

DATA FROMAGES

1/ Compréhension des métiers:

- Objectif Métier (Détecter les caractéristiques):

Classification des Fromages en Fonction de Leurs Caractéristiques Nutritionnelles

L'objectif principal est de classer 29 fromages en catégories distinctes en utilisant leurs caractéristiques nutritionnelles.

- Objectif Data Mining:

L'objectif du data mining est d'extraire à partir des données des fromages des relations nutritionnelles entre les fromages et créer des clusters qui nous aideront dans la classification des fromages.

2/ Compréhension des données:

On a deux possibilités:

- *Éliminer les données qualitatives (les noms des fromages)*
- *Spécifier la première colonne en tant que nom des lignes*

	calories	sodium	calcium	lipides	retinol	folates	proteines	cholesterol	magnesium
CarreleEst	314	353.5	72.6	26.3	51.6	30.3	21.0	70	20
Babybel	314	238.0	209.8	25.1	63.7	6.4	22.6	70	27
Beaufort	401	112.0	259.4	33.3	54.9	1.2	26.6	120	41
Bleu	342	336.0	211.1	28.9	37.1	27.5	20.2	90	27
Camembert	264	314.0	215.9	19.5	103.0	36.4	23.4	60	20
Cantal	367	256.0	264.0	28.8	48.8	5.7	23.0	90	30
Chabichou	344	192.0	87.2	27.9	90.1	36.3	19.5	80	36
Chausource	292	276.0	132.9	25.4	116.4	32.5	17.8	70	25
Cheddar	406	172.0	182.3	32.5	76.4	4.9	26.0	110	28
Comte	399	92.0	220.5	32.4	55.9	1.3	29.2	120	51
Coulommiers	308	222.0	79.2	25.6	63.6	21.1	20.5	80	13
Edam	327	148.0	272.2	24.7	65.7	5.5	24.7	80	44
Emmental	378	60.0	308.2	29.4	56.3	2.4	29.4	110	45
Fr.chevrepate molle	206	160.0	72.8	18.5	150.5	31.0	11.1	50	16
Fr.fondu.45	292	390.0	168.5	24.0	77.4	5.5	16.8	70	20
Fr.frais20nat.	80	41.0	146.3	3.5	50.0	20.0	8.3	10	11
Fr.frais40nat.	115	25.0	94.8	7.8	64.3	22.6	7.0	30	10
Maroilles	338	311.0	236.7	29.1	46.7	3.6	20.4	90	40
Morbier	347	285.0	219.0	29.5	57.6	5.8	23.6	80	30

```
> #génère un résumé statistique des variables contenues dans fromages
> summary(fromages)
  calories      sodium      calcium      lipides      retinol      folates      proteines
