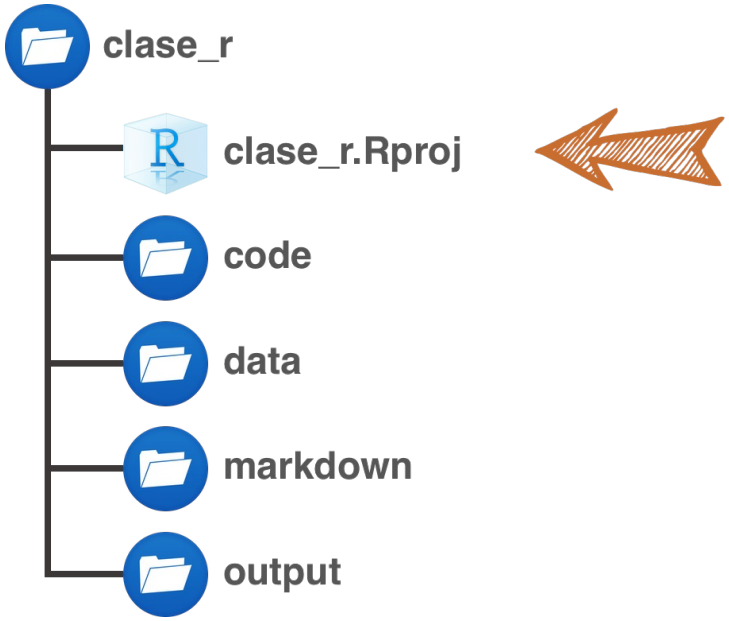
Tidyverse

José Luis Texcalac Sangrador

Procesamiento y visualización de datos espaciales en R

Trabajar siempre con el proyecto de la clase



Clic a este archivo para abrir Rstudio y

- Generar script de clase
- Generar R Quarto para su laboratorio
- Generar R Markdown

Si la malla de datos está dentro de la carpeta data

- R Quarto (.qmd) del laboratorio dentro de la carpeta markdown.

```
read_csv("../data/malla.csv")
```

- Script de R (.r) de clase dentro de la carpeta code.

```
read_csv("./data/malla.csv")
```

Encabezado de script vs YAML

```
# ============================================================================ #
# Procesamiento y visualización de datos espaciales en R
# Sesión - 06
# Autor: José Luis Texcalac Sanrador
# Fecha: 07 de octubre 2021
# ============================================================================ #
```

Encabezado de script de R (.r)

```
---
title: "Introducción a Tidyverse"
subtitle: "Primer vistazo a tidyverse"
author: "Tex"
title-block-banner: true
date: today
format:
  html:
    theme: cosmo
editor: source
---
```

YAML R Quarto (.qmd)

Sugerencias para su scripts

```
# ========================================================================== #
# Procesamiento y visualización de datos espaciales en R
# Sesión - 06
# Autor: José Luis Texcalac Sanrador
# Fecha: 07 de octubre 2021
# ========================================================================== #
library(tidyverse)
library(janitor)
library(foreign)
library(haven)
### Inicio procesamiento de datos
malla <- read_csv("./tabla.csv") %>% clean_names() %>% print()
```

