Parcours AI Engineer

SOUTENANCE PROJET 5

“Segmentation de la clientèle d’un site de e-commerce”

Stéphanie Duhem - Octobre 2024

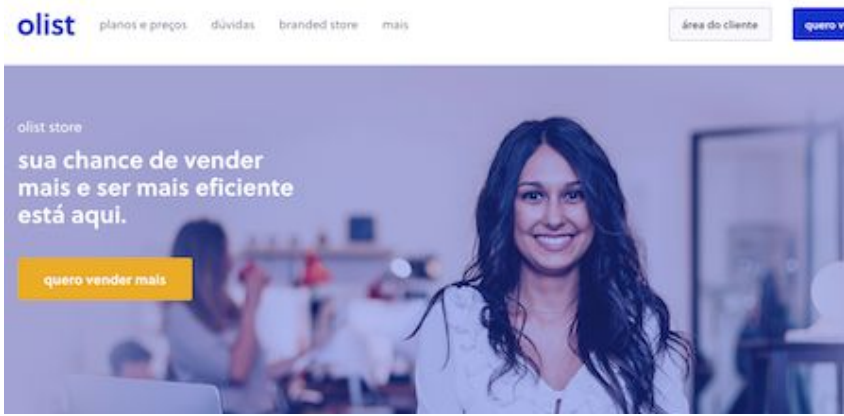
01. CONTEXTE & DÉROULÉ DU PROJET

LE CONTEXTE

Olist est un site bérislien de marketplaces en ligne, ouvert depuis septembre 2016.
L'entreprise souhaite une segmentation des clients que les équipes commerciales pourront utiliser au quotidien dans leurs campagnes de communication, ainsi qu'une proposition de contrat de maintenance basée sur une analyse de la stabilité des segments au cours du temps.

LES 5 GRANDES ÉTAPES DU PROJET

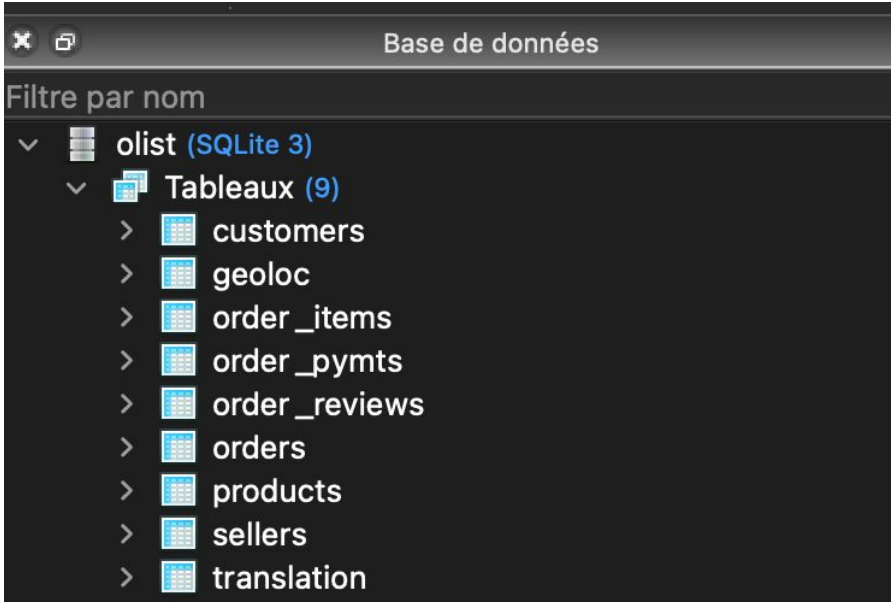
- Récupération des données depuis les BDD (SQLiteStudio)
- Analyse exploratoire des données
- Segmentation RFM "classique"
- Segmentations de ML et sélection du meilleur modèle
- Simulation d'un contrat de maintenance sur le meilleur modèle



01. La base de données olist

LES DONNÉES

- Tables de dimensions : customers, sellers, products, geoloc, products
- Tables de faits : orders, order_items, order_pymts, order_reviews



