Segmentez des clients d'un site e-commerce

Parcours Data Scientist
Projet 5

Matthieu Gschwend

Présentation de la problématique

- **Olist** souhaite que vous fournissiez à ses équipes d'e-commerce une **segmentation des clients** qu'elles pourront utiliser au quotidien pour leurs campagnes de communication.
- Vous devrez **fournir à l'équipe marketing une description actionable** de votre segmentation
- Ainsi qu'une **proposition de contrat de maintenance** basée sur une analyse de la stabilité des segments au cours du temps

- **Mise en situation :** Je présente mon travail à un collègue, ces critiques et suggestions amélioreront la pertinence de mon travail.

Sommaire

1. Analyse exploratoire
   1.1. Description du jeu du jeu de données
   1.2. Première segmentation

2. Apprentissage non supervisé
   2.1. Descriptions des méthodes de clustering
   2.2. Présentation des Résultats

3. Contrat de maintenance
4. Conclusion et ouverture

Analyse exploratoire

Description des jeux de données

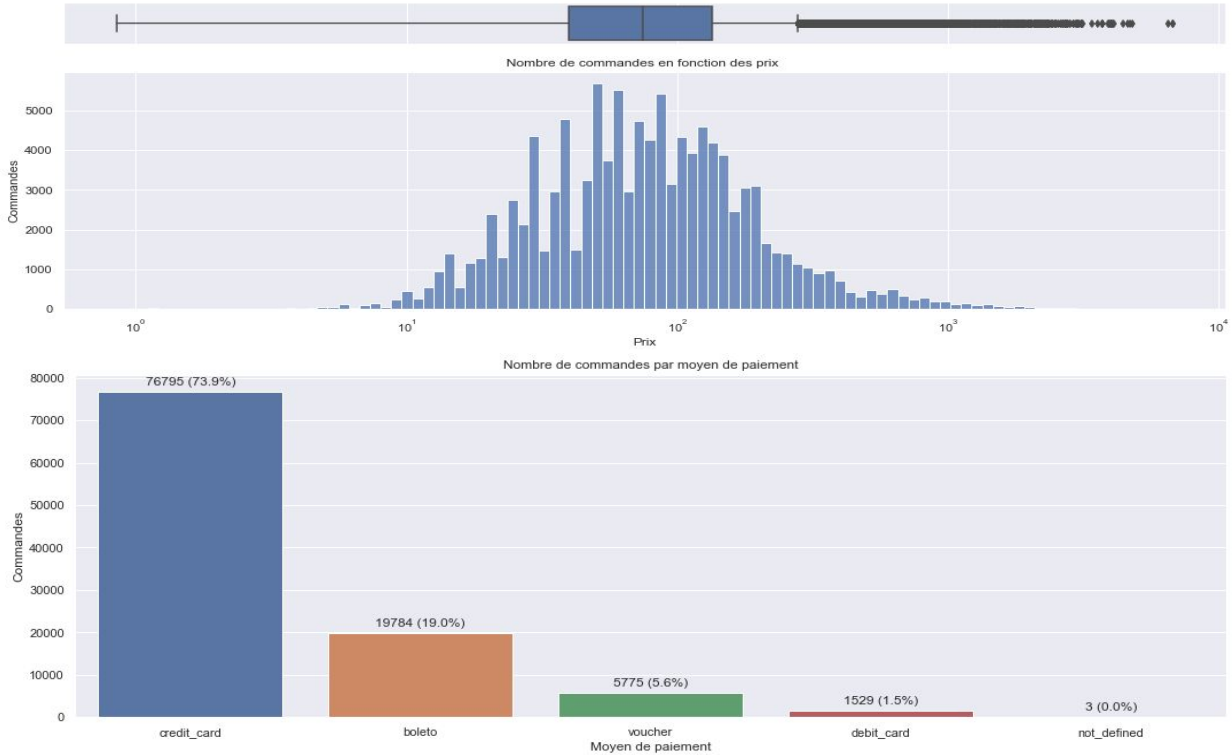
- olist_customer :
  - dimension (99441, 5)
  - identifiants des clients/adresse
- olist_geolocation :
  - (1000163, 5)
  - adresse détaillée à partir du code dans olist_customer
- olist_orders :
  - (99441, 8)
  - Liste des commandes reliées à un client. Infos sur les dates de commande/livraison
  - octobre 2016 à août 2018
- olist_order_items:
  - (112650, 7)
  - liste des produits reliée à une commande et à un vendeur. Prix du produit frais de port
- olist_order_payments:
  - (103886, 5)
  - méthodes de payement reliées à une commande.
  - supprimer les 3 not defined
  - pour les paiements en plusieurs fois faire des groupes : 1, [2,5], [5 et plus]
- olist_order_reviews:
  - (99224, 7)
  - note de satisfaction client reliée à une commande
- olist_products:
  - (32951, 9)
  - détails des produits (catégories, dimensions)
  - regrouper éventuellemnt les petites catégories dans des plus grandes (20 catégories représentent 85% de produits), nous allons le faire à l'étape suivante, car à postériorie il y a vaiment un trop grand nombre de catégories. En plus, certaines sont vraiment très similaires.
  - calculer plutot le volume pour ne pas garder hauteur, longueur, largeur
- olist_sellers:
  - (3095, 4)
  - Information sur les vendeurs en fonction de leur identifiant
- olist_translation:
  - (71, 2)
  - traduction de la catégorie des produits en anglais