línea en blanco aquí

Respetar patrón en nombre de archivos

- **S06_Tex**
- **L06_Tex**

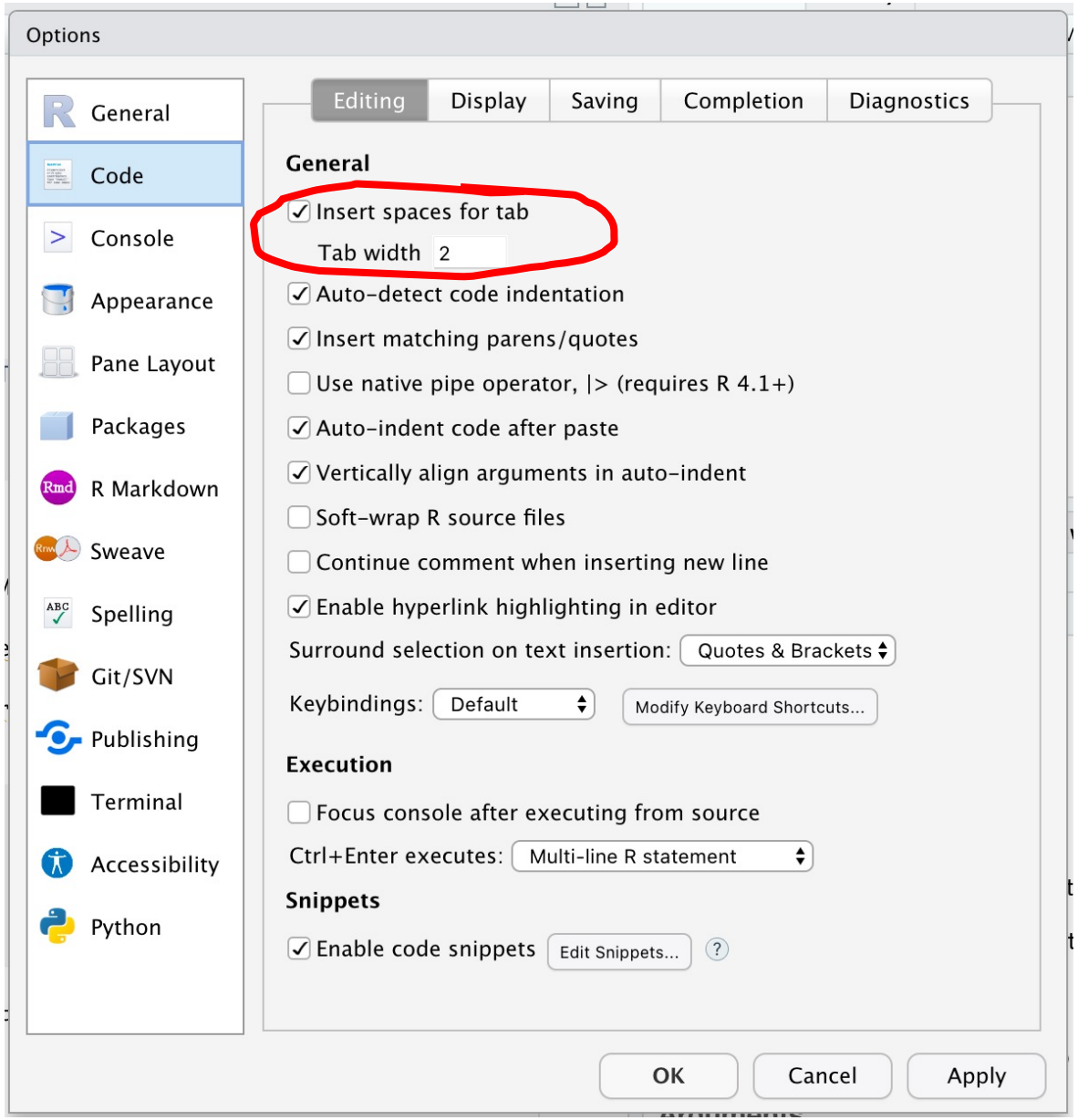
```
# ========================================================================== #
# Procesamiento y visualización de datos espaciales en R
# Sesión - 06
# Autor: José Luis Texcalac Sanrador
# Fecha: 07 de octubre 2021
# ========================================================================== #

library(tidyverse)
library(janitor)
library(foreign)
library(haven)

### Inicio procesamiento de datos
malla <-
  read_csv("./tabla.csv") %>%
  clean_names() %>%
  print()
```

Nombre completo

Cargar todas las librerías a utilizar al inicio de su script

Use tabulaciones

INSP/ESPM
Escuela de Salud Pública de México



INSP/ESPM
Escuela de Salud Pública de México

YouTube MX

tabs vs space

ACCEDER

Principal

Explorar

Suscripcion...

Biblioteca

Historial

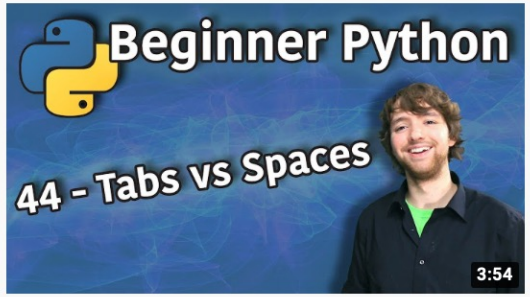
Elon Musk en Tabs vs Spaces, Vim vs Emacs y Light vs Dark Mode

46,443 vistas • hace 1 año

Hack Club

Mira el AMA completo con Elon: https://www.youtube.com/watch?v=riru9OzScwk Vea más clips de AMA de Elon: https://www.youtube ...

