SEGMENTATION DES CLIENTS D'UN SITE E-COMMERCE OLIST

Zouheir HMIDI

SOMMAIRE

1. Problématique
2. Nettoyage et analyse des données
3. Segmentation
4. Maintenance du modèle

1- PROBLÉMATIQUE

Olist, le site E-Commerce, souhaite mettre en place une segmentation de sa clientèle qui soit utilisable au quotidien par leur équipe marketing lors de leurs campagnes de communication. L'objectif principal est de mieux comprendre les différents types d'utilisateurs en se basant sur leur comportement et leurs données personnelles anonymisées.

Pour cela, nous allons élaborer une description opérationnelle de la segmentation ainsi que sa logique sous-jacente, afin de permettre une utilisation optimale de cette information par l'équipe marketing. Nous utiliserons des méthodes non supervisées pour regrouper les clients ayant des profils similaires, ce qui nous permettra de créer des segments homogènes.

En plus de la segmentation, nous fournirons une analyse de la stabilité des segments dans le temps. Cette analyse sera essentielle pour établir un contrat de maintenance solide. En surveillant l'évolution des segments au fil du temps, Olist pourra adapter ses stratégies de communication et de marketing pour mieux cibler les besoins changeants de sa clientèle.

En résumé, notre approche consistera à créer des groupes de clients ayant des caractéristiques similaires grâce à des méthodes non supervisées, tout en fournissant une analyse détaillée de la stabilité de ces segments au cours du temps pour assurer une utilisation optimale et pertinente de la segmentation par l'équipe marketing d'Olist.

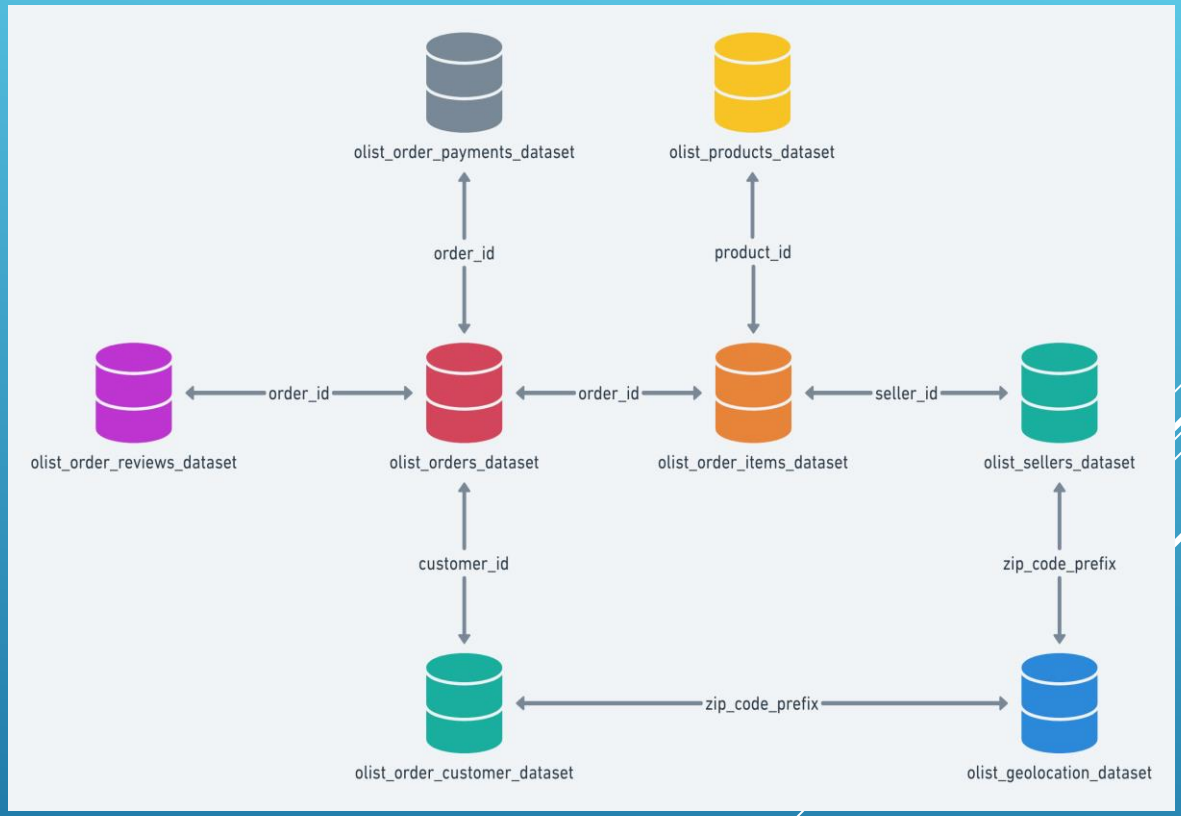
Haut du formulaire

2- ANALYSE EXPLORATOIRE DES DONNÉES

Analyse exploratoire des données - suite

L'ensemble de données contient des informations sur 100 000 commandes de 2016 à 2018

	Lignes	Colonnes
Customers	99441	5
Geolocation	1000163	7
Items	112650	5
Payments	103886	5
reviews	99224	7
Orders	96478	8
Products	32951	9
sellers	3095	4

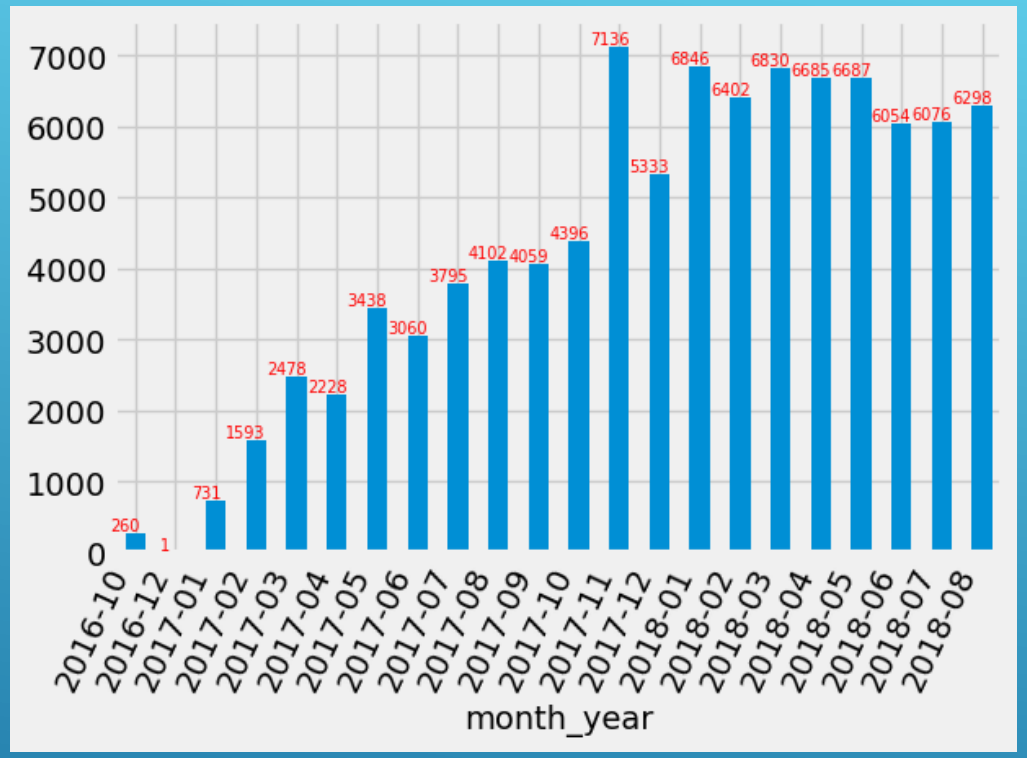


Analyse exploratoire des données - suite

Suite à la fusion des dataframes à l'aide de la méthode **merge**, ainsi qu'à l'élimination des doublons et des colonnes non pertinentes pour notre analyse, notre jeu de données est désormais composé de :

