WZB

Wissenschaftszentrum Berlin
für Sozialforschung

Einführung in die Quantitative Datenanalyse

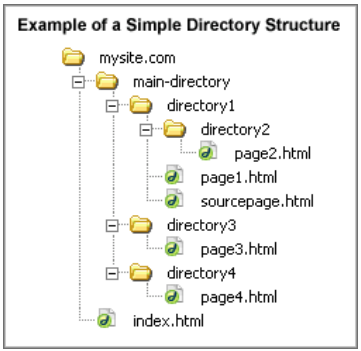
Sitzung 6: Datenquellen und Datenmanipulation

Proseminar an der Freien Universität Berlin
12.06.2017 - Marcus Spittler

Inhalt der 6. Sitzung

- **Grundlagen**
  - Pfadangaben
  - R Projekte
  - Dateiformate
- **Datenquellen**
  - Woher sind Daten zu beziehen?
  - Codebücher und Methodenreports
- **Datenmanipulation**
  - `mutate`-Befehl
  - `ifelse`-Befehl
  - Wie mit Faktoren umgehen

Pfadangaben



Pfadangaben

- **R** verfügt über eine Reihe Funktionen mit denen wir Dateien importieren oder exportieren können. Dazu nutzen wir **Pfadangaben** zu Verzeichnissen und Ordnern
- Pfadangaben können entweder
  - **absolut** sein, wie z.B.: `C://Users/meinName/RProjekt/daten/`
  - **relativ** sein, wie z.B.: `./RProjekt/daten/`
- Unter Windows wird normalerweise ein *back slash* `\` als Verzeichnisstrenner verwendet, in **R** kommt wie in MacOS or Linux der *forward slash* `/` zum Einsatz.

Pfadangaben

- Der aktuelle Arbeitspfad von **R** lässt sich mit `getwd()` (steht für *get working directory*) anzeigen.

```
getwd()
```

```
## [1] "C:/Users/spittler/GitProjects/CourseDataAnalysis"
```

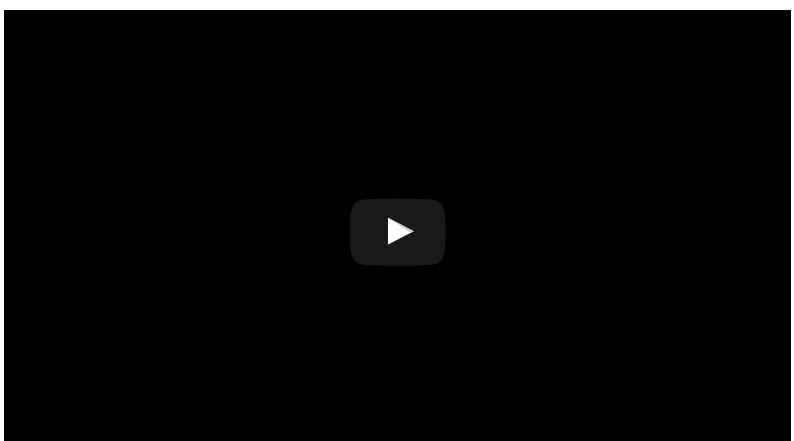
- Mit `setwd("C://Order/NochEinOrdner/")` kann der Arbeitspfad geändert werden.

- Innerhalb des Arbeitspfads können Dateien mit *relativen Pfadangaben* aufgerufen werden

```
daten <- read_csv("./OrdnermitDateien/dateiname.csv")
```

R Projekte

- *R Projekte* bestehen aus einem Ordner und einer Projektdatei **".Rproj"**
- Mit Projekten können wir in einer abgeschlossenen Entwicklungsumgebung arbeiten
- Pfadangaben können wir immer *relativ* zum Ordnerpfad des Projektes angeben



Dateiformate

- **Dateien** bestehen aus einem **Dateinamen** und haben ein **Dateiformat**. Das Format wird über die **Dateiendung** angegeben (z.B. **".mp3, .docx, .pdf"**).

