TRABAJO GRUPAL

ENCUESTA CIS

Hugo Alonso, Gonzalo Blanca, Pablo Galarón y Raúl Palomo

TABLA DE CONTENIDOS

01

PREPARACIÓN DATOS

NS/NC → NA

```r
data_num <- data_num |>
  mutate(
    INGRESHOG = ifelse(INGRESHOG %in% c(8, 9), NA, INGRESHOG),
    TARHOGENTREV_HH = ifelse(TARHOGENTREV_HH %in% c(98, 99), NA, TARHOGENTREV_HH),
    SITLAB = ifelse(SITLAB == 9, NA, SITLAB),
    NIVELESTENTREV = ifelse(NIVELESTENTREV %in% c(98, 99), NA, NIVELESTENTREV),
    P6_2 = ifelse(P6_2 %in% c(8, 9), NA, P6_2),
    P4 = ifelse(P4 %in% c(8, 9), NA, P4),
    P7_1 = ifelse(P7_1 %in% c(8, 9), NA, P7_1),
    P7_2 = ifelse(P7_2 %in% c(8, 9), NA, P7_2),
    P7_3 = ifelse(P7_3 %in% c(8, 9), NA, P7_3),
    ESCFEMINIS = ifelse(ESCFEMINIS %in% c(98, 99), NA, ESCFEMINIS),
    ESCIDEOL = ifelse(ESCIDEOL %in% c(98, 99), NA, ESCIDEOL),
    RELIGION = ifelse(RELIGION == 9, NA, RELIGION),
    CUIDADOHIJOS_HH = ifelse(CUIDADOHIJOS_HH %in% c(96, 98, 99), NA, CUIDADOHIJOS_HH)
    RECUVOTOG = ifelse(RECUVOTOG == 9998, NA, RECUVOTOG)
  )
```

Mirar los porcentajes y significado ⚠️ de cada variable para tomar la decisión

PASAMSOS A FACTOR Y RECODIFICAMOS LAS CUALIS

```r
data_num <- data_num |>
  mutate(
    SEXO = factor(SEXO, levels = c(1, 2), labels = c("Hombre", "Mujer")),  # nominal
    TAMUNI = factor(TAMUNI,
                    levels = 7:1,
                    labels = c(">1000000 habitantes",
                               "400001-1000000 habitantes",
                               "100001-400000 habitantes",
                               "50001-100000 habitantes",
                               "10001-50000 habitantes",
                               "2001-10000 habitantes",
                               "≤2000 habitantes"),
                    ordered = TRUE),  # ordinal
```

🔢 Cuidado con las ordinales, asignarles el orden.

