

# Herramientas Cuantitativas para el Análisis Político

[CP44] Maestría en Ciencia Política

Juan Pablo Ruiz Nicolini

Universidad Torcuato Di Tella

29/09/2020

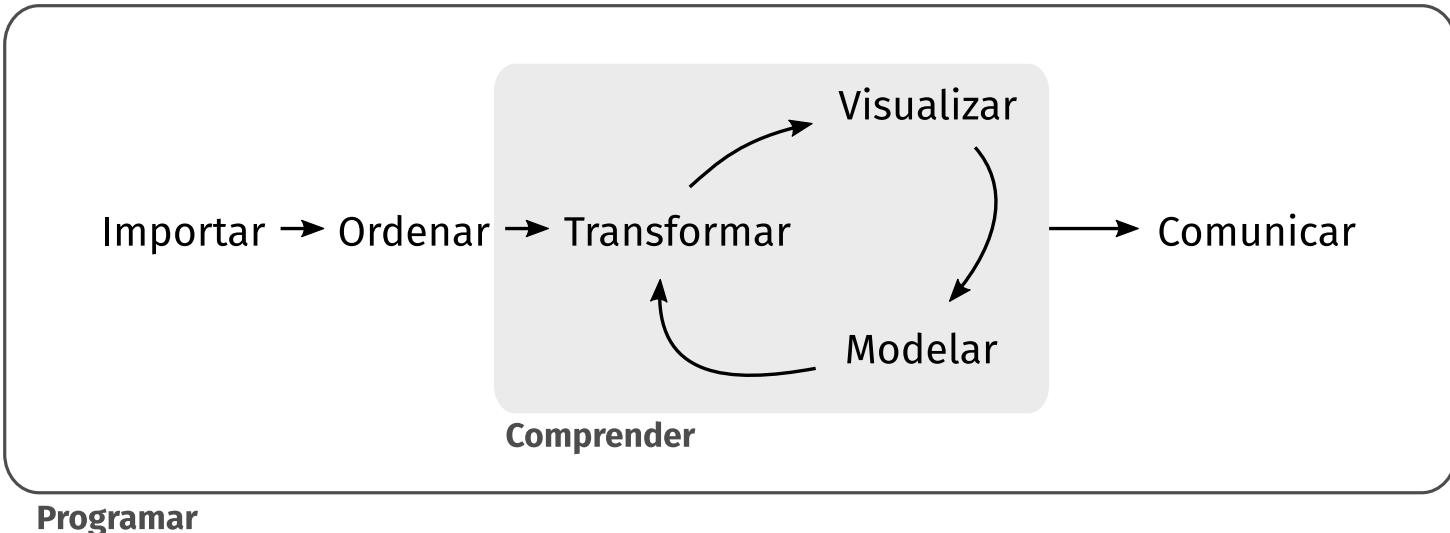
# SESIÓN 3

## Domar los datos (I)

/MetodosCiPol/

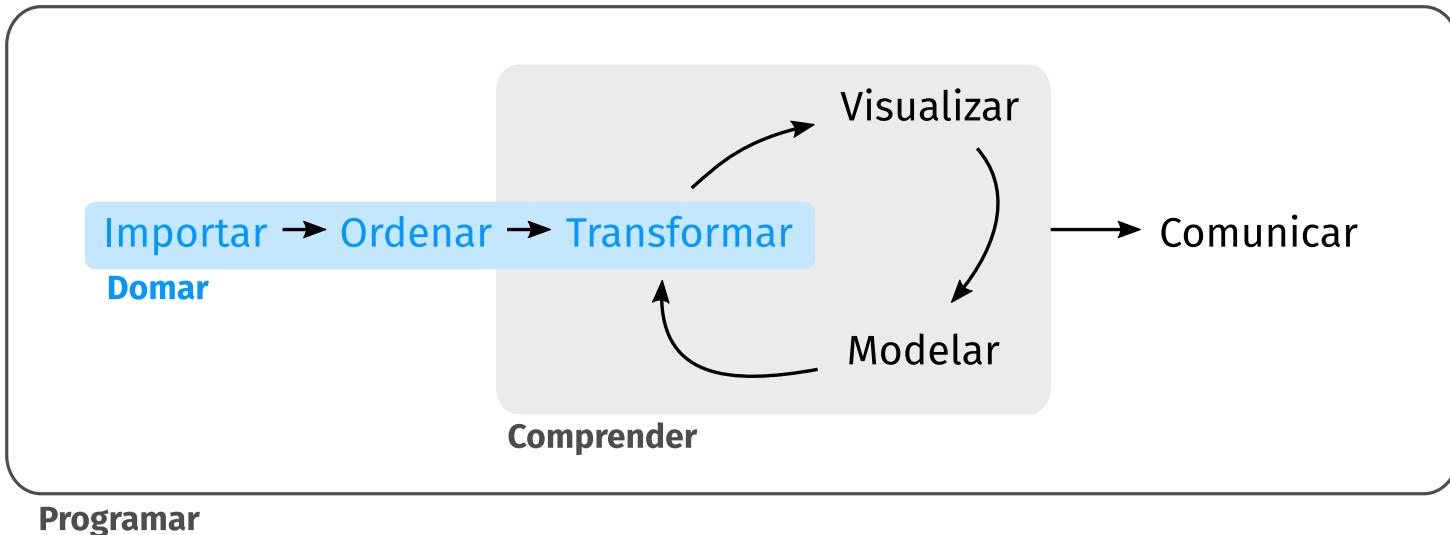
/MetodosCiPol/

# Ciencia de Datos



Recurso: <https://es.r4ds.hadley.nz/domardatos-intro.html>

# Ciencia de Datos - Domar Datos



Recurso: <https://es.r4ds.hadley.nz/domardatos-intro.html>

# Referencias

## IMPORTAR

1. Importar Datos, en [Wickham y Golemnud](#).
2. Carga de bases en [Urdinez y Cruz](#)
3. Importando datos a R en [Montané](#).

## PROCESAR

1. Transformación de Datos, en [Wickham y Golemnud](#)
2. Transformando nuestros Datos en [Montané](#)
3. Manejo de datos en [Urdinez y Cruz](#)
4. Poniendo los datos en forma en [Vazquez Brust](#)

# Importar | base

```
# no es necesario cargar paquetes para read.csv()

read.csv(file = "data/arg_presi_gral2003.csv")
##      codprov          depto coddepto circuito mesa electo
## 1           1    Balvanera Norte      11 0102   981
## 2           1    Balvanera Norte      11 0102   982
## 3           1    Balvanera Norte      11 0102   983
## 4           1    Balvanera Norte      11 0102   984
## 5           1    Balvanera Norte      11 0102   985
## 6           1    Balvanera Norte      11 0102   986
## 7           1    Balvanera Norte      11 0102   987
## 8           1    Balvanera Norte      11 0102   988
## 9           1    Balvanera Norte      11 0102   989
## 10          1    Balvanera Norte      11 0102   990
## 11          1    Balvanera Norte      11 0102 4003
## 12          