Classification automatique

TP4

CAH sous R

1. Chargement des données

Charge les données du fichier hotel.txt, identifier les différentes variables du fichier et leur nature.

```r
data = read.csv('HOTEL.txt')

z <- data.frame(variable = names(data),
                classe = sapply(data, class),
                distinct_value = sapply (data, function(x) paste0(length(unique(x)))),
                first_values = sapply(data, function(x) paste0(head(x)[1:4], collapse = ', ')),
                row.names=NULL)
z
```

variable	classe	distinct_value	first_values
NOM	factor	39	Appolpon, Caravel, Christina, Economy
PAYS	factor	5	Grèce, Grèce, Grèce, Grèce
ETOILE	integer	6	1, 4, 2, 1
CONFORT	integer	8	4, 7, 7, 3
CHAMBRE	integer	33	56, 471, 93, 56
CUISINE	integer	9	2, 7, 3, 1
SPORT	integer	10	0, 6, 0, 0
PLAGE	integer	8	8, 5, 5, 8
PRIX	integer	35	390, 468, 427, 369

```r
data$NOM <- toString(data$NOM)
data$ETOILE <- as.factor(data$ETOILE)
data$PAYS <-as.factor(data$PAYS)

head(data)
```

NOM
Appolpon, Caravel, Christina, Economy, Eden Beach, Hanikian Beach, Marina Beach, Xenia, Agdal, Almohades, Atlas, Atla
Appolpon, Caravel, Christina, Economy, Eden Beach, Hanikian Beach, Marina Beach, Xenia, Agdal, Almohades, Atlas, Atl
Appolpon, Caravel, Christina, Economy, Eden Beach, Hanikian Beach, Marina Beach, Xenia, Agdal, Almohades, Atlas, Atl
Appolpon, Caravel, Christina, Economy, Eden Beach, Hanikian Beach, Marina Beach, Xenia, Agdal, Almohades, Atlas, Atl
Appolpon, Caravel, Christina, Economy, Eden Beach, Hanikian Beach, Marina Beach, Xenia, Agdal, Almohades, Atlas, Atl
Appolpon, Caravel, Christina, Economy, Eden Beach, Hanikian Beach, Marina Beach, Xenia, Agdal, Almohades, Atlas, Atl

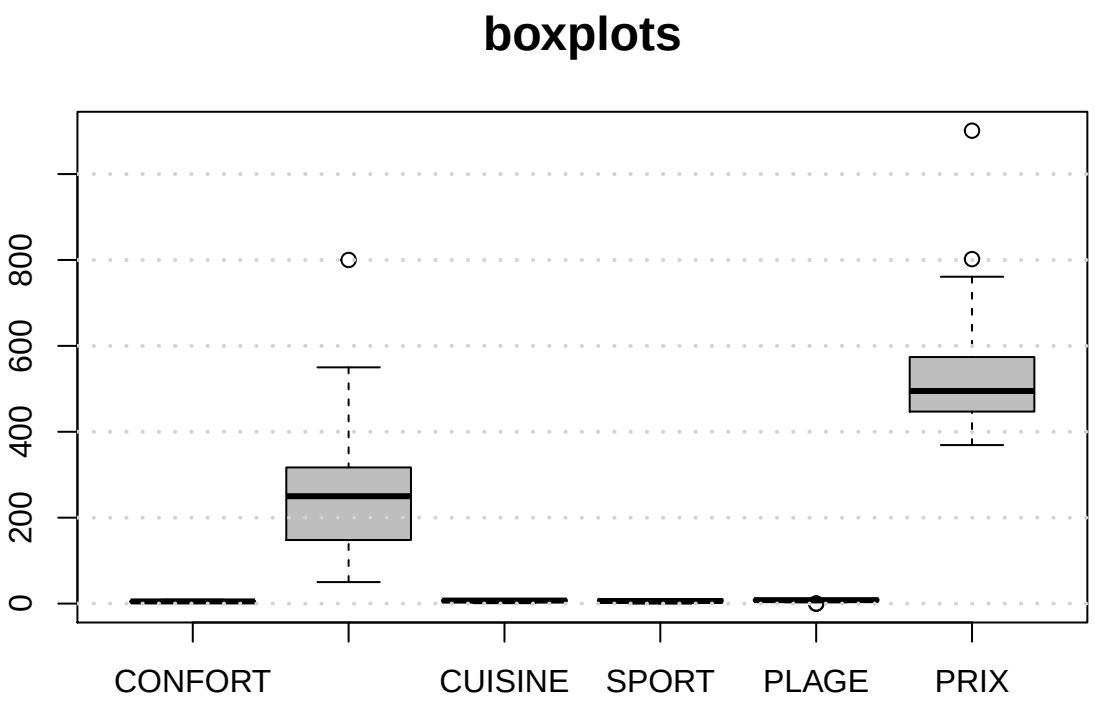
2.Analyse descriptive

Donner les différentes statistiqus descriptives et liasons entre les variables quantitatives du fichier.

```r
# extraire les données quantitatives
ind.quanti <- sapply(data, function(x) is.numeric(x) | is.integer(x))

# variables quantitative
data.quanti <- data[ ,ind.quanti]

# extraire les données quanlitatives
ind.quali <- sapply(data, function(x) is.factor(x))

# variables quantitative
data.quali <- data[ ,ind.quali]


dtf <- data.frame(round(sapply(data.quanti, each(min, max, mean, sd, var, median, IQR)),2))
xtable(dtf)
```

	CONFORT	CHAMBRE	CUISINE	SPORT	PLAGE	PRIX
min	2.00	50.00	1.00	0.00	0.00	369.00
max	9.00	800.00	10.00	10.00	10.00	1101.00
mean	5.18	261.21	6.67	6.23	7.77	529.90
sd	1.57	149.83	2.65	3.44	2.72	137.87
var	2.47	22449.75	7.02	11.87	7.39	19006.99
median	5.00	250.00	7.00	6.00	8.00	495.00
IQR	2.00	169.00	4.00	6.00	3.50	127.00