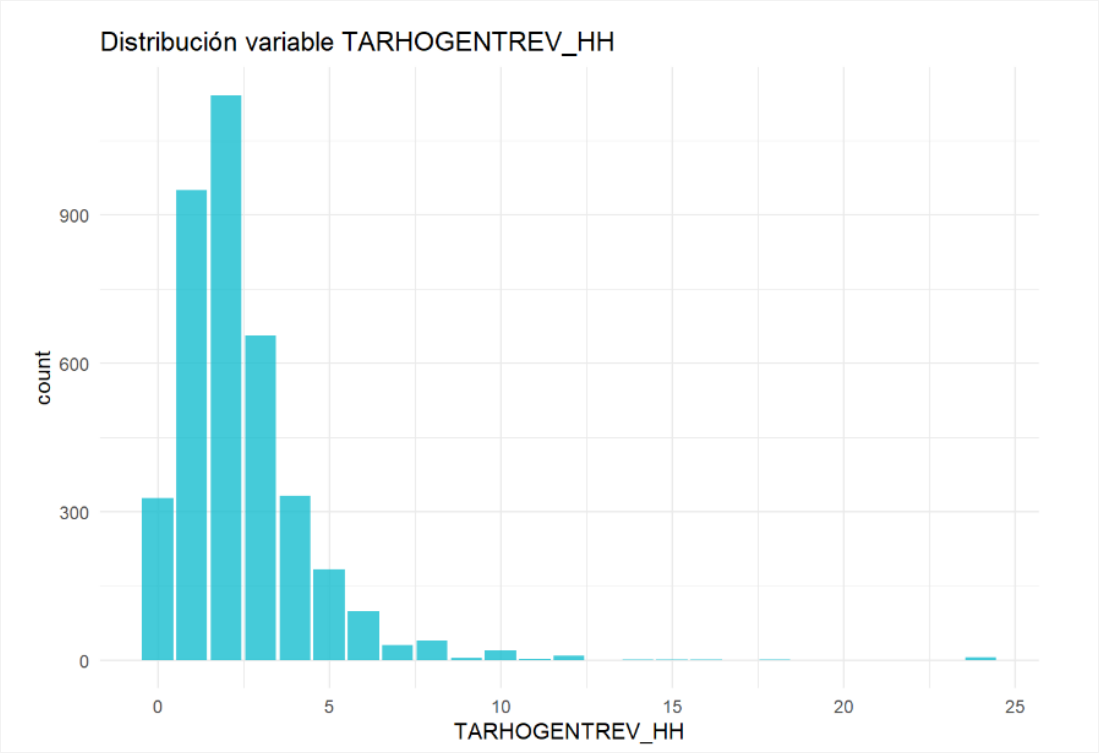
02

ERRORES INICIALES

TARHOGENTREV_HH

Horas dedicadas a tareas del hogar 🏠

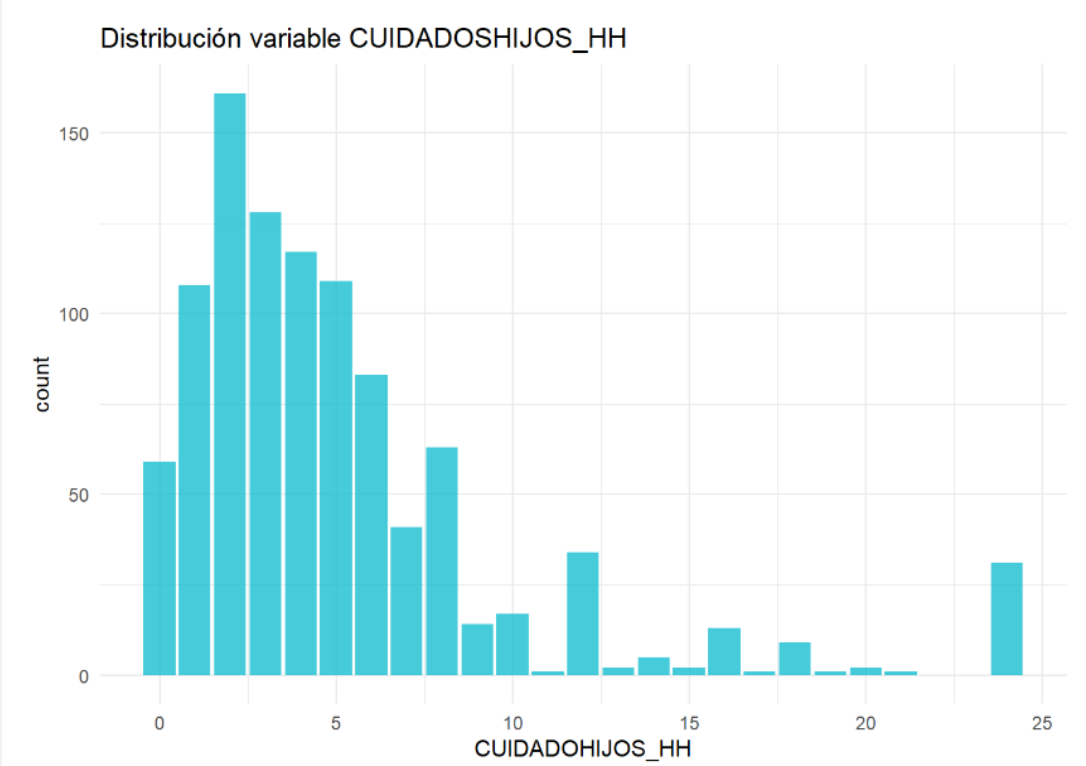
```
TARHOGENTREV_HH
Min.    : 0.000
1st Qu.: 1.000
Median : 2.000
Mean    : 2.401
3rd Qu.: 3.000
Max.    :24.000
NA's    :204
```



CUIDADOHIJOS_HH

Horas dedicadas al cuidado de los hijos 👨‍👩‍👧‍👦

```
CUIDADOHIJOS_HH
Min.    : 0.000
1st Qu.: 2.000
Median : 4.000
Mean    : 5.158
3rd Qu.: 6.000
Max.    :24.000
NA's    :3003
```