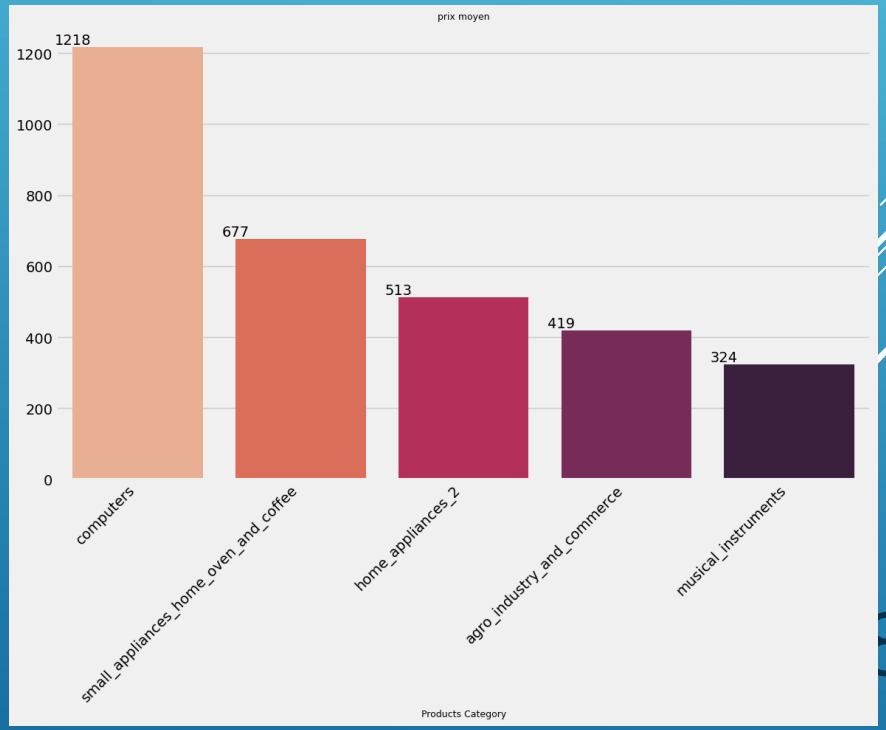
lignes	colonnes
94 488	13

Analyse exploratoire des données - suite

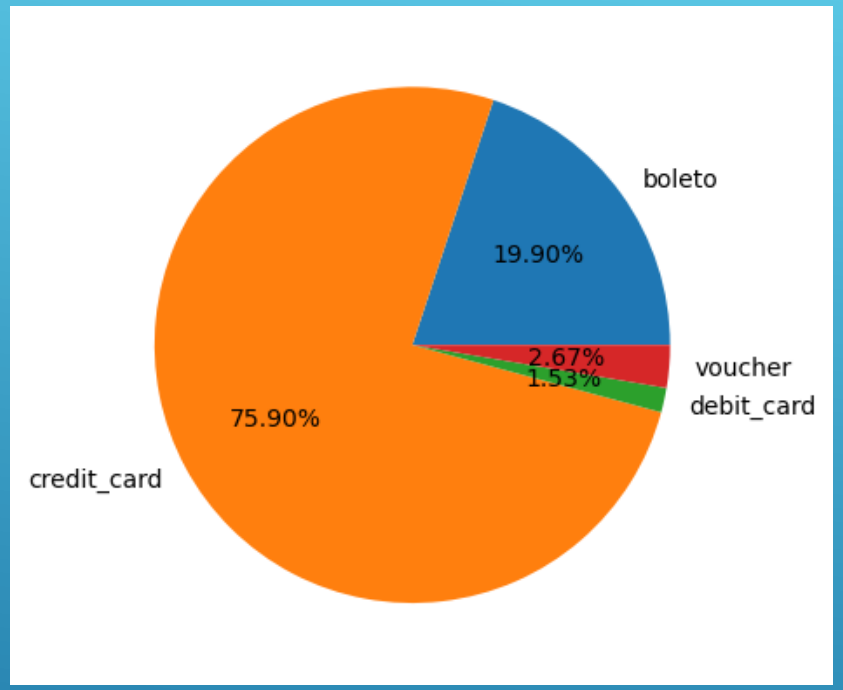


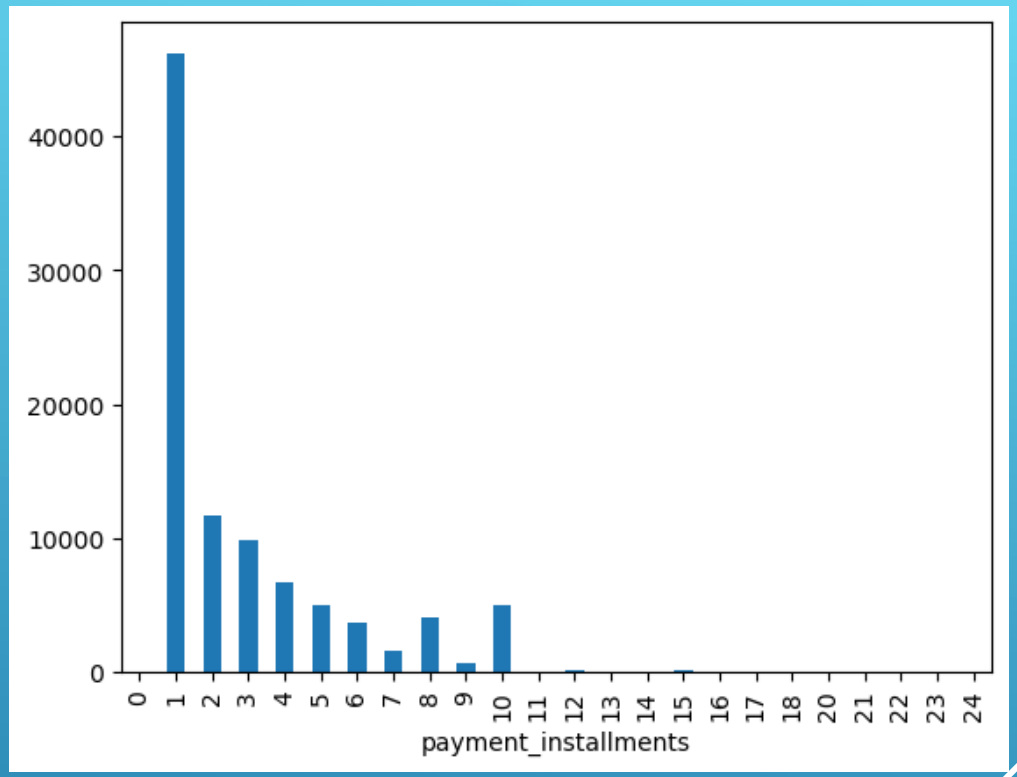
Nous avons 22 mois non consécutifs, car le mois de Novembre pour l'année 2016 est manquant. Cependant, nous avons enregistré le pic maximum de commandes en Novembre 2017.

La figure montre quelle catégorie de produits est principalement achetée par le client. On y trouve des ordinateurs , appareils électroménagers, L'agroalimentaire et les instruments de musique.

