Análisis exploratorio

Wikipedia:

> El análisis exploratorio de datos es una forma de analizar datos definido por John W. Tukey (E.D.A.: Exploratory data analysis) es el tratamiento estadístico al que se someten las muestras recogidas durante un proceso de investigación en cualquier campo científico. Para mayor rapidez y precisión, todo el proceso suele realizarse por medios informáticos, con aplicaciones específicas para el tratamiento estadístico. Los E.D.A., no necesariamente, se llevan a cabo con una base de datos al uso, ni con una hoja de cálculo convencional; no obstante el programa SPSS y R (lenguaje de programación) son las aplicaciones más utilizadas, aunque no las únicas.

Tras recoger una muestra de datos, antes de procesarlos para la tarea correspondiente, es necesario hacer primero un análisis exploratorio. En este análisis lo que hacemos es *explorar* los datos, mirándolos "a ojo" para descubrir su estructura y relación.

Pero los números por sí solos no dicen nada, o peor, pueden llevarnos a conclusiones equivocadas. Por ello, aunque el análisis sea "a ojo", no quiere decir que no se pueda hacer de manera formal. Existen herramientas estadísticas que nos permiten hacer este análisis de manera rigurosa, para que las conclusiones (**¡eso sí, preliminares!**) que extraigamos de los datos sean sólidas.

Si los datos que tenemos son muchos, este análisis es arduo de hacer a mano. Afortunadamente, la mayor parte de las herramientas estadísticas que queramos usar están implementadas computacionalmente. Veamos un ejemplo:

Dataset

El conjunto de datos "*Iris*" es un conjunto de datos multivariante usado muy a menudo como un ejemplo de análisis discriminante. Nosotros lo vamos a utilizar para el análisis exploratorio, debido a su pequeño tamaño y facilidad de uso, y a que está contenido en *R* por defecto.

Este conjunto de datos se refiere a una población de especímenes de flor de tres especies distintas. Sobre estos *individuos*, se realizaron cuatro medidas: la lóngitud y anchura de los sépalos, y las de los pétalos. Además, cada especimen viene identificado con el nombre de la especie a la que pertenece.

