Segmentez des clients d'un site e-commerce

Projet 4 parcours IML OpenClassrooms
juillet 2022

OPENCLASSROOMS

TABLE DES MATIÈRES

Introduction

Contexte

Problématique

Objectif

Contexte

- Toute entreprise aspire à un retour de ces clients afin d'exploiter mieux leurs comportements
- Dans le cas des sites d'e-commerces, les entreprises préconisent la segmentation pour réduire les ressources allouées et viser de façon efficace leurs clients.
- La segmentation des clients consiste à découper automatiquement un dataset en sous-ensemble homogènes
- Ce types de segments constituent la base d'exploitation des campagnes de communication des équipes de marketing,

Problématique et objectif

Mission de consultant pour Olist, une entreprise de vente e-commerce

Objectifs:

- Fournir aux équipes d'e-commerce une segmentation de l'ensemble de ses clients pour les campagnes de communication
- Comprendre les différentes caractéristiques d'utilisateurs
- Etablir une description exploitable de la segmentation
- Etablir une période de maintenance de la segmentation

Préparation et exploration

Cleaning et exploration

Feature engineering

Cleaning et exploration

Données réparties en 9 tables:
clients / géolocalisation / commandes / paiements / produits / vendeurs / catégories de produits

Principales tâches

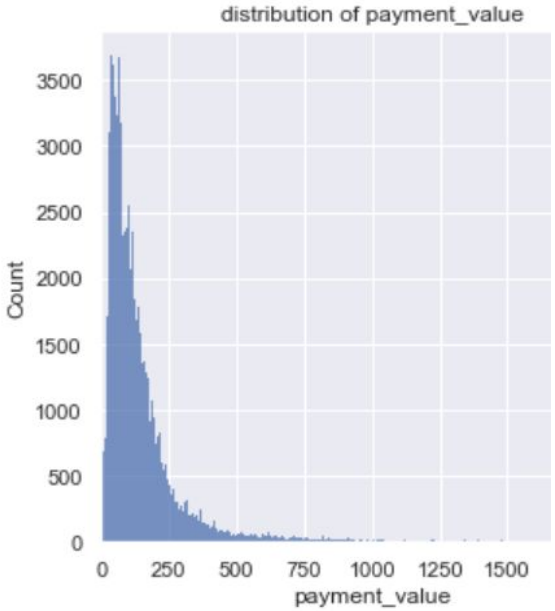
- Types de données
- Imputation des données manquantes
- suppression des colonnes non importantes
- Analyse exploratoire
- Création de nouvelles features
- Création d’un dataset de clients étendue.
- ACP réduction

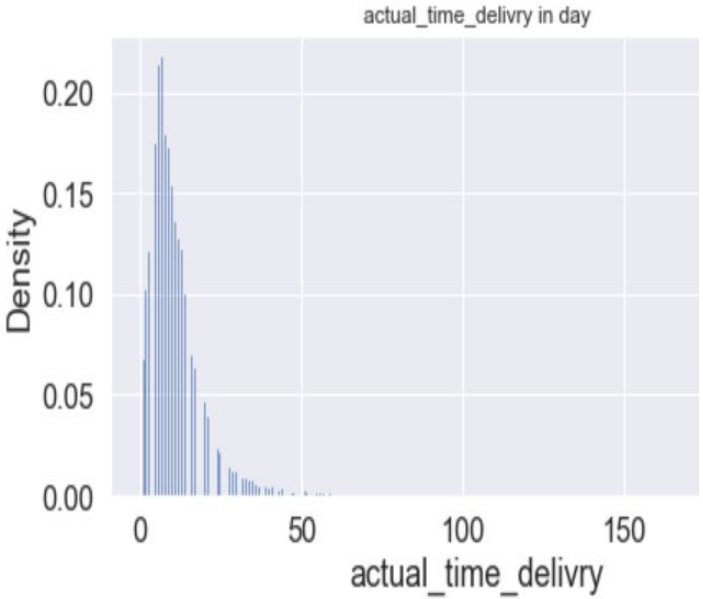
Feature engineering

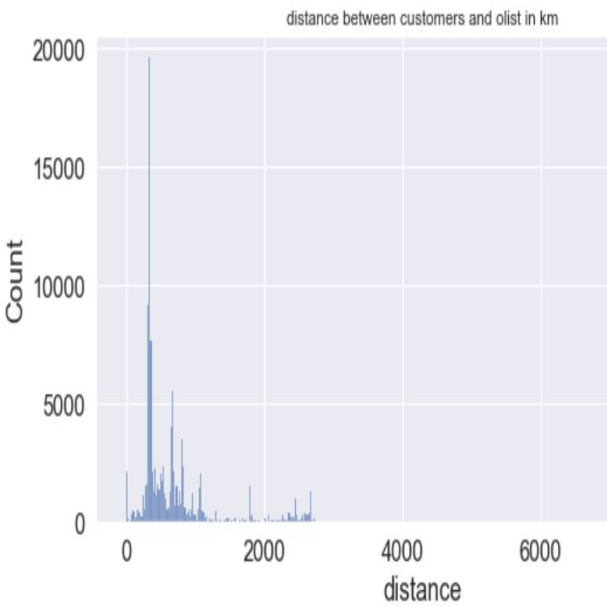
Features caractérisant le comportement des clients

