



Universidad del
Rosario

Facultad de
Economía

Laboratorio de
Finanzas

Escuela de Ingeniería,
Ciencia y Tecnología

TALLER DE **DATA SCIENCE EN R**



Sueño**SER**

Potencializa tu perfil
académico y profesional.

6:00 p.m a 8:00 p.m

10 al **12** de octubre

Trasciende,
construye
y Lidera

Rosarios
con
Propósito

Contenido

- ¿Qué es ordenar los datos?
- Herramientas a usar
- Reglas para ordenar un dataset
- Operaciones principales de tidy



¿Qué es ordenar los datos?

Representación de los datos

¿Qué es ordenar los datos?

Existen varias formas de representar la misma información

	Prueba A	Prueba B
Manuel Martinez	-	2
Jhon Salcedo	16	11
Carlos Zarate	3	1

Persona	Prueba	Resultado
Manuel Martinez	A	-
Jhon Salcedo	A	16
Carlos Zarate	A	3
Manuel Martinez	B	2
Jhon Salcedo	B	11
Carlos Zarate	B	1

	Manuel Martinez	Jhon Salcedo	Carlos Zarate
Prueba A	-	16	3
Prueba B	2	11	1

Semántica

¿Qué es ordenar los datos?

Dataset: Es una colección de valores. Atributos medidos en unidades observacionales.

División según el tipo de atributo:

- Cualitativos (strings)
- Cuantitativos (números)

Cada valor en un dataset pertenece a una **variable** y a una **observación**

- **Variable:** Todos los valores que miden el mismo atributo en todas las unidades observacionales
- **Observación:** Todos los valores que toman los atributos, medidos en una unidad observacional

¿Qué son variables y qué son observaciones en este dataset?

	Prueba A	Prueba B
Manuel Martinez	-	2
Jhon Salcedo	16	11
Carlos Zarate	3	1

Una transformación de ordenamiento

¿Qué es ordenar los datos?

	Prueba A	Prueba B
Manuel Martinez	-	2
Jhon Salcedo	16	11
Carlos Zarate	3	1



Persona	Prueba	Resultado
Manuel Martinez	A	-
Jhon Salcedo	A	16
Carlos Zarate	A	3
Manuel Martinez	B	2
Jhon Salcedo	B	11
Carlos Zarate	B	1

Una transformación de ordenamiento

¿Qué es ordenar los datos?

	Prueba A	Prueba B
Manuel Martinez	-	2
Jhon Salcedo	16	11
Carlos Zarate	3	1



Persona	Prueba	Resultado
Manuel Martinez	A	-
Jhon Salcedo	A	16
Carlos Zarate	A	3
Manuel Martinez	B	2
Jhon Salcedo	B	11
Carlos Zarate	B	1

Una transformación de ordenamiento

¿Qué es ordenar los datos?

	Prueba A	Prueba B
Manuel Martinez	-	2
Jhon Salcedo	16	11
Carlos Zarate	3	1



Persona	Prueba	Resultado
Manuel Martinez	A	-
Jhon Salcedo	A	16
Carlos Zarate	A	3
Manuel Martinez	B	2
Jhon Salcedo	B	11
Carlos Zarate	B	1

Una transformación de ordenamiento

¿Qué es ordenar los datos?

	Prueba A	Prueba B
Manuel Martinez	-	2
Jhon Salcedo	16	11
Carlos Zarate	3	1



Persona	Prueba	Resultado
Manuel Martinez	A	-
Jhon Salcedo	A	16
Carlos Zarate	A	3
Manuel Martinez	B	2
Jhon Salcedo	B	11
Carlos Zarate	B	1

Una transformación de ordenamiento

¿Qué es ordenar los datos?

	Prueba A	Prueba B
Manuel Martinez	-	2
Jhon Salcedo	16	11
Carlos Zarate	3	1



Persona	Prueba	Resultado
Manuel Martinez	A	-
Jhon Salcedo	A	16
Carlos Zarate	A	3
Manuel Martinez	B	2
Jhon Salcedo	B	11
Carlos Zarate	B	1

Una transformación de ordenamiento

¿Qué es ordenar los datos?