Subtítulos Inglés



Beginner Python Tutorial 44 - Tabs vs Spaces

7,062 vistas • hace 1 año

Caleb Curry

~~~~~~~~~~~~~~~~ CONNECT ~~~~~~~~~~~~~~~~ Newsletter - https://calcur.tech/newsletter Instagram ...

Tabs vs Spaces on Silicon Valley

2,267 vistas • hace 5 años

Corey Donohoe

Kanye inspired **tabs vs spaces** scene in silicon valley on hbo.

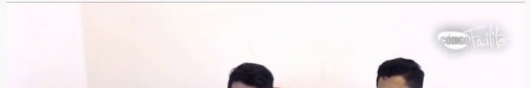
Douglas Crockford Tabs vs Spaces

1,174 vistas • hace 1 año

downthecrop

De: The Post JavaScript Apocalypse - Douglas Crockford 2017 https://www.youtube.com/watch?v=NPB34IDZj3E Crockford ofrece ...

Subtítulos Inglés



Pregúntale a un Desarrollador #2 - Tabs vs Spaces
~~~~~~~~~~~~~~~~

Renombrar columnas: `rename()`

Sin guardar el resultado

```
pm10 %>%
  rename(est_dif = dif,
         est_cr = cruz_roja)
```

Primero el nuevo nombre

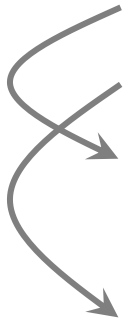
Después el viejo nombre

```
# A tibble: 8,764 × 5
   fecha      hora  est_dif est_cr nativitas
   <date>     <fct>   <dbl>  <dbl>     <dbl>
 1 2014-01-01 0          NA     NA        NA
 2 2014-01-01 1          68     57       224
 3 2014-01-01 2          74     52       118
 4 2014-01-01 3          58     45        59
 5 2014-01-01 4          50     51        40
 6 2014-01-01 5          48     39        43
 7 2014-01-01 6          49     25        39
 8 2014-01-01 7          41     42        43
 9 2014-01-01 8          49     48        46
10 2014-01-01 9          60     50        47
# … with 8,754 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Identificar duplicados {janitor}: `get_dupes()`

```
malla
```

Site	2013	2014	2015
CAM	51.0	42.8	39.9
FAC	48.3	39.0	36.6
CAM	51.0	42.8	39.9
IZT	44.6	39.3	35.0
FAC	48.3	39.0	36.6



```
malla %>% get_dupes()
```

Site	2013	2014	2015	dupe_count
CAM	51.0	42.8	39.9	2
FAC	48.3	39.0	36.6	2

Eliminar duplicados {dplyr}: `distinct()`

malla

Site	2013	2014	2015	
CAM	51.0	42.8	39.9	
FAC	48.3	39.0	36.6	
CAM	51.0	42.8	39.9	✗
IZT	44.6	39.3	35.0	
FAC	48.3	39.0	36.6	✗

Elimina filas duplicadas

malla %>% distinct()

Site	2013	2014	2015
CAM	51.0	42.8	39.9
FAC	48.3	39.0	36.6
IZT	44.6	39.3	35.0

Identificar duplicados {janitor}: `get_dupes()`

```
nombres
```

nombre	edad	peso	pelo
Luis	49	81.3	Rizado
Juan	58	79.4	Lacio
Pedro	51	80.7	Lacio
Luis	49	81.3	
Pedro	51	80.7	

```
nombres %>% get_dupes(nombre, edad, peso)
```

nombre	edad	peso	dupe_count	pelo
Luis	49	81.3	2	Rizado
Luis	49	81.3	2	""
Pedro	51	80.7	2	Lacio
Pedro	51	80.7	2	""

Eliminar duplicados en columnas seleccionadas

nombre	edad	peso	pelo	
Luis	49	81.3	Rizado	
Juan	58	79.4	Lacio	
Pedro	51	80.7	Lacio	
Luis	49	81.3		✗
Pedro	51	80.7		✗

```
malla %>%
  distinct(nombre, edad, peso, .keep_all = TRUE)
```

nombre	edad	peso	pelo
Luis	49	81.3	Rizado
Juan	58	79.4	Lacio
Pedro	51	80.7	Lacio

- El resultado es una tabla sin duplicados en columnas seleccionadas
- El argumento .keep_all = TRUE permite mantener todas las columnas

Contar repeticiones de valores: count()

malla

Site	2013	2014	2015
CAM	51.0	42.8	39.9
FAC	48.3	39.0	36.6
CAM	51.0	42.8	39.9
IZT	44.6	39.3	35.0
FAC	48.3	39.0	36.6

```
malla %>%
  count(site, name = "tot_rep")
```

```
# A tibble: 3 × 2
  site  tot_rep
  <chr>   <int>
1 CAM         2
2 FAC         2
3 IZT         1
```

- Alternativa para identificar y contabilizar los duplicados
- Útil con pocas variables

```
malla %>%
  count(site, 2013, 2014, 2015, name = "tot_rep")
```

Exportar malla de datos

.CSV

