

Clase_1__Alan__Amolef

Alan Amolef

2025-05-29

Ayudantía 1: R-studio.

A continuación se presenta el trabajo sobre la base de datos.

- 1.- Ver las primeras 6 filas de la base de datos original.
- 2.- Cambiarle el nombre a la base de datos, recomendación: db; df.
- 3.- ¿cuántas filas tiene la base de datos original?
- 4.- ¿cuántas columnas tiene la base de datos original?
- 5.- Mencione el nombre de las columnas de la base de datos original.
- 6.- Seleccione las columnas: “region”, “edad”, “sexo”, “estudia_actual” y nombre esa nueva base de datos de la siguiente manera: nombre_apellido (usando su nombre y su apellido como aparece en U-cusos.)
- 7.- ¿Qué tipo de variables son?
- 8.- ¿Qué tipo de variables deberían ser?
- 9.- Corrija las variables y dele el formato correcto.
- 10.- Ver datos faltantes o errneos.
- 11.- Gráficos de la columna región, edad, sexo, boxplot de edad con sexo.
- 12.- Test de medias: Las medias de edad entre hombres (H) y mujeres (M) son iguales.

Importamos la base de datos

```
library(readr) #librería para leer archivo
```

```
ene.2023.06.mjj <- read.csv2("/cloud/project/Clase_1/ene-2023-06-mjj.csv", header=TRUE) #importamos el
```

Vemos las primeras 6 filas

```
head(ene.2023.06.mjj, 6) # vemos las 6 primeras filas
```

```
##   ano_trimestre mes_central ano_encuesta mes_encuesta region provincia r_p_c
## 1      2023         6      2023         5      4         41  4101
## 2      2023         6      2023         5      6         61  6117
## 3      2023         6      2023         5      3         32  3201
## 4      2023         6      2023         5     15        151 15101
## 5      2023         6      2023         5     11        111 11101
## 6      2023         6      2023         5     13        132 13202
##   estrato tipo conglomerado id_identificacion hogar   idrph nro_linea edad
## 1  4100113   1      13990      123820         1  240906      3      8
## 2  6100310   1      25530      107004         1  604902      1     76
## 3  3100220   3      24403      121763         1  680083      1     67
## 4 15000012   1      43796      113405         1  970777      3     20
```

## 5	11100110	1	26749		113849	1	1023056	4	8									
## 6	13200210	1	34003		159057	1	1035050	4	6									
##	tramo_edad	sexo	parentesco	curso	nivel	termino_nivel	estudia_actual	cine										
## 1	NA	2	4	1	3	2		1	3									
## 2	12	1	1	2	6	2		2	4									
## 3	11	1	1	0	0	2		2	1									
## 4	2	1	4	4	4	1		2	5									
## 5	NA	2	4	2	3	2		1	3									
## 6	NA	1	4	1	2	1		2	2									
##	est_conyugal	proveedor	orig1	orig2	orig3	nacionalidad	mig1	mig2_cod	mig3_cod									
## 1	0	0	2	NA		152	1	NA	NA									
## 2	1	1	2	NA		152	1	NA	NA									
## 3	3	1	2	NA		152	1	NA	NA									
## 4	3	0	1	2		68	3	NA	68									
## 5	0	0	2	NA		152	1	NA	NA									
## 6	0	0	2	NA		152	1	NA	NA									
##	mig4	mig5_cod	mig6_cod	a1	a2	a3	a4	a5	a6	a6_otro	a7	a8	b1	b2	b3	b4	b5	b6
## 1	1	NA	NA	NA	NA	NA	NA	NA	NA		NA	NA	NA	NA	NA	NA	NA	NA
## 2	2	7101	NA	2	2	NA	NA	2	NA		NA	NA	NA	NA	NA	NA	NA	NA
## 3	2	4101	NA	2	2	NA	NA	2	NA		NA	NA	NA	NA	NA	NA	NA	NA
## 4	3	NA	68	2	2	NA	NA	2	NA		NA	NA	NA	NA	NA	NA	NA	NA
## 5	1	NA	NA	NA	NA	NA	NA	NA	NA		NA	NA	NA	NA	NA	NA	NA	NA
## 6	1	NA	NA	NA	NA	NA	NA	NA	NA		NA	NA	NA	NA	NA	NA	NA	NA
##	b7a_1	b7a_2	b7a_3	b7b_1	b7b_2	b7b_3	b7b_4	b8	b9	b10	b11_proxy	b12						
## 1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA		NA	NA					
## 2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA		NA	NA					
## 3	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA		NA	NA					
## 4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA		NA	NA					
## 5	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA		NA	NA					
## 6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA		NA	NA					
##	b13_rev4cl_caenes	i1	i2	i3	i3_v	i4	i5	i6	i7	f7a	f7b	f7c	f7d	f7e				
## 1		NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA				
## 2		NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA				
## 3		NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA				
## 4		NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA				
## 5		NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA				
## 6		NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA				
##	b14_rev4cl_caenes	b15_1	b15_2	b16	b16_otro	b17_mes	b17_ano	b18_codigo										
## 1		NA	NA	NA	NA			NA	NA									
## 2		NA	NA	NA	NA			NA	NA									
## 3		NA	NA	NA	NA			NA	NA									
## 4		NA	NA	NA	NA			NA	NA									
## 5		NA	NA	NA	NA			NA	NA									
## 6		NA	NA	NA	NA			NA	NA									
##	b18_region	b18_varias	ocup_honorarios	plataformas_digitales	pd_especifique													
## 1	NA		NA		NA			NA										
## 2	NA		NA		NA			NA										
## 3	NA		NA		NA			NA										
## 4	NA		NA		NA			NA										
## 5	NA		NA		NA			NA										
## 6	NA		NA		NA			NA										
##	b19_dependencia_segunda	sda_pd	sda_pd_especifique	turno	turno_d	turno_de												
## 1	NA		NA		NA			NA										
## 2	NA		NA		NA			NA										

