Segmenter des clients d’un site e-commerce

Guilhem Berthou - Pierre-Antoine Ganaye (mentor)

Introduction

Problématique : Olist est un site de e-commerce Segmentation site e-commerce.

Comprendre les différents **types d'utilisateurs**

Descriptions actionnables de **segmentation Marketing**

Contrat de maintenance

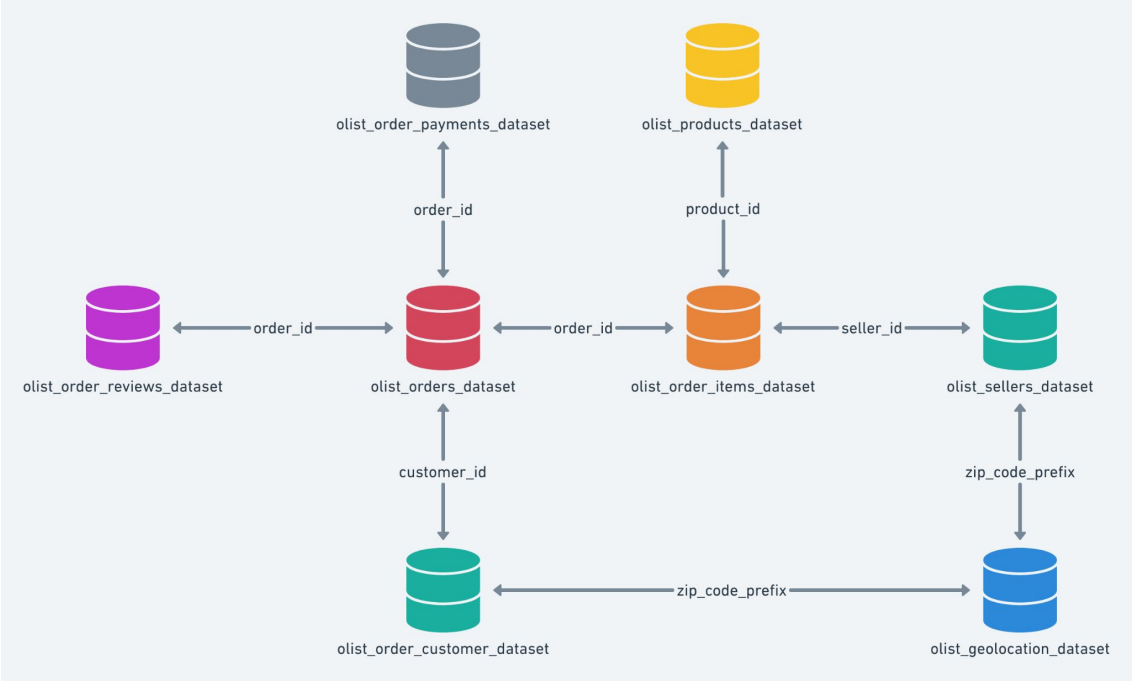
⇒ **Interprétation :** Il s'agit d'un problème de **classification** qui pourra être résolu grâce à l'utilisation d'**algorithmes non supervisés**.

Sommaire

Sommaire

Préparation de la donnée

➢ Structure du dataset




```
Orders [99441, 8]
Order Items [112650, 7]
Customers [99441, 5]
Products Items [32951, 9]
Product Catagory Translation [71, 2]
Payments [103886, 5]
Review [100000, 7]
Geolocation [1000163, 5]
Sellers [3095, 4]
```

