ANT 6973: DATA VISUALIZATION AND EXPLORATION

RESHAPING DATA

TODAY'S TOPICS

- Principles of "tidy data"
- Reshaping from wide to long, and vice versa
- Splitting and combining cells/columns

WRANGLING

RESHAPING

MANIPULATION

MUNGING

TRANSFORMATION

MAYBE >50% OF YOUR TIME

TYPICAL GOALS

- Clean and error-check the data
- Make the data suitable to use with particular software
  - Plotting
  - Statistical tests
- Reveal information

PACKAGES FOR WORKING WITH DATA

tidyr

dplyr

Both are part of core

```
library("tidyverse")
```

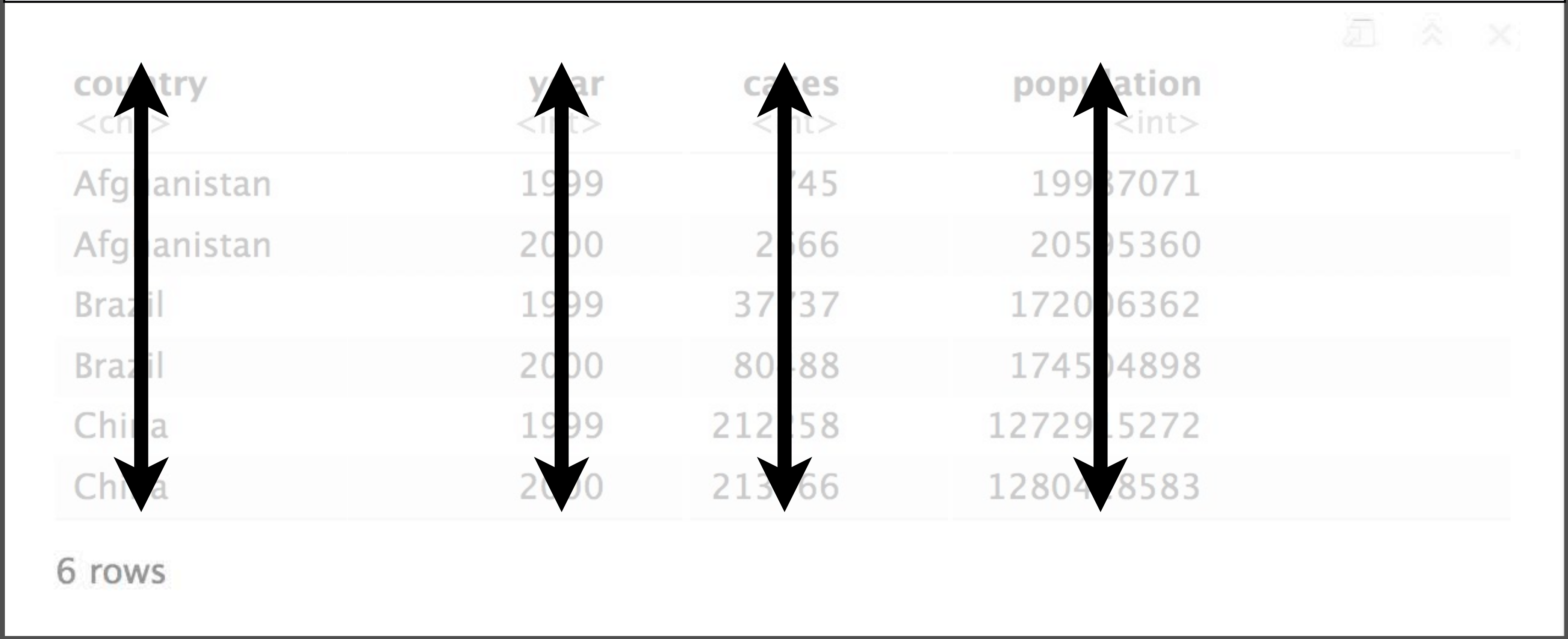
PACKAGES FOR WORKING WITH DATA

`tidyr` — Create tidy data by reshaping

`dplyr` — Manipulate and summarize data

WHAT ARE THE VARIABLES IN THIS DATA SET?

```
tidyr::table1
```

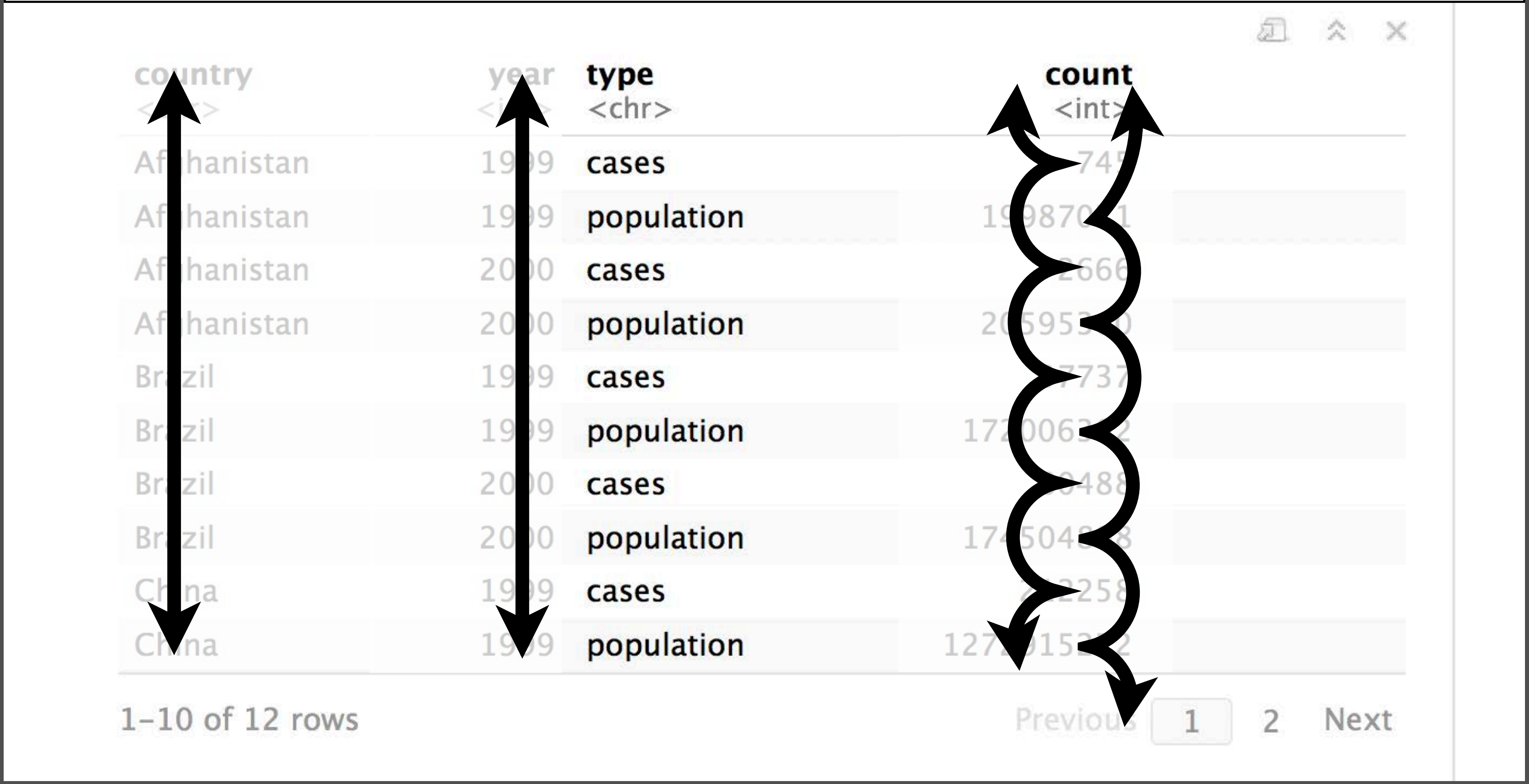


country	year	cases	population
<chr>	<int>	<int>	<int>
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

6 rows

WHAT ARE THE VARIABLES IN THIS DATA SET?

```
tidyr::table2
```

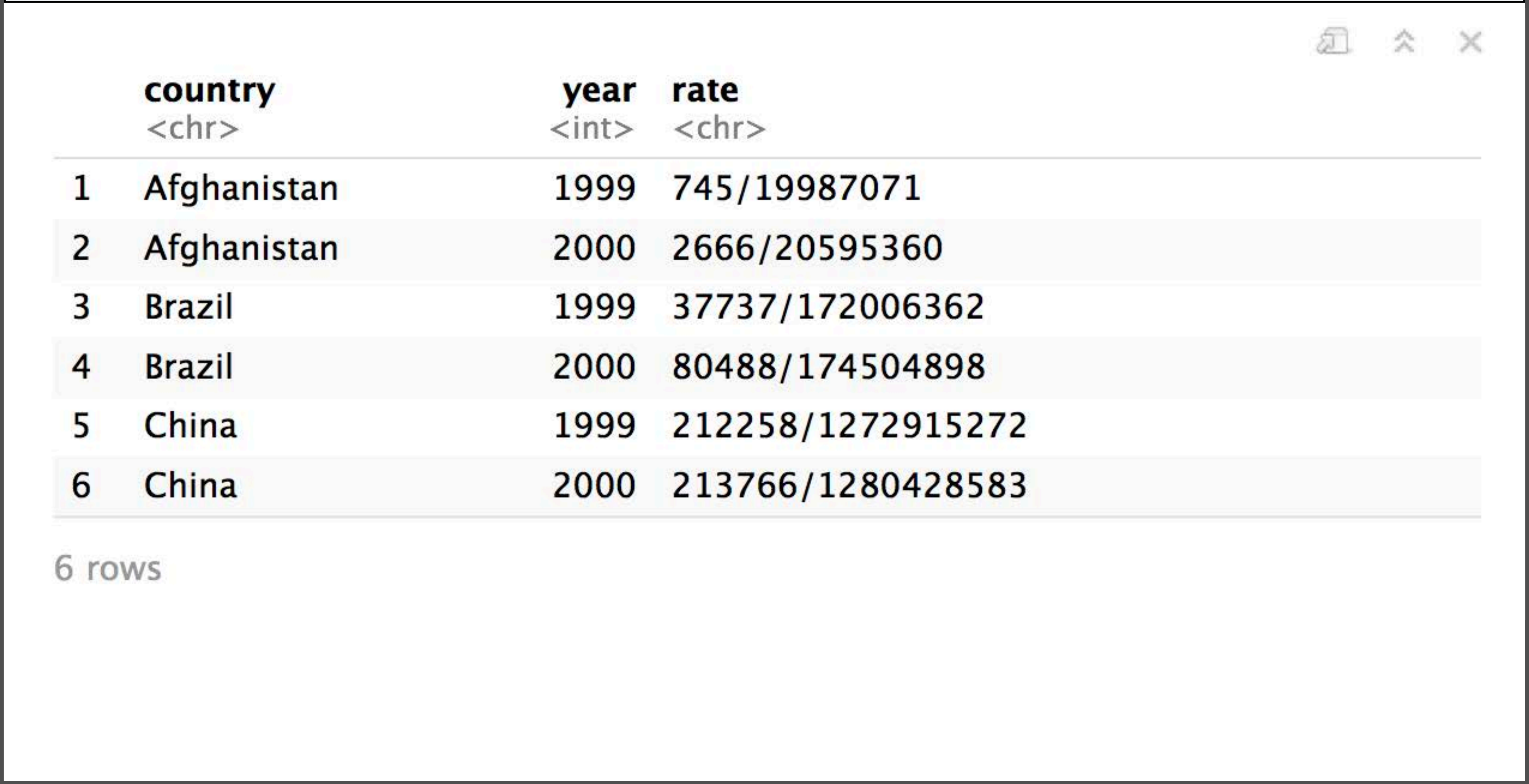


country <chr>	year <int>	type <chr>	count <int>
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999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

1–10 of 12 rows

OTHER (BAD) IDEAS

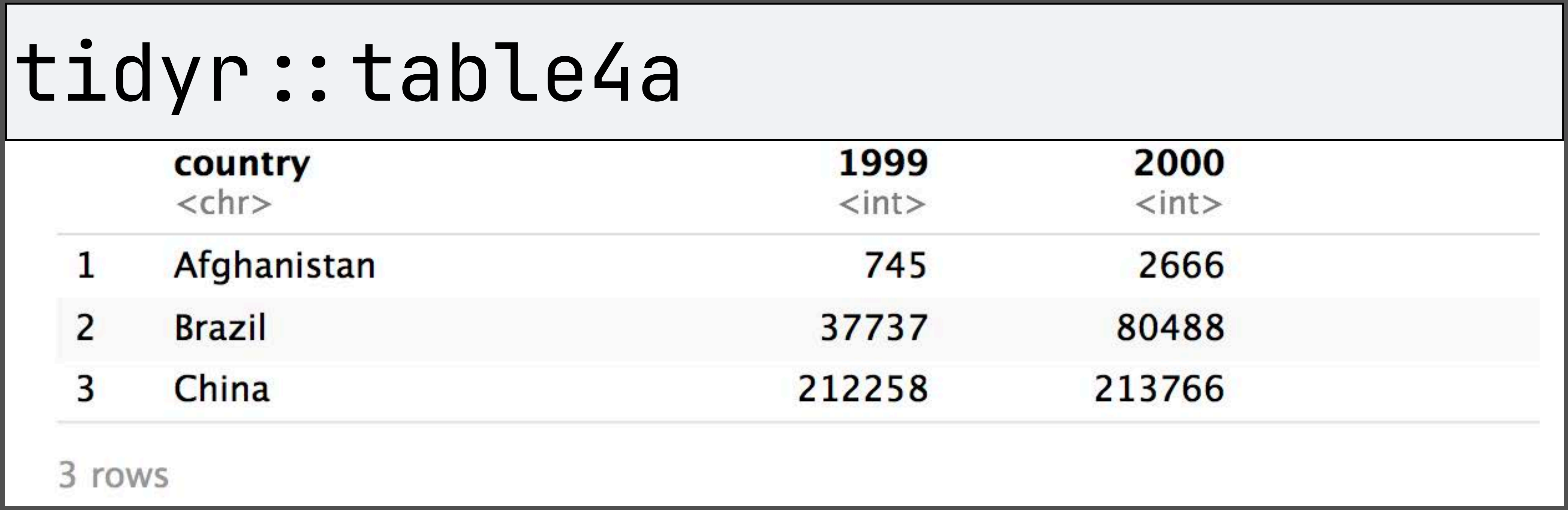
```
tidyr::table3
```



	country	year	rate
	<chr>	<int>	<chr>
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

