Projet 5 OC :

“Segmentez des clients d'un site e-commerce”

- Problématique
- EDA
- Classification non supervisée
- Conclusion

Problématique

Mission de consultant pour Olist, solution de vente sur les marketplaces en ligne

Objectifs

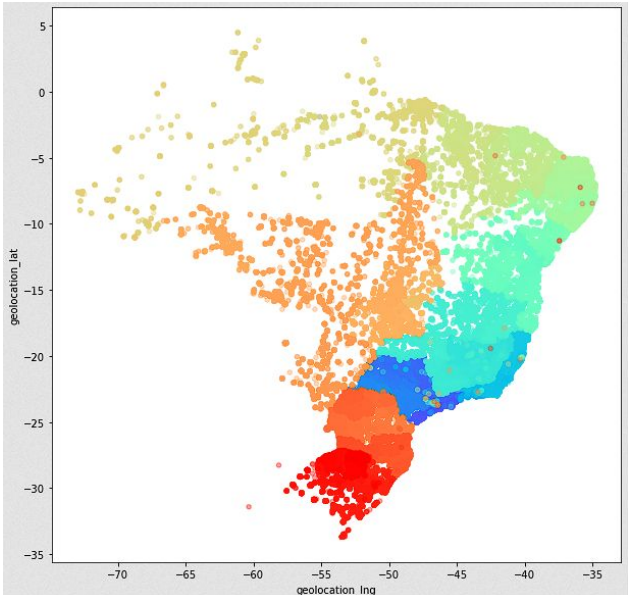
- Fournir aux équipes d'e-commerce une segmentation des clients pour les campagnes de communication
- Comprendre les différents types d'utilisateurs
- Fournir une description actionable de la segmentation
- Faire une proposition de contrat de maintenance

Olist : une plateforme d'e-commerce

- Plateforme d'e-commerce
- Créée au Brésil en 2016
- Propose une vitrine en ligne pour les vendeurs
- Met en lien acheteurs et vendeurs
- Suivi du paiement et de la livraison
- Notation des vendeurs et des produits

Données du dataset brute

- Données réparties en 9 fichiers .csv de 120 Mo, globalement bien complétées.
- des données variées (textuelles, chiffrées, catégorielles, géographiques)
- 96096 clients unique, 99441 commandes distinctes, 118310 produits vendus
- Enregistrements sur 2 ans, 38 variables, environs 120 000 lignes

