Projet 5 - Openclassrooms Segmentez des clients d'un site E-commerce

Dabidin Keshika
03.05.2023

OPENCLASSROOMS

olist

Plan

1. Contexte de l'étude
2. Présentation des données
3. Démarche de nettoyage et feature engineering
4. Analyse exploratoire des données
5. Présentation des différentes pistes de modélisation effectuées et du modèle final sélectionné
6. Contrat de maintenance
7. Conclusion

olist

Introduction

- **Contexte :**

Olist est une entreprise brésilienne qui propose une solution de vente sur les marketplaces en ligne.

Olist souhaite que l'on fournisse à ses équipes une segmentation des clients qu'elles pourront utiliser au quotidien pour leurs campagnes de communication.

L'ensemble de données contient des informations sur 100 000 commandes de 2016 à 2018 effectuées sur plusieurs marchés au Brésil.

- **Objectifs:**

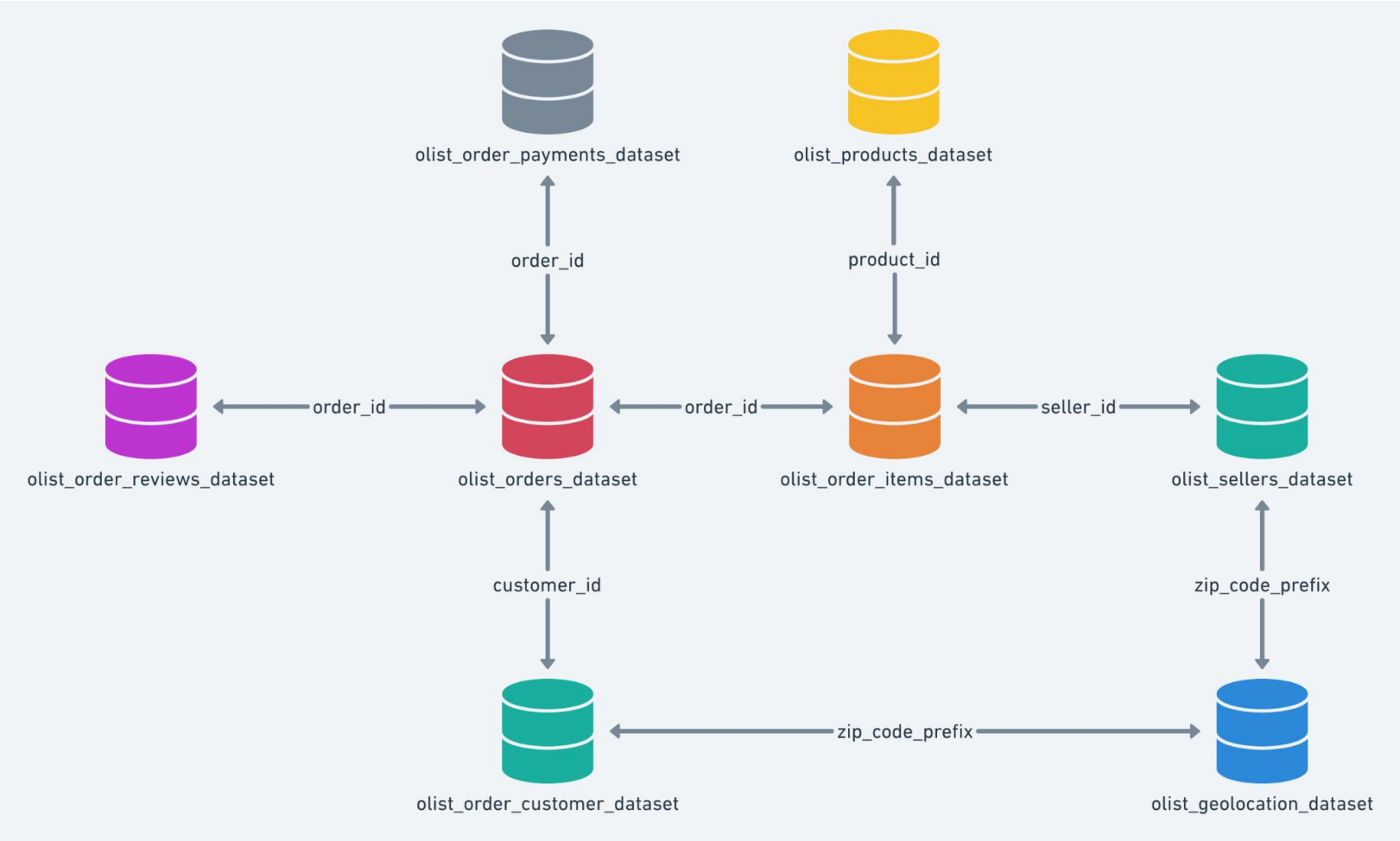
  - Identifier les différents types d'utilisateurs à partir d'un modèle de segmentation
  - Fournir à l'équipe de marketing une description des clients
  - Mise en place d'un contrat de maintenance

olist

Présentation des données

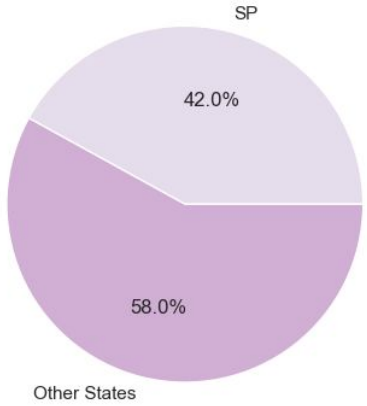
- **Résumé :**

9 fichiers comportent des informations sur le statut de la commande, le prix, les performances de paiement, le coût de transport à l'emplacement du client, les attributs du produit et enfin les avis rédigés par les clients.

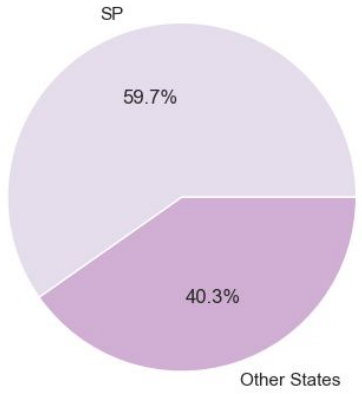


Présentation des données

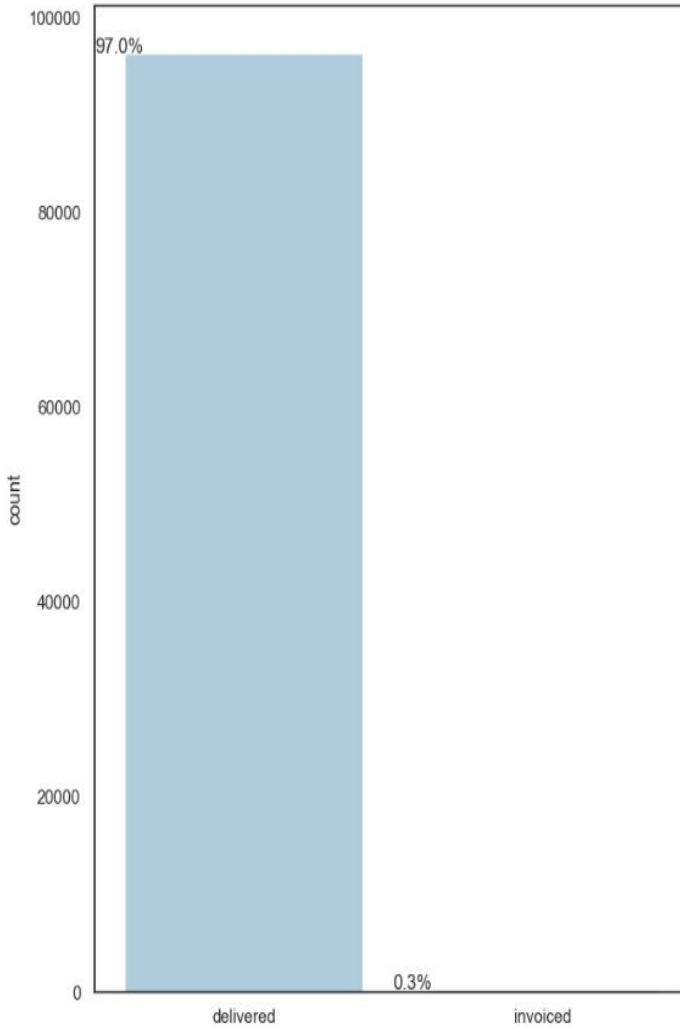
Acheteurs :



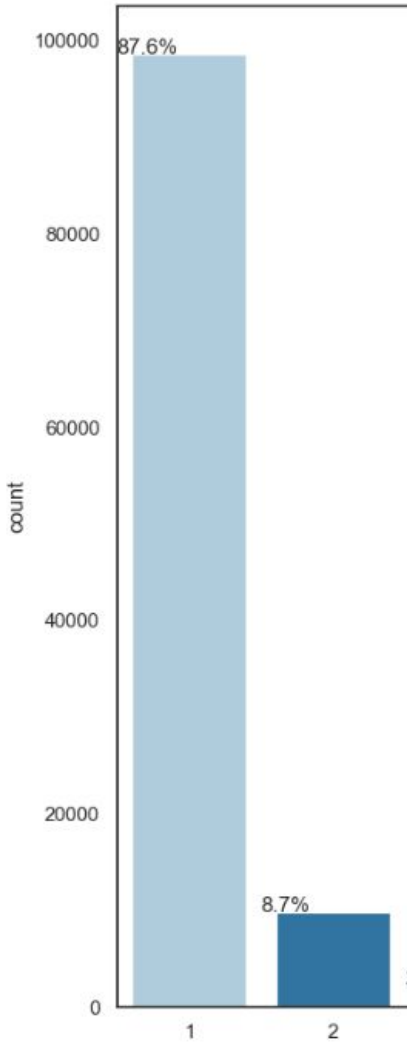
Vendeurs :



- **Commandes :**



- **Articles :**



olist

Démarche nettoyage et feature engineering

Nettoyage :

- Conversion en datetime pour les dates
- Remplacement des valeurs manquantes

Fusion des fichiers :

- Localisation et vendeurs/acheteurs
- Article et avis clients
- Création variables RFM (Récence : durée moyenne depuis la dernière commande, Fréquence : nombre de commandes, Montant : montant moyen des commandes)

• Fichier client final :

- customer_unique_id

Sur les commandes et le moyen de paiement
- 'nb_orders' (Fréquence)
- 'mean_price_order' (Montant)
- 'mean_payment_installments' (Nombre d'échéances)
- 'order_mean_delay' (Récence)

Sur les produits vendus
- 'mean_product_volume_cm3' (Volume moyen des produits)
- 'mean_review_score' (Note moyenne)
- 'total_items' (Nombre total d'items)