Préparation de la donnée

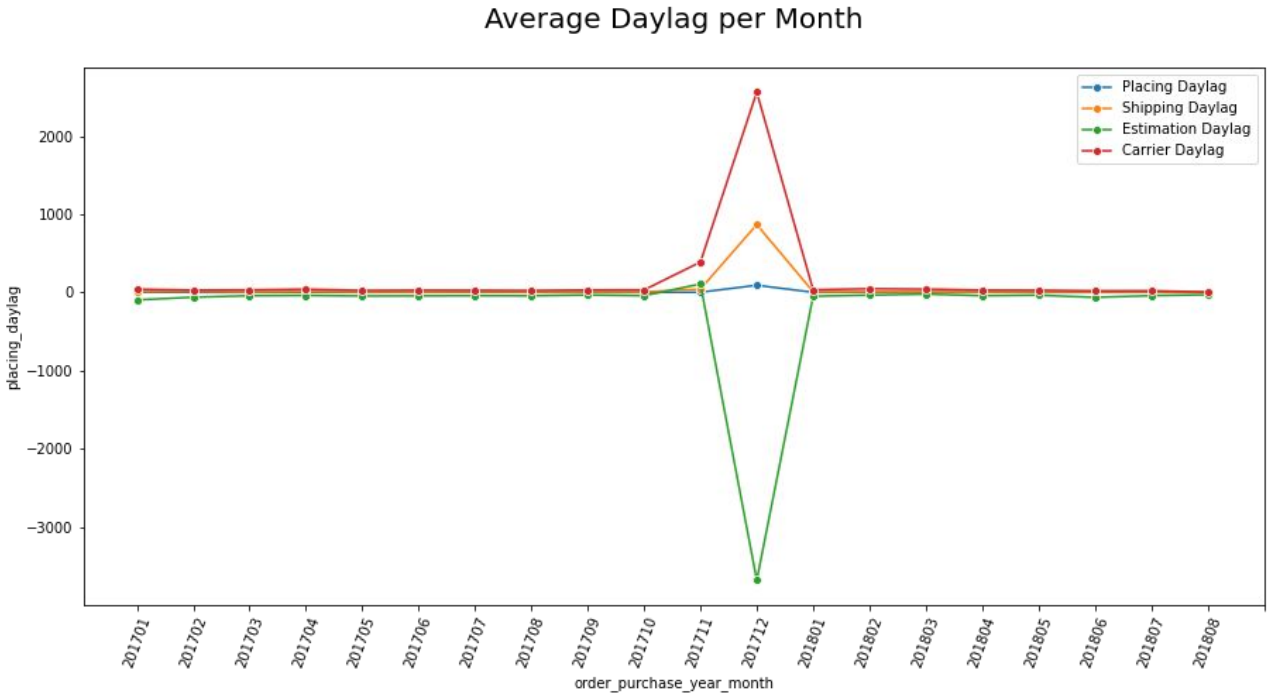
➢ Nettoyage

- Commandes livrées (98%)
- Suppression des outliers (dernier quantile - 99%)
- Suppression des périodes bornes

Préparation de la donnée

➢ Feature Engineering

- Temps de livraison



Sommaire

Clustering

➢ Segmentation RFM

- *Description de la méthode :*

  - Méthode marketing
    - Récence
    - Fréquence
    - Monétaire

  - Mode de calcul des scores
    - *R-M :* Division selon les 3 premiers quantiles
    - *Fréquence* : utilisation d'un booléen selon La valeur de la fréquence (1 pour 1 ou 2)

	R	F	M
count	90164.000000	90164.000000	90164.000000
mean	236.380884	1.032840	159.077784
std	151.946197	0.207759	218.678270
min	0.000000	1.000000	9.590000
25%	113.000000	1.000000	61.960000
50%	219.000000	1.000000	105.115000
75%	347.000000	1.000000	175.330000
max	602.000000	15.000000	13664.080000

=> Conclusion : Nous constatons que les clients de la plateforme n'ont pas renouvelés leurs achats (***fréquence = 0***). Nous allons donc privilégier l'**ajout d'autres variables** pour réaliser une segmentation pertinente.

Clustering

➢ Segmentation RFM

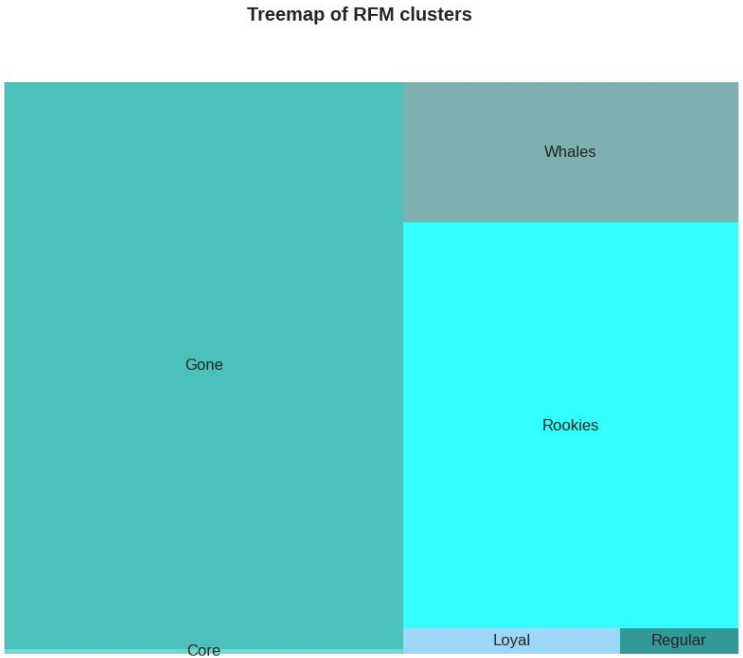
- *Interprétation des segments obtenus :*