```
write_excel_csv(malla, "ruta y nombre de archivo.csv")
```

```
write_excel_csv(malla, "./data/malla.csv")
```

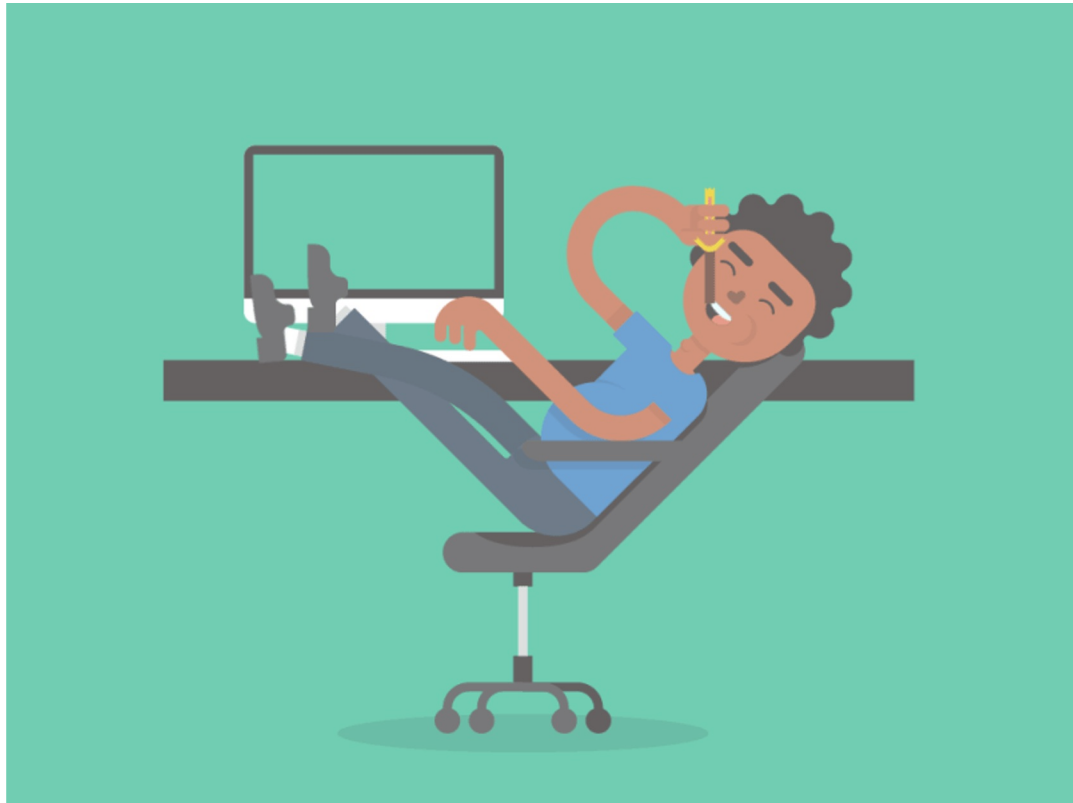
.RDS

