Capacitación - Estadística Aplicada.

Pablo Herrera Gálvez

2025-01-02

Universidad de los Lagos

Sesión 3 - Clase 1 de programación con R

Sección 1. Análisis de población en países sudamericanos.

Códigos iniciales y creación de objetos.

```
# Para limpiar el ambiente
rm(list=ls())

# Creación de vectores
pais <- c("Chile", "Ecuador", "Bolivia", "Paraguay", "Uruguay")
poblacion <- c(20, 18, 12, 6, 3)

# Se muestran los dos vectores
pais
```

```
## [1] "Chile"    "Ecuador"  "Bolivia"  "Paraguay" "Uruguay"
```

```
poblacion
```

```
## [1] 20 18 12  6  3
```

```
# Creación de data frame con las variables
tpaises <- data.frame(pais, poblacion)

# Se muestra el data frame creado con dos vectores (variables)
print(tpaises)
```

```
##       pais poblacion
## 1    Chile        20
## 2  Ecuador        18
## 3  Bolivia        12
## 4 Paraguay         6
## 5  Uruguay         3
```

Estadísticos básicos

Media (o promedio).

```
mean(poblacion) # Cálculo
```

```
## [1] 11.8
```

```r
media_poblacion <- mean(poblacion) # Cálculo es almacenado en un objeto
media_poblacion # Se muestra el valor almacenado
```

```
## [1] 11.8
```

Mediana (o percentil 50).

```r
median(poblacion)
```

```
## [1] 12
```

Varianza (muestral).

```r
# Varianza
var(poblacion)
```

```
## [1] 54.2
```

Visualización

Gráfico de barras.

```r
barplot(poblacion, names.arg = pais)
```

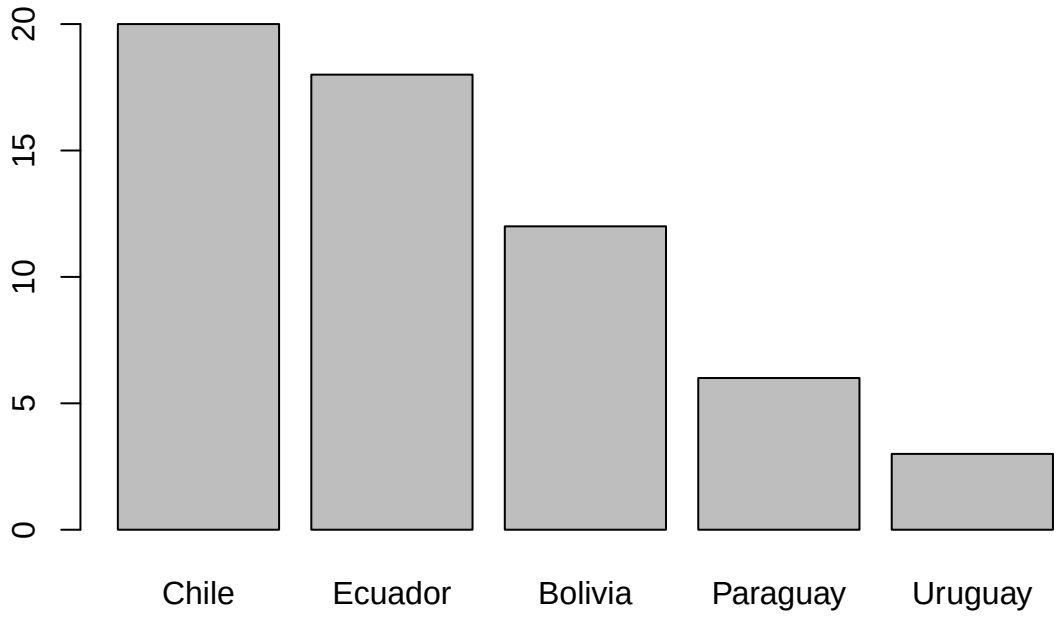
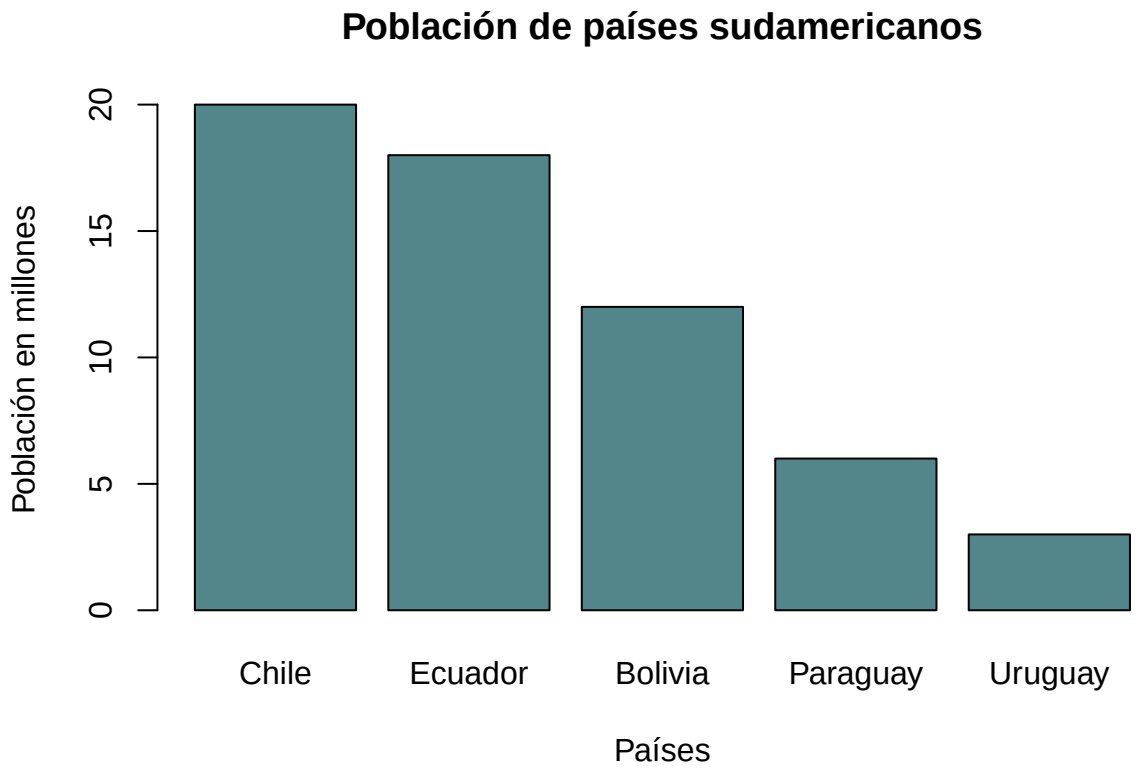


Gráfico de barras personalizado.

```r
barplot(poblacion, names.arg = pais,
        main = "Población de países sudamericanos",
        xlab = "Países",
```

```
        ylab = "Población en millones",
        col = "cadetblue4")
```



Para ver códigos de colores, visitar: https://r-charts.com/es/colores/

Sección 2. Rendimiento académico de estudiantes.