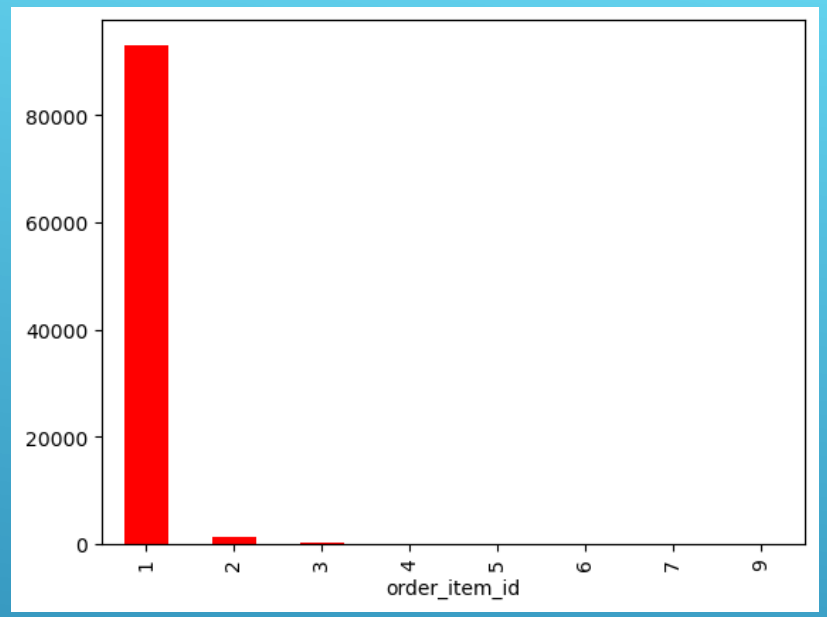


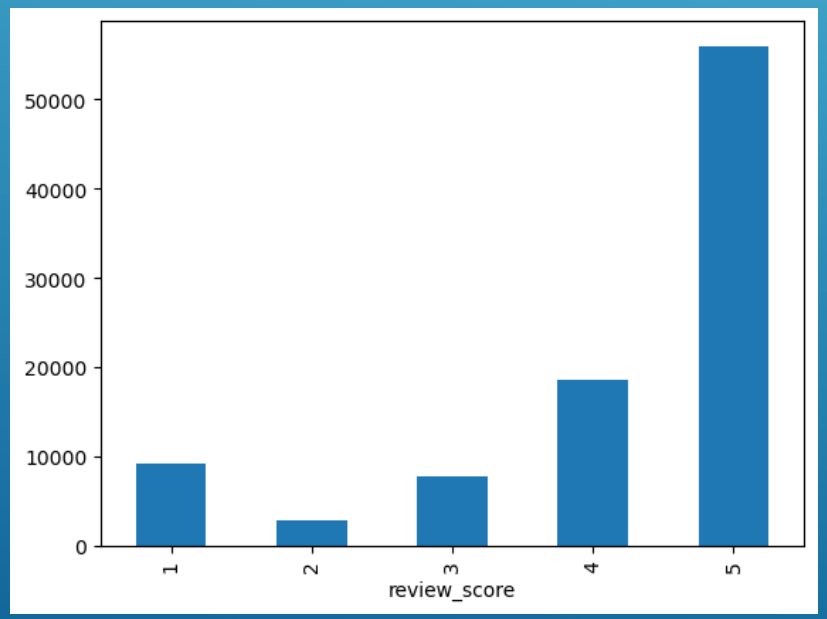
Analyse exploratoire des données - suite





Près de la totalité des clients effectuent leurs paiements par carte de crédit en une seule fois.





La grande majorité des clients ne commandent que 1 seule fois.

