Einführung in R

S. Trahasch, S. Niro

Einordnung

R

S, S-Plus, R

Vorteile von R

Nachteile von R

Rechnen mit R

Vektoren

Übung I

Übung II

Data Frame

Übung III (Teil1)

Einschub

Übung III (Teil2)

# Einführung in R

S. Trahasch, S. Niro

8. Oktober 2017

# Ziele

Einführung in R

S. Trahasch, S. Niro

Einordnung

R

S, S-Plus, R

Vorteile von R

Nachteile von R

Rechnen mit R

Vektoren

Übung I

Übung II

Data Frame

Übung III (Teil1)

Einschub

Übung III (Teil2)

- Grundlegende Konzepte kennen
- Wichtige Datenstrukturen und Befehle verwenden können
- Einfache Datenvisualisierungen erstellen können
- Packages installieren und verwenden können

Einordnung

R

S, S-Plus, R

Vorteile von R

Nachteile von R

Rechnen mit R

Vektoren

Übung I

Übung II

Data Frame

Übung III (Teil1)

Einschub

Übung III (Teil2)

1. Einordnung
2. R
   - S, S-Plus, R
   - Vorteile von R
   - Nachteile von R
3. Rechnen mit R
4. Vektoren
5. Übung I
6. Übung II
7. Data Frame
8. Übung III (Teil1)
9. Einschub
10. Übung III (Teil2)

# Einordnung (kdnuggets.com Umfrage 2015-2017)

Einführung in R

S. Trahasch, S. Niro

Einordnung

R

S, S-Plus, R

Vorteile von R

Nachteile von R

Rechnen mit R

Vektoren

Übung I

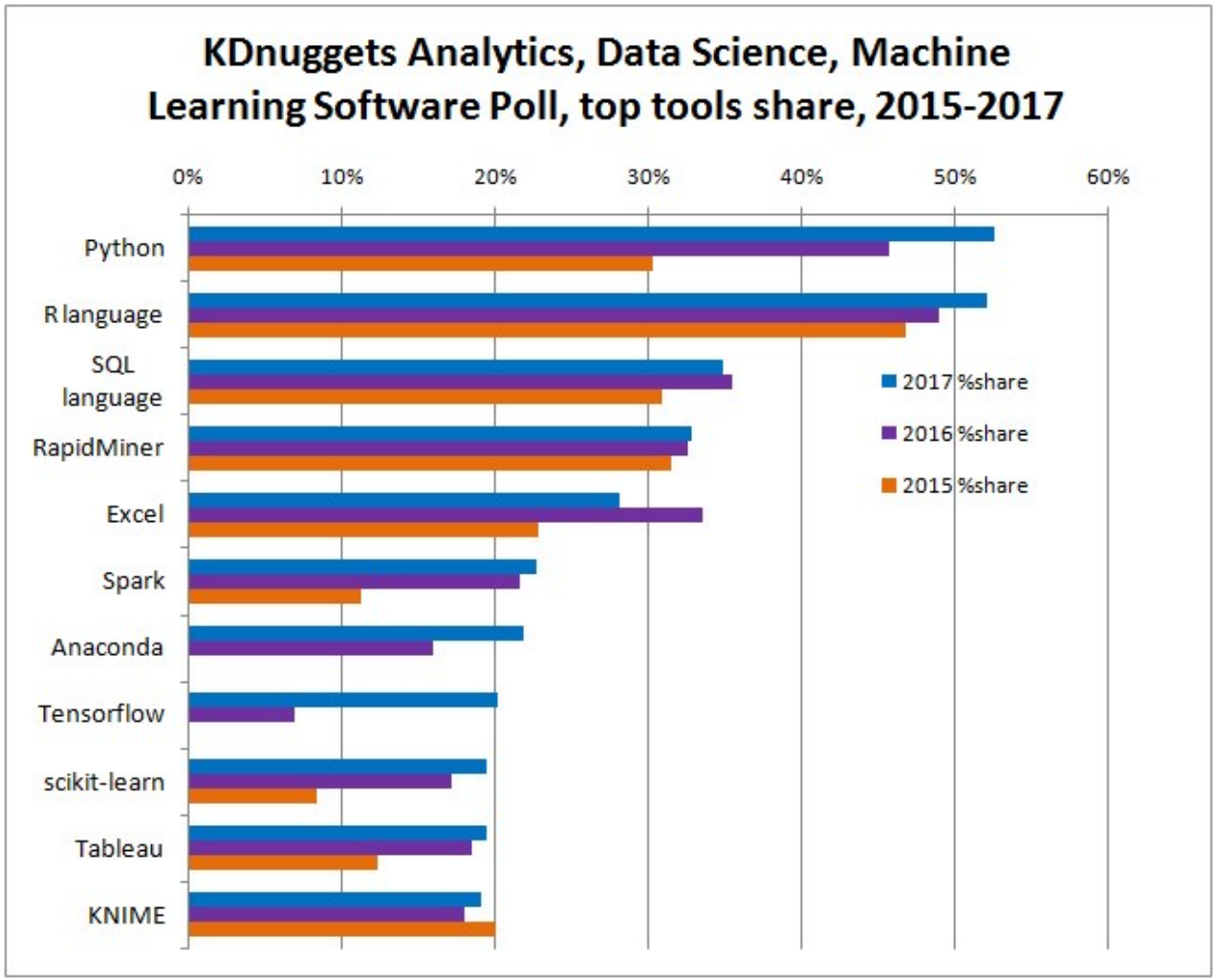
Übung II

Data Frame

Übung III (Teil1)

Einschub

Übung III (Teil2)



# Einordnung (kdnuggets.com Umfrage 2015-2017)

Einführung in R

S. Trahasch, S. Niro

Einordnung

R

S, S-Plus, R

Vorteile von R

Nachteile von R

Rechnen mit R

Vektoren

Übung I

Übung II

Data Frame

Übung III (Teil1)

Einschub

Übung III (Teil2)

Table 1: Top Analytics/Data Science Tools in 2017 KDnuggets Poll

| Tool | 2017 % Usage | % change 2017 vs 2016 | % alone |
|---|---|---|---|
| Python | 52.6% | 15% | 0.2% |
| R language | 52.1% | 6.4% | 3.3% |
| SQL language | 34.9% | -1.8% | 0% |
| RapidMiner | 32.8% | 0.7% | 13.6% |
| Excel | 28.1% | -16% | 0.1% |
| Spark | 22.7% | 5.3% | 0.2% |
| Anaconda | 21.8% | 37% | 0.8% |
| Tensorflow | 20.2% | 195% | 0% |
| scikit-learn | 19.5% | 13% | 0% |
| Tableau | 19.4% | 5.0% | 0.4% |
| KNIME | 19.1% | 6.3% | 2.4% |

# Einordnung

Einführung in R

S. Trahasch, S. Niro

Einordnung

R

S, S-Plus, R

Vorteile von R

Nachteile von R

Rechnen mit R

Vektoren

Übung I

Übung II

Data Frame

Übung III (Teil1)

Einschub

Übung III (Teil2)

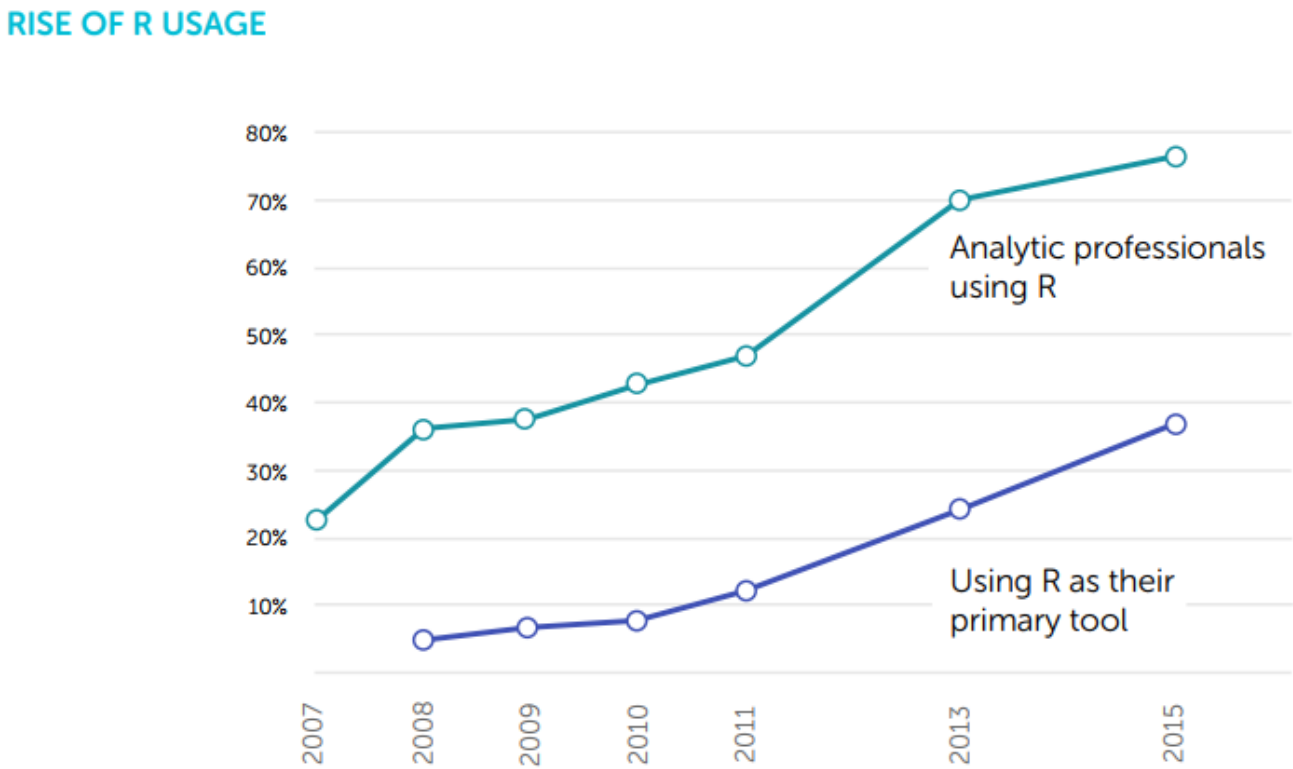



Abbildung: Rexer Analytics Data Miner Survey 2015

# Einordnung

Einführung in R
S. Trahasch, S. Niro
Einordnung
R
S, S-Plus, R
Vorteile von R
Nachteile von R
Rechnen mit R
Vektoren
Übung I
Übung II
Data Frame
Übung III (Teil1)
Einschub
Übung III (Teil2)

