Segmentation client appliquée à la clientèle de Olist

Présenté par Gabriel Chehade

Mise en contexte

Contexte :

Olist souhaite fournir à ses équipes d'e-commerce une segmentation des clients pour leurs campagnes de communication.

Objectifs :

- comprendre les différents types d’utilisateurs et les segmenter à partir de leurs données personnelles.
- proposer un contrat de maintenance basée sur une analyse de la stabilité des segments au cours du temps

SOMMAIRE

- **Présentation des données / Preprocessing**

- **Segmentation**

- **Maintenance**

Présentation du jeu de données

Source des données :

https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce (vrai données)

Fichiers .csv :

- olist_customers_dataset : fichier clients (id client, id commande, ville, état)
- olist_order_items_dataset : fichier articles (id commande, nb d'articles, prix)
- olist_order_reviews_dataset : fichier avis (id commande, note, commentaire, date)
- olist_orders_dataset : fichier commandes (id commande, date de la commande, date de livraison...)
- olist_products_dataset : fichier produits (id produit, catégorie, dimensions, poids)
- olist_sellers_dataset : fichier vendeurs (id vendeur, ville, état)

Preprocessing

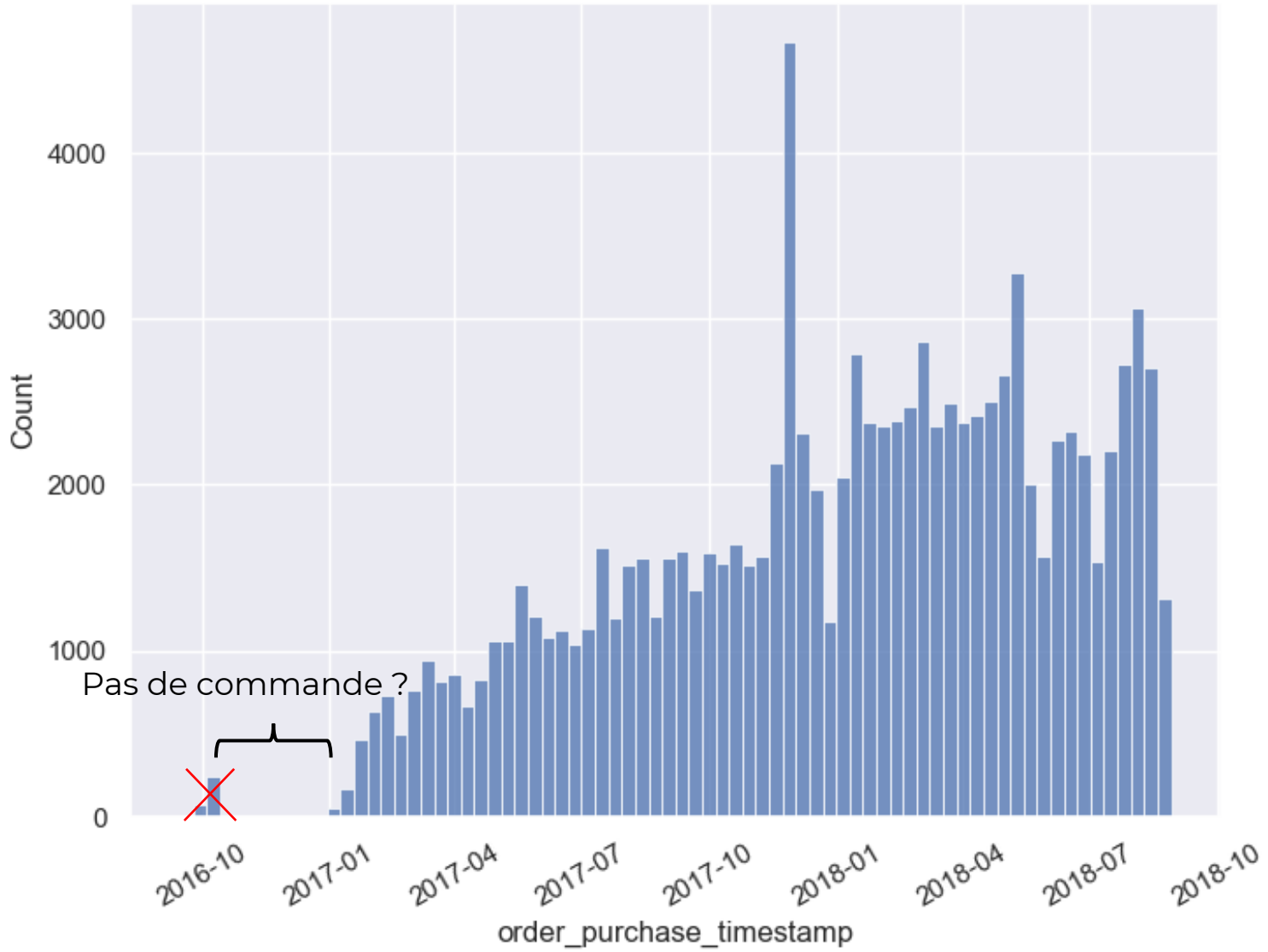
- Regroupement des données en un seul fichier
- Suppression des variables non pertinentes
- Suppression des valeurs manquantes
- Suppression des doublons
- Combinaison de variables

→ Dimensions : 107 800 lignes x 10 colonnes (chaque ligne représente l'article d'une commande)

	order_id	customer_unique_id	order_item_id	order_purchase_timestamp	delivered_estimated_interval	price	freigh
0	00e7ee1b050b8499577073aeb2a297a1	861eff4711a542e4b93843c6dd7febb0	1.0	2017-05-16	-11 days	124.99	
1	29150127e6685892b6eab3eec79f59c7	290c77bc529b7ac935b93aa66c333dc3	1.0	2018-01-12	-8 days	289.00	
2	b2059ed67ce144a36e2aa97d2c9e9ad2	060e732b5b29e8181a18229c7b0b2b5e	1.0	2018-05-19	1 days	139.94	
3	951670f92359f4fe4a63112aa7306eba	259dac757896d24d7702b9acbbff3f3c	1.0	2018-03-13	-13 days	149.94	
4	6b7d50bd145f6fc7f33cebabd7e49d0f	345ecd01c38d18a9036ed96c73b8d066	1.0	2018-07-29	-6 days	230.00	

