Análisis exploratorio de los datos Iris

Máster Data Science

22 de septiembre de 2022

DSLAB
DATA SCIENCE LABORATORY

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:

```
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). De no ser así, habría que guardarlas con el formato correspondiente.

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

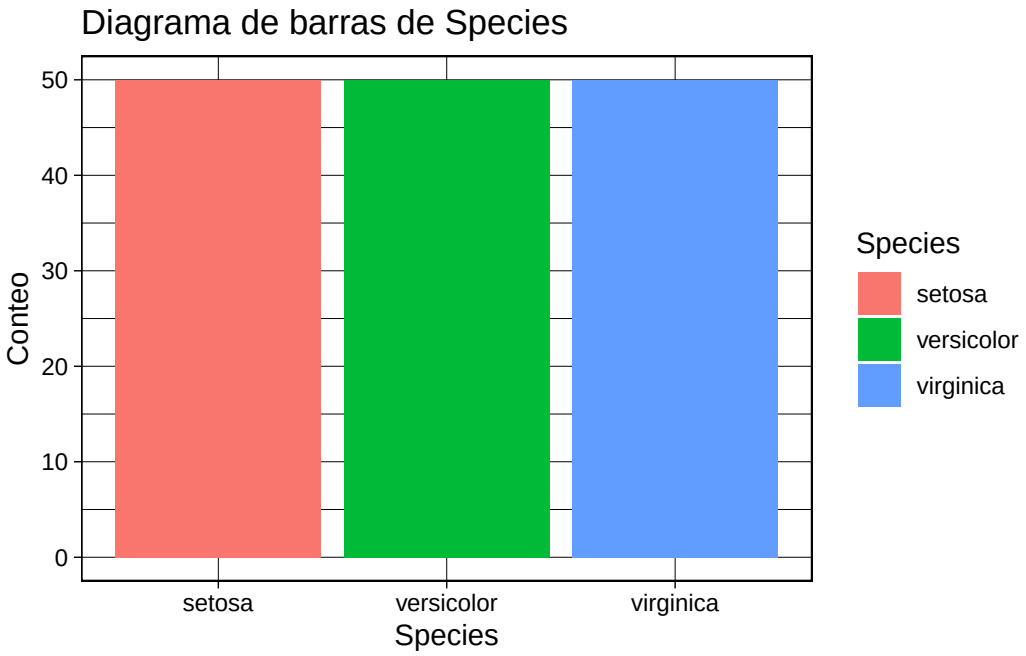
```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   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
##
##
##
```

El resumen indica que hay 50 datos de cada tipo de flor. Las variables de longitud toman valores mayores que las de anchura y, en ambos casos, los sépalos se mueven en unos valores en general más altos que los pétalos. Por ejemplo, tanto la media y la mediana de *Sepal.Length* son mayores que la media y la mediana de *Petal.Length*. Lo mismo ocurre con *Sepal.Width* y *Petal.Width*. En las variables sobre los sépalos sucede que la media y la mediana están próximas, esto significa que su distribución es bastante simétrica. Los casos menos simétricos serían los de las variables de pétalos puesto que es donde más difieren la media y la mediana.

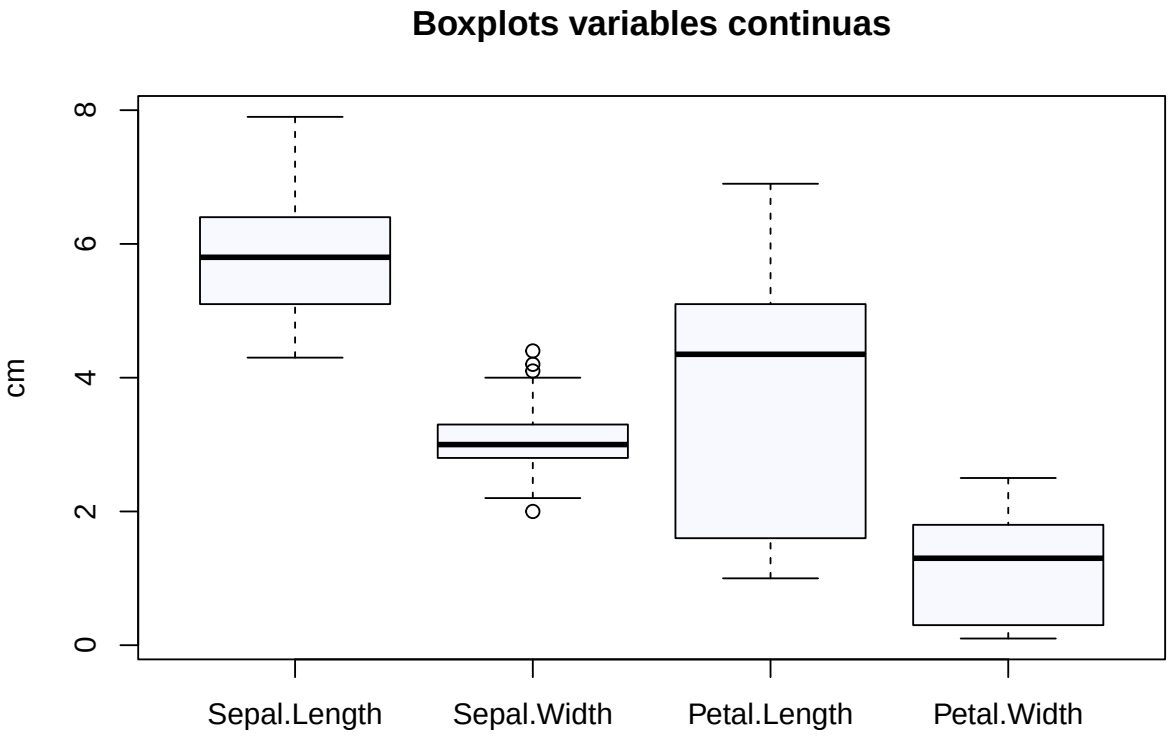
Vemos además que ninguna de las variables cuenta con datos faltantes puesto que no aparecen en el summary. Veamos un ejemplo de cómo aparecería en el caso de haber NA's:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   4.300   5.100   5.800   5.843   6.400   7.900       1
```