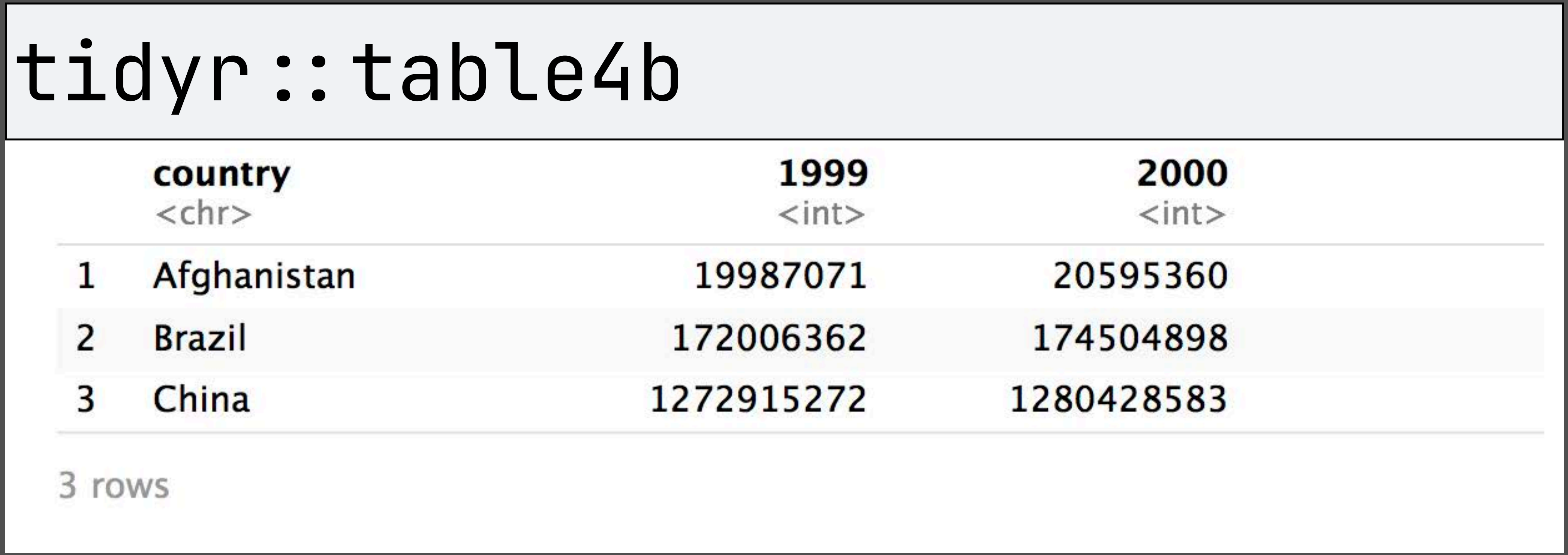
OTHER (BAD) IDEAS



`tidyr::table4a`

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

3 rows



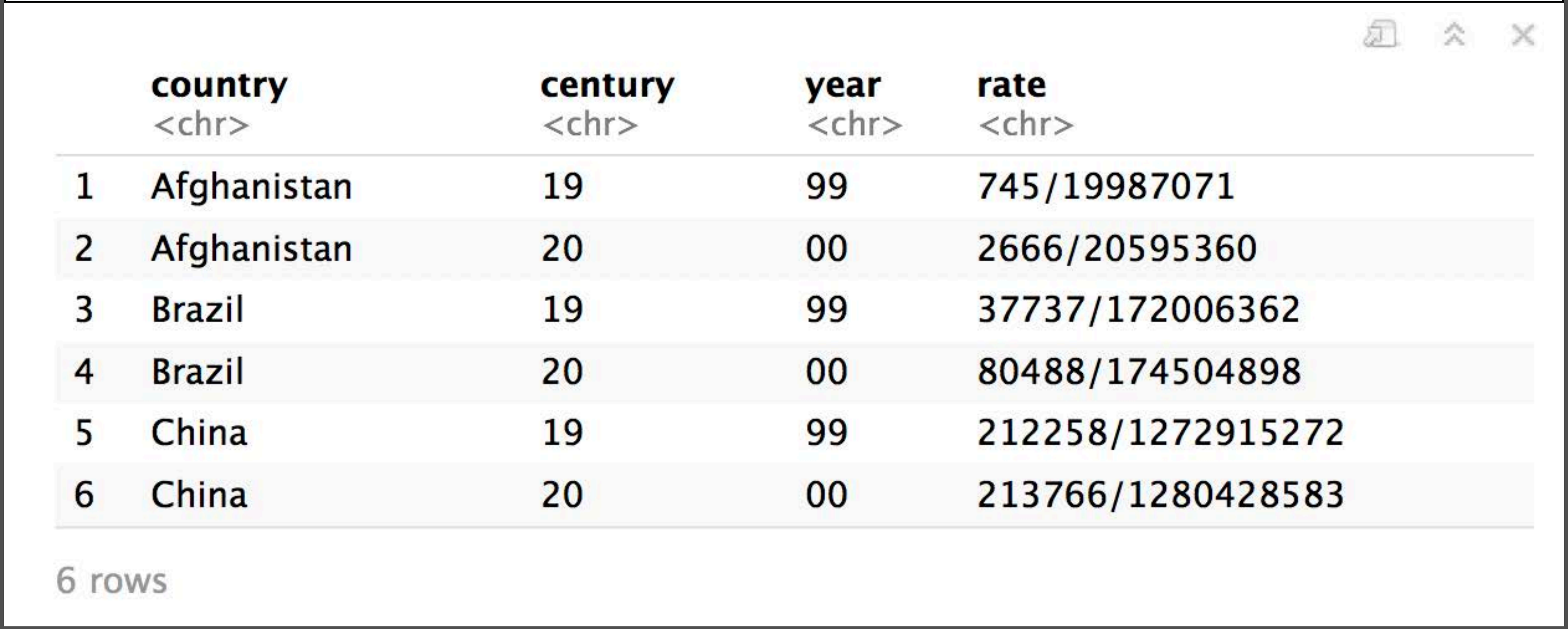
`tidyr::table4b`

	country <chr>	1999 <int>	2000 <int>
1	Afghanistan	19987071	20595360
2	Brazil	172006362	174504898
3	China	1272915272	1280428583

3 rows

OTHER (BAD) IDEAS

```
tidyr::table5
```



	country	century	year	rate
	<chr>	<chr>	<chr>	<chr>
1	Afghanistan	19	99	745/19987071
2	Afghanistan	20	00	2666/20595360
3	Brazil	19	99	37737/172006362
4	Brazil	20	00	80488/174504898
5	China	19	99	212258/1272915272
6	China	20	00	213766/1280428583

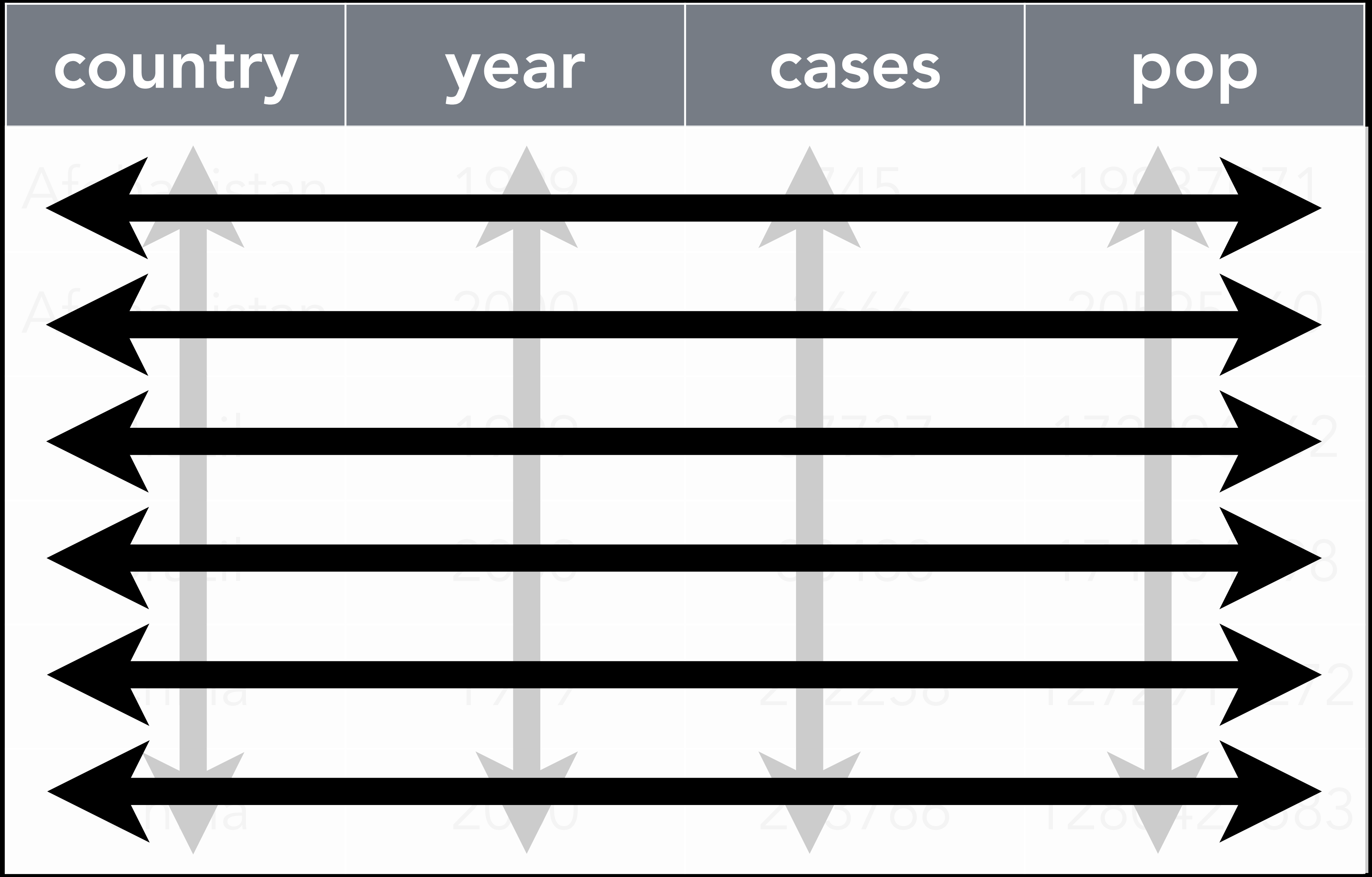
6 rows

TIDY DATA

- Data sets come in **many** different formats.
- Often data are in a format that facilitates data *entry* rather than data *analysis*.
- Most software for scientific computing (including R and SPSS) prefers just **one** format.

A data set is **tidy** if:

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



EXAMPLE: CONTINGENCY TABLE

	Survived	Died
Drug	15	3
Placebo	4	12

Is this tidy?

REORGANIZE TO MAKE IT TIDY

	Survived	Died
Drug	15	3
Placebo	4	12

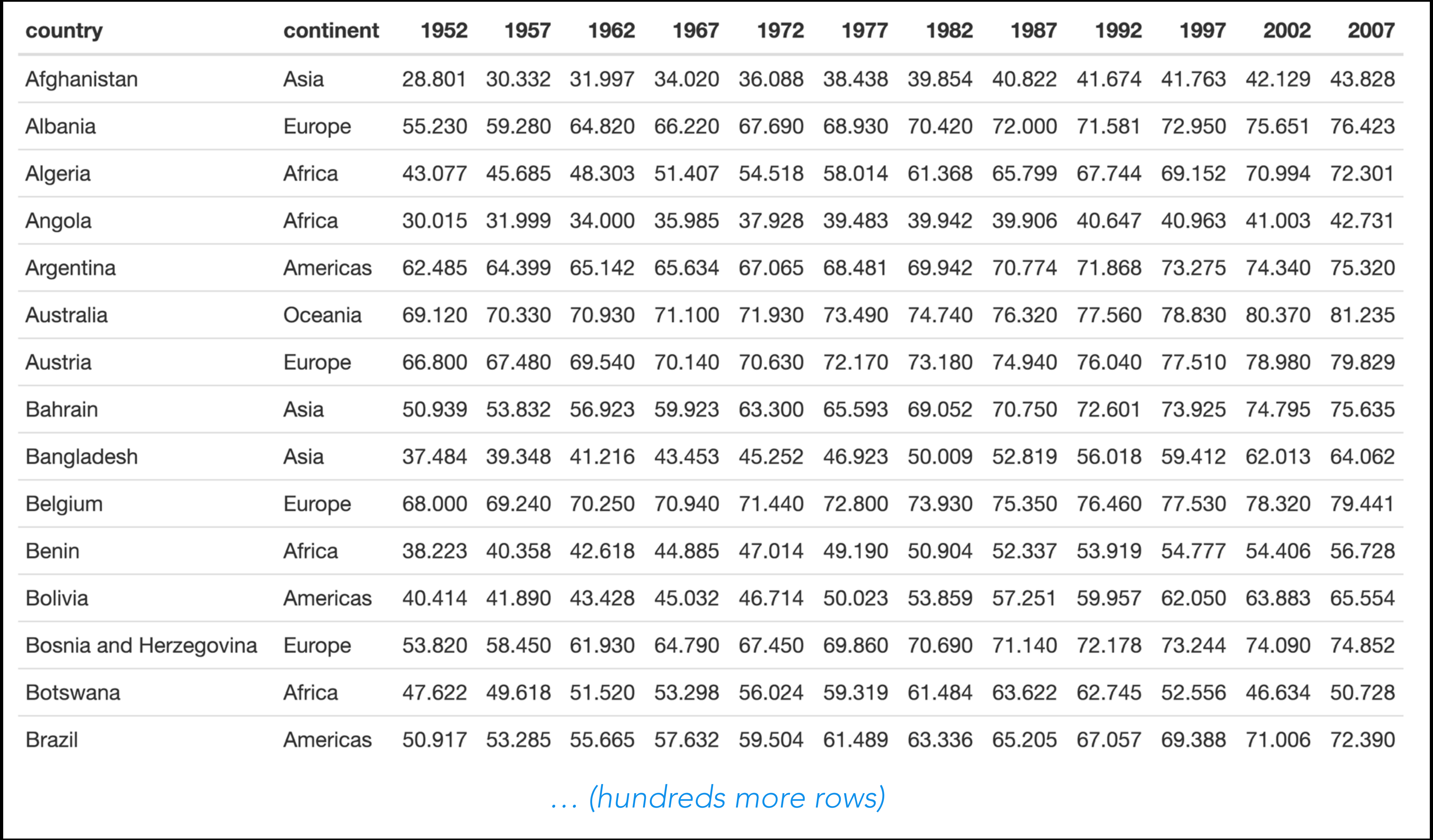
Not tidy

Treatment	Outcome	Count
Drug	Survived	15
Drug	Died	3
Placebo	Survived	4
Placebo	Died	12

Tidy

RESHAPING BY HAND IS NOT ALWAYS SO SIMPLE…

A wide version of gapminder life expectancy data



country	continent	1952	1957	1962	1967	1972	1977	1982	1987	1992	1997	2002	2007
Afghanistan	Asia	28.801	30.332	31.997	34.020	36.088	38.438	39.854	40.822	41.674	41.763	42.129	43.828
Albania	Europe	55.230	59.280	64.820	66.220	67.690	68.930	70.420	72.000	71.581	72.950	75.651	76.423
Algeria	Africa	43.077	45.685	48.303	51.407	54.518	58.014	61.368	65.799	67.744	69.152	70.994	72.301
Angola	Africa	30.015	31.999	34.000	35.985	37.928	39.483	39.942	39.906	40.647	40.963	41.003	42.731
Argentina	Americas	62.485	64.399	65.142	65.634	67.065	68.481	69.942	70.774	71.868	73.275	74.340	75.320
Australia	Oceania	69.120	70.330	70.930	71.100	71.930	73.490	74.740	76.320	77.560	78.830	80.370	81.235
Austria	Europe	66.800	67.480	69.540	70.140	70.630	72.170	73.180	74.940	76.040	77.510	78.980	79.829
Bahrain	Asia	50.939	53.832	56.923	59.923	63.300	65.593	69.052	70.750	72.601	73.925	74.795	75.635
Bangladesh	Asia	37.484	39.348	41.216	43.453	45.252	46.923	50.009	52.819	56.018	59.412	62.013	64.062
Belgium	Europe	68.000	69.240	70.250	70.940	71.440	72.800	73.930	75.350	76.460	77.530	78.320	79.441
Benin	Africa	38.223	40.358	42.618	44.885	47.014	49.190	50.904	52.337	53.919	54.777	54.406	56.728
Bolivia	Americas	40.414	41.890	43.428	45.032	46.714	50.023	53.859	57.251	59.957	62.050	63.883	65.554
Bosnia and Herzegovina	Europe	53.820	58.450	61.930	64.790	67.450	69.860	70.690	71.140	72.178	73.244	74.090	74.852
Botswana	Africa	47.622	49.618	51.520	53.298	56.024	59.319	61.484	63.622	62.745	52.556	46.634	50.728
Brazil	Americas	50.917	53.285	55.665	57.632	59.504	61.489	63.336	65.205	67.057	69.388	71.006	72.390

… (hundreds more rows)

HOW CAN WE CONVERT BETWEEN WIDE AND LONG FORMATS?

