



PROJET N°4

Parcours ingénieur machine learning

SEGMENTER DES CLIENTS D'UN SITE E-COMMERCE



Présenté par :
Mounira Abderrahmani

Mentor :
Denis Lecoeuche



PLAN

01

Problématique

02

Récupération
des données

03

Nettoyage et
exploration

04

Modélisation

05

Période de
maintenance

06

Conclusion





1. Problématique :

Contexte : • Consultant pour Olist

Olist :

- Entreprise brésilienne
- Marketplace
- Solution de vente en ligne

Mission: • Fournir une segmentation des clients
pour les campagnes de communication





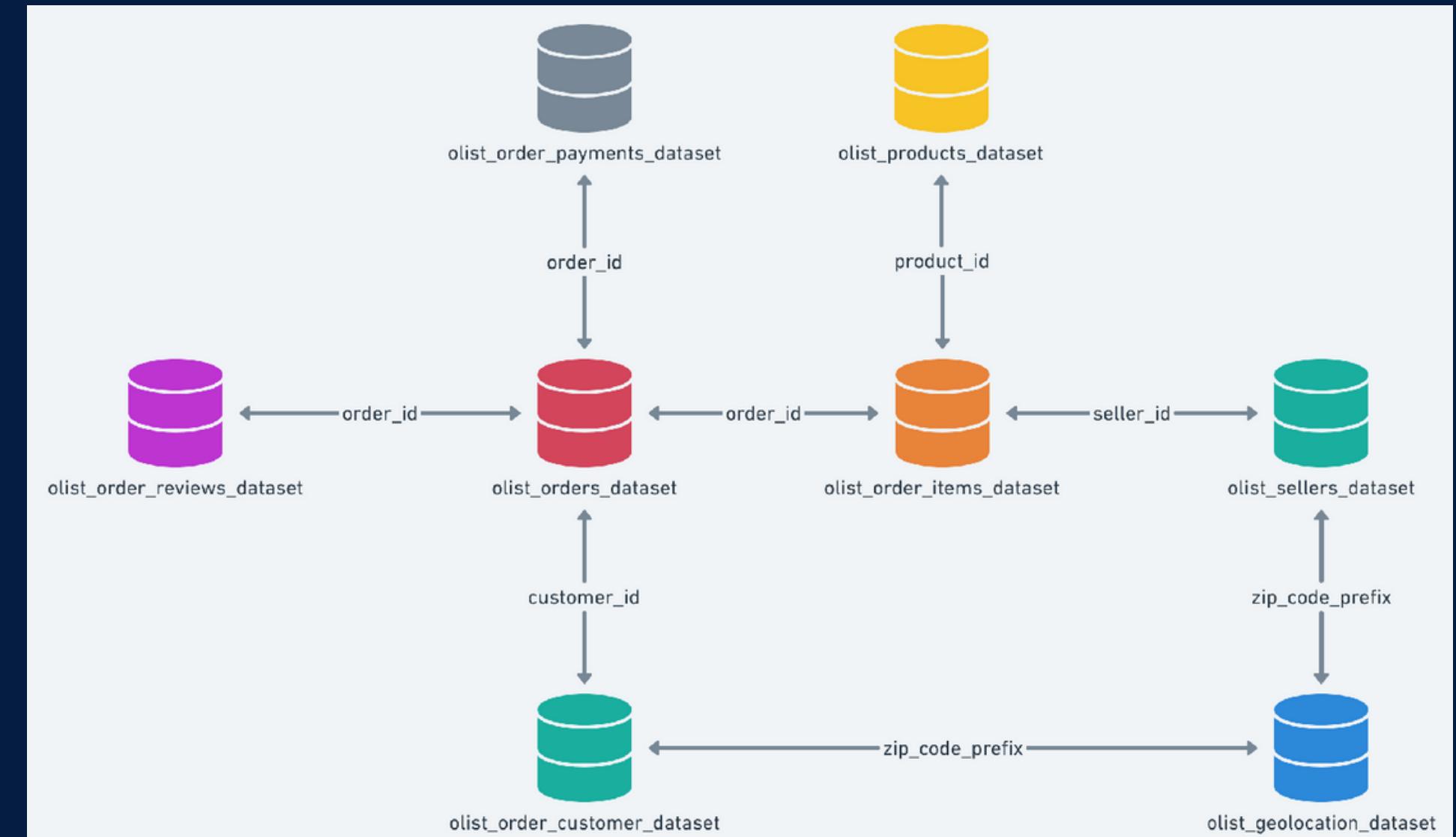
1. Problématique :

Objectifs

- 01 Comprendre les différents types d'utilisateurs
- 02 Fournir une description actionnable de la segmentation
- 03 Fournir une proposition de contrat de maintenance

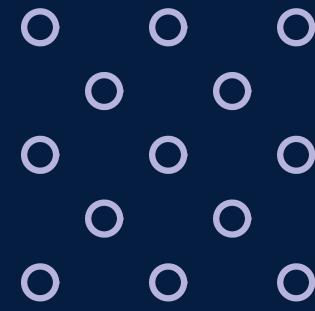
2. Récupération des données

- Données anonymisées disponible sur Kaggle
- 9 fichiers csv :
 - Commandes : 100K de 2016 à 2018,
 - Clients : 96K uniques,
 - Localisation,
 - Produits,
 - Produits par commande
 - Commentaires clients,
 - Paiement,
 - Vendeurs,
 - Traduction catégories produits





2. Récupération des données

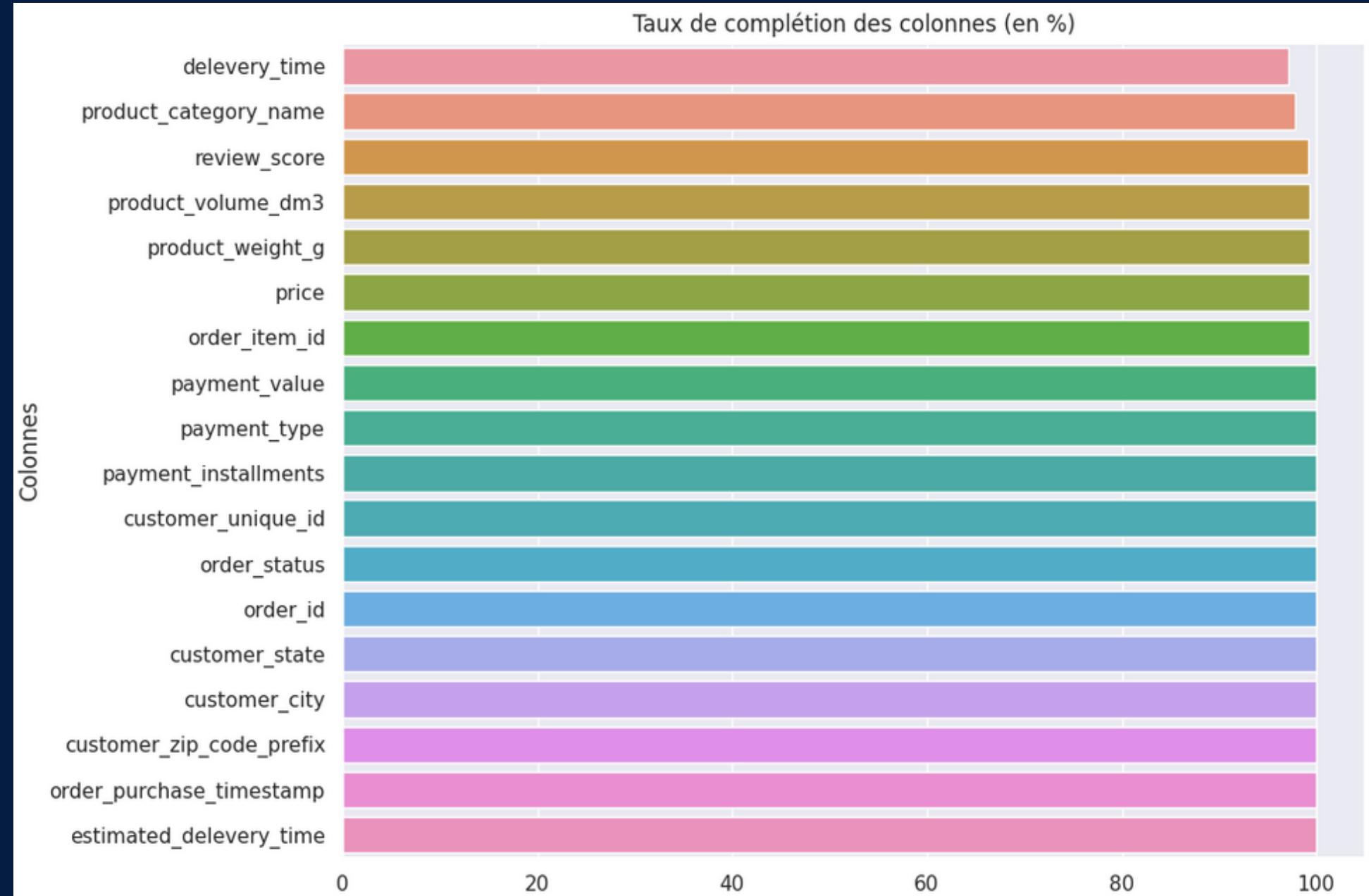


Etapes :

1. Assemblage des tables : commandes, clients, produits, paiements, commentaires, traduction catégories
2. Sélection des variables pertinentes

Table finale :

- 18 variables
- 119 143 lignes





3. Nettoyage et exploration

Étapes de nettoyage :

Méthode des percentiles 1 - 99

Table finale :
Une ligne = un client

Table finale :
96 096 lignes

Réduction des modalités

Suppression d'outliers

Nouvelles variables

Aggrégation

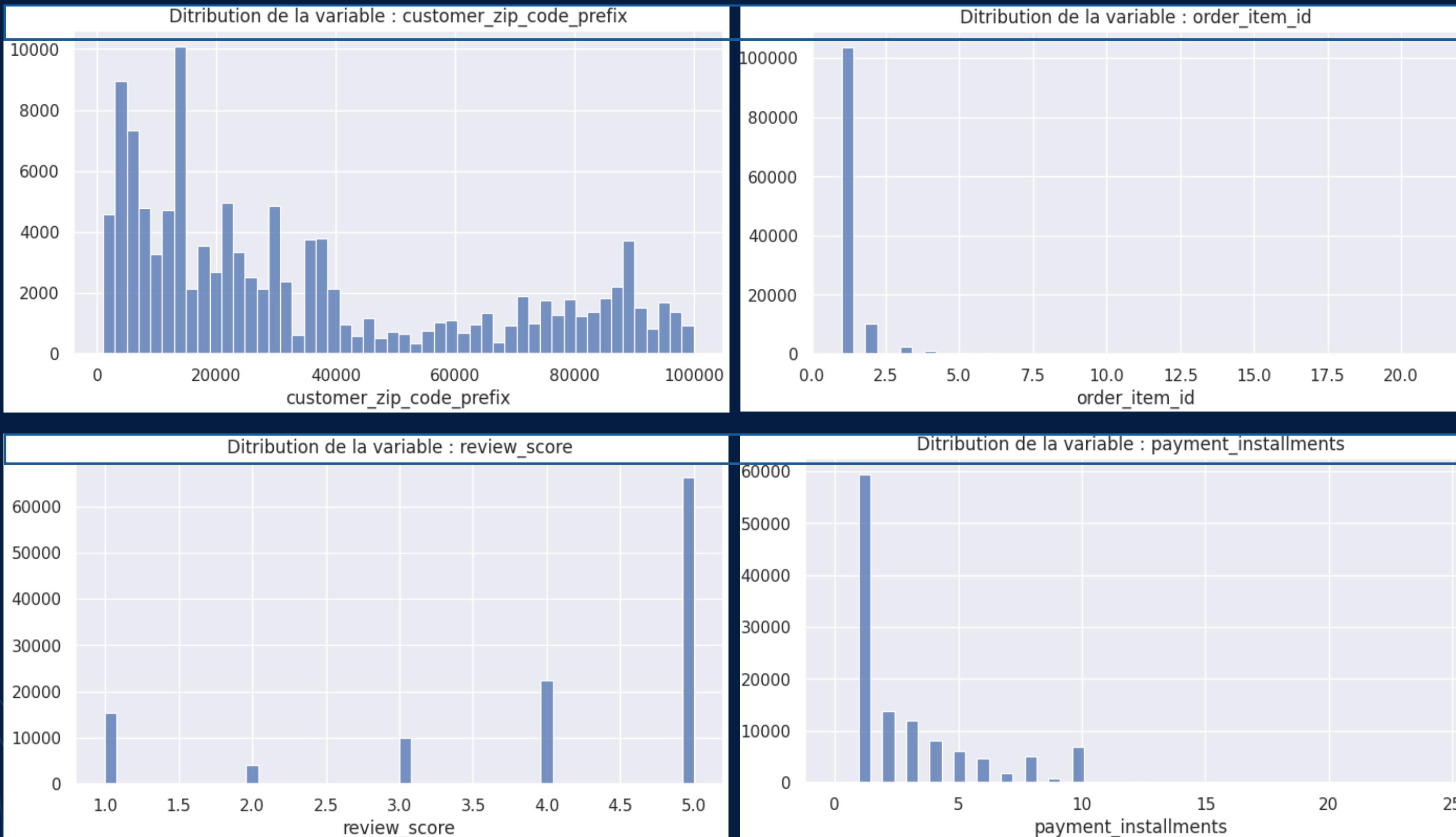
Catégories de produits
72 → 14

- Volume produits dm3
- Délai de livraison estimé (jour)
- Délai de livraison effectif (jour)
- Variables RFM
 - Récence (jours écoulés depuis la dernière visite)
 - Fréquence (nombre de visite)
 - Montant total (real brésilien)

