Procesando datos con el paquete tidyverse

Big Data e Inteligencia Territorial

¿Qué es Tidyverse?

Tidyverse

`Tidyverse` es una colección de paquetes de R, pensados para denominada "ciencia de datos".

Comparten la misma filosofía de uso, por lo que trabajan en armonía entre unos y otros.

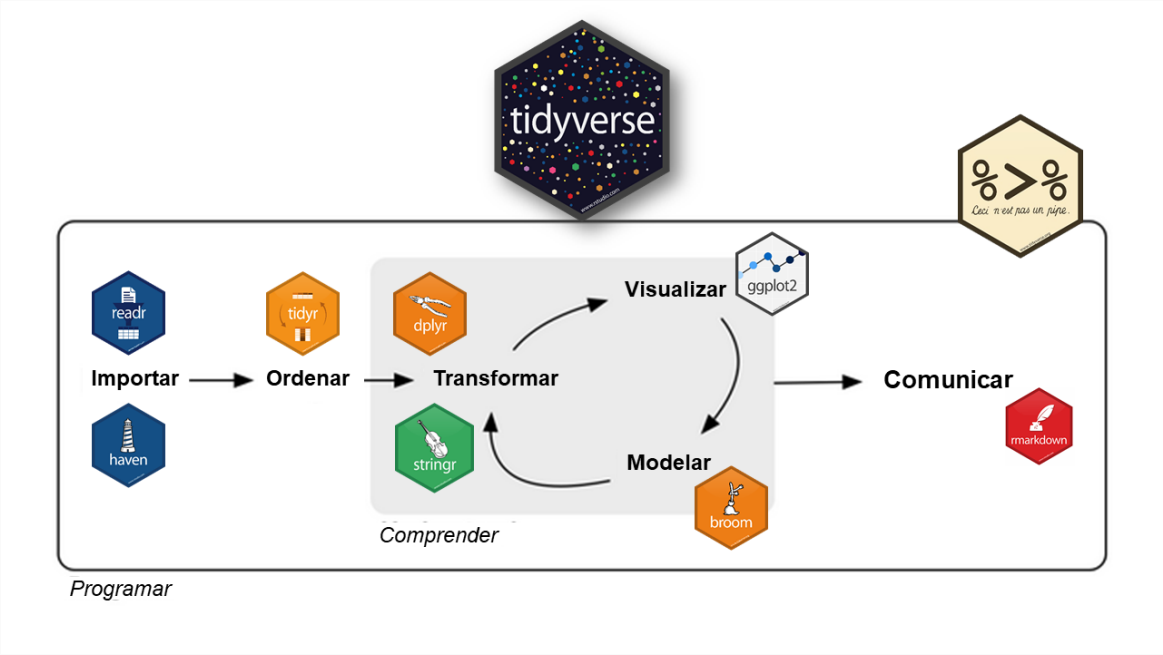
¿Por qué tidyverse?

¿Por qué tidyverse?

- Orientado a ser leído y escrito por y para seres humanos

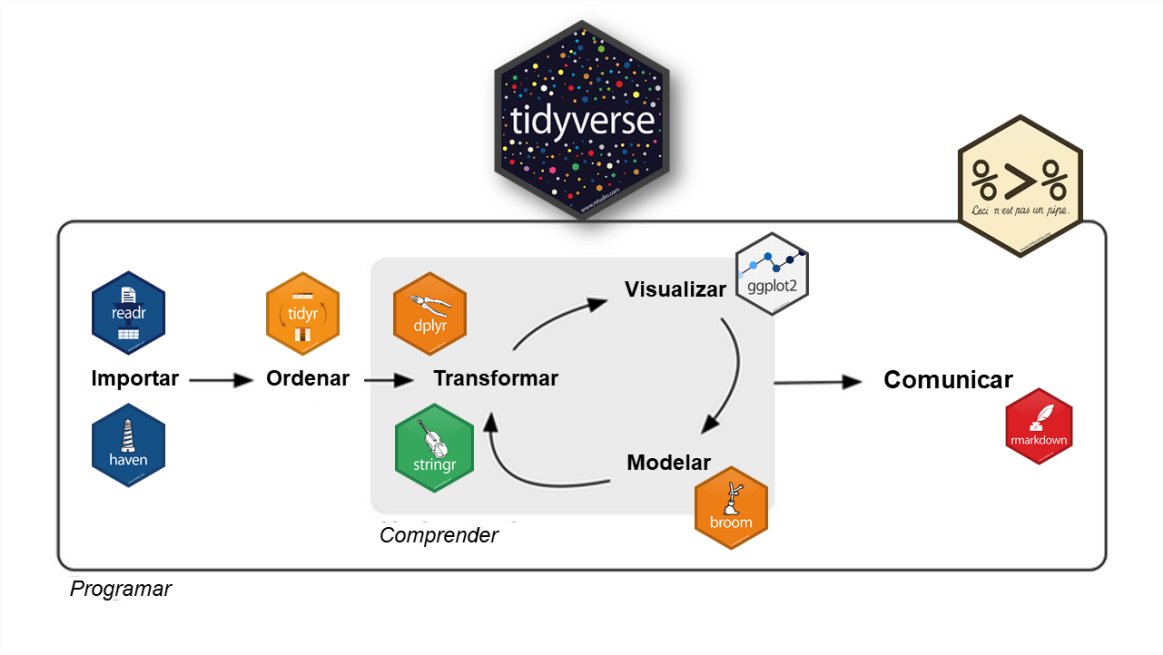
¿Por qué tidyverse?

- Orientado a ser leído y escrito por y para seres humanos
- Funciones no pensadas para una tarea específica sino para un proceso de trabajo



¿Por qué tidyverse?

- Orientado a ser leído y escrito por y para seres humanos
- Funciones no pensadas para una tarea específica sino para un proceso de trabajo



- Su comunidad, basada en los principios del código abierto y trabajo colaborativo

Instalación y uso

- Sólo una vez (por computadora):

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

Instalación y uso

- Sólo una vez (por computadora):

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

- En cada inicio de sesión de R o Rstudio:

```
library(tidyverse)
```

Instalación y uso

- Sólo una vez (por computadora):

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

- En cada inicio de sesión de R o Rstudio:

```
library(tidyverse)
```

No es necesario esto:

```
install.packages("dplyr")
install.packages("tidyr")
install.packages("ggplot2")
```

Hoja de ruta

Presentación de los paquetes `dplyr` y `tidyr`

✔ dplyr

- [x] `select()`
- [x] `filter()`
- [x] `mutate()`
- [x] `rename()`
- [x] `arragne()`
- [x] `summarise()`
- [x] `group_by()`

✔ tidyr

- [x] `pivot_longer()`
- [x] `pivot_wider()`

✔ magrittr

- [x] `%>%`

```
base_covid ← read.table("entradas/base_covid_muestra.txt", sep = ",", header = T, fileEncoding
```

EL PIPE

Una forma de escribir

magrittr

Sin EL PIPE:

```
table(base_covid$sexo)
```

--

```
    F     M    NR 
93963 86742  1975 
```

magrittr

Sin EL PIPE:

```
table(base_covid$sexo)
```

--

```
    F     M    NR 
93963 86742  1975 
```

Con EL PIPE

```
base_covid$sexo %>%
  table()
```

--

.

```
    F     M    NR 
93963 86742  1975 
```

magrittr - una forma de escribir

Caso: Deseo obtener la distribución relativa de casos por sexo:

```
base_prueba ← data.frame(var1 = c("varon","mujer","mujer","varon","mujer","mujer"),
                         var2 = c(23,67,42,25,73,11))
```

Funciones:

`table()` - `prop.table()` - `round()`

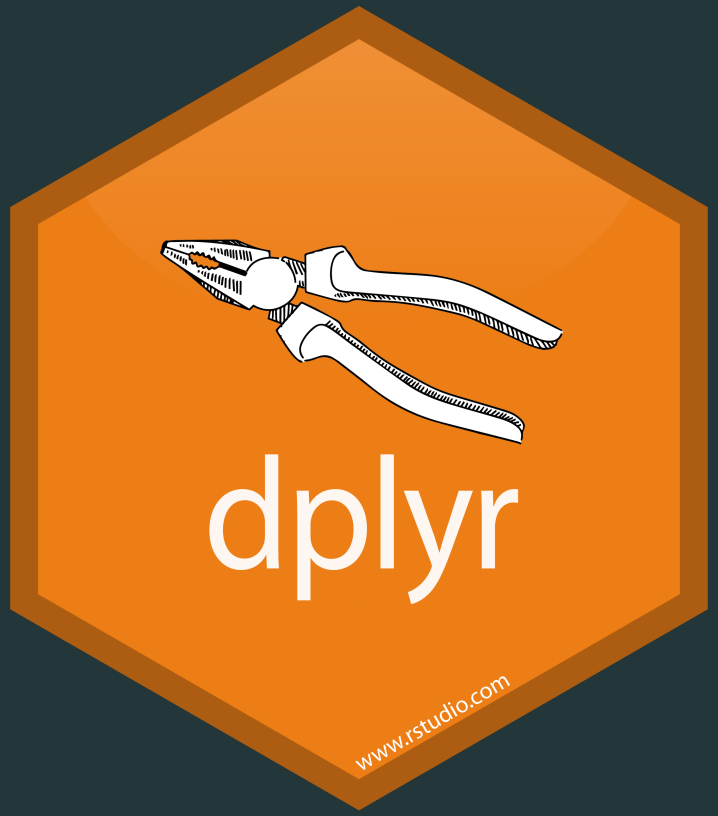
magrittr - una forma de escribir

Caso: Deseo obtener la distribución relativa de casos por sexo:

```
base_prueba ← data.frame(var1 = c("varon","mujer","mujer","varon","mujer","mujer"),
                         var2 = c(23,67,42,25,73,11))
```

Funciones:

`table()` - `prop.table()` - `round()`


dplyr
www.rstudio.com

Funciones del paquete dplyr:

Función	Acción
`select()`	*selecciona o descarta variables*
`filter()`	*selecciona filas*
`mutate()`	*crea / edita variables*
`rename()`	*renombra variables*
`group_by()`	*segmenta en funcion de una variable*
`summarize()`	*genera una tabla de resúmen*

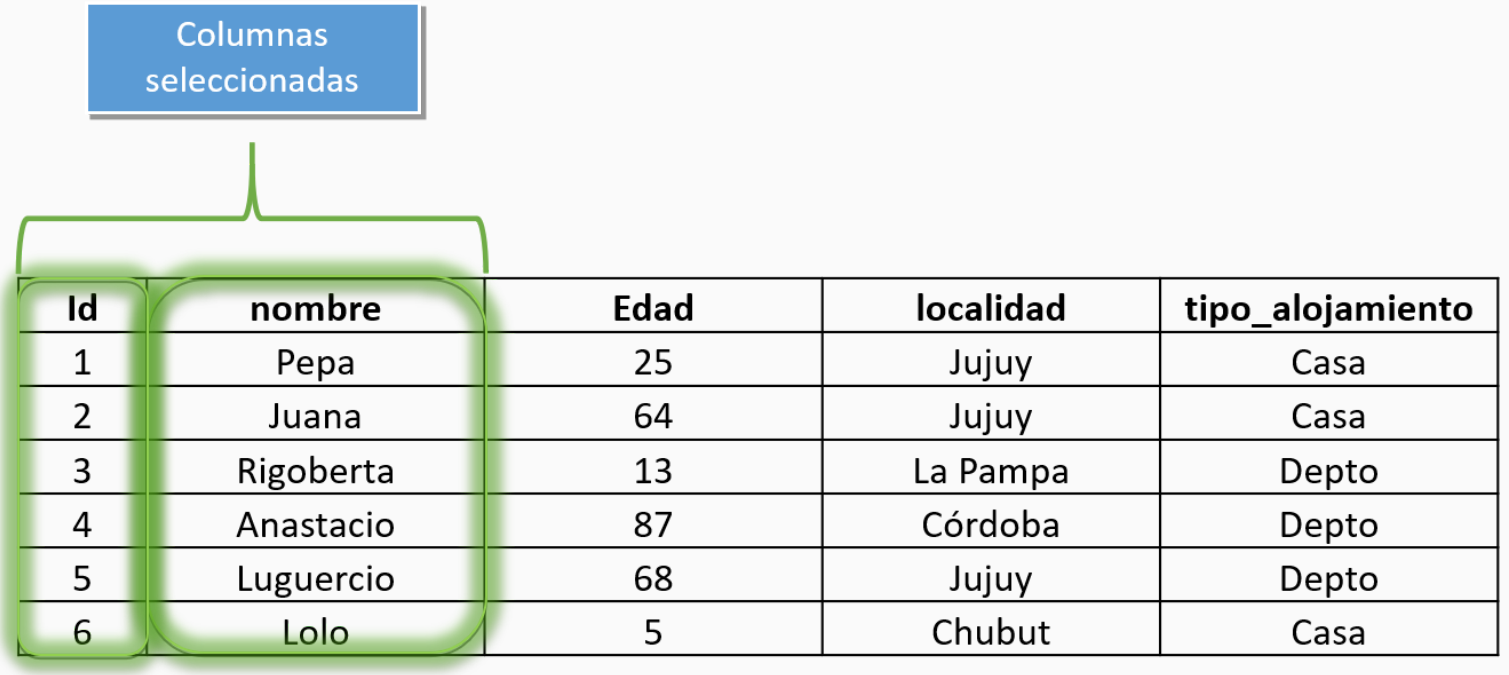
select()

Elije o descarta columnas de una base de datos

select()

La función tiene el siguiente esquema:

```
base_de_datos %>%
  select(id, nombre)
```



Id	nombre	Edad	localidad	tipo_alojamiento
1	Pepa	25	Jujuy	Casa
2	Juana	64	Jujuy	Casa
3	Rigoberta	13	La Pampa	Depto
4	Anastacio	87	Córdoba	Depto
5	Luguercio	68	Jujuy	Depto
6	Lolo	5	Chubut	Casa

select()

Supongamos que debo realizar un pequeño informe para caracterizar los CASOS COVID-19 registrados en el país, según la **edad** y **sexo**.

select()

Supongamos que debo realizar un pequeño informe para caracterizar los CASOS COVID-19 registrados en el país, según la **edad** y **sexo**.

Para ello, en base a la metadata, identifico los nombres de las variables en cuestión: **sexo** / **edad**

select()

Supongamos que debo realizar un pequeño informe para caracterizar los CASOS COVID-19 registrados en el país, según la **edad** y **sexo**.

Para ello, en base a la metadata, identifico los nombres de las variables en cuestión: **sexo** / **edad**

Luego, procedemos a cargar las librerías que voy a utilizar:

```
library(tidyverse)
```

select()

Supongamos que debo realizar un pequeño informe para caracterizar los CASOS COVID-19 registrados en el país, según la **edad** y **sexo**.

Para ello, en base a la metadata, identifico los nombres de las variables en cuestión: **sexo** / **edad**

Luego, procedemos a cargar las librerías que voy a utilizar:

```
library(tidyverse)
```

Finalmente, creamos un objeto en donde importo la base de datos con la función `read.table()`:

```
base_codiv ← read.table("entradas/base_covid_muestra.txt", sep = ",", header = T, fileEncoding
```

select() - nombre de las variables

selecciono las columnas que deseo de la base de datos:

```
base_covid_seleccion ← base_covid %>%
  select(sexo, edad)
```

select() - nombre de las variables

selecciono las columnas que deseo de la base de datos:

```
base_covid_seleccion ← base_covid %>%
  select(sexo, edad)
```

Chequeo la operación:

```
colnames(base_covid_seleccion)
```

```
[1] "sexo" "edad"
```

select() - por posición de la columna

Supongamos que quiero las columnas **"id_evento_caso"**, **"edad"** y **"edad_años_meses"**

select() - por posición de la columna

Supongamos que quiero las columnas **"id_evento_caso"**, **"edad"** y **"edad_años_meses"**

1) Chequeo la posición de las columnas que deseo:

```
colnames(base_covid)
```

select() - por posición de la columna

2) Aplico la función `select()` en base a la posición de las columnas:

```
base_covid_seleccion ← base_covid %>%
  select(1, 3, 4)
```

select() - por posición de la columna

2) Aplico la función `select()` en base a la posición de las columnas:

```
base_covid_seleccion ← base_covid %>%
  select(1, 3, 4)
```

chequeo seleccion:

```
colnames(base_covid_seleccion)
```

```
[1] "id_evento_caso"  "edad"            "edad_años_meses"
```

Otra forma de selecionar

```
base_covid
```

```
# A tibble: 182,680 x 25
   id_evento_caso sexo   edad edad_años_meses residencia_pais_~ residencia_p
            <dbl> <chr> <dbl> <chr>           <chr>             <chr>
 1         748361 NR       23 Años            Líbano            SIN ESPECIFI
 2         748780 F        53 Años            Argentina         CABA
 3         751658 M        44 Años            Argentina         CABA
 4         755897 F        29 Años            Argentina         CABA
 5         756503 M        54 Años            Argentina         CABA
 6         758578 M         2 Años            Argentina         CABA
 7         762704 M        41 Años            Argentina         CABA
 8         763097 M        53 Años            Argentina         CABA
 9         764087 F        70 Años            Argentina         CABA
10         765127 M        30 Años            Argentina         CABA
# ... with 182,670 more rows, and 19 more variables:
#   residencia_departamento_nombre <chr>, carga_provincia_nombre <chr>,
#   fecha_inicio_sintomas <date>, fecha_apertura <date>, sepi_apertura <dbl>
#   fecha_internacion <date>, cuidado_intensivo <chr>,
#   fecha_cui_intensivo <lgl>, fallecido <chr>, fecha_fallecimiento <date>,
#   asistencia_respiratoria_mecanica <chr>, carga_provincia_id <chr>,
#   origen_financiamiento <chr>, clasificacion <chr>,
#   clasificacion_resumen <chr>, residencia_provincia_id <chr>,
#   fecha_diagnostico <date>, residencia_departamento_id <chr>,
#   ultima_actualizacion <date>
```

Otra forma de selecionar

```
base_covid %>%
  select(2:4)
```

```
# A tibble: 182,680 x 3
   sexo   edad edad_años_meses
   <chr> <dbl> <chr>
 1 NR       23 Años
 2 F        53 Años
 3 M        44 Años
 4 F        29 Años
 5 M        54 Años
 6 M         2 Años
 7 M        41 Años
 8 M        53 Años
 9 F        70 Años
10 M        30 Años
# ... with 182,670 more rows
```

Una más!

