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

Guille Anaïs – Parcours Data Scientist

Mentor : Ahmed Tidiane Balde

Sommaire

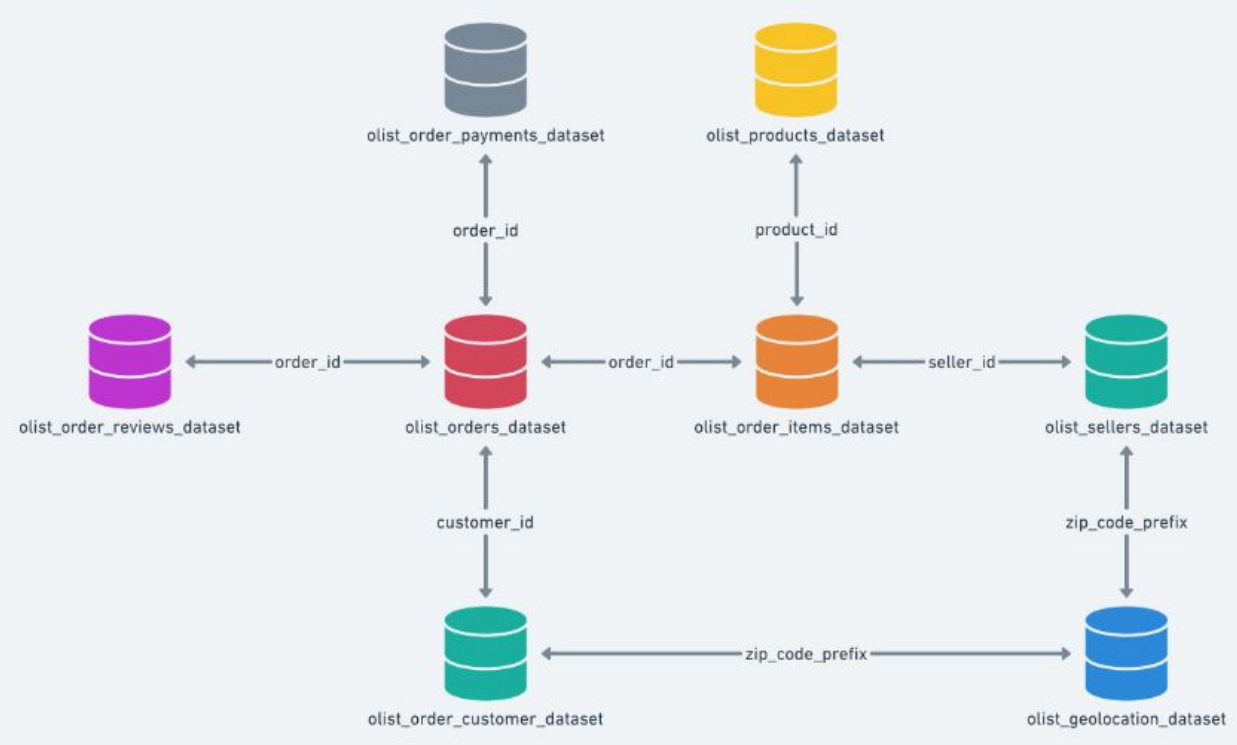
I- Problématique

- ***Olist*** : Entreprise brésilienne spécialisée dans la vente sur les marketplace en ligne

* *Fournir une segmentation des clients pour leurs équipes*

* *Recommander la fréquence de mise à jour de cette segmentation pour rester pertinente*

II- Présentation du jeu de données



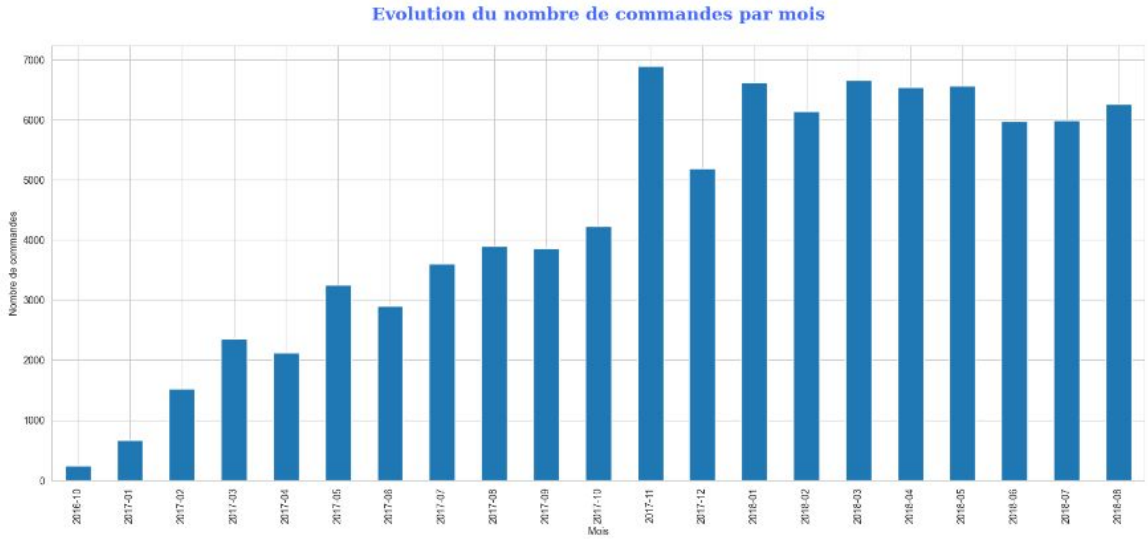
- 9 fichiers csv
- Base de données anonymisée
- Commande de 2016 à 2018

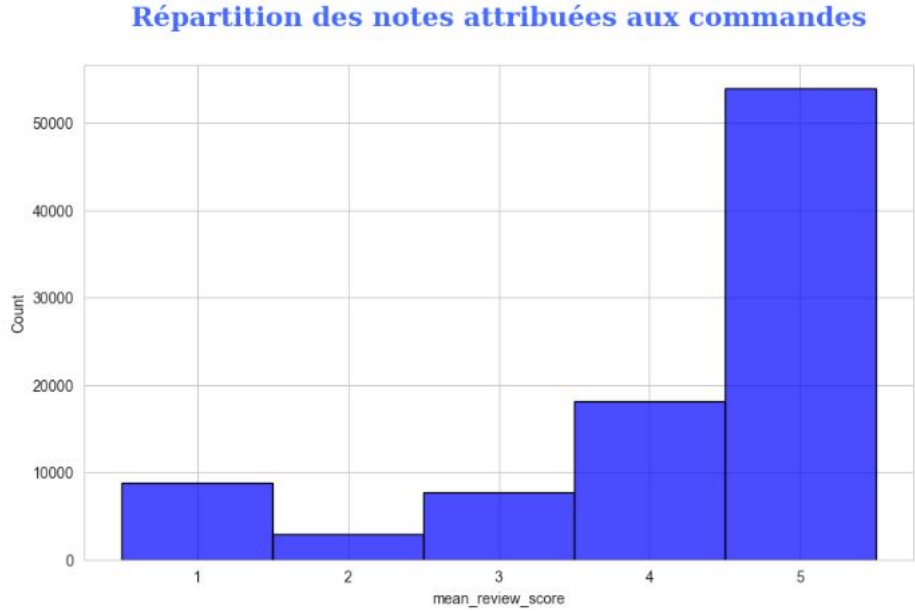
III- Nettoyage des données

- Suppression des données dupliquées
- Suppression des colonnes inutilisées pour le projet
- Conversion des colonnes temporelles en datetime
- Conservation des commandes ‘delivered’ uniquement
- Agrégation et regroupement du dataframe par identifiant client unique

