R: wide, long, join

INSP/ESPM
Escuela de Salud Pública de México

José Luis Texcalac Sangrador

Procesamiento y visualización de datos espaciales en R

Exportar mallas 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")
```

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

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

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

pivot_wider(...)

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

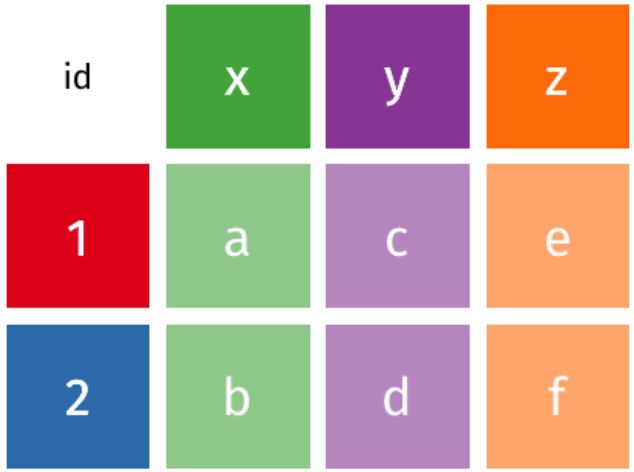
Transformar de wide a long

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

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

pivot_longer()

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



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

