Ratio de altas en hospital de día por diagnóstico, sexo y edad

Alicia Perdices Guerra

3 de mayo, 2021

Contents

1.PROCESAMIENTO DE LOS DATOS.

- En primer lugar leemos el fichero:

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

- Realicemos una breve inspección de los datos

```
str(ratio_altasHD)
```

```
## 'data.frame':    24750 obs. of  9 variables:
##  $ TIME              : int  2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...
##  $ GEO               : Factor w/ 33 levels "Austria","Belgium",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ AGE               : Factor w/ 5 levels "From 1 to 4 years",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ INDIC_HE          : Factor w/ 1 level "Day cases (total number)": 1 1 1 1 1 1 1 1 1 1 ...
##  $ UNIT              : Factor w/ 1 level "Per hundred thousand inhabitants": 1 1 1 1 1 1 1 1 1 1 ...
##  $ SEX               : Factor w/ 3 levels "Females","Males",..: 3 3 3 3 3 2 2 2 2 2 ...
##  $ ICD10             : Factor w/ 5 levels "All causes of diseases (A00-Z99) excluding V00-Y98",..: 1
##  $ Value             : Factor w/ 5938 levels ":","0.0","0.1",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Flag.and.Footnotes: Factor w/ 2 levels "","b": 1 1 1 1 1 1 1 1 1 1 ...
```

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

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

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

```
## [1] 24750
```

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

```
## [1] 9
```

*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.RAtio (por 100.000 hab)
- **AGE**: variable cualitativa. Indica la edad del paciente.
- **INDIC_HE**: variable cualitativa. Hace referencia al número total de casos en hospital de día.
- **ICD10**:variable cualitativa. En relación al tipo de enfermedad diagnosticada.
- **Value**: Variable cuantitativa. Indica el ratio de pacientes en hospital de día por diagnóstico.

- **Fal.and.footnotes**. Notas sobre etiquetas. Eliminamos esta columna.

*Años de las mediciones:

```
unique(ratio_altasHD$TIME)
```

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

*Paises:

```
unique(ratio_altasHD$GEO)
```

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

*Unidad de las mediciones:

```
unique(ratio_altasHD$UNIT)
```

```
## [1] Per hundred thousand inhabitants
## Levels: Per hundred thousand inhabitants
```

- Edad del paciente.

```
unique(ratio_altasHD$AGE)
```

```
## [1] Total               Less than 1 year    From 1 to 4 years
```

```
## [4] From 5 to 9 years   From 10 to 14 years
## 5 Levels: From 1 to 4 years From 10 to 14 years ... Total
```

- Número total de casos en hospital de día.

```
unique(ratio_altasHD$INDIC_HE)
```

```
## [1] Day cases (total number)
## Levels: Day cases (total number)
```

- En relación al tipo de enfermedad diagnosticada

```
unique(ratio_altasHD$ICD10)
```

```
## [1] All causes of diseases (A00-Z99) excluding V00-Y98
## [2] All causes of diseases (A00-Z99) excluding V00-Y98 and Z38
## [3] Certain infectious and parasitic diseases (A00-B99)
## [4] Tuberculosis
## [5] Intestinal infectious diseases except diarrhoea
## 5 Levels: All causes of diseases (A00-Z99) excluding V00-Y98 ...
```

- Eliminamos la columna Fal.and.footnotes.

```
ratio_altasHD<-ratio_altasHD[,-9]
```

- 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.

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

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

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

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

```
##
## 717751 754188 786597 832232 905996   <NA>
##      1      1      1      1      1   9046
```

- Observamos que tenemos **9046 valores perdidos**.Guardamos en la variable **idx** los índices de los registros con valores **NA** de la variable **Value**.

