Projet 5 – Parcours Data Scientist

Segmentez des clients d'un site e-commerce

OPENCLASSROOMS

Robin Perbet – juillet 2022

Sommaire

1. Introduction

Introduction

Problématique du projet

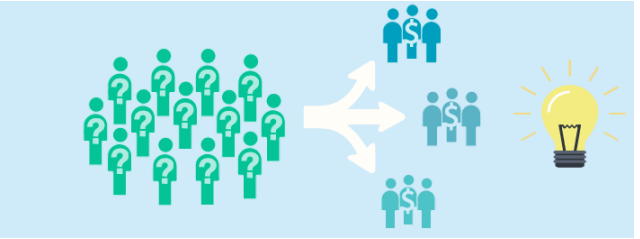
Présentation de la problématique

Olist : une entreprise brésilienne de e-commerce

La société Olist est une entreprise brésilienne qui propose une solution de vente sur les marketplaces en ligne

Une volonté de réaliser un clustering

Afin de mieux comprendre sa clientèle et d'effectuer des campagnes marketing ciblées, la société souhaite réaliser une segmentation des différents types d'utilisateurs grâce à leur comportement et à leurs données personnelles



Avec la possibilité d'effectuer des maintenances régulières

Afin que cette segmentation demeure pertinente, la société souhaite également procéder à des maintenances

→ L'objectif sera à partir de réaliser un algorithme de machine learning permettant de segmenter les clients et de proposer un délai de maintenance pour maintenir cette segmentation pertinente

Introduction

Présentation du jeu de données

Description des fichiers à disposition

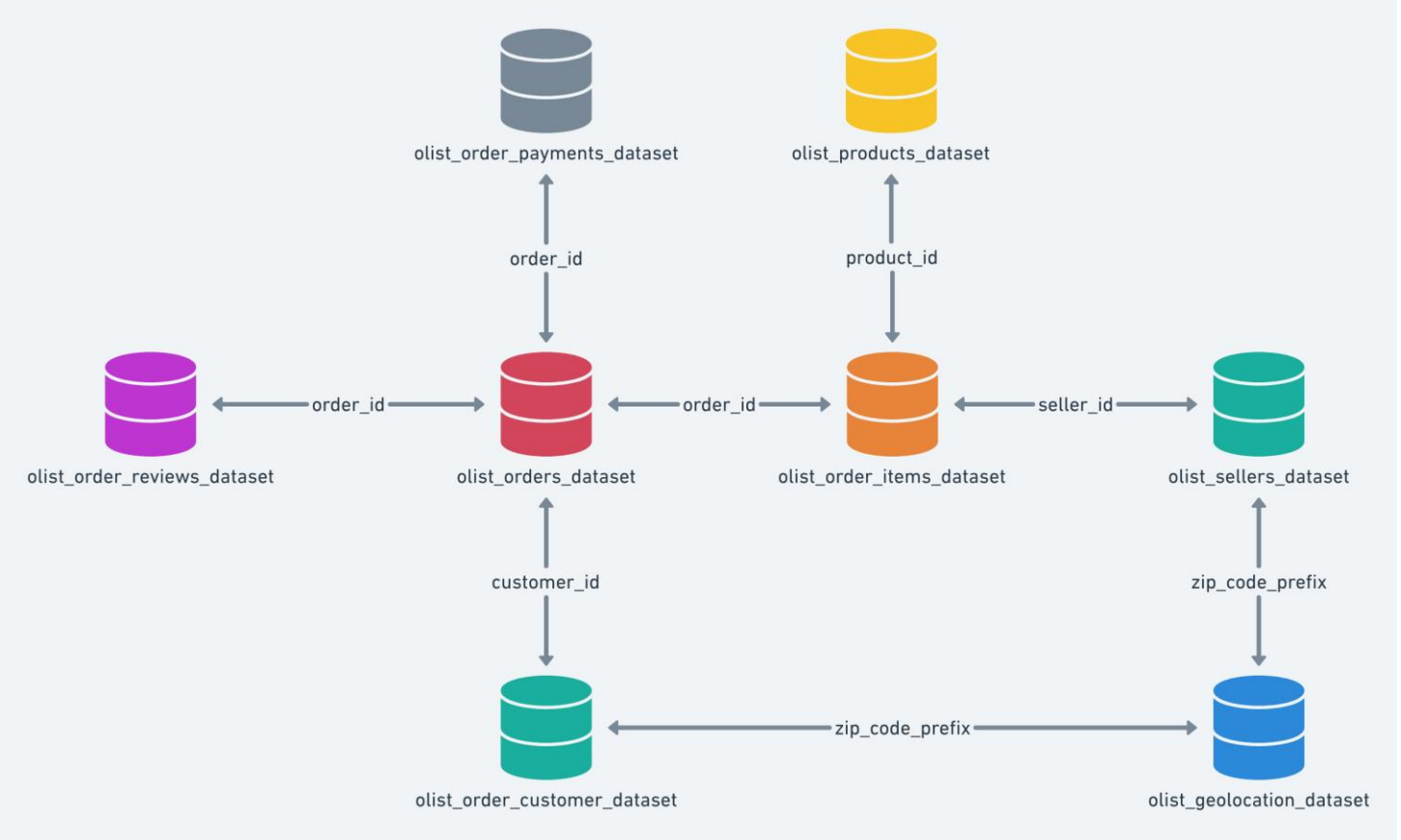
Fichier	Taille	Description
customers	99 441 x 5	Informations sur les clients (localisation et ID client)
geolocation	1 000 163 x 5	Informations détaillées de localisation en fonction du code postal (latitude, longitude, ville, état)
orders	99 441 x 8	Informations sur les commandes (ID client, ID commande, statut, chronologie des étapes)
product_category_name_translation	71 x 2	Traduction des catégories de produits du portugais à l'anglais
order_items	98 666 x 7	Table permettant d'associer ID commande, ID vendeur et ID produits, ainsi que des informations sur la commande (prix et date)
products	32 951 x 9	Informations sur les produits (type, description, taille, etc.)
order_reviews	99 224 x 7	Informations sur les évaluations des commandes (note, commentaires, date)
order_payments	103 886 x 5	Informations sur le paiement des commandes (nombre de paiements, moyen utilisé, montant)
sellers	3 095 x 4	Informations sur les vendeurs (localisation et ID vendeur)

2. Nettoyage des données

Nettoyage des données

Fusion des tableaux dans un jeu de données unique

Présentation du schéma du jeu de données



Objectifs

1. Conserver une ligne par commande avec toutes les informations associées
2. Repérer les éventuelles incohérences afin que la fusion soit effectuée proprement
3. Synthétiser certaines variables afin d'avoir une seule ligne par commande
4. Le tableau vendeurs n'a pas été fusionné car cela ne semblait pas pertinent pour une segmentation clients
5. Le tableau permettra ensuite de faire un groupby afin d'obtenir les informations pertinentes sélectionnées par client unique

Nettoyage des données

Harmonisation des formats et traitement des outliers

Actions réalisées pour le nettoyage des données

1 Vérification du type des variables

- Les variables concernant des dates étaient sous un format objet, elles ont été converties en format temps
- Le type des autres variables était adapté

2 Vérification de la cohérence des variables numériques

- Vérification que les valeurs pour chaque variables soient contenues dans un intervalle cohérent
- Il n'existait globalement pas de valeurs aberrantes, mais quelques valeurs atypiques

3 Vérification de la cohérence des variables catégorielles

- Vérification de l'absence de doublons sur les catégories (catégories désignant la même chose mais avec des différences d'orthographe)
- Pas de valeurs aberrantes décelées

4 Analyse et traitement des valeurs manquantes

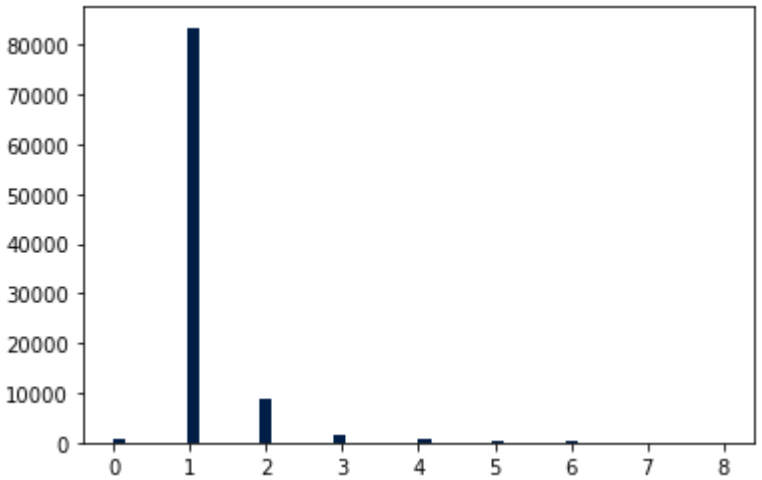
- Observation des valeurs manquantes et de l'éventuelle possibilité de manuellement inférer ou de remplacer ces valeurs
- Quelques valeurs ont pu être remplacées

