

# Segmentez des clients d'un site e-commerce



par Ethan VUILLEMIN



# Contexte & Objectifs

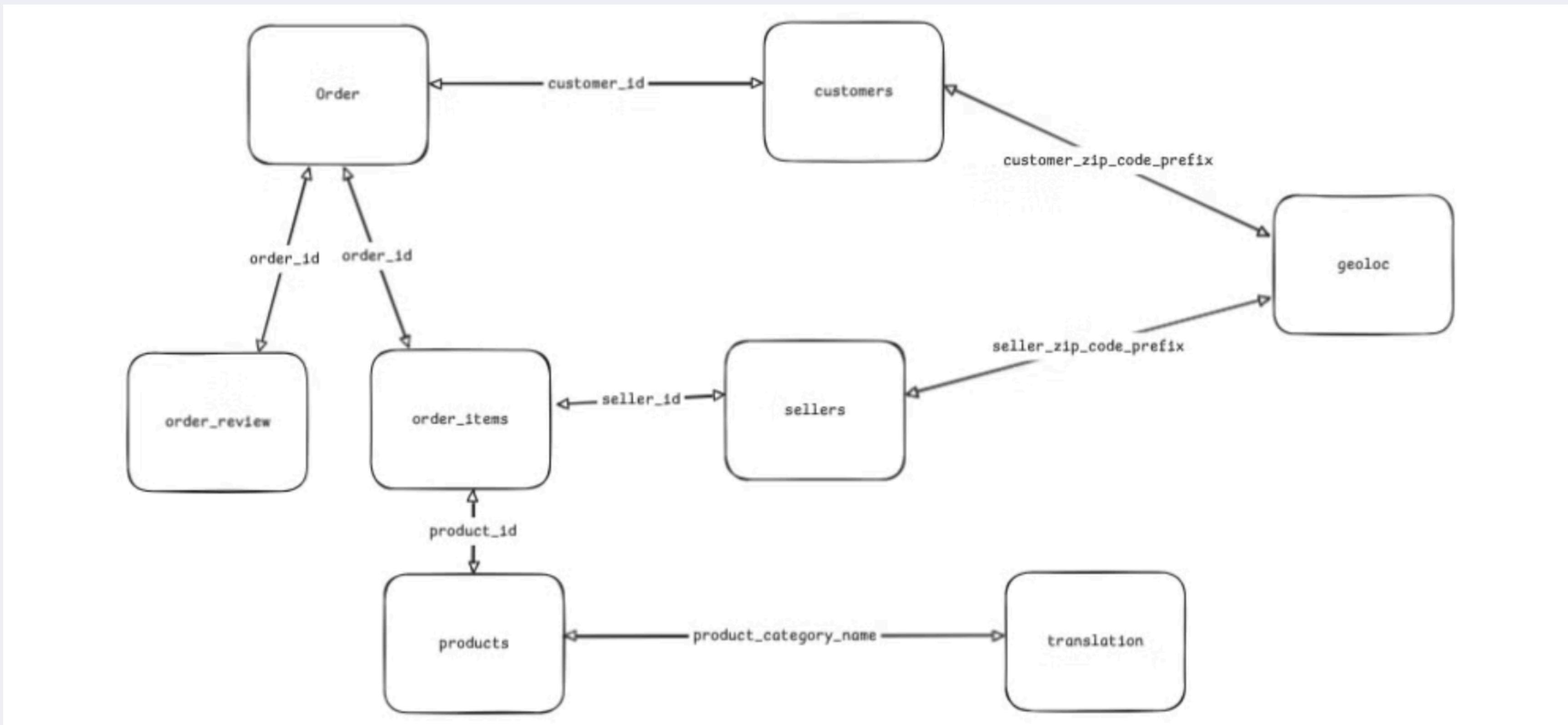
## Contexte

Dans un marché e-commerce compétitif, **Olist** cherche à mieux comprendre ses clients pour **améliorer ses services** et **augmenter sa fidélisation**. En analysant les données clients, l'entreprise vise à **créer des segments pertinents** pour des actions marketing ciblées..

## Objectif

**Segmenter les clients** d'Olist à l'aide d'algorithmes de clustering, permettant ainsi une **personnalisation** des offres et une **amélioration** de **l'expérience client**.

# Structure des



# Requêtes SQL



## Commandes récentes en retard

Commandes de moins de 3 mois, reçues avec plus de 3 jours de retard (hors commandes annulées).

## Meilleurs vendeurs Olist

Vendeurs ayant généré plus de 100 000 Real de chiffre d'affaires.

## Nouveaux vendeurs engagés

Nouveaux vendeurs (moins de 3 mois) ayant vendu plus de 30 produits.

## Pires codes postaux (reviews)

5 codes postaux (plus de 30 reviews) avec le pire score moyen sur 12 mois.

# Requêtes SQL

**257**

## Commandes récentes en retard

Commandes de moins de 3 mois, reçues avec plus de 3 jours de retard (hors commandes annulées).

**17**

## Meilleurs vendeurs Olist

Vendeurs ayant généré plus de 100 000 Real de chiffre d'affaires.

**71**

## Nouveaux vendeurs engagés

Nouveaux vendeurs (moins de 3 mois) ayant vendu plus de 30 produits.

**5**

## Pires codes postaux (reviews)

5 codes postaux (plus de 30 reviews) avec le pire score moyen sur 12 mois.

# Aperçu des fonctions utilisées

- Select, GroupBy, Join, OrderBy, WithAs

```
-- Utilisation de WITH AS pour créer une vue temporaire nommée filtered_reviews
WITH filtered_reviews AS (
    -- Sélection des colonnes nécessaires et calcul des agrégats
    SELECT
        customers.customer_zip_code_prefix,
        AVG(order_reviews.review_score) AS avg_review_score,
        COUNT(order_reviews.review_id) AS review_count
```



# Exploration des données (EDA)

## Objectifs

Comprendre les caractéristiques des clients d'Olist et identifier des patterns pour éclairer la segmentation.

## Méthodes

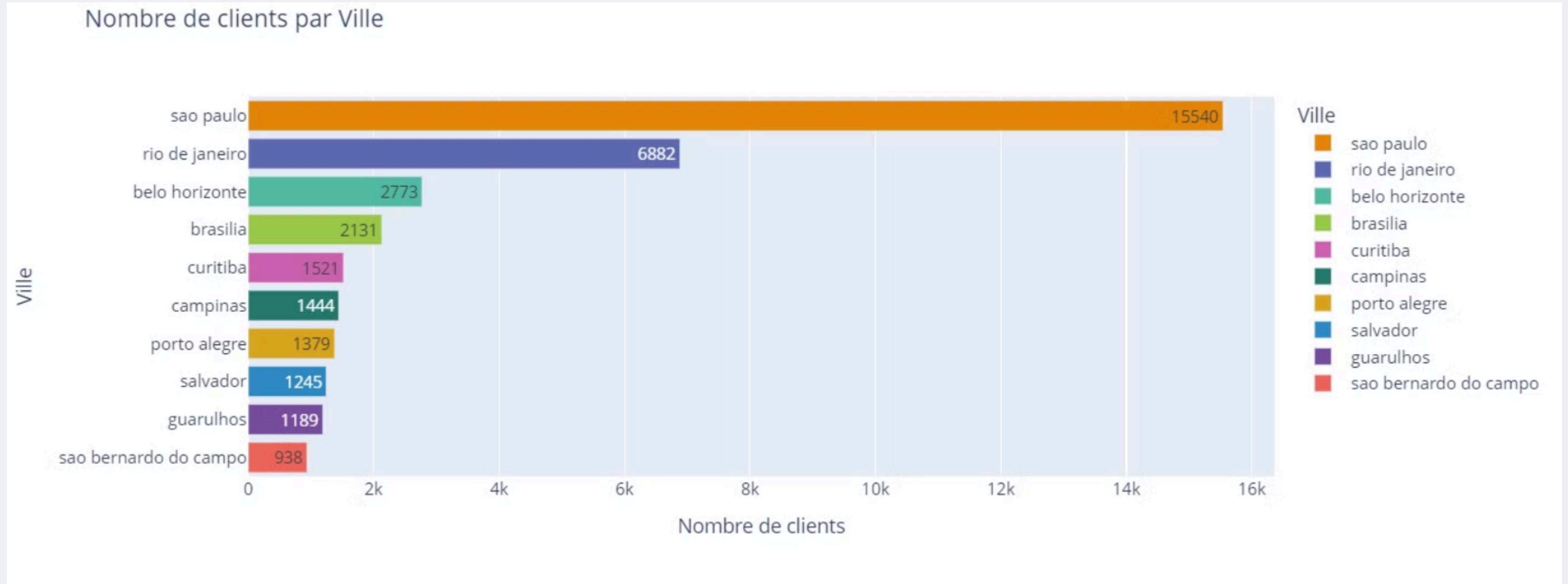
Analyse statistique, visualisation de données et requêtes SQL pour extraire ce qui est significatif.

