Rendu TP1 : Apprentissage non supervisé

Analyse en composantes principales

On centre les données pour garder leurs variations et non leurs valeurs absolues.

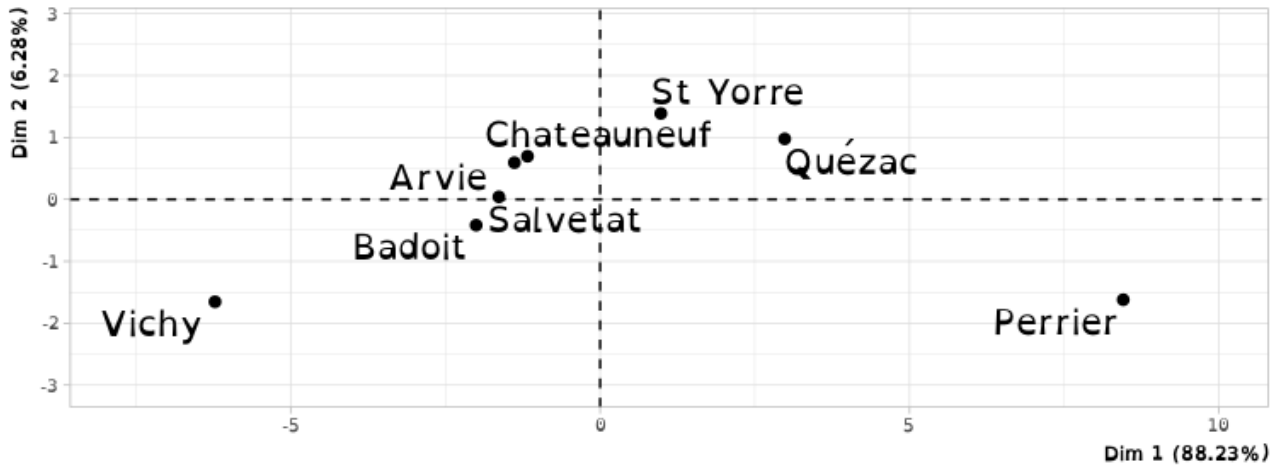
Sur la matrice de corrélation, on voit que les saveurs `alcaline` et `sucrée` indiquent une caractéristique commune et que l'`acide` et l'`amère` indiquent une caractéristique proche. À l'inverse la relation entre les saveurs `amère` et `sucrée`, `amère` et `alcaline`, `acide` et `alcaline` indique des caractéristiques très distinctes.

La matrice des distances entre les individus nous indique les individus qui ont les données les moins similaires. Par exemple, `Vichy` et `Perrier` sont très différentes avec une valeur de `14.7`. On peut le vérifier directement dans les données brutes.

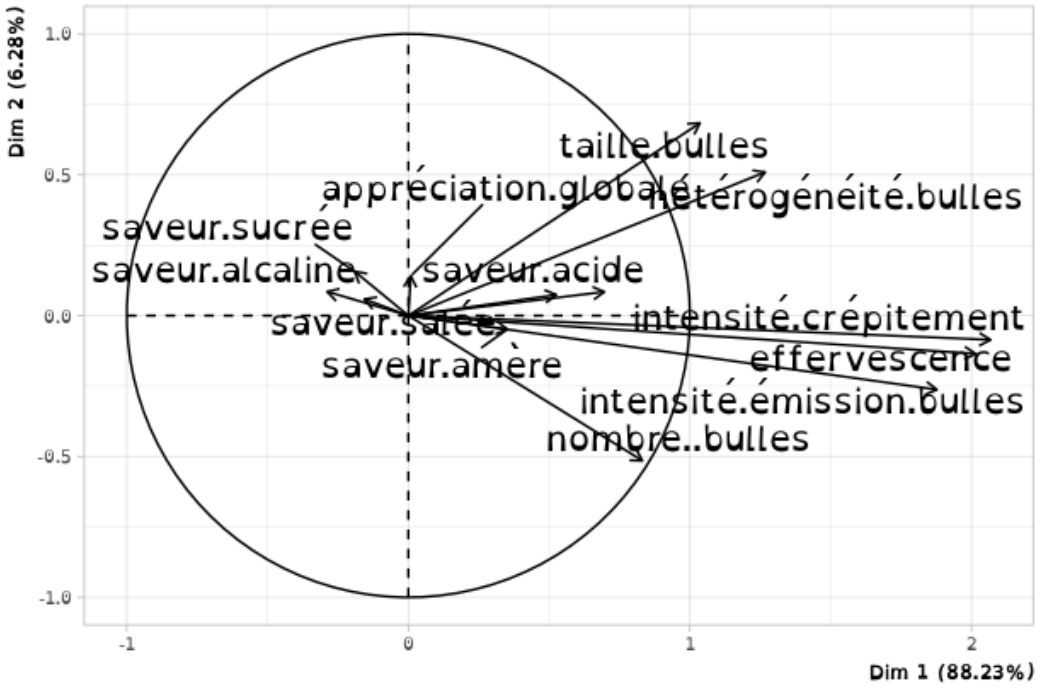
	saveur.amère	saveur.sucrée	saveur.acide	saveur.salée	saveur.alcaline	appréciation.globale	intensité.émission.bulles
St Yorre	-0.5072997	1.03143052	-0.2709005	1.6833150	0.60915786	-0.5651411	0.012321
Badoit	0.3043798	-0.55538567	-0.6043166	-0.6059934	-0.20305262	-0.5651411	-0.381980
Vichy	-1.5218990	0.39670405	-1.6045647	1.1446542	1.15063151	-0.9129202	-1.219873
Quézac	0.5072997	-0.55538567	1.2294716	-0.6059934	-0.74452627	1.5215337	0.603775
Arvie	-1.1160592	1.34879376	-0.1041925	0.0673326	1.15063151	-0.5651411	-0.381980
Chateauneuf	0.1014599	0.07934081	-0.1041925	0.0673326	0.06768421	0.8259754	-0.283405
Salvetat	0.7102195	0.07934081	-0.1041925	-1.4139846	-0.20305262	1.1737546	-0.480556
Perrier	1.5218990	-1.82483862	1.5628877	-0.3366630	-1.82747358	-0.9129202	2.131698

Pour l'ACP

- Sur la figure des individus, on voit des individus groupés au centre et quelques individus éloignés.

