Segmentez des clients d'un site e-commerce

Parcours Data Scientist : Projet P5

Par May CHOUEIB

Présenté le 17 Juillet 2022

Problématique

Olist, une entreprise brésilienne qui propose une solution de vente sur les marketplaces en ligne souhaite obtenir une segmentation de ses clients utilisable au quotidien par leur équipe marketing dans leurs campagnes de communication

Objectifs :

- Comprendre les différents types d'utilisateurs grâce à leur comportement
- Fournir à Olist une segmentation de ses clients avec une description actionnable de cette segmentation
- Fournir à Olist une proposition de contrat de maintenance basée sur une analyse de la stabilité des segments au cours du temps

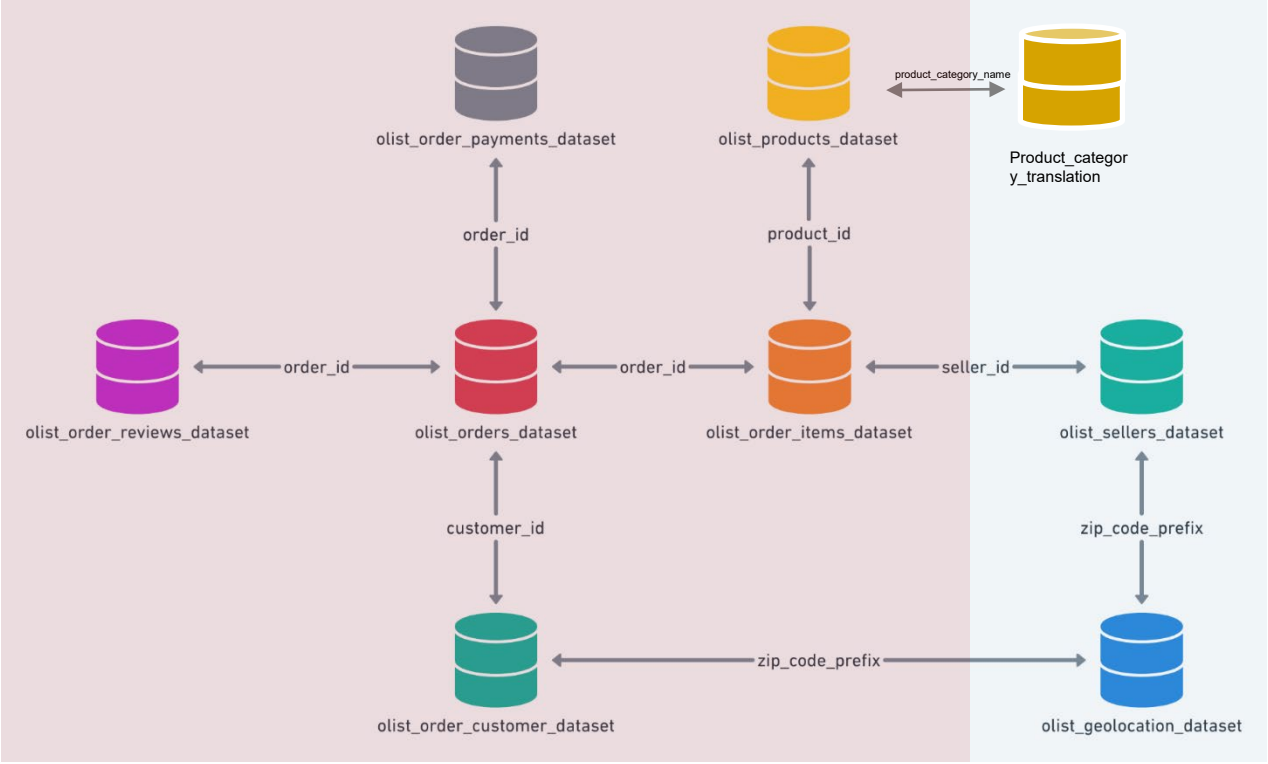
Plan

- Présentation du jeu de données
- Nettoyage et exploration des données
- Modélisations effectuées
- Modèle sélectionné et résultats de la segmentation
- Stabilité de la segmentation et proposition de contrat de maintenance

Présentation de jeux de données

Données réparties en 9 tables:

- clients / commandes / paiements / vendeurs / produits / traduction de produits en anglais / géolocalisation
- Ils sont liés entre eux par une ou plusieurs variables comme illustré sur le schéma ci-dessous du site kaggle de Olist



Nettoyage

- Suppression des colonnes avec un faible taux de remplissage
- Suppression des doublons
- Garder les commandes livrées et évaluées
- Jointure entre les variables communes des différents datasets
- Complétion des valeurs manquantes quand applicable avec la médiane de la variable
- Assemblage dans une table unique avec pour index l'id unique de client (customer_unique_id)
- Sélectionner les variables pertinentes pour la segmentation

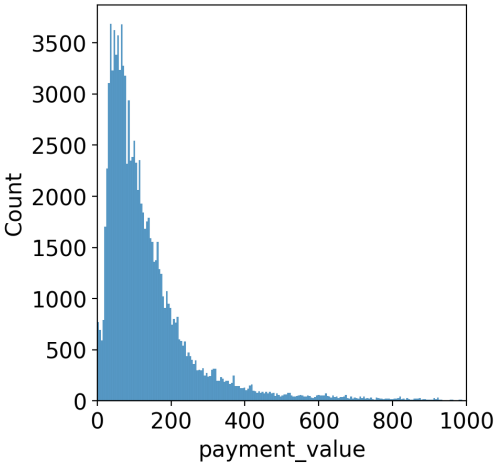
Nettoyage Variables pertinentes

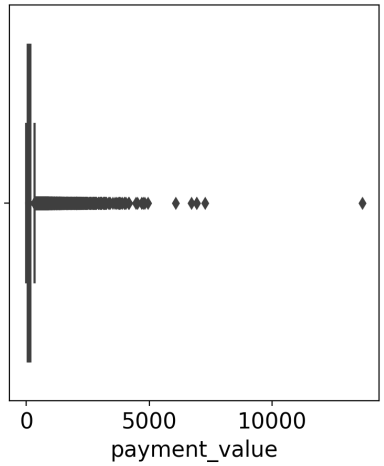
Indicateurs sélectionnés par client :

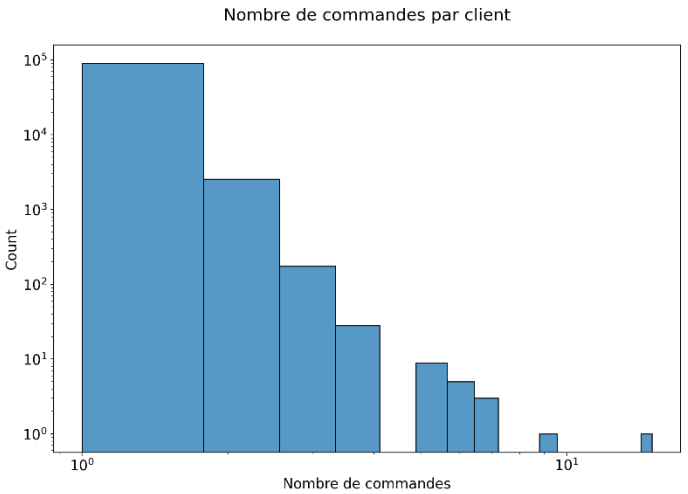
- Date du dernier achat (Récence/Recency)
- Nombre de commande (Fréquence/Frequency)
- Dépense totale (Montant/Monetary)
- Nombre de produits achetés (nb_item)
- Facilités de paiement (installment_payment)
- Moyen de paiement (payment_type)
- Note moyenne (review_score)
- Géolocalisation (State)