1    Balvanera Norte      11 0102 4004
## 13          1    Balvanera Norte      11 0102 4005
## 14          1    Balvanera Norte      11 0102 4006
## 15          1    Balvanera Norte      11 0102 4007
## 16          1    Balvanera Norte      11 0102 4008
## 17          1    Balvanera Norte      11 0102 4009
## 18          1    Balvanera Norte      11 0102 4010
## 19          1    Balvanera Norte      11 0102 4011
```

# Importar | Tidy



(Ver **tibble** en sección *Ordenar*)

```
library(readr)

read_csv(file = "data/arg_presi_gral2003.csv")
## # A tibble: 62,323 x 26
##   codprov depto coddepto circuito mesa electores `0001` `0003` `00
##   <chr>    <chr>  <chr>    <chr>    <chr>    <dbl>    <dbl>    <dbl>    <dbl>
## 1 01       Balv~ 011     0102    0981      416      0       4
## 2 01       Balv~ 011     0102    0982      415      1       2
## 3 01       Balv~ 011     0102    0983      415      0       3
## 4 01       Balv~ 011     0102    0984      415      0       1
## 5 01       Balv~ 011     0102    0985      415      0       3
## 6 01       Balv~ 011     0102    0986      415      1       5
## 7 01       Balv~ 011     0102    0987      415      0       2
## 8 01       Balv~ 011     0102    0988      415      2       0
## 9 01       Balv~ 011     0102    0989      415      1       1
## 10 01      Balv~ 011     0102    0990      415      1       1
## # ... with 62,313 more rows, and 16 more variables: `0022` <dbl>, `00
## #   `0037` <dbl>, `0050` <dbl>, `0051` <dbl>, `0053` <dbl>, `0131` <
## #   `0132` <dbl>, `0133` <dbl>, `0134` <dbl>, `0135` <dbl>, `0136` <
## #   `0137` <dbl>, `0138` <dbl>, blancos <dbl>, nulos <dbl>
```

