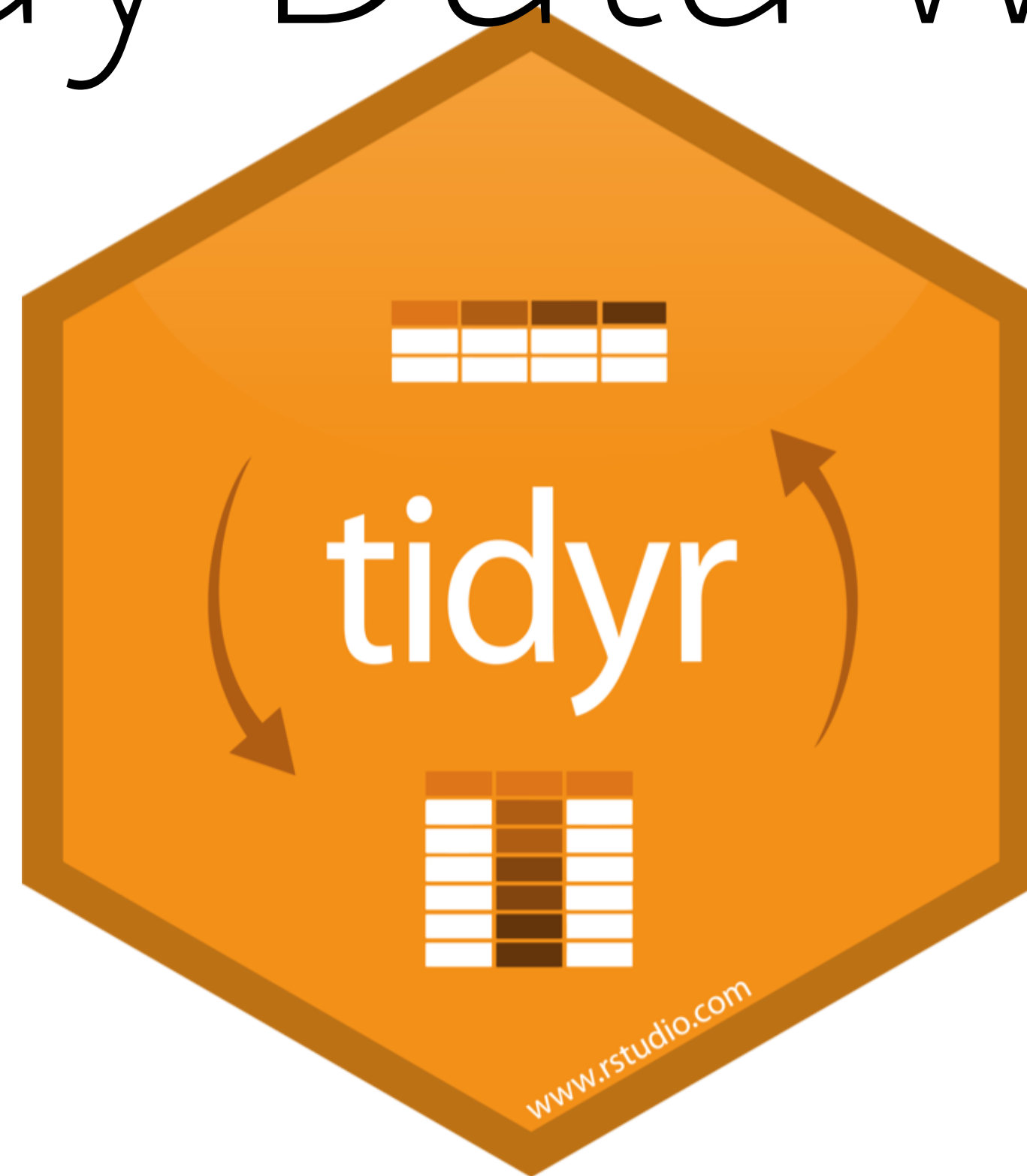


Tidy Data with



"Data comes in many formats, but R prefers just one: tidy data. "

- Garrett Grolemund

Tidy data

country	year	cases	pop
Afghanistan	1999	745	10127000
Afghanistan	2000	666	10125000
Afghanistan	2001	787	10125000
Afghanistan	2002	1123	10125000
Afghanistan	2003	2230	10125000
Afghanistan	2004	3760	10125000

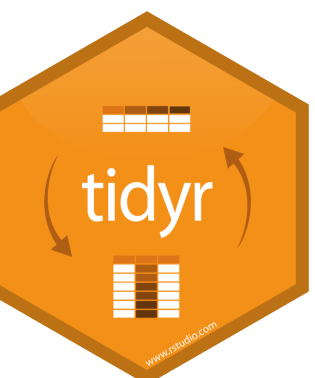
A data set is **tidy** iff:

1. Each **variable** is in its own **column**
2. Each **observation** is in its own **row**
3. Each **value** is in its own **cell**

Also see these papers, in your “other resources” folder:

Wickham, 2014: *Tidy Data*

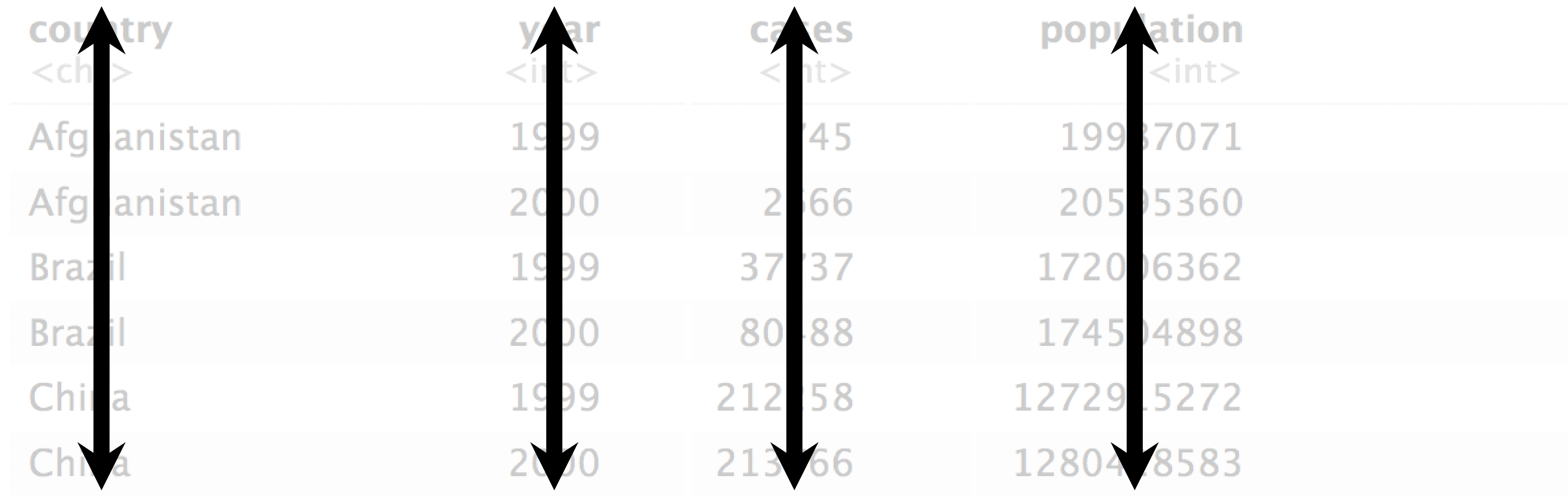
Broman and Woo, 2017: *Data Organization in Spreadsheets*