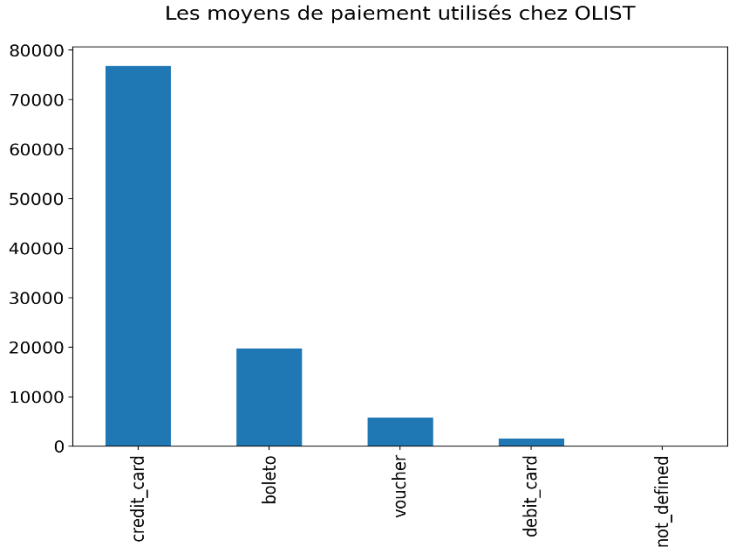
Dataset final :
92755 clients
8 attributs

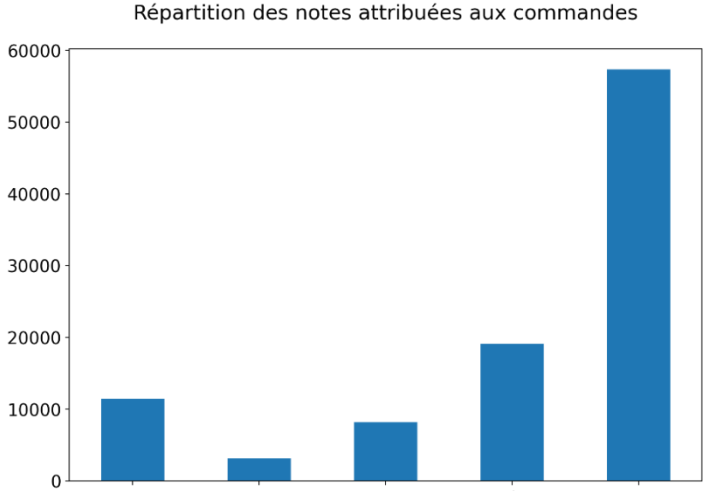
Analyse exploratoire





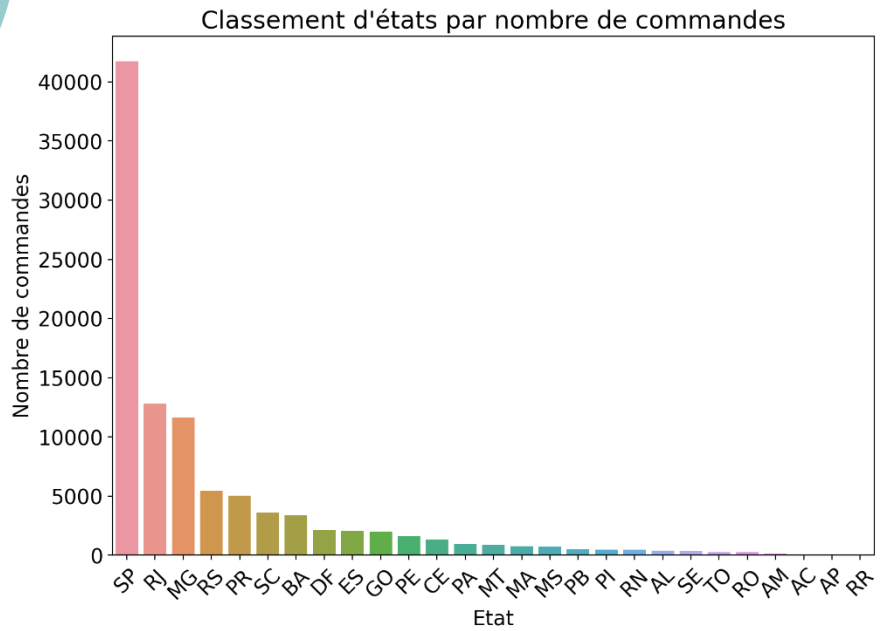




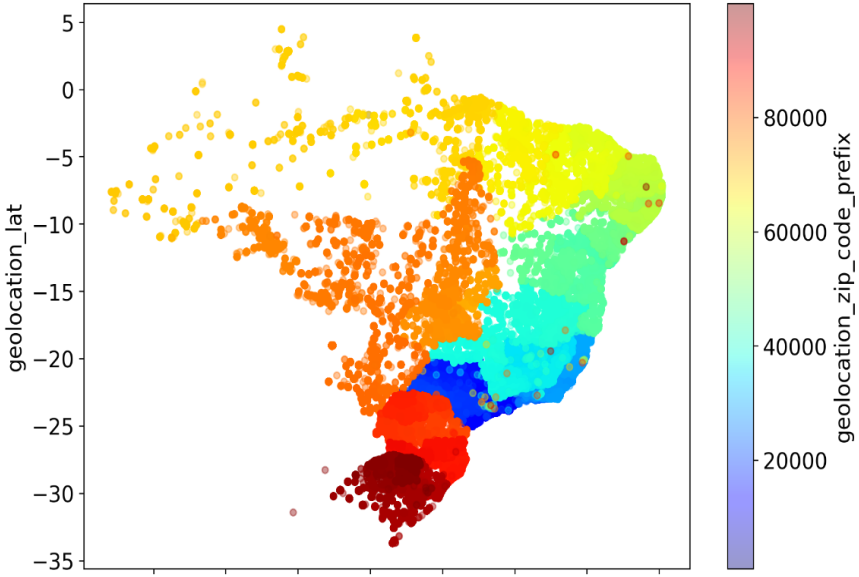


Analyse exploratoire

- Géolocalisation des clients



La majorité de commandes sont placées dans l'état de São Paulo



chevauchements dans les codes postaux

- Encodage binaire de l'état des acheteurs, Sao Paulo (1) ou Non (0)

Modèles testés

- 3 Modèles de classification non supervisée ont été testés

Kmeans
(Centroïdes)

- Permet de rechercher efficacement une partition des données dont la variance intra-cluster est minimale
- Le nombre k de clusters doit être fixé
- Détermine les centres de de k clusters
- Affecte chaque donnée au centre le plus proche
- Met à jour les centres au fur et à mesure de l'évolution des clusters (si nouvelle donnée y est affectée)

