

Importación de datos

Utilizando las opciones de importación de RStudio, selecciono el tipo de fichero que quiero importar, en este caso, un fichero .sav de SPSS. Copio y pego el código generado para que la próxima vez ocurra de manera automática y miro las seis primeras líneas del fichero.

```
setwd("~/R/introstatsconr")
library(haven)
empleados <- read_sav("instrostat-v2/data/EMPLEADOS.sav")

head(empleados)

## # A tibble: 6 x 17
##   id sexo   fechnac      educ  catlab salario salini tiempemp expprev
##   <dbl> <chr+1> <date>      <dbl+1> <dbl+1> <dbl+1> <dbl+> <dbl+1b> <dbl+1>
## 1     1  h [Hom~ 1952-02-03 15 [15] 3 [Dir~   57000   27000      98     144
## 2     2  h [Hom~ 1958-05-23 16 [16] 1 [Adm~   40200   18750      98      36
## 3     3  h [Hom~ 1955-02-09 15 [15] 1 [Adm~   45000   21000      98     138
## 4     4  h [Hom~ 1958-08-22 15 [15] 1 [Adm~   32100   13500      98      67
## 5     5  h [Hom~ 1956-04-26 15 [15] 1 [Adm~   36000   18750      98     114
## 6     6 15 h [Hom~ 1962-08-29 12 [12] 1 [Adm~   27300   13500      97      66
## # ... with 8 more variables: minoría <dbl+1b>, salinico <dbl>,
## #   sexo_rec <dbl+1b>, PRE_1 <dbl>, RES_1 <dbl>, PRE_2 <dbl>,
## #   RES_2 <dbl>, `filter_$` <dbl+1b>
```

Exploración de datos

Vamos a comprobar ahora el número de variables y casos que tenemos, así como si tenemos casos incompletos o no.

En primer lugar, comprobamos si se trata de un data.frame, que es el tipo de objeto que necesitaremos para nuestros análisis.

```
is.data.frame(empleados)
```

```
## [1] TRUE
```

