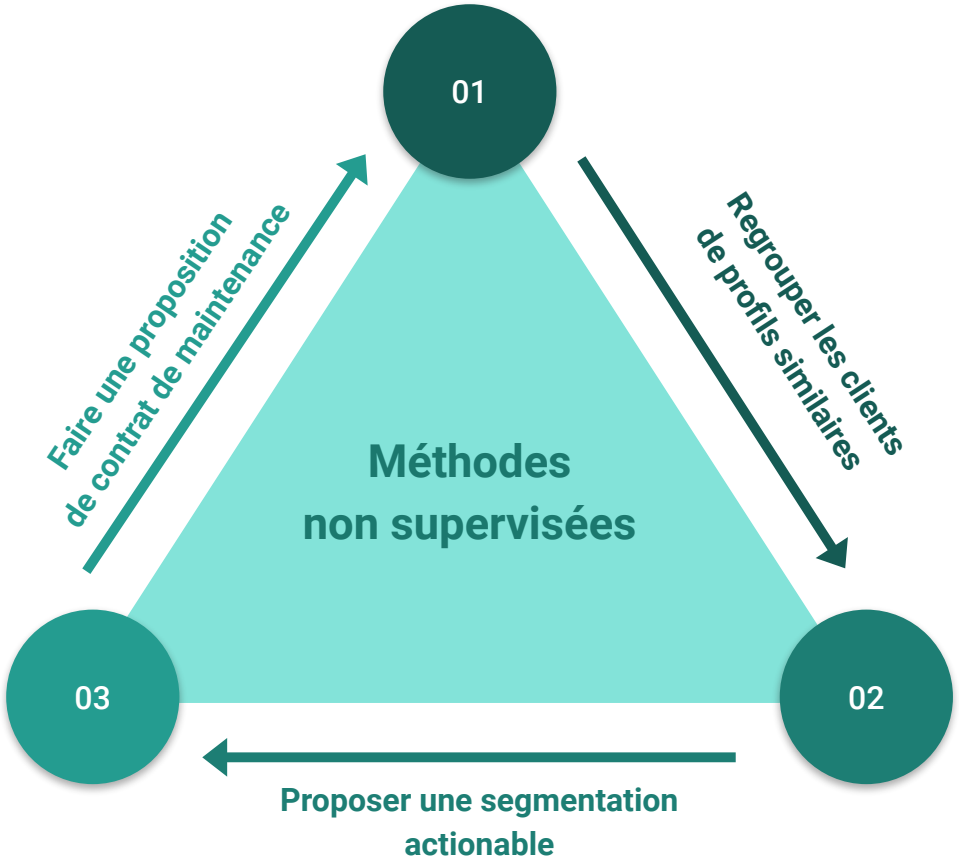
Segmentation des clients du site e-commerce OLIST

P5 - Sylvia BANKOWSKA - Juin 2022

Sommaire

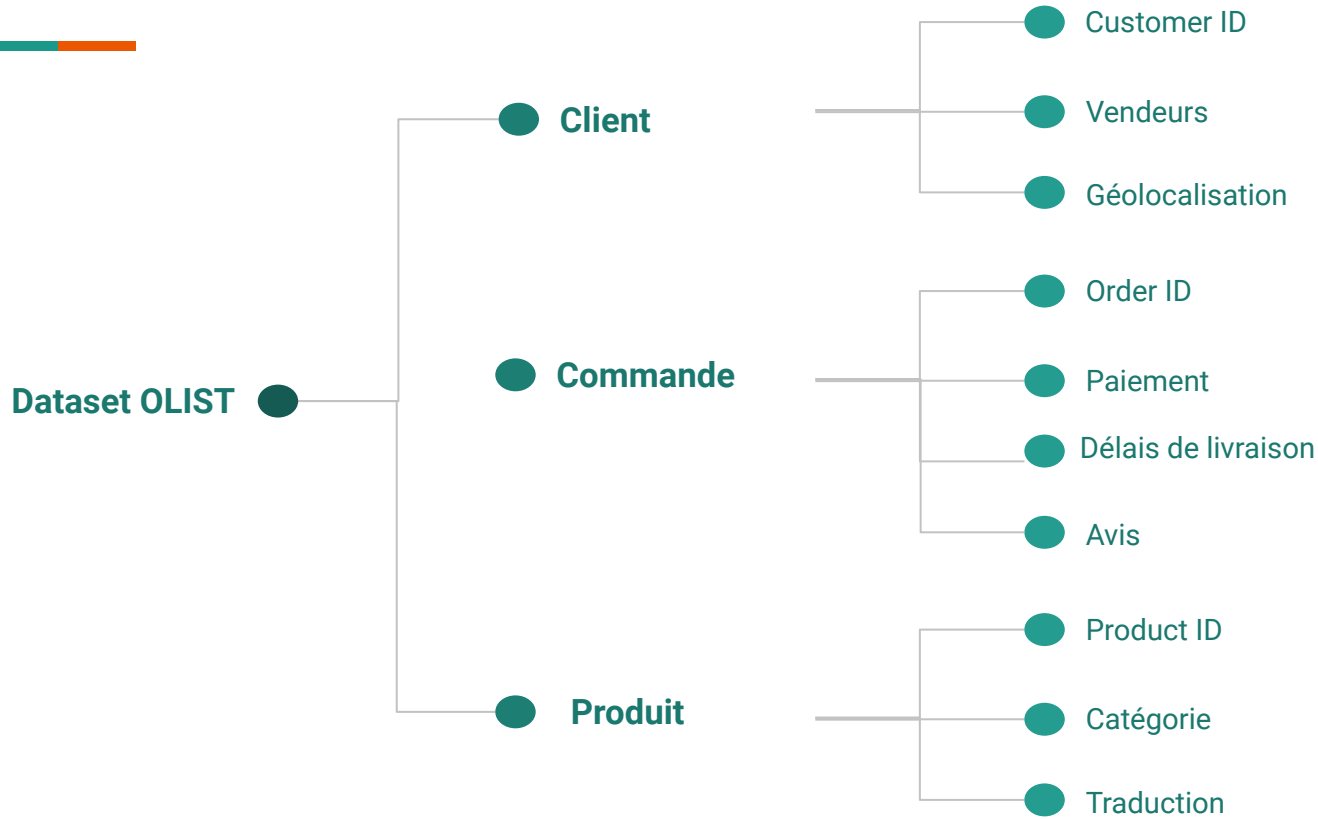
Enjeux

Fournir aux employés OLIST **un outil de segmentation des clients**



Analyse exploratoire

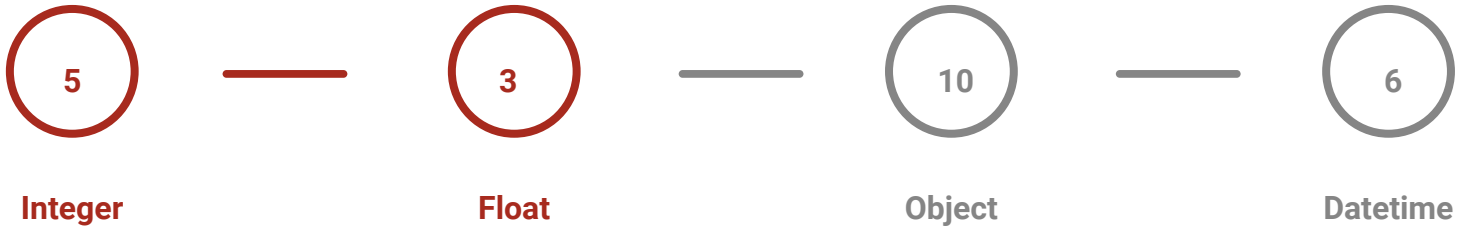
Base de données



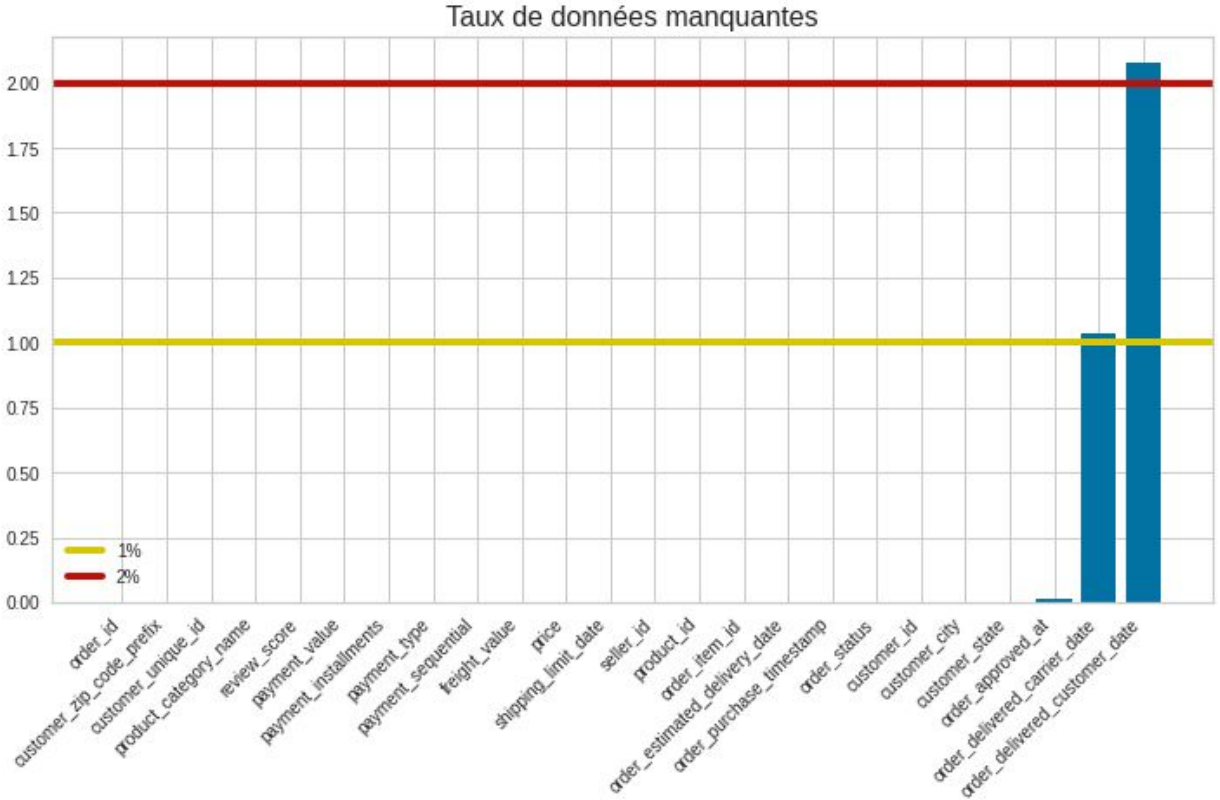
Jeu de données final

df_commande + df_client + df_produit ⇨ **df_final**

- 115 609 lignes
- 24 variables



Valeurs manquantes



Traitement des NaN

NaN avant

order_purchase_timestamp	0
order_approved_at	14
order_delivered_carrier_date	1195
order_delivered_customer_date	2400
order_estimated_delivery_date	0

NaN vs 'order_status'

shipped	1138
canceled	529
invoiced	358
processing	357
delivered	8
unavailable	7
approved	3

'order_status'

