Segmentez des clients d'un site e-commerce

Projet 5

SOMMAIRE

- Le rappel de la problématique
- La démarche de nettoyage et le feature engineering
- L'analyse exploratoire
- Les pistes de modélisation
- Le contrat de maintenance
- La conclusion

Rappel de la problématique

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

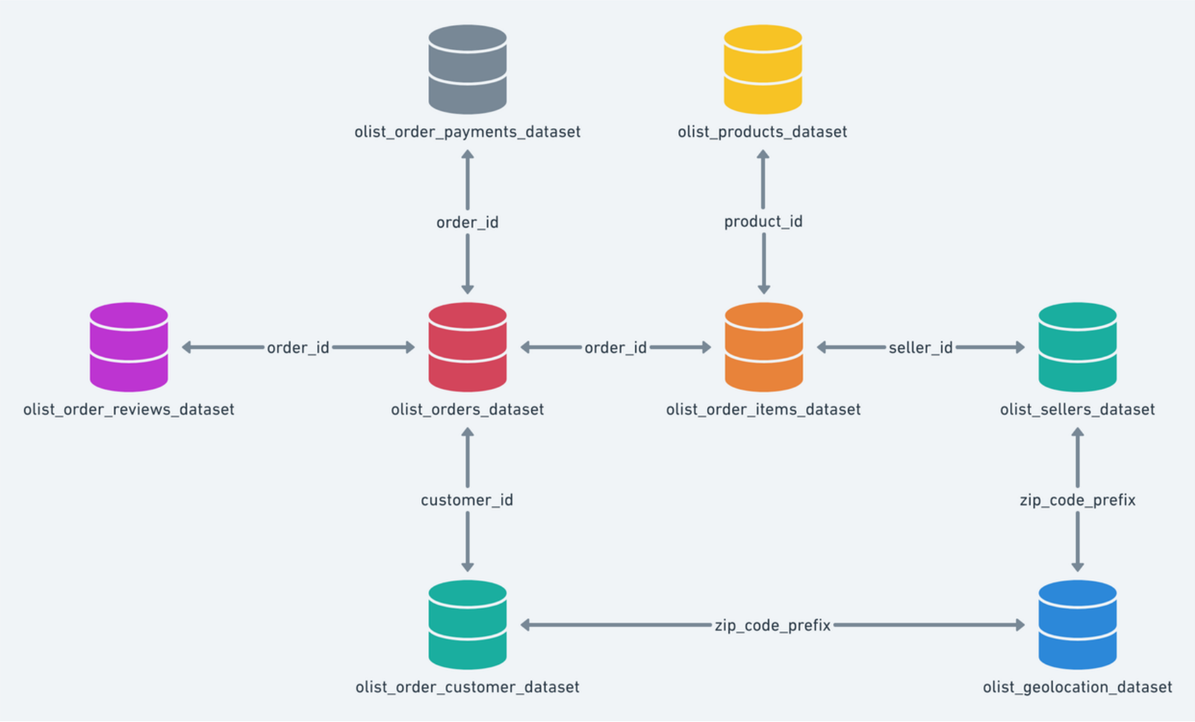
Les objectifs:

- **Comprendre les différents types d’utilisateurs**
- **Fournir à l’équipe marketing une description actionable**
- **Proposer un contrat de maintenance**

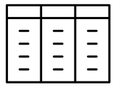
Les données

Brazilian E-Commerce Public Dataset by Olist

9 fichiers comportant des informations sur l'historique de commandes, les produits achetés, les commentaires de satisfaction, et la localisation des clients entre 2016 et 2018



La démarche de nettoyage et le feature engineering



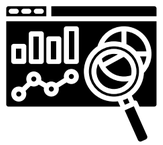
la jointure des différentes tables

Création de features par client:

- Le nombre de commandes
- Le panier moyen
- la catégorie la plus achetée
- Moyenne du nombre de paiements
- Note moyenne des commentaires

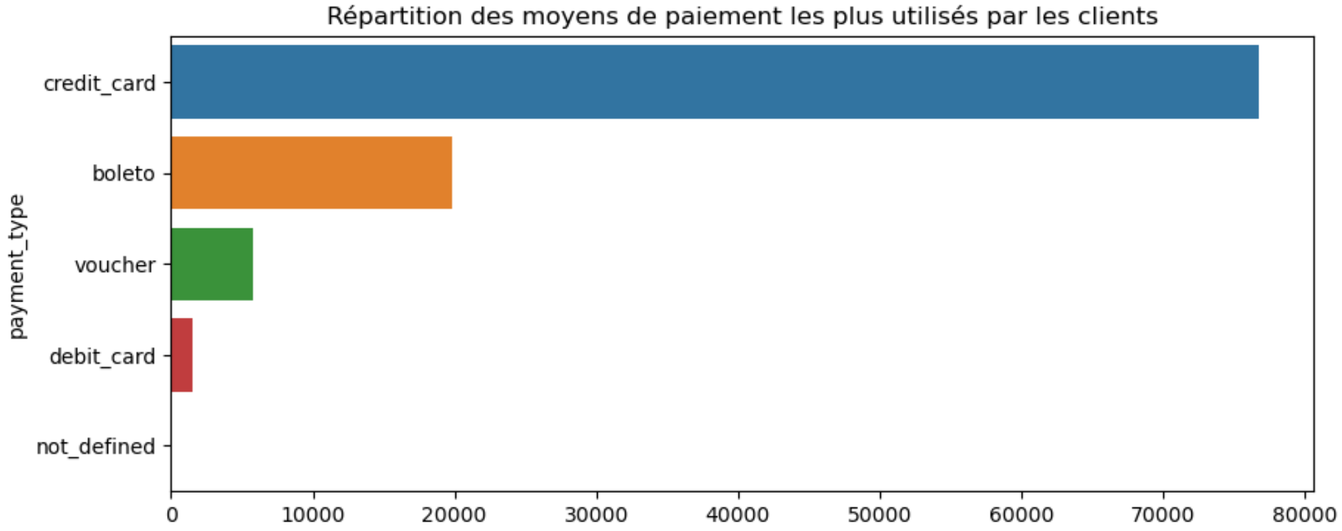
Création de variables pour la segmentation RFM:

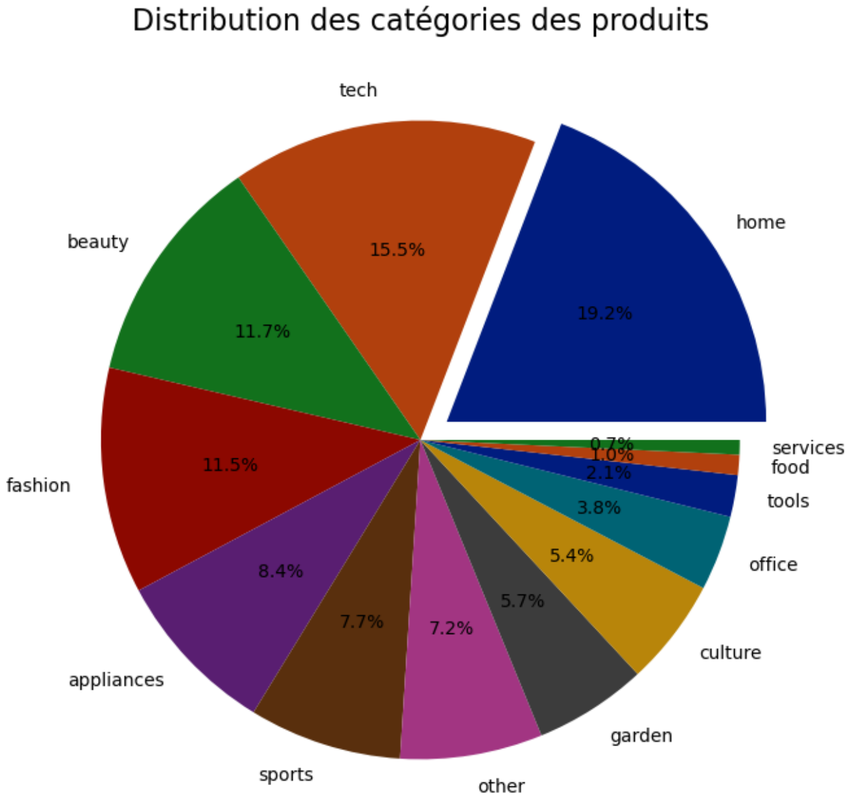
- Récence
- Frequence
- Montant



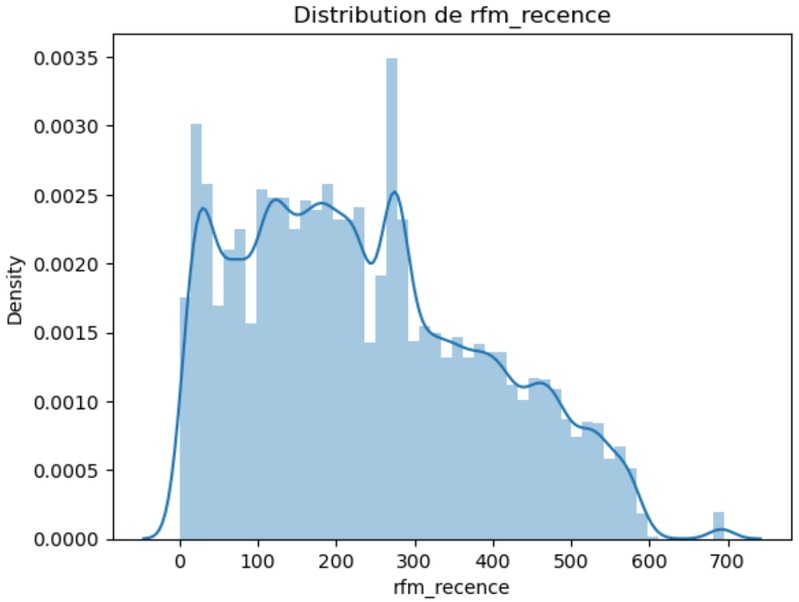
Analyse exploratoire

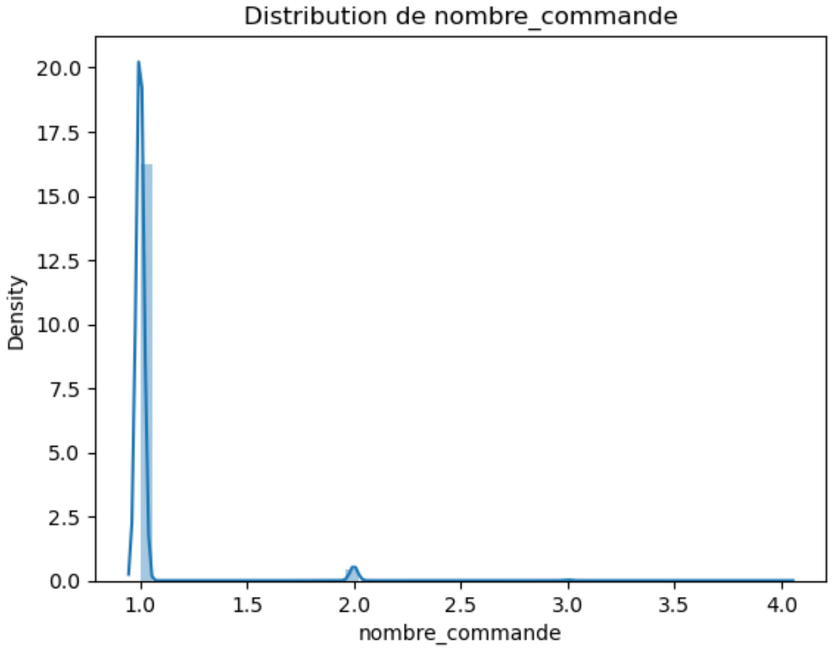
Analyse exploratoire

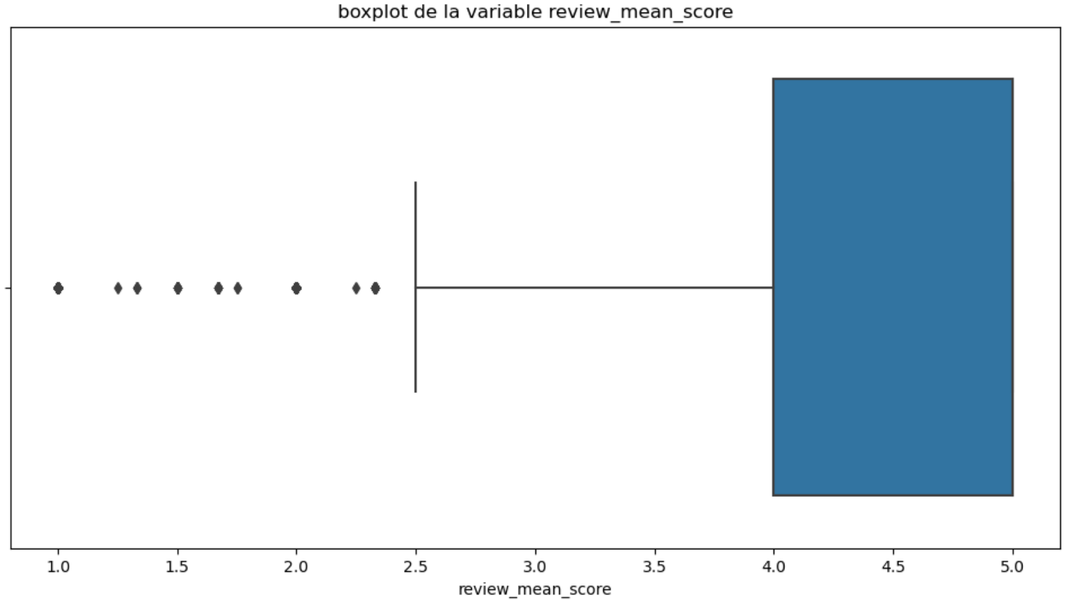




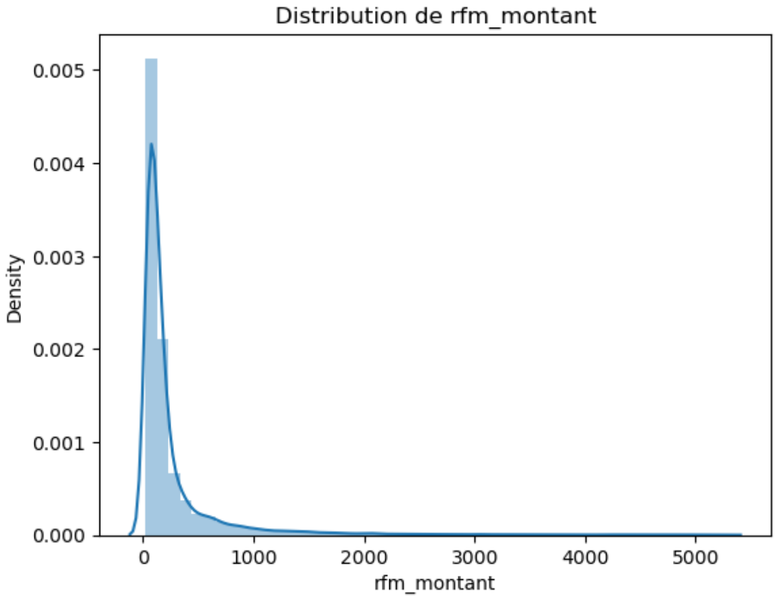
Analyse exploratoire

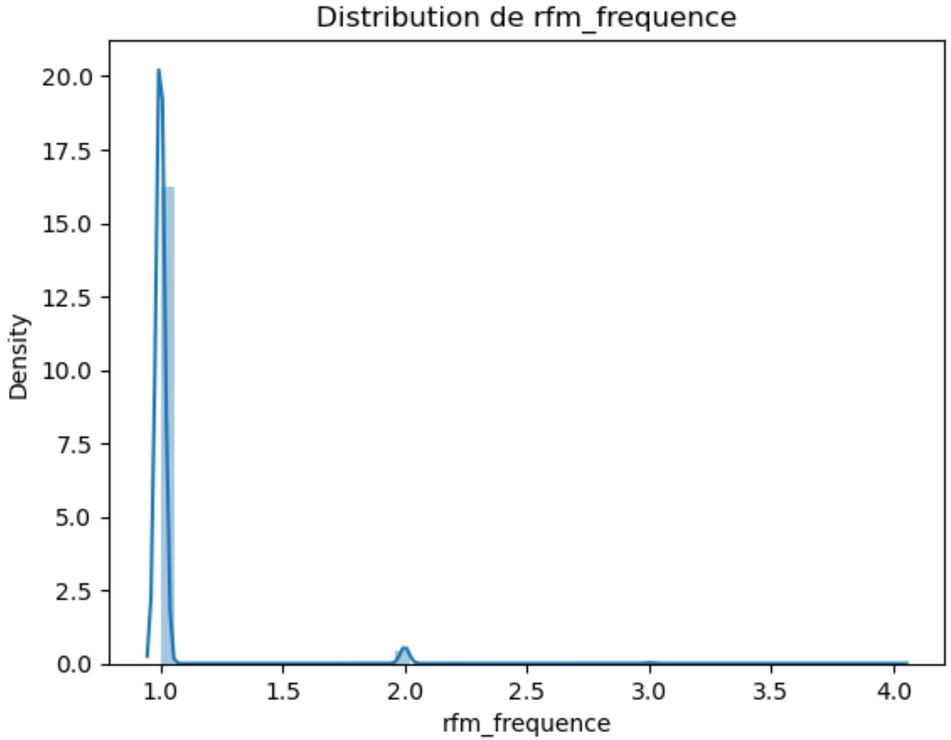


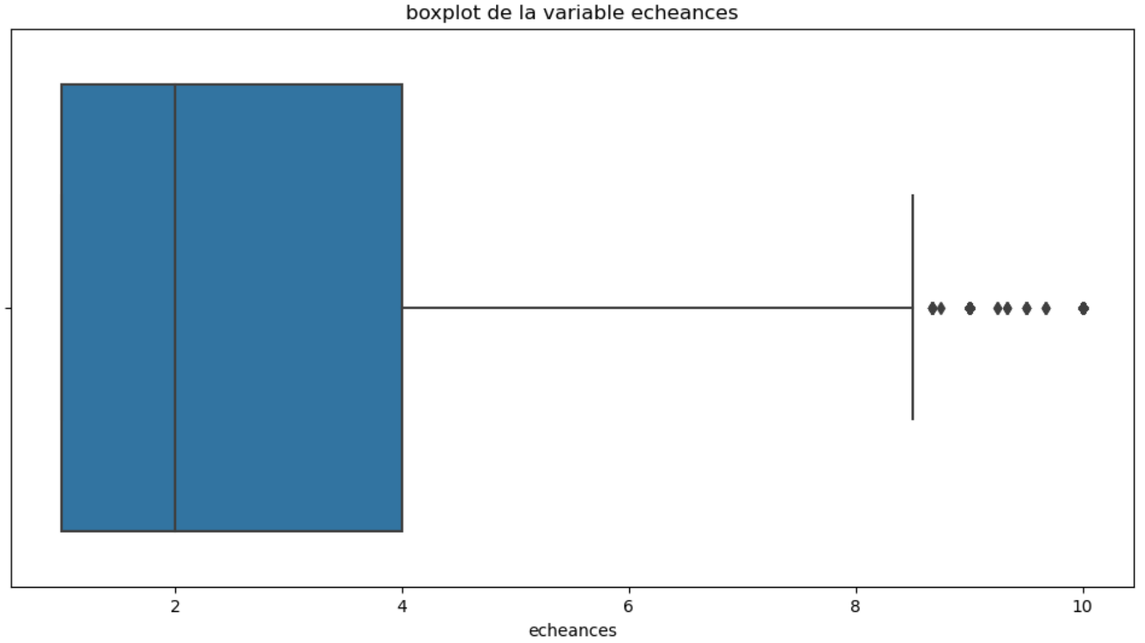




Analyse exploratoire







Modélisation

01

Preprocessing:

- Transformation en log
- StandardScaler

02

K-Means

- Détermination de la valeur K
  - La méthode d’Elbow,
  - Le coefficient de silhouette
- La répartition des clients
- Visualisation des clusters
- Interprétation des profils des clients

