ANEXO: Limpieza y filtrado de la base de datos

Jeje

```
#Limpieza de las bases de datos ####

rm(list=ls()) #Limpiamos la memoria

##Librerias####
library(tidyverse)    #Para manejar bases de datos.
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(eph)          #Libreria hecha por cientificxs argentinxs.
library(tinytable)

##Descarga de base####
df <- get_microdata(year = 2023, #del paquete EPH
                    trimester = 4,
                    type = 'individual'
```

```
                    )

## Transformaciones y filtros ####

### Transformaciones de tipo de variable ####
df <- df %>% mutate_at(vars(NIVEL_ED,
                            AGLOMERADO,
                            CH07, #estado civil
                            ESTADO,
                            REGION
                            ),
                       ~as.factor(.)
                       )

### Generacion de la variable de años de educacion ####
df <- df %>%
  mutate(CH14bis = replace_na(CH14, 0)) %>%
  mutate(CH14bis = replace(CH14bis, CH14bis == "", 0))

df <- df %>%
  mutate(ult_anio = as.numeric(CH14bis)) %>%
  mutate(ult_anio = case_when(
    ult_anio == 98 ~ 0,  # Educacion especial
    ult_anio == 99 ~ 0,  # Ns/Nr
    TRUE ~ ult_anio      # Keep original values for other cases
  ))

df <- df %>%
  mutate(educn = case_when(               #Nivel mas alto cursado:
    CH12 == 1 ~ 0,                           #Jardin o preescolar
    CH12 == 2 & CH13 == 1 ~ 7,               #Primario completo
    CH12 == 2 & CH13 == 2 ~ (0 + ult_anio),  #Primario incompleto
    CH12 == 3 & CH13 == 1 ~ 9,               #EGB completo
    CH12 == 3 & CH13 == 2 ~ (0 + ult_anio),  #EGB incompleto
    CH12 == 4 & CH13 == 1 ~ 12,              #Secundario completo
    CH12 == 4 & CH13 == 2 ~ (7 + ult_anio),  #Secundario incompleto
    CH12 == 5 & CH13 == 1 ~ 12,              #Polimodal completo
    CH12 == 5 & CH13 == 2 ~ (9 + ult_anio),  #Polimodal incompleto
    CH12 == 6 & CH13 == 1 ~ 15,              #Terciario completo
    CH12 == 6 & CH13 == 2 ~ (12 + ult_anio), #Terciario incompleto
    CH12 == 7 & CH13 == 1 ~ 18,              #Universitario completo
    CH12 == 7 & CH13 == 2 ~ (12 + ult_anio), #Universitario incompleto
```

```
    CH12 == 8 & CH13 == 1 ~ 22,              #Posgrado completo
    CH12 == 8 & CH13 == 2 ~ (18 + ult_anio), #Posgrado incompleto
    TRUE ~ 0
  ))

#Cambios de nombres de variables
df <- df %>%
  rename(edad = CH06) %>%
  mutate(est_civ = recode_factor(CH07,
                                 "1" = "Unido",
                                 "2" = "Casado",
                                 "3" = "Separado/Divorciado",
                                 "4" = "Viudo",
                                 "5" = "Soltero"
                                 )
         ) %>%
  mutate(region = recode_factor(REGION,
                                 "1" = "GBA",
                                "40" = "Noroeste",
                                "41" = "Noreste",
                                "42" = "Cuyo",
                                "43" = "Pampeana",
                                "44" = "Patagonia"
                                )
         ) %>%
  mutate(aglomerado = recode_factor(AGLOMERADO,
                                 "2" = "Gran La Plata",
                                "02" = "Gran La Plata",
                                 "3" = "Bahia Blanca - Cerri",
                                "03" = "Bahia Blanca - Cerri",
                                 "4" = "Gran Rosario",
                                "04" = "Gran Rosario",
                                 "5" = "Gran Santa Fe",
                                "05" = "Gran Santa Fe",
                                 "6" = "Gran Parana",
                                "06" = "Gran Parana",
                                 "7" = "Posadas",
                                "07" = "Posadas",
                                 "8" = "Gran Resistencia",
                                "08" = "Gran Resistencia",
