SOUTENANCE PROJET 5

Segmentez la clientèle d'un site d'e-commerce

GAUTHIER RAULT
PARCOURS DATA SCIENTIST
CHEZ OPENCLASSROOMS

EVALUATEUR
MONSIEUR JULIEN HEIDUK

"Les mégadonnées sonneront le glas de la segmentation de la clientèle et obligeront le spécialiste du marketing à comprendre chaque client en tant qu'individu dans les 18 mois ou risqueront d'être laissés pour compte."
Par Ginni Rometty

ORDRE DU JOUR

Ouverture de la problématique - 5 min

Modélisation des données - 10 min

Maintenance prévisionnelle - 5 min

Discussion - 5-10 min

OUVERTURE DE LA PROBLEMATIQUE

olist est une entreprise e-commerce (marketplace) Brésilienne. Elle souhaite fournir à ses équipes une segmentation de leur clientèle pour réaliser des actions marketing.

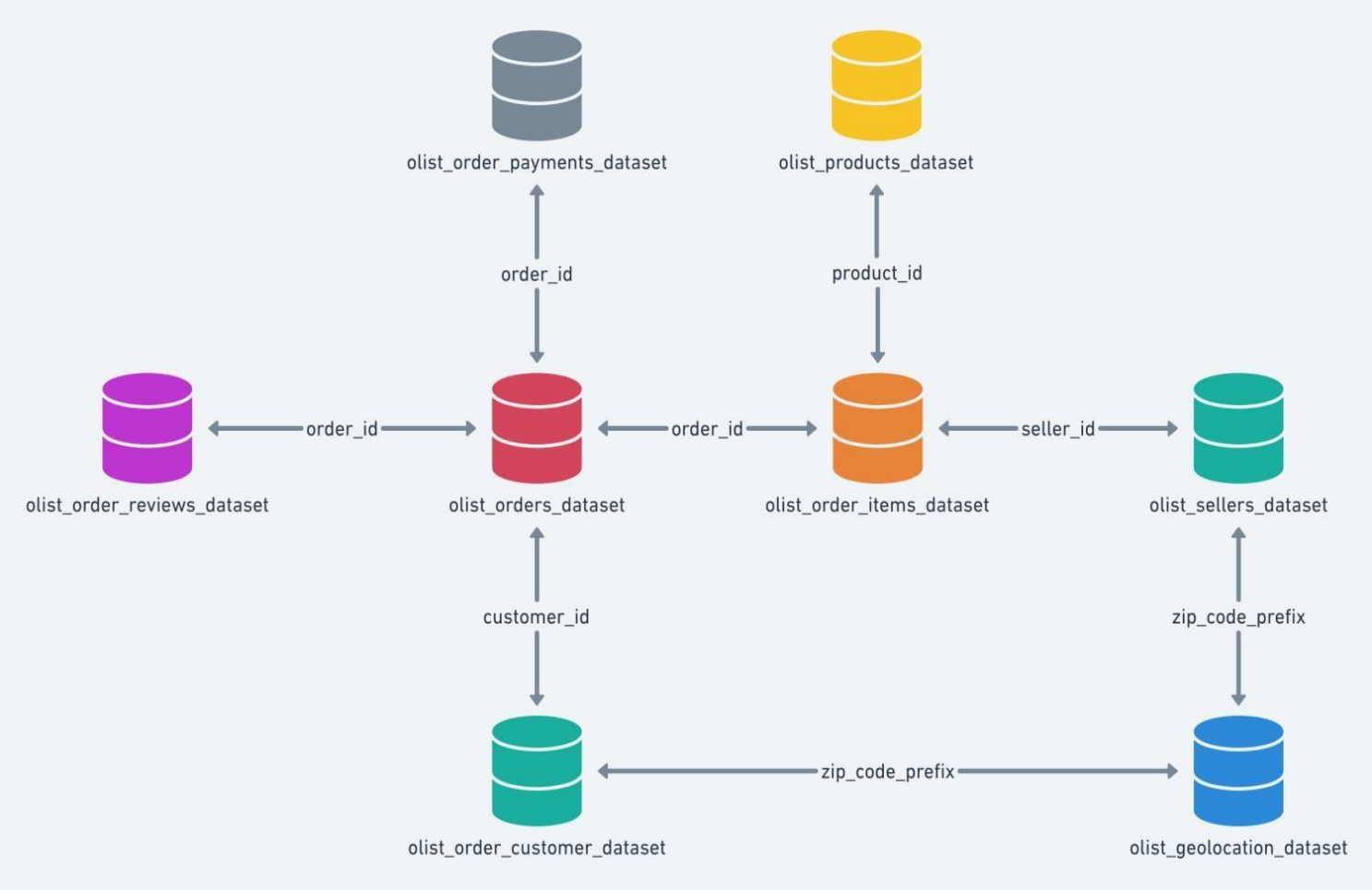
<u>Objectifs principaux:</u>

- - Comprendre ses différents utilisateurs
- - Réaliser une segmentation pertinente
- - Proposer un contrat de maintenance

<u>Moyens pour y parvenir:</u>

- - Mise à disposition d'une base de donnée
  - Premier jour de la BDD : 2016-09-15
  - Dernier jour de la BDD : 2018-08-29
- - Réalisation d'un clustering en testant différents algorithmes (Kmeans, DBscan, CAH)

OUVERTURE DE LA PROBLEMATIQUE



<u>9 datasets reliés par des variables clés</u>

Les données fournies par olist sont réparties dans plusieurs tables:

- **Les clients**
  Customers
  Geolocation

- **Les commandes**
  Orders
  Items
  Payments
  Reviews

- **Les produits**
  Products
  Categories_en

- **Les vendeurs**
  Sellers

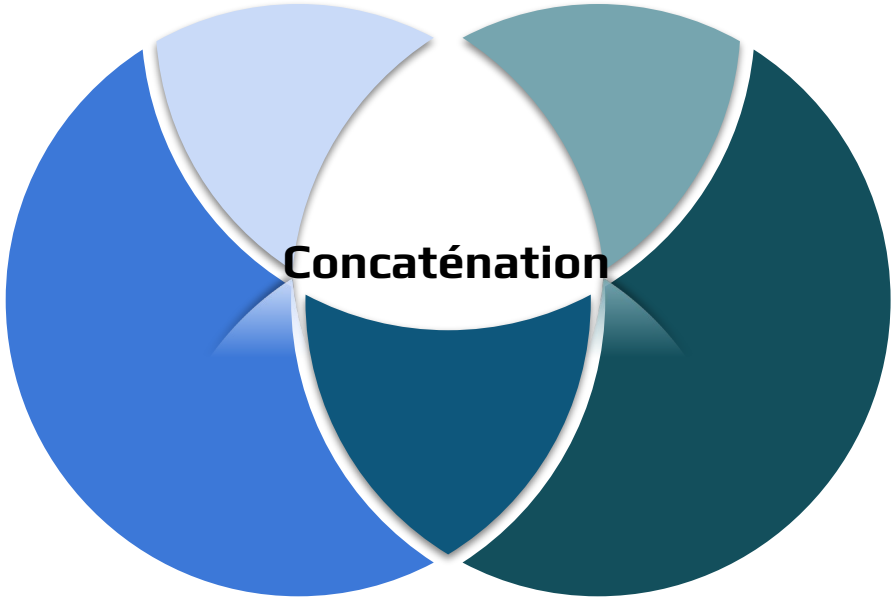
DESCRIPTION DES DONNÉES

Nom du fichier	Taille	Description	typologie
customers	99 441 x 5	Informations sur les clients (localisation et ID client)	0 duplicate 100% rempli 3/4 object 1/4 float
geolocation	1 000 163 x 5	Informations détaillées de localisation en fonction du code postal (latitude, longitude, ville, état)	261831 duplicates 100% rempli 1/3 object 1/3 float 1/3 int
orders	99 441 x 8	Informations sur les commandes (ID client, ID commande, statut, chronologie des étapes)	0 duplicates 99% rempli 100% object
category	71 x 2	Traduction des catégories de produits du portugais à l'anglais	0 duplicates 100% rempli 100% object
items	98 666 x 7	Table permettant d'associer ID commande, ID vendeur et ID produits, ainsi que des informations sur la commande (prix et date)	0 duplicates 100% rempli 2/4 object 1/4 float 1/4 int
products	32 951 x 9	Informations sur les produits (type, description, taille, etc.)	0 duplicates 99% rempli 1/4 object 3/4 float
reviews	99 224 x 7	Informations sur les évaluations des commandes (note, commentaires, date)	0 duplicates 78% rempli 3/4 object 1/4 int
payments	103 886 x 5	Informations sur le paiement des commandes (nombre de paiements, moyen utilisé, montant)	0 duplicates 100% rempli 1/3 object 1/3 float 1/3 int
sellers	3 095 x 4	Informations sur les vendeurs (localisation et ID vendeur)	0 duplicates 100% rempli 3/4 object 1/4 int

