

# **Segmentez des clients d'un site e-commerce**

**Joyce Kuoh Moukouri,  
P5, Soutenance du 08/05/2023**

# Ordre du jour

## Segmentez des clients d'un site e-commerce

1. La mission
  2. Feature engineering et exploration
  3. Approche de modélisation
    - A. Segmentation RFM
    - B. Segmentation RFM + profil économique
    - C. Segmentation RFM + préférence
    - D. Résultats
  4. Maintenance
- Conclusion

# 1. La mission

# La mission

Rappel des objectifs fixés par Juan de l'équipe marketing de Olist

- **Segmenter les clients du site Olist**
- Comprendre les différents utilisateurs et **transmettre les informations recueillies à l'équipe marketing**
- **Proposer un contrat de maintenance du modèle de segmentation**

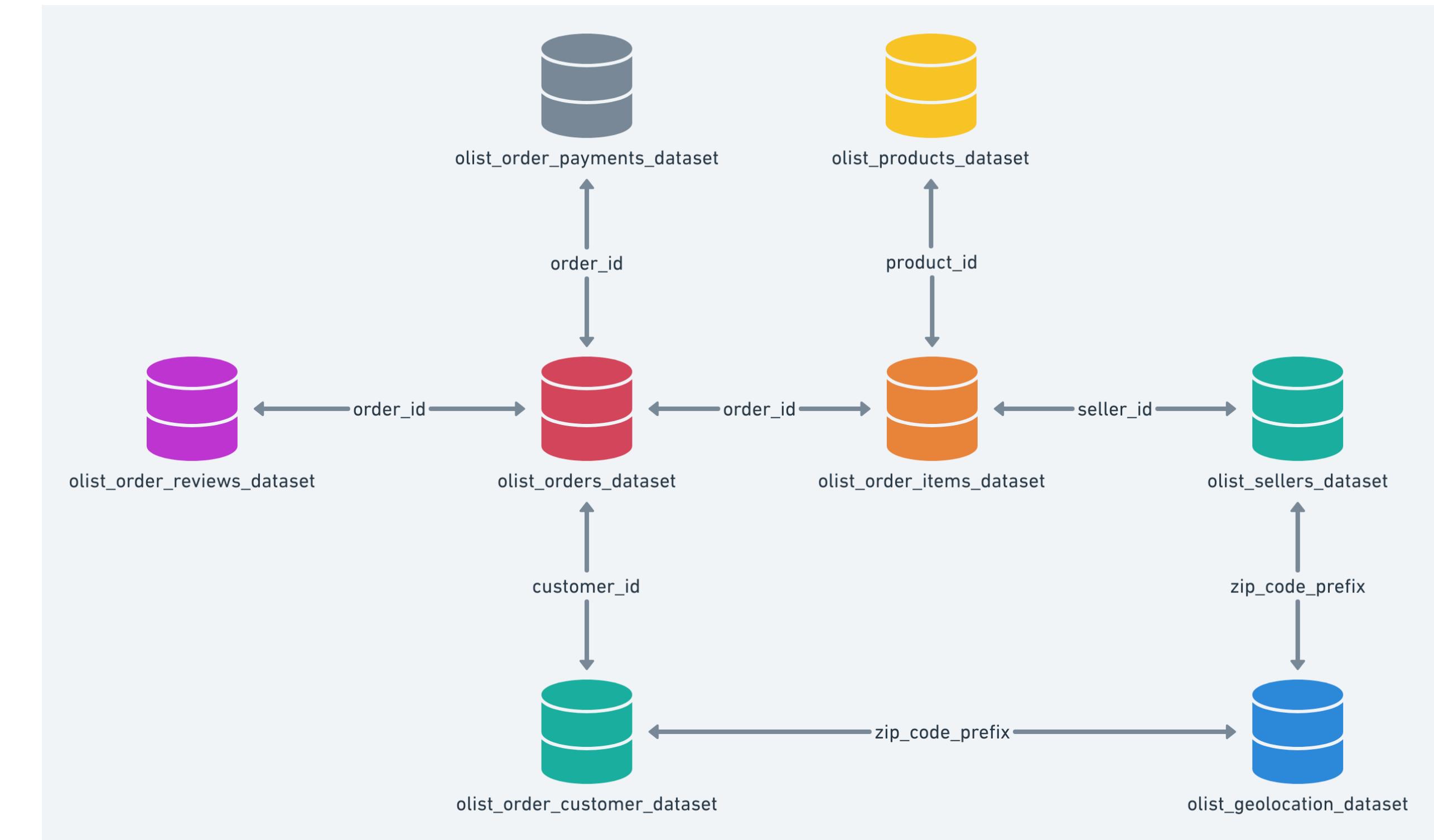
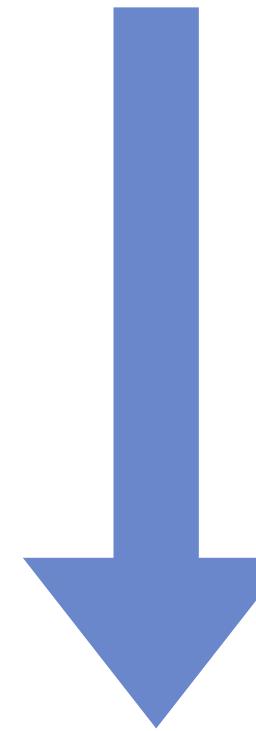
## **2. La base de données**

## 2. La base de données

- Base de données publiques de Olist
- Données accessibles en Open Data et gérées par Olist
- Environ 100k commandes faites entre 2016 et 2018
- 8 tables

## 2. La base de données

**Objectif :** Obtenir un fichier ‘client’ pour étudier le comportement de chaque individus



**Démarche :**

1. Création d'une base de données SQL
2. Jointure JOIN pour obtenir deux tables : **commande.csv** et **zipcode.csv**
3. Agrégation de la table commande pour obtenir la table : **clients.csv**

## 2. La base de données

### Nettoyage

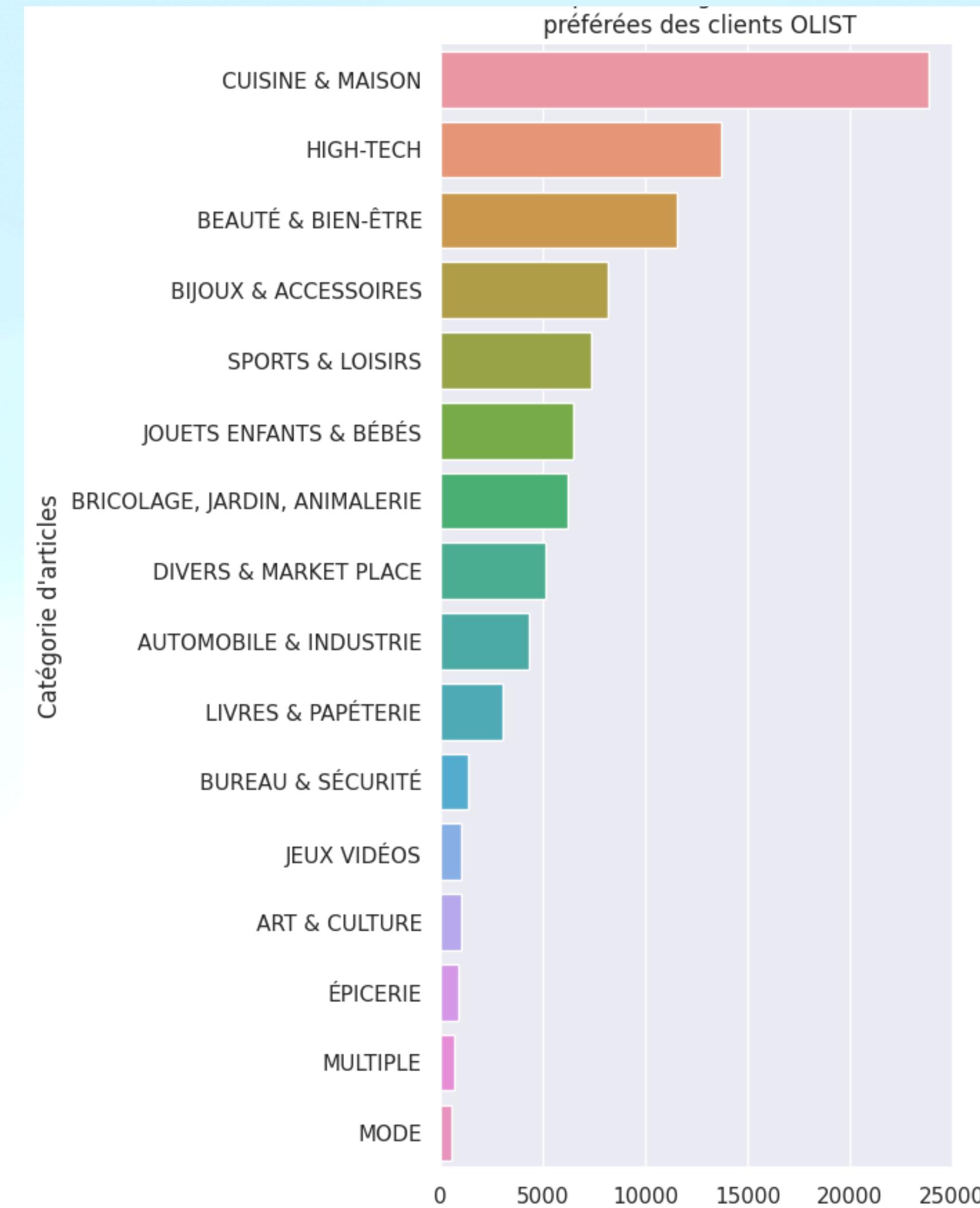
- Base de données de bonne qualité, peu de valeurs aberrantes (valeur aberrante de géolocalisation)
- Imputation des valeurs manquantes des données de géolocalisation avec la distance textuelle de Levenshtein

### Feature engineering

Création de variables	Modification de la variables catégorielles
Variables RFM : ' <b>récence_j</b> ', ' <b>fréquence</b> ' et ' <b>montant_moy</b> '	' <b>etat_client</b> ' : Chaque état fédéral est remplacé par le revenu moyen mensuel correspondant, données du gouvernement brésilien
' <b>Reste_échéance_mois</b> ' : Nombre de paiement restant depuis la dernière commande.	' <b>categorie</b> ' : Chaque catégorie favoris est remplacé par le montant moyen des transactions correspondant