Quiz

What are the variables in this data set?

table1



country <chr>	year <int>	cases <int>	population <int>
Afghanistan	1999	745	19987071
Afghanistan	2000	266	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212158	1272915272
China	2000	213766	1280478583

6 rows

Tidy data

country	year	cases	pop
Afghanistan	1999	745	19927000
Afghanistan	2000	689	20005000
Afghanistan	2001	757	19965000
Afghanistan	2002	673	20000000
Afghanistan	2003	2230	12743000
Afghanistan	2004	3700	12743000

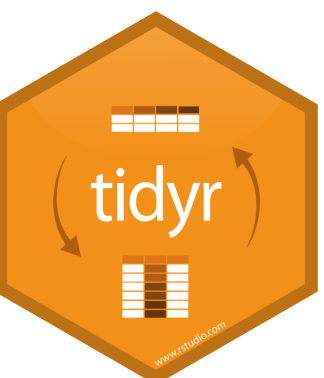
A data set is **tidy** iff:

1. Each **variable** is in its own **column**
2. Each **observation** is in its own **row**
3. Each **value** is in its own **cell**

variable: all values that measure the same underlying *attribute*

observation: all values measured on the same *unit*

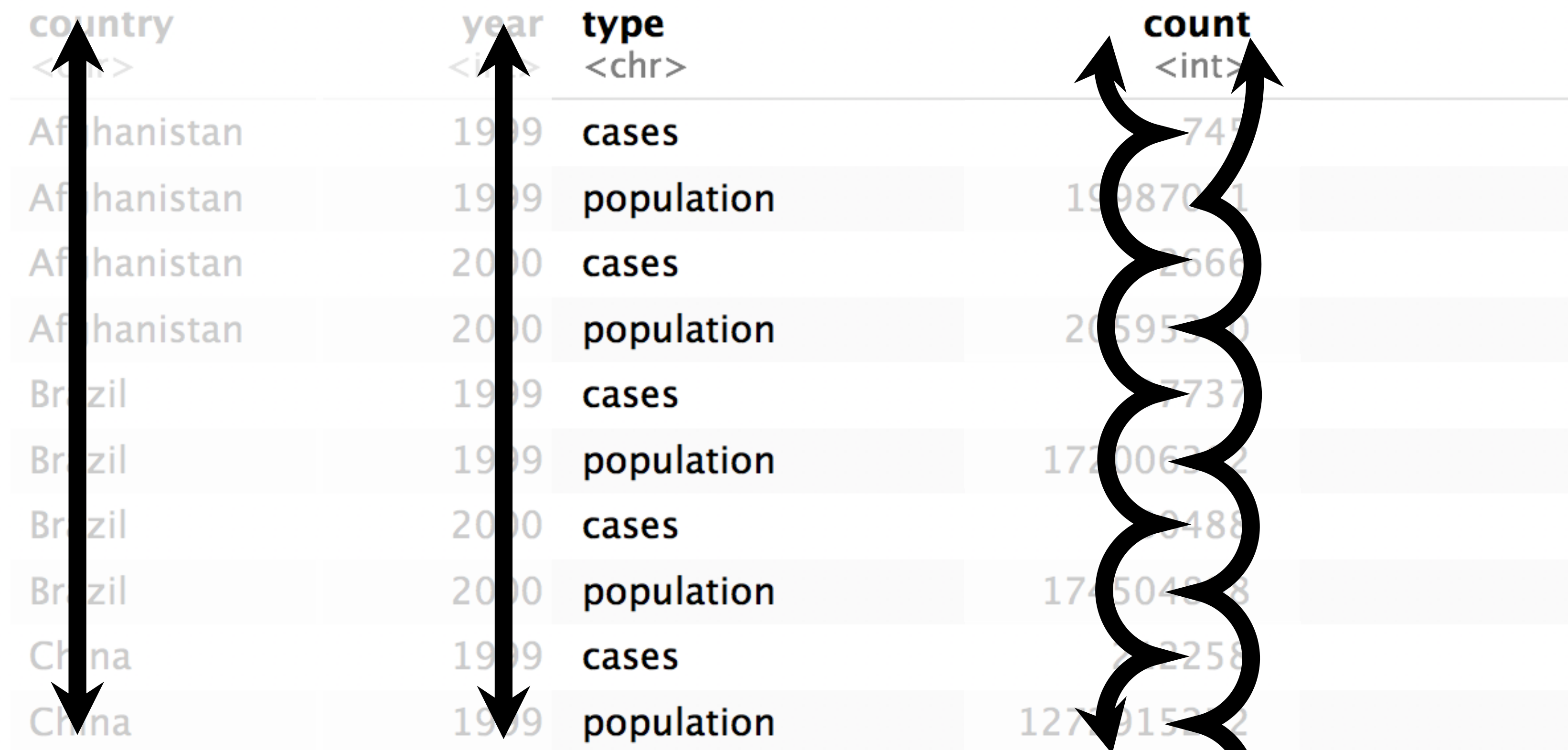
value: belongs to one **variable** and one **observation**



Quiz

What are the variables in this data set?

table2



country <chr>	year <int>	type <chr>	count <int>
Afghanistan	1999	cases	745
Afghanistan	1999	population	1998701
Afghanistan	2000	cases	2666
Afghanistan	2000	population	2059530
Brazil	1999	cases	7737
Brazil	1999	population	17200632
Brazil	2000	cases	9488
Brazil	2000	population	17450488
China	1999	cases	2258
China	1999	population	12791522

1-10 of 12 rows

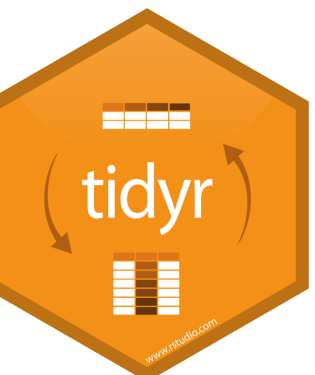
Previous 1 2 Next


Tidy data

country	year	cases	pop
Afghanistan	1999	745	10027000
Afghanistan	2000	666	10025000
Algeria	1999	1583	14520000
Algeria	2000	1553	14515000
Algeria	1999	22266	12747000
Algeria	2000	23766	12742000

A data set is **tidy** iff:

1. Each **variable** is in its own **column**
2. Each **observation** is in its own **row**
3. Each **value** is in its own **cell**

