PCA en R

Trabajaremos con el dataset iris, el mismo que hemos utilizado para hacer el PCA en Weka. Estos son los datos generales del dataset.

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

En R, la función que ejecuta el análisis de componentes principales (PCA) de nuestro problema es prcomp. Puede tomar varios parámetros pero los únicos que nos interesan en este caso son por un lado, los propios datos del dataset (en este caso cogemos las cuatro primeras columnas, puesto que la ultima es la clase) y el scale, que indica que queremos normalizar los datos antes de ejecutar el PCA.

```
pca_iris <- prcomp(iris[1:4],scale=TRUE)
summary(pca_iris)
```

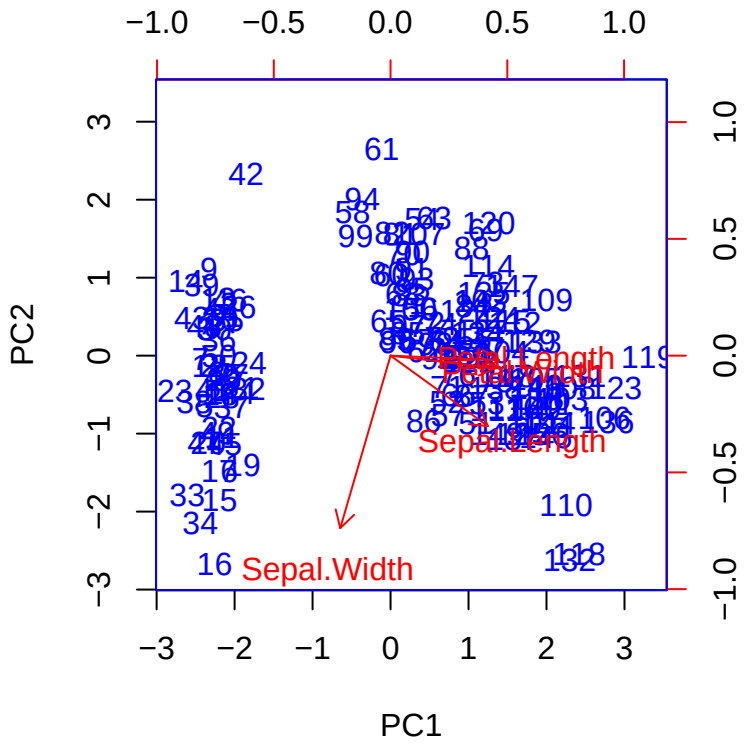
```
## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.7084 0.9560 0.38309 0.14393
## Proportion of Variance 0.7296 0.2285 0.03669 0.00518
## Cumulative Proportion  0.7296 0.9581 0.99482 1.00000
```

¿Qué porcentaje de la varianza recoge la primera componente? ¿Y la segunda?

- La primera componente (PCA1) recoge el 72.96% de la varianza, y la segunda (PCA2) recoge un 22.85%.

Si entre las dos se recoge el menos un 90-95% de la varianza, la visualización en 2-D de todos los puntos sobre ambas componentes, está "permitida".

```
biplot(pca_iris,scale=0,col=c("blue","red"))
```



¿Qué representan los números en la visualización?

- Cada número es una instancia del dataset, que están numeradas del 1 al 150.

¿Qué representan los dos ejes de la visualización?

- Los ejes son los valores que toma cada variable para la PCA1 y la PCA2.

¿Qué representan las direcciones rojas en la visualización?

- Las direcciones de las flechas representan los ejes originales de las variables del dataset. De dichas flechas y direcciones se puede deducir cual es el signo y la proporción de las variables originales que toman los nuevos componentes principales.

Gráficos más visuales

Varias visualizaciones "elegantes". Necesario utilizar dos paquetes: instalación y carga

```
install.packages("ggfortify")
```

```
library(ggfortify)
```

