Mortalidad Tratable y Prevenible por Paises y Sexos

Alicia Perdices Guerra

3 de mayo, 2021

Contents

1.PROCESAMIENTO DE LOS DATOS.

- En primer lugar leemos el fichero:

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

- Realicemos una breve inspección de los datos

```
str(mortalidad)
```

```
## 'data.frame':    4536 obs. of  8 variables:
##  $ TIME              : int  2011 2011 2011 2011 2011 2011 2011 2011 2011 2011 ...
##  $ GEO               : Factor w/ 36 levels "Austria","Belgium",..: 9 9 9 9 9 9 9 9 9 9 ...
##  $ MORTALIT          : Factor w/ 3 levels "Preventable mortality",..: 2 2 2 2 2 2 1 1 1 1 ...
##  $ SEX               : Factor w/ 3 levels "Females","Males",..: 3 3 2 2 1 1 3 3 2 2 ...
##  $ ICD10             : Factor w/ 1 level "Total": 1 1 1 1 1 1 1 1 1 1 ...
##  $ UNIT              : Factor w/ 2 levels "Number","Rate": 1 2 1 2 1 2 1 2 1 2 ...
##  $ Value             : Factor w/ 4271 levels ":","1 009.5",..: 11 2089 3658 2667 2586 1215 3525 1148
##  $ Flag.and.Footnotes: Factor w/ 2 levels "","p": 1 1 1 1 1 1 1 1 1 1 ...
```

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

```
## [1] "TIME"               "GEO"                "MORTALIT"
## [4] "SEX"                "ICD10"              "UNIT"
## [7] "Value"              "Flag.and.Footnotes"
```

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

```
## [1] 4536
```

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

```
## [1] 8
```

*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.Número y Ratio
- **MORTALIT**: variable cualitativa. Indica el tipo de mortalidad, Tratable , Prevenible o Total.
- **SEX**: Variable cualitativa. Indica el sexo de la población estudiada.
- **ICD10**: Variable cualitativa. En la clasificación de enfermedades, en este apartado indica el Total de ellas.
- **Value**: Variable cuantitativa. Indica número o ratio de causas de muerte tratable o prevenible.

- **Fal.and.footnotes**. Notas sobre etiquetas. Eliminamos esta columna.
- Años de las mediciones:

```
unique(mortalidad$TIME)
```

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

- Paises:

```
unique(mortalidad$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] Liechtenstein
## [32] Norway
## [33] Switzerland
## [34] United Kingdom
## [35] Serbia
## [36] Turkey
## 36 Levels: Austria Belgium Bulgaria Croatia Cyprus Czechia Denmark ... United Kingdom
```

- Unidad de las mediciones:

```
unique(mortalidad$UNIT)
```

```
## [1] Number Rate
## Levels: Number Rate
```

- Tipo de mortalidad:

```
unique(mortalidad$MORTALIT)
```

```
## [1] Total                 Preventable mortality Treatable mortality
## Levels: Preventable mortality Total Treatable mortality
```

- Sexo de la población estudiada.

```
unique(mortalidad$SEX)
```

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

- En la clasificación de enfermedades, en este apartado indica el Total de ellas

```
unique(mortalidad$ICD10)
```

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

- Eliminamos la columna Fal.and.footnotes.

```
mortalidad<-mortalidad[,-8]
```

- Tendríamos que convertir la columna Value a numérico porque se ha cargado como factor y es erróneo. El resto de variables tienen el tipo correcto.

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

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

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

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

```
##
## 1163083 1175433 1183590 1200902 1210890    <NA>
##       1       1       1       1       1      54
```

- Observamos que tenemos **54 valores perdidos**.Guardamos en la variable **idx** los índices de los registros con valores **NA** de la variable **Value**.