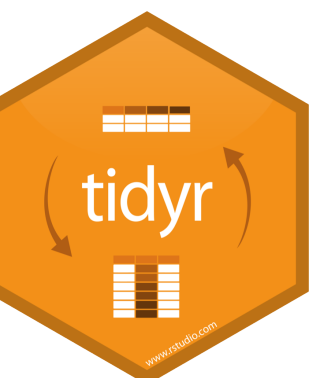




country <chr>	year <int>	cases <int>	population <int>
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

6 rows

```
table1$country
table1$year
table1$cases
table1$population
```

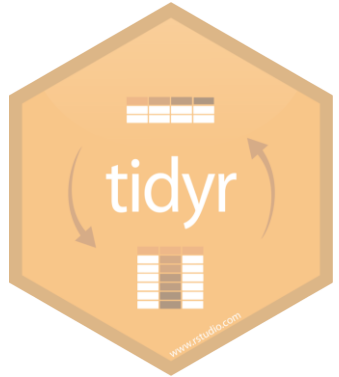


country	year	type	count
<chr>	<int>	<chr>	<int>
Afghanistan	1999	cases	745
Afghanistan	1999	n	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	n	95360
Brazil	1999	cases	737
Brazil	1999	n	62
Brazil	2000	cases	88
Brazil	2000	n	8
China	1999	cases	58
China	1999	n	72

1-10 of 12 rows

Previous12Next

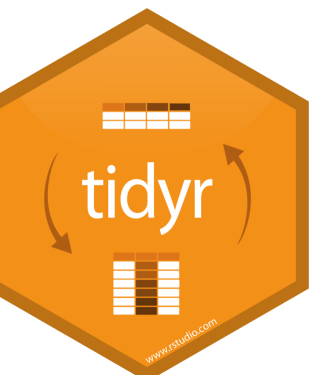
```
table2$count
table2$year
table2$count[c(1,3,5,7,9,11)]
table2$count[c(2,4,6,8,10,12)]
```



country <chr>	year <int>	cases <int>	population <int>	rate <dbl>
Afghanistan	1999	745	19987071	0.0000372741
Afghanistan	2000	2666	20595360	0.0001294466
Brazil	1999	37737	172006362	0.0002193930
Brazil	2000	80488	174504898	0.0004612363
China	1999	212258	1272915272	0.0001667495
China	2000	213766	1280428583	0.0001669488

6 rows

```
table1$cases / table1$population -> table1$rate
```



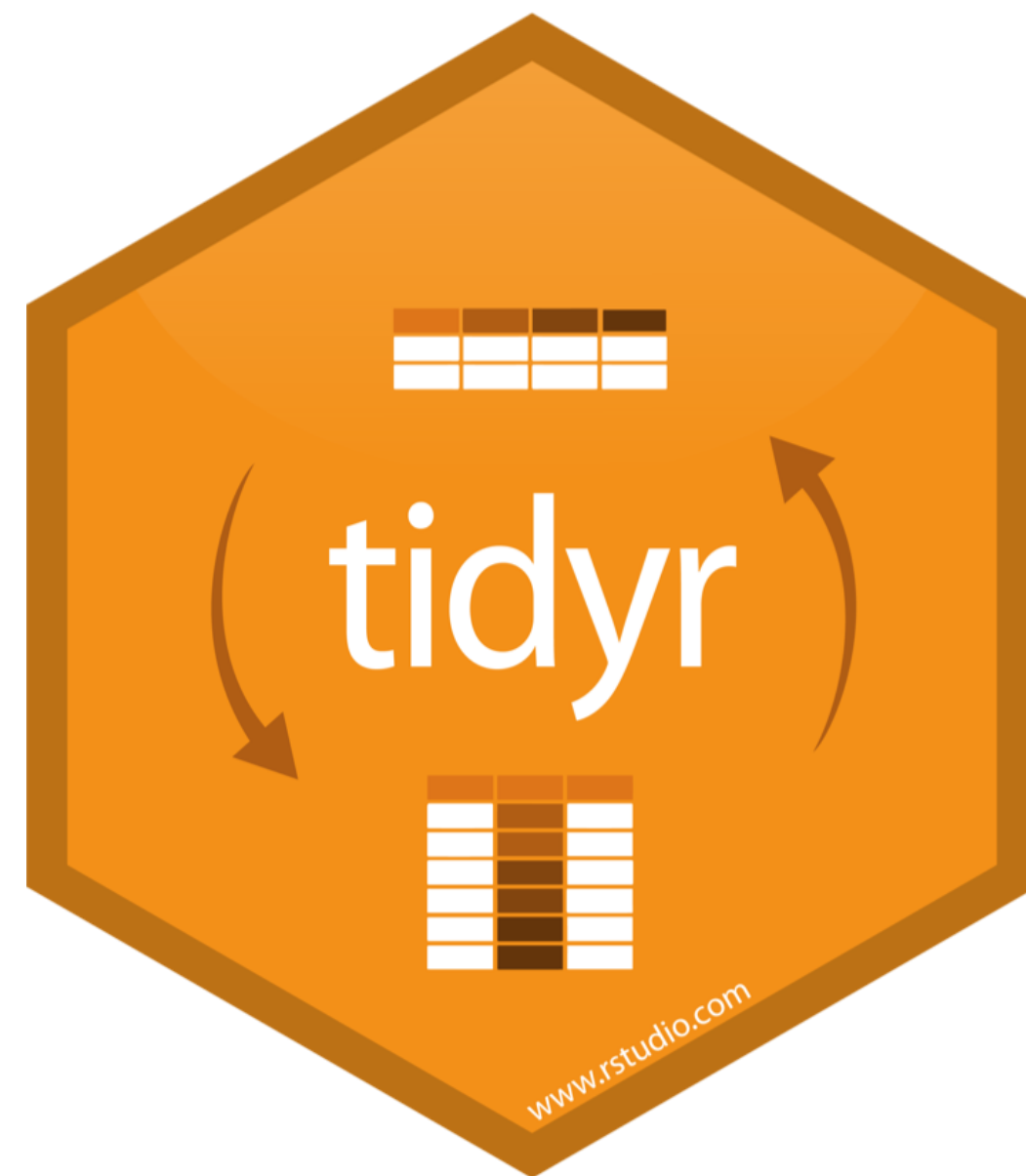
"Tidy data sets are all alike; but every messy data set is messy in its own way."

- Hadley Wickham

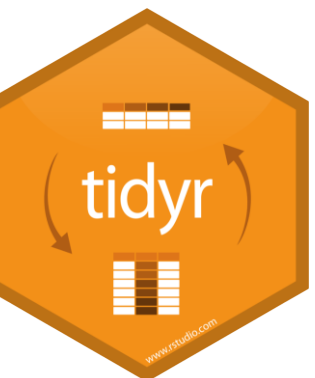
tidyr



tidyr



A package that reshapes the layout of tabular data.