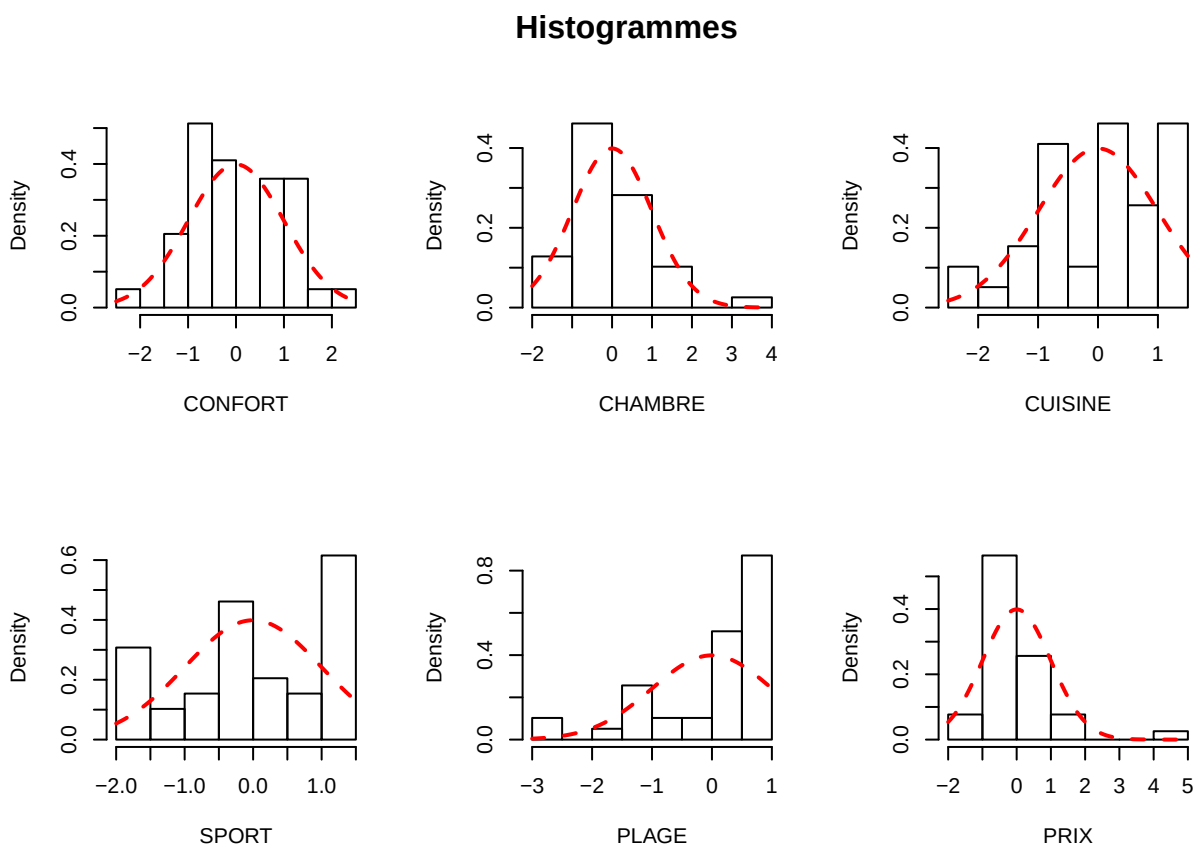
```r
par(mfrow=c(1,1))
boxplot(data.quanti,main='boxplots',cex.main=1.5,col='grey')
grid(nx = 1, ny = NULL,lwd = 2)
```



```r
#standardisation des données
data.quanti.norm <- scale(data.quanti,center=T,scale=T)
```

- Visualisation de la distribution des variables :

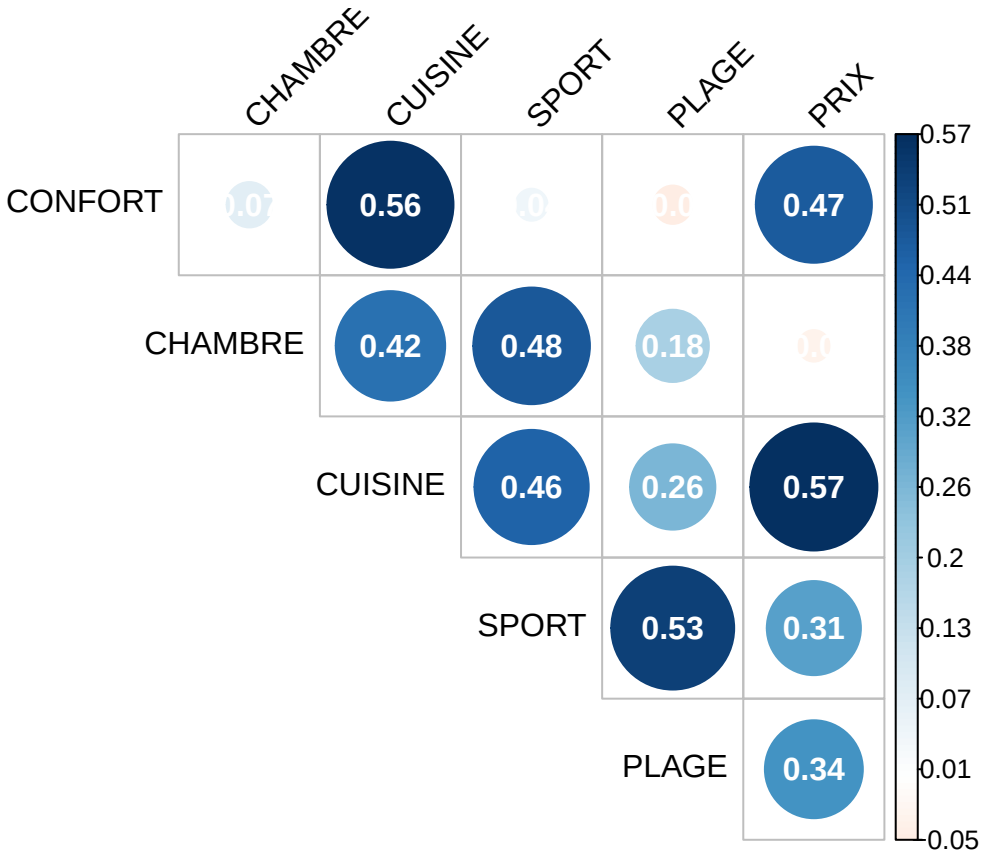
```r
l = ncol(data.quanti.norm)
par(mfrow=c(2,3))
for (i in 1:l) {
  hist(data.quanti.norm[,i],probability=TRUE,xlab=colnames(data.quanti.norm)[i],main='')
  curve(dnorm(x), add=T, col="red", lwd=2, lty=2)
}
title("Histogrammes", outer=TRUE, line=-1,cex.main=1.5)
```




```r
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```r
R=cor(data.quanti.norm)
corrplot(R, tl.col="black", tl.srt=45,is.corr = FALSE,addCoef.col = "white",diag=FALSE,type='upper')
```