Códigos iniciales y creación de objetos.

```
# Se eliminan los elementos existentes en el ambiente
rm(list=ls())

# Creación de vectores
nota <- c(44, 32, 38, 33, 40, 32, 43, 62, 63, 23, 36, 30, 46, 24, 34, 35, 28, 56, 52, 68)
horas_estudiadas <- c(28, 22, 12, 8, 14, 2, 17, 22, 28, 2, 21, 10, 24, 10, 20, 5, 0, 23, 21, 45)

# Creación de data frame con las variables
tnotas <- data.frame(horas_estudiadas, nota)
tnotas # Revisión tabla
```

```
##    horas_estudiadas nota
## 1                28   44
## 2                22   32
## 3                12   38
## 4                 8   33
## 5                14   40
## 6                 2   32
```

```
## 7                17   43
## 8                22   62
## 9                28   63
## 10                2   23
## 11               21   36
## 12               10   30
## 13               24   46
## 14               10   24
## 15               20   34
## 16                5   35
## 17                0   28
## 18               23   56
## 19               21   52
## 20               45   68
```

```
tnotas <- data.frame(hrs = horas_estudiadas, not = nota) # Se asignan nuevos nombres a las columnas
tnotas # Revisión tabla
```

```
##    hrs not
## 1   28  44
## 2   22  32
## 3   12  38
## 4    8  33
## 5   14  40
## 6    2  32
## 7   17  43
## 8   22  62
## 9   28  63
## 10   2  23
## 11  21  36
## 12  10  30
## 13  24  46
## 14  10  24
## 15  20  34
## 16   5  35
## 17   0  28
## 18  23  56
## 19  21  52
## 20  45  68
```

Estadísticos básicos.

Medias.

```
mean(horas_estudiadas)
```

```
## [1] 16.7
```

```
mean(nota)
```

```
## [1] 40.95
```

```
media_horas <- mean(nota) # Podemos almacenar en objetos los valores calculados
media_nota <- mean(nota)
```

Varianzas.

```r
var(horas_estudiadas) # Ojo, esta función calcula la varianza muestral (no poblacional)
```

```
## [1] 120.8526
```

```r
var(nota)
```

```
## [1] 172.9974
```

Desviaciones estándar.

```r
# D.E. horas de estudio
sd(horas_estudiadas) # Fórmula de desviación estándar muestral
```

```
## [1] 10.9933
```

```r
sqrt(var(horas_estudiadas)) # Mismo resultado, pero calcula la raíz cuadrada de la varianza estimada en
```

```
## [1] 10.9933
```

```r
# D.E. notas
sd(nota)
```

```
## [1] 13.15285
```

```r
sqrt(var(nota))
```

```
## [1] 13.15285
```

Covarianza.

```r
var(horas_estudiadas, nota)
```

```
## [1] 115.9316
```

```r
cov(horas_estudiadas, nota)
```

```
## [1] 115.9316
```

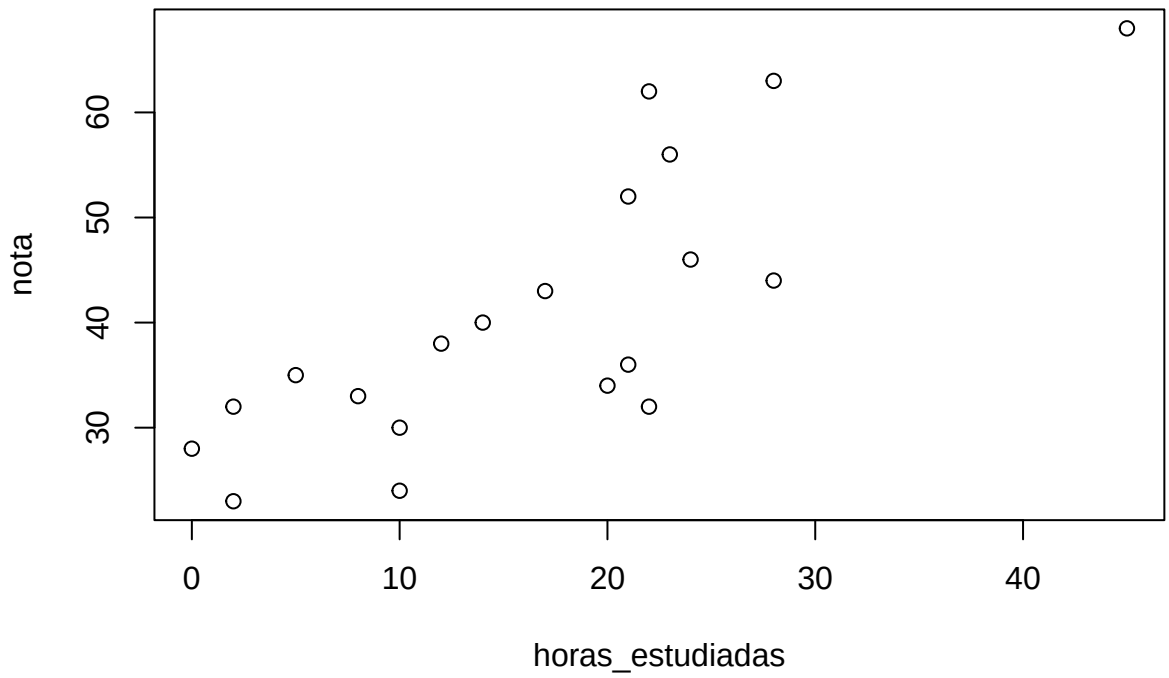
```r
# Correlación (de Pearson).
cor(horas_estudiadas, nota)
```

```
## [1] 0.8017777
```

Visualización.

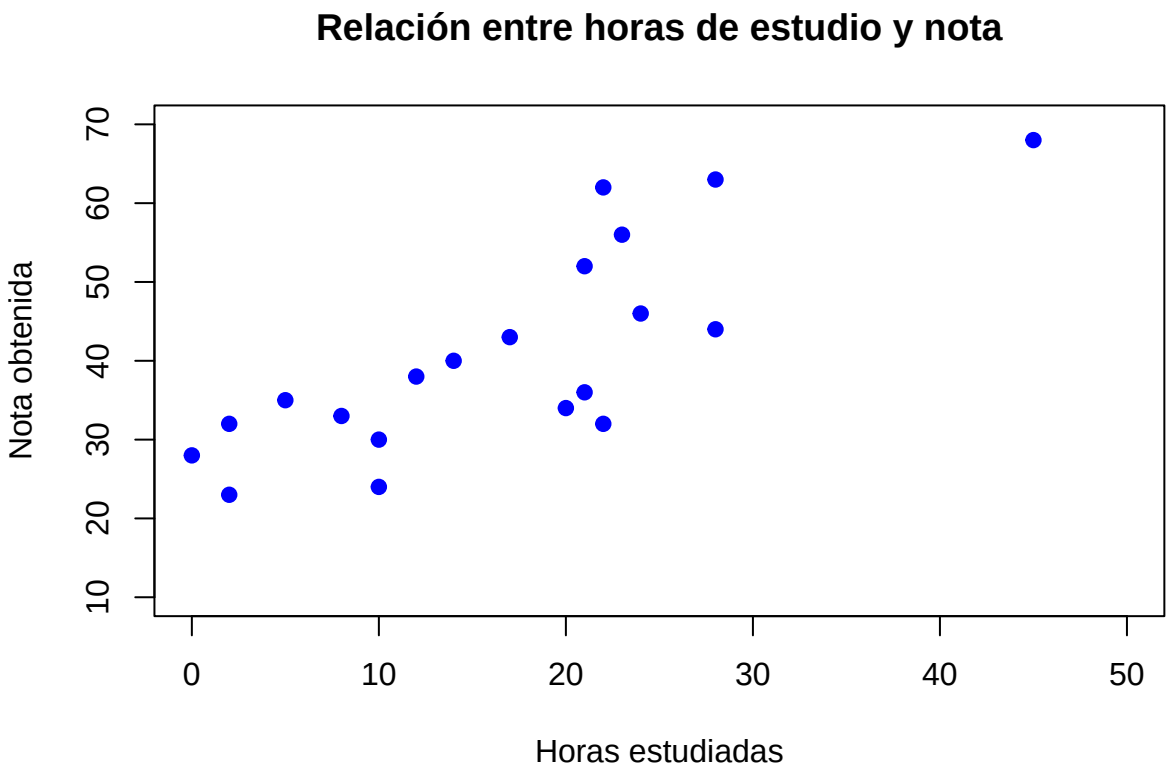
Scatter plot.

```r
plot(horas_estudiadas, nota)
```



