Segmentez des clients d'un site de e-commerce

PLAN

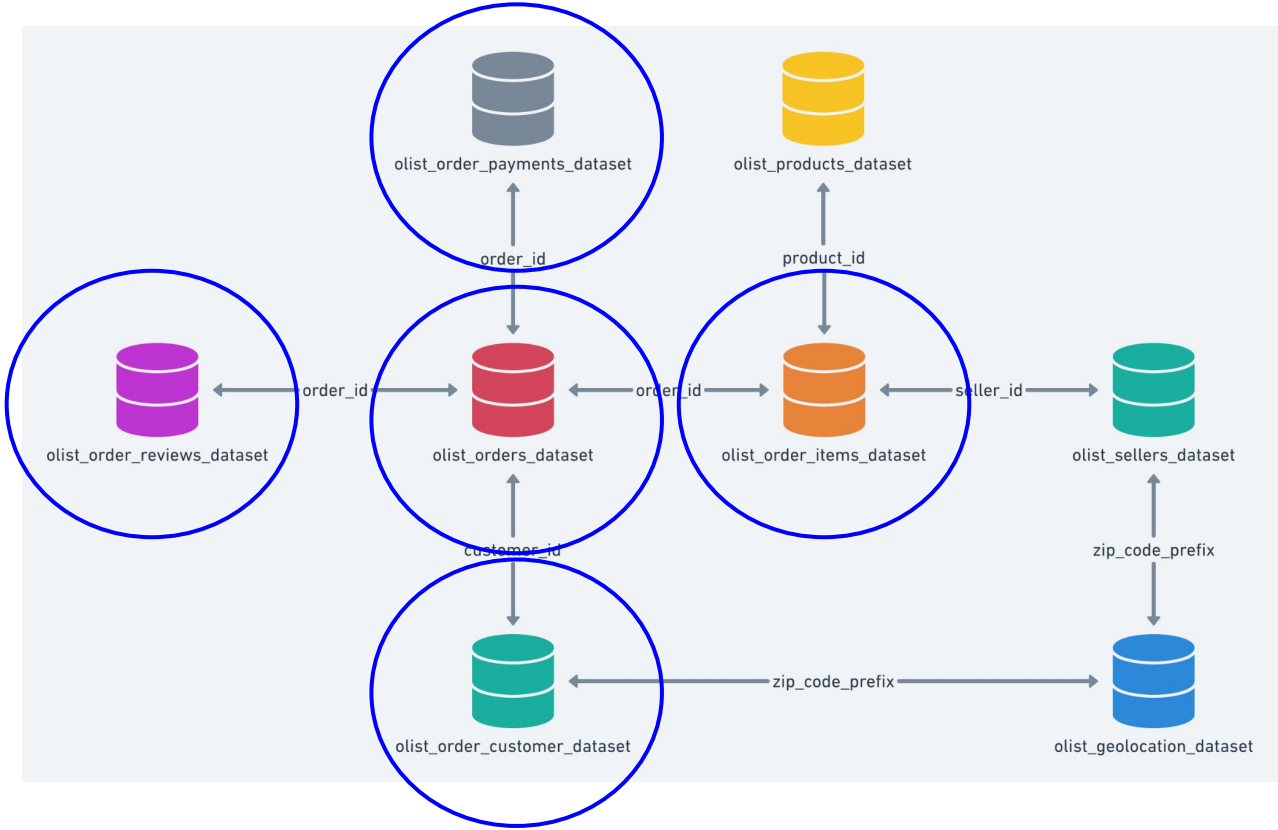
1/ Présentation de la problématique
2/ Présentation des jeux de données
3/ Exploration du jeu de données
4/ Méthodes de clustering
5/ Stabilité du clustering
6/ Conclusion
7/ Suite du projet

Consultant entreprise brésilienne qui propose une solution de vente sur les marketplaces en ligne : **Olist**

Travailler avec équipes e-commerce :

- Segmentation des clients
- Comprendre les types d'utilisateurs
- Fournir description actionnable
- Proposition de contrat de maintenance