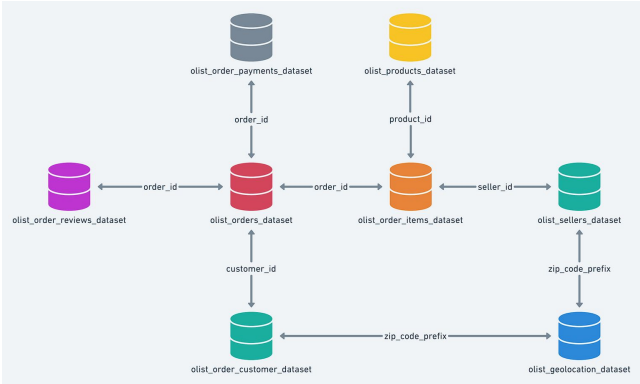


Feuille de route

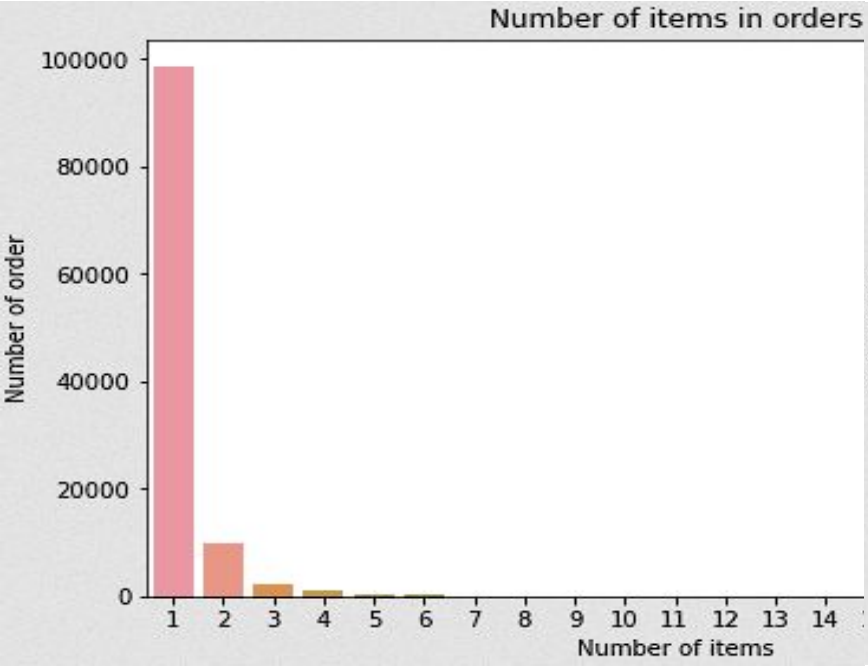
- Extraire les données de la base de donnée Olist permettant de caractériser les clients
- Utiliser des outils de machine learning non supervisés pour réaliser un partitionnement des clients en fonction de ces caractéristiques
- Interpréter les segments obtenus d'un point de vue métier
- Analyser la stabilité temporelle du processus de partitionnement pour évaluer une fréquence de maintenance

