

Altas, Estancias y Camas Hospitalarias en Pacientes Ingresados y por Servicios Curativos

Alicia Perdices Guerra

3 de mayo, 2021

Contents

1.PROCESAMIENTO DE LOS DATOS.

- En primer lugar leemos el fichero:

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

- Realicemos una breve inspección de los datos

```
str(aec)
```

```
## 'data.frame': 2340 obs. of 7 variables:
## $ TIME : int 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...
## $ GEO : Factor w/ 39 levels "Albania","Austria",...: 10 10 10 10 10 10 11 11 11 11 ...
## $ ICHA_HC : Factor w/ 2 levels "In-patient care",...: 2 2 2 1 1 1 2 2 2 1 ...
## $ INDIC_HE : Factor w/ 3 levels "Bed-days","Discharges",...: 2 3 1 2 3 1 2 3 1 2 ...
## $ UNIT : Factor w/ 1 level "Number": 1 1 1 1 1 1 1 1 1 1 ...
## $ Value : Factor w/ 1141 levels ":", "1 009 687.00",...: 1 1 1 946 1 1 1 1 1 1053 ...
## $ Flag.and.Footnotes: Factor w/ 6 levels "","b","bd","d",...: 1 1 1 1 1 1 1 1 1 1 ...
```

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

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

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

```
## [1] 2340
```

```
ncol(aec) #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.
- **ICHA_HC**: variable cualitativa. Indica si la medida realizada es sobre paciente ingresado o no.
- **INDIC_HE**: variable cualitativa. Hace referencia a las altas, estancia media hospitalaria y ocupación de camas.
- **Value**: Variable cuantitativa. Indica el número de altas, estancias hospitalarias y camas. Se ha cargado mal como factor.
- **Fal.and.footnotes**. Notas sobre etiquetas. Eliminamos esta columna.

*Años de las mediciones:

```
unique(aec$TIME)
```

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

*Países:

```
unique(aec$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] Montenegro
## [36] North Macedonia
## [37] Albania
## [38] Serbia
## [39] Turkey
## 39 Levels: Albania Austria Belgium Bulgaria Croatia Cyprus Czechia ... United Kingdom
```

*Unidad de las mediciones:

```
unique(aec$UNIT)
```

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

- Medida sobre paciente ingresado o relacionado con servicios curativos en general:

```
unique(aec$ICHA_HC)
```

```
## [1] Services of curative care In-patient care  
## Levels: In-patient care Services of curative care
```

*Hacer referencia a altas, a la estancia media del paciente ingresado y a la ocupación de camas (hospital de día)

```
unique(aec$INDIC_HE)
```

```
## [1] Discharges  
## [2] In-patient average length of stay (in days)  
## [3] Bed-days  
## Levels: Bed-days Discharges In-patient average length of stay (in days)
```

- Eliminamos la columna Fal.and.footnotes.

```
aec<-aec[, -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.

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

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

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

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

```
##  
## 145019306 145019320 145355219 145701290 146048193 <NA>  
##          1          1          1          1          1      892
```

- Observamos que tenemos **892 valores perdidos**. Guardamos en la variable **idx** los índices de los registros con valores **NA** de la variable **Value**.