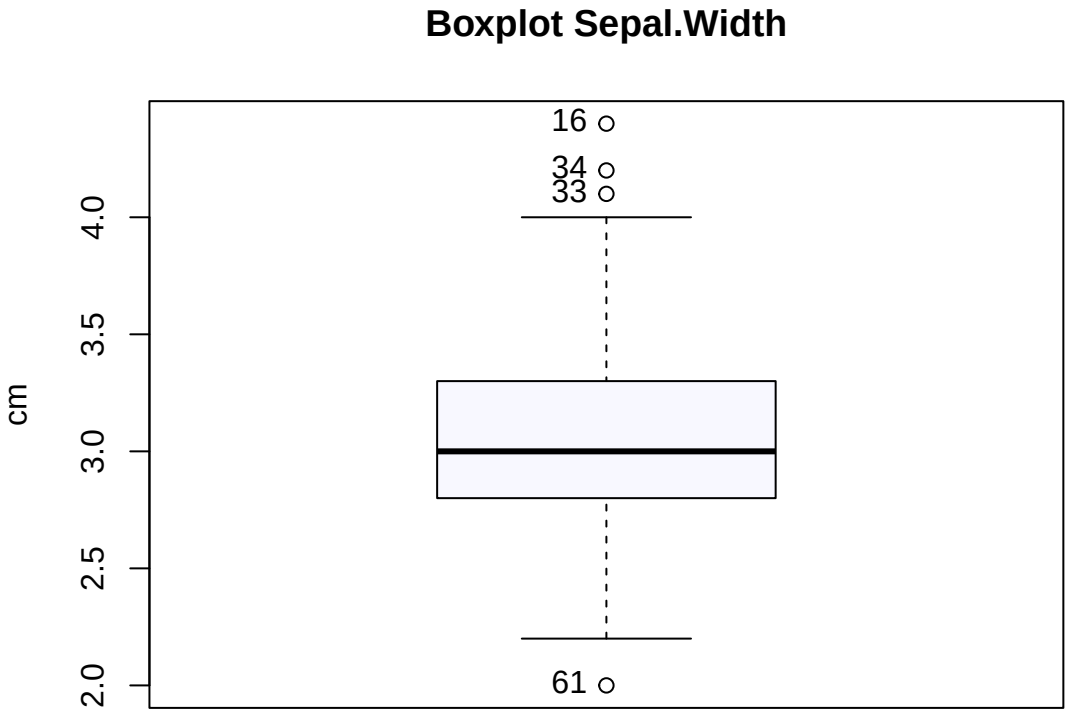
Analizamos ahora las variables desde el punto de vista gráfico. Como la variable categórica *Species* tiene 50 datos en cada clase, realizar un gráfico resulta innecesario. En este caso lo mostramos para tener un ejemplo.



