



OLIST

Nom du projet : **Segmentez des clients d'un site e-commerce**

Présenté par : Nathan FARDIN

Plan

I. Introduction

II. Données et feature
engineering

III. Machine learning
et
Maintenance

IV. Conclusion

Le besoin

Proposer à l'entreprise Olist une solution de clusterisation de sa base de données pour mieux appréhender les profils clients.

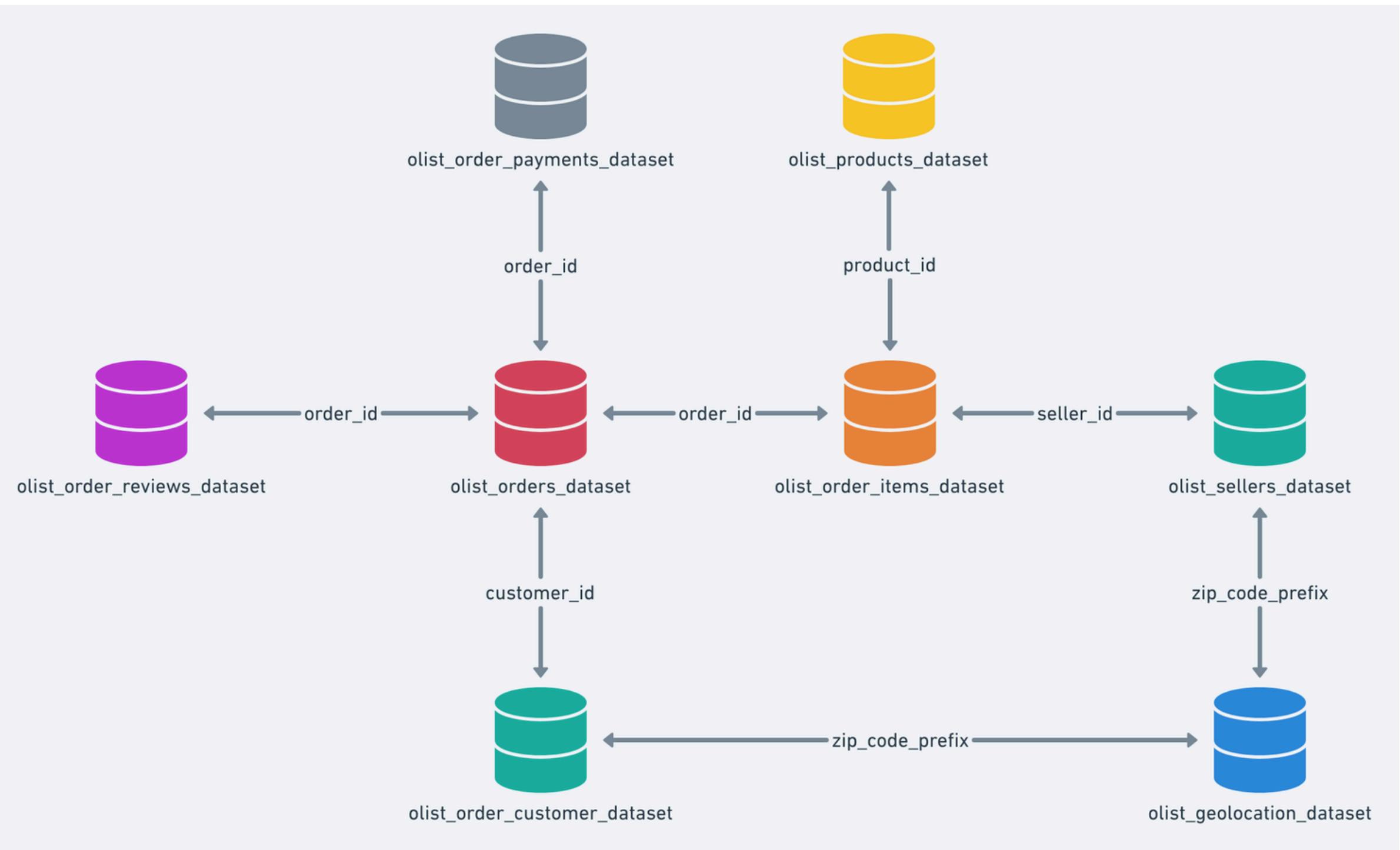
Simuler un contrat de maintenance.



Les données

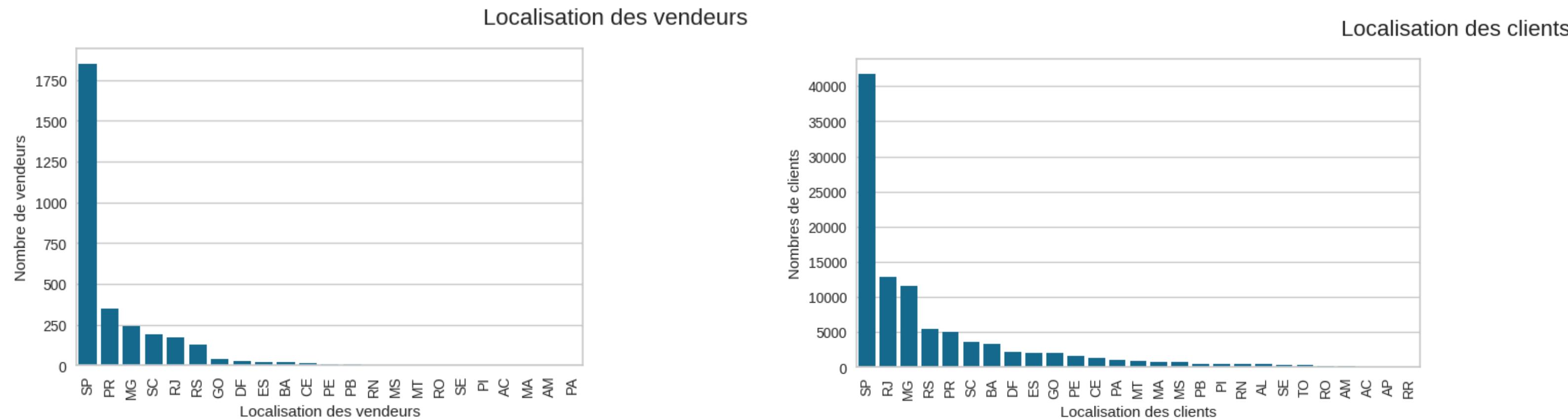
8 tables de données contenant des informations variées sur les clients, vendeurs produits et commande issu d'une plateforme e commerce.

SQL



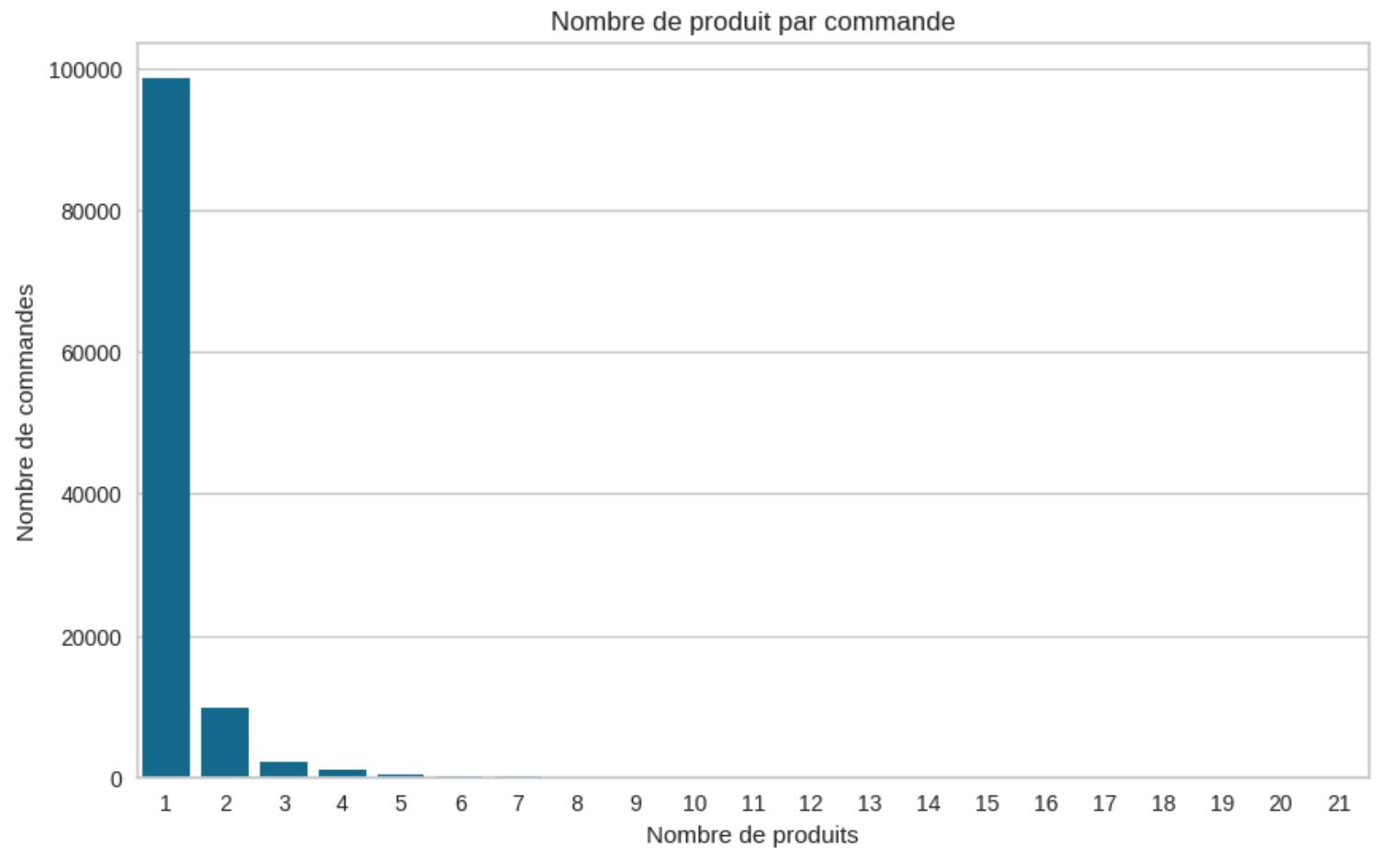
4 scripts SQL
demandés par Olist
pour permettre une
première visualisation
de ses clients

Localisation



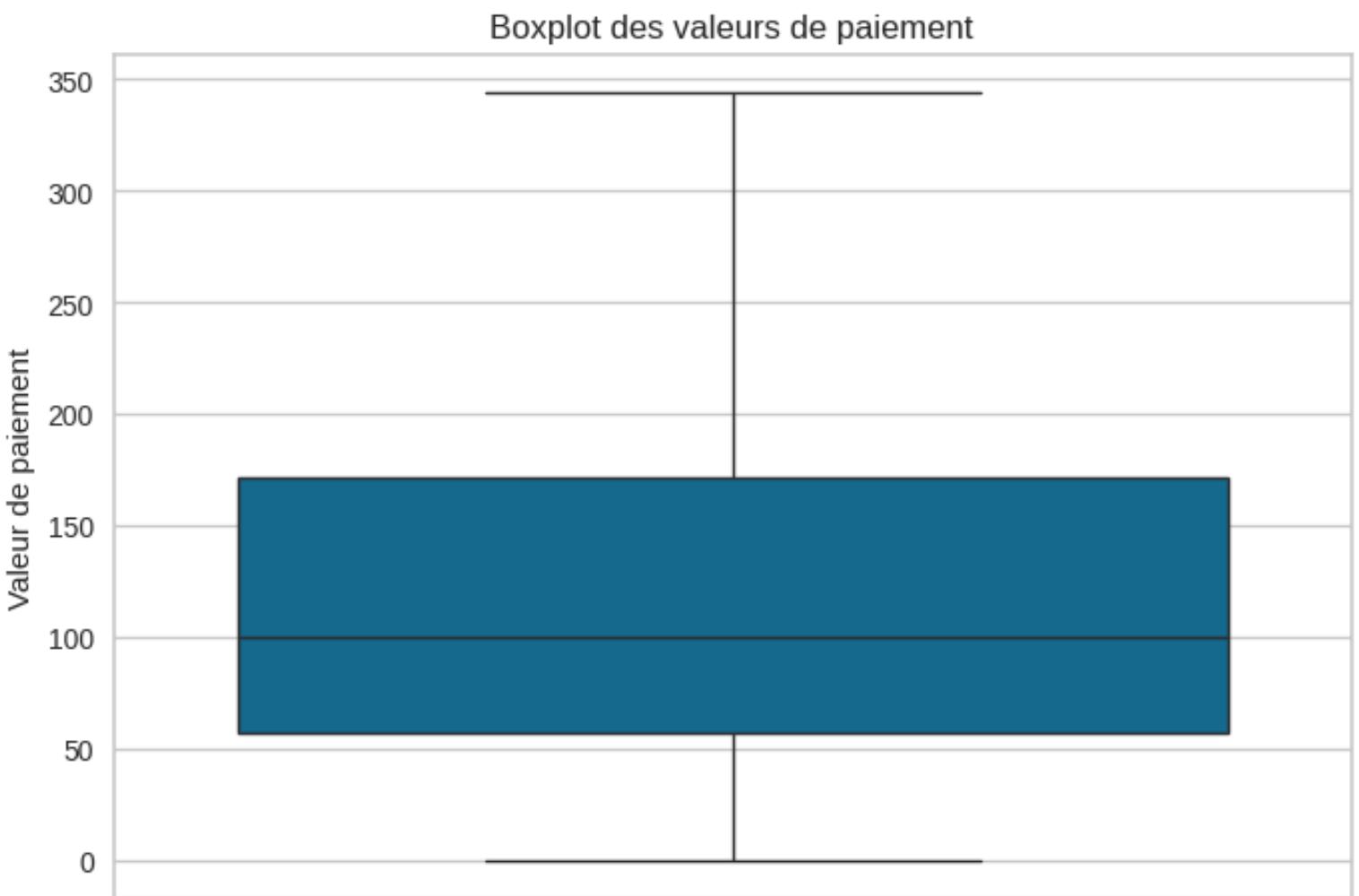
La majorité des vendeurs et des acheteurs sont localisé à Sao Paulo
Une des villes les plus peuplés et un centre financier important

