Detección Enfermedades Cardiovasculares y Diabetes

Alicia Perdices Guerra

3 de mayo, 2021

Contents

1.PROCESAMIENTO DE LOS DATOS.

- En primer lugar leemos el fichero:

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

- Realicemos una breve inspección de los datos

```
str(cardio_diabetes)
```

```
## 'data.frame':    576 obs. of  10 variables:
##  $ HLTHCARE          : Factor w/ 3 levels "Blood cholesterol measurement",..: 2 2 2 2 2 2 2 2 2 2 ..
##  $ GEO               : Factor w/ 32 levels "Austria","Bulgaria",..: 8 8 8 8 8 8 9 9 9 9 ...
##  $ UNIT              : Factor w/ 1 level "Percentage": 1 1 1 1 1 1 1 1 1 1 ...
##  $ TIME              : int  2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 ...
##  $ DURATION          : Factor w/ 6 levels "5 years or over",..: 4 2 3 5 1 6 4 2 3 5 ...
##  $ ISCED11           : Factor w/ 1 level "All ISCED 2011 levels ": 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 ...
##  $ AGE               : Factor w/ 1 level "Total": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Value             : num  67.1 19.2 4.1 90.4 3.8 5.8 66.9 19.1 4.3 90.3 ...
##  $ Flag.and.Footnotes: Factor w/ 2 levels "","e": 1 1 1 1 1 1 2 2 2 2 ...
```

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

```
##  [1] "HLTHCARE"           "GEO"                "UNIT"
##  [4] "TIME"               "DURATION"           "ISCED11"
##  [7] "SEX"                "AGE"                "Value"
## [10] "Flag.and.Footnotes"
```

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

```
## [1] 576
```

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

```
## [1] 10
```

*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.
- **HLTHCARE**: variable cualitativa. Indica el tipo de medida.
- **ISCED11**:Variable cualitativa. Estándar en estadísticas de educación en el que se hacen las mediciones.
- **DURATION**: variable cualitativa. Indica durante cuánto tiempo se hace la medición.
- **SEX**: Variable cualitativa. En cómputo total.No hay distinción entre sexos.
- **AGE**: Variable cualitativa. En Cómputo total. No hay distinción en edades.

- **Value**: Variable cuantitativa. Indica el porcentaje de población a la que se le ha hecho cada medida en sangre por países.Se ha cargado bien como valor numérico.
- **Fal.and.footnotes**. Notas sobre etiquetas. Eliminamos esta columna.

*Años de las mediciones:

```
unique(cardio_diabetes$TIME)
```

```
## [1] 2014
```

*Paises:

```
unique(cardio_diabetes$GEO)
```

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

*Unidad de las mediciones:

```
unique(cardio_diabetes$UNIT)
```

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

- Tipo de medida en sangre:

```r
unique(cardio_diabetes$HLTHCARE)
```

```
## [1] Blood pressure measurement    Blood cholesterol measurement
## [3] Blood sugar measurement
## 3 Levels: Blood cholesterol measurement ... Blood sugar measurement
```

*Estándar de las mediciones.

```r
unique(cardio_diabetes$ISCED11)
```

```
## [1] All ISCED 2011 levels
## Levels: All ISCED 2011 levels
```

- Duración de las mediciones

```r
unique(cardio_diabetes$DURATION)
```

```
## [1] Less than 1 year  From 1 to 3 years From 3 to 5 years Less than 5 years
## [5] 5 years or over   Never
## 6 Levels: 5 years or over From 1 to 3 years ... Never
```

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

```r
cardio_diabetes<-cardio_diabetes[,-10]
cardio_diabetes<-cardio_diabetes[,-8]
cardio_diabetes<-cardio_diabetes[,-7]
```

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

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

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

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

```
##
## 95.1 95.5   96 96.4 96.5 97.3
##    2    1    2    1    1    1
```

- Observamos que no tenemos valores perdidos.Guardamos en la variable **idx** los índices de los registros con valores **NA** de la variable **Value**.Observamos que es cero.

