

Projet 5 : Segmentez des clients d'un site e-commerce



- **Problématique** : permettre à Olist, une entreprise brésilienne de vente en ligne de connaître ses clients, d'identifier leurs besoins spécifiques .
 - but : Ajuster l'offre commerciale pour les satisfaire.
- **Objectif** :
 - utiliser les données fournies par l'entreprise pour réaliser une segmentation de sa clientèle en groupes homogènes → mise en place de campagnes de communication ciblées.
 - Déterminer la fréquence à laquelle la segmentation doit être mise à jour.

Méthodologie de travail

- Choix de features pertinentes pour connaître le comportement des utilisateurs de la plateforme.
- Utilisation de modèles de machine learning non supervisés (les groupes de clients étant complètement ignorés au départ):
 - but : segmentation des clients en groupes homogènes et bien définis.
- Choix du meilleur modèle et étude de la fréquence à laquelle on doit réaliser sa maintenance:
 - but : garantir la stabilité du modèle au cours du temps pour maintenir la qualité du clustering.

Plan d'étude

- Présentation des données
- Nettoyage des données
- Feature engineering : création de nouvelles variables utiles à notre étude :
 - étude de la segmentation RFM
 - distribution des features
- Segmentation de la clientèle à partir de la RFM par plusieurs modèles:
 - K-means
 - DBSCAN
 - Clustering hiérarchique
- Ajout de 2 variables à la RFM : delivery_time et review_score
- Maintenance du modèle de machine learning

Présentation des données

- Les données sont anonymisées et mise à disposition par Olist.
- Les données sont disponibles sur le site internet Kaggle sur le lien suivant:
<https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce>
- Informations sur plus de 100 000 commandes client
- 9 tableaux csv : chaque tableau

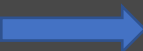
Présentation des données

Tableau	Type d'informations
olist_customers_dataset	▪ Identifiants des clients, leur ville et leur Etat
olist_geolocation_dataset	▪ Codes postaux des villes, leur longitude et latitude
olist_order_items_dataset	▪ Identifiants des commandes passées, des produits achetés et des vendeurs ▪ Prix des produits ▪ Date d'expédition ▪ Frais de livraison
olist_order_payments_dataset	▪ Moyen de paiement utilisé ▪ valeur de l'achat ▪ Paiement en une ou plusieurs fois
olist_order_reviews_dataset	▪ Commentaire et score de satisfaction laissés par les clients
olist_orders_dataset	▪ Identifiants, date des commandes et leur date d'expédition
olist_products_dataset	▪ Identifiants, noms, catégorie et tailles des produits
olist_sellers_dataset	▪ Identifiants des vendeurs avec leur code postal, leur ville et leur Etat
product_category_name_translation	▪ Traduction en anglais des catégories de produits

Nettoyage des données

- Jointures successives entre les tableaux par des variables communes pertinentes.
- On ne garde que les commandes livrées :
 - Cas où la variable 'order_status' correspond à la qualité 'delivered'
- Suppression des valeurs manquantes
- Suppression des doublons

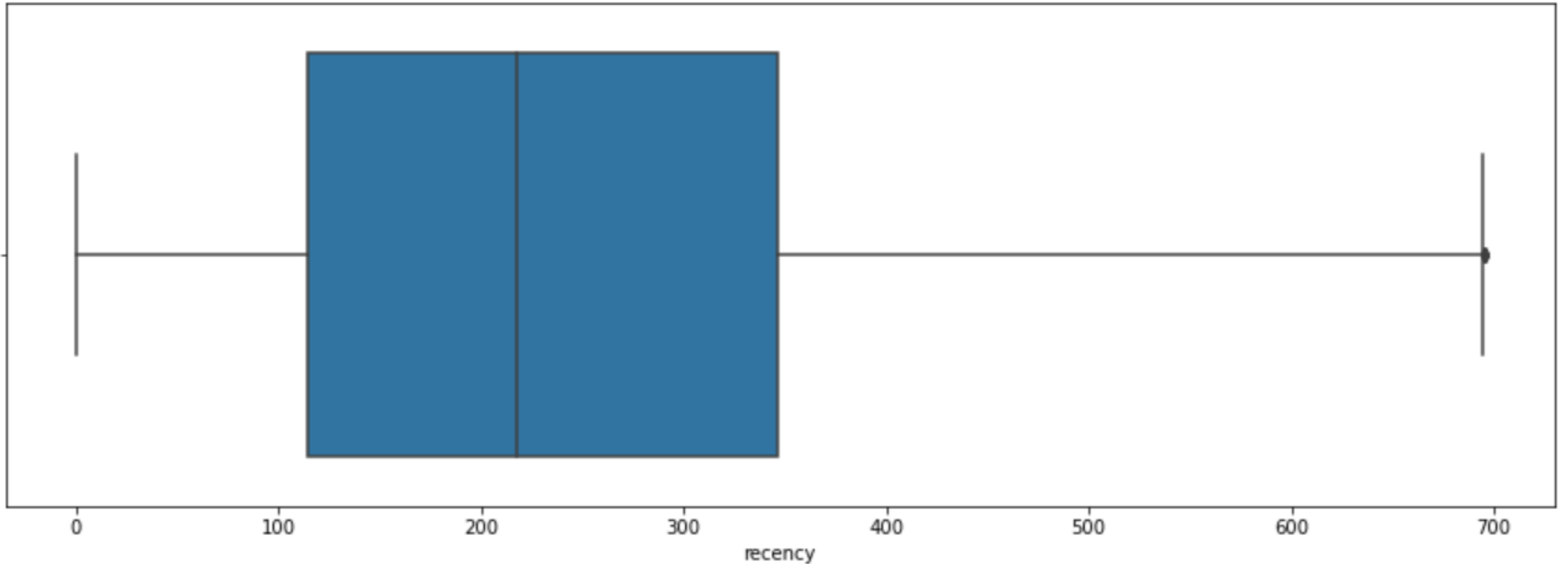
Segmentation RFM

- Segmentation très utilisée dans le domaine de vente directe.
- De l'historique des commandes faites par les clients, on déduit les comportements donnés par la RFM :
 - R : Récency (la récence) = nombre de jours écoulés depuis la dernière commande.
 - F : Fréquency (la fréquence) = nombre d'achat effectués par le client (le client achète t-il souvent?)
 - M : Montant = montant total dépensé.
-  création des 3 variables ' recency ', ' frequency ', ' monetary_value '

Distribution des variables RFM

- La récence :