Agglomerative
(Hiérarchique ascendant)

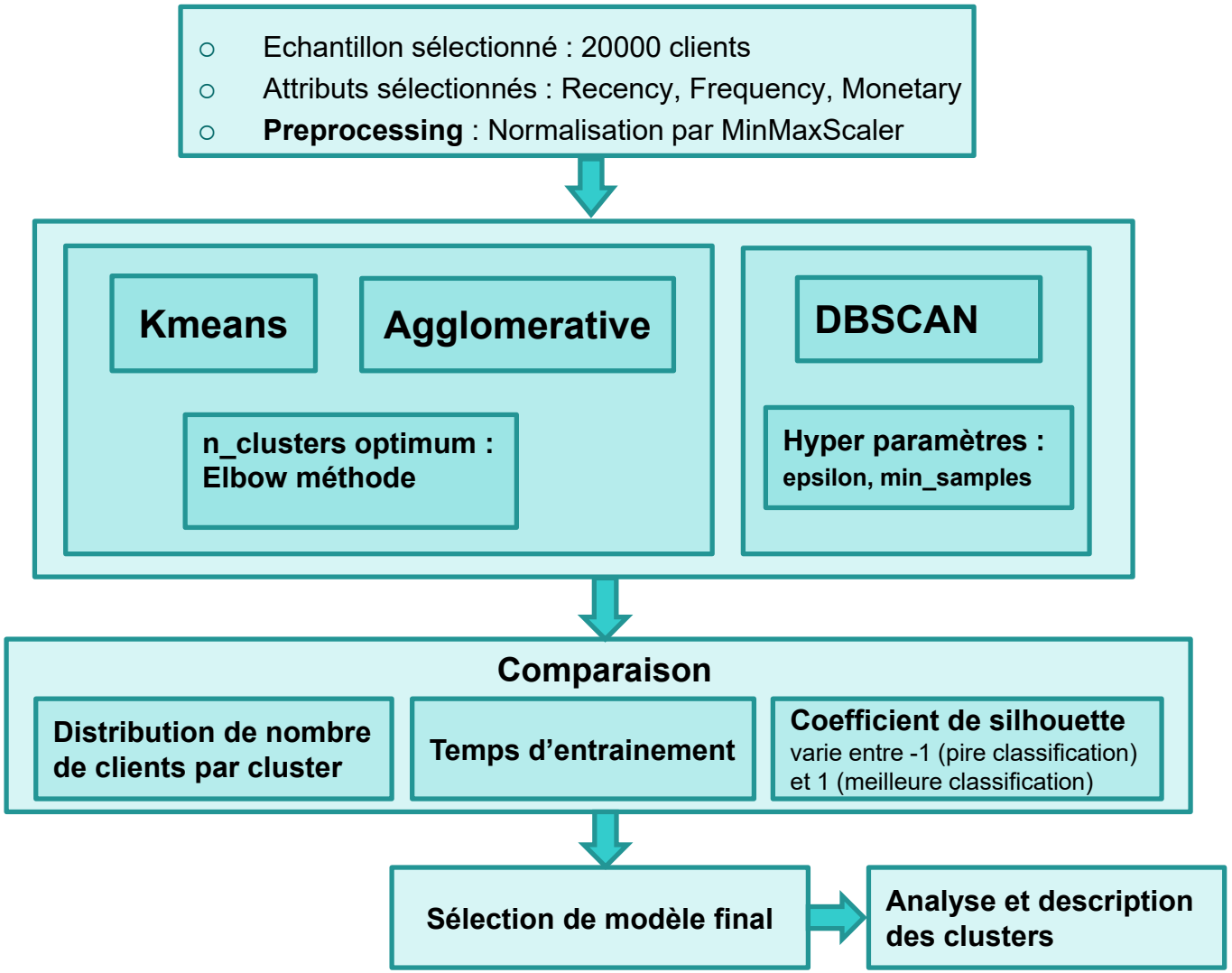
- Initialement chaque point est un cluster
- Agglomérer les observations proches (mesure de similarité-ressemblance)
- Itérer jusqu'à l'obtention d'un seul cluster

- Nombre de clusters non fixé (A établir)

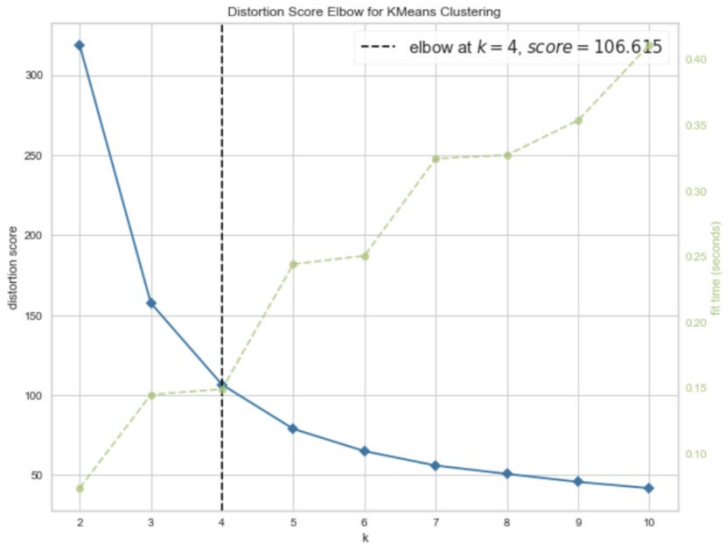
DBSCAN
(A densité)

- Il s'appuie sur la densité estimée des clusters pour effectuer le partitionnement
- Il prend deux paramètres en entrée : Ɛ la distance maximale qui peut définir deux individus comme voisins, et N le nombre minimum de points nécessaires pour former un cluster
- En plus de former des clusters, il repère les valeurs hors du commun (qualifiés de noise).

Modèles testés : Démarches



Résultats : KMeans

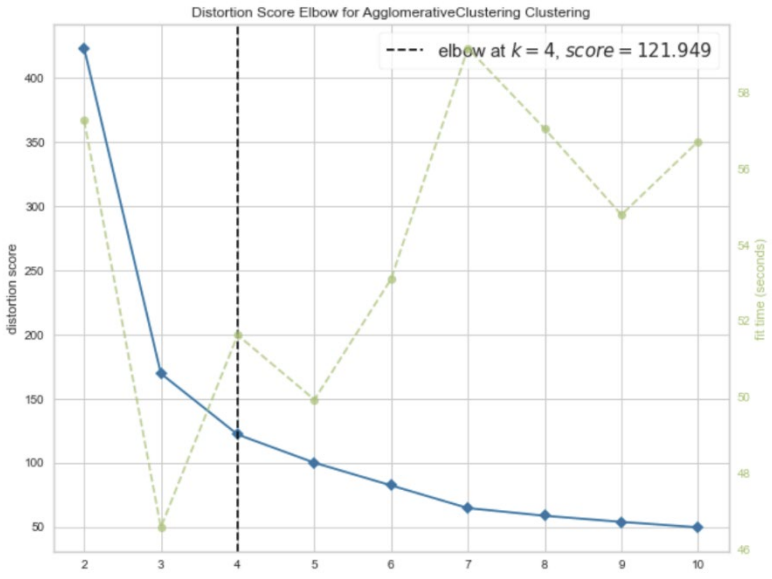


Elbow method basé sur le score de distorsion (somme moyenne des carrés des distances aux centres) K optimum 4