                                 "9" = "Comodoro Rivadavia - Rada Tilly",
                                "09" = "Comodoro Rivadavia - Rada Tilly",
```

```
                                  "10" = "Gran Mendoza",
                                  "12" = "Corrientes",
                                  "13" = "Gran Cordoba",
                                  "14" = "Concordia",
                                  "15" = "Formosa",
                                  "17" = "Neuquen - Plottier",
                                  "18" = "Santiago del Estero - La Banda",
                                  "19" = "Jujuy - Palpala",
                                  "20" = "Rio Gallegos",
                                  "22" = "Gran Catamarca",
                                  "23" = "Gran Salta",
                                  "25" = "La Rioja",
                                  "26" = "Gran San Luis",
                                  "27" = "Gran San Juan",
                                  "29" = "Gran Tucuman - Tafi Viejo",
                                  "30" = "Santa Rosa - Toay",
                                  "31" = "Ushuaia - Rio Grande",
                                  "32" = "Ciudad Autonoma de Buenos Aires",
                                  "33" = "Partidos del GBA",
                                  "34" = "Mar del Plata",
                                  "36" = "Rio Cuarto",
                                  "38" = "San Nicolas - Villa Constitucion",
                                  "91" = "Rawson - Trelew",
                                  "93" = "Viedma - Carmen de Patagones"
                                  )
        ) %>%
  mutate(educf = recode_factor(NIVEL_ED,
                                  "1" = "Primario incompleto",
                                  "2" = "Primario completo",
                                  "3" = "Secundario incompleto",
                                  "4" = "Secundario completo",
                                  "5" = "Superior universitario incompleto",
                                  "6" = "Superior universitario completo",
                                  "7" = "Sin instruccion"
                                  )
        )

### Generacion de la variable de intervalos de edad ####
df <- df %>%
  mutate(edadi = case_when(
    edad >= 25 & edad <= 29 ~ "De 25 a 29 años",
    edad >= 30 & edad <= 34 ~ "De 30 a 34 años",
```

```
    edad >= 35 & edad <= 39 ~ "De 35 a 39 años",
    edad >= 40 & edad <= 44 ~ "De 40 a 44 años",
    edad >= 45 & edad <= 49 ~ "De 45 a 49 años",
    edad >= 50 & edad <= 54 ~ "De 50 a 54 años",
    edad >= 55 & edad <= 59 ~ "De 55 a 59 años",
    edad >= 60 & edad <= 65 ~ "De 60 a 65 años",
    TRUE ~ NA_character_
  )) %>%
  mutate(edadi = factor(edadi))

### Rebase de los factores
df$educf <- relevel(df$educf,"Secundario completo")
df$est_civ <- relevel(df$est_civ,"Casado")
df$region <- relevel(df$region,"GBA")
df$edadi <- relevel(df$edadi,"De 25 a 29 años")


## Base del primer punto ####
### Filtramos segan la consigna del primer punto
df1 <- df %>%
  filter(CH03 == 1,           #Jefes/as de hogar
         CH04 == 1,           #Hombres
         edad >= 25,          #Entre 25...
         edad <= 65,          #...y 65 años
         ESTADO == 1,         #Ocupados
         CAT_OCUP == 3,       #Asalariados
         P21 > 0,             #Salario positivo
         CH12 != 9            #¬Educacion especial
  )

#Logaritmo del Salario (P21)
df1 <- df1 %>%
  mutate(logSal = log(df1$P21)
  )

### Seleccionamos las columnas para el 1abc) ####
df1 <- df1 %>%
  select(logSal,
         educn,
         educf,
         edad,
         edadi,
```

```
          est_civ,
          region,
          aglomerado,
          PONDERA,
          PONDIIO,
          PONDII
  )

### ## Filtramos para los puntos 1d y 1e ####
df2 <- df %>%
  filter(CH03 == 1,           #Jefes/as de hogar
          CH04 == 1,          #Hombres
          edad >= 25,         #Entre 25...
