Importación y Tratamiento de datos

Yendi Lestrade Rodal

11/10/2021

Vamos a utilizar como ejemplo: **iris**, que es una matriz de datos precargada en R.

Abrir matriz de datos.

```
library(datasets)
```

```
data(iris)
```

Exploración de los datos iris.

1.- Dimensión de la matriz.

```
dim(iris)
```

```
## [1] 150   5
```

2.- Nombre de las columnas.

```
colnames(iris)
```

```
## [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
```

```
names(iris)
```

```
## [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
```

3.- Clase a la que pertenece la matriz de datos.

```
class(iris)
```

```
## [1] "data.frame"
```

4.- Estructura interna.

```
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 ...
```

5.- Observación de una variable específica.

```
iris$Species
```

```
##   [1] setosa     setosa     setosa     setosa     setosa     setosa
##   [7] setosa     setosa     setosa     setosa     setosa     setosa
##  [13] setosa     setosa     setosa     setosa     setosa     setosa
##  [19] setosa     setosa     setosa     setosa     setosa     setosa
##  [25] setosa     setosa     setosa     setosa     setosa     setosa
##  [31] setosa     setosa     setosa     setosa     setosa     setosa
##  [37] setosa     setosa     setosa     setosa     setosa     setosa
##  [43] setosa     setosa     setosa     setosa     setosa     setosa
##  [49] setosa     setosa     versicolor versicolor versicolor versicolor
##  [55] versicolor versicolor versicolor versicolor versicolor versicolor
##  [61] versicolor versicolor versicolor versicolor versicolor versicolor
##  [67] versicolor versicolor versicolor versicolor versicolor versicolor
##  [73] versicolor versicolor versicolor versicolor versicolor versicolor
##  [79] versicolor versicolor versicolor versicolor versicolor versicolor
##  [85] versicolor versicolor versicolor versicolor versicolor versicolor
##  [91] versicolor versicolor versicolor versicolor versicolor versicolor
##  [97] versicolor versicolor versicolor versicolor virginica  virginica
## [103] virginica  virginica  virginica  virginica  virginica  virginica
## [109] virginica  virginica  virginica  virginica  virginica  virginica
## [115] virginica  virginica  virginica  virginica  virginica  virginica
## [121] virginica  virginica  virginica  virginica  virginica  virginica
## [127] virginica  virginica  virginica  virginica  virginica  virginica
## [133] virginica  virginica  virginica  virginica  virginica  virginica
## [139] virginica  virginica  virginica  virginica  virginica  virginica
## [145] virginica  virginica  virginica  virginica  virginica  virginica
## Levels: setosa versicolor virginica
```

6.- Visualización de tabla.

```
View(iris)
```

7.- Estadística descriptiva básica.

```
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
```

```
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

Datos faltantes.

1.- Búsqueda de datos faltantes.

```
anyNA(iris)
```

```
## [1] FALSE
```

Nota: Se le pregunta a R si hay datos faltantes **(NA)**, R me responde con **FALSE** en caso de NO HABER NA's y **TRUE** en caso de HABER NA's.

2.- Suma de datos faltantes.

```
sum(is.na(iris))
```

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

3.- Librería **mice**.

a) Instalar el paquete **mice**

b) Función **md.pattern(iris)**, pero no me compila en el pdf, así que voy a insertar el gráfico de NA's como imagen. Nota: Se activó la función desde un script sencillo.

Datos atípicos.

a) Detección. Se detectan con el gráfico boxplot.