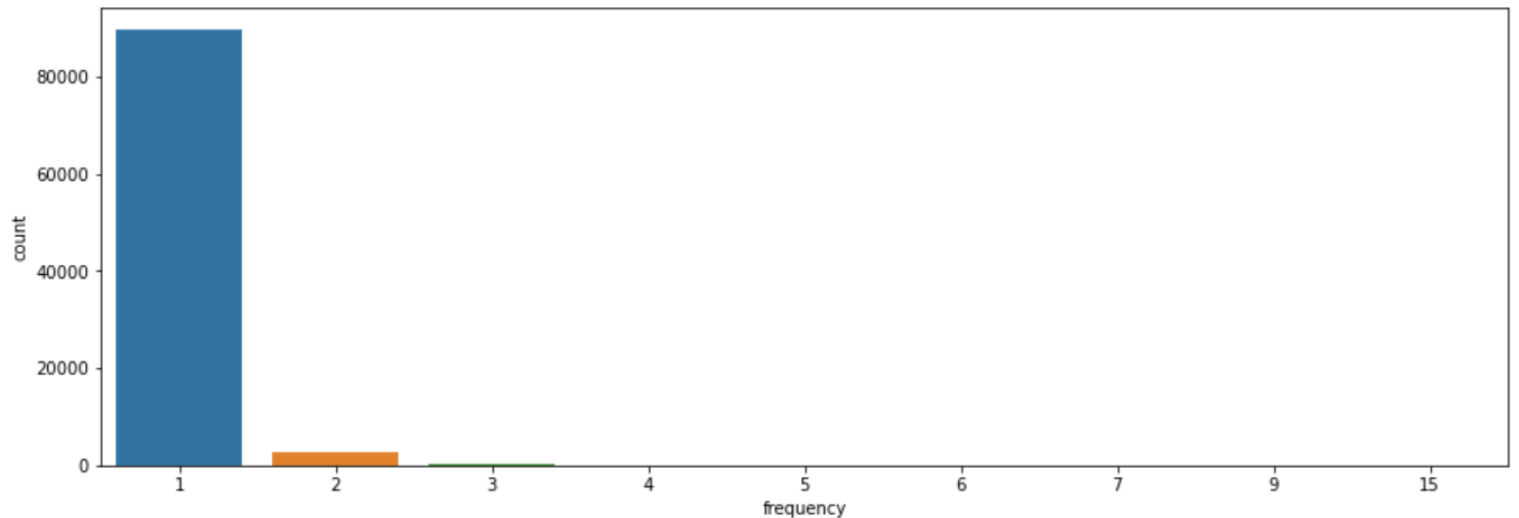
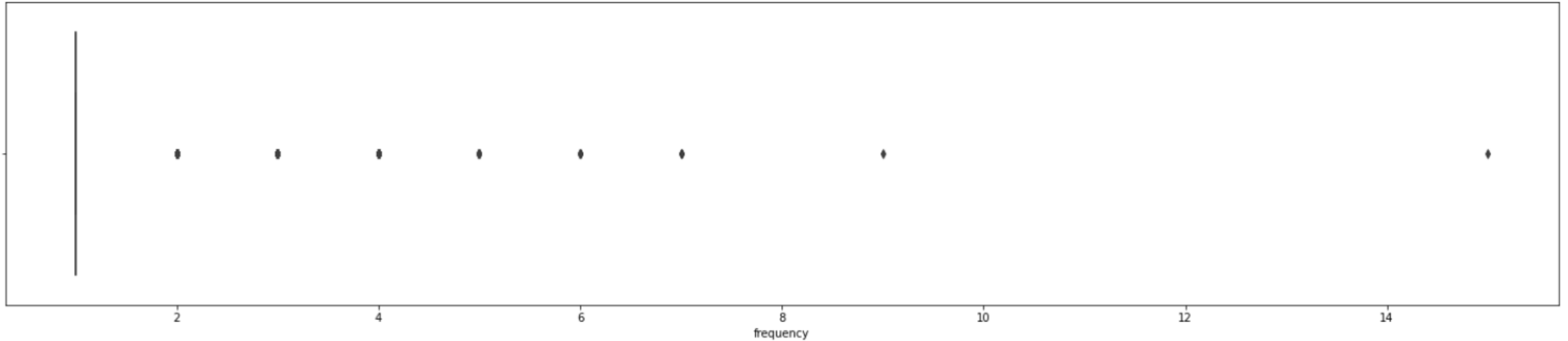
Distribution du nombre de jours écoulés depuis le dernier achat



Distribution des variables RFM

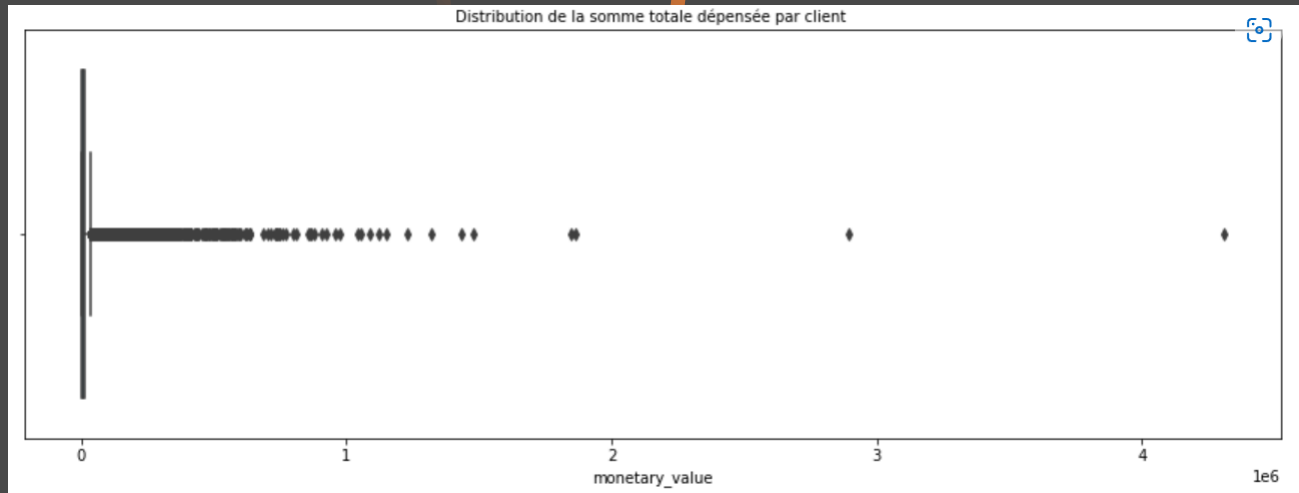
- La fréquence : un seul achat effectué en très grande majorité

Distribution du nombre d'achats effectués par les clients

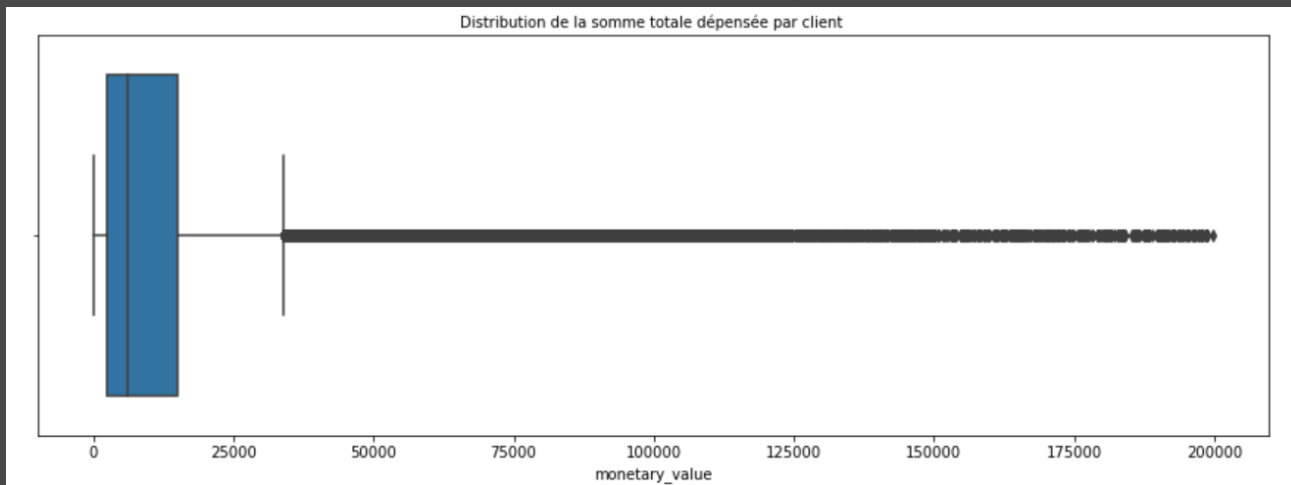


Distribution des variables RFM

- La somme totale dépensée :

