

R für die Sozialwissenschaften - Teil 1

Jan-Philipp Kolb

04 August, 2017

Pluspunkte von R

- Als Weg kreativ zu sein ...
- Graphiken, Graphiken, Graphiken
- In Kombination mit anderen Programmen nutzbar
- Zur Verbindung von Datenstrukturen
- Zum Automatisieren
- Um die Intelligenz anderer Leute zu nutzen ;-)
- ...

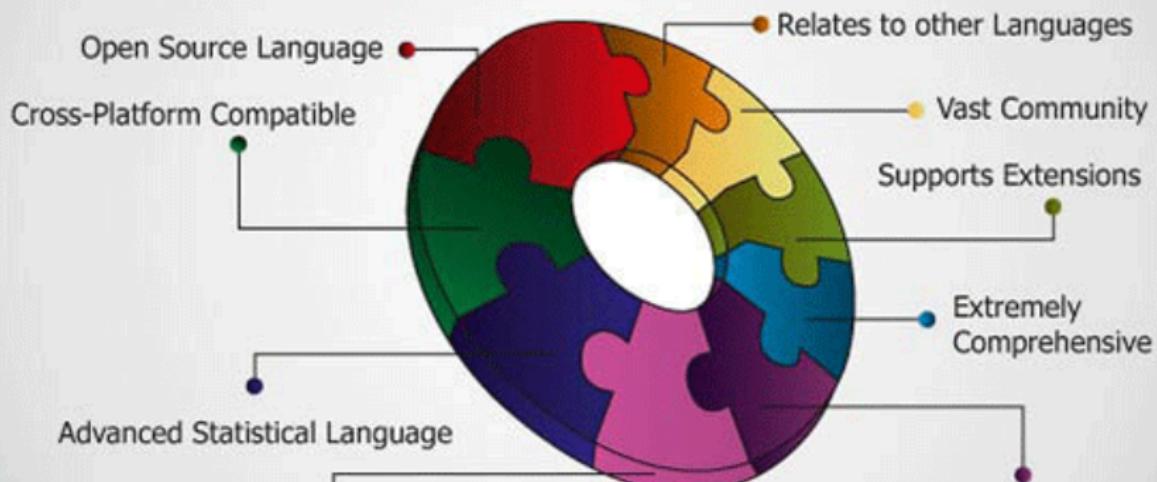
Einführung und Motivation

Gründe

- R ist frei verfügbar. Es kann umsonst runtergeladen werden.
- R ist eine Skriptsprache / Popularität von R

Why Learn R?

edureka!



Erste Schritte mit R

R ist eine Objekt-orientierte Sprache

Vektoren und Zuweisungen

- R ist eine Objekt-orientierte Sprache
- <- ist der Zuweisungsoperator (Shortcut: "Alt" + "-")

```
b <- c(1,2) # erzeugt ein Objekt mit den Zahlen 1 und 2
```

- Eine Funktion kann auf dieses Objekt angewendet werden:

```
mean(b) # berechnet den Mittelwert
```

```
## [1] 1.5
```

Mit den folgenden Funktionen können wir etwas über die Eigenschaften des Objekts lernen:

```
length(b) # b hat die Länge 2
```

Wie bekommt man Hilfe?

Wie bekommt man Hilfe?

- Um generell Hilfe zu bekommen:

`help.start()`

- Online Dokumentation für die meisten Funktionen:

`help(name)`

- Nutze `? name` um Hilfe zu bekommen.