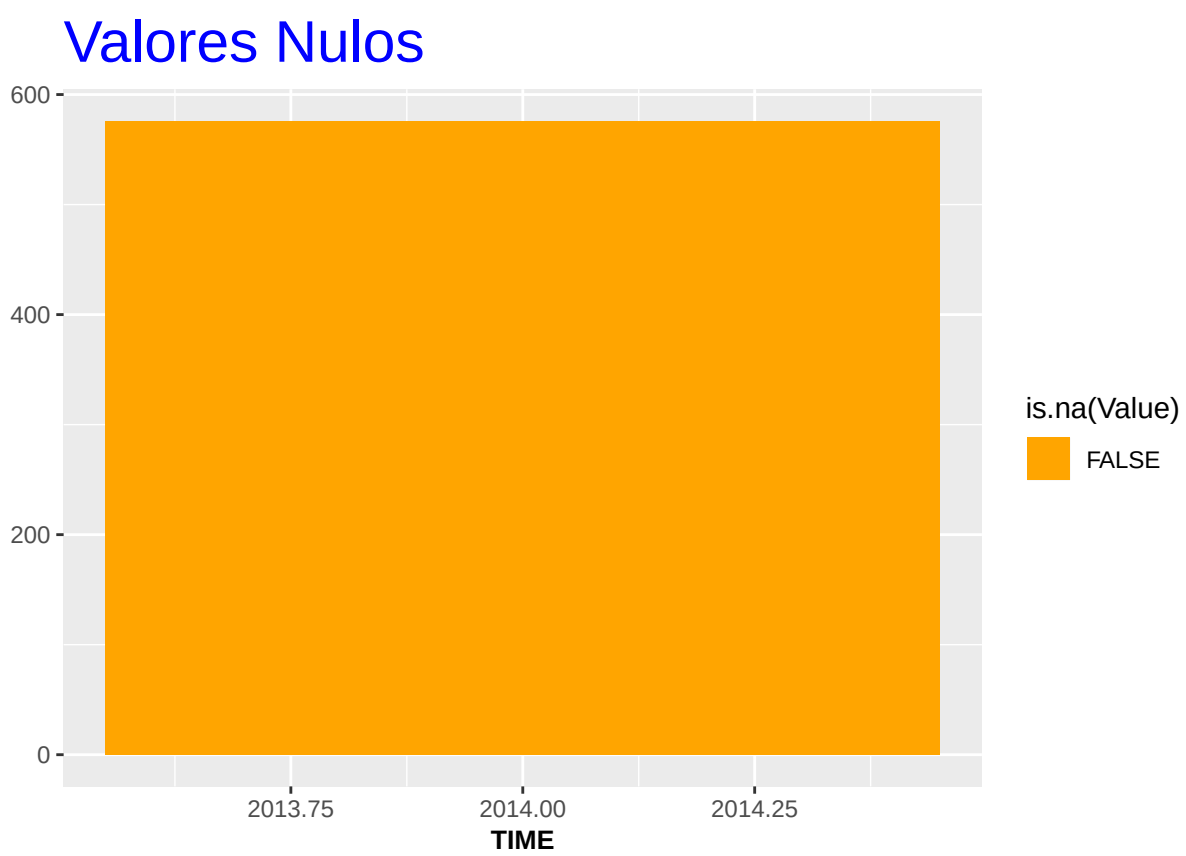
```r
idx<-which(is.na(cardio_diabetes$Value))
length(idx)
```

```
## [1] 0
```

- Grafiquemos la información que contiene la variable **Value**. No hay valores perdidos.