- 80% des notes sont 4 et 5
- 20% des notes sont en dessous de 4

3- Segmentation

3-1 RFM : Méthode statistique de marketing

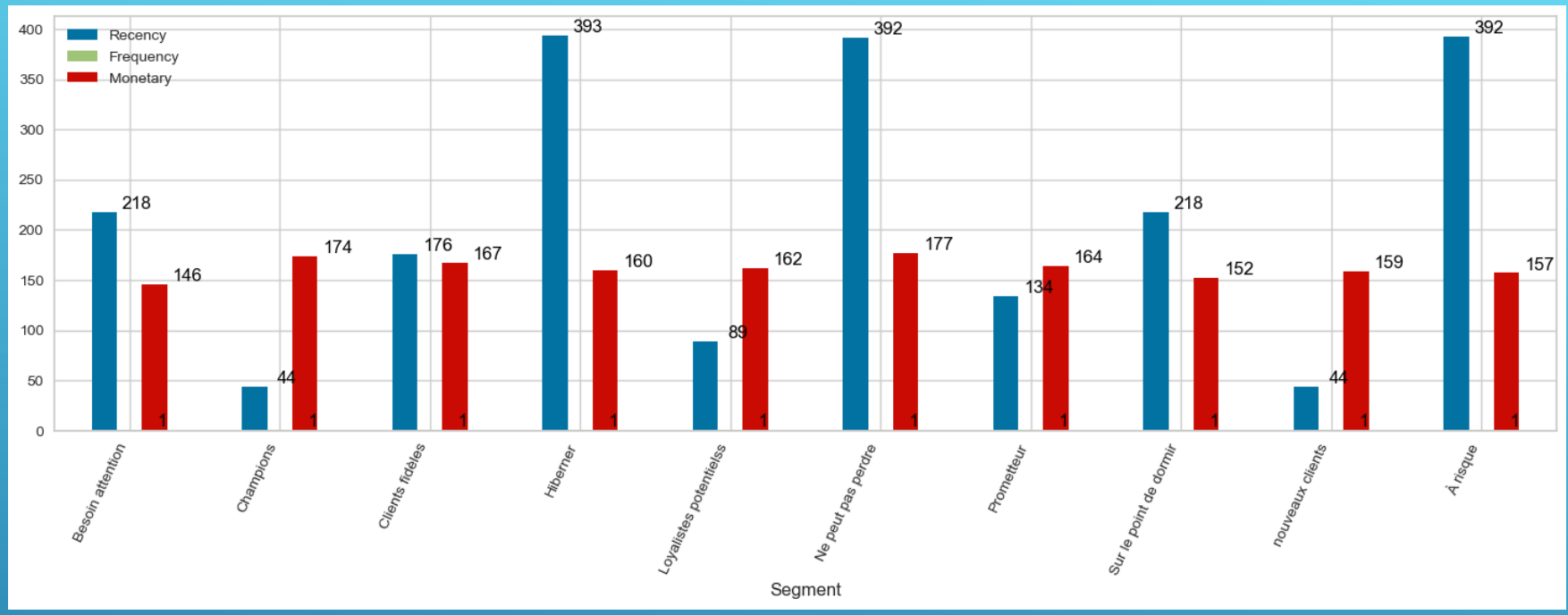
Segmentation - suite

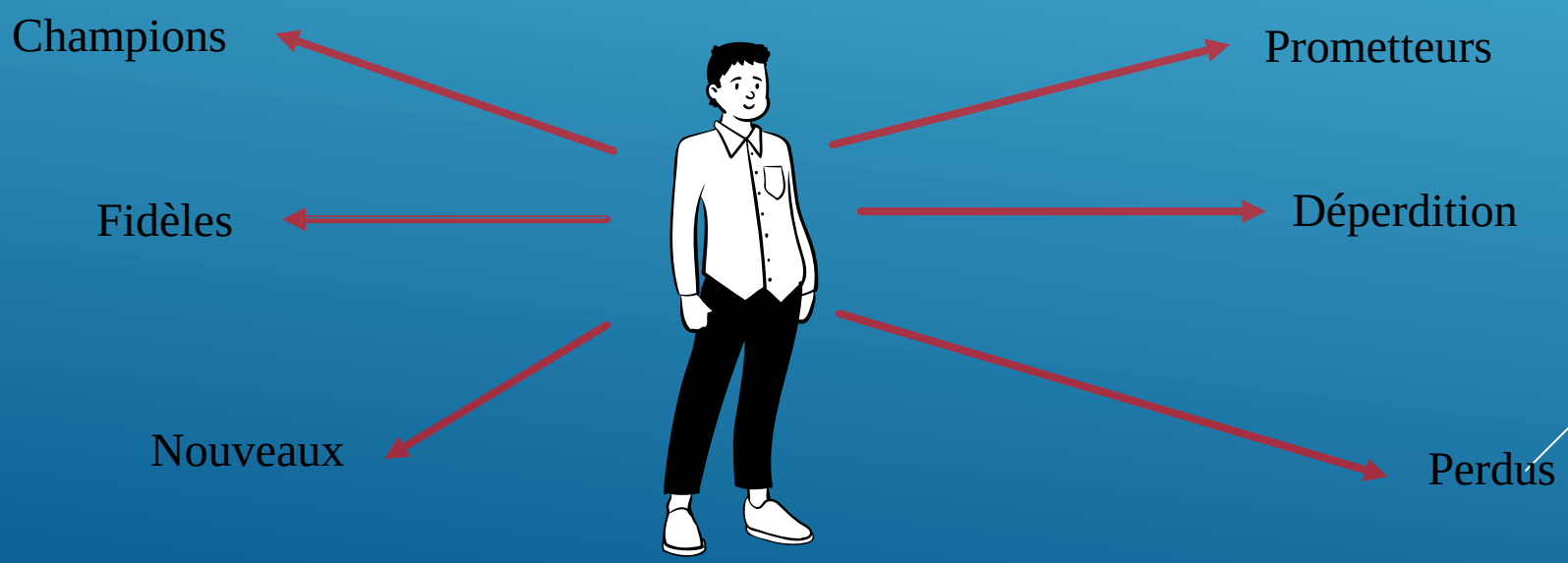
Les segments de clientèle RFM sont un outil d'analyse marketing qui permet de segmenter les clients en fonction de leur comportement d'achat.

RFM est l'acronyme de trois indicateurs clés : Récence (Recency), Fréquence (Frequency) et Montant (Monetary).

1. Récence (Recency) : Cet indicateur mesure la période de temps écoulée depuis le dernier achat d'un client.
2. Fréquence (Frequency) : Cet indicateur mesure le nombre total d'achats effectués par un client sur une période donnée.
3. Montant (Monetary) : Cet indicateur mesure le montant total dépensé par un client sur une période donnée.

Segmentation - suite





Segmentation - suite

- Meilleurs clients (Champions) : Ce sont les clients qui ont une récence élevée, une fréquence d'achat élevée et qui ont dépensé beaucoup d'argent. Ce segment regroupe les clients les plus précieux et les plus actifs pour l'entreprise.

- Clients en voie de déperdition (**Hibernating**) : Ce sont les clients qui ont une récence élevée, mais une fréquence d'achat et un montant dépensé faibles. Ils ont été actifs par le passé mais ne sont plus aussi engagés .

- Nouveaux clients (**New**) : Ce sont les clients qui ont récemment effectué leur premier achat.
  Ils sont considérés comme prometteurs car ils ont montré un certain intérêt pour la marque.

- Clients fidèles (**Loyal**) : Ce sont les clients qui ont une fréquence d'achat élevée et une récence moyenne. Ils peuvent ne pas dépenser autant que les meilleurs clients, mais ils sont loyaux et réguliers dans leurs achats.

- Clients perdus (**Lost**) : Ce sont les clients qui ont une récence, une fréquence d'achat et un montant dépensé faibles. Ils sont inactifs depuis longtemps et pourraient être considérés comme perdus.

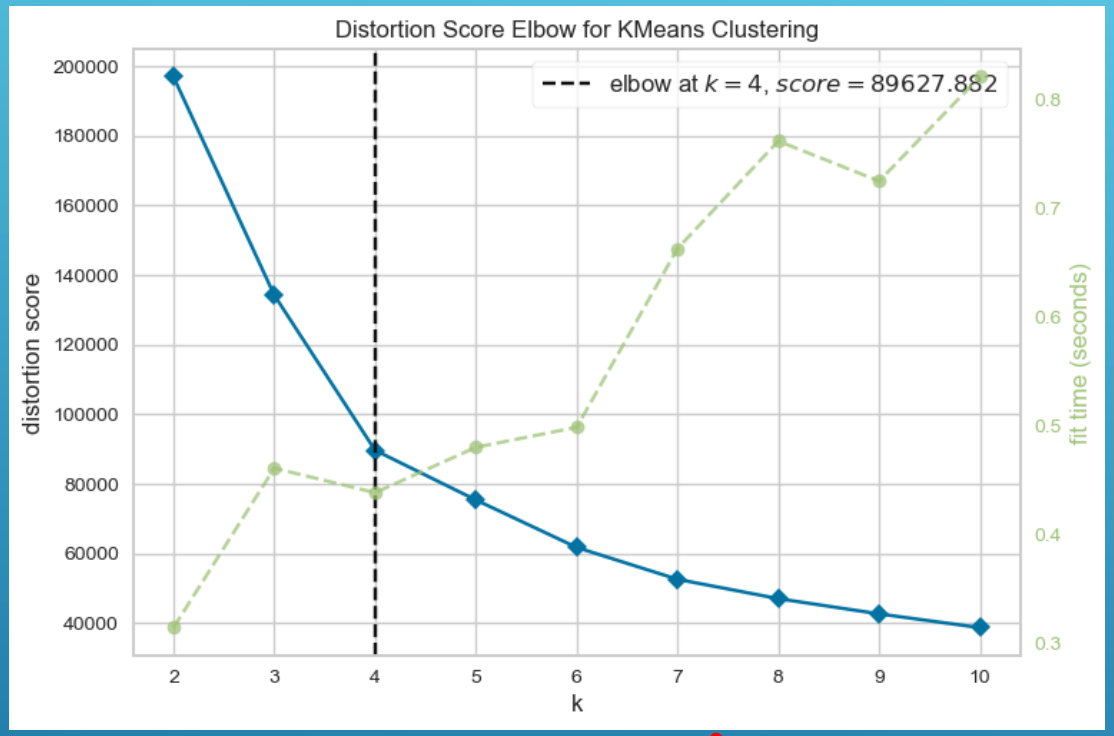
- Clients prometteurs (**Promising**) : Ce sont les clients qui ont une fréquence d'achat élevée, mais une récence et un montant dépensé moyens. Ils sont actifs, mais il y a encore un potentiel pour augmenter leur engagement et leur valeur.