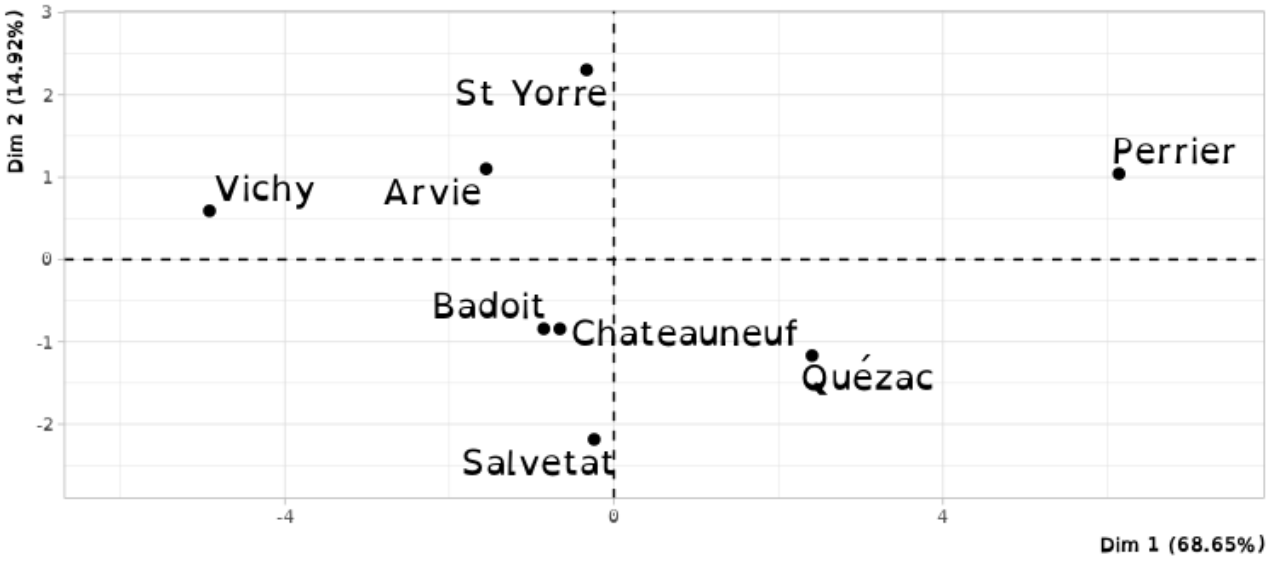


- Sur la figure des variables, on voit que des vecteurs sorte du cercle °~°.

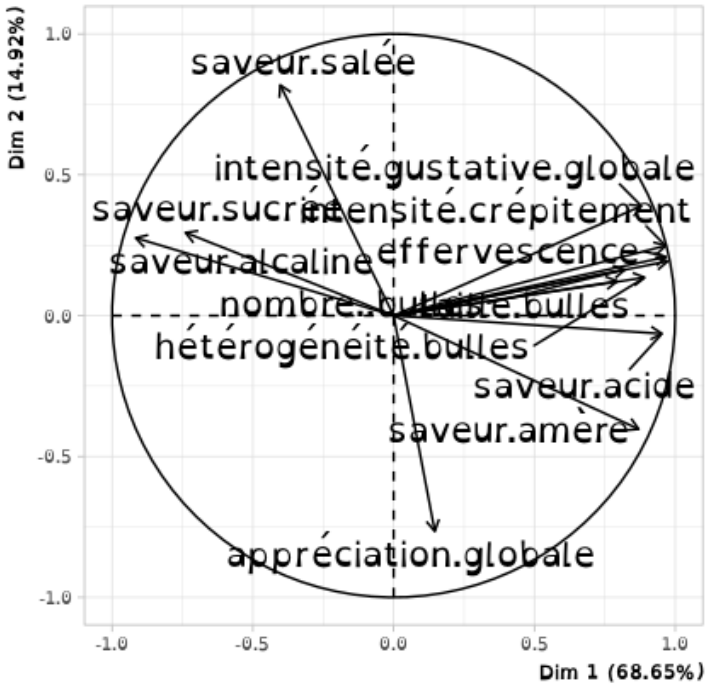


Pour l'ACP normé

- Sur la figure des individus, on voit que les individus sont plus dispersés, mais quelques individus restent éloignés comme `Vichy` et `Perrier`.



- Sur la figure des variables, on voit que des vecteurs ne sortent plus du cercle °~°.



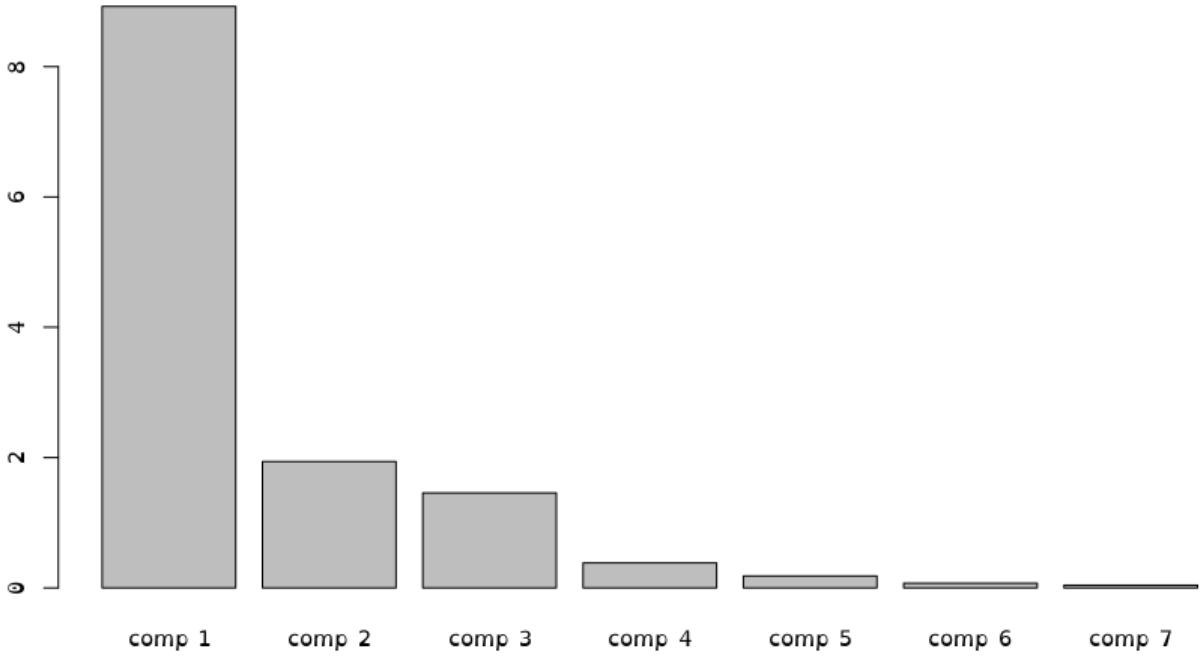
Choix du nombre de composantes à retenir

Avec la `règle de Kaiser` ont garde `3` composantes, car les `3` premières valeurs propres sont supérieur à `1`.