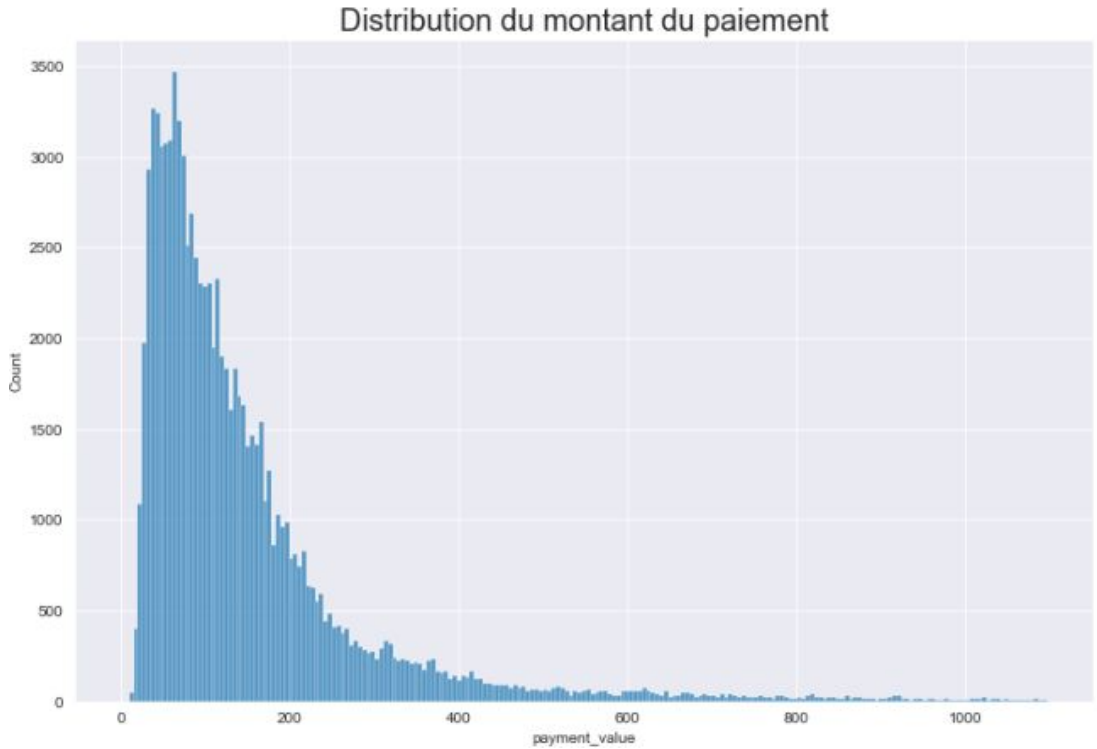
Présentation du jeu de données



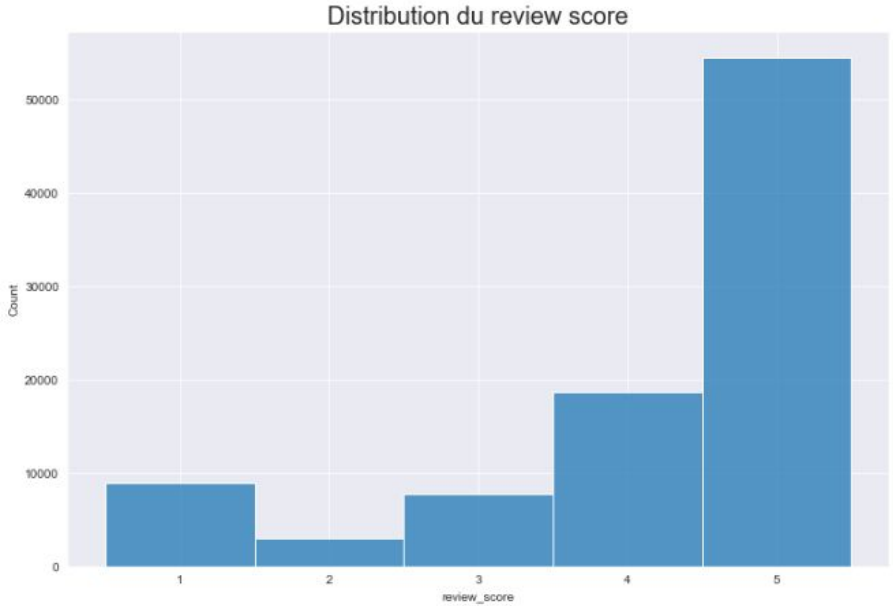
- **Assemblage**
  Base de données clients :
  - *Client unique*
  - *Date dernier achat*
  - *Récence*
  - *Nombre de commandes*
  - *Dépenses totales*
  - *Review score*
  - *Nombre de produits commandés*
  - *Délai de livraison*
  - *Type de paiement*
  - *Nombre de paiement*