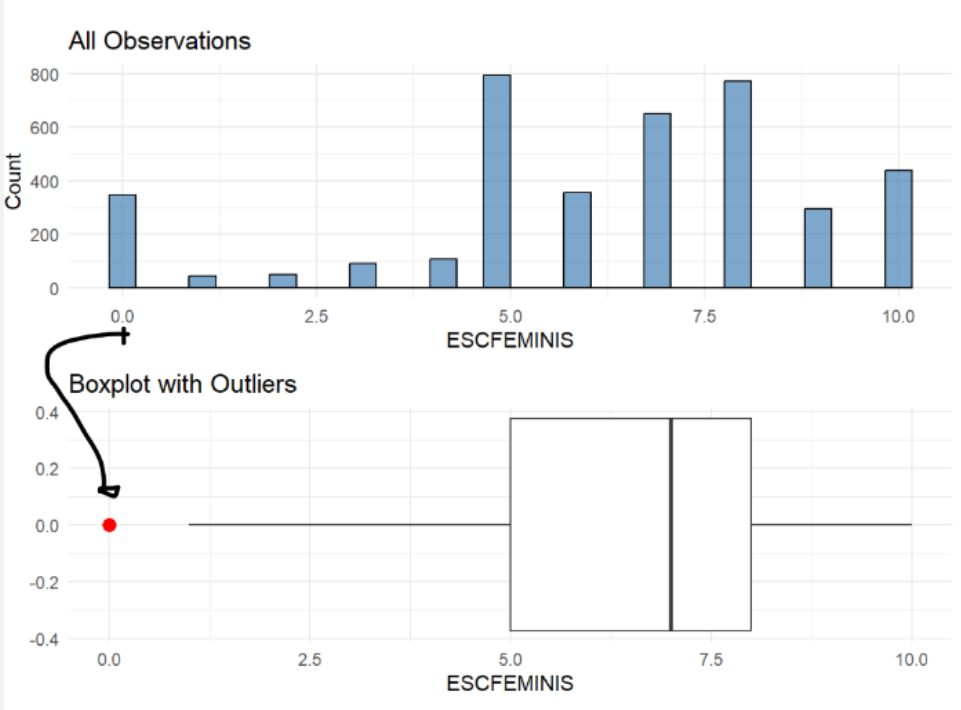


03

DETECCIÓN ATÍPICOS

Variable ESCFEMINIS

TRATAMIENTO UNIVARIANTE



📌 Outliers identified in ESCFEMINIS : 346 outliers
📊 Proportion (%) of outliers: 8.77 %

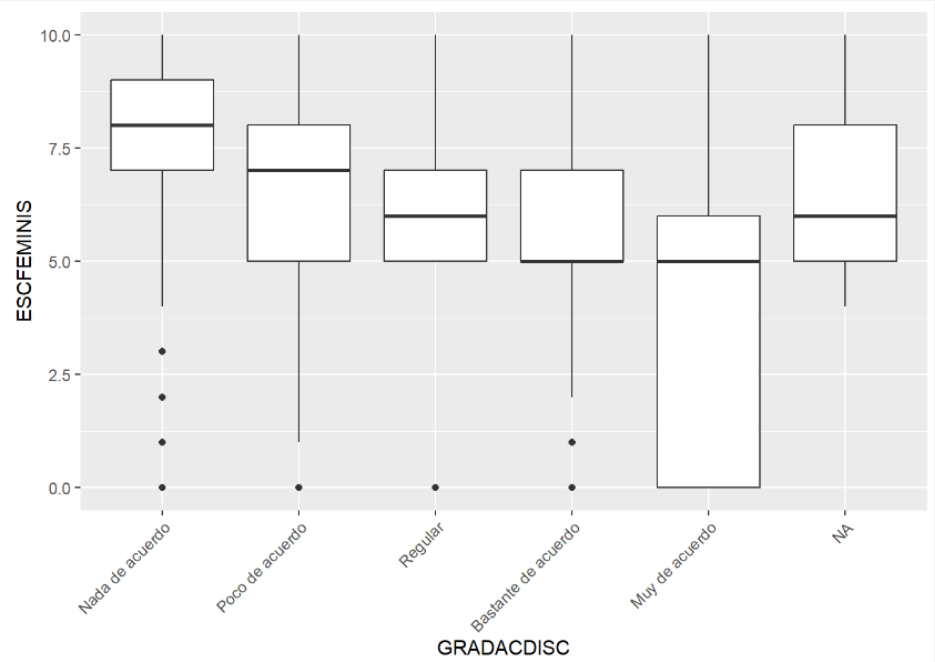
🚨 Extreme values identified in ESCFEMINIS : 0 extreme values
📊 Proportion (%) of extreme values: 0 %

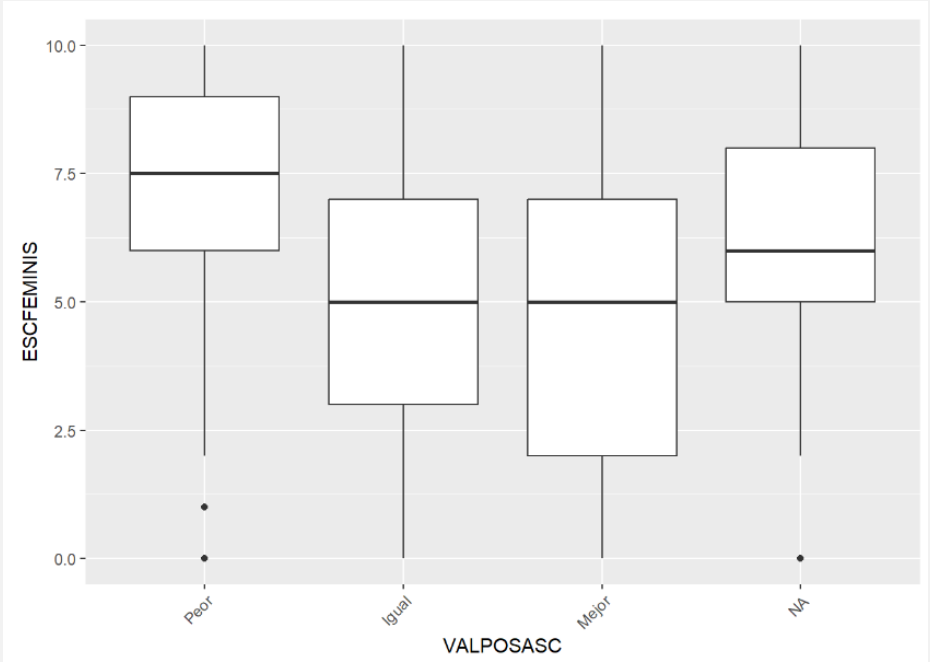
- **No hay valores extremos**, pero sí un **8,77 % de outliers**. Alto %.