##	3	NA			NA	NA				NA	NA	NA			
##	4	NA			NA	NA				NA	NA	NA			
##	5	NA			NA	NA				NA	NA	NA			
##	6	NA			NA	NA				NA	NA	NA			
##		turno_h	turno_t	c2_1_1	c2_1_2	c2_1_3	c2_2_1	c2_2_2	c2_2_3	c3_1	c3_2	c3_3			
##	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA			
##	2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA			
##	3	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA			
##	4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA			
##	5	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA			
##	6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA			
##		turno_cont_d	turno_cont_de	turno_cont_h	turno_cont_t	c4	c5	c6	c7	c8	c9				
##	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA			
##	2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA			
##	3	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA			
##	4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA			
##	5	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA			
##	6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA			
##		c9_otro	adicionales_h	adicionales_d	adicionales_t	c10	c11	c12	e2	e3_1	e3_2				
##	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA			
##	2	NA	NA	NA	NA	NA	NA	NA	2	NA	NA	NA			
##	3	NA	NA	NA	NA	NA	NA	NA	2	NA	NA	NA			
##	4	NA	NA	NA	NA	NA	NA	NA	1	0	0	NA			
##	5	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA			
##	6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA			
##		e3_3	e3_4	e3_5	e3_6	e3_7	e3_8	e3_9	e3_10	e3_11	e3_12	e3_total	e4	e5	e5_dia
##	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
##	2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
##	3	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
##	4	1	0	0	0	0	0	0	0	NA	1	NA	1	22	NA
##	5	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
##	6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
##		e5_sem	e5_mes	e5_ano	e6_mes	e6_ano	e7	e9	e9_otro	e10	deseo_trabajar	e11	e12		
##	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
##	2	NA	NA	NA	NA	NA	NA	8	NA	NA	NA	2	2	6	NA
##	3	NA	NA	NA	NA	NA	NA	8	NA	NA	NA	2	2	8	NA
##	4	4	5	2023	1	2023	1	NA	NA	NA	NA	NA	1	NA	NA
##	5	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
##	6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
##		e12_otro	e13	e21_mes	e21_ano	e21_tramo	e22	e23	e23_otro	e24	e24_otro	activ			
##	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
##	2	1	6	2019	3	4	NA	NA	NA	NA	NA	3	NA	NA	NA
##	3	1	3	2020	1	4	NA	NA	NA	NA	NA	3	NA	NA	NA
##	4	1	12	2022	1	3	NA	NA	NA	NA	NA	2	NA	NA	NA
##	5	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
##	6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
##		obe	tpi	id	ftp	cae_general	cae_especifico	categoria_ocupacion	habituales						
##	1	NA	NA	NA	NA	0	0	0	0	0	0	NA	NA	NA	NA
##	2	NA	NA	0	0	9	21	0	0	0	0	NA	NA	NA	NA
##	3	NA	NA	0	0	9	21	0	0	0	0	NA	NA	NA	NA
##	4	NA	NA	NA	NA	4	8	0	0	0	0	NA	NA	NA	NA
##	5	NA	NA	NA	NA	0	0	0	0	0	0	NA	NA	NA	NA
##	6	NA	NA	NA	NA	0	0	0	0	0	0	NA	NA	NA	NA
##		efectivas	ocup_form	r_p_rev4cl_caenes	sector	asocia	fact_cal								

```
## 1      NA      NA      NA      NA      NA 127.32695
## 2      NA      NA      NA      NA      NA 161.89138
## 3      NA      NA      NA      NA      NA 27.24410
## 4      NA      NA      NA      NA      NA 91.42452
## 5      NA      NA      NA      NA      NA 41.99031
## 6      NA      NA      NA      NA      NA 488.01655
```