# Importar : datos tabulares



Leer datos tabulares - Estas funciones comparten estos argumentos:

```
read_*(file, col_names = TRUE, col_types = NULL, locale = default_locale(), na = c("", "NA"),
quoted_na = TRUE, comment = "", trim_ws = TRUE, skip = 0, n_max = Inf, guess_max = min(1000,
n_max), progress = interactive())
```

a,b,c  
1,2,3  
4,5,NA

A	B	C
1	2	3
4	5	NA

**Archivo separado por comas**  
**read\_csv("archivo.csv")**

Para generar archivo .csv ejecuta:  
`write_file(x = "a,b,c\n1,2,3\n4,5,NA", path = "archivo.csv")`

a;b;c  
1;2;3  
4;5;NA

A	B	C
1	2	3
4	5	NA

**Archivo separado por punto y coma**  
**read\_csv2("archivo2.csv")**

`write_file(x = "a;b;c\n1;2;3\n4;5;NA", path = "archivo2.csv")`

a|b|c  
1|2|3  
4|5|NA

A	B	C
1	2	3
4	5	NA

**Archivo con cualquier separador**  
**read\_delim("archivo.txt", delim = "|")**

`write_file(x = "a|b|c\n1|2|3\n4|5|NA", path = "archivo.txt")`

a b c  
1 2 3  
4 5 NA

A	B	C
1	2	3
4	5	NA

**Archivos de ancho fijo**  
**read\_fwf("archivo.fwf", col\_positions = c(1, 3, 5))**

`write_file(x = "a b c\n1 2 3\n4 5 NA", path = "archivo.fwf")`

**Archivo separado por tabulaciones**  
**read\_tsv("archivo.tsv")** también con **read\_table()**.  
`write_file(x = "a\tb\tc\n1\t2\t3\n4\t5\tNA", path = "archivo.tsv")`



# Importar .xls y .xlsx

```
library(readxl)

read_xlsx(path = "data/arg_presi_gral2003.xlsx")
## # A tibble: 62,323 x 26
##   codprov depto coddepto circuito mesa electores `1` `3` `5`
##   <dbl> <chr>   <dbl>    <dbl> <dbl>    <dbl> <dbl> <dbl> <dbl>
## 1      1 Balv~     11     102    981     416     0     4     1
## 2      1 Balv~     11     102    982     415     1     2     2
## 3      1 Balv~     11     102    983     415     0     3     0
## 4      1 Balv~     11     102    984     415     0     1     0
## 5      1 Balv~     11     102    985     415     0     3     1
## 6      1 Balv~     11     102    986     415     1     5     1
## 7      1 Balv~     11     102    987     415     0     2     2
## 8      1 Balv~     11     102    988     415     2     0     1
## 9      1 Balv~     11     102    989     415     1     1     1
## 10     1 Balv~     11     102    990     415     1     1     2
## # ... with 62,313 more rows, and 15 more variables: `30` <dbl>, `37` 
## # `50` <dbl>, `51` <dbl>, `53` <dbl>, `131` <dbl>, `132` <dbl>, `1` 
## # `134` <dbl>, `135` <dbl>, `136` <dbl>, `137` <dbl>, `138` <dbl>,
## # blancos <dbl>, nulos <dbl>
```

# Importar (otros)



Posibilita leer y escribir archivos en formatos utilizados por otros programas estadísticos:

- [SAS](#)
- [SPSS](#)
- [STATA](#)