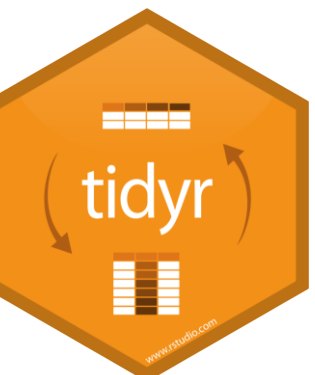
pivot_longer()



Toy data

```
03-Tidy-Data.Rmd x
1 |---
2 title: "Tidy Data"
3 output: html_notebook
4 ---
5
6 ```{r setup}
7 library(tidyverse)
8 library(babynames)
9
10 # Toy data
11 cases <- tribble(
12   ~Country, ~"2011", ~"2012", ~"2013",
13   "FR",      7000,    6900,    7000,
14   "DE",      5800,    6000,    6200,
15   "US",     15000,   14000,    13000,
16 )
17
18 pollution <- tribble(
19   ~city, ~size, ~amount,
20   "New York", "large", 23,
21   "New York", "small", 14,
22   "London", "large", 22,
23   "London", "small", 16,
24   "Beijing", "large", 121,
25   "Beijing", "small", 121,
26 )
27
28 x <- tribble(
29   ~x1, ~x2,
30   "A", 1,
31   "B", NA,
32   "C", NA,
33   "D", 3,
34   "E", NA,
35 )
```

```
cases <- tribble(
  ~Country, ~"2011", ~"2012", ~"2013",
  "FR",      7000,    6900,    7000,
  "DE",      5800,    6000,    6200,
  "US",     15000,   14000,    13000
)
```



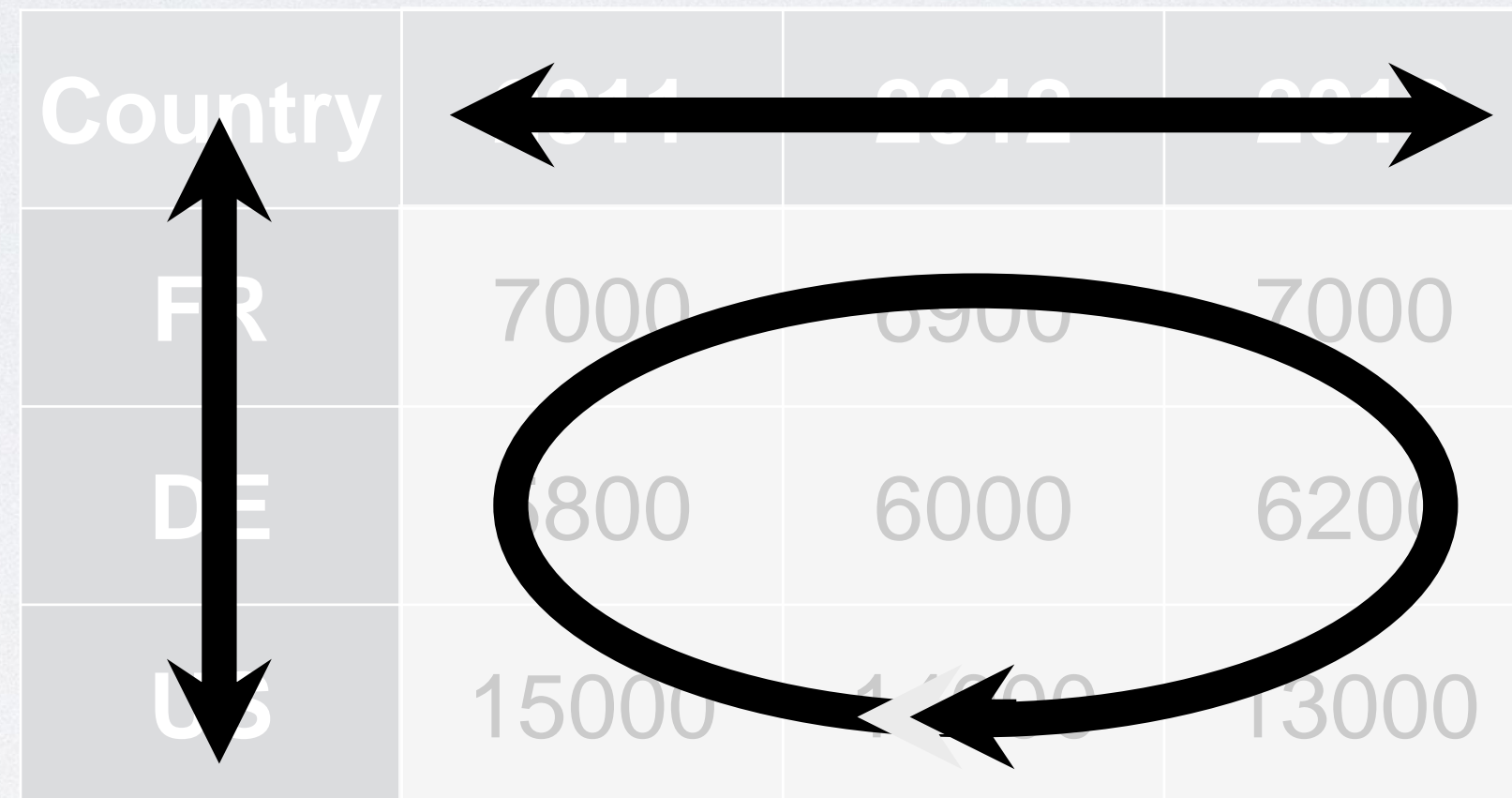
Quiz

What are the variables in cases?

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

Quiz

What are the variables in cases?



Country	2011	2012	2013
FR	7000	6900	7000
DE	800	6000	6200
US	15000	14000	13000

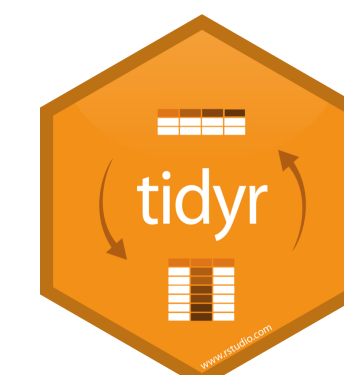
- Country
- Year
- Count

Your Turn 7

On a sheet of paper, draw how the cases data set would look if it had the same values grouped into three columns: *country*, *year*, *n*

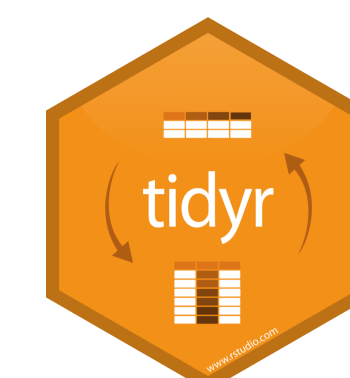
Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000



Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

Country	Year	n
---------	------	---



Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

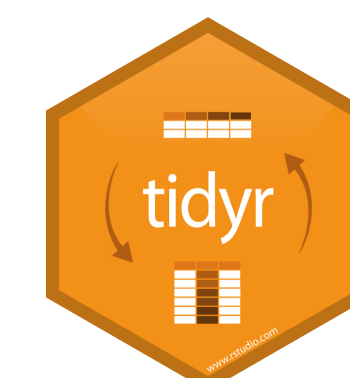
Country	Year	n
FR	2011	7000

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

Country	Year	n
FR	2011	7000
DE	2011	5800

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

Country	Year	n
FR	2011	7000
DE	2011	5800
US	2011	15000



Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

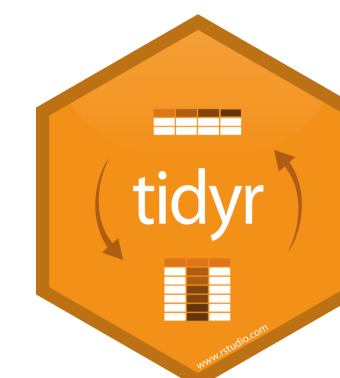
Country	Year	n
FR	2011	7000
DE	2011	5800
US	2011	15000
FR	2012	6900

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

Country	Year	n
FR	2011	7000
DE	2011	5800
US	2011	15000
FR	2012	6900
DE	2012	6000

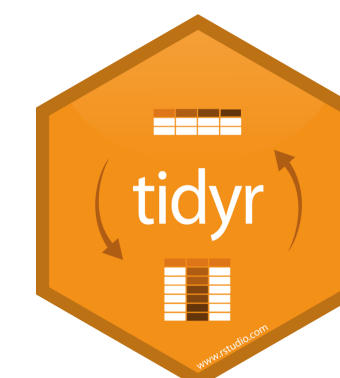
Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

Country	Year	n
FR	2011	7000
DE	2011	5800
US	2011	15000
FR	2012	6900
DE	2012	6000
US	2012	14000



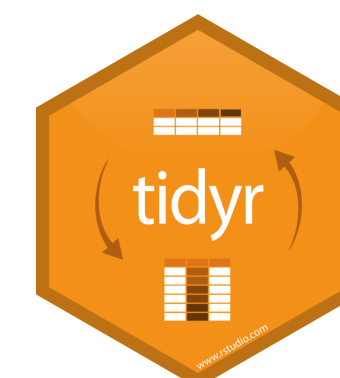
Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

Country	Year	n
FR	2011	7000
DE	2011	5800
US	2011	15000
FR	2012	6900
DE	2012	6000
US	2012	14000
FR	2013	7000



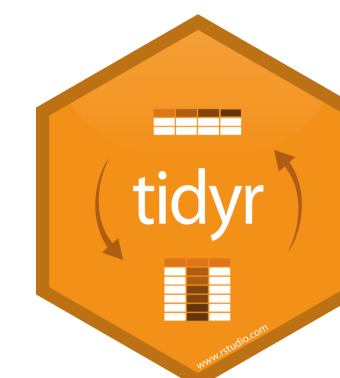
Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

Country	Year	n
FR	2011	7000
DE	2011	5800
US	2011	15000
FR	2012	6900
DE	2012	6000
US	2012	14000
FR	2013	7000
DE	2013	6200



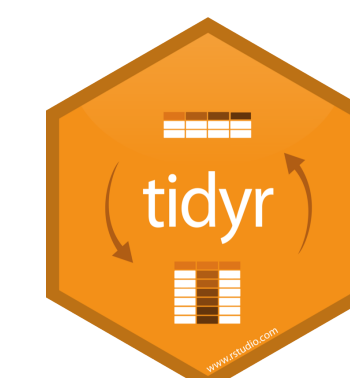
Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

Country	Year	n
FR	2011	7000
DE	2011	5800
US	2011	15000
FR	2012	6900
DE	2012	6000
US	2012	14000
FR	2013	7000
DE	2013	6200
US	2013	13000



Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

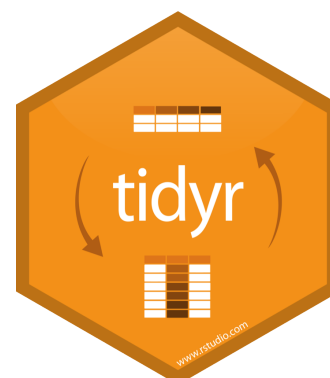
Country	Year	Pop
FR	2011	7000
DE	2011	5800
US	2011	15000
FR	2012	6900
DE	2012	6000
US	2012	14000
FR	2013	7000
DE	2013	6200
US	2013	13000



Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

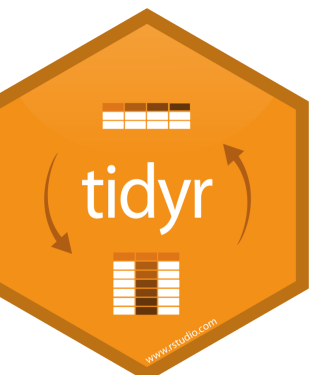
`pivot_longer()`

Country	Year	n
FR	2011	7000
DE	2011	5800
US	2011	15000
FR	2012	6900
DE	2012	6000
US	2012	14000
FR	2013	7000
DE	2013	6200
US	2013	13000



Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

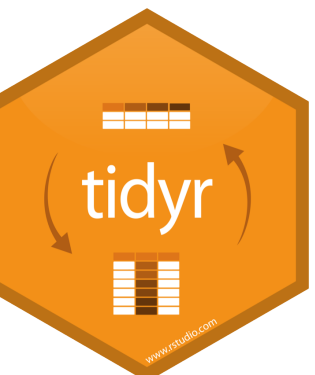
	1	2
Country	Year	n
FR	2011	7000
DE	2011	5800
US	2011	15000
FR	2012	6900
DE	2012	6000
US	2012	14000
FR	2013	7000
DE	2013	6200
US	2013	13000



names_to (former column names)

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

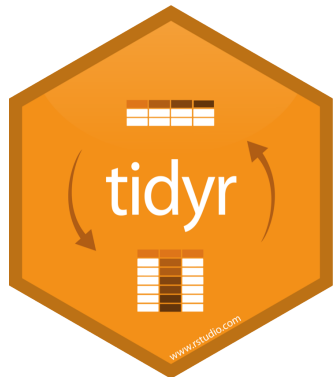
Country	Year	n
FR	2011	7000
DE	2011	5800
US	2011	15000
FR	2012	6900
DE	2012	6000
US	2012	14000
FR	2013	7000
DE	2013	6200
US	2013	13000



values_to (former cells)

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

Country	Year	n
FR	2011	7000
DE	2011	5800
US	2011	15000
FR	2012	6900
DE	2012	6000
US	2012	14000
FR	2013	7000
DE	2013	6200
US	2013	13000



pivot_longer()

```
cases %>% pivot_longer(cols = 2:4, names_to = "year", values_to = "n")
```