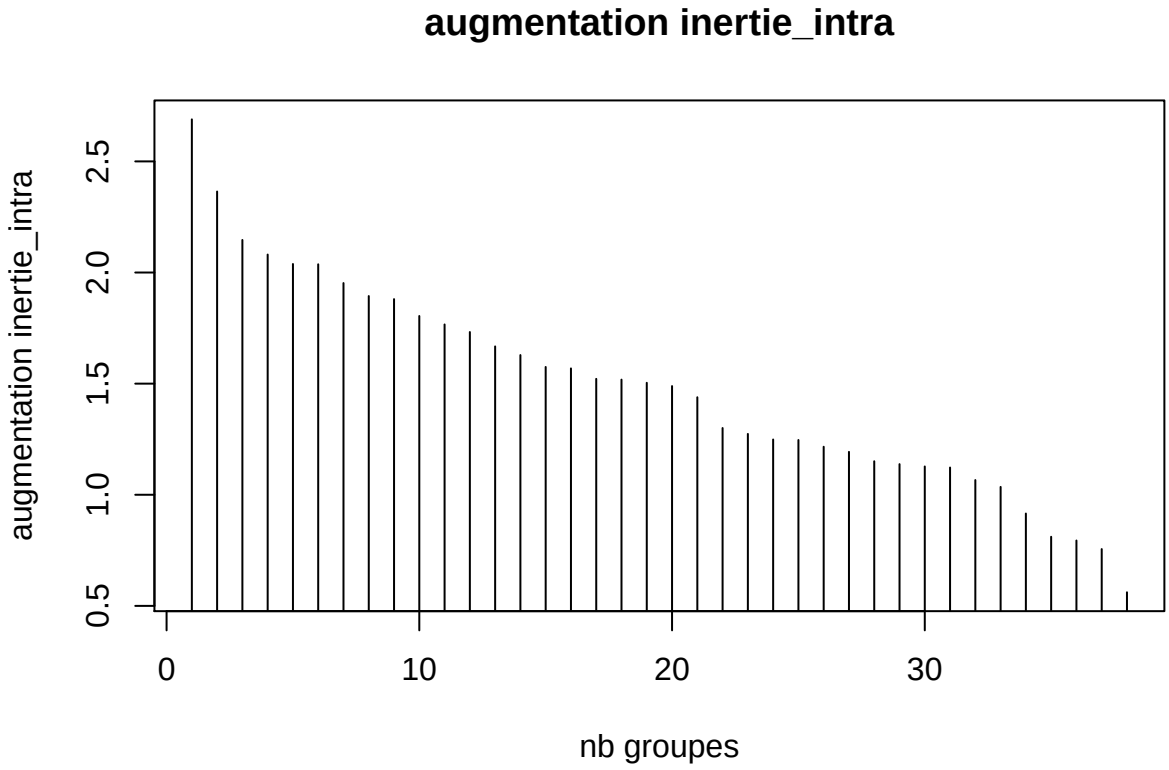


3.

4. Classification ascendante

4.1 CAH distance=euclidienne, methode=simple

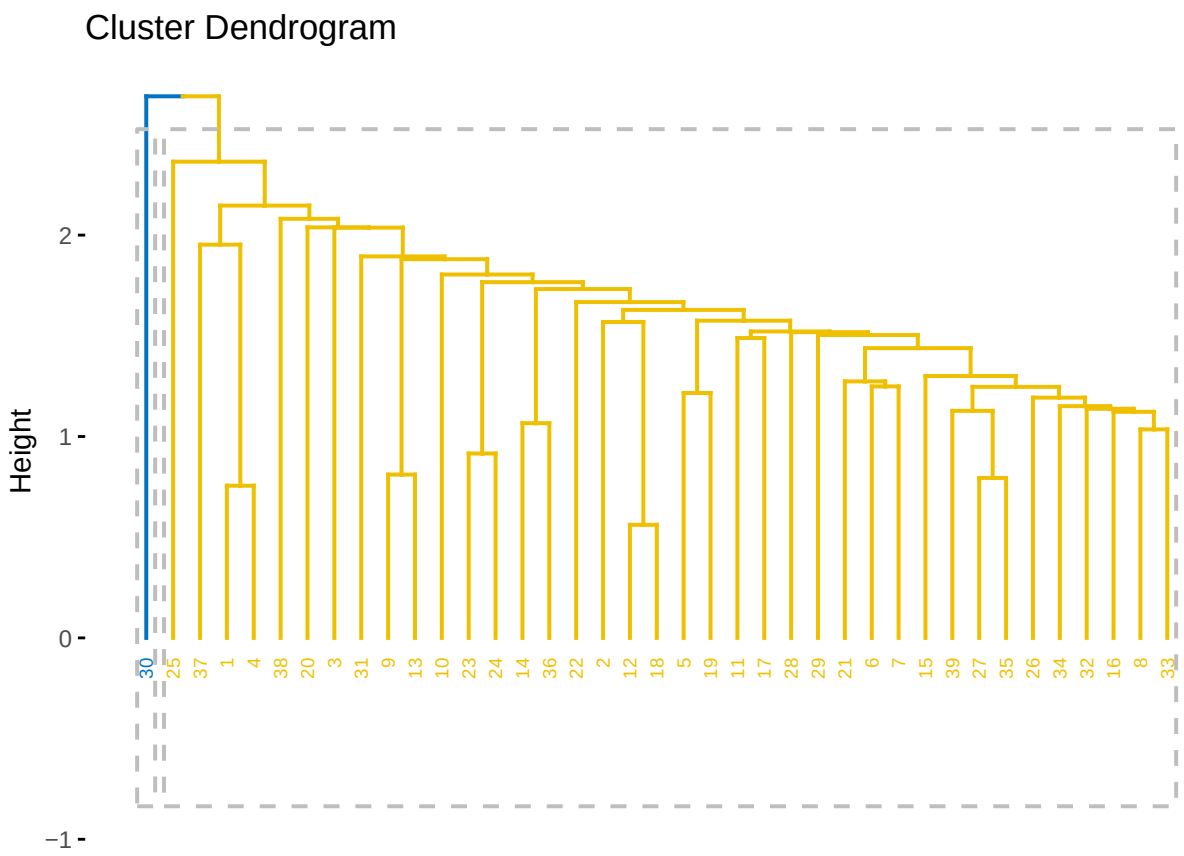
```
CAH.eucli.simple <- hclust(dist(data.quanti.norm,method = 'euclidian'),
                           method='single')

h <- CAH.eucli.simple$height
plot((nrow(data.quanti.norm)-1):1,
     h,
     xlab="nb groupes",
     ylab="augmentation inertie_intra",
     type="h")
title ("augmentation inertie_intra")
```




```r
library(factoextra)
```

```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ
```

