Data manipulation in R

Eduard Bukin

February 21, 2018

Plan of the first meeting of the R-Users at IAMO

1. **Get data in R**: `base::read.csv()` - why we should never use it; `readr::read_csv()`; `readxl:read_excel()`
2. Glance at data in R: `str()`; `glimpse()`; `tibble::tbl_df()`
3. Basic grammar of data manipulation `dplyr`: `select()`, `filter()`, `mutate()`, `summaries()`, `group_by()`
4. ***tidy data***

To the R code

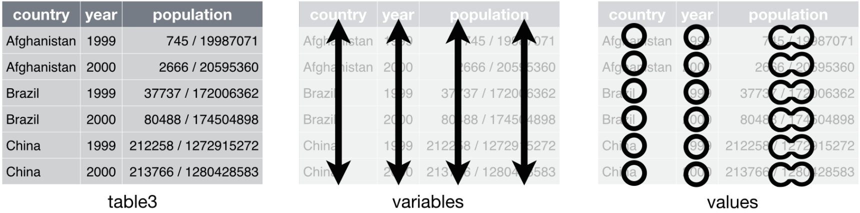
Tidy data

- What is tidy?
- How to make it tidy?
- Get data in R: `readr::read_csv()`; `readxl:read_excel()`

Data sets examples 1 (1)

```
##       country year              rate
## 1 Afghanistan 1999      745/19987071
## 2 Afghanistan 2000     2666/20595360
## 3      Brazil 1999   37737/172006362
## 4      Brazil 2000   80488/174504898
## 5       China 1999 212258/1272915272
## 6       China 2000 213766/1280428583
```

Data sets examples 1 - structure



country	year	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

table3

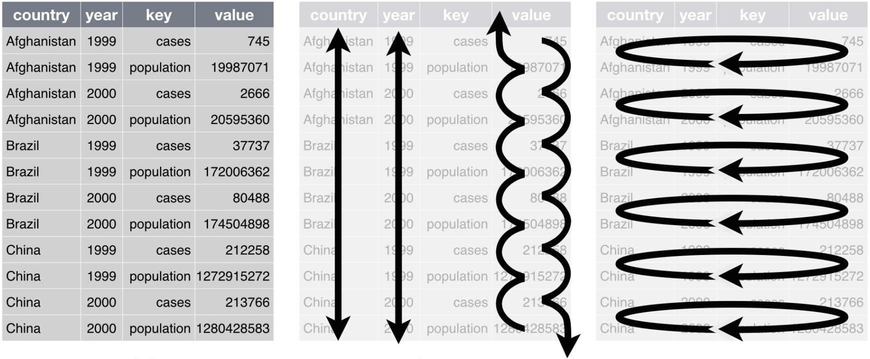
variables

values

Data sets examples 2

```
##        country year        key      value
## 1  Afghanistan 1999      cases        745
## 2  Afghanistan 1999 population   19987071
## 3  Afghanistan 2000      cases       2666
## 4  Afghanistan 2000 population   20595360
## 5       Brazil 1999      cases      37737
## 6       Brazil 1999 population  172006362
## 7       Brazil 2000      cases      80488
## 8       Brazil 2000 population  174504898
## 9        China 1999      cases     212258
## 10       China 1999 population 1272915272
## 11       China 2000      cases     213766
## 12       China 2000 population 1280428583
```

Data sets examples 2 - structure



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

table2

variables

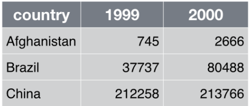
observations

Data sets examples 3

```
##       country   1999   2000
## 1 Afghanistan    745   2666
## 2      Brazil  37737  80488
## 3       China 212258 213766
```

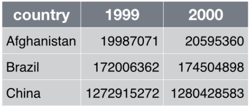
```
##       country       1999       2000
## 1 Afghanistan   19987071   20595360
## 2      Brazil  172006362  174504898
## 3       China 1272915272 1280428583
```

Data sets examples 3 - structure



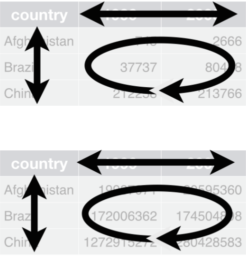
country	1999	2000
Afghanistan	745	2666
Brazil	37737	80488
China	212258	213766

table4

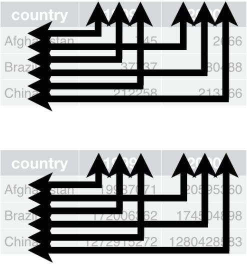


country	1999	2000
Afghanistan	19987071	20595360
Brazil	172006362	174504898
China	1272915272	1280428583

table5



variables



observations

Tidy data

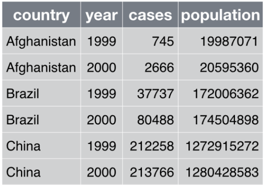
Your data will be easier to work with in R if it follows three rules:

- Each variable in the data set is placed in its own column
- Each observation is placed in its own row
- Each value is placed in its own cell

Data that satisfies these rules is known as tidy data.

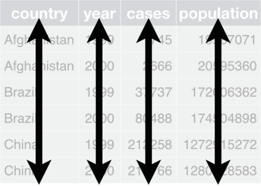
Borrowed from Data science with R: Tidying

Tidy data example - structure

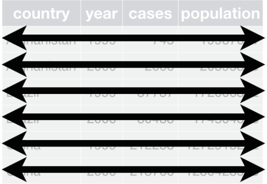


country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

table1



variables



observations

Tidy data example

```
DSR::table1
```

```
##       country year  cases population
## 1 Afghanistan 1999    745   19987071
## 2 Afghanistan 2000   2666   20595360
## 3      Brazil 1999  37737  172006362
## 4      Brazil 2000  80488  174504898
## 5       China 1999 212258 1272915272
## 6       China 2000 213766 1280428583
```

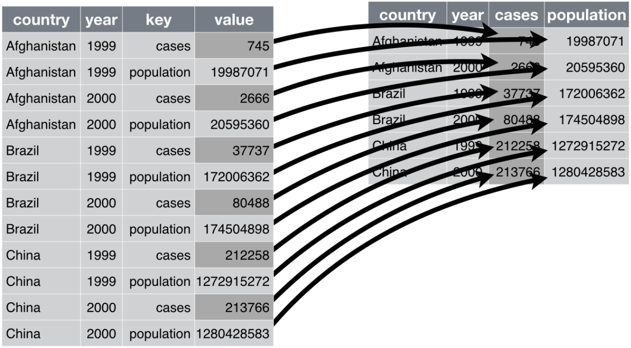
How to make tidy data?

Use R package `tidyr`.

Functions:

- `spread()`
- `gather()`

spread()

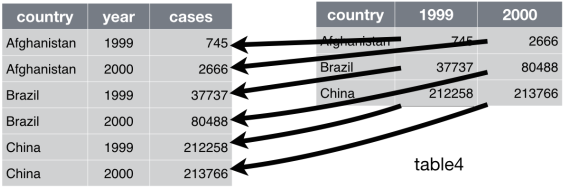


country	year	key	value
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999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

table2

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

gather()



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

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

table4

Non-tidy data

Be aware, that sometimes, data cannot be tidy and in fact is it easier to work with such data.

For more information, see Non-tidy data.

Where to go next?

Data manipulation with `dplyr` - next meeting