Min.   : 70      Min.   : 22.0      Min.   : 72.6      Min.   : 3.40      Min.   : 37.10      Min.   : 1.20      Min.   : 4.10
1st Qu.:292      1st Qu.:140.0      1st Qu.:132.9      1st Qu.:23.40      1st Qu.: 51.60      1st Qu.: 4.90      1st Qu.:17.80
Median :321      Median :223.0      Median :202.3      Median :26.30      Median : 62.30      Median : 6.40      Median :21.00
Mean   :300      Mean   :210.1      Mean   :185.7      Mean   :24.16      Mean   : 67.56      Mean   :13.01      Mean   :20.17
3rd Qu.:355      3rd Qu.:276.0      3rd Qu.:220.5      3rd Qu.:29.10      3rd Qu.: 76.40      3rd Qu.:21.10      3rd Qu.:23.40
Max.   :406      Max.   :432.0      Max.   :334.6      Max.   :33.30      Max.   :150.50      Max.   :36.40      Max.   :35.70

  cholesterol      magnesium
Min.   :10.00      Min.   :10.00
1st Qu.: 70.00      1st Qu.:20.00
Median : 80.00      Median :26.00
Mean   : 74.59      Mean   :26.97
3rd Qu.: 90.00      3rd Qu.:30.00
Max.   :120.00      Max.   :51.00

> #renvoie les noms des colonnes (variables) de fromages
> names(fromages)
[1] "calories" "sodium" "calcium" "lipides" "retinol" "folates" "proteines" "cholesterol"
[9] "magnesium"

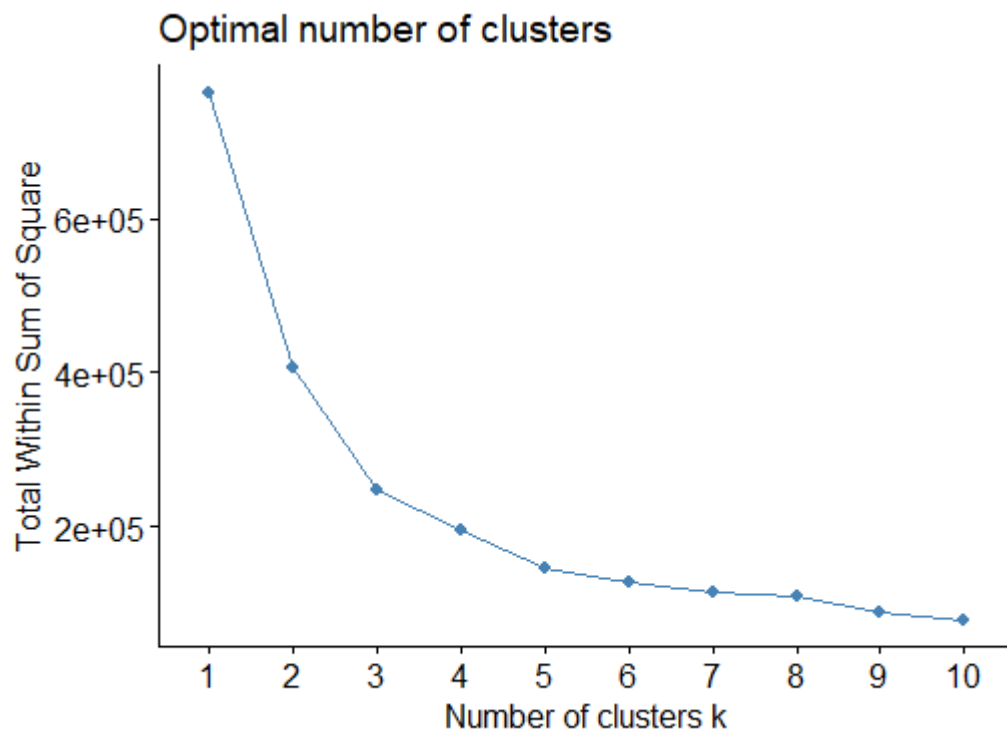
> #affiche la structure de fromages
> str(fromages)
'data.frame': 29 obs. of 9 variables:
 $ calories : int 314 314 401 342 264 367 344 292 406 399 ...
 $ sodium : num 354 238 112 336 314 ...
 $ calcium : num 72.6 209.8 259.4 211.1 215.9 ...
 $ lipides : num 26.3 25.1 33.3 28.9 19.5 28.8 27.9 25.4 32.5 32.4 ...
 $ retinol : num 51.6 63.7 54.9 37.1 103 ...
 $ folates : num 30.3 6.4 1.2 27.5 36.4 5.7 36.3 32.5 4.9 1.3 ...
 $ proteines : num 21 22.6 26.6 20.2 23.4 23 19.5 17.8 26 29.2 ...
 $ cholesterol : int 70 70 120 90 60 90 80 70 110 120 ...
 $ magnesium : int 20 27 41 27 20 30 36 25 28 51 ...

> dim(fromages)
[1] 29 9
```

3/Préparation des données:

K-MEANS

Choix du nombre des clusters: (on remarque que la stabilisation de l' "elbow" entre 3 et 4 (donc on va choisir soit 3 ou 4 clusters)



utiliser K-means pour la segmentation des données (4 clusters)

```

R 4.3.1 ~ /
> fviz_nbclust(fromages.seg, kmeans, method="wss")
> k<-kmeans(fromages.seg, 4) #4 c'est le nombre de classes
> k
K-means clustering with 4 clusters of sizes 8, 8, 4, 9

Cluster means:
  calories  sodium  calcium  lipides  retinol  folates  proteines  cholesterol  magnesium
1 322.0000 336.6875 186.0125 26.63750 66.25000 16.312500 20.47500 80.00000 26.50000
2 305.0000 224.3750 130.3750 25.15000 81.38750 18.037500 19.60000 73.75000 23.00000
3 101.7500 44.7500 133.7500 6.27500 55.15000 16.475000 7.20000 18.25000 11.25000
4 364.2222 158.3333 257.8000 29.02222 61.95556 4.066667 26.16667 95.55556 37.88889

Clustering vector:
[1] 1 2 4 1 1 4 2 2 4 4 2 4 4 2 1 3 3 1 1 4 3 2 2 1 1 4 2 4 3

Within cluster sum of squares by cluster:
[1] 53718.38 52015.46 18690.22 70549.80
 (between_SS / total_SS = 74.5 %)

Available components:

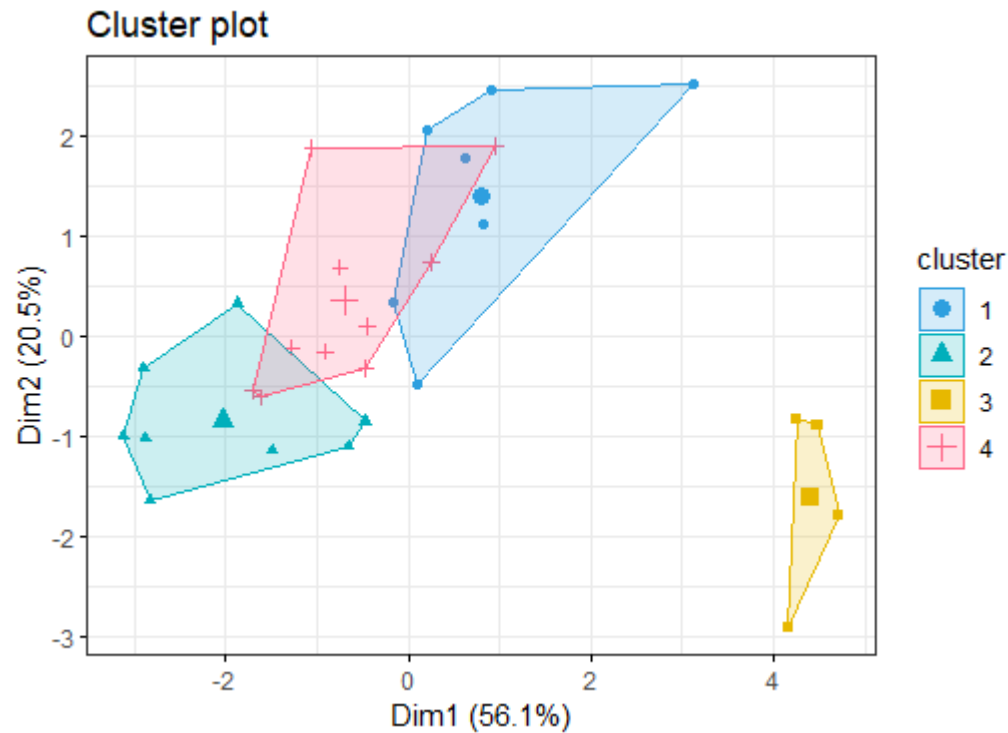
[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
[8] "iter"         "ifault"

> k$centers
  calories  sodium  calcium  lipides  retinol  folates  proteines  cholesterol  magnesium
1 322.0000 336.6875 186.0125 26.63750 66.25000 16.312500 20.47500 80.00000 26.50000
2 305.0000 224.3750 130.3750 25.15000 81.38750 18.037500 19.60000 73.75000 23.00000
3 101.7500 44.7500 133.7500 6.27500 55.15000 16.475000 7.20000 18.25000 11.25000
4 364.2222 158.3333 257.8000 29.02222 61.95556 4.066667 26.16667 95.55556 37.88889

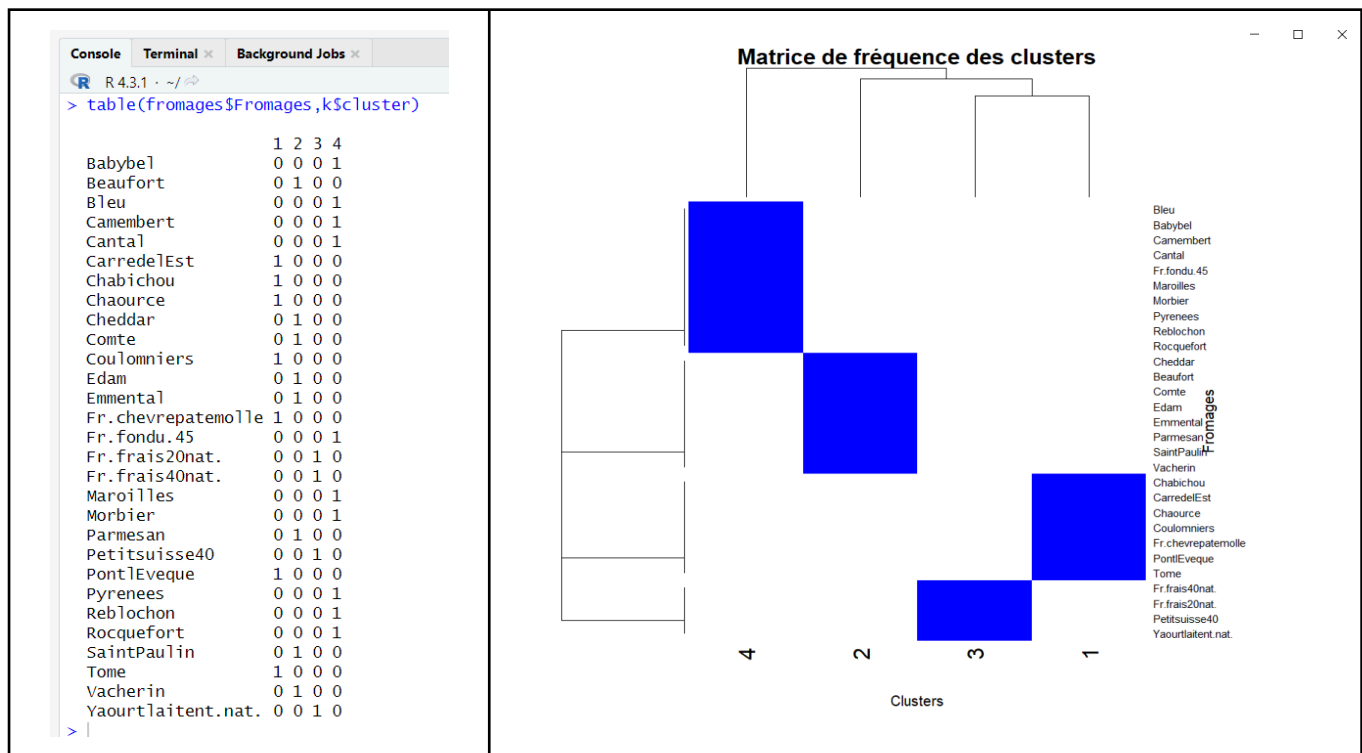
```