Preprocessing

Distribution des dates des commandes



Pas de commande entre octobre 2016 et janvier 2017

Suppression des commandes effectuées avant janvier 2017

Segmentation

Segmentation RFM

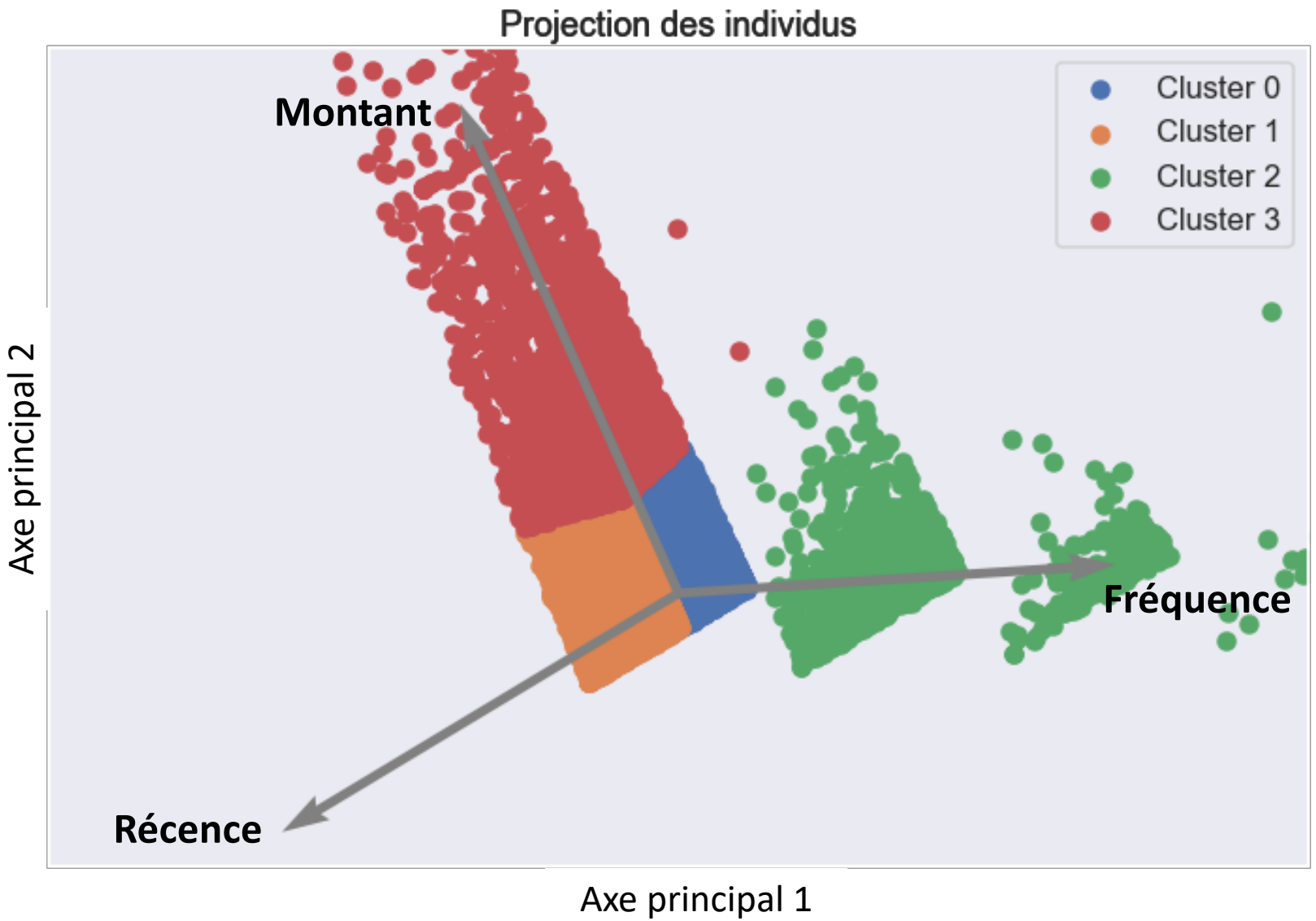
- **Récence** : Nombre de jours écoulés depuis la dernière commande
- **Fréquence** : Nombre de commandes effectuées
- **Montant** : Montant dépensé par commande en moyenne

customer_unique_id	recency	frequency	mean_spending
4b3207464f5f7a48a7f63fa0b1251d86	601	1	9.9
527cd2850ef91088969ffbef0103dec3	601	1	11.9
29a63a400c3ca9982907ce8de1f19527	601	1	10.9
b6b2c3c8fd76769b478618a3c2505009	601	1	10.9
f7be9bec658c62ab6240b44cd26c0b84	601	2	10.4

Segmentation RFM – Choix du modèle

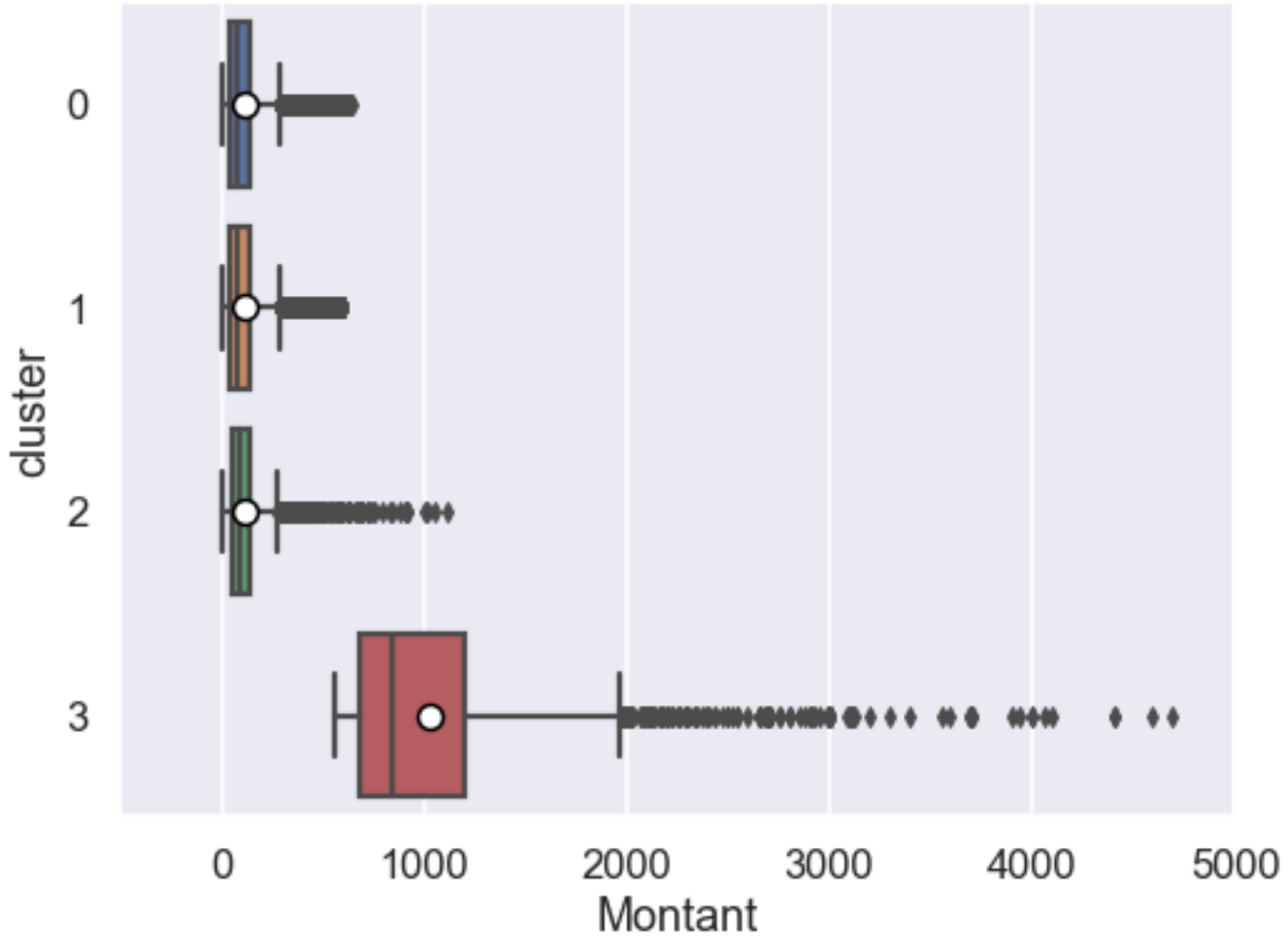
	KMeans	Clustering agglomératif	DBSCAN
Segmentation	✓	✓	✗
Temps de calcul (sur 10k clients)	0.1 s <	2.4 s	