- **Todos los outliers son el valor 0** ("nada feminista").

- **Visualmente el 0 no destaca** en el histograma: su barra es **similar a otras**. Distrib asimétrica

TRATAMIENTO BIVARIANTE

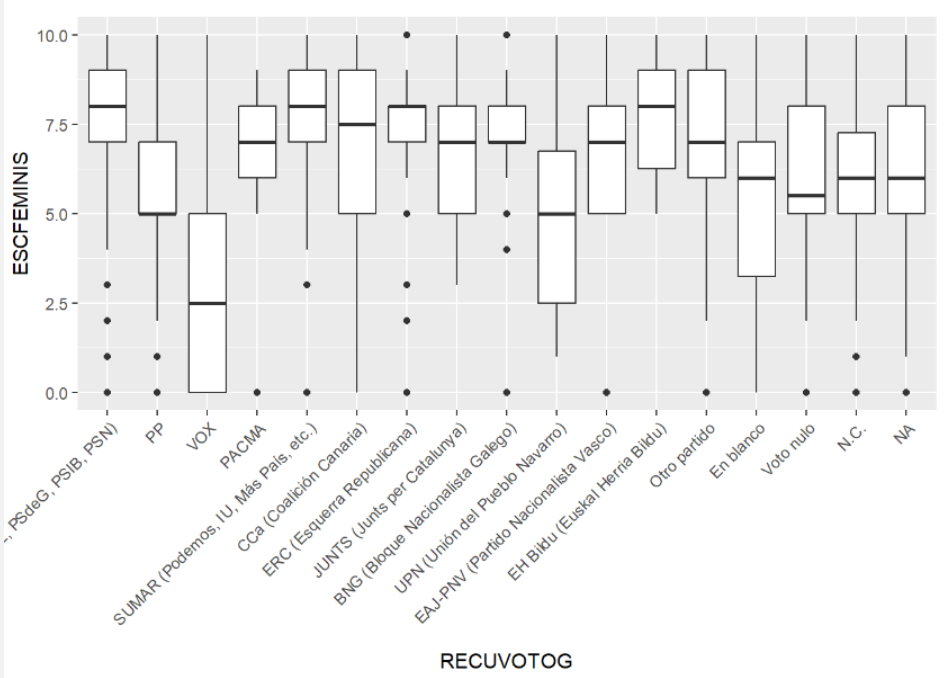


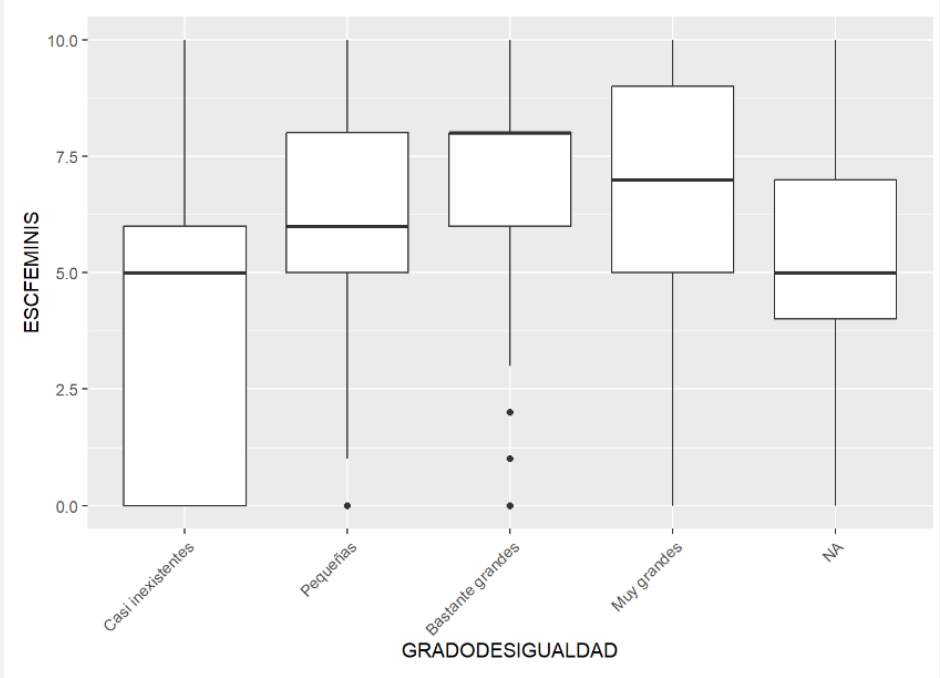


• Alta puntuación en "se discrimina a los hombres" → **baja puntuación en feminismo**.

