

OLIST – Segmentation de clients

▶ Classification de
la clientèle



Préambule

- OLIST
 - Entreprise brésilienne
 - Solution de vente sur les marketplaces en ligne
 - Améliorer les campagnes de communication
- Objectif
 - Etudier le comportement des clients
 - Réaliser une segmentation des clients
 - Développement d'une modélisation
 - Maintenance du modèle

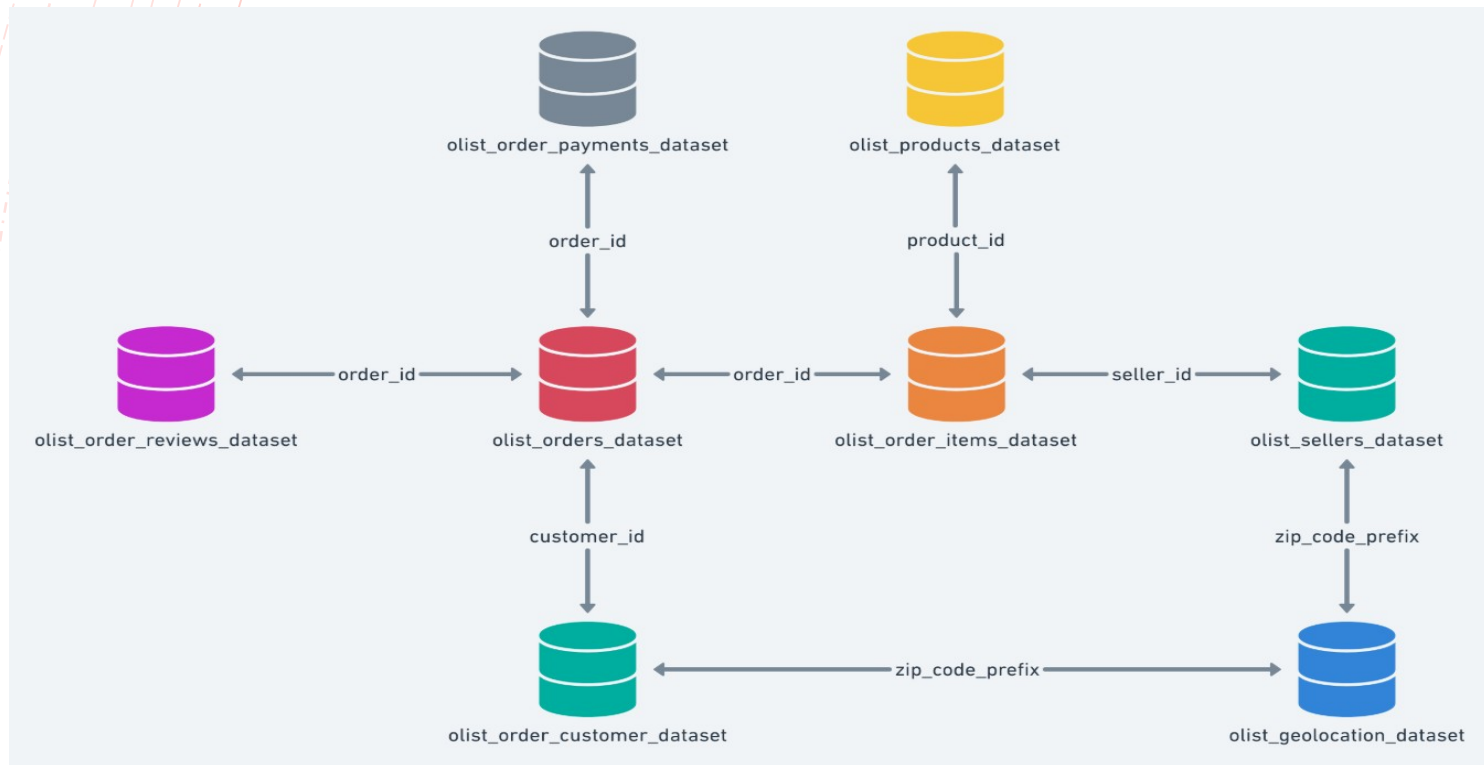


Sommaire

- Présentation du jeu de données
- Nettoyage
- Feature engineering
- Exploration
- Modélisation
- Simulation maintenance

Présentation des données

- 8 tables dans la bases de données
- Lien important entre les tables
- Tri selon la pertinence




Architecture de la base de données



Nettoyage des données

- Jointure des tables
 - Unique dataframe
- Suppression données inutiles
 - Descriptions produits
 - Expédition
 - Paiement (hors somme)
 - Commentaires (hors note)
- Commandes livrées
- Duplicata des numéros de commande

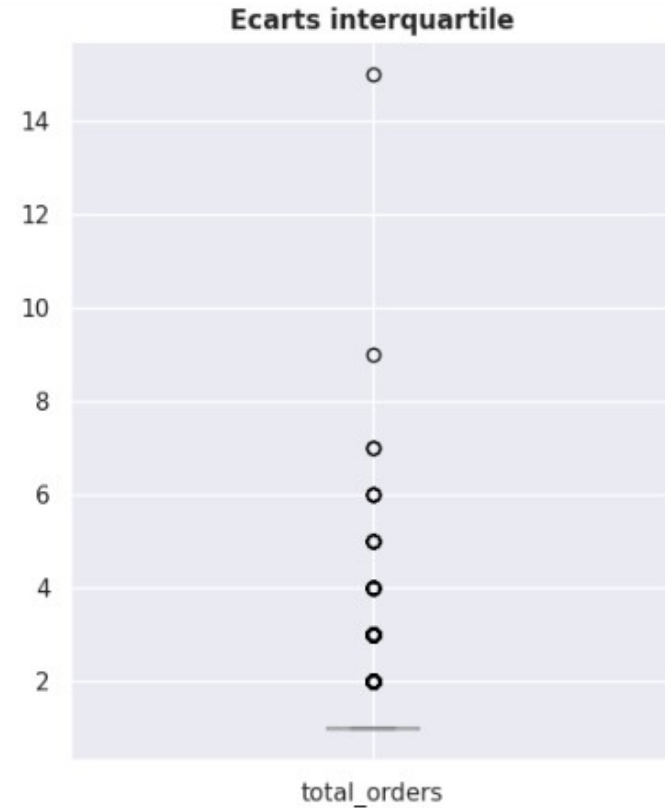
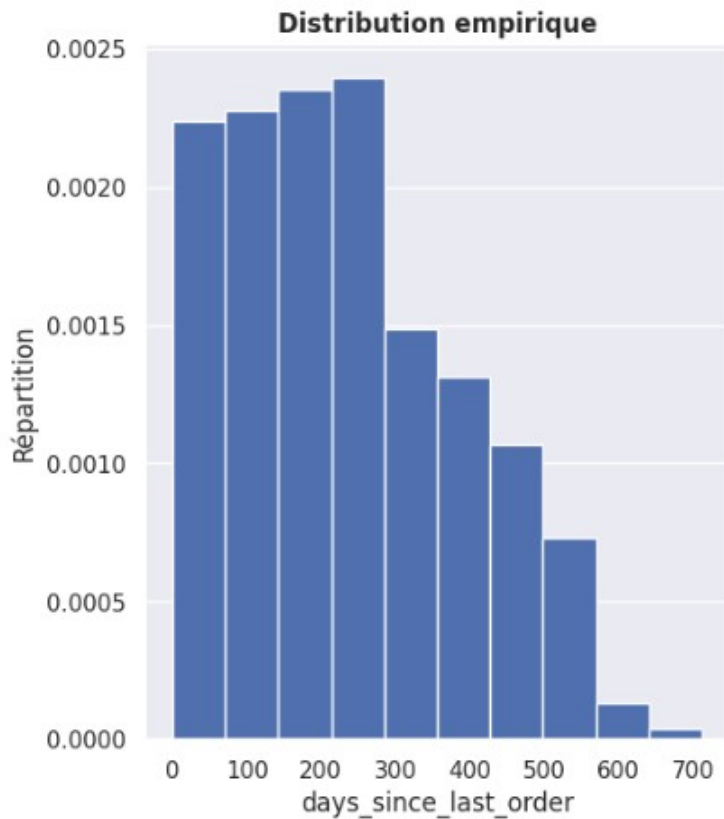