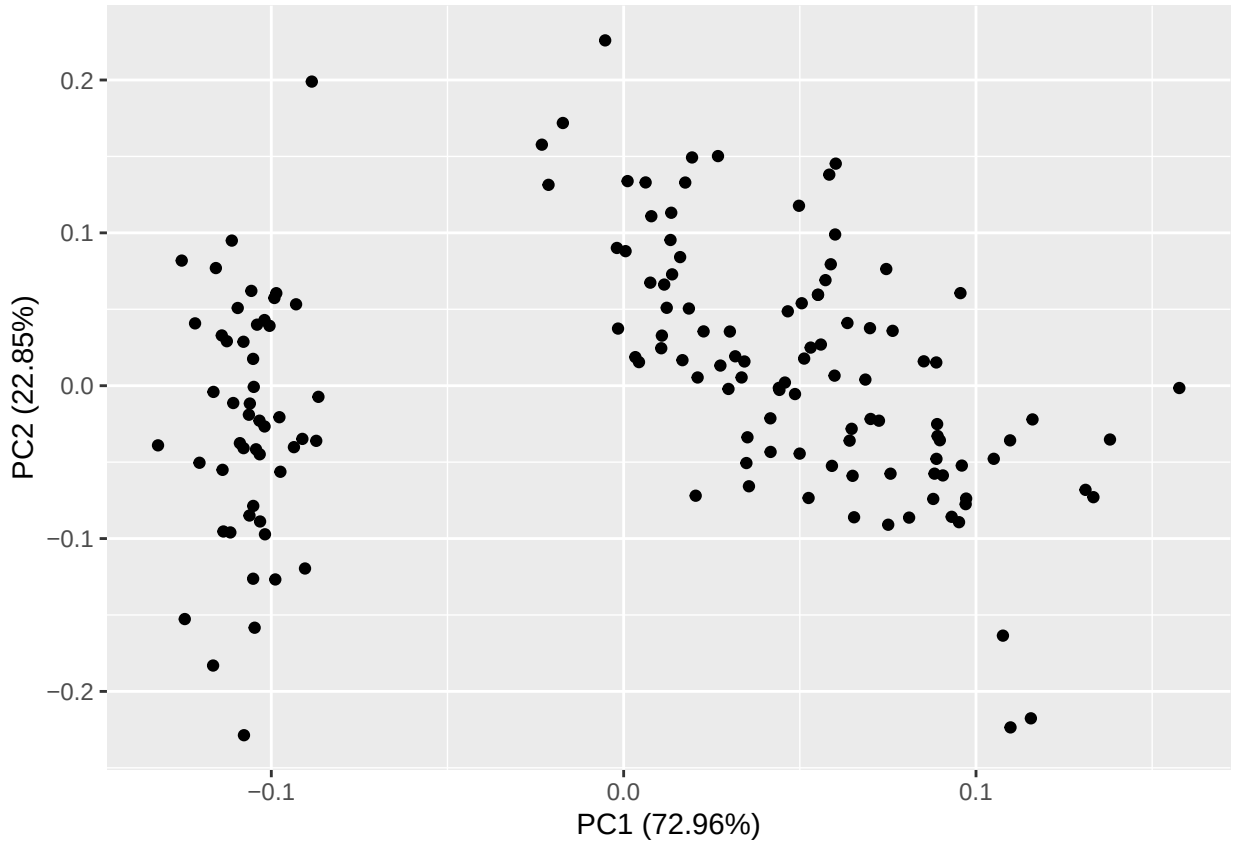
```
## Loading required package: ggplot2
```

```
install.packages("ggplot2")
```

```
library(ggplot2)
```

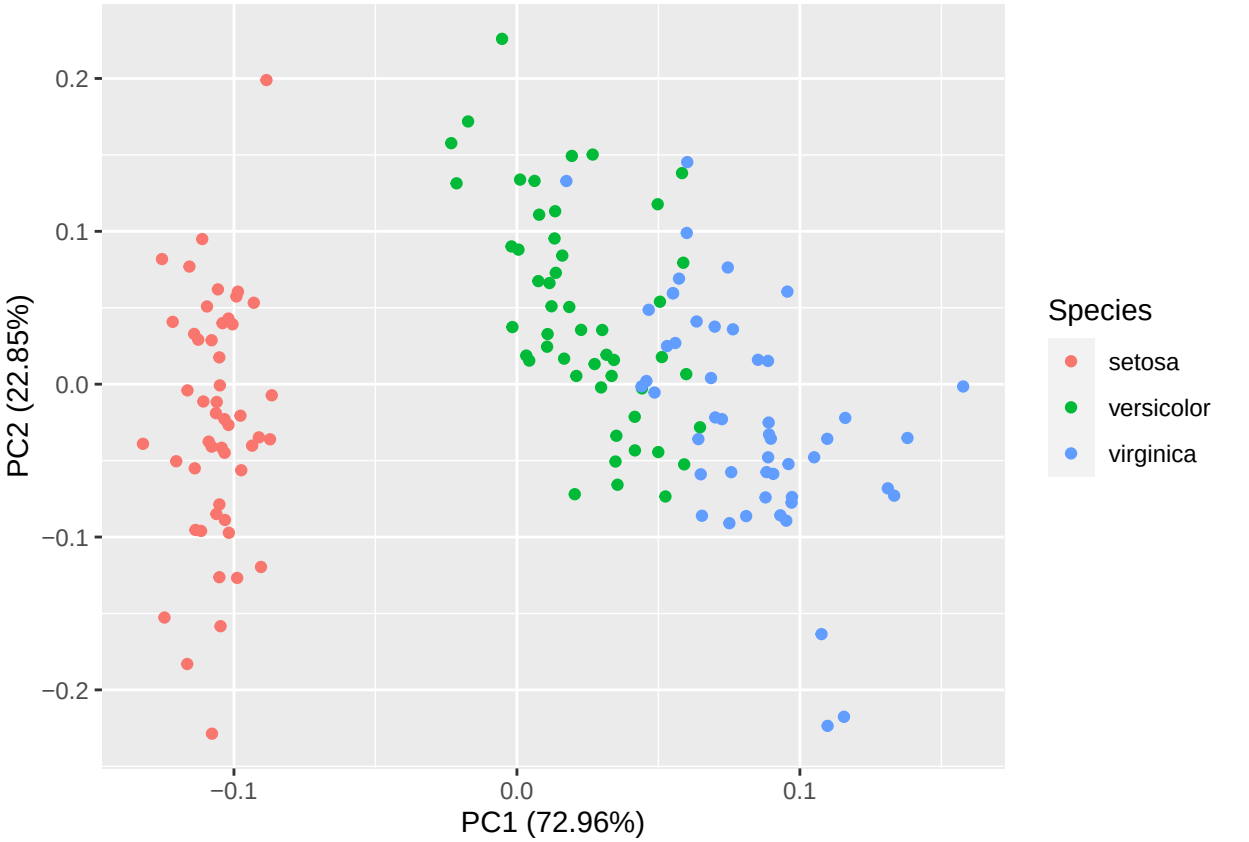
Indica las diferencias entre las siguientes 3 visualizaciones: qué reflejan, y en qué se diferencian

```
autoplot(pca_iris)
```