CONCATENATION DES DONNEES

Analyse et traitement des valeurs manquantes

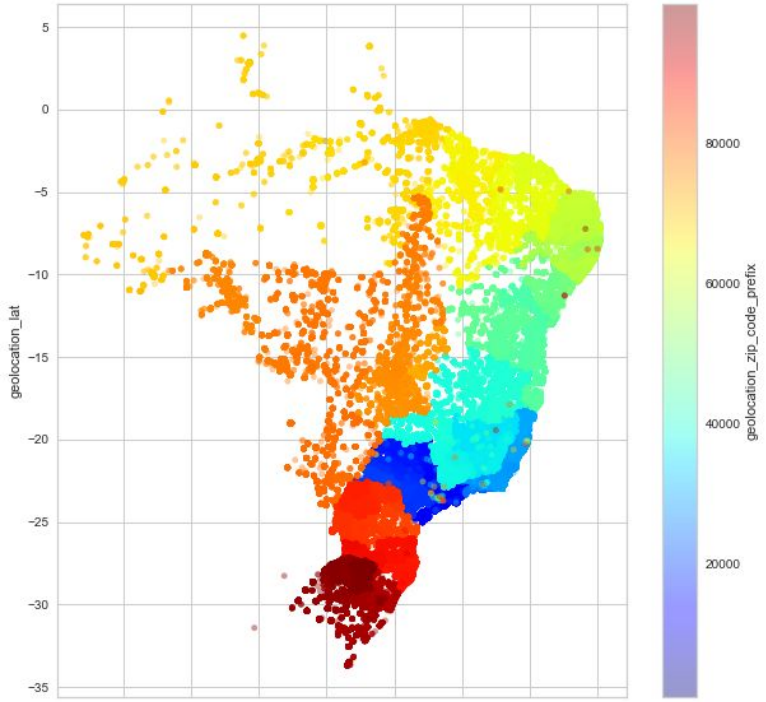
Correspondance Renommage, des colonnes

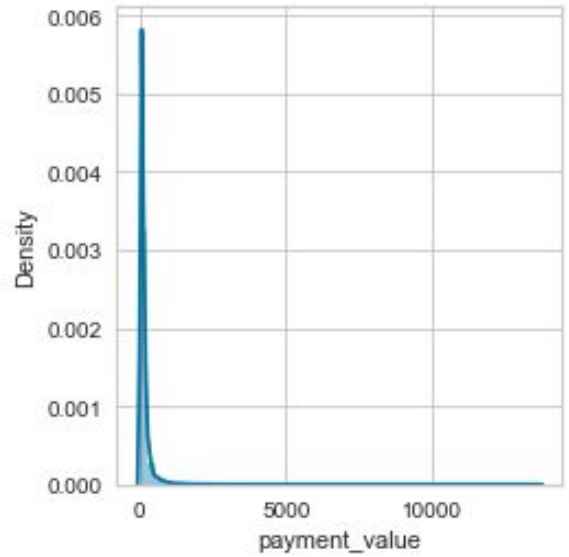


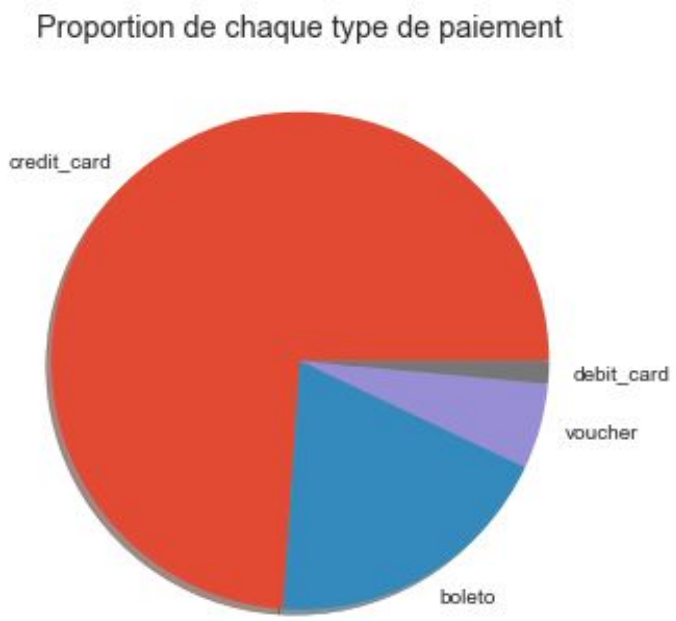
Transformation variable temps

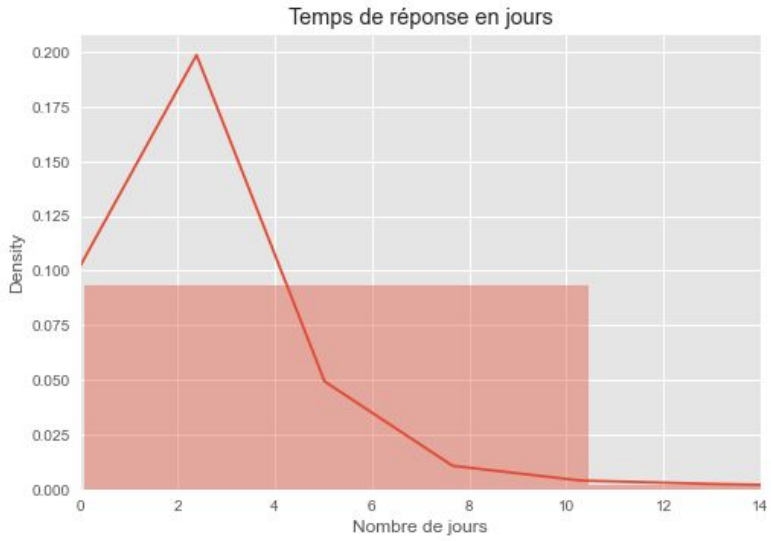
Vérification de la cohérence

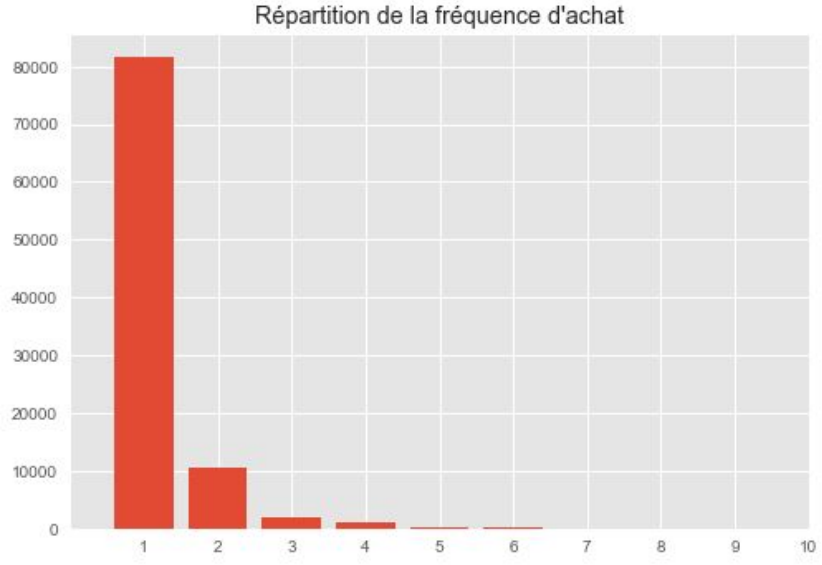
EXPLORATION DES DONNÉES



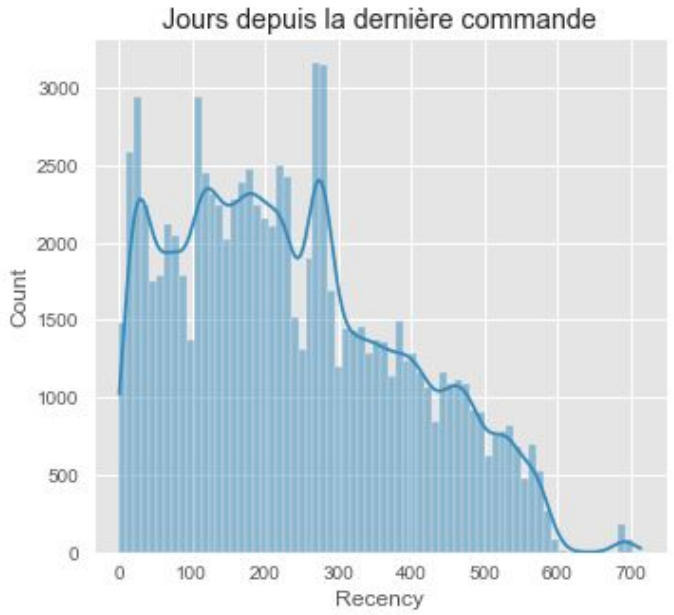


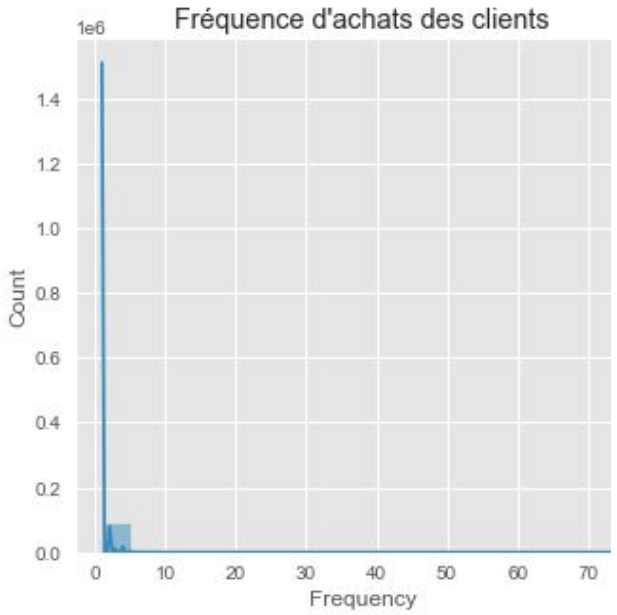


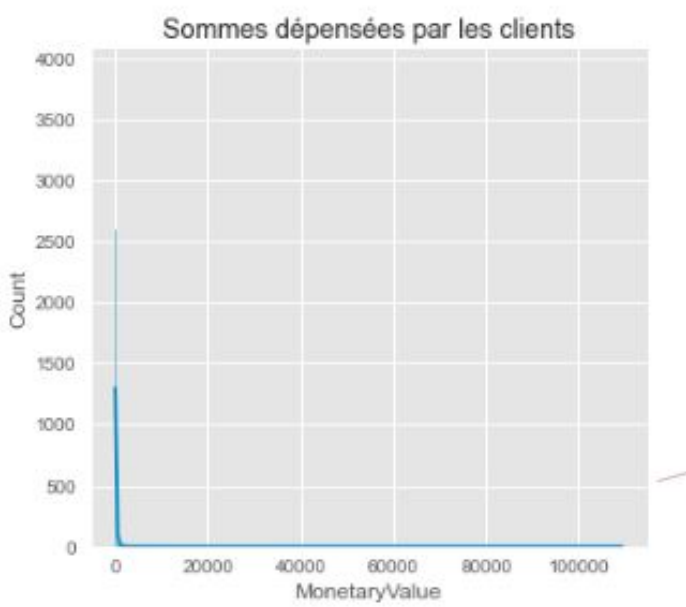


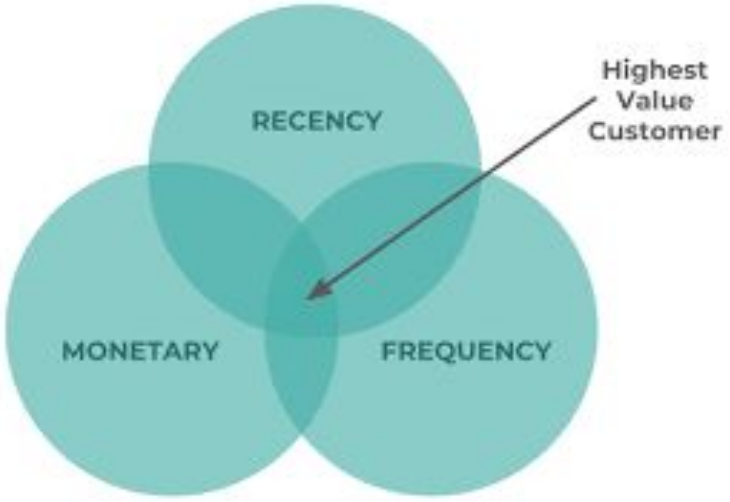


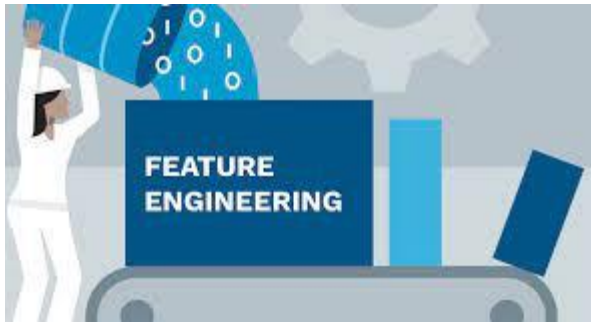
CREATION DE LA SEGMENTATION RFM
(RÉCENCE, FRÉQUENCE, MONTANT)