• El valor **0 en ESCFEMINIS** se da sobre todo en personas que **no creen que las mujeres tengan más dificultades para ascender**.

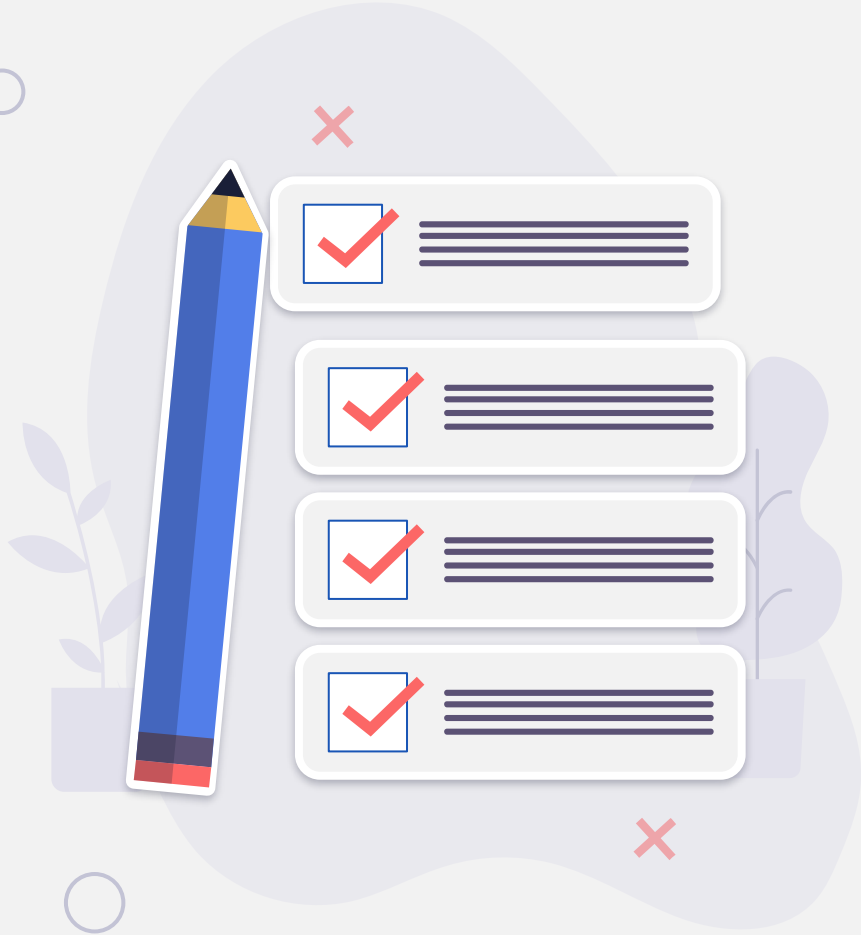
TRATAMIENTO BIVARIANTE





• Los valores de 0 se concentran en votantes de **VOX,** partido crítico con el feminismo.

• Quienes **perciben poca desigualdad de género** tienden a puntuar bajo en ESCFEMINIS.



Por tanto:

- Veníamos viendo en univariante
- Alto porcentaje de atípicos
- Asociación outliers con variables

NO LOS QUITAMOS

04

ALGORITMO LOF

LOF no permite NA'S

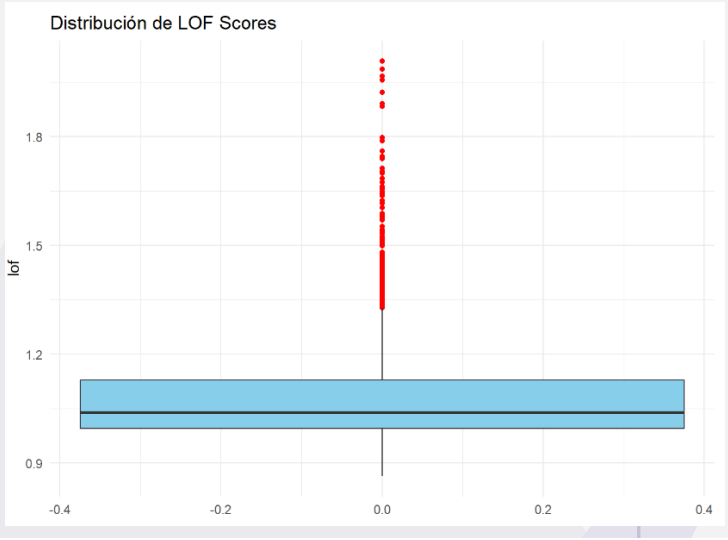
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
EDAD	0	1.00	50.02	17.31	16	37	50	63	97	
TARHOGENTREV_HH	210	0.95	2.37	1.81	0	1	2	3	18	
ESCFEMINIS	59	0.99	6.27	2.74	0	5	7	8	10	
ESCIDEOL	218	0.95	4.77	2.52	1	3	5	7	10	
CUIDADOHIJOS_HH	3003	⚠ 0.25	5.16	4.95	0	2	4	6	24	