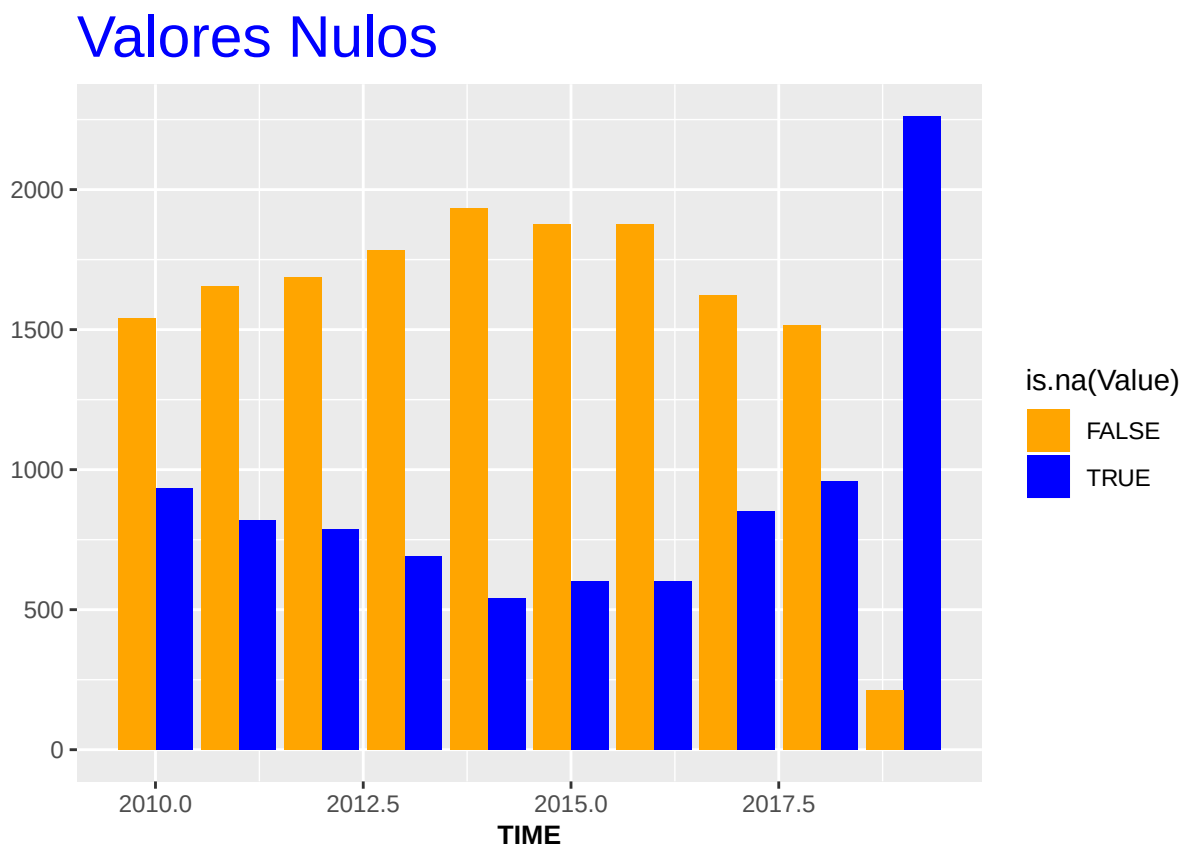
```
idx<-which(is.na(ratio_altasHD$Value))
length(idx)
```

```
## [1] 9046
```

