

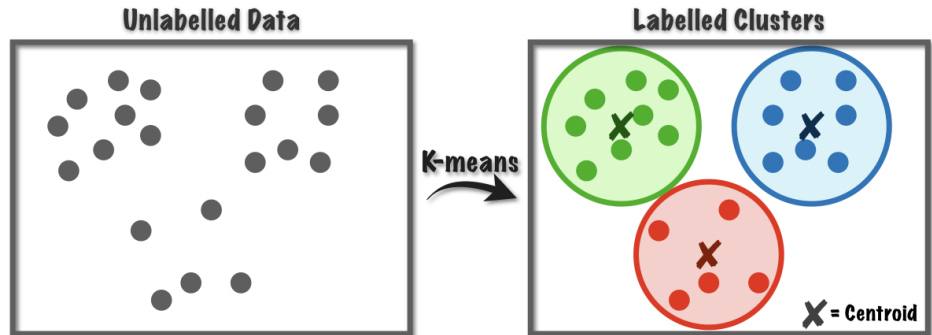
METHODES DE CLUSTERING

INTRODUCTION

Définitions

Technique d'analyse de données qui consiste à regrouper un ensemble de données non étiquetées en plusieurs groupes.

Objectif : trouver des groupes de données similaires (homogénéité intra-classes) et les séparer des autres groupes (hétérogénéité inter-classes)



Cas d'usages

Segmentation de la clientèle : séparer les clients en fonction de leur comportement, de leur équipement...

Détection de fraudes : identification de comportements/transactions suspects dans un cluster isolé

Traitement du langage : création de clusters pour identifier des textes (verbatim) qui peuvent se regrouper

LES ALGORITHMES LES PLUS UTILISÉS

Clustering hierarchique

Structure d'arbre de données (dendrogramme) qui regroupe progressivement les éléments similaires en clusters

Deux types :

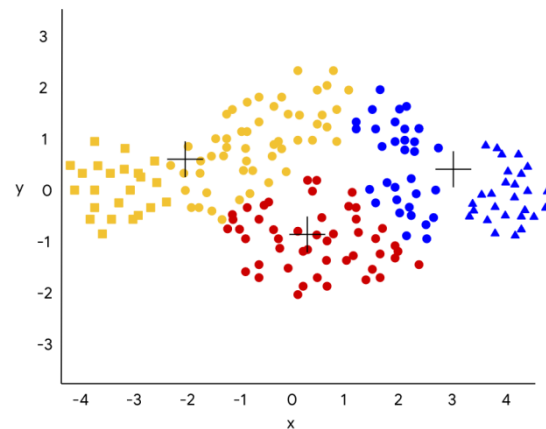
- ❑ Ascendante (agglomerative) : chaque individu est un cluster et on agrège
- ❑ Descendante (divise) : un seul cluster au départ et on divise



Clustering basé sur le centroïde

Kmeans

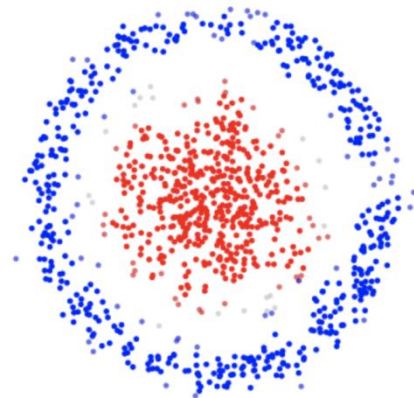
- ❑ divise les données en k-clusters, avec k connu à l'avance.
- ❑ utilise la distance euclidienne pour mesurer la similarité entre les données