- K optimum 4
- 4 clusters répartis équitablement
- Temps d' exécution : **7.94sec**
- Coefficient de Silhouette : **0.51**

Résultats : Agglomerative clustering

- Temps d'exécution longue

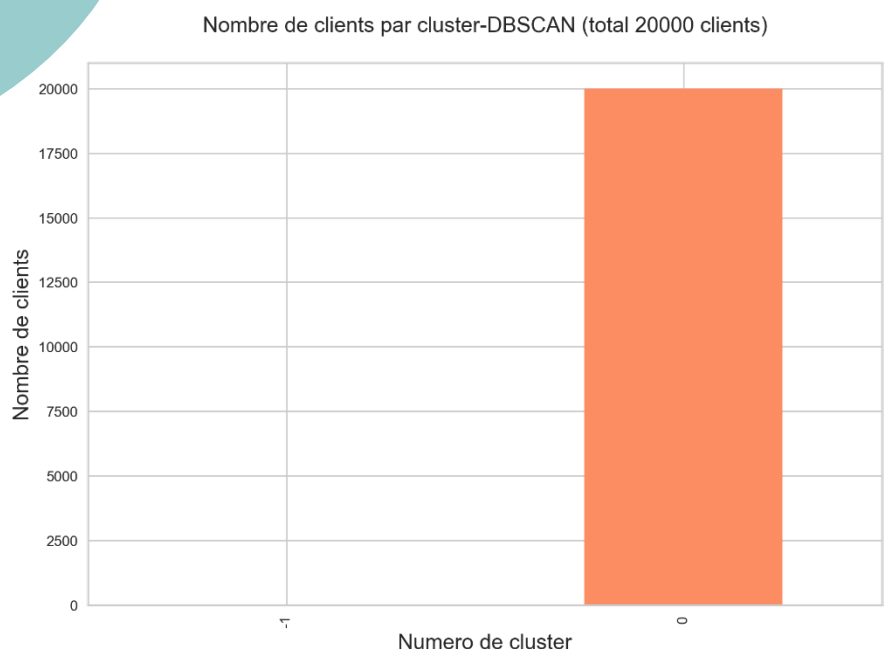


- K optimum 4
- 4 clusters répartis équitablement
- Temps d'exécution : **63.89sec** **(8x temps kmeans)**
- Coefficient de Silhouette : **0.43** **(< kmeans)**

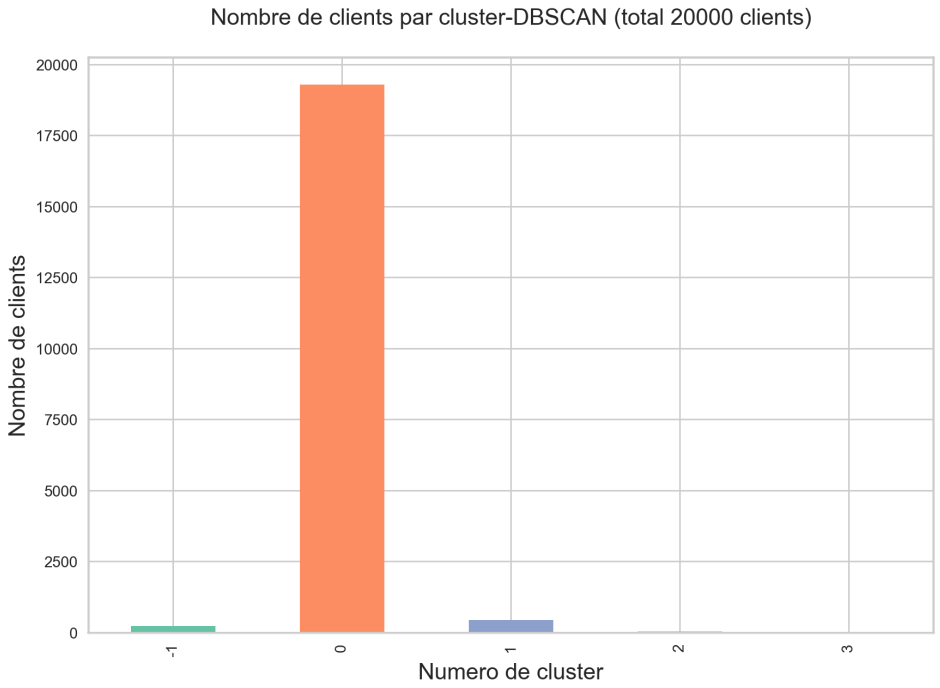
Résultats : DBSCAN

- Segmentation non pertinente même en variant les hyper paramètres

  - DBSCAN(eps = 0.4, min_samples = 100)



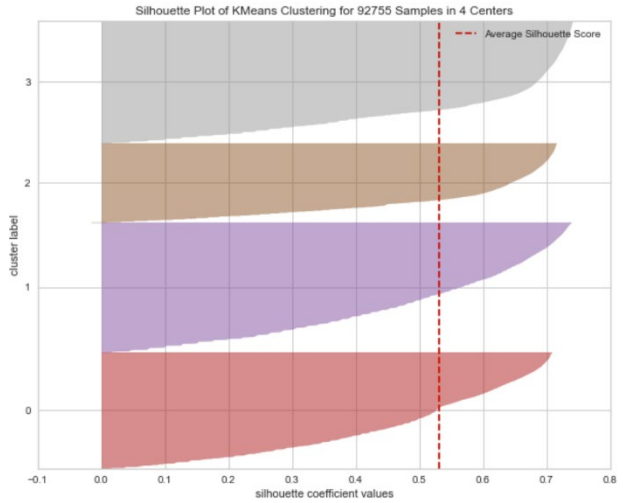
  - DBSCAN(eps = 0.40, min_samples = 10)



  - Le cluster -1 correspond aux clients non classés (noise)

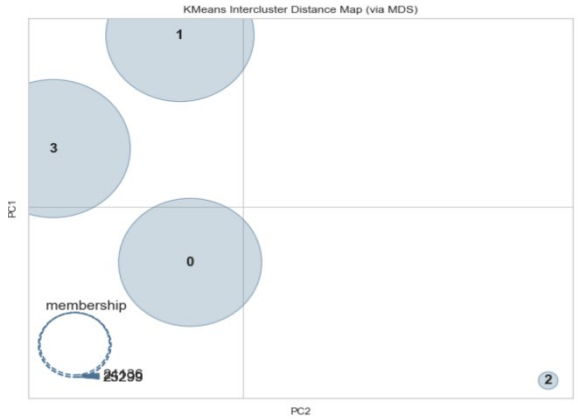
Modèle final : Kmeans

- Répartition des clusters pour le dataset complet : Les clusters semblent relativement bien répartis et les séparations sont claires