3. Exploration des données

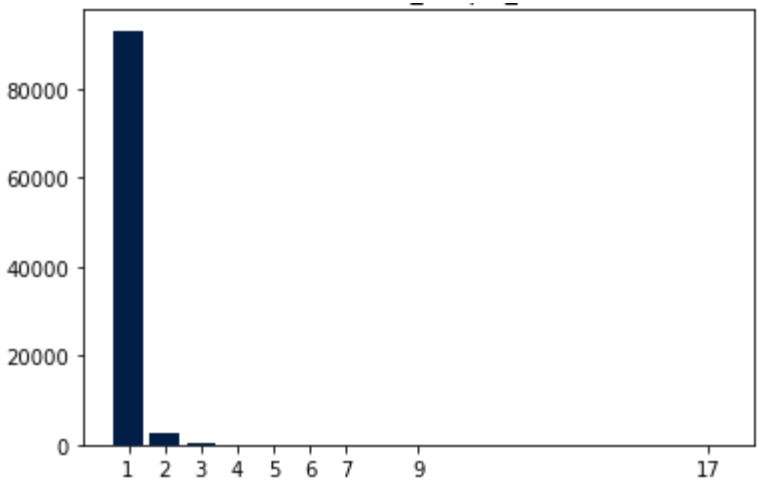
Exploration des données

Analyse de la distribution de quelques variables

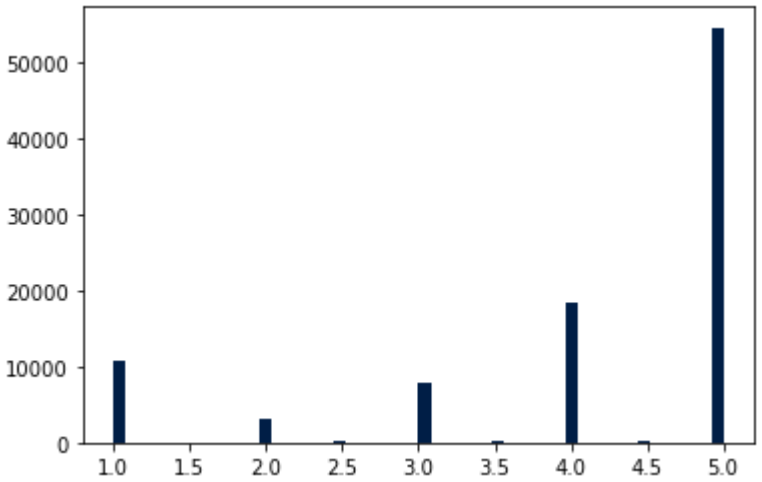
Nombre de produits achetés par client unique



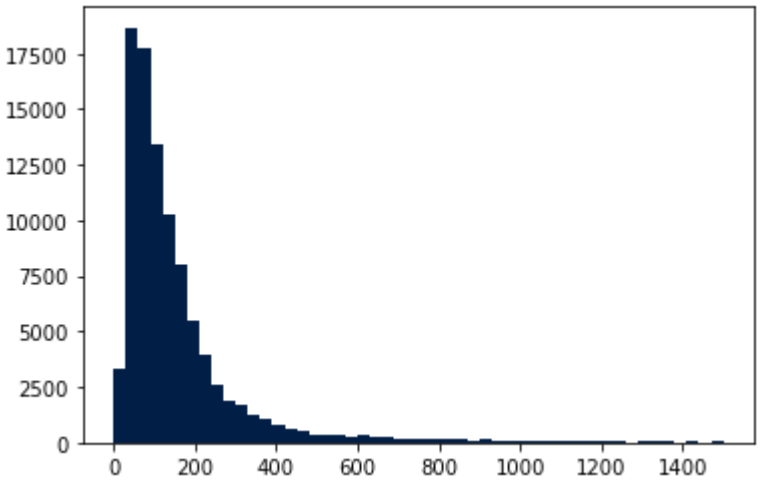
Nombre de commandes réalisées par client unique



Note de satisfaction moyenne par client unique



Montant total payé par client unique



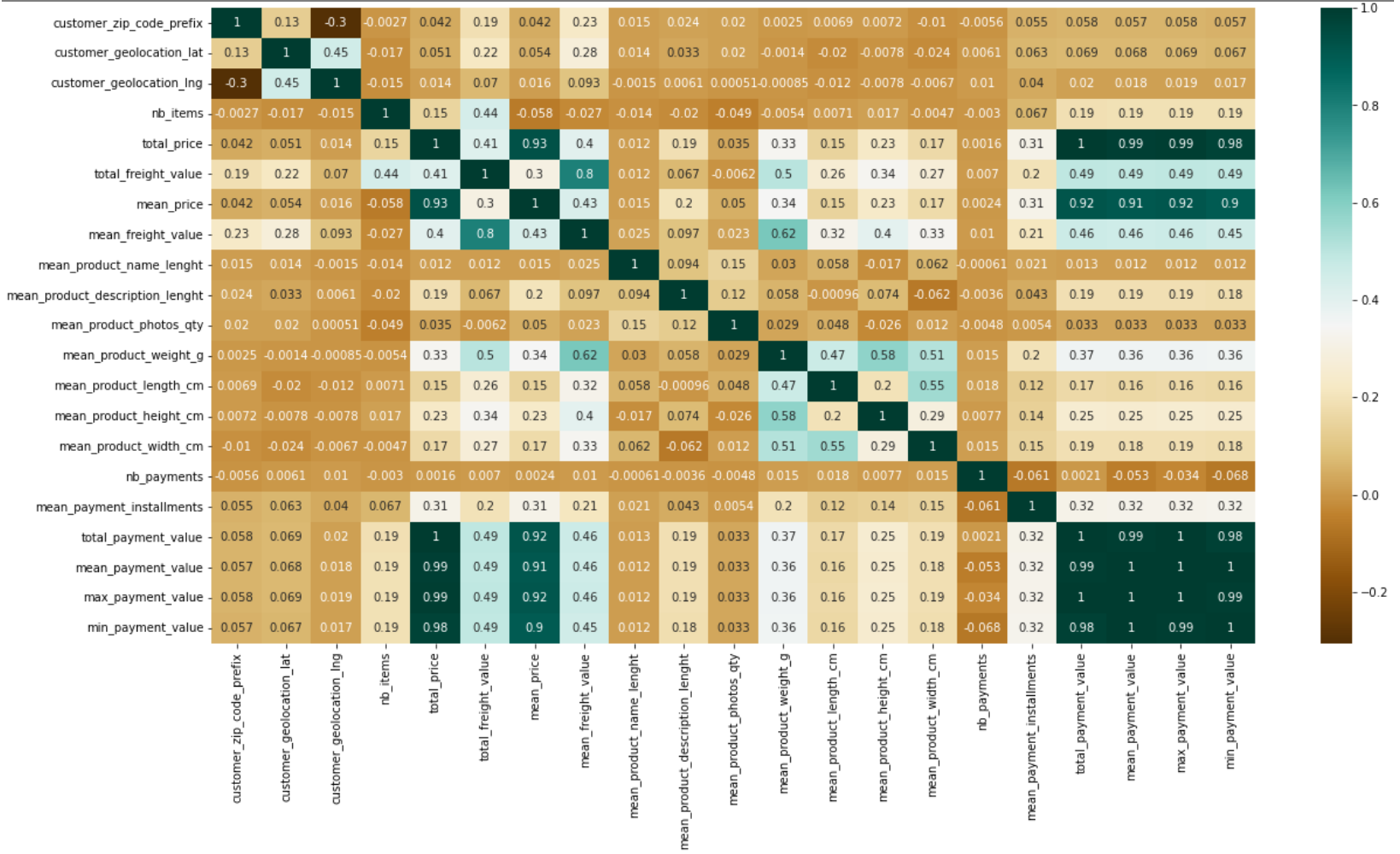
Exploration des données

Matrice des corrélations

Commentaires

- Les variables très fortement corrélées (> 0,8) correspondent à des indicateurs d'une même variable initiale (moyenne, max, min)
- Autrement, nous observons peu de corrélation entre les variables à l'exception :
  - Des variables concernant les dimensions des produits entres elles (poids, taille, etc.)
  - Des frais de port avec le nombre d'objets, le prix et les dimensions

Matrice des corrélations des variables numériques



4. Pre-processing et feature engineering

Pre-processing et feature engineering

Méthodologie

Présentation des actions réalisées

Objectif

- Créer des variables synthétiques par client unique afin de pouvoir effectuer une classification métier pertinente

Variables retenues

- Date de la dernière commande (Récence)
- Nombre de commandes (Fréquence)
- Montant total dépensé sur la plateforme (Montant)
- Note moyenne donnée par commande (Satisfaction)
- Etat de résidence du client (Localisation)

Variables non retenues

- Date de la première commande
- Temps moyen entre les commandes
- Nombre de produits total et nombre moyen par commande
- Prix moyen et frais de ports moyen par commande
- Catégorie de produits la plus achetée
- % d'évaluations laissées
- Temps entre la réception du produit et l'évaluation
- Mode de paiement le plus utilisé

- Preprocessing :
  - Variables numériques : imputation par la moyenne des valeurs manquantes puis normalisation par un standard scaler
  - Variables catégorielles : imputation par le mode puis encodage ordinal