Cambio de nombre a la base de datos, pasa a llamarse db

```
db = ene.2023.06.mjj # cambiamos el nombre a db
```

```
head(db,5)
```

```
##      ano_trimestre mes_central ano_encuesta mes_encuesta region provincia r_p_c
## 1      2023      6      2023      5      4      41 4101
## 2      2023      6      2023      5      6      61 6117
## 3      2023      6      2023      5      3      32 3201
## 4      2023      6      2023      5     15     151 15101
## 5      2023      6      2023      5     11     111 11101
##      estrato tipo conglomerado id_identificacion hogar idrph nro_linea edad
## 1 4100113      1      13990      123820      1 240906      3      8
## 2 6100310      1      25530      107004      1 604902      1     76
## 3 3100220      3      24403      121763      1 680083      1     67
## 4 15000012      1      43796      113405      1 970777      3     20
## 5 11100110      1      26749      113849      1 1023056      4     8
##      tramo_edad sexo parentesco curso nivel termino_nivel estudia_actual cine
## 1      NA      2      4      1      3      2      1      3
## 2      12      1      1      2      6      2      2      4
## 3      11      1      1      0      0      2      2      1
## 4      2      1      4      4      4      1      2      5
## 5      NA      2      4      2      3      2      1      3
##      est_conyugal proveedor orig1 orig2 orig3 nacionalidad mig1 mig2_cod mig3_cod
## 1      0      0      2      NA      152      1      NA      NA
## 2      1      1      2      NA      152      1      NA      NA
## 3      3      1      2      NA      152      1      NA      NA
## 4      3      0      1      2      68      3      NA      68
## 5      0      0      2      NA      152      1      NA      NA
##      mig4 mig5_cod mig6_cod a1 a2 a3 a4 a5 a6 a6_otro a7 a8 b1 b2 b3 b4 b5 b6
## 1      1      NA      NA NA NA NA NA NA NA      NA NA NA NA NA NA NA NA
## 2      2      7101      NA 2 2 NA NA 2 NA      NA NA NA NA NA NA NA NA
## 3      2      4101      NA 2 2 NA NA 2 NA      NA NA NA NA NA NA NA NA
## 4      3      NA      68 2 2 NA NA 2 NA      NA NA NA NA NA NA NA NA
## 5      1      NA      NA NA NA NA NA NA NA      NA NA NA NA NA NA NA NA
##      b7a_1 b7a_2 b7a_3 b7b_1 b7b_2 b7b_3 b7b_4 b8 b9 b10 b11_proxy b12
## 1      NA      NA      NA      NA      NA      NA      NA NA NA NA      NA NA
## 2      NA      NA      NA      NA      NA      NA      NA NA NA NA      NA NA
## 3      NA      NA      NA      NA      NA      NA      NA NA NA NA      NA NA
## 4      NA      NA      NA      NA      NA      NA      NA NA NA NA      NA NA
## 5      NA      NA      NA      NA      NA      NA      NA NA NA NA      NA NA
##      b13_rev4cl_caenes i1 i2 i3 i3_v i4 i5 i6 i7 f7a f7b f7c f7d f7e
## 1      NA NA NA NA      NA NA NA NA NA NA      NA NA NA NA
## 2      NA NA NA NA      NA NA NA NA NA NA      NA NA NA NA
## 3      NA NA NA NA      NA NA NA NA NA NA      NA NA NA NA
## 4      NA NA NA NA      NA NA NA NA NA NA      NA NA NA NA
## 5      NA NA NA NA      NA NA NA NA NA NA      NA NA NA NA
```



```
## obe tpi id ftp cae_general cae_especifico categoria_ocupacion habituales
## 1 NA NA NA NA 0 0 0 NA
## 2 NA NA 0 0 9 21 0 NA
## 3 NA NA 0 0 9 21 0 NA
## 4 NA NA NA NA 4 8 0 NA
## 5 NA NA NA NA 0 0 0 NA
## efectivas ocup_form r_p_rev4cl_caenes sector asocia fact_cal
## 1 NA NA NA NA NA 127.32695
## 2 NA NA NA NA NA 161.89138
## 3 NA NA NA NA NA 27.24410
## 4 NA NA NA NA NA 91.42452
## 5 NA NA NA NA NA 41.99031
```