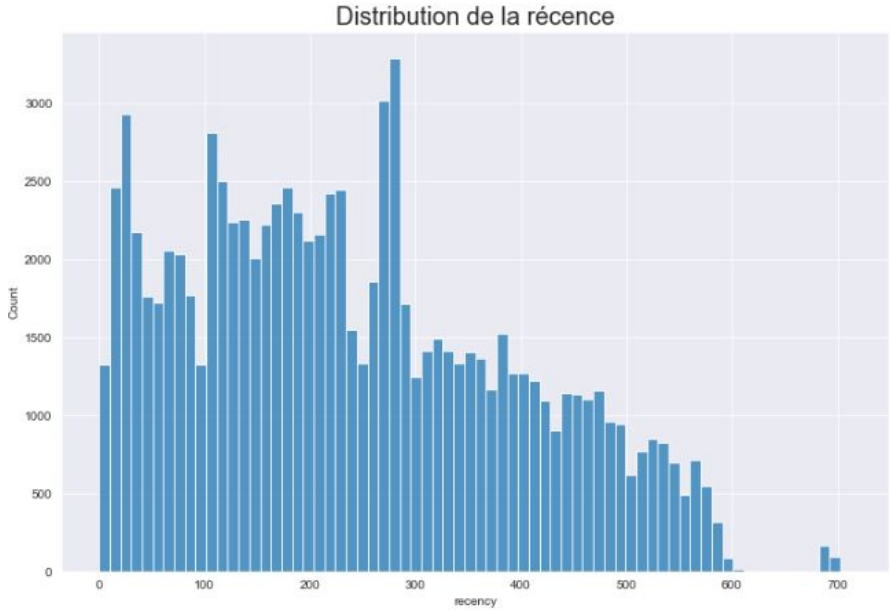
- **Nettoyage**

```
Etendue du paiement : 13664.08
Valeur max. du paiement : 13664.08
Valeur min. du paiement : 0.0
```