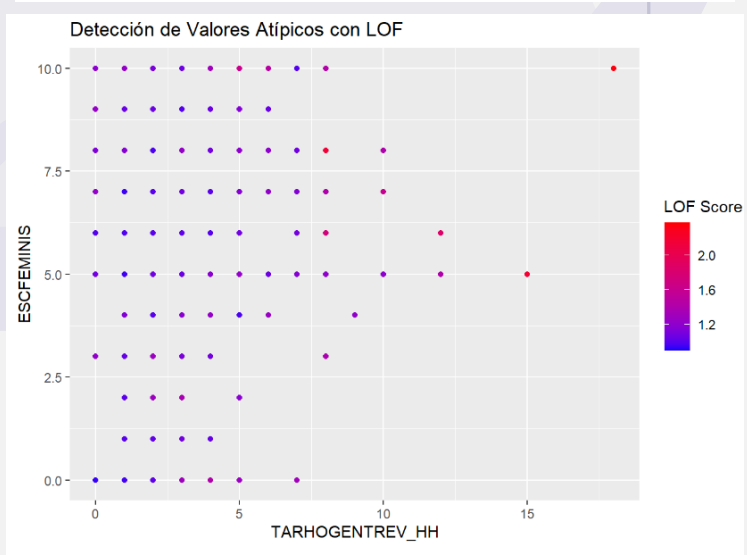
LOF sin cuidado_hijos

LOF sin hijos	Valor
Minimo LOF	1.50238
Máximo LOF	2.00844

LOF con cuidado_hijos

LOF scon hijos	Valor
Minimo LOF	1.502268
Máximo LOF	2.37662





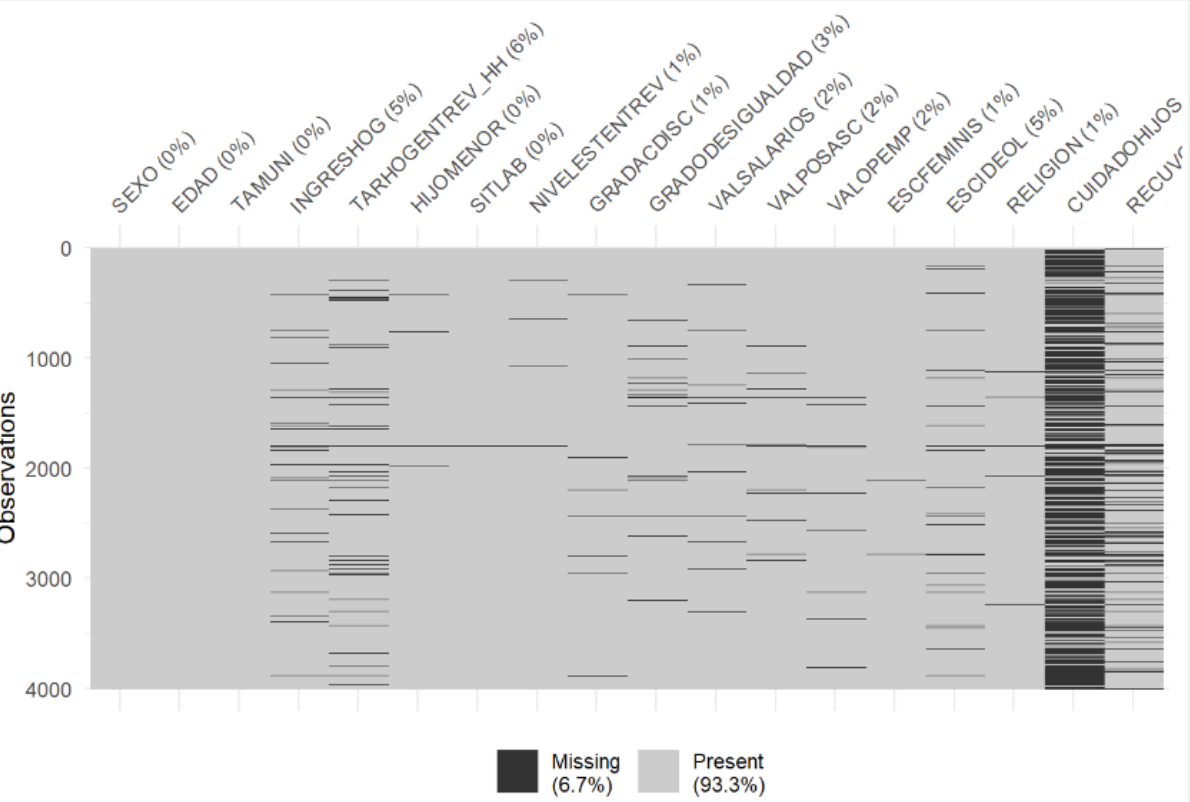
- **No hay observaciones extremadamente atípicas** que deban eliminarse.
- Se observaron algunos valores alejados, pero **sin grandes saltos** en las distancias LOF.

