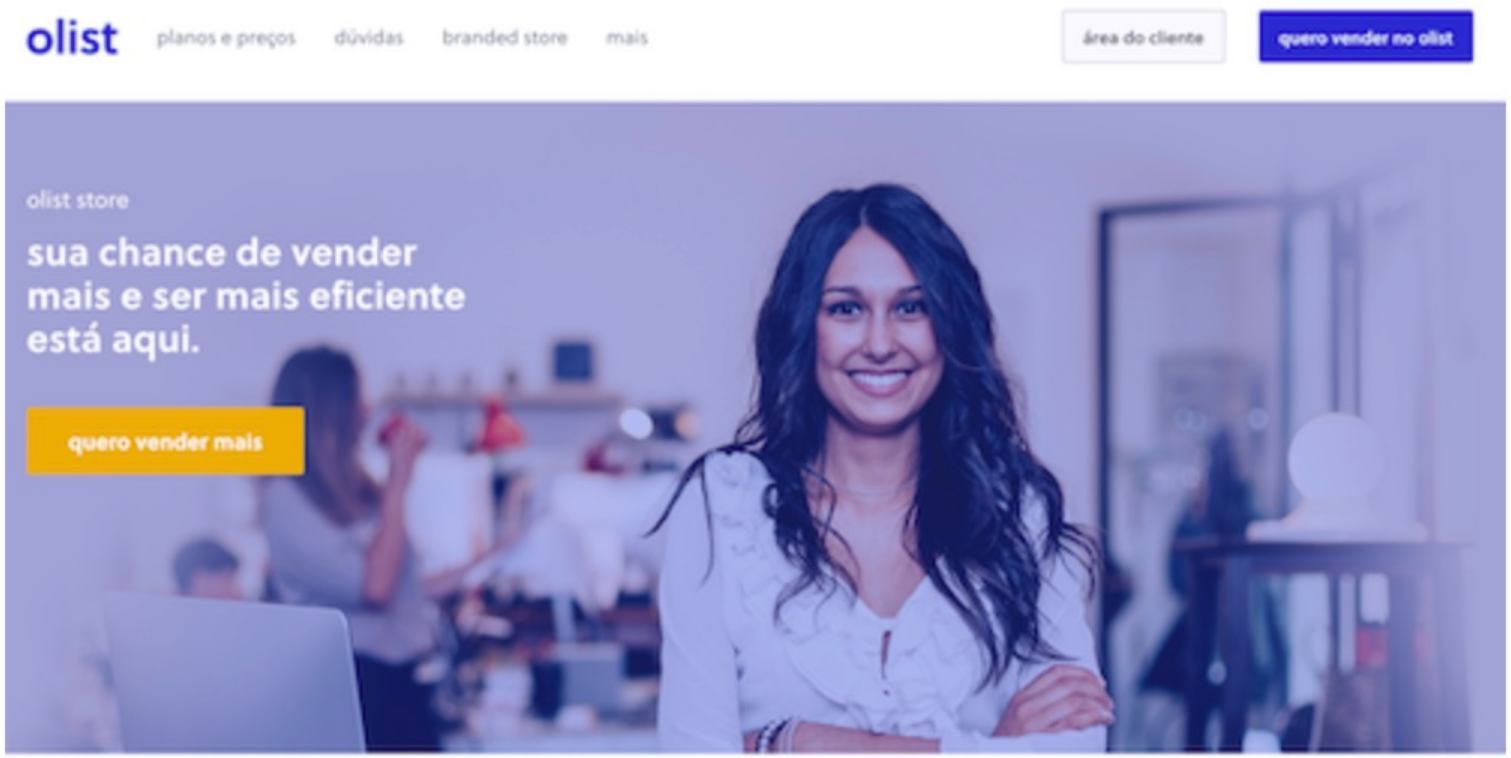
Segmentation des clients d'un site e-commerce



La page d'accueil du site Olist

Sommaire

Présentation de la problématique, de son interprétation et des pistes de recherche envisagées.

Présentation du cleaning effectué, du feature engineering et de l'exploration.

Présentation des différentes pistes de modélisation effectuées.

Présentation du modèle final sélectionné ainsi que des améliorations effectuées.

Demande client

La segmentation proposée doit être exploitable et facile d'utilisation pour l'équipe marketing.

Il faut évaluer de la fréquence à laquelle la segmentation doit être mise à jour. L'objectif est de pouvoir effectuer un devis de contrat de maintenance.

Le code fourni doit respecter la convention PEP8, pour être utilisable par Olist.

Interprétation

Choix des variables qui vont caractériser les groupes de clients

Fréquence de mise à jour du modèle.