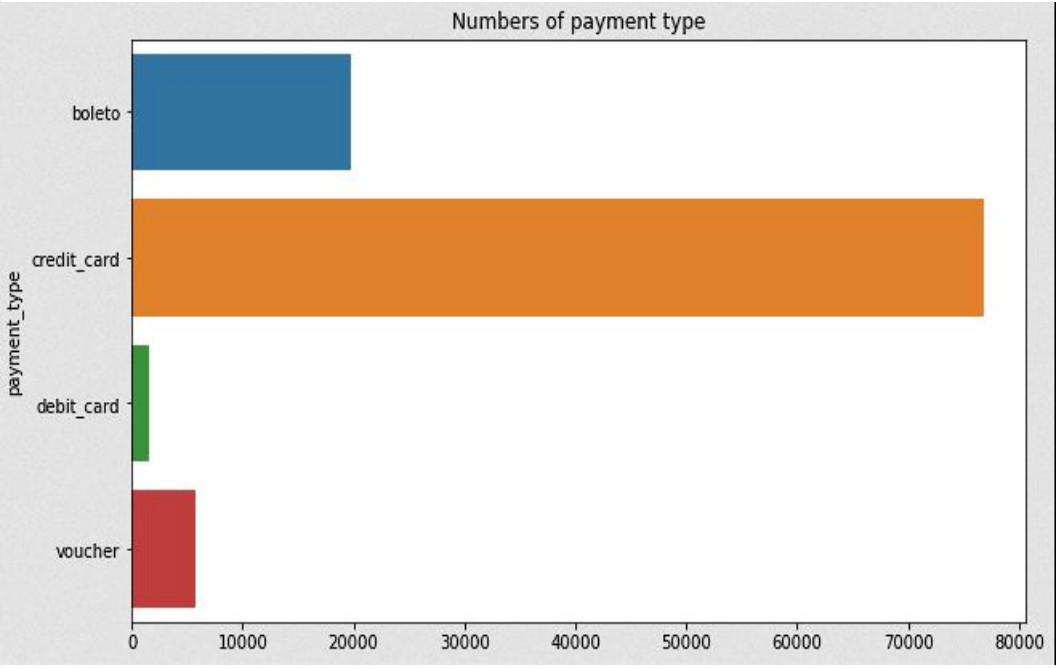
Nettoyage et merge du dataset

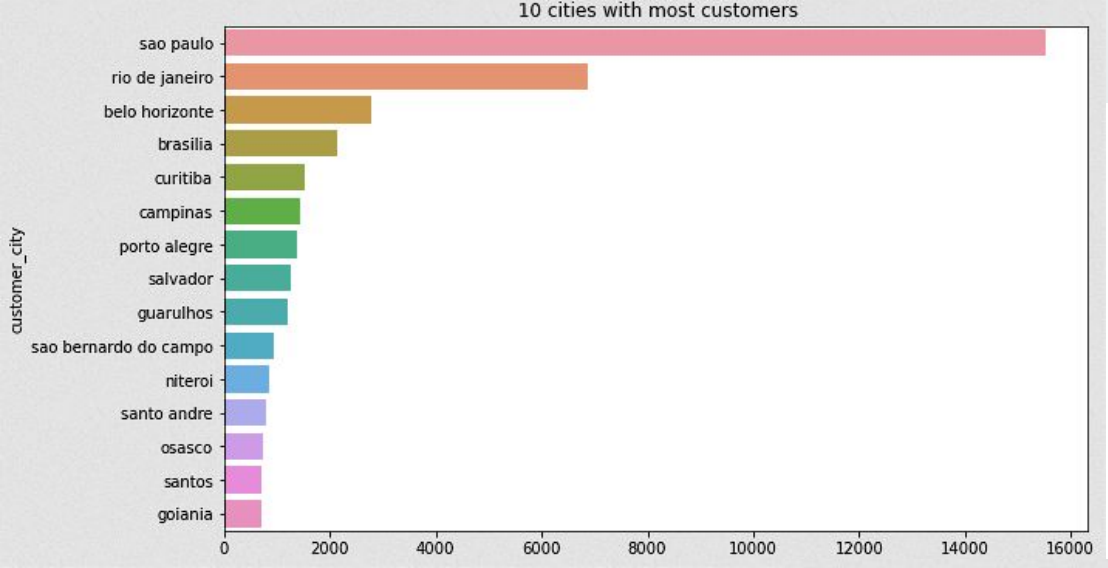
- Suppression des lignes 100% “NaN” et dupliqué
- Suppression des colonnes avec 100% et >80% de valeurs manquantes
- Imputation des informations manquantes par médiane (entre 0,1 % et 3% NaN)
- Transtypage de données en “datetime” et “category”
- Suppression de features non pertinentes “review_comment_message" , “customer_zip_code_prefix”
- Création de nouvelles features “volume“, “estimated_delivery_time”, “delivery_time” et RFM
- Transformation log de “F”, “M”, "price", "freight", “delivery_time” (car biaisées)
- Assemblage dans une table unique avec pour index l’id client unique

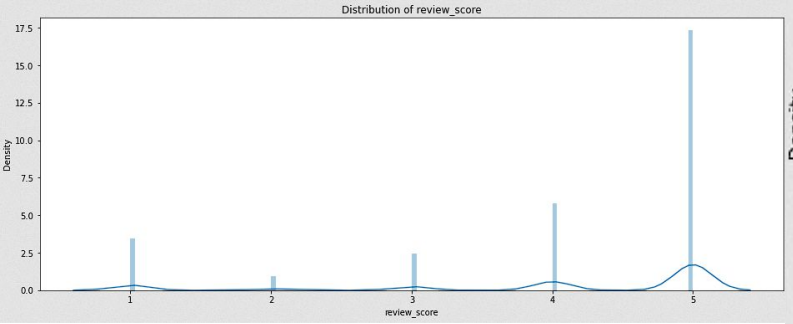


EDA


Number of items in orders
Number of order
100000
80000
60000
40000
20000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Number of items