3. Nettoyage et exploration

Exploration

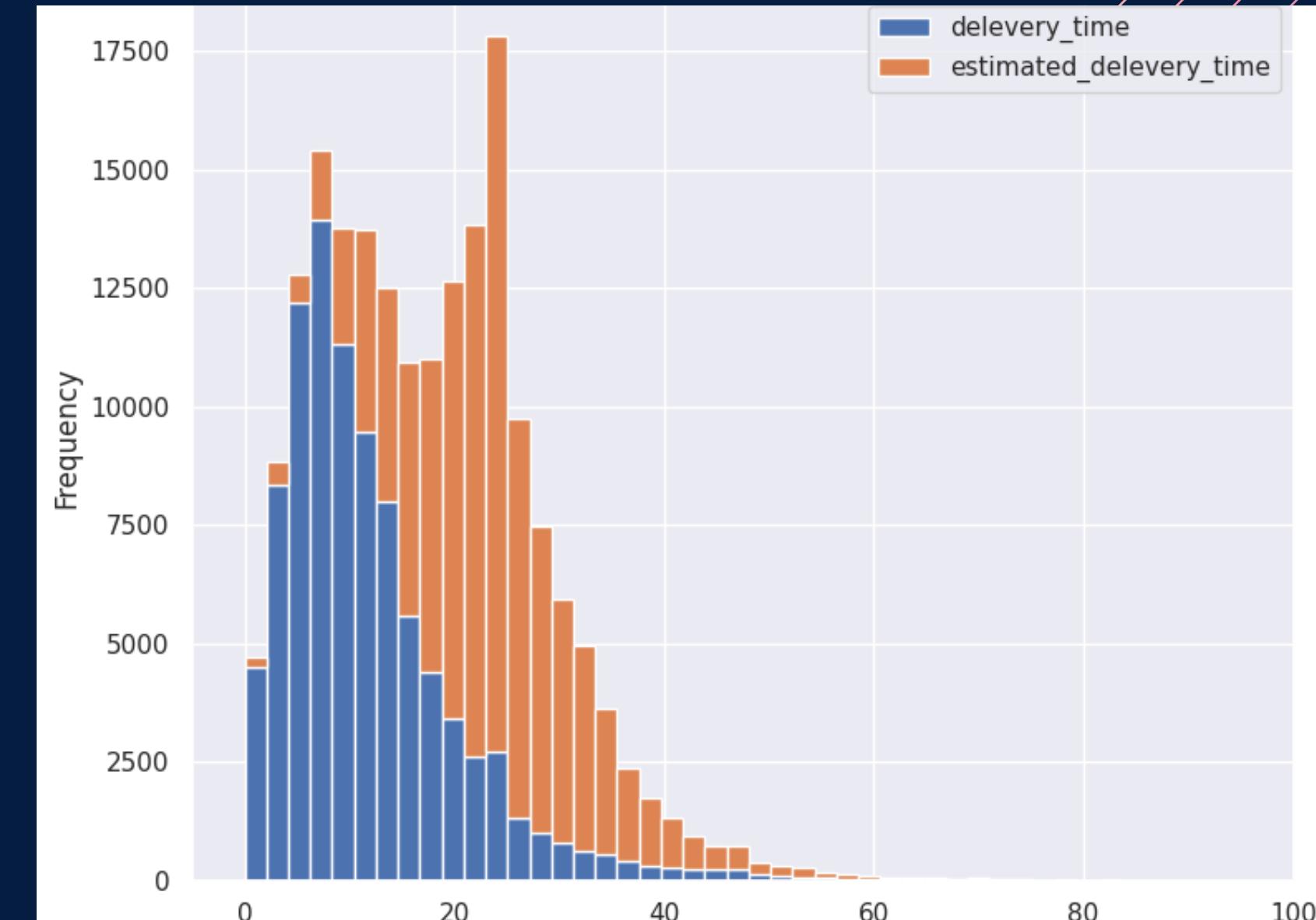
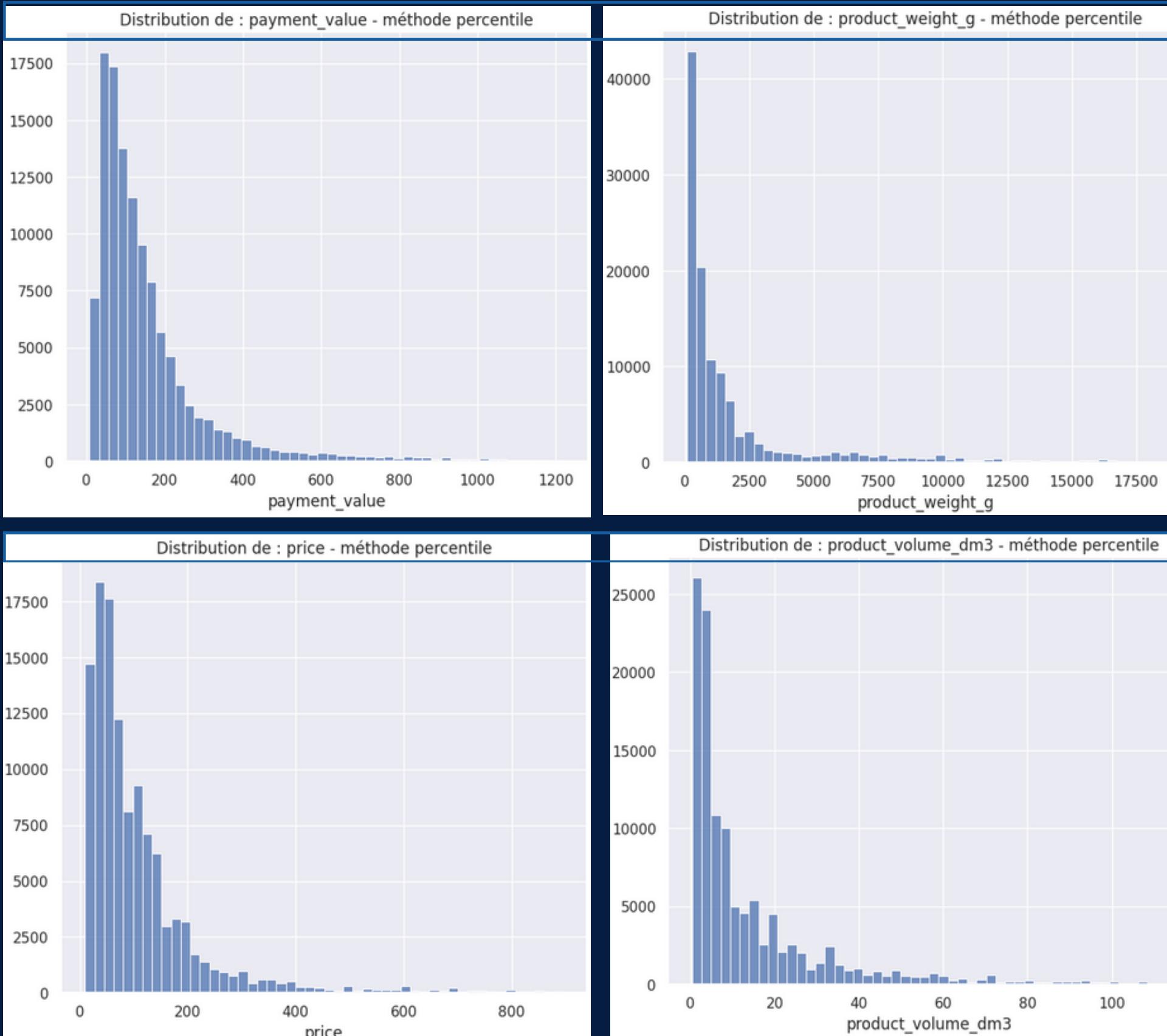
Variables numériques discrètes



3. Nettoyage et exploration

Exploration

Variables numériques continues



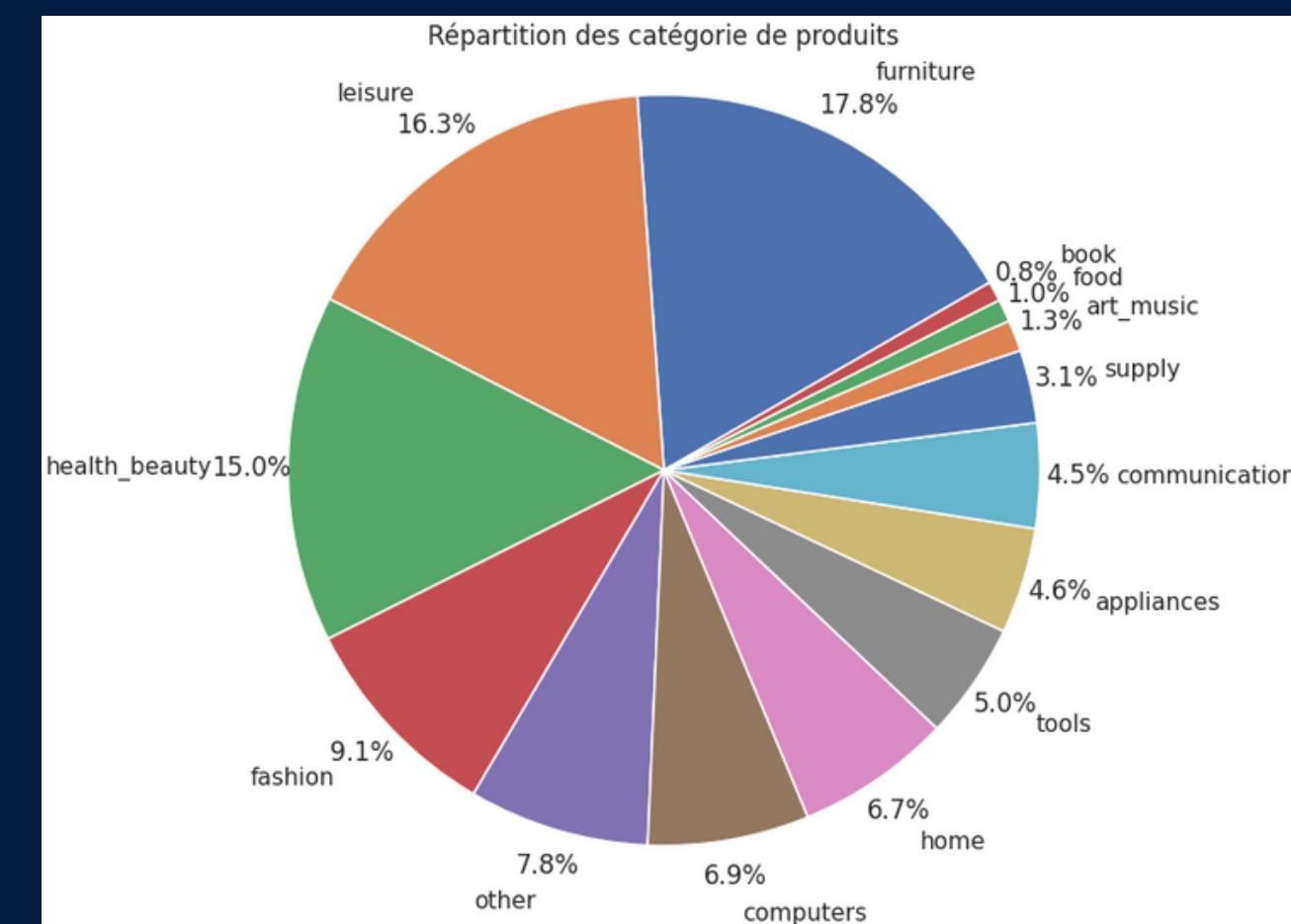
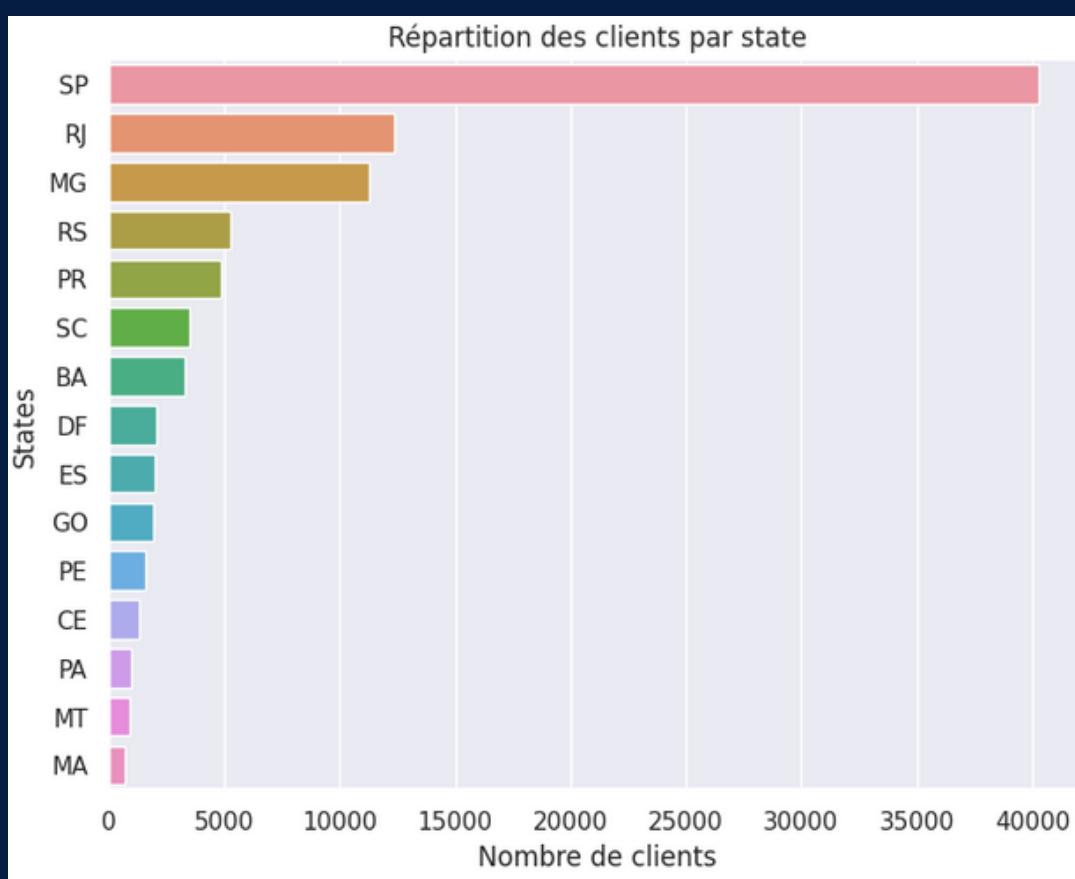
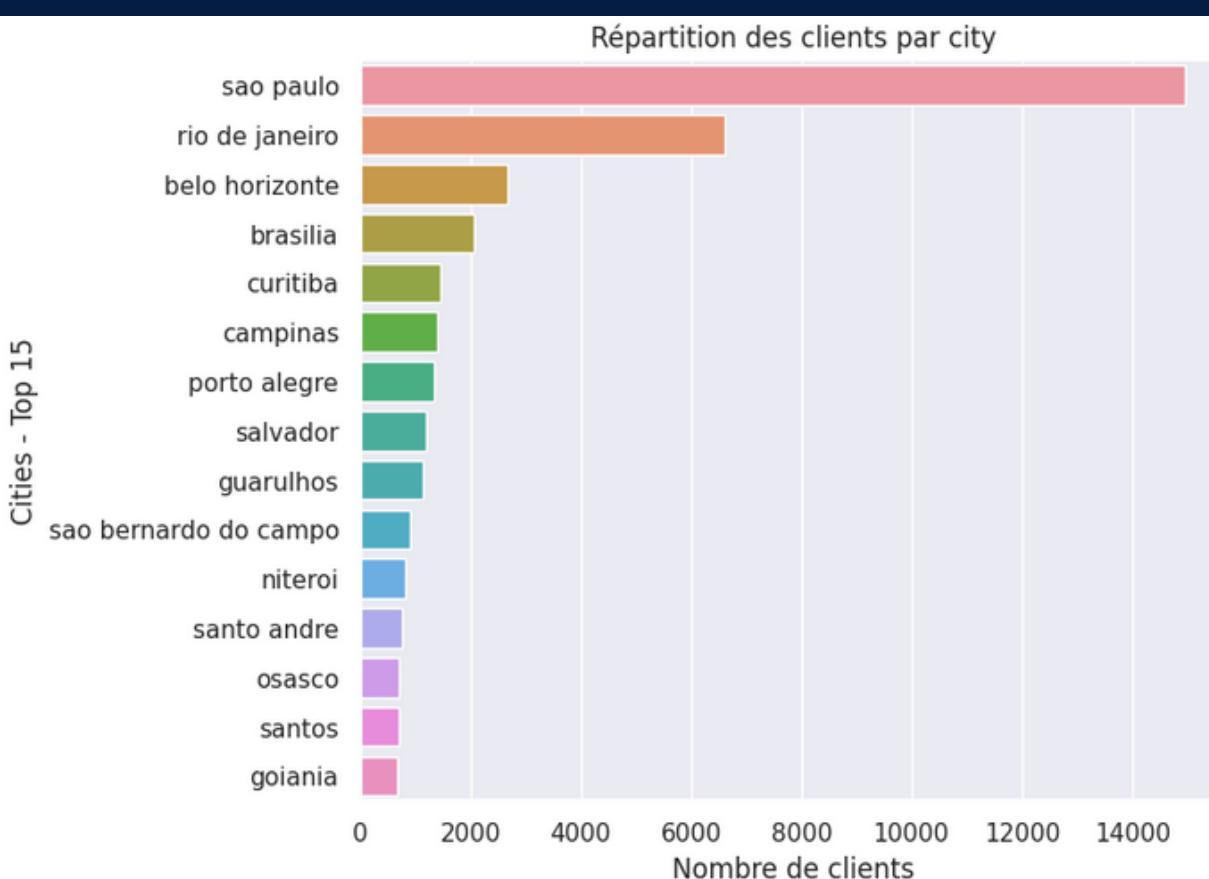


