Data representation with R

BioMath licence L2 - Université Paris Saclay

Elodie Marchadier

R software is required to load and visualize the data. It can be downloaded for free on https://cran.r-project.org/

1. How to load data

Most of the time, you will need to load your data from a .csv or .txt file using the `read.table()` function. The symbol used to delimite the columns must be specified using `sep=`, the decimal symbol must be specified using `dec=` and if there is an header it has to be mentionned using `header=TRUE`.

```
#to go to the right folder
#setwd(path/to/the/file)

# to read the table
#tab=read.table("nomduficier.csv",sep=";",header =TRUE)
```

In this lesson, the data set `iris` is included in R and just need to be loaded

```
# this line allows to remove all the variables in the environment
rm(list=ls())

data("iris")

# to visualize the head of the dataset :
head(iris)
```

```
##   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
```

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

2. The variables

Each of the variables of the dataset can be called using the `$` followed by the name of the column or using the column number

```
iris$Sepal.Length
```

```
##   [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9 5.4 4.8 4.8 4.3 5.8 5.7 5.4 5.1
##  [19] 5.7 5.1 5.4 5.1 4.6 5.1 4.8 5.0 5.0 5.2 5.2 4.7 4.8 5.4 5.2 5.5 4.9 5.0
##  [37] 5.5 4.9 4.4 5.1 5.0 4.5 4.4 5.0 5.1 4.8 5.1 4.6 5.3 5.0 7.0 6.4 6.9 5.5
##  [55] 6.5 5.7 6.3 4.9 6.6 5.2 5.0 5.9 6.0 6.1 5.6 6.7 5.6 5.8 6.2 5.6 5.9 6.1
##  [73] 6.3 6.1 6.4 6.6 6.8 6.7 6.0 5.7 5.5 5.5 5.8 6.0 5.4 6.0 6.7 6.3 5.6 5.5
##  [91] 5.5 6.1 5.8 5.0 5.6 5.7 5.7 6.2 5.1 5.7 6.3 5.8 7.1 6.3 6.5 7.6 4.9 7.3
## [109] 6.7 7.2 6.5 6.4 6.8 5.7 5.8 6.4 6.5 7.7 7.7 6.0 6.9 5.6 7.7 6.3 6.7 7.2
## [127] 6.2 6.1 6.4 7.2 7.4 7.9 6.4 6.3 6.1 7.7 6.3 6.4 6.0 6.9 6.7 6.9 5.8 6.8
## [145] 6.7 6.7 6.3 6.5 6.2 5.9
```

```
# is equivalent to
iris[,1]
```

```
##   [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9 5.4 4.8 4.8 4.3 5.8 5.7 5.4 5.1
##  [19] 5.7 5.1 5.4 5.1 4.6 5.1 4.8 5.0 5.0 5.2 5.2 4.7 4.8 5.4 5.2 5.5 4.9 5.0
##  [37] 5.5 4.9 4.4 5.1 5.0 4.5 4.4 5.0 5.1 4.8 5.1 4.6 5.3 5.0 7.0 6.4 6.9 5.5
##  [55] 6.5 5.7 6.3 4.9 6.6 5.2 5.0 5.9 6.0 6.1 5.6 6.7 5.6 5.8 6.2 5.6 5.9 6.1
##  [73] 6.3 6.1 6.4 6.6 6.8 6.7 6.0 5.7 5.5 5.5 5.8 6.0 5.4 6.0 6.7 6.3 5.6 5.5
##  [91] 5.5 6.1 5.8 5.0 5.6 5.7 5.7 6.2 5.1 5.7 6.3 5.8 7.1 6.3 6.5 7.6 4.9 7.3
## [109] 6.7 7.2 6.5 6.4 6.8 5.7 5.8 6.4 6.5 7.7 7.7 6.0 6.9 5.6 7.7 6.3 6.7 7.2
## [127] 6.2 6.1 6.4 7.2 7.4 7.9 6.4 6.3 6.1 7.7 6.3 6.4 6.0 6.9 6.7 6.9 5.8 6.8
## [145] 6.7 6.7 6.3 6.5 6.2 5.9
```

```
# the first number indicated in brackets is the row number
iris[1,]
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
```

Calculations such as mean, variance,… can be done on quantitative variables using dedicated functions.

```
# mean and variance
mean(iris$Petal.Length)
```

```
## [1] 3.758
```

```
var(iris$Petal.Length)
```

```
## [1] 3.116278
```

```
# correlation between two quantitative variables
cor(iris$Petal.Length,iris$Sepal.Length)
```

```
## [1] 0.8717538
```

For quantitative variables, the number of observations of the modalities can be computed.

```r
# number of individuals of each species
table(iris$Species)
```