05

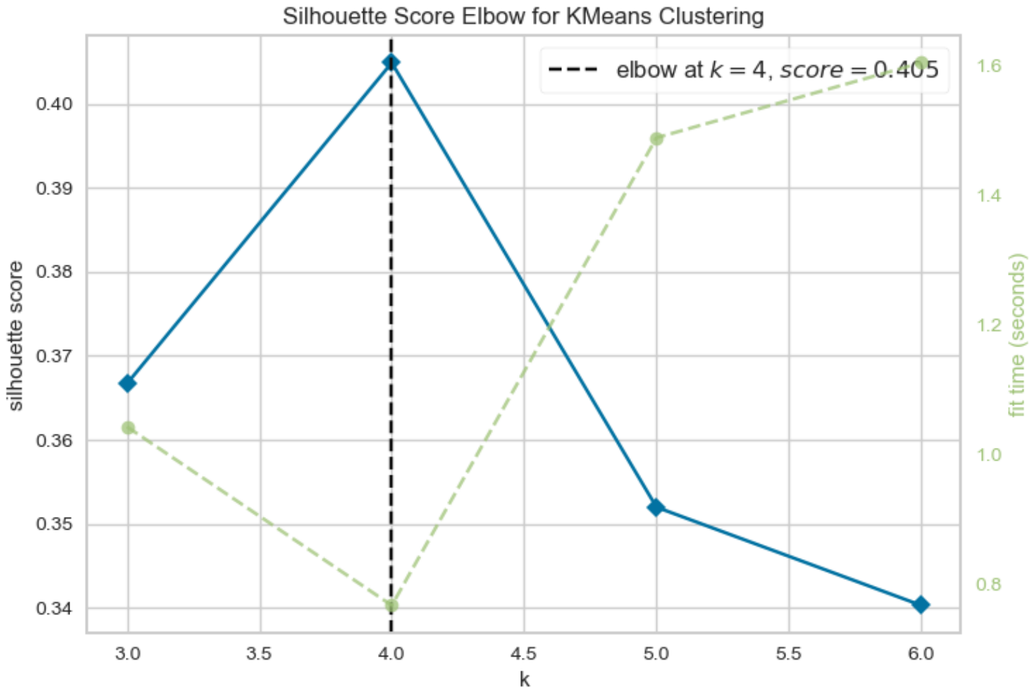
Test d'autres algorithmes

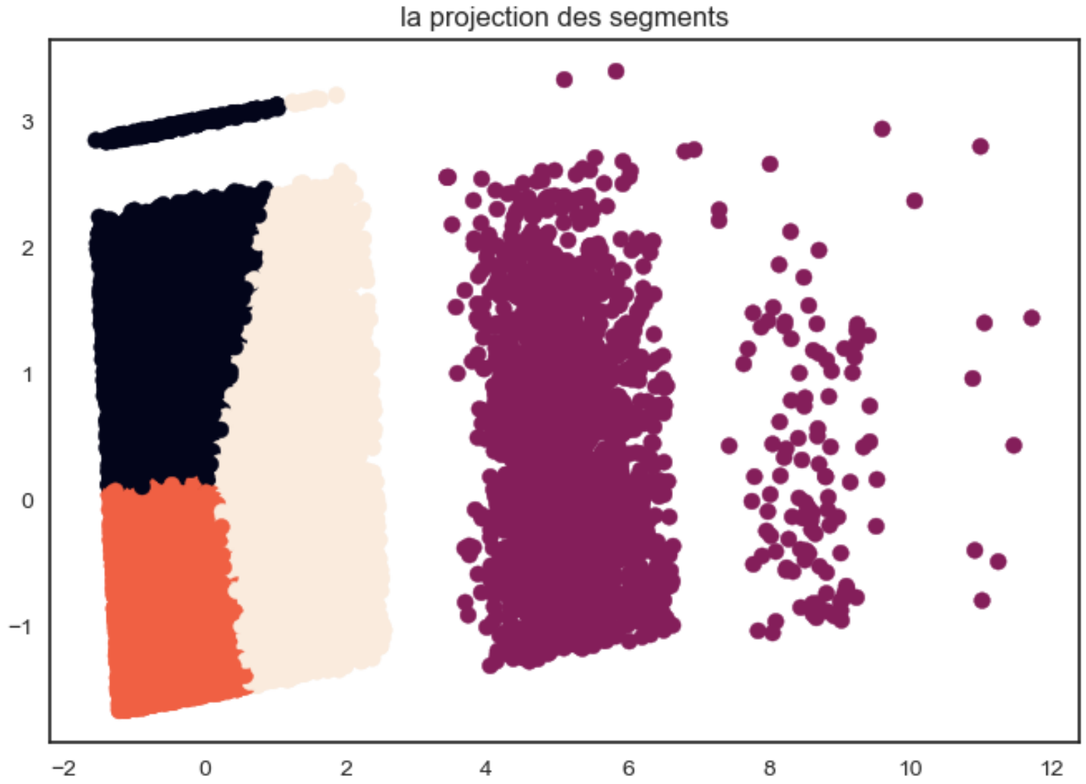
- DBSCAN
- Agglomerative Clustering

Modélisation

K-means Segmentation

RFM

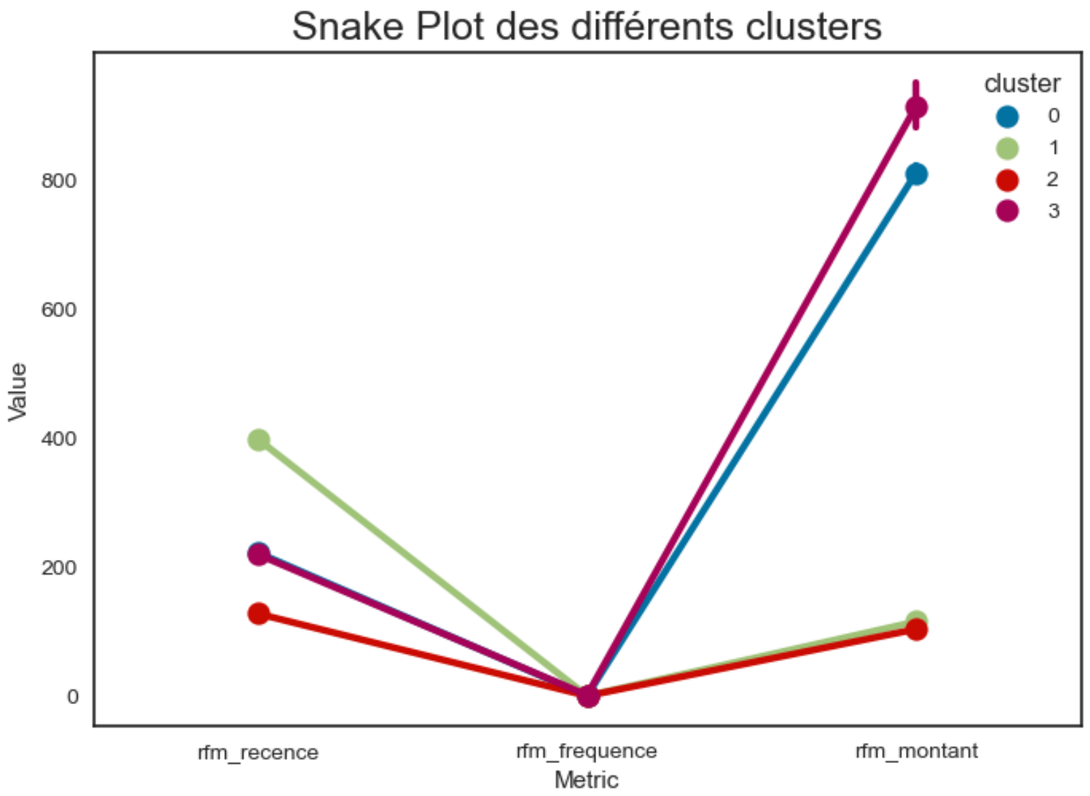


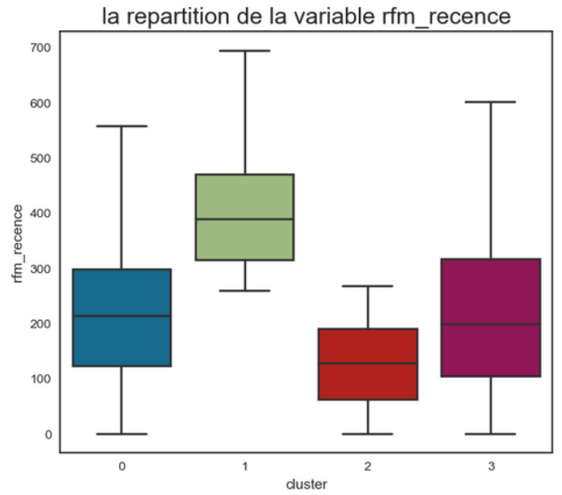