Cantidad de filas

```
nrow(db)
```

```
## [1] 101592
```

Cantidad de columnas

```
ncol(db)
```

```
## [1] 178
```

Nombre de las columnas

```
colnames(db)
```

```
## [1] "ano_trimestre" "mes_central" "ano_encuesta"
## [4] "mes_encuesta" "region" "provincia"
## [7] "r_p_c" "estrato" "tipo"
## [10] "conglomerado" "id_identificacion" "hogar"
## [13] "idrph" "nro_linea" "edad"
## [16] "tramo_edad" "sexo" "parentesco"
## [19] "curso" "nivel" "termino_nivel"
## [22] "estudia_actual" "cine" "est_conyugal"
## [25] "proveedor" "orig1" "orig2"
## [28] "orig3" "nacionalidad" "mig1"
## [31] "mig2_cod" "mig3_cod" "mig4"
## [34] "mig5_cod" "mig6_cod" "a1"
## [37] "a2" "a3" "a4"
## [40] "a5" "a6" "a6_otro"
## [43] "a7" "a8" "b1"
## [46] "b2" "b3" "b4"
## [49] "b5" "b6" "b7a_1"
## [52] "b7a_2" "b7a_3" "b7b_1"
## [55] "b7b_2" "b7b_3" "b7b_4"
## [58] "b8" "b9" "b10"
## [61] "b11_proxy" "b12" "b13_rev4cl_caenes"
## [64] "i1" "i2" "i3"
## [67] "i3_v" "i4" "i5"
## [70] "i6" "i7" "f7a"
## [73] "f7b" "f7c" "f7d"
## [76] "f7e" "b14_rev4cl_caenes" "b15_1"
## [79] "b15_2" "b16" "b16_otro"
## [82] "b17_mes" "b17_ano" "b18_codigo"
## [85] "b18_region" "b18_varias" "ocup_honorarios"
```

```
## [88] "plataformas_digitales" "pd_especifique" "b19"
## [91] "dependencia_segunda" "sda_pd" "sda_pd_especifique"
## [94] "turno" "turno_d" "turno_de"
## [97] "turno_h" "turno_t" "c2_1_1"
## [100] "c2_1_2" "c2_1_3" "c2_2_1"
## [103] "c2_2_2" "c2_2_3" "c3_1"
## [106] "c3_2" "c3_3" "turno_cont_d"
## [109] "turno_cont_de" "turno_cont_h" "turno_cont_t"
## [112] "c4" "c5" "c6"
## [115] "c7" "c8" "c9"
## [118] "c9_otro" "adicionales_h" "adicionales_d"
## [121] "adicionales_t" "c10" "c11"
## [124] "c12" "e2" "e3_1"
## [127] "e3_2" "e3_3" "e3_4"
## [130] "e3_5" "e3_6" "e3_7"
## [133] "e3_8" "e3_9" "e3_10"
## [136] "e3_11" "e3_12" "e3_total"
## [139] "e4" "e5" "e5_dia"
## [142] "e5_sem" "e5_mes" "e5_ano"
## [145] "e6_mes" "e6_ano" "e7"
## [148] "e9" "e9_otro" "e10"
## [151] "deseo_trabajar" "e11" "e12"
## [154] "e12_otro" "e13" "e21_mes"
## [157] "e21_ano" "e21_tramo" "e22"
## [160] "e23" "e23_otro" "e24"
## [163] "e24_otro" "activ" "obe"
## [166] "tpi" "id" "ftp"
## [169] "cae_general" "cae_especifico" "categoria_ocupacion"
## [172] "habituales" "efectivas" "ocup_form"
## [175] "r_p_rev4cl_caenes" "sector" "asocia"
## [178] "fact_cal"
```

Selección de columnas a trabajar

```
db2 = db[, c("region", "edad", "sexo", "estudia_actual")]
```

```
head(db2,5) # vemos las primeras 5 filas de db2
```

```
##   region edad sexo estudia_actual
## 1     4    8    2              1
## 2     6   76    1              2
## 3     3   67    1              2
## 4    15   20    1              2
## 5    11    8    2              1
```

Tipos de variable

```
str(db2)
```

```
## 'data.frame': 101592 obs. of 4 variables:
## $ region      : int  4 6 3 15 11 13 7 8 10 7 ...
## $ edad        : int  8 76 67 20 8 6 8 39 35 9 ...
## $ sexo        : int  2 1 1 1 2 1 2 2 2 1 ...
## $ estudia_actual: int  1 2 2 2 1 2 1 2 2 1 ...
```

¿Qué tipo de variables deberían ser?