III- Nettoyage des données et feature engineering

#	Column	Non-Null Count
0	order_purchase_timestamp	91481 non-null
1	customer_state	91481 non-null
2	order_id	91481 non-null
3	nb_total_item	91481 non-null
4	price	91481 non-null
5	freight_value	91481 non-null
6	payment_type	91481 non-null
7	mean_payment_installments	91481 non-null
8	total_payment_value	91481 non-null
9	mean_payment_value	91481 non-null
10	mean_review_score	91481 non-null
11	seller_state	91481 non-null
12	product_category_name	91481 non-null

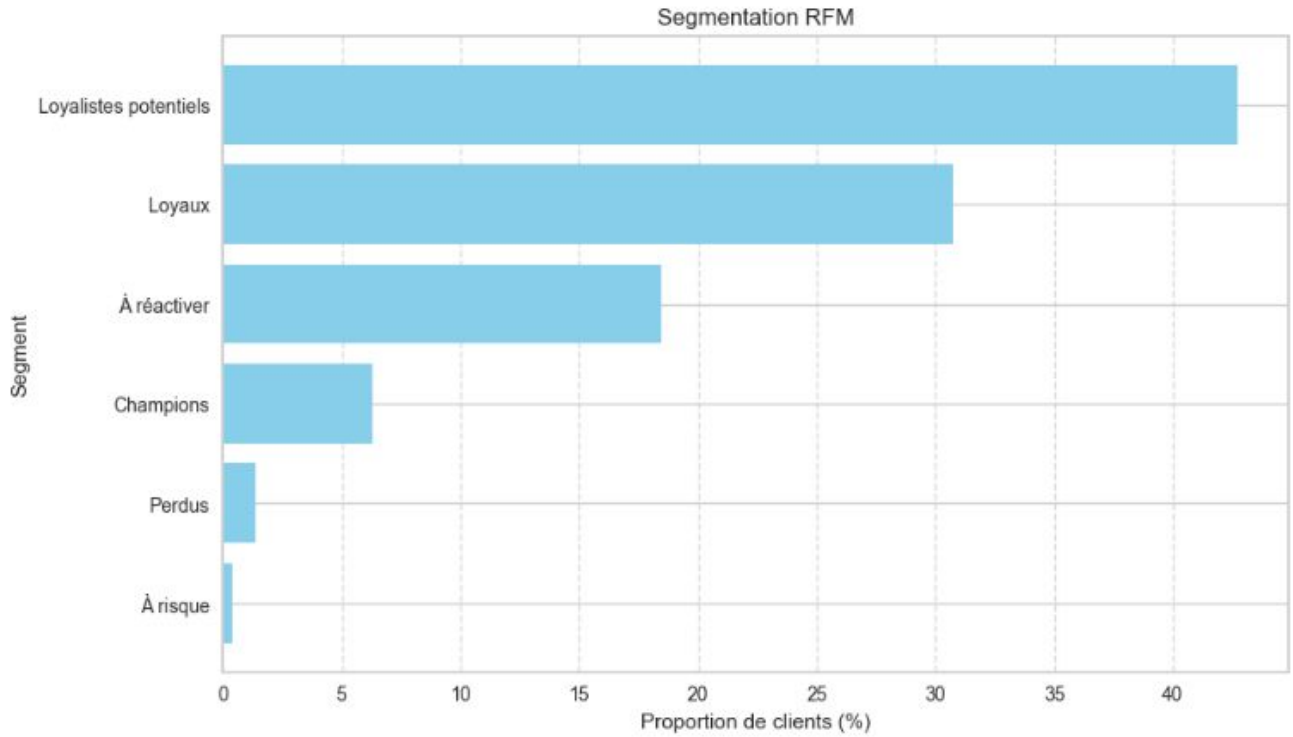




III- Nettoyage des données et feature engineering

La segmentation RFM (Recency, Frequency, Monetary)

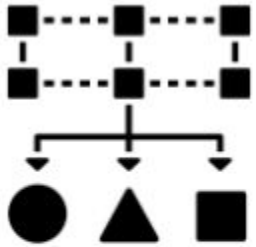
	recency	frequency	monetary
count	91481.000000	91481.000000	91481.000000
mean	236.108875	1.032870	173.092009
std	152.586572	0.206215	257.592652
min	0.000000	1.000000	10.070000
25%	113.000000	1.000000	64.000000
50%	217.000000	1.000000	110.170000
75%	344.000000	1.000000	188.500000
max	694.000000	14.000000	13664.080000