3-2 Méthode de clustering

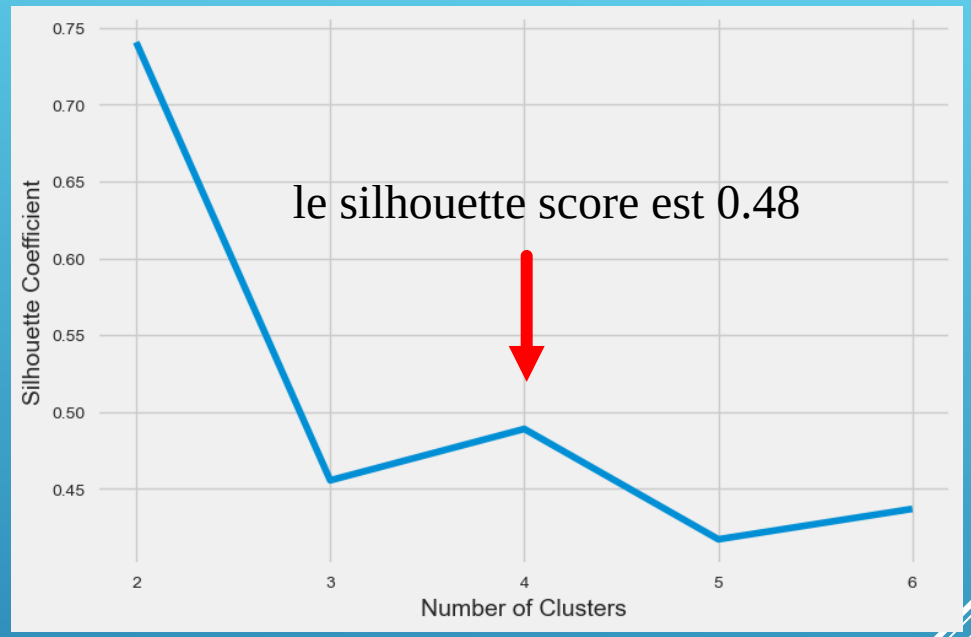
Segmentation - suite

A - KMeans

La méthode du coude

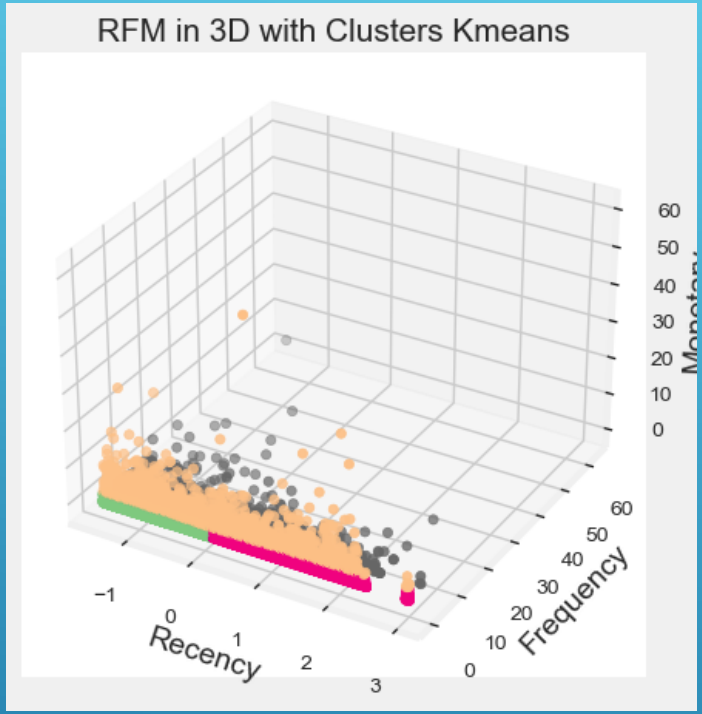


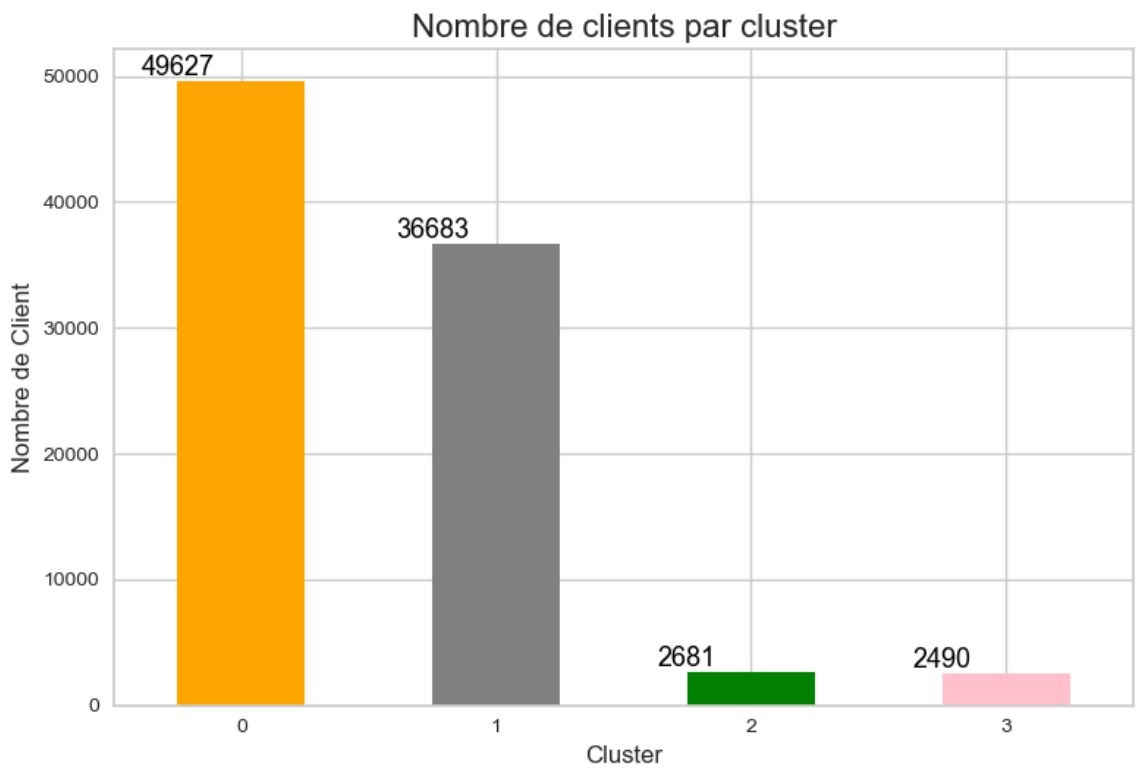
silhouette score



Le nombre de clusters à calculer est de 4

Segmentation - suite





Dans les deux graphiques, une séparation relativement nette des clusters est visible.

Il est évident que le cluster 0 et 1 sont plus grands en taille, tandis que les clusters 2 et 3 comptent moins de clients.

