Detección Cáncer de Mama y Cérvix uterino

Alicia Perdices Guerra

3 de mayo, 2021

Contents

1.PROCESAMIENTO DE LOS DATOS.

- En primer lugar leemos el fichero:

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

- Realicemos una breve inspección de los datos

```
str(cancer)
```

```
## 'data.frame':    1440 obs. of  7 variables:
##  $ TIME              : int  2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...
##  $ GEO               : Factor w/ 36 levels "Austria","Belgium",..: 2 2 2 2 3 3 3 3 6 6 ...
##  $ UNIT              : Factor w/ 1 level "Percentage": 1 1 1 1 1 1 1 1 1 1 ...
##  $ SOURCE            : Factor w/ 2 levels "Programme data",..: 2 2 1 1 2 2 1 1 2 2 ...
##  $ ICD10             : Factor w/ 2 levels "Malignant neoplasm of breast",..: 1 2 1 2 1 2 1 2 1 2 ...
##  $ Value             : Factor w/ 391 levels ":","0.04","0.20",..: 1 1 218 226 1 1 1 1 1 1 ...
##  $ Flag.and.Footnotes: Factor w/ 8 levels "","b","bd","d",..: 1 1 1 1 1 1 1 1 1 1 ...
```

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

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

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

```
## [1] 1440
```

```
ncol(cancer) #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.Porcentaje.
- **SOURCE**: variable cualitativa. Indica la fuente del estudio.
- **ICD10**: variable cualitativa. Indica el tipo de cancer foco de estudio: cáncer de mama o cérvix uterino.
- **Value**: Variable cuantitativa. Indica el porcentaje de casos de cáncer (de mama o cérvix uterino) por países.Se ha cargado mal como factor.
- **Fal.and.footnotes**. Notas sobre etiquetas. Eliminamos esta columna.

*Años de las mediciones:

```
unique(cancer$TIME)
```

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

*Paises:

```
unique(cancer$GEO)
```

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

*Unidad de las mediciones:

```
unique(cancer$UNIT)
```

```
## [1] Percentage
## Levels: Percentage
```

- Tipo de fuente:

```
unique(cancer$SOURCE)
```

```
## [1] Survey data    Programme data
## Levels: Programme data Survey data
```

*Tipo de cáncer foco de estudio.

```
unique(cancer$ICD10)
```

```
## [1] Malignant neoplasm of breast       Malignant neoplasm of cervix uteri
## Levels: Malignant neoplasm of breast Malignant neoplasm of cervix uteri
```

- Eliminamos la columna Fal.and.footnotes.

```
cancer<-cancer[,-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.

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

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

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

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

```
##
## 85.6 87.2 87.3 87.6 90.4 <NA>
##    1    1    1    1    1  915
```

- Observamos que tenemos **915 valores perdidos**.Guardamos en la variable **idx** los índices de los registros con valores **NA** de la variable **Value**.

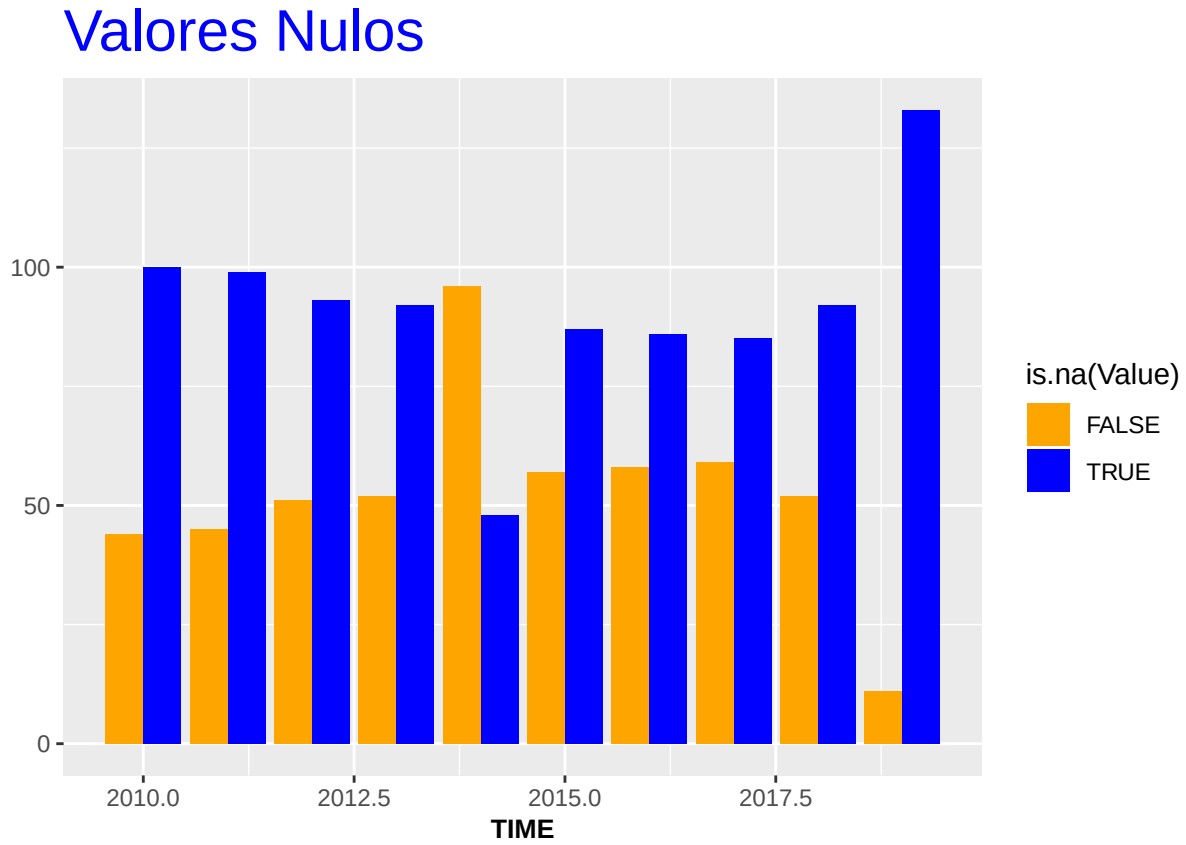
```
idx<-which(is.na(cancer$Value))
length(idx)
```

```
## [1] 915
```