IV - Elaboration d'un modèle de clustering

Preprocessing

StandardScaler



Classification non-supervisée

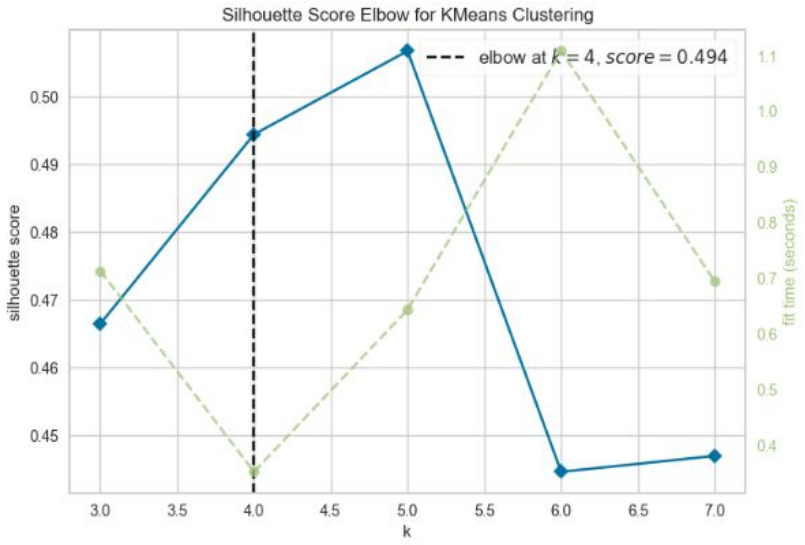
Kmeans
DBSCAN
Agglomerative Clustering

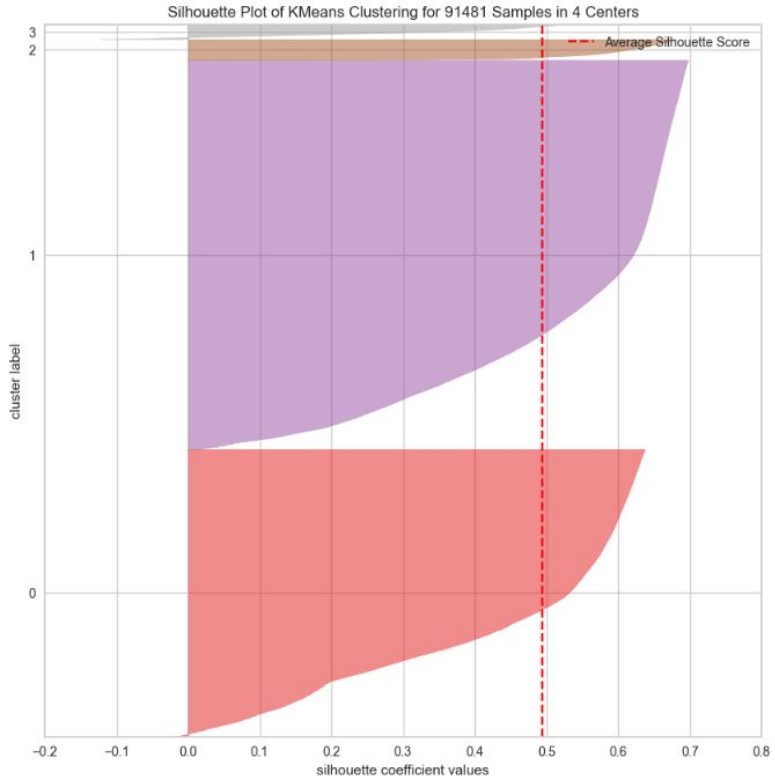
Choix des paramètres

Analyse des clusters

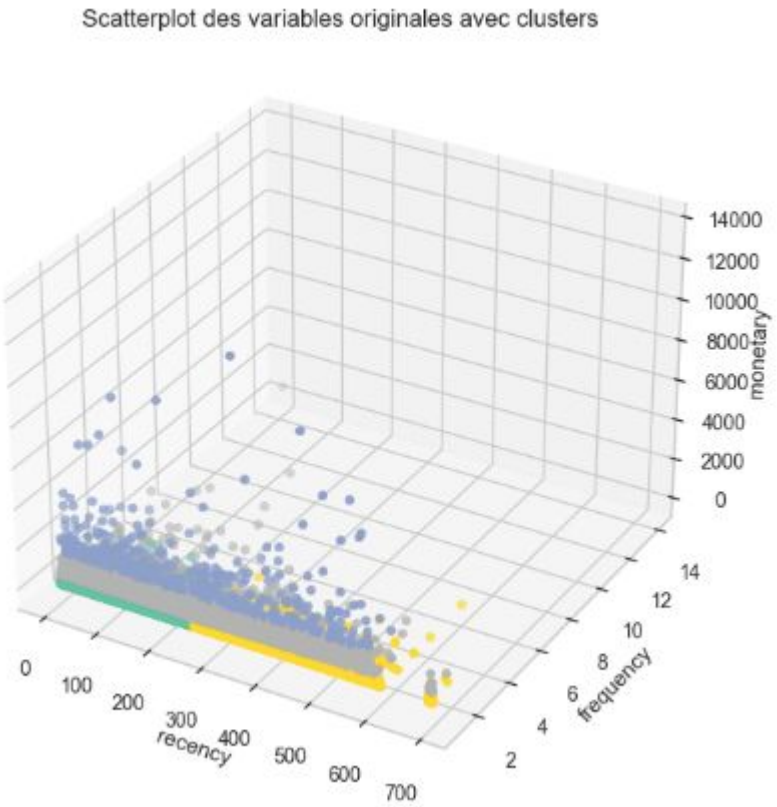
IV - Elaboration d'un modèle de clustering

1) *Kmeans*





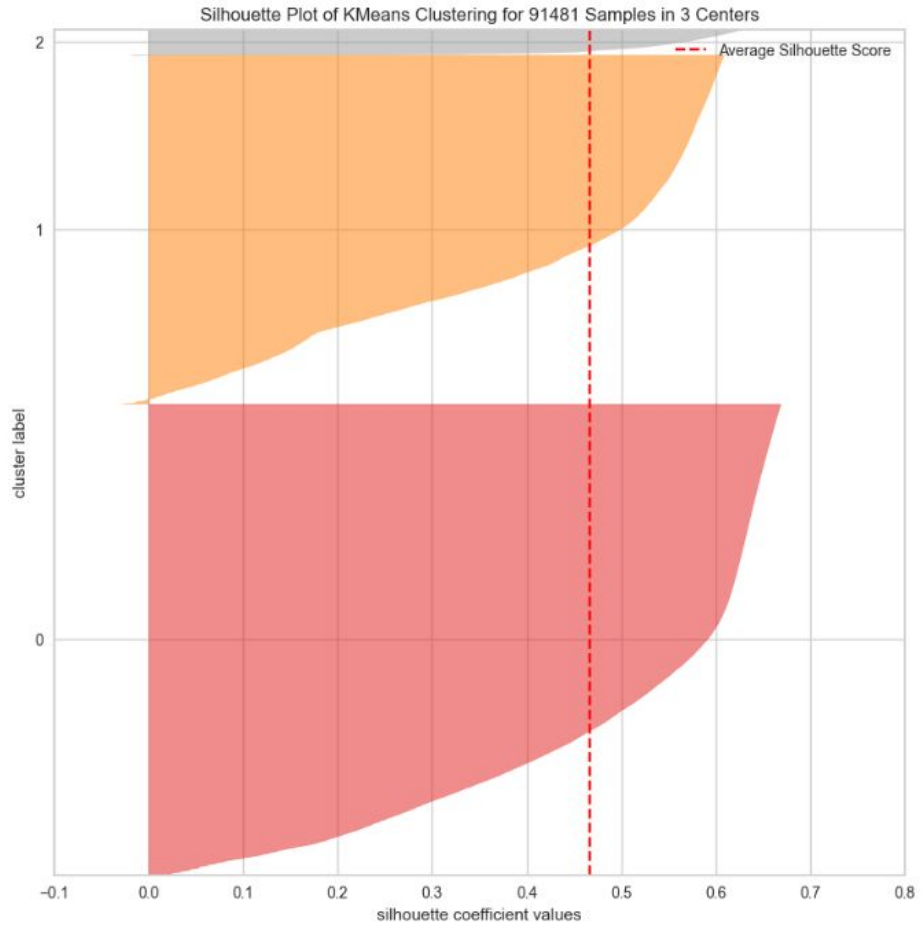
	k	sil_scores	temps
0	3	0.466489	0.713639
1	4	0.494458	0.352348
2	5	0.506776	0.643865
3	6	0.444635	1.108587
4	7	0.446988	0.694277




```
Nombre de clients par cluster :
cluster
0    49264
1      464
2    36476
3     5277
Name: count, dtype: int64
```