Analyse exploratoire

Origine des clients

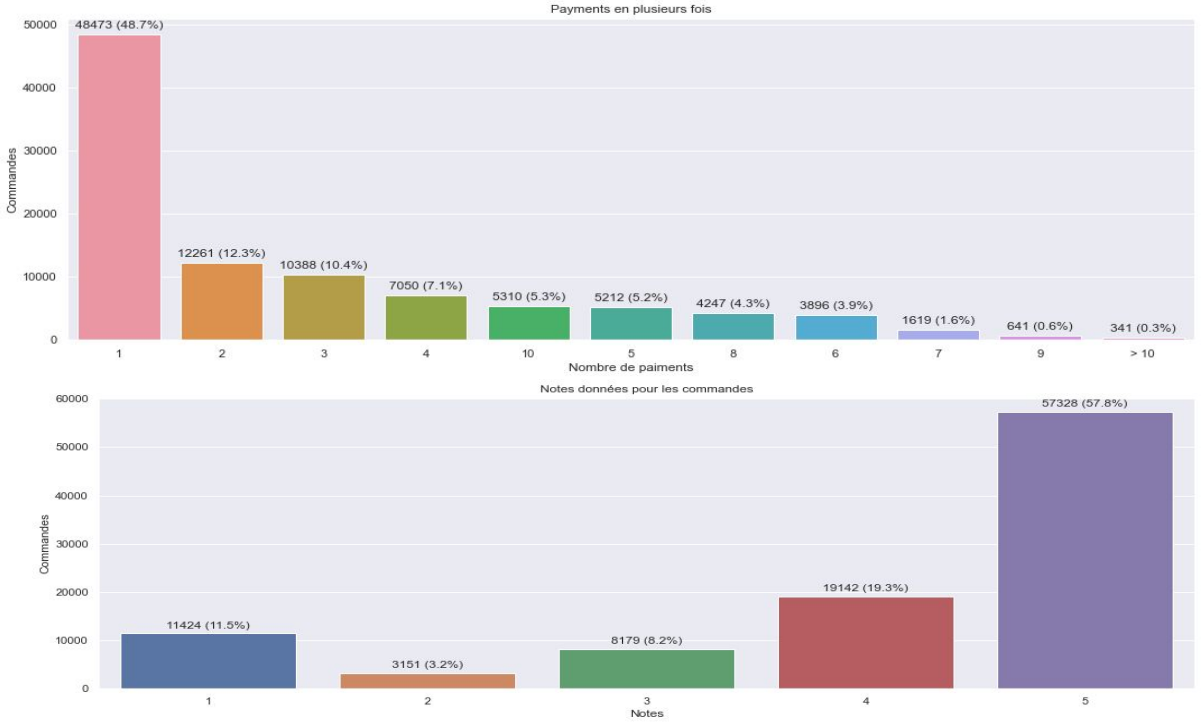
Analyse exploratoire.

Prix des paniers et moyens de paiement



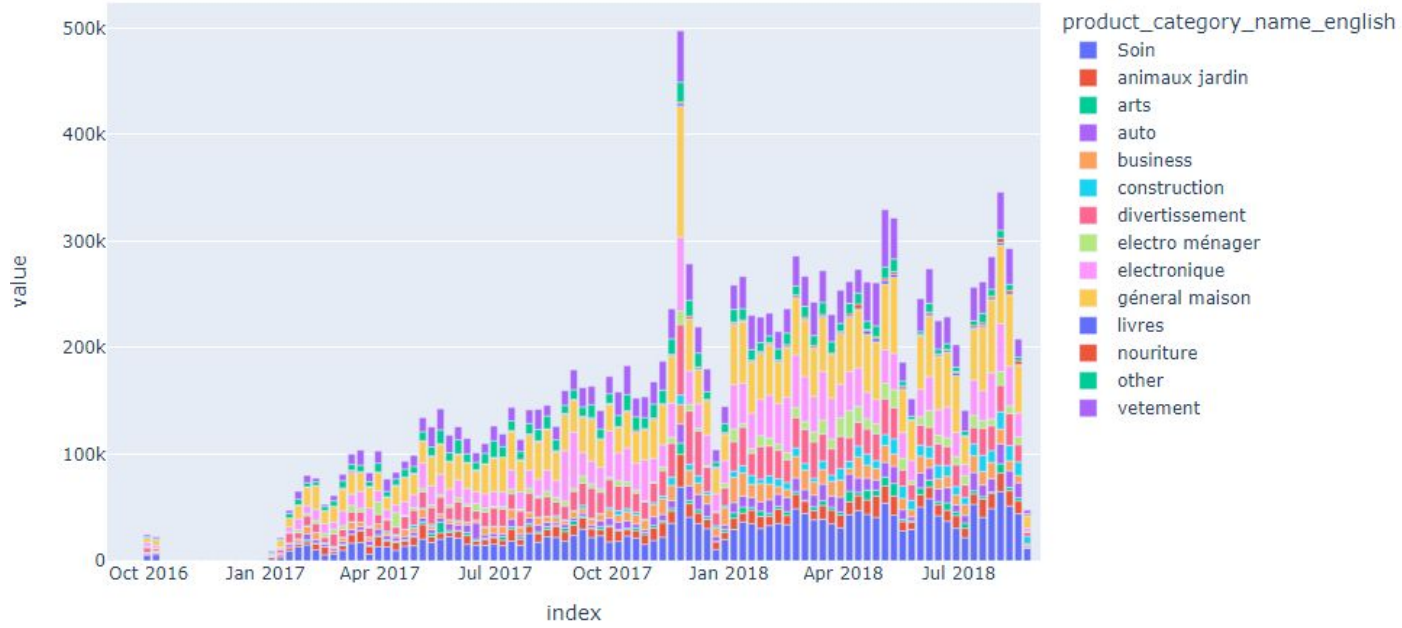
Analyse exploratoire.

Paiements en plusieurs fois, note de satisfaction



Analyse exploratoire.

Chiffre d'affaires par semaine



Analyse exploratoire.

Fréquence des commandes par client

Une première segmentation RFM

Pour initialiser la démarche de clustering nous mettons en place une première segmentation basée sur le récence, la fréquence et montant des dépenses associés aux clients :

Fonctionnement :
Nous attribuons une note entre 1 et 3 pour chaque caractéristique RFM.

- **Fréquence :** 1 si une unique commande, 2 sinon
- **Récence :** 3 si < 6 mois, 2 si < 1 an, 1 sinon
- **Montant :** Fonctionnement par tertiles, 3 premiers tiertiles, 2 deuxième, 1 dernier.

