olist

Segmentation des clients d'un site e-commerce

Présentée par: LYNDA HADJEMI
Parcours: Data science

Sommaire

- Contexte et problématique .
- Piste suivis
- Structure de données
- Nettoyages et regroupement des datasets
- Créations des colonnes pertinentes pour l’analyse
- Analyse exploratoire
- Modélisations
- Contrat de maintenance
- Evolution de score ARI au fil du temps.
- Conclusion

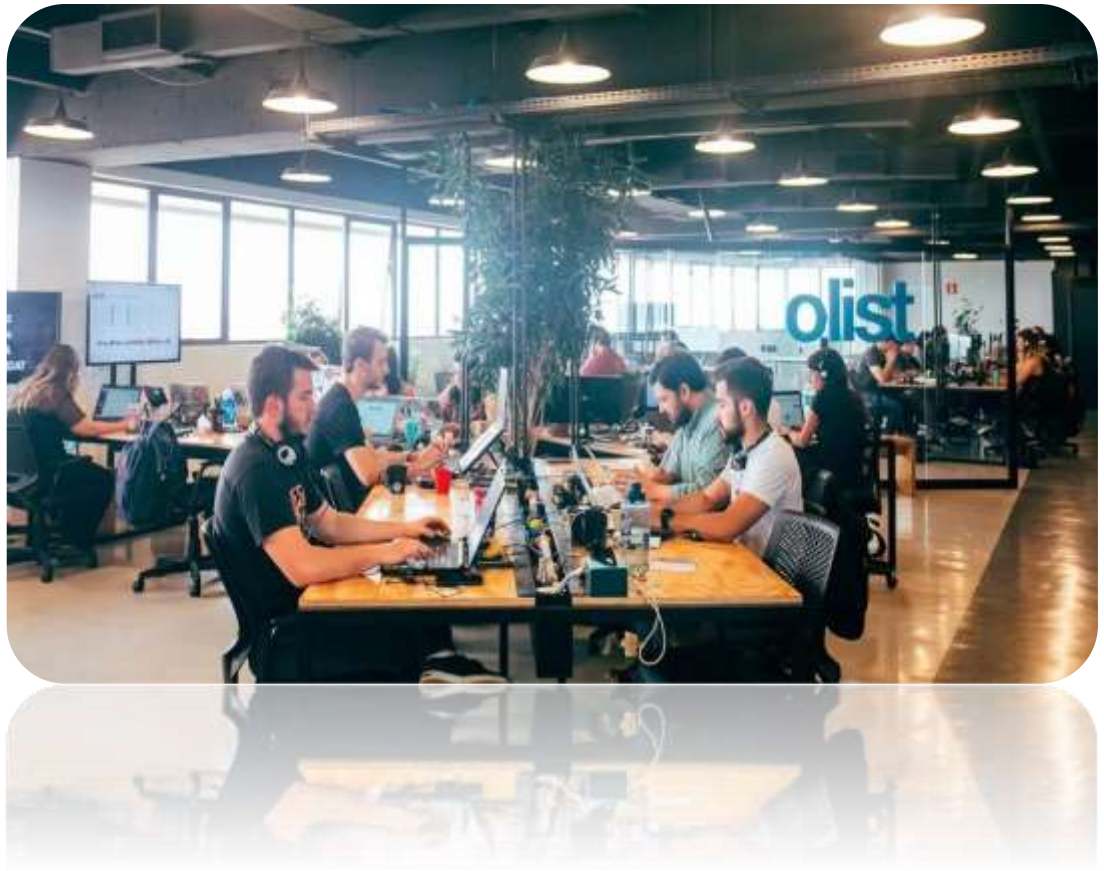
olist

Contexte et problématique:

- Olist , Spécialiste E-commerce Brésilien, Souhaite fournir à ses équipes marketing une segmentation des clients utilisable au quotidien pour leur campagnes de communication.

- L’objectif est de comprendre les différents types d’utilisateurs grâces a leurs données personnelles et à leur comportements.

- Fournir une description actionnable de cette segmentation et de sa logique sous-jacente ,ainsi qu’une proposition de contrat de maintenance basée sur une analyse de la stabilité des segments au cours de temps.



olist

Pistes suivies:

- En premier point , on pars sur une analyse précise de comportement de client par seulement les données RFM.

- En suite adapté ce modèle tout on analysons d'autres features , pour indiquée par la suite est ce que le clients est bien satisfait de son achat, le temps de livraison est important ou non…..

- En suite, adapter les modèle de machine Learning pour la segmentation , et puis on va mesurer la similarité des clusters au fil de temps

olist

olist

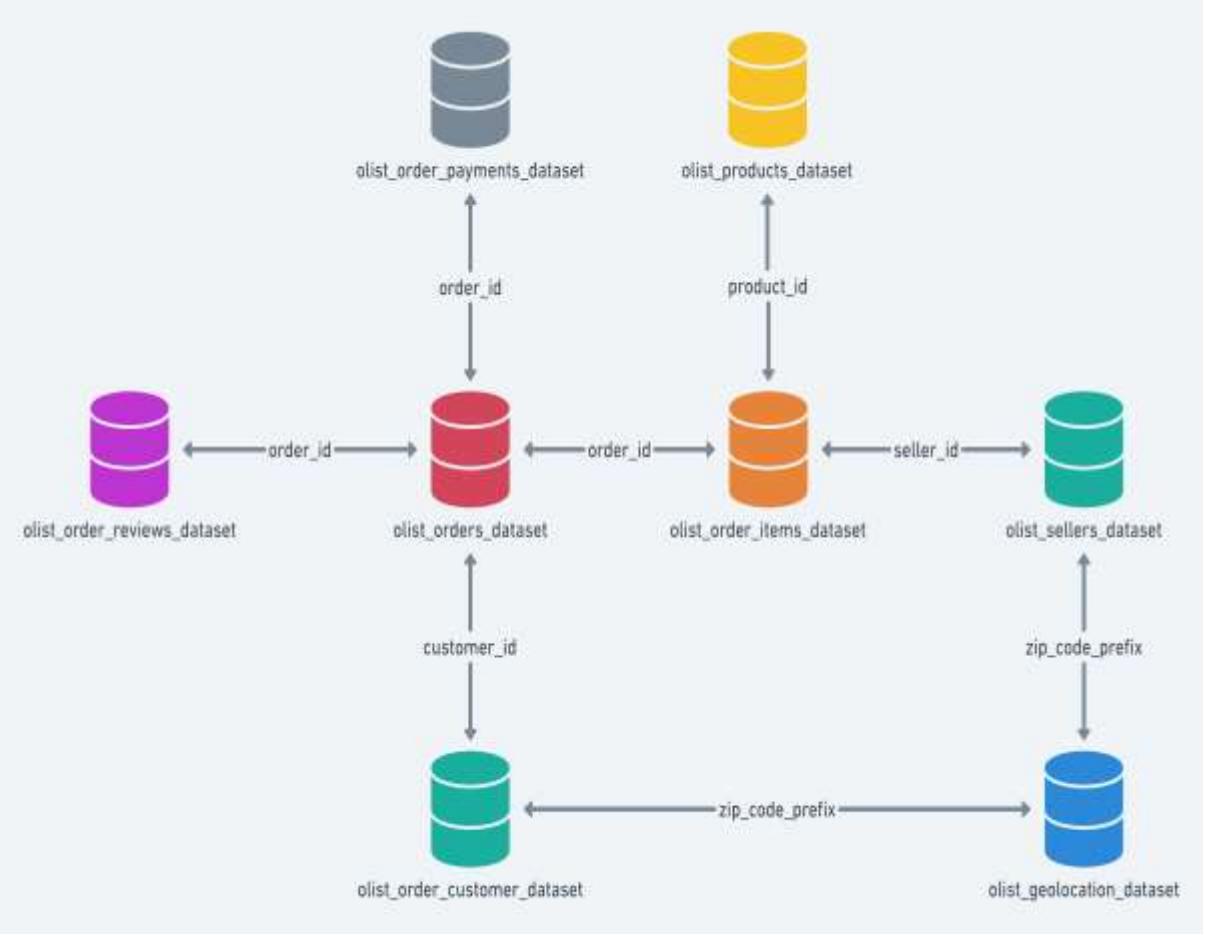
ANALYSES DES DONNEES

- Data cleanning.
- Feature engineering.
- Analyse exploratoire.

Structure de données:

Les données sont divisées en plusieurs ensembles de datasets:

- **Olist_customers** : Contient des informations sur le client et son emplacement
- **Olist_geolocalisation** : contient des informations sur les codes postaux brésiliens et ses coordonnées lat/lng
- **Olist_order_items** : Données sur les articles achetés dans chaque commande.
- **Orders_payments** : Données sur les options de paiement des commandes.
- **Olist_orders_reviews**: Données sur les avis rédigés par les clients
- **Orders_dataset** : données de base. De chaque commande
- **Olist_products** : Données sur les produits vendus par Olist



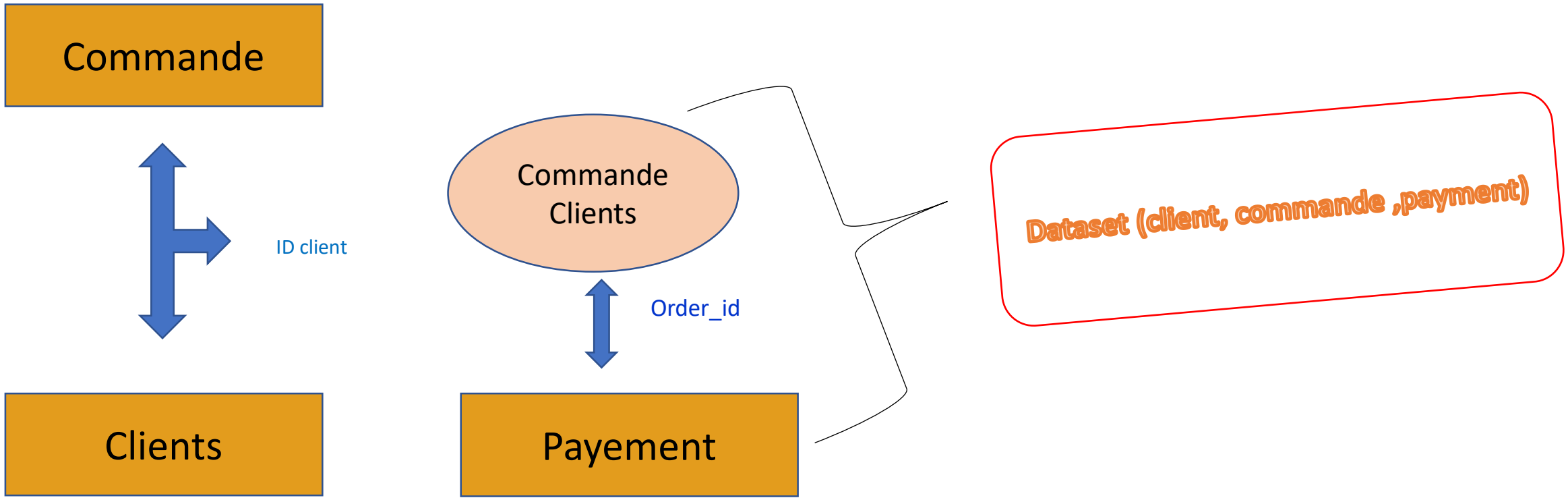
- **Olist_sellers** : Données sur les vendeurs qui ont exécuté les commandes passées chez Olist
- **Product_category_name_translation** : Traduit le product_category_name en anglais.

Nettoyage et regroupement des datasets:

Les données sont bien complètes et complémentaire entre eux,

1-Très peu de données manquantes 

2- Les données attendus ont été crée 



Création des colonnes importantes:

Récence : c'est la mesure de la période de temps écoulée depuis la dernière activité d'un clients.

Fréquence : cela fait référence au nombre total d'activités réalisée par un clients sur une période donnée. Donc c'est le nombre d'achats effectuées par un client.