Otra forma de selecionar

```
base_covid
```

```
# A tibble: 182,680 x 25
   id_evento_caso sexo   edad edad_años_meses residencia_pais_~ residencia_p
            <dbl> <chr> <dbl> <chr>           <chr>             <chr>
 1         748361 NR       23 Años            Líbano            SIN ESPECIFI
 2         748780 F        53 Años            Argentina         CABA
 3         751658 M        44 Años            Argentina         CABA
 4         755897 F        29 Años            Argentina         CABA
 5         756503 M        54 Años            Argentina         CABA
 6         758578 M         2 Años            Argentina         CABA
 7         762704 M        41 Años            Argentina         CABA
 8         763097 M        53 Años            Argentina         CABA
 9         764087 F        70 Años            Argentina         CABA
10         765127 M        30 Años            Argentina         CABA
# ... with 182,670 more rows, and 19 more variables:
#   residencia_departamento_nombre <chr>, carga_provincia_nombre <chr>,
#   fecha_inicio_sintomas <date>, fecha_apertura <date>, sepi_apertura <dbl>
#   fecha_internacion <date>, cuidado_intensivo <chr>,
#   fecha_cui_intensivo <lgl>, fallecido <chr>, fecha_fallecimiento <date>,
#   asistencia_respiratoria_mecanica <chr>, carga_provincia_id <chr>,
#   origen_financiamiento <chr>, clasificacion <chr>,
#   clasificacion_resumen <chr>, residencia_provincia_id <chr>,
#   fecha_diagnostico <date>, residencia_departamento_id <chr>,
#   ultima_actualizacion <date>
```

Otra forma de selecionar

```
base_covid %>%
  select(edad:residencia_departamento_id)
```

```
# A tibble: 182,680 x 22
    edad edad_años_meses residencia_pais_~ residencia_provin~ residencia_dep
   <dbl> <chr>           <chr>             <chr>               <chr>
 1    23 Años            Líbano            SIN ESPECIFICAR     SIN ESPECIFICA
 2    53 Años            Argentina         CABA                SIN ESPECIFICA
 3    44 Años            Argentina         CABA                SIN ESPECIFICA
 4    29 Años            Argentina         CABA                SIN ESPECIFICA
 5    54 Años            Argentina         CABA                SIN ESPECIFICA
 6     2 Años            Argentina         CABA                SIN ESPECIFICA
 7    41 Años            Argentina         CABA                SIN ESPECIFICA
 8    53 Años            Argentina         CABA                SIN ESPECIFICA
 9    70 Años            Argentina         CABA                SIN ESPECIFICA
10    30 Años            Argentina         CABA                SIN ESPECIFICA
# ... with 182,670 more rows, and 17 more variables:
#   carga_provincia_nombre <chr>, fecha_inicio_sintomas <date>,
#   fecha_apertura <date>, sepi_apertura <dbl>, fecha_internacion <date>,
#   cuidado_intensivo <chr>, fecha_cui_intensivo <lgl>, fallecido <chr>,
#   fecha_fallecimiento <date>, asistencia_respiratoria_mecanica <chr>,
#   carga_provincia_id <chr>, origen_financiamiento <chr>, clasificacion <ch
#   clasificacion_resumen <chr>, residencia_provincia_id <chr>,
#   fecha_diagnostico <date>, residencia_departamento_id <chr>
```

Una más!

Otra forma de selecionar

```
base_covid
```

```
# A tibble: 182,680 x 25
   id_evento_caso sexo   edad edad_años_meses residencia_pais_~ residencia_p
            <dbl> <chr> <dbl> <chr>           <chr>             <chr>
 1         748361 NR       23 Años            Líbano            SIN ESPECIFI
 2         748780 F        53 Años            Argentina         CABA
 3         751658 M        44 Años            Argentina         CABA
 4         755897 F        29 Años            Argentina         CABA
 5         756503 M        54 Años            Argentina         CABA
 6         758578 M         2 Años            Argentina         CABA
 7         762704 M        41 Años            Argentina         CABA
 8         763097 M        53 Años            Argentina         CABA
 9         764087 F        70 Años            Argentina         CABA
10         765127 M        30 Años            Argentina         CABA
# ... with 182,670 more rows, and 19 more variables:
#   residencia_departamento_nombre <chr>, carga_provincia_nombre <chr>,
#   fecha_inicio_sintomas <date>, fecha_apertura <date>, sepi_apertura <dbl>
#   fecha_internacion <date>, cuidado_intensivo <chr>,
#   fecha_cui_intensivo <lgl>, fallecido <chr>, fecha_fallecimiento <date>,
#   asistencia_respiratoria_mecanica <chr>, carga_provincia_id <chr>,
#   origen_financiamiento <chr>, clasificacion <chr>,
#   clasificacion_resumen <chr>, residencia_provincia_id <chr>,
#   fecha_diagnostico <date>, residencia_departamento_id <chr>,
#   ultima_actualizacion <date>
```

Otra forma de selecionar

```
base_covid %>%
  select(starts_with("residencia"))
```

```
# A tibble: 182,680 x 5
   residencia_pais_n~ residencia_provinc~ residencia_departam~ residencia_pr
   <chr>              <chr>               <chr>                <chr>
 1 Líbano             SIN ESPECIFICAR     SIN ESPECIFICAR      99
 2 Argentina          CABA                SIN ESPECIFICAR      02
 3 Argentina          CABA                SIN ESPECIFICAR      02
 4 Argentina          CABA                SIN ESPECIFICAR      02
 5 Argentina          CABA                SIN ESPECIFICAR      02
 6 Argentina          CABA                SIN ESPECIFICAR      02
 7 Argentina          CABA                SIN ESPECIFICAR      02
 8 Argentina          CABA                SIN ESPECIFICAR      02
 9 Argentina          CABA                SIN ESPECIFICAR      02
10 Argentina          CABA                SIN ESPECIFICAR      02
# ... with 182,670 more rows, and 1 more variable:
#   residencia_departamento_id <chr>
```

Una más!

Otra forma de selecionar

```
base_covid
```

```
# A tibble: 182,680 x 25
   id_evento_caso sexo   edad edad_años_meses residencia_pais_~ residencia_p
            <dbl> <chr> <dbl> <chr>           <chr>             <chr>
 1         748361 NR       23 Años            Líbano            SIN ESPECIFI
 2         748780 F        53 Años            Argentina         CABA
 3         751658 M        44 Años            Argentina         CABA
 4         755897 F        29 Años            Argentina         CABA
 5         756503 M        54 Años            Argentina         CABA
 6         758578 M         2 Años            Argentina         CABA
 7         762704 M        41 Años            Argentina         CABA
 8         763097 M        53 Años            Argentina         CABA
 9         764087 F        70 Años            Argentina         CABA
10         765127 M        30 Años            Argentina         CABA
# ... with 182,670 more rows, and 19 more variables:
#   residencia_departamento_nombre <chr>, carga_provincia_nombre <chr>,
#   fecha_inicio_sintomas <date>, fecha_apertura <date>, sepi_apertura <dbl>
#   fecha_internacion <date>, cuidado_intensivo <chr>,
#   fecha_cui_intensivo <lgl>, fallecido <chr>, fecha_fallecimiento <date>,
#   asistencia_respiratoria_mecanica <chr>, carga_provincia_id <chr>,
#   origen_financiamiento <chr>, clasificacion <chr>,
#   clasificacion_resumen <chr>, residencia_provincia_id <chr>,
#   fecha_diagnostico <date>, residencia_departamento_id <chr>,
#   ultima_actualizacion <date>
```

Otra forma de selecionar

```
base_covid %>%
  select(ends_with("id"))
```

```
# A tibble: 182,680 x 3
   carga_provincia_id residencia_provincia_id residencia_departamento_id
   <chr>              <chr>                    <chr>
 1 06                 99                       0
 2 02                 02                       0
 3 06                 02                       0
 4 02                 02                       0
 5 02                 02                       0
 6 06                 02                       0
 7 02                 02                       0
 8 02                 02                       0
 9 02                 02                       0
10 06                 02                       0
# … with 182,670 more rows
```

Una más!

Otra forma de selecionar

```
base_covid
```

```
# A tibble: 182,680 x 25
   id_evento_caso sexo   edad edad_años_meses residencia_pais_~ residencia_p
            <dbl> <chr> <dbl> <chr>           <chr>             <chr>
 1         748361 NR       23 Años            Líbano            SIN ESPECIFI
 2         748780 F        53 Años            Argentina         CABA
 3         751658 M        44 Años            Argentina         CABA
 4         755897 F        29 Años            Argentina         CABA
 5         756503 M        54 Años            Argentina         CABA
 6         758578 M         2 Años            Argentina         CABA
 7         762704 M        41 Años            Argentina         CABA
 8         763097 M        53 Años            Argentina         CABA
 9         764087 F        70 Años            Argentina         CABA
10         765127 M        30 Años            Argentina         CABA
# ... with 182,670 more rows, and 19 more variables:
#   residencia_departamento_nombre <chr>, carga_provincia_nombre <chr>,
#   fecha_inicio_sintomas <date>, fecha_apertura <date>, sepi_apertura <dbl>
#   fecha_internacion <date>, cuidado_intensivo <chr>,
#   fecha_cui_intensivo <lgl>, fallecido <chr>, fecha_fallecimiento <date>,
#   asistencia_respiratoria_mecanica <chr>, carga_provincia_id <chr>,
#   origen_financiamiento <chr>, clasificacion <chr>,
#   clasificacion_resumen <chr>, residencia_provincia_id <chr>,
#   fecha_diagnostico <date>, residencia_departamento_id <chr>,
#   ultima_actualizacion <date>
```

Otra forma de selecionar

```
base_covid %>%
  select(contains("provincia"))
```

```
# A tibble: 182,680 x 4
   residencia_provincia~ carga_provincia_no~ carga_provincia~ residencia_pro
   <chr>                 <chr>               <chr>            <chr>
 1 SIN ESPECIFICAR       Buenos Aires        06               99
 2 CABA                  CABA                02               02
 3 CABA                  Buenos Aires        06               02
 4 CABA                  CABA                02               02
 5 CABA                  CABA                02               02
 6 CABA                  Buenos Aires        06               02
 7 CABA                  CABA                02               02
 8 CABA                  CABA                02               02
 9 CABA                  CABA                02               02
10 CABA                  Buenos Aires        06               02
# ... with 182,670 more rows
```

PRÁCTICA

Práctica

1) Armar una carpeta y un proyecto de trabajo con la estructura vista en clase. Luego ubicar la base de datos en la carpeta correspondiente.

2) Abrir un script y crear un objeto en donde importamos la base (recordar tener en cuenta la extensión del archivo).

