Segmentez des clients d’un site e-commerce

olist

Lubin Stéphane - Mars 2023

Problématique et objectifs

La société Olist souhaite réaliser une segmentation des clients

Objectifs:

- Proposer un clustering de clients adapté
- Proposer une fréquence de maintenance pour le modèle de segmentation proposée

Objectif secondaire:

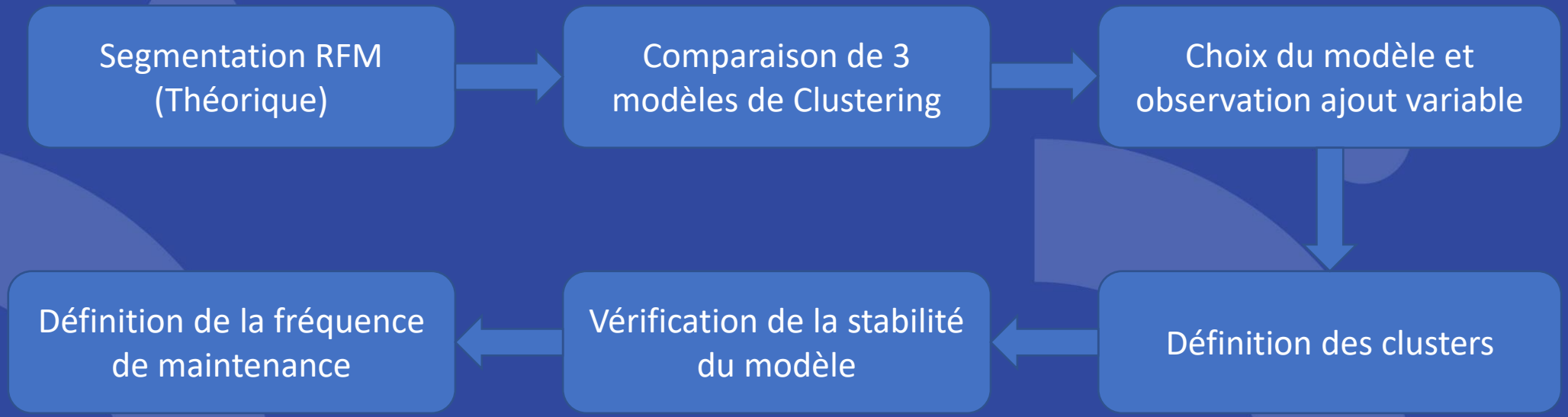
- Observer l'effet des variables sur le modèle

olist

Stratégie choisie

La stratégie choisie pour traiter le dataset est la segmentation RFM qui se base sur:

- La Récence du dernier achat
- La Fréquence d'achat
- Le Montant total des achats

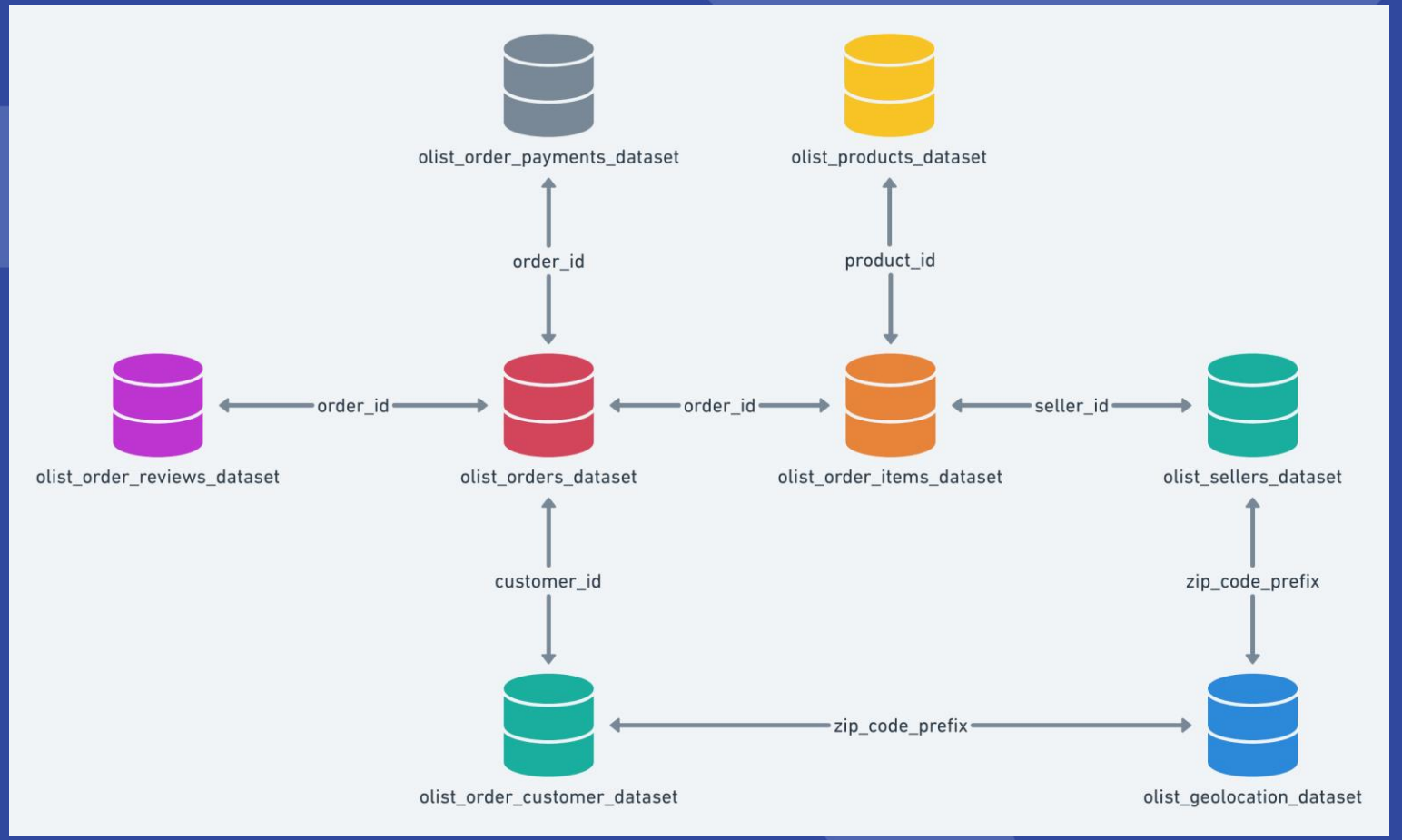


olist

Jeu de données

9 fichiers:

- Liste des clients
- Liste des commandes
- Liste des notes et commentaires
- Liste des vendeurs
- Liste des produits
- Liste des paiements
- Liste des géolocalisation
- Liste des produits par commandes
- Liste des catégories dans les 2 langues



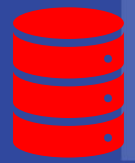
olist

Nettoyage et traitements des données

Sélections des fichiers et variables d'intérêt:

Customers:

- Customer id
- Customer unique id
- City
- State
- Zip Code



Orders:

- Order id
- Order purchase time
- Order status

Payments:

- Payments id
- Payment sequential
- Payment installments
- Payment value

Reviews:

- Order id
- Reviews Score

Items:

- Order id
- Product id
- Order items

Products:

- Product id
- Category

olist

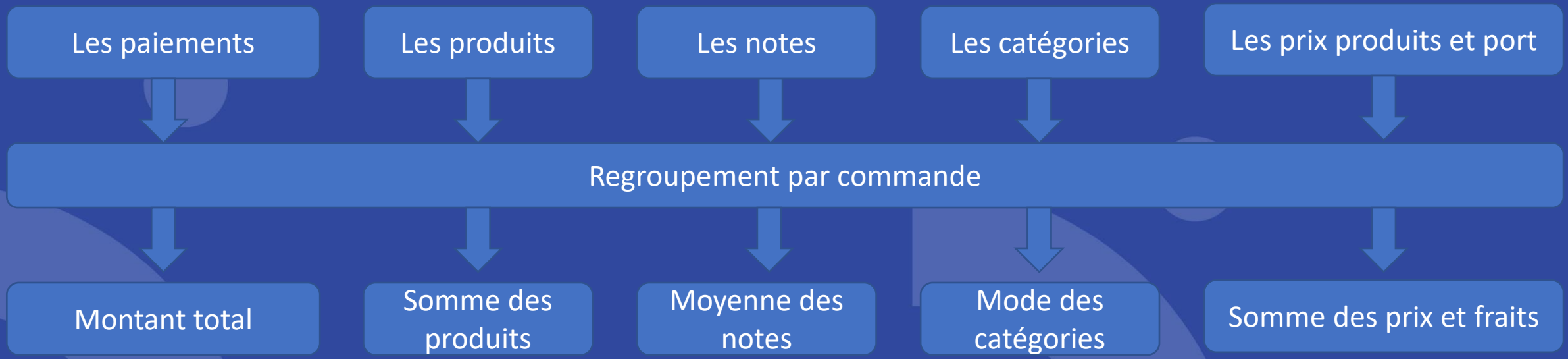
- Imputation

Variable note imputée par 3

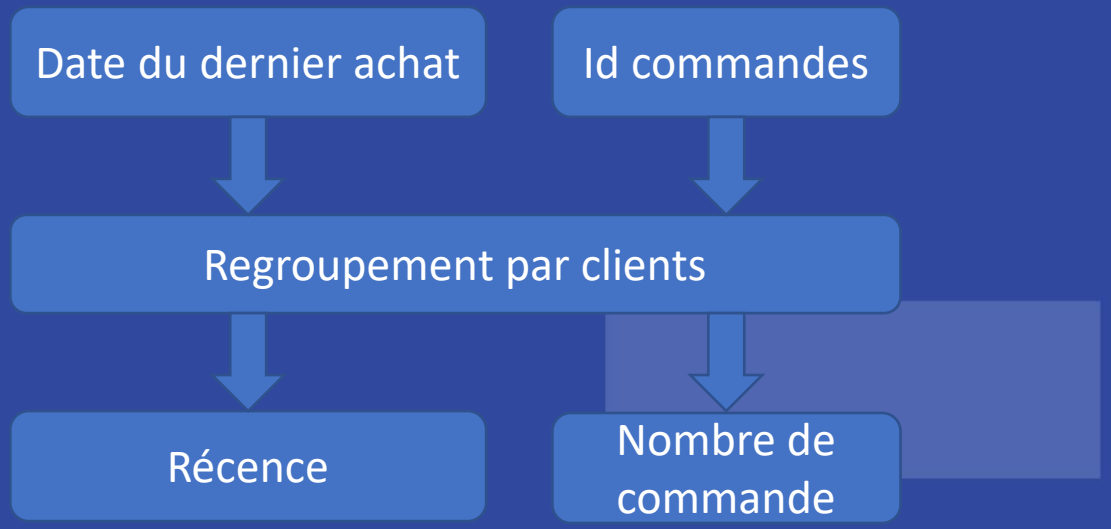
Les variables prix produit, et frais de port imputées par le mode

La variable catégorie par ‘unknown’

- Traitements:

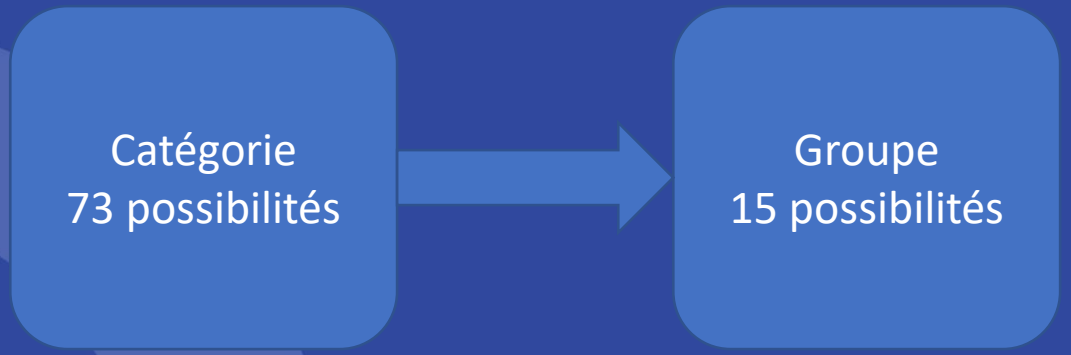


olist



On se suppose que la date du test est le 01/10/2018 étant donné que la date la plus proche étant fin septembre

Nous avons nos:
- R -> Récence
- F -> Nombre de commande
- M -> Montant total



Répartition des catégories dans 15 groupes

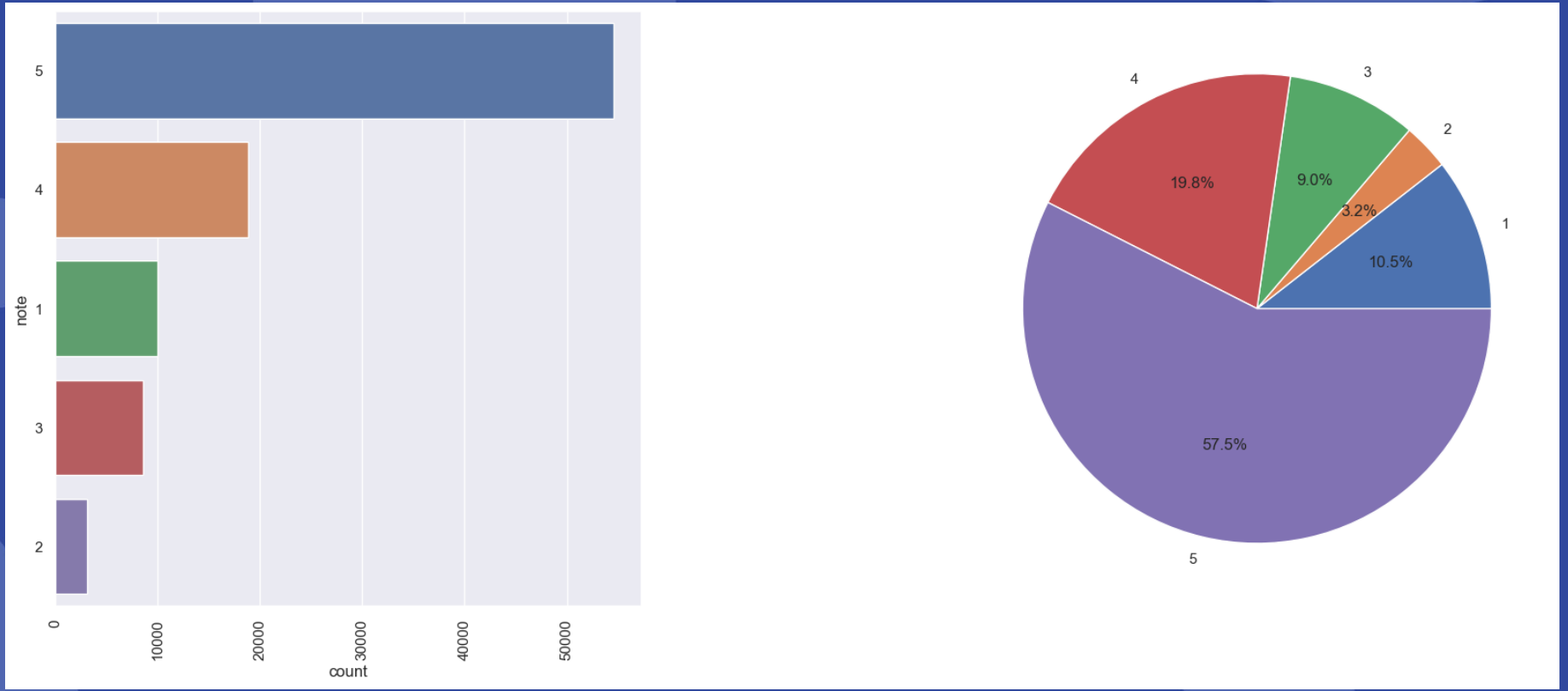
Il y avait des catégories qui semblaient représenter la même gamme de produits

Préparation pour One-hot-Encoder

olist

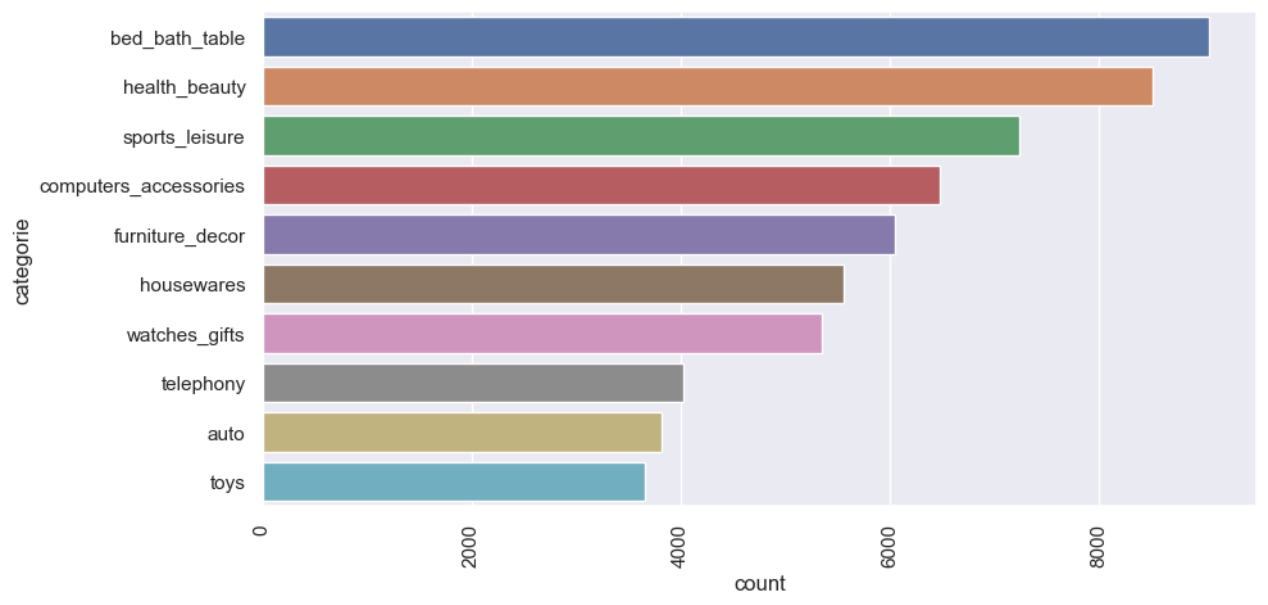
Analyse exploratoire

Répartition des notes:

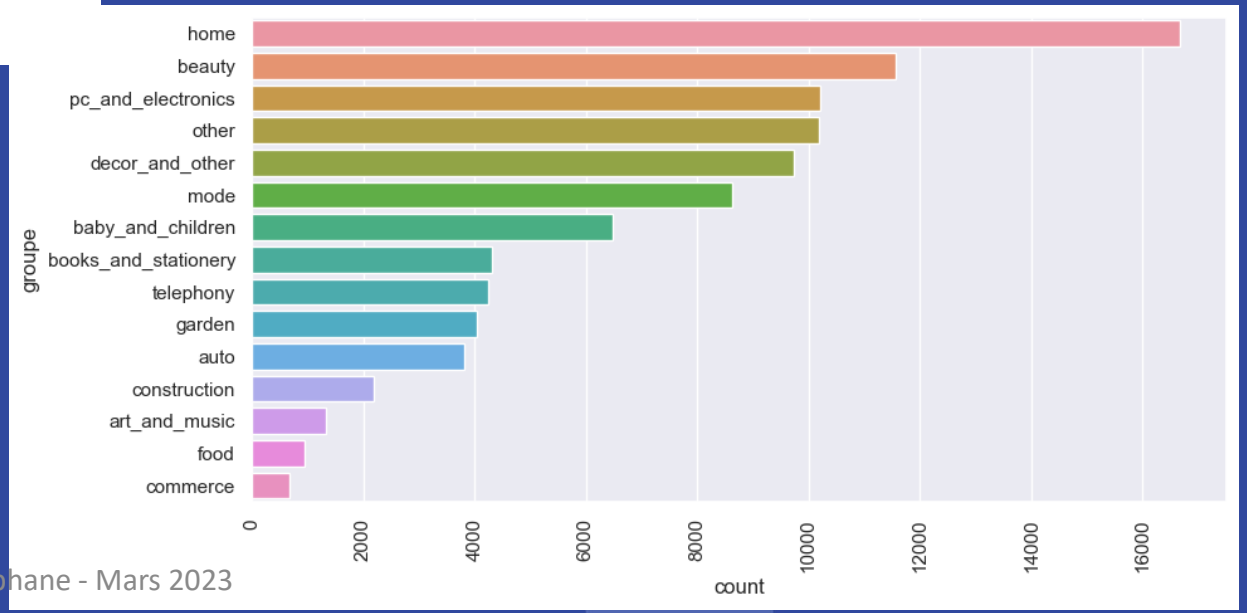


olist

Répartition des catégories et des groupes:



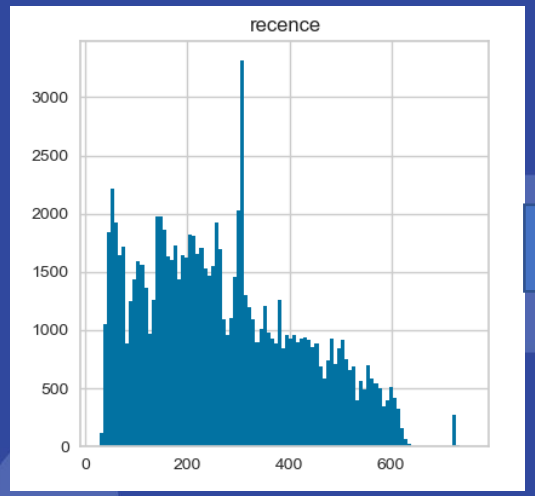
Les commandes concernent en majorité les accessoires de maison, la beauté et l'électronique



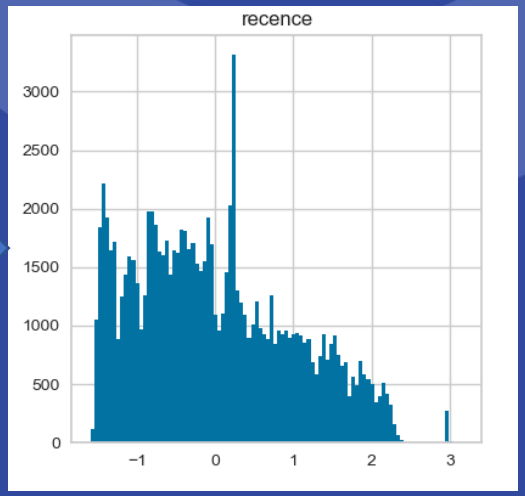
Les groupes représentent bien les 10 catégories les plus représentées

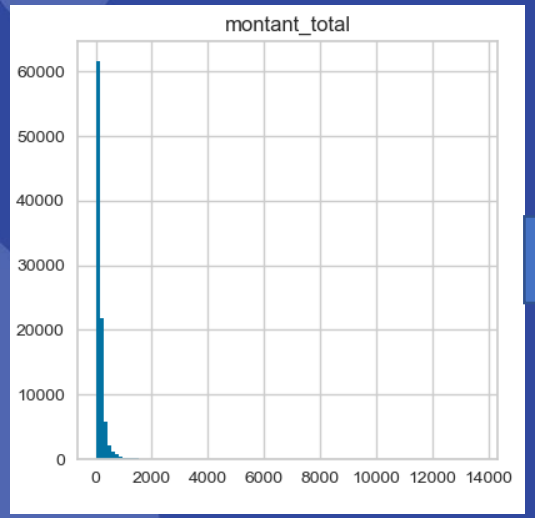
olist

Transformations et encodage

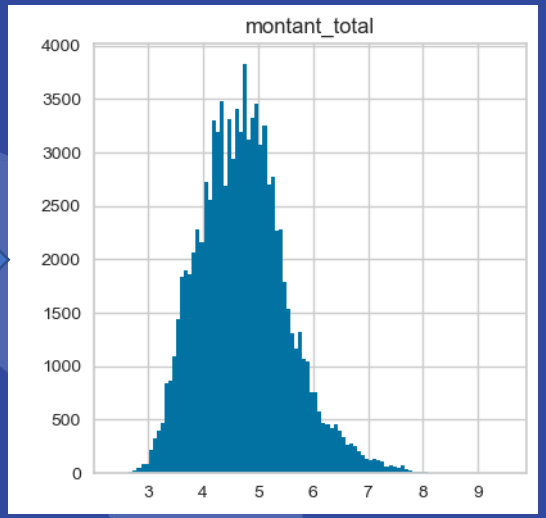


Standard Scaler



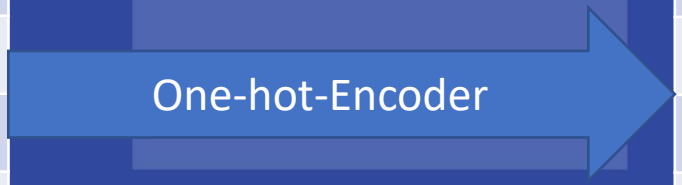


Transformation log



olist

groupe
Home
Beauty
Mode
….
Auto



Home	Beauty	Mode	….	Auto
1	0	0	….	0
0	1	0	….	0
0	0	1	….	0
0	0	0	….	0
0	0	0	….	1

Une colonne

15 colonnes

Les variables, nombre de commandes, moyenne de nombre de paiement, moyenne de nombre de produit, ne subissent aucune transformation pour préserver leur variance

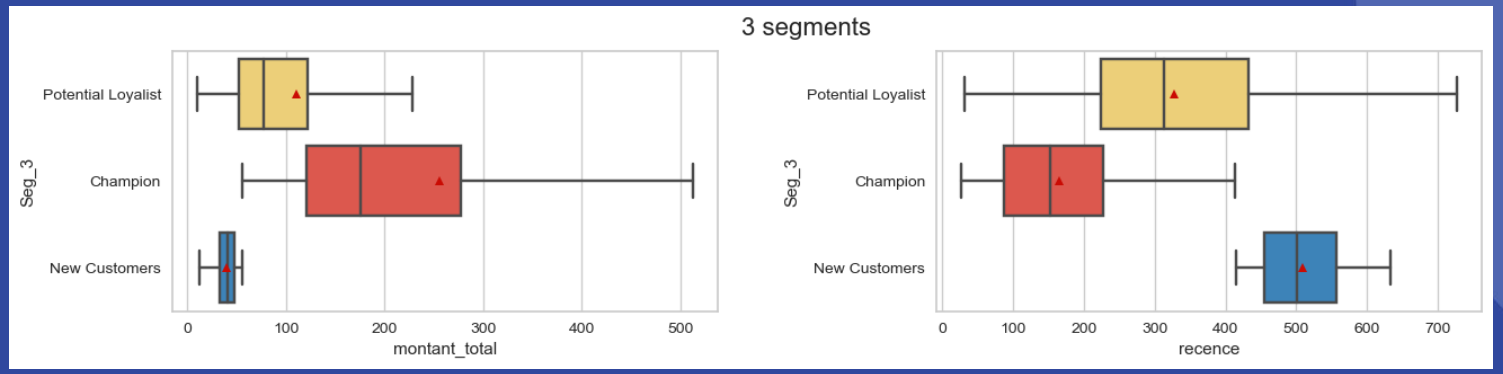
olist

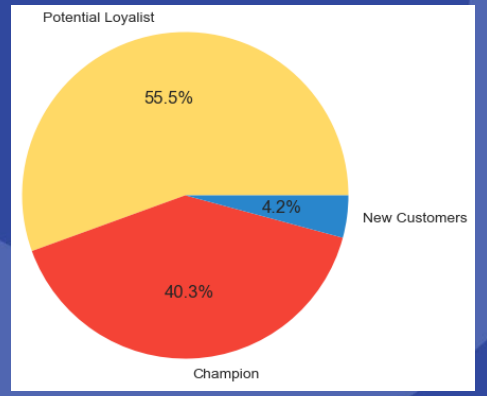
Segmentation RFM

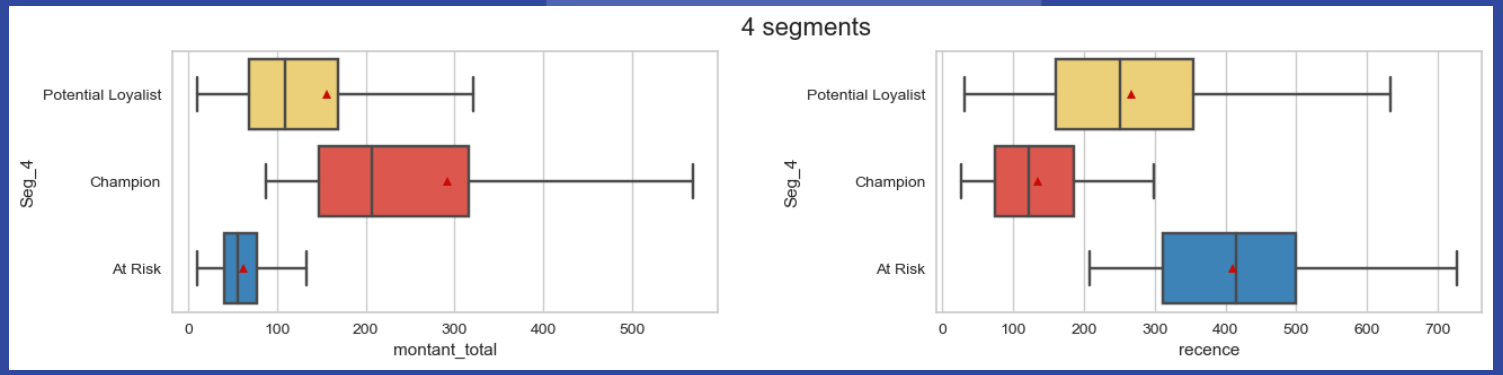
- La segmentation RFM basée sur un score utilisant la méthode des quantiles:
  - Répartition des clients selon la somme de leur score établi sur les variables Récence, Fréquence et Montant total
  - Score variant de 1 à 5 pour les variables soit un score total variant de 3 à 15

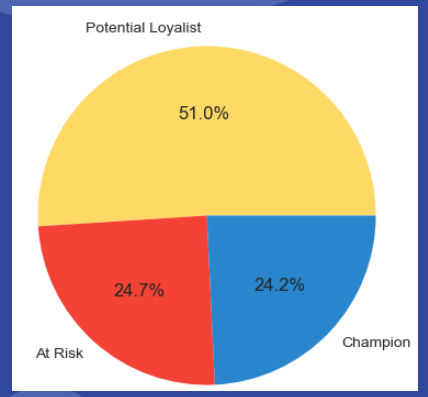
- La segmentation RFM basée sur la technique des 11 segments:
  - Score variant de 1 à 5 pour les variables, puis création de segments RFM (ex: 111, 545 ou 354)
  - Répartitions des segments dans 11 clusters

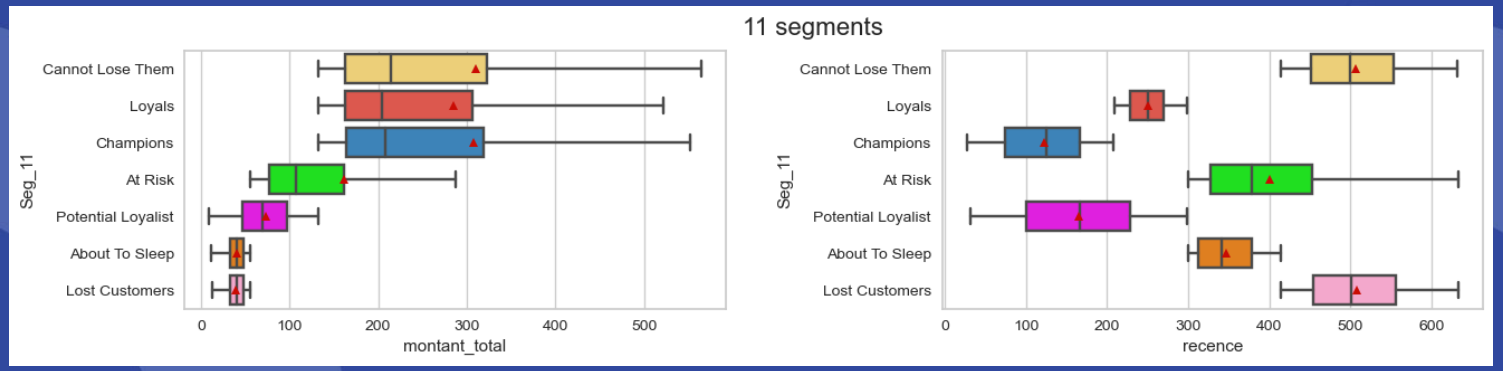
olist

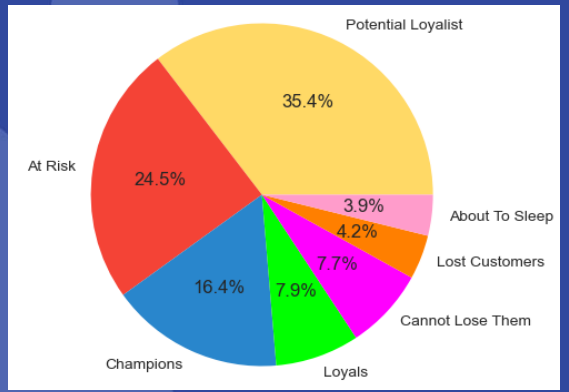

3 segments
Potential Loyalist
Champion
New Customers
Seg_3
montant_total
recence
0
100
200
300
400
500
600
700


Potential Loyalist
55.5%
4.2%
New Customers
40.3%
Champion


4 segments
Potential Loyalist
Champion
At Risk
Seg_4
montant_total
recence
0
100
200
300
400
500
600
700

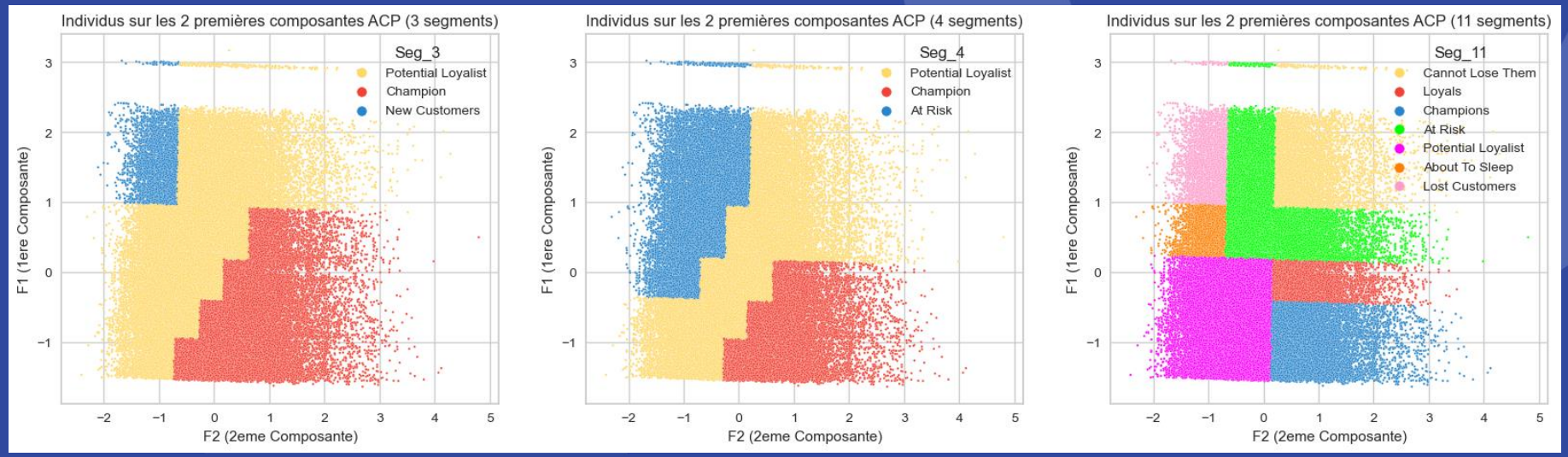

Potential Loyalist
51.0%
24.7%
24.2%
At Risk
Champion


11 segments
Cannot Lose Them
Loyals
Champions
At Risk
Potential Loyalist
About To Sleep
Lost Customers
Seg_11
montant_total
recence
0
100
200
300
400
500
600