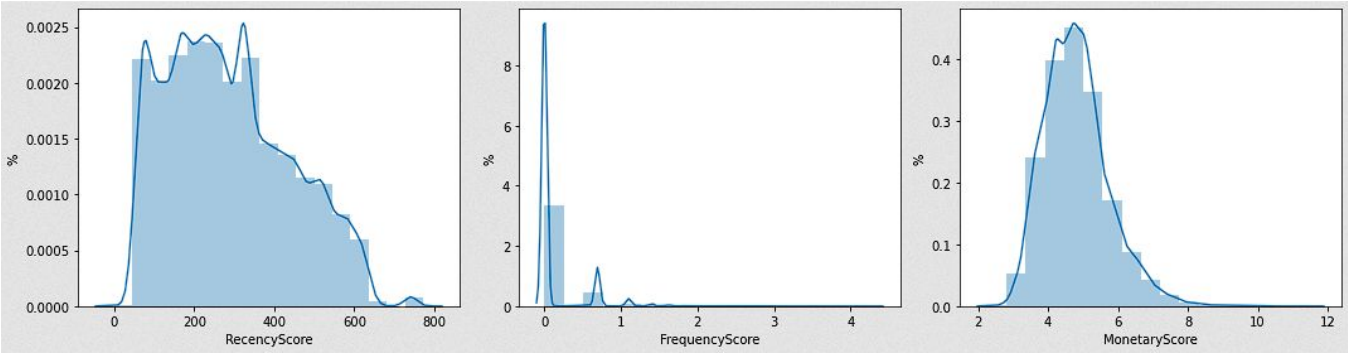

Numbers of payment type
payment_type
boleto
credit_card
debit_card
voucher
0
10000
20000
30000
40000
50000
60000
70000
80000

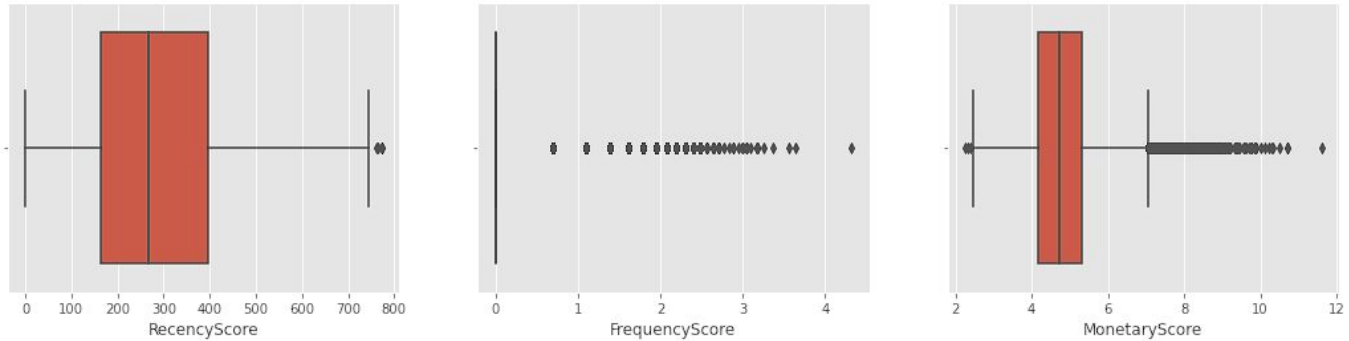

10 cities with most customers
customer_city
sao paulo
rio de janeiro
belo horizonte
brasilia
curitiba
campinas
porto alegre
salvador
guarulhos
sao bernardo do campo
niteroi
santo andre
osasco
santos
goiania
0
2000
4000
6000
8000
10000
12000
14000
16000


Distribution of review_score
Density
review_score
1
2
3
4
5


Delivery time in days
Delivery time
Estimated delivery time
Density
estimated_delivery_time
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0
20
40
60
80
100

RFM (pour établir des segments de clients homogènes et de cibler les offres)





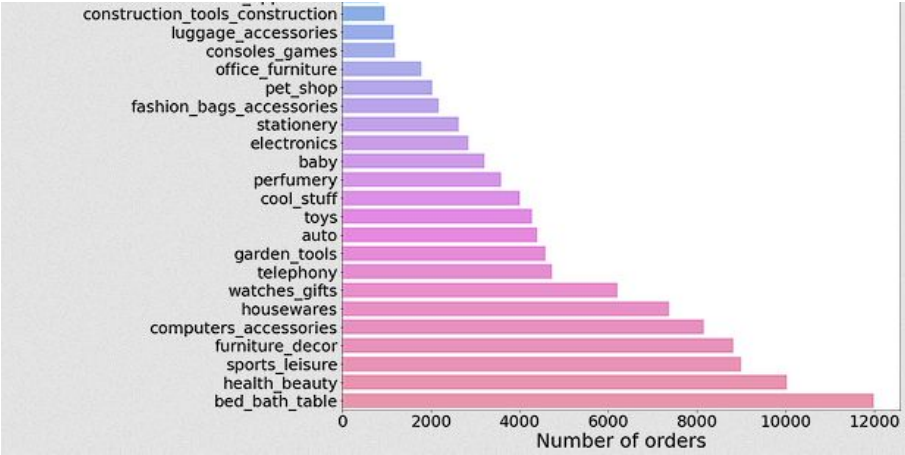
Recency — nombre de jouer depuis la dernière commande