3) Crear un objeto llamado **seleccion1** cuyo contenido sea dos columnas de la base original, **seleccionadas según su nombre**

4) Crear un objeto llamado **seleccion2** cuyo contenido sea dos columnas de la base original, **seleccionadas según su posición**

5) Crear un objeto llamado **selecicon3** cuyo contenido sea las columnas en base a un cáracter específico en su nombre. Recordar las funciones *starts_with()*, *contains()* y *ends_with()*

filter()

Define los casos (filas) en base a una condición

filter()

La función tiene el siguiente esquema:

```
base_de_datos %>%
  filter(condicion)
```

Id	nombre	Edad	localidad	tipo_alojamiento
1	Pepa	25	Jujuy	Casa
2	Juana	64	Jujuy	Casa
3	Rigoberta	13	La Pampa	Depto
4	Anastacio	87	Córdoba	Depto
5	Luguercio	68	Jujuy	Depto
6	Lolo	5	Chubut	Casa

filter()

- Por ejemplo:

```
base %>%
  filter(Edad > 70)
```

Id	nombre	Edad	localidad	tipo_alojamiento
1	Pepa	25	Jujuy	Casa
2	Juana	64	Jujuy	Casa
3	Rigoberta	13	La Pampa	Depto
4	Anastacio	87	Córdoba	Depto
5	Luguercio	68	Jujuy	Depto
6	Lolo	5	Chubut	Casa

filter()

Caso: Quiero quedarme unicamente con aquella población que tuvieron **asistencia respiratoria mecánica**.

Según la metadata, la variable que necesito para filtrar se llama `asistencia_respiratoria_mecanica`, cuyas categorías son `SI / NO`:

filter()

```
base_covid
```

```
# A tibble: 182,680 x 25
   id_evento_caso sexo   edad edad_años_meses residencia_pais_~ reside
            <dbl> <chr> <dbl> <chr>           <chr>              <chr>
 1         748361 NR       23 Años            Líbano             SIN ES
 2         748780 F        53 Años            Argentina          CABA
 3         751658 M        44 Años            Argentina          CABA
 4         755897 F        29 Años            Argentina          CABA
 5         756503 M        54 Años            Argentina          CABA
 6         758578 M         2 Años            Argentina          CABA
 7         762704 M        41 Años            Argentina          CABA
 8         763097 M        53 Años            Argentina          CABA
 9         764087 F        70 Años            Argentina          CABA
10         765127 M        30 Años            Argentina          CABA
# ... with 182,670 more rows, and 19 more variables:
#   residencia_departamento_nombre <chr>, carga_provincia_nombre <chr>
#   fecha_inicio_sintomas <date>, fecha_apertura <date>, sepi_apertura
#   fecha_internacion <date>, cuidado_intensivo <chr>,
#   fecha_cui_intensivo <lgl>, fallecido <chr>, fecha_fallecimiento <da
#   asistencia_respiratoria_mecanica <chr>, carga_provincia_id <chr>,
#   origen_financiamiento <chr>, clasificacion <chr>,
#   clasificacion_resumen <chr>, residencia_provincia_id <chr>,
#   fecha_diagnostico <date>, residencia_departamento_id <chr>,
#   ultima_actualizacion <date>
```

filter()

```
base_covid %>%
  select(id_evento_caso,
         asistencia_respiratoria_mecanica)
```

```
# A tibble: 182,680 x 2
   id_evento_caso asistencia_respiratoria_mecanica
            <dbl> <chr>
 1         748361 NO
 2         748780 NO
 3         751658 NO
 4         755897 NO
 5         756503 NO
 6         758578 NO
 7         762704 NO
 8         763097 NO
 9         764087 NO
10         765127 NO
# ... with 182,670 more rows
```

filter()

```
base_covid %>%
  select(id_evento_caso,
         asistencia_respiratoria_mecanica) %>%
  filter(asistencia_respiratoria_mecanica == "SI")
```

```
# A tibble: 364 x 2
   id_evento_caso asistencia_respiratoria_mecanica
            <dbl> <chr>
 1         773376 SI
 2         816251 SI
 3         836895 SI
 4         852977 SI
 5         870899 SI
 6         901291 SI
 7         904720 SI
 8         937062 SI
 9         998935 SI
10        1125035 SI
# ... with 354 more rows
```

filter()

Operadores para filtrar:

Condición	Acción
`==`	*igual*
`%in%`	*incluye*
`≠`	*distinto*
`>`	*mayor que*
`<`	*menor que*
`⩾`	*mayor o igual que*
`⩽`	*menor o igual que*

Operador	Descripción
`&`	*y* - Cuando se cumplen ambas condiciones
|	*o* - Cuando se cumple una u otra condición

filter()

Caso: Quiero quedarme con la población de la *Ciudad Autónoma de buenos Aires* **o** de *la provincia Buenos aires* **y** que haya recibido asistencia respiratoria mecánica:

- Provincias CABA o Buenos Aires (`residencia_provincia_nombre %in% c("CABA", "Buenos Aires")`);
- Asistencia respiratorio mecanica (`asistencia_respiratoria_mecanica == "SI"`);

filter

```
base_covid
```

```
# A tibble: 182,680 x 25
   id_evento_caso sexo   edad edad_años_meses residenc
            <dbl> <chr> <dbl> <chr>           <chr>
 1         748361 NR       23 Años            Líbano
 2         748780 F        53 Años            Argentin
 3         751658 M        44 Años            Argentin
 4         755897 F        29 Años            Argentin
 5         756503 M        54 Años            Argentin
 6         758578 M         2 Años            Argentin
 7         762704 M        41 Años            Argentin
 8         763097 M        53 Años            Argentin
 9         764087 F        70 Años            Argentin
10         765127 M        30 Años            Argentin
# ... with 182,670 more rows, and 19 more variables:
#   residencia_departamento_nombre <chr>, carga_provin
#   fecha_inicio_sintomas <date>, fecha_apertura <date
#   fecha_internacion <date>, cuidado_intensivo <chr>,
#   fecha_cui_intensivo <lgl>, fallecido <chr>, fecha_
#   asistencia_respiratoria_mecanica <chr>, carga_prov
#   origen_financiamiento <chr>, clasificacion <chr>,
#   clasificacion_resumen <chr>, residencia_provincia_
#   fecha_diagnostico <date>, residencia_departamento_
#   ultima_actualizacion <date>
```

filter

```
base_covid %>%
  select(id_evento_caso,
         residencia_provincia_nombre,
         asistencia_respiratoria_mecanica)
```

```
# A tibble: 182,680 x 3
   id_evento_caso residencia_provincia_nombre asistenc
            <dbl> <chr>                       <chr>
 1         748361 SIN ESPECIFICAR             NO
 2         748780 CABA                        NO
 3         751658 CABA                        NO
 4         755897 CABA                        NO
 5         756503 CABA                        NO
 6         758578 CABA                        NO
 7         762704 CABA                        NO
 8         763097 CABA                        NO
 9         764087 CABA                        NO
10         765127 CABA                        NO
# ... with 182,670 more rows
```

filter

```
base_covid %>%
  select(id_evento_caso,
         residencia_provincia_nombre,
         asistencia_respiratoria_mecanica) %>%
  filter(residencia_provincia_nombre %in% c("CABA","Buenos Aires")
```

```
# A tibble: 99,818 x 3
   id_evento_caso residencia_provincia_nombre asistenc
            <dbl> <chr>                       <chr>
 1         748780 CABA                        NO
 2         751658 CABA                        NO
 3         755897 CABA                        NO
 4         756503 CABA                        NO
 5         758578 CABA                        NO
 6         762704 CABA                        NO
 7         763097 CABA                        NO
 8         764087 CABA                        NO
 9         765127 CABA                        NO
10         766173 CABA                        NO
# ... with 99,808 more rows
```

filter

```
base_covid %>%
  select(id_evento_caso,
         residencia_provincia_nombre,
         asistencia_respiratoria_mecanica) %>%
  filter(residencia_provincia_nombre %in% c("CABA","Buenos Aires")) %>%
  filter(asistencia_respiratoria_mecanica == "SI")
```

```
# A tibble: 200 x 3
   id_evento_caso residencia_provincia_nombre asistenc
            <dbl> <chr>                       <chr>
 1         773376 CABA                        SI
 2         816251 CABA                        SI
 3         836895 CABA                        SI
 4         852977 CABA                        SI
 5         870899 Buenos Aires                SI
 6         901291 CABA                        SI
 7         904720 CABA                        SI
 8         937062 CABA                        SI
 9         998935 Buenos Aires                SI
10        1125035 CABA                        SI
# ... with 190 more rows
```

PRÁCTICA

Práctica

- Crear un objeto que contenga las variables **id_evento_viaje**, **residencia_provincia_nombre** y **asistencia_respiratoria_mecanica** y cuya población sea sólo aquella de las **provincias de la Patagonia** que **NO** recibió ayuda respiratoria mecánica.

mutate()

Creoa / edita variables (columnas)

mutate()

- En R base:

```
base_de_dato$var_nueva ← base_de_datos$var_1 + base_de_datos$var_2
```

- En `tidyverse`:

```
base_de_datos %>%
mutate(var_nueva = var_1 + var_2)
```

mutate()

