Análisis exploratorio de los datos Iris

Máster Data Science

29 de septiembre de 2021

Conjunto de datos

Trabajaremos con el conjunto de datos **Iris** de Fisher que contiene información sobre 3 clases de flores: *setosa*, *virgínica* y *versicolor*. Estos datos ya están disponibles en R por lo que simplemente hay que cargarlos:

```
library(datasets)
data(iris)
```

Visualizamos las primeras líneas del conjunto de datos mediante el comando *head()* de R:

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

Está formado por un total de 150 datos y 5 variables:

- Species: setosa, virginica, versicolor
- Sepal.Length: longitud en cm del sépalo
- Sepal.Width: anchura en cm del sépalo
- Petal.Length: longitud en cm del pétalo
- Petal.Width: anchura en cm del pétalo

De este modo, el conjunto de datos dispone de una variable categórica (*Species*) que toma 3 posibles valores y de 4 variables continuas (*Sepal.Length*, *Sepal.Width*, *Petal.Length* y *Petal.Width*).

Análisis univariante

Comenzamos realizando un análisis univariante de los datos. Esto nos permitirá saber el rango de valores de cada variable, sus valores medios, su dispersión, si tienen datos faltantes, valores atípicos, etc.

Lo primero es verificar que las variables están guardadas en el formato correspondiente, es decir, que las variables categóricas están guardadas como categóricas y las continuas como continuas. Esto se comprueba con mediante el siguiente comando de R:

```
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 1 ...
```

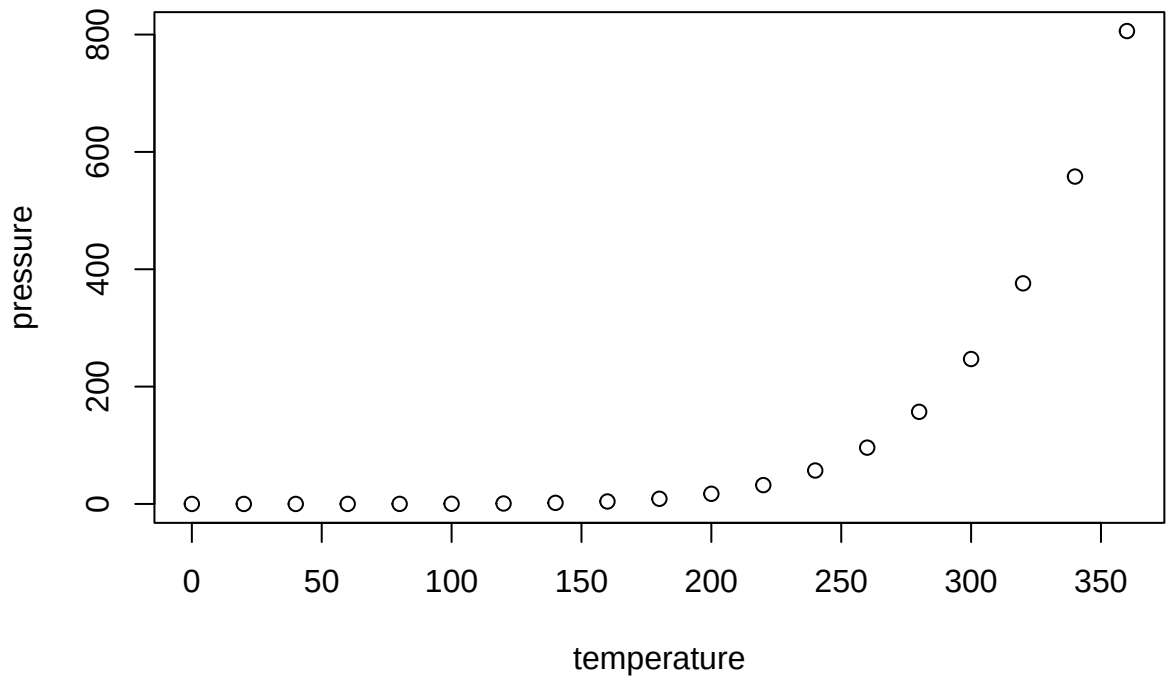
Vemos que *Sepal.Length*, *Sepal.Width*, *Petal.Length* y *Petal.Width* aparecen como variables numérica y *Species* como un factor (variable categórica con 3 niveles).

Para tener un resumen rápido de las variables usamos la función *summary()*:

```
##   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
##
##
##
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.