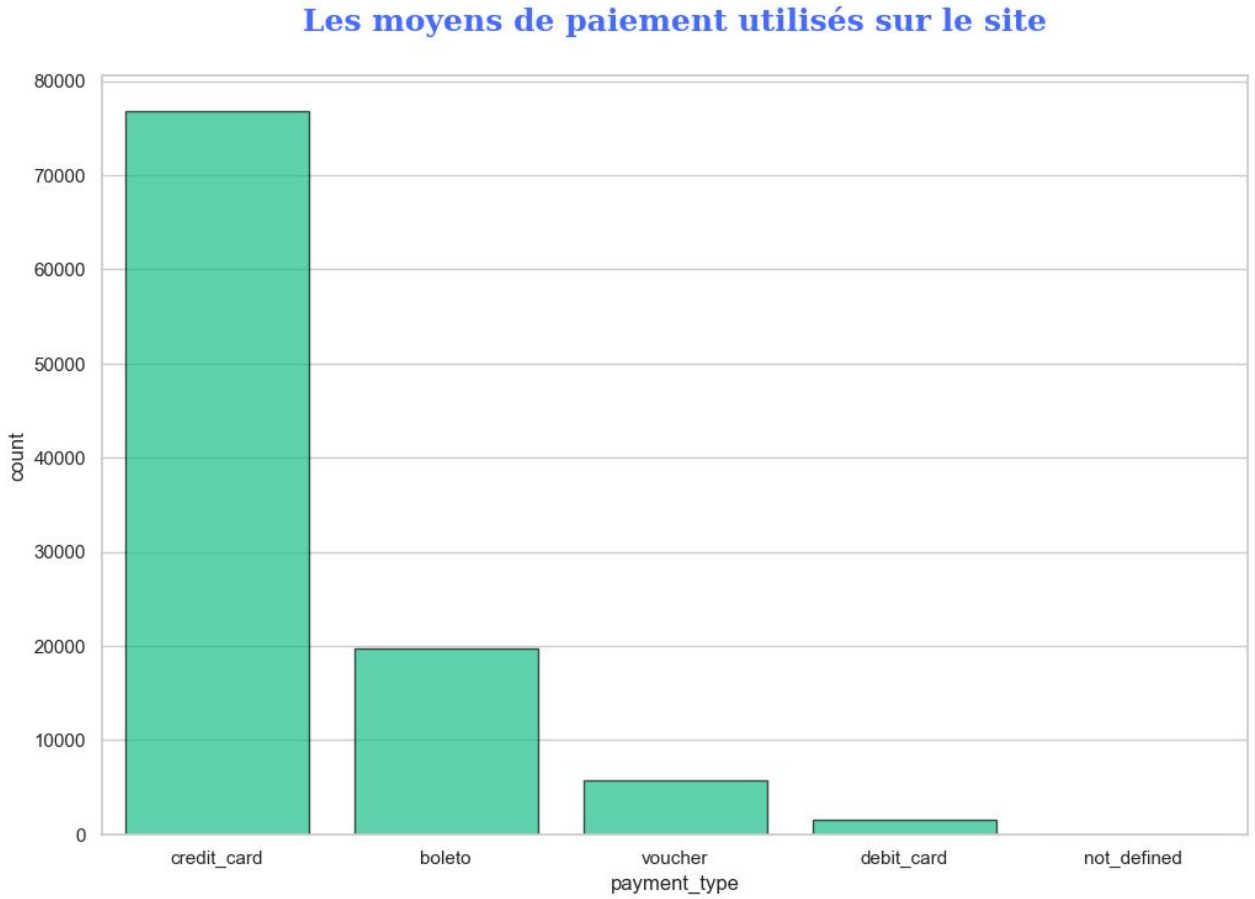
Sur les catégories de produits

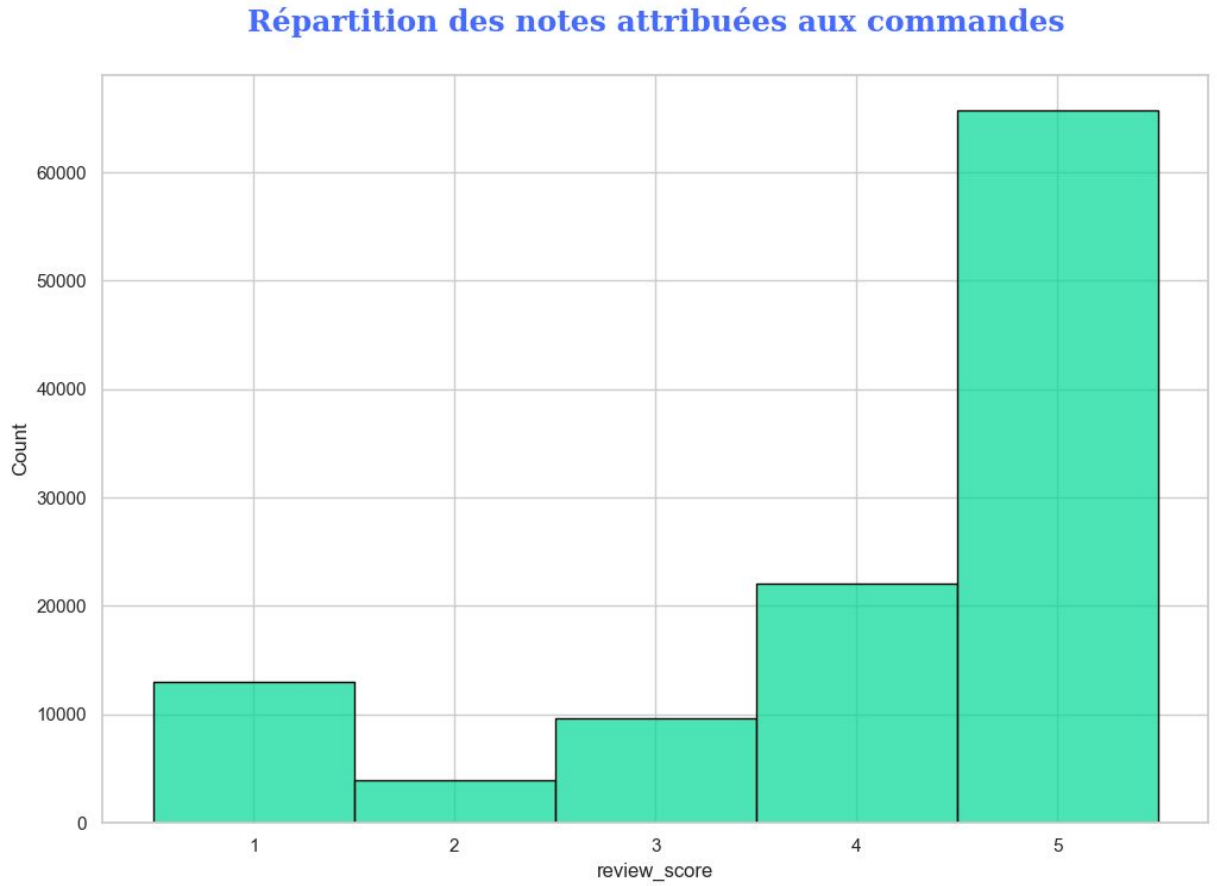
Sur la livraison
- 'mean_delivery_days' (Délai de livraison moyenne)
- 'customer_state' (État du client)
- 'order_freight_ratio' (Ratio du coût du fret par rapport au coût total)

Autres paramètres
- 'favorite_sale_month' (Mois préféré)
- 'distance_seller_customer' (Distance vendeur-acheteur)

