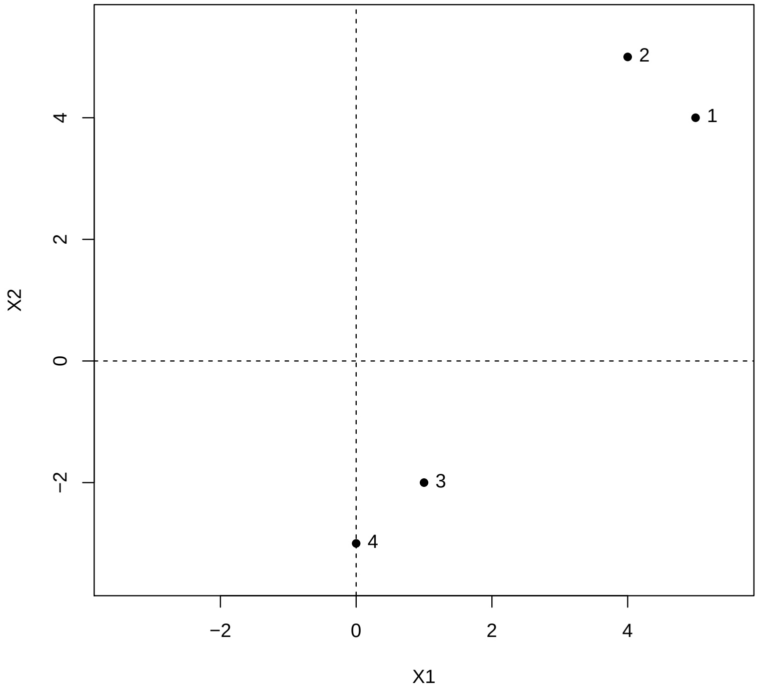
TP3: Clustering

Exercice 1. The k-means algorithm

Let X be a data matrix where a set $\Omega = \{1, 2, 3, 4\}$ of $n = 4$ observations described by $p = 2$ variables. The observations are weighted by $w_i = 1$.

1. Apply *by hand* the k-means algorithm to Ω with $K = 2$ clusters and with the two first rows of X chosen as initial centers. Perform the within-cluster sum of squares of the final partition.

```
##   X1 X2
## 1  5  4
## 2  4  5
## 3  1 -2
## 4  0 -3
```

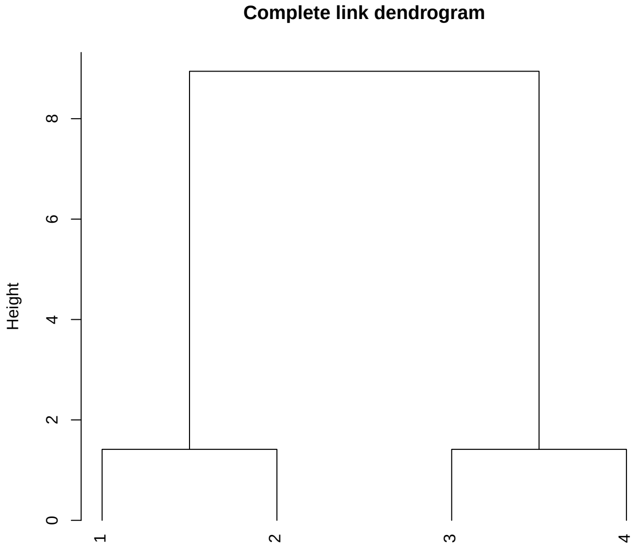


2. Use now the R function **kmeans()** to repeat the previous question. Check that you find the same results.
3. Perform the total sum of squares T of the data. Check that $T = B + W$ where B is the between-clusters sum of squares and W is the within-clusters sum of squares of the final partition.
4. Perform the proportion of variance explained by the final partition.

Exercice 2. The complete link ascendant hierarchical clustering algorithm.