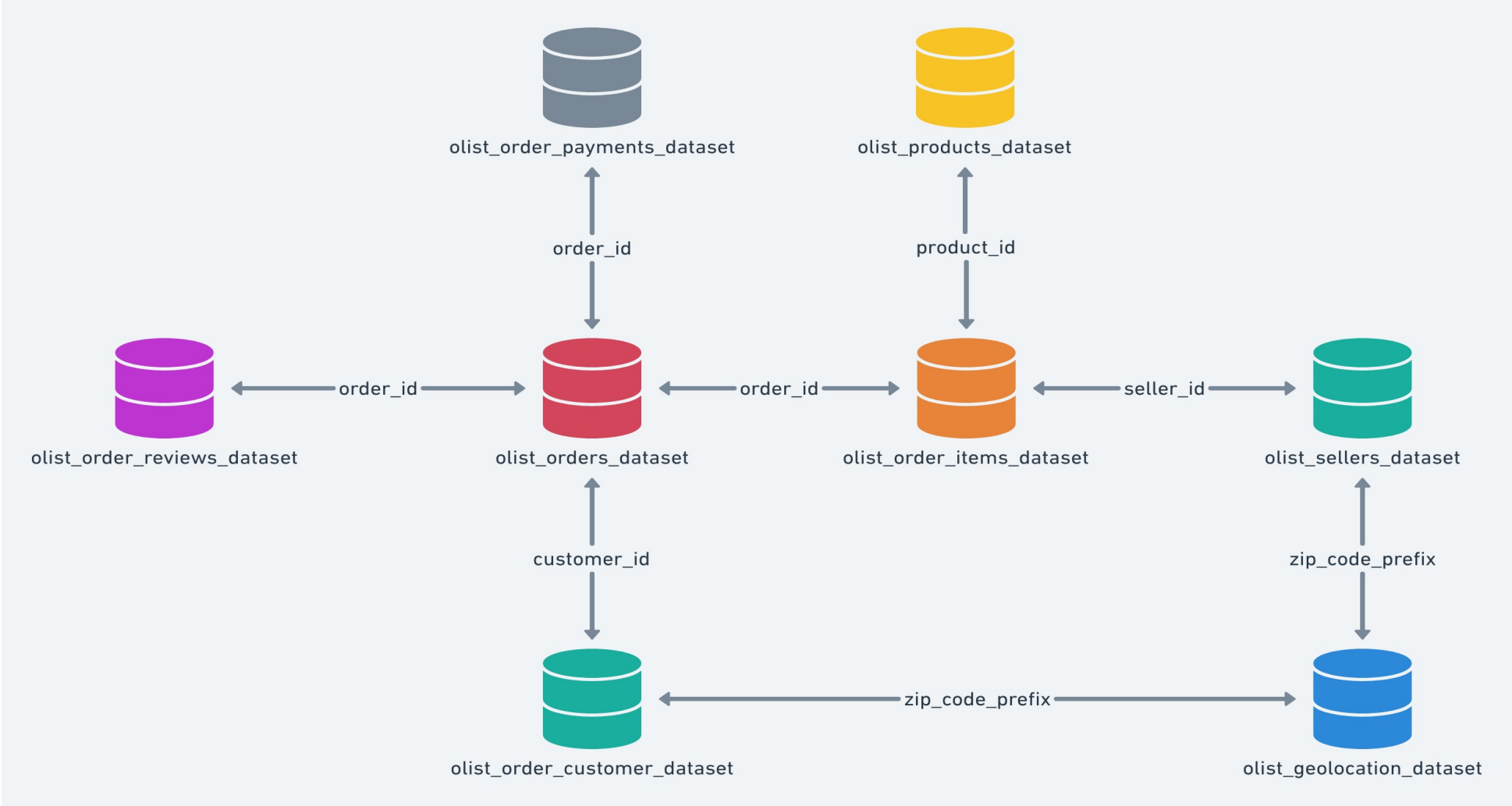
Pistes de recherche envisagées

- Nettoyage des données
- Analyse des variables
- Segmentation RFM
- Algorithme K-means
- Algorithme DBSCAN
- Algorithme CAH

Présentation des données

Période du 09/04/2016 au 17/10/2018.

96 096 clients référencés
99 441 commandes



Cleaning effectué

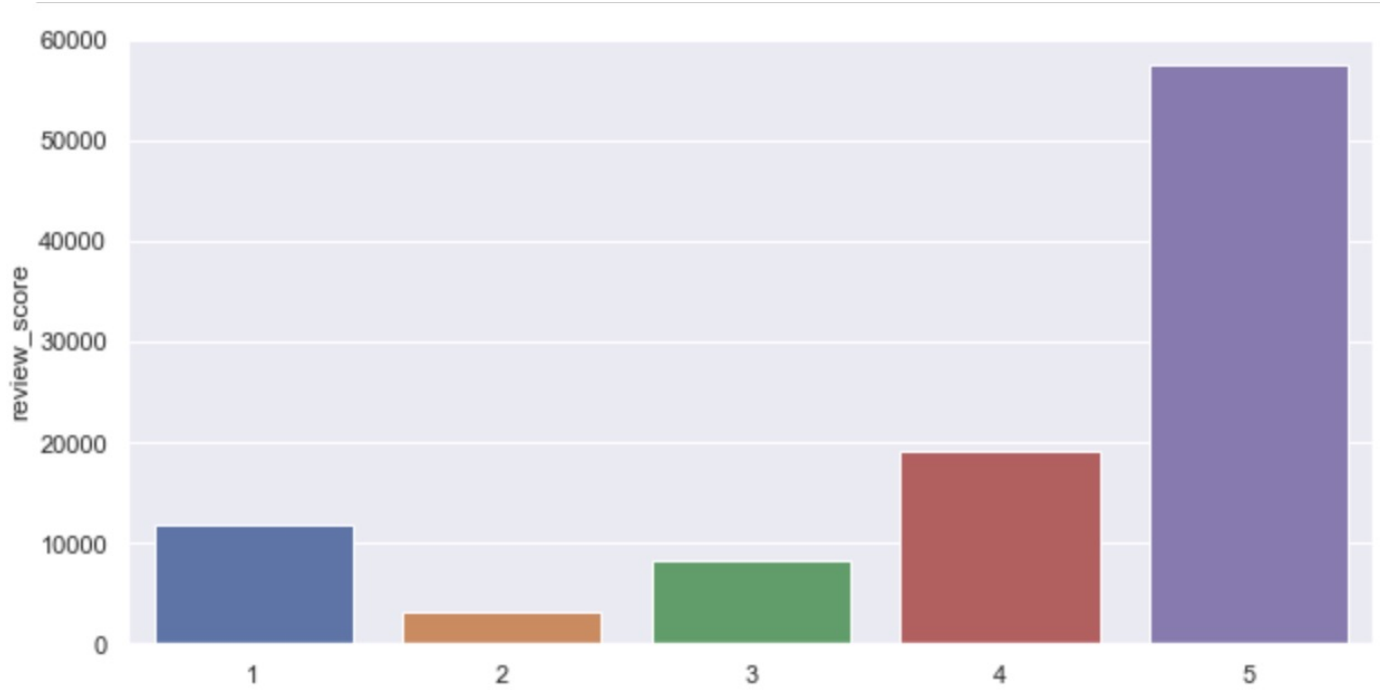
- Suppression des valeurs manquantes
- Conservation des seules commandes livrées
- Suppression des variables de géolocalisation
- Suppression des variables relatives aux vendeurs

Feature engineering

- Création des variables RFM
- Passage au log de la variable Montant
- Création d’une variable de délai de livraison d’une commande (en jours)
- Création d’une variable sur le volume d’un produit
- Création de deux variables pour déterminer si le client a laissé un commentaire avec titre et/ou texte