Les valeurs propres sont nommées $λ_d$ avec d la dimension (c'est la variance empirique). La valeur propre par dimension :

- Avec `1` composante on couvre `68.65%` des données significatives et une valeur propre de $λ_1=8.9$.
- Avec `2` composantes on couvre `83.57%` des données significatives et une valeur propre de $λ_2=1.9$.
- Avec `3` composantes on couvre `94.77%` des données significatives et une valeur propre de $λ_3=1.5$.
- Avec `4` composantes on couvre `97.73%` des données significatives et une valeur propre de $λ_4=0.3$.

Avec la règle du coude ont garde `2` composantes. Visuellement, on voit que la cassure est dès la `2`ème composante.



Par le calcul, on tombe sur 2 aussi, on arrête le calcul dès que $δ_x$ est négatif puis on compte le nombre de $δ_x$.

Calcul des différences premières :

- $ε_1 = (λ_1-λ_2), ε_2 = (λ_2-λ_3), ε_3 = (λ_3-λ_4)$
- $ε_1 = (8.92-1.93), ε_2 = (1.93-1.45), ε_3 = (1.45-0.38)$
- $ε_1 = 7.99, ε_2 = 0.48, ε_3 = 1.07$

Calcul des différences secondes :

- $δ_1 = (ε_1-ε_2), δ_2 = (ε_2-ε_3)$
- $δ_1 = (7.99-0.48), δ_2 = (0.48-1.07)$
- $δ_1 = 7.51, δ_2 = -0.59$

On a une bonne qualité de la projection et contribution des individus avec 3 dimensions, car ils contribuent tous à plus de 71%, ce qui est mieux que 2 dimensions où la contribution la plus basse est de 55% :

```
resnorm$ind$cos2[,1]+resnorm$ind$cos2[,2]
 St Yorre      Badoit       Vichy      Quézac     Arvie Chateauneuf    Salvetat      Perrier
0.6912732   0.3777300   0.8882563   0.7323157   0.6012887   0.5518255   0.8856256   0.9360553
# Et les 3 premières ?
resnorm$ind$cos2[,1]+resnorm$ind$cos2[,2]+resnorm$ind$cos2[,3]
 St Yorre      Badoit       Vichy      Quézac     Arvie Chateauneuf    Salvetat      Perrier
0.9405832   0.7955333   0.9777641   0.9271582   0.7145190   0.7536751   0.8925276   0.9988988
```

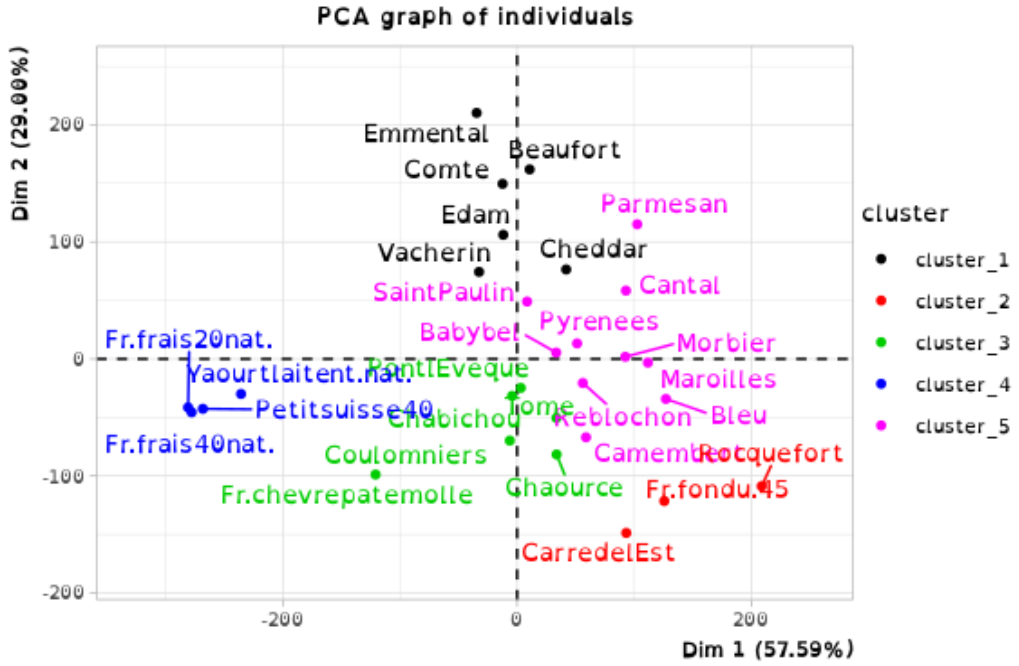
On a une bonne qualité de la projection et contribution des variables avec 3 dimensions, car elles contribuent toutes à plus de 84%, ce qui est mieux que 2 dimensions où la contribution la plus basse est de 60% :

```
             saveur.amère              saveur.sucrée               saveur.acide
                0.9193755                  0.6333039                  0.9122874
             saveur.salée            saveur.alcaline        appréciation.globale
                0.8321232                  0.9153633                  0.6073619
intensité.émission.bulles           nombre..bulles                taille.bulles
                0.9842741                  0.7013436                  0.6455454
     hétérogénéité.bulles           effervescence intensité.gustative.globale
                0.8128302                  0.9801694                  0.9252784
    intensité.crépitement
                0.9945345
# Et les 3 ?
resnorm$var$cos2[,1]+resnorm$var$cos2[,2]+resnorm$var$cos2[,3]
             saveur.amère              saveur.sucrée               saveur.acide
                0.9362833                  0.9247121                  0.9568183
             saveur.salée            saveur.alcaline        appréciation.globale
                0.8409480                  0.9733153                  0.9419222
intensité.émission.bulles           nombre..bulles                taille.bulles
                0.9917772                  0.9214411                  0.9139851
     hétérogénéité.bulles           effervescence intensité.gustative.globale
                0.9842623                  0.9802432                  0.9605466
    intensité.crépitement
                0.9948487
```