Segmentation RFM - Clustering



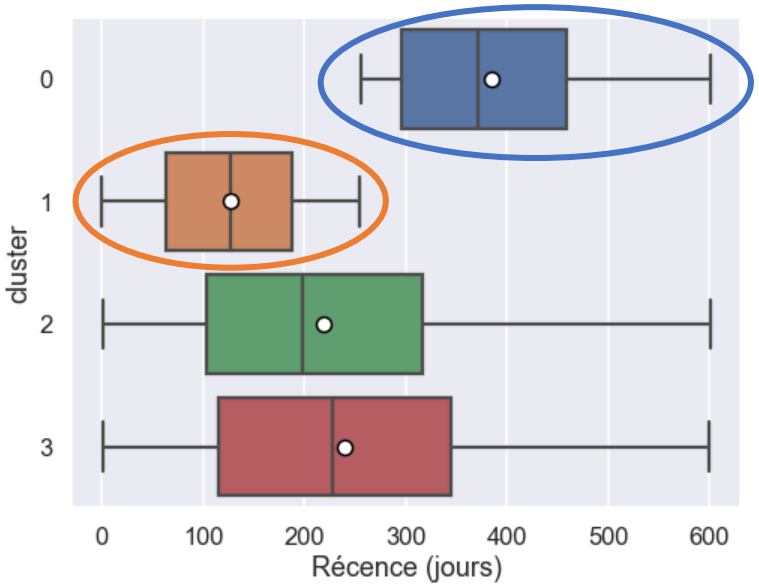
Représentation graphique

Boxplots (boîtes à moustaches)

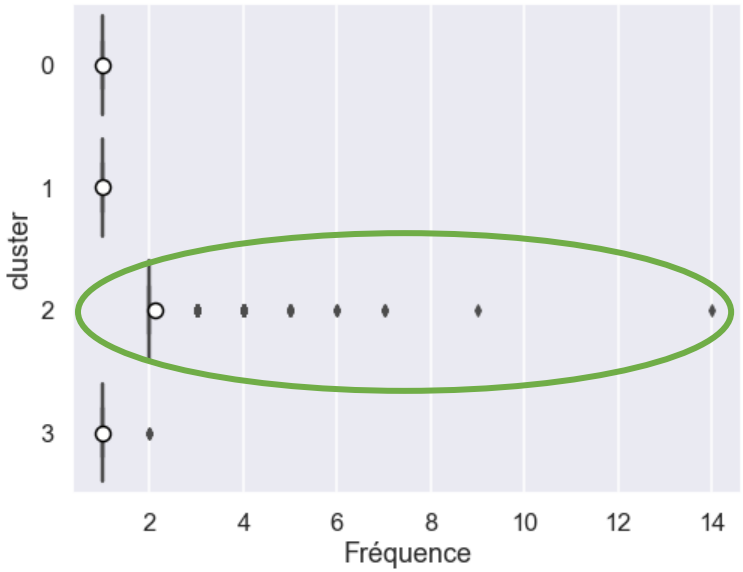


Identification des clusters

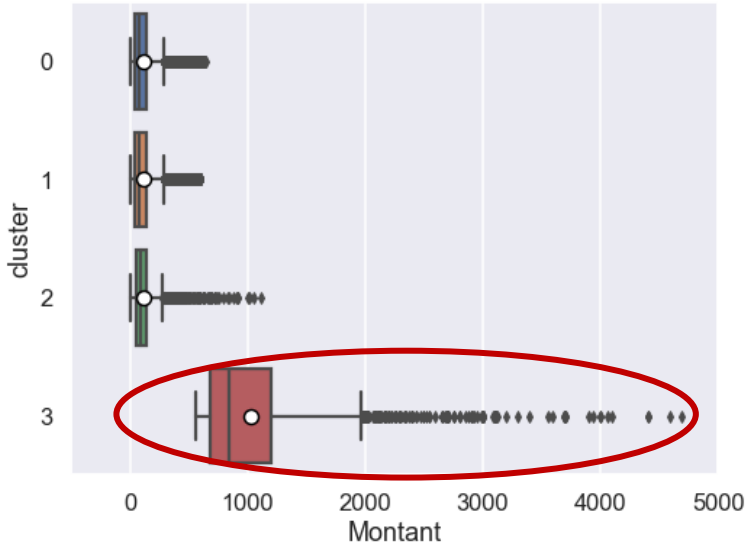
Boxplots pour la Récence



Boxplots pour la Fréquence



Boxplots pour le Montant



Cluster 0 : clients anciens

Cluster 1 : clients récents

Cluster 2 : clients fréquents

Cluster 3 : clients qui dépensent beaucoup

Segmentation RFMSA

- **Récence** : Nombre de jours écoulés depuis la dernière commande
- **Fréquence** : Nombre de commandes effectuées
- **Montant** : Montant dépensé par commande en moyenne
- **Score** : Note de satisfaction attribuée par le client en moyenne
- **Articles par commande** : Nombre d'articles par commande en moyenne

customer_unique_id	recency	frequency	mean_spending	review_score	items_per_order
4b3207464f5f7a48a7f63fa0b1251d86	601	1	9.9	5.0	1
527cd2850ef91088969ffbef0103dec3	601	1	11.9	5.0	1
29a63a400c3ca9982907ce8de1f19527	601	1	10.9	5.0	1
b6b2c3c8fd76769b478618a3c2505009	601	1	10.9	5.0	1
f7be9bec658c62ab6240b44cd26c0b84	601	2	10.4	5.0	1