1. Apply now *by hand* the complete link hierarchical clustering algorithm to $\Omega = \{1, 2, 3, 4\}$ using the Euclidean distance. Give the hierarchy H and represent the dendrogram. What partition in two clusters is obtained by cutting this dendrogram ?

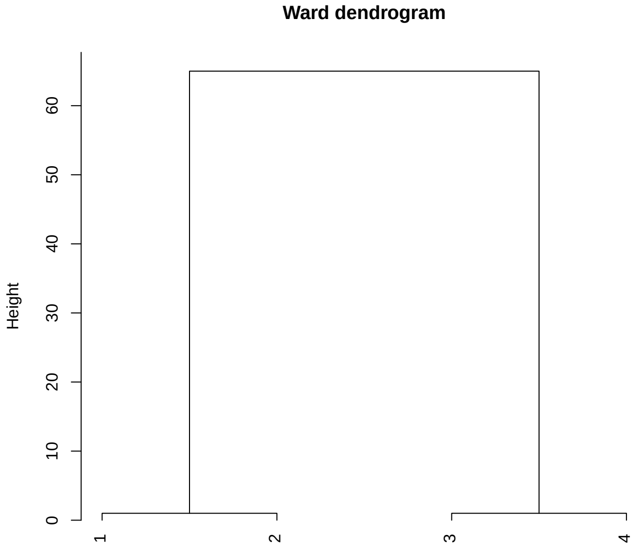
2. Use now the R functions **hclust()**, **plot()** and **cutree()** to repeat the previous question. Check that you find the same results and then the following dendrogram.



3. Build now the complete link dendrogram obtained with the Manhattan distance instead of the Euclidean distance.

Exercice 3. The Ward minimum variance hierarchical clustering algorithm.

1. Apply now *by hand* the Ward minimum variance method to $\Omega = \{1, 2, 3, 4\}$ where the observations are still weighted by $w_i = 1$. Plot the dendrogram obtained in that way.
2. Use now the R function **hclust()** with the recommandations given in Appendix (at the end of the TP) to find the results of question 1. and then the following dendrogram.

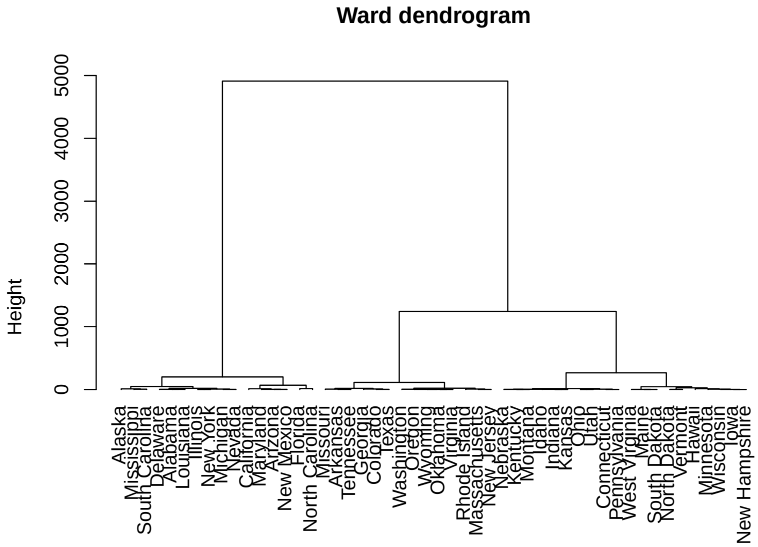


Exercice 4. Zoom the upper part of the Ward dendrogram.