- ✓ cols = c(pob_tot, pob_fem, 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

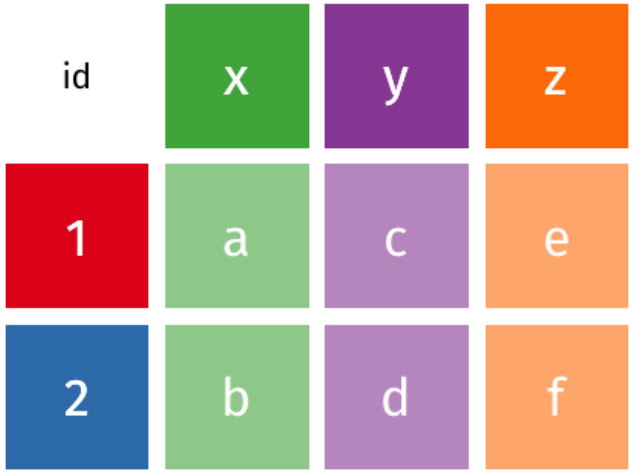
Su turno...

Importe la malla ozono, nombre al objeto como ozono

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

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...

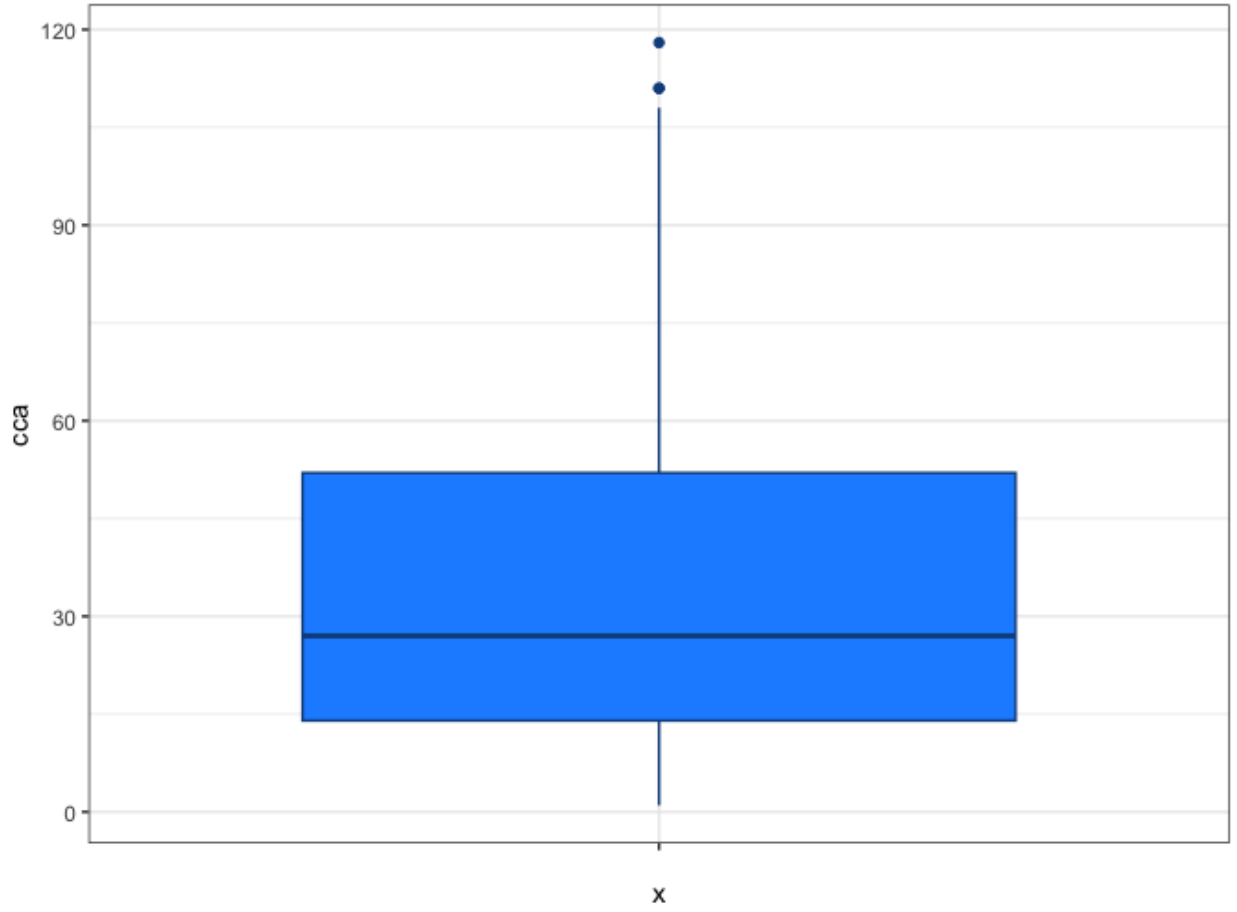
Trabaje con la malla ozono_l

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

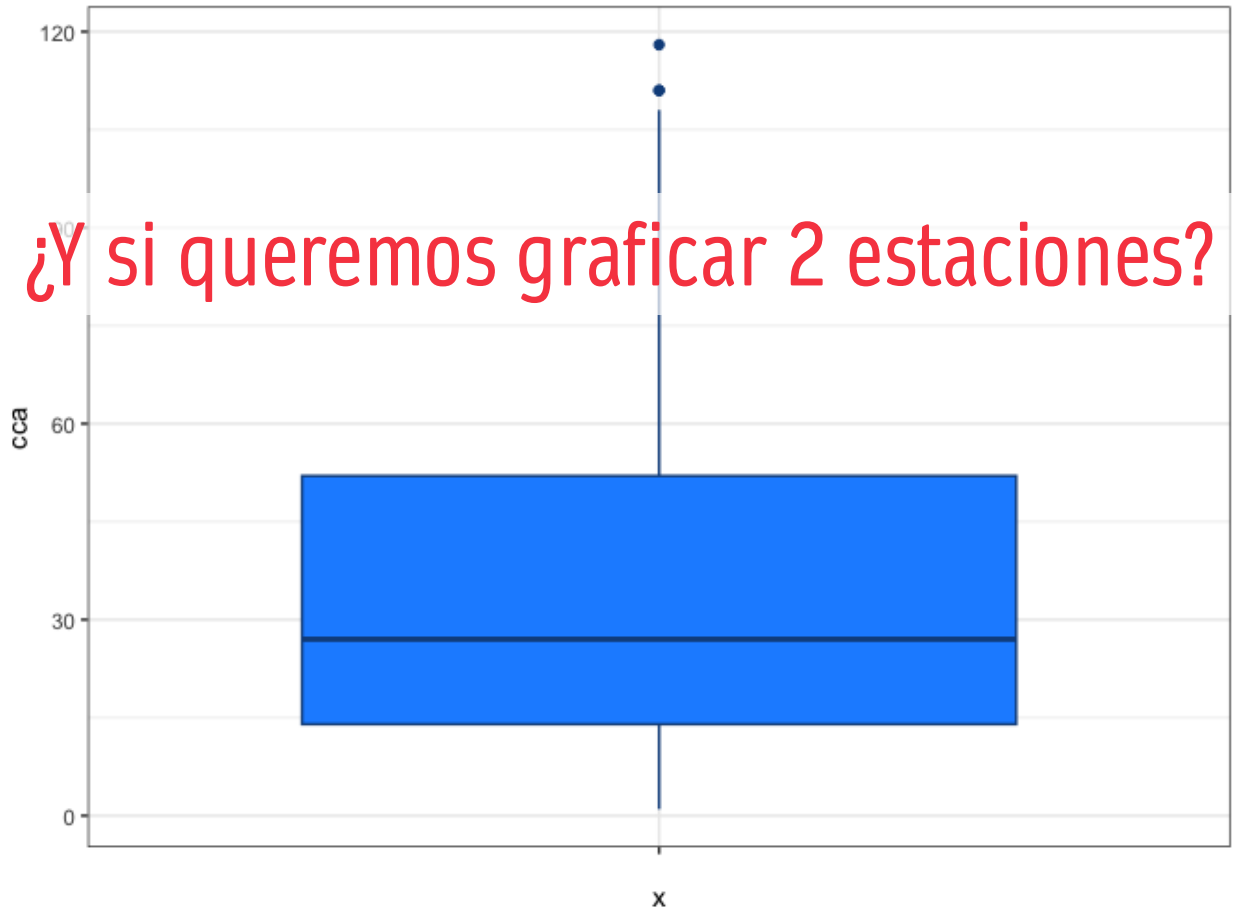
```
ggplot(data = ozono_w) +
    geom_boxplot(aes(x = "", cca), colour = "dodgerblue4", fill = "dodgerblue1") +
    theme_light()
```

Argumentos "color" y "fill"

- color: define el color de un borde.
- fill: define el color de relleno.

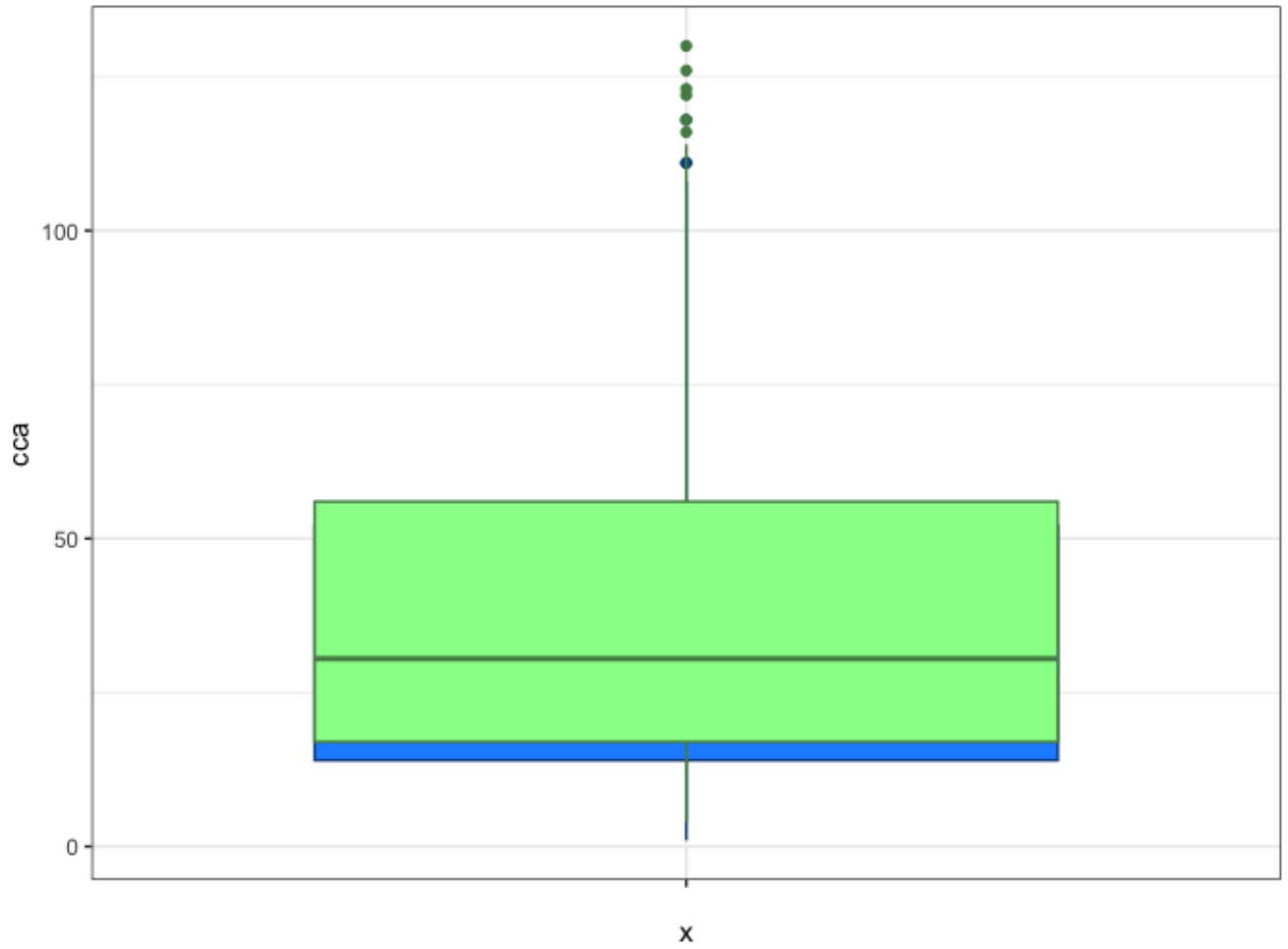