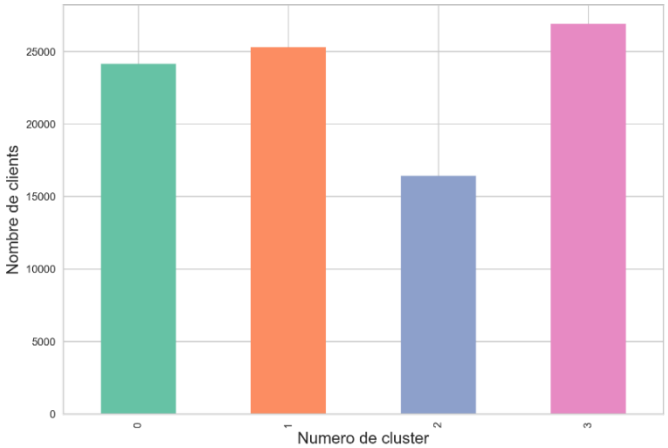


Coefficient de silhouette :
permet de visualiser la densité et la séparation des clusters

MAP distance entre clusters

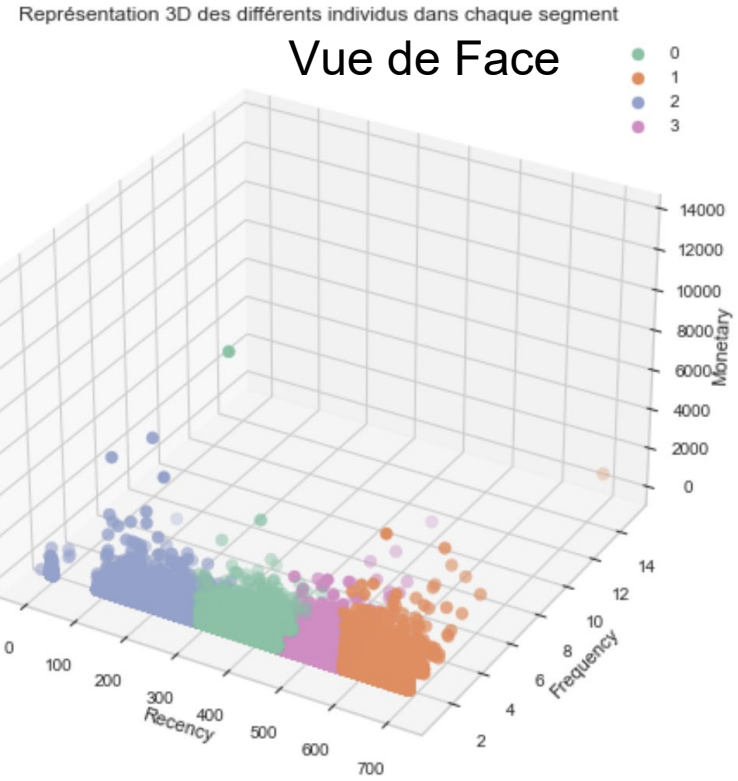


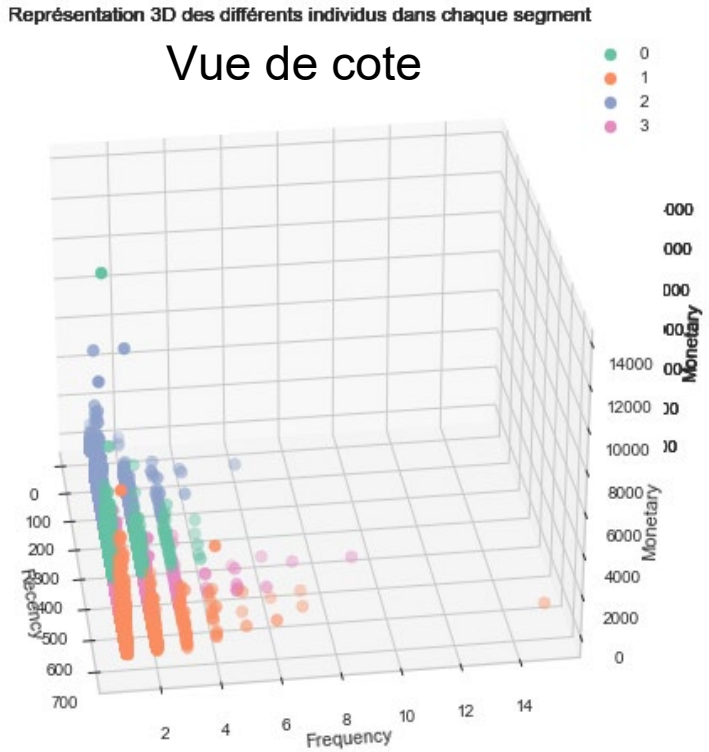
Nombre de clients par cluster-KMeans



Modèle final : Kmeans

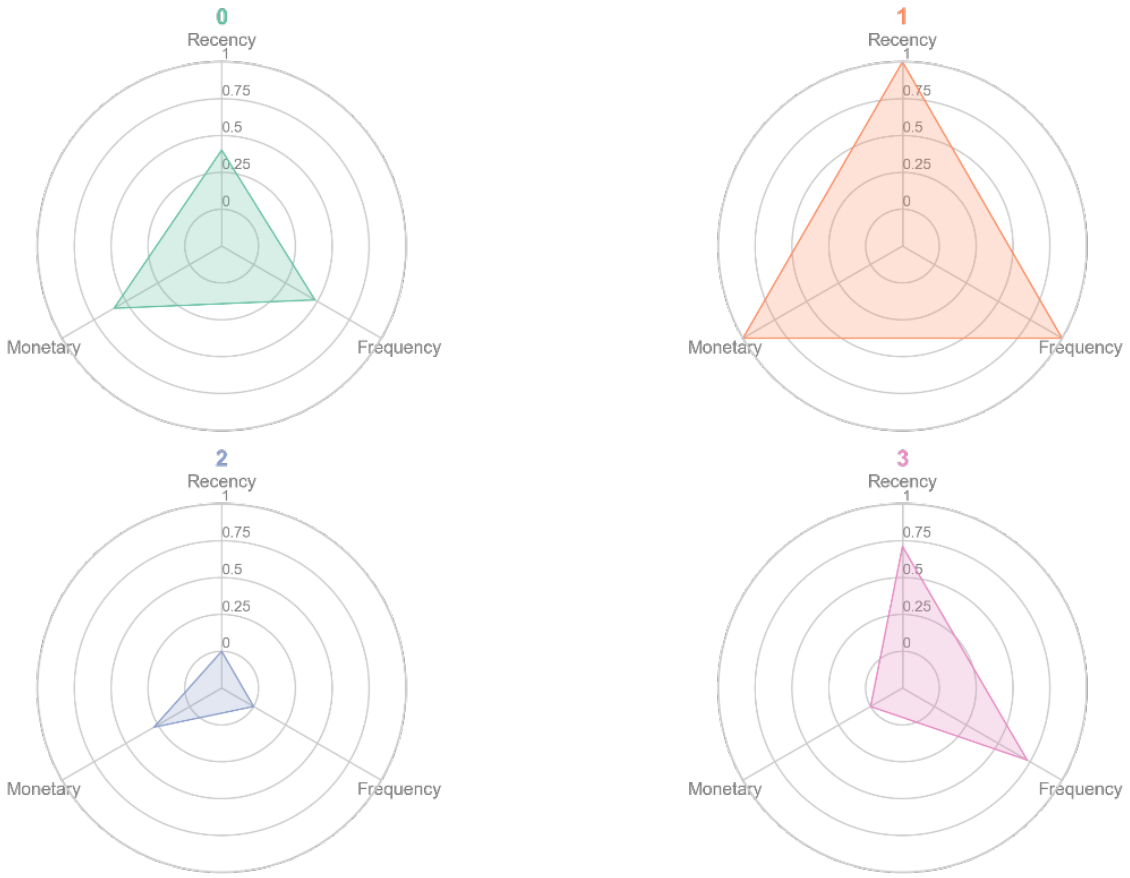
- Projection 3D des 4 clusters





Modèle final : Kmeans

- Projection radar des moyennes RFM pour chaque cluster pour une visualisation plus simple pour l'équipe marketing