Caso: Quiero una variable nueva con todos los nombres de provincia en minúscula (y evitar errores por filtrar mal el nombre):

mutate()

```
base_covid
```

```
# A tibble: 182,680 x 25
   id_evento_caso sexo   edad edad_años_meses residenc
            <dbl> <chr> <dbl> <chr>           <chr>
 1         748361 NR       23 Años            Líbano
 2         748780 F        53 Años            Argentin
 3         751658 M        44 Años            Argentin
 4         755897 F        29 Años            Argentin
 5         756503 M        54 Años            Argentin
 6         758578 M         2 Años            Argentin
 7         762704 M        41 Años            Argentin
 8         763097 M        53 Años            Argentin
 9         764087 F        70 Años            Argentin
10         765127 M        30 Años            Argentin
# ... with 182,670 more rows, and 19 more variables:
#   residencia_departamento_nombre <chr>, carga_provin
#   fecha_inicio_sintomas <date>, fecha_apertura <date
#   fecha_internacion <date>, cuidado_intensivo <chr>,
#   fecha_cui_intensivo <lgl>, fallecido <chr>, fecha_
#   asistencia_respiratoria_mecanica <chr>, carga_prov
#   origen_financiamiento <chr>, clasificacion <chr>,
#   clasificacion_resumen <chr>, residencia_provincia_
#   fecha_diagnostico <date>, residencia_departamento_
#   ultima_actualizacion <date>
```

mutate()

```
base_covid %>%
  select(id_evento_caso, residencia_provincia_nombre)
```

```
# A tibble: 182,680 x 2
   id_evento_caso residencia_provincia_nombre
            <dbl> <chr>
 1         748361 SIN ESPECIFICAR
 2         748780 CABA
 3         751658 CABA
 4         755897 CABA
 5         756503 CABA
 6         758578 CABA
 7         762704 CABA
 8         763097 CABA
 9         764087 CABA
10         765127 CABA
# ... with 182,670 more rows
```

mutate()

```
base_covid %>%
  select(id_evento_caso, residencia_provincia_nombre) %>%
  mutate(res_prov_nom_minus = tolower(residencia_provincia_nombre)
```

```
# A tibble: 182,680 x 3
   id_evento_caso residencia_provincia_nombre res_prov
            <dbl> <chr>                       <chr>
 1         748361 SIN ESPECIFICAR             sin espe
 2         748780 CABA                        caba
 3         751658 CABA                        caba
 4         755897 CABA                        caba
 5         756503 CABA                        caba
 6         758578 CABA                        caba
 7         762704 CABA                        caba
 8         763097 CABA                        caba
 9         764087 CABA                        caba
10         765127 CABA                        caba
# ... with 182,670 more rows
```

mutate()

Caso: Supongamos que quiero crear la variable de año y mes de fallecimiento (variable **fecha_fallecimiento**):

mutate()

```
base_covid
```

```
# A tibble: 182,680 x 25
   id_evento_caso sexo   edad edad_años_meses residenc
            <dbl> <chr> <dbl> <chr>           <chr>
 1         748361 NR       23 Años            Líbano
 2         748780 F        53 Años            Argentin
 3         751658 M        44 Años            Argentin
 4         755897 F        29 Años            Argentin
 5         756503 M        54 Años            Argentin
 6         758578 M         2 Años            Argentin
 7         762704 M        41 Años            Argentin
 8         763097 M        53 Años            Argentin
 9         764087 F        70 Años            Argentin
10         765127 M        30 Años            Argentin
# ... with 182,670 more rows, and 19 more variables:
#   residencia_departamento_nombre <chr>, carga_provin
#   fecha_inicio_sintomas <date>, fecha_apertura <date
#   fecha_internacion <date>, cuidado_intensivo <chr>,
#   fecha_cui_intensivo <lgl>, fallecido <chr>, fecha_
#   asistencia_respiratoria_mecanica <chr>, carga_prov
#   origen_financiamiento <chr>, clasificacion <chr>,
#   clasificacion_resumen <chr>, residencia_provincia_
#   fecha_diagnostico <date>, residencia_departamento_
#   ultima_actualizacion <date>
```

mutate()

```
base_covid %>%
  select(fallecido, fecha_fallecimiento)
```

```
# A tibble: 182,680 x 2
   fallecido fecha_fallecimiento
   <chr>     <date>
 1 NO        NA
 2 NO        NA
 3 NO        NA
 4 NO        NA
 5 NO        NA
 6 NO        NA
 7 NO        NA
 8 NO        NA
 9 NO        NA
10 NO        NA
# ... with 182,670 more rows
```

mutate()

```
base_covid %>%
  select(fallecido, fecha_fallecimiento) %>%
  filter(fallecido == "SI")
```

```
# A tibble: 1,402 x 2
   fallecido fecha_fallecimiento
   <chr>     <date>
 1 SI        2020-03-27
 2 SI        2020-04-10
 3 SI        2020-04-19
 4 SI        2020-05-01
 5 SI        2020-07-25
 6 SI        2020-05-02
 7 SI        2020-05-01
 8 SI        2020-05-29
 9 SI        2020-05-05
10 SI        2020-05-30
# ... with 1,392 more rows
```

mutate()

```
base_covid %>%
  select(fallecido, fecha_fallecimiento) %>%
  filter(fallecido == "SI") %>%
  mutate(anio = substr(x = fecha_fallecimiento,
                       start = 1,
                       stop = 4))
```

```
# A tibble: 1,402 x 3
   fallecido fecha_fallecimiento anio
   <chr>     <date>              <chr>
 1 SI        2020-03-27          2020
 2 SI        2020-04-10          2020
 3 SI        2020-04-19          2020
 4 SI        2020-05-01          2020
 5 SI        2020-07-25          2020
 6 SI        2020-05-02          2020
 7 SI        2020-05-01          2020
 8 SI        2020-05-29          2020
 9 SI        2020-05-05          2020
10 SI        2020-05-30          2020
# ... with 1,392 more rows
```

mutate()

```
base_covid %>%
  select(fallecido, fecha_fallecimiento) %>%
  filter(fallecido == "SI") %>%
  mutate(anio = substr(x = fecha_fallecimiento,
                       start = 1,
                       stop = 4)) %>%
  mutate(mes = substr(x = fecha_fallecimiento,
                      start = 6,
                      stop = 7))
```

```
# A tibble: 1,402 x 4
   fallecido fecha_fallecimiento anio  mes
   <chr>     <date>              <chr> <chr>
 1 SI        2020-03-27          2020  03
 2 SI        2020-04-10          2020  04
 3 SI        2020-04-19          2020  04
 4 SI        2020-05-01          2020  05
 5 SI        2020-07-25          2020  07
 6 SI        2020-05-02          2020  05
 7 SI        2020-05-01          2020  05
 8 SI        2020-05-29          2020  05
 9 SI        2020-05-05          2020  05
10 SI        2020-05-30          2020  05
# ... with 1,392 more rows
```

mutate() - case_when()

Función complementaria: `case_when()`, mayormente utilizada para recodificación de variables

Recodificando con mutate() y case_when()

```
base_covid
```

```
# A tibble: 182,680 x 25
   id_evento_caso sexo   edad edad_años_meses residenc
            <dbl> <chr> <dbl> <chr>           <chr>
 1         748361 NR       23 Años            Líbano
 2         748780 F        53 Años            Argentin
 3         751658 M        44 Años            Argentin
 4         755897 F        29 Años            Argentin
 5         756503 M        54 Años            Argentin
 6         758578 M         2 Años            Argentin
 7         762704 M        41 Años            Argentin
 8         763097 M        53 Años            Argentin
 9         764087 F        70 Años            Argentin
10         765127 M        30 Años            Argentin
# ... with 182,670 more rows, and 19 more variables:
#   residencia_departamento_nombre <chr>, carga_provin
#   fecha_inicio_sintomas <date>, fecha_apertura <date
#   fecha_internacion <date>, cuidado_intensivo <chr>,
#   fecha_cui_intensivo <lgl>, fallecido <chr>, fecha_
#   asistencia_respiratoria_mecanica <chr>, carga_prov
#   origen_financiamiento <chr>, clasificacion <chr>,
#   clasificacion_resumen <chr>, residencia_provincia_
#   fecha_diagnostico <date>, residencia_departamento_
#   ultima_actualizacion <date>
```

Recodificando con mutate() y case_when()

```
base_covid %>%
  select(id_evento_caso, fallecido)
```

```
# A tibble: 182,680 x 2
   id_evento_caso fallecido
            <dbl> <chr>
 1         748361 NO
 2         748780 NO
 3         751658 NO
 4         755897 NO
 5         756503 NO
 6         758578 NO
 7         762704 NO
 8         763097 NO
 9         764087 NO
10         765127 NO
# ... with 182,670 more rows
```

Recodificando con mutate() y case_when()

```
base_covid %>%
  select(id_evento_caso, fallecido) %>%
  mutate(fallecidos_rec = case_when(fallecido == "SI" ~ 1,
                                    fallecido == "NO" ~ 2))
```

```
# A tibble: 182,680 x 3
   id_evento_caso fallecido fallecidos_rec
            <dbl> <chr>               <dbl>
 1         748361 NO                      2
 2         748780 NO                      2
 3         751658 NO                      2
 4         755897 NO                      2
 5         756503 NO                      2
 6         758578 NO                      2
 7         762704 NO                      2
 8         763097 NO                      2
 9         764087 NO                      2
10         765127 NO                      2
# ... with 182,670 more rows
```

Recodificando con mutate() y case_when()

```
base_covid
```

```
# A tibble: 182,680 x 25
   id_evento_caso sexo   edad edad_años_meses residenc
            <dbl> <chr> <dbl> <chr>           <chr>
 1         748361 NR       23 Años            Líbano
 2         748780 F        53 Años            Argentin
 3         751658 M        44 Años            Argentin
 4         755897 F        29 Años            Argentin
 5         756503 M        54 Años            Argentin
 6         758578 M         2 Años            Argentin
 7         762704 M        41 Años            Argentin
 8         763097 M        53 Años            Argentin
 9         764087 F        70 Años            Argentin
10         765127 M        30 Años            Argentin
# ... with 182,670 more rows, and 19 more variables:
#   residencia_departamento_nombre <chr>, carga_provin
#   fecha_inicio_sintomas <date>, fecha_apertura <date
#   fecha_internacion <date>, cuidado_intensivo <chr>,
#   fecha_cui_intensivo <lgl>, fallecido <chr>, fecha_
#   asistencia_respiratoria_mecanica <chr>, carga_prov
#   origen_financiamiento <chr>, clasificacion <chr>,
#   clasificacion_resumen <chr>, residencia_provincia_
#   fecha_diagnostico <date>, residencia_departamento_
#   ultima_actualizacion <date>
```

