Econ 106: Data Analyis for Economics

Lecture 11

slides adapted from: https://r4ds.had.co.nz/tidy-data.html

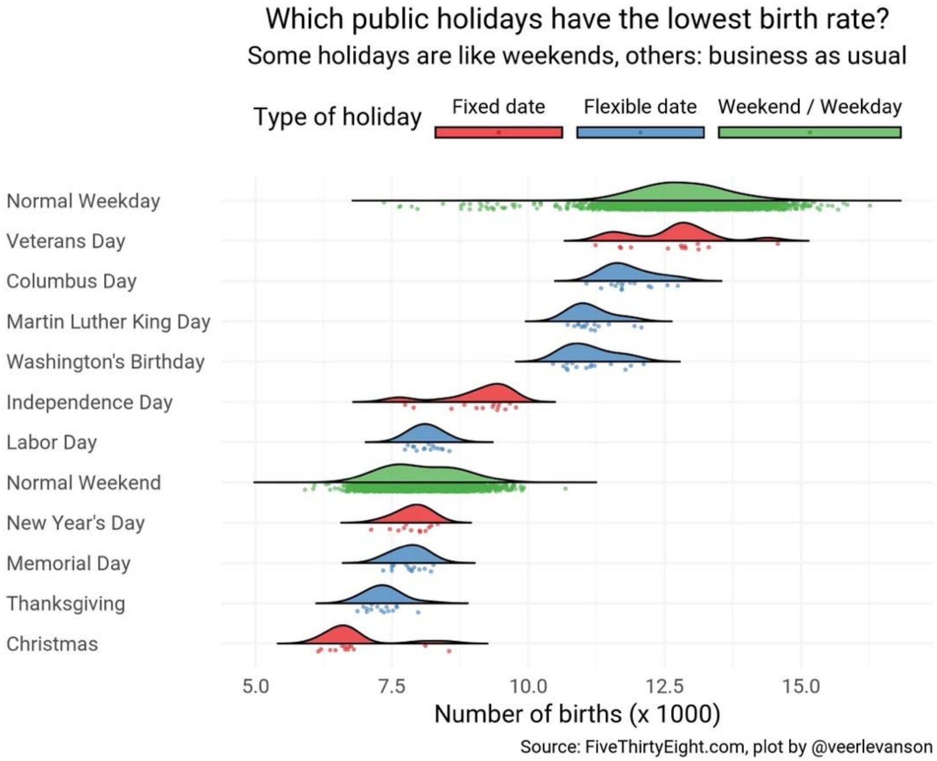
Reminders

- Lab #3 is due Friday, 11:59pm
- Please remember to turn in MS #1 by 11:59pm tonight (5% late penalty is better than getting a zero)

https://pollev.com/vsovero

#tidytuesday

- Remember, you need to also be able to interpret your visualization



Outline

- Tidy data
- reshaping
- separate

Data in the Wild

- Unfortunately, ”real data” is going to be messy
- Each data set is messy in its own unique way
- Our job as analysts is to untangle the mess before we can conduct any sort of analysis
- This week, we are going to learn some tools for tidying data

What is Tidy Data?

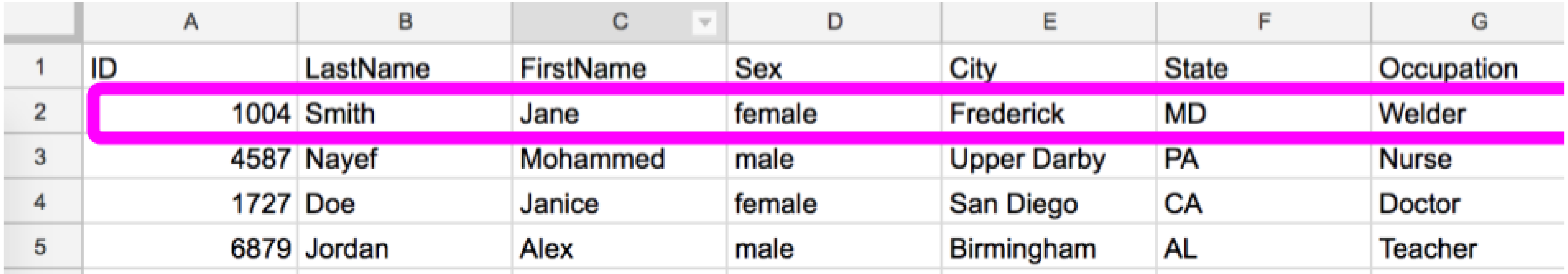
- each row is an observation
- each column is a single variable
- data is rectangular
- if there are multiple data tables, they should have an identifier that allows them to be joined together

Why do we want Tidy Data?

- Tidy data require only a *small set of tools to be learned*:
  - When using a consistent data format, only a small set of tools is required (dplyr for example)
  - these tools can be reused easily from one project to the next

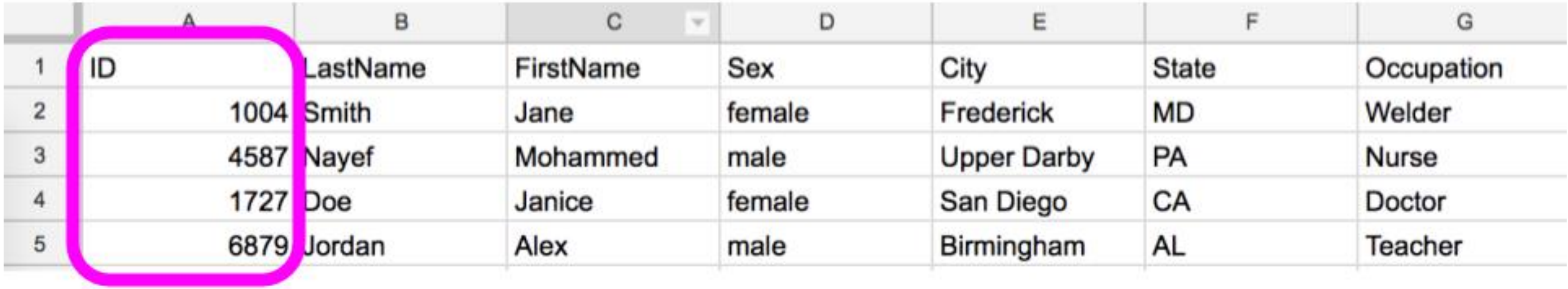
- Tidy data allow for *datasets to be combined:*
  - Data are often stored in multiple tables or in different locations.
  - By getting each table into a tidy format, combining across tables or sources is easy

Each row is an observation



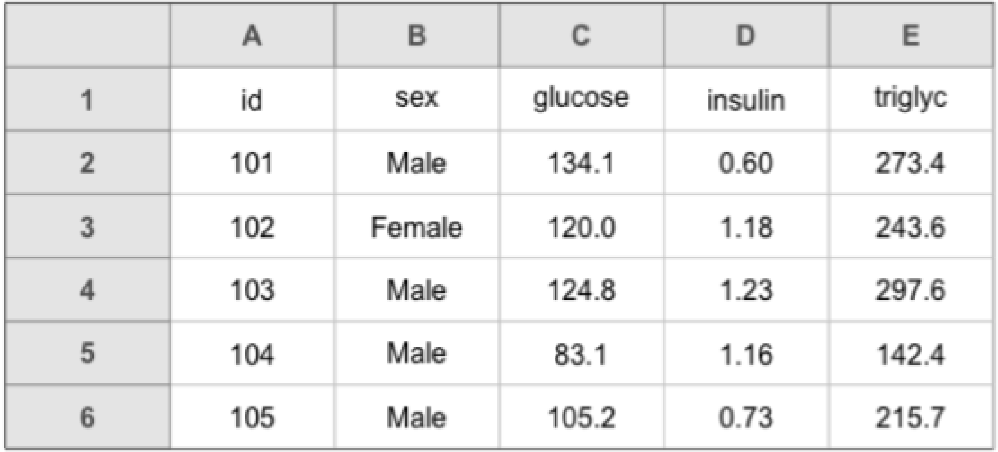
	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Sex	City	State	Occupation
2	1004	Smith	Jane	female	Frederick	MD	Welder
3	4587	Nayef	Mohammed	male	Upper Darby	PA	Nurse
4	1727	Doe	Janice	female	San Diego	CA	Doctor
5	6879	Jordan	Alex	male	Birmingham	AL	Teacher

Each variable has it’s own column



	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Sex	City	State	Occupation
2	1004	Smith	Jane	female	Frederick	MD	Welder
3	4587	Nayef	Mohammed	male	Upper Darby	PA	Nurse
4	1727	Doe	Janice	female	San Diego	CA	Doctor
5	6879	Jordan	Alex	male	Birmingham	AL	Teacher

Data is rectangular

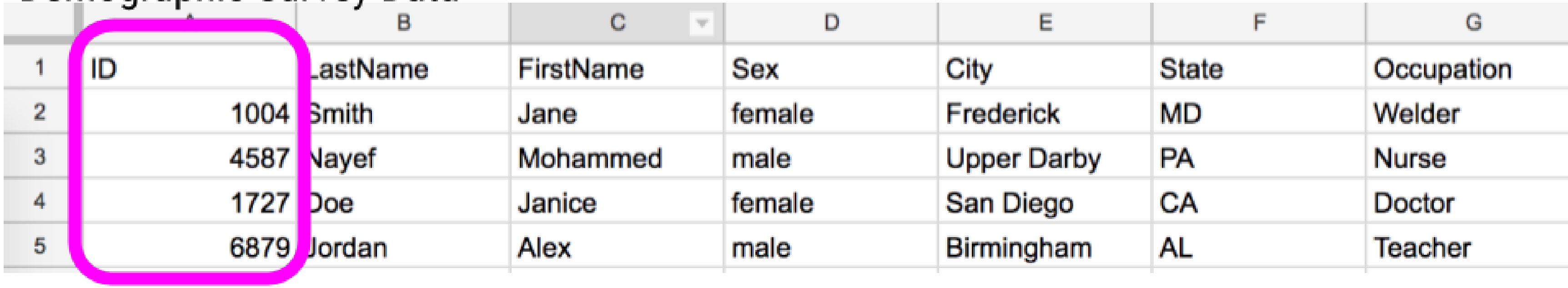


	A	B	C	D	E
1	id	sex	glucose	insulin	triglyc
2	101	Male	134.1	0.60	273.4
3	102	Female	120.0	1.18	243.6
4	103	Male	124.8	1.23	297.6
5	104	Male	83.1	1.16	142.4
6	105	Male	105.2	0.73	215.7

Broman KW, Woo KH. (2017) Data organization in spreadsheets. *PeerJ Preprints*5:e3183v1 https://doi.org/10.7287/peerj.preprints.3183v1

