

Projet 5

Segmentez des clients d'un site e-commerce

Zeynep Erdem
08-07-2023





Sommaire

- Rappel de la Problématique et Environnement
- Présentation et nettoyage des données
- Analyses Exploratoires des Données
- Présentation de démarche de modélisation
- RFM modélisation
- RFM + Satisfaction modélisation
- RFM + Satisfaction + Durée de Livraison modélisation
- Modèle final sélectionné et actions à faire
- Le délai de maintenance
- Conclusion et Recommandations

Rappel de la Problématique

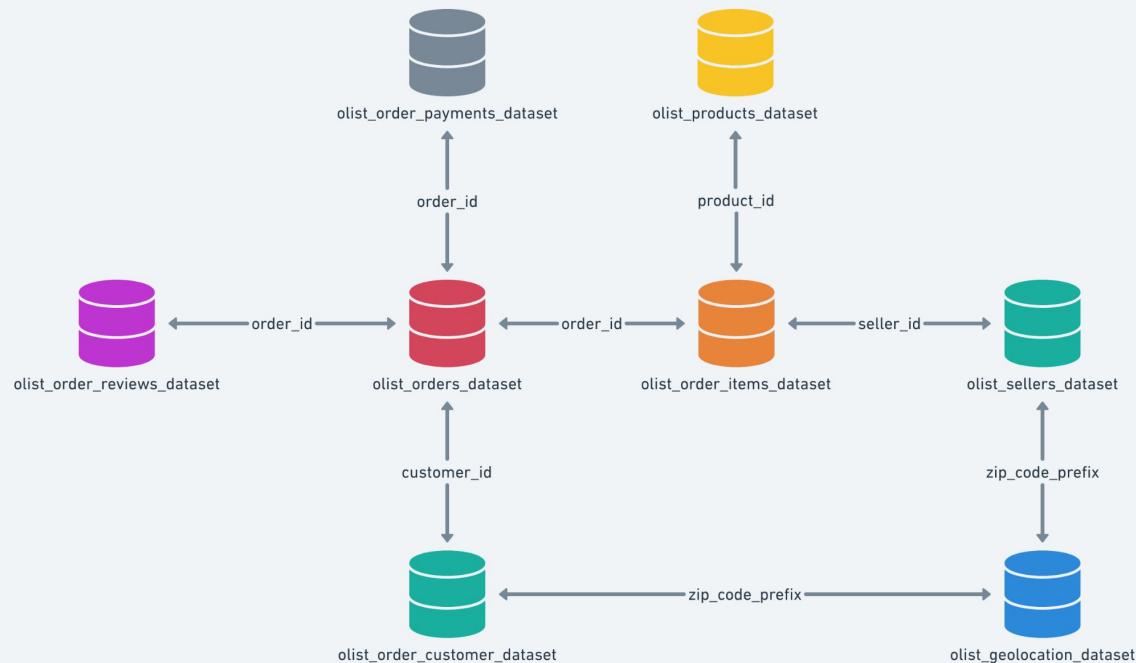


- ❖ Fournir à l'équipe e-commerce une segmentation des clients pour leurs campagnes de communication.
- ❖ Description détaillée et actionnable de chaque segment identifié.
- ❖ Différenciation des bons et moins bons clients en termes de commandes et de satisfaction.
- ❖ Proposition d'un contrat de maintenance.
- ❖ Respect de la convention PEP8 pour le code fourni.

Environnement

- Python: 3.9.16
- Pandas: 2.0.2
- Numpy: 1.24.3
- Seaborn: 0.12.2
- Matplotlib: 3.7.1
- Missingno: 0.5.2
- Sklearn: 1.2.2
- Scipy: 1.10.1
- Plotly: 5.15.0
- Black pour PEP8

Présentation du Jeu de Données



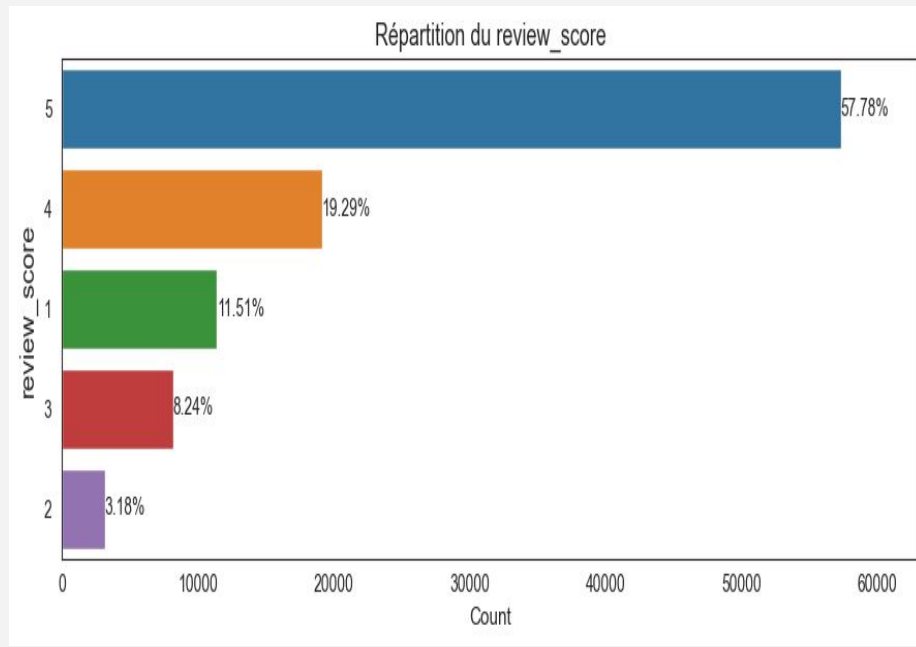
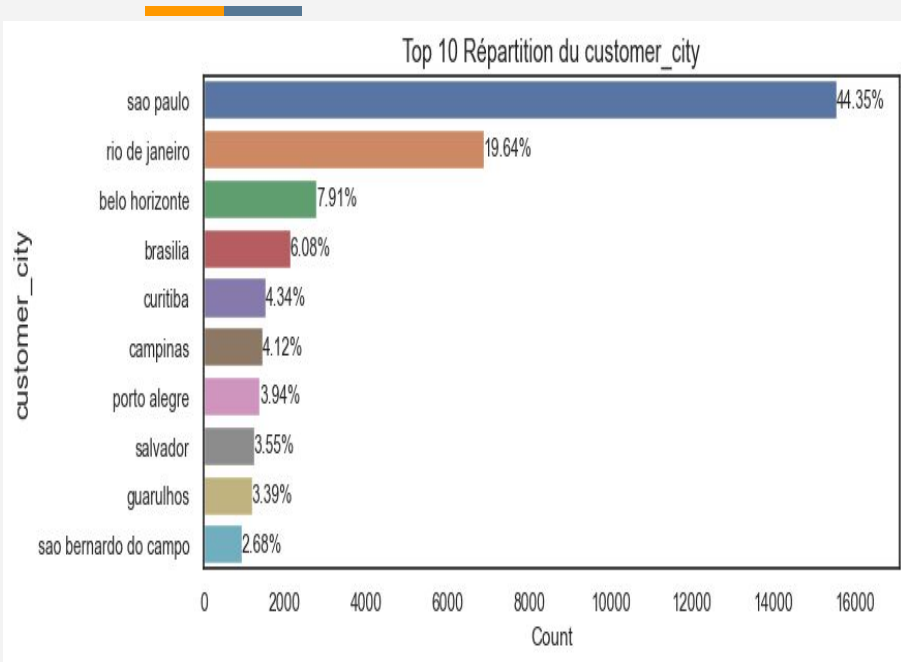
- Il y a environ 96000 clients.
- Seuls 3 % ont réalisé plusieurs commandes.
- Les commandes sont entre Septembre 2016 et Octobre 2018