Exploration du jeu de données

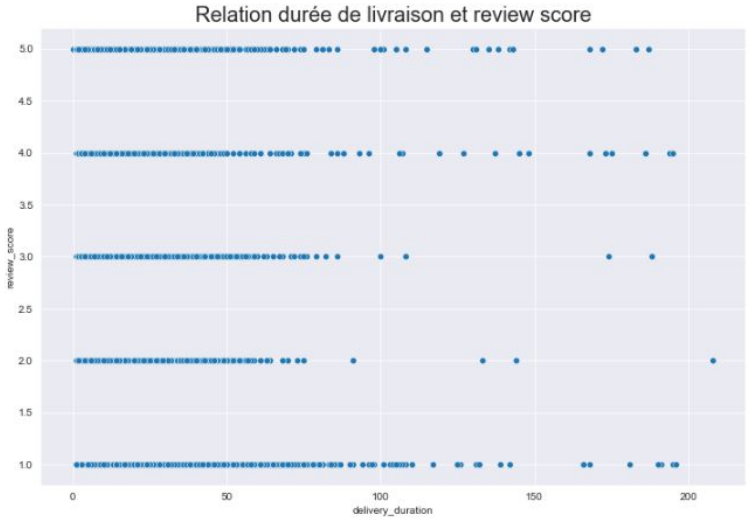


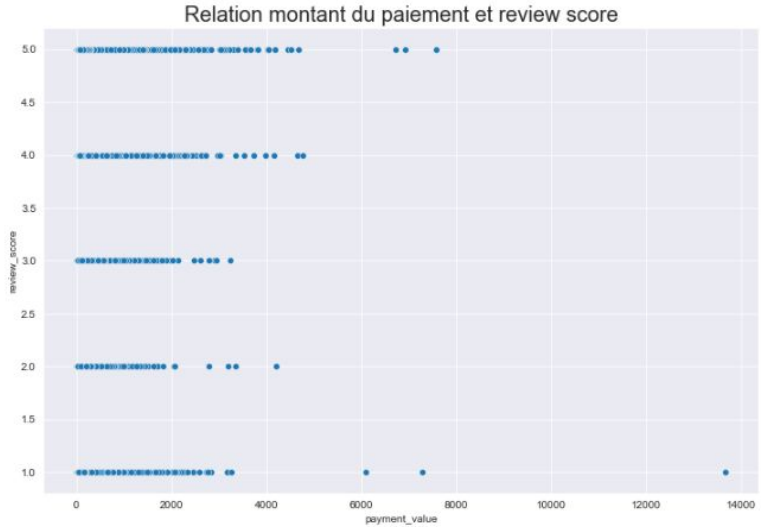


Exploration du jeu de données

```
Etendue du délai de livraison : 210.0
Valeur max. du délai de livraison : 210.0
Valeur min. du délai de livraison : 0.0
```

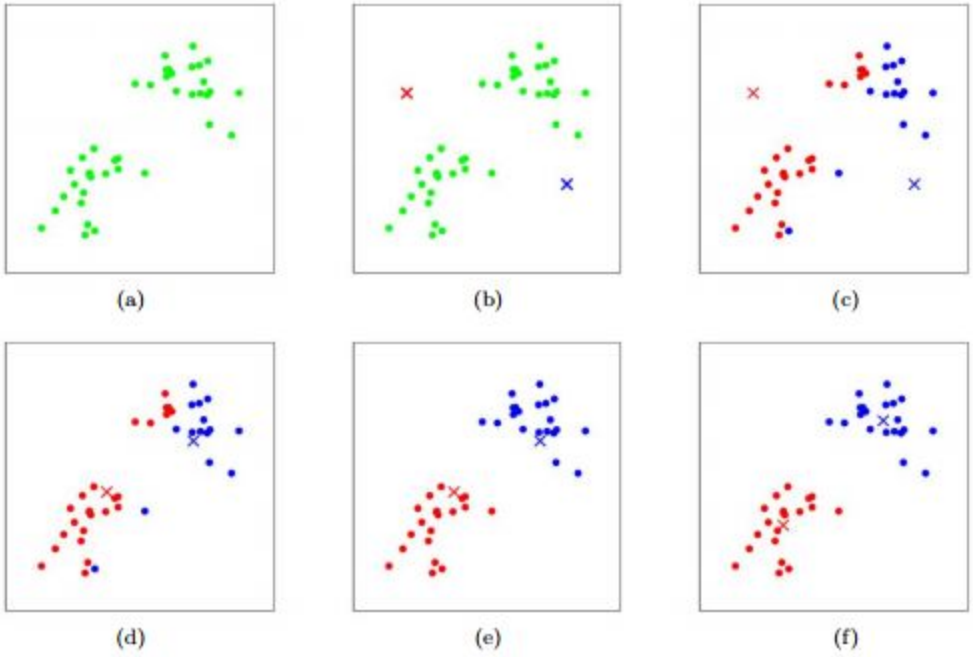
Exploration du jeu de données






payment_value
60
50
40
30
20
10
0
number_orders
70
60
50
40
30
20
10
0
recency
3
2
1
0
−1

- 3 méthodes de clustering : KMeans, Classification Hiérarchique, DBSCAN
- Segmentation technique
- Segmentation métier
- Caractérisation et identification des clusters
- Variables : RFM + Review score + Délai de livraison (5 variables)


(a)
(b)
(c)
(d)
(e)
(f)

