Clase 1.3
Acomodo de datos

Marcos Rosetti y Luis Pacheco-Cobos

Estadística y Manejo de Datos con R (EMDR) — Virtual

Wrangling (arreo) de datos

Acomodo de datos

Acomodo de datos: `tidyr`

Acomodo de datos: `tidyr`

- La filosofia de `tidyr` es tener una forma estandarizada de acomodar los datos.
  - Cada *variable* es una **columna**.
  - Cada *observación* es una **fila**.
  - Tenemos *un solo un tipo* de información en **cada celda**.
  - Así la información resulta fácil de convertir y graficar.

Acomodo de datos: `tidyr`

- Formato largo/longitudinal vs ancho.

Formato "ancho"

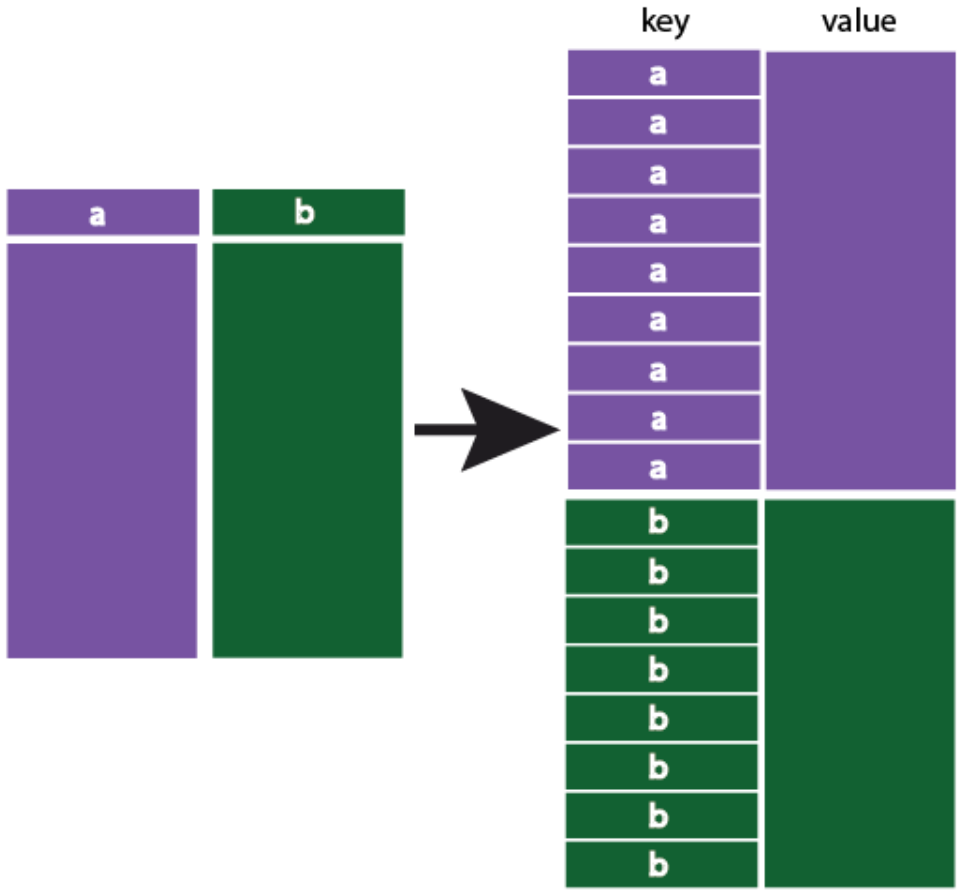
name	gender	age	height	weight
Matt	M	20	180	80
Thomas	M	60	200	100
Cate	F	30	150	50

Formato "largo"

name	gender	feature	value
Matt	M	age	20
Matt	M	height	180
Matt	M	weight	80
Thomas	M	age	60
Thomas	M	height	200
Thomas	M	weight	100
Cate	F	age	30
Cate	F	height	150
Cate	F	weight	50

Acomodo de datos: `tidyr`

- `gather()` sirve para pasar de formato **ancho** a formato **largo**.