3. Nettoyage et exploration

Exploration

Répartition des clients

Etat



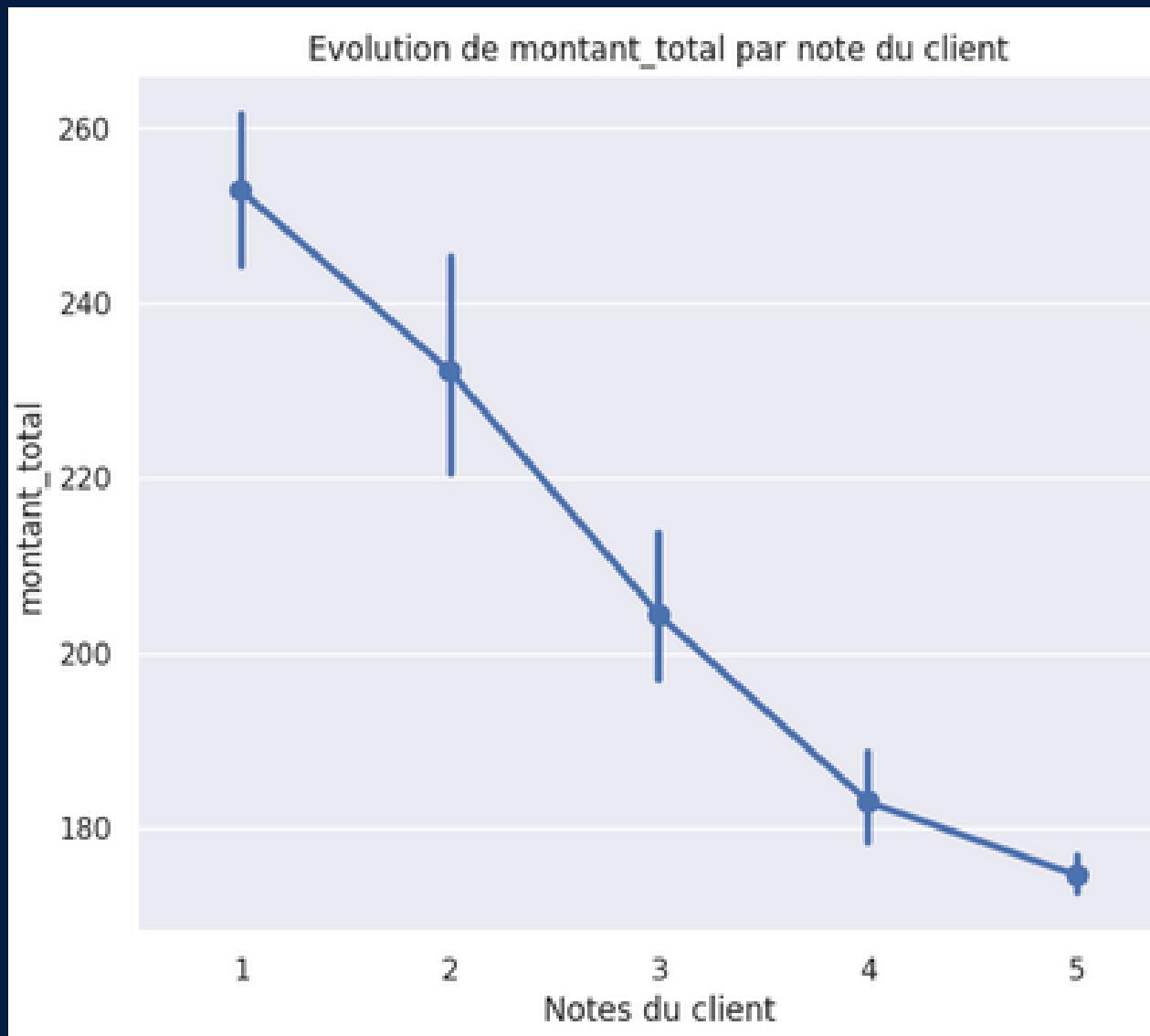


3. Nettoyage et exploration



Exploration

Montant total



Délai de livraison



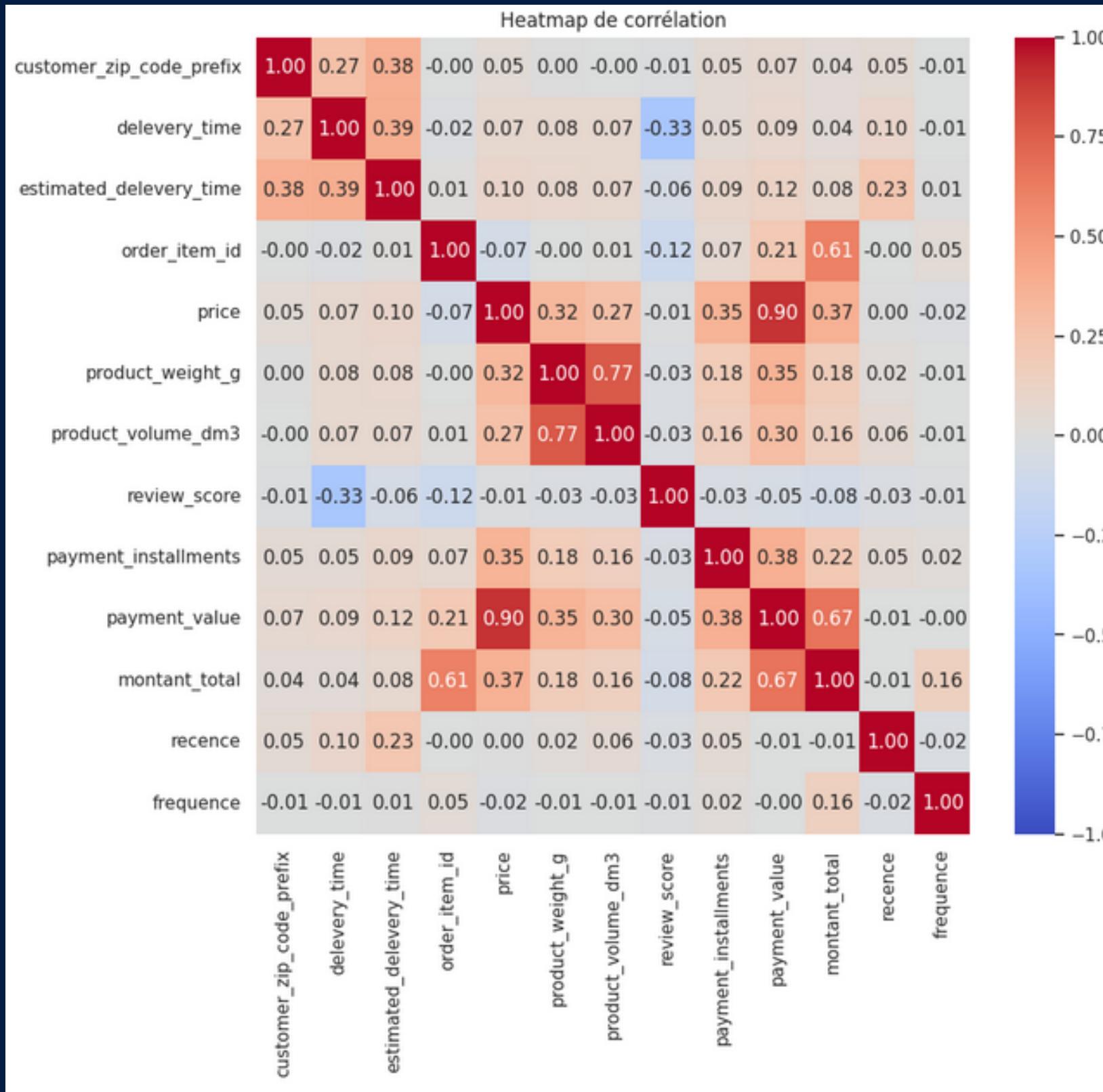
En fonction de la note du client



3. Nettoyage et exploration

Exploration

Corrélation



Fortes corrélation :

- Prix par commande - prix par article
- Montant total - nombre d'article
- poids par produit - volume du produit

Légère corrélation :

- Prix par article - poids du produit

Corrélation négative :

- note client - délai de livraison



4. Modélisation

Méthode :

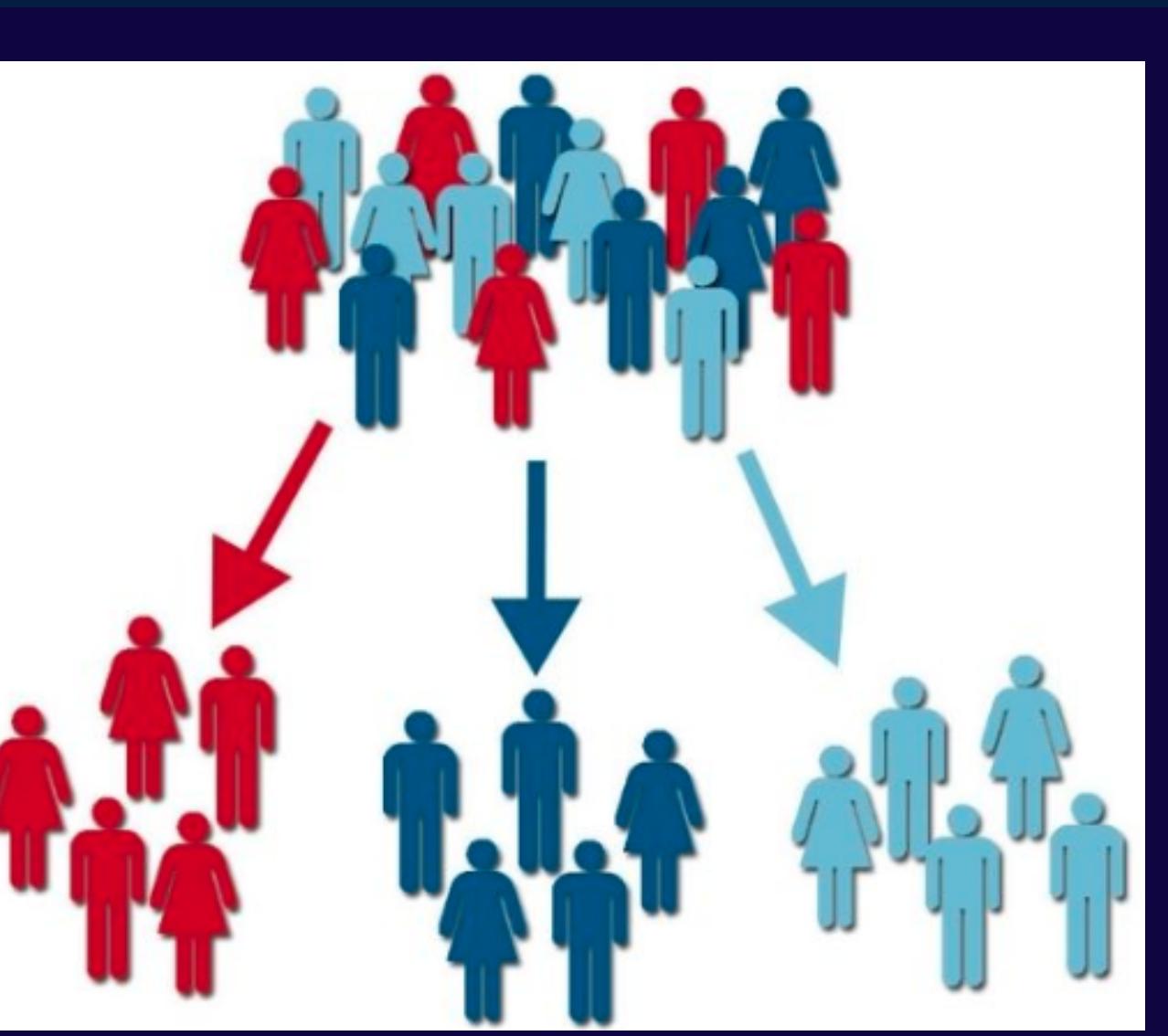
- Clustering
- Algorithmes non supervisés

Evaluation :

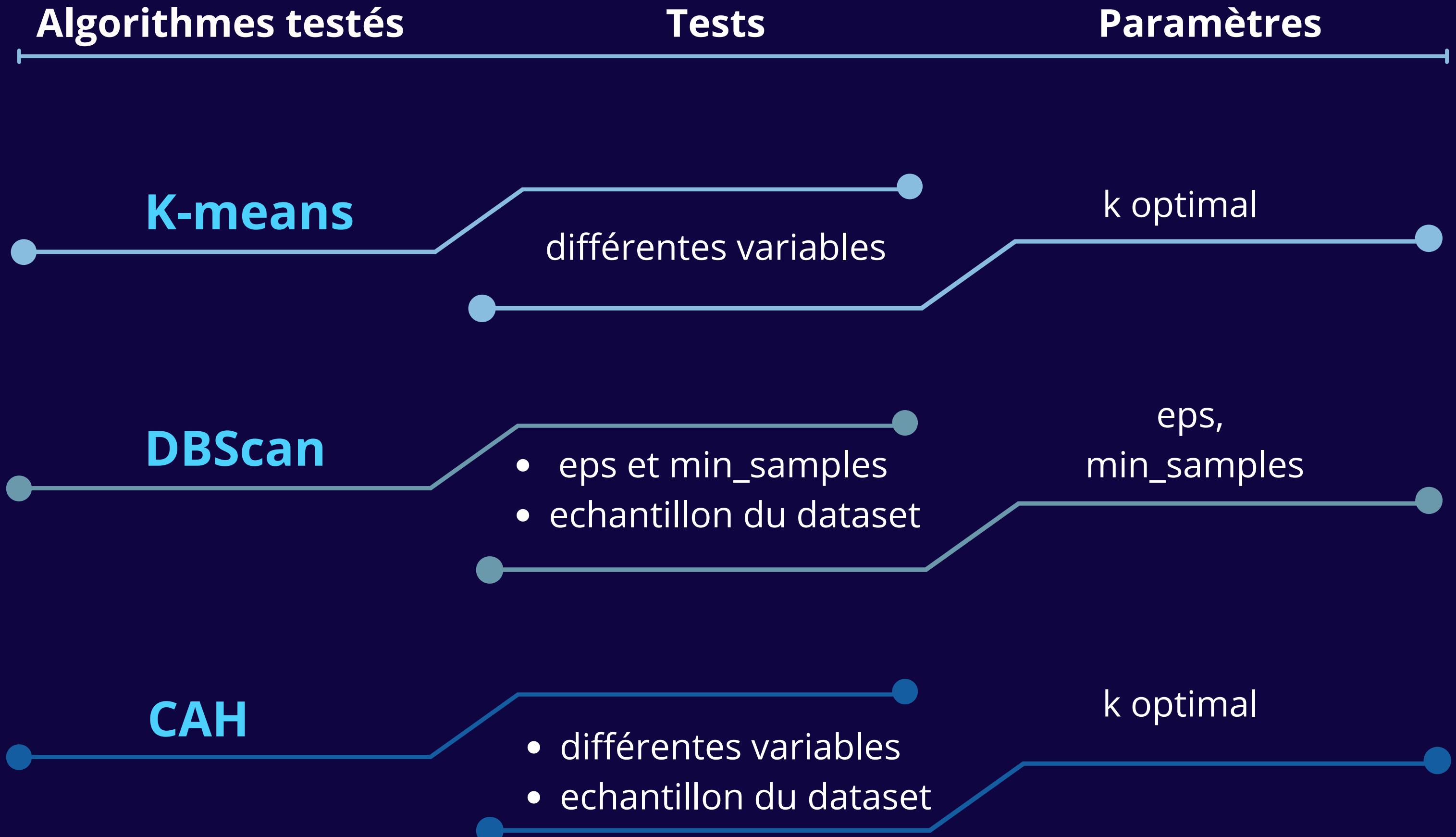
- Coefficient de silhouette
- Indice de Calinski-Harabasz
- Indice de Davies-Bouldin
- Distorsion Elbow

Interprétation

- Graphiques (boxplot, piechart violinplot, pairplot, scatterplot)