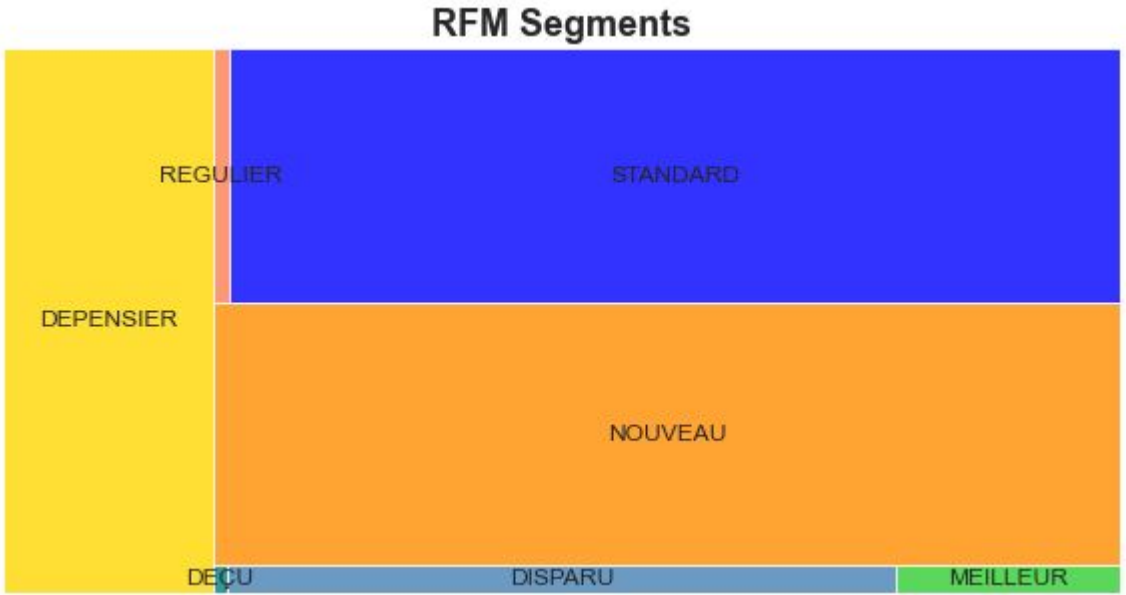
Résultats de la segmentation RFM

Il y a 18 combinaisons possibles : - 3 pour la récence (1 : inactif, 2 : standard, 3 : actif) - 2 pour la fréquence (1 : unique, 2 : fréquent) - 3 pour le montant (1 : faible, 2 : standard, 3 : dépensier)

- MEILLEUR "323" - les plus récents, les plus fréquents, les plus générateurs de revenus - les clients de base qui doivent être considérés comme les clients les plus précieux.
- DISPARU - '111', '112', '113' - des clients uniques il y a longtemps- ces clients ont probablement disparu.
- NOUVEAU - '311', '312', '313' - viennent de s'inscrire - nouveaux clients qui se sont inscrits récemment
- DEPENSIER - '213', '223 - les plus générateurs de revenus - ceux qui génèrent les plus gros revenus et qui ne sont pas inactifs.
- REGULIER - '221', '222', '321', '322', - utilisateurs fidèles ayant commandés aux moins deux fois, avec une commande relativement récente
- DEÇU - '121', '123','122', , - utilisateurs ayant passé plusieurs commandes mais qui n'ont pas recommandé depuis longtemps. On peut supposer qu'il ont été peu satisfait de leurs dernières commandes et qu'il sont allés chez la concurrence.
- STANDARD - '212', '211' rien à déclarer sur ces utilisateurs, commande ni très récente ni ancienne, montant peu élevés

Taille des segments

DEÇU	58
REGULIER	636
MEILLEUR	909
DISPARU	2706
DEPENSIER	17244
STANDARD	34374
NOUVEAU	36049



Apprentissage non supervisé

Dans la partie précédente nous avons effectué un premier clustering. Cette méthode comporte néanmoins des défauts :

1. Nous avons défini de manière relativement subjective nos seuils pour l'attribution des scores (1-2-3)
2. Augmenter le nombre de caractéristiques risque devenir vraiment lourd d'un point de vu métier (si on ajoute par exemple la note de satisfaction on se retrouve avec 18 * 3 = 54 catégories de clients)
3. La structure des données n'est pas prise en compte. Toute notre segmentation se fonde sur un raisonnement a priori des différents profils de client.

Changement de stratégie :

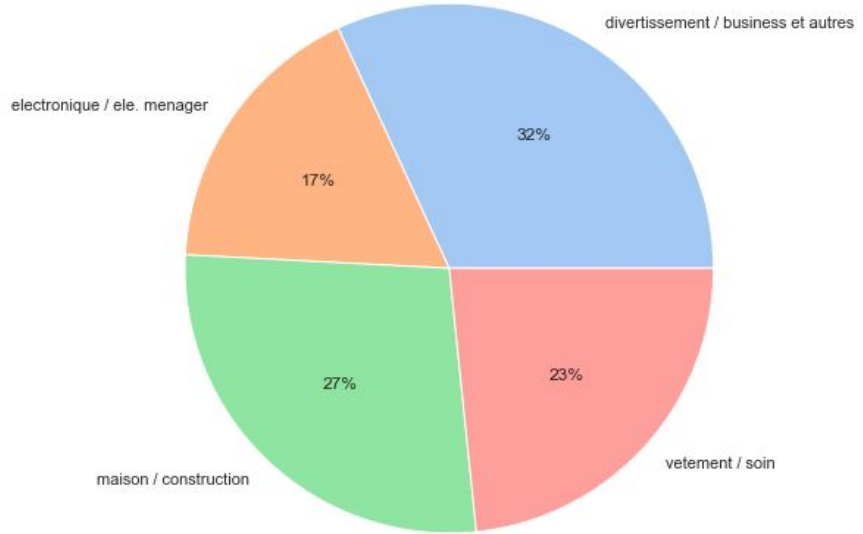
L'utilisation des méthodes d'apprentissage non supervisée considère la structure de notre jeu de données pour ensuite en extraire des sous groupes (clusters).

