Mortalidad Por Enfermedades Infecciosas

Alicia Perdices Guerra

3 de mayo, 2021

Contents

1.PROCESAMIENTO DE LOS DATOS.

- En primer lugar leemos el fichero:

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

- Realicemos una breve inspección de los datos

```
str(infeccion)
```

```
## 'data.frame':    1332 obs. of  7 variables:
##  $ TIME              : int  2011 2011 2011 2011 2011 2011 2011 2011 2011 2011 ...
##  $ GEO               : Factor w/ 37 levels "Austria","Belgium",..: 9 9 9 9 10 10 10 10 2 2 ...
##  $ UNIT              : Factor w/ 1 level "Number": 1 1 1 1 1 1 1 1 1 1 ...
##  $ SEX               : Factor w/ 1 level "Total": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ICD10             : Factor w/ 4 levels "Certain infectious diseases (A00-A40, A42-B99)",..: 1 3 2
##  $ Value             : Factor w/ 949 levels ":","0","1","1 000",..: 573 545 927 884 616 558 208 230 
##  $ Flag.and.Footnotes: logi  NA NA NA NA NA NA ...
```

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

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

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

```
## [1] 1332
```

```
ncol(infeccion) #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.Número
- **SEX**: Variable cualitativa. Indica el sexo de la población estudiada.Total
- **ICD10**: Variable cualitativa.Indica la clasificación de enfermedades infecciosas que se han estudiado en la población.
- **Value**: Variable cuantitativa. Indica número de muertes por cada tipo de enfermedad infecciosa.
- **Fal.and.footnotes**. Notas sobre etiquetas. Eliminamos esta columna.
- Años de las mediciones:

```r
unique(infeccion$TIME)
```

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

- Paises:

```r
unique(infeccion$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] France (metropolitan)
## [14] Croatia
## [15] Italy
## [16] Cyprus
## [17] Latvia
## [18] Lithuania
## [19] Luxembourg
## [20] Hungary
## [21] Malta
## [22] Netherlands
## [23] Austria
## [24] Poland
## [25] Portugal
## [26] Romania
## [27] Slovenia
## [28] Slovakia
## [29] Finland
## [30] Sweden
## [31] Iceland
## [32] Liechtenstein
## [33] Norway
## [34] Switzerland
## [35] United Kingdom
## [36] Serbia
## [37] Turkey
## 37 Levels: Austria Belgium Bulgaria Croatia Cyprus Czechia Denmark ... United Kingdom
```

- Unidad de las mediciones:

```r
unique(infeccion$UNIT)
```

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

- Sexo de la población estudiada.

```r
unique(infeccion$SEX)
```

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

- En la clasificación de enfermedades tenemos:

```r
unique(infeccion$ICD10)
```

```
## [1] Certain infectious diseases (A00-A40, A42-B99)
## [2] Other sepsis
## [3] Other infectious diseases (G00, G03-G04, G06, G08-G09, H00-H01, H10, H16, H20, H30, H46, H60, H6
## [4] Pneumonia, organism unspecified
## 4 Levels: Certain infectious diseases (A00-A40, A42-B99) ...
```

- Eliminamos la columna Fal.and.footnotes.

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

- 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.

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

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

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

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