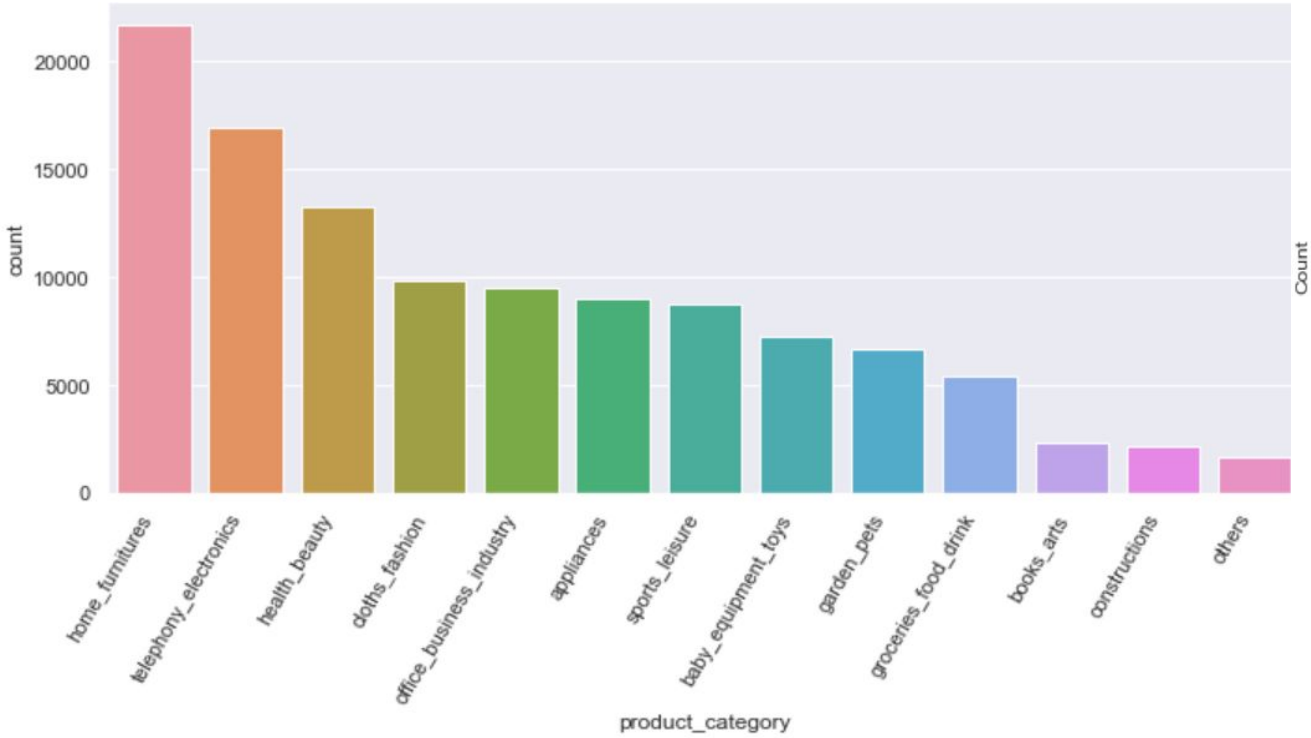
- Nombre d’achats
- Catégories des produits plus condensées
- RFM features
- périodes favorites d’achat (jour, heure, mois)
- Facilités de paiement (moyenne)
- Moyen de paiement (moyenne)
- score moyenne
- Etc