Segment name	Description	Marketing Strategies
1. CORE - Best Customers	Highly engaged customers who have bought the most recent, the most often, and generated the most revenue.	Focus on loyalty programs and new product introductions. These customers have proven to have a higher willingness to pay, so don't use discount pricing to generate incremental sales. Instead, focus on value added offers through product recommendations based on previous purchases.
2. LOYAL - Your Most Loyal Customers	Customers who buy the most often from your store.	Loyalty programs are effective for these repeat visitors. Advocacy programs and reviews are also common strategies. Lastly, consider rewarding these customers with Free Shipping or other like benefits.
3. WHALES - Your Highest Paying Customers	Customers who have generated the most revenue for your store.	These customers have demonstrated a high willingness to pay. Consider premium offers, subscription tiers, luxury products, or value add cross/up-sells to increase AOV. Don't waste margin on discounts.
4. REGULAR - Faithful customers	Customers who return often, but do not spend a lot.	You've already succeeded in creating loyalty. Focus on increasing monetization through product recommendations based on past purchases and incentives tied to spending thresholds.
5. ROOKIES - Your Newest Customers	First time buyers on your site.	Most customers never graduate to loyal. Having clear strategies in place for first time buyers such as triggered welcome emails will pay dividends.
6. GONE - Once Loyal, Now Gone	Great past customers who haven't bought in a while.	Customers leave for a variety of reasons. Depending on your situation price deals, new product launches, or other retention strategies.

Clustering

➢ Segmentation RFM

- *Interprétation des segments obtenus :*



segments	RecencyMean	FrequencyMean	MonetaryMean	GroupSize
Core	73.047904	2.194611	271.292712	334
Gone	338.257627	1.000000	127.955765	48640
Loyal	307.323859	2.100775	84.392004	1161
Regular	74.031496	2.127559	86.252458	635
Rookies	74.259326	1.000000	164.344075	29272
Whales	223.104624	1.056115	302.833424	10122

Clustering

➢ Ajout de variables supplémentaires

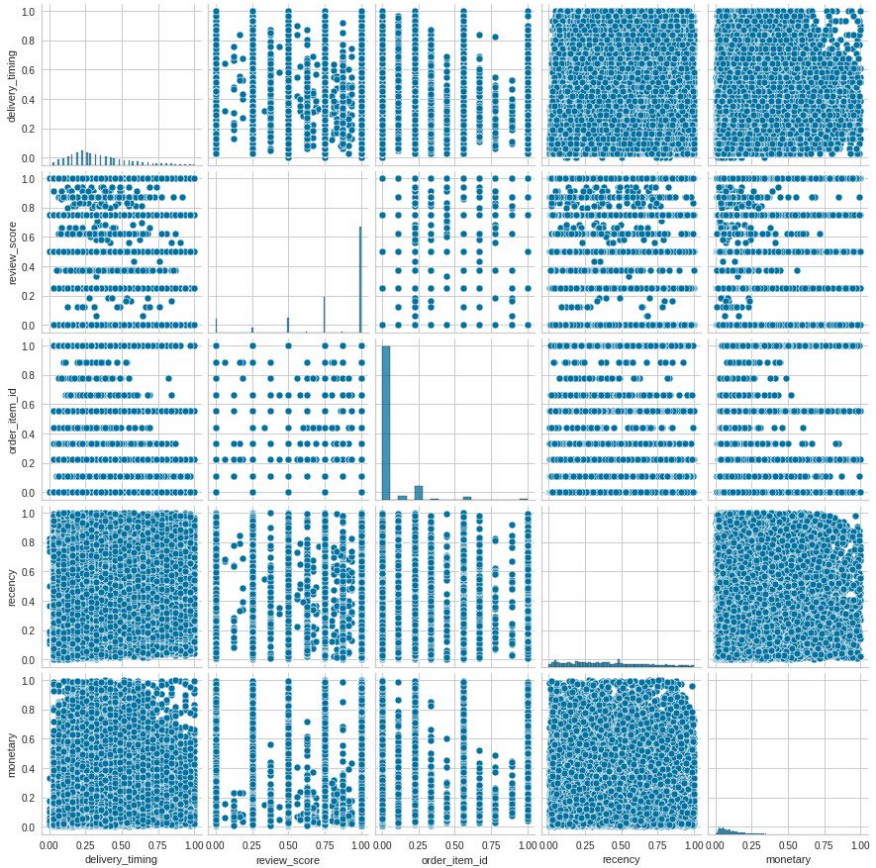
Actual lead time (days)
- Nombre de jours nécessaires pour qu'un produit soit livré

Monetary
- *Issu de la segmentation marketing RFM*

Recency
- *Issu de la segmentation marketing RFM*

Review Score (Score 1-5)
- Avis client

Nombre de produit achetés (par client)

