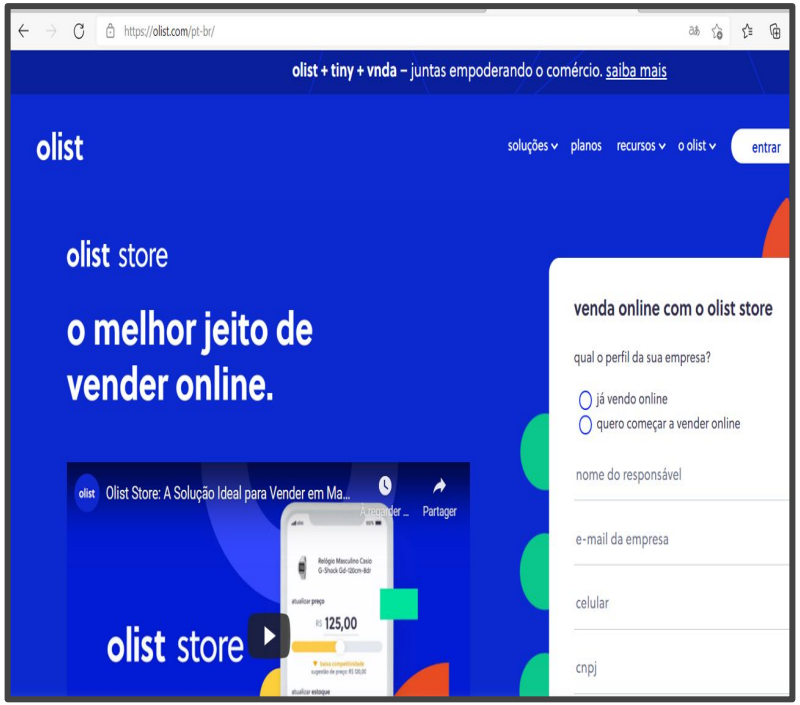
Projet 5 : Segmentez des clients d'un site e-commerce.

Ilham NOUMIR | Parcours Data Science | Date : 19 /11/2021

Sommaire

1. Présentation de la problématique .
2. Préparation et nettoyage des données.
3. Pistes de modélisation effectuées.
4. Présentation du modèle final.

1. Présentation de la problématique :



- **Olist** c'est une solution de vente sur les marketplaces en ligne .

- **Olist** souhaite réaliser une segmentation des clients qui sera destinée aux campagnes de communication.

1. Présentation de la problématique :

Objectif :

- **Comprendre les différents types d'utilisateurs**
- **Fournir à l'équipe marketing une description actionable**
- **Proposition de contrat de maintenance basée sur une analyse de la stabilité des segments au cours du temps.**

Présentation des données :

Le jeu de données comportent 9 dataframes:

1. Clients : **(99441, 5)** **Identifiant unique des clients , localisation (city & state)**
2. Géolocalisation : **(1000163, 5)** **ZipCode , Latitude , longitude , city , state**
3. Commande : **(99441, 8)** **Statue de la commande , Date de l'achat, date de l'approbation , Date de la livraison.**
4. Article par commande : **(112650, 7)** **Nbr article par commande , prix , date limite d'expédition**
5. Paiement : **(103886, 5)** **Type de paiement , valeur de paiement , Séquence de paiement , Nbr de versements.**
6. Revue de la commande **(99224, 7)** **Score de la revue , Commentaire , date de la revue**
7. Produits **(32951, 9)** **Catégorie du produit , longueur , hauteur**
8. Vendeurs :**(3095, 4)** **Identifiant unique du vendeur , zipCode, State , City**
9. Traduction des produits : **(71, 2)** **nom de la catégorie des produits , traduction**

Assemblage des données :

- Client
- Géolocalisation
- Commande
- Article/ Commande
- Paiement
- Revue de la commande
- Produits
- Vendeurs
- Catégorie de produits

Une seule Dataframe
(113218, 28)

Informations Générale :

Intervalle de temps :

Entre Octobre 2016 et Août 2018

Géolocalisation :

Pays : Le Brésil (27 état)

Nombre de client :

91467 Clients uniques

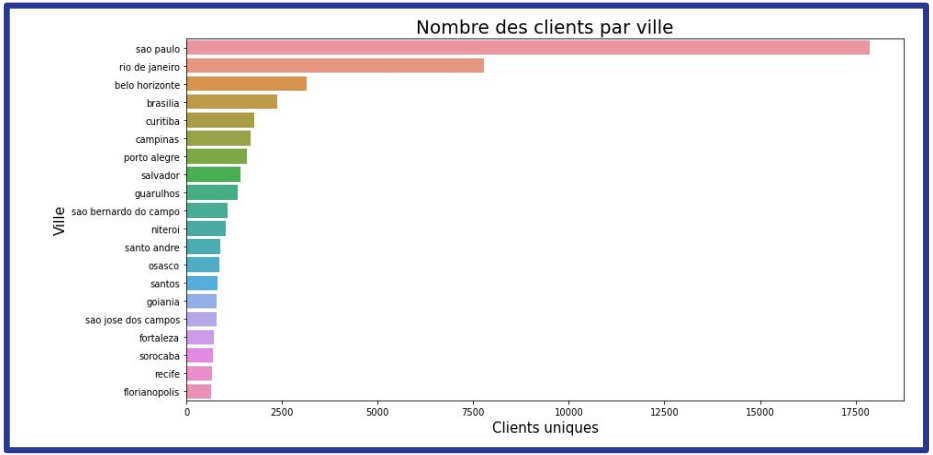
Nombre de Commande :

94473 Commandes passées

Catégorie de produits:

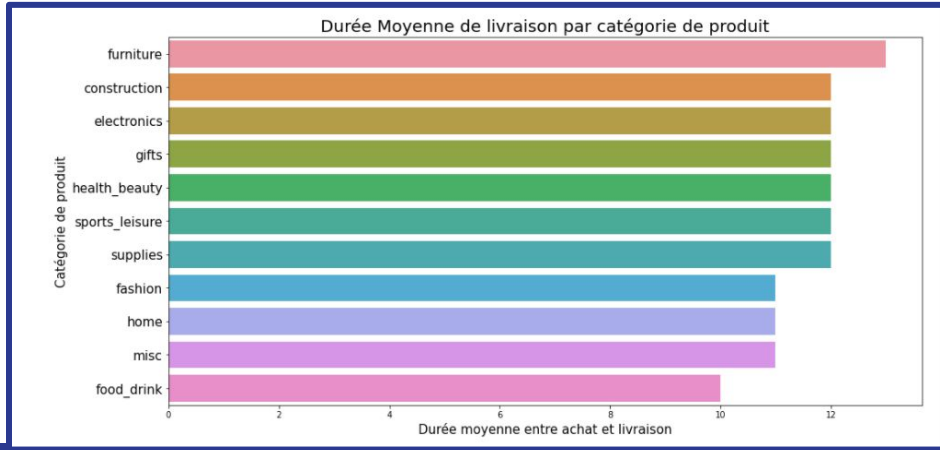
71 Différentes catégories

Exploration des données :