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

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

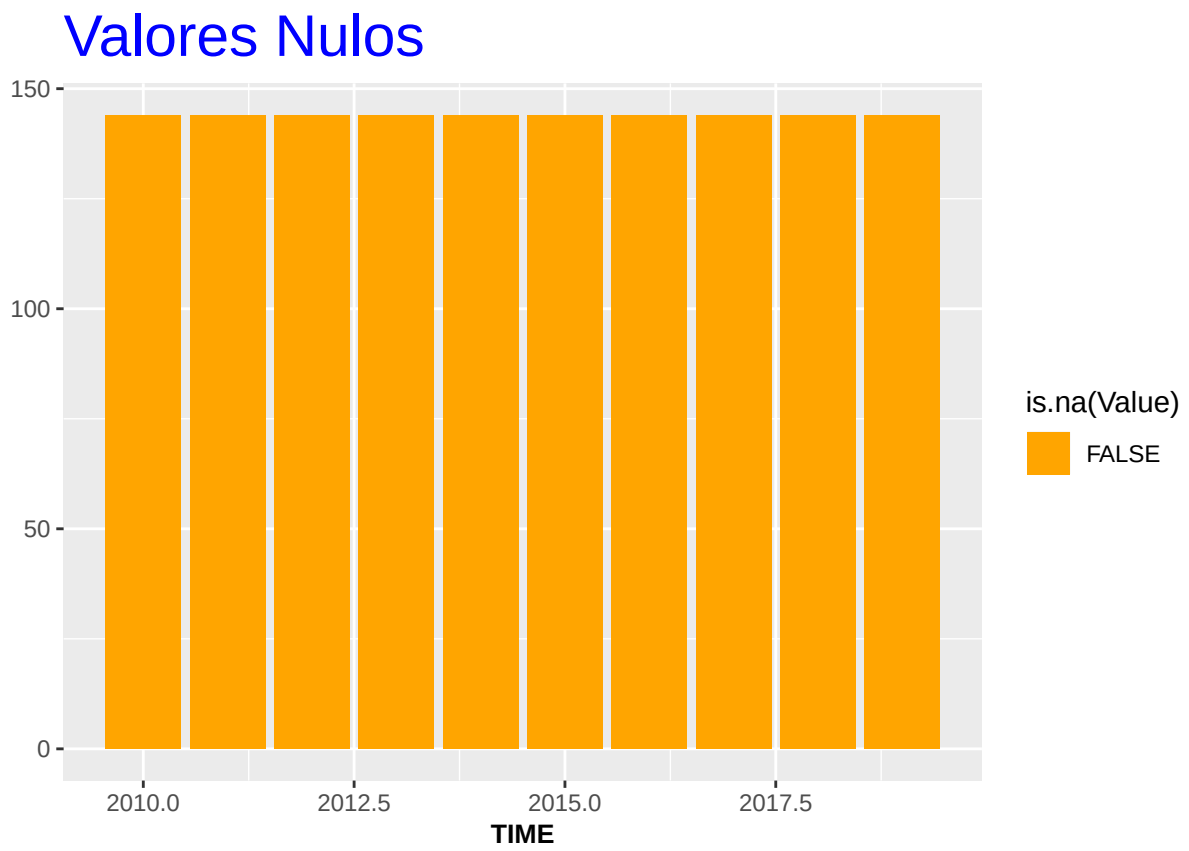
```
## Loading required package: colorspace
## Loading required package: grid
## VIM is ready to use.
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
##
## Attaching package: 'VIM'
## The following object is masked from 'package:datasets':
##
##     sleep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

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

- Comprobamos que no tenemos valores nulos después de la imputación

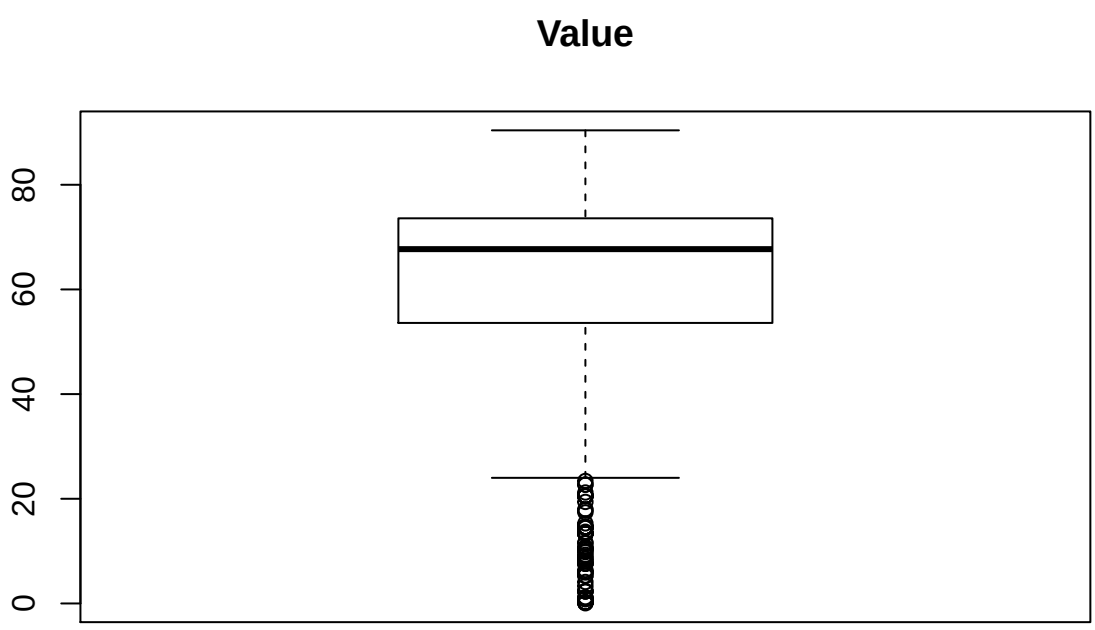
```
g = ggplot(cancer, aes(TIME, fill=is.na(Value)) ) +
labs(title = "Valores Nulos")+ylab("") +
theme(plot.title = element_text(size = rel(2), colour = "blue"))
```

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



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

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

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
##  144  144  144  144  144  144  144  144  144  144
```