Les commandes

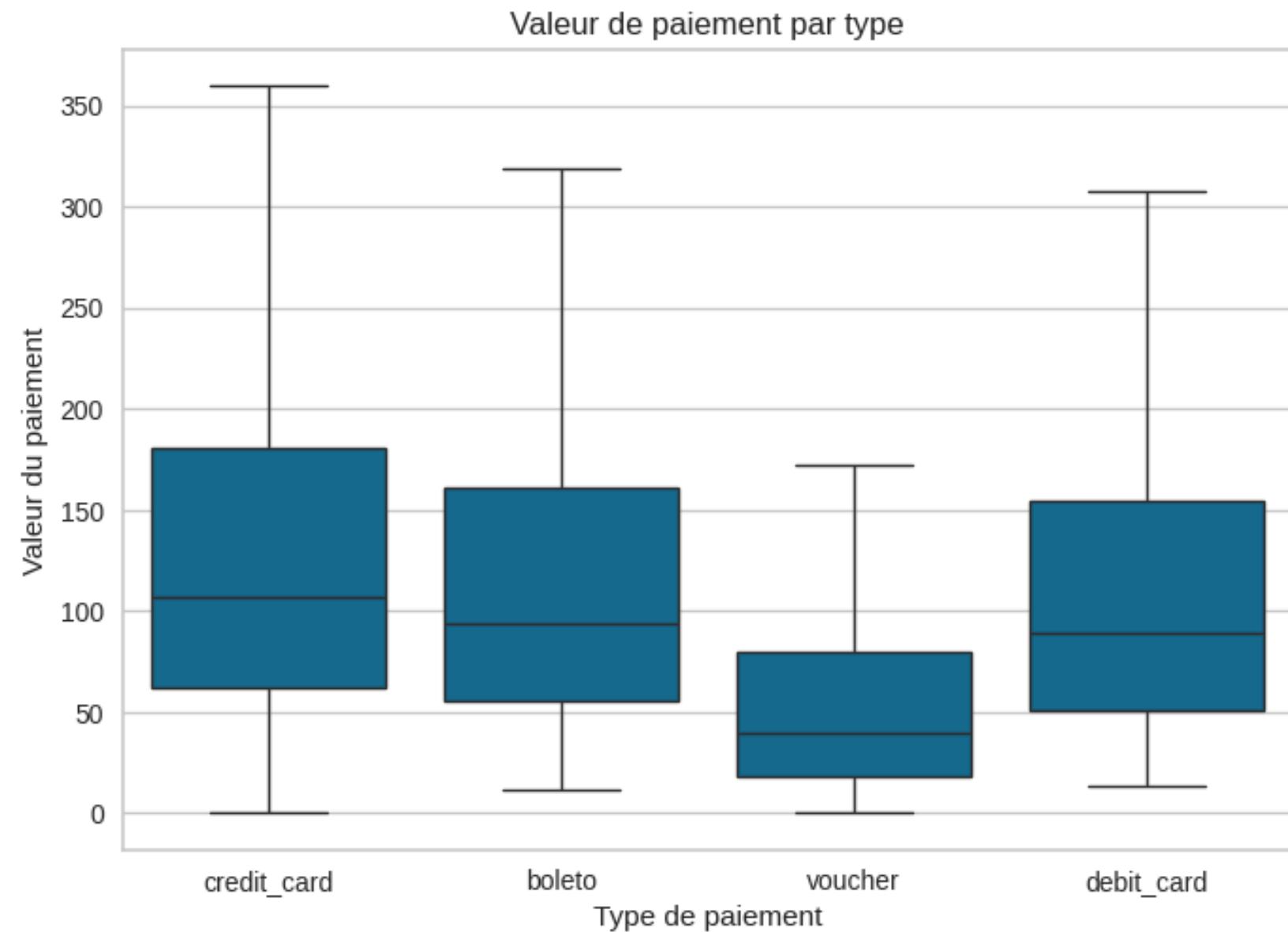


La grande majorité des commandes ne comporte qu'un article

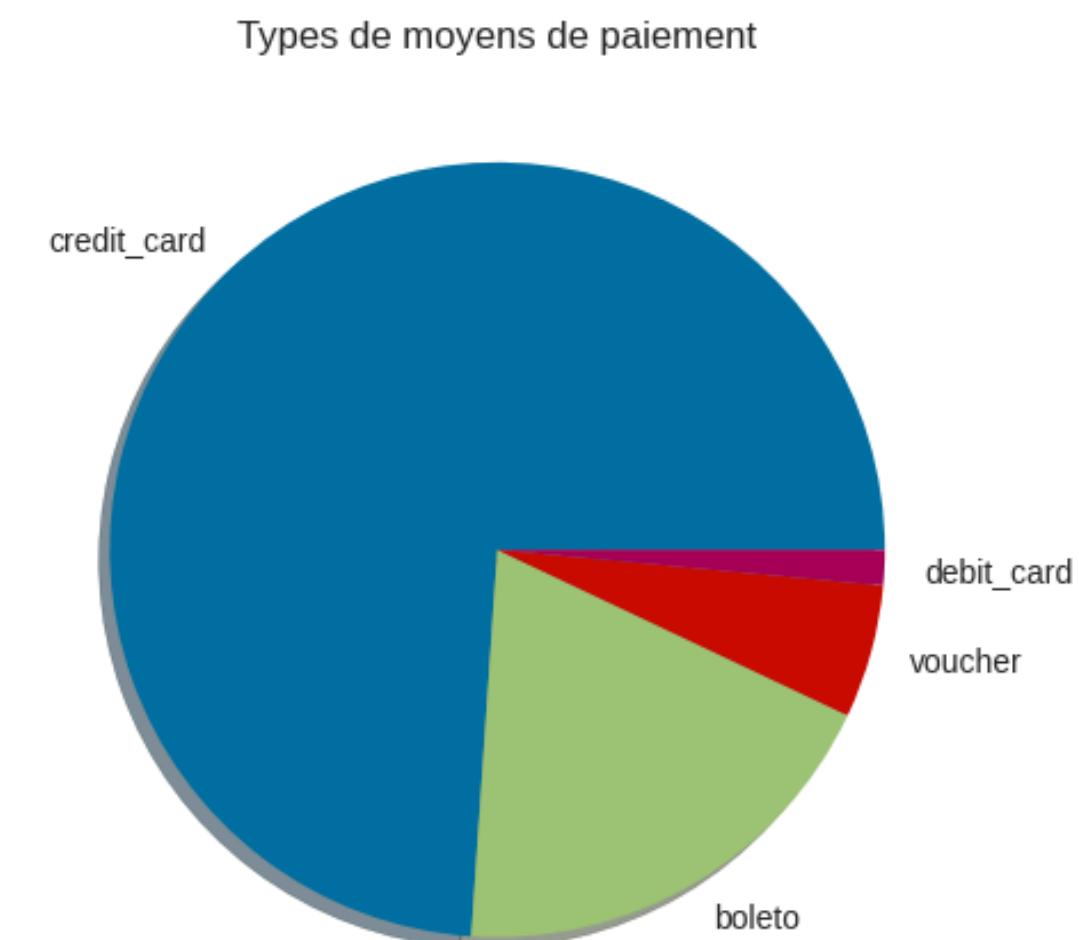
La médiane des valeurs de commande est à 100 breal



Transaction



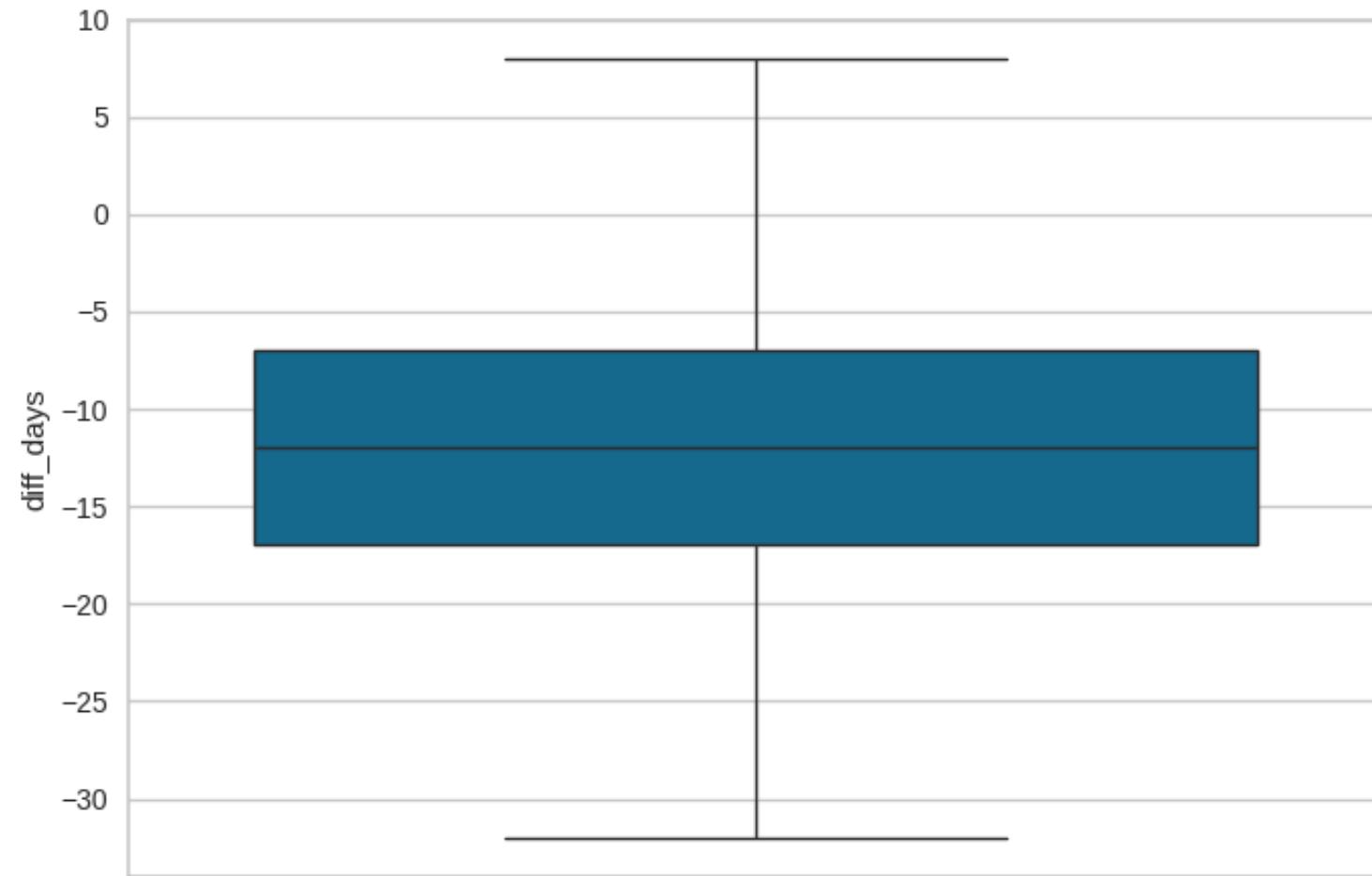
Le moyen de paiement le plus utilisé est la carte de crédit et c'est également avec celui-ci que les clients dépenses le plus.



