Estado de Salud por Paises y por Sexo

Alicia Perdices Guerra

3 de mayo, 2021

Contents

1.PROCESAMIENTO DE LOS DATOS.

- En primer lugar leemos el fichero:

```
salud<-read.csv("C:/temp/EstadoDeSalud_Sexo.csv",sep= ",")
```

- Realicemos una breve inspección de los datos

```
str(salud)
```

```
## 'data.frame':    990 obs. of  7 variables:
##  $ TIME              : int  2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...
##  $ GEO               : Factor w/ 33 levels "Austria","Belgium",..: 9 9 9 10 10 10 2 2 2 3 ...
##  $ UNIT              : Factor w/ 1 level "Year": 1 1 1 1 1 1 1 1 1 1 ...
##  $ SEX               : Factor w/ 3 levels "Females","Males",..: 3 2 1 3 2 1 3 2 1 3 ...
##  $ INDIC_HE          : Factor w/ 1 level "Healthy life years in absolute value at birth": 1 1 1 1 1 1
##  $ Value             : Factor w/ 216 levels ":","50.6","51.0",..: 100 95 104 104 99 108 115 122 108 1
##  $ Flag.and.Footnotes: Factor w/ 2 levels "","b": 1 1 1 1 1 1 1 1 1 1 ...
```

```
colnames(salud) #Nombre de las variables
```

```
## [1] "TIME"               "GEO"                "UNIT"
## [4] "SEX"                "INDIC_HE"           "Value"
## [7] "Flag.and.Footnotes"
```

```
nrow(salud) #Número de registros
```

```
## [1] 990
```

```
ncol(salud) #Número de variables
```

```
## [1] 7
```

*Observamos las siguientes variables:

- **TIME**: variable cuantitativa. Indica el año en el que se ha realizado la medida, en este caso el valor de la variable "Value". Se ha cargado bien como número entero.
- **GEO**: variable cualitativa. Indica el país o región en el que se ha realizado la medida.Se ha cargado bien como factor.
- **UNIT**: variable cualitativa. Indica la medida de la variable valor.Se ha cargado bien como factor.Años
- **SEX**: Variable cualitativa. Indica el sexo de la población estudiada, Males, Females o Total.
- **INDIC_HE**: Variable cualitativa. Explica el valor de la variable "Value". Años de vida sana en valores absolutos desde el nacimiento.
- **Value**: Variable cuantitativa. Indica los años de vida sana en valores absolutos desde el nacimiento. Se ha cargado mal como factor.
- **Fal.and.footnotes**. Notas sobre etiquetas. Eliminamos esta columna.

*Años de las mediciones:

```
unique(salud$TIME)
```

```
##  [1] 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
```

*Paises:

```
unique(salud$GEO)
```

```
##  [1] European Union - 27 countries (from 2020)
##  [2] European Union - 28 countries (2013-2020)
##  [3] Belgium
##  [4] Bulgaria
##  [5] Czechia
##  [6] Denmark
##  [7] Germany (until 1990 former territory of the FRG)
##  [8] Estonia
##  [9] Ireland
## [10] Greece
## [11] Spain
## [12] France
## [13] Croatia
## [14] Italy
## [15] Cyprus
## [16] Latvia
## [17] Lithuania
## [18] Luxembourg
## [19] Hungary
## [20] Malta
## [21] Netherlands
## [22] Austria
## [23] Poland
## [24] Portugal
## [25] Romania
## [26] Slovenia
## [27] Slovakia
## [28] Finland
## [29] Sweden
## [30] Iceland
## [31] Norway
## [32] Switzerland
## [33] United Kingdom
## 33 Levels: Austria Belgium Bulgaria Croatia Cyprus Czechia Denmark ... United Kingdom
```

*Unidad de las mediciones:

```
unique(salud$UNIT)
```

```
## [1] Year
## Levels: Year
```

*Años de vida Sana en valores absolutos desde el nacimiento.

```
unique(salud$INDIC_HE)
```

```
## [1] Healthy life years in absolute value at birth
## Levels: Healthy life years in absolute value at birth
```

- Sexo de la población estudiada.

```r
unique(salud$SEX)
```

```
## [1] Total   Males   Females
## Levels: Females Males Total
```

- Eliminamos la columna Fal.and.footnotes y AGE ya que no nos aporta información relevante.

```r
salud<-salud[,-7]
```

- Tendríamos que resolver las posibles inconsistencias en relación al formato del valor numérico de la variable **Value** y convertirla a valor numérico.

```r
salud$Value<-as.character(salud$Value)
salud$Value<-as.numeric (gsub(',','.',salud$Value) )
```

```
## Warning: NAs introducidos por coerción
```

- Comprobamos que valores tenemos en la columna **Value**:

```r
tail(table(salud$Value, useNA = "ifany"))
```

```
##
##   74 74.5 74.6 75.1 76.1 <NA>
##    1    1    1    1    1   18
```

- Observamos que tenemos **18 valores perdidos**.Guardamos en la variable **idx** los índices de los registros con valores **NA** de la variable **Value**.