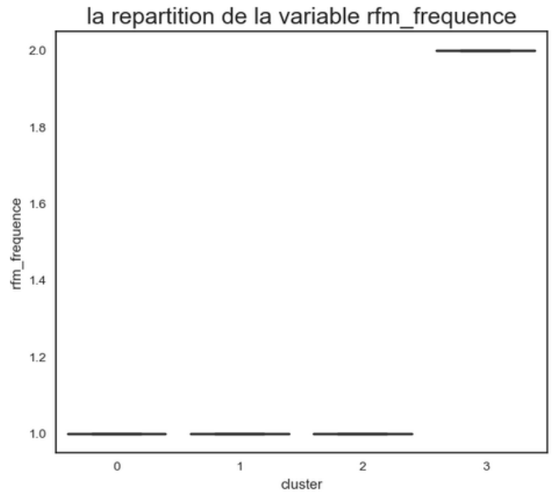
Modélisation

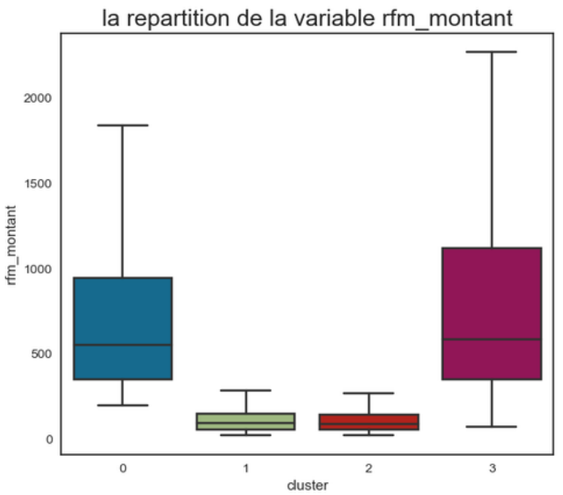
K-means Segmentation

RFM





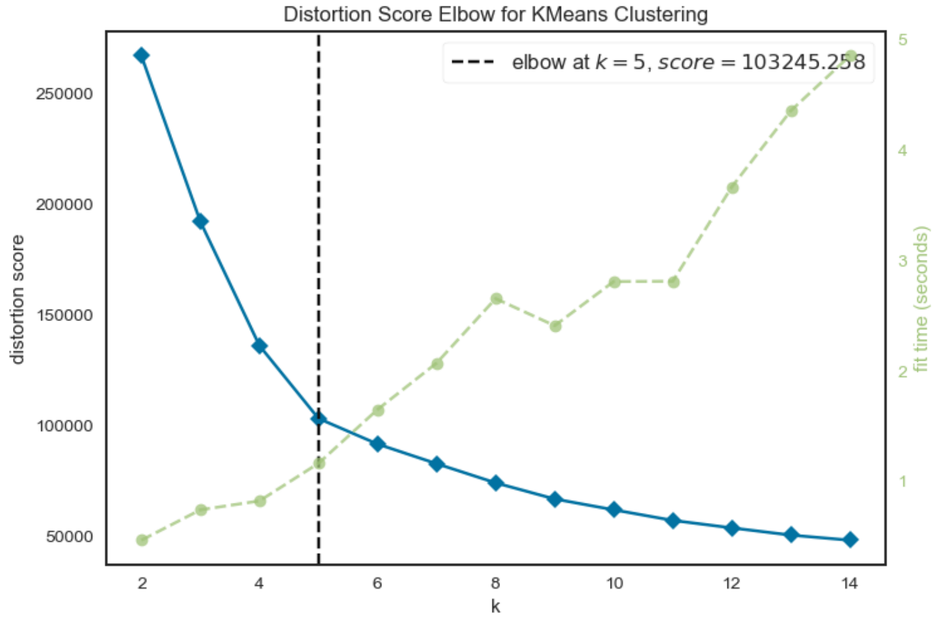


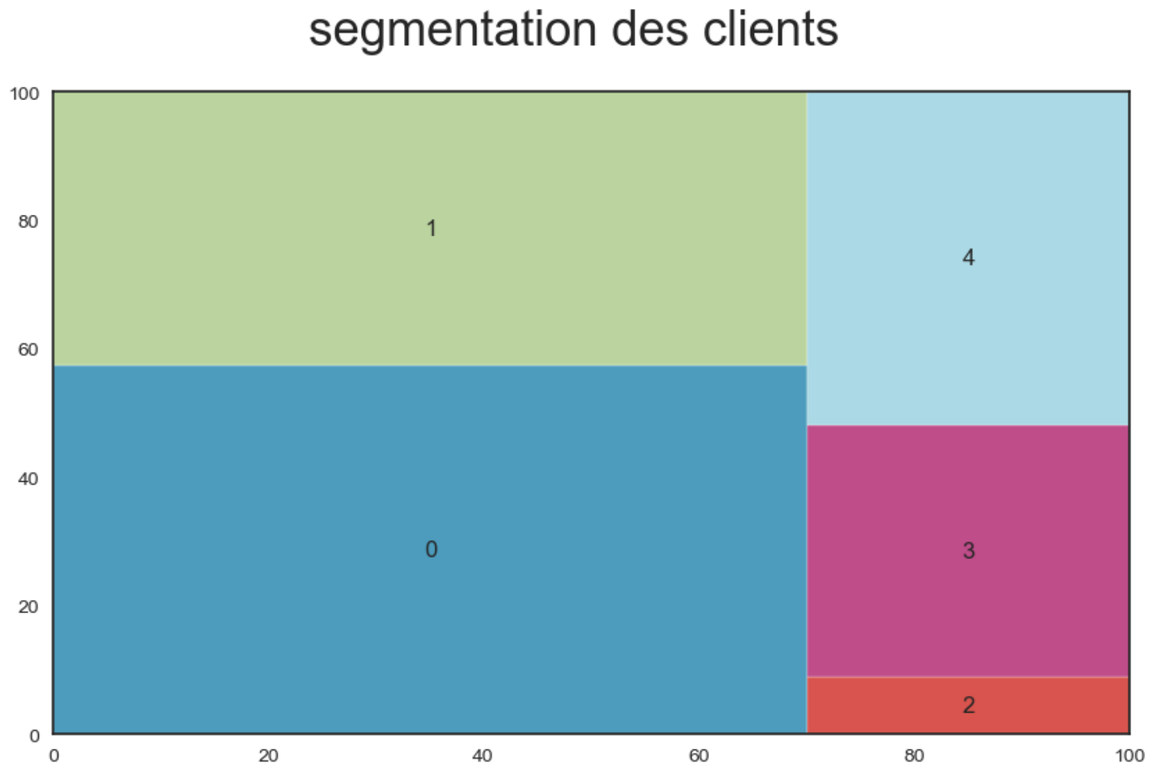


Modélisation

K-means 4 variables

RFM +'review_mean_score'