**data frame to
reshape**

**numeric
indices of
columns to
collapse
(or names)**

**name of the
new key
column
(a character
string)**

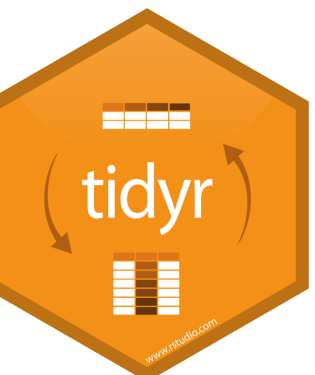
**name of the
new value
column
(a character
string)**

pivot_longer()

```
cases %>% pivot_longer(2:4, "year", "n")
```

numeric
indices

Country <chr>	2	3	4
	2011 <dbl>	2012 <dbl>	2013 <dbl>
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

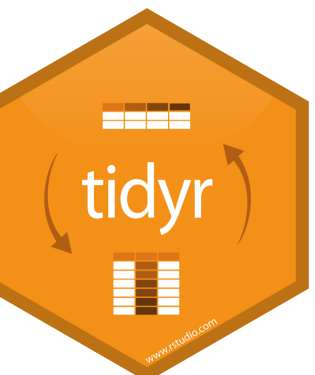


pivot_longer()

```
cases %>% pivot_longer(c("2011", "2012", "2013"), "year", "n")
```

names

Country <chr>	2011	2012	2013
	2011 <dbl>	2012 <dbl>	2013 <dbl>
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000



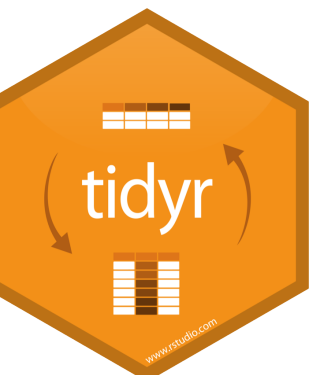
pivot_longer()

```
cases %>% pivot_longer(-Country, "year", "n")
```

Everything
except...

Not Country Not Country Not Country

Country <chr>	2011 <dbl>	2012 <dbl>	2013 <dbl>
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000



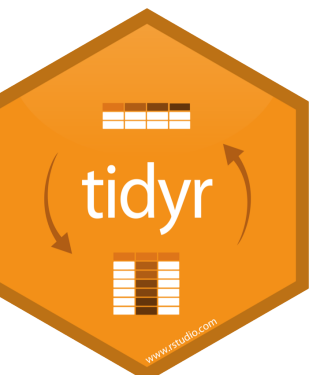
pivot_longer()

```
cases %>% pivot_longer(2:4, "year", "n")
```

```
cases %>% pivot_longer(c("2011", "2012", "2013"), "year", "n")
```

```
cases %>% pivot_longer(starts_with("201"), "year", "n")
```

```
cases %>% pivot_longer(-Country, "year", "n")
```



Your Turn 8

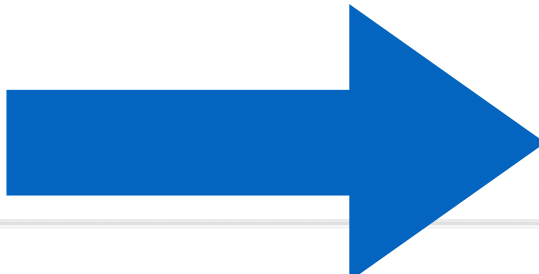
Use **pivot_longer()** to reorganize **table4a** into three columns: *country*, *year*, and *cases*.

	country <chr>	1999 <int>	2000 <int>
1	Afghanistan	745	2666
2	Brazil	37737	80488
3	China	212258	213766

3 rows

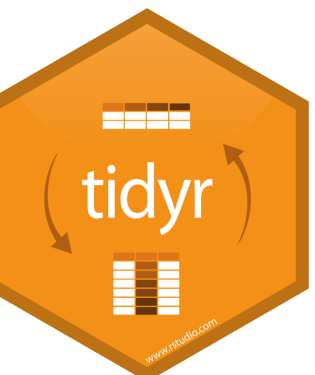

```
table4a %>%
```

```
  pivot_longer(cols = 2:3, names_to = "year", values_to = "n")
```



country <chr>	year <chr>	n <int>
Afghanistan	1999	745
Brazil	1999	37737
China	1999	212258
Afghanistan	2000	2666
Brazil	2000	80488
China	2000	213766

6 rows



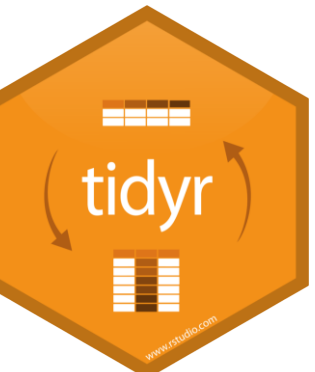
pivot_wider()



Toy data

```
03-Tidy-Data.Rmd x
1 |---
2 title: "Tidy Data"
3 output: html_notebook
4 ---
5
6 ```{r setup}
7 library(tidyverse)
8 library(babynames)
9
10 # Toy data
11 cases <- tribble(
12   ~Country, ~"2011", ~
13     "FR", 7000,
14     "DE", 5800,
15     "US", 15000,
16 )
17
18 pollution <- tribble(
19   ~city, ~size, ~
20     "New York", "large",
21     "New York", "small",
22     "London", "large",
23     "London", "small",
24     "Beijing", "large",
25     "Beijing", "small",
26 )
27
28 x <- tribble(
29   ~x1, ~x2,
30     "A", 1,
31     "B", NA,
32     "C", NA,
33     "D", 3,
34     "E", NA
35 )
```

```
pollution <- tribble(
  ~city, ~size, ~amount,
  "New York", "large", 23,
  "New York", "small", 14,
  "London", "large", 22,
  "London", "small", 16,
  "Beijing", "large", 121,
  "Beijing", "small", 56
)
```