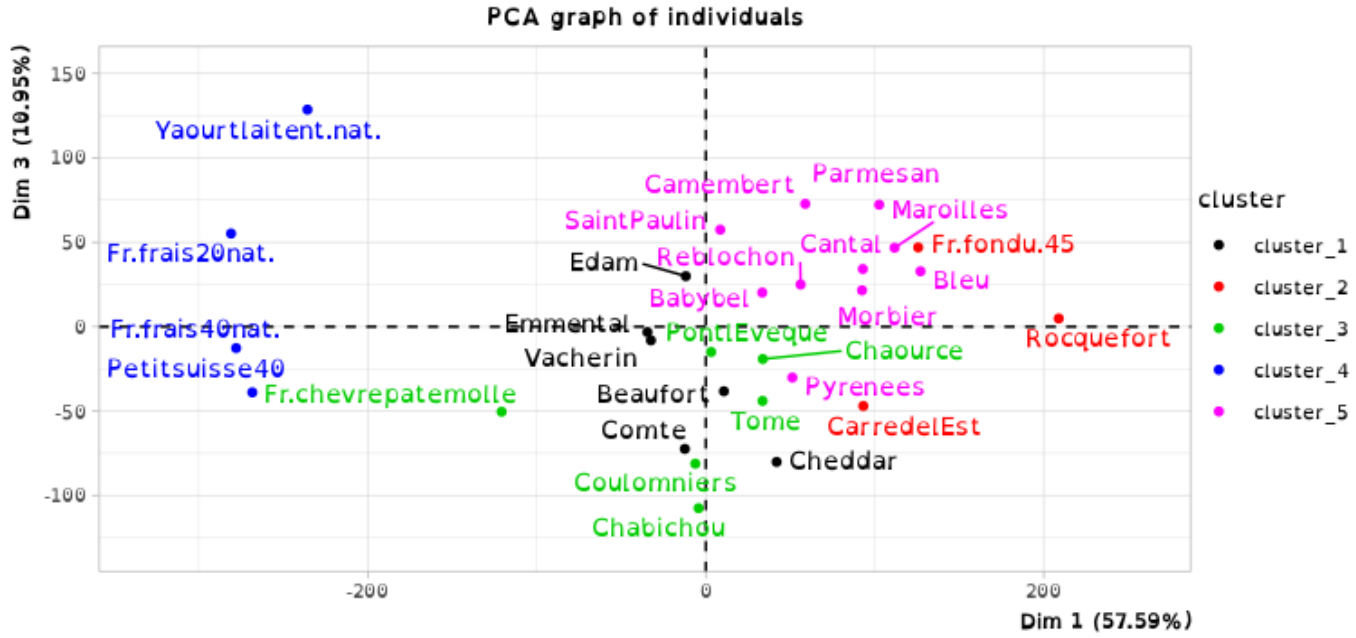
Partitionnement

Les individus sont mieux répartis pour visuellement voir les clusters avec le `k-mean` brute.

`K-mean` sur les données brute :

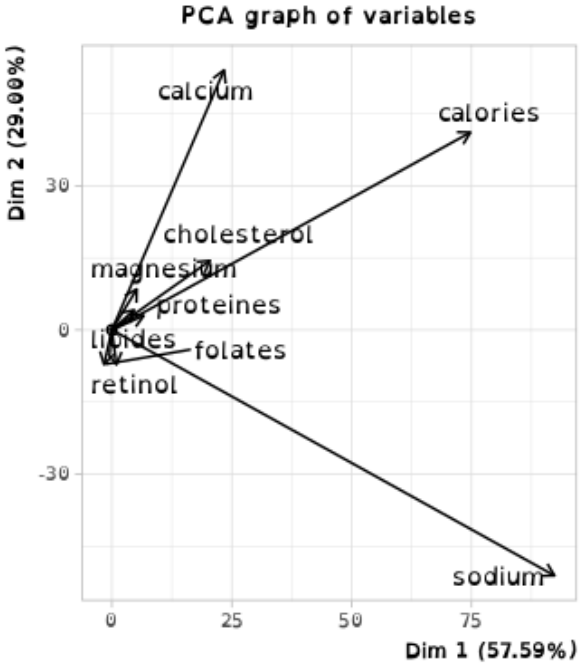


`K-mean` sur les données centrées-réduites :

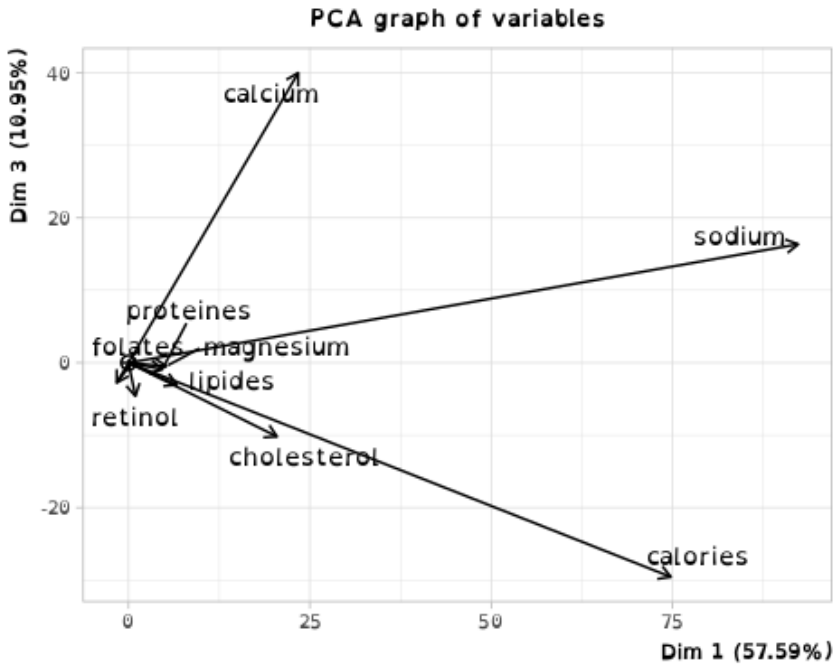


On distingue mieux les variables avec le K-mean centrées-réduites.