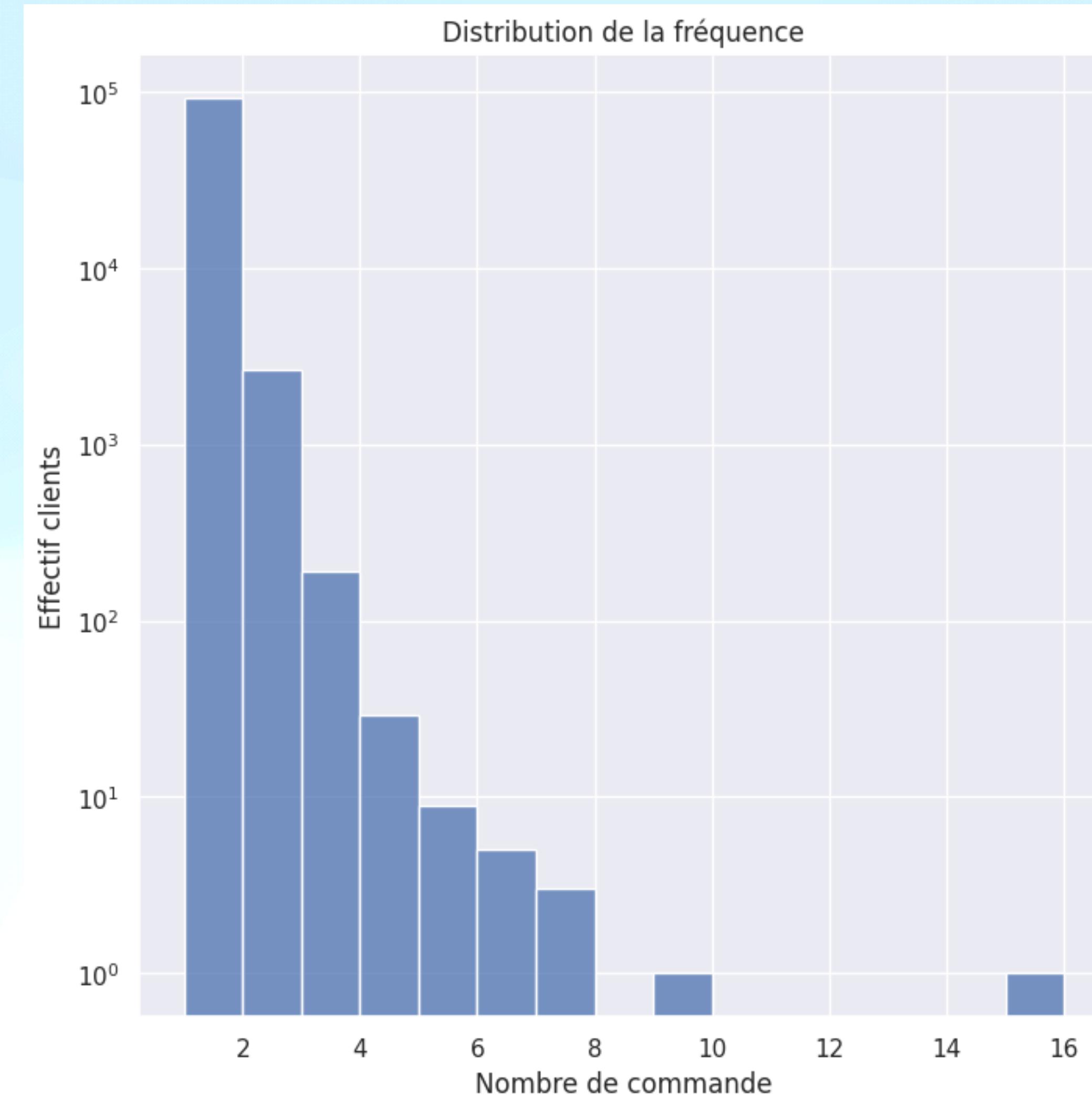
## 2. La base de données

### Exploration



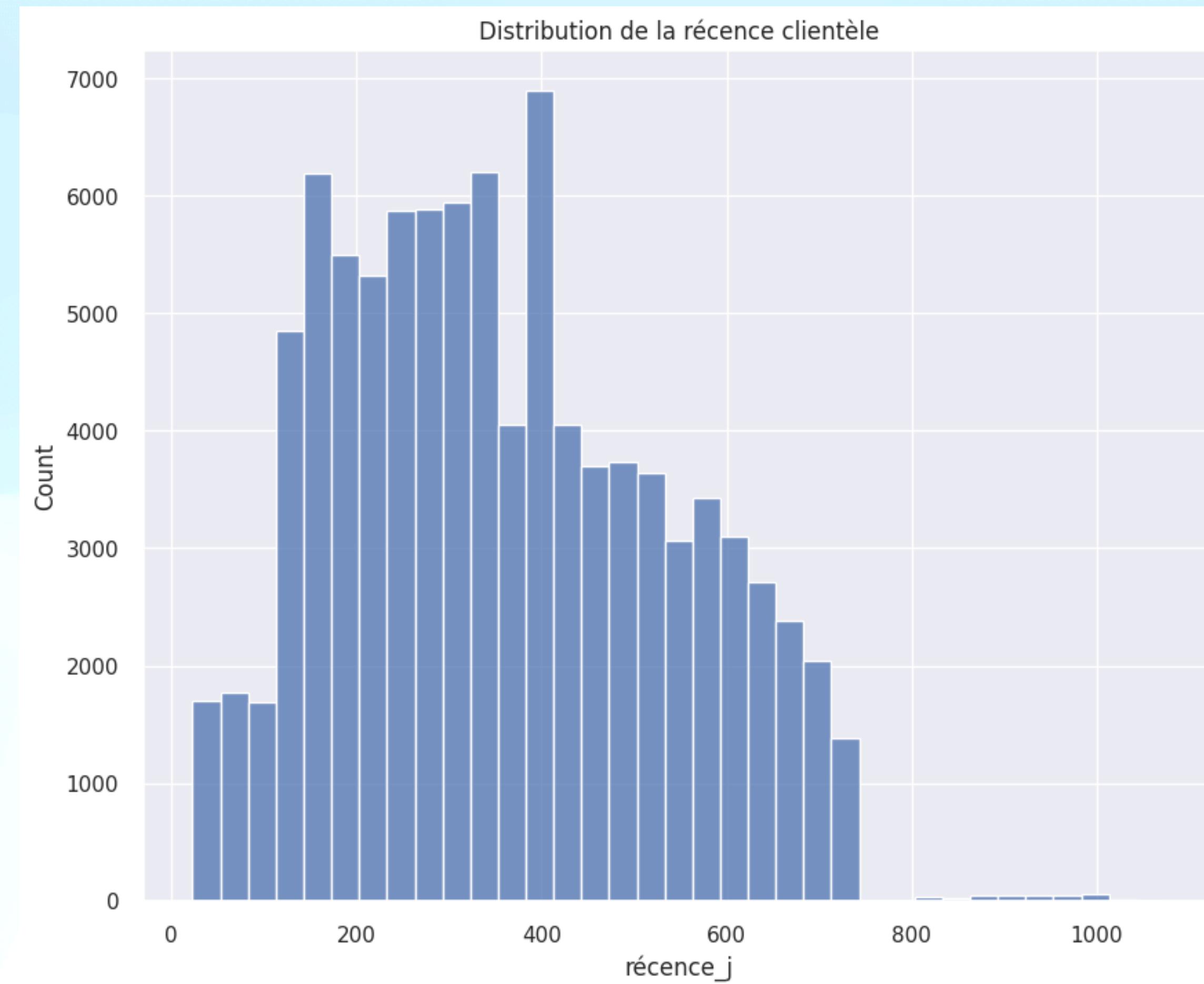
## 2. La base de données

### Exploration



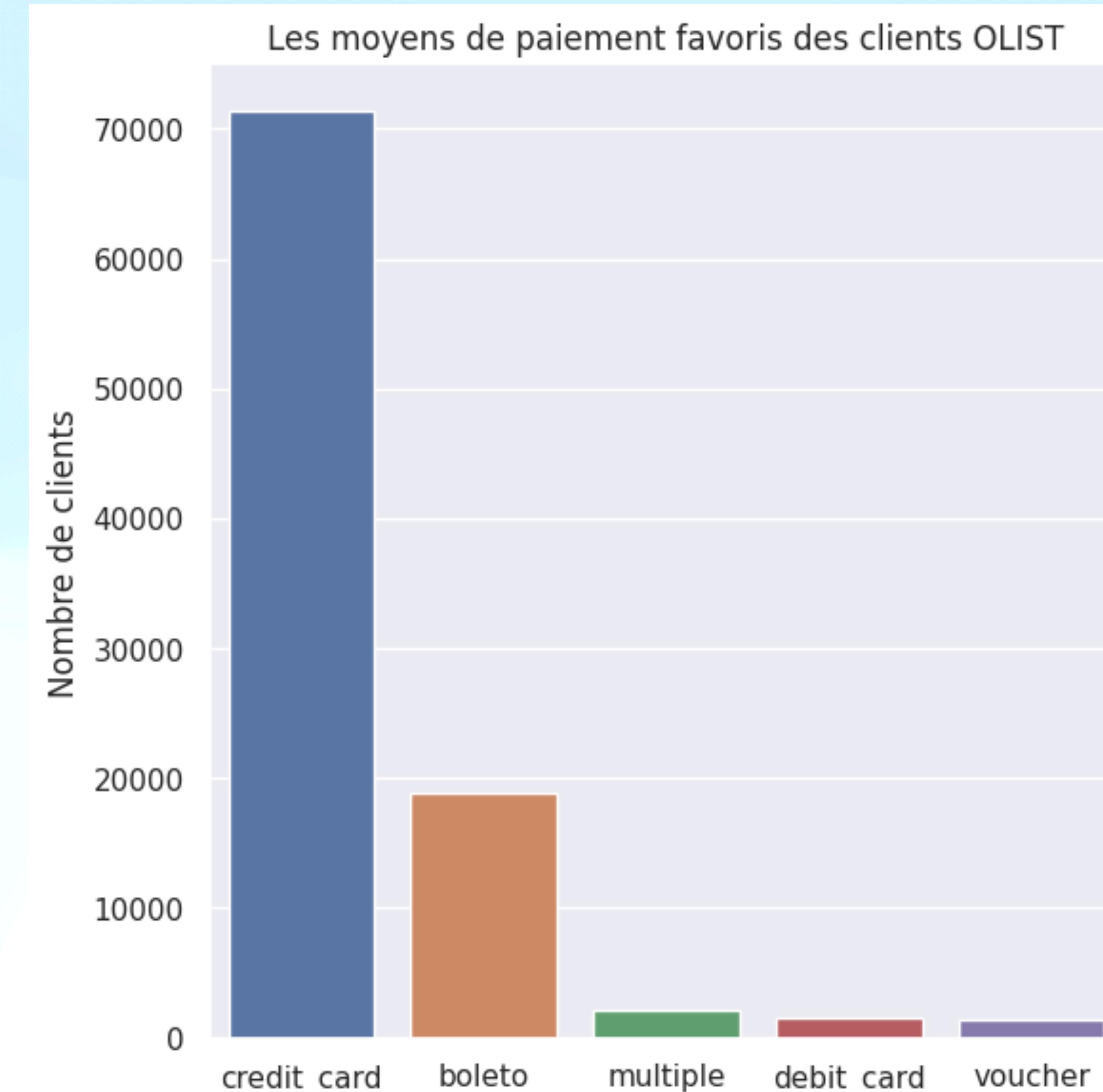
## 2. La base de données

### Exploration



## 2. La base de données

### Exploration



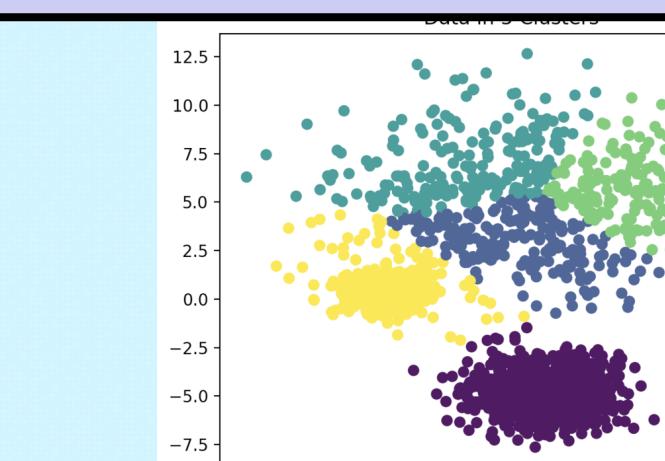
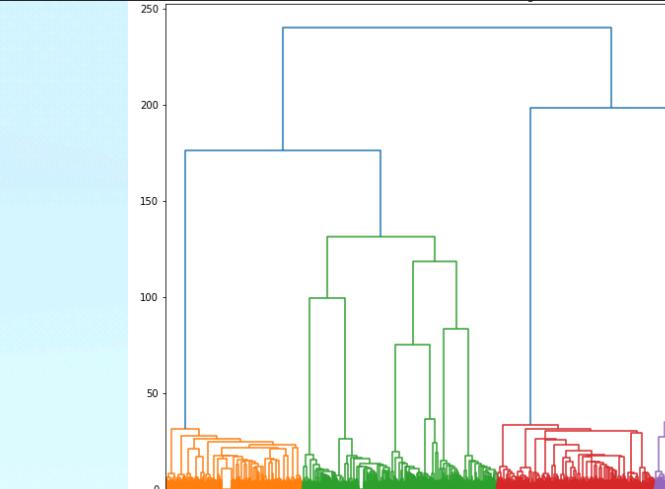
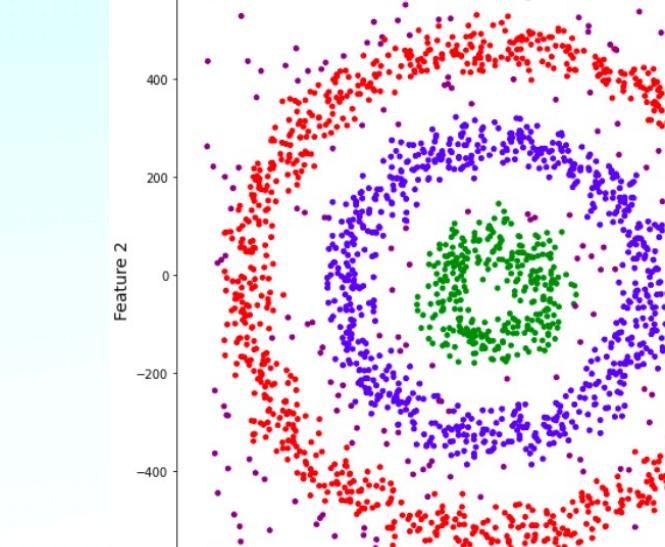
# **3. Approche de modélisation**

### 3. Approche de la modélisation

**Objectifs : Segmenter les clients et identifier les « meilleurs » clients**

- Diviser la population en sous-groupe et identifier des comportements moyens pour chacun des groupes
- Technique utilisé : le ***clustering***

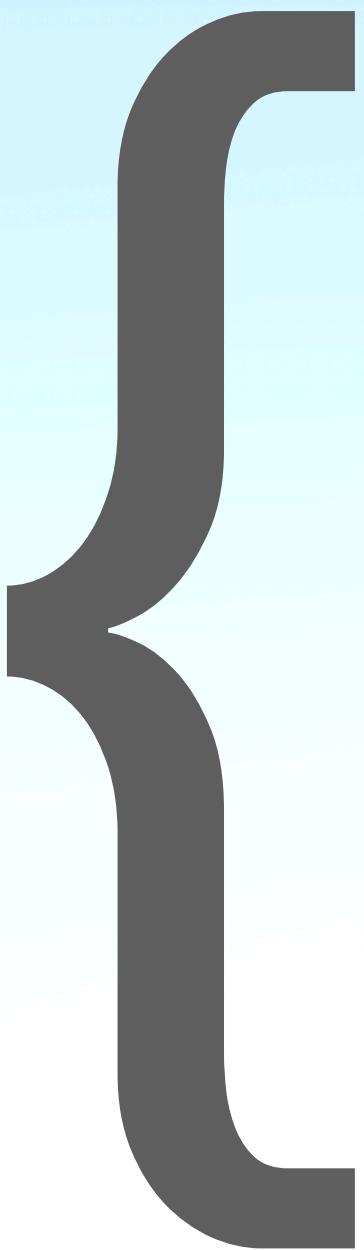
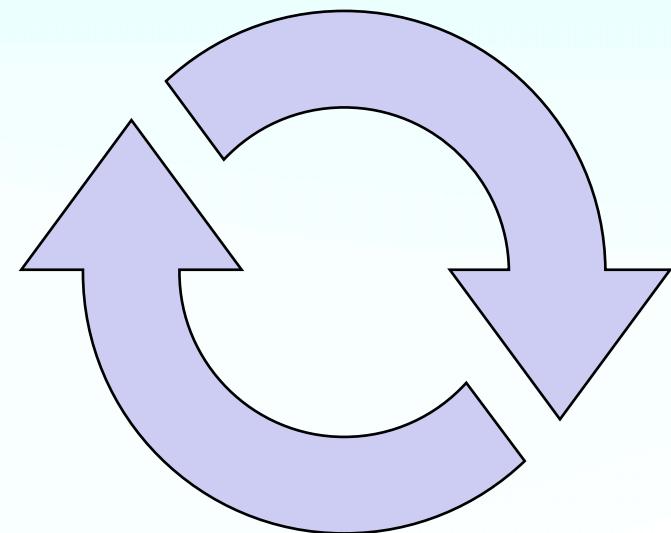
### 3. Approche de modélisation

Les modèles de clustering testés		Les paramètres sous sklearn
Kmeans()		<ul style="list-style-type: none"><li><code>k</code>, le nombre de cluster</li><li><code>init</code>, initialisation des centroïdes</li><li><code>algorithm</code>, algorithme de clustering</li><li><code>n_init</code>, le nombre d'initialisation</li></ul>
AgglomerativeClustering() - CAH		<ul style="list-style-type: none"><li><code>k</code>, le nombre de cluster</li><li><code>linkage</code>, la mesure de dissimilarité entre deux individus</li></ul>
DBSCAN()		<ul style="list-style-type: none"><li><code>epsilon</code>, le rayon maximum de la boule contenant un cluster</li><li><code>min_samples</code>, le nombre d'individus minimal pour définir un cluster</li><li><code>p</code>, l'indice définissant la norme de distance <math>p=1</math> norme 1, <math>p=2</math>, norme 2</li></ul>
Kmeans() x CAH x K_means()	<p>Méthode mixte.</p> <ol style="list-style-type: none"><li>1. On lance le K_means en stipulant <code>k_init = 1000</code> classes.</li><li>2. On utilise les <code>k_init</code> centroïdes pour effectuer un clustering de type CAH en cherchant le bon nombre de cluster <code>k_final</code></li><li>3. Je relance K_means en précisant les centroïdes <code>k_final</code> obtenu à l'étape 2.</li></ol>	

# 3. Approche de modélisation

Les segmentation clients testés		Les features
Set 1	Segmentation RFM	<ul style="list-style-type: none"><li>• <code>récence_j</code>, nombre de jour entre le 31 décembre 2018 et la dernière commande du client</li><li>• <code>fréquence</code>, nombre de commande entre 2016 et 2018</li><li>• <code>montant_moy</code>, montant moyen par commande</li></ul>
Set 2	Segmentation RFM + profil économique	<p>RFM</p> <ul style="list-style-type: none"><li>+ <code>'reste_echeance_mois'</code>, nombre d'échéance de paiement approximatif restant depuis la dernière commande</li><li>+ <code>'etat_client'</code>, Chaque état fédéral est remplacé par le revenu moyen mensuel correspondant, données du gouvernement brésilien</li></ul>
Set 3	Segmentation RFM +préférences	<p>RFM</p> <ul style="list-style-type: none"><li>+ <code>'catégorie'</code>, Chaque catégorie favoris est remplacé par le montant moyen des transactions correspondant</li><li>+ <code>'critique_note'</code>, note moyenne attribué par chaque client</li></ul>

Pour chaque modèle (excepté DBSCAN, phase 3 directement)



## PHASE 1

### Préparation du dataset

Choix des features de segmentation

Feature Engineering

Échantillonnage du dataset (10% du dataset)

## PHASE 2

### Recherche du nombre de classe $k$ optimal approximatif

Courbe de l'inertie et technique du coude

Courbe du score de silhouette

## PHASE 3

### Recherche des hyperparamètres précision du $k$ optimal

Grille de recherche à l'aide de *itertools*.  
Évaluation du score de silhouette

Recherche de la distribution du score de silhouette optimale

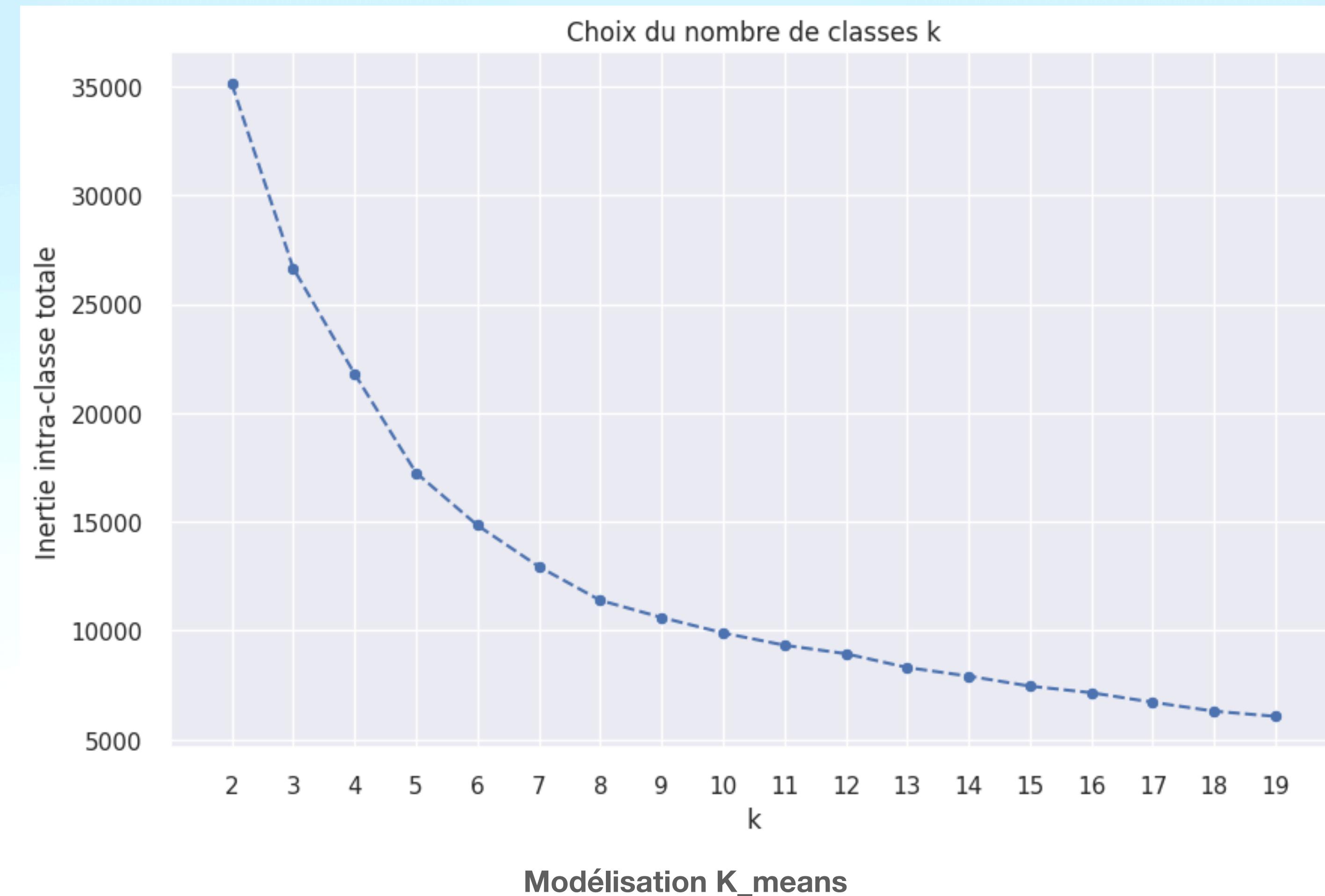
Choix d'un point de vu métier

## PHASE 4

### Choix du modèle optimal

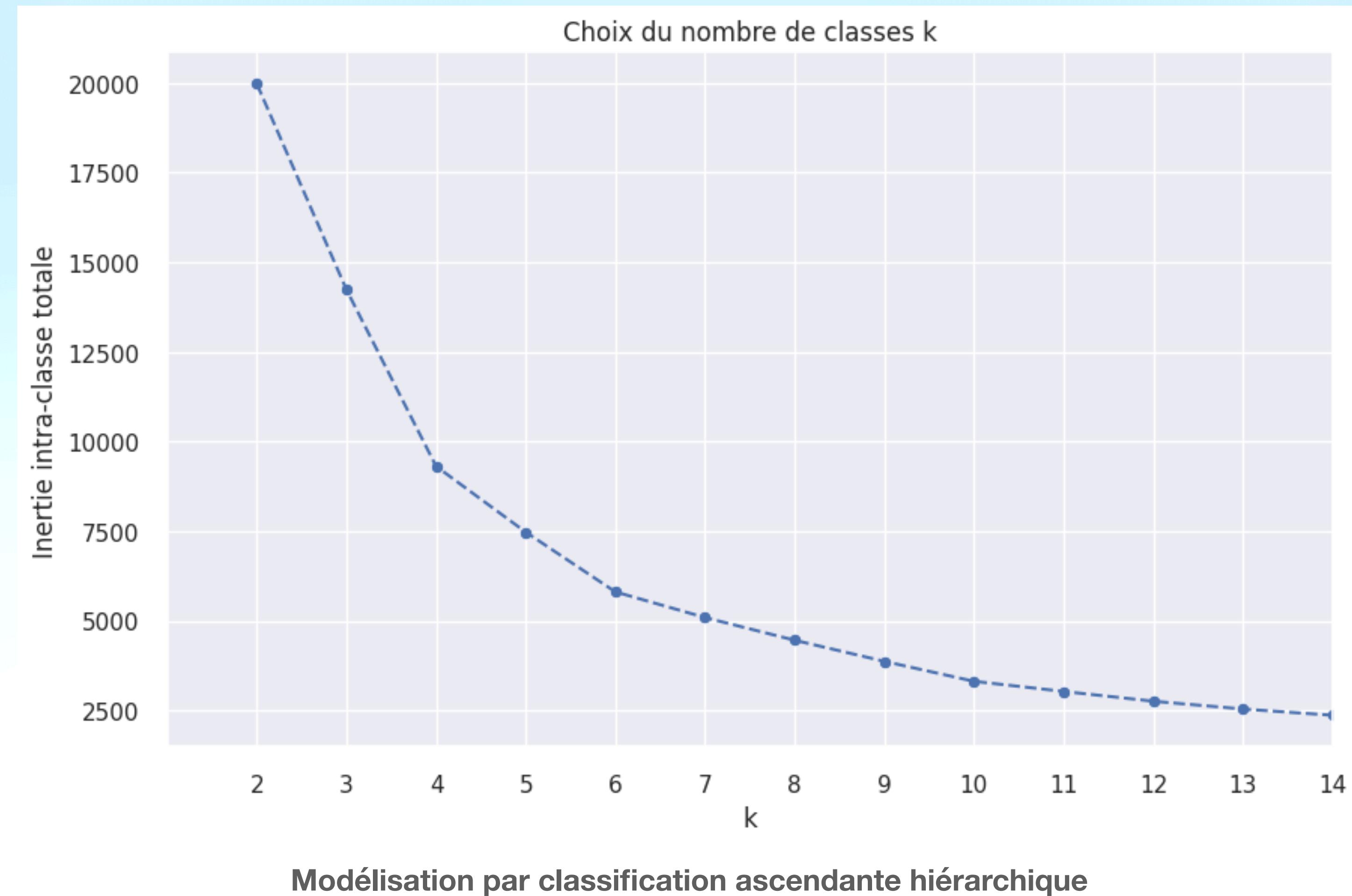
# A. Segmentation RFM

Choix du nombre de classes  $k$  (Kmeans, CAH, méthode mixte)



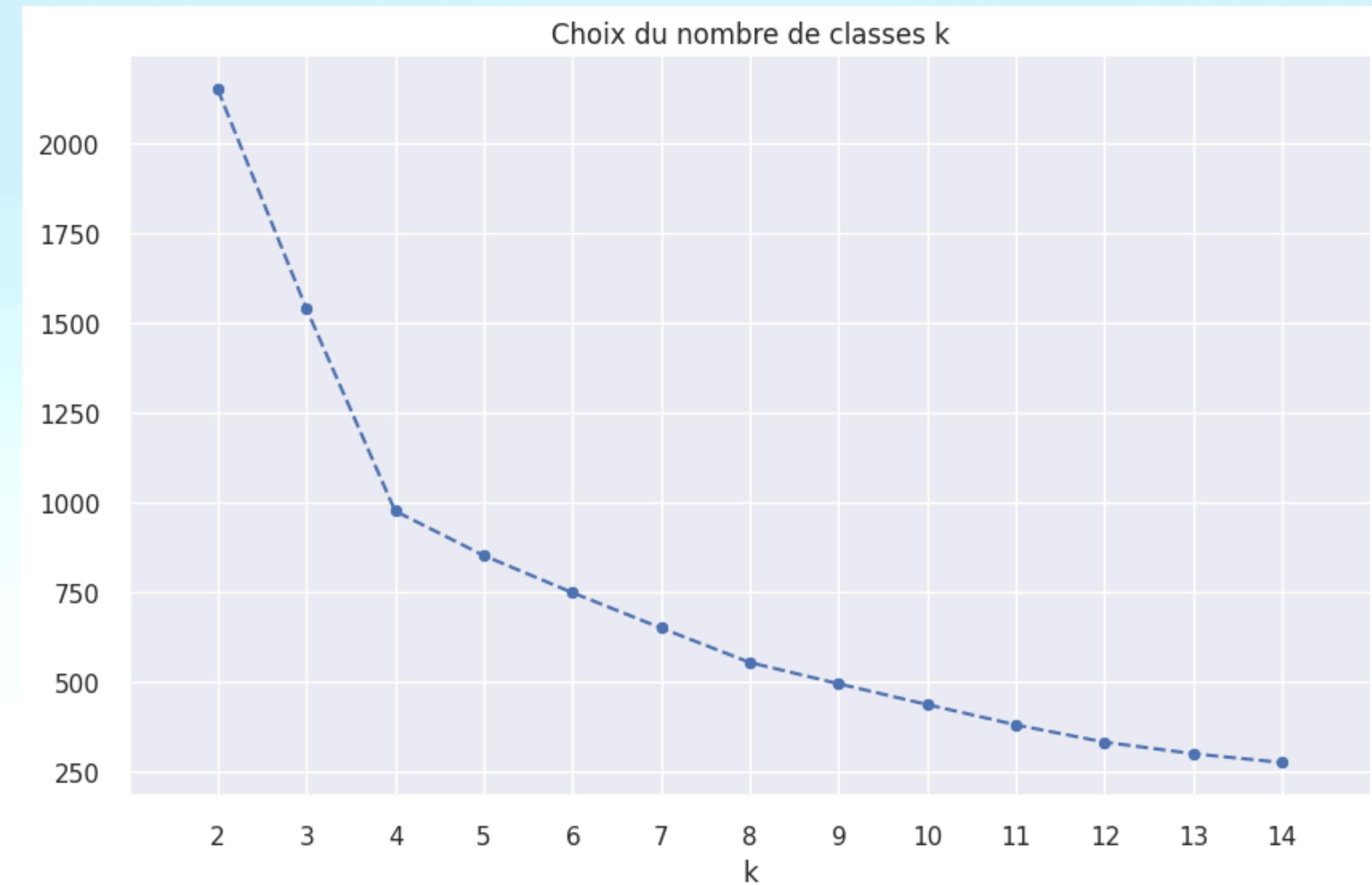
# A. Segmentation RFM

Choix du nombre de classes  $k$  (Kmeans, CAH, méthode mixte)