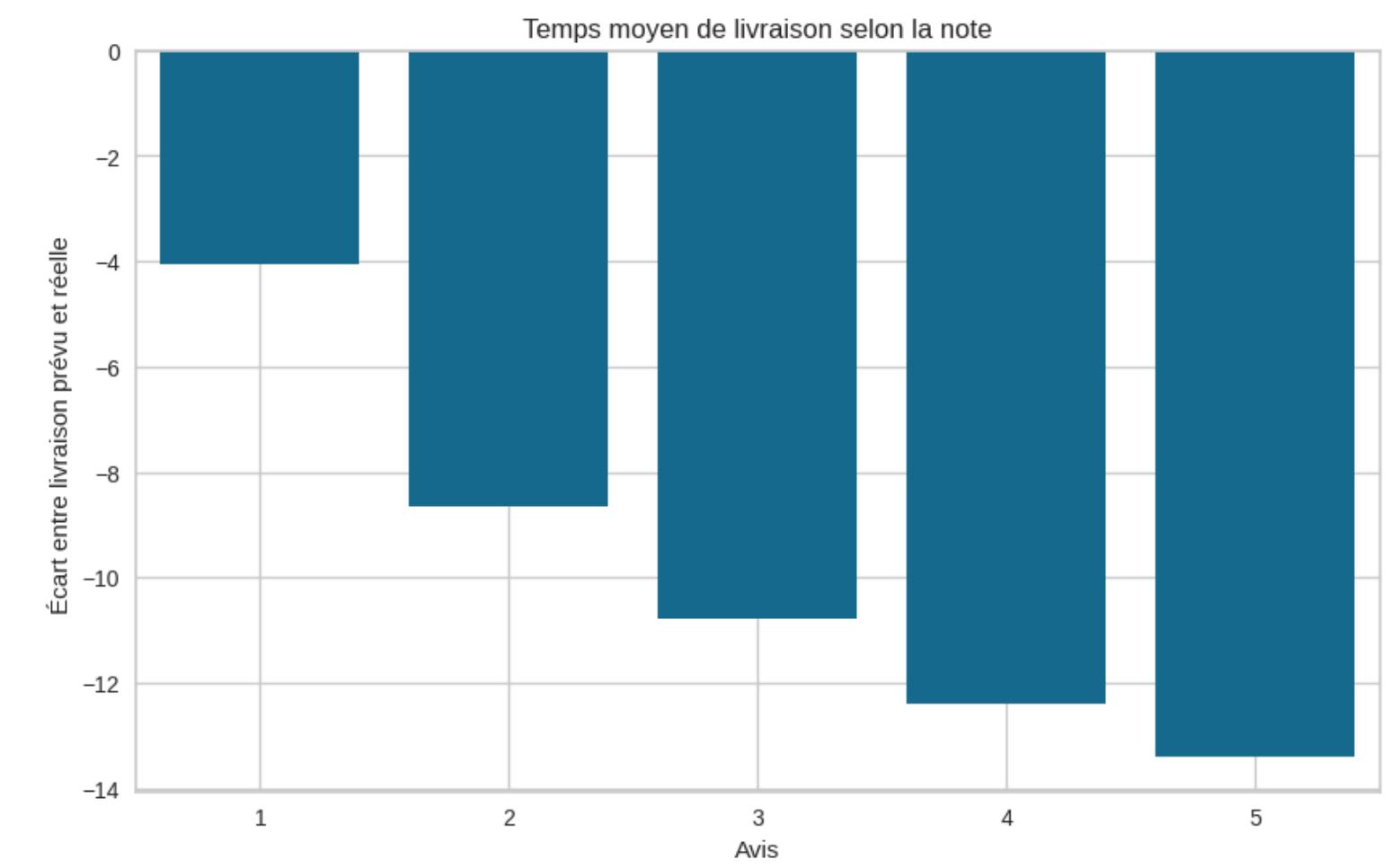
Voucher : bon de paiement

Boleto : moyen de paiement en espèce
très prisé au Brésil.

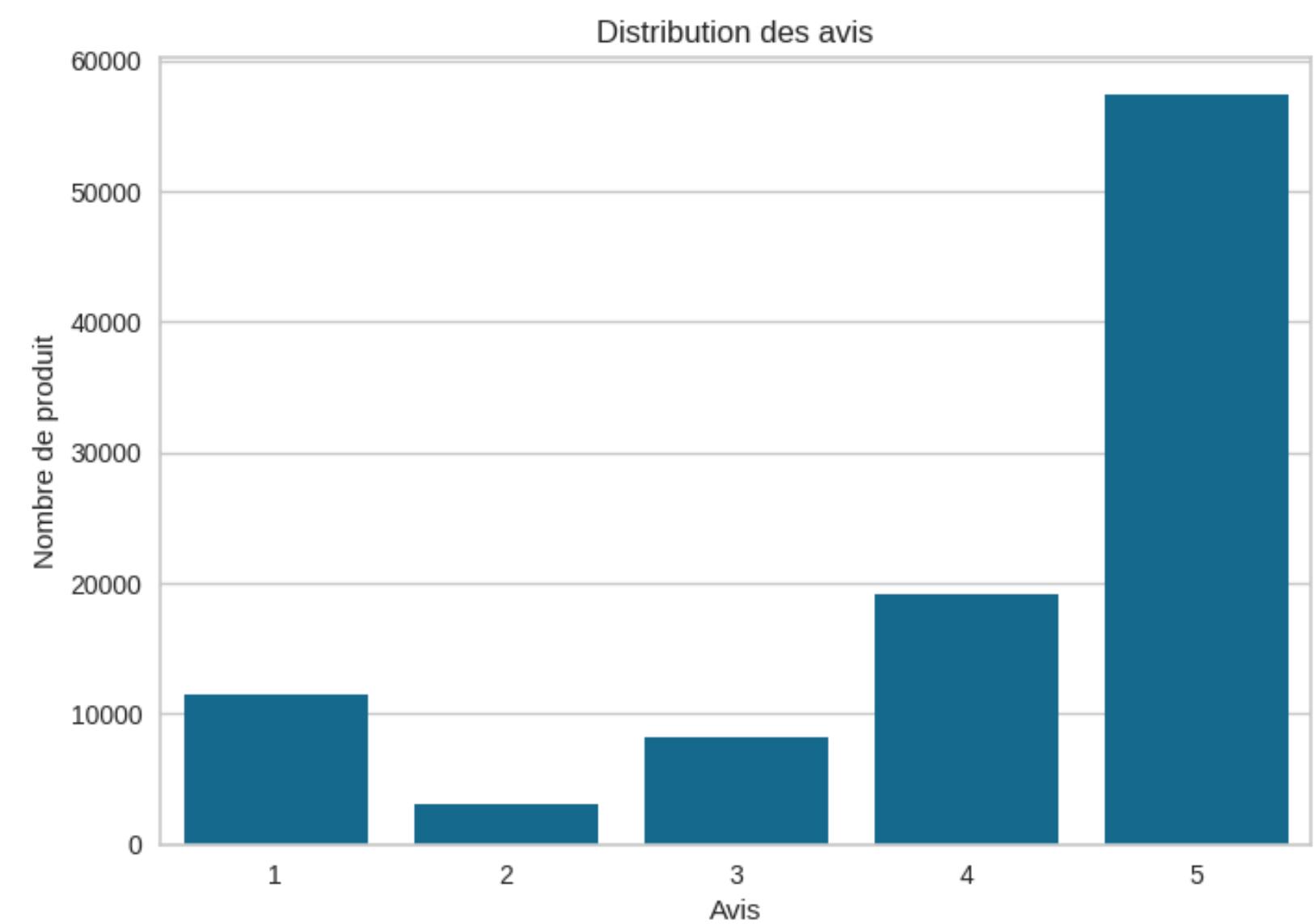
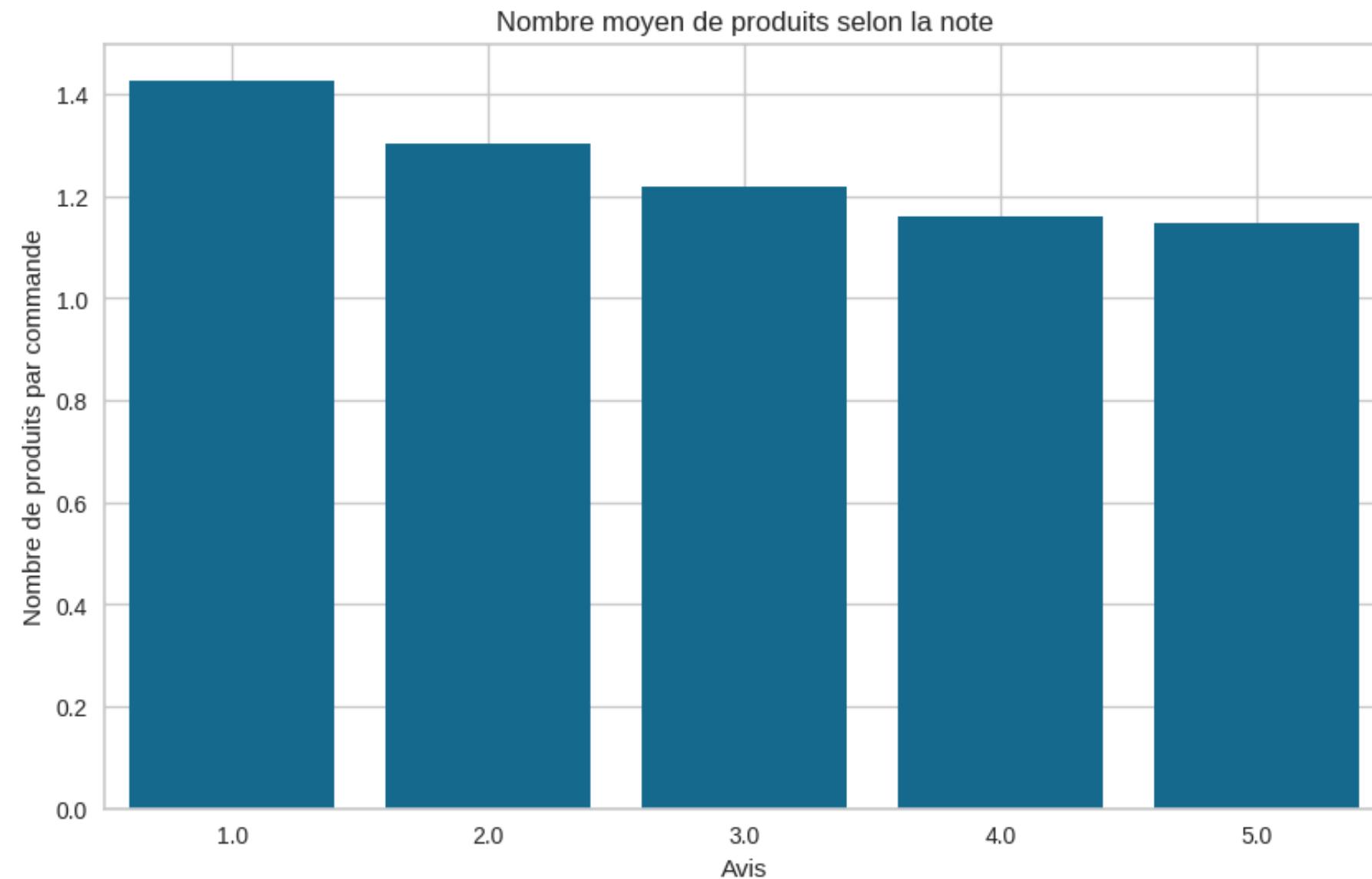
Respect des délais de livraison



Le plus souvent, les commandes arrivent avant la date annoncée, ce qui a tendance à influer positivement la note des avis données par les clients.



Satisfaction clients



Plus un grand nombre d'articles est commandé plus les clients ont tendance à donner une note faible.

Néanmoins, la majorité des clients est très satisfaite de ses achats.

Feature engineering

Création de 5 variables

Recency : date du dernier achat.

Frequency : nombre d'achats sur la période.

Monetary : montant des achats sur la période.

Delay : délai moyen entre la date de livraison annoncée et la date réelle.

Average_review_score : note moyenne sur la période.