	Prueba A	Prueba B
Manuel Martinez	-	2
Jhon Salcedo	16	11
Carlos Zarate	3	1



Persona	Prueba	Resultado
Manuel Martinez	A	-
Jhon Salcedo	A	16
Carlos Zarate	A	3
Manuel Martinez	B	2
Jhon Salcedo	B	11
Carlos Zarate	B	1

Una transformación de ordenamiento

¿Qué es ordenar los datos?

	Prueba A	Prueba B
Manuel Martinez	-	2
Jhon Salcedo	16	11
Carlos Zarate	3	1



Persona	Prueba	Resultado
Manuel Martinez	A	-
Jhon Salcedo	A	16
Carlos Zarate	A	3
Manuel Martinez	B	2
Jhon Salcedo	B	11
Carlos Zarate	B	1



Herramientas a usar

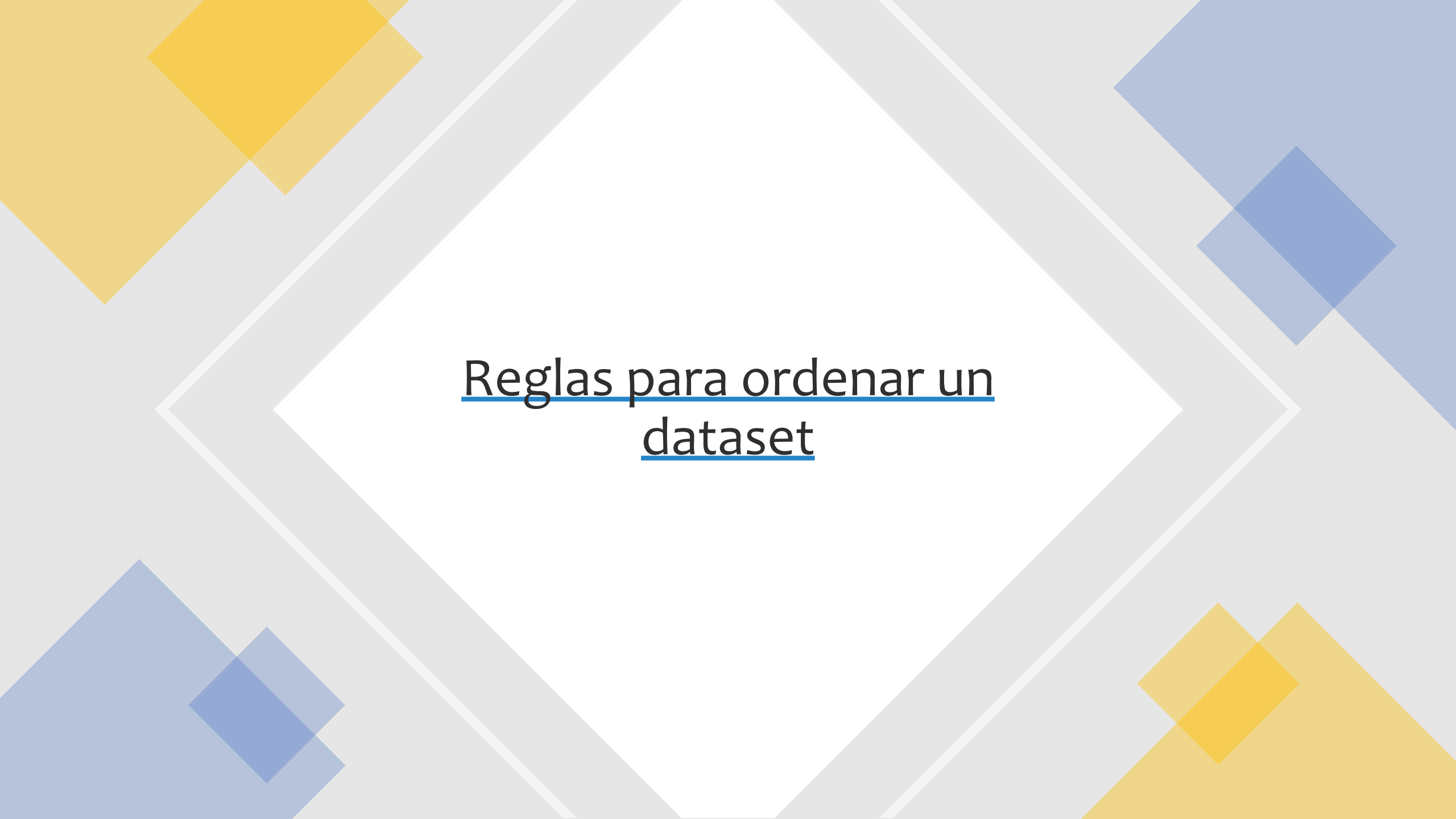
Programas y paquetes a usar

Algunos conceptos preliminares

■ Librería tidyverse

```
# install.packages("tidyverse")  
library(tidyverse)
```

```
> library(tidyverse)  
-- Attaching packages ----- tidyverse 1.3.1 --  
v ggplot2 3.3.3      v purrr  0.3.4  
v tibble  3.1.1      v dplyr  1.0.5  
v tidyr   1.1.3      v stringr 1.4.0  
v readr   1.4.0      v forcats 0.5.1  
-- Conflicts ----- tidyverse_conflicts() --  
x dplyr::filter() masks stats::filter()  
x dplyr::lag()    masks stats::lag()
```



Reglas para ordenar un dataset

Características de los datos ordenados

Reglas para ordenar un dataset

- Cada variable forma una columna
- Cada observación forma una fila
- Cada tipo de unidad observacional forma una tabla

	Prueba A	Prueba B
Manuel Martinez	-	2
Jhon Salcedo	16	11
Carlos Zarate	3	1



Persona	Prueba	Resultado
Manuel Martinez	A	-
Jhon Salcedo	A	16
Carlos Zarate	A	3
Manuel Martinez	B	2
Jhon Salcedo	B	11
Carlos Zarate	B	1



Reglas para ordenar datos

Reglas para ordenar un dataset

- Cada variable forma una columna