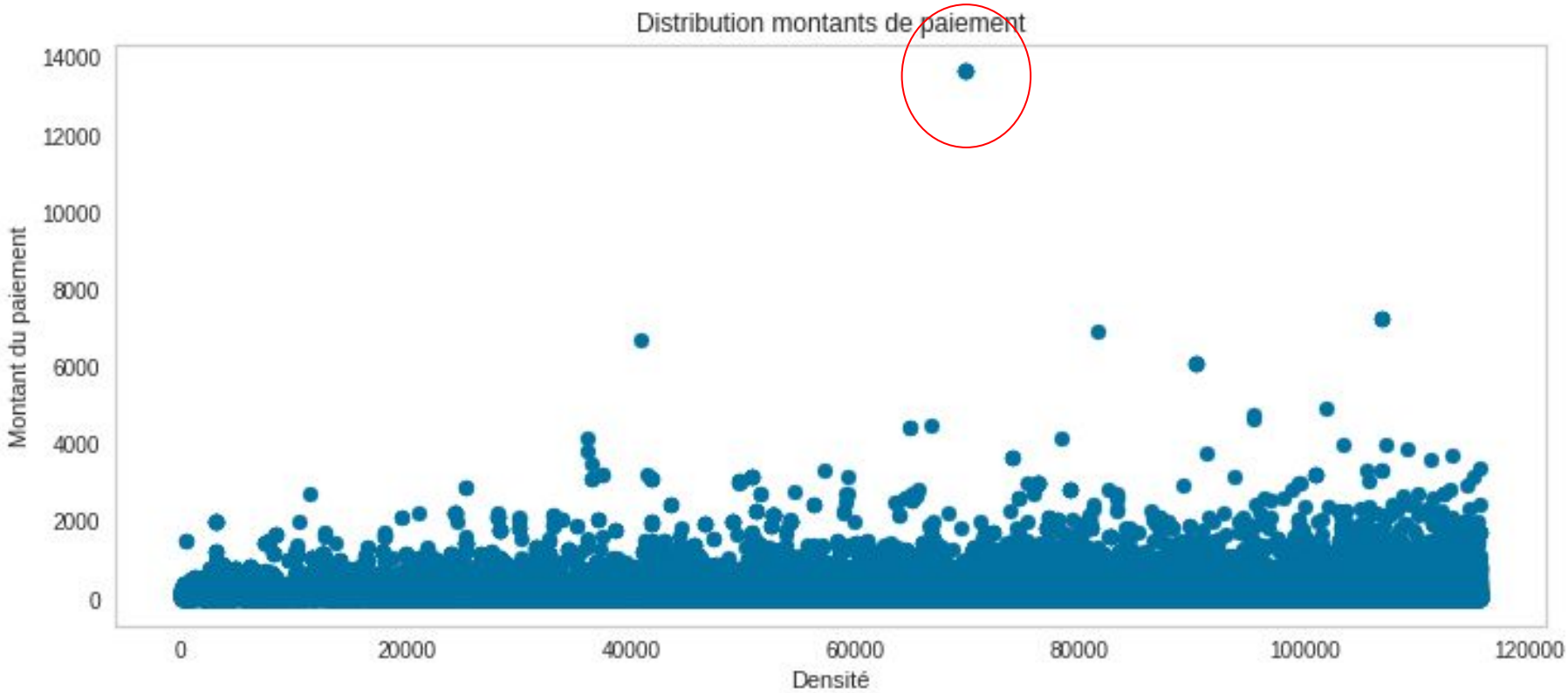
delivered	0.979
shipped	0.010
canceled	0.005
invoiced	0.003
processing	0.003
unavailable	0.000
approved	0.000

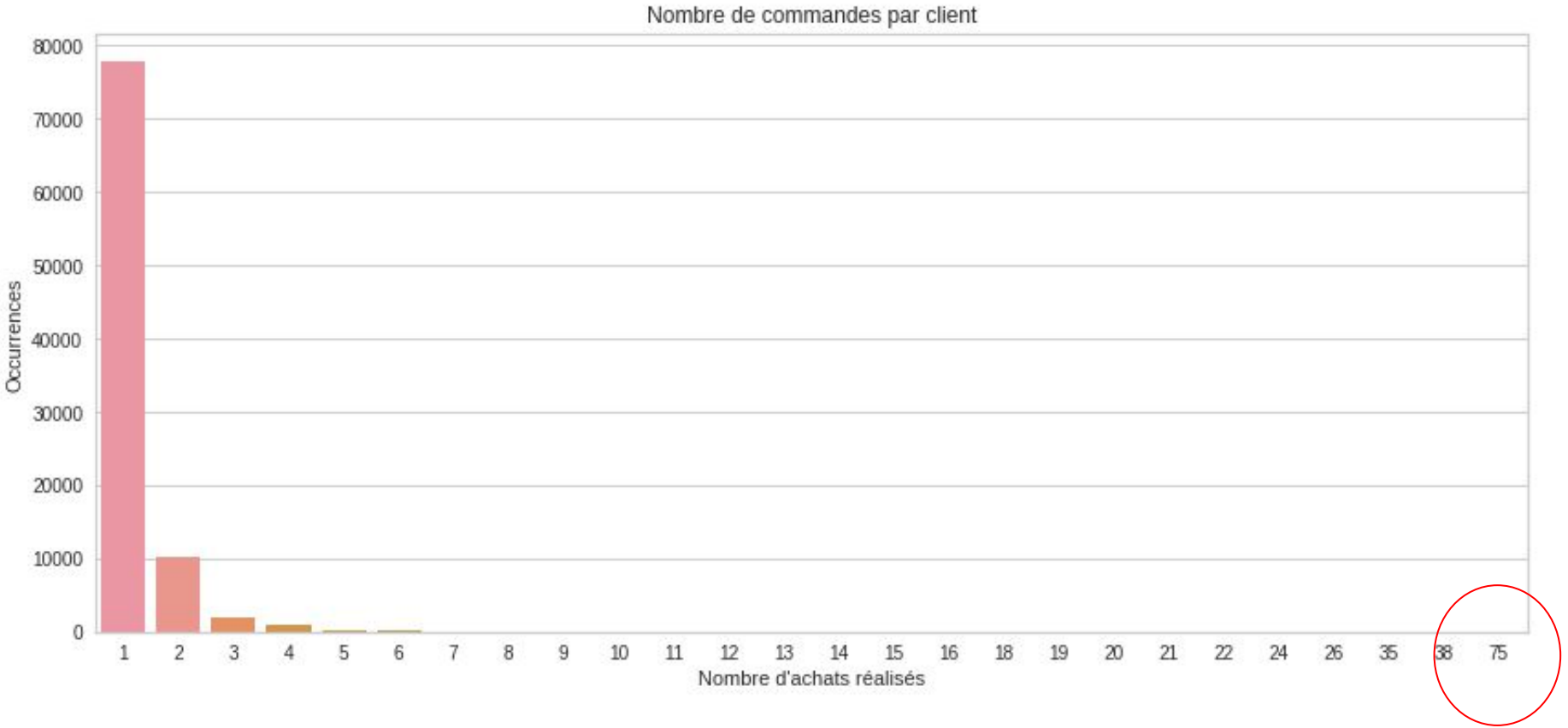
drop

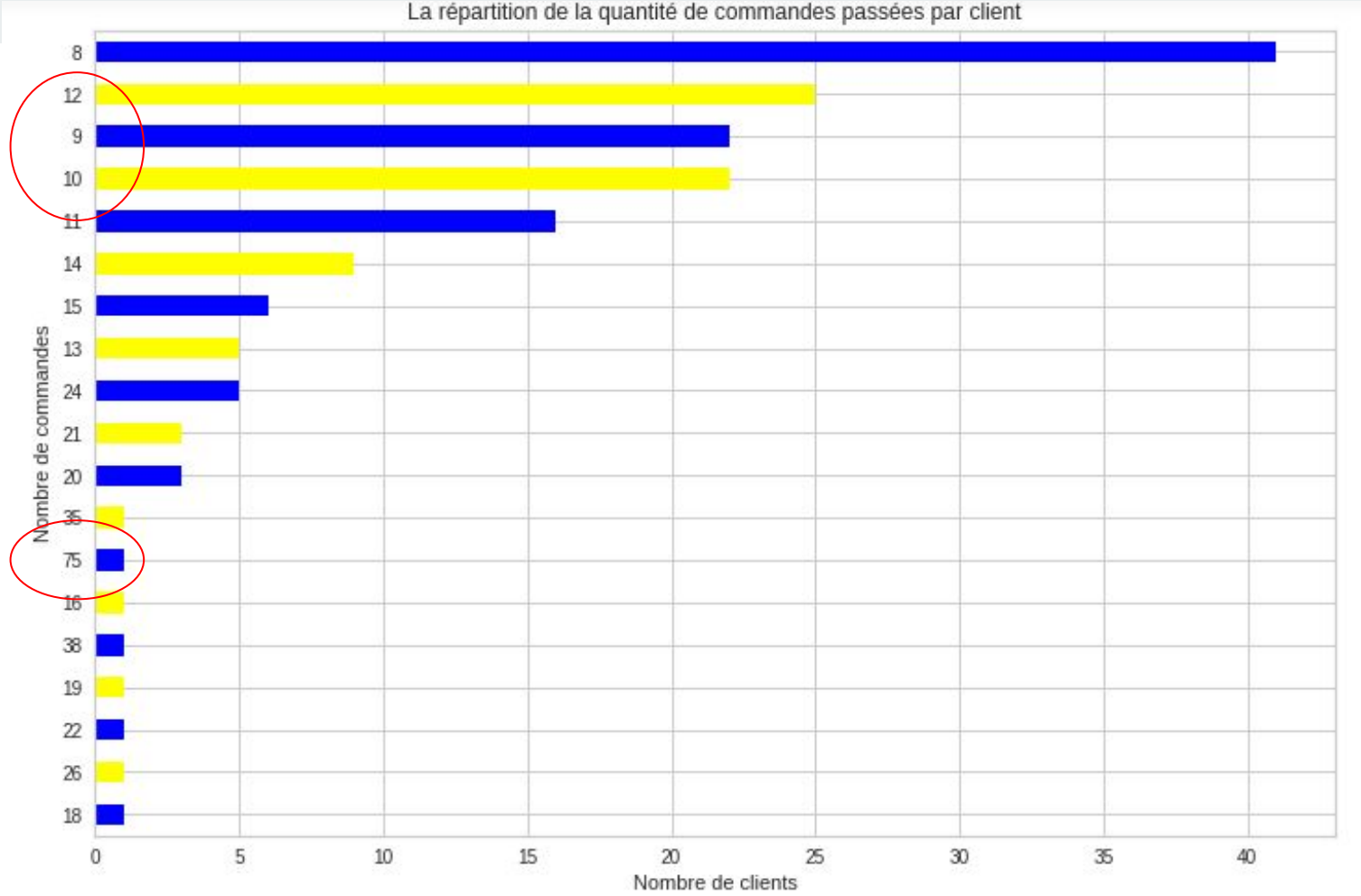
NaN après

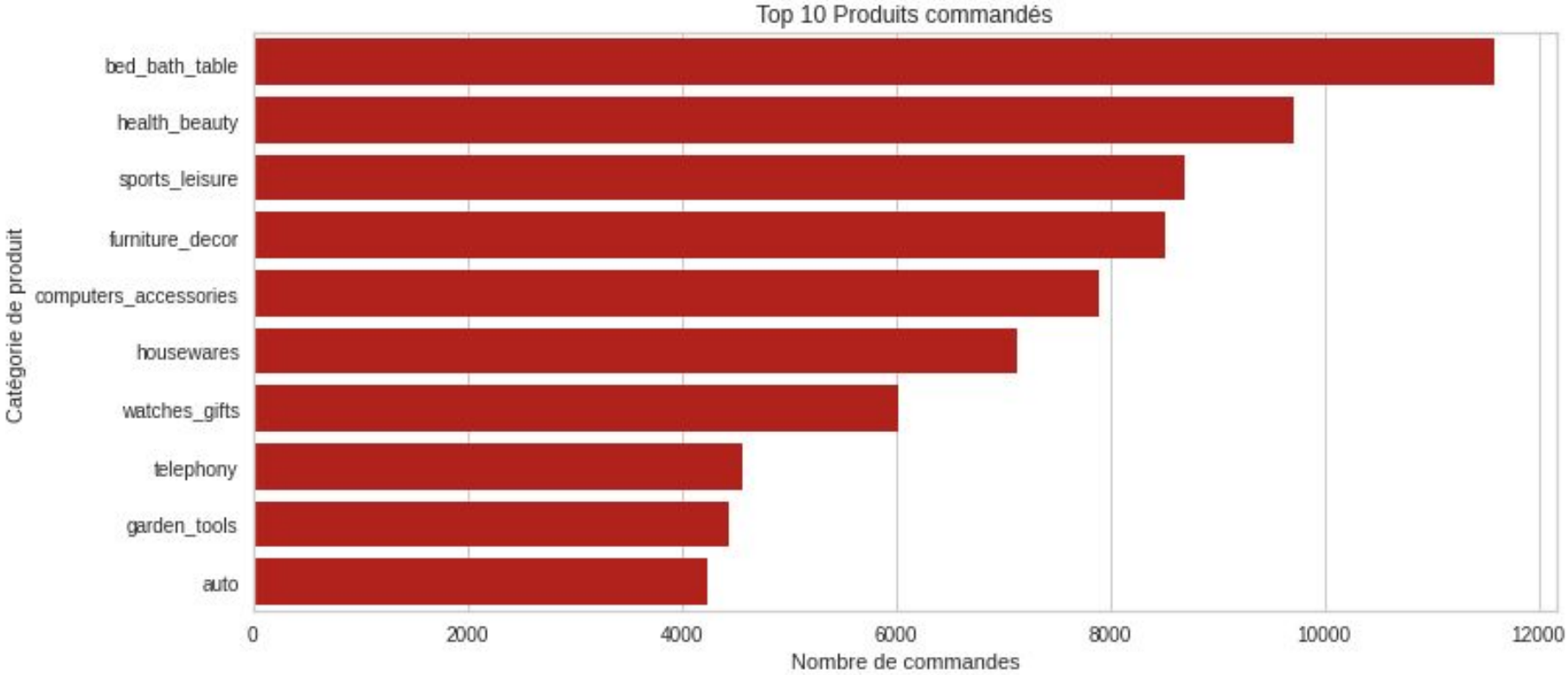
order_status	0
order_purchase_timestamp	0
order_approved_at	14
order_delivered_carrier_date	2
order_delivered_customer_date	8

