Taller 1

2024-02-07

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::combine() masks gridExtra::combine()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ readr::guess\_encoding() masks rvest::guess\_encoding()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

segun Smith (2010) .

url\_base<-my\_url <- "https://ignaciomsarmiento.github.io/GEIH2018\_sample/pages/geih\_page\_"  
  
tables\_list <- list()  
  
for (i in 1:10){  
 bucle\_url<-paste0(url\_base, i, ".html") # Bucle  
 my\_html=read\_html(bucle\_url)  
 table<-my\_html %>% html\_table()  
 tables\_list[[i]] <- table  
}  
  
data<-dplyr::bind\_rows(tables\_list) # Dataframe

## New names:  
## New names:  
## New names:  
## New names:  
## New names:  
## New names:  
## New names:  
## New names:  
## New names:  
## New names:  
## • `` -> `...1`

t2<-Sys.time()  
Tiempo\_ejecucion<-t2-t1  
Tiempo\_ejecucion

## Time difference of 4.512639 mins

La estrategia propuesta para conservar la representatividad de la muestra que permite el factor de expansión consiste en transformar esta variable en un factor de ponderación de manera que conservemos su proporción con respecto al universo sin afectar el tamaño original de la muestra, de esta manera se ahorra recursos computacionales la ejecución de los algoritmos y el tratamiento de los datos.

base\_no\_ponderada<-length(data$fex\_c)  
universo\_representado<-sum(data$fex\_c)  
  
data<- data %>%  
 mutate(data, peso=fex\_c/universo\_representado) %>%   
 mutate(data,fponderacion=peso\*base\_no\_ponderada)  
  
base\_ponderada<-sum(data$fponderacion)  
unidades\_muestrales\_fex<-length(unique(data$fex\_c))  
unidades\_muestrales\_fpon<-length(unique(data$fponderacion))  
  
tabla1<-t(tibble(base\_no\_ponderada,base\_ponderada,unidades\_muestrales\_fex,unidades\_muestrales\_fpon,universo\_representado))  
tabla1

## [,1]  
## base\_no\_ponderada 32177  
## base\_ponderada 32177  
## unidades\_muestrales\_fex 3536  
## unidades\_muestrales\_fpon 3536  
## universo\_representado 8164164

Para trabajar con los datos ponderados se replica cada registro tantas unidades muestrales represente.

data\_ponderada<-uncount(data,weights = round(data$fponderacion))  
base\_ponderada<-length(data\_ponderada$directorio)  
base\_ponderada

## [1] 33466

base\_no\_ponderada

## [1] 32177

# Introduccion

El objetivo de este informe es encontrar los determinantes del salario de las personas encuestadas en Bogota y realizar una prediccion de este.

Por lo que ya sabemos podemos iniciar con un analisis de las siguientes variables:

* Edad : age
* Cuenta propia
* Estrato
* Formal
* Horas trabajadas: Actual hours worked previus week
* Usual weekly hours wworked
* Informal
* Ingreso total
* educacion: max. Education level attained
* Oficio
* educacion: p6210
* Experiencia: p6426
* Salario: p6500
* Sexo: sex
* Horas trabajadas: totalHoursWor
* Tamaño de la empresa

# Datos

# Limpieza de mayores de edad: Este chunk lo usamos solo como limpieza de datos   
  
  
  
  
  
data <- data %>%  
 rename(  
 Edad= age,  
 Estrato=estrato1,  
 Ingreso=ingtot,  
 Educacion=p6210,  
 Experiencia= p6426,  
 Salario=p6500,  
 Sexo=sex,  
 Horas\_trabajadas=totalHoursWorked  
 ) %>%  
   
 mutate(  
 across(c(cuentaPropia,Estrato,formal,informal,maxEducLevel,oficio,Educacion,Sexo),as.character)   
 )  
  
### Filtro por población economica activa  
  
data <- data %>%  
 filter(Edad >18 & dsi !="1")   
  
  
data <-data %>%  
 filter(p6240==1)

## AEDE individual

En este apartado observamos como se comportan las variables incluidas en el modelo , con el objetivo de analizar su distribucion e iniciar con el proceso de limpieza