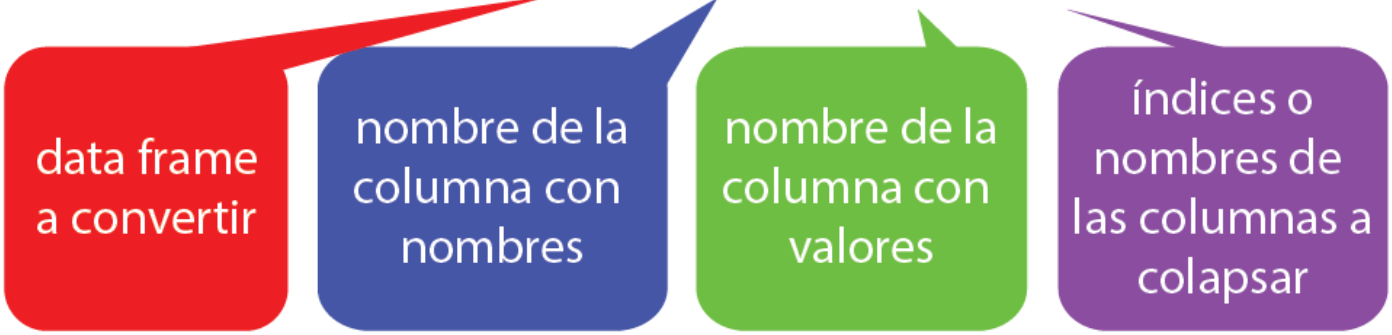


Acomodo de datos: `tidyr`

gather() colapsa múltiples columnas en dos columnas con:

1. una columna **key**, que contiene los nombres
2. una columna **value**, que contiene los valores

gather(casos, key, value, i:j)



Acomodo de datos: `tidyr`

- Veamos unos datos: ¿qué problema tienen?

```
library(tidyr)
```

```
table4b
```

```
## # A tibble: 3 × 3
##   country          `1999`     `2000`
## * <chr>             <int>      <int>
## 1 Afghanistan   19987071   20595360
## 2 Brazil       172006362  174504898
## 3 China       1272915272 1280428583
```

Acomodo de datos: `tidyr`

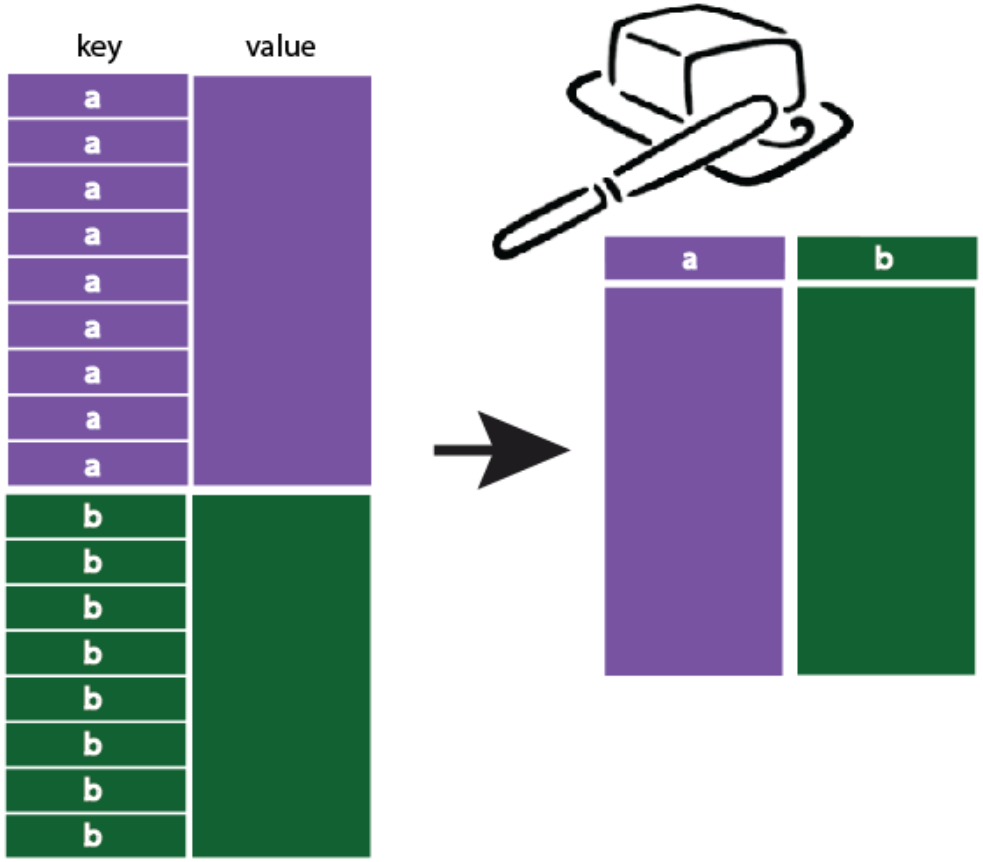
- Los acomodamos o dejamos `tidy` (ordenados) con:

```
tidy.4b <- gather (table4b, key ="year", value="population", -country )
tidy.4b
```

```
## # A tibble: 6 × 3
##   country     year  population
##   <chr>       <chr>      <int>
## 1 Afghanistan 1999    19987071
## 2 Brazil      1999   172006362
## 3 China       1999  1272915272
## 4 Afghanistan 2000    20595360
## 5 Brazil      2000   174504898
## 6 China       2000  1280428583
```