5. KMeans

KMeans

Méthodologie

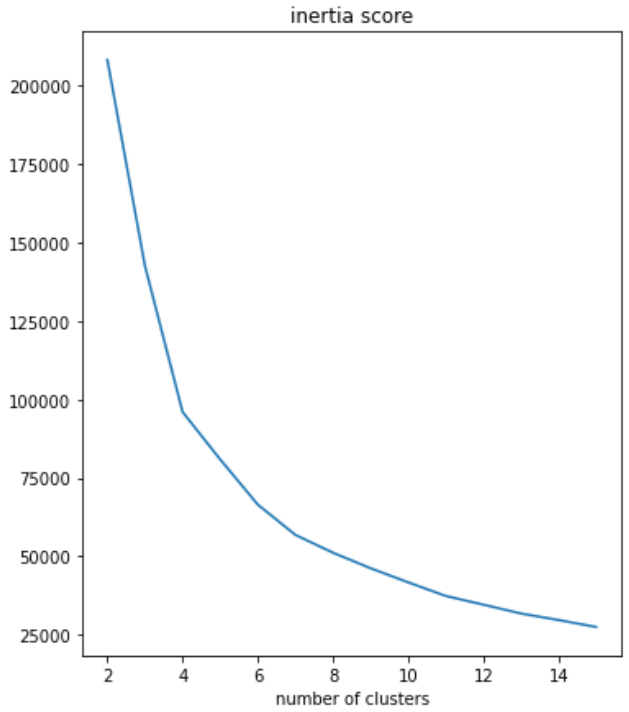
1. Utiliser les variables les plus pertinentes afin de décrire le comportement des utilisateurs
2. Optimiser le paramètre k de l'algorithme afin d'obtenir une inertie faible avec un silhouette score élevé
3. Interpréter les résultats
4. Ajouter de nouvelles variables afin de voir si cela permet d'obtenir une segmentation qui reste pertinente d'un point de vue métier
5. Itération 1 : variables nombre de commandes, date de la dernière commande et montant total dépensé sur la plateforme (RFM)
6. Itération 2 : ajout de la note moyenne par commande
7. Itération 3 : ajout de la localisation du consommateur

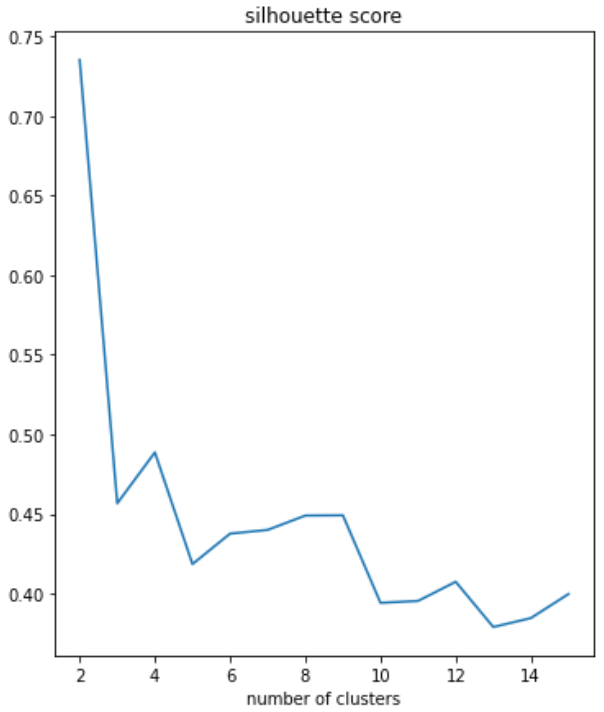
5.1. Première itération

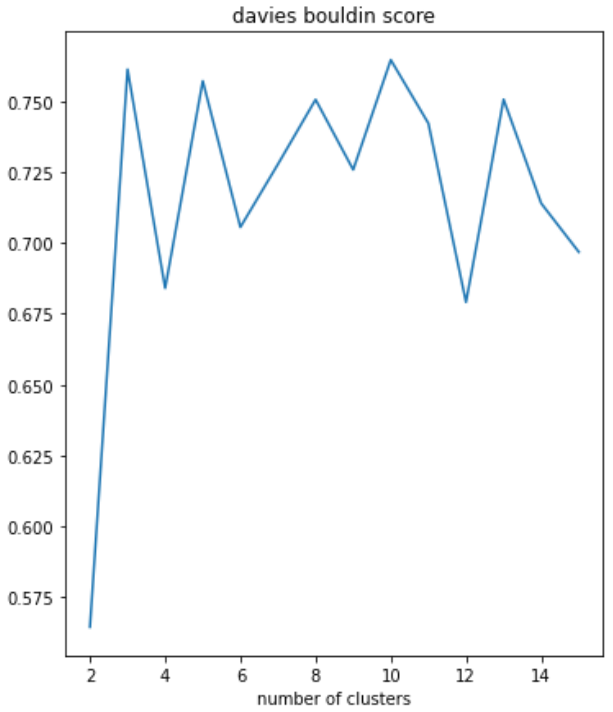
KMeans

Première itération – choix du nombre de clusters

Présentation des scores en fonction du nombre de clusters







Choix du nombre de clusters

- Nous observons un léger « coude » à 4 clusters pour le score d'inertie : au-delà de 4 clusters la minimisation de la variance intra-cluster est moins rapide
- Pour la séparation des clusters, le silhouette score et le score davies bouldin mettent en avant 2 clusters. Toutefois l'inertie est élevée avec 2 clusters, de plus cela n'offrirait pas une distinction très intéressante d'un point de vue métier
- 4 clusters permet à la fois de minimiser la variance intra-cluster et d'obtenir des clusters relativement éloignés les uns des autres, nous retenons donc cette valeur pour k

KMeans

Première itération – interprétation des résultats

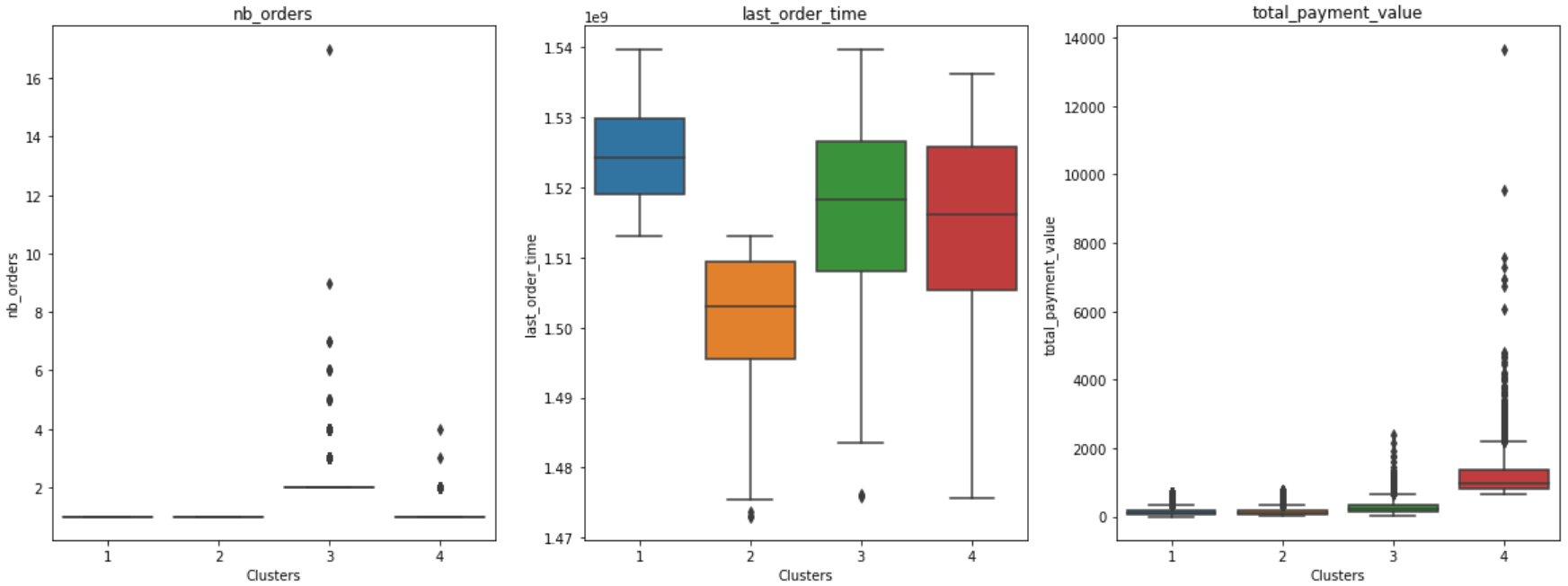
Répartition des clients par cluster

- Cluster 1 : 52 154 clients (54,3%)
- Cluster 2 : 38 549 clients (40,1%)
- Cluster 3 : 2 962 clients (3,1%)
- Cluster 4 : 2 421 clients (2,5%)

Interprétation

	Récence	Fréquence	Montant
Cluster 1	+++	+	+
Cluster 2	+	+	+
Cluster 3	++	+++	++
Cluster 4	++	++	+++