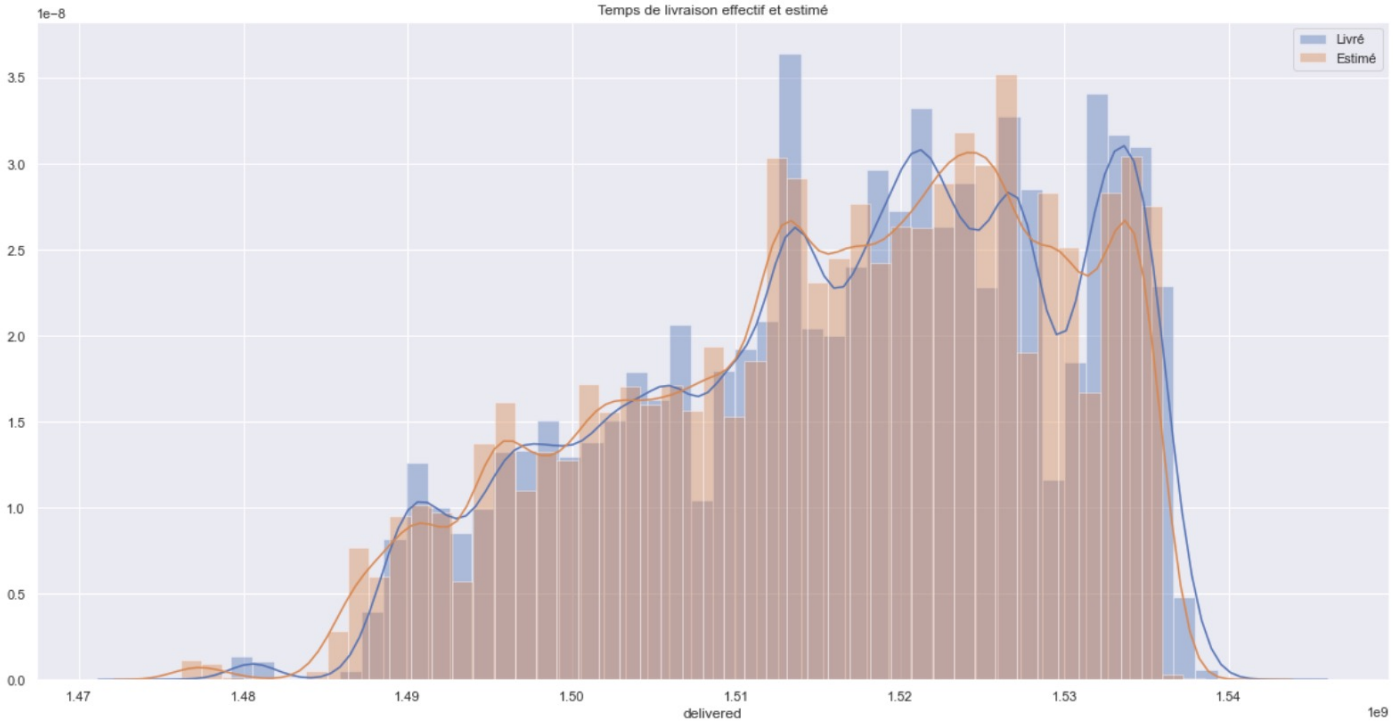
Exploration

Questionnaire de satisfaction



Exploration

Comparaison entre le temps de livraison effectif et le temps de livraison estimé



Estimation des livraisons de commandes :

- 87005 estimations sont au dessus de la livraison réelle
- 1459 estimations correspondent exactement avec la livraison réelle
- 7824 estimations sont en dessous de la livraison réelle

Modélisation

Modélisation sur 3 mois
du 29/08/2017 au 29/11/2017

15 572 clients

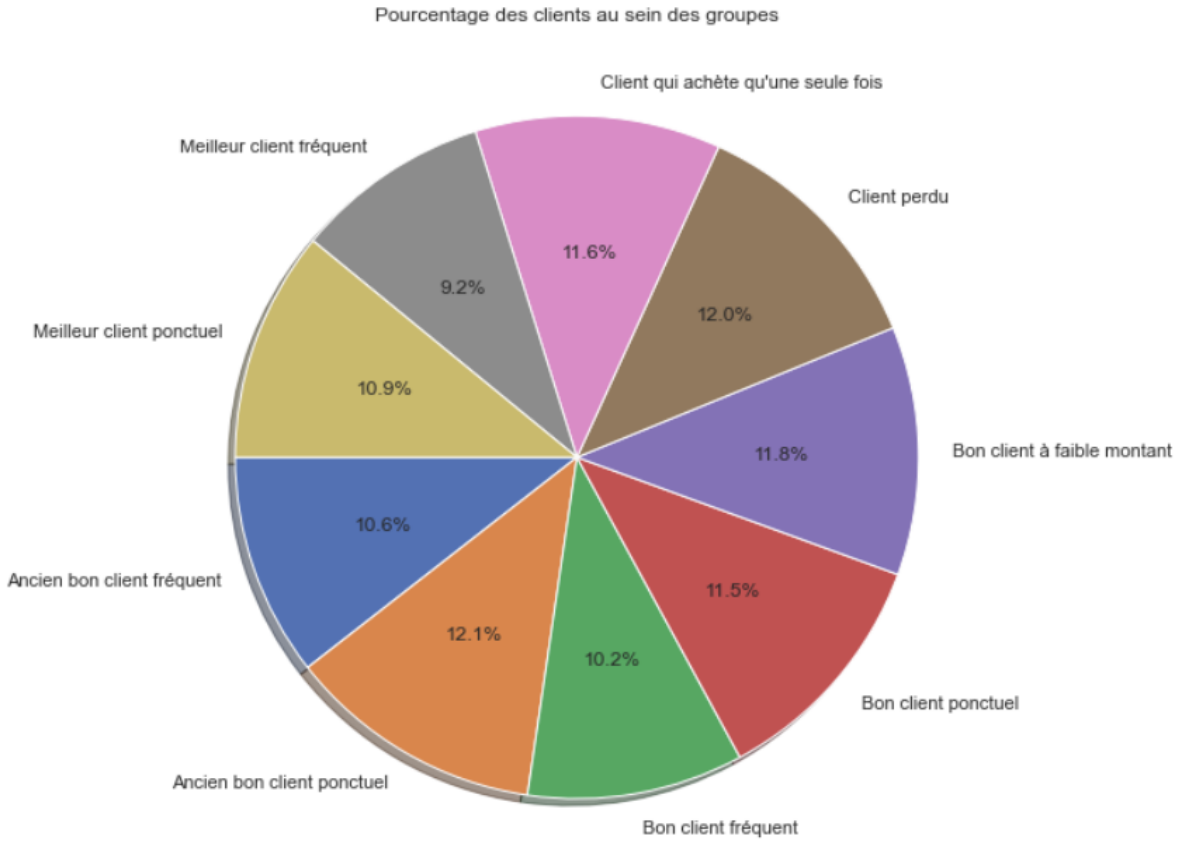
15 898 commandes.