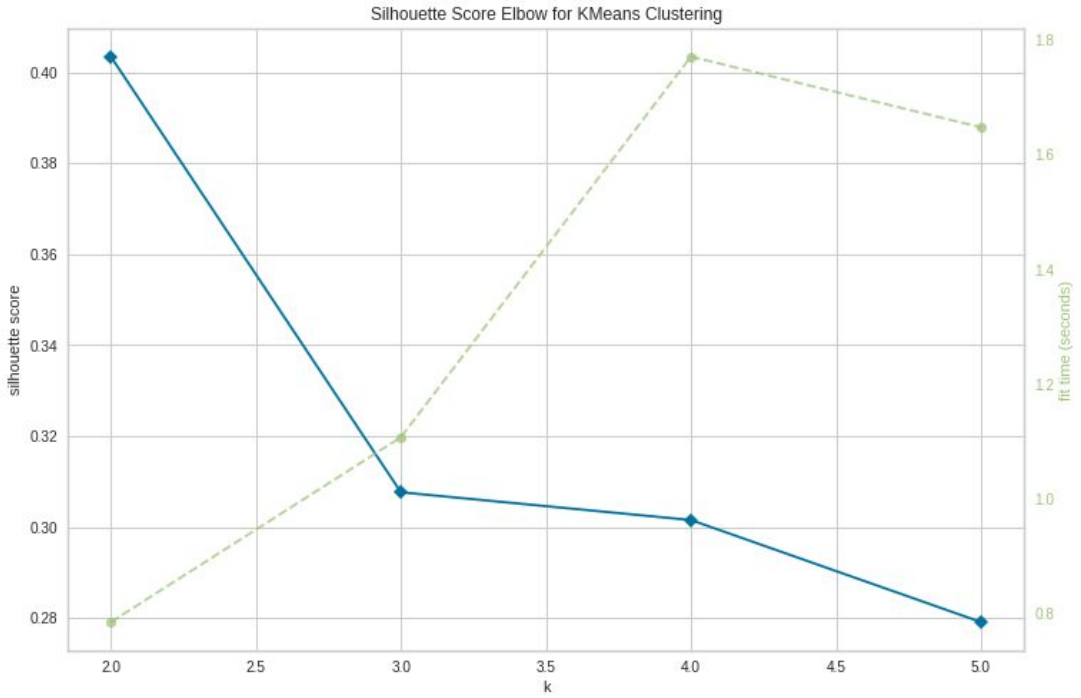


=> Conclusion : Nous n'identifions pas de tendances visuelles nettes.

Clustering

➢ **k-means**

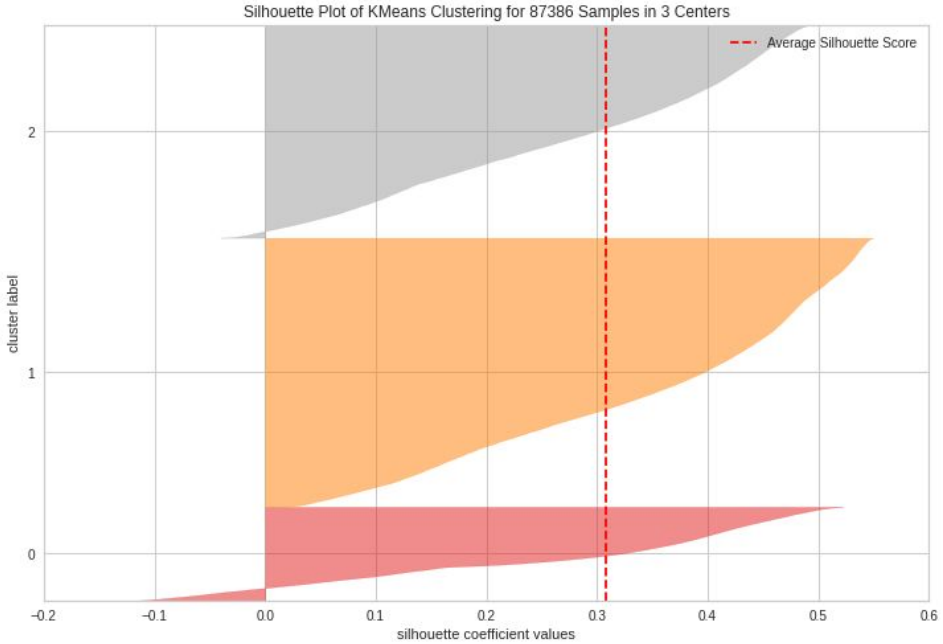
- **Paramètre à harmoniser**
  - k voisins (K = 3)
- **Métrique d'évaluation**
  - Score silhouette