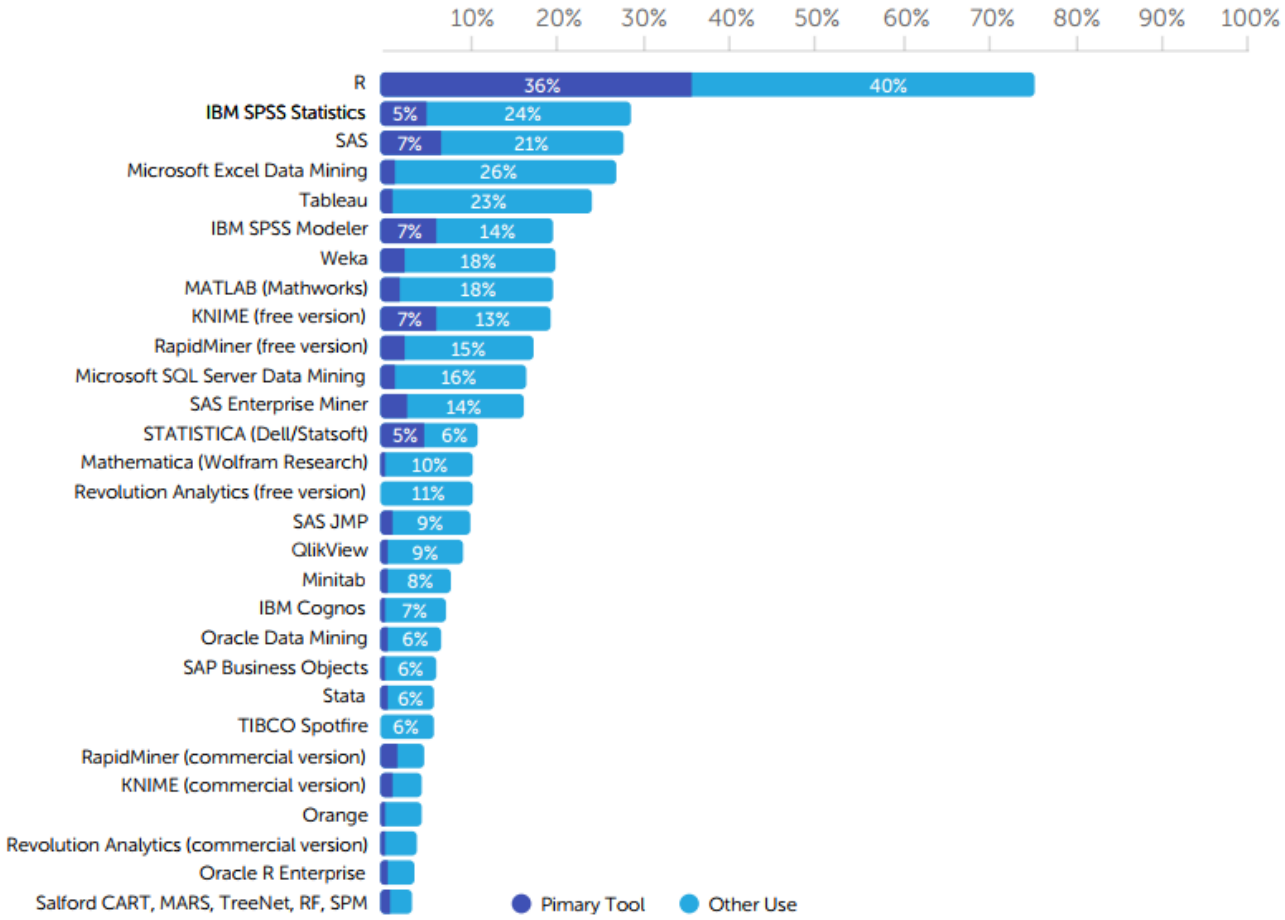



Abbildung: Rexer Analytics Data Miner Survey 2015

# Geschichtliches: S, S-Plus, R

Einführung in R

S. Trahasch, S. Niro

Einordnung

R

S, S-Plus, R

Vorteile von R

Nachteile von R

Rechnen mit R

Vektoren

Übung I

Übung II

Data Frame

Übung III (Teil1)

Einschub

Übung III (Teil2)

- Becker, R. A. und Chambers, J. M. veröffentlichten 1984 Sprache **S** für **Datenanalyse (Statistik)** und **Grafik**
- **S-PLUS** ist eine kommerzielle Implementation von S
- **R** ist eine Open Source Implementation (GNU GPL) von S, 1992 von **R**oss Ihaka und **R**obert Gentleman entwickelt

# Vorteile von R

Einführung in R

S. Trahasch, S. Niro

Einordnung

R

S, S-Plus, R

Vorteile von R

Nachteile von R

Rechnen mit R

Vektoren

Übung I

Übung II

Data Frame

Übung III (Teil1)

Einschub

Übung III (Teil2)

- Domänenspezifische Sprache für Datenanalyse und Visualisierung
- Open Source, keine Lizenzgebühren (GNU GPL)
- Große aktive Community
- Crossplattform: Windows, Linux, Solaris, usw.
- Sehr viele ($>$ 5500) R-Pakete. Neue statistische Methoden werden oft als (kostenlose) R-Pakete angeboten
- Schneller als S-Plus
- Programmierschnittstellen zu R für viele Sprachen verfügbar (Java, Python, ... )
- Integration von R durch andere Datenanalysesoftware (Rapidminer, SAP HANA, SPSS, SAS, ... )

# Grafiken mit R - Beispiele (Entwicklung der R-Packages)

Einführung in R

S. Trahasch, S. Niro

Einordnung

R

S, S-Plus, R

Vorteile von R

Nachteile von R

Rechnen mit R

Vektoren

Übung I

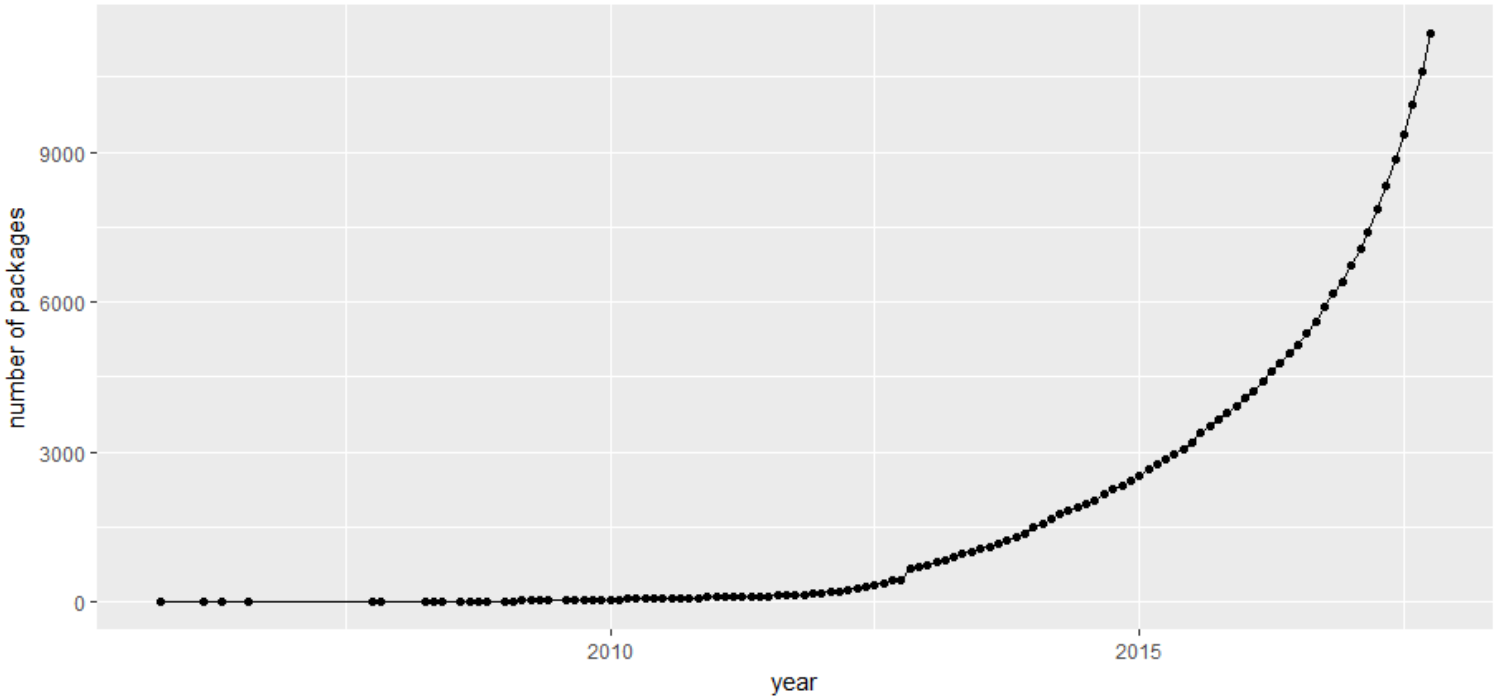
Übung II

Data Frame

Übung III (Teil1)

Einschub

Übung III (Teil2)