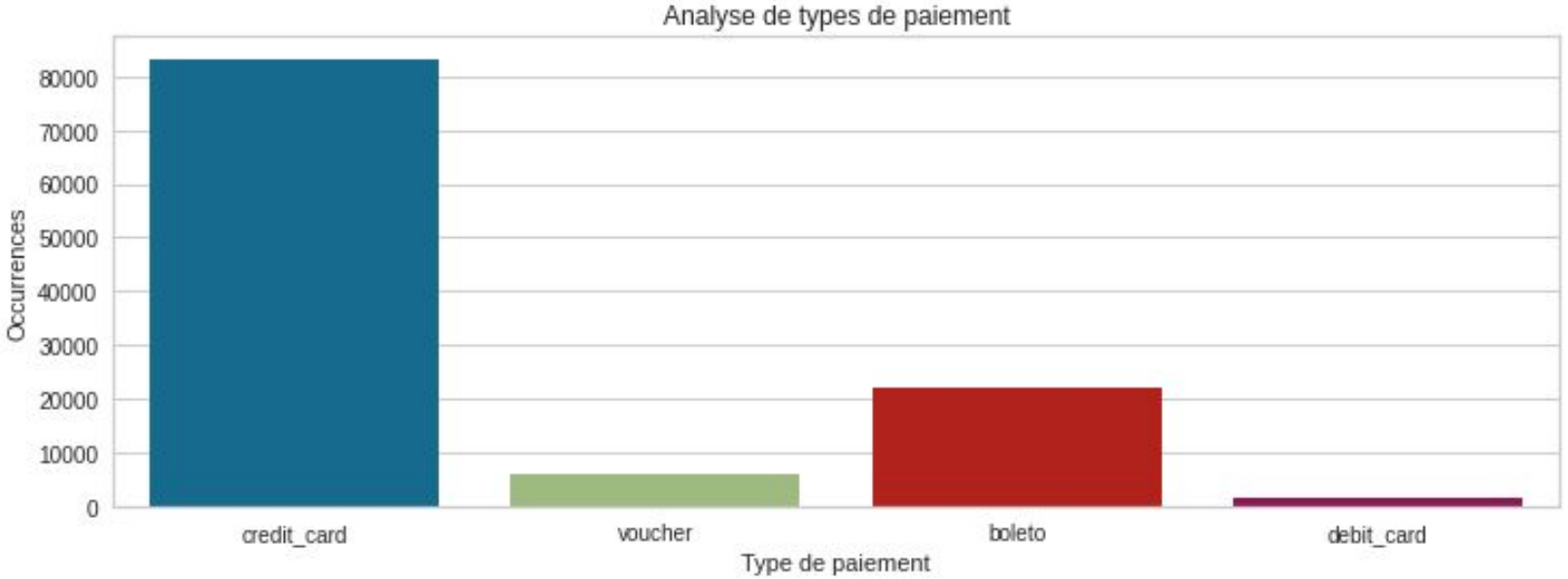
Exploration des variables

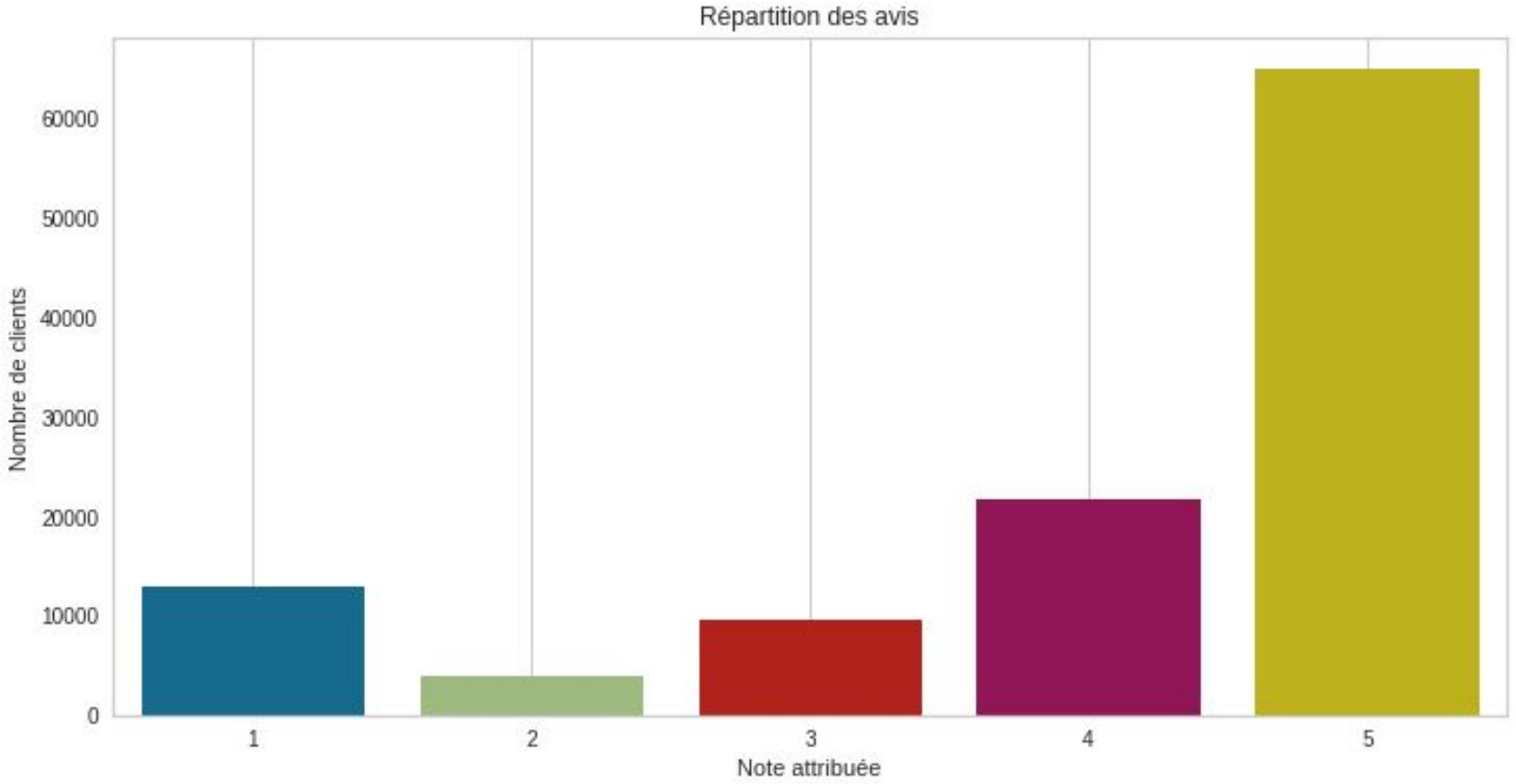

Distribution montants de paiement
Montant du paiement
14000
12000
10000
8000
6000
4000
2000
0
0
20000
40000
60000
80000
100000
120000
Densité

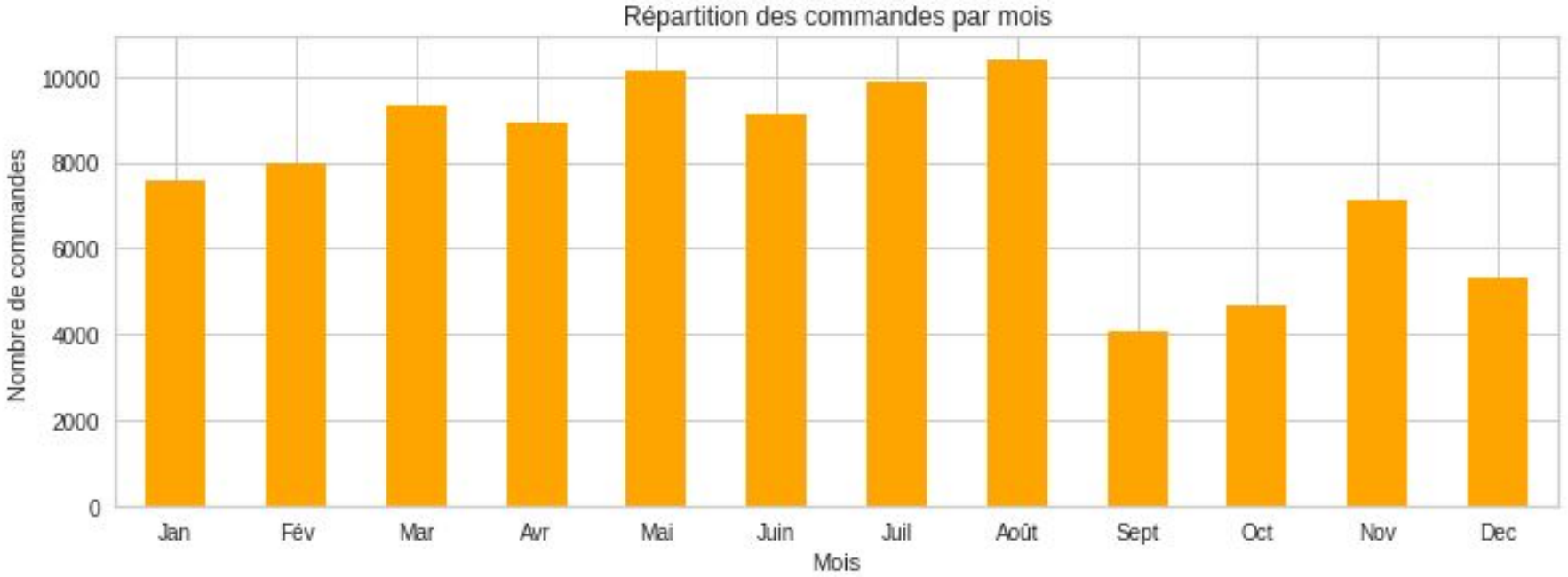

Nombre de commandes par client
Occurrences
80000
70000
60000
50000
40000
30000
20000
10000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
18
19
20
21
22
24
26
35
38
75
Nombre d'achats réalisés