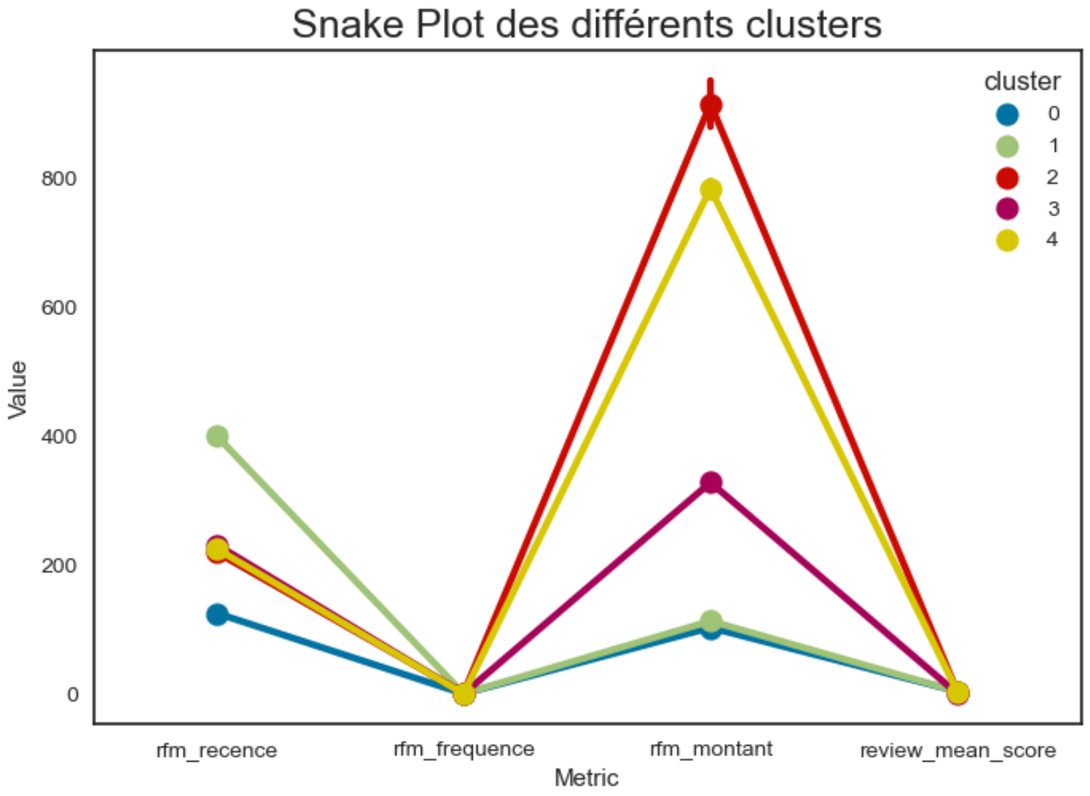


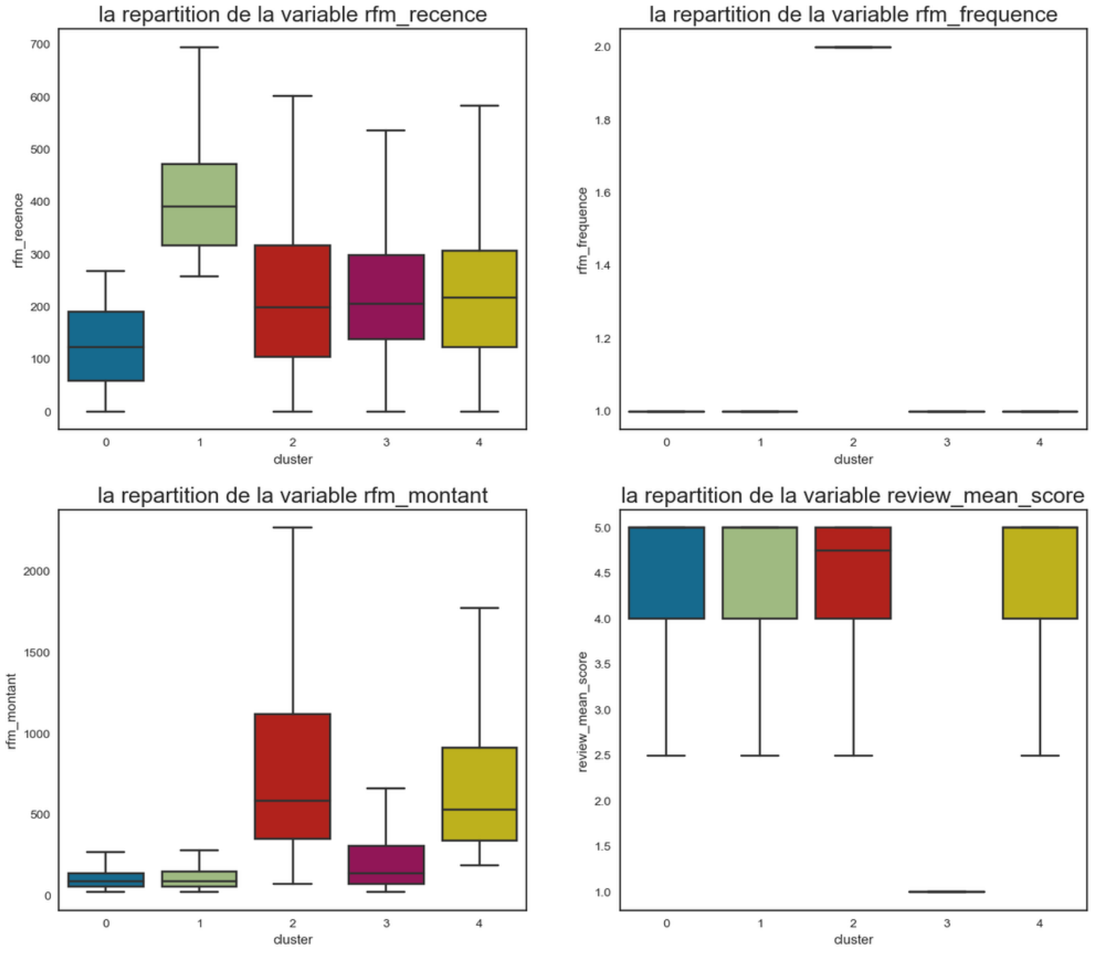


Modélisation

K-means 4 variables

RFM +'review_mean_score'