In [14]:
```
iris[0:10,]
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa

El primer análisis que podemos hacer es observar la distribución observada de las variables, usando el commando `summary`.

In [2]:
```
summary(iris)
```

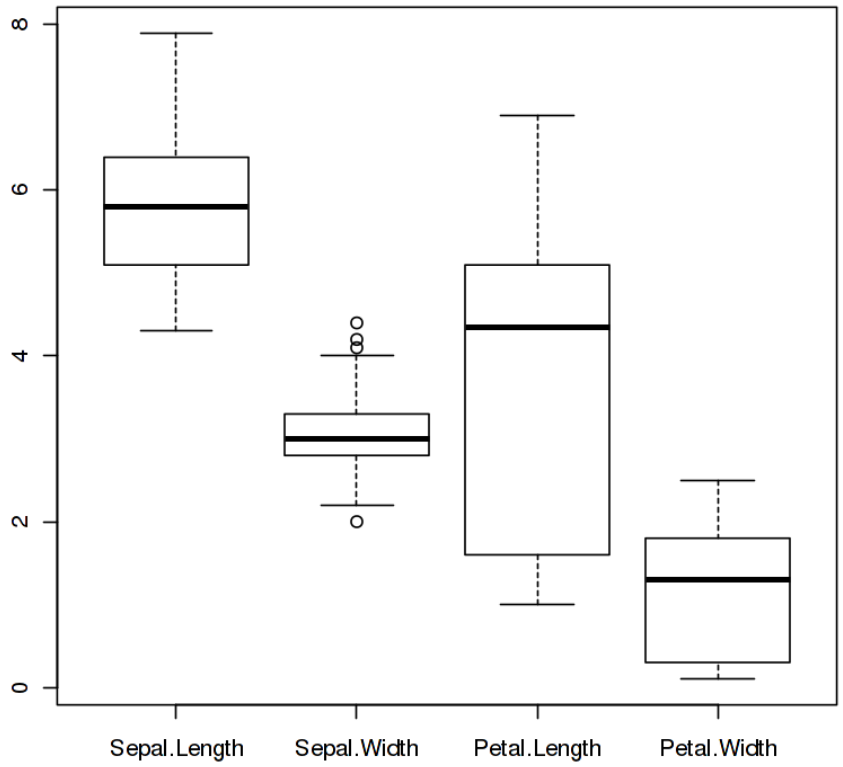
```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Wid
th
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.
100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.
300
 Median :5.800   Median :3.000   Median :4.350   Median :1.
300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   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
```

Separamos las medidas continuas de la etiqueta de especie para facilitar el análisis posterior.

In [3]:
```
medidas <- iris[1:4]
especies <- iris$Species
```

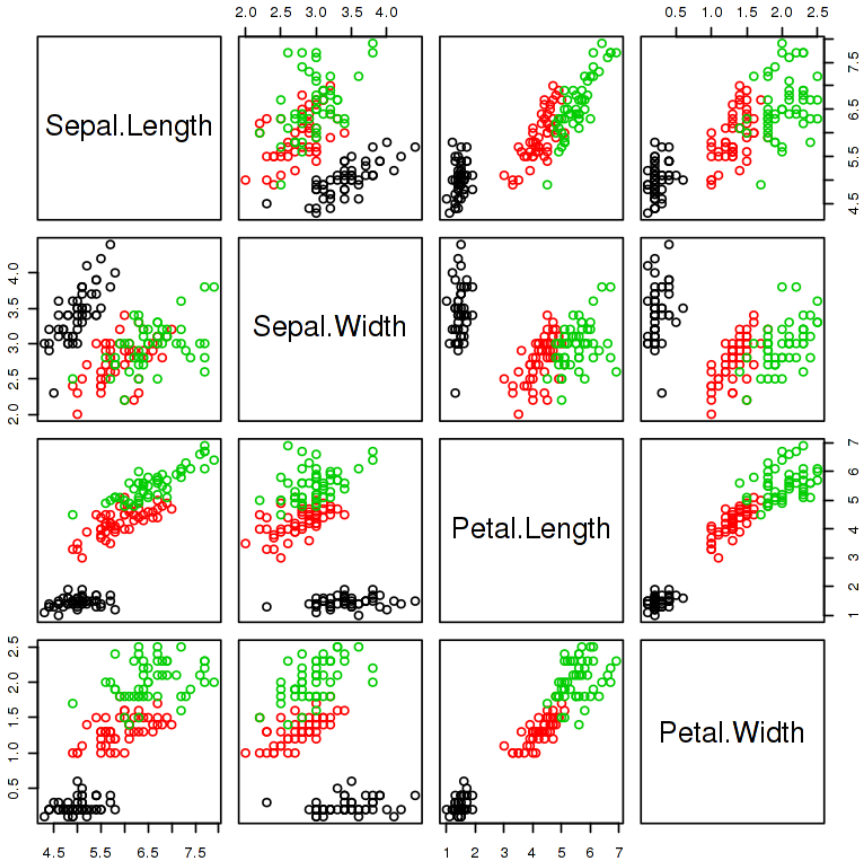
Una buena gráfica siempre resulta más fácil de visualizar que una tabla de números. En este caso, para observar la distribución de las variables, podemos usar un **diagrama de cajas y bigotes**. Este diagrama muestra los cuartiles y varianzas, con lo que nos da una muy buena idea de la distribución.

In [4]:
```
boxplot(medidas)
```



Para visualizar la relación de las variables, podemos también usar **diagramas de puntos**, poniendo en los ejes las variables a pares. En este caso, podemos pedirle a R también que nos coloree los puntos según la especie. Aquí empezamos a ver los **clusters**, agrupaciones de individuos que en este caso se corresponden claramente con las especies.

