PROJET 5 – « SEGMENTEZ DES CLIENTS D'UN SITE E-COMMERCE »

Soutenance de projet

2023

Sommaire

I - PROBLÉMATIQUE

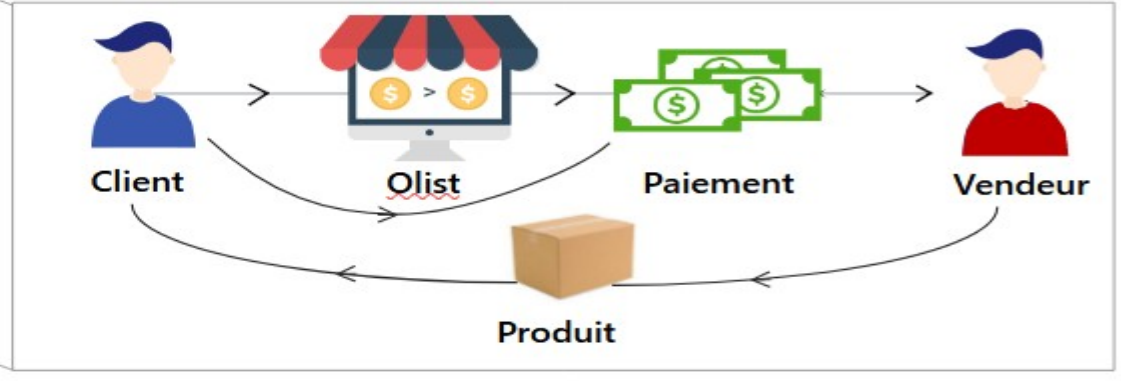
Rappel de la problématique

Interprétation

Pistes de recherche envisagées

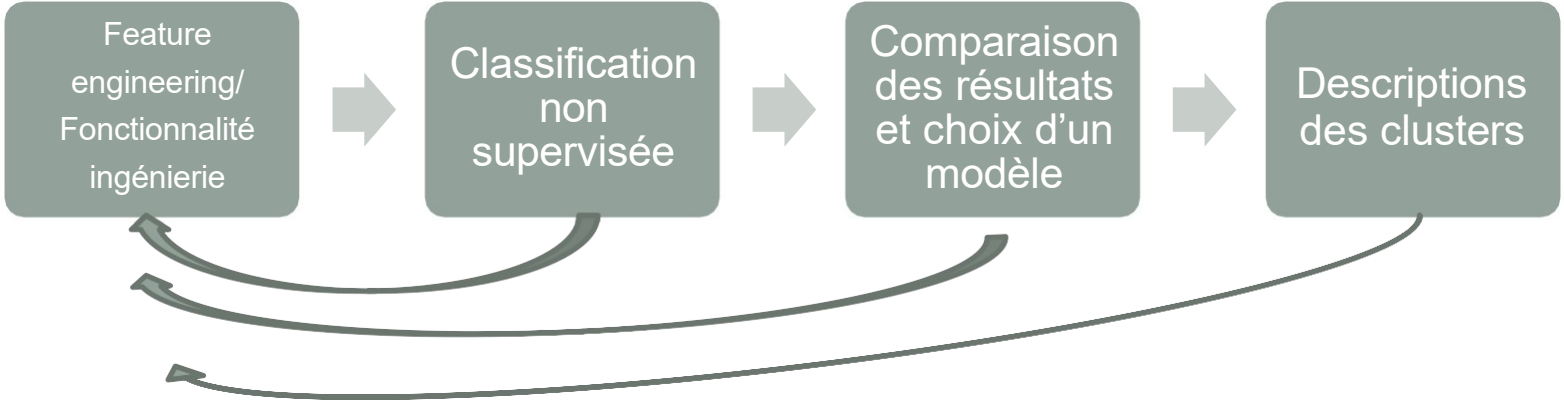
Présentation de la problématique

- Olist : plateforme e-commerce crée en 2016.
- Propose une vitrine en ligne qui met en relation les vendeurs et clients avec suivi de paiement et de livraison, et avec un système de notation.
- Mission de consultant pour Olist (solution de vente en ligne).
- Objectifs:
  - Fournir aux équipes d'e-commerce une segmentation des clients pour les campagnes de communication
  - Comprendre les différents types d'utilisateurs
  - Fournir une description utilisable de la segmentation



Interprétation de la problématique et pistes de recherche envisagées

- Exploration des données et choix de features adaptées
- Problème de classification non supervisée
- Les clusters devront être explicables et réutilisables pour des campagnes de communication



Problématique
Feuille de route

- **Extraire les données** de la base de donnée Olist permettant de caractériser les clients
- Utiliser des outils de machine learning non supervisés pour réaliser un **partitionnement des clients** en fonction de ces caractéristiques
- **Interpréter les segments** obtenus d'un point de vue métier
- Analyser la stabilité temporelle du processus de partitionnement pour **évaluer une fréquence de maintenance**

II – PRÉPARATION DU JEU DE DONNÉES

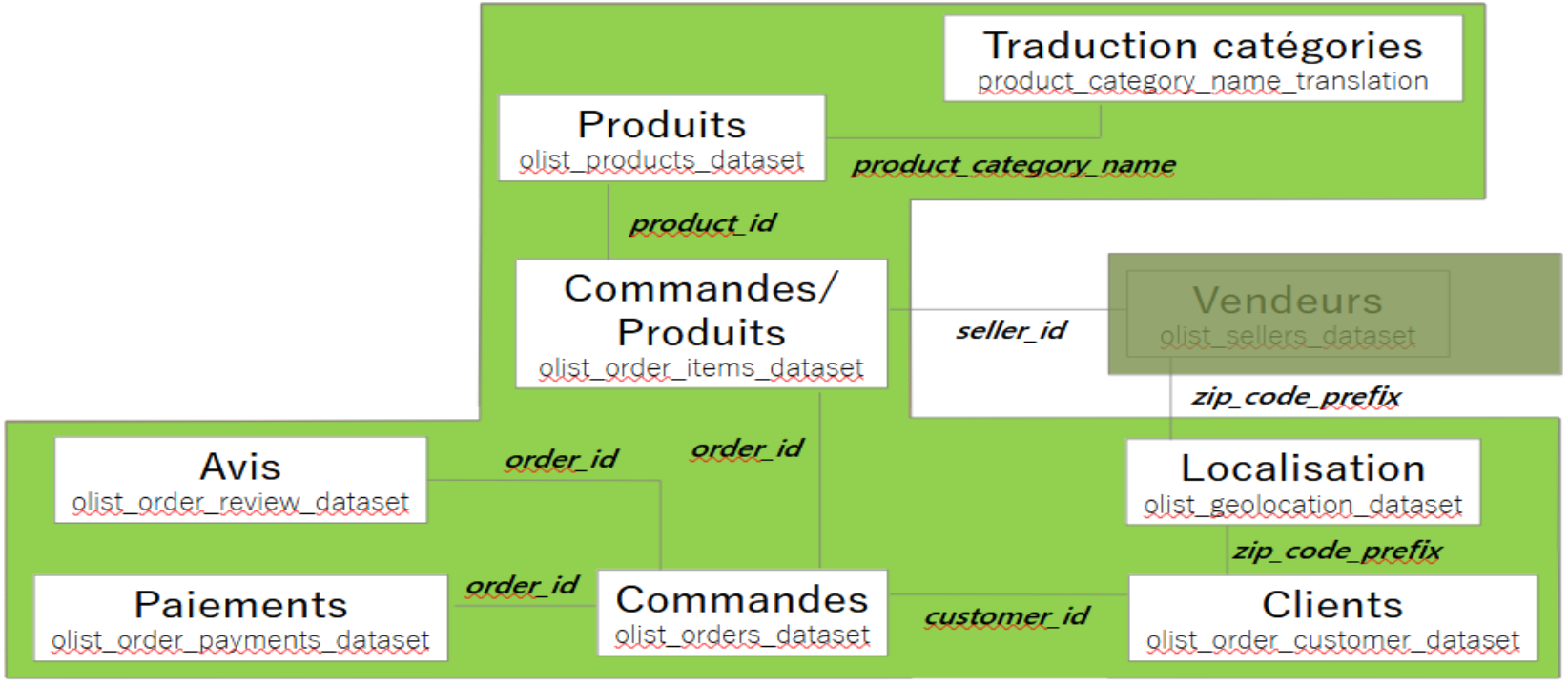
Cleaning

Feature engineering

Exploration

Données

Structure de la base de données



Cleaning

- Données réparties en 9 tables:

clients / geolocalisation / commandes / paiements / produits / vendeurs / traduction des catégories de produits