Feature engineering

- Ancienneté commande
 - Date la plus récente du dataframe – Date de la commande
- Fréquence
 - Somme des commandes par client
- Montant
 - Somme des valeurs de commande par client
- Note moyenne
 - Moyenne des commentaires par client

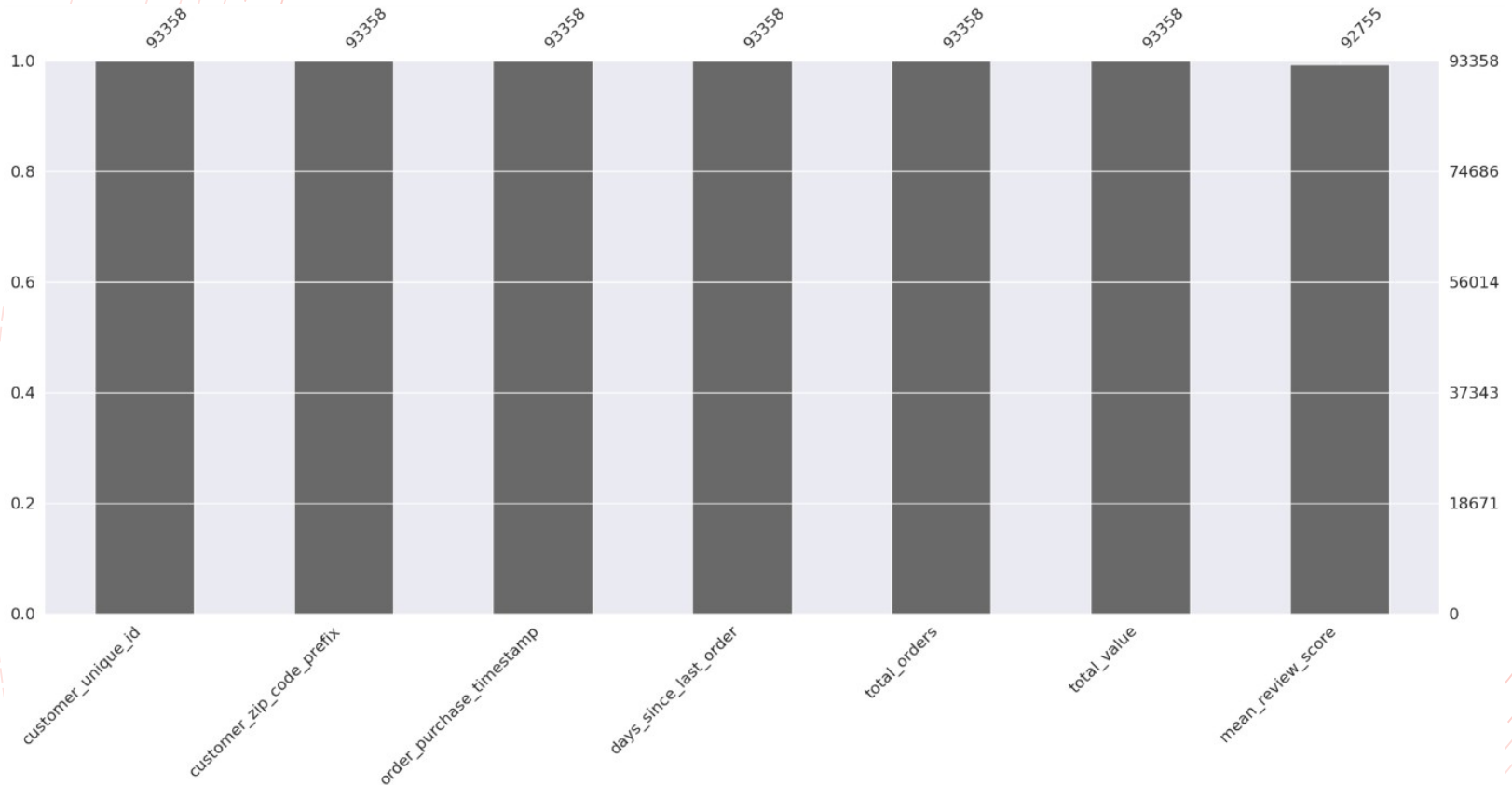
Analyse exploratoire

- Distributions non normales
- Présence d'outliers



Valeurs manquantes

- Imputation mean_review_score par la moyenne



Nombre de valeurs manquantes par feature

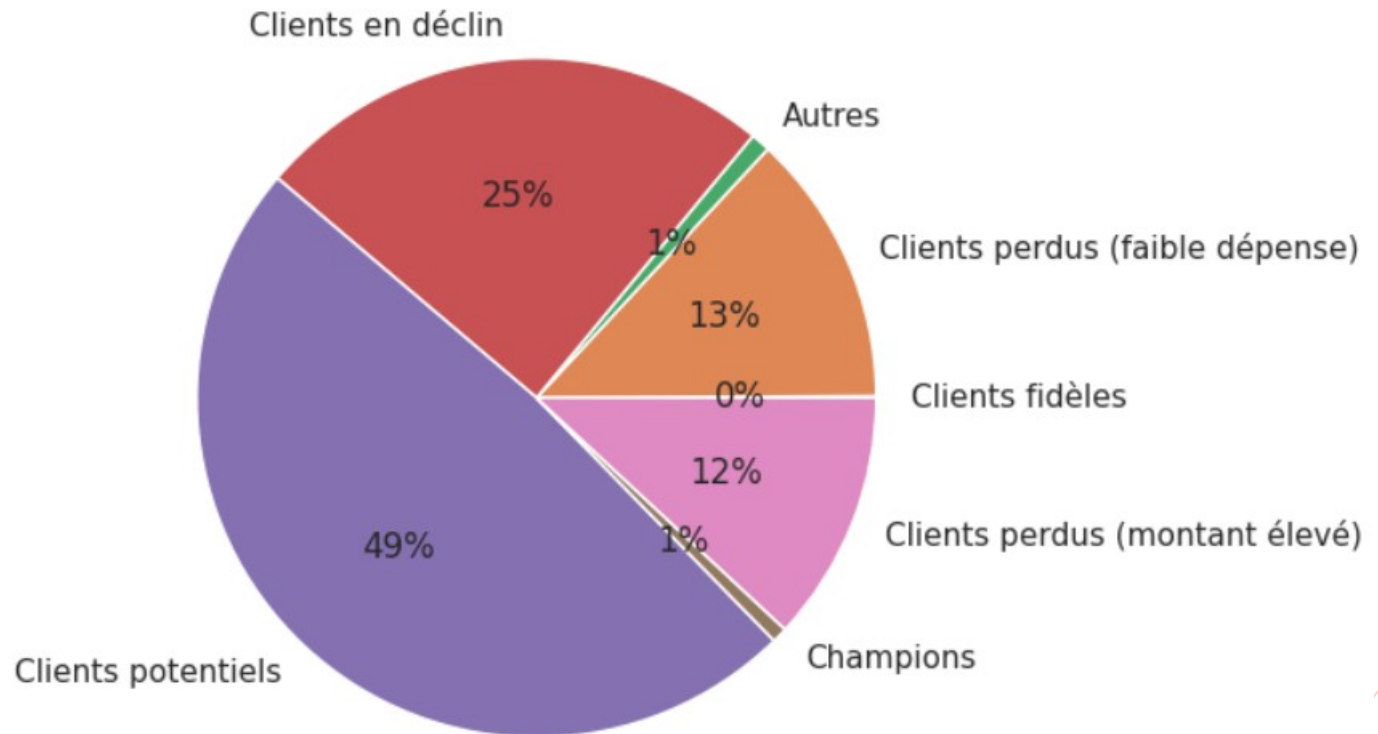
Segmentation RFM - Baseline

- Score pour chaque composante
 - Méthode des quantiles
 - Adaptation pour la fréquence
- Classification en fonction des scores de chaque composante

Segment	Récence (R)	Fréquence (F)	Montant (M)	Description
Champions	4	2 à 4	≥ 3	Clients très récents, actifs et avec un montant élevé.
Clients fidèles	4	2 à 4	2 à 3	Clients récents, plusieurs commandes, montant modéré.
Clients potentiels	3 ou 4	1	≥ 1	Nouveaux clients, faible fréquence mais potentiel d'achat.
Clients en déclin	2	1 à 2	≥ 1	Moins récents, faible fréquence mais montant modéré à élevé.
Clients perdus (faible dépense)	1 ou 2	1 à 2	≤ 2	Clients peu récents, faible fréquence et faible montant.
Clients perdus (montant élevé)	1 ou 2	1 à 2	> 2	Clients peu récents, faible fréquence mais montant élevé.