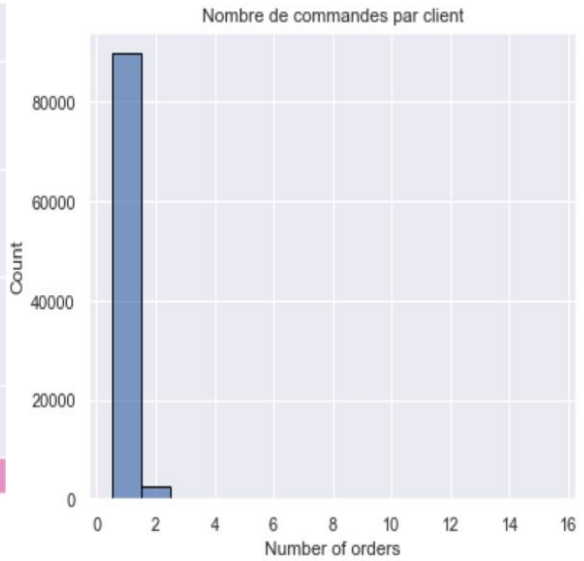
Exploration

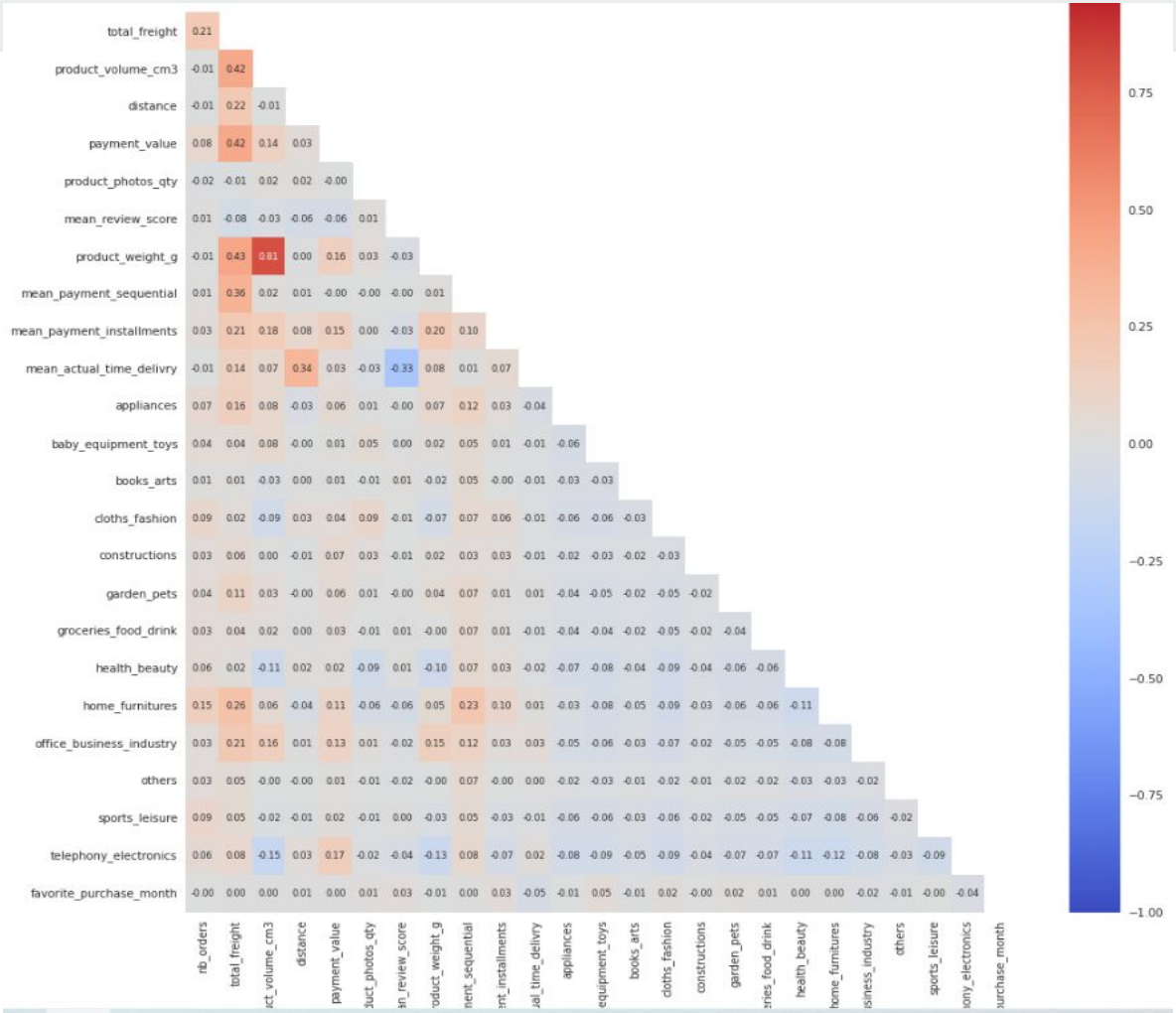







count
20000
15000
10000
5000
0
home_furnitures
telephony_electronics
health_beauty
cloths_fashion
office_business_industry
appliances
sports_leisure
baby_equipment_toys
garden_pets
groceries_food_drink
books_arts
constructions
others
product_category


Nombre de commandes par client
Count
80000
60000
40000
20000
0
0
2
4
6
8
10
12
14
16
Number of orders



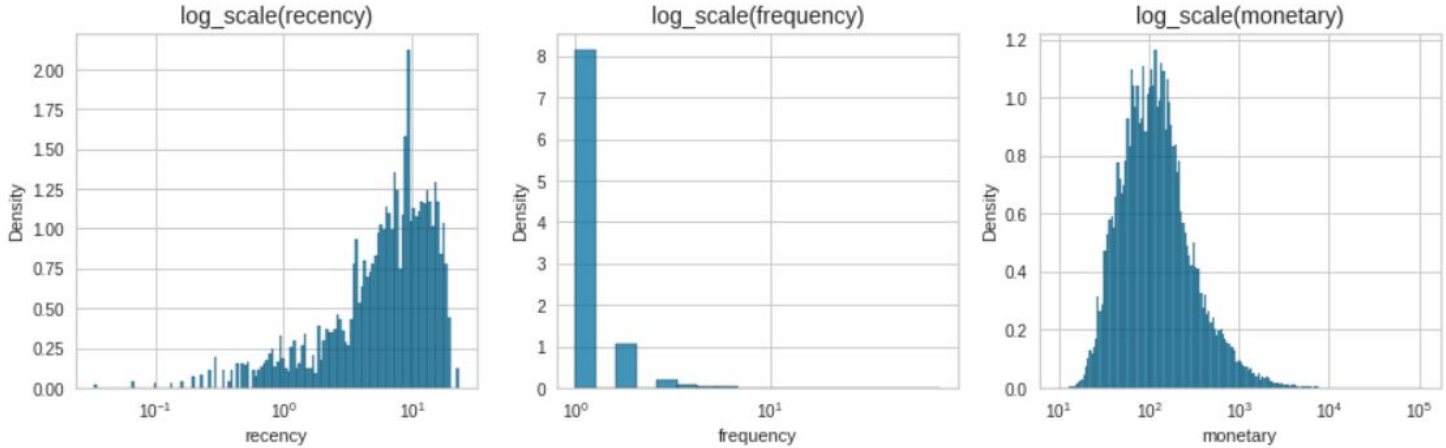
Absence de corrélation linéaire poids et volume des produits significativement corrélé