Recodificando con mutate() y case_when()

```
base_covid %>%
  select(id_evento_caso,fallecido, edad)
```

```
# A tibble: 182,680 x 3
   id_evento_caso fallecido  edad
            <dbl> <chr>     <dbl>
 1         748361 NO           23
 2         748780 NO           53
 3         751658 NO           44
 4         755897 NO           29
 5         756503 NO           54
 6         758578 NO            2
 7         762704 NO           41
 8         763097 NO           53
 9         764087 NO           70
10         765127 NO           30
# ... with 182,670 more rows
```

Recodificando con mutate() y case_when()

```
base_covid %>%
  select(id_evento_caso,fallecido, edad) %>%
  mutate(edad_rango = case_when(edad %in% c(0:18) ~  "0 a 18",
                                edad %in% c(19:29) ~ "19 a 29",
                                edad %in% c(30:39) ~ "30 a 39",
                                edad %in% c(40:49) ~ "40 a 49",
                                edad %in% c(50:59) ~ "50 a 59",
                                edad ⩾ 60 ~ "60 o más"))
```

```
# A tibble: 182,680 x 4
   id_evento_caso fallecido  edad edad_rango
            <dbl> <chr>     <dbl> <chr>
 1         748361 NO           23 19 a 29
 2         748780 NO           53 50 a 59
 3         751658 NO           44 40 a 49
 4         755897 NO           29 19 a 29
 5         756503 NO           54 50 a 59
 6         758578 NO            2 0 a 18
 7         762704 NO           41 40 a 49
 8         763097 NO           53 50 a 59
 9         764087 NO           70 60 o más
10         765127 NO           30 30 a 39
# ... with 182,670 more rows
```

PRÁCTICA

Práctica

1) Recodificar la variable **edad**, en 4 rangos.

2) Dadas las siguientes regiones, crear una nueva variable cuyo contenido sean las regiones a la que corresponde cada provincia:

```
patagonia ← c("Neuquén", "Río Negro", "Chubut", "Santa Cruz", "Tierra Del Fuego")
cuyo ← c("Mendoza", "San Juan", "San Luis")
pampeana ← c("Buenos Aires", "La Pampa", "Entre Ríos", "Córdoba", "Santa Fé", "CABA")
noroeste ← c("La Rioja", "Catamarca", "Jujuy", "Tucumán", "Santiago Del Estero", "Salta")
noreste ← c("Formosa", "Chaco", "Misiones", "Corrientes")
```

3) Comprobar que la operación haya sido un éxito.

Práctica

4) Completar los espacios en `_____` de la siguiente sentencia, con el fin de re-codificar la variable numérica **edad** en 4 rangos:

```
base_nueva ← base_covid %>%
  mutate(edad_rango = _____(edad < 1 ~ "_____",
                            edad _____ ~ "Entre 2 y 60 años",
                            edad > _____ "Mayor a 60",
                            _____ ~ "¿y esto?"))
```

summarise()

Resume la información en una nueva tabla

summarise()

```
base_de_datos %>%
        summarise(var1_resumen = sum(var1),
                  var2_media   = mean(var2),
                  var2_desvio  = sd(var2),
                  var2_cv      = var2_desvio / var2_media * 100)
```

summarise()

```
base_covid
```

```
# A tibble: 182,680 x 25
   id_evento_caso sexo   edad edad_años_meses residencia_pais_~ reside
            <dbl> <chr> <dbl> <chr>           <chr>              <chr>
 1         748361 NR       23 Años            Líbano             SIN ES
 2         748780 F        53 Años            Argentina          CABA
 3         751658 M        44 Años            Argentina          CABA
 4         755897 F        29 Años            Argentina          CABA
 5         756503 M        54 Años            Argentina          CABA
 6         758578 M         2 Años            Argentina          CABA
 7         762704 M        41 Años            Argentina          CABA
 8         763097 M        53 Años            Argentina          CABA
 9         764087 F        70 Años            Argentina          CABA
10         765127 M        30 Años            Argentina          CABA
# ... with 182,670 more rows, and 19 more variables:
#   residencia_departamento_nombre <chr>, carga_provincia_nombre <chr>
#   fecha_inicio_sintomas <date>, fecha_apertura <date>, sepi_apertura
#   fecha_internacion <date>, cuidado_intensivo <chr>,
#   fecha_cui_intensivo <lgl>, fallecido <chr>, fecha_fallecimiento <d
#   asistencia_respiratoria_mecanica <chr>, carga_provincia_id <chr>,
#   origen_financiamiento <chr>, clasificacion <chr>,
#   clasificacion_resumen <chr>, residencia_provincia_id <chr>,
#   fecha_diagnostico <date>, residencia_departamento_id <chr>,
#   ultima_actualizacion <date>
```

summarise()

```
base_covid %>%
  select(id_evento_caso, edad)
```

```
# A tibble: 182,680 x 2
   id_evento_caso  edad
            <dbl> <dbl>
 1         748361    23
 2         748780    53
 3         751658    44
 4         755897    29
 5         756503    54
 6         758578     2
 7         762704    41
 8         763097    53
 9         764087    70
10         765127    30
# ... with 182,670 more rows
```

summarise()

```
base_covid %>%
  select(id_evento_caso, edad) %>%
  summarise(min = min(edad, na.rm = TRUE),
            max = max(edad, na.rm = TRUE),
            media = mean(edad, na.rm = TRUE),
            mediana = median(edad, na.rm = TRUE),
            desvio = sd(edad, na.rm = TRUE),
            cv = desvio / media * 100)
```

```
# A tibble: 1 x 6
    min   max media mediana desvio    cv
  <dbl> <dbl> <dbl>   <dbl>  <dbl> <dbl>
1   -12   121  39.1      37   18.1  46.4
```

group_by()

Aplica una operación sobre la población de forma segmentada

group_by()

```
base_de_datos %>%
        group_by(variable_de_corte) #<<
```

group_by()

```
base_covid
```

```
# A tibble: 182,680 x 25
   id_evento_caso sexo   edad edad_años_meses residencia_pais_~ reside
            <dbl> <chr> <dbl> <chr>           <chr>              <chr>
 1         748361 NR       23 Años            Líbano             SIN ES
 2         748780 F        53 Años            Argentina          CABA
 3         751658 M        44 Años            Argentina          CABA
 4         755897 F        29 Años            Argentina          CABA
 5         756503 M        54 Años            Argentina          CABA
 6         758578 M         2 Años            Argentina          CABA
 7         762704 M        41 Años            Argentina          CABA
 8         763097 M        53 Años            Argentina          CABA
 9         764087 F        70 Años            Argentina          CABA
10         765127 M        30 Años            Argentina          CABA
# ... with 182,670 more rows, and 19 more variables:
#   residencia_departamento_nombre <chr>, carga_provincia_nombre <chr>
#   fecha_inicio_sintomas <date>, fecha_apertura <date>, sepi_apertura
#   fecha_internacion <date>, cuidado_intensivo <chr>,
#   fecha_cui_intensivo <lgl>, fallecido <chr>, fecha_fallecimiento <d
#   asistencia_respiratoria_mecanica <chr>, carga_provincia_id <chr>,
#   origen_financiamiento <chr>, clasificacion <chr>,
#   clasificacion_resumen <chr>, residencia_provincia_id <chr>,
#   fecha_diagnostico <date>, residencia_departamento_id <chr>,
#   ultima_actualizacion <date>
```

group_by()

```
base_covid %>%
  group_by(residencia_provincia_nombre)
```

```
# A tibble: 182,680 x 25
# Groups:   residencia_provincia_nombre [25]
   id_evento_caso sexo   edad edad_años_meses residencia_pais_~ reside
            <dbl> <chr> <dbl> <chr>           <chr>             <chr>
 1         748361 NR       23 Años            Líbano            SIN ES
 2         748780 F        53 Años            Argentina         CABA
 3         751658 M        44 Años            Argentina         CABA
 4         755897 F        29 Años            Argentina         CABA
 5         756503 M        54 Años            Argentina         CABA
 6         758578 M         2 Años            Argentina         CABA
 7         762704 M        41 Años            Argentina         CABA
 8         763097 M        53 Años            Argentina         CABA
 9         764087 F        70 Años            Argentina         CABA
10         765127 M        30 Años            Argentina         CABA
# ... with 182,670 more rows, and 19 more variables:
#   residencia_departamento_nombre <chr>, carga_provincia_nombre <chr>
#   fecha_inicio_sintomas <date>, fecha_apertura <date>, sepi_apertura
#   fecha_internacion <date>, cuidado_intensivo <chr>,
#   fecha_cui_intensivo <lgl>, fallecido <chr>, fecha_fallecimiento <d
#   asistencia_respiratoria_mecanica <chr>, carga_provincia_id <chr>,
#   origen_financiamiento <chr>, clasificacion <chr>,
#   clasificacion_resumen <chr>, residencia_provincia_id <chr>,
#   fecha_diagnostico <date>, residencia_departamento_id <chr>,
```

group_by()

```
base_covid %>%
  group_by(residencia_provincia_nombre) %>%
  summarise(min = min(edad, na.rm = TRUE),
            max = max(edad, na.rm = TRUE),
            media = mean(edad, na.rm = TRUE),
            mediana = median(edad, na.rm = TRUE),
            desvio = sd(edad, na.rm = TRUE),
            cv = desvio / media * 100)
```

```
# A tibble: 25 x 7
   residencia_provincia_nombre   min   max media mediana desvio    cv
   <chr>                       <dbl> <dbl> <dbl>   <dbl>  <dbl> <dbl>
 1 Buenos Aires                    0   121  38.6      37   18.2  47.2
 2 CABA                          -12   121  40.3      38   18.8  46.5
 3 Catamarca                       1    94  37.3      36   17.4  46.7
 4 Chaco                           1    97  37.7      35   17.3  46.0
 5 Chubut                          1    97  38.8      36   16.2  41.7
 6 Córdoba                         1   121  39.0      37   18.8  48.1
 7 Corrientes                      1    95  37.7      35   15.4  40.8
 8 Entre Ríos                      1   101  39.0      37   18.5  47.3
 9 Formosa                         1    89  37.2      36   17.1  46.0
10 Jujuy                           1   121  40.7      39   18.1  44.4
# … with 15 more rows
```

summarise()