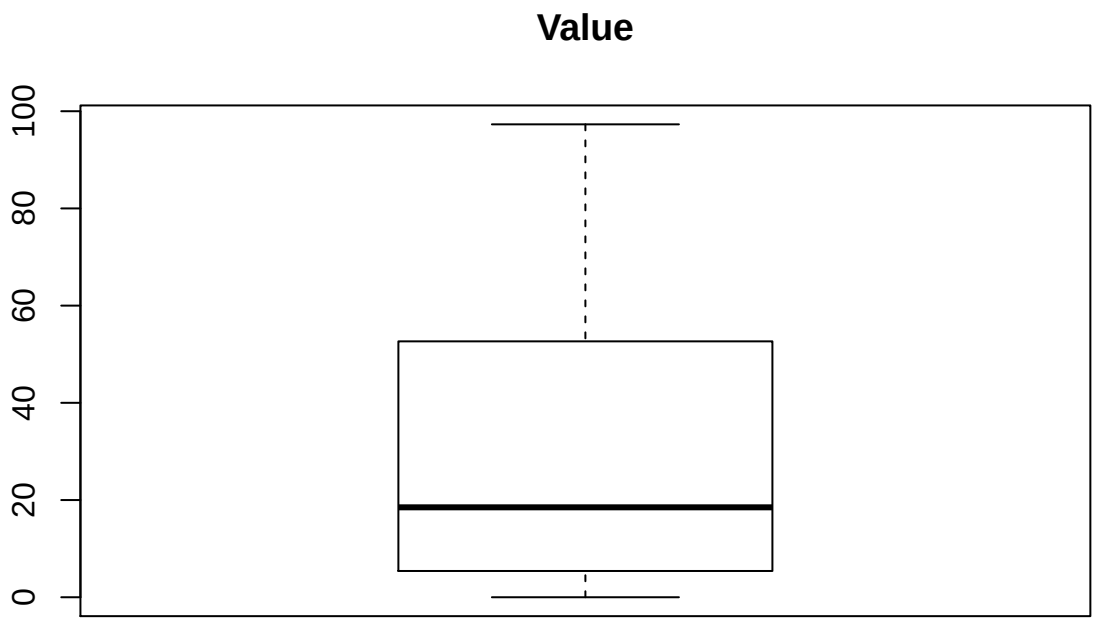
```r
library(ggplot2)
library(scales)
g = ggplot(cardio_diabetes, aes(TIME, fill=is.na(Value)) ) +
labs(title = "Valores Nulos")+ylab("") +
theme(plot.title = element_text(size = rel(2), colour = "blue"))

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** no tiene outliers o valores extremos

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



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

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

```
##
## 2014
##  576
```

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

```
##
##                                        Austria
##                                             18
##                                       Bulgaria
##                                             18
##                                        Croatia
##                                             18
##                                         Cyprus
##                                             18
##                                        Czechia
##                                             18
##                                        Denmark
##                                             18
##                                        Estonia
##                                             18
##        European Union - 27 countries (from 2020)
##                                             18
##        European Union - 28 countries (2013-2020)
```

```
##                                           18
##                                      Finland
##                                           18
##                                       France
##                                           18
## Germany (until 1990 former territory of the FRG)
##                                           18
##                                       Greece
##                                           18
##                                      Hungary
##                                           18
##                                      Iceland
##                                           18
##                                      Ireland
##                                           18
##                                        Italy
##                                           18
##                                       Latvia
##                                           18
##                                    Lithuania
##                                           18
##                                   Luxembourg
##                                           18
##                                        Malta
##                                           18
##                                  Netherlands
##                                           18
##                                       Norway
##                                           18
##                                       Poland
##                                           18
##                                     Portugal
##                                           18
##                                      Romania
##                                           18
##                                     Slovakia
##                                           18
##                                     Slovenia
##                                           18
##                                        Spain
##                                           18
##                                       Sweden
##                                           18
##                                       Turkey
##                                           18
##                                United Kingdom
##                                           18
```

```r
table(cardio_diabetes$UNIT, useNA = "ifany")
```

```
##
## Percentage
##        576
```

```
table(cardio_diabetes$HLTHCARE, useNA = "ifany")
```

```
##
## Blood cholesterol measurement    Blood pressure measurement
##                           192                           192
##       Blood sugar measurement
##                           192
```

```
table(cardio_diabetes$ISCED11, useNA = "ifany")
```

```
##
## All ISCED 2011 levels
##                   576
```

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(cardio_diabetes)
```

```
## 'data.frame':    576 obs. of  7 variables:
##  $ HLTHCARE: Factor w/ 3 levels "Blood cholesterol measurement",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ GEO     : Factor w/ 32 levels "Austria","Bulgaria",..: 8 8 8 8 8 8 9 9 9 9 ...
##  $ UNIT    : Factor w/ 1 level "Percentage": 1 1 1 1 1 1 1 1 1 1 ...
##  $ TIME    : int  2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 ...
##  $ DURATION: Factor w/ 6 levels "5 years or over",..: 4 2 3 5 1 6 4 2 3 5 ...
##  $ ISCED11 : Factor w/ 1 level "All ISCED 2011 levels ": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Value   : num  67.1 19.2 4.1 90.4 3.8 5.8 66.9 19.1 4.3 90.3 ...
```

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

```
write.csv(cardio_diabetes, file="Deteccion_enfCardio_Diabetes_clean.csv", row.names = FALSE)
```