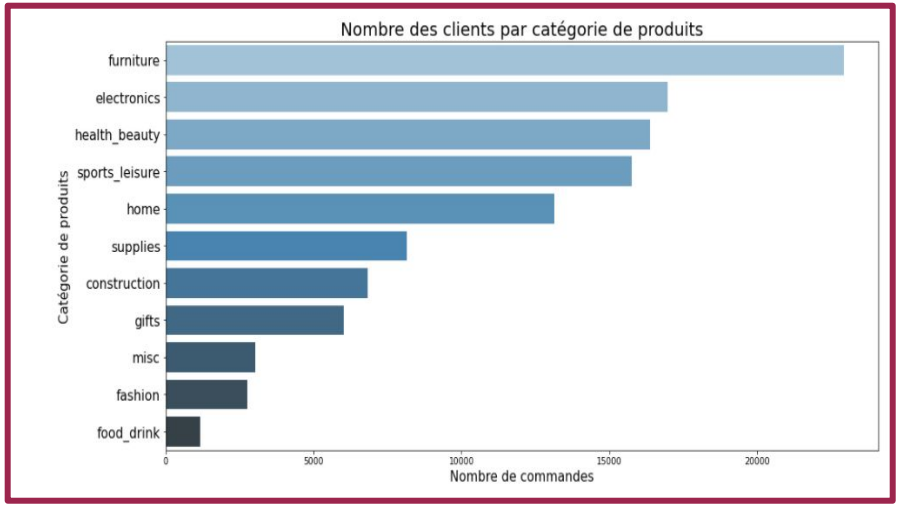


La ville de sao paolo est la ville avec le plus grand nombre de clients qui ont commandé

La catégorie la plus commandé sur le site est la catégorie de fourniture

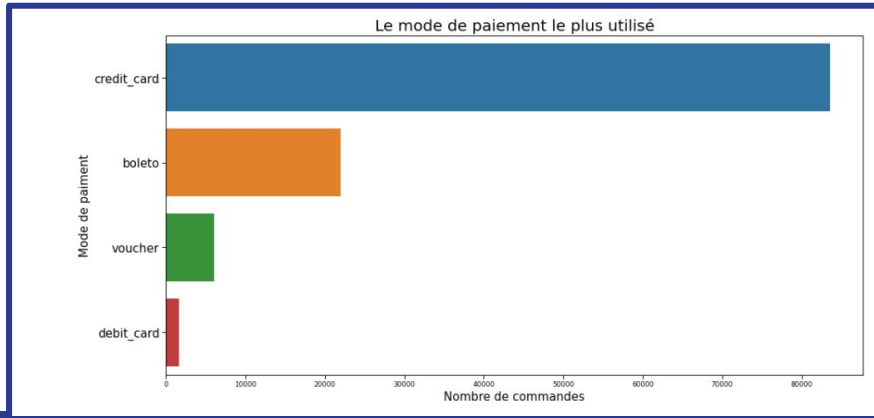


Exploration des données :



La catégorie fourniture est la catégorie qui a plus de client

Le mode de paiement le plus utilisé est le crédit card



Opération de nettoyage effectuée:

Assemblage des données
Vérification de l’existence des valeurs manquantes
Réduction du nombre des catégories des produits
Traitement des outliers
Exploration des données

Feature engineering :

Création d’une nouvelle dataframe à partir des id clients (Identifiant unique pour chaque client)

Dataframe après l’assemblage des 9 tableaux initiales

Dataframe crée en se basant sur les identifiants uniques des clients avec des nouvelles variables

Segmentation RFM :

Fréquence

Elle indique le nombre de fois où un client a fait une transaction sur la période d' étude

Récence

permet de situer le dernier achat dans le temps(en nombre de jours)

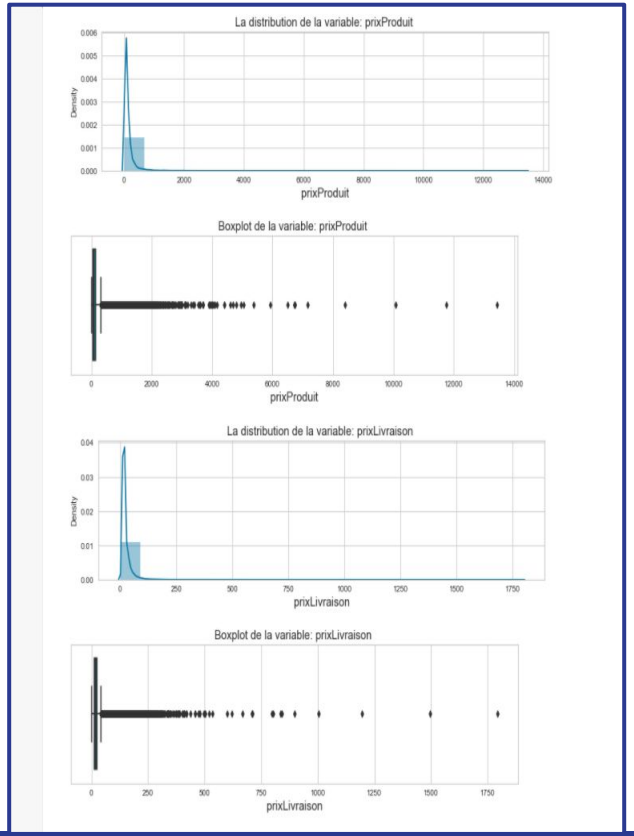
Montant:

Correspond aux sommes des dépenses d'un client

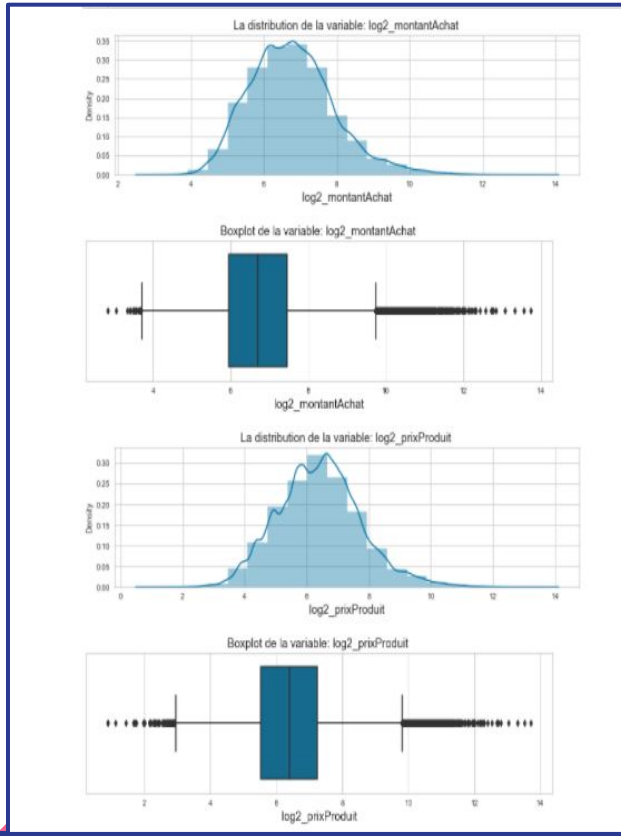
Segmentation RFM : La variable montant

Montants:

- Prix produit
- Frais livraison
- Montant achat



Passage au log

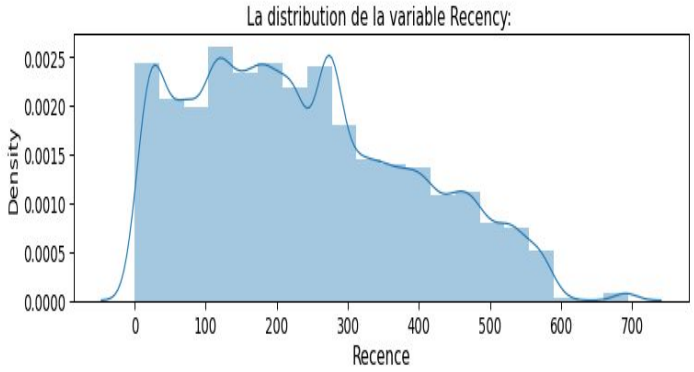


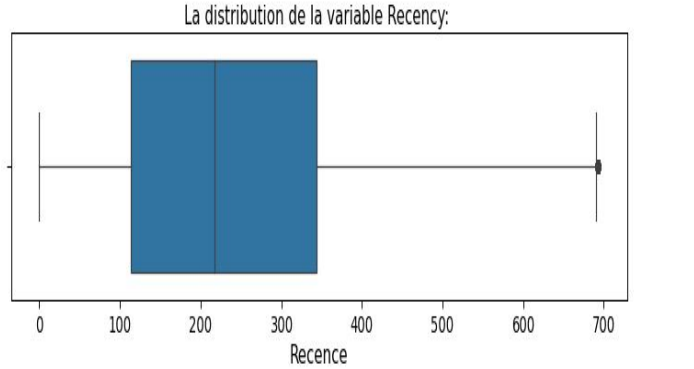
Segmentation RFM : La variable récence

Récence:

Pour calculer la récence, nous devons connaître la date d'achat la plus récente de chaque client et voir pendant combien de jours ils sont inactifs

	customer_unique_id	Recency
0	0000366f3b9a7992bf8c76cfdf3221e2	111
1	0000b849f77a49e4a4ce2b2a4ca5be3f	114
2	0000f46a3911fa3c0805444483337064	536
3	0000f6ccb0745a6a4b88665a16c9f078	320
4	0004aac84e0df4da2b147fca70cf8255	287
5	0004bd2a26a76fe21f786e4fbd80607f	145
6	00050ab1314c0e55a6ca13cf7181fecf	131
7	00053a61a98854899e70ed204dd4bafe	182
8	0005e1862207bf6ccc02e4228effd9a0	542
9	0005ef4cd20d2893f0d9fbd94d3c0d97	169
10	0006fdc98a402fceb4eb0ee528f6a8d4	407
11	00082cbe03e478190aadbea78542e933	282
12	00090324bbad0e9342388303bb71ba0a	158
13	000949456b182f53c18b68d6babc79c1	128
14	000a5ad9c4601d2bbdd9ed765d5213b3	383
15	000bfa1d2f1a41876493be685390d6d3	334
16	000c8bdb58a29e7115cfc257230fb21b	259
17	000d460961d6dbfa3ec6c9f5805769e1	233
18	000de6019bb59f34c099a907c151d855	376
19	000e309254ab1fc5ba99dd469d36bdb4	65



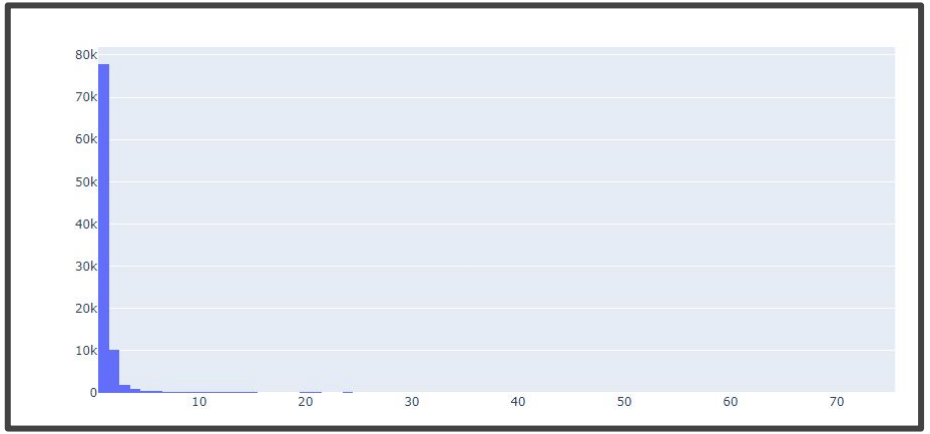


Segmentation RFM : La variable récence

Fréquence :

La fréquence indique le nombre de fois où le client a fait une transaction . Plus celle-ci sera élévée plus sera la valeur du client.

	frequence	nbrClient
0	1	77784
1	2	10120
2	3	1877
3	4	875
4	5	306
5	6	270
6	7	70
7	8	41
8	12	25
9	9	22
10	10	22
11	11	16
12	14	9
13	15	6
14	24	5
15	13	5
16	20	3
17	21	3
18	16	1
19	18	1
20	19	1
21	22	1
22	75	1
23	26	1
24	35	1

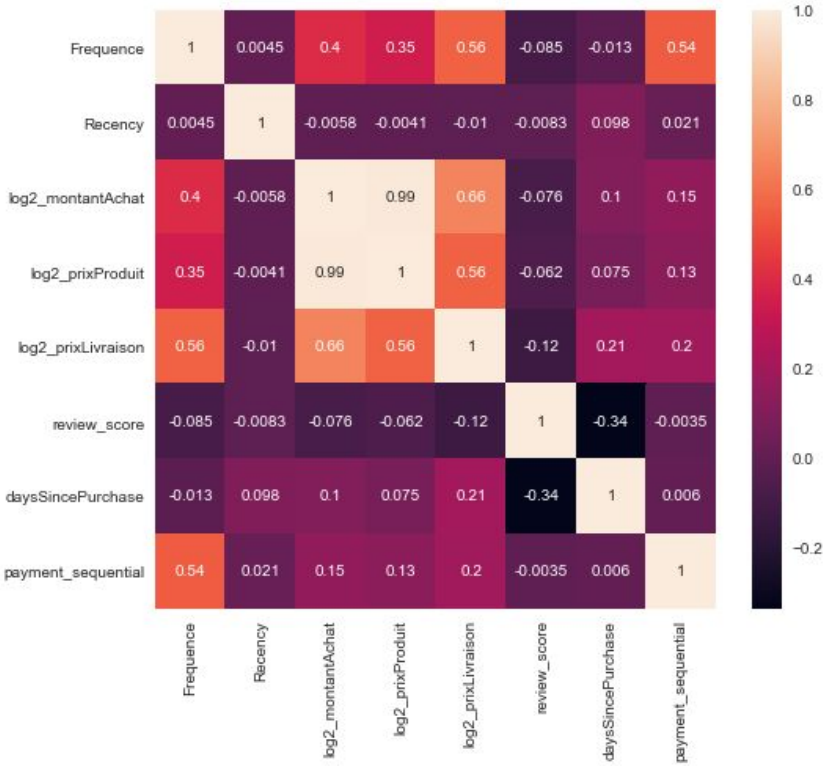


Autres variables :

A partir de l'identifiant unique des clients j'ai créé d'autres variables :

- review_score moyen par client unique ;
- product_categorie por la catégorie la plus commandée pour chaque client ;
- payment_type : pour la catégorie de paiement la plus utilisée pour chaque client ;
- customer_state : pour la ville de chaque client unique;
- payment_sequential : qui informe sur combien de fois un client a payé la somme de l'achat