Montant : Il s'agit du montant total dépensé par un client sur une période donnée, ou toute autre mesure monétaire liée a son activité.

Récence

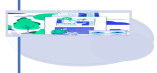
- Colonne:
- last_purchase_timestamp,
- Customer_unique_id
- Fonction: max()
- Date max – date d'achat

Fréquence

- Colonne : customer_unique_id, order_id
- Fonction : Groupby,
- Count : (order_id)

Montant

- Colonne:
- Customer_unique_id, payement_values
- Fonction : Groupby
- Sum : (payement_id)



Création des colonnes importantes:

- On va crée un des colonne en basant sur la clé unique de clients, pour tester la segmentation et puis choisir un meilleurs modèle.

Score : Ajouter le score données par chaque clients

Temps de livraison : Date de livraison - date d'achat de la commande

Délai de livraison : Date de la livraison pour le clients - Date estimée

Prix total d'une commande : Grouper par order_id et puis fais la somme de la colonne 'Price'.

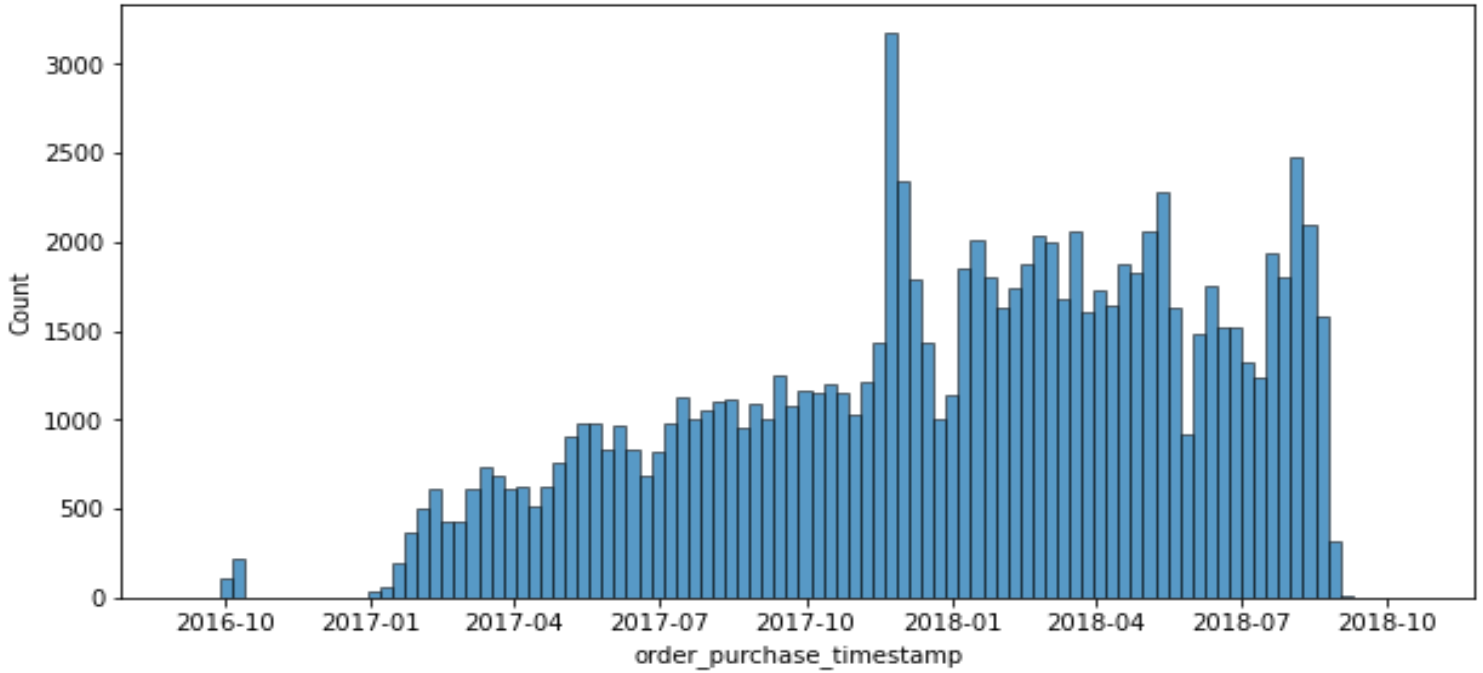
Poids et la taille des produits : Grouper par order_id et puis faire la somme de la colonne 'product_weight_g'

Volume des produits : appliquée la fonction lambda et puis faire la multiplication pour chaque ligne de ses colonnes 'product_length_cm ', 'product_height_cm' , 'product_width_cm'

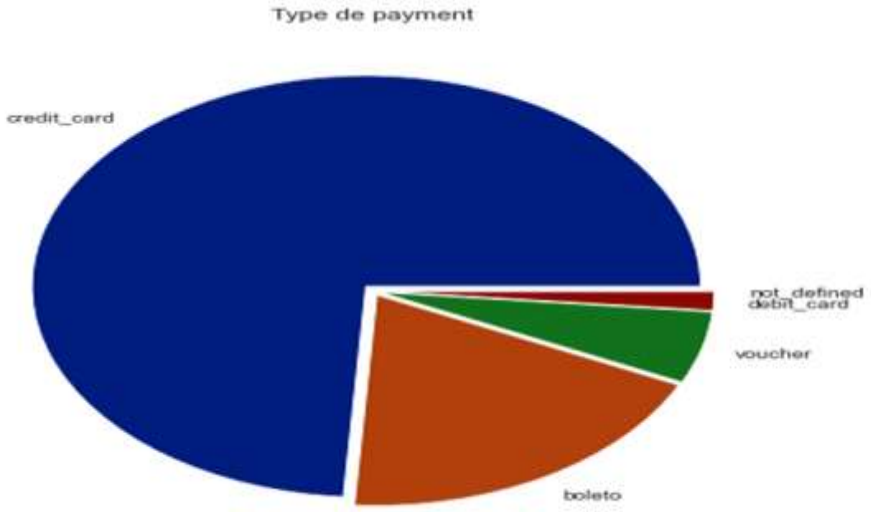
Les données manquantes : amputation avec la moyenne comme la colonne (score),
Utiliser StandardScaler pour le passage a l'échelle

- Features autre que RFM, trop de features , modèles difficiles à interprété.
- Pour un modèle facile à interpreter, modélisation avec un minimum de features.

Analyse exploratoire:

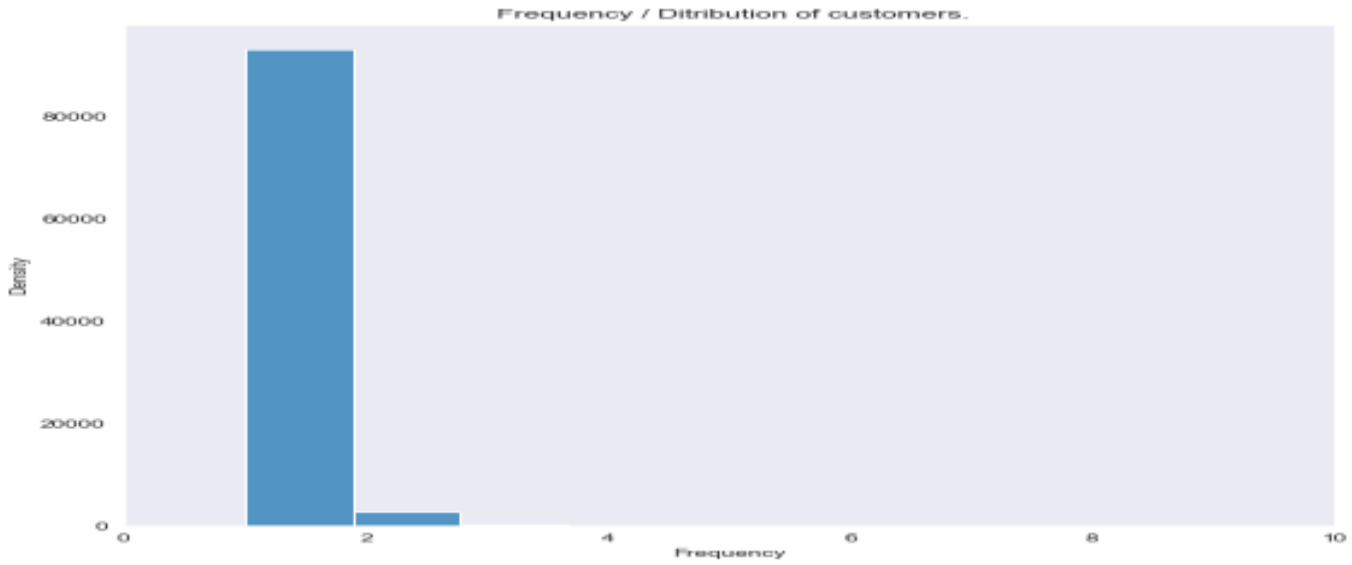


On remarque un pic anormal autour de Novembre / Décembre 2017 beaucoup de commande sont passée en cette période ,ainsi qu'un plateau sans commandes entre Octobre 2016 et Janvier 2017.

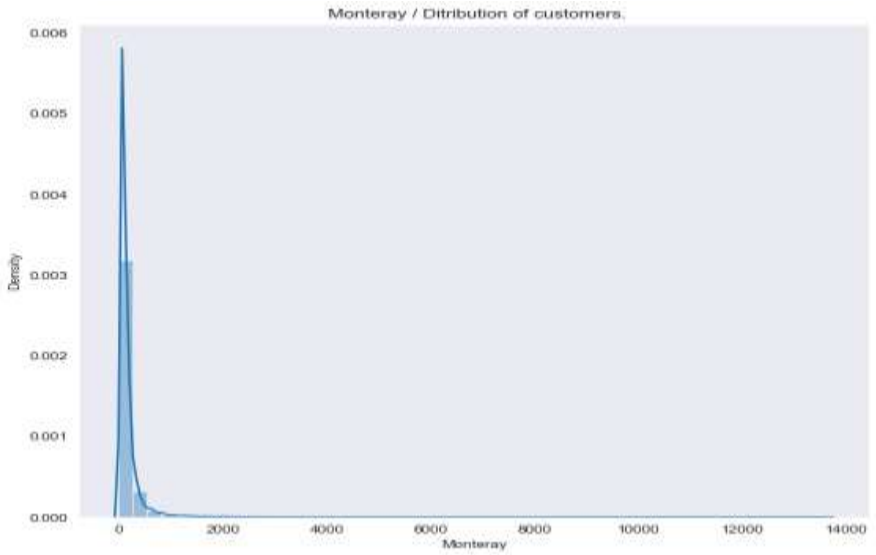


Ce graphique montre la diversification des options de paiement :
la majorité des commandes sont payées par carte de crédit (credit_card).
Peu de clients qui utilise le type débit_card

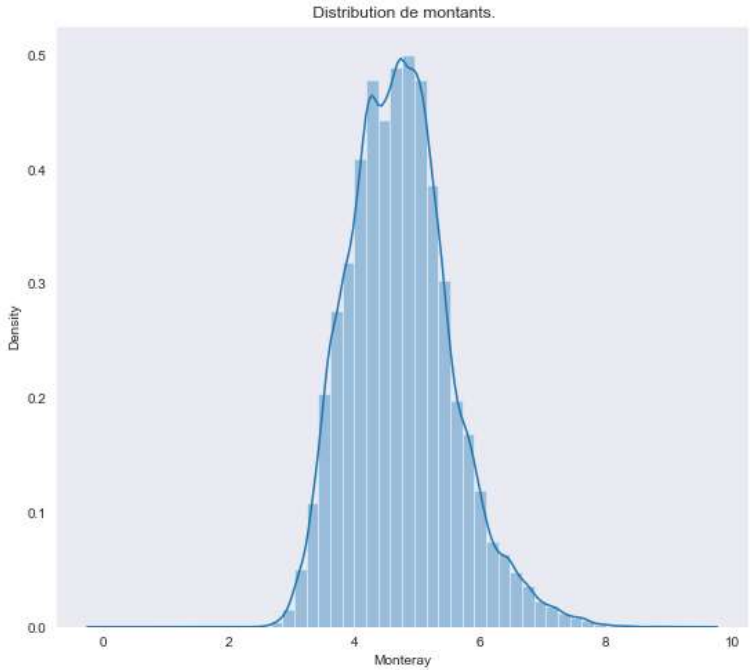
Analyse exploratoire:



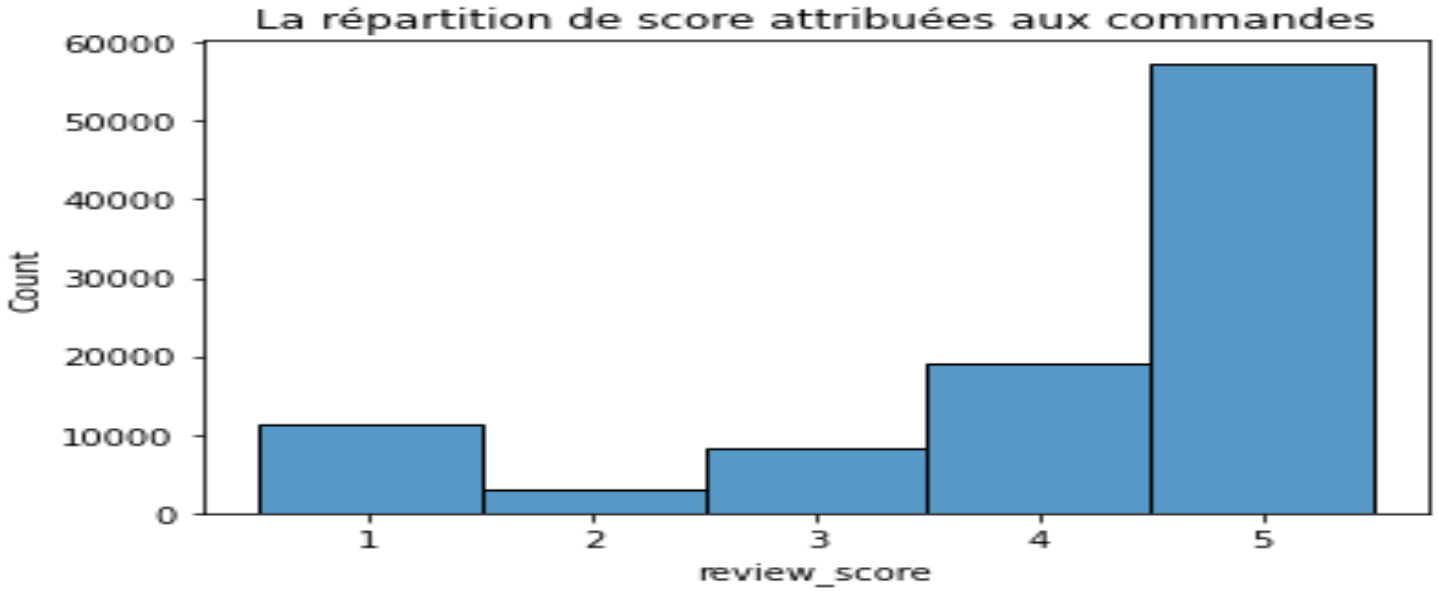
On remarque que 90% de clients n’a commandé qu’un seule fois ,une seule commande passée,



Une distribution anormal pour les Montant dépensé par les clients



Analyse exploratoire:

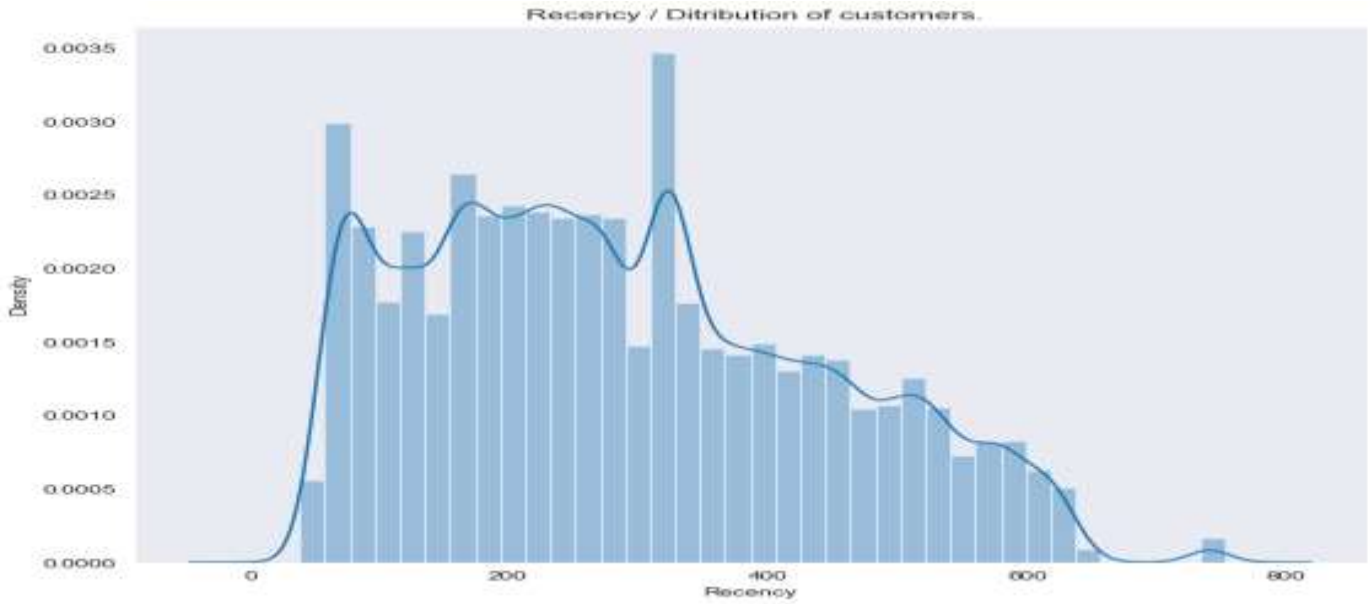


Ce graphique montre la répartition de review_score,

- le score de note 5 est attribué par un grand nombre de clients.
- un faible nombre de clients ayant donné des notes de 2.

Une distribution anormale pour la récence,

- La valeur moyenne de la récence se situe autour de 250 à 280 jours.
- un faible nombre de clients qui commandent au-delà de 600 jours.



MODÉLISATIONS

Analyse et modélisation:

Dans un premier temps , nous avons testé les différents modèles sur seulement les features (R,F,M).et puis au fur et à mesure on intègre features par features et on regarde quelle sont les données qui influence vraiment sur la segmentation.

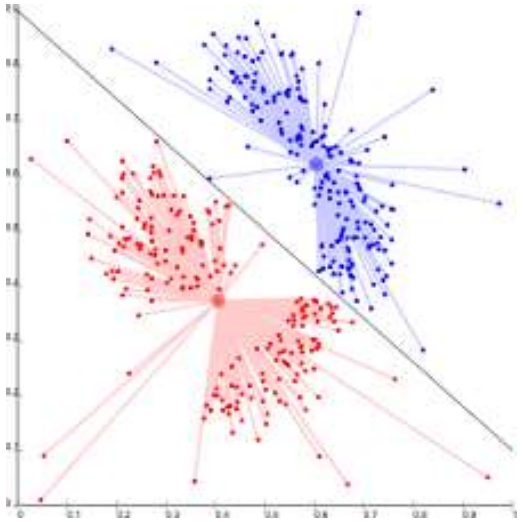
Modèles testé:

KMEANS:

Algorithme Kmeans est un algorithme itératif qui fonctionne en deux étapes:

1- **Affectations** des points au centre le **plus proche**.

2- **Déplacement** du centre à la **moyenne** du clusters

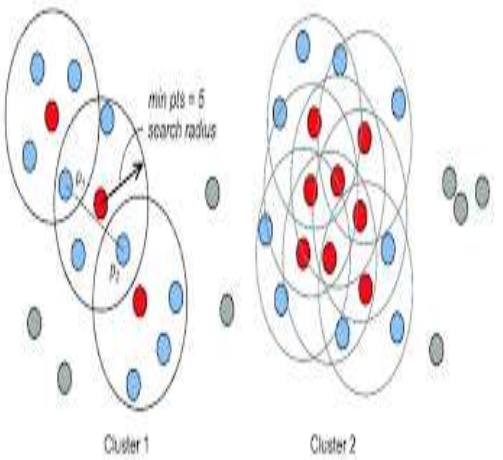


DBSCAN

Utilisé pour des clusters de haute densité.

il crée des clusters en fonction de paramètres **epsilon**, les **points min** et le **bruits**.

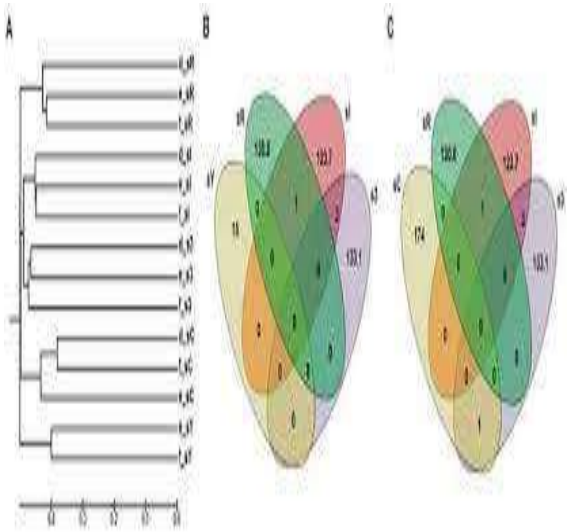
Prédit séparément les points centraux.



Hiérarchique:

Consiste à créer une **arborescence** de cluster .

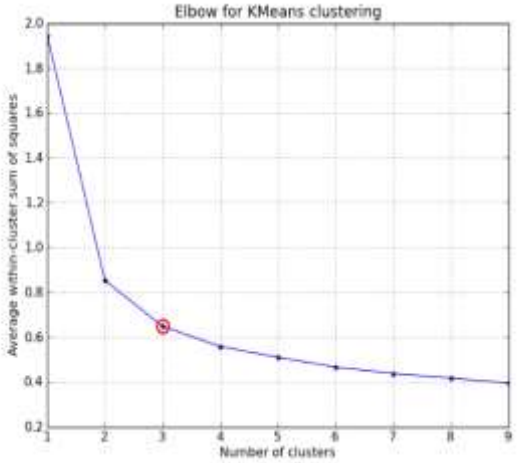
La classification hiérarchique est un algorithme déterministe, elle produira toujours le même dendrogramme.



Les métriques pour évaluer les clusters:

Ebow Method:

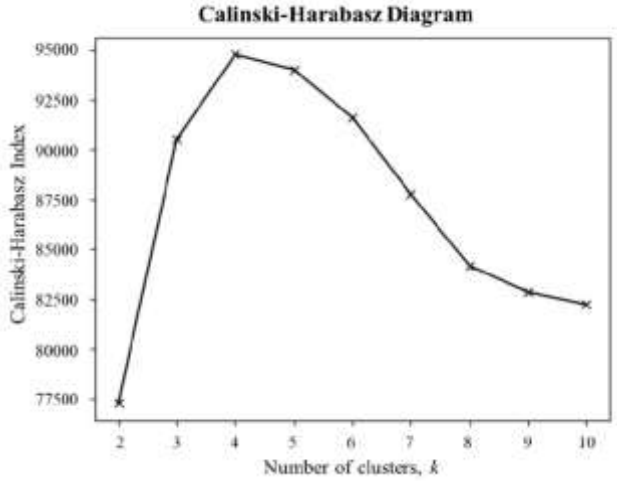
Permet de détecter une zone de coude dans la minimisation du cout (inertia_)



Calinski_Harabasz:

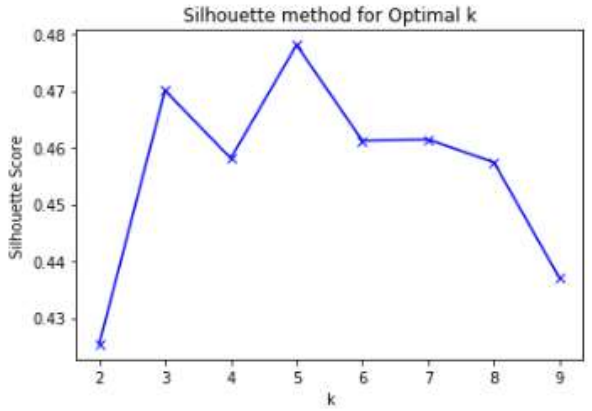
Mesure de distinction qui compare la dispersion entre les clusters avec la dispersion pour tous les clusters.