Les étapes du nettoyage



- Elimination des colonnes trop de valeurs manquantes et ne sont pas nécessaires
- Dédoublonner les clé primaires
- Imputation avec médian ou moyenne des colonnes
- Les valeurs atypiques
- On les a laissé
- Récence, Fréquence, Montant
- Durée de livraison
- Mois et jour

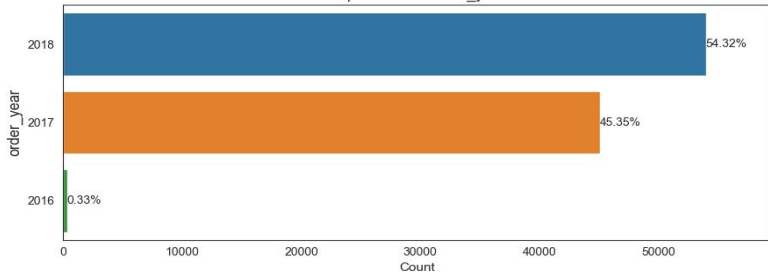
Analyses Exploratoires des Données



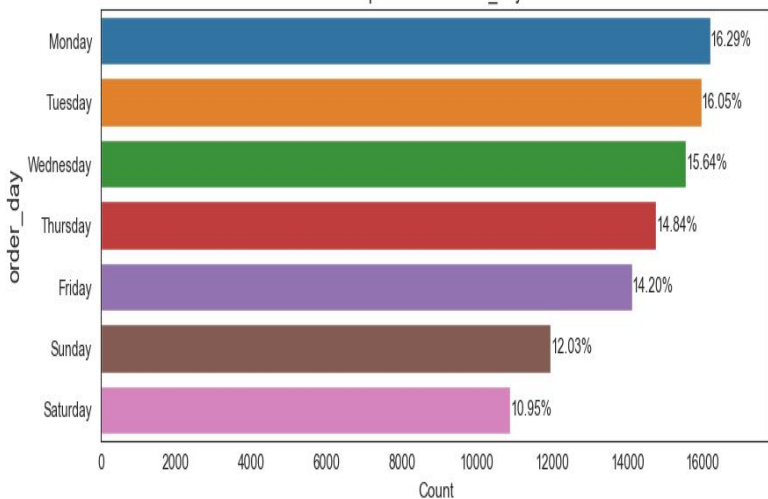
La plupart des clients se situent sur la côte et ils sont satisfaits de nos services.