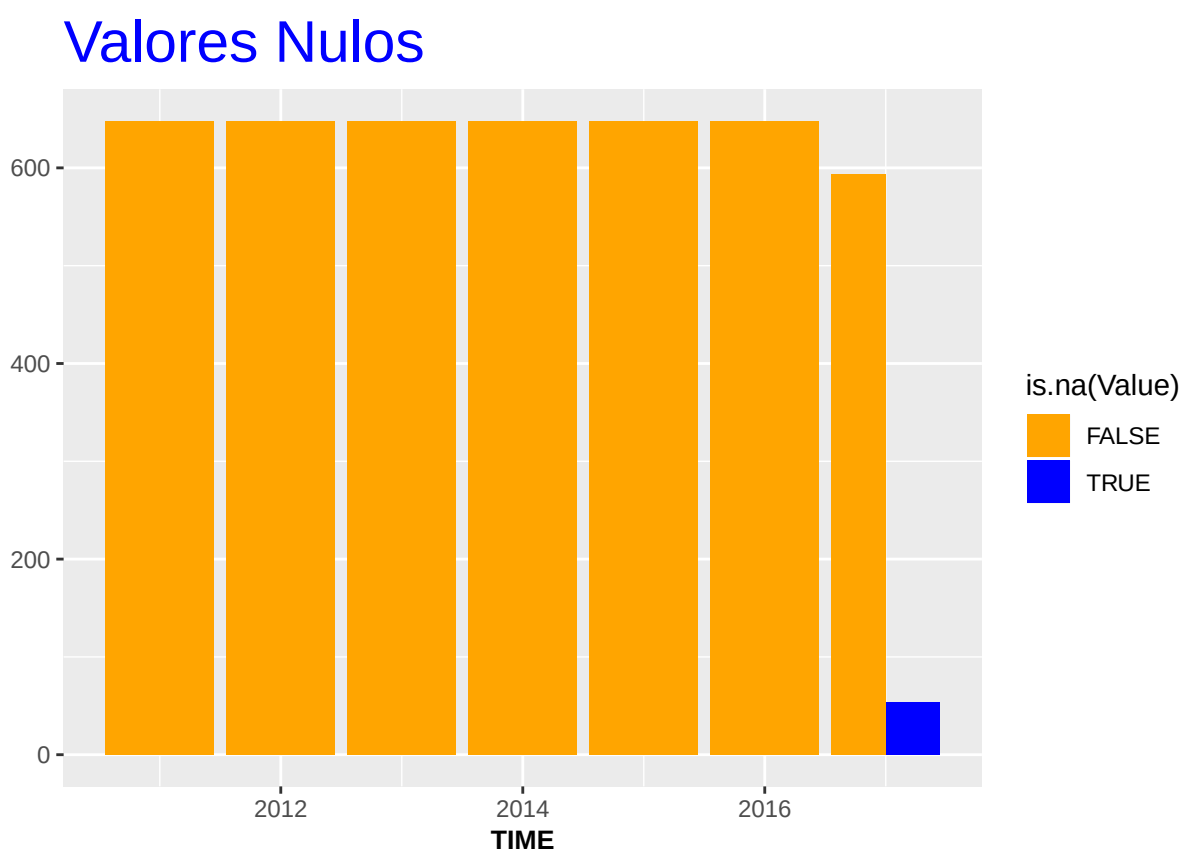
```
idx<-which(is.na(mortalidad$Value))
length(idx)
```

```
## [1] 54
```

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

```
library(ggplot2)
library(scales)
g = ggplot(mortalidad, 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)
```