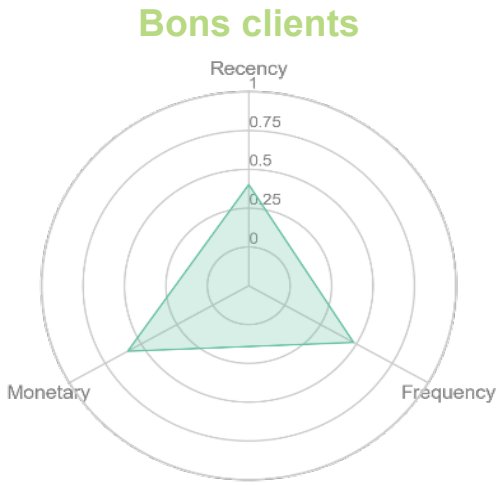


Identification des clusters (RFM)

Cluster Bons clients :

Achat midterme, plus d'une fois pour un bon montant.

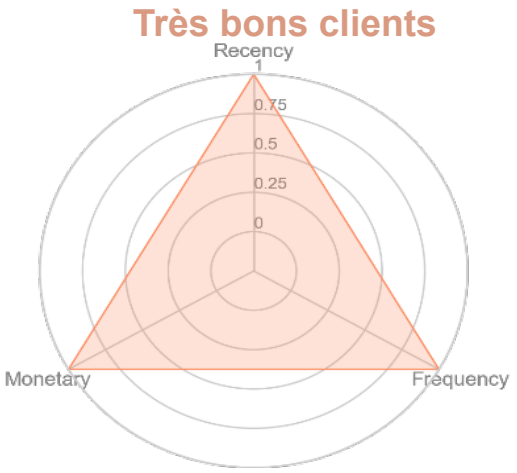
Action : Développement de la fréquence d'achat ou du montant par commande : coupler les actions de fidélisation et les promotions



Cluster Très bon clients:

Achat récent, plus d'une fois pour une bonne somme d'argent.

Action : fidélisation avec une carte premium, invitations VIP, etc.

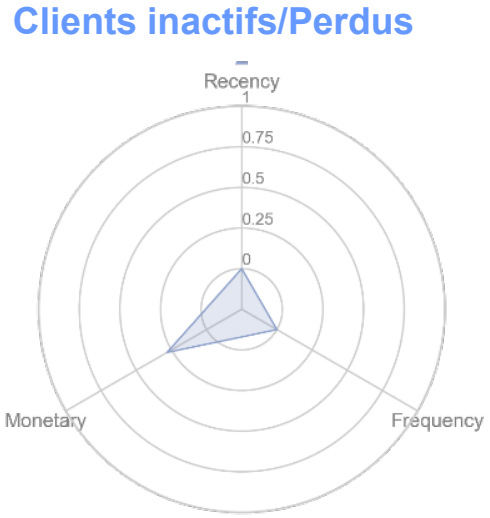


Identification des clusters (RFM)

Cluster Clients inactifs/Perdus :

pas d'achat fréquent ni récent et pour une faible somme d'argent.

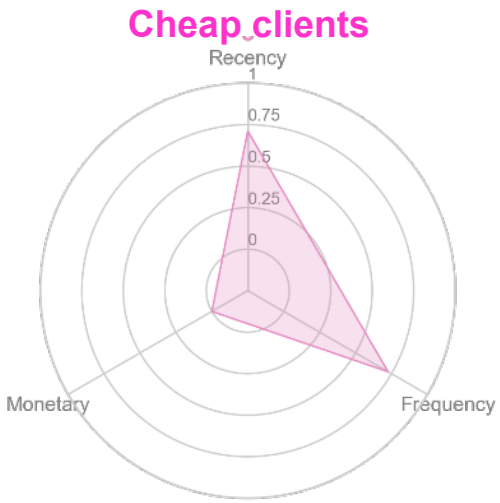
Action : Ne pas passer trop de temps à essayer de les récupérer. Séparer les dans une liste et cibler-les avec une compagne de réductions. Enlever ensuite tous ceux qui n'ont pas réagi.



Cluster Cheap clients :

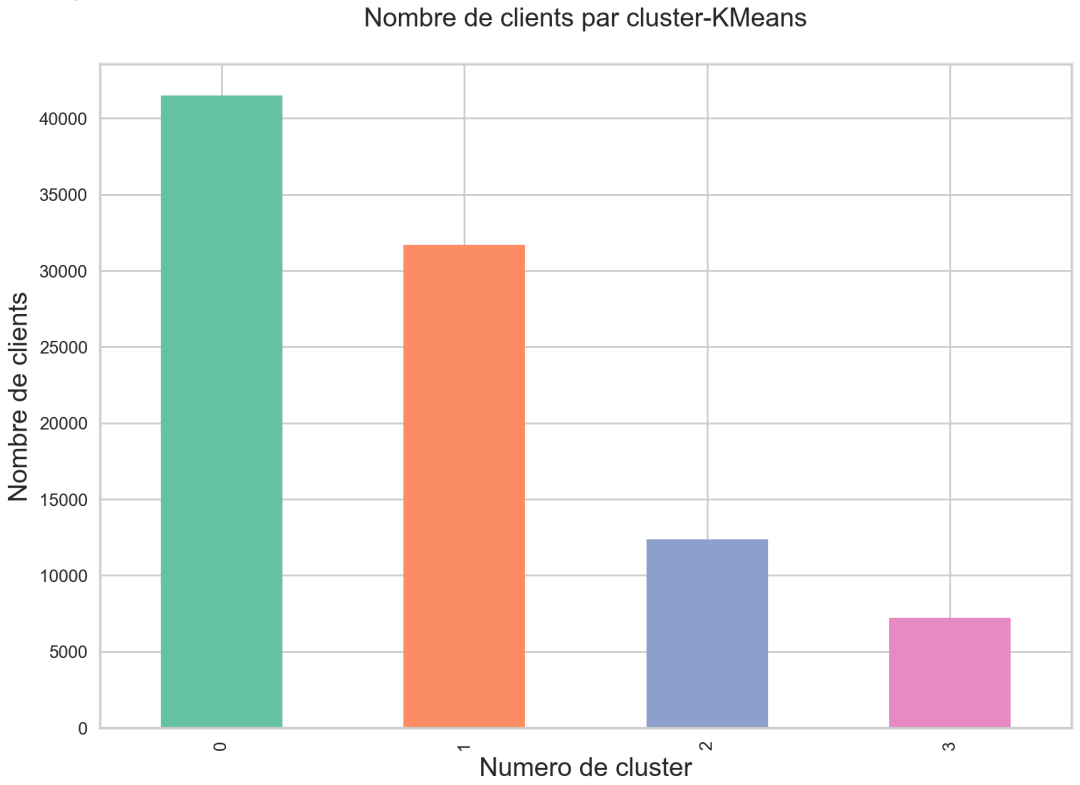
Achat récent, plus d'une fois mais pour une somme modique d'argent.