Analyses Exploratoires des Données

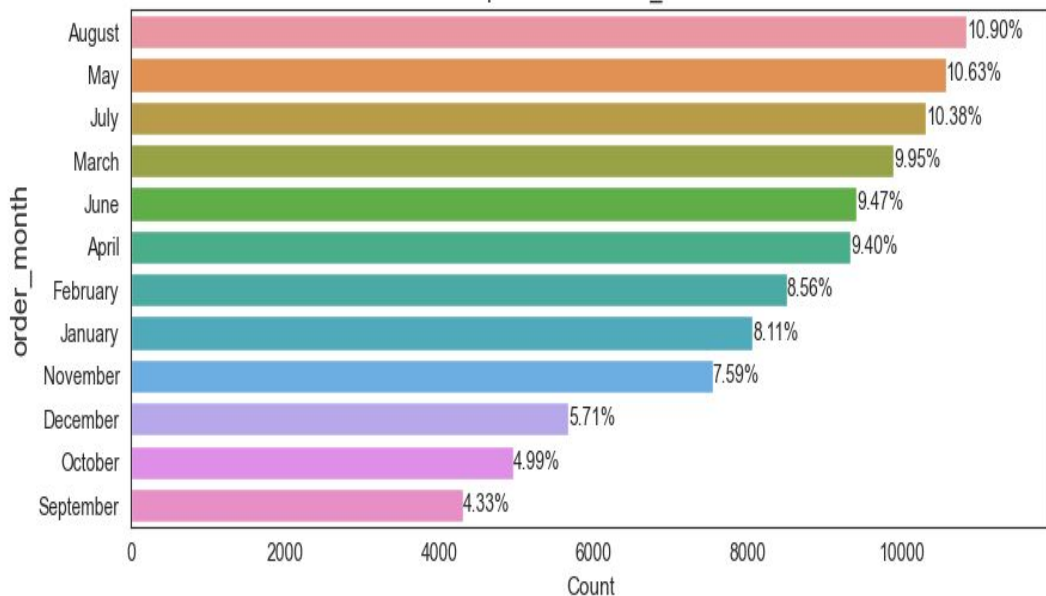
Répartition du order_year



Répartition du order_day

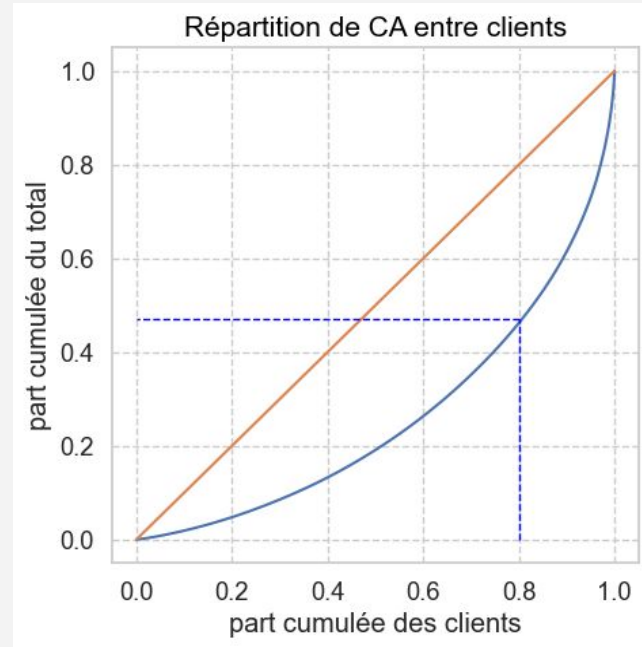
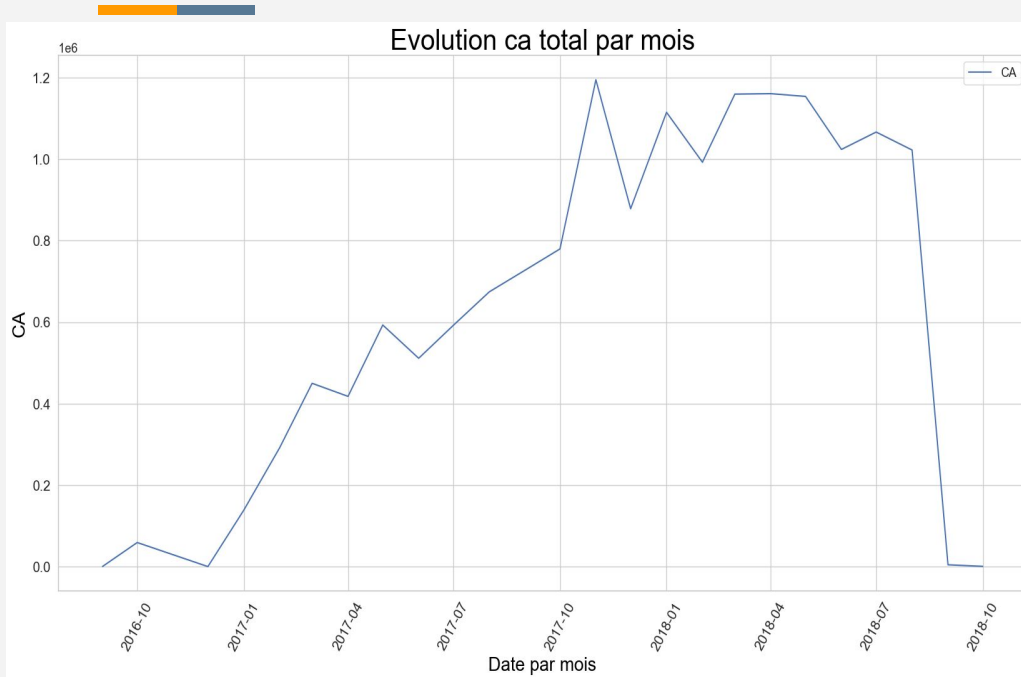


Répartition du order_month



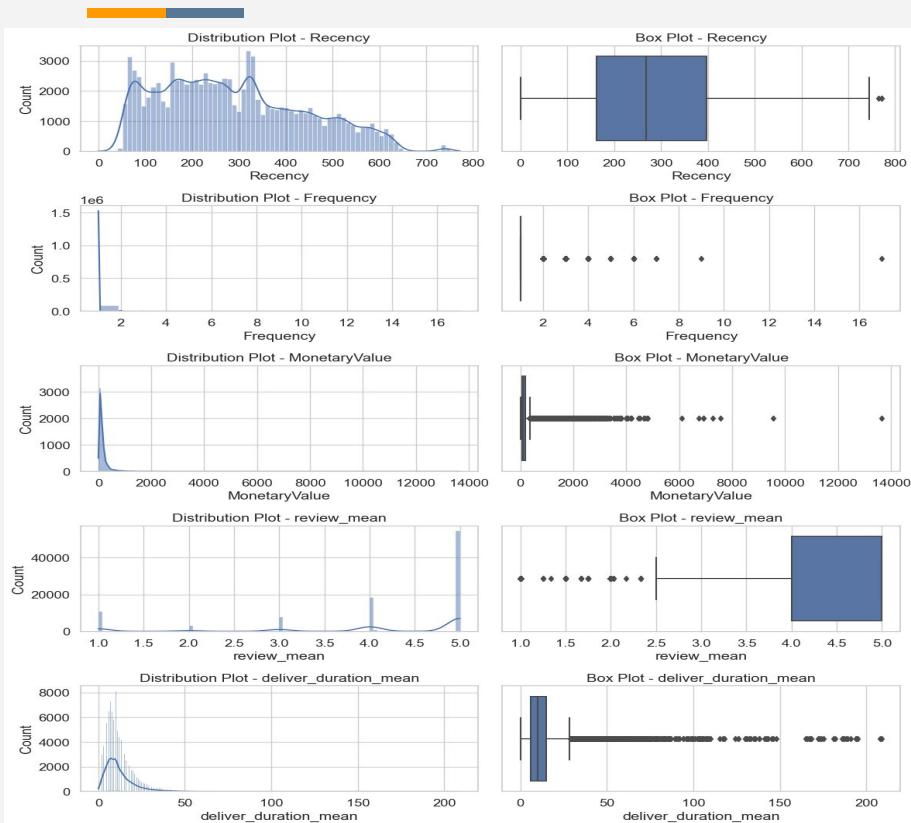
Plupart des commandes ont été effectuée pendant la semaine et les mois de l'été

Analyses Exploratoires des Données



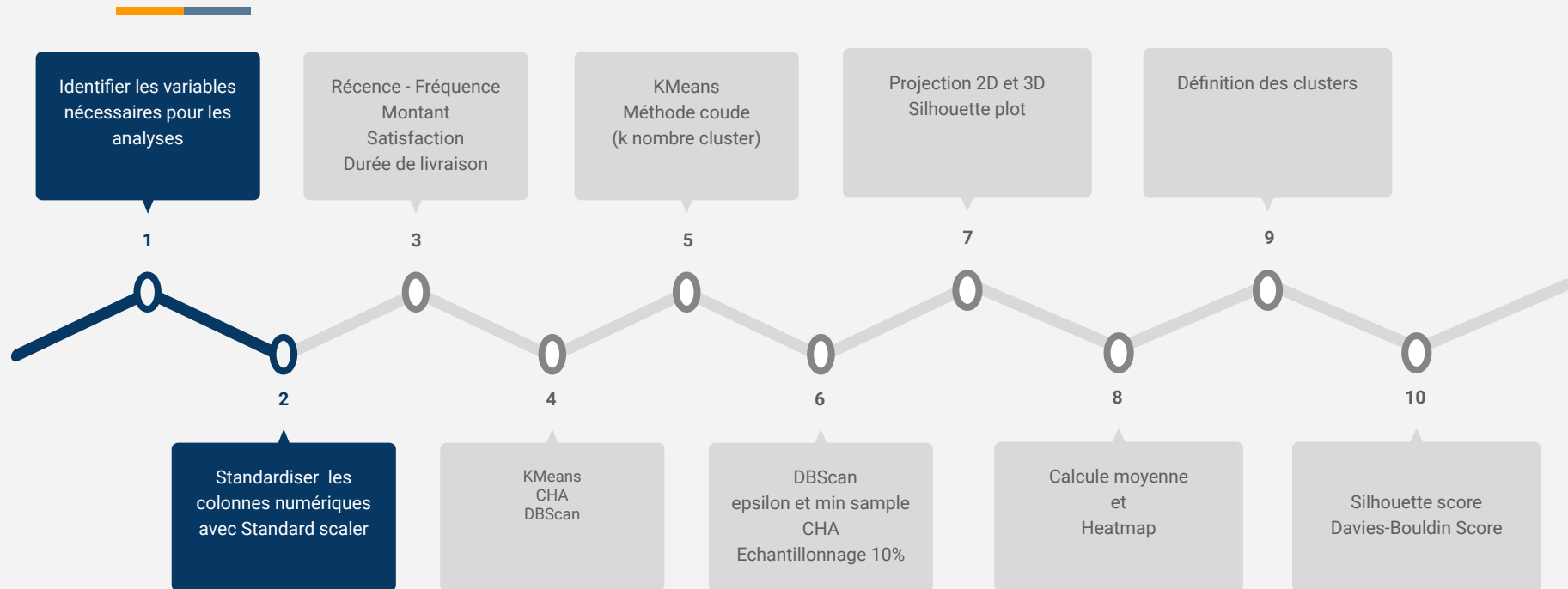
- On a plus de chiffre d'affaire pendant l'année 2018
- Environ 20% de clients génèrent 50% de chiffre d'affaire