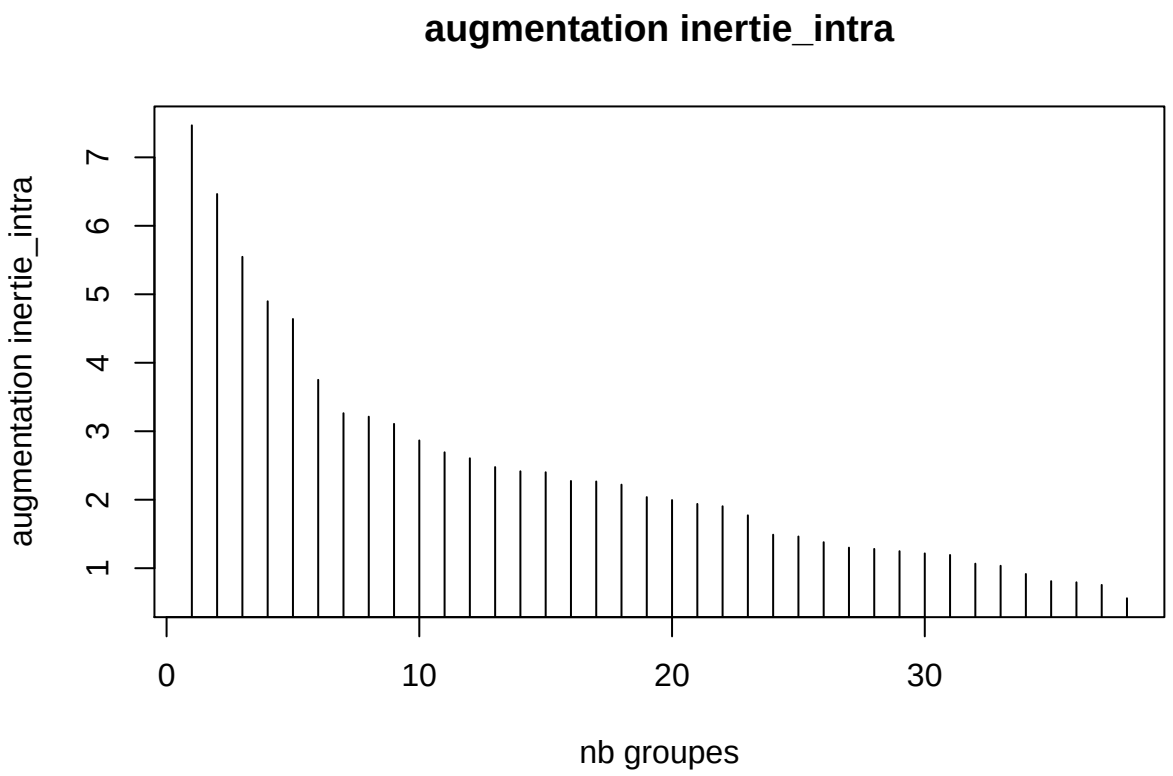
```r
fviz_dend(CAH.eucli.simple, # cluster
          k = 2, # nombre de classe
          cex = 0.5, # taille du label
          palette = "jco", # choix couleurs
          color_labels_by_k = TRUE, # couleur par label
          rect = TRUE # rectangle autoutr des classes
          )
```



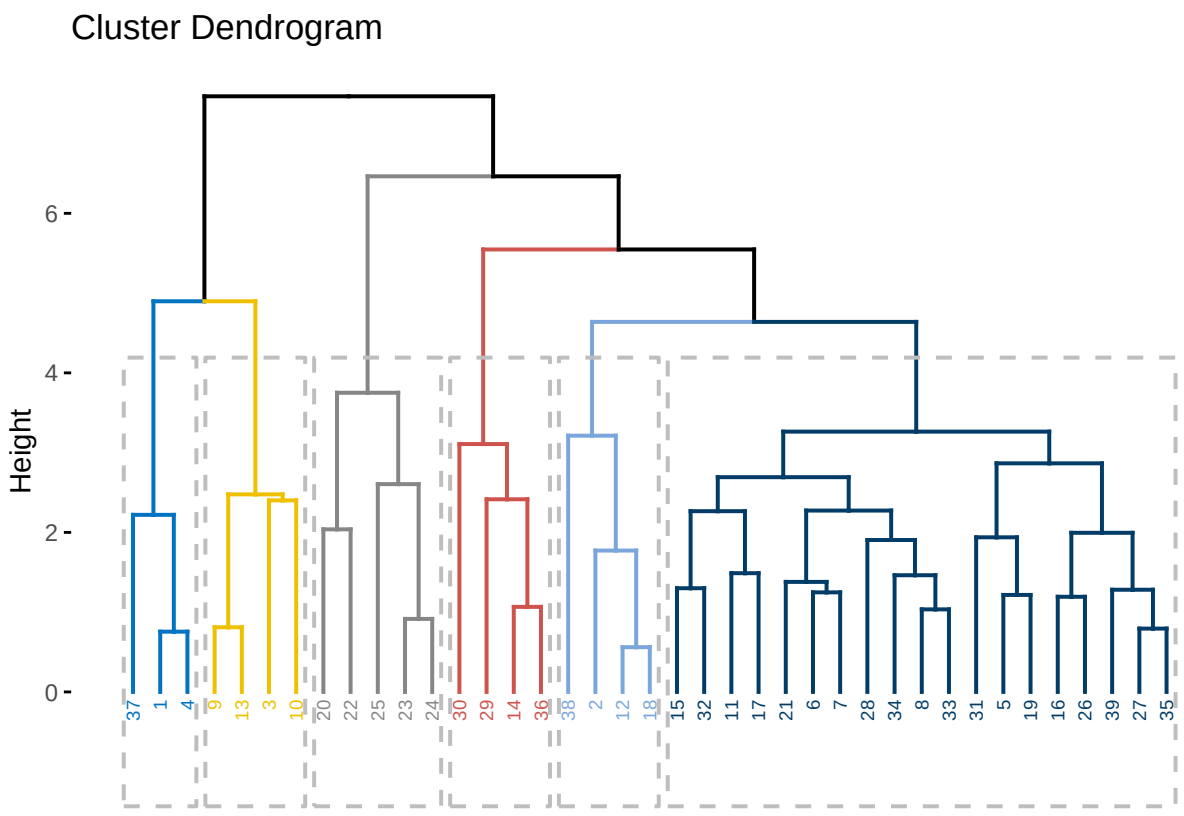
Commentaires:
Le saut d'inertie intra, n'étant pas élevé, les classes ne sont pas homogénes

4.2 CAH distance=euclidienne, methode=complete

```
CAH.eucli.complete <- hclust(dist(data.quanti.norm,method = 'euclidian'),
                             method='complete')

h <- CAH.eucli.complete$height
plot((nrow(data.quanti.norm)-1):1,
     h,
     xlab="nb groupes",
     ylab="augmentation inertie_intra",
     type="h")
title ("augmentation inertie_intra")
```




```
fviz_dend(CAH.eucli.complete , # cluster
          k = 6, # nombre de classe
          cex = 0.5, # taille du label
          palette = "jco", # choix couleurs
          color_labels_by_k = TRUE, # couleur par label
          rect = TRUE # rectangle autoutr des classes
          )
```