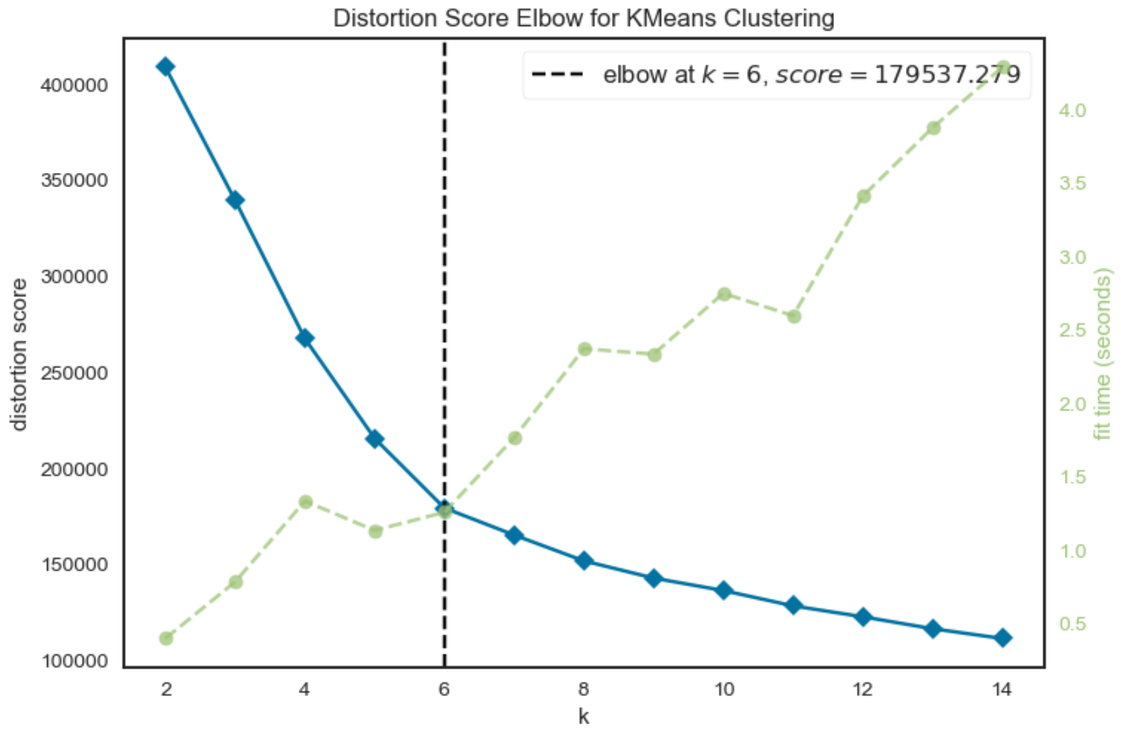


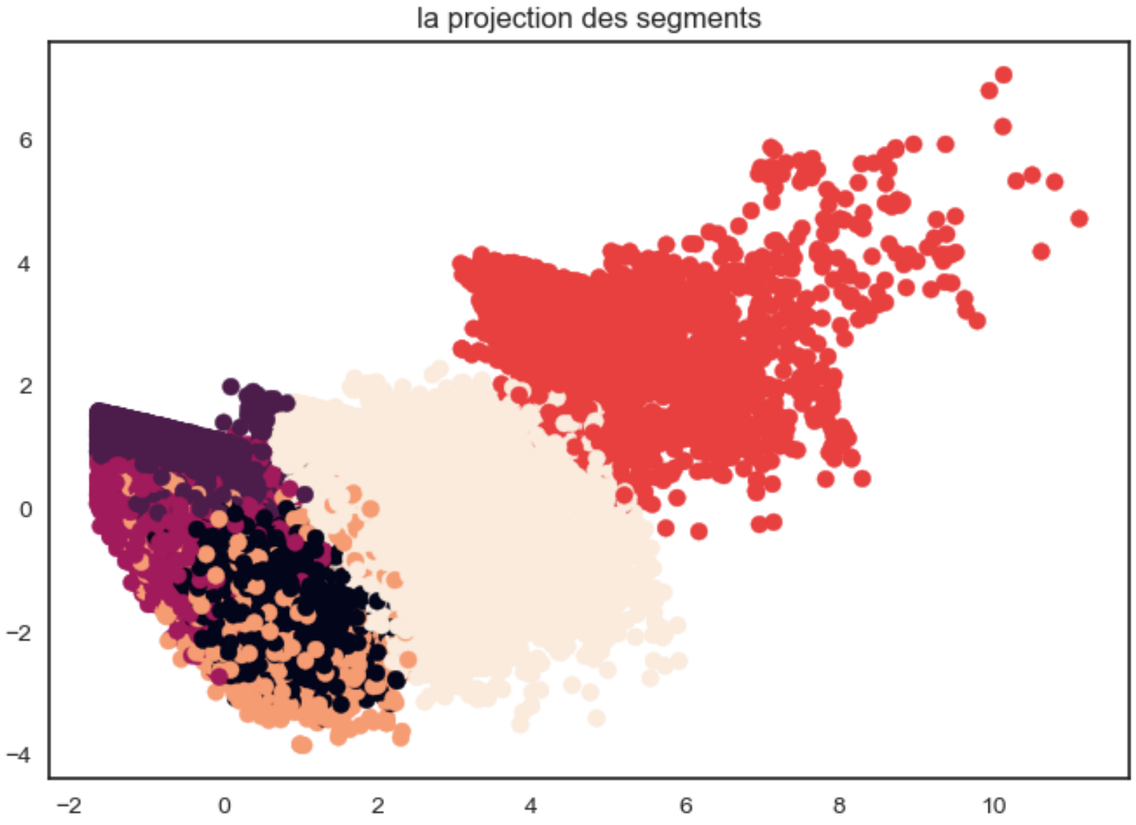


Modélisation

K-means

RFM +'review_mean_score'+'nombre_produits'+'echéances'

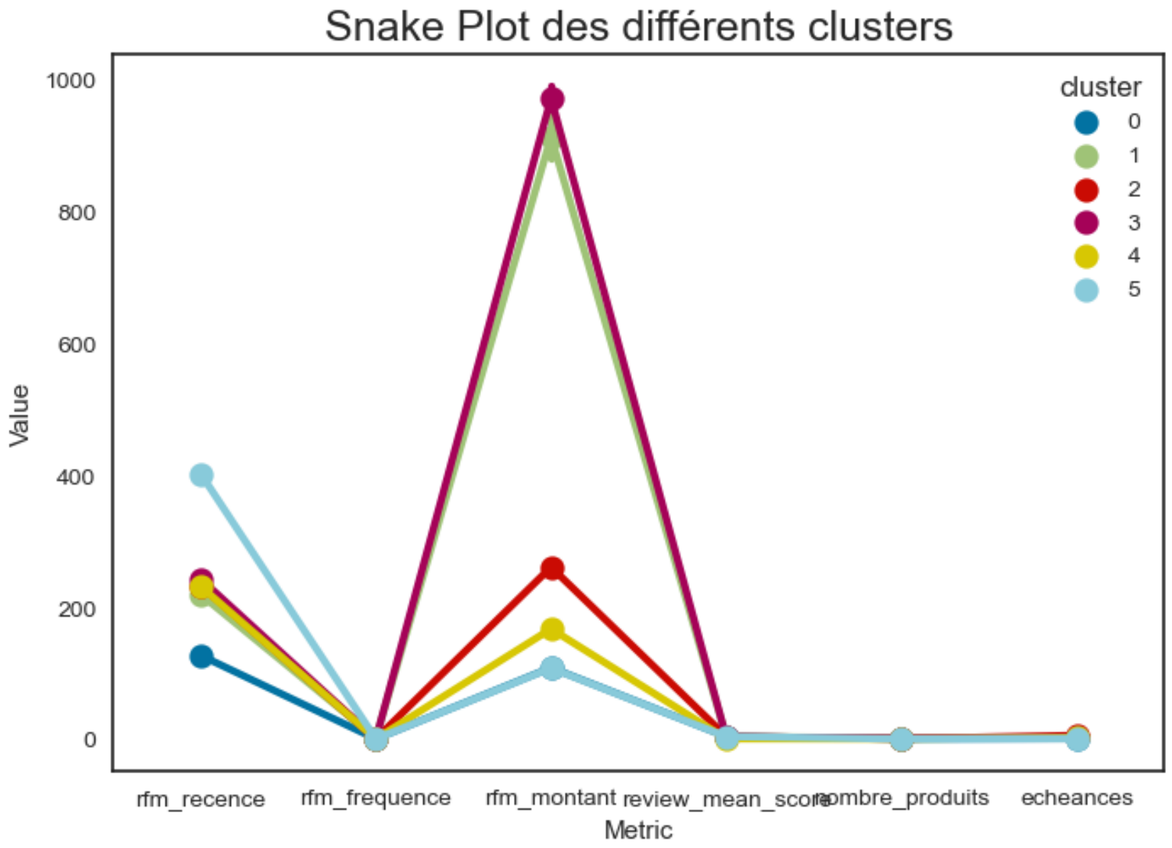


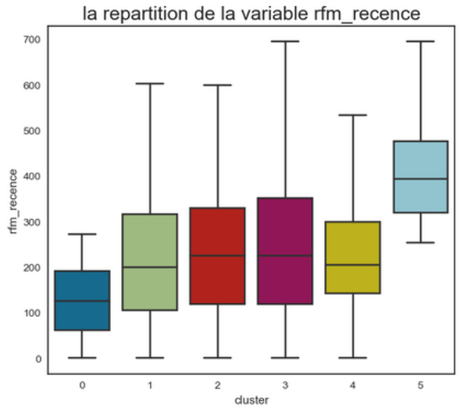


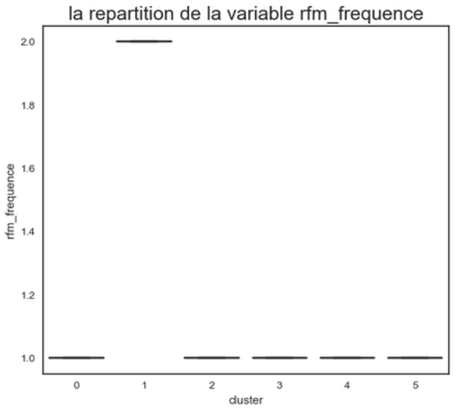
Modélisation

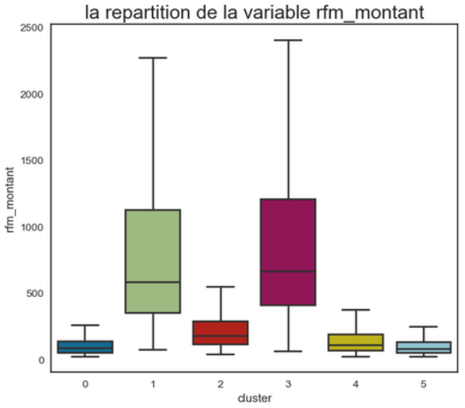
K-means

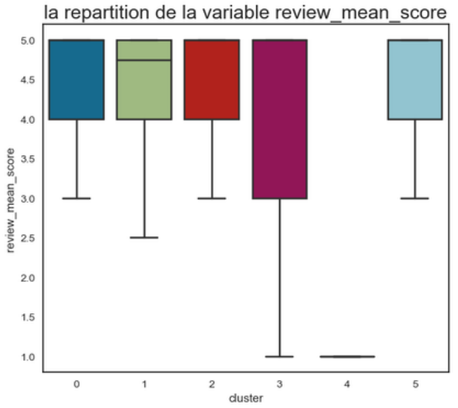
RFM+'review_mean_score'+'nombre_produits'+'echéances'