02. Analyse exploratoire

LES DONNÉES

- **Historique disponible de septembre 2016 à août 2018**
- **119 085 lignes** (1 ligne = 1 produit dans une commande)
- **33 colonnes** (informations client, commande, vendeur, produit, paiement et avis client)
- **Peu de NaN** : avant le nettoyage des données, seulement 5% de valeurs manquantes. Ce sont les informations concernant les avis client qui sont le moins renseigné (+ de 50% de NaN pour les colonnes des commentaires et des titres des commentaires)

PREMIÈRE SÉLECTION DES DONNÉES

- Pour ce projet, nous avons besoin d'avoir des informations concernant les clients pour lesquels Olist a fourni un service complet, nous ne gardons donc que les commandes dont le statut est "*delivered*"

FEATURES ENGINEERING

- **Plusieurs agrégations** (ex.: total de commande par client, total de produit par commande, montant total d'une commande, etc.)
- **Des calculs d'écarts entre les différentes dates disponibles** (ex.: écart entre les dates de livraison estimée et de livraison réelle, délai entre la date de validation de la commande et sa livraison, etc.)

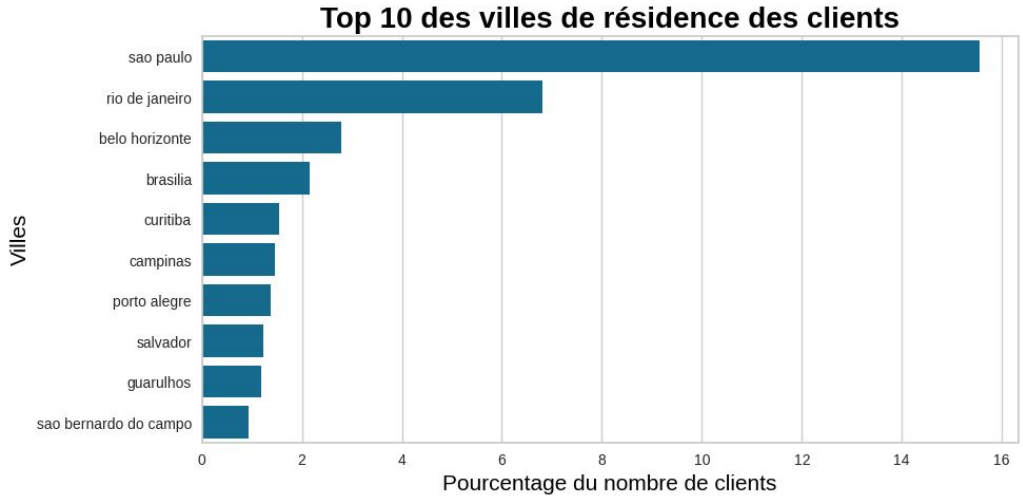
02. Analyse exploratoire

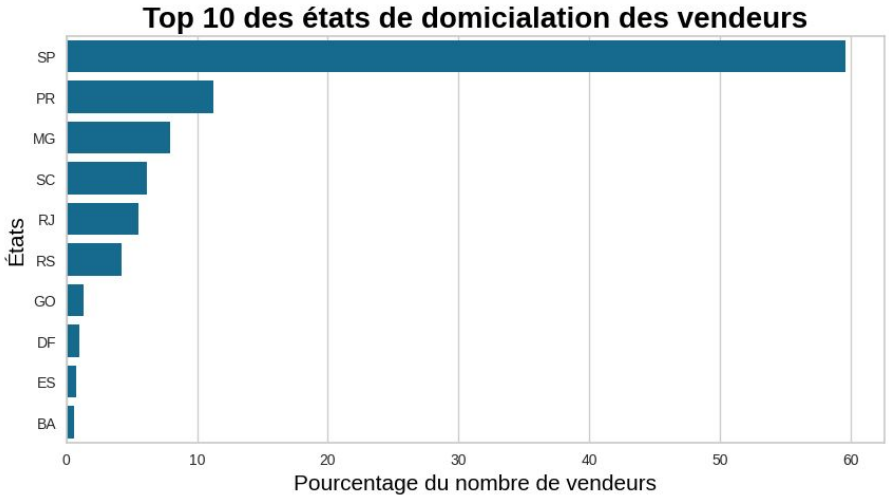
- **près de 94% des clients n'ont effectué qu'une seule commande**
- **près de 87% des commandes ne comportent qu'un seul article**
- **Peu de renouvellement et de fidélisation de la part des clients = manque de rétention client manifeste**

Le comportement d'achat unique (One-time purchase behavior) est manifestement la base de clientèle de Olist.

