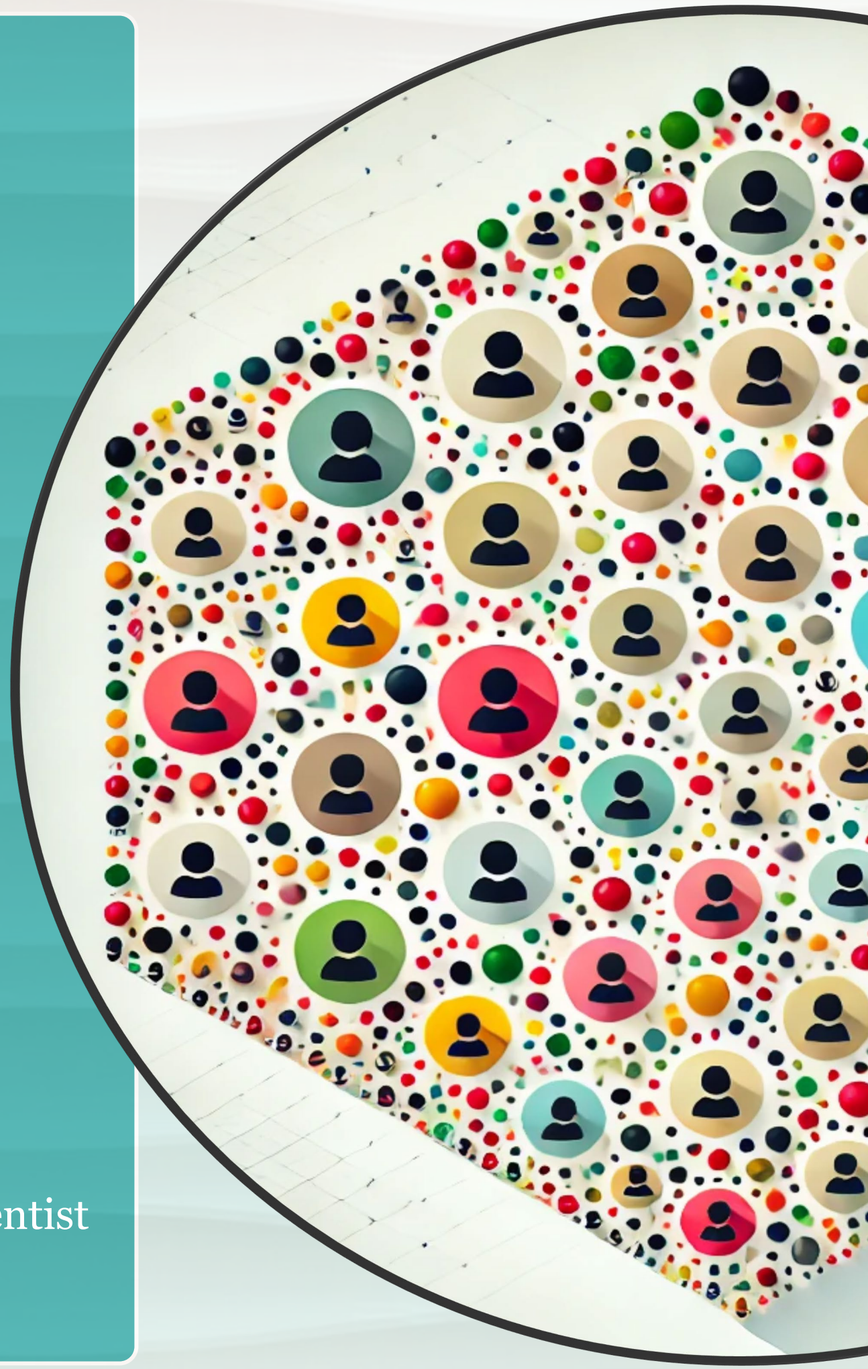
Projet_05 :

Segmentez des clients d’un site e-commerce

Presentation

Jérôme LE GAL
Etudiant OpenClassRooms – parcours Data Scientist
Le 21/09/2024



Contexte du projet :

- Mission pour Olist
- Analyse de ses clients

Objectifs de l’analyse :

- Segmenter les clients
- Compréhension des comportements

Description du jeu de données :

Base de données SQL anonymisée :

- Historique de commandes
- Produits achetés
- Commentaires
- Localisation

Dates : du 2016-09-04 au 2018-09-03
Période : 2 années

Problématique et import des données :

1. Objectifs
    - *Personnaliser les offres*
    - *Identifier les clients*
2. Définition des critères de segmentation
    - *Choix des features initiales*
    - *Feature engineering à prévoir*

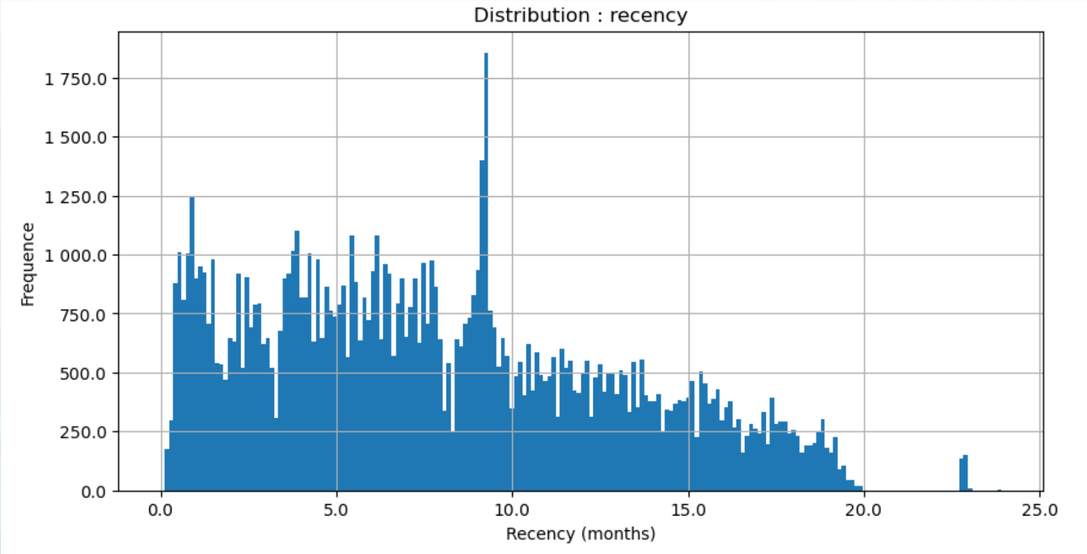
Feature Engineering :