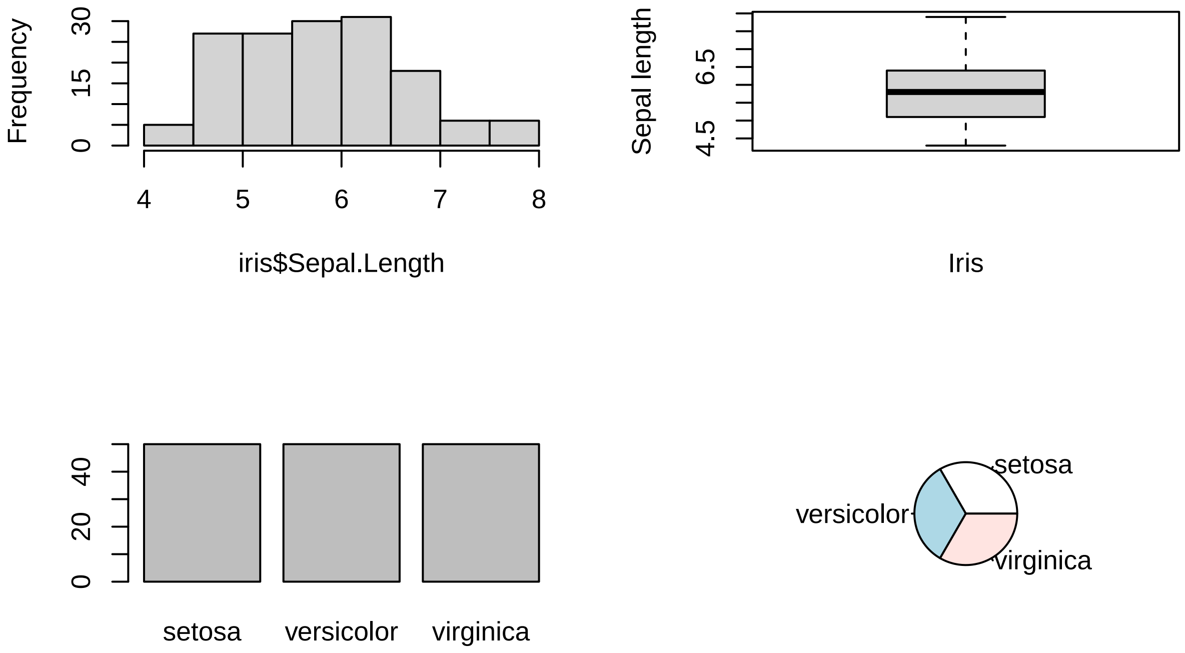
```
##
##     setosa versicolor  virginica
##         50         50         50
```

3. Graphical representation of variables

3.a Graphical representation of a single variable

There are different ways to plot a given variable. A quantitative variable can be plotted as an histogram or a boxplot. The distribution of a quantitative variable can be observed as a barplot.

```r
par(mfrow=c(2,2))
# quantitative variables
hist(iris$Sepal.Length,main="")
boxplot(iris$Sepal.Length,ylab="Sepal length",xlab="Iris")

# qualitatie variables
barplot(table(iris$Species))
pie(table(iris$Species))
```