          edad <= 65,         #...y 65 años
          ESTADO == 1 | ESTADO == 2,         #Ocupados
          #CAT_OCUP == 3,       #Asalariados
          #P21 > 0,            #Salario positivo
          CH12 != 9           #¬Educacion especial
  )

df2 <- df2 %>% mutate(estado = case_when(
  ESTADO == 1 ~ 0,  #0 si está empleado
  ESTADO == 2 ~ 1,  #1 si está desempleado
  TRUE ~ NA_real_))

### Seleccionamos las columnas para el 1de) ####
df2 <- df2 %>%
  select(estado,
          educn,
          educf,
          edad,
          edadi,
          est_civ,
          region,
          aglomerado
  )

## Base del segundo punto ####
# PP3E_TOT N(5,1)
# Total de horas que trabajó en la semana
# en la ocupación principal
#
```

```r
# PP3F_TOT N(5,1)
# Total de horas que trabajó en la semana
# en otras ocupaciones

### Generación de la variable de horas trabajadas:
df <- df %>%
  mutate(PP3E_TOT = replace_na(PP3E_TOT, 0)) %>%
  mutate(PP3E_TOT = replace(PP3E_TOT, as.numeric(PP3E_TOT) == 999, 0)) %>%

  mutate(PP3F_TOT = replace_na(PP3F_TOT, 0)) %>%
  mutate(PP3F_TOT = replace(PP3F_TOT, as.numeric(PP3F_TOT) == 999, 0)) %>%

  mutate(horas = as.numeric(PP3E_TOT) + as.numeric(PP3F_TOT))

# Acá mostraría en una tabla la decisión de sumar las horas, que me parece interesante.
# O incluso, si decidimos no mostrarlas, mostrar igual la tablita esta:
tabla_full_time <- data.frame(
  Según_PP3F_TOT = sum(df$PP3F_TOT >= 35),
  Según_PP3E_TOT = sum(df$PP3E_TOT >= 35),
  Según_Ambas = sum(df$horas >= 35)
)

rownames(tabla_full_time) <- "Trabajadores Jornada Completa"

tabla_full_time
```

```
                            Según_PP3F_TOT Según_PP3E_TOT Según_Ambas
Trabajadores Jornada Completa             87          12486       13209
```

```r
# tt(tabla_full_time)

### Generación de variable de formalidad que funciona pero mal ####
df <- df %>%
  mutate(formalidad = case_when(
    df$PP07I == 1 ~ "Aportes Propios",
    df$PP07H == 1 ~ "Aportes Empleador",
    (df$PP07H == 2) | (df$PP07I == 2)~ "Sin Aportes",
    TRUE ~ as.character(NA)
  )) %>%
  mutate(formalidad = factor(formalidad,
                             levels = c("Sin Aportes",
```

```
                                        "Aportes Propios",
                                        "Aportes Empleador"
                                ),
                                ordered = TRUE
                                )
        )



# df <- df %>%
#   mutate(formalidad = case_when(
#     (df$PP07H == 2) | (df$PP07H == 0) ~ "Sin aportes",
#     (as.numeric(df$PP07H) == 1) & (as.numeric(df$PP07I) == 1) ~ "Aportes Propios",
#     (as.numeric(df$PP07H) == 1) & (as.numeric(df$PP07I) == 2) ~ "Aportes Empleador",
#     (as.numeric(df$PP07H) == 1) & (is.na(df$PP07I) | as.numeric(df$PP07H) == 0) ~ "Aportes
#     TRUE ~ as.character(NA)
#     )) %>%
  # mutate(formalidad = factor(formalidad,
  #                            levels = c("Sin aportes",
  #                                       "Aportes Propios",
  #                                       "Aportes Empleador"
  #                                       ),
  #                            