olist

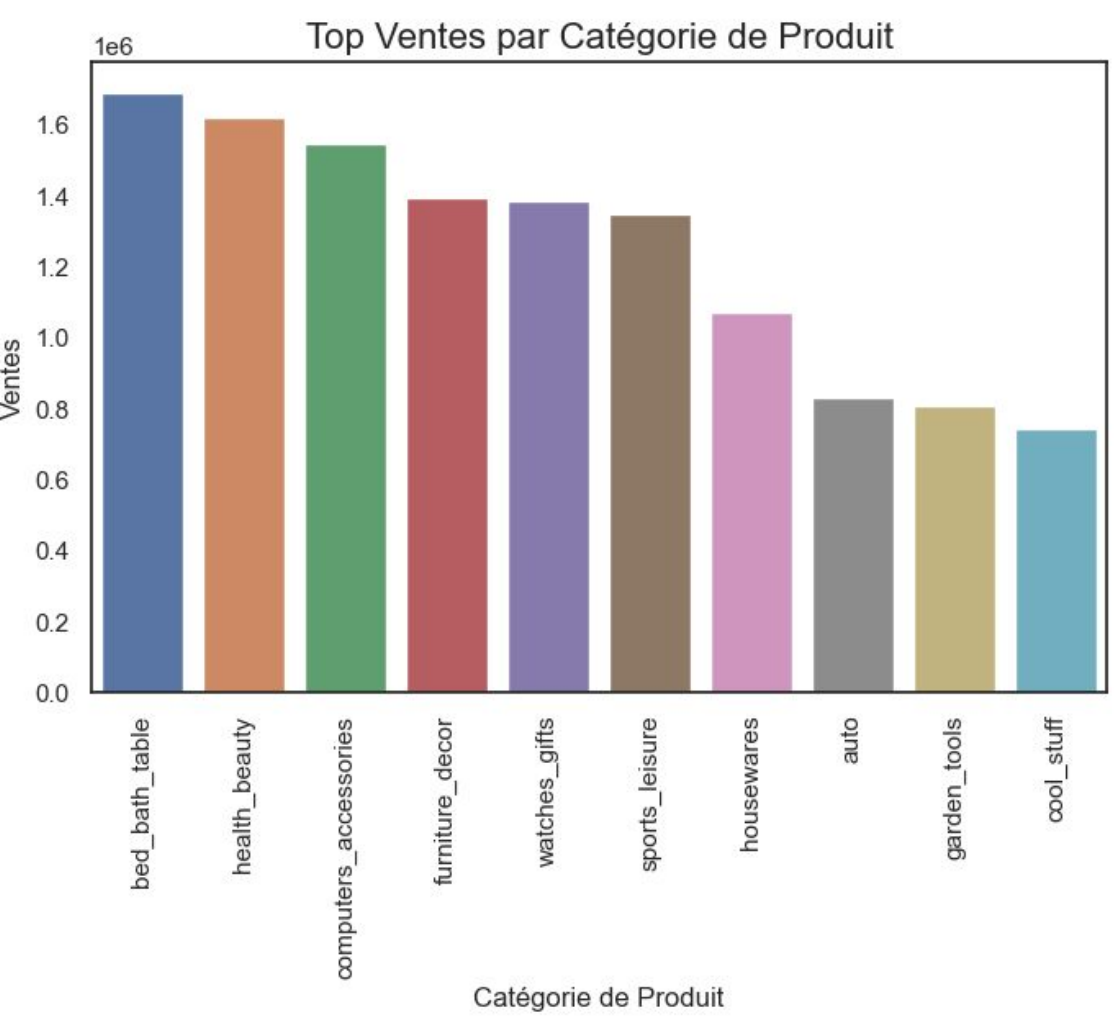
Analyse exploratoire des données

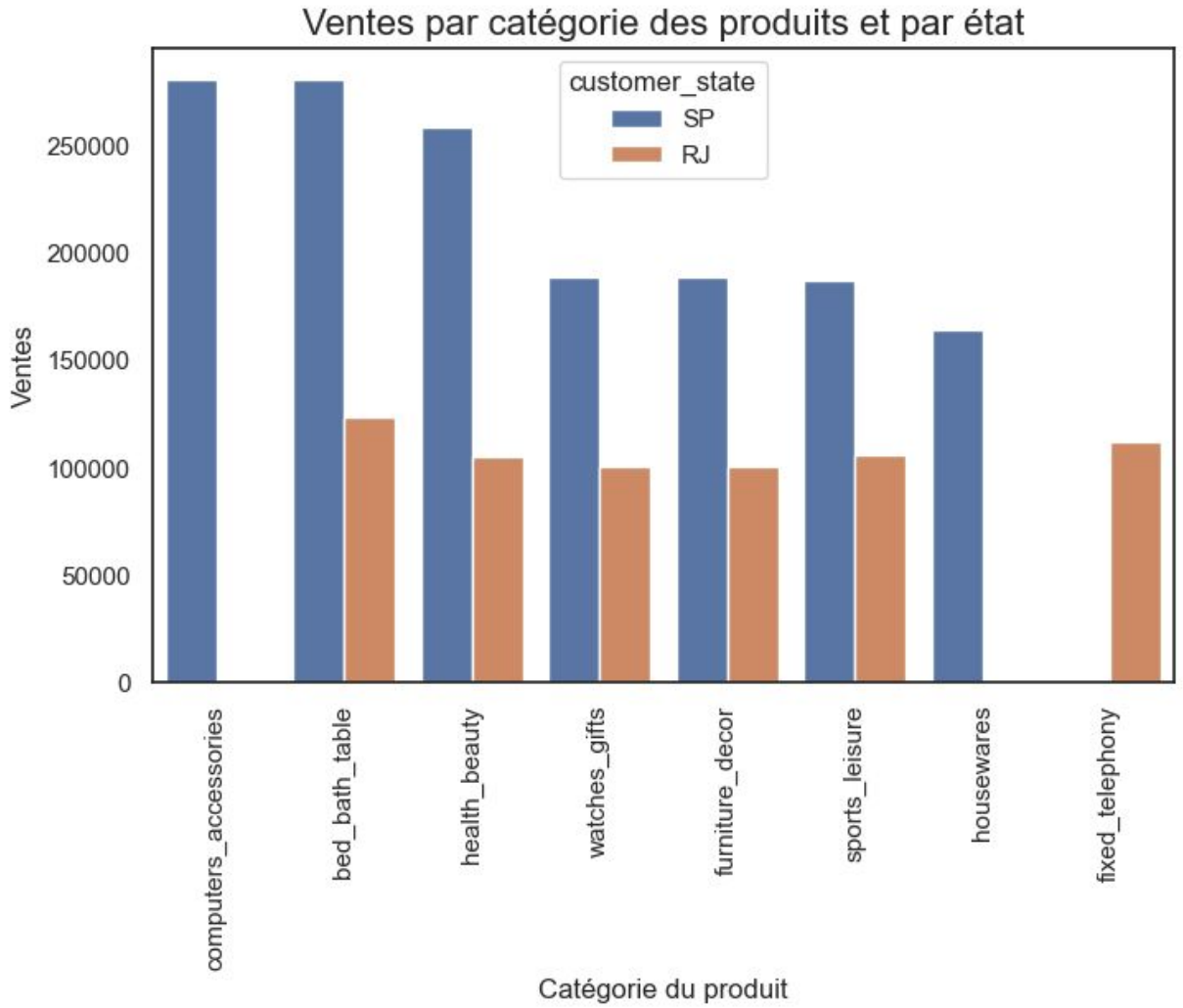




olist

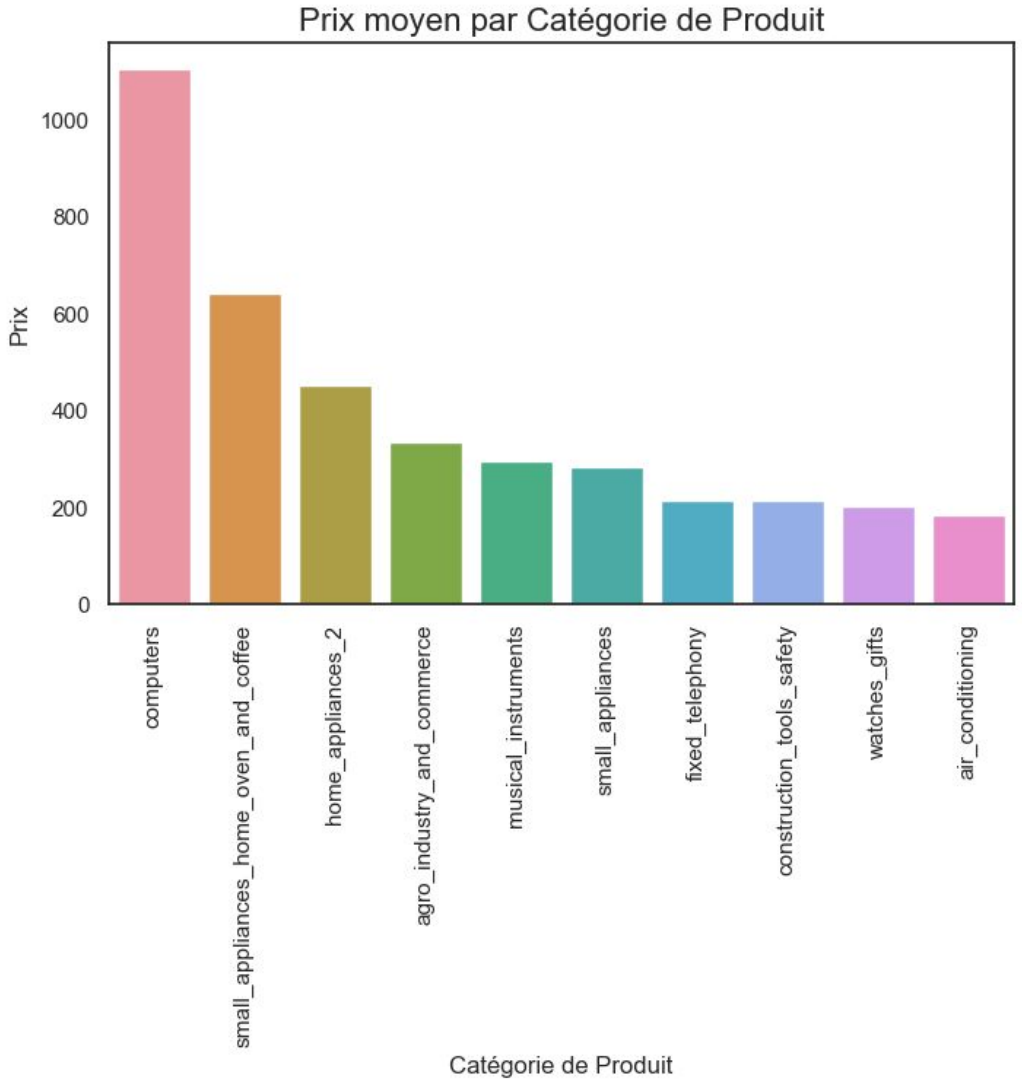
Analyse exploratoire des données

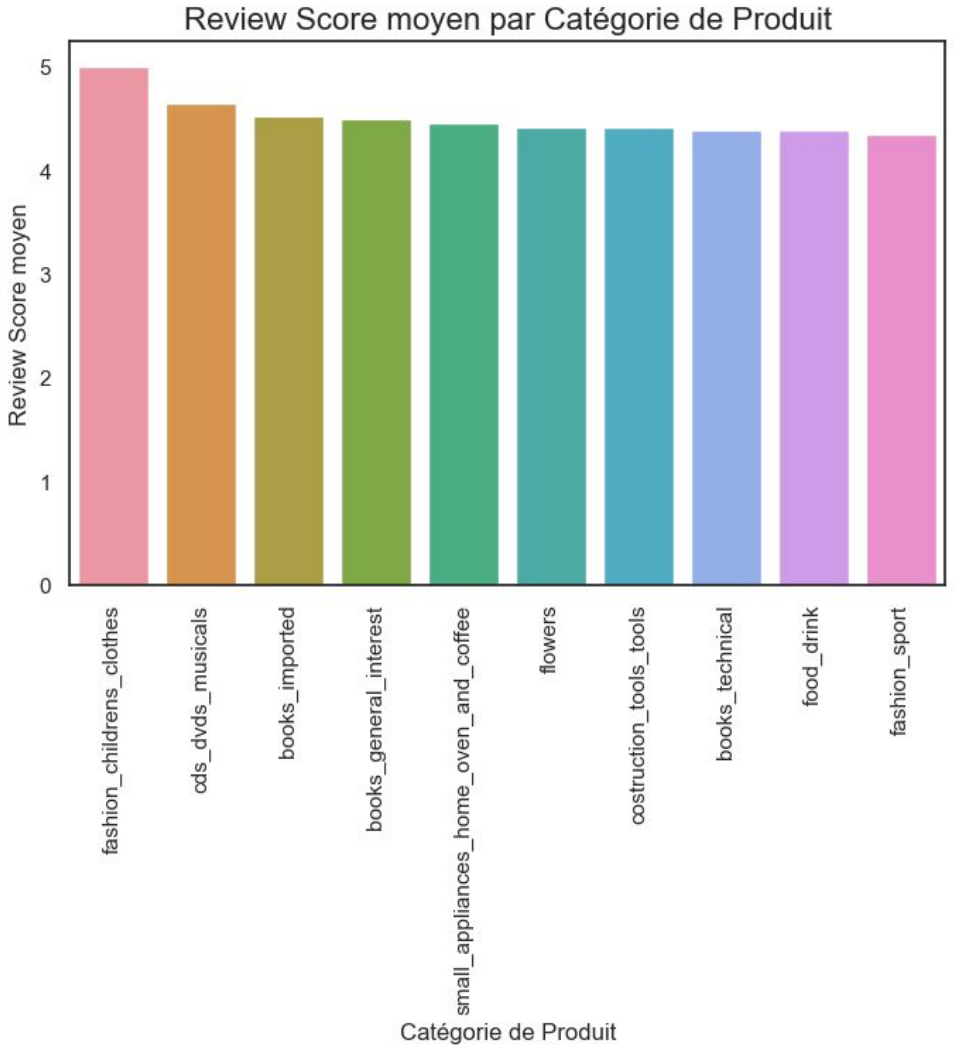




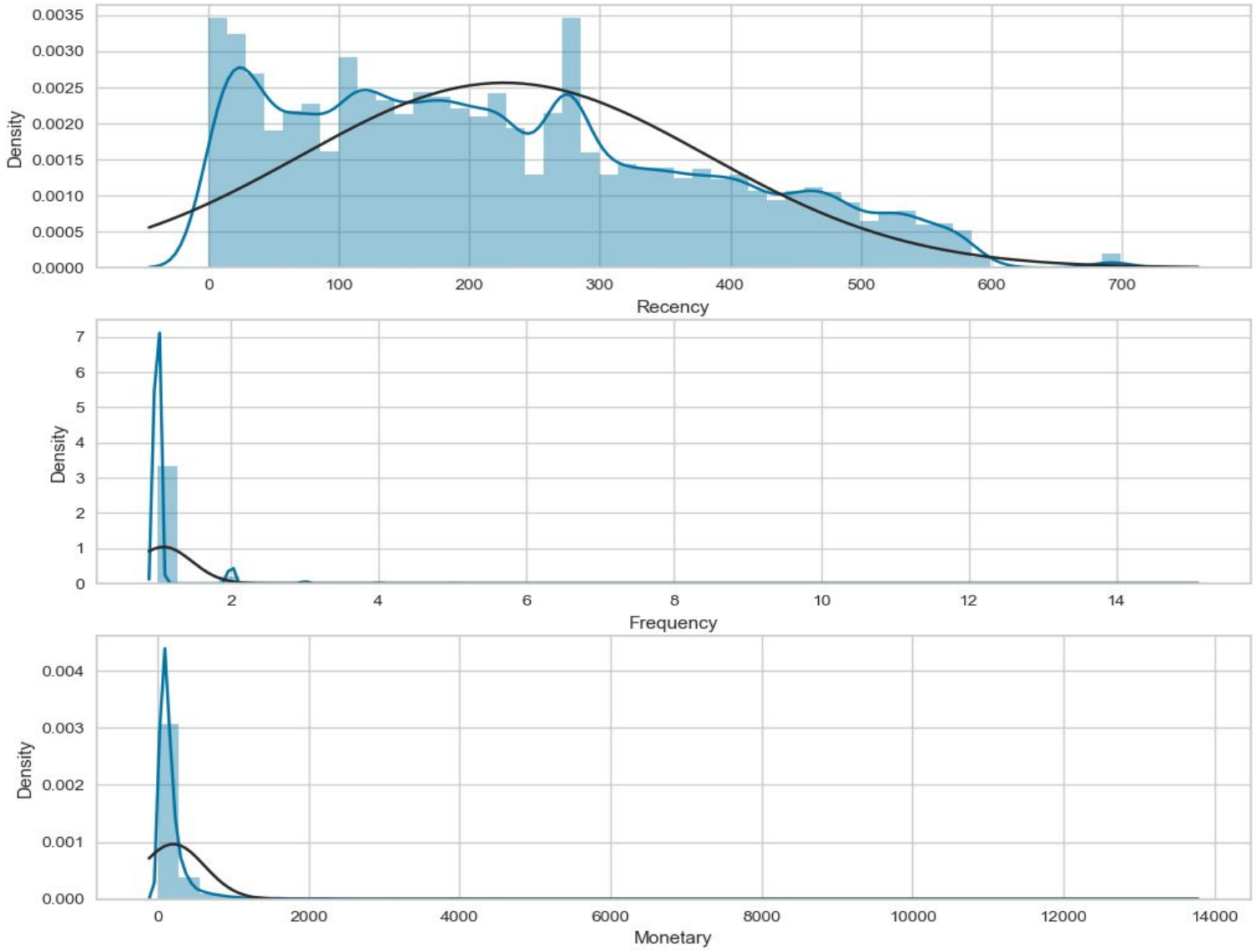
olist

Analyse exploratoire des données





Analyse exploratoire des données



olist

Pistes de modélisation

1. Clustering RFM Manuel