Segmentation RFM features

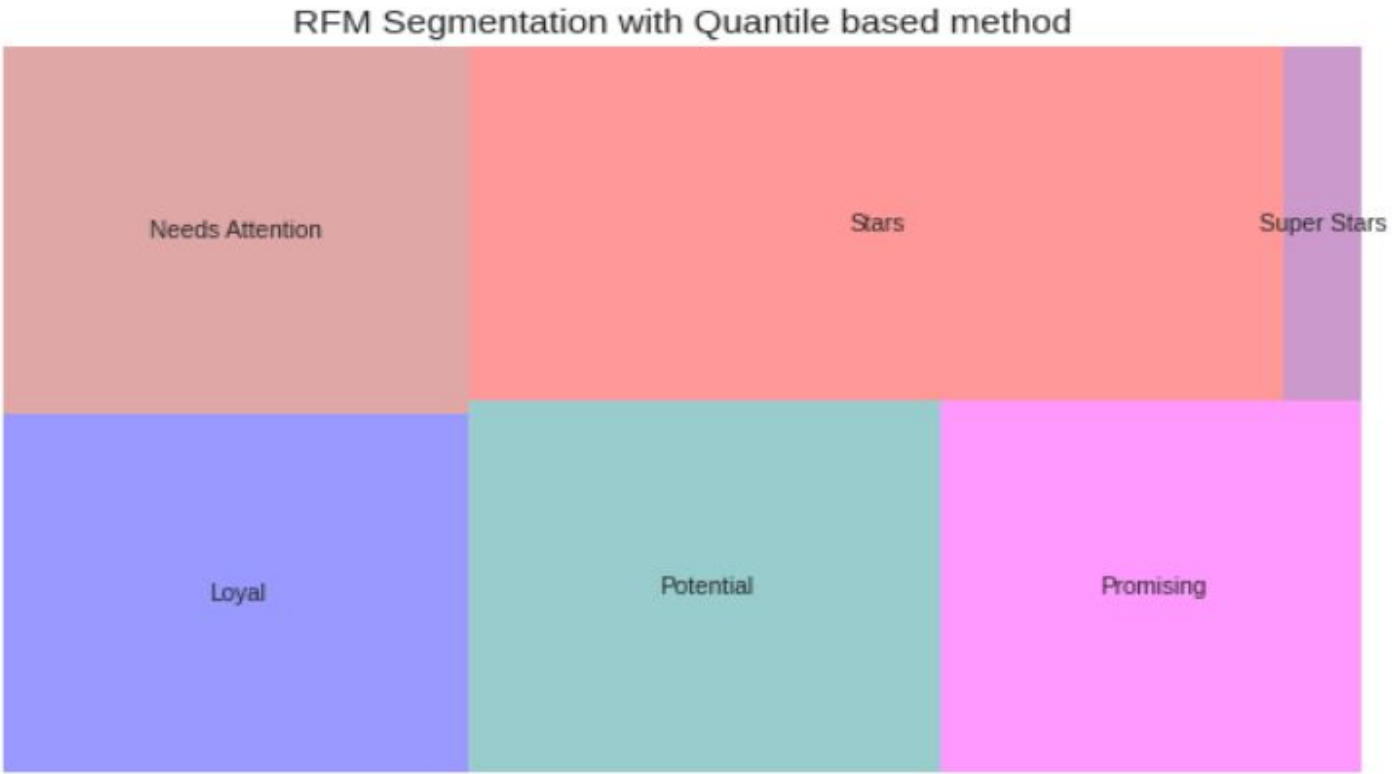
RFM segmentation

1. R : nombre de jour écoulé du dernier achat
2. F : nombre d'achats effectués.
3. M : somme totale dépensée.

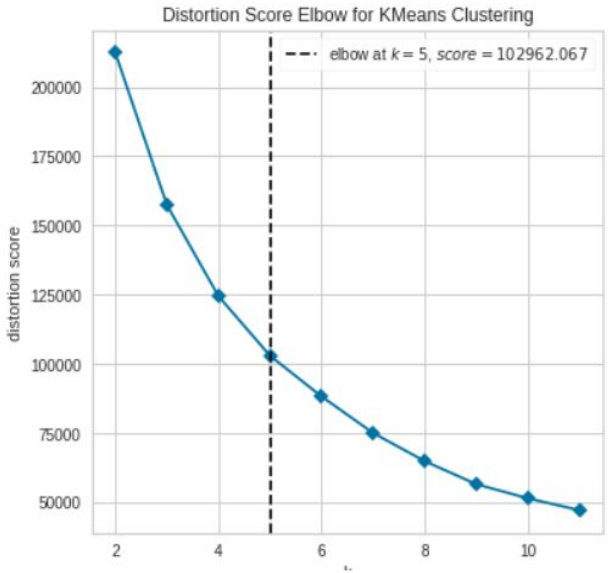
```
from datetime import datetime, timedelta

df.order_approved_at = pd.to_datetime(df.order_approved_at)
# Create snapshot date
max_date = df.order_approved_at.max() + timedelta(days=1)
# Grouping by customer_unique_id
rfm = df.groupby('customer_unique_id').agg(
                    recency = ('order_approved_at', lambda x: (max_date-x.max()).days/30),
                    frequency = ('order_id', 'count'),
                    monetary = ('payment_value', 'sum'))
```