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

```
##
##                                          Austria
##                                               40
##                                          Belgium
##                                               40
##                                         Bulgaria
##                                               40
##                                          Croatia
##                                               40
##                                           Cyprus
##                                               40
##                                          Czechia
##                                               40
##                                          Denmark
##                                               40
##                                          Estonia
##                                               40
##                                          Finland
```

```
##                                             40
##                                         France
##                                             40
## Germany (until 1990 former territory of the FRG)
##                                             40
##                                         Greece
##                                             40
##                                        Hungary
##                                             40
##                                        Iceland
##                                             40
##                                        Ireland
##                                             40
##                                          Italy
##                                             40
##                                         Latvia
##                                             40
##                                  Liechtenstein
##                                             40
##                                      Lithuania
##                                             40
##                                     Luxembourg
##                                             40
##                                          Malta
##                                             40
##                                     Montenegro
##                                             40
##                                    Netherlands
##                                             40
##                                  North Macedonia
##                                             40
##                                         Norway
##                                             40
##                                         Poland
##                                             40
##                                       Portugal
##                                             40
##                                        Romania
##                                             40
##                                         Serbia
##                                             40
##                                       Slovakia
##                                             40
##                                       Slovenia
##                                             40
##                                          Spain
##                                             40
##                                         Sweden
##                                             40
##                                    Switzerland
##                                             40
##                                         Turkey
##                                             40
##                                 United Kingdom
```

```
##                                                 40
```

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

```
##
## Percentage
##       1440
```

```
table(cancer$SOURCE, useNA = "ifany")
```

```
##
## Programme data    Survey data
##            720            720
```

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

```
##
##       Malignant neoplasm of breast Malignant neoplasm of cervix uteri
##                                720                               720
```

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

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

```
write.csv(cancer, file="Deteccion_Cancer_Mama_Cervix_clean.csv", row.names = FALSE)
```