Analyse bivariée

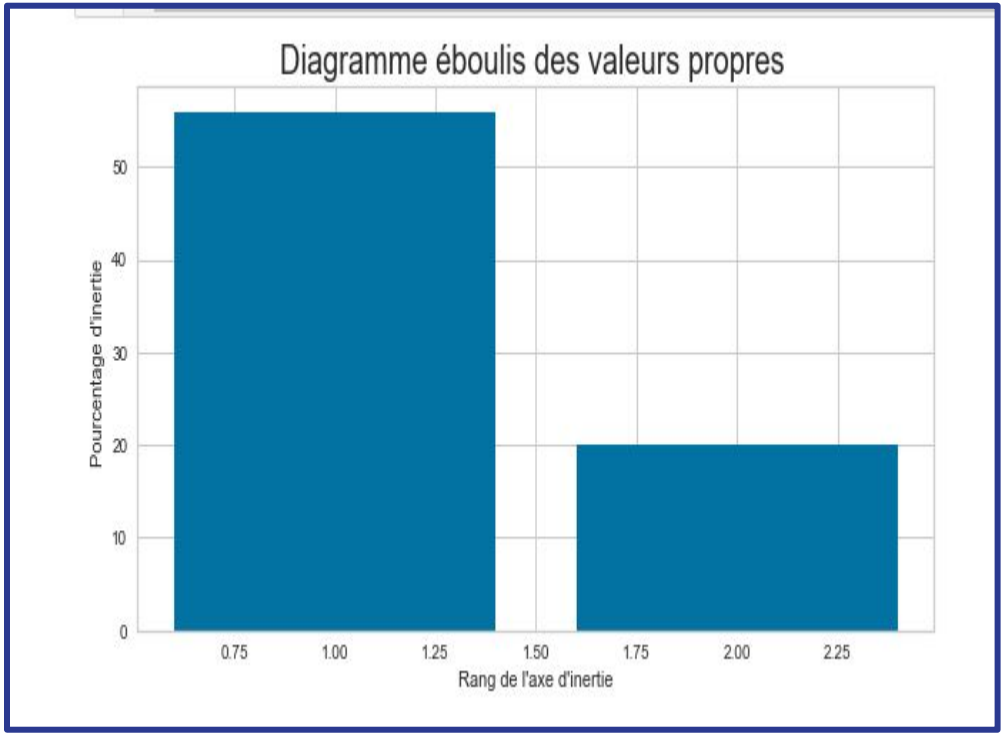


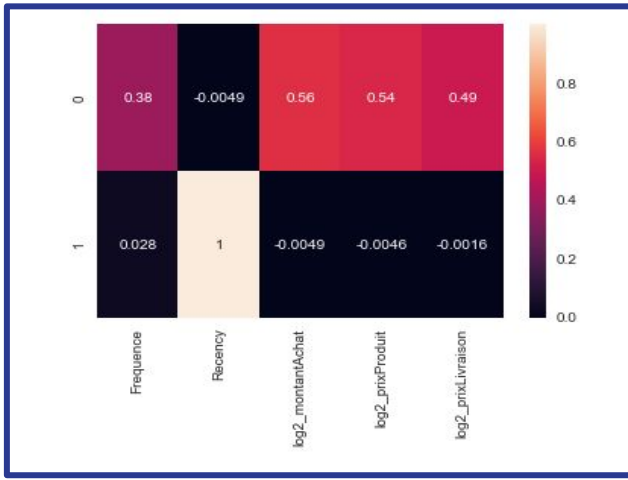
Réduction dimensionnelle :

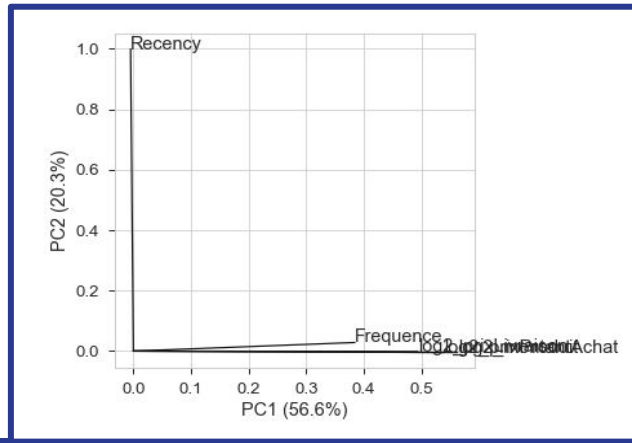
La réduction dimensionnelle est utilisée dans ce projet pour visualiser et comprendre mes données :

- Réduction dimensionnelle linéaire : l’ACP (Analyse en composante principale);
- Réduction dimensionnelle non linéaire : t-SNE (une méthode non linéaire qui favorise la structure locale et s'intéresse aux voisinages de chacun des points)

Réduction dimensionnelle : ACP



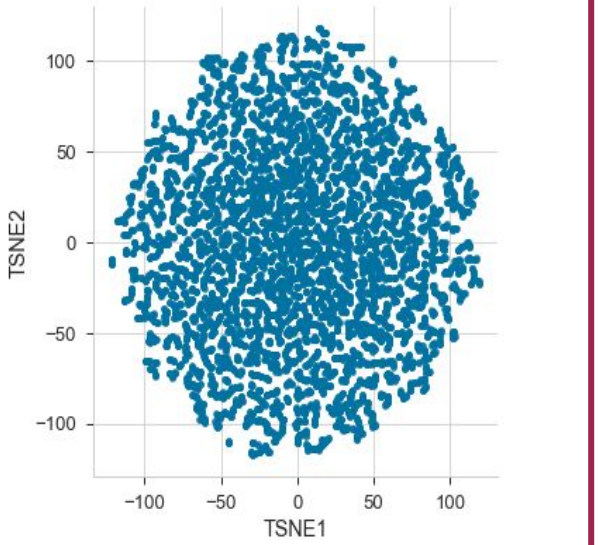




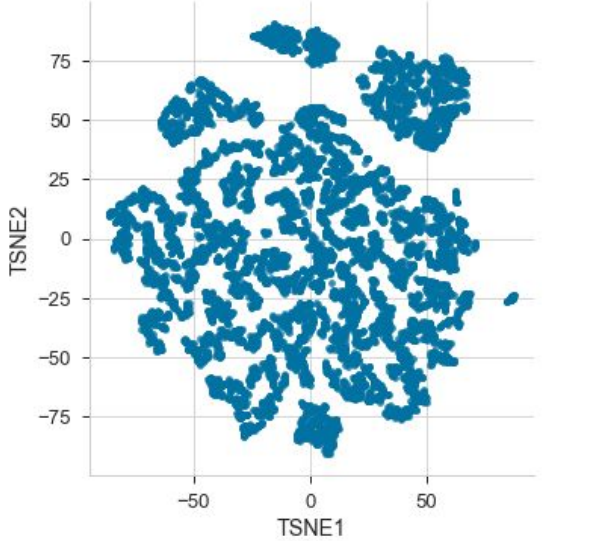
Réduction dimensionnelle : t-SNE

En variant l'hyper paramètre de la perplexité (le voisinage des points) : 5 , 30 , 50

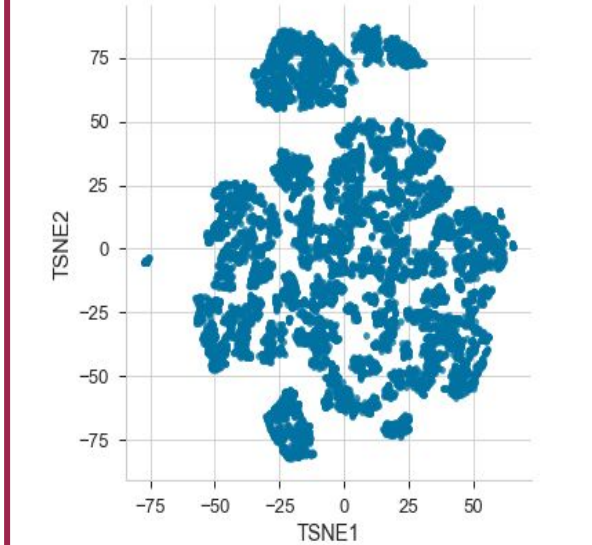
Perplexity : 5



Perplexity :30



Perplexity :50



Pistes de modélisation choix du nombre de cluster :

Score de silhouette : Différence entre les distances intra_cluster et les distances au cluster le plus proche (à maximiser).