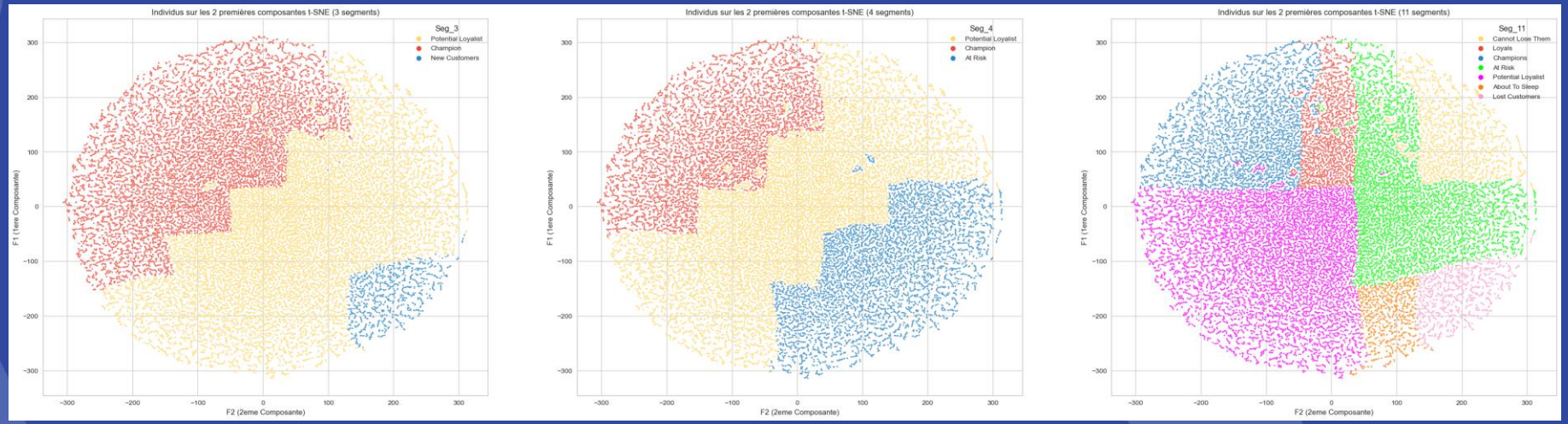

Potential Loyalist
35.4%
At Risk
24.5%
3.9%
About To Sleep
4.2%
Lost Customers
7.7%
Cannot Lose Them
16.4%
7.9%
Champions
Loyals

olist

ACP



t-SNE



olist

Comparaison des modèles

La comparaison est faite sur 50% du dataset choisie aléatoirement, à cause des limites matérielles.

- Le clustering via k-Means:

  Minimise la variance intra-cluster (l'inertie)

- Le clustering hiérarchique (Agglomerative Clustering):

  Chaque individus est considéré comme un cluster, puis ils sont regroupés pour minimiser la variance intra-cluster en utilisant la méthode Ward

- Le clustering via DBScan:

  Relie les points d'un même cluster en passant de voisin à voisin

olist

Les métriques utilisées:

- L'inertie:
  - Détermine si les points d'un même cluster sont proches
  - Représente la variance intra-cluster
  - Si petit bon cluster
- Le score silhouette:
  - Détermine si chaque point est dans le bon cluster
  - Si grand bon cluster
- L'Indice de Rand ajusté:
  - Calcule une mesure de similarité entre deux regroupements en considérant toutes les paires d'échantillons et en comptant les paires qui sont attribuées dans le même groupe ou dans des groupes différents dans les regroupements prédits et vrais.
  - Il est ajusté pour avoir une valeur proche de 0 si clusterings aléatoires, proche de 1 si clusterings identiques et négatif si les clusterings ne s'accordent pas

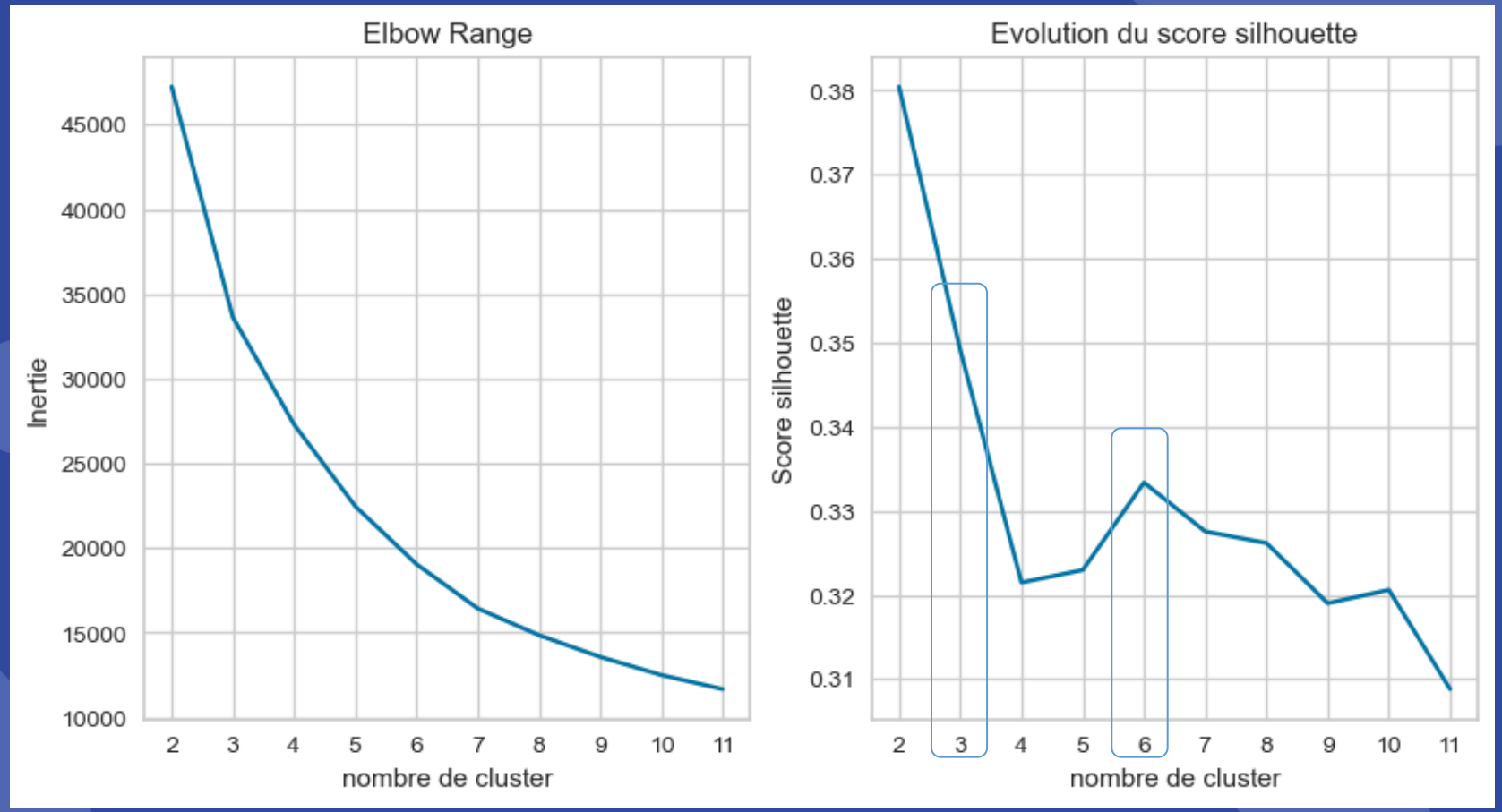
Autres métriques:

Détermine si les clusters sont bien éloignés les uns des autres et bien regroupés sur eux-mêmes

- Indice de Davies-Bouldin (Rapport entre la taille du cluster et la distance entre les clusters)
- Indice de Calinski-Harabasz (Variance inter-clusters / Variance intra-cluster)

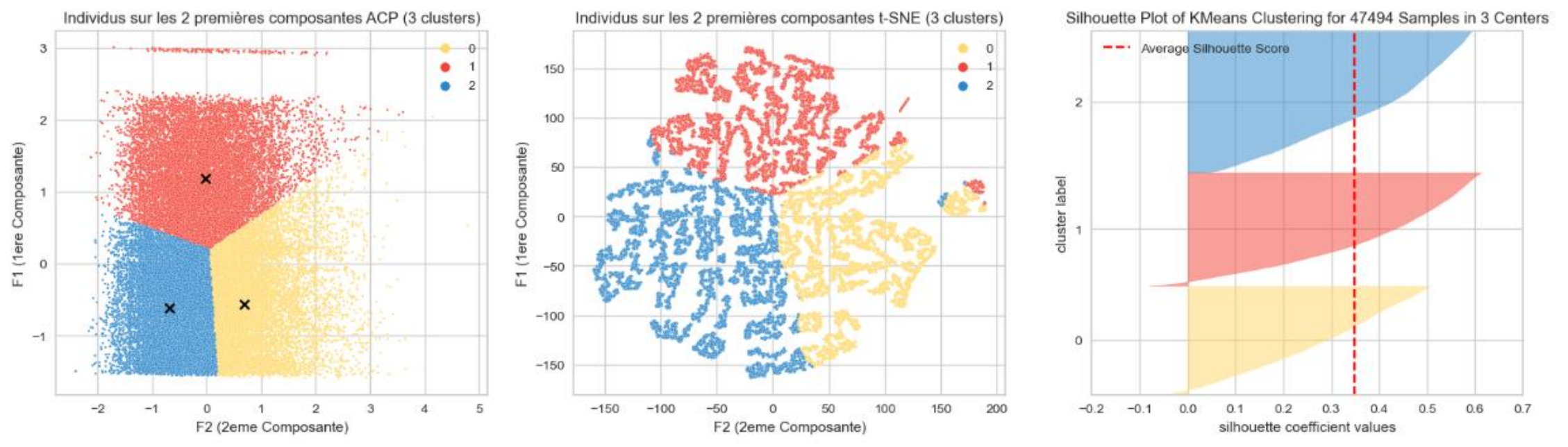
olist

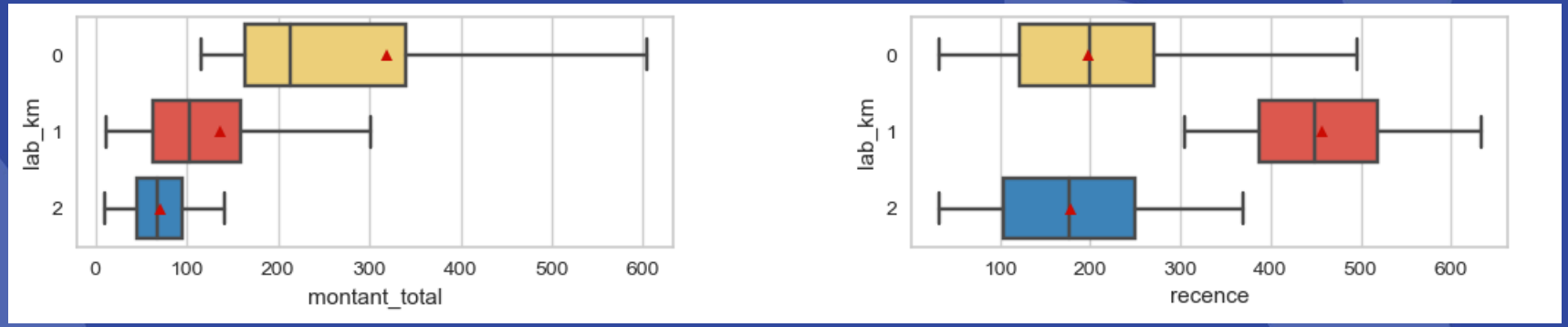
Clustering via k-Means



olist

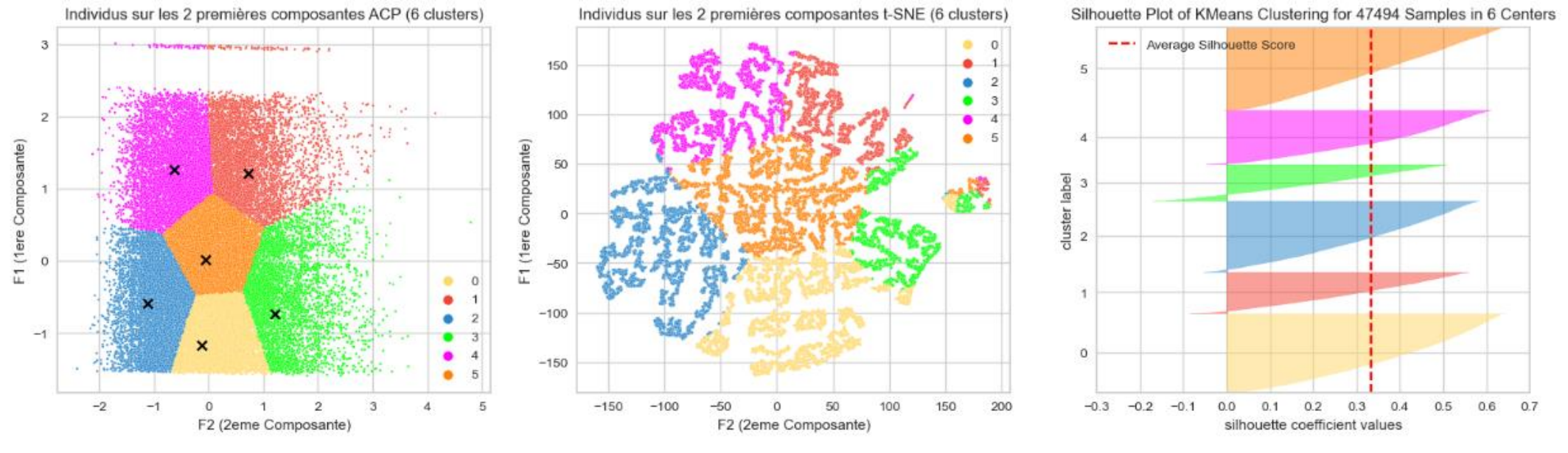
- 3 clusters:

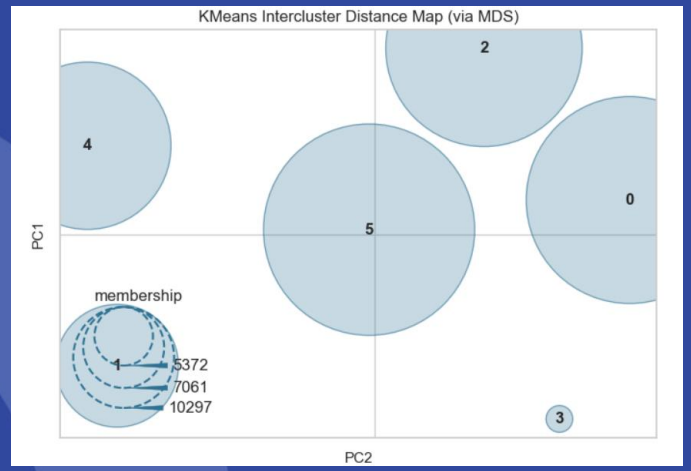


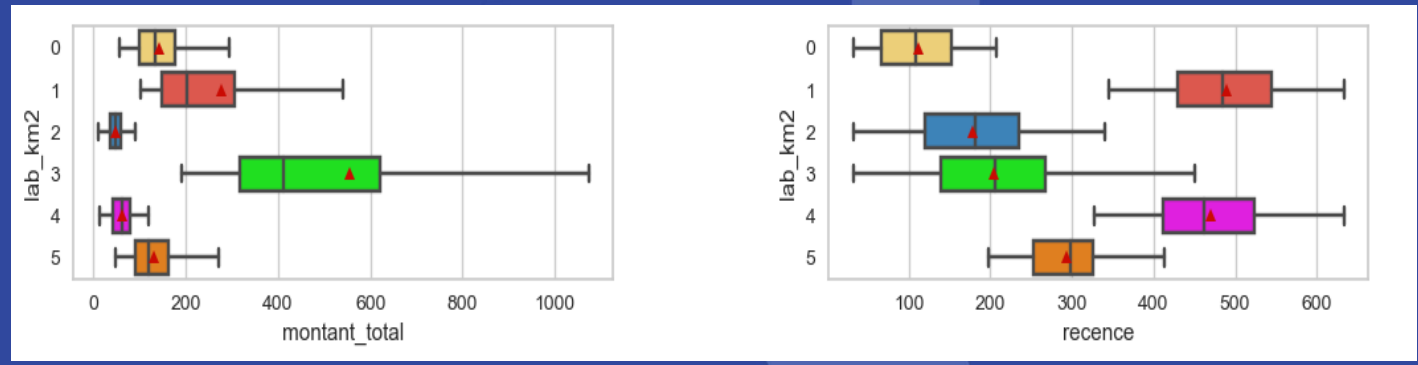


olist

- 6 clusters:



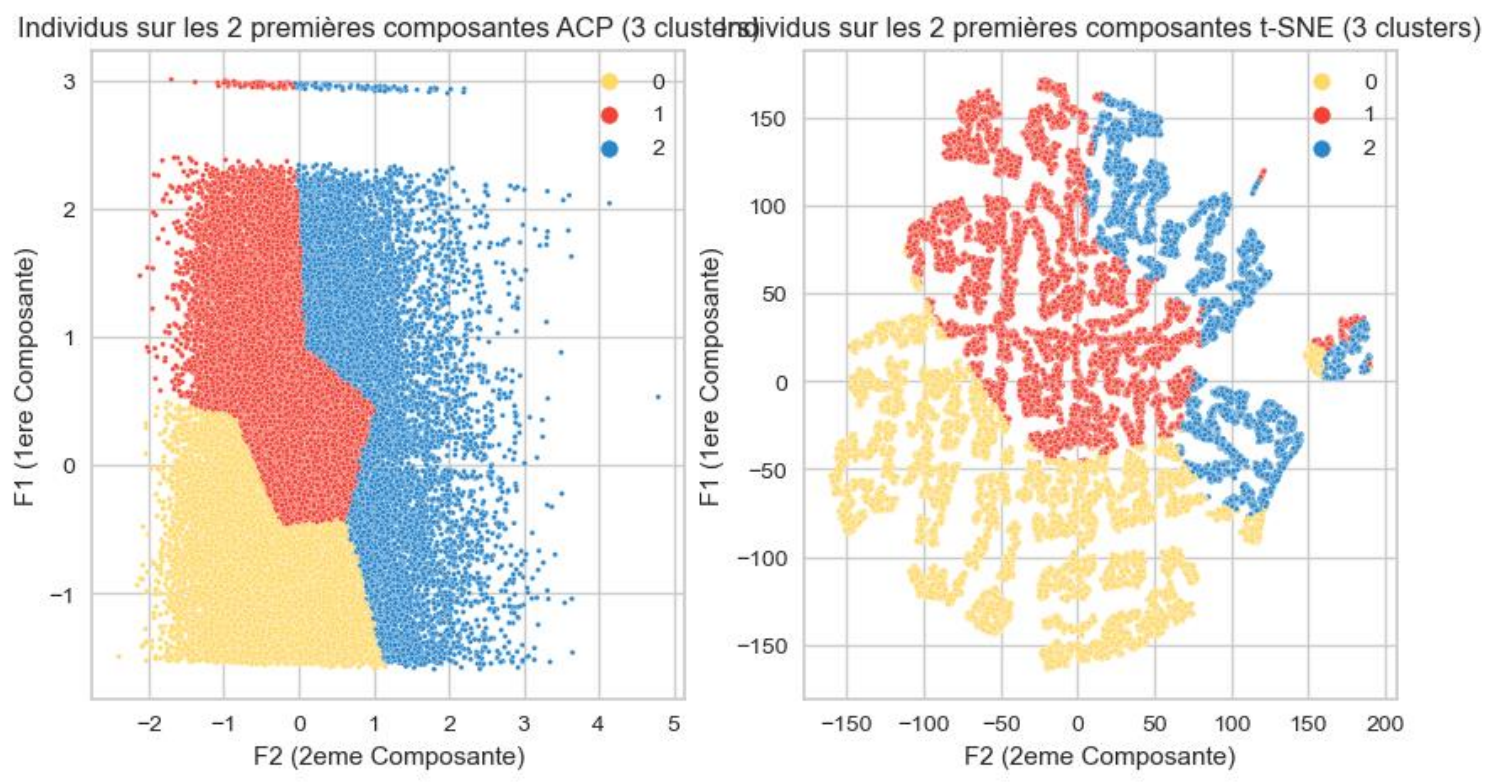


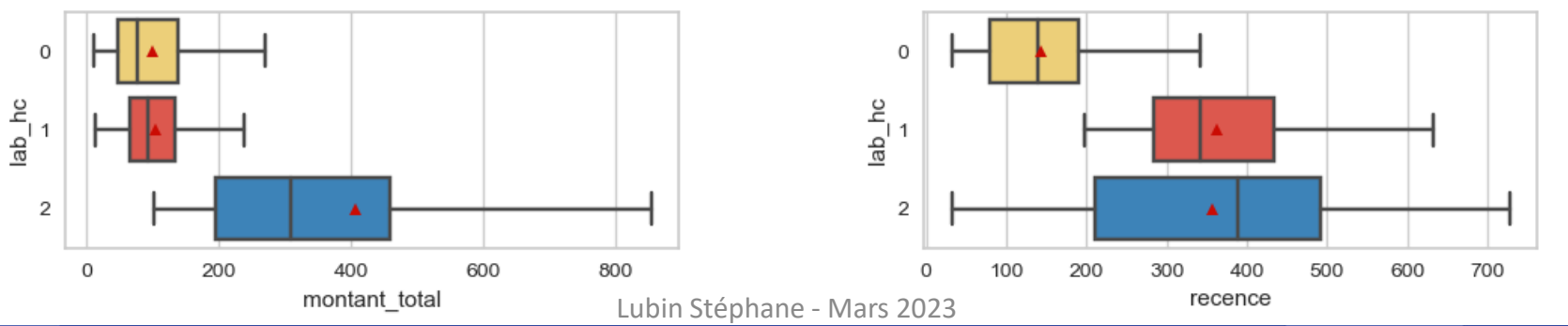


olist

Clustering Hiérarchique

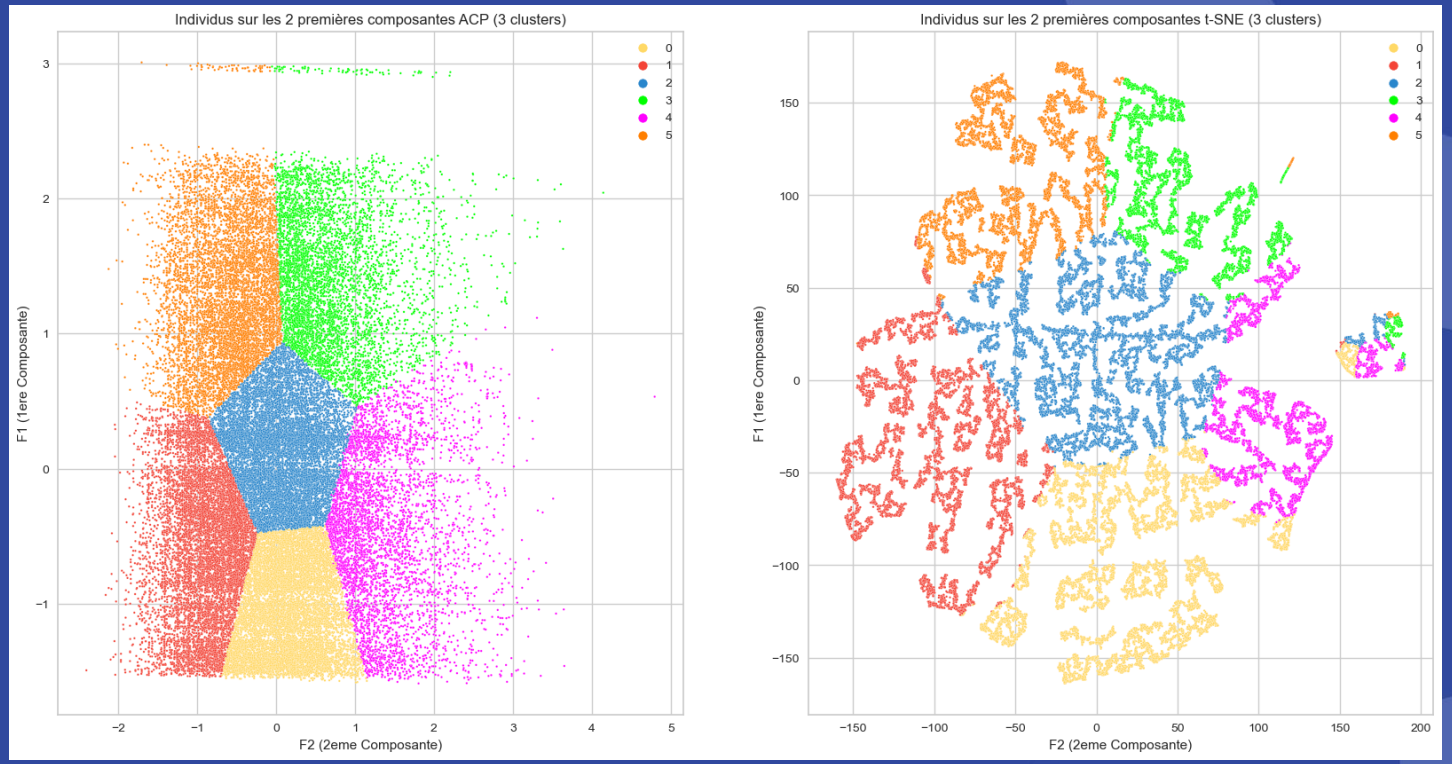
- 3 clusters:

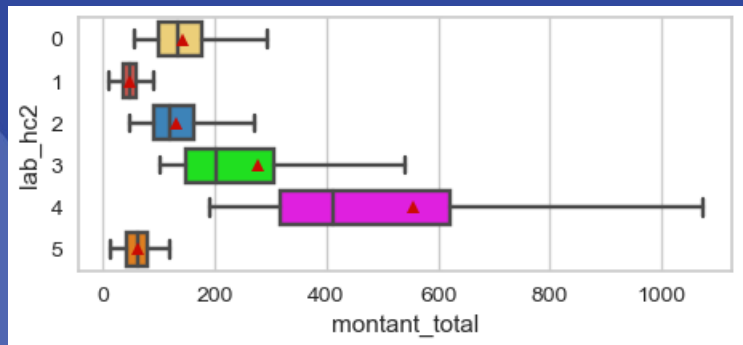


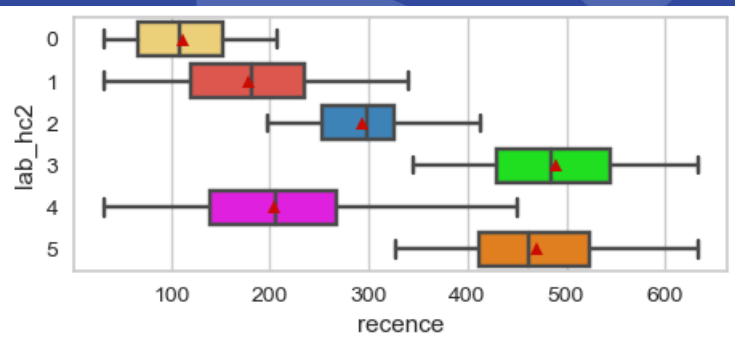


olist

- 6 clusters:







olist

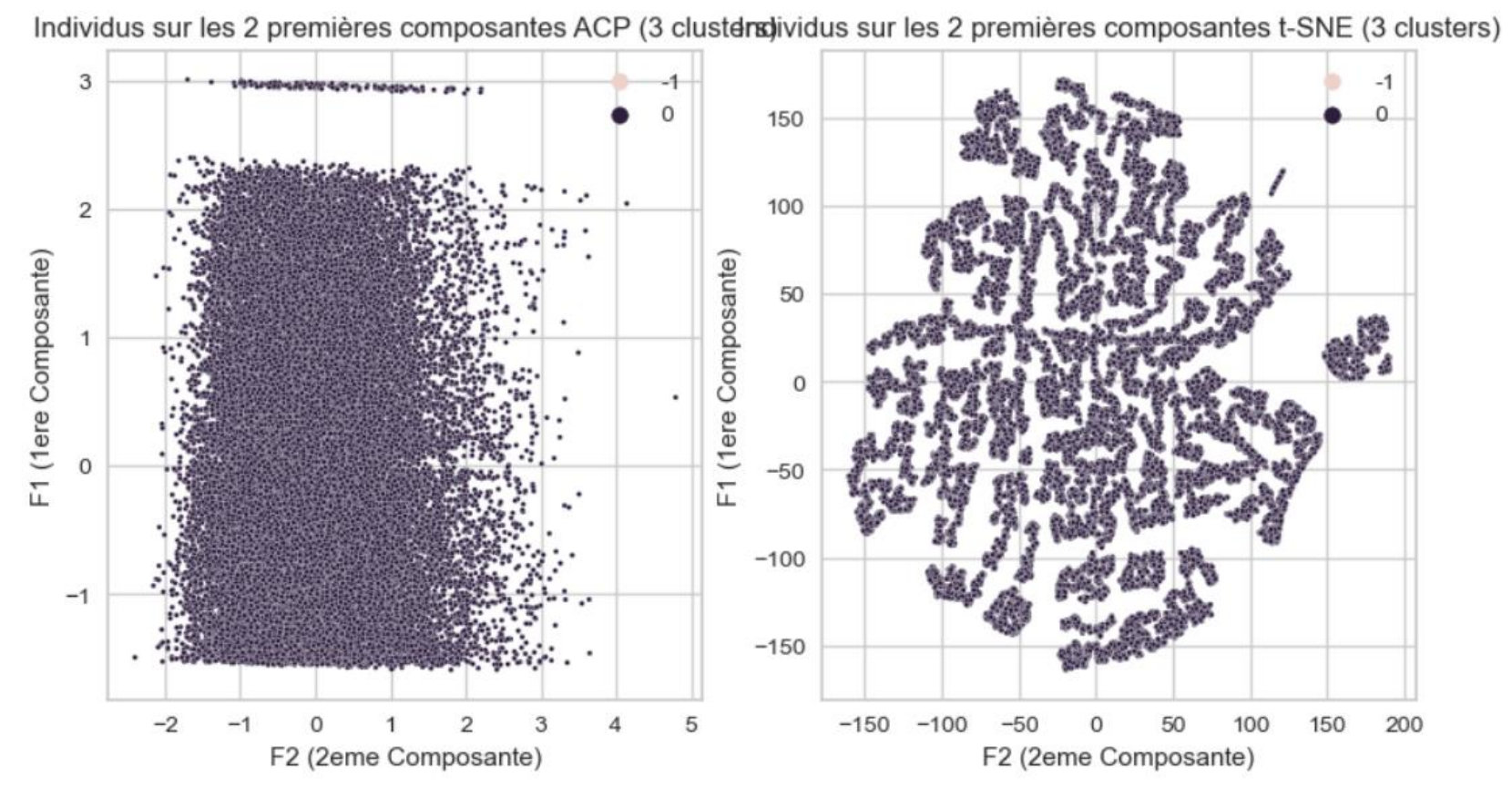
ARI entre les prédictions de k-Means et le clustering hiérarchique:

- Pour 3 clusters : ARI = 0,28
- Pour 6 clusters: ARI = 1

Pour des raisons matérielles, le modèle K-Means est préféré

olist

Clustering via DBScan



DBScan n’est pas adapté, points trop proches les uns des autres. Quasiment tous réunis dans le même cluster ou tous considérés comme du bruit

Ajout de variables

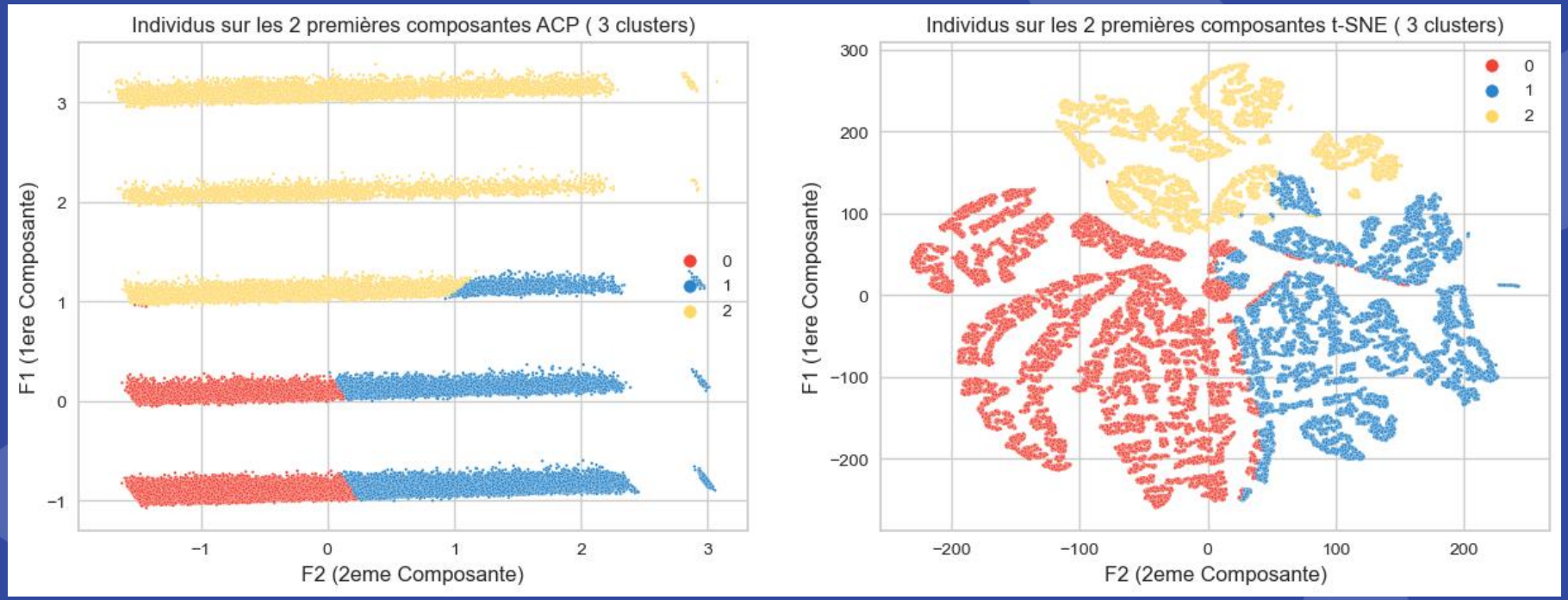
Le modèle utilisé sera k-Means, on regardera les 2 meilleurs choix de clustering, la répartitions et la stabilité des prédictions sur 25 répétitions.