```
##
## 120389 121701 129257 130816 133225   <NA>
##      1      1      1      1      1    156
```

- Observamos que tenemos **156 valores perdidos**.Guardamos en la variable **idx** los índices de los registros con valores **NA** de la variable **Value**.

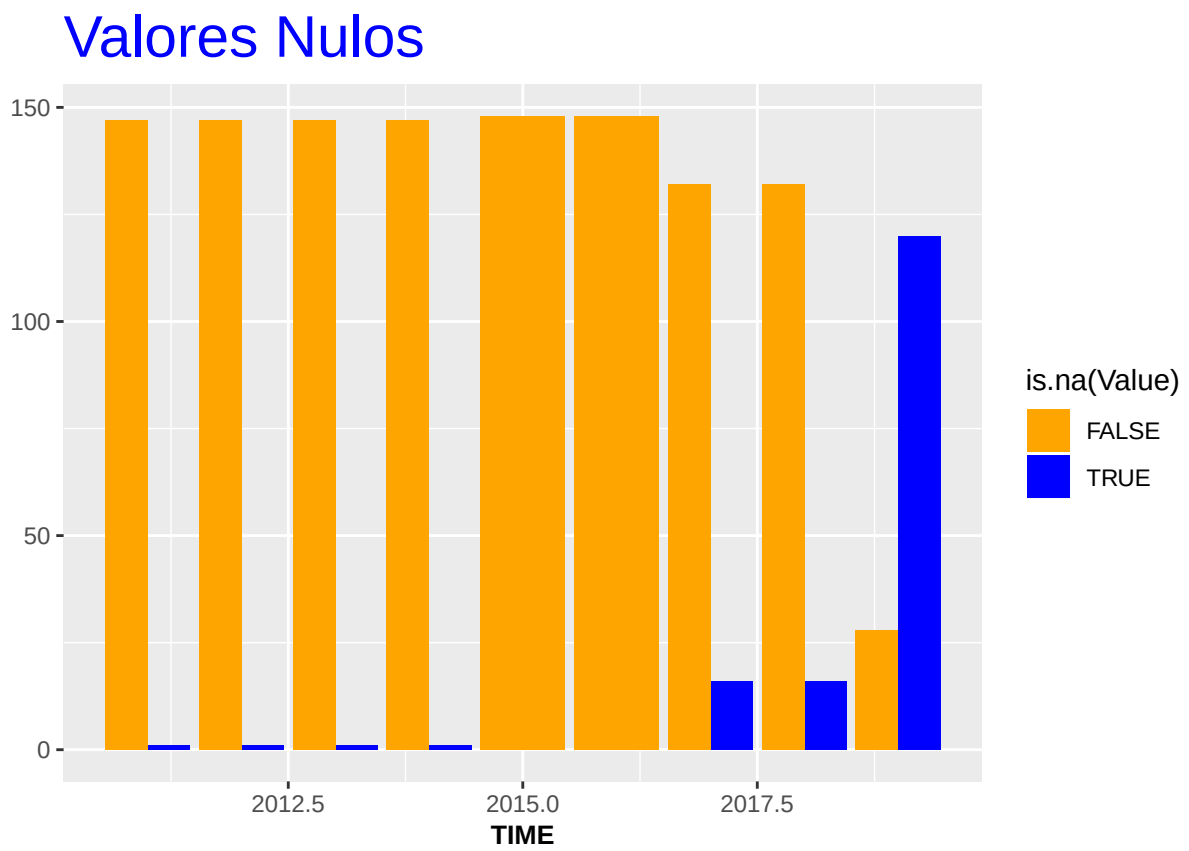
```r
idx<-which(is.na(infeccion$Value))
length(idx)
```