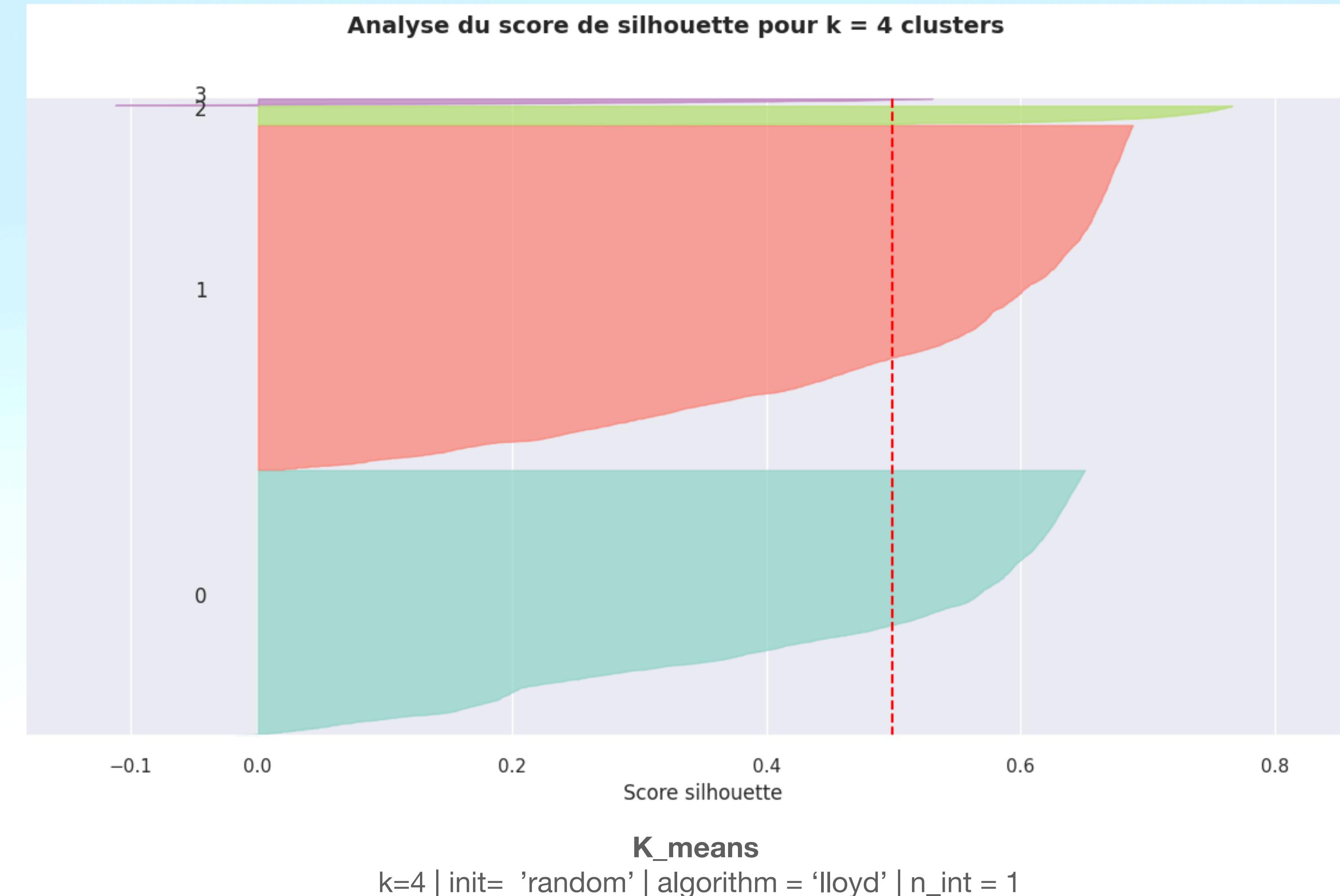
# A. Segmentation RFM

Choix du nombre de classes  $k$  (Kmeans, CAH, méthode mixte)



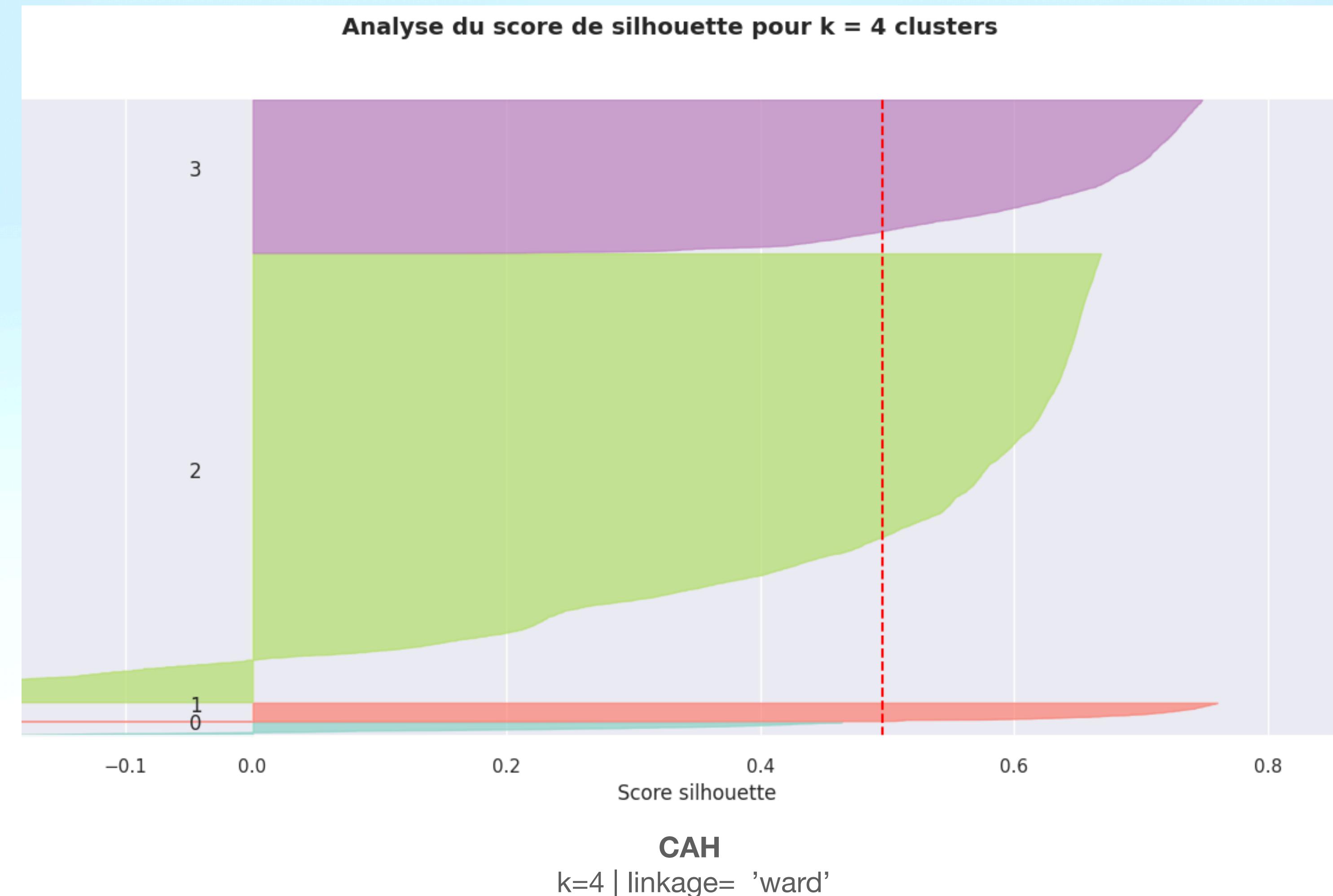
# A. Segmentation RFM

## Recherche des hyperparamètres optimaux



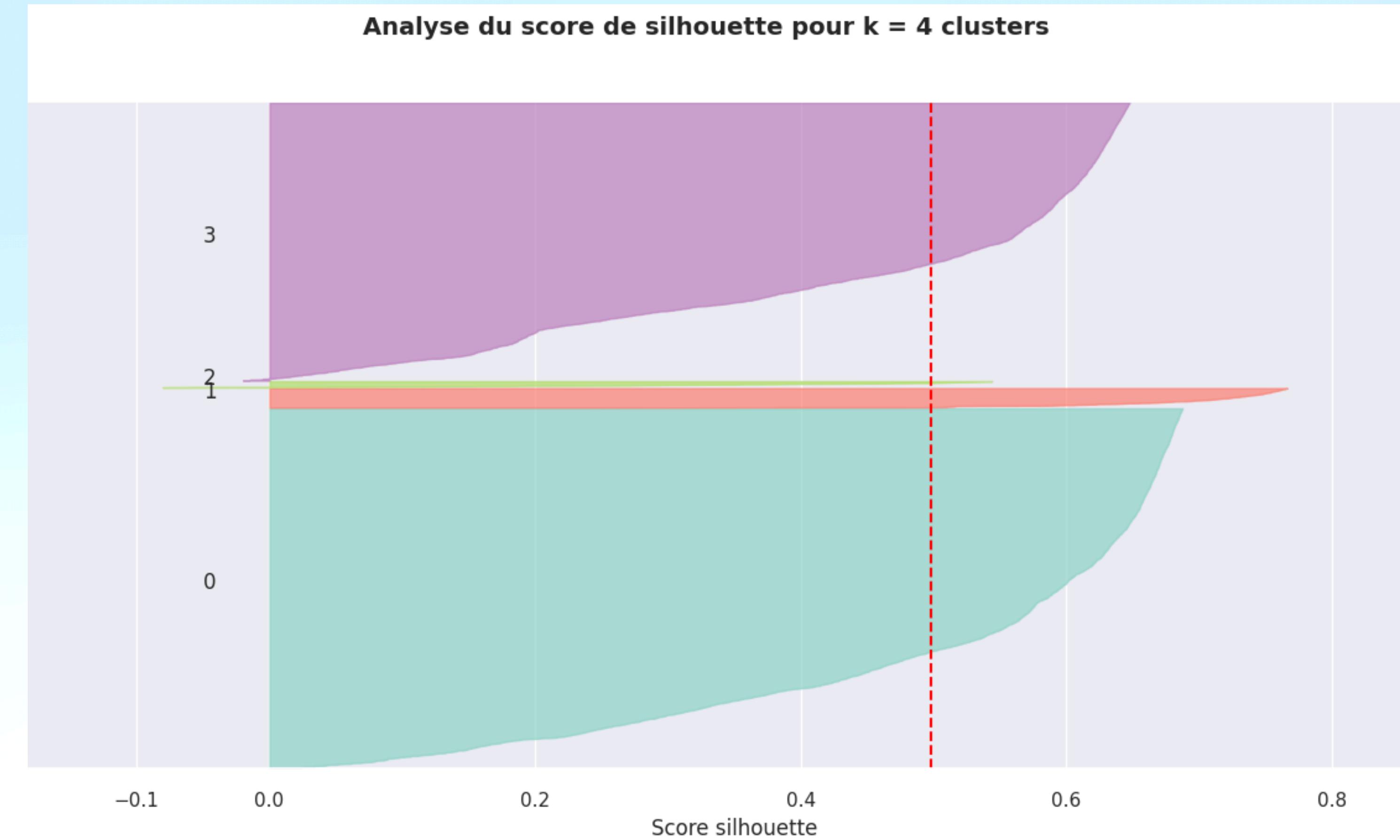
# A. Segmentation RFM

## Recherche des hyperparamètres optimaux



# A. Segmentation RFM

## Recherche des hyperparamètres optimaux



### Méthode mixte

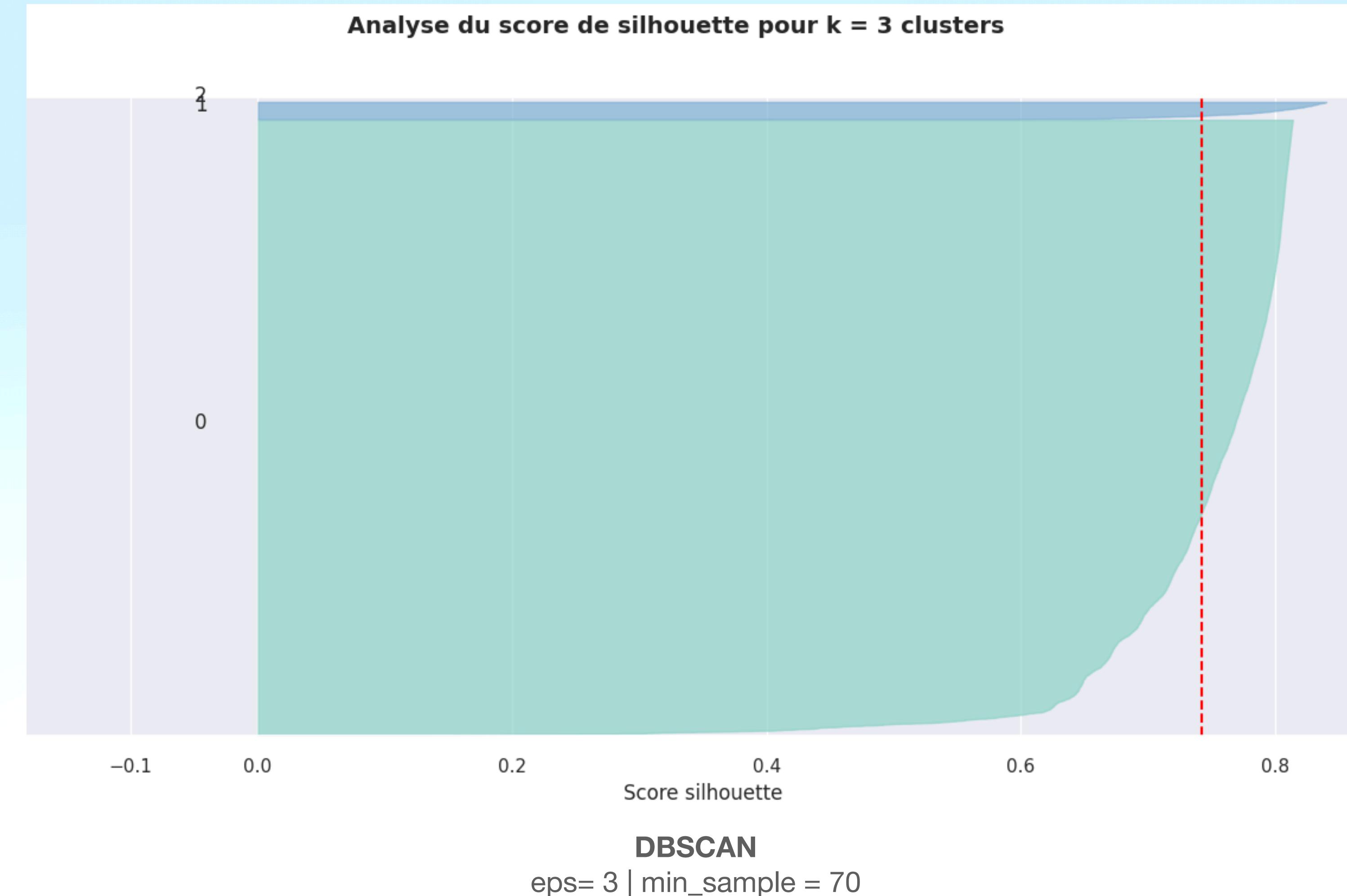
**K\_means 1 :** init= 'k\_means++' | algorithm = 'lloyd' | n\_int = 1, k\_init = 1000,

**CAH :** k = 4, linkage = 'ward'

**K\_means 2 :** init= les centroïdes issus de CAH | algorithm = 'lloyd' | n\_int = 1, k = 4

# A. Segmentation RFM

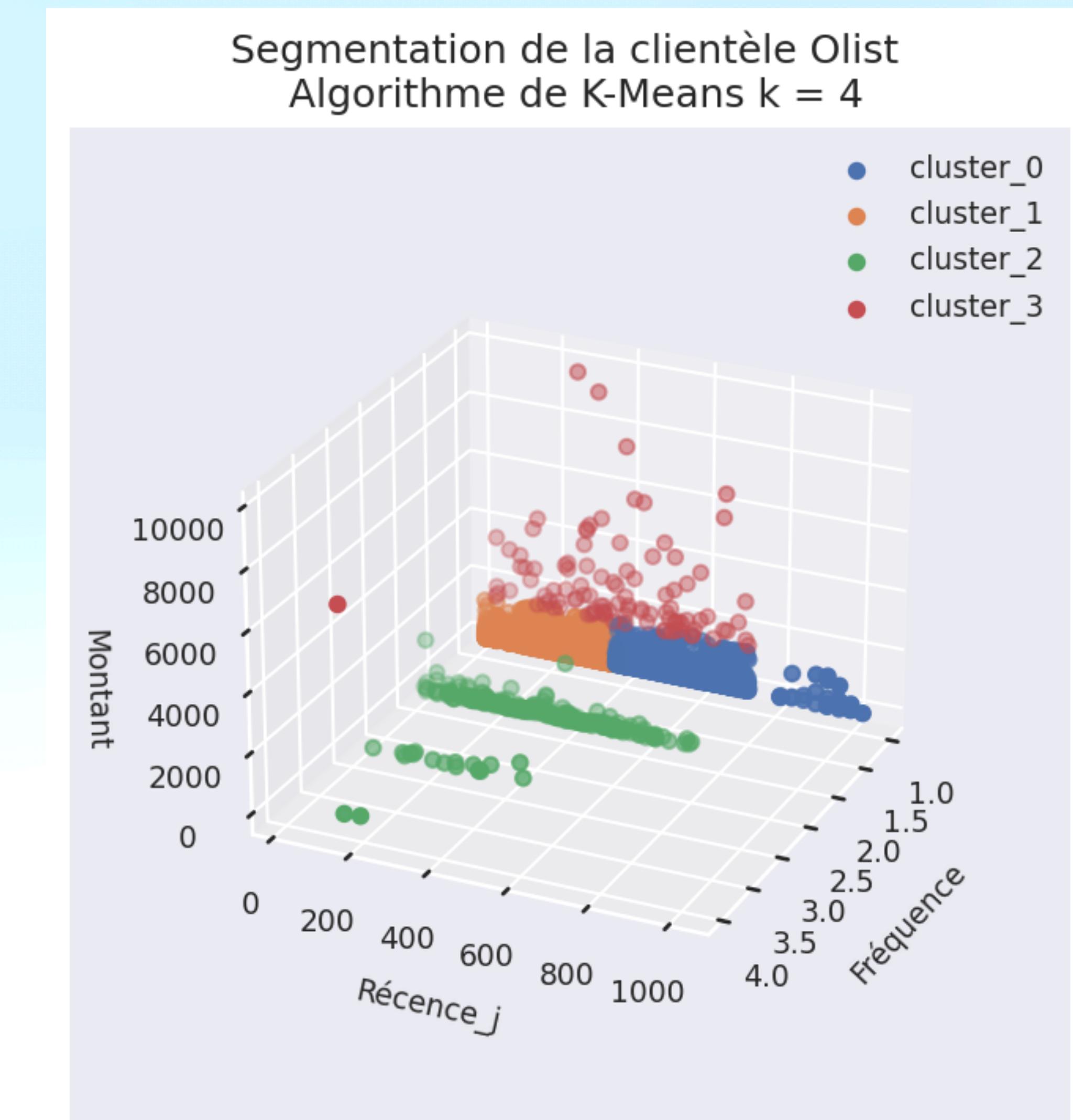
## Recherche des hyperparamètres optimaux