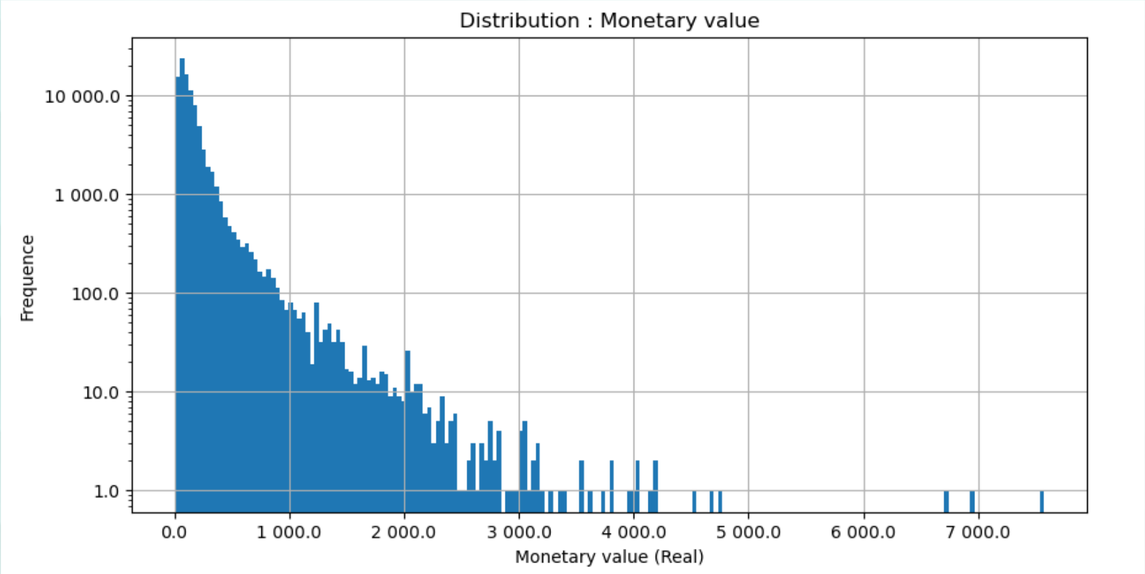
- Récence
- Fréquence
- Montant total
- Satisfaction

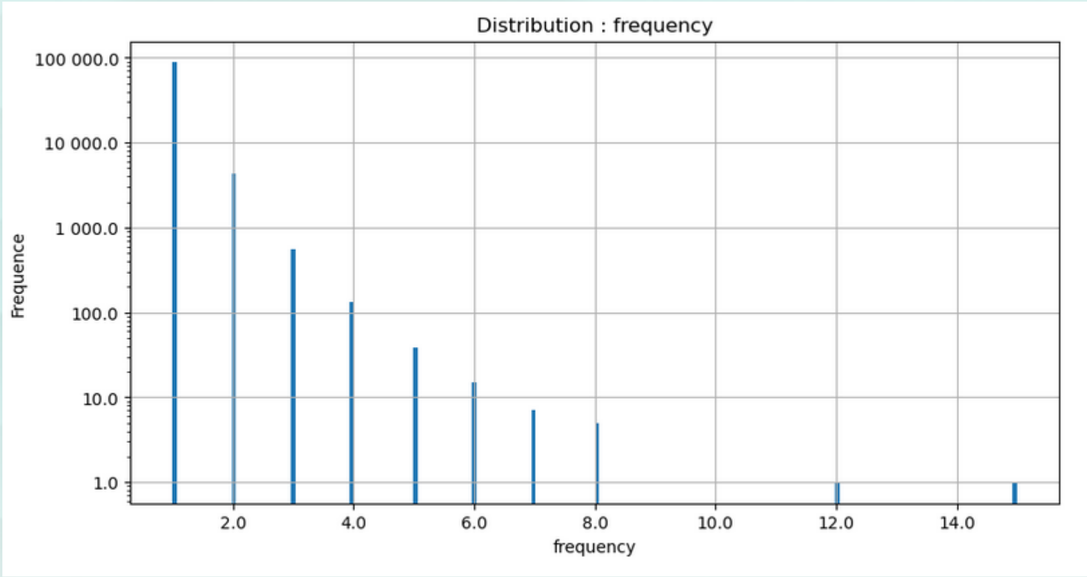
- Standardisation
- Transformation logarithmique

Analyse exploratoire des données :

- Récence
- Fréquence
- Montant







Stratégie de modélisation :

