Segmentation des clients d'un site de e-commerce

19/01/2024

olist

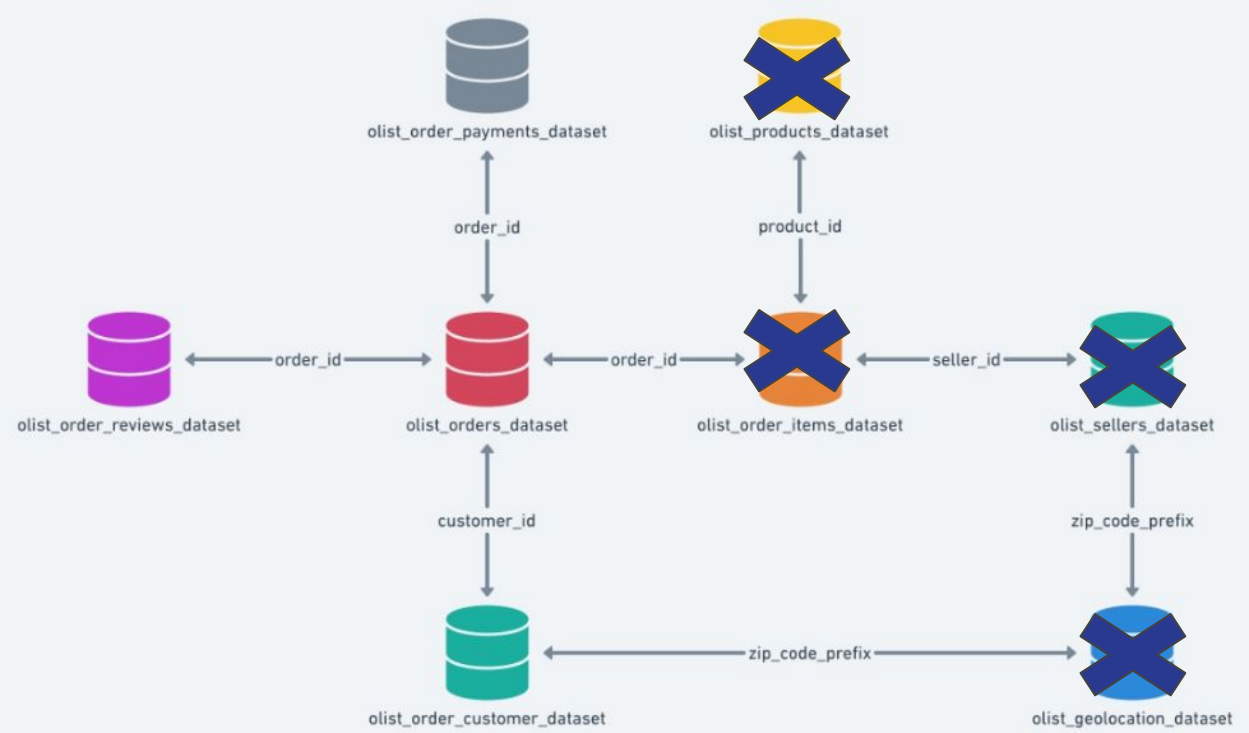
Introduction

- Objectif d'Olist: optimiser ses campagnes de communication

- Contexte:
  - Consultant Olist
  - Equipe marketing: nouvelle segmentation clients pour mieux comprendre les différents types d'utilisateurs
  - Segmentation RFM réalisée par prédécesseur

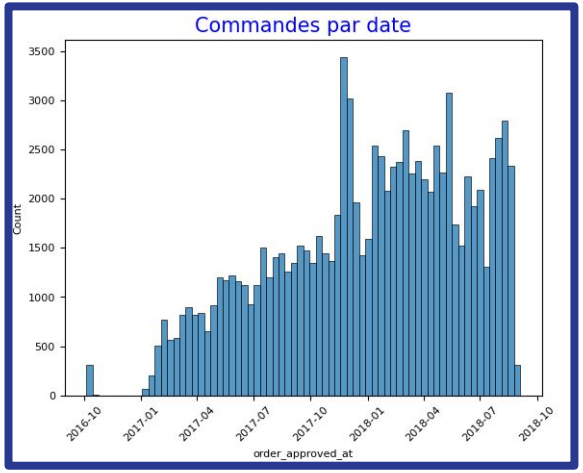
- But de cette étude:
  - Tester différentes approches de modélisation
  - Description actionable de la segmentation
  - Proposer un contrat de maintenance