```
write_rds(malla, "ruta y nombre de archivo.rds")
```

```
write_rds(malla, "./data/malla.rds")
```

Su turno. ..

- Abra su proyecto y genere su script de clase
- Exporte el objeto ozono
- Guarde el archivo como ozono_wide.rds
- Reinicie su sesión de R
  - Session\Restart R
- Continúe en su script
- Importe el archivo ozono_wide.rds y nombre al objeto como ozono_w



Estructura de la malla de datos

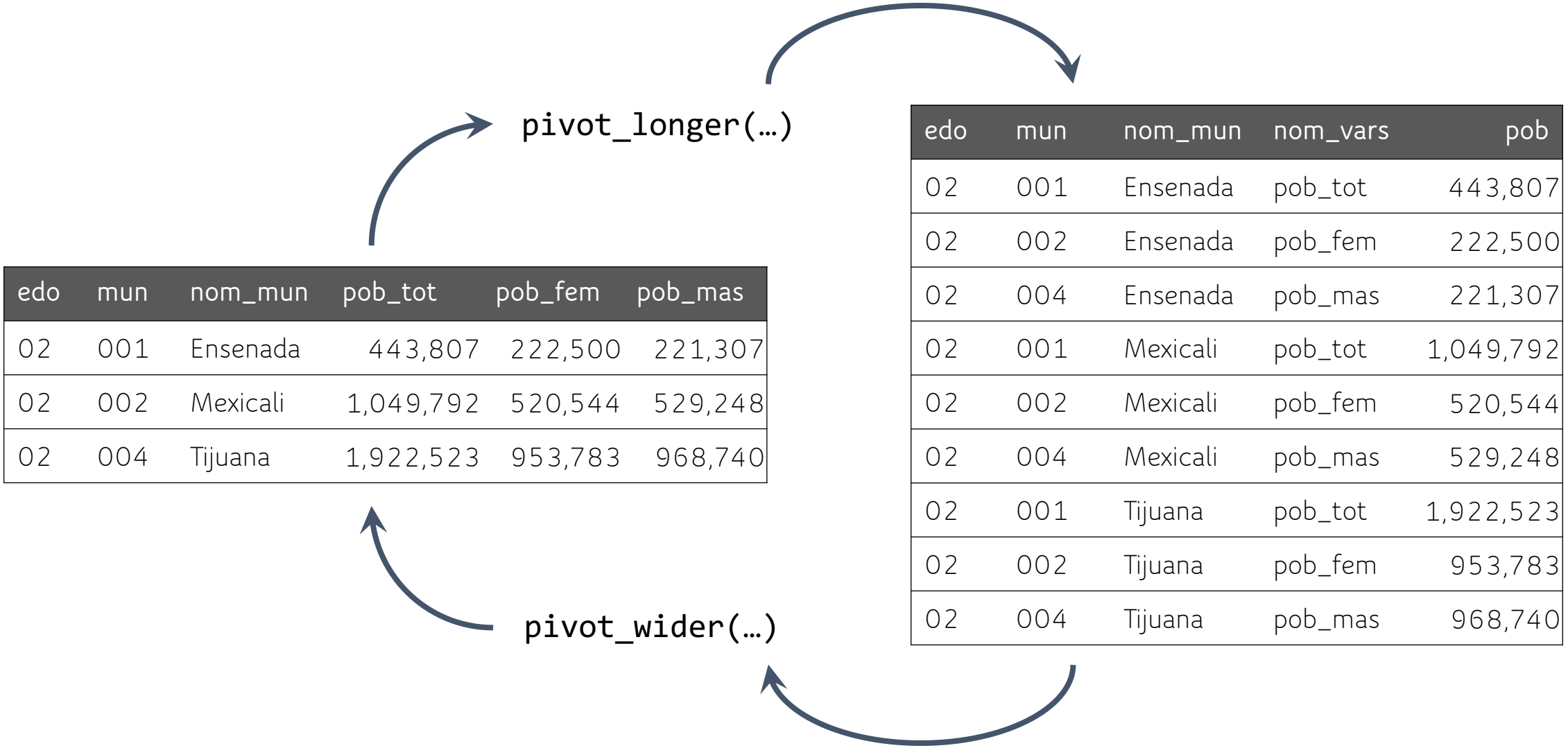
edo	mun	nom_mun	pob_tot	pob_fem	pob_mas
02	001	Ensenada	443,807	222,500	221,307
02	002	Mexicali	1,049,792	520,544	529,248
02	004	Tijuana	1,922,523	953,783	968,740

wide

long

edo	mun	nom_mun	nom_vars	pob
02	001	Ensenada	pob_tot	443,807
02	002	Ensenada	pob_fem	222,500
02	004	Ensenada	pob_mas	221,307
02	001	Mexicali	pob_tot	1,049,792
02	002	Mexicali	pob_fem	520,544
02	004	Mexicali	pob_mas	529,248
02	001	Tijuana	pob_tot	1,922,523
02	002	Tijuana	pob_fem	953,783
02	004	Tijuana	pob_mas	968,740

Transformar malla de datos a wide y long

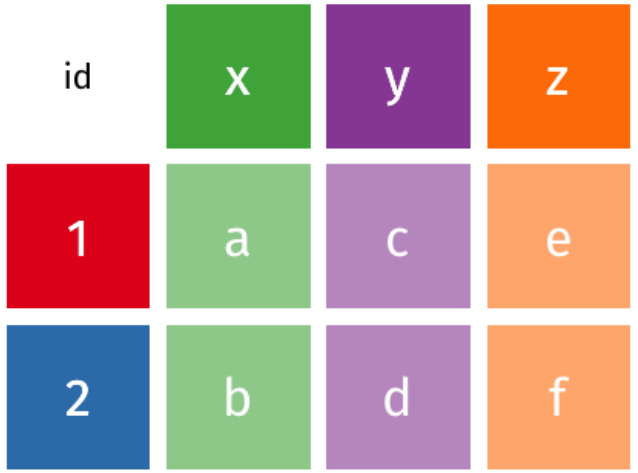


edo	mun	nom_mun	pob_tot	pob_fem	pob_mas
02	001	Ensenada	443,807	222,500	221,307
02	002	Mexicali	1,049,792	520,544	529,248
02	004	Tijuana	1,922,523	953,783	968,740

edo	mun	nom_mun	nom_vars	pob
02	001	Ensenada	pob_tot	443,807
02	002	Ensenada	pob_fem	222,500
02	004	Ensenada	pob_mas	221,307
02	001	Mexicali	pob_tot	1,049,792
02	002	Mexicali	pob_fem	520,544
02	004	Mexicali	pob_mas	529,248
02	001	Tijuana	pob_tot	1,922,523
02	002	Tijuana	pob_fem	953,783
02	004	Tijuana	pob_mas	968,740

pivot_longer()