Répondre aux exigences d’Olist

Segmentation pertinente

- Forme des clusters
- Stabilité
- Cohérence métier

Nature des données

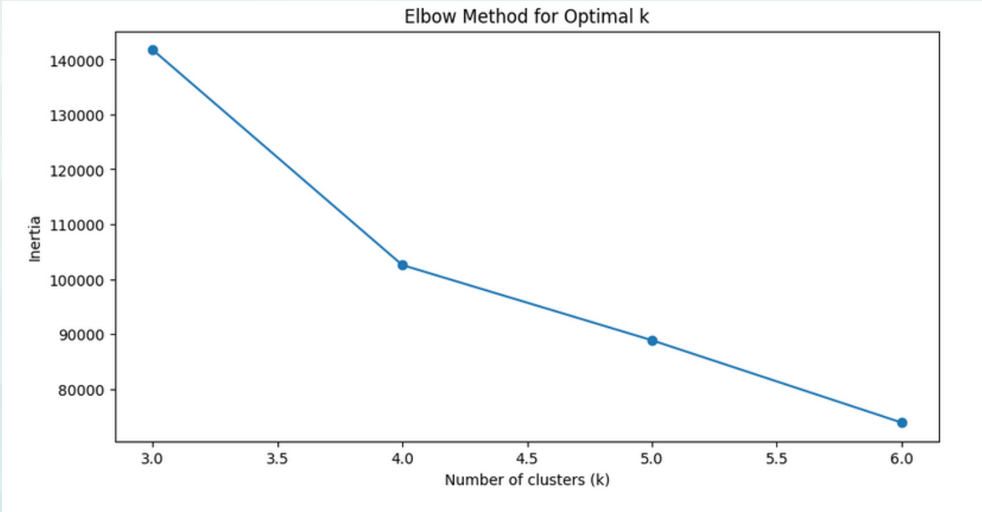
Modèles

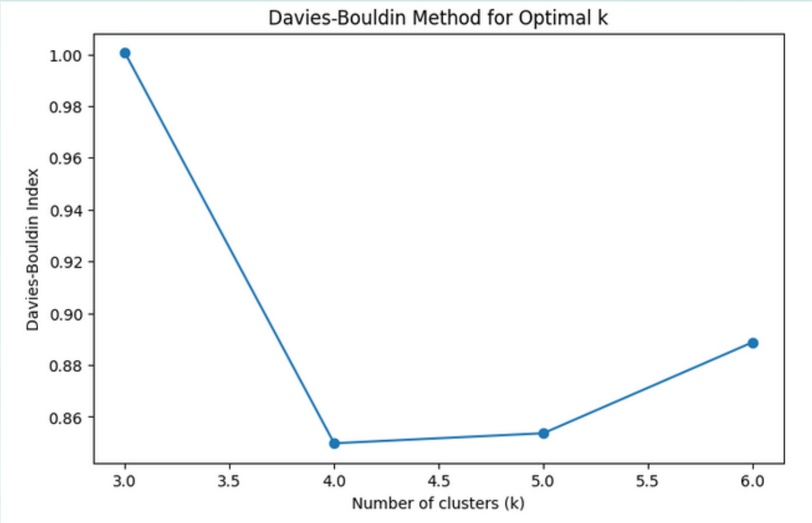
- Kmeans
- DBSCAN
- Agglomerative Clustering

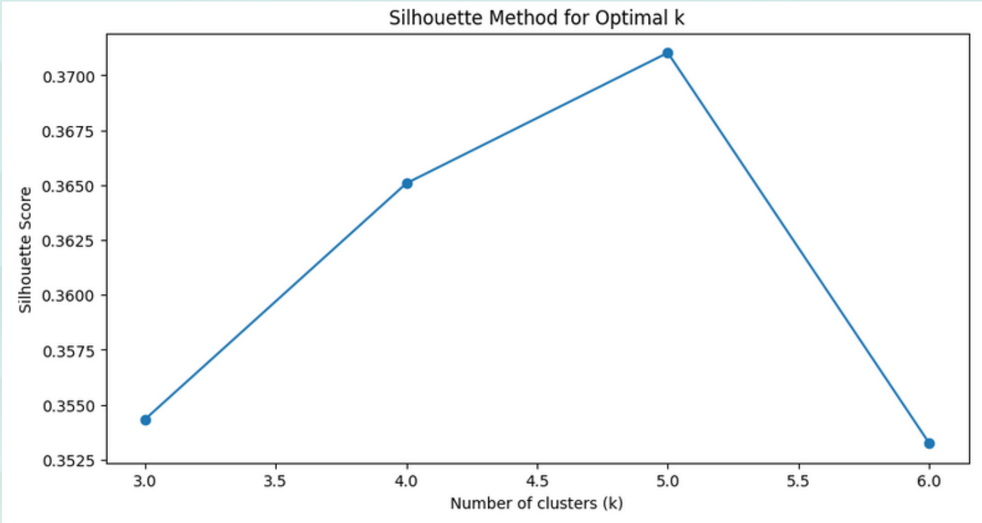
Recherche du nombre de cluster optimal :

Optimisation par les métriques de *k = 3 à k = 6* :

- Inertie (méthode du coude)
- Coefficient de silhouette (optimal proche de 1)
- Index de Davies-Bouldin (optimal au plus faible)







Modèle 1 – KMeans :

