
$$P(E2) = P(E2|E1)$$
$$P(E1 \cdot E2) = P(E1)P(E2)$$

Beispiel Skat: 32 Karten, 4 Farben

$$P(As) = \frac{4}{32} = \frac{1}{8}$$

Hilft mir die Info, wenn ich weiß, dass ich eine Pik-Karte gezogen habe?

$$P(As|Pik) = \frac{1}{8}$$

Nein, denn die Ereignisse sind unabhängig.

Totale Wahrscheinlichkeit

Annahmen:

- $E1, \dots, En$ können nicht zusammen auftreten

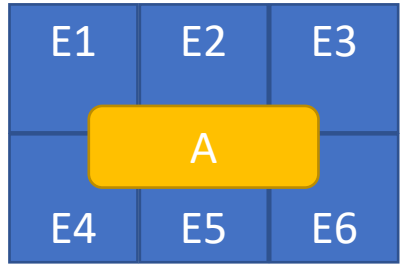
$$P\left(E_i \cdot \mathrm{E_j}\right) = 0 \; \forall \; i, j \in 1, \dots, n.$$

- Irgendein Ereignis muss auftreten

$$\sum P(E_i) = 1$$

Es gilt dann:

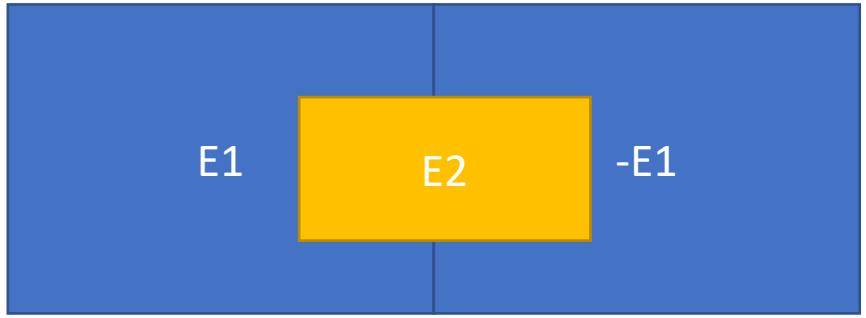
$$P(A) = \sum P(A|E_i)P(E_i)$$



Beispiel: n=6
A überlagert die anderen Ereignisse, tritt also immer bedingt auf.

Beispiel Lostopf: Totale Wahrscheinlichkeit

- $E1$ = erstes gezogenes Los gewinnt
- $E2$ = zweites gezogenes Los gewinnt



$$P(E2) = P(E1) \cdot P(E2|E1) + P(-E1) \cdot P(E2|-E1)$$

Satz von Bayes

1. Bedingte Wahrscheinlichkeit $P(E_j|A) = \frac{P(E_j \cdot A)}{P(A)}$
2. UND-Verknüpfung $P(E_j \cdot A) = P(E_j) \cdot P(A|E_j)$
3. Totale Wahrscheinlichkeit $P(A) = \sum P(E_i) \cdot P(A|E_i)$

2. und 3. in 1. eingesetzt:

$$P(E_j|A) = \frac{P(E_j) \cdot P(A|E_j)}{\sum P(E_i) \cdot P(A|E_i)}$$

Aufg. 2 (Skript): Medizinische Diagnostik

Positiver Test (Ereignis B), wie wahrscheinlich ist Patient dann wirklich krank (Ereignis A)?

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|-A)P(-A)}$$

$$= \frac{0{,}98 \cdot 0{,}001}{0{,}98 \cdot 0{,}001 + 0{,}003 \cdot (1 - 0{,}001)}$$

$$= 0{,}032$$

2. Beispiel **Satz von Bayes** in der medizinischen Diagnostik

$A =$ Patient ist krank

$B =$ Testergebnis ist positiv

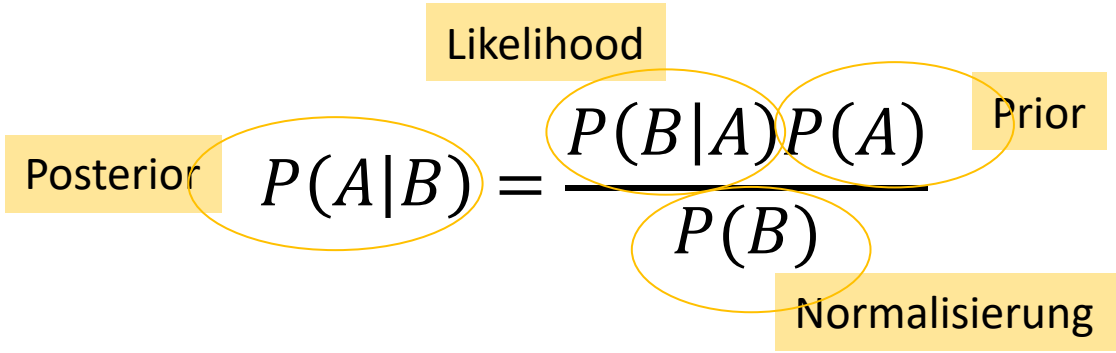
$W(B|A) = W(\text{Testergebnis ist positiv bei Kranken}) = 0.98$

$W(B|-A) = W(\text{Testergebnis ist positiv bei Nicht-Kranken}) = 0.03$

$W(A) = 0.001(\text{Krankheit ist sehr selten!})$

Wie wahrscheinlich ist es, dass Patient wirklich krank ist, wenn Testergebnis positiv ist ($W(A|B)$ ist gesucht)? Zeichnen Sie einen Ereignisbaum.