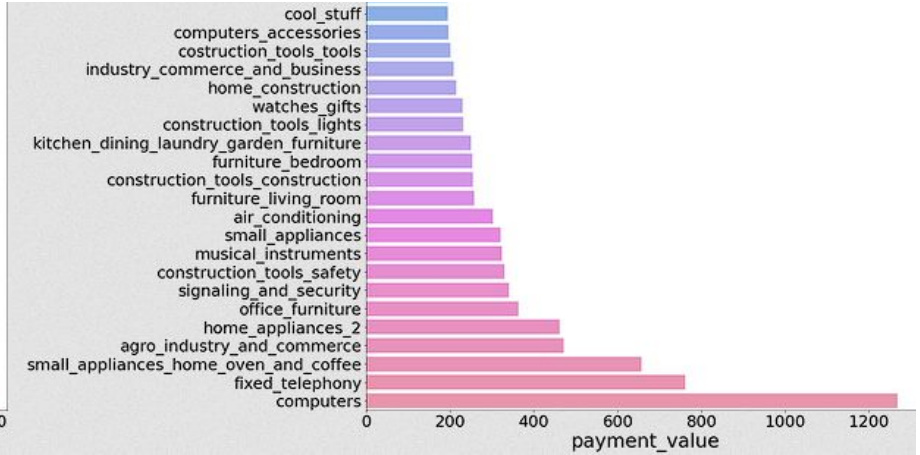
Frequency — nombre de commande sur toute la période

Monetary — total valeur des commandes sur toute la période

Conclusion exploratoire

- Un jeu de donnée de clients principalement à commande unique (97% d'article unique par commande)
- Clients globalement satisfait
- Les articles les plus vendus sont "ameublement, multimédia, équipement de sport, décoration"
- Les articles les plus chère "produits électroniques"
- Les notes de reviews sont un peu meilleurs sur les produits peu chers
- Les délais estimés sont plus longs que les délais de livraison
- Pas de forte corrélation entre les variables