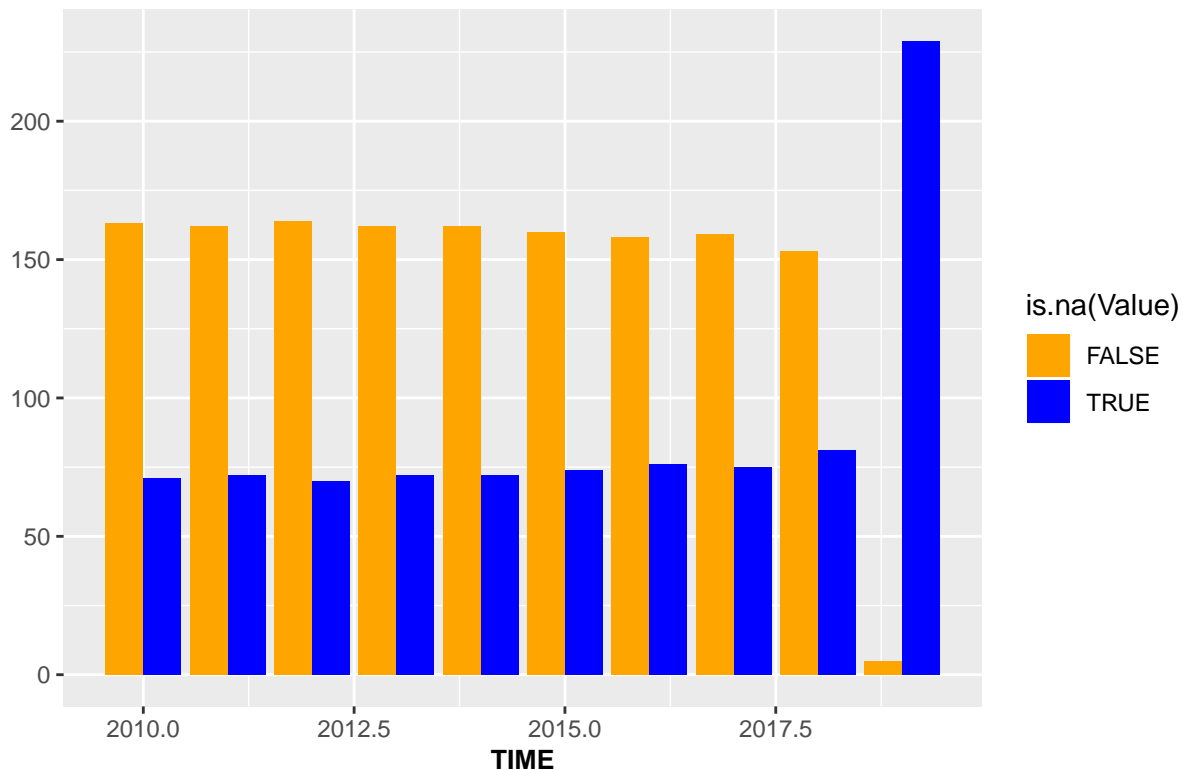
```
idx<-which(is.na(aec$Value))  
length(idx)
```

```
## [1] 892
```