# A. Segmentation RFM

## Interprétation des résultats

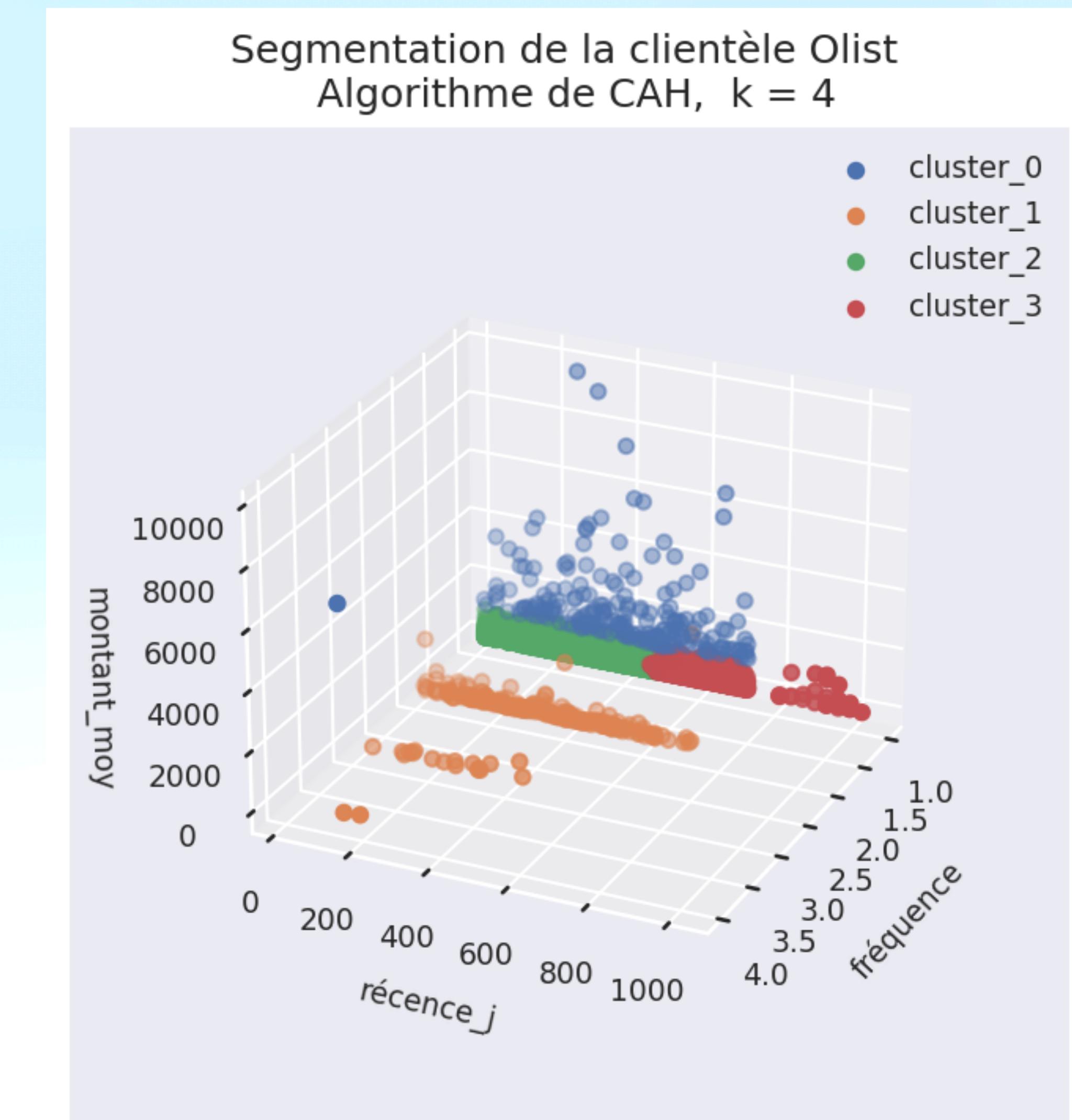
- La courbe d'inertie de l'algorithme mixte m'a aidé à choisir le bon nombre de classes,  $k = 4$
- Les trois modèles, K\_means, CAH et la méthode mixte offre une partition cohérente.  
**On distingue : les clients fréquents, les clients récents, les clients anciens, les dépenses élevées.**
- La différence majeur entre K\_means et CAH est dans la frontière des « clients récents ».
- L'algorithme DBSCAN distingue 2 clusters et un cluster d'outliers.



# A. Segmentation RFM

## Interprétation des résultats

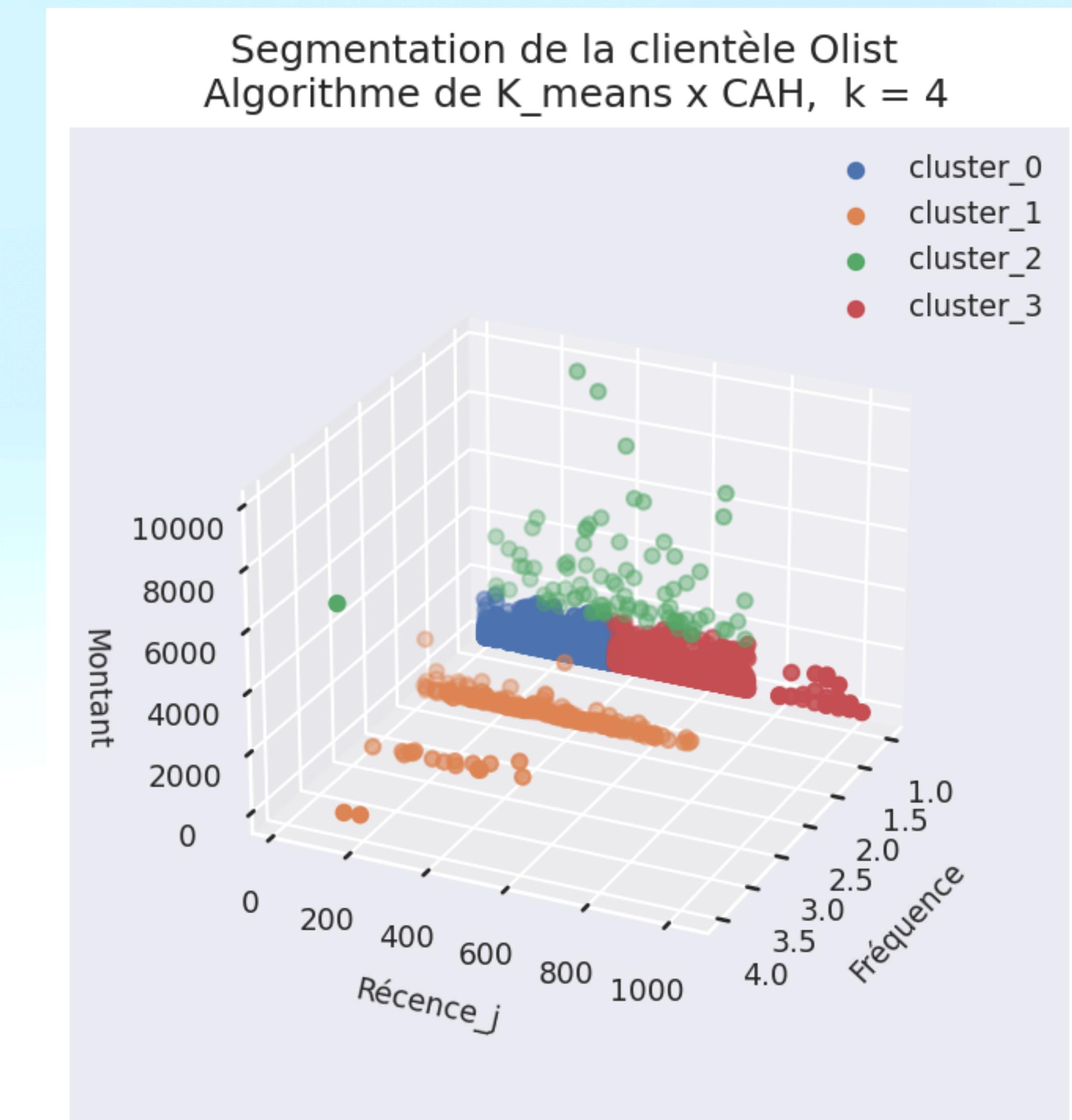
- La courbe d'inertie de l'algorithme mixte m'a aidé à choisir le bon nombre de classes,  $k = 4$
- Les trois modèles, K\_means, CAH et la méthode mixte offre une partition cohérente.  
**On distingue : les clients fréquents, les clients récents, les clients anciens, les dépenses élevées.**
- La différence majeur entre K\_means et CAH est dans la frontière des « clients récents ».
- L'algorithme DBSCAN distingue 2 clusters et un cluster d'outliers.



# A. Segmentation RFM

## Interprétation des résultats

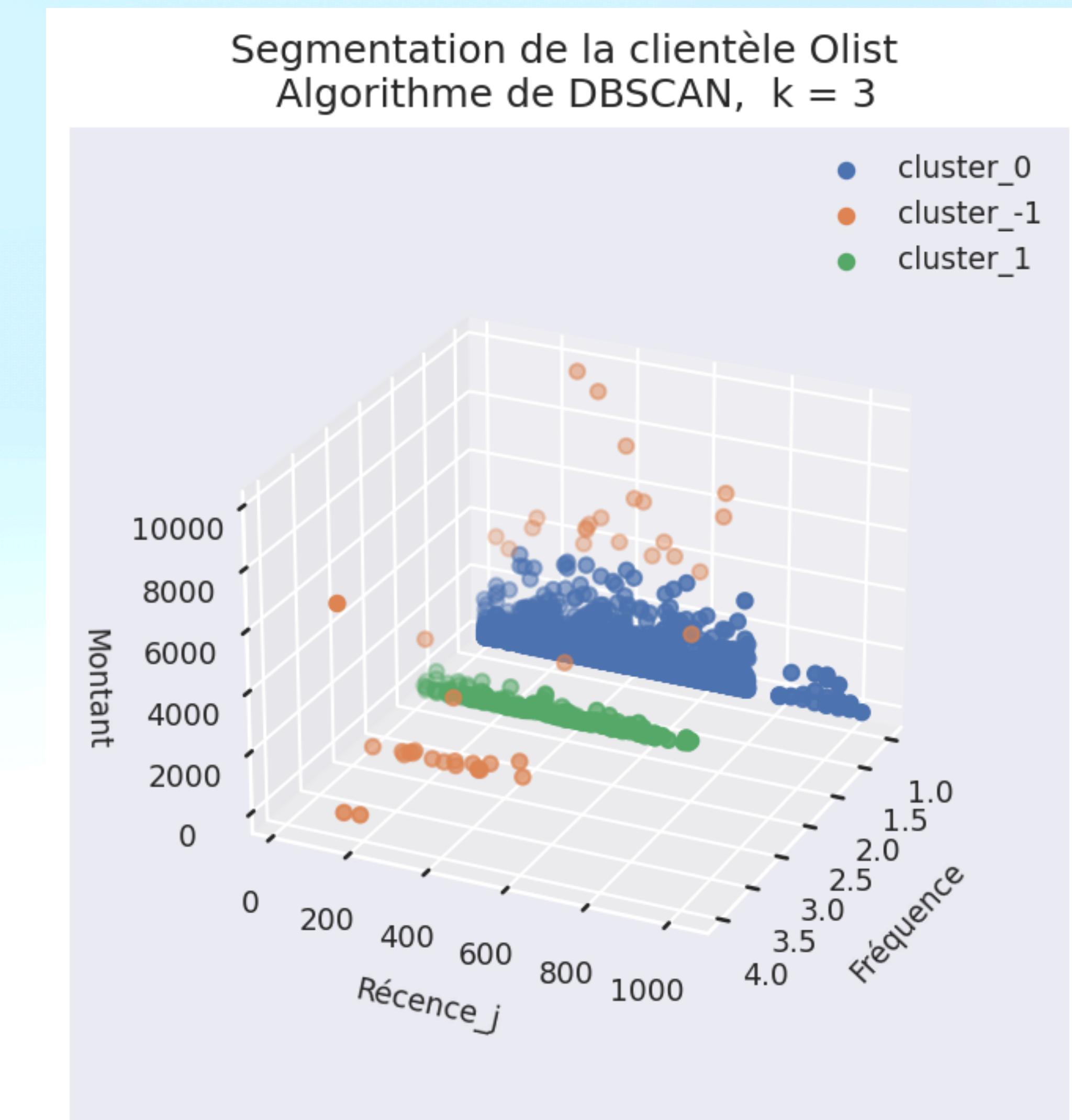
- La courbe d'inertie de l'algorithme mixte m'a aidé à choisir le bon nombre de classes,  $k = 4$
- Les trois modèles, K\_means, CAH et la méthode mixte offre une partition cohérente.  
**On distingue : les clients fréquents, les clients récents, les clients anciens, les dépenses élevées.**
- La différence majeur entre K\_means et CAH est dans la frontière des « clients récents ».
- L'algorithme DBSCAN distingue 2 clusters et un cluster d'outliers.



# A. Segmentation RFM

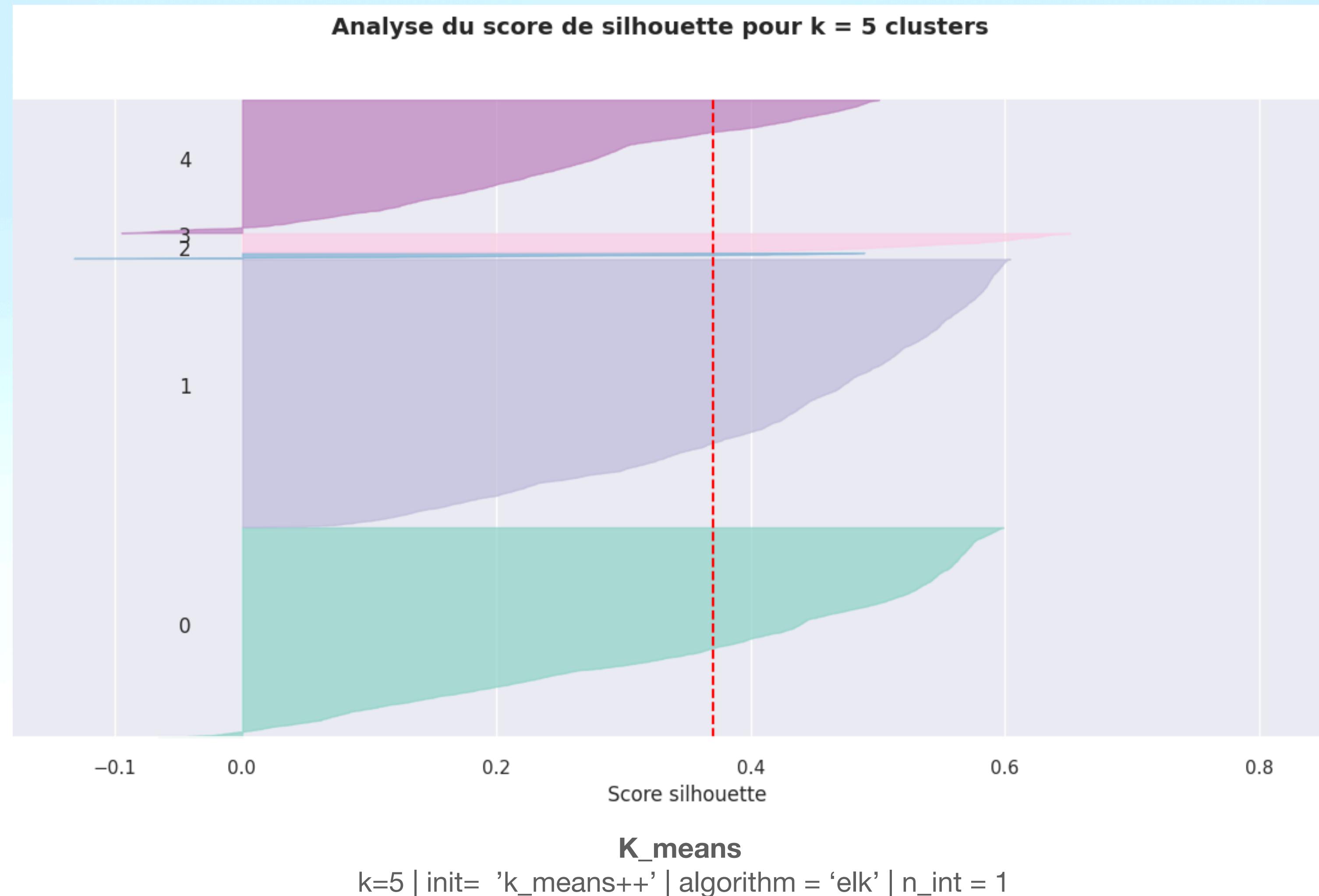
## Interprétation des résultats

- La courbe d'inertie de l'algorithme mixte m'a aidé à choisir le bon nombre de classes,  $k = 4$
- Les trois modèles, K\_means, CAH et la méthode mixte offre une partition cohérente.  
**On distingue : les clients fréquents, les clients récents, les clients anciens, les dépenses élevées.**
- La différence majeur entre K\_means et CAH est dans la frontière des « clients récents ».
- L'algorithme DBSCAN distingue 2 clusters et un cluster d'outliers.



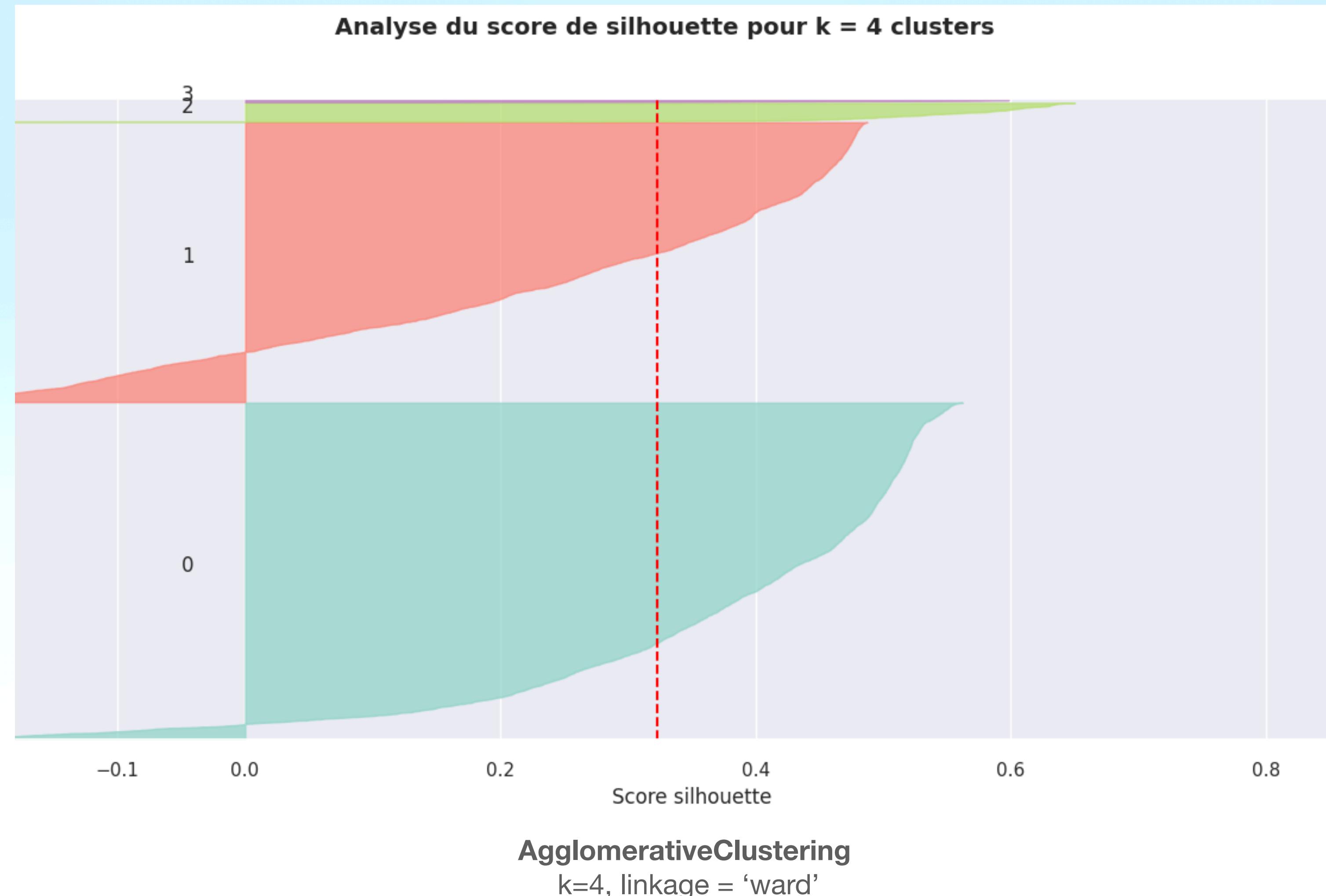
# B. Segmentation RFM + profil économique