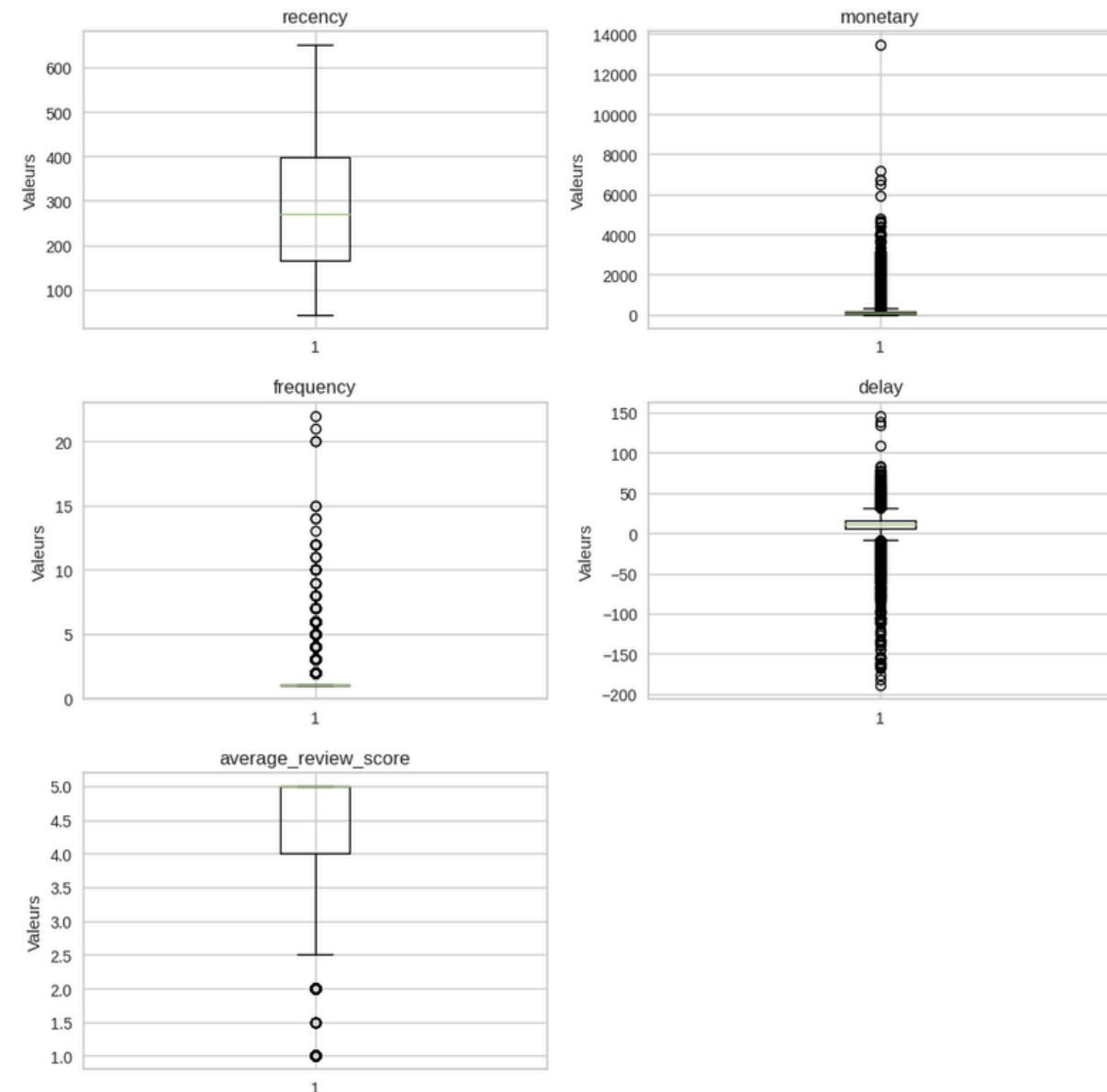


Données manquantes

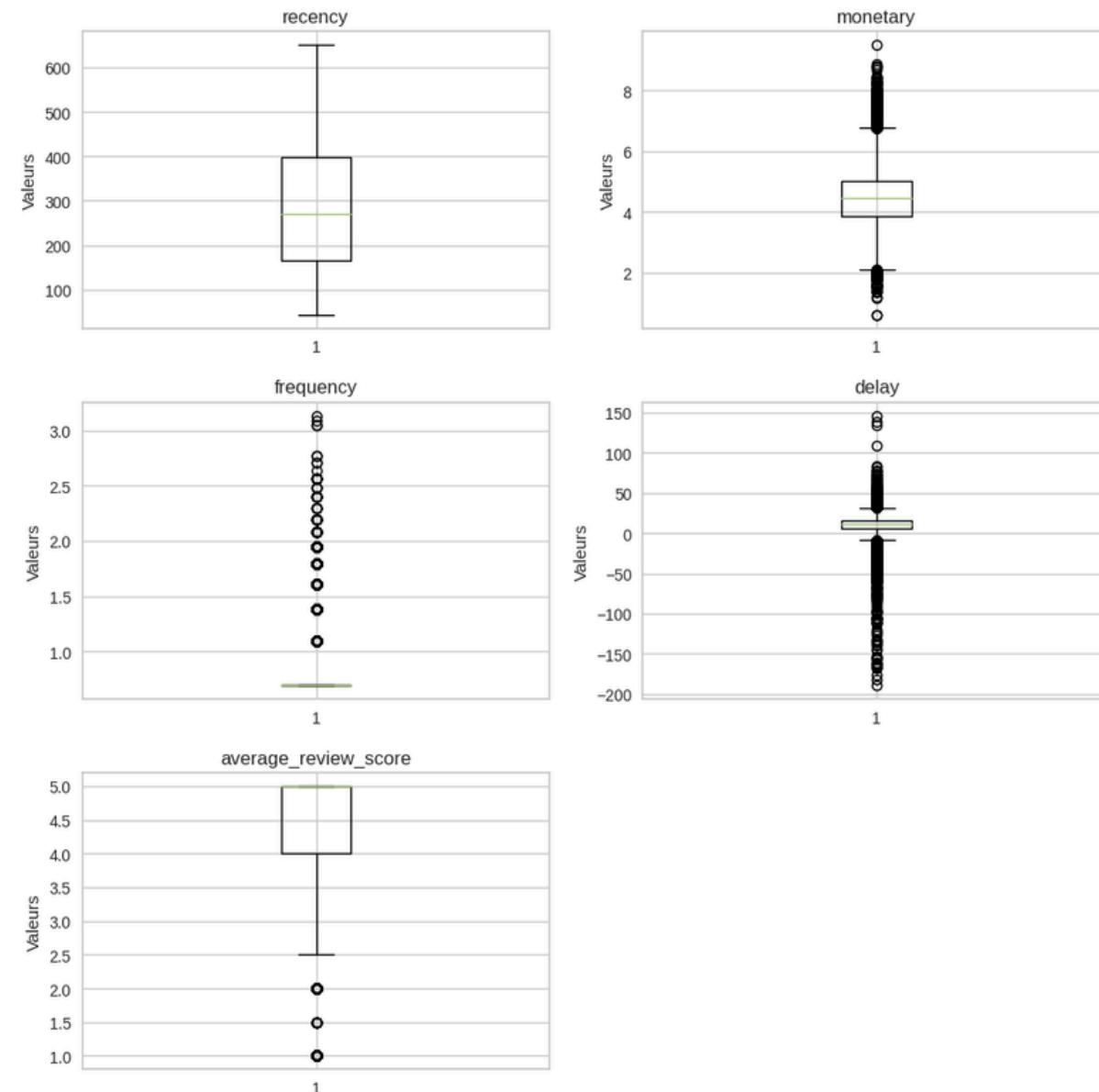
Imputation à la moyenne pour les données manquantes.

404

Etalement



Avant



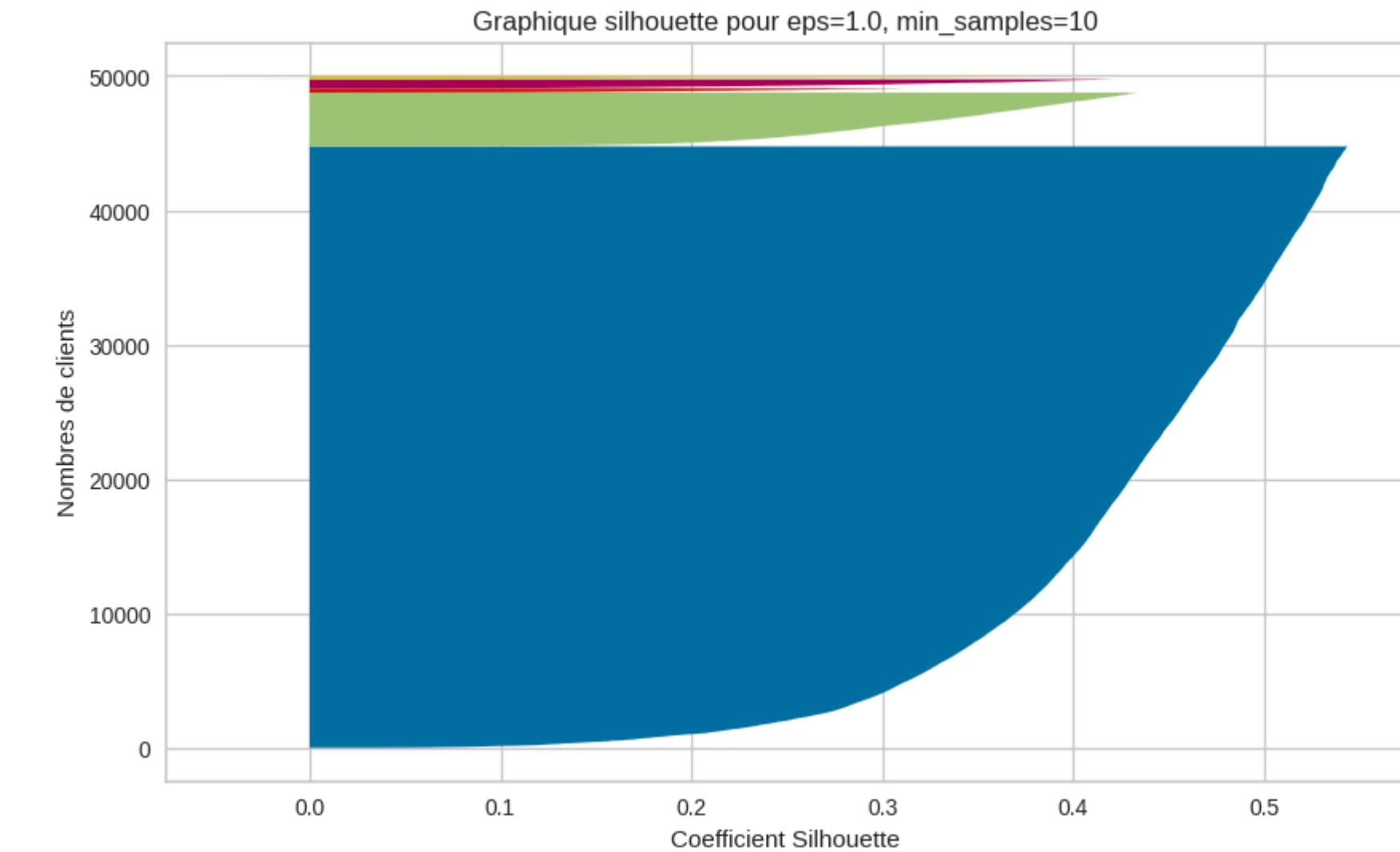
Après

DBscan

RFM

Hyperparamètres

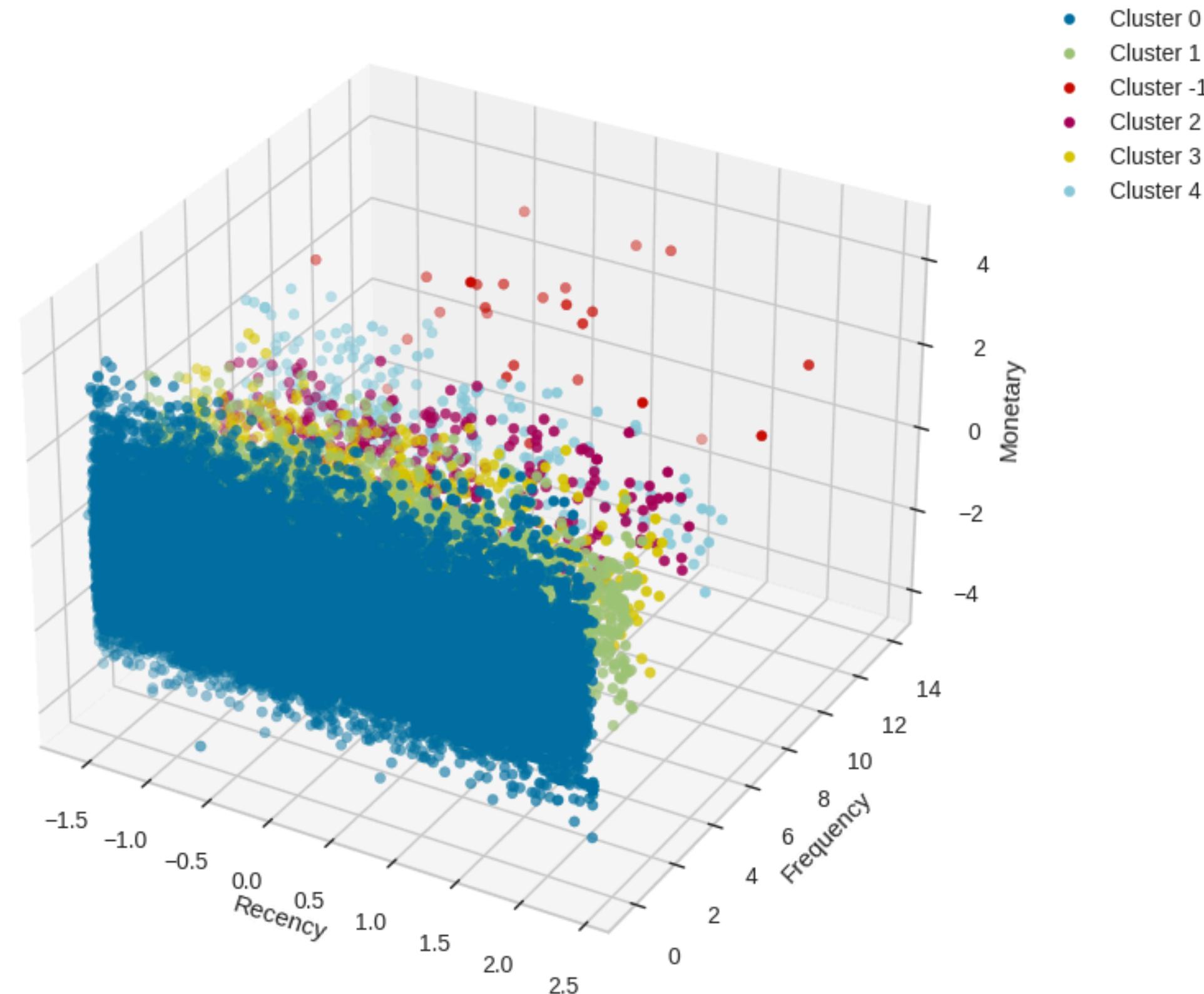
Eps	range(0.2, 1, 0.2)
min_samples	3, 5, 10, 20