Segmentation RFM - Baseline

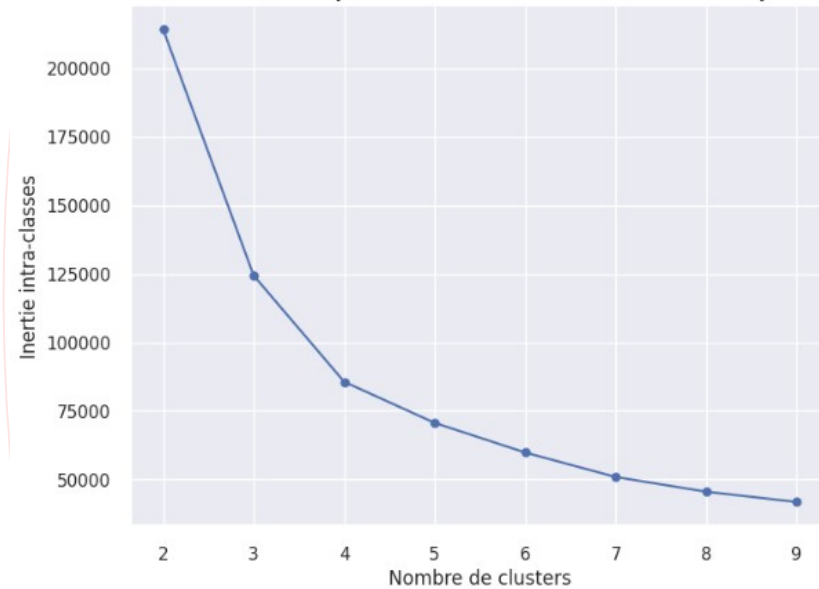
■ Résultats



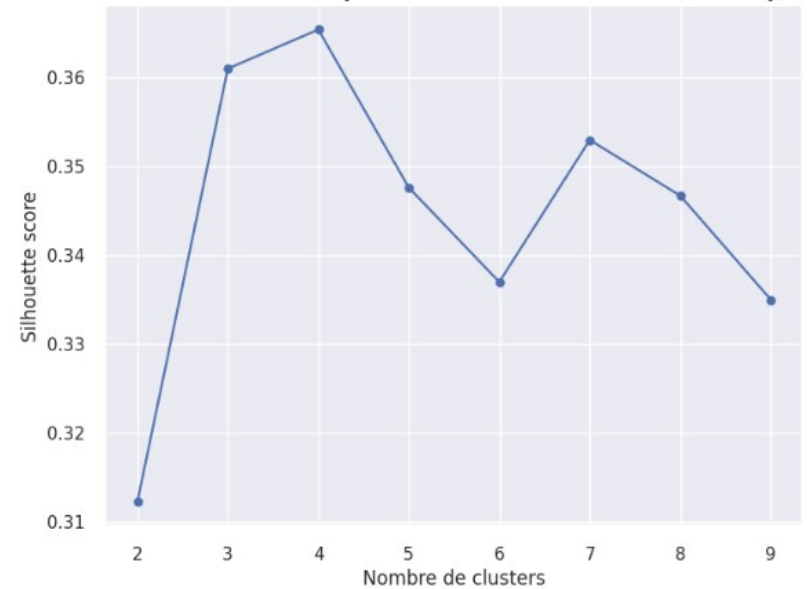
Kmeans avec 3 variables

- Choix du nombre de clusters
 - Méthode du coude
 - Méthode de la silhouette

Méthode du coude pour recherche du nombre de clusters optimal

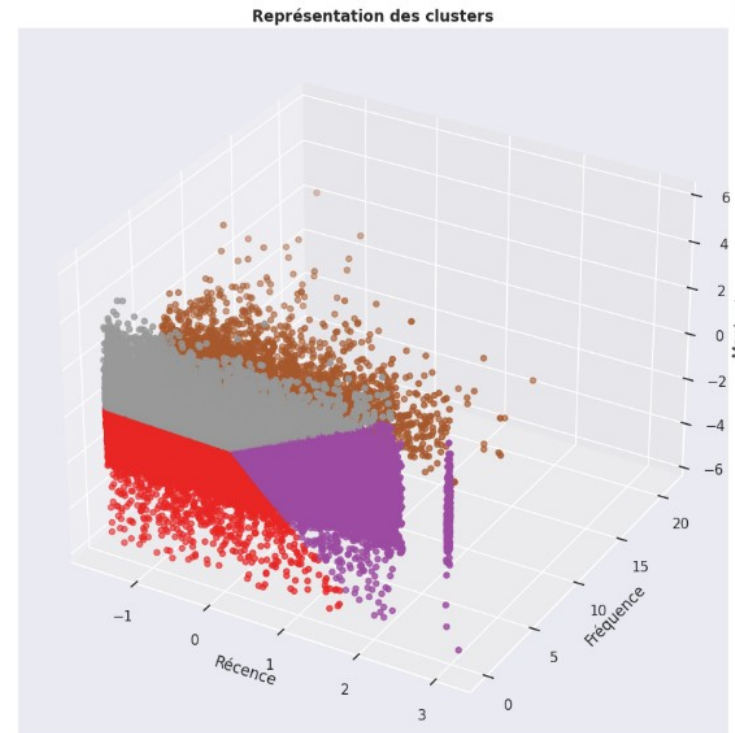
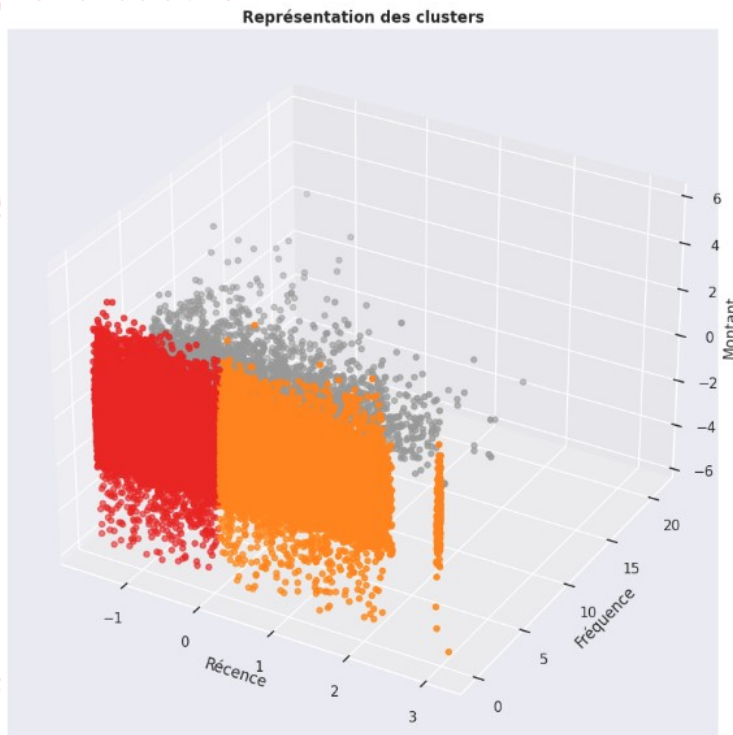