Clustering basé sur la densité

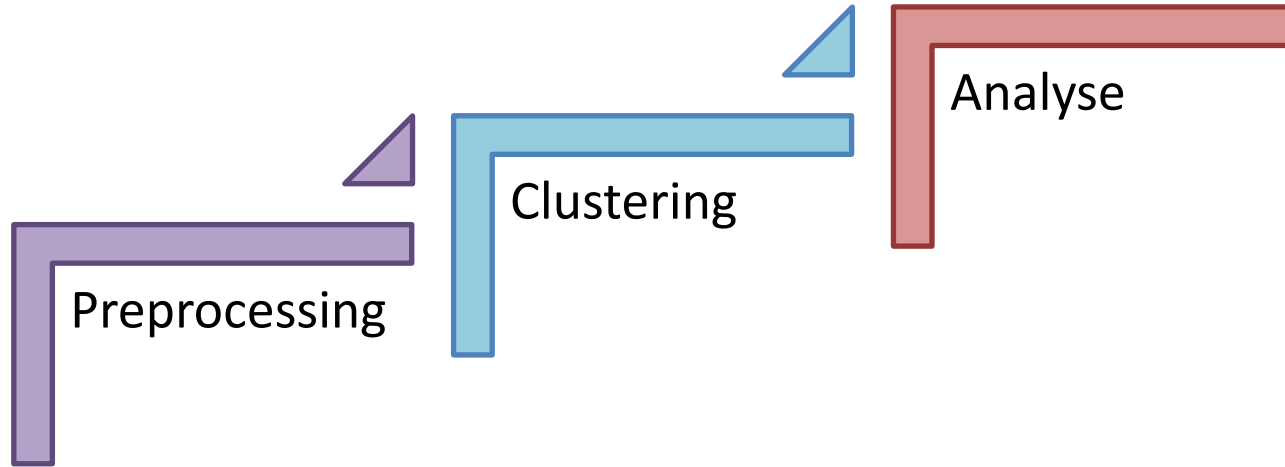
DBScan

- ❑ Identifie de manière arbitraire des clusters dans les données
- ❑ Se base sur la densité de points dans un espace multidimensionnel



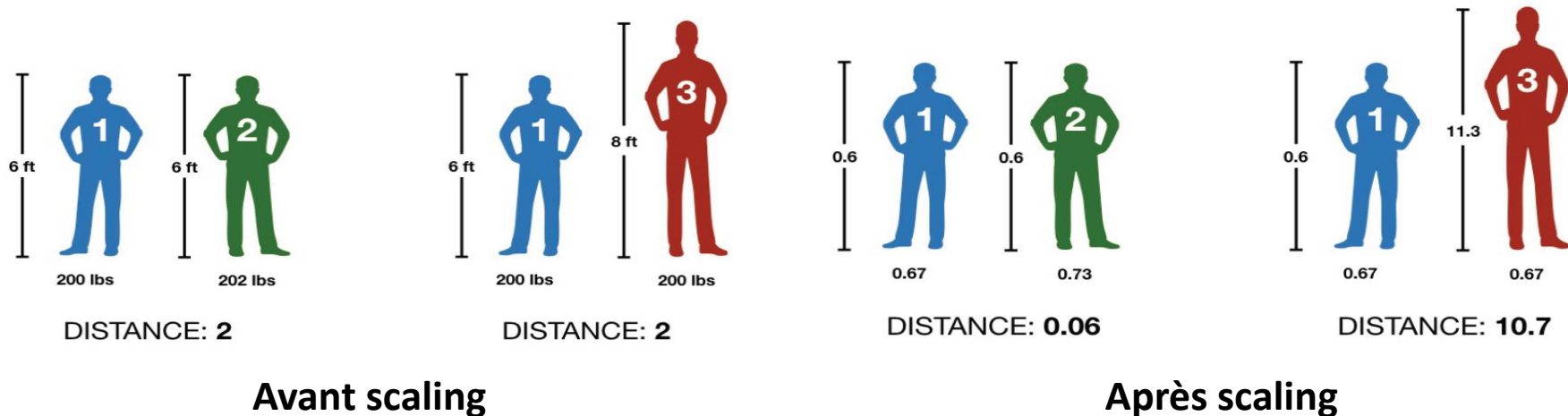
WORKFLOW DE CLUSTERING

APPROCHE STANDARD POUR UN CLUSTERING



Preprocessing - standardiser

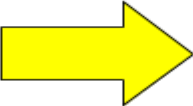
- L'affectation des individus aux clusters repose sur la distance, il est donc important de comparer des variables de même ordre de