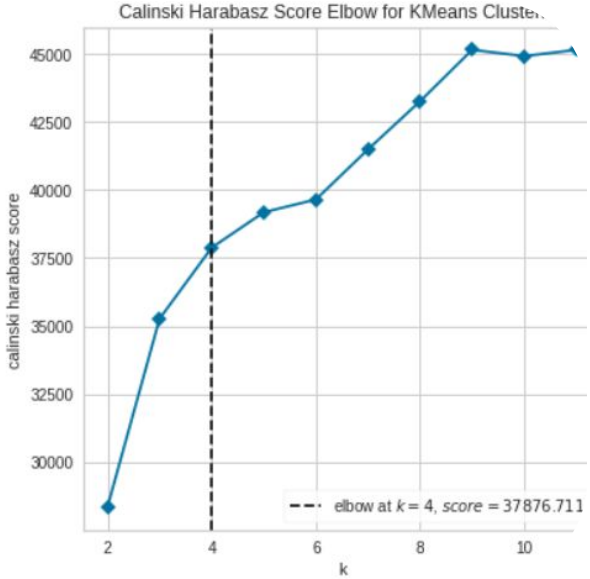


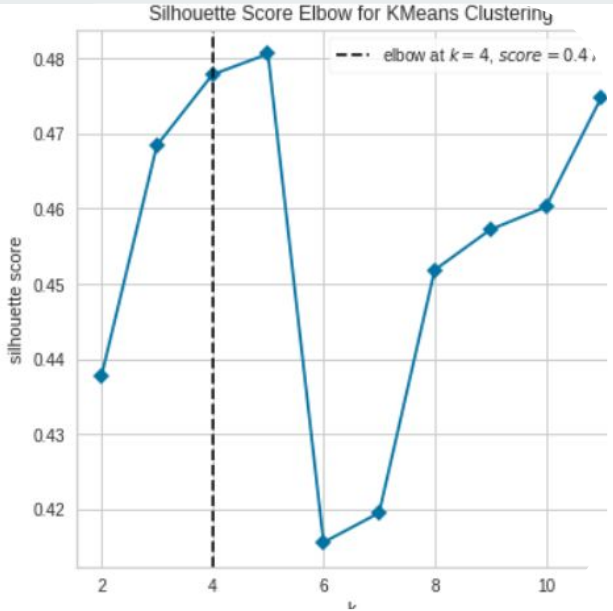
RFM avec discrétisation du Quantile



RFM-Kmeans clustering

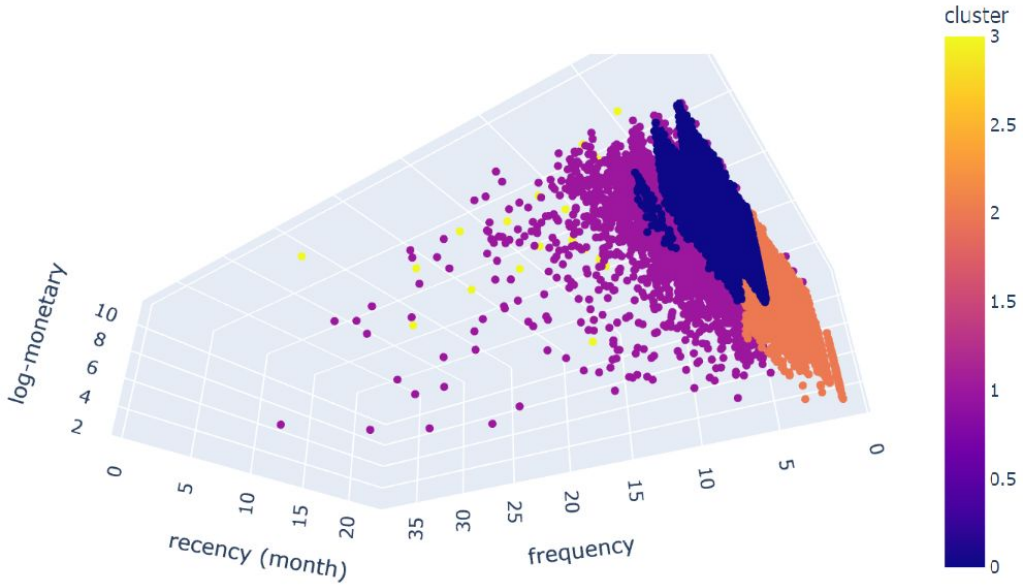


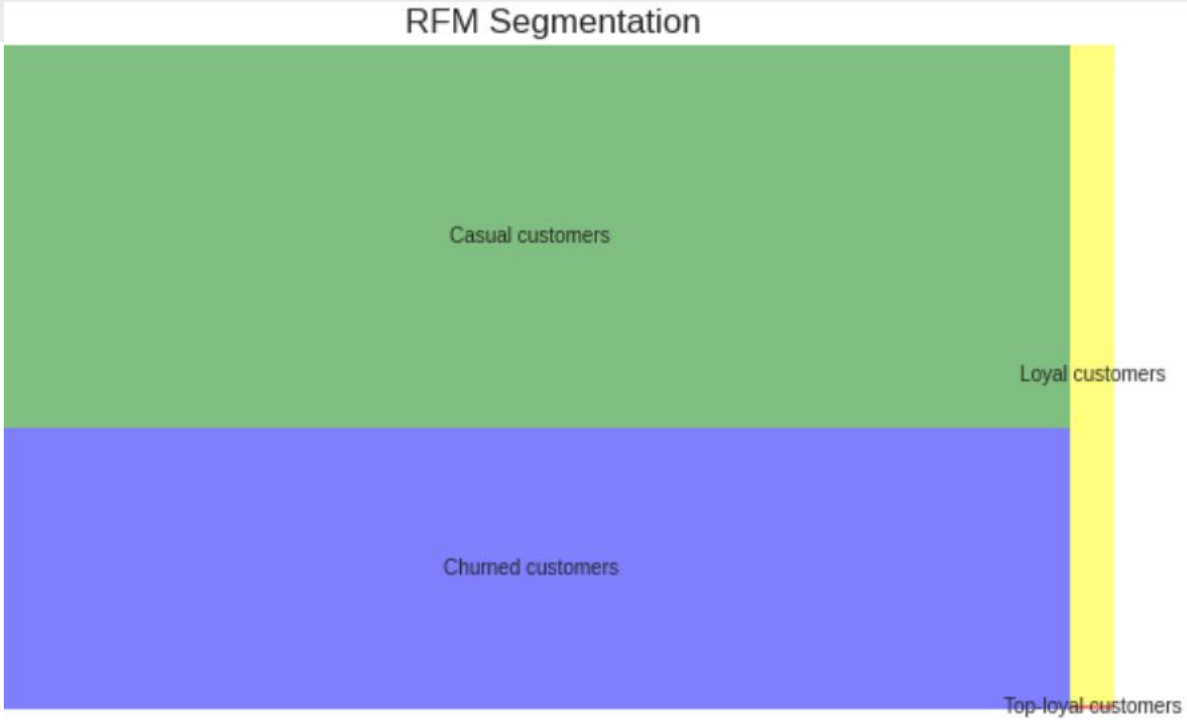




K=5 silhouette maximale (cluster avec peu de clients)
K=4 le nombre de cluster choisi

	recency		frequency		monetary		number_customer
	mean	std	mean	std	mean	std	
cluster							
0	12.910595	3.199097	1.123105	0.340857	169.656791	215.163690	37667
1	4.247419	2.409993	1.114302	0.322212	169.284811	209.054645	51145
2	7.645833	4.338122	14.250000	14.034894	23827.422917	20235.140205	24
3	7.689066	4.732042	4.088938	2.164893	1096.495183	1384.908121	3643

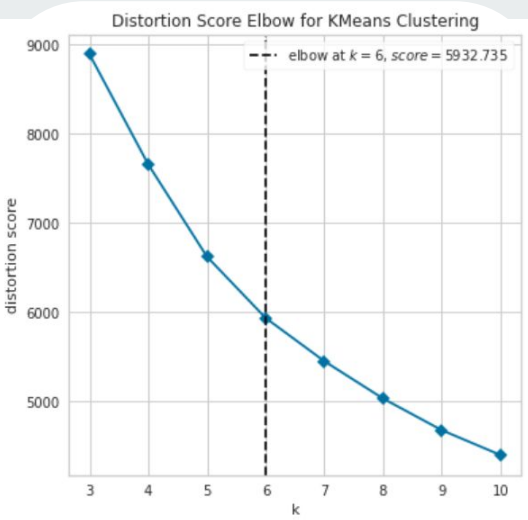


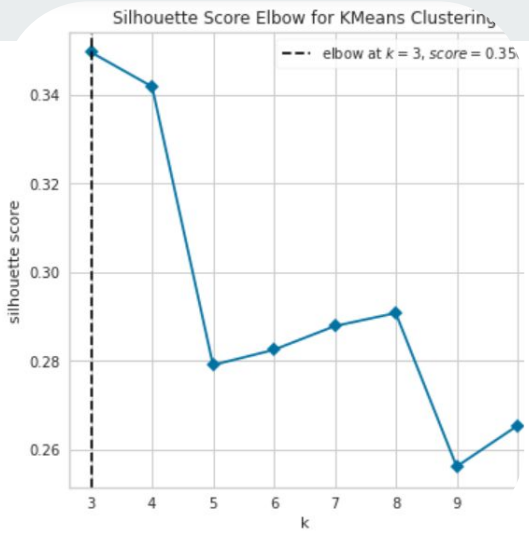


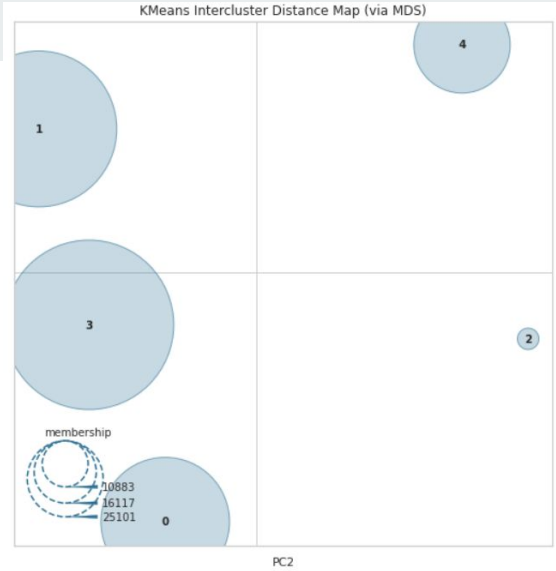
- *Top-loyal customers* : Les promotions ne sont pas nécessaires pour ce type de clients.
- *Loyal customers* : Aussi les promotions ne sont pas nécessaires pour ce type de clients, des points de fidélité peut suffire pour les garder.
- *Churned customers* On peut les attirer à reprendre l'abonnement par une proposition de certaines bénéfices attirantes.
- *Casual customers* : Afin de les encourager, on peut fournir à ce type de clients des bons sous forme de cashback avec un seuil d'éligibilité.