Mostramos ahora el boxplot para cada variable continua.



En este caso, hemos pintado los 4 boxplots en el mismo gráfico porque las 4 variables continuas se miden en centímetros y porque el rango de valores de las 4 no es muy dispar. Gracias al gráfico confirmamos que las distribuciones más simétricas son las de las variables de los sépalos. La longitud de pétalo es la que mayor variabilidad de valores muestra y la anchura de pétalos contiene algunos valores atípicos. Para estudiar con más detalle dichos atípicos, hacemos un boxplot independiente de dicha variable en el que identificamos esos puntos.



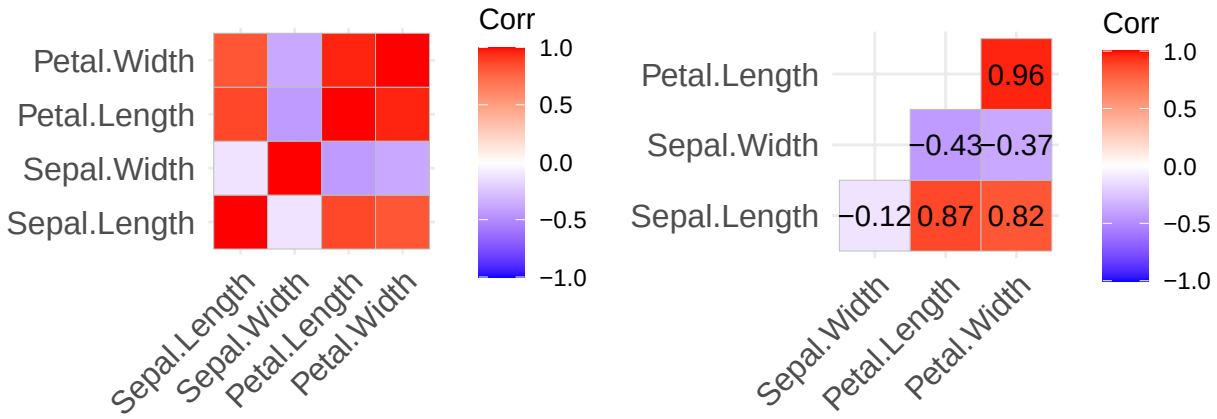

```
## [1] "61" "16" "33" "34"
```

Obtenemos ahora los histogramas de las 4 variables continuas.

Los histogramas muestran que la variable más simétrica es *Sepal.Width*, con media y mediana en torno al 3 y unas colas similares a ambos lados. *Sepal.Length* tiene mayor concentración de valores a la izquierza de la media, es decir, hay menos observaciones con longitud de sépalo mayor que 6 que menor. Los histogramas sobre las variables sobre los pétalos muestran que hay un número importante de flores (más de 40) con valores bajitos de longitud y anchura de pétalo y el resto de flores ya se concentran en valores más altos (una longitud mayor de 3 y una anchura en torno a 1 y mayor que 1).

Análisis multivariante

Hacemos un gráfico de correlaciones para estudiar la relación linear existente entre las variables. Ambos gráficos muestran la misma información, la diferencia es que en el de la derecha se ha eliminado la información superflua puesto que la matriz de correlaciones es simétrica. Además, en el gráfico de la derecha se ha añadido una etiqueta con el valor de la correlación de Pearson para tener más información. Vemos que *Petal.Length* y *Petal.Width* están muy correladas y *Sepal.Length* y *Sepal.Width* tienen una correlación prácticamente nula. Las variables de longitudes también muestran una correlación positiva alta (*Petal.Length* y *Sepal.Length*), esto es, cuando una crece la otra también. La anchura de pétalo y la longitud del sépalo también tienen correlación positiva alta. *Sepal.Width* es la variable que menor correlación muestra con el resto.



