03 Datenaufbereitung mit tidyR

Martin Hanewald

2019-02-19

Packages

```
library(tidyverse)
library(knitr)
```

Überblick

Die gesamte Paketbibliothek `tidyverse` basiert auf dem Konzept von `tidy data`, (im Gegensatz zu `messy data`).

Ein Datensatz ist `tidy`, wenn

- Jede Variable eine eigene Spalte hat
- Jede Messung einer Variable in einer eigenen Zeile zu finden ist
- Mehrere Tabellen über eine eindeutige ID verknüpft werden können.

Die wichtigsten Funktionen von `tidyr`:

- `gather()`: Spaltenüberschriften zu Variablen
- `spread()`: Umkehrfunktion zu `gather`
- `separate()`: Textspalten auftrennen
- `unite()`: Umkehrfunktion zu `unite`

Beispiele von messy-Data

Variable “Year” ist in Spaltenüberschriften

```
data("table4a")
table4a %>% kable()
```

country	1999	2000
Afghanistan	745	2666
Brazil	37737	80488
China	212258	213766

Anwendung von `gather`:

```
table4a %>%
    gather(year, count, -country) %>%
    kable()
```

country	year	count
Afghanistan	1999	745
Brazil	1999	37737
China	1999	212258

country	year	count
Afghanistan	2000	2666
Brazil	2000	80488
China	2000	213766

In Spalte `rate` sind mehrere Variablen enthalten und die Variable `year` ist auf zwei Spalten verteilt.

```
data("table5")
table5 %>% kable()
```

country	century	year	rate
Afghanistan	19	99	745/19987071
Afghanistan	20	00	2666/20595360
Brazil	19	99	37737/172006362
Brazil	20	00	80488/174504898
China	19	99	212258/1272915272
China	20	00	213766/1280428583

Anwendung von `separate` und `unite`

```
table5 %>%
    separate(rate, into = c("cases", "population")) %>%
    unite(year, century, year, sep="") %>%
    kable()
```

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

Aufgabe: Transformiere den WHO Datensatz in das tidy Format

```
who_messy <- read_csv2('who_messy.csv')
```

Show 10 entries Search:

	country	iso2	iso3	year	ep_f_014	ep_f_1524	ep_f_2534	ep_f_35
1	Luxembourg	LU	LUX	1997				
2	Marshall Islands	MH	MHL	2005				

3	Micronesia (Federated States of)	FM	FSM	2007	1	0	2	
4	Yemen	YE	YEM	1997				
5	Cook Islands	CK	COK	2000				
6	Tunisia	TN	TUN	2005				
7	Angola	AO	AGO	2003				
8	Cook Islands	CK	COK	2003				
9	Ghana	GH	GHA	2002				
10	Seychelles	SC	SYC	2009	0	0	0	

Showing 1 to 10 of 100 entries Previous 1 2 3 4 5 … 10 Next

Aus Dokumentation:

> The data uses the original codes given by the World Health Organization. The column names for columns five through 60 are made by combining a code for method of diagnosis (rel = relapse, sn = negative pulmonary smear, sp = positive pulmonary smear, ep = extrapulmonary) to a code for gender (f = female, m = male) to a code for age group (014 = 0-14 yrs of age, 1524 = 15-24 years of age, 2534 = 25 to 34 years of age, 3544 = 35 to 44 years of age, 4554 = 45 to 54 years of age, 5564 = 55 to 64 years of age, 65 = 65 years of age or older).

Lösung

```
who_tidy <- who_messy %>%
    gather(col, value, -(country:year)) %>%
    separate(col, into = c('method', 'sex', 'age')) %>%
    drop_na()
```

Show 10 entries Search:

	country	iso2	iso3	year	method	sex	age	value
1	Hungary	HU	HUN	2010	sp	f	014	0
2	Argentina	AR	ARG	2012	ep	m	1524	128
3	Solomon Islands	SB	SLB	2009	sn	m	4554	6
4	Germany	DE	DEU	2001	sp	m	1524	3
5	El Salvador	SV	SLV	2012	sn	m	2534	28
6	Saint Vincent and the Grenadines	VC	VCT	2005	sp	f	5564	0
7	China, Hong Kong SAR	HK	HKG	2005	sn	f	014	14
8	Canada	CA	CAN	2000	sp	f	65	66

9	Nauru	NR	NRU	2003	sp		m	014	0
10	Lebanon	LB	LBN	1996	sp		m	1524	28

Showing 1 to 10 of 100 entries Previous 1 2 3 4 5 … 10 Next

```
```