```
ggplot(data = ozono_w) +
    geom_boxplot(aes(x = "", cca), colour = "dodgerblue4", fill = "dodgerblue1") +
    theme_light()
```

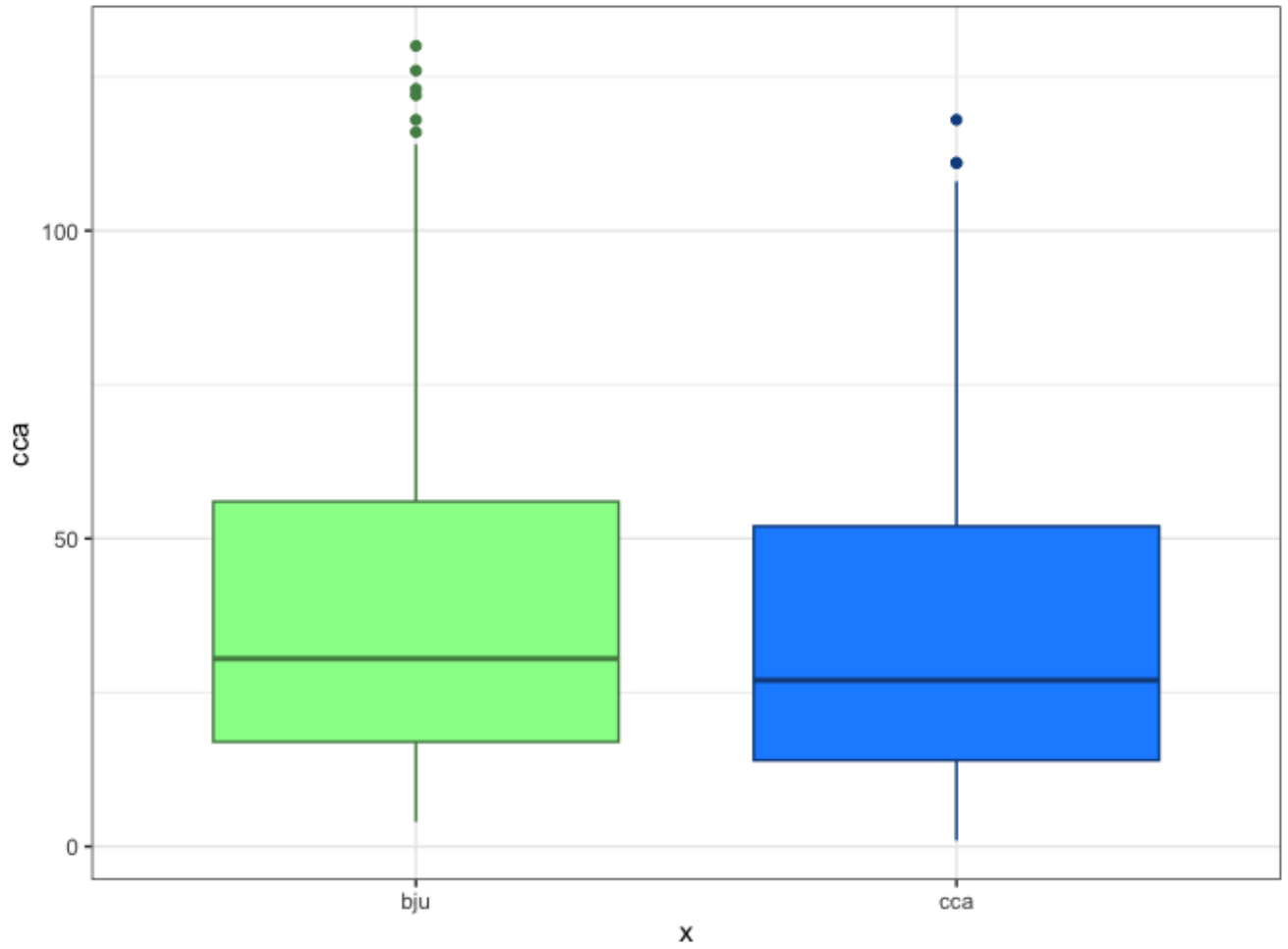


¿Y si queremos graficar 2 estaciones?

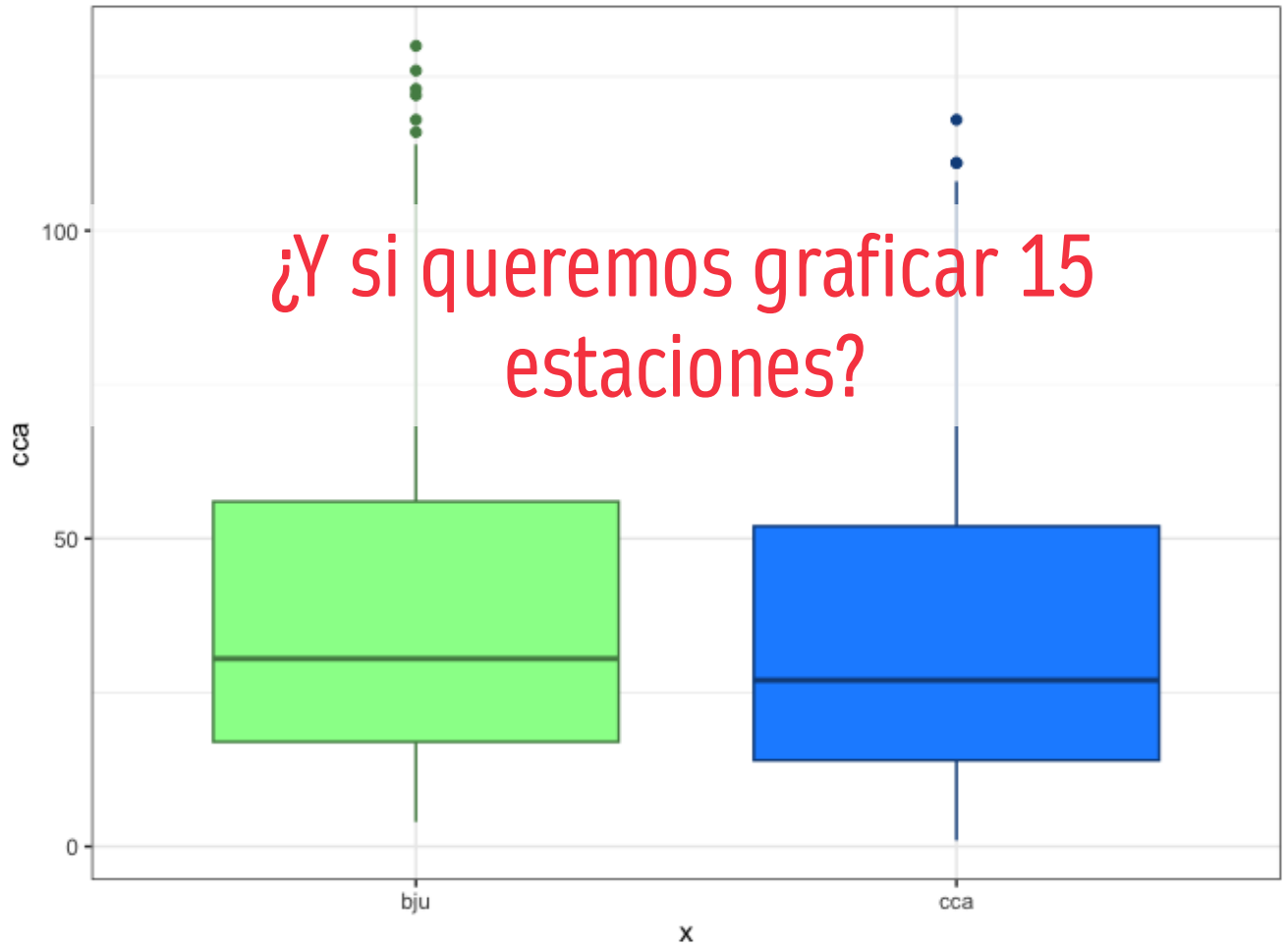
```
ggplot(data = ozono_w) +
    geom_boxplot(aes(x = "", cca), colour = "dodgerblue4", fill = "dodgerblue1") +
    geom_boxplot(aes(x = "", bju), colour = "#548B54", fill = "#98FB98") +
    theme_light()
```