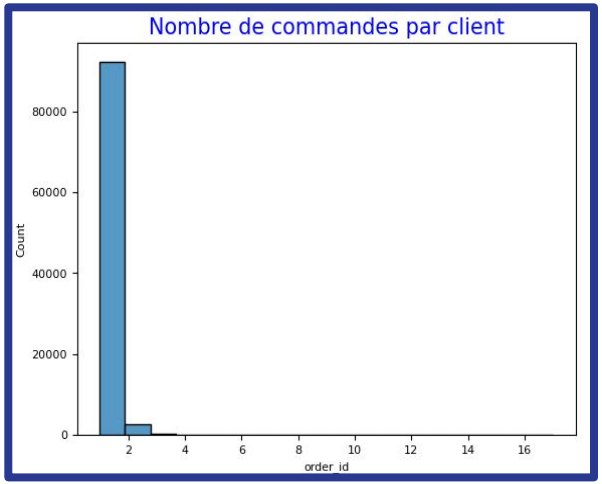
Description du jeu de données

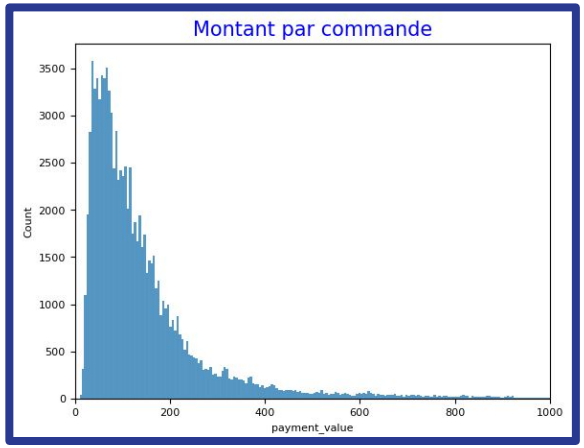


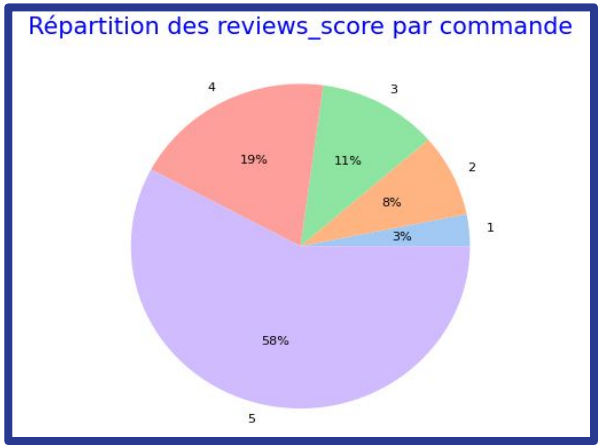
- Infos principales: commandes, produits achetés, notes de satisfaction…
- 9 fichiers csv => sélection
- Fusion nécessaire
- Nettoyage:
  - Valeurs manquantes
  - Pas de valeurs aberrantes
  - Peu de doublons