Méthode de la silhouette pour recherche du nombre de clusters optimal



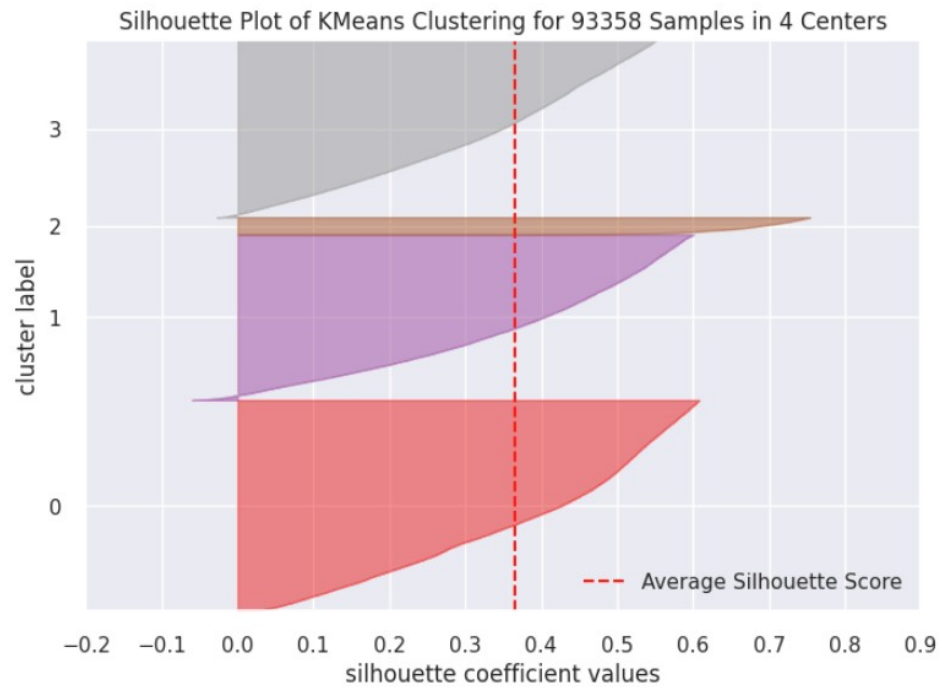
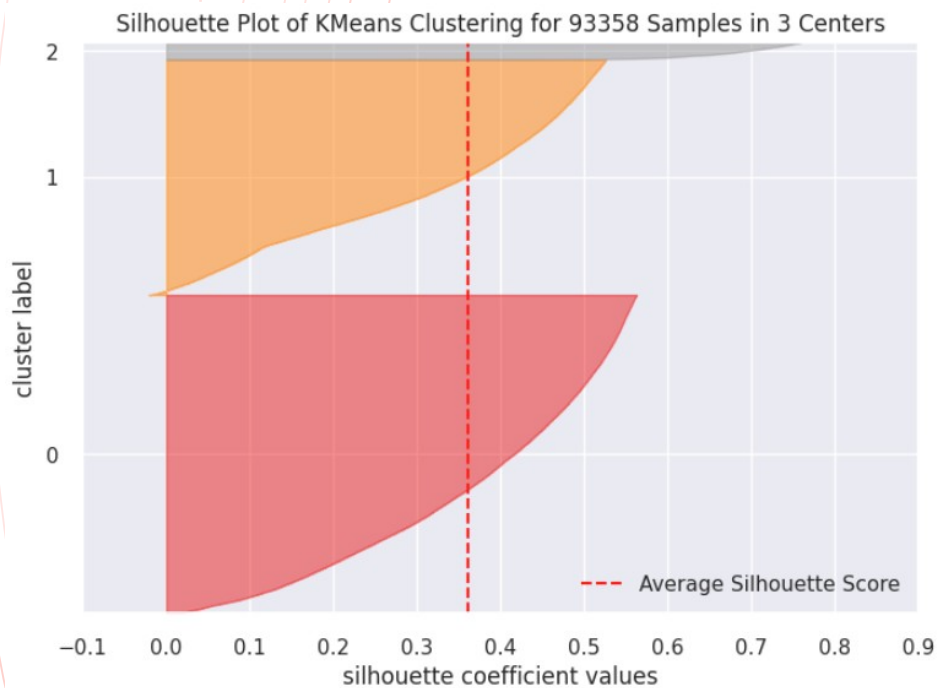
Kmeans avec 3 variables

- Représentation des clusters en 3D



Kmeans avec 3 variables

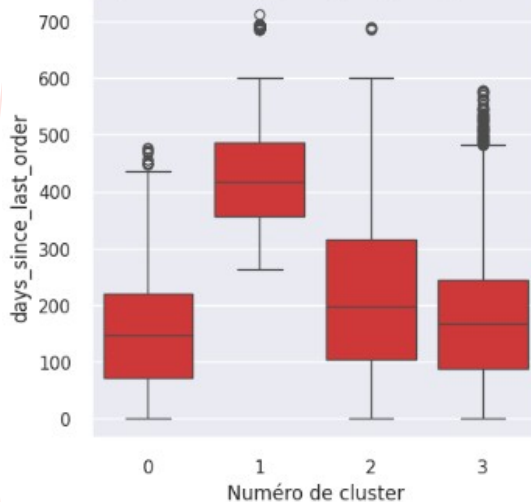
- Représentation des scores de silhouette par cluster



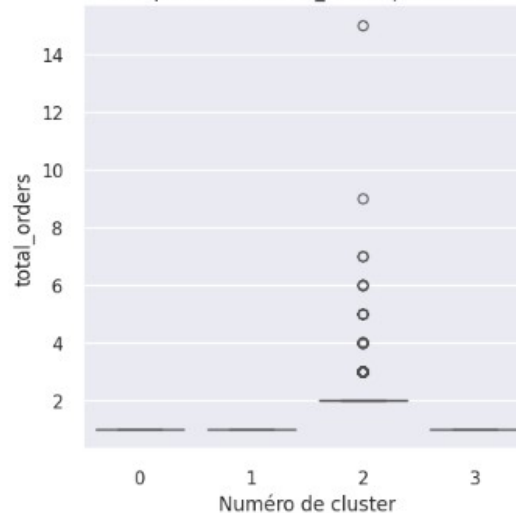
Kmeans avec 3 variables

- Analyse métier
 - 1 box plot pour chaque cluster

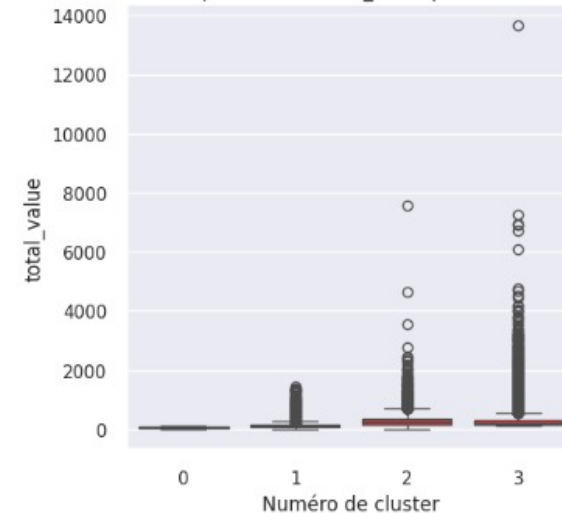
Boxplot feature days_since_last_order par cluster



Boxplot feature total_orders par cluster



Boxplot feature total_value par cluster

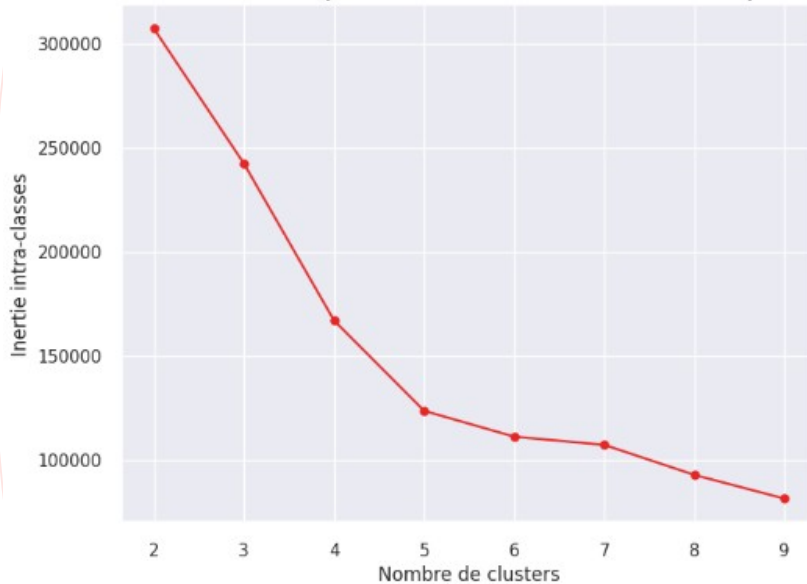


- **Premier cluster** : clients moins récents, faible fréquence et montant => **clients en déclin**
- **Deuxième cluster** : clients peu récents avec une faible fréquence et un faible montant => **clients perdus**
- **Troisième cluster** : clients très récents, actifs, avec un montant élevé => **champions**
- **Quatrième cluster** : nouveaux clients avec une faible fréquence mais un potentiel d'achat => **clients potentiels**