# Orders



# Customers

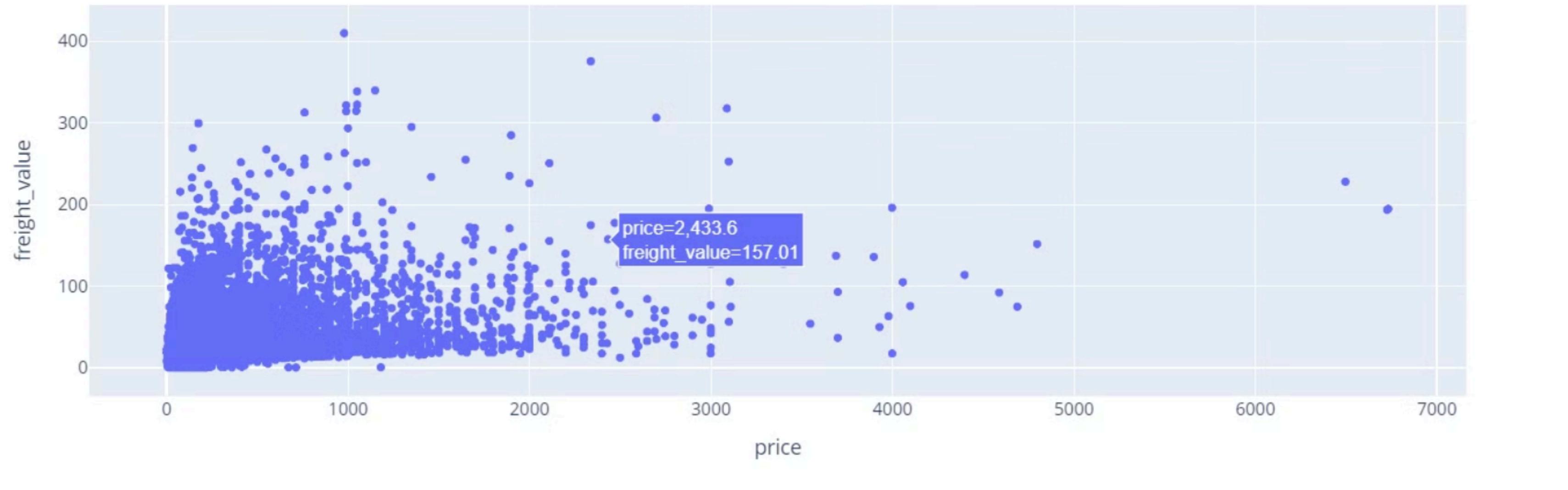
Nombre de clients par Ville



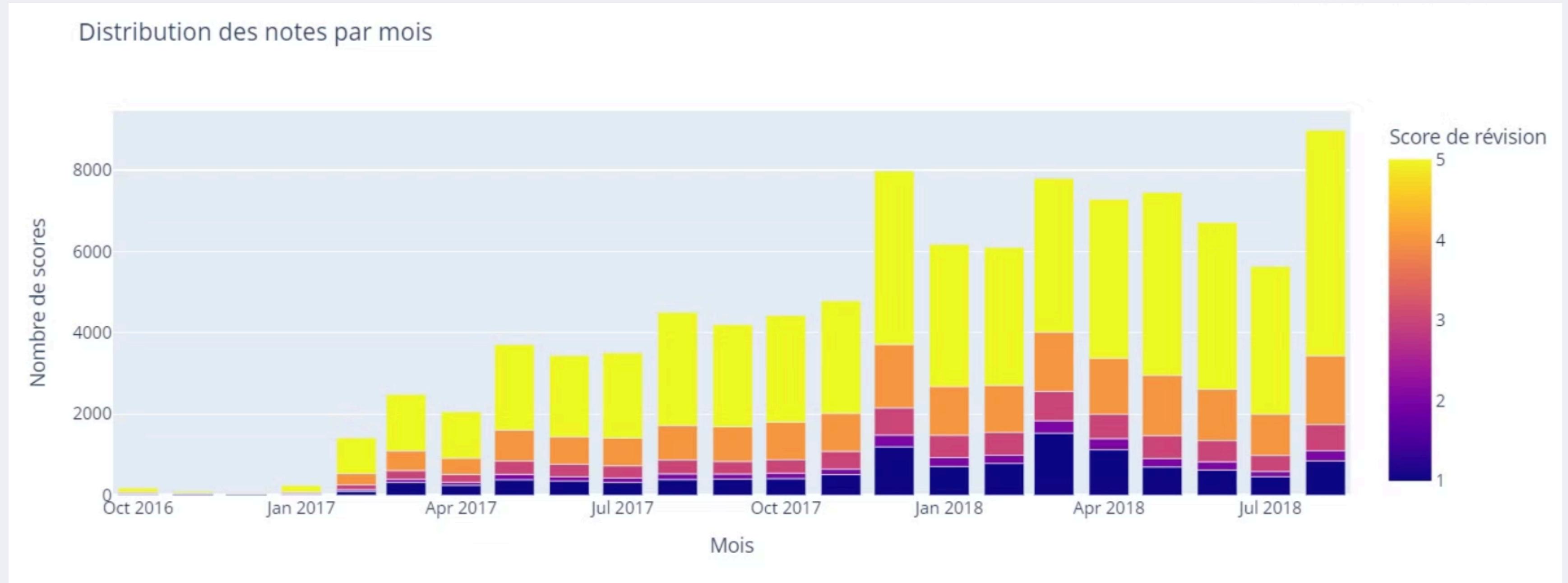
# Order\_items



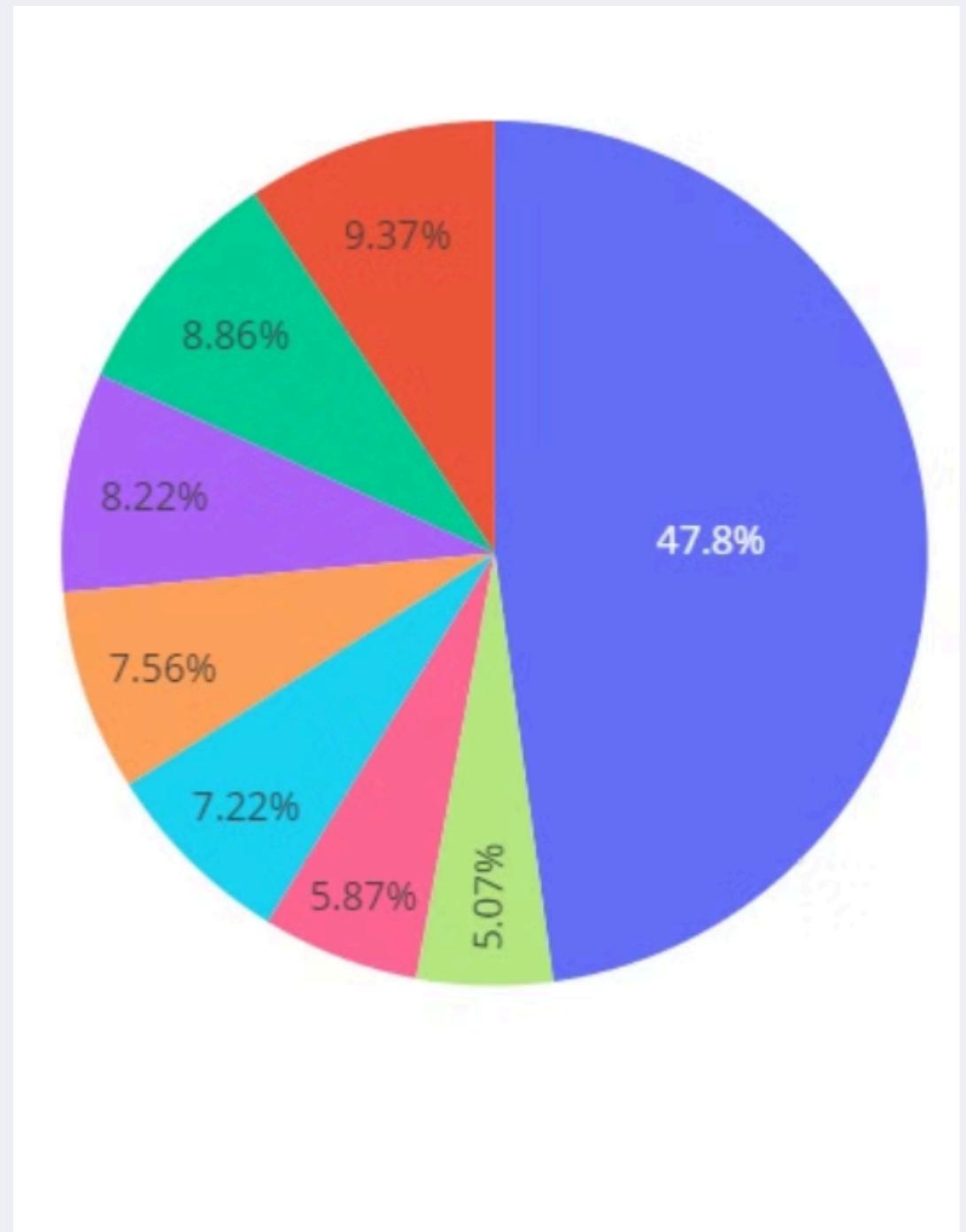
## Prix VS Fret



# Order\_reviews



# Product

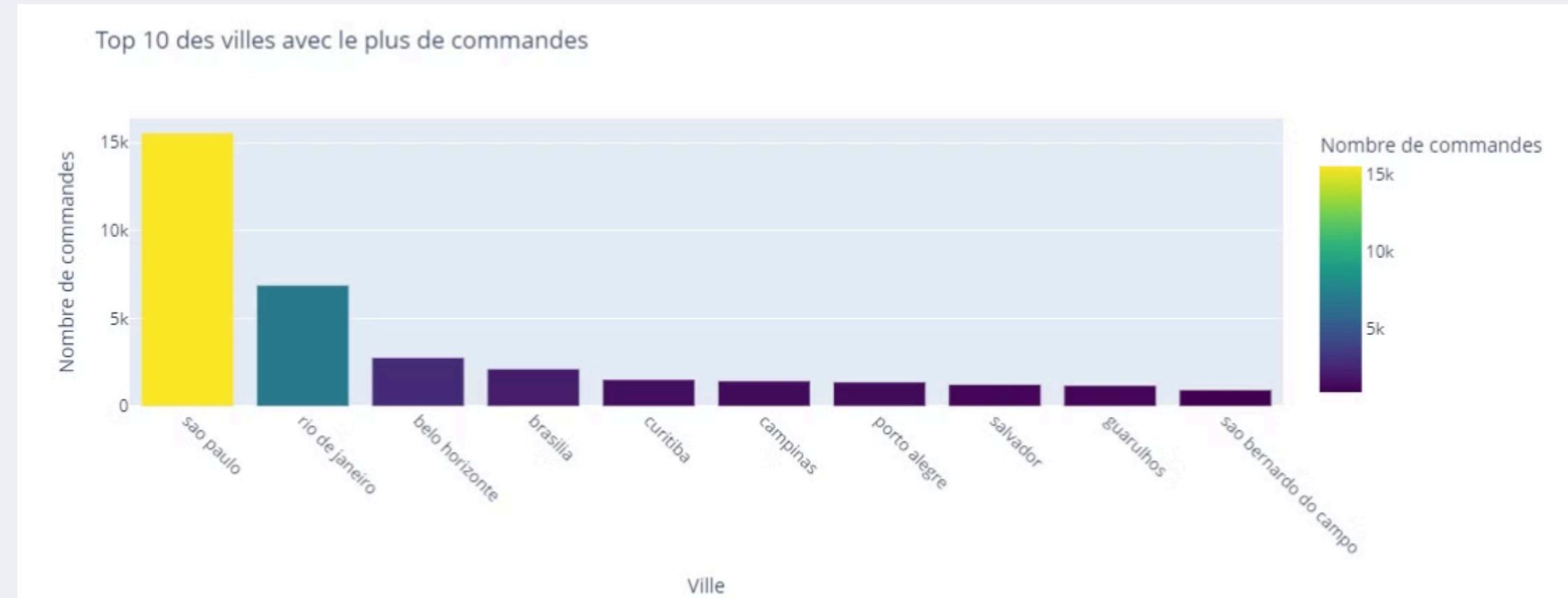
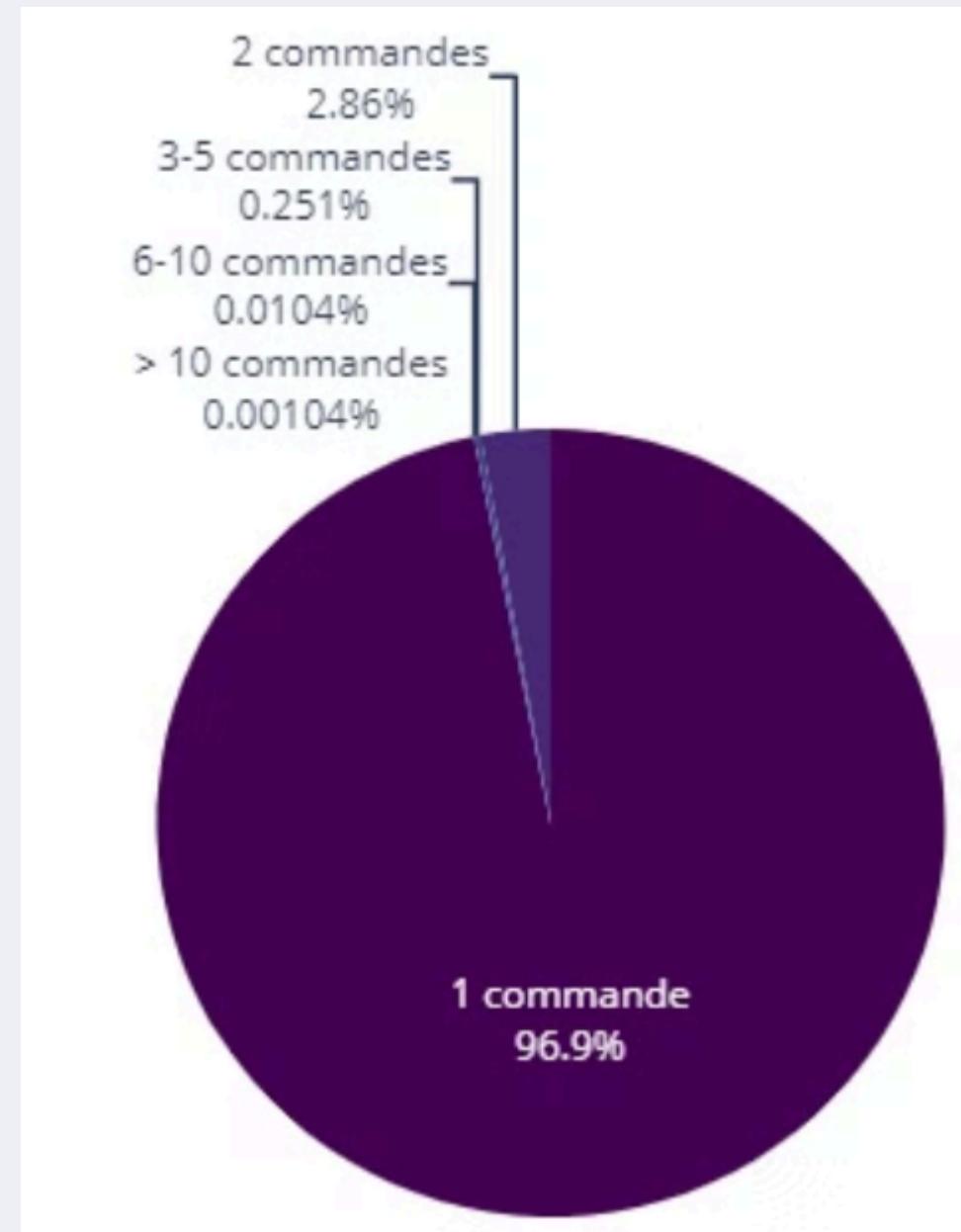


- **Cama mesa banho:** Lit et salle de bain
- **beleza saude:** beauté santé
- **esporte lazer:** sport loisirs

- **moveis decorado:** decoration film
- **informatica acessorios:** accessoire informatiques
- **utilidades domesticas:** produit domestiques
- **relogios presentes:** cadeaux horlogers
- **telefonia:** telephonie
- **ferramentas jardim:** outils de jardin
- **automotivo:** automobile