## Recherche des hyperparamètres optimaux



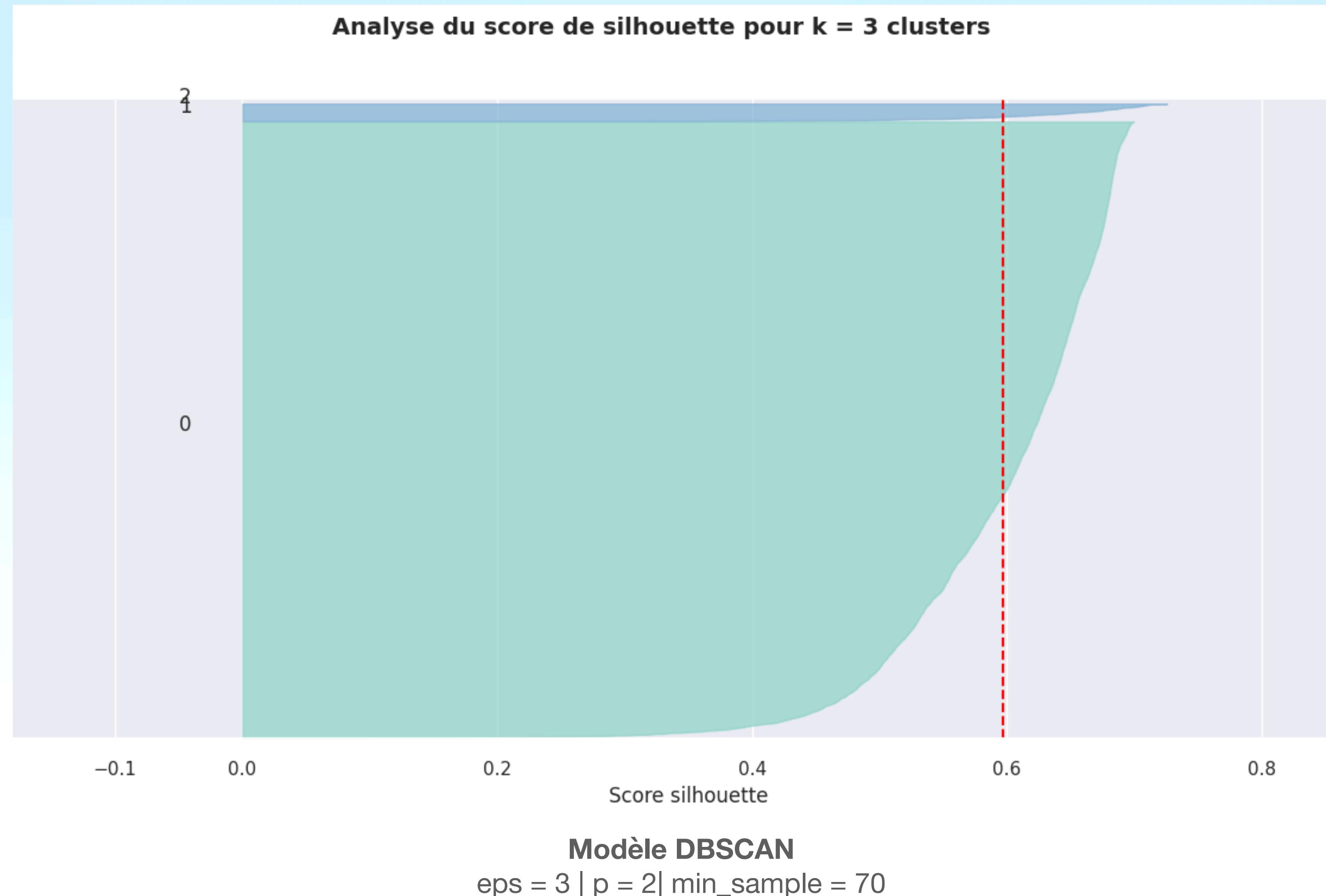
# B. Segmentation RFM + profil économique

## Recherche des hyperparamètres optimaux



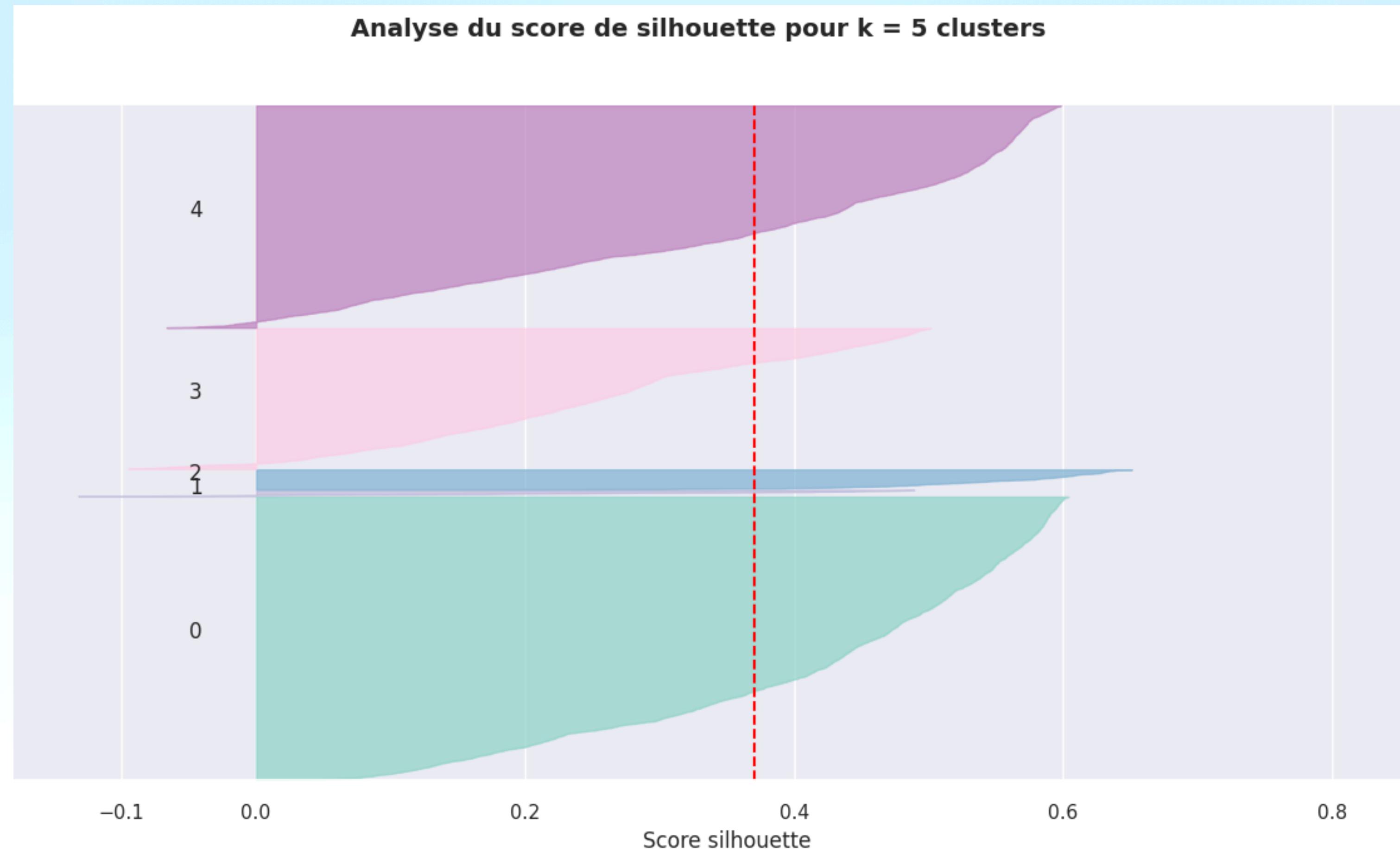
# B. Segmentation RFM + profil économique

## Recherche des hyperparamètres optimaux



# B. Segmentation RFM + profil économique

## Recherche des hyperparamètres optimaux



### Méthode mixte

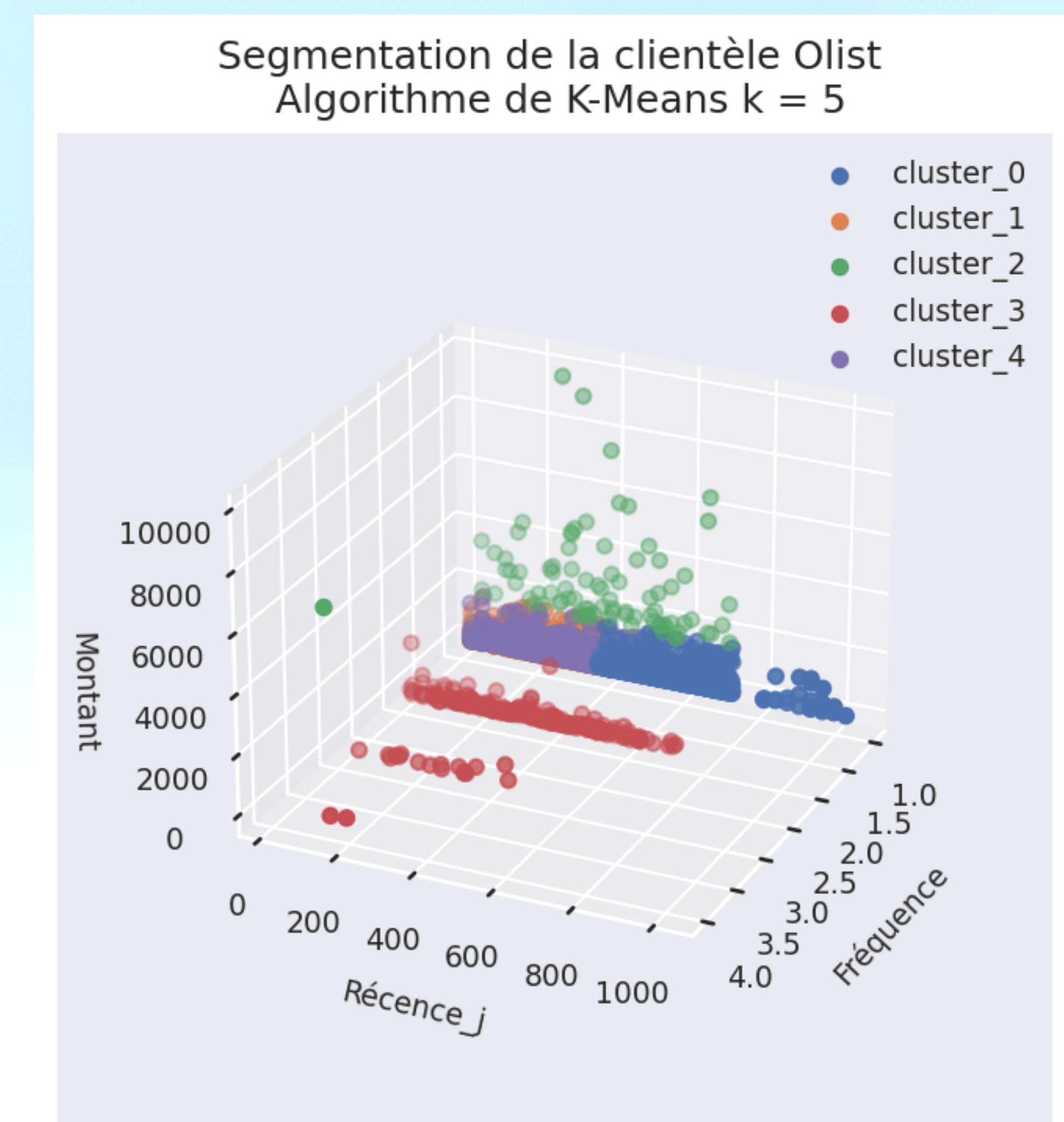
**K\_means 1 :** init= 'k\_means++' | algorithm = 'lloyd' | n\_int = 1, k\_init = 1000,  
**CAH :** k = 5, linkage = 'ward'

**K\_means 2 :** init= les centroïdes issus de CAH | algorithm = 'lloyd' | n\_int = 1, k = 5

# B. Segmentation RFM + profil économique

## Interprétation des résultats

- Les modèles K\_means et mixte permettent d'identifier les clients résidents dans des état ayant un faible revenu moyen. Les clients fréquents, ceux qui ont des dépenses très élevées, ceux qui ont fini de payer leur dernière commande et ceux qui ont encore des échéances.
- Le modèle CAH permet de d'identifier les clients qui ont des dépenses très élevés. Néanmoins les revenus et le nombre d'échéances influent peu sur le clustering CAH.
- L'algorithme DBSCAN distingue 2 clusters et un cluster d'outliers en fonction de la fréquence.



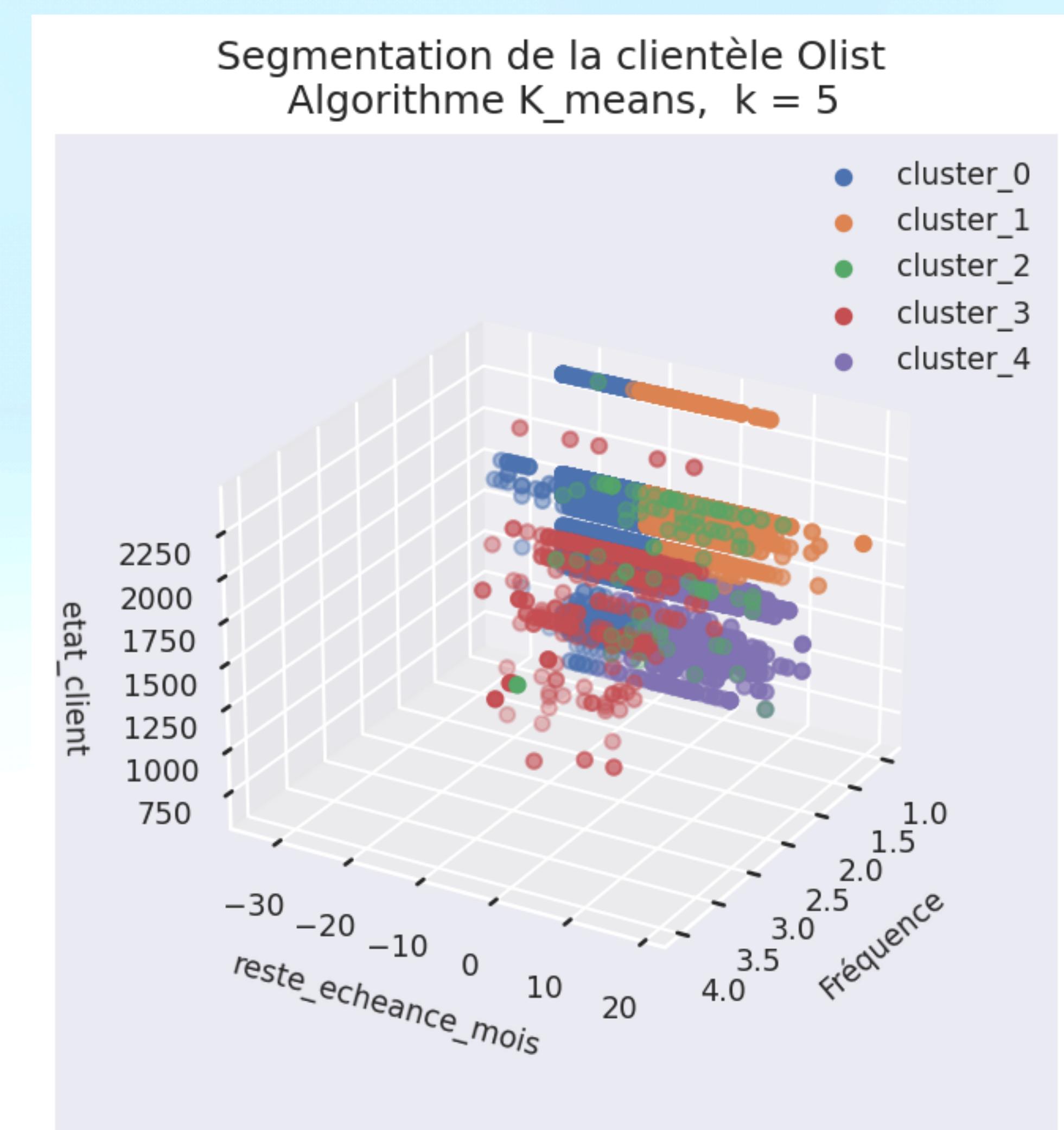
K\_means

k=5 | init= 'k\_means++' | algorithm = 'lloyd' | n\_int = 1

# B. Segmentation RFM + profil économique

## Interprétation des résultats

- Les modèles K\_means et mixte permettent d'identifier les clients résidents dans des état ayant un faible revenu moyen. Les clients fréquents, ceux qui ont des dépenses très élevées, ceux qui ont fini de payer leur dernière commande et ceux qui ont encore des échéances.
- Le modèle CAH permet de d'identifier les clients qui ont des dépenses très élevés. Néanmoins les revenus et le nombre d'échéances influent peu sur le clustering CAH.
- L'algorithme DBSCAN distingue 2 clusters et un cluster d'outliers en fonction de la fréquence.

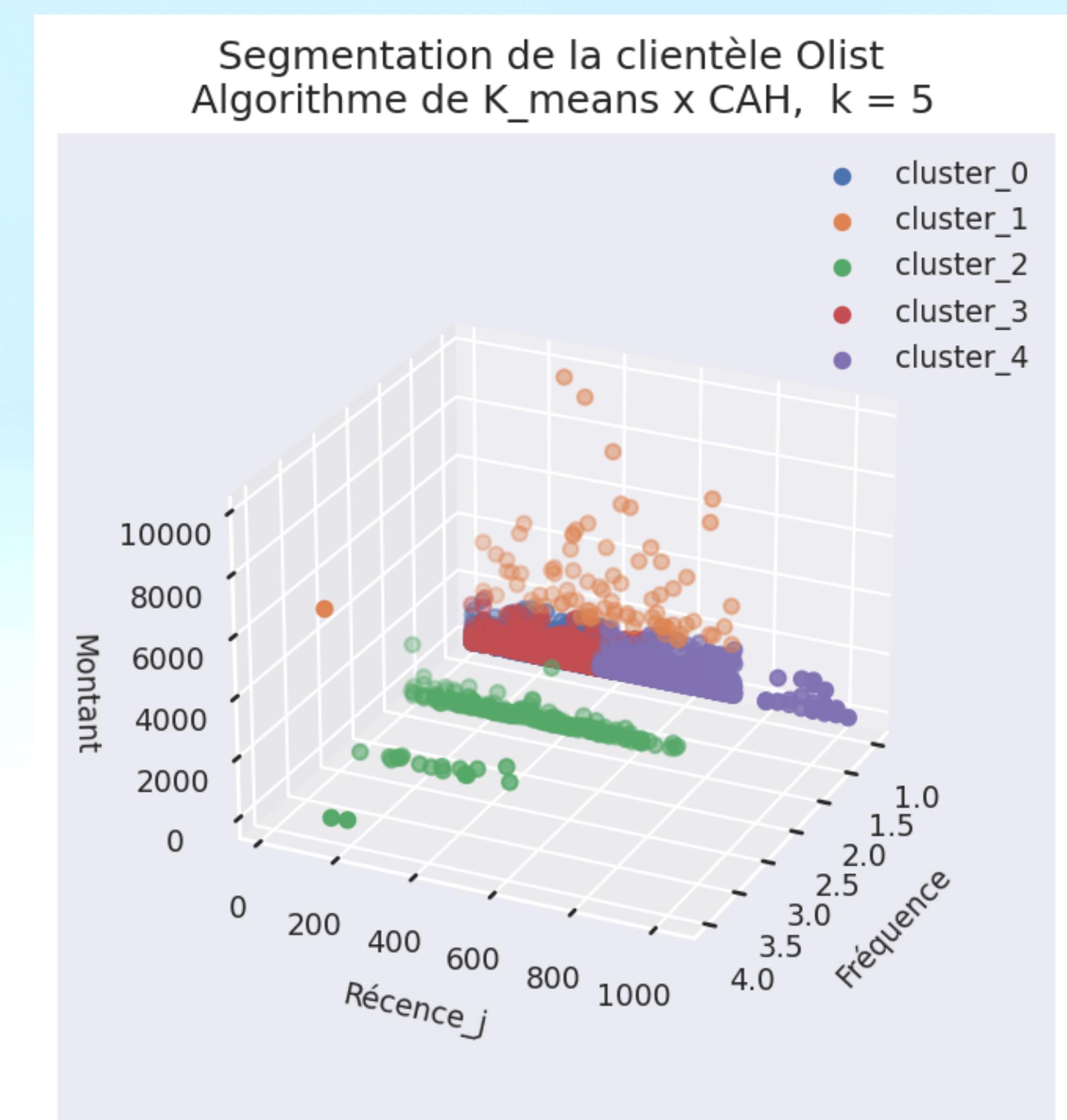


**K\_means**  
k=5 | init= 'k\_means++' | algorithm = 'lloyd' | n\_int = 1

# B. Segmentation RFM + profil économique

## Interprétation des résultats

- Les modèles K\_means et mixte permettent d'identifier les clients résidents dans des état ayant un faible revenu moyen. Les clients fréquents, ceux qui ont des dépenses très élevées, ceux qui ont fini de payer leur dernière commande et ceux qui ont encore des échéances.
- Le modèle CAH permet de d'identifier les clients qui ont des dépenses très élevés. Néanmoins les revenus et le nombre d'échéances influent peu sur le clustering CAH.
- L'algorithme DBSCAN distingue 2 clusters et un cluster d'outliers en fonction de la fréquence.

