Datenaufbereitung

Jan-Philipp Kolb

07 April 2016

Die Daten editieren

```
load("data/refugeeTab.RData")
```

```
mean(refugeeTab[,2])
```

```
## Warning in mean.default(refugeeTab[, 2]): argument is no
## logical: returning NA
```

```
## [1] NA
```

um dies zu ändern ist ein wenig Kosmetik notwendig:

```
refugeeTab[,2] <- as.numeric(refugeeTab[,2])
```

```
mean(refugeeTab[,2])
```

```
## [1] 17.16129
```

Die weiteren Spalten bearbeiten

- In R wird der Punkt als Dezimaltrennzeichen verwendet.
- Wenn ein Komma im Ausdruck ist, wird der Eintrag als `character` behandelt
- dann kann bspw. kein Mittelwert berechnet werden

```
refugeeTab[,3] <- gsub(",",".",refugeeTab[,3])
refugeeTab[,3] <- as.numeric(refugeeTab[,3])
```

Erste Spalte bearbeiten

```
ab <- as.character(refugeeTab[,1])
info <- round(nchar(ab)/2)
Namen <- substr(ab,1,info)
Namen[1:29] <- gsub(" ","",Namen[1:29])
Namen[31]  <- "Zypern"
refugeeTab[,1] <- Namen
```

Spaltennamen verändern

```
colnames(refugeeTab) <- c("Land","2015",
                          "pro_tsd_Einwohner")
```

Das Ergebnis

	Land	2015	pro_tsd_Einwohner
3	Bulgarien	14	2.83
4	Danemark	15	3.70
5	Deutschland	29	5.87
6	Estland	16	0.18
7	Finnland	20	5.91
8	Frankreich	30	1.14

Das Editing ist also aufwändiger als das eigentliche Scraping

CO2 Verbrauch

```
link <- "https://en.wikipedia.org/wiki/
List_of_countries_by_carbon_dioxide_
emissions_per_capita"
```

```
link_data <- read_html(link)
doc <- htmlParse(link_data)
tab <- readHTMLTable(doc)
```

```
str(tab)
tab[[1]]
tab[[2]]
```

Auf die Daten schauen

```
tab2 <- tab[[2]]
head(tab2)
```

Daten speichern

```
write.csv(tab2,file="CO2emissions.csv")
```