Scatter plot personalizado.

```
plot(horas_estudiadas, nota,
     xlim = c(0,50),
     ylim = c(10,70),
     main = "Relación entre horas de estudio y nota",
     xlab = "Horas estudiadas",
     ylab = "Nota obtenida",
     col = "blue", # Color de los puntos
     pch = 19) # Tipo de punto (círculo sólido). PUEDEN JUGAR CON ESTAS ESPECIFICACIONES
```



Columnas adicionales.

Se crean dos nuevos vectores (comuna y sexo). Estos serán incorporados en el data frame y utilizados para profundizar el análisis.

```
# Definición de vectores Comuna y Sexo
comuna <- c("san_juan", "san_pablo", "san_juan", "san_juan", "san_juan", "purranque", "rio_negro", "oso
sexo <- c("mujer", "hombre", "hombre", "hombre", "mujer", "hombre", "hombre", "mujer", "hombre", "hombr

# Tablas de frecuencias de los dos nuevos vectores
table(comuna)
```

```
## comuna
##    osorno purranque rio_negro  san_juan san_pablo
##         5         2         2         7         4
```

```
table(sexo)
```

```
## sexo
## hombre  mujer
##     12      8
```

```
# Modificación del datafram existente (incluyendo las 4 variables)
tnotas <- data.frame(hrs = horas_estudiadas, not = nota, com = comuna, sex = sexo) # Se asignan nuevos
```

Análisis preliminar del dataframe.

Head arroja las primeras seis filas del dataframe

```r
head(tnotas)
```

```
##   hrs not       com    sex
## 1  28  44  san_juan  mujer
## 2  22  32 san_pablo hombre
## 3  12  38  san_juan hombre
## 4   8  33  san_juan hombre
## 5  14  40  san_juan  mujer
## 6   2  32 purranque hombre
```

Tail arroja las últimas seis filas del dataframe

```r
tail(tnotas)
```

```
##    hrs not       com    sex
## 15  20  34    osorno  mujer
## 16   5  35    osorno hombre
## 17   0  28 purranque hombre
## 18  23  56    osorno hombre
## 19  21  52 san_pablo  mujer
## 20  45  68  san_juan  mujer
```

Structure muestra el tipo de objeto que analizamos. Adicionalmente muestra las dimensiones (20 x 4), tipo de variables que existen, y los primeros datos de cada variable.

```r
str(tnotas)
```

```
## 'data.frame':    20 obs. of  4 variables:
##  $ hrs: num  28 22 12 8 14 2 17 22 28 2 ...
##  $ not: num  44 32 38 33 40 32 43 62 63 23 ...
##  $ com: chr  "san_juan" "san_pablo" "san_juan" "san_juan" ...
##  $ sex: chr  "mujer" "hombre" "hombre" "hombre" ...
```

Summary entrega estadística descriptiva rápida de las variables del objeto analizado. Si hay variables cuantitativas se entregan el mínimo, máximo, cuartiles y media.

```r
summary(tnotas)
```

```
##       hrs             not            com                sex
##  Min.   : 0.00   Min.   :23.00   Length:20          Length:20
##  1st Qu.: 9.50   1st Qu.:32.00   Class :character   Class :character
##  Median :18.50   Median :37.00   Mode  :character   Mode  :character
##  Mean   :16.70   Mean   :40.95
##  3rd Qu.:22.25   3rd Qu.:47.50
##  Max.   :45.00   Max.   :68.00
```

Summary también puede ser ejecutado para una sola variable del data frame.

```r
summary(tnotas$not)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23.00   32.00   37.00   40.95   47.50   68.00
```

Tablas de frecuencia.

Tablas de frecuencia separadas por UNA categoría.

```r
table(tnotas$sex)
```

```
##
```

```
## hombre  mujer
##     12      8
```

```
table(tnotas$com)
```

```
##
##    osorno purranque rio_negro  san_juan san_pablo
##         5         2         2         7         4
```

Addmargins, prop.table y round para una tabla mejorada.

```
addmargins(table(tnotas$com)) # Con "addmargins" se incluye un conteo del total
```

```
##
##    osorno purranque rio_negro  san_juan san_pablo       Sum
##         5         2         2         7         4        20
```

```
addmargins(prop.table(table(tnotas$com))) # Con "prop.table" se reemplazan las frecuencias por proporci
```

```
##
##    osorno purranque rio_negro  san_juan san_pablo       Sum
##      0.25      0.10      0.10      0.35      0.20      1.00
```

```
round(addmargins(prop.table(table(tnotas$com))), 2) # Con "round" se redondea a la cantidad de decimale
```

```
##
##    osorno purranque rio_negro  san_juan san_pablo       Sum
##      0.25      0.10      0.10      0.35      0.20      1.00
```

Tablas de frecuencia separadas por DOS categorías.

```
table(tnotas$sex,tnotas$com)
```

```
##
##          osorno purranque rio_negro san_juan san_pablo
##   hombre      2         2         1        4         3
##   mujer       3         0         1        3         1
```

```
table(tnotas$com,tnotas$sex) # Misma información pero intercambiando filas por columnas
```

```
##
##             hombre mujer
##   osorno         2     3
##   purranque      2     0
##   rio_negro      1     1
##   san_juan       4     3
##   san_pablo      3     1
```

Addmargins y prop.table

```
addmargins(table(tnotas$com,tnotas$sex)) # Addmargins para totales
```

```
##
##             hombre mujer Sum
##   osorno         2     3   5
##   purranque      2     0   2
##   rio_negro      1     1   2
##   san_juan       4     3   7
##   san_pablo      3     1   4
##   Sum           12     8  20
```

```r
prop.table(table(tnotas$com,tnotas$sex)) # prop.table para proporción
```

```
##
##             hombre mujer
##   osorno      0.10  0.15
##   purranque   0.10  0.00
##   rio_negro   0.05  0.05
##   san_juan    0.20  0.15
##   san_pablo   0.15  0.05
```