Distribution des variables par cluster

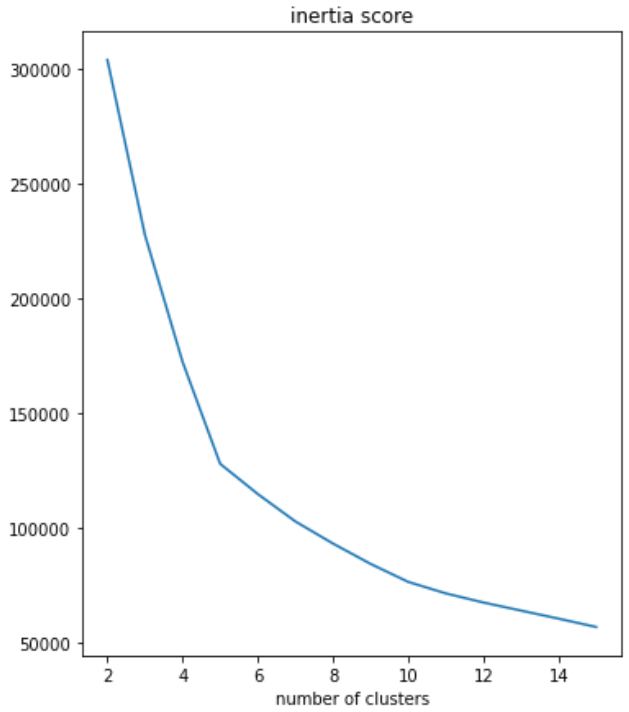


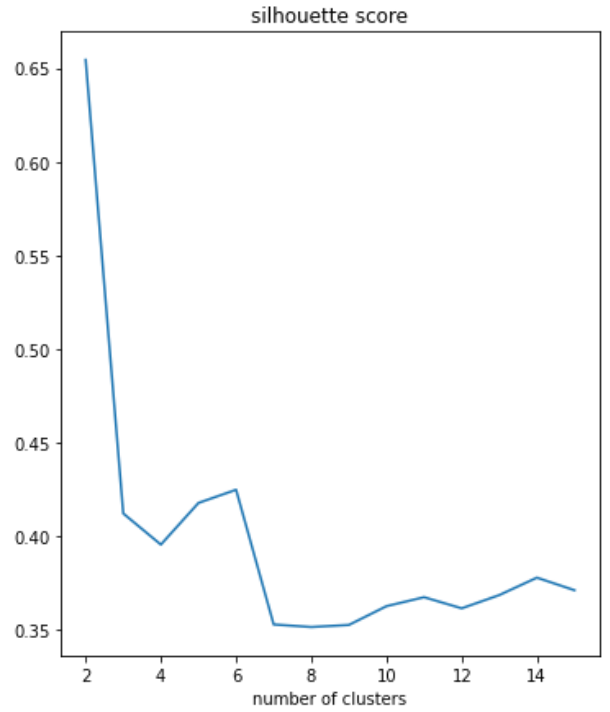
5.2. Seconde itération

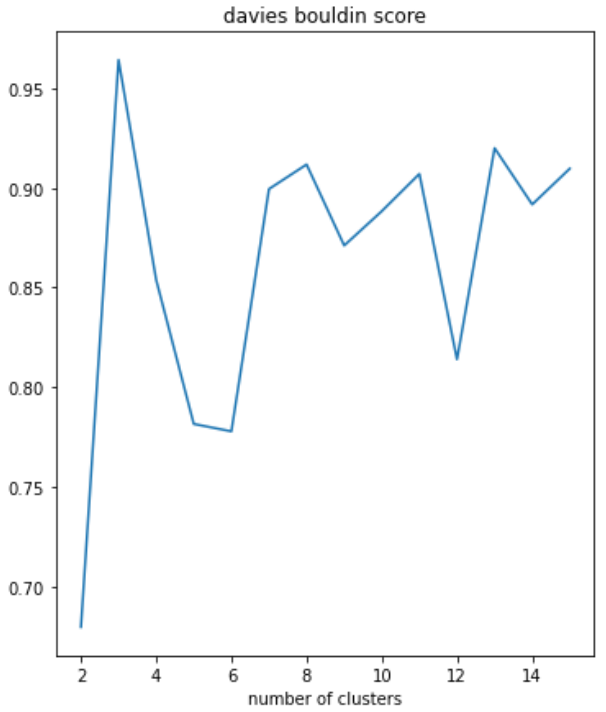
KMeans

Seconde itération – choix du nombre de clusters

Présentation des scores en fonction du nombre de clusters







Choix du nombre de clusters

- Nous observons un léger « coude » à 5 clusters pour le score d'inertie
- Nous obtenons de bons scores silhouette et davies bouldin à 5 ou 6 clusters
- Avec 6 clusters, il s'avère que le KMeans créé un cluster avec trop peu de clients, nous allons donc retenir k = 5

KMeans

Seconde itération – interprétation des résultats

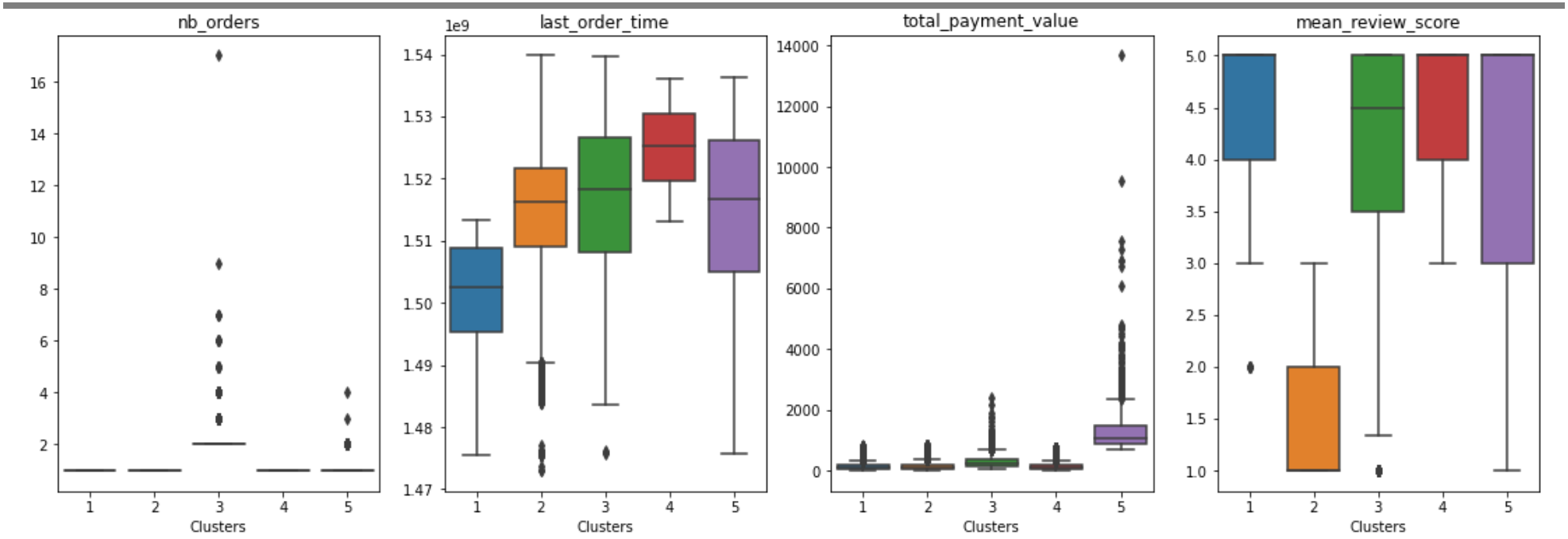
Répartition des clients par cluster

- Cluster 1 : 31 993 clients (33,3%)
- Cluster 2 : 16 714 clients (17,4%)
- Cluster 3 : 2 962 clients (3,1%)
- Cluster 4 : 42 398 clients (44,1%)
- Cluster 5 : 2 019 clients (2,1%)

Interprétation

	Récence	Fréquence	Montant	Satisfaction
Cluster 1	+	+	+	+++
Cluster 2	+	+	+	+
Cluster 3	++	+++	++	+++
Cluster 4	+++	+	+	+++
Cluster 5	++	++	+++	++