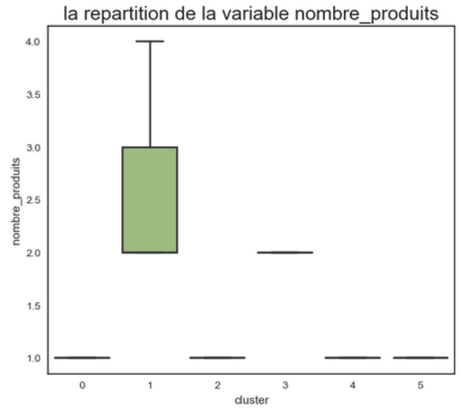


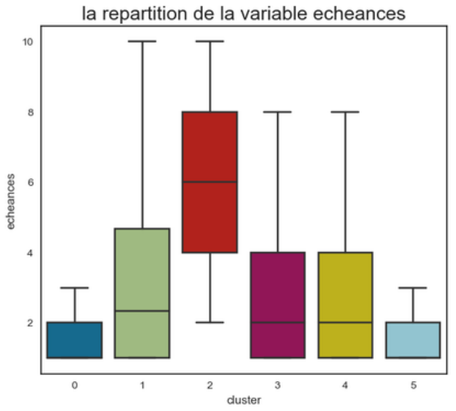










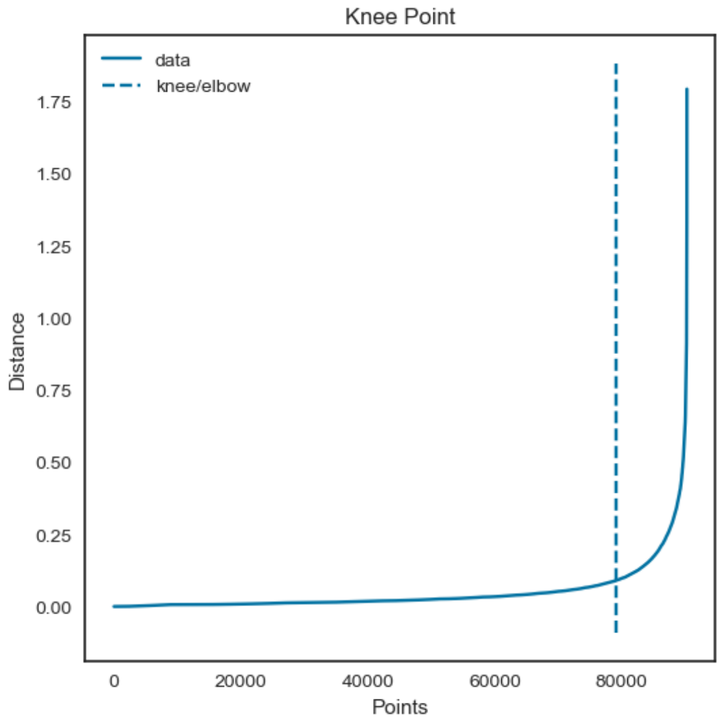


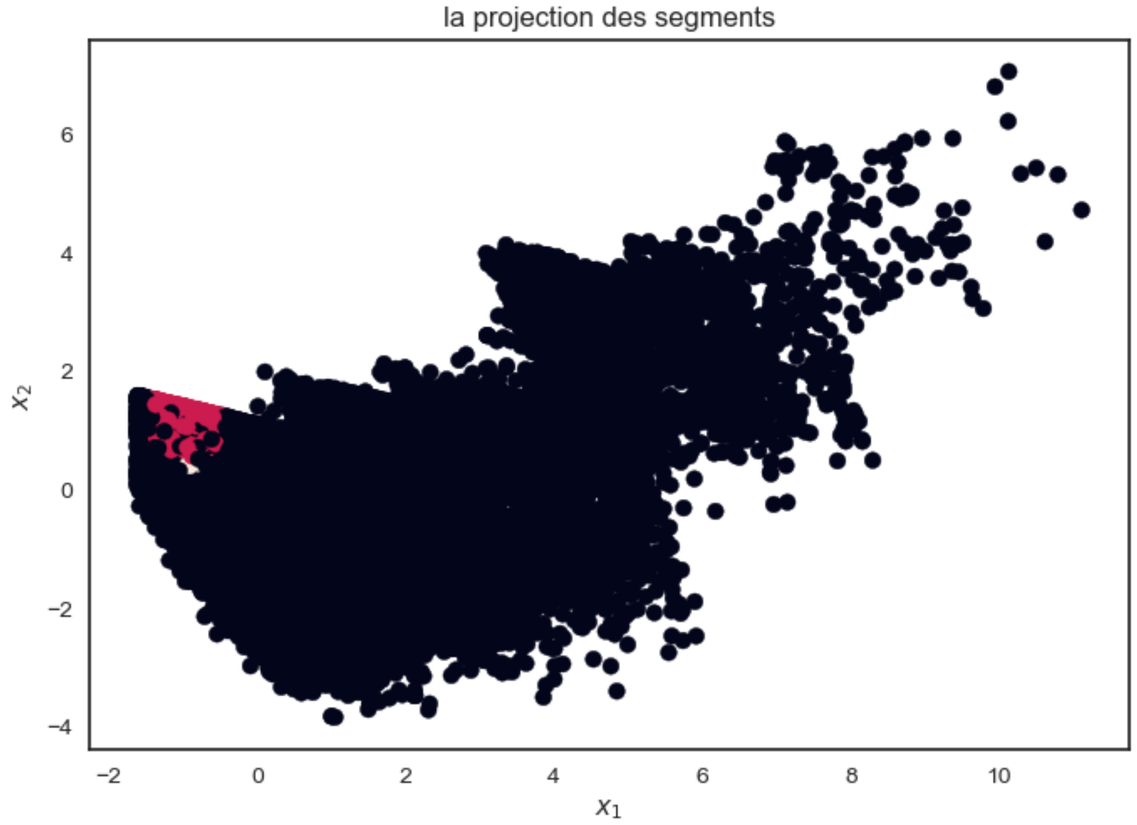
Modélisation

Test d'autres algorithmes

DBscan

RFM +'review_mean_score'+'nombre_produits'+'echéances'