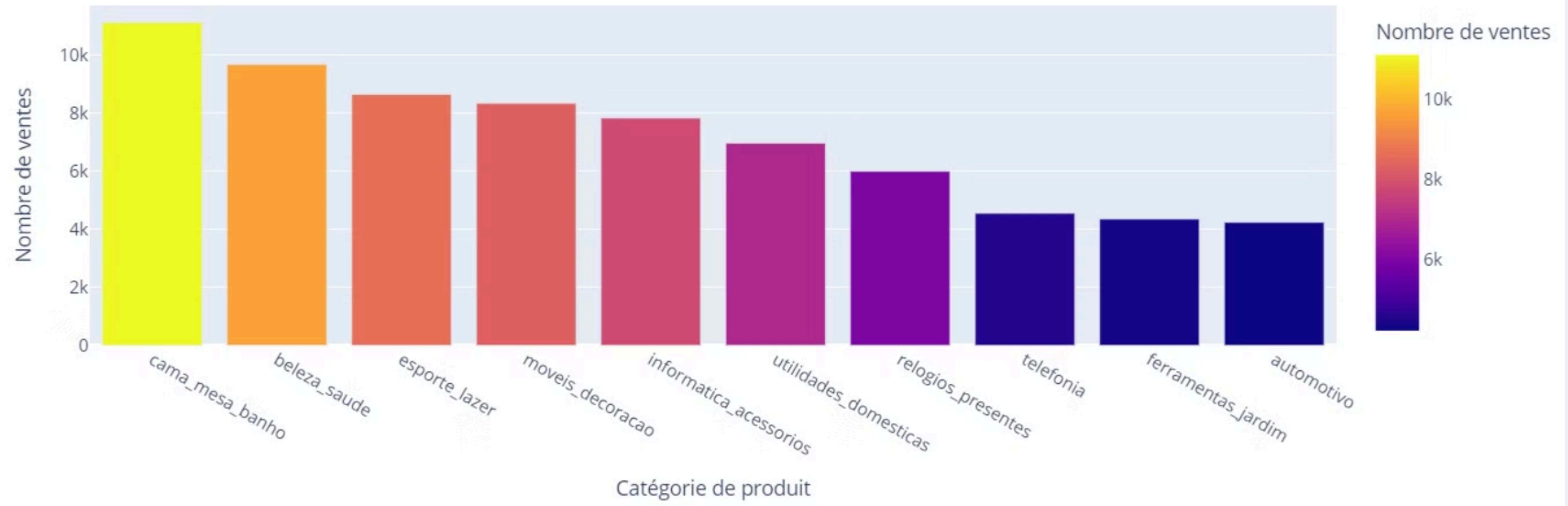
# Multi-Tables

## Client / Commande



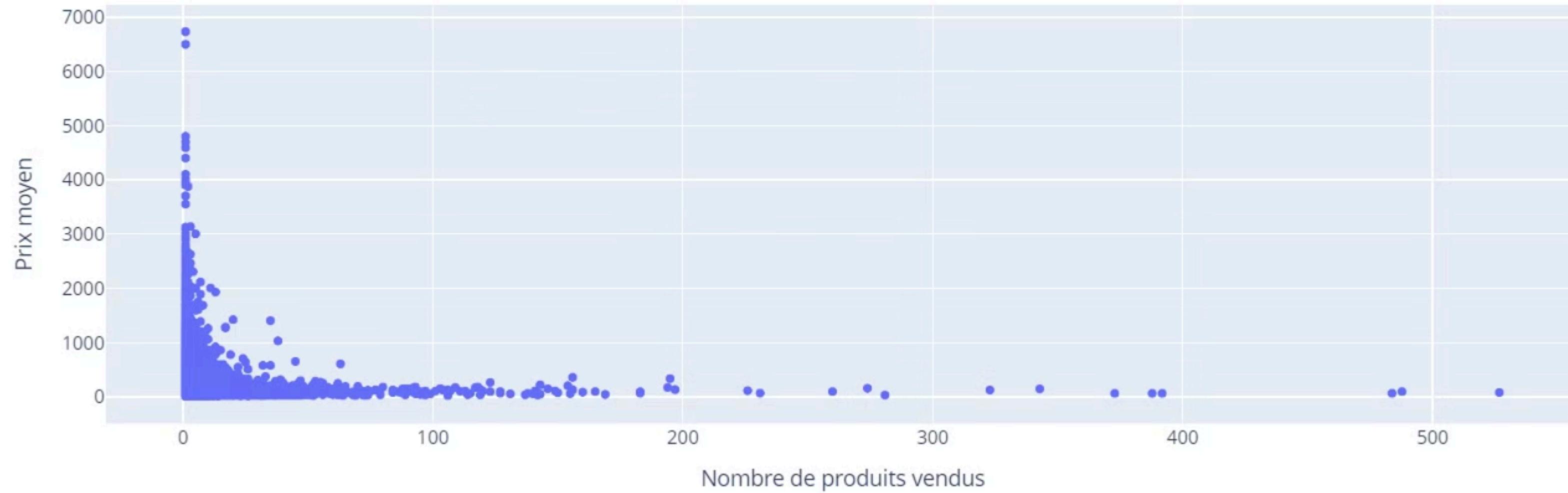
**customers / geolocalisation**

## Top des catégories de produits les plus vendues



**products / orders**

## Relation entre le nombre de produits vendus et leur prix moyen



**products / orders**



## Créations des variables

# RFM



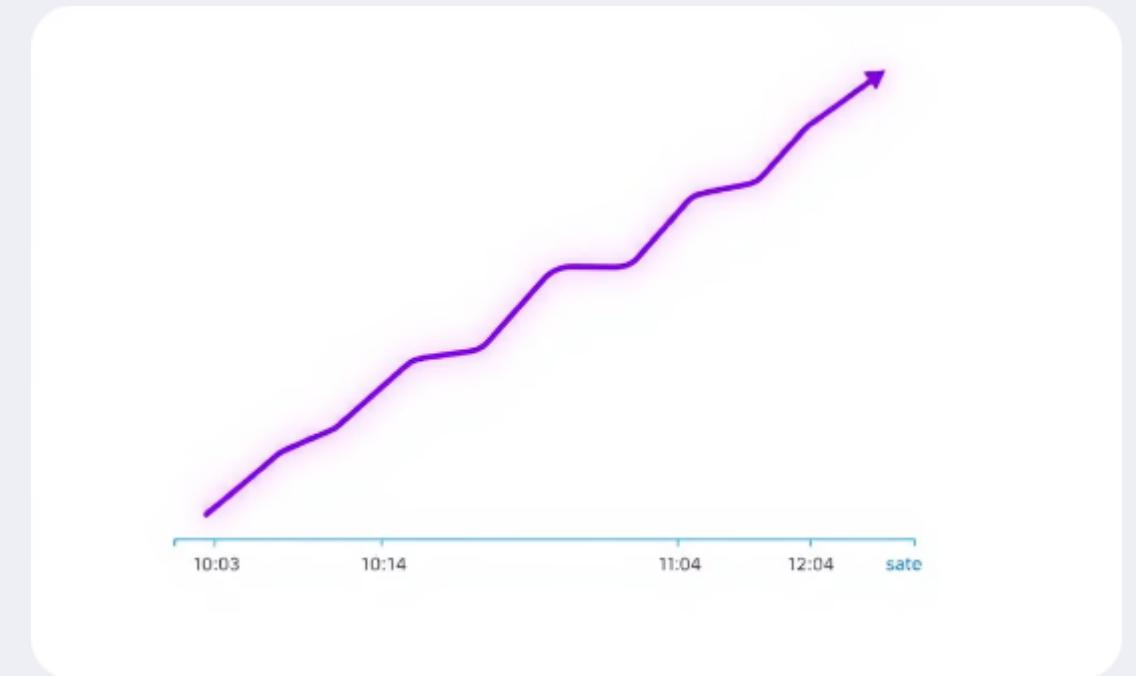
## Récence (R)

Temps écoulé depuis le dernier achat, indiquant l'activité récente du client.



## Montant (M)

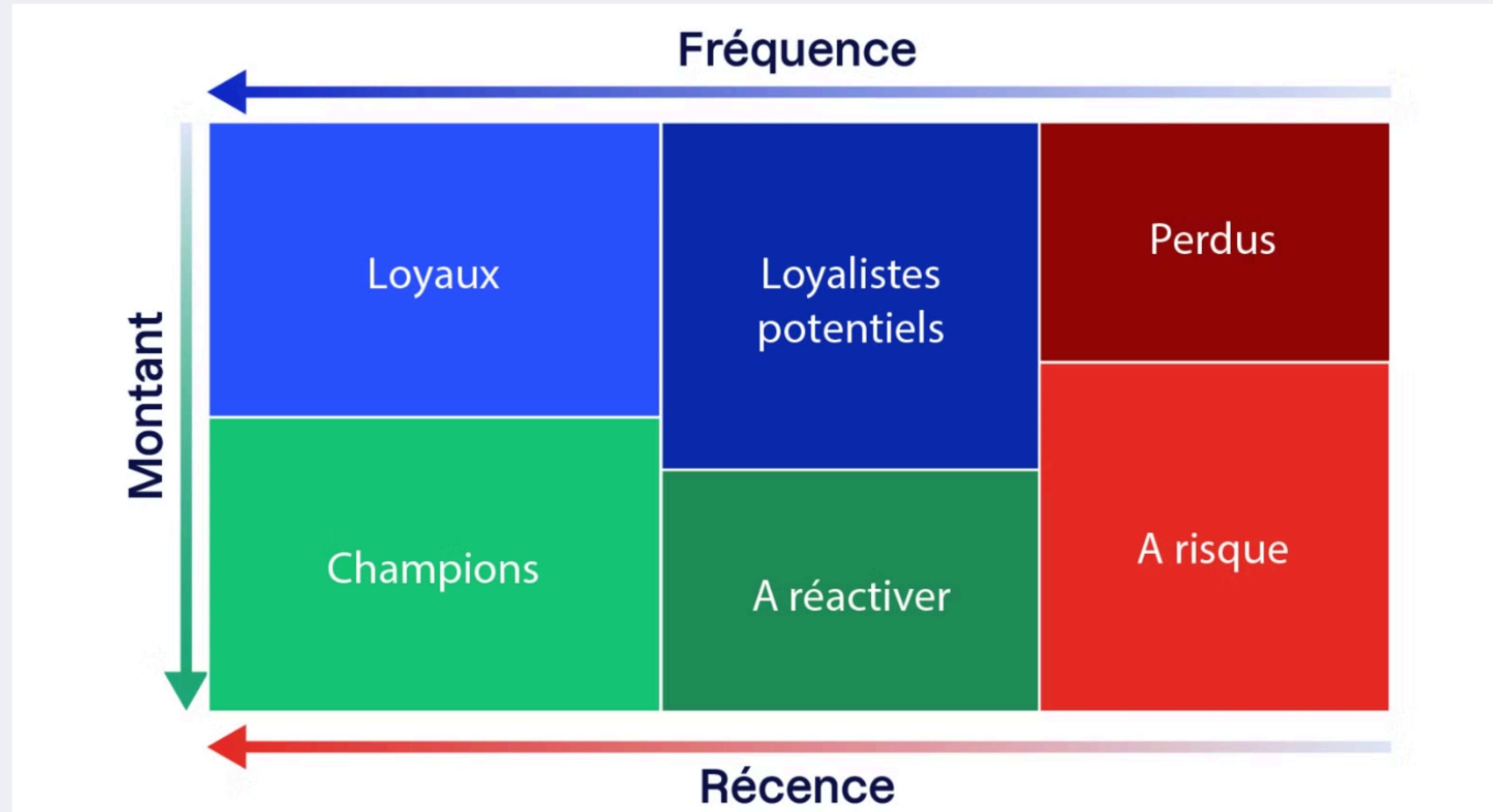
Valeur totale des achats, reflétant l'importance financière du client.



## Fréquence (F)

Nombre total d'achats, montrant la fidélité du client.

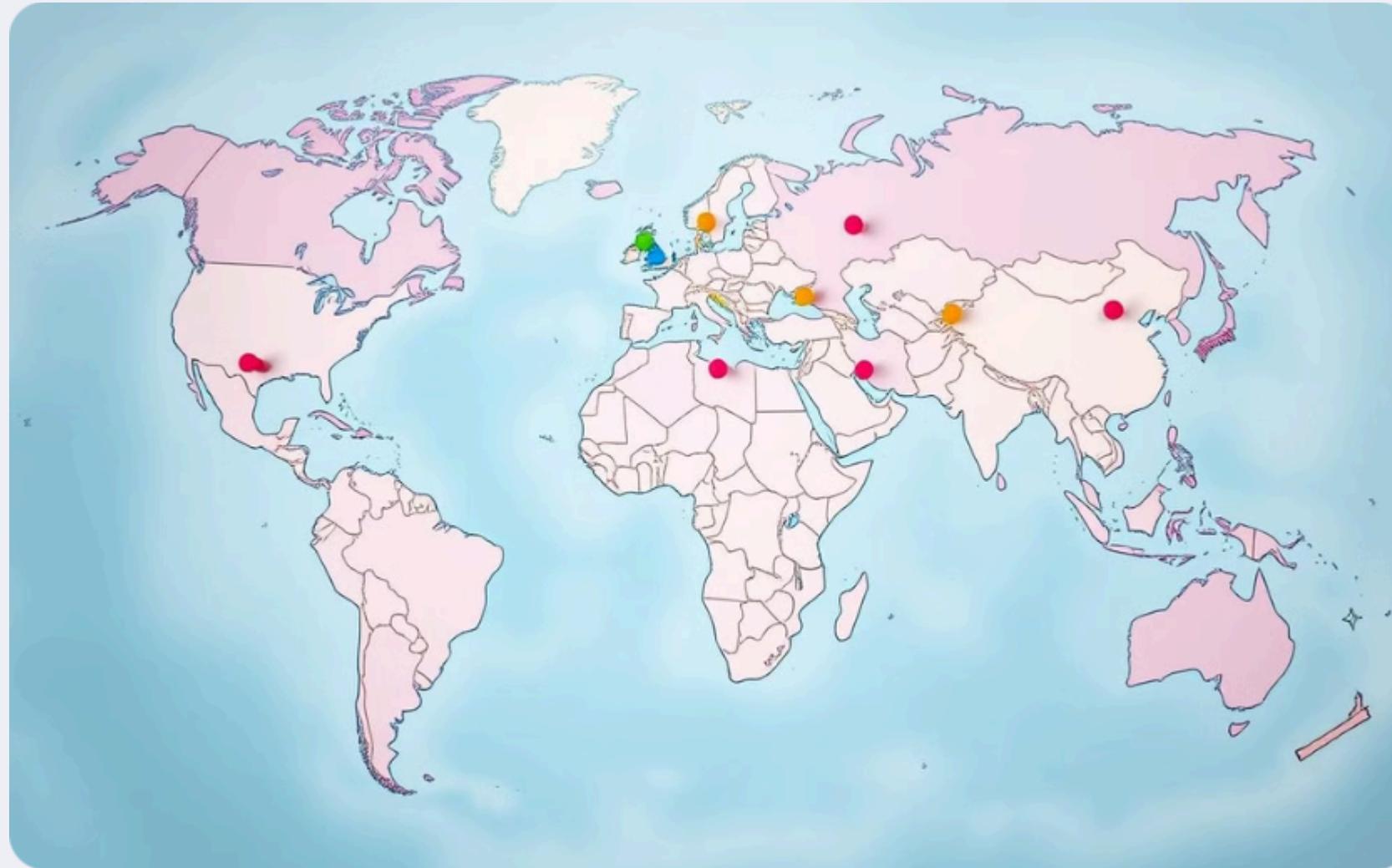
# Treemap de la segmentation clients



# Application de la Segmentations Clients (Treemap)



# Features



## Localisation (L)

Emplacement géographique des clients, pour adapter les stratégies marketing localement.