Clustering

Cluster 0 : clients anciens

Cluster 1 : clients récents

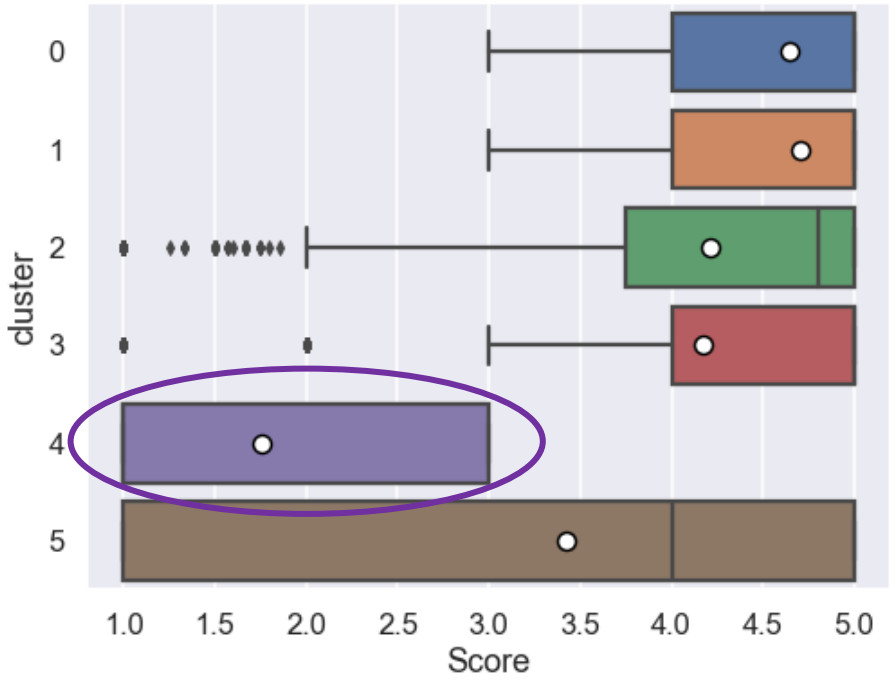
Cluster 2 : clients fréquents

Cluster 3 : clients qui dépensent beaucoup

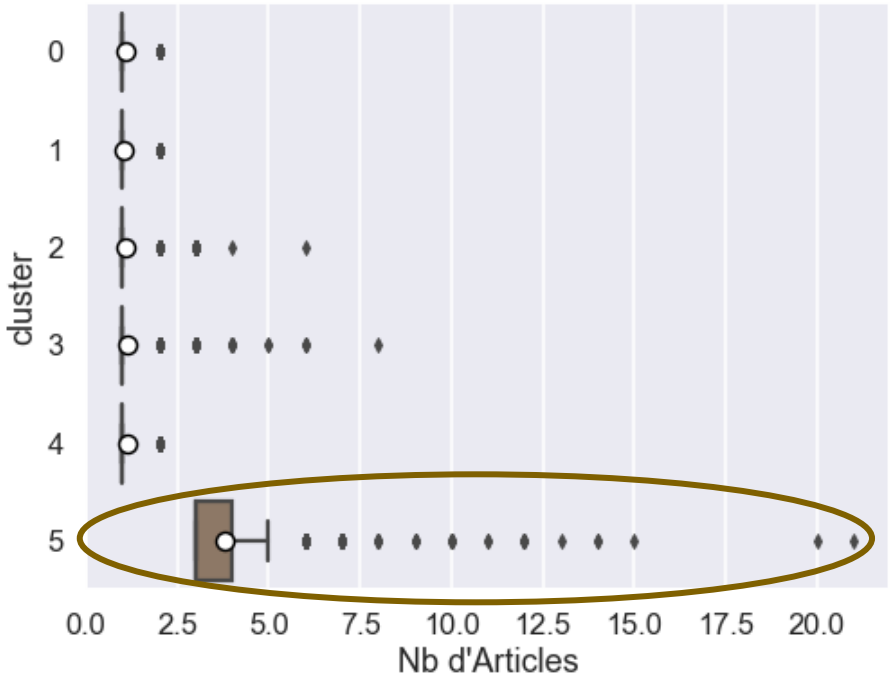
Cluster 4 : clients insatisfaits

Cluster 5 : clients commandant beaucoup d'articles

Boxplots pour le Score

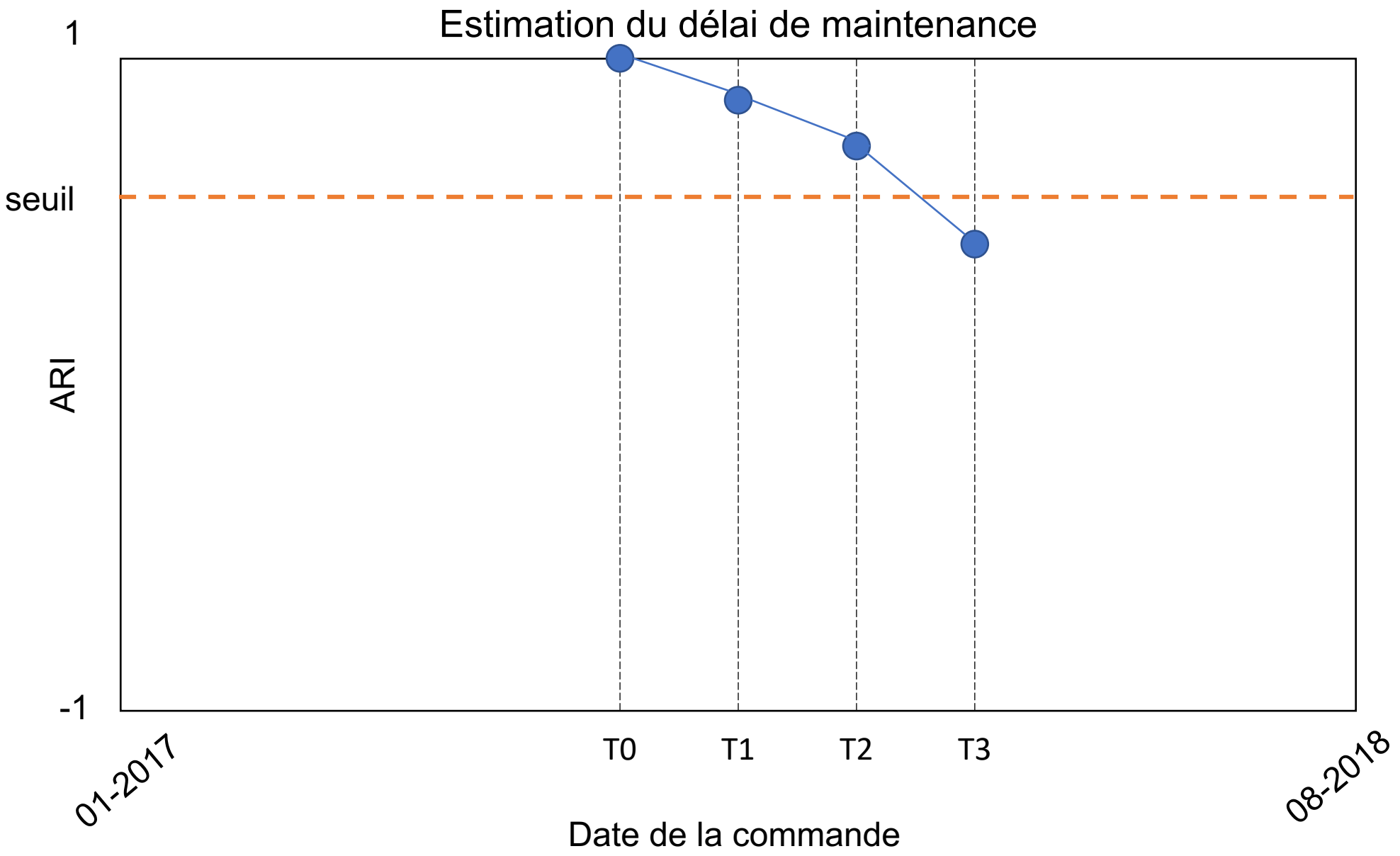


Boxplots pour le nombre d'Articles