Distribution des variables par cluster

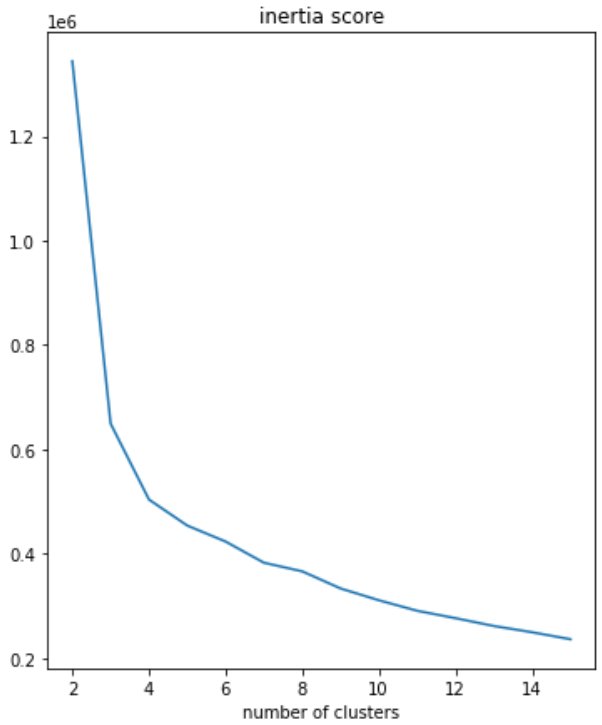


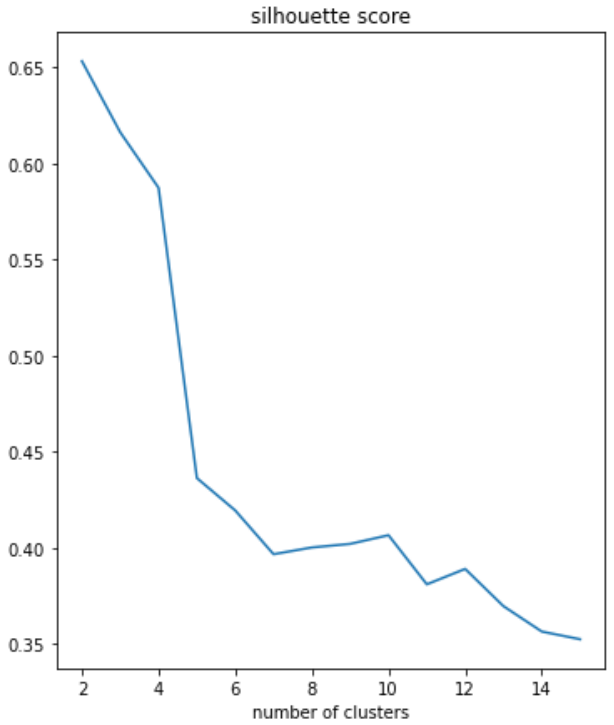
5.3. Troisième itération

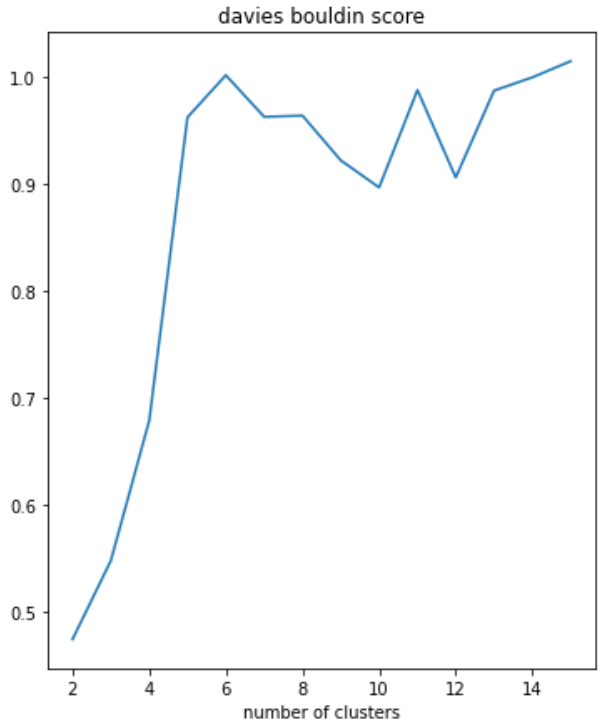
KMeans

Troisième itération – choix du nombre de clusters

Présentation des scores en fonction du nombre de clusters







Choix du nombre de clusters

- Nous observons un « coude » à 4 clusters pour le score d'inertie
- k = 4 donne également de bons scores silhouette et davies bouldin, nous retenons ce paramètre

KMeans

Troisième itération – interprétation des résultats

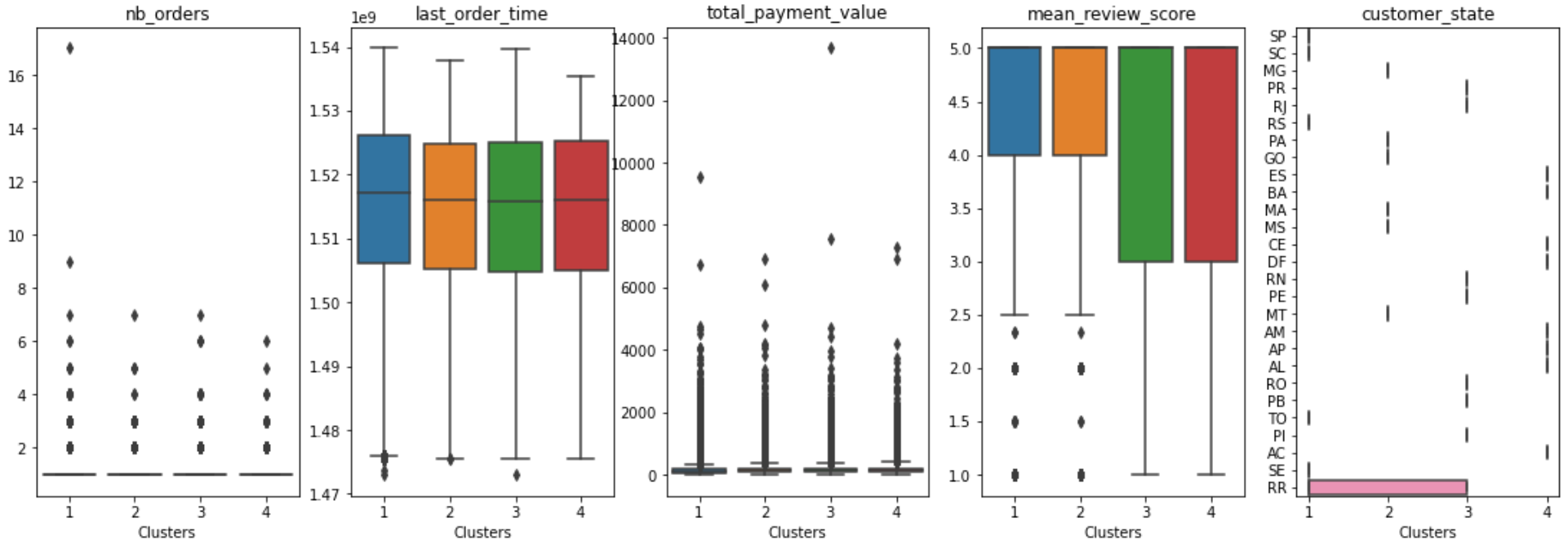
Répartition des clients par cluster

- Cluster 1 : 49 733 clients (51,8%)
- Cluster 2 : 16 448 clients (17,1%)
- Cluster 3 : 20 594 clients (21,4%)
- Cluster 4 : 9 311 clients (9,7%)

Interprétation

	Récence	Fréquence	Montant	Satisfaction	Localisation
Cluster 1	++	++	++	+++	**SP**
Cluster 2	++	++	++	+++	**MG**
Cluster 3	++	++	++	++	**RJ**
Cluster 4	++	++	++	++	**BA**

Distribution des variables par cluster



6. DBSCAN

DBSCAN

Méthodologie

Présentation du DBSCAN

Variables

- Utilisation des variables donnant un résultat optimal pour le KMeans :
  - Date de la dernière commande (Récence)
  - Nombre de commandes (Fréquence)
  - Montant total dépensé sur la plateforme (Montant)
  - Note moyenne donnée par commande (Satisfaction)

Paramètres

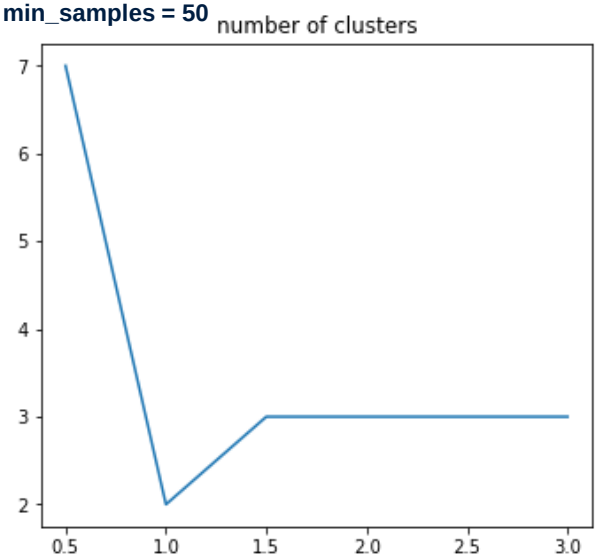
- Paramètres à optimiser :
  - Rayon de recherche autour de chaque point (avec la distance euclidienne)
  - Nombre minimum de points dans le rayon nécessaires pour qu'un point soit considéré dans un cluster
- Valeurs testées :
  - Rayon : test d'un rayon très large afin d'observer à partir de quelle valeur nous commençons à avoir plusieurs clusters. Cette valeur étant 3, nous testons ensuite les valeurs 0,5, 1, 1,5, 2, 2,5 et 3
  - Min sample : un nombre trop faible ne devrait pas nous permettre d'avoir des clusters de taille satisfaisante, nous utiliserons 50 et 100

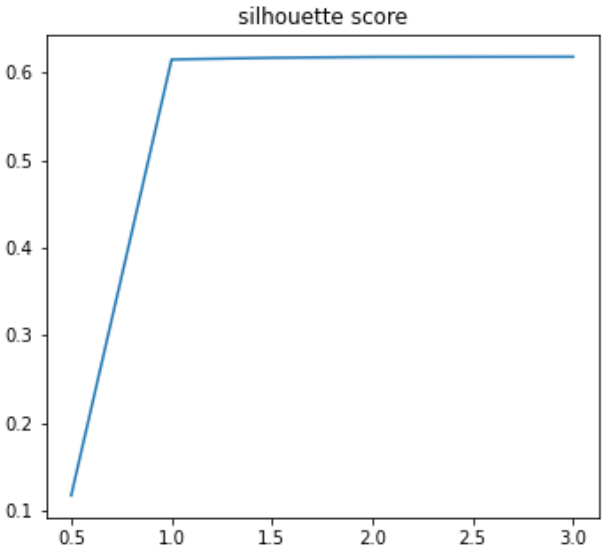
L'algorithme DBSCAN nécessitait une consommation de mémoire vive trop importante pour que je puisse le lancer sur le jeu de données complet avec mon ordinateur. J'ai donc effectué le travail sur la moitié du jeu de données (avec sélection aléatoire des points)