```
ggplot(data = ozono_w) +
    geom_boxplot(aes(x = "cca", cca), colour = "dodgerblue4", fill = "dodgerblue1") +
    geom_boxplot(aes(x = "bju", bju), colour = "#548B54", fill = "#98FB98") +
    theme_light()
```

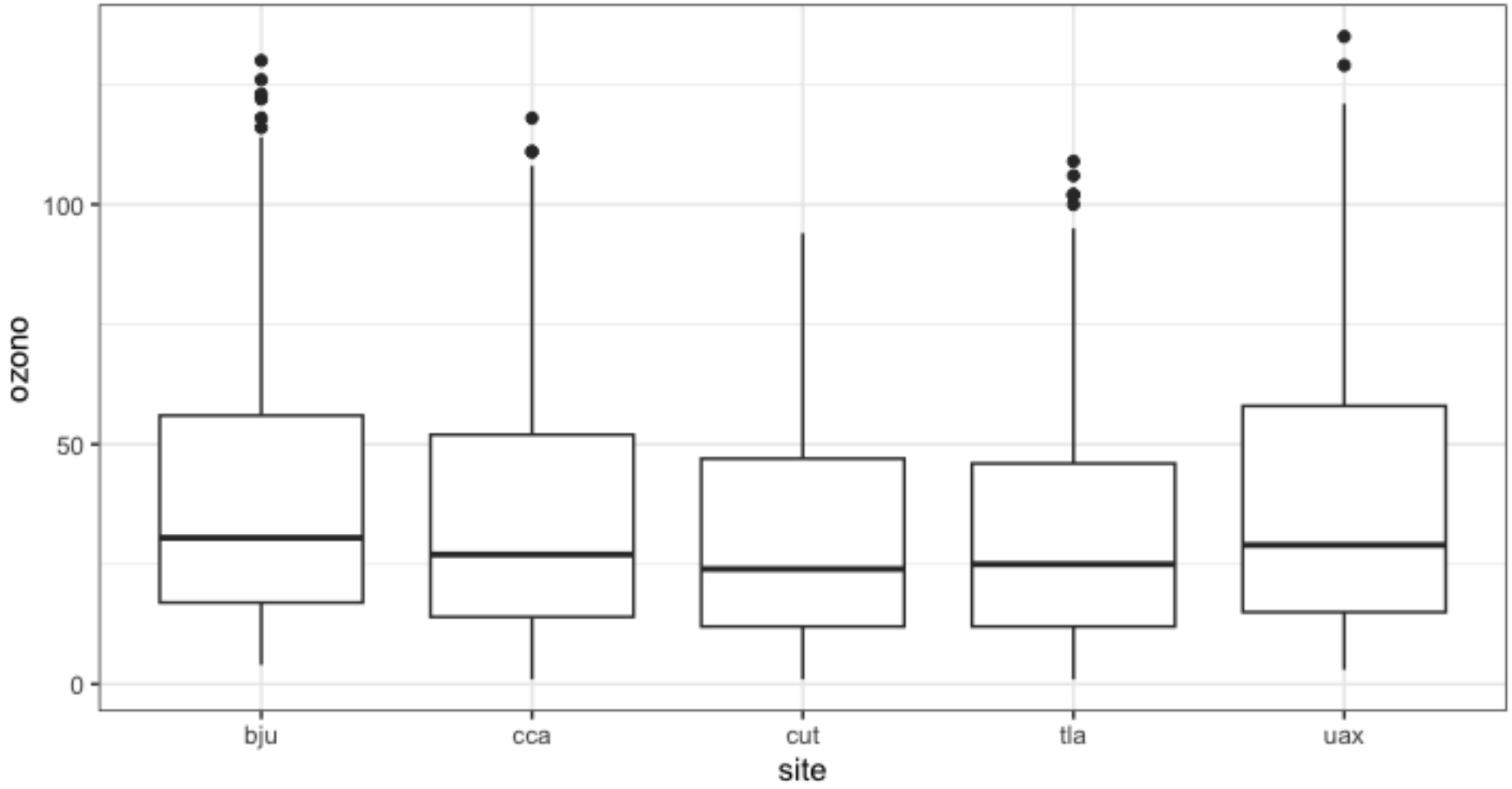