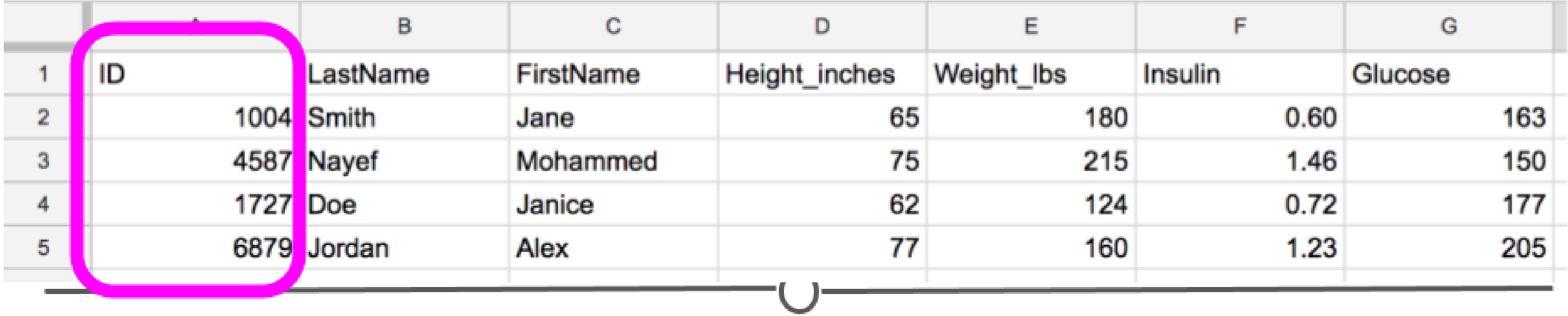
Each dataset has an identifier for joins

Demographic Survey Data



	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Sex	City	State	Occupation
2	1004	Smith	Jane	female	Frederick	MD	Welder
3	4587	Nayef	Mohammed	male	Upper Darby	PA	Nurse
4	1727	Doe	Janice	female	San Diego	CA	Doctor
5	6879	Jordan	Alex	male	Birmingham	AL	Teacher

Doctor's Office Measurements Data



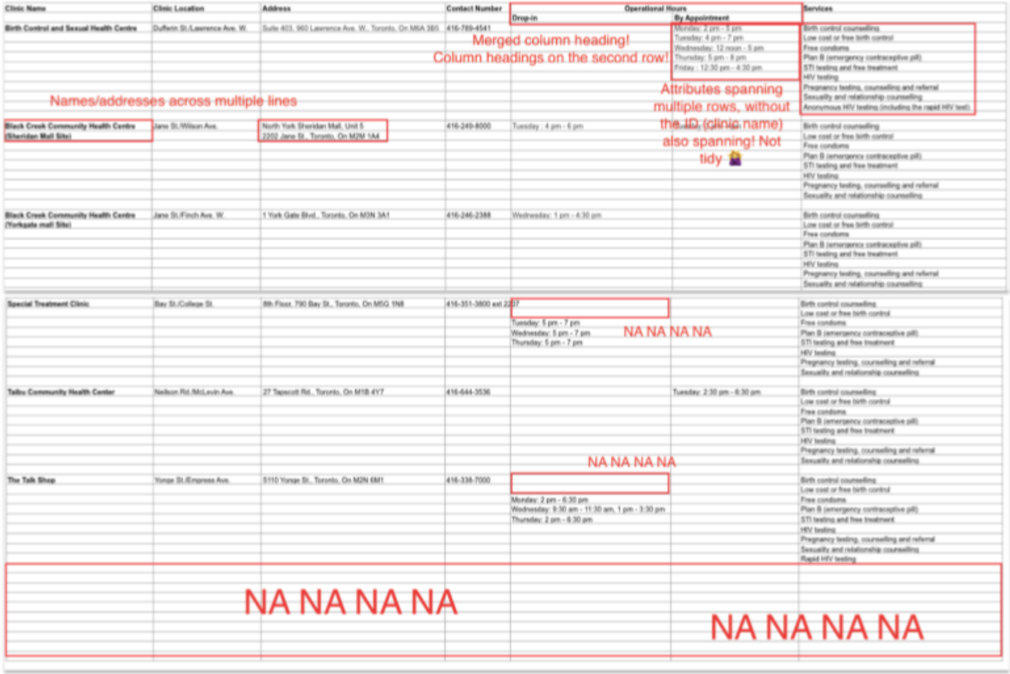
	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Height_inches	Weight_lbs	Insulin	Glucose
2	1004	Smith	Jane	65	180	0.60	163
3	4587	Nayef	Mohammed	75	215	1.46	150
4	1727	Doe	Janice	62	124	0.72	177
5	6879	Jordan	Alex	77	160	1.23	205

Messy Data

Common problems with untidy data:

1. Column headers are values but should be variable names.
2. A single column has multiple variables.
3. Variables have been entered in both rows and columns.
4. Multiple “types” of data are in the same spreadsheet.
5. A single observation is stored across multiple spreadsheets.

Messy Data Example



Clinic Name	Clinic Location	Address	Contact Number	Operational Hours		Services
				Drop-in	By Appointment	
Birth Control and Sexual Health Centre	Dufferin St./Lawrence Ave. W.	Suite 403, 960 Lawrence Ave. W., Toronto, On M6A 3B5	416-789-4541		Monday: 2 pm - 5 pm Tuesday: 4 pm - 7 pm Wednesday: 12 noon - 5 pm Thursday: 5 pm - 8 pm Friday : 12:30 pm - 4:30 pm	Birth control counselling Low cost or free birth control Free condoms Plan B (emergency contraceptive pill) STI testing and free treatment HIV testing Pregnancy testing, counselling and referral Sexuality and relationship counselling Anonymous HIV testing (including the rapid HIV test)
Black Creek Community Health Centre (Sheridan Mall Site)	Jane St./Wilson Ave.	North York Sheridan Mall, Unit 5 2202 Jane St., Toronto, On M2M 1A4	416-249-8000	Tuesday : 4 pm - 6 pm	[illegible]	Birth control counselling Low cost or free birth control Free condoms Plan B (emergency contraceptive pill) STI testing and free treatment HIV testing Pregnancy testing, counselling and referral Sexuality and relationship counselling
Black Creek Community Health Centre (Yorkgate mall Site)	Jane St./Finch Ave. W.	1 York Gate Blvd., Toronto, On M3N 3A1	416-246-2388	Wednesday: 1 pm - 4:30 pm		Birth control counselling Low cost or free birth control Free condoms Plan B (emergency contraceptive pill) STI testing and free treatment HIV testing Pregnancy testing, counselling and referral Sexuality and relationship counselling
Special Treatment Clinic	Bay St./College St.	8th Floor, 790 Bay St., Toronto, On M5G 1N8	416-351-3800 ext 2207	Tuesday: 5 pm - 7 pm Wednesday: 5 pm - 7 pm Thursday: 5 pm - 7 pm		Birth control counselling Low cost or free birth control Free condoms Plan B (emergency contraceptive pill) STI testing and free treatment HIV testing Pregnancy testing, counselling and referral Sexuality and relationship counselling
Taibu Community Health Center	Neilson Rd./McLevin Ave.	27 Tapscott Rd., Toronto, On M1B 4Y7	416-644-3536		Tuesday: 2:30 pm - 6:30 pm	Birth control counselling Low cost or free birth control Free condoms Plan B (emergency contraceptive pill) STI testing and free treatment HIV testing Pregnancy testing, counselling and referral Sexuality and relationship counselling
The Talk Shop	Yonge St./Empress Ave.	5110 Yonge St., Toronto, On M2N 6M1	416-338-7000	Monday: 2 pm - 6:30 pm Wednesday: 9:30 am - 11:30 am, 1 pm - 3:30 pm Thursday: 2 pm - 6:30 pm		Birth control counselling Low cost or free birth control Free condoms Plan B (emergency contraceptive pill) STI testing and free treatment HIV testing Pregnancy testing, counselling and referral Sexuality and relationship counselling Rapid HIV testing

Merged column heading!
Column headings on the second row!

Names/addresses across multiple lines

Attributes spanning multiple rows, without the ID (clinic name) also spanning! Not tidy 🤦‍♀️

NA NA NA NA

NA NA NA NA

NA NA NA NA

NA NA NA NA

https://sharlagelfand.netlify.com/posts/tidying-toronto-open-data/

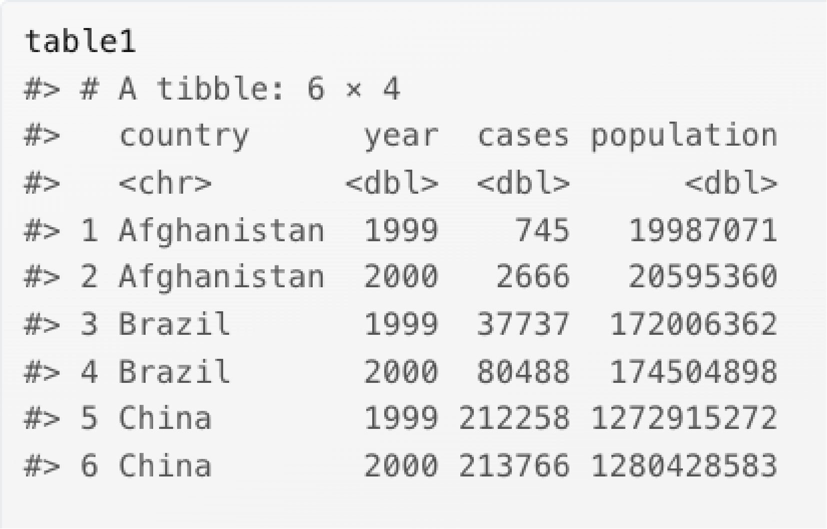
Example: TB data

- Each dataset shows the same values of four variables:
  - country
  - year
  - population
  - number of documented cases of TB (tuberculosis)
- However, each dataset organizes the values in a different way

```
data(table1)
data(table2)
```

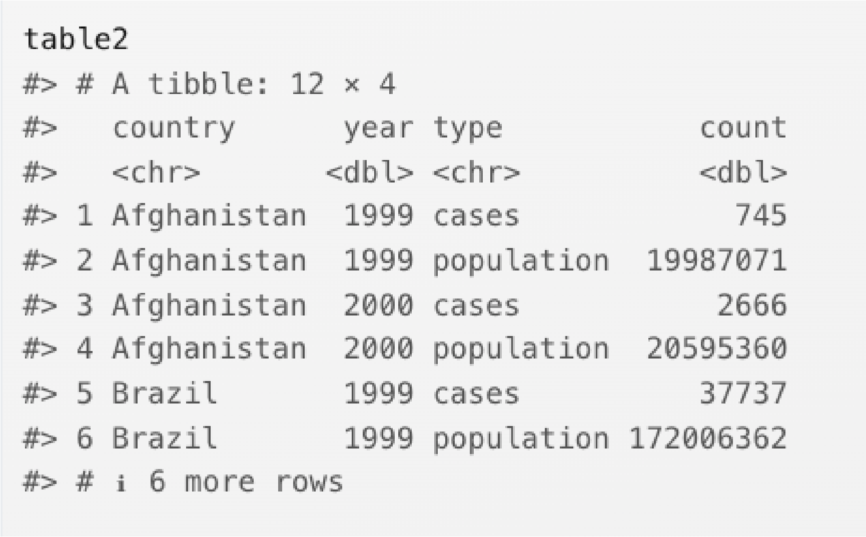
Which data set is tidy?

table1



```
table1
#> # A tibble: 6 × 4
#>   country      year  cases population
#>   <chr>       <dbl>  <dbl>      <dbl>
#> 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
```

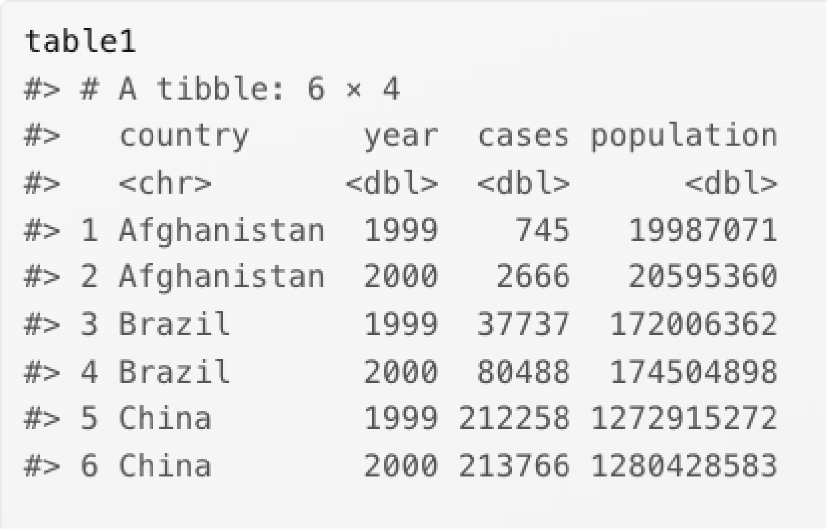
table2



```
table2
#> # A tibble: 12 × 4
#>   country      year type            count
#>   <chr>       <dbl> <chr>           <dbl>
#> 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
#> # ℹ 6 more rows
```

Exercise: How would you calculate the case rate by country?