Ce premier avis est à tempérer car le site est encore "jeune" (historique de - de 3 ans)

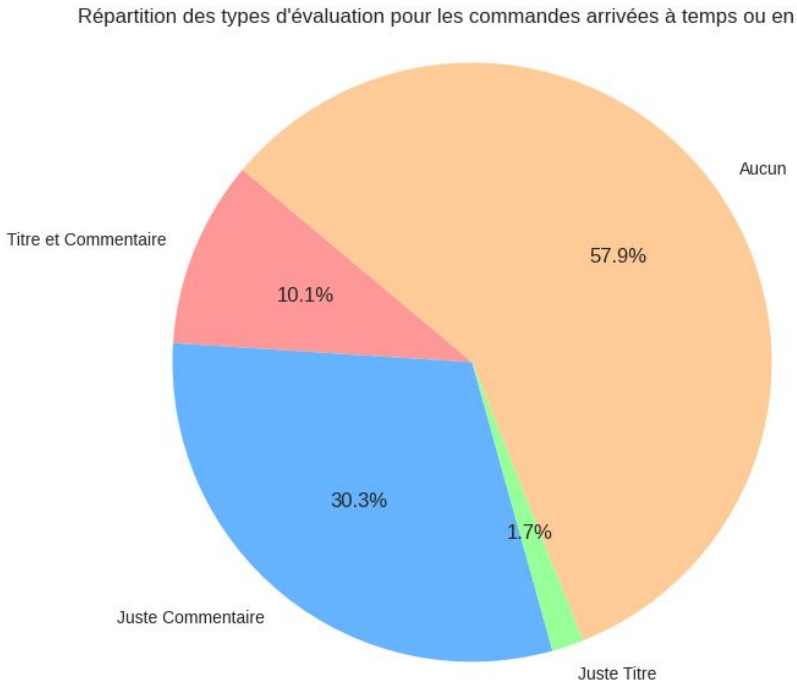
02. Analyse exploratoire

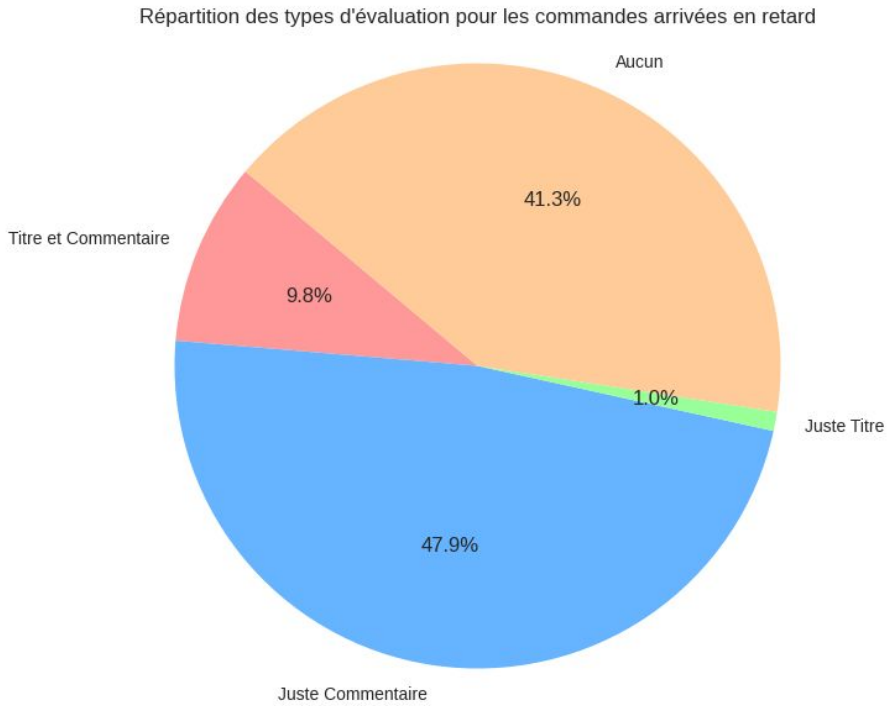




- la grande majorité des clients d'Olist est urbaine, issue des plus grandes ville du pays.
- la grande majorité des vendeurs d'Olist se trouve en zone urbaine, proche de la majorité des clients.