- KMeans

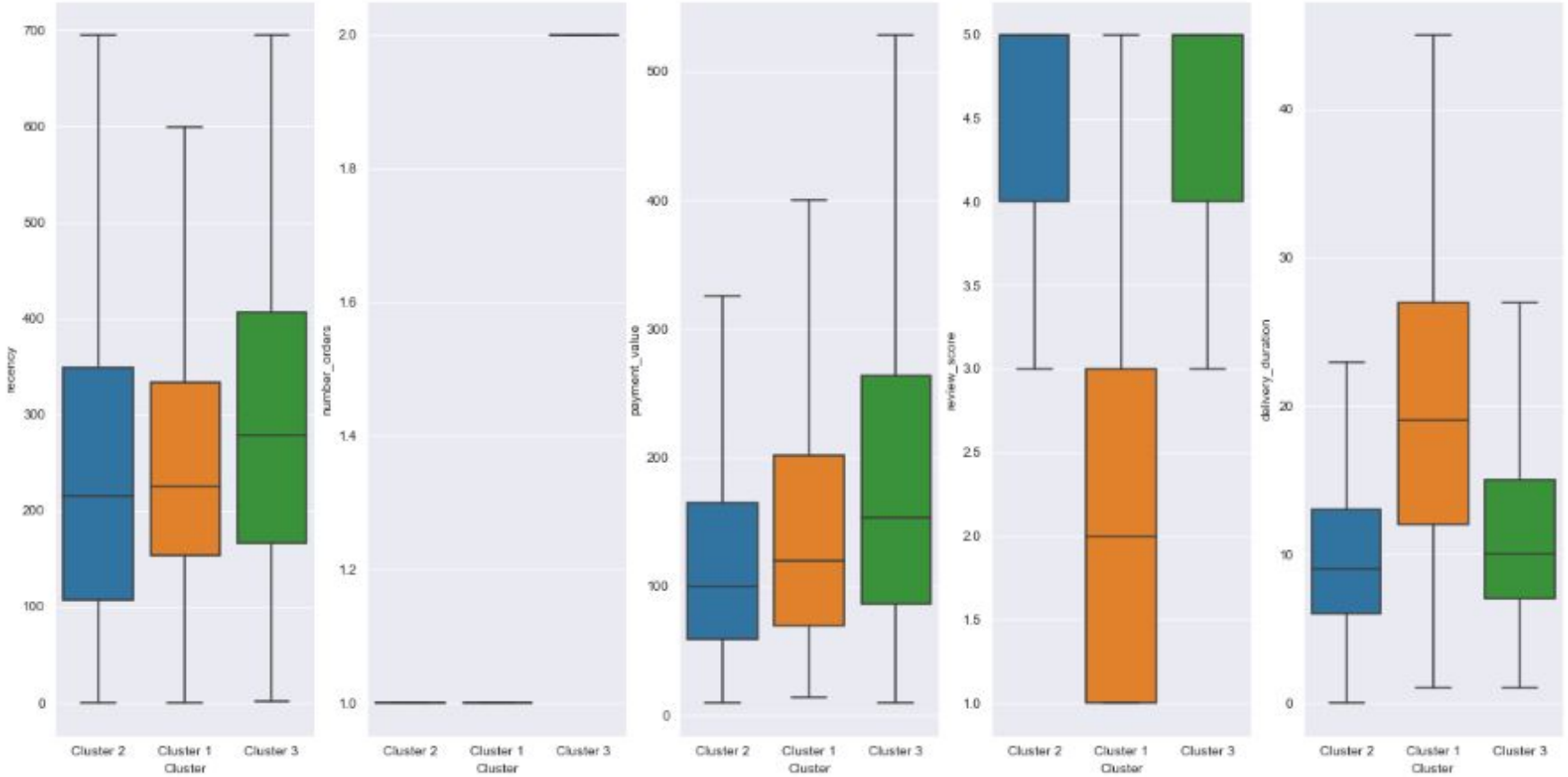
```
Features dans KMeans : ['recency', 'number_orders', 'payment_value', 'review_score', 'delivery_duration']
Nombre de clusters : 3
Score de silhouette : 0.3554525091453411

Features dans KMeans : ['recency', 'number_orders', 'payment_value', 'review_score', 'delivery_duration']
Nombre de clusters : 4
Score de silhouette : 0.2644541437966453

Features dans KMeans : ['recency', 'number_orders', 'payment_value', 'review_score', 'delivery_duration']
Nombre de clusters : 5
Score de silhouette : 0.2867329916981656

Features dans KMeans : ['recency', 'number_orders', 'payment_value', 'review_score', 'delivery_duration']
Nombre de clusters : 6
Score de silhouette : 0.29632119188577266
```