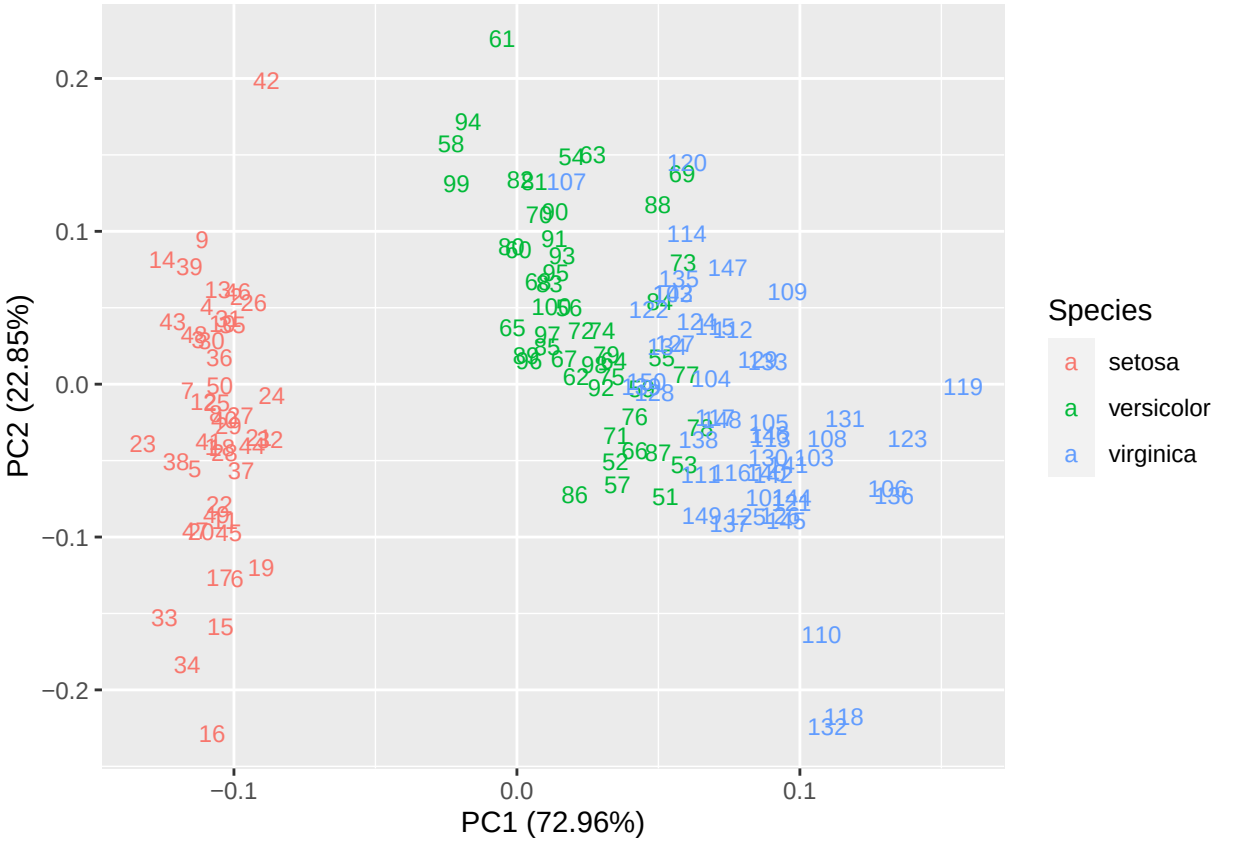
En este primer gráfico se nos muestrán las instancias del dataset en un gráfico que representa cual es el valor que toman para las nuevas componentes principales. Ademas también muestra cual es la varianza que recoge cada componente.

```
autoplot(pca_iris, data = iris, colour = 'Species')
```



En este segunda visualización se refleja además del valor de las instancias para las componentes principales, que ya se nos mostraba en en anterior, la clase a la pertenecen en el dataset iris, en función de su color.

```
autoplot(pca_iris, data = iris, colour = 'Species', shape = FALSE, label.size = 3)
```



Por último este dataset refleja lo mismo que el anterior con la diferencia de que en lugar de representar cada individuo con un punto en el gráfico se le representa con el número que le correspone en el dataset.