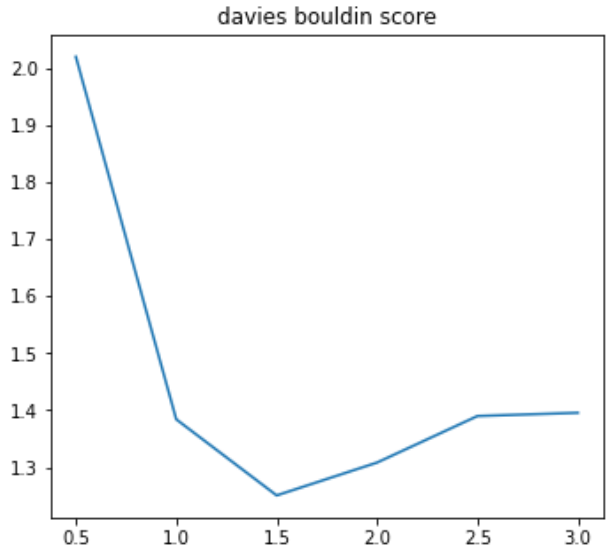
DBSCAN

Sélection des paramètres

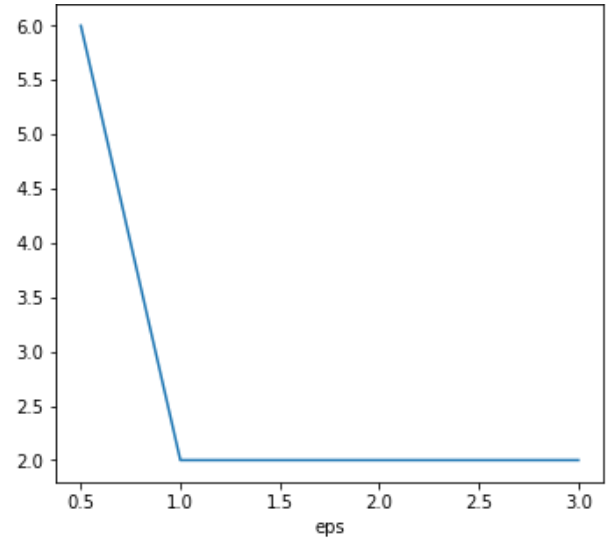
min_samples = 50

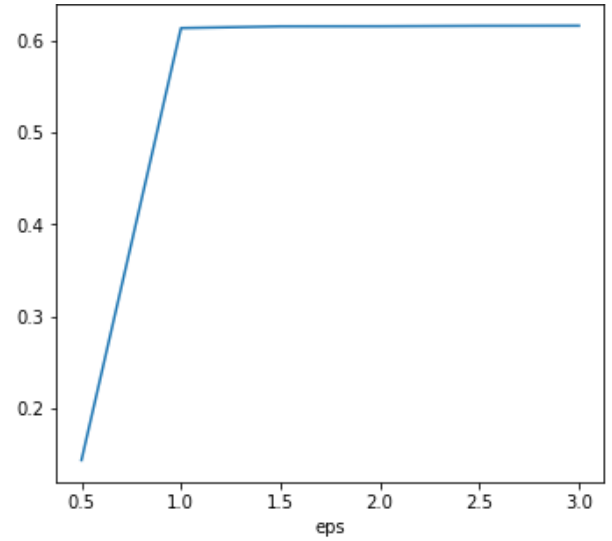


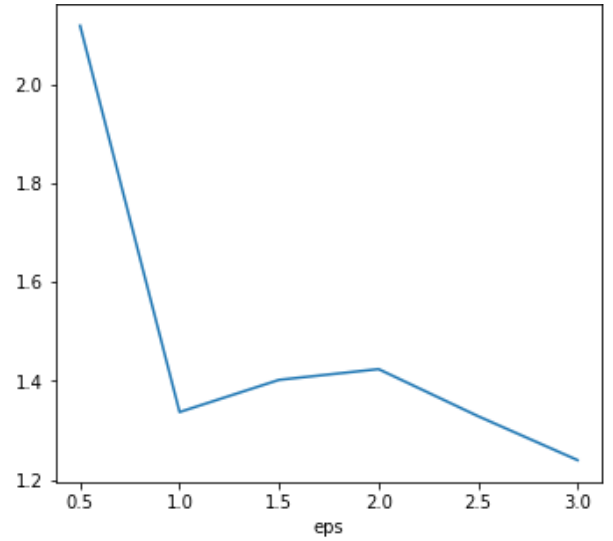




min_samples = 100







DBSCAN

Interprétation des résultats

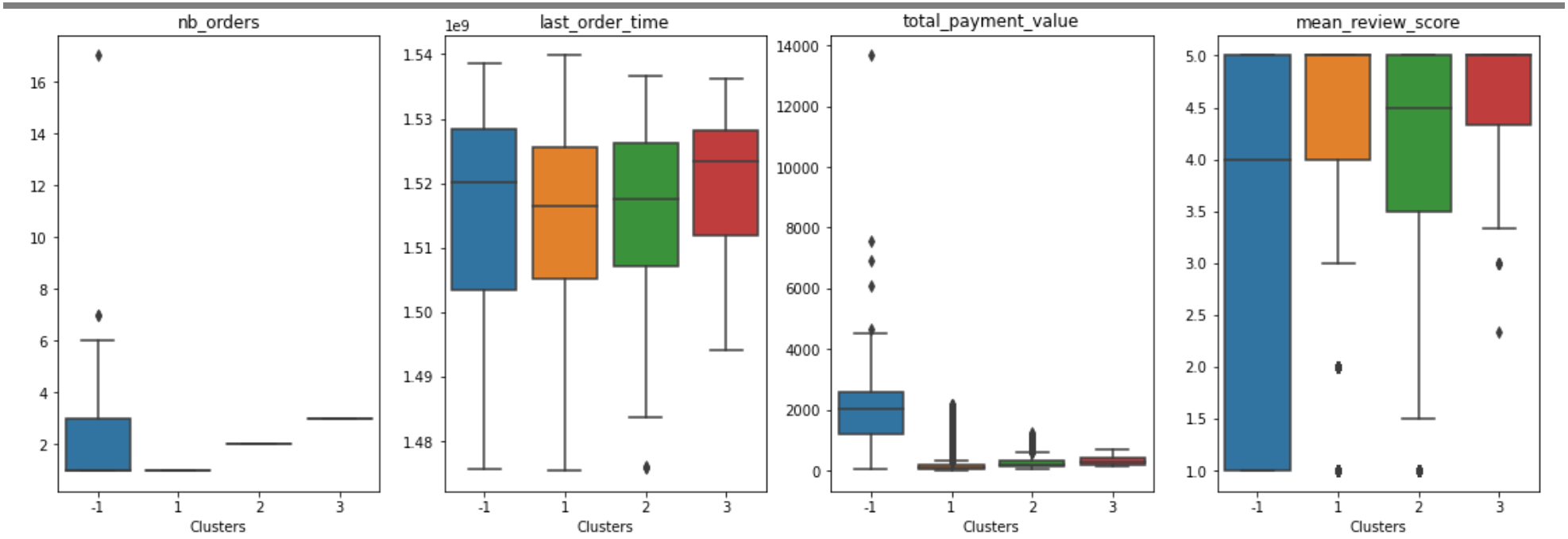
Répartition des clients par cluster

- Cluster 1 : 46 450 clients (96,7%)
- Cluster 2 : 1 347 clients (2,8%)
- Cluster 3 : 82 clients (0,2%)
- Les autres clients (0,3%) sont classifiés en tant que bruit (cluster -1)

Interprétation

	Récence	Fréquence	Montant	Satisfaction
Cluster 1	++	++	++	+++
Cluster 2	++	++	++	+++
Cluster 3	++	++	++	+++