- Score plus élevé indique des clusters mieux définis



Silhouette:

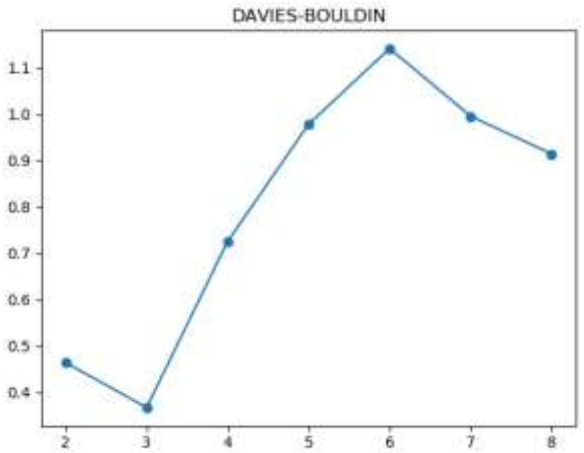
Mesure de distinction qui compare les distances au sein du cluster avec celles des clusters voisin les plus proches



Davies_Bouldin:

Mesure de distinction qui compare la distance entre les clusters avec la taille des clusters elle-même,

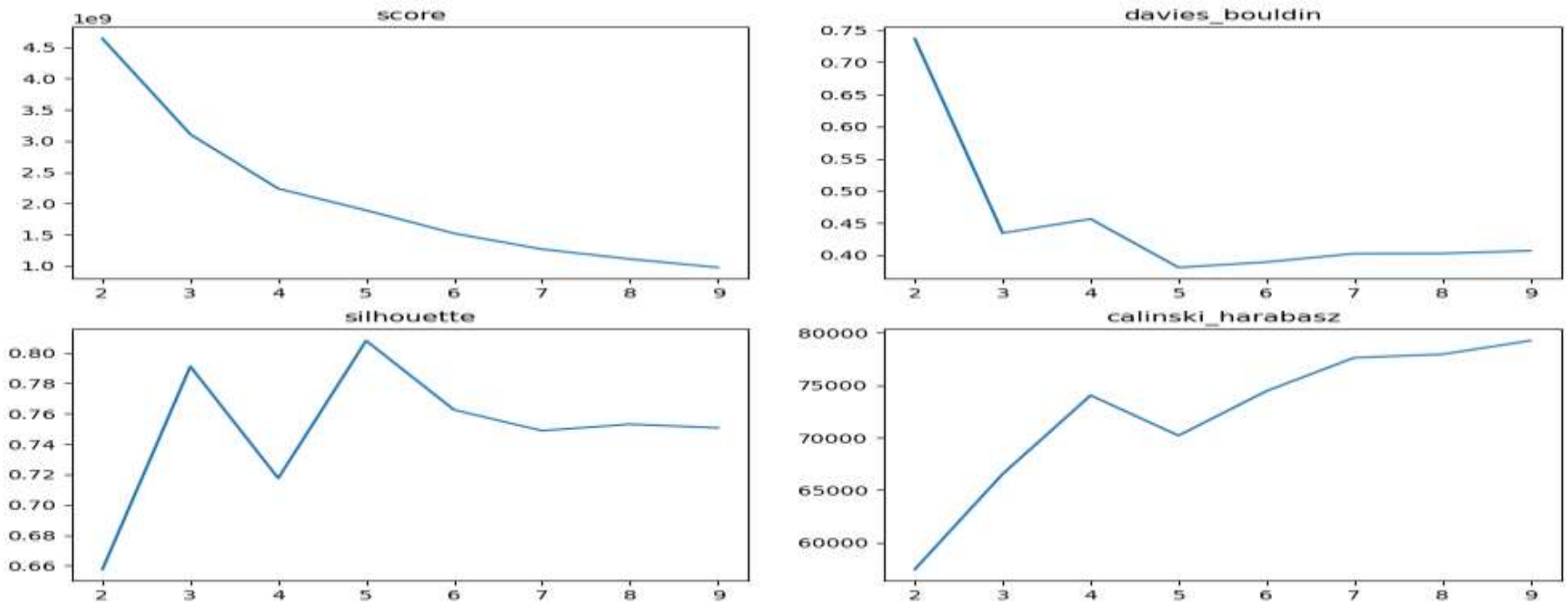
- Score plus faible indique une meilleure partition



MODÉLISATIONS DE RFM

Analyse sur les données RFM:

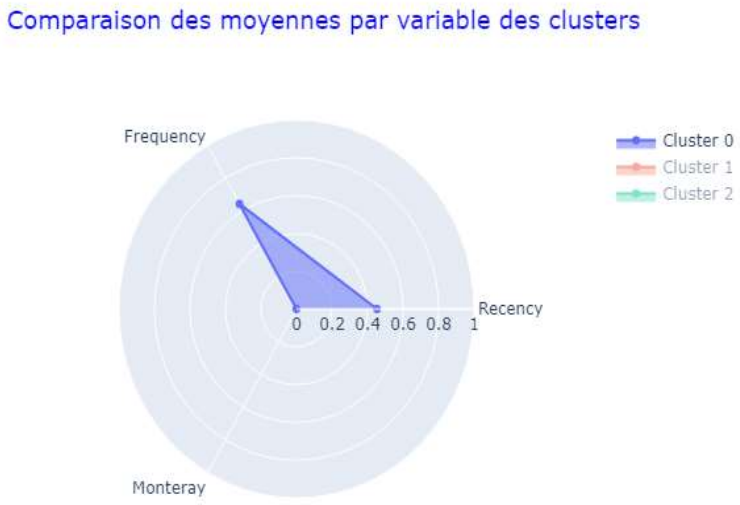
Détermination de nombre de clusters:

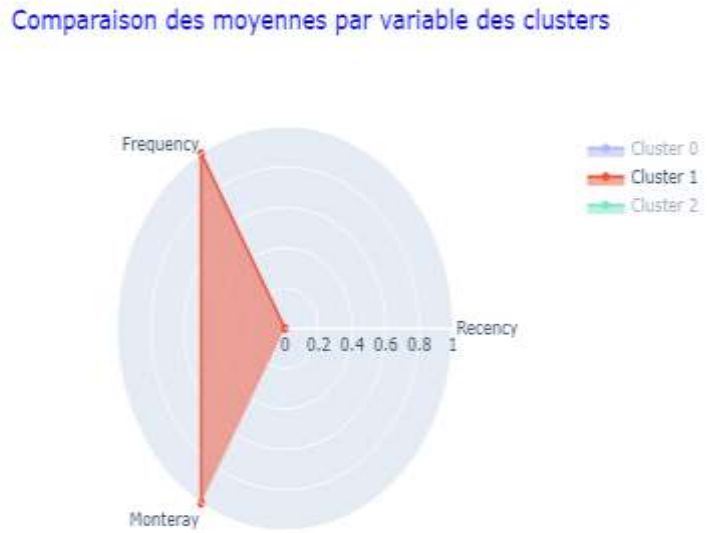


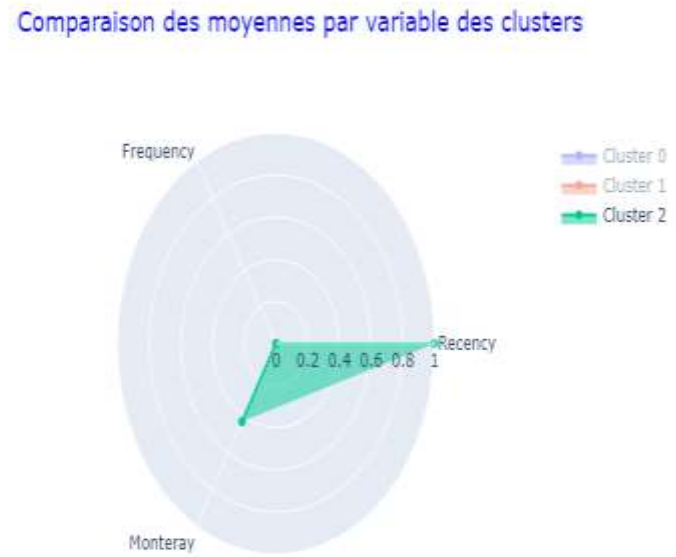
Nombre de clusters déterminé par la méthode de coude et davies_bouldin nous montre que le nombre de clusters idéal est k =3

Interprétation des clusters:

KMEANS







Clusters 0 :
Des clients qui commande moyennement sur une période donnée et qui ont réalisé des achats moyennement récentes

Clusters 1 :
Ce sont des clients qui passent beaucoup de commande et qui paye beaucoup d'argents sur leur achats

Clusters 2 :
Des clients qui ont réaliser des commandes très récente avec un montant moyen,

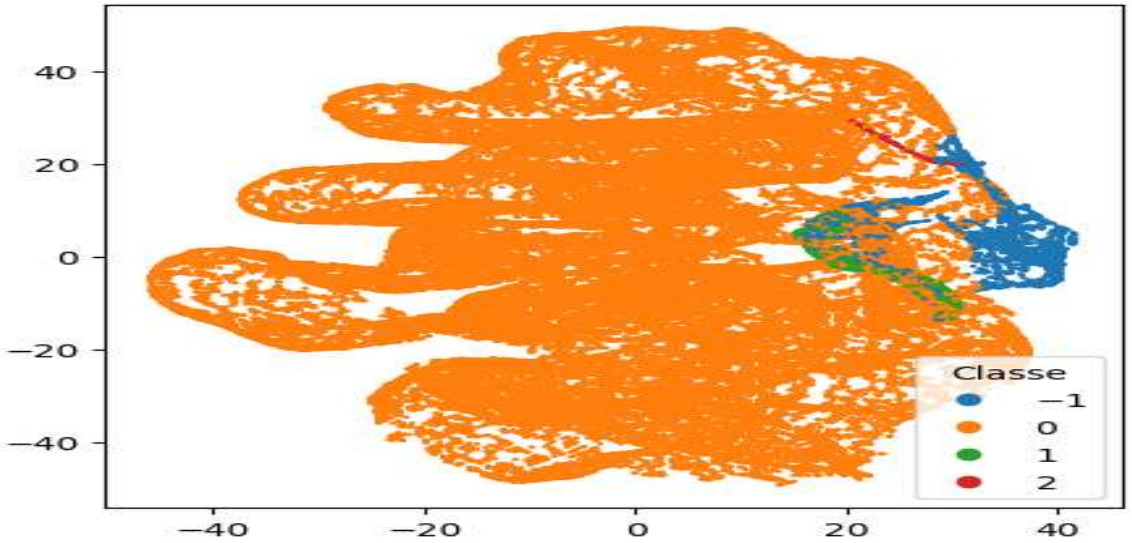
Interprétation des clusters:

DBSCAN pour RFM:

- Gros déséquilibre des clusters avec DBSCAN:

- Clusters 0 contient la plus grand partie des clients environ(93%).
- Une partie bleu n'appartient à aucun clusters.