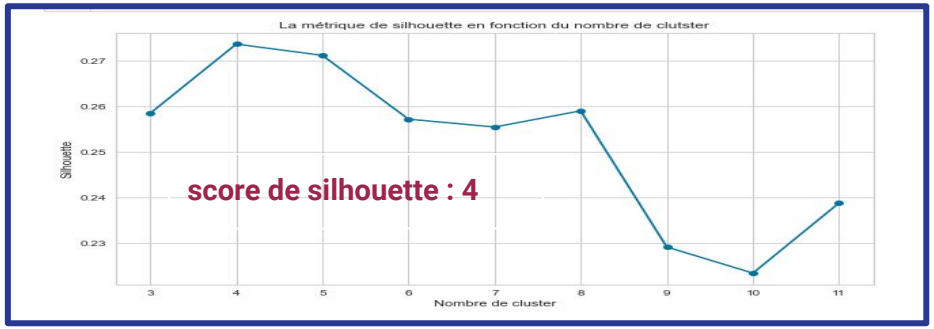
Score de distorsion : La moyenne de la somme des carrés des distances au centroïde le plus proche(coude/Elbow)

Score de Calinski_harabaz: Rapport entre la variance inter-groupes et la variance intra-groupe(à maximiser)

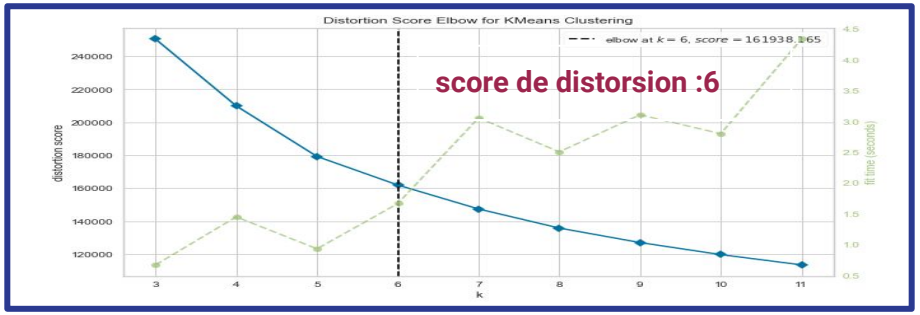
Score de Davis Bouldin : Moyenne du rapport maximal entre la distance entre deux centres de groupes (à minimiser)

Modélisation avec kMeans : Choix du nombre de clusters

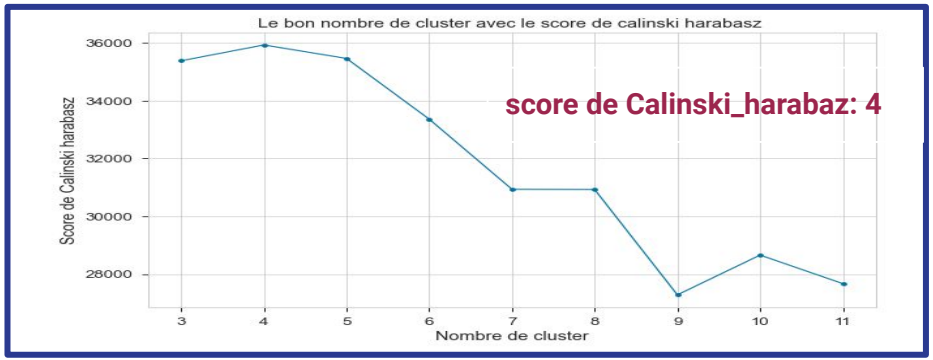
Différence entre les distances intra_cluster et les distances au cluster le plus proche (à maximiser)



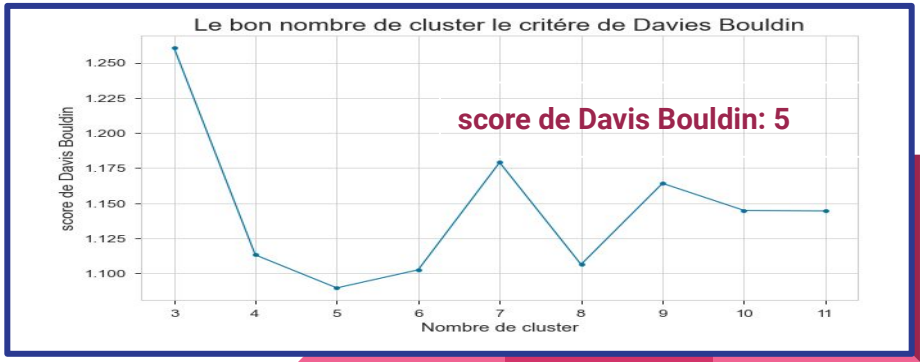
La moyenne de la somme des carrés des distances au centroïde le plus proche(coude/Elbow)



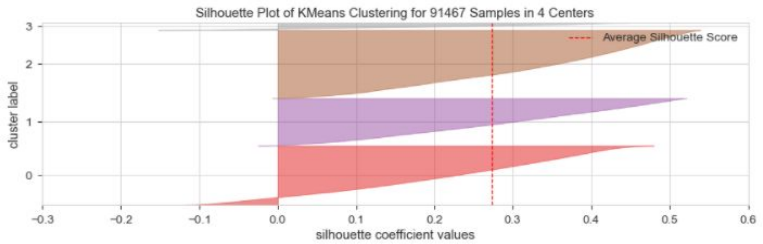
Rapport entre la variance inter-groupes et la variance intra-groupe(à maximiser)

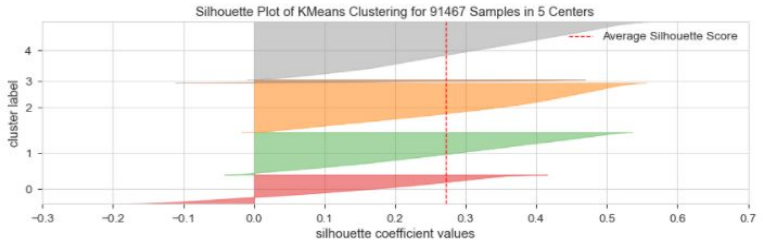


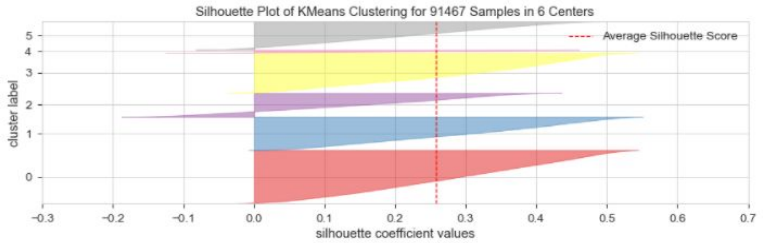
Moyenne du rapport maximal entre la distance entre deux centres de groupes (à minimiser)