Action : Concentrez-vous sur l'augmentation de la monétisation grâce à des recommandations de produits basées sur des achats antérieurs et des promotions liées à des seuils de dépenses



Modèle Kmeans : RFM++

Attributs ajoutes au RFM : moyenne par client (satisfaction, nombre de produits achetés, nombre de moyens de paiement et des échéances) et son état de provenance (Saint Paulo ou pas)



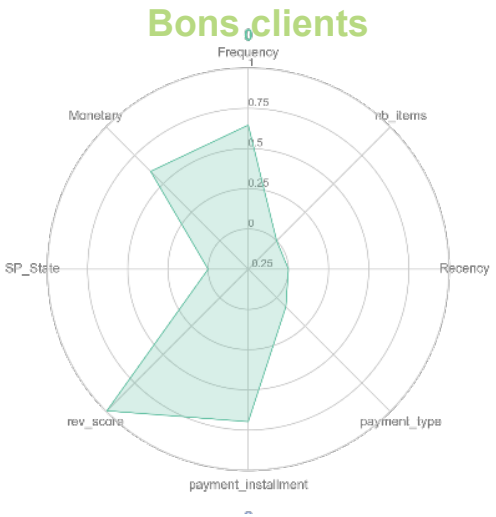
On remarque que les clusters ne sont pas repartis équitablement

Identification des clusters

Cluster Bons clients :

Achat plus d'une fois pour un bon montant avec un nombre important d'échéances. Ils ne sont pas de SP et leurs avis sont très bons.

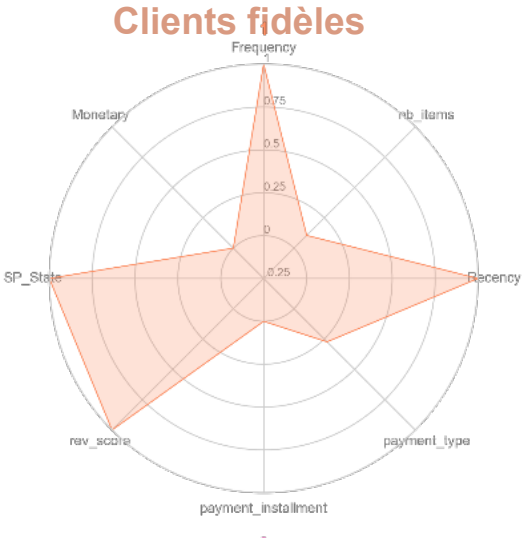
Action : Action de développement de la fréquence d'achat et du montant par commande: coupler les actions de fidélisation et les promotions



Cluster Clients fidèles :

De SP, ont acheté récemment et fréquemment mais pour un montant faible d'argent qu'ils le règlent en une seule échéance. Leurs avis sont très bons

Action : Action d'augmentation de la monétisation en la couplant a la fidélisation (une carte premium, invitations VIP, etc.) avec des promotions liées à des seuils de dépenses

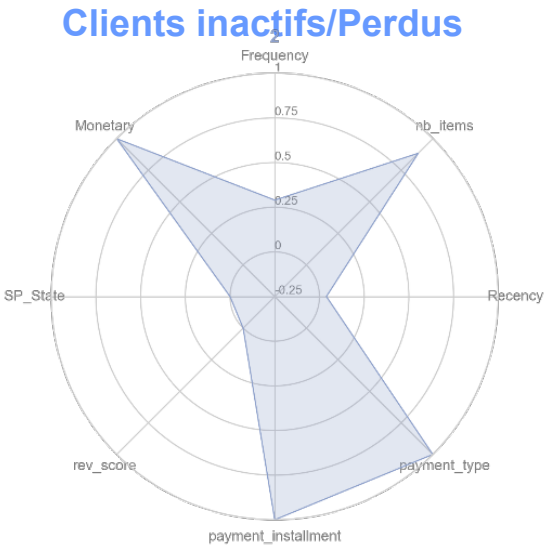


Identification des clusters

Cluster Clients inactifs/Perdus :

Achat plusieurs fois mais pas récemment. Ils ont payé une somme importante d'argent pour plusieurs produits avec plusieurs moyens de paiement et un nombre important d'échéances. Ils sont mécontents

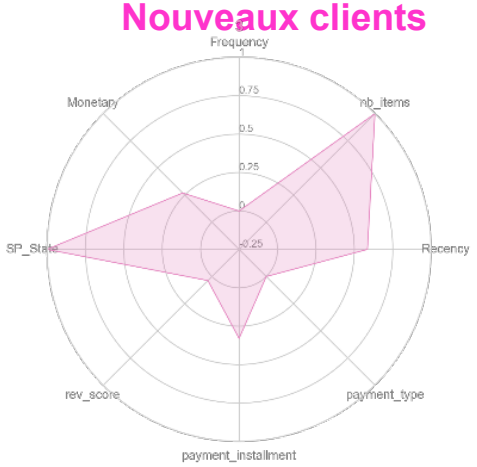
Action Il faut les récupérer en les réconciliant. Répondre a leurs avis négatifs et récompenser-les avec des bons de réductions conséquents.



Cluster Nouveaux clients :

Achat récent une fois, plusieurs produits mais pour un montant moyen et avec un nombre faible d'échéances. Ils sont de SP et ils sont mécontents.