Analyses Exploratoires des Données

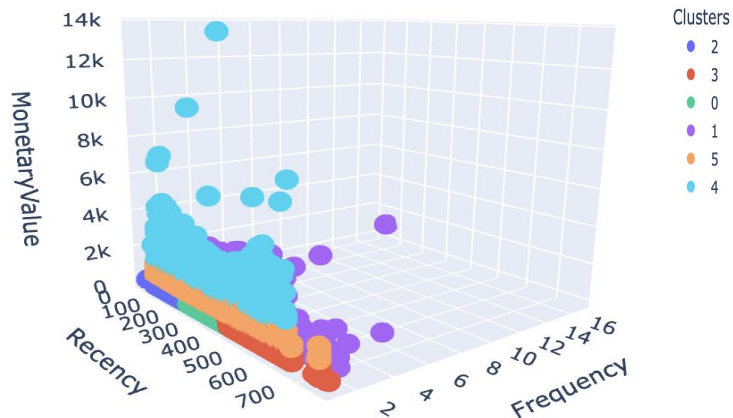
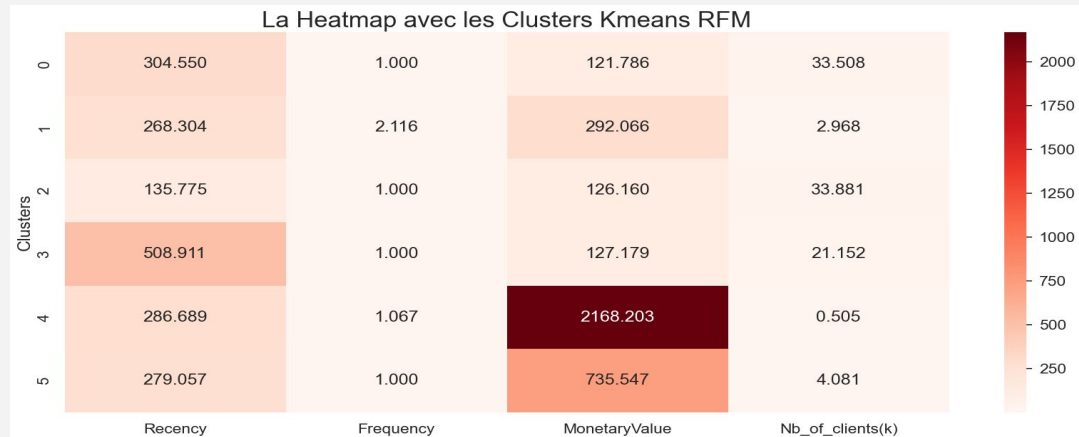
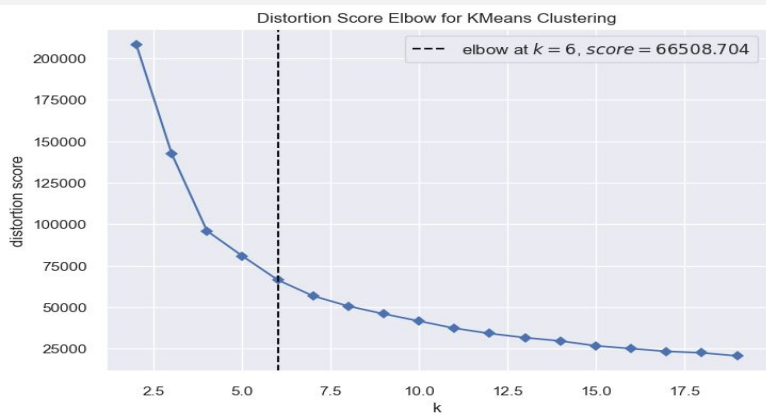


- Seuls 25% des clients ont fait une commande depuis cinq mois.(moyen 287 jours)
- Le client le plus actif a fait 17 fois de commande
- Le moyen de montant total des achats est d'environ 165 réal brésilien
- Le plupart des clients sont satisfaits
- La durée moyenne de livraison est de 12 jours