➢ DBSCAN donc n'est pas approprié pour ce clustering

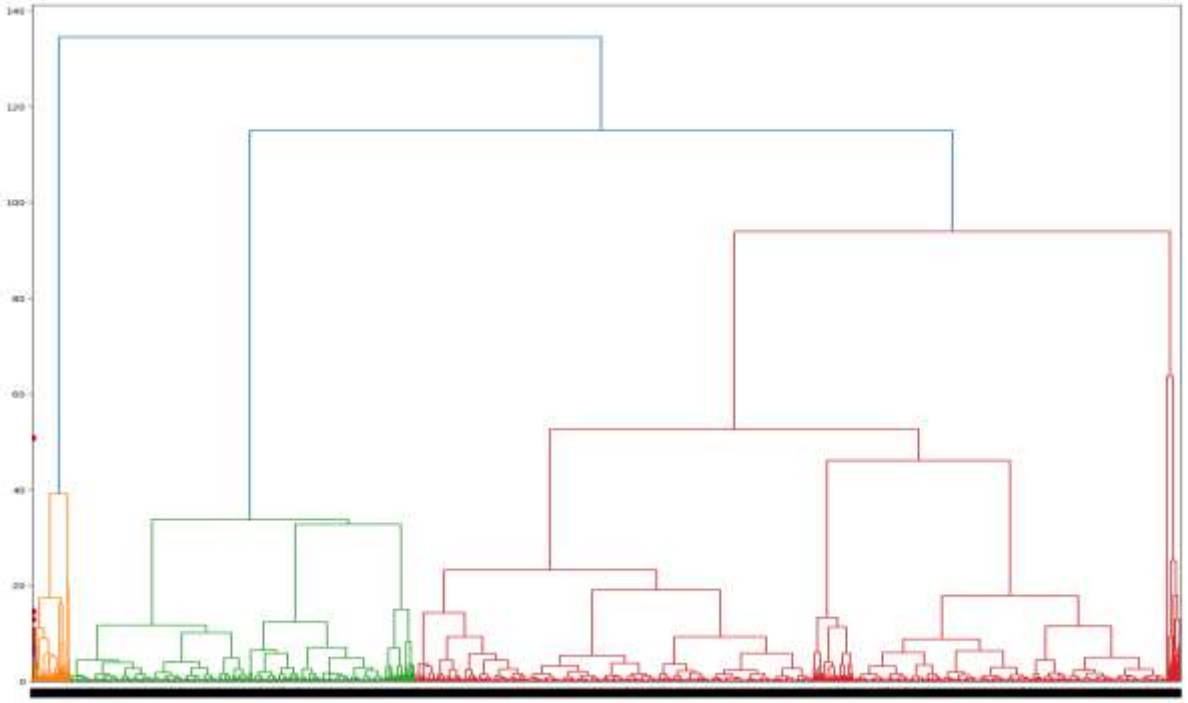


Clustering hiérarchique pour RFM:

Ce graphique montre la répartition des clusters dans l'arbre de clustering hiérarchique:

une concentration plus élevée de points similaires dans cette zone rouge.

Clustering hiérarchique donne les même résultats que KMEANS contrairement au DBSCAN



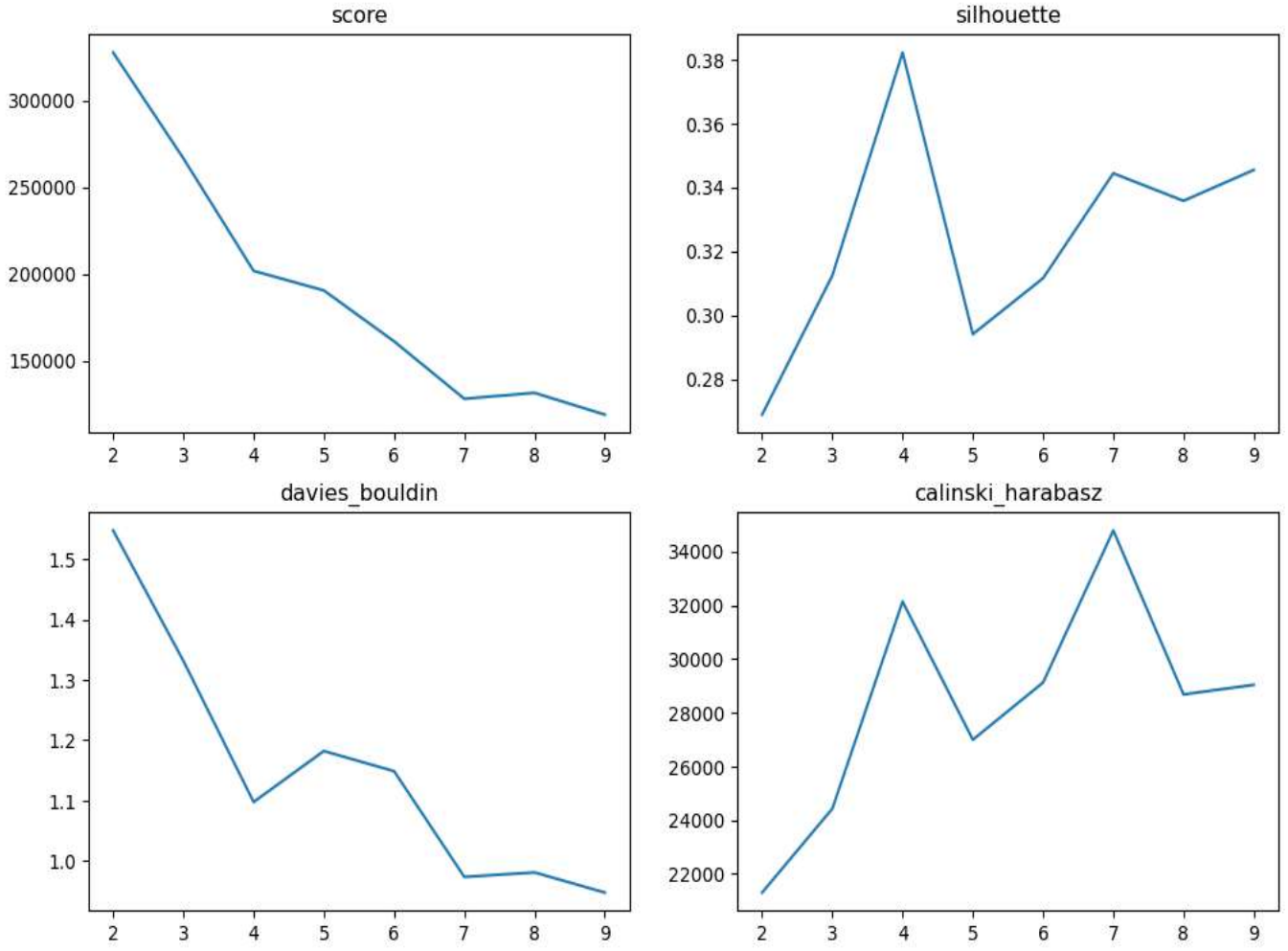
MODÉLISATIONS DE RFM + SCORE

Détermination de nombre de clusters:

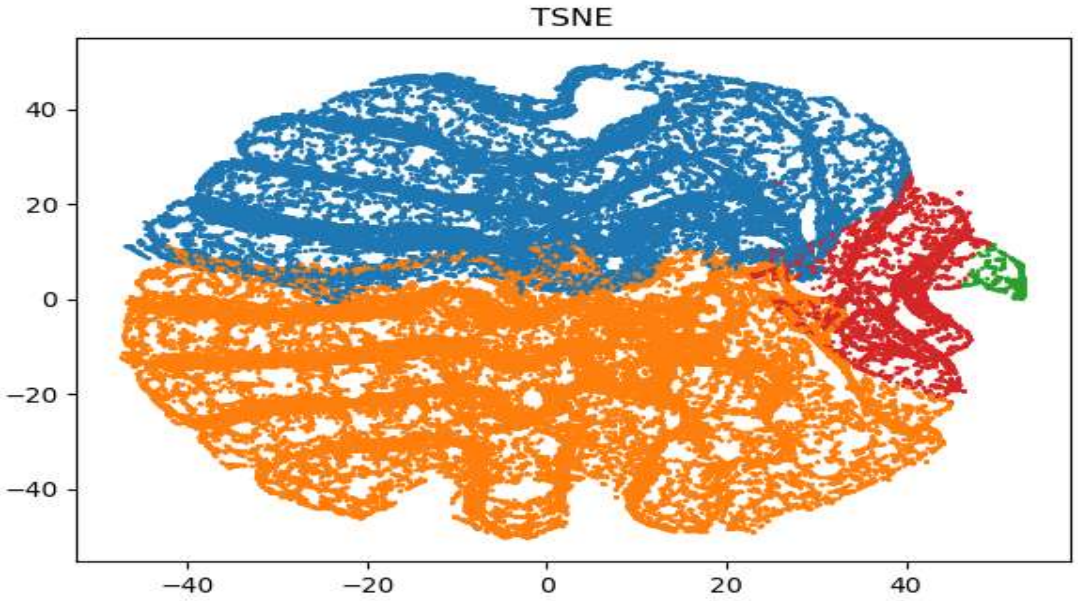
- Pour la réalisation de clustering pour les différents
- donnée, j'ai décidé de continuer avec le KMEANS

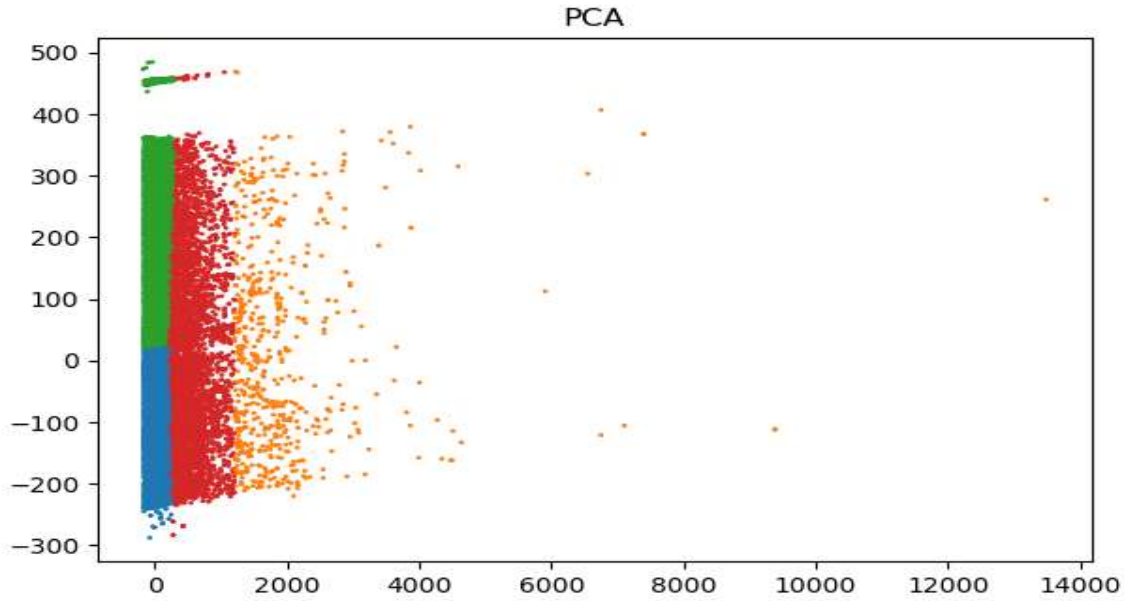
Pour les donnée RFM + review_score:

- Nombre de clusters déterminé par la méthode de coude et davies_bouldin et silhouette nous montre que le nombre de clusters idéal est k =7 ou 4.
- J'ai décidé de prendre le k = 4 pour une bonne interprétation.