Acomodo de datos: `tidyr`

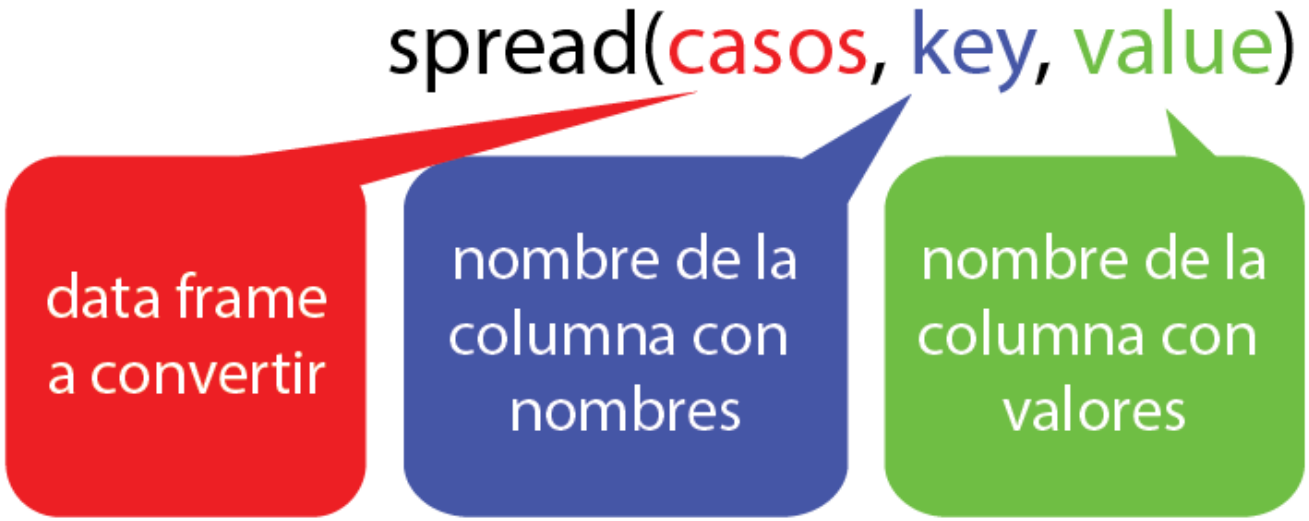
- `spread()` sirve para pasar de formato **largo** a formato **ancho**.