La pertinence métier de ces clusters est ensuite analysée.

Apprentissage non supervisé
Construction du jeu de données

- La date de la première commande
- Le nombre de jours entre la première et la dernière commande
- La somme des dépenses.
- La région du vendeur le plus présent avec sa latitude/longitude
- Le poid moyen des articles achetés
- Le volume moyen des articles achetés
- La région du client et sa longitude/latitude
- Le nombre de payment_installments moyen
- La note de satisfaction moyenne
- Le moyen de paiement le plus utilisé
- Le nombre d'articles achetés par catégories

Un travail a consisté à réunir tous les produits en 4 grandes catégories pour que cela soit plus facilement interprétable



Apprentissage non supervisé
Démarche

Plan d'action:

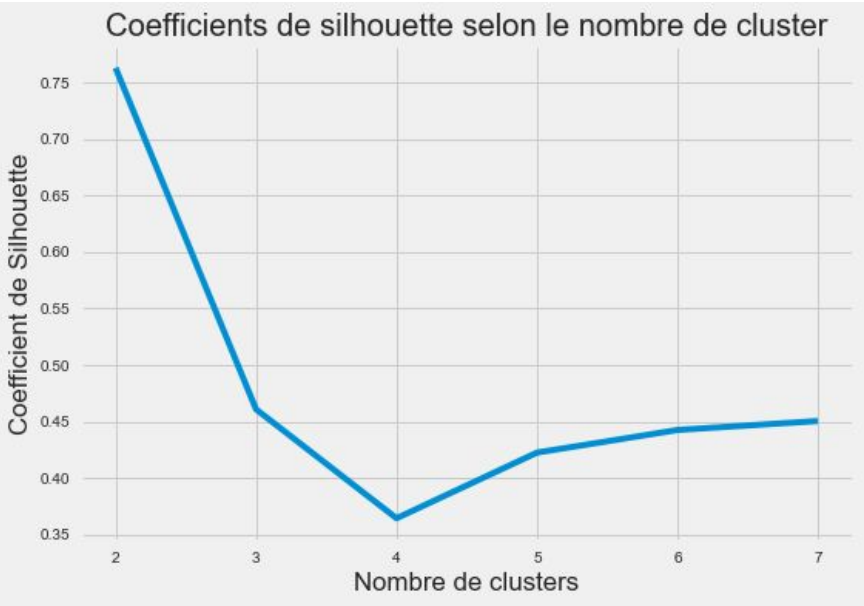
- Différentes méthodes de clustering avec différentes features.
- L'objectif reste bien de pouvoir identifier des clusters utiles du point de vue métier, nous analyserons donc les clusters pour en identifier leurs caractéristiques.

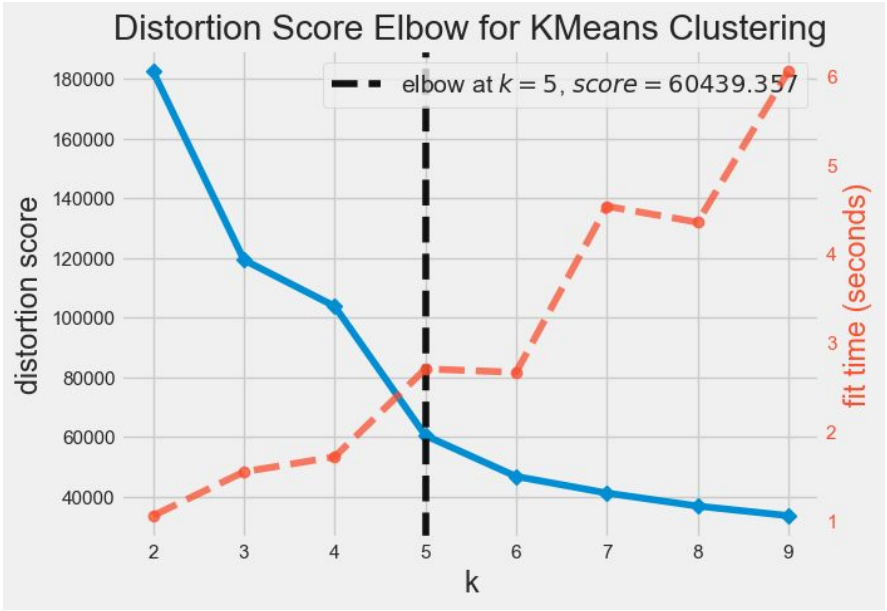
Dans un premier temps nous allons utiliser l'algorithme k-means avec les 3 features de la RFM. Nous ajouterons ensuite certaines caractéristiques en fonction de leurs pertinences.

Les méthodes DBSCAN et clustering agglomératif seront également étudiées.

Apprentissage non supervisé

K Means avec les variables RFM





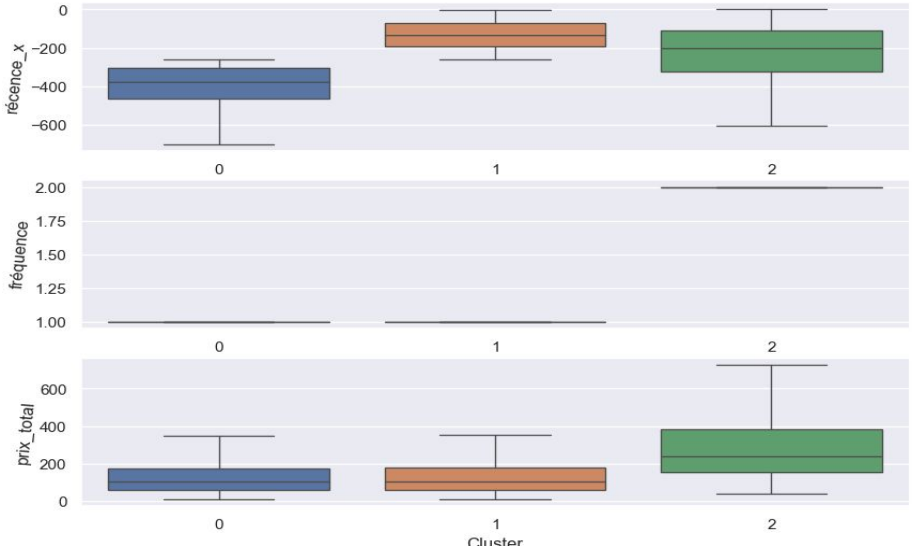
Apprentissage non supervisé
Choix du meilleur nombre de cluster