## Satisfaction Client (S)

Niveau de satisfaction des clients, mesuré par des enquêtes ou des avis.

# Comparaison des modèles (benchmark)

## 1 — Metrique de scoring

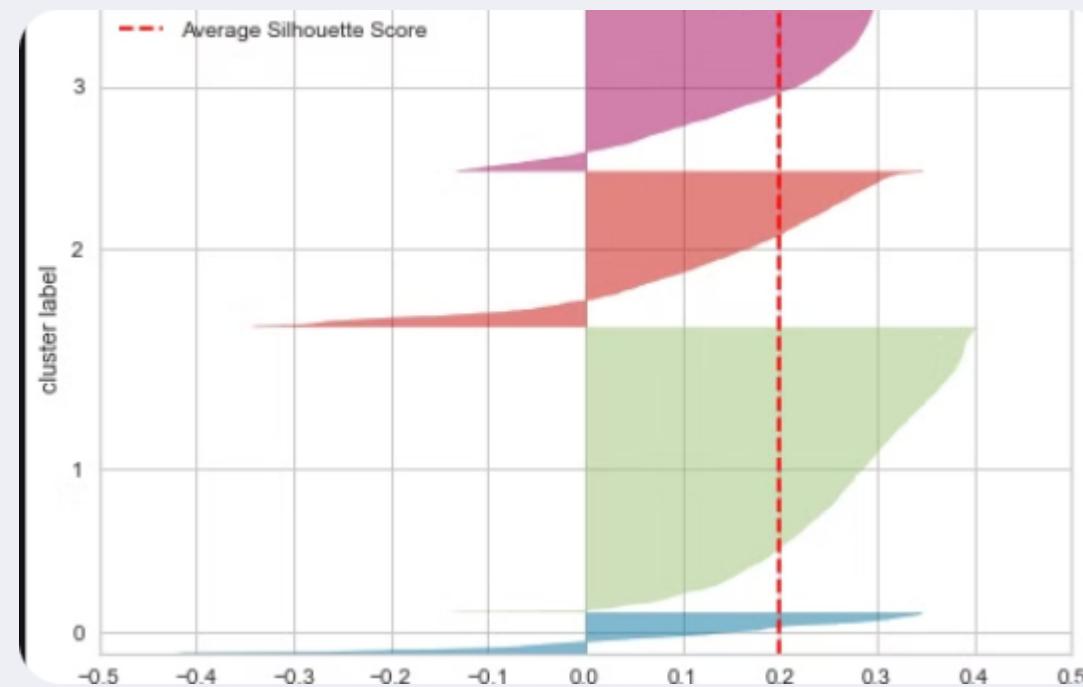
Sihlouette, davies bouldin, calinski

## 2 — Modèles de clustering

Benchmark des modèles

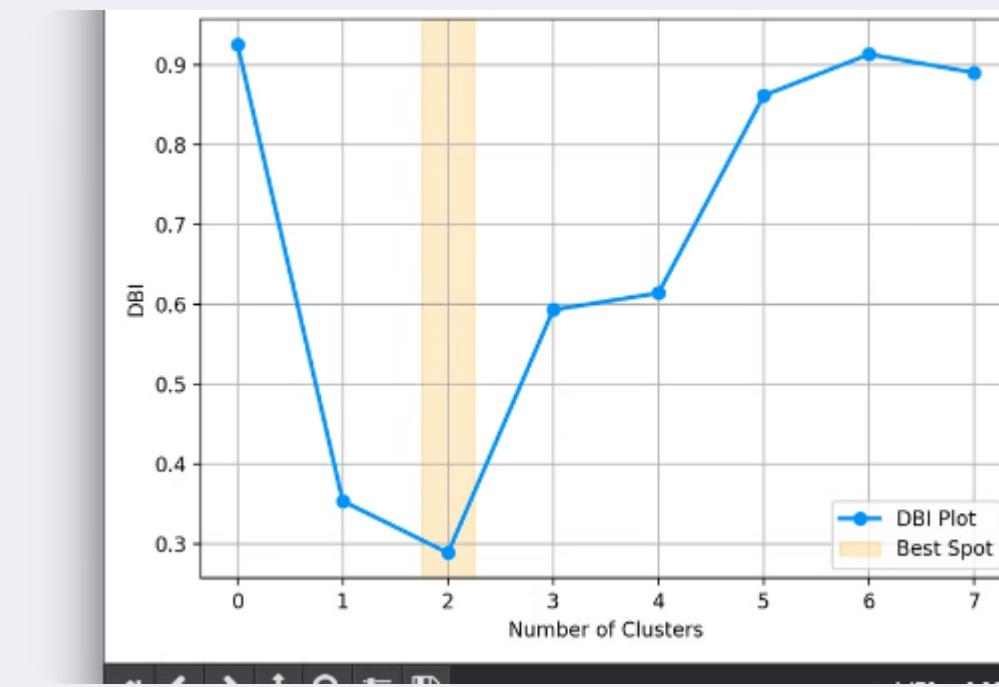


# Métriques de Scoring



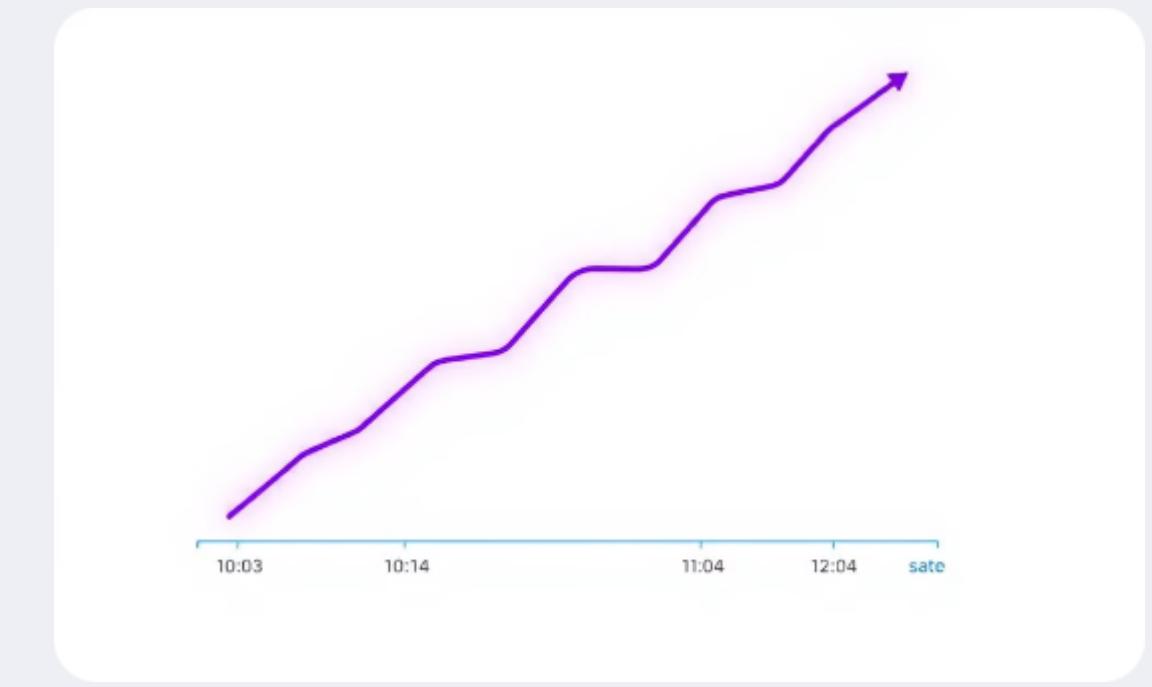
## Sihlouette Score

Évalue la similarité d'un point avec son propre cluster par rapport aux autres clusters, indiquant la cohésion et la séparation.



## Davies Bouldin

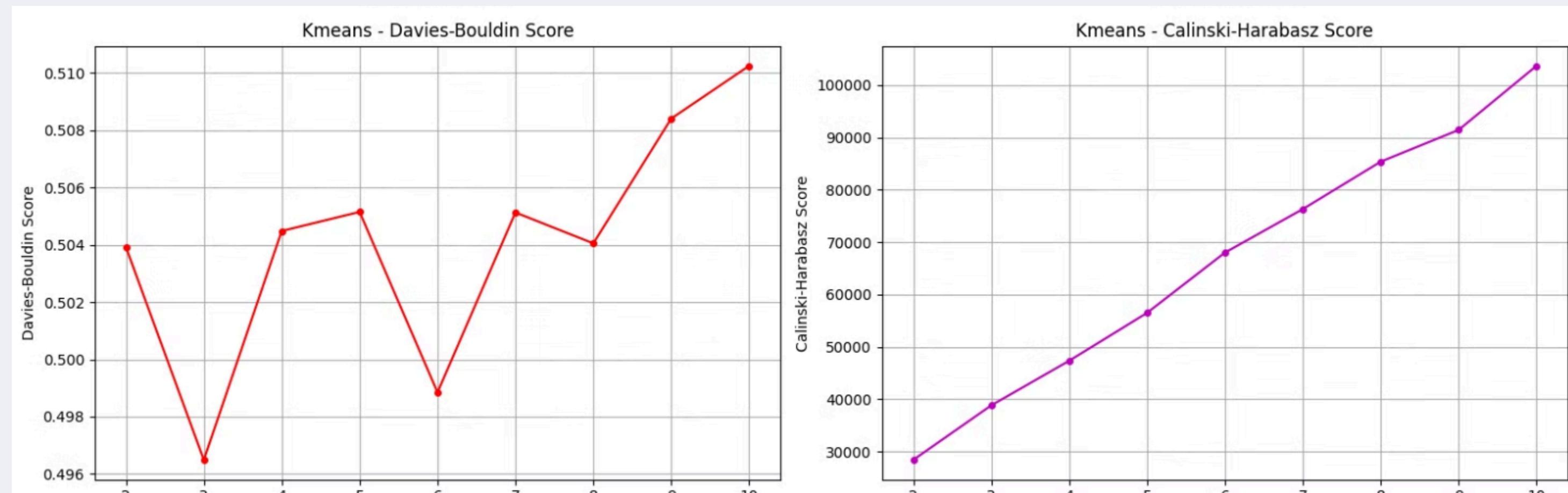
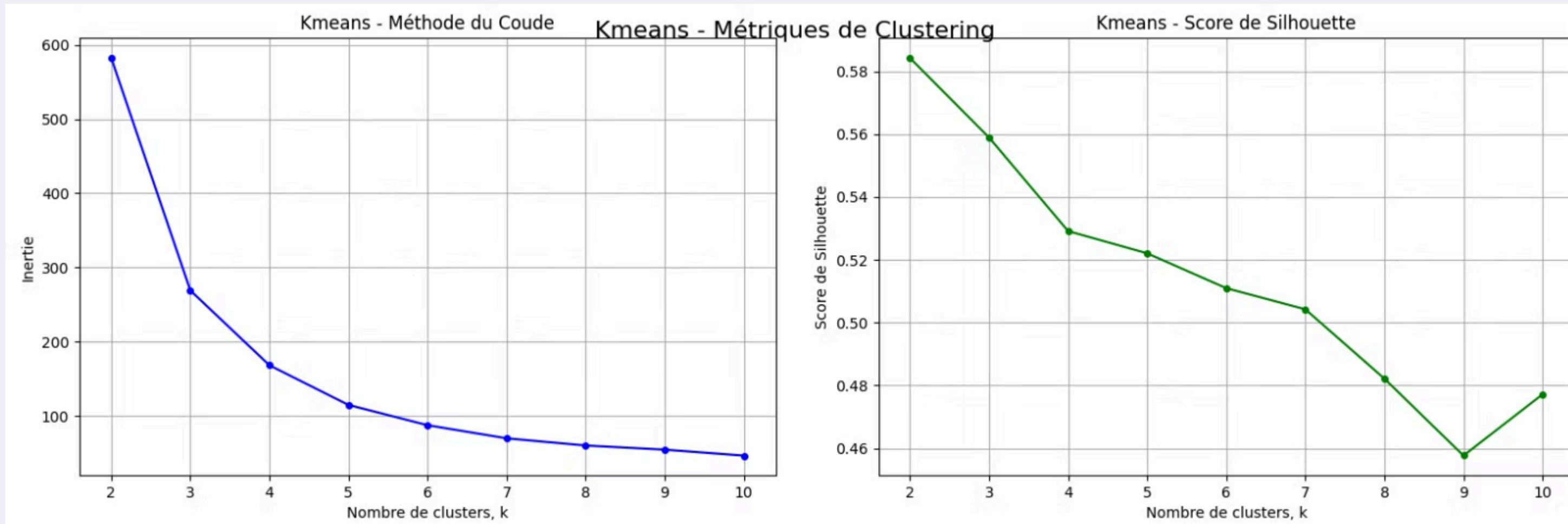
Mesure la qualité de la séparation entre les clusters, en comparant la dispersion intra-cluster et inter-