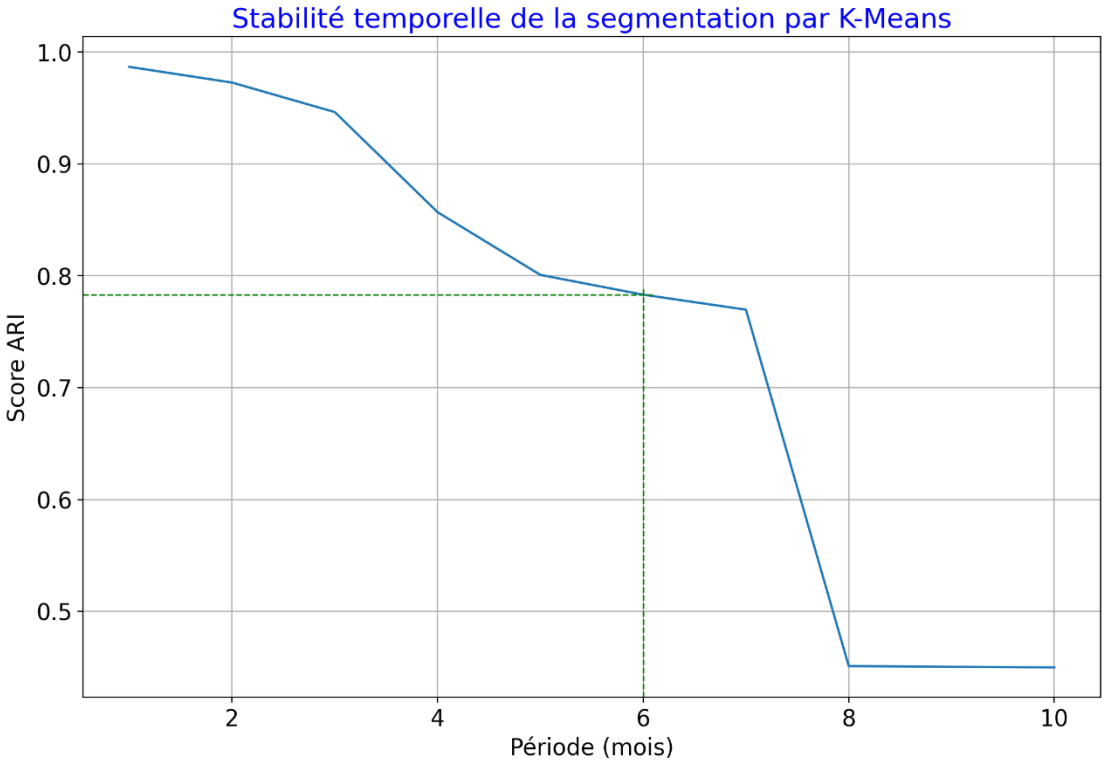
Action : Action de fidélisation et de réconciliation. Recommandations de produits basées sur des achats antérieurs avec des bons de réductions



Stabilité : Contrat de maintenance

- Adjusted Rand Index (ARI) est une mesure de similarité entre deux partitions d'un ensemble qui va nous permettre de déterminer à quel moment les clients changent de Cluster

- Calcul de ARI pour la première année et ensuite pour les 11 mois suivants avec une itération de 1 mois
- On remarque une forte inflexion a partir du 7eme mois sur les clients initiaux
- A prévoir la maintenance du programme de segmentation tous les 6 mois (ARI < 0.8)

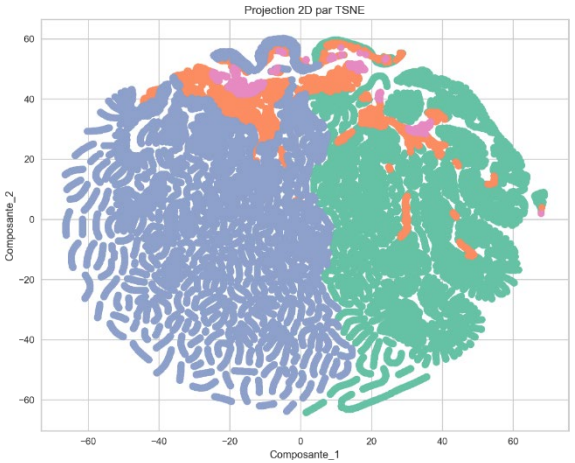


Conclusion

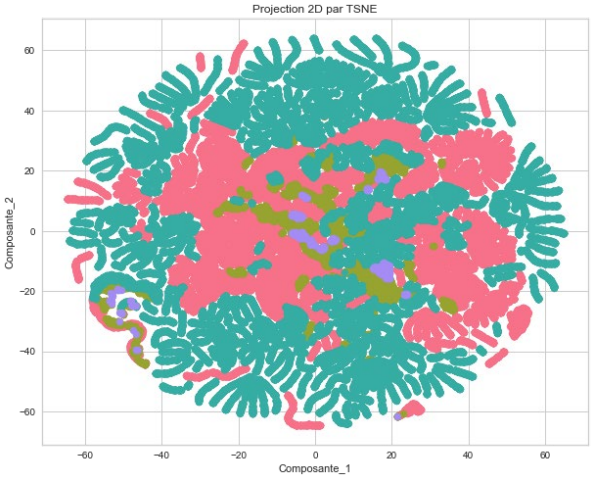
- Différents modèles de classification non-supervisé ont été testés
- Résultat optimal obtenu avec Kmeans(K=4)
- 4 segments équilibrés ont été identifiés avec la segmentation RFM
- 4 segments aussi ont été identifiés et interprétés avec 5 attributs de plus
- Etude de la stabilité des segments a montré qu'un contrat de maintenance tous les 6 mois s'impose

TSNE

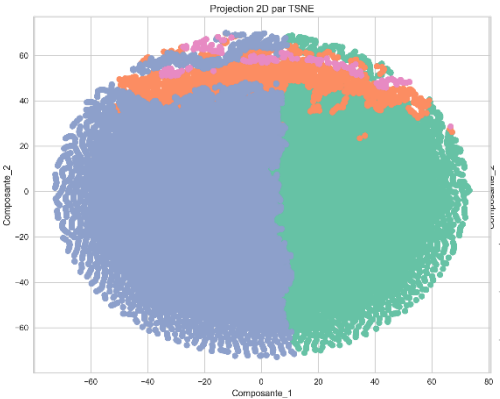
Init pca, def prex



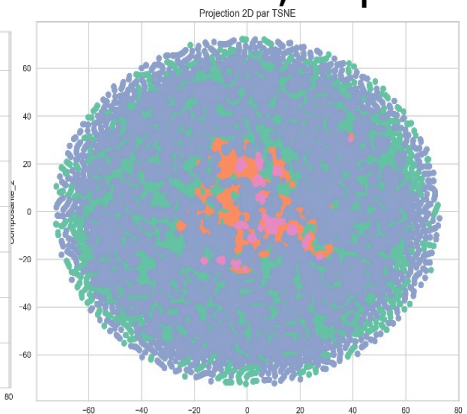
Init random, def prex



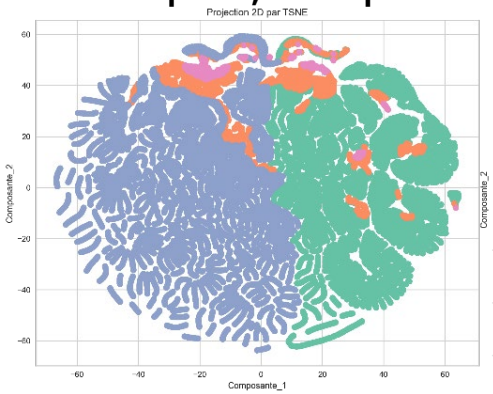
Init pca, 5 prex



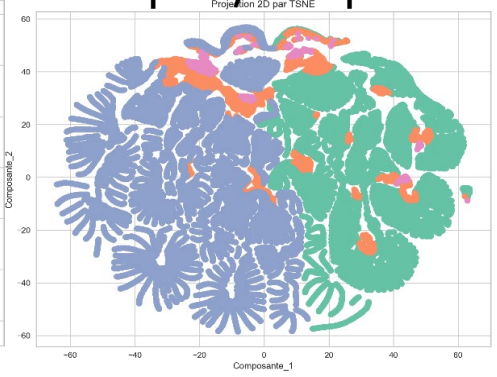
Init random, 5 prex



Init pca, 35 prex



Init pca, 50 prex



Dendograms