```
## [1] 156
```

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

```r
library(ggplot2)
library(scales)
g = ggplot(infeccion, 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)
```

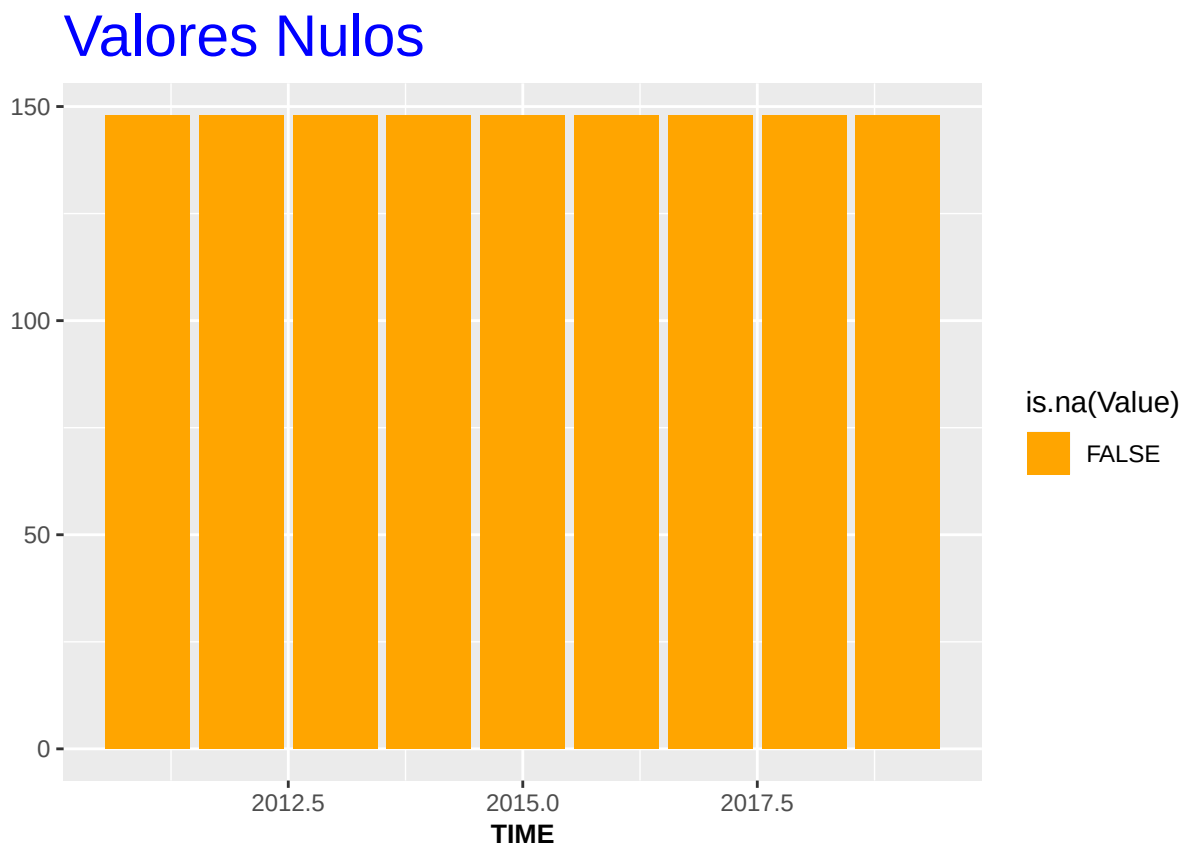
```
## Loading required package: colorspace
## Loading required package: grid
## VIM is ready to use.
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
##
## Attaching package: 'VIM'
## The following object is masked from 'package:datasets':
##
##     sleep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

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

- Comprobamos que no tenemos valores nulos después de la imputación

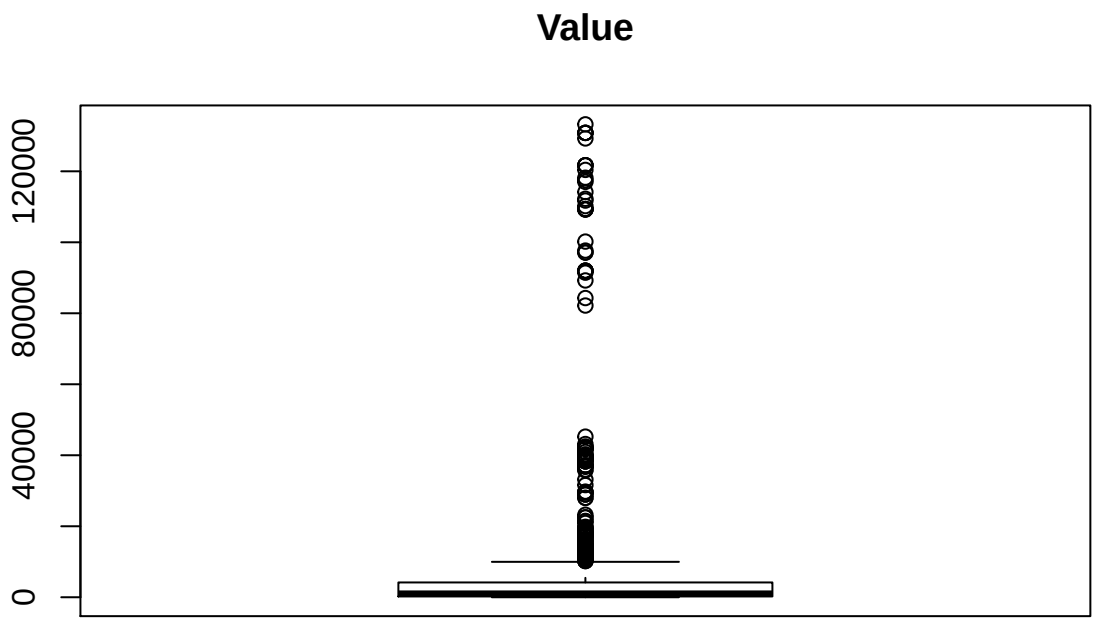
```
g = ggplot(infeccion, aes(TIME, fill=is.na(Value)) ) +
labs(title = "Valores Nulos")+ylab("") +
theme(plot.title = element_text(size = rel(2), colour = "blue"))
```