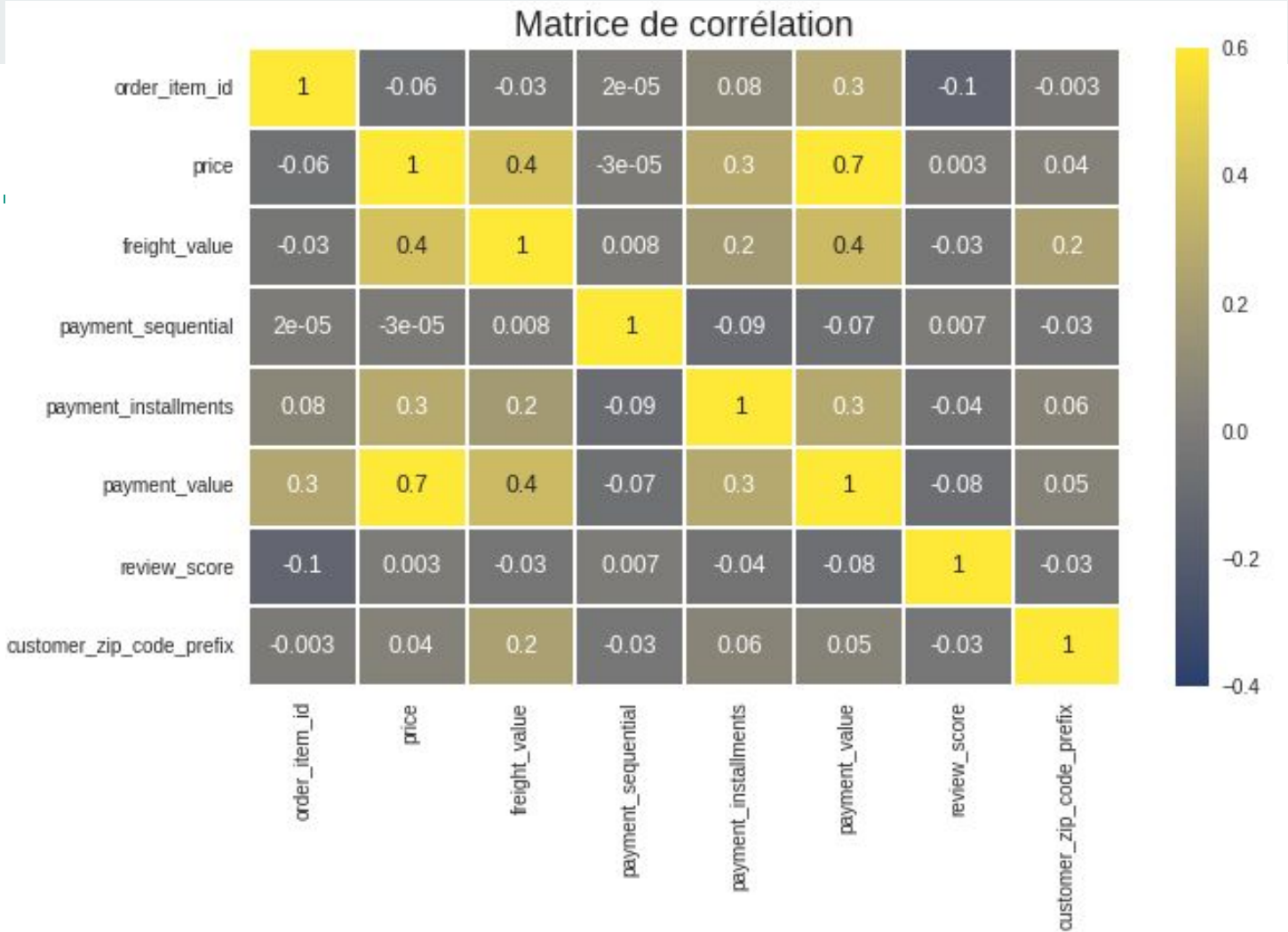

La répartition de la quantité de commandes passées par client
Nombre de commandes
8
12
9
10
11
14
15
13
24
21
20
35
75
16
38
19
22
26
18
0
5
10
15
20
25
30
35
40
Nombre de clients


Top 10 Produits commandés
bed_bath_table
health_beauty
sports_leisure
furniture_decor
computers_accessories
housewares
watches_gifts
telephony
garden_tools
auto
Catégorie de produit
0
2000
4000
6000
8000
10000
12000
Nombre de commandes


Analyse de types de paiement
Occurrences
80000
70000
60000
50000
40000
30000
20000
10000
0
credit_card
voucher
boleto
debit_card
Type de paiement


Répartition des avis
Nombre de clients
60000
50000
40000
30000
20000
10000
0
1
2
3
4
5
Note attribuée

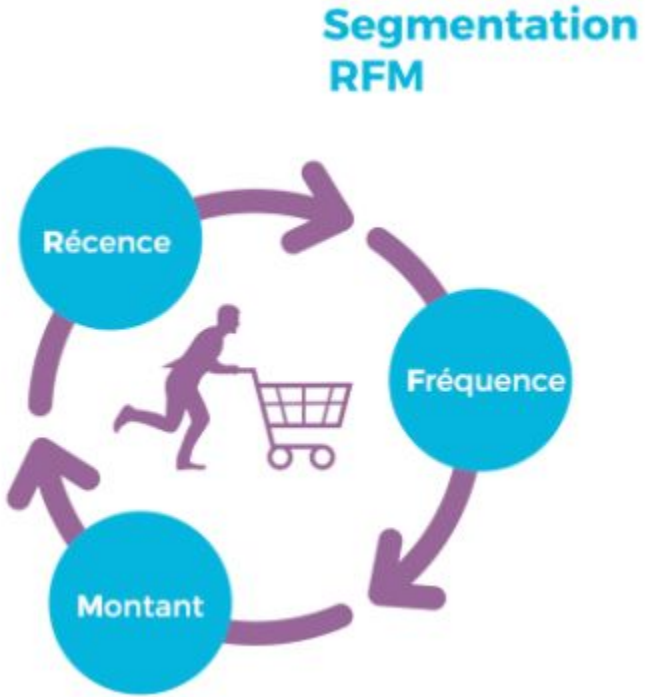

Répartition des commandes par mois
Nombre de commandes
0
2000
4000
6000
8000
10000
Jan
Fév
Mar
Avr
Mai
Juin
Juil
Août
Sept
Oct
Nov
Dec
Mois


Matrice de corrélation
order_item_id 1 -0.06 -0.03 2e-05 0.08 0.3 -0.1 -0.003
price -0.06 1 0.4 -3e-05 0.3 0.7 0.003 0.04
freight_value -0.03 0.4 1 0.008 0.2 0.4 -0.03 0.2
payment_sequential 2e-05 -3e-05 0.008 1 -0.09 -0.07 0.007 -0.03
payment_installments 0.08 0.3 0.2 -0.09 1 0.3 -0.04 0.06
payment_value 0.3 0.7 0.4 -0.07 0.3 1 -0.08 0.05
review_score -0.1 0.003 -0.03 0.007 -0.04 -0.08 1 -0.03
customer_zip_code_prefix -0.003 0.04 0.2 -0.03 0.06 0.05 -0.03 1
order_item_id
price
freight_value
payment_sequential
payment_installments
payment_value
review_score
customer_zip_code_prefix
0.6
0.4
0.2
0.0
−0.2
−0.4

Pistes de modélisation

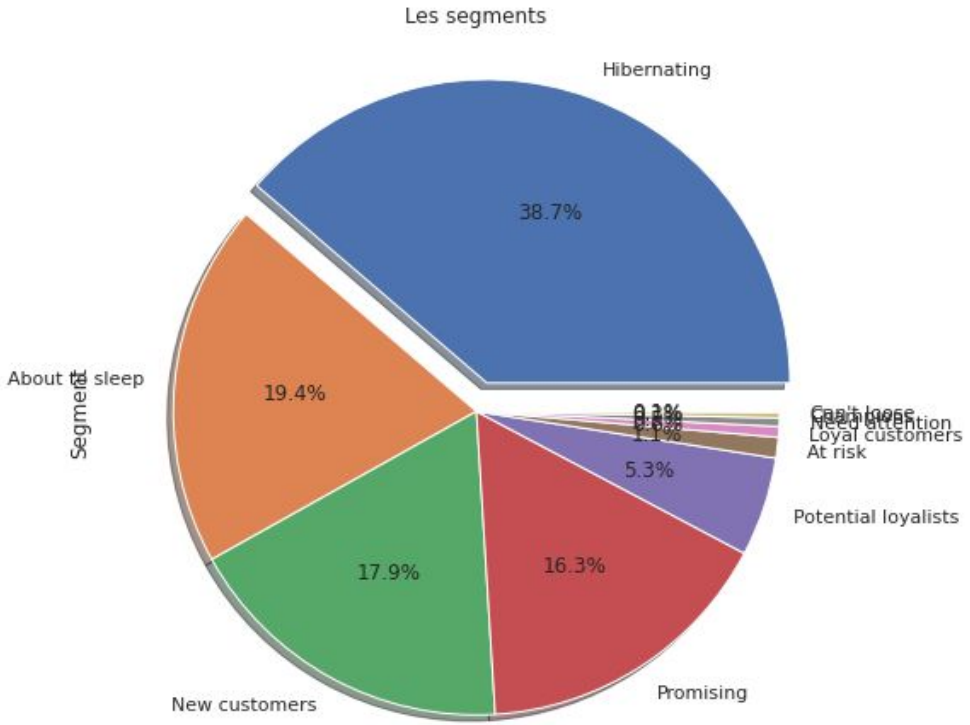
Segmentation RFM

Récence	Ceux qui ont acheté le plus récemment
Fréquence	Ceux qui ont acheté le plus fréquemment
Montant	Ceux qui dépensent le plus



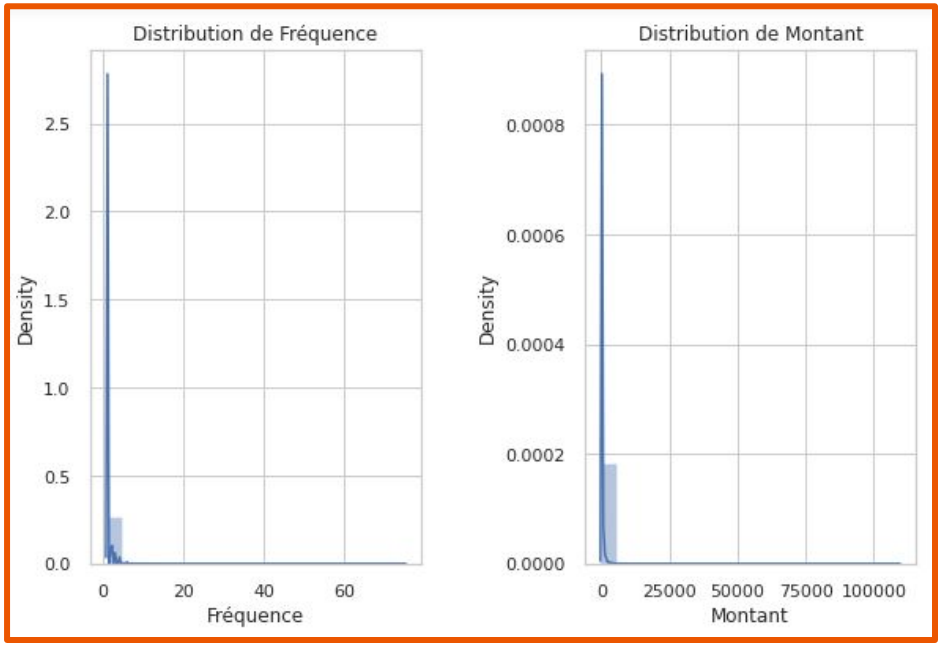
Profils clients

Segment	Récence mean	Fréquence mean	Montant mean	Montant count
About to sleep	122.500	1.100	165.100	7865
At risk	237.900	3.700	501.200	447
Can't loose	224.400	8.500	413.200	21
Champions	26.100	4.800	567.500	117
Hibernating	236.900	1.100	159.600	15723
Loyal customers	91.800	4.700	596.900	232
Need attention	122.300	3.000	496.200	183
New customers	24.100	1.000	137.200	7261
Potential loyalists	43.700	2.200	342.100	2144
Promising	62.600	1.000	149.400	6604