La démarche de modélisation



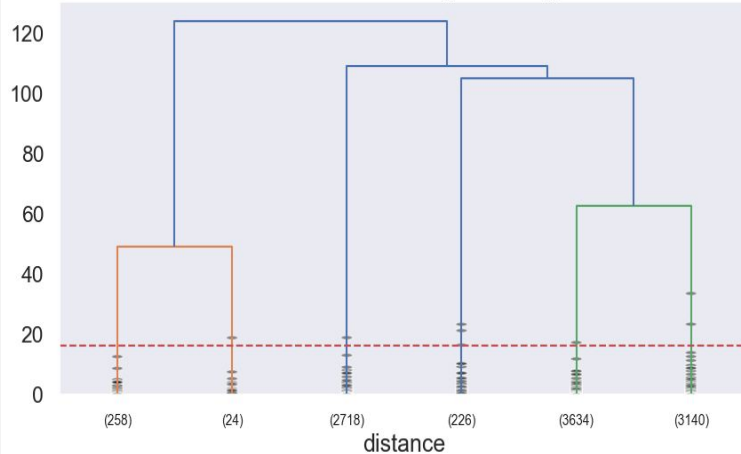
KMeans pour RFM



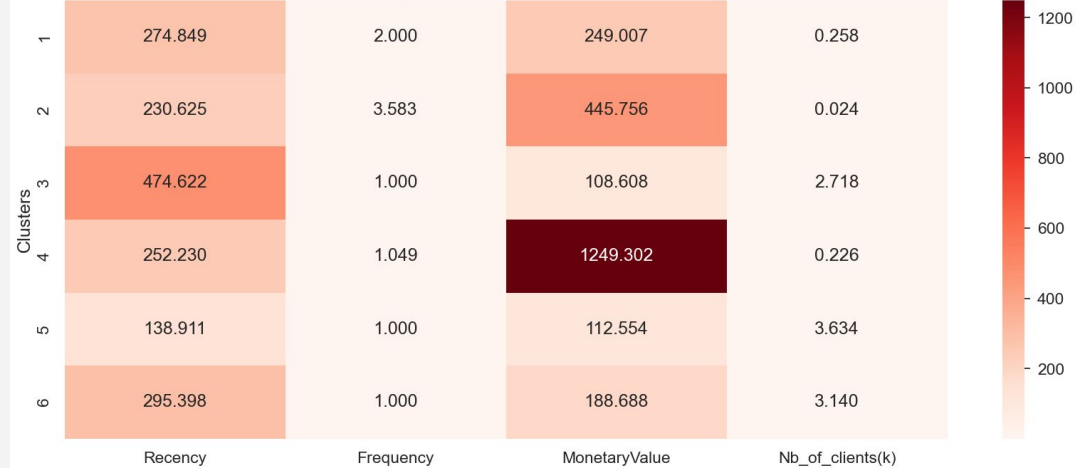
Cluster 0 : "Clients peu récents et à faible dépense"
 Cluster 1 : "Clients actifs et de valeur moyen"
 Cluster 2 : "Clients nouveaux et à faible dépense"
 Cluster 3 : "Clients anciens et à faible dépense"
 Cluster 4 : "Clients à forte valeur (flambeurs) et peu nombreux"
 Cluster 5 : "Clients peu récents, dépenses plutôt élevé avec potentiel"

CHA pour RFM

Hierarchical Clustering Dendrogram



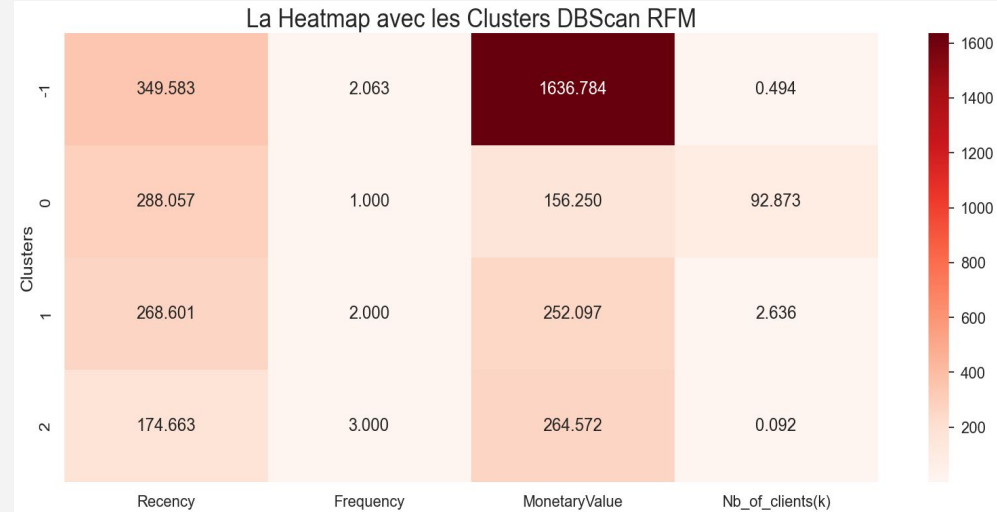
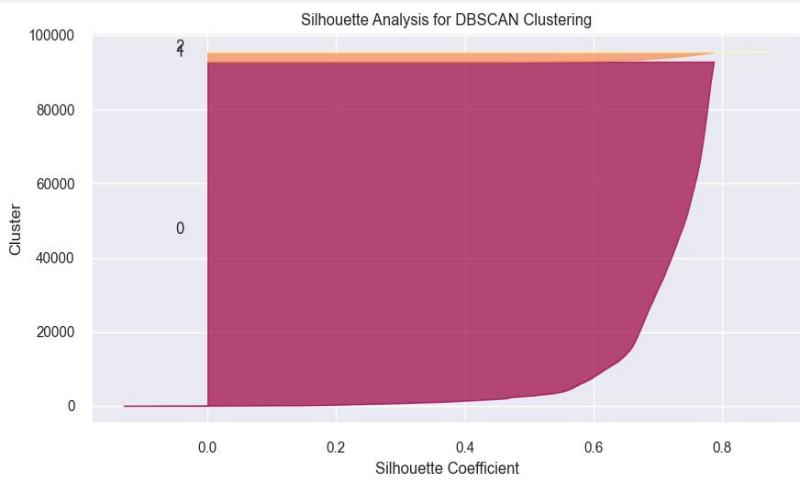
La Heatmap avec les Clusters avec CAH



A cause du temps trop long de calcul on a pris 10000 individus comme échantillon

Cluster 1 : "Clients actifs et de valeur moyen"
 Cluster 2 : "Acheteurs réguliers et à valeur plutôt élevée"
 Cluster 3 : "Clients anciens et à faible dépense"
 Cluster 4 : "Clients à forte valeur mais peu nombreux"
 Cluster 5 : "Clients nouveaux et à faible dépense"
 Cluster 6 : "Acheteurs peu récents et de valeur modérée"

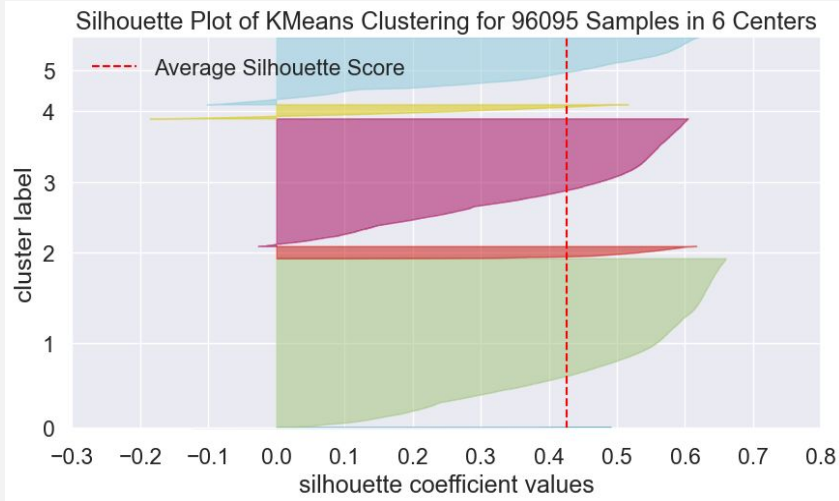
DBScan pour RFM



Malgré le score élevé du coefficient de silhouette, le DBScan divise les clients en 3 clusters qui ne sont pas équilibrés.