Región: en Chile hay 16 regiones, pero su valor debe ser de tipo carácter (texto), porque no tiene sentido sumar las regiones, como por ejemplo el rut, si eliminamos el guión (dígito verificador) no tiene sentido hacer operaciones con los ruts, por ejemplo, si promedias tu rut con el de tu amigo. ¿Qué dice el resultado? Entonces debemos pasarlo a *tipo texto* (carácter)

Edad: Indica que es tipo numérico, pero debemos convertirlo a tipo *numérico*, porque la edad es discreta, no continua.

Sexo: dice tipo numérico, y según el documento compartido en la página del curso, 1 = Hombre, 2 = Mujer, pero de nuevo, no tiene sentido que sean numéricos porque no podemos hacer aritmética sobre ellos. Entonces conviene convertirlo a tipo *booleano*, y para mayor simplicidad, H y M, o directamente cambiarlo por la palabra Hombre o Mujer.

Estudia_actual: 1= estudia, 2 = no estudia, nuevamente es de tipo *booleano*, ya que pertenece a una u otra condición.

Antes de hacer cualquier análisis, cálculo, estadística, gráficos, etc, sobre estos datos, debemos formatearlos para no cometer errores por el tipo de formato del dato.

Formateo de los datos

```
db2$region = as.character(db2$region) #lo pasamos a caracter
db2$sexo = as.factor(db2$sexo) # lo pasamos a factor
db2$estudia_actual = as.factor(db2$estudia_actual)
```

```
#install.packages("dplyr")
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union
```

```
db2 = db2 %>%
  mutate(sexo = case_when(
    sexo == 1 ~ "H",
    sexo == 2 ~ "M",
    TRUE ~ as.character(sexo)
  ))

db2 = db2 %>%
  mutate(estudia_actual = case_when(
    estudia_actual == 1 ~ "Estudia",
    estudia_actual == 2 ~ "No estudia",
    TRUE ~ as.character(estudia_actual)
  ))
```

Vemos nuevamente el tipo de variables

```
str(db2)
```

```
## 'data.frame':   101592 obs. of  4 variables:
## $ region       : chr  "4" "6" "3" "15" ...
## $ edad         : int   8 76 67 20 8 6 8 39 35 9 ...
```

```
## $ sexo          : chr "M" "H" "H" "H" ...
## $ estudia_actual: chr "Estudia" "No estudia" "No estudia" "No estudia" ...
```

Ahora está correcto para próximos análisis.

```
head(db2,10)
```

```
##      region edad sexo estudia_actual
## 1         4   8   M      Estudia
## 2         6  76   H     No estudia
## 3         3  67   H     No estudia
## 4        15  20   H     No estudia
## 5        11   8   M      Estudia
## 6        13   6   H     No estudia
## 7         7   8   M      Estudia
## 8         8  39   M     No estudia
## 9        10  35   M     No estudia
## 10        7   9   H      Estudia
```

Ver datos faltantes (colSums para ver el resumen)

```
head(is.na(db2), 12)
```

```
##      region edad sexo estudia_actual
## [1,] FALSE FALSE FALSE          FALSE
## [2,] FALSE FALSE FALSE          FALSE
## [3,] FALSE FALSE FALSE          FALSE
## [4,] FALSE FALSE FALSE          FALSE
## [5,] FALSE FALSE FALSE          FALSE
## [6,] FALSE FALSE FALSE          FALSE
## [7,] FALSE FALSE FALSE          FALSE
## [8,] FALSE FALSE FALSE          FALSE
## [9,] FALSE FALSE FALSE          FALSE
## [10,] FALSE FALSE FALSE          FALSE
## [11,] FALSE FALSE FALSE          FALSE
## [12,] FALSE FALSE FALSE          FALSE
```

Vemos un resumen de los datos faltantes

```
colSums(is.na(db2))
```

```
##      region      edad      sexo estudia_actual
##         0         0         0         0
```

Vemos que no hay datos faltantes.