- Die Kenntnis des Dateiformats ist essentiell für die Interpretation der in einer Datei abgelegten Information. Betriebssysteme ordnen Dateien über das Dateiformat Anwendungen zu, die die Dateien interpretieren können.

Dateiformate in der Datenanalyse

Format	Beschreibung	R-Befehl	Paket
.RData	R Datendatei	`load("daten.RData")`	-
.csv	Comma seperated values	`read_csv("daten.csv")`	`foreign`
	Deutsche **csv's**	`read_csv2("daten.csv")`	`foreign`
.xls/x	Excel Dateien	`read_excel("daten.xlsx")`	`readxl`
.dta	STATA Dateien	`read_dta("daten.dta")`	`haven`
.sav	SPSS Dateien	`read_sav("daten.sav")`	`haven`
.sas/7	SAS Dateien	`read_sas("daten.sas7")`	`haven`

- Deutsche **csv's** werden mit Semikolon getrennt und müssen anders geöffnet werden.
- Daten können auch bequem über das GUI von RStudio eingelesen werden

Datenquellen

- Quality of Government
  - Aggregatdatensatz auf Land/Jahr-Basis der eine Vielzahl kleinerer Datensätze vereint
- GESIS bzw. https://dbk.gesis.org/dbksearch/
  - Das Archiv der Leibniz-Gemeinschaft, sehr gut für u.a. Bevölkerungsbefragungen
  - z.B.: ALLBUS, Eurobarometer, EVS, ISSP, GLES, Politbarometer
- **Statistisches Bundesamt** und Landesämter z.B. Berlin
- World Value Survey
  - Weltweit vergleichende Studie zu Werten und Einstellungen, Demokratiezufriedenheit etc.

Datenquellen

- V-Dem: Global Standards, Local Knowledge
  - Expertendatensatz mit Demokratieindizes
- Eurostat
  - Vor allem ökonomische Daten zur EU-Staaten
  - Auch Zugriff über ein R-Paket möglich
- TwitteR
  - Möglichkeit Tweets direkt mit R aus Twitter auszulesen und zu analysieren. Die Analyse erfolgt z.B. mit dem Paket quanteda

Weitere wichtige Dokumente

- Fragebogen
- Codebücher
- Methodenreport

Datenmanipulation

Dem Fragenbogen haben wir entnommen, dass die Antwort auf die Frage nach Geschlecht der Befragten in der Variable **`d10`** gespeichert ist. Unser Datensatz hat den Namen **`E`**. Das `$`-Zeichen trennt Datensatz und Datenspalte.

Mit `table()` lassen wir uns die abs. Häufigkeit anzeigen. `str()` gibt uns Informationen zum Datentyp.

```
table(E$d10)
```

```
##
##   1   2
## 814 834
```

```
str(E$d10)
```

```
## Class 'labelled'  atomic [1:1648] 1 1 1 1 1 1 2 1 1 1 ...
##   ..- attr(*, "label")= chr "D10 Gender."
##   ..- attr(*, "format.stata")= chr "%8.0g"
##   ..- attr(*, "labels")= Named num [1:2] 1 2
##   .. ..- attr(*, "names")= chr [1:2] "Male" "Female"
```

Datenmanipulation

Da wir den Datensatz mit dem `haven`-Paket geladen haben (empfohlen!), können wir die numerischen Werte von Geschlecht wieder in einen Faktor umwandeln. Dies geht mit `as_factor()`.