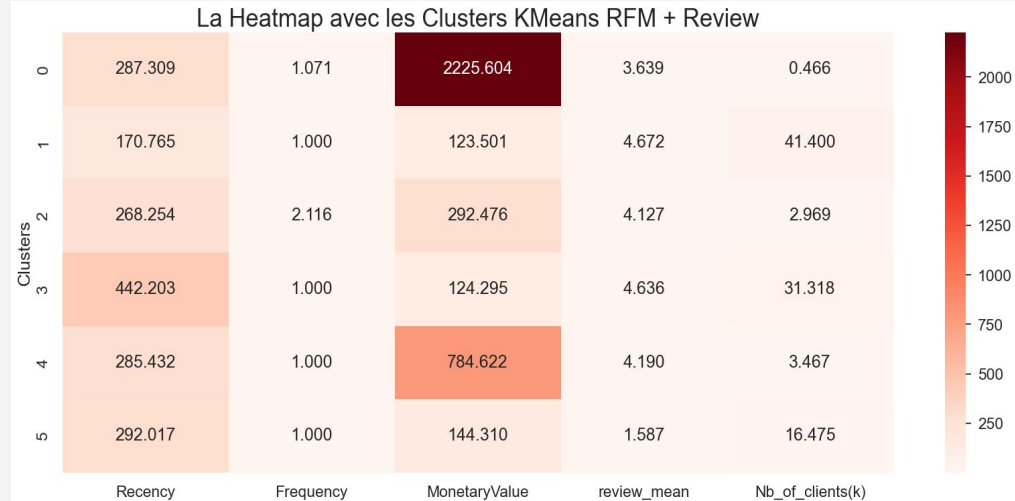
Cluster -1 : "Clients à forte valeur mais peu nombreux"
 Cluster 0 : "Clients anciens et à faible dépense"
 Cluster 1 : "Clients actifs et de valeur moyen"
 Cluster 2 : "Acheteurs réguliers et à valeur plutôt élevée"

KMeans pour RFM + Satisfaction



Les clusters ont un score de silhouette supérieur à la moyenne.

Ils sont plutôt équilibrés, à l'exception des clusters qui contiennent des valeurs extrêmes.



Cluster 0 : "Clients à forte valeur mais satisfaction modéré"

Cluster 1 : "Clients nouveaux à faible dépense et satisfaits"

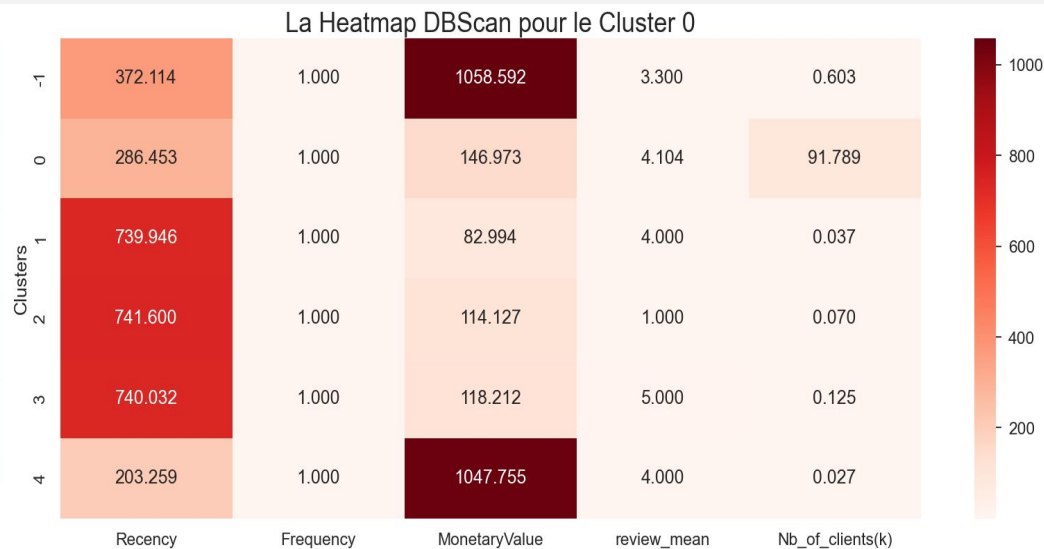
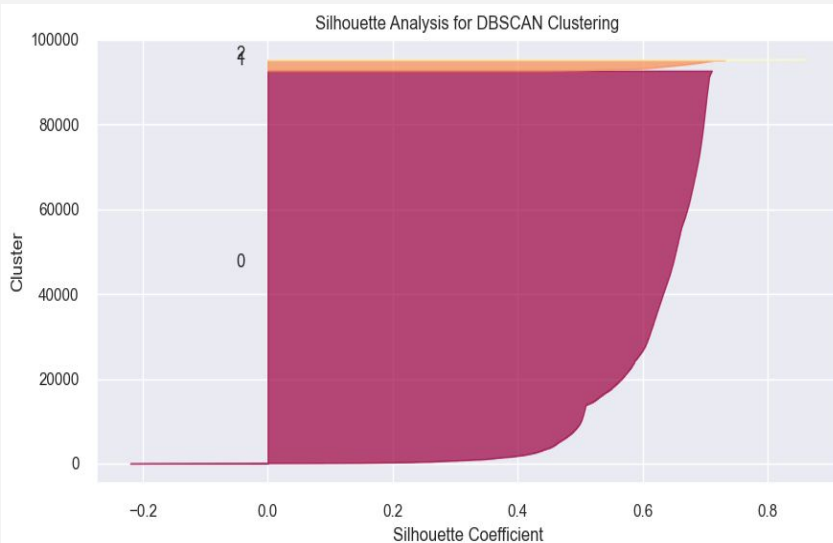
Cluster 2 : "Clients actifs et plutôt satisfaits"

Cluster 3 : "Clients anciens et à faible dépense et satisfaits"

Cluster 4 : "Clients peu récents dépenses plutôt élevé et satisfaits"

Cluster 5 : "Clients peu récents à faible dépense et mécontents"

