

METHODES DE CLUSTERING

zoom sur les kmeans

INTRODUCTION

Définition

Méthode d'analyse exploratoire de données où les **observations** sont séparées en groupe **significatifs** qui partagent des caractéristiques communes

Définition

Regrouper les observations en classes homogènes

- ✓ créer des groupes de personnes (marketing)
- ✓ détection d'anomalies (fraudes, pannes...)



Différence avec la classification

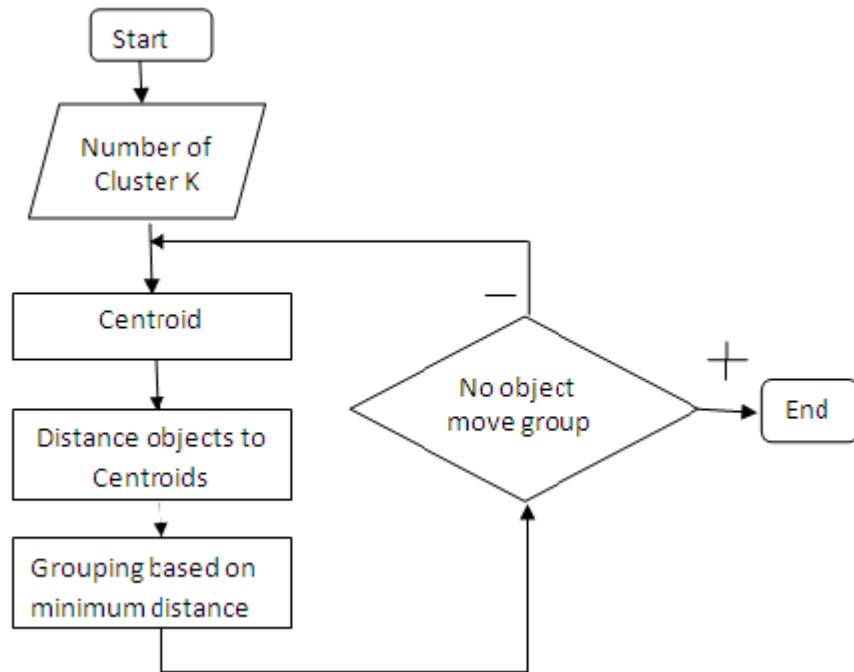
Classification : les classes sont pré-définies

- ✓ cet email est un spam O/N
- ✓ ce commentaire facebook est positif/négatif
- ✓ reconnaissance d'image



LE FONCTIONNEMENT DES KMEANS

L'algorithme des kmeans



Step 1 :

On choisit k éléments aléatoires dans le plan – ce sont les « centres » des clusters

Step 2 :

on affecte chaque individu à un cluster

Step 3 :