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

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

Valores Nulos



- 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 robusta 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(aec, variable=c("Value"),k=3)
```

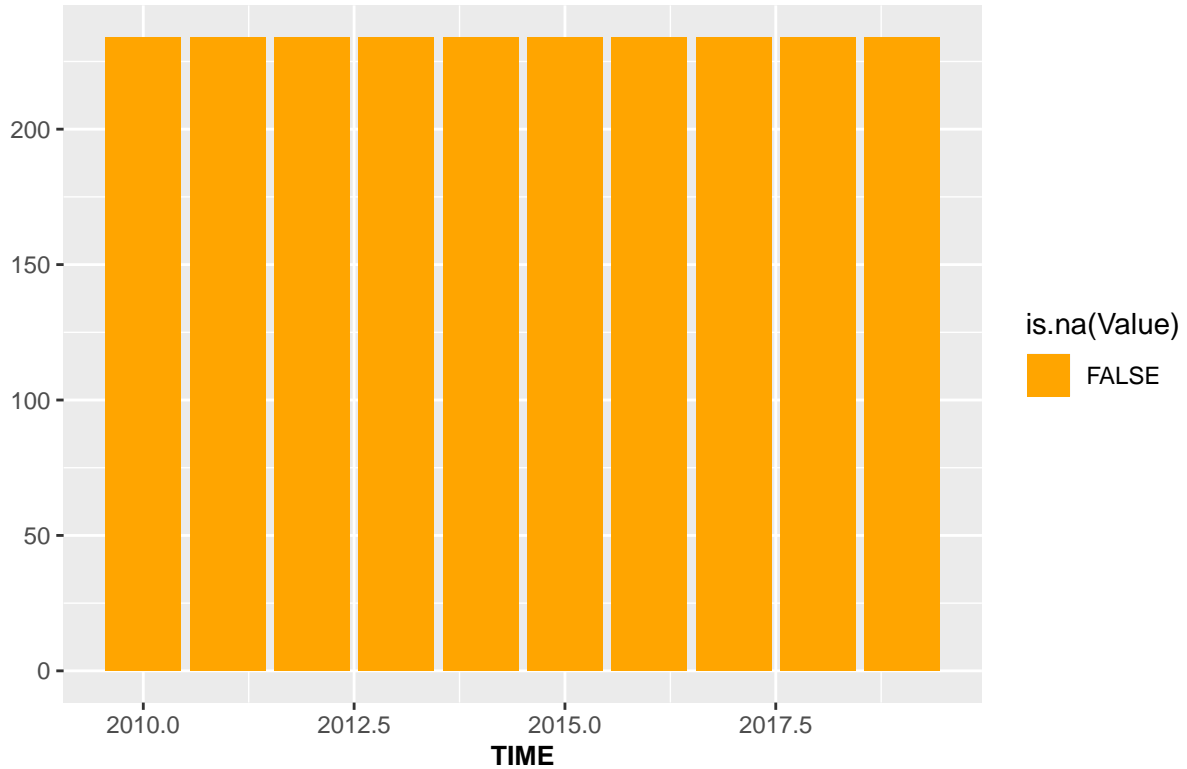
```
aec<-output
```

- Comprobamos que no tenemos valores nulos después de la imputación

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

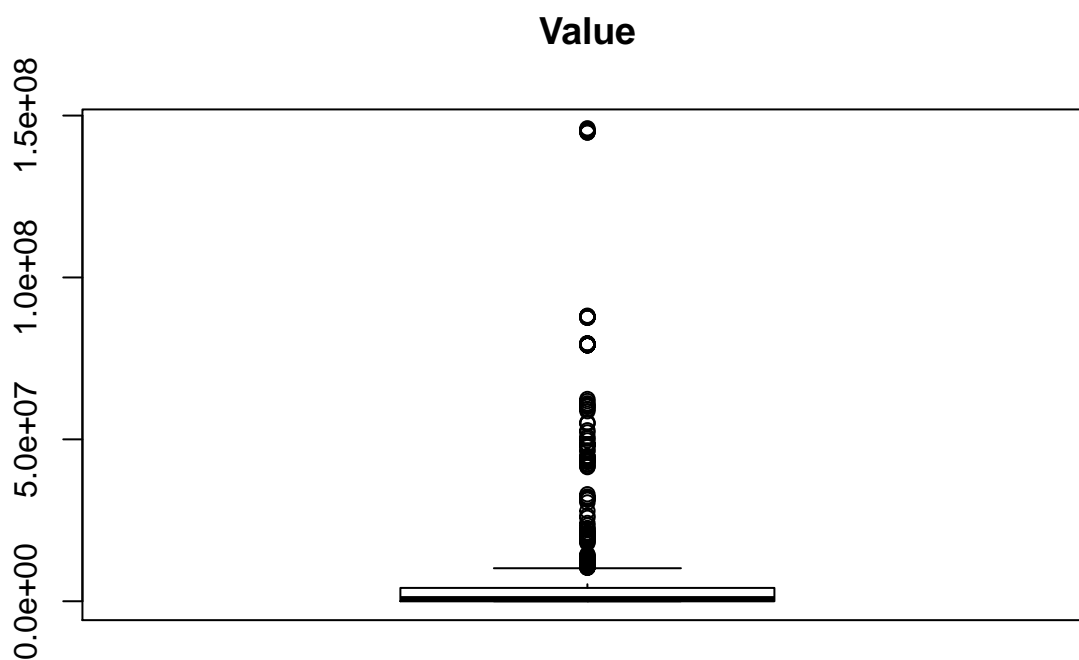
```
g+geom_bar(position="dodge") + scale_fill_manual(values = alpha(c("orange", "blue"), 1)) +
theme(axis.title.x = element_text(face="bold", size=10))
```

Valores Nulos



- Con el siguiente gráfico, observaremos que la variable **Value** no tiene outliers o valores extremos

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



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

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

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
## 234 234 234 234 234 234 234 234 234 234
```

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

```
##
## Albania
## 60
## Austria
## 60
## Belgium
## 60
## Bulgaria
## 60
## Croatia
## 60
## Cyprus
## 60
## Czechia
## 60
## Denmark
## 60
## Estonia
```

##		60
##	European Union - 27 countries (from 2020)	
##		60
##	European Union - 28 countries (2013-2020)	
##		60
##	Finland	
##		60
##	France	
##		60
##	Germany (until 1990 former territory of the FRG)	
##		60
##	Greece	
##		60
##	Hungary	
##		60
##	Iceland	
##		60
##	Ireland	
##		60
##	Italy	
##		60
##	Latvia	
##		60
##	Liechtenstein	
##		60
##	Lithuania	
##		60
##	Luxembourg	
##		60
##	Malta	
##		60
##	Montenegro	
##		60
##	Netherlands	
##		60
##	North Macedonia	
##		60
##	Norway	
##		60
##	Poland	
##		60
##	Portugal	
##		60
##	Romania	
##		60
##	Serbia	
##		60
##	Slovakia	
##		60
##	Slovenia	
##		60
##	Spain	
##		60
##	Sweden	

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

```
table(aec$ICHA_HC, useNA = "ifany")
```

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

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

8