Principales étapes du nettoyage
• Imputation pour les informations manquantes
• Types de données
• Suppression des outliers univariés et multivariés
• Création de nouvelles features
• Assemblage dans une table unique avec pour index l'id client

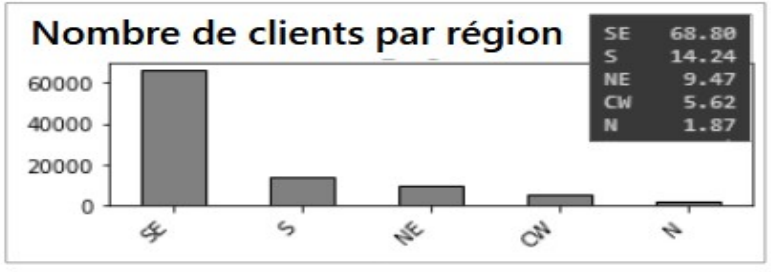
Feature engineering

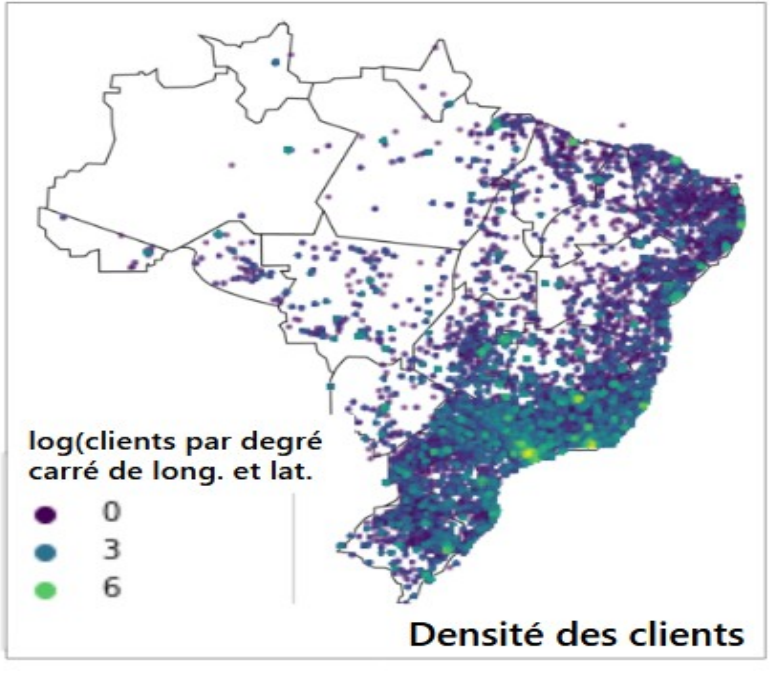
- **Intuitions : Indicateurs généraux par client**
  - Date du dernier achat : récence
  - Fréquence d'achat : fréquence
  - Total Dépensé par le Client : montant
  - Prix Moyen du Panier
  - Nombre de Produits Achetés
  - Catégorie la plus achetée
  - Note moyenne
  - Volume Total des Produits Achetés
  - Évaluation Moyenne par Client
  - Nombre de Commentaires
  - Type de Paiement Préféré
  - Satisfaction Client