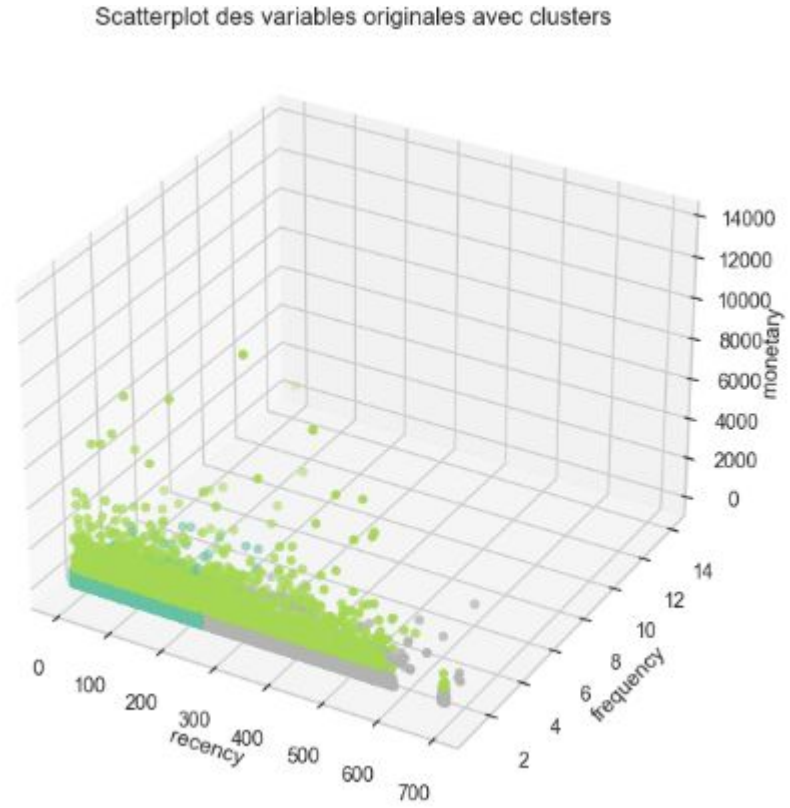
IV - Elaboration d'un modèle de clustering

1) *Kmeans*



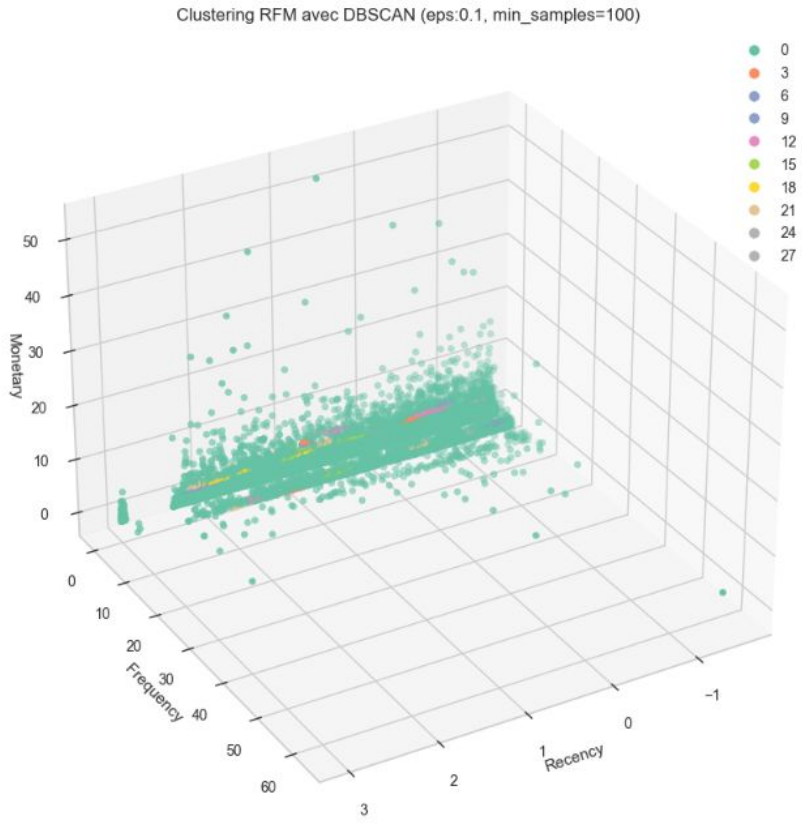
	k	sil_scores	temps
0	3	0.466489	0.713639
1	4	0.494458	0.352348
2	5	0.506776	0.643865
3	6	0.444635	1.108587
4	7	0.446988	0.694277

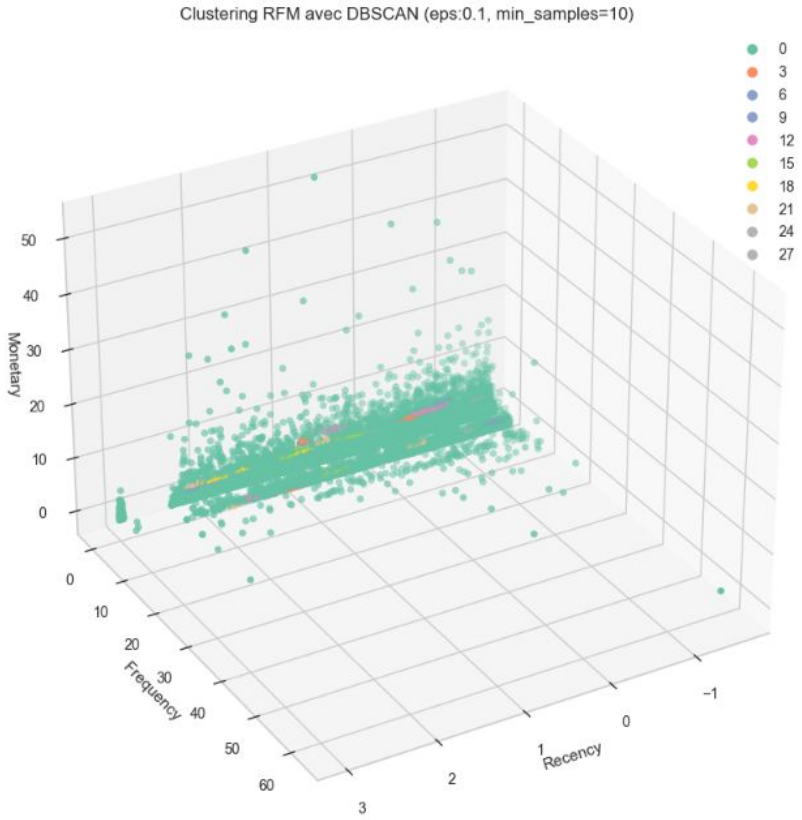
```
Nombre de clients par cluster :
cluster
0    51208
1     2583
2    37770
```