Acomodo de datos: `tidyr`

spread() expande dos columnas en múltiples columnas

1. una columna **key**, que contiene los nombres
2. una columna **value**, que contiene los valores



Acomodo de datos: `tidyr`

- Veamos unos datos: ¿qué problema tienen?

```
head(table2)
```

```
## # A tibble: 6 × 4
##   country      year type           count
##   <chr>       <int> <chr>          <int>
## 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
```

Acomodo de datos: `tidyr`

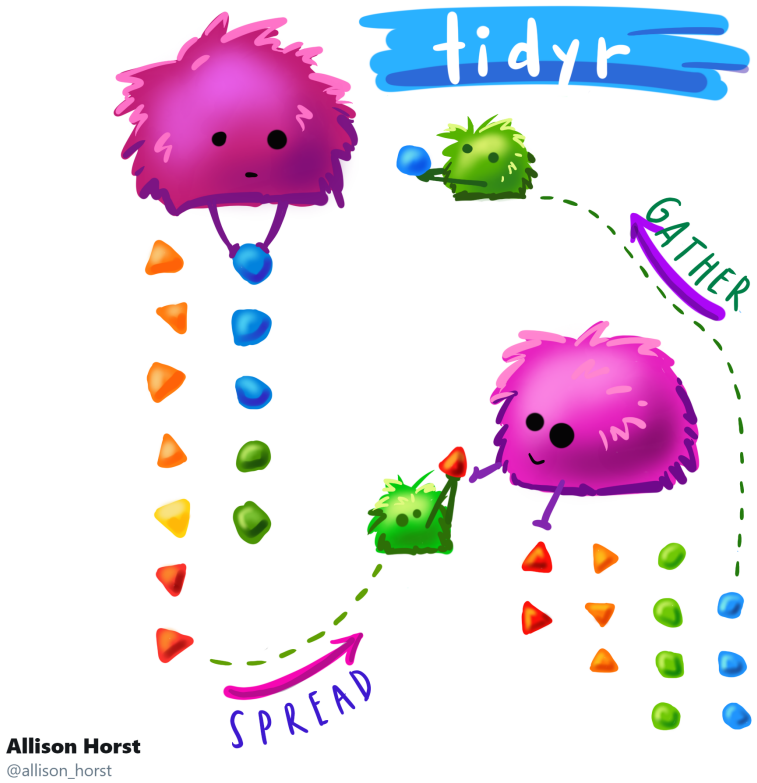
- Los acomodamos o dejamos `tidy` (ordenados) con:

```
tidy.2 <- spread(table2, type, count)
tidy.2
```

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

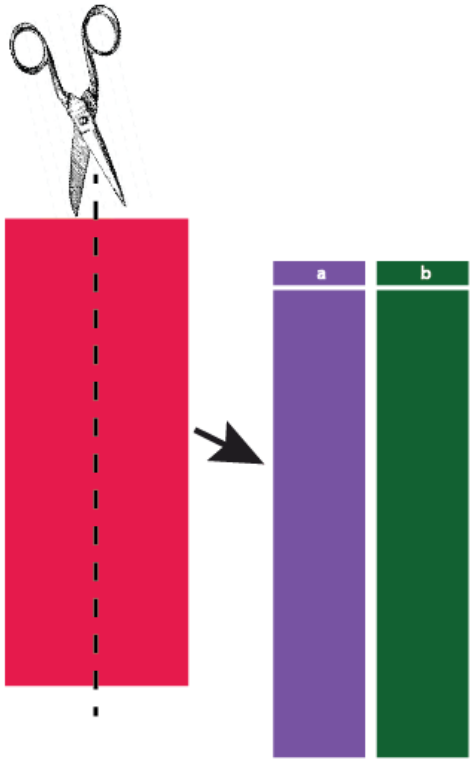
Acomodo de datos: `tidyr`

- Y de largo a ancho y viceversa.