Visualisation des clusters en 2 D:



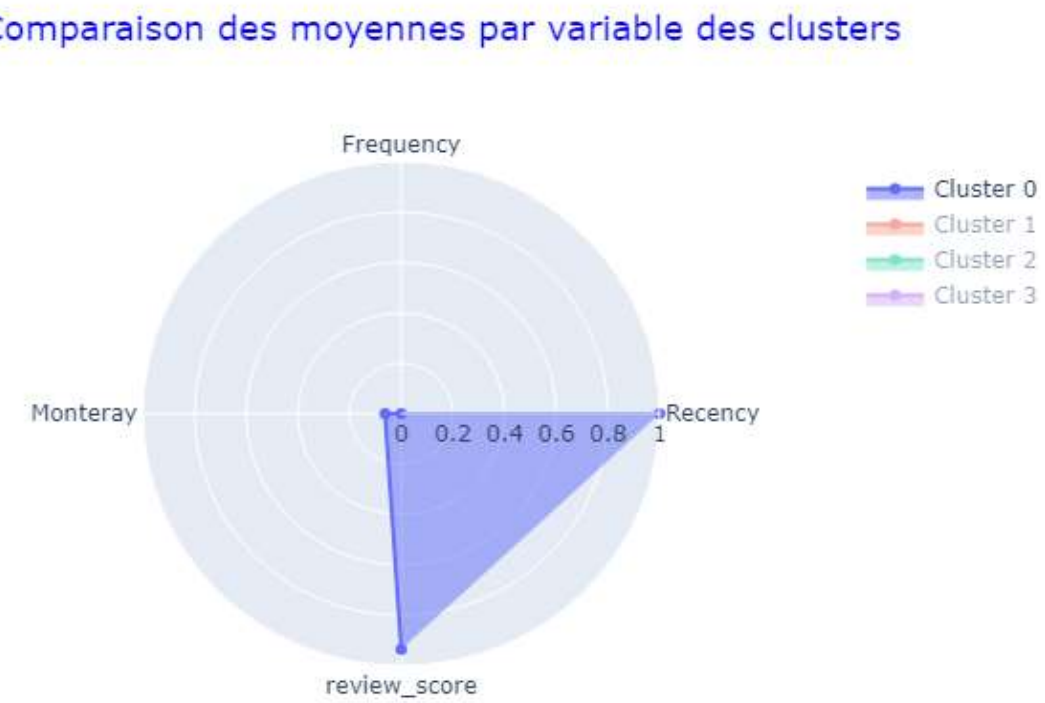


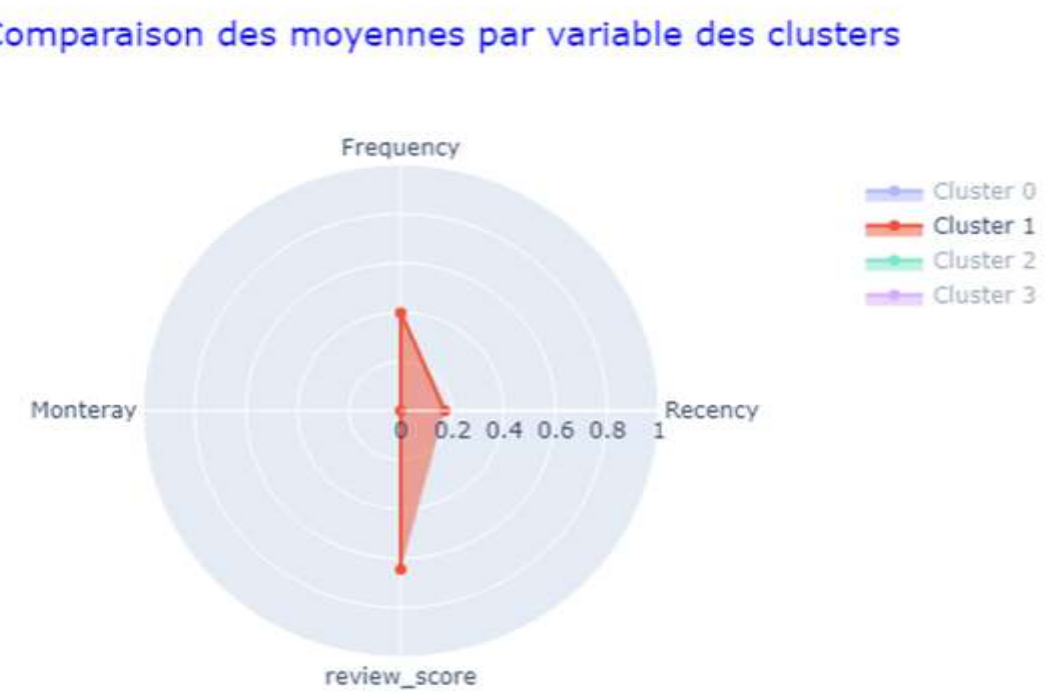
La visualisation en 2 dimension avec TSNE ET PCA :

- 4 clusters qui sont différents de forme et très variée.
- Pour TSNE : le clusters bleu et orange ont même forme et ils sont bien apparaitre.
- Peu de clients dans le clusters vert , le clusters rouge important par rapport au clusters vert.

- Une similarité entre le clusters rouge et orange , et on vois que le bleu et le vert sur le même axe

Interprétation des clusters:

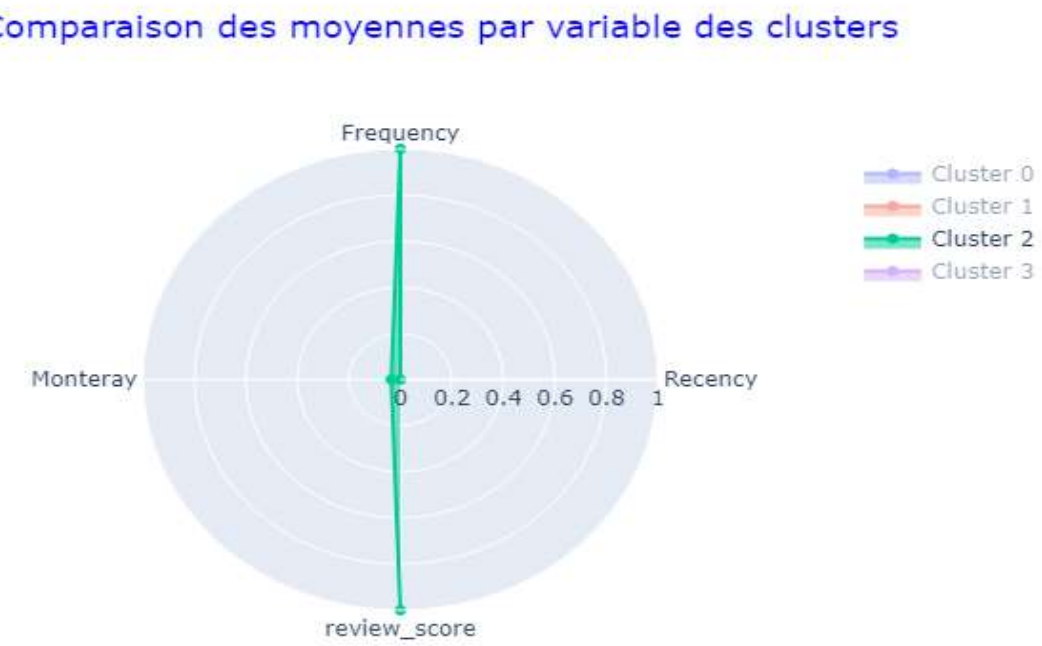


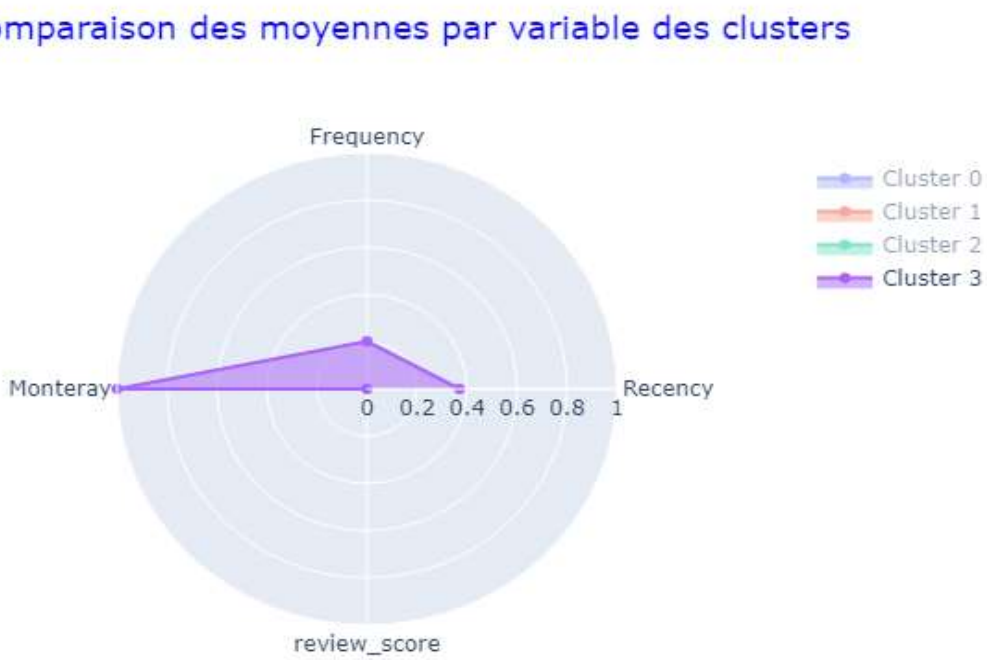


Clusters 0 :C’est des clients qui ont réaliser des commandes très récente et qui sont très satisfait de leur achats

Clusters 1 : c’est des clients qui achéte rarement et qui ont réalisés des achats moins récente et qui sont moyennement satisfait de leurs achats.

Interprétation des clusters:





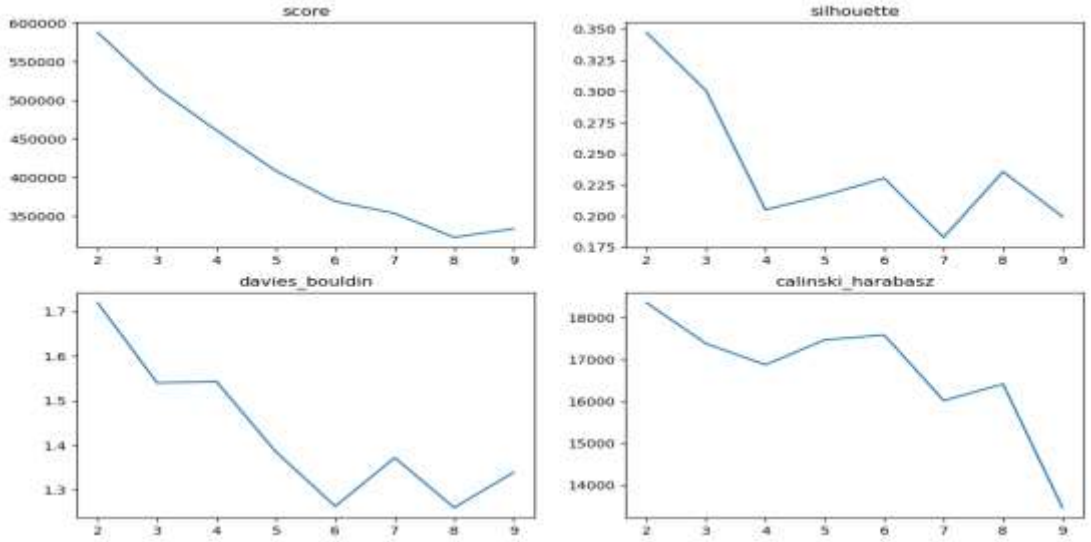
Clusters 2 : C'est des clients achètent beaucoup et sont beaucoup satisfait de leurs achats