# Grafiken mit R - Beispiele (R-User Herkunft)

Einführung in R

S. Trahasch, S. Niro

Einordnung

R

S, S-Plus, R

Vorteile von R

Nachteile von R

Rechnen mit R

Vektoren

Übung I

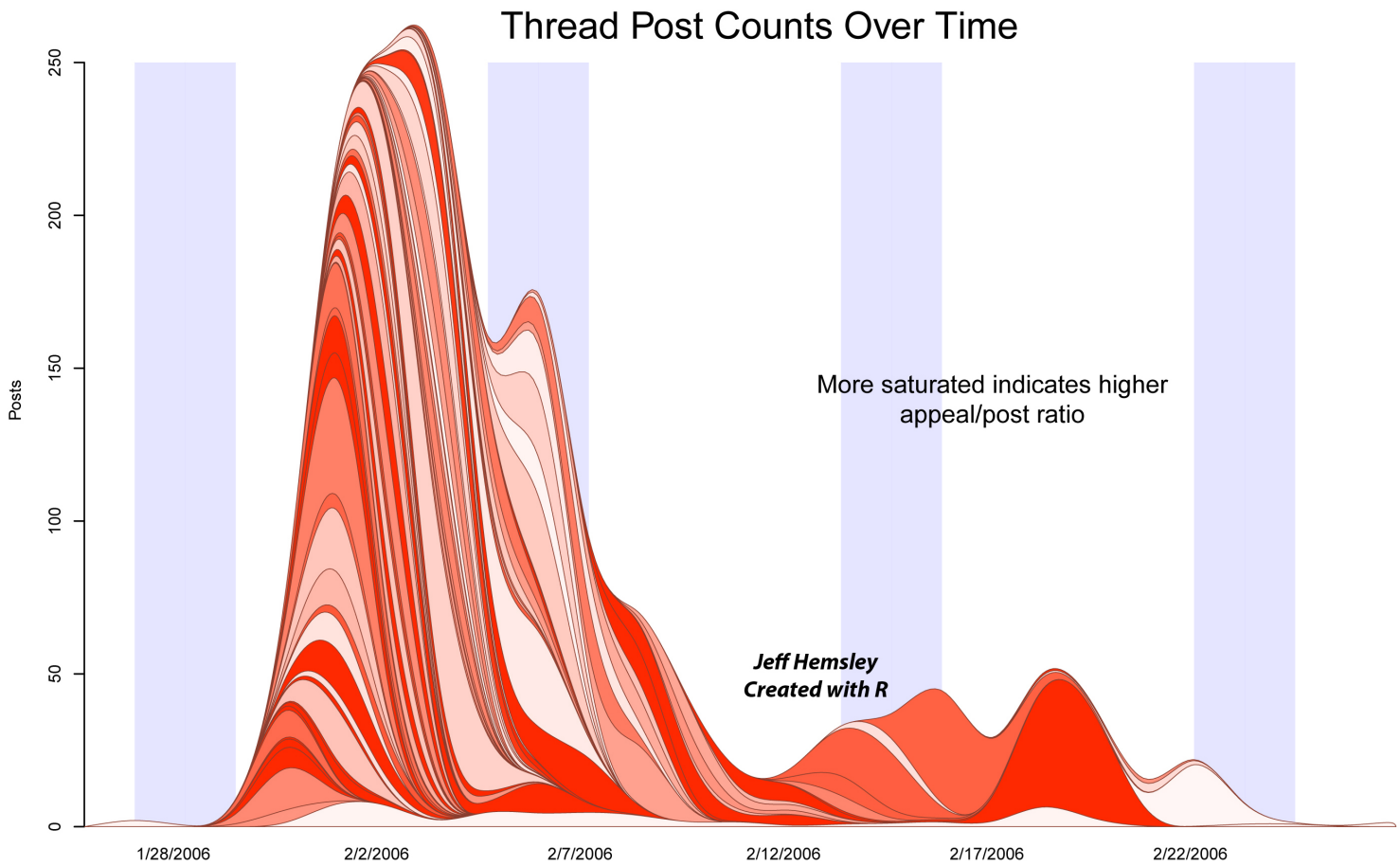
Übung II

Data Frame

Übung III (Teil1)

Einschub

Übung III (Teil2)



# Grafiken mit R - Beispiele

Einführung in R

S. Trahasch, S. Niro

Einordnung

R

S, S-Plus, R

Vorteile von R

Nachteile von R

Rechnen mit R

Vektoren

Übung I

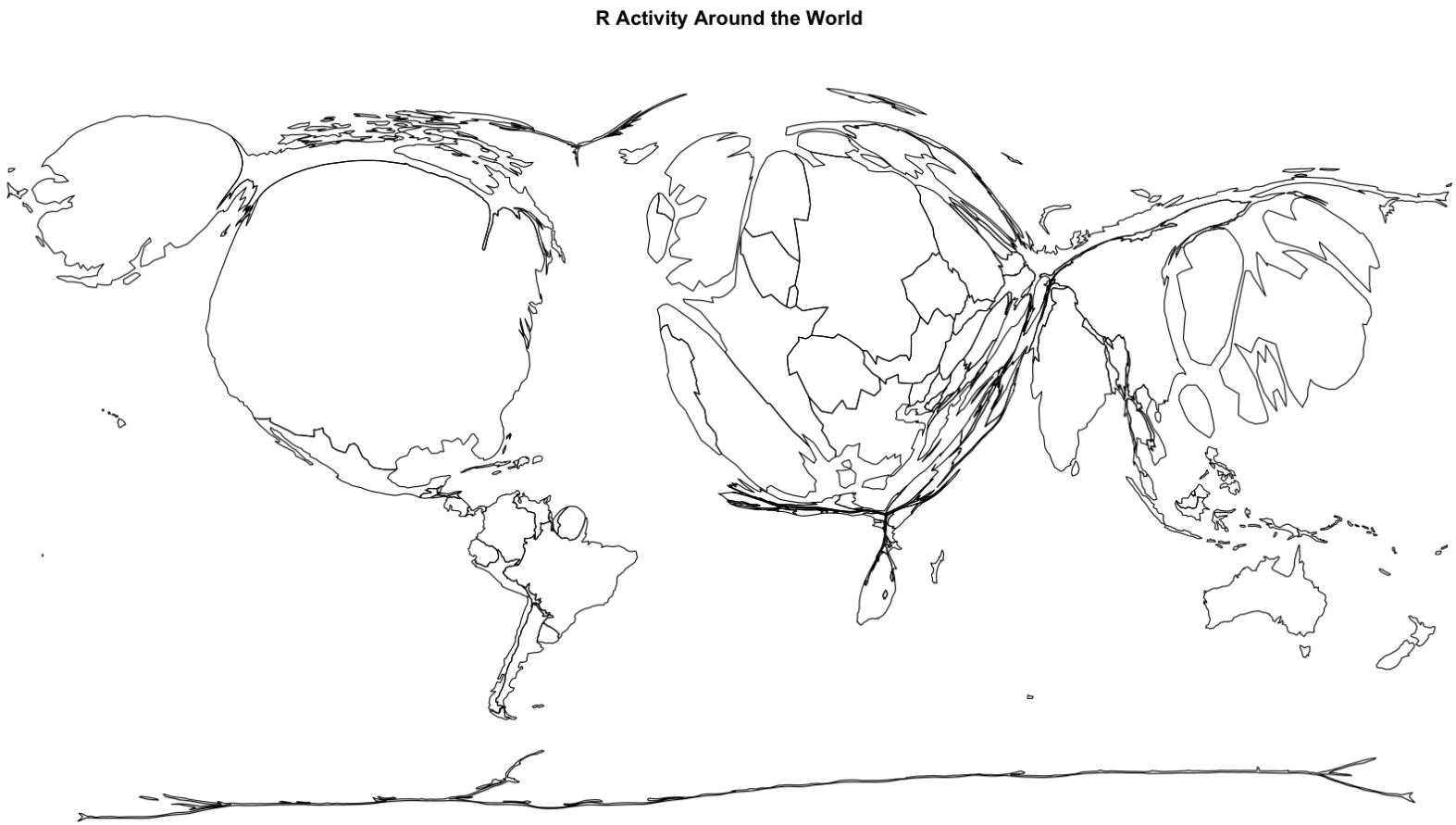
Übung II

Data Frame

Übung III (Teil1)

Einschub

Übung III (Teil2)



# Grafiken mit R - Beispiele

Einführung in R

S. Trahasch, S. Niro

Einordnung

R

S, S-Plus, R

Vorteile von R

Nachteile von R

Rechnen mit R

Vektoren

Übung I

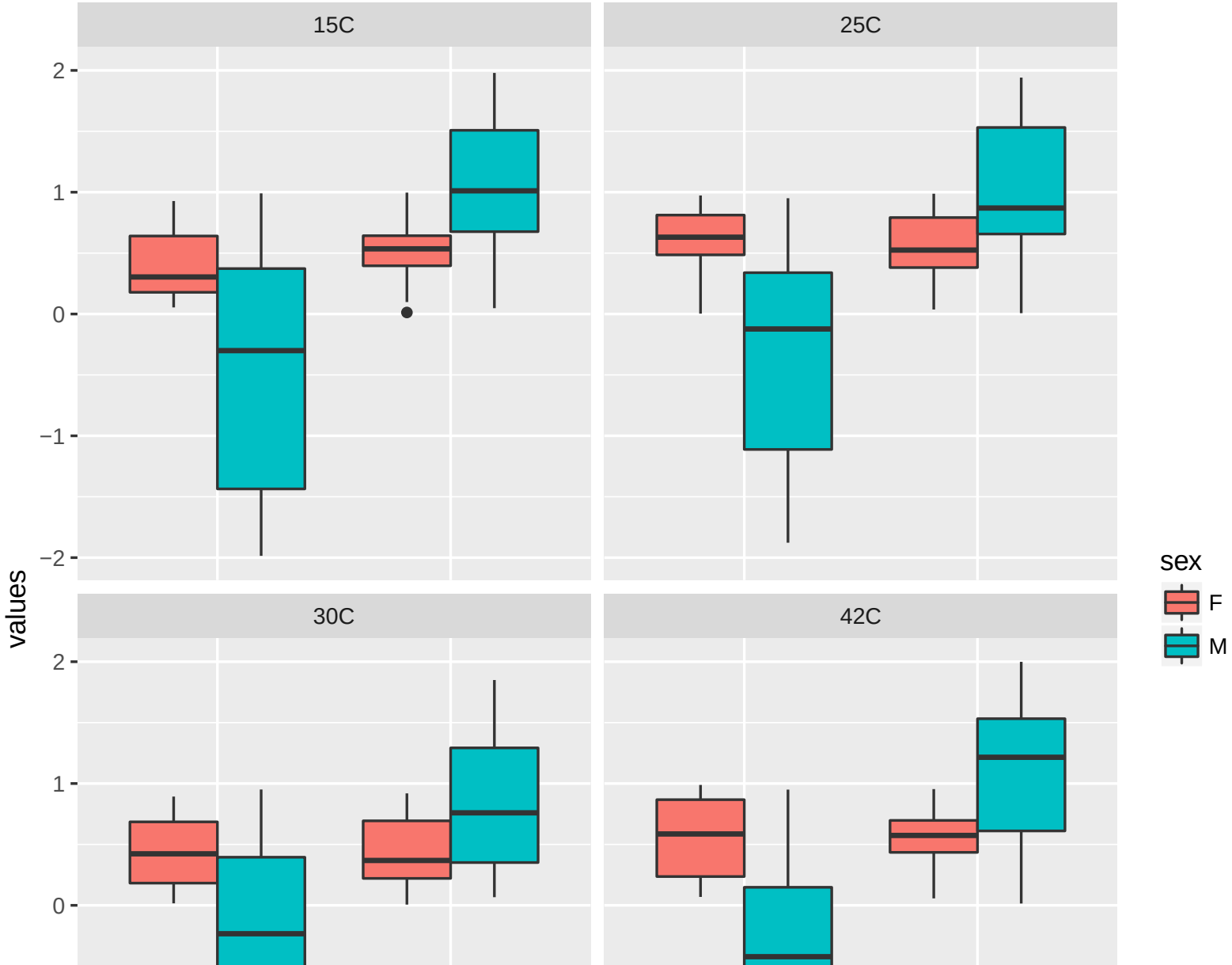
Übung II

Data Frame

Übung III (Teil1)