Pistes de modélisation : La forme de silhouette et visualisation avec le t-SNE



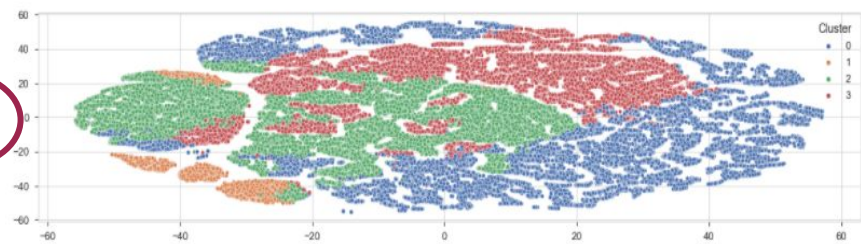


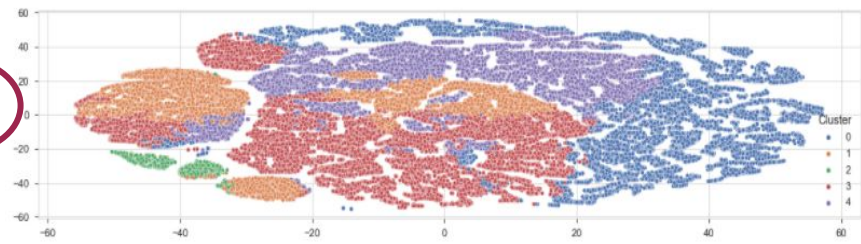


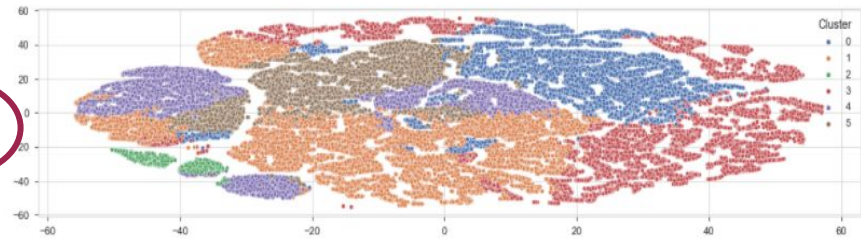
4

5

6



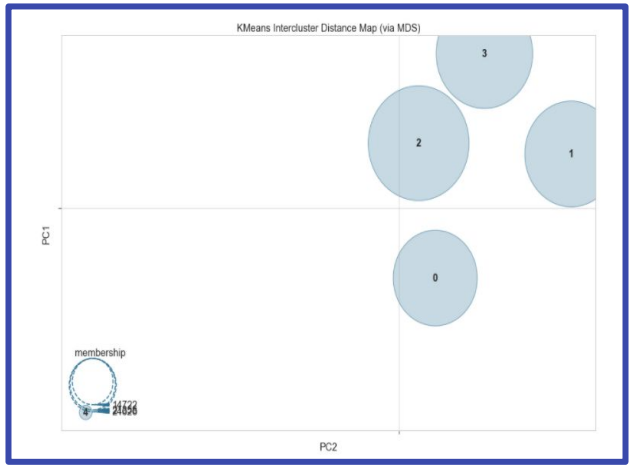




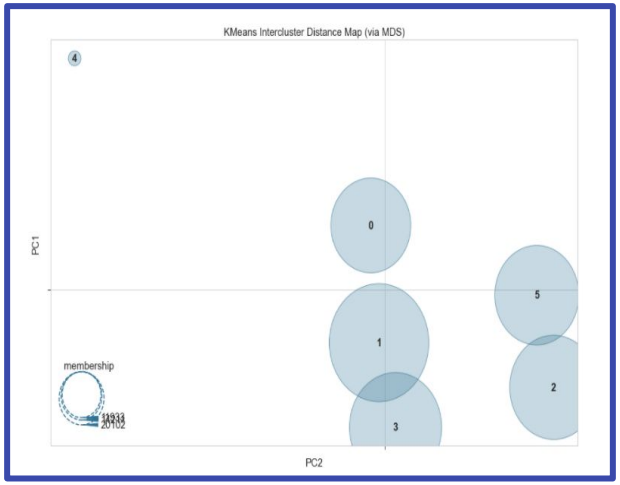
Pistes de modélisation : Séparation des clusters

4

5



6



Stabilité à l'initialisation du KMeans :

	nbrCluster	ARI_moyen
0	4	0.97
1	5	0.94
2	6	0.95

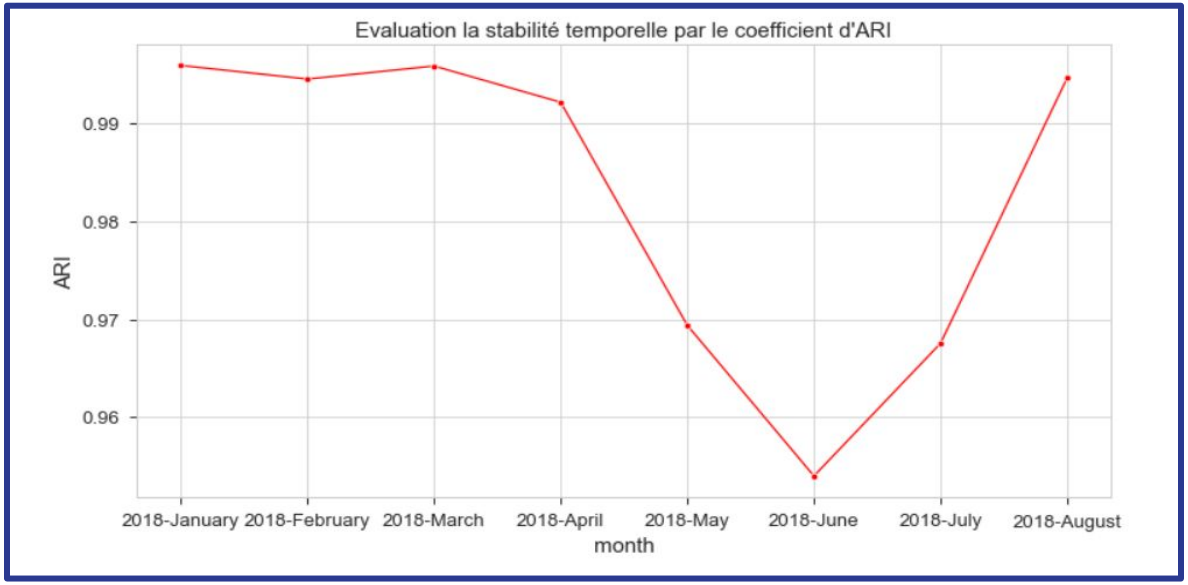
En prenant deux partitions des données et on comparant les ARI . On remarque que tous les clusters 4, 5, et 6 sont stables avec un petit avantage pour le nombre de cluster 4

Stabilité temporelle des clusters :

Pour évaluer la stabilité des clusters au fil du temps .J'ai utilisé la méthode décrit ci-après :

1. Fixer une période de base (l'année 2017 avec 44000 clients).
2. Prédire les labels sur cette période . Ces labels vont servir de référence par la suite.
3. Ajouter des périodes (un mois à la fois) et fiter à nouveau l'algorithme.
4. Faire appel aux labels de base(les labels de la période de 2017).
5. Calculer l'ARI entre les labels de base et les nouveaux labels .

Stabilité temporelle des clusters :



	month	ARI
0	2018-January	0.995929
1	2018-February	0.994537
2	2018-March	0.995870
3	2018-April	0.992166
4	2018-May	0.969368
5	2018-June	0.953975
6	2018-July	0.967534
7	2018-August	0.994658

On voit une rechute à partir du mois de Juin.

Il faudra donc prévoir la maintenance du programme de segmentation tous les 6 mois et de redéfinir les segments clients après chaque période de maintenance.