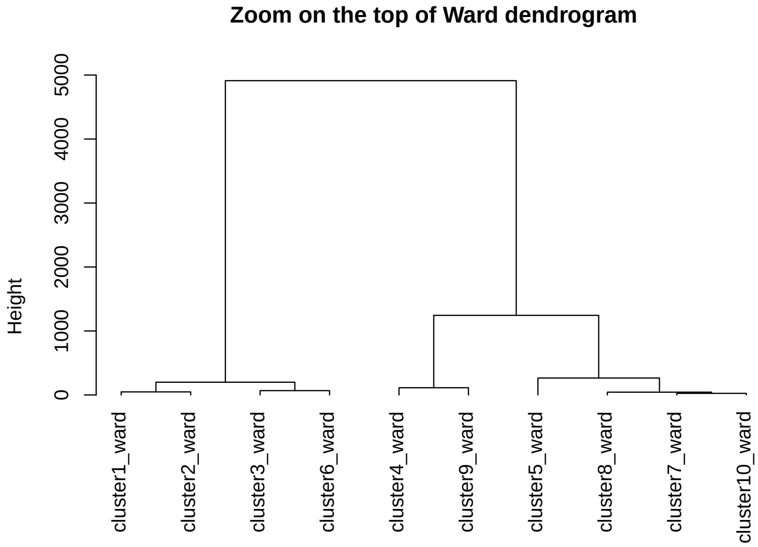
In this exercice, the $n = 50$ american states described in the **USArrests** dataset are weighted by $\frac{1}{n}$

```
#Violent crime rates by US state
help(USArrests)
```

1. Build with **hclust()** the Ward dendrogram of the $n = 50$ american states.



2. Cut the tree into ten clusters. Perform a new data matrix with 10 rows (the 10 centers of the clusters) and the vector $(\mu_1, ..., \mu_{10})$ of the weights of the 10 centers (the weights of the 10 clusters).
3. Reconstruct the upper part of the Ward dendrogram of question 1 using the cluster centers, their weights and the recommandations in Apprendix (at the end of the TP).

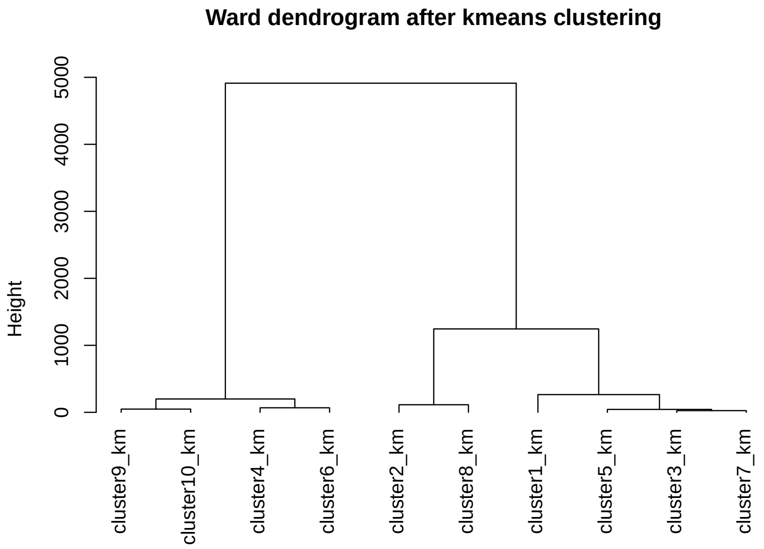


Exercice 5. Combine k-means and Ward.

In this exercice, the $n = 50$ states of the **USArrests** dataset are weighted by $\frac{1}{n}$.

First part : Ward after k-means.

1. Find with the k-means method a partition in $K = 10$ clusters (choose **nstart=200** in the **kmeans** function).
2. Build the Ward dendrogram starting from the $K = 10$ clusters obtained with the k-means method.



3. When do you think this methodology is usefull ?

Second part : k-means after Ward.