Quiz

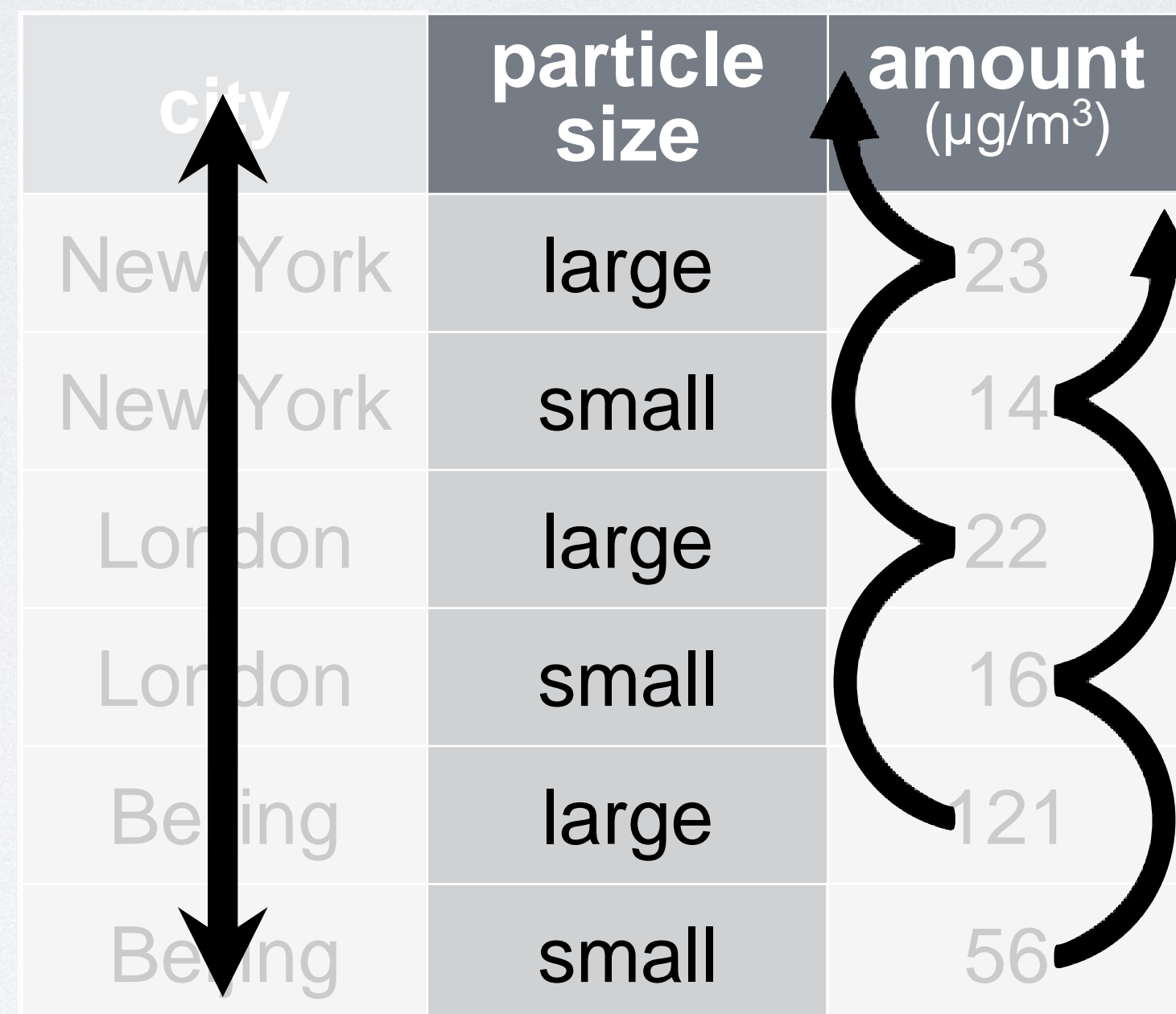
What are the variables in pollution?

city	particle size	amount ($\mu\text{g}/\text{m}^3$)
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

Quiz

What are the variables in pollution?

city	particle size	amount ($\mu\text{g}/\text{m}^3$)
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56



The diagram illustrates the relationships between the variables in the table. A long vertical arrow on the left points from the 'city' header to the 'Beijing' rows, indicating that 'city' is a variable. A curved arrow on the right points from the 'amount' header to the '23' and '14' values, indicating that 'amount' is a variable. Another curved arrow on the right points from the 'amount' header to the '121' and '56' values, indicating that 'amount' is a variable. A third curved arrow on the right points from the 'amount' header to the '22' and '16' values, indicating that 'amount' is a variable.

- City
- Amount of large particulate
- Amount of small particulate

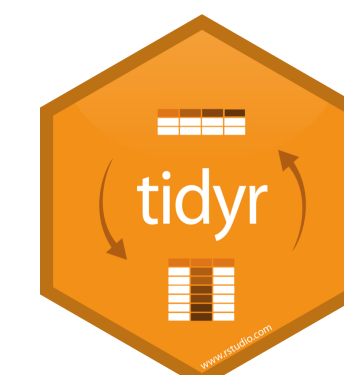
Your Turn 3

On a sheet of paper, draw how this data set would look if it had the same values grouped into three columns: *city*, *large*, *small*

city	size	amount
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

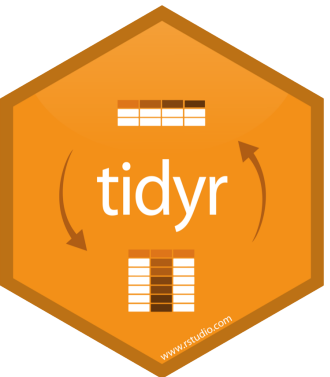
03:00

city	size	amount
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56



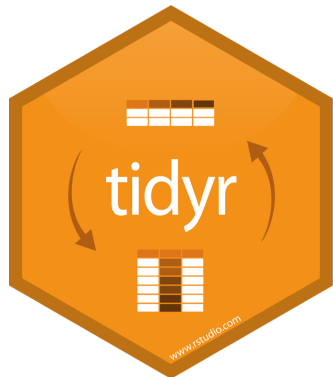
city	size	amount
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

city	large	small
------	-------	-------



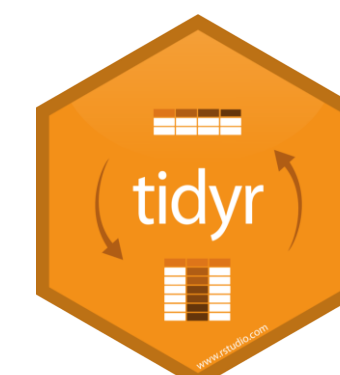
city	size	amount
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

city	large	small
New York	23	



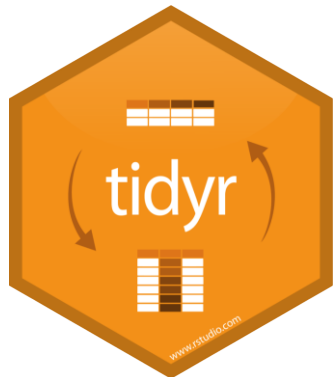
city	size	amount
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

city	large	small
New York	23	14



city	size	amount
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

city	large	small
New York	23	14
London	22	



city	size	amount
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

city	large	small
New York	23	14
London	22	16

city	size	amount
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

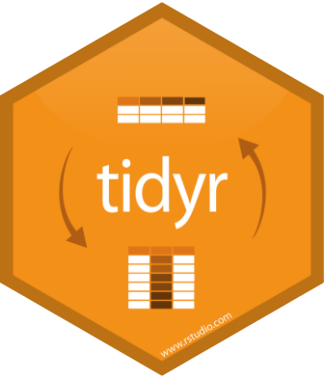
city	large	small
New York	23	14
London	22	16
Beijing	121	

city	size	amount
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

city	large	small
New York	23	14
London	22	16
Beijing	121	56

city	size	amount
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

city	large	small
New York	23	14
London	22	16
Beijing	121	56



city	size	amount
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

`pivot_wider()`

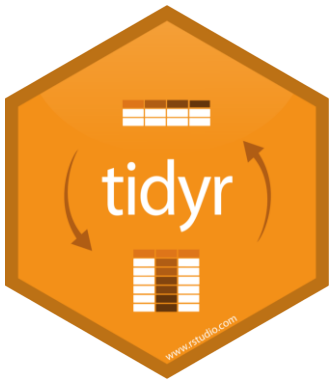
city	large	small
New York	23	14
London	22	16
Beijing	121	56