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

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

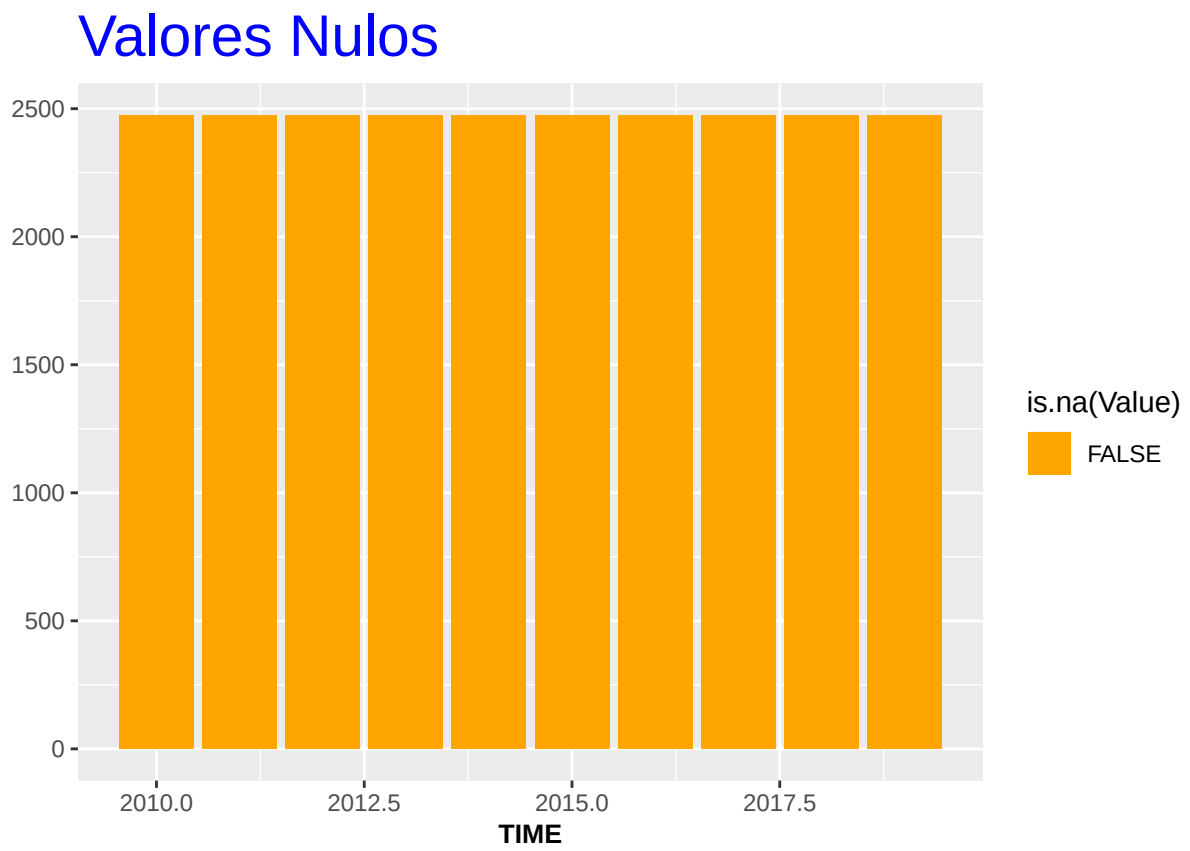
```
## Loading required package: colorspace
## Loading required package: grid
## VIM is ready to use.
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
##
## Attaching package: 'VIM'
## The following object is masked from 'package:datasets':
##
##     sleep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

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

- Comprobamos que no tenemos valores nulos después de la imputación

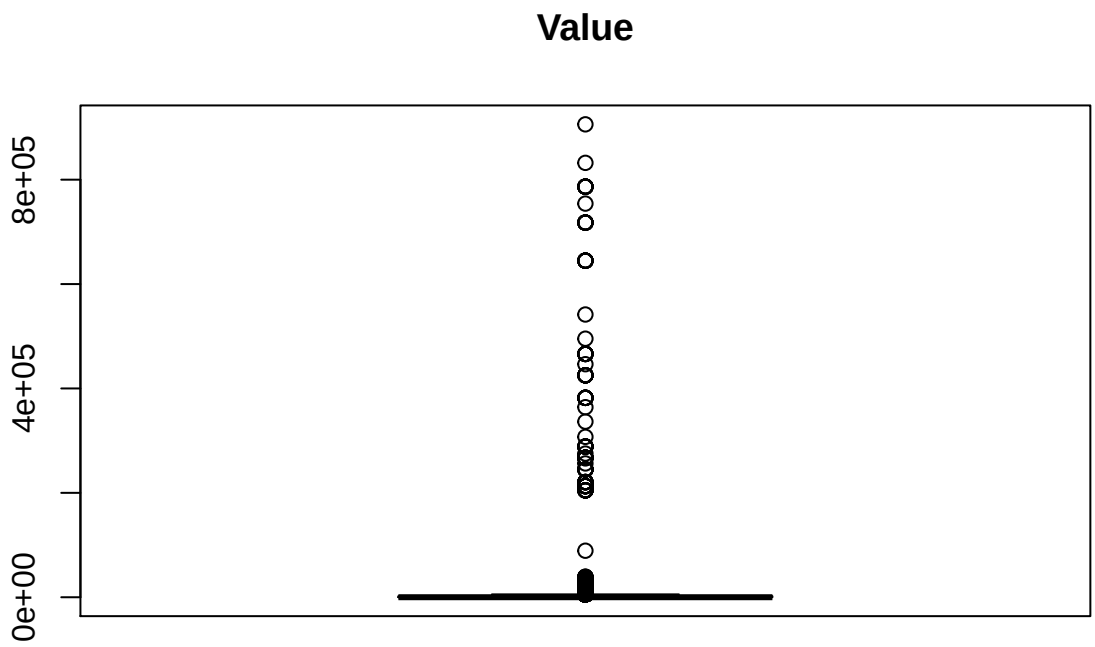
```
g = ggplot(ratio_altasHD, aes(TIME, fill=is.na(Value)) ) +
labs(title = "Valores Nulos")+ylab("") +
theme(plot.title = element_text(size = rel(2), colour = "blue"))
```

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



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

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

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
## 2475 2475 2475 2475 2475 2475 2475 2475 2475 2475
```