De proporción a porcentaje

```r
prop.table(table(tnotas$com,tnotas$sex))*100 # Multiplicamos *100 para pasarlo a porcentajes
```

```
##
##             hombre mujer
##   osorno        10    15
##   purranque     10     0
##   rio_negro      5     5
##   san_juan      20    15
##   san_pablo     15     5
```

Proporciones por FILAS.

```r
prop.table(table(tnotas$com,tnotas$sex), margin = 1)*100 # Porcentaje por filas
```

```
##
##                hombre     mujer
##   osorno     40.00000  60.00000
##   purranque 100.00000   0.00000
##   rio_negro  50.00000  50.00000
##   san_juan   57.14286  42.85714
##   san_pablo  75.00000  25.00000
```

```r
round(prop.table(table(tnotas$com,tnotas$sex), margin = 1)*100, 2) # Porcentaje por filas, redondeado a
```

```
##
##             hombre  mujer
##   osorno     40.00  60.00
##   purranque 100.00   0.00
##   rio_negro  50.00  50.00
##   san_juan   57.14  42.86
##   san_pablo  75.00  25.00
```

```r
prop_table_filas <- prop.table(table(tnotas$com,tnotas$sex), margin = 1) # Podemos almacenar este objet
addmargins(prop_table_filas, 2) # Aplicamos addmargins al objeto (tabla con proporciones por fila)
```

```
##
##                hombre     mujer       Sum
##   osorno    0.4000000 0.6000000 1.0000000
##   purranque 1.0000000 0.0000000 1.0000000
##   rio_negro 0.5000000 0.5000000 1.0000000
##   san_juan  0.5714286 0.4285714 1.0000000
##   san_pablo 0.7500000 0.2500000 1.0000000
```

Proporciones por COLUMNAS.

```r
prop.table(table(tnotas$com,tnotas$sex), margin = 2)*100 # Porcentaje por columnas
```

```
##
##                hombre     mujer
##   osorno    16.666667 37.500000
##   purranque 16.666667  0.000000
##   rio_negro  8.333333 12.500000
##   san_juan  33.333333 37.500000
##   san_pablo 25.000000 12.500000
```

```r
round(prop.table(table(tnotas$com,tnotas$sex), margin = 2)*100, 2) # Porcentaje por columnas, redondead
```

```
##
##             hombre mujer
##   osorno     16.67 37.50
##   purranque  16.67  0.00
##   rio_negro   8.33 12.50
##   san_juan   33.33 37.50
##   san_pablo  25.00 12.50
```

```r
prop_table_colum <- prop.table(table(tnotas$com,tnotas$sex), margin = 2) # Podemos almacenar este objet
addmargins(prop_table_colum, 1) # Aplicamos addmargins al objeto (tabla con proporciones por columna)
```

```
##
##                 hombre      mujer
##   osorno    0.16666667 0.37500000
##   purranque 0.16666667 0.00000000
##   rio_negro 0.08333333 0.12500000
##   san_juan  0.33333333 0.37500000
##   san_pablo 0.25000000 0.12500000
##   Sum       1.00000000 1.00000000
```

Análisis estadístico separando por categorías.

Media (agregada y agrupada por categorías).

```r
mean(tnotas$not)
```

```
## [1] 40.95
```

```r
tapply(tnotas$not, tnotas$sex, mean) # tapply permite calcular estadísticos como la media según una VAR
```

```
## hombre  mujer
##  38.25  45.00
```

```r
tapply(tnotas$not, tnotas$com, mean) # media de NOTAS según COMUNA
```

```
##    osorno purranque rio_negro  san_juan san_pablo
##  44.60000  30.00000  33.50000  47.42857  34.25000
```

Varianzas.

```r
tapply(tnotas$not, tnotas$sex, var) # varianza de NOTAS según SEXO
```

```
##   hombre    mujer
## 139.2955 219.4286
```

```r
tapply(tnotas$not, tnotas$com, var) # varianza de NOTAS según COMUNA
```

```
##    osorno purranque rio_negro  san_juan san_pablo
```

```
##  177.8000    8.0000  180.5000  171.9524  154.9167
```

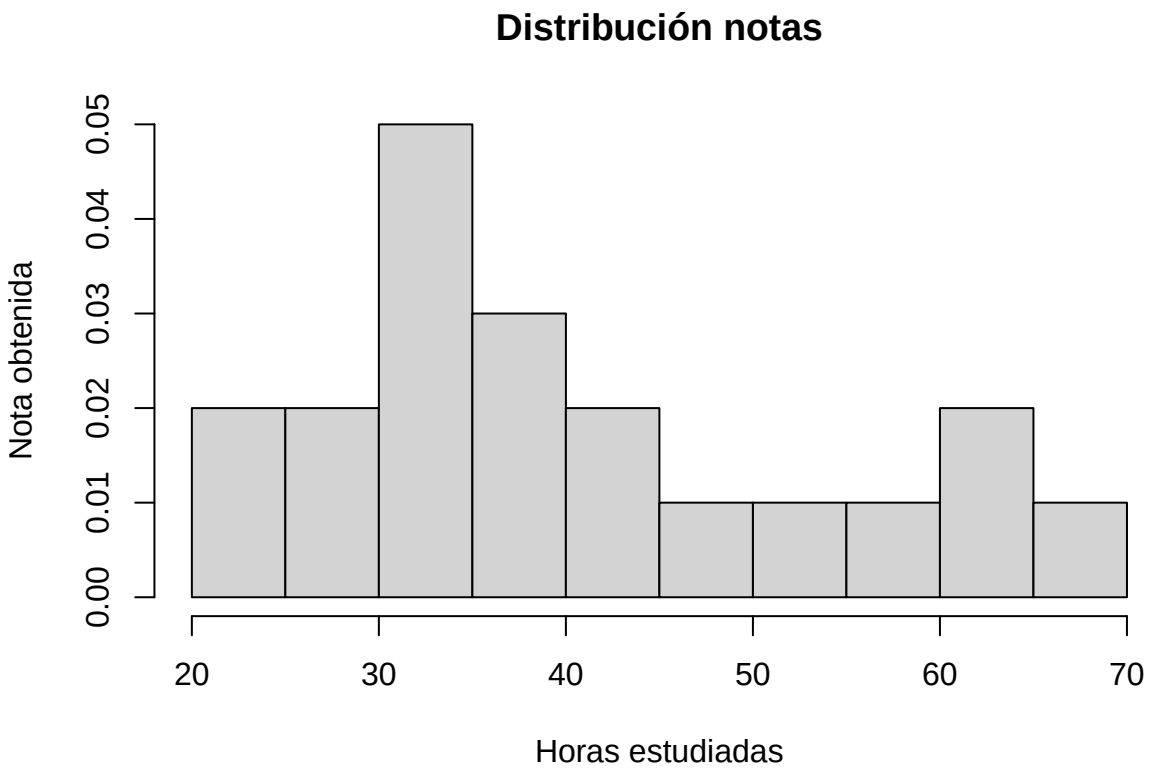
Visualizaciones separando por categorías