K = 4
Init = « k-means++ »
n_init=1
Random_state = 22

Nombre de clusters : 4

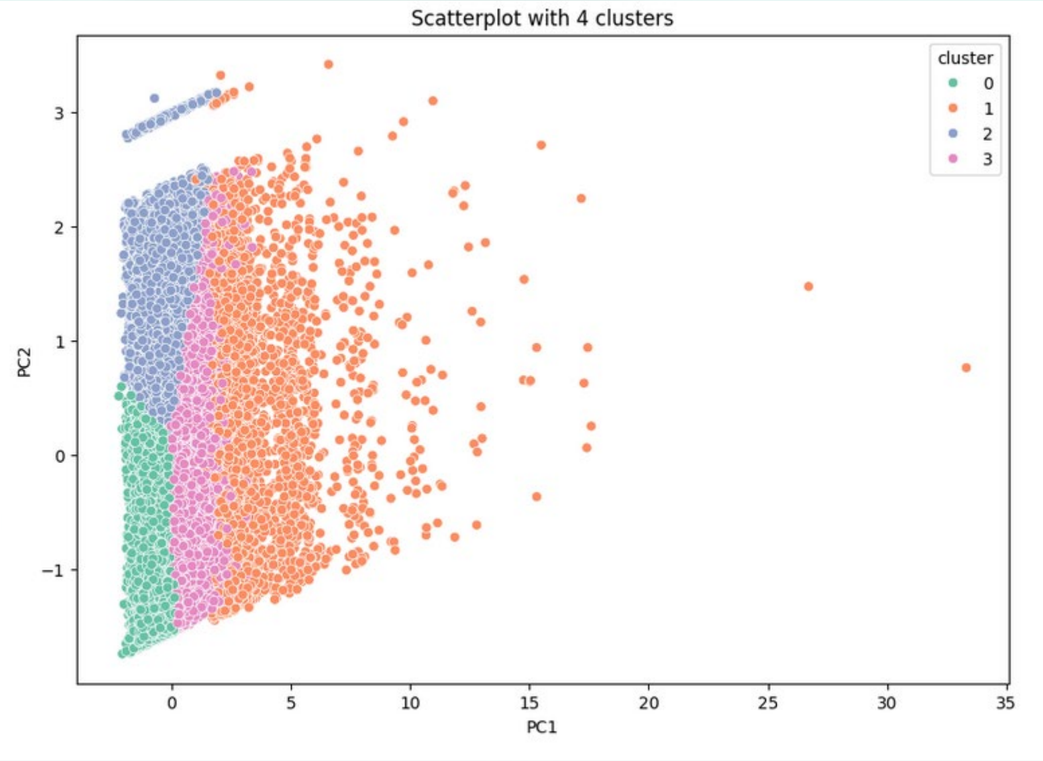
Nombre de points ‘noise’ : 0

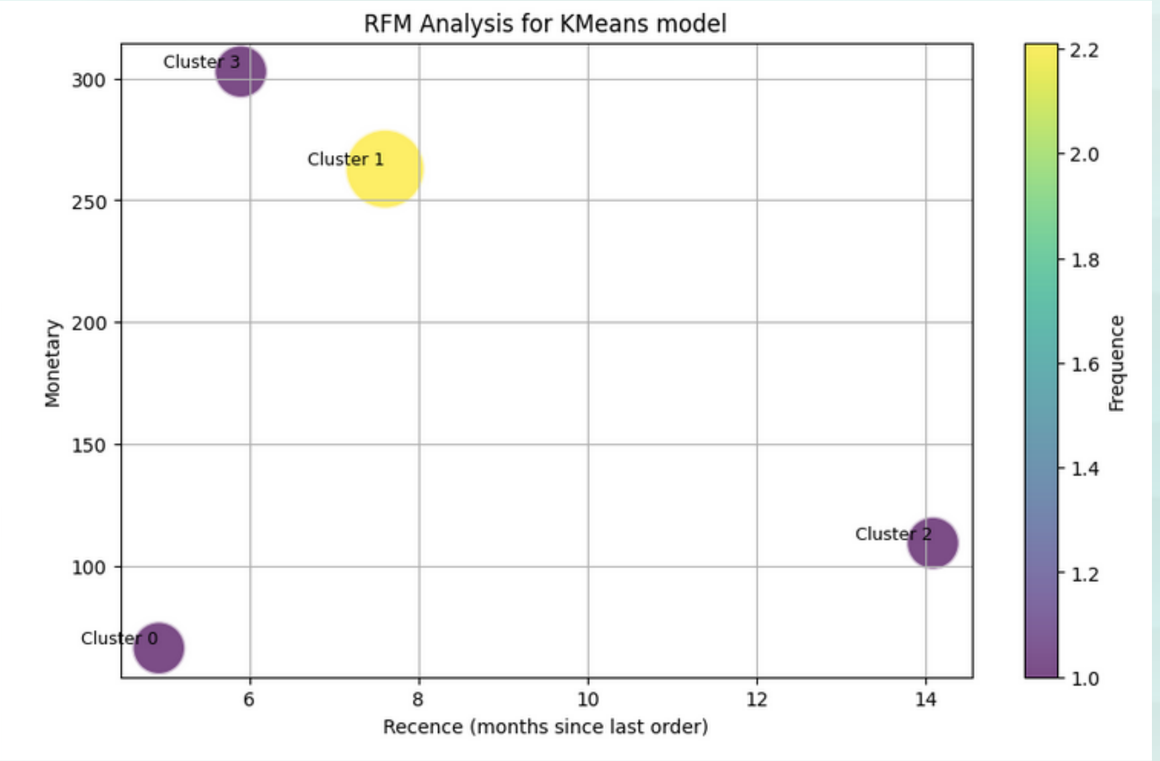
Coefficient de silhouette : 0.37

Index Davies-Bouldin : 0.85

Profiling des clusters KMeans :

- Cluster 0 (36049 pts) : "Nouveaux clients à faibles dépenses"
- Cluster 1(5111 pts)"Clients potentiellement fidèles"
- Cluster 2 (26473 pts) : "Clients inactifs/perdus"
- Cluster 3 (25763 pts) : "Clients réguliers et dépensiers."





Modèle 2 – DBSCAN :