2. Clustering avec l'algorithme du K-Means (RFM seulement, RFM avec quelques variables, toutes les variables sans les catégories de produits)

3. K-Means après réduction de dimensions

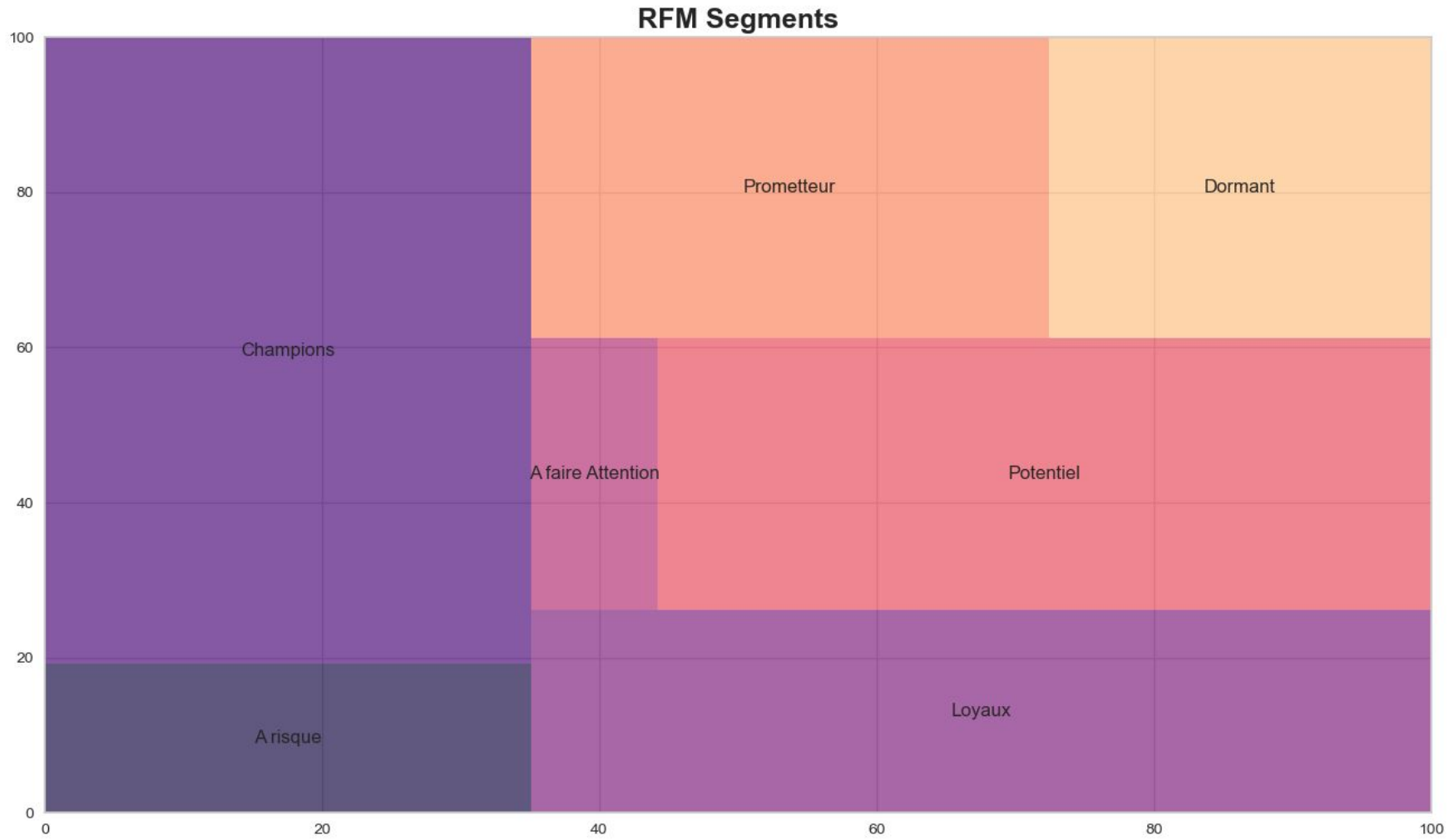
4. Clustering Hiérarchique

5. Clustering DBscan

Pistes de modélisation - Clustering RFM Manuel

Les meilleurs clients ont une fréquence d'achats élevée, une récence basse et un montant de dépense élevé.

Le calcul est basé sur le RFM Score (méthode des quintiles).