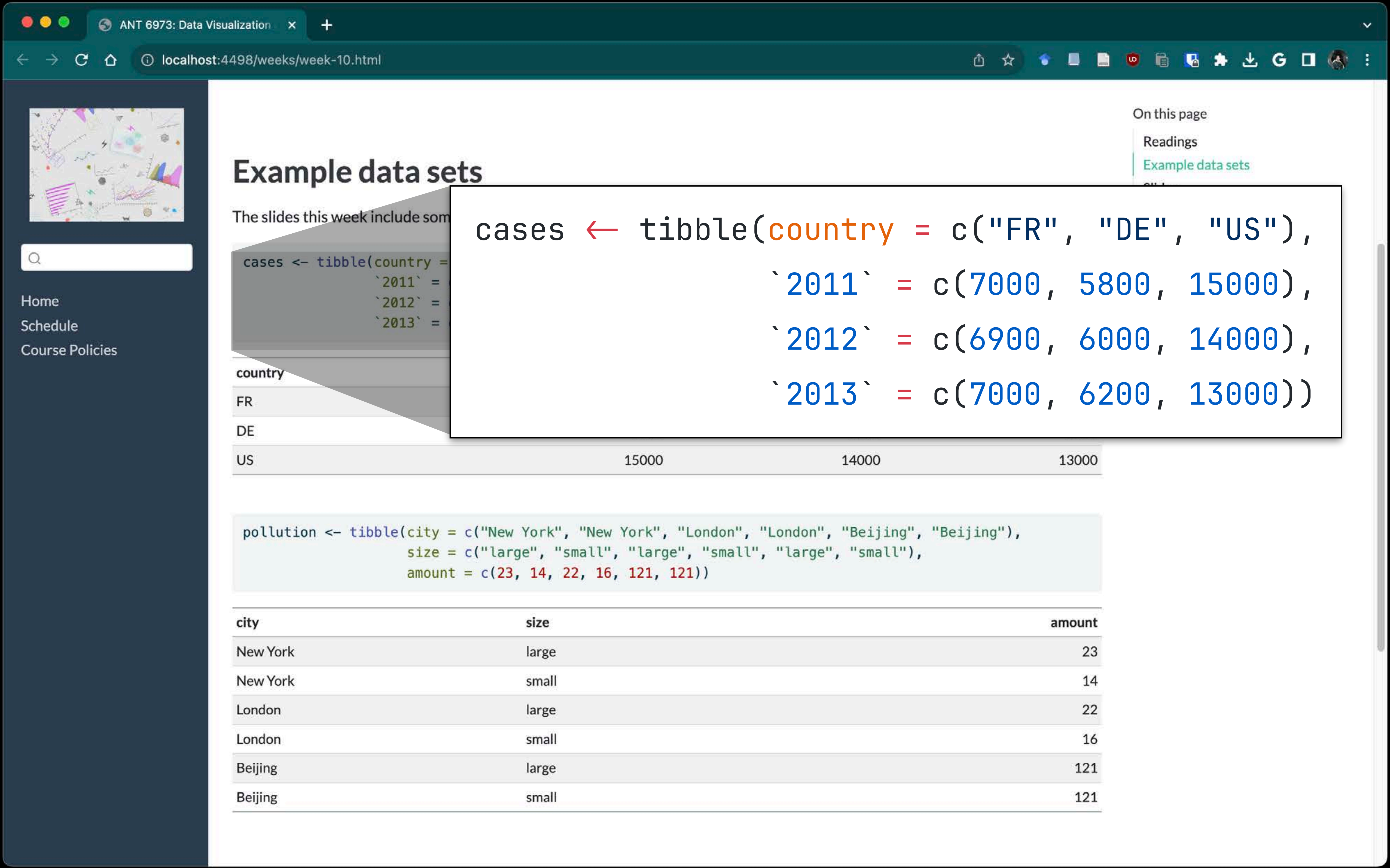
RESHAPE DATA

RESHAPING "VERBS"

- `pivot_longer()`: reshape from wide to long
- `pivot_wider()`: reshape from long to wide

pivot_longer()

PRACTICE DATA



Example data sets

The slides this week include som

```r
cases ← tibble(country = c("FR", "DE", "US"),
               `2011` = c(7000, 5800, 15000),
               `2012` = c(6900, 6000, 14000),
               `2013` = c(7000, 6200, 13000))
```

country			
FR			
DE			
US	15000	14000	13000

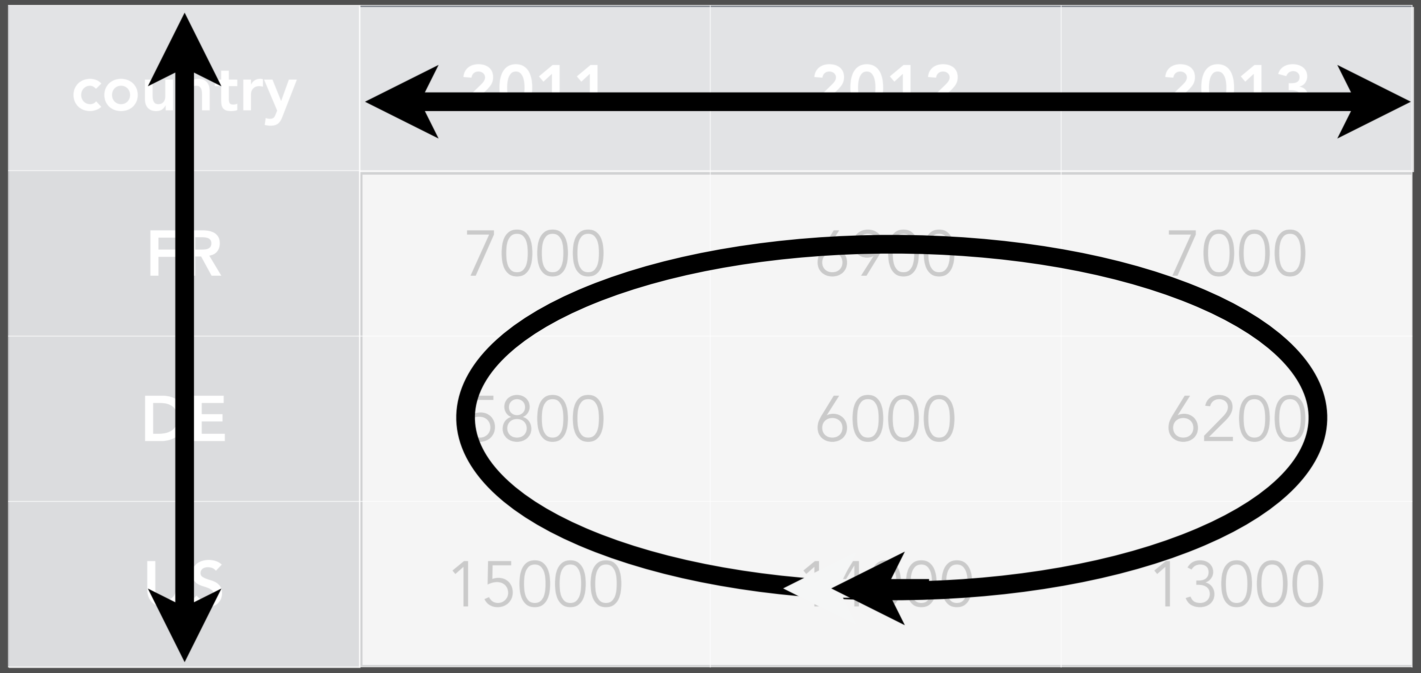
```r
pollution <- tibble(city = c("New York", "New York", "London", "London", "Beijing", "Beijing"),
                    size = c("large", "small", "large", "small", "large", "small"),
                    amount = c(23, 14, 22, 16, 121, 121))
```

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

WHAT ARE THE VARIABLES?

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

WHAT ARE THE VARIABLES?



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

- Country
- Year
- Count

ACTIVITY 1

- Plan how the data would look if it were organized in 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
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
FR	2012	6900
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
FR	2012	6900
FR	2013	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
FR	2012	6900
FR	2013	7000
DE	2011	5800
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
FR	2012	6900
FR	2013	7000
DE	2011	5800
DE	2012	6000
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
FR	2012	6900
FR	2013	7000
DE	2011	5800
DE	2012	6000
DE	2013	6200
US	2011	15000

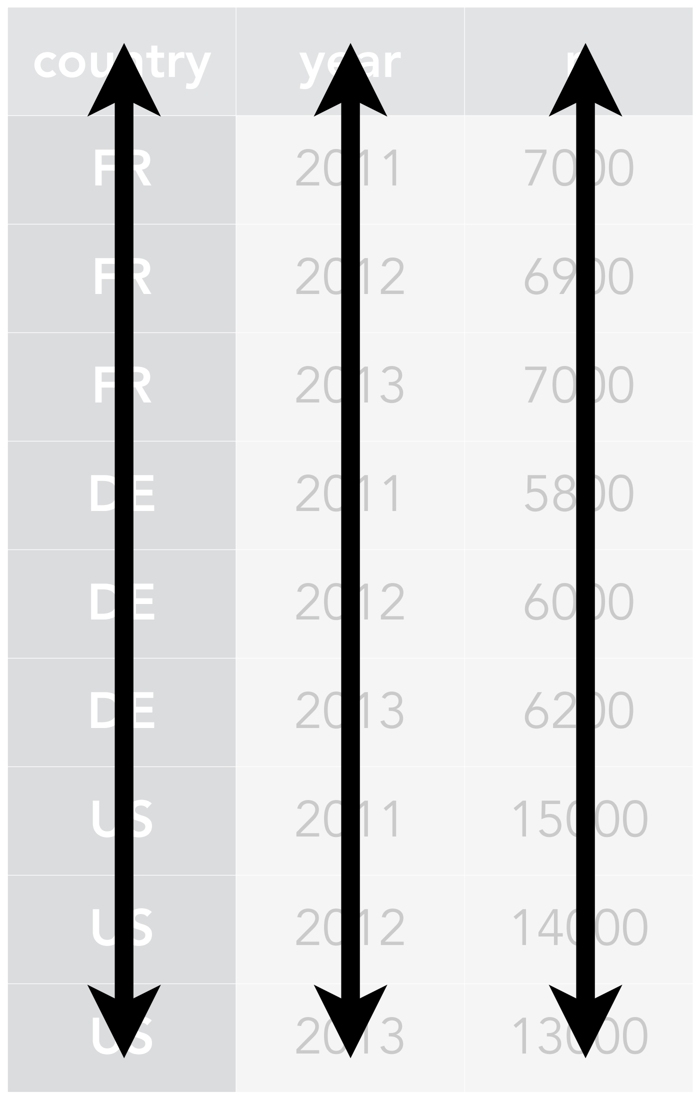
country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

country	year	n
FR	2011	7000
FR	2012	6900
FR	2013	7000
DE	2011	5800
DE	2012	6000
DE	2013	6200
US	2011	15000
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
FR	2012	6900
FR	2013	7000
DE	2011	5800
DE	2012	6000
DE	2013	6200
US	2011	15000
US	2012	14000
US	2013	13000

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



country	year	n
FR	2011	7000
FR	2012	6900
FR	2013	7000
DE	2011	5800
DE	2012	6000
DE	2013	6200
US	2011	15000
US	2012	14000
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
FR	2012	6900
FR	2013	7000
DE	2011	5800
DE	2012	6000
DE	2013	6200
US	2011	15000
US	2012	14000
US	2013	13000

Column names

New variable "year"

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

Transformation logic: column names **TO** new variable

Cell values

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

New variable "n"

(former cell values)

country	year	n
FR	2011	7000
FR	2012	6900
FR	2013	7000
DE	2011	5800
DE	2012	6000
DE	2013	6200
US	2011	15000
US	2012	14000
US	2013	13000

Transformation logic: cell values **TO** new variable

pivot_longer()

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

pivot_longer()

Data to reshape

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

Columns to pivot *(more later)*

Name of new variable for data stored in column names

Name of new variable for data stored in cells values

pivot_longer()

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

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

3 rows

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

9 rows

pivot_longer()

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

Numeric indices

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

3 rows

pivot_longer()

Actual column names

```r
cases ▷
  pivot_longer(cols = c("2011","2012","2013"),
               names_to = "year",
               values_to = "n")
```

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

3 rows

pivot_longer()

```r
cases ▷
  pivot_longer(cols = -country,
               names_to = "year",
               values_to = "n")
```