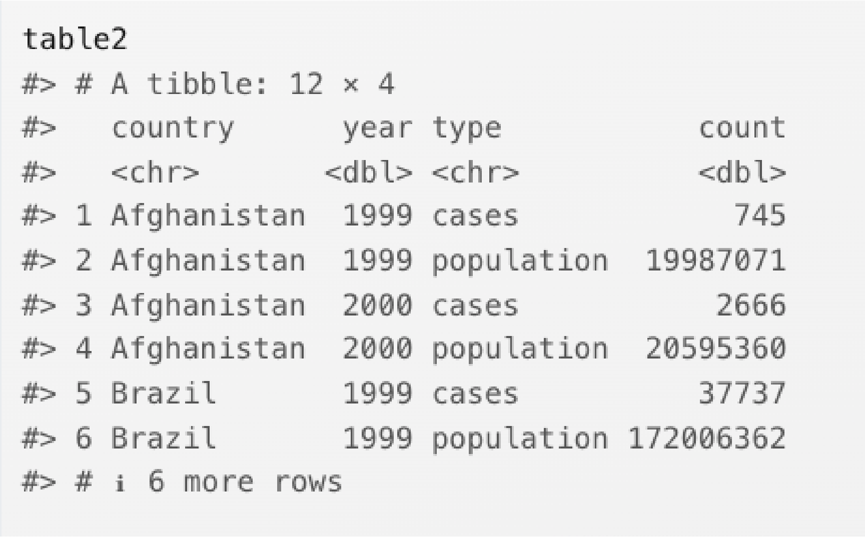
table1



```
table1
#> # A tibble: 6 × 4
#>   country      year  cases population
#>   <chr>       <dbl>  <dbl>      <dbl>
#> 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 would you calculate the case rate by country?

table2