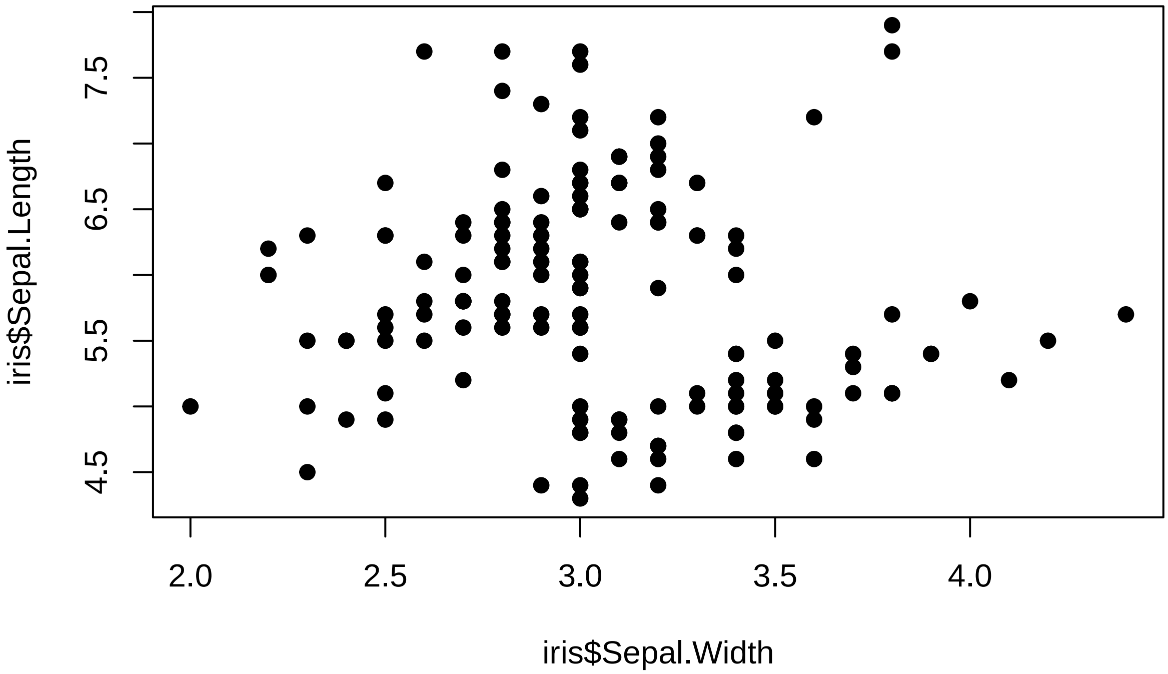


3.b Graphical representation of several variables

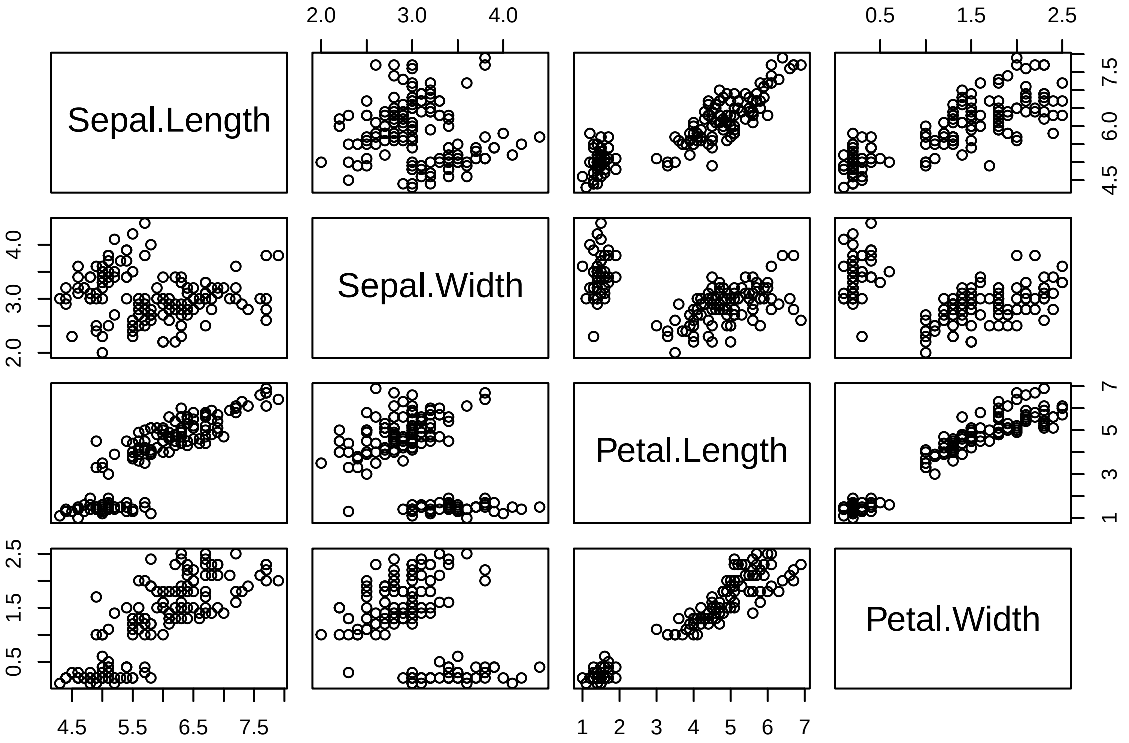
When several variable are evaluated on the same individuals, the relations between these varaibles can be observed. Depending on the nature of the variables, several representations can be done.

A dot plot can be used to observe the relation between two quantitative variables.

```r
# dotplot ad scatterplot
plot(iris$Sepal.Length~iris$Sepal.Width,pch=19)
```

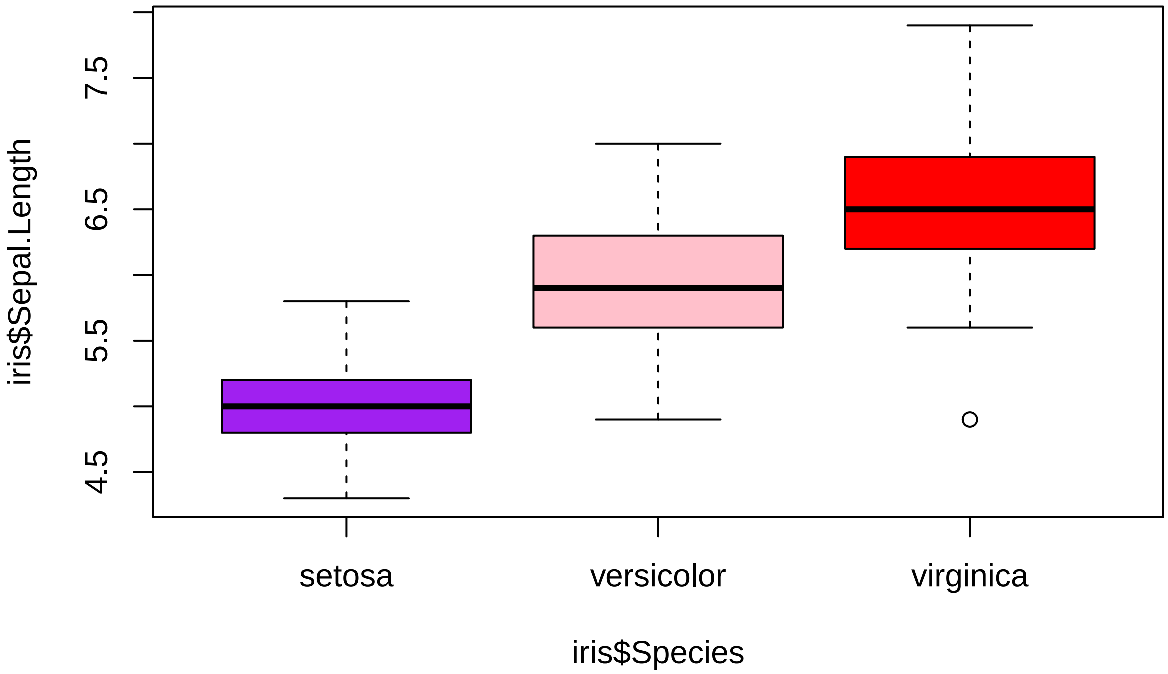


```
plot(iris[,c("Sepal.Length","Sepal.Width","Petal.Length","Petal.Width")])
```