02. Analyse exploratoire

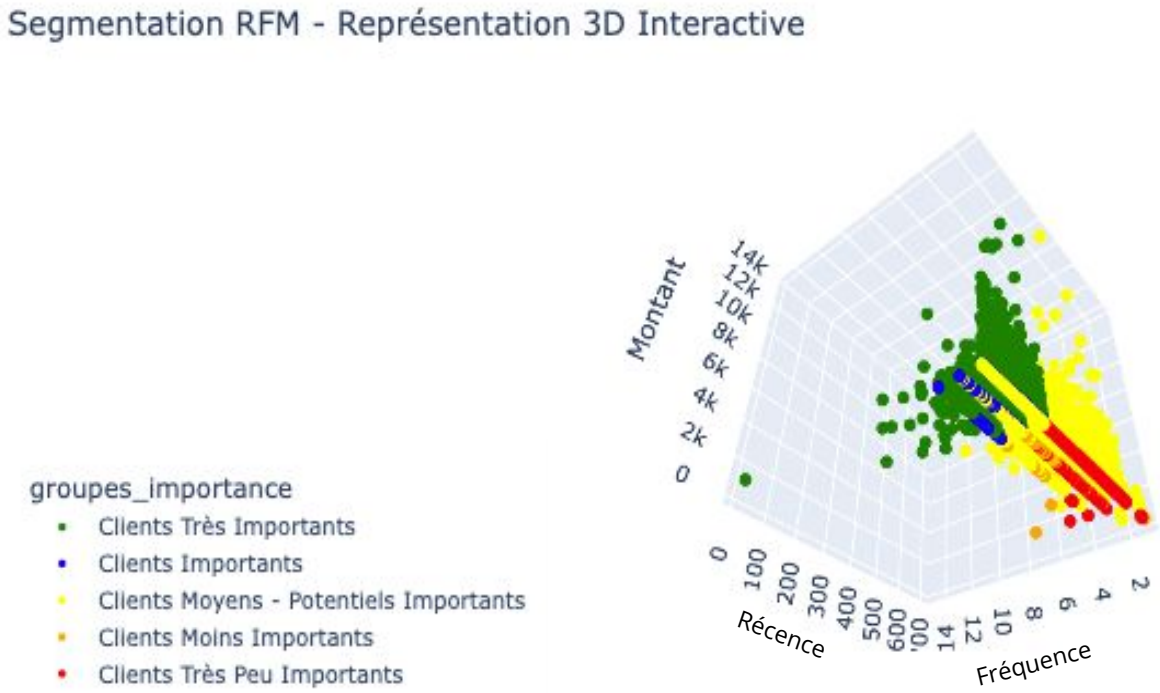




- **L'insatisfaction des clients s'exprime plus** en quantité (nombres d'évaluation écrites) et en qualité (longueur des textes) que leur satisfaction

03. Segmentation RFM “classique”

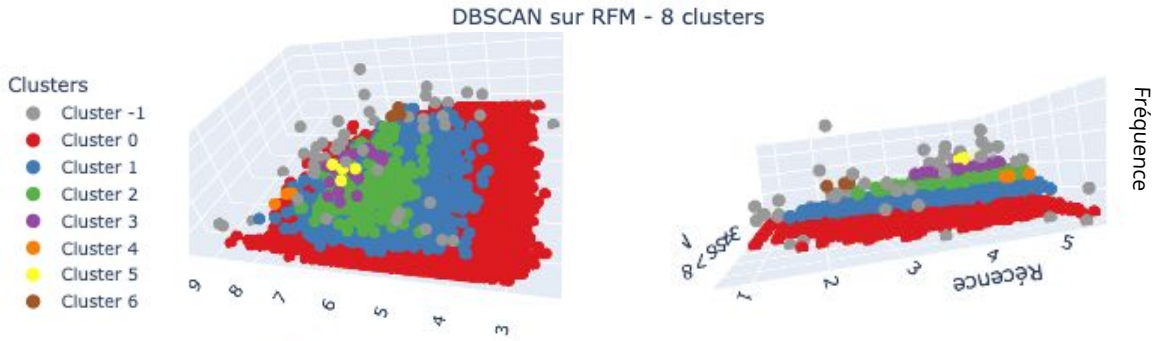
- Segmentation basée sur la méthode des quantiles des variables : **Récence, Fréquence et Montant**
- Score de silhouette de **-0,086**
- **L’imbrication visible des clusters**, un chevauchement des groupes.
- Pas de détermination de groupes homogènes de clients, manque de précision