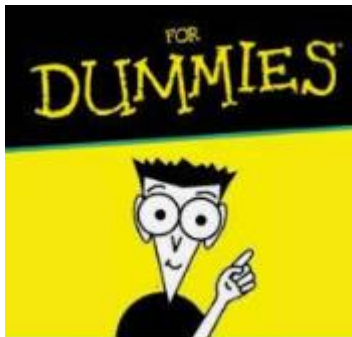
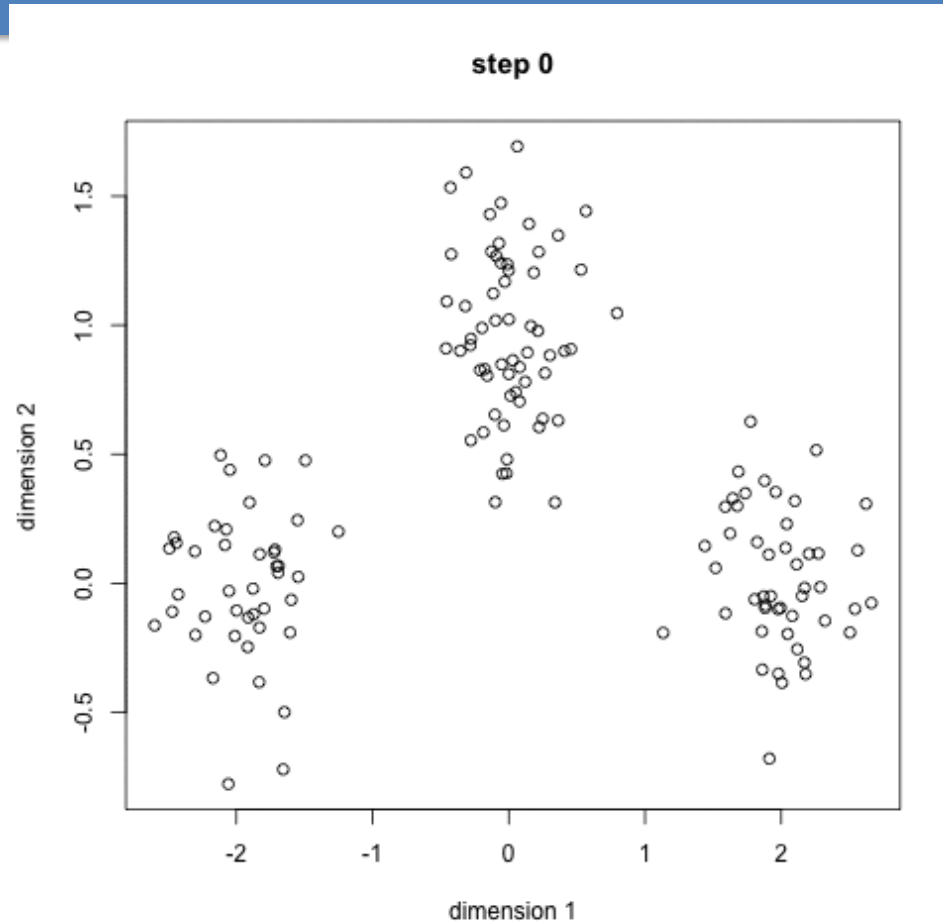
Chaque groupe constitué permet de recalculer un élément « centre » (centre de gravité)

Step 4 :