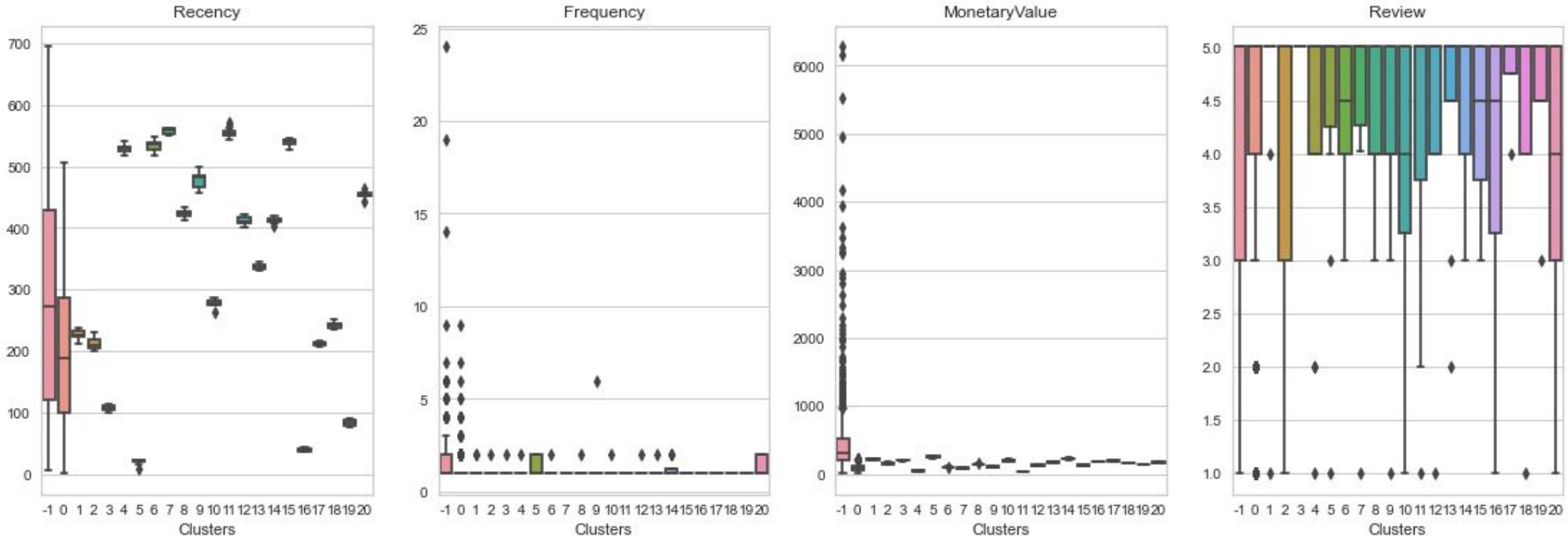




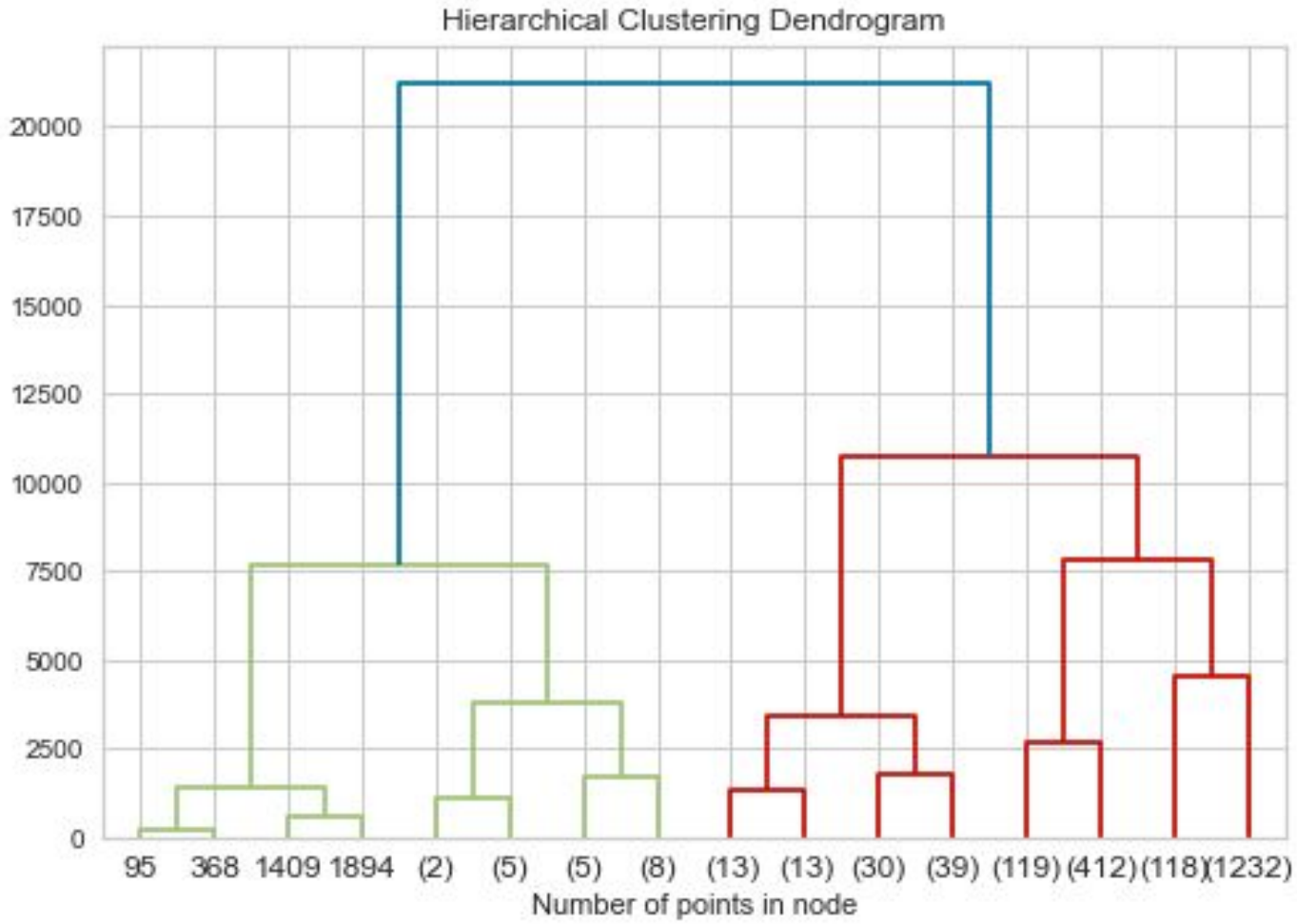




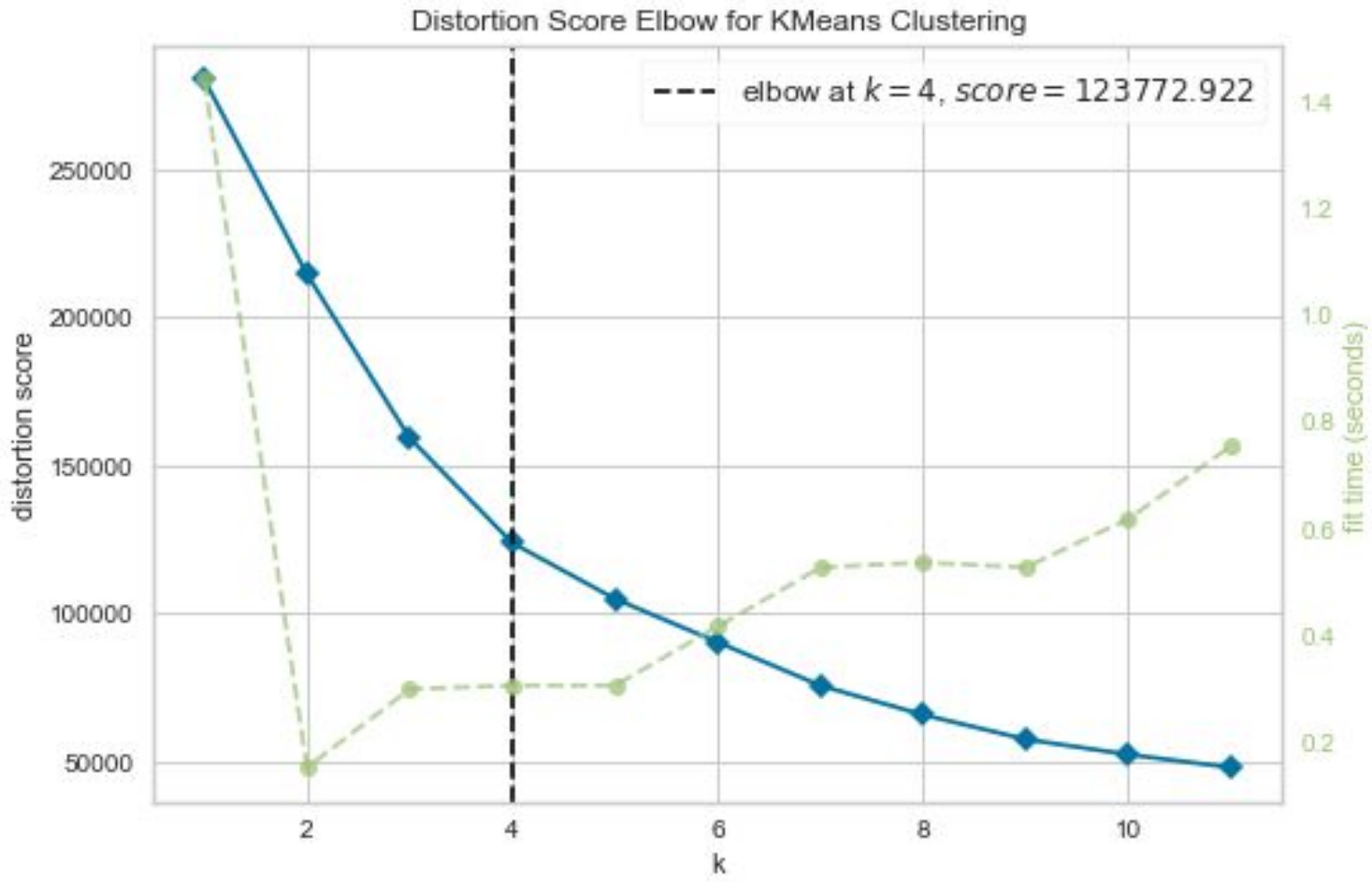
TEST DE DBSCAN SUR SAMPLE DE 2000 LIGNES



