Segmentez des clients d'un site e-commerce

Réaliser une segmentation des clients de la plateforme brésilienne Olist

Date de la soutenance : 05/03/2024

Antoine Arragon

Objectifs

Réaliser une **segmentation des clients** de la plateforme Olist afin d'**aider l'équipe marketing** à mieux **cibler ses campagnes** de communication.

- ➢ Compréhension et mise en forme du jeu de données et création des variables pertinentes (Via SQL puis Python) ;

- ➢ Entrainement d'algorithmes non supervisés en vue d'établir un clustering pertinent (RFM puis d'autres variables telles que la satisfaction) ;

- ➢ Objectifs :
  - Aboutir à une segmentation satisfaisante à la fois selon certains indicateurs théoriques et pertinentes et aisément applicable d'un point de vue pratique.
  - Etablir une proposition de contrat de maintenance basé sur l'analyse de la stabilité du modèle de clustering retenu.

Plan :

1. Présentation du jeu de données et analyse exploratoire
    1.1. Jeu de données Olist
    1.2. Analyse exploratoire

2. Méthodologie et résultats des différents modèles de clustering
    2.1. Les différents modèles : KMeans, Clustering Hiérarchique et DBSCAN
    2.2. Evaluation des performances
    2.3. Itérations, base RFM puis ajout et suppression de variables
    2.4. Meilleurs modèles
    2.5. Stabilité

3. Maintenance du modèle choisi

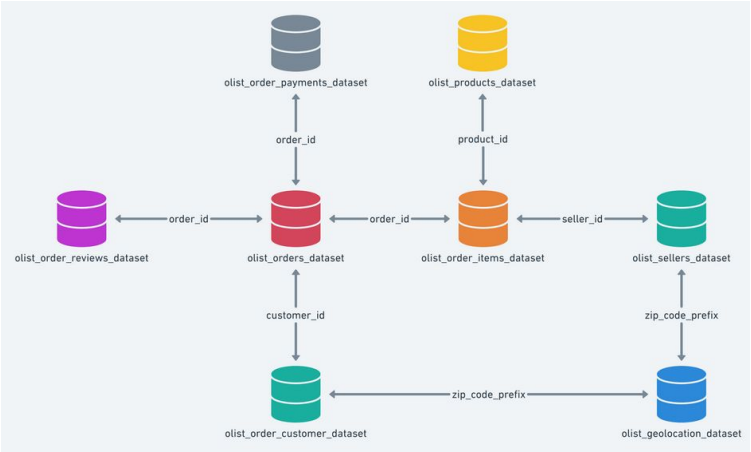
Conclusion

1. Présentation du jeu de données et analyse exploratoire

1.1. Jeu de données Olist

Provenance des données : plateforme Olist - téléchargement du jeu de données mis à disposition

- 9 tables de données correspondant à différents éléments : les clients, les vendeurs, les produits, les commandes, leur prix, etc.



- Travail de jointure et de transformation de variables, en SQL, afin d'aboutir à un jeu de données comprenant 96096 lignes (une par client), et 16 variables.
- Les données vont du **04/09/2016** au **17/10/2018**.

1. Présentation du jeu de données et analyse exploratoire

1.2. Analyse exploratoire

- **Types des variables** : **numériques** (nombre de commandes, montant dépensé), **catégorielles** (type de produits commandés, types de paiement utilisés) ;

- **Suppression des données manquantes** (très faible taux), jeu de données “final” : **91481 lignes**

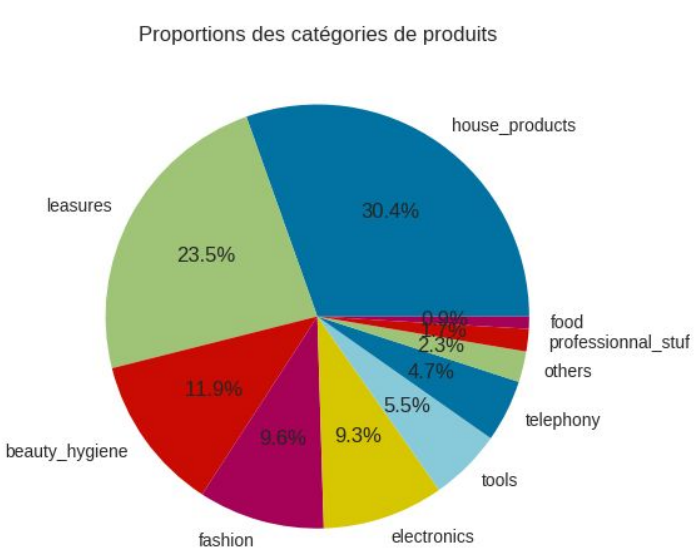
- Réduction du nombre de modalités concernant le type de produit, **création de la variable “recency”** pour l’**analyse RFM**. Création de variables **binaires** : retard de livraison ou non, paiement échelonnées ou non, etc. **23 variables au total**.

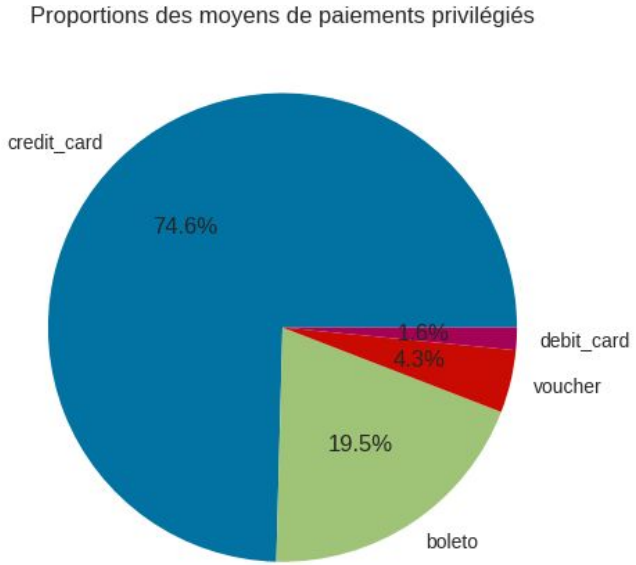
- Sélection d’un **sous-échantillon** de données pour l'entraînement ultérieur des **modèles de clustering** : conservation de l’ensemble des clients ayant réalisé plusieurs commandes et sélection aléatoire de 20 % des autres individus : 20667 lignes.