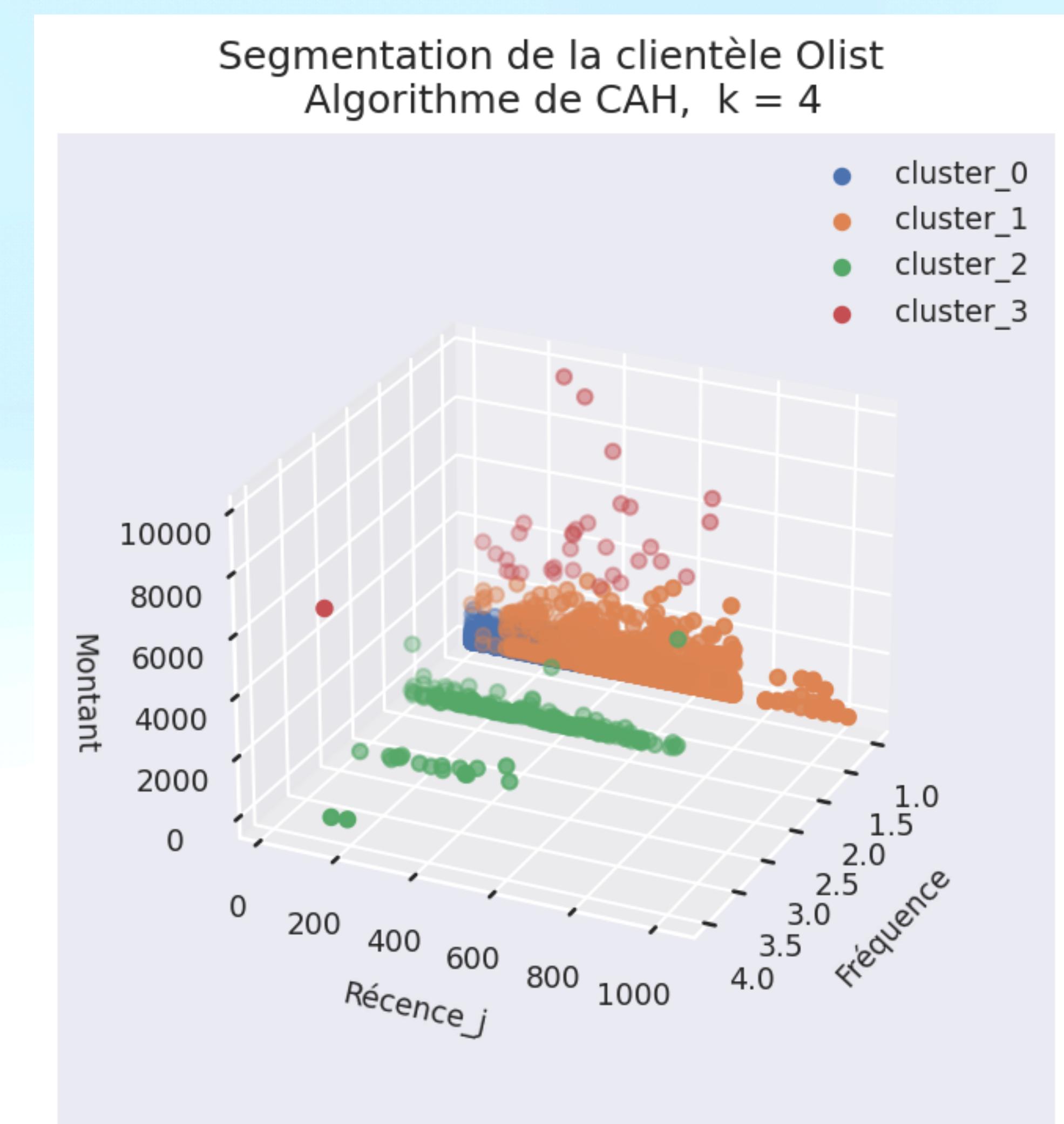


**K\_means 1 :** init= 'k\_means++' | algorithm = 'lloyd' | n\_int = 1, k\_init = 1000,  
**CAH :** k = 5, linkage = 'ward'  
**K\_means 2 :** init= les centroïdes issus de CAH | algorithm = 'lloyd' | n\_int = 1,  
k = 5

# B. Segmentation RFM + profil économique

## Interprétation des résultats

- Les modèles K\_means et mixte permettent d'identifier les clients résidents dans des état ayant un faible revenu moyen. Les clients fréquents, ceux qui ont des dépenses très élevées, ceux qui ont fini de payer leur dernière commande et ceux qui ont encore des échéances.
- Le modèle CAH permet de d'identifier les clients qui ont des dépenses très élevés. Néanmoins les revenus et le nombre d'échéances influent peu sur le clustering CAH.
- L'algorithme DBSCAN distingue 2 clusters et un cluster d'outliers en fonction de la fréquence.

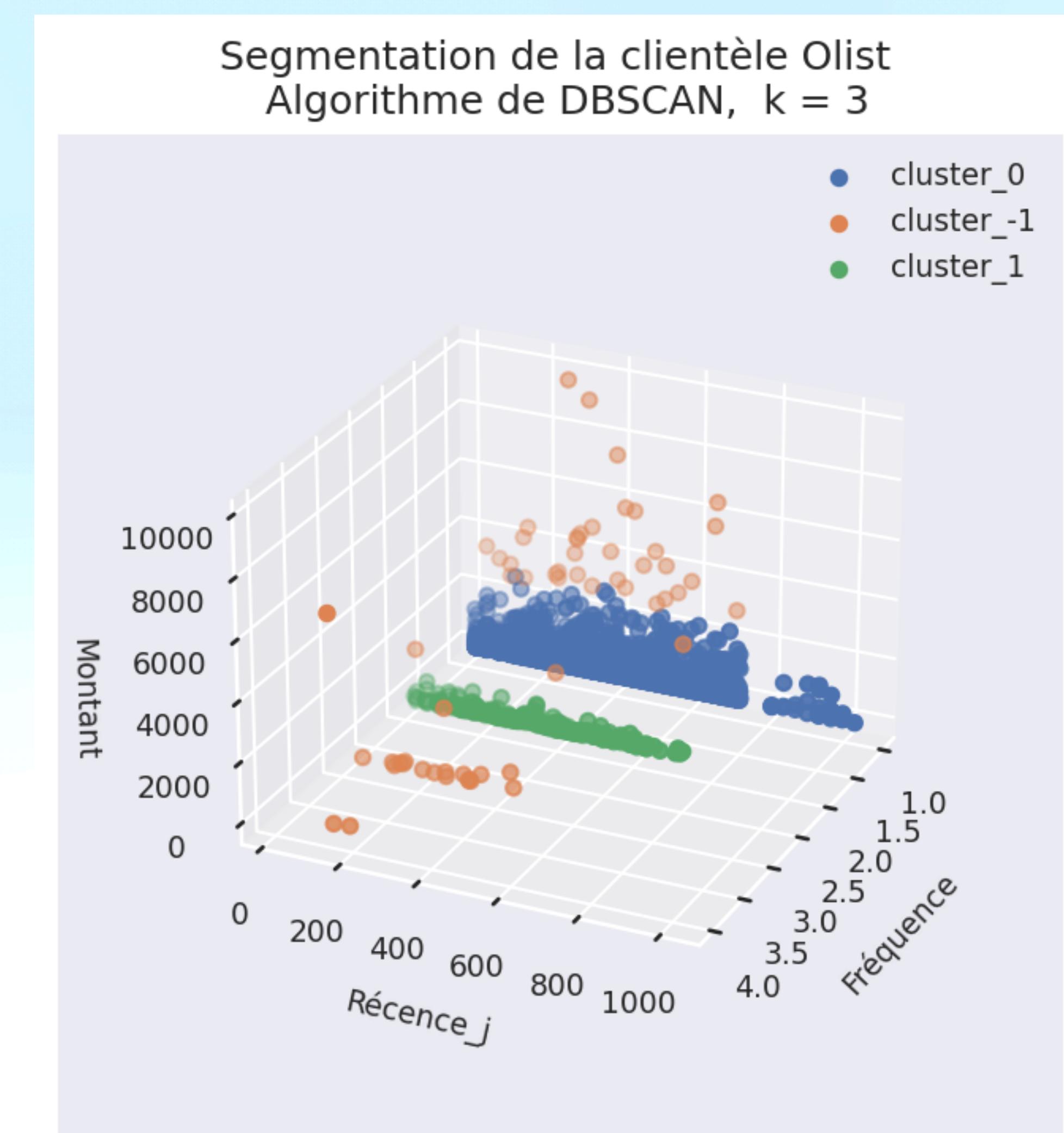


AgglomerativeClustering  
k=4, linkage = 'ward'

# B. Segmentation RFM + profil économique

## Interprétation des résultats

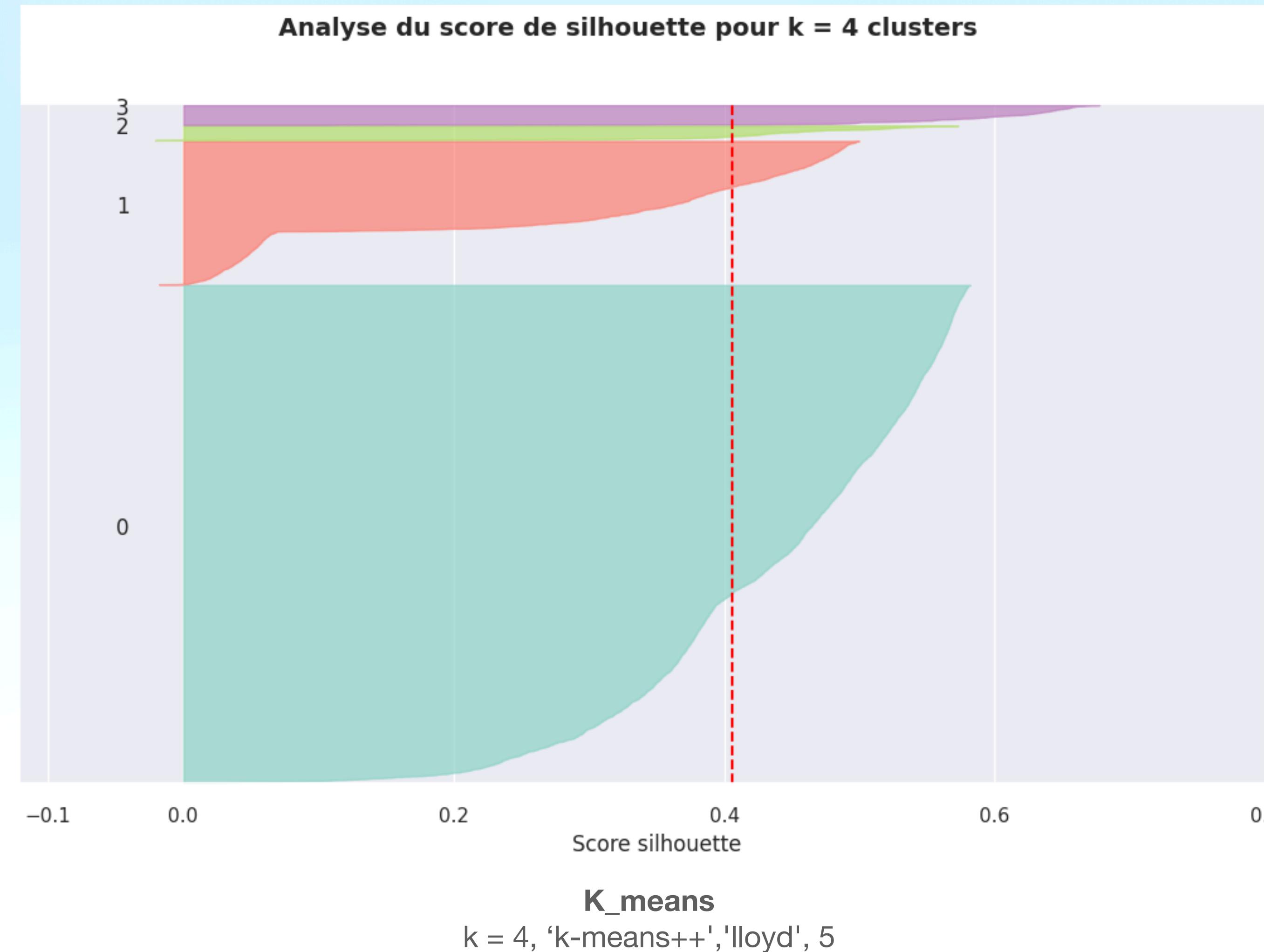
- Les modèles K\_means et mixte permettent d'identifier les clients résidents dans des état ayant un faible revenu moyen. Les clients fréquents, ceux qui ont des dépenses très élevées, ceux qui ont fini de payer leur dernière commande et ceux qui ont encore des échéances.
- Le modèle CAH permet de d'identifier les clients qui ont des dépenses très élevés. Néanmoins les revenus et le nombre d'échéances influent peu sur le clustering CAH.
- L'algorithme DBSCAN distingue 2 clusters et un cluster d'outliers en fonction de la fréquence.



Modèle DBSCAN  
eps = 3 | p = 2 | min\_sample = 70

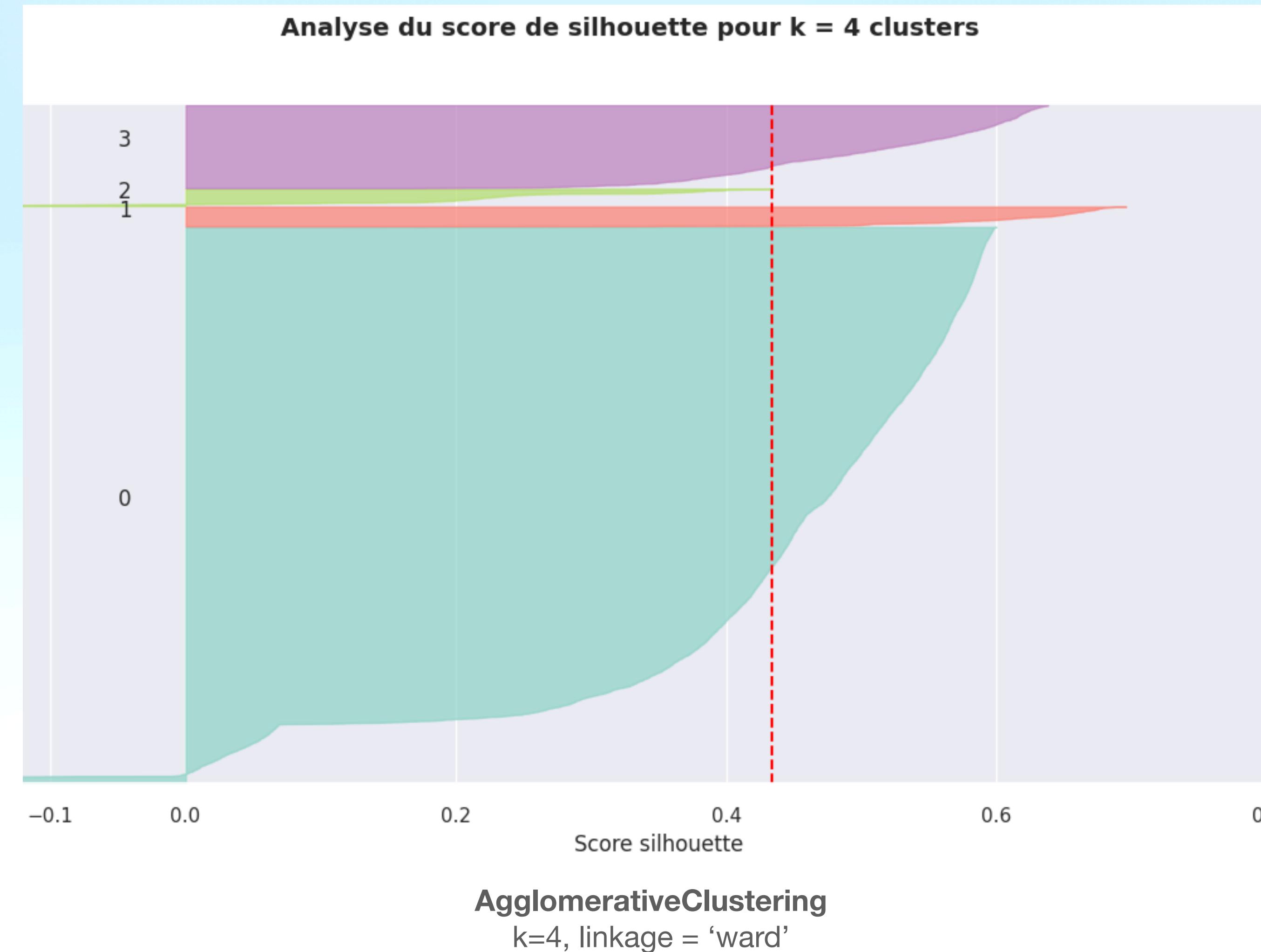
# C. Segmentation RFM + profil préférence

## Recherche des hyperparamètres optimaux



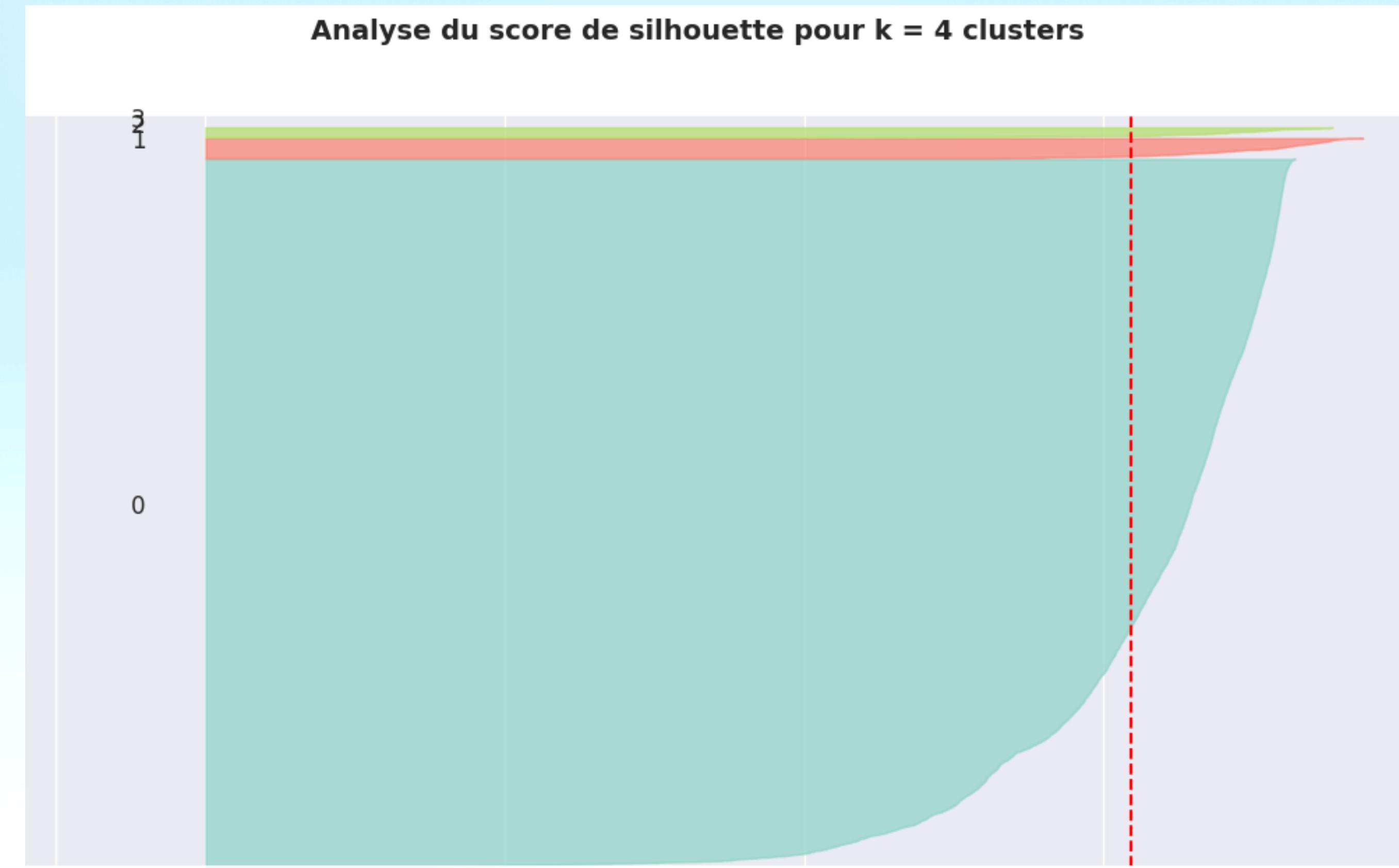
# C. Segmentation RFM + profil préférence