`?mean`

- `example(lm)` gibt ein Beispiel für die lineare Regression

`example(lm)`

Vignetten

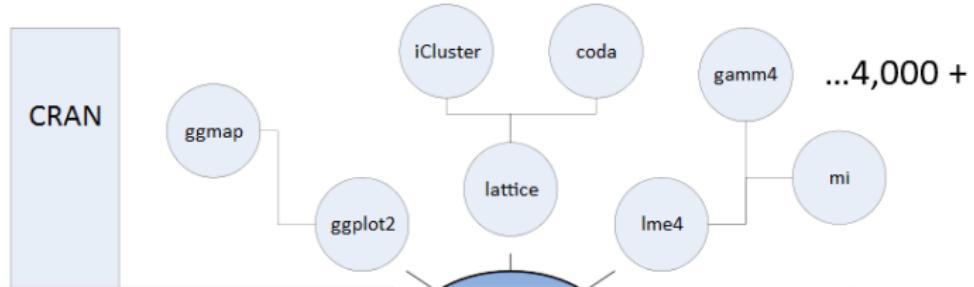
- Dokumente zur Veranschaulichung und Erläuterung von Funktionen im

Modularer Aufbau

Wo sind die Funktionen enthalten

- Viele Funktionen sind im Basis-R enthalten
- Viele spezifische Funktionen sind in zusätzlichen Bibliotheken integriert
- R kann modular erweitert werden durch sog. packages bzw. libraries
- Auf CRAN werden die wichtigsten packages gehostet (im Moment 11020)
- Mehr Pakete (v.a. Biostatistik, Medizin) finden sich z.B. bei bioconductor

Übersicht R-Pakete



Datenimport

Datenimport



SPSS®

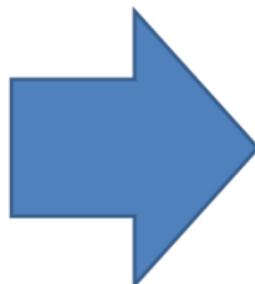
STATA®

dBASE™

Minitab ▶™



MySQL™



Datenaufbereitung

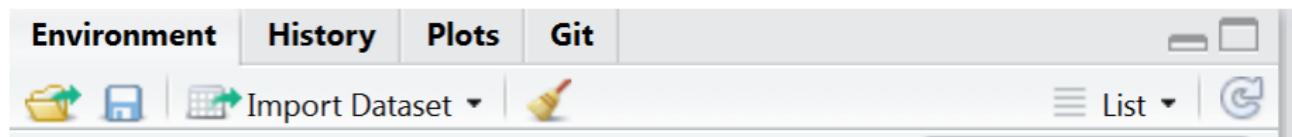
Data Frames

Beispieldaten einlesen:

```
library(foreign)
dat<-read.dta("https://github.com/Japhilko/RSocialScience/
               raw/master/data/GPanel.dta")
```

- Auf dem Github Verzeichnis liegt eine verkleinerte Version des Campus Files.
- Alle Operationen sollten aber auch mit dem größeren Datensatz funktionieren

Übersicht mittels Rstudio



Datenexport

Die Exportformate von R

- In R werden offene Dateiformate bevorzugt
- Genauso wie `read.X()` Funktionen stehen viele `write.X()` Funktionen zur Verfügung
- Das eigene Format von R sind sog. Workspaces (`.RData`)

Beispieldatensatz erzeugen

```
A <- c(1,2,3,4)  
B <- c("A","B","C","D")
```

```
mydata <- data.frame(A,B)
```

	A	B
1	A	
2	B	

Basisgraphiken

Ein Plot sagt mehr als 1000 Worte

- Grafisch gestützte Datenanalyse ist toll
- Gute Plots können zu einem besseren Verständnis beitragen
- Einen Plot zu generieren geht schnell
- Einen guten Plot zu machen kann sehr lange dauern
- Mit R Plots zu generieren macht Spaß
- Mit R erstellte Plots haben hohe Qualität
- Fast jeder Plottyp wird von R unterstützt
- R kennt eine große Menge an Exportformaten für Grafiken

Plot ist nicht gleich Plot

- Bereits das base Package bringt eine große Menge von Plot Funktionen mit
- Das lattice Packet erweitert dessen Funktionalität
- Eine weit über diese Einführung hinausgehende Übersicht findet sich in Murrell, P (2006): R Graphics.