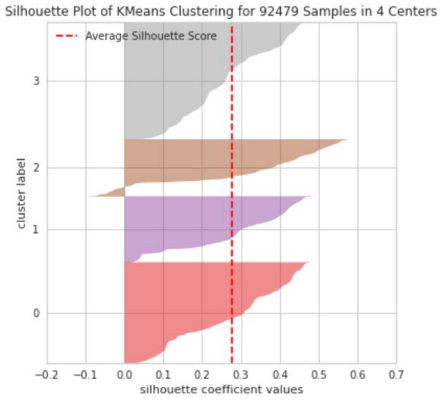
Segmentation avec KMeans algorithme (dataset total)

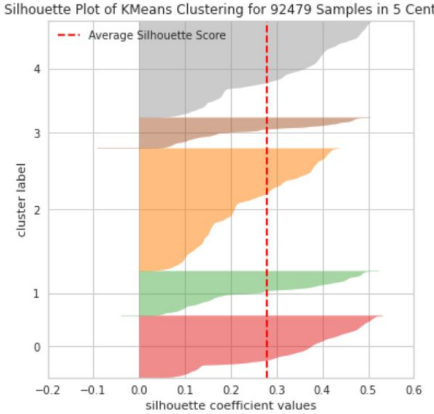
KMeans clusternig (Elbow method)