country	year	cases	population
Afghanistan	1999	17	19994071
Afghanistan	2000	1666	20005360
Brazil	1999	3737	172006362
Brazil	2000	8488	174004898
China	1999	21258	1272015272
China	2000	21766	128002583

variables

- Cada observación forma una fila

country	year	cases	population
Afghanistan	1999	17	19994071
Afghanistan	2000	1666	20005360
Brazil	1999	3737	172006362
Brazil	2000	8488	174004898
China	1999	21258	1272015272
China	2000	21766	128002583

observations

- Cada tipo de unidad observacional forma una tabla

country	year	cases	population
Afghanistan	99	17	19994071
Afghanistan	00	1666	20005360
Brazil	99	3737	172006362
Brazil	00	8488	174004898
China	99	21258	1272015272
China	00	21766	128002583

values

Problemas más frecuentes

Reglas para ordenar un dataset

- Los encabezados de las columnas son valores
- Hay múltiples variables en una columna
- Los valores están almacenados en filas y columnas
- Varios tipos de unidades observacionales en la misma tabla
- Una única unidad observacional está almacenada en tablas múltiples

Algunos ejemplos

Reglas para ordenar un dataset

Cargar la librería tidyverse

```
library(tidyverse)
```

```
table1  
table2  
table3  
table4a  
table4b
```

- Los encabezados de las columnas son valores
- Hay múltiples variables en una columna
- Los valores están almacenados en filas y columnas
- Varios tipos de unidades observacionales en la misma tabla
- Una única unidad observacional está almacenada en tablas múltiples

Ventajas

Reglas para ordenar un dataset

- Es una forma consistente de almacenar los datos
- La mayoría de funciones de R trabajan con vectores de datos