Boxplots are usually used to plot a quantitative variable as function of a qualitative one. Violon plots can be used to better visualize the dsitribution of the quantitatvie variables.

```
plot(iris$Sepal.Length~iris$Species,col=c("purple","pink","red"))
```



```
install.packages("vioplot") # a package is a set of additionnal function that need to be independantly
```

```
## Installation du package dans '/home/elodie/R/x86_64-pc-linux-gnu-library/4.3'
## (car 'lib' n'est pas spécifié)
```

```
library(vioplot) # and then loaded
```

```
## Le chargement a nécessité le package : sm
```

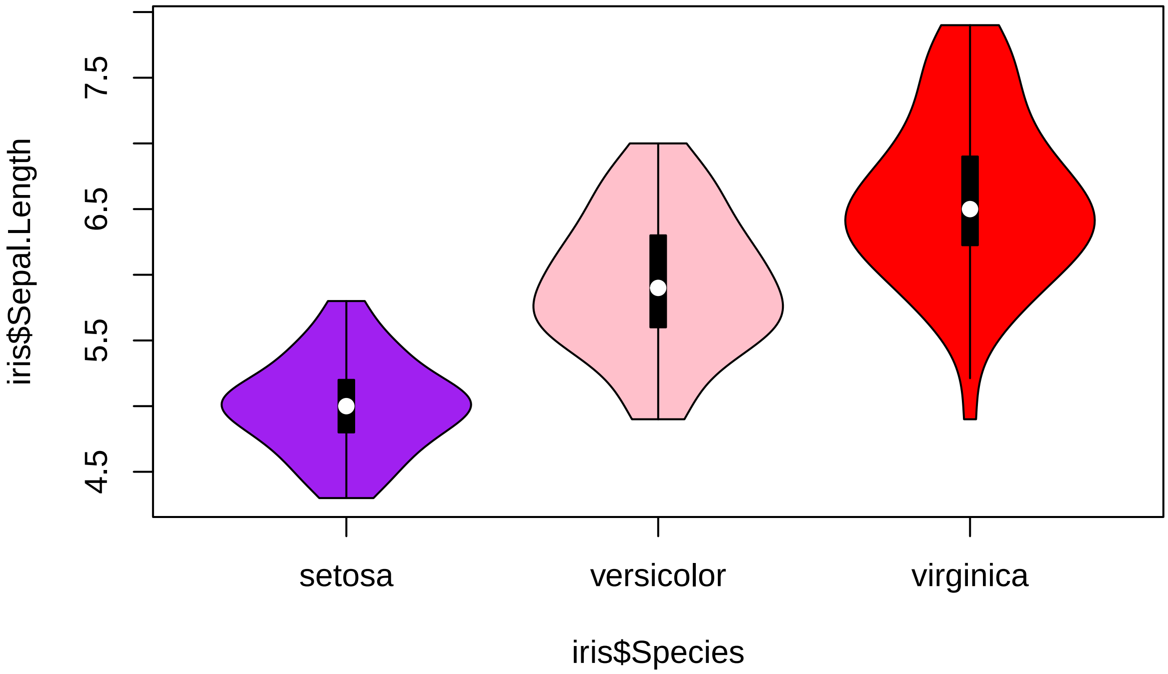
```
## Package 'sm', version 2.2-5.7: type help(sm) for summary information
```

```
## Le chargement a nécessité le package : zoo
```

```
##
## Attachement du package : 'zoo'
```