```
malla %>%
   pivot_longer(cols = ...,
                names_to = ...,
                values_to = ...,
                names_prefix = ...)
```



https://tidyr.tidyverse.org/reference/pivot_longer.html

pivot_longer(...)

edo	mun	nom_mun	pob_tot	pob_fem	pob_mas
02	001	Ensenada	443,807	222,500	221,307
02	002	Mexicali	1,049,792	520,544	529,248
02	004	Tijuana	1,922,523	953,783	968,740

edo	mun	nom_mun	nom_vars	pob
02	001	Ensenada	pob_tot	443,807
02	002	Ensenada	pob_fem	222,500
02	004	Ensenada	pob_mas	221,307
02	001	Mexicali	pob_tot	1,049,792
02	002	Mexicali	pob_fem	520,544
02	004	Mexicali	pob_mas	529,248
02	001	Tijuana	pob_tot	1,922,523
02	002	Tijuana	pob_fem	953,783
02	004	Tijuana	pob_mas	968,740

https://tidyr.tidyverse.org/reference/pivot_longer.html

- ✓ cols = c(pob_tot, pob_mas)
- ✓ cols = pob_tot:pob_mas

```
malla %>%
    pivot_longer(cols = starts_with("pob_"),
        names_to = "nom_vars",
        names_prefix = "pob_",
        values_to = "pob")
```

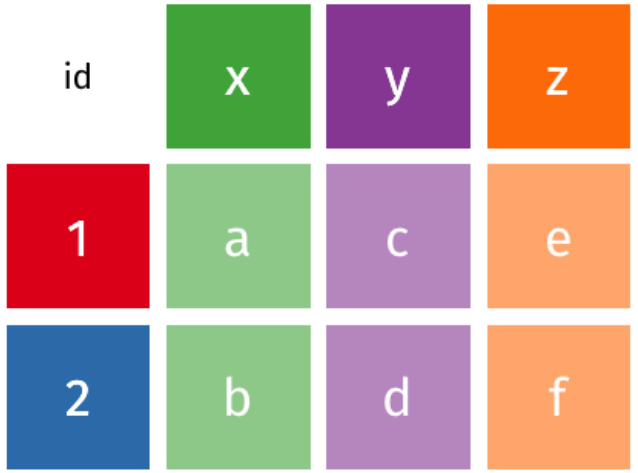
edo	mun	nom_mun	pob_tot	pob_fem	pob_mas
02	001	Ensenada	443807	222500	221307
02	002	Mexicali	1049792	520544	529248
02	004	Tijuana	1922523	953783	968740

Si no se especifica no se borra el prefijo

edo	mun	nom_mun	nom_vars	pob
02	001	Ensenada	tot	443,807
02	002	Ensenada	fem	222,500
02	004	Ensenada	mas	221,307
02	001	Mexicali	tot	1,049,792
02	002	Mexicali	fem	520,544
02	004	Mexicali	mas	529,248
02	001	Tijuana	tot	1,922,523
02	002	Tijuana	fem	953,783
02	004	Tijuana	mas	968,740

pivot_wider()

```
malla %>%
  pivot_wider(names_from = ...,
              names_prefix = ...,
              values_to = ...)
```



https://tidyr.tidyverse.org/reference/pivot_wider.html

pivot_wider(...)

edo	mun	nom_mun	nom_vars	pob
02	001	Ensenada	pob_tot	443,807
02	002	Ensenada	pob_fem	222,500
02	004	Ensenada	pob_mas	221,307
02	001	Mexicali	pob_tot	1,049,792
02	002	Mexicali	pob_fem	520,544
02	004	Mexicali	pob_mas	529,248
02	001	Tijuana	pob_tot	1,922,523
02	002	Tijuana	pob_fem	953,783
02	004	Tijuana	pob_mas	968,740

Transformar de long a wide

edo	mun	nom_mun	pob_tot	pob_fem	pob_mas
02	001	Ensenada	443,807	222,500	221,307
02	002	Mexicali	1,049,792	520,544	529,248
02	004	Tijuana	1,922,523	953,783	968,740