```
In [5]: plot(medidas, col=especies)
```

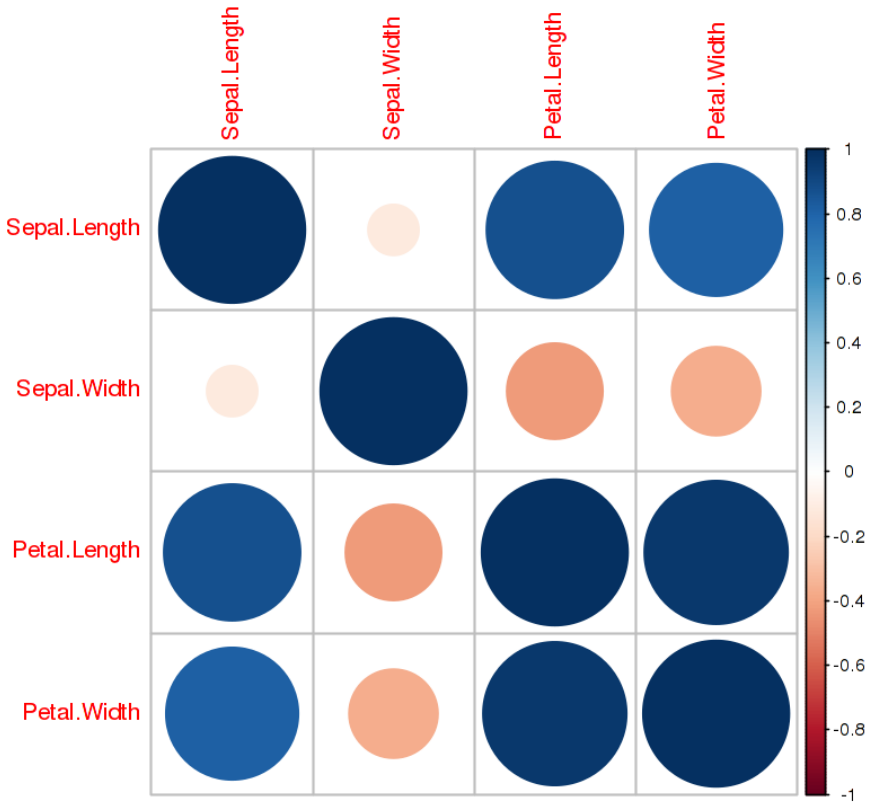


Otra cuestión que podemos ver en los diagramas anteriores, es que hay variables que presentan una fuerte relación lineal. Por ejemplo, observad la longitud y anchura de los pétalos.

Para analizar esto formalmente, podemos obtener la correlación de las medidas, e incluso visualizarla en una gráfica.

In [6]:
```
cor(medidas)
library(corrplot)
corrplot(cor(medidas))
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.0000000	-0.1175698	0.8717538	0.8179411
Sepal.Width	-0.1175698	1.0000000	-0.4284401	-0.3661259
Petal.Length	0.8717538	-0.4284401	1.0000000	0.9628654
Petal.Width	0.8179411	-0.3661259	0.9628654	1.0000000



Como hemos visto antes, hay variables que muestran una gran correlación. Por ejemplo, el ancho y largo de los pétalos. Intuitivamente, esto podríamos verlo como "el tamaño de los pétalos". Pensamos que los pétalos tienden a tener una proporción definida, y simplemente pueden ser más grandes o pequeños. Esta intuición se puede trasladar formalmente de varias maneras. En nuestro aálisis exploratorio, una herramienta muy habitual es el Análisis de Componentes Principales.

ACP

De la wikipedia:

> [El A]nálisis de componentes principales (en español ACP, en inglés, PCA) es una técnica utilizada para describir un set de datos en términos de nuevas variables ("componentes") no correlacionadas. Los componentes se ordenan por la cantidad de varianza original que describen, por lo que la técnica es útil para reducir la dimensionalidad de un conjunto de datos.

De manera resumida, el ACP busca maneras de "reorientar el espacio" en el que las nuevas dimensiones (componentes) sean independientes. Es decir, cambia las medidas originales por otras, combinación de las anteriores, que no dependan entre ellas. En nuestro caso, cambiaría longitud y anchura de los pétalos por tamaño y proporción, por ejemplo.

Pero, ¿qué ganamos con esto? La idea más importante del ACP es que podemos ordenar las nuevas componentes por "importancia" (calculada por ejemplo a partir de la desviación estándar o la varianza). Así, vemos que la componente "tamaño del pétalo" presenta una gran varianza, y "proporción" poca. Por tanto, podemos eliminar la componente "proporción", y quedarnos sólo con "tamaño". Así, reducimos el número de variables a tratar.