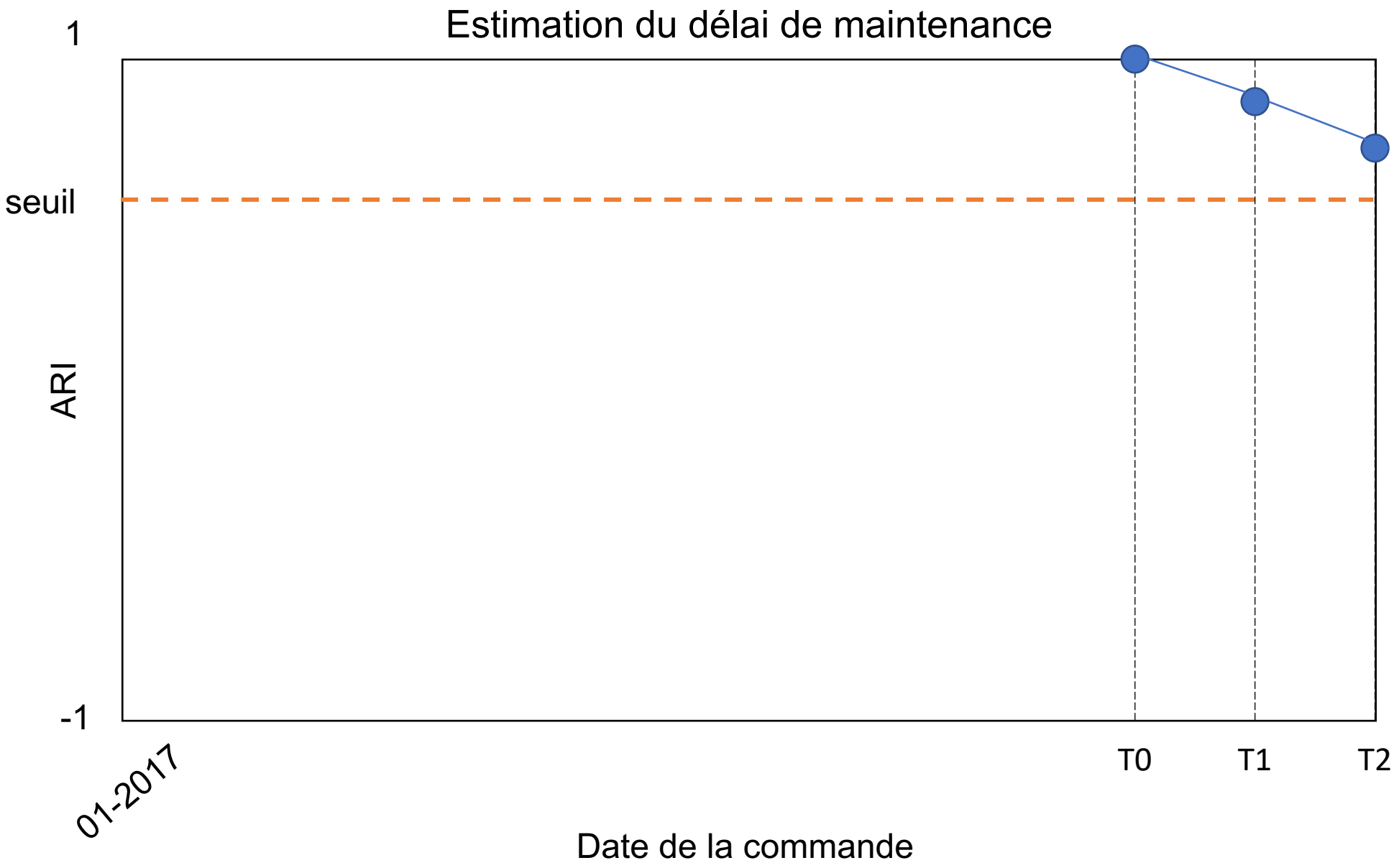


Fréquence de maintenance

Maintenance

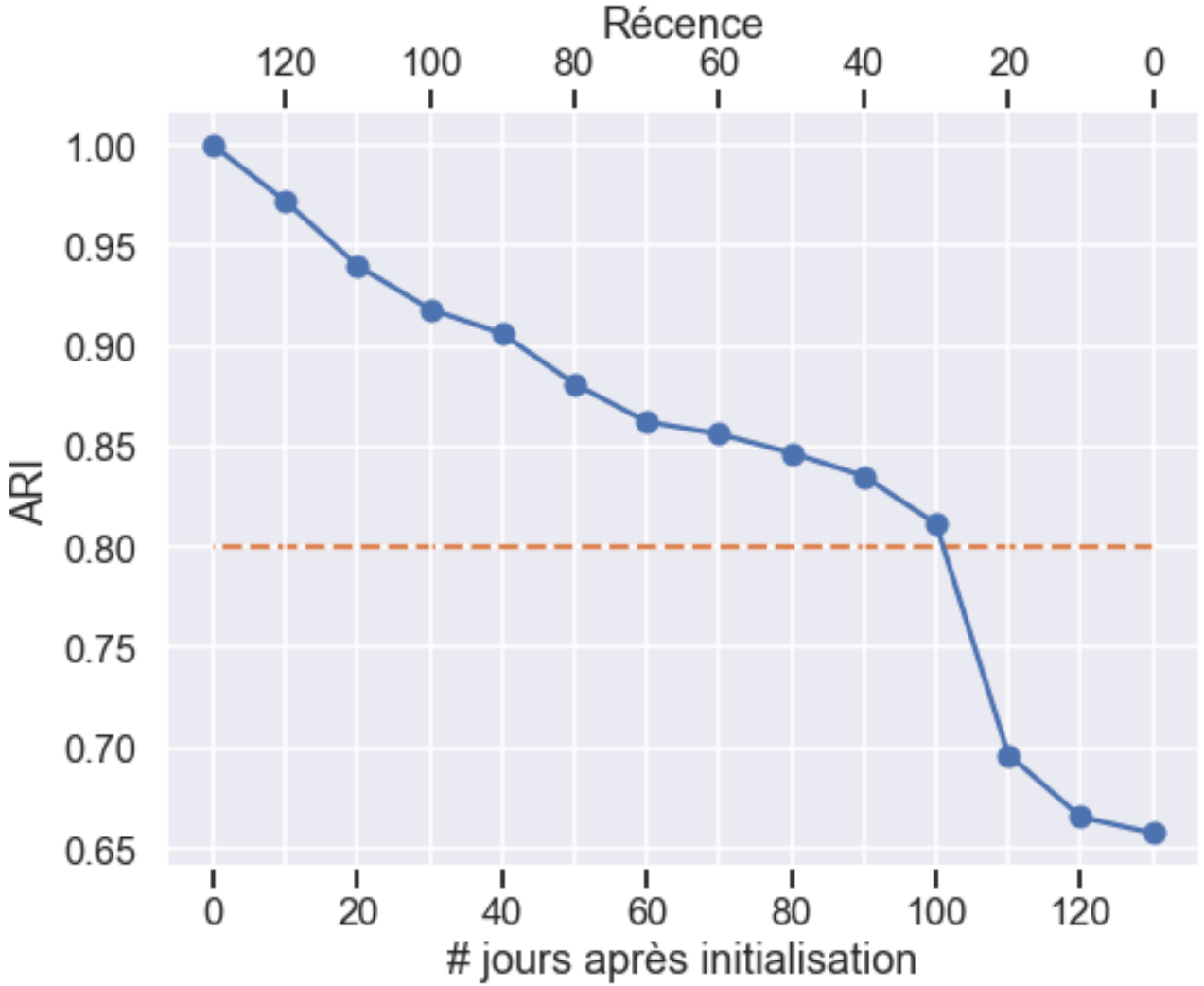


Maintenance



Maintenance

Estimation du délai de maintenance



Fréquence de maintenance :

100 jours

Conclusion

- La segmentation RFM donne des résultats très intéressants permet d'identifier les clients les plus intéressants (fréquents et dépensiers)