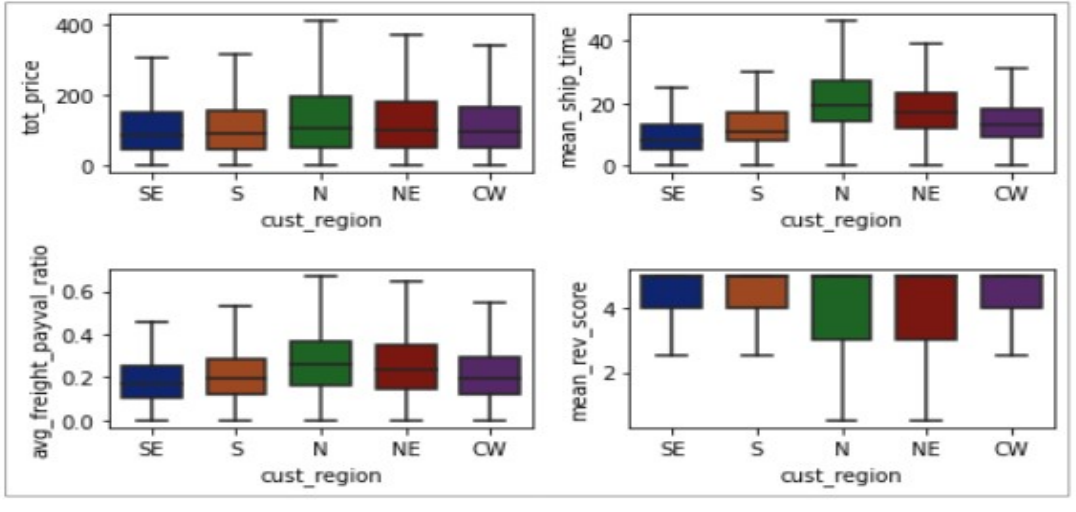
Jeu final

- 95260 lignes
- 12 colonnes

Exploration

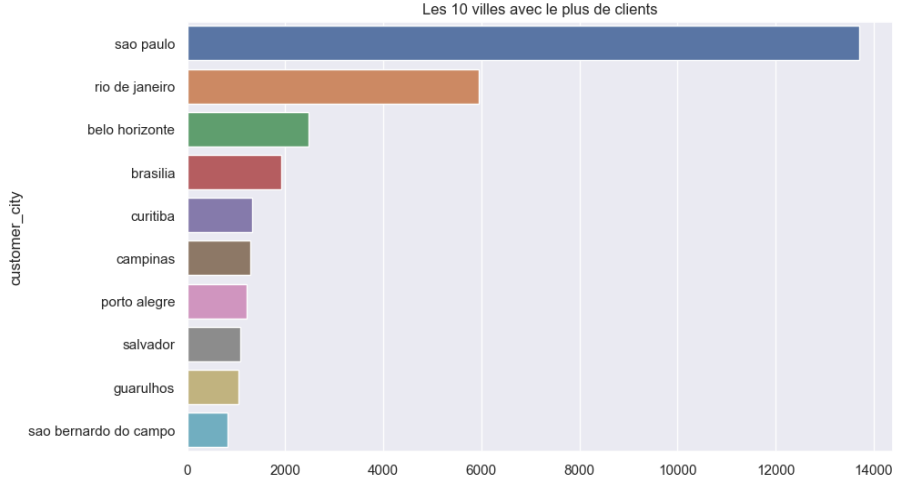


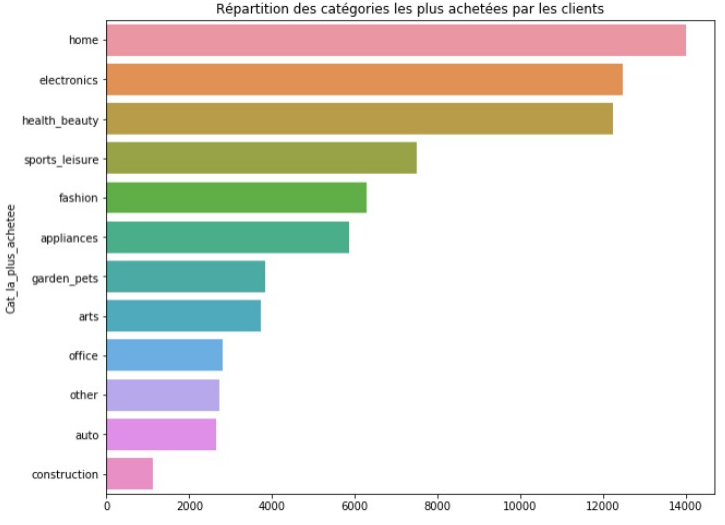


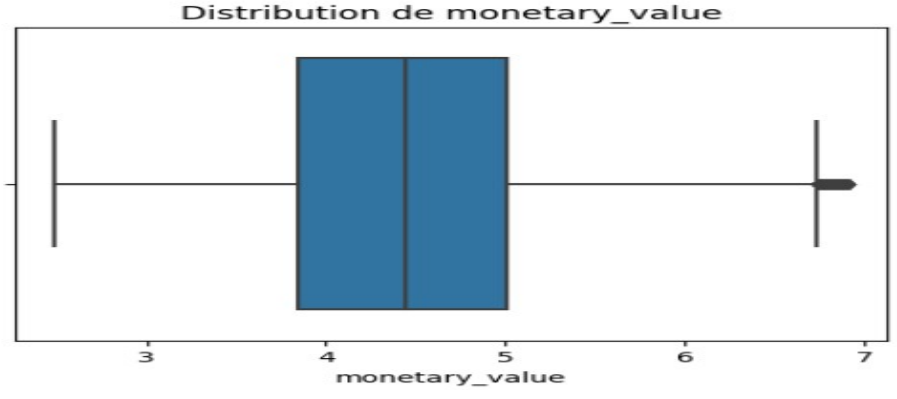


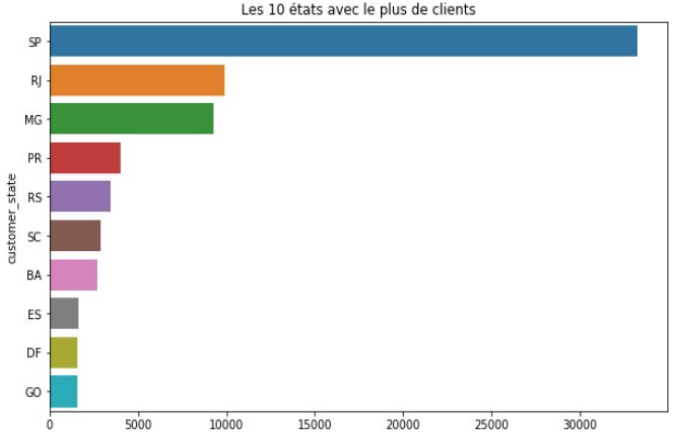
- La région du **sud-est** est sur-représentée (**70%**)
- La région a un effet sur le **temps de livraison**, sur la **satisfaction** et sur les **frais de port**
- Assez peu d'effet sur la **valeur** des produits commandés

Exploration





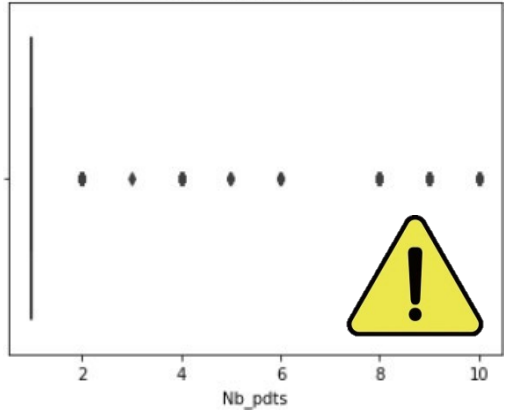




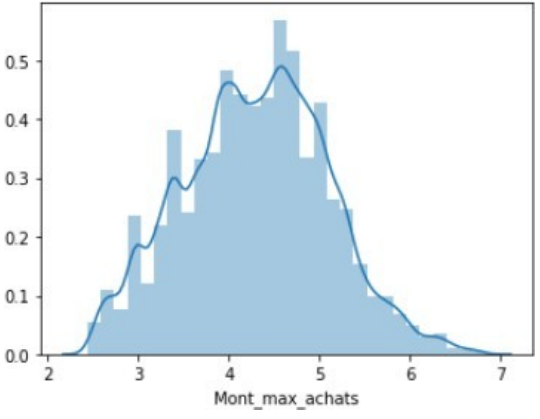
Exploration

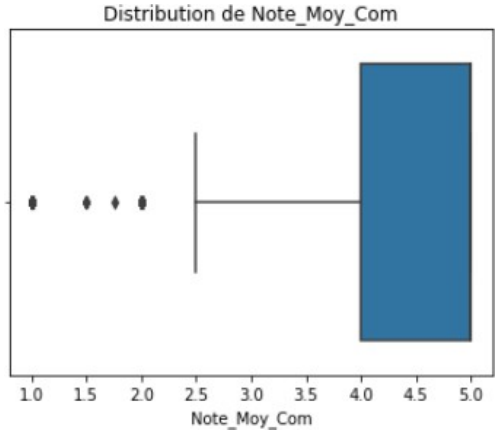
Autres informations:
- 2 années d'historique
- 100 000 clients