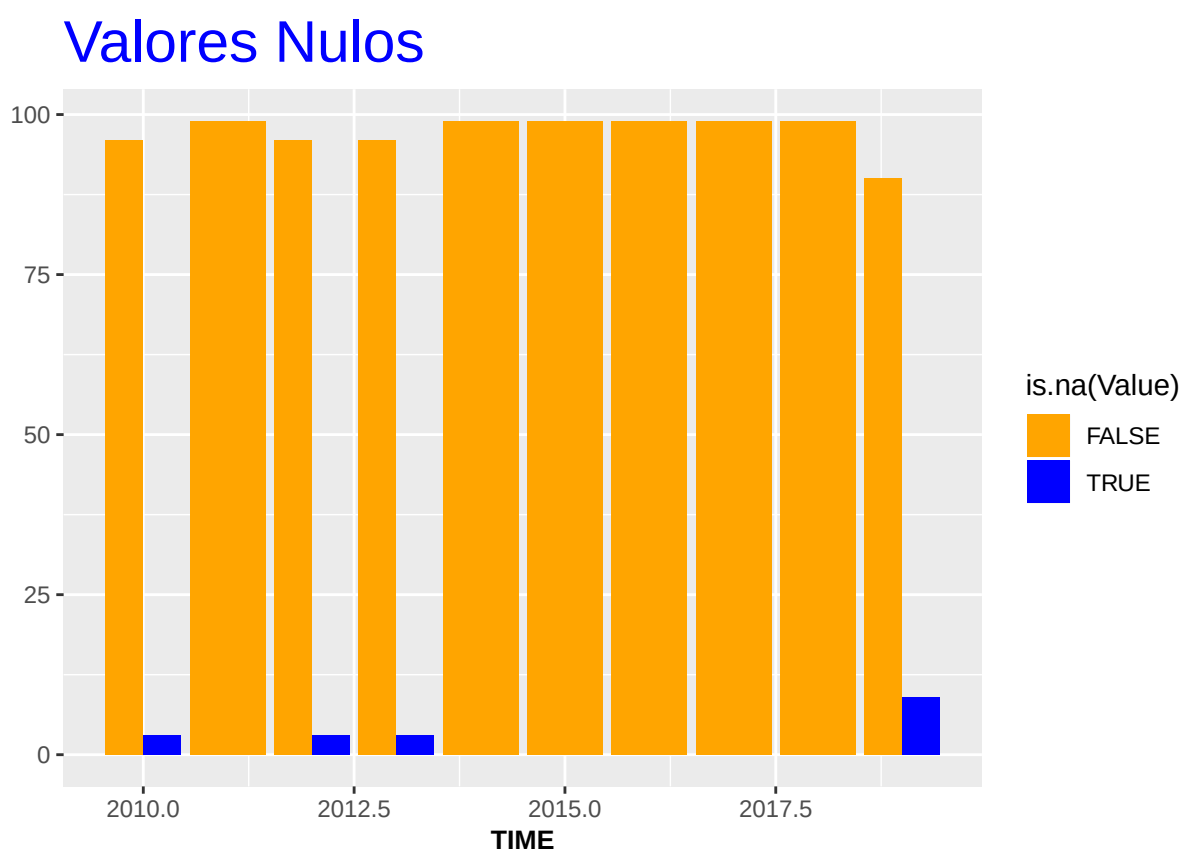
```r
idx<-which(is.na(salud$Value))
length(idx)
```

```
## [1] 18
```

- Grafiquemos la información que contiene la variable **Value**.

```r
library(ggplot2)
library(scales)
g = ggplot(salud, aes(TIME, fill=is.na(Value)) ) +
labs(title = "Valores Nulos")+ylab("") +
theme(plot.title = element_text(size = rel(2), colour = "blue"))

g+geom_bar(position="dodge") + scale_fill_manual(values = alpha(c("orange", "blue"), 1)) +
theme(axis.title.x = element_text(face="bold", size=10))
```



- En caso de detectar algún valor anómalo (en nuestro caso los NAS) en las variables tendríamos que realizar una imputación de esos valores o bien sustituyéndolos por la media o usando el algoritmo KNN (k-Nearest Neighbour) con los 3 vecinos más cercanos usando la distancia que consideremos, en este caso usaremos Gower(Mediana), por ser una medida más robusa frente a extremos.

```
library(VIM)
```