- L'ajout de deux variables permet de identifier les clients mécontents et ceux qui achètent beaucoup d'articles

- La période de maintenance est estimée à 100 jours en prenant un seuil de 0.8 pour l'ARI (Adjusted Rand Index)

MERCI POUR VOTRE ATTENTION

Adjusted Rand Index (ARI)

Pour mesurer la similarité entre deux clusterings K_1 et K_2 on peut utiliser l'ARI qui se définit de la manière suivante :

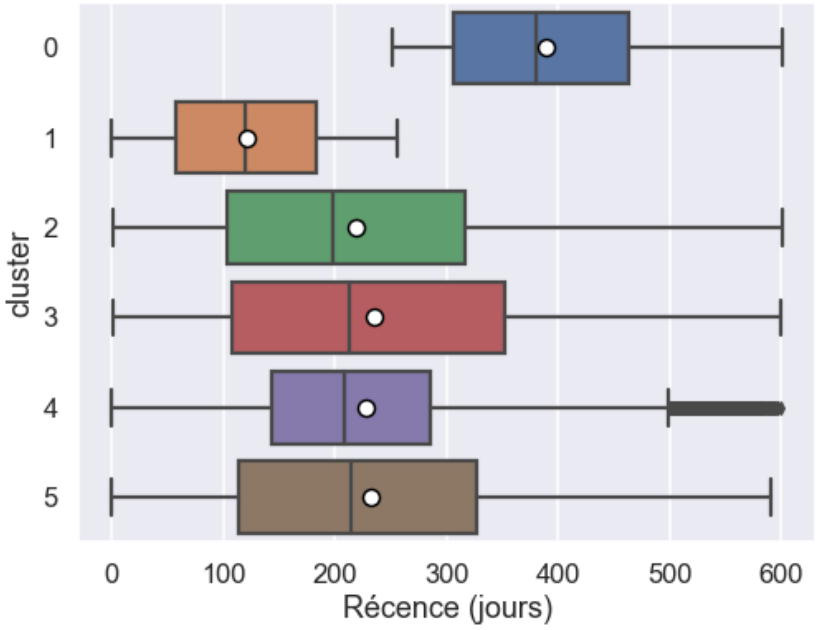
$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]}$$

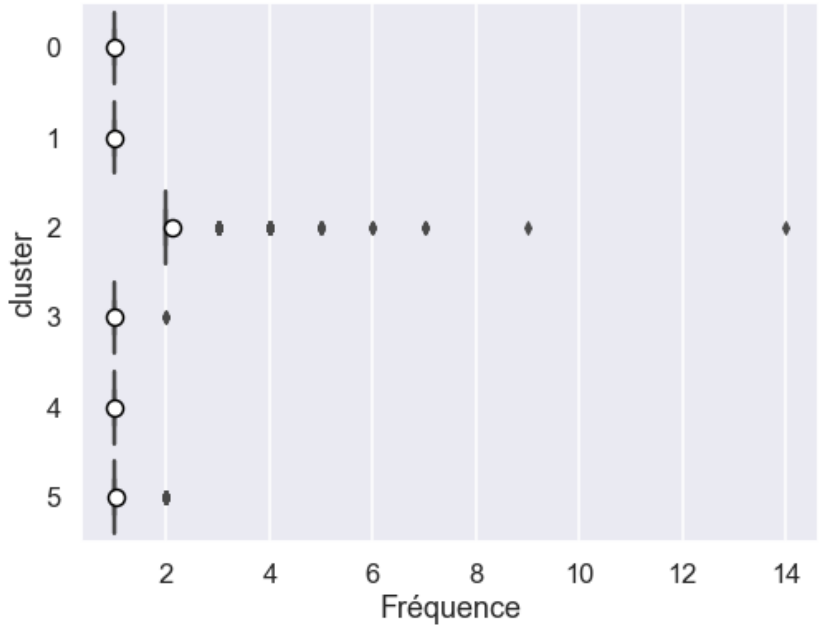
avec

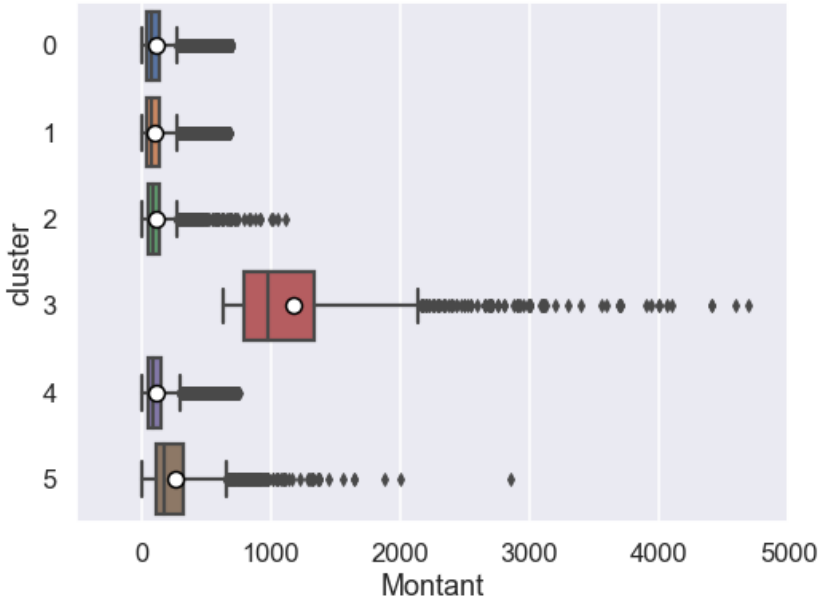
$$RI = \frac{a + b}{C_2^n}$$

RI : indice de Rand (Rand index)
$E[RI]$: espérance de la valeur de RI, c'est-à-dire la valeur de RI obtenue si on partitionne les données au hasard dans K_2
a : nombre de paires d'individus qui sont dans le même cluster à la fois dans K_1 et dans K_2.
b : nombre de paires d'individus qui sont dans des clusters différents à la fois dans K_1 et dans K_2.
C_2^n : nombre total de paires possibles dans le jeu de données (n correspond au nombre d'individus).