Segmentation - suite

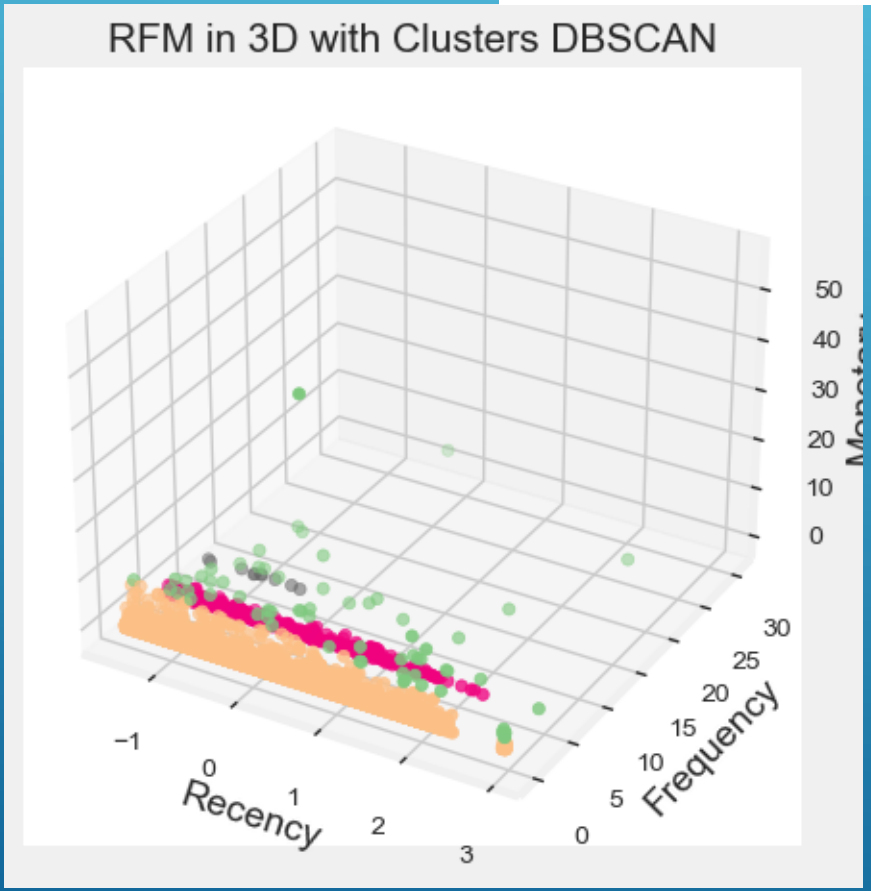
Évaluation de la qualité du clustering :

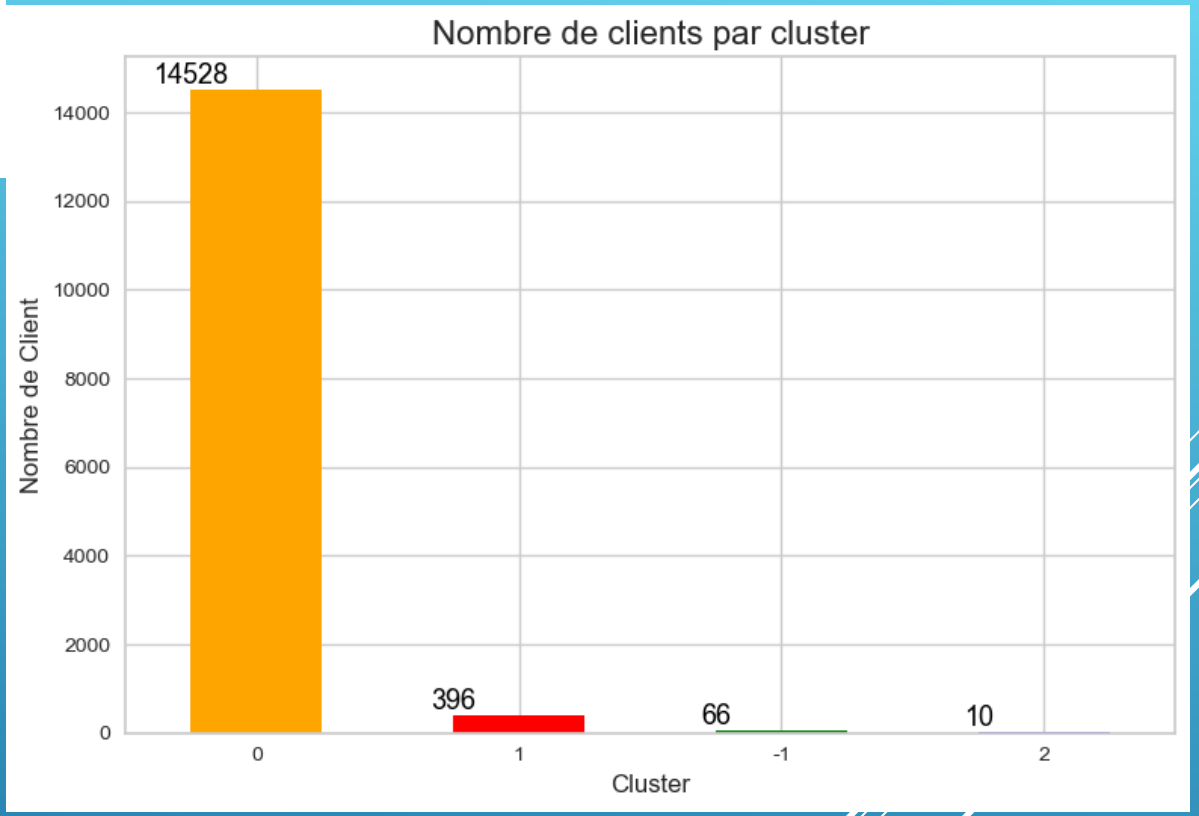
Metrics	Score
Silhouette Score	0.49
Calinski harabasz Score	62876.02
Davies Bouldin Score	0.68

Sur la base du score d'évaluation ci-dessus, la qualité du clustering à l'aide de KMeans avec 4 clusters est bonne.

B - DBscan

```
eps=0.6
min_samples=5
Clusters present: [-1  0  1  2]
Cluster sizes: [   66 14528   396    10]
Silhouette Score: 0.7296447606842419
```





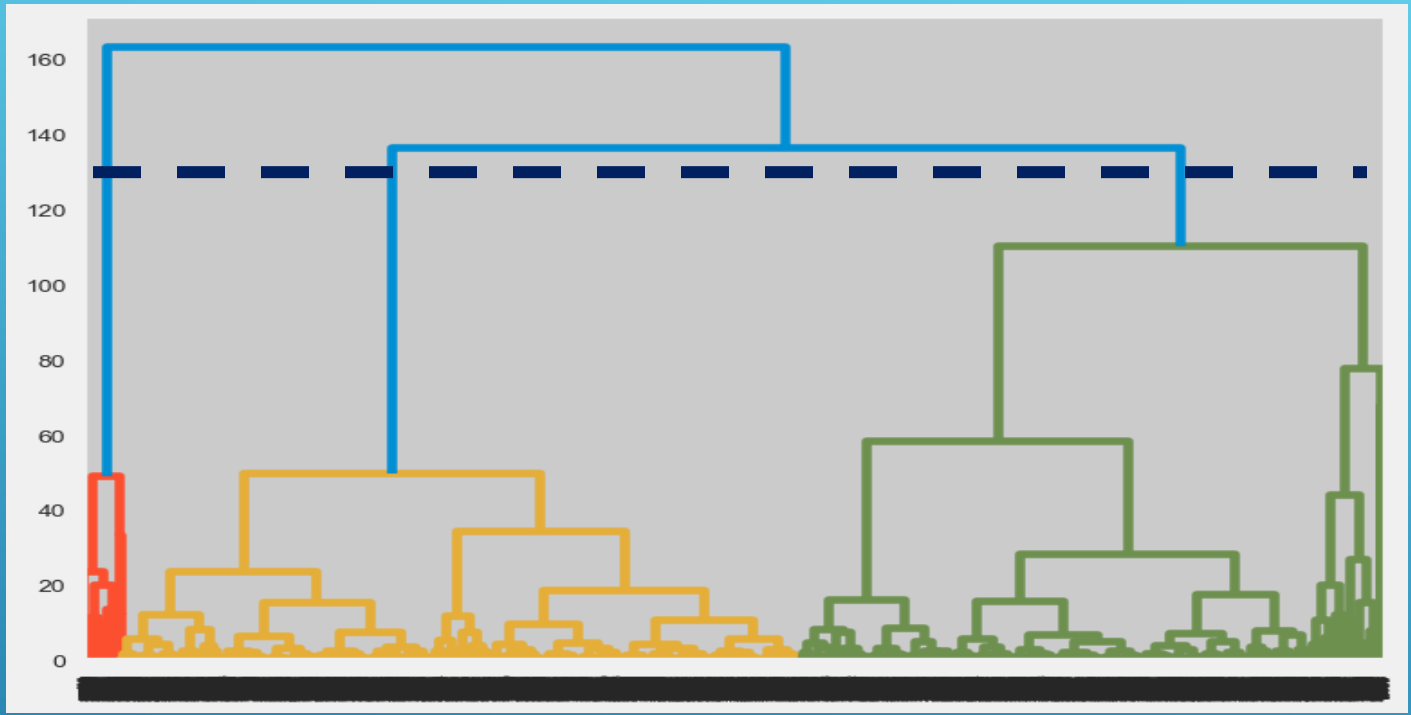
Il y a trois clusters formés. Le cluster 0 contient le plus grand nombre de points de données par rapport au cluster 1 et 2. Cependant, il y a des valeurs aberrantes détectées, car certains points sont situés beaucoup plus loin des autres points de données.