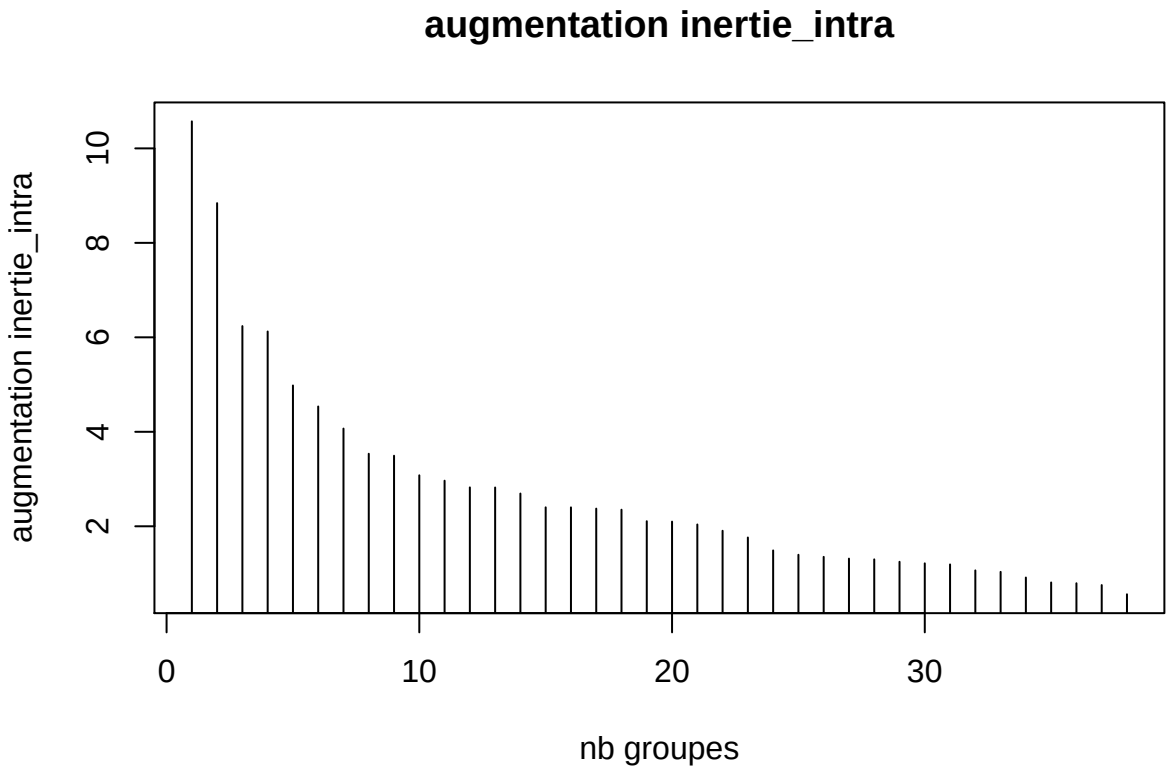


4.3 CAH distance=euclidienne, methode=ward

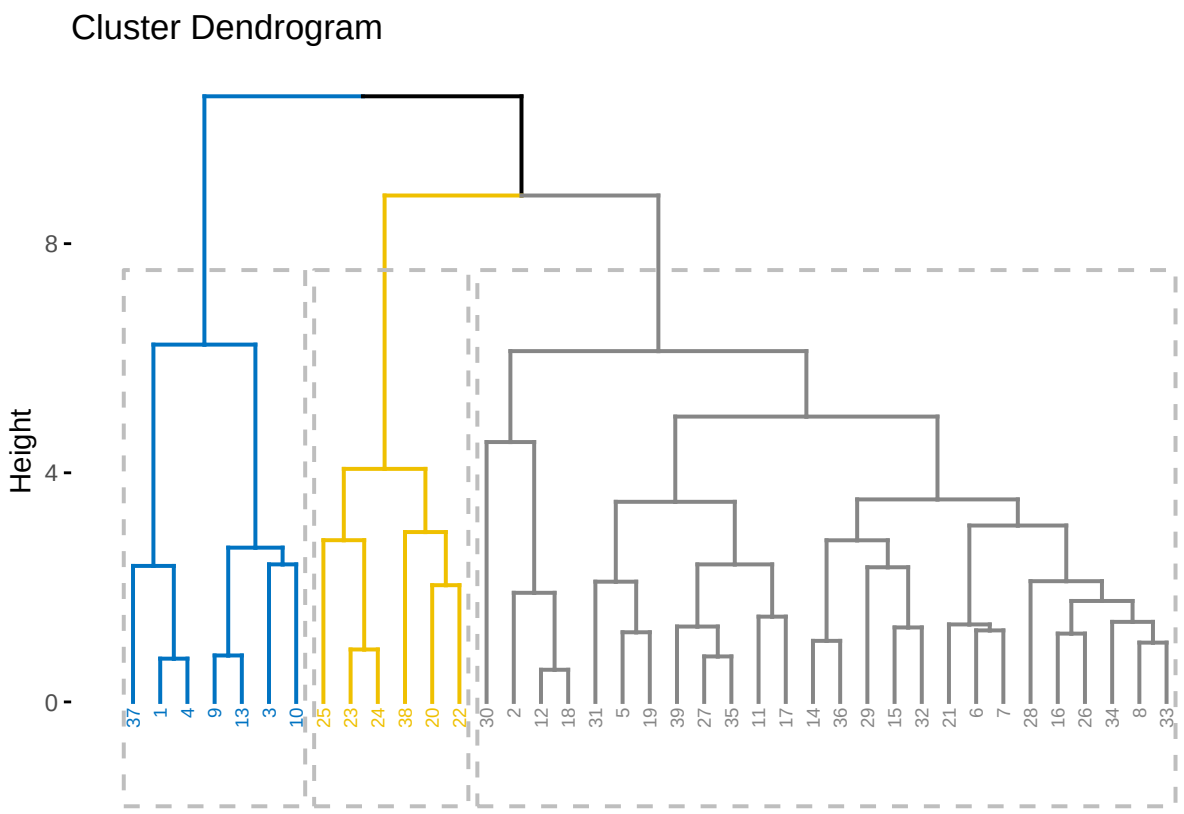
```
CAH.eucli.ward <- hclust(dist(data.quanti.norm,method = 'euclidian'),

                                                    method='ward.D2')
summary(CAH.eucli.ward)
```

```
##             Length Class  Mode
## merge       76     -none- numeric
## height      38     -none- numeric
## order       39     -none- numeric
## labels       0     -none- NULL
## method       1     -none- character
## call         3     -none- call
## dist.method  1     -none- character
```

```
h <- CAH.eucli.ward$height
plot((nrow(data.quanti.norm)-1):1,
     h,
     xlab="nb groupes",
     ylab="augmentation inertie_intra",
     type="h")
title ("augmentation inertie_intra")
```