Histograma de notas para el total de estudiantes.

```
hist(tnotas$not)
```

Histograma de notas para el total de estudiantes personalizada.

```
hist(tnotas$not,
     breaks = 10,
     freq = F,
     main = "Distribución notas",
     xlab = "Horas estudiadas",
     ylab = "Nota obtenida")
```



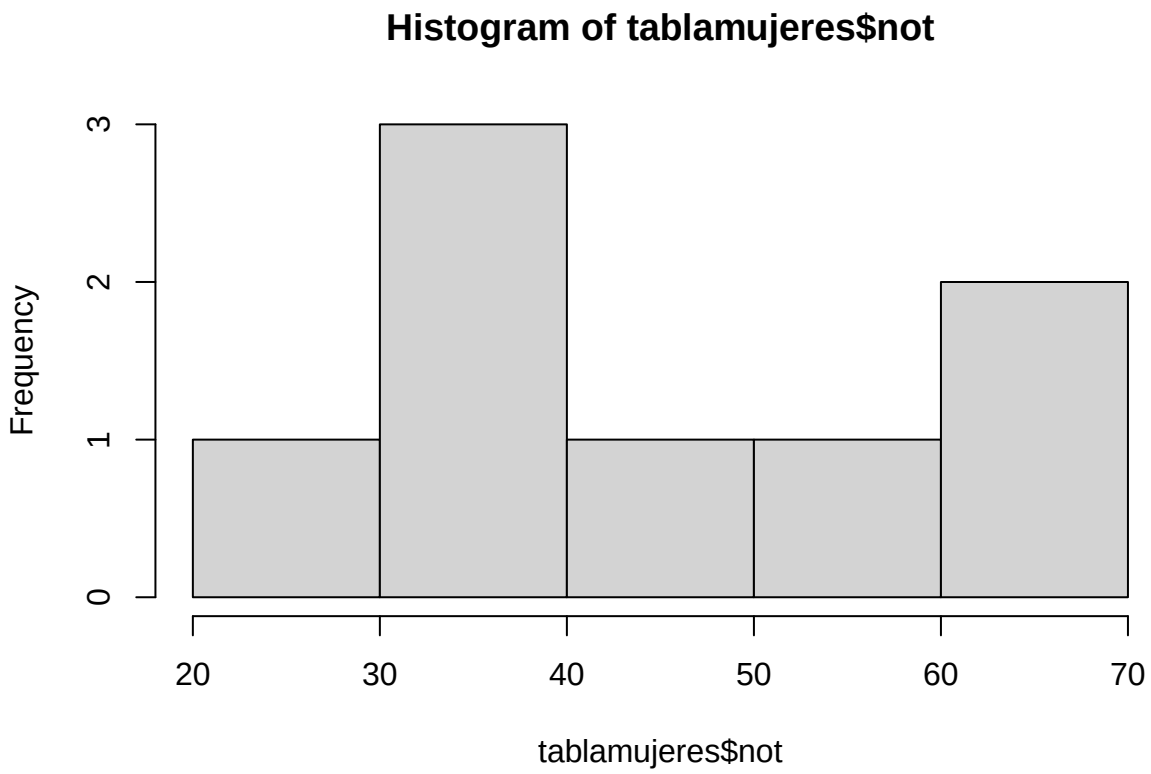
Creación sub sets (mini dataframes) para un análisis por separado.

```
tablahombres <- subset(tnotas, sex == "hombre") # Data frame solo de HOMRES
tablamujeres <- subset(tnotas, sex == "mujer") # Data frame solo de MUJERES
```

Histogramas de notas según SEXO.

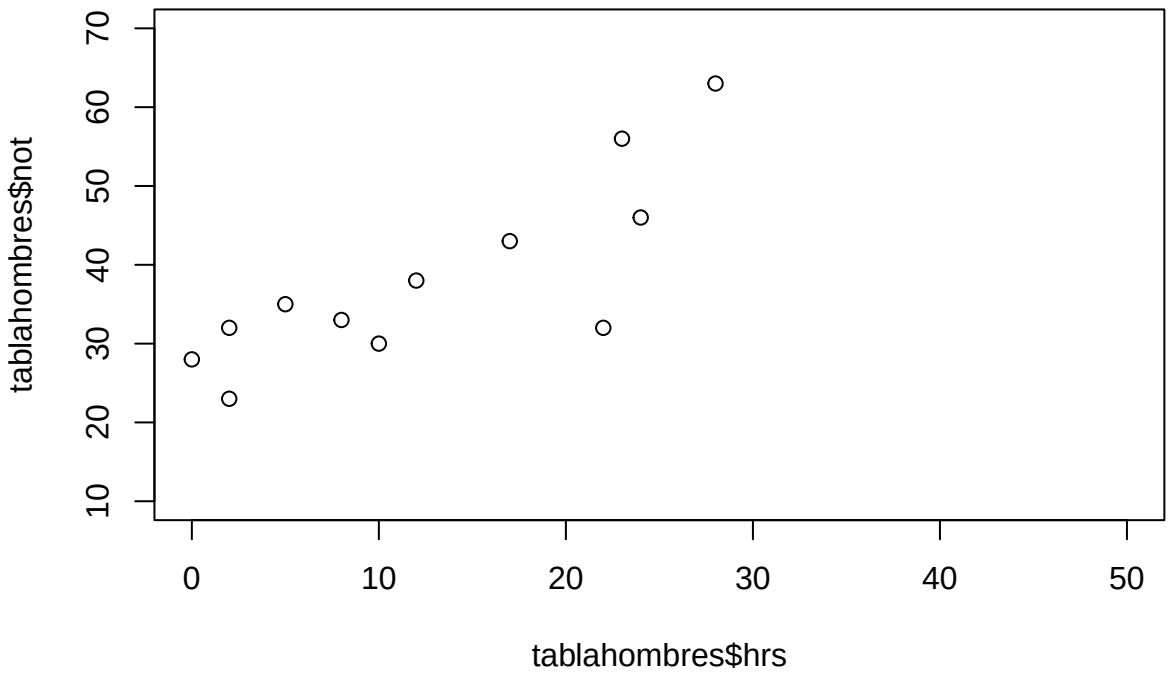
```
hist(tablahombres$not) # Histograma de notas en HOMBRES
```


```r
hist(tablamujeres$not) # Histograma de notas en MUJERES
```