Segmentation - suite

Metrics	Score
Silhouette Score	0.73
Calinski harabasz Score	2549.46
Davies Bouldin Score	1.15

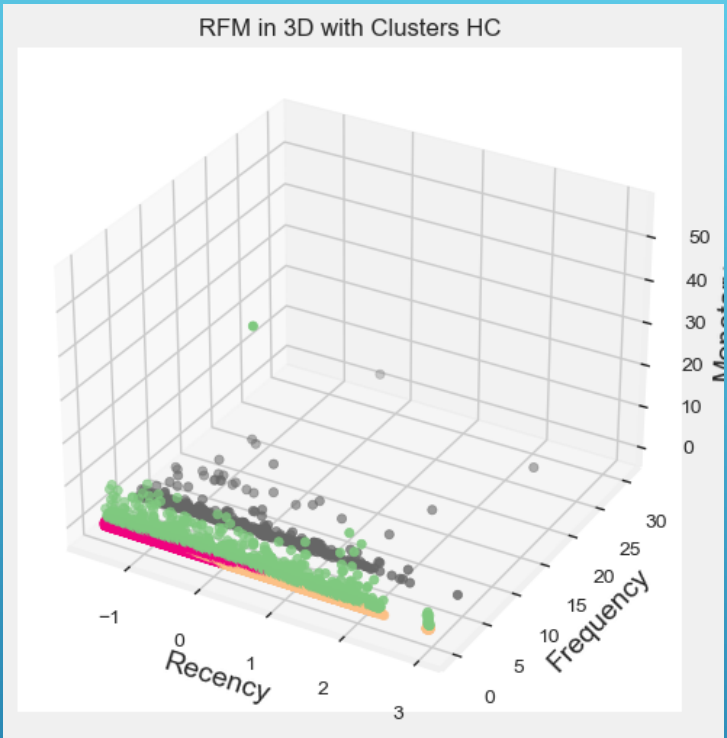
La qualité de clustering à l'aide de DBSCAN avec 3 clusters et des valeurs aberrantes est équitable selon la note d'évaluation ci-dessus. Le score de silhouette est meilleur que KMeans car il y a un grand cluster et 2 petits cluster formé, bien que l' indice Davies-Bouldin soit supérieur à KMeans , ce qui indique un clustering équitable. Cependant, l' indice de Calinski-Harabasz obtenu est bien inférieur à KMeans

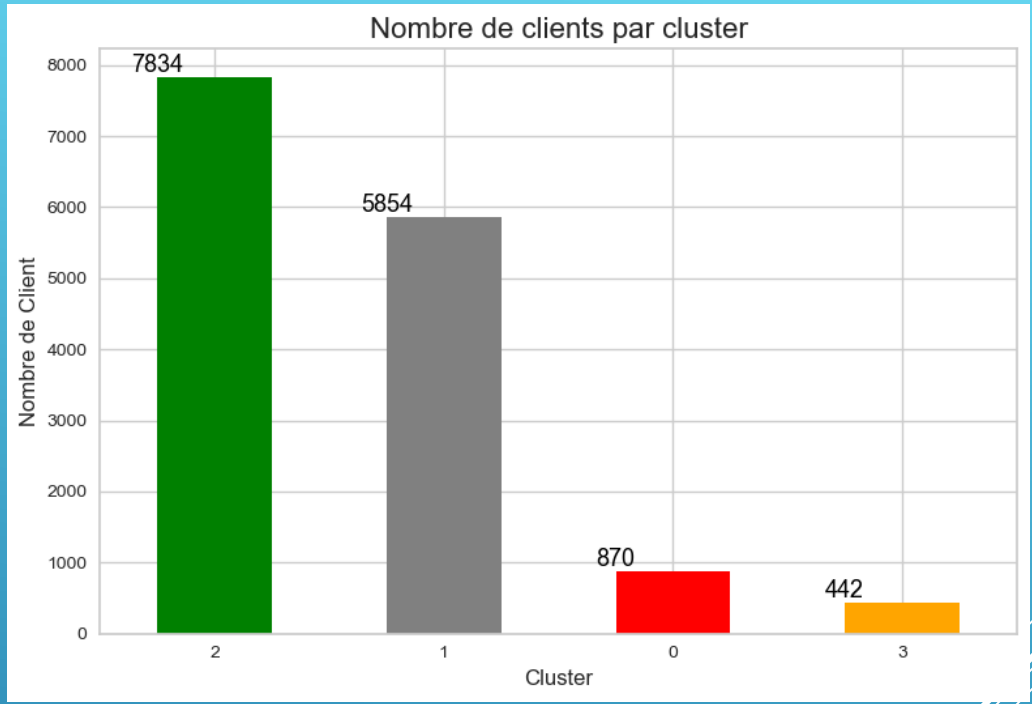
C- Hierarchical Clustering



nous déterminons la plus grande distance verticale qui ne croise aucun des autres clusters. le nombre optimal de clusters est égal au nombre de lignes verticales passant par la ligne horizontale. Par exemple, dans le cas ci-dessus, le meilleur choix pour nous. des clusters seront 4.

Segmentation - suite





Dans les deux graphiques, on observe une séparation assez distincte entre les clusters. Il est clair que le cluster 2 et le cluster 1 sont plus grands en taille, tandis que les clusters 0 et 3 comptent moins de clients.

Segmentation - suite

Metrics	Score
Silhouette Score	0.47
Calinski harabasz Score	8745.9
Davies Bouldin Score	0.82

Sur la base des résultats de l'évaluation de la qualité du clustering à l'aide du clustering hiérarchique, on peut voir que les résultats obtenus sont différents de K-Means . Un indice Davies-Bouldin élevé indique qualité de regroupement acceptable.
L' indice Calinski-Harabasz obtenu est légèrement inférieur par rapport à K-Means, mais supérieur par rapport à DBSCAN.