K-mean sur les données brute :



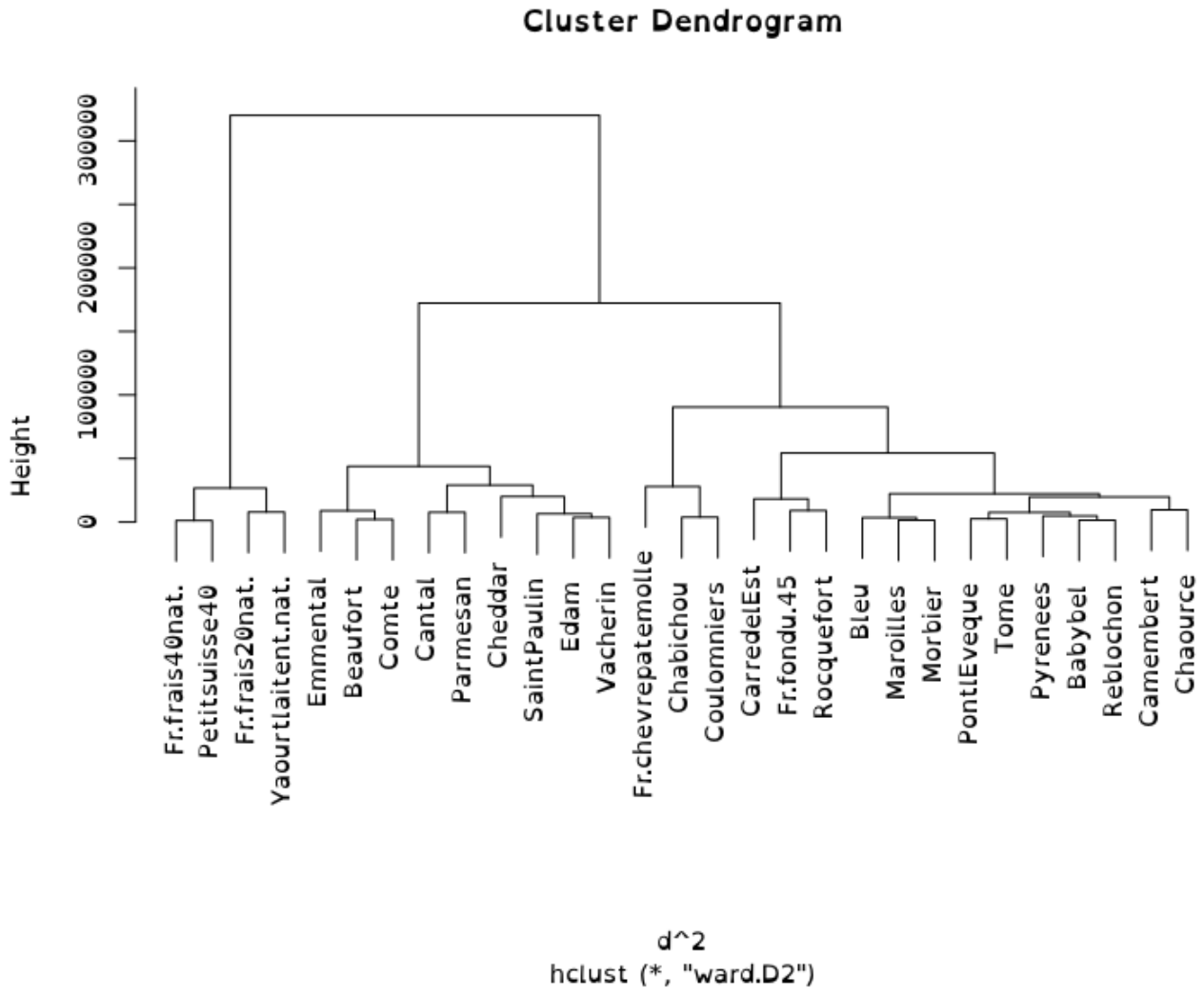
K-mean sur les données centrées-réduites :



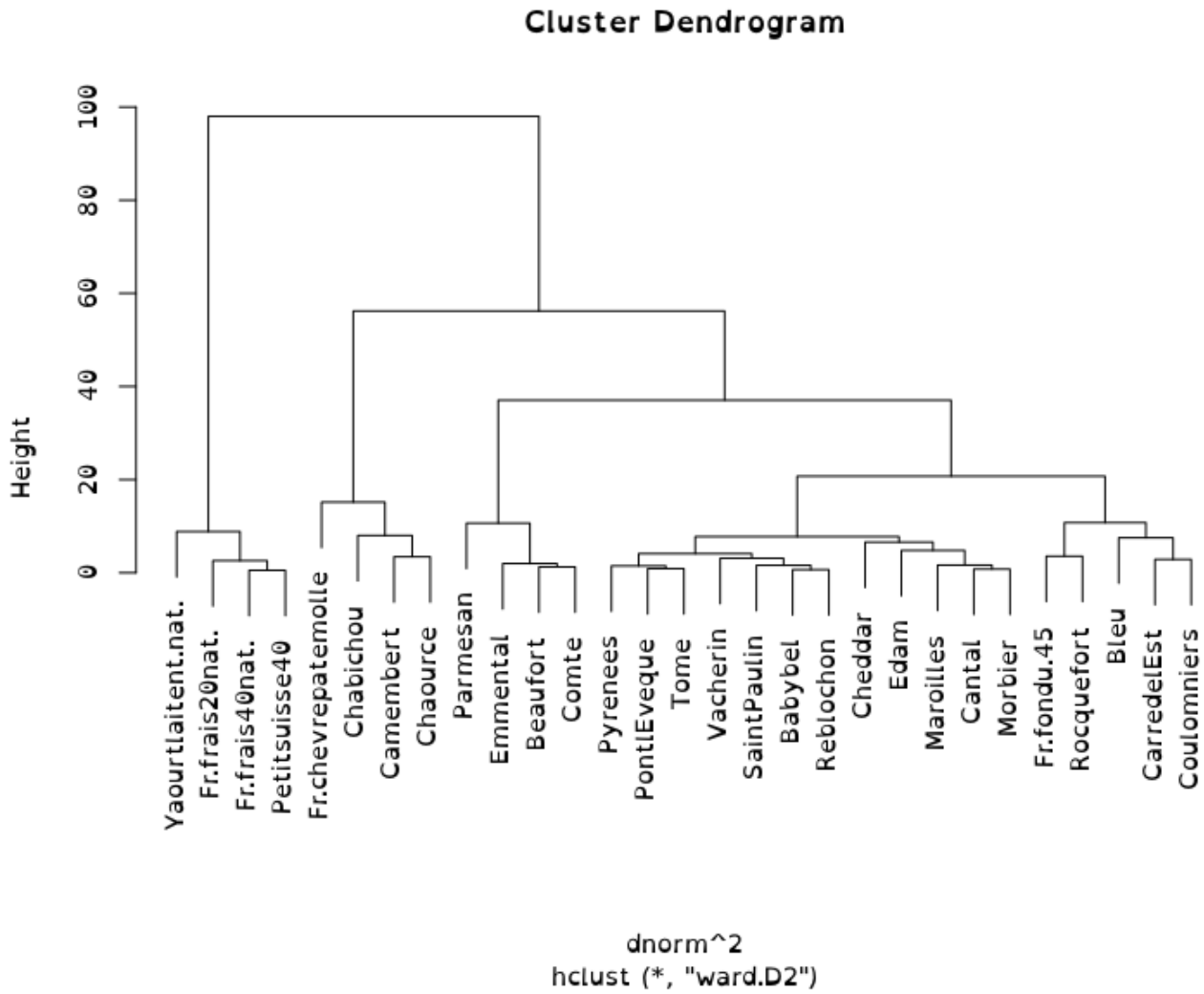
Partitionnement hiérarchique (distance de Ward)