On réaffecte les individus aux nouveaux centres définis au step 3

On itère jusqu'à ce que les groupe d'individus soient stables

En image



On choisit un nombre de clusters et
l'algorithme fait le boulot

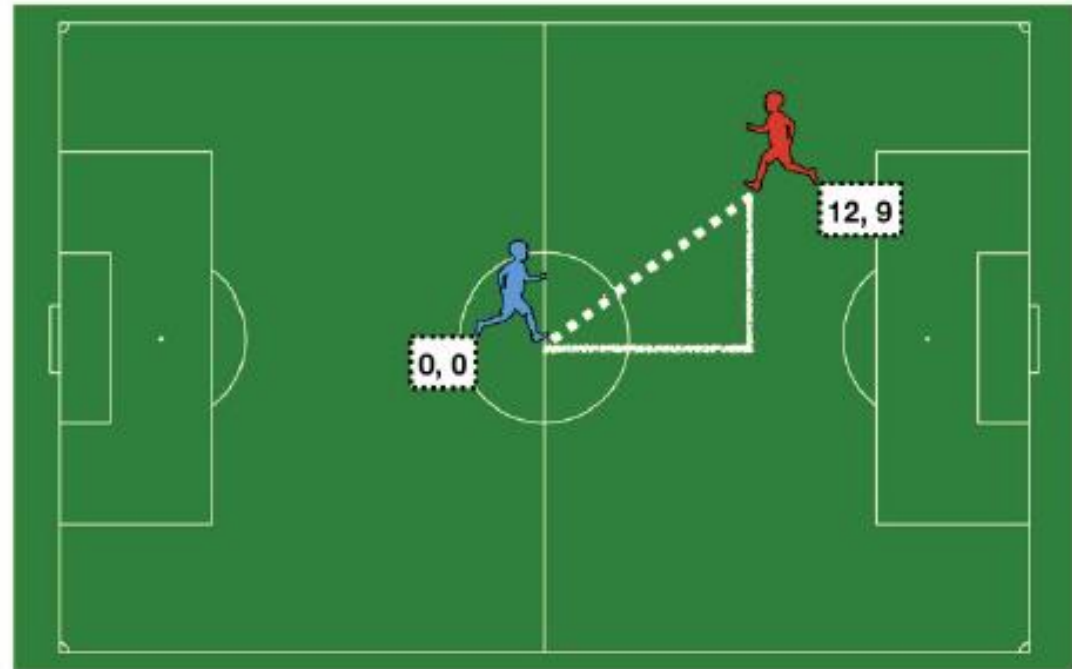
UN PEU DE THEORIE

Le groupe le plus proche ?

L'algorithme s'appuie sur la mesure de la distance euclidienne d'un point à son centre le plus proche.

Calcul de la distance

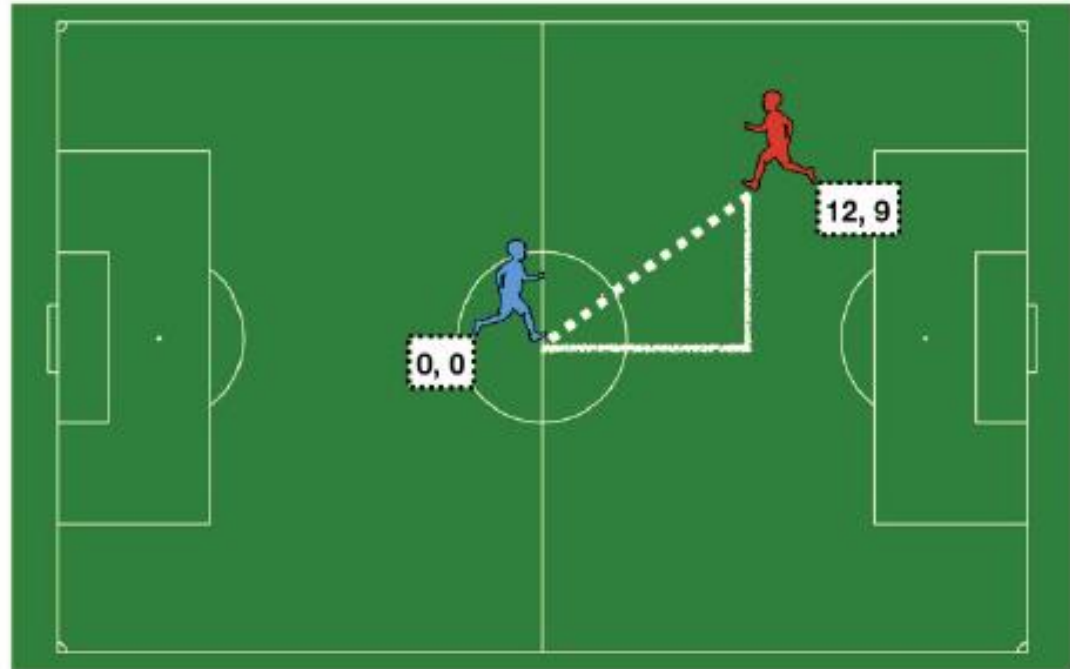
	X	Y
Blue	0	0
Red	12	9



Réponse ?

Calcul de la distance

	X	Y
Blue	0	0
Red	12	9



$$\sqrt{(xRed - xBlue)^2 + (yRed - yBlue)^2}$$
$$=15$$

Minimiser la variance intra

L'objectif d'affecter un individu à son « centre » le plus proche revient à minimiser la variance intra du centroïde.

Ce point sera utile pour mesurer la performance du modèle.