epsilon = 1
min_sample = 30

Nombre de clusters : 4

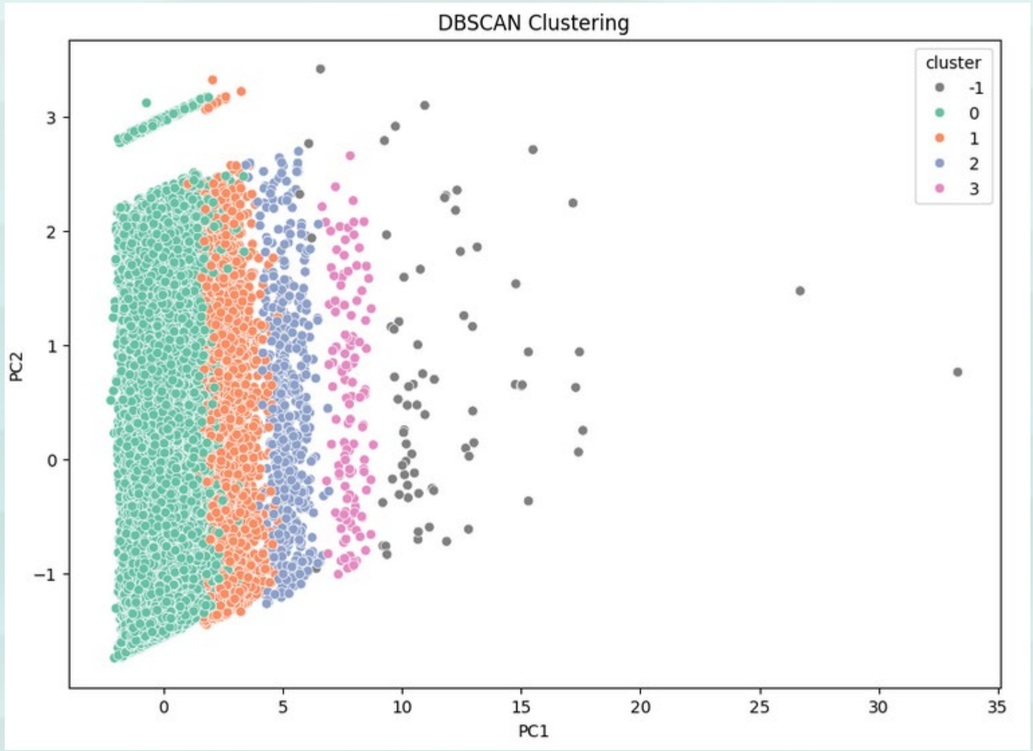
Nombre de points ‘noise’ : 75

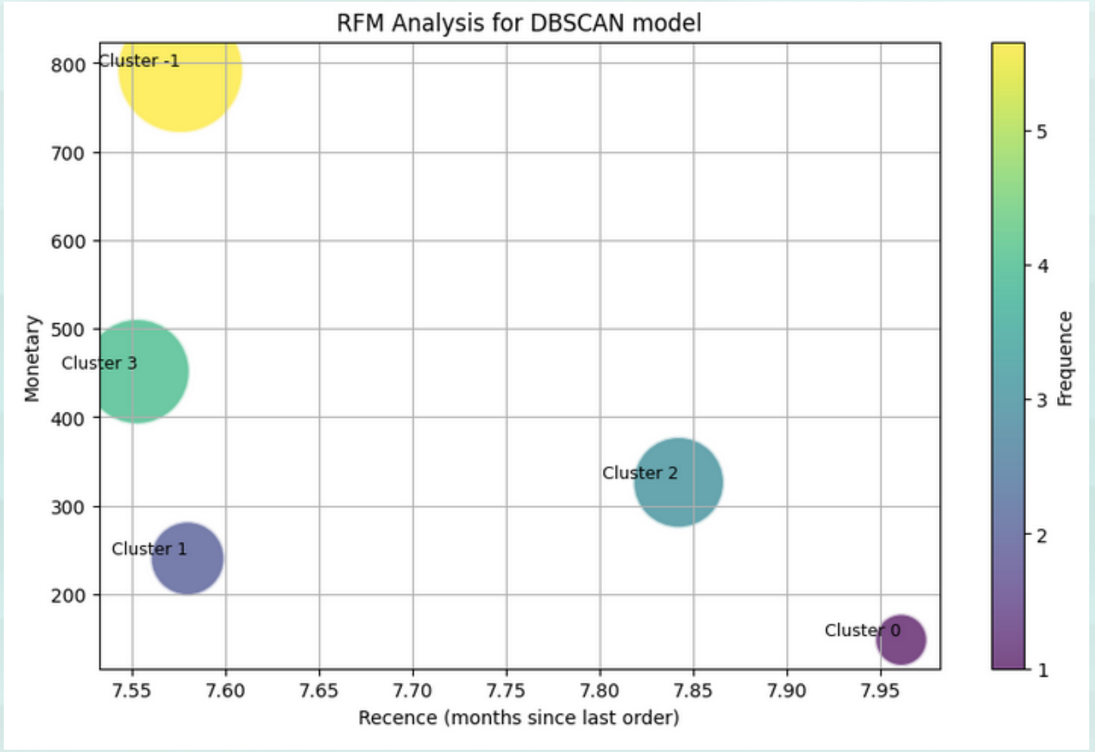
Coefficient de silhouette : 0.54

Index Davies-Bouldin : 0.81

Profiling des clusters DBSCAN :

- Cluster 0 (88285 pts) : « Clients ayant réalisé 1 commande»
- Cluster 1(4356 pts) : « Clients ayant réalisé 2 commandes»
- Cluster 2 (554 pts) : « Clients ayant réalisé 3 commandes»
- Cluster 3 (126 pts): « Clients ayant réalisé 4 commandes»
- NOISE(75 pts): « Clients ayant réalisé plus de 5 commandes»





Modèle 3 – Agglomerative Clustering :

n_clusters = 4
linkage = « single »

Nombre de clusters : 4

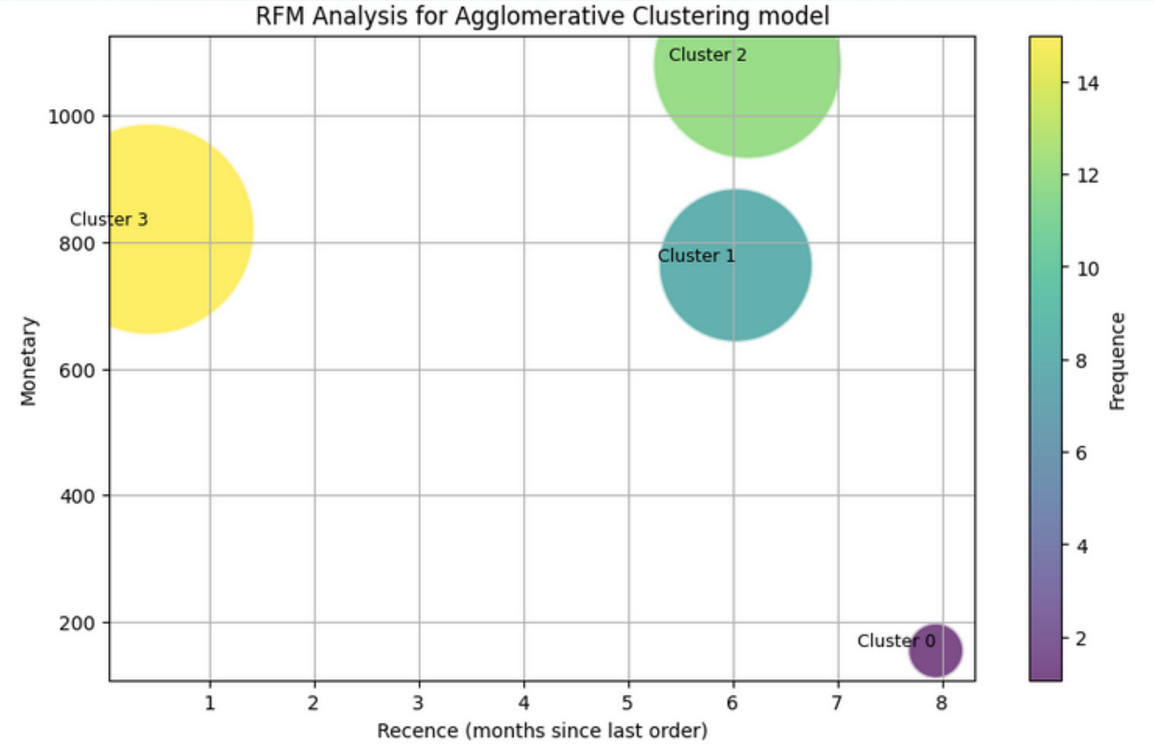
Nombre de points ‘noise’ : 0

Coefficient de silhouette : 0.91

Index Davies-Bouldin : 0.07

Profiling des clusters Agglomerative Clustering :

- Cluster 0 (93389 pts) : « Majorité des clients »
- Cluster 1(5 pts) : « Clients ayant réalisé 8 commandes»
- Cluster 2 (1 pt) : « Clients ayant réalisé 12 commandes»
- Cluster 3 (1 pt) : « Clients ayant réalisé 15 commandes»



Modèle Kmeans : RFM + satisfaction :