IV - Elaboration d'un modèle de clustering

2) DBSCAN

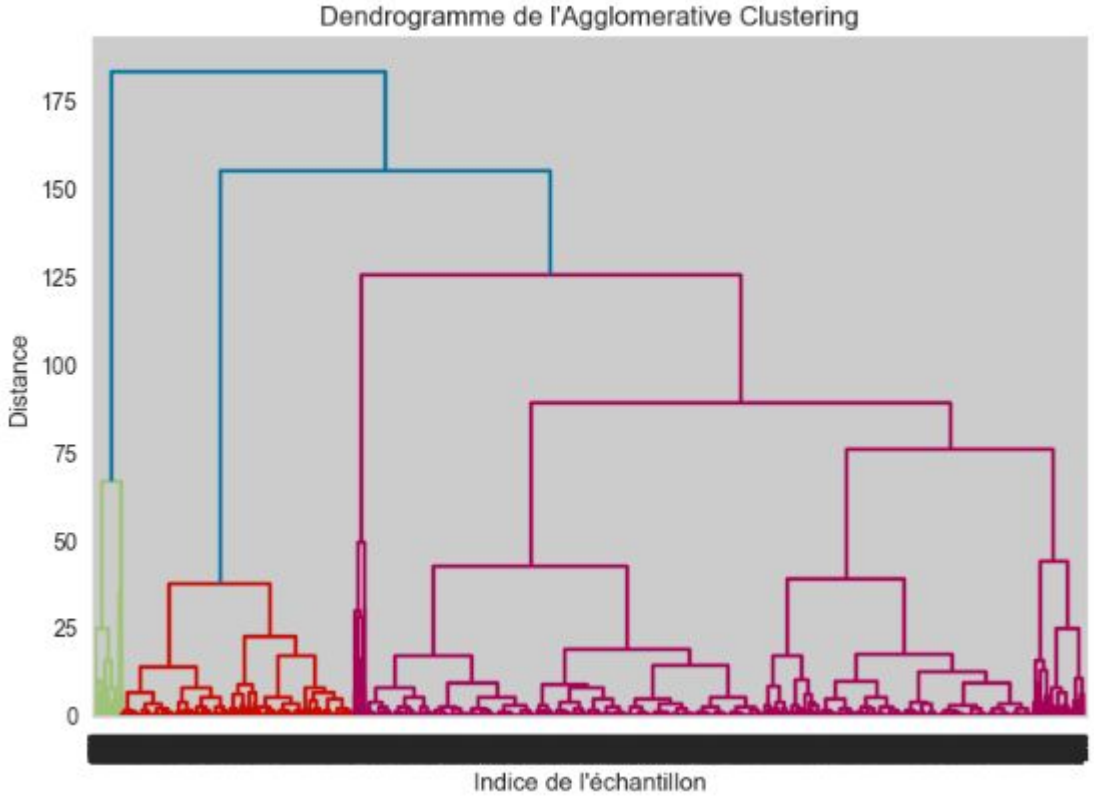


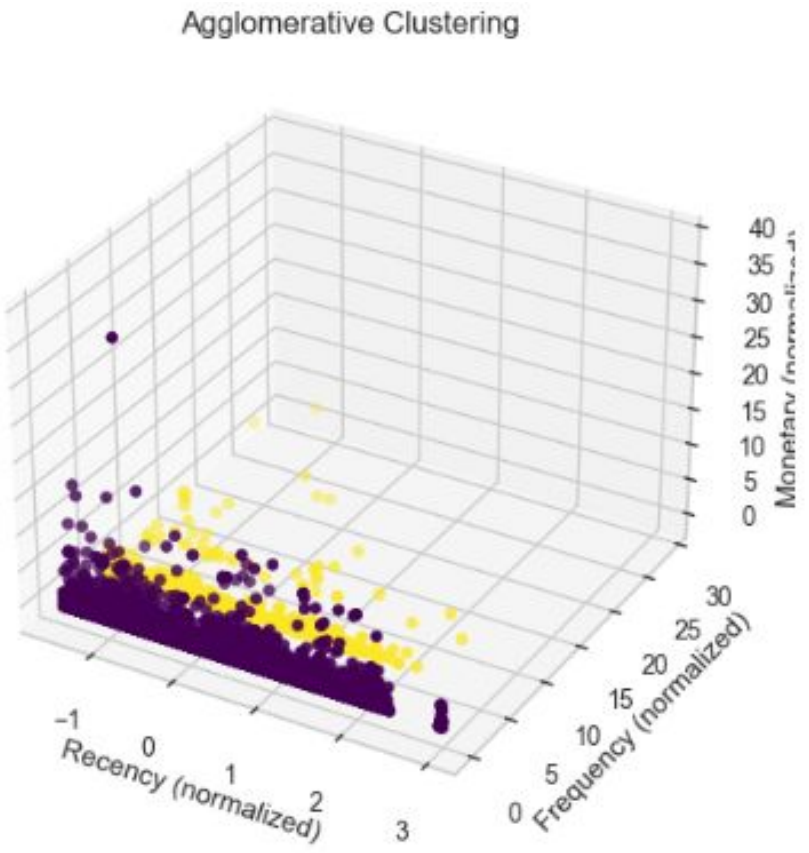


IV - Elaboration d'un modèle de clustering

3) Agglomerative Clustering

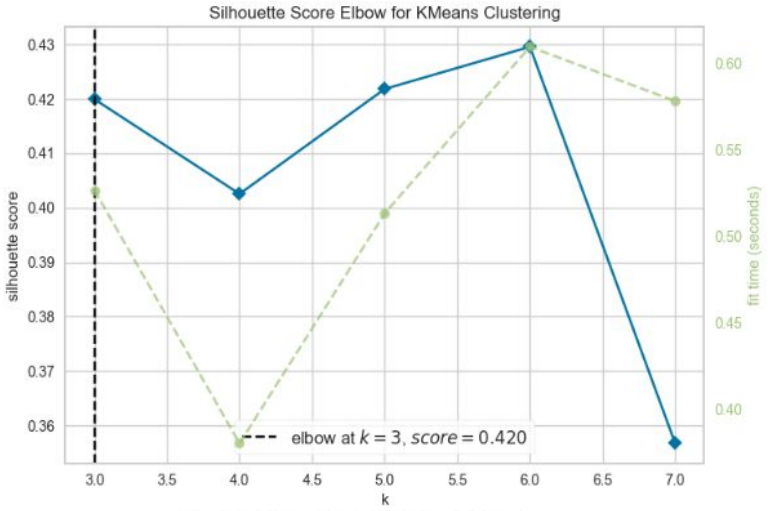
```
Cluster 0: 19429 points
Cluster 1: 571 points
```