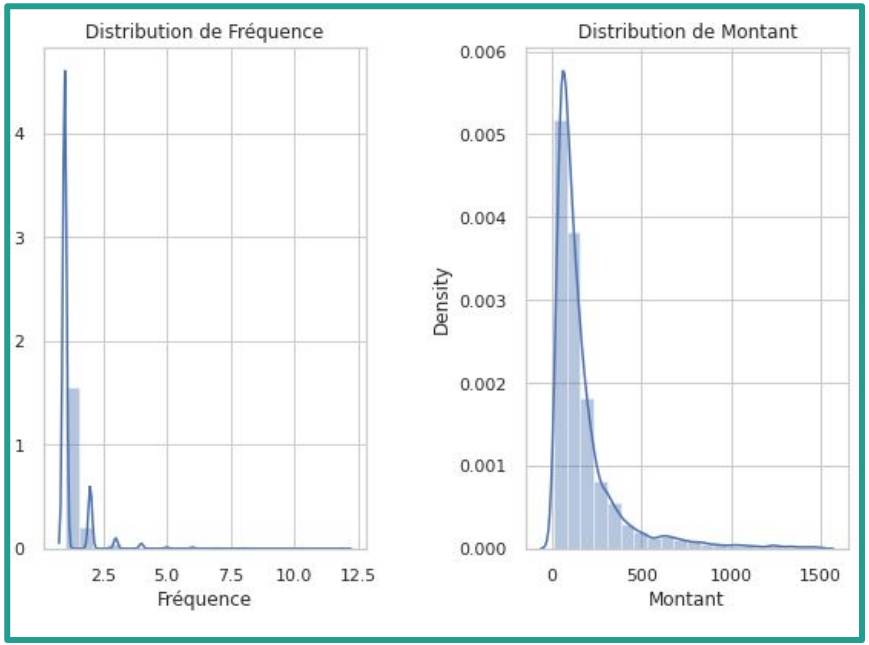


Outliers

AVANT

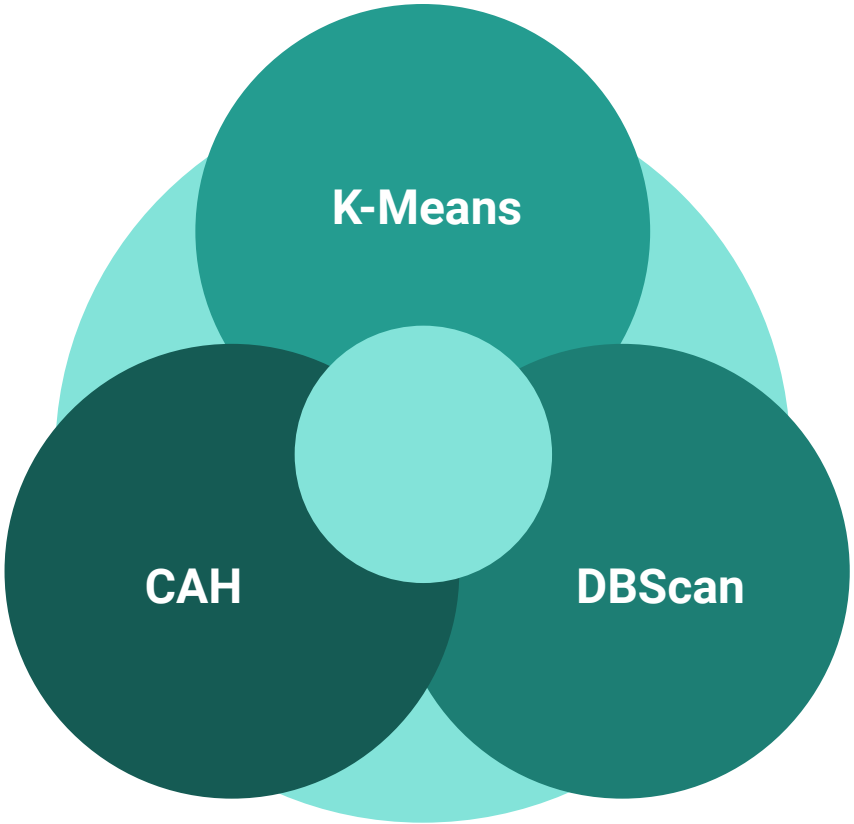


APRES

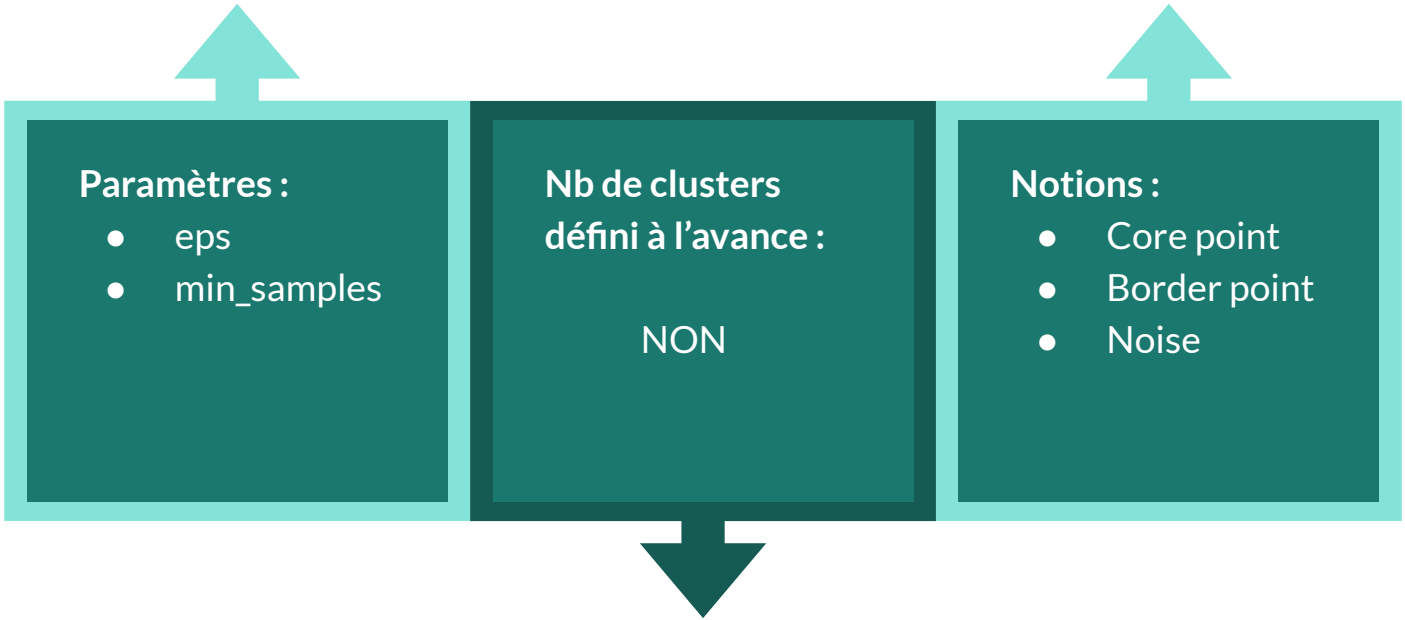


Clustering

Algorithmes utilisés

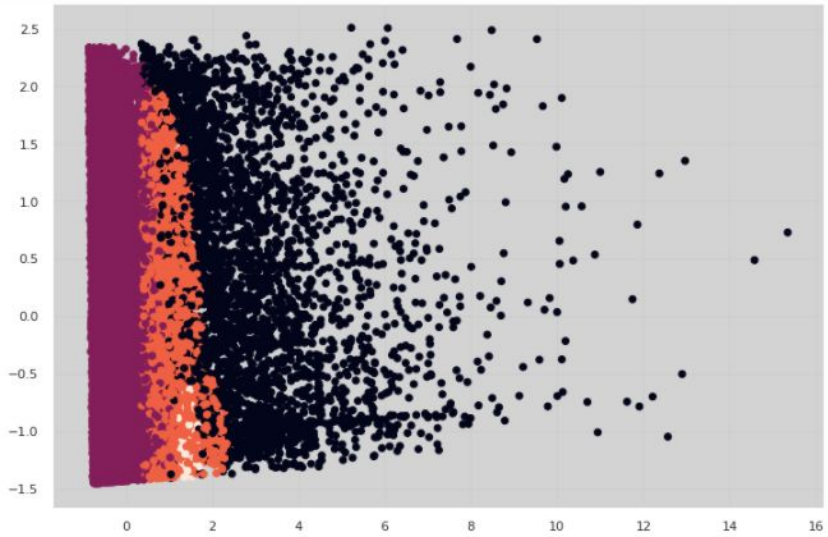


DBScan


Paramètres :
eps
min_samples
Nb de clusters défini à l'avance :
NON
Notions :
Core point
Border point
Noise

Essai n° 1 :

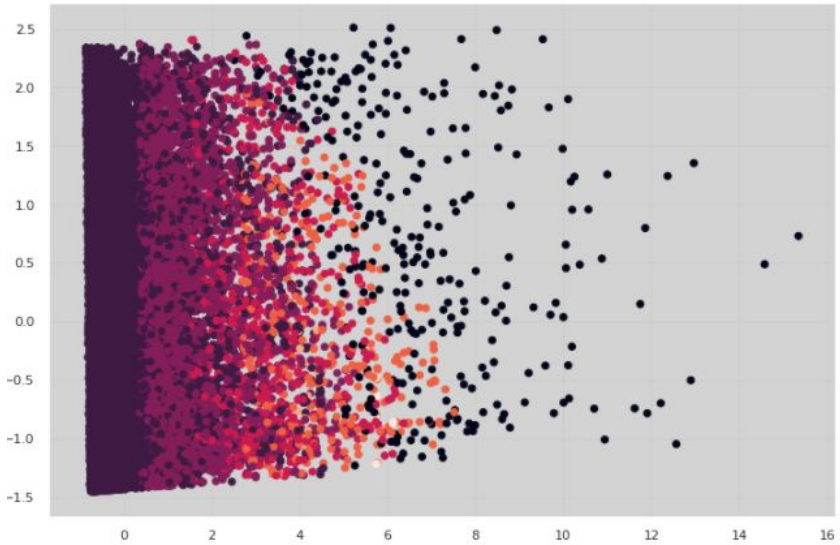
- min_samples = 100
- eps = 0.3, 0.4, 0.5



- ➢ eps=0.3 : 3 clusters et 3549 anomalies
- ➢ eps=0.4 : 2 clusters et 2759 anomalies
- ➢ eps=0.5 : 2 clusters et 2311 anomalies

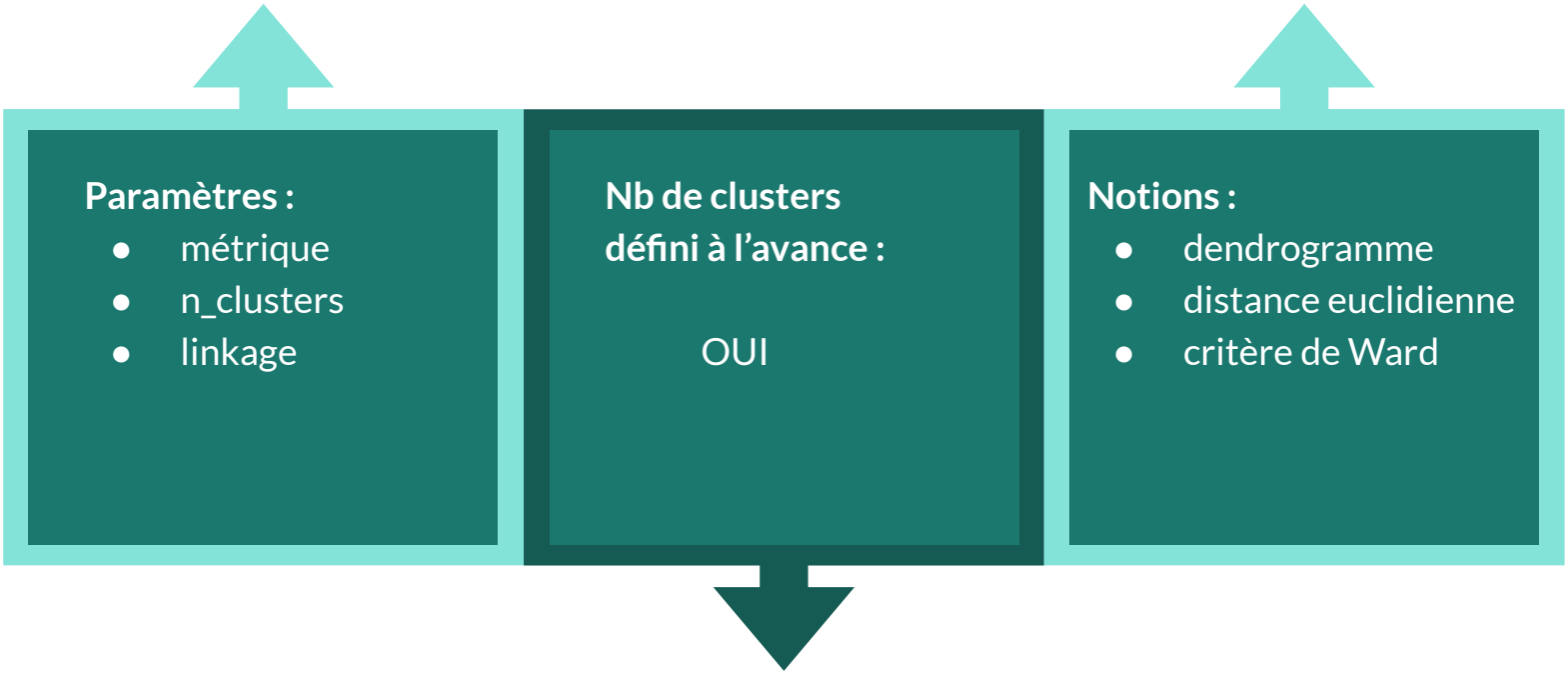
Essai n° 2 :