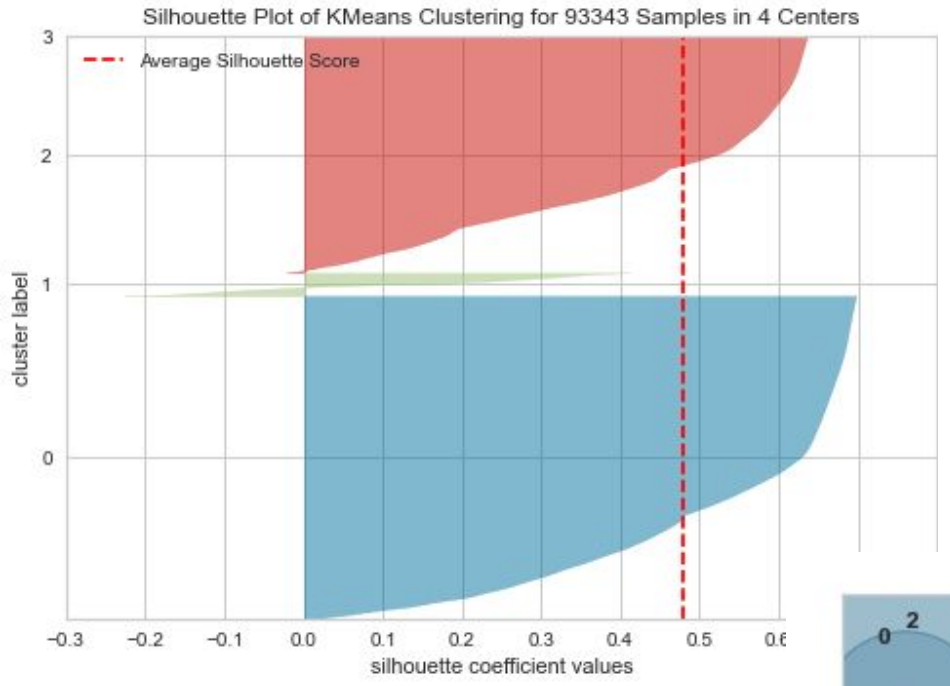
TEST DE **AGGLOMERATIVE CLUSTERING - CAH**
SUR SAMPLE DE 2000 LIGNES

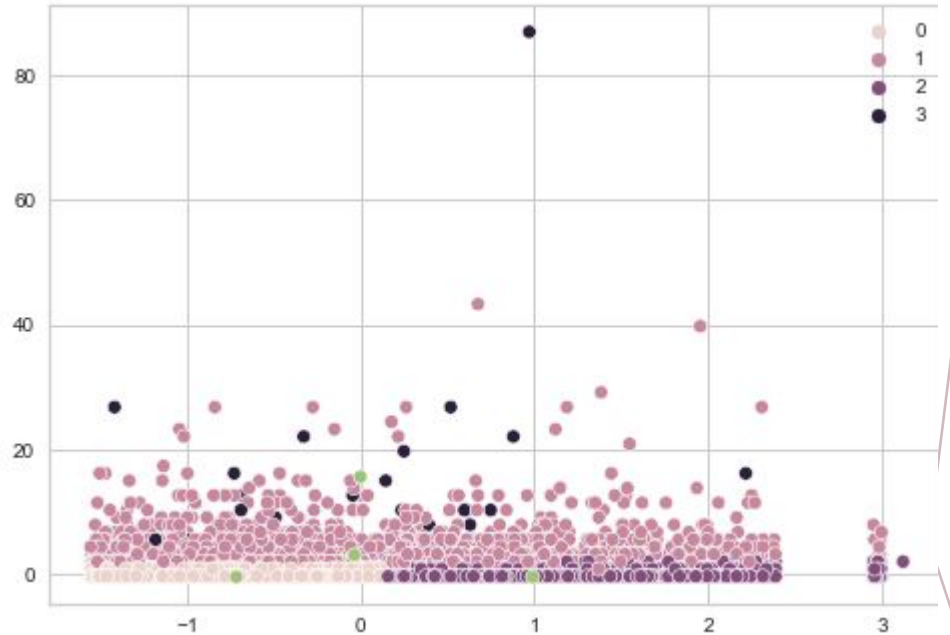


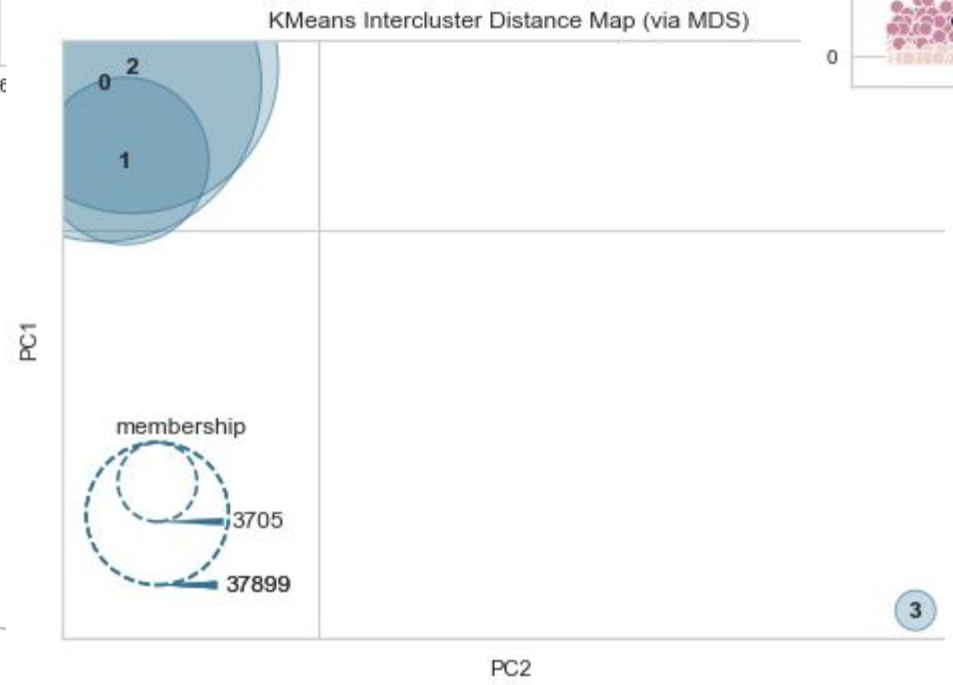
KMEANS SUR LA SEGMENTATION RFM



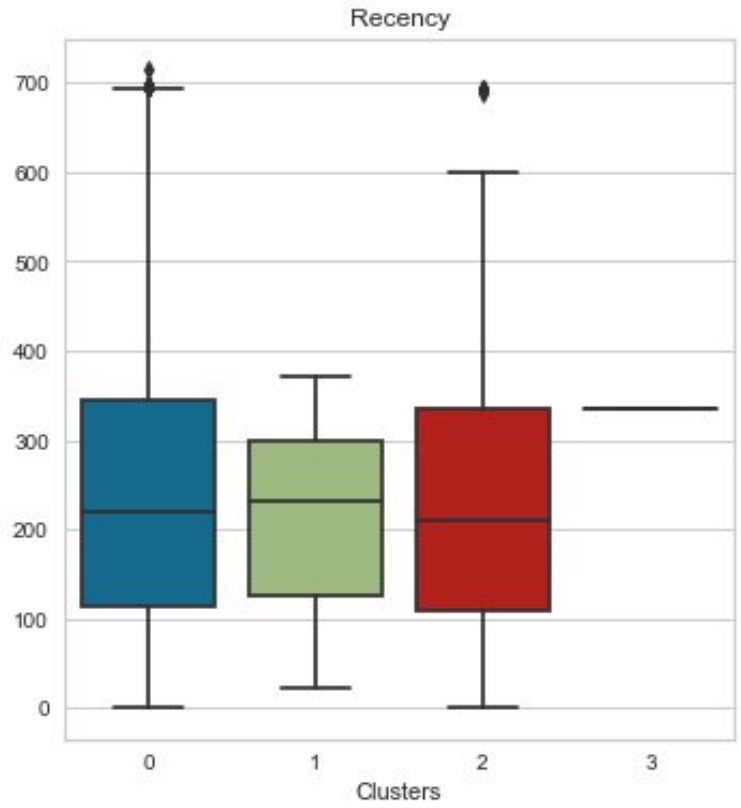
KMEANS (4 CLUSTERS)