4. Build with **hclust()** the Ward dendrogram of the $n = 50$ american states.
5. Cut the tree in 2 clusters and perform the proportion of variance explained by this 2-clusters partition.

```
prop_inert_cutree <- function(tree, K)
{
  #tree= Ward minimum variance tree
  n <- length(tree$order)
  P <- cutree(tree, k=K)
  W <- sum(tree$height[1:(n-K)])
  Tot <- sum(tree$height)
  return(1-W/Tot)
}
```

6. Find a partition in 2 clusters with the k-means method starting from the Ward's 2-clusters partition.
7. Perform the proportion of variance explained by this partition. Compare with the result of question 5. Why is this result expected ?

Exercice 6. Combine clustering and PCA.

In this exercice, the $n = 25$ european countries of the **protein** dataset are weighted by $\frac{1}{n}$.

```
library(PCAmixdata)
data(protein)
```

Let X be a numerical data matrix of dimension $n \times p$. The Ward and k-means clustering methods give same results when applied

- to the data matrix X (resp. standardized data matrix Z) of dimension $n \times p$,
- to the matrix of all the principal components F of the non normalized PCA (resp. normalized PCA) of dimension $n \times r$ where r is the rank of X.

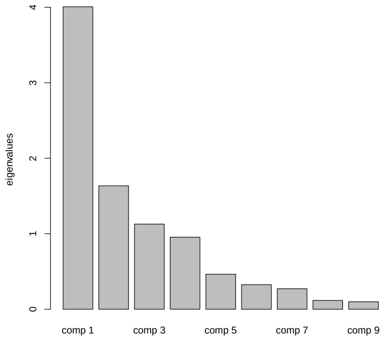
First part: Clustering on all the principal components.

1. Buil the Ward dendrogram of the $n = 25$ european countries on the **standardized data**. Check that the sum of the heights is equal to the total inertia.
2. Build the Ward dendrogram of the $n = 25$ european countries on **all the principal components** of **normalized PCA**. Check that the sum of the heights is equal to the total inertia.
3. Compare the heights of the dendrograms of questions 1 and 2.

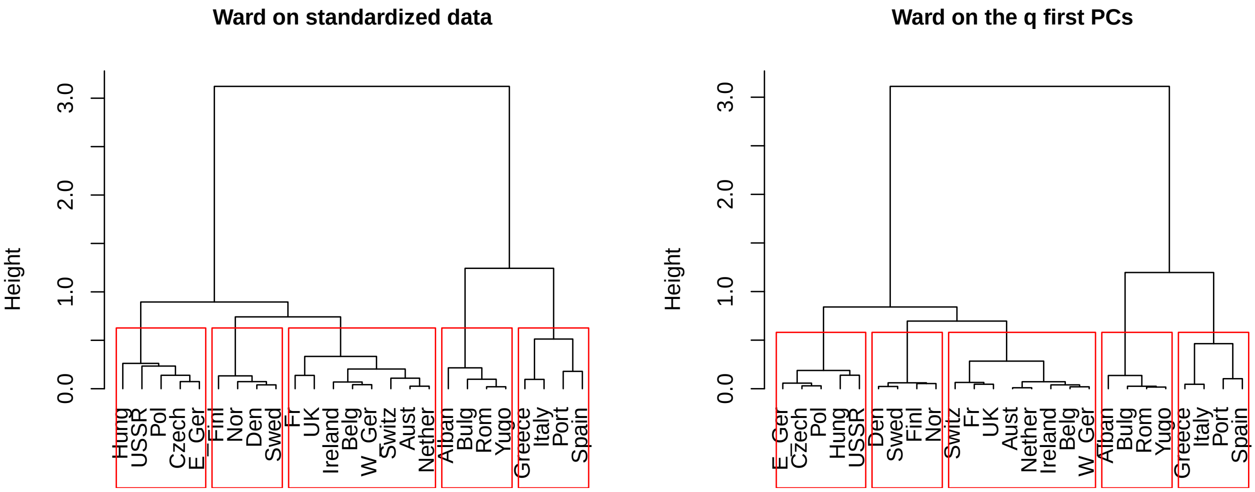
```
all.equal(tree_F$height,tree_Z$height)
```

Second part: Clustering on $q \leq r$ principal components.