```
ggplot(data = ozono_w) +
    geom_boxplot(aes(x = "cca", cca), colour = "dodgerblue4", fill = "dodgerblue1") +
    geom_boxplot(aes(x = "bju", bju), colour = "#548B54", fill = "#98FB98") +
    theme_light()
```

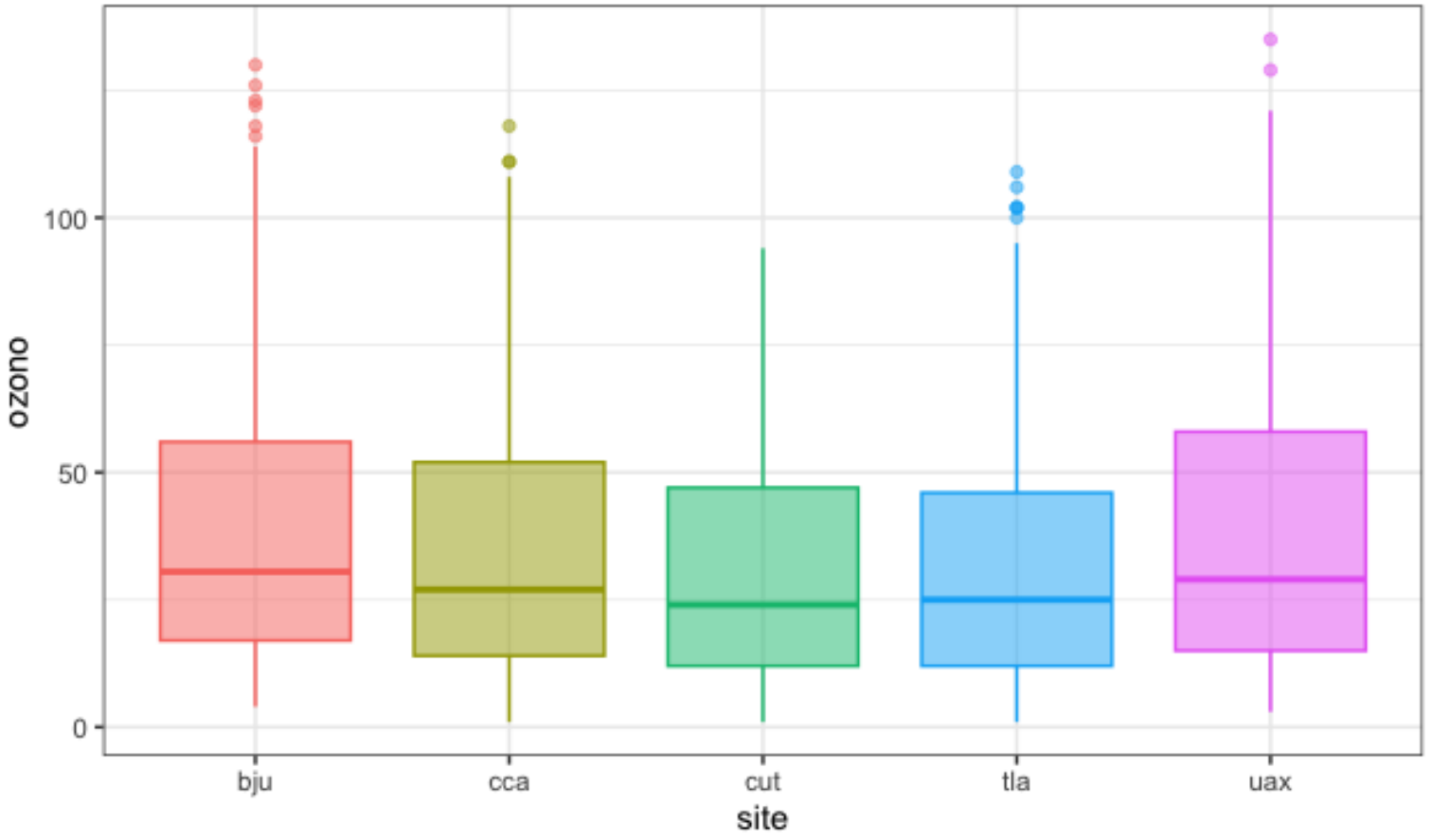