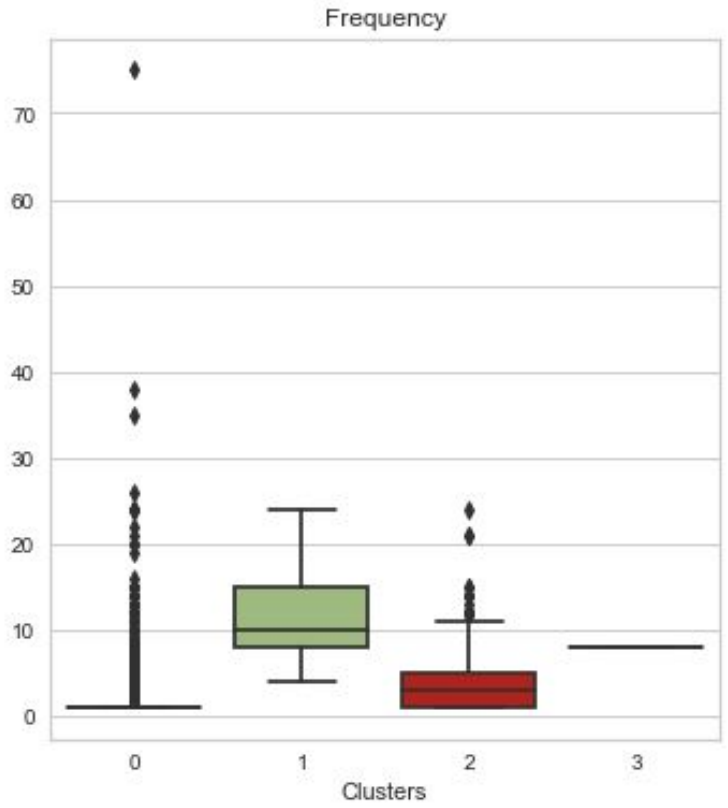


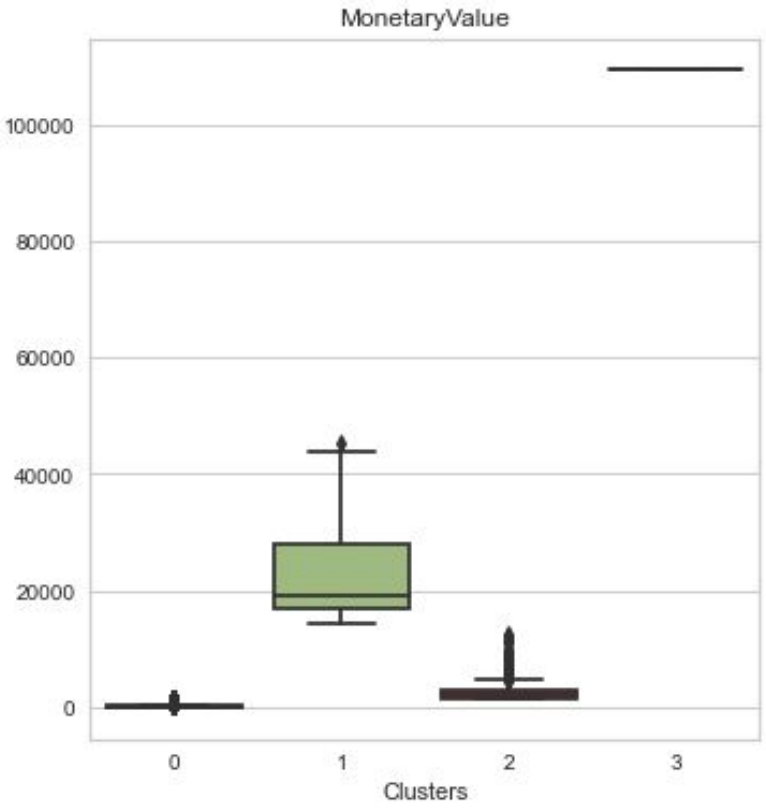




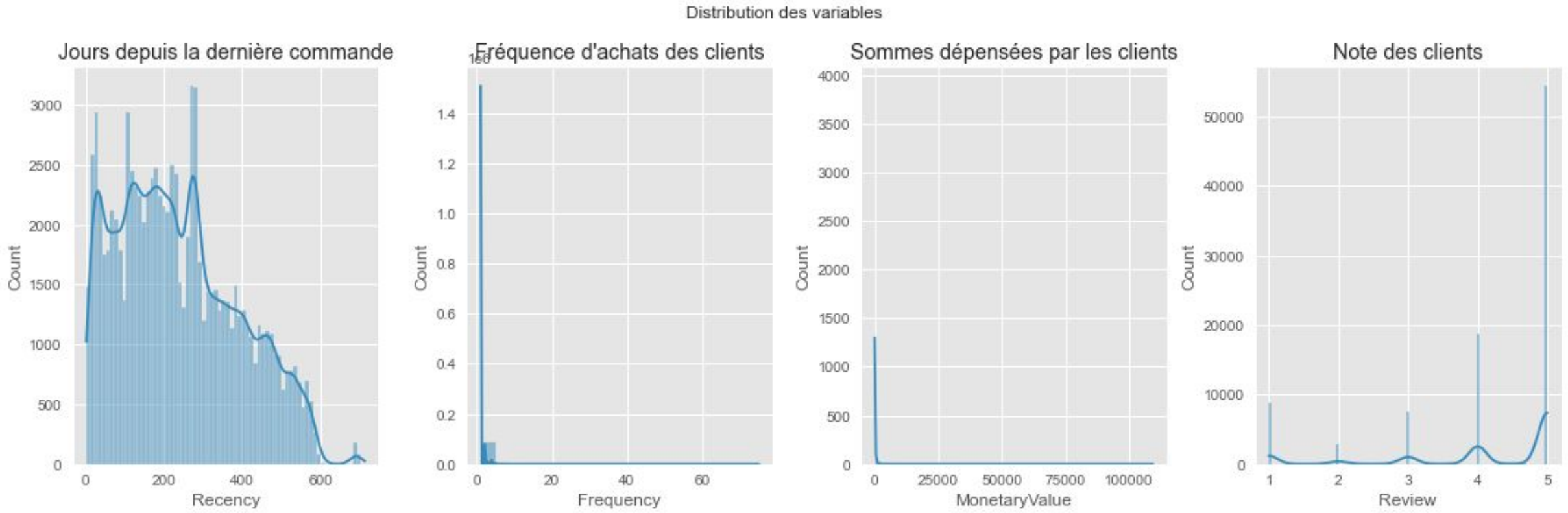
KMEANS (4 CLUSTERS)