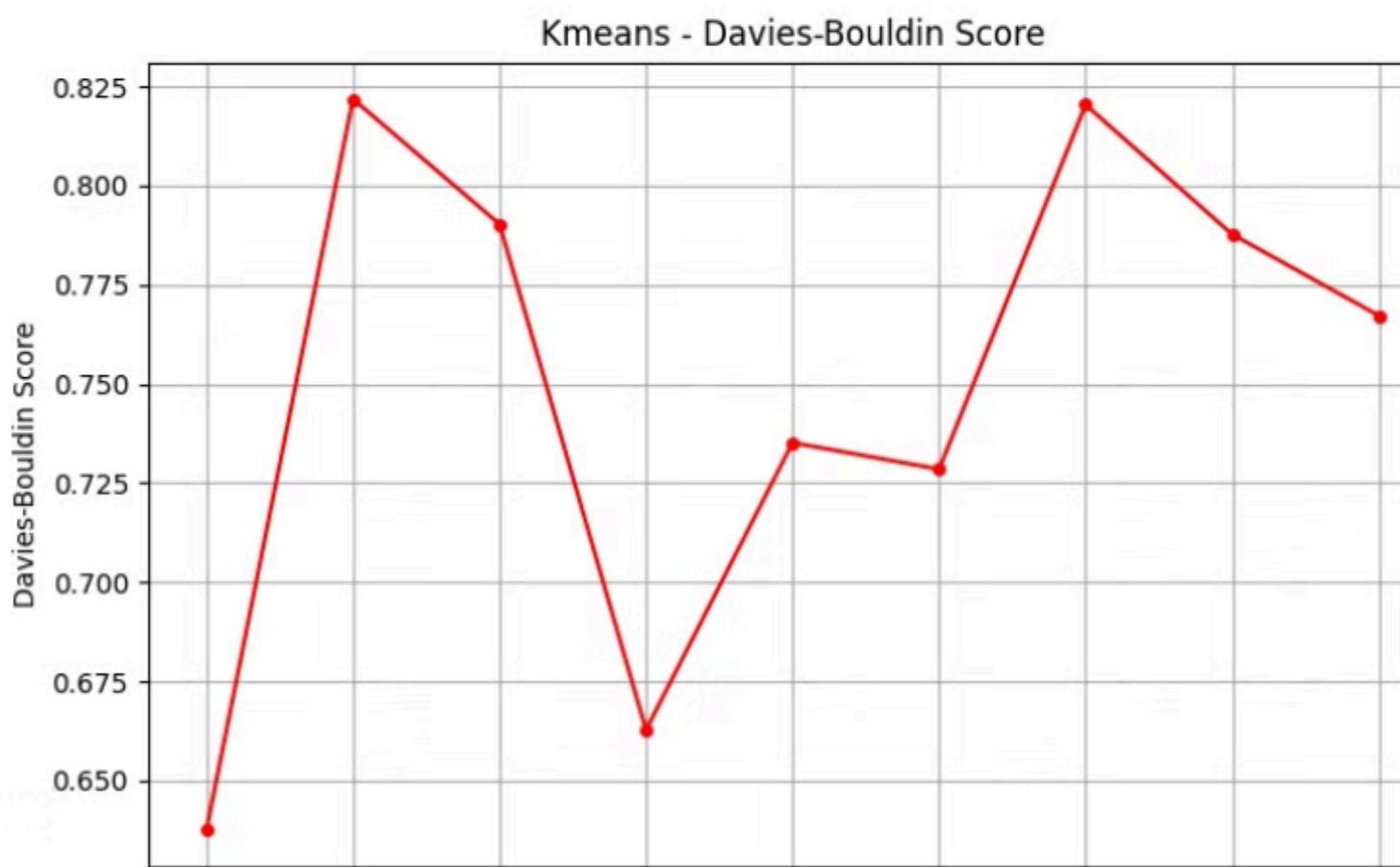
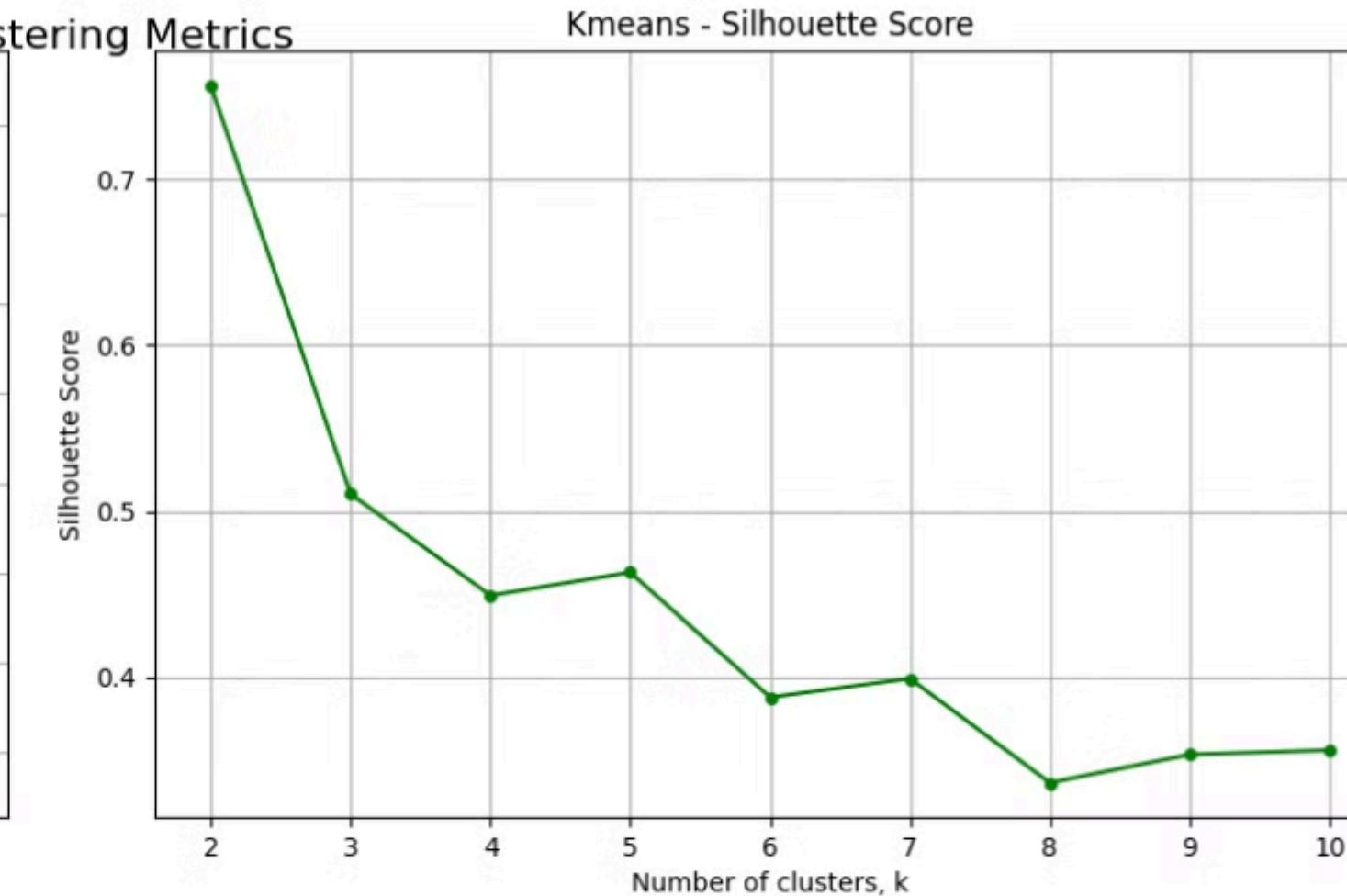
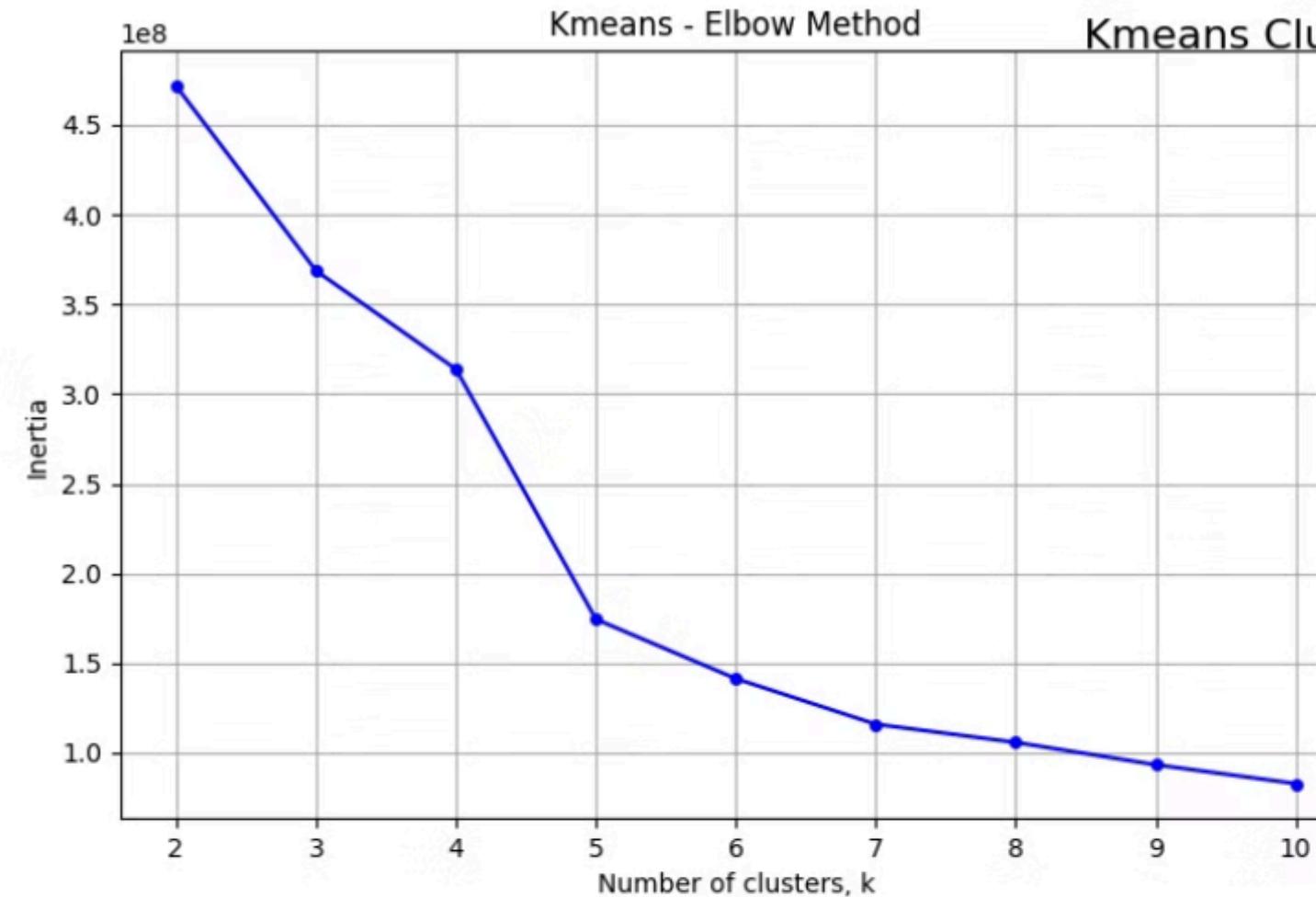
## Calinski

Évalue la densité et la séparation des clusters en comparant la variance intra-cluster et inter-cluster.