Meilleur hyperparamètres :
eps : 1, min_samples : 10
Silhouette score :
0.413

Visualisation clusters DBscan

Scatter Plot 3D des Clusters DBSCAN

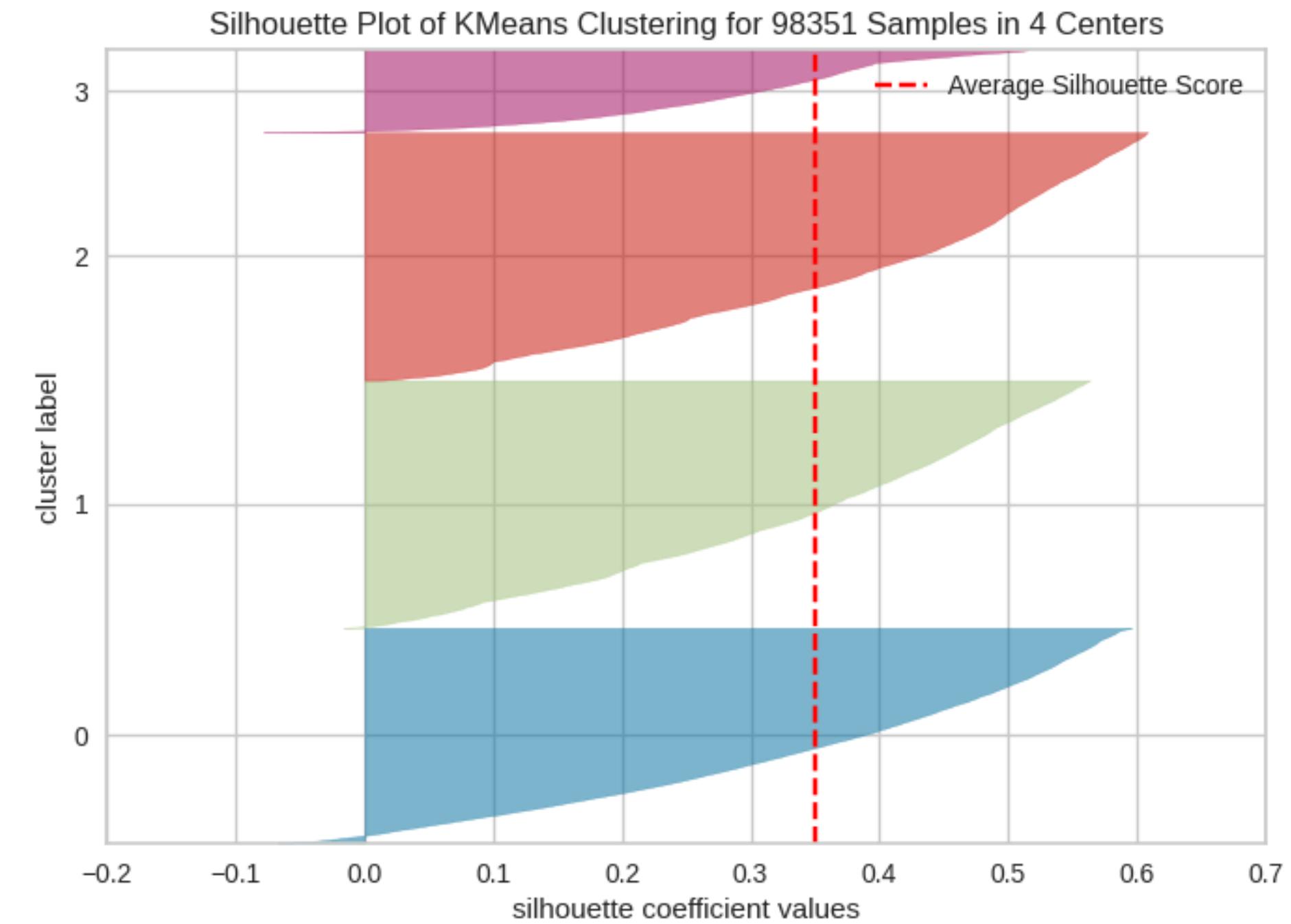


Kmeans

RFM

Grille
d'hyperparamètres

n_clusters	range(4, 10, 1)
init	'k-means++', 'random'
n_init	10, 20, 30

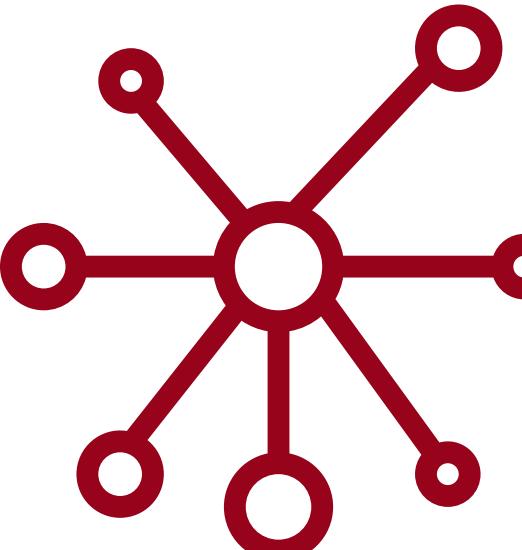
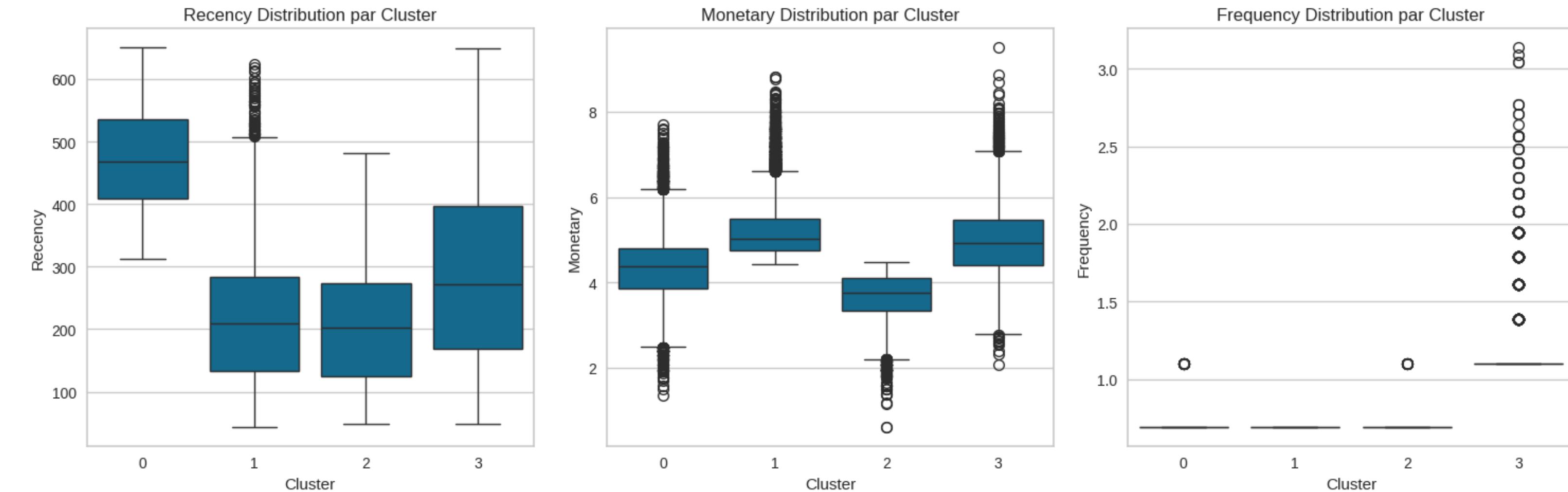