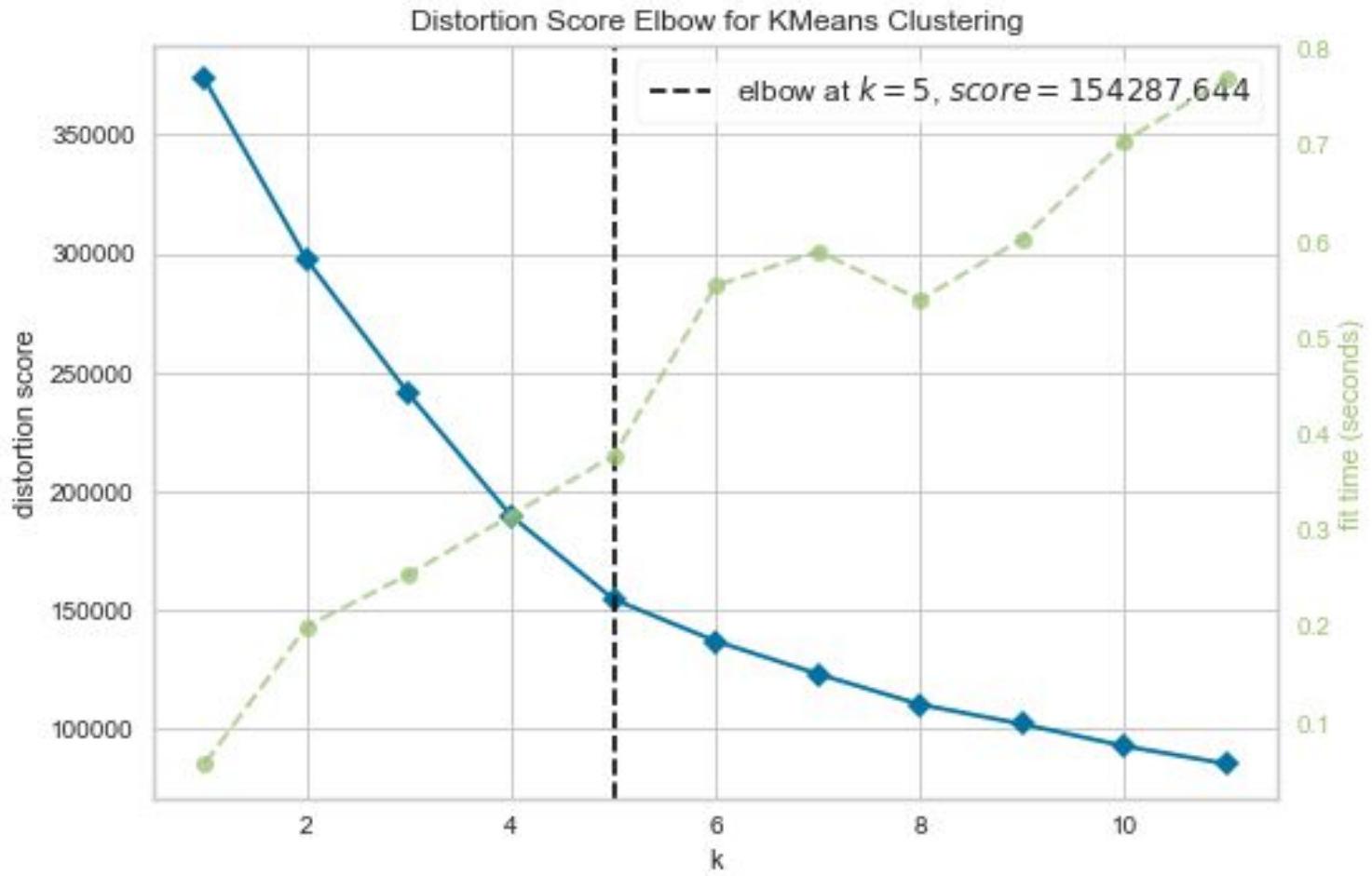




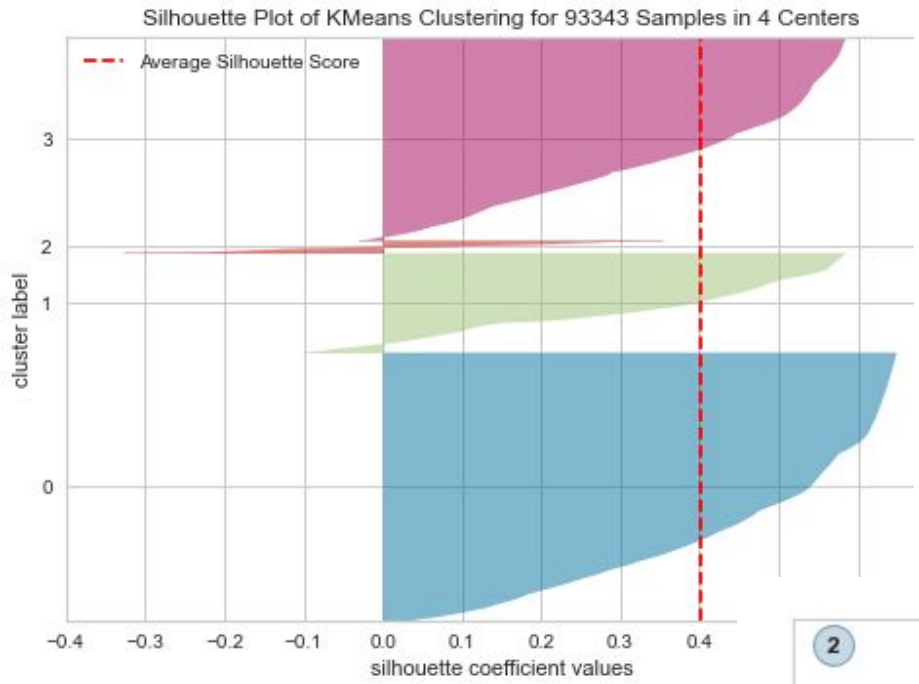
KMEANS SUR LA SEGMENTATION RFM AJOUT DE LA VARIABLE REVIEW

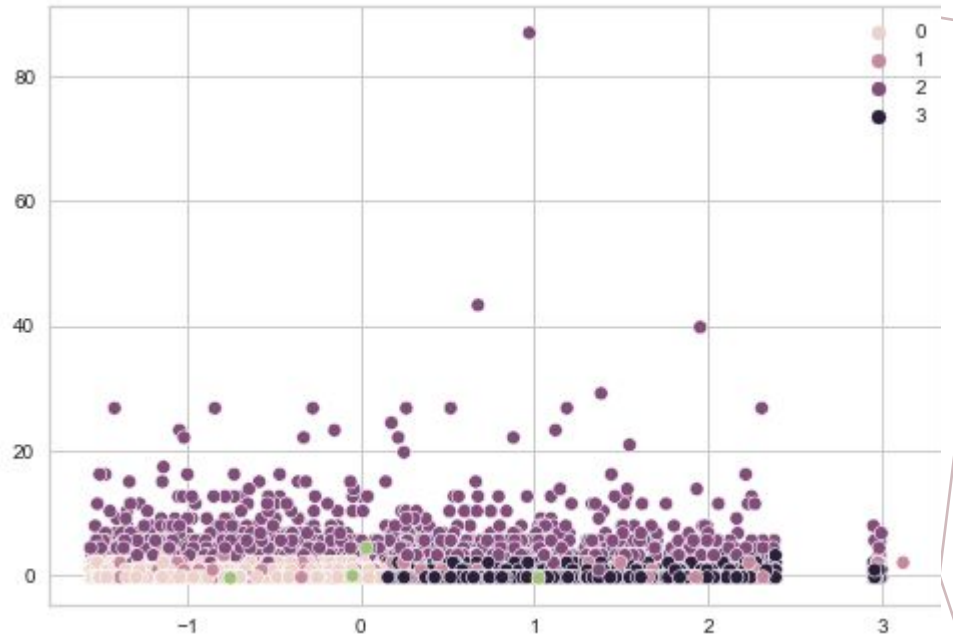


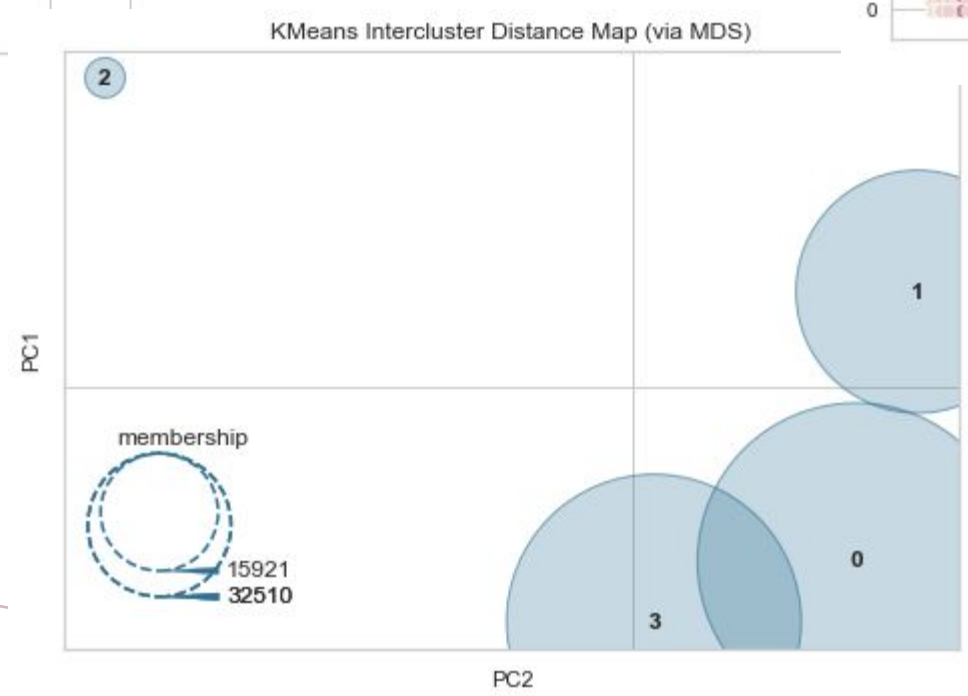
KMEANS SUR LA SEGMENTATION RFM AJOUT DE LA VARIABLE REVIEW



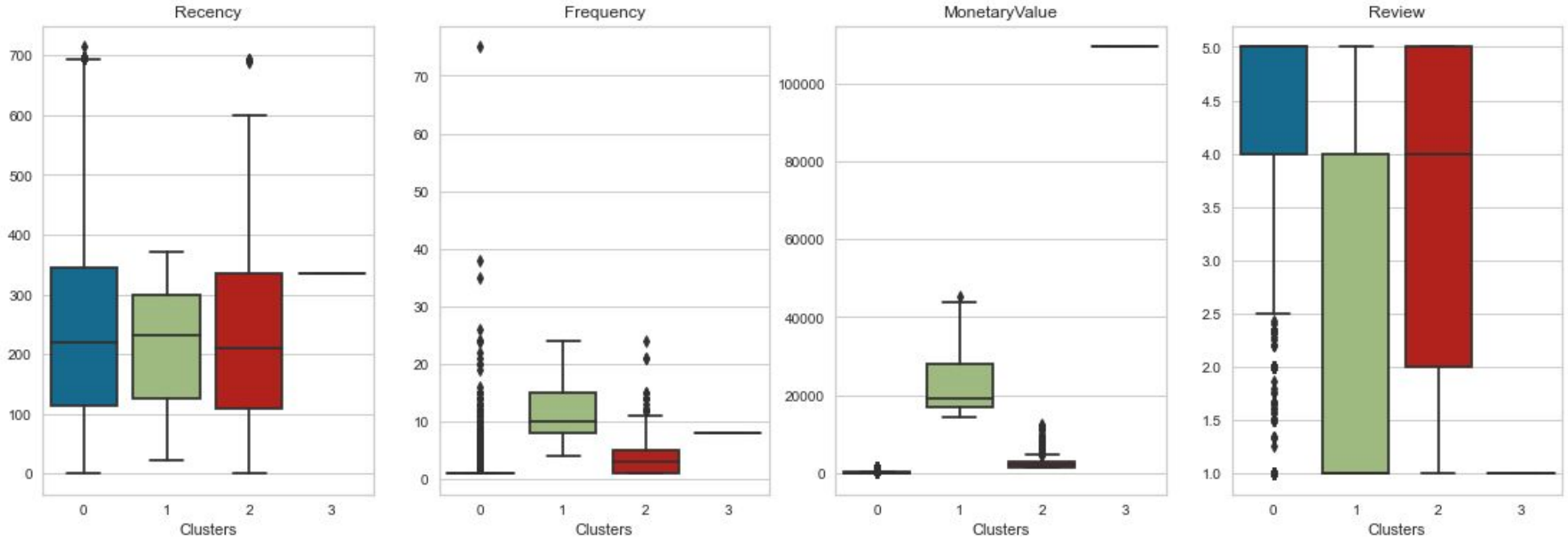
KMEANS (4 CLUSTERS)







KMEANS (4 CLUSTERS) AVEC REVIEW

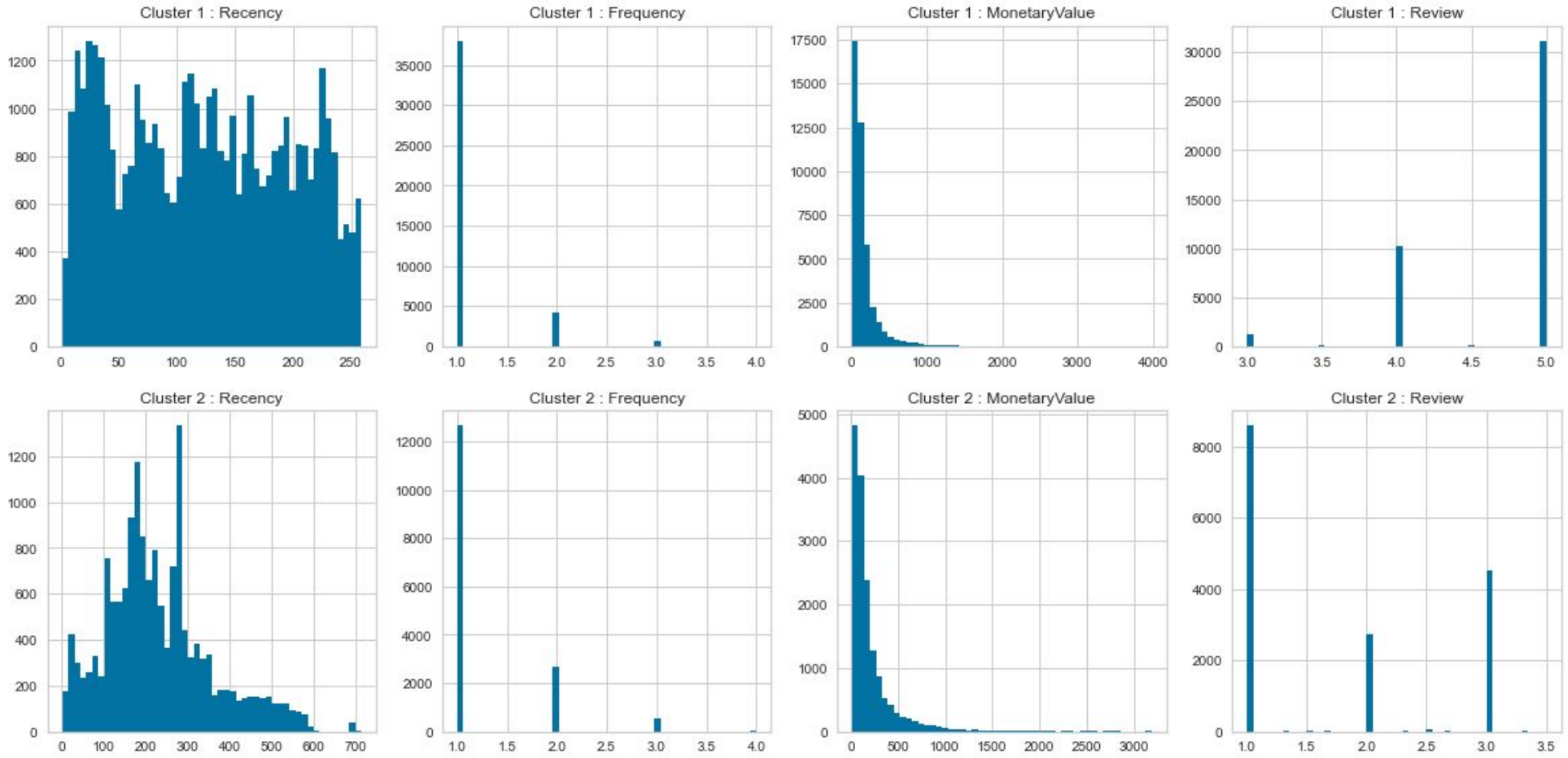


KMEANS (4 CLUSTERS) AVEC REVIEW

	nb_customers	prop_customers	mean_Recency	mean_Frequency	mean_MonetaryValue	mean_Review
Cluster 1	42910	0.459702	122.948171	1.131624	170.279445	4.699727
Cluster 2	15921	0.170564	230.487469	1.244520	208.604483	1.746087
Cluster 3	2002	0.021448	242.247253	5.153846	1838.933072	3.722998
Cluster 4	32510	0.348285	393.069825	1.138634	171.346187	4.634692