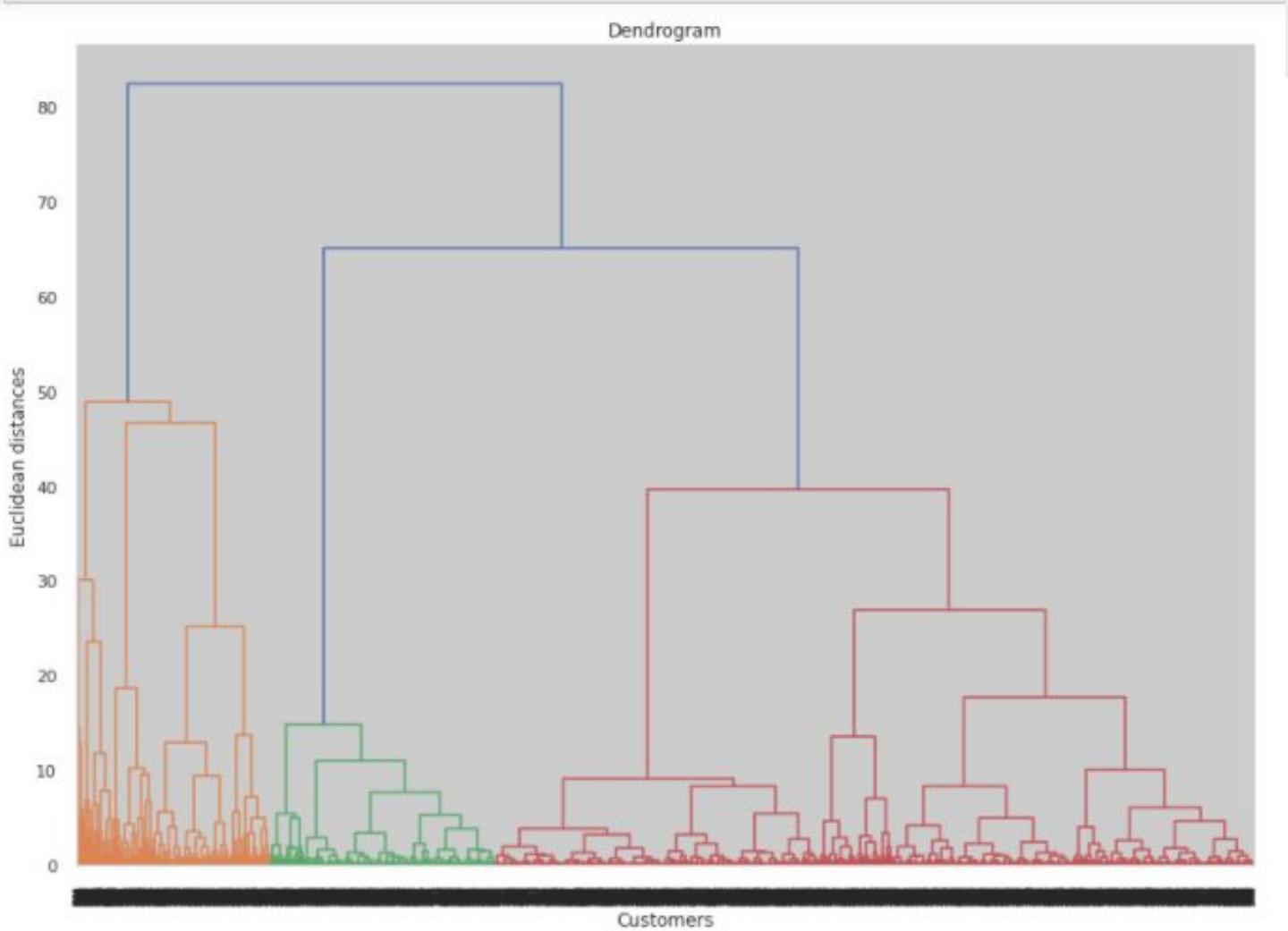
- min_samples = 10
- eps = 0.3, 0.4, 0.5

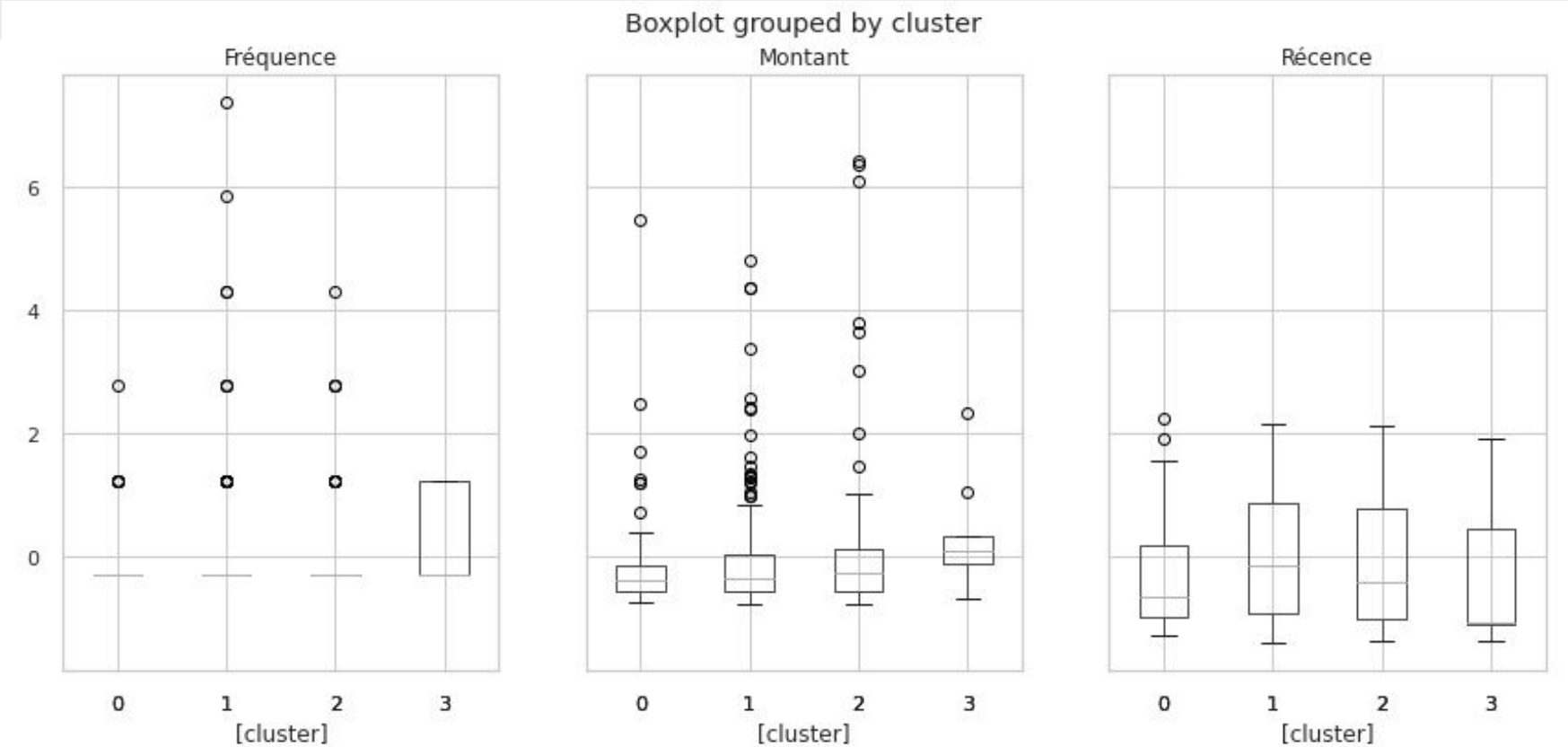


- ➢ eps=0.3 : 17 clusters et 854 anomalies
- ➢ eps=0.4 : 9 clusters et 450 anomalies
- ➢ eps=0.5 : 6 clusters et 298 anomalies