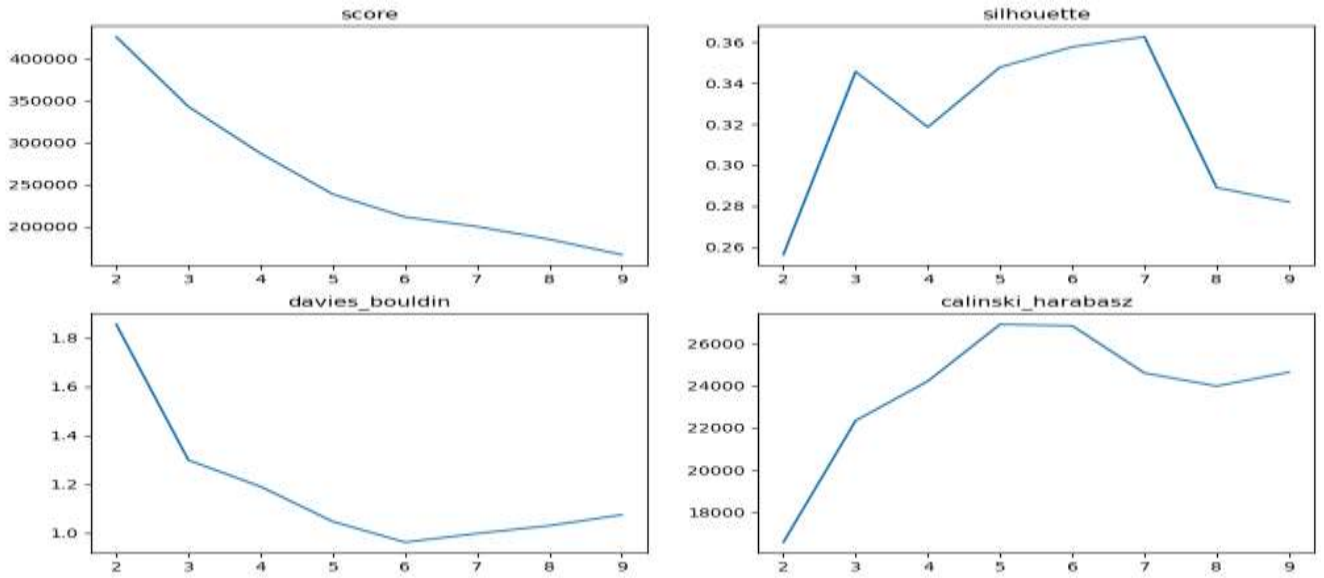
Clusters 3 : Des clients qui ont dépensé beaucoup d'argents sur leurs commande, c'est clients achètent rarement et ils sont pas satisfait de leurs achats.

MODÉLISATIONS DE RFM,SCORE,POIDS,VOLUME

Analyse des données:
Nombres de clustres:

[Recence, Frequence, Montant, score +poids de la commande , le volume de la commande]





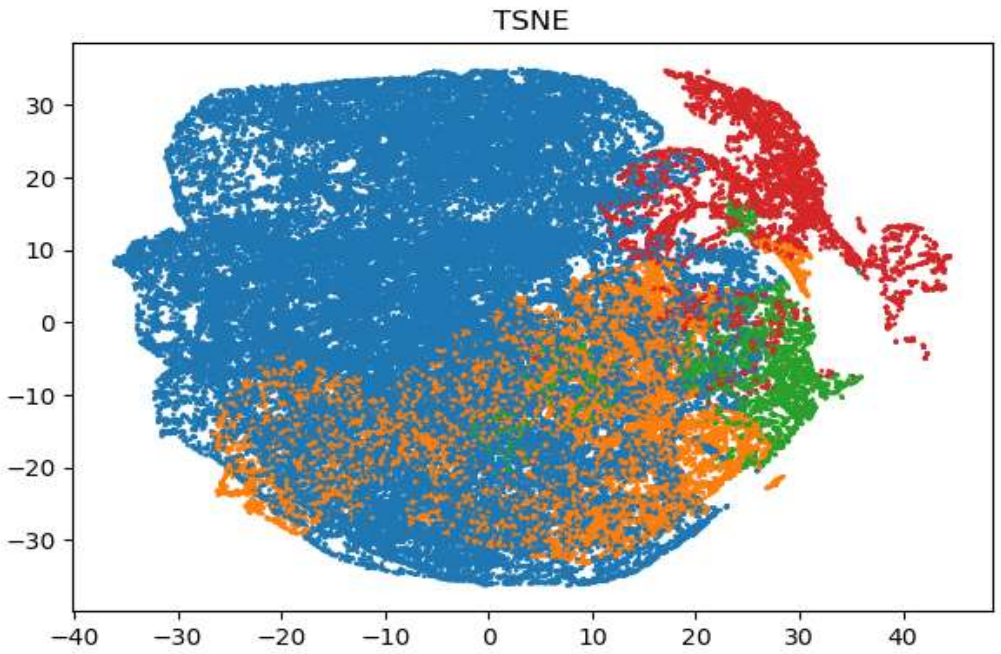
['Recency','Frequency','Monteray','review_score','geolocation_lat', 'geolocation_lng','delivry_time']

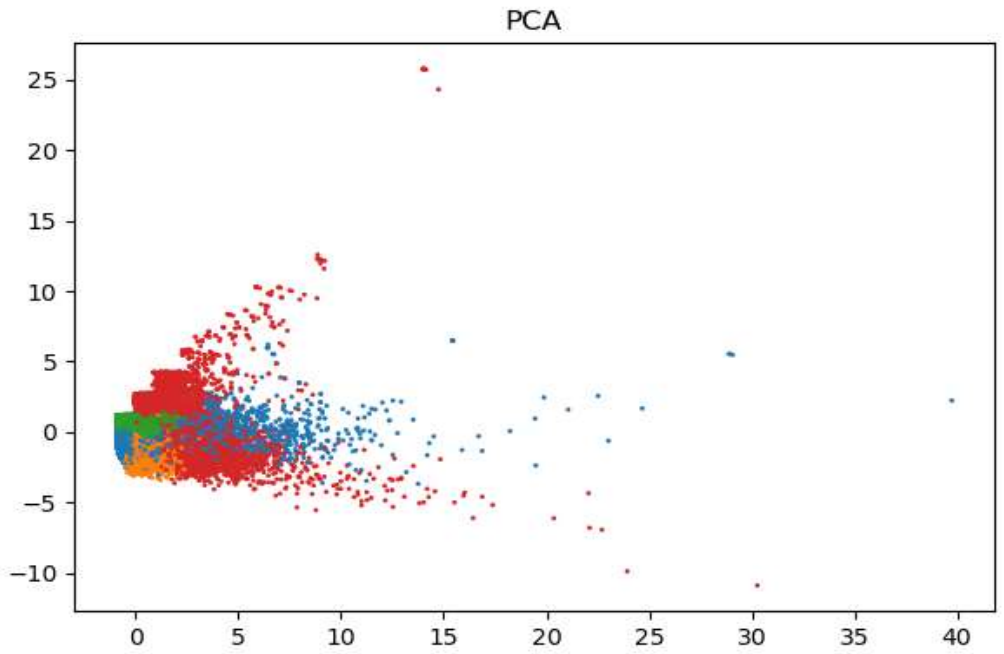
Nombre de clusters déterminé par la méthode de coude et davies_bouldin et silhouette nous montre que:
le nombre de clusters idéal est k =6 ou 8.

- J'ai décidé de prendre le k = 6 pour les deux dataframe pour une bonne interprétation.

Visualisation des clusters en 2D:

RFM, score +poids de la commande , le volume de la commande





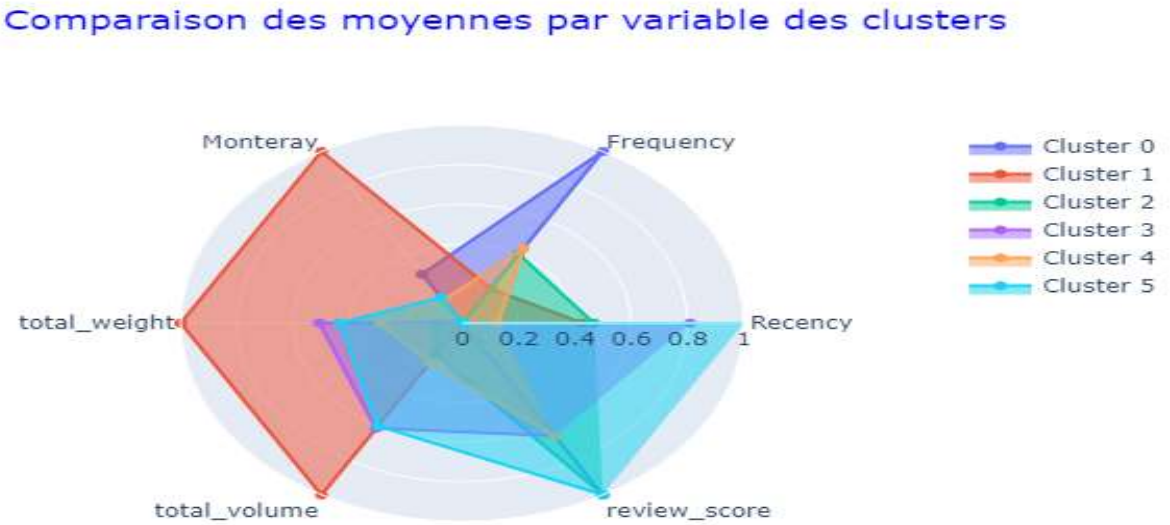
La majorité de clients sont dans le clusters bleu,

- On vois la partie de clients qui sont dans le clusters rouge on des caractéristiques différentes ils est éloigné des autres clusters

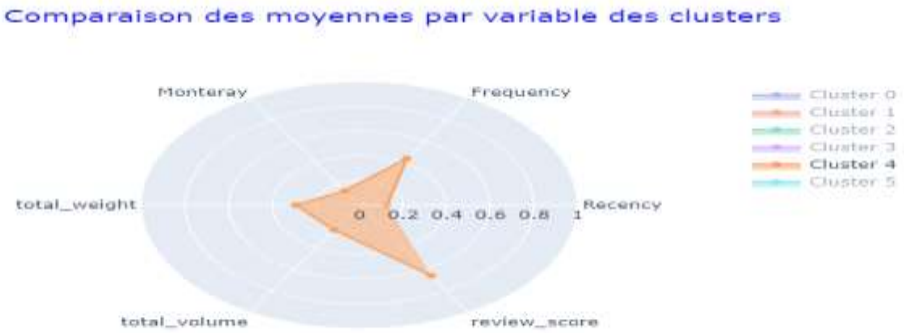
- Le clusters orange superpose sur le bleu , c'est des clients qui ont les même caractéristique,
- On va également que orange superpose sur le vert.

Interprétation des clusters :

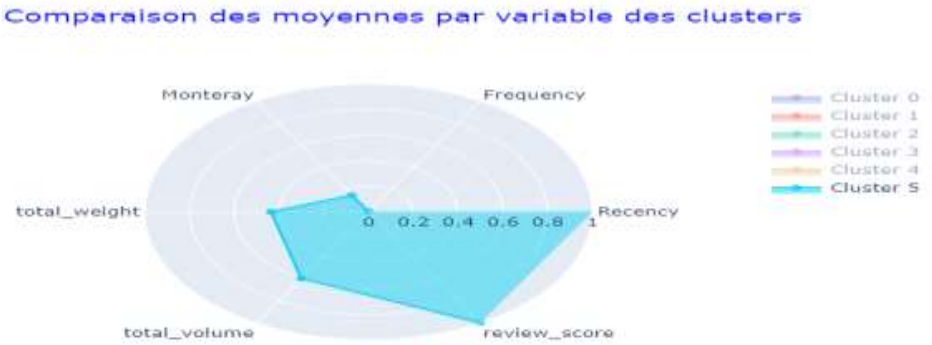
.