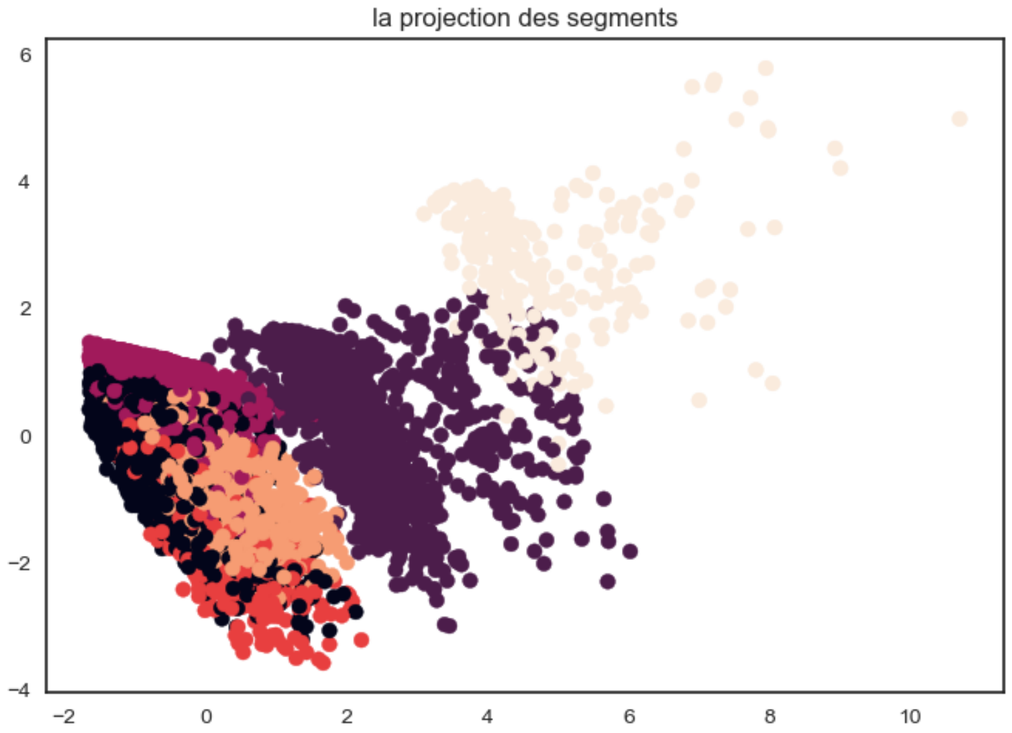


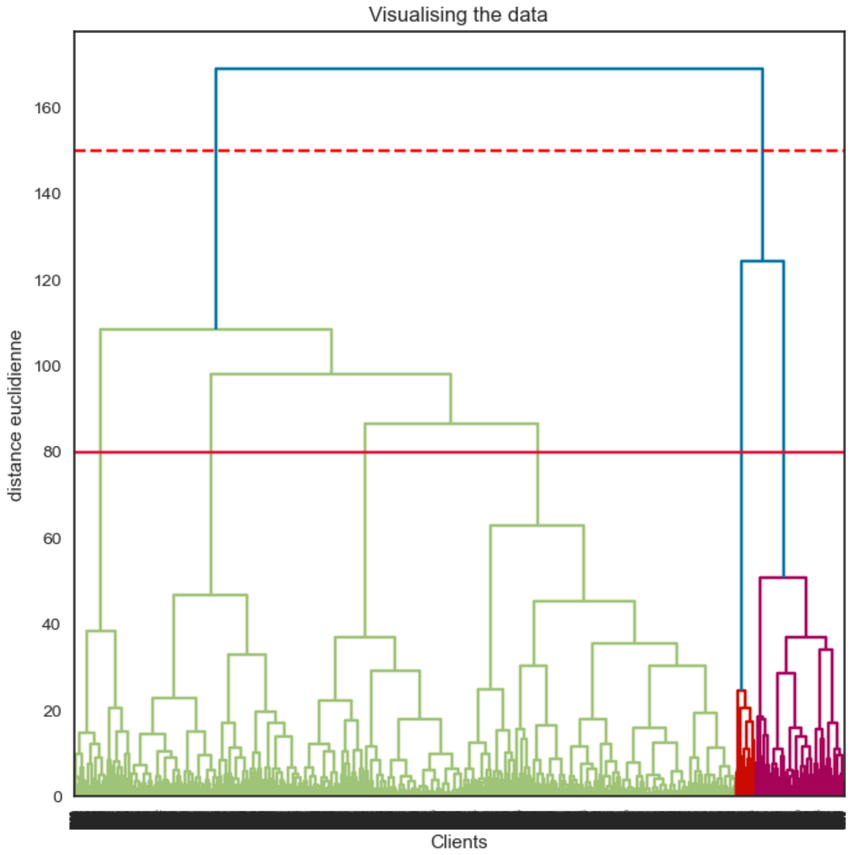


Modélisation

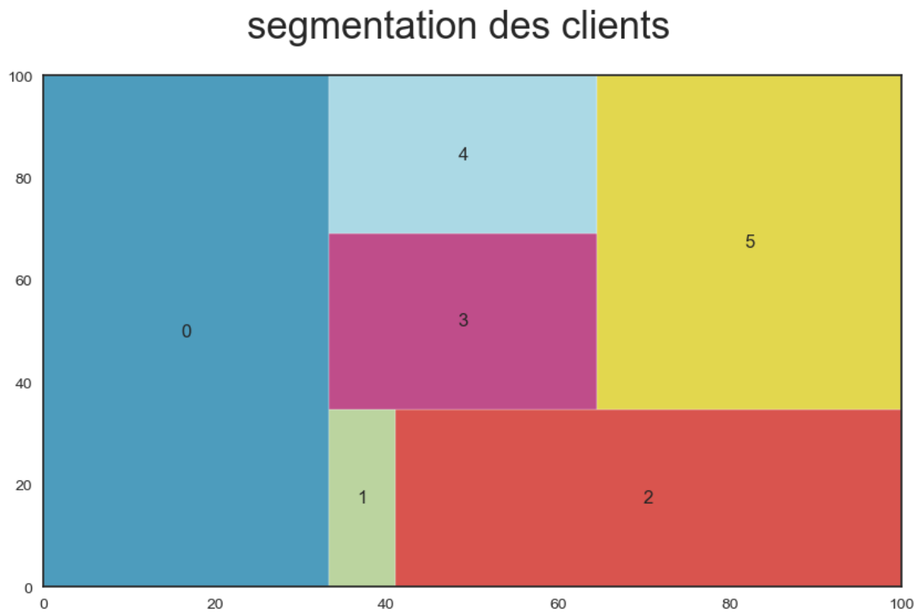
Test d'autres algorithmes
Agglomerative Clustering

RFM +'review_mean_score'+'nombre_produits'+'echéances'





Profils des clients par clusters



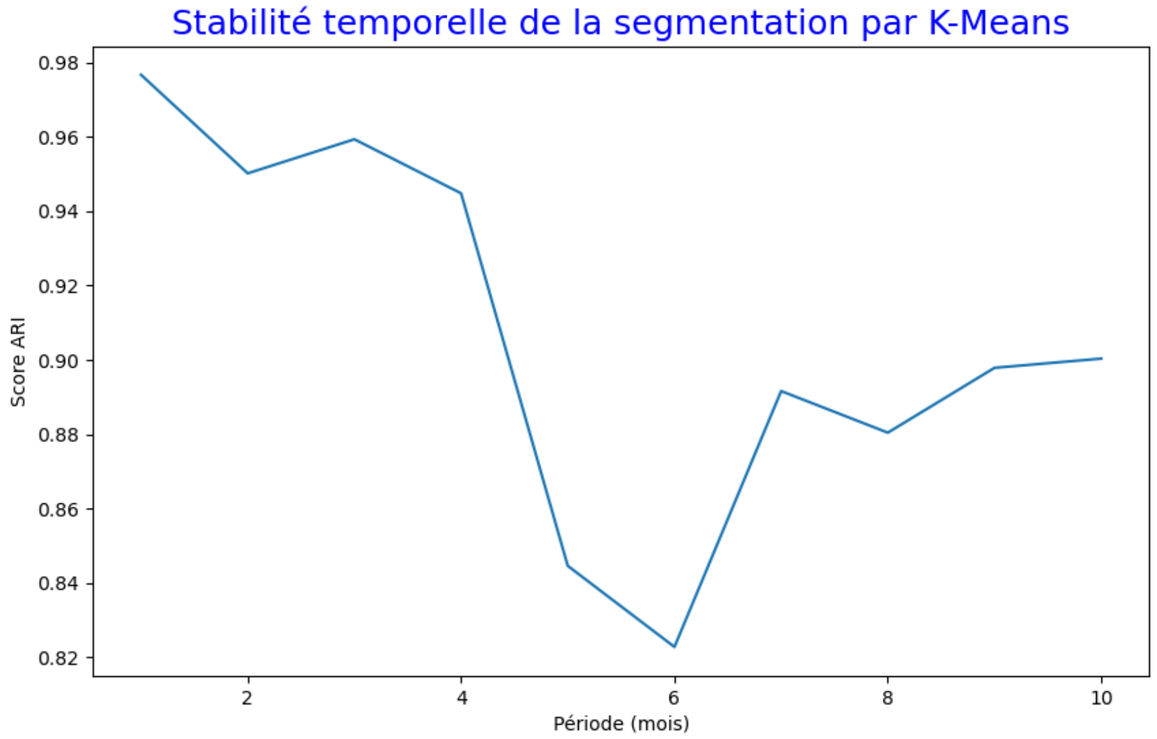
- Les recrues
- Les clients dépensiers fidèles
- Les client qui bénéficient de facilités de paiement
- Les clients dépensiers
- Les clients non satisfaits
- Les clients perdus

Contrat de maintenance

Analyse de la stabilité des clusters au cours du temps

- 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
- Evaluer la cohérence entre les clusters de départ et les partitionnements trouvé en utilisant l'indice Ajusted_Rand_Score

On remarque une forte inflexion après 4 mois sur les clients initiaux



Conclusion

Modèle retenu: **K-means**

Identification de **6 clusters**

DBscan n'est pas adapté à notre problématique, la densité des 3 000 bons clients (qui ont commandé plusieurs fois) étant faible

L’agglomerative clustering donne des résultats similaire au K_means

Prévoir la maintenance du programme de segmentation tous les 4 mois dans un premier temps puis re-tester cette stabilité temporelle au fil du temps afin de l'affiner.

Merci !