- Au delà de k = 5 le clustering comporte des clusters avec moins de 1000 individus
- Pour k = 4 le coefficient de silhouette est le plus faible et score de Davies Bouldin élevé.

Le choix se pose donc entre l'utilisation de 5 ou 3 clusters.

Nous constatons que le montant n'est pas déterminant pour la séparation des clusters

Cas K=3



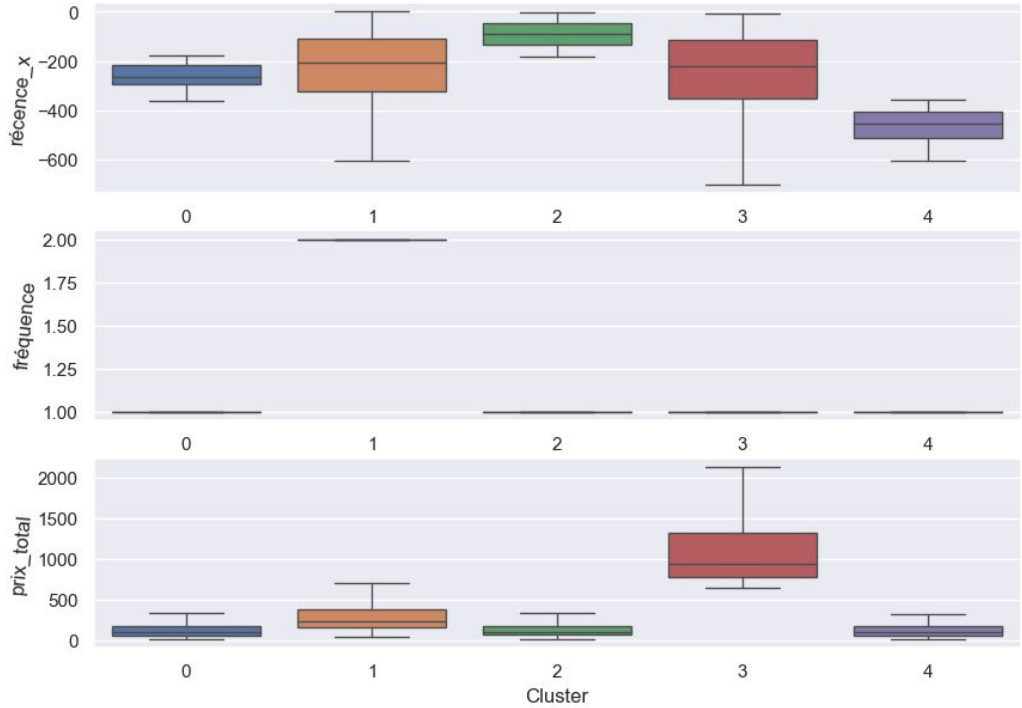
Apprentissage non supervisé

Etude du clustering retenu k = 5

Analyse :
La récence : 3 groupes se distinguent en fonction de la date du dernier achat :

- Le cluster 2 pour les clients les plus récents
- Le cluster 4 pour les clients les plus anciens
- Le cluster 0 pour les clients moyennement anciens
- La fréquence : le cluster 1 se différencie car il comporte les clients ayant commandé plus d'une fois
- Le montant : Le cluster 3 représente les clients qui dépensent le plus

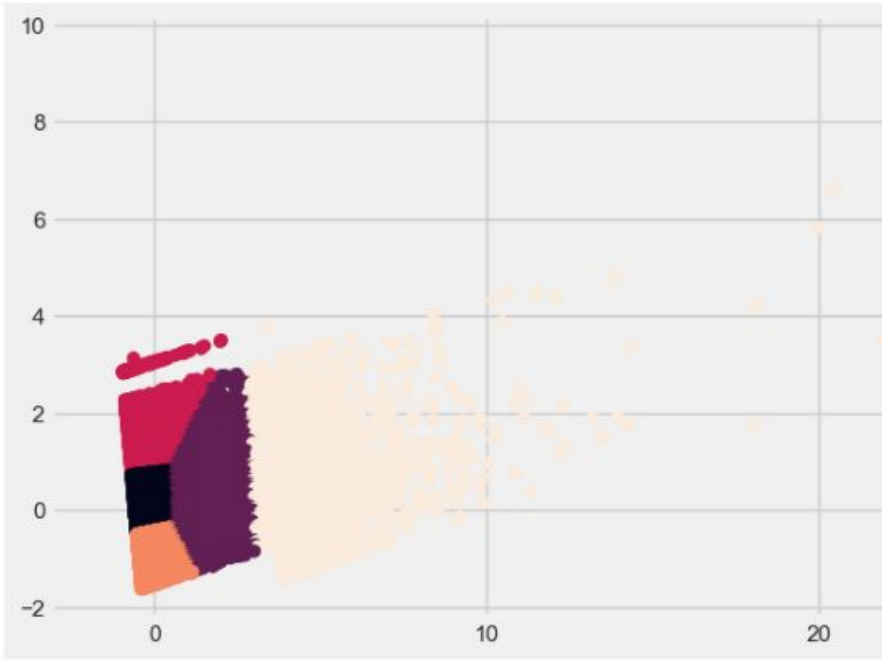
En conclusion : Avec k=5 nous avons obtenu un clustering simplement interprétable et pertinent d'un point de vue métier.