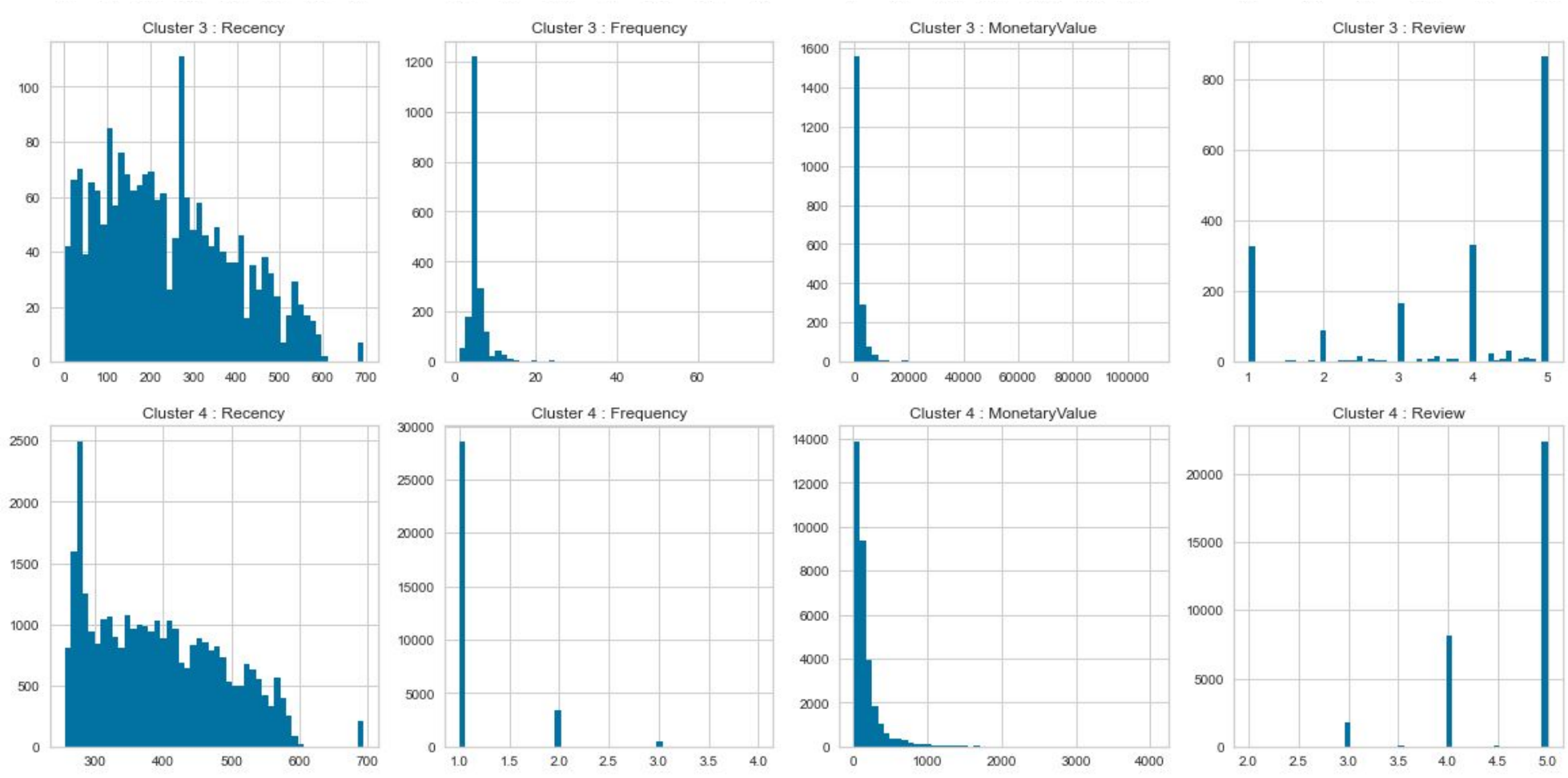
KMEANS (4 CLUSTERS) AVEC REVIEW

Clusters 1 et 2



KMEANS (4 CLUSTERS) AVEC REVIEW

Clusters 3 et 4



PERSONAE DU K-MEANS

Cluster 1 - Nouveau client

- 45% des clients
- ~1 achat
- ~120 derniers jours
- ~170 réals
- note 4,7/5

Cluster 2 -Le mécontent

- 17% des clients
- ~1 achat
- ~230 derniers jours
- ~208 réals
- note 1,7/5

Cluster 3 - le gros client

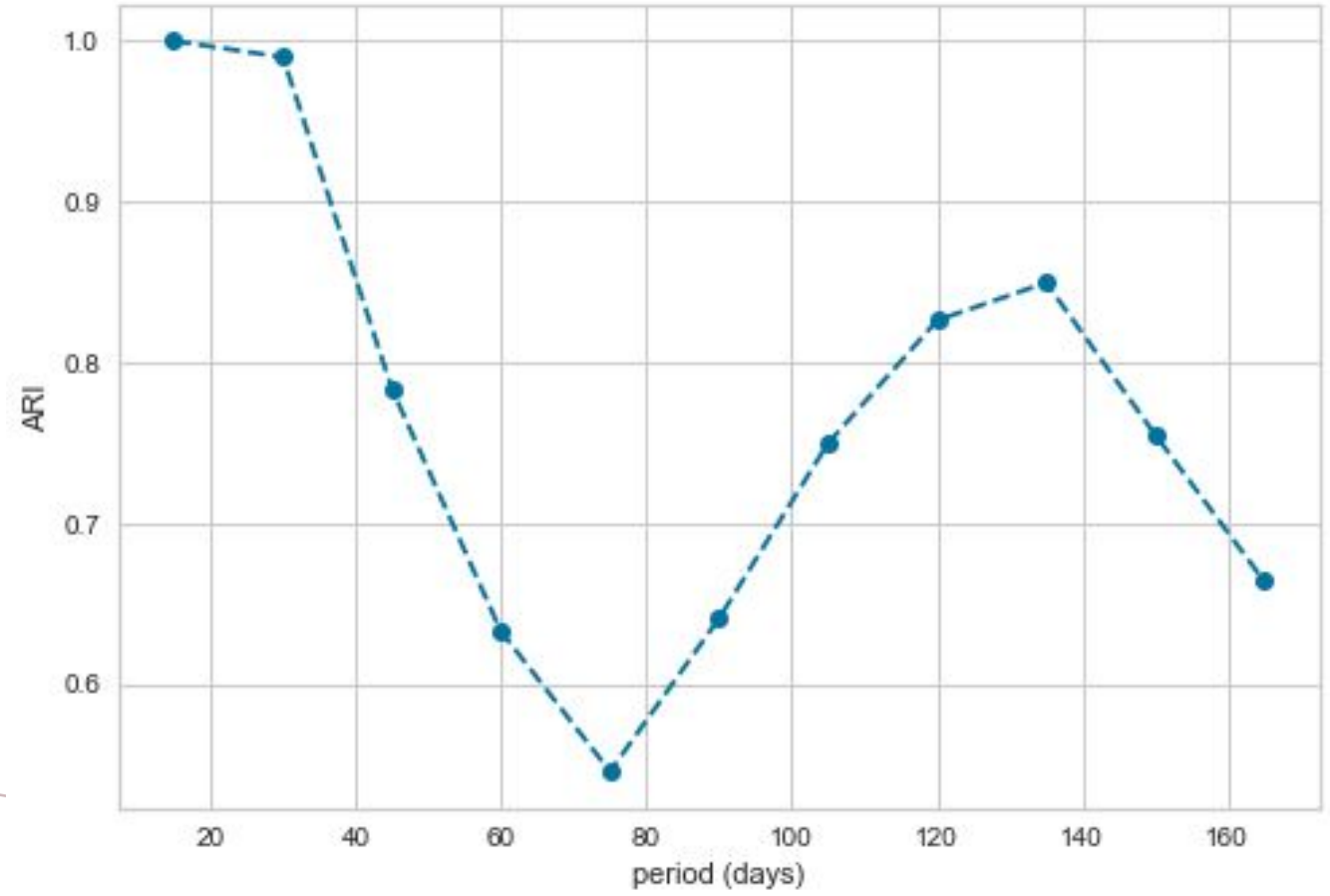
- 2% des clients
- ~5 achats moyen
- ~1838 réals
- note 3.7/5

Cluster 4 -Le client dormant

- 35% des clients
- ~1 achat
- ~390 derniers jours
- ~170 réals
- note 4,6

Maintenance prévisionnelle Kmeans avec 4clusters basée sur l'ARI *(adjusted Rand index)*