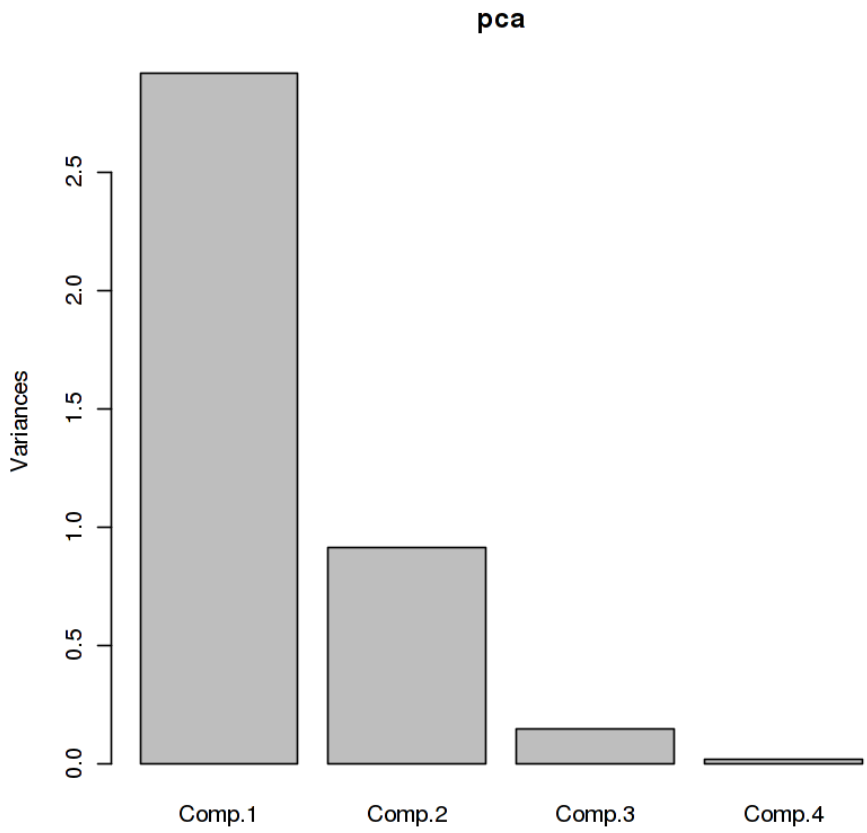
Sin embargo, hay que tener cuidado cuando damos nombres a las componentes. El ACP sólo nos da componente 1, 2, 3, etc. La interpretación que le demos nosotros luego es eso, una interpretación, y por tanto hay que ser cuidadoso con ella. Es más, en ocasiones tendremos componentes muy explicativas (es decir, que justifican gran parte de la varianza de los individuos) pero que no podremos interpretar claramente con un concepto intuitivo.

In [7]:
```
pca <- princomp(medidas, cor = T)
summary(pca)
```

```
Importance of components:
                          Comp.1    Comp.2     Comp.3
Comp.4
Standard deviation     1.7083611 0.9560494 0.38308860 0.143
926497
Proportion of Variance 0.7296245 0.2285076 0.03668922 0.005
178709
Cumulative Proportion  0.7296245 0.9581321 0.99482129 1.000
000000
```

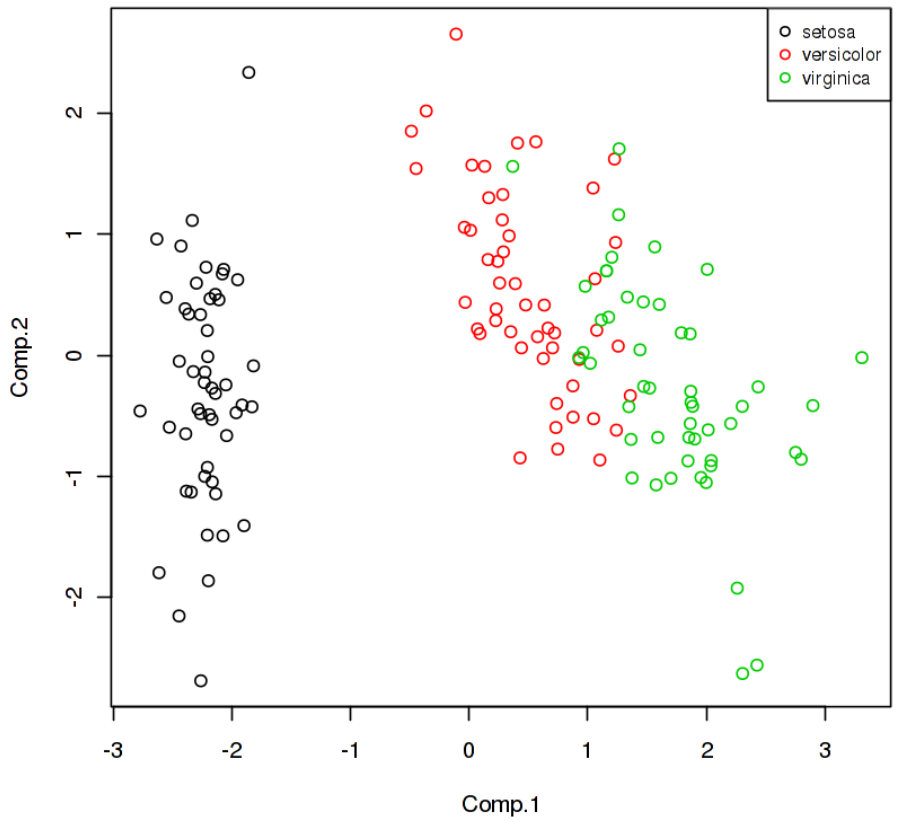
Podemos dibujar las varianzas, para visualizar la "importancia" de las componentes. Tras este paso, lo habitual es sólo quedarnos con las más explicativas (principales).

