## Daten laden und speichern

Dirk Seidensticker/Clemens Schmid

7. Juli 2017

## Daten aus dem Dateisystem laden

### Daten aus dem Dateisystem laden

```
df <- read.csv("../data/AtlantPottery.csv")</pre>
```

bei csv evtl. zu beachten:

- sep = identifiziert das Trennzeichen
  - ":" Semikolon
  - "\t" Tabulator
- dec = "," identifiziert das Komma als Dezimalzeichen
- row.names = 1 definiert erste Spalte als Zeilennamen (index)
- encoding = 'latin1' Zeichenkodierung für mit Windows-Programmen geschriebene Dateien (siehe auch 'utf8')

### Daten aus dem Dateisystem laden

head(df)

	site	feature	e obje	ect	clas	ssher	d qty	w.	t size	wall	muendun	gsD	m
1	C	surface	9	54	]	(	G 1	1448	3 500	7	1	7.5	
2	C	surface	9	36	]	ζ Ι	R 1	. 50	3 120	) 5	1	7.0	
3	C	surface	9	32	]	ζ Ι	R 1	. 9:	2 120	) 9	2	0.0	
4	C	surface	9	26	]	ζ Ι	R 1	. 4	5 70	11	2	2.5	
5	E	surface	9	86	]	ζ (	G 1	129	3 500	8	1	8.5	
6	Ε	surface	9	63	]	(	G 1	. 66	7 200	) NA		6.0	
	${\tt minD}$	${\tt minD\_H}$	${\tt maxD}$	max	D_H 1	oodenD	temp	erSi	ze ves	selSh	ape		
1	16.0	16.5	21.0	1	11.0	7.5			M		3		
2	14.5	NA	16.0		NA	NA			M		7		
3	18.5	NA	24.5		NA	NA		,	VF		7		
4	NA	NA	27.0		NA	NA		,	VF		8		
5	22.5	9.5	22.5		9.5	0.0		1	VC		8		
6	5.5	11.5	10.5		5.5	6.0			C		1		
	2 3 4	1 C 2 C 3 C 4 C 5 E 6 E minD 1 16.0 2 14.5 3 18.5 4 NA 5 22.5	1 C surface 2 C surface 3 C surface 4 C surface 5 E surface 6 E surface minD minD_H 1 16.0 16.5 2 14.5 NA 3 18.5 NA 4 NA NA 5 22.5 9.5	1 C surface 2 C surface 3 C surface 4 C surface 5 E surface 6 E surface minD minD_H maxD 1 16.0 16.5 21.0 2 14.5 NA 16.0 3 18.5 NA 24.5 4 NA NA 27.0 5 22.5 9.5 22.5	1 C surface 54 2 C surface 36 3 C surface 32 4 C surface 26 5 E surface 86 6 E surface 63 minD minD_H maxD max 1 16.0 16.5 21.0 1 2 14.5 NA 16.0 3 18.5 NA 24.5 4 NA NA 27.0 5 22.5 9.5 22.5	1 C surface 54 R 2 C surface 36 R 3 C surface 32 R 4 C surface 26 R 5 E surface 86 R 6 E surface 63 R minD minD_H maxD maxD_H R 1 16.0 16.5 21.0 11.0 R 2 14.5 NA 16.0 NA 3 18.5 NA 24.5 NA 4 NA NA 27.0 NA 6 22.5 9.5 22.5 9.5	1 C surface 54 K C 2 C surface 36 K H 3 C surface 32 K H 4 C surface 26 K H 5 E surface 86 K C 6 E surface 63 K C minD minD_H maxD maxD_H bodenD 1 16.0 16.5 21.0 11.0 7.5 2 14.5 NA 16.0 NA NA 3 18.5 NA 24.5 NA NA 4 NA NA 27.0 NA NA 5 22.5 9.5 22.5 9.5 0.0	1 C surface 54 K G 1 2 C surface 36 K R 1 3 C surface 32 K R 1 4 C surface 26 K R 1 5 E surface 86 K G 1 6 E surface 63 K G 1 minD minD_H maxD maxD_H bodenD temp 1 16.0 16.5 21.0 11.0 7.5 2 14.5 NA 16.0 NA NA 3 18.5 NA 24.5 NA NA 4 NA NA 27.0 NA NA 5 22.5 9.5 22.5 9.5 0.0	1 C surface 54 K G 1 1444 2 C surface 36 K R 1 56 3 C surface 32 K R 1 93 4 C surface 26 K R 1 93 5 E surface 86 K G 1 1298 6 E surface 63 K G 1 66 minD minD_H maxD maxD_H bodenD temperSiz 1 16.0 16.5 21.0 11.0 7.5 2 14.5 NA 16.0 NA NA 3 18.5 NA 24.5 NA NA NA 5 22.5 9.5 22.5 9.5 0.0	1 C surface 54 K G 1 1448 500 2 C surface 36 K R 1 56 120 3 C surface 32 K R 1 92 120 4 C surface 26 K R 1 45 70 5 E surface 86 K G 1 1298 500 6 E surface 63 K G 1 667 200 minD minD_H maxD maxD_H bodenD temperSize ves 1 16.0 16.5 21.0 11.0 7.5 M 2 14.5 NA 16.0 NA NA M 3 18.5 NA 24.5 NA NA VF 4 NA NA 27.0 NA NA VF 5 22.5 9.5 22.5 9.5 0.0 VC	1 C surface 54 K G 1 1448 500 7 2 C surface 36 K R 1 56 120 5 3 C surface 32 K R 1 92 120 9 4 C surface 26 K R 1 45 70 11 5 E surface 86 K G 1 1298 500 8 6 E surface 63 K G 1 667 200 NA minD minD_H maxD maxD_H bodenD temperSize vesselSha 1 16.0 16.5 21.0 11.0 7.5 M 2 14.5 NA 16.0 NA NA NA WF 4 NA NA 27.0 NA NA NA VF 5 22.5 9.5 22.5 9.5 0.0 VC	1 C surface 54 K G 1 1448 500 7 1 2 C surface 36 K R 1 56 120 5 1 3 C surface 32 K R 1 92 120 9 2 4 C surface 26 K R 1 45 70 11 2 5 E surface 86 K G 1 1298 500 8 1 6 E surface 63 K G 1 667 200 NA minD minD_H maxD maxD_H bodenD temperSize vesselShape 1 16.0 16.5 21.0 11.0 7.5 M 3 2 14.5 NA 16.0 NA NA NA WF 7 3 18.5 NA 24.5 NA NA NA VF 7 4 NA NA 27.0 NA NA NA VF 8 5 22.5 9.5 22.5 9.5 0.0 VC 8	2 C surface 36 K R 1 56 120 5 17.0 3 C surface 32 K R 1 92 120 9 20.0 4 C surface 26 K R 1 45 70 11 22.5 5 E surface 86 K G 1 1298 500 8 18.5 6 E surface 63 K G 1 667 200 NA 6.0 minD minD_H maxD maxD_H bodenD temperSize vesselShape 1 16.0 16.5 21.0 11.0 7.5 M 3 2 14.5 NA 16.0 NA NA NA M 77 3 18.5 NA 24.5 NA NA NA VF 7 4 NA NA 27.0 NA NA NA VF 8 5 22.5 9.5 22.5 9.5 0.0 VC 8