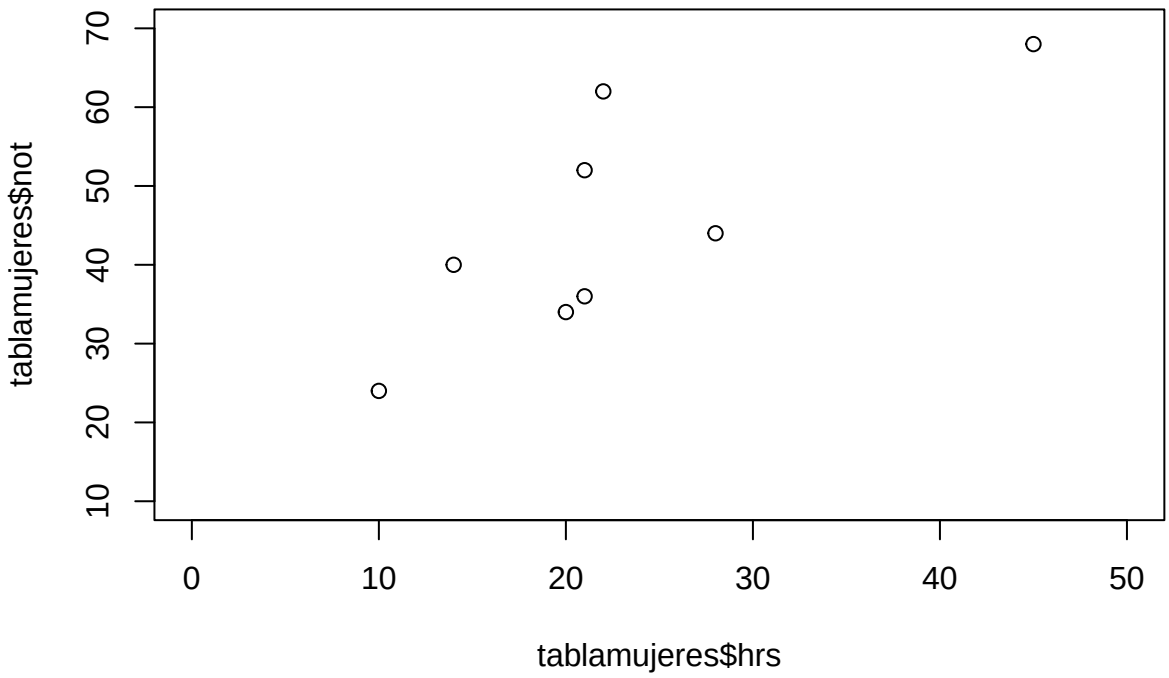


Scatter plot por SEXO.

```
plot(tablahombres$hrs, tablahombres$not, # Scatter plot de HOMBRES
     xlim = c(0,50),
     ylim = c(10,70))
```




```
plot(tablamujeres$hrs, tablamujeres$not, # Scatter plot de MUJERES
     xlim = c(0,50),
     ylim = c(10,70))
```




```
# Scatter plot agregado (HOMBRES Y MUJERES EN CONJUNTO)
colors <- ifelse(tnotas$sex == "mujer", "red", "blue")
plot(tnotas$hrs, tnotas$not,
     col = colors,
     pch = 19)
legend("bottomright", legend = c("Mujeres", "Hombres"), col = c("red", "blue"), pch = 19)
```

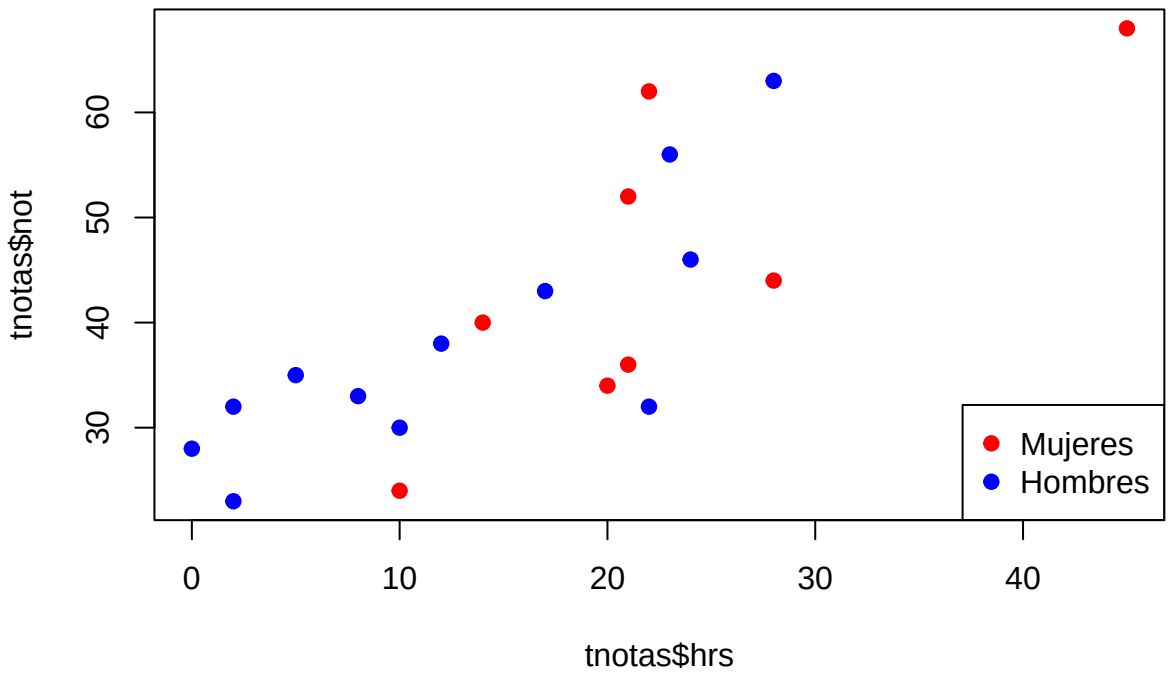


Gráfico de barras para hombres.

```
contador_comuna_hom <- table(tablahombres$com) # Conteo de comunas de HOMBRES. Será insumo para el gráf

barplot(contador_comuna_hom, main = "Comunas de origen de estudiantes hombres", # Gráfico para HOMBRES
        xlab = "Comuna", ylab = "Frecuencia", ylim = c(0,5))
```