Datenanalyse

Den Datensatz laden

```
## Warning: NAs durch Umwandlung erzeugt
```

```
library(foreign)
dat <- read.dta(
  "https://github.com/Japhilko/RSocialScience/blob/master/data/
  GPanel.dta?raw=true")
dat$bazq020a <- as.numeric(dat$bazq020a)
```

Streuungsmaße

- Varianz: `var()`
- Standardabweichung: `sd()`
- Minimum und Maximum: `min()` und `max()`
- Range: `range()`

```
var(dat$bazq020a)
```

Grafiken und Zusammenhang

Die Daten laden

```
library(foreign)
dat <- read.dta("https://github.com/Japhilko/RSocialScience/
blob/master/data/GPanel.dta?raw=true")
```

Eine Kreuztabelle erstellen

```
Beruf_Gefordert <- dat$a11c109a
Beruf_Anerkannt <- dat$a11c111a
```

```
table(Beruf_Gefordert,Beruf_Anerkannt)
```

```
##                                     Beruf_Anerkannt
## Beruf_Gefordert      Missing by design Ja Nein Weiß nicht
##   Missing by design                               93  0    0     0
##   Ja                                         0  7    0     0
##   Nein                                         0  0    0     0
##   Weiß                                         0  0    0     0
##   nicht                                         0  0    0     0
```

Das lattice Paket

Intro lattice-Paket

It is designed to meet most typical graphics needs with minimal tuning, but can also be easily extended to handle most nonstandard requirements.

[http://stat.ethz.ch/R-manual/R-devel/library/lattice/html/
Lattice.html](http://stat.ethz.ch/R-manual/R-devel/library/lattice/html/Lattice.html)

Der Datensatz - Scores on A-level Chemistry in 1997

```
library("mlmRev")
data(Chem97)
```

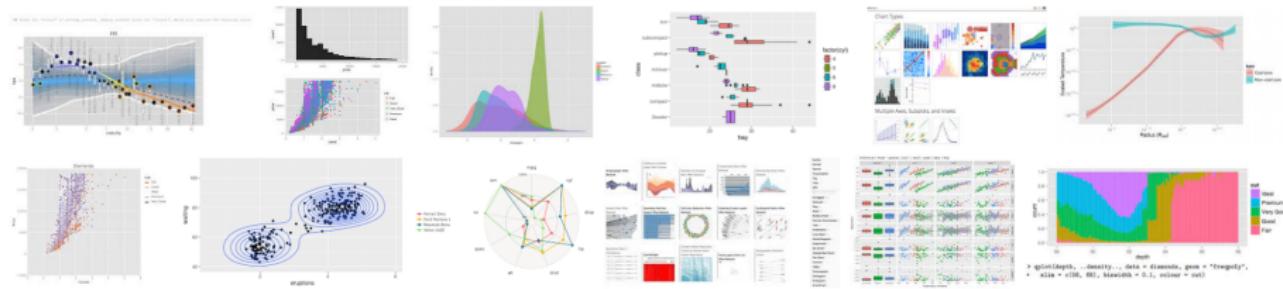
variables	categories
-----------	------------

lea	Local Education Authority
school	School identifier
student	Student identifier

Die Pakete `ggplot2` und `ggmap`

Das Paket `ggplot2`

- Entwickelt von Hadley Wickham
- Viele Informationen unter:
<http://ggplot2.org/>
- Den Graphiken liegt eine eigene Grammatik zu Grunde



Das Paket `ggplot2` installieren und laden

- Basiseinführung `ggplot2`

```
install.packages("ggplot2")
```

Die lineare Regression

Literatur - lineare Regression

Maindonald - Data Analysis

- Einführung in R
- Datenanalyse
- Statistische Modelle
- Inferenzkonzepte
- Regression mit einem Prädiktor
- Multiple lineare Regression
- Ausweitung des linearen Modells
- ...