1. Présentation du jeu de données et analyse exploratoire

1.2. Analyse exploratoire

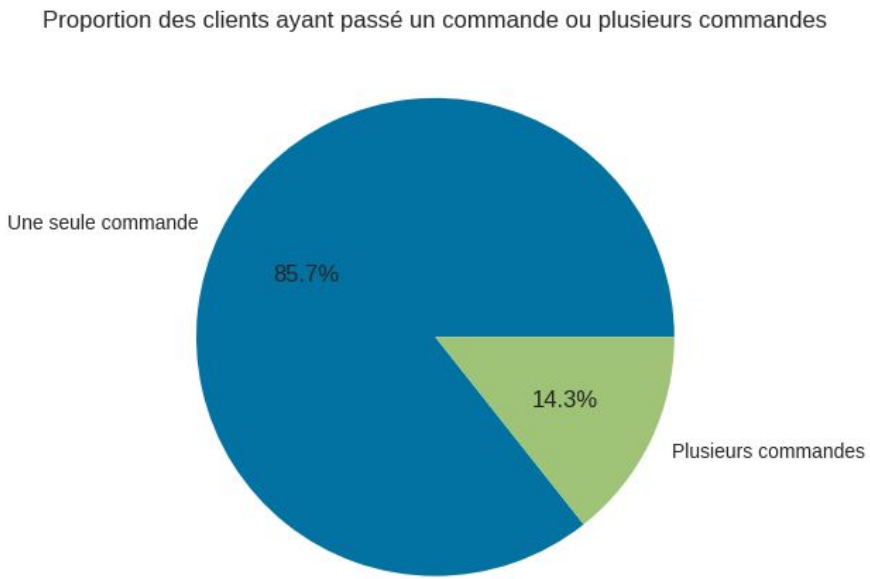
Distribution de variables non retenues par la suite mais pouvant intéresser l'équipe marketing :

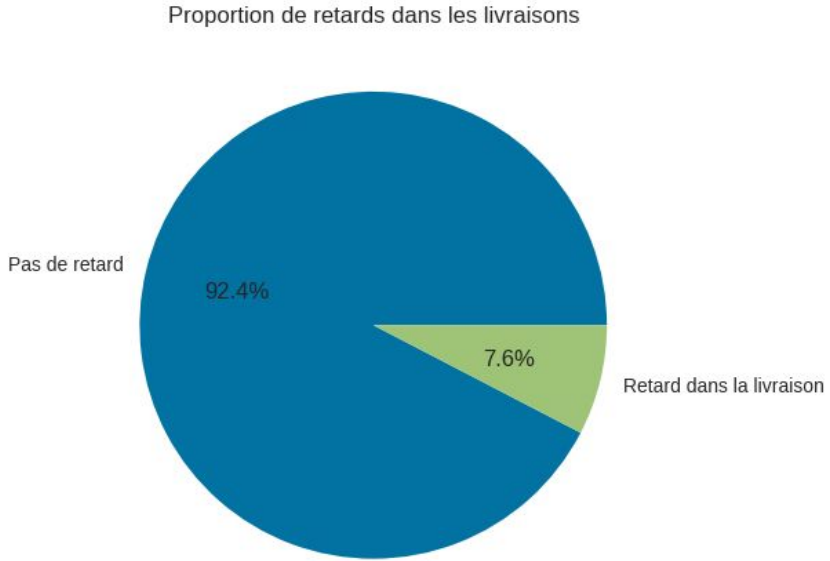




1. Présentation du jeu de données et analyse exploratoire

1.2. Analyse exploratoire

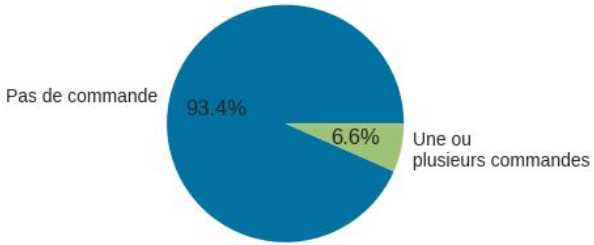




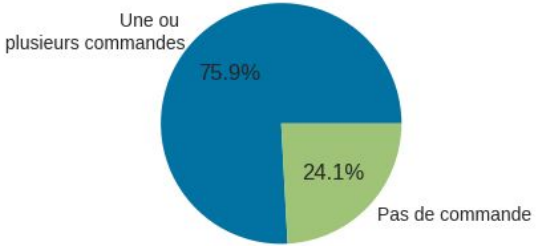
1. Présentation du jeu de données et analyse exploratoire

1.2. Analyse exploratoire

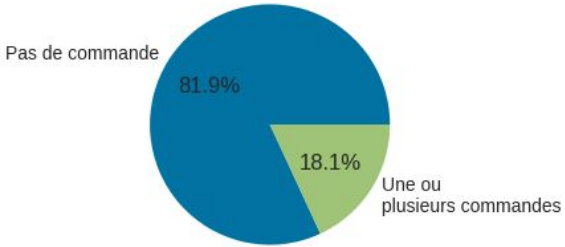
Proportions des clients ayant commandé ou non lors du dernier mois



Proportions des clients ayant commandé ou non lors de la dernière année



Proportions des clients ayant commandé ou non lors du dernier trimestre

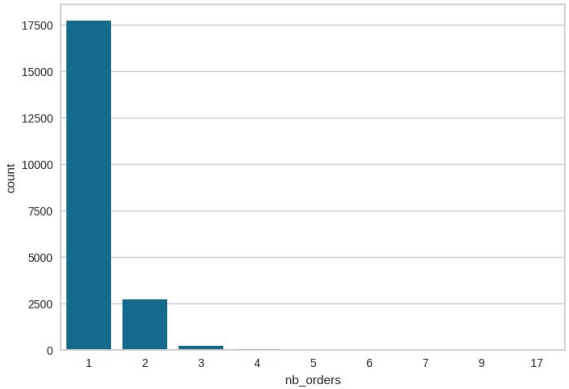


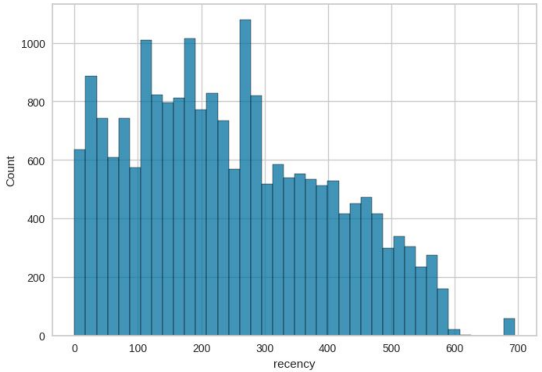
Beaucoup de clients ont commandé lors de la dernière année (comprise dans le jeu de données).