Preprocessing – variables catégorielles

Possible d'utiliser des variables qualitatives en les recodant (dummy encoding)

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



Preprocessing - ACP

- Dans le cas de dataset avec un nombre important de variables (features), il est conseillé de réaliser une ACP pour réduire le nombre de dimensions
- On réalisera le clustering sur les coordonnées des individus sur chaque composante principale retenue

ACP - Rappels

Méthode statistique qui permet de :

- décrire et visualiser des données ;
- les décorrélérer ; la nouvelle base est constituée d'axes qui ne sont pas corrélés entre eux ;
- effectuer une réduction de dimension des données d'entraînement

TD - PREPROCESSING

Les attendus

- ☐ Lire le fichier csv dans un dataframe
 - ☐ Donner le nombre de lignes et colonnes du fichier
 - ☐ Sortir des statistiques (mean, med, percentile, null) par variable
 - ☐ Afficher la courbe de densité de la variable overall
 - ☐ Afficher la courbe de densité de la variable weight_kg par Role
-
- ☐ Identifier les variables sans variance
 - ☐ Identifier les variables corrélées
 - ☐ Afficher la matrice de corrélation

Les attendus

- ☐ Réaliser une ACP
- ☐ Afficher le screeplot => combien d'axes on retient ?
- ☐ Afficher les contributions des variables sur les axes retenus
- ☐ Décrire chaque axe
- ☐ Afficher le graph des individus sur les axes 1 et 2

KMEANS CLUSTERING

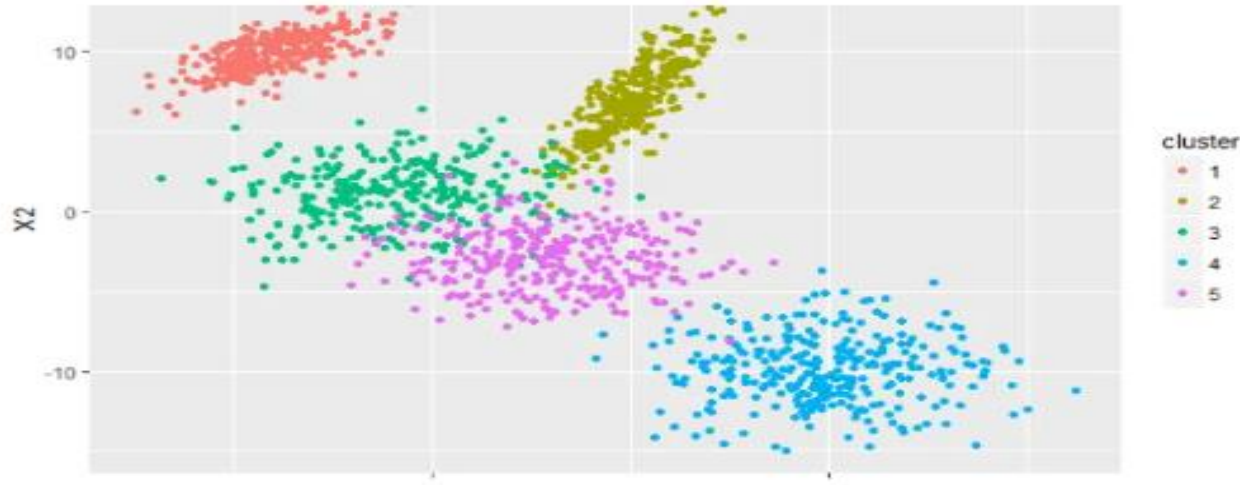
La méthode des kmeans

Méthode **d'apprentissage non supervisé** utilisée sur des **données non labellisées** (sans target connue)