Nombre de produits achetés par client

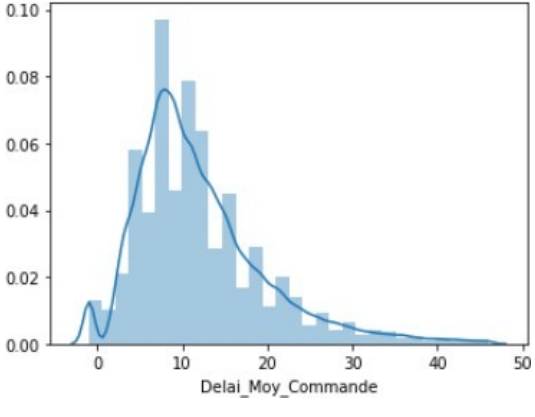


Distribution du montant maximum dépensé par commande

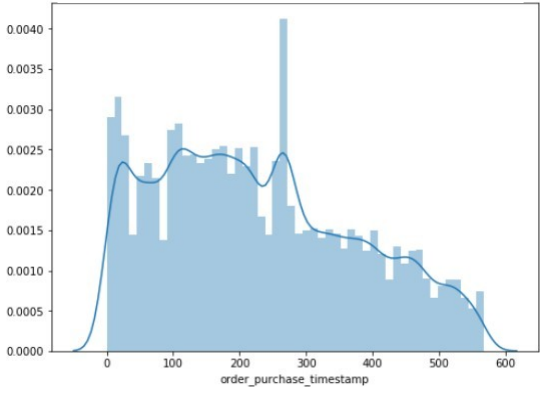




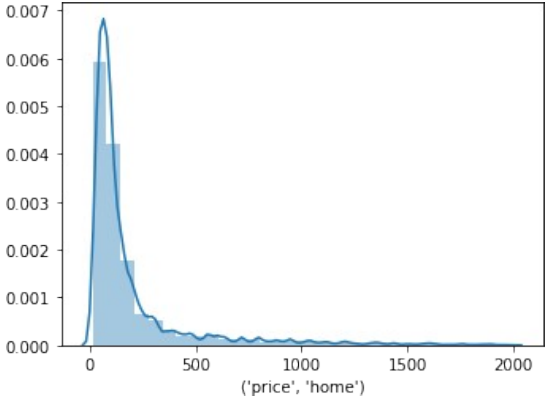
Délai moyen de commande en jours



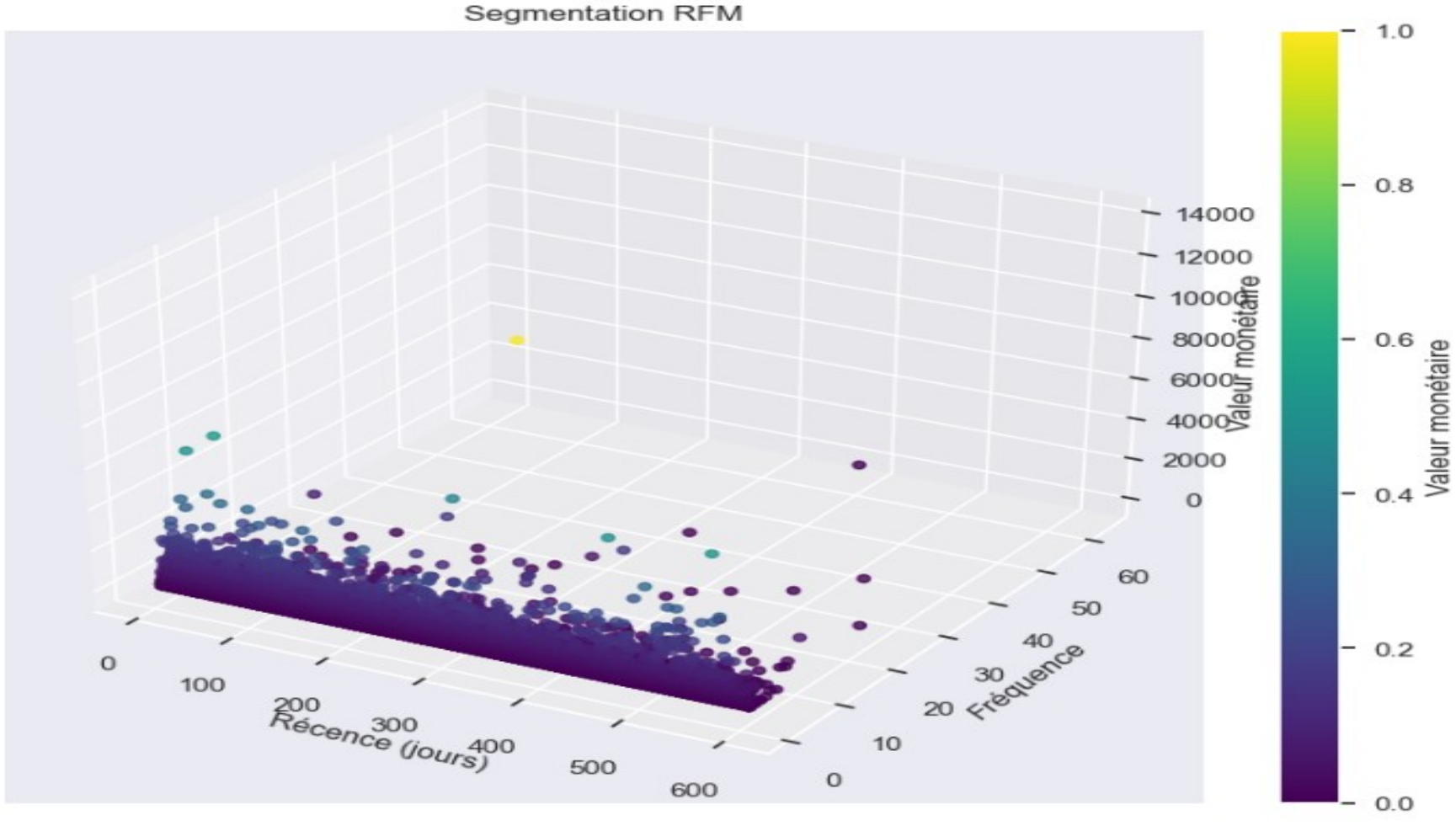
Nombre de jours écoulés depuis la dernière commande

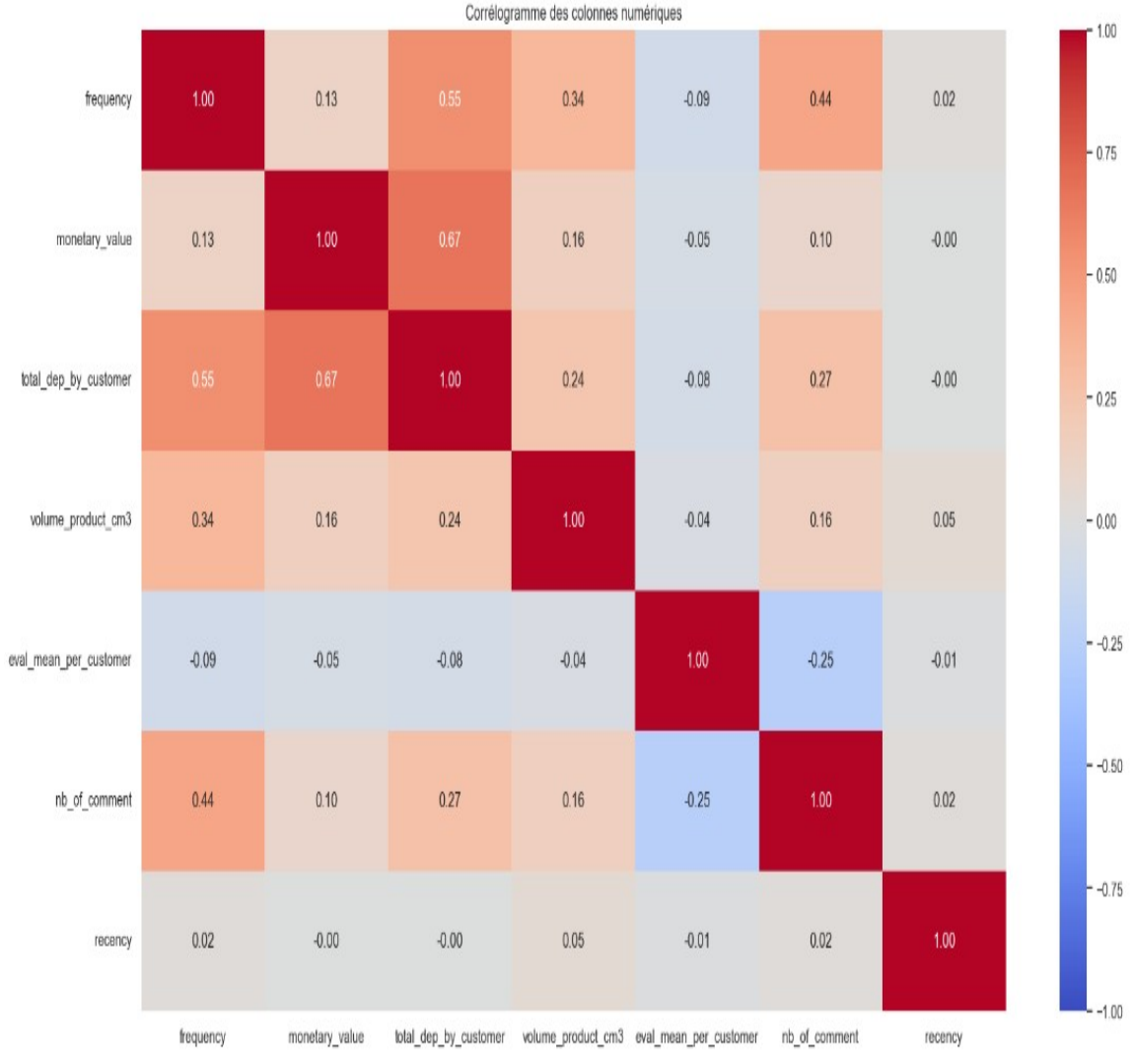


Distribution des dépenses dans la catégorie « home »



Segmentation RFM





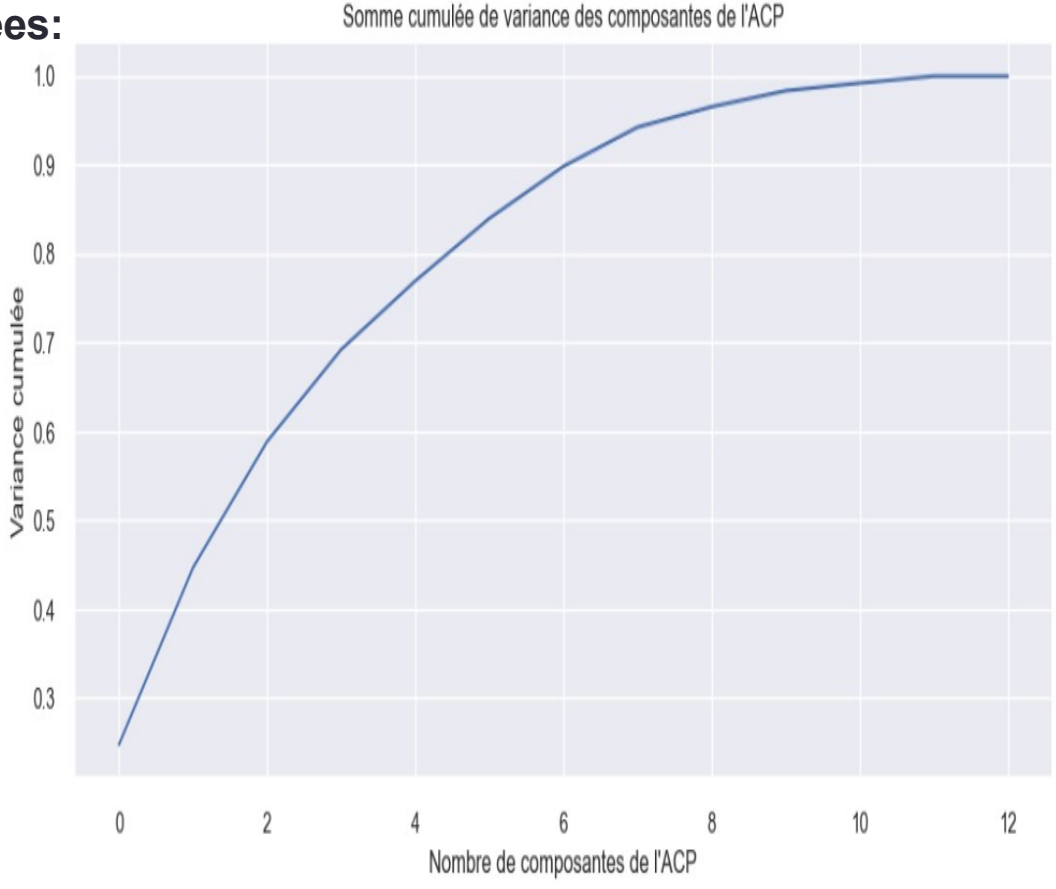
Exploration : Corrélations

- **Très peu de corrélation entre les variables**

⇒ Non pertinence de certaines features dues au jeu de données comprenant plus de 95 % de clients avec un seul achat