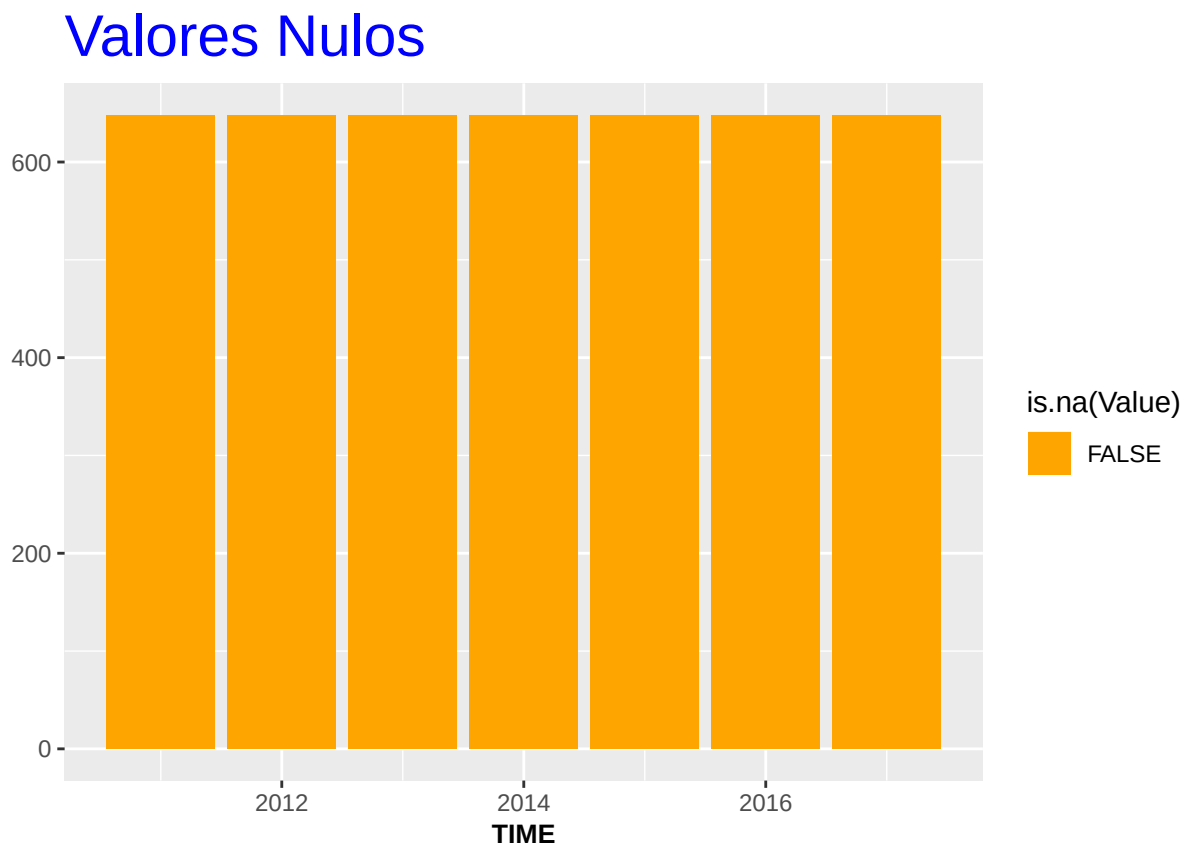
```
## Loading required package: colorspace
## Loading required package: grid
## VIM is ready to use.
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
##
## Attaching package: 'VIM'
## The following object is masked from 'package:datasets':
##
##     sleep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

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

- Comprobamos que no tenemos valores nulos después de la imputación

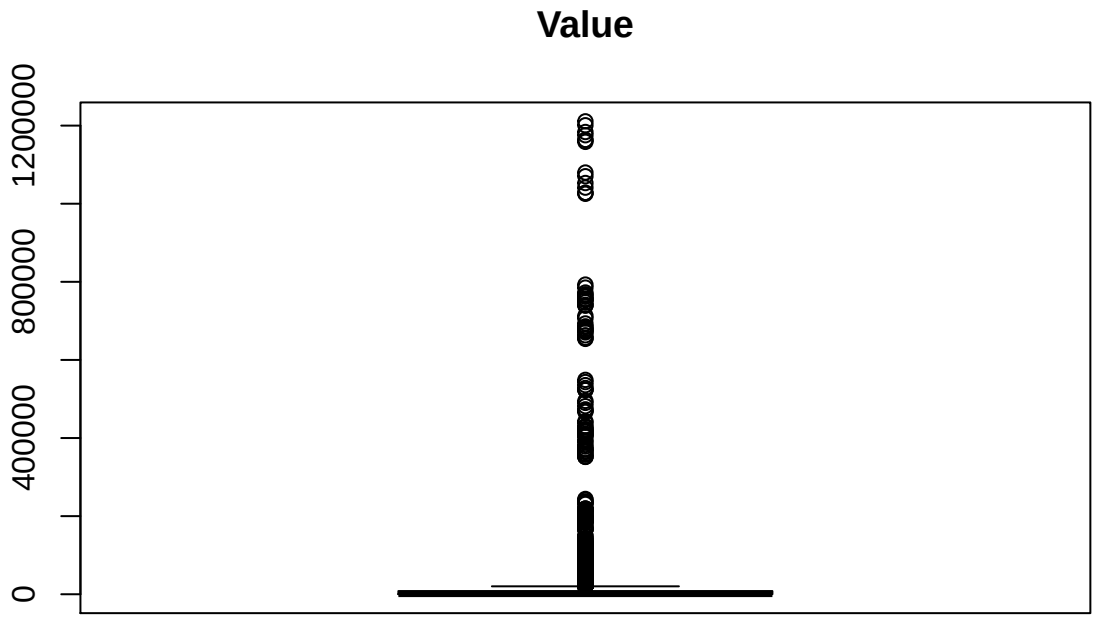
```
g = ggplot(mortalidad, aes(TIME, fill=is.na(Value)) ) +
labs(title = "Valores Nulos")+ylab("") +
theme(plot.title = element_text(size = rel(2), colour = "blue"))
```

```
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(mortalidad$Value, main="Value")
```



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

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

```
##
## 2011 2012 2013 2014 2015 2016 2017
##  648  648  648  648  648  648  648
```

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

```
##
##                                        Austria
##                                            126
##                                        Belgium
##                                            126
##                                       Bulgaria
##                                            126
##                                        Croatia
##                                            126
##                                         Cyprus
##                                            126
##                                        Czechia
##                                            126
##                                        Denmark
##                                            126
##                                        Estonia
##                                            126
##        European Union - 27 countries (from 2020)
```

```
##                                             126
##        European Union - 28 countries (2013-2020)
##                                             126
##                                         Finland
##                                             126
##                                          France
##                                             126
## Germany (until 1990 former territory of the FRG)
##                                             126
##                                          Greece
##                                             126
##                                         Hungary
##                                             126
##                                         Iceland
##                                             126
##                                         Ireland
##                                             126
##                                           Italy
##                                             126
##                                          Latvia
##                                             126
##                                   Liechtenstein
##                                             126
##                                       Lithuania
##                                             126
##                                      Luxembourg
##                                             126
##                                           Malta
##                                             126
##                                     Netherlands
##                                             126
##                                          Norway
##                                             126
##                                          Poland
##                                             126
##                                        Portugal
##                                             126
##                                         Romania
##                                             126
##                                          Serbia
##                                             126
##                                        Slovakia
##                                             126
##                                        Slovenia
##                                             126
##                                           Spain
##                                             126
##                                          Sweden
##                                             126
##                                     Switzerland
##                                             126
##                                          Turkey
##                                             126
##                                  United Kingdom
```

```
##                                          126
```

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

```
##
## Number   Rate
##   2268   2268
```

```
table(mortalidad$MORTALIT, useNA = "ifany")
```

```
##
## Preventable mortality                 Total   Treatable mortality
##                  1512                  1512                  1512
```

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

```
##
## Females   Males   Total
##    1512    1512    1512
```

```
table(mortalidad$ICD10, useNA = "ifany")
```

```
##
## Total
##  4536
```

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 del siguiente modo:

```
str(mortalidad)
```

```
## 'data.frame':    4536 obs. of  8 variables:
##  $ TIME     : int  2011 2011 2011 2011 2011 2011 2011 2011 2011 2011 ...
##  $ GEO      : Factor w/ 36 levels "Austria","Belgium",..: 9 9 9 9 9 9 9 9 9 9 ...
##  $ MORTALIT : Factor w/ 3 levels "Preventable mortality",..: 2 2 2 2 2 2 1 1 1 1 ...
##  $ SEX      : Factor w/ 3 levels "Females","Males",..: 3 3 2 2 1 1 3 3 2 2 ...
##  $ ICD10    : Factor w/ 1 level "Total": 1 1 1 1 1 1 1 1 1 1 ...
##  $ UNIT     : Factor w/ 2 levels "Number","Rate": 1 2 1 2 1 2 1 2 1 2 ...
##  $ Value    : num  1079803 281 713132 391 366671 ...
##  $ Value_imp: logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

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

```
write.csv(mortalidad, file="Mortalidad_Tratable_Prevenible_clean.csv", row.names = FALSE)
```