Einschub

Übung III (Teil2)



# Grafiken mit R - Beispiele



Einordnung

R

S, S-Plus, R

Vorteile von R

Nachteile von R

Rechnen mit R

Vektoren

Übung I

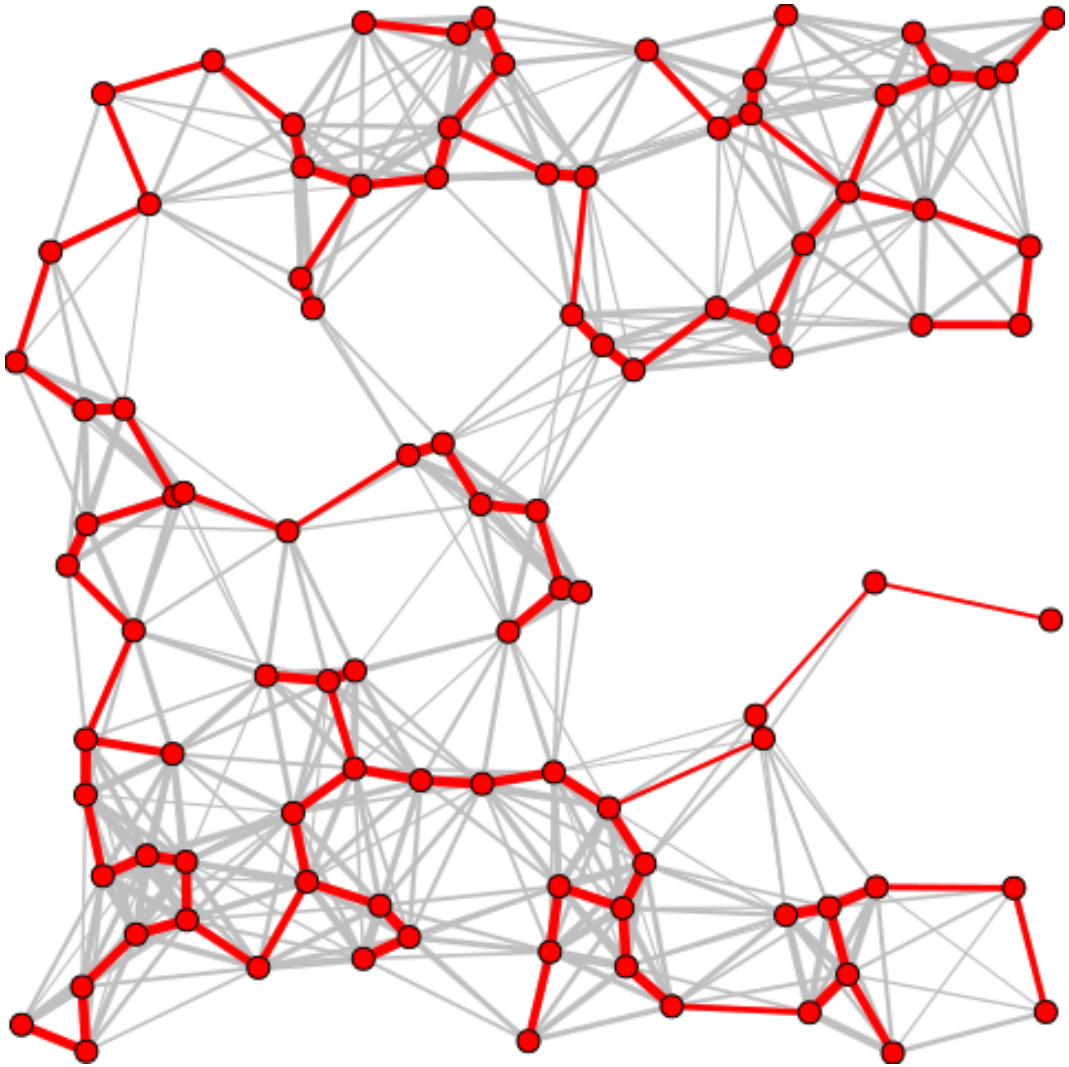
Übung II

Data Frame

Übung III (Teil1)

Einschub

Übung III (Teil2)



# Grafiken mit R - Beispiele


Einführung in R

S. Trahasch, S. Niro

Einordnung

R

S, S-Plus, R

Vorteile von R

Nachteile von R

Rechnen mit R

Vektoren

Übung I

Übung II

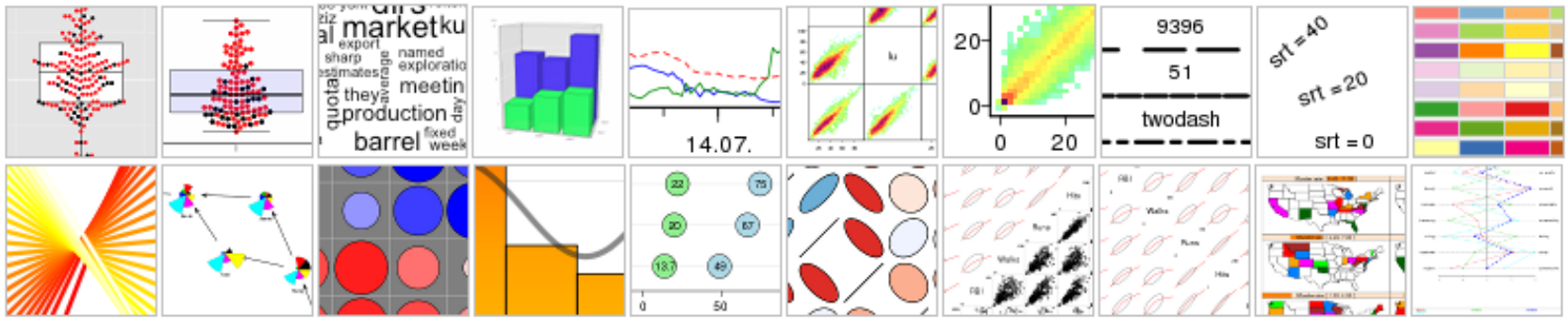
Data Frame

Übung III (Teil1)

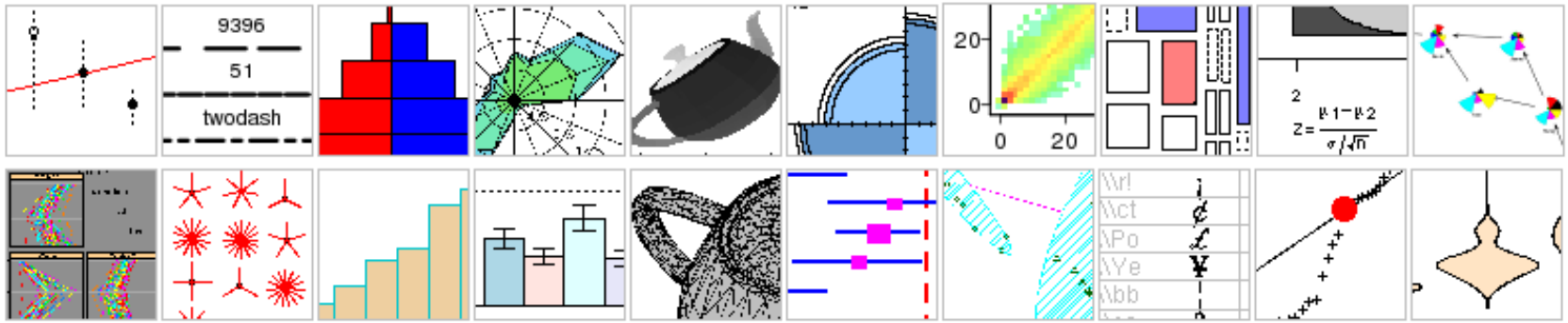
Einschub

Übung III (Teil2)


» Last entries ...




» Random entries




Weitere: http://www.sr.bham.ac.uk/ ajrs/R/r-gallery.html
http://addictedtor.free.fr/graphiques/

# Grafiken (Dilbert)

Einführung in R

S. Trahasch, S. Niro

Einordnung

R

S, S-Plus, R

Vorteile von R

Nachteile von R

Rechnen mit R

Vektoren

Übung I

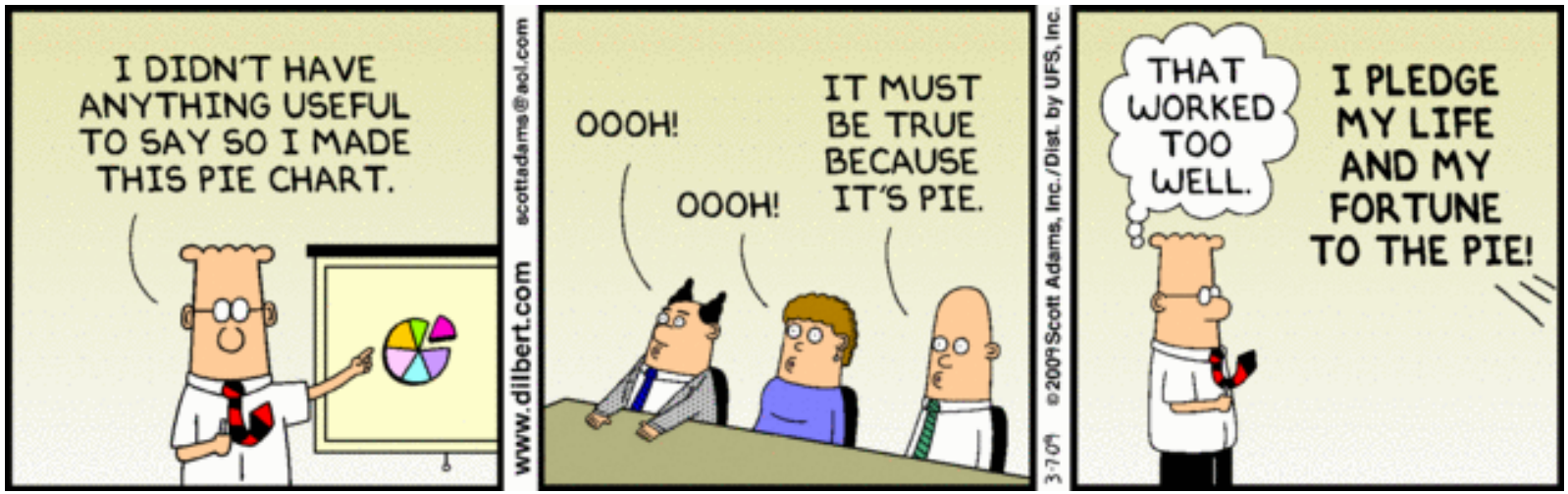
Übung II

Data Frame

Übung III (Teil1)