Préparation des données : Finalisation

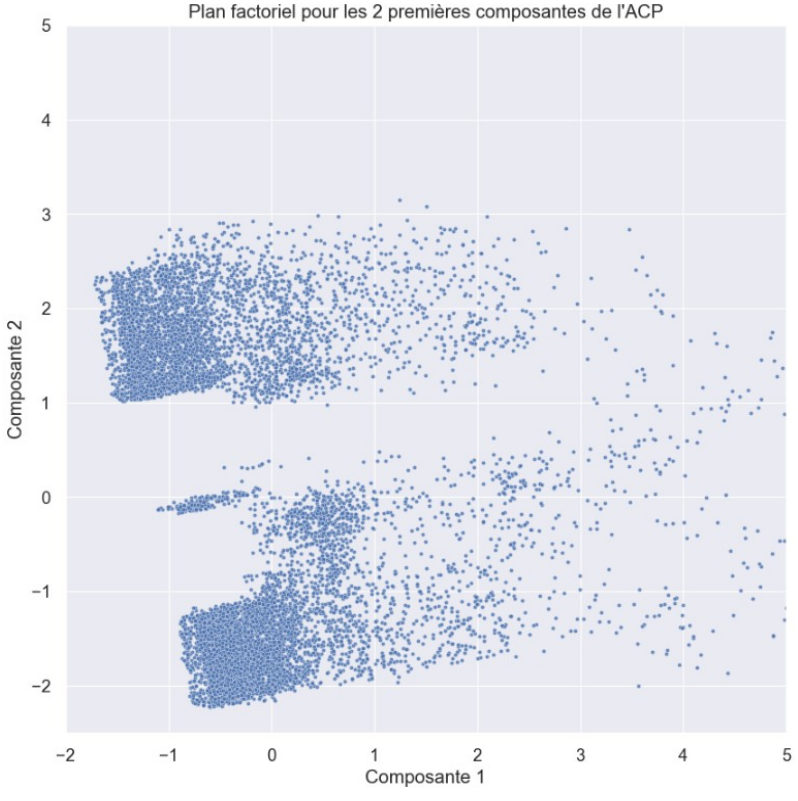
- **Suppression des features non adaptées:**
  - Montant maximum d'une commande
  - Nombre moyen de produits par commande
- **Préparation**
  - One hot encoder
  - StandardScaler
- **Réduction de dimension par ACP**
  - Réduction à 7 features avec 94 % de variance



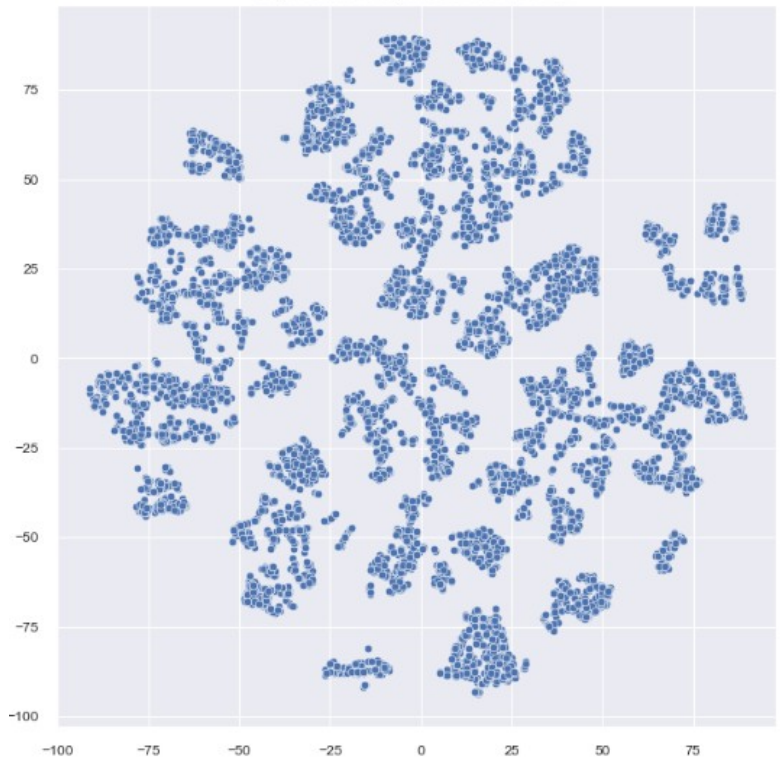
III – PISTES DE MODÉLISATIONS

Intuition : Visualisation

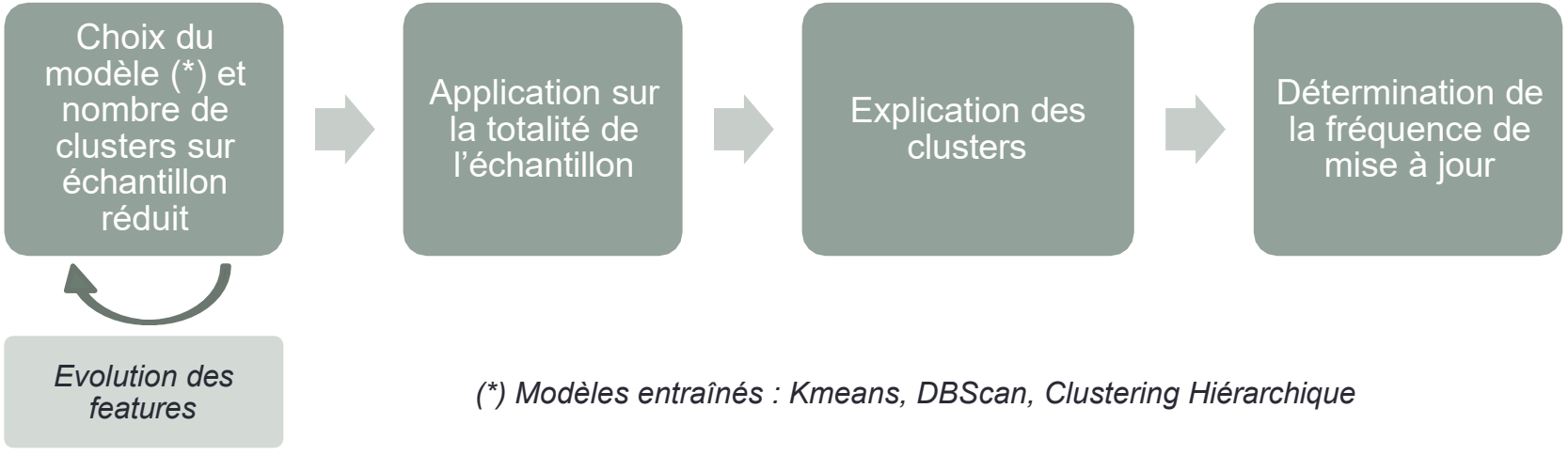
Projection des données sur les 2 premières composantes de l'ACP