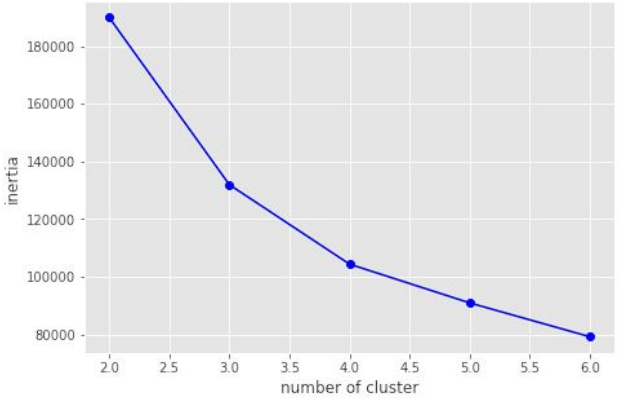


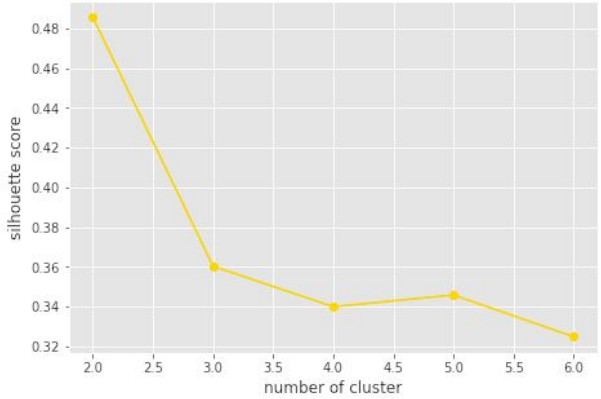
Classification non supervisée

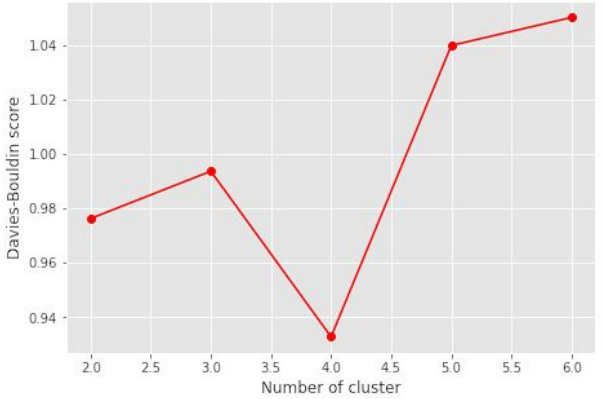
- Récupération de l'identifiant, de la date et de la valeur de chaque commande.
- Après jointure sur l'identifiant client on peut calculer les variables RFM
- Le dataset est agrégé par clients unique. Les variables catégorielles ne peuvent pas être agrégés, elle seront supprimés
- Approche retenue : le Kmeans

Clustering Kmeans et métrique

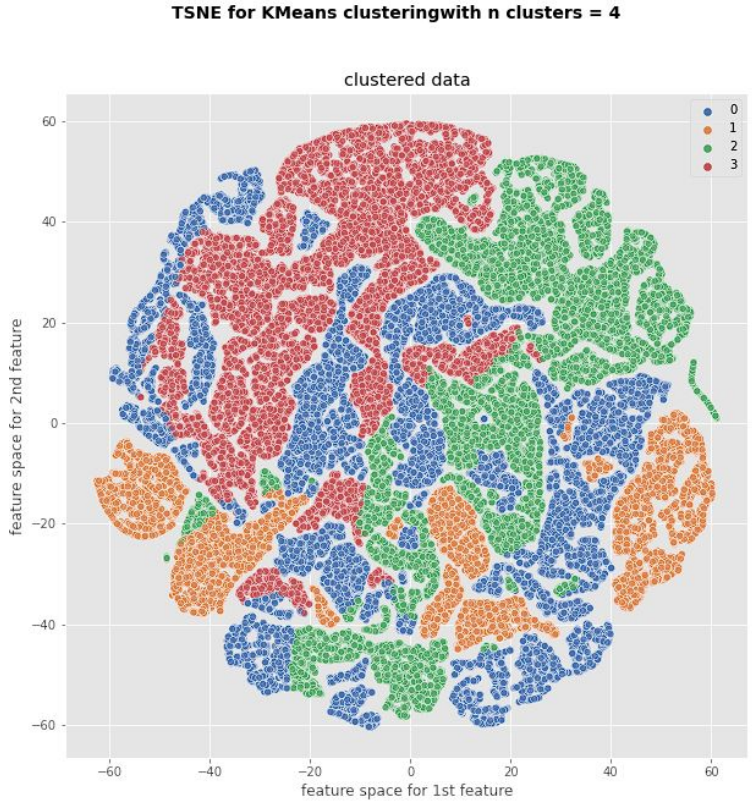
- Un pipeline : avec SimpleImputer(médiane) et normalisation avec StandardScaler()
- Itération sur un nombre borné K
- Méthode du coude (inflexion nette a K= 4)
- Score de silhouette (clusters correctement séparés)
- Score de Davies-Bouldin (score minimisé)