# **Andere Datenquellen**

#### Datenbanken und XML

- Datenbanken
  - **SQLite**: library(RSQLite)
  - PostgreSQL: library(RPostgreSQL)
  - MySQL, Oracle, ODBC (library(RODBC)), MongoDB, ...
- hierarchisch strukturierter Daten
  - XML
  - JSON

### **SQLite**

```
# nicht lauffähig
library(RSQLite)
drv <- dbDriver("SQLite")
con <- dbConnect(drv, "../data/DB.sqlite")

df = dbGetQuery(con, "SELECT * FROM ...")
head(df)</pre>
```

```
library(XML)
library(RCurl)
## Loading required package: bitops
url <- "https://www.w3schools.com/xml/simple.xml"</pre>
data <- getURL(url)</pre>
doc <- xmlParse(data)</pre>
class(doc)
## [1] "XMLInternalDocument" "XMLAbstractDocument"
```

```
doc <- xmlRoot(doc) #qives content of root</pre>
doc[1]
## $food
## <food>
     <name>Belgian Waffles</name>
##
     <price>$5.95</price>
##
     <description>Two of our famous Belgian Waffles with plenty of real
##
     <calories>650</calories>
##
## </food>
##
## attr(,"class")
## [1] "XMLInternalNodeList" "XMLNodeList"
```

#### **XML**

```
df <- xmlToDataFrame(nodes=getNodeSet(doc,"//food"))
df[,c(1,2,4)]</pre>
```

```
##
                           name price calories
## 1
                 Belgian Waffles $5.95
                                            650
## 2
     Strawberry Belgian Waffles $7.95
                                            900
## 3 Berry-Berry Belgian Waffles $8.95
                                            900
## 4
                    French Toast $4.50
                                            600
## 5
             Homestyle Breakfast $6.95
                                            950
```