```
malla %>%
pivot_wider(names_from = "nom_vars",
            names_prefix = "pob_",
            values_from = "pob")
```

edo	mun	nom_mun	nom_vars	pob
02	001	Ensenada	tot	443,807
02	002	Ensenada	fem	222,500
02	004	Ensenada	mas	221,307
02	001	Mexicali	tot	1,049,792
02	002	Mexicali	fem	520,544
02	004	Mexicali	mas	529,248
02	001	Tijuana	tot	1,922,523
02	002	Tijuana	fem	953,783
02	004	Tijuana	mas	968,740

edo	mun	nom_mun	pob_tot	pob_fem	pob_mas
02	001	Ensenada	443,807	222,500	221,307
02	002	Mexicali	1,049,792	520,544	529,248
02	004	Tijuana	1,922,523	953,783	968,740

Su turno...

- Trabaje con la malla ozono_w
  - Transforme la malla de wide a long
  - Guarde el resultado como ozono_l

Agrupar datos: `group_by()`

```
malla %>% group_by(...) %>% summarise(...)
```

Malla de datos a agrupar

Argumentos de agrupación

Argumentos de resumen

```
malla %>%
  group_by(var1)
  summarise(tot_var2 = sum(var2, na.rm = TRUE))
```

Agrupar datos... group_by( )

`malla` `%>%` `group_by(entidad)` `%>%` `summarise(...)`

entidad	municipio	población
Aguascalientes	Asientos	51,536
Aguascalientes	Calvillo	58,250
Baja California	Ensenada	443,807
Baja California	Mexicali	1,049,792
Ciudad de México	Azcapotzalco	432,205
Ciudad de México	Benito Juárez	434,153

Aguascalientes	Asientos	51,536
Aguascalientes	Calvillo	58,250

Baja California	Ensenada	443,807
Baja California	Mexicali	1,049,792

Ciudad de México	Azcapotzalco	432,205
Ciudad de México	Benito Juárez	434,153

entidad	mean	sum	n
Aguascalientes	54,893.0	109,786	2
Baja California	433,179.0	1,493,599	2
Ciudad de México	740,998.5	866,358	2

```
malla %>%
   group_by(entidad) %>%
   summarise(mean = mean(población),
             sum = sum(población),
             n = n())
```

Su turno...

- Trabaje con la malla ozono_l
- Agrupe por fecha y calcule el promedio y desviación estándar.
- Guarde el resultado como o3_dia

Tidy Animated Verbs

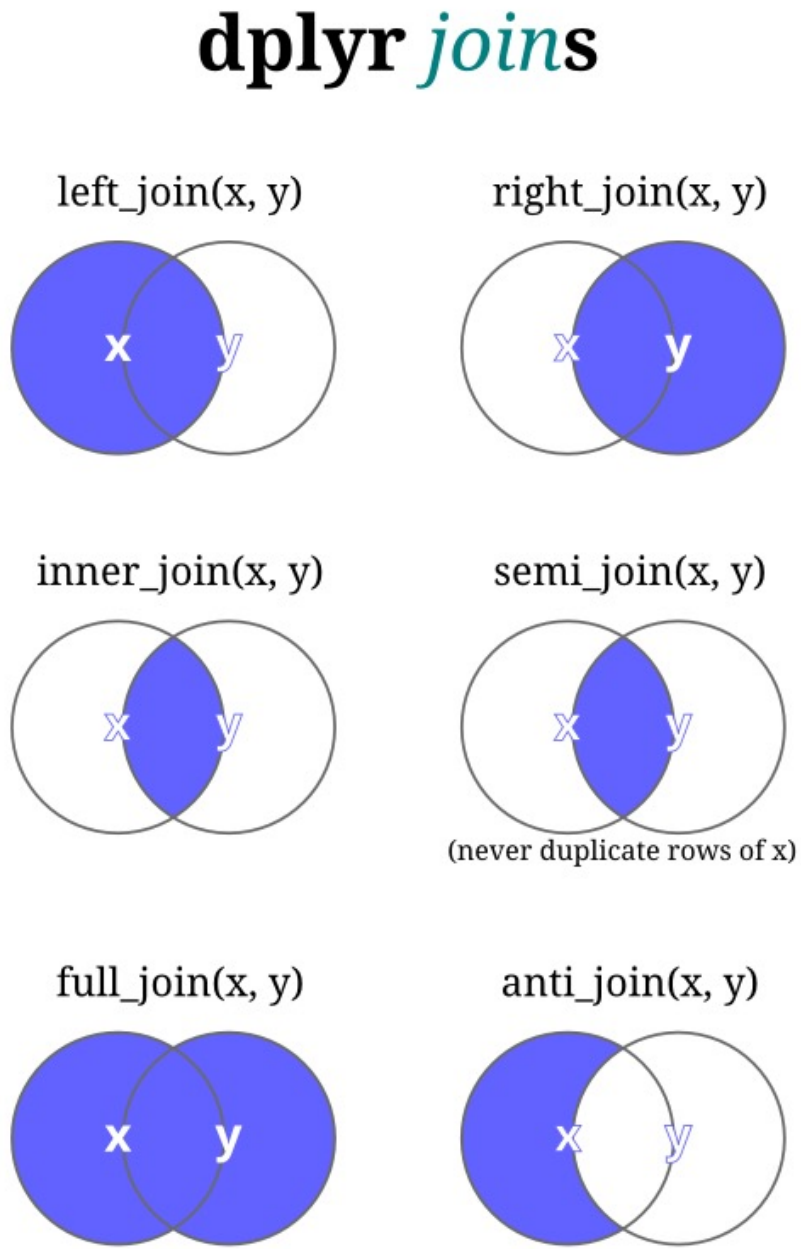
Garrick Aden-Buie – @grrrck – garrickadenbuie.com. Set operations contributed by Tyler Grant Smith.

- **Mutating Joins** — `inner_join()`, `left_join()`, `right_join()`, `full_join()`
- **Filtering Joins** — `semi_join()`, `anti_join()`
- **Set Operations** — `union()`, `union_all()`, `intersect()`, `setdiff()`
- **Tidy Data** — `spread()` and `gather()`
- Learn more about
  - Using the animations and images
  - Relational Data
  - gganimate

https://dplyr.tidyverse.org/reference/join.html

INSP/ESPM
Escuela de Salud Pública de México

Unión de mallas de datos



inner_join(tbl_1, tbl_2, by = "cvegeo")

tbl_1

cvegeo	nom_mun
17007	Cuernavaca
17020	Tepoztlán
17028	Xochitepec

tbl_2

cvegeo	cve_sun
17009	M17.02
17011	M17.02
17020	M17.02

cvegeo	nom_mun	cve_sun
17020	Tepoztlán	M17.02

inner_join(tbl_1, tbl_2, by = "cvegeo")

tbl_1

cvegeo	nom_mun
17017	Puente de Ixtla
17020	Tepoztlán
17028	Xochitepec

tbl_2

cvegeo	cve_sun
17009	M17.02
17011	M17.02
17017	P17.01
17017	C17.02

cvegeo	nom_mun	cve_sun
17017	Puente de Ixtla	P17.01
17017	Puente de Ixtla	C17.02

Argumento by

Si el nombre del identificador de ambas columnas no coincide entonces:

X Y