Ahora vemos los tipos de variables que tiene:

```
str(empleados)

## Classes 'tbl_df', 'tbl' and 'data.frame':   474 obs. of  17 variables:
##  $ id      : num  1 2 5 6 7 15 16 17 18 19 ...
##  ..- attr(*, "label")= chr "Código de empleado"
##  ..- attr(*, "format.spss")= chr "F4.0"
##  $ sexo    : 'haven_labelled' chr  "h" "h" "h" "h" ...
##  ..- attr(*, "label")= chr "Sexo"
##  ..- attr(*, "format.spss")= chr "A1"
##  ..- attr(*, "display_width")= int 5
##  ..- attr(*, "labels")= Named chr  "h" "m"
##  .. ..- attr(*, "names")= chr  "Hombre" "Mujer"
##  $ fechnac : Date, format: "1952-02-03" "1958-05-23" ...
##  $ educ    : 'haven_labelled' num  15 16 15 15 15 12 12 15 16 12 ...
##  ..- attr(*, "label")= chr "Nivel educativo"
##  ..- attr(*, "format.spss")= chr "F2.0"
##  ..- attr(*, "labels")= Named num  0 8 12 14 15 16 17 18 19 20 ...
```

```

## ..- attr(*, "names")= chr "0 (Ausente)" "8" "12" "14" ...
## $ catlab : 'haven_labelled' num 3 1 1 1 1 1 1 1 3 1 ...
## ..- attr(*, "label")= chr "Categoría laboral"
## ..- attr(*, "format.spss")= chr "F1.0"
## ..- attr(*, "labels")= Named num 0 1 2 3
## ..- attr(*, "names")= chr "0 (Ausente)" "Administrativo" "Seguridad" "Directivo"
## $ salario : 'haven_labelled' num 57000 40200 45000 32100 36000 ...
## ..- attr(*, "label")= chr "Salario actual"
## ..- attr(*, "format.spss")= chr "DOLLAR8"
## ..- attr(*, "labels")= Named num 0
## ..- attr(*, "names")= chr "Ausente"
## $ salini : 'haven_labelled' num 27000 18750 21000 13500 18750 ...
## ..- attr(*, "label")= chr "Salario inicial"
## ..- attr(*, "format.spss")= chr "DOLLAR8"
## ..- attr(*, "display_width")= int 6
## ..- attr(*, "labels")= Named num 0
## ..- attr(*, "names")= chr "Ausente"
## $ tiempemp: 'haven_labelled' num 98 98 98 98 98 97 97 97 97 97 ...
## ..- attr(*, "label")= chr "Meses desde el contrato"
## ..- attr(*, "format.spss")= chr "F2.0"
## ..- attr(*, "labels")= Named num 0
## ..- attr(*, "names")= chr "Ausente"
## $ expprev : 'haven_labelled' num 144 36 138 67 114 66 24 48 70 103 ...
## ..- attr(*, "label")= chr "Experiencia previa (meses)"
## ..- attr(*, "format.spss")= chr "F6.0"
## ..- attr(*, "labels")= Named num 0
## ..- attr(*, "names")= chr "Ausente"
## $ minoría : 'haven_labelled' num 0 0 0 0 0 0 0 0 0 0 ...
## ..- attr(*, "label")= chr "Clasificación de minorías"
## ..- attr(*, "format.spss")= chr "F1.0"
## ..- attr(*, "labels")= Named num 0 1 9
## ..- attr(*, "names")= chr "No" "Sí" "9 (Ausente)"
## $ salinico: num 53730 37313 41790 26865 37313 ...
## ..- attr(*, "label")= chr "Salario inicial corregido"
## ..- attr(*, "format.spss")= chr "DOLLAR8"
## ..- attr(*, "display_width")= int 10
## $ sexo_rec: 'haven_labelled' num 1 1 1 1 1 1 1 1 1 1 ...
## ..- attr(*, "format.spss")= chr "F8.2"
## ..- attr(*, "display_width")= int 10
## ..- attr(*, "labels")= Named num 1 2
## ..- attr(*, "names")= chr "HOMBRE" "MUJER"
## $ PRE_1 : num 53483 37730 42027 27706 37730 ...
## ..- attr(*, "label")= chr "Unstandardized Predicted Value"
## ..- attr(*, "format.spss")= chr "F11.5"
## ..- attr(*, "display_width")= int 13
## $ RES_1 : num 3517 2470 2973 4394 -1730 ...
## ..- attr(*, "label")= chr "Unstandardized Residual"
## ..- attr(*, "format.spss")= chr "F11.5"
## ..- attr(*, "display_width")= int 13
## $ PRE_2 : num 53769 38587 42727 28926 38587 ...
## ..- attr(*, "label")= chr "Unstandardized Predicted Value"
## ..- attr(*, "format.spss")= chr "F11.5"
## ..- attr(*, "display_width")= int 13
## $ RES_2 : num 3231 1613 2273 3174 -2587 ...

```

```
##   ..- attr(*, "label")= chr "Unstandardized Residual"
##   ..- attr(*, "format.spss")= chr "F11.5"
##   ..- attr(*, "display_width")= int 13
## $ filter_$: 'haven_labelled' num  1 0 0 0 0 0 0 0 1 0 ...
##   ..- attr(*, "label")= chr "catlab = 3 (FILTER)"
##   ..- attr(*, "format.spss")= chr "F1.0"
##   ..- attr(*, "display_width")= int 10
##   ..- attr(*, "labels")= Named num  0 1
##   .. ..- attr(*, "names")= chr  "No seleccionado" "Seleccionado"
## - attr(*, "label")= chr "05.00.00"
```

Al tratarse de datos extraídos de SPSS, las variables están etiquetadas con la definición de cada una de ellas. Vamos a contar el número de observaciones incompletas:

```
sum(!complete.cases(empleados))
```