1

2

city	size	amount
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

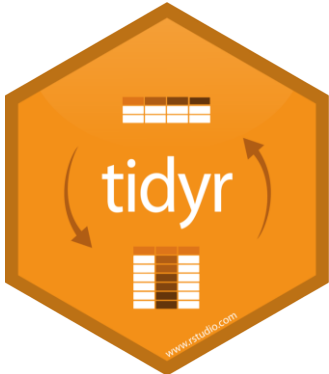
city	large	small
New York	23	14
London	22	16
Beijing	121	56



names_from (new column names)

city	size	amount
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

city	large	small
New York	23	14
London	22	16
Beijing	121	56



values_from (new cells)

city	size	amount
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

city	large	small
New York	23	14
London	22	16
Beijing	121	56

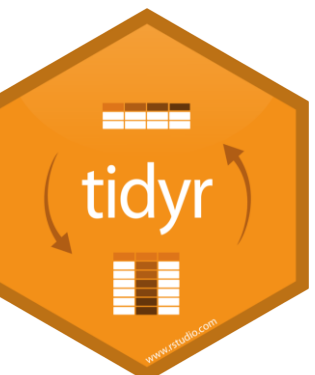
pivot_wider()

```
pollution %>% pivot_wider(names_from = size, values_from = amount)
```

**data frame to
reshape**

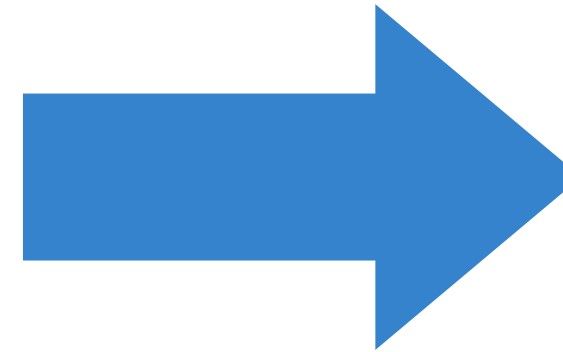
column to use for keys
(becomes new
column names)

column to use for values
(becomes new
column cells)



```
pollution %>% pivot_wider(names_from = size, values_from = amount)
```

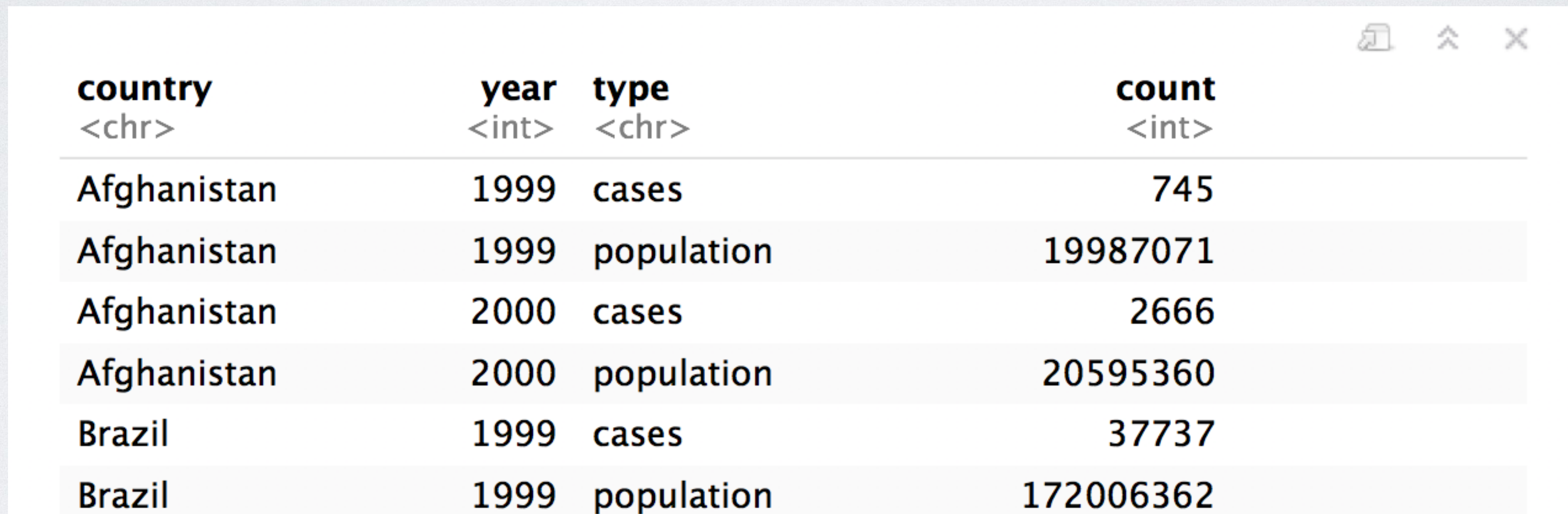
city	size	amount
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56



city	large	small
New York	23	14
London	22	16
Beijing	121	56

Your Turn 9


Use **pivot_wider()** to reorganize **table2** into four columns: *country*, *year*, *cases*, and *population*.

A screenshot of an RStudio console window. The window has a title bar with a file icon, an up arrow, and a close 'X' button. Inside the window, a data table is displayed with four columns: 'country', 'year', 'type', and 'count'. The 'country' column has a data type of '<chr>', 'year' is '<int>', 'type' is '<chr>', and 'count' is '<int>'. The table contains six rows of data for Afghanistan and Brazil, with rows for 'cases' and 'population' for each year (1999 and 2000 for Afghanistan, 1999 for Brazil).

country <chr>	year <int>	type <chr>	count <int>
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362

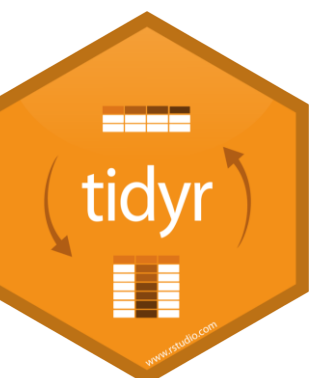

```
table2 %>%
```

```
  pivot_wider(names_from = type, values_from = count)
```



	country <chr>	year <int>	cases <int>	population <int>
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

6 rows



Tidy Data with