```
bx1<-boxplot(iris)
```

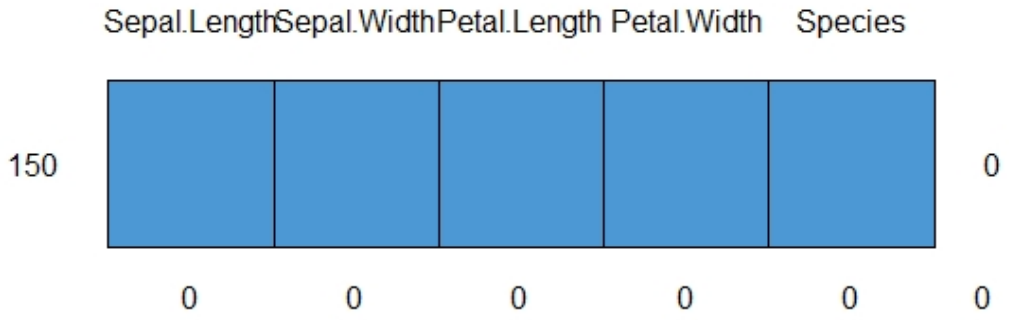
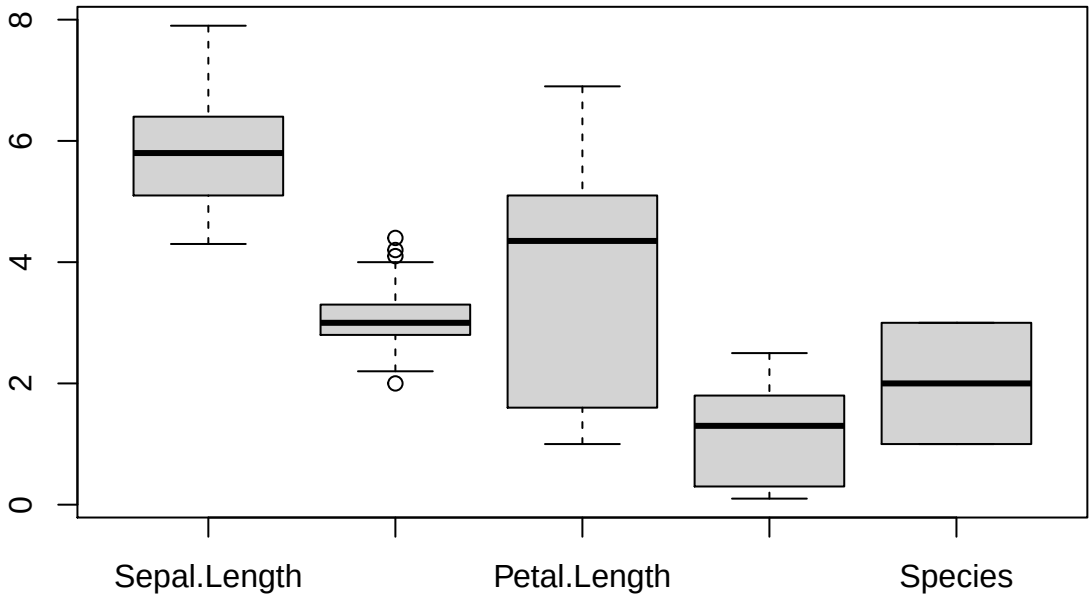


Figure 1: Gráfico de datos perdidos

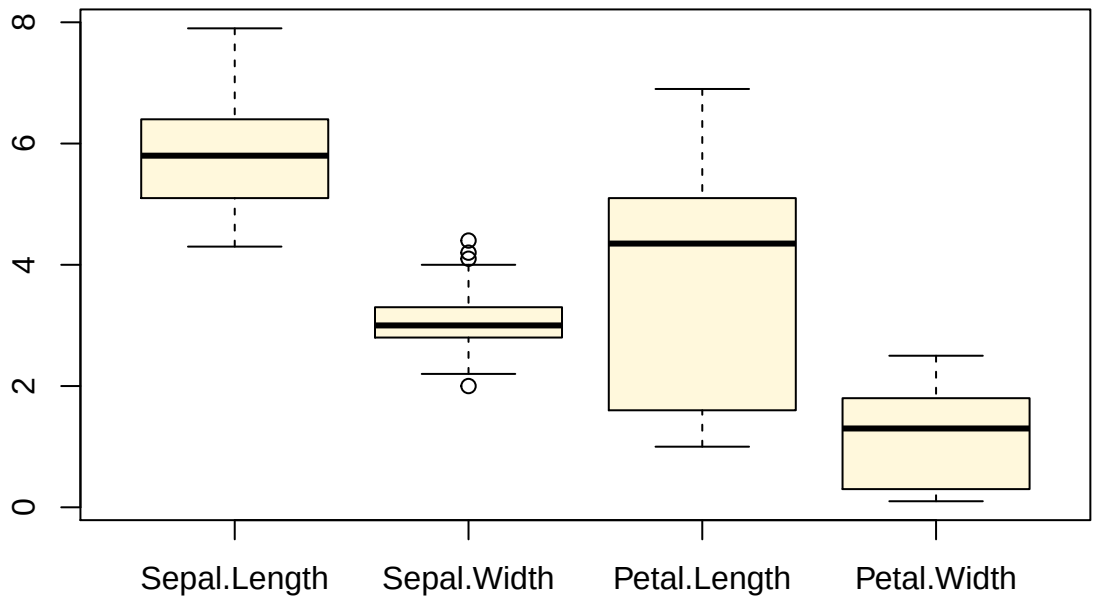


```
bx1
```

```
## $stats
##      [,1] [,2] [,3] [,4] [,5]
## [1,]  4.3  2.2 1.00  0.1    1
## [2,]  5.1  2.8 1.60  0.3    1
## [3,]  5.8  3.0 4.35  1.3    2
## [4,]  6.4  3.3 5.10  1.8    3
## [5,]  7.9  4.0 6.90  2.5    3
##
## $n
## [1] 150 150 150 150 150
##
## $conf
##          [,1]     [,2]     [,3]    [,4]     [,5]
## [1,] 5.632292 2.935497 3.898477 1.10649 1.741987
## [2,] 5.967708 3.064503 4.801523 1.49351 2.258013
##
## $out
## [1] 4.4 4.1 4.2 2.0
##
## $group
## [1] 2 2 2 2
##
## $names
## [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
```

b) filtrado de variables para realizar el boxplot

```
bx2<-boxplot(iris[,c(1:4)], col = "cornsilk1")
```




```
bx2
```

```
## $stats
##      [,1] [,2] [,3] [,4]
## [1,]  4.3  2.2 1.00  0.1
## [2,]  5.1  2.8 1.60  0.3
## [3,]  5.8  3.0 4.35  1.3
## [4,]  6.4  3.3 5.10  1.8
## [5,]  7.9  4.0 6.90  2.5
##
## $n
## [1] 150 150 150 150
##
## $conf
##          [,1]     [,2]     [,3]    [,4]
## [1,] 5.632292 2.935497 3.898477 1.10649
## [2,] 5.967708 3.064503 4.801523 1.49351
##
## $out
## [1] 4.4 4.1 4.2 2.0
##
## $group
## [1] 2 2 2 2
##
## $names
## [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"
```