CHOIX DE K ET PERFORMANCE DU CLUSTERING

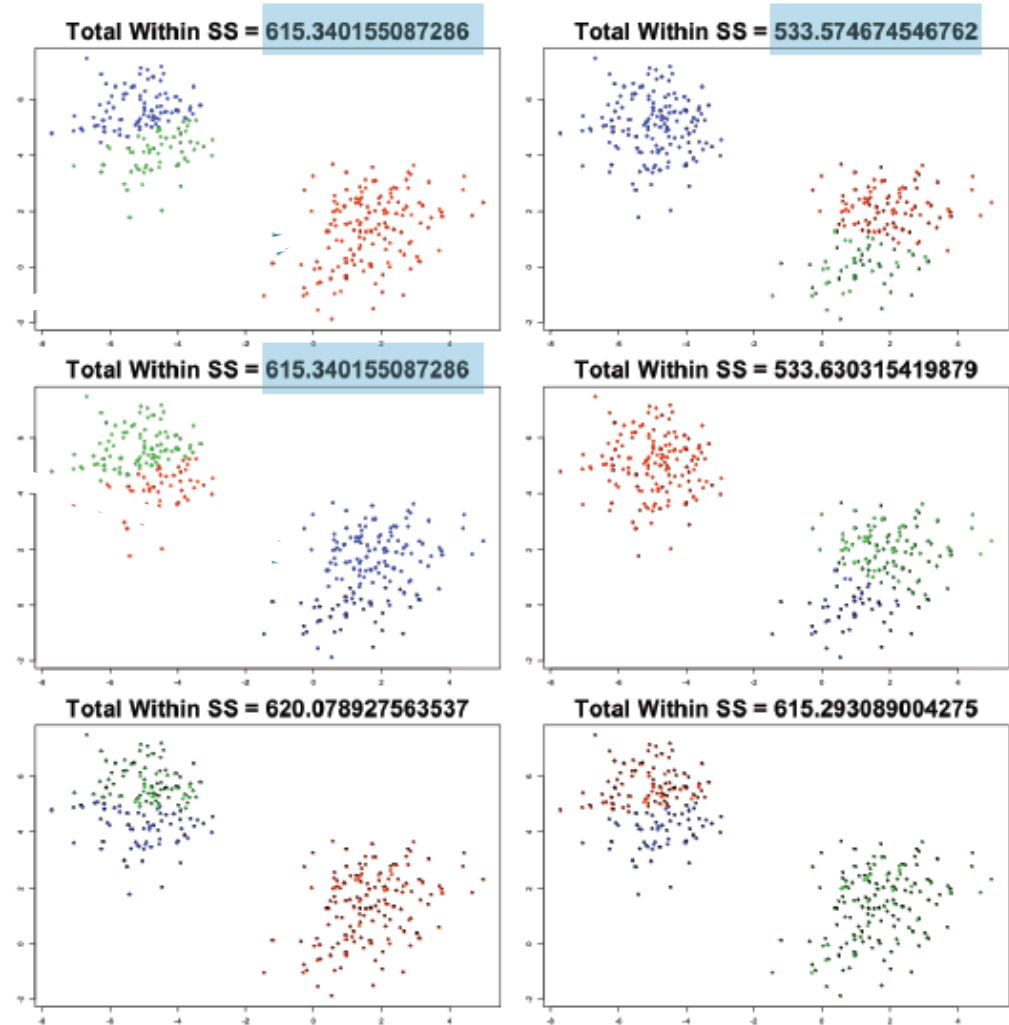
Variance intra (*total within SS*)

Une mesure de la performance du modèle est la variance intra totale, *ie* la somme des variances intra de chaque cluster.

