Paquetes de R

Las herramientas del científico de datos

Juan Manuel Moreno — jmmoreno@profesores.imf.com

IMF
Smart Education

ÍNDICE

IMF
Smart Education

01 Objetivos unidad 5

IMF
Smart Education

1.- Objetivos Unidad 5

- Saber filtrar los valores de un dataframe..
- Trabajar con las columnas de un dataframe.
- Agrupar un dataframe por columnas y mostrar ciertas estadísticas personalizadas como resultado de la previa agrupación.
- Obtener muestras sobre los datos de diferentes maneras.
- Obtener nuevas columnas en base a transformaciones sobre las ya existentes.
- Dividir una columna en n columnas nuevas.
- Compactar n columnas en una sola columna.
- Identificar conjuntos de clave:valor como columnas para pivotar sus valores.
- Identificar columnas que pueden ser representadas a través de una única columna con valores categóricos.

02

Procesamiento de datos con DPLYR

IMF
Smart Education

2.– Procesamiento de datos con DPLYR

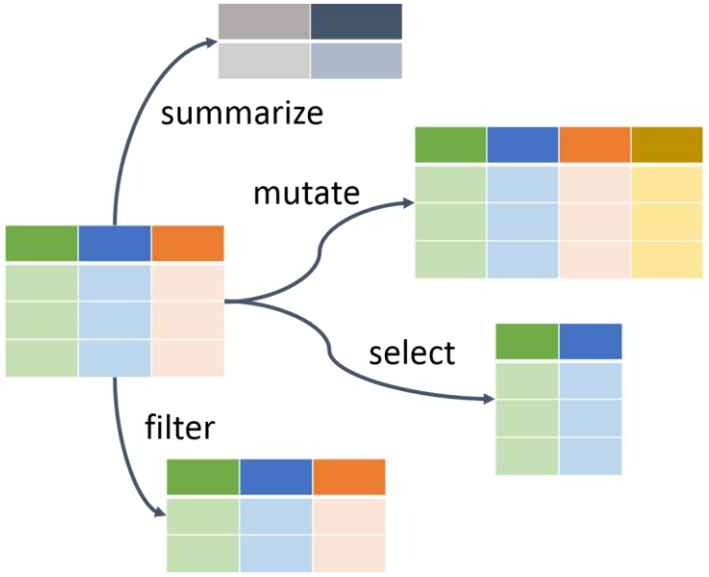
2.1.– DPLYR

•Se utilizan funciones similares a verbos para aplicar funciones a los datos.

• Mediante el operador (pipe) `%>%` podemos concatenar funciones.

• Funciones que se estudiarán:

- `filter`
- `select`
- `mutate`
- `transmute`
- `summarise`
- `group_by`
- `distinct`
- `rename`
- `arrange`
- `sample_n`
- `sample_frac`
- `recode`
- `ifelse`
- `case_when`



03

Procesamiento de datos con TIDYR

IMF Smart Education

3.- Procesamiento de datos con TIDYR

3.1.- TIDYR

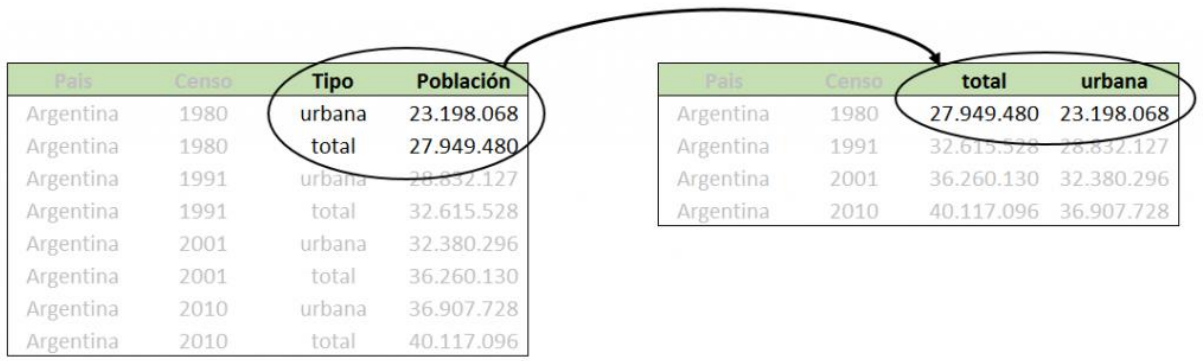
• Para un científico de datos, más del 80% de los datos se trabaja en arreglar (tidy) ordenar los datos, para ello, la librería tidyr incorpora potentes funciones para organizar los datos.