- Los **valores LOF más altos** se asocian con **extremos de TARHOGARENTREV_HH(variable para cual habíamos observado atípicos)**
- **No se elimina toda la observación**, solo los **valores atípicos que hacen la obs rara.**

05

TRATAMIENTO DE AUSENTES

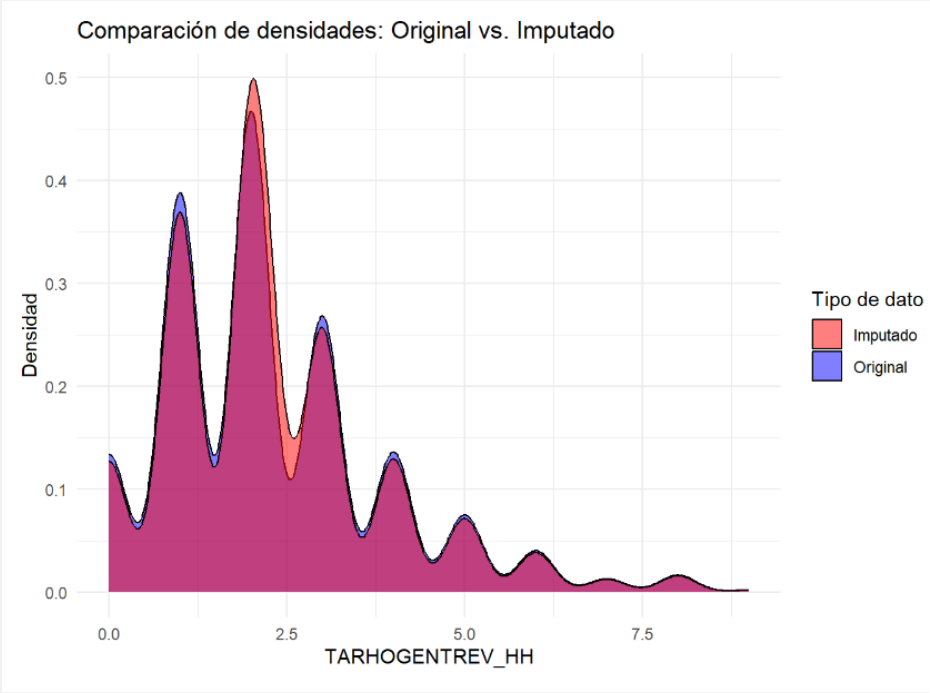
IMPUTACIÓN SIMPLE



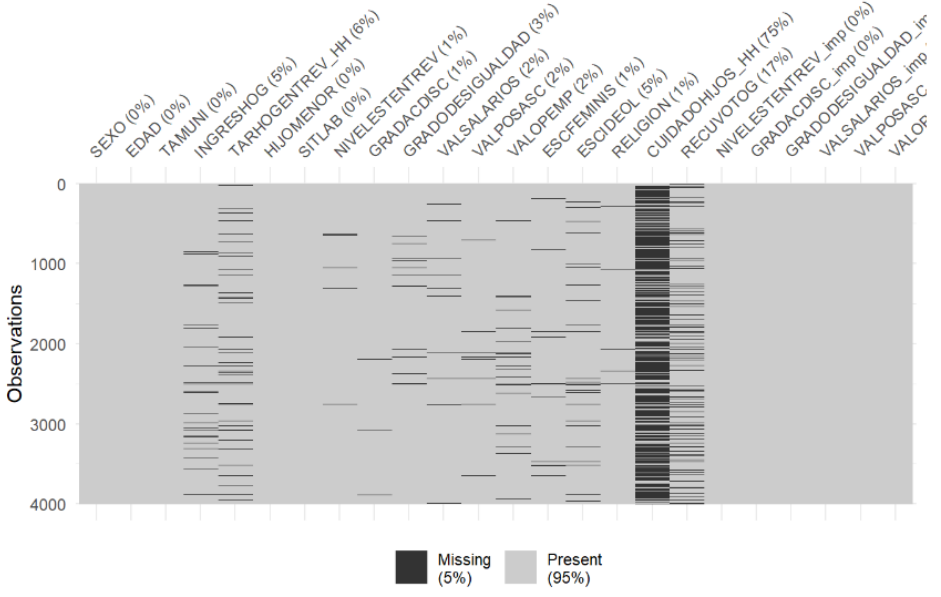
- Cualis <5% -> **imp MODA**
- Todas Cuantis -> **imp MEDIA**
- Todas las variables son **MCAR**