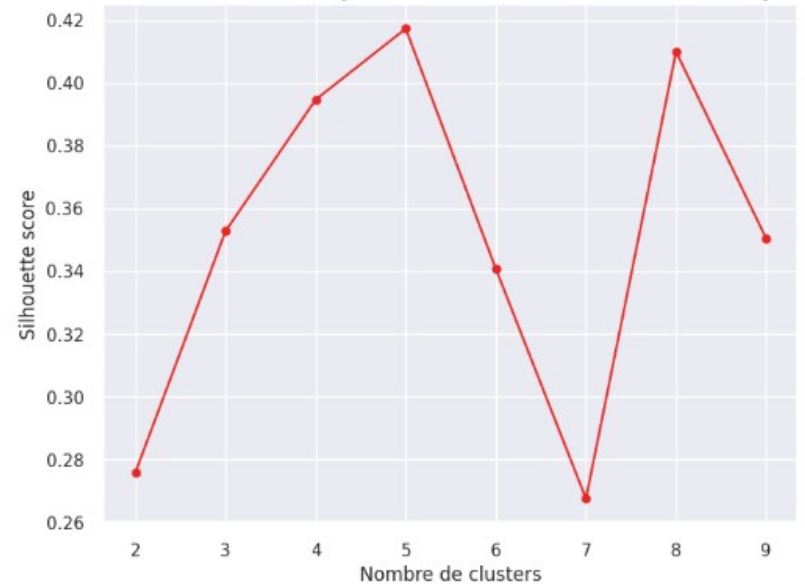
Kmeans avec 4 variables

- Choix du nombre de clusters
 - Méthode du coude
 - Méthode de la silhouette

Méthode du coude pour recherche du nombre de clusters optimal

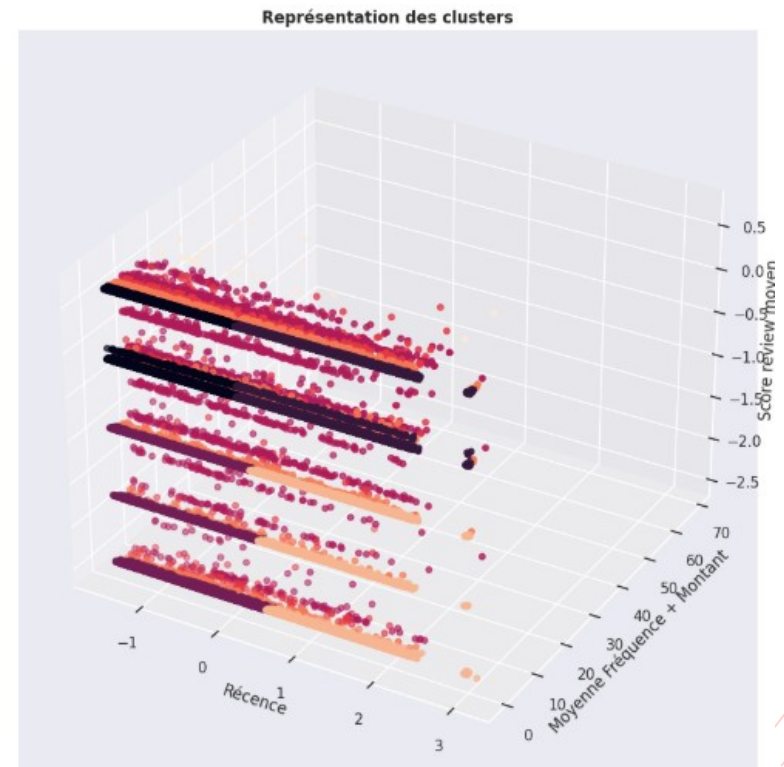
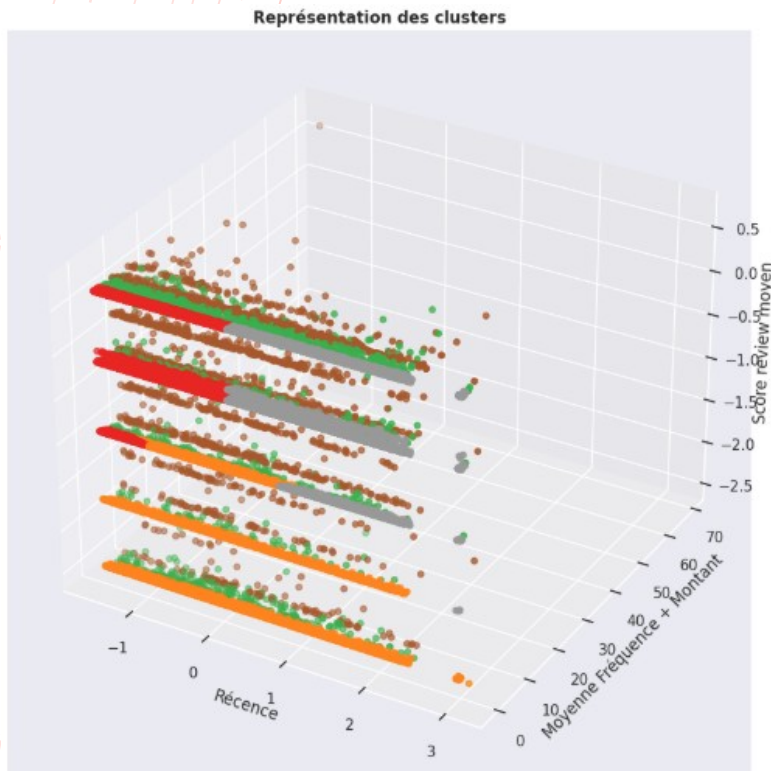