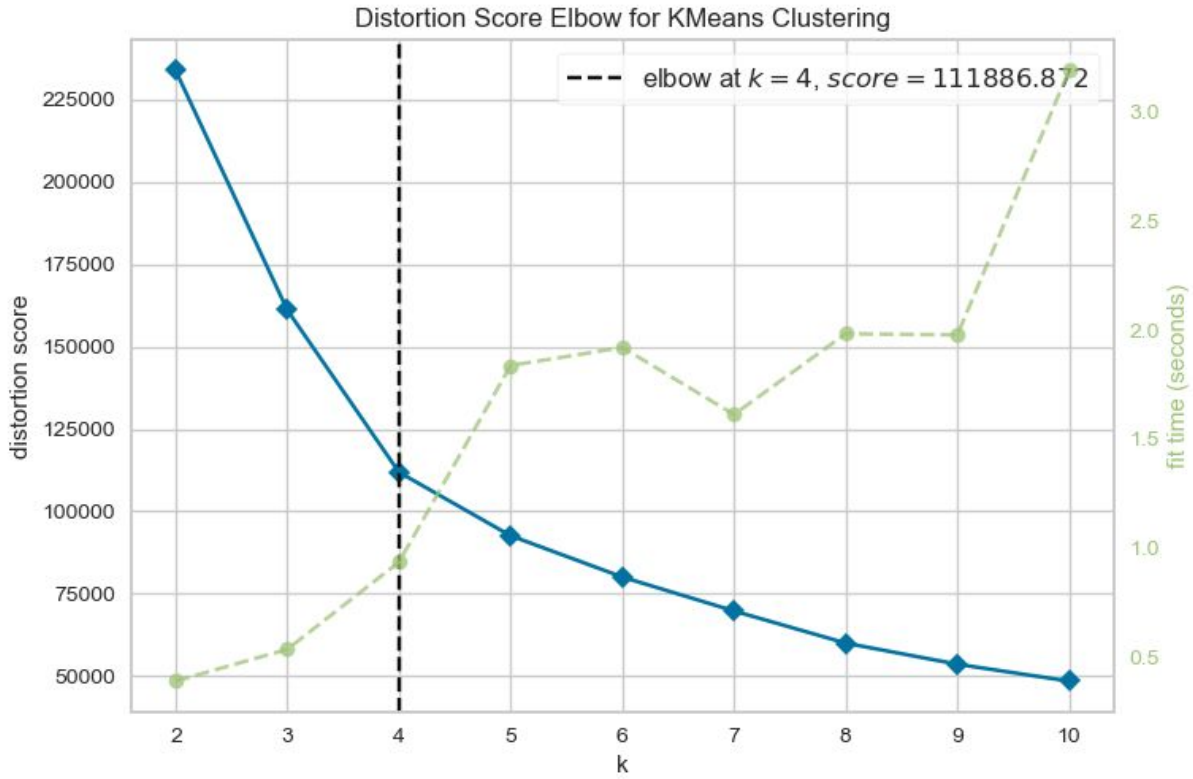


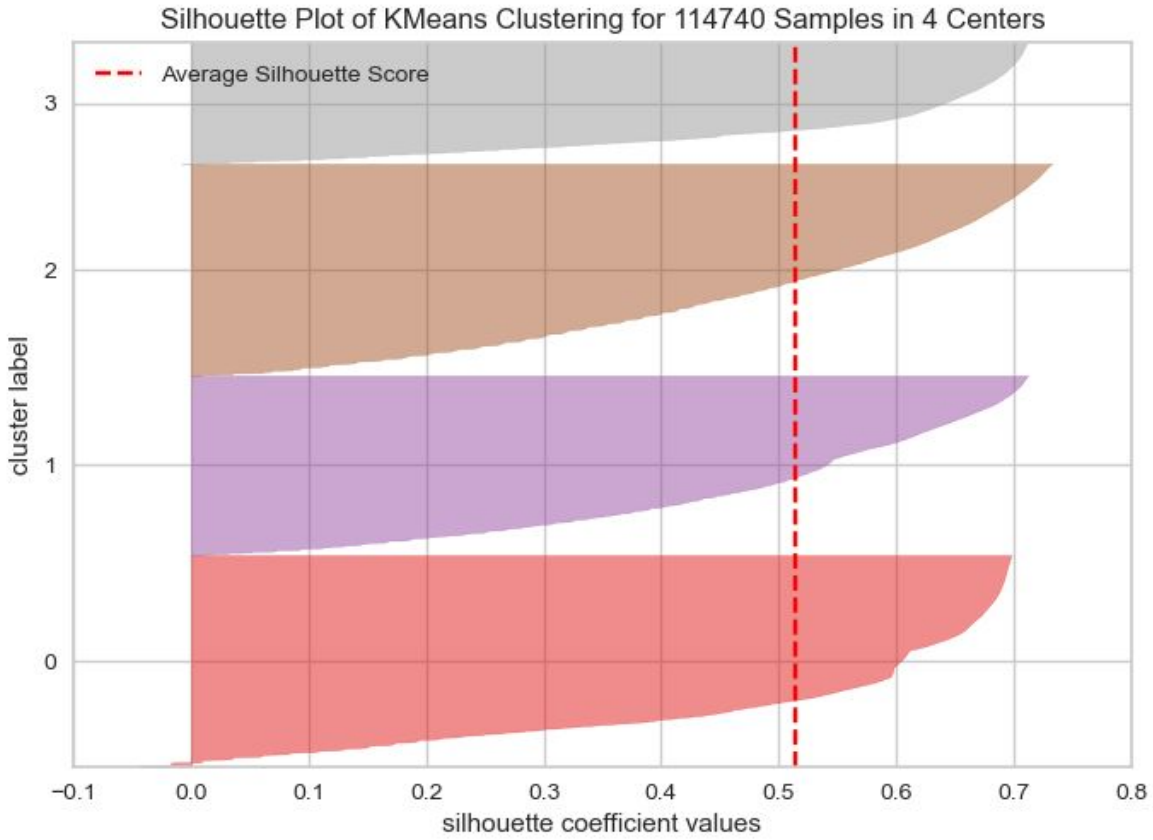
olist

Pistes de modélisation - Clustering Kmeans

KMeans avec RFM seulement :

- Preprocessing : Log + Normalisation pour un skewness > 1
- KMeans de 2 à 12 centroïdes

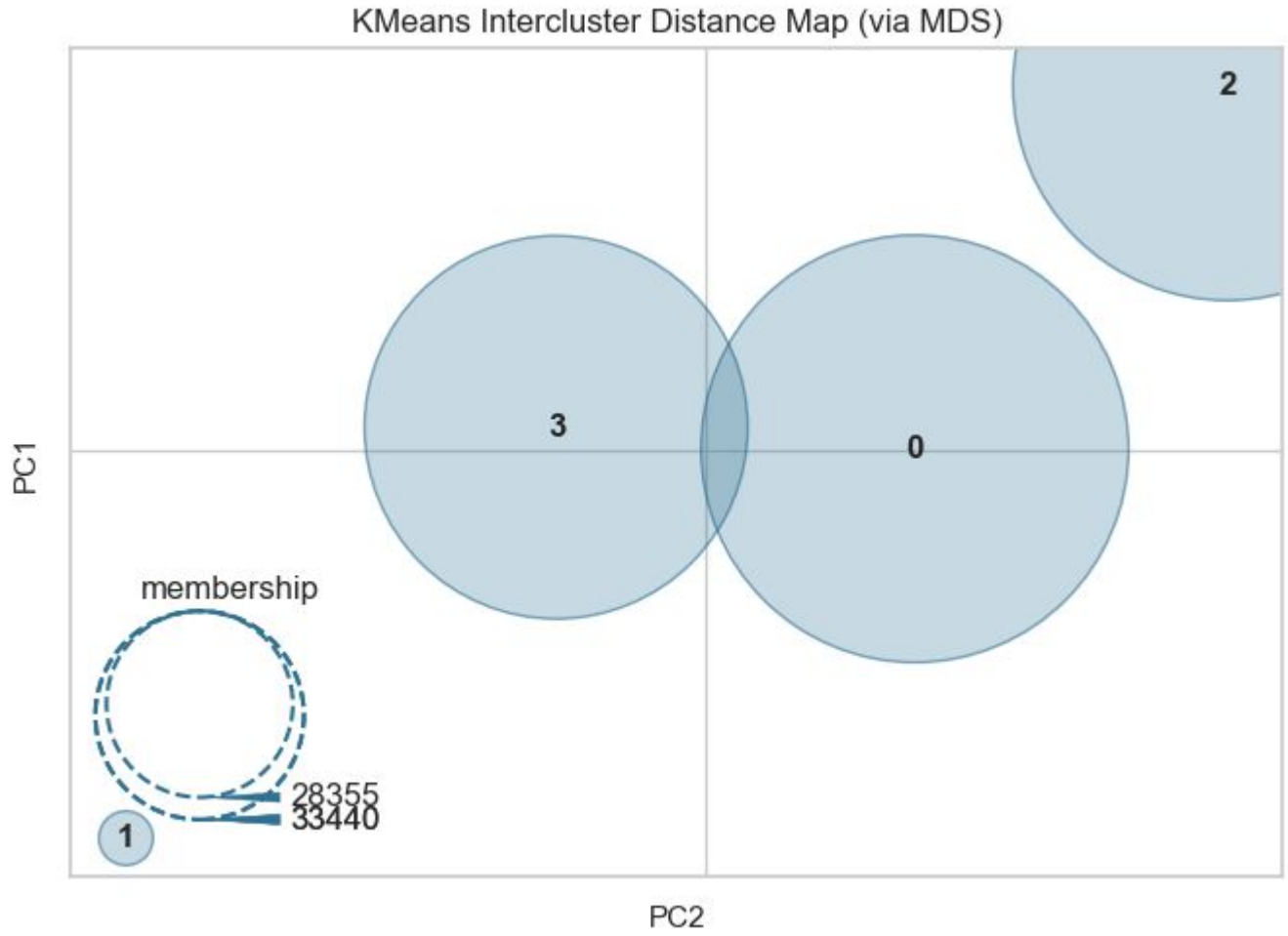




olist

Pistes de modélisation - Clustering Kmeans

KMeans avec RFM seulement :

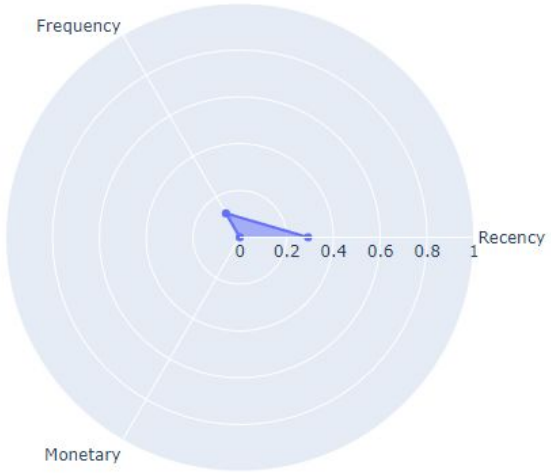


olist

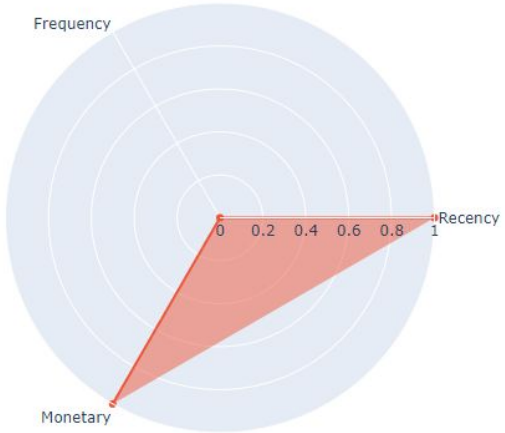
Pistes de modélisation - Clustering Kmeans

KMeans avec RFM seulement (Radar plot) :

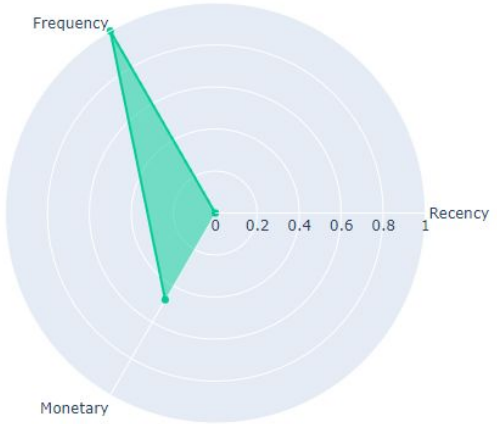
Clients nouveaux :



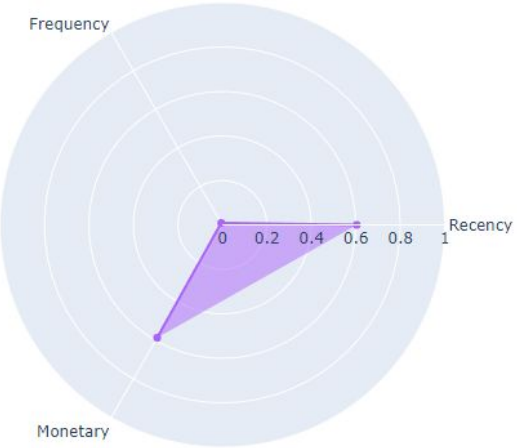
Clients à risque :



Clients prometteurs :