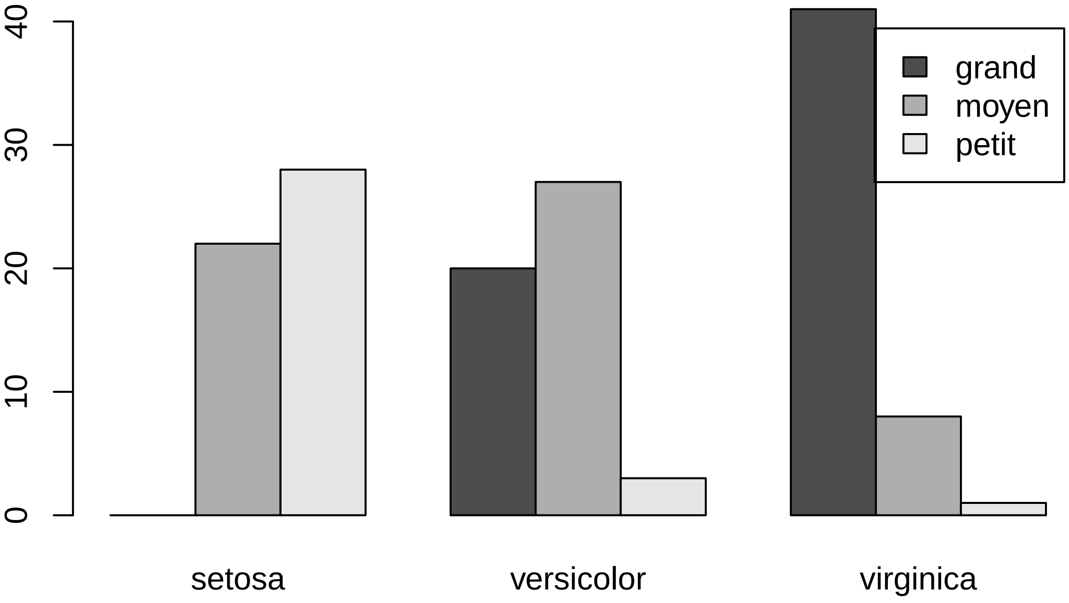
```
## Les objets suivants sont masqués depuis 'package:base':
##
##     as.Date, as.Date.numeric
```

```
vioplot(iris$Sepal.Length~iris$Species,col=c("purple","pink","red"))
```



This data set only include one qualitative variable. A second one will be created to illustrate a graphical visualization of two qualitative variables.

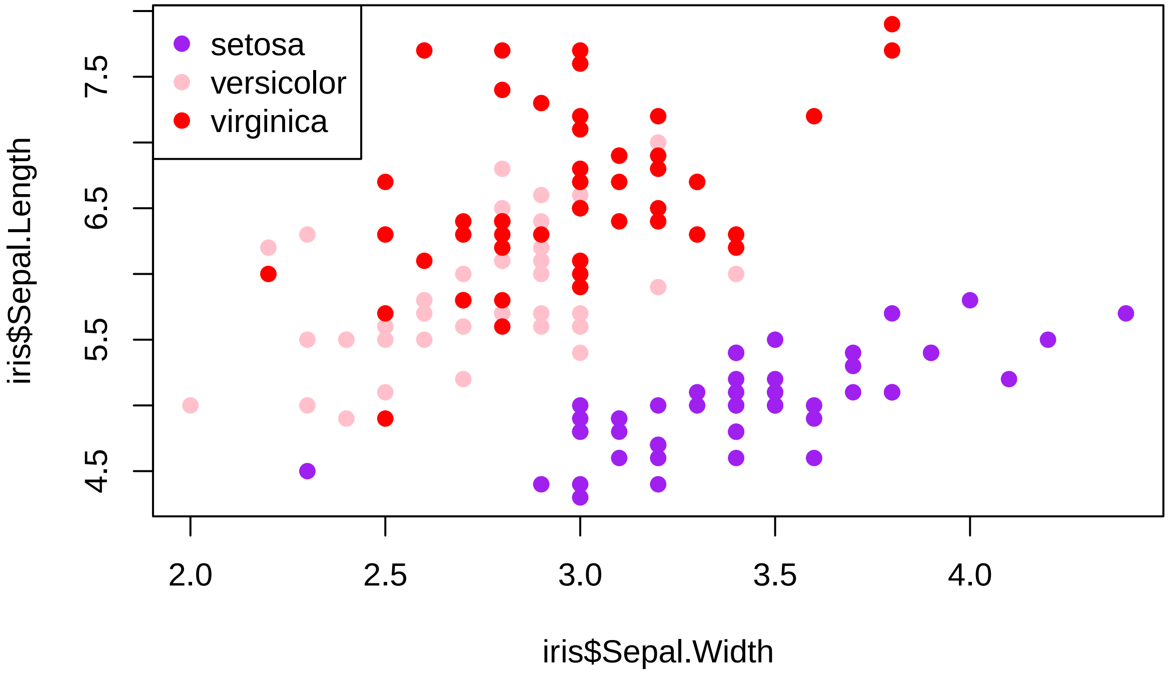
```
# a new variable is created for 3 levels of sepal length
iris2<-iris
iris2$Sepal.Length_cat[iris2$Sepal.Length<=5]<-"petit"
iris2$Sepal.Length_cat[iris2$Sepal.Length<=6 & iris2$Sepal.Length>5]<-"moyen"
iris2$Sepal.Length_cat[iris2$Sepal.Length>6]<-"grand"

# the distribution of the created variable can be observed
conting=table(iris2[,c("Sepal.Length_cat","Species")])
barplot(conting,beside=T,legend=T)
```