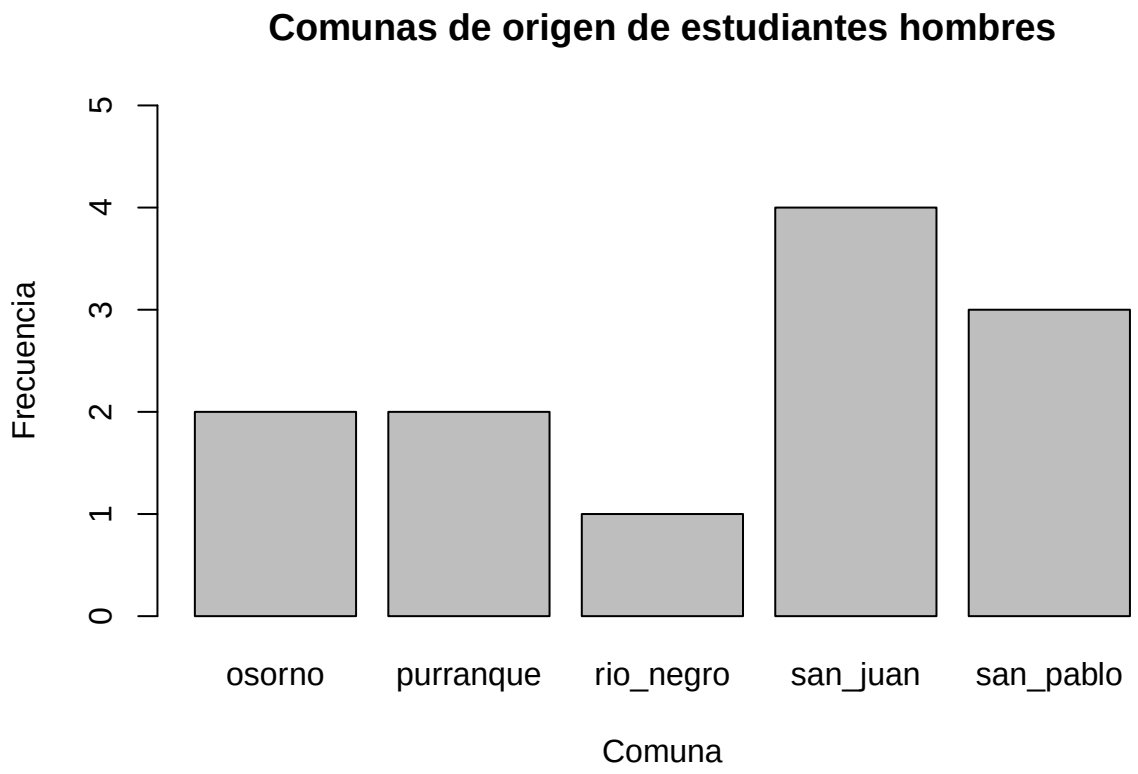
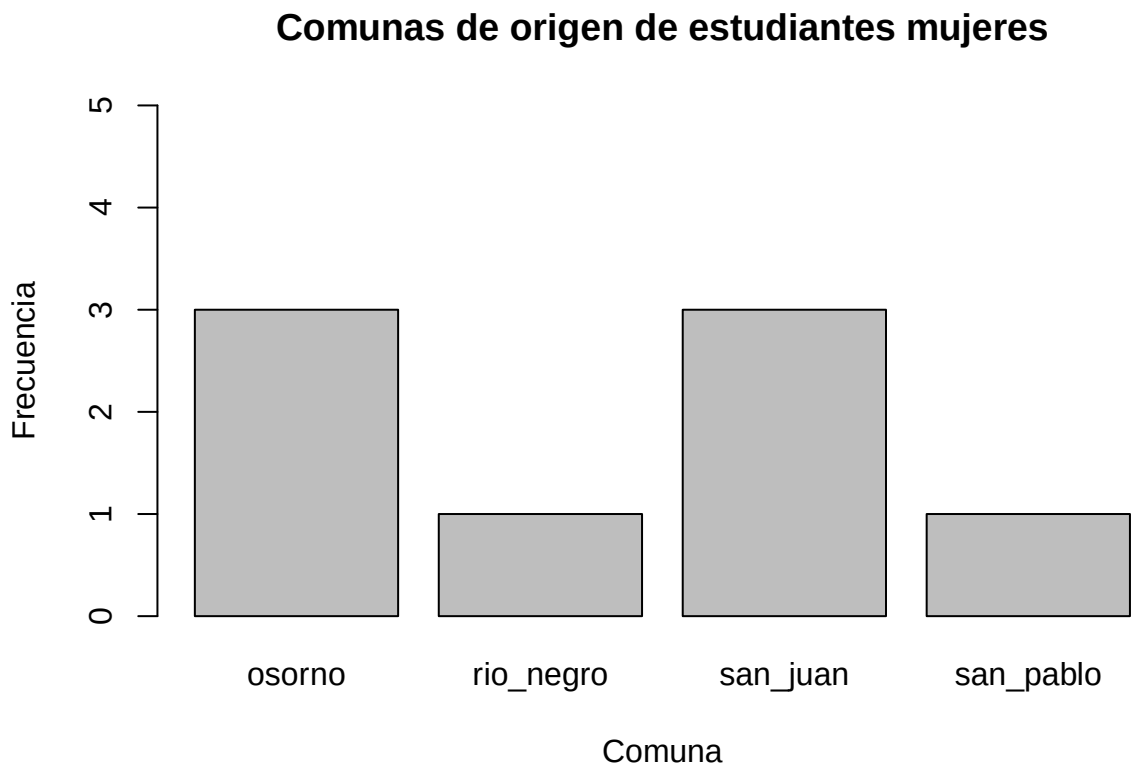


Gráfico de barras para mujeres.

```
contador_comuna_muj <- table(tablamujeres$com) # Conteo de comunas de MUJERES. Será insumo para el gráf

barplot(contador_comuna_muj, main = "Comunas de origen de estudiantes mujeres", # Gráfico para MUJERES
        xlab = "Comuna", ylab = "Frecuencia", ylim = c(0,5))
```



Sección 3. Análisis de una base de datos.

Códigos iniciales.

```
# Se eliminan los elementos existentes
rm(list=ls())
```

Carga de una base de datos.

- Paquetes (instalación y carga).
- Working directory
- Se importa la misma base de datos (desde una tabla EXCEL y desde un CSV).

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

#Se define el working directory
setwd("C:/00 Pablo/00 Universidad de los Lagos/Clases R")
# Se consulta el working directory
#getwd()

tabla_excel <- read_excel("Ejercicio Autos.xlsx")
tabla_csv <- read.csv("Ejercicio Autos.csv", sep = ";", dec = ".")
```

Se exploran ambos data frame y se verifica que ambos tienen la misma información (es lo mismo).

```r
# HEAD
head(tabla_excel)
```

```
## # A tibble: 6 x 12
##   fabricante pais  modelo cilindrada  anio cilindros transmision traccion ciudad
##   <chr>      <chr> <chr>  <chr>      <dbl>     <dbl> <chr>       <chr>     <dbl>
## 1 audi       no j~ a4     1.8         1999         4 auto(l5)    d            18
## 2 audi       no j~ a4     1.8         1999         4 manual(m5)  d            21
## 3 audi       no j~ a4     2           2008         4 manual(m6)  d            20
## 4 audi       no j~ a4     2           2008         4 auto(av)    d            21
## 5 audi       no j~ a4     2.8         1999         6 auto(l5)    d            16
## 6 audi       no j~ a4     2.8         1999         6 manual(m5)  d            18
## # i 3 more variables: autopista <dbl>, combustible <chr>, clase <chr>
```

```r
head(tabla_csv)
```

```
##   fabricante     pais modelo cilindrada anio cilindros transmision traccion
## 1       audi no japon     a4        1.8 1999         4    auto(l5)        d
## 2       audi no japon     a4        1.8 1999         4  manual(m5)        d
## 3       audi no japon     a4        2.0 2008         4  manual(m6)        d
## 4       audi no japon     a4        2.0 2008         4    auto(av)        d
## 5       audi no japon     a4        2.8 1999         6    auto(l5)        d
## 6       audi no japon     a4        2.8 1999         6  manual(m5)        d
##   ciudad autopista combustible    clase
## 1     18        29           p compacto
## 2     21        29           p compacto
## 3     20        31           p compacto
## 4     21        30           p compacto
## 5     16        26           p compacto
## 6     18        26           p compacto
```

