Módulo 2 — Tidyverse I

select() y filter()

CEPAL - Unidad de Estadísticas Sociales

2025-10-30

Introducción

Enfoque de esta sección

- Manipulación de datos con `dplyr`.
- Uso de **`select()`** para elegir variables.
- Uso de **`filter()`** para seleccionar observaciones.
- Ambas funciones se combinan frecuentemente con el operador `%>%`.

Lectura de base de ejemplos

```
datos <- readRDS("data/base_personas_gasto.rds")
library(dplyr)
glimpse(datos)
```

```
Rows: 19,427
Columns: 17
$ id_hogar   <dbl> 262, 262, 265, 265, 265, 277, 277, 277, 277, 288, 288, 1
$ id_pers    <dbl> 1, 2, 1, 2, 3, 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 5, 1
$ upm        <dbl> 1100100006, 1100100006, 1100100006, 1100100006, 11001000
$ estrato    <dbl> 11001, 11001, 11001, 11001, 11001, 11001, 11001, 11001,
$ area       <chr> "1", "1", "1", "1", "1", "1", "1", "1", "1", "1", "1",
$ fep        <dbl> 19, 19, 16, 16, 16, 16, 16, 16, 16, 19, 19, 19, 19, 30,
$ pobreza    <chr> "3", "3", "3", "3", "3", "3", "3", "3", "3", "3", "3",
$ ingreso_hh <dbl> 10783.053, 10783.053, 21259.723, 21259.723, 21259.723, 9
$ gasto_hh   <dbl> 10783.053, 10783.053, 21259.723, 21259.723, 21259.723, 9
$ parentesco <chr> "1", "2", "1", "2", "3", "1", "3", "3", "3", "1", "2",
$ edad       <dbl> 51, 46, 26, 24, 7, 42, 20, 17, 13, 60, 32, 13, 5, 39, 3
$ sexo       <fct> Hombre, Mujer, Mujer, Hombre, Hombre, Mujer, Mujer, Homb
$ etnia      <fct> 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
```

`select()`

¿Para qué sirve?

Permite **seleccionar columnas** (variables) de un conjunto de datos.

```
temp <- select(datos, sexo, parentesco, edad)
head(temp)
```

sexo	parentesco	edad
Hombre	1	51
Mujer	2	46
Mujer	1	26
Hombre	2	24
Hombre	3	7
Mujer	1	42

Devuelve solo las variables indicadas.

Selección por posición

```
temp <- select(datos, 1:3)
head(temp)
```

id_hogar	id_pers	upm
262	1	1100100006
262	2	1100100006
265	1	1100100006
265	2	1100100006
265	3	1100100006
277	1	1100100006

Selección por rango

```
temp <-  select(datos, parentesco:anoest)
head(temp, 10)
```

parentesco	edad	sexo	etnia	anoest
1	51	Hombre	0	18
2	46	Mujer	0	12
1	26	Mujer	0	17
2	24	Hombre	0	15
3	7	Hombre	1	0
1	42	Mujer	0	15
3	20	Mujer	1	13
3	17	Hombre	1	8
3	13	Mujer	1	5
1	60	Hombre	0	10

Selección por exclusión

```
temp <-  select(datos,-c("id_hogar",  "id_pers","upm",  "estrato",
                  "area", "fep",  "pobreza",
                  "ingreso_hh","gasto_hh")) # excluye variables
head(temp, 5)
```

parentesco	edad	sexo	etnia	anoest	niveduc_ee	ingreso	gasto
1	51	Hombre	0	18	7	5391.527	5391.527
2	46	Mujer	0	12	5	5391.527	5391.527
1	26	Mujer	0	17	7	7077.083	7077.083
2	24	Hombre	0	15	6	7105.557	7105.557
3	7	Hombre	1	0	1	7077.083	7077.083

`select()` acepta rangos y exclusiones, lo que permite construir subconjuntos rápidamente.

Selectores auxiliares

```
select(datos,
       starts_with("e"),    # variables que inician con "e"
       ends_with("o"),      # variables que terminan con "o"
       contains("ed"),      # variables que contienen "ed"
       everything())        # todas las variables
```

*Los **selectores** permiten patrones más flexibles, útiles con grandes bases.*

Selectores auxiliares: variables que inician con “e”

```
select(datos, starts_with("e")) %>% head()
```

estrato	edad	etnia
11001	51	0
11001	46	0
11001	26	0
11001	24	0
11001	7	1
11001	42	0

Selectores auxiliares: variables que terminan con “o”

```
select(datos, ends_with("o")) %>% head()
```

estrato	parentesco	sexo	ingreso	gasto
11001	1	Hombre	5391.527	5391.527
11001	2	Mujer	5391.527	5391.527
11001	1	Mujer	7077.083	7077.083
11001	2	Hombre	7105.557	7105.557
11001	3	Hombre	7077.083	7077.083
11001	1	Mujer	2418.750	2418.750

Selectores auxiliares: variables que contienen “ed”

```
select(datos, contains("ed")) %>% head()
```

edad	niveduc_ee
51	7
46	5
26	7
24	6
7	1
42	6

Renombrar variables al seleccionar

Puedes renombrar variables dentro de `select()`, sin usar `rename()` aparte.

```
select(datos,
       Nivel_Educativo = niveduc_ee,
       Años_estudio = anoest,
       Sexo = sexo
       ) %>%
       head()
```

Nivel_Educativo	Años_estudio	Sexo
7	18	Hombre
5	12	Mujer
7	17	Mujer
6	15	Hombre
1	0	Hombre
6	15	Mujer

filter()

¿Para qué sirve?