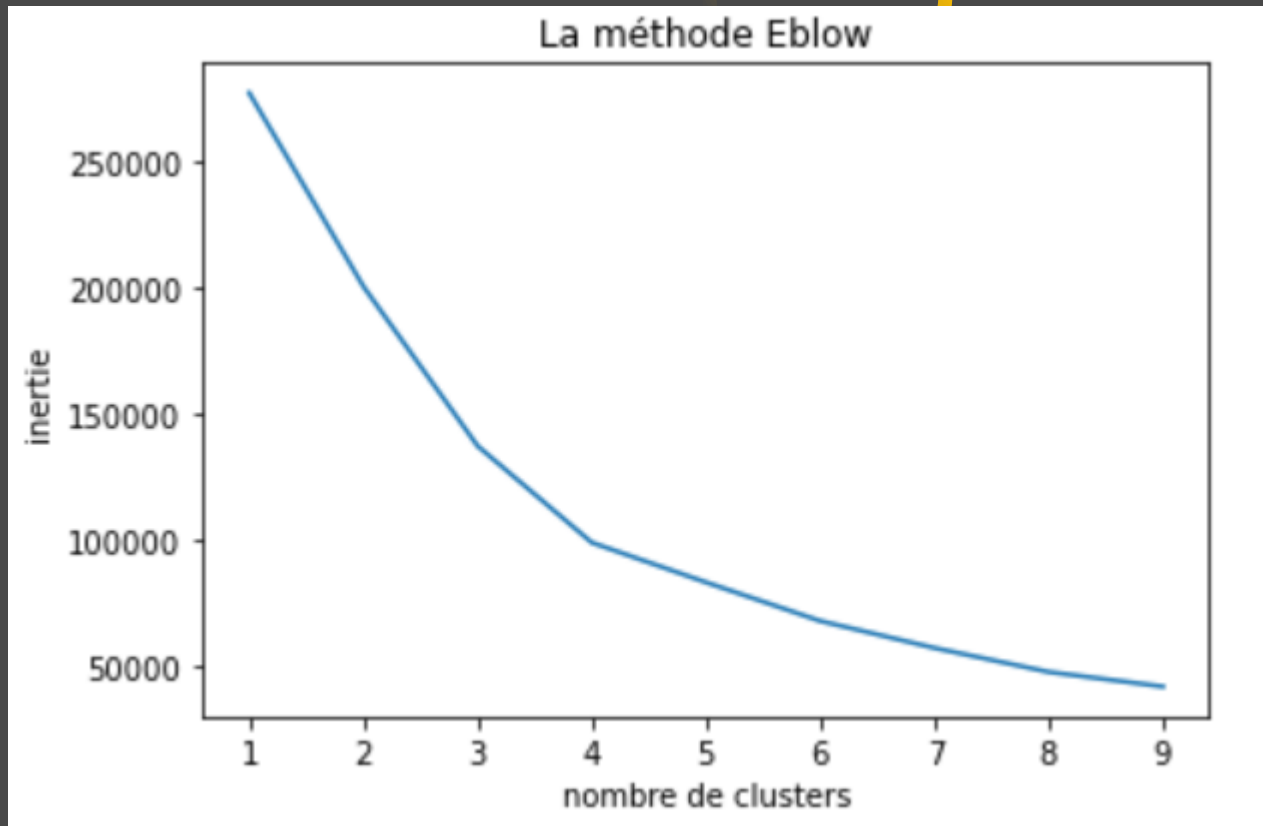


- Distribution pour la variable 'monetary-value' $\leq 2 \cdot 10^5$:



Modèle non supervisé : K – Means appliquée à la RFM

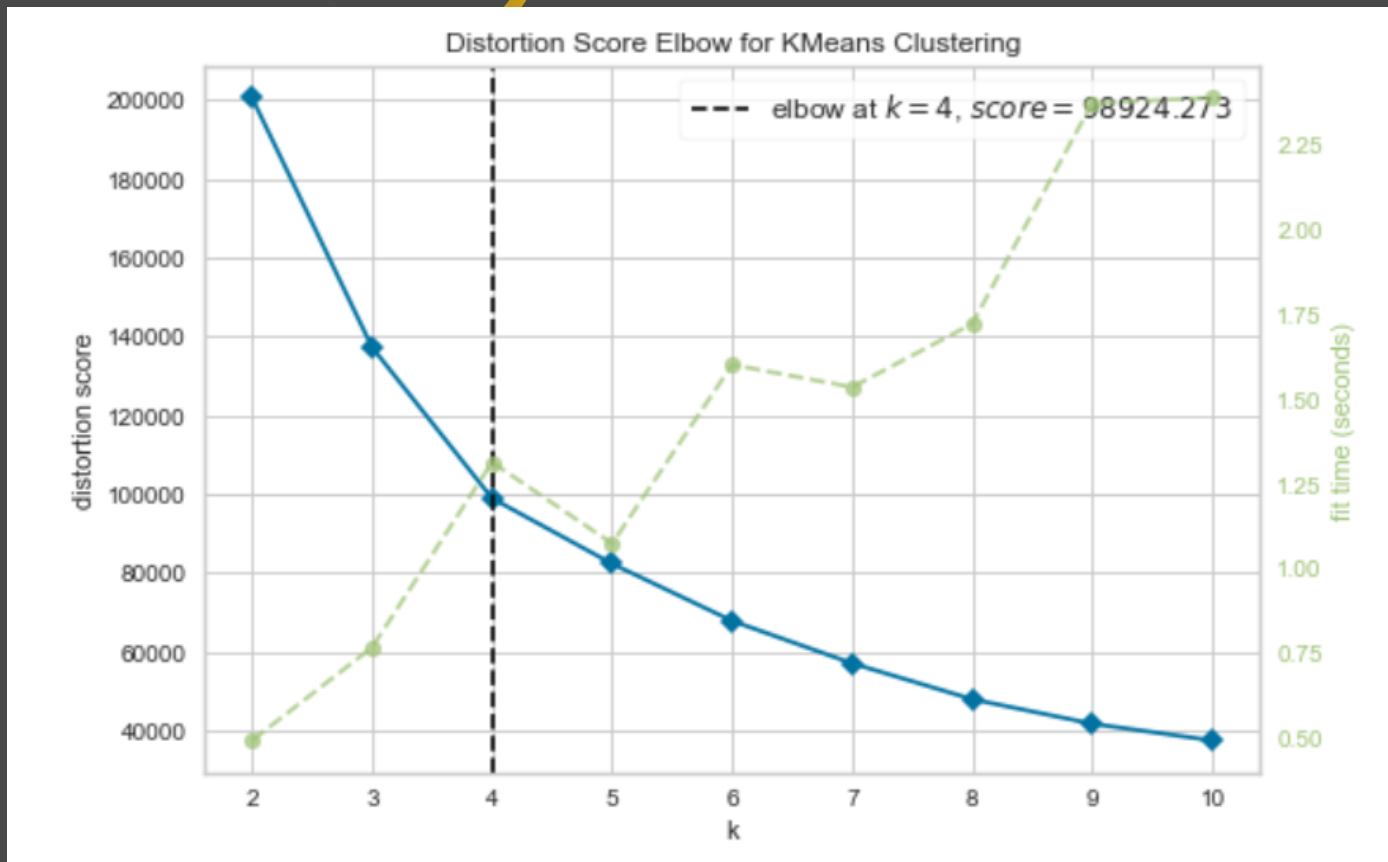
- Détermination du nombre de clusters par la méthode du coude:



- Nombre de clusters = 4

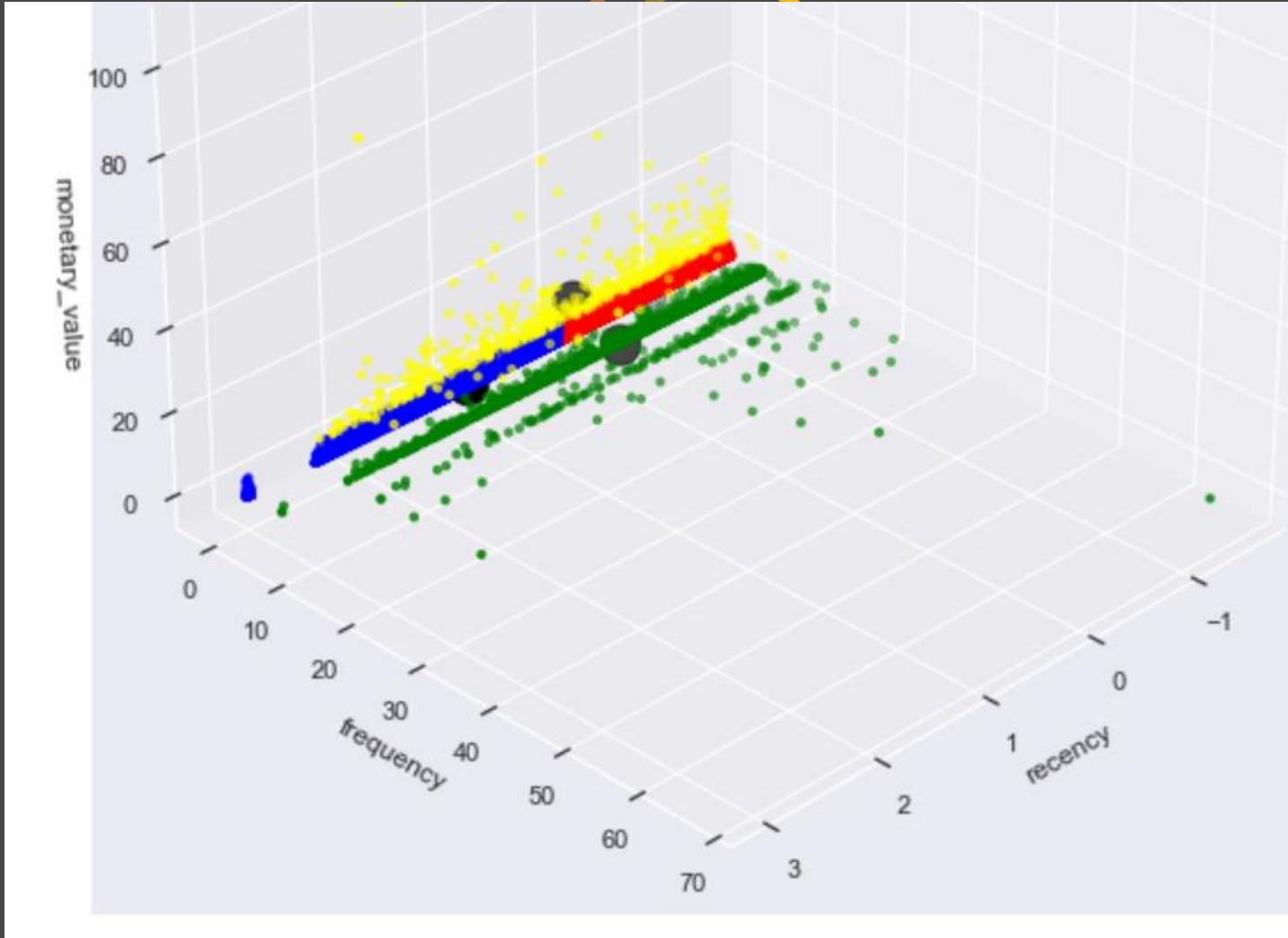
Modèle non supervisé : K – Means appliquée à la RFM