Clustering

➢ **k-means**

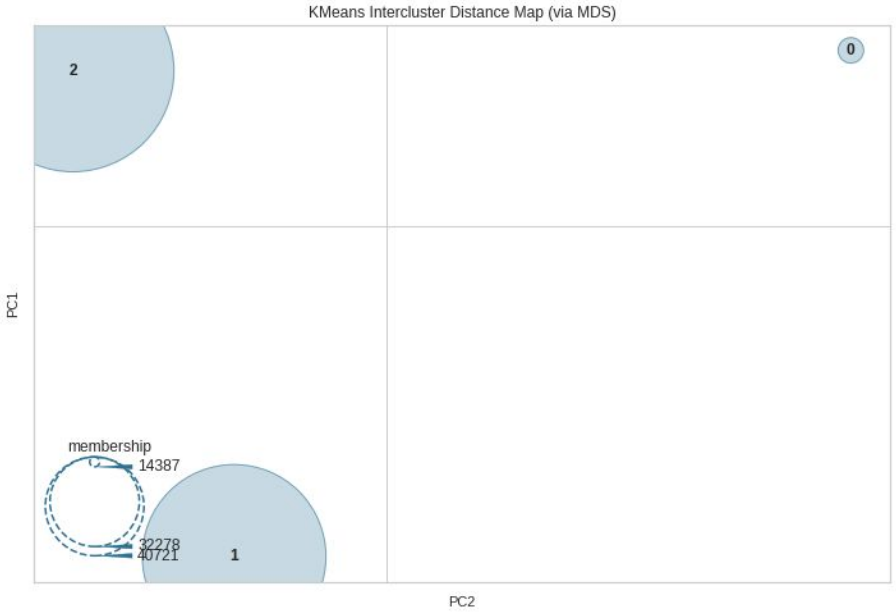
- **Paramètre à harmoniser**
  - k voisins (K = 3)
- **Métrique d'évaluation**
  - Score silhouette



Clustering

➢ **k-means**

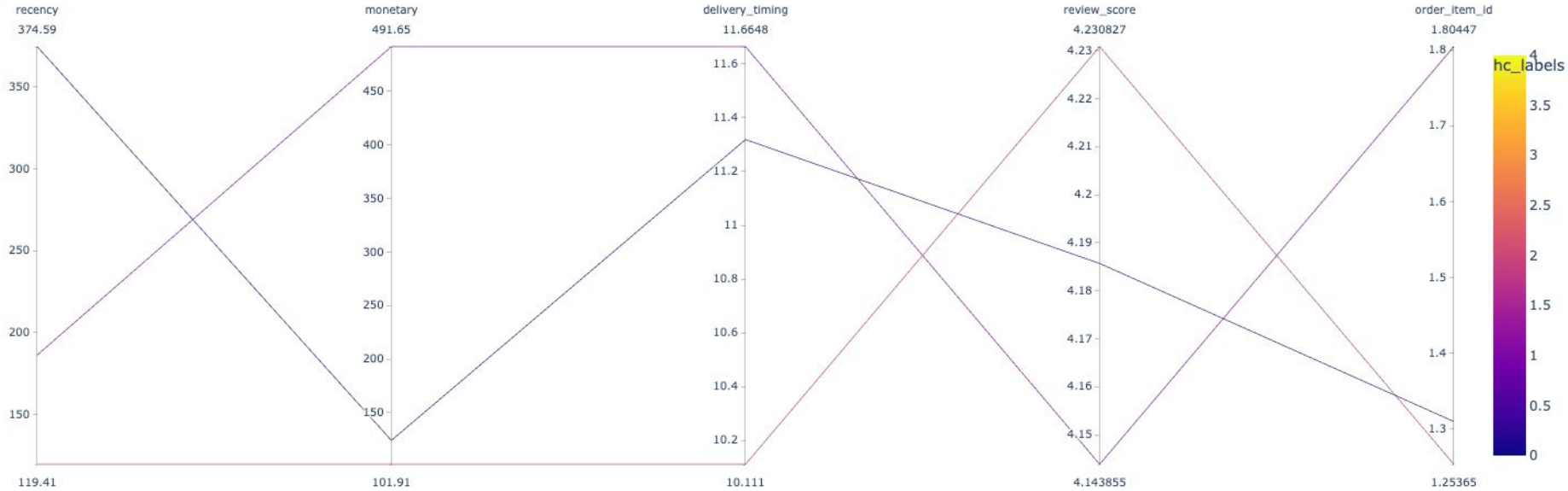
- **Paramètre à harmoniser**
  - k voisins (K = 3)
- **Métrique d'évaluation**
  - Score silhouette
  - Distance inter-cluster