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

Para poder visualizarla, tendremos que filtrar utilizando la función `subset()`.

```
incompleto <- subset(empleados, !complete.cases(empleados))
incompleto
```

```
## # A tibble: 1 x 17
##   id sexo fechnac educ catlab salario salini tiempemp expprev
##   <dbl> <chr+1> <date>   <dbl+1> <dbl+1> <dbl+1> <dbl+> <dbl+1b> <dbl+1>
## 1  434 h [Hom~ NA      16 [16] 1 [Adm~ 34950 20250      66      55
## # ... with 8 more variables: minoría <dbl+1b1>, salinico <dbl>,
## #   sexo_rec <dbl+1b1>, PRE_1 <dbl>, RES_1 <dbl>, PRE_2 <dbl>,
## #   RES_2 <dbl>, `filter_$` <dbl+1b1>
```

Vamos a sacar estadísticas descriptivas de cada una de las variables de nuestro set de datos. Aquí también indica el número de casos incompletos.

```
summary(empleados)
```

```
##           id           sexo           fechnac
## Min.      : 1.0   Length:474   Min.      :1929-02-10
## 1st Qu.:119.2   Class :haven_labelled 1st Qu.:1948-01-03
## Median :237.5   Mode  :character   Median :1962-01-23
## Mean     :237.5                      Mean   :1956-10-08
## 3rd Qu.:355.8                      3rd Qu.:1965-07-06
## Max.     :474.0                      Max.   :1971-02-10
##                                     NA's    :1
##           educ           catlab           salario           salini
## Min.      : 8.00   Min.      :1.000   Min.      : 15750   Min.      : 9000
## 1st Qu.:12.00   1st Qu.:1.000   1st Qu.: 24000   1st Qu.:12488
## Median :12.00   Median :1.000   Median : 28875   Median :15000
## Mean     :13.49   Mean     :1.411   Mean     : 34420   Mean     :17016
## 3rd Qu.:15.00   3rd Qu.:1.000   3rd Qu.: 36938   3rd Qu.:17490
## Max.     :21.00   Max.     :3.000   Max.     :135000   Max.     :79980
##
##           tiempemp           expprev           minoría           salinico
## Min.      :63.00   Min.      : 0.00   Min.      :0.0000   Min.      : 17910
## 1st Qu.:72.00   1st Qu.: 19.25   1st Qu.:0.0000   1st Qu.: 24850
## Median :81.00   Median : 55.00   Median :0.0000   Median : 29850
## Mean     :81.11   Mean      : 95.86   Mean     :0.2194   Mean     : 33862
## 3rd Qu.:90.00   3rd Qu.:138.75   3rd Qu.:0.0000   3rd Qu.: 34805
```

```

## Max.      :98.00    Max.      :476.00    Max.      :1.0000    Max.      :159160
##
##      sexo_rec      PRE_1      RES_1      PRE_2
## Min.      :1.000    Min.      : 19113    Min.      :-35424    Min.      : 18033
## 1st Qu.:1.000    1st Qu.: 25772    1st Qu.: -4031    1st Qu.: 25363
## Median :1.000    Median : 30570    Median : -1154    Median : 31686
## Mean      :1.456    Mean      : 34420    Mean      :      0    Mean      : 34420
## 3rd Qu.:2.000    3rd Qu.: 35324    3rd Qu.:  2584    3rd Qu.: 35627
## Max.      :2.000    Max.      :154646    Max.      : 49293    Max.      :151263
##
##      RES_2      filter_$
## Min.      :-33944    Min.      :0.0000
## 1st Qu.: -3875    1st Qu.:0.0000
## Median : -1124    Median :0.0000
## Mean      :      0    Mean      :0.1772
## 3rd Qu.:  2696    3rd Qu.:0.0000
## Max.      : 49043    Max.      :1.0000
##

```