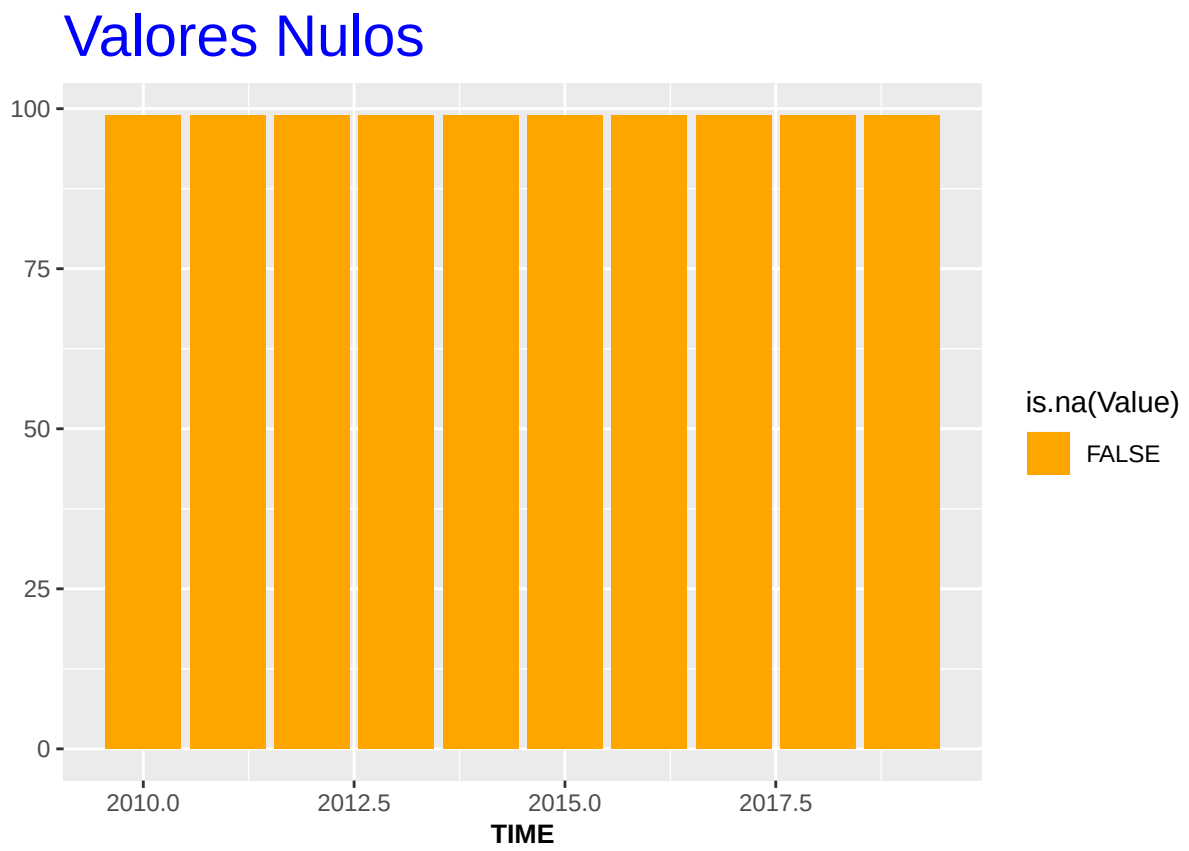
```
## Loading required package: colorspace
## Loading required package: grid
## VIM is ready to use.
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
##
## Attaching package: 'VIM'
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
output<-kNN(salud, variable=c("Value"),k=3)
salud<-output
```