Analyse exploratoire



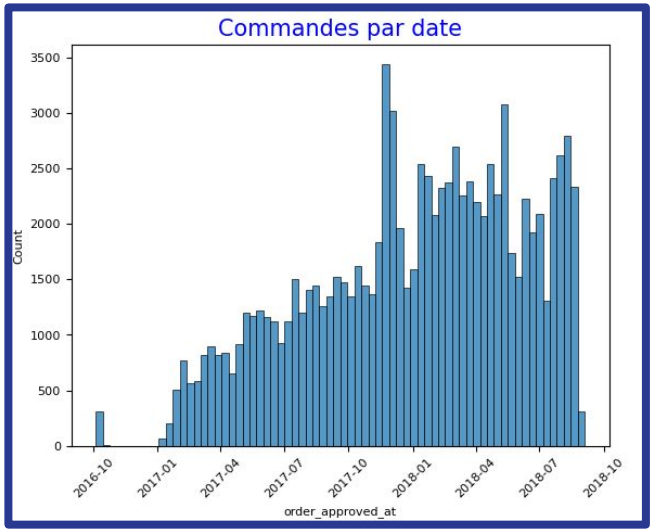




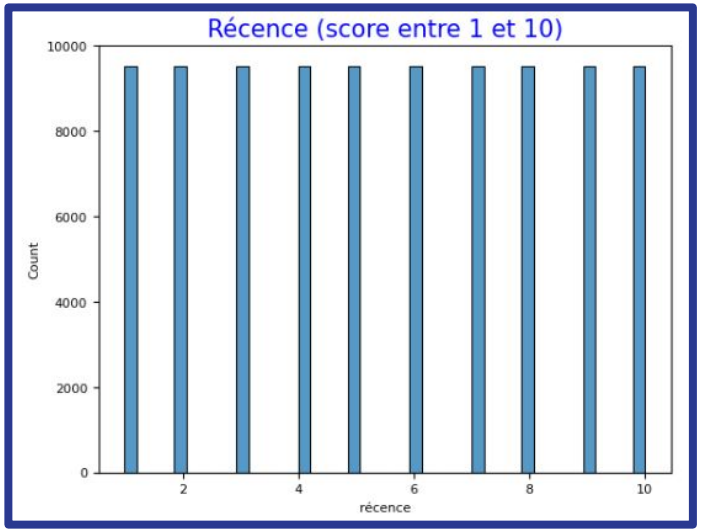


Partie 1: Feature engineering

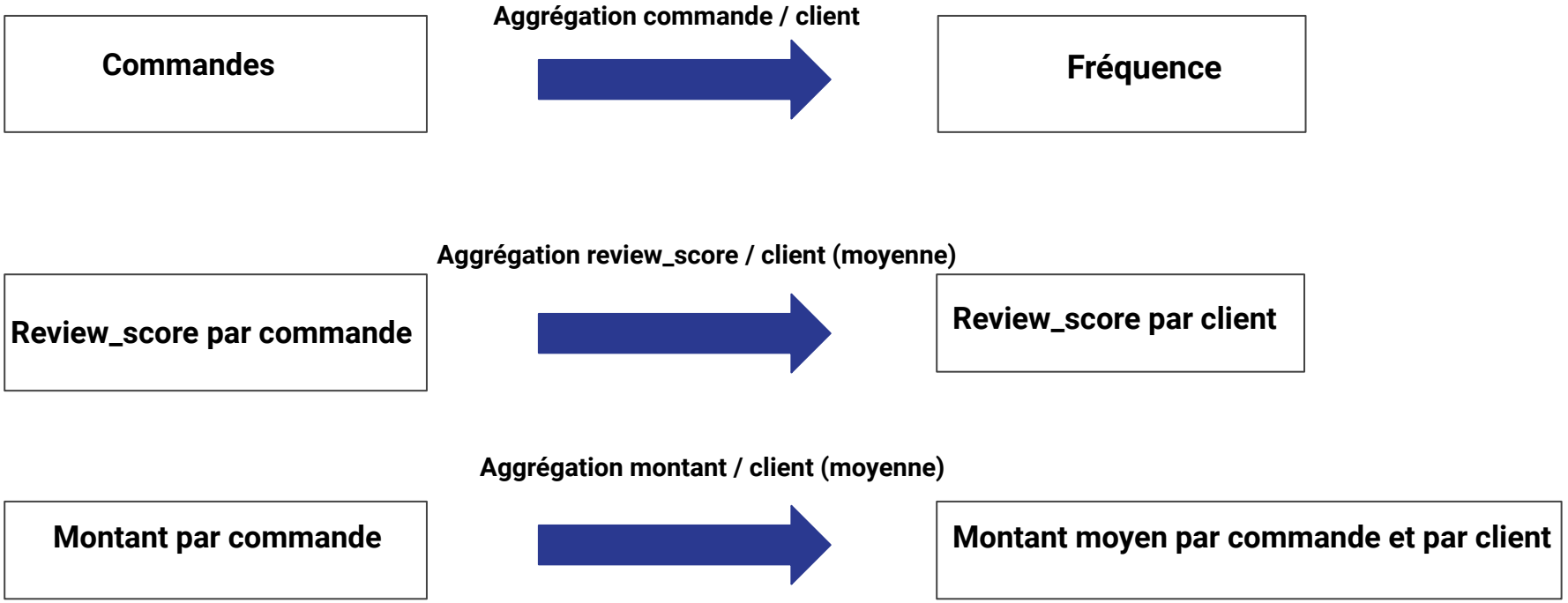
Transformation des dates d’achat => Récence



nb days, qcut (10 quantiles)

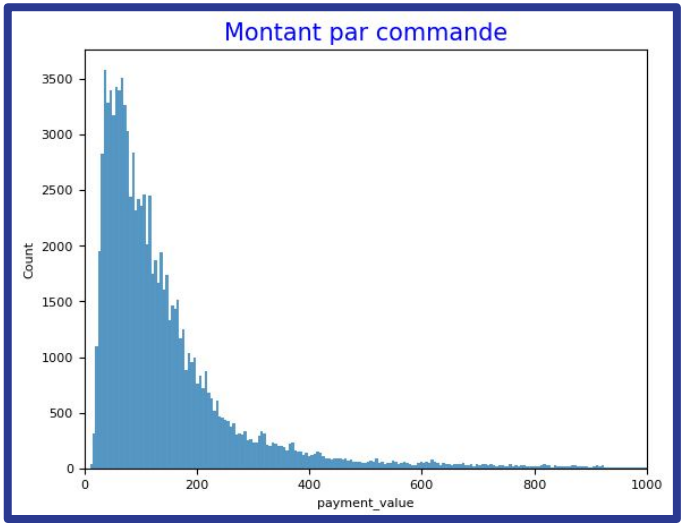


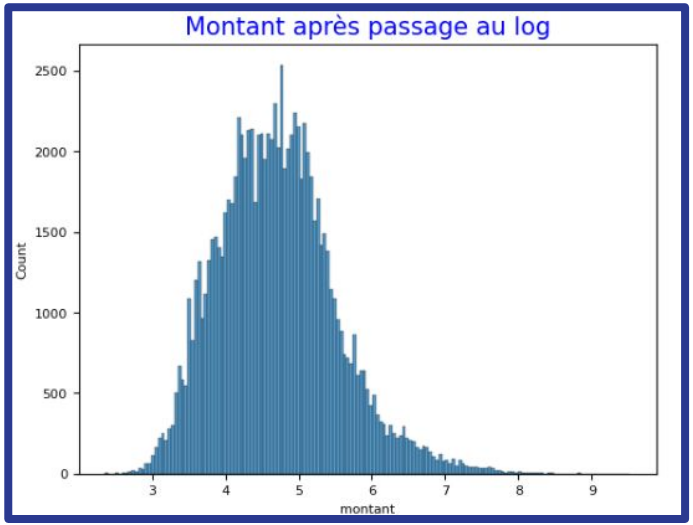
Création de features



Passage au log du montant

- Réduire l'amplitude de variables sans perte d'information
- Réduire l'influence des valeurs atypiques





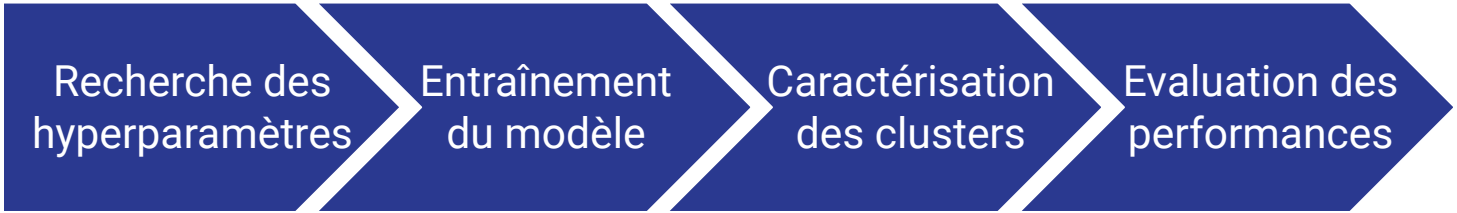
Partie 2: Essais de modélisation

Démarche

- Essais de 3 modèles de clustering non-supervisés avec RFM
  - K-means
  - Clustering hiérarchique
  - DBSCAN