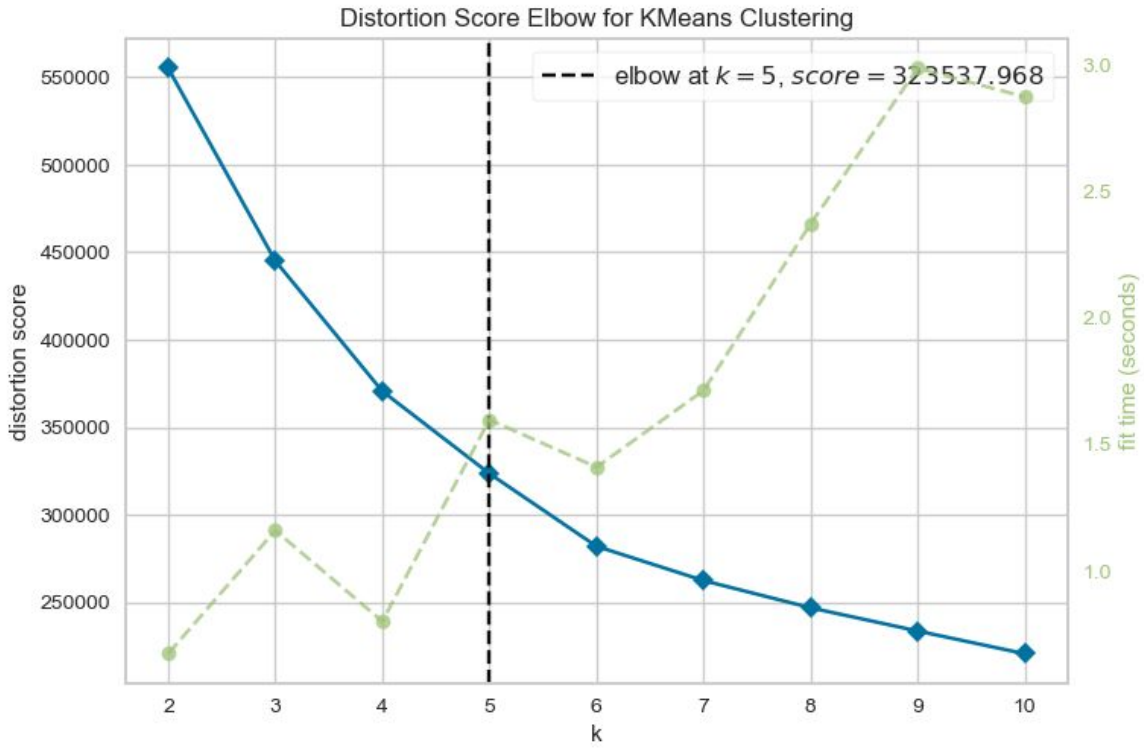


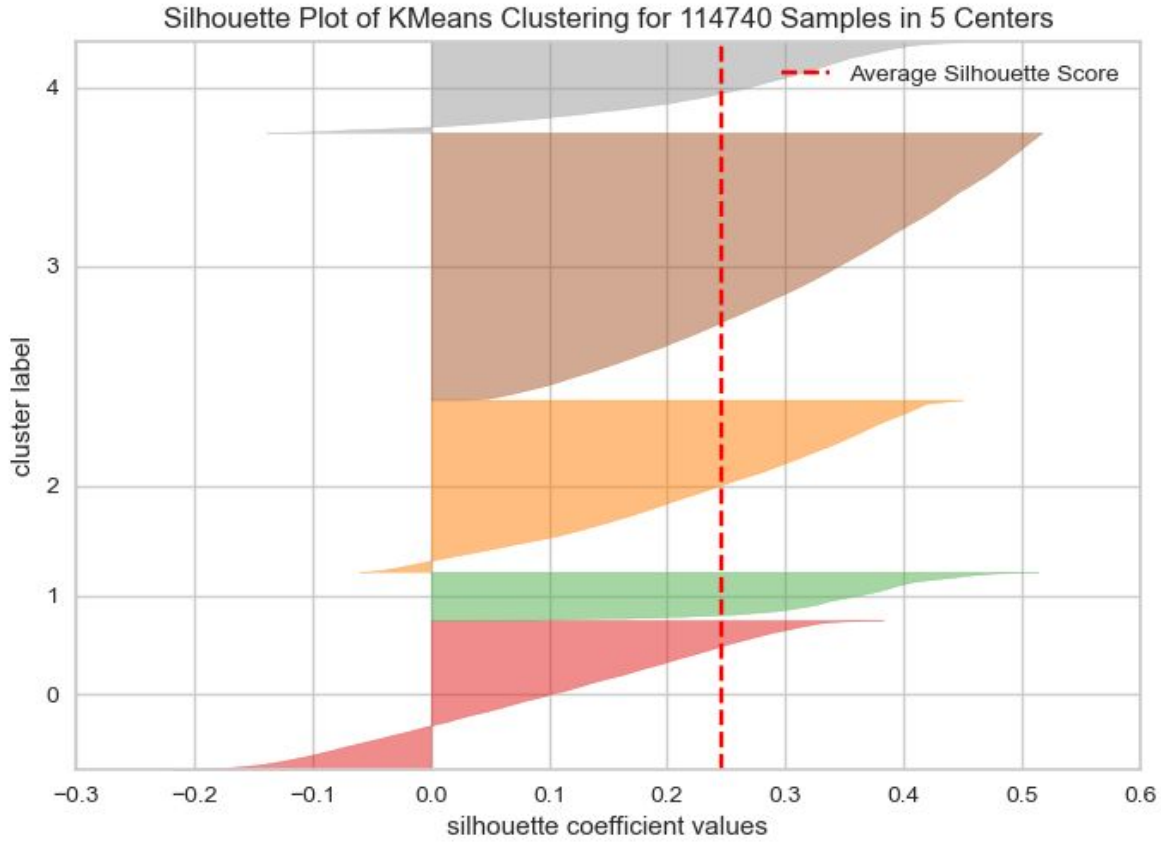
Clients à activer :



Pistes de modélisation - Clustering Kmeans

KMeans avec RFM + quelques paramètres : Note produits, nombre d'items, échéances de paiement



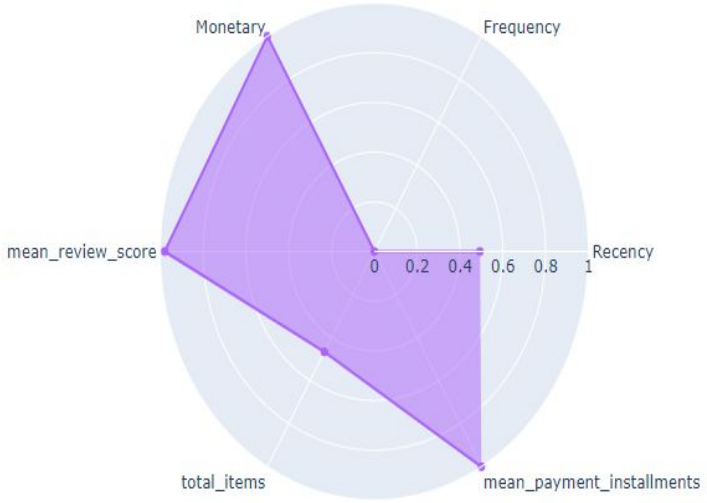


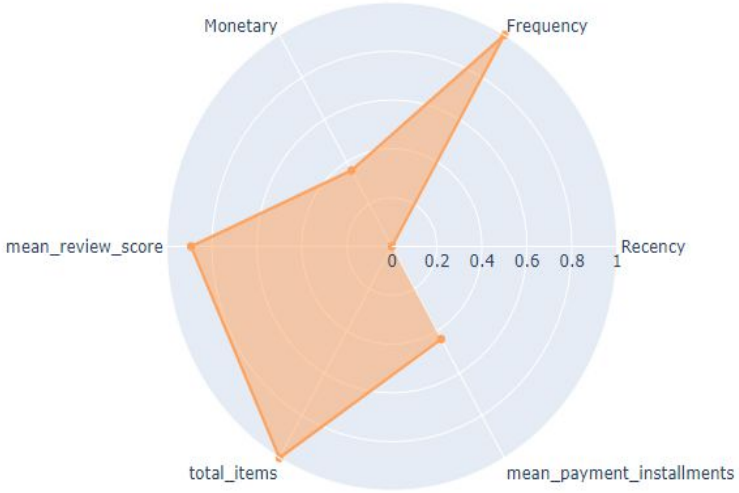
olist

Pistes de modélisation - Clustering Kmeans

KMeans avec RFM + quelques paramètres (Radar plot) : Note produits, nombre d'items, échéances paiement, distance, saisonnier ou pas