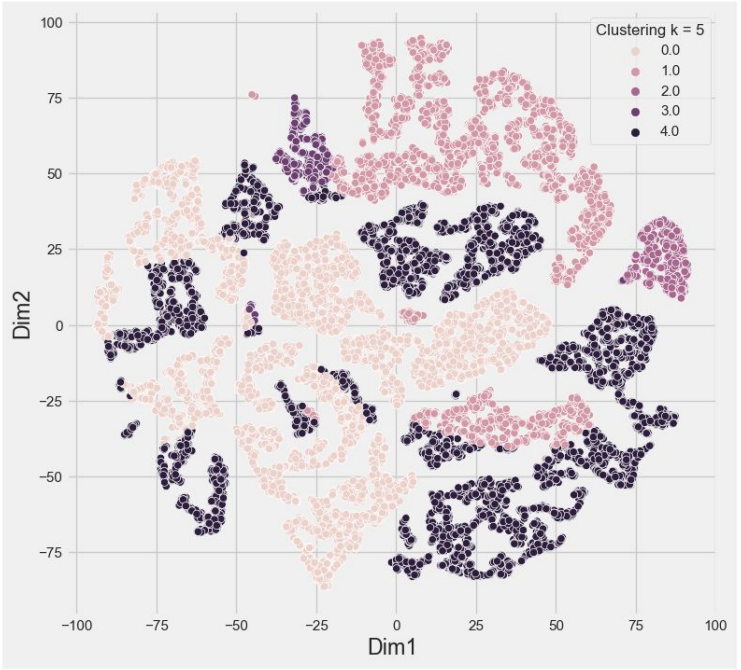
Apprentissage non supervisé

Visualisation des clusters

ACP

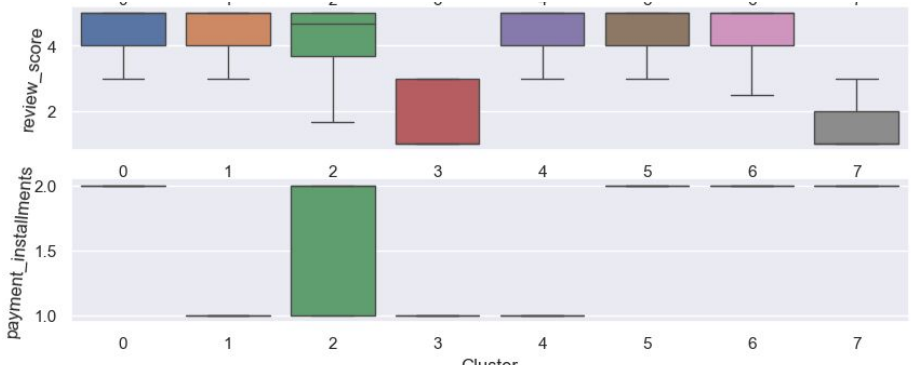


T-SNE

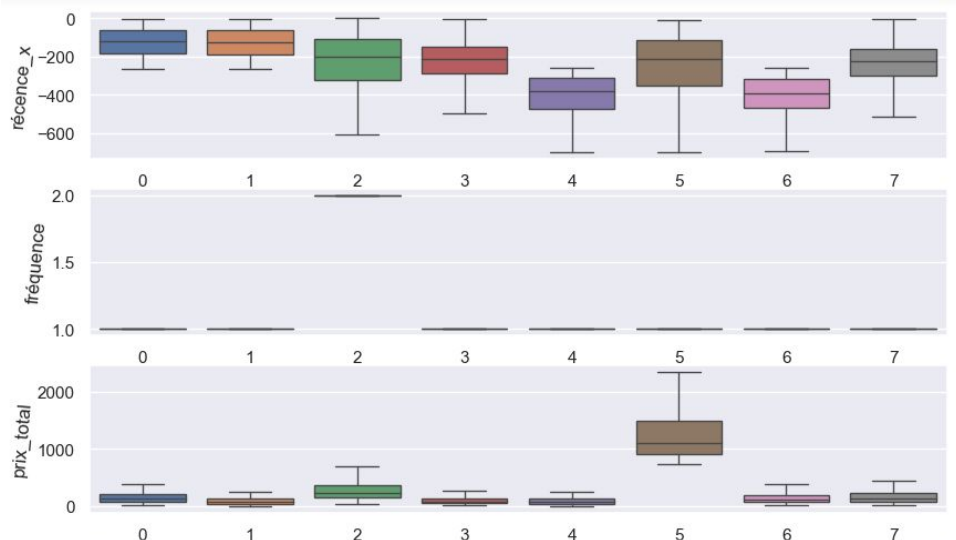


Apprentissage non supervisé
Augmenter le nombre de variables

Nous avons toujours notre cluster avec les clients qui ont passé plus d'une seule commande (cluster numéro 2) ainsi que celui dont le montant total est élevé (cluster 5).
Les clients insatisfait (3 et 7)
Les nouveaux clients (0 et 1)
Les anciens clients (4 et 6)



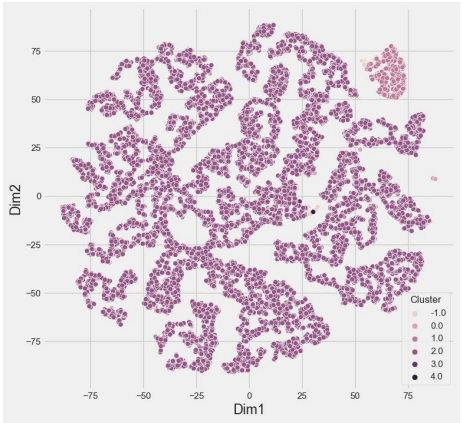
Nous ajoutons deux nouvelles variables à notre clustering : Le review score et le paiement en plusieurs fois.