4/Modélisation:

Visualisation des clusters choisis



5/Evaluation: tester l'algorithme



6/Déploiement : un dataset labellisé

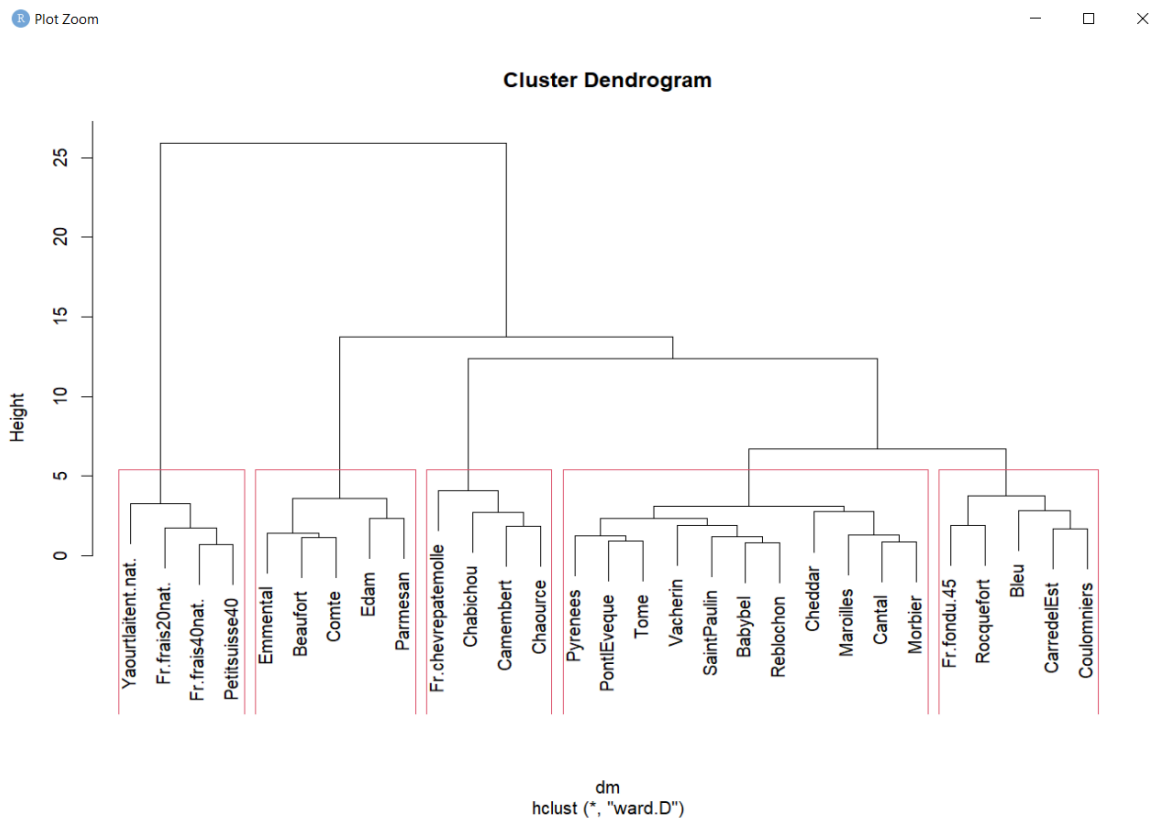
CAH

```

> dm<-dist(fromages.cr)
> dendo<-hclust(dm,method = "ward.D")
> summary(dendo)
      Length Class  Mode
merge      56  -none- numeric
height     28  -none- numeric
order      29  -none- numeric
labels     29  -none- character
method      1  -none- character
call        3  -none- call
dist.method 1  -none- character
> plot(dendo)
> rect.hclust(dendo,k=5)
>

```

On va afficher dendrogramme avec matérialisation des groupes (5 groupes)



A travers cette modélisation, on va classer les fromages sous 5 types:

FROMAGE	TYPE
Fr.frais20nat. Fr.frais40nat. Petitsuis se40 Yaourtlaitent.nat.	Fromages frais et yaourts naturels
Beaufort Comte Emmental Parmesan	Fromages à pâte dure
Camembert Chabichou Chaource Fr.chevrepatemolle	Fromages à pâte molle et de chèvre
Babybel Cantal Cheddar Edam Maroilles Morbier Pontl'Eveque Pyrénées Reblochon SaintPaulin Tome Vacherin	Fromages variés à pâte dure et molle
Fr.fondu.45 Bleu Rocquefort Carre del Es t Coulomniens	Fromages à pâte fondu et fromages bleus

Afin de vérifier notre choix de nombre de classes est optimal on va calculer l'inertie (L'inertie est une mesure de la similarité des clusters)