Modélisation RFM

Évaluation selon 3 aspects:

- la récence du dernier achat (R)
- la fréquence des achats (F)
- le montant dépensé par transaction (M).

Segment	Catégorie	clientèle
1	Meilleur	Fréquent
2	Meilleur	Ponctuel
3	Bon	Fréquent
4	Bon	Ponctuel
5	Ancien bon	Fréquent
6	Ancien bon	Ponctuel
7	Bon à faible montant	Fréquent
8	Qui n'achète qu'une seule fois	Ponctuel
9	Perdu	-



Modélisation choix des variables

- Délais de livraison en jours
- Note d’appréciation
- Fréquence
- Récence
- Montant total (passage au log)

Pistes de recherche

- Algorithme K-means
- Algorithme DBSCAN
- Algorithme CAH

Méthode de sélection des clusters

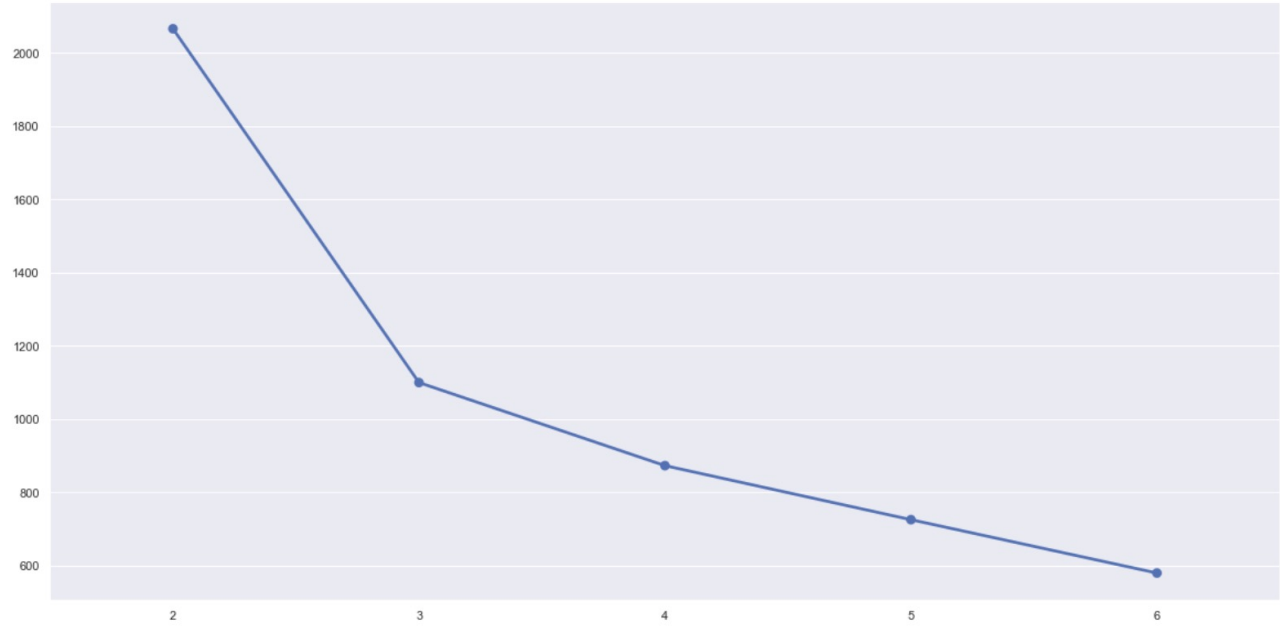
- Coefficient de silhouette
- Méthode « Elbow »

Coefficient de silhouette

Coefficient de silhouette mesure de la qualité d'une partition.

- la différence entre la distance moyenne avec les points du même groupe
- la distance moyenne avec les points des autres groupes voisins.

Méthode « Elbow »



Sélection du point d’inflexion

Modélisation K-means

Regroupe en clusters les observations similaires (distance de séparation).

- Silhouette Score(n=2): 0.450
- Silhouette Score(n=3): 0.454
- Silhouette Score(n=4): 0.456
- Silhouette Score(n=5): 0.419
- Silhouette Score(n=6): 0.392