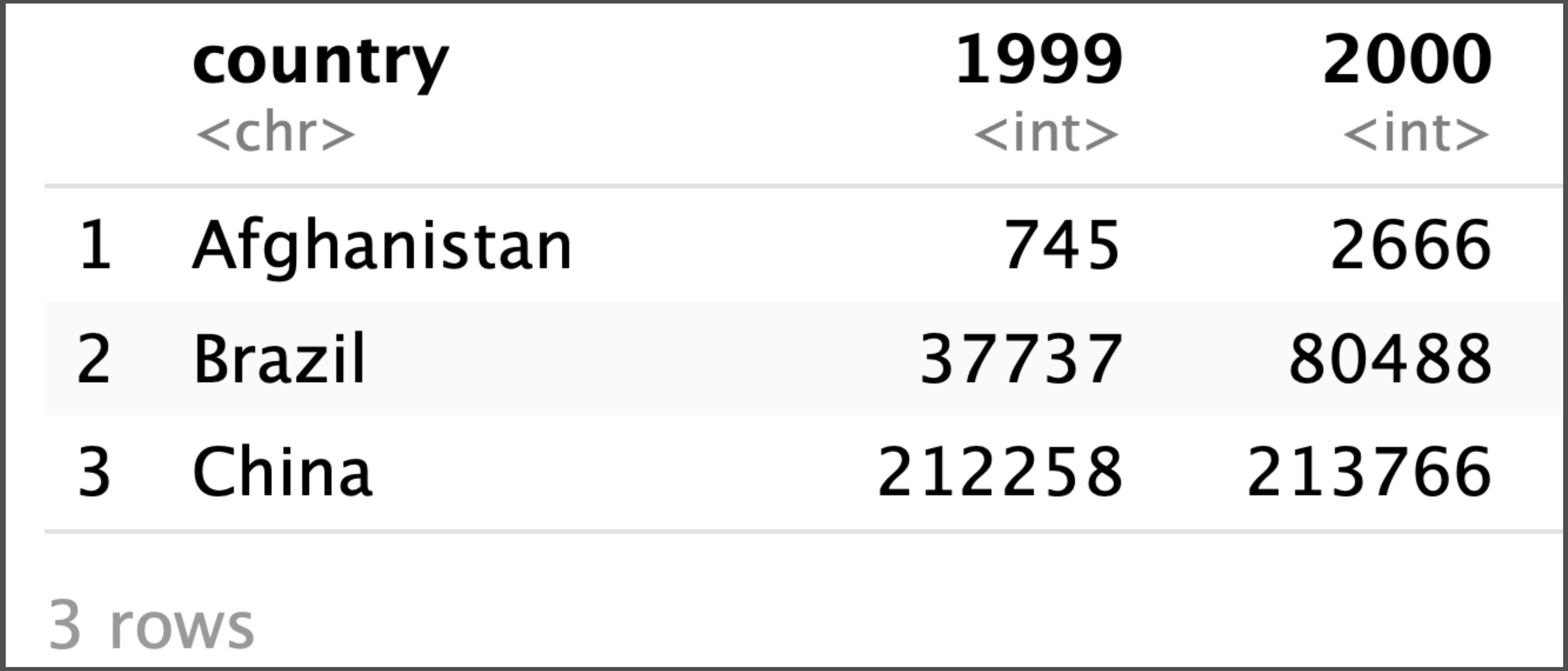
Everything except...

country	Not country 2011	Not country 2012	Not country 2013
<chr>	<dbl>	<dbl>	<dbl>
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

3 rows

ACTIVITY 2

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

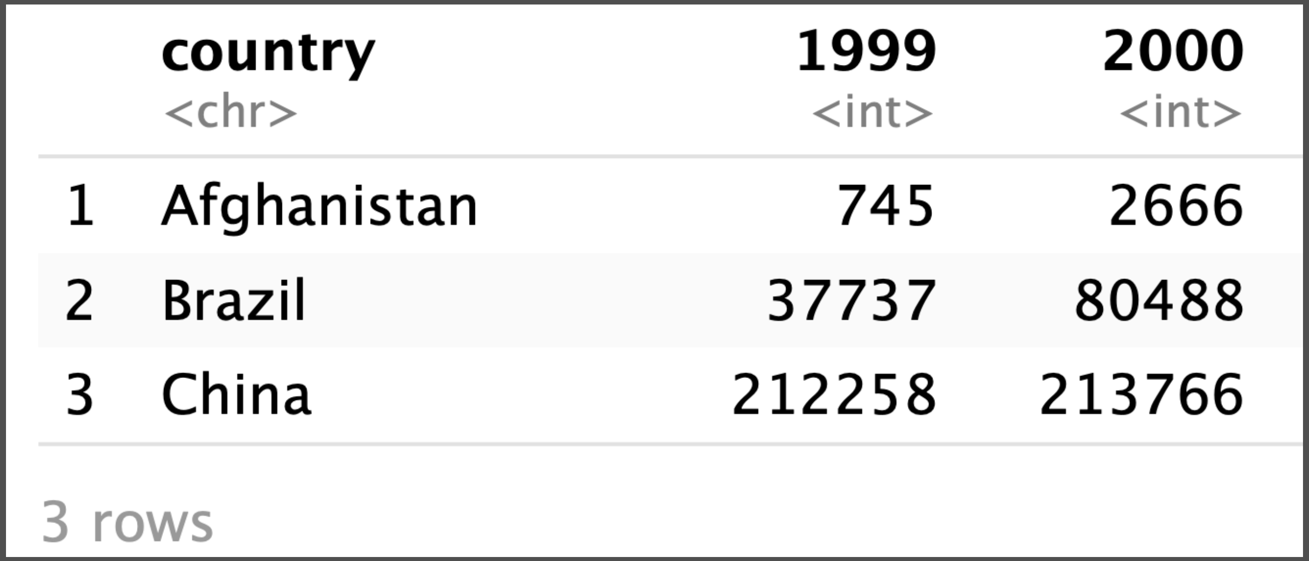


	country	1999	2000
	<chr>	<int>	<int>
1	Afghanistan	745	2666
2	Brazil	37737	80488
3	China	212258	213766

3 rows

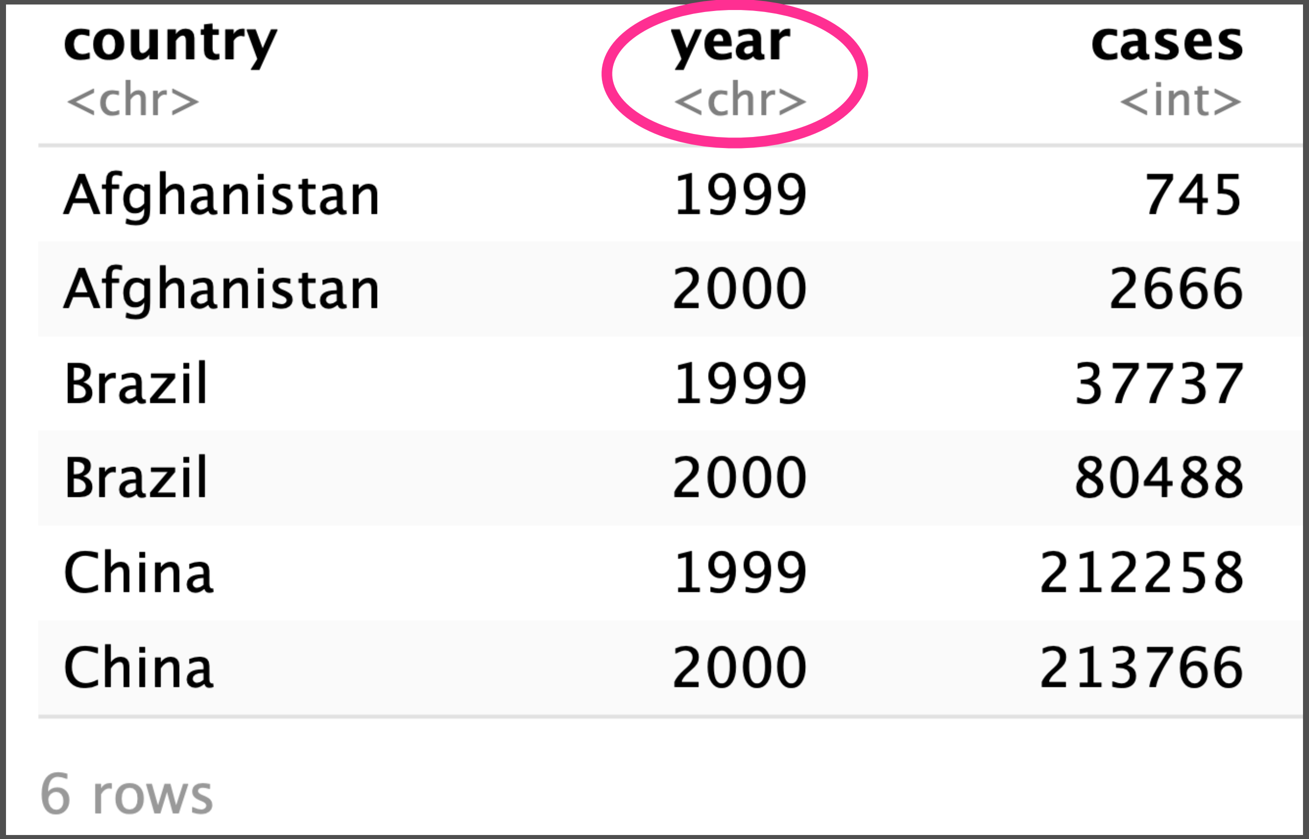
SOLUTION

```
table4a |>
  pivot_longer(cols = 2:3, names_to = "year", values_to = "cases")
```



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

3 rows

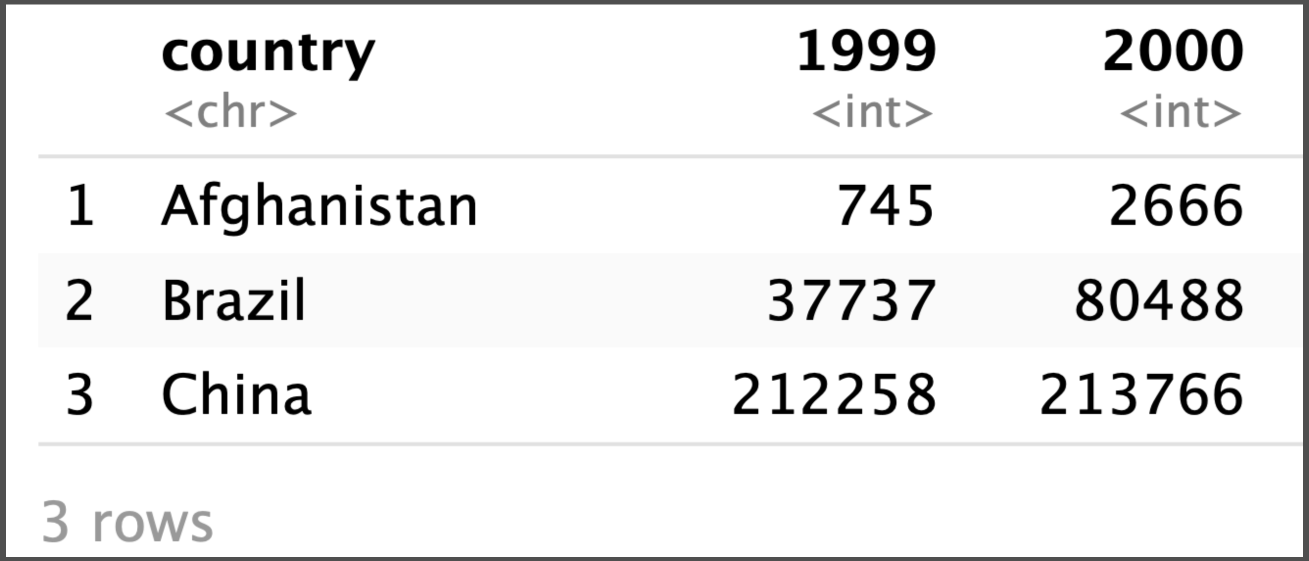


country <chr>	year <chr>	cases <int>
Afghanistan	1999	745
Afghanistan	2000	2666
Brazil	1999	37737
Brazil	2000	80488
China	1999	212258
China	2000	213766

6 rows

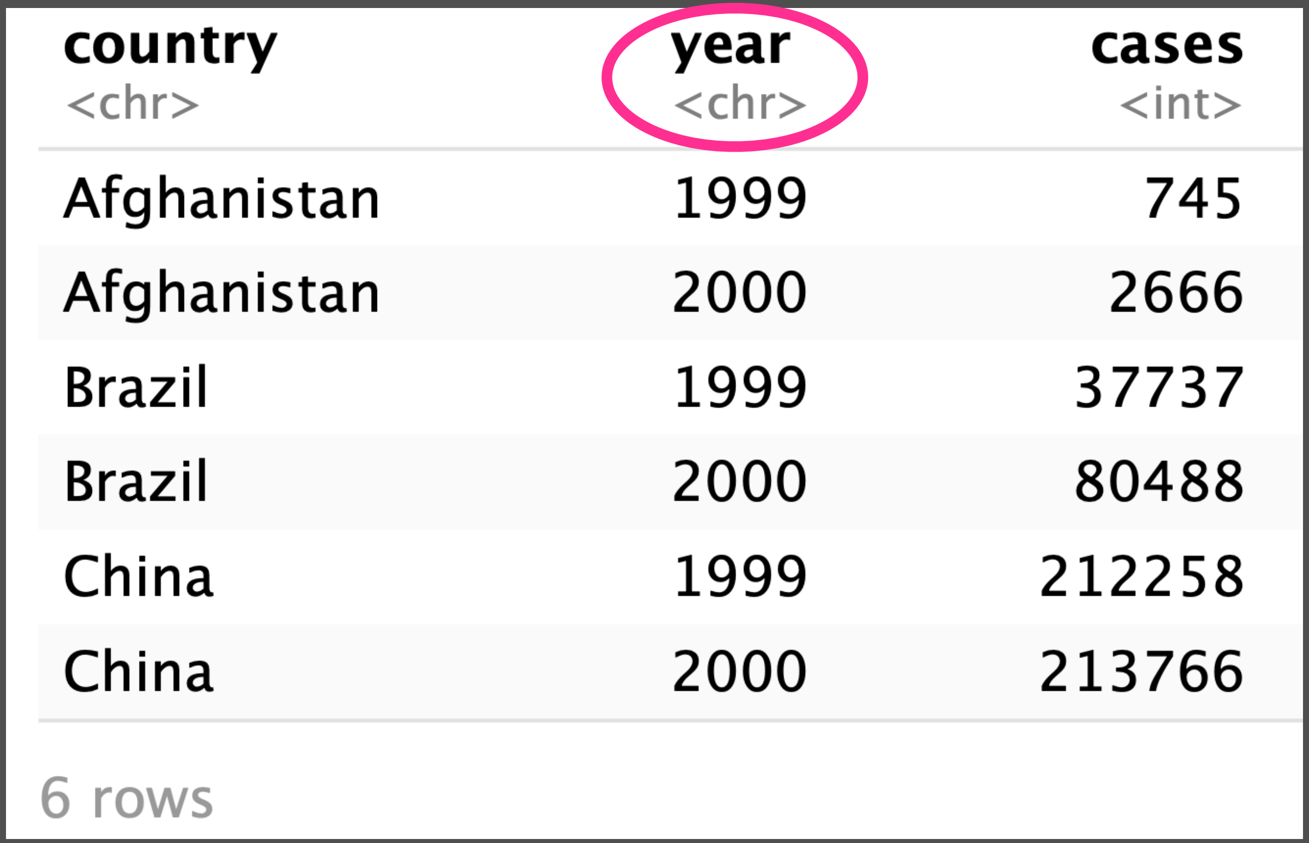
SOLUTION

```
table4a |>
  pivot_longer(cols = 2:3, names_to = "year", values_to = "cases",
                              names_transform = as.integer)
```