Einschub

Übung III (Teil2)



# Nachteile von R



- Keine vollwertige grafische Benutzeroberfläche
- Lernkurve etwas flacher als bei anderer SW
- Qualität der Packete hängt von der Anzahl der Benutzer ab
- Fehlermeldungen nicht immer hilfreich

# R als Taschenrechner

Einführung in R
S. Trahasch, S. Niro
Einordnung
R
S, S-Plus, R
Vorteile von R
Nachteile von R
Rechnen mit R
Vektoren
Übung I
Übung II
Data Frame
Übung III (Teil1)
Einschub
Übung III (Teil2)

```
3.5 + 1.5

[1] 5
```

```
x <- 6 * (1/3)   # Zuweisung
                 # (empfohlen)
x

[1] 2
```

```
x = 2^2          # Zuweisung
print(x)

[1] 4
```

| Operator | |
|---|---|
| + | Addition |
| − | Subtraktion |
| * | Multiplikation |
| / | Division |
| ^ | Potenz |
| %% | Modulo (Rest) |

Weitere math. Funktionen: sin(x), sqrt(x), exp(x), ...

# Vektoren

Einführung in R

S. Trahasch, S. Niro

Einordnung

R

S, S-Plus, R

Vorteile von R

Nachteile von R

Rechnen mit R

Vektoren

Übung I

Übung II

Data Frame

Übung III (Teil1)

Einschub

Übung III (Teil2)

## Geordnete Menge von Elementen gleichen Typs

```
a <- c(4, 5, 6) # combine

a

[1] 4 5 6
```

```
length(a) # Länge von a

[1] 3
```

```
a[2]        # zweites Element in a

[1] 5
```

# Vektoren: Arithmetik

Einführung in R
S. Trahasch, S. Niro
Einordnung
R
S, S-Plus, R
Vorteile von R
Nachteile von R
Rechnen mit R
Vektoren
Übung I
Übung II
Data Frame
Übung III (Teil1)
Einschub
Übung III (Teil2)

```
a <- seq(from = 1, to = 3, by = 1) # entspricht c(1,2,3)
b <- 9:7 # entspricht c(9, 8, 7)
a

[1] 1 2 3

b

[1] 9 8 7
```

## händisch

```
  c <- c(0,0,0)
  for(i in 1:length(a))
  {
c[i] <- a[i] + b[i]
  }
  c
```

## Vektorisiert (empfohlen)

```
c <- a + b
c

[1] 10 10 10
```

# Vektoren: Recycling

Einführung in R
S. Trahasch, S. Niro
Einordnung
R
S, S-Plus, R
Vorteile von R
Nachteile von R
Rechnen mit R
Vektoren
Übung I
Übung II
Data Frame
Übung III (Teil1)
Einschub
Übung III (Teil2)

```
a <- 1:6
a

[1] 1 2 3 4 5 6

a + c(1,2) # ???

[1] 2 4 4 6 6 8
```

$$\begin{pmatrix}1\\2\\3\\4\\5\\6\end{pmatrix} + \begin{pmatrix}1\\2\end{pmatrix} \longrightarrow^{recycling} \begin{pmatrix}1\\2\\3\\4\\5\\6\end{pmatrix} + \begin{pmatrix}1\\2\\1\\2\\1\\2\end{pmatrix} = \begin{pmatrix}2\\4\\4\\6\\6\\8\end{pmatrix}$$

# Vektoren und Funktionen

Einführung in R
S. Trahasch, S. Niro
Einordnung
R
S, S-Plus, R
Vorteile von R
Nachteile von R
Rechnen mit R
Vektoren
Übung I
Übung II
Data Frame
Übung III (Teil1)
Einschub
Übung III (Teil2)

```
a <- 1:4
```

Funktionen, die auf Skalare angewandt werden, werden auf jedes Element des Vektors angewandt

```
sqrt(a) # Wurzel
```

```
[1] 1.000000 1.414214 1.732051 2.000000
```

```
max(a^2) # größtes element
```

```
[1] 16
```

```
sum(a^2) # summe aller elemente
```

```
[1] 30
```

# R-Studio

Einführung in R

S. Trahasch, S. Niro

Einordnung

R

S, S-Plus, R

Vorteile von R

Nachteile von R

Rechnen mit R

Vektoren

Übung I

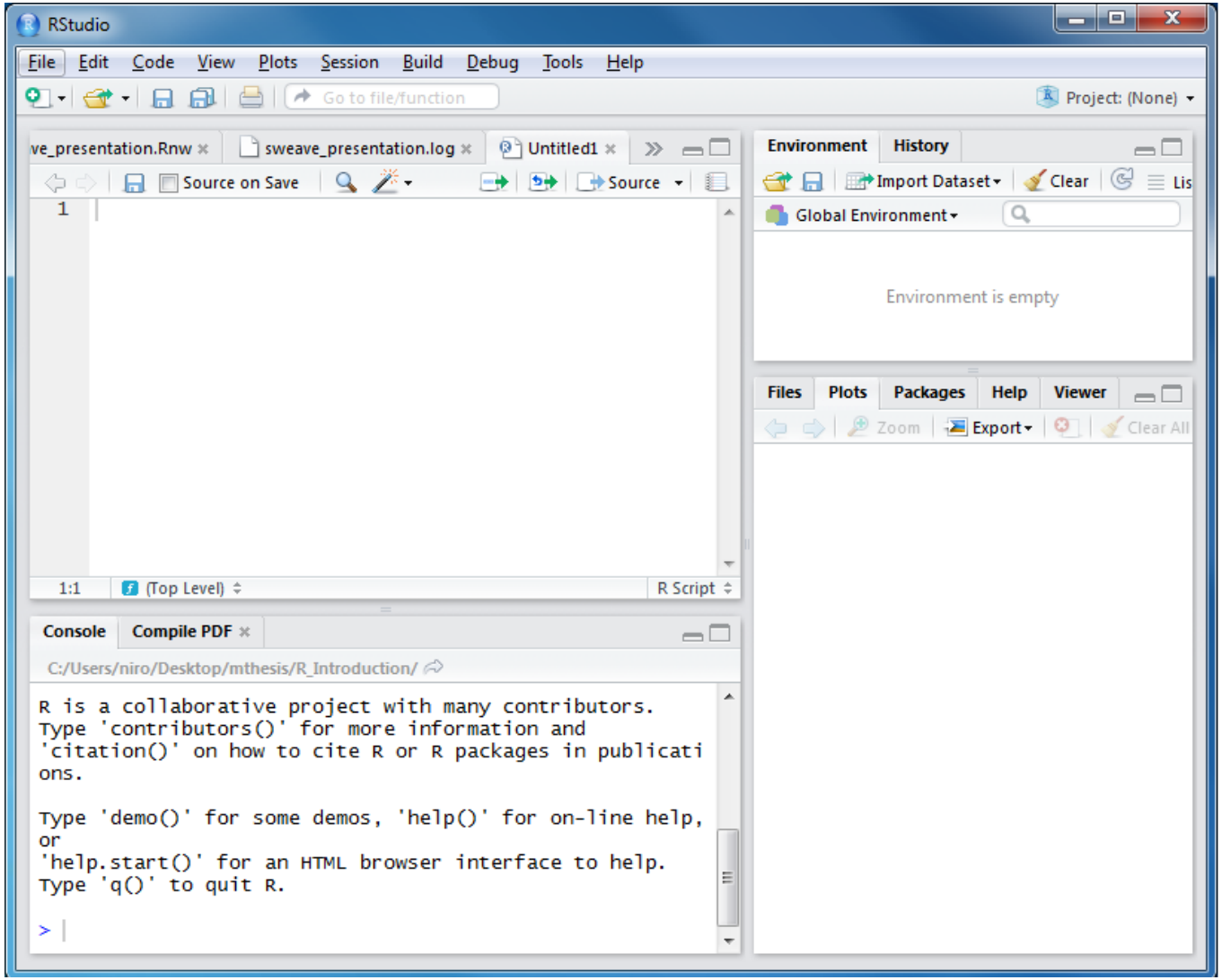
Übung II

Data Frame

Übung III (Teil1)

Einschub

Übung III (Teil2)



# Übung I

Einführung in R
S. Trahasch, S. Niro
Einordnung
R
S, S-Plus, R
Vorteile von R
Nachteile von R
Rechnen mit R
Vektoren
Übung I
Übung II
Data Frame
Übung III (Teil1)
Einschub
Übung III (Teil2)

- Erstellen Sie einen Vektor $x$ von Ganzzahlen im Intervall $[-10; 10]$
- Wieviele Elemente sind in $x$ (length)?
- Welche Werte haben das 10.te und das 22.te Element?
- Berechnen Sie $y(x) = -x^2 + 20$
- Was ist der kleinste/größte Funktionswert von $y(x)$ (min/max)?
- Plotten Sie die Funktion mit plot(x, y)
- Fügen Sie dem Funktionsaufruf das Argument

```
type = "l"
```

hinzu. Wie verändert sich der Plot für

```
type = "b"
```

```
type = "p"
```

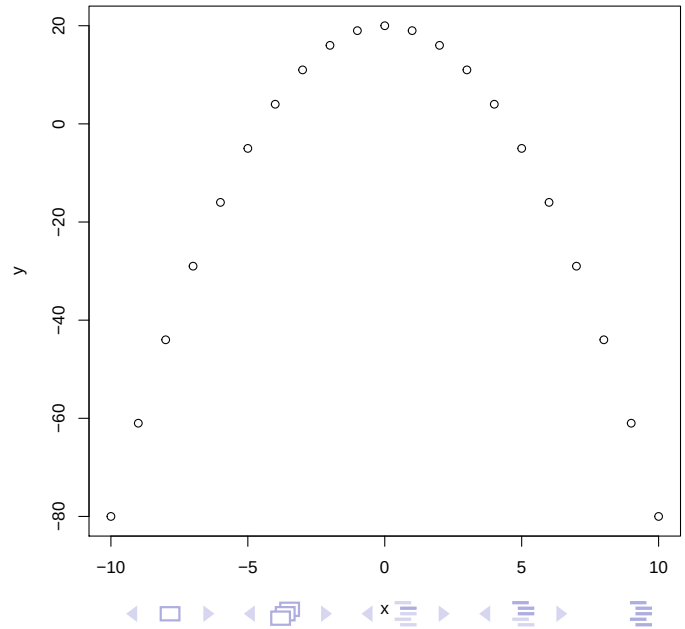
- Optional: Berechnen Sie $\bar{y} = \frac{1}{N} * \sum_{i=1}^{N}(y_i)$