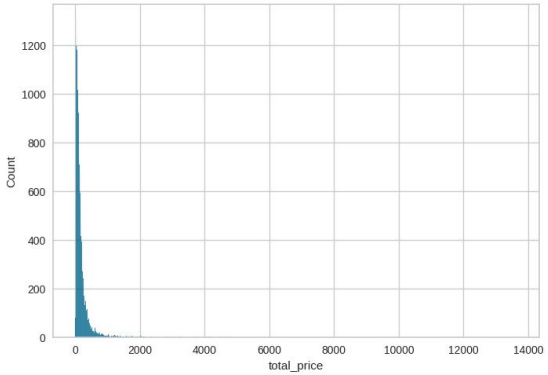
Assez peu lors du dernier trimestre et encore moins lors du dernier mois.

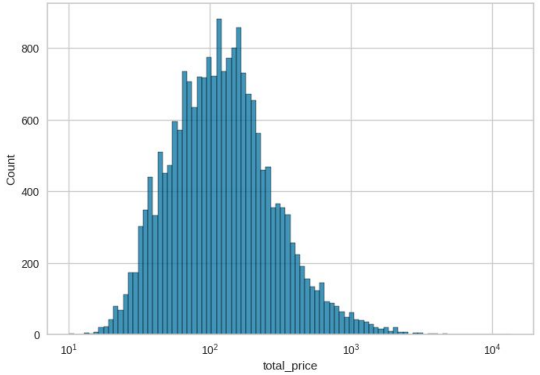
1. Présentation du jeu de données et analyse exploratoire

1.2. Analyse exploratoire









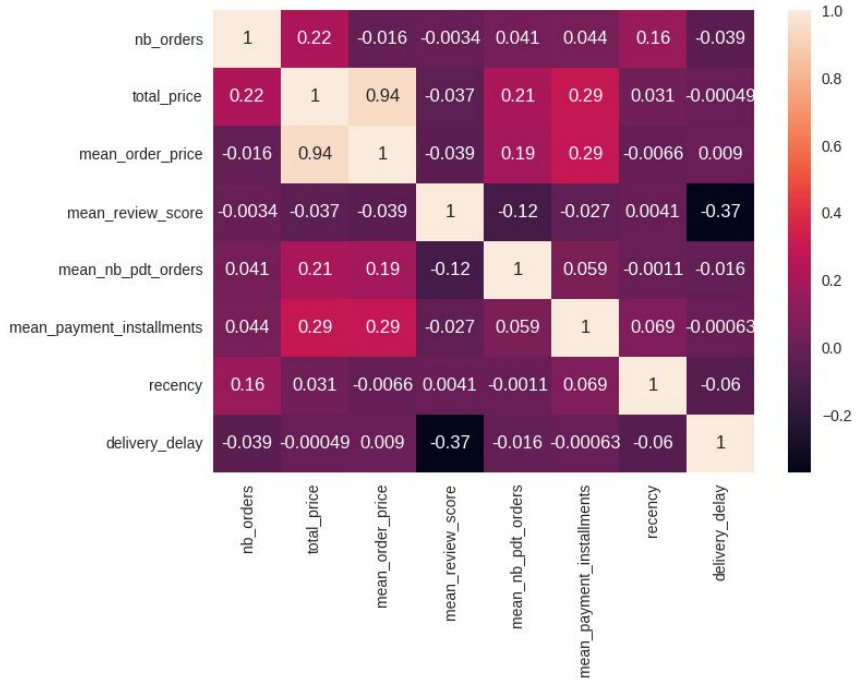
Distribution des variables RFM -> indications sur futures transformations : passage au log et / ou réduction d'échelle.

1. Présentation du jeu de données et analyse exploratoire

1.2. Analyse exploratoire

- corrélations positives entre le montant total dépensé et le faire d'avoir recours aux paiements échelonnées.

- corrélation négative entre la satisfaction et le fait d'avoir subi un retard de livraison



2. Méthodologie et résultats des différents modèles de clustering

2.1. Les différents modèles : KMeans, Clustering Hiérarchique et DBCAN

Modèles	Principes	Avantages	Inconvénients
KMeans	1) Initialisation aléatoires de centroïdes 2) Attribution de chaque point au cluster correspondant au centroïde le plus proche 3) Re-calcul des centroïdes -> itération 2) et 3) jusqu'à convergence ou atteinte du max d'itération	Rapidité d'exécution Recherche efficace d'une partition de données dont la variance intracluster est minimale	Fixation au préalable du nombre de clusters Non déterministe, dépendant des initialisation des centroïdes.
Clustering hiérarchique (distance Ward)	1) Au départ chaque point est considéré comme un cluster 2) Les points sont agglomérés en fonction d'un critère de distance prédéfini 3) On aboutit à un seul cluster regroupant tous les points	Visualisation par dendrogramme Nombre de clusters non défini à l'avance, permet d'explorer différentes possibilités	Gourmand en temps de calcul et en mémoire
DBSCAN	1) On définit un nombre minimal de points (min_samples) et un paramètre epsilon mesurant la distance entre les points 2) L'algorithme visite alors tous les points et les classes en 3 catégories: points "principaux", points "frontière" (pas suffisamment de voisin pour être "principal mais dans le voisinnage de l'un d'eux) et "bruits"	Peut identifier des clusters de formes arbitraires, et notamment non convexes. Efficace en temps de calcul Nombre de clusters non défini à l'avance	Choix délicat des paramètres epsilon et "min_samples" Difficile à utiliser en très grande dimension Ne peut pas trouver des clusters de densité différente

2. Méthodologie et résultats des différents modèles de clustering

2.2. Evaluation des performances

Indicateurs théoriques :