Modélisation K-means

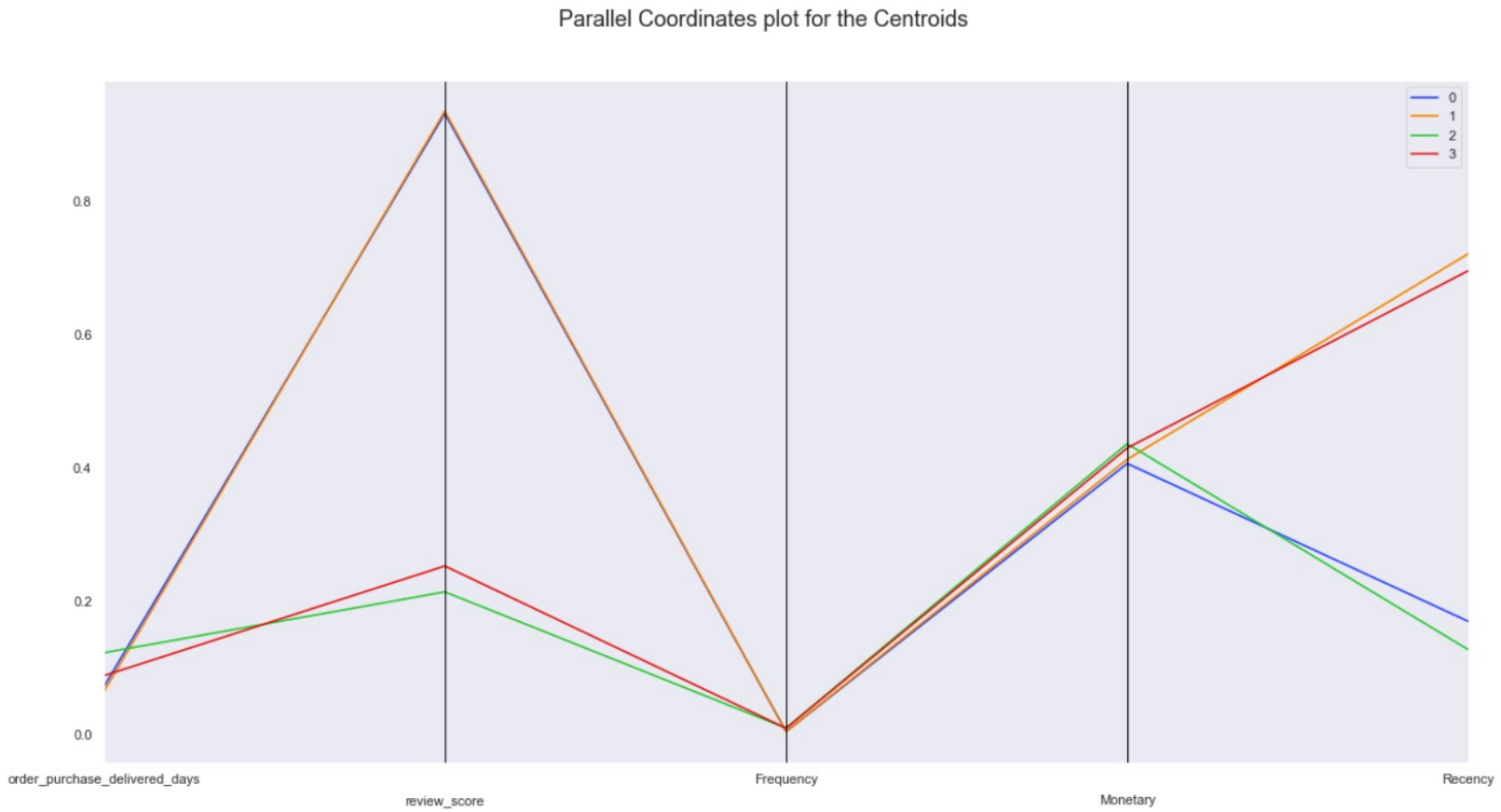
Segmentation en 4 groupes

Groupe 0 :
- Livraison rapide
- Bonne note
- Ne commande pas souvent
- Faible montant
- N'a pas commandé récemment

Groupe 1 :
- Livraison standard
- Mauvaise note
- Ne commande pas souvent
- Montant moyen
- N'a pas commandé récemment

Groupe 2 :
- Livraison rapide
- Bonne note
- Commande souvent
- Montant importants
- N'a pas commandé récemment

Groupe 3 :
- Livraison très lente
- Mauvaise note
- Ne commande pas souvent
- Montant moyen
- A commandé récemment



Modélisation DBSCAN

L'algorithme DBSCAN utilise 2 paramètres : la distance ε et le nombre minimum de points « MinPts » devant se trouver dans un rayon ε pour que ces points soient considérés comme un cluster.

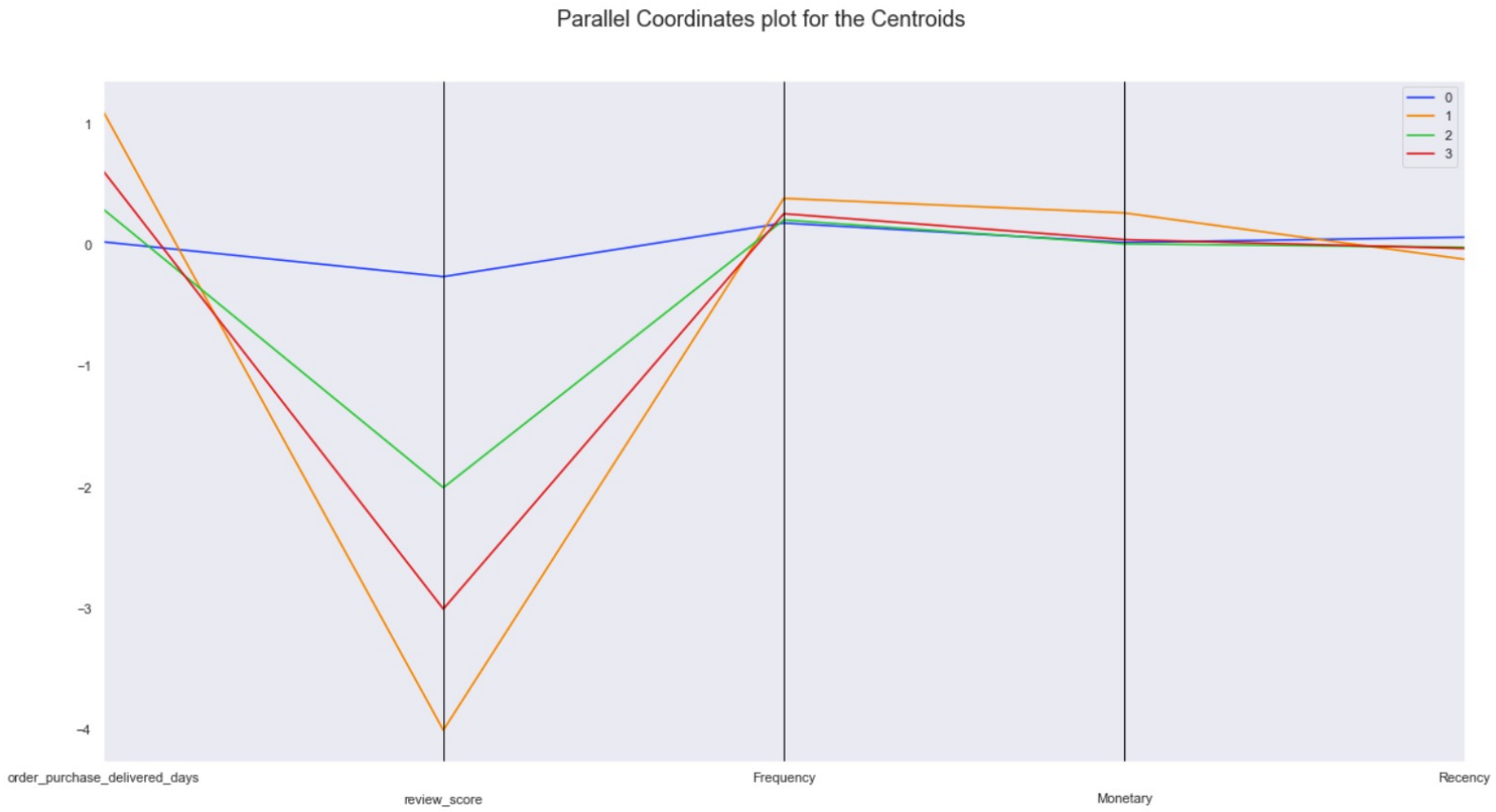
Groupe 0 :
- Livraison rapide
- Bonne note
- Ne commande pas souvent
- Faible montant
- N'a pas commandé récemment