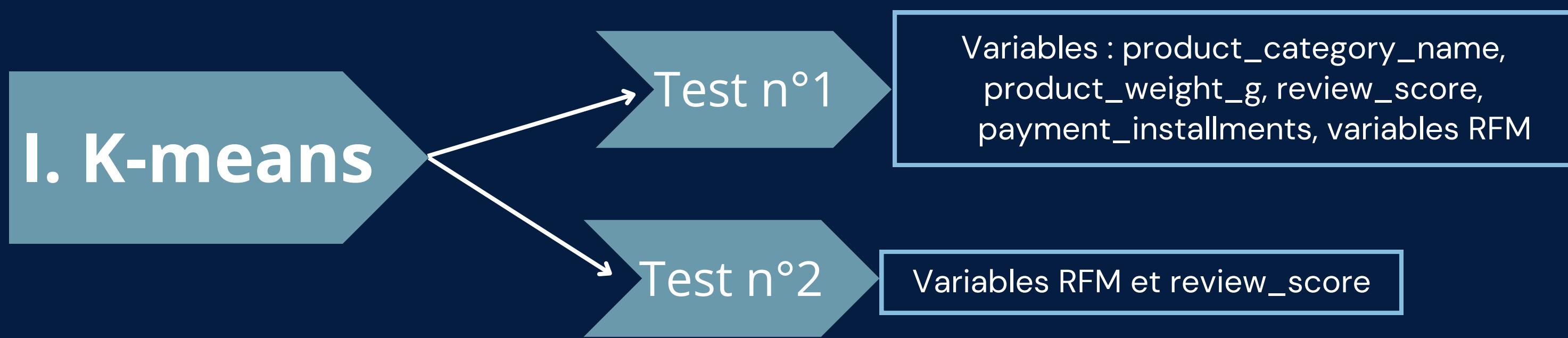


4. Modélisation





4. Modélisation



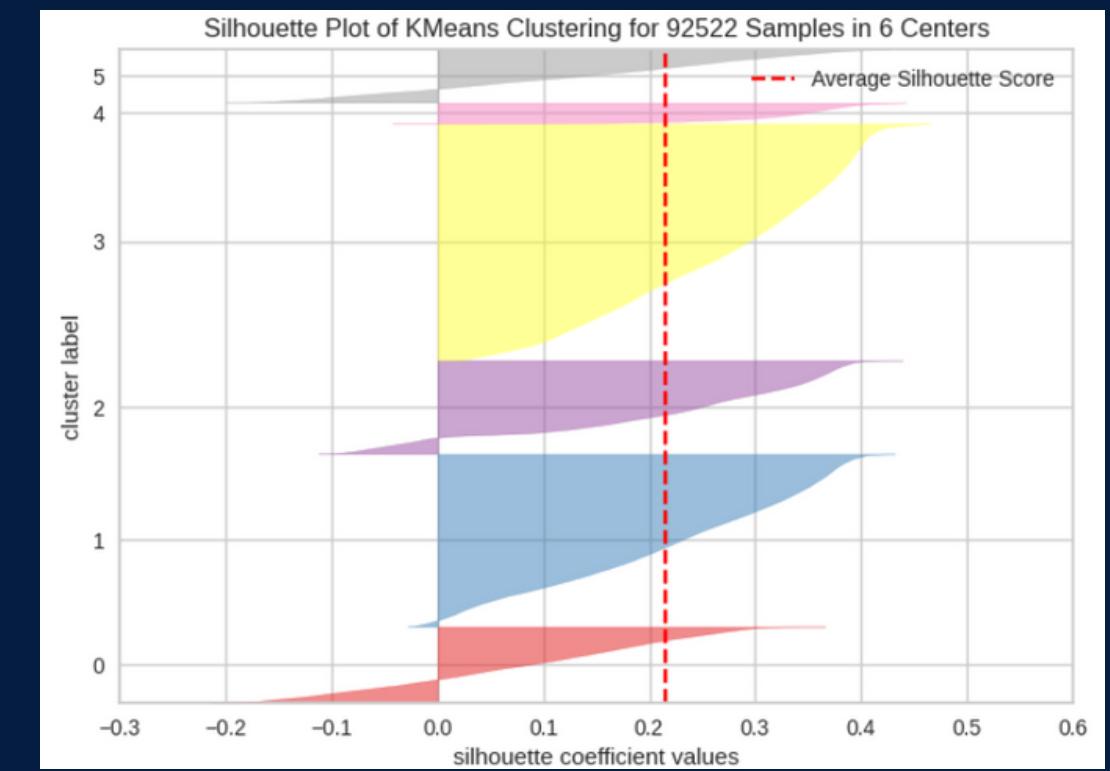
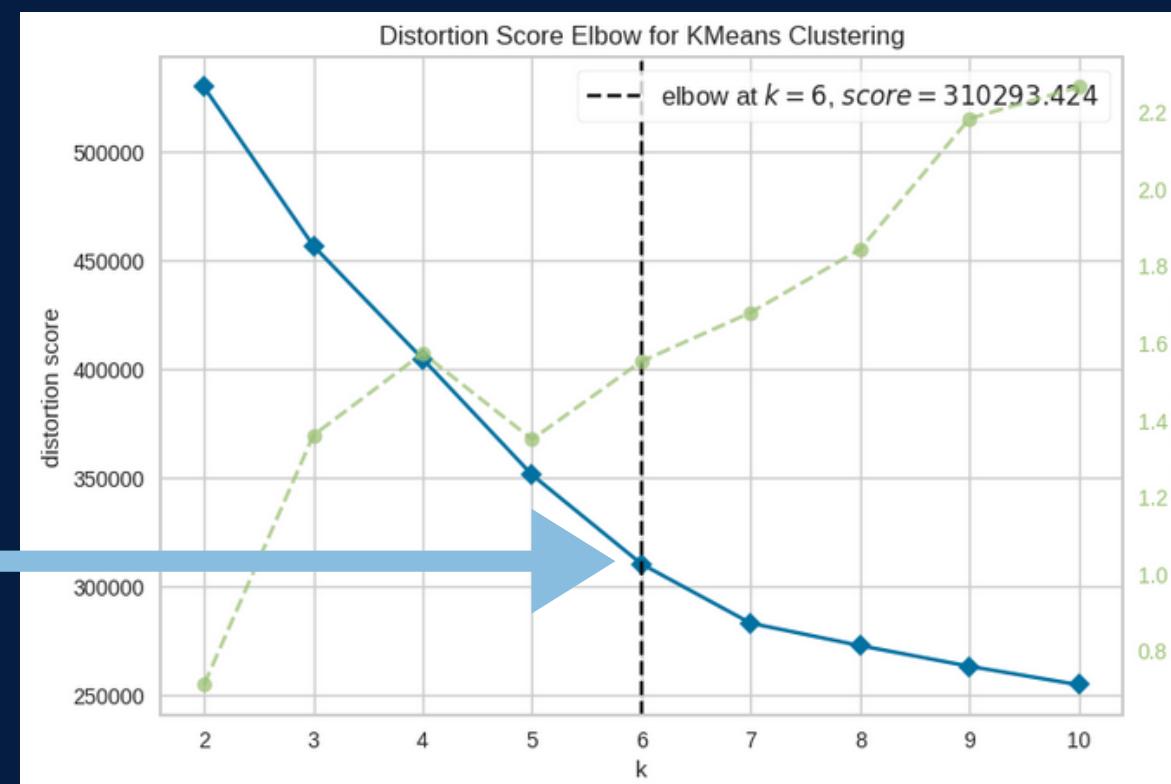


4. Modélisation

K-means

1. Recherche k optimal

$k = 6$

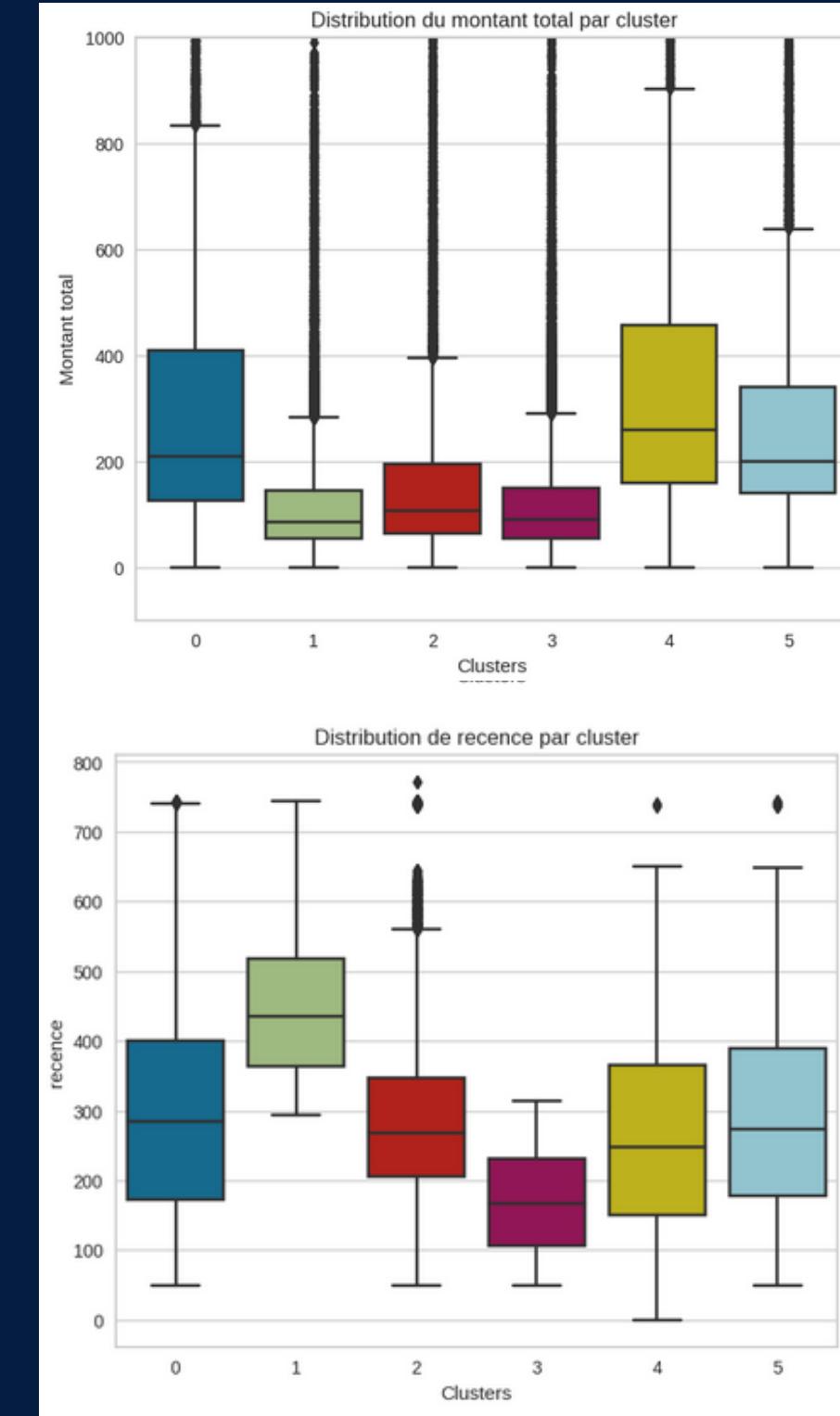
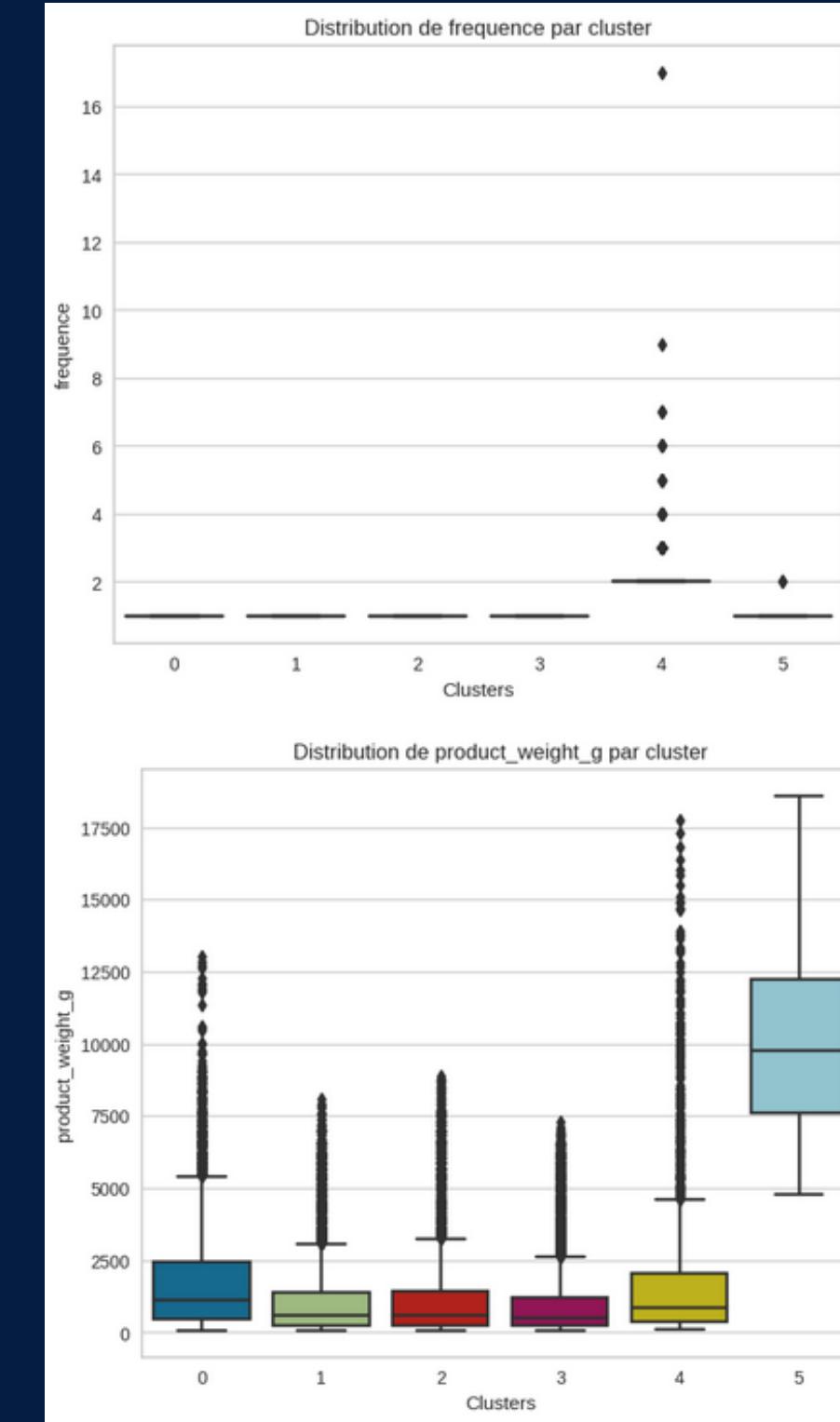
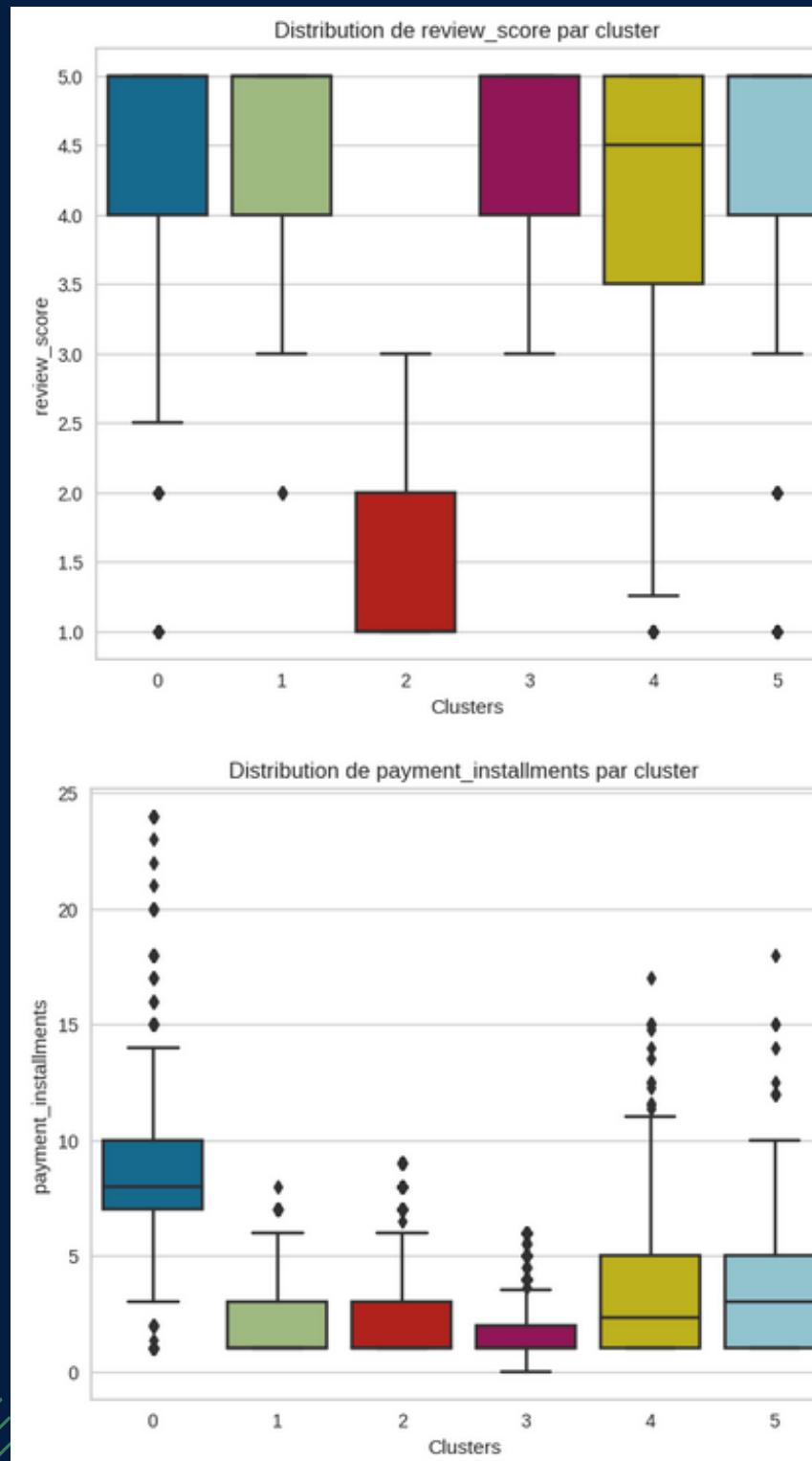