	Principes	Intervalle de valeurs	Signification
Silhouette Score	Compare la distance intra-clustering à la distance interclustering	Entre -1 et 1	Plus le score est proche de 1 et meilleur est clustering Un score négatif indique l'attribution d'un point au mauvais cluster, il doit s'en trouver un plus "proche".
Indice de Calinski-Harabasz	Établit un ratio entre la variance intercluster et la variance intracluster.	prend des valeurs positives	Plus l'indicateur est élevé et plus les clusters sont homogènes et bien séparés.
Indice de Davies-Bouldin	Mesure d'homogénéité/séparabilité	prend des valeurs positives	Plus la valeur est faible et meilleur est le clustering

En pratique : Statistiques et utilisations de boxplots / clusters, visualisation des clusters (PCA et T-SNE)

2. Méthodologie et résultats des différents modèles de clustering

2.3. Itérations, base RFM puis ajout et suppression de variables

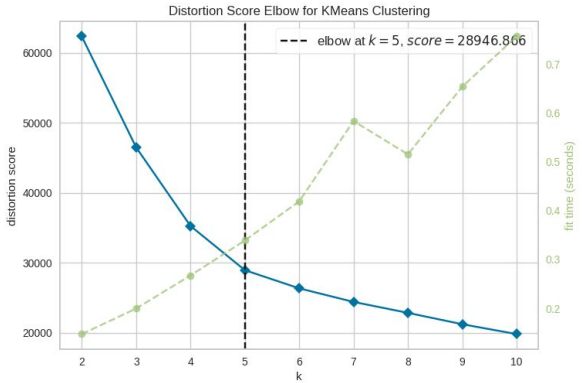
- Processus itératif, d'abord 3 variables : Nombre de commandes / Montant total dépensé / Récence
- Tentatives avec :
  - Satisfaction moyenne ;
  - Échelonnement des paiements ou non ;
  - Retard dans la livraison ou non ;
  - Moyenne du nombre de paiements ;
  - Moyenne du nombre de produits par commande ;
  - Catégorie de produits principale ;

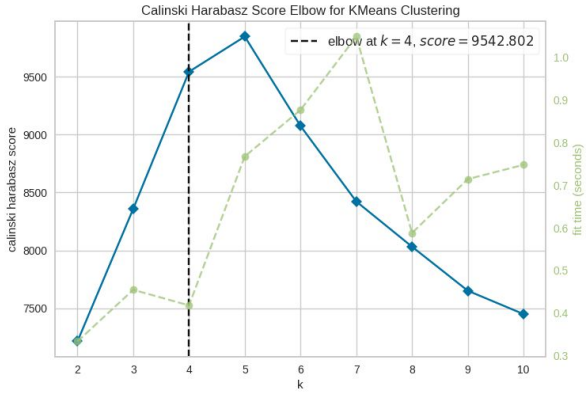
- **Variables retenues** : RFM + Satisfaction moyenne + Retard dans la livraison ou non.
- Le modèle DBSCAN n'a jamais donné de résultats satisfaisants.

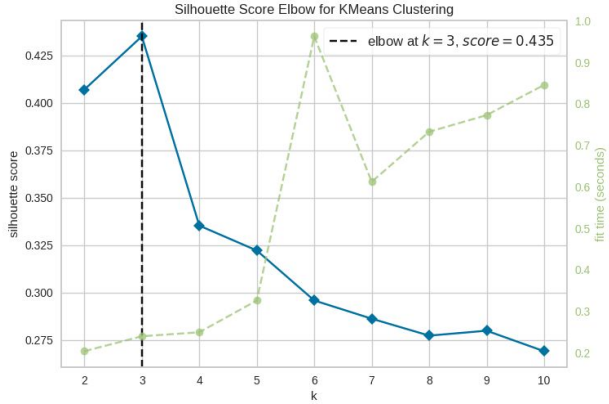
2. Méthodologie et résultats des différents modèles de clustering

2.4. Meilleurs modèles

KMeans :





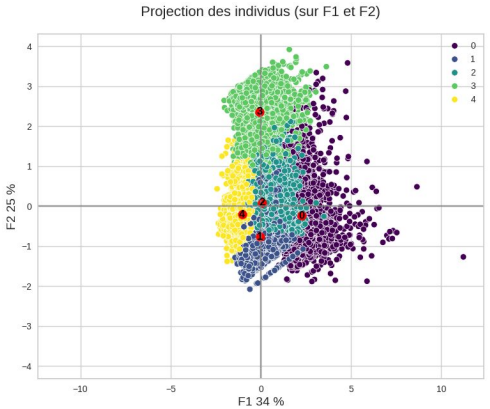


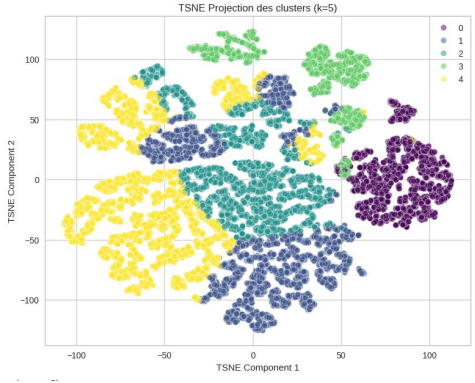
D'un point de vue théorique on voit qu'on aboutit à un nombre de clusters allant de 3 à 5. Une segmentation en 7 groupe a également été testée pour en voir l'intérêt d'un point de vue métier.