Apprentissage non supervisé
D’autres méthodes : DBSCAN

L’algorithme DBSCAN fonctionne par densité. Deux individus peuvent être assignés au même cluster s’il existe une suite de points distants d’au plus ϵ permettant de les reliés.

- Cette approche n’a pas été fructueuse : définir le nombre de clusters a priori n’étant pas possible nous avons tâtonné pour trouver des bons hyperparamètres (Nombre de cluster raisonnables, peu de points non assignés à un cluster)



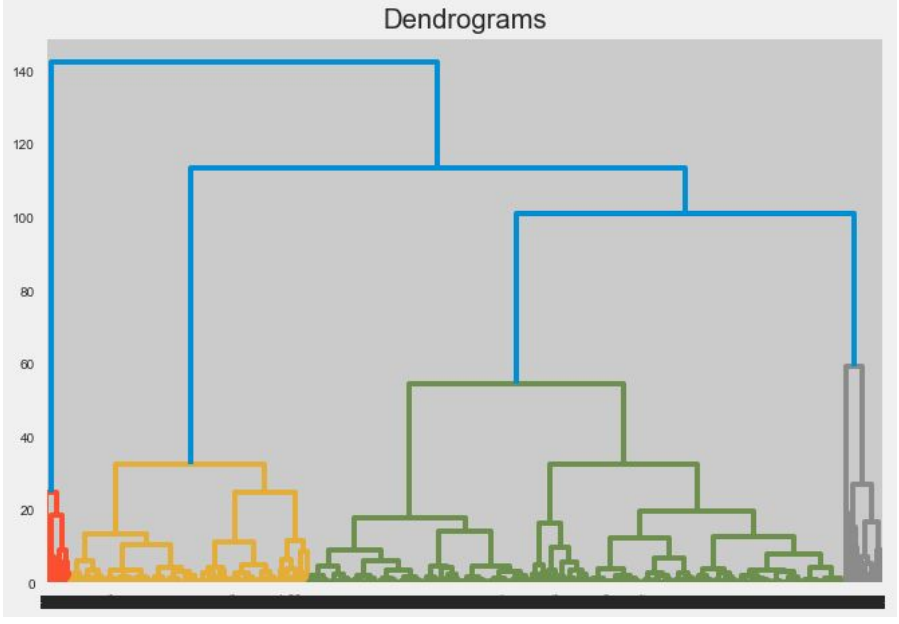

```
2    88666
1     2586
0      234
4       71
3       19
Name: labels,
```

Un cluster rassemble la grande majorité des points, nous avons abandonné cette méthode

Apprentissage non supervisé

D’autres méthodes : Clustering Agglomératif

Initialement chaque point est considéré comme un cluster. Les deux clusters les plus proches, et on les agglomère en un seul cluster. On répète cette étape jusqu'à ce que tous les points appartiennent à un unique cluster.



Les tests réalisés avec les 3 variables RFM, puis avec l’ajout des 2 autres variables nous montrent des résultats très similaires au K Means.

L'inconvénient de cette méthode est qu'elle nous force à travailler avec une petite partie des données (ici 10%) pour des raisons de temps de calcul.

Un des atouts de cette méthode est qu'elle permet de visualiser la distance d’un cluster à l’autre.

Exemple : ici les variables RFM : 4 clusters suggérés.

Contrat de maintenance
Introduction

L’entreprise nous suggère de lui indiquer la fréquence à laquelle nous devons actualiser notre clustering pour que celui-ci reste pertinent dans le temps.

Démarche :

Pour évaluer la qualité d’un clustering entraîné à l’instant T à l'instant T+1 nous utilisons 2 choses :

- La méthode predict de notre modèle de l’instant T sur les données présentes à T+1 pour obtenir clustering C(T,T+1)
- Faire un nouveau clustering à T+1 avec les données de T+1 : C(T+1,T+1).
  Calculer le score **ARI** entre C(T,T+1) et C(T+1,T+1)

Contrat de maintenance
Mise en pratique

- Nous décidons de commencer la simulation 10 mois avant la fin du jeu de données
- Construction d'une fonction permettant d'avoir un jeu de données antérieur à une date donnée
- Fonction permettant de diminuer automatiquement le Delta de simulation lorsque le score ARI passe en dessous d'un certain seuil.

Problèmes rencontrés : Notre approche est de faire cette simulation avec un Kmeans avec K=5. Or à la période du début de la simulation les 5 clusters ne constituent pas les mêmes groupes de clients. Nous utilisons donc k = 4 (cf notebook)