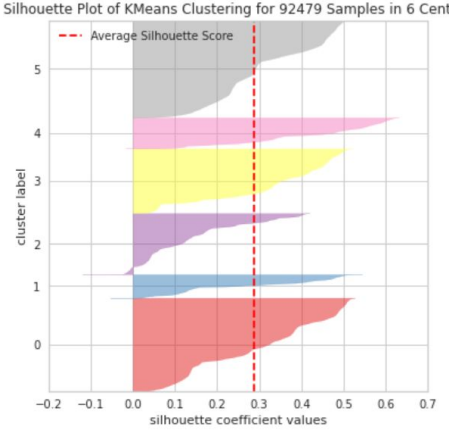


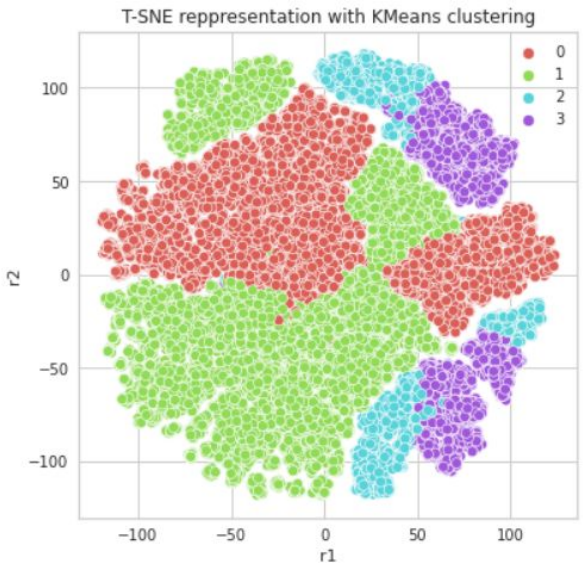


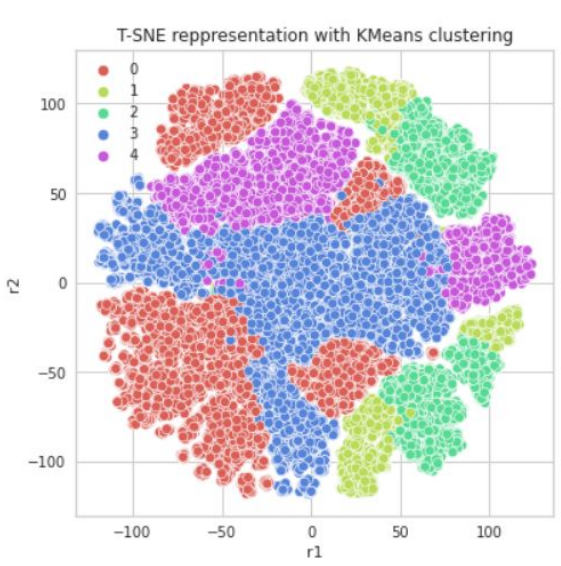


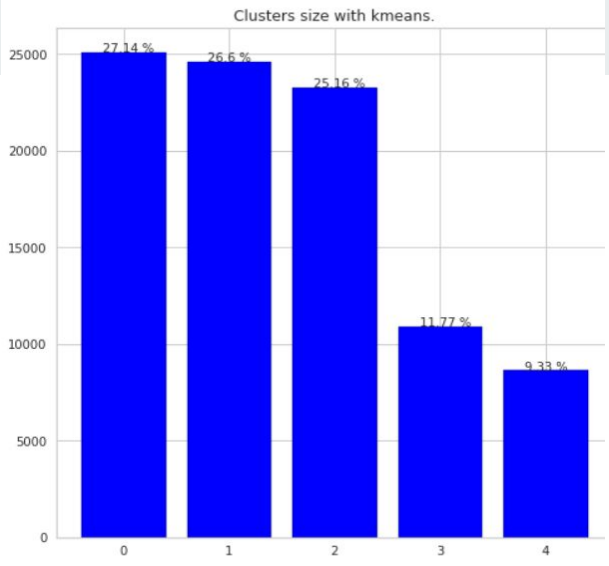






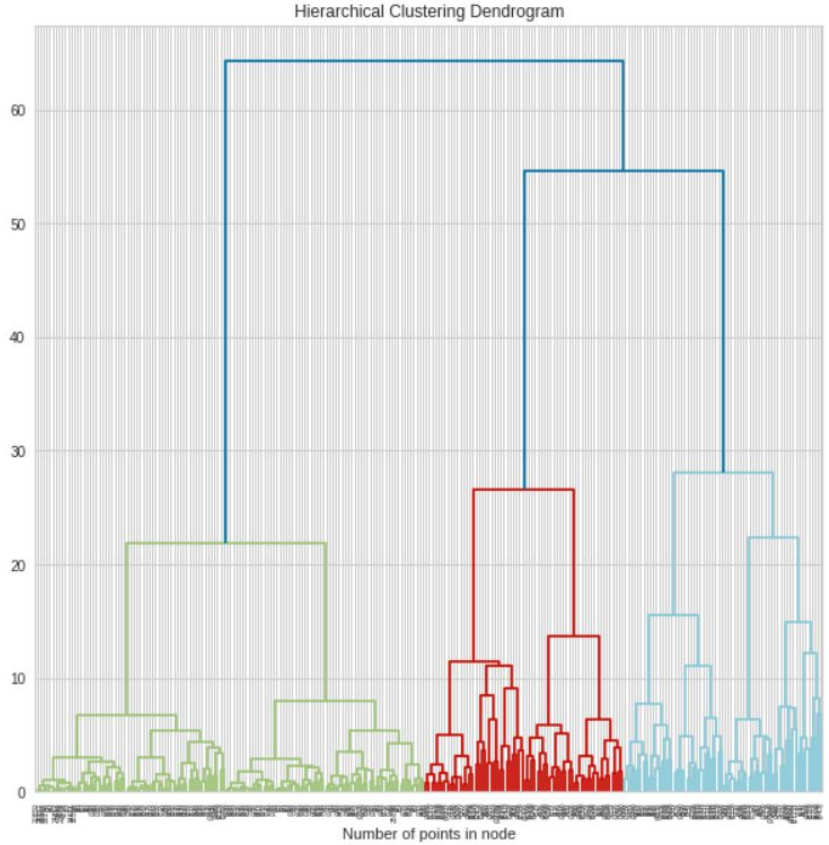


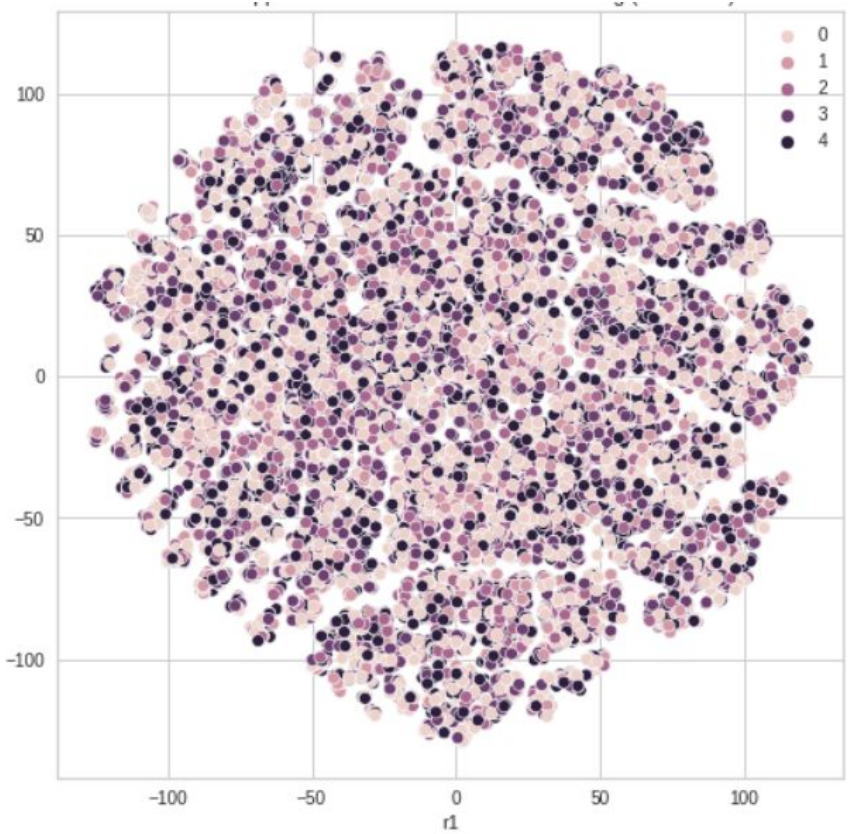




- Répartition en 5 cluster 5 : (silhouette score : 0.28)
- Répartition compacte selon la distance intercluster
- Stabilité à l'initiation

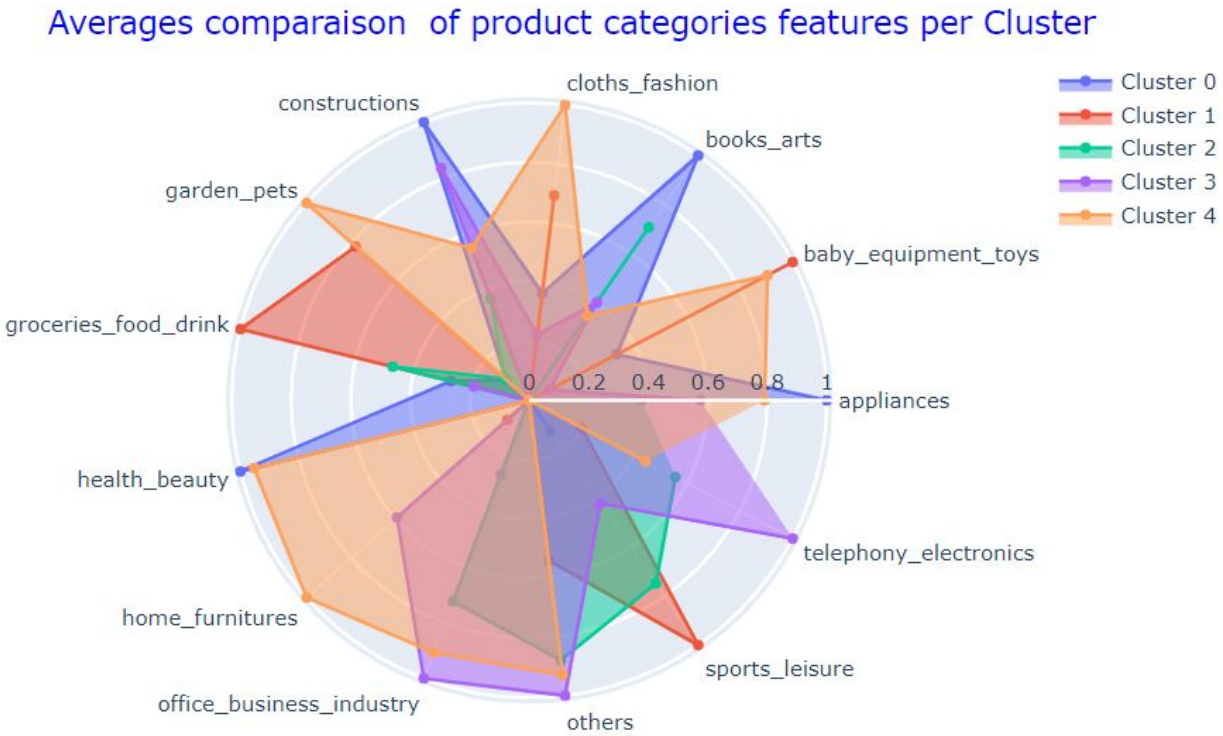
Clustering hiérarchique (5 clusters)