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

3 rows

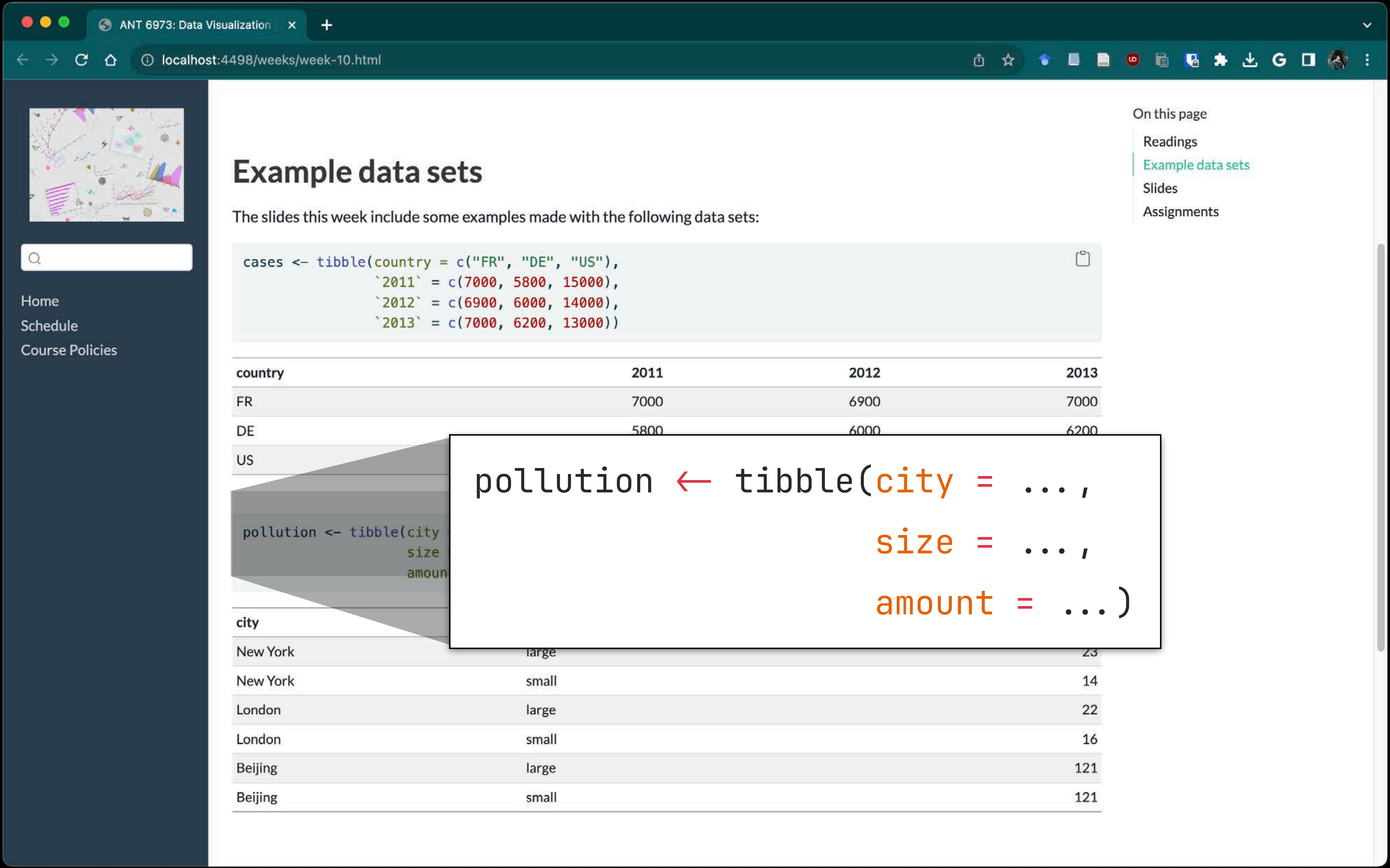


country	year	cases
<chr>	<chr>	<int>
Afghanistan	1999	745
Afghanistan	2000	2666
Brazil	1999	37737
Brazil	2000	80488
China	1999	212258
China	2000	213766

6 rows

pivot_wider()

PRACTICE DATA



Example data sets

The slides this week include some examples made with the following data sets:

```r
cases <- tibble(country = c("FR", "DE", "US"),
                `2011` = c(7000, 5800, 15000),
                `2012` = c(6900, 6000, 14000),
                `2013` = c(7000, 6200, 13000))
```

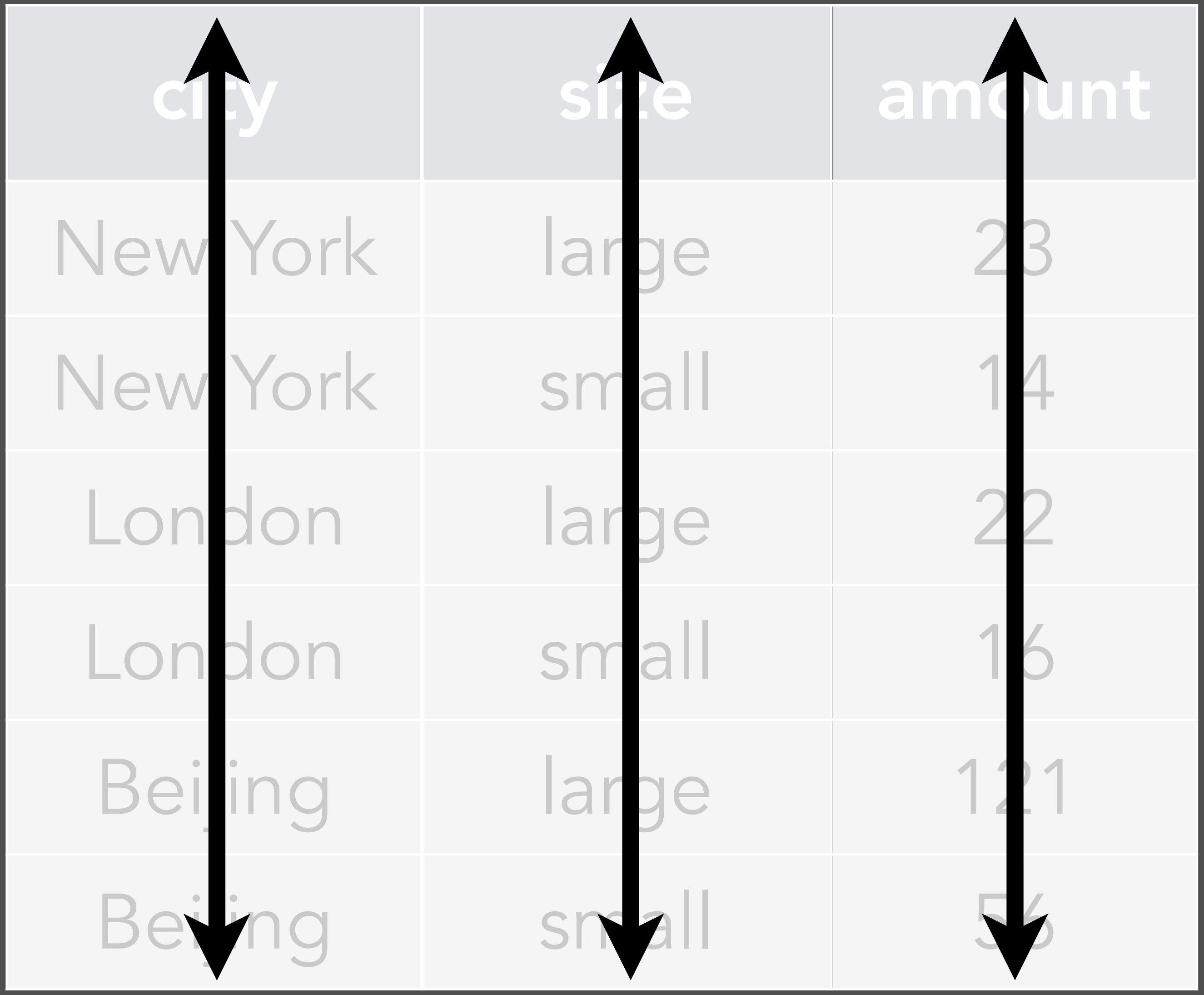
country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US			

```r
pollution <- tibble(city
                    size
                    amoun
```

```r
pollution ← tibble(city = ...,
                   size = ...,
                   amount = ...)
```

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

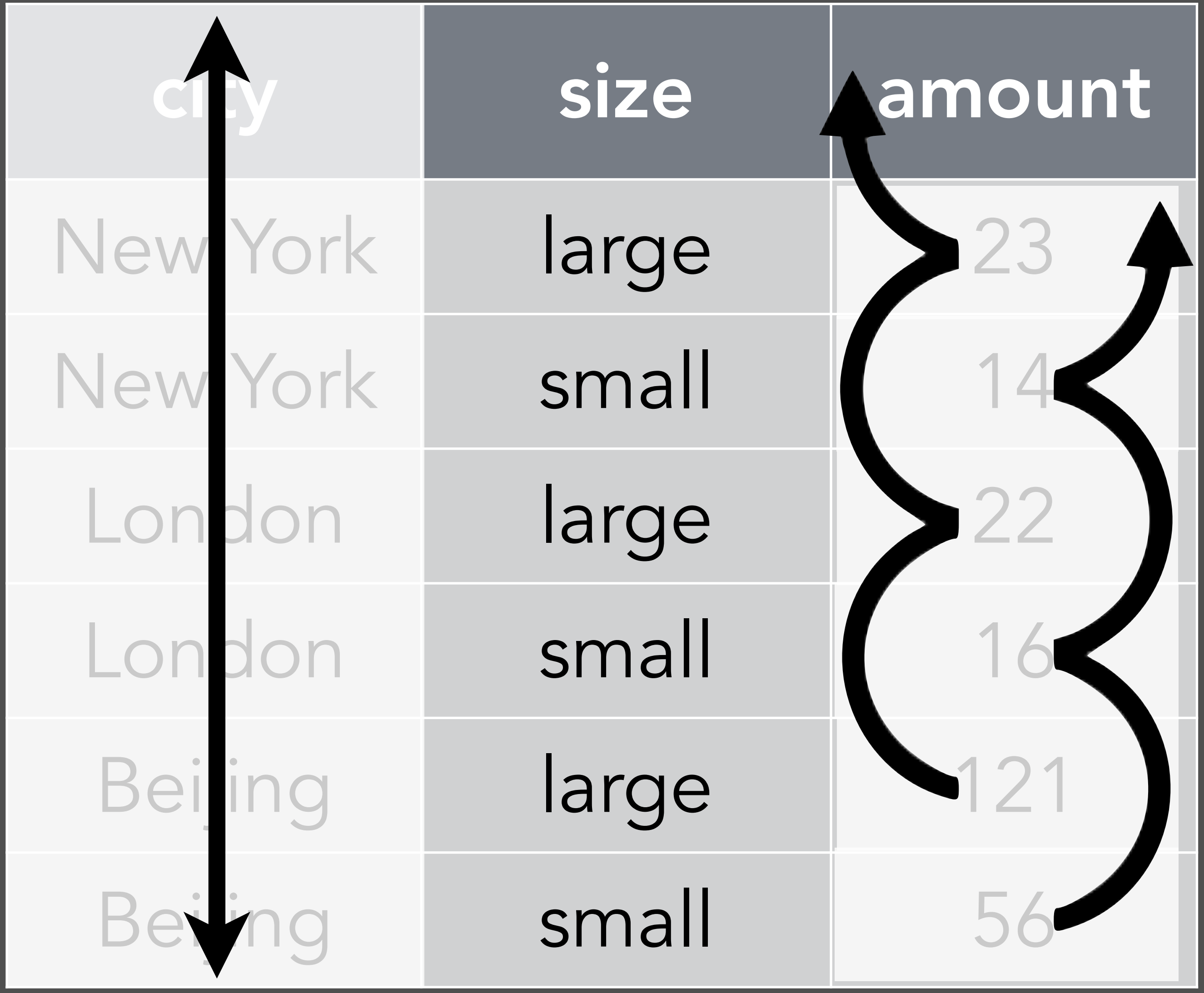
WHAT ARE THE VARIABLES?



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

- City
- Particle size
- Amount of particulate

To make a scatter plot of of large vs. small particle amounts, what columns would we need?



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

- City
- Amount of large
- Amount of small

What is a variable and an observation may depend on your immediate goal!

ACTIVITY 3

- Plan 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

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

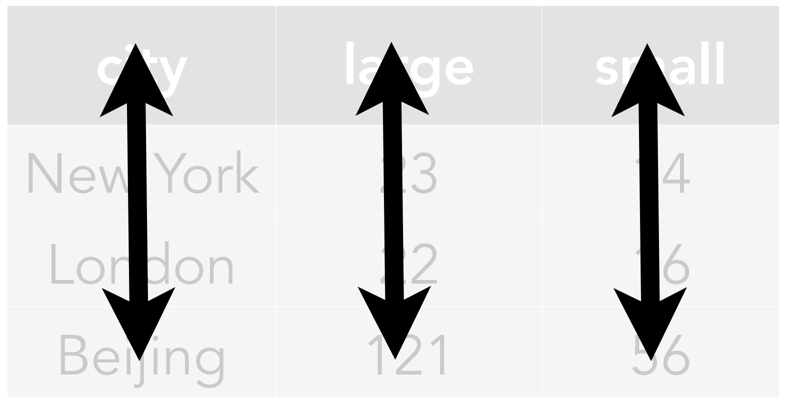
city	size	amount
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

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

Variable with 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

New columns

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

Transformation logic: column names **FROM** old variable

Variable with new cell values

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

New cell values

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

Transformation logic: cell values **FROM** old variable

pivot_wider()

```
pollution ▷
  pivot_wider(names_from = size,
              values_from = amount)
```

pivot_wider()

Data to reshape

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

Old variable that becomes new column names

Old variable that becomes new cell values

pivot_wider()

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

city <chr>	size <chr>	amount <dbl>
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	121