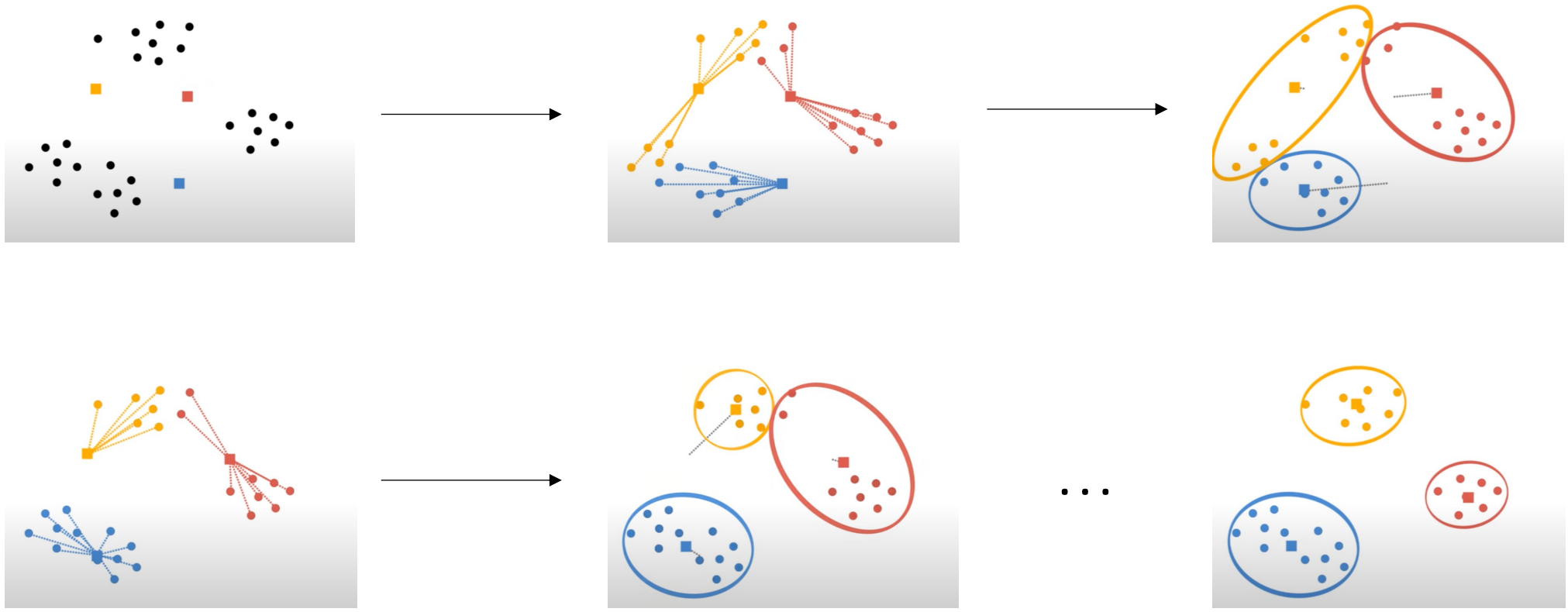
Segmentation RFMSA





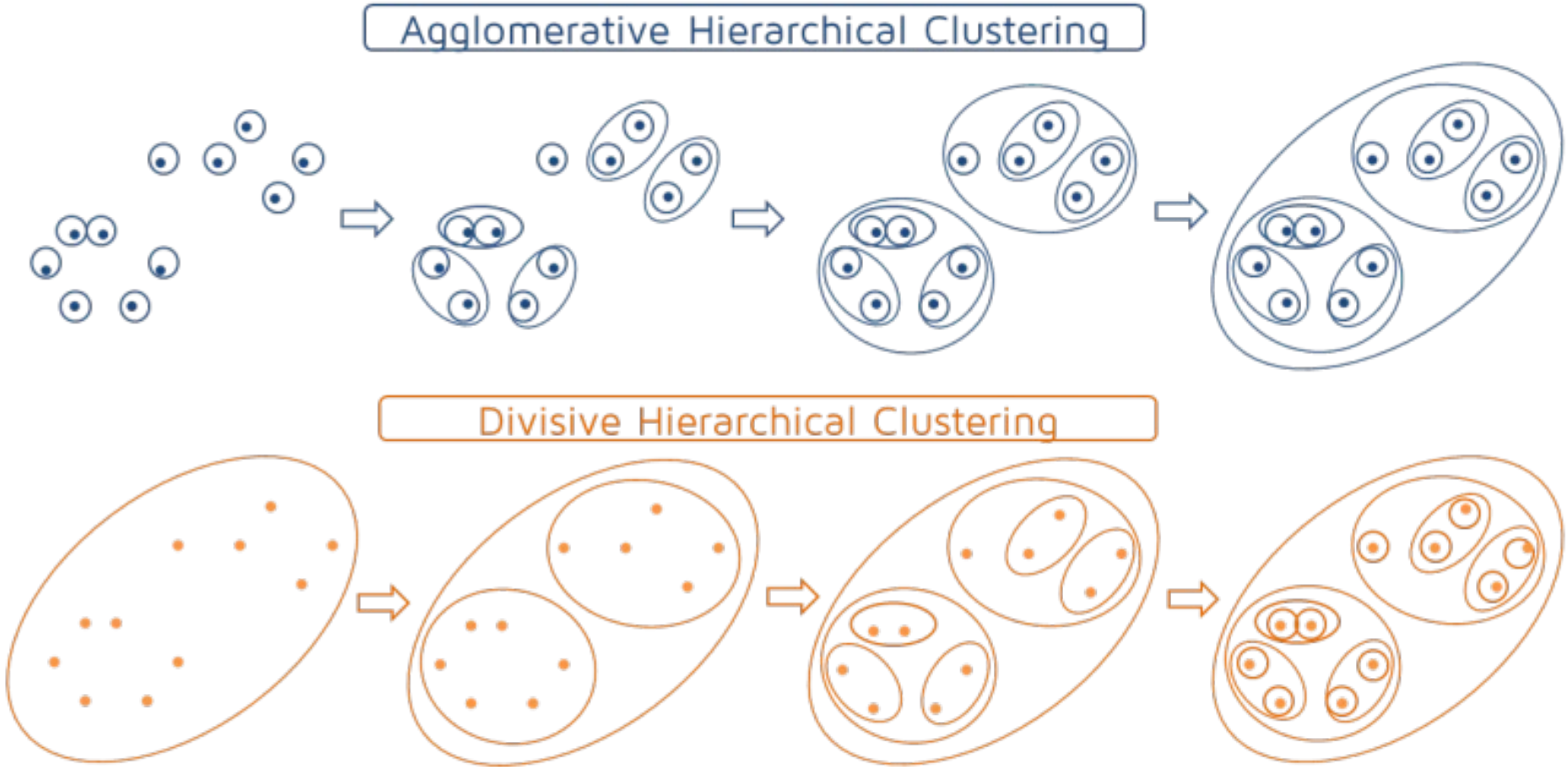


KMeans



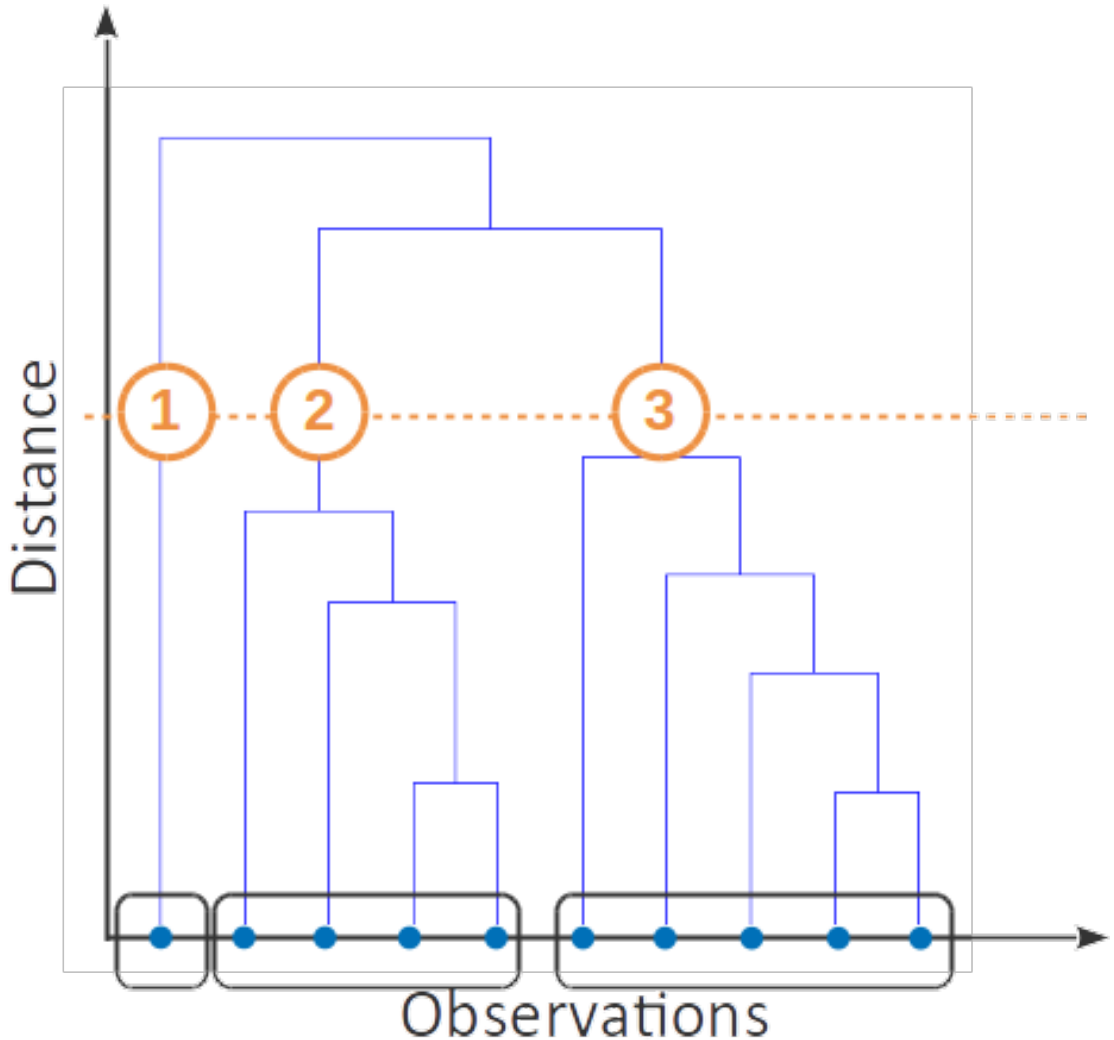
Source : https://www.youtube.com/watch?v=FTtzd31IAOw&list=PLO_fdPEVlfKqMDNmCFzQISI2H_nJcEDJq&index=28 (Machine Learnia)

Clustering hiérarchique



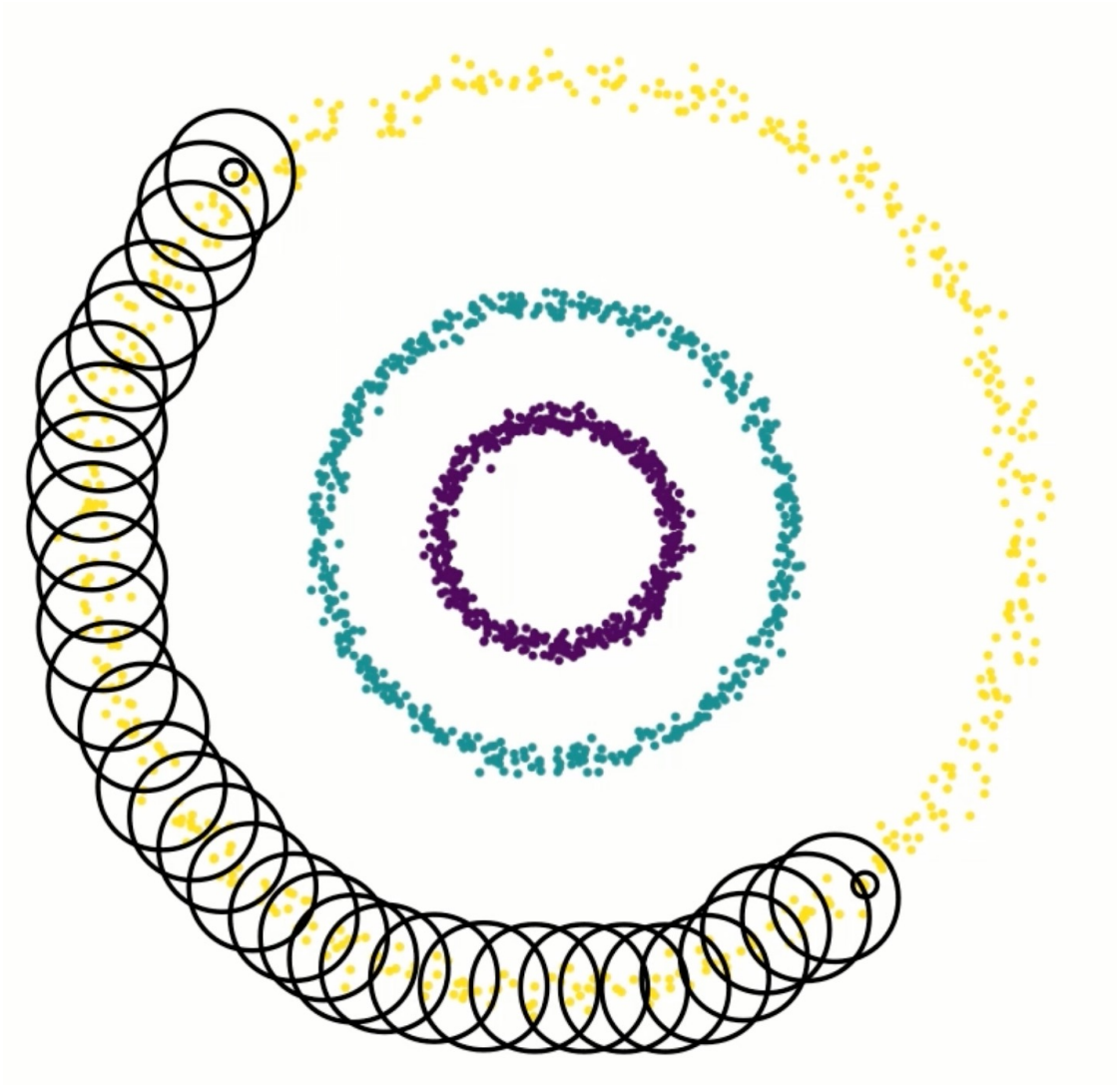
Source : https://quantdare.com/hierarchical-clustering/

Dendogramme



Source : https://openclassrooms.com/fr/courses/4379436-explorez-vos-donnees-avec-des-algorithmes-non-supervises/4379561-partitionnez-vos-donnees-avec-un-algorithme-de-clustering-hierarchique

DBSCAN



Source https://openclassrooms.com/fr/courses/4379436-explorez-vos-donnees-avec-des-algorithmes-non-supervises/4379571-partitionnez-vos-donnees-avec-dbscan

Comparaison des modèles

KMeans

- Efficace en temps de calcul
- Nombre de clusters fixe

- Sensible aux outliers

Clustering hiérarchique

- Possibilité de choisir le nombre de clusters en fonction du dendogramme
- Possibilité d'utiliser différents types de distance

- Pas efficace en temps de calcul

DBSCAN

- Peut former des clusters non convexes (clustering par densité)
- Efficace en temps de calcul
- Pas besoin de prédéfinir le nombre de clusters

- Difficile à utiliser en grande dimension
- Choix des paramètres ε et n_{min}