Clustering

➢ **k-means**

○ **Interprétation des segments (3 clusters)**



Clustering

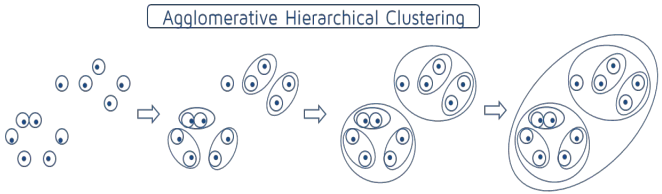
- **k-means**
  - **Interprétation des segments (3 clusters)**
    - **Cluster 1** *(label = 0)* **: Clients récents livrés rapidement**
    - **Cluster 2** *(label = 1)* **: Clients fidèles et tolérants**
    - **Cluster 3** *(label = 2)* **: Clients exigeants déçus de la livraison**

kmeans_label	delivery_timing	review_score	order_item_id	recency	monetary
0	9.303846	4.720106	1.244721	119.051722	139.369774
1	11.322123	4.656238	1.253527	380.069183	141.019543
2	13.911063	1.872737	1.603809	225.614583	154.388790

Clustering

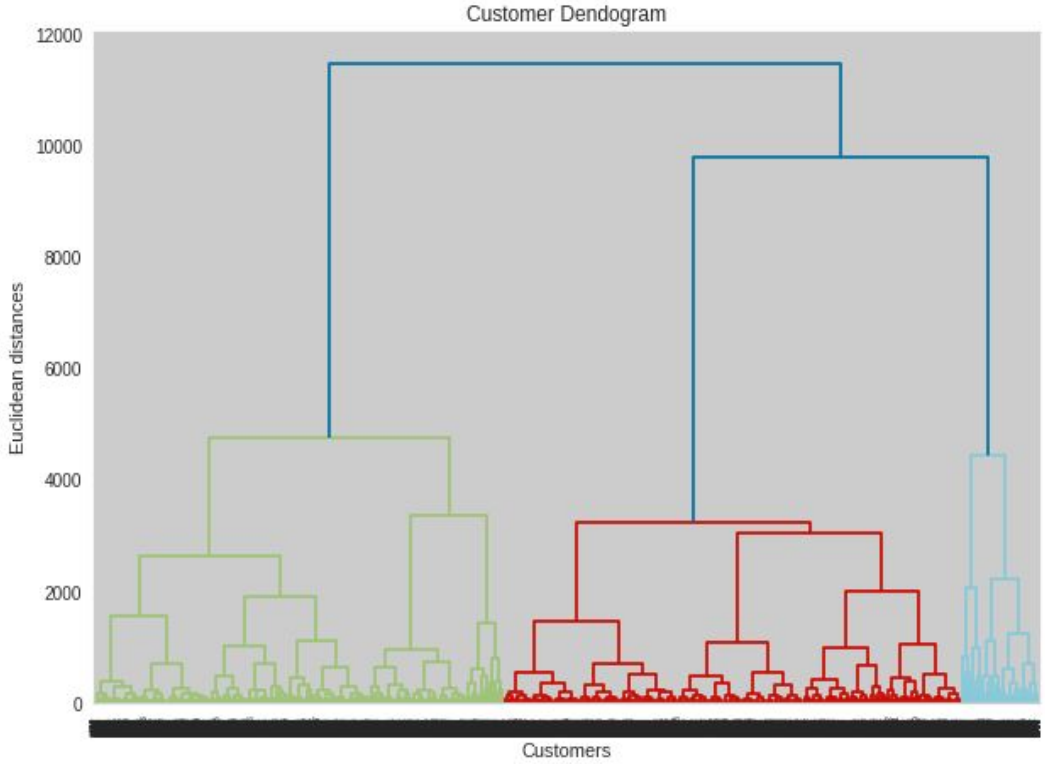
➢ CAH

- **Clustering agglomératif**