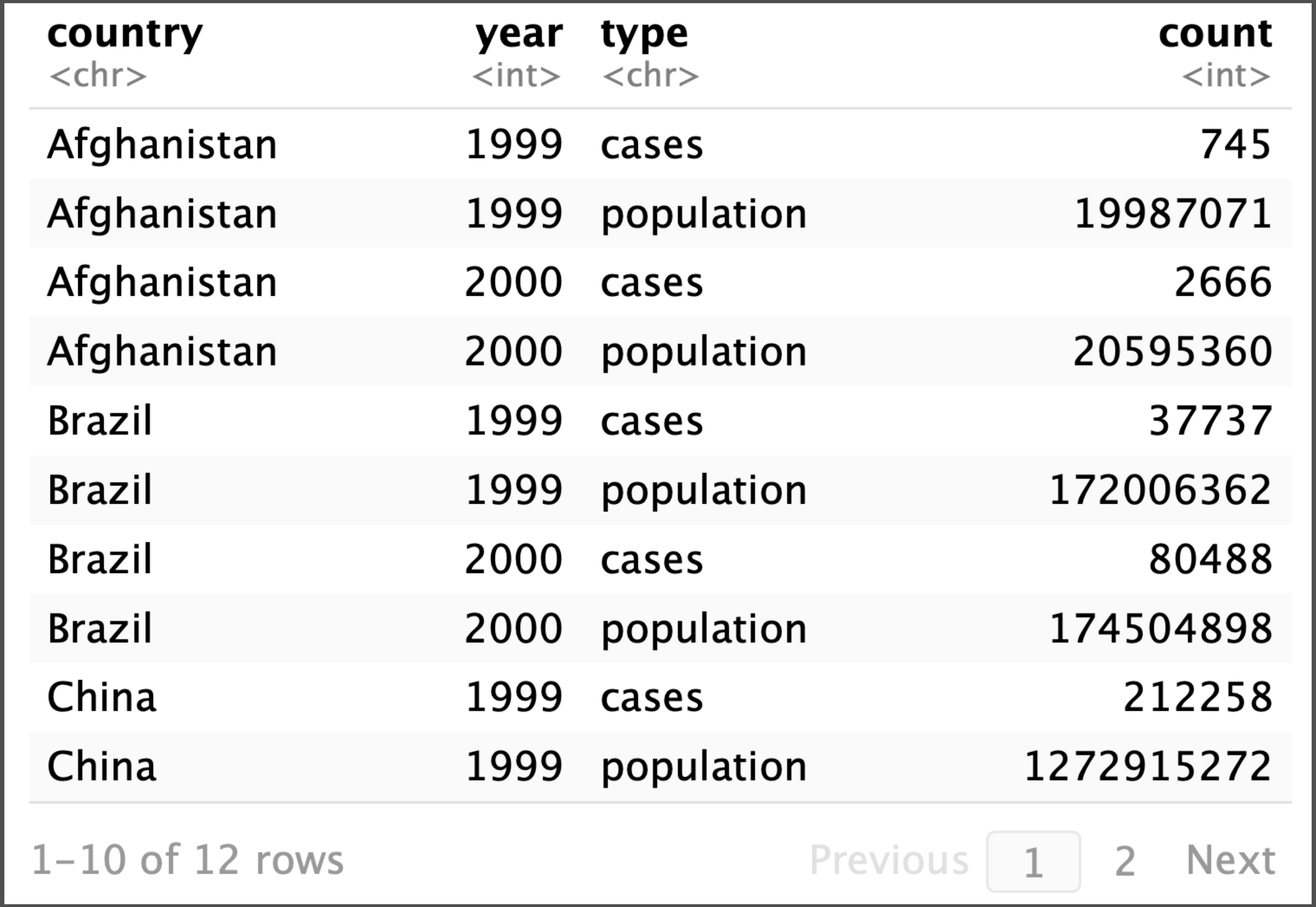
6 rows

city <chr>	large <dbl>	small <dbl>
New York	23	14
London	22	16
Beijing	121	121

3 rows

ACTIVITY 4

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

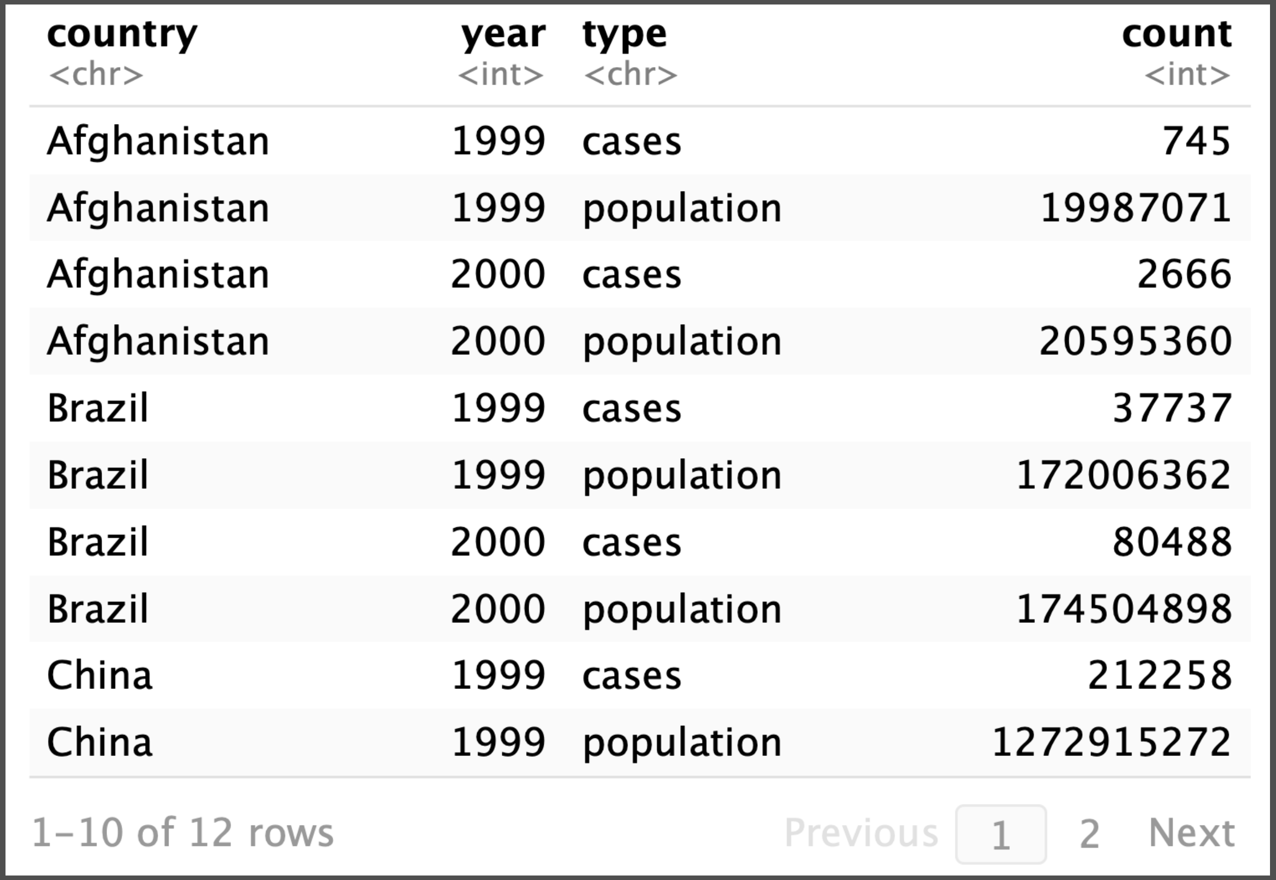


country <chr>	year <int>	type <chr>	count <int>
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999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

1–10 of 12 rows

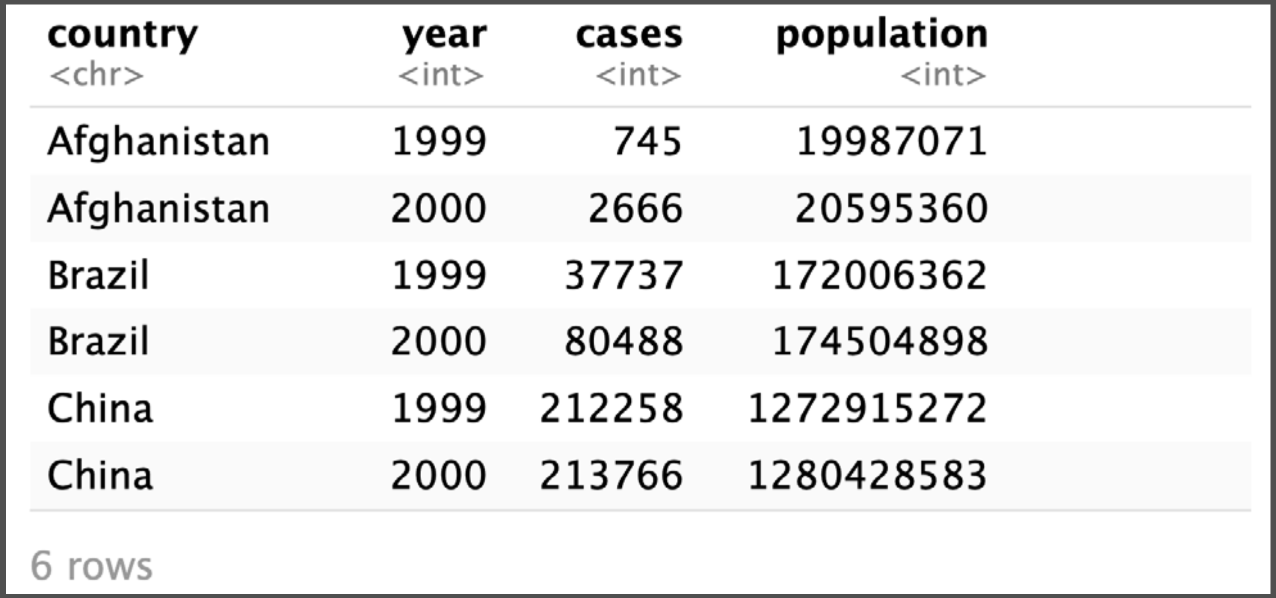
SOLUTION

```
table2 |>
  pivot_wider(names_from = type, values_from = count)
```



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
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272

1–10 of 12 rows Previous 1 2 Next

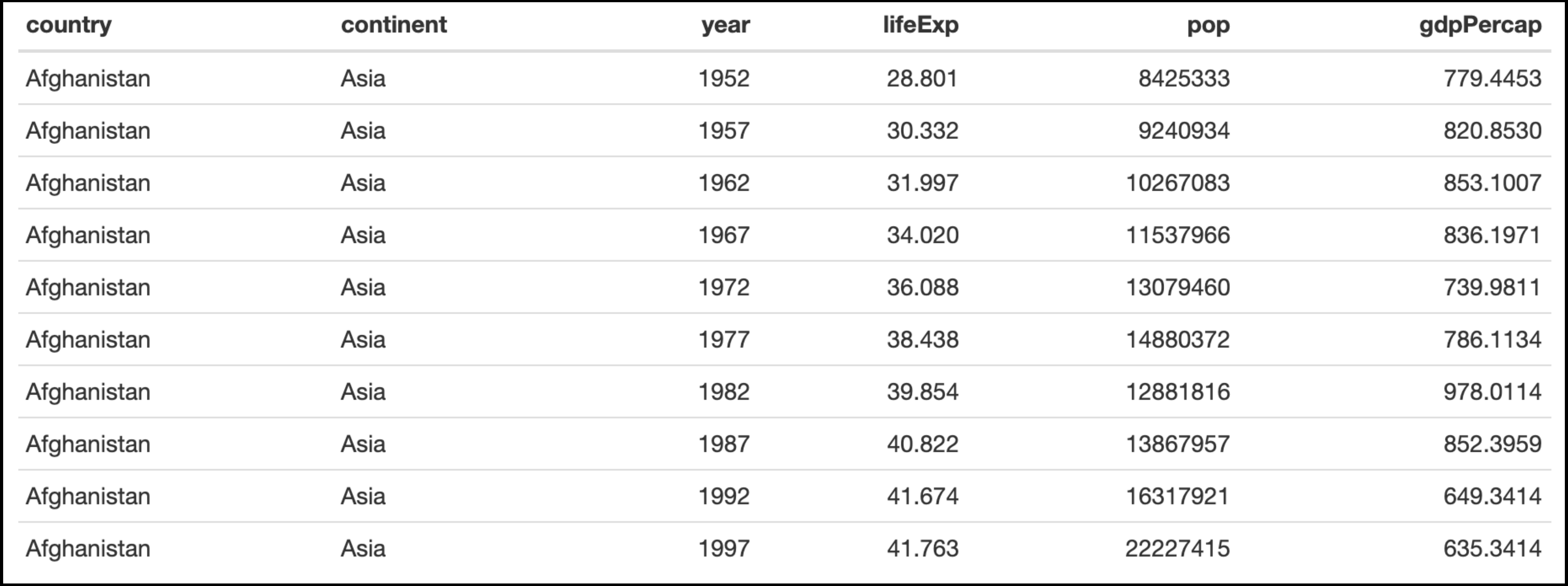


country <chr>	year <int>	cases <int>	population <int>
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

6 rows

GAPMINDER

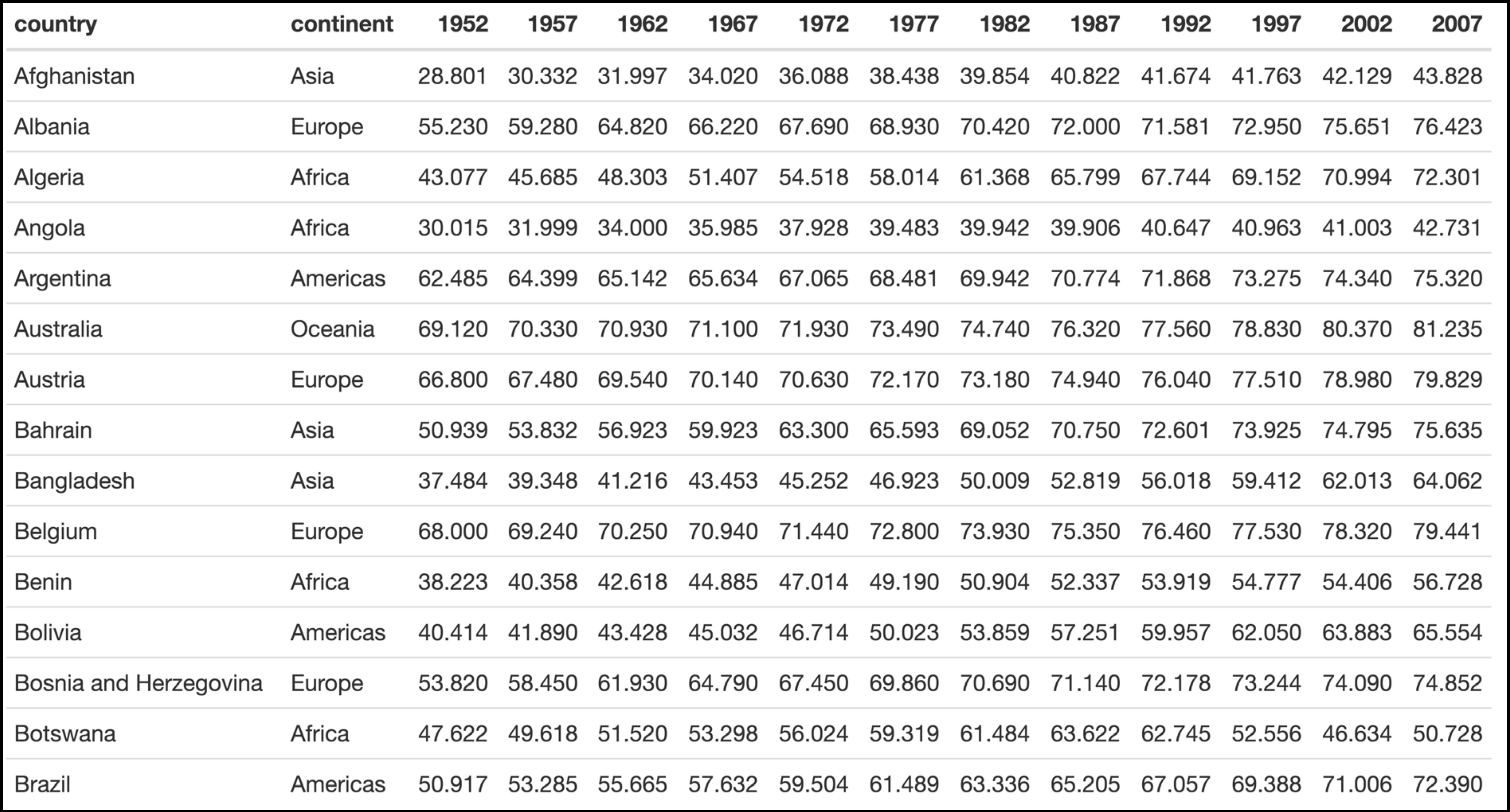
```
gapminder
```



country	continent	year	lifeExp	pop	gdpPercap
Afghanistan	Asia	1952	28.801	8425333	779.4453
Afghanistan	Asia	1957	30.332	9240934	820.8530
Afghanistan	Asia	1962	31.997	10267083	853.1007
Afghanistan	Asia	1967	34.020	11537966	836.1971
Afghanistan	Asia	1972	36.088	13079460	739.9811
Afghanistan	Asia	1977	38.438	14880372	786.1134
Afghanistan	Asia	1982	39.854	12881816	978.0114
Afghanistan	Asia	1987	40.822	13867957	852.3959
Afghanistan	Asia	1992	41.674	16317921	649.3414
Afghanistan	Asia	1997	41.763	22227415	635.3414