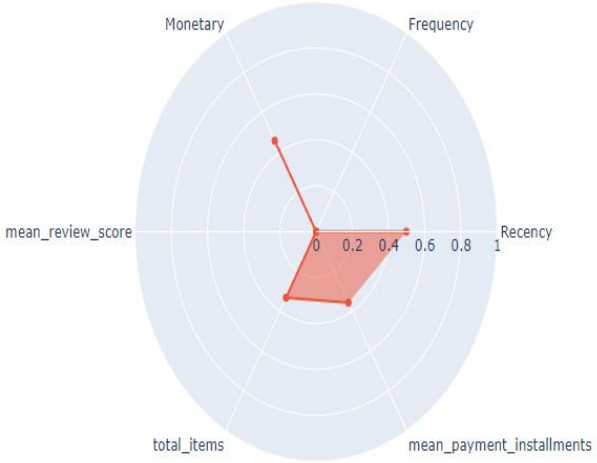
Clients prometteurs



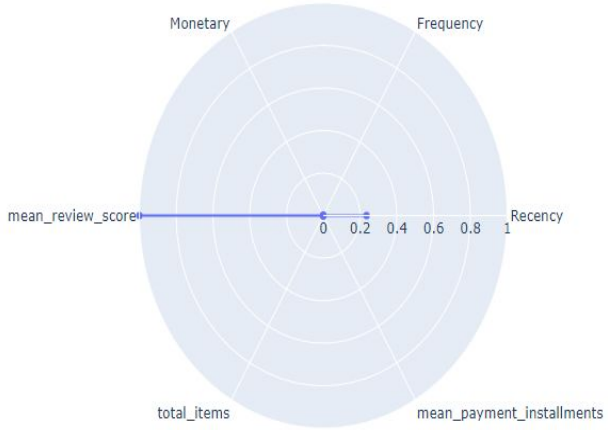


Clients fidèles

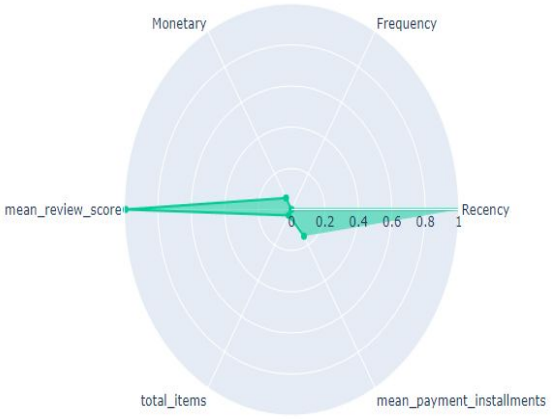
Clients à activer



Clients à risque

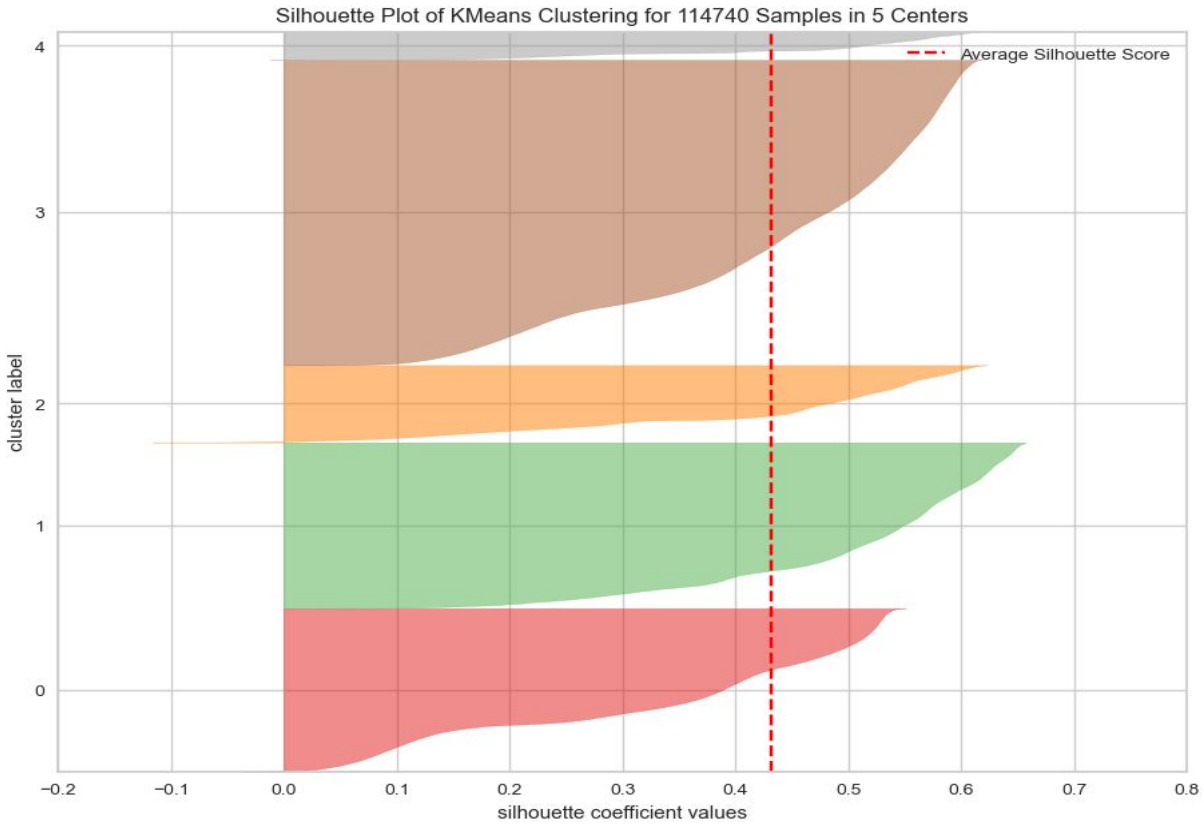


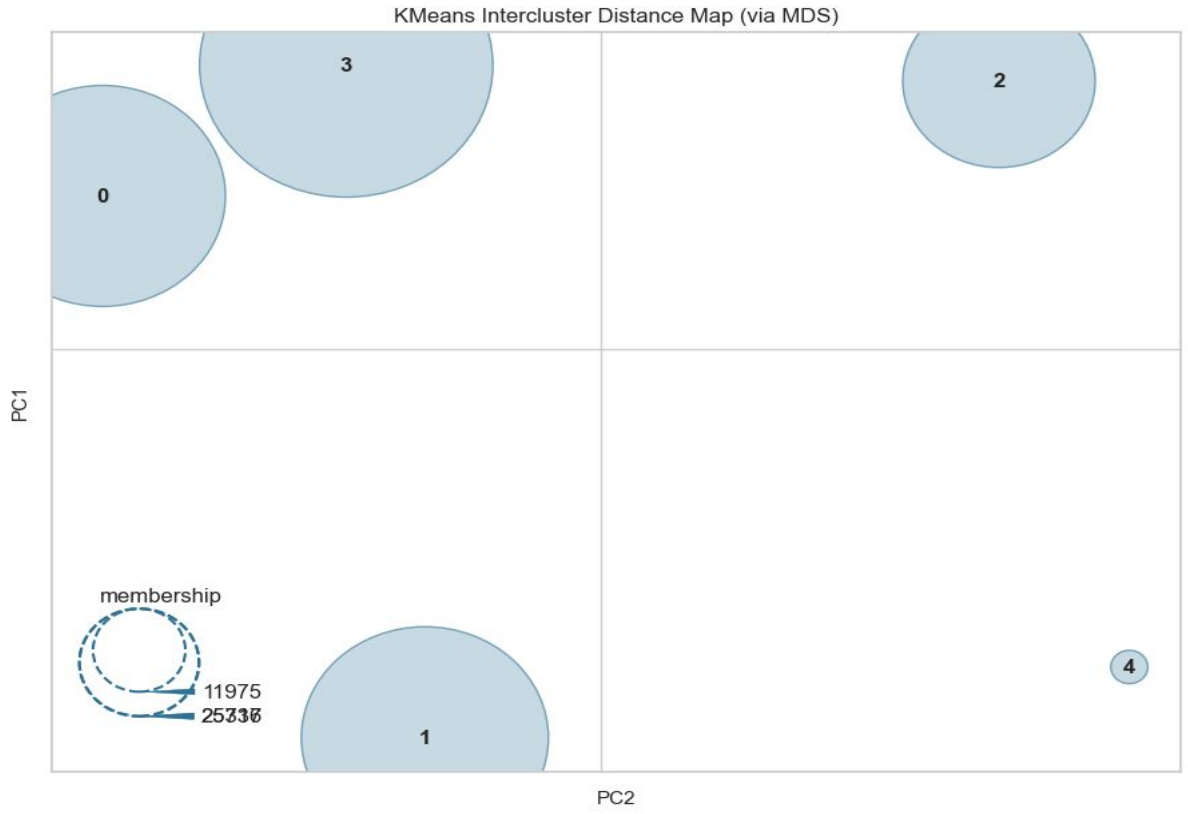
Clients perdus



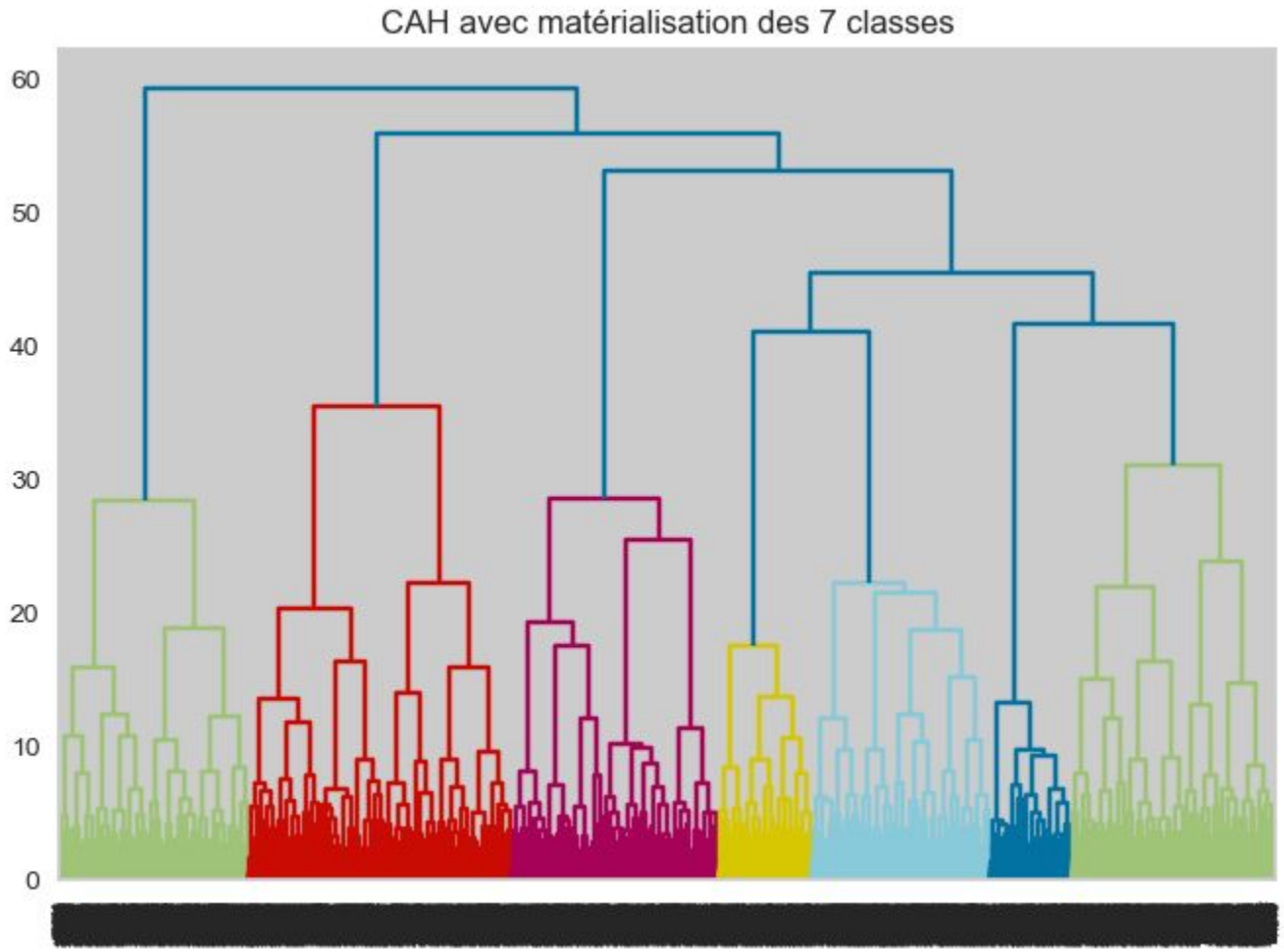
Pistes de modélisation - Clustering Kmeans

KMeans avec tous les paramètres (sauf catégories de produits) + PCA appliquée :