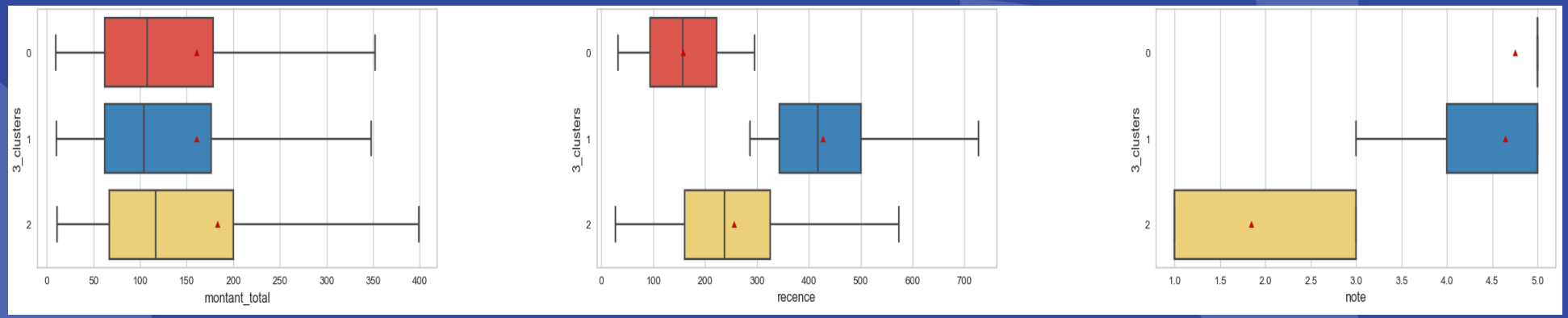
Ajout une à une de variable:

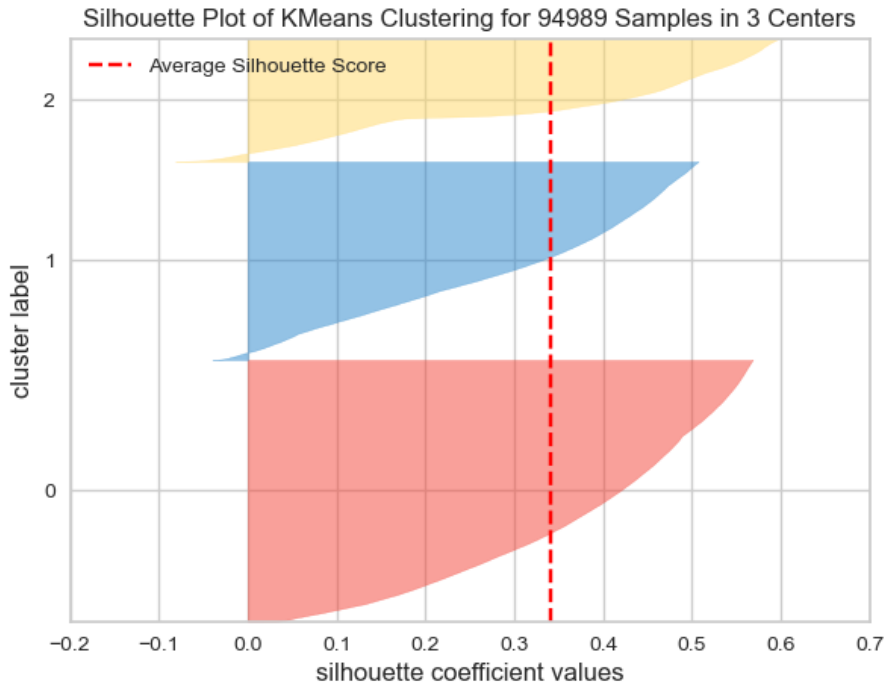
- Note , qui reflète la satisfaction de client
- Moyenne du nombre de produits, qui permet de voir si les clients font des grosses commandes
- Moyenne du nombre de paiement, qui permet de voir si l'acheteur est compulsif ou s'il fractionne ses paiements
- Groupe, qui permet de différencier la catégorie favori des clients

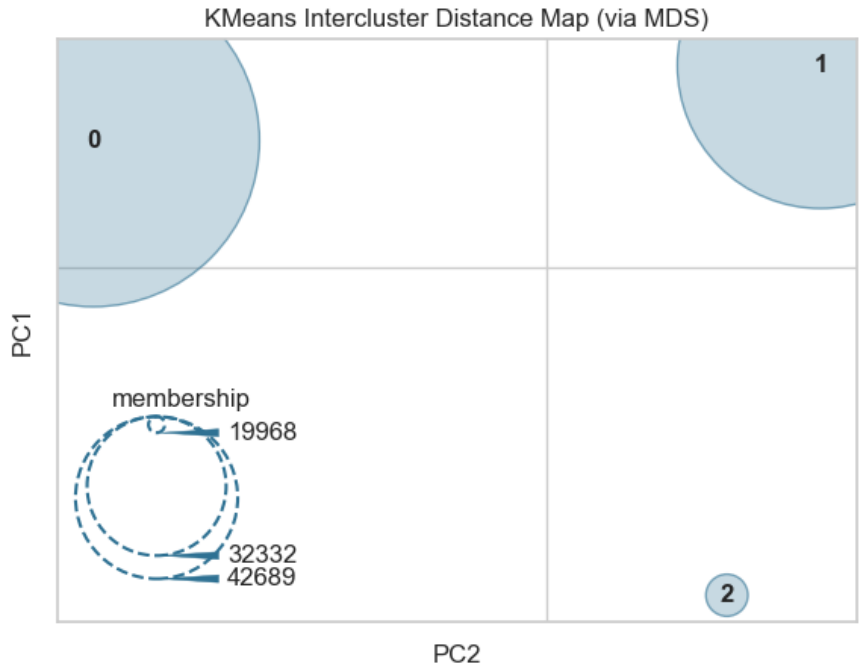
Variable Note

3 Clusters:

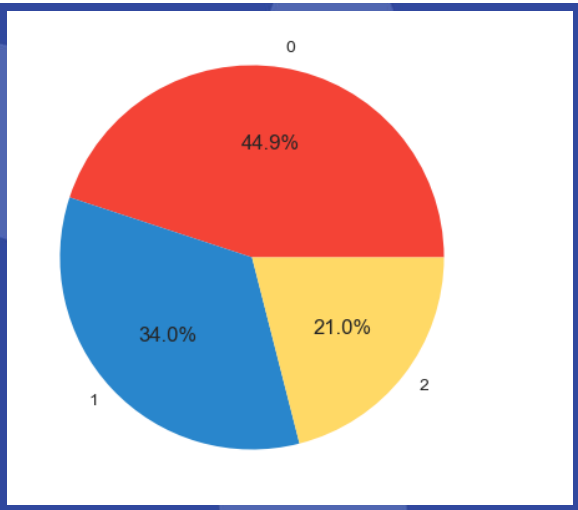






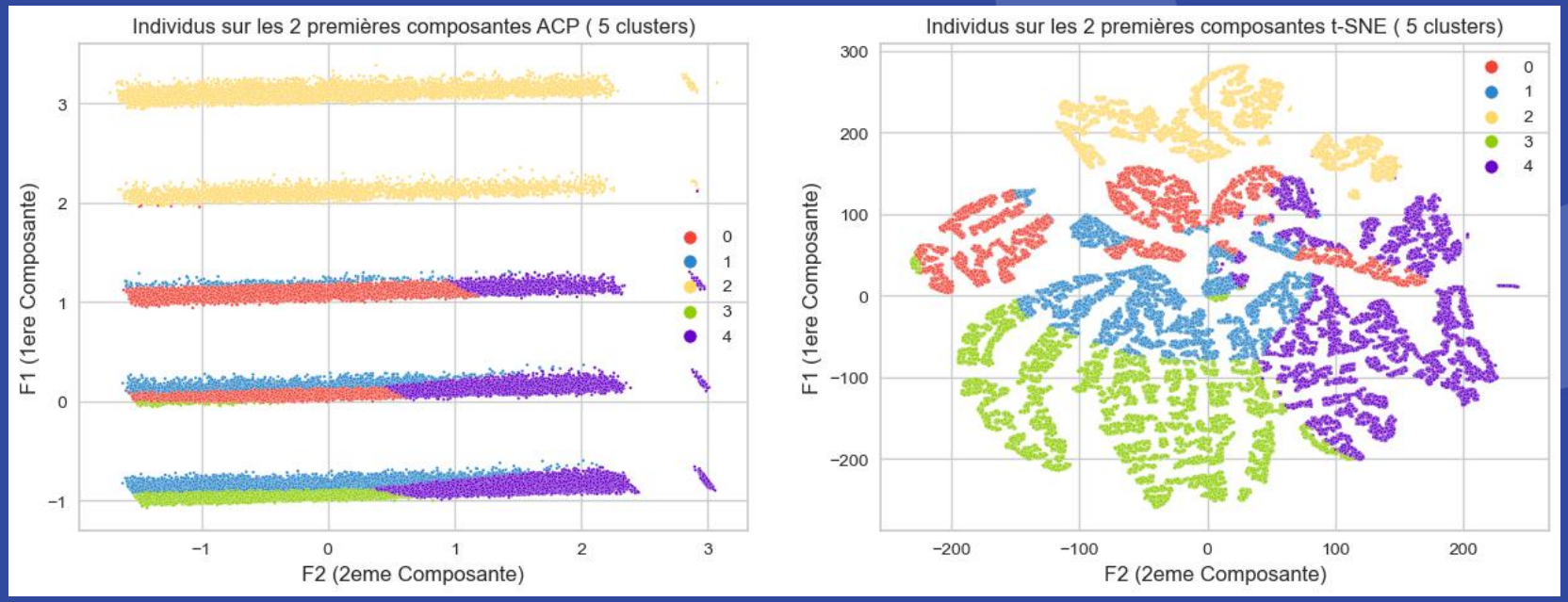


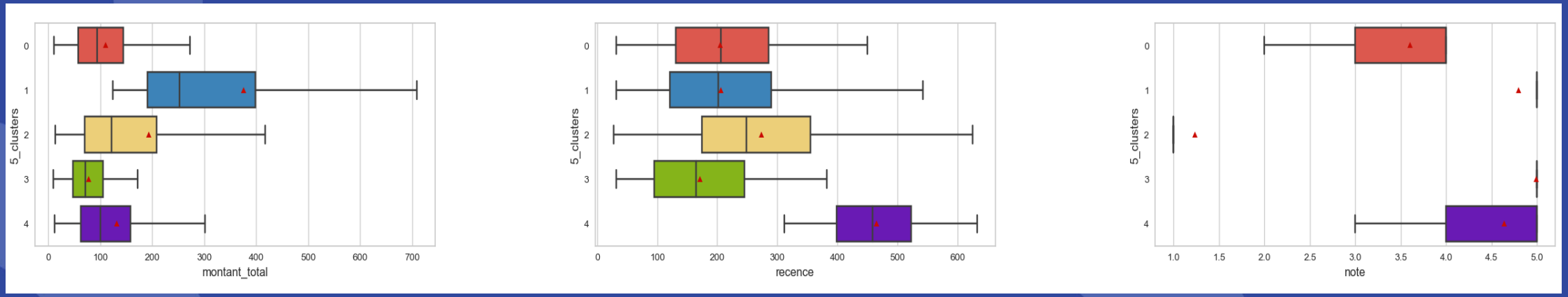
ARI moy = 1,0 (arrondi)
Stable mais trop généraliste



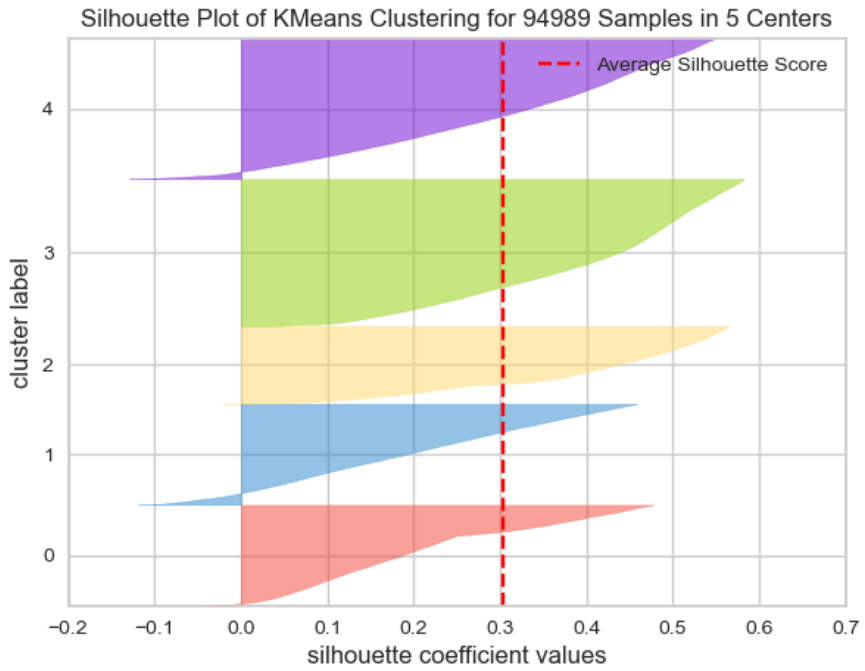
olist

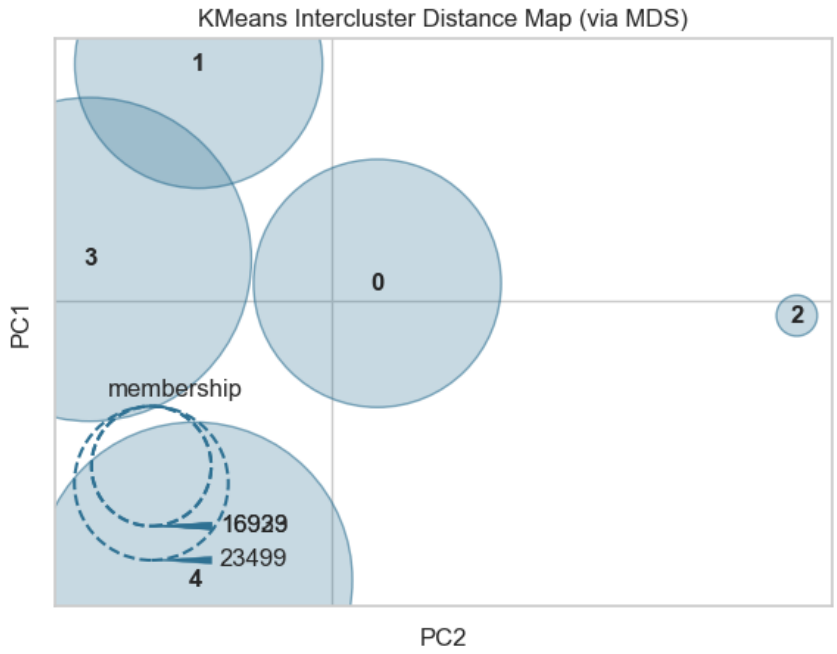
5 Clusters:





olist

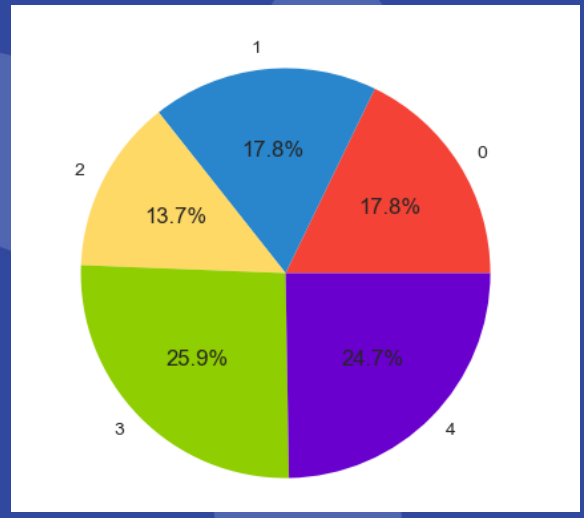




ARI moy = 0,94

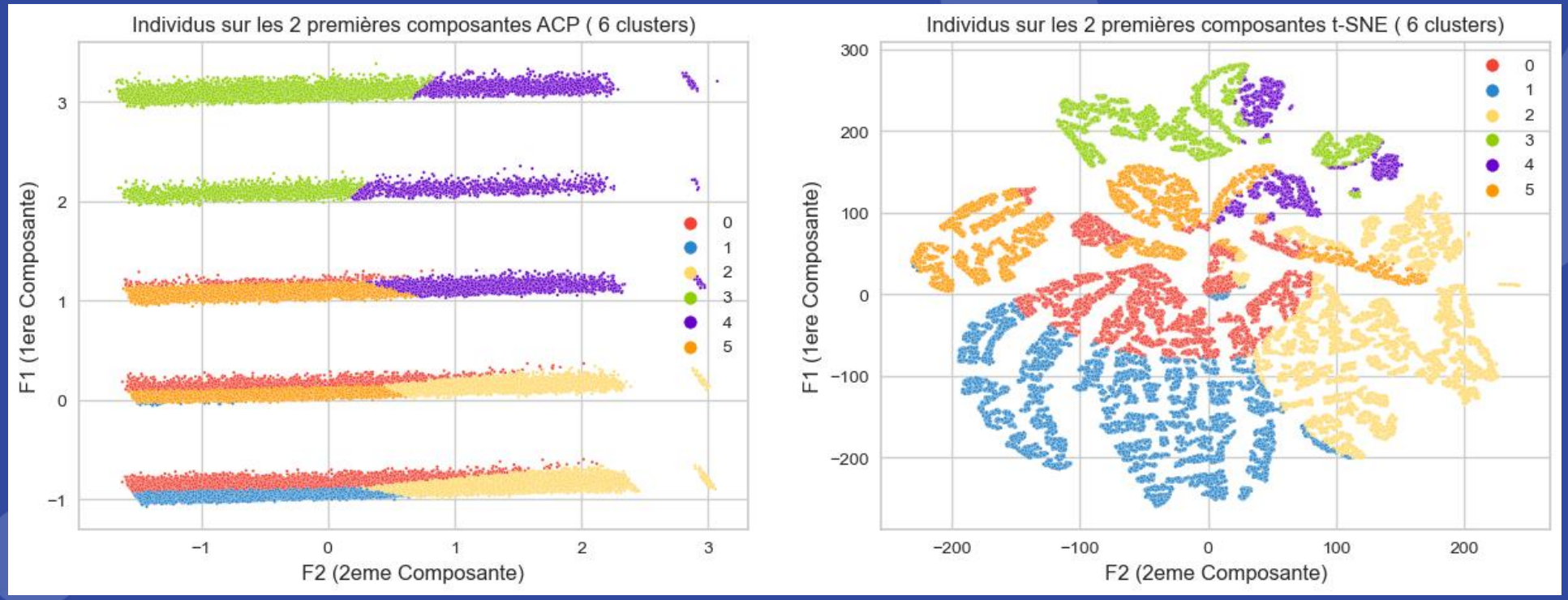
ARI Min = 0,74

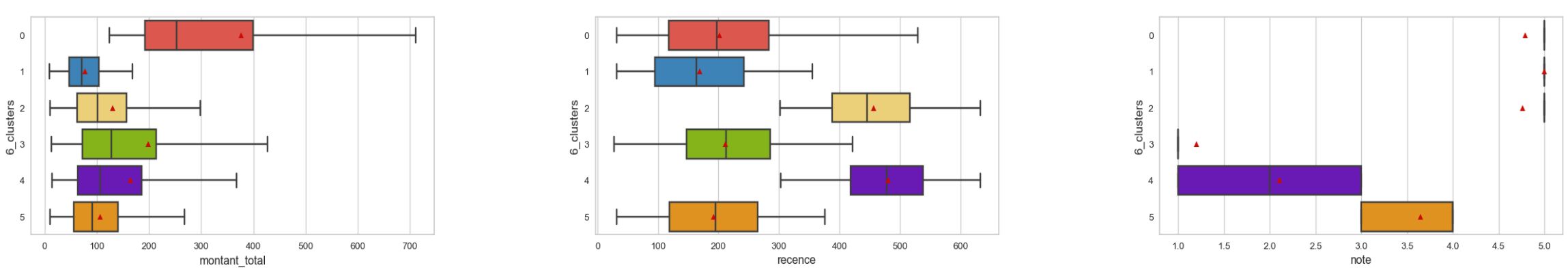
Perd en stabilité

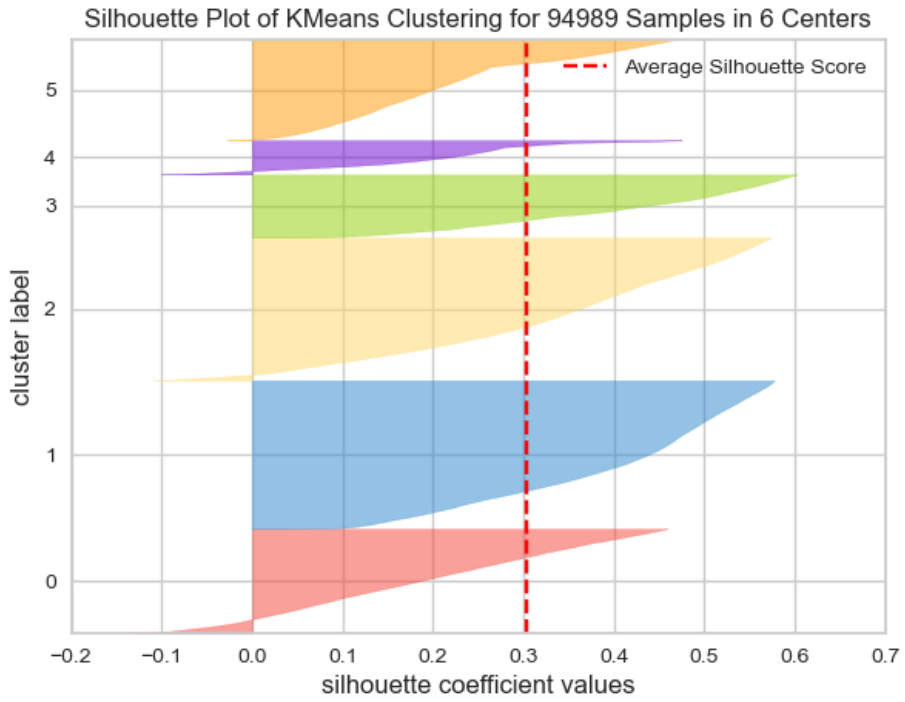


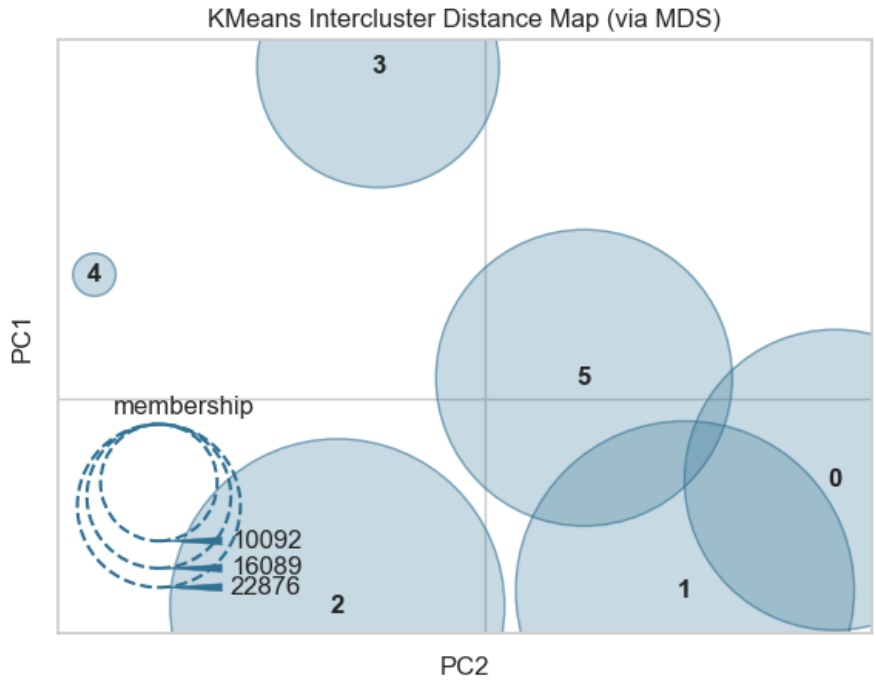
olist

6 Clusters:


Individus sur les 2 premières composantes ACP (6 clusters)
F1 (1ere Composante)
F2 (2eme Composante)
0
1
2
3
4
5
Individus sur les 2 premières composantes t-SNE (6 clusters)
F1 (1ere Composante)
F2 (2eme Composante)


6_clusters
montant_total
recence
note

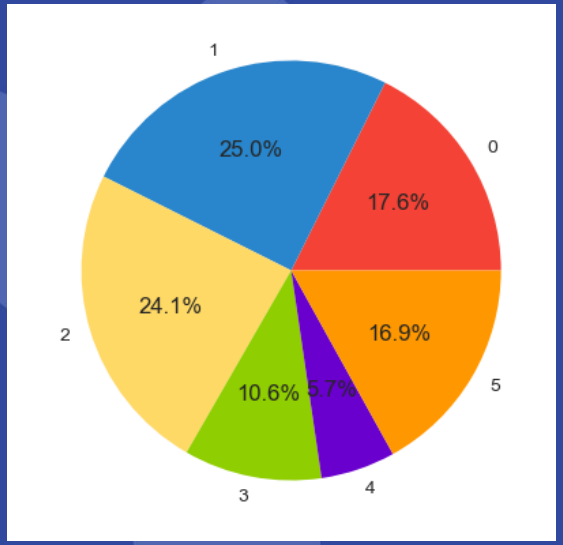




ARI moy = 0,96

ARI Min = 0,56

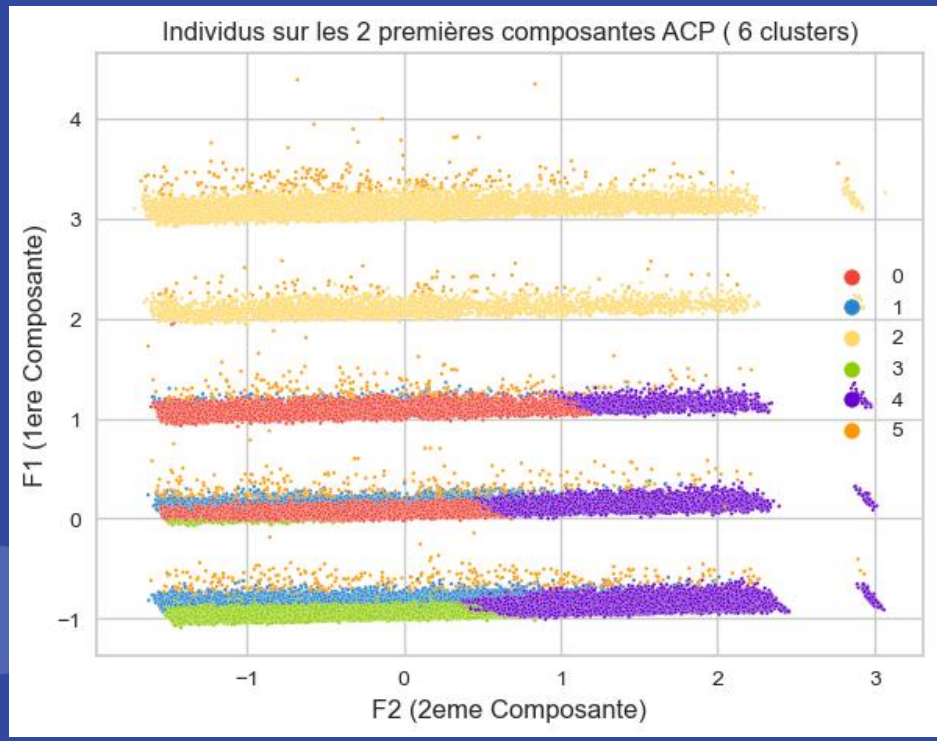
Perd encore en stabilité

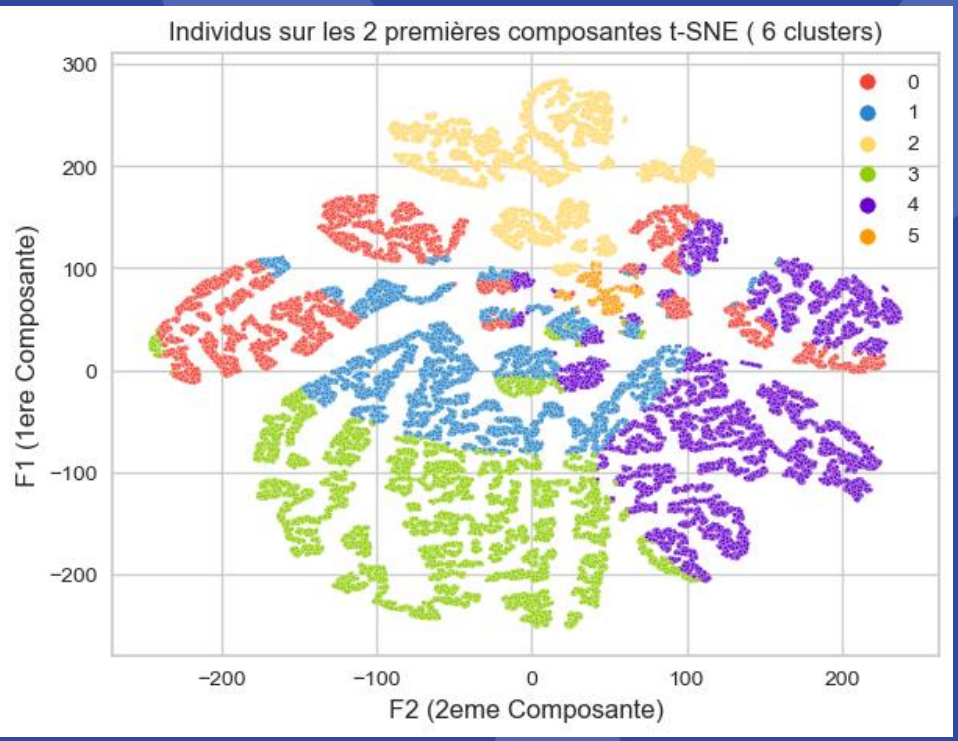


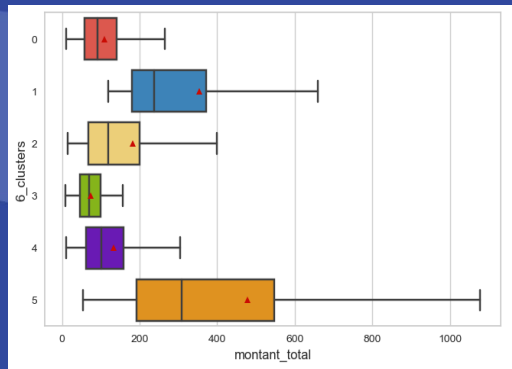
olist

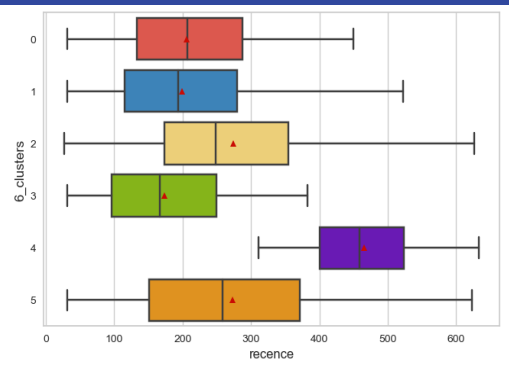
Variable nombre de produit moyen

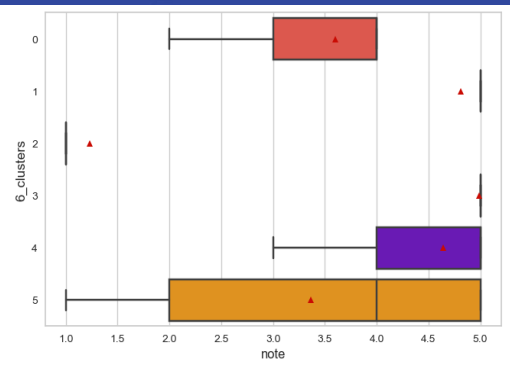
6 Clusters:

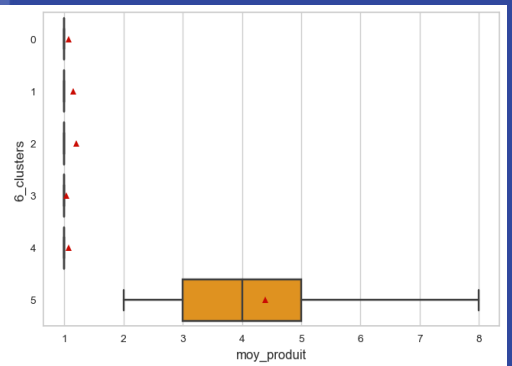




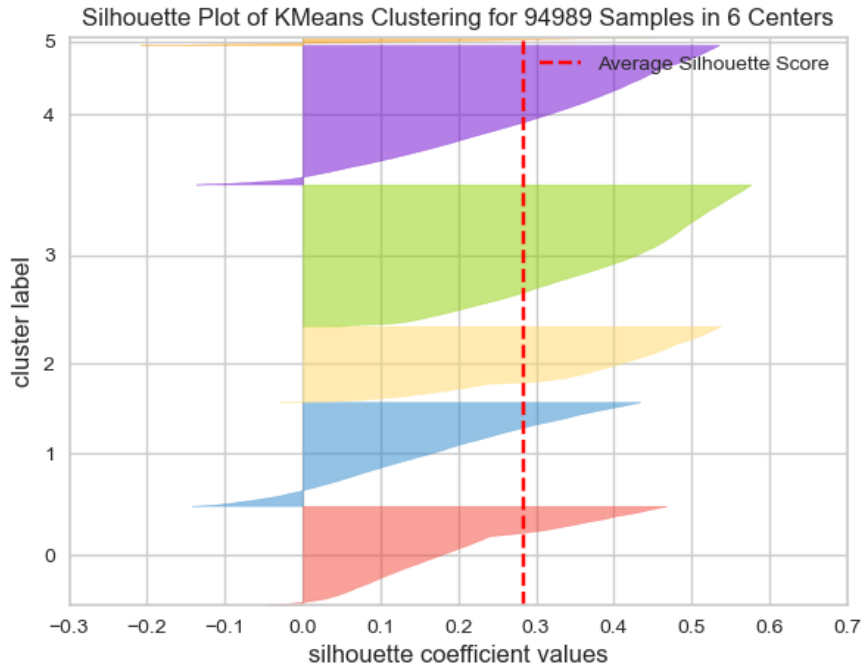


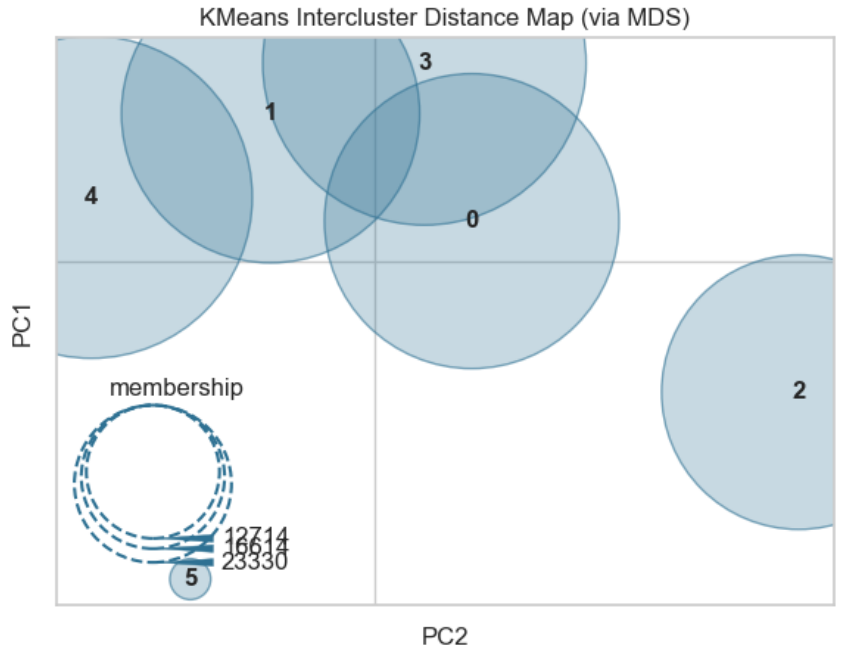






olist

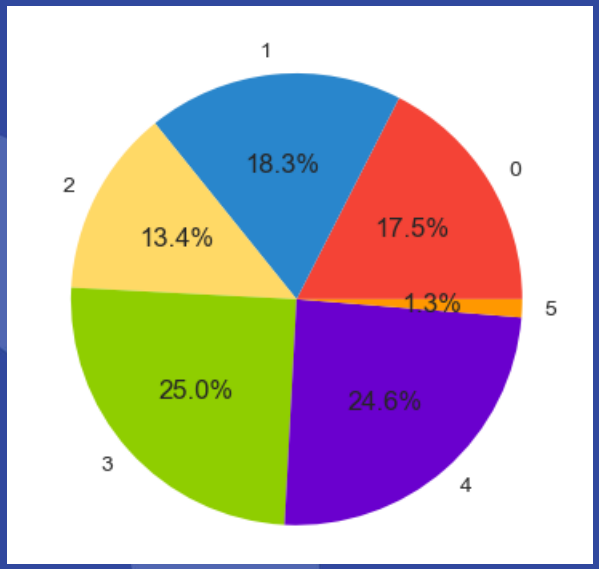




ARI moy = 0,9

ARI Min = 0,62

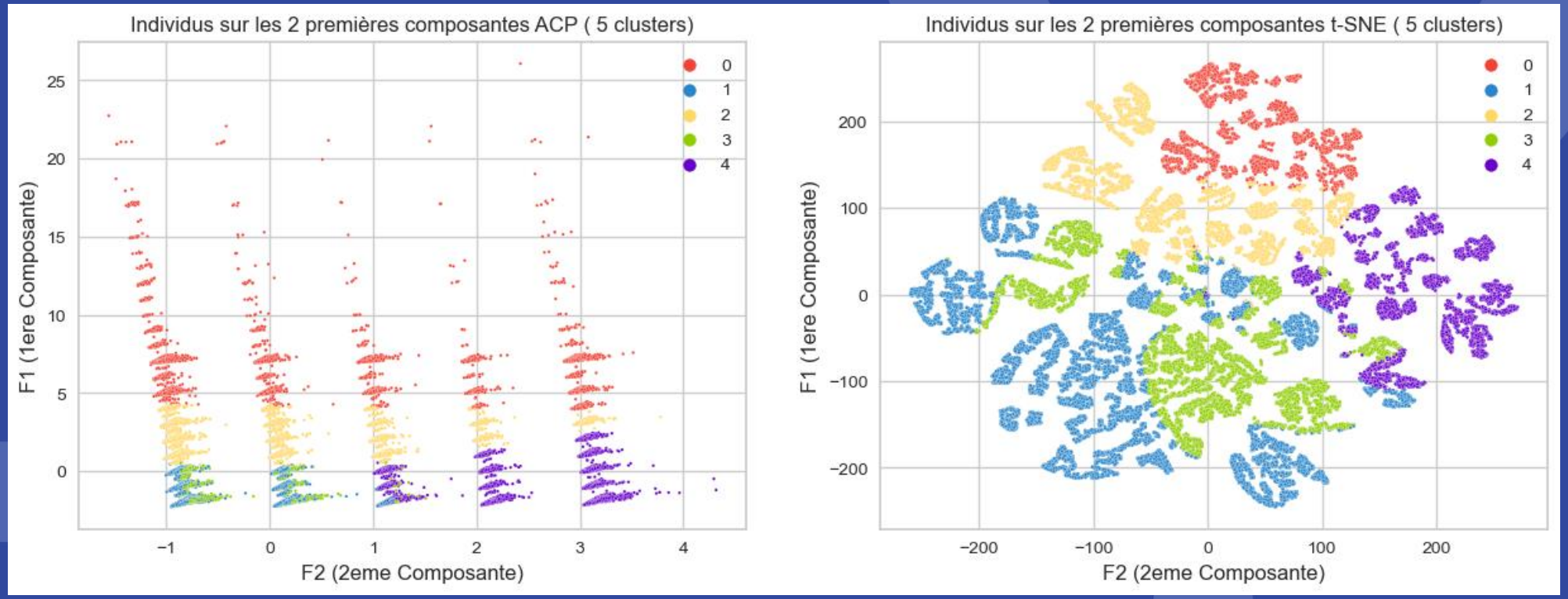
Pas très stable

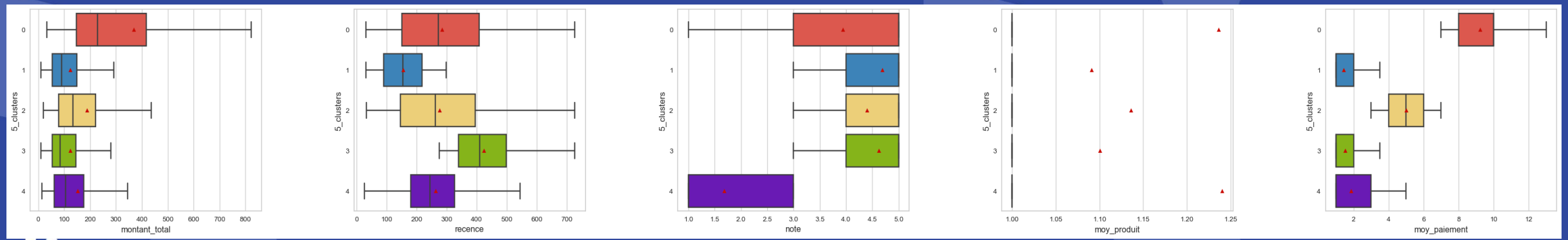


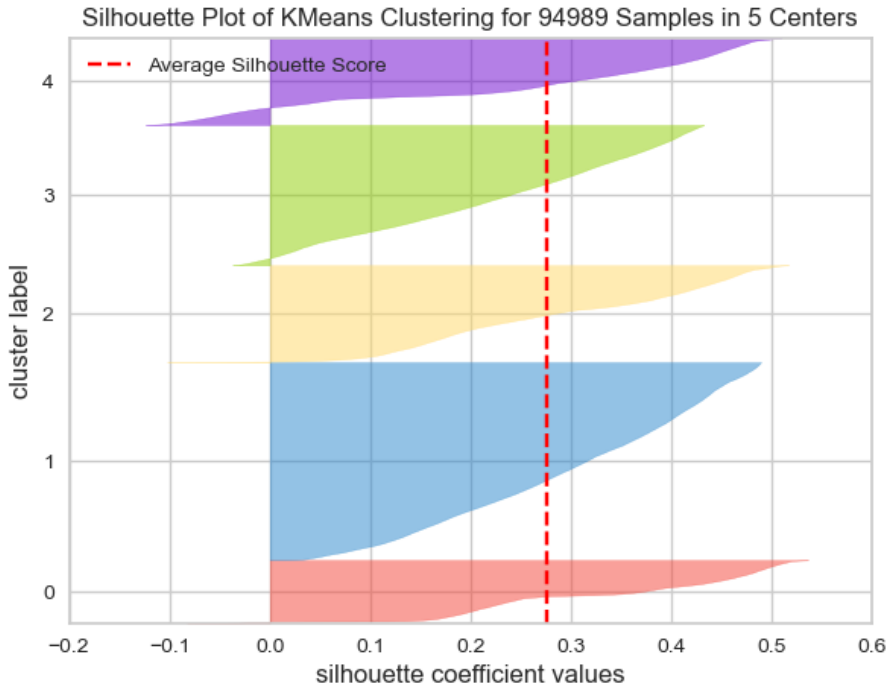
olist

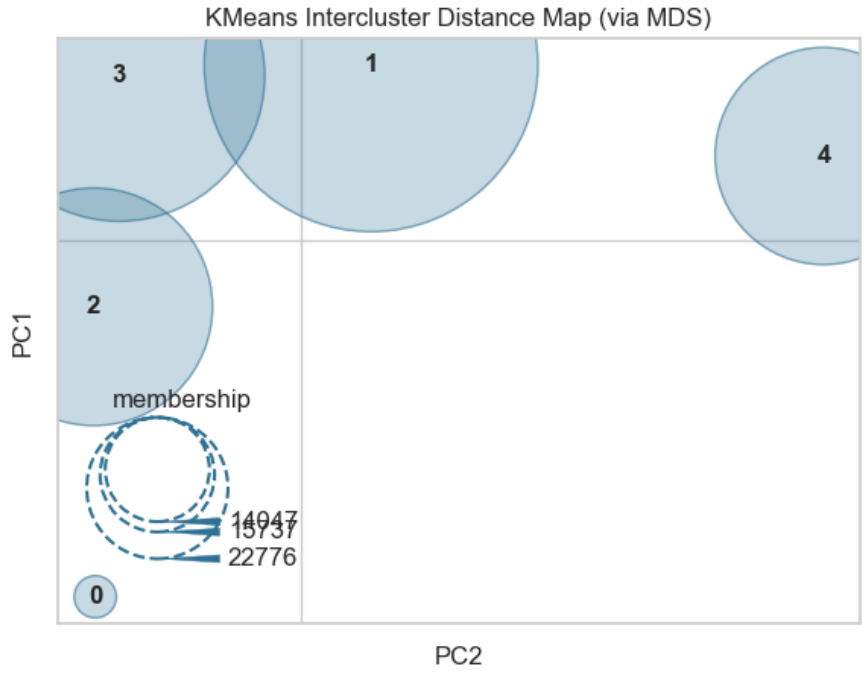
Variable nombre de paiement moyen

5 Clusters:





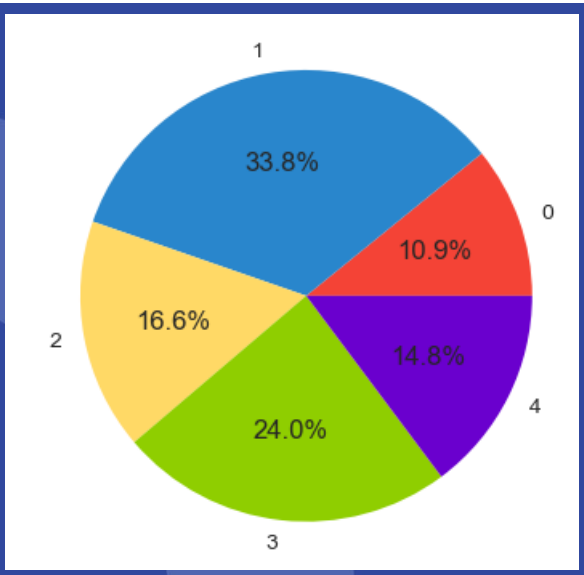




ARI moy = 0,99

ARI Min = 0,94

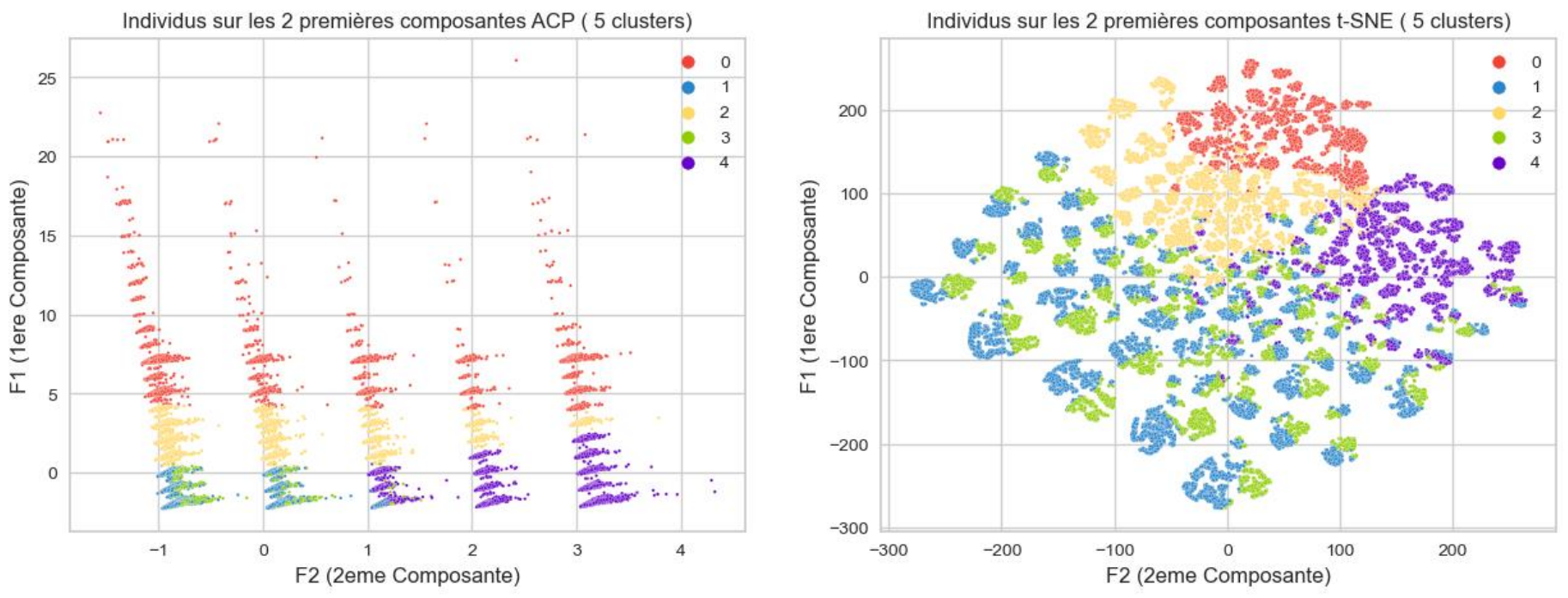
Très stable

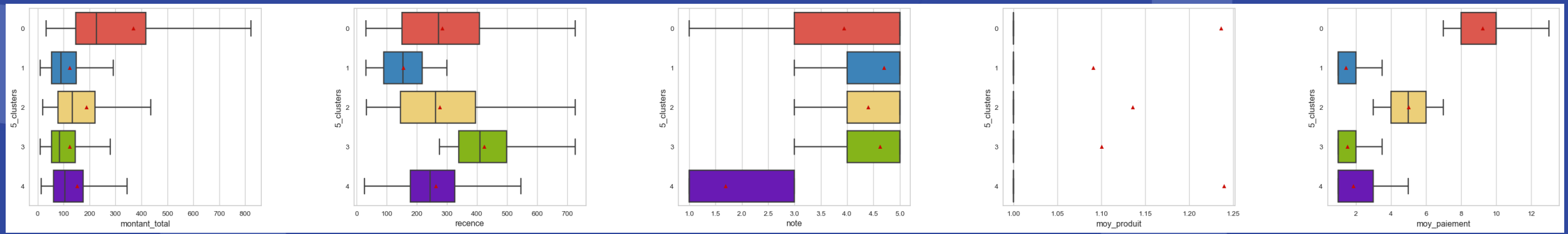


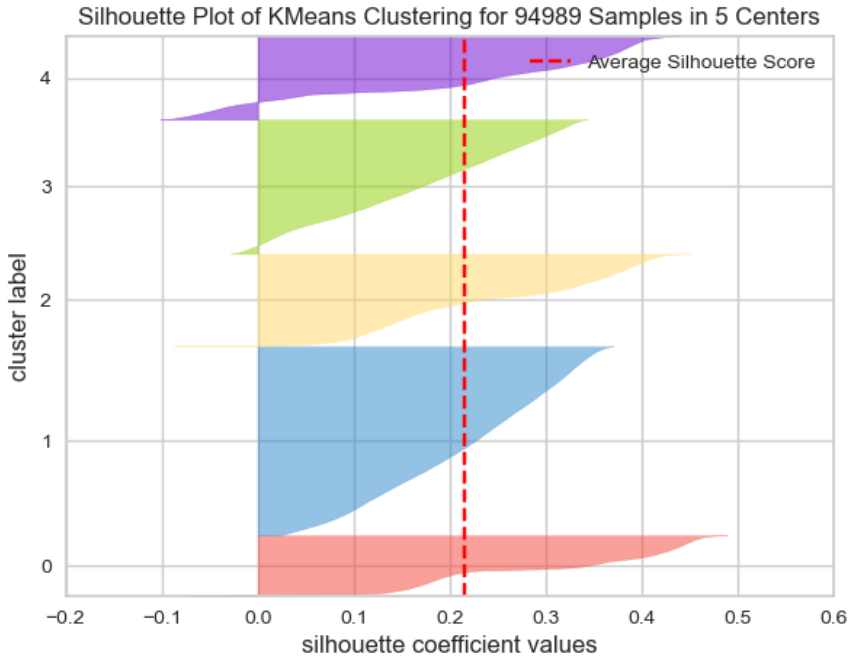
olist

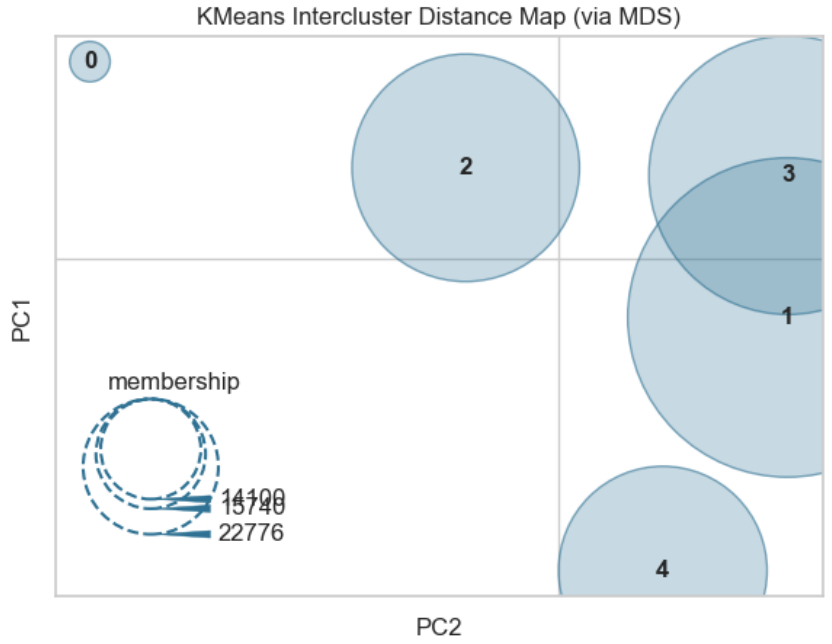
Variable groupe

5 Clusters:





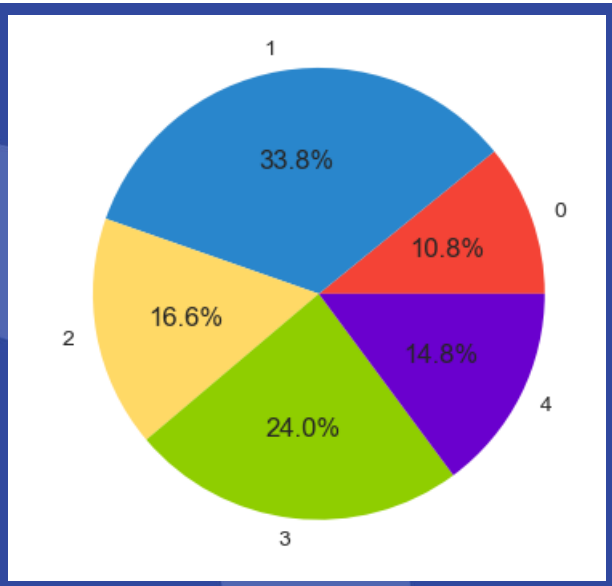




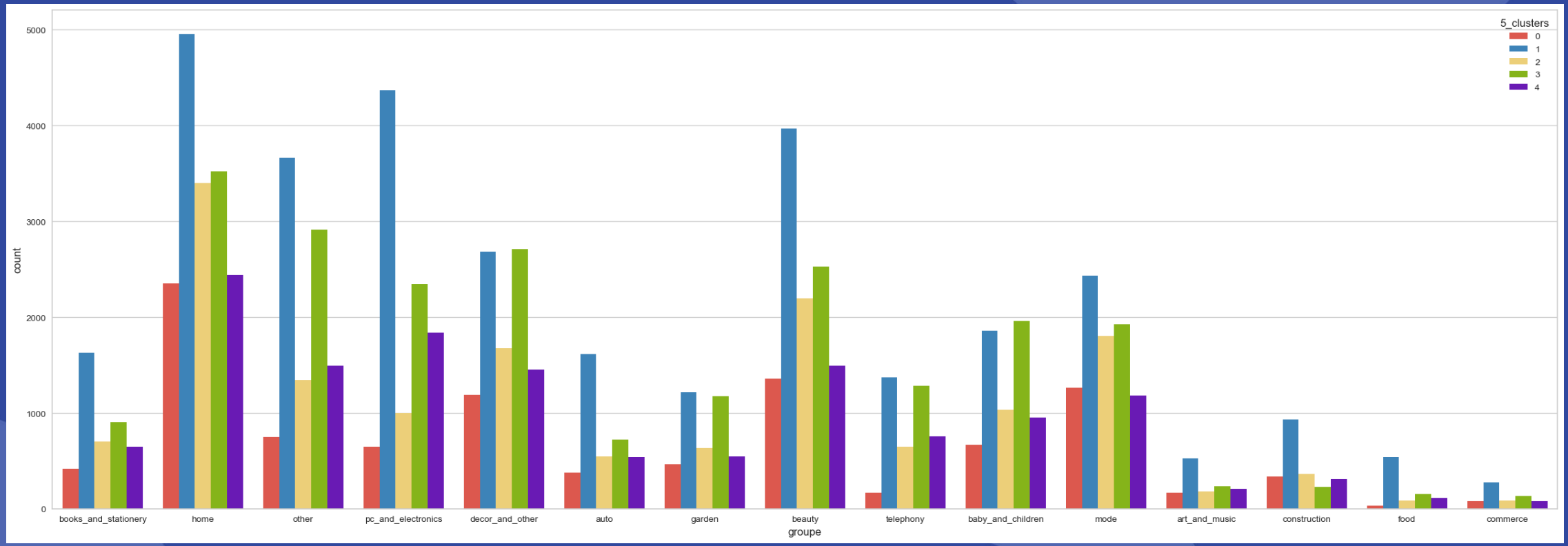
ARI moy = 1

ARI Min = 0,99

Très stable



olist

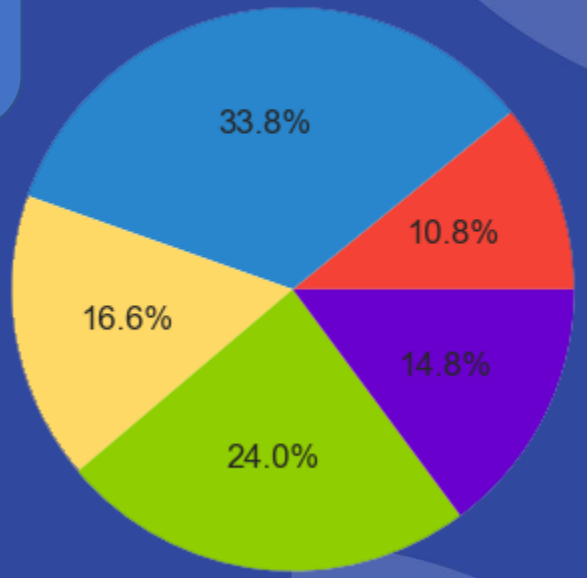

5_clusters
0
1
2
3
4
count
5000
4000
3000
2000
1000
0
books_and_stationery
home
other
pc_and_electronics
decor_and_other
auto
garden
beauty
telephony
baby_and_children
mode
art_and_music
construction
food
commerce
groupe

olist

Clustering sélectionné (k-Means)

Le nouveau client (régulier à en devenir):
- Dépense peu
- À commander dans les 6-7 derniers mois
- Satisfait des produits