Gráficos exploratorios

```
library(ggplot2)
```

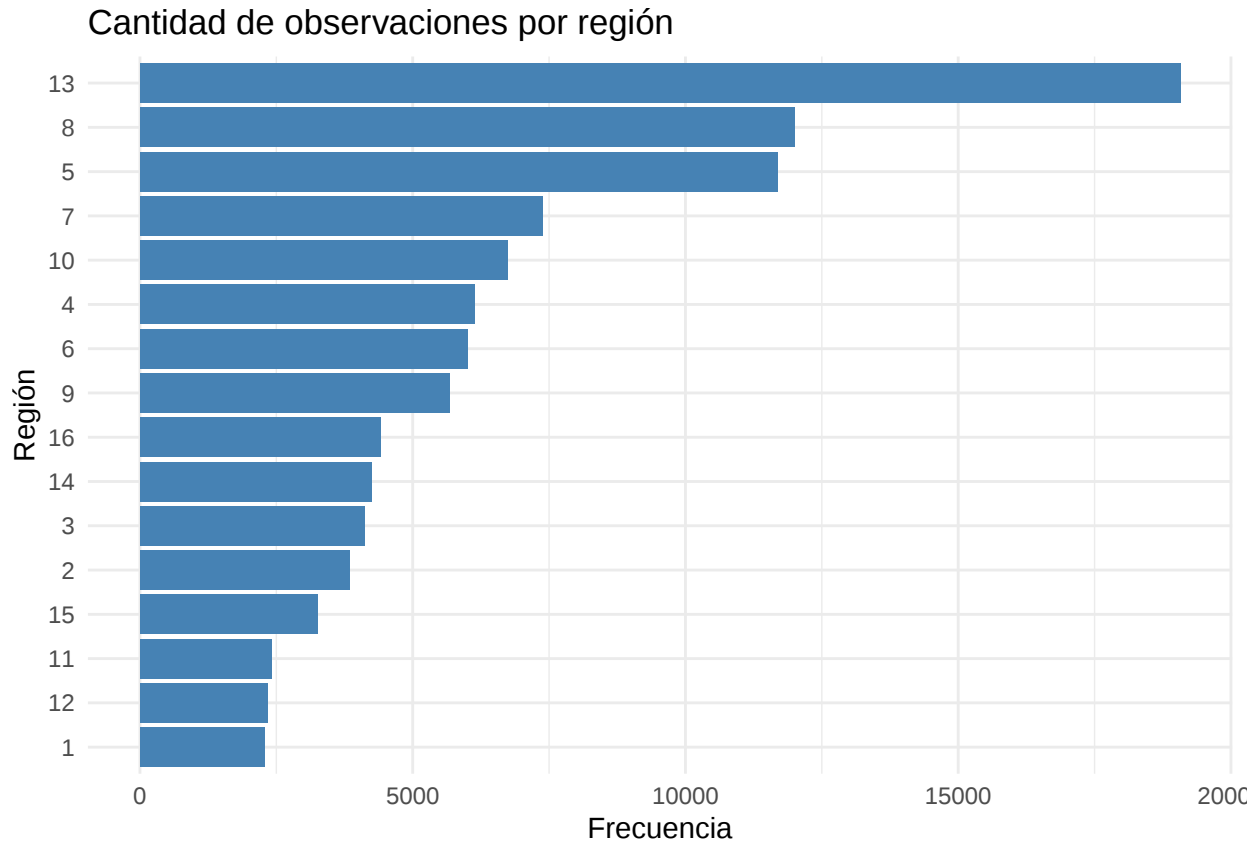
```
library(dplyr)
```

```
# Contar la cantidad de veces que aparece cada región
region_count <- db2 %>%
  count(region) %>%
  arrange(desc(n)) # Ordenar de mayor a menor
```

```
# Crear gráfico de barras
```

```
ggplot(region_count, aes(x = reorder(region, n), y = n)) +
  geom_bar(stat = "identity", fill = "steelblue") +
```