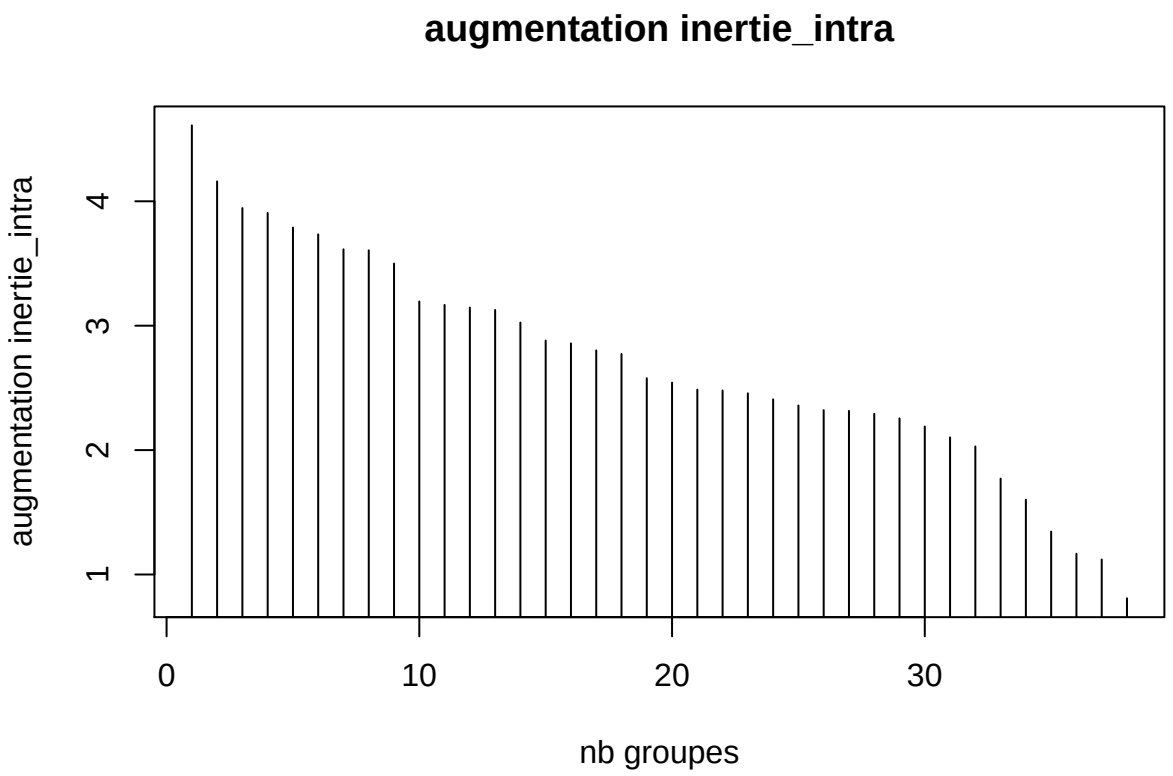



```
fviz_dend(CAH.eucli.ward , # cluster
          k = 3, # nombre de classe
          cex = 0.5, # taille du label
          palette = "jco", # choix couleurs
          color_labels_by_k = TRUE, # couleur par label
          rect = TRUE # rectangle autoutr des classes
          )
```

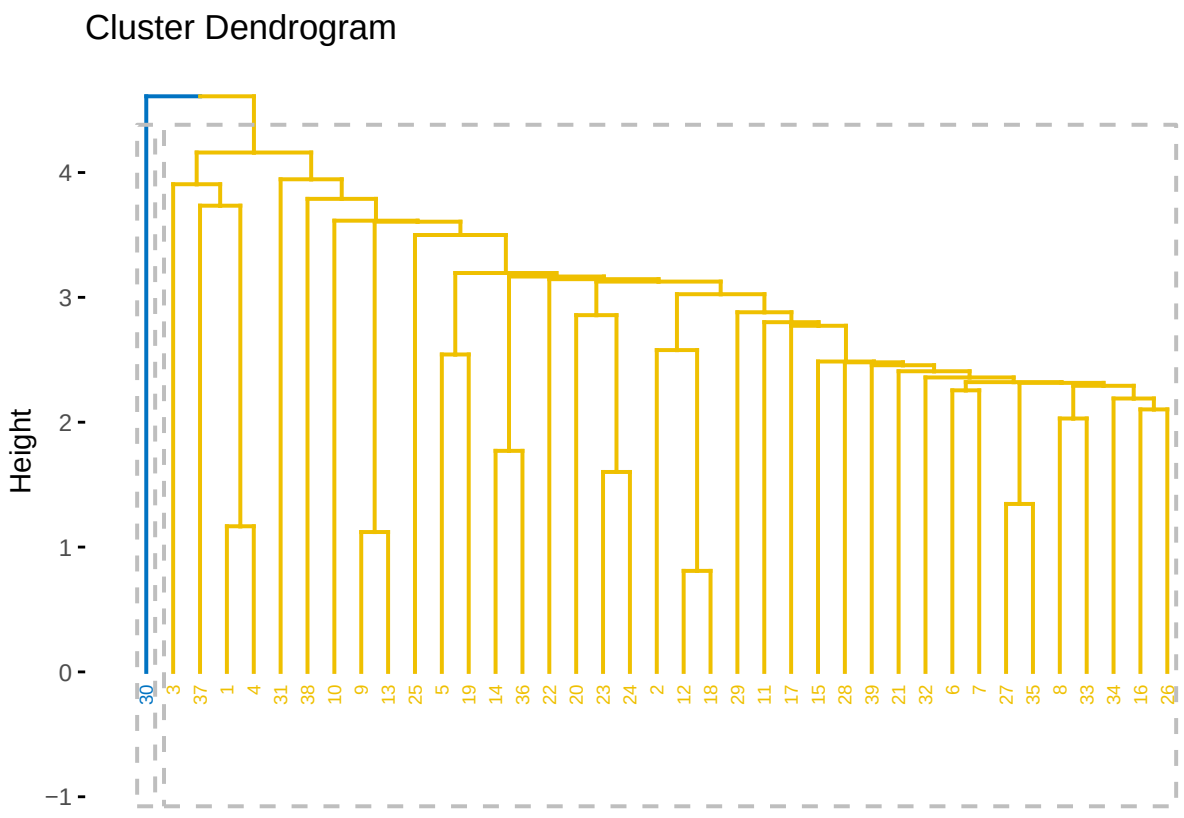


4.4 CAH distance=manhatan, methode=simple

```
CAH.man.simple <- hclust(dist(data.quanti.norm,method = 'manhattan'),
                         method='single')

h <- CAH.man.simple$height
plot((nrow(data.quanti.norm)-1):1,
     h,
     xlab="nb groupes",
     ylab="augmentation inertie_intra",
     type="h")
title ("augmentation inertie_intra")
```