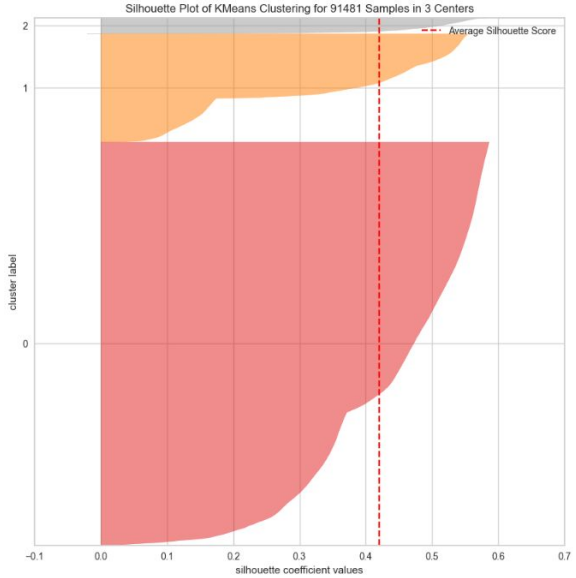


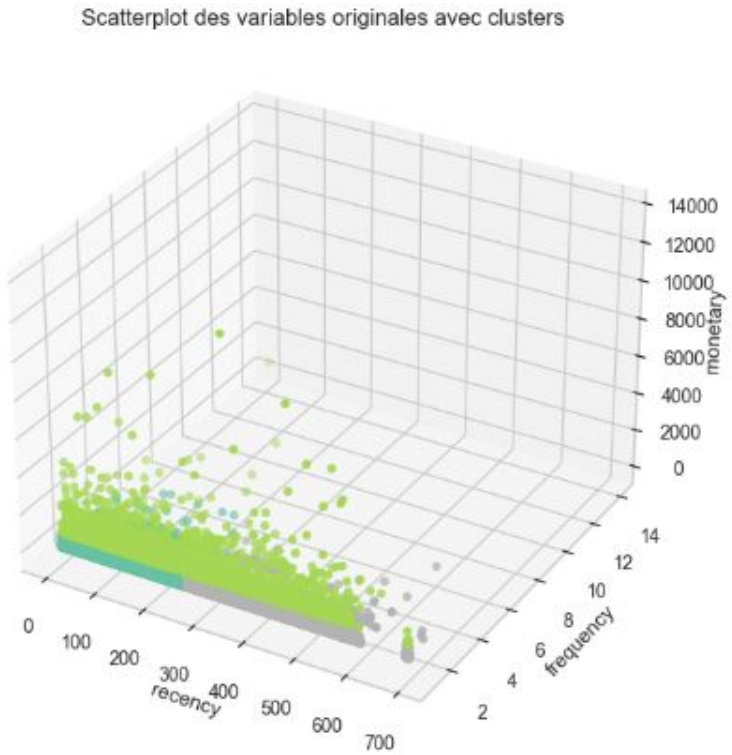


IV - Elaboration d'un modèle de clustering

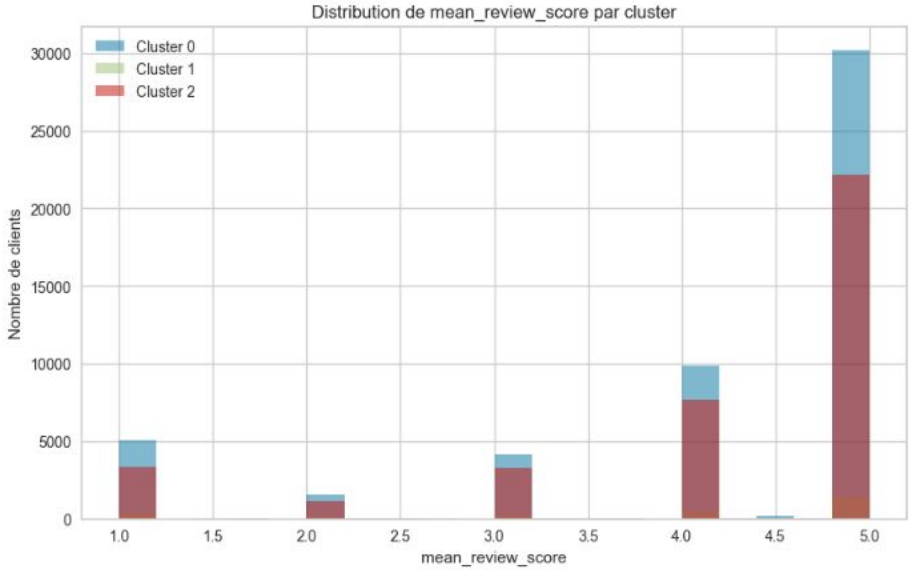
4) Ajout du review score moyen





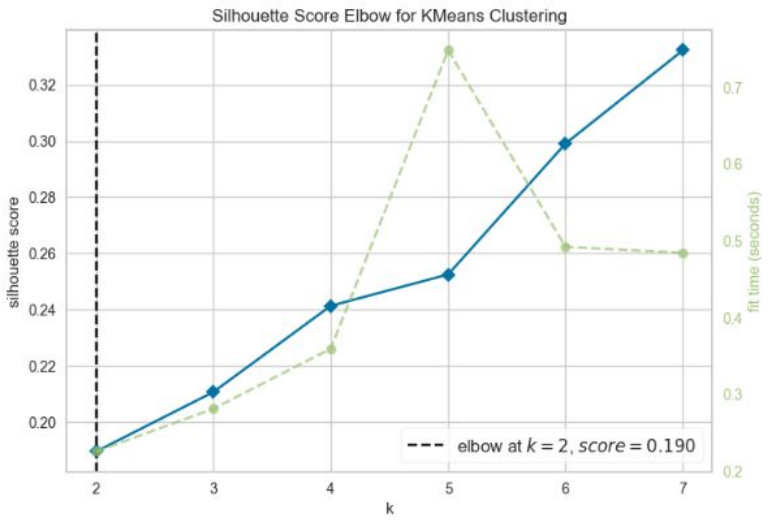


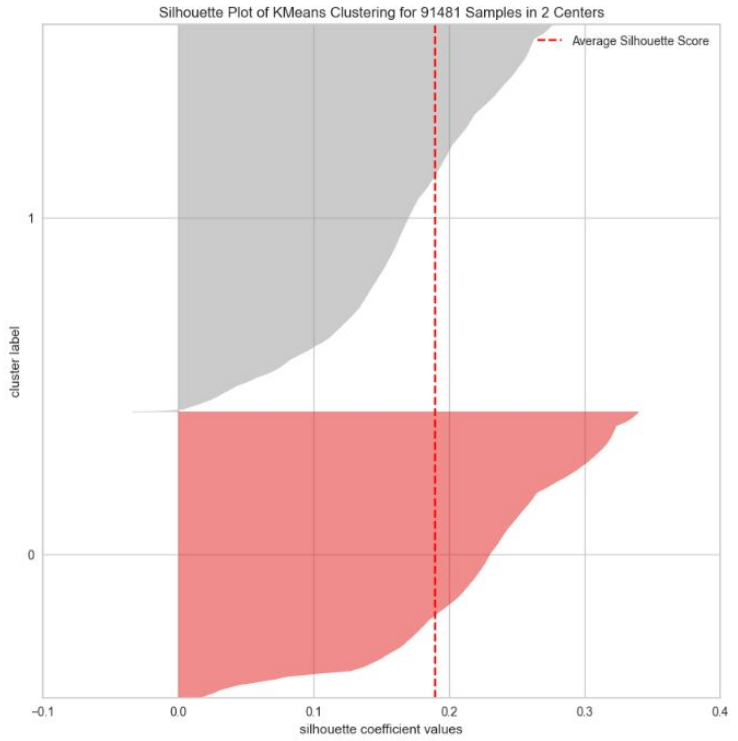

```
Nombre de clients par cluster :
cluster
0    51208
1     2503
2    37770
```