```
table2
#> # A tibble: 12 × 4
#>   country      year type            count
#>   <chr>       <dbl> <chr>           <dbl>
#> 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
#> # ℹ 6 more rows
```

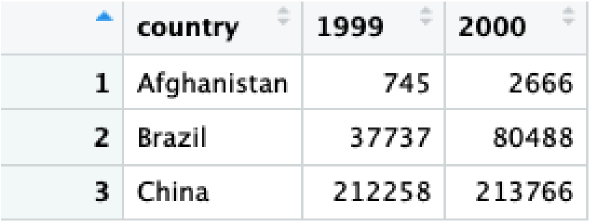
Reshaping Data

- Sometimes our data requires reshaping in order to be tidy
  - converting columns into rows (make your data longer)
  - converting rows into columns (make your data wider)
- There are tools in the tidyverse that can do this reshaping for you

Example: TB Cases (Table 4a)

Problems:

- Column names in this dataset represent years
- Values in the 1999 and 2000 columns actually represent TB cases



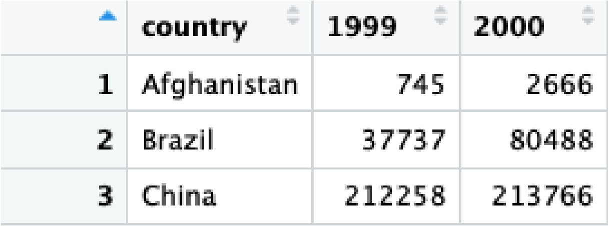
	country	1999	2000
1	Afghanistan	745	2666
2	Brazil	37737	80488
3	China	212258	213766

Solution:

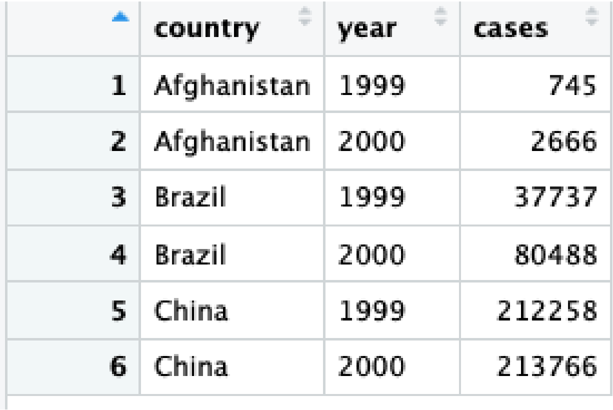
- We need to reshape from <u>wide</u> to <u>long</u>

Reshaping Data: Wide to Long

- What are the changes need to go from the table on the left to the table on the right?



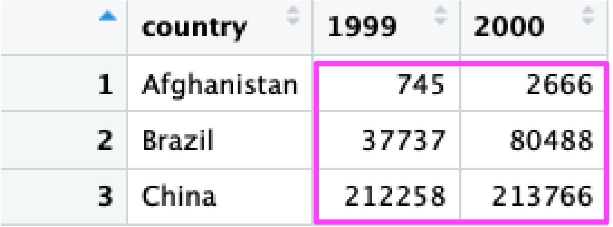
	country	1999	2000
1	Afghanistan	745	2666
2	Brazil	37737	80488
3	China	212258	213766



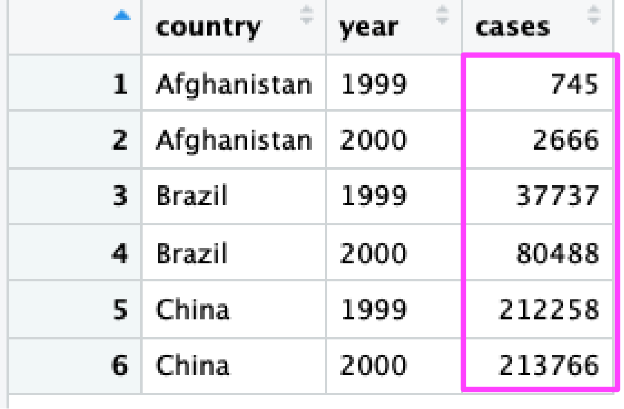
	country	year	cases
1	Afghanistan	1999	745
2	Afghanistan	2000	2666
3	Brazil	1999	37737
4	Brazil	2000	80488
5	China	1999	212258
6	China	2000	213766

Wide to Long Breakdown

- there are two columns that include information on cases
- this should be a single column called cases



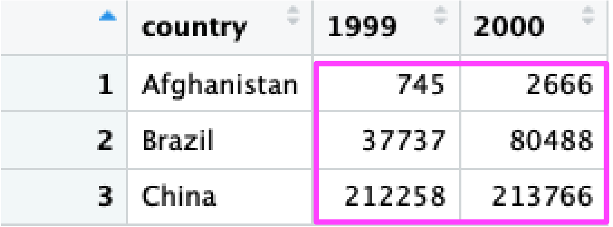
	country	1999	2000
1	Afghanistan	745	2666
2	Brazil	37737	80488
3	China	212258	213766



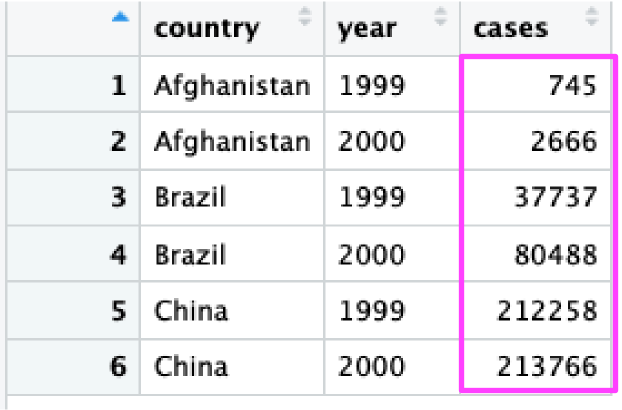
	country	year	cases
1	Afghanistan	1999	745
2	Afghanistan	2000	2666
3	Brazil	1999	37737
4	Brazil	2000	80488
5	China	1999	212258
6	China	2000	213766

Wide to Long Breakdown

- This is called “wide to long” because we started with a wide table (many columns) and ended up with a long table (fewer columns, more rows)



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

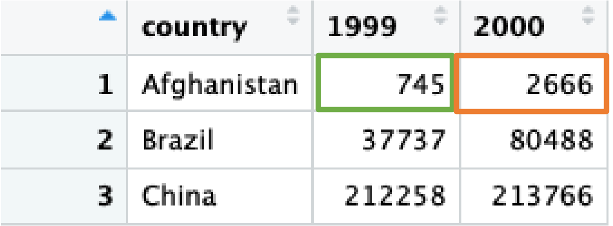


	country	year	cases
1	Afghanistan	1999	745
2	Afghanistan	2000	2666
3	Brazil	1999	37737
4	Brazil	2000	80488
5	China	1999	212258
6	China	2000	213766

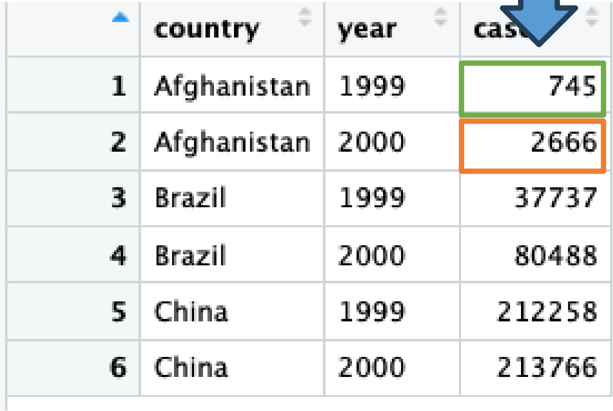
Wide to Long Breakdown

How do we move over the values?

- Afghanistan has cases from 1999 and 2000
- We need stack the values into a single column
- This creates two rows of data



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

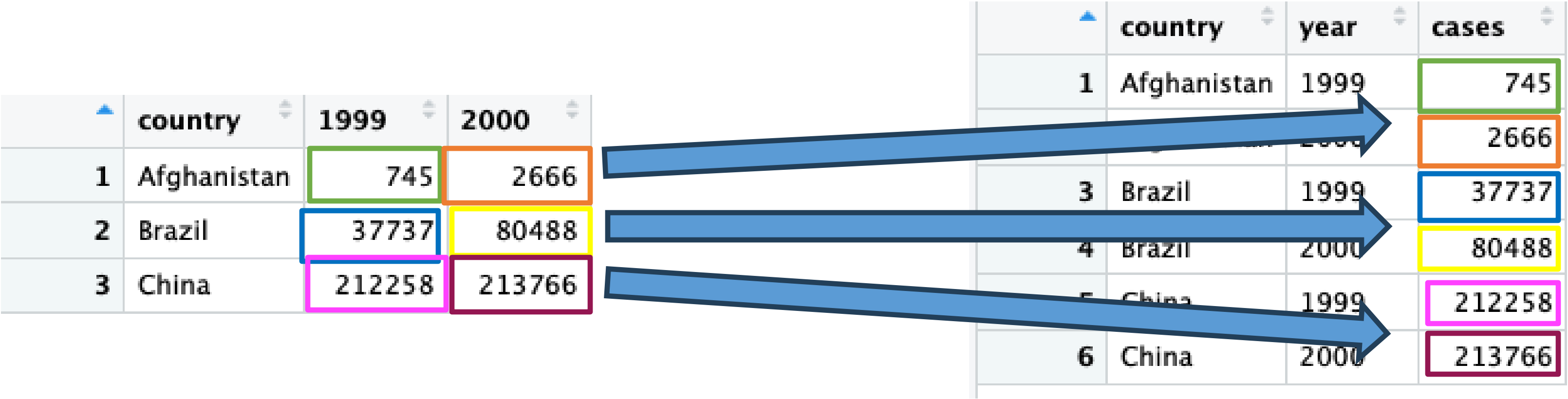


	country	year	cas
1	Afghanistan	1999	745
2	Afghanistan	2000	2666
3	Brazil	1999	37737
4	Brazil	2000	80488
5	China	1999	212258
6	China	2000	213766

Wide to Long Breakdown

We do this process one row at a time:

- Afghanistan cases stack into a single column (two rows)
- Brazil cases stack into a single column (two rows)
- Chinca cases stack into a single column (two rows)

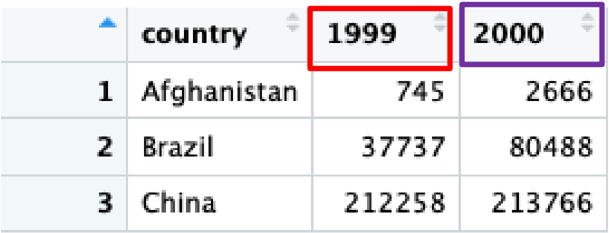


	country	1999	2000
1	Afghanistan	745	2666
2	Brazil	37737	80488
3	China	212258	213766

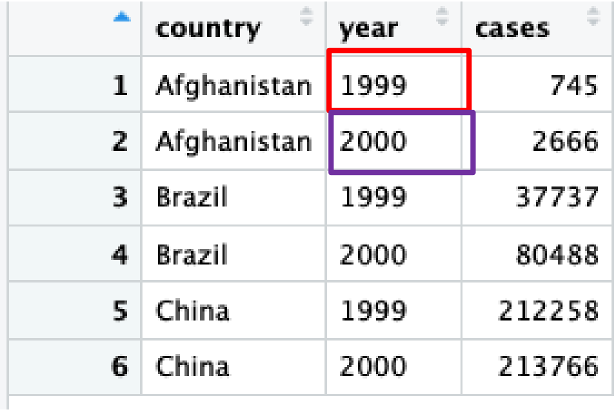
	country	year	cases
1	Afghanistan	1999	745
2	Afghanistan	2000	2666
3	Brazil	1999	37737
4	Brazil	2000	80488
5	China	1999	212258
6	China	2000	213766

Wide to Long Breakdown

- The 1999 and 2000 column names go into a new column called “year”



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

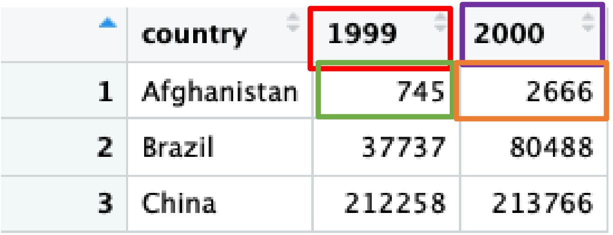


	country	year	cases
1	Afghanistan	1999	745
2	Afghanistan	2000	2666
3	Brazil	1999	37737
4	Brazil	2000	80488
5	China	1999	212258
6	China	2000	213766

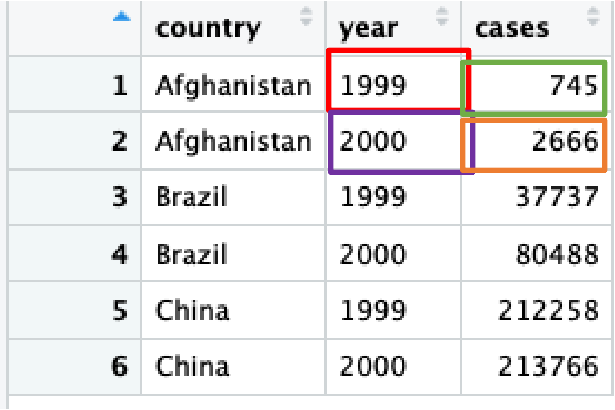
Wide to Long Breakdown

The year column helps us tell which year the case data comes from:

- 745 cases in Afghanistan in 1999
- 2666 cases in Afghanistan in 2000



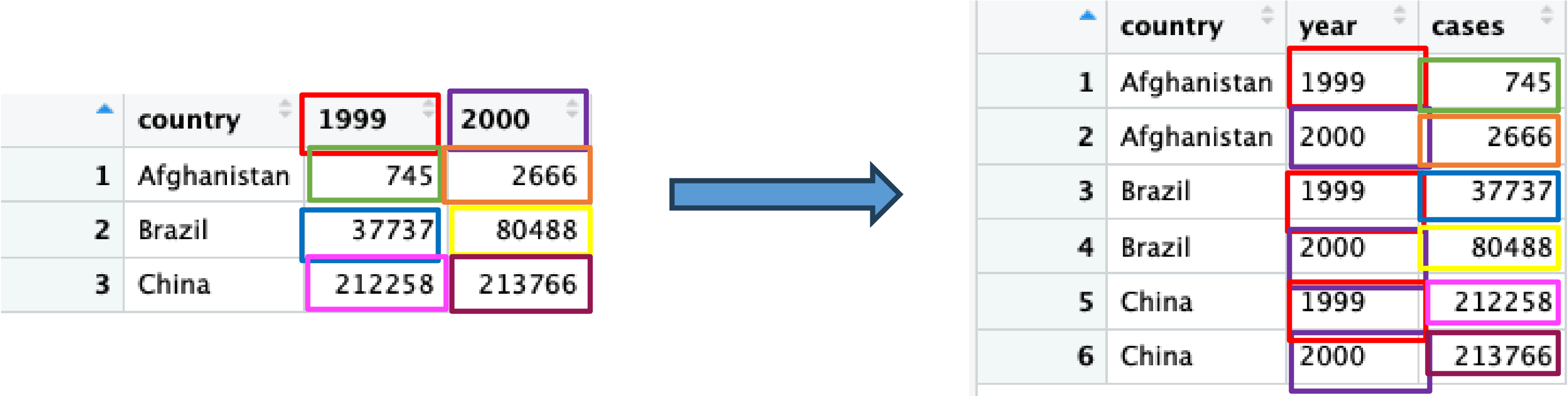
	country	1999	2000
1	Afghanistan	745	2666
2	Brazil	37737	80488
3	China	212258	213766



	country	year	cases
1	Afghanistan	1999	745
2	Afghanistan	2000	2666
3	Brazil	1999	37737
4	Brazil	2000	80488
5	China	1999	212258
6	China	2000	213766

Wide to Long Breakdown

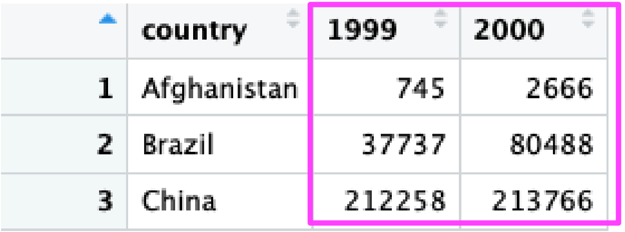
Everything all together:



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

	country	year	cases
1	Afghanistan	1999	745
2	Afghanistan	2000	2666
3	Brazil	1999	37737
4	Brazil	2000	80488
5	China	1999	212258
6	China	2000	213766

pivot_longer()

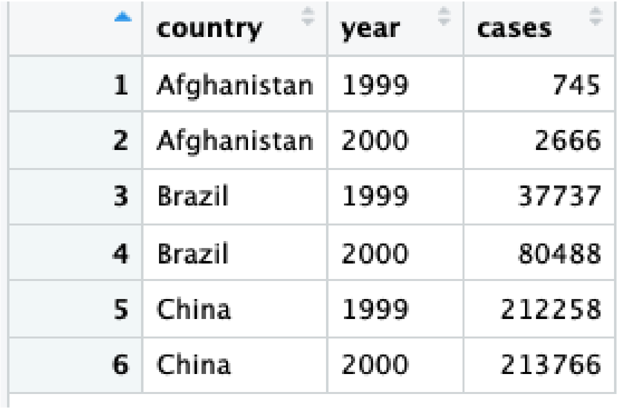


	country	1999	2000
1	Afghanistan	745	2666
2	Brazil	37737	80488
3	China	212258	213766

We need to specify:

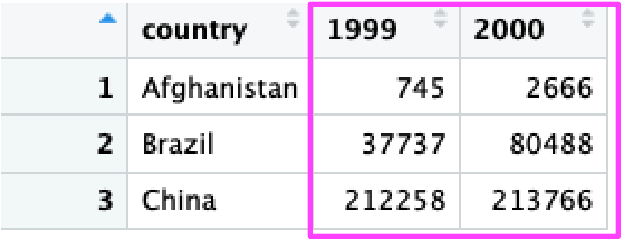
- The set of columns that need to be stacked into a single row

```
case_data <- table4a %>%
 pivot_longer(cols= c(`1999`, `2000`),
                 names_to = "year",
                 values_to = "cases”)
```



	country	year	cases
1	Afghanistan	1999	745
2	Afghanistan	2000	2666
3	Brazil	1999	37737
4	Brazil	2000	80488
5	China	1999	212258
6	China	2000	213766

pivot_longer()

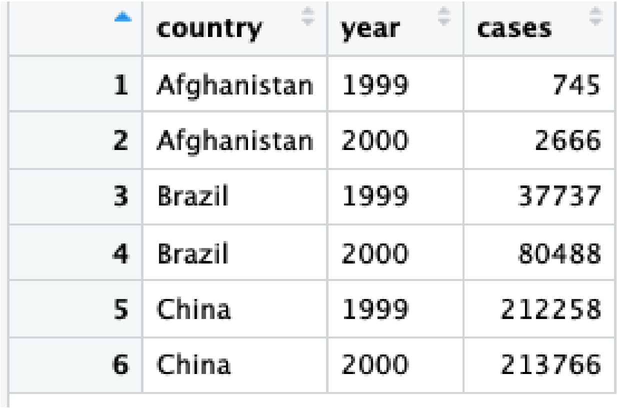


	country	1999	2000
1	Afghanistan	745	2666
2	Brazil	37737	80488
3	China	212258	213766

We need to specify:

- Alternatively, you can deselect the variable you don’t want to pivot (country)