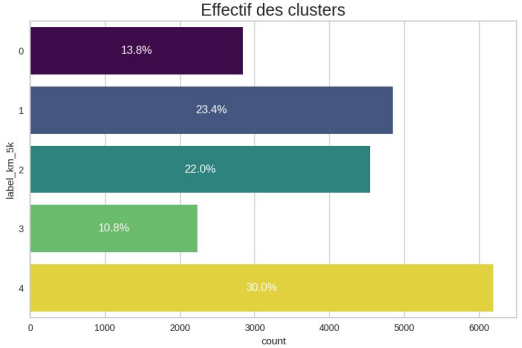
2. Méthodologie et résultats des différents modèles de clustering

2.4. Meilleurs modèles

KMeans - 5 clusters :



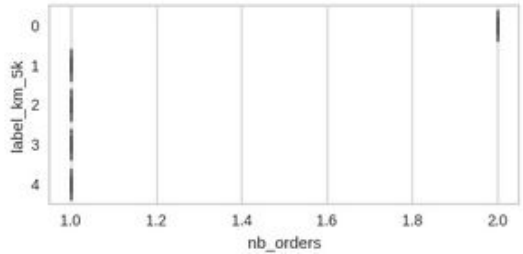


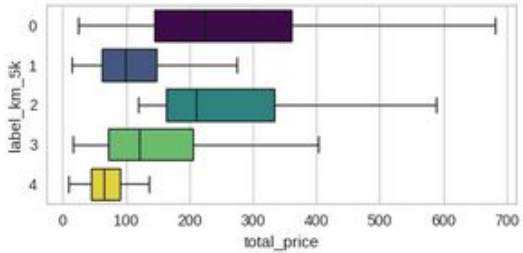


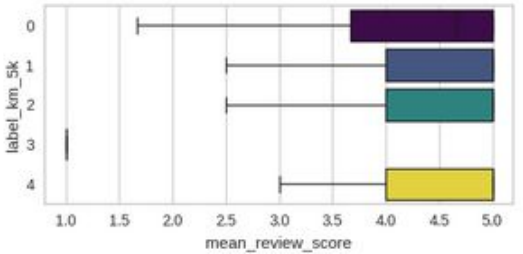
2. Méthodologie et résultats des différents modèles de clustering

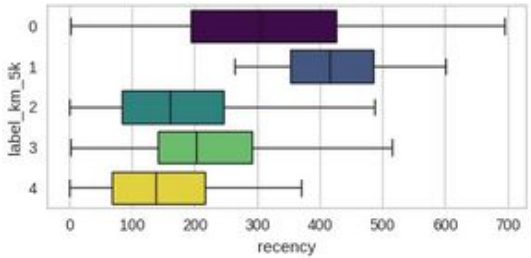
2.4. Meilleurs modèles

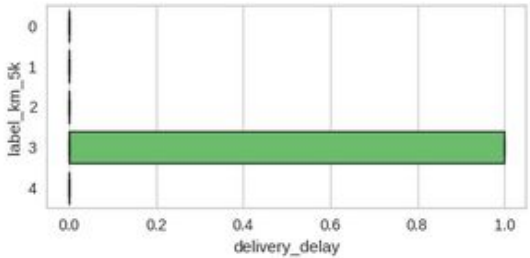
KMeans - 5 clusters :









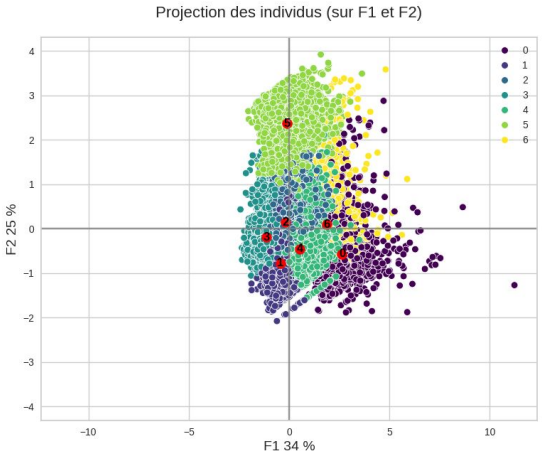


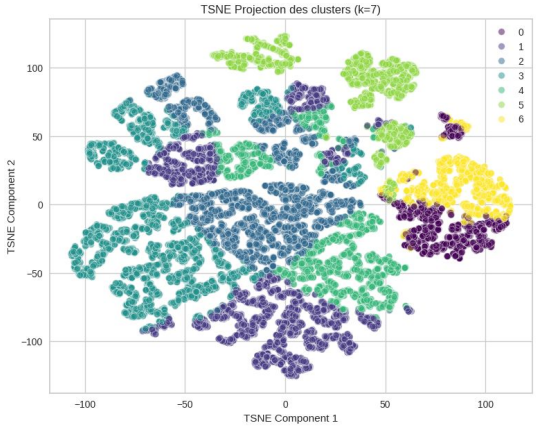
Meilleurs clients (13.8 %) :groupe 0
Mauvais clients / Perdus (23.7 %) : groupe 1
Insatisfaits (10.8%) : groupe 3
Nouveaux clients plutôt dépensiers (21.9 %): groupe 2
“Loyalistes potentiels” - peu dépensiers (29.8 %) : groupe 4

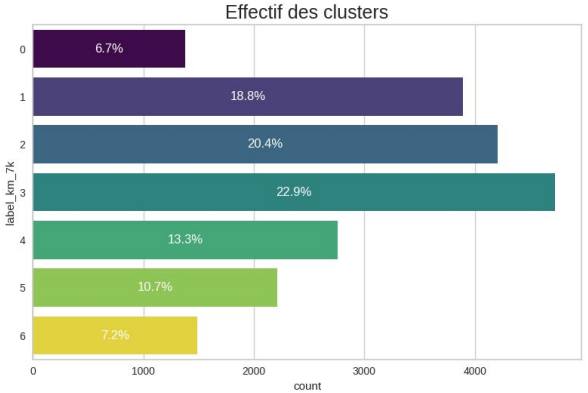
2. Méthodologie et résultats des différents modèles de clustering

2.4. Meilleurs modèles

KMeans - 7 clusters :



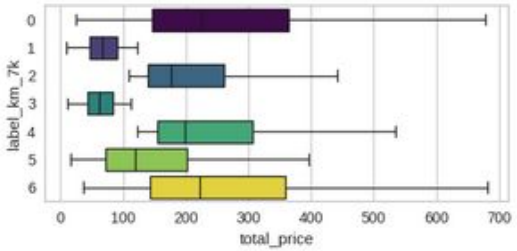


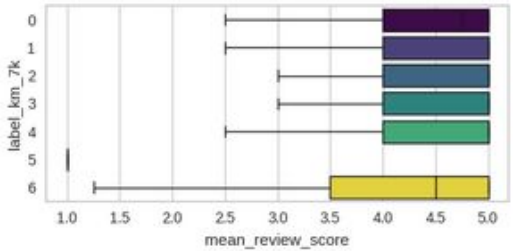


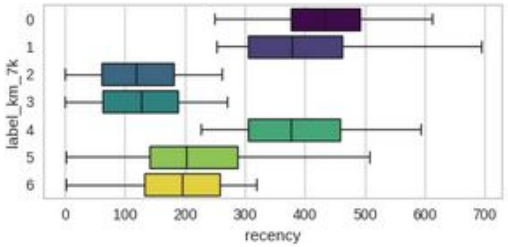
2. Méthodologie et résultats des différents modèles de clustering