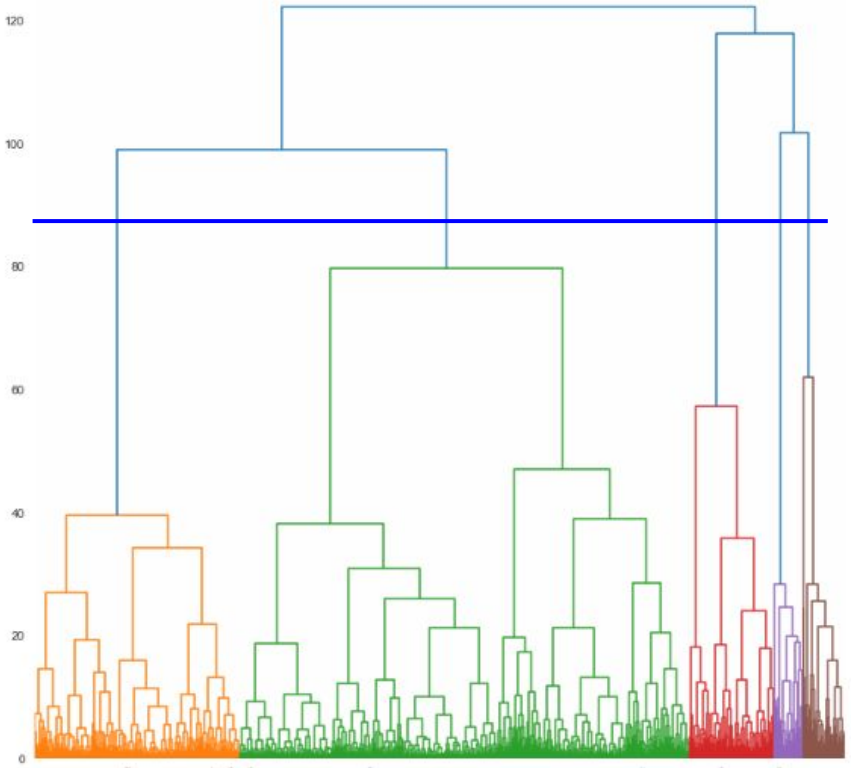
Méthodes de clustering



Point de vue métier, 3 clusters semblent être le bon paramètre pour distinguer les clients :