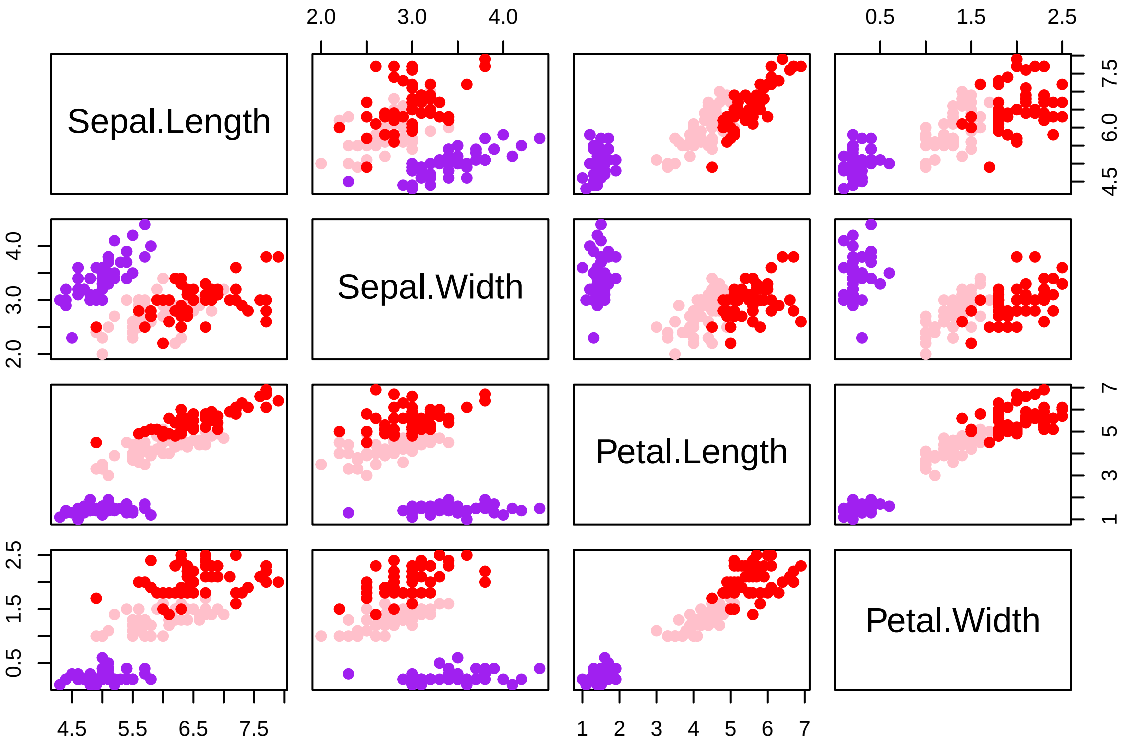
Using colors, more than 2 variables can be observed.

For instance, the species can be indicated with a color on a dotplot to plot two quantitative variables and a qualitative one.

```
plot(iris$Sepal.Length~iris$Sepal.Width,pch=19,col=c("purple","pink","red")[as.factor(iris$Species)])
legend("topleft",levels(iris$Species),pch=19,col=c("purple","pink","red"))
```

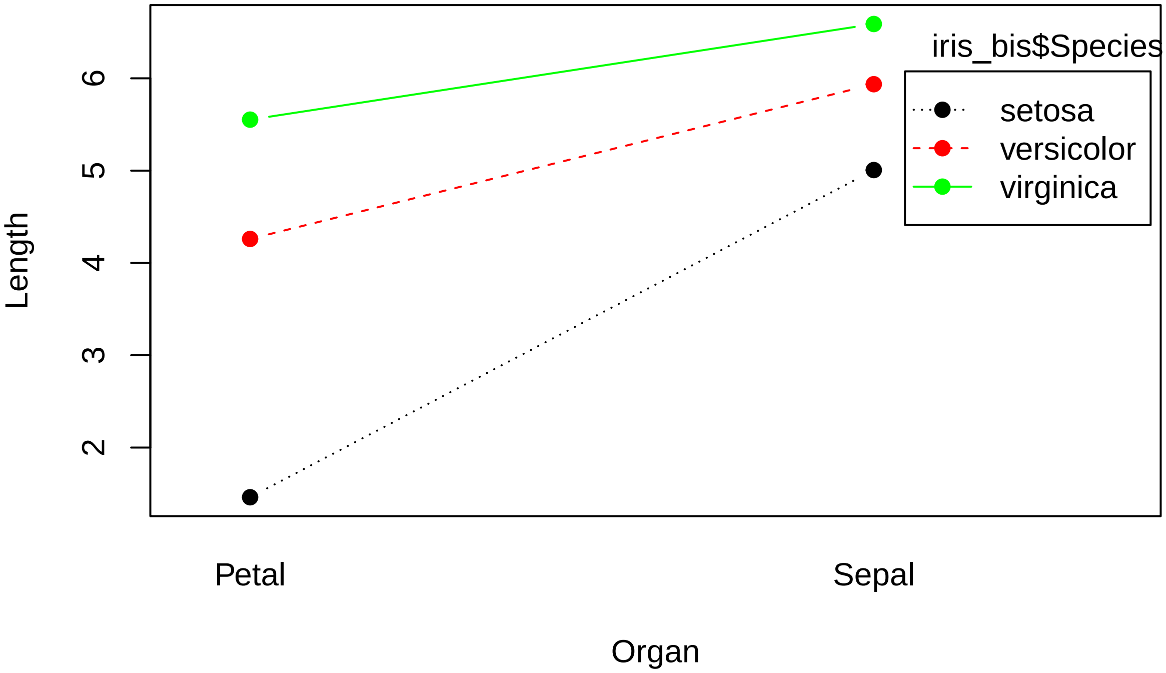


```
plot(iris[,c("Sepal.Length","Sepal.Width","Petal.Length","Petal.Width")],
     pch=19,col=c("purple","pink","red")[as.factor(iris$Species)])
```