- Détermination du nombre de clusters :
 - méthode du coude basée sur le score de distorsion (somme moyenne des carrés des distances aux centres) → segmentation en K=4 clusters



Modèle non supervisé : K – Means appliquée à la RFM

- Visualisation des clusters pour les variables RFM standardisées



- Cluster vert : clients réguliers qui achètent souvent
- Cluster jaune : clients dont le montant des dépenses est le plus grand
- Cluster rouge : clients dont la commande est la plus récente
- Cluster bleu : clients dont la commande est la plus ancienne

Modèle non supervisé : K – Means appliquée à la RFM

- Tracé silhouette pour 4 clusters :

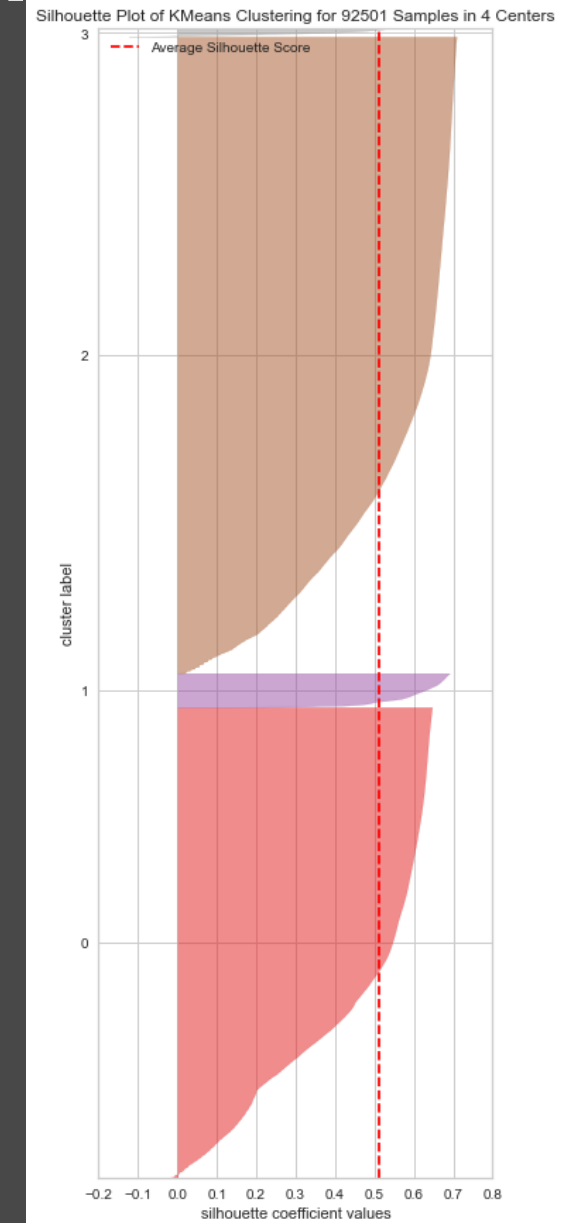
- score = 1 : grappes très denses et bien séparés.
- score = 0 : chevauchement des clusters.
- score < 0 : erreurs au niveau des clusters.

- Bonne séparation des clusters

- Les densités des clusters sont différentes

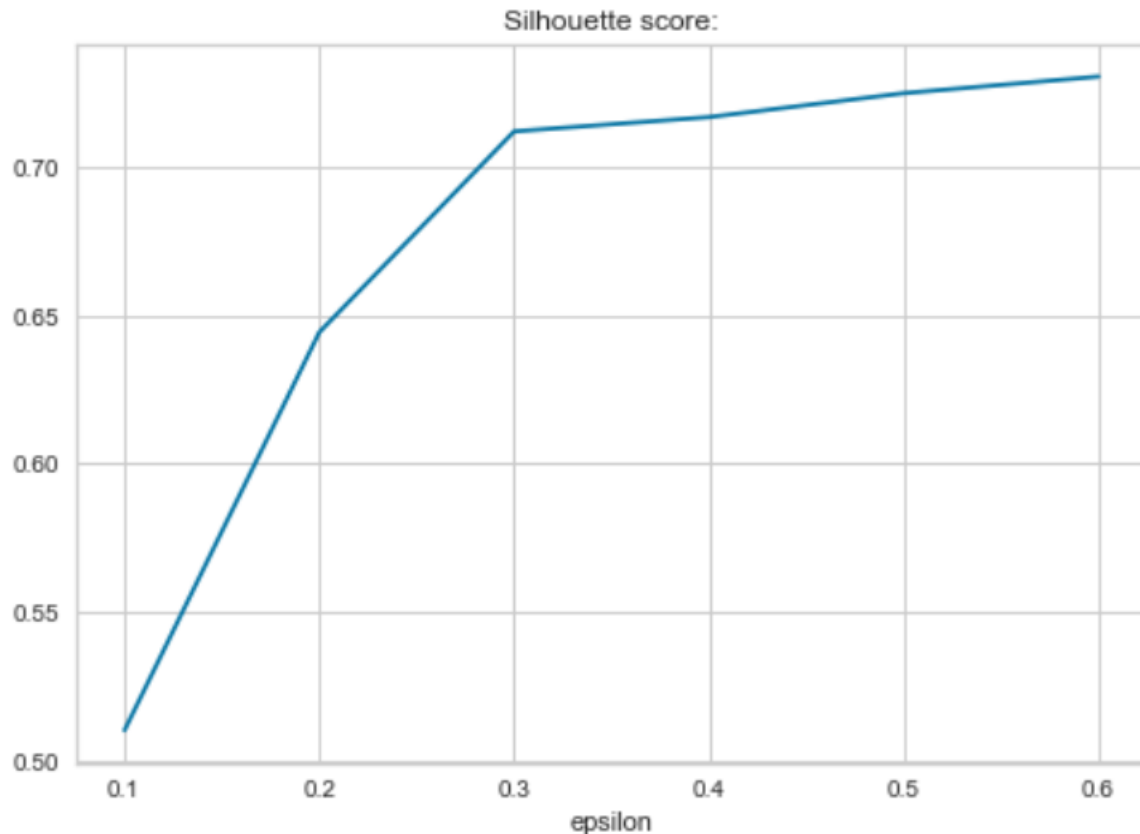
```
Nombre d'individus du claster 0 : 37900  
Nombre d'individus du claster 1 : 2730  
Nombre d'individus du claster 2 : 51266  
Nombre d'individus du claster 3 : 605
```