4. Modélisation

K-means

2. Interprétation



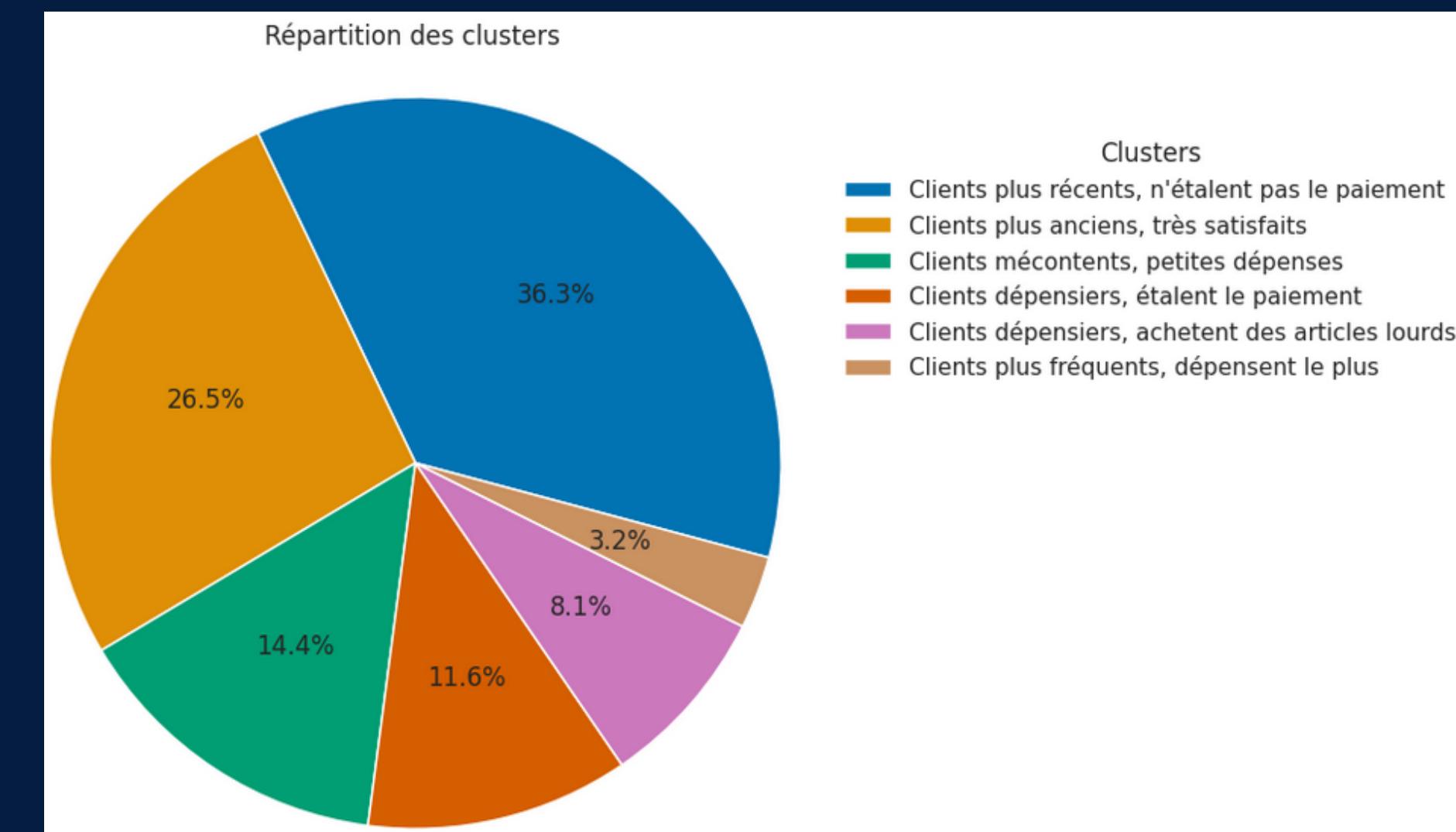
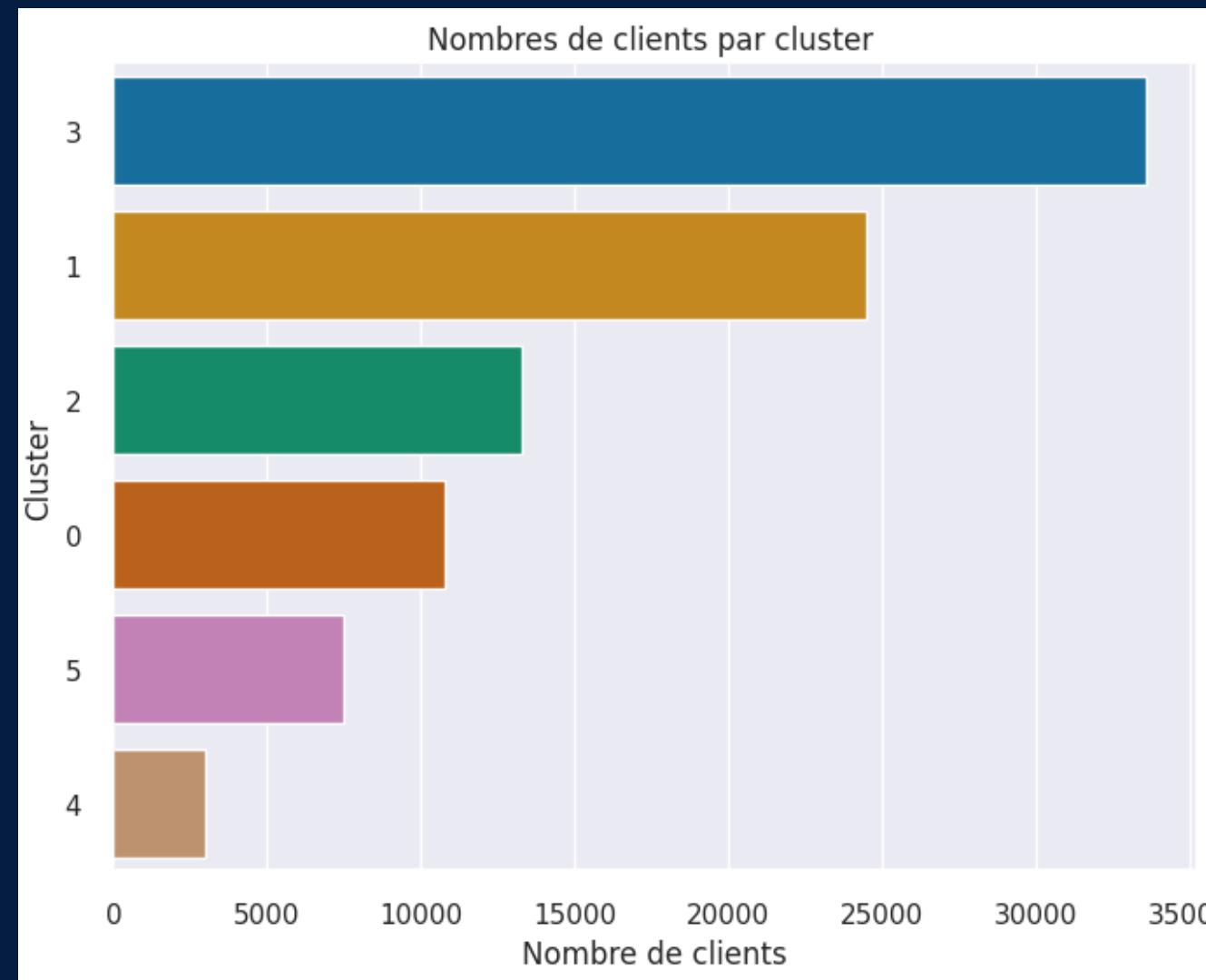


4. Modélisation

Test n°1

K-means

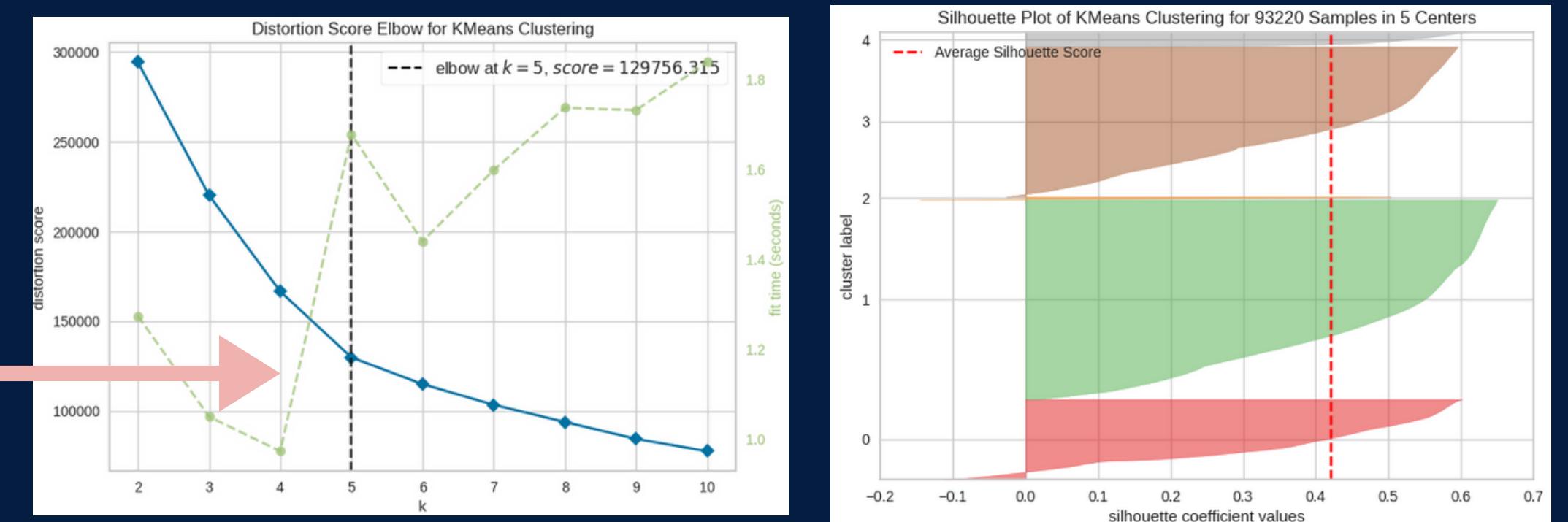
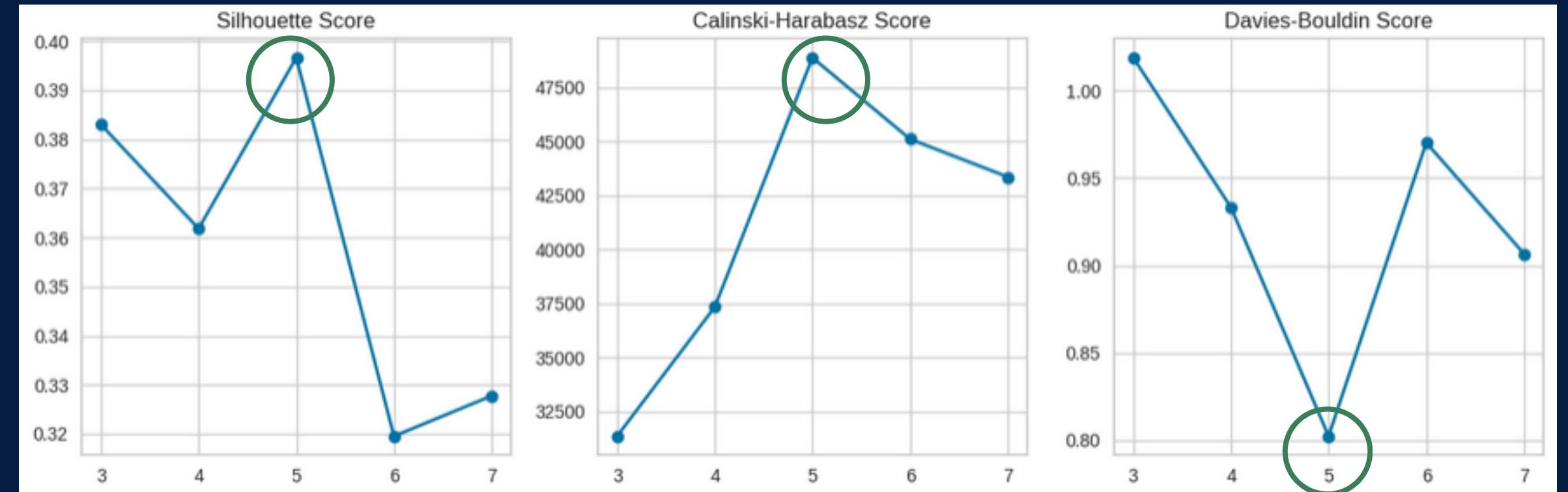
2. Interprétation