```
case_data <- table4a %>%
 pivot_longer(-country),
                    names_to = "year",
                    values_to = "cases")
```

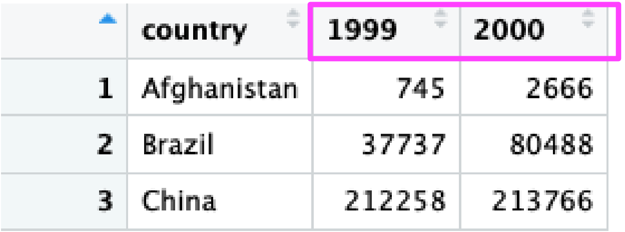


	country	year	cases
1	Afghanistan	1999	745
2	Afghanistan	2000	2666
3	Brazil	1999	37737
4	Brazil	2000	80488
5	China	1999	212258
6	China	2000	213766

pivot_longer()

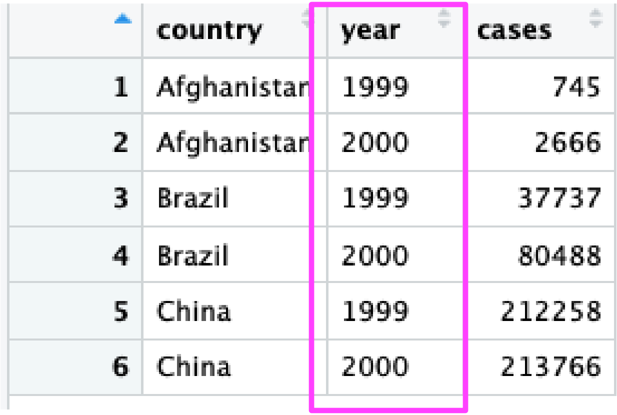
We need to specify:

- names_to: the name of the variable to move the column names to



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

```
case_data <- table4a %>%
  pivot_longer(-country),
               names_to = "year",
               values_to = "cases")
```

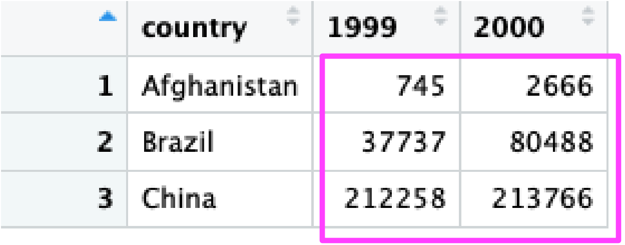


	country	year	cases
1	Afghanistan	1999	745
2	Afghanistan	2000	2666
3	Brazil	1999	37737
4	Brazil	2000	80488
5	China	1999	212258
6	China	2000	213766

pivot_longer()

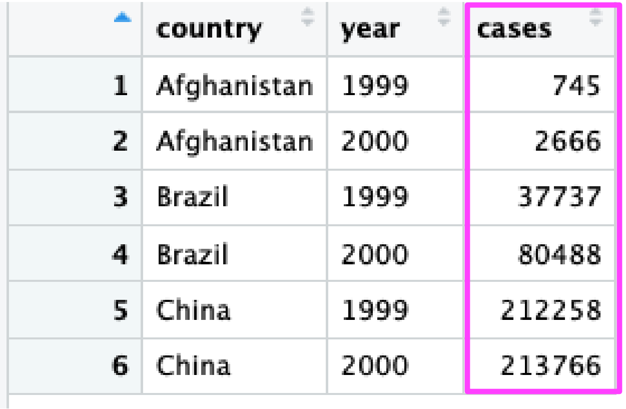
We need to specify:

- values_to: the name of the variable to move the column values to



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

```
case_data <- table4a %>%
  pivot_longer(-country),
                        names_to = "year",
                        values_to = "cases")
```



	country	year	cases
1	Afghanistan	1999	745
2	Afghanistan	2000	2666
3	Brazil	1999	37737
4	Brazil	2000	80488
5	China	1999	212258
6	China	2000	213766

pivot_longer() additional options

https://tidyr.tidyverse.org/reference/pivot_longer.html#ref-examples

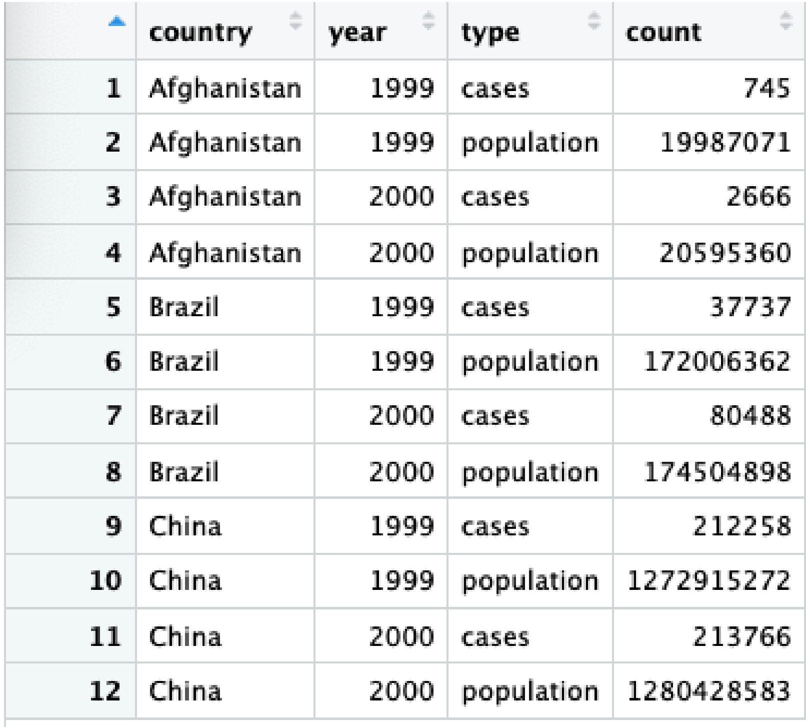
Example (table2)

Problems:

- The Type variable contains variable names instead of values
- Count has values of more than one variable

Solution:

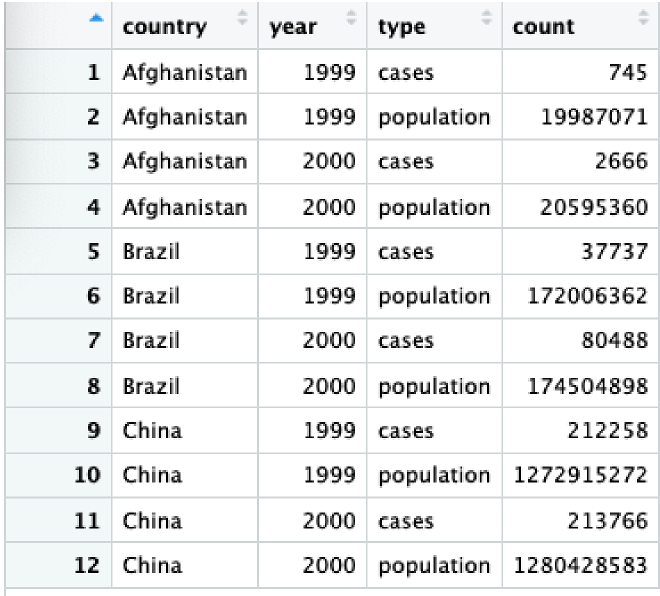
- we need to reshape from <u>long</u> to <u>wide</u>



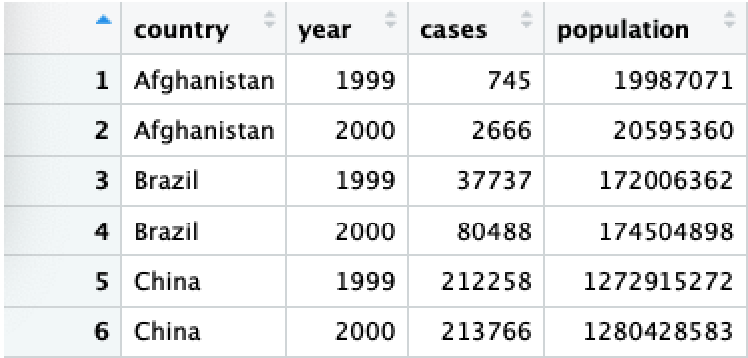
	country	year	type	count
1	Afghanistan	1999	cases	745
2	Afghanistan	1999	population	19987071
3	Afghanistan	2000	cases	2666
4	Afghanistan	2000	population	20595360
5	Brazil	1999	cases	37737
6	Brazil	1999	population	172006362
7	Brazil	2000	cases	80488
8	Brazil	2000	population	174504898
9	China	1999	cases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

Long to Wide Breakdown

- This is called “long to wide” because we are reducing the number of rows and instead adding columns



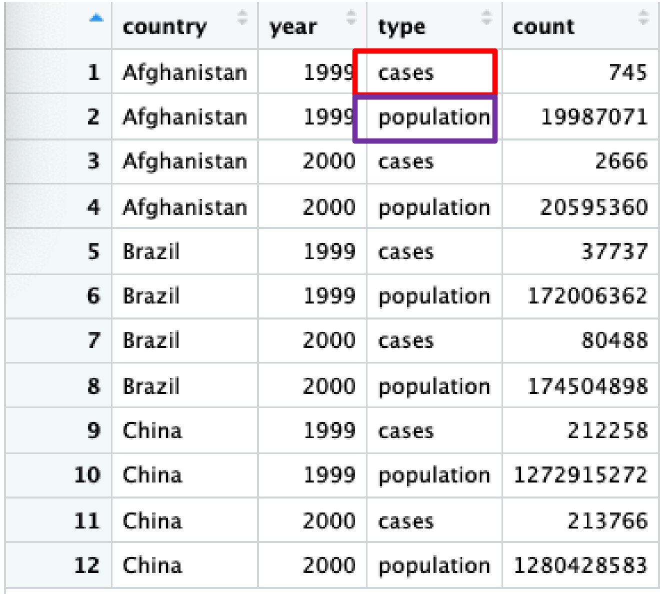
	country	year	type	count
1	Afghanistan	1999	cases	745
2	Afghanistan	1999	population	19987071
3	Afghanistan	2000	cases	2666
4	Afghanistan	2000	population	20595360
5	Brazil	1999	cases	37737
6	Brazil	1999	population	172006362
7	Brazil	2000	cases	80488
8	Brazil	2000	population	174504898
9	China	1999	cases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



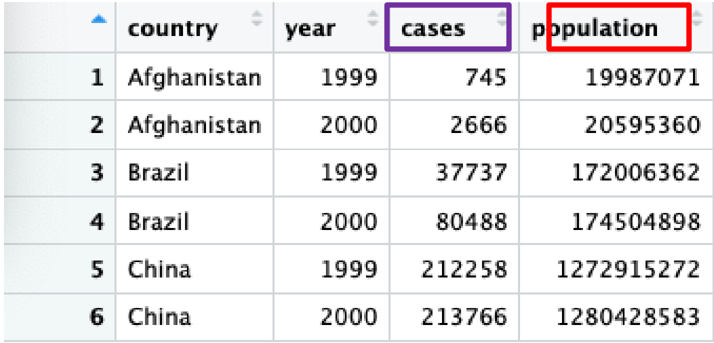
	country	year	cases	population
1	Afghanistan	1999	745	19987071
2	Afghanistan	2000	2666	20595360
3	Brazil	1999	37737	172006362
4	Brazil	2000	80488	174504898
5	China	1999	212258	1272915272
6	China	2000	213766	1280428583

Long to Wide Breakdown

- We take values from the type column and using them as new variable names



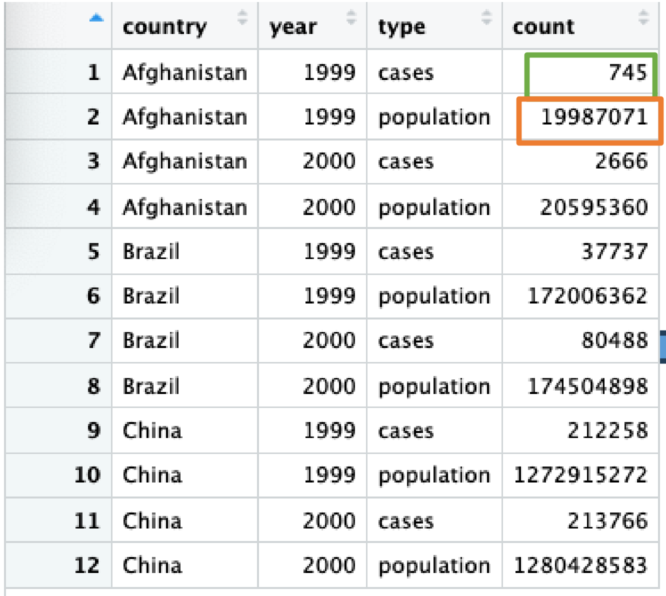
	country	year	type	count
1	Afghanistan	1999	cases	745
2	Afghanistan	1999	population	19987071
3	Afghanistan	2000	cases	2666
4	Afghanistan	2000	population	20595360
5	Brazil	1999	cases	37737
6	Brazil	1999	population	172006362
7	Brazil	2000	cases	80488
8	Brazil	2000	population	174504898
9	China	1999	cases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



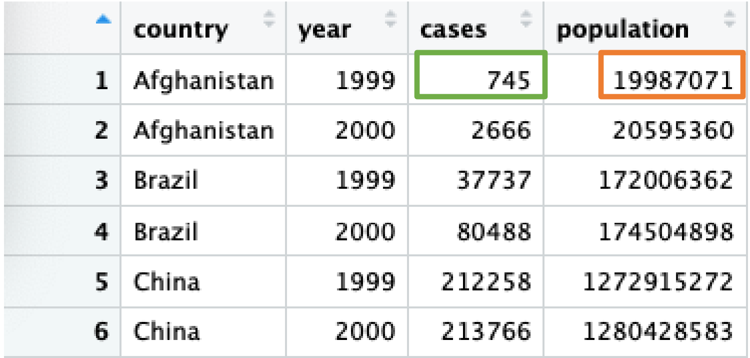
	country	year	cases	population
1	Afghanistan	1999	745	19987071
2	Afghanistan	2000	2666	20595360
3	Brazil	1999	37737	172006362
4	Brazil	2000	80488	174504898
5	China	1999	212258	1272915272
6	China	2000	213766	1280428583

Long to Wide Breakdown

- We move the values in count column to the respective cases and population columns



	country	year	type	count
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



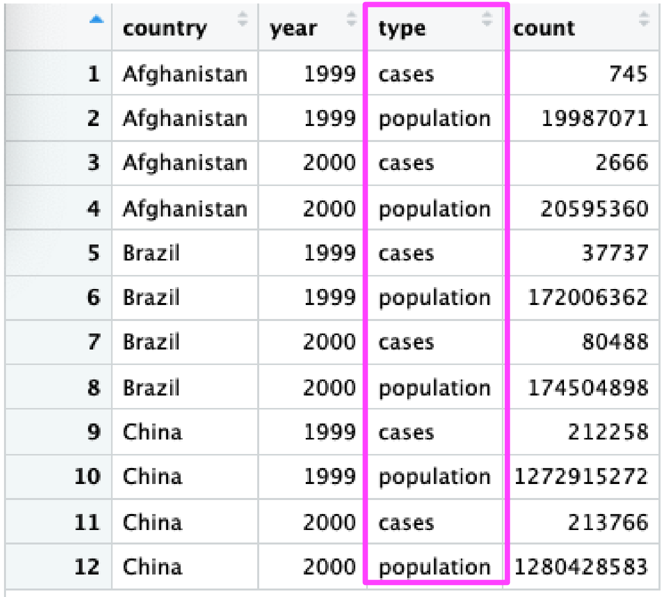
	country	year	cases	population
1	Afghanistan	1999	745	19987071
2	Afghanistan	2000	2666	20595360
3	Brazil	1999	37737	172006362
4	Brazil	2000	80488	174504898
5	China	1999	212258	1272915272
6	China	2000	213766	1280428583

pivot_wider()

We need to specify:

- names_from : The column to take the variable names from