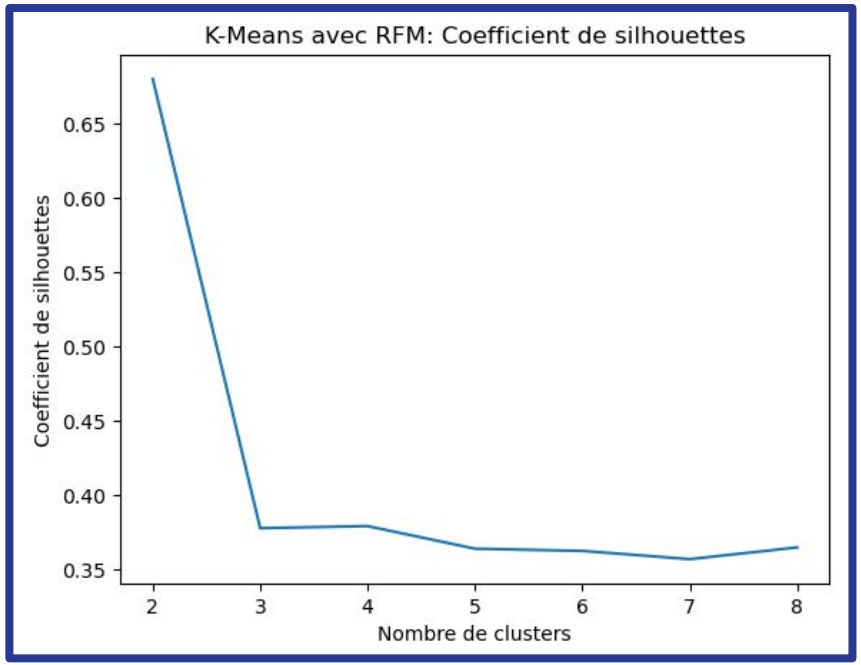
- Essais du meilleur modèle avec RFM + review_score