Usando la función `ggpairs()` de ggplot podemos obtener un cuadro general con diversos gráficos multivariantes. En particular, contiene gráficos boxplots para cada variable continua diferenciando por clases, diagramas de dispersión con todos los cruces de variables continuas diferenciando por clase, gráficos de densidad e histogramas también distinguiendo por clases y los valores de las correlaciones filtrando y sin filtrar por clase. Este cuadro revela que las flores más parecidas entre sí son versicolor y virgínica puesto que en todos los caos muestran gráficos más similares y comparten valores. Como reflejan los diagramas de dispersión, setosa está claramente más separada y su identificación es más sencilla.

A modo de ejemplo, ampliamos uno de los gráficos de dispersión y le añadimos líneas de densidad:

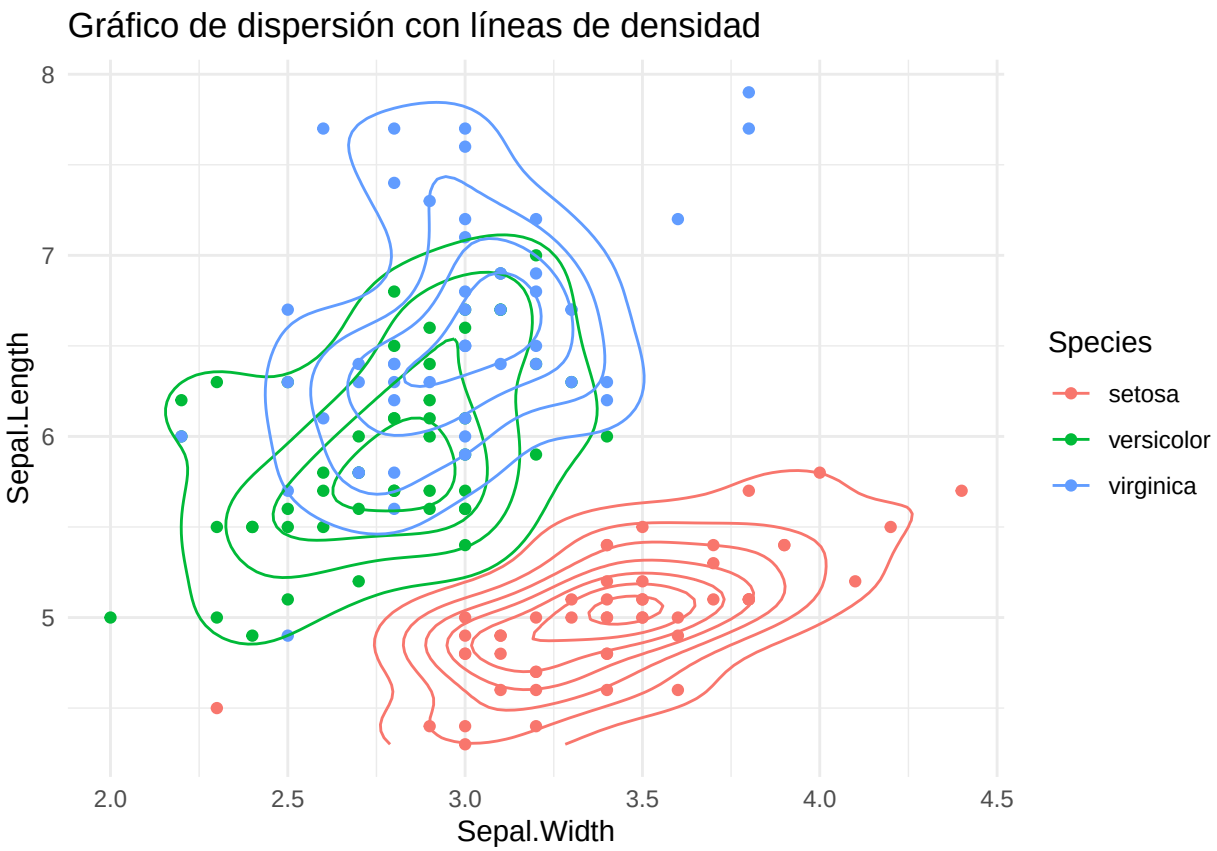


Gráfico de dispersión con líneas de densidad
Sepal.Length
Sepal.Width
8
7
6
5
2.0
2.5
3.0
3.5
4.0
4.5
Species
setosa
versicolor
virginica