{haven}

# Importar (otros)



Lupu, Noam and Susan C. Stokes. 2009. "The Social Bases of Political Parties in Argentina, 1912-2003." Latin American Research Review 44 (1): 58-87. Fuente: [www.noamlupu.com/data.html](http://www.noamlupu.com/data.html)

```
library(haven)
```

```
read_dta(file = "data/argentina_ecological_data.dta")
## # A tibble: 16,683 x 17
##   recid province department dcode year month dip pres total uo
##   <dbl> <chr>     <chr>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 6046 Buenos ~ 25 de Mayo 1000 1914 NA     1     0 1446 72
## 2 6227 Buenos ~ 25 de Mayo 1000 1916 NA     1     0 2725 135
## 3 6539 Buenos ~ 25 de Mayo 1000 1918 NA     1     0 2698 186
## 4 6798 Buenos ~ 25 de Mayo 1000 1920 NA     1     0 2664 162
## 5 136  Buenos ~ 25 de Mayo 1000 1922 NA     0     1 2097 158
## 6 7076 Buenos ~ 25 de Mayo 1000 1926 NA     1     0 1863 139
## 7 472  Buenos ~ 25 de Mayo 1000 1928 NA     0     1 5553 294
## 8 7448 Buenos ~ 25 de Mayo 1000 1930 NA     1     0 5239 239
## 9 929  Buenos ~ 25 de Mayo 1000 1946 NA     0     1 8808 269
## 10 14900 Buenos ~ 25 de Mayo 1000 1948 3      1     0 6350 134
## # ... with 16,673 more rows, and 6 more variables: soc <dbl>, con <dbl>
## #   enp <dbl>, urb <dbl>, lit <dbl>, pop <dbl>
```

# Importar - Directo de la URL!



Lupu, Noam and Susan C. Stokes. 2009. "The Social Bases of Political Parties in Argentina, 1912-2003." Latin American Research Review 44 (1): 58-87. Fuente: [www.noamlupu.com/data.html](http://www.noamlupu.com/data.html)

```
library(haven)
```

```
read_dta(file = "https://www.noamlupu.com/argentina_ecological_data.dta")
## # A tibble: 16,683 x 17
##   recid province department dcode year month dip pres total uo
##   <dbl> <chr>     <chr>      <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 6046 Buenos ~ 25 de Mayo  1000  1914    NA     1     0  1446  7.
## 2 6227 Buenos ~ 25 de Mayo  1000  1916    NA     1     0  2725 13.
## 3 6539 Buenos ~ 25 de Mayo  1000  1918    NA     1     0  2698 18.
## 4 6798 Buenos ~ 25 de Mayo  1000  1920    NA     1     0  2664 16.
## 5 136  Buenos ~ 25 de Mayo  1000  1922    NA     0     1  2097 15.
## 6 7076 Buenos ~ 25 de Mayo  1000  1926    NA     1     0  1863 13.
## 7 472  Buenos ~ 25 de Mayo  1000  1928    NA     0     1  5553 29.
## 8 7448 Buenos ~ 25 de Mayo  1000  1930    NA     1     0  5239 23.
## 9 929  Buenos ~ 25 de Mayo  1000  1946    NA     0     1  8808 26.
## 10 14900 Buenos ~ 25 de Mayo 1000  1948     3     1     0  6350 13.
## # ... with 16,673 more rows, and 6 more variables: soc <dbl>, con <dbl>
## #   enp <dbl>, urb <dbl>, lit <dbl>, pop <dbl>
```

# Importar: planillas de *Google*



```
library(googlesheets4)

gs4_deauth() # Eliminar auth interactivo (link publico)

read_sheet("https://docs.google.com/spreadsheets/d/1J84PviVxVMsCDzup0cD(
# A tibble: 4 x 9
#   `Apellido y Nom~ Profesión  Edad `Lugar de Nacim~ `Fecha de Nacimien-
#   <chr>           <chr>     <dbl> <chr>           <dttm>
#1 Fasola, Juan Pe~ Amigo       37 Hurlingam    1983-09-15 00:00:00
#2 Soze, Kayser      Narcotra~  58 Los Angeles  1962-02-28 00:00:00
#3 Bigpear, John S~ General     75 Lobos        1945-10-17 00:00:00
#4 Tower, JC         Sociólogo   74 Bahía Blanca 1946-09-24 00:00:00
# ... with 1 more variable: `Software Utilizado` <chr>
```

🔗 {googlesheets4}

# Importar: Otros

- `vroom::vroom()`

<https://www.tidyverse.org/blog/2019/05/vroom-1-0-0/>

- `data.table::fread()`

<https://rdatatable.gitlab.io/data.table/>

- `DB` (<https://db.rstudio.com/>)

# Importar con {datapasta}



## Desde el portapapeles (*clipboard*)

Brisbane area  
Partly cloudy. Light winds.