```r
# STRUCTURE
str(tabla_excel)
```

```
## tibble [234 x 12] (S3: tbl_df/tbl/data.frame)
##  $ fabricante : chr [1:234] "audi" "audi" "audi" "audi" ...
##  $ pais       : chr [1:234] "no japon" "no japon" "no japon" "no japon" ...
##  $ modelo     : chr [1:234] "a4" "a4" "a4" "a4" ...
##  $ cilindrada : chr [1:234] "1.8" "1.8" "2" "2" ...
##  $ anio       : num [1:234] 1999 1999 2008 2008 1999 ...
##  $ cilindros  : num [1:234] 4 4 4 4 6 6 6 4 4 4 ...
##  $ transmision: chr [1:234] "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ traccion   : chr [1:234] "d" "d" "d" "d" ...
##  $ ciudad     : num [1:234] 18 21 20 21 16 18 18 18 16 20 ...
##  $ autopista  : num [1:234] 29 29 31 30 26 26 27 26 25 28 ...
##  $ combustible: chr [1:234] "p" "p" "p" "p" ...
##  $ clase      : chr [1:234] "compacto" "compacto" "compacto" "compacto" ...
```

```r
str(tabla_csv)
```

```
## 'data.frame':    234 obs. of  12 variables:
##  $ fabricante : chr  "audi" "audi" "audi" "audi" ...
##  $ pais       : chr  "no japon" "no japon" "no japon" "no japon" ...
##  $ modelo     : chr  "a4" "a4" "a4" "a4" ...
##  $ cilindrada : num  1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ anio       : int  1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
```

```
##  $ cilindros  : int  4 4 4 4 6 6 6 4 4 4 ...
##  $ transmision: chr  "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ traccion   : chr  "d" "d" "d" "d" ...
##  $ ciudad     : int  18 21 20 21 16 18 18 18 16 20 ...
##  $ autopista  : int  29 29 31 30 26 26 27 26 25 28 ...
##  $ combustible: chr  "p" "p" "p" "p" ...
##  $ clase      : chr  "compacto" "compacto" "compacto" "compacto" ...
```

```
# VARIABLES NAMES
variable.names(tabla_excel)
```

```
##  [1] "fabricante"  "pais"        "modelo"      "cilindrada"  "anio"
##  [6] "cilindros"   "transmision" "traccion"    "ciudad"      "autopista"
## [11] "combustible" "clase"
```

```
variable.names(tabla_csv)
```

```
##  [1] "fabricante"  "pais"        "modelo"      "cilindrada"  "anio"
##  [6] "cilindros"   "transmision" "traccion"    "ciudad"      "autopista"
## [11] "combustible" "clase"
```

```
# SUMMARY DEL DATA FRAME
summary(tabla_excel)
```

```
##   fabricante            pais              modelo           cilindrada
##  Length:234         Length:234         Length:234         Length:234
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##       anio        cilindros     transmision          traccion
##  Min.   :1999   Min.   :4.000   Length:234         Length:234
##  1st Qu.:1999   1st Qu.:4.000   Class :character   Class :character
##  Median :2004   Median :6.000   Mode  :character   Mode  :character
##  Mean   :2004   Mean   :5.889
##  3rd Qu.:2008   3rd Qu.:8.000
##  Max.   :2008   Max.   :8.000
##      ciudad        autopista     combustible           clase
##  Min.   : 9.00   Min.   :12.00   Length:234         Length:234
##  1st Qu.:14.00   1st Qu.:18.00   Class :character   Class :character
##  Median :17.00   Median :24.00   Mode  :character   Mode  :character
##  Mean   :16.86   Mean   :23.44
##  3rd Qu.:19.00   3rd Qu.:27.00
##  Max.   :35.00   Max.   :44.00
```

```
summary(tabla_csv)
```

```
##   fabricante            pais              modelo            cilindrada
##  Length:234         Length:234         Length:234         Min.   :1.600
##  Class :character   Class :character   Class :character   1st Qu.:2.400
##  Mode  :character   Mode  :character   Mode  :character   Median :3.300
##                                                           Mean   :3.472
##                                                           3rd Qu.:4.600
##                                                           Max.   :7.000
##       anio        cilindros     transmision          traccion
##  Min.   :1999   Min.   :4.000   Length:234         Length:234
```

```
##  1st Qu.:1999   1st Qu.:4.000   Class :character   Class :character
##  Median :2004   Median :6.000   Mode  :character   Mode  :character
##  Mean   :2004   Mean   :5.889
##  3rd Qu.:2008   3rd Qu.:8.000
##  Max.   :2008   Max.   :8.000
##      ciudad        autopista     combustible           clase
##  Min.   : 9.00   Min.   :12.00   Length:234         Length:234
##  1st Qu.:14.00   1st Qu.:18.00   Class :character   Class :character
##  Median :17.00   Median :24.00   Mode  :character   Mode  :character
##  Mean   :16.86   Mean   :23.44
##  3rd Qu.:19.00   3rd Qu.:27.00
##  Max.   :35.00   Max.   :44.00
```

```
# SUMMARY DE UNA VARIABLE
summary(tabla_excel$ciudad)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   14.00   17.00   16.86   19.00   35.00
```

```
summary(tabla_csv$ciudad)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   14.00   17.00   16.86   19.00   35.00
```

Comando REMOVE. Dado que ambas tablas son iguales, borramos una de ellas y continuamos con la otra.

```
rm(tabla_csv)
```

Estadística descriptiva.

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

```
##
## Adjuntando el paquete: 'dplyr'
```

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

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

```
# Crear tabla descriptiva
tabla_descriptiva <- tabla_excel %>%
  group_by(pais) %>%
  summarise(
    N = n(),                                    # Tamaño del grupo
    Porcentaje = round((N / sum(N)) * 100, 2),  # Porcentaje del total
    Media = mean(ciudad, na.rm = TRUE), # Media
    Mediana = median(ciudad, na.rm = TRUE), # Mediana
    Varianza = var(ciudad, na.rm = TRUE) # Varianza
  )

# # Mostrar la tabla
print(tabla_descriptiva)
```

```
## # A tibble: 2 x 6
```

```
##   pais         N Porcentaje Media Mediana Varianza
##   <chr>    <int>      <dbl> <dbl>   <dbl>    <dbl>
## 1 japon       70        100  19.4      19     14.5
## 2 no japon   164        100  15.8      15     15.9
```

```r
# T Test

ttest_rendimiento_ciudad <- t.test(ciudad ~ pais, data = tabla_excel)
ttest_rendimiento_autopista <- t.test(autopista ~ pais, data = tabla_excel)

print(ttest_rendimiento_ciudad)
```

```
##
##  Welch Two Sample t-test
##
## data:  ciudad by pais
## t = 6.4627, df = 136.24, p-value = 1.691e-09
## alternative hypothesis: true difference in means between group japon and group no japon is not equal
## 95 percent confidence interval:
##  2.473764 4.655156
## sample estimates:
##    mean in group japon mean in group no japon
##               19.35714               15.79268
```

```r
print(ttest_rendimiento_autopista)
```

```
##
##  Welch Two Sample t-test
##
## data:  autopista by pais
## t = 4.5159, df = 137.47, p-value = 1.344e-05
## alternative hypothesis: true difference in means between group japon and group no japon is not equal
## 95 percent confidence interval:
##  2.030227 5.193118
## sample estimates:
##    mean in group japon mean in group no japon
##               25.97143               22.35976
```