- Comprobamos que no tenemos valores nulos después de la imputación

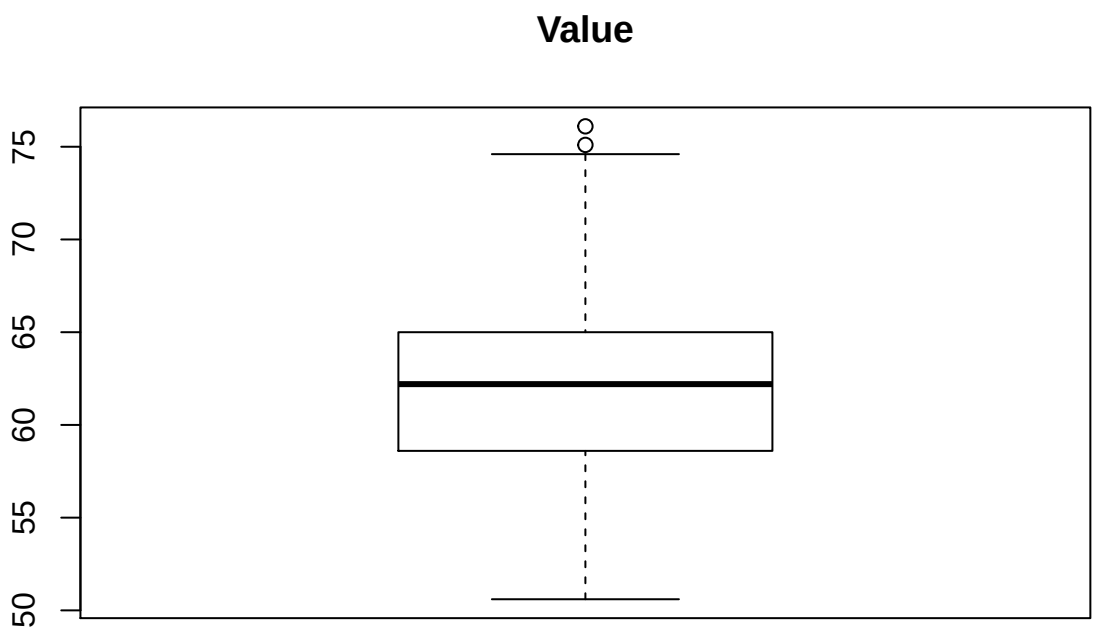
```
g = ggplot(salud, aes(TIME, fill=is.na(Value)) ) +
labs(title = "Valores Nulos")+ylab("") +
theme(plot.title = element_text(size = rel(2), colour = "blue"))
```

```
g+geom_bar(position="dodge") + scale_fill_manual(values = alpha(c("orange", "blue"), 1)) +
theme(axis.title.x = element_text(face="bold", size=10))
```



- Con el siguiente gráfico, observaremos que la variable **Value** tiene outliers o valores extremos

```
boxplot(salud$Value, main="Value")
```



- Por otro lado, revisamos para el resto de columnas si tenemos valores NA.(desconocidos o perdidos)

```
table(salud$TIME, useNA = "ifany")
```

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
##   99   99   99   99   99   99   99   99   99   99
```

```
table(salud$GEO, useNA = "ifany")
```

```
##
##                                        Austria
##                                             30
##                                        Belgium
##                                             30
##                                       Bulgaria
##                                             30
##                                        Croatia
##                                             30
##                                         Cyprus
##                                             30
##                                        Czechia
##                                             30
##                                        Denmark
##                                             30
##                                        Estonia
##                                             30
##        European Union - 27 countries (from 2020)
```

```
##                                                30
##        European Union - 28 countries (2013-2020)
##                                                30
##                                           Finland
##                                                30
##                                            France
##                                                30
## Germany (until 1990 former territory of the FRG)
##                                                30
##                                            Greece
##                                                30
##                                           Hungary
##                                                30
##                                           Iceland
##                                                30
##                                           Ireland
##                                                30
##                                             Italy
##                                                30
##                                            Latvia
##                                                30
##                                         Lithuania
##                                                30
##                                        Luxembourg
##                                                30
##                                             Malta
##                                                30
##                                       Netherlands
##                                                30
##                                            Norway
##                                                30
##                                            Poland
##                                                30
##                                          Portugal
##                                                30
##                                           Romania
##                                                30
##                                          Slovakia
##                                                30
##                                          Slovenia
##                                                30
##                                             Spain
##                                                30
##                                            Sweden
##                                                30
##                                       Switzerland
##                                                30
##                                    United Kingdom
##                                                30
```

```
table(salud$UNIT, useNA = "ifany")
```

```
##
## Year
##  990
```

```
table(salud$SEX, useNA = "ifany")
```

```
##
## Females   Males   Total
##     330     330     330
```

```
table(salud$INDIC_HE, useNA = "ifany")
```

```
##
## Healthy life years in absolute value at birth
##                                           990
```

Observamos que no existen ahora valores perdidos después de la imputación.La suma de las cantidades de cada variable, suman el total.

La estructura de los datos quedaría:

```
str(salud)
```

```
## 'data.frame':    990 obs. of  7 variables:
##  $ TIME     : int  2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...
##  $ GEO      : Factor w/ 33 levels "Austria","Belgium",..: 9 9 9 10 10 10 2 2 2 3 ...
##  $ UNIT     : Factor w/ 1 level "Year": 1 1 1 1 1 1 1 1 1 1 ...
##  $ SEX      : Factor w/ 3 levels "Females","Males",..: 3 2 1 3 2 1 3 2 1 3 ...
##  $ INDIC_HE : Factor w/ 1 level "Healthy life years in absolute value at birth": 1 1 1 1 1 1 1 1 1 1
##  $ Value    : num  61.8 61.3 62.2 62.2 61.7 62.6 63.3 64 62.6 65 ...
##  $ Value_imp: logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

- Finalmente, creamos un fichero con toda la información corregida.

```
write.csv(salud, file="EstadoDeSalud_Sexo_clean.csv", row.names = FALSE)
```