4. Modélisation

K-means

1. Recherche k optimal

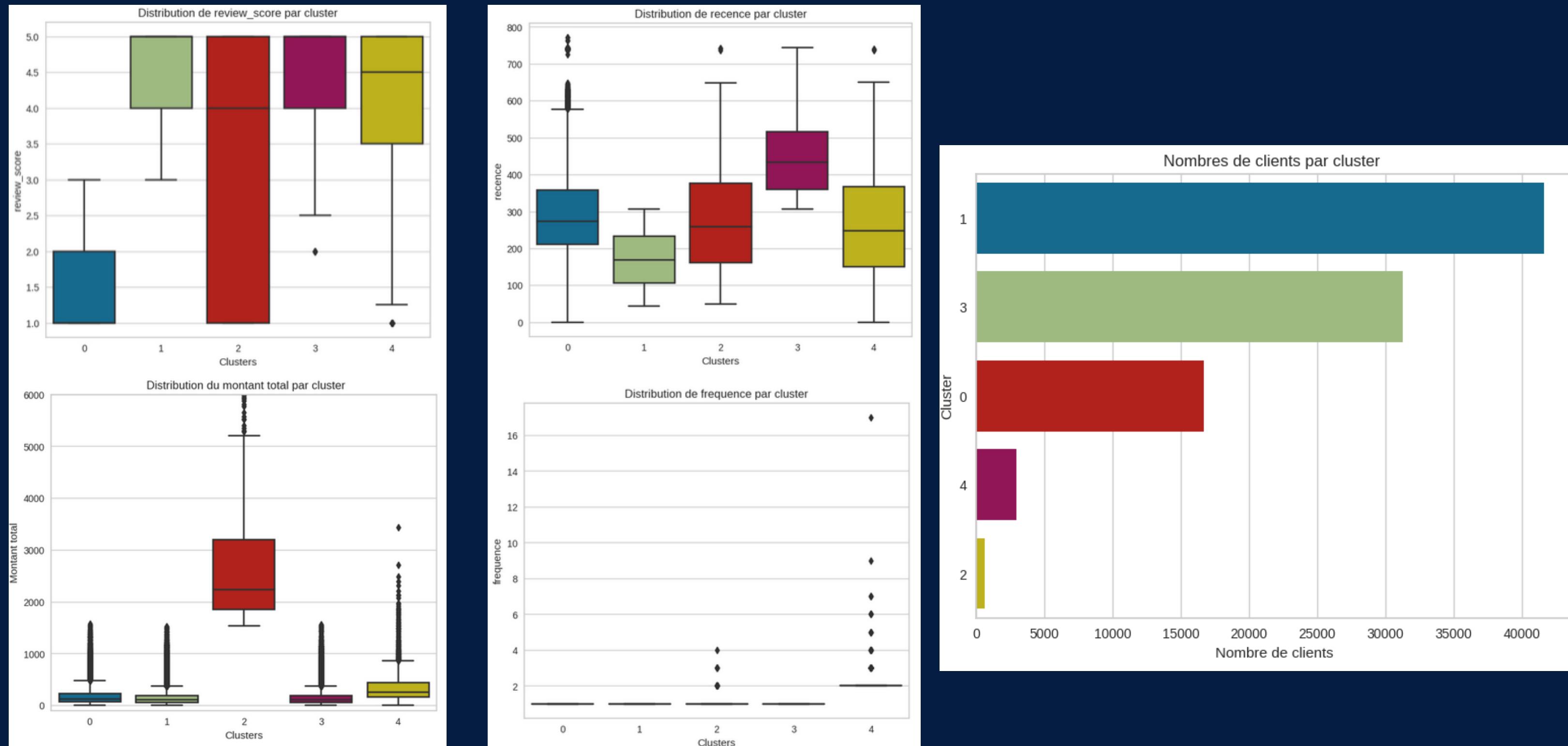


$k = 5$

4. Modélisation

K-means

2. Interprétation



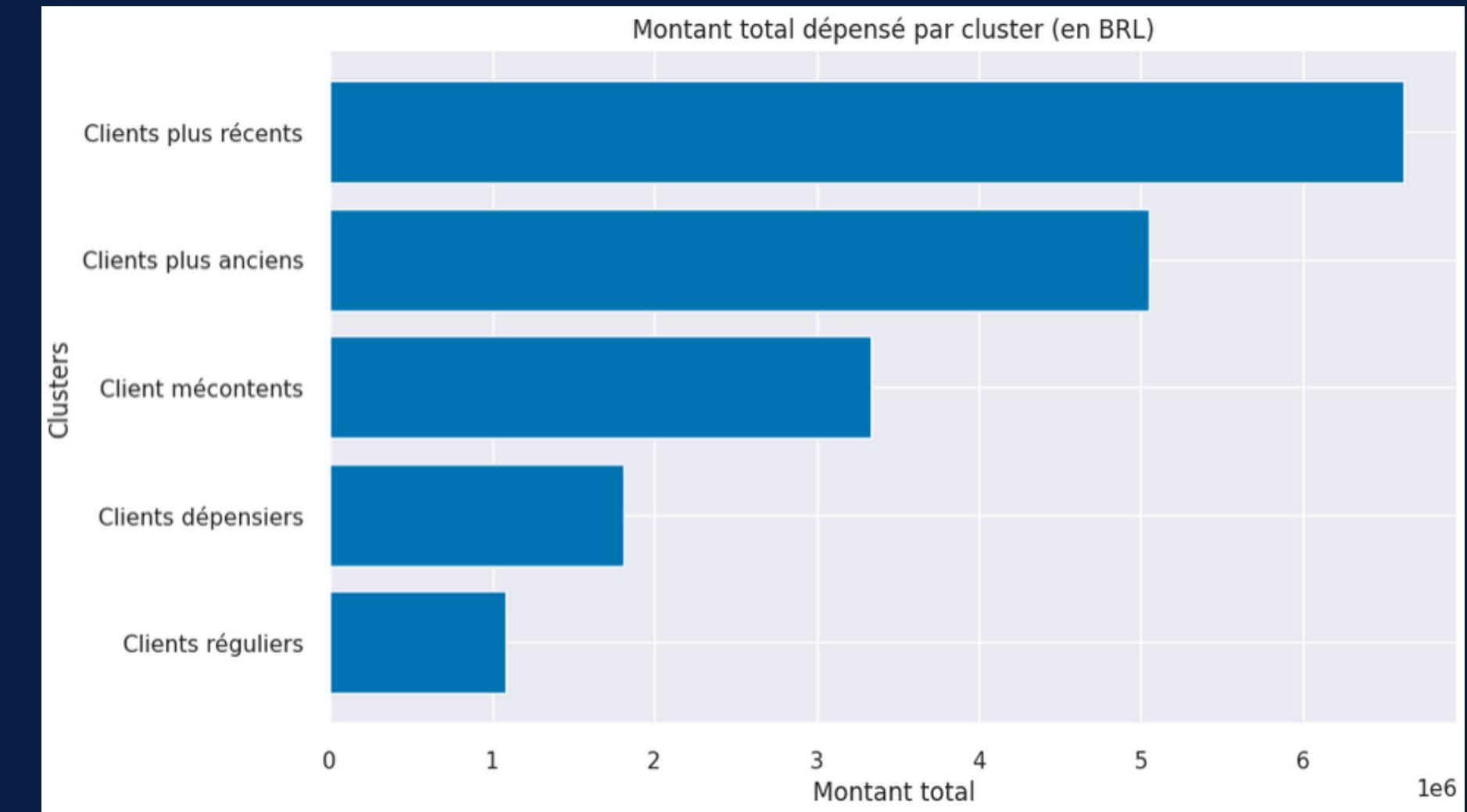
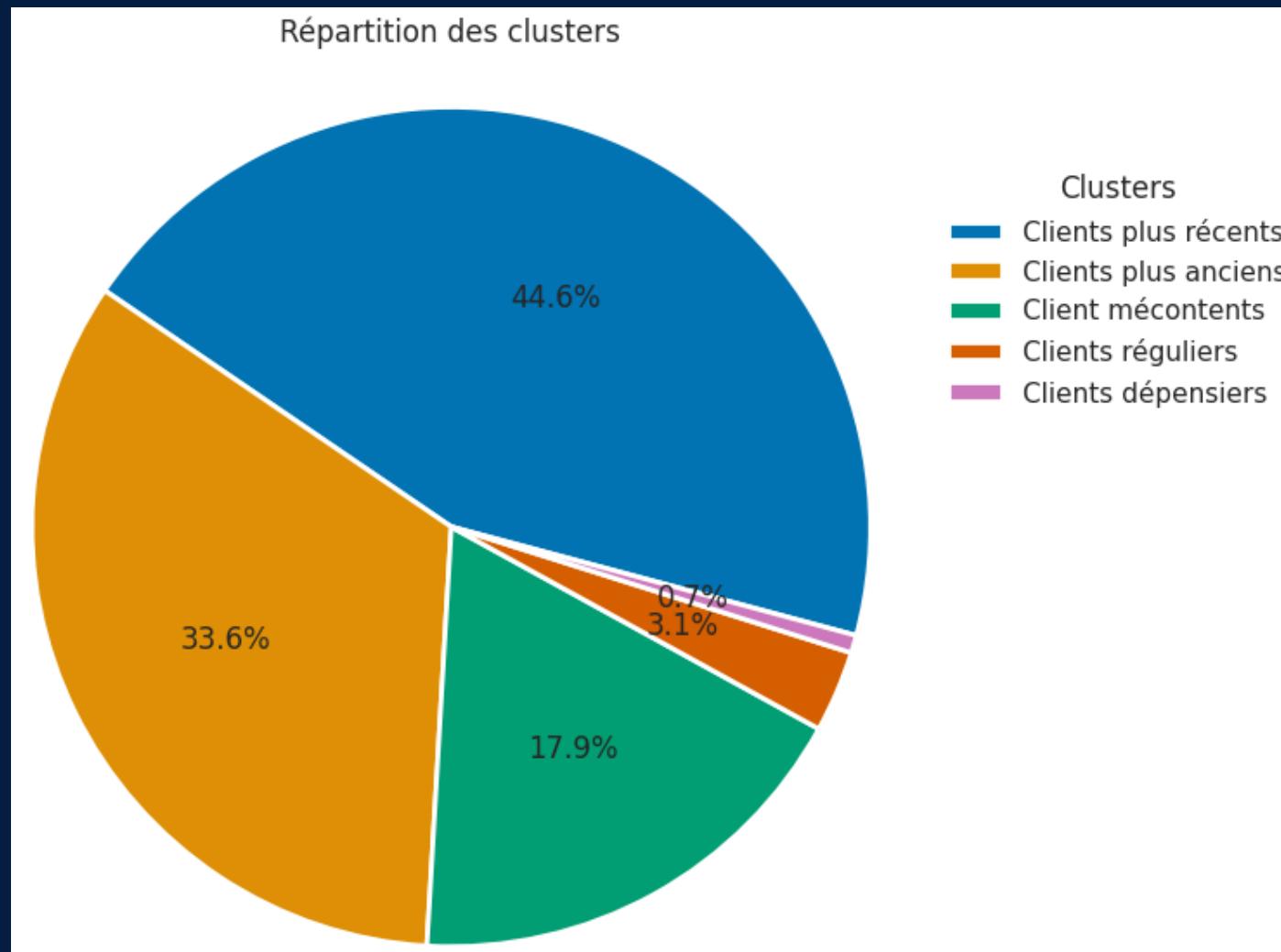


K-means

4. Modélisation

Test n°2

2. Interprétation





4. Modélisation

II. DBScan

Deux paramètres :

- eps : taille du voisinage d'un point
- min_sample : le nombre de voisins dans le voisinage pour être point central.

Plusieurs tests

01

Variation des paramètres
puis calcul du coefficient
de silhouette

02

Echantillon du dataset,
recherche de paramètres
optimaux en maximisant le
coefficient de silhouette

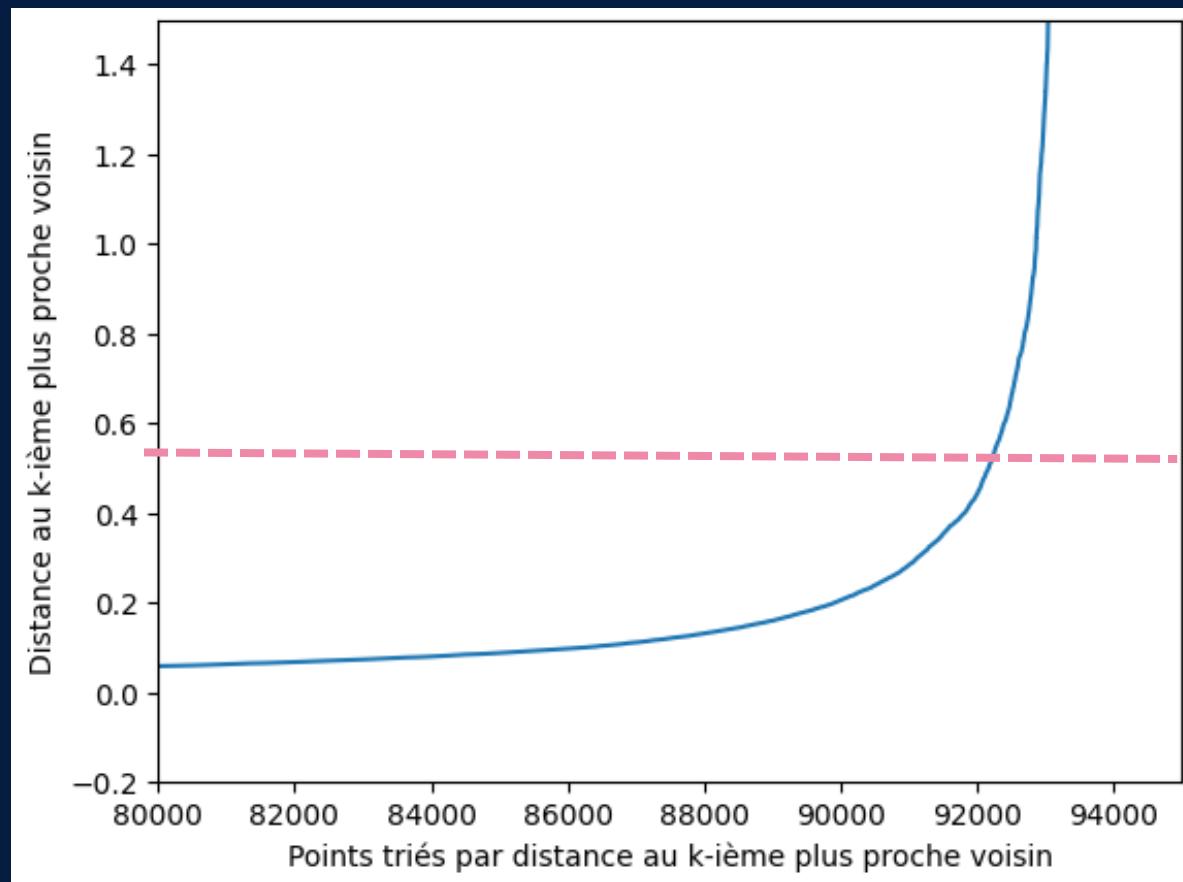


4. Modélisation

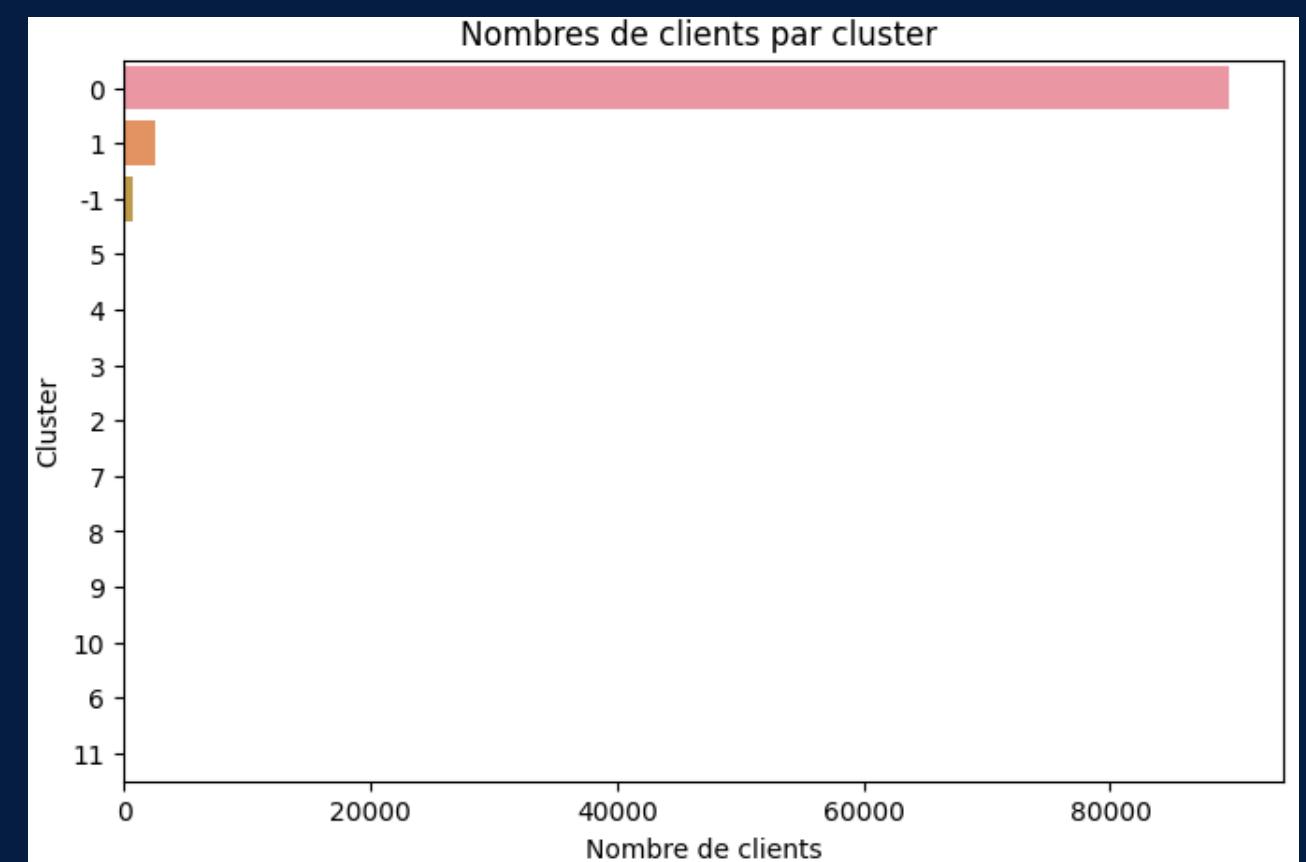
II. DBScan

Test n°1

Distance au k-ième plus proche voisin



13 clusters obtenus



Ici,
eps = 0.5, min_samples = 8

Score de silhouette : 0.282



- Un cluster dominant
- Beaucoup de clusters
- Interprétation difficile

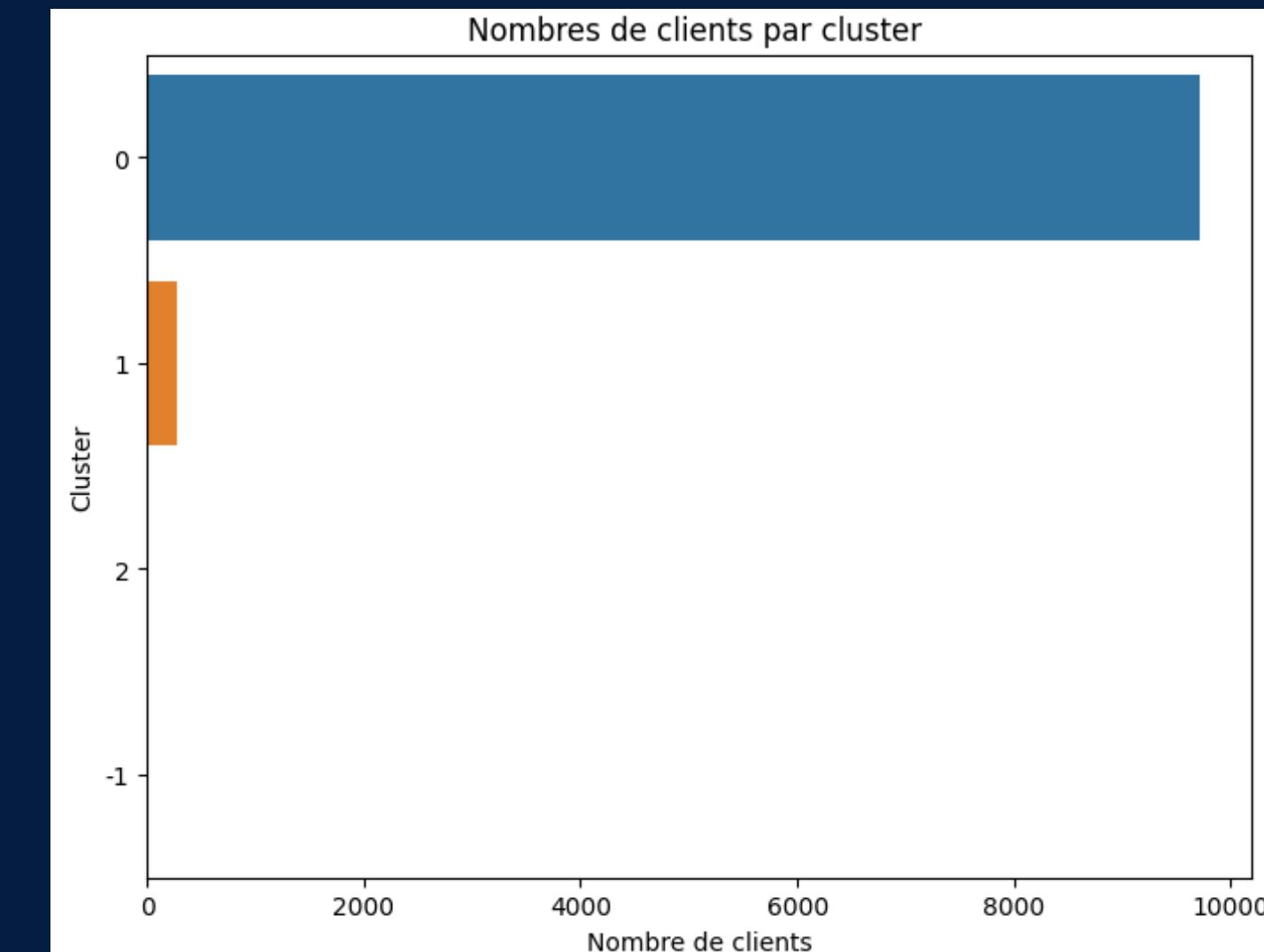
4. Modélisation

II. DBScan - Test n°2

- Échantillon du dataset : 10k lignes
- Recherche de paramètres maximisant le coefficient de silhouette



- $\text{eps} = 3$, $\text{min_samples} = 5$.
- Coefficient de silhouette = 0.587
- nombre de clusters = 4