Groupe 1 :
- Livraison lente
- Mauvaise note
- Commande peu souvent
- Montant élevé
- A commandé récemment

Groupe 2 :
- Livraison moyenne
- Bonne note
- Commande peu souvent
- Montant importants
- N'a pas commandé récemment

Groupe 3 :
- Livraison moyenne
- Note moyenne
- Commande souvent
- Montant importants
- N'a pas commandé récemment

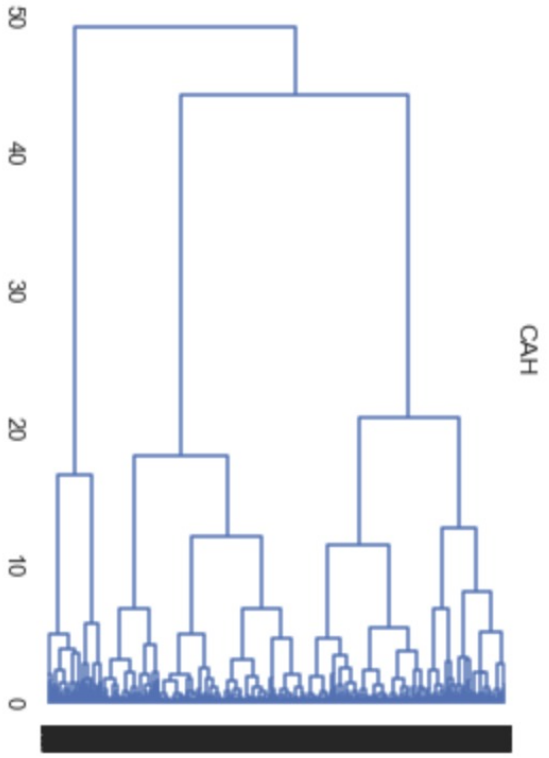
Estimation de 4 clusters
191 point considérés comme du bruit
Silhouette Score: 0.25



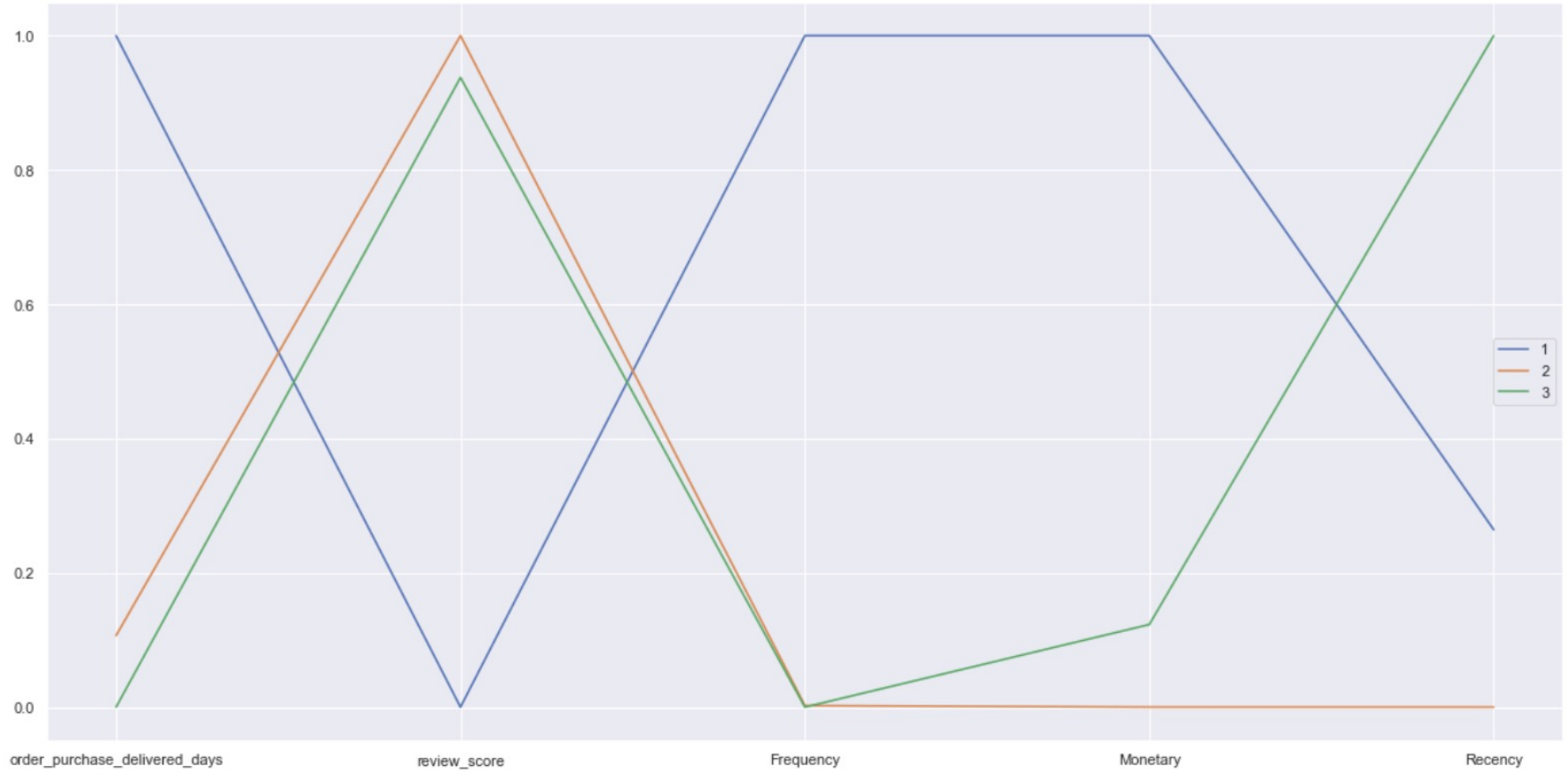
Modélisation CAH
(Classification Ascendante Hiérarchique)