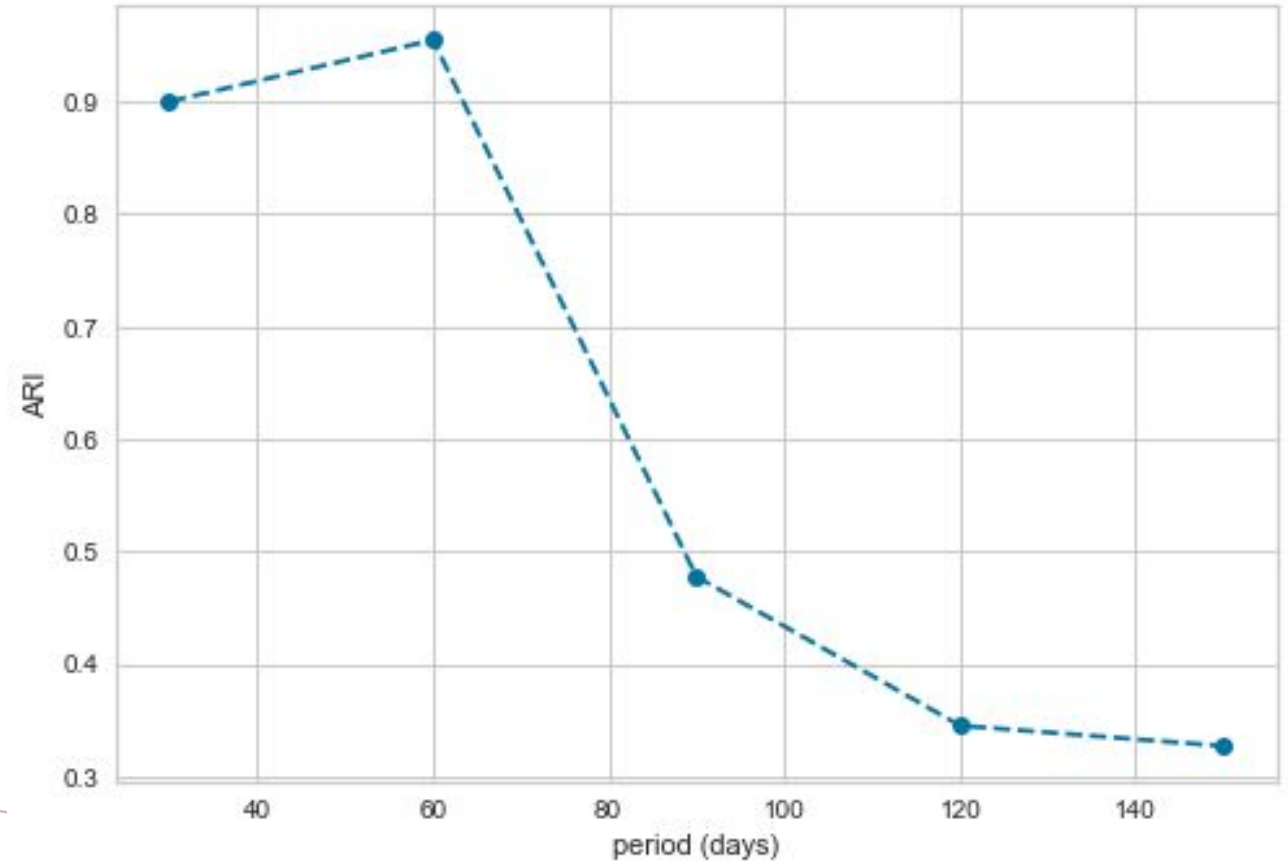
2016 sur 180 jours



ARI >= 0,90 excellente récupération
0,80 =< ARI < 0,90 bonne récupération
0,65 =< ARI < 0,80 récupération modérée
ARI < 0,65 mauvaise récupération

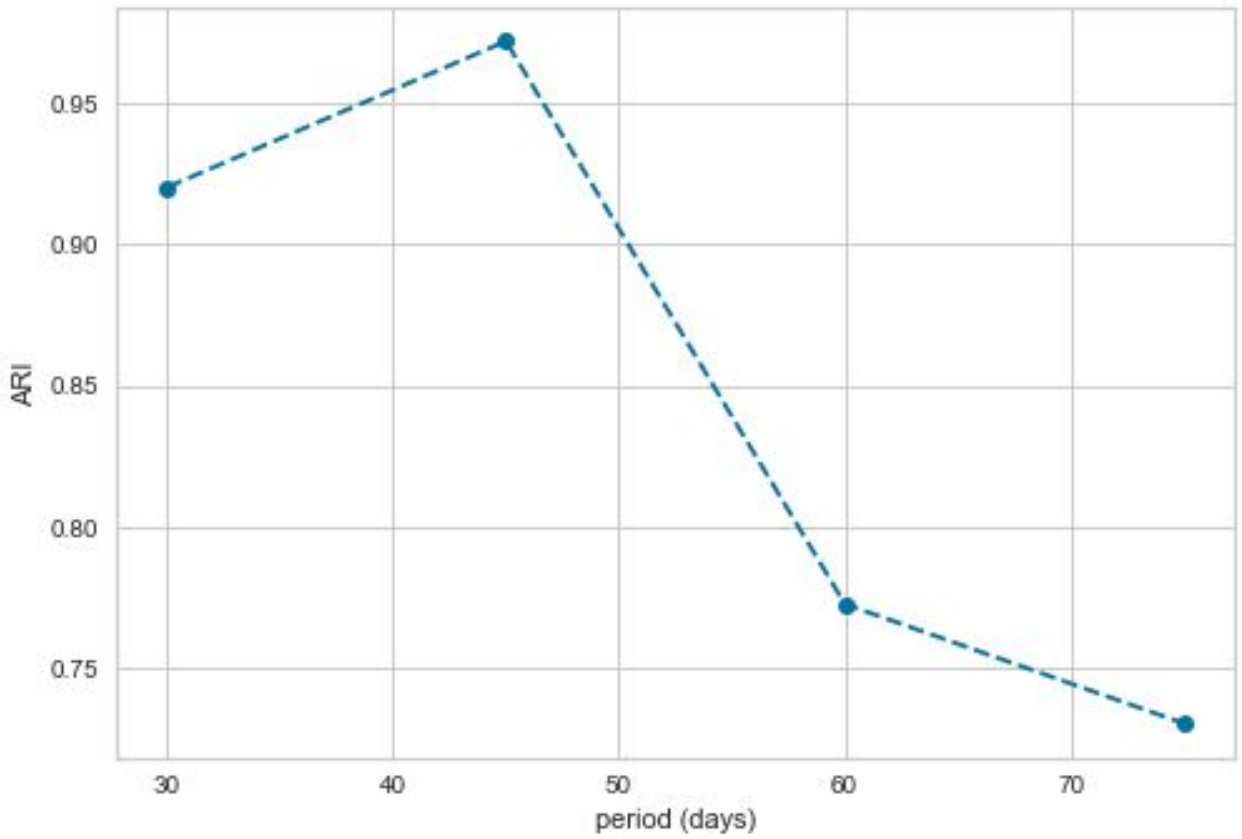
Maintenance prévisionnelle Kmeans avec 4 clusters basée sur l'ARI

2017 sur 180 jours



Maintenance prévisionnelle Kmeans avec 4 clusters basée sur l'ARI

2018 sur 90 jours



Une maintenance tous les 2 mois est recommandée afin que la segmentation reste pertinente

CONCLUSION

-base de données de qualité et facilement transformable

-segmentation RFM permet d'obtenir une

analyse rapide et facilement interprétable

-Kmeans permet de différencier les bons et mauvais clients

-Kmeans est très efficace/efficient mais versatile

-3% clients > 1 commande

-manque de données en volume

-manque de qualification des clients (sexe, age, CSP)

-impossibilité de faire fonctionner DBscan et CAH sur l'ensemble des clients

AXES D'AMÉLIORATIONS DE L'ÉTUDE

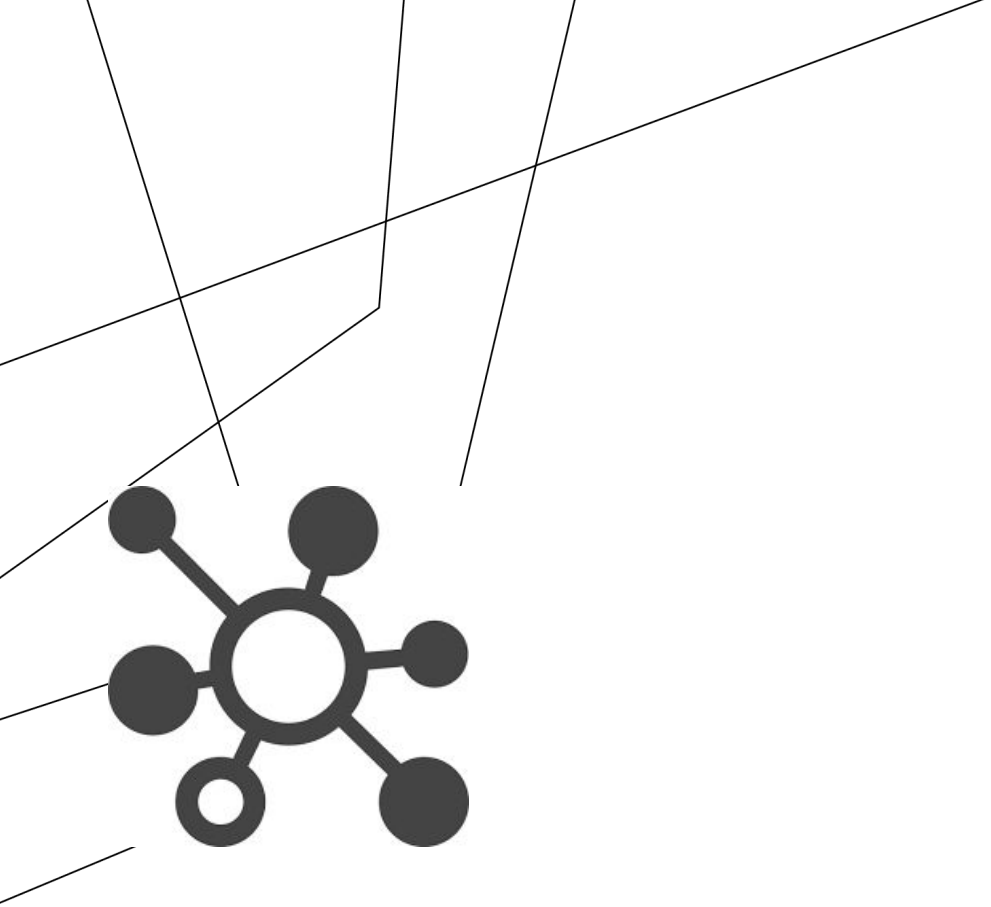
1. Augmenter la taille du dataset pour améliorer les résultats des algorithmes et permettre d'en tester d'autres

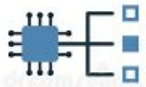
2. Déporter la puissance de calcul pour permettre l'utilisation de DBSCAN et CAH

3. Méthode de tracking (cookies) des clients pour avoir une meilleure estimation de leur fréquence d'achat et visite du site

4. Échanger avec les métiers pour identifier sur quel axe accentuer l'analyse afin de proposer de meilleurs leviers

MERCI POUR VOTRE ATTENTION





CLUSTERING