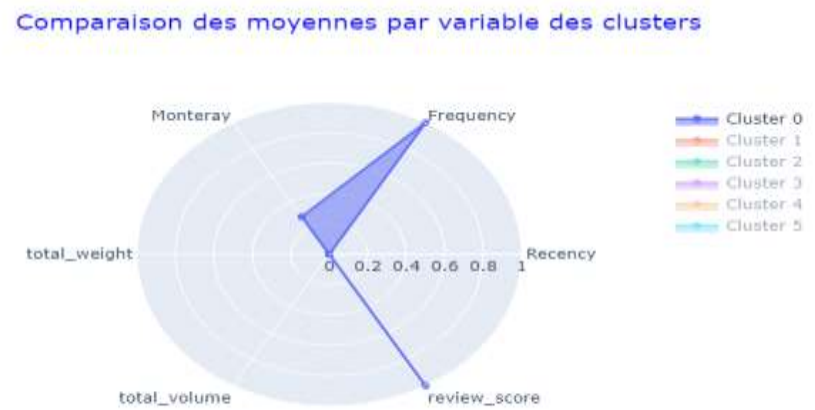
Clusters 4 :clients moyennement satisfait poids et volume faible, commande pas récente



Clusters 5 : clients qui ont réaliser des commandes très récentes et qui sont satisfait , et moyennement en terme de poids et volume



Visualisation avec radar plot:



Clusters 0 :
clients qui passent beaucoup de commande et très satisfait

Clusters 1 :
clients qui achète moins, leurs commande contient des achats très volumineux et bcp de poids et qui gaspille bcp d'argents,

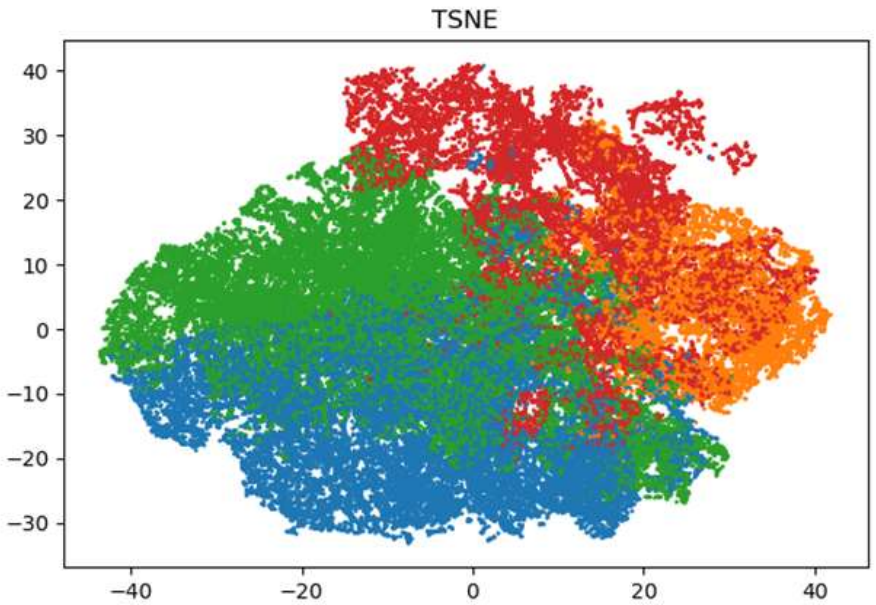
Clusters 2:clients
qui commande en moyen
Des commandes
pas très récente et qui
sont beaucoup satisfait

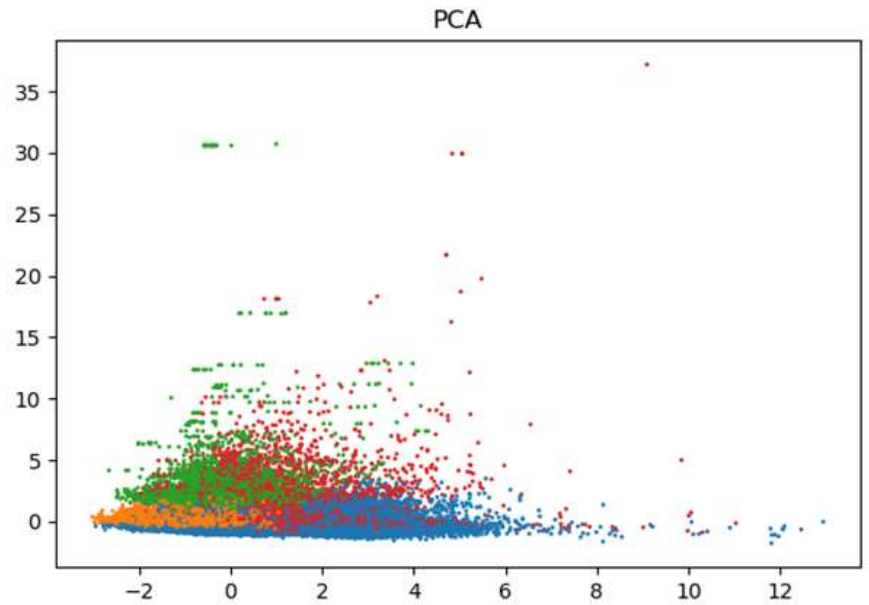
Clusters 3 :
clients qui commande rarement qui sont moyennement satisfait , des commande moyennent récente , poids et volume moyens

MODÉLISATIONS DE

RFM, SCORE, GEOLOCATION_LNG , GEOLOCATION_LAT , DELEVRY_TIME

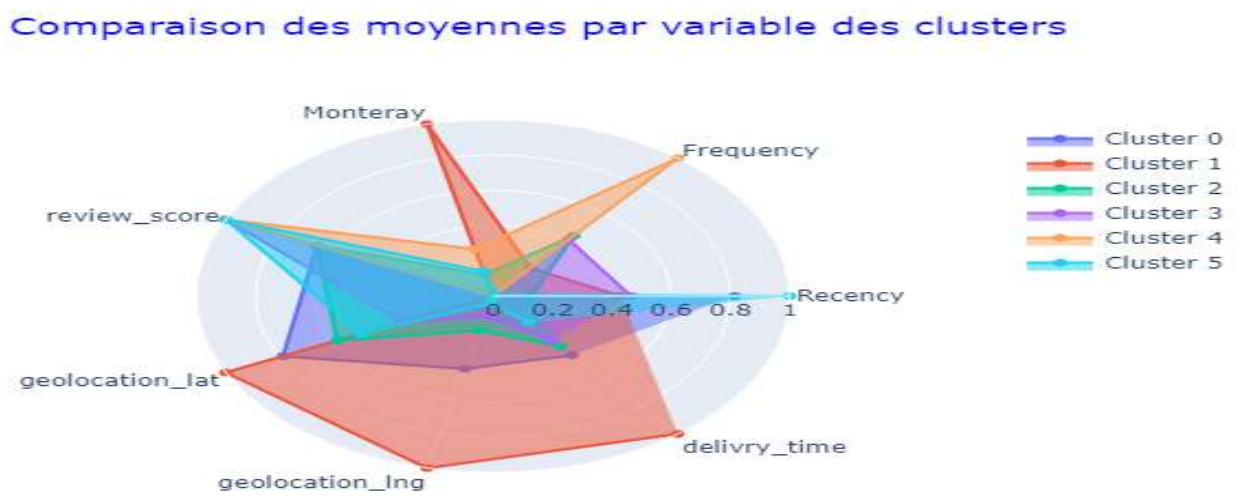
Visualisation on 2D:

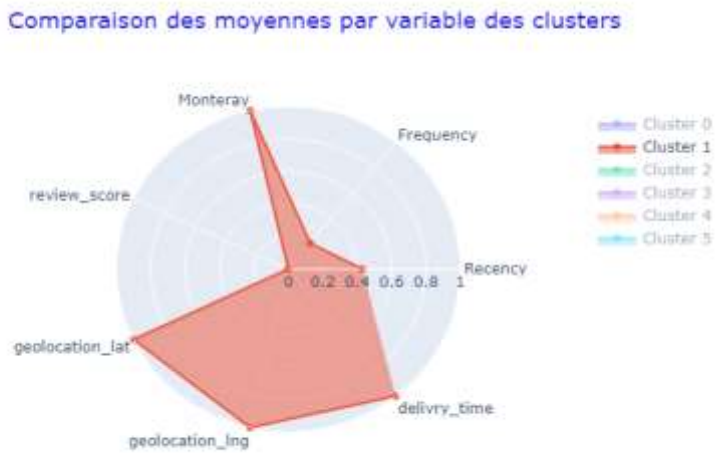




On vois 6 clusters qui apparaitre, peu de clients dans le clusters rouge et orange ,
et l'intégralité en bleu et vert , une partie de clusters verts superpose sur le bleu,

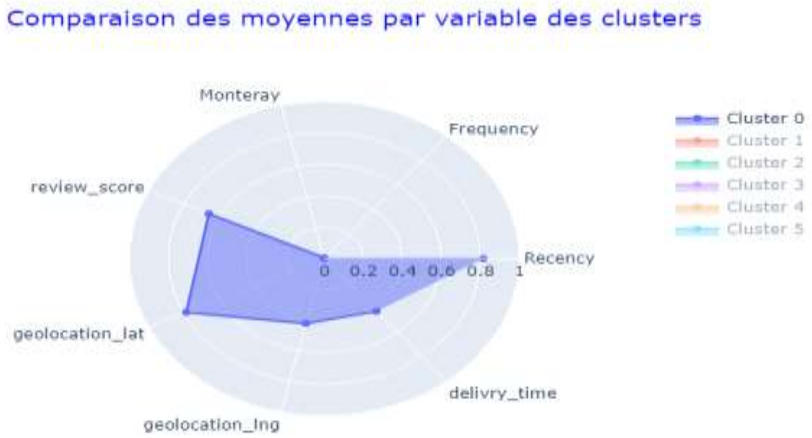
Visualisation avec radar plot:





Clusters0:

Clients moyennement satisfait,
cours temps de livraison
et proche géographiquement,
leurs commandes et très récentes

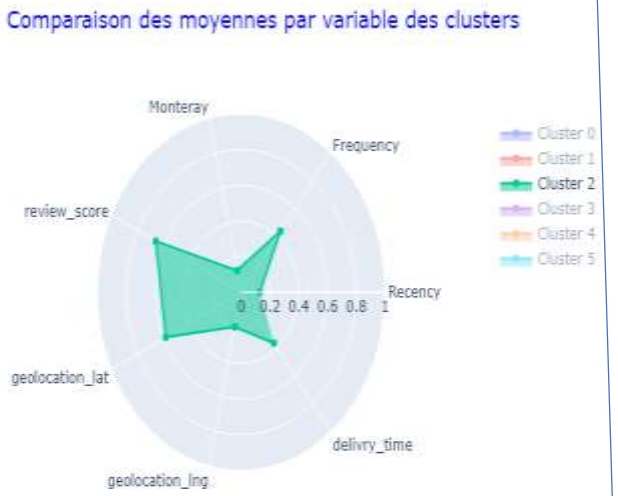


Clusters1:

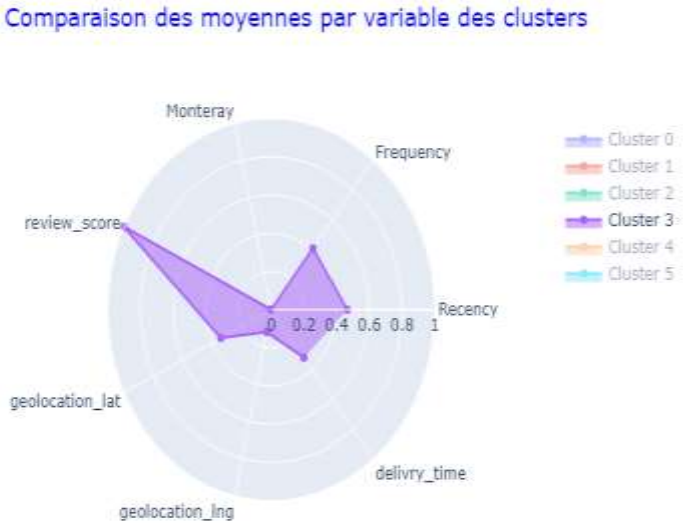
Clients qui commande rarement et paye