- **Paramètre à harmoniser**
  - Choix de la distance

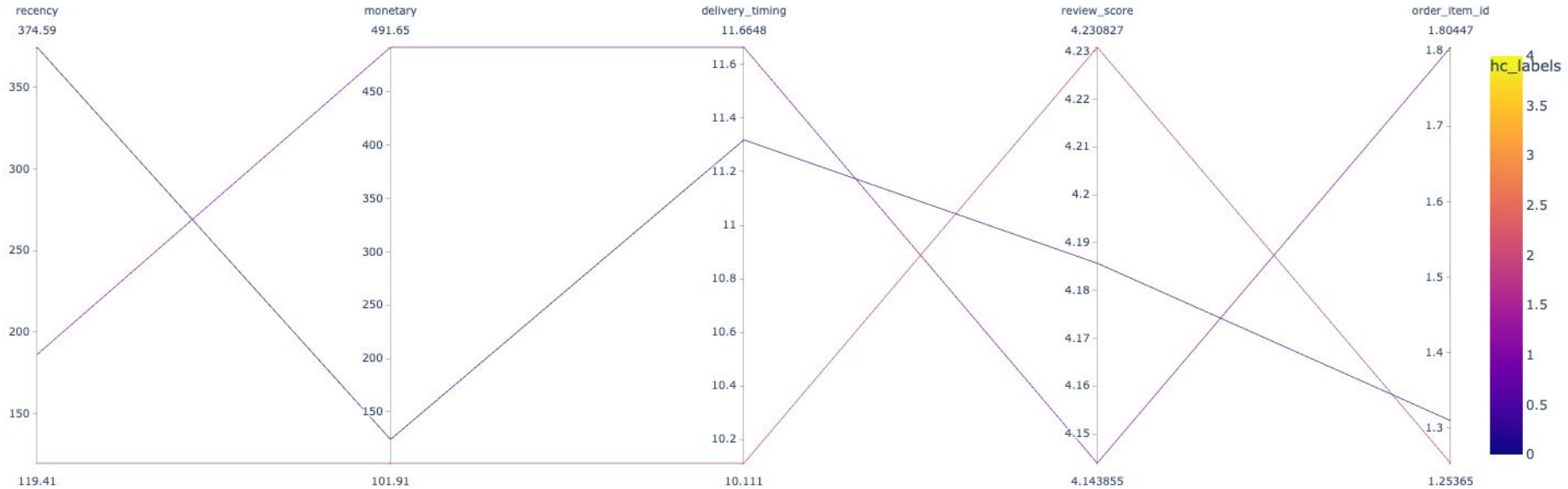
    (linkage criterion = ward)



Clustering

➢ CAH

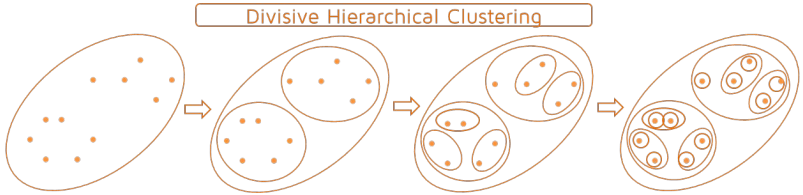
- **Interprétation des segments (3 clusters)**



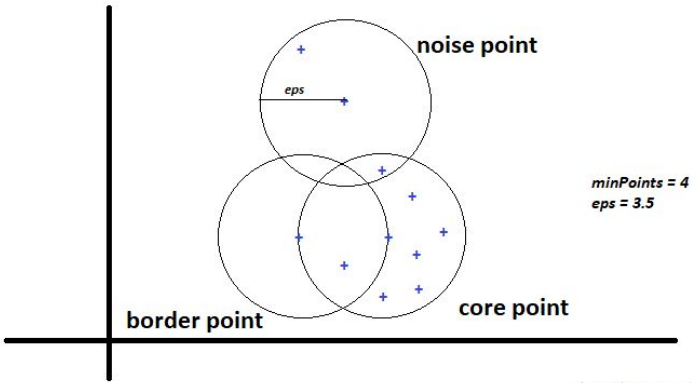
Clustering

➢ DBSCAN

- **Clustering basé sur la densité**



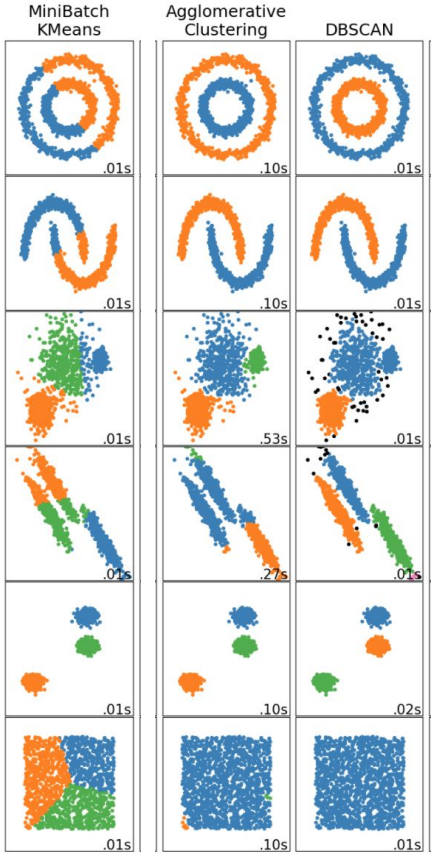
- **Paramètre à harmoniser**
  - Epsilon
  - Minimum sample size



Clustering

➢ **Choix de l'algorithme**

- **K-means**
  - **Meilleur choix**
- CAH :
  - Gourmand en mémoire (distance quadratique)
- DBSCAN :
  - Pas adapté à la structure des données



Sommaire

Analyse de stabilité

➢ Pourquoi une analyse de stabilité ?