Plus cette variance est faible, meilleur est le modèle

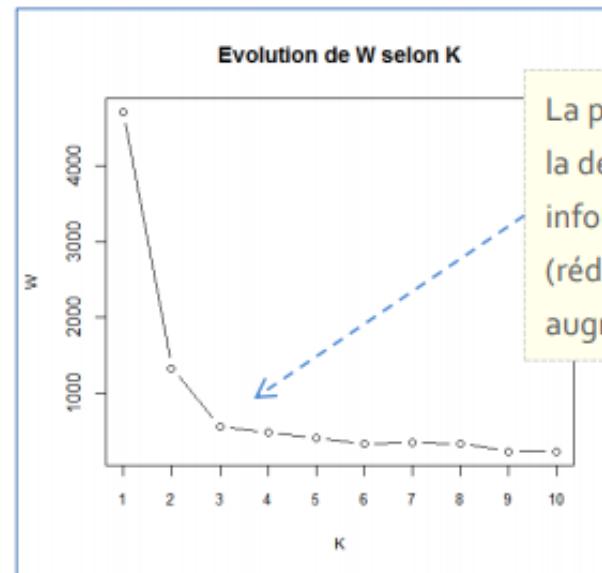
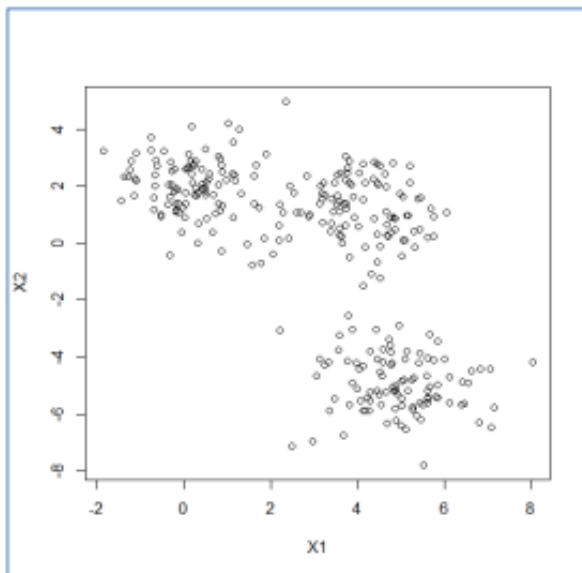
Variance intra (*total within SS*)

Pour rappel, le choix des clusters initiaux étant aléatoire, il faut procéder à plusieurs exécutions pour trouver le modèle qui minimise la *Vintra*



Choix du nombre de classes k

Principe : Une stratégie simple pour identifier le nombre de classes consiste à faire varier K et surveiller l'évolution de l'inertie intra-classes W . L'idée est de visualiser le « coude » où l'adjonction d'une classe ne correspond à rien dans la structuration des données.



La partition en $K = 3$ classes est la dernière à induire un gain informationnel significatif (réduction inertie intra → augmentation de l'inertie inter)

Méthode de la silhouette

Permet de vérifier la pertinence de l'affectation d'un individu à une classe en calculant :

- sa distance moyenne aux individus de sa classe (C)
- sa distance moyenne aux individus de la classe la plus proche (N)