Distribution des variables par cluster

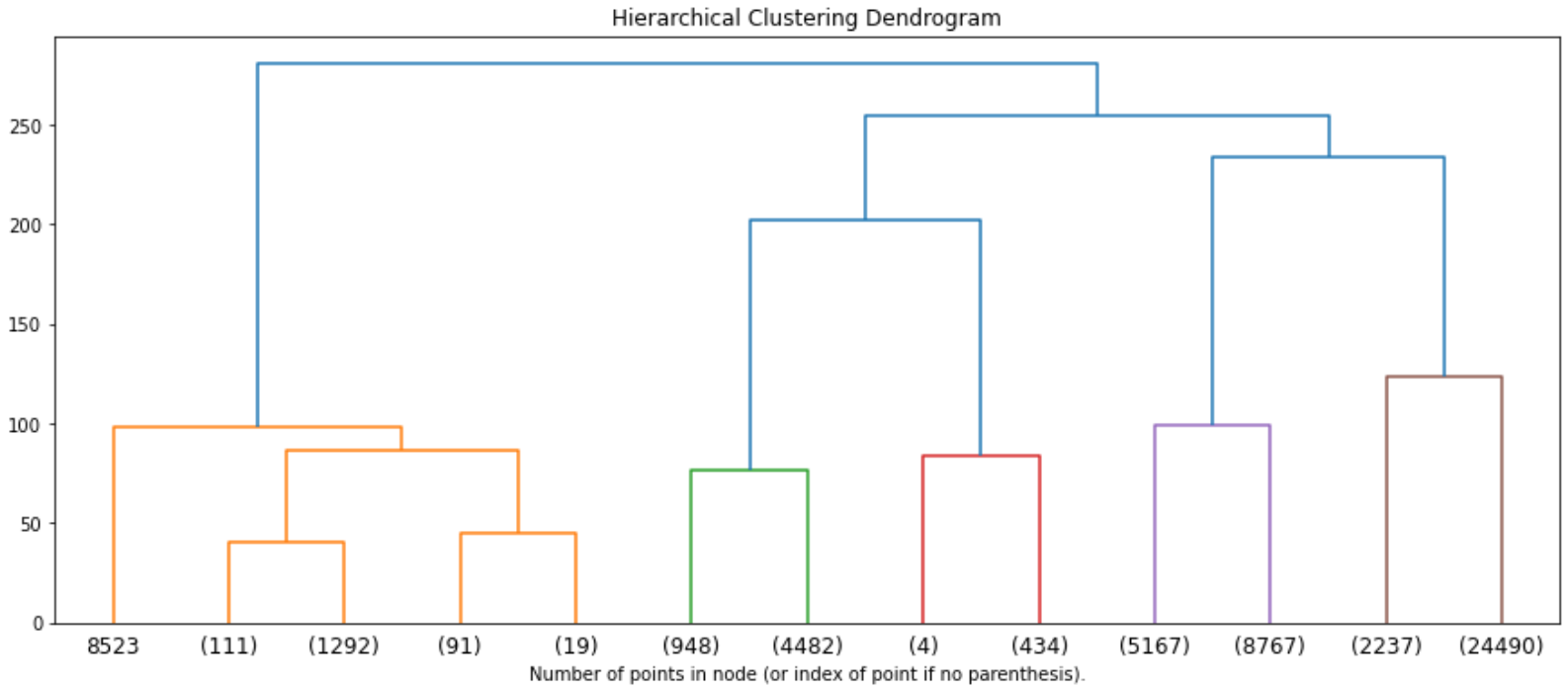


7. Clustering hiérarchique

Clustering hiérarchique

Sélection des paramètres

Dendrogramme du clustering hiérarchique



Choix du nombre de clusters

- Nous observons que le nombre de cluster maximisant la distance inter-clusters est 5
- Néanmoins, cela semble créer un cluster avec trop peu de clients (438), nous allons donc retenir 4 clusters qui semble également un bon candidat d'après ce dendrogramme
- **Comme pour l'algorithme DBSCAN, le clustering hiérarchique nécessitait une consommation de mémoire vive trop importante pour que je puisse le lancer sur le jeu de données complet avec mon ordinateur. J'ai donc effectué le travail sur la moitié du jeu de données (avec sélection aléatoire des points)**

Clustering hiérarchique

Interprétation des résultats

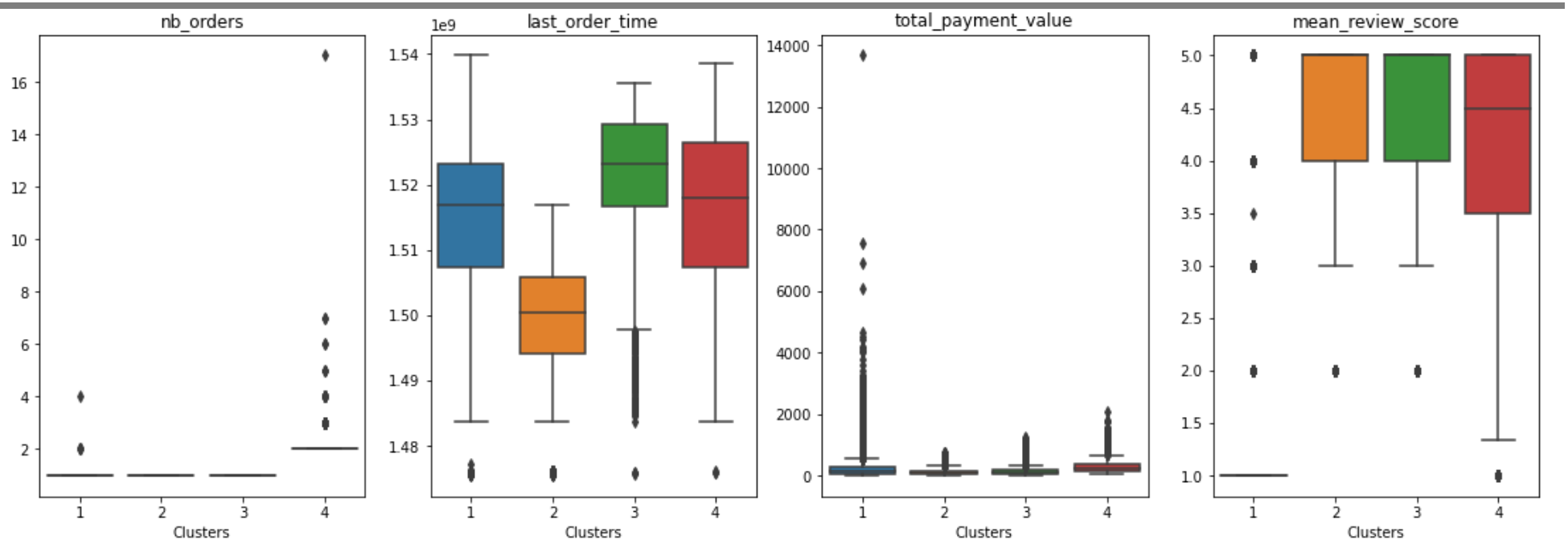
Répartition des clients par cluster

- Cluster 1 : 5 868 clients (12,2%)
- Cluster 2 : 13 934 clients (29,0%)
- Cluster 3 : 26 727 clients (55,6%)
- Cluster 4 : 1 514 clients (3,2%)

Interprétation

	Récence	Fréquence	Montant	Satisfaction
Cluster 1	++	+	++	+
Cluster 2	+	+	+	+++
Cluster 3	+++	+	+	+++
Cluster 4	++	+++	++	+++

Distribution des variables par cluster



8. Stabilité et maintenance du modèle

Stabilité et maintenance du modèle

Méthodologie

Présentation de la méthodologie d'évaluation de la stabilité du modèle

Objectif

- Calculer la durée à partir de laquelle un modèle de clustering devient obsolète considérant l'arrivée de nouveaux clients ou de nouvelles commandes effectuées par des clients existants
- Notre modèle a été conçu à partir de l'intégralité des données disponibles à fin 2018
- Nous allons essayer d'estimer la durée à partir de laquelle les prédictions de cet algorithme ne sont plus pertinente et nécessiteraient d'entraîner de nouveau l'algorithme sur les nouvelles données

Mesures

- Nous allons utiliser l'adjusted rand score, qui permet de comparer deux classifications sur le même jeu de données et d'estimer leur similarité
- Plus les classifications seront similaires, plus le score sera proche de 1
- Nous allons estimer que si le score est inférieur à 0,8, alors notre algorithme initial ne permet plus d'effectuer des prédictions pertinentes

Etapes

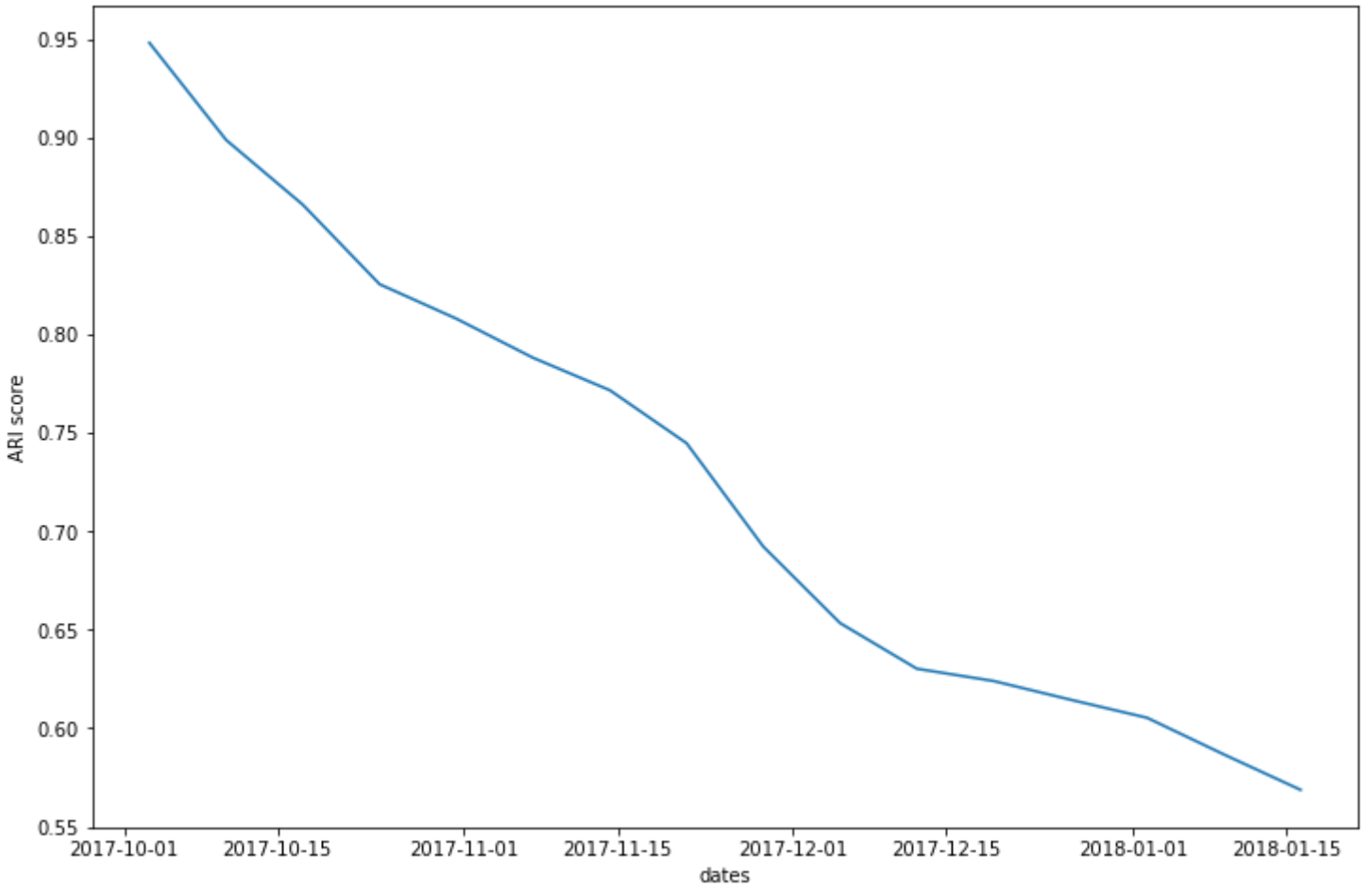
1. Créer et entrainer un modèle (modèle initial) à une certaine date, sur la base des données disponibles à cette date
2. Une semaine plus tard, le jeu de données va être modifié avec l'intégration de nouvelles commandes sur cette semaine :
   - Effectuer une prédiction avec le modèle initial sur ce nouveau jeu de données
   - Entraîner un nouveau modèle sur ce nouveau jeu de données
   - Comparer les prédictions de deux modèles avec un adjusted rand score
3. Répéter l'opération jusqu'à obtenir un ARI < 0,8