Le découpage est plus distinct avec les données centrées-réduites.

Classification hiérarchique de Ward sur les données brutes :

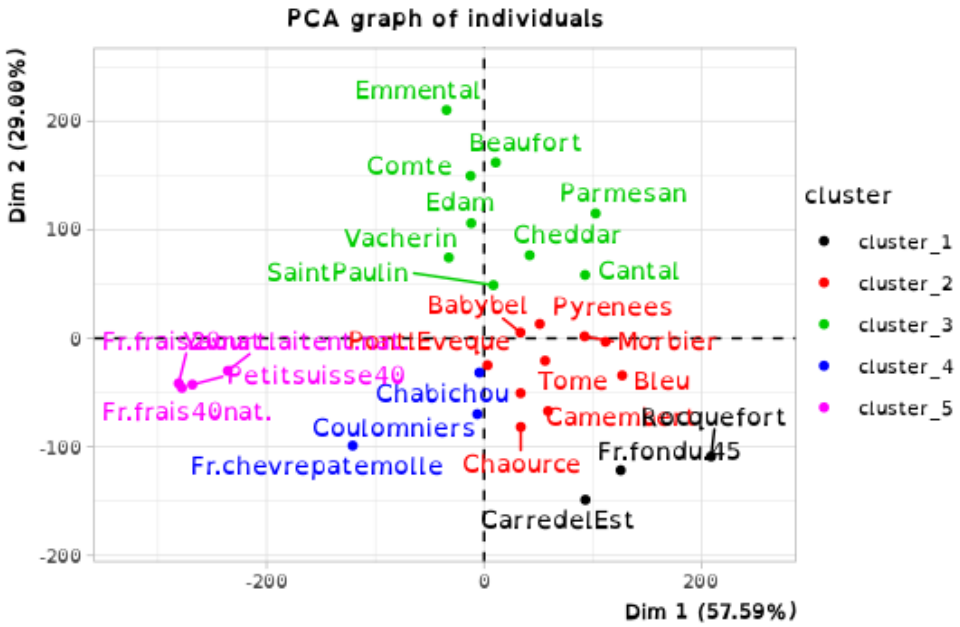


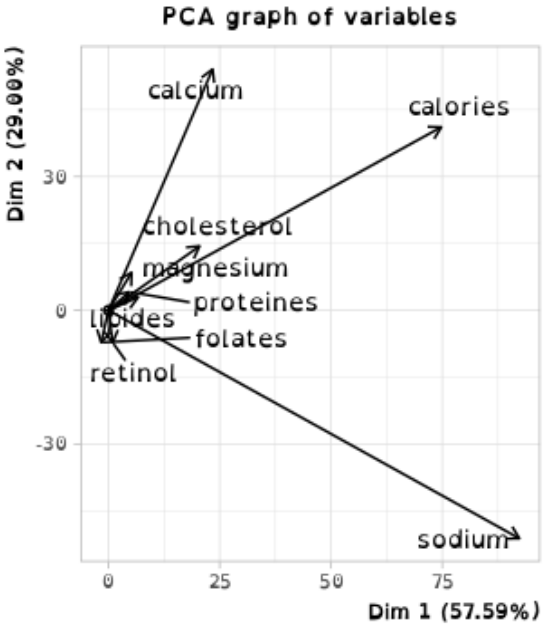
Classification hiérarchique de Ward sur les données centrées-réduites :