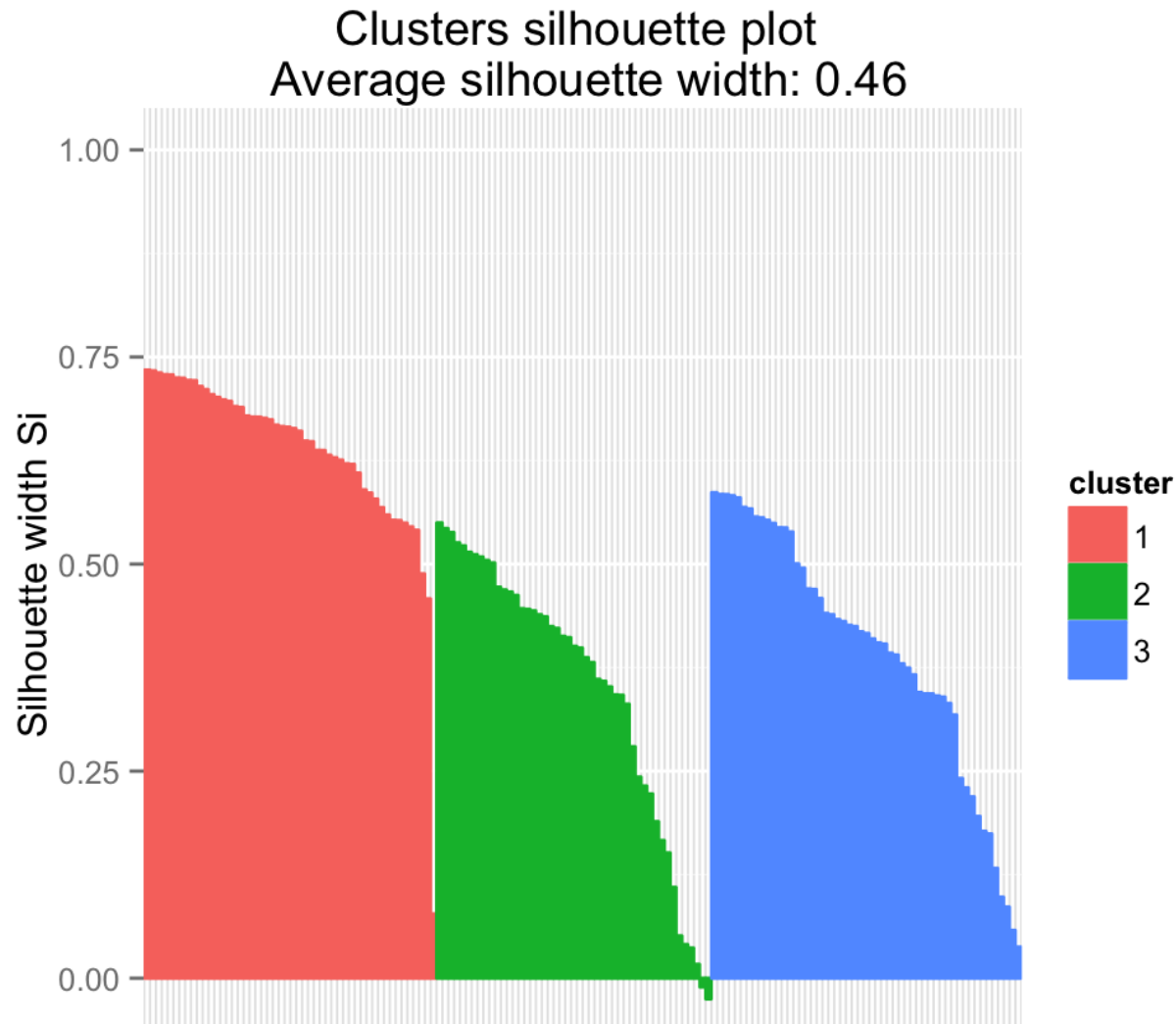
$$s(i) = \begin{cases} 1 - C(i)/N(i), & \text{if } C(i) < N(i) \\ 0, & \text{if } C(i) = N(i) \\ N(i)/C(i) - 1, & \text{if } C(i) > N(i) \end{cases}$$

Plus s est proche de 1, plus l'affectation est bonne

Si s est proche de 0, l'individu est à la frontière de deux clusters

Si s est <0, l'individu est mal classé

Méthode de la silhouette



Méthode de la silhouette

La moyenne des silhouettes des individus permet de définir une métrique de performance :