04. Segmentation ML : sélection du modèle sur RFM

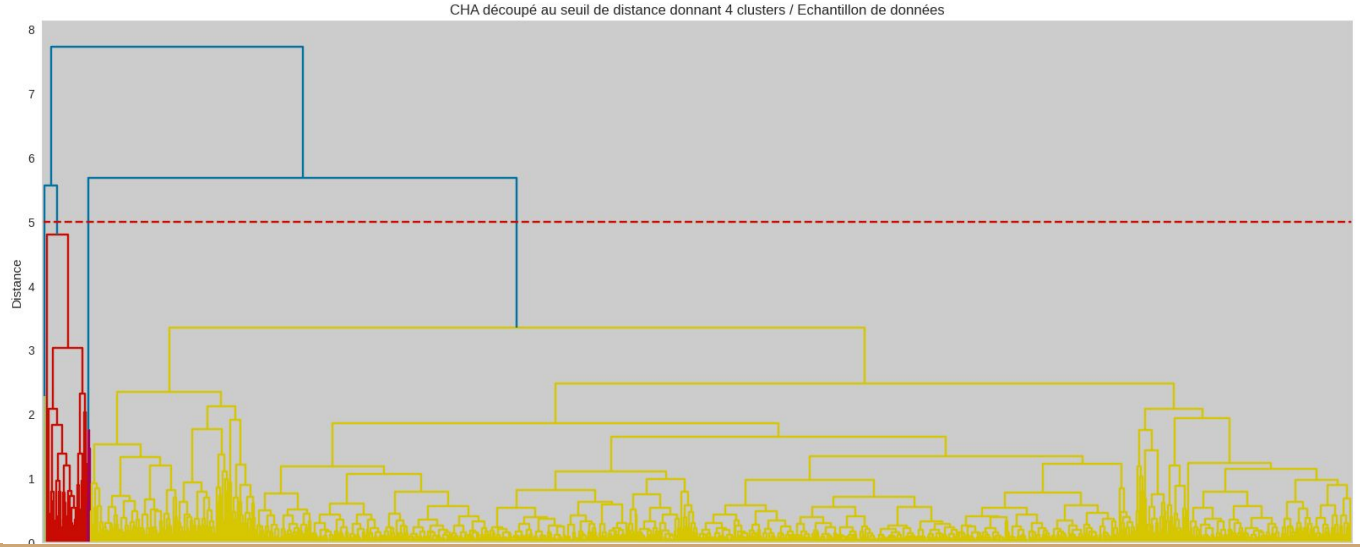
DBSCAN :

- partition peu fiable car surtout basée sur la fréquence d'achat.



CHA :

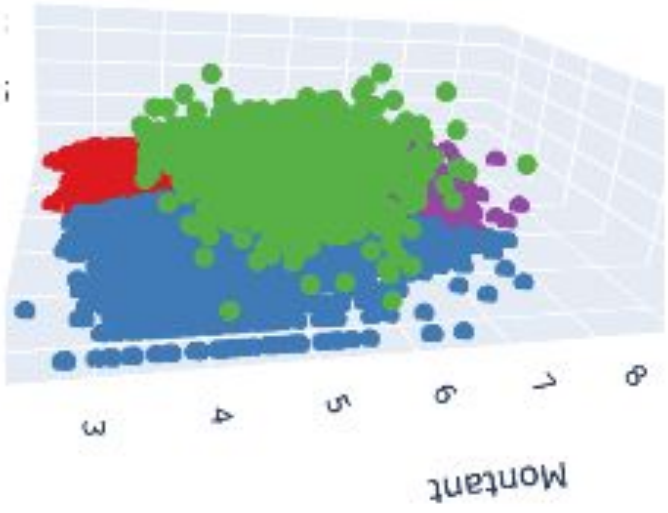
- clusters très déséquilibrés sur l'échantillon de données
- impossible de passer à l'échelle et d'appliquer la CHA sur l'ensemble des données