K = 4
Init = « k-means++ »
n_init=1
Random_state = 22

Nombre de clusters : 4

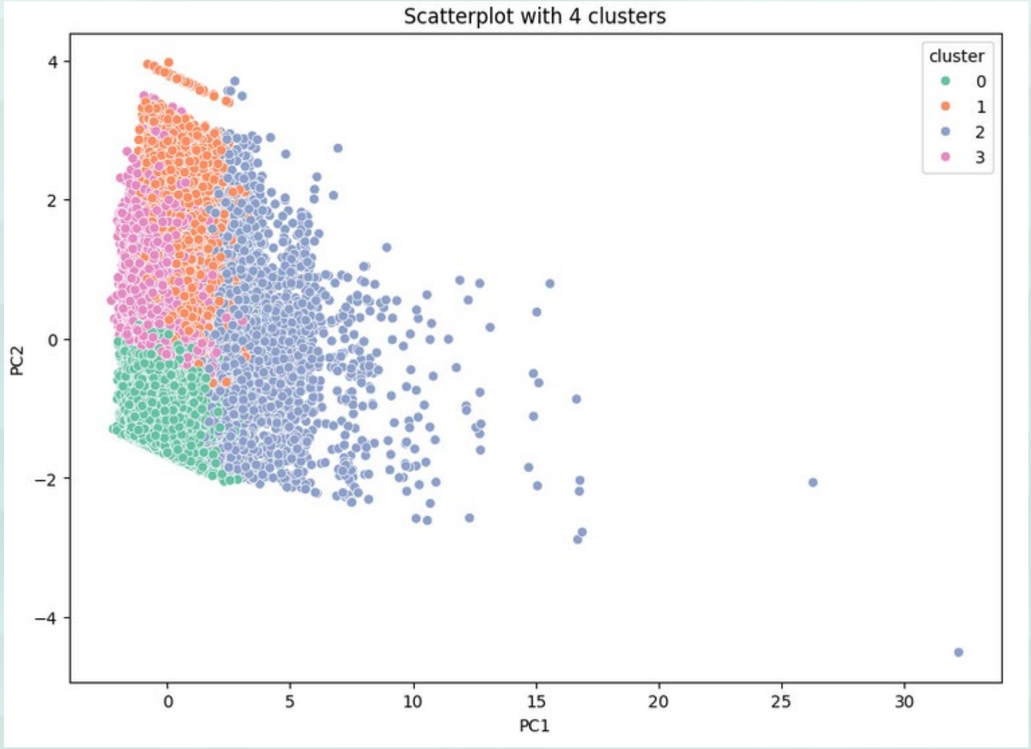
Nombre de points ‘noise’ : 0

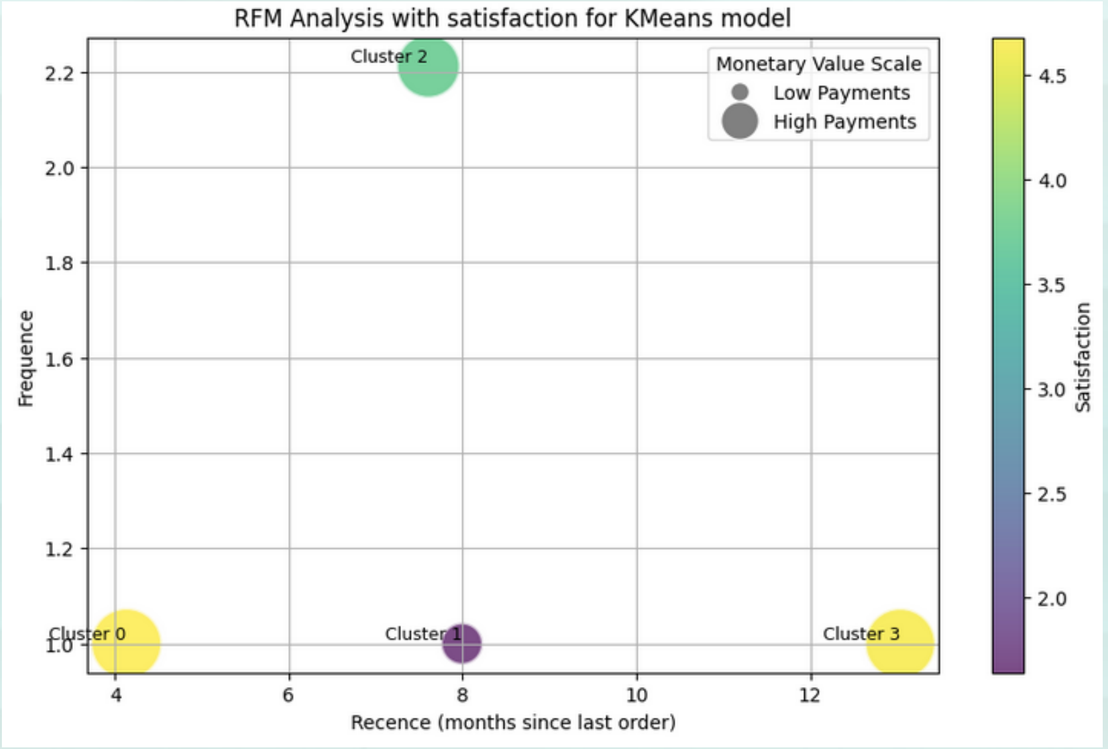
Coefficient de silhouette : 0.32

Index Davies-Bouldin : 1,06

Profiling des clusters KMeans :

- Cluster 0 (41450 pts): «Nouveaux clients vraiment satisfaits»
- Cluster 1(15675 pts): «Clients non fidèles et insatisfaits»
- Cluster 2 (5111 pts)«Clients fidèles mais modérément satisfaits»
- Cluster 3 (31160 pts) «Clients inactifs/perdus mais satisfaits»