Cette méthode démarre avec un cluster par individu puis se regroupe à chaque étape selon un critère jusqu'à l'obtention d'un seul cluster contenant l'ensemble des individus.

Critère utilisé : distance



Silhouette Score : 0.15



Sélection du modèle

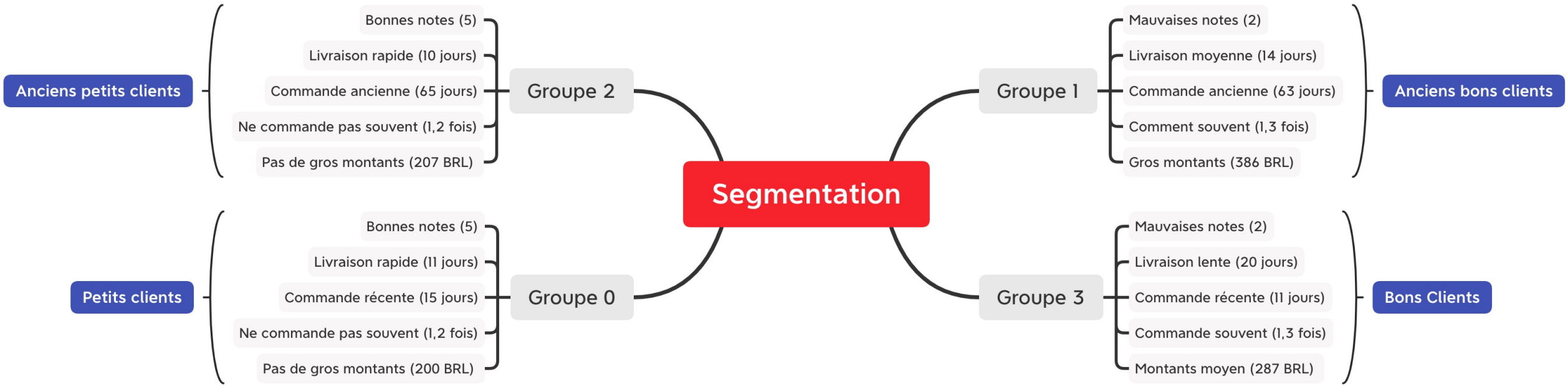
Modèle KMEANS avec 5 variables, une normalisation de type "MinMax" et une segmentation en 4 groupes

Modèle	nb groupes	Silhouette	Elbow	Commentaire
RFM				
KMEANS	2	0,58	3	Variables RFM uniquement
KMEANS	**2**	**0,48**	**4**	**5 variables**
KMEANS	2	0,48	3	normalisation robust
KMEANS	2	0,38	4	normalisation standard
KMEANS	2	0,31	4	variables numériques
CAH	3	0,28		
KMEANS	3	0,26	3	PCA
DBSCAN	5	0,24		
KMEANS	2	0,07	4	Toutes les variables

- Délai moyen entre la commande et la livraison (purchase_delivered)
- Moyenne de la récence (Recency)
- Moyenne des notes d'appréciation (review_score)
- Moyenne de la fréquence (Frequency)
- Montant total moyen (Monetary)

Modèle

Modèle KMEANS avec 5 variables, une normalisation de type "MinMax" et une segmentation en 4 groupes