# Übung I

Einführung in R
S. Trahasch, S. Niro
Einordnung
R
S, S-Plus, R
Vorteile von R
Nachteile von R
Rechnen mit R
Vektoren
Übung I
Übung II
Data Frame
Übung III (Teil1)
Einschub
Übung III (Teil2)

```
x <- -10:10
length(x)

[1] 21

x[10]

[1] -1

y <- -x^2 + 20
min(y)

[1] -80

max(y)

[1] 20
```

```
plot(x,y)
1/length(y) * sum(y)
mean(y)

[1] -16.66667
[1] -16.66667
```



# Übung II

Einführung in R

S. Trahasch, S. Niro

Einordnung

R

S, S-Plus, R

Vorteile von R

Nachteile von R

Rechnen mit R

Vektoren

Übung I

Übung II

Data Frame

Übung III (Teil1)

Einschub

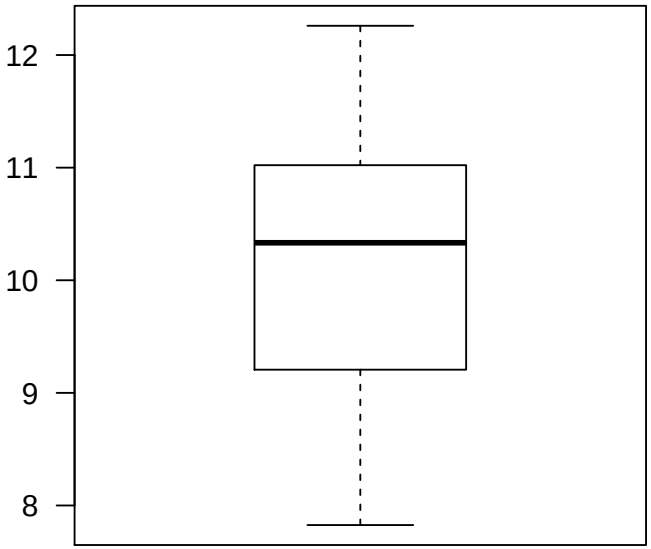
Übung III (Teil2)

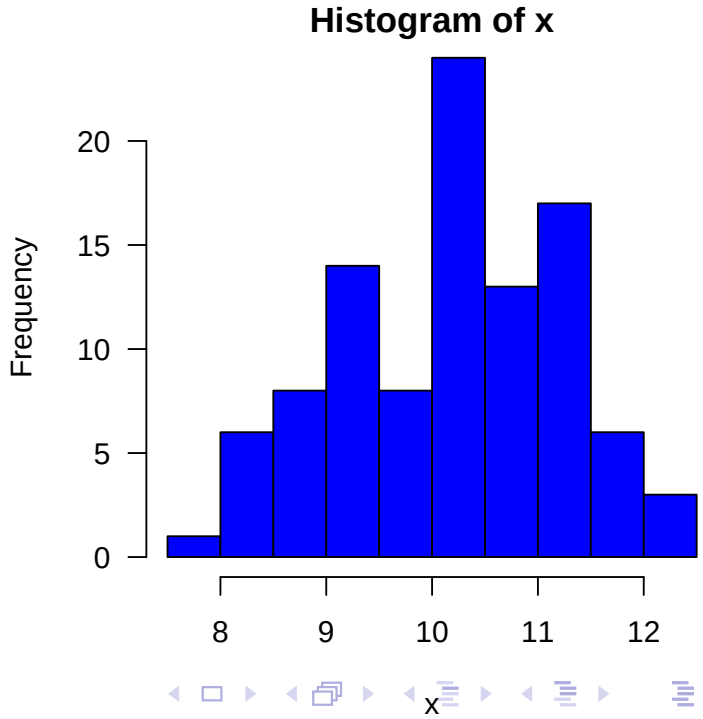
- Erstellen Sie mit **rnorm** $n = 100$ normalverteilte Zufallswerte mit Mittelwert 10 und Standardabweichung 1 (Hilfe über **?rnorm** oder help(rnorm))
- Überprüfen Sie den Mittelwert (**mean**) und die Standardabweichung (**sd**)
- Erstellen Sie einen Boxplot (**boxplot**) und ein Histogram (**hist**)
- Wiederholen Sie das Ganze mit $n = 10000$. Was fällt Ihnen auf?
- Optional: Verwenden Sie gleichverteilte Zufallswerte (**runif**)

# Übung II (Lösung)

Einführung in R
S. Trahasch, S. Niro
Einordnung
R
S, S-Plus, R
Vorteile von R
Nachteile von R
Rechnen mit R
Vektoren
Übung I
Übung II
Data Frame
Übung III (Teil1)
Einschub
Übung III (Teil2)

```
x <- rnorm(100, mean=10, sd=1)
mean(x); sd(x)
par(las =1, mar=c(4,4,1,.1))
boxplot(x); hist(x, col="blue")

[1] 10.19487
[1] 1.051083
```

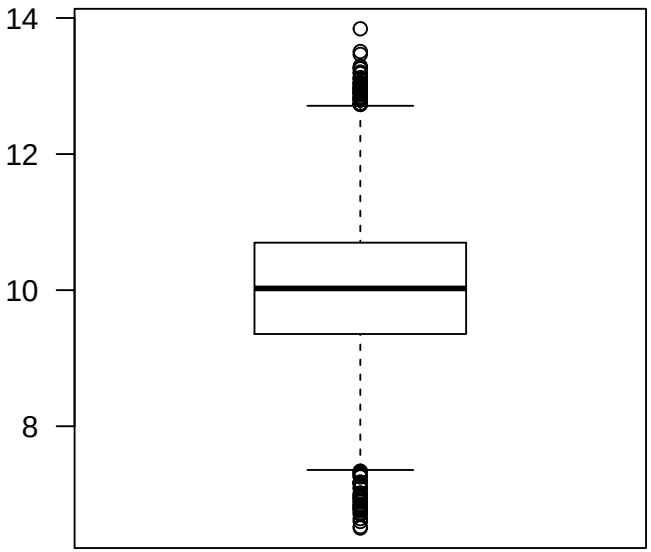


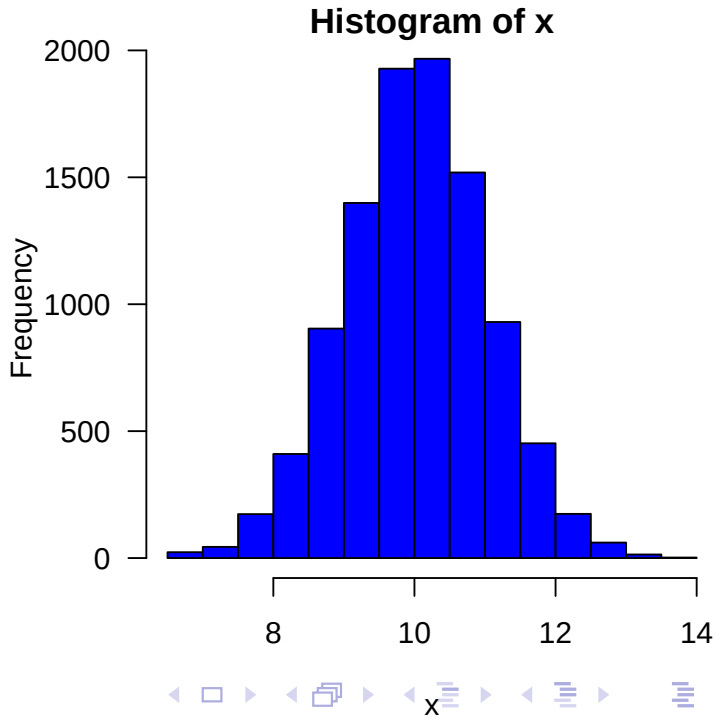



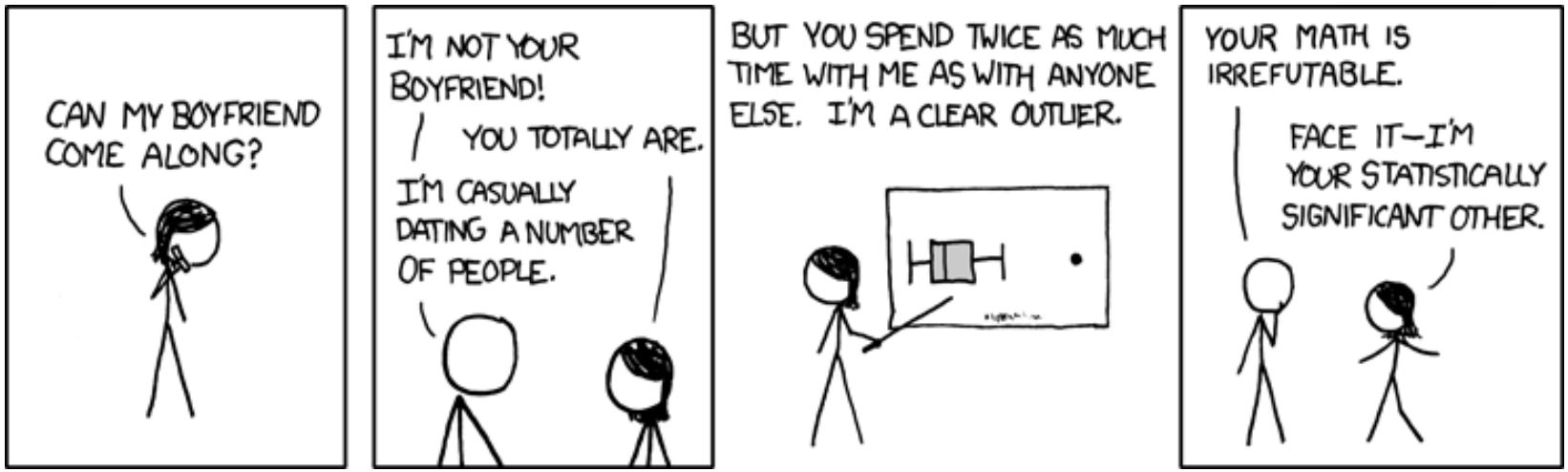
# Übung II (Lösung)

Einführung in R
S. Trahasch, S. Niro
Einordnung
R
S, S-Plus, R
Vorteile von R
Nachteile von R
Rechnen mit R
Vektoren
Übung I
Übung II
Data Frame
Übung III (Teil1)
Einschub
Übung III (Teil2)

```
x <- rnorm(10000, mean=10, sd=1)
mean(x); sd(x)
par(las =1, mar=c(4,4,1,.1))
boxplot(x); hist(x, col="blue")

[1] 10.02118
[1] 1.008185
```




CAN MY BOYFRIEND COME ALONG?
I'M NOT YOUR BOYFRIEND!
YOU TOTALLY ARE.
I'M CASUALLY DATING A NUMBER OF PEOPLE.
BUT YOU SPEND TWICE AS MUCH TIME WITH ME AS WITH ANYONE ELSE. I'M A CLEAR OUTLIER.
YOUR MATH IS IRREFUTABLE.
FACE IT—I'M YOUR STATISTICALLY SIGNIFICANT OTHER.