Contrat de maintenance

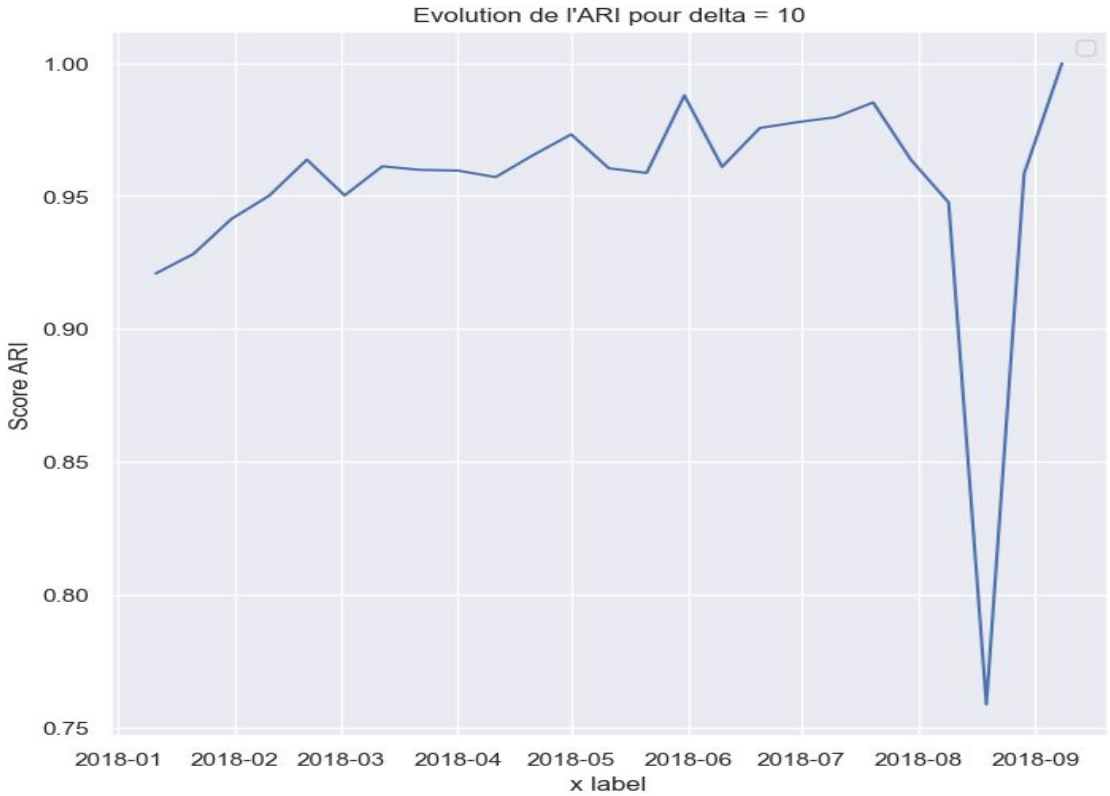
Evolution de l'ARI

La fonction de calcul automatique du delta nous donne un résultat surprenant :

L'ARI chute brutalement le 10 Août 2018.

Deux cas possibles :
-1: Il y a un gros outlier à cette date qui perturbe le standartscaler. (soit sur la fréquence soit sur le montant)

- 2 : Le problème provient de la récence, les clusters qui définissent les nouveaux clients des anciens se "touchent" et il y a une instabilité lorsqu'on ajoute de nouveaux clients.

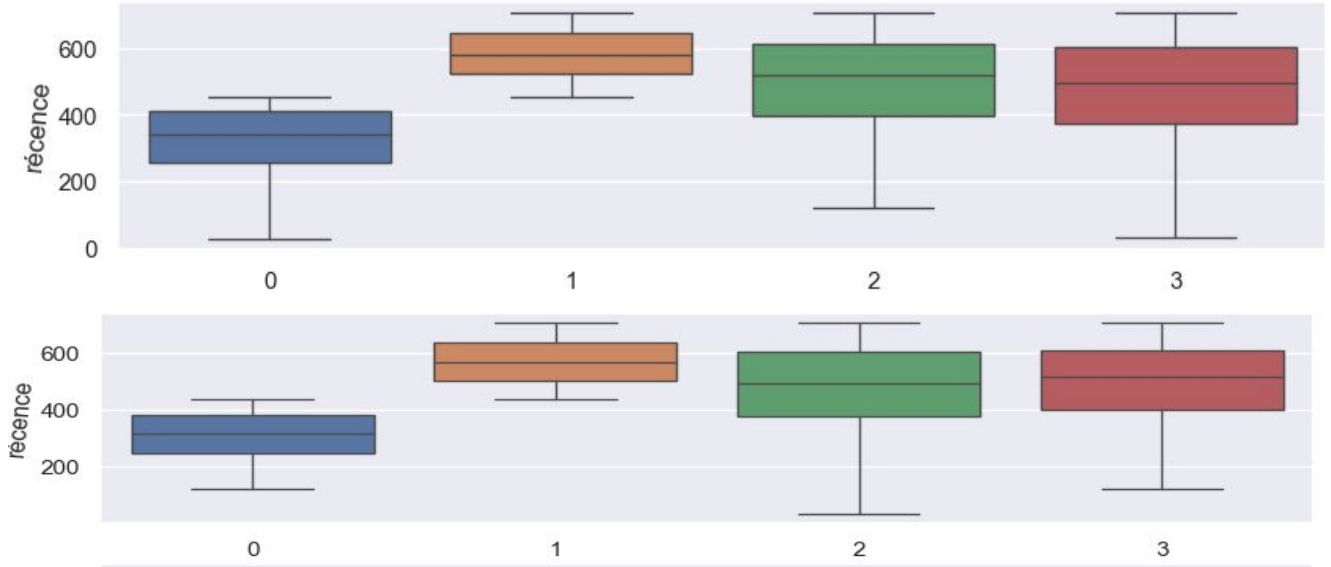


Contrat de maintenance

Outlier ou cluster instable ?

Nous n’avons pas d'outlier à cette période. Le phénomène reste le même si on enlève la semaine qui pose problème

Comparaisons des clusters juste avant la chute de l’ARI



Contrat de maintenance
Résultat

Nous avons donc décidé de recommencer la manœuvre en s'arrêtant juste avant la période qui pose problème.

Avec un seuil ARI = 0.75 nous obtenons un delta = 10 semaines.

Précisions :

- Plus il y aura de nouvelles données, plus le score ARI deviendra stable d'une période à une autre. Il faudra alors sans doute diminuer la fréquence des maintenances au fil du temps.

- Lorsque nous observerons une chute brutale de l'ARI se sera sans doute le bon moment pour ajouter des nouveaux clusters. Typiquement dans notre cas il faudrait passer à 5 clusters avec 3 pour la récence et non 2.

Conclusion et ouverture

- Suite à l'exploration des jeux de données nous avons identifié de nombreuses variables possibles pour segmenter les clients de l'entreprise.

- L'implémentation des différentes méthodes de clustering a permis d'identifier le Kmeans comme étant le plus adéquat.

- L'utilisation de beaucoup de variables rend difficile les interprétations des clusters.

- Le contrat de maintenance est pour l'instant fixé à 10 semaines.

- Il faudra travailler main dans la main avec l'équipe marketing de Olist :
  L'intérêt d'utiliser plus de variables pour le clustering.
  Évolution des clusters à chaque maintenance.