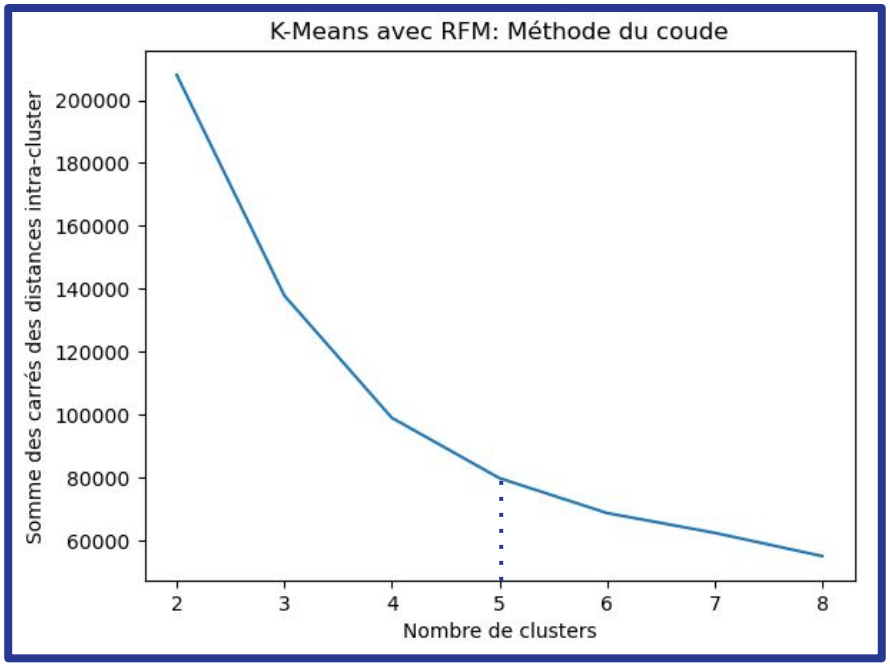
- Etapes de modélisation à répéter



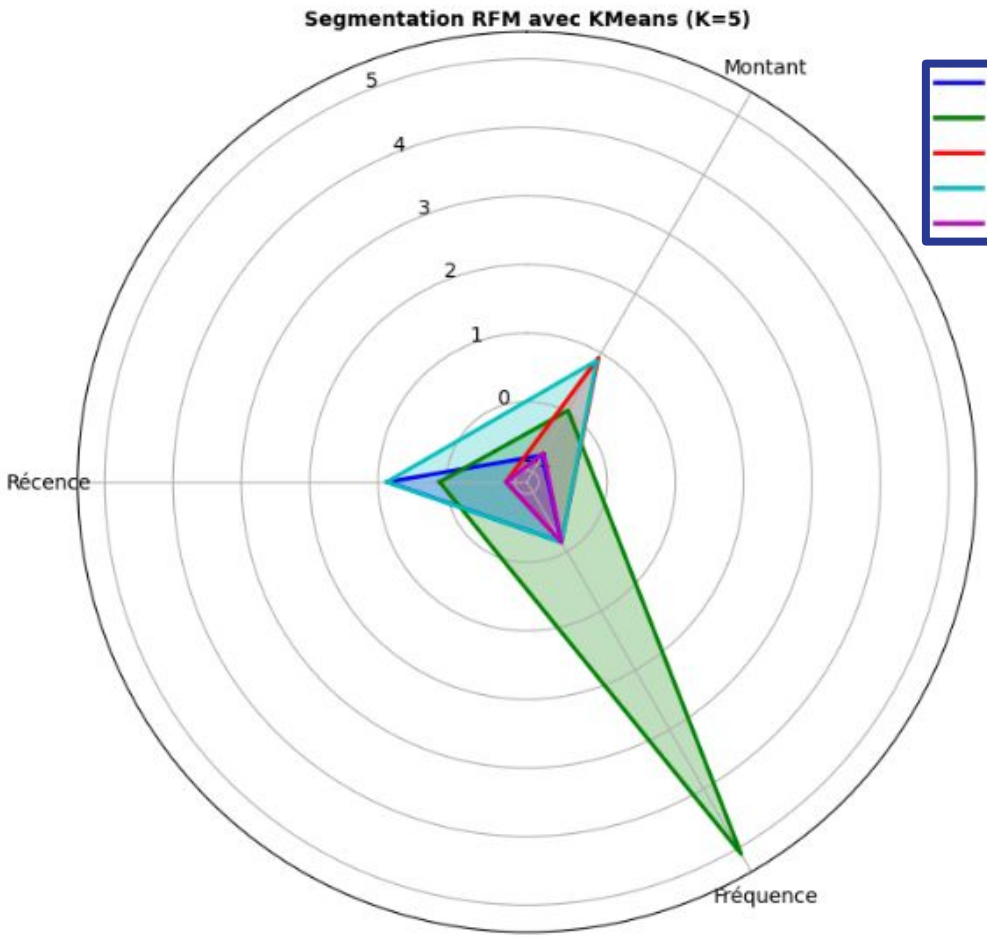
K-means avec RFM (½) - Hyperparamètres

- Qualité clustering: Homogénéité et la séparation des clusters
- Recherche du nombre optimal de clusters (K=5)
- Initialisation: k-means++





K-means avec RFM (2/2) - Caractérisation

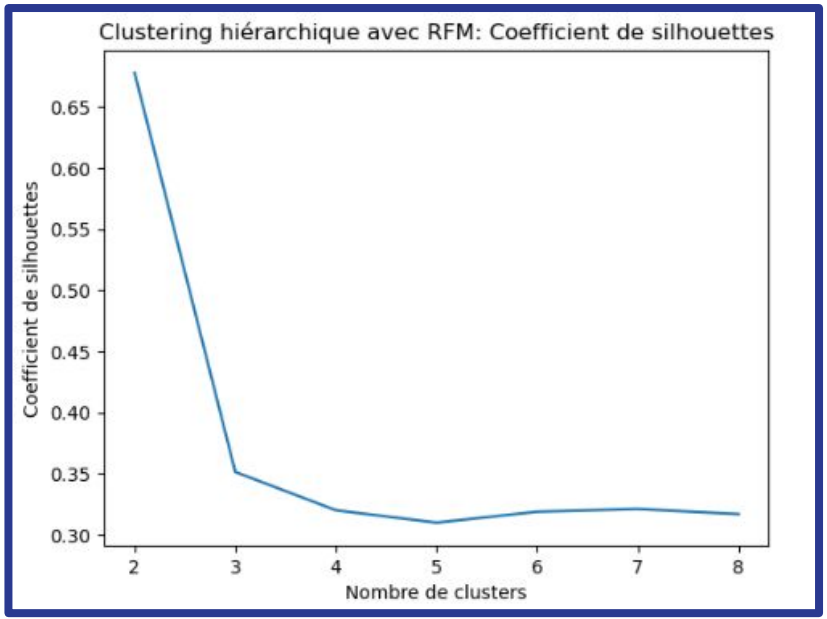


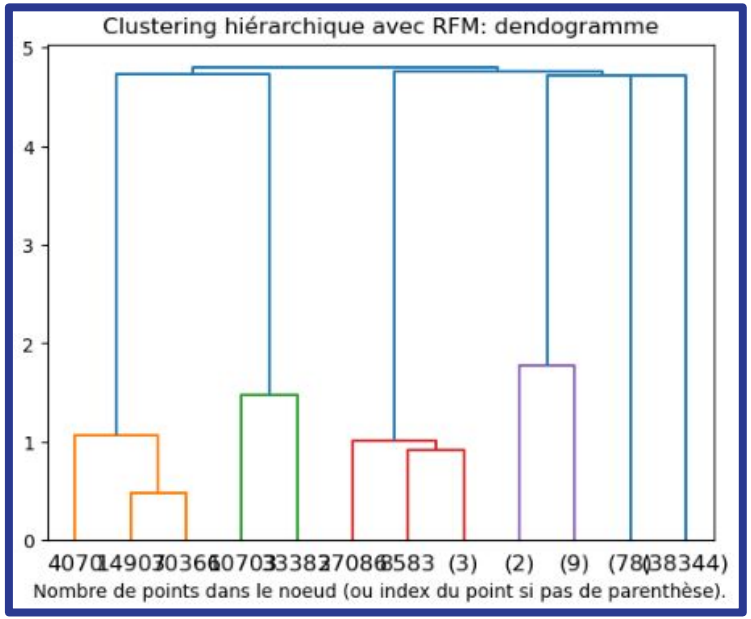
- Cluster 1: 24978 clients
- Cluster 2: 2903 clients
- Cluster 3: 19546 clients
- Cluster 4: 21070 clients
- Cluster 5: 26787 clients