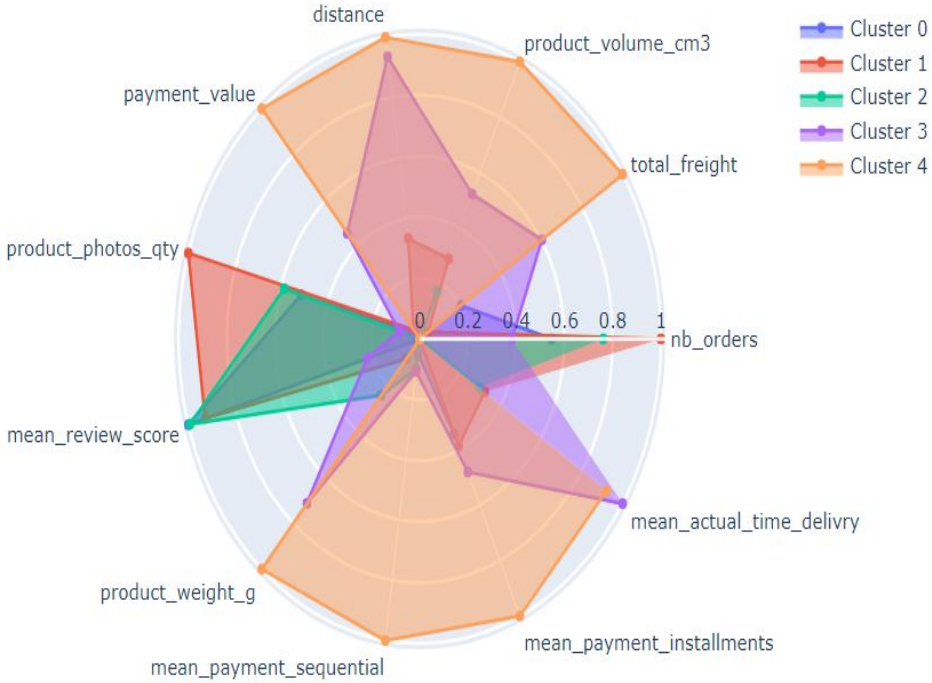


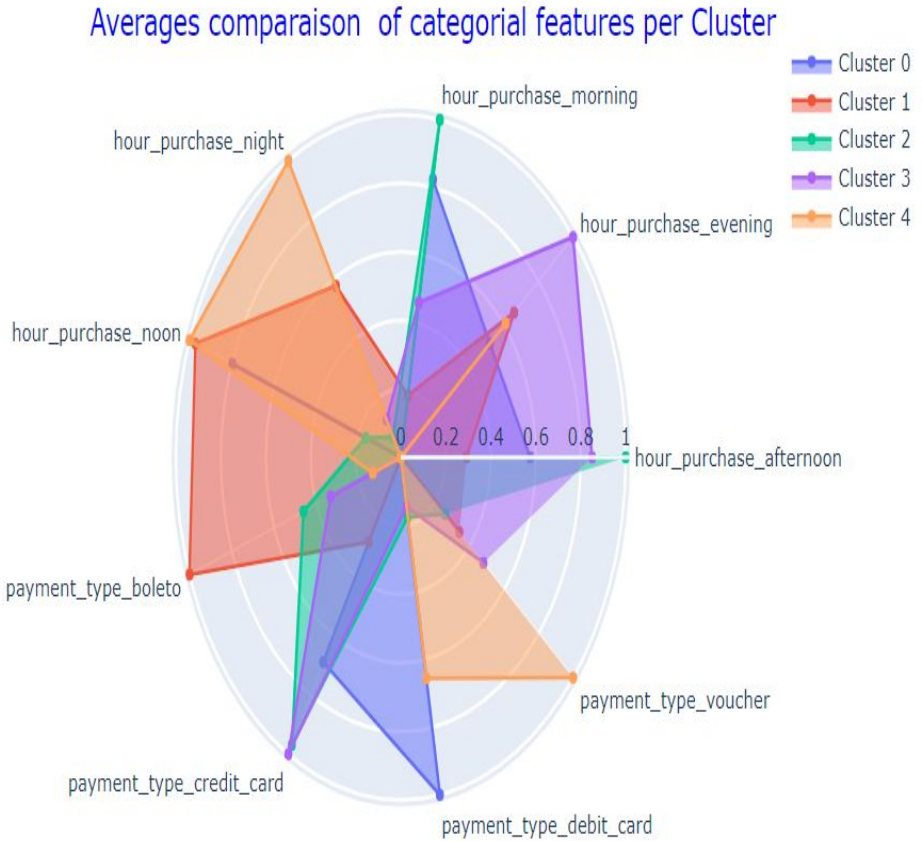
Modèle retenu (Kmeans avec 5 clusters)

5 Clusters identifiés avec des caractéristiques des clients de chaque cluster
Projection Radar des features



Modèle retenu (Kmeans avec 5 clusters)





Sensibilité temporelle du modèle

Identification de la période de maintenance:

- Vérification de la stabilité des clients dans le temps, avec ARI score, du coefficient de silhouette sur une durée d'une année tous les 2 mois

- période de mise à jour : 2 mois
- ARI score décroît avec un point d'inflexion au point 2 mois.
- Coefficient de silhouette quasi-stable sur les deux premiers mois

Conclusion

- Application des approches de classification non supervisée à un problème métier
- L'analyse RFM est plus pratique, permet d'identifier différentes clients (ex: **les meilleurs**)
- L'analyse RFM à des limites lorsque les clients ont un seul achats (frequency proche de 1)
- Les algorithmes automatiques sont flexibles avec plus de choix de features

Merci de votre attention!