Prior- und Posterior-Information



Prior: unsere Informationen, die wir vorher haben (Beobachtung der Krankheit)

Likelihood: wenn unsere Prior-Information richtig ist, dann ist das unsere WSK, B zu beobachten (*WSK, das zu erhalten, was wir erhalten haben*)

Posterior: Update unserer Prior-Beobachtung aufgrund der gemachten Beobachtung

Normalisierung: Gewährleistung der Grenzen 0 und 1.

Update Prior-Information

- Vor dem Test:

Patient hat Krankheit mit P(A)=0,1% Wahrscheinlichkeit

- Nach dem positiven Test: Patient hat Krankheit mit P(A|B)=3,2% Wahrscheinlichkeit
- Zweiter positiver Test mit Update des Priors:

$$P(A|B) = \frac{0{,}98 \cdot 0{,}032}{0{,}98 \cdot 0{,}032 + 0{,}03 \cdot (1 - 0{,}032)} = 0{,}519$$

Prior kann weitere Informationen berücksichtigen.

Bayes Statistik vs. Frequent Statistik.

Assoziationsregeln

Warenkorbanalyse
Korb 1: Nudeln, Tomaten
Korb 2: Tomaten, Parmesan
Korb 3: Bratwurst
Korb 4: Nudeln, Tomaten, Parmesan
Korb 5: Nudeln, Tomaten

$X \Rightarrow Y$

Support:

$$\sup(X \cdot Y) = \frac{Anzahl\ Transaktionen\ mit\ X\ und\ Y}{Gesamtanzahl\ Transaktionen} \approx P(X \cdot Y)$$

Wie oft kommen die Produkte zusammen überhaupt vor?

ANNÄHERUNG an die Wahrscheinlichkeit!

Regel: Nudeln ⇒ Tomaten, Support(Nudeln · Tomaten)=3/5

Assoziationsregeln

Warenkorbanalyse
Korb 1: Nudeln, Tomaten
Korb 2: Tomaten, Parmesan
Korb 3: Bratwurst
Korb 4: Nudeln, Tomaten, Parmesan
Korb 5: Nudeln, Tomaten

$X \Rightarrow Y$

Confidence: Wie oft war die Regel richtig?

$$conf(X) = \frac{Anzahl\ Transaktionen\ mit\ X\ und\ Y}{Anzahl\ Transaktionen\ mit\ X}$$

Regel: Nudeln ⇒ Tomaten, Confidence = 3/3

Assoziationsregeln

Warenkorbanalyse
Korb 1: Nudeln, Tomaten
Korb 2: Tomaten, Parmesan
Korb 3: Bratwurst
Korb 4: Nudeln, Tomaten, Parmesan
Korb 5: Nudeln, Tomaten

$X \Rightarrow Y$

Lift: Wie stark ist Abhängigkeit von LHS und RHS?

$$Lift \approx \frac{P(X \cdot Y)}{P(X) \cdot P(Y)}$$

Bei Unabhängigkeit gilt Lift = 1, bei Abhängigkeit Lift > 1

Regel: Nudeln ⇒ Tomaten, $Lift = \frac{\frac{3}{5}}{\frac{3}{5} \cdot \frac{4}{5}} = \frac{5}{4} > 1$

Assoziationsregeln

Warenkorbanalyse
Korb 1: Nudeln, Tomaten
Korb 2: Tomaten, Parmesan
Korb 3: Bratwurst
Korb 4: Nudeln, Tomaten, Parmesan
Korb 5: Nudeln, Tomaten

$X \Rightarrow Y$

<u>Coverage</u>:

Wie oft kommt LHS (X) in den Transaktionen vor – wie oft kann die Regel also angewendet werden?

$$\sup(X) = \frac{Anzahl\ Transaktionen\ mit\ X}{Gesamtanzahl\ Transaktionen}$$

Regel: Nudeln ⇒ Tomaten, $\sup(Nudeln) = 3/5$

Aufg. 8 (Skript): Assoziationsregeln

8. Assoziationsregeln

a) Installieren und laden Sie das **R**-Paket **arules**.

b) Suchen sie auf der Hilfeseite die zugehörige Vignette *Introduction to arules*. Gehen Sie durch die Beispiele der Vignette.

c) Analysieren Sie den Datensatz **Groceries** hinsichtlich der Artikel, die am häufigsten gekauft worden sind und welche Assoziationsregeln sich ableiten lassen.

Dataset 33 http://casas.wsu.edu/datasets/hh111.zip

- Überlegt euch eine Strategie, wie Assoziationsregeln auch helfen können, das gemeinsame Verändern des Status von Sensoren bei bestimmten Aktivitäten zu beschreiben.