# Weitere Typen (*mode*)

Einführung in R
S. Trahasch, S. Niro
Einordnung
R
S, S-Plus, R
Vorteile von R
Nachteile von R
Rechnen mit R
Vektoren
Übung I
Übung II
Data Frame
Übung III (Teil1)
Einschub
Übung III (Teil2)

## Logical (Boolscher Wert)

```
verheiratet <- c(TRUE, FALSE, T, F, T)
print(verheiratet)

[1]  TRUE FALSE  TRUE FALSE  TRUE
```

## Character (Zeichenkette)

```
name <- c("Max", "Fritz")
print(name)

[1] "Max"   "Fritz"
```

## Factor (Nominalwert):

```
geschlecht <- factor(c("m","m","w","m","w","w"))
print(geschlecht)

[1] m m w m w w
Levels: m w
```

# Logische und relationale Operatoren

Einführung in R
S. Trahasch, S. Niro
Einordnung
R
S, S-Plus, R
Vorteile von R
Nachteile von R
Rechnen mit R
Vektoren
Übung I
Übung II
Data Frame
Übung III (Teil1)
Einschub
Übung III (Teil2)

```
10 == (2 + 8)
```

```
[1] TRUE
```

```
(10 %% 3 != 0) && (4 < 5)
```

```
[1] TRUE
```

```
!FALSE
```

```
[1] TRUE
```

| Operator | Bedeutung |
|---|---|
| == | Gleichheit |
| != | Ungleichheit |
| > | größer |
| <= | kleiner gleich |
| Logisch: | |
| ! | NOT |
| && | UND |
| \|\| | ODER |

```
c(5,8,10) > 5
```

```
[1] FALSE  TRUE  TRUE
```

# Bedingte Ausführung

Einführung in R

S. Trahasch, S. Niro

Einordnung

R

S, S-Plus, R

Vorteile von R

Nachteile von R

Rechnen mit R

Vektoren

Übung I

Übung II

Data Frame

Übung III (Teil1)

Einschub

Übung III (Teil2)

```
if(2+2==5)
{
  print("gleich")
}else
{
  print("ungleich")
}

[1] "ungleich"
```

kurz:

```
ifelse(2+2==5, "gleich", "ungleich")

[1] "ungleich"

ifelse(1:10==5, "gleich", "ungleich") # vektorisiert

 [1] "ungleich" "ungleich" "ungleich" "ungleich" "gleich"
 [6] "ungleich" "ungleich" "ungleich" "ungleich" "ungleich"
```

# Bedingte Auswahl

Einführung in R
S. Trahasch, S. Niro
Einordnung
R
S, S-Plus, R
Vorteile von R
Nachteile von R
Rechnen mit R
Vektoren
Übung I
Übung II
Data Frame
Übung III (Teil1)
Einschub
Übung III (Teil2)

```
a <- c(2,4,6,8,10)
```

```
a[1:3]  # indexbasierte Auwahl

[1] 2 4 6
```

```
a[c(T,T,T,F,F)] # bedingte Auswahl

[1] 2 4 6
```

```
a[a < 7] # bedingte Auswahl

[1] 2 4 6
```

# Data Frame

Einführung in R
S. Trahasch, S. Niro
Einordnung
R
S, S-Plus, R
Vorteile von R
Nachteile von R
Rechnen mit R
Vektoren
Übung I
Übung II
Data Frame
Übung III (Teil1)
Einschub
Übung III (Teil2)

- Liste aus Vektoren gleicher Länge (=Spalten), die Namen haben
- Wichtigste Datenstruktur
- Beispiel

| Name | Gruppe | Schuhgröße |
|---|---|---|
| Dennis | APC | 42 |
| Ralf | SIB | 43 |
| Stefan | IS | 42 |
| Susanne | APC | 39 |
| Swen | SIB | 42 |
| Werner | SIB | 43 |

Tabelle: Teilnehmerliste als CSV-Datei

- Zwei Indices: df[ Zeile(n) , Spalte(n) ]

# Data Frame

Einführung in R

S. Trahasch, S. Niro

Einordnung

R

S, S-Plus, R

Vorteile von R

Nachteile von R

Rechnen mit R

Vektoren

Übung I

Übung II

Data Frame

Übung III (Teil1)

Einschub

Übung III (Teil2)

```
df <- read.csv("teilnehmer.csv", sep=";")
df

     Name Gruppe Schuhgröße
1  Dennis    APC         42
2    Ralf    SIB         43
3  Stefan     IS         42
4 Susanne    APC         39
5    Swen    SIB         42
6  Werner    SIB         43

names(df)  # Spaltennamen

[1] "Name"       "Gruppe"     "Schuhgröße"

dim(df)        # dimensionen (zeilen, spalten)

[1] 6 3
```

# Data Frame: Zugriff auf Inhalte

Einführung in R
S. Trahasch, S. Niro
Einordnung
R
S, S-Plus, R
Vorteile von R
Nachteile von R
Rechnen mit R
Vektoren
Übung I
Übung II
Data Frame
Übung III (Teil1)
Einschub
Übung III (Teil2)

```
df[1, ] # erste Zeile, alle Spalten
```

```
    Name Gruppe Schuhgröße
1 Dennis    APC         42
```

```
df[1,3] # erste Zeile, dritte Spalte
```

```
[1] 42
```

```
df[,2] # alle Zeilen, zweite Spalte
```

```
[1] APC SIB IS  APC SIB SIB
Levels: APC IS SIB
```

```
df[,"Schuhgröße"] # Spalte nach Name
```

```
[1] 42 43 42 39 42 43
```

# Data Frame: Zugriff auf Inhalte II

Einführung in R

S. Trahasch, S. Niro

Einordnung

R

S, S-Plus, R

Vorteile von R

Nachteile von R

Rechnen mit R

Vektoren

Übung I

Übung II

Data Frame

Übung III (Teil1)

Einschub

Übung III (Teil2)

```
df$Name # Spalte nach Name II

[1] Dennis  Ralf    Stefan  Susanne Swen    Werner
Levels: Dennis Ralf Stefan Susanne Swen Werner

df[df$Schuhgröße < 41,]

     Name Gruppe Schuhgröße
4 Susanne    APC         39

df[df$Gruppe == "APC", "Name"]

[1] Dennis  Susanne
Levels: Dennis Ralf Stefan Susanne Swen Werner
```

```
str(df)

'data.frame': 6 obs. of  3 variables:
 $ Name      : Factor w/ 6 levels "Dennis","Ralf",..: 1 2 3 4 5 6
 $ Gruppe    : Factor w/ 3 levels "APC","IS","SIB": 1 3 2 1 3 3
 $ Schuhgröße: int  42 43 42 39 42 43
```

```
summary(df)

      Name    Gruppe    Schuhgröße
 Dennis :1   APC:2   Min.   :39.00
 Ralf   :1   IS :1   1st Qu.:42.00
 Stefan :1   SIB:3   Median :42.00
 Susanne:1           Mean   :41.83
 Swen   :1           3rd Qu.:42.75
 Werner :1           Max.   :43.00
```

# Übung III (Teil1): Packages und Data Frames

Einführung in R
S. Trahasch, S. Niro
Einordnung
R
S, S-Plus, R
Vorteile von R
Nachteile von R
Rechnen mit R
Vektoren
Übung I
Übung II
Data Frame
Übung III (Teil1)
Einschub
Übung III (Teil2)

- Installieren Sie die beiden packages **rpart** und **rpart.plot** (Hinweis: *Tools/Install Packages ...* in RStudio oder via Console mit *install.packages("PACKAGE_NAME")*)
- Laden Sie die beiden Packages mit *library(PACKAGE_NAME)*
- Laden Sie das Beispieldatenset **ptitanic** mit *data*(*ptitanic*)
- Untersuchen Sie das Data frame **ptitanic** mit den Funktionen **summary** und **str**
- Erstellen Sie einen Scatterplot mit **plot** (Einfärben der Datenpunkte mit dem Parameter

```
col = ifelse(ptitanic$survived==" survived", "green", "red")
```

# Übung III (Teil1 Lösung)

Einführung in R
S. Trahasch, S. Niro
Einordnung
R
S, S-Plus, R
Vorteile von R
Nachteile von R
Rechnen mit R
Vektoren
Übung I
Übung II
Data Frame
Übung III (Teil1)
Einschub
Übung III (Teil2)

```
library(rpart);library(rpart.plot);data(ptitanic); options(width=
str(ptitanic)

'data.frame': 1309 obs. of  6 variables:
 $ pclass  : Factor w/ 3 levels "1st","2nd","3rd": 1 1 1 1 1 1 1
 $ survived: Factor w/ 2 levels "died","survived": 2 2 1 1 1 2 2
 $ sex     : Factor w/ 2 levels "female","male": 1 2 1 2 1 2 1 2
 $ age     :Class 'labelled'  atomic [1:1309] 29 0.917 2 30 25 ..
  .. ..- attr(*, "units")= chr "Year"
  .. ..- attr(*, "label")= chr "Age"
 $ sibsp   :Class 'labelled'  atomic [1:1309] 0 1 1 1 1 0 1 0 2 0
  .. ..- attr(*, "label")= chr "Number of Siblings/Spouses Aboard
 $ parch   :Class 'labelled'  atomic [1:1309] 0 2 2 2 2 0 0 0 0 0
  .. ..- attr(*, "label")= chr "Number of Parents/Children Aboard
```