- Un cluster dominant (plus de 90% des données)
- Interprétation difficile



4. Modélisation

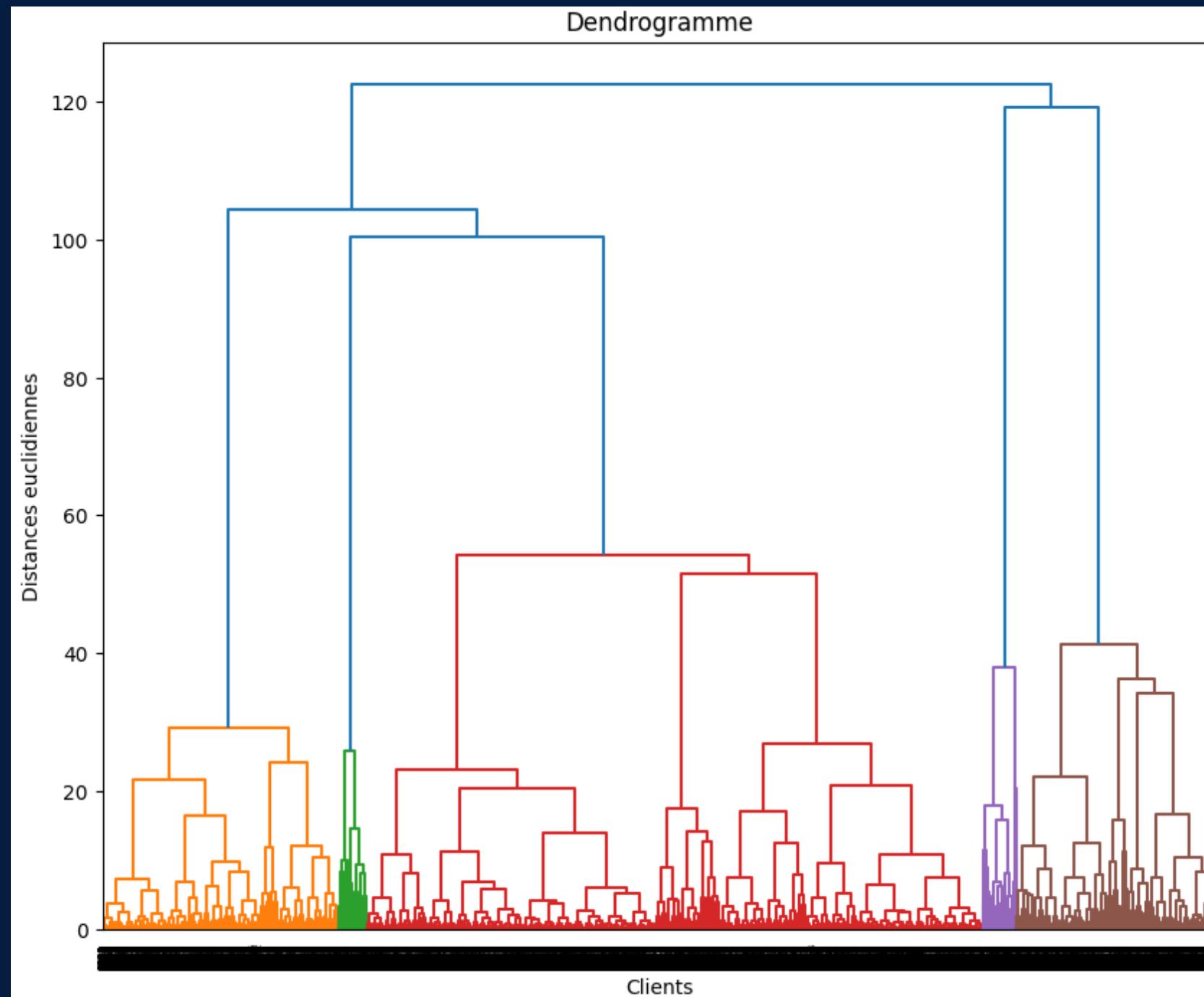
III. Classification ascendante hiérarchique

- 01 Échantillon du dataset 10K observations
- 02 Chercher K optimal (coefficient de silhouette, Calinski, Davies)
- 03 Interpréter les cluster

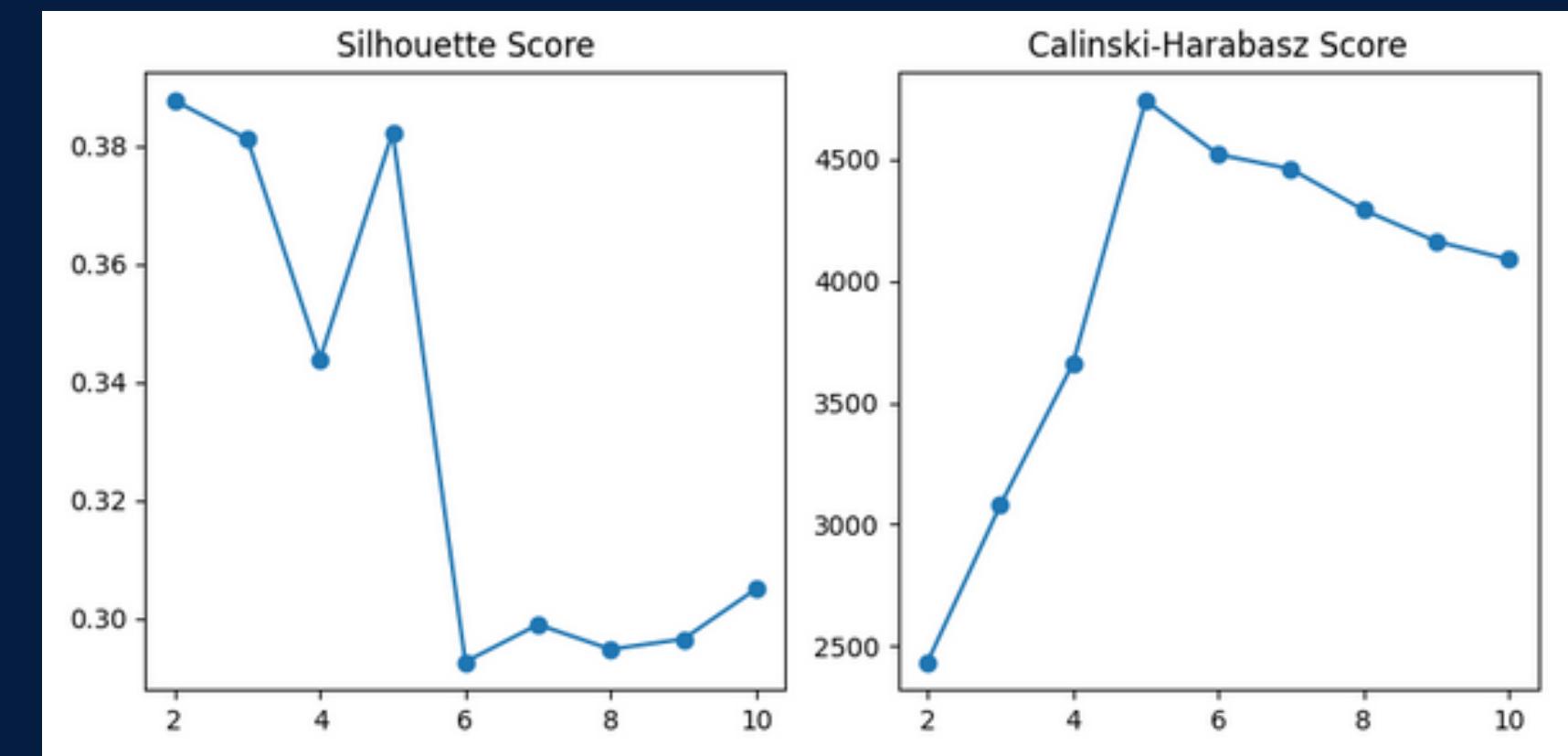


4. Modélisation

III. CAH



K optimal = 5



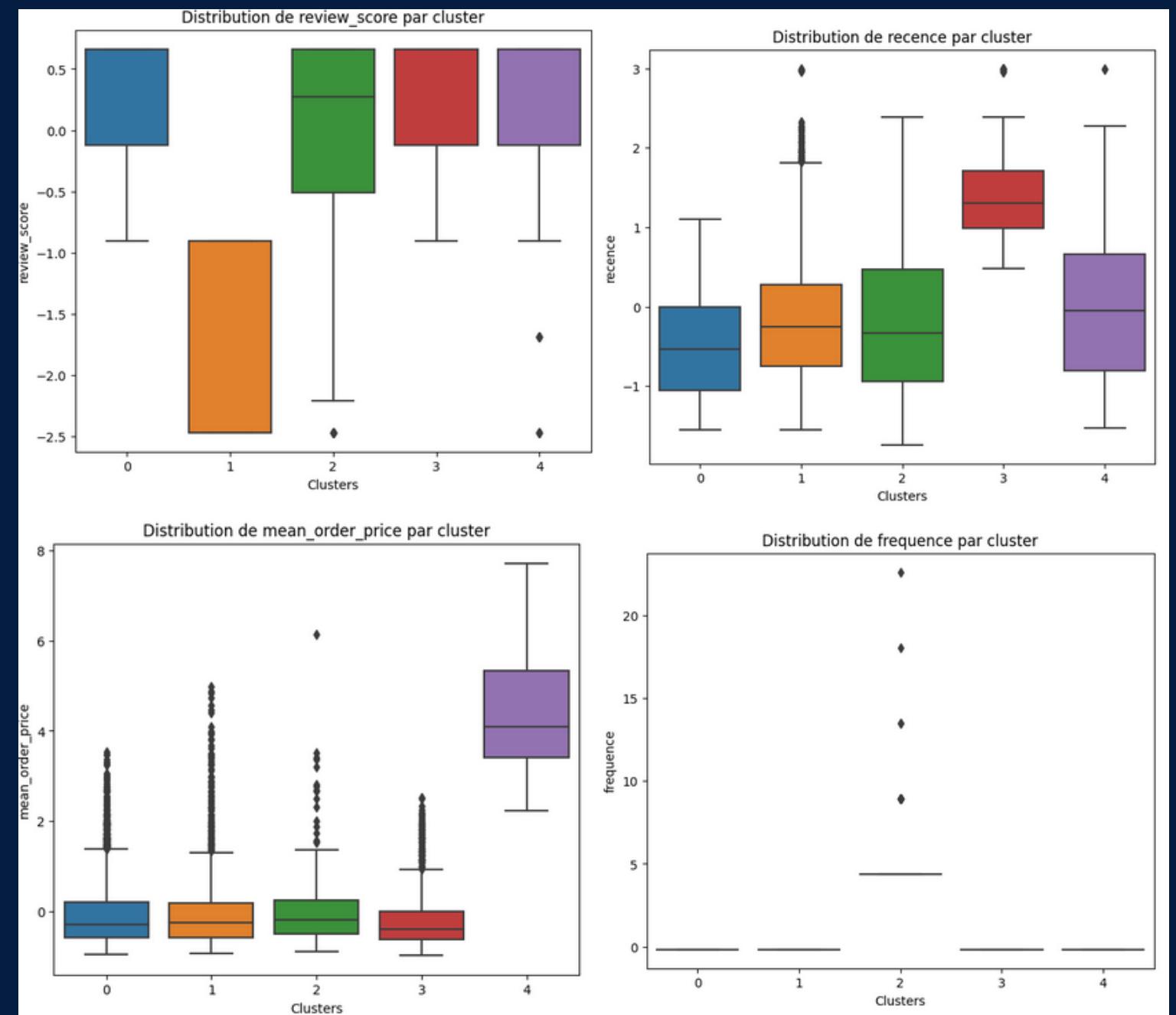
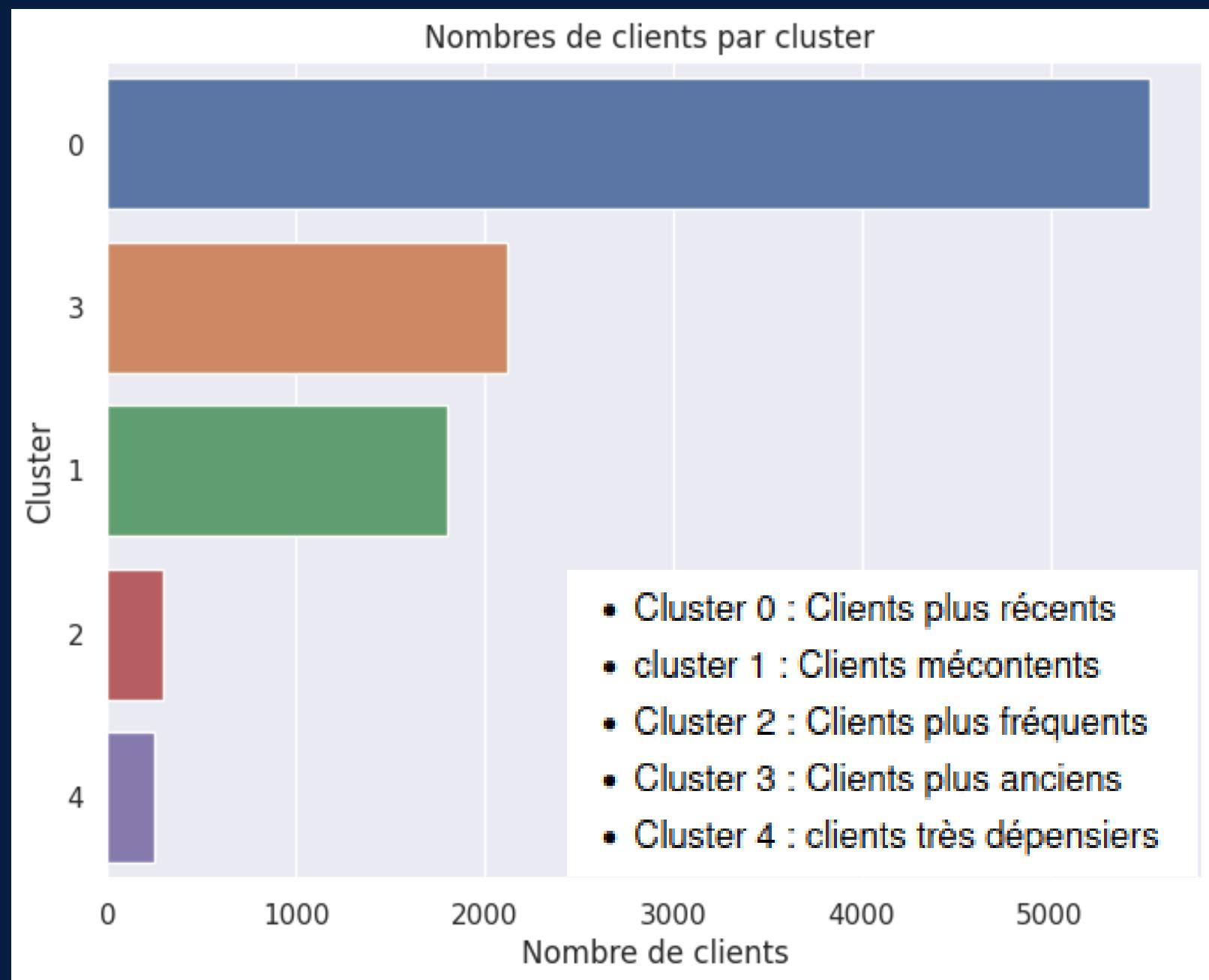
metric = euclidean, method = ward