Représentation des données via t-SNE



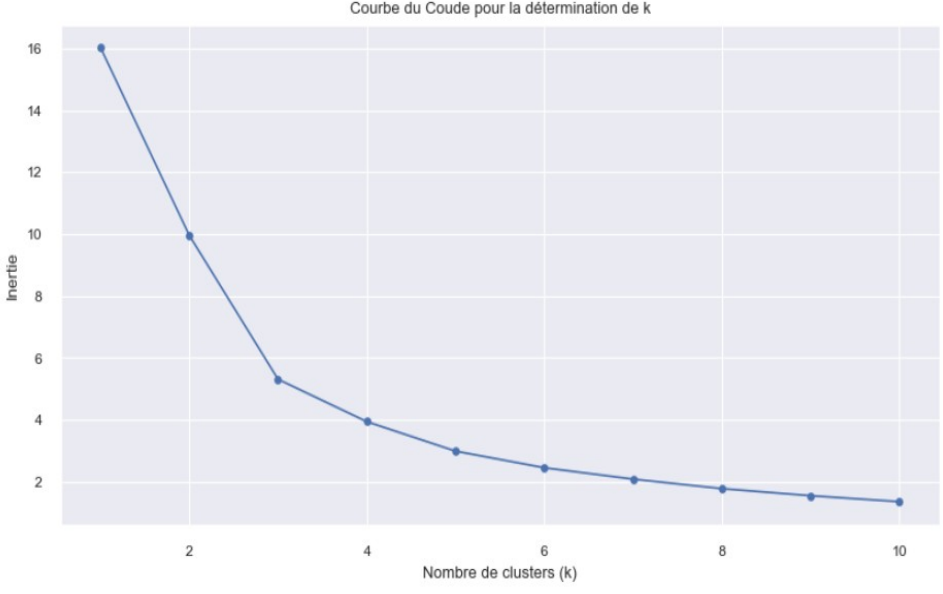
Processus de clustering



(*) Modèles entraînés : Kmeans, DBScan, Clustering Hiérarchique

Détermination optimum du nombre de clusters

- Entraînement de modèles avec 1 à 10 clusters

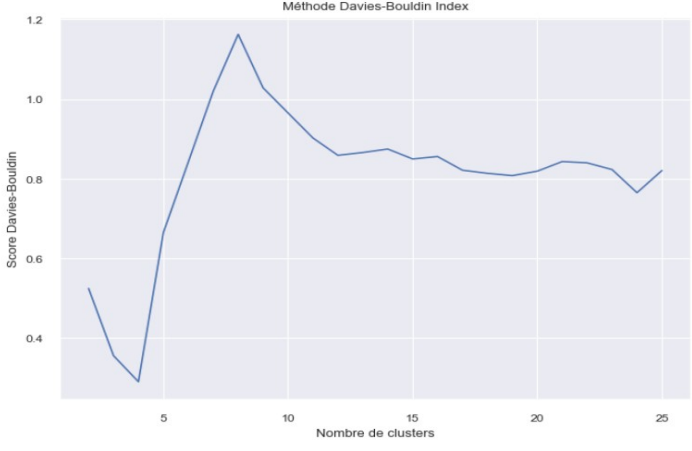


« Méthode du coude »

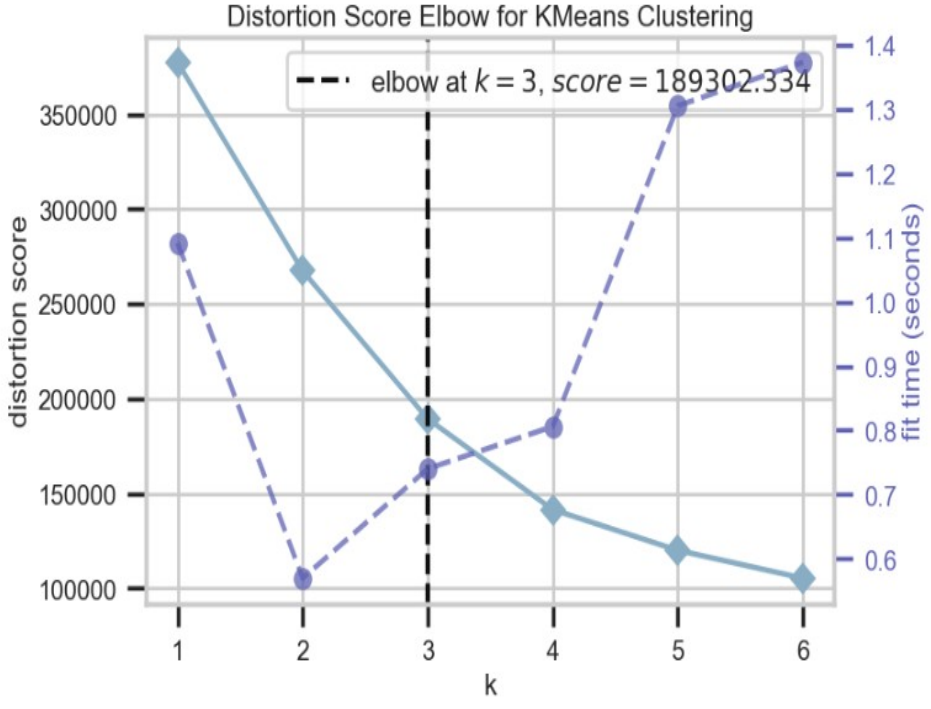
Optimum retenu : 3 clusters

Coefficient de silhouette en fonction du nb de clusters

Davies bouldin score en fonction du nb de clusters



La méthode Elbow



- La méthode elbow donne un nombre de cluster égale à 3

GridSearchCV

- La méthode GridSearchCV donne comme meilleurs paramètres trouvés : {'init': 'k-means++', 'max_iter': 100, 'n_clusters': 3}

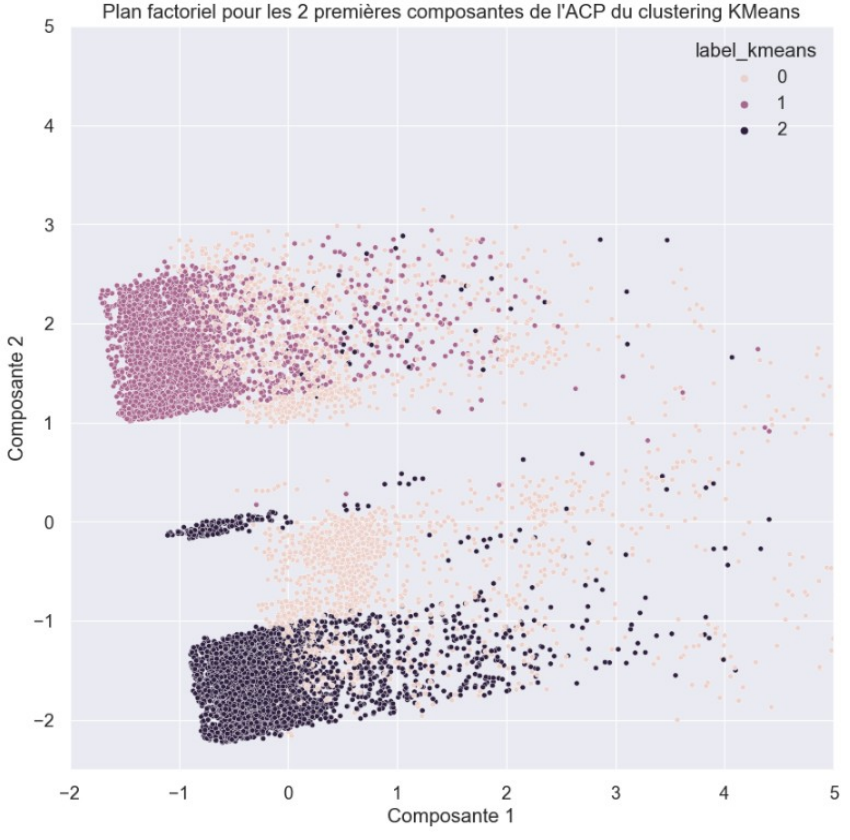
Catégorie client

- 1. **Les Bons Clients** : Ce sont ceux qui ont les valeurs les plus élevées pour les caractéristiques associées à un bon comportement d'achat (par exemple, recency faible, frequency élevée, monetary élevée, etc.).

- 2. **Clients Moyens** : Ce sont ceux qui ont des valeurs proches de la moyenne pour la plupart des caractéristiques.