```
ggplot(data = ozono_l) +
    geom_boxplot(aes(site, ozono)) +
    theme_light()
```

Usamos malla en formato long



```
ggplot(data = ozono_l) +
    geom_boxplot(aes(site, ozono, color = site, fill = site), alpha = 0.6) +
    theme_light()
```




```
> ozono_w
# A tibble: 500 × 7
   date        hour   bju   cca   cut   tla   uax
   <date>     <int> <dbl> <dbl> <dbl> <dbl> <dbl>
 1 2023-09-01     1    NA    NA    NA    NA    NA
 2 2023-09-01     2    NA    NA    NA    NA    NA
 3 2023-09-01     3    NA    NA    NA    NA    NA
 4 2023-09-01     4    28    29    24    24    28
 5 2023-09-01     5    27    25    18     9    15
 6 2023-09-01     6    23    23     1     9    10
 7 2023-09-01     7    15    15     1     1     5
 8 2023-09-01     8    10    11     3     4     9
 9 2023-09-01     9    20    15     6     9    19
10 2023-09-01    10    29    27    15    23    28
```

```
ozono_w %>% summarise_at(vars(-date), mean, na.rm = TRUE)

ozono_w %>% summarise(across(-date, ~mean(.x, na.rm = TRUE)))

ozono_w %>% summarise_if(is.numeric, mean, na.rm = TRUE)
```

```
    bju   cca   cut   tla   uax
  <dbl> <dbl> <dbl> <dbl> <dbl>
1  39.3  35.5  30.4  31.3  38.3
```

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 de la variable ozono.
- Guarde el resultado como o3_dia
- Exporte su archivo con extensión csv y rds

Tidy Animated Verbs

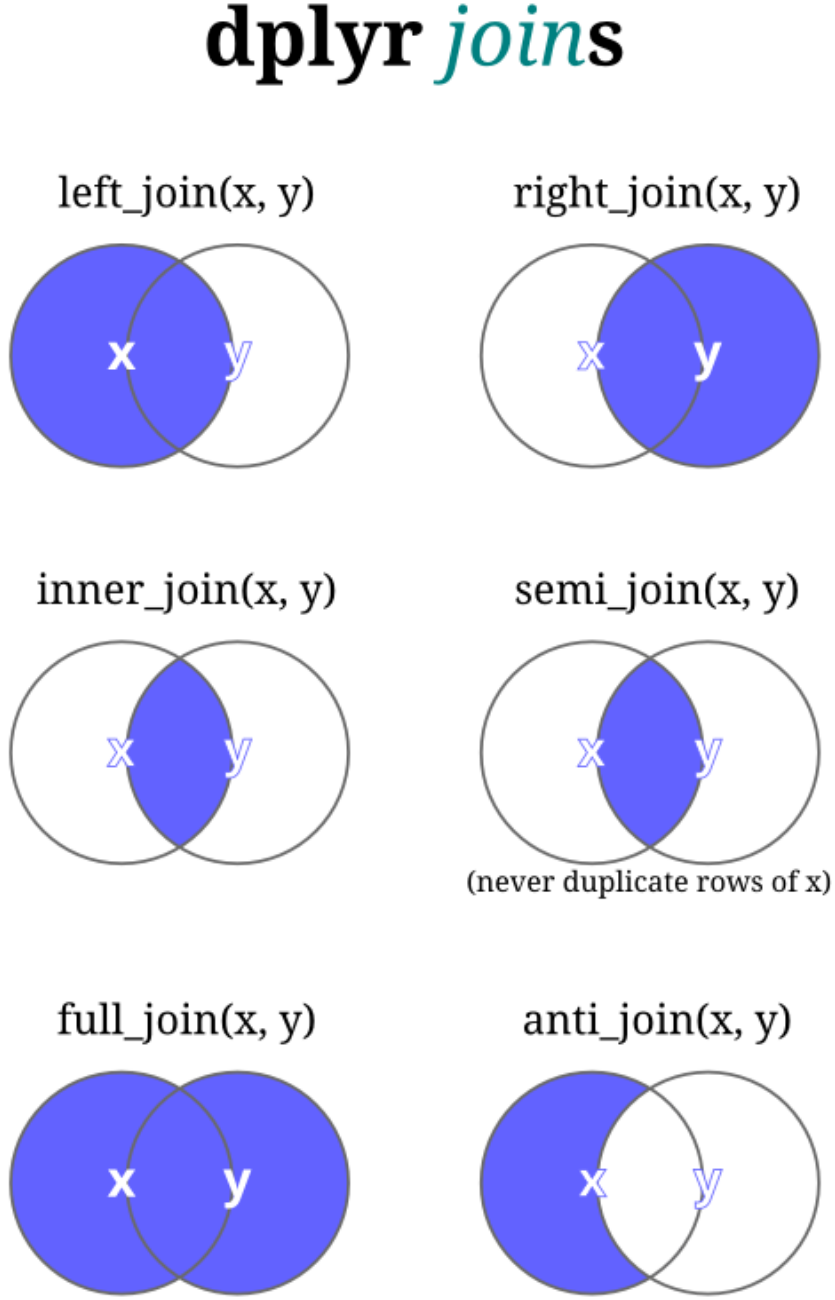
Garrick Aden-Buie – @grrrck – garrickadenbuie.com. Set operations contributed by Tyler Grant Smith.