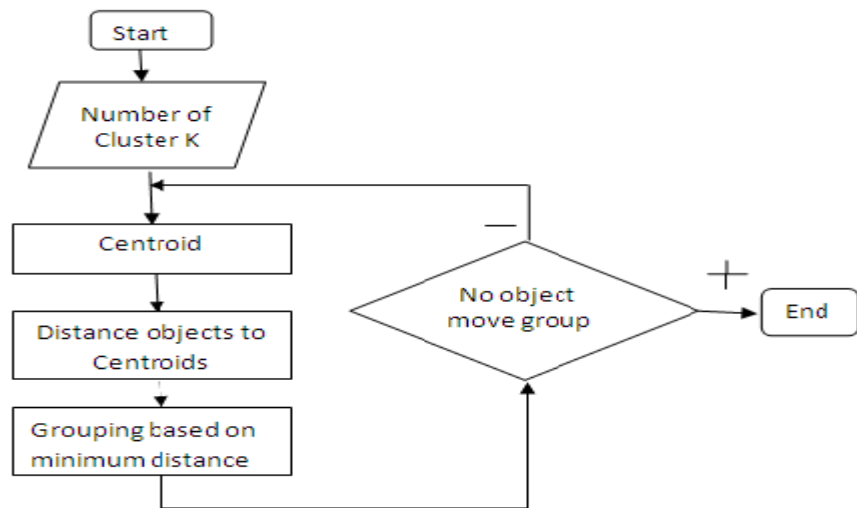
Objectif : regrouper des observations similaires ensemble dans un nombre (k) de clusters prédéfinis.

Objectif

Définir k groupes homogènes parmi les observations



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 :

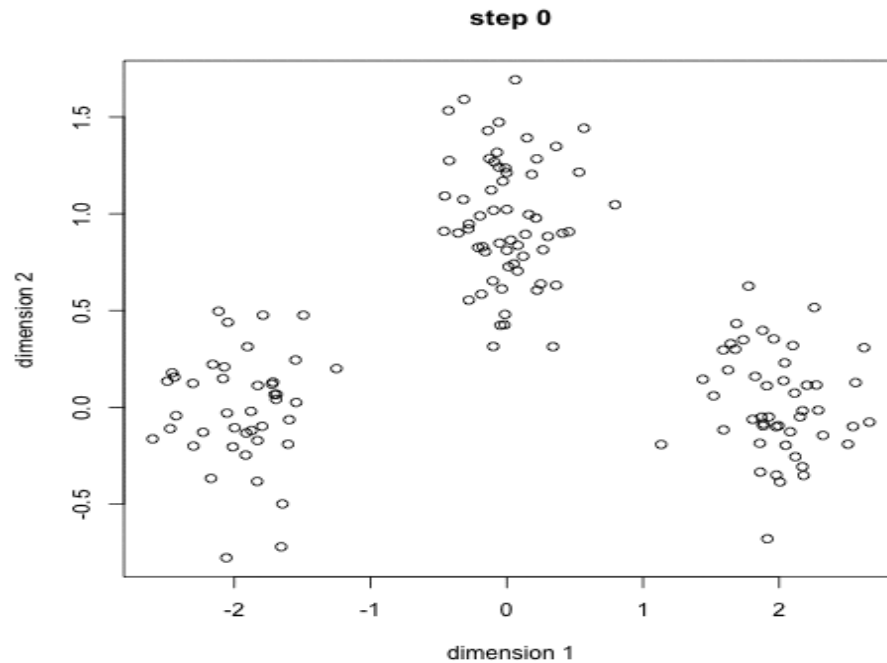
Chaque groupe constitué permet de recalculer un élément « centre » (centre de gravité)

Step 4 :

On réaffecte les individus aux nouveaux centres définis au step 3

On itère jusqu'à ce que les groupes d'individus soient stables (*ie aucun individu ne change plus de groupe*)

Visuellement



Inconvénients de l'algorithme

1. Le nombre de classe doit être fixé au départ
2. Le résultat dépend du tirage aléatoire initial des centres des classes
3. Labels des classes pas stables d'une exécution à l'autre
4. Les clusters générés sont sphériques (pas adapté à tous les datasets)