Wie immer gibt es mehr als eine Möglichkeit dies zu tun:

```
E$gender <- as_factor(E$d10)
table(E$gender)
```

```
##
##   Male Female
##    814    834
```

Alternative:

```
E <- E %>% mutate(
  gender = as_factor(d10)
)
```

Datenmanipulation

Mit `head()` können die ersten sechs Zeilen des Datensatzes ausgegeben werden.

```
E <- E %>% mutate(
  gender = as_factor(d10)
)
head(E)
```

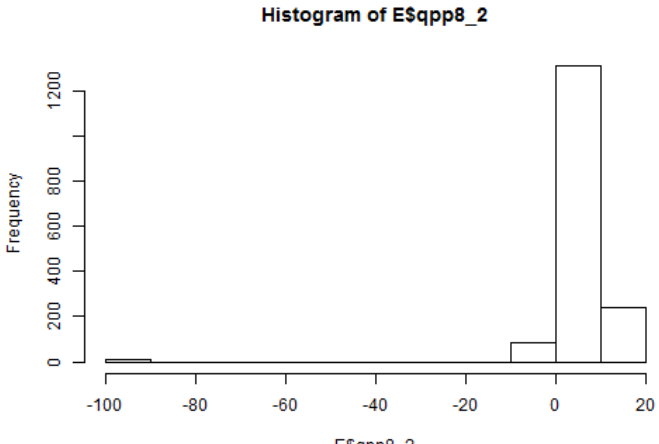
```
## # A tibble: 6 x 3
##         d10    qpp8_2 gender
##   <dbl+lbl> <dbl+lbl> <fctr>
## 1         1        11   Male
## 2         1        10   Male
## 3         1         3   Male
## 4         1         5   Male
## 5         1         9   Male
## 6         1         4   Male
```

In der Variable `qpp8_2` ist die **Wahlwahrscheinlichkeit der SPD** gespeichert. Wieder nutzen wir `table()` für einen Überblick. Mit `hist()` erstellen wir ein sehr einfaches Histogramm.

```
table(E$qpp8_2)
```

```
##
## -99  -9  -8   1   2   3   4   5   6   7   8   9  10  11
##  11  33  52 281  45  83 112 121 226 102 117 141  81 243
```

```
hist(E$qpp8_2)
```



ifelse()

Der Datendokumentation entnehmen wir, dass negative Werte **fehlende Werte/Missings** repräsentieren. Diese müssen wir von der Analyse ausschließen. Dazu nutzen wir den `ifelse()`-Befehl.

`ifelse()` funktioniert nach folgendem Schema:

```
ifelse( BEDINGUNG, falls WAHR, falls FALSCH )
```

- Eine Bedingung ist etwa:
  - `x < 2` = Variable ist kleiner 2
  - `x == 1` = Variable ist gleich 1
  - `x != 5` = Variable ist ungleich 5
  - `is.na(x)` = Variable ist fehlend
  - `x %in% c("Deutschland", "Schweiz")` = Variable ist ein Text mit dem Inhalt Deutschland oder Schweiz

ifelse()

```
E <- E %>% mutate(
  ptv.spd = ifelse(qpp8_2 < 0, NA, qpp8_2)
)
set.seed(50);sample_n(E,6)
```

```
## # A tibble: 6 x 4
##         d10    qpp8_2 gender ptv.spd
##   <dbl+lbl> <dbl+lbl> <fctr>   <dbl>
## 1         2         4 Female       4
## 2         1        -8   Male      NA
## 3         1         7   Male       7
## 4         2         5 Female       5
## 5         2         4 Female       4
## 6         1         1   Male       1
```

European Parliament Election Study 2014

- Voter Study, First Post-Election Survey
- Grundgesamtheit: Wohnbevölkerung ab 18 Jahren der jeweiligen 28 EU-Mitgliedsstaaten, die die Staatsbürgerschaft besitzen oder Staatsbürger eines anderen EU-Mitgliedsstaates sind (Österreich: ab 16 Jahren).Befragte mussten ausreichende Sprachkenntnisse besitzen, um die Fragen in einer der jeweiligen Landessprachen zu beantworten.
- https://dbk.gesis.org/dbksearch/SDesc2.asp?DB=D&no=5160
- Download für den Kurs: https://www.dropbox.com/sh/qkoniyemrsbscz4/AAABHBT2jjedhYamFhBaLUzBa?dl=0