- *A partir de quand est-ce que le client doit renouveler le modèle ?*
- *A partir de quand est-ce que les données sur lesquelles le modèle a été entrainé n'est plus représentatif ?*
- *Que les clusters ne reflètent plus les comportements des clients (évolution en fonction des saisons etc.)*

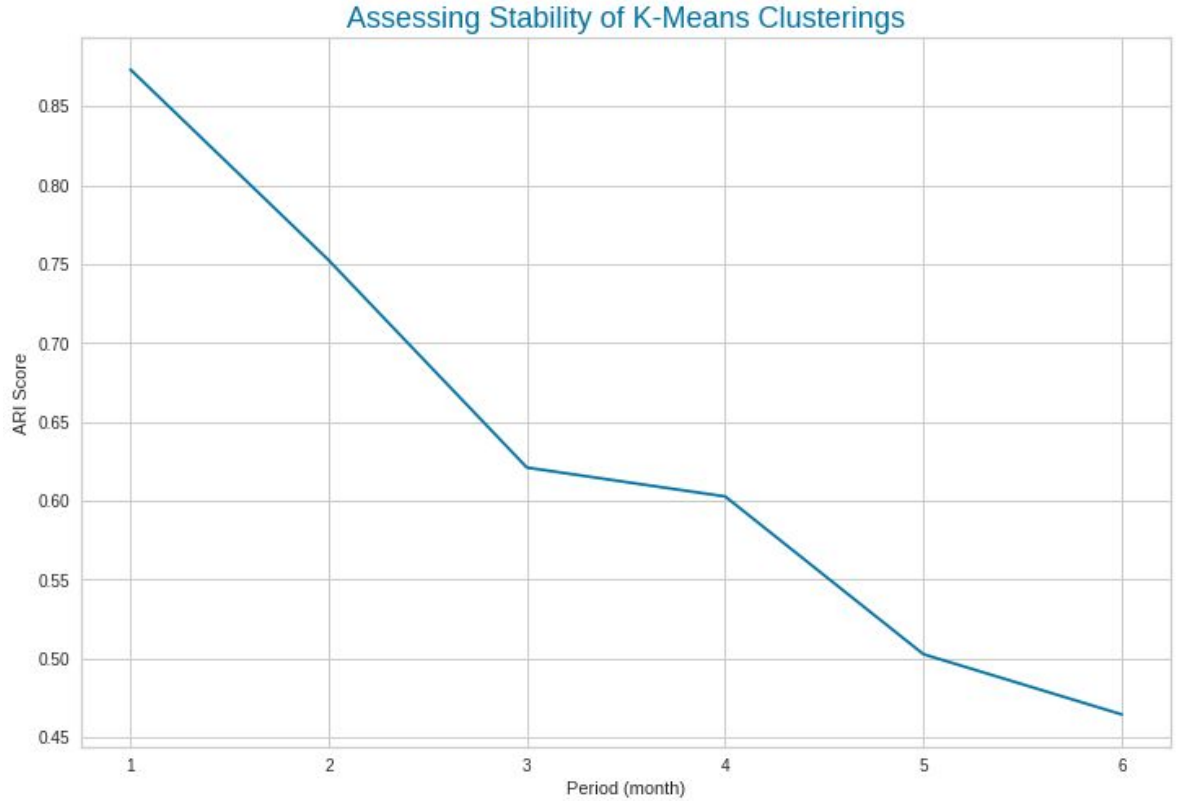
➢ Quand est-ce que l'on renouvelle l'entraînement du modèle ?

➢ Algorithme d'analyse de stabilité

- **Division de la base de donnée :**
  - ***B0 :*** *Base de donnée initiale (12 mois)*
  - ***B1 … BN :*** *Base de données subséquentes (12 mois + k mois).*
- **Clustering selon chaque sous-base de données :**
  - ***C0*** *: clustering initiale livré au client*
  - ***C1 … cN :*** *Clustering subséquents*
- **Comparaison des clusterings subséquents**

Analyse de stabilité

➢ Maintenance préconisé à horizon 3 mois



Sommaire

Conclusion

- **Comparaisons de 3 méthodes de clustering automatiques :**
  - **K-means (k=3)** le plus adapté à la **structure** de nos données

- **Axes d'amélioration :**
  - Connecter des données où il existe une **fréquence (RFM)**
  - **Nouveaux indicateurs catégoriels** pourraient être ajoutés pour raffiner l'analyse (k-prototype)