# Determination du nombre de cluster adequats por RFM



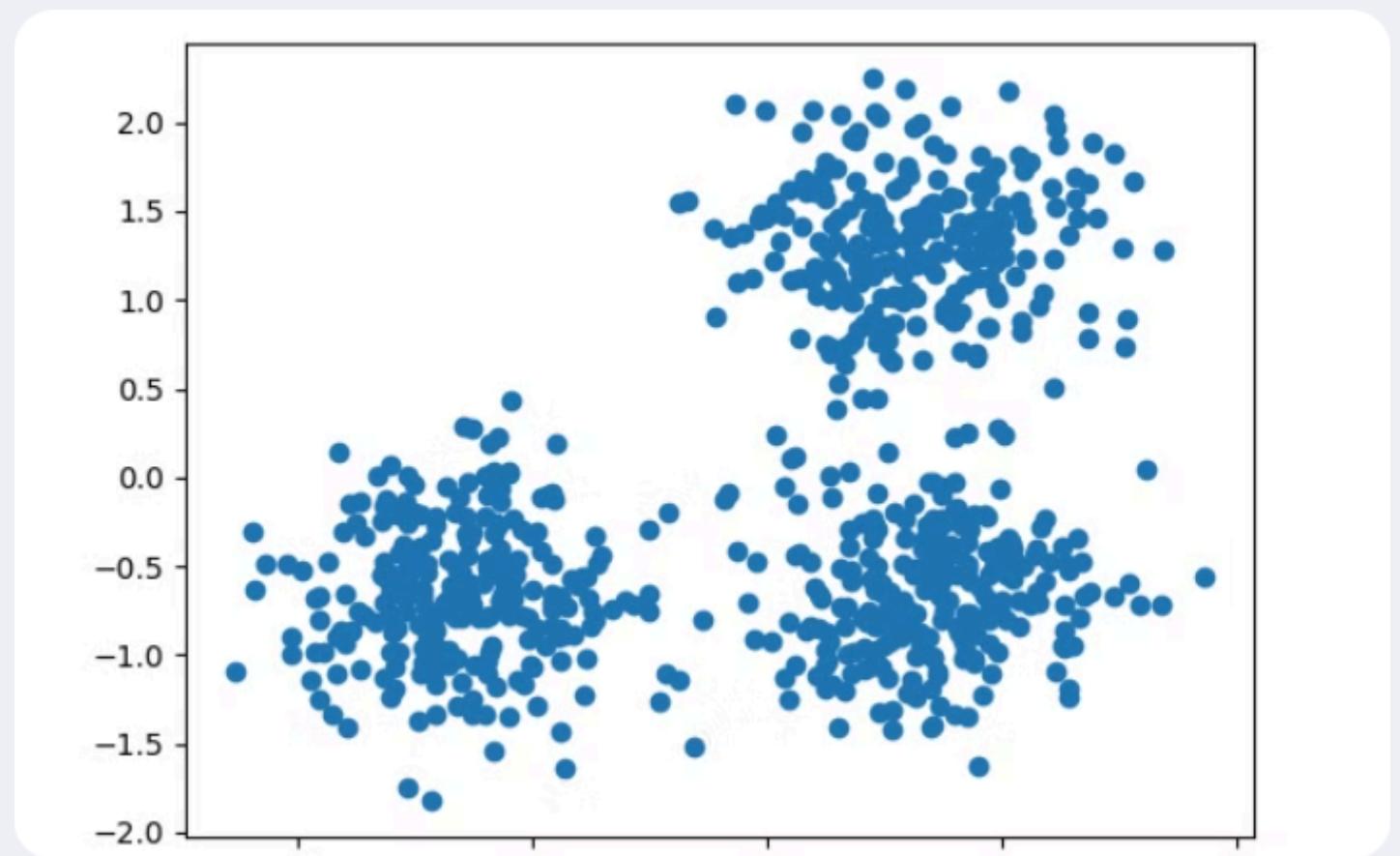
# Determination du nombre de cluster adequats por RFM + L + S



# Evaluations des models de clustering

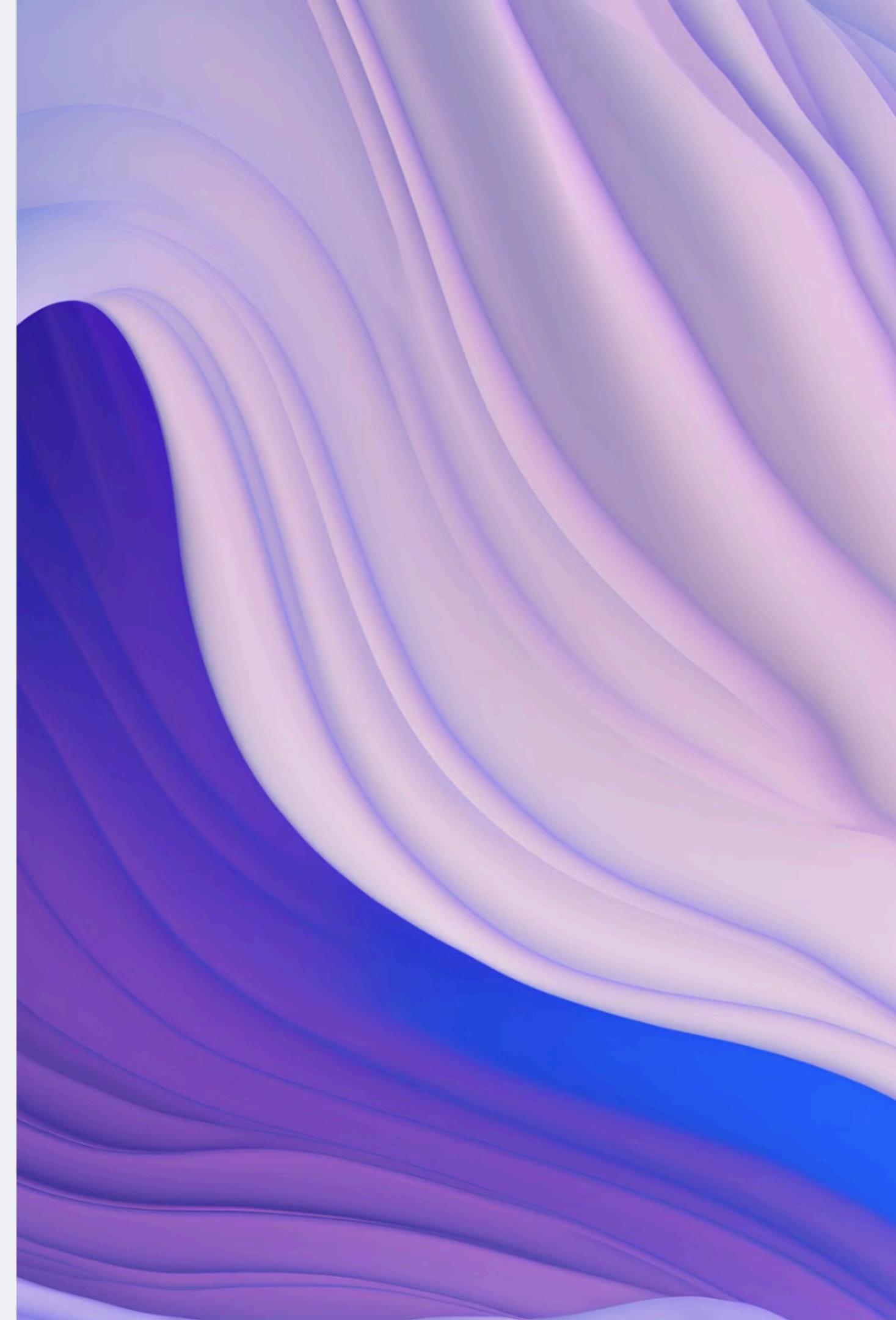
	Algorithm	Time (s)	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
0	kmeans	0.171138	0.594249	0.496498	38798.572844
1	minibatchkmeans	0.171398	0.594121	0.496599	38792.233582
4	birch	2.906458	0.593265	0.496840	38681.788157
5	agglomerative	3.109138	0.593265	0.496840	38681.788157
8	meanshift	49.161351	0.538539	0.503867	27051.463914
3	gmm	0.229274	-0.008243	79.012811	1.430951
2	hdbscan	0.223837	-0.296055	3.306890	48.244279
7	optics	7.588027	-0.446664	3.117610	12.360714
6	dbSCAN	0.047138	NaN	NaN	NaN

# Modèles non retenu



**Modèles basés sur la densité des points de données**

DBSCAN, HDBSCAN, OPTICS

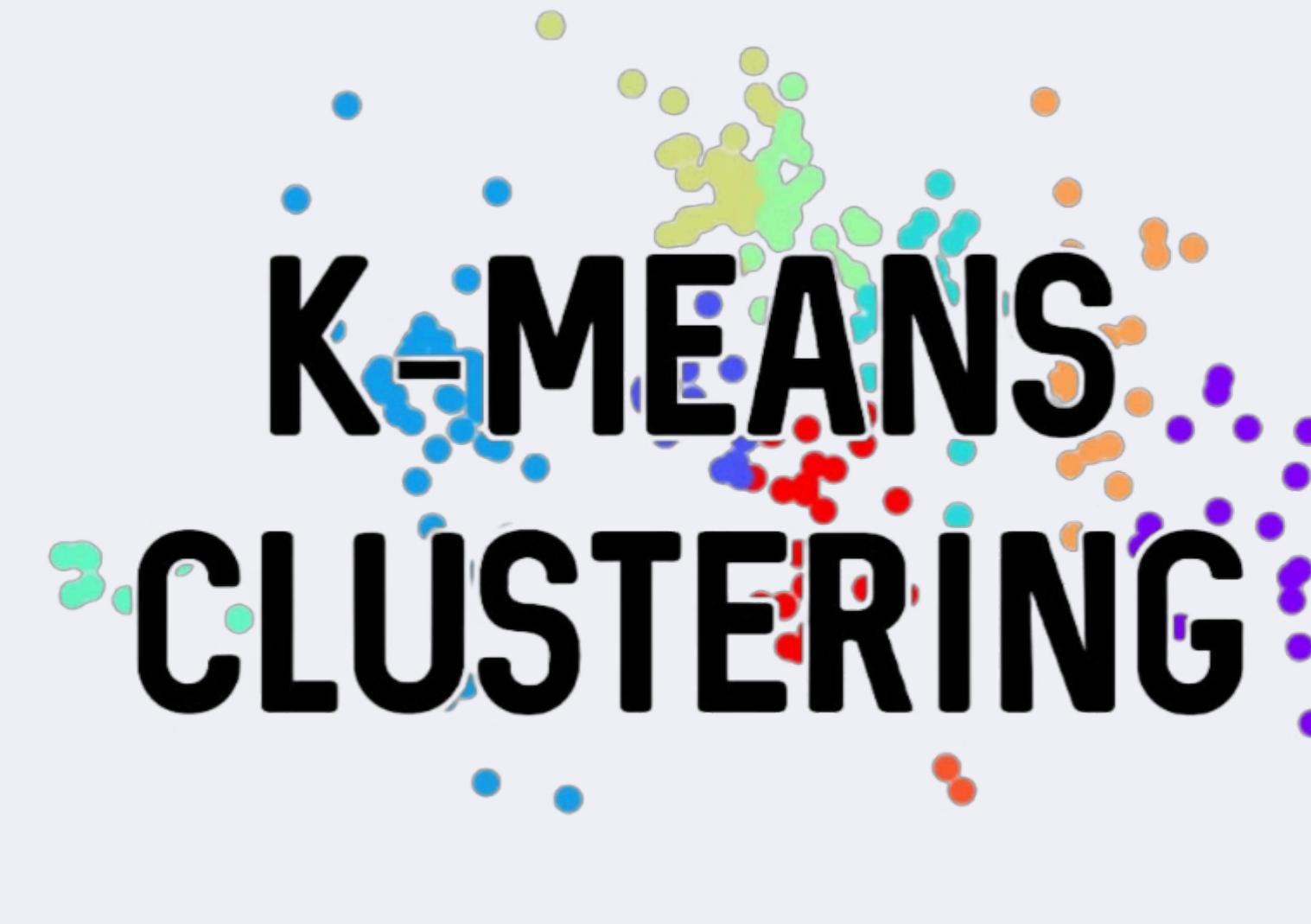


## Modèles non retenu - Faibles metrics

5	agglomerative	14.134435	0.593267	0.496833	38681.787083
8	meanshift	143.184301	0.538532	0.503872	27050.627977
2	hdbscan	2.644889	0.268603	2.955818	3581.264838
3	gmm	0.083935	-0.000662	190.360082	0.174185
6	dbscan	0.146302	NaN	NaN	NaN
7	optics	30.450814	NaN	NaN	NaN

# Modèle retenu

Le modèle qui a été retenu est Kmeans car il possède de meilleure performance globale



Algorithm	Time (s)	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
0	kmeans	0.054837	0.594251	0.496492