- Pas de score silhouette négatif → pas d'erreurs sur les clusters



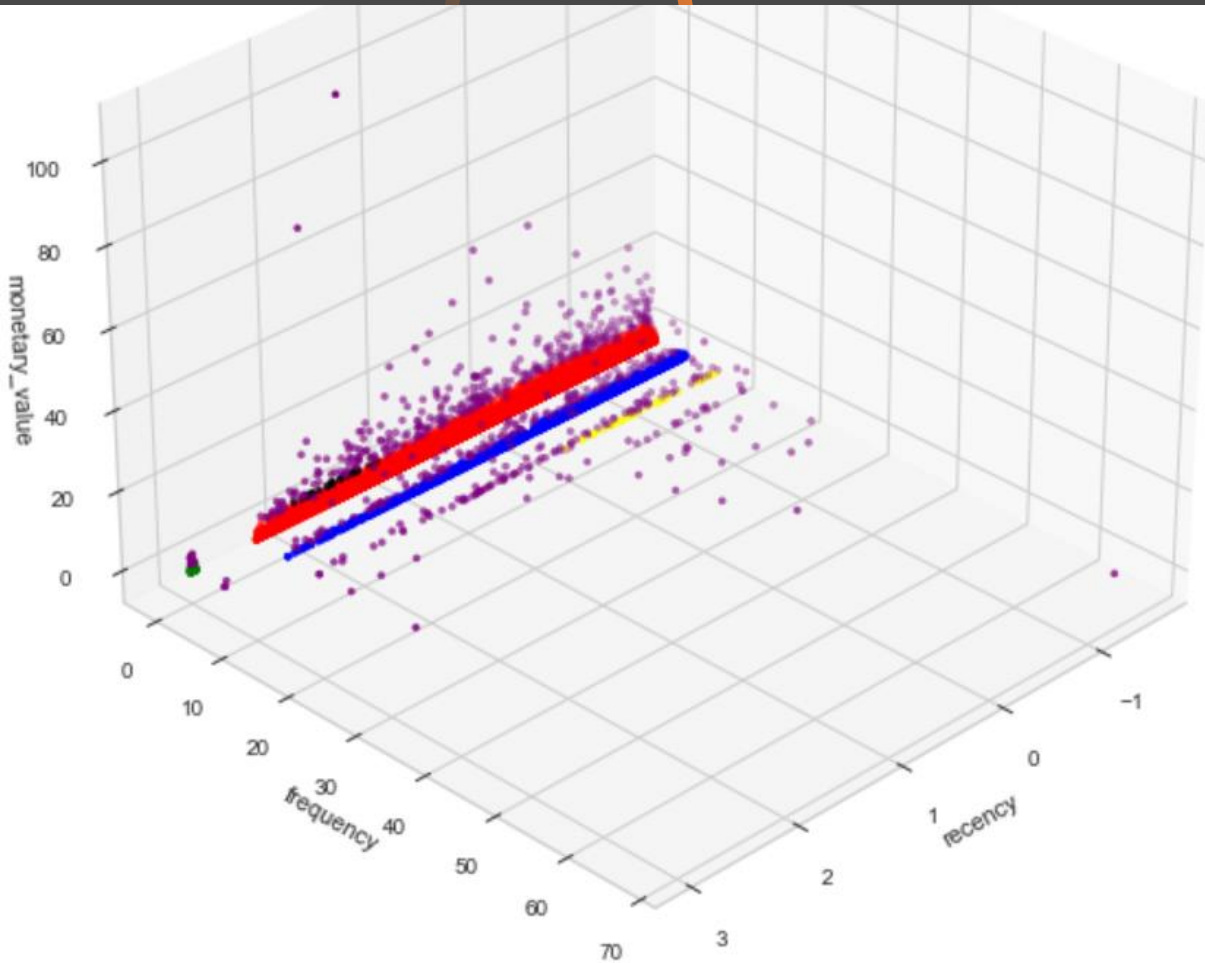
Modèle non supervisé : DBSCAN appliqué à la RFM

- Détermination des hyperparamètres par la méthode du coude:
 - epsilon = 0,3 obtenu pour min_sample = 30



```
(array([0.1, 0.2, 0.3, 0.4, 0.5, 0.6]),  
 [0.5100819053778827,  
  0.6443724743744625,  
  0.7121343556050999,  
  0.7169479476853694,  
  0.7250125425559933,  
  0.7306266318792131],  
 [25, 25, 20, 25, 25, 10])
```

Modèle non supervisé : DBSCAN appliqué à la RFM

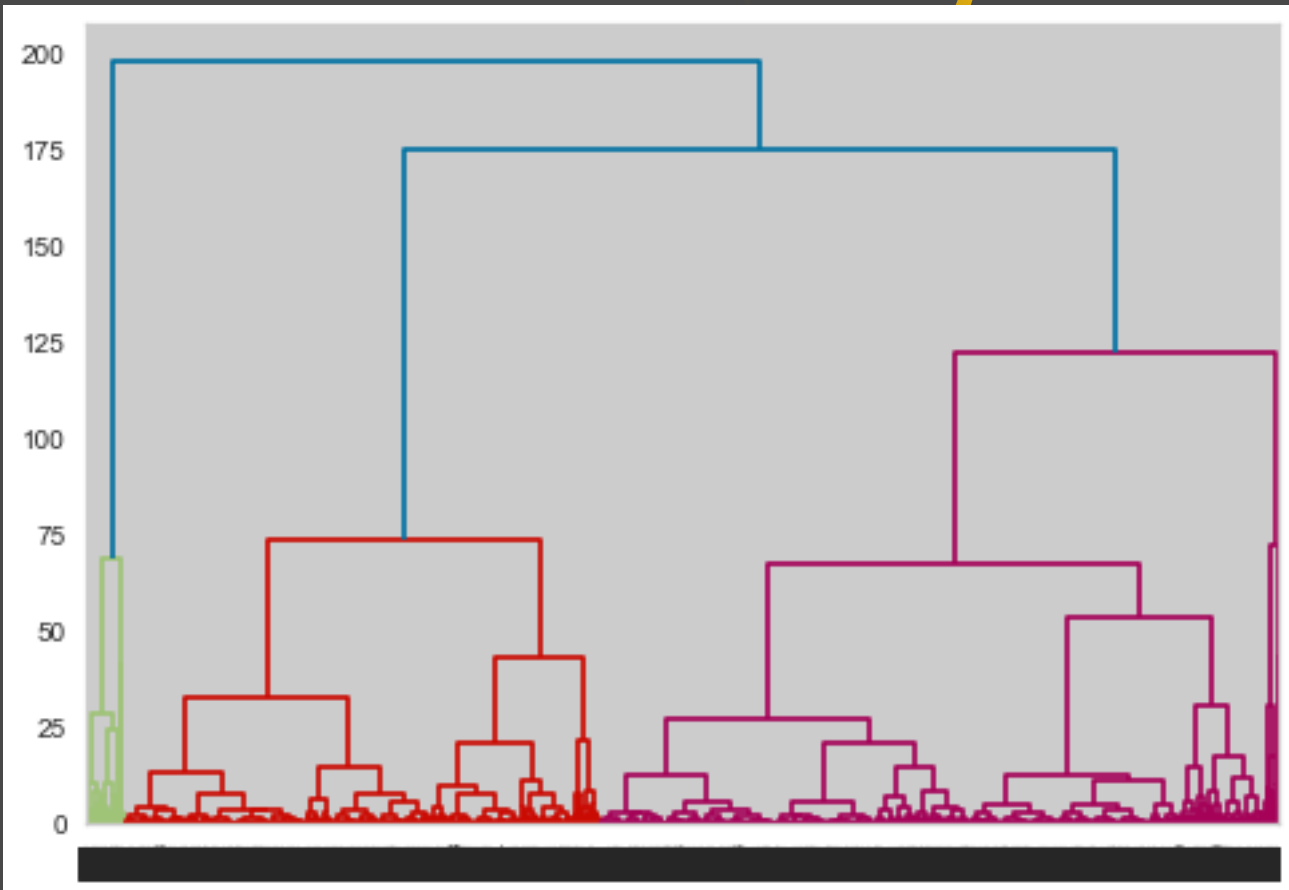


Estimated number of clusters: 6
Estimated number of noise points: 693
Silhouette Coefficient: 0.486

- L'algorithme n'est pas stable : En relançant plusieurs fois l'algorithme, on obtiens des résultats différents