2.4. Meilleurs modèles

KMeans - 7 clusters :





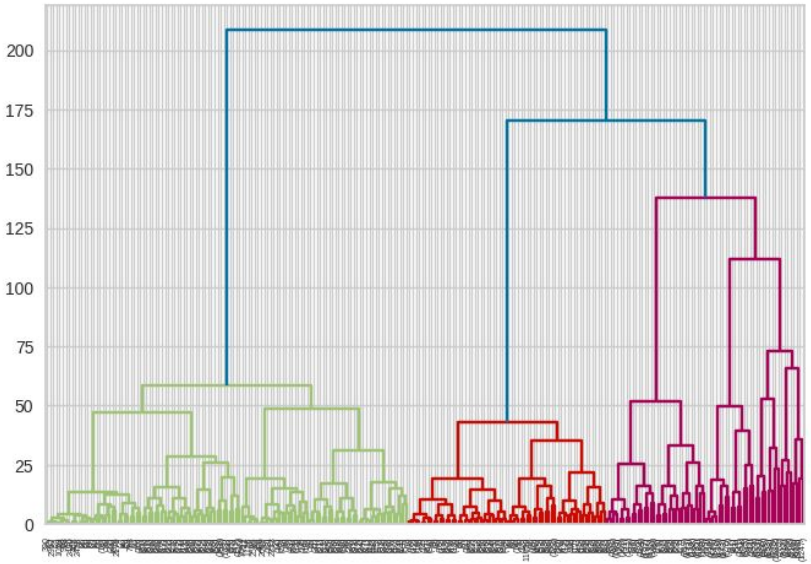


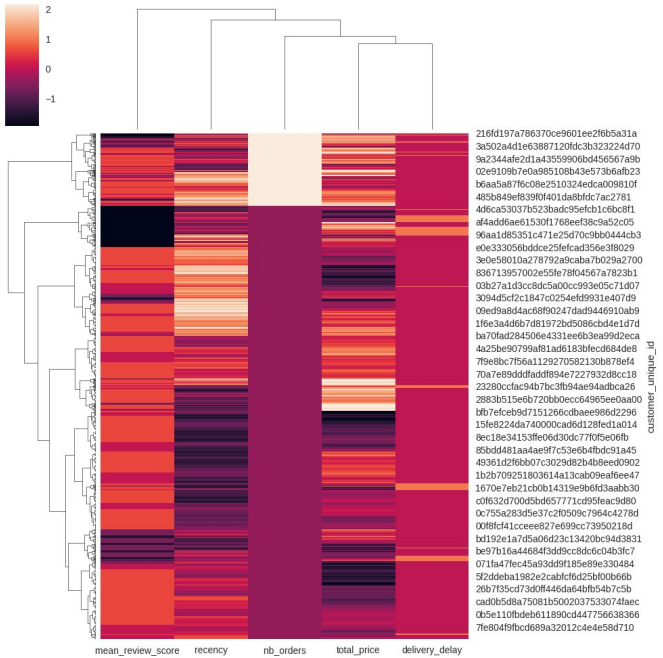
Meilleurs clients (7.8 %) : groupe 6
Bons clients mais à risque (6.7 %) : groupe 0
Perdus (18.8 % %) : groupe 1
A réactiver (13.3 %) ? : groupe 4
Insatisfaits (10.7%) : groupe 5
Loyaux (20.4 %) : groupe 2
Clients loyaux mais peu dépensiers (22.9 % %) : groupe 3

2. Méthodologie et résultats des différents modèles de clustering

2.4. Meilleurs modèles

Clustering Hiérarchique Ward :



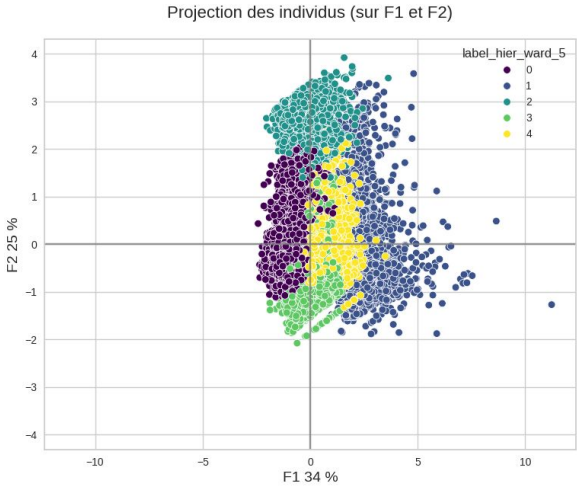


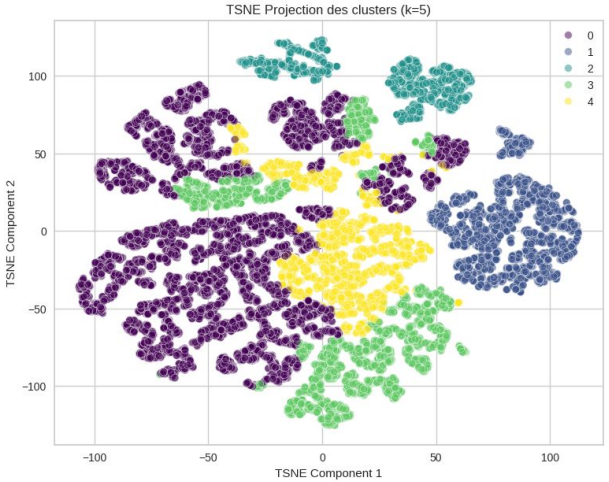
Les visualisations suggèrent 4 ou 5 clusters, les résultats les plus probants étant pour 5.