ordered = TRUE
  #                            ))

# ### Generación de variable de formalidad ####
# df <- df %>%
#   mutate(formalidad = case_when(
#     PP07H == 2 ~ "Sin aportes",
#     PP07H == 1 & PP07I == 1 ~ "Aportes Propios",
#     PP07H == 1 & PP07I == 2 ~ "Aportes Empleador",
#     PP07H == 1 & is.na(PP07I) ~ "Aportes indefinidos",
#     TRUE ~ as.character(NA)
#     )) %>%
#   mutate(formalidad = factor(formalidad,
#                              levels = c("Sin aportes",
#                                         "Aportes indefinidos",
#                                         "Aportes Propios",
#                                         "Aportes Empleador"
#                                         ),
#                              ordered = TRUE
#                              ))
```

```r
## Filtramos para el punto 1e ####
df3 <- df %>%
  filter(CH03 == 1,           #Jefes/as de hogar
         CH04 == 1,           #Hombres
         edad >= 25,          #Entre 25...
         edad <= 65,          #...y 65 años
         CAT_OCUP == 3,       #Asalariados
         ESTADO == 1 | ESTADO == 2, #Ocupadosy desocupados
         # P21  0,            #Salario positivo
         CH12 != 9            #¬Educacion especial
  )

df3 <- df3 %>% mutate(estado = case_when(
  ESTADO == 1 ~ 1,  #1 si está empleado
  ESTADO == 2 ~ 0,  #0 si está desempleado
  TRUE ~ NA_real_))

#Logaritmo del Salario (P21)
df3 <- df3 %>%
  mutate(logSal = log(df3$P21)
  )
```

```
Warning: There was 1 warning in `mutate()`.
i In argument: `logSal = log(df3$P21)`.
Caused by warning in `log()`:
! Se han producido NaNs
```

```r
df3 <- df3 %>% mutate(logSal = case_when(
  logSal == -Inf ~ 0,
  TRUE ~ logSal))

# df3 <- df3 %>% filter(!is.na(df3$logSal))

df3 <- df3 %>%
  select(logSal,
         estado,
         educn,
         educf,
         edad,
         edadi,
         est_civ,
```

```
         region,
         aglomerado
  )
# data <- na.omit(data)

## Filtramos para el punto 2 ####
df4 <- df %>%
  filter(CH03 == 1,           #Jefes/as de hogar
         CH04 == 1,           #Hombres
         edad >= 25,          #Entre 25...
         edad <= 65,          #...y 65 años
         CAT_OCUP == 3,       #Asalariados
         #PP3E_TOT >= 35,     #Solo los que trabajan jornada completa en ocup. ppal.
         horas >= 35,         #Quienes trabajan jornada completa en total
         #P21 > 0,            #Salario positivo
         CH12 != 9            #¬Educacion especial
         )

sum(is.na(df4$formalidad))
```

```
[1] 0
```

```
### Seleccionamos las columnas para el 2ab) ####
df4 <- df4 %>%
  select(formalidad,
         educn,
         educf,
         edad,
         edadi,
         est_civ,
         region,
         aglomerado
  )

saveRDS(df1, file = "Bases/eph_1abc.RDS")
saveRDS(df2, file = "Bases/eph_1d.RDS")
saveRDS(df3, file = "Bases/eph_1e.RDS")
saveRDS(df4, file = "Bases/eph_2ab.RDS")

# sjlabelled::write_stata(df1, "Bases/eph_1abc.dta")
# sjlabelled::write_stata(df2, "Bases/eph_1d.dta")
```

```
# sjlabelled::write_stata(df3, "Bases/eph_1e.dta")
# sjlabelled::write_stata(df4, "Bases/eph_2ab.dta")
```