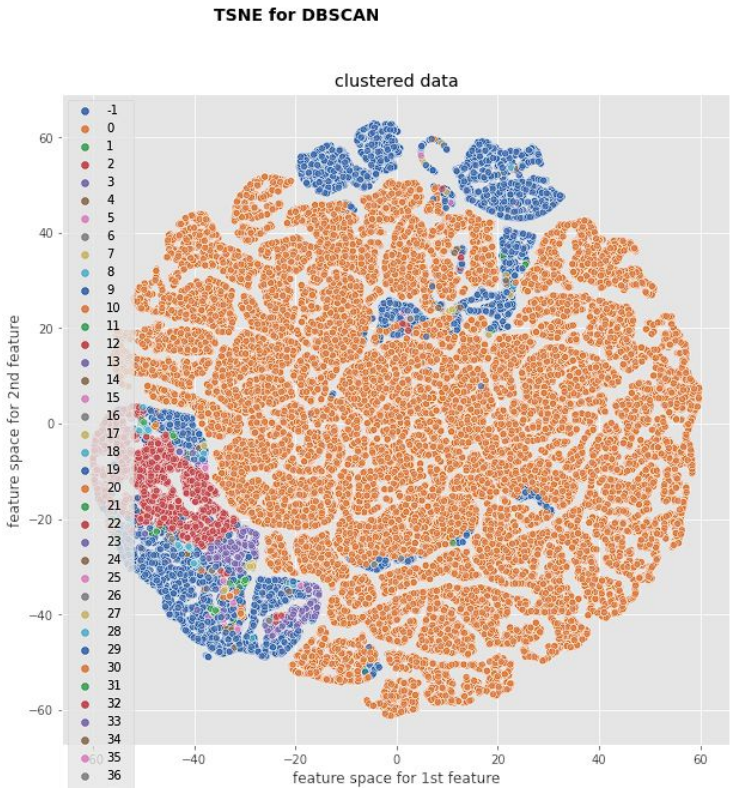




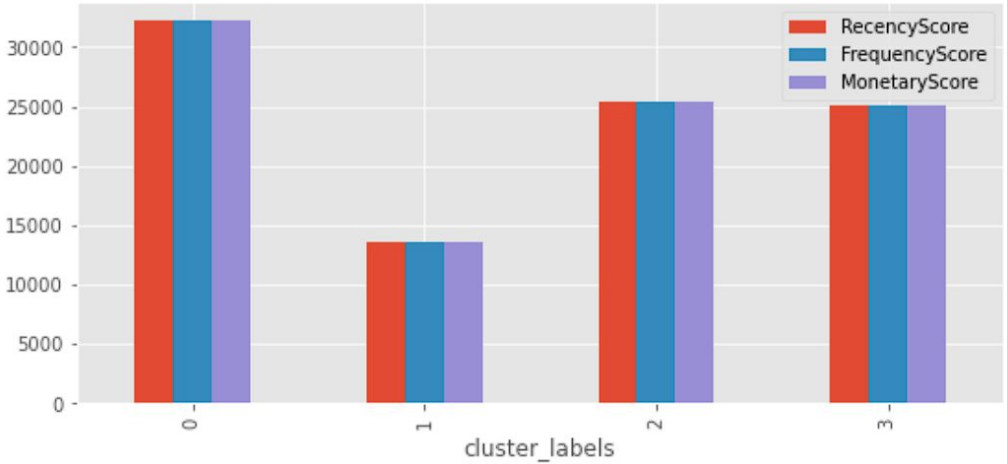


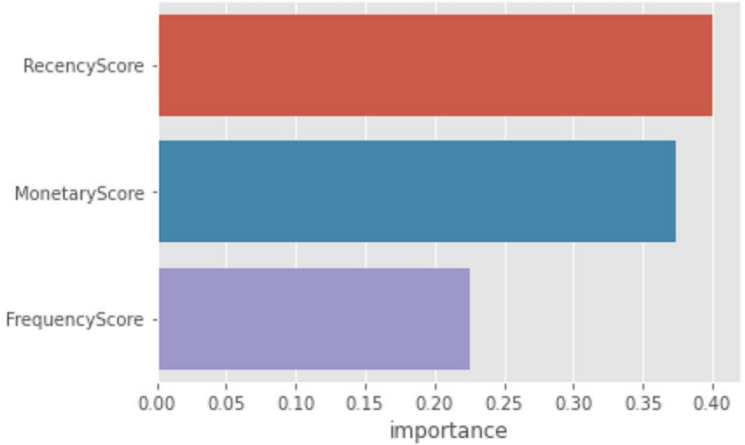
Réduction dimensionnelle (T-SNE)