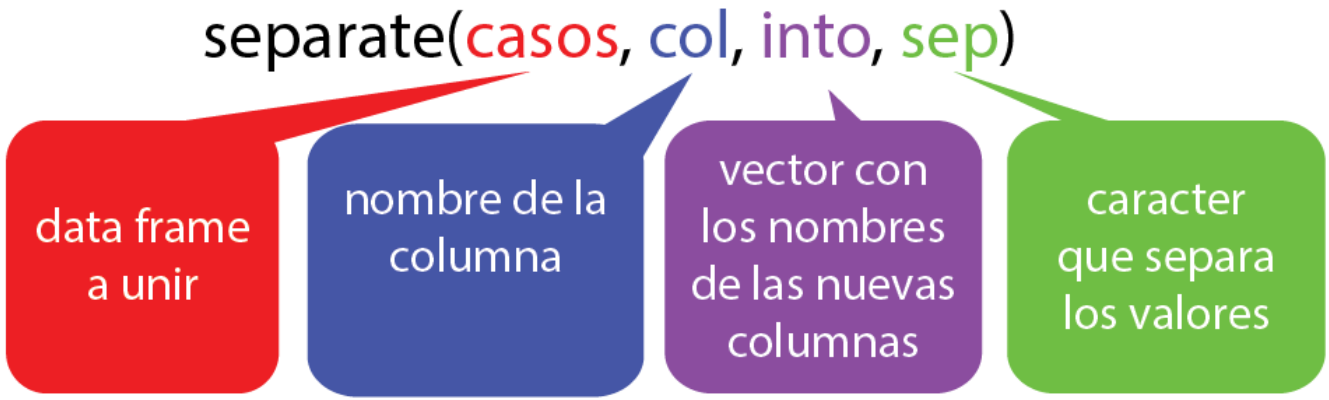


Acomodo de datos: `tidyr`

- `separate()` divide variables complejas en variables individuales.



Acomodo de datos: `tidyr`



Acomodo de datos: `tidyr`

- Veamos unos datos: ¿qué problema tienen?

```
head(table3)
```

```
## # A tibble: 6 × 3
##   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
```

Acomodo de datos: `tidyr`

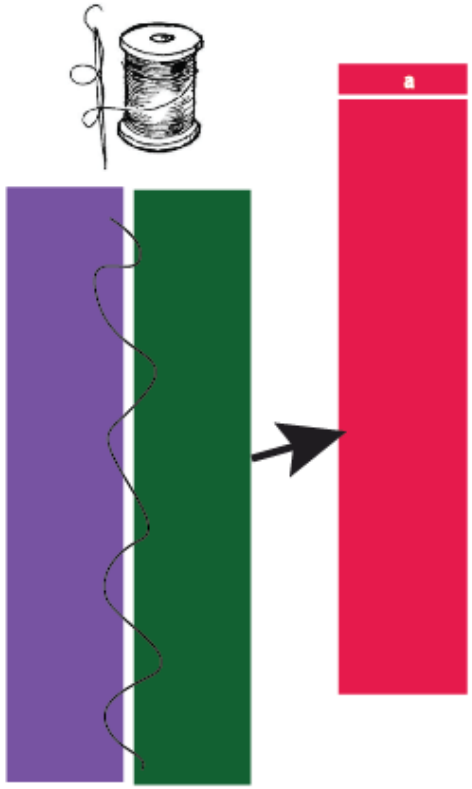
- Los acomodamos o dejamos `tidy` (ordenados) con:

```
tidy.3 <- separate(table3, rate, c("cases", "population"), sep = "/")
tidy.3
```