Meilleur hyperparamètres :
'init': 'k-means++', 'n_clusters': 4, 'n_init': 10

Silhouette score :
0.3502

Caractéristiques des clusters

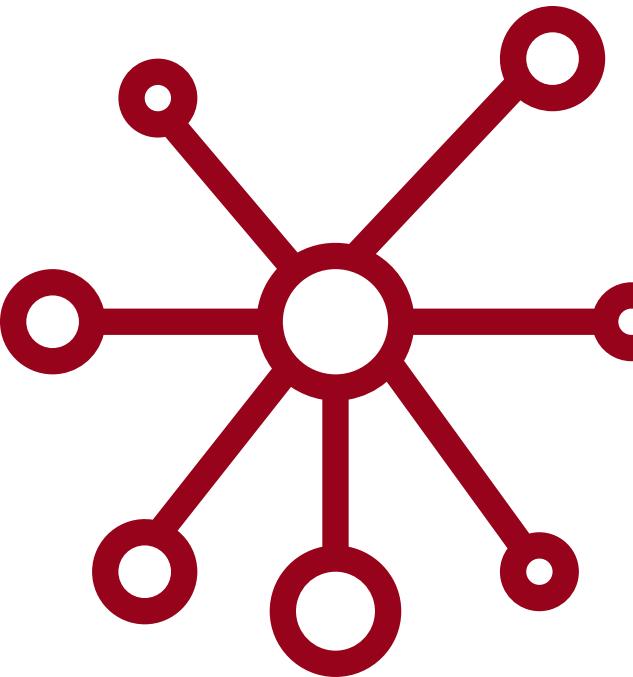
Boxplot



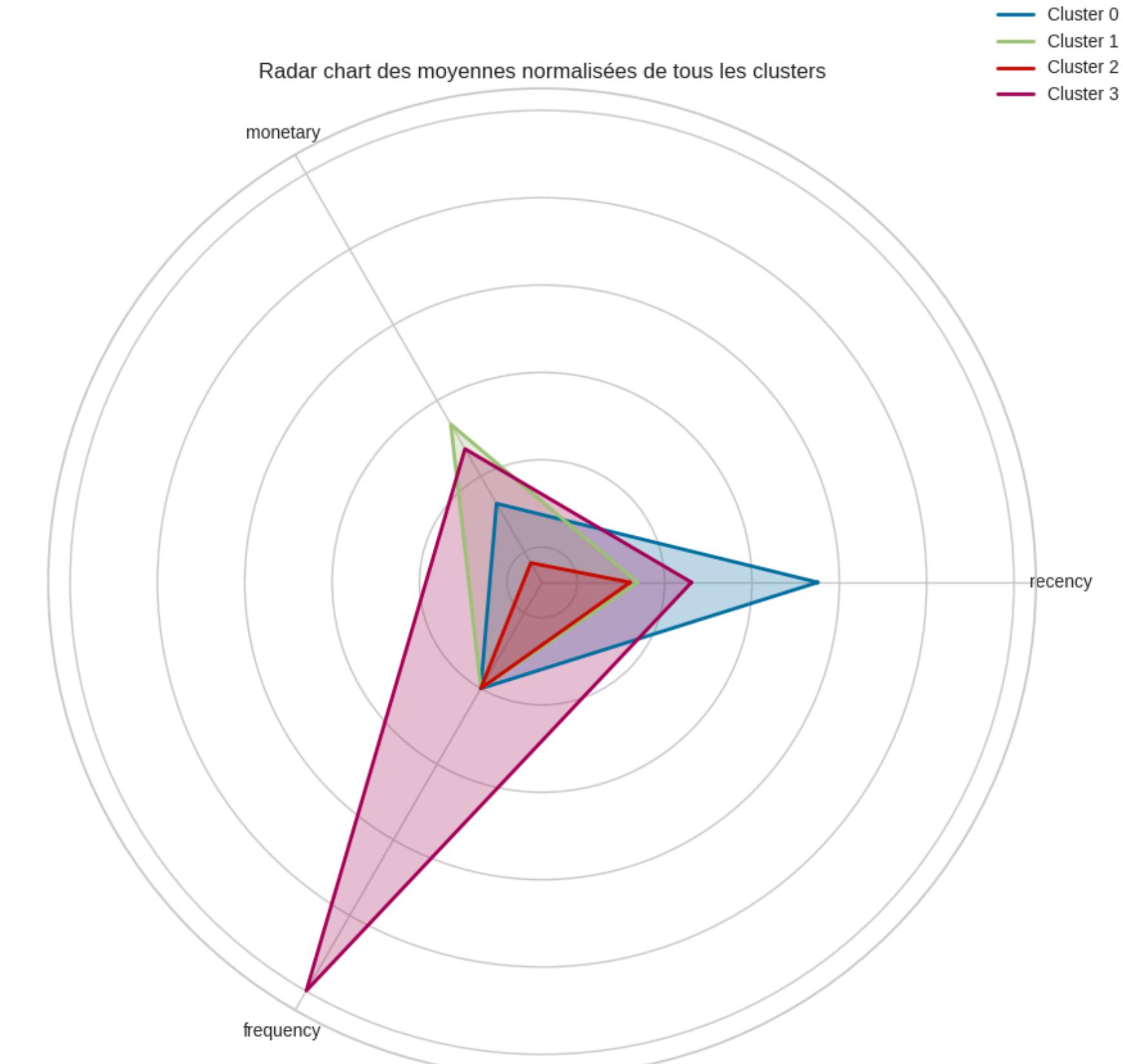
Kmeans

Caractéristiques des clusters

Radar plot



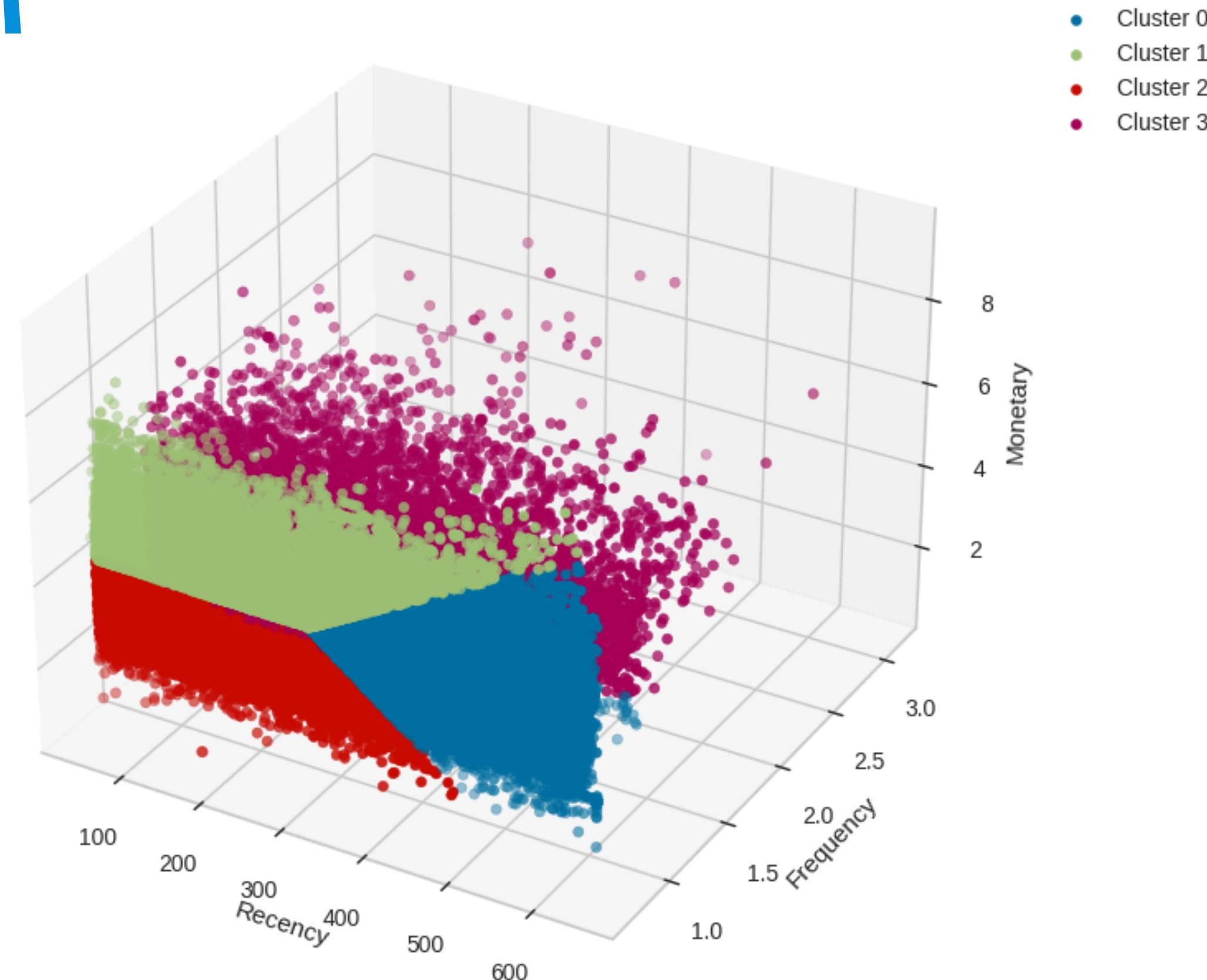
RFM



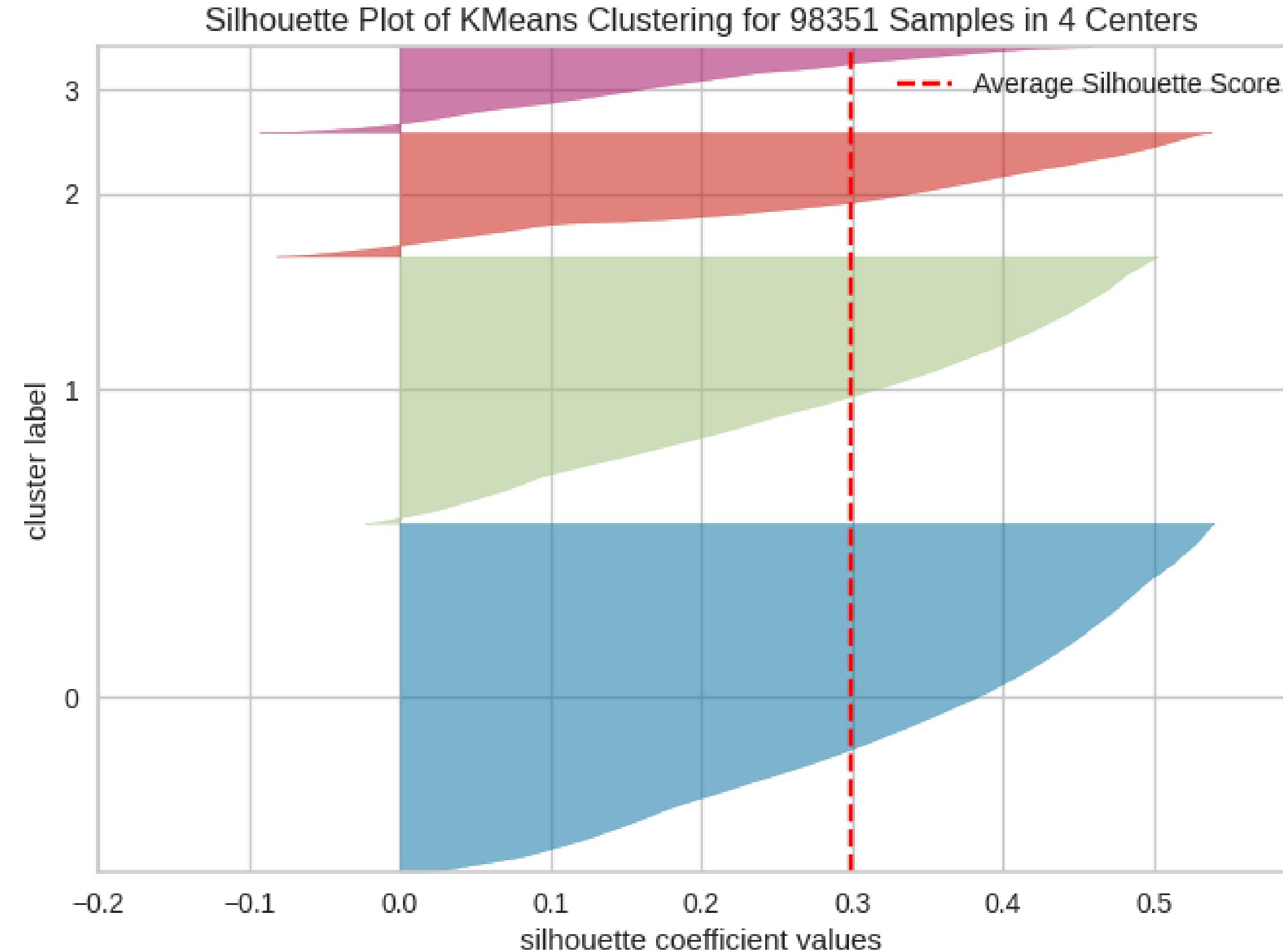
Visualisation clusters

RFM

Scatter Plot 3D des clusters du KMeans



Clusters RFM + reviews



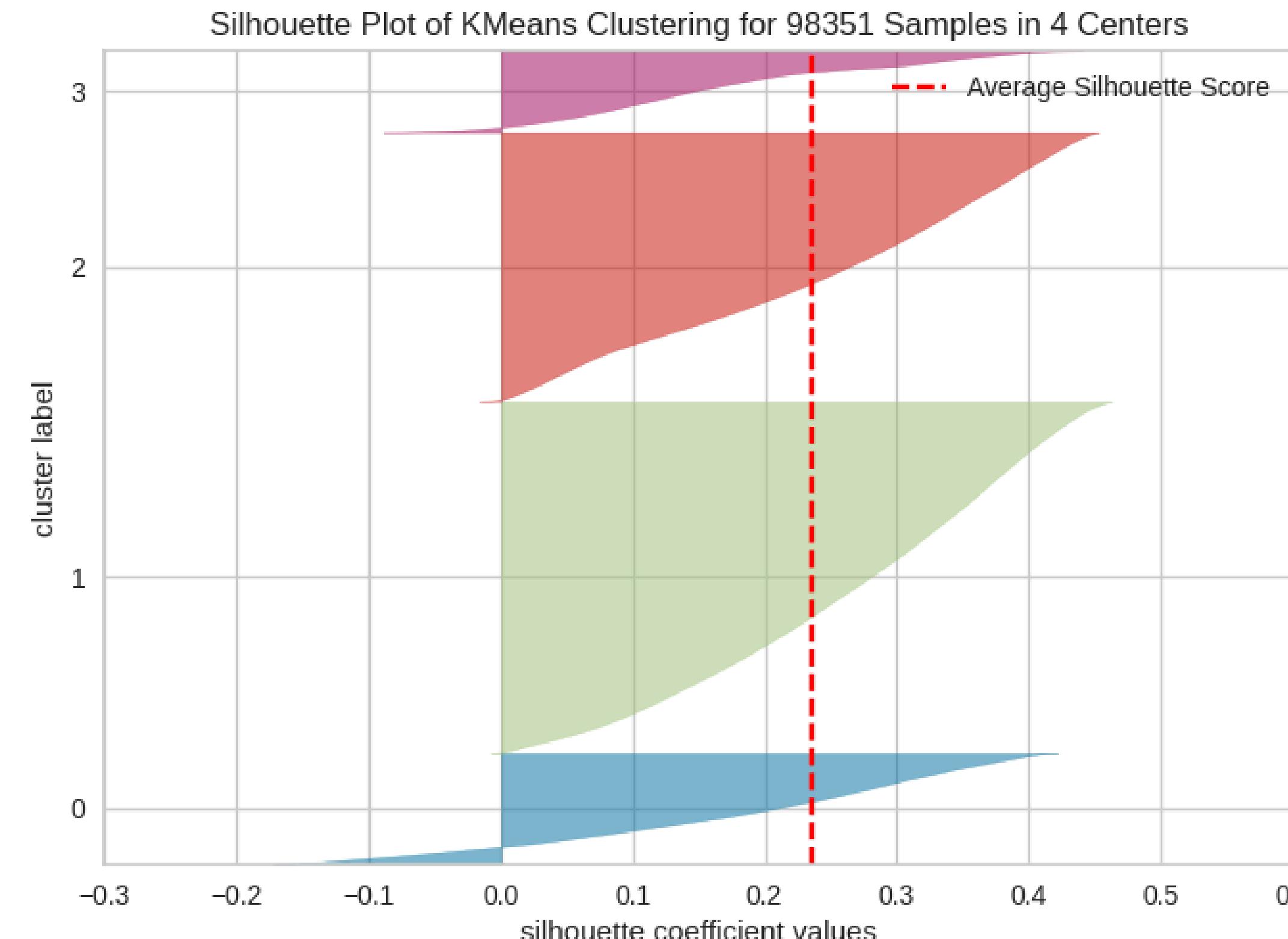
légère baisse du score silhouette
mais ajout d'une information pertinente,
la satisfaction client

Radar plot

légère perte de score
silhouette mais ajout
d'une information
pertinente, la
satisfaction client