```
case_pop_data <- table2 %>%
 pivot_wider(   names_from =type,
                values_from =count)
```



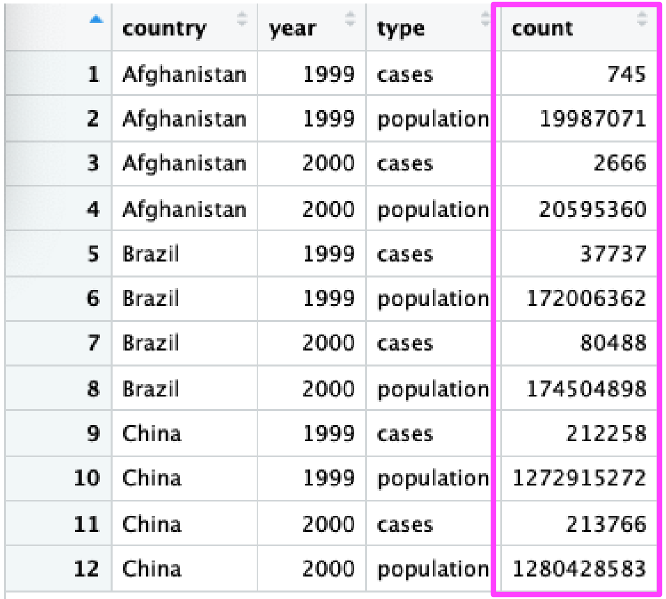
	country	year	type	count
1	Afghanistan	1999	cases	745
2	Afghanistan	1999	population	19987071
3	Afghanistan	2000	cases	2666
4	Afghanistan	2000	population	20595360
5	Brazil	1999	cases	37737
6	Brazil	1999	population	172006362
7	Brazil	2000	cases	80488
8	Brazil	2000	population	174504898
9	China	1999	cases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

pivot_wider()

```
case_pop_data <- table2 %>%
  pivot_wider(   names_from =type,
                 values_from =count)
```

We need to specify:

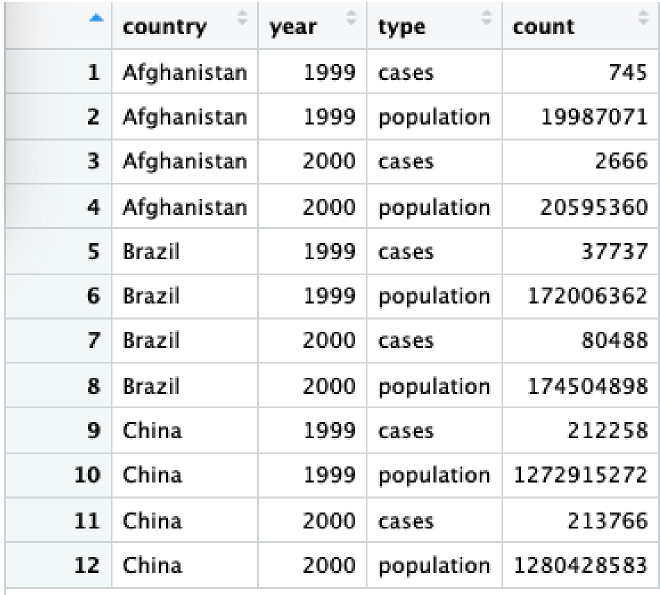
- The column to take values from



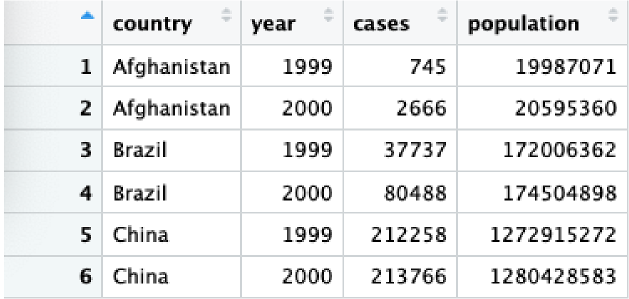
	country	year	type	count
1	Afghanistan	1999	cases	745
2	Afghanistan	1999	population	19987071
3	Afghanistan	2000	cases	2666
4	Afghanistan	2000	population	20595360
5	Brazil	1999	cases	37737
6	Brazil	1999	population	172006362
7	Brazil	2000	cases	80488
8	Brazil	2000	population	174504898
9	China	1999	cases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

pivot_wider()

```
case_pop_data <- table2 %>%
  pivot_wider(   names_from =type,
                values_from =count)
```



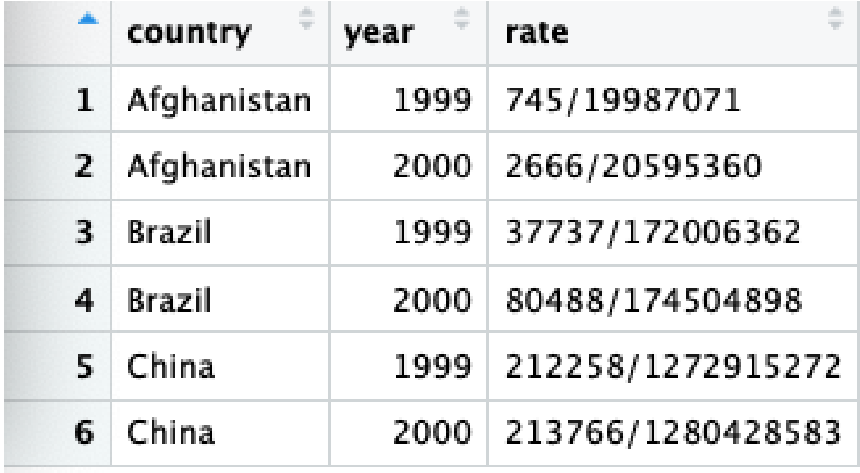
	country	year	type	count
1	Afghanistan	1999	cases	745
2	Afghanistan	1999	population	19987071
3	Afghanistan	2000	cases	2666
4	Afghanistan	2000	population	20595360
5	Brazil	1999	cases	37737
6	Brazil	1999	population	172006362
7	Brazil	2000	cases	80488
8	Brazil	2000	population	174504898
9	China	1999	cases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



	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

Splitting values into multiple columns

- In table3, the rate variable needs to be split into two columns



	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

Separate()

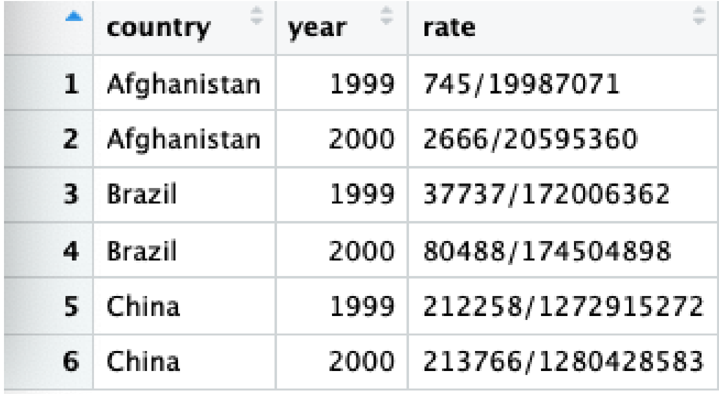
Arguments

1. col: The name of the existing variable whose values you want to split
2. into: The name of new variables where the split values will be moved into
3. sep: The string used to identify where to make the split
4. convert: whether you want the new variables to be converted to numeric

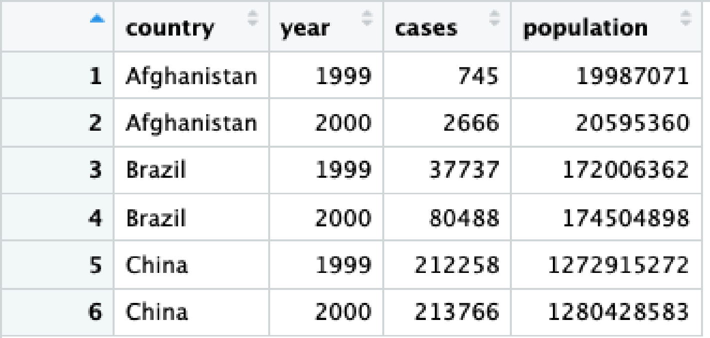
```
table3_separated <- table3%>%
 separate(col=rate,
                   into = c("cases", "population"),
                   sep = "/”,
                  convert=TRUE)
```

Separate()