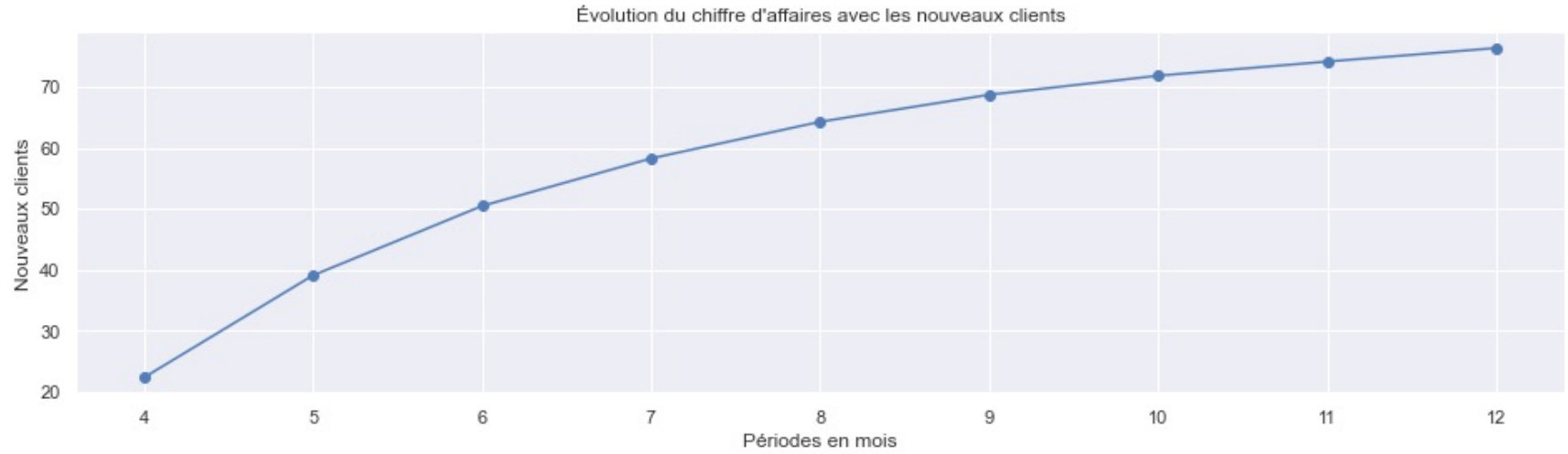
Stabilité

Période de référence sur 3 mois
du 29/08/2017 au 29/11/2017

9 périodes de 1 mois

Stabilité

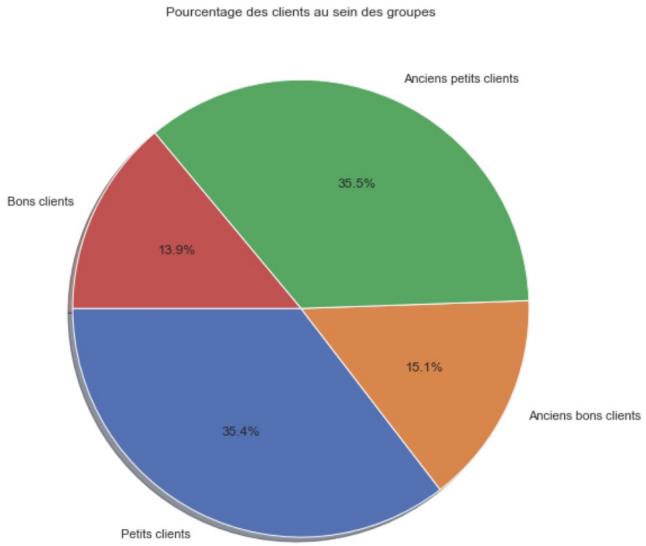
Évolution des nouveaux clients



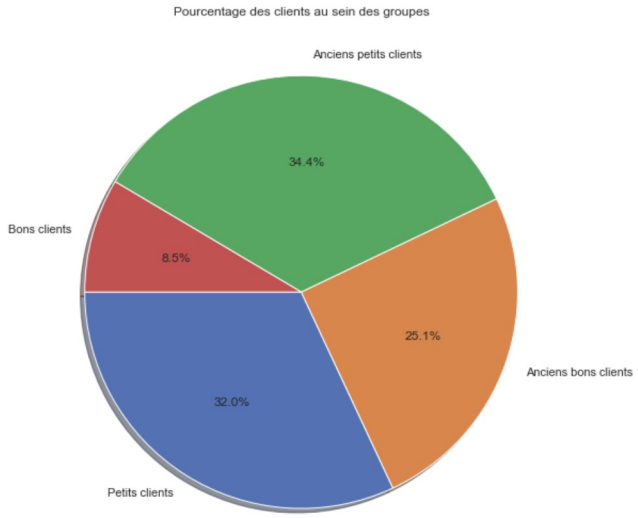
Stabilité

Analyse de la première période

Période de référence



Période de 4 mois

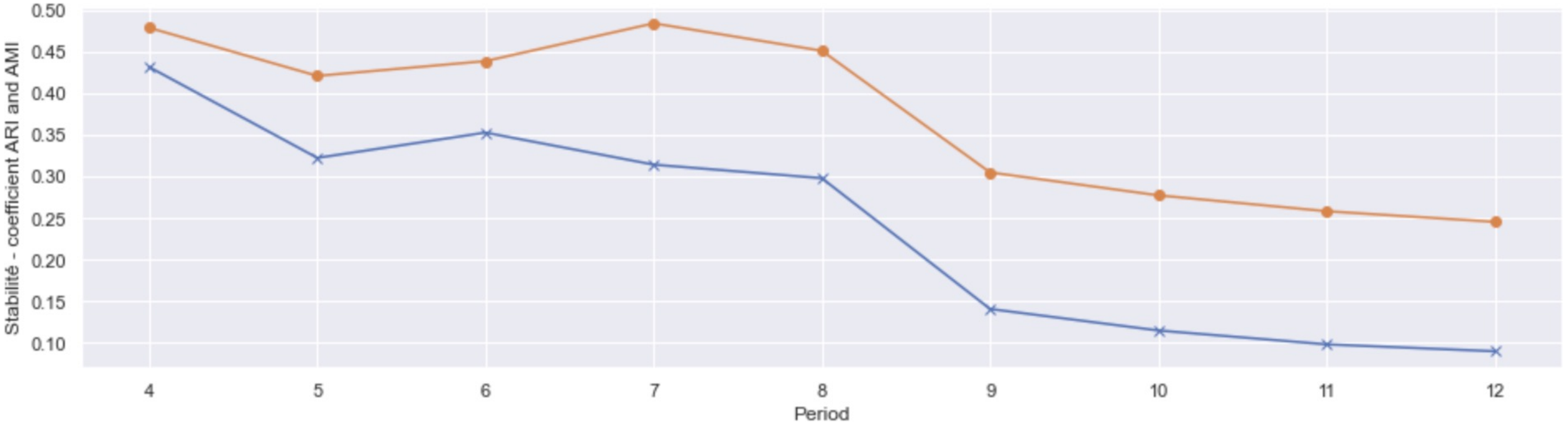


Migration des clients du groupe « petits clients » vers le groupe « anciens petits clients »

Migration des clients du groupe « bons clients » vers le groupe « anciens bons clients »

Analyse de la stabilité

Chute du modèle après 8 mois.



L'**indice de Rand** (ARI) mesure la similarité entre les groupes des deux périodes.

L'**information mutuelle** (AMI) est une quantité mesurant la dépendance statistique des groupes des deux périodes.

Récapitulatif

Variables

- Délai moyen entre la commande et la livraison (purchase_delivered)
- Moyenne de la récence (Recency)
- Moyenne des notes d'appréciation (review_score)
- Moyenne de la fréquence (Frequency)
- Montant total moyen (Monetary)

Chute du modèle après 8 mois