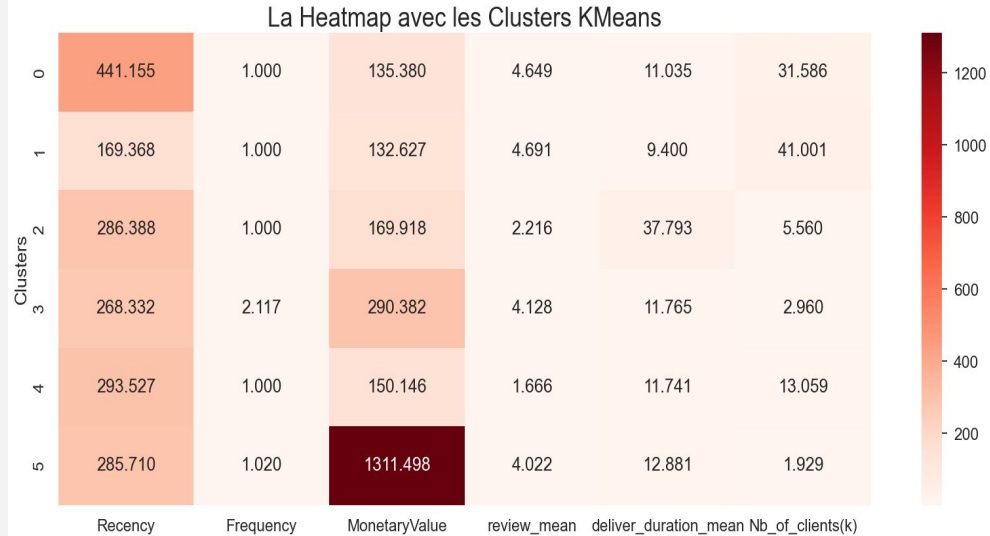
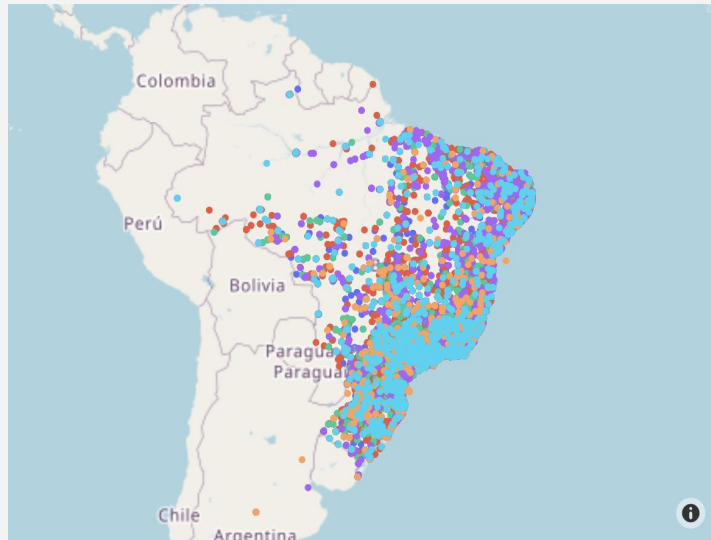
DBScan RFM + Satisfaction



On a effectué une deuxième segmentation pour le cluster 0

Cluster -1 : "Clients à forte valeur peu récents mais satisfaction modéré"
 Cluster 0 : "Clients peu récents à faible dépense et satisfaits"
 Cluster 1 : "Clients anciens à faible dépense et plutôt satisfaits"
 Cluster 2 : "Clients anciens à faible dépense et mécontents"
 Cluster 3 : "Clients anciens à faible dépense et satisfaits"
 Cluster 4 : "Clients à forte valeur plutôt récents et plutôt satisfaits"

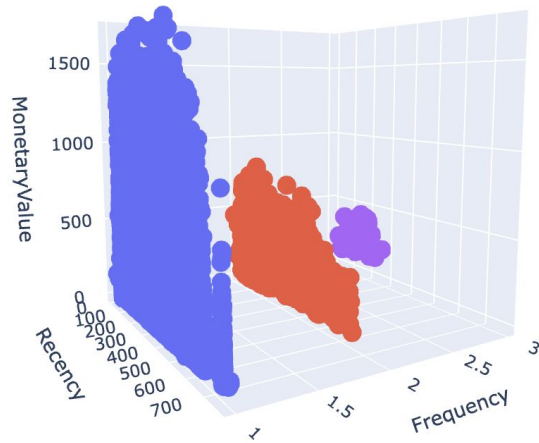
KMeans pour RFM + Satisfaction + Durée de livraison



La plupart des clients se situent sur la côte

Cluster 0 : Clients anciens à faible dépense, satisfaits et livraison moyen
 Cluster 1 : Clients nouveaux à faible dépense, satisfaits et livraison rapide
 Cluster 2 : Clients peu récents à faible dépense, mécontents et livraison long
 Cluster 3 : Clients actifs, plutôt satisfaits et engagés
 Cluster 4 : Clients peu récents à faible dépense, mécontents et livraison moyen
 Cluster 5 : Clients à forte valeur plutôt satisfaits livraison moyen

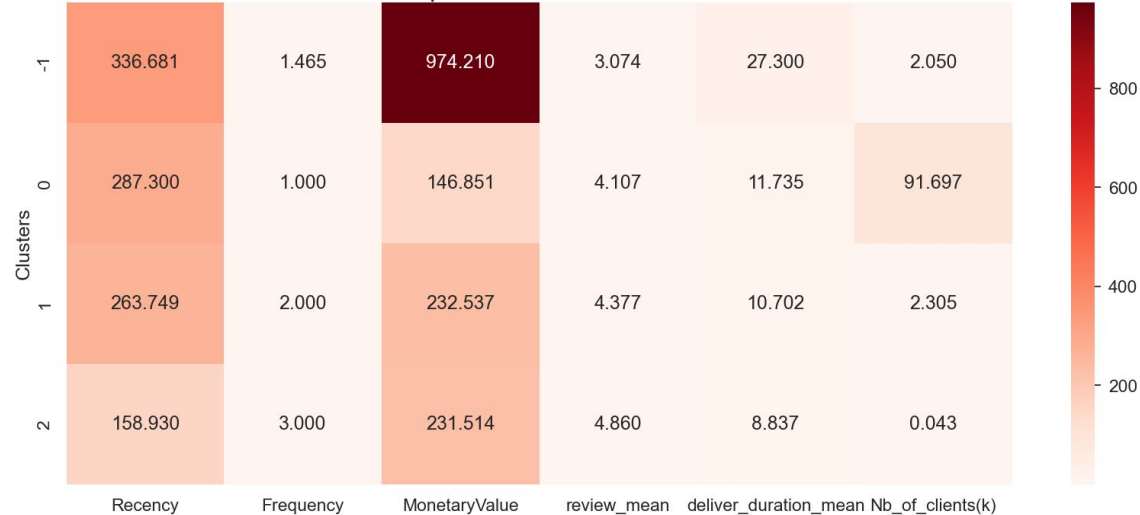
DBScan RFM + Satisfaction + Durée de livraison



Clusters

- 0
- 1
- -1
- 2

La Heatmap avec les Clusters avec ACP



Quand on exclue les valeurs extrêmes le modèle fait des segmentation par rapport à fréquence d'achat


Cluster -1 : Clients à forte valeur, satisfaction modéré, livraison long

Cluster 0 : Clients peu récents à faible dépense, satisfaits et livraison moyen une seule achat

Cluster 1 : Clients peu récents à dépense moyen, satisfaits et livraison moyen, deux achats

Cluster 2 : Clients nouveaux à dépense moyen, satisfaits et livraison rapide trois achats