*Selecciona **filas (observaciones)** que cumplen condiciones lógicas.*

```
filter(datos, parentesco == 1) %>% head()
```

id_hogar	id_pers	upm	estrato	area	fep	pobreza	ingreso_hh	gasto_hh	parentesco	edad	sexo	etnia	anoest	niveduc_e	ingreso	gasto
262	1	1100100006	1001	1	19	3	10783.053	10783.053	1	51	Hombre	0	18	7	5391.527	5391.527
265	1	1100100006	1001	1	16	3	21259.723	21259.723	1	26	Mujer	0	17	7	7077.083	7077.083
277	1	1100100006	1001	1	16	3	9618.053	9618.053	1	42	Mujer	0	15	6	2418.750	2418.750
288	1	1100100006	1001	1	19	3	5414.580	5414.580	1	60	Hombre	0	10	4	1375.000	1375.000
289	1	1100100006	1001	1	30	3	7807.633	7807.633	1	39	Hombre	0	12	5	1561.527	1561.527
291	1	1100100006	1001	1	28	3	7337.633	7337.633	1	43	Hombre	0	12	5	1467.527	1467.527
295	1	1100100006	1001	1	36	3	50214.580	47566.942	1	71	Mujer	0	11	4	12546.527	12546.527
296	1	1100100006	1001	1	21	3	7971.527	7971.527	1	67	Hombre	0	10	4	4000.000	4000.000
298	1	1100100006	1001	1	24	3	7261.040	7261.040	1	47	Mujer	1	12	5	2420.347	2420.347
302	1	1100100006	1001	1	26	3	14717.796	14717.796	1	61	Mujer	0	12	5	2943.559	2943.559

Devuelve solo los jefes de hogar.

Condiciones múltiples

```
filter(datos, parentesco == 1, ingreso > 7100) %>%
  select(id_hogar, parentesco,
         ingreso, sexo) %>% head(10)
```

id_hogar	parentesco	ingreso	sexo
295	1	12546.527	Mujer
359	1	14729.860	Mujer
362	1	7521.527	Hombre
406	1	8971.527	Hombre
495	1	8703.053	Mujer
495	1	8760.000	Mujer
549	1	19783.335	Hombre
555	1	7583.330	Hombre
557	1	8258.330	Hombre
572	1	7164.772	Hombre

Operadores lógicos más usados

Operador	Significado	Ejemplo
==	Igual	`parentesco == 1`
!=	Diferente	`pobreza != 3`
> / <	Mayor / Menor	`edad > 17`
>= / <=	Mayor/igual / Menor/igual	`anoest >= 12`
%in%	Pertenencia	`edad %in% c(1:10)`
& / |	Y / O lógicos	`edad > 17 & anoest > 20`

Uso de operadores de pertenencia

```
datos %>%  filter(edad %in% c(1:10)) %>%
  select(id_hogar, parentesco, ingreso, sexo, edad) %>%
  head()
```

id_hogar	parentesco	ingreso	sexo	edad
265	3	7077.083	Hombre	7
288	3	1346.527	Mujer	5
289	3	1561.527	Mujer	5
289	3	1561.527	Mujer	3
291	3	1467.527	Mujer	10
302	4	2943.559	Hombre	8

Filtra solo las observaciones con edad de 1 a 10 años.

Flujo de trabajo visual

1. `select()` → selecciona variables
2. `filter()` → aplica condiciones
3. Resultado: subconjunto depurado y legible

Piensa en estas funciones como una "lupa": una observa **qué columnas**, la otra **qué filas**.

Variantes útiles

select()

- `rename_with(fn, .cols)` → aplicar función a nombres
- `relocate(col, .before/.after)` → cambiar orden de columnas

filter()

- `filter(.data, between(var, a, b))`
- `filter(.data, near(var, value, tol))`

```
filter(datos, between(edad, 1, 10))
```

En resumen

Función	Actúa sobre	Objetivo principal	Ejemplo
`select()`	Columnas	Reducir o reorganizar variables	`select(datos, edad, sexo)`
`filter()`	Filas	Seleccionar observaciones específicas	`filter(datos, edad >= 18)`

Combinadas, son la base de todo flujo de transformación en el tidyverse.