Segmentation - suite

	Model_Name	silhouette score	calinski_harabasz score	davies_bouldin score
0	KMeans	0.49	62876.02	0.68
1	DBSCAN	0.73	2549.46	1.15
2	Hierarchical	0.47	8745.9	0.82

Le tableau ci-dessus montre que l'algorithme KMeans a l'indice Davies-Bouldin le plus bas par rapport aux deux autres algorithmes, on peut donc en conclure que **KMeans a la qualité de clustering acceptable** par rapport aux deux autres algorithmes. Cependant, par score de silhouette, **K-Means a le deuxième score de silhouette le plus élevé.**

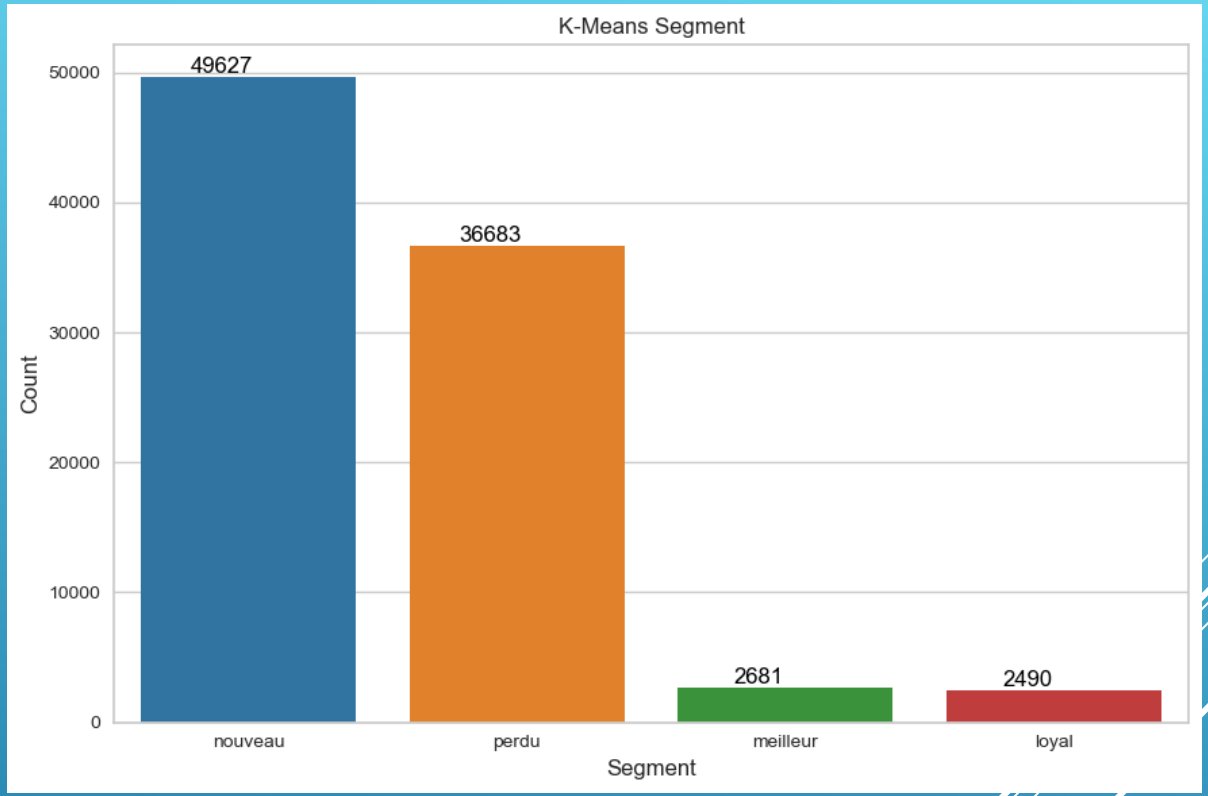
DBSCAN montre le pire indice Davies-Bouldin mais le meilleur score de silhouette par rapport aux autres algorithmes .

D'après les résultats de l' **indice Calinski-Harabasz** , on peut voir que **KMeans a l'indice le plus élevé** par rapport aux autres algorithmes. Cela indique que **KMeans est plus performant et plus dense que les autres algorithmes** .

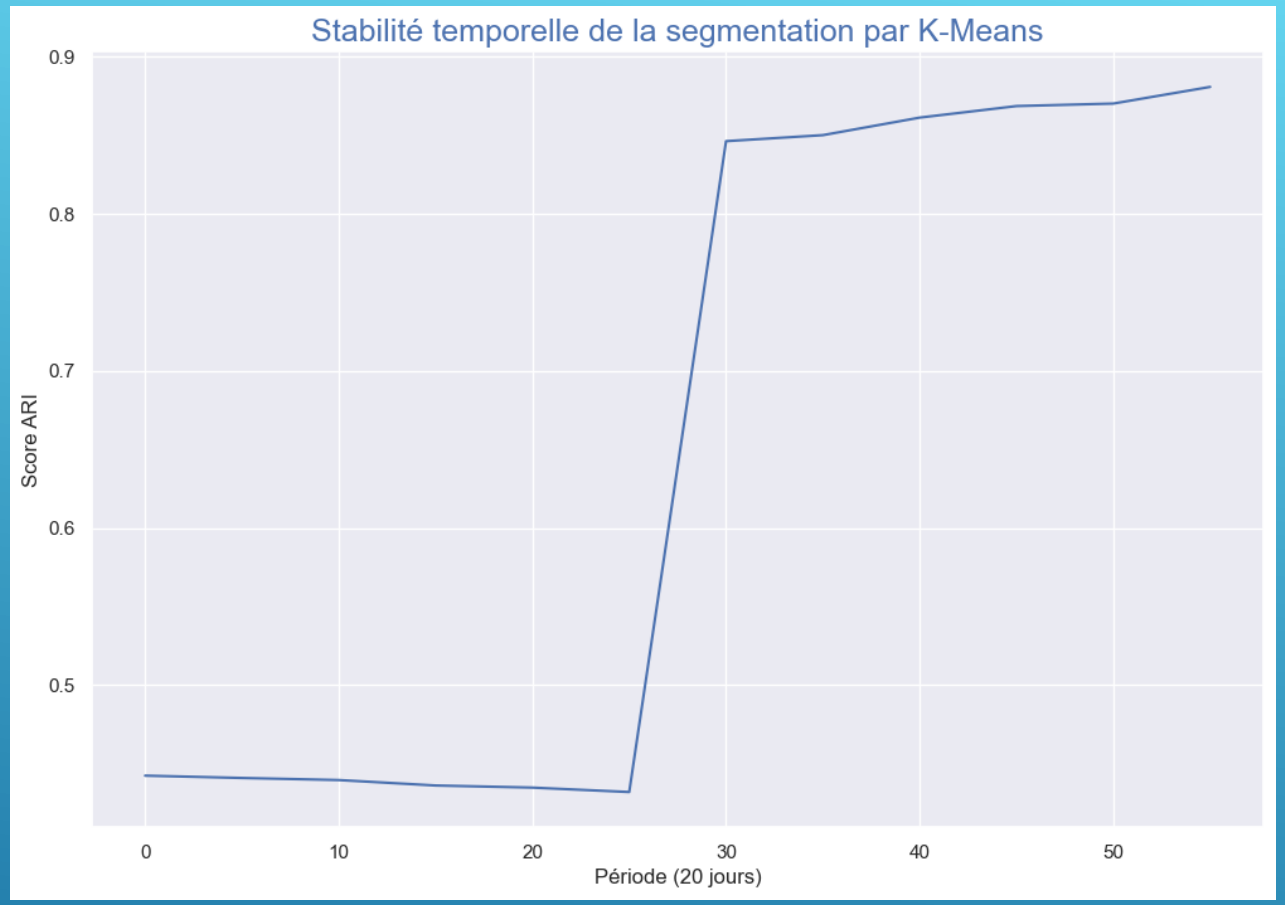
On peut conclure que **KMeans à la meilleure qualité de clustering des trois algorithmes**

Segmentation - suite

Segmentation finale



4- Contrat de Maintenance



Au bout de 30 jours, nous devrions proposer au client une nouvelle segmentation

MERCI