LOYAUX: Montant faible, achat récent
FIDELES: Plusieurs commandes, montant moyen
A REACTIVER: Montant élevé, achat pas récent
CHAMPIONS: Montant élevé, achat très récent
PERDUS: Montant faible, achat pas récent

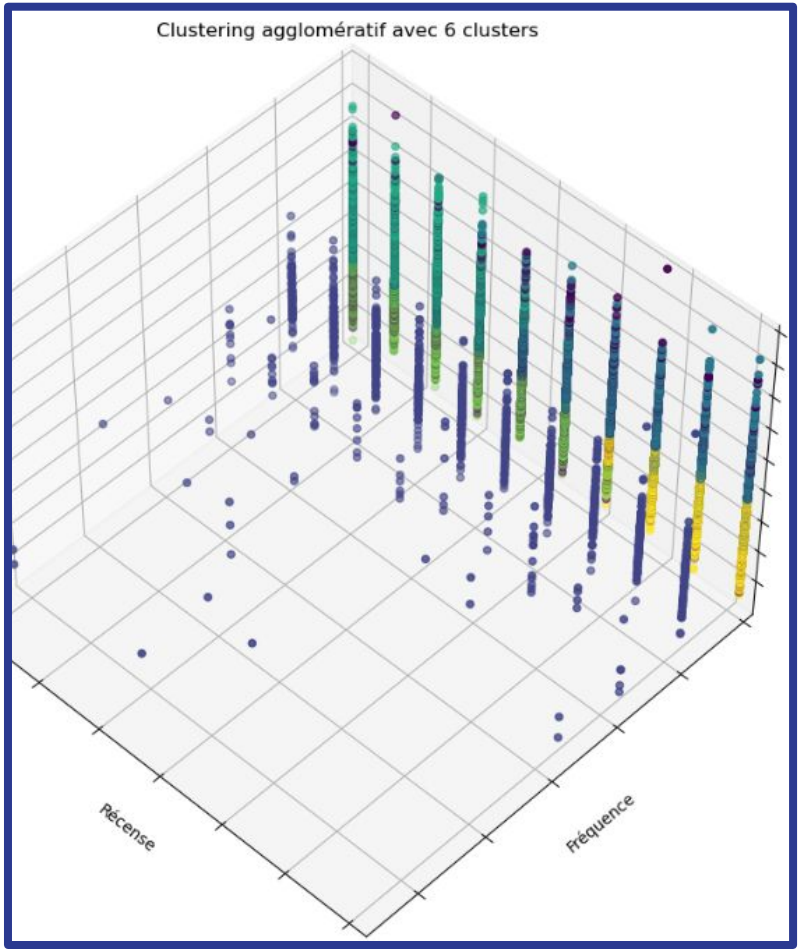
Clustering hiérarchique avec RFM (½)

- Qualité clustering: Homogénéité et la séparation des clusters
- Recherche du nombre optimal de clusters (K=6)





Clustering hiérarchique avec RFM (2/2) - Caractérisation

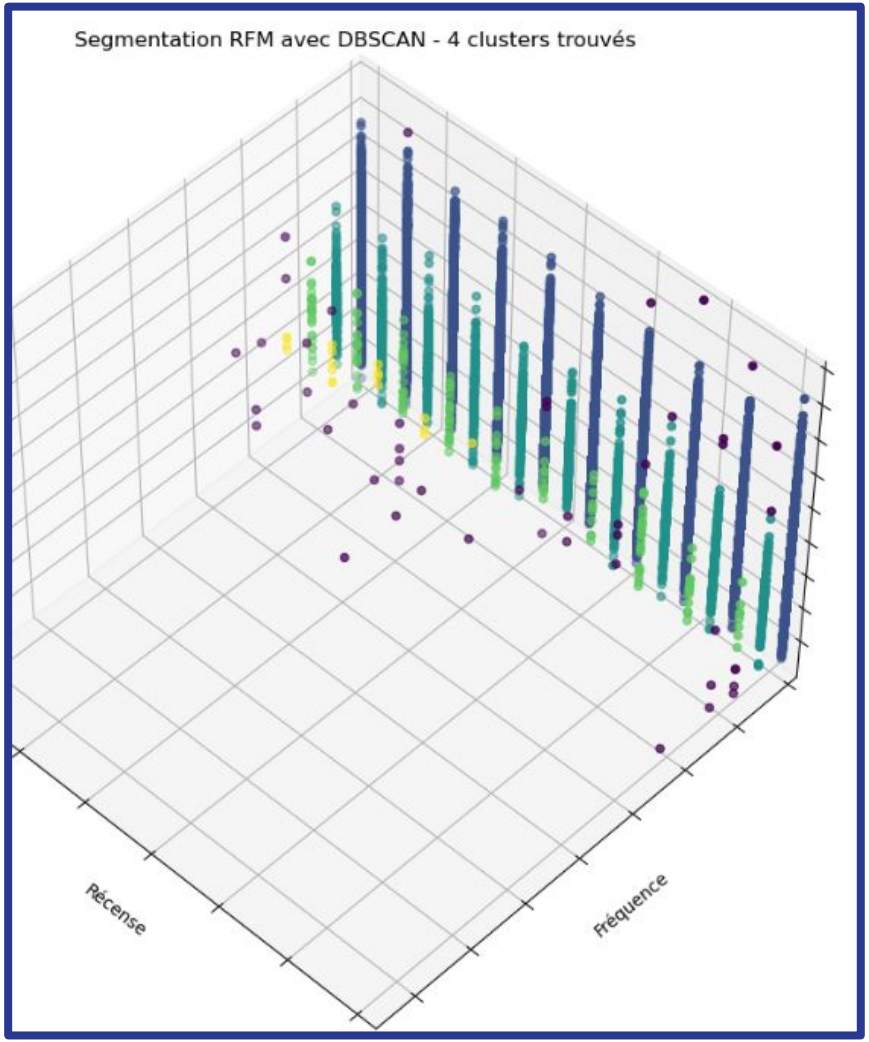


- **Exigence CAH: ressources mémoires élevées => on se limite à 40% des clients**