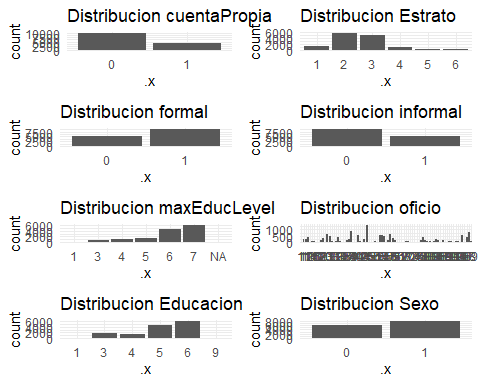
data\_h <- data %>%  
 select(Edad,cuentaPropia,Estrato,formal,informal,Ingreso,maxEducLevel,oficio,Educacion,Experiencia,Salario,Sexo,Horas\_trabajadas)

* Nota importante willian: Hay que discutir muy bien como vamos a trabajar con los factores de expansion.

str(data\_h)

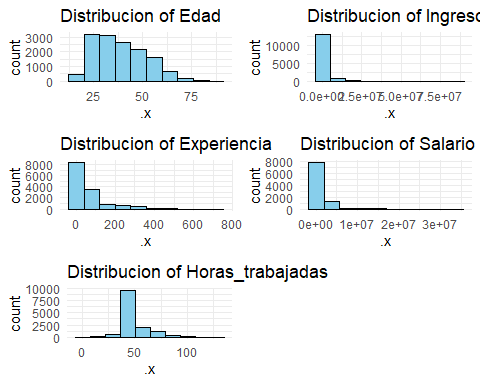
## tibble [13,993 × 13] (S3: tbl\_df/tbl/data.frame)  
## $ Edad : int [1:13993] 36 51 45 61 35 55 56 23 30 50 ...  
## $ cuentaPropia : chr [1:13993] "0" "0" "0" "0" ...  
## $ Estrato : chr [1:13993] "2" "2" "2" "2" ...  
## $ formal : chr [1:13993] "1" "1" "1" "0" ...  
## $ informal : chr [1:13993] "0" "0" "0" "1" ...  
## $ Ingreso : num [1:13993] 1840833 1878973 883357 1200000 1083000 ...  
## $ maxEducLevel : chr [1:13993] "6" "7" "4" "3" ...  
## $ oficio : chr [1:13993] "39" "85" "45" "53" ...  
## $ Educacion : chr [1:13993] "5" "6" "3" "3" ...  
## $ Experiencia : int [1:13993] 166 12 15 120 36 2 30 51 8 72 ...  
## $ Salario : num [1:13993] 1300000 1200000 737717 700000 950000 ...  
## $ Sexo : chr [1:13993] "1" "1" "1" "1" ...  
## $ Horas\_trabajadas: int [1:13993] 45 48 56 40 40 48 48 40 54 48 ...

character\_cols <- data\_h %>%  
 select(where(is.character))  
  
  
character\_plots <- map2(character\_cols, names(character\_cols), ~ ggplot(data\_h, aes(x = .x)) +  
 geom\_bar() +  
 ggtitle(paste("Distribucion", .y)) +   
 theme\_minimal() +  
 theme(plot.title = element\_text(color = "black", hjust = 0)))  
  
  
character\_grid <- grid.arrange(grobs = character\_plots, ncol = 2)



numeric\_cols <- data\_h %>%  
 select(where(is.numeric))  
  
numeric\_plots <- map2(numeric\_cols, names(numeric\_cols), ~ ggplot(data\_h, aes(x = .x)) +  
 geom\_histogram(fill = "skyblue", color = "black", bins = 10) +  
 ggtitle(paste("Distribucion of", .y)) + # Set the title to the column name  
 theme\_minimal() +  
 theme(plot.title = element\_text(color = "black", hjust = 0)))  
  
  
numeric\_grid <- grid.arrange(grobs = numeric\_plots, ncol = 2)

## Warning: Removed 4519 rows containing non-finite values (`stat\_bin()`).



# Analisis de regresion

EL proposito es estimar el siguiente modelo:

# Modelo de genero

# 5 Prediccion

El dia de hoy estime

# Bibliografía