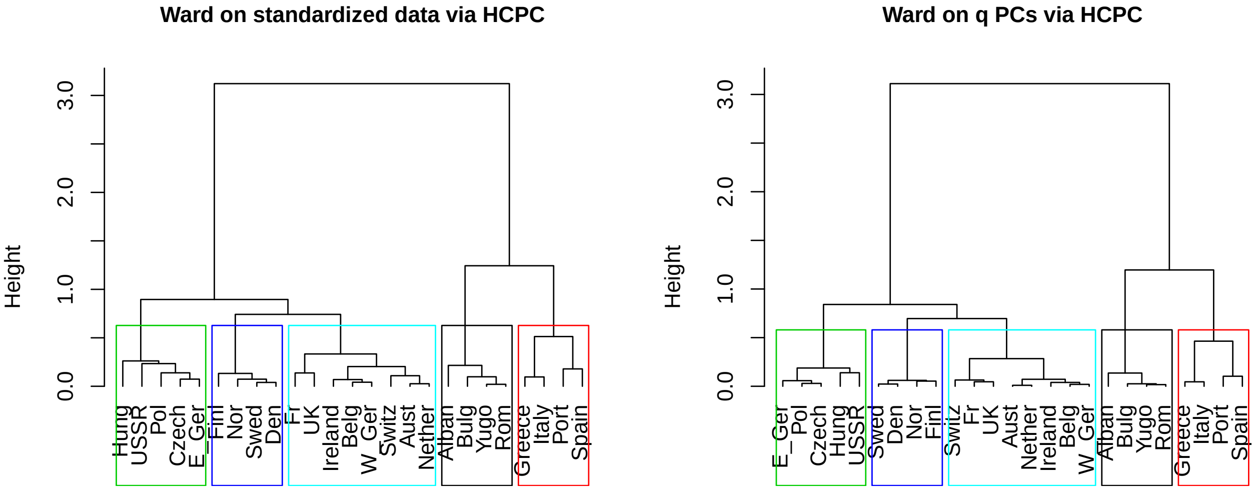
4. **Choose the number** q of principal components that summarizes "well" the data. What is the proportion of the variance of the data explained with these q principal components ?



5. Build the Ward dendrogram of the $n = 25$ european countries on the q **first principal components** of the **normalized PCA**. Use the function **rect.hclust** to obtain the graphics below and compare the 5-clusters partitions of the two dendrograms (the one on the standardized data and the one on the q first principal components).



6. Same question but using the function **HCPC()** of the **FactoMineR** package.



Exercice 7. Clustering numerical data: the cheeses dataset.

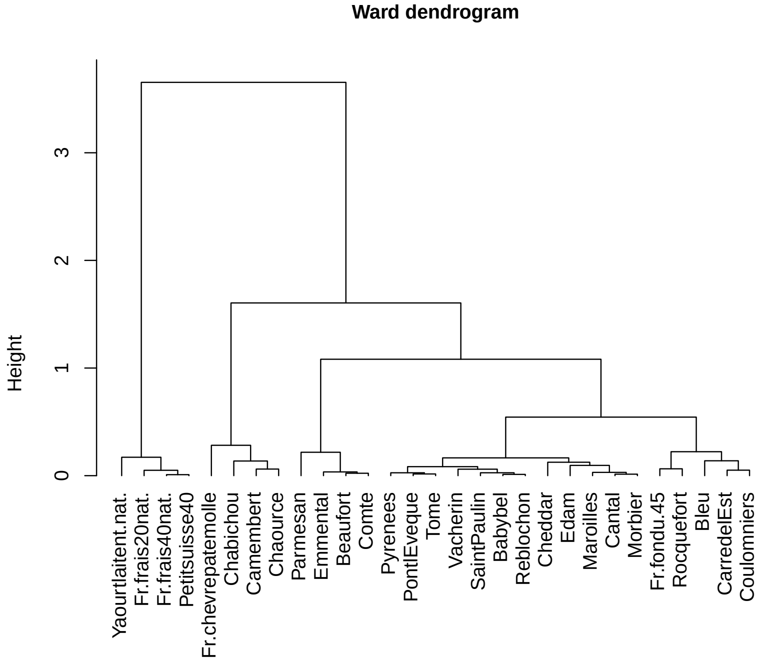
The dataset describes $n = 29$ cheeses on $p = 9$ numerical variables.