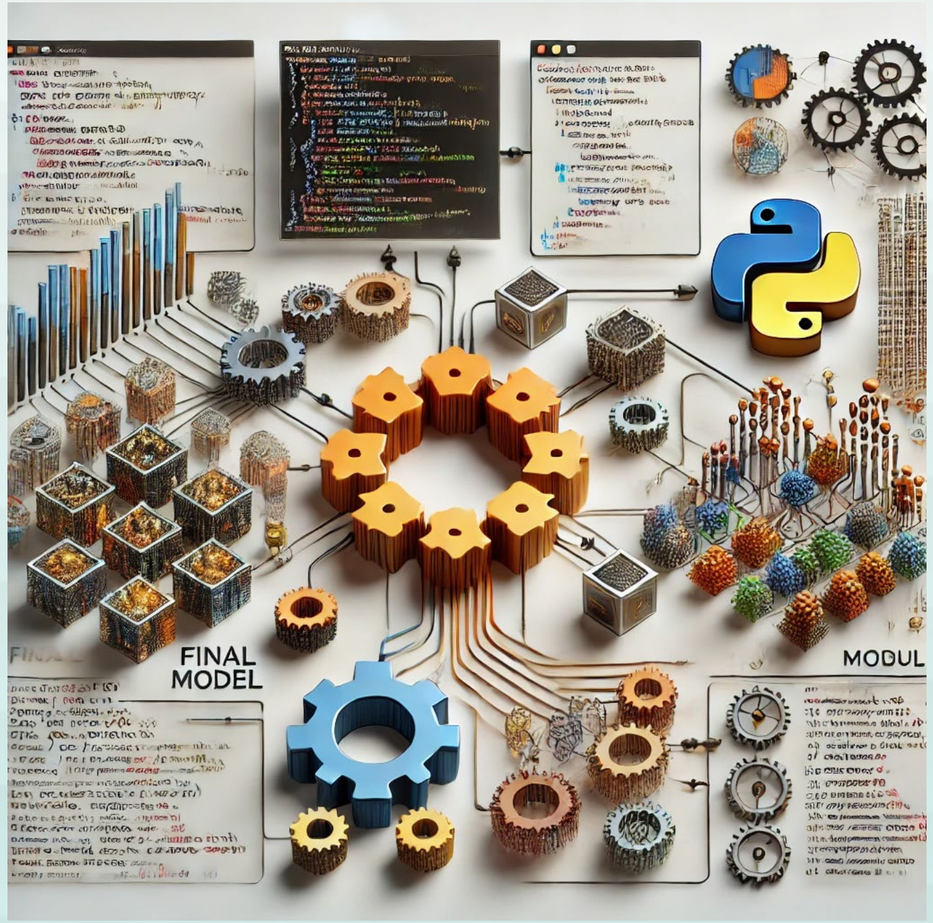
Choix modèle final et classe :

Modèle sélectionné : **KMeans RFM + Satisfaction**
k = 4

Nouveaux clients ?

Class OlistClustering(df) :

- clustering.get_labels()
- clustering.metrics()
- clustering.plot_2D()
- clustering.plot_correlation_circle()
- clustering.summary_clustering()
- clustering.interpretation_graph()



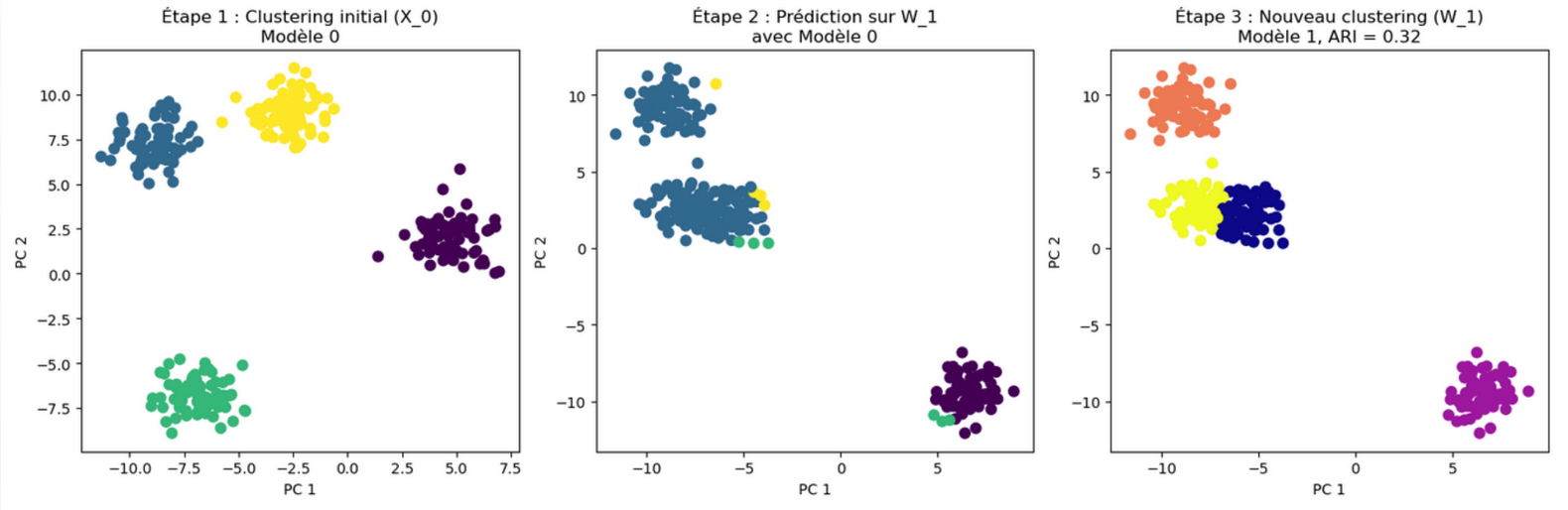
Contrat de maintenance :

Stratégie de simulation :

1. Suivre la stabilité des clusters
   - *Adjusted Rand Index (hebdomadaire)*
2. Analyse de l'évolution des distributions des features
   - *Tests de Kolmogorov-Smirnov (hebdomadaire)*
   - *Boxplots (mensuel)*
3. Recommandations de réévaluation périodique

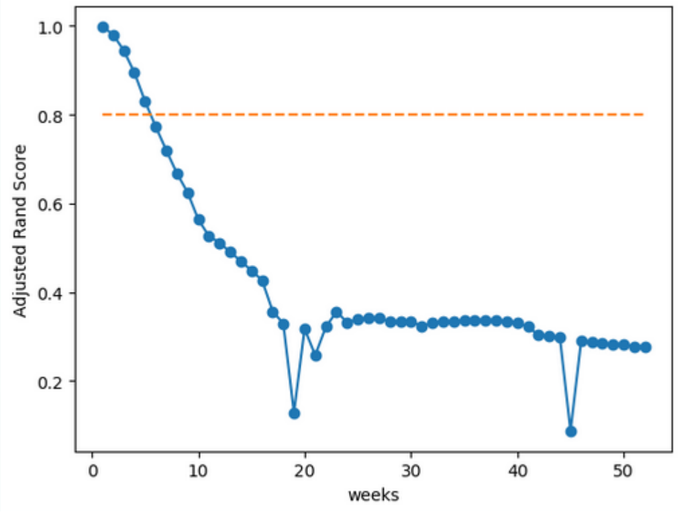
Stabilité des clusters :

Le concept de calcul avec **l'Adjusted Rand Index**



X_0 : données les plus récentes sur 1 an
W_1 : données d'une période d'1an glissée d'1 semaine / X_0
…
W_52 : données d'une période d'1an glissée de 52 semaines / X_0

Résultats sur 52 semaines :