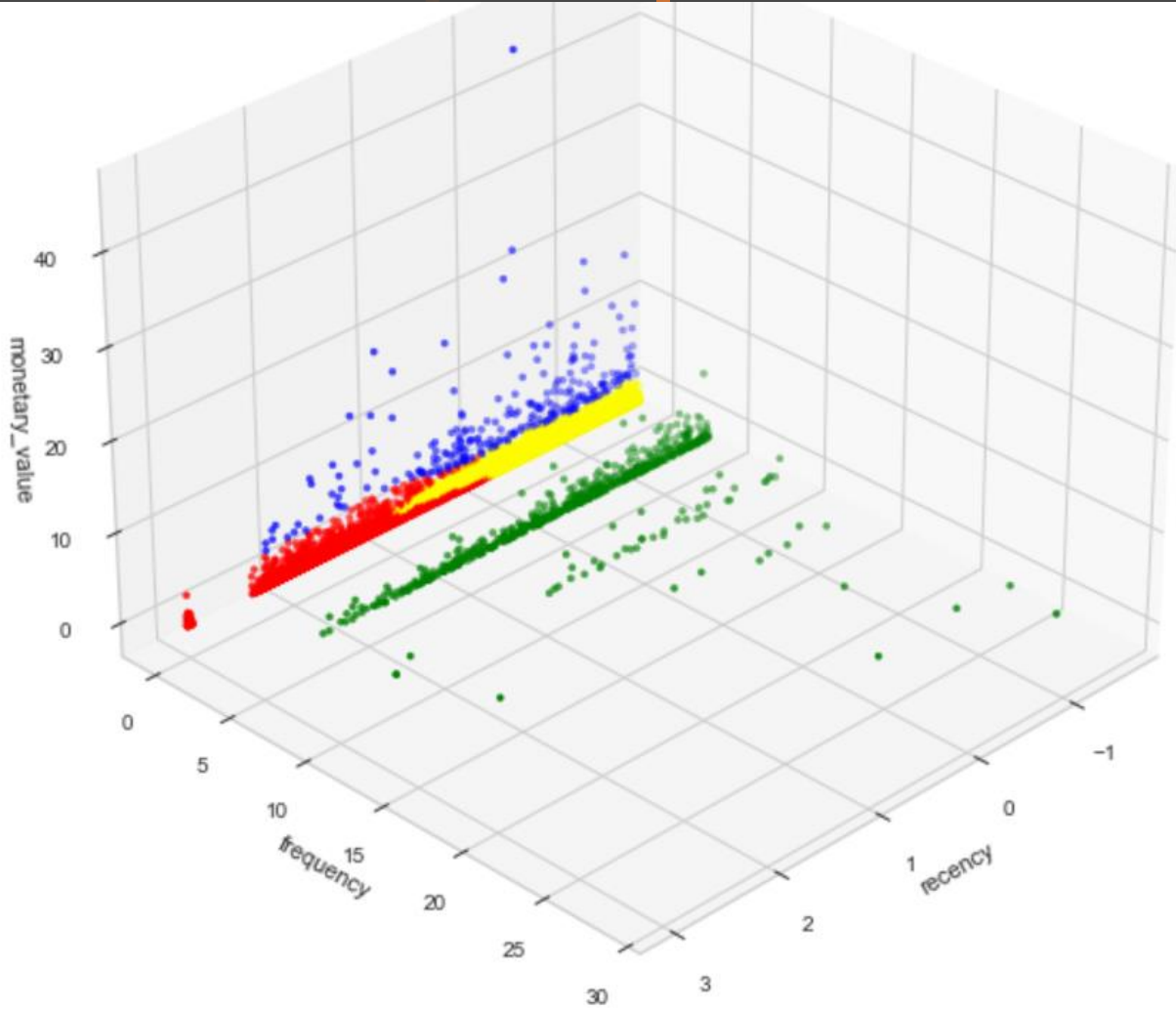
Modèle non supervisé : le clustering hiérarchique

- Utilisation du dendrogramme pour trouver le nombre de clusters optimal
n_clusters = 4



Score de la silhouette 0.5053902418304833

Modèle non supervisé : le clustering hiérarchique



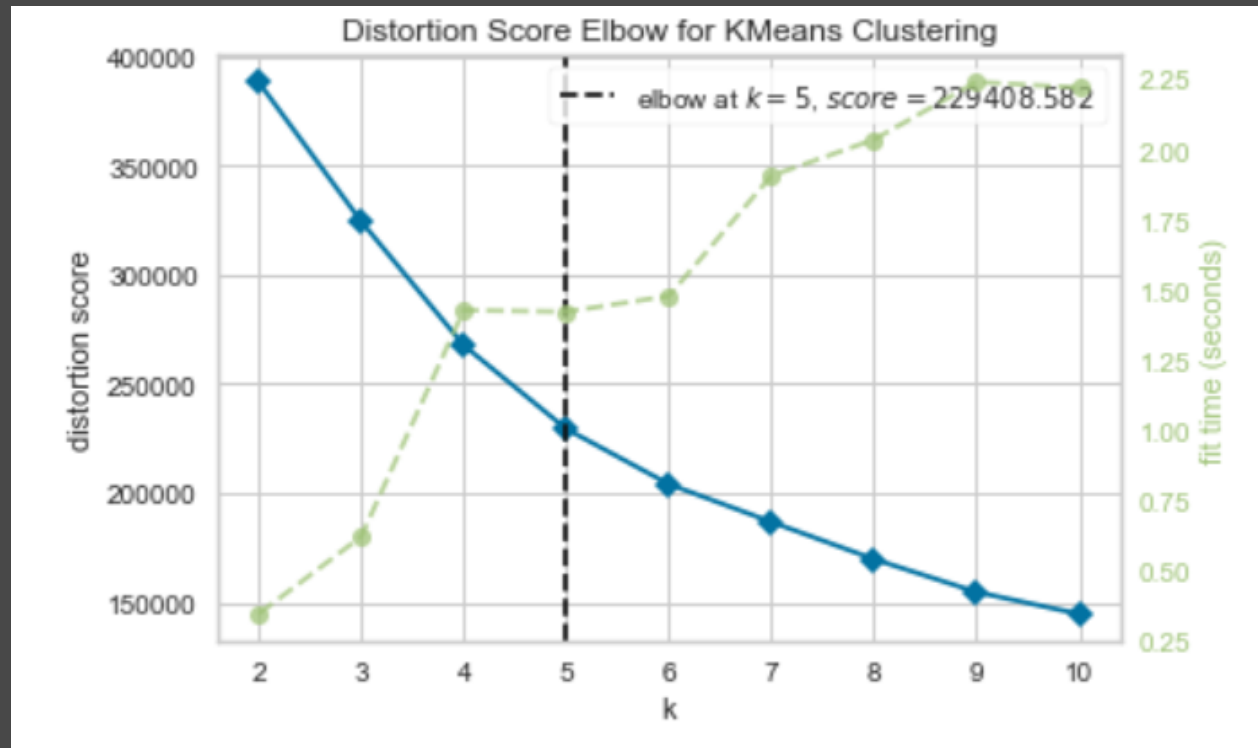
Score de la silhouette 0.5053902418304833

Synthèse

- Facilité de mise en place du K-means
- K-means est bien adapté dans notre cas car on a un très grand nombre de données (Il est toutefois sensible aux outliers).
- Difficulté de mise en place de Dbscan et du hiérarchique clustering qui demande beaucoup de temps de calcul
- Stabilité du K-means.
- La qualité des clusters obtenus par K-means est meilleure (score de silhouette plus important)

Appréciation de la satisfaction des clients

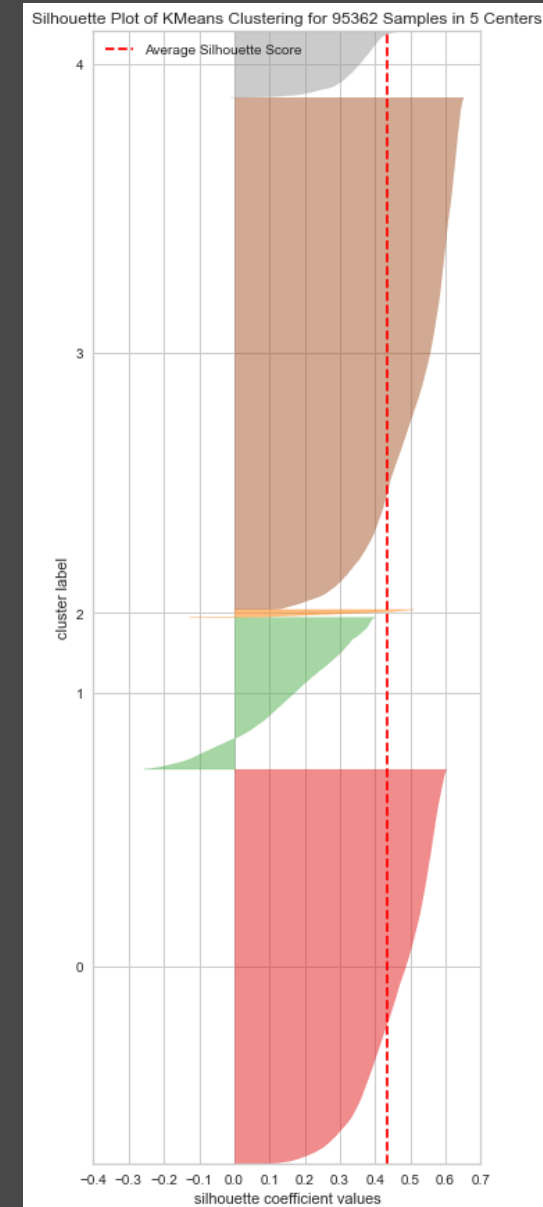
- Etude de la satisfaction des clients par l'ajout de 2 variables:
 - delivery_time (Temps de livraison)
 - review_score (note de satisfaction)
- Segmentation par le modèle K-means:
 - obtention de 5 clusters