Ejemplo: Cálculo de información derivada

```
# Casos por cada 10,000 habitantes  
table1 %>%  
  mutate(rate = cases / population * 10000)
```

```
# Casos totales por año  
table1 %>%  
  count(year, wt = cases)
```

Ventajas

Reglas para ordenar un dataset

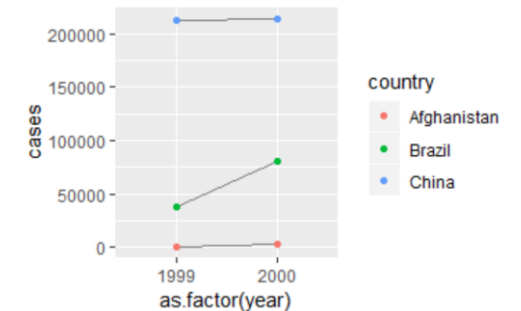
- Compatibilidad extra con otros paquetes del tidyverse

Ejemplo: Gráfico de los datos

Visualizar cambios a través del tiempo

```
library(ggplot2)
```

```
ggplot(table1, aes(year, cases)) +  
  geom_line(aes(group = country), colour = "grey50") +  
  geom_point(aes(colour = country))
```





Operaciones principales de tidyr

Cuatro operaciones de ordenamiento

- Separate
- Unite
- Gather
- Spread

table5

Operaciones principales de tidyr

country	century	year	rate
Afghanistan	19	99	745/19987071
Afghanistan	20	00	2666/20595360
Brazil	19	99	37737/172006362
Brazil	20	00	80488/174504898
China	19	99	212258/1272915272
China	20	00	213766/1280428583

Separate

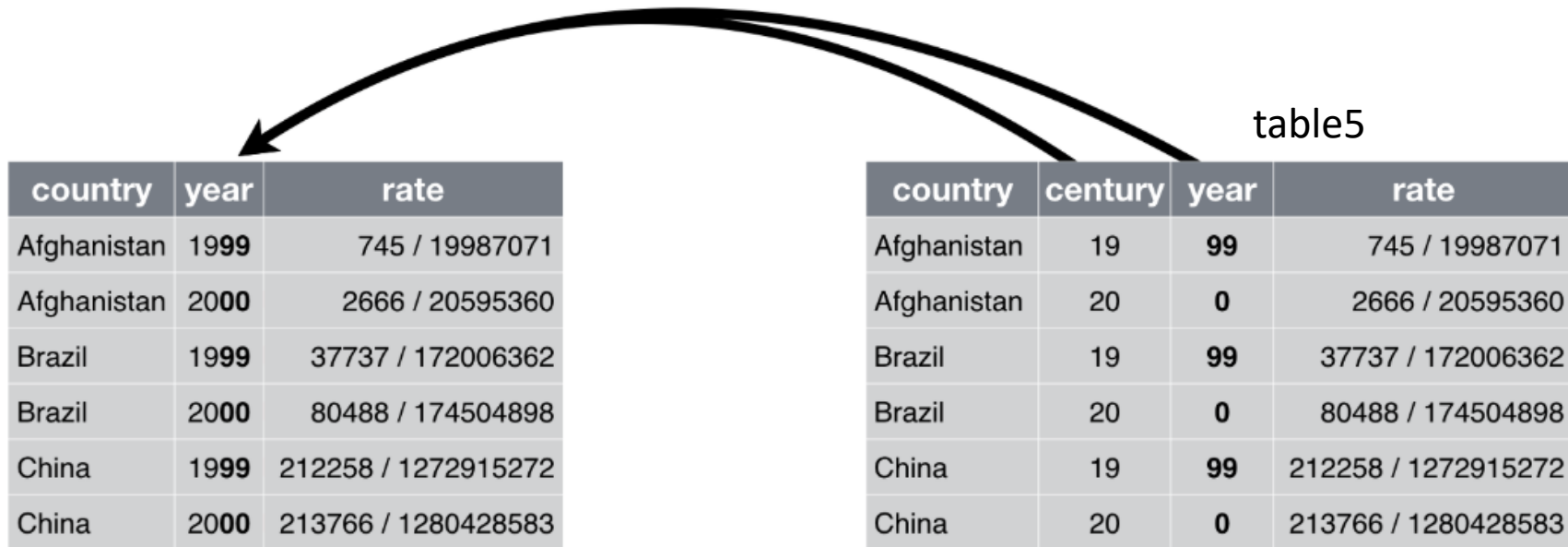
Operaciones principales de tidyr



```
table5 %>%  
  separate(rate, into = c("cases", "population"), sep = "/")
```