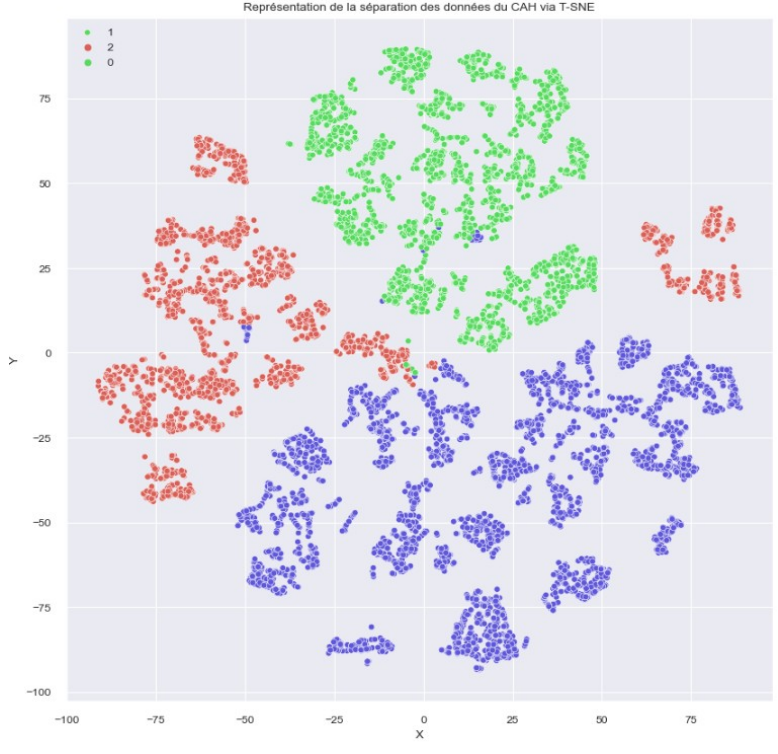
- 3. **Mauvais Clients** : Ce sont ceux qui ont des valeurs inférieures à la moyenne pour la plupart des caractéristiques.

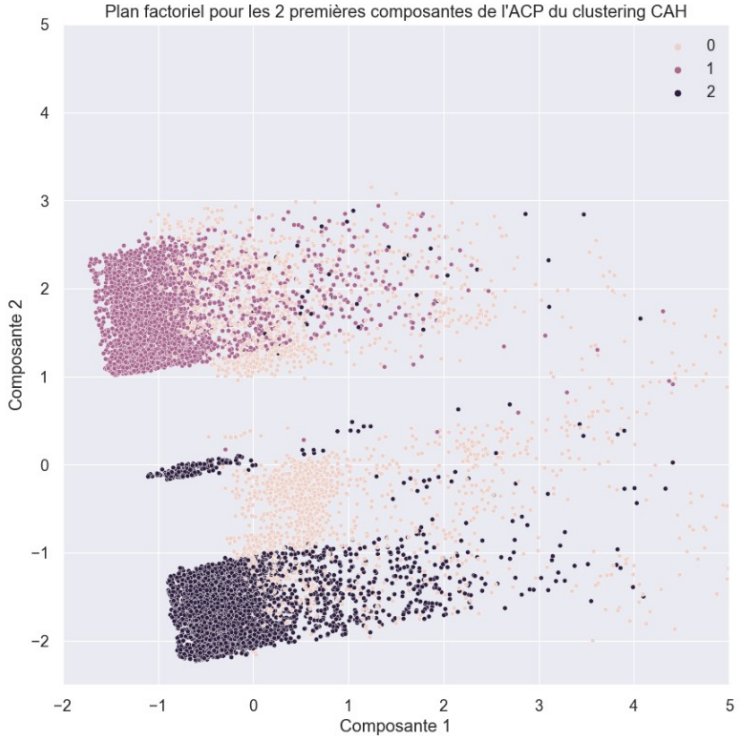
Kmeans : représentation graphique



Clustering hiérarchique

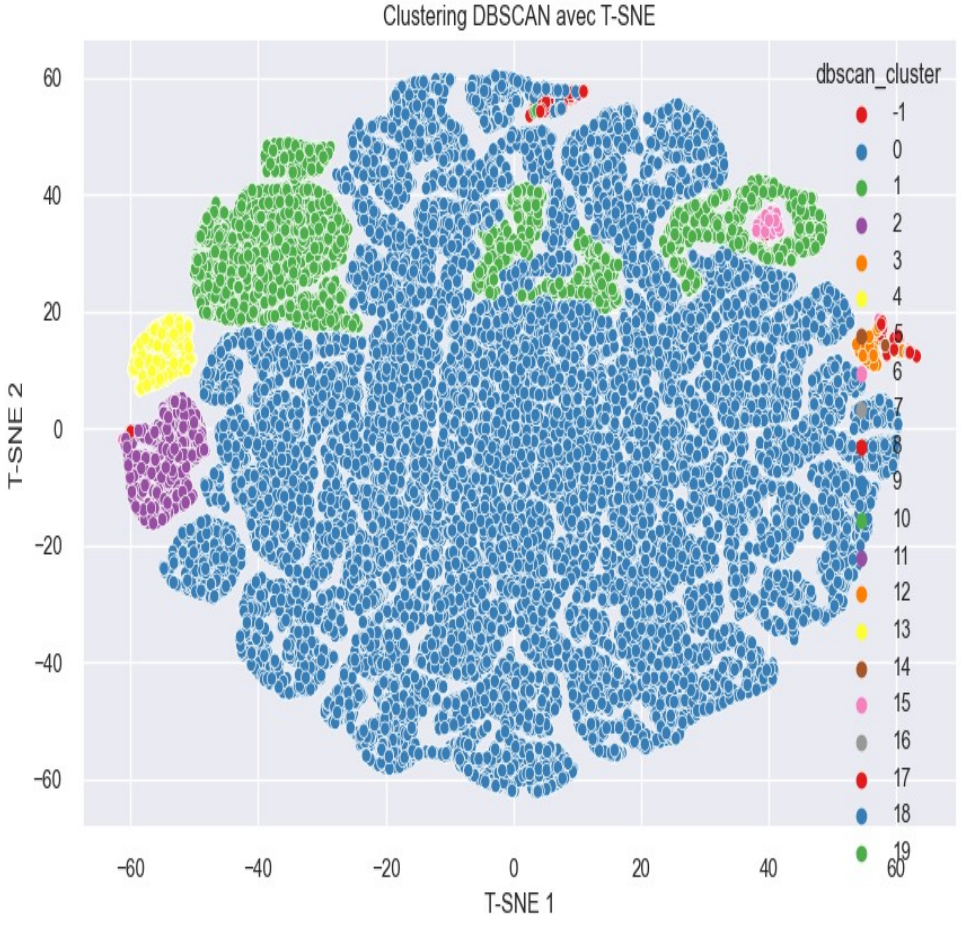
- Avec 3 clusters





DBScan

- Exemple ci-contre:
  - Epsilon = 1
  - Min_samples = 5



Choix de la méthode : Kmeans

Connaissance du nombre de clusters, commencez par K-Means.
K-Means peut être plus efficace car vous pouvez spécifier ce nombre.

Si l'interprétation des clusters en termes de centres est importante, K-Means peut être préférable en raison de sa simplicité d'interprétation.

K-Means est souvent plus rapide

Définition : Kmeans

Le K-Means est un algorithme de clustering utilisé pour regrouper un ensemble de données en clusters ou en groupes similaires. Voici comment il fonctionne de manière simple et concise :

1. **Initialisation** : Sélectionnez aléatoirement K points comme centres de cluster initiaux, où K est le nombre de clusters souhaité.

2. **Assignation** : Pour chaque point de données, attribuez-le au cluster dont le centre est le plus proche. Cela se fait généralement en utilisant la distance euclidienne.

3. **Mise à jour du Centre** : Recalculez le centre de chaque cluster en utilisant la moyenne des points qui lui sont assignés.

4. **Répétez** : Répétez les étapes 2 et 3 jusqu'à ce que les centres des clusters ne changent que très peu ou que le nombre maximum d'itérations soit atteint.

5. **Résultat** : Vous obtenez ainsi K clusters, chacun ayant ses propres points de données similaires.