- Taille des clusters
  - Cluster 1: 6503 clients
  - Cluster 2: 1185 clients
  - Cluster 3: 8266 clients
  - Cluster 4: 8083 clients
  - Cluster 5: 6762 clients
  - Cluster 6: 7644 clients

DBSCAN avec RFM

- Pas besoin de déterminer au préalable nombre de clusters
- K=4 choisi
- Qualité clustering (homogénéité et séparation des clusters): coefficient de silhouette élevé
- **Inconvénient métier: 97% des clients dans le même cluster (1 commande)**

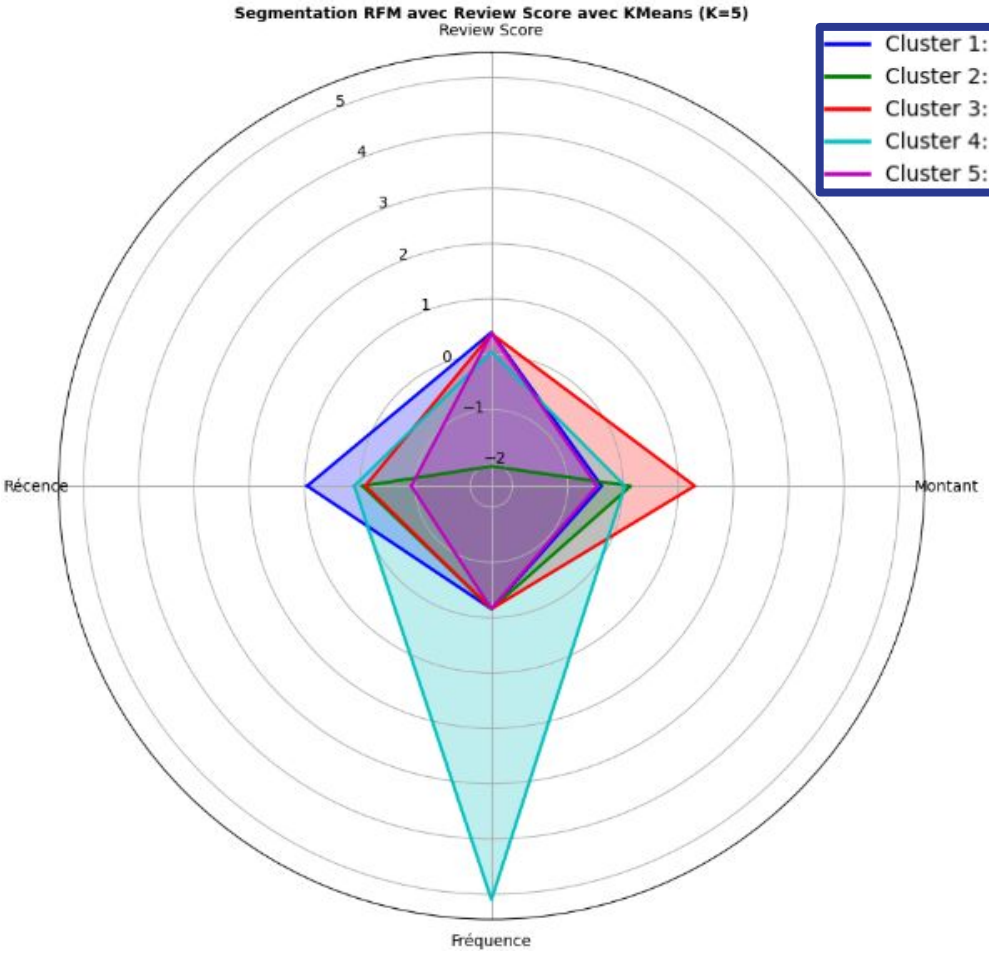


Comparaison des performances

Modèle de clustering	Recherche nb de clusters	Nb optimal de clusters	Coefficient de silhouette	Taille des partitions	Inconvénient majeur
K-Means	Silhouette + coude	5 clusters	0.37	ClusterMin=2903, ClusterMax=26787	AUCUN
Clustering hiérarchique	Silhouette + dendogramme	6 clusters	0.33	ClusterMin=1185, ClusterMax=8266	Ressources mémoires élevés (40% du dataset)
DBSCAN	Ne peut pas être prédéfini	4 clusters	0.65	ClusterMin=18, ClusterMax=92377	97% des clients dans le même cluster

- **K-means: meilleur modèle pour notre segmentation**
  - **Qualité**: le coefficient de silhouette atteint 0.37, ce qui est acceptable pour de la segmentation de clients
  - Métier: les 5 clusters trouvés par k-means sont **caractérisables**
  - Pas d'inconvénient majeur contrairement aux 2 autres modèles