```
table3_separated <- table3%>%
 separate(col=rate,
                   into = c("cases", "population"),
                   sep = "/",
                  convert=TRUE)
```



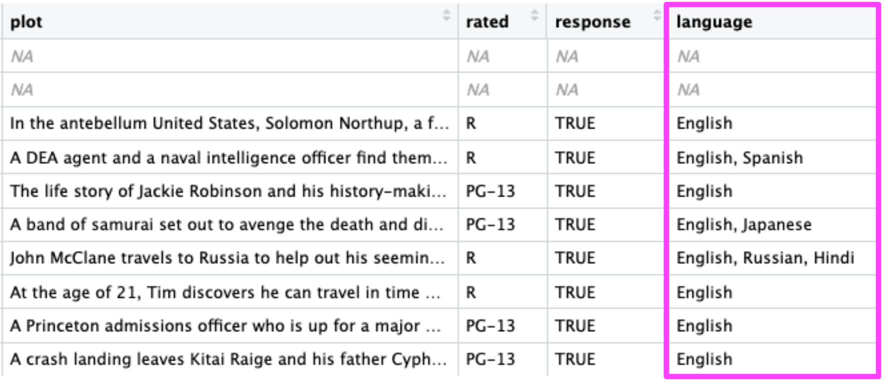
	country	year	rate
1	Afghanistan	1999	745/19987071
2	Afghanistan	2000	2666/20595360
3	Brazil	1999	37737/172006362
4	Brazil	2000	80488/174504898
5	China	1999	212258/1272915272
6	China	2000	213766/1280428583



	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

Splitting values into multiple rows

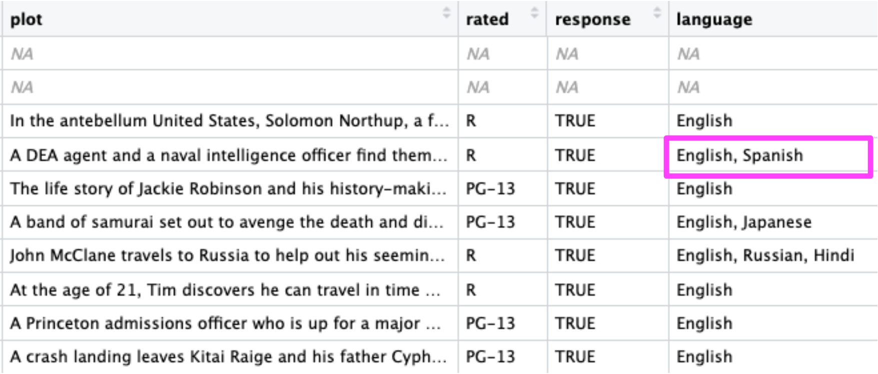
- In this table, the language column has multiple languages listed



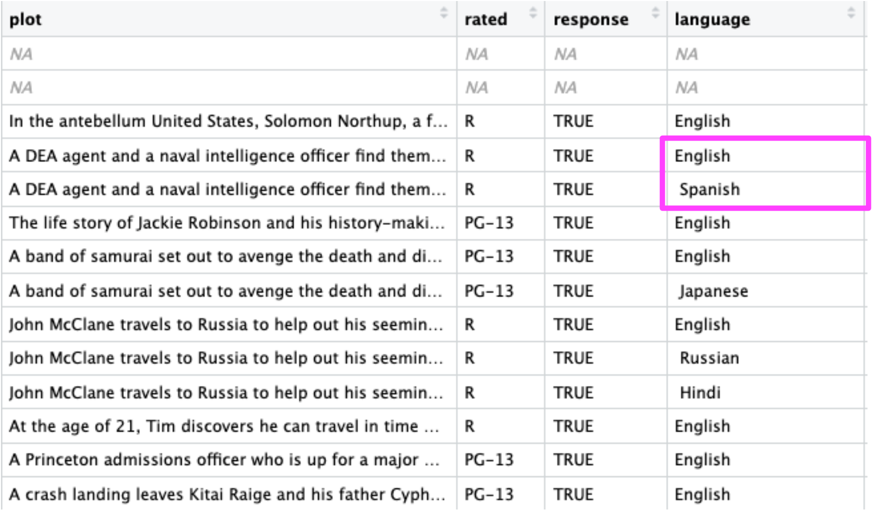
plot	rated	response	language
NA	NA	NA	NA
NA	NA	NA	NA
In the antebellum United States, Solomon Northup, a f...	R	TRUE	English
A DEA agent and a naval intelligence officer find them...	R	TRUE	English, Spanish
The life story of Jackie Robinson and his history-maki...	PG-13	TRUE	English
A band of samurai set out to avenge the death and di...	PG-13	TRUE	English, Japanese
John McClane travels to Russia to help out his seemin...	R	TRUE	English, Russian, Hindi
At the age of 21, Tim discovers he can travel in time ...	R	TRUE	English
A Princeton admissions officer who is up for a major ...	PG-13	TRUE	English
A crash landing leaves Kitai Raige and his father Cyph...	PG-13	TRUE	English

Splitting values into multiple rows

- Let’s create a new row for every language listed in a movie



plot	rated	response	language
NA	NA	NA	NA
NA	NA	NA	NA
In the antebellum United States, Solomon Northup, a f…	R	TRUE	English
A DEA agent and a naval intelligence officer find them…	R	TRUE	English, Spanish
The life story of Jackie Robinson and his history–maki…	PG–13	TRUE	English
A band of samurai set out to avenge the death and di…	PG–13	TRUE	English, Japanese
John McClane travels to Russia to help out his seemin…	R	TRUE	English, Russian, Hindi
At the age of 21, Tim discovers he can travel in time …	R	TRUE	English
A Princeton admissions officer who is up for a major …	PG–13	TRUE	English
A crash landing leaves Kitai Raige and his father Cyph…	PG–13	TRUE	English



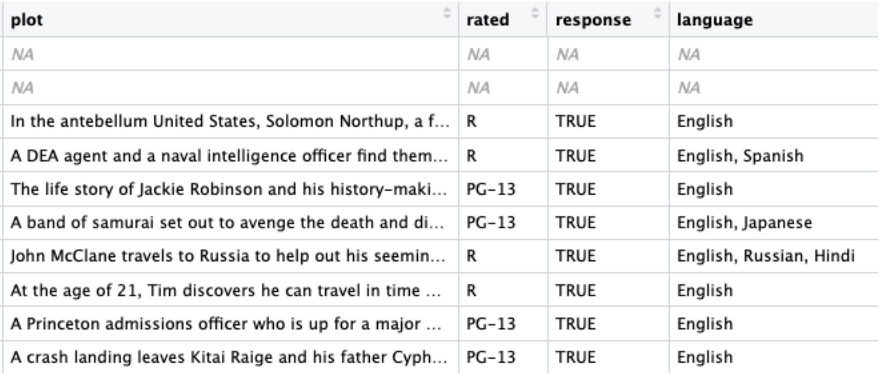
plot	rated	response	language
NA	NA	NA	NA
NA	NA	NA	NA
In the antebellum United States, Solomon Northup, a f…	R	TRUE	English
A DEA agent and a naval intelligence officer find them…	R	TRUE	English
A DEA agent and a naval intelligence officer find them…	R	TRUE	Spanish
The life story of Jackie Robinson and his history–maki…	PG–13	TRUE	English
A band of samurai set out to avenge the death and di…	PG–13	TRUE	English
A band of samurai set out to avenge the death and di…	PG–13	TRUE	Japanese
John McClane travels to Russia to help out his seemin…	R	TRUE	English
John McClane travels to Russia to help out his seemin…	R	TRUE	Russian
John McClane travels to Russia to help out his seemin…	R	TRUE	Hindi
At the age of 21, Tim discovers he can travel in time …	R	TRUE	English
A Princeton admissions officer who is up for a major …	PG–13	TRUE	English
A crash landing leaves Kitai Raige and his father Cyph…	PG–13	TRUE	English

separate_rows()

We need to specify:

1. The variable whose values you want to split
2. The character used to identify where to make the split

```
movies_language <- movies%>%
 separate_rows(language,
                        sep = ”,")
```



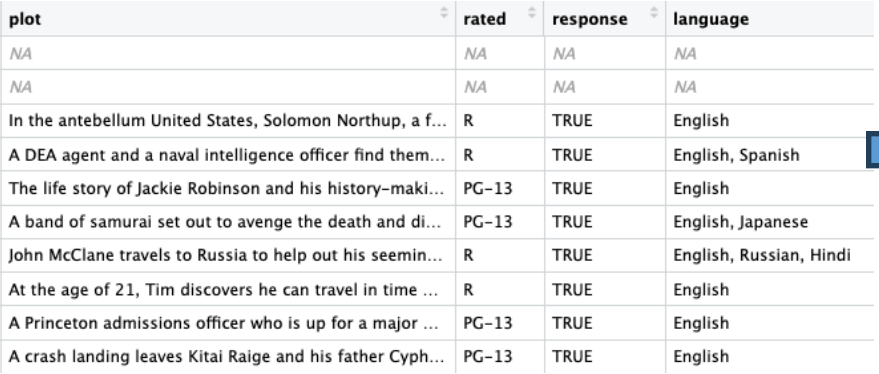
plot	rated	response	language
NA	NA	NA	NA
NA	NA	NA	NA
In the antebellum United States, Solomon Northup, a f...	R	TRUE	English
A DEA agent and a naval intelligence officer find them...	R	TRUE	English, Spanish
The life story of Jackie Robinson and his history-maki...	PG-13	TRUE	English
A band of samurai set out to avenge the death and di...	PG-13	TRUE	English, Japanese
John McClane travels to Russia to help out his seemin...	R	TRUE	English, Russian, Hindi
At the age of 21, Tim discovers he can travel in time ...	R	TRUE	English
A Princeton admissions officer who is up for a major ...	PG-13	TRUE	English
A crash landing leaves Kitai Raige and his father Cyph...	PG-13	TRUE	English

https://pollev.com/vsovero

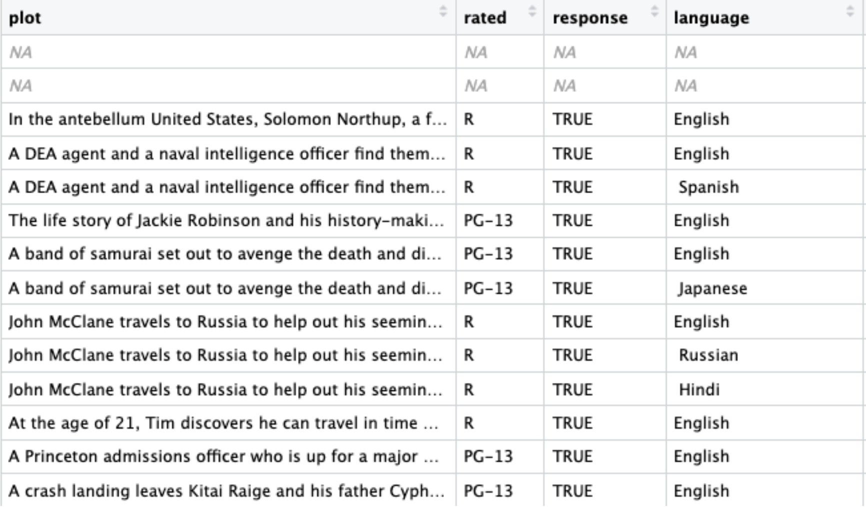
separate_rows()

- Result: row for every movie-language pair