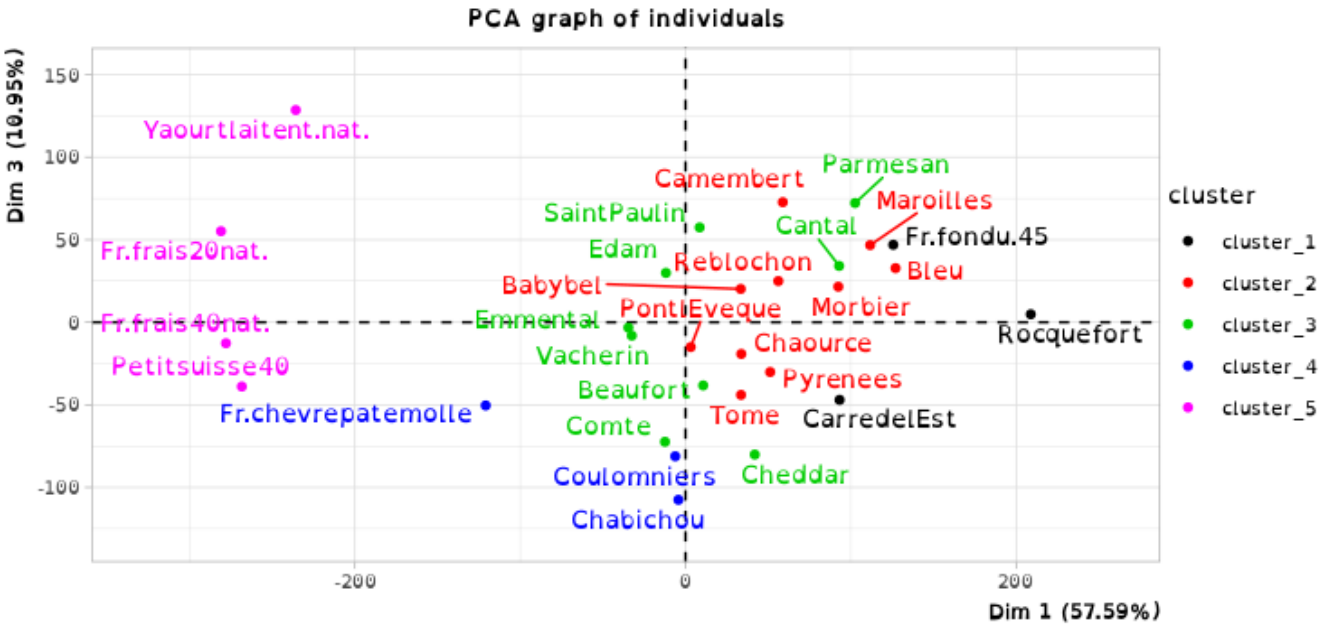


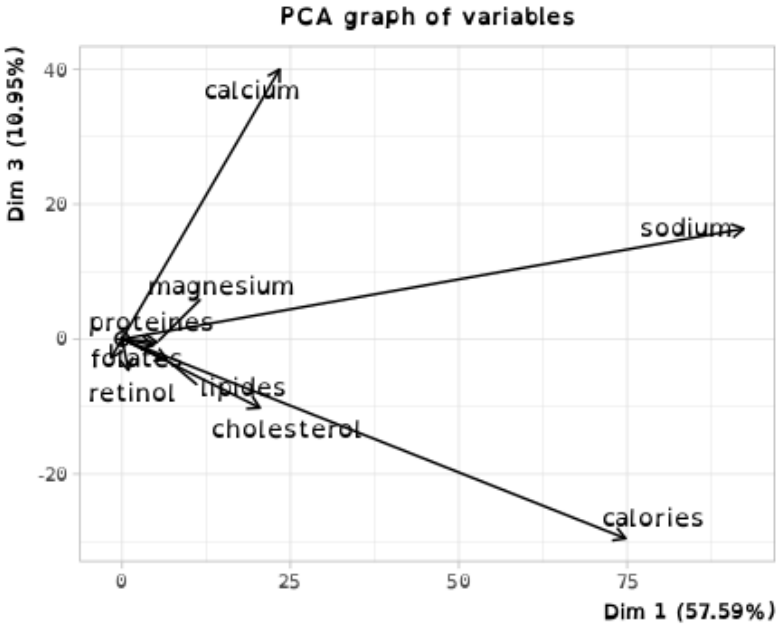
Avec la `distance de Ward` sur les données centrées-réduites, on voit que les clusters sont mieux répartis et définis.

`distance de Ward` sur les données brute :

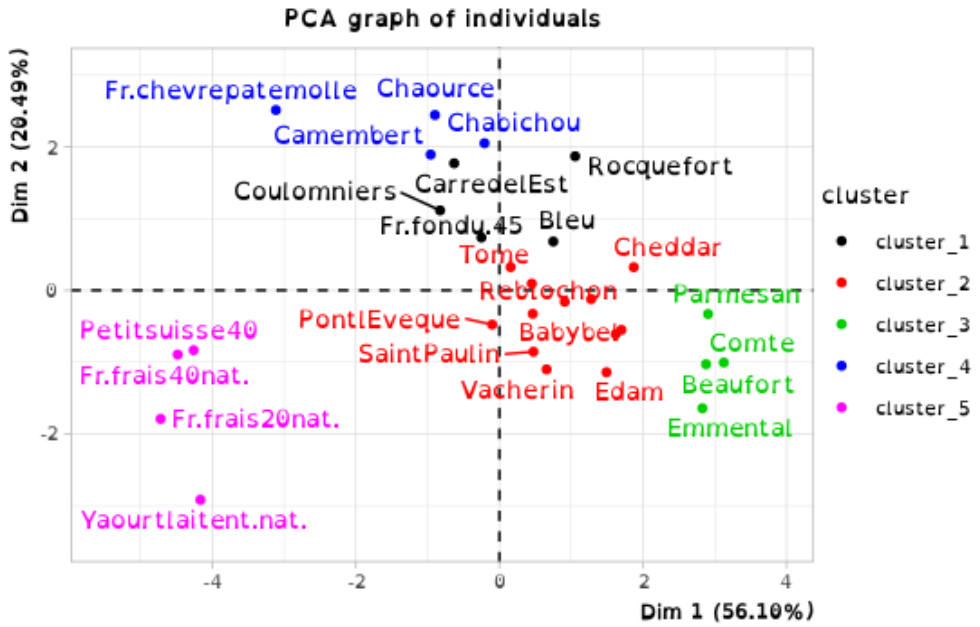

PCA graph of individuals
Dim 2 (29.00%)
Dim 1 (57.59%)
cluster
cluster_1
cluster_2
cluster_3
cluster_4
cluster_5
Emmental
Beaufort
Comte
Parmesan
Edam
Vacherin
Cheddar
SaintPaulin
Cantal
Babybel
Pyrenees
Morbier
Petitsuisse40
Tome
Bleu
Chabichou
Fr.frais40nat.
Coulomniers
Roquefort
Fr.fondu45
Fr.chevrepatemolle
Chaource
CarredelEst

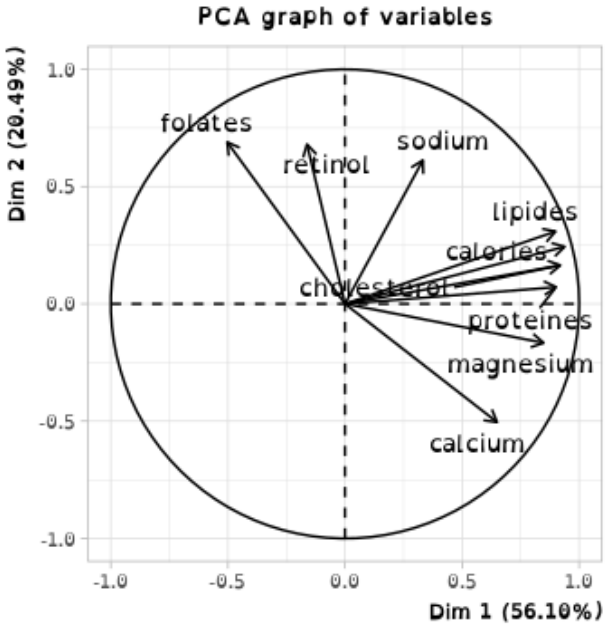

PCA graph of variables
Dim 2 (29.00%)
Dim 1 (57.59%)
calcium
calories
cholesterol
magnesium
proteines
lipides
folates
retinol
sodium