- **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



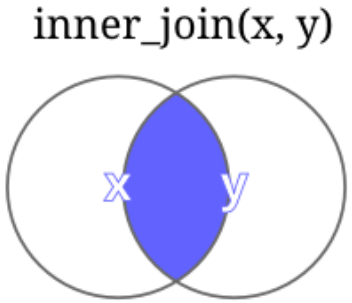
inner_join(tbl_1, tbl_2, by = "cvegeo")

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

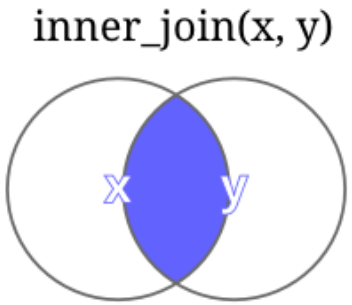
inner_join(tbl_1, tbl_2, by = "cvegeo")

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 en la posición X

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

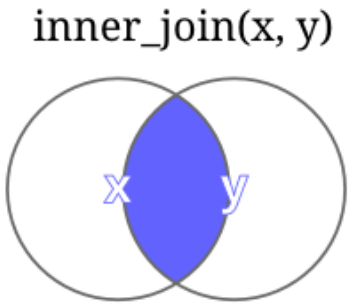
inner_join(tbl_1, tbl_2, by = c("cvegeo" = "cve_mun"))

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

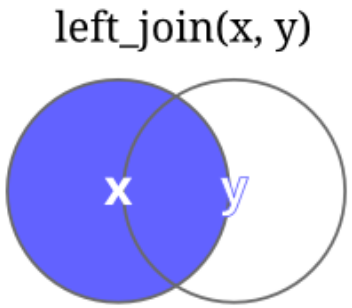
left_join(tbl_1, tbl_2, by = "cvegeo")

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

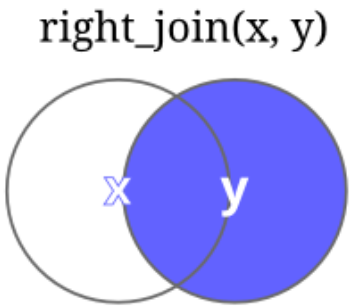
right_join(tbl_1, tbl_2, by = "cvegeo")

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")

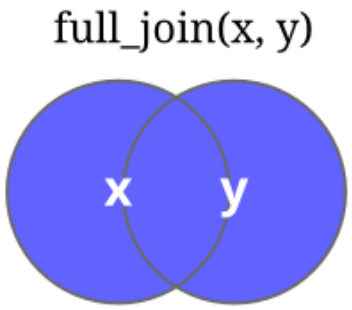
tbl_1

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



INSP/ESPM
Escuela de Salud Pública de México

cdmx

mun_cve	mun_nom	defun	pob
09002	Azcapotzalco	5860	432205
09003	Coyoacán	7242	614447
09006	Iztapalapa	21334	1835486
09012	Tlalpan	6456	699928
09015	Cuauhtémoc	7008	545884

sun

mun_cve	cve_sun	cve_nom	pob
09002	M09.01	Valle de México	400567
09009	M09.01	Valle de México	137931
09012	M09.01	Valle de México	667260
09013	M09.01	Valle de México	415541

¿cómo agrego la zona metropolitana a cada municipio de CDMX y elimino los municipios en los que no tenga esa información?

¿Cuántas filas tendrá esa nueva tabla?