GAPMINDER

Our target:



country	continent	1952	1957	1962	1967	1972	1977	1982	1987	1992	1997	2002	2007
Afghanistan	Asia	28.801	30.332	31.997	34.020	36.088	38.438	39.854	40.822	41.674	41.763	42.129	43.828
Albania	Europe	55.230	59.280	64.820	66.220	67.690	68.930	70.420	72.000	71.581	72.950	75.651	76.423
Algeria	Africa	43.077	45.685	48.303	51.407	54.518	58.014	61.368	65.799	67.744	69.152	70.994	72.301
Angola	Africa	30.015	31.999	34.000	35.985	37.928	39.483	39.942	39.906	40.647	40.963	41.003	42.731
Argentina	Americas	62.485	64.399	65.142	65.634	67.065	68.481	69.942	70.774	71.868	73.275	74.340	75.320
Australia	Oceania	69.120	70.330	70.930	71.100	71.930	73.490	74.740	76.320	77.560	78.830	80.370	81.235
Austria	Europe	66.800	67.480	69.540	70.140	70.630	72.170	73.180	74.940	76.040	77.510	78.980	79.829
Bahrain	Asia	50.939	53.832	56.923	59.923	63.300	65.593	69.052	70.750	72.601	73.925	74.795	75.635
Bangladesh	Asia	37.484	39.348	41.216	43.453	45.252	46.923	50.009	52.819	56.018	59.412	62.013	64.062
Belgium	Europe	68.000	69.240	70.250	70.940	71.440	72.800	73.930	75.350	76.460	77.530	78.320	79.441
Benin	Africa	38.223	40.358	42.618	44.885	47.014	49.190	50.904	52.337	53.919	54.777	54.406	56.728
Bolivia	Americas	40.414	41.890	43.428	45.032	46.714	50.023	53.859	57.251	59.957	62.050	63.883	65.554
Bosnia and Herzegovina	Europe	53.820	58.450	61.930	64.790	67.450	69.860	70.690	71.140	72.178	73.244	74.090	74.852
Botswana	Africa	47.622	49.618	51.520	53.298	56.024	59.319	61.484	63.622	62.745	52.556	46.634	50.728
Brazil	Americas	50.917	53.285	55.665	57.632	59.504	61.489	63.336	65.205	67.057	69.388	71.006	72.390

GAPMINDER

country	continent	year	lifeExp	pop	gdpPercap
Afghanistan	Asia	1952	28.801	8425333	779.4453
Afghanistan	Asia	1957	30.332	9240934	820.8530
Afghanistan	Asia	1962	31.997	10267083	853.1007
Afghanistan	Asia	1967	34.020	11537966	836.1971
Afghanistan	Asia	1972	36.088	13079460	739.9811
Afghanistan	Asia	1977	38.438	14880372	786.1134
Afghanistan	Asia	1982	39.854	12881816	978.0114
Afghanistan	Asia	1987	40.822	13867957	852.3959
Afghanistan	Asia	1992	41.674	16317921	649.3414
Afghanistan	Asia	1997	41.763	22227415	635.3414

country	continent	1952	1957	1962	1967	1972	1977	1982	1987	1992	1997	2002	2007
Afghanistan	Asia	28.801	30.332	31.997	34.020	36.088	38.438	39.854	40.822	41.674	41.763	42.129	43.828
Albania	Europe	55.230	59.280	64.820	66.220	67.690	68.930	70.420	72.000	71.581	72.950	75.651	76.423
Algeria	Africa	43.077	45.685	48.303	51.407	54.518	58.014	61.368	65.799	67.744	69.152	70.994	72.301
Angola	Africa	30.015	31.999	34.000	35.985	37.928	39.483	39.942	39.906	40.647	40.963	41.003	42.731
Argentina	Americas	62.485	64.399	65.142	65.634	67.065	68.481	69.942	70.774	71.868	73.275	74.340	75.320
Australia	Oceania	69.120	70.330	70.930	71.100	71.930	73.490	74.740	76.320	77.560	78.830	80.370	81.235
Austria	Europe	66.800	67.480	69.540	70.140	70.630	72.170	73.180	74.940	76.040	77.510	78.980	79.829
Bahrain	Asia	50.939	53.832	56.923	59.923	63.300	65.593	69.052	70.750	72.601	73.925	74.795	75.635
Bangladesh	Asia	37.484	39.348	41.216	43.453	45.252	46.923	50.009	52.819	56.018	59.412	62.013	64.062
Belgium	Europe	68.000	69.240	70.250	70.940	71.440	72.800	73.930	75.350	76.460	77.530	78.320	79.441
Benin	Africa	38.223	40.358	42.618	44.885	47.014	49.190	50.904	52.337	53.919	54.777	54.406	56.728
Bolivia	Americas	40.414	41.890	43.428	45.032	46.714	50.023	53.859	57.251	59.957	62.050	63.883	65.554
Bosnia and Herzegovina	Europe	53.820	58.450	61.930	64.790	67.450	69.860	70.690	71.140	72.178	73.244	74.090	74.852
Botswana	Africa	47.622	49.618	51.520	53.298	56.024	59.319	61.484	63.622	62.745	52.556	46.634	50.728
Brazil	Americas	50.917	53.285	55.665	57.632	59.504	61.489	63.336	65.205	67.057	69.388	71.006	72.390

Which column do we get our names from?

Which column do we get our values from?

GAPMINDER

```
gapminder |>
  pivot_wider(names_from = year, values_from = lifeExp)
```

Why is this happening?

country	continent	pop	gdpPercap	1952	1957	1962	1967	1972	197						2007
Afghanistan	Asia	8425333	779.4453	28.801	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Afghanistan	Asia	9240934	820.8530	NA	30.332	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Afghanistan	Asia	10267083	853.1007	NA	NA	31.997	NA	NA	NA	NA	NA	NA	NA	NA	NA
Afghanistan	Asia	11537966	836.1971	NA	NA	NA	34.02	NA	NA	NA	NA	NA	NA	NA	NA
Afghanistan	Asia	13079460	739.9811	NA	NA	NA	NA	36.088	NA	NA	NA	NA	NA	NA	NA
Afghanistan	Asia	14880372	786.1134	NA	NA	NA	NA	NA	38.438	NA	NA	NA	NA	NA	NA
Afghanistan	Asia	12881816	978.0114	NA	NA	NA	NA	NA	NA	39.854	NA	NA	NA	NA	NA
Afghanistan	Asia	13867957	852.3959	NA	NA	NA	NA	NA	NA	NA	40.822	NA	NA	NA	NA
Afghanistan	Asia	16317921	649.3414	NA	NA	NA	NA	NA	NA	NA	NA	41.674	NA	NA	NA
Afghanistan	Asia	22227415	635.3414	NA	NA	NA	NA	NA	NA	NA	NA	NA	41.763	NA	NA

GAPMINDER

country	continent	year	lifeExp	pop	gdpPercap
Afghanistan	Asia	1952	28.801	8425333	779.4453
Afghanistan	Asia	1957	30.332	9240934	820.8530
Afghanistan	Asia	1962	31.997	10267083	853.1007
Afghanistan	Asia	1967	34.020	11537966	836.1971
Afghanistan	Asia	1972	36.088	13079460	739.9811
Afghanistan	Asia	1977	38.438	14880372	786.1134
Afghanistan	Asia	1982	39.854	12881816	978.0114
Afghanistan	Asia	1987	40.822	13867957	852.3959
Afghanistan	Asia	1992	41.674	16317921	649.3414
Afghanistan	Asia	1997	41.763	22227415	635.3414

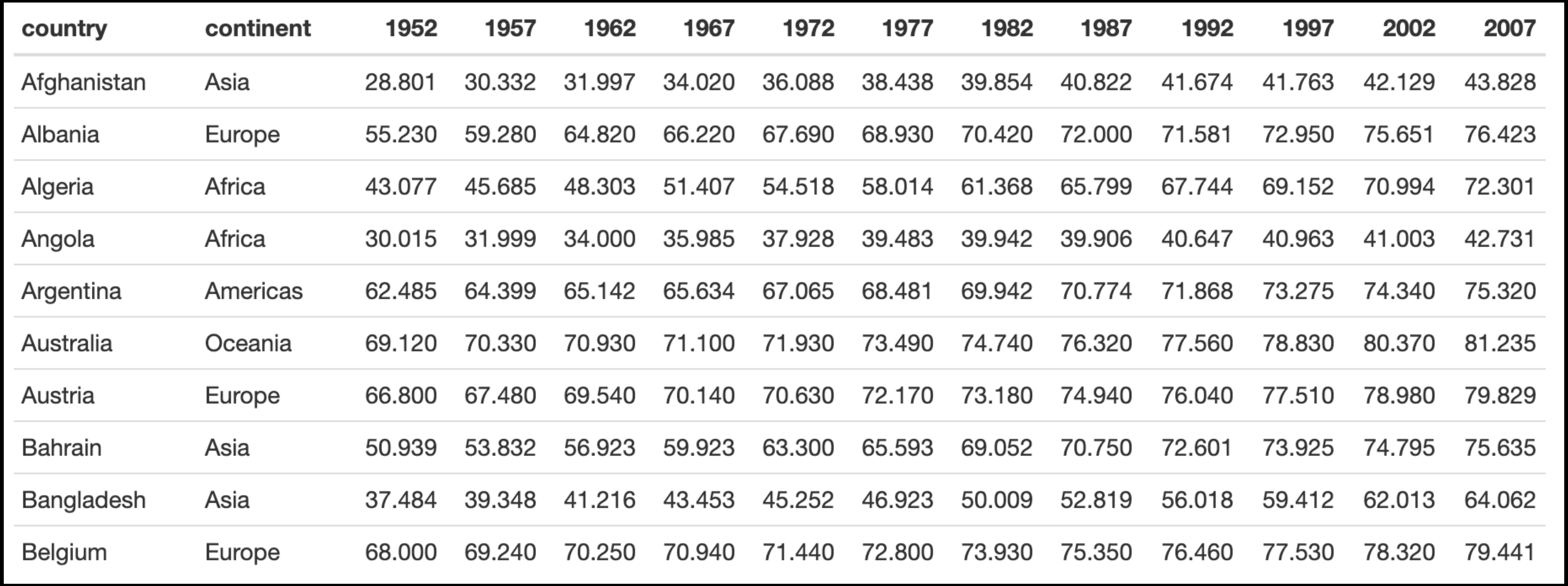
country	continent	1952	1957	1962	1967	1972	1977	1982	1987	1992	1997	2002	2007
Afghanistan	Asia	28.801	30.332	31.997	34.020	36.088	38.438	39.854	40.822	41.674	41.763	42.129	43.828
Albania	Europe	55.230	59.280	64.820	66.220	67.690	68.930	70.420	72.000	71.581	72.950	75.651	76.423
Algeria	Africa	43.077	45.685	48.303	51.407	54.518	58.014	61.368	65.799	67.744	69.152	70.994	72.301
Angola	Africa	30.015	31.999	34.000	35.985	37.928	39.483	39.942	39.906	40.647	40.963	41.003	42.731
Argentina	Americas	62.485	64.399	65.142	65.634	67.065	68.481	69.942	70.774	71.868	73.275	74.340	75.320
Australia	Oceania	69.120	70.330	70.930	71.100	71.930	73.490	74.740	76.320	77.560	78.830	80.370	81.235
Austria	Europe	66.800	67.480	69.540	70.140	70.630	72.170	73.180	74.940	76.040	77.510	78.980	79.829
Bahrain	Asia	50.939	53.832	56.923	59.923	63.300	65.593	69.052	70.750	72.601	73.925	74.795	75.635
Bangladesh	Asia	37.484	39.348	41.216	43.453	45.252	46.923	50.009	52.819	56.018	59.412	62.013	64.062
Belgium	Europe	68.000	69.240	70.250	70.940	71.440	72.800	73.930	75.350	76.460	77.530	78.320	79.441
Benin	Africa	38.223	40.358	42.618	44.885	47.014	49.190	50.904	52.337	53.919	54.777	54.406	56.728
Bolivia	Americas	40.414	41.890	43.428	45.032	46.714	50.023	53.859	57.251	59.957	62.050	63.883	65.554
Bosnia and Herzegovina	Europe	53.820	58.450	61.930	64.790	67.450	69.860	70.690	71.140	72.178	73.244	74.090	74.852
Botswana	Africa	47.622	49.618	51.520	53.298	56.024	59.319	61.484	63.622	62.745	52.556	46.634	50.728
Brazil	Americas	50.917	53.285	55.665	57.632	59.504	61.489	63.336	65.205	67.057	69.388	71.006	72.390

Which columns *uniquely identify* each observation (row) that we want to keep?