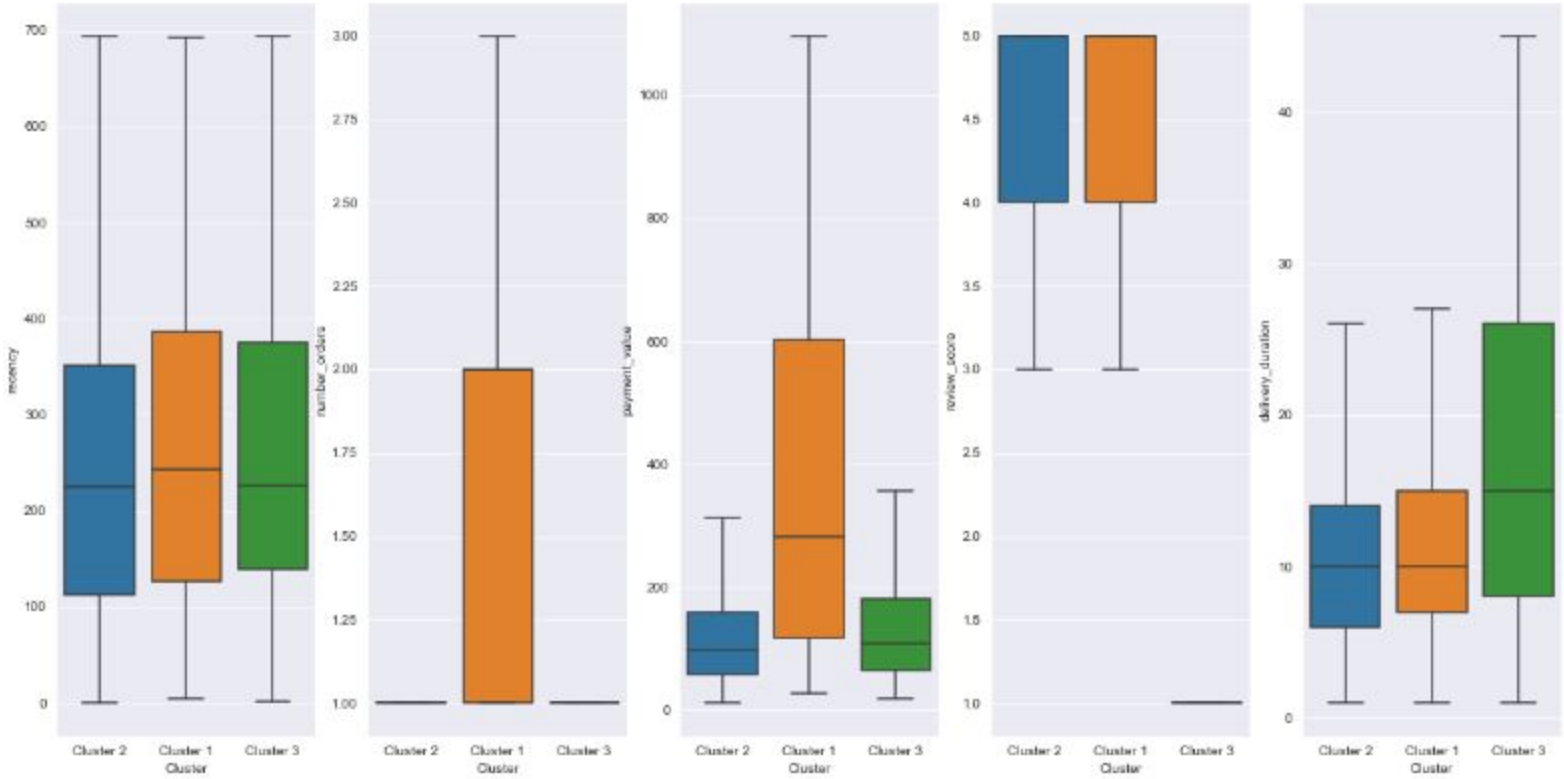
- **Cluster 1 :** attend le plus et note le moins bien (**clients non satisfaits**)
- **Cluster 2 :** plus récent, paie le moins, note mieux et attend le moins (**nouveaux clients satisfaits**)
- **Cluster 3 :** moins récent, paie plus, note mieux et achète plus (**anciens clients**)

- <u>Classification Ascendante Hiérarchique</u>



5 clusters
Score de silhouette = 0.24
(inférieur au KMeans)

Méthodes de clustering



Point de vue métier, 3 clusters semblent être le bon paramètre pour distinguer les clients (meilleur score de silhouette, 0.38) :

- **Cluster 1 :** achète le plus, paie le plus et note le mieux (**très bons clients**)
- **Cluster 2 :** paie le moins, note le mieux (**bons clients à fidéliser**)
- **Cluster 3 :** paie le moins, note le moins bien et attend le plus (**clients insatisfaits**)

Cependant CAH réalisée sur un échantillon : méthode peu adapté pour un grand fichier de clients

- DBSCAN

DBSCAN n'est pas adapté au projet :

Méthode se base sur les densités : hors clients "atypiques" qui paient plus et/ou achètent plus (bons clients) sont classés comme des outliers et n'appartiennent donc à aucun cluster

De plus, méthode réalisée sur un échantillon (peu adapté à un grand fichier client)

Stabilité du KMeans :