## Recherche des hyperparamètres optimaux



# C. Segmentation RFM + profil préférence

## Recherche des hyperparamètres optimaux

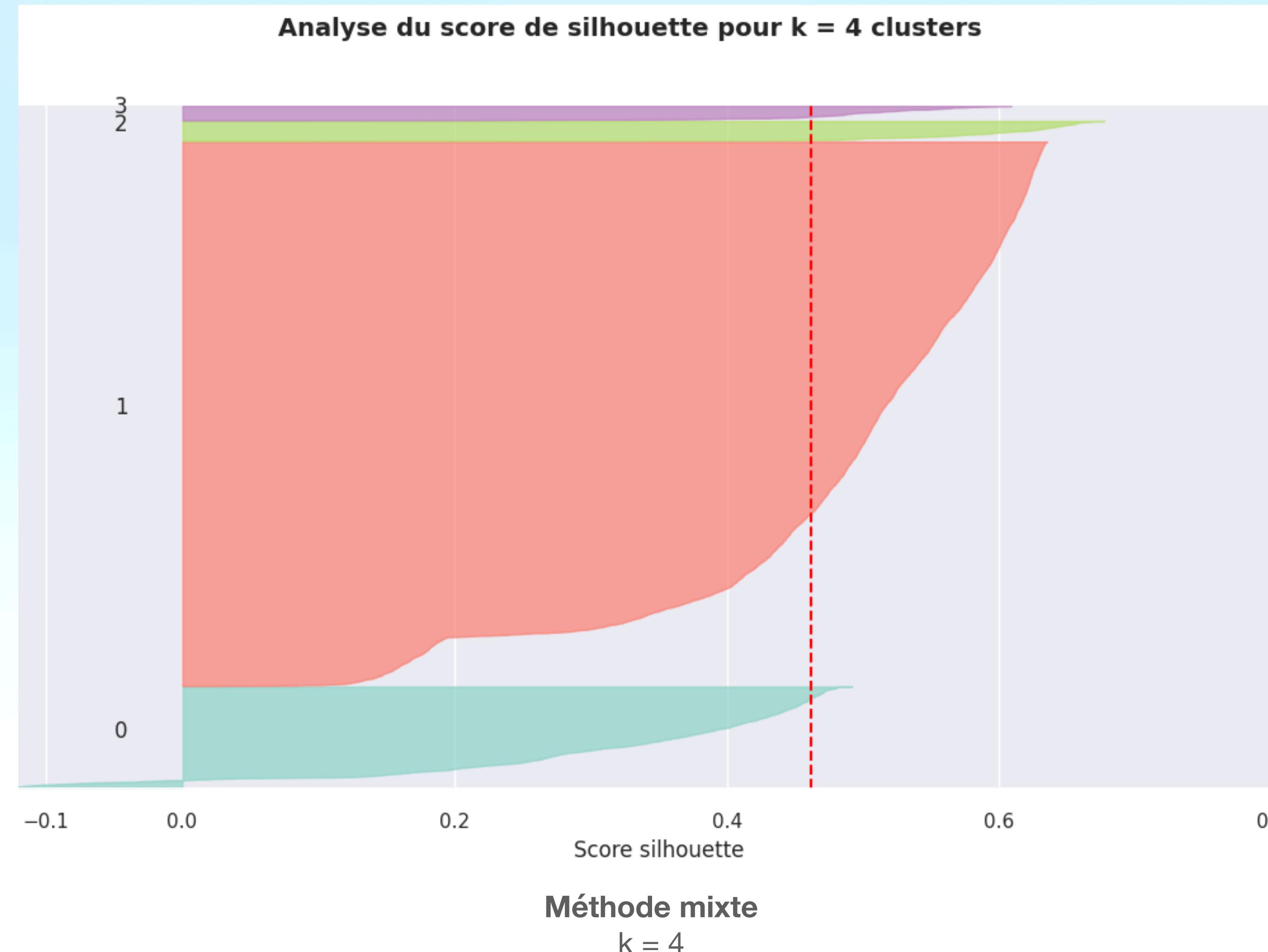


### DBSCAN

3 clusters et 1 cluster d'outliers  
eps = 2.5, min\_samples = 70, p = 2

# C. Segmentation RFM + profil préférence

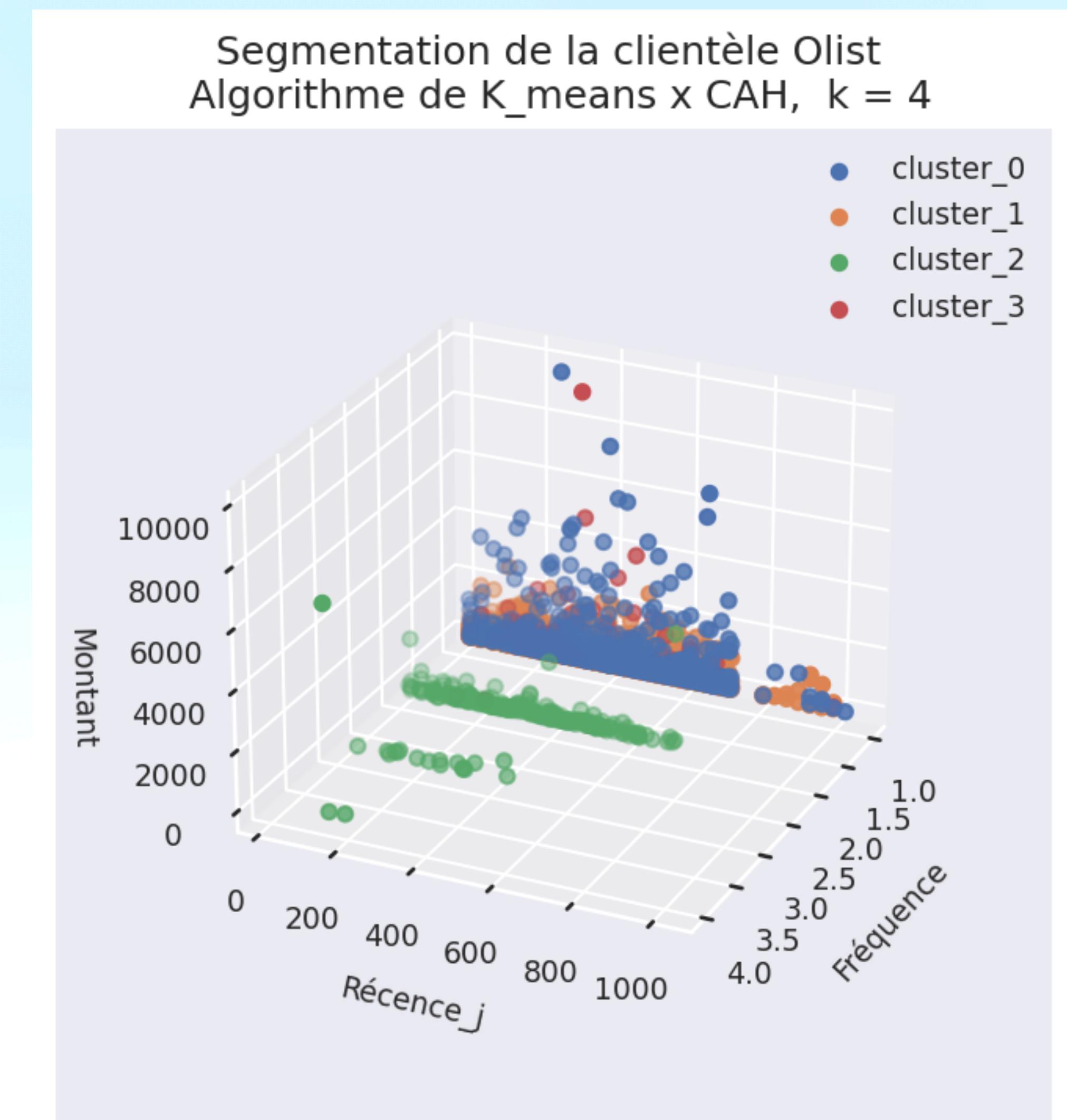
## Recherche des hyperparamètres optimaux



# C. Segmentation RFM + profil préférence

## Interprétation des résultats

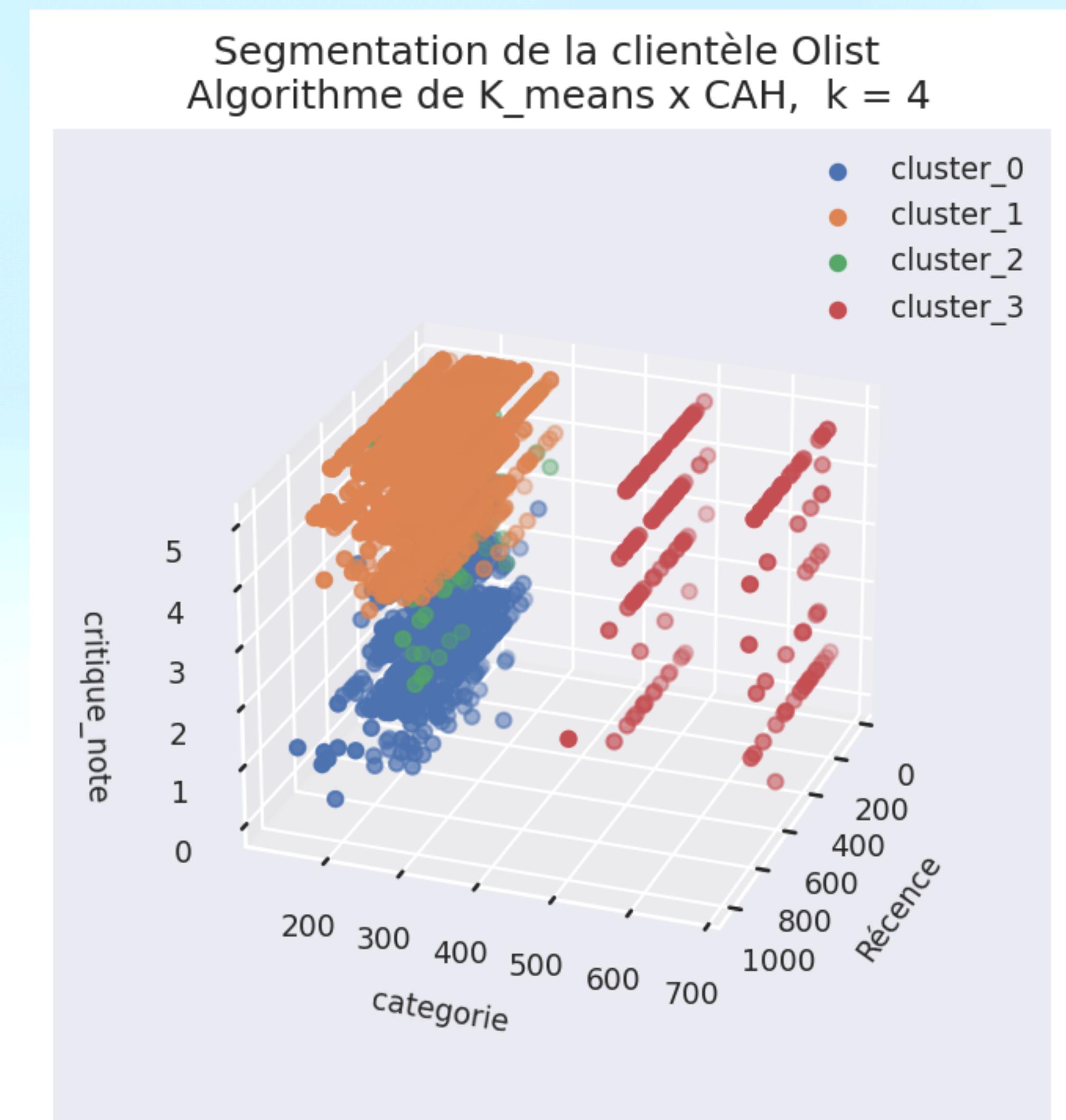
- L'algorithme mixte donnent les meilleurs résultats et distingue les clients fréquents, les clients dont les catégories préférées garantissent des dépenses élevées, ceux qui sont satisfaits et les ceux qui ne le sont pas.
- L'algorithme DBSCAN a une meilleure performance avec cette segmentation, et permet de distinguer les clients dont les catégories préférées garantissent des dépenses élevées.
- Les modèles K\_means et CAH présentent des partitions difficilement applicables aux problématiques métiers



# C. Segmentation RFM + profil préférence

## Interprétation des résultats

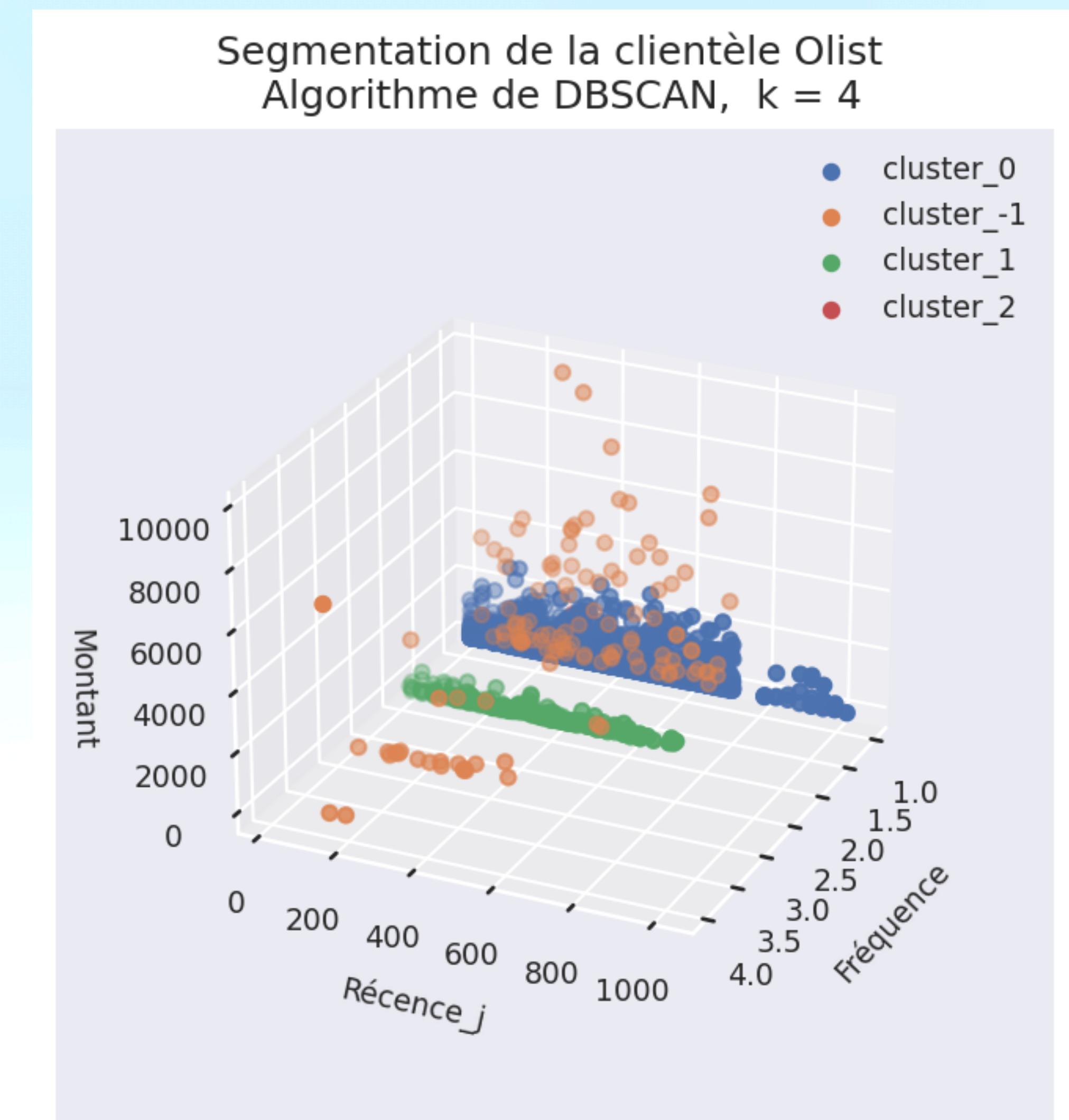
- L'algorithme mixte donnent les meilleurs résultats et distingue les clients fréquents, les clients dont les catégories préférées garantissent des dépenses élevées, ceux qui sont satisfaits et les ceux qui ne le sont pas.
- L'algorithme DBSCAN a une meilleure performance avec cette segmentation, et permet de distinguer les clients dont les catégories préférées garantissent des dépenses élevées.
- Les modèles K\_means et CAH présentent des partitions difficilement applicables aux problématiques métiers



# C. Segmentation RFM + profil préférence

## Interprétation des résultats

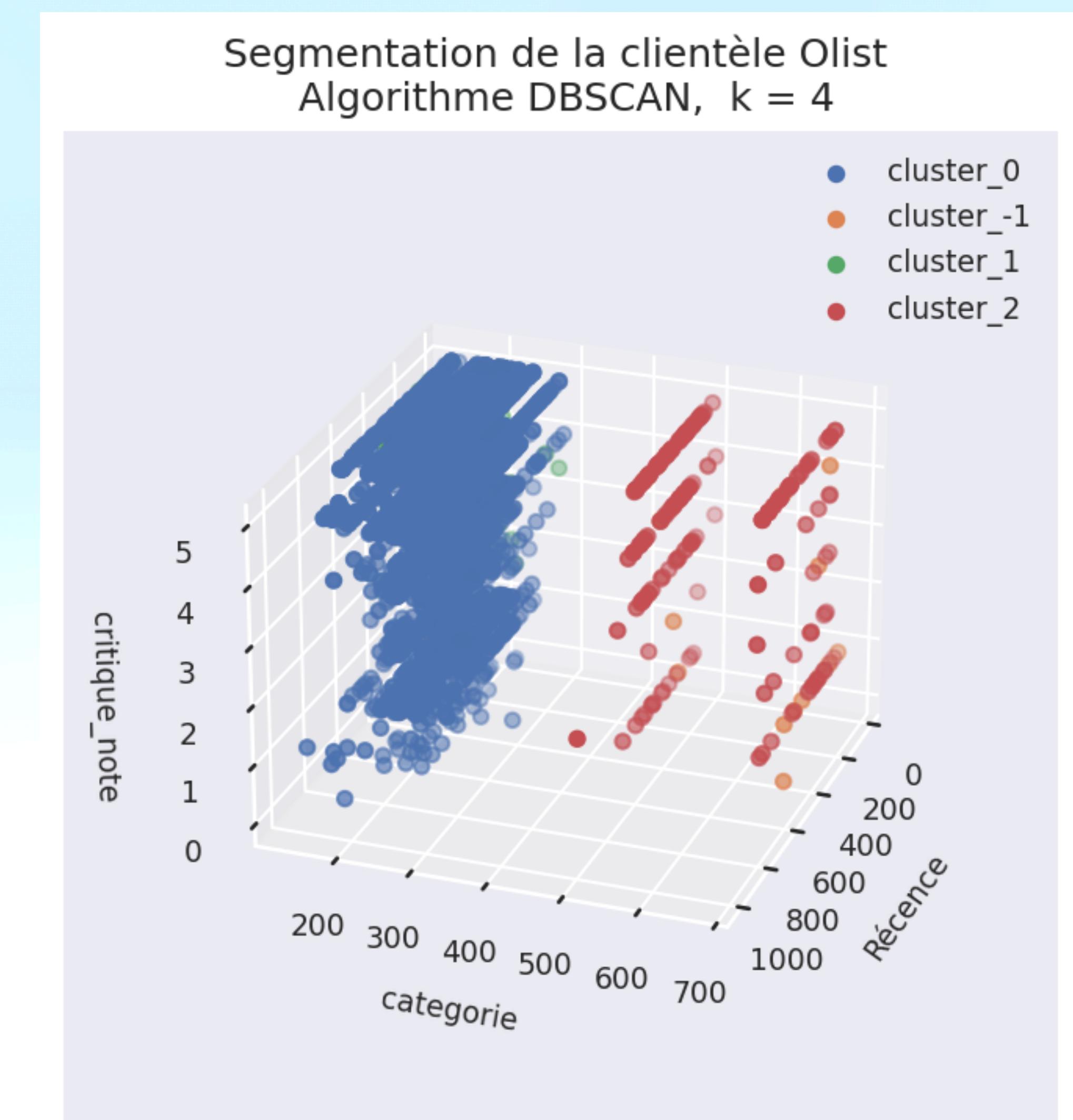
- L'algorithme mixte donnent les meilleurs résultats et distingue les clients fréquents, les clients dont les catégories préférées garantissent des dépenses élevées, ceux qui sont satisfaits et les ceux qui ne le sont pas.
- L'algorithme DBSCAN a une meilleure performance avec cette segmentation, et permet de distinguer les clients dont les catégories préférées garantissent des dépenses élevées.
- Les modèles K\_means et CAH présentent des partitions difficilement applicables aux problématiques métiers