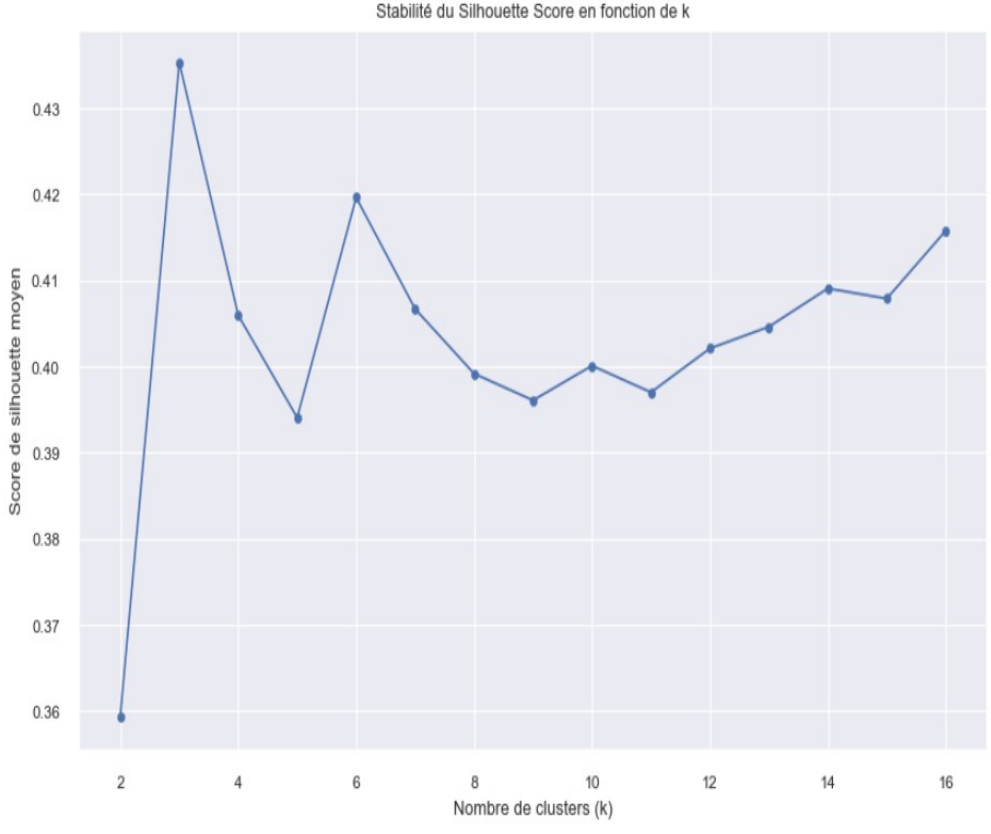
L'objectif principal du K-Means est de regrouper les données de manière à ce que les points à l'intérieur de chaque cluster soient similaires les uns aux autres tout en étant différents des points des autres clusters. Il est largement utilisé dans divers domaines, notamment la segmentation de la clientèle, la détection d'anomalies, la compression d'images et l'analyse de données, pour découvrir des structures cachées dans les données et prendre des décisions basées sur ces regroupements.

IV –PRÉSENTATION DU MODÈLE FINAL

ainsi que des améliorations effectuées.

Kmeans

Stabilité du silhouette



Stabilité du silhouette score

- ...
- 2 clusters : 0.359
- **3 clusters : 0.435**
- 4 clusters : 0.405
- ...

Clusters identifiés : améliorations?

- Suppression de certaines features : pas d'amélioration du clustering constatée

- Proposition de clustering « manuel »:
  - clients insatisfaits (note moyenne de 1/5)
  - clients avec un très long délai de livraison (> 1 mois)
  - 4600 clients qui achètent le plus et sous clustering par application d'un kmeans :
    - Identification des meilleurs clients par catégorie
    - Identification des catégories complémentaires pour cibler la publicité

Définition : ARI

L'ARI (Adjusted Rand Index) est une mesure de similarité entre deux ensembles de données étiquetés, souvent utilisée pour évaluer la qualité d'un algorithme de clustering. Voici comment il fonctionne de manière simple et concise :

Comment ça marche :
1. L'ARI commence par comparer chaque paire de points dans les ensembles étiquetés et mesure combien de paires se trouvent dans le même cluster à la fois dans les ensembles réels et prédits par l'algorithme de clustering.

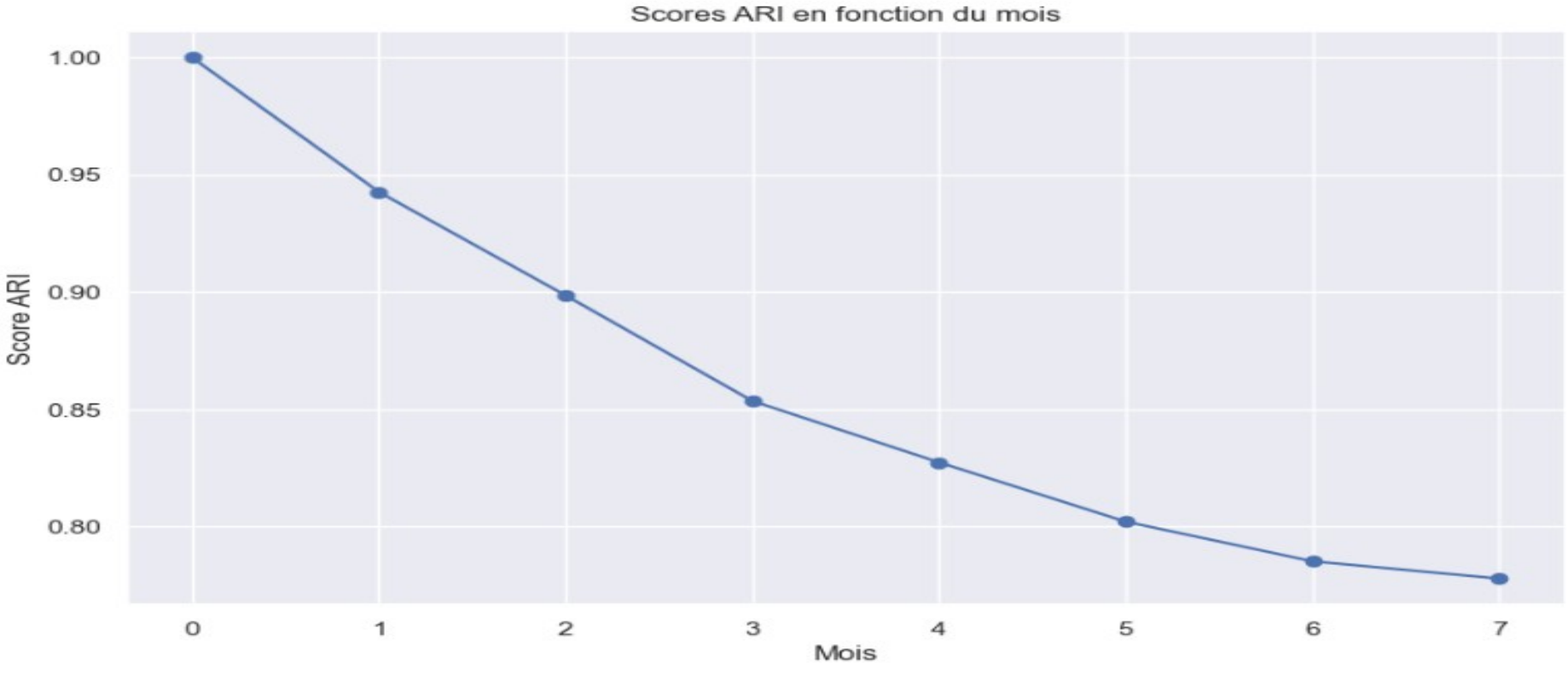
2. Il calcule également le nombre de paires de points qui se trouvent dans des clusters différents à la fois dans les ensembles réels et prédits.

3. Ensuite, l'ARI combine ces informations pour calculer une mesure de similarité qui varie de -1 (pas du tout similaire) à 1 (très similaire).

À quoi ça sert :
L'ARI sert à évaluer la qualité des clusters produits par un algorithme de clustering en comparant ses résultats avec un ensemble de données étiqueté. Plus l'ARI est proche de 1, meilleure est la qualité du clustering, car cela signifie que les clusters créés correspondent étroitement aux véritables groupes dans les données.

En résumé, l'ARI est un moyen de mesurer à quel point un algorithme de clustering est performant en comparant ses résultats aux groupes réels. C'est utile pour évaluer la précision d'un modèle de clustering et choisir le meilleur algorithme ou le meilleur nombre de clusters pour vos données.

Courbe du Score ARI en fonction du mois



- Le score ari pourrait nous indiquer que la période de maintenance est de 3 mois environs

Contrat de maintenance

- Identification de la période de maintenance:
  - Réduction du jeu de données sur la dimension « durée » (exemple : 3 mois)
  - Vérification de la stabilité du nombre de clusters, du coefficient de silhouette et des valeurs des features

- Compromis identifié : 3 mois
  - Nombre de clusters optimal sur Kmeans : 3
  - Coefficient de silhouette stable
  - Conservation des caractéristiques principales des clusters (notes, etc.)

Conclusion

- Mise en application des algorithmes de classification non supervisée et application à un problème métier

- Limites du clustering proposé
  - Pas ou peu d'apport des algorithmes
  - Un cluster avec 50 % des clients, peu intelligibles

- Opportunités d'amélioration du clustering
  - Nouvelles features / clients ayant acheté plusieurs articles
  - Caractérisation dans le détail des produits des champs textuels
  - Données plus précises sur les clients (à anonymiser) : âge, sexe

MERCI DE VOTRE ATTENTION