Segmentation par l'algorithme K – Means

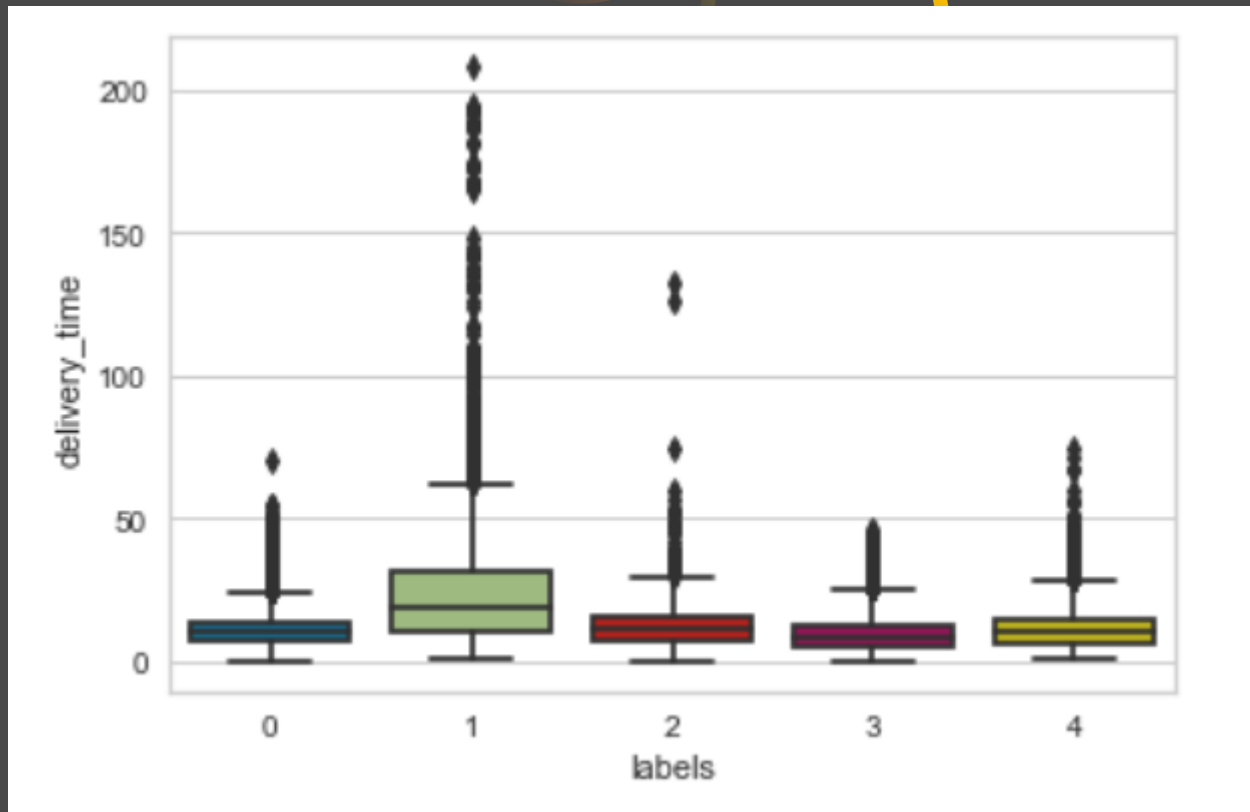
- Bonne séparation globale des clusters
- Quelques erreurs au niveau du cluster 1
- Les densités des clusters sont différentes

```
Nombre d'individus du cluster 0 : 33252  
Nombre d'individus du cluster 1 : 12796  
Nombre d'individus du cluster 2 : 656  
Nombre d'individus du cluster 3 : 43144  
Nombre d'individus du cluster 4 : 5514
```



Caractéristiques des 5 clusters

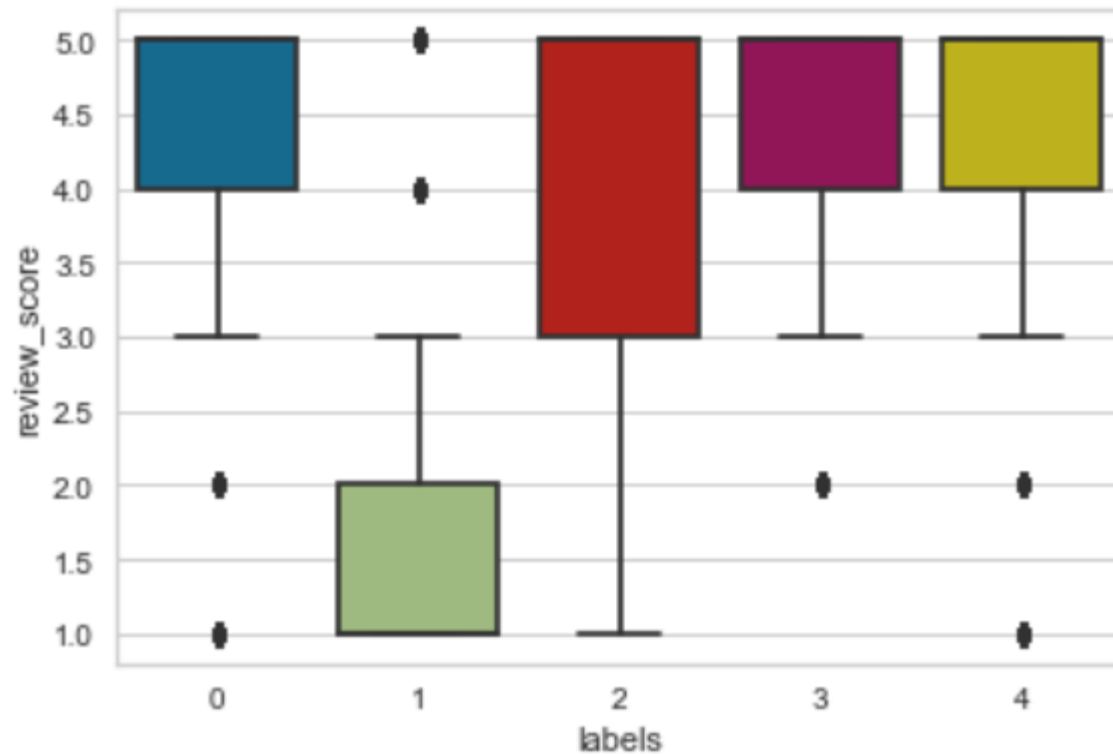
- Distribution du temps de livraison pour les 5 clusters:



- Temps de livraison plus important pour le cluster 1

Caractéristiques des 5 clusters

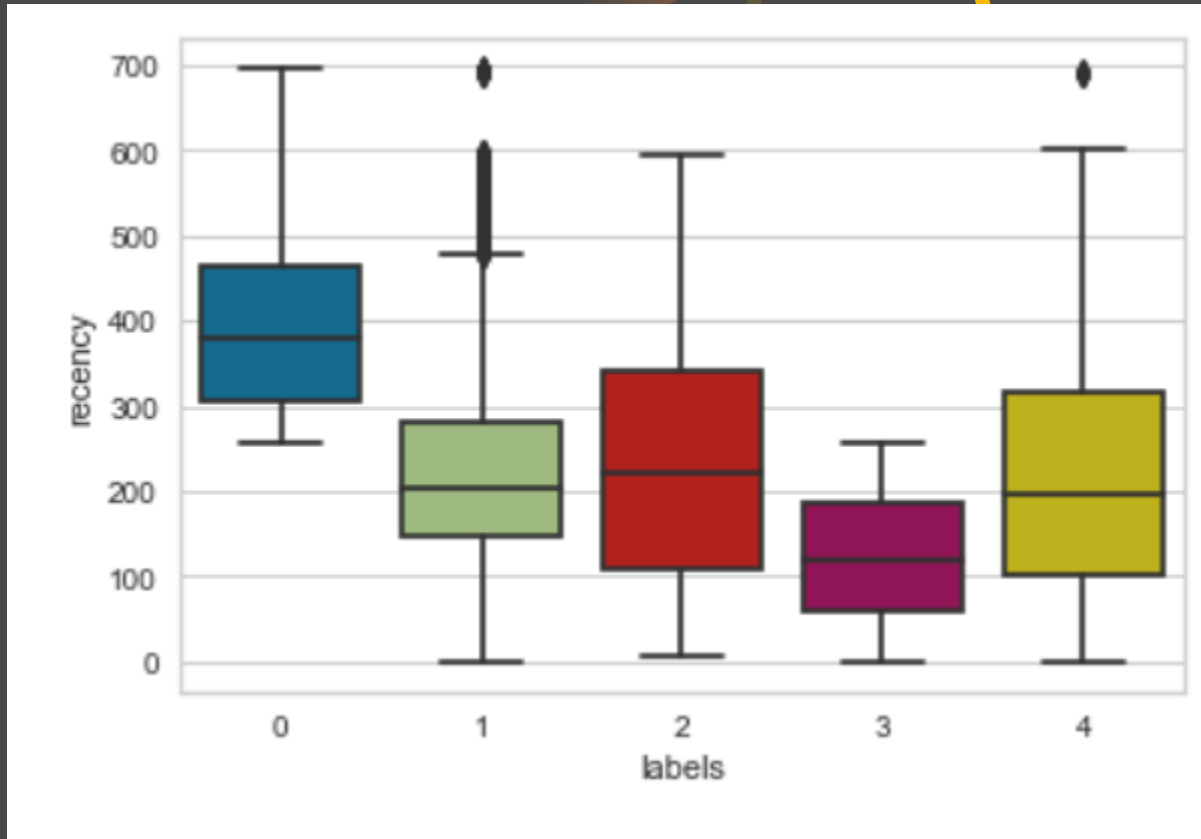
- Distribution de la variable review_score pour les 5 clusters:



- Cluster 1 : clients les moins satisfaits (score_review le plus faible)
(possible lien entre le délai de livraison et la satisfaction du client)

Caractéristiques des 5 clusters

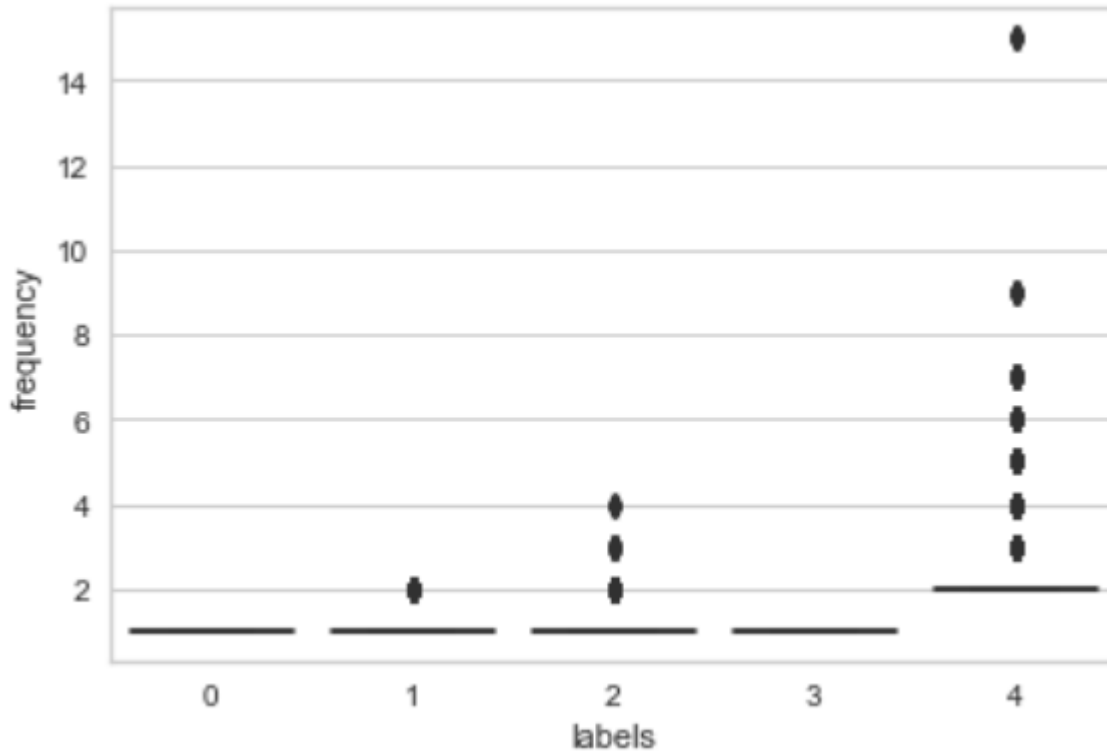
- Distribution de la variable recency pour les 5 clusters:



- Cluster 3 : clients dont la commande est la plus récente

Caractéristiques des 5 clusters

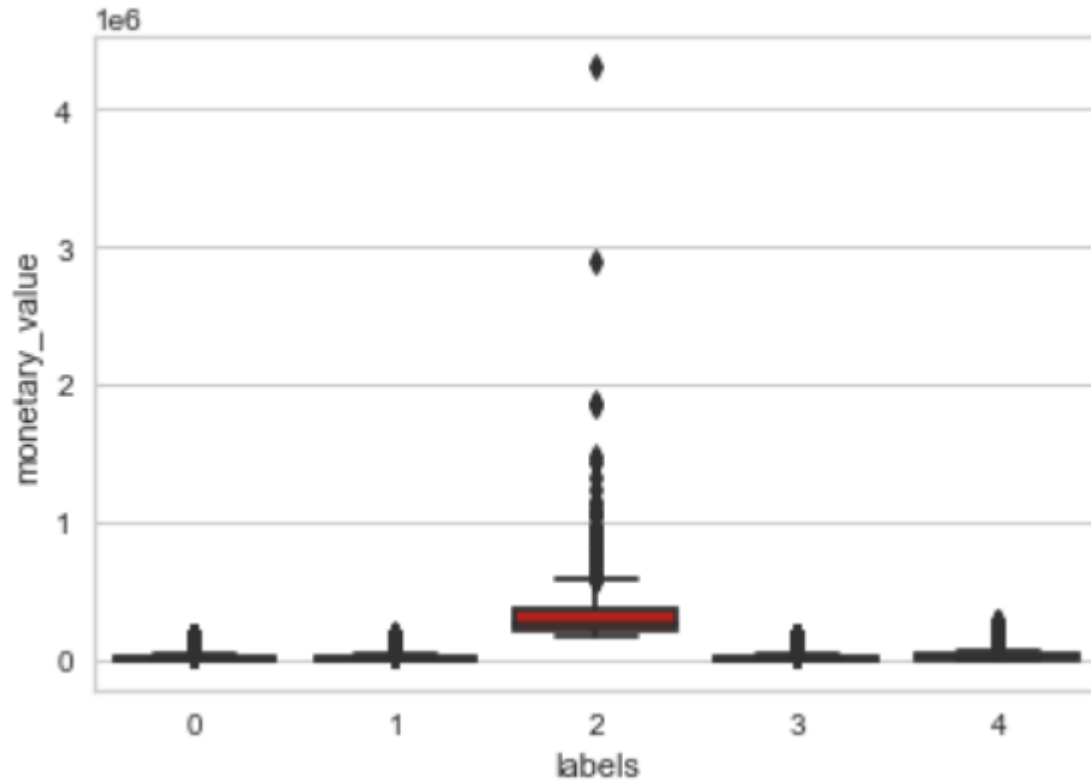
- Distribution de la variable frequency pour les 5 clusters:



- Cluster 4 : clients réguliers, qui commandent plus souvent

Caractéristiques des 5 clusters

- Distribution de la variable monetary_value pour les 5 clusters:



- Cluster 2 : clients qui ont dépensé le plus pour leur commandes

Maintenance du modèle

- Etude de la stabilité du modèle dans le temps :
 - classification des données, de la commande la plus ancienne à la plus récente
 - séparation du jeu données en périodes successives pour simuler l'évolution du temps (première période : 3 premiers mois, deuxième période: 6 premiers mois etc.)
 - comparaison sur ces périodes entre le modèle entraîné une seule fois (uniquement sur la première période) et celui qui est entraîné à toute les périodes

Maintenance du modèle K-means

- Maintenance du modèle K-means appliqué à la RFM :
- ARI : mesure de la similarité entre les 2 clusterings (celui du modèle entraîné une seule fois et celui réentraîné à chaque période).

	ARI1	ARI2	ARI3	ARI4	ARI5	ARI6	ARI7
0	1.0	0.986393	0.963614	0.931915	0.91282	0.85851	0.97523

- $ARI < 90\%$ après 1 an et 3 mois = fréquence de maintenance du modèle

Maintenance du modèle K-means

- Maintenance du modèle K-means appliqué à la RFM associée à `delivery_time` et `review_score` :

	ARI1ter	ARI2ter	ARI3ter	ARI4ter	ARI5ter	ARI6ter	ARI7ter
0	1.0	0.988551	0.966055	0.928476	0.871811	0.888854	0.955095

- $ARI < 90\%$ après 1 an = fréquence de maintenance du modèle

Conclusions

- Notre modèle K-means nous a permis de segmenter la clientèle en clusters dont on a pu définir facilement la nature.
- Ce modèle nous offre une description facilement actionnable que peut utiliser l'équipe marketing pour faire des campagnes de communication ciblées.
- Toutefois une maintenance régulière du modèle s'impose pour garantir sa stabilité au cours du temps à mesure que d'autres données s'ajoutent
- Possibilité de développer notre modèle en y intégrant d'autres variables pour avoir des groupes de clients avec d'autres caractéristiques et une connaissance du client encore plus nuancée pour adapter encore plus la communication