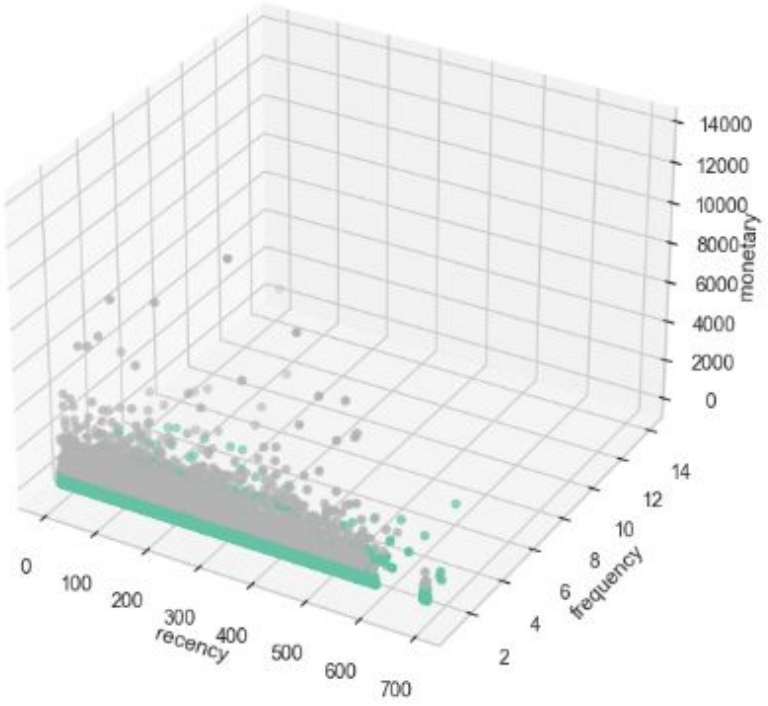


IV - Elaboration d'un modèle de clustering

5) Ajout des catégories de produits



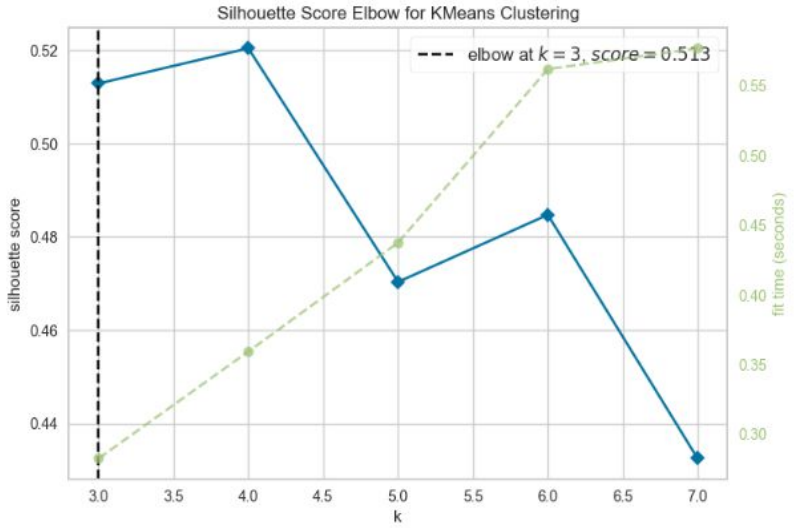


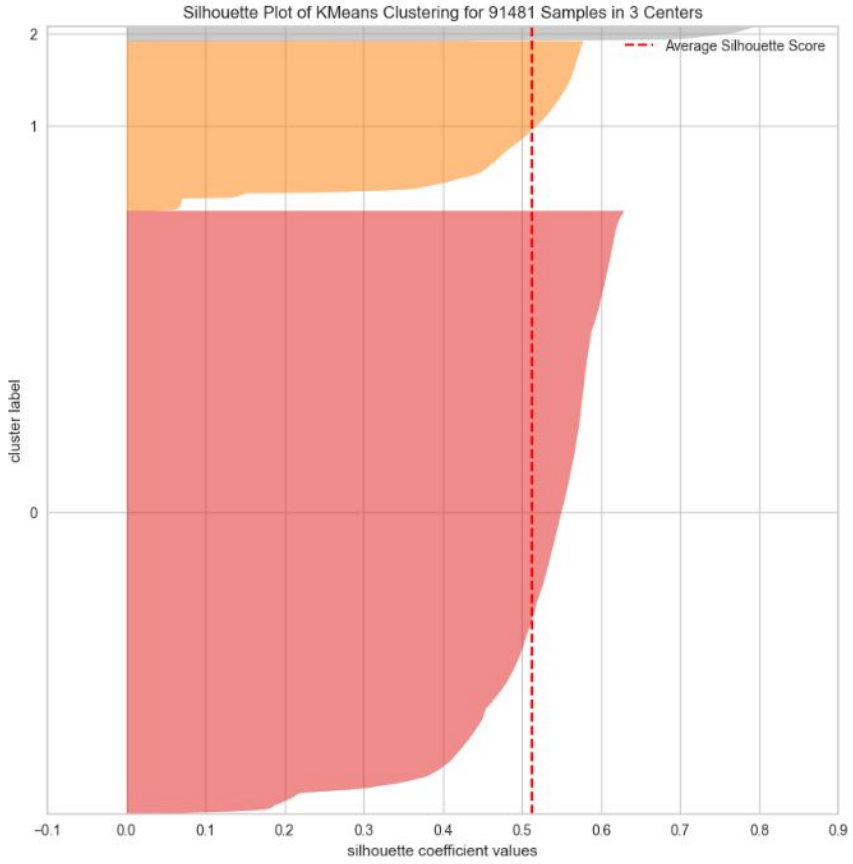


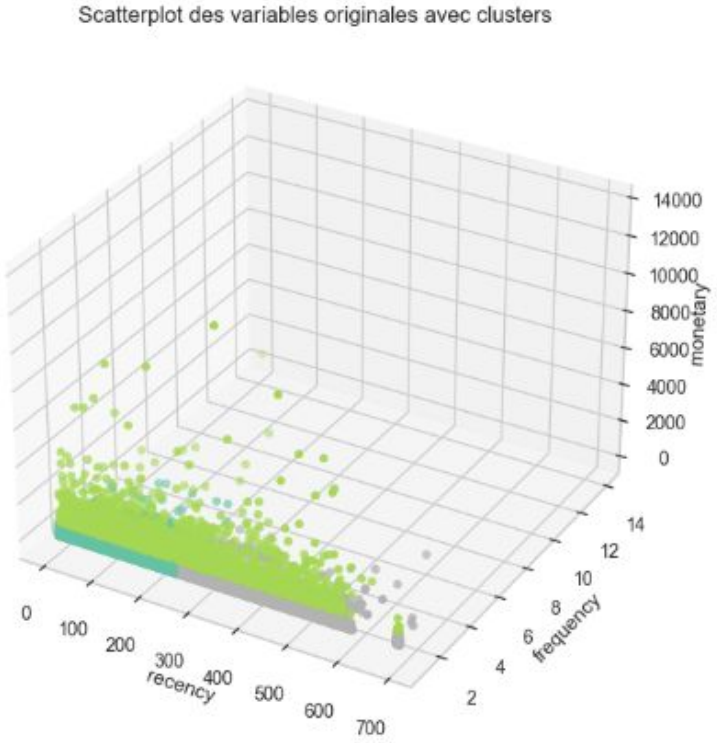

```
Nombre de clients par cluster :
cluster
0    88782
1     2699
Name: count, dtype: int64
```

IV - Elaboration d'un modèle de clustering

6) Ajout des variables de paiement








```
Nombre de clients par cluster :
cluster
0    51208
1     2503
2    37770
```

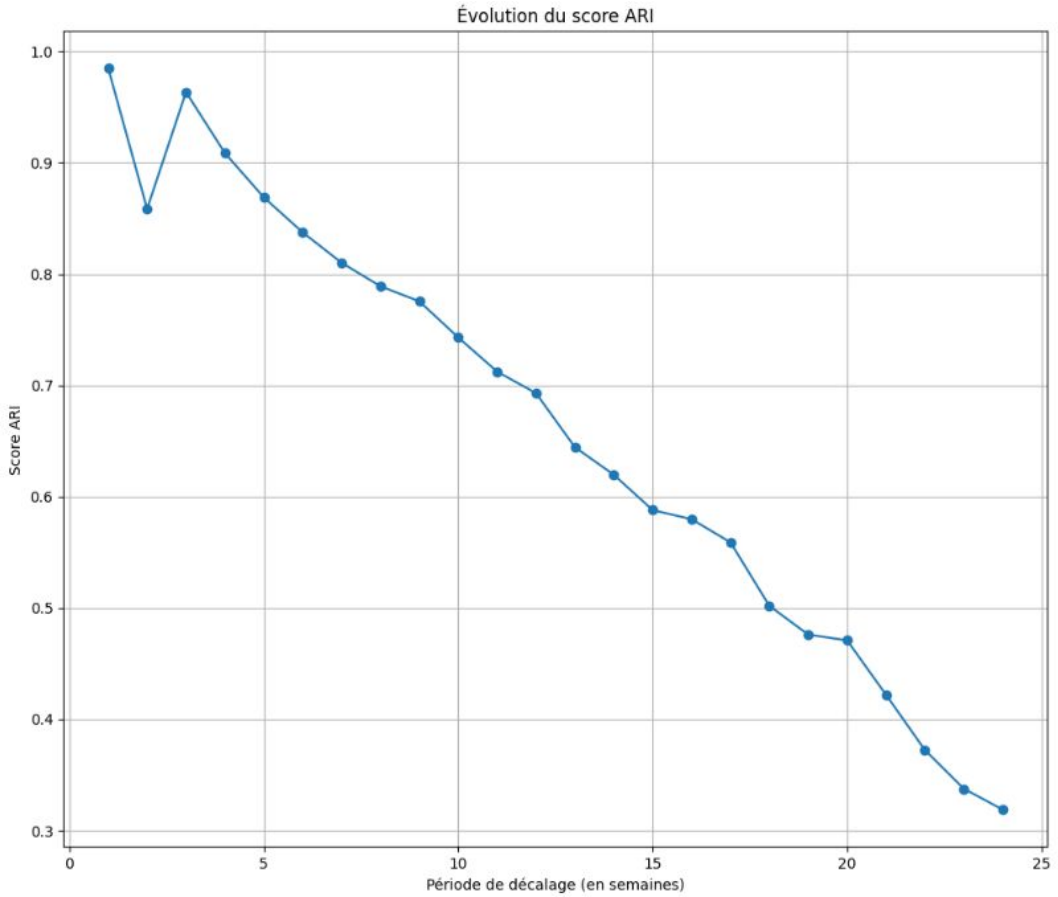
IV - Elaboration d'un modèle de clustering

7) Conclusion

- Trois cluster principaux :
  - Achat récent pour un faible montant
  - Achat ancien pour un faible montant (A REACTIVER)
  - Achat avec un montant élevé
- Ajout de variable n'affecte pas fondamentalement la segmentation des clients
- Piste pour la compréhension du comportement d'achat des clients et pour le ciblage de chaque segment

V - Simulation d'évolution de la stabilité du clustering dans le temps

- Entraîner modèle initial (M0) jusqu'au 12/2017
- Décalage par semaine des données et réentraînement de nouveaux modèles sur ces données décalées (M1)
- Calcul du score ARI à chaque décalage



Réentraînement du modèle nécessaire toutes les 7 semaines environ

VI- Conclusion

- Kmeans avec k = 3
- Trois cluster principaux :
  - Achat récent pour un faible montant
  - Achat ancien pour un faible montant (A REACTIVER)
  - Achat avec un montant élevé
- Variables Recency, Frequency, Monetary
- Modèle à réentraîner toutes les 7 semaines environ

Merci pour votre attention