# Übung III ((Teil1 Lösung))

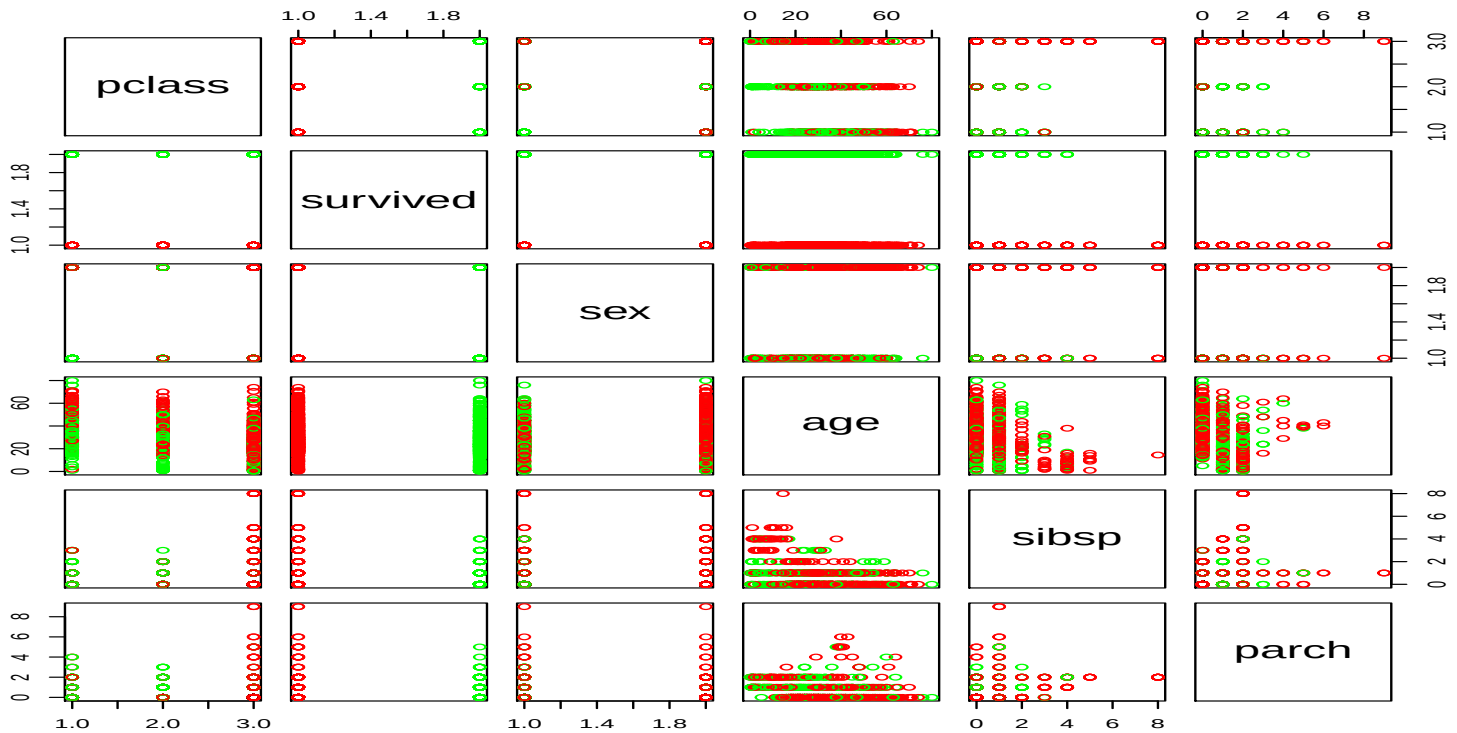
Einführung in R
S. Trahasch, S. Niro
Einordnung
R
S, S-Plus, R
Vorteile von R
Nachteile von R
Rechnen mit R
Vektoren
Übung I
Übung II
Data Frame
Übung III (Teil1)
Einschub
Übung III (Teil2)

```
summary(ptitanic)

 pclass        survived       sex           age
 1st:323   died    :809   female:466   Min.    : 0.1667
 2nd:277   survived:500   male  :843   1st Qu.:21.0000
 3rd:709                               Median :28.0000
                                       Mean   :29.8811
                                       3rd Qu.:39.0000
                                       Max.   :80.0000
                                       NA's   :263

     sibsp           parch
 Min.   :0.0000   Min.   :0.000
 1st Qu.:0.0000   1st Qu.:0.000
 Median :0.0000   Median :0.000
 Mean   :0.4989   Mean   :0.385
 3rd Qu.:1.0000   3rd Qu.:0.000
 Max.   :8.0000   Max.   :9.000
```

# Übung III (Teil1 Lösung)

Einführung in R
S. Trahasch, S. Niro
Einordnung
R
S, S-Plus, R
Vorteile von R
Nachteile von R
Rechnen mit R
Vektoren
Übung I
Übung II
Data Frame
Übung III (Teil1)
Einschub
Übung III (Teil2)

```
plot(ptitanic, col = ifelse(ptitanic$survived == "survived",
    "green", "red"))
```



# Einschub: Data-Mining I

Einführung in R
S. Trahasch, S. Niro
Einordnung
R
S, S-Plus, R
Vorteile von R
Nachteile von R
Rechnen mit R
Vektoren
Übung I
Übung II
Data Frame
Übung III (Teil1)
Einschub
Übung III (Teil2)

- Ziel: Unbekannte Zusammenhänge mithilfe von Algorithmen in den Daten finden
- Hier: Herausfinden was maßgeblich dafür war, ob ein Passagier die Katastrophe überlebt hat.

| $x_1$ | ... | $x_i$ | $y$ |
|---|---|---|---|
| 1 | .. | 1 | survived |
| 1 | .. | 1 | died |
| 2 | .. | 1 | survived |
| 1 | .. | 1 | ? |
| 1 | .. | 1 | ? |
| 2 | .. | 5 | ? |

Tabelle: Daten liegen in Tabelleform vor

# Einschub: Data-Mining II

Einführung in R
S. Trahasch, S. Niro
Einordnung
R
S, S-Plus, R
Vorteile von R
Nachteile von R
Rechnen mit R
Vektoren
Übung I
Übung II
Data Frame
Übung III (Teil1)
Einschub
Übung III (Teil2)

- Wir haben einen Data Frame
- Wir müssen dem Algorithmus (Funktion) zeigen was unabhängige Variablen $x$ sind und was die abhängige Variable $y$ ist
- Zwei Möglichkeiten:
  - Aufteilung des Data Frames in $x$ und $y$
  - Verwendung von R-Formeln und Übergabe des Datenframes:
    Formel (Prinzip):
    $y \sim x_1 + ... + x_i$
    $y$ und $x$ sind die Namen der Spalten im data frame

# Übung III (Teil2): Machine Learning from Disaster

Einführung in R
S. Trahasch, S. Niro
Einordnung
R
S, S-Plus, R
Vorteile von R
Nachteile von R
Rechnen mit R
Vektoren
Übung I
Übung II
Data Frame
Übung III (Teil1)
Einschub
Übung III (Teil2)

- Wir wollen rausfinden was maßgeblich dafür ist, ob ein Passagier überlebt hat oder gestorben ist
- Dazu erstellen Sie mit *rpart* einen Entscheidungsbaum. Wir betrachten nur die drei Attribute *sex*, *age*, und *pclass*:

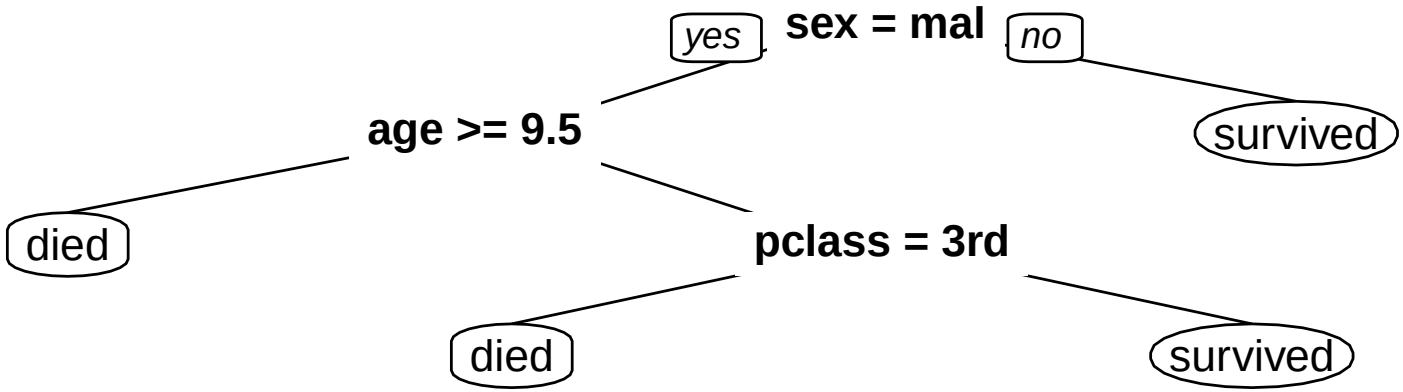
```
rtree <- rpart(survived ~ sex + age + pclass
               # entspricht  y ~ x1 + .. + x2

               , data = ptitanic) # data frame
```

- Zeichnen Sie den Entscheidungsbaum mit **prp**
- Hätten Sie überlebt?

# Übung III (Teil2): Machine Learning from Disaster (Lösung)

Einführung in R
S. Trahasch, S. Niro
Einordnung
R
S, S-Plus, R
Vorteile von R
Nachteile von R
Rechnen mit R
Vektoren
Übung I
Übung II
Data Frame
Übung III (Teil1)
Einschub
Übung III (Teil2)

```
rtree <- rpart(survived ~ sex + age + pclass
               ,data = ptitanic)
prp(rtree)
```



# Ende



Einordnung

R

S, S-Plus, R

Vorteile von R

Nachteile von R

Rechnen mit R

Vektoren

Übung I

Übung II

Data Frame

Übung III (Teil1)

Einschub

Übung III (Teil2)