3:30 pm, UV Index predicted to reach 11 [Extreme]

Brisbane area  
Partly cloudy. Medium (50%) chance of showers, most likely in the late morning and afternoon. Light winds becoming easterly 15 to 20 km/h in the late afternoon then becoming light in the evening.

3:30 pm, UV Index predicted to reach 11 [Extreme]

Brisbane area  
Partly cloudy. Light winds.

7 day Town Forecasts			
	Location	Min	Max
	<a href="#">Brisbane</a>	23	30
	<a href="#">Brisbane Airport</a>	22	29
	<a href="#">Beaudesert</a>	21	30
	<a href="#">Chermside</a>	22	30
	<a href="#">Gatton</a>	21	30
	<a href="#">Ipswich</a>	21	31
	<a href="#">Logan Central</a>	22	30
	<a href="#">Manly</a>	23	28
	<a href="#">Mount Gravatt</a>	22	29
	<a href="#">Oxley</a>	22	31
	<a href="#">Redcliffe</a>	23	28



# {datapasta} *live coding*

🔗 <https://docs.google.com/spreadsheets/>

# Ordenar



una versión de data frame que facilita el trabajo con el tidyverse - R4DS - ñ

- no modifica clase de las variables
- genera más advertencias por problemas ([warnings](#))
- método de impresión ([print](#)) más prolífico y comentado

# Ordenar



-> live coding: limpiemos nombres de variables de nuestro Google Sheet

# Ordenar: *Tidy Data*



“Todas las familias felices se parecen unas a otras, pero cada familia infeliz lo es a su manera.” — León Tolstoy

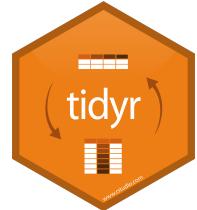
“Todos los set de datos ordenados se parecen unos a otros, pero cada set de datos desordenado lo es a su manera” — Hadley Wickham

## Referencias

🔗 R4DS en español

🔗 Wickham, H; *Tidy Data*, Journal of Statistical Software

# Datos Ordenados



país	anio	casos	poblacion
Afganistán	1999	745	19987071
Afganistán	2000	2666	20595360
Brasil	1999	37737	172006362
Brasil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

variables

país	anio	casos	poblacion
Afganistán	1999	745	19987071
Afganistán	2000	2666	20595360
Brasil	1999	37737	172006362
Brasil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

observaciones

país	anio	casos	poblacion
Afganistán	1999	745	19987071
Afganistán	2000	2666	20595360
Brasil	1999	37737	172006362
Brasil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

valores

1. Cada variable debe tener su propia columna.
2. Cada observación debe tener su propia fila.
3. Cada valor debe tener su propia celda.

# Datos Ordenados: *pivotear*



Entre los distintos verbos, se destacan:

- `pivot_longer()`: reduce cantidad de columnas y aumenta las filas
- `pivot_wider()`: reduce cantidad de filas y aumenta columnas

-> *live coding* con datos de `{polAr}`

- Más verbos:

`complete / fill / replace_na / drop_na,`

`nest / unnest,`

`unite /separate / extract`

🔗`{tidyr}`

# Transformar



## Una caja de herramientas

Verbos principales de `{dplyr}` para manipular la *data*

- `filter()`: reduce la cantidad de filas (observaciones)
- `select()`: reduce la cantidad de columnas (variables)
- `mutate()`: crea o modifica variables
- `arrange()`: ordena (sort)
- `group_by()`: agrupa observaciones
- `summarize()`: reduce múltiples observaciones a un valor

-> *live coding* con datos de `{polAr}`

# EJERCICIO

1. crear un proyecto
2. descargar datos según indicaciones
3. procesar la información en un reporte que responda a la siguiente pregunta:

*Cuántos puntos porcentuales de diferencia hubo entre el primero y segundo en la elección a presidente 2003 en \_su provincia?\_*