GAPMINDER

Columns that uniquely identify each observation

```
gapminder ▷
  pivot_wider(id_cols = c(country, continent),
              names_from = year, values_from = lifeExp)
```



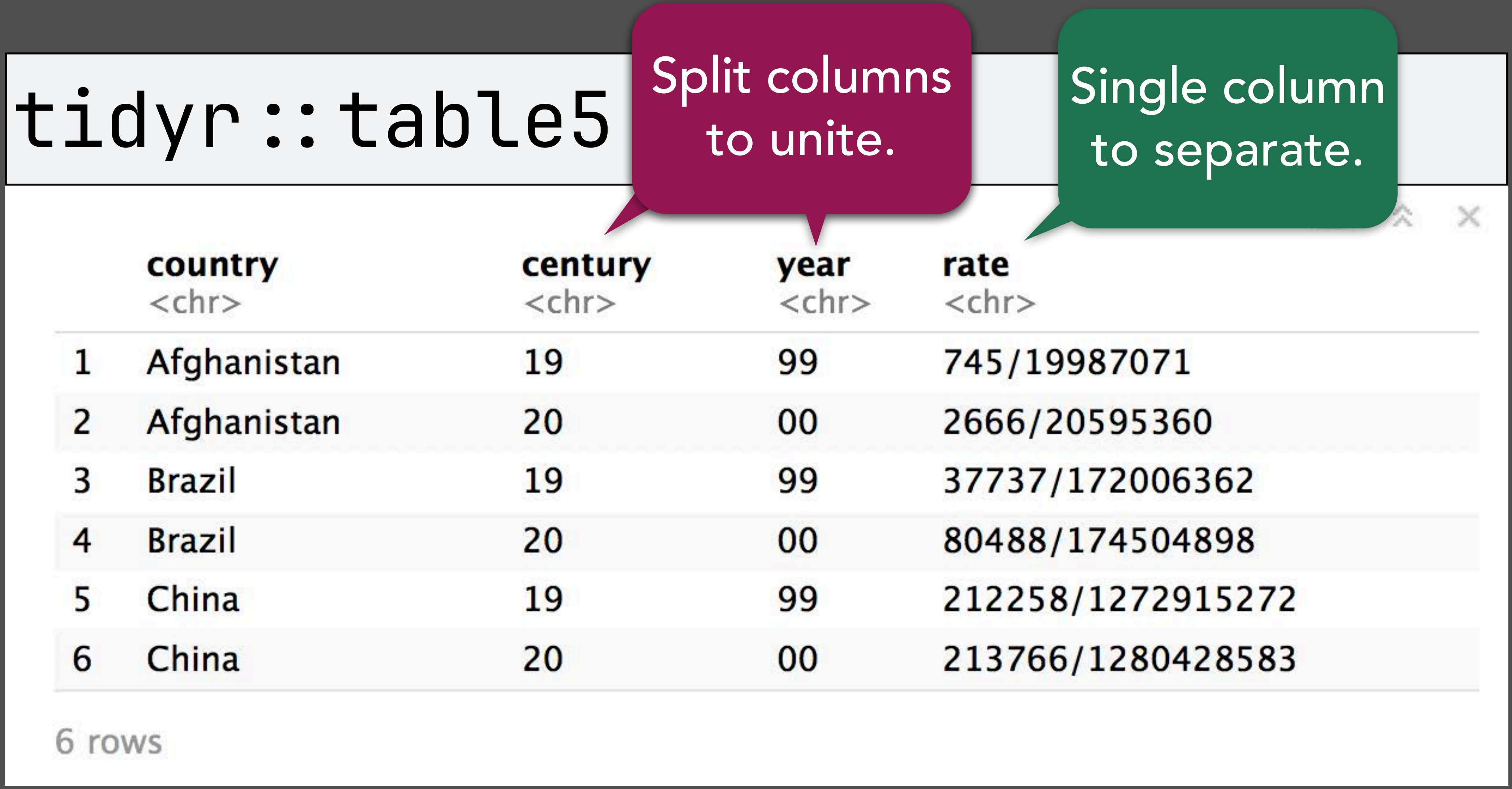
country	continent	1952	1957	1962	1967	1972	1977	1982	1987	1992	1997	2002	2007
Afghanistan	Asia	28.801	30.332	31.997	34.020	36.088	38.438	39.854	40.822	41.674	41.763	42.129	43.828
Albania	Europe	55.230	59.280	64.820	66.220	67.690	68.930	70.420	72.000	71.581	72.950	75.651	76.423
Algeria	Africa	43.077	45.685	48.303	51.407	54.518	58.014	61.368	65.799	67.744	69.152	70.994	72.301
Angola	Africa	30.015	31.999	34.000	35.985	37.928	39.483	39.942	39.906	40.647	40.963	41.003	42.731
Argentina	Americas	62.485	64.399	65.142	65.634	67.065	68.481	69.942	70.774	71.868	73.275	74.340	75.320
Australia	Oceania	69.120	70.330	70.930	71.100	71.930	73.490	74.740	76.320	77.560	78.830	80.370	81.235
Austria	Europe	66.800	67.480	69.540	70.140	70.630	72.170	73.180	74.940	76.040	77.510	78.980	79.829
Bahrain	Asia	50.939	53.832	56.923	59.923	63.300	65.593	69.052	70.750	72.601	73.925	74.795	75.635
Bangladesh	Asia	37.484	39.348	41.216	43.453	45.252	46.923	50.009	52.819	56.018	59.412	62.013	64.062
Belgium	Europe	68.000	69.240	70.250	70.940	71.440	72.800	73.930	75.350	76.460	77.530	78.320	79.441

SPLITTING AND COMBINING CELLS

SPLIT/COMBINE "VERBS"

- `unite()`: collapse cells across several columns into one column
- `separate_wider_delim()`: separate each cell in a column into several columns
    - Variants `separate_wider_position()` and `separate_wider_regex()`
    - Also some `separate_longer_` functions, but they are rarely useful.

REMEMBER THIS TABLE?



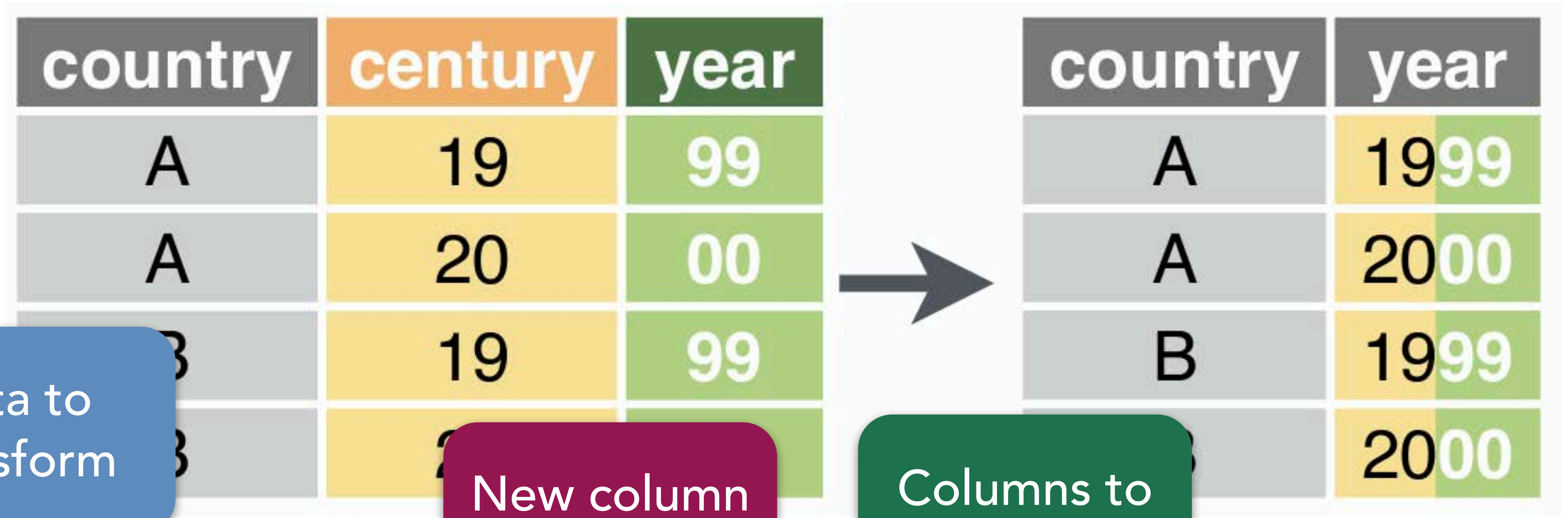
```
tidyr::table5
```

	country	century	year	rate
	<chr>	<chr>	<chr>	<chr>
1	Afghanistan	19	99	745/19987071
2	Afghanistan	20	00	2666/20595360
3	Brazil	19	99	37737/172006362
4	Brazil	20	00	80488/174504898
5	China	19	99	212258/1272915272
6	China	20	00	213766/1280428583

6 rows

Step 1:

unite()



Data to transform

New column create

Columns to unite

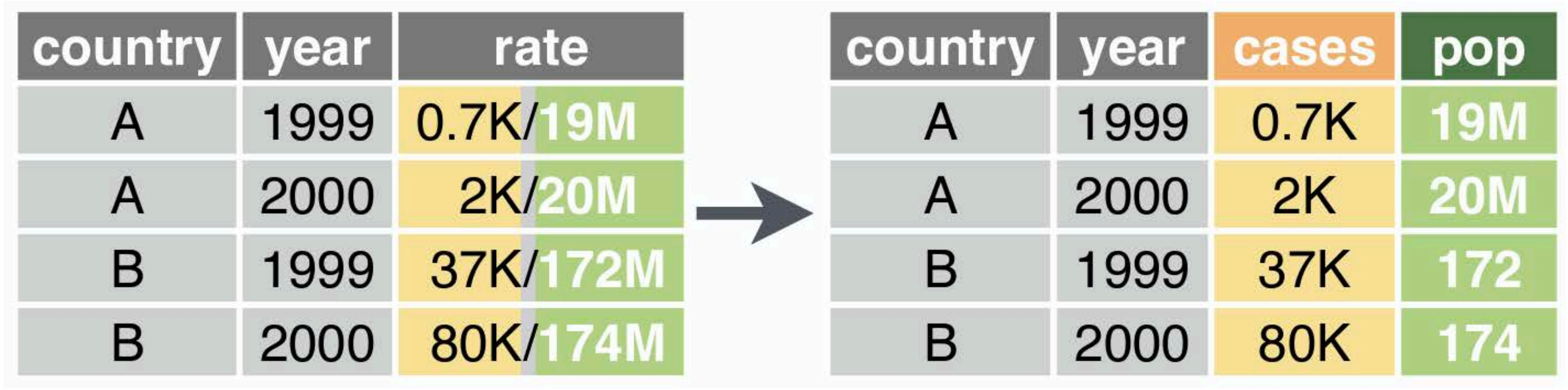
```
table5 ▷
  unite(col = "year", c(century, year),
        sep = "", remove = TRUE)
```

Separator to use between values

Remove old columns?

Step 2:

separate_…()



country	year	rate
A	1999	0.7K/19M
A	2000	2K/20M
B	1999	37K/172M
B	2000	80K/174M

country	year	cases	pop
A	1999	0.7K	19M
A	2000	2K	20M
B	1999	37K	172
B	2000	80K	174

```
table5 ▷
  unite(col = "year",              ye
        sep = "", remov            ▷
  separate_wider_delim(cols = rate, delim = "/",
                       names = c("cases", "pop"))
```

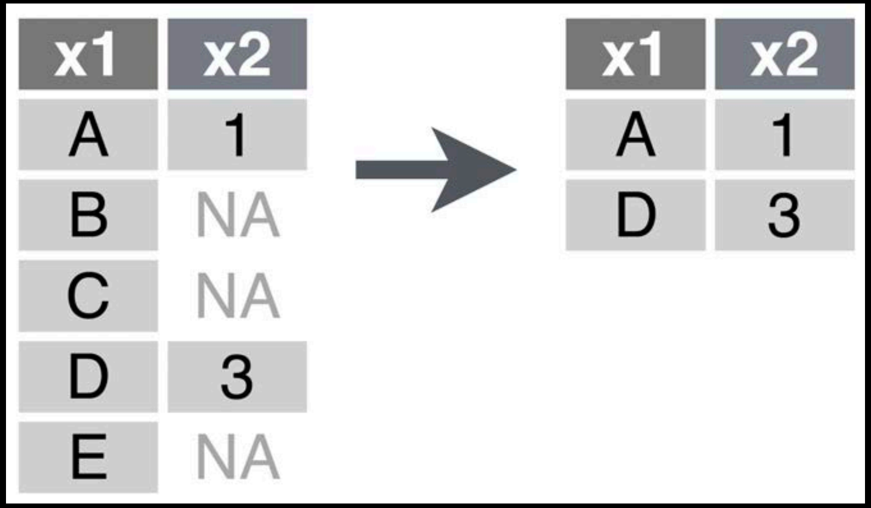
Column to separate

Separator used between values

New columns to create

A FEW OTHER USEFUL THINGS

MISSING VALUES



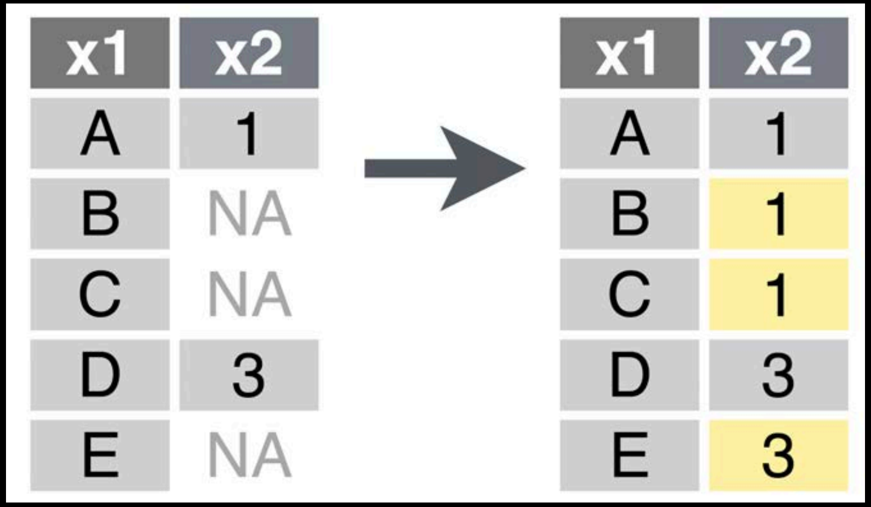
x1	x2
A	1
B	NA
C	NA
D	3
E	NA

→

x1	x2
A	1
D	3

Drop rows containing NAs in columns

```
drop_na(dat, x2)
```



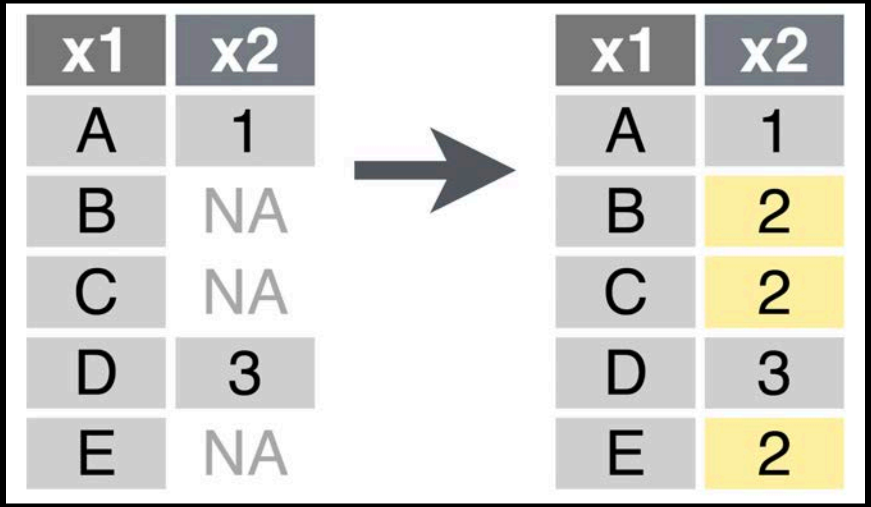
x1	x2
A	1
B	NA
C	NA
D	3
E	NA

→

x1	x2
A	1
B	1
C	1
D	3
E	3

Fill in NAs using next or previous value.

```
fill(dat, x2)
```



x1	x2
A	1
B	NA
C	NA
D	3
E	NA

→

x1	x2
A	1
B	2
C	2
D	3
E	2

Specify a value to replace NAs.

```
replace_na(dat, list(x2 = 2))
```

ACTIVITY 5: TIDY DATA

- Go to this week's assignments on the course website.
- Download tidy-data.qmd and follow the instructions to complete the assignment.