Interactions plots allow to visualize two qualitative variables and a quantitative one. This representation requires to change the format of the data.

```
# a single column indicate the length, the species and the organ are indicated in two other columns
iris3=iris[,c("Sepal.Length","Species")]
colnames(iris3)[1]="length"
iris3$organ="Sepal"

iris4=iris[,c("Petal.Length","Species")]
colnames(iris4)[1]="length"
iris4$organ="Petal"
iris_bis=rbind(iris3,iris4)

interaction.plot(x.factor=iris_bis$organ,trace.factor=iris_bis$Species,response=iris_bis$length,
                 fun=mean,type="b", col=c("black","red","green"), pch=19, fixed=TRUE, leg.bty = "o",
                 xlab="Organ", ylab="Length")
```



3.c Principal component analysis

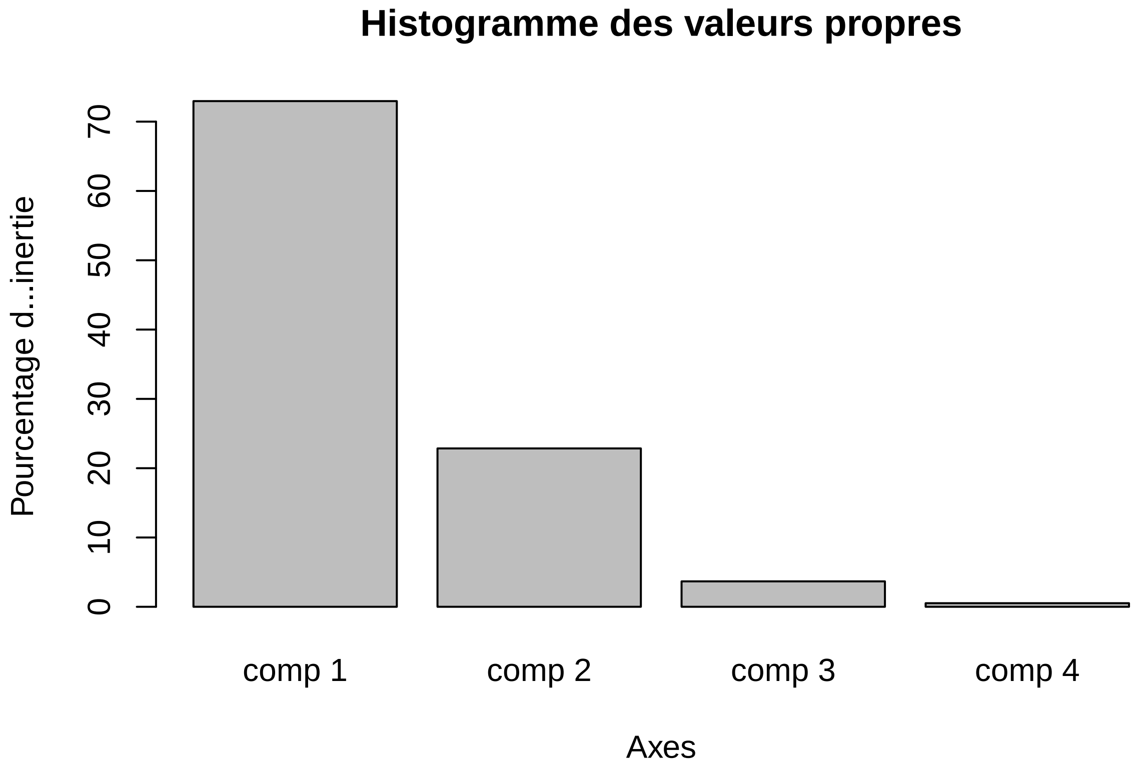
```
library(FactoMineR)

head(iris)
```

```
##   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
```

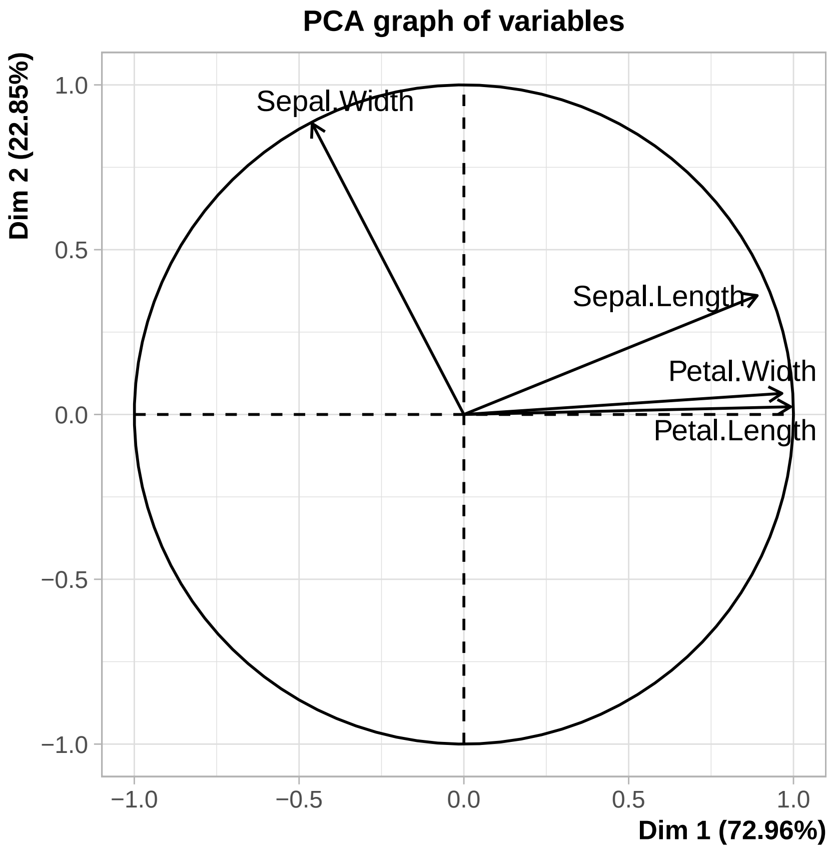
```
myPCA=PCA(iris, scale.unit=TRUE, graph=F,quali.sup=5)
```

```
barplot(myPCA$eig[,2], main="Histogramme des valeurs propres", names.arg=rownames(myPCA$eig), xlab="Axes
```