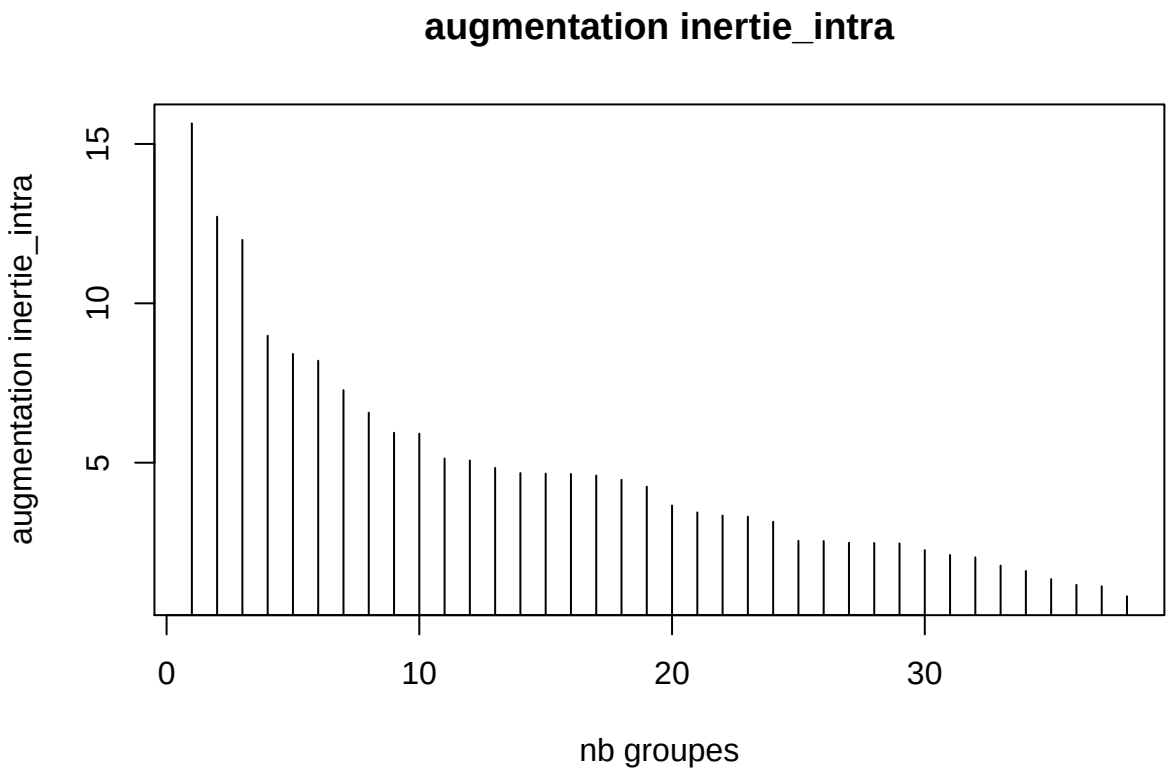



```
fviz_dend(CAH.man.simple , # cluster
          k = 2, # nombre de classe
          cex = 0.5, # taille du label
          palette = "jco", # choix couleurs
          color_labels_by_k = TRUE, # couleur par label
          rect = TRUE # rectangle autoutr des classes
          )
```

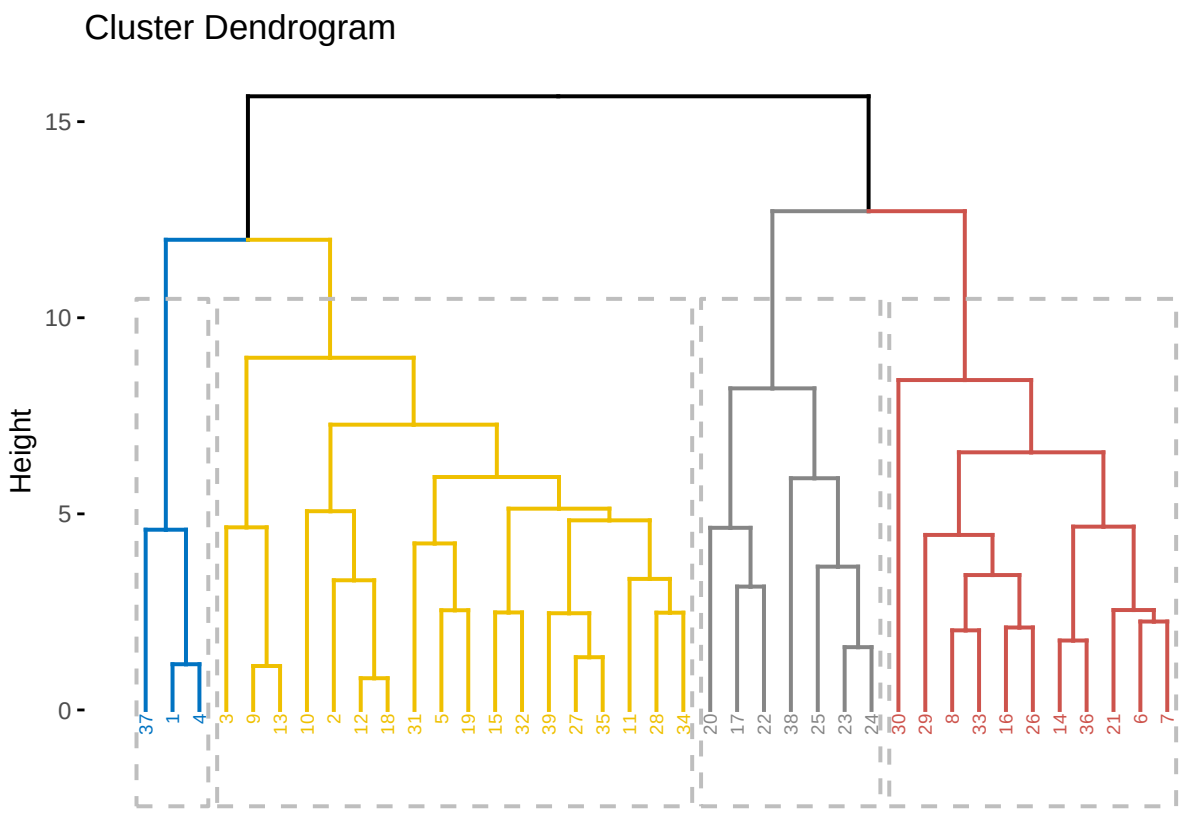


4.5 CAH distance=manhattan, methode=complete

```
CAH.man.complete <- hclust(dist(data.quanti.norm,method = 'manhattan'),
                           method='complete')

h <- CAH.man.complete$height
plot((nrow(data.quanti.norm)-1):1,
     h,
     xlab="nb groupes",
     ylab="augmentation inertie_intra",
     type="h")
title ("augmentation inertie_intra")
```