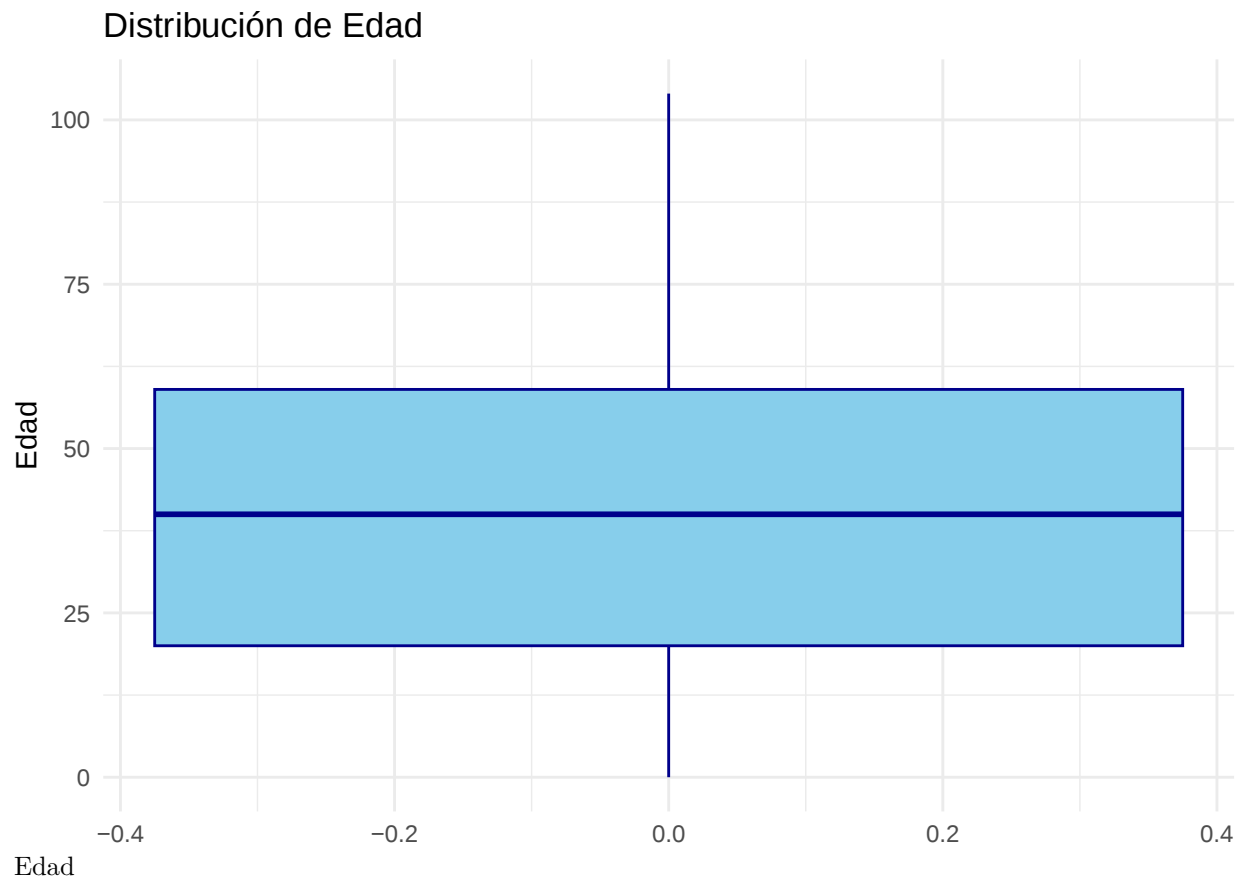
```
coord_flip() + # Gira el gráfico para que sea horizontal
labs(
  title = "Cantidad de observaciones por región",
  x = "Región",
  y = "Frecuencia"
) +
theme_minimal()
```



Boxplot para la edad

```
library(ggplot2)

ggplot(db2, aes(y = edad)) +
  geom_boxplot(fill = "skyblue", color = "darkblue") +
  labs(
    title = "Distribución de Edad",
    y = "Edad"
  ) +
  theme_minimal()
```