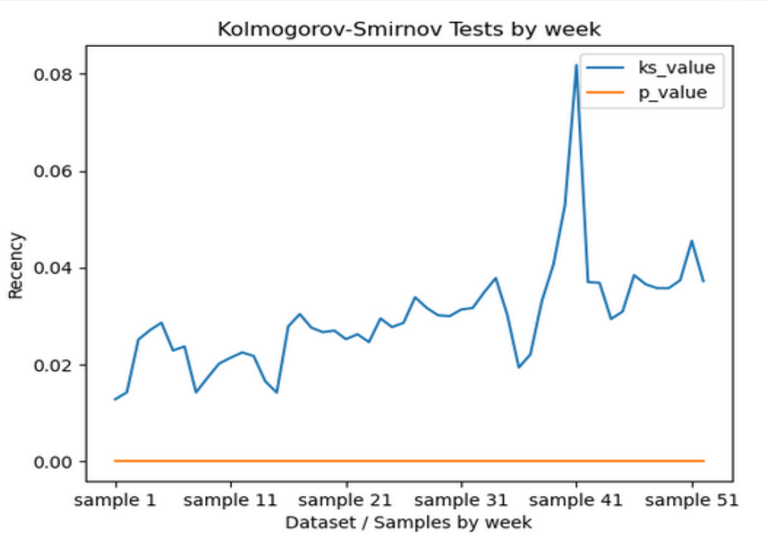


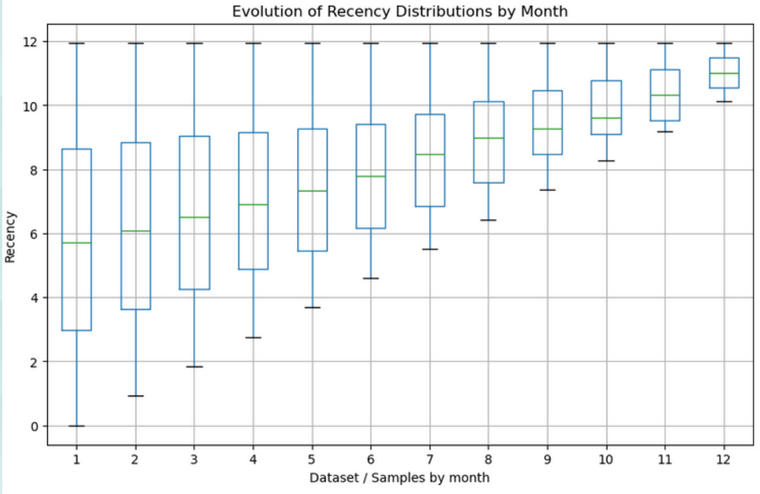
	week	ari_score	limit
0	1	0.998576	0.8
1	2	0.980462	0.8
2	3	0.944798	0.8
3	4	0.894912	0.8
4	5	0.830295	0.8
5	6	0.773432	0.8
6	7	0.719753	0.8
7	8	0.667857	0.8
8	9	0.625017	0.8
9	10	0.564480	0.8

- **Données** : 1 année glissante
- **Seuil ARI** : 0.8

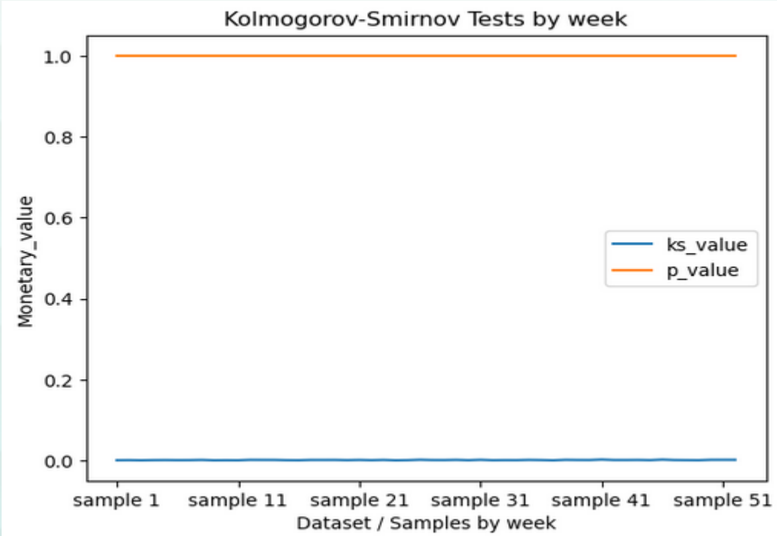
Evolution des distributions des features :

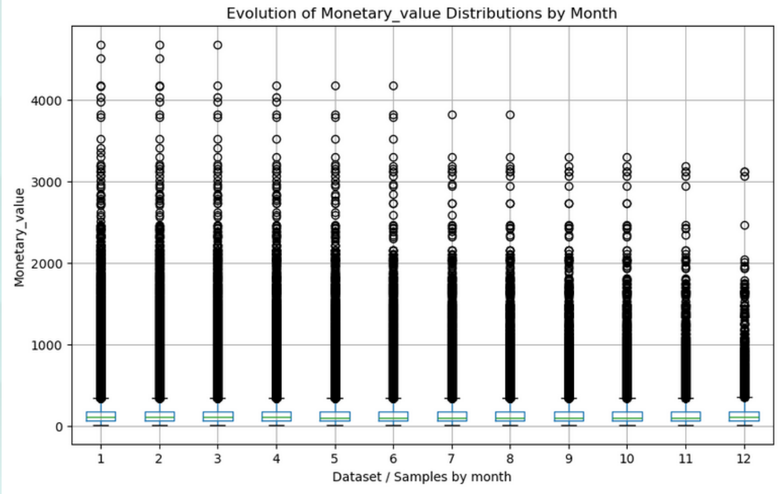
Tests de Kolmogorov-Smirnov :





Distributions :





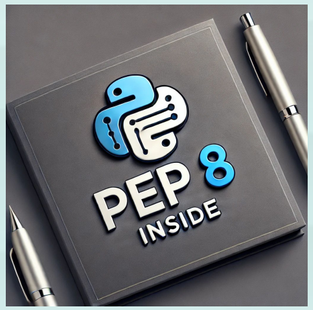
Recommandations :

Maintenance :

- **Réentrainement** : 5 semaines
- **Automatisation** : fourniture d'une classe

Conclusion :

- Clustering KMeans à 4 clusters
- RFM + satisfaction
- Réentrainement : 5 semaines



Questions ?