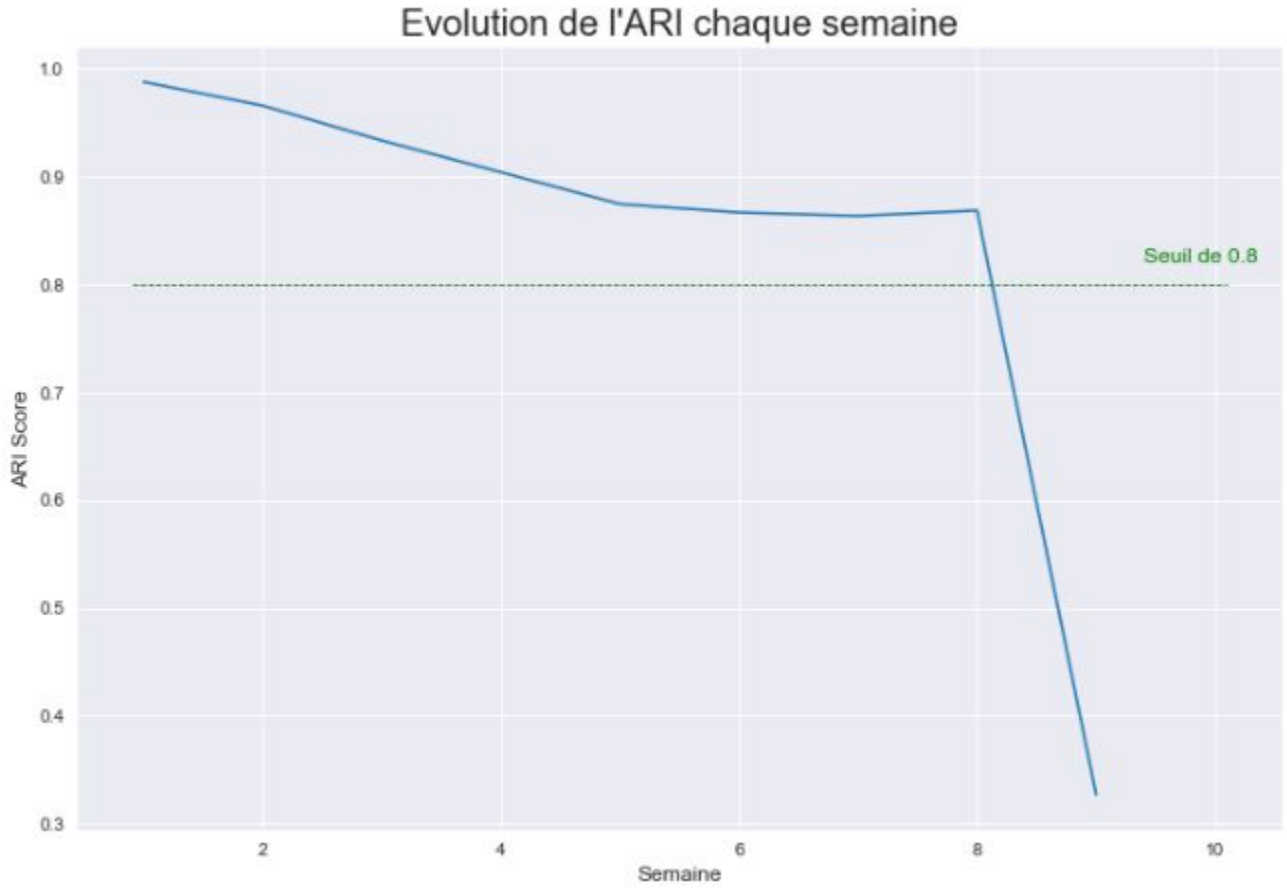
Pas d'une semaine

Comparer la période t0 avec la nouvelle période tn (t0 + n semaine)

Evolution de l'ARI

Seuil 0.8

Stabilité du clustering



- Méthode de clustering KMeans est la plus adaptée à la problématique du projet
- Segmentation métier valide la segmentation technique (pour les données à disposition)
- Stabilité du clustering : mise à jour après 8 semaines

- Présentation des résultats au client (Olist) et en discuter
- Développement d'une API pour le clustering automatisé
- Proposition de maintenance de l'API (mise à jour du clustering)