# Entrainement des modèles



## Entrainement

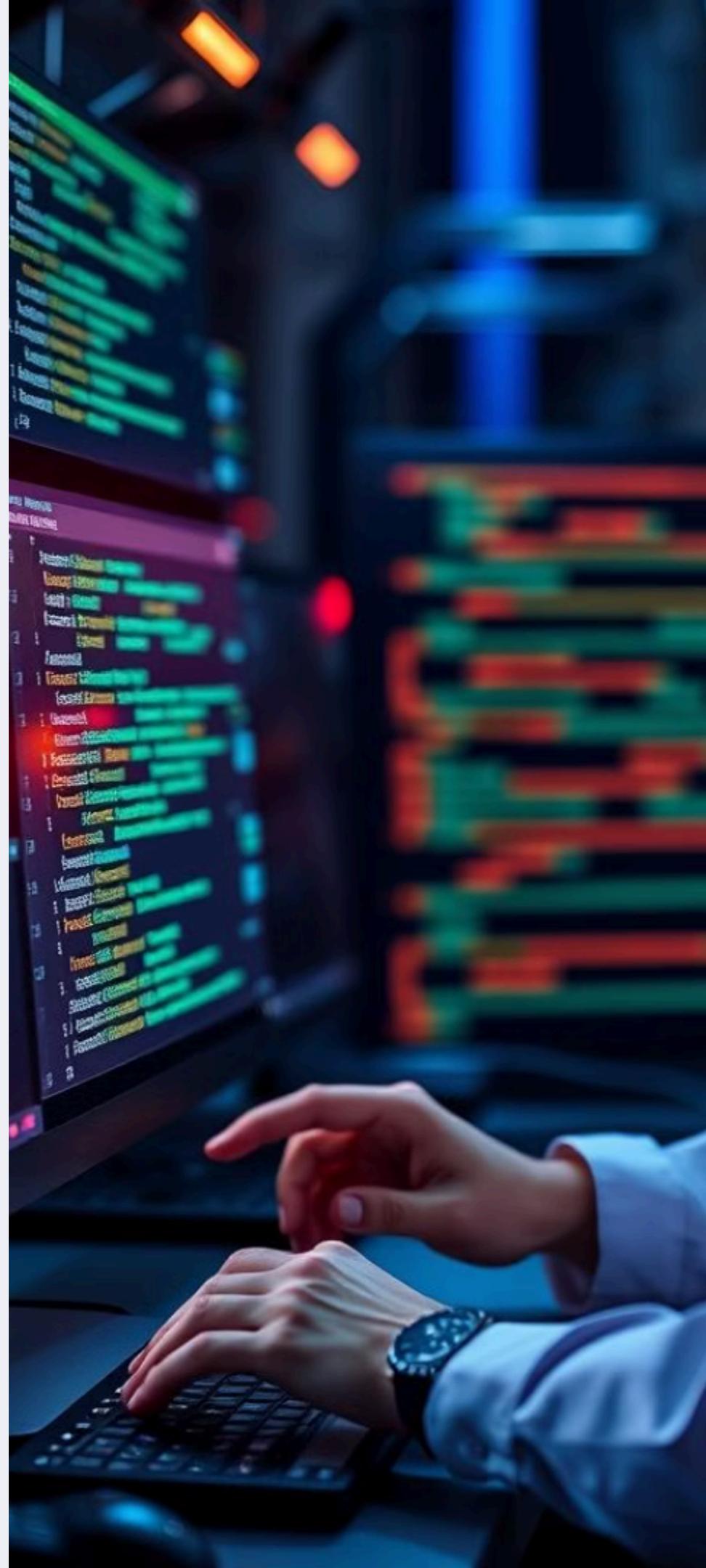
Entraînement du modèle sur un ensemble de données étiquetées. Ajuster les paramètres du modèle pour minimiser l'erreur de prédiction.

## Évaluation

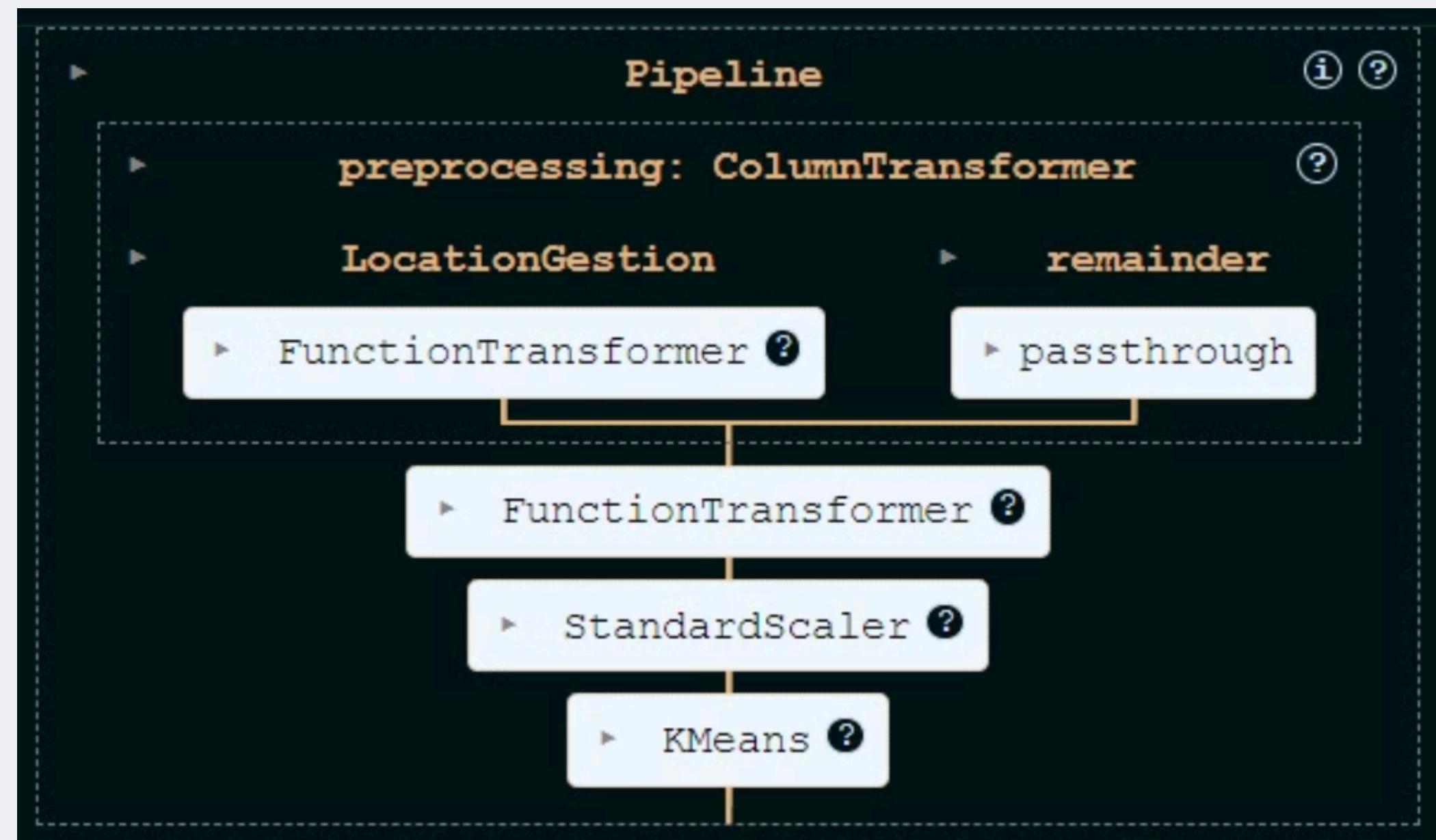
Évaluation du modèle sur un ensemble de données distinct. Mesurer la performance du modèle à l'aide des métriques définies.

## Sélection du modèle

Sélection du modèle le plus performant en fonction des métriques et de l'analyse des résultats.



# Création d'une pipeline Sklearn

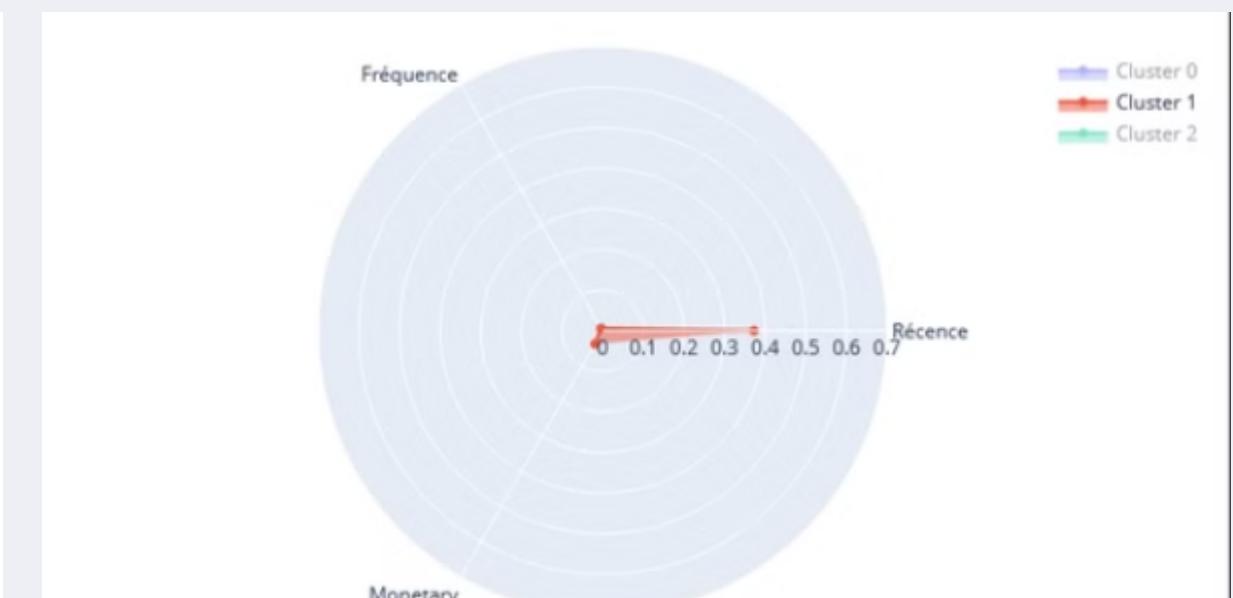
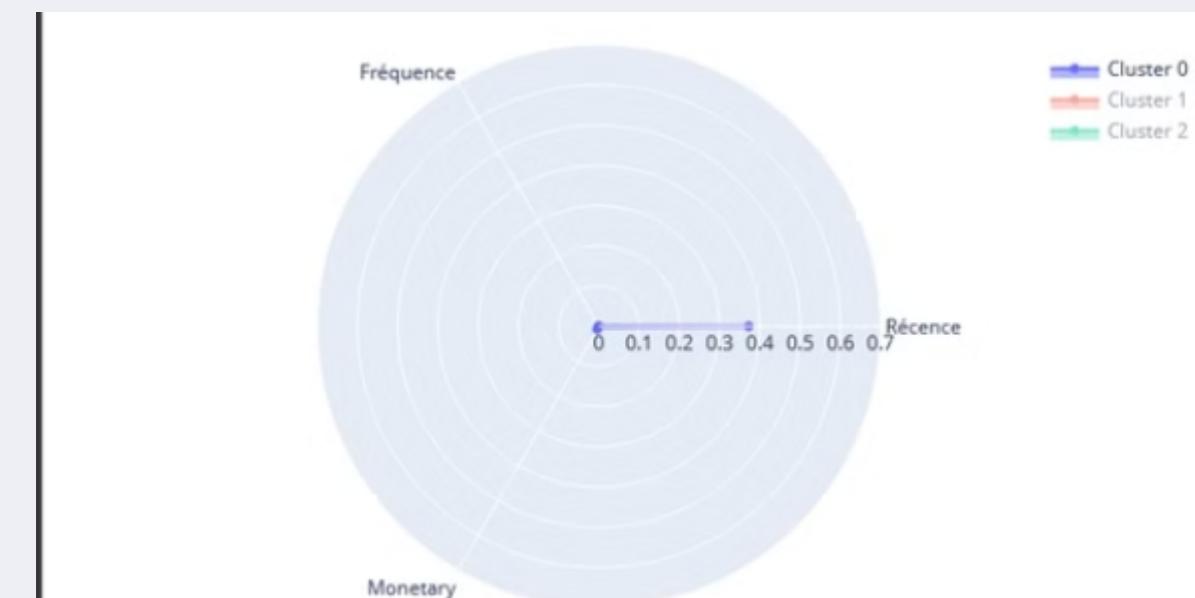


# Identification des clusters

Radar plot RFM des Segments



Radarplot clustrers



# Persona à partir des clusters

## Cluster 0

- **Description** : Nouvel utilisateurs avec une fréquence d'utilisation faible et une monétisation faible.
- **Objectifs** : Augmenter la monétisation en proposant des fonctionnalités premium ou des abonnements.