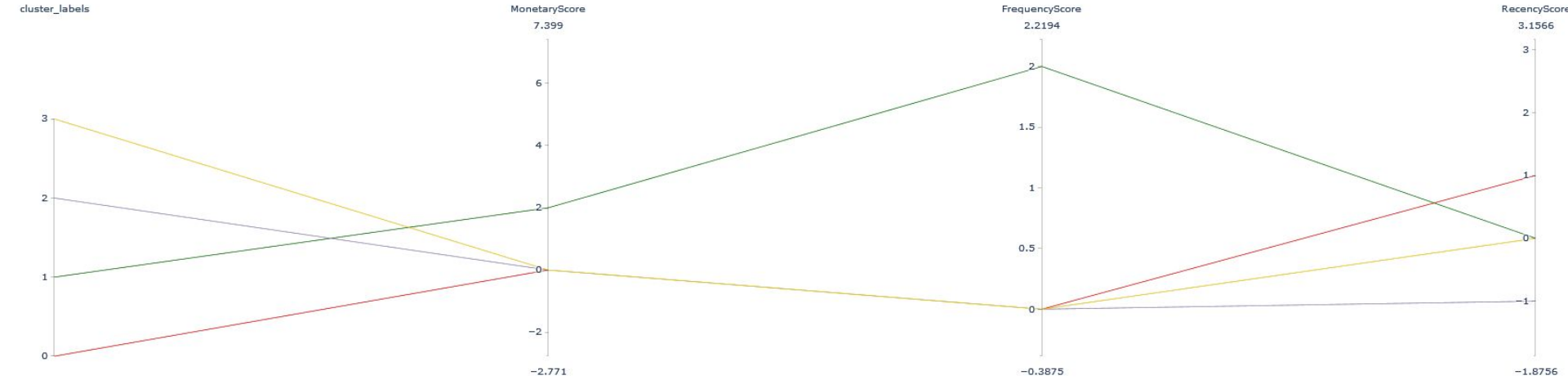


Importance des variables latentes





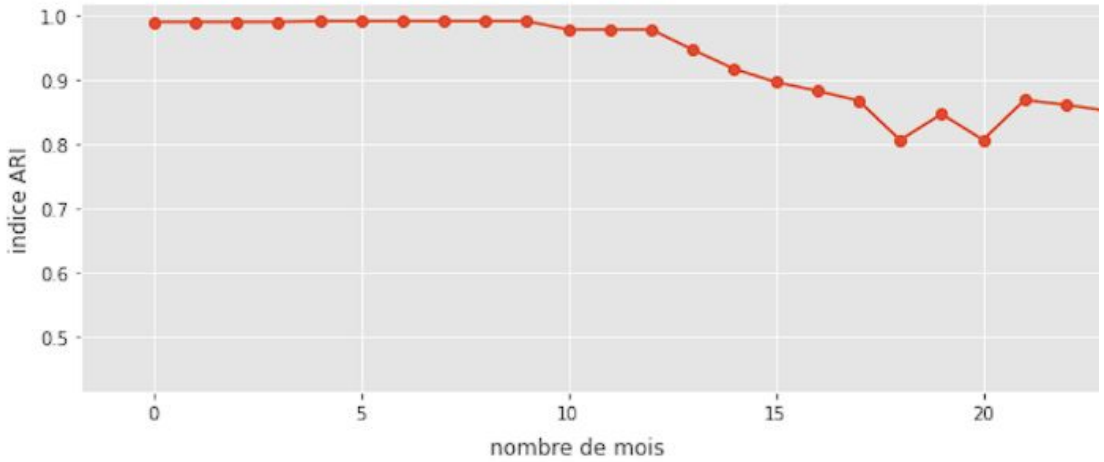
Interprétation des résultats



Cluster 1 : les commandes sont moins fréquentes, moins onéreuses mais très récentes
Cluster 2 : les commandes sont très fréquentes, plus onéreuse mais moins récentes
Cluster 3 : les commandes sont moins fréquentes, moins onéreuses et pas du tout récentes
Cluster 4 : les commandes sont moins fréquentes, moins onéreuses et moyennement récentes

Métriques ARI (refit périodique)

Pour établir un contrat de maintenance de l'algorithme de segmentation client, nous testerons sa stabilité sur 23 mois, avec un K-Means itéré tous les mois comparé au dataset global.



Suivi de la prédiction du numéro de cluster des mêmes clients d'une période à l'autre
Changement de l'attribution des clusters à partir de 9 mois

Conclusion

- Mise en application des algorithmes de classification non supervisée adaptée à notre problème métier
- Opportunités d'amélioration du clustering en ajoutant de nouvelles features ("volume", "delivery_time", etc
- Mise en relation avec les catégories de produits dominants pour chaque clusters (mais commande unique)
- Analyse des clients ayant acheté plusieurs articles
- Données plus précises sur les clients (à anonymiser) : âge, sexe
- Proposition de maintenance à 9 mois