Minimiser la variance intra

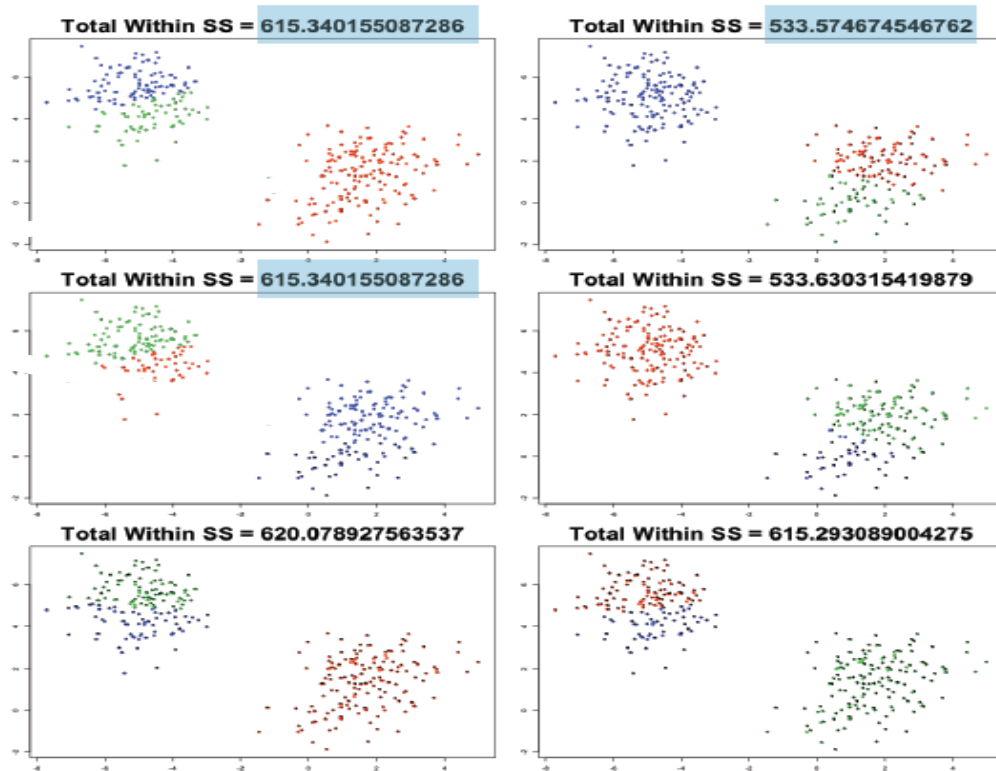
Affecter un individu à son « centre » le plus proche revient à minimiser la variance intra du centroïde.

Formule variance intra

$$\sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

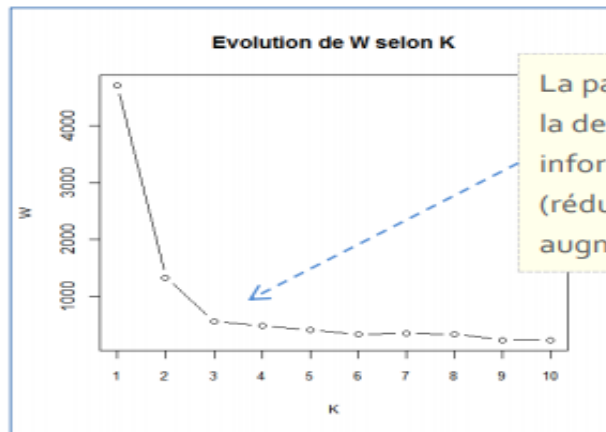
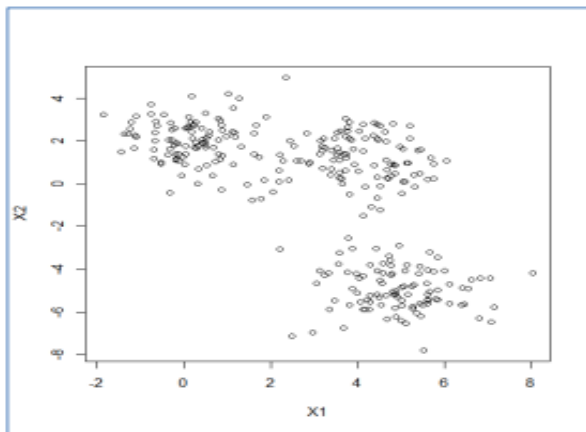
Itérer pour minimiser la variance