In [8]:
```
plot(pca)
```



Si nos quedamos sólo con las componentes 1 y 2, podemos dibujar una nube de puntos con los especímenes (coloreados por especie) y observar que con estas nuevas medidas las especies se dintinguen bien.

In [9]:
```
plot(pca$scores, col=especies)
legend('topright', legend = levels(especies), col = 1:3, ce
x = 0.8, pch = 1)
```

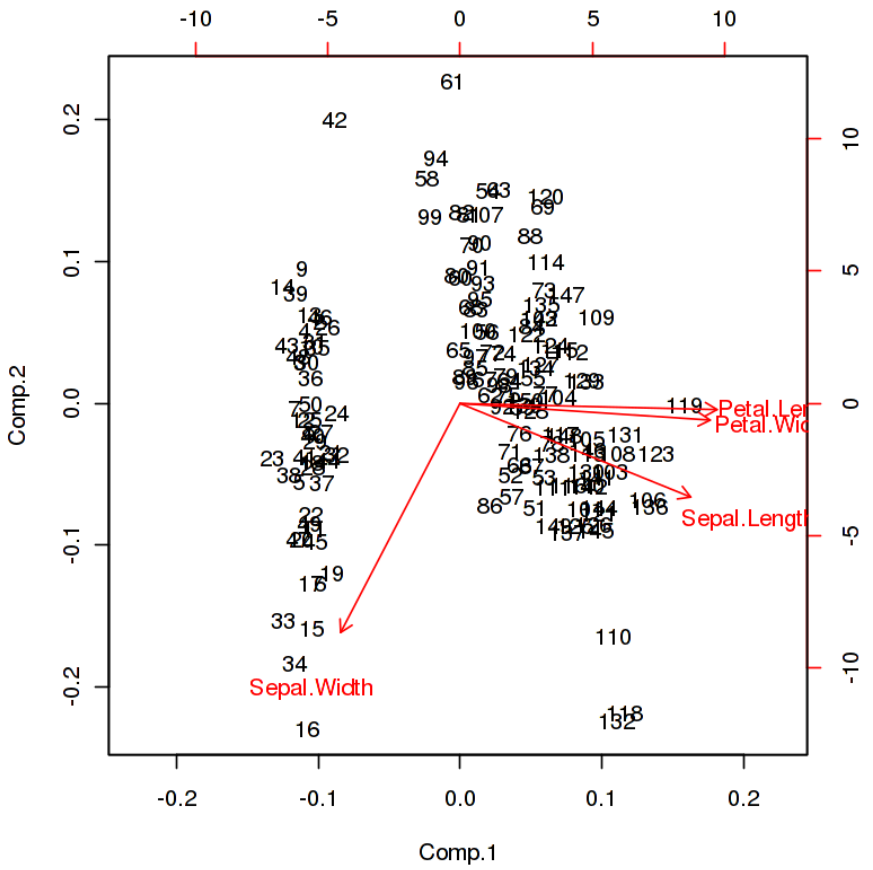


También podemos dibujar, en el espacio de las nuevas componentes, cómo varían las medidas originales.

En este caso, podemos ver justificada nuestra intuición de que el tamaño del pétalo es una componente importante, y que el ancho y la longitud dependen de ella. Esto se ve en que estas medidas son casi horizontales (es decir, paralelas a la componente 1), y coinciden en gran medida.

Sin embargo, ¿cuál sería la interpretación para la componente 2? Esto no es tan inmediato, y es posible que no haya una interpretación natural con mejor nombre que "segunda componente principal".

In [10]:
```
biplot(pca)
```



Preguntas

- ¿Cómo describirías tú la componente 2?
- ¿Qué quiere decir que los especímenes de "*setosa*" formen una línea vertical?
- A la vista del diagrama de componentes, y de las nubes de puntos en las medidas dos a dos (segundo diagrama), ¿qué herramientas estadísticas usarías para decidir, dada una nueva flor de la que no conocemos la especie, a qué especie pertenece?