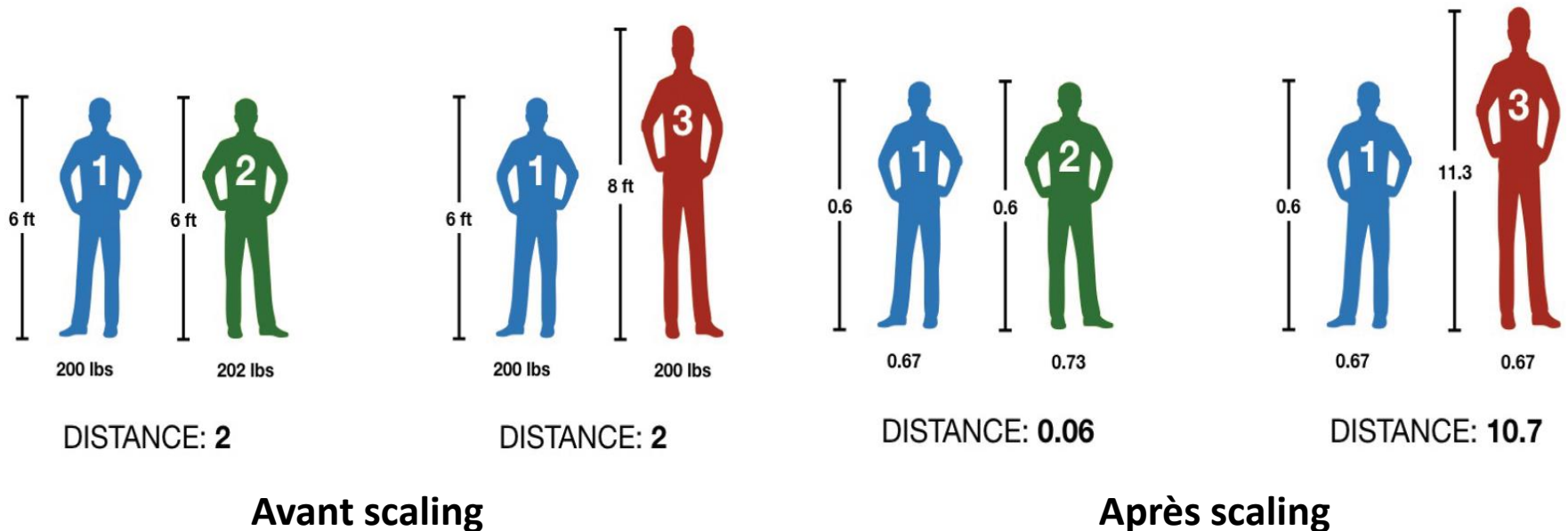
- plus cette valeur est proche de 1 meilleur est le modèle

On peut ainsi calculer les silhouettes pour $k:1\dots n$ clusters et choisir la valeur de k qui maximise la silhouette moyenne

AMELIORER LA PERFORMANCE

Standardiser les données

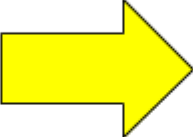
- L'affectation des individus aux clusters repose sur la distance, il est donc important de comparer des variables de même ordre de grandeur



Variables qualitatives

- Il est possible d'utiliser des variables qualitatives en les recodant

Color	Red	Red	Yellow	Green	Yellow
-------	-----	-----	--------	-------	--------



Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

The diagram illustrates the recoding of a qualitative variable 'Color' into three binary variables: 'Red', 'Yellow', and 'Green'. The first table shows the original data with five rows of color values. A yellow arrow points to the second table, which shows the same data recoded into a 3x5 grid of binary values (0 or 1). The first row of the recoded table corresponds to the first row of the original table, and so on. The last row of the recoded table is empty.