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

```
##
##                                         Austria
##                                             750
##                                         Belgium
##                                             750
##                                         Croatia
##                                             750
##                                          Cyprus
##                                             750
##                                         Czechia
##                                             750
##                                         Denmark
##                                             750
##                                         Estonia
##                                             750
##                                         Finland
##                                             750
##                                          France
```

```
##                                             750
## Germany (until 1990 former territory of the FRG)
##                                             750
##                                          Greece
##                                             750
##                                         Hungary
##                                             750
##                                         Iceland
##                                             750
##                                         Ireland
##                                             750
##                                           Italy
##                                             750
##                                          Latvia
##                                             750
##                                       Lithuania
##                                             750
##                                      Luxembourg
##                                             750
##                                           Malta
##                                             750
##                                     Netherlands
##                                             750
##                                  North Macedonia
##                                             750
##                                          Norway
##                                             750
##                                          Poland
##                                             750
##                                        Portugal
##                                             750
##                                         Romania
##                                             750
##                                          Serbia
##                                             750
##                                        Slovakia
##                                             750
##                                        Slovenia
##                                             750
##                                           Spain
##                                             750
##                                          Sweden
##                                             750
##                                     Switzerland
##                                             750
##                                          Turkey
##                                             750
##                                  United Kingdom
##                                             750
```

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

```
##
## Per hundred thousand inhabitants
##                            24750
```

```
table(ratio_altasHD$AGE, useNA = "ifany")
```

```
##
##   From 1 to 4 years From 10 to 14 years   From 5 to 9 years    Less than 1 year
##                4950                4950                4950                4950
##               Total
##                4950
```

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

```
##
## Day cases (total number)
##                   24750
```

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

```
##
##         All causes of diseases (A00-Z99) excluding V00-Y98
##                                                       4950
## All causes of diseases (A00-Z99) excluding V00-Y98 and Z38
##                                                       4950
##        Certain infectious and parasitic diseases (A00-B99)
##                                                       4950
##            Intestinal infectious diseases except diarrhoea
##                                                       4950
##                                                 Tuberculosis
##                                                       4950
```

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(ratio_altasHD, file="RatioAltas_HospitalDia_Diagnostico_clean.csv", row.names = FALSE)
```