Méthode de la silhouette pour recherche du nombre de clusters optimal



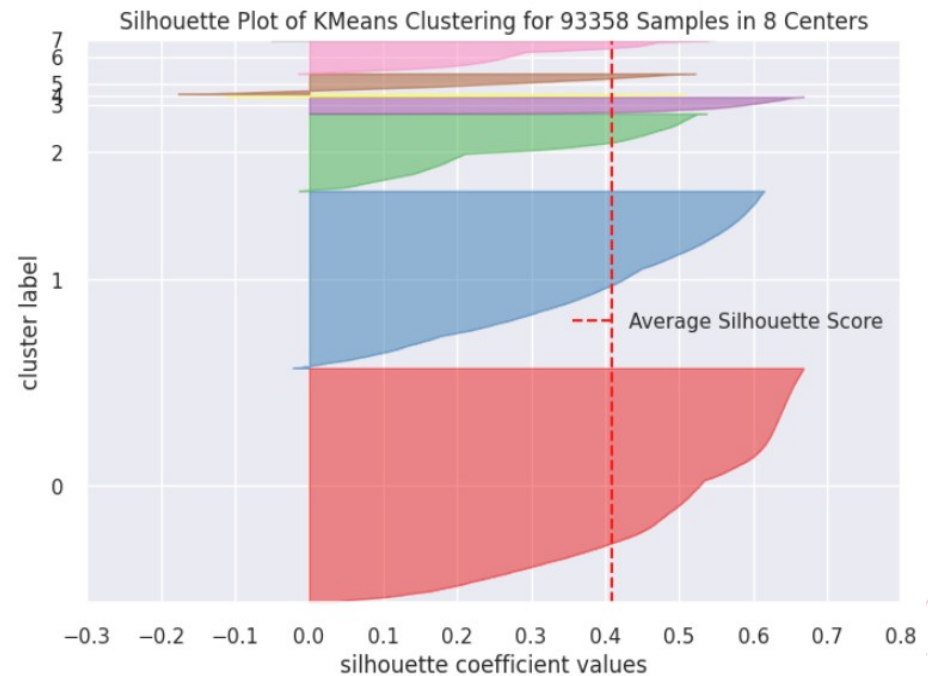
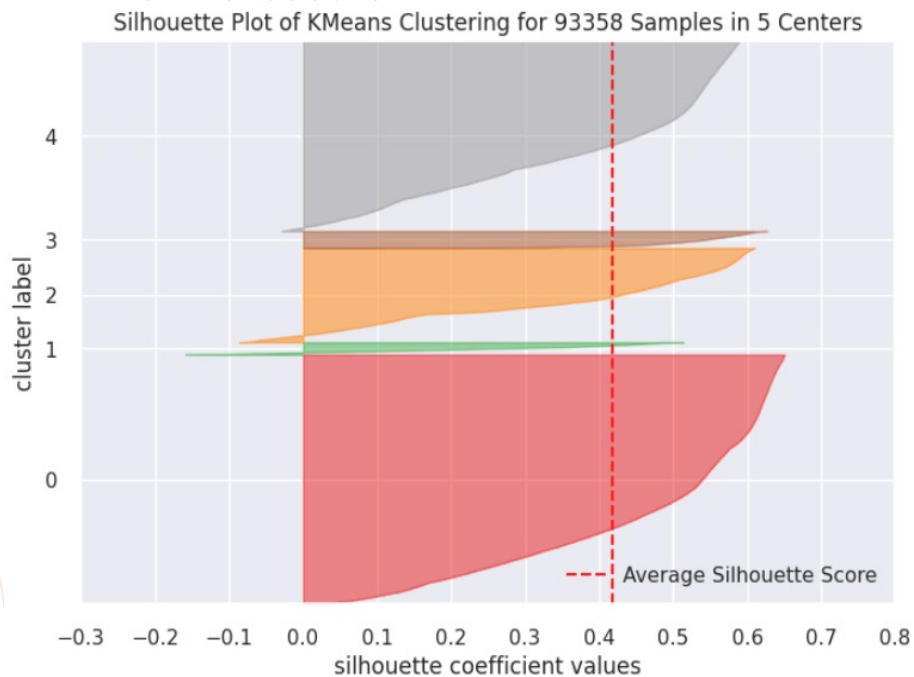
Kmeans avec 4 variables

- Représentation des clusters en 3D



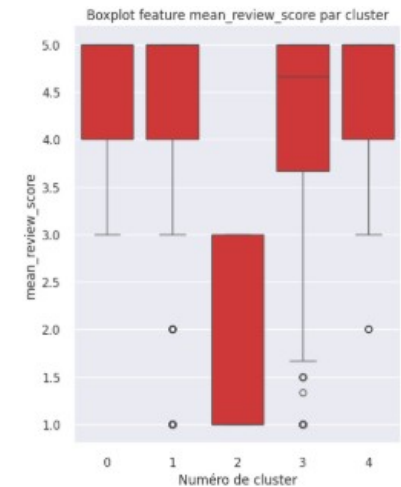
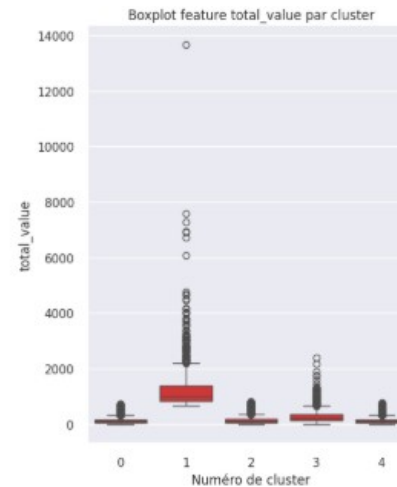
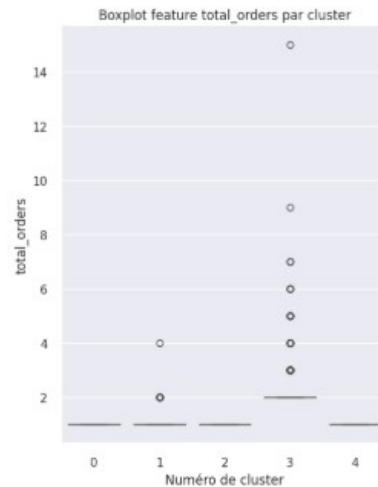
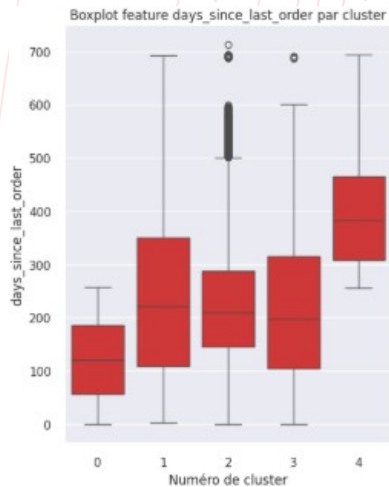
Kmeans avec 4 variables

- Représentation des scores de silhouette par cluster



Kmeans avec 4 variables

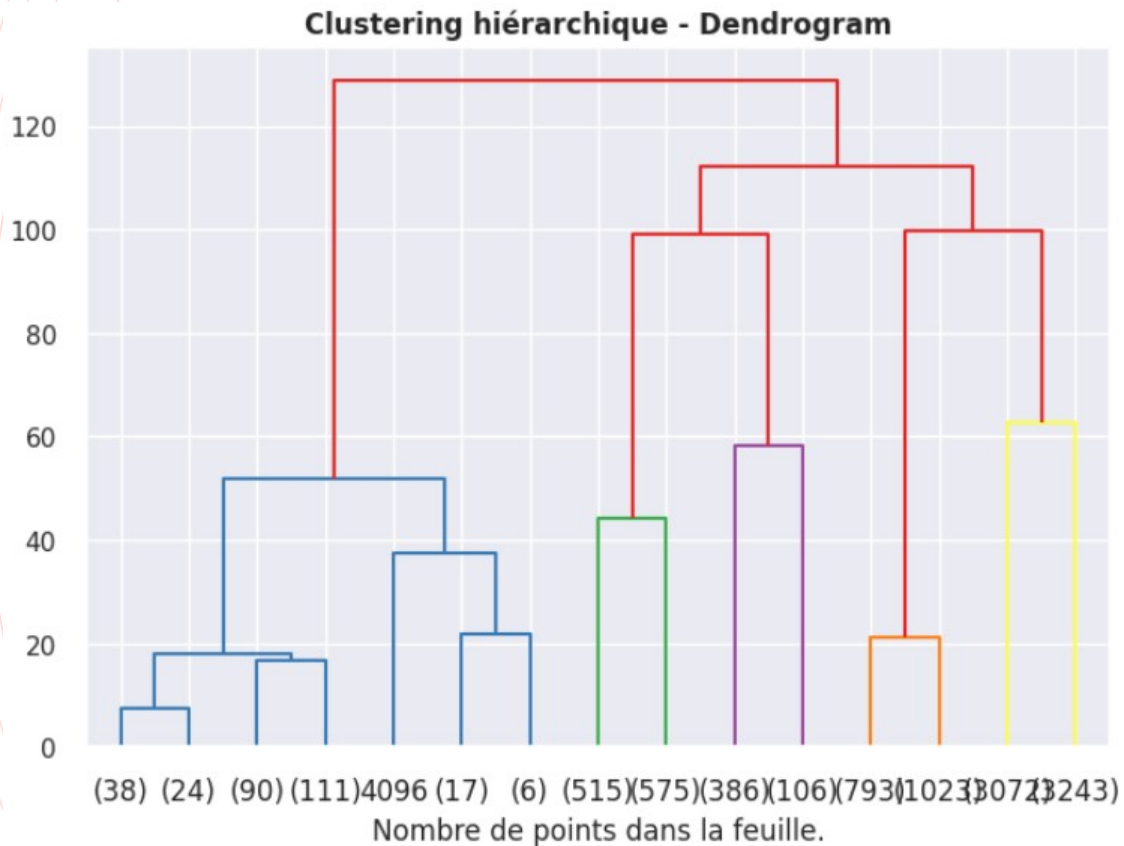
- Analyse métier
 - 1 box plot pour chaque cluster