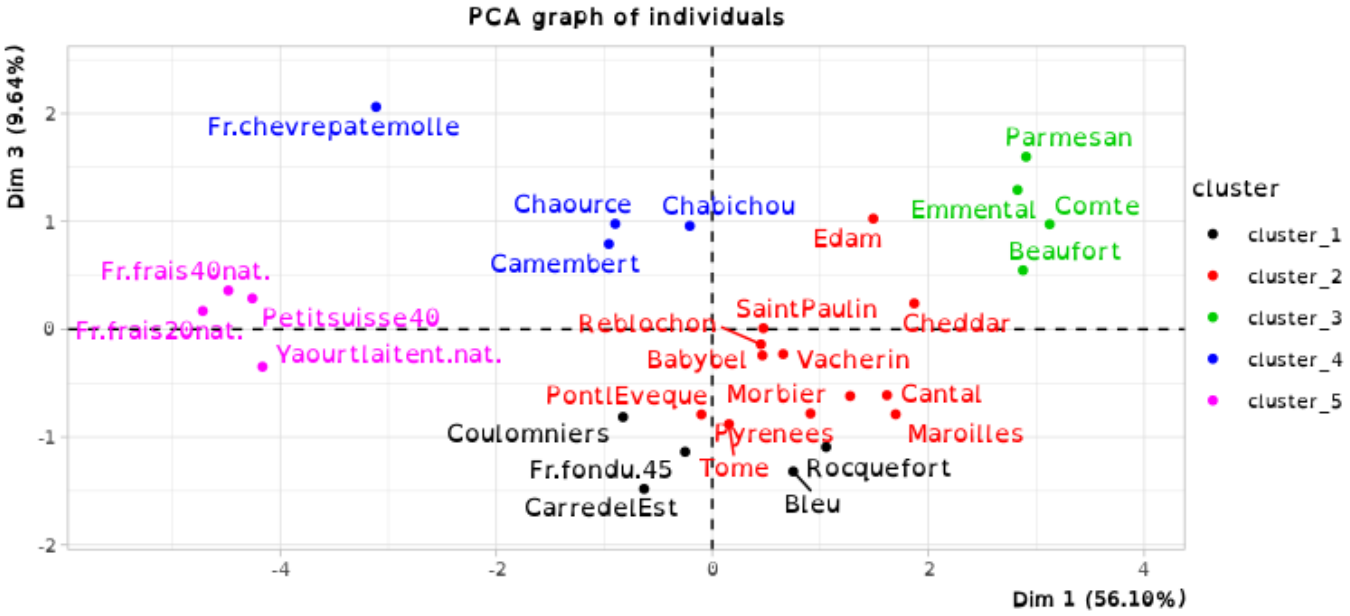


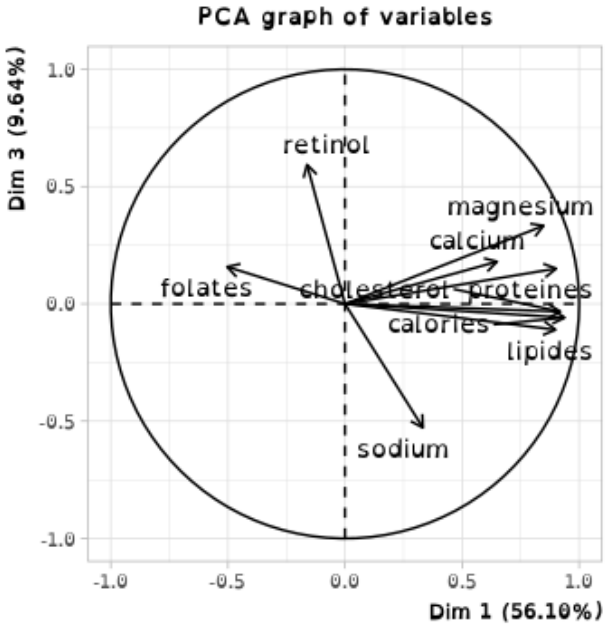


`distance de Ward` sur les données centrées-réduites :


PCA graph of individuals
Dim 2 (20.49%)
Dim 1 (56.10%)
Fr.chevrepatemolle
Chaource
Camembert
Chabichou
Rocquefort
CarredelEst
Coulomniers
Fr.fondu.45
Bleu
Cheddar
Tome
Reblochon
Parmesan
PontlEveque
Babybel
Comte
SaintPaulin
Petitsuisse40
Fr.frais40nat.
Vacherin
Edam
Beaufort
Fr.frais20nat.
Emmental
Yaourtlaitent.nat.
cluster
cluster_1
cluster_2
cluster_3
cluster_4
cluster_5


PCA graph of variables
Dim 2 (20.49%)
Dim 1 (56.10%)
folates
sodium
retinol
lipides
calories
cholesterol
proteines
magnesium
calcium


PCA graph of individuals
Dim 3 (9.64%)
Dim 1 (56.10%)
cluster
cluster_1
cluster_2
cluster_3
cluster_4
cluster_5
Fr.chevrepatemolle
Parmesan
Chaource
Chabichou
Emmental
Comte
Edam
Beaufort
Camembert
Fr.frais40nat.
Petitsuisse40
Fr.frais20nat.
Yaourtlaitent.nat.
Reblochon
SaintPaulin
Cheddar
Babybel
Vacherin
PontlEveque
Morbier
Cantal
Coulomniers
Pyrenees
Maroilles
Fr.fondu.45
Tome
Rocquefort
CarredelEst
Bleu


PCA graph of variables
Dim 3 (9.64%)
Dim 1 (56.10%)
retinol
magnesium
calcium
folates
cholesterol
proteines
calories
lipides
sodium