Construction des clusters

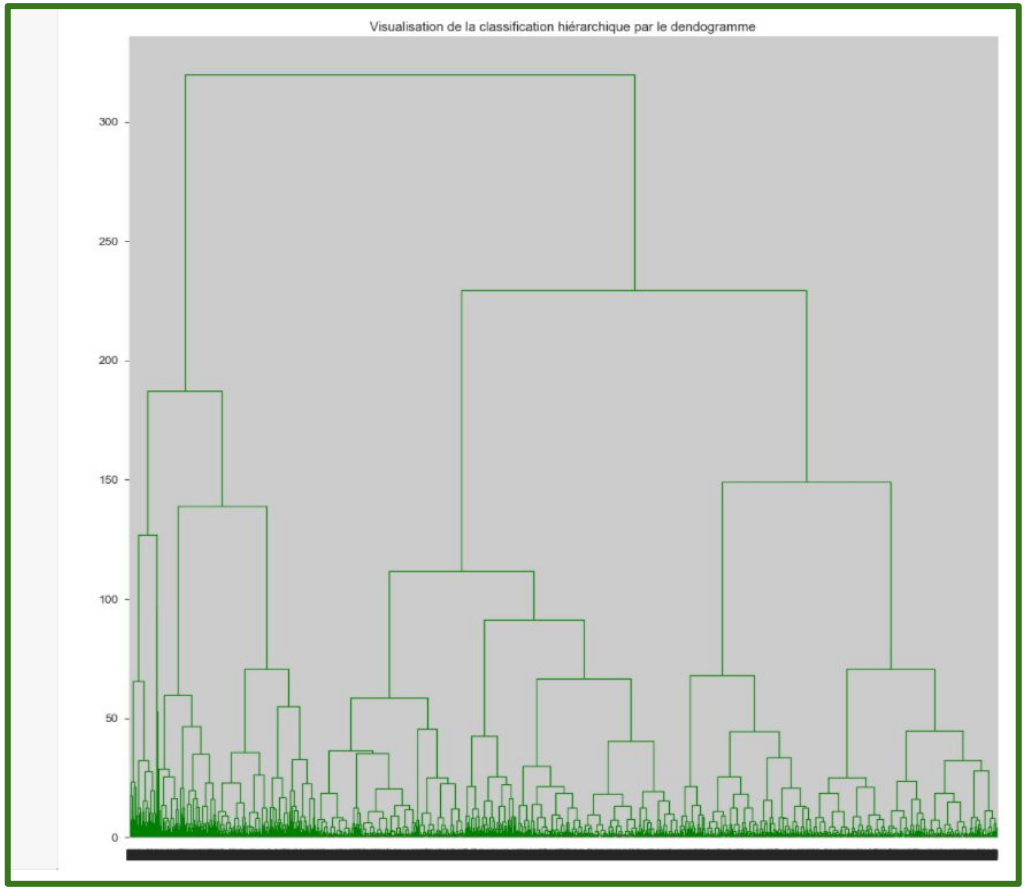
nbrClients	prixMoyen	livraisonMoyenne	montantAchatMoyen	frequence	recence	scoreRevueMoyen	categorie	state	payment_type
34212	56.38	16.39	71.77	1.0	144.0	4.2	electronics	SP	credit_card
3645	860.47	174.81	1034.28	4.0	238.0	3.6	furniture	SP	credit_card
23841	85.18	18.99	103.18	1.0	416.0	4.2	furniture	SP	credit_card
29769	278.92	36.98	314.91	1.0	197.0	4.1	furniture	SP	credit_card

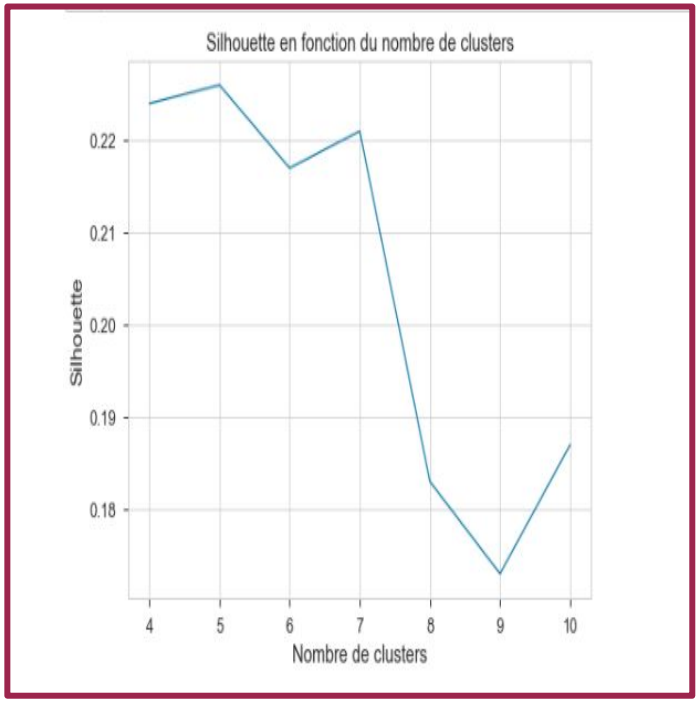
Cluster 1 : Nombre de client le plus élevé avec la catégorie “electronics”, ayant le montant d’achat moyen le plus bas et ils sont plus actifs sur le site(La récence la plus basse)

Cluster 2 : Nombre de client est le mois élevé avec la catégorie “furniture” ,ayant le montant d’achat le plus élevé et ils sont moyennement présents sur le site(La récence la plus basse) .

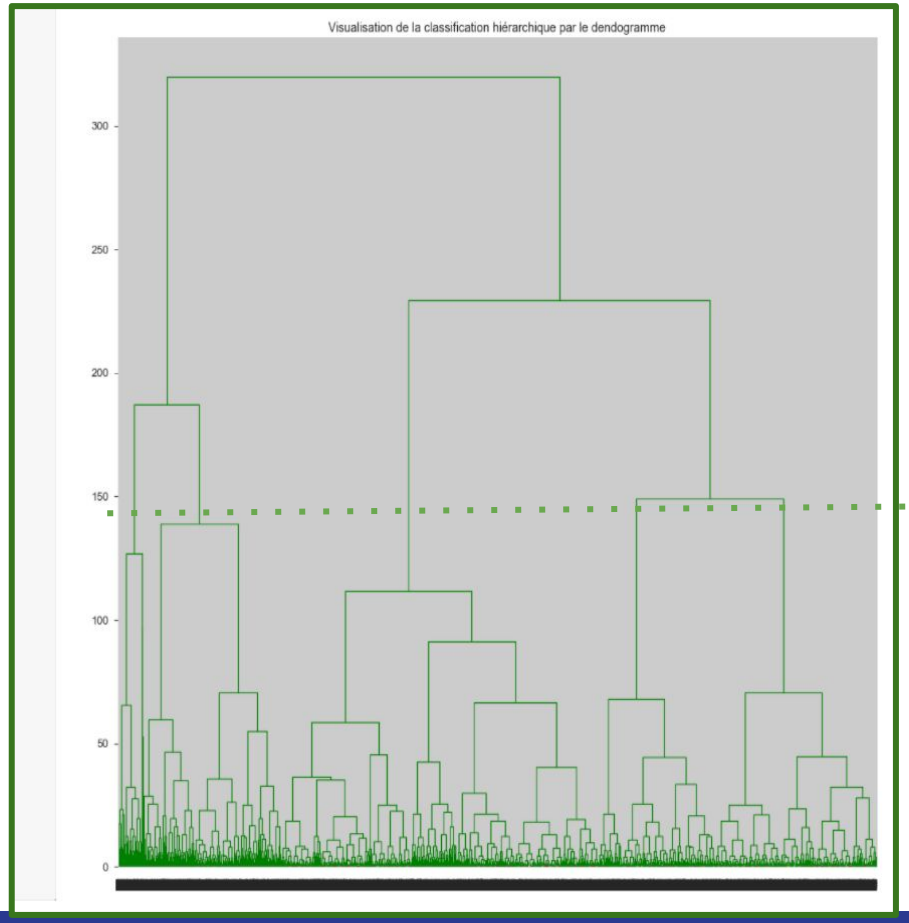
Cluster 3 et **Cluster 4** : sont rapprochés en terme de nombre des clients mais avec des montants d’achat différents et différence en terme d’activité et d’inactivité dans le site

Autres algorithmes : Classification ascendante hiérarchique (CAH)