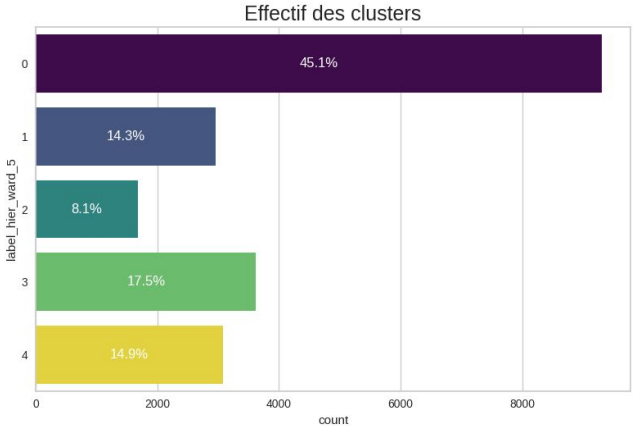
2. Méthodologie et résultats des différents modèles de clustering

2.4. Meilleurs modèles

Clustering Hiérarchique Ward - 5 clusters :



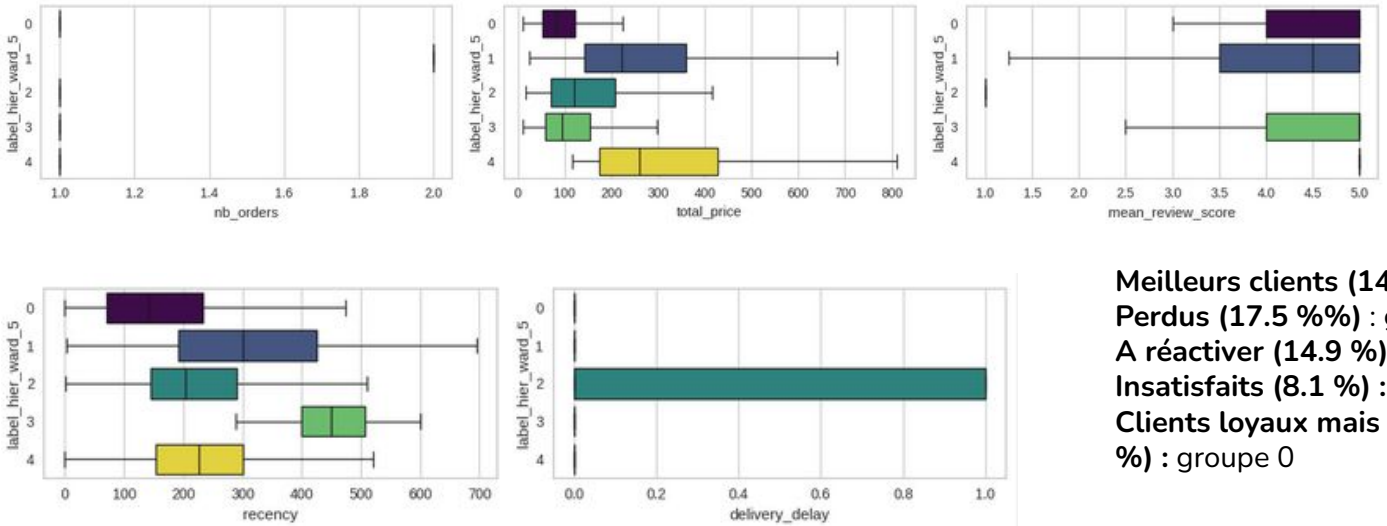




2. Méthodologie et résultats des différents modèles de clustering

2.4. Meilleurs modèles

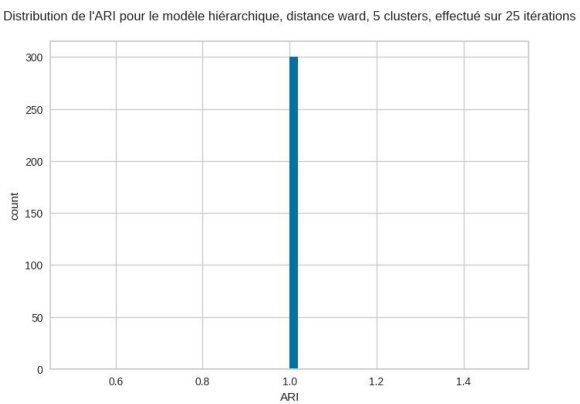
Clustering Hiérarchique Ward - 5 clusters :

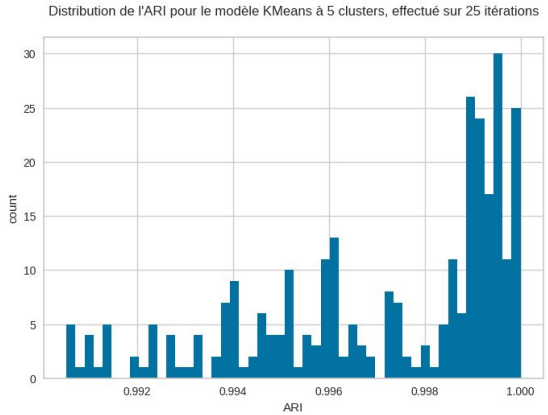


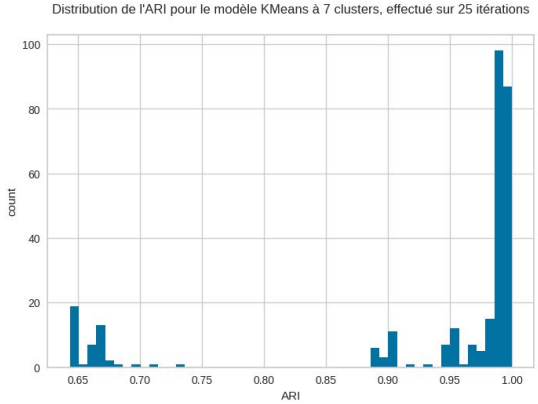
Meilleurs clients (14.3 %) : groupe 1
Perdus (17.5 %%) : groupe 3
A réactiver (14.9 %) ? : groupe 4
Insatisfaits (8.1 %) : groupe 2
Clients loyaux mais peu dépensiers (45.1 %%) : groupe 0