Classification Ascendante Hiérarchique

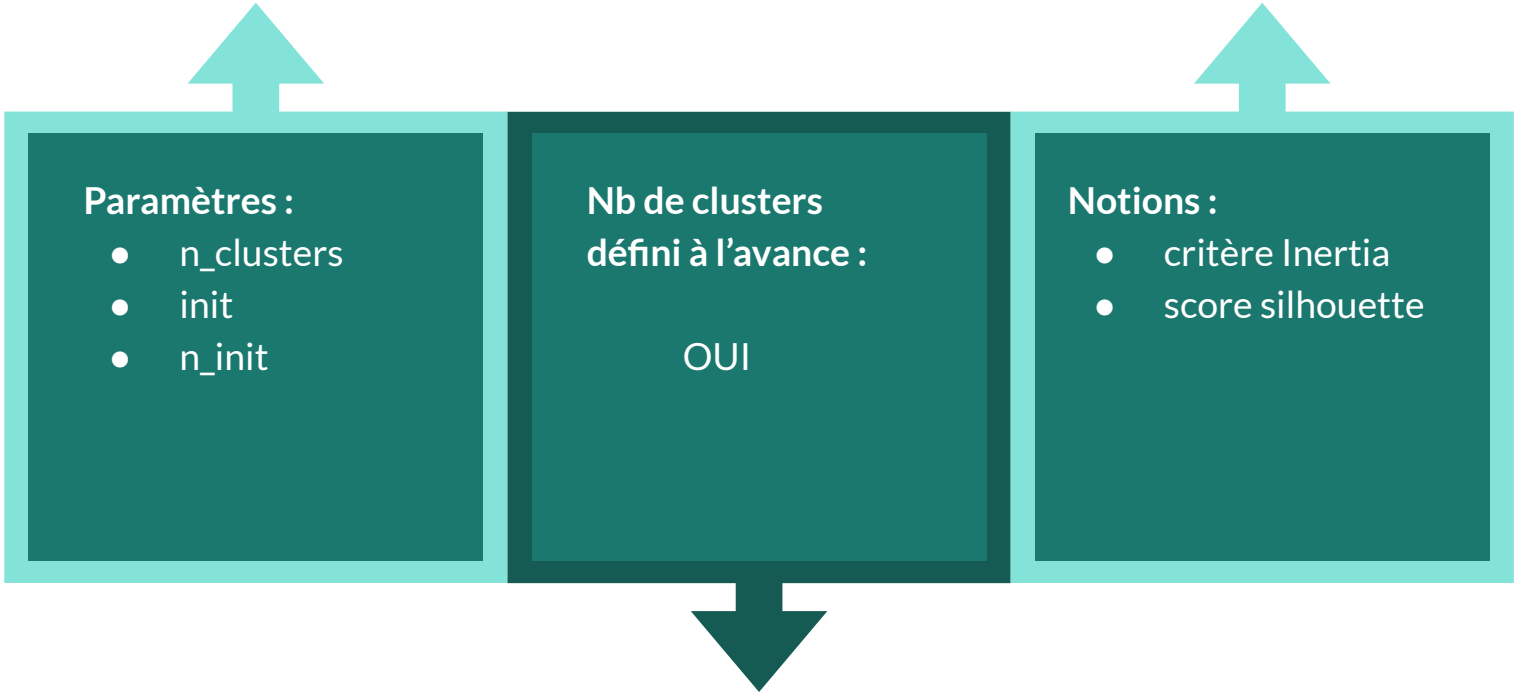



Dendrogram
Euclidean distances
0
10
20
30
40
50
60
70
80
Customers



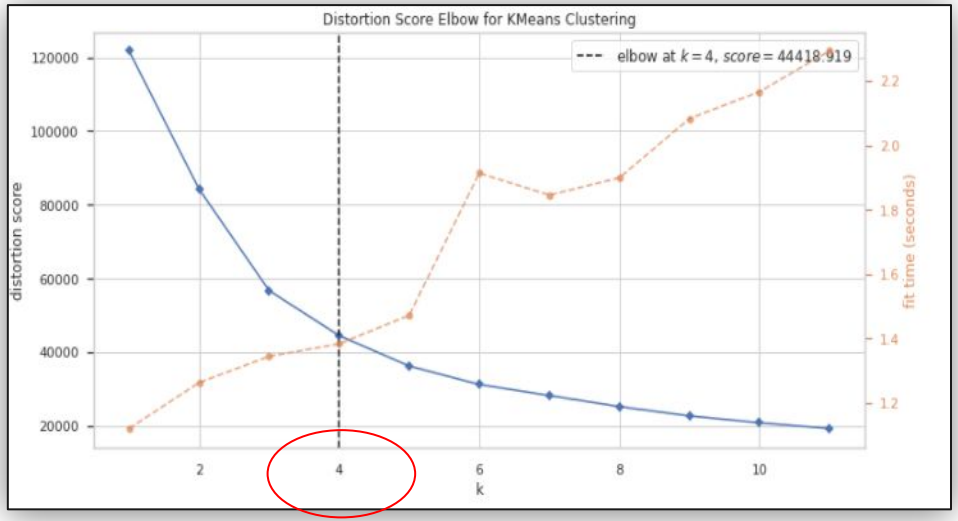
Résultats satisfaisants MAIS lenteurs à l'exécution…

K-Means

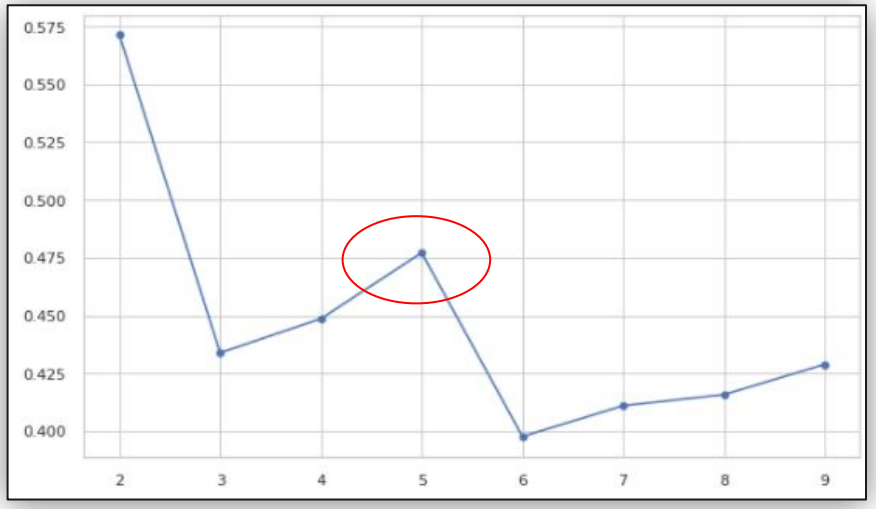

Paramètres :
n_clusters
init
n_init
Nb de clusters défini à l'avance :
OUI
Notions :
critère Inertia
score silhouette