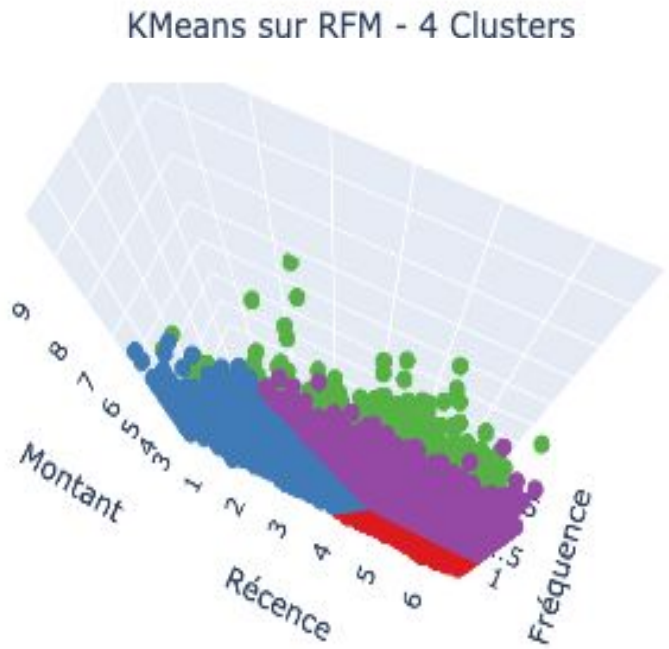


04. Segmentation ML : sélection du modèle sur RFM

Kmeans :

- Clusters mieux définis et homogènes
- Répartition équilibrée
- Nombre de clusters raisonnable