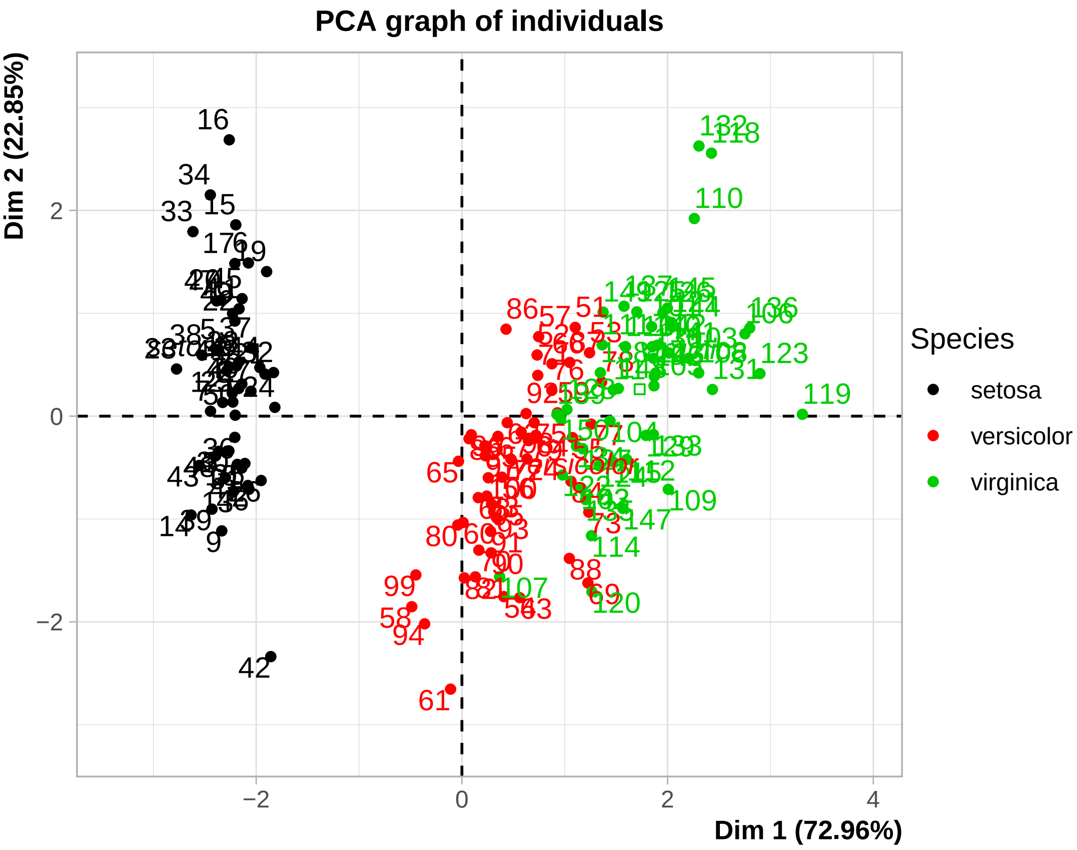
```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...): erreur
## de conversion de 'Pourcentage d’inertie' dans 'mbcsToSbcs' : le point est
## substitué pour <e2>

## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...): erreur
## de conversion de 'Pourcentage d’inertie' dans 'mbcsToSbcs' : le point est
## substitué pour <80>

## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...): erreur
## de conversion de 'Pourcentage d’inertie' dans 'mbcsToSbcs' : le point est
## substitué pour <99>
```



```
plot.PCA(myPCA, axes=c(1, 2), choix="var")
```



```r
plot.PCA(myPCA, axes=c(1, 2), choix="ind",habillage=5)
```


PCA graph of individuals
Dim 2 (22.85%)
Dim 1 (72.96%)
−2
0
2
4
2
0
−2
Species
● setosa
● versicolor
● virginica
16
34
33
15
42
119
110
132
118
61
65
99
58
94
86
51
136
106
123
109
114
120
147
107