Le choix des clusters initiaux étant aléatoire, il faut procéder à plusieurs exécutions pour trouver le modèle qui minimise la variance *intra* (total within Sum of Squares)



Choix de k - Méthode du coude

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)

Choix de k – 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)

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

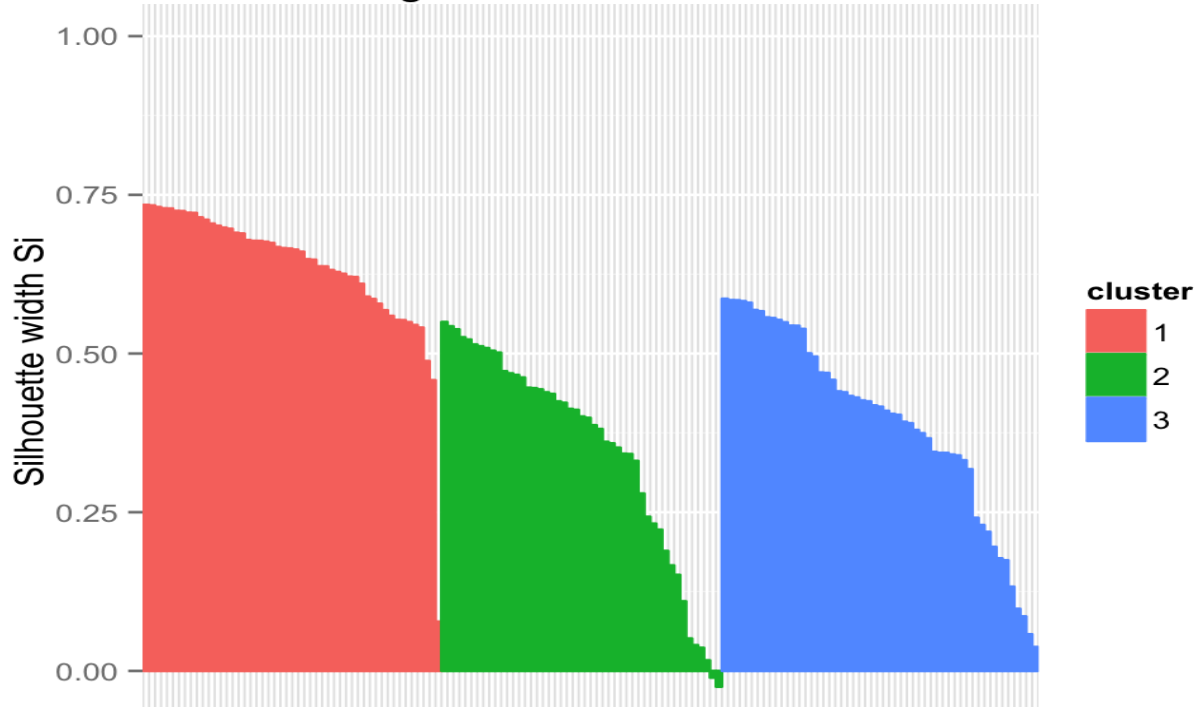
Plus s est proche de 1, plus l'affectation est bonne