1. Import this dataset from the file "fromages.txt".

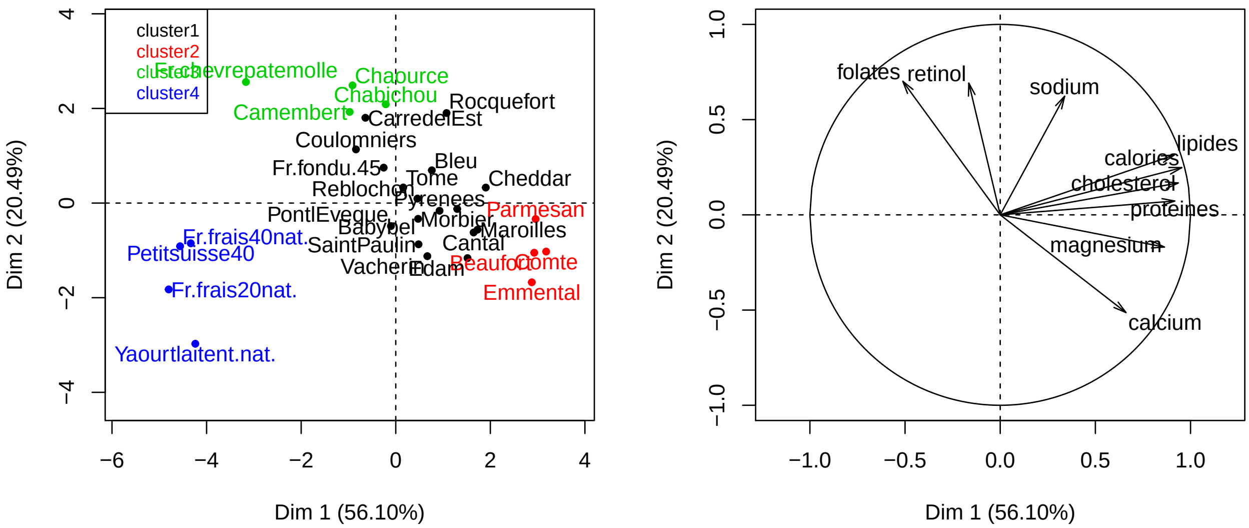
```
X <- read.table("../data/fromage.txt", header=TRUE,row.names=1)
```

2. Do you think these data shoud be scaled before clustering ?
3. Build the Ward dendrogram (with the cheeses weighted by $\frac{1}{n}$) on the **standatdized data**. Check that the sum of the heights is equal to the total inertia.

4. Plot the dendrogram and choose the number K of clusters that seems relevant to cut the tree.



5. Cut this tree and interpret the partition in K clusters using PCA (principal component analysis).



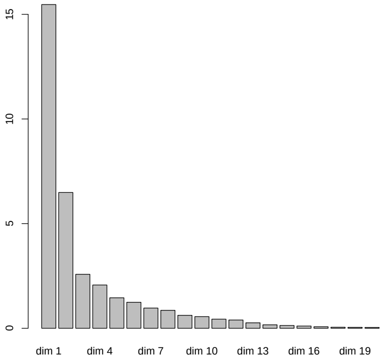
6. Confirm this interpretation using the **catdes()** R function.

Exercice 8. Clustering mixed data: the wines dataset.