```
by = c("edo" = "estado")
```

En la unión se mantiene el nombre de la columna de la tabla X

```
by = c("edo" = "estado", "mun" = "mun")
```

INSP/ESPM
Escuela de Salud Pública de México

inner_join(tbl_1, tbl_2, by = c("cvegeo" = "cve_mun"))

tbl_1

cvegeo	nom_mun
17017	Puente de Ixtla
17020	Tepoztlán
17028	Xochitepec

tbl_2

cve_mun	cve_sun
17009	M17.02
17011	M17.02
17017	P17.01
17017	C17.02

cvegeo	nom_mun	cve_sun
17017	Puente de Ixtla	P17.01
17017	Puente de Ixtla	C17.02

left_join(tbl_1, tbl_2, by = "cvegeo")

tbl_1

cvegeo	nom_mun
17007	Cuernavaca
17020	Tepoztlán
17028	Xochitepec

tbl_2

cvegeo	cve_sun
17009	M17.02
17011	M17.02
17020	M17.02

cvegeo	nom_mun	cve_sun
17007	Cuernavaca	NA
17020	Tepoztlán	M17.02
17028	Xochitepec	NA

right_join(tbl_1, tbl_2, by = "cvegeo")

tbl_1

cvegeo	nom_mun
17007	Cuernavaca
17020	Tepoztlán
17028	Xochitepec

tbl_2

cvegeo	cve_sun
17009	M17.02
17011	M17.02
17020	M17.02

cvegeo	nom_mun	cve_sun
17009	NA	M17.02
17011	NA	M17.02
17020	Tepoztlán	M17.02

full_join(tbl_1, tbl_2, by = "cvegeo")

tbl_1

cvegeo	cve_mun
17007	Cuernavaca
17020	Tepoztlán
17028	Xochitepec

tbl_2

cvegeo	cve_sun
17009	M17.02
17011	M17.02
17020	M17.02

cvegeo	nom_mun	cve_sun
17007	Cuernavaca	NA
17009	NA	M17.02
17011	NA	M17.02
17020	Tepoztlán	M17.02
17028	Xochitepec	NA