Lineare Regression in R - Beispieldatensatz

John H. Maindonald and W. John Braun

DAAG - Data Analysis and Graphics Data and Functions

`install.packages("DAAG")`

Die logistische Regression

Agresti - Categorical Data Analysis (2002)



- Sehr intuitiv geschriebenes Buch
- Sehr ausführliches begleitendes Skript von Thompson
- Das Skript eignet sich um die kategoriale Datenanalyse nachzuvollziehen

Mehrebenenmodelle

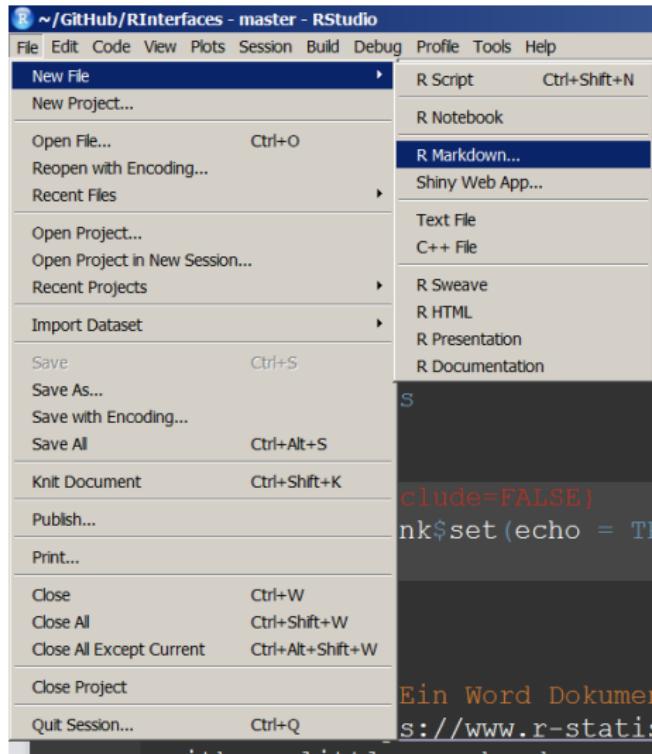
Wie sehen die Daten aus?

- Beispiel Mehrebenenstruktur der Daten



Word Dokumente mit R erstellen

Ein Markdown Dokument mit Rstudio erzeugen



PDF Dokumente und Präsentationen mit LaTeX, Beamer und Sweave

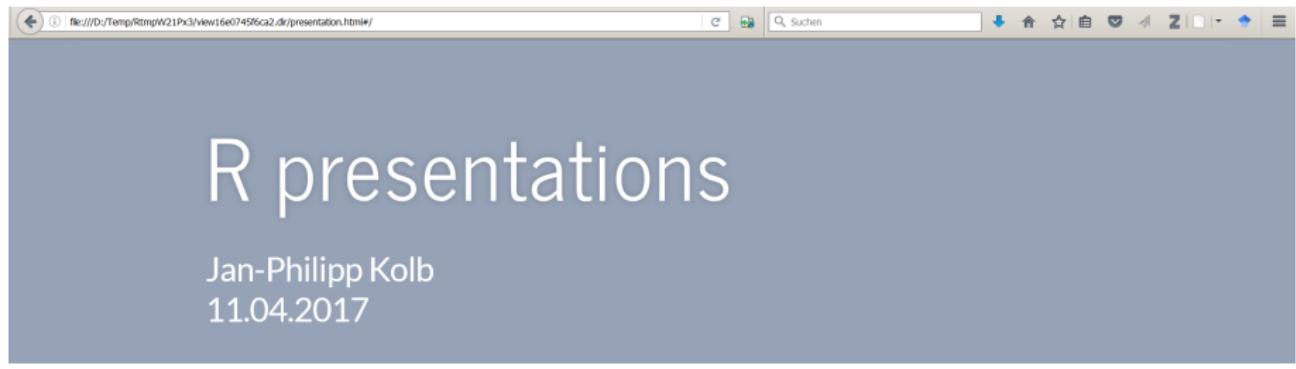
Präsentationen mit Rmarkdown - beamer Präsentationen