- **Premier cluster** : clients récents, faible fréquence et montant mais satisfaits => **clients potentiels**
- **Deuxième cluster** : clients moins récents, actifs, avec un montant élevé et satisfaits => **champions**
- **Troisième cluster** : clients moins récents, faible fréquence et montant mais non satisfaits => **clients perdus**
- **Quatrième cluster** : clients moins récents, actifs, avec un montant moyen et plutôt satisfaits => **clients fidèles**
- **Cinquième cluster** : clients peu récents, faible fréquence et montant mais satisfaits => **clients en déclin**

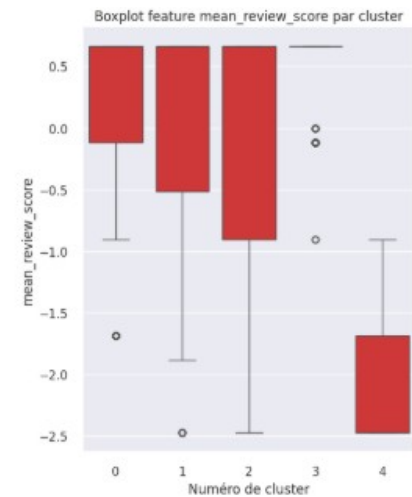
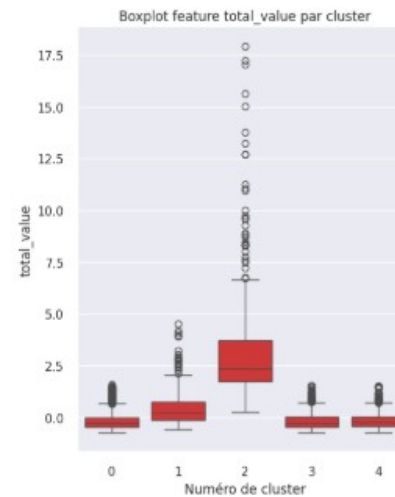
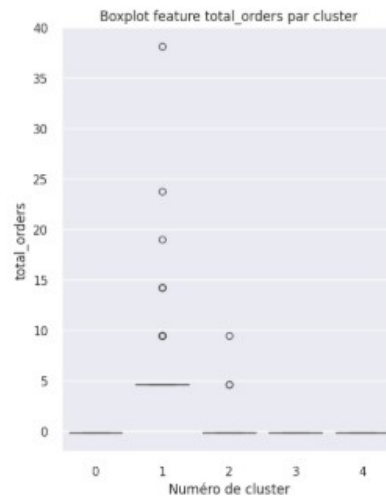
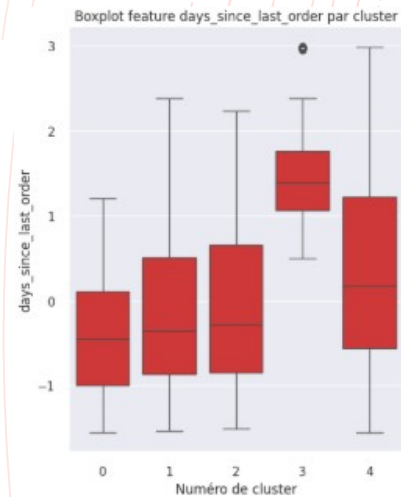
Dendrogram

- Travail sur un échantillon des données
- Score de silhouette moins performant



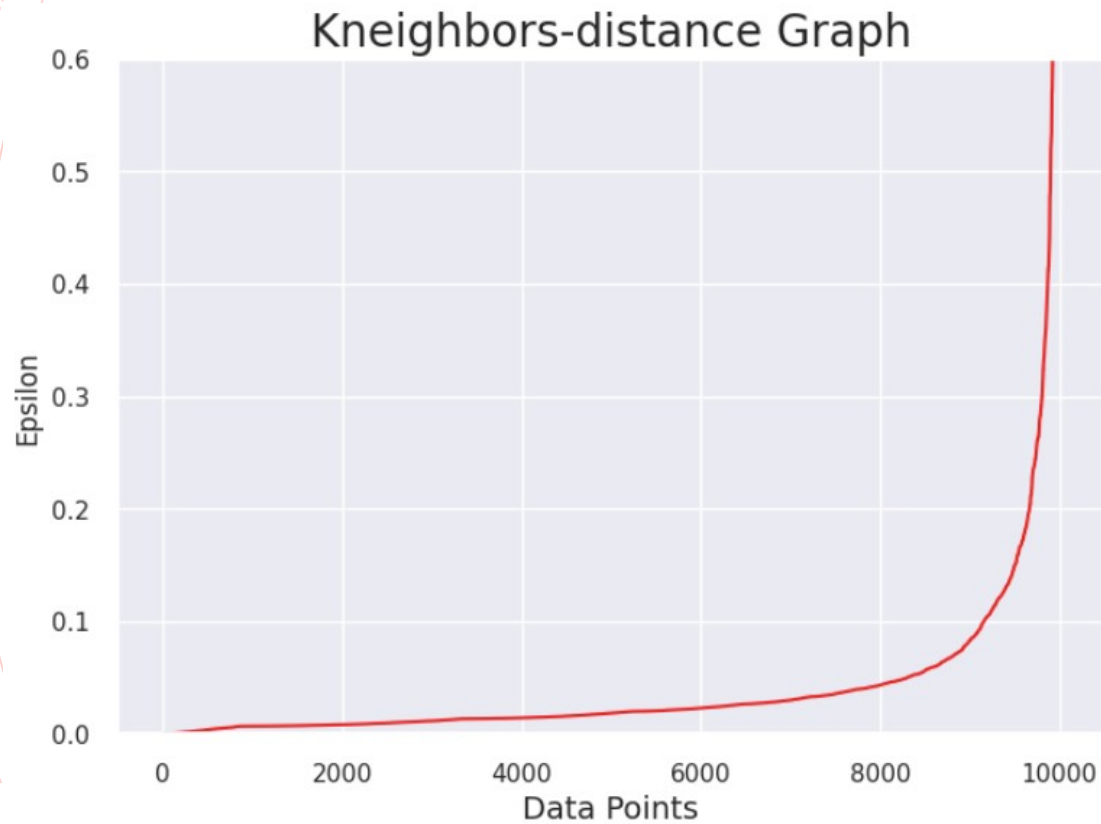
Dendrogram

- Analyse métier
 - 1 box plot pour chaque cluster
 - Moins lisible



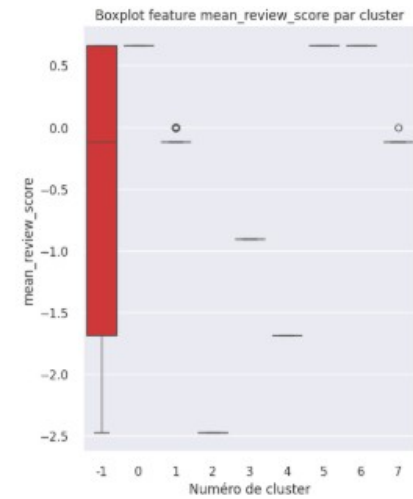
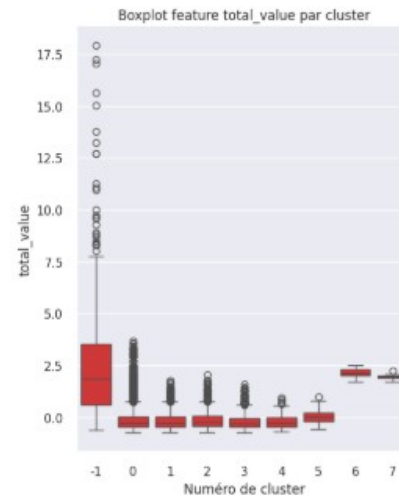
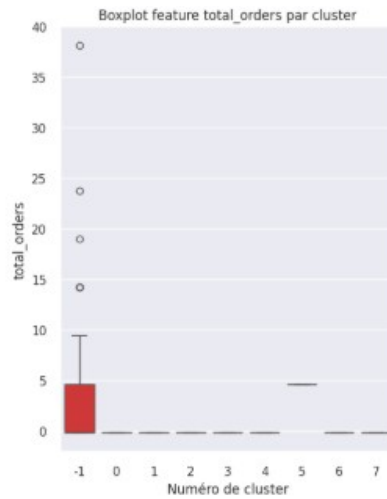
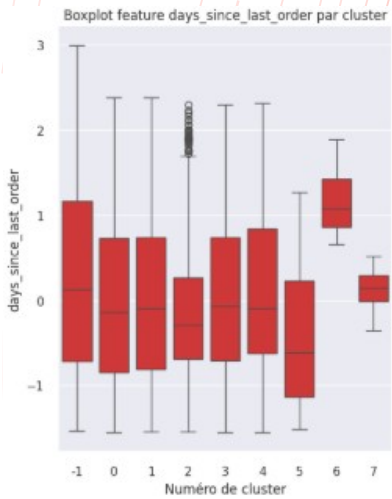
DBSCAN