```
inner_join(cdmx, sun, by = "mun_cve")
```

mun_cve	mun_nom	defun	pob	cve_sun	nom_sun	pob_sun
09002	Azcapotzalco	5860	432205	M09.01	Valle de México	400567
09012	Tlalpan	6456	699928	M09.01	Valle de México	667260

INSP/ESPM
Escuela de Salud Pública de México

cdmx

mun_cve	mun_nom	defun	pob
09002	Azcapotzalco	5860	432205
09003	Coyoacán	7242	614447
09006	Iztapalapa	21334	1835486
09012	Tlalpan	6456	699928
09015	Cuauhtémoc	7008	545884

sun

mun_cve	cve_sun	cve_nom	pob
09002	M09.01	Valle de México	400567
09009	M09.01	Valle de México	137931
09012	M09.01	Valle de México	667260
09013	M09.01	Valle de México	415541

¿cómo agrego la zona metropolitana a cada municipio de CDMX y preservo los municipios aunque no tengan esa información?

¿Cuántas filas tendrá esa nueva tabla?

```
left_join(cdmx, sun, by = "mun_cve")
```

mun_cve	mun_nom	defun	pob	cve_sun	nom_sun	pob_sun
09002	Azcapotzalco	5860	432205	M09.01	Valle de México	400567
09003	Coyoacán	7242	614447	NA	NA	NA
09006	Iztapalapa	21334	1835486	NA	NA	NA
09012	Tlalpan	6456	699928	M09.01	Valle de México	667260
09015	Cuauhtémoc	7008	545884	NA	NA	NA

cdmx

mun_cve	mun_nom	defun	pob
09002	Azcapotzalco	5860	432205
09003	Coyoacán	7242	614447
09006	Iztapalapa	21334	1835486
09012	Tlalpan	6456	699928
09015	Cuauhtémoc	7008	545884

sun

mun_cve	cve_sun	cve_nom	pob
09002	M09.01	Valle de México	400567
09009	M09.01	Valle de México	137931
09012	M09.01	Valle de México	667260
09013	M09.01	Valle de México	415541

right_join(cdmx, sun, by = "mun_cve")

mun_cve	mun_nom	defun	pob	cve_sun	nom_sun	pob_sun
09002	Azcapotzalco	5860	432205	M09.01	Valle de México	400567
09012	Tlalpan	6456	699928	M09.01	Valle de México	667260
09009	NA	NA	NA	M09.01	Valle de México	137931
09013	NA	NA	NA	M09.01	Valle de México	415541

anti_join(cdmx, sun, by = "mun_cve")

mun_cve	mun_nom	defun	pob
09003	Coyoacán	7242	614447
09006	Iztapalapa	21334	1835486
09015	Cuauhtémoc	7008	545884

INSP/ESPM
Escuela de Salud Pública de México

cdmx

mun_cve	mun_nom	defun	pob
09002	Azcapotzalco	5860	432205
09003	Coyoacán	7242	614447
09006	Iztapalapa	21334	1835486
09012	Tlalpan	6456	699928
09015	Cuauhtémoc	7008	545884

sun

mun_cve	cve_sun	cve_nom	pob
09002	M09.01	Valle de México	400567
09009	M09.01	Valle de México	137931
09012	M09.01	Valle de México	667260
09013	M09.01	Valle de México	415541

```
full_join(cdmx, sun, by = "mun_cve")
```

mun_cve	mun_nom	defun	pob	cve_sun	nom_sun	pob_sun
09002	Azcapotzalco	5860	432205	M09.01	Valle de México	400567
09003	Coyoacán	7242	614447	NA	NA	NA
09006	Iztapalapa	21334	1835486	NA	NA	NA
09012	Tlalpan	6456	699928	M09.01	Valle de México	667260
09015	Cuauhtémoc	7008	545884	NA	NA	NA
09009	NA	NA	NA	M09.01	Valle de México	137931
09013	NA	NA	NA	M09.01	Valle de México	415541