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

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

Choix de k – la silhouette

Clusters silhouette plot
Average silhouette width: 0.46



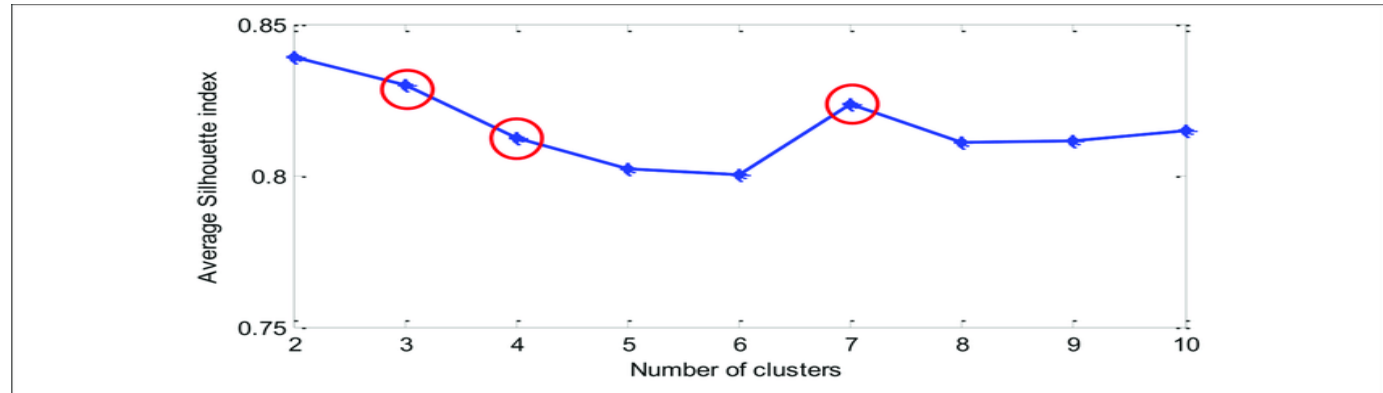
Un bâton->
un individu

Des mal
classés ?

Choix de k – 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

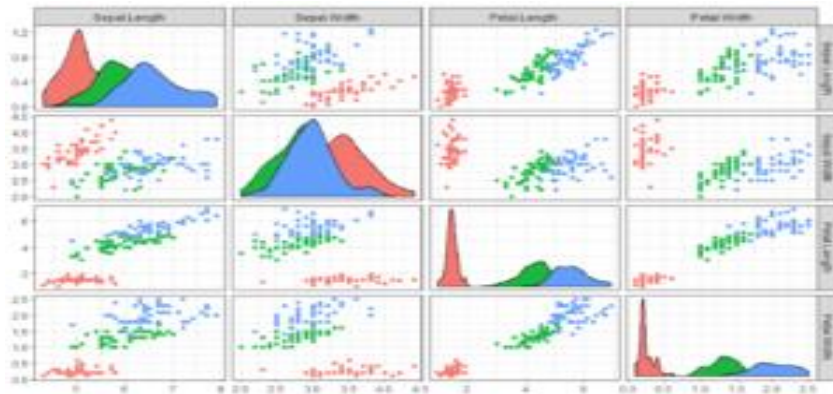
On peut ainsi calculer les silhouettes pour $k:1\dots n$ clusters et choisir la valeur de k qui maximise la silhouette moyenne



Analyser les clusters

Pour chaque cluster

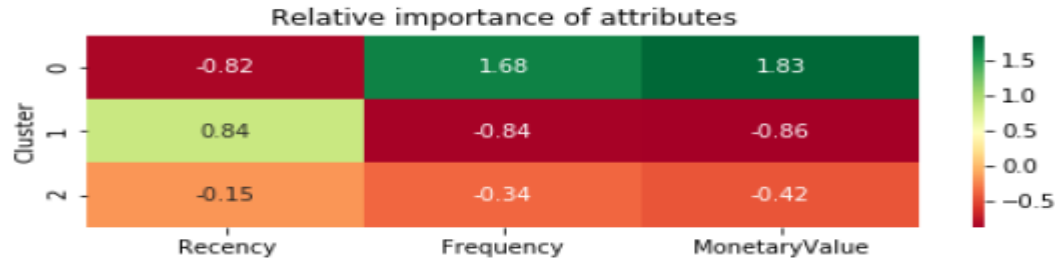
- sortir les statistiques descriptives
- étudier les différences entre les clusters



Analyser les clusters

Identifier les variables marquantes de chaque cluster

Heatmap plot:



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

TD - KMEANS

CLUSTERING HIERARCHIQUE

Clustering hiérarchique

- Basé sur le calcul de la distance entre des individus
- Deux approches : ascendante (de l'individu au groupe) et descendant (du groupe à l'individu)
- Avantage : pas besoin de connaître le nombre de classes

TD – CLASSIFICATION HIERARCHIQUE

DBSCAN

TD – DBSCAN

TD - DBSCAN