• Funciones que se estudiarán:

- **Spread**
- **Gather**
- **Separate**
- **Unite**

3.2.– TIDYR – Spread

- **`spread`**: Toma columnas a modo clave:valor para crear nuevas columnas en función de los valores únicos del campo clave



País	Censo	Tipo	Población
Argentina	1980	urbana	23.198.068
Argentina	1980	total	27.949.480
Argentina	1991	urbana	28.832.127
Argentina	1991	total	32.615.528
Argentina	2001	urbana	32.380.296
Argentina	2001	total	36.260.130
Argentina	2010	urbana	36.907.728
Argentina	2010	total	40.117.096

País	Censo	total	urbana
Argentina	1980	27.949.480	23.198.068
Argentina	1991	32.615.528	28.832.127
Argentina	2001	36.260.130	32.380.296
Argentina	2010	40.117.096	36.907.728

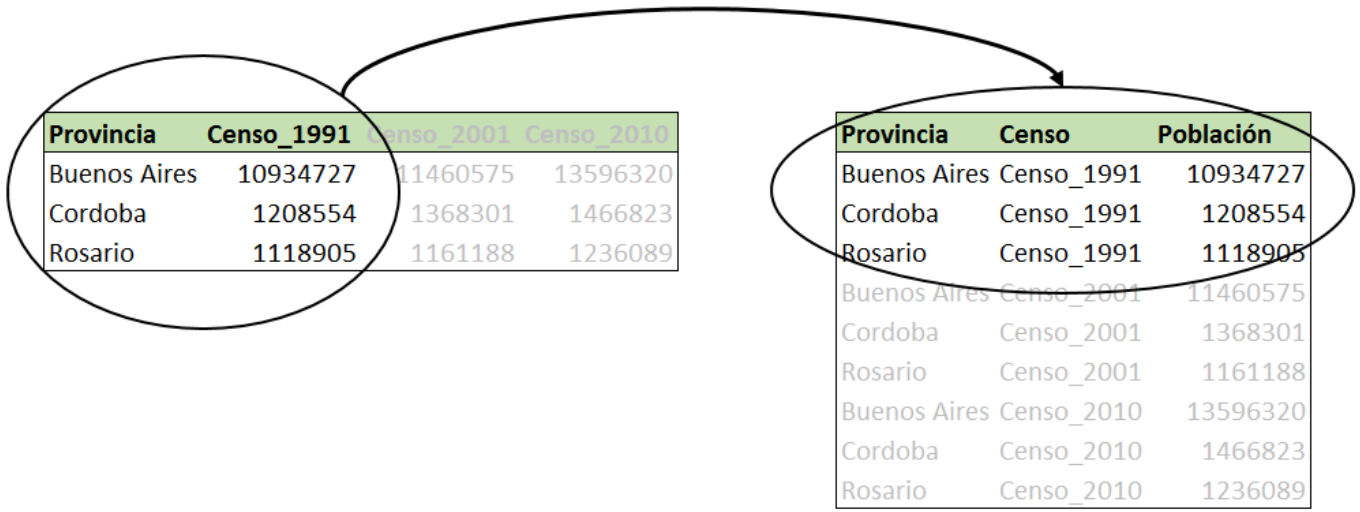
country	year	key	value
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999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

table2

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

3.3.– TIDYR – Gather

- **`gather`**: Lo contrario que `spread`, en lugar de pivotar por filas, pivota por columnas (aumenta la dimensionalidad vertical y reduce la horizontal)



Provincia	Censo_1991	Censo_2001	Censo_2010
Buenos Aires	10934727	11460575	13596320
Cordoba	1208554	1368301	1466823
Rosario	1118905	1161188	1236089

Provincia	Censo	Población
Buenos Aires	Censo_1991	10934727
Cordoba	Censo_1991	1208554
Rosario	Censo_1991	1118905
Buenos Aires	Censo_2001	11460575
Cordoba	Censo_2001	1368301
Rosario	Censo_2001	1161188
Buenos Aires	Censo_2010	13596320
Cordoba	Censo_2010	1466823
Rosario	Censo_2010	1236089

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

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

table4

3.4.– TIDYVERSE

- Tanto **dplyr** como **tidyr** y otras muchas librerías de procesamiento de datos se encuentran en el ecosistema de paquetes **tidyverse.** https://www.tidyverse.org/

R packages for data science

The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Install the complete tidyverse with:

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

Seguimiento práctico del contenido

A partir de aquí, veremos tanto dplyr como tidyr con los siguientes notebooks.

5_1_Dplyr.RMD
5_2_Tidyr.RMD

IMF

Smart Education