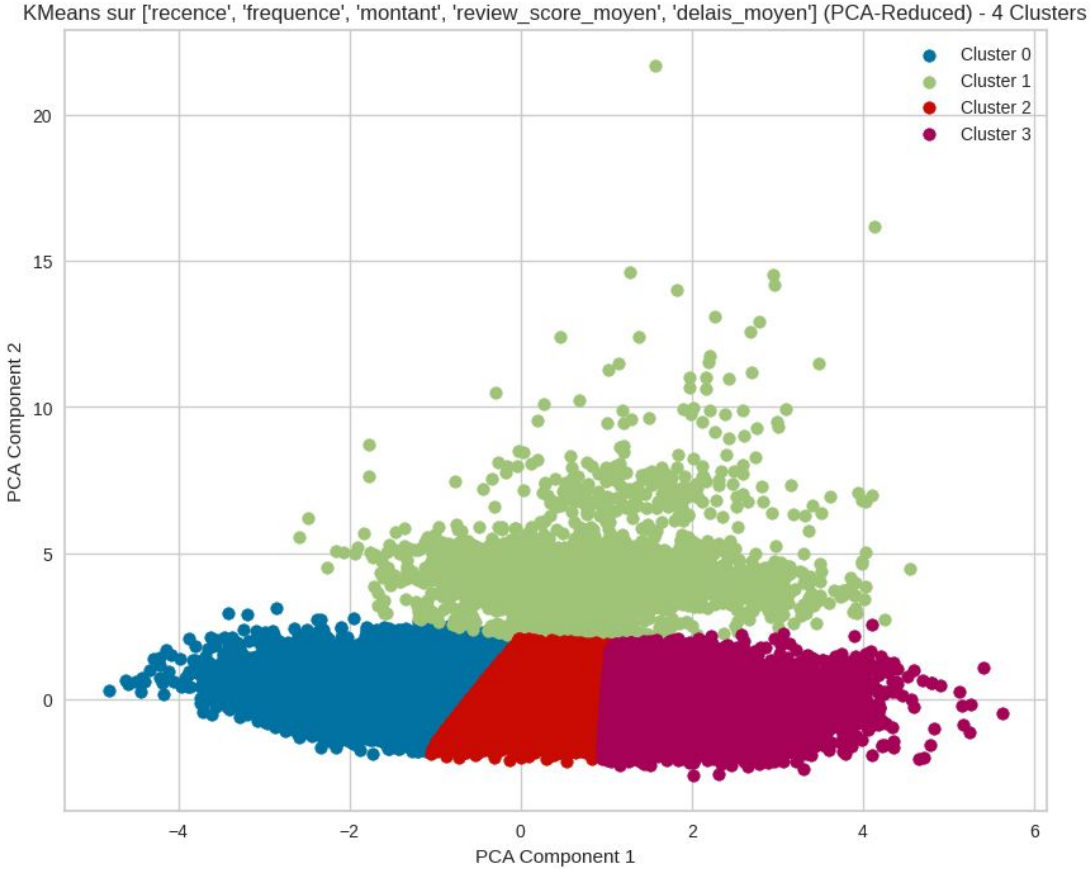




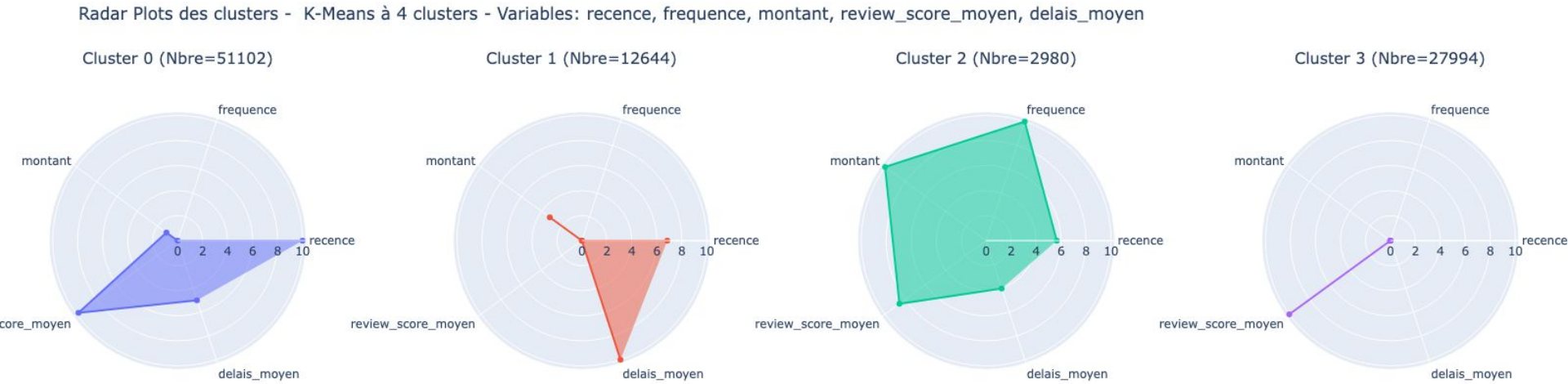
04. Segmentation ML : Amélioration du KMeans

Kmeans :

- Variables RFM + review_score_moyen + delais_moyen
- 4 clusters
- Facilement interprétable



04. Segmentation ML : Amélioration du KMeans



Interprétation des clusters :

- CLUSTER 0 (51102 clients) = clients anciens à relancer
- CLUSTER 1 (12644 clients) = clients perdus
- CLUSTER 2 (2980 clients) = clients fidèles, satisfaits et dépensiers
- CLUSTER 3 (27994 clients) = nouveaux arrivants à relancer