ACP

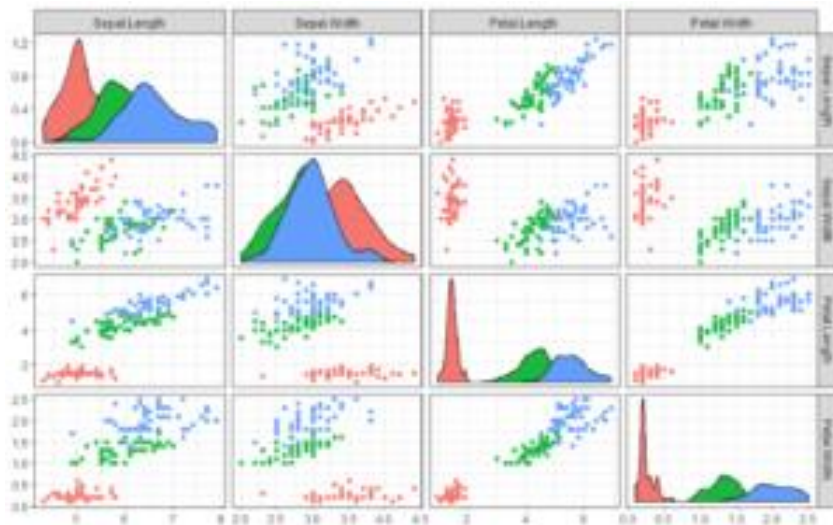
- Dans le cas de dataset avec un nombre important de features, il peut être intéressant de procéder à une ACP pour réduire le nombre de dimensions.
- On réalisera le clustering sur les coordonnées des individus sur chaque composante principale

DONNER DU SENS AU CLUSTERING

Analyser les caractéristiques des clusters

Pour chaque cluster

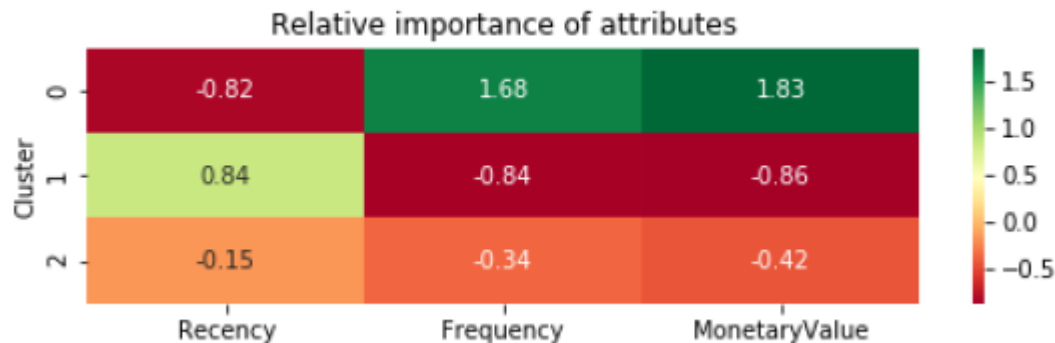
- ✓ sortir les statistiques descriptives
- ✓ étudier les différences entre les clusters



Analyser les caractéristiques des clusters

Identifier les variables marquantes de chaque cluster

Heatmap plot:



Écart relatif de la moyenne du cluster à la moyenne de la population

CRITIQUE DE L'ALGORITHME

Inconvénients de l'algorithme


1. Le nombre de classe doit être fixé au départ
2. Le résultat dépend du tirage initial des centres des classes
3. label des classes pas stables d'une exécution à l'autre

APPROCHE METHODOLOGIQUE

Ma première k-means



Choix d'un dataset complet

- 
- Choix et transformation des features (*le fameux feature engineering*)
 - Lancement de l'algorithme de clustering
 - Evaluation des métriques du modèle

-> process itératif

En approche ML : on déploie le modèle appris sur un nouveau dataset

LET'S CODE

R

```
irisCluster <- stats::kmeans(iris[, 3:4], 3,  
nstart = 20)
```

```
PROC FASTCLUS <MAXCLUSTERS= n> <RADIUS= t>  
<options>;  
VAR variables;  
ID variables;  
FREQ variable;  
WEIGHT variable;  
BY variables;
```

Python

```
from sklearn.cluster import Kmeans  
kmeans = KMeans(n_clusters=4)  
kmeans.fit(X)  
y_kmeans = kmeans.predict(X)
```



A LA SEMAINE PROCHAINE