Clusters RFM review score et délai de livraison



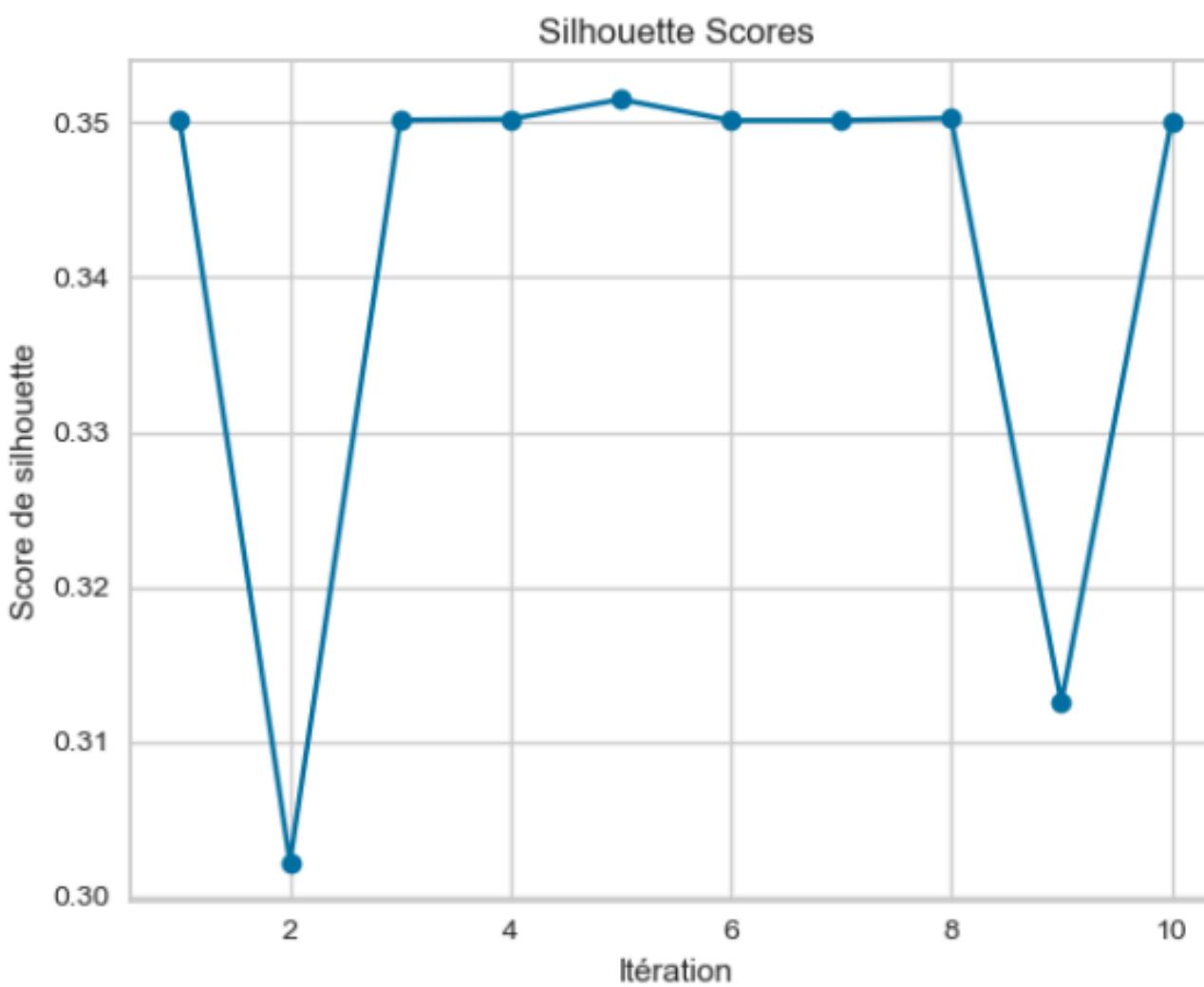
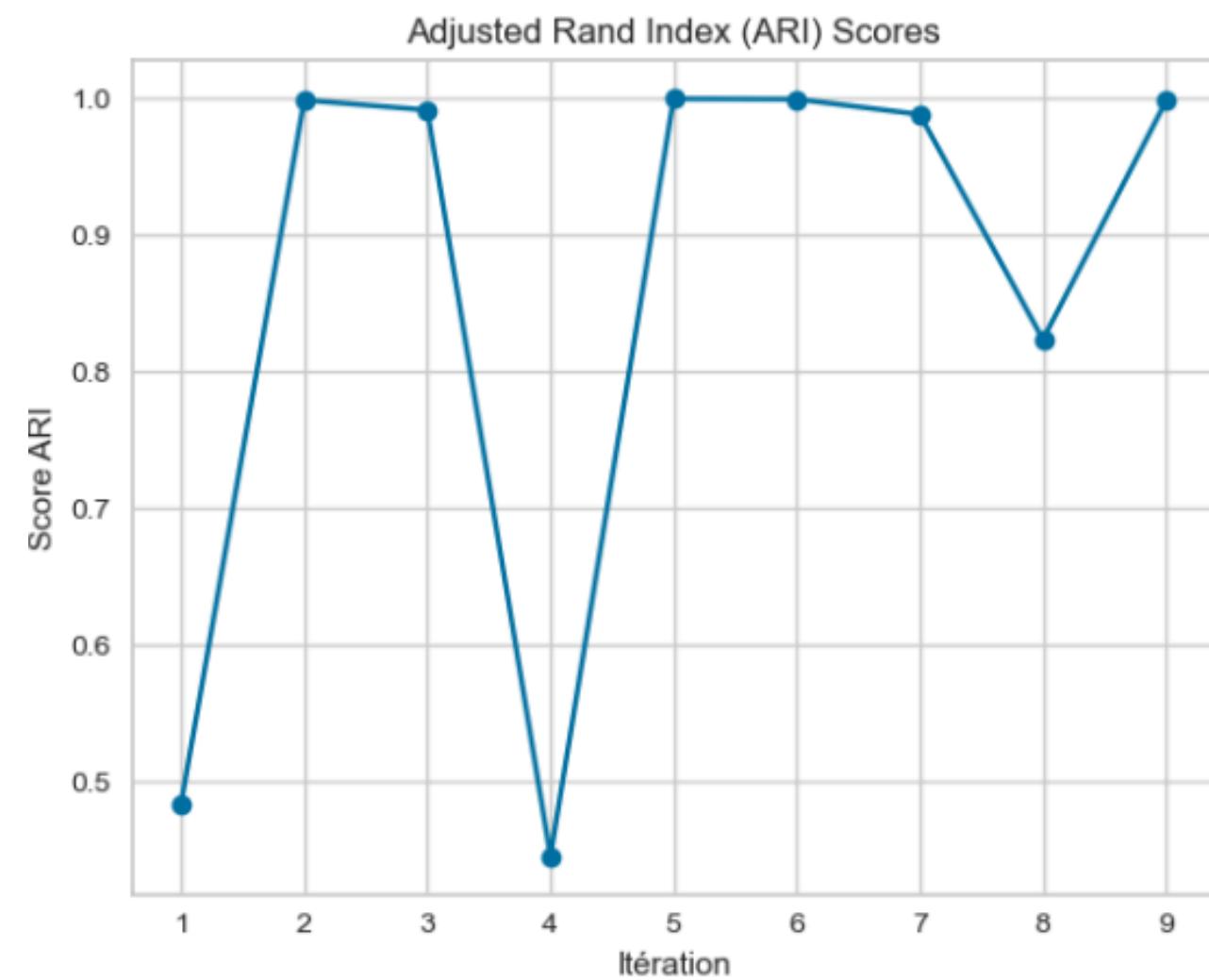
Baisse du score silhouette et peu d'informations ajoutées par la variables délai

Radar plot

Baisse du score silhouette
et peu d'informations
ajoutées par la variables
délai



Stabilité



Moyenne du score ARI : 0.86

Moyenne du score de silhouette : 0.34



Profil clients

Cluster 0 : À réactiver en raison de la faible fréquence et du faible score moyen des avis.

Cluster 1: Clients loyaux car ils ont un score moyen des avis très élevé et une récence faible.

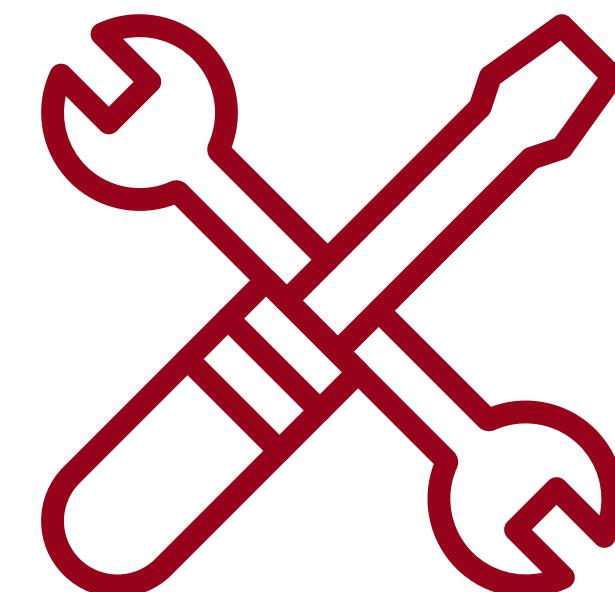
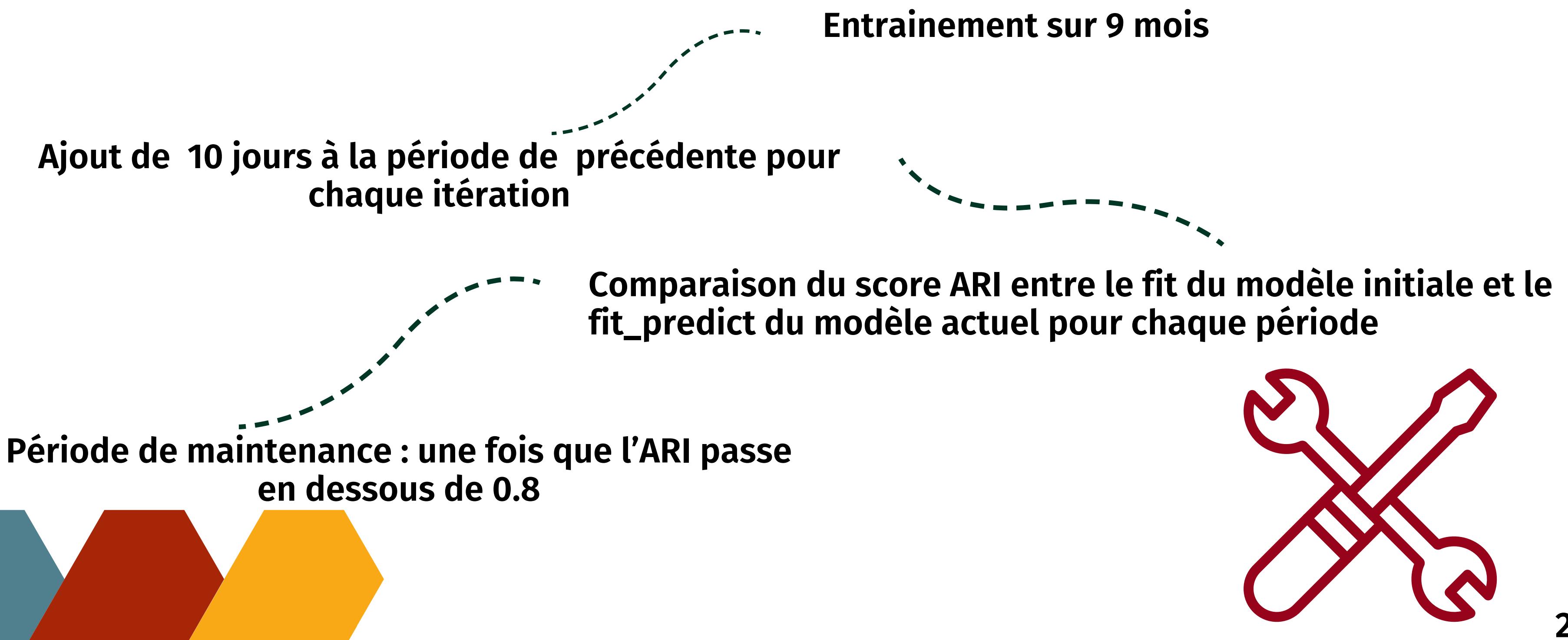
Cluster 2 : Clients perdus étant donné la haute récence.

Cluster 3 : Loyalistes potentiels en raison de la haute fréquence, du montant élevé mais des avis moyens.