- Travail sur un échantillon des données
- Score de silhouette non satisfaisant



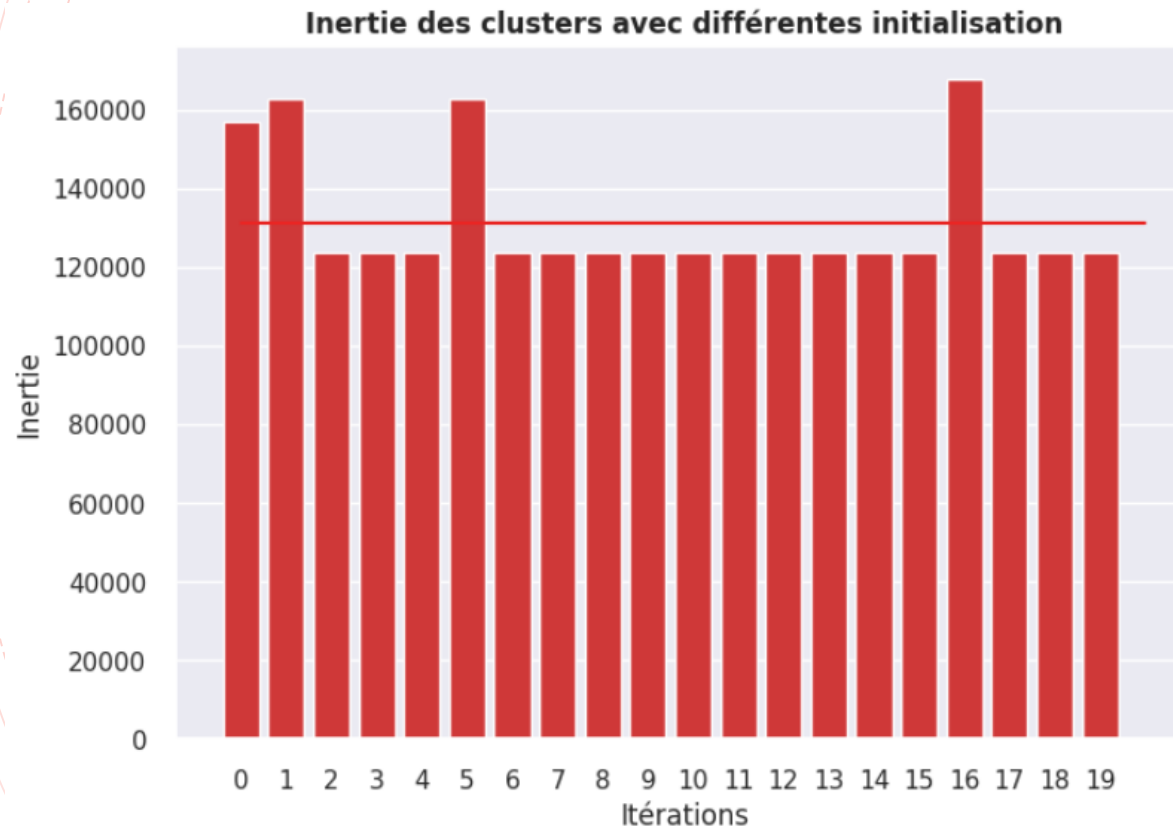
DBSCAN

- Analyse métier
 - 1 box plot pour chaque cluster
 - Peu lisible



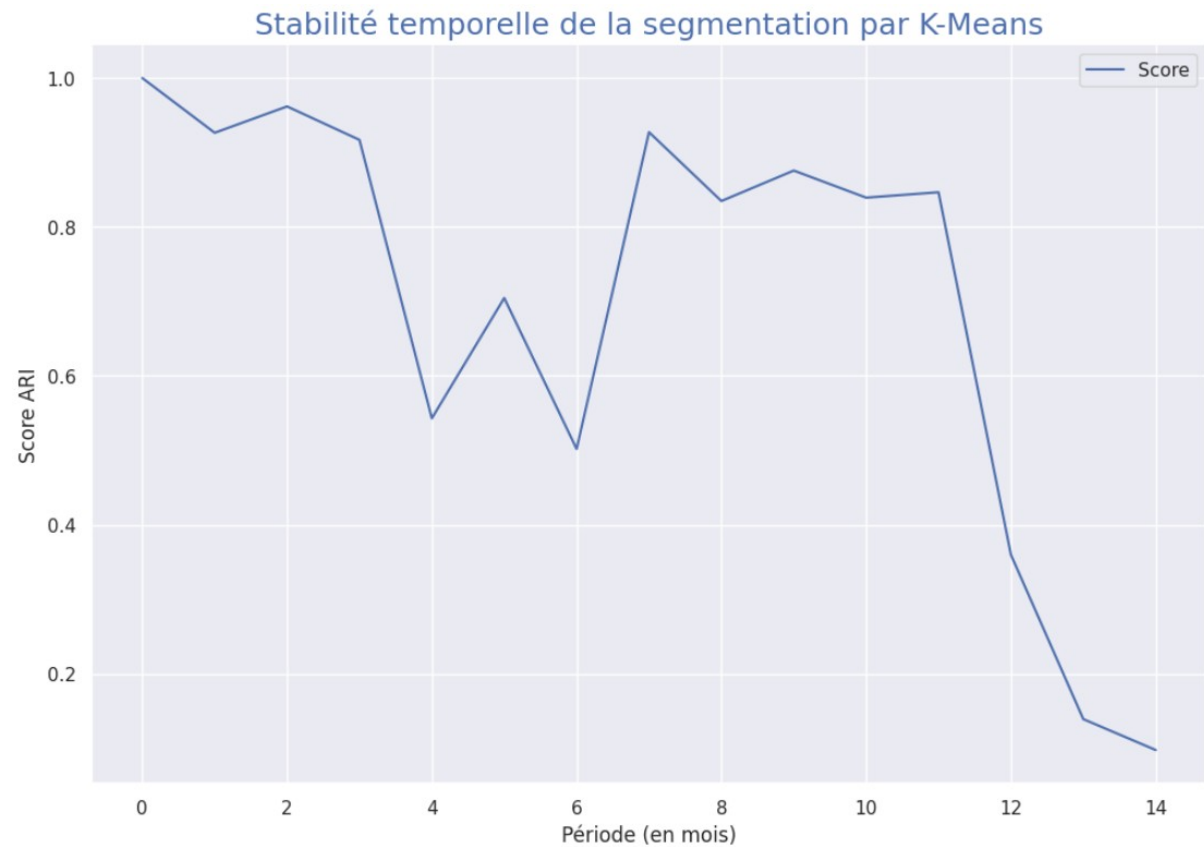
Stabilité du clustering

- Initialisation maîtrisée
- Suppression random_state



Besoin de maintenance

- Evolution du score ARI dans le temps





Conclusion

- Pistes d'améliorations
 - Calcul des features lors de l'ajout de nouveaux clients
 - Mis à jour régulière des features de récence et fréquence
 - Exploration d'autres variables
 - Réduction de la temporalité pour calcul besoin de maintenance