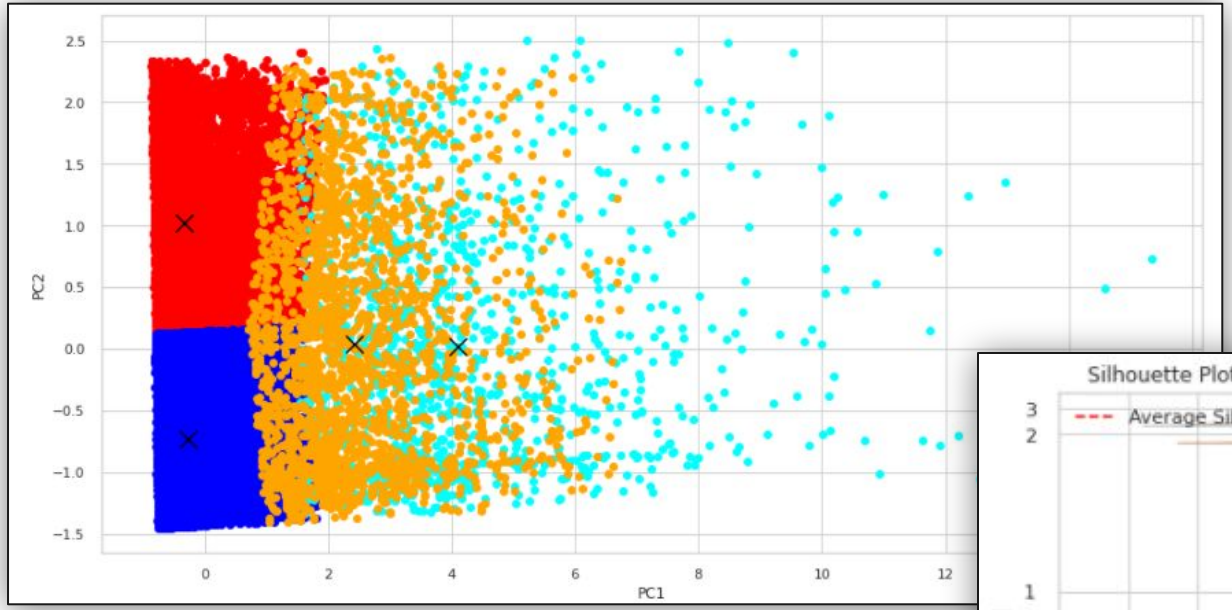
K-Means sur 3 variables : RFM

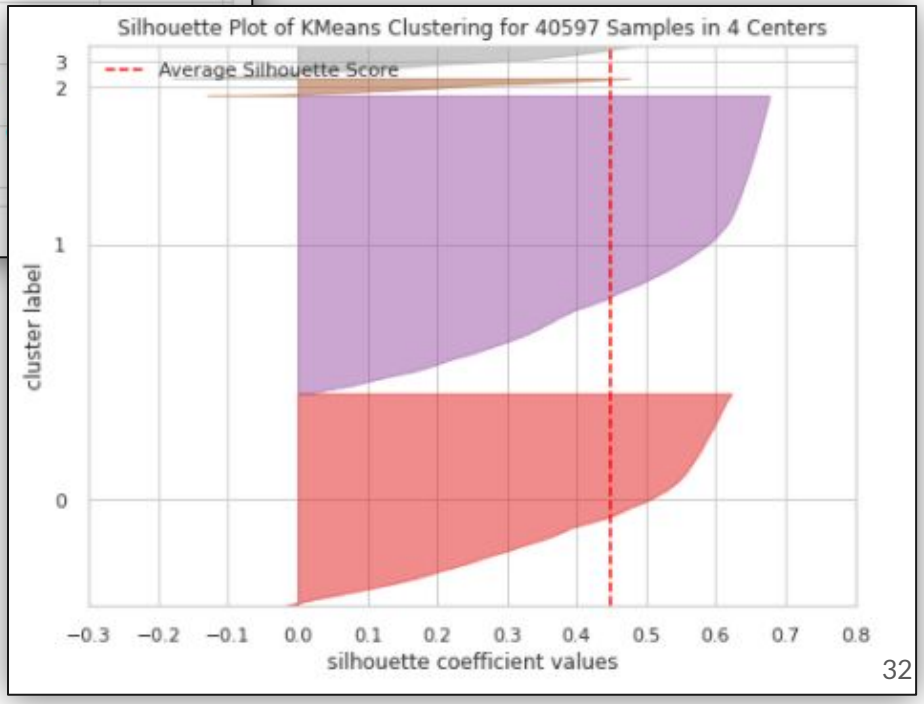
Méthode du coude



Score Silhouette

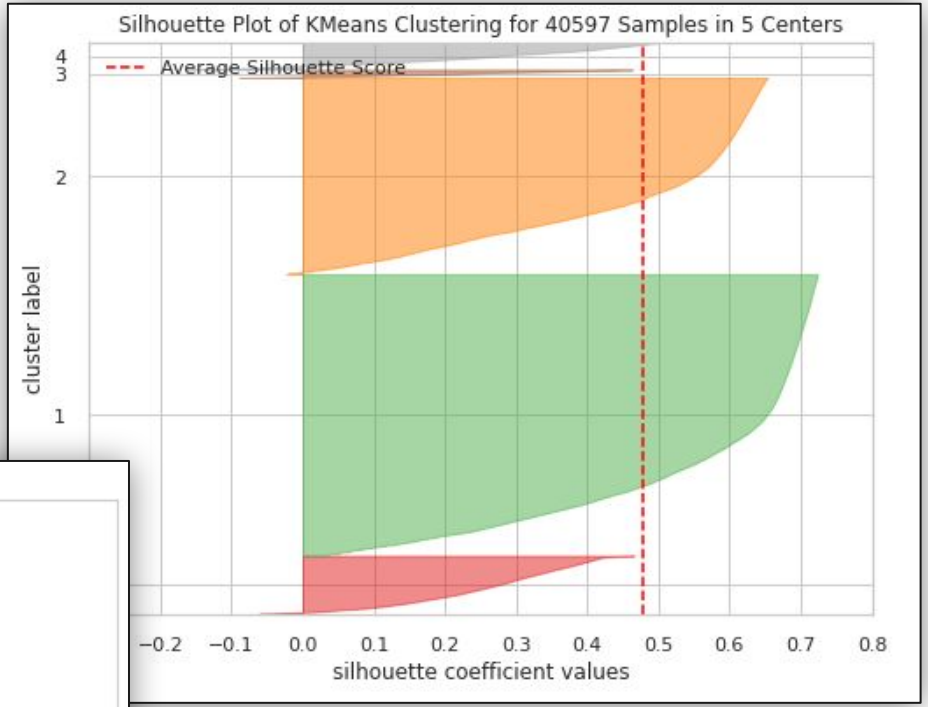


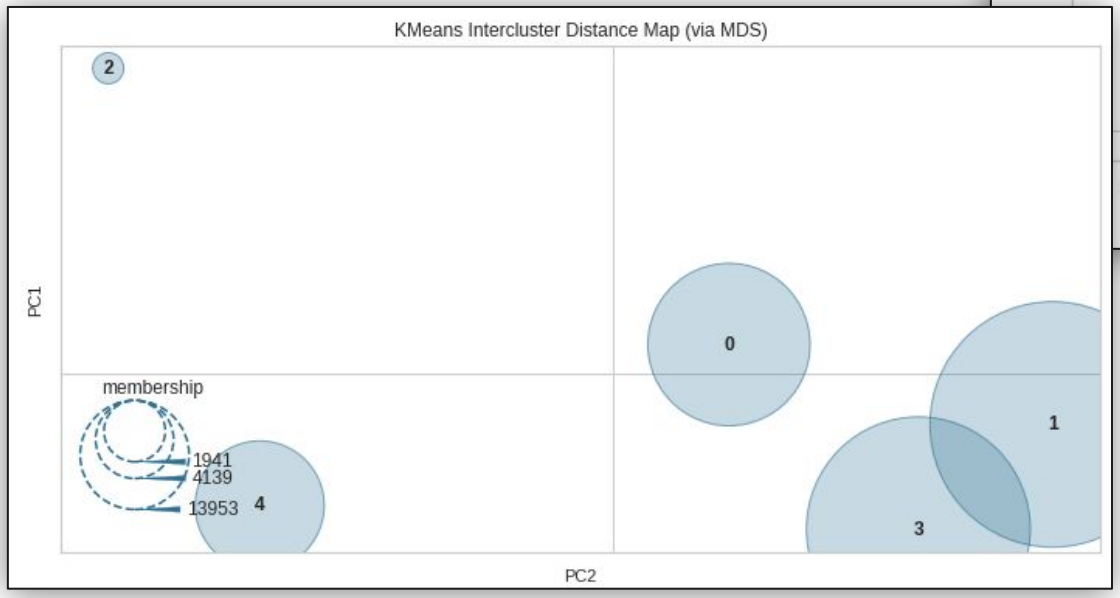




Choix optimal de clusters : 4

Essai avec 5 clusters

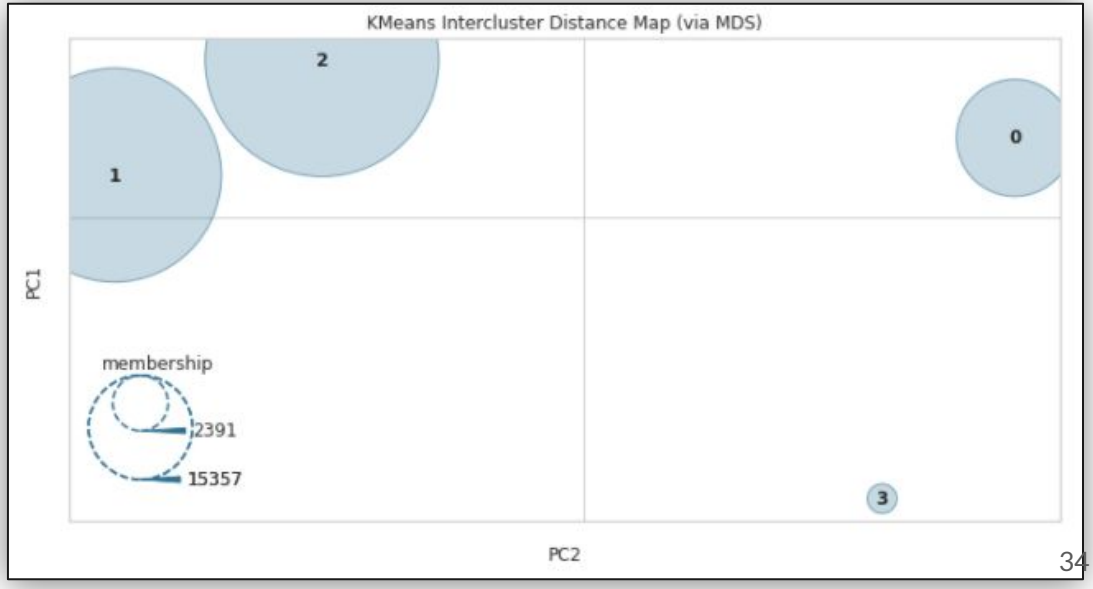


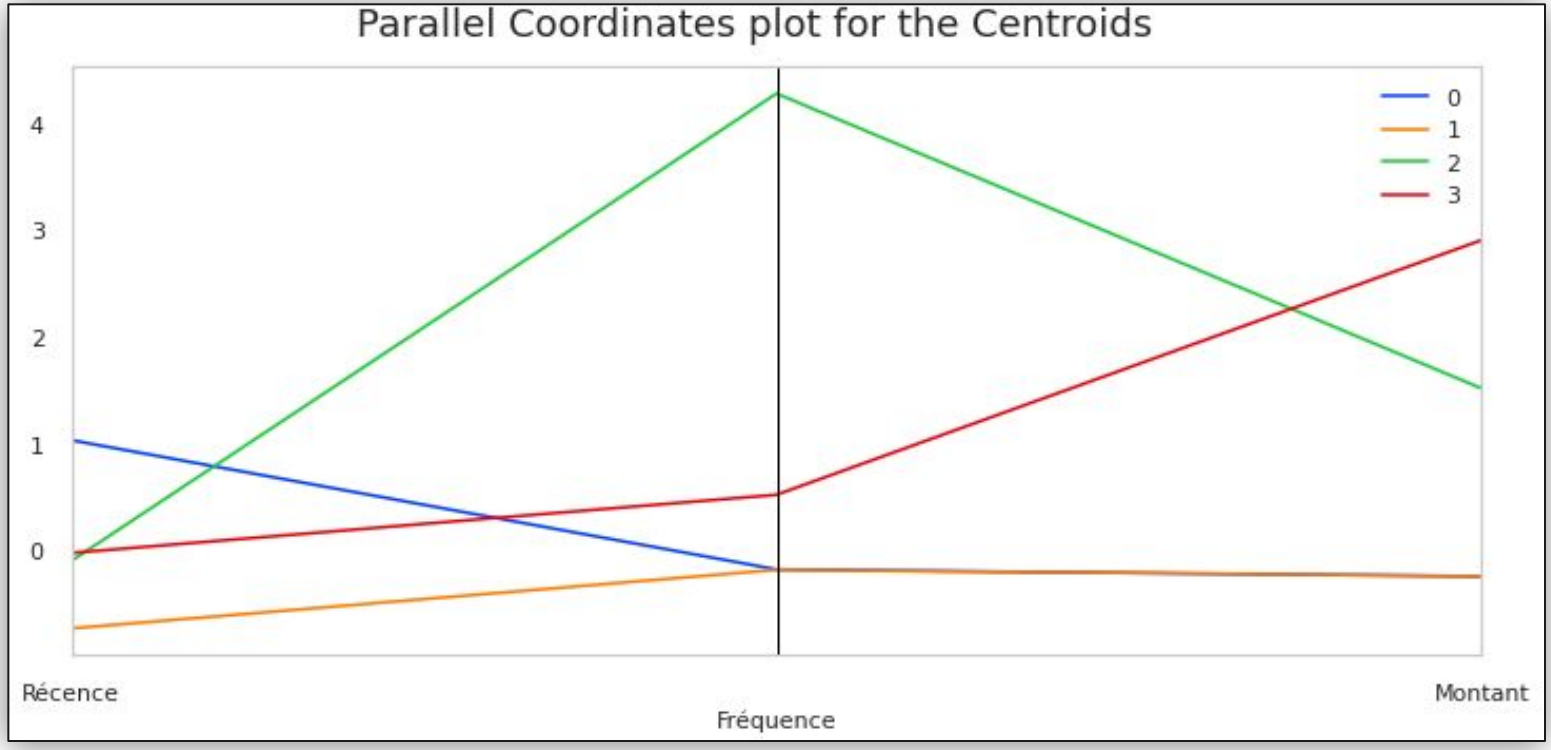


Cluster	Récence mean	Fréquence mean	Montant mean	Segment unique	Segment count
0	234.000	1.000	127.000	[Hibernating, About to sleep, At risk]	15357
1	67.000	1.000	127.000	[Promising, About to sleep, New customers, Pot...	21612
2	127.000	4.000	473.000	[At risk, Potential loyalists, Loyal customers...	1237
3	134.000	2.000	746.000	[About to sleep, Potential loyalists, Hibernat...	2391

Critères :

- ➢ taille des silhouettes
- ➢ volumétrie par cluster
- ➢ présence d'erreurs





- **Cluster 0** : les clients dont la Récence est élevée (le dernier achat date d'il y a longtemps) et qui ont dépensé peu d'argent
- **Cluster 1** : les clients dont la Récence est basse (le dernier achat est récent) et qui ont dépensé peu d'argent
- **Cluster 2** : les clients qui achètent souvent et dont le panier moyen est important
- **Cluster 3** : les clients qui n'achètent pas souvent et dont le panier moyen est très élevé

Etude de stabilité

Objectifs et méthodologie

Tester la stabilité de l'algorithme **dans le temps** pour savoir à quel moment **les clients changent de cluster.**

Etape 1

Jeux de données du futur

Créer plusieurs datasets : B0 et les B1, B2, B3 etc du futur.

Etape 2

Clustering

Effectuer une segmentation C1, C2, C3 etc sur les bases du futur.

Etape 3

Prédictions

Générer des labels B1, B2, B3 etc.

Etape 4

Comparaison

Comparer la segmentation entre C0 et Cx grâce à l'indice de Rand ajusté.

Indice de Rand (ajusté)

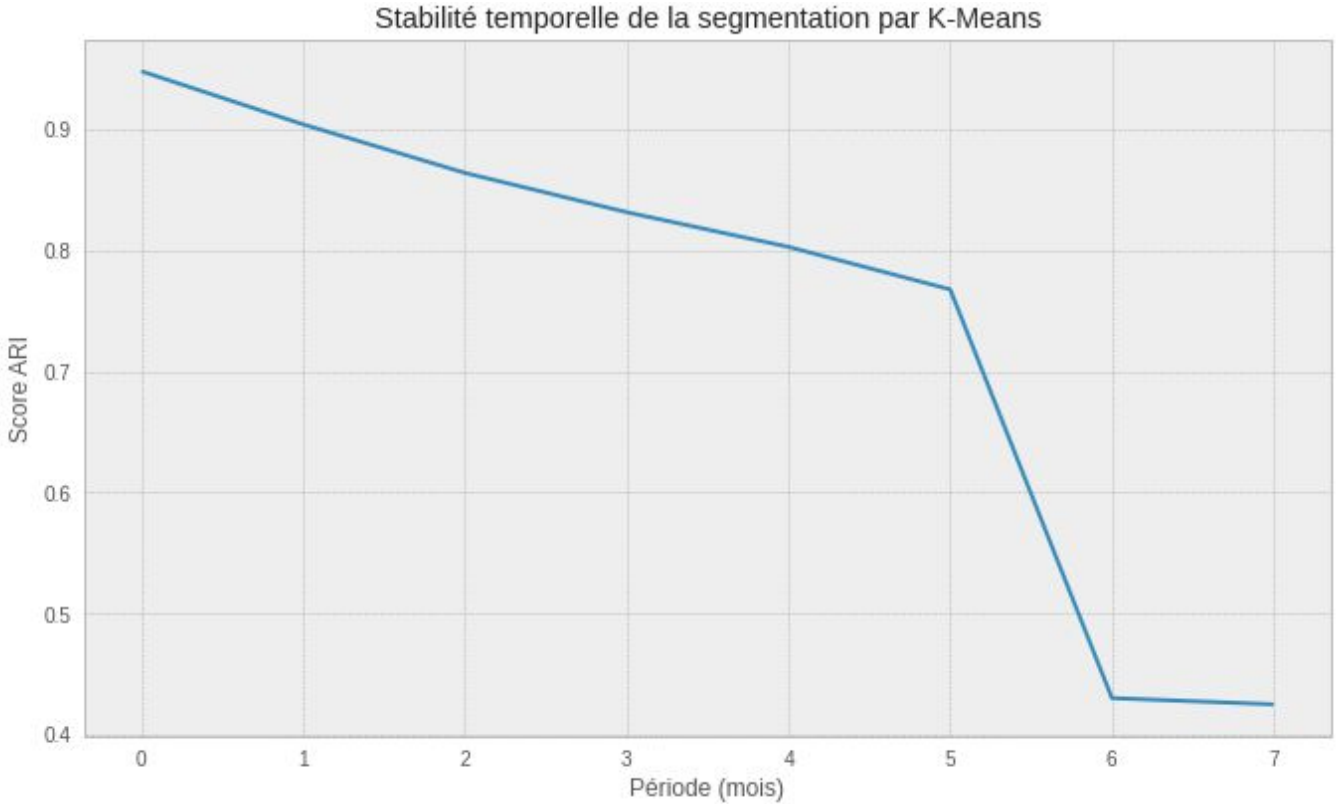
Indice de Rand :

utilisé pour mesurer la **similarité des points de données** présents dans les clusters.

ARI :

- proche de 0 pour un clustering aléatoire
- égal à 1 uniquement quand le clustering correspond exactement à la partition initiale

Calcul du score ARI



Conclusion

- Parmi toutes les méthodes de clustering appliquées, **c'est le K-Means qui semble être le plus efficace.**
- Le CAH donne des résultats similaires, en revanche **les temps de performances laissent à désirer.**
- Les résultats de K-Means sont parfaitement **interprétables et utilisables** par l'équipe marketing.
- La segmentation RFM semble fournir **une analyse des profils client tout aussi pertinente** (note : rétrécir la liste des comportements d'achat en fusionnant certains profils - ceux similaires ou ceux les moins nombreux).

Avez-vous des questions ?