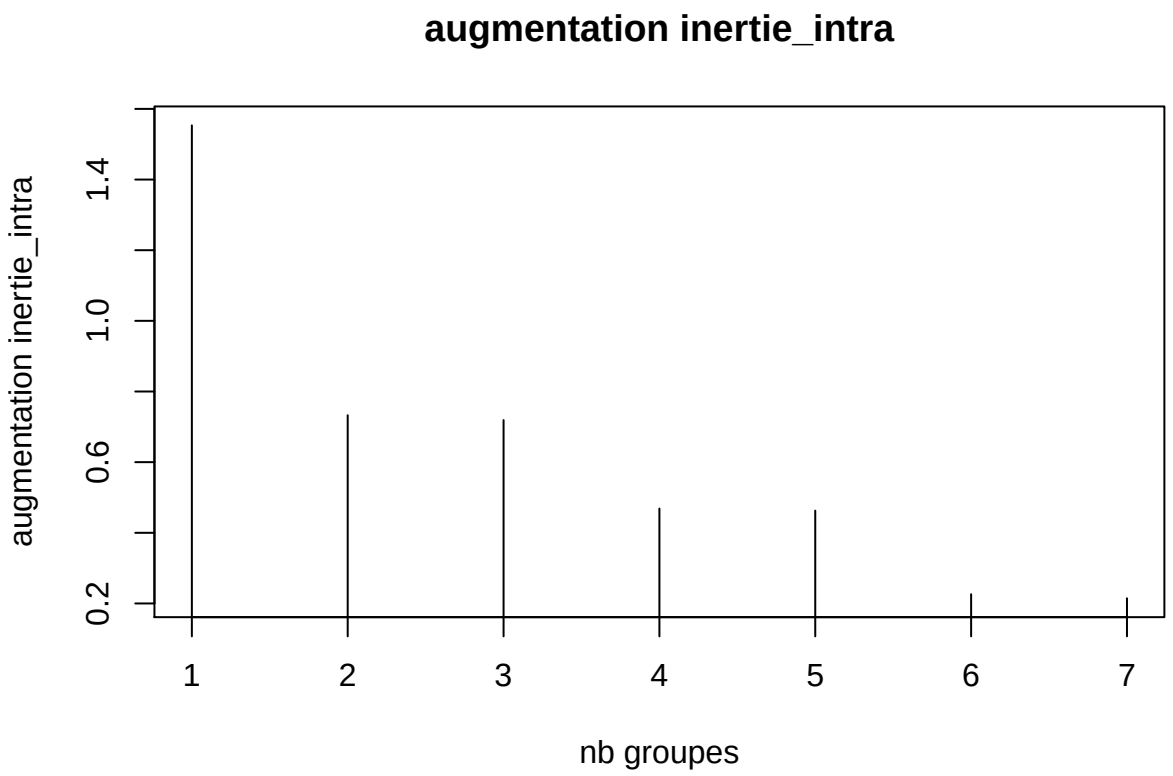

```
fviz_dend(CAH.man.complete , # cluster
          k = 4, # nombre de classe
          cex = 0.5, # taille du label
          palette = "jco", # choix couleurs
          color_labels_by_k = TRUE, # couleur par label
          rect = TRUE # rectangle autoutr des classes
          )
```



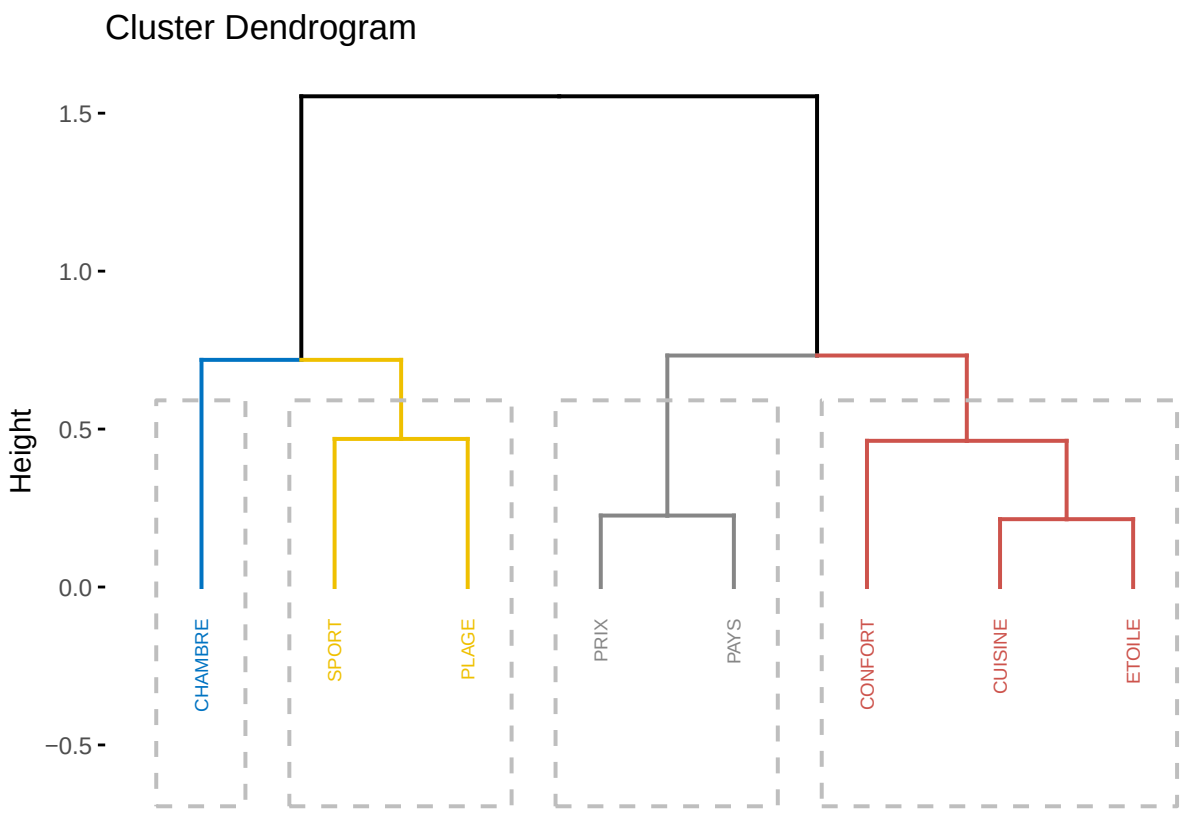
5. CAH - Variables

Reprendre la question précédente sur les variables.

```
CAH.var<-hclustvar(X.quanti=data.quanti.norm,X.quali=data.quali)

h <- CAH.var$height
plot((ncol(data.quanti.norm)+ncol(data.quali)-1):1,
     h,
     xlab="nb groupes",
     ylab="augmentation inertie_intra",
     type="h")
title ("augmentation inertie_intra")
```




```
fviz_dend(CAH.var ,
          k = 4, # nombre de classe
          cex = 0.5, # taille du label
          palette = "jco", # choix couleurs
          color_labels_by_k = TRUE, # couleur par label
          rect = TRUE # rectangle autoutr des classes
          )
```


Cluster Dendrogram
Height
1.5
1.0
0.5
0.0
−0.5
CHAMBRE
SPORT
PLAGE
PRIX
PAYS
CONFORT
CUISINE
ETOILE