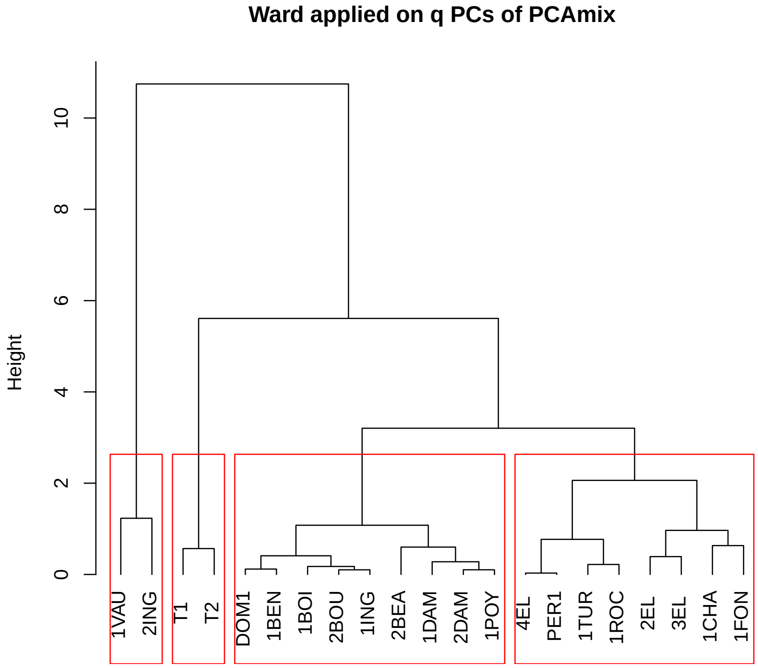
The wines dataset describes $n = 21$ wines on a mixture of $p = 31$ numerical and categorical variables.

```
library(PCAmixdata)
data(wine)
```

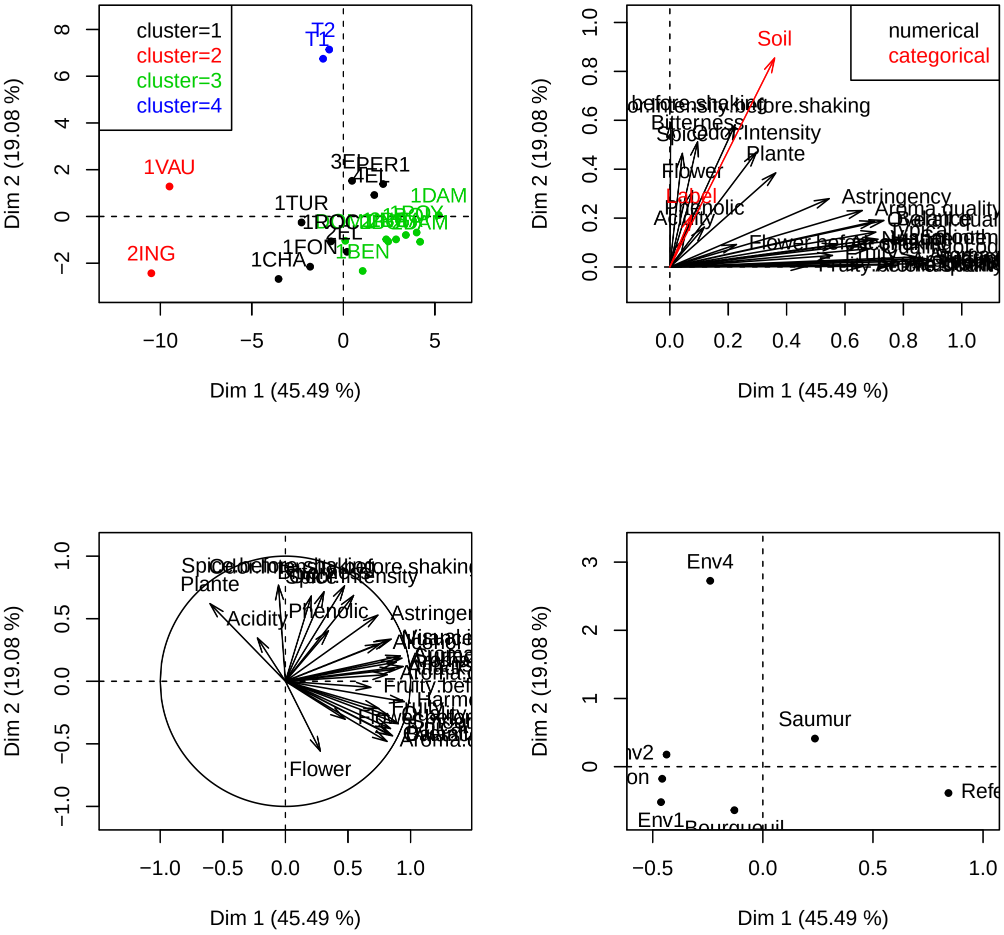
1. How many variables are categorical and how many are numerical ? How many levels for each categorical variable ?
2. Transform this dataset into a numerical dataset using the funcion **PCAmix()** of the R package **PCAmixdata** and choosing a number q of principal components.