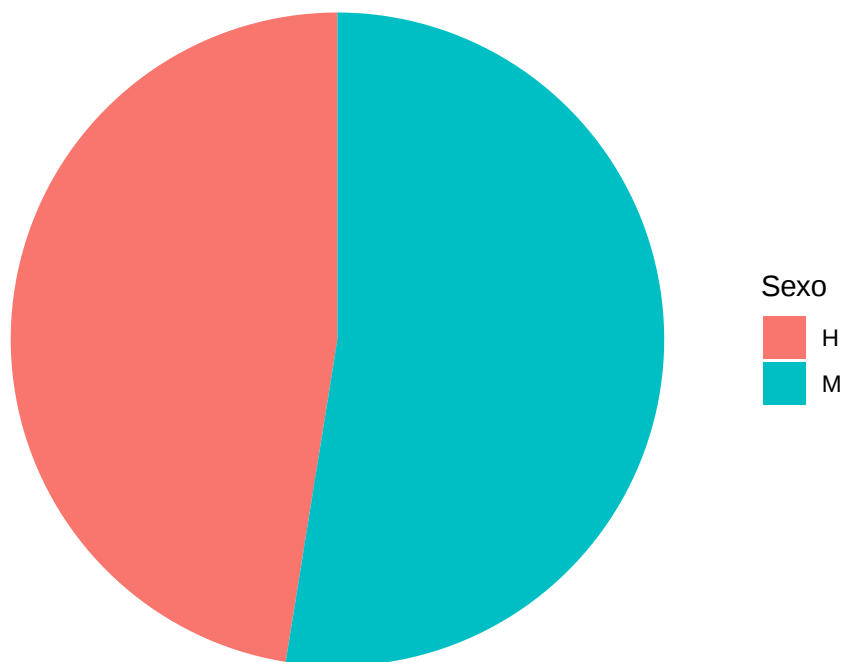


```
library(ggplot2)
library(dplyr)

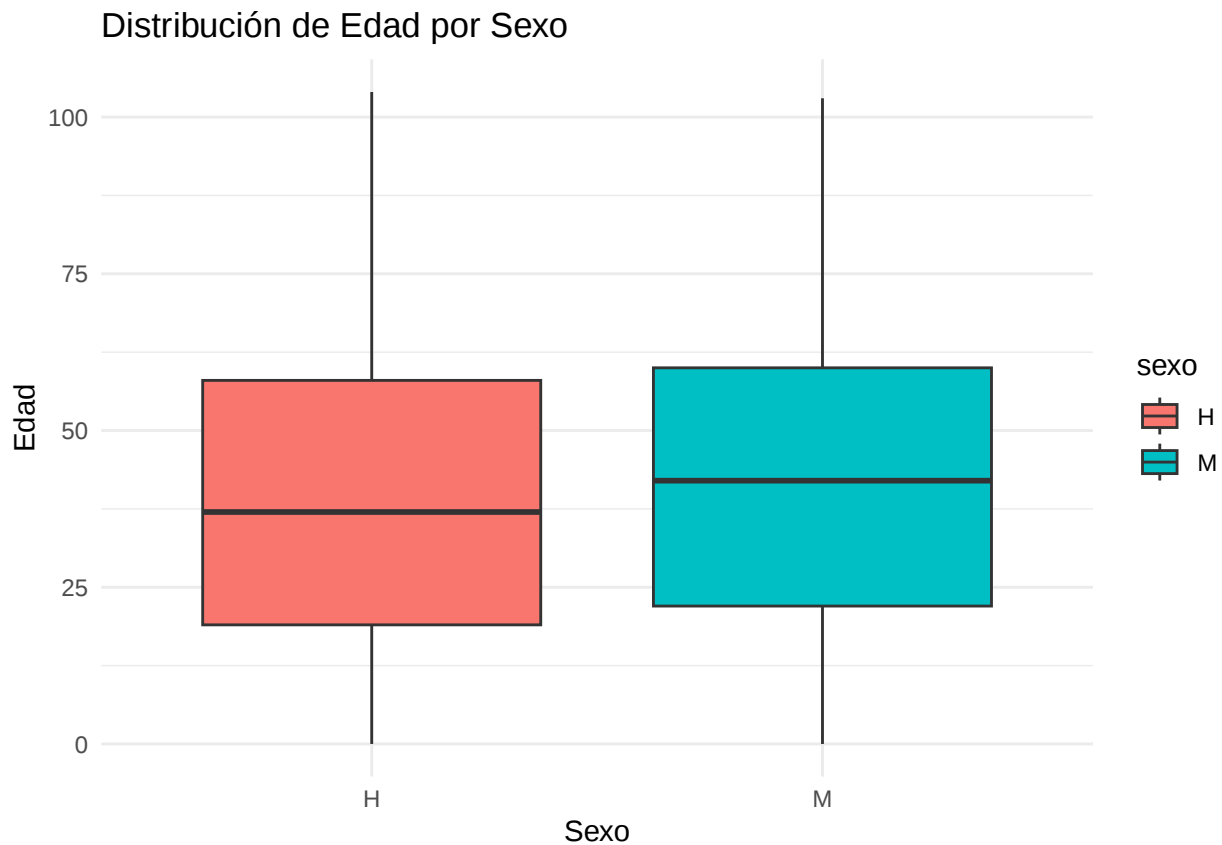
# Contar cantidad por sexo
sexo_count <- db2 %>%
  count(sexo)

# Crear gráfico de torta
ggplot(sexo_count, aes(x = "", y = n, fill = sexo)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  labs(
    title = "Distribución por Sexo",
    fill = "Sexo"
  ) +
  theme_void()
```

Distribución por Sexo



```
ggplot(db2, aes(x = sexo, y = edad, fill = sexo)) +  
  geom_boxplot() +  
  labs(  
    title = "Distribución de Edad por Sexo",  
    x = "Sexo",  
    y = "Edad"  
  ) +  
  theme_minimal()
```



Hipótesis nula (H_0): Las medias de edad entre hombres (H) y mujeres (M) son iguales. ¿Es correcto?

```
t.test(edad ~ sexo, data = db2)
```

```
##
##  Welch Two Sample t-test
##
## data:  edad by sexo
## t = -18.918, df = 100651, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group H and group M is not equal to 0
## 95 percent confidence interval:
##  -3.033405 -2.463860
## sample estimates:
## mean in group H mean in group M
##      38.74800      41.49664
```

Se rechaza la hipótesis de que el promedio de las edades entre Hombres y Mujeres es igual.