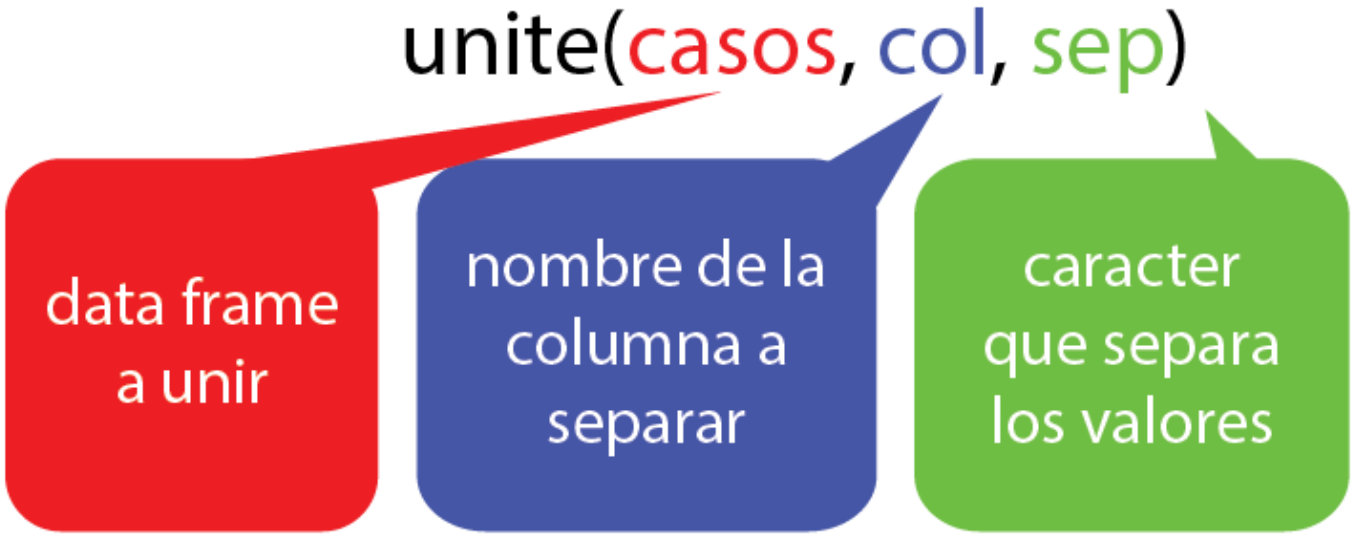
```
## # A tibble: 6 × 4
##   country      year cases  population
##   <chr>       <int> <chr>  <chr>
## 1 Afghanistan  1999 745    19987071
## 2 Afghanistan  2000 2666   20595360
## 3 Brazil       1999 37737  172006362
## 4 Brazil       2000 80488  174504898
## 5 China        1999 212258 1272915272
## 6 China        2000 213766 1280428583
```

Acomodo de datos: `tidyr`

- `unite()` une variables en una sola columna.



Acomodo de datos: `tidyr`



Acomodo de datos: `tidyr`

- Veamos unos datos: ¿qué problema tienen?

```
head(table5)
```

```
## # A tibble: 6 × 4
##   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
```

Acomodo de datos: `tidyr`

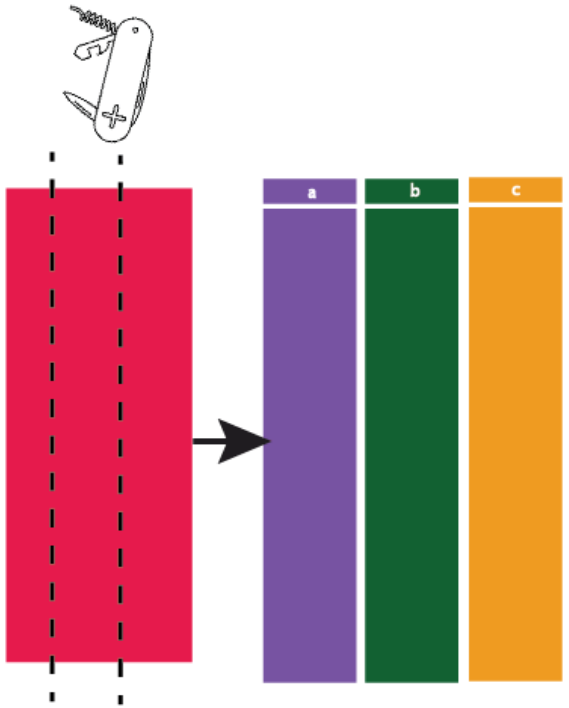
- Los acomodamos o dejamos `tidy` (ordenados) con:

```
tidy.5 <- unite(table5, "year", 2:3, sep ="")
tidy.5
```

```
## # A tibble: 6 × 3
##   country     year  rate
##   <chr>       <chr> <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
```

Acomodo de datos: `tidyr`

- `extract()` divide una variable usando múltiples criterios y genera múltiples columnas.



Licencia CC BY