Import csv

```
url <- "https://raw.githubusercontent.com/Japhilko/  
GeoData/master/2015/data/whcSites.csv"  
  
whcSites <- read.csv(url)
```

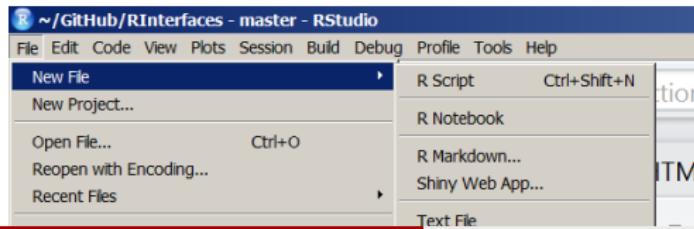
HTML Dokumente, Präsentationen und Dashboards mit Rmarkdown

Präsentationen - Rpres der einfachste Weg



The screenshot shows a web browser window displaying an R presentation. The URL in the address bar is `file:///D:/Temp/RtmpW21Px3/view16e0745f6ca2.dir/presentation.html#`. The main content of the slide is "R presentations" in large white font, followed by the author's name "Jan-Philipp Kolb" and the date "11.04.2017".

Eine erste Präsentation



Eine ioslides Präsentation

Eine ioslides Präsentation

The screenshot shows a web browser window with a white slide against a black background. The slide contains the following text:

Präsentationen mit R und
Rstudio

Jan-Philipp Kolb
11 April 2017

The browser's address bar shows the path: file:///C:/Users/kolbjp/Documents/GitHub/RInterfaces/slides/R2pdf.html#1. The top navigation bar includes standard icons for back, forward, search, and refresh.

Eine slidy Präsentation

slidy Präsentationen



Präsentationen mit R und Rstudio

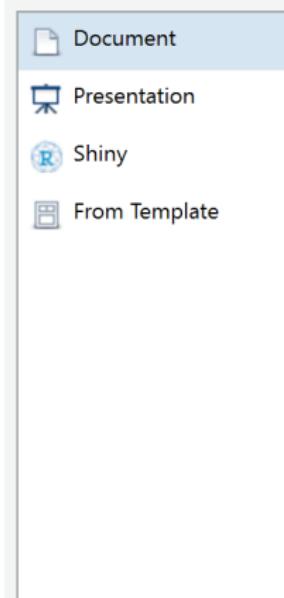
Jan-Philipp Kolb

11 April 2017

HTML Dokumente

Ein HTML Dokument erzeugen

New R Markdown



Title: Neues HTML Dokument

Author: Jan-Philipp Kolb

Default Output Format:

HTML

Recommended format for authoring (you can switch to PDF or Word output anytime).

PDF

PDF output requires TeX (MiKTeX on Windows, MacTeX 2013+ on OS X, TeX Live 2013+ on Linux).