Stabilité et maintenance du modèle

Première itération

Présentations des résultats de la première itération

Date du modèle initial	26/09/2017
Date de dépassement du seuil	07/11/2017
Temps d'obsolescence en jours	42
Temps d'obsolescence en semaines	6

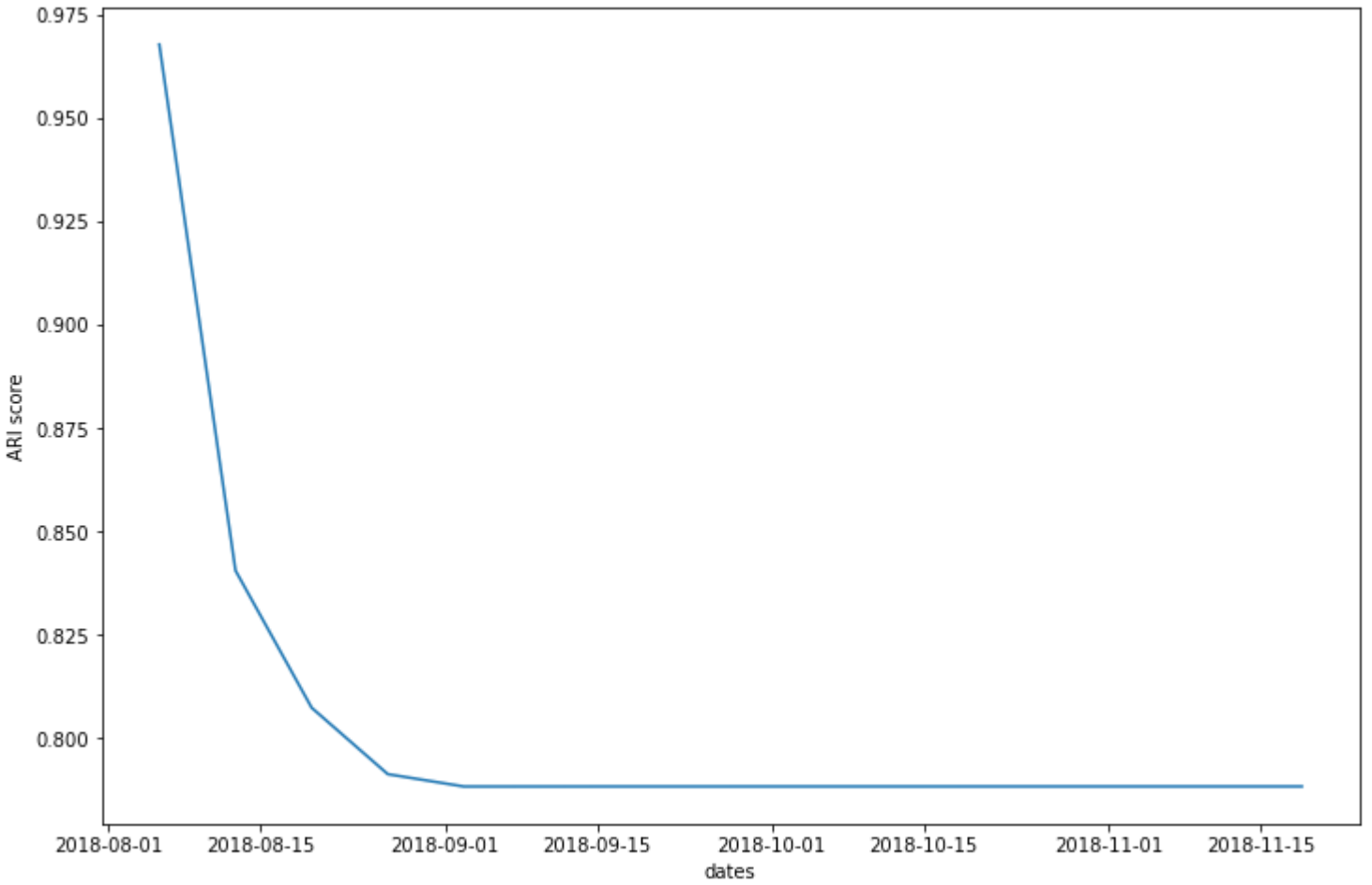


Stabilité et maintenance du modèle

Seconde itération

Présentations des résultats de la seconde itération

Date du modèle initial	29/07/2018
Date de dépassement du seuil	26/08/2018
Temps d'obsolescence en jours	28
Temps d'obsolescence en semaines	4



9. Conclusion

Conclusion

Sélection du modèle

Performances des modèles utilisés

	KMeans	DBSCAN	Clustering hiérarchique
Silhouette score	0,42	0,62	0,37
Davies Bouldin score	0,78	1,25	0,92
Nombre de cluster optimal	5	3	4
Temps de calcul	■ Entraînement : 0,73s ■ Prédiction : 0,04s	■ Entraînement : 17,67s (50% du dataset) ■ Prédiction : n.a.	■ Entraînement : 314,17s (50% du dataset) ■ Prédiction : n.a.
Commentaires	■ Algorithme rapide dans l'exécution ■ Scores moyens mais segmentation intéressante d'un point de vue métier	■ Consommation de mémoire vive trop importante : nécessité d'utiliser seulement la moitié du jeu de données ■ Un cluster dispose de trop peu de clients (0,17%)	■ Consommation de mémoire vive trop importante : nécessité d'utiliser seulement la moitié du jeu de données

- Bien que ses scores soient bons l'algorithme DBSCAN ne permet pas d'effectuer une segmentation satisfaisante pour une interprétation métier avec ce jeu de données
- Le clustering hiérarchique et le KMeans permettent une segmentation intéressante, toutefois le KMeans affiche des temps de calcul bien inférieurs et permet par ailleurs d'utiliser l'intégralité du jeu de données : nous retenons donc le KMeans

Conclusion

Interprétation des clusters et suggestions d'actions

Présentation des résultats de la segmentation client

	# clients	Récence	Fréquence	Montant	Satisfaction	Typologie	Actions possibles
Cluster 1	31 993 clients (33,3%)	+	+	+	+++	Clients « one-shot », satisfaits	Grosse promotion pour les attirer de nouveau sur le site
Cluster 2	16 714 clients (17,4%)	+	+	+	+	Clients « one-shot » insatisfaits	Aucune action : clients qui ne reviendront à priori pas sur le site
Cluster 3	2 962 clients (3,1%)	++	+++	++	+++	Bons clients (dépenses moyennes mais régulières) satisfaits	Campagne de mailing régulière
Cluster 4	42 398 clients (44,1%)	+++	+	+	+++	Nouveaux clients satisfaits	Suivi actif avec promotions pour fidélisation
Cluster 5	2 019 clients (2,1%)	++	++	+++	++	Bons clients (grosses dépenses irrégulières) moyennement satisfaits	Campagne de mailing régulière, sondages sur des possibles axes d'amélioration

Points d'attentions / limites sur la segmentation

- Le jeu de données contient seulement 3% de clients avec plus d'une commande : cela complexifie la tâche de la classification et l'interprétation métier
- Le silhouette score montre que la séparation des clusters n'est pas parfaite
- La durée de maintenance est faible, montrant que le modèle se dégrade relativement rapidement et nécessite d'être réentraîné régulièrement