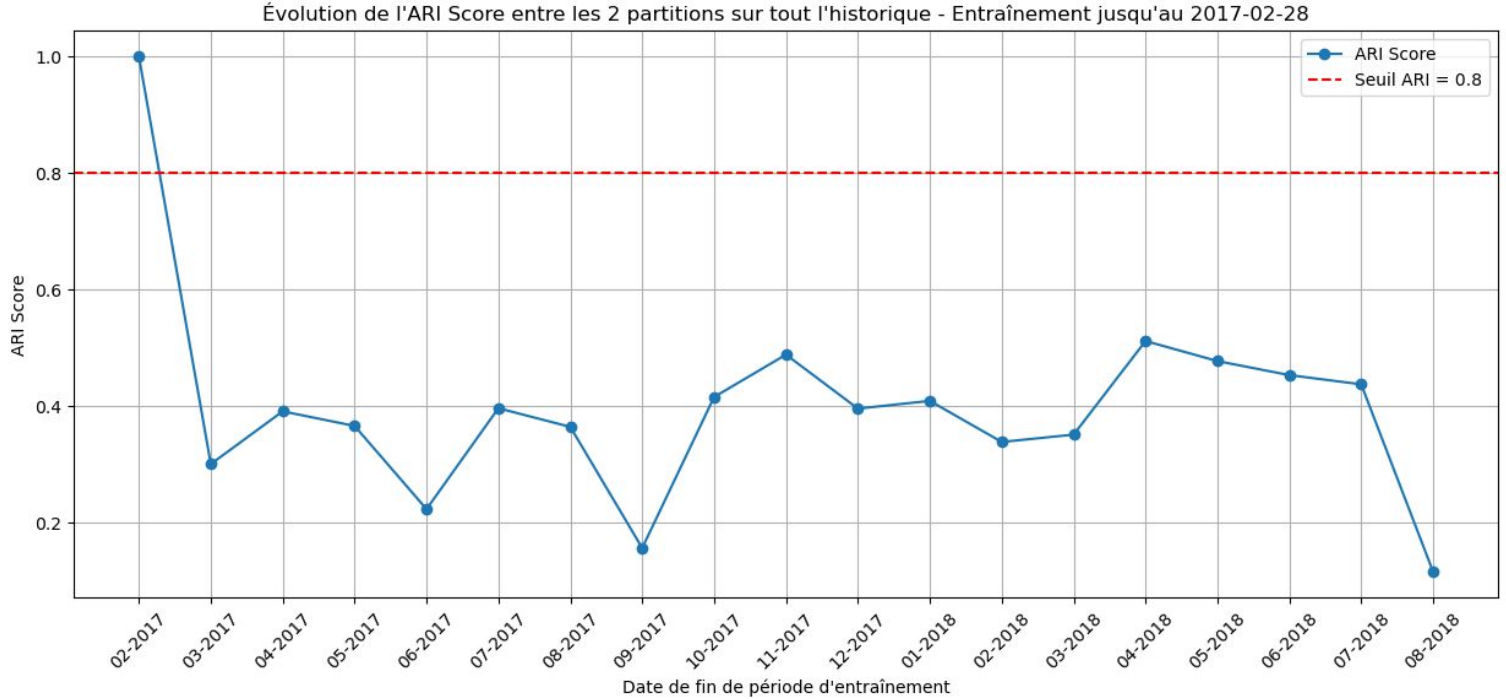
Informations pertinentes :

- Impact du délai de livraison sur la fidélisation client
- Les montants élevés ne sont pas un frein à la fidélisation.
- Identification de clients 'VIP' : à prévoir des actions marketing ciblées avec de belles retombées.

05. Simulation contrat de maintenance

Méthodologie:

- Un Kmeans de référence entraîné sur une 1ère période de donnée, puis utilisée pour prédire les labels ‘futurs’ sur le reste de l’historique.
- Les clusters de référence seront ensuite comparés à ceux d’un 2nd Kmeans “réalité terrain”, entraîné sur toute la période.
- La comparaison est faite en utilisant le ARI Score : en dessous d’un seuil, il faut prévoir la maintenance.



06. Conclusions

Simulation de Maintenance :

- Suite à notre simulation, il semble qu'il faudrait mettre en place une maintenance tous les mois.
- Pour affiner notre estimation nous pourrions simuler les effets d'une maintenance régulière sur l'historique.
- Certaines variables peuvent faire évoluer le nombre de cluster (par exemple la fréquence qui augmenterait drastiquement), il serait opportun de mettre en place des tests statistiques pour les évaluer et ainsi mettre en place des règles pour permettre une maintenance préventive

La Segmentation :

- Avec une augmentation des clients et donc des données, on pourrait s'intéresser à d'autres variables pour améliorer le modèle.
- Par exemples :
    - la géolocalisation (on a vu l'impact du délai de livraison sur l'insatisfaction des clients, les grandes distances entre les vendeurs et les acheteurs rajoutent du délai de livraison)
    - les catégories de produits (certaines catégories à fort CA et Volume de ventes pourraient être mises en avant et affiner notre connaissance de la clientèle)
    - les avis clients pourraient être étudiés plus en profondeur avec une analyse de sentiment