Caso: Queremos conocer la estructura etárea de las personas residentes en Capital Federal y la Provincia de Buenos Aires, comparando entre aquellas que fallecieron y las que no, y por sexo.

group_by()

```
base_covid
```

```
# A tibble: 182,680 x 25
   id_evento_caso sexo   edad edad_años_meses residenc
            <dbl> <chr> <dbl> <chr>           <chr>
 1         748361 NR       23 Años            Líbano
 2         748780 F        53 Años            Argentin
 3         751658 M        44 Años            Argentin
 4         755897 F        29 Años            Argentin
 5         756503 M        54 Años            Argentin
 6         758578 M         2 Años            Argentin
 7         762704 M        41 Años            Argentin
 8         763097 M        53 Años            Argentin
 9         764087 F        70 Años            Argentin
10         765127 M        30 Años            Argentin
# ... with 182,670 more rows, and 19 more variables:
#   residencia_departamento_nombre <chr>, carga_provin
#   fecha_inicio_sintomas <date>, fecha_apertura <date
#   fecha_internacion <date>, cuidado_intensivo <chr>,
#   fecha_cui_intensivo <lgl>, fallecido <chr>, fecha_
#   asistencia_respiratoria_mecanica <chr>, carga_prov
#   origen_financiamiento <chr>, clasificacion <chr>,
#   clasificacion_resumen <chr>, residencia_provincia_
#   fecha_diagnostico <date>, residencia_departamento_
#   ultima_actualizacion <date>
```

group_by()

```
base_covid %>%
  select(edad,sexo, fallecido,
         res_provincia = residencia_provincia_nombre)
```

```
# A tibble: 182,680 x 4
    edad sexo  fallecido res_provincia
   <dbl> <chr> <chr>     <chr>
 1    23 NR    NO        SIN ESPECIFICAR
 2    53 F     NO        CABA
 3    44 M     NO        CABA
 4    29 F     NO        CABA
 5    54 M     NO        CABA
 6     2 M     NO        CABA
 7    41 M     NO        CABA
 8    53 M     NO        CABA
 9    70 F     NO        CABA
10    30 M     NO        CABA
# ... with 182,670 more rows
```

group_by()

```
base_covid %>%
  select(edad,sexo, fallecido,
         res_provincia = residencia_provincia_nombre) %>%
  mutate(edad_rango = case_when(edad %in% c(0:18) ~  "0 a 18",
                                edad %in% c(19:29) ~ "19 a 29",
                                edad %in% c(30:39) ~ "30 a 39",
                                edad %in% c(40:49) ~ "40 a 49",
                                edad %in% c(50:59) ~ "50 a 59",
                                edad >= 60 ~ "60 o más"),
         sexo = case_when(sexo == "F" ~ "Femenino",
                          sexo == "M" ~ "Masculino",
                          sexo == "NR" ~ "No responde"))
```

```
# A tibble: 182,680 x 5
    edad sexo        fallecido res_provincia   edad_ra
   <dbl> <chr>       <chr>     <chr>           <chr>
 1    23 No responde NO        SIN ESPECIFICAR 19 a 29
 2    53 Femenino    NO        CABA            50 a 59
 3    44 Masculino   NO        CABA            40 a 49
 4    29 Femenino    NO        CABA            19 a 29
 5    54 Masculino   NO        CABA            50 a 59
 6     2 Masculino   NO        CABA            0 a 18
 7    41 Masculino   NO        CABA            40 a 49
 8    53 Masculino   NO        CABA            50 a 59
 9    70 Femenino    NO        CABA            60 o má
10    30 Masculino   NO        CABA            30 a 39
# ... with 182,670 more rows
```

group_by()

```
base_covid %>%
  select(edad,sexo, fallecido,
         res_provincia = residencia_provincia_nombre) %>%
  mutate(edad_rango = case_when(edad %in% c(0:18) ~  "0 a 18",
                                edad %in% c(19:29) ~ "19 a 29",
                                edad %in% c(30:39) ~ "30 a 39",
                                edad %in% c(40:49) ~ "40 a 49",
                                edad %in% c(50:59) ~ "50 a 59",
                                edad >= 60 ~ "60 o más"),
         sexo = case_when(sexo == "F" ~ "Femenino",
                          sexo == "M" ~ "Masculino",
                          sexo == "NR" ~ "No responde")) %>%
  filter(res_provincia %in% c("Buenos Aires",
                              "CABA"))
```

```
# A tibble: 99,818 x 5
    edad sexo      fallecido res_provincia edad_rango
   <dbl> <chr>     <chr>     <chr>          <chr>
 1    53 Femenino  NO        CABA           50 a 59
 2    44 Masculino NO        CABA           40 a 49
 3    29 Femenino  NO        CABA           19 a 29
 4    54 Masculino NO        CABA           50 a 59
 5     2 Masculino NO        CABA           0 a 18
 6    41 Masculino NO        CABA           40 a 49
 7    53 Masculino NO        CABA           50 a 59
 8    70 Femenino  NO        CABA           60 o más
 9    30 Masculino NO        CABA           30 a 39
10    28 Masculino NO        CABA           19 a 29
# ... with 99,808 more rows
```

group_by()

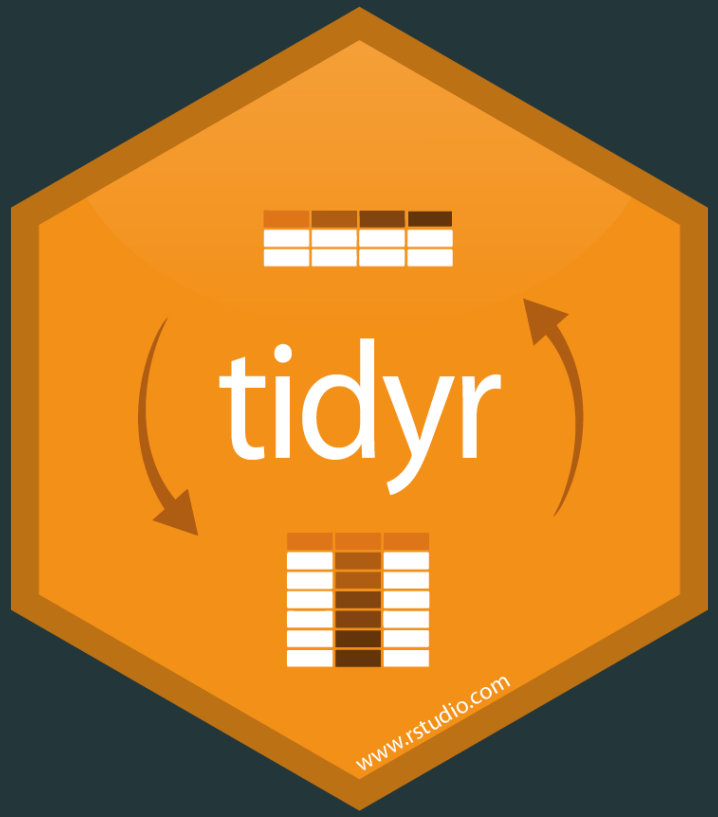
```
base_covid %>%
  select(edad,sexo, fallecido,
         res_provincia = residencia_provincia_nombre) %>%
  mutate(edad_rango = case_when(edad %in% c(0:18) ~  "0 a 18",
                                edad %in% c(19:29) ~ "19 a 29",
                                edad %in% c(30:39) ~ "30 a 39",
                                edad %in% c(40:49) ~ "40 a 49",
                                edad %in% c(50:59) ~ "50 a 59",
                                edad >= 60 ~ "60 o más"),
         sexo = case_when(sexo == "F" ~ "Femenino",
                          sexo == "M" ~ "Masculino",
                          sexo == "NR" ~ "No responde")) %>%
  filter(res_provincia %in% c("Buenos Aires",
                              "CABA")) %>%
  group_by(sexo, edad_rango)
```

```
# A tibble: 99,818 x 5
# Groups:   sexo, edad_rango [21]
    edad sexo      fallecido res_provincia edad_rango
   <dbl> <chr>     <chr>     <chr>         <chr>
 1    53 Femenino  NO        CABA          50 a 59
 2    44 Masculino NO        CABA          40 a 49
 3    29 Femenino  NO        CABA          19 a 29
 4    54 Masculino NO        CABA          50 a 59
 5     2 Masculino NO        CABA          0 a 18
 6    41 Masculino NO        CABA          40 a 49
 7    53 Masculino NO        CABA          50 a 59
 8    70 Femenino  NO        CABA          60 o más
 9    30 Masculino NO        CABA          30 a 39
10    28 Masculino NO        CABA          19 a 29
# ... with 99,808 more rows
```

group_by()

```
base_covid %>%
  select(edad,sexo, fallecido,
         res_provincia = residencia_provincia_nombre) %>%
  mutate(edad_rango = case_when(edad %in% c(0:18) ~  "0 a 18",
                                edad %in% c(19:29) ~ "19 a 29",
                                edad %in% c(30:39) ~ "30 a 39",
                                edad %in% c(40:49) ~ "40 a 49",
                                edad %in% c(50:59) ~ "50 a 59",
                                edad >= 60 ~ "60 o más"),
         sexo = case_when(sexo == "F" ~ "Femenino",
                          sexo == "M" ~ "Masculino",
                          sexo == "NR" ~ "No responde")) %>%
  filter(res_provincia %in% c("Buenos Aires",
                              "CABA")) %>%
  group_by(sexo, edad_rango) %>%
  summarise(min = min(edad, na.rm = TRUE),
            max = max(edad, na.rm = TRUE),
            media = mean(edad, na.rm = TRUE),
            mediana = median(edad, na.rm = TRUE),
            desvio = sd(edad, na.rm = TRUE),
            cv = desvio / media * 100)
```

```
# A tibble: 21 x 8
# Groups:   sexo [3]
   sexo      edad_rango   min   max media mediana desv
   <chr>     <chr>      <dbl> <dbl> <dbl>   <dbl>  <db
 1 Femenino  0 a 18         1    18  11.2      13   5.
 2 Femenino  19 a 29       19    29  24.6      25   3.
 3 Femenino  30 a 39       30    39  34.3      34   2.
 4 Femenino  40 a 49       40    49  44.3      44   2.
 5 Femenino  50 a 59       50    59  54.2      54   2.
 6 Femenino  60 o más      60   120  71.7      69   9.
 7 Femenino  <NA>         -12    -8 -10      -10   2.
 8 Masculino 0 a 18         0    18  10.4      11   5.
 9 Masculino 19 a 29       19    29  24.6      25   3.
10 Masculino 30 a 39       30    39  34.4      34   2.
# ... with 11 more rows
```


tidyr
www.rstudio.com

Funciones del paquete tidyr:

Función	Acción
`pivot_longer()`	*Transforma en filas varias columnas*
`pivot_wider()`	*transforma en columnas varias filas*

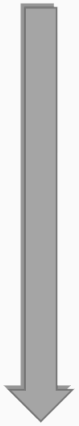
estructura de datos

A lo Ancho

id	provincia	edad_media	edad_desvio	edad_cv
1	BsAs	38	18	47
2	CABA	40	18	46

A lo Largo

id	provincia	Variable	valor
1	BsAs	edad_media	38
2	CABA	edad_media	40
1	BsAs	edad_desvio	18
2	CABA	edad_desvio	18
1	BsAs	edad_cv	47
2	CABA	edad_cv	46



pivot_longer()

Reestructura la base, apilando varias columnas en una. De ancho a largo

pivot_longer()

```
base_covid
```

```
# A tibble: 182,680 x 25
   id_evento_caso sexo   edad edad_años_meses residencia_pais_~ reside
            <dbl> <chr> <dbl> <chr>           <chr>              <chr>
 1         748361 NR       23 Años            Líbano             SIN ES
 2         748780 F        53 Años            Argentina          CABA
 3         751658 M        44 Años            Argentina          CABA
 4         755897 F        29 Años            Argentina          CABA
 5         756503 M        54 Años            Argentina          CABA
 6         758578 M         2 Años            Argentina          CABA
 7         762704 M        41 Años            Argentina          CABA
 8         763097 M        53 Años            Argentina          CABA
 9         764087 F        70 Años            Argentina          CABA
10         765127 M        30 Años            Argentina          CABA
# ... with 182,670 more rows, and 19 more variables:
#   residencia_departamento_nombre <chr>, carga_provincia_nombre <chr>
#   fecha_inicio_sintomas <date>, fecha_apertura <date>, sepi_apertura
#   fecha_internacion <date>, cuidado_intensivo <chr>,
#   fecha_cui_intensivo <lgl>, fallecido <chr>, fecha_fallecimiento <d
#   asistencia_respiratoria_mecanica <chr>, carga_provincia_id <chr>,
#   origen_financiamiento <chr>, clasificacion <chr>,
#   clasificacion_resumen <chr>, residencia_provincia_id <chr>,
#   fecha_diagnostico <date>, residencia_departamento_id <chr>,
#   ultima_actualizacion <date>
```

pivot_longer()

```
base_covid %>%
  group_by(residencia_provincia_nombre)
```

```
# A tibble: 182,680 x 25
# Groups:   residencia_provincia_nombre [25]
   id_evento_caso sexo   edad edad_años_meses residencia_pais_~ reside
            <dbl> <chr> <dbl> <chr>           <chr>              <chr>
 1         748361 NR       23 Años            Líbano             SIN ES
 2         748780 F        53 Años            Argentina          CABA
 3         751658 M        44 Años            Argentina          CABA
 4         755897 F        29 Años            Argentina          CABA
 5         756503 M        54 Años            Argentina          CABA
 6         758578 M         2 Años            Argentina          CABA
 7         762704 M        41 Años            Argentina          CABA
 8         763097 M        53 Años            Argentina          CABA
 9         764087 F        70 Años            Argentina          CABA
10         765127 M        30 Años            Argentina          CABA
# ... with 182,670 more rows, and 19 more variables:
#   residencia_departamento_nombre <chr>, carga_provincia_nombre <chr>
#   fecha_inicio_sintomas <date>, fecha_apertura <date>, sepi_apertura
#   fecha_internacion <date>, cuidado_intensivo <chr>,
#   fecha_cui_intensivo <lgl>, fallecido <chr>, fecha_fallecimiento <d
#   asistencia_respiratoria_mecanica <chr>, carga_provincia_id <chr>,
#   origen_financiamiento <chr>, clasificacion <chr>,
#   clasificacion_resumen <chr>, residencia_provincia_id <chr>,
#   fecha_diagnostico <date>, residencia_departamento_id <chr>,
```

pivot_longer()

```
base_covid %>%
  group_by(residencia_provincia_nombre) %>%
  summarise(min = min(edad, na.rm = TRUE),
            max = max(edad, na.rm = TRUE),
            media = mean(edad, na.rm = TRUE),
            mediana = median(edad, na.rm = TRUE),
            desvio = sd(edad, na.rm = TRUE),
            cv = desvio / media * 100)
```

```
# A tibble: 25 x 7
   residencia_provincia_nombre   min   max media mediana desvio    cv
   <chr>                       <dbl> <dbl> <dbl>   <dbl>  <dbl> <dbl>
 1 Buenos Aires                    0   121  38.6      37   18.2  47.2
 2 CABA                          -12   121  40.3      38   18.8  46.5
 3 Catamarca                       1    94  37.3      36   17.4  46.7
 4 Chaco                           1    97  37.7      35   17.3  46.0
 5 Chubut                          1    97  38.8      36   16.2  41.7
 6 Córdoba                         1   121  39.0      37   18.8  48.1
 7 Corrientes                      1    95  37.7      35   15.4  40.8
 8 Entre Ríos                      1   101  39.0      37   18.5  47.3
 9 Formosa                         1    89  37.2      36   17.1  46.0
10 Jujuy                           1   121  40.7      39   18.1  44.4
# … with 15 more rows
```

pivot_longer()

```
base_covid %>%
  group_by(residencia_provincia_nombre) %>%
  summarise(min = min(edad, na.rm = TRUE),
            max = max(edad, na.rm = TRUE),
            media = mean(edad, na.rm = TRUE),
            mediana = median(edad, na.rm = TRUE),
            desvio = sd(edad, na.rm = TRUE),
            cv = desvio / media * 100) %>%
  select(residencia_provincia_nombre,
         media, mediana, desvio)
```

```
# A tibble: 25 x 4
   residencia_provincia_nombre media mediana desvio
   <chr>                         <dbl>   <dbl>  <dbl>
 1 Buenos Aires                   38.6      37   18.2
 2 CABA                           40.3      38   18.8
 3 Catamarca                      37.3      36   17.4
 4 Chaco                          37.7      35   17.3
 5 Chubut                         38.8      36   16.2
 6 Córdoba                        39.0      37   18.8
 7 Corrientes                     37.7      35   15.4
 8 Entre Ríos                     39.0      37   18.5
 9 Formosa                        37.2      36   17.1
10 Jujuy                          40.7      39   18.1
# … with 15 more rows
```

pivot_longer()

```
base_covid %>%
  group_by(residencia_provincia_nombre) %>%
  summarise(min = min(edad, na.rm = TRUE),
            max = max(edad, na.rm = TRUE),
            media = mean(edad, na.rm = TRUE),
            mediana = median(edad, na.rm = TRUE),
            desvio = sd(edad, na.rm = TRUE),
            cv = desvio / media * 100) %>%
  select(residencia_provincia_nombre,
         media, mediana, desvio) %>%
  pivot_longer(cols = c(media, mediana, desvio),
               names_to = "variable",
               values_to = "valor")
```

```
# A tibble: 75 x 3
   residencia_provincia_nombre variable valor
   <chr>                       <chr>    <dbl>
 1 Buenos Aires                media     38.6
 2 Buenos Aires                mediana   37
 3 Buenos Aires                desvio    18.2
 4 CABA                        media     40.3
 5 CABA                        mediana   38
 6 CABA                        desvio    18.8
 7 Catamarca                   media     37.3
 8 Catamarca                   mediana   36
 9 Catamarca                   desvio    17.4
10 Chaco                       media     37.7
# ... with 65 more rows
```

pivot_wider()

Reestructura la base, encolumnando varias filas de una variable. De largo a ancho

pivot_wider()

```
base_largo
```

```
# A tibble: 75 x 3
   residencia_provincia_nombre variable valor
   <chr>                       <chr>    <dbl>
 1 Buenos Aires                media     38.6
 2 Buenos Aires                mediana   37  
 3 Buenos Aires                desvio    18.2
 4 CABA                        media     40.3
 5 CABA                        mediana   38  
 6 CABA                        desvio    18.8
 7 Catamarca                   media     37.3
 8 Catamarca                   mediana   36  
 9 Catamarca                   desvio    17.4
10 Chaco                       media     37.7
# ... with 65 more rows
```

pivot_wider()

```
base_largo %>%
  pivot_wider(names_from = "variable",  #<<
              values_from = "valor")
```

```
# A tibble: 25 x 4
   residencia_provincia_nombre media mediana desvio
   <chr>                         <dbl>   <dbl>  <dbl>
 1 Buenos Aires                   38.6      37   18.2
 2 CABA                           40.3      38   18.8
 3 Catamarca                      37.3      36   17.4
 4 Chaco                          37.7      35   17.3
 5 Chubut                         38.8      36   16.2
 6 Córdoba                        39.0      37   18.8
 7 Corrientes                     37.7      35   15.4
 8 Entre Ríos                     39.0      37   18.5
 9 Formosa                        37.2      36   17.1
10 Jujuy                          40.7      39   18.1
# ... with 15 more rows
```