```
movies_language <- movies%>%
 separate_rows(language,
                          sep = ”,")
```



plot	rated	response	language
NA	NA	NA	NA
NA	NA	NA	NA
In the antebellum United States, Solomon Northup, a f...	R	TRUE	English
A DEA agent and a naval intelligence officer find them...	R	TRUE	English, Spanish
The life story of Jackie Robinson and his history-maki...	PG-13	TRUE	English
A band of samurai set out to avenge the death and di...	PG-13	TRUE	English, Japanese
John McClane travels to Russia to help out his seemin...	R	TRUE	English, Russian, Hindi
At the age of 21, Tim discovers he can travel in time ...	R	TRUE	English
A Princeton admissions officer who is up for a major ...	PG-13	TRUE	English
A crash landing leaves Kitai Raige and his father Cyph...	PG-13	TRUE	English



plot	rated	response	language
NA	NA	NA	NA
NA	NA	NA	NA
In the antebellum United States, Solomon Northup, a f...	R	TRUE	English
A DEA agent and a naval intelligence officer find them...	R	TRUE	English
A DEA agent and a naval intelligence officer find them...	R	TRUE	Spanish
The life story of Jackie Robinson and his history-maki...	PG-13	TRUE	English
A band of samurai set out to avenge the death and di...	PG-13	TRUE	English
A band of samurai set out to avenge the death and di...	PG-13	TRUE	Japanese
John McClane travels to Russia to help out his seemin...	R	TRUE	English
John McClane travels to Russia to help out his seemin...	R	TRUE	Russian
John McClane travels to Russia to help out his seemin...	R	TRUE	Hindi
At the age of 21, Tim discovers he can travel in time ...	R	TRUE	English
A Princeton admissions officer who is up for a major ...	PG-13	TRUE	English
A crash landing leaves Kitai Raige and his father Cyph...	PG-13	TRUE	English

Exercise

Use the movies data frame to do the following:

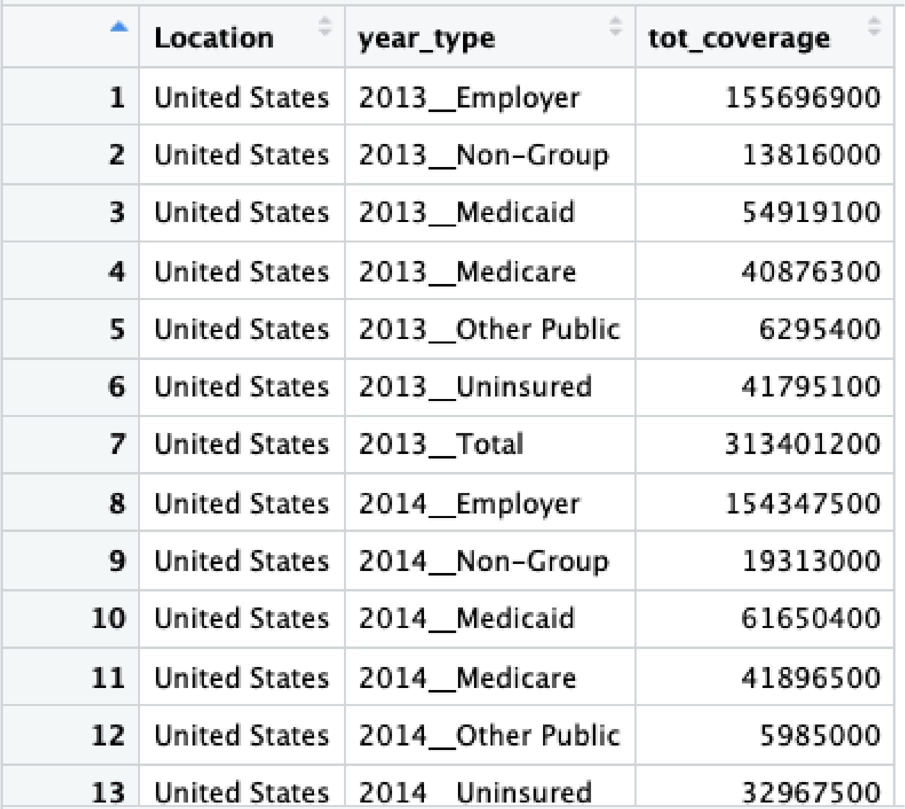
a) calculate the average domestic gross by genre (keep the top 5)

b) separate the genre variable, then calculate the average domestic gross by genre (keep the top 5)

c) Why are these tables different?

Final Exercise

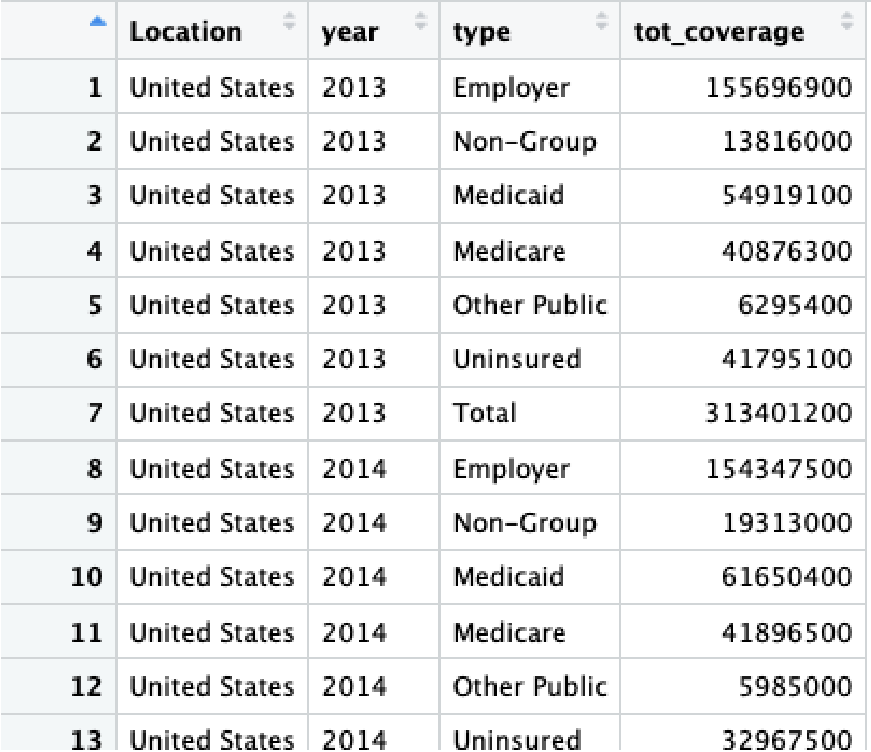
Step 1: tidy the data so it looks like the table on the right



	Location	year_type	tot_coverage
1	United States	2013__Employer	155696900
2	United States	2013__Non-Group	13816000
3	United States	2013__Medicaid	54919100
4	United States	2013__Medicare	40876300
5	United States	2013__Other Public	6295400
6	United States	2013__Uninsured	41795100
7	United States	2013__Total	313401200
8	United States	2014__Employer	154347500
9	United States	2014__Non-Group	19313000
10	United States	2014__Medicaid	61650400
11	United States	2014__Medicare	41896500
12	United States	2014__Other Public	5985000
13	United States	2014__Uninsured	32967500

Final Exercise

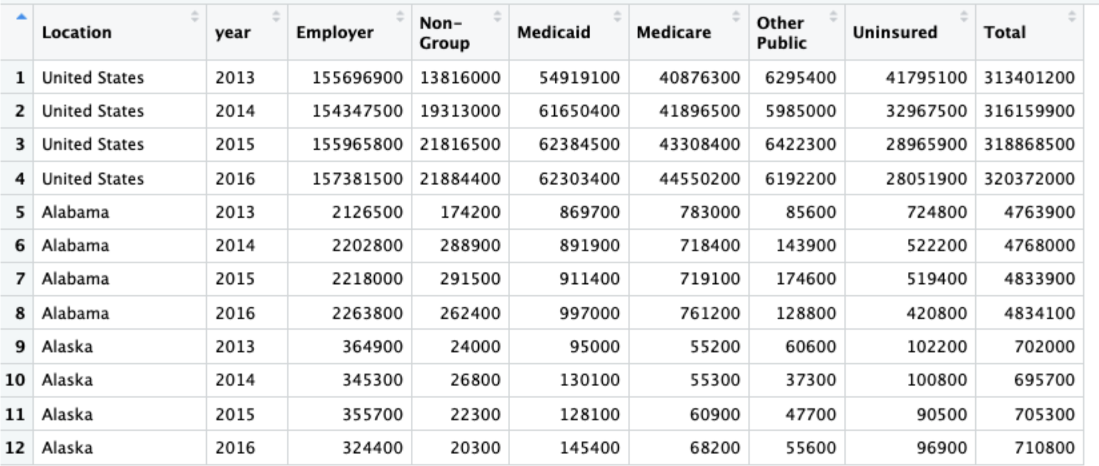
Step 2: tidy the data so it looks like the table on the right



	Location	year	type	tot_coverage
1	United States	2013	Employer	155696900
2	United States	2013	Non-Group	13816000
3	United States	2013	Medicaid	54919100
4	United States	2013	Medicare	40876300
5	United States	2013	Other Public	6295400
6	United States	2013	Uninsured	41795100
7	United States	2013	Total	313401200
8	United States	2014	Employer	154347500
9	United States	2014	Non-Group	19313000
10	United States	2014	Medicaid	61650400
11	United States	2014	Medicare	41896500
12	United States	2014	Other Public	5985000
13	United States	2014	Uninsured	32967500

Final Exercise

Step 3: tidy the data so it looks like the table on the right



	Location	year	Employer	Non-Group	Medicaid	Medicare	Other Public	Uninsured	Total
1	United States	2013	155696900	13816000	54919100	40876300	6295400	41795100	313401200
2	United States	2014	154347500	19313000	61650400	41896500	5985000	32967500	316159900
3	United States	2015	155965800	21816500	62384500	43308400	6422300	28965900	318868500
4	United States	2016	157381500	21884400	62303400	44550200	6192200	28051900	320372000
5	Alabama	2013	2126500	174200	869700	783000	85600	724800	4763900
6	Alabama	2014	2202800	288900	891900	718400	143900	522200	4768000
7	Alabama	2015	2218000	291500	911400	719100	174600	519400	4833900
8	Alabama	2016	2263800	262400	997000	761200	128800	420800	4834100
9	Alaska	2013	364900	24000	95000	55200	60600	102200	702000
10	Alaska	2014	345300	26800	130100	55300	37300	100800	695700
11	Alaska	2015	355700	22300	128100	60900	47700	90500	705300
12	Alaska	2016	324400	20300	145400	68200	55600	96900	710800