2. Méthodologie et résultats des différents modèles de clustering

2.5. Stabilité



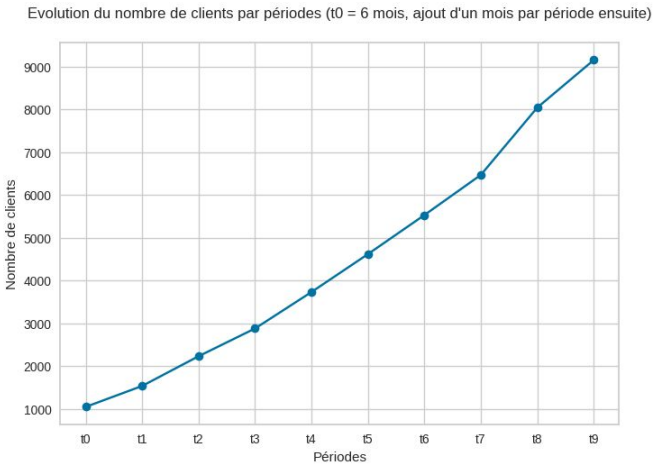


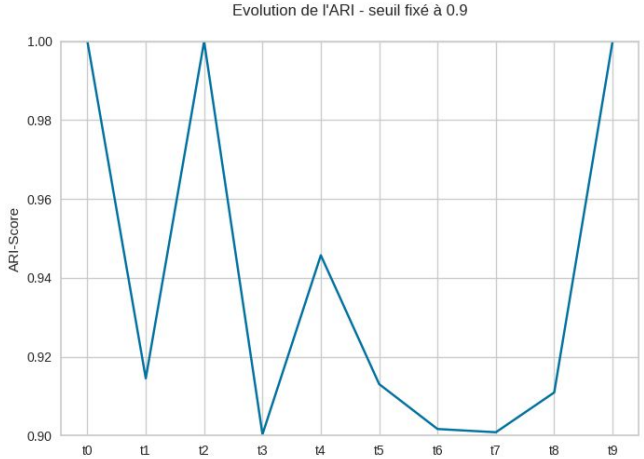


Utilisation de l'**Adjusted Rand Index** : métrique qui évalue la **proportion de paires de points qui sont groupées de la même façon** lors de la comparaison entre 2 itérations d'un modèle de clustering.
Le **KMeans à 7 clusters** donnait une segmentation intéressante mais est **trop instable**.
Le **modèle hiérarchique** est par construction **très stable** mais il offre une segmentation moins intéressante que le KMeans à 5 clusters.
Nous conserverons donc le modèle KMeans à 5 clusters. Sa **stabilité** est **satisfaisante** et la **segmentation** proposée nous paraît **pertinente**.

3. Maintenance du modèle KMeans à 5 clusters

Idée : Création d'un Dataframe comprenant les 6 premiers mois du dataset, puis d'autres Dataframes ajoutant des périodes d'un mois (t0 = 6 mois , t1 = 7 mois, t2 = 8 mois, etc.)
-> Fixer un seuil d'ARI score minimal, calculer l'ARI score entre un modèle entraîné sur les 6 premiers mois et un modèle entraîné à la période 't', tant que le résultat est supérieur au seuil on conserve le modèle, sinon on le met à jour.

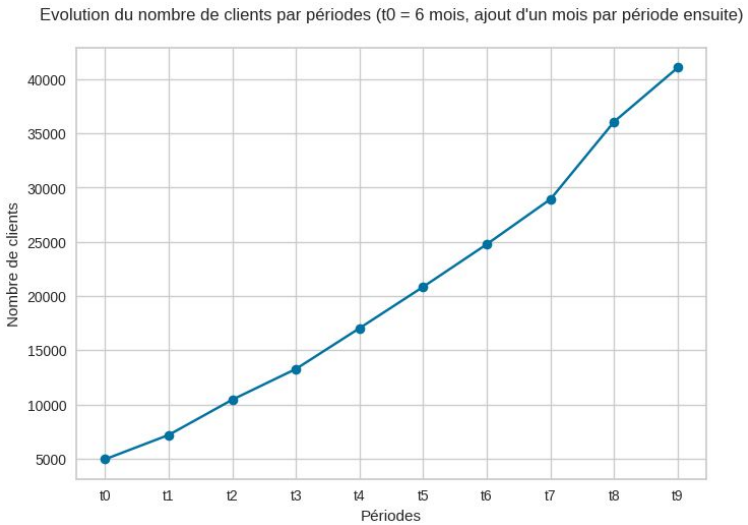


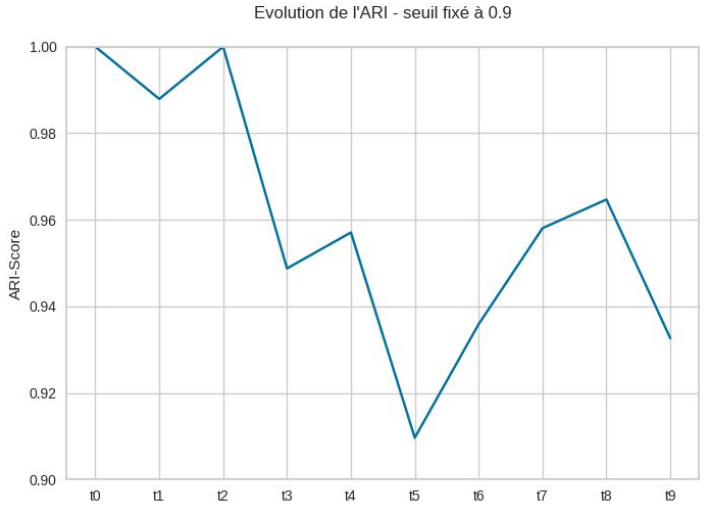


-> sur le jeu de données échantillon
-> Avec un seuil fixé à 0.9, une mise à jour tous les mois ou les 2 mois peut s'envisager.

3. Maintenance du modèle KMeans à 5 clusters

Sur le dataset complet cela donne :





-> Avec un seuil fixé à 0.9, une mise à jour tous les mois ou les 2 mois peut également s'envisager.

Conclusion

- Constitution, via SQL, d’un jeu de données pertinent ;
- Application d’un modèle de clustering proposant une segmentation facilement exploitable : groupes de clients bien distincts ;
- Segmentation qui affine la segmentation RFM en y ajoutant un indicateur de satisfaction des clients, et d’un éventuel retard lors de la livraison ;
- Fréquence de mise à jour de la segmentation : mensuelle.

Merci de votre attention.