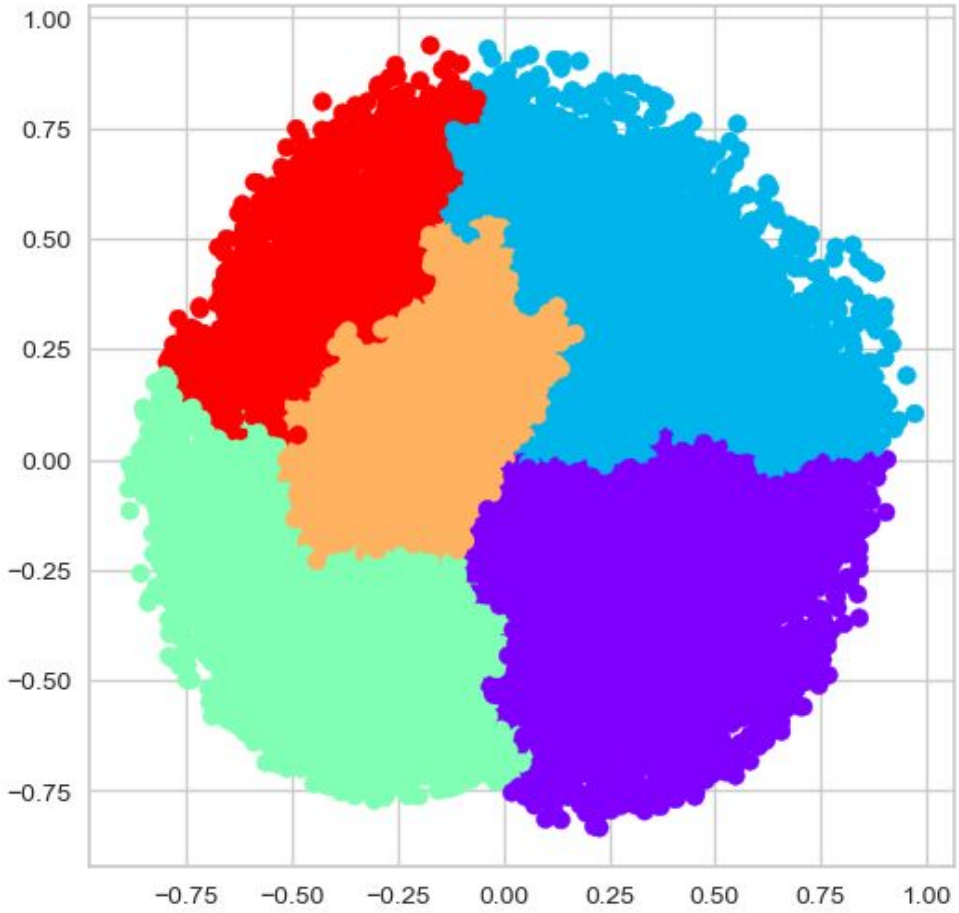




Pistes de modélisation - Clustering Hiérarchique et agglomératif



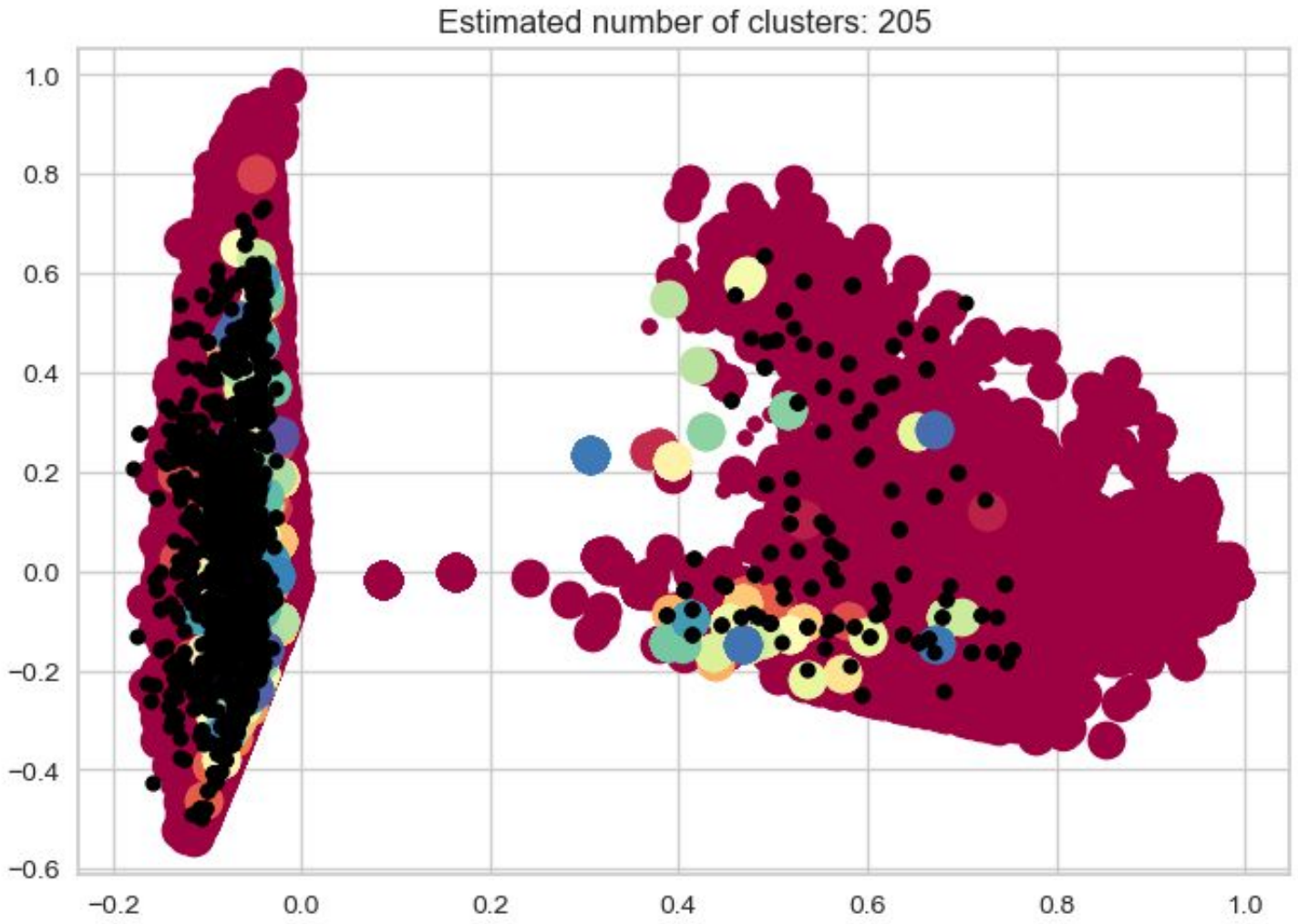
Clustering hiérarchique



Clustering agglomératif K=5 (meilleur Silhouette Score)

olist

Pistes de modélisation - Clustering DBScan avec PCA



- DBscan n'est pas adapté à notre problématique, la densité des bons clients (qui ont commandé plusieurs fois) étant faible.

olist

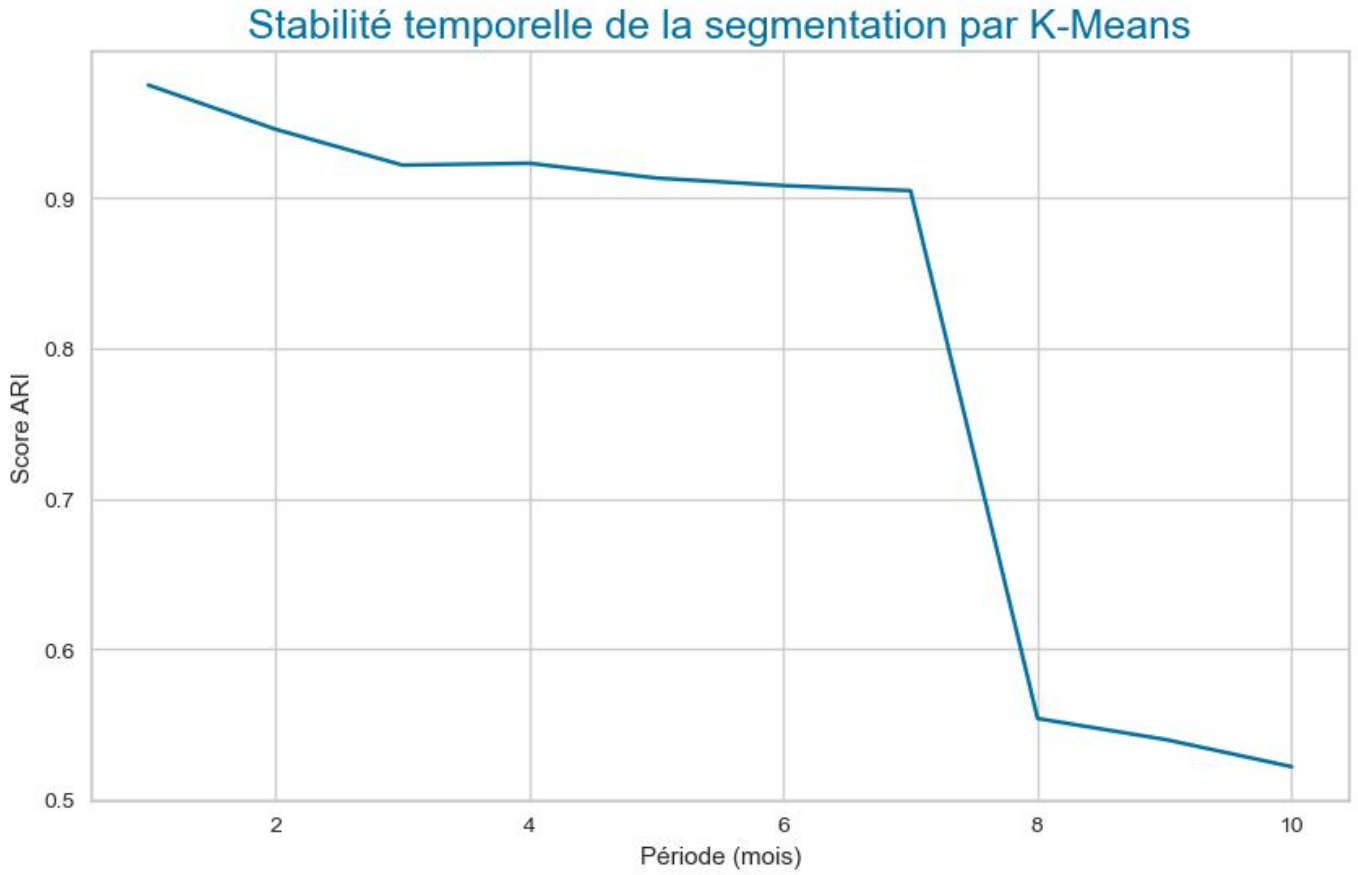
Modèle sélectionné

Clustering avec l'algorithme du K-Means :

- Cluster 0 : Clients à risque - achètent peu, notent bien les produits.
- Cluster 1 : Clients à activer - achats récents, profitent du paiement en plusieurs échéances, achètent plusieurs items.
- Cluster 2 : Clients perdus - n'ont pas acheté depuis longtemps même s'ils ont bien noté les produits.
- Cluster 3 : Clients prometteurs - dépensent beaucoup et notent bien les produits.
- Cluster 4 : Clients fidèles - achètent régulièrement.

Contrat de maintenance

- La période d'achat est de 23 mois
- Première simulation avec les données existantes à t = 12 mois.
- Refaire des simulations en ajoutant 1 mois supplémentaire.
- Évaluer la cohérence entre les clusters de départ et les partitionnements trouvés en utilisant l'indice ARI (Ajusted Rand Index).
- On remarque une forte inflexion après 8 mois sur les clients initiaux.



olist

Conclusion

Modèles de clustering :

- Identification des différents types d'utilisateurs à partir d'un modèle de segmentation afin de fournir à l'équipe de marketing une description des clients.
- Le RFM Manuel nous permet de définir des groupes, cependant cela nécessite de la connaissance métier pour établir les seuils.
- Le KMeans nous donne 5 clusters avec RFM et quelques autres paramètres.
- Le clustering hiérarchique et le clustering agglomératif donnent des résultats similaires au KMeans.
- DBscan n'est pas adapté à notre problématique, la densité des bons clients (qui ont commandé plusieurs fois) étant faible.

Evaluation de la stabilité temporelle du modèle :

Prévoir la maintenance du programme de segmentation tous les 8 mois dans un premier temps puis re-tester cette stabilité temporelle au fil du temps afin de l'affiner.

olist

Merci pour votre attention.

OPENCLASSROOMS

olist