2. Build the Ward dendrogram on the q first principal components and choose the number K of clusters that seems relevant to cut the tree.



3. Cut the tree and interpret the partition in K using PCAmix (PCA of a mixture of numerical and categorical variables).



4. Confirm this interpretation using the **catdes()** R function.

Appendix

The R function **hclust()** implements the ascendant hierarchichal clustering algorithm using the Lance & Williams formula. The Ward agregation measure $D(A, B) = \frac{\mu_{\mathcal{A}}\mu_{\mathcal{B}}}{\mu_{\mathcal{A}}+\mu_{\mathcal{B}}} d^2(g_{\mathcal{A}}, g_{\mathcal{B}})$ is then used only in the initialisation step where the aggregation measures between the singletons of the partition P_n are performed and stored in the $n \times n$ matrix $\Delta = [\delta_{ij}]$ knowing that:

$$\delta_{ij} := D(\{i\}, \{j\}) = \frac{w_i w_j}{w_i + w_j} d_{ij}^2.$$

When all the weights w_i are uniform (all equal to 1 or all equal to $\frac{1}{n}$ for instance) the function **hclust** implements the Ward minimum variance algorithm with the following arguments:

- `method = "ward.D"`,
- `d =` Δ,
- `members = NULL`.

The argument **members=NULL** (by default) means that the weights of the observations are considered as uniform. The argument **d** must be the matrix Δ of the *agregation measures* between the singletons. If all the observations are weighted by $1/n$, the argument **d** must then be the matrix $\Delta = \frac{\mathbf{D}^2}{2n}$ where $\mathbf{D} = [d_{ij}]$ is the matrix of the Euclidean distance between the observations. The R code is then:

```
> D <- dist(X)
> tree <- hclust(D^2/(2*n),method="ward.D")
```

If all the observations are weighted by 1, the argument **d** must be the matrix $\Delta = \frac{\mathbf{D}^2}{2}$.

When the weights w_i are non uniform the function **hclust** implements the Ward minimum variance algorithm with the following arguments:

- `method = "ward.D"`,
- `d =` Δ,
- `members = w`.

The argument **members=w** with **w**! =NULL means that the weights w_i of the observations are non uniform. The argument $d = \Delta$ is then more complicated to perform. For instance the following R code can be used:

```
> Delta <-  D
> for (i in 1:(n-1)) {
    for (j in (i+1):n) {
      Delta[n*(i-1) - i*(i-1)/2 + j-i] <-
        Delta[n*(i-1) - i*(i-1)/2 + j-i]^2*w[i]*w[j]/(w[i]+w[j])}}
> tree <- hclust(Delta,method="ward.D",members=w)
```