Word

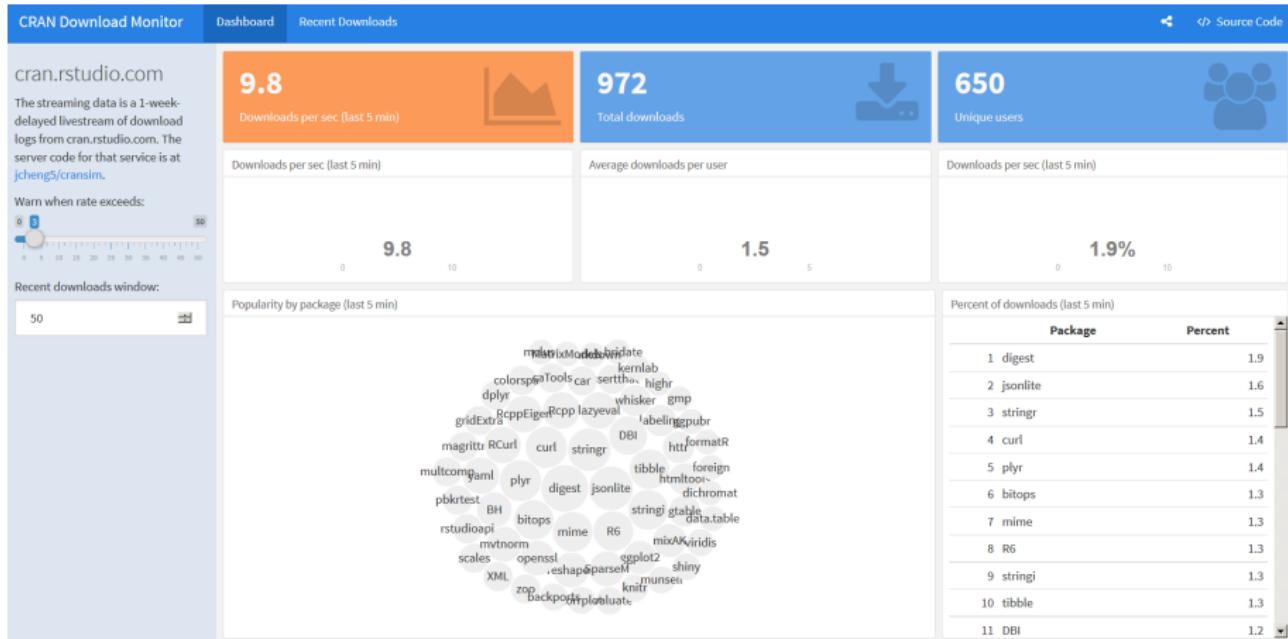
Previewing Word documents requires an installation of MS Word (or Libre/Open Office on Linux).

OK

Cancel

Dashboards

Beispiel R-Pakete



Paket installieren

Notebooks zur Integration von anderen Programmiersprachen (Python,LaTeX,Julia)

Notebooks

- Warum R Notebook nutzen

The screenshot shows the RStudio interface with an R notebook file named "nb-demo.Rmd". The code editor pane contains the following R code:

```
9
10 `r` summary(iris)
11 ...
12
13 Sepal.Length Sepal.Width Petal.Length Petal.Width Species
14 Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100 setosa :50
15 1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300 versicolor:50
16 Median :5.800 Median :3.000 Median :4.350 Median :1.300 virginica :50
17 Mean :5.843 Mean :3.057 Mean :4.358 Mean :1.199
18 3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800
19 Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500
20
21 `r` library(ggplot2)
22 ggplot(Sepal.Length, Petal.Length, data = iris, color = Species, size =
23 Petal.Width)
24 ...
25
```

The code uses the `summary` function to print descriptive statistics for the Iris dataset, and then generates a scatter plot using ggplot2. The plot shows Petal Length on the y-axis versus Sepal Length on the x-axis. Data points are colored by species (setosa in red, versicolor in green) and sized by Petal Width. A legend on the right side indicates that point size corresponds to Petal Width values of 0.5, 1.0, 1.5, 2.0, and 2.5.

Rnotebooks

Ein Rnotebook anlegen

The screenshot shows the RStudio interface with the following details:

- File Menu:** Shows options like New File, New Project..., Open File..., Save, Print, Close, and Quit Session...
A dropdown menu is open under "New File" showing "R Notebook".
- Code Editor:** Displays R code:

```
tation:  
beaver  
structurebold  
tango  
ridgeUS  
: false  
clude=FALSE)  
ink.set(echo = TRUE)
```
- Toolbar:** Includes Insert, Run, and other common tools.
- Environment Tab:** Shows "R presentations (1/4)"
- Plots Tab:** Shows a slide titled "R presentations" by Jan-Philipp Kolb from 11.04.2017.
- Git Tab:** Shows the repository structure: GitHub > RInterfaces > slides.
- Presentation Tab:** Shows a file list with the following files:

Name	Size
presentHTML.md	1.1 KB
presentHTML.Rmd	857 B
presentHTML_files	
R2pdf.html	292.3 KB
R2pdf.pdf	775 KB
R2pdf.Rmd	1.1 KB
rcpp.html	692.7 KB
rcpp.md	372 B
rcpp.Rmd	862 B
Rexcel.html	692.5 KB
- Console Tab:** Shows the command "R Markdown" and the status "18:4 (Top Level)".
- Task View:** Shows various R packages and tools available.
- Bottom Taskbar:** Shows the Windows taskbar with multiple open applications.

Andere Notebooks

Jupyter Notebook

- Anaconda installieren
- folgenden Befehl in die Eingabeaufforderung eingeben
- Bei Windows findet man diese, wenn man cmd in Suche eingibt.

```
jupyter notebook
```

Start Jupyter Notebook

Jupyter Notebook im Intranet

<http://intranet.gesis.intra/AGs/iedi/Seiten/Jupyter.aspx>
jupyter

The screenshot shows the Jupyter Notebook interface. At the top, there are three tabs: 'Files' (selected), 'Running', and 'Clusters'. Below the tabs, a message says 'Select items to perform actions on them.' On the right side, there are buttons for 'Upload', 'New', and a refresh icon. The main area is a file browser with a sidebar. The sidebar contains two entries: 'Anaconda3' and 'AppData'. The main area shows a single entry: 'Untitled.ipynb'.

Beaker Notebook

Beaker Notebook

- Auch bei Beaker kann man R-code einbauen



Beaker starten

- Beaker installieren ...
- ... und mit `beaker.command.bat` starten

Interaktive Tabellen mit DataTables

The R-package DT

- DT: An R interface to the DataTables library

```
install.packages('DT')
```

```
library('DT')
```

```
exdat <- read.csv("data/exdat.csv")
```

```
datatable(exdat)
```

Beispiel für interaktive Tabelle

Hier ist das Ergebnis - Beispiel für eine interaktive Tabelle

The screenshot shows a web browser displaying an Rpubs page. At the top, there are navigation icons for back, forward, and search, along with the URL rpubs.com/Japhilko82/osmplzbe. Below the header, the Rpubs logo is visible, followed by the text "brought to you by RStudio". The main content area contains an interactive table generated by the DT package. The table has columns for PLZ99, PLZORT99, area_d, bakery, bar, cafe, and clothes. The first row shows data for PLZ 680 (Berlin (östl. Stadtbezirke)) with values 0.000405556447886959, 29, 36, 55, and 41 respectively. The last row is a footer for Jan-Philip Kolb with the text "R für die Sozialwissenschaften - Teil 1" and the date "04 August, 2017". At the bottom left, there is a dropdown menu for "Show 10 entries" and a search bar. On the right side of the table, there are sorting and filtering icons for each column.

PLZ99	PLZORT99	area_d	bakery	bar	cafe	clothes
680	Berlin (östl. Stadtbezirke)	0.000405556447886959	29	36	55	41
Jan-Philip Kolb	R für die Sozialwissenschaften - Teil 1					

Interaktive Karten mit dem Javascript Paket leaflet

Die Daten - Weltkulturerbe

- die Daten einlesen:

```
url <- "https://raw.githubusercontent.com/Japhilko/  
GeoData/master/2015/data/whcSites.csv"
```

```
whcSites <- read.csv(url)
```

- die Daten werden eingeschränkt:

```
whcSitesDat <- with(whcSites,data.frame(name_en,  
category))
```

Eine Tabelle erzeugen mit knitr

```
library(knitr)  
kable(head(whcSitesDat))
```