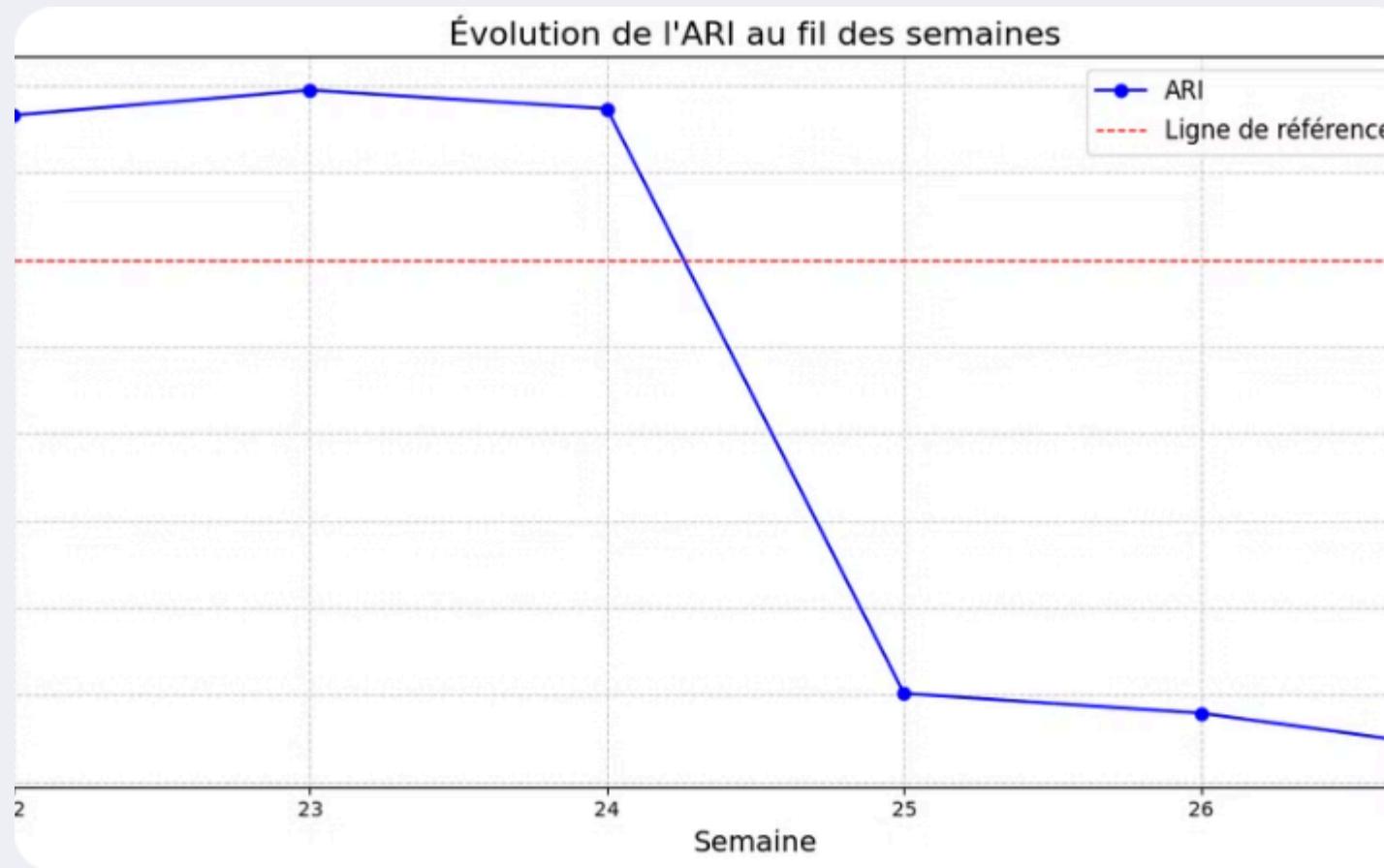
## Cluster 1

- **Description** : Utilisateurs avec une fréquence d'utilisation élevée et une monétisation modérée.
- **Objectifs** : Maintenir l'engagement et explorer des opportunités de monétisation supplémentaire, comme des achats intégrés ou des services premium.

## Cluster 2

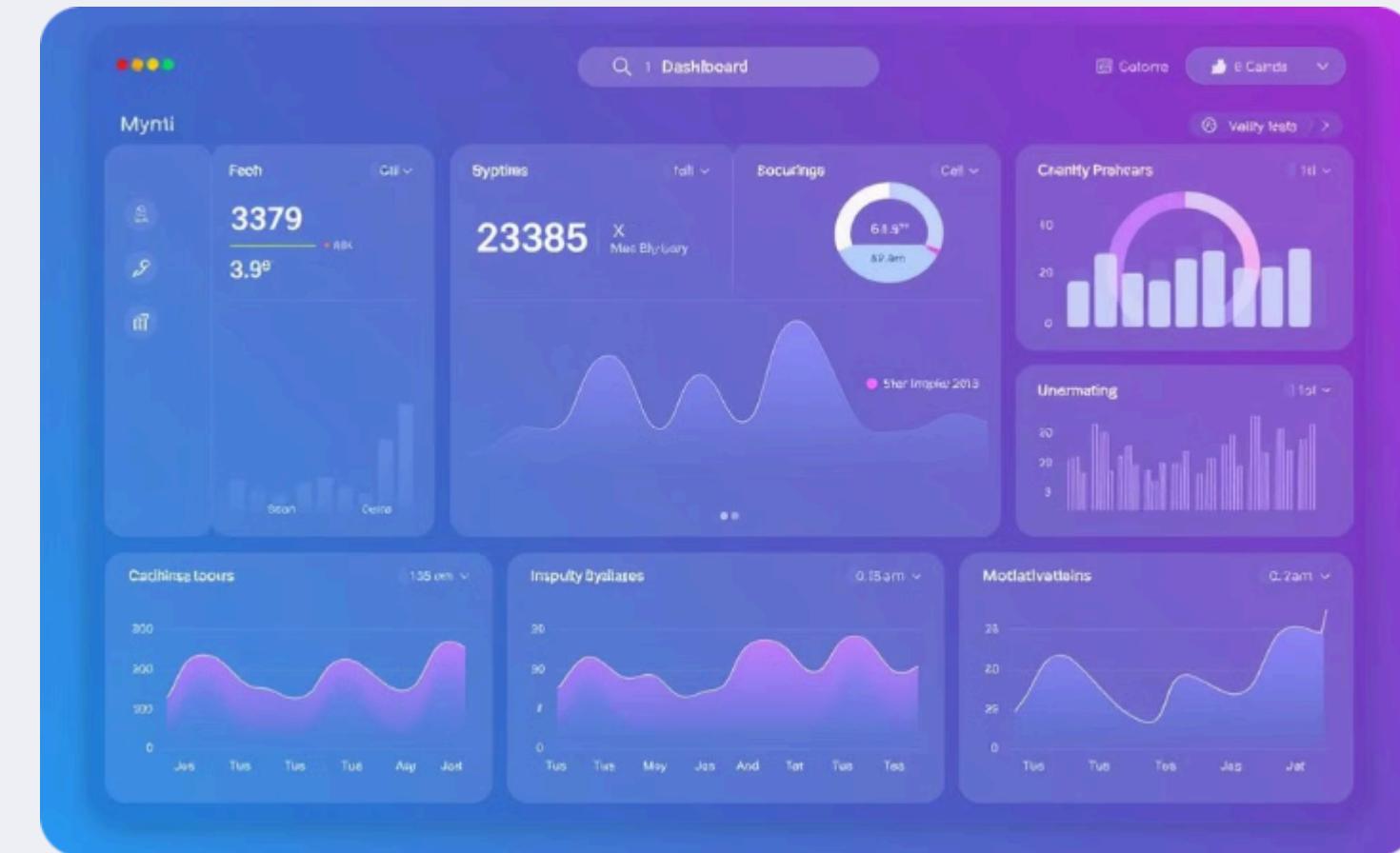
- **Description** : Utilisateurs avec une recence élevée et une monétisation modérée.
- **Objectifs** : Maintenir l'engagement et explorer des opportunités de monétisation supplémentaire, comme des achats intégrés ou des services premium.

# Contrat de maintenance



## Adjusted Rand Index (ARI)

Évalue la similarité entre deux partitions de données, ajustée pour le hasard.



## Surveillance des métriques

Assure le suivi et l'analyse continue des performances des modèles.

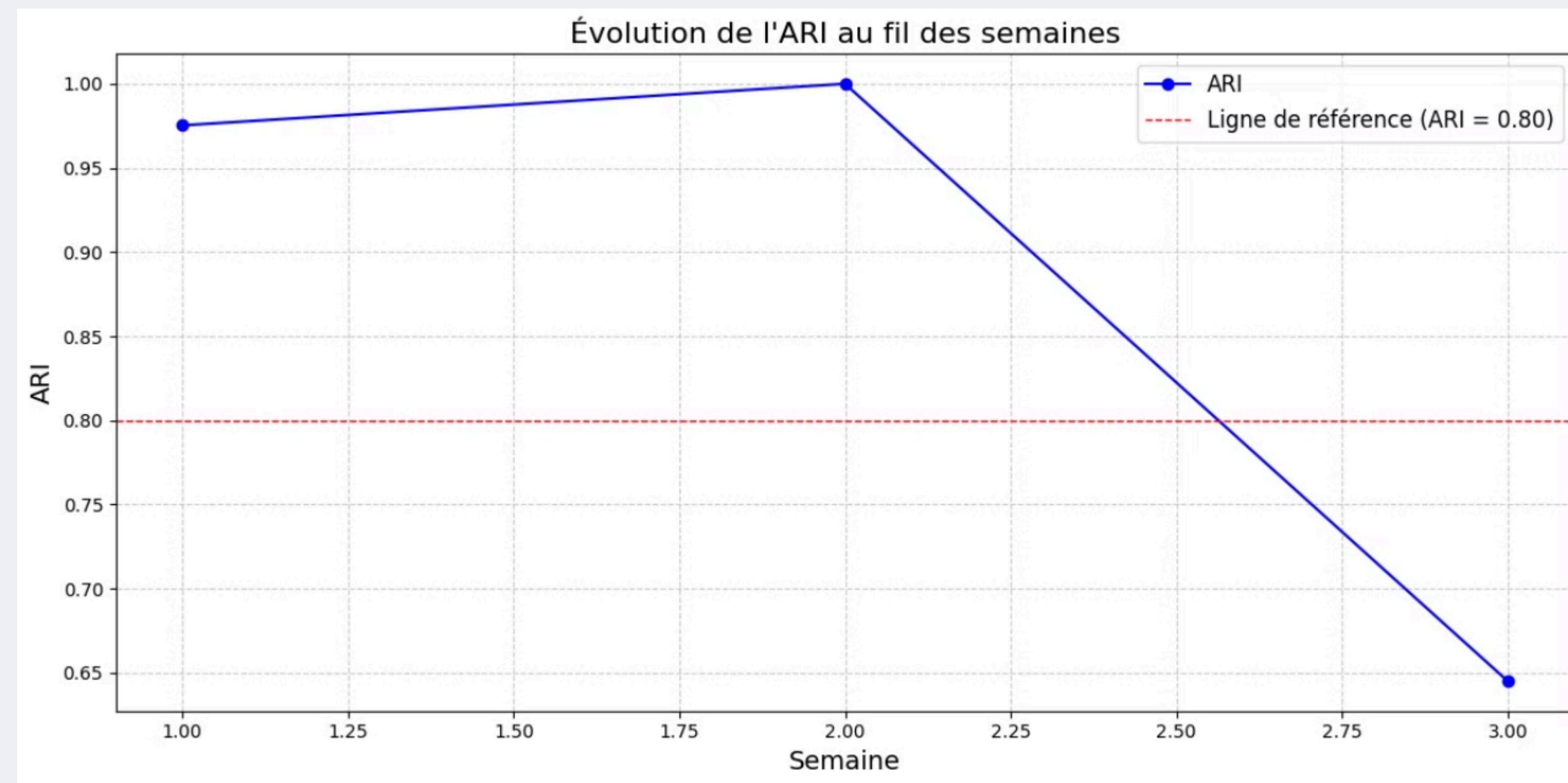


# Ajusted Rand Index

L'Adjusted Rand Index (ARI) mesure la similarité entre deux partitions d'un ensemble de données. Il compare les groupes formés par deux méthodes différentes, en tenant compte du hasard. Un ARI de 1 indique des partitions identiques, 0 des partitions aléatoires, et un ARI négatif des partitions moins similaires que le hasard.

# Quand doit-on re-entraîner le modèle

D'après le graphique ci dessous montrant l'évolution du modèle au cours du temps un reentraînement sera nécessaire tous les 15 jours en moyenne pour s'assurer que le modèle n'a pas de période de faible performance



# Persona Important à Prendre en Compte

## Le Nouvel Acheteur

Un client qui effectue son premier achat. L'objectif est de comprendre ce qui motive son achat et comment l'inciter à revenir

## Le Client Perdu

Un client qui n'a pas commandé depuis longtemps. L'objectif est de comprendre pourquoi et comment le faire revenir

## Le Client à Risque

Un client occasionnel ayant effectué peu d'achats récemment et avec un faible montant total dépensé. L'objectif est de le réengager à travers des offres personnalisées augmenter sa fréquence d'achat et sa valeur client

# Conclusion



# Améliorations proposées

## Qualité des Données

S'assurer de la bonne qualité des données au cours du temps et ainsi éviter le data drift

## Validation des Segments

Tester les segments avec de nouvelles données pour assurer leur pertinence

## Personnalisation des Recommandations

Adapter les campagnes marketing et améliorer l'expérience client en fonction des personas