Unite

Operaciones principales de tidyr



```
table5 %>%  
  unite(new, century, year)
```

table4a

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

gather

Operaciones principales de tidyr

table4a

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

The diagram illustrates the 'gather' operation in tidyr. It shows a wide table on the right with columns 'country', '1999', and '2000'. Arrows point from the '1999' and '2000' columns to the 'year' column of a long table on the left. Another arrow points from the 'country' column to the 'country' column of the long table. The long table has columns 'country', 'year', and 'cases'. The data is reshaped from wide to long format.

```
table4a %>%
```

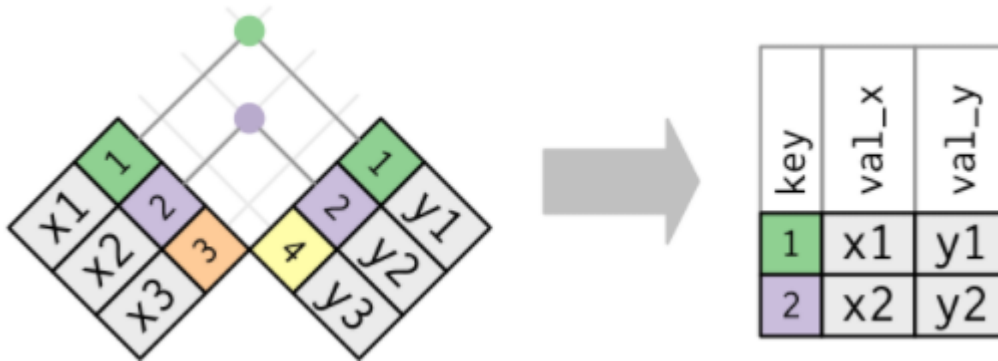
```
  gather(`1999`, `2000`, key = "year", value = "cases")
```

gather

Operaciones principales de tidyr

```
tidy4a <- table4a %>%  
  gather(`1999`, `2000`, key = "year", value = "cases")
```

```
tidy4b <- table4b %>%  
  gather(`1999`, `2000`, key = "year", value = "population")
```



```
left_join(tidy4a, tidy4b)
```

Universidad del Rosario | Escuela de Ingeniería, Ciencia y Tecnología

INTRODUCCIÓN A TIDYR

SÁBADO
16 DE OCTUBRE | 10:00 AM A 12:00 PM

Ordenamiento de datos para optimización de procesos

SALA KNUTH
MODALIDAD ACCESO REMOTO

Entrada Libre

ict_urosario • Siguiendo
Escuela de Ingeniería, Ciencia y Tecnol...

ict_urosario ¡Descubre una forma consistente de organizar tus datos en R usando el paquete tidyR!

No te pierdas el Taller Knuth de este sábado, insíbete en el link de la biografía o envíame un DM!

2 d

4 Me gusta
HACE 2 DÍAS

Agrega un comentario... Publicar

table2

Operaciones principales de tidyr

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

Separate

Operaciones principales de tidyr

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

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

```
table2 %>%  
  spread(key = type, value = count)
```


Gracias por tu asistencia y participación 😊

Contacto

✉ miguela.orjuela@urosario.edu.co

🌐 <https://www.linkedin.com/in/miguel-orjuela/>

🐙 <https://github.com/maorjuela73>

Links de interés

- <https://bookdown.org/rdpeng/rprogdatascience/>
- <https://www.rstudio.com/resources/cheatsheets/>