Maintenance

Process



Maintenance

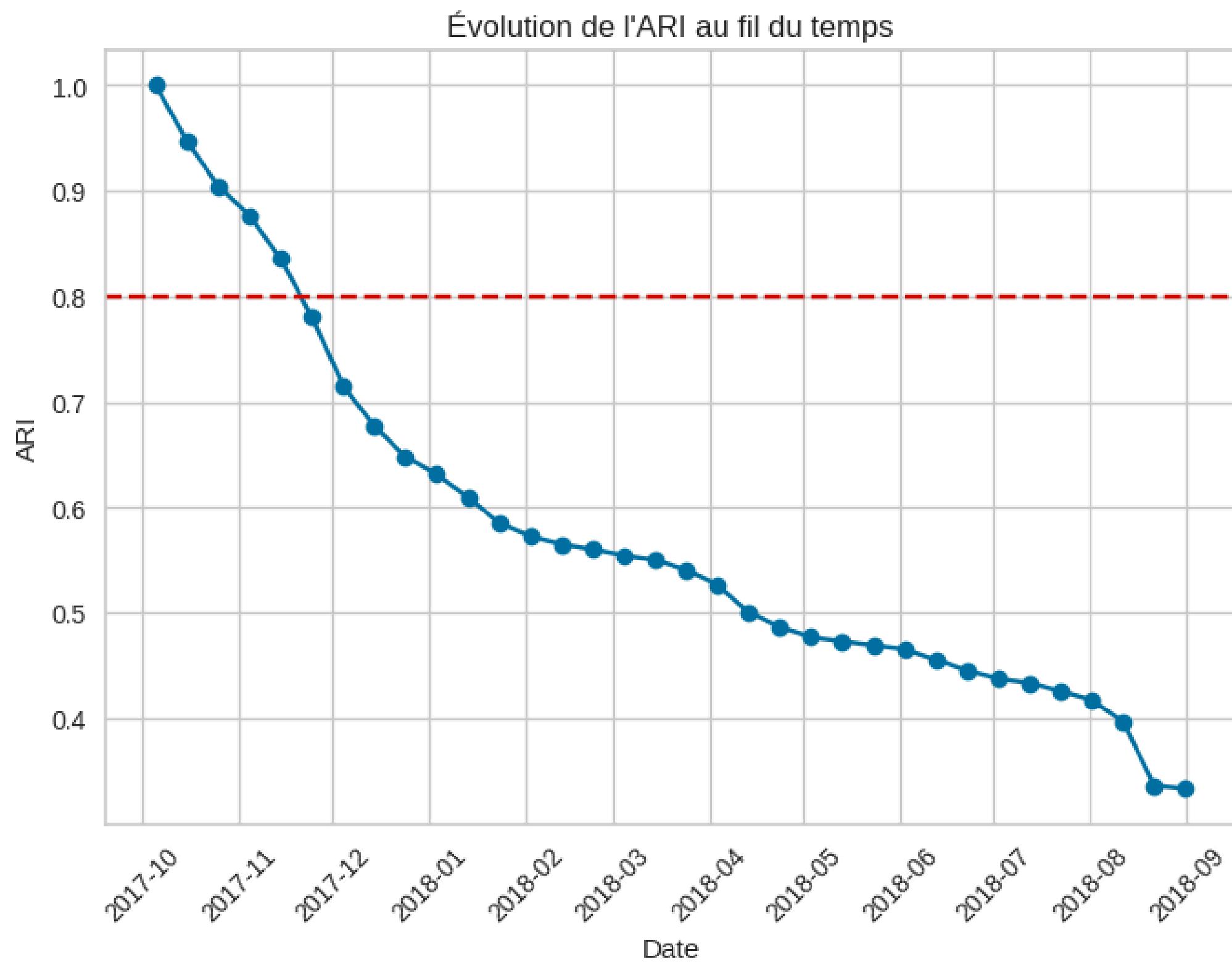
Résultats

Ensemble des valeurs

RFM avec ajout du review score moyen

RFM

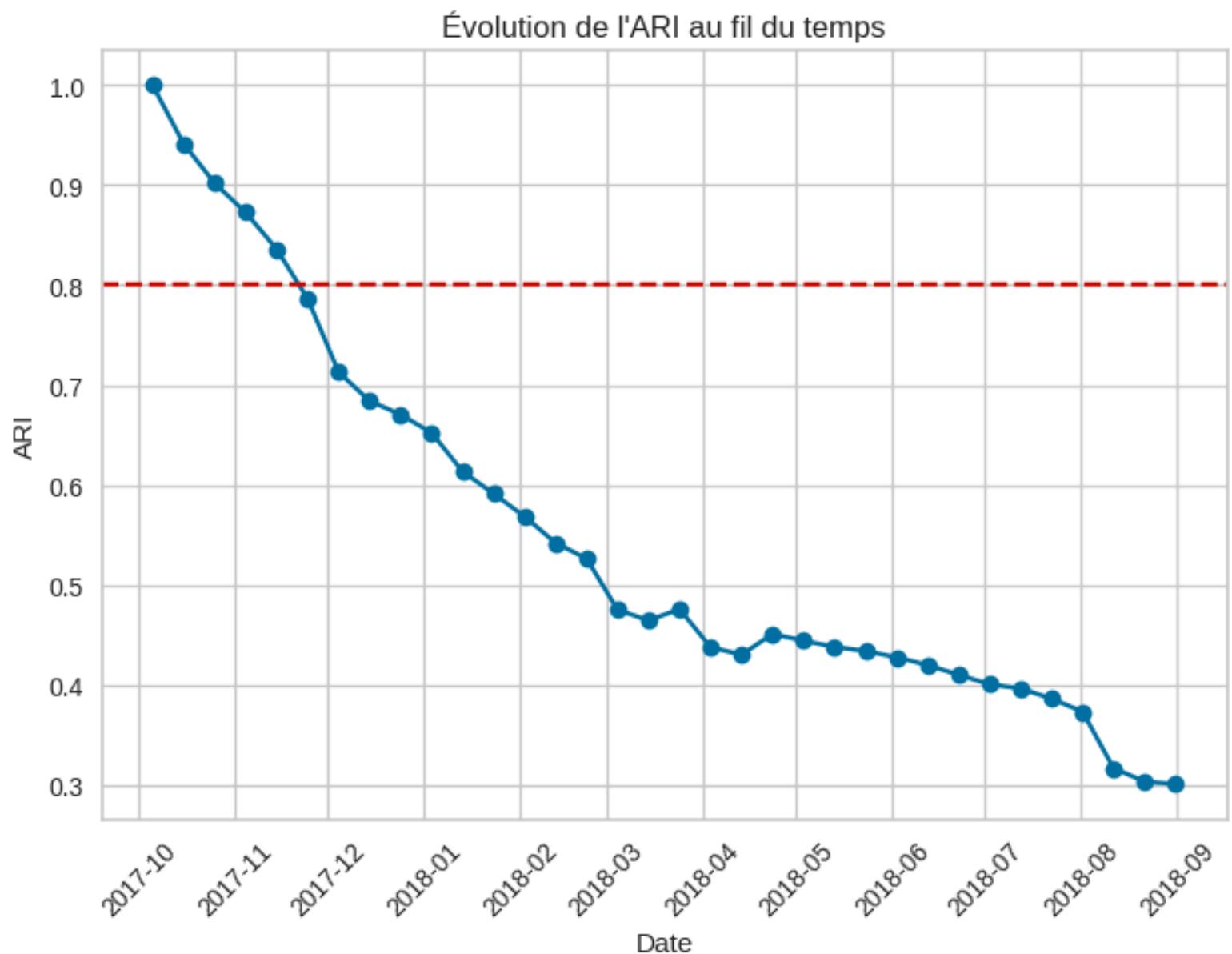
50 jours



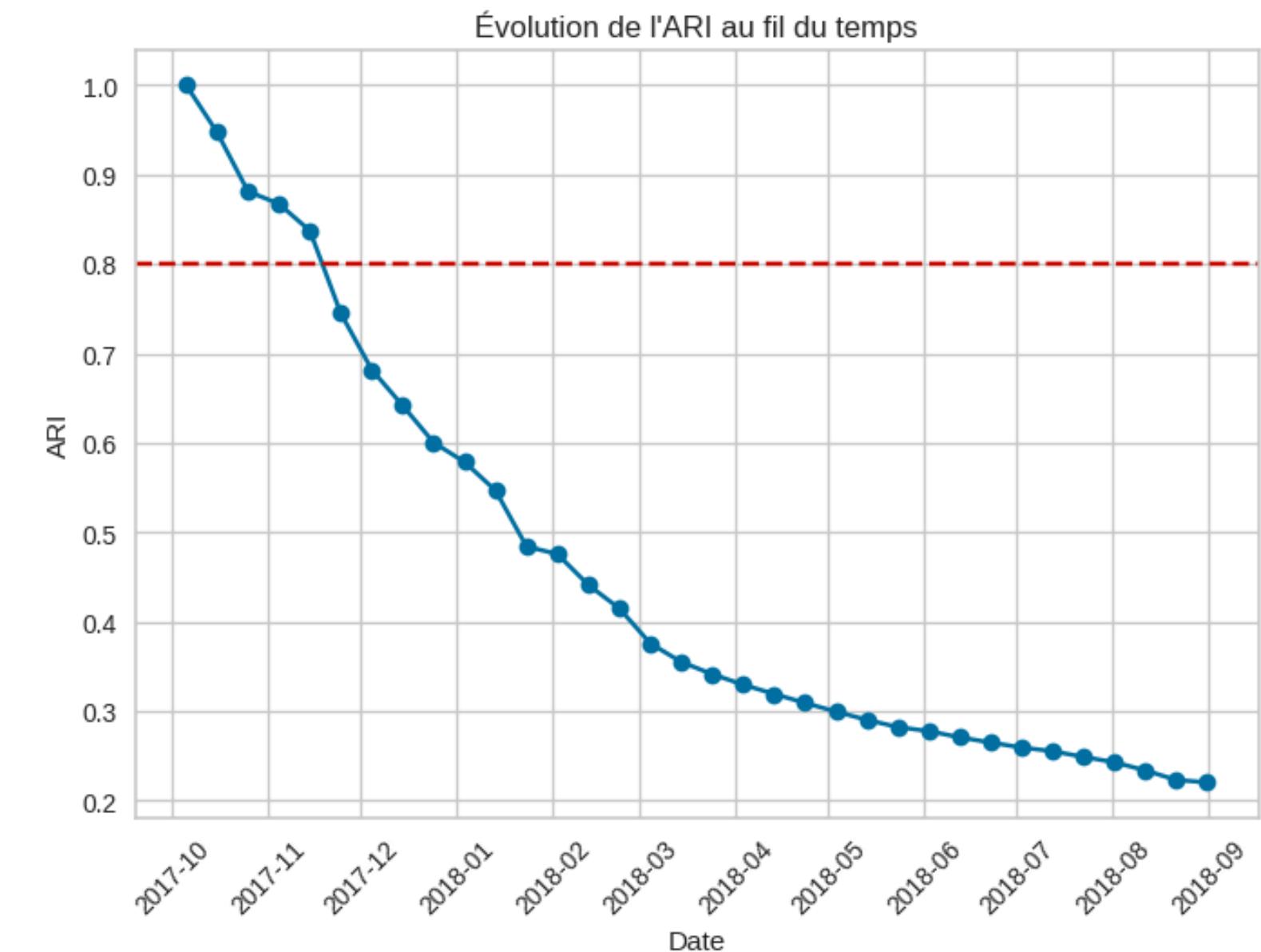
Evolution ARI ensemble des variables

Maintenance

Résultats



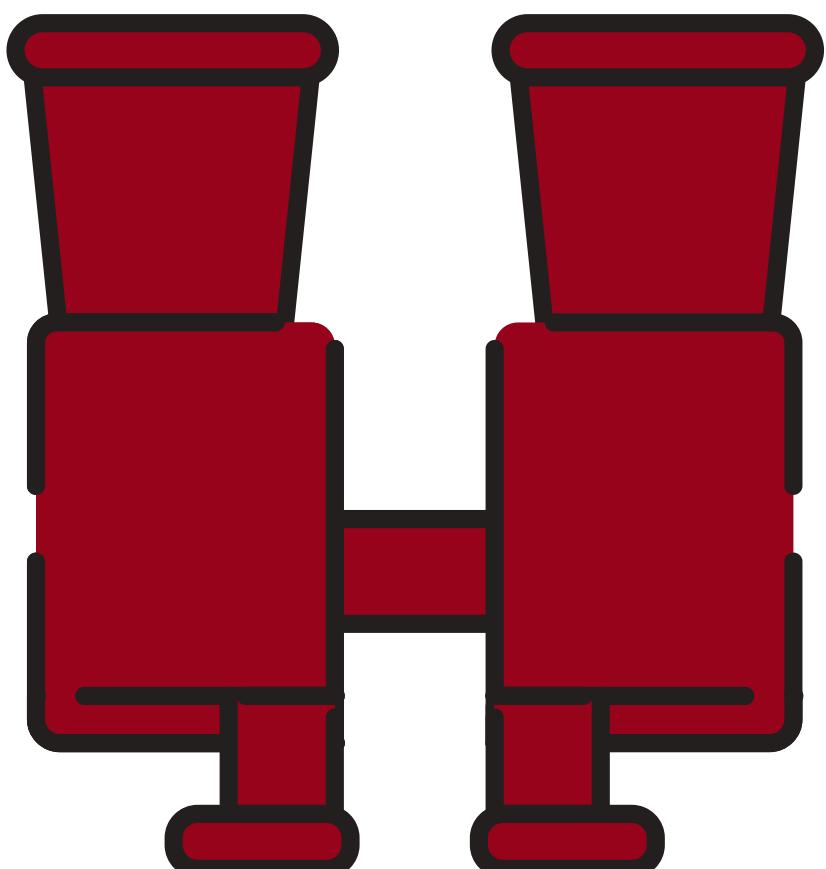
Evolution ARI + review score



Evolution ARI RFM



Conclusion



**Il est possible de mettre en place un système de clustering pour les clients de Olist en utilisant le Kmeans.
Aux variables classique RFM, nous sommes susceptibles d'ajouter les avis clients moyens induisant une légère perte de performance.**

Pour rester pertinent le modèle devra bénéficier d'une mise à jour tous les 50 jours.