4. Modélisation

III. CAH

Interprétation





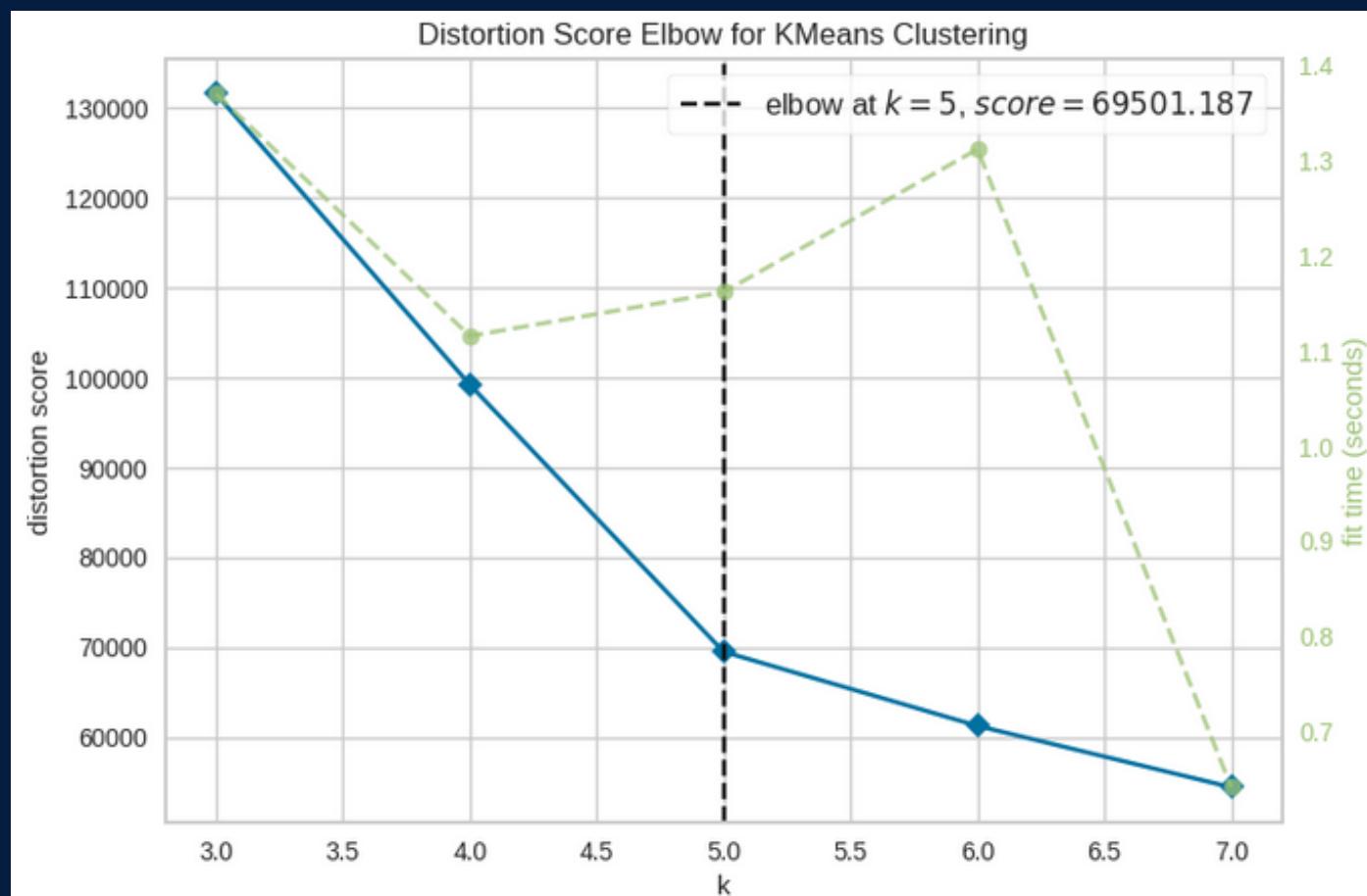
5. Période de maintenance

Etapes d'analyse de la stabilité temporelle

1. Définir un premier modèle basé sur une période de référence
2. Comparer le modèles aux modèles entraînés en ajoutant les données progressivement (+ n mois, n semaines) en utilisant le score ARI.
3. En déduire graphiquement la période de maintenance
4. Refaire une simulation avec :
période de référence = date_dernier_achat - période de maintenance

5. Période de maintenance

Modèle choisi, période de référence



Modèle étudié :
KMeans(`n_clusters = 5, init='k-means++', random_state=21`)

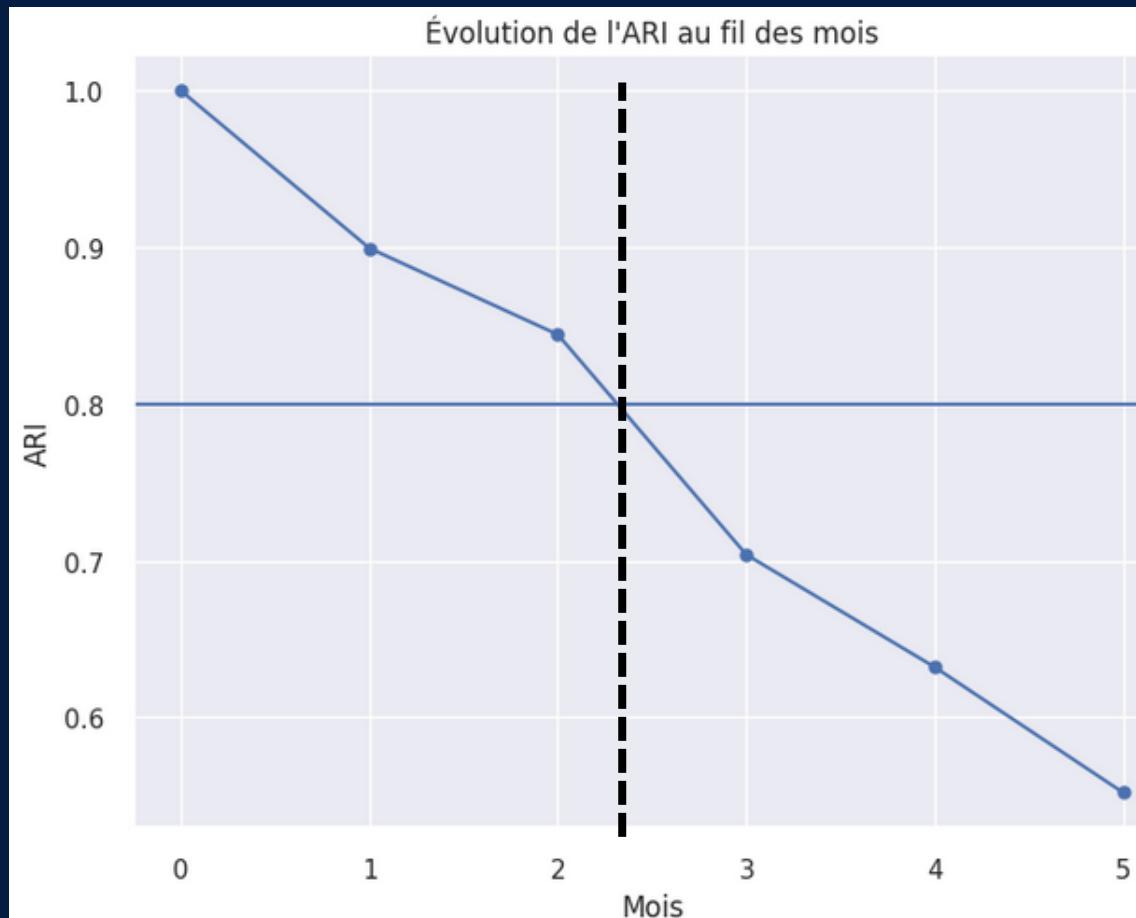
Période de référence :
1 an et 6 mois de données



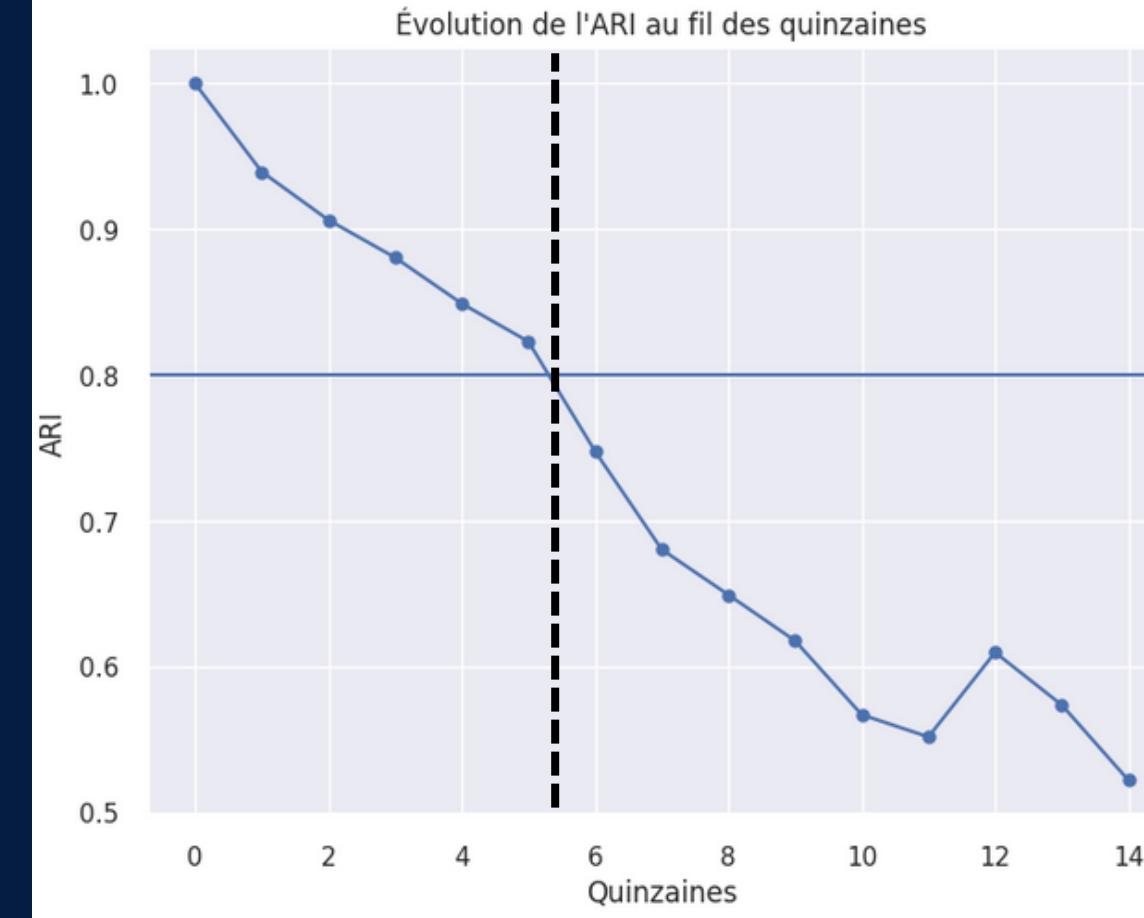
5. Période de maintenance

Première simulation

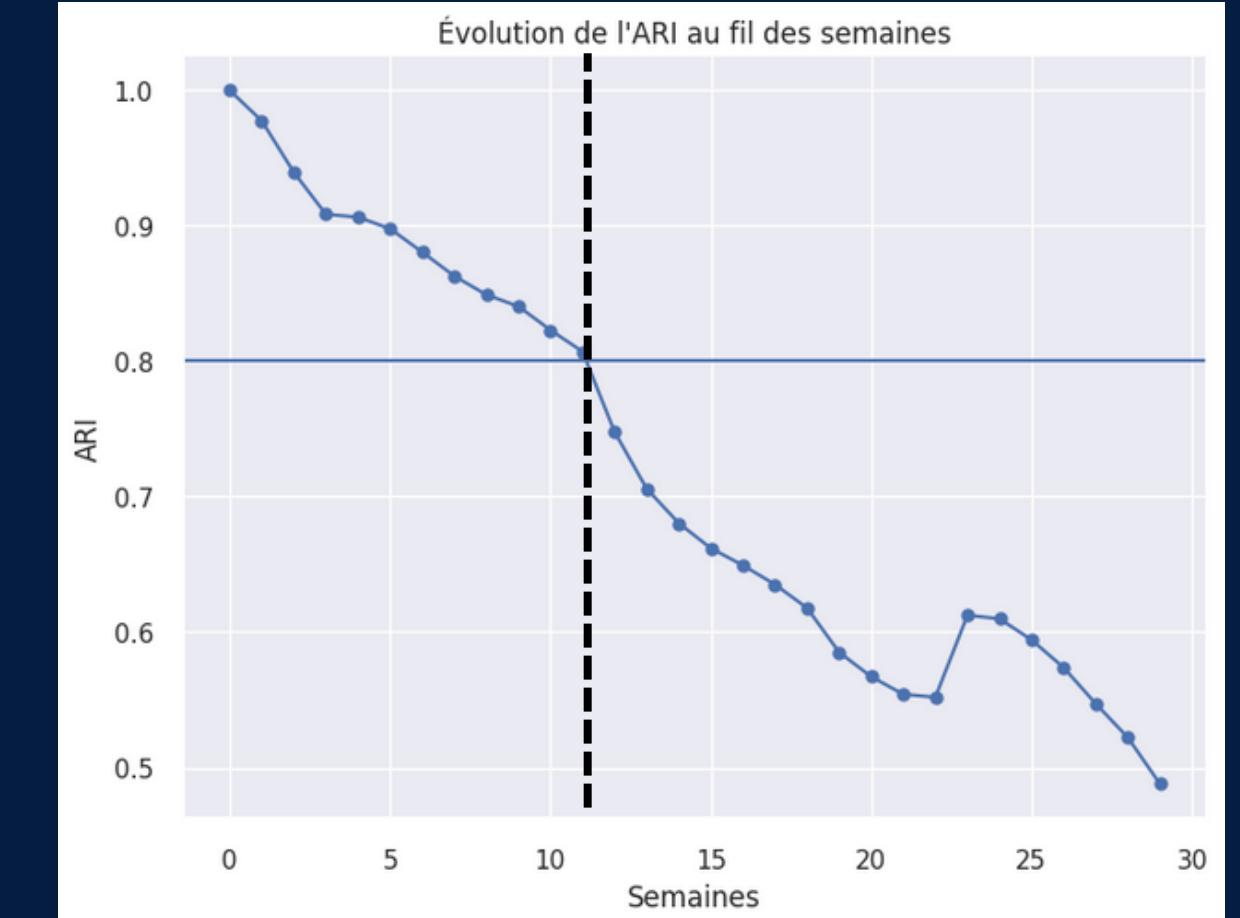
Période de référence + n mois



Période de référence + n quinzaines



Période de référence + n semaines



Période de maintenance : 11 semaines

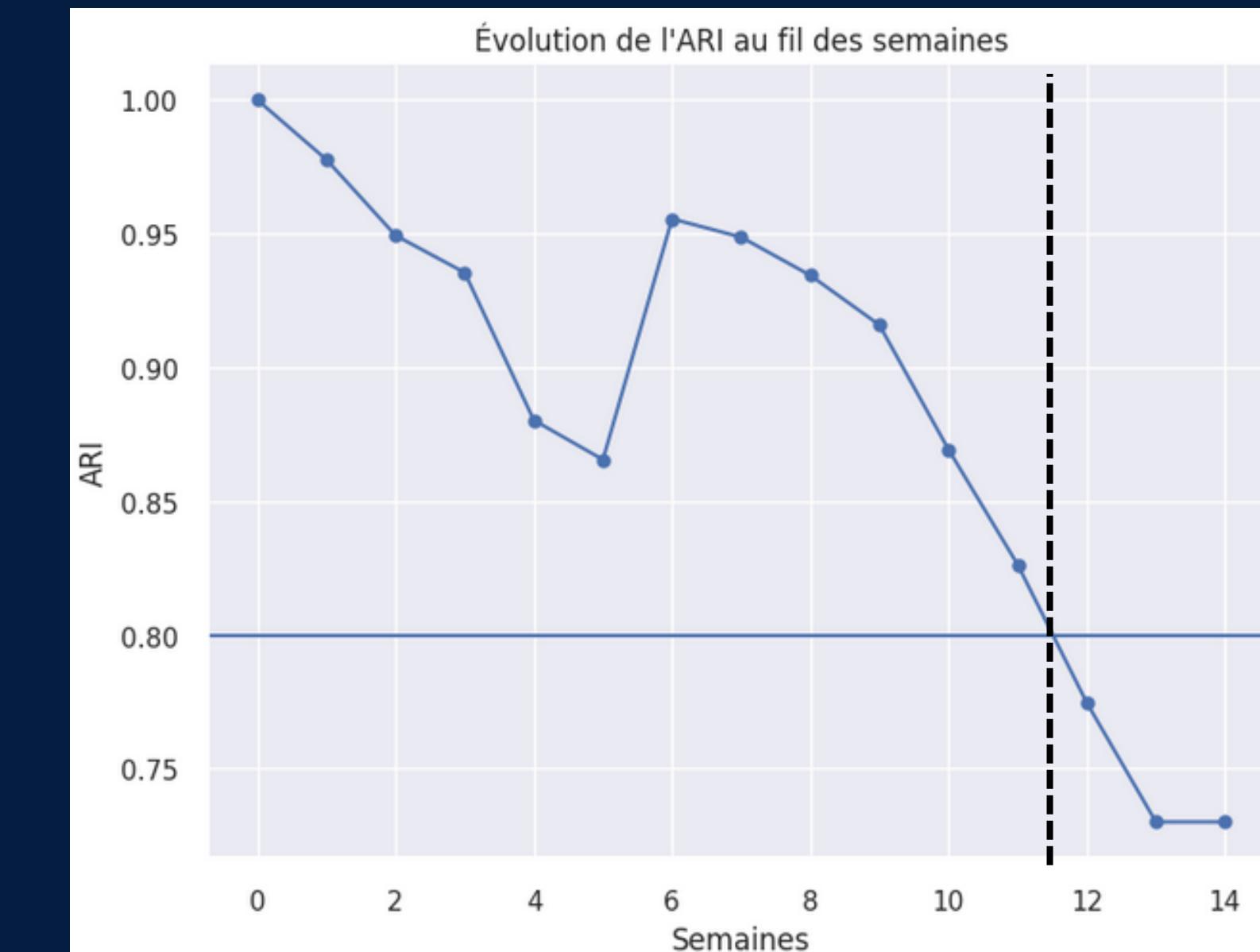
5. Période de maintenance

Deuxième simulation

Période de référence :
1 an et 10 mois de données

Période de maintenance conseillée :
11 semaine

Période de référence + n semaines





6. Conclusion

- Algorithmes d'apprentissage non supervisés
- Modèle : K-means, variable RFM et review_score
- Segmentation proposée : 5 clusters
- Période de maintenance : 11 semaines
- Pistes d'amélioration :
 - Un plus grand nombres de données
 - la fréquence des clients plus exploitable
 - Des informations supplémentaires sur les clients
(âge, sexe,...)