IMPUTACIÓN SIMPLE

Cuanti por media

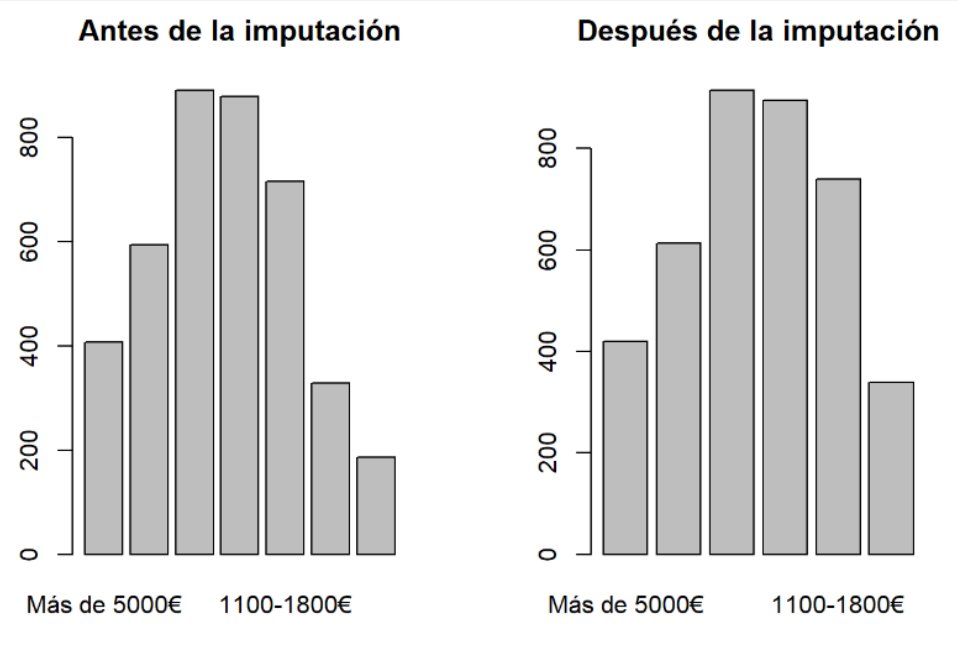


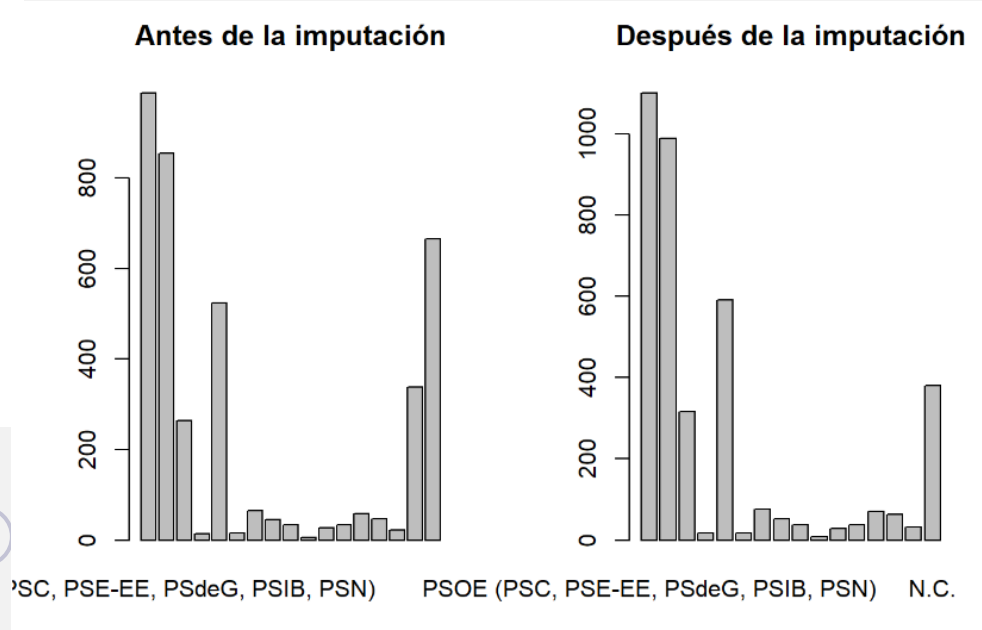
Imp por moda cualis



IMPUTACIÓN MÚLTIPLE

Las cualis de >5% ->> imp mice





GRACIAS!

Hugo Alonso, Gonzalo Blanca,
Pablo Galarón y Raúl Palomo