K-means avec RFM + review_score



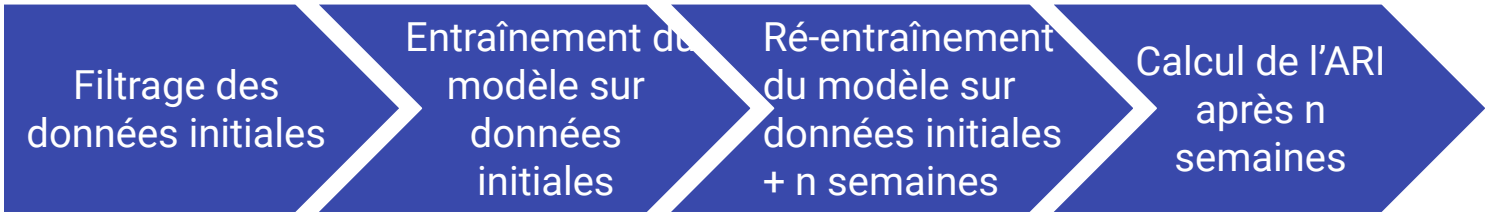
Cluster 1: 30562 clients	LOYAUX: Achat récent, montant faible, très satisfaits
Cluster 2: 14745 clients	MECONTENTS: achat assez récent, montant moyen, pas satisfaits
Cluster 3: 18030 clients	CHAMPIONS: Achat assez récent, montant élevé, très satisfait
Cluster 4: 2903 clients	FIDELES: Plusieurs commandes, dernier achat assez récent, montant moyen
Cluster 5: 29044 clients	A REACTIVER: Achat pas récent, montant faible, très satifaits

Partie 3: Simulation contrat de maintenance

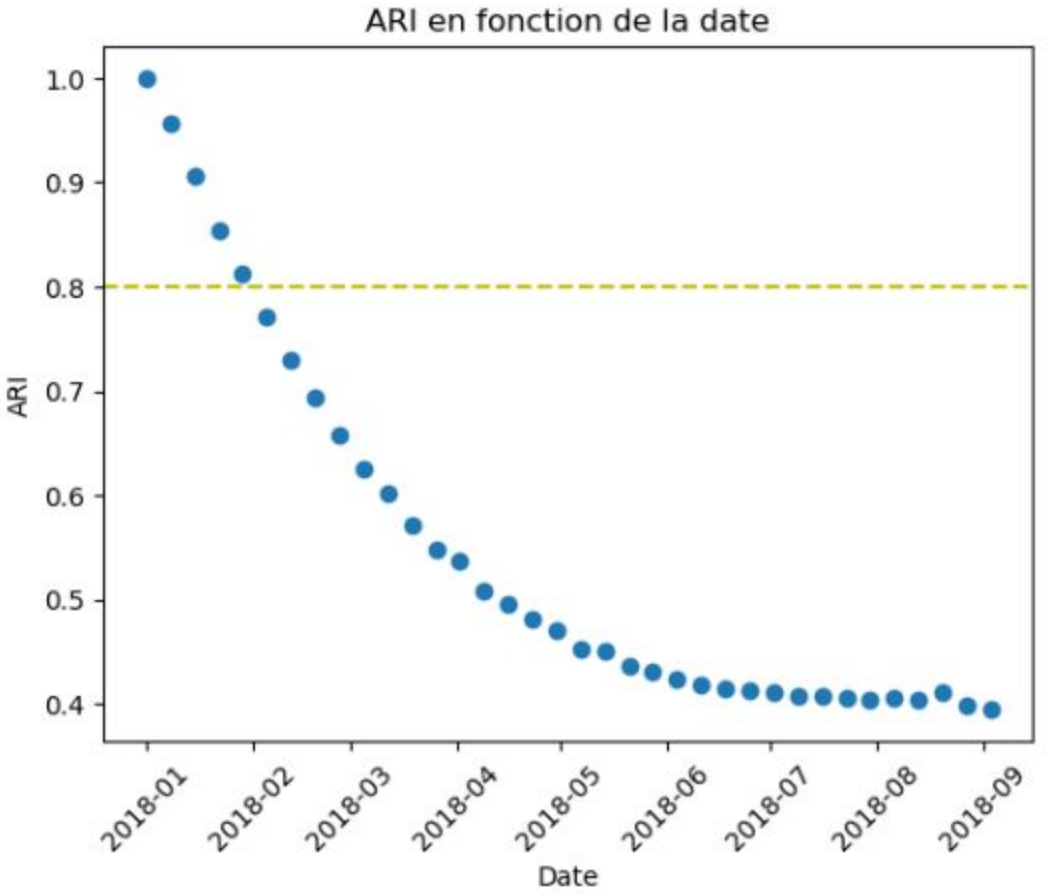
Principes et étapes de la simulation

- **Principes**
  - Evaluation de la stabilité du modèle dans le temps afin d'en assurer la maintenance
  - Cette simulation nous permettra de déterminer la fréquence nécessaire de mise à jour du modèle de segmentation

- **Etapes**



Stabilité du modèle dans le temps



- L'ARI décroît et passe sous la valeur de 0.8 dès la 5ème semaine
- Fréquence nécessaire de mise à jour du modèle de segmentation: toutes les 5 semaines

Conclusion

- Le K-means est le plus adapté à notre segmentation client => 5 segments

- Stabilité du modèle dans le temps: mise à jour toutes les 5 semaines

- Axes d'amélioration:
  - plus de données (plusieurs commandes par client)
  - intégrer plus de features (géographique, saison achat, …)