# C. Segmentation RFM + profil préférence

## Interprétation des résultats

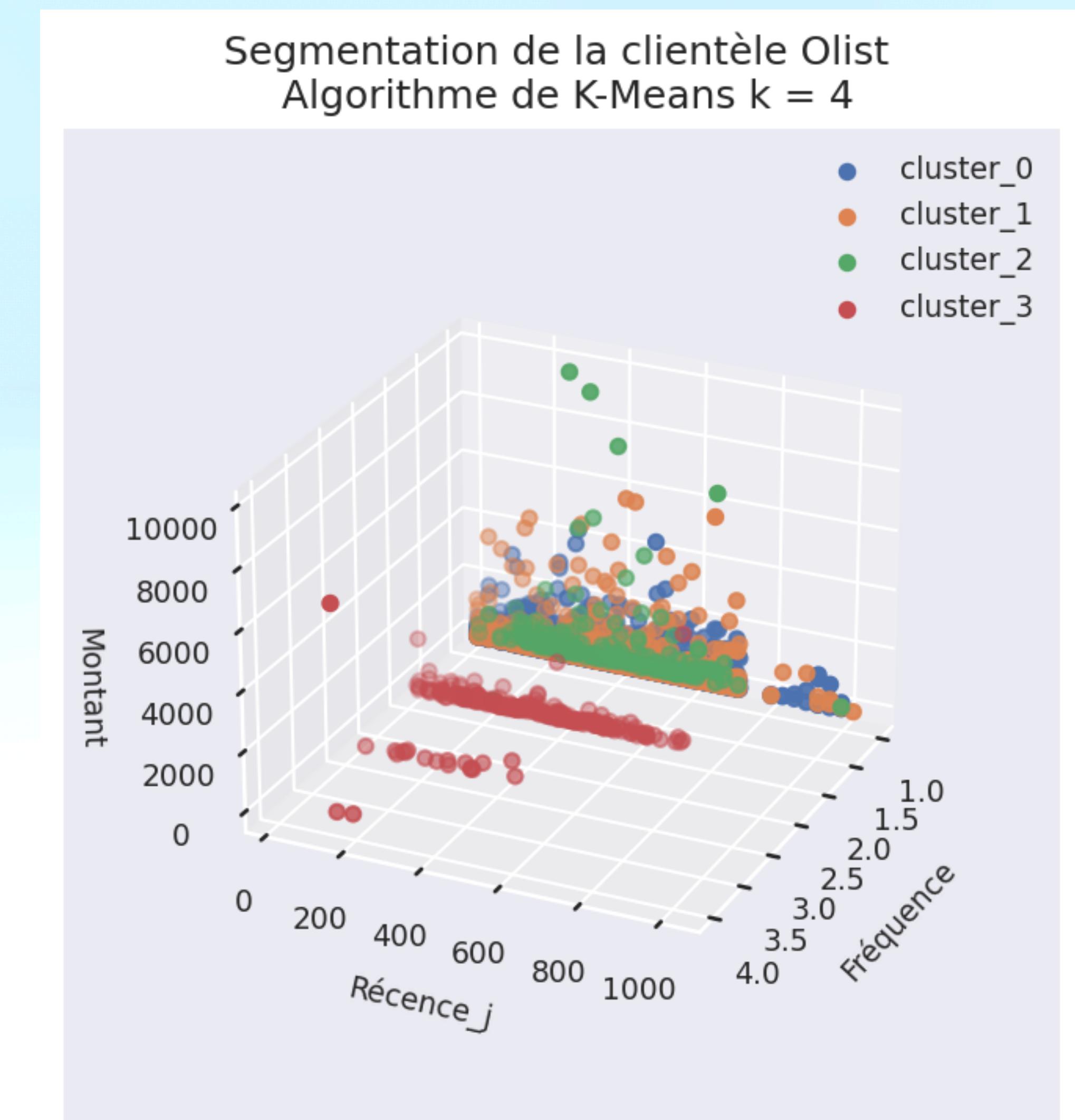
- L'algorithme mixte donnent les meilleurs résultats et distingue les clients fréquents, les clients dont les catégories préférées garantissent des dépenses élevées, ceux qui sont satisfaits et les ceux qui ne le sont pas.
- L'algorithme DBSCAN a une meilleure performance avec cette segmentation, et permet de distinguer les clients dont les catégories préférées garantissent des dépenses élevées.
- Les modèles K\_means et CAH présentent des partitions difficilement applicables aux problématiques métiers



# C. Segmentation RFM + profil préférence

## Interprétation des résultats

- L'algorithme mixte donnent les meilleurs résultats et distingue les clients fréquents, les clients dont les catégories préférées garantissent des dépenses élevées, ceux qui sont satisfaits et les ceux qui ne le sont pas.
- L'algorithme DBSCAN a une meilleure performance avec cette segmentation, et permet de distinguer les clients dont les catégories préférées garantissent des dépenses élevées.
- Les modèles K\_means et CAH présentent des partitions difficilement applicables aux problématiques métiers

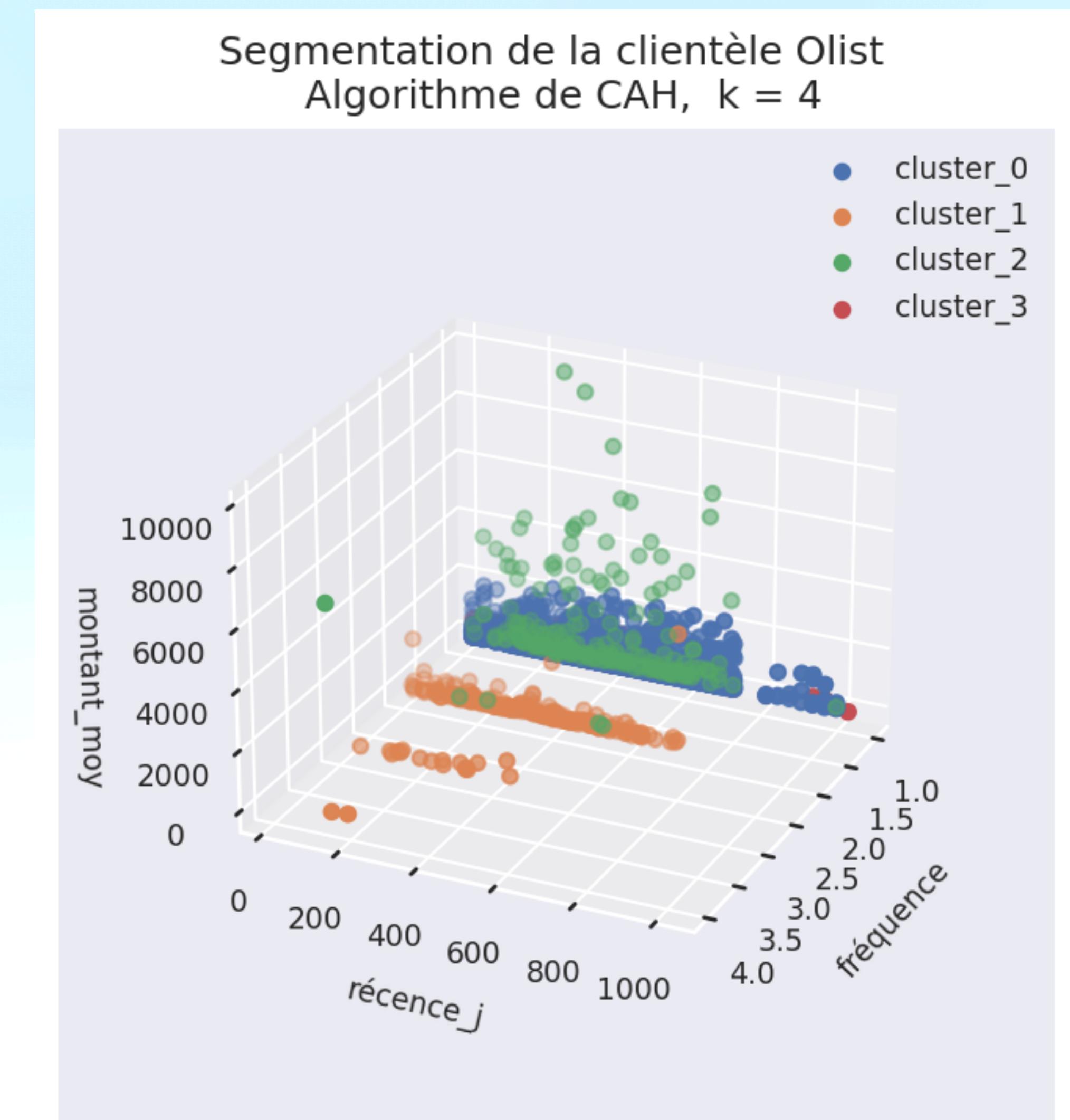


K\_means  
k = 4, 'k-means++', 'lloyd', 5

# C. Segmentation RFM + profil préférence

## Interprétation des résultats

- L'algorithme mixte donnent les meilleurs résultats et distingue les clients fréquents, les clients dont les catégories préférées garantissent des dépenses élevées, ceux qui sont satisfaits et les ceux qui ne le sont pas.
- L'algorithme DBSCAN a une meilleure performance avec cette segmentation, et permet de distinguer les clients dont les catégories préférées garantissent des dépenses élevées.
- Les modèles K\_means et CAH présentent des partitions difficilement applicables aux problématiques métiers



AgglomerativeClustering  
k=4, linkage = 'ward'

# D. Choix du modèle optimal et résultats

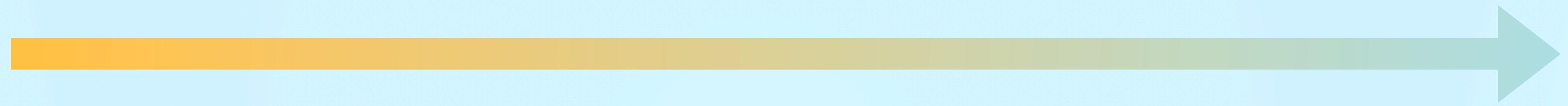
	Modèle	k	Silhouette	Temps_calcul (s)	Cohérence métier
Set 1 RFM	K_means	4	0,499	0,076	OK
	CAH	4	0,496	4,203	OK. Coûteux en calcul.
	DBSCAN	2 + outliers	0,742	2,423	NOK. Identification de deux clusters distinct selon la fréquence
	K_means x CAH	4	0,498	0,064	OK. Peu coûteux pour l'emploi d'une CAH
Set 2 RFM + profil économique	K_means	5	0,37	0,019	OK
	CAH	4	0,323	5,051	Moyenne. Coûteux en calcul.
	DBSCAN	2 + outliers	0,742	1,991	NOK
	K_means x CAH	5	0,371	0,080	OK. Peu coûteux pour l'emploi d'une CAH
Set 3 RFM + préférence	K_means	4	0,410	0,023	Moyenne.
	CAH	4	0,433	4,898	Moyenne.
	DBSCAN	3 + outliers	0,636	3,424	Moyenne.
	K_means x CAH	4	0,462	0,104	OK. Peu coûteux pour l'emploi d'une CAH

# D. Choix du modèle optimal et résultats

## Segmentation RFM classique | k = 4

Feature / Cluster	0	1	2 ★★★★★	3 ★★★★★
Proportion de la population	Environ 42 %	Environ 54 %	Environ 3%	Environ 1%
Fréquence	Une commande entre 2016 et 2018	Une commande entre 2016 et 2018	Plusieurs commandes entre 2016 et 2018	Une commande entre 2016 et 2018
Récence	Récence médiane = 18 mois	Récence médiane = 7 mois	Récence médiane = 10 mois	Récence médiane = 12 mois
Montant moyen par commande	Montant moyen médian = 105 R\$	Montant moyen médian = 111 R\$	Montant moyen médian = 126 R\$	Montant moyen médian = 2160 R\$

# 4. Maintenance



## PHASE 1

Initialisation du modèle au temps  $T_0$  = instant de la commande la plus récente

Récupération des coordonnées des centroïdes

## PHASE 2

Filtre de la table *client* en utilisant la ‘récence\_j’, bon en arrière de  $k$  jours

Prédiction des cluster de l'instant  $T_k$  via le modèle initial

Prédiction des cluster de l'instant  $T_k$  via le modèle entraîné à l'instant  $T_k$  en initialisant les centroïdes

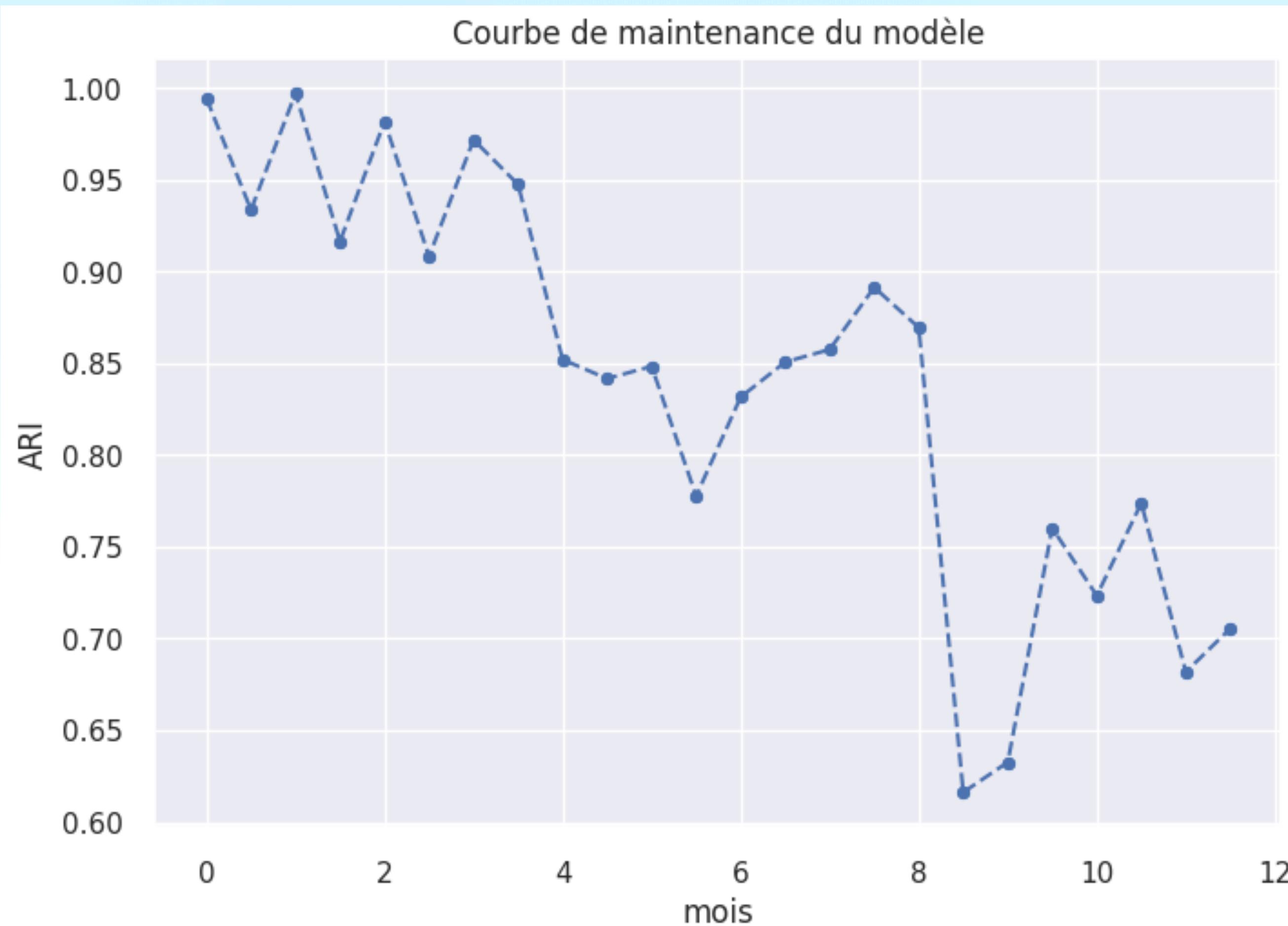
## PHASE 3

Comparaison des clusters à l'aide du score Adjusted Rand



Pour chaque instant  $T_k$

# 4. Maintenance

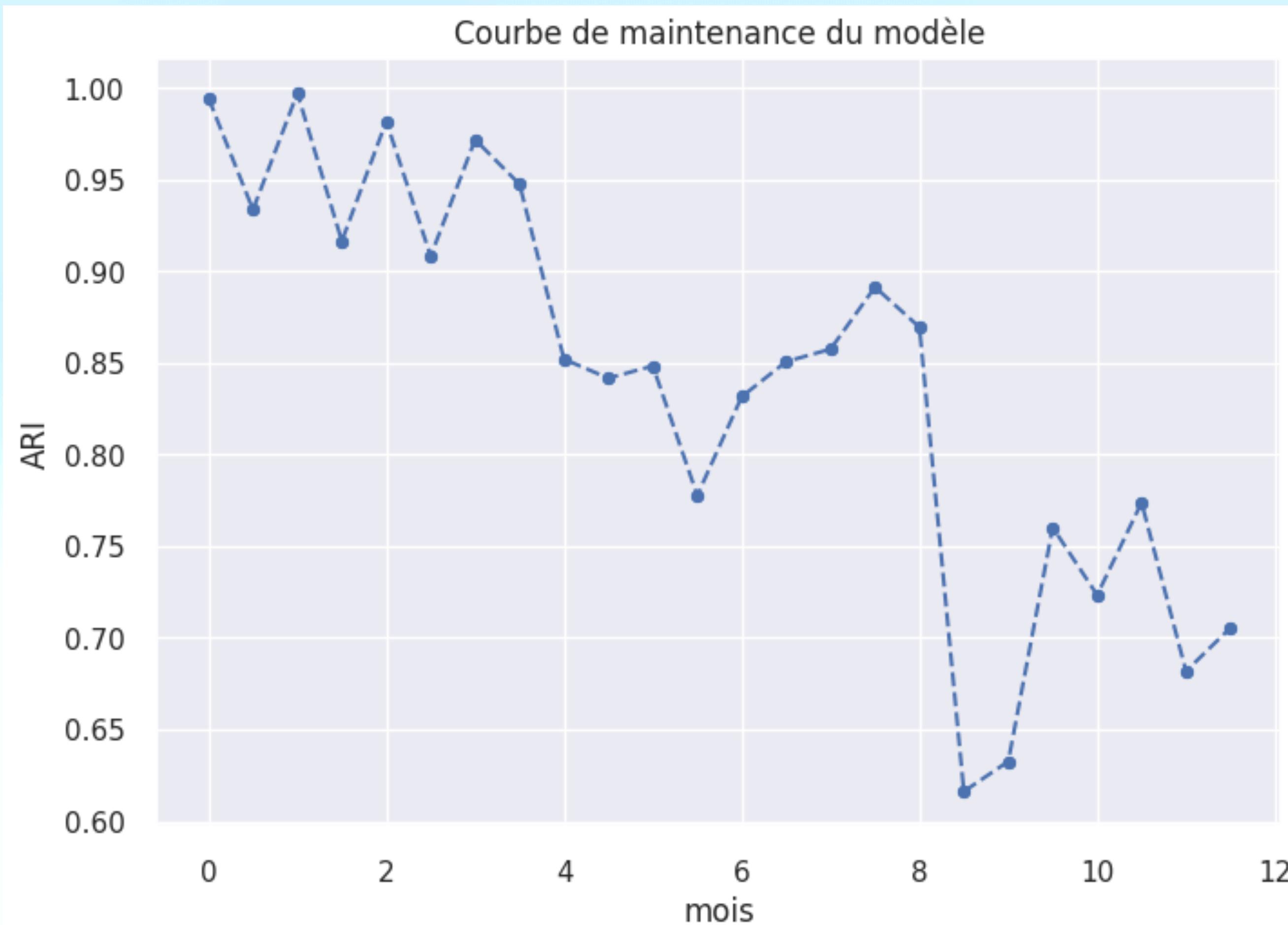


K\_means

k=4 | init= 'random' puis centroïdes | algorithm = 'lloyd' | n\_int = 1

La segmentation RFM  
nécessite une maintenance au  
bout de **5 mois et demi !**

# 4. Maintenance



K\_means

k=4 | init= 'random' puis centroïdes | algorithm = 'lloyd' | n\_int = 1

La segmentation RFM  
nécessite une maintenance au  
bout de **5 mois et demi !**

# Conclusion

- Essais de **3 types de segmentations**.
- La segmentation RFM classique a été choisie.
- **4 modèles de clustering** testés. Le modèle K\_means a été retenu avec **k = 4**
- Les « meilleurs clients » : (cluster 2 & 3)
  - ont acheté plusieurs fois entre 2016 et 2018 avec récence = 10 mois
  - ont acheté une fois, un an avant la collecte des données pour des montants moyens très élevés (= 2160 R\$)
- Une simulation de contrat de maintenance a été effectué. Le modèle choisi est valable pendant 5 mois et demi.