Comparaison des résultats Silhouette et Davies-Bouldin Scores



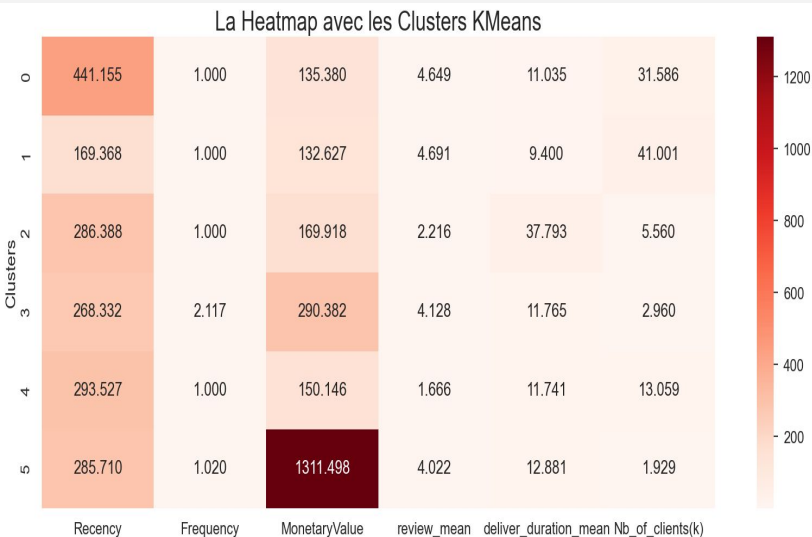
	Silhouette Score	Davies-Bouldin Score
KMeans_rfm	0.44	0.71
CAH_rfm	0.39	0.87
DBScan_rfm	0.70	1.16
KMeans_review	0.43	0.78
CAH_review	0.37	0.94
DBScan_review	0.61	1.23
Kmeans_plus	0.34	0.98
CAH_plus	0.30	1.05
DBScan_plus	0.55	1.34

Le modèle KMeans avec les variables RFM+Satisfaction+Durée livraison a donné les meilleurs résultats du point de vue métier

Choix de modèle - Clusters - Actions

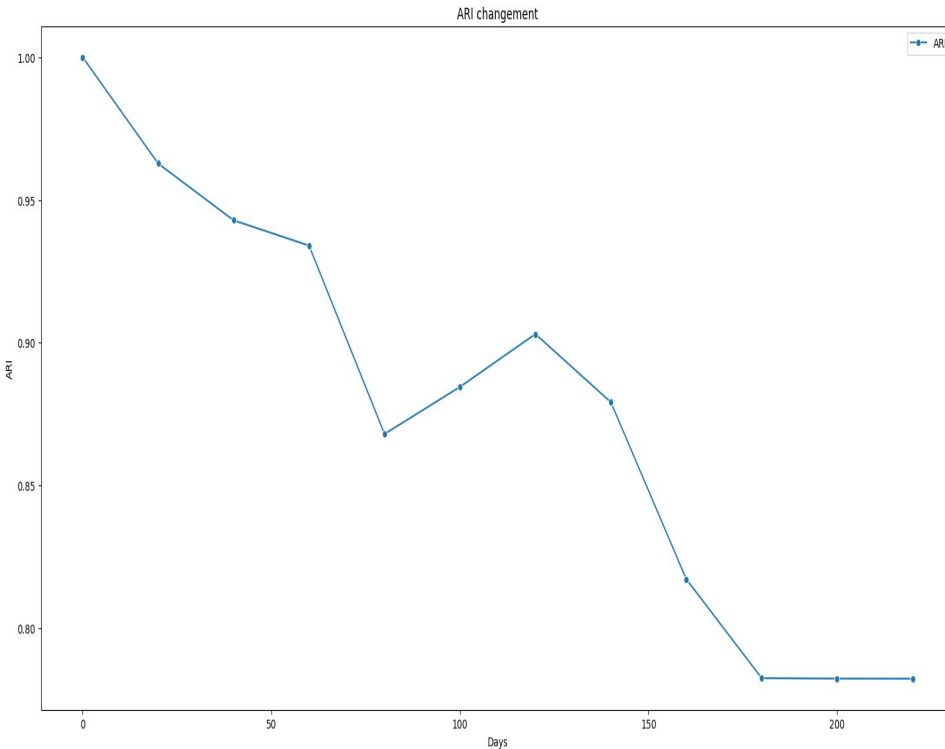


KMeans pour RFM + Satisfaction + Durée de livraison



- Cluster 0 : "**Clients anciens avec une satisfaction élevée et livraison moyen**"
 - Envoyer des publicités
 - Les encourager à effectuer des achats à nouveau
- Cluster 1 : "**Clients nouveaux avec faible dépense, satisfaits et livraison rapide**"
 - Utiliser des programmes de fidélisation pour augmenter leur valeur
 - Maintenir le délai de livraison court.
- Cluster 2 : "**Clients pas satisfaits, peu récent et délai de livraison long**"
 - Réduire le délai de livraison
 - Les communiquer de manière proactive afin de gérer leurs attentes.
- Cluster 3 : "**Clients actifs, plutôt satisfaits et engagés**"
 - Offrir des avantages exclusifs, des programmes de fidélisation
 - Les encourager à partager leurs expériences positives et à recommander nos services.
- Cluster 4 : "**Clients peu récents à faible dépense, pas satisfaits et livraison moyen**"
 - Trouver la raison de leur mécontentement
 - Proposer des offres personnalisées et des recommandations de produits.
- Cluster 5 : "**Clients à forte valeur plutôt satisfaits livraison moyen**"
 - Offrir des services premium ou des avantages exclusifs pour renforcer leur fidélité.
 - Communication proactives pour maintenir leur engagement et leur valeur.

La simulation pour le délai de maintenance



Modèle 0

période_0 : 9
mois de
données
(06/17-03/18)
KMeans_0
6 clusters
RFM +
Satisfaction +
Durée Livraison



Définir clusters intervalles des 20 jours

période_1 : 9
mois + 20 jours
KMeans_1
6 clusters
RFM +
Satisfaction +
Durée Livraison



Comparaison : ARI score

On compare les résultats de
période_0 avec M1 sur
période_1

On recommande une maintenance à partir
de 160 jours

Conclusion et Recommandations



- On a effectué les segmentations RFM et les variables satisfaction et durée de livraison .
 - On a utilisé KMeans, CHA et DBScan pour les modèles.
 - On a comparé les clusters par rapport aux scores silhouettes Davies- Bouldin et l'approche métier.
 - Le modèle KMeans avec 6 clusters a donné les meilleurs résultats du point de vue métier.
 - Les définitions des clusters et les actions possibles ont été définies.
 - Un délai d'environ 6 mois est prévu pour la maintenance du modèle.
-
- On peut améliorer nos résultats en faisant des hyperparamètres tuning plus fin pour les modèles.
 - On peut essayer de collecter des données supplémentaires.