Acheteur Compulsif:
- Dépense beaucoup
- Étale les paiements sur un grand nombre de mensualités



Commande ce dont il a besoin:
- Dépense raisonnablement
- Étale les paiements sur peu de mensualités

Client à séduire:
- Peu satisfait des produits

Client à récupérer:
- N'as pas commandé durant la dernière année
- Dépense peu mais, est satisfait des produits
- N'étale pas les paiements

olist

Lubin Stéphane - Mars 2023

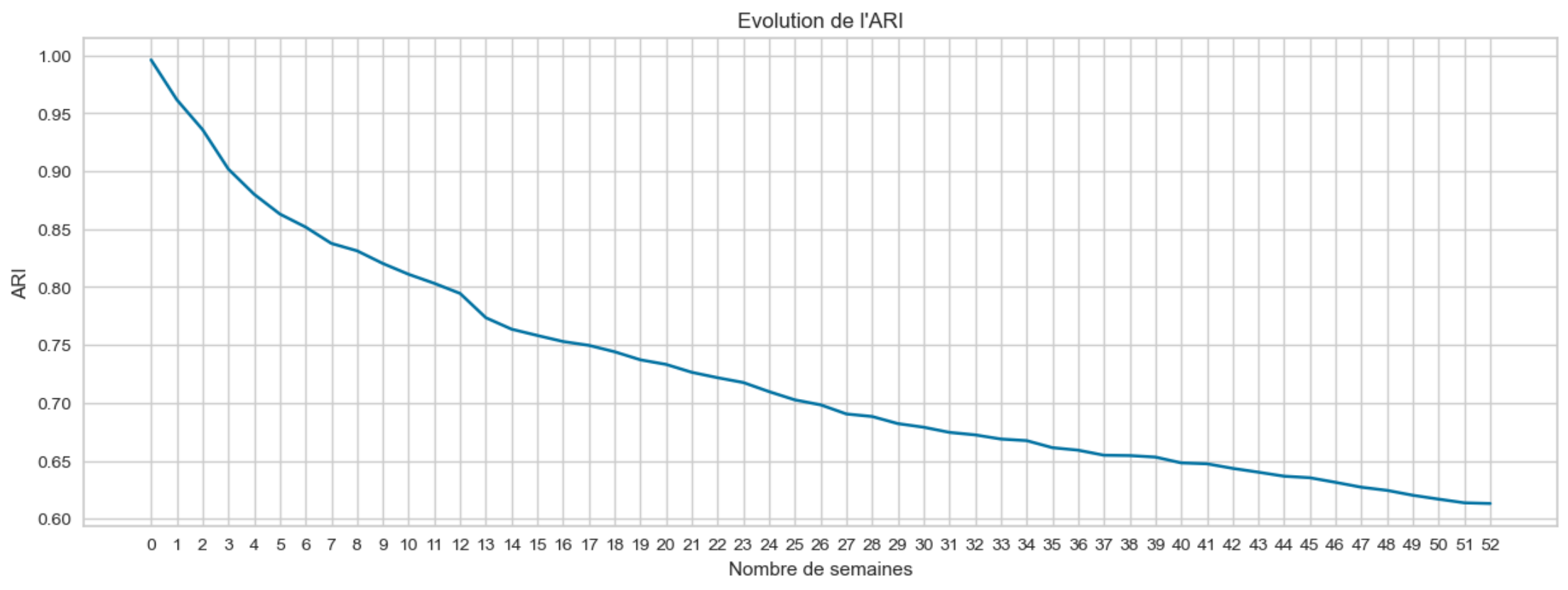
Maintenance du modèle

Procédure:

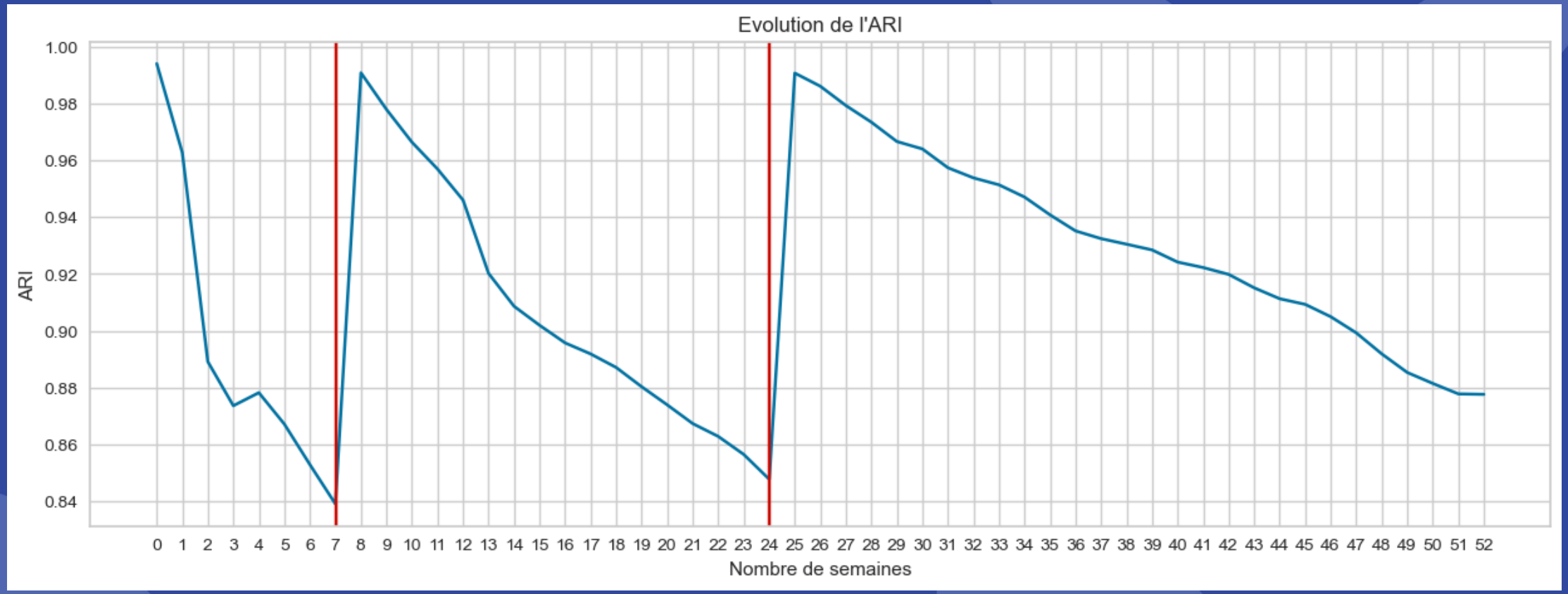
- On commence avec les commandes entre septembre 2016 et septembre 2017 ce qui représentent 23% de l'effectif total des clients.
- Puis on ajoute les commandes semaine par semaine sur 53 semaines, on observe l'évolution de l'ARI.
- On fixe un seuil à 0,85 pour l'ARI.
- On observe la fréquence à laquelle l'ARI atteint ce seuil.
- On définit un fréquence de mise à jour du cluster.

olist

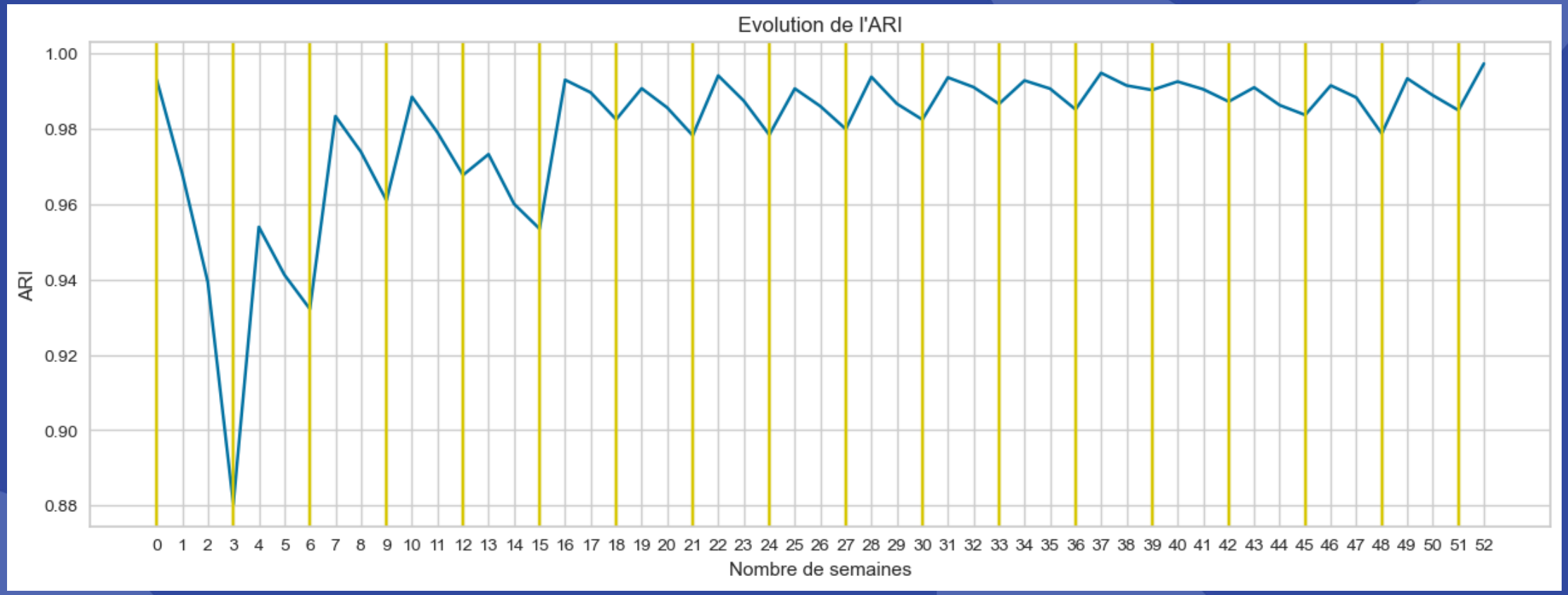
Evolution de l’ARI sur 53 semaines



olist


Evolution de l'ARI
ARI
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52
Nombre de semaines

olist



Une mise à jour avec une fréquence de 3 semaines semble bien adaptée,
ce qui conserve l'ARI entre 1 et 0,88

olist

MERCI DE VOTRE ATTENTION

olist

Lubin Stéphane - Mars 2023