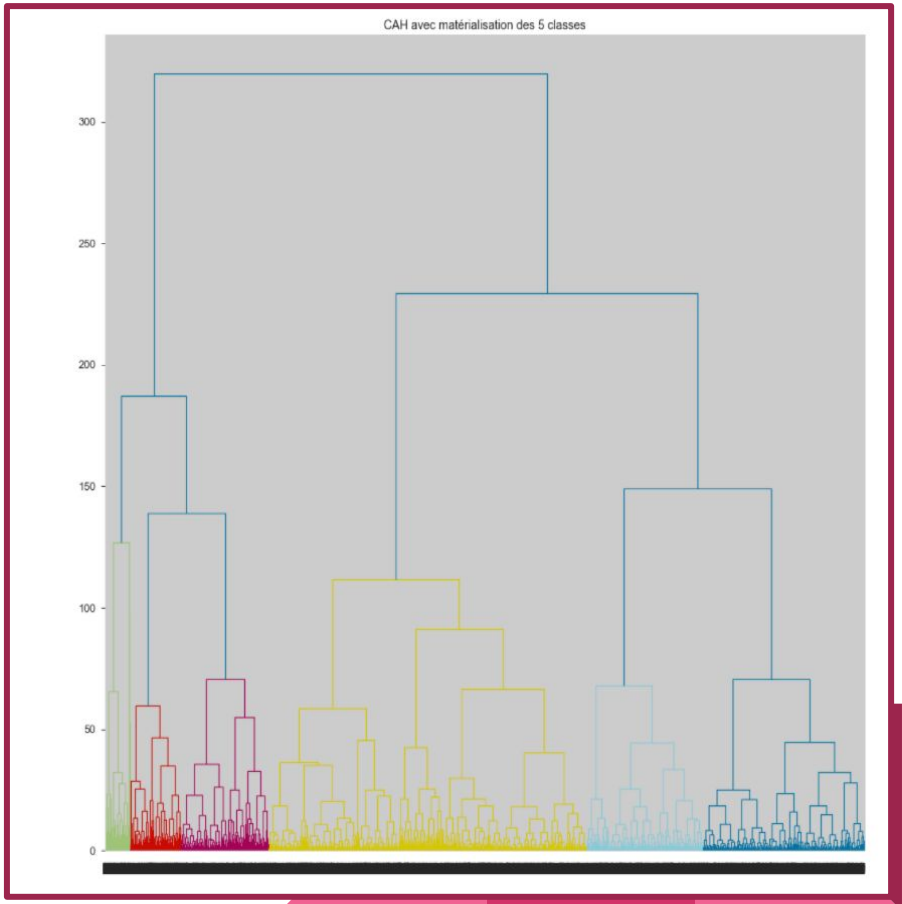




Autres algorithmes : Classification ascendante hiérarchique (CAH)



138

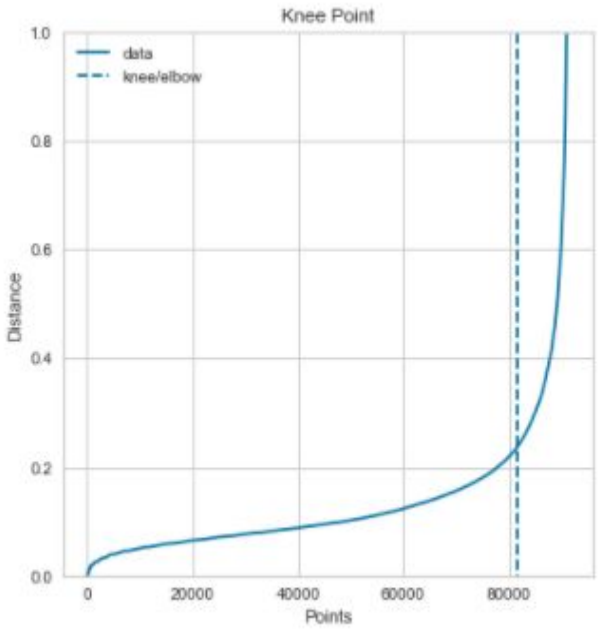


Autres algorithmes : DBSCAN

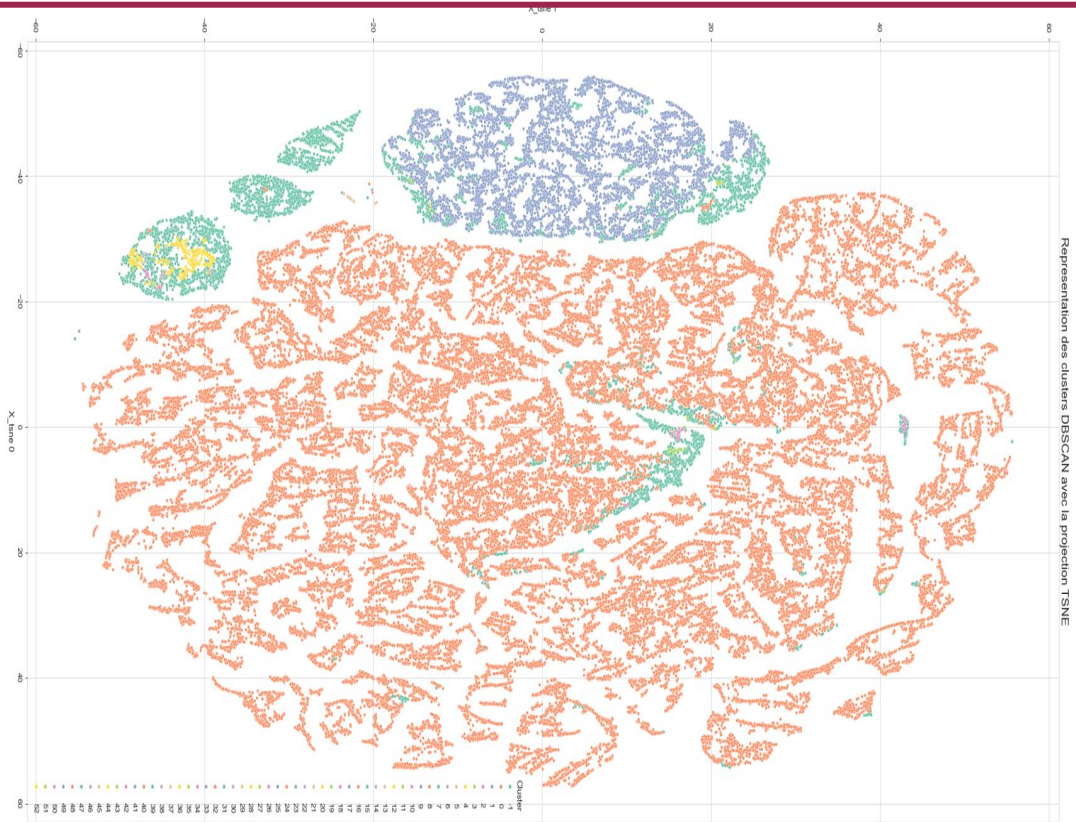
min_samples : Ce paramètre fait référence au nombre de points voisins requis pour qu'un point soit considéré comme une région dense ou un cluster valide.

eps: C'est la distance la plus éloignée à laquelle un point choisira ses voisins.Intuitivement, cela décidera du nombre de voisins qu'un point découvrira

Autres algorithmes : DBSCAN



0.23639103088525346



Conclusion :

La segmentation RFM et le clustering KMeans nous a permis de déceler 4 segments de clients :

- Des clients qui dépensent beaucoup avec une fréquence élevée et une récence élevée.
  Des clients (mais ils sont moins nombreux 4%).
- Des clients qui consomment le moins sur le site mais ils sont plus actifs et avec un nombre très élevé (37,4%)
- Des clients dont la consommation est moyenne mais ils sont plus actifs et très nombreux (33%)
- Des clients dont la consommation est faible ils sont très inactifs et moyennement (nombreux (26%)

La maintenance de la segmentation doit être effectuée dans le deuxième trimestre de l'année.

En comparaison des autres algorithmes de clustering le KMeans et la classification hiérarchique dont des résultats très proches (4 pour le KMeans et 5 pour le DBSCAN).