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



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

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

```
##
## 2011 2012 2013 2014 2015 2016 2017 2018 2019
##  148  148  148  148  148  148  148  148  148
```

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

```
##
##                                        Austria
##                                             36
##                                        Belgium
##                                             36
##                                       Bulgaria
##                                             36
##                                        Croatia
##                                             36
##                                         Cyprus
##                                             36
##                                        Czechia
##                                             36
##                                        Denmark
##                                             36
##                                        Estonia
##                                             36
##        European Union - 27 countries (from 2020)
```

```
##                                                36
##        European Union - 28 countries (2013-2020)
##                                                36
##                                           Finland
##                                                36
##                                            France
##                                                36
##                               France (metropolitan)
##                                                36
## Germany (until 1990 former territory of the FRG)
##                                                36
##                                            Greece
##                                                36
##                                           Hungary
##                                                36
##                                           Iceland
##                                                36
##                                           Ireland
##                                                36
##                                             Italy
##                                                36
##                                            Latvia
##                                                36
##                                     Liechtenstein
##                                                36
##                                         Lithuania
##                                                36
##                                        Luxembourg
##                                                36
##                                             Malta
##                                                36
##                                       Netherlands
##                                                36
##                                            Norway
##                                                36
##                                            Poland
##                                                36
##                                          Portugal
##                                                36
##                                           Romania
##                                                36
##                                            Serbia
##                                                36
##                                          Slovakia
##                                                36
##                                          Slovenia
##                                                36
##                                             Spain
##                                                36
##                                            Sweden
##                                                36
##                                       Switzerland
##                                                36
##                                            Turkey
```

```
##                                          36
##                                United Kingdom
##                                          36
```

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

```
##
## Number
##   1332
```

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

```
##
## Total
##  1332
```

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

```
##
##
##
## Other infectious diseases (G00, G03-G04, G06, G08-G09, H00-H01, H10, H16, H20, H30, H46, H60, H65-H6
##
##
##
##
##
```

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(infeccion)
```

```
## 'data.frame':    1332 obs. of  7 variables:
##  $ TIME     : int  2011 2011 2011 2011 2011 2011 2011 2011 2011 2011 ...
##  $ GEO      : Factor w/ 37 levels "Austria","Belgium",..: 9 9 9 9 10 10 10 10 2 2 ...
##  $ UNIT     : Factor w/ 1 level "Number": 1 1 1 1 1 1 1 1 1 1 ...
##  $ SEX      : Factor w/ 1 level "Total": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ICD10    : Factor w/ 4 levels "Certain infectious diseases (A00-A40, A42-B99)",..: 1 3 2 4 1 3 2 4
##  $ Value    : num  37138 33171 92037 84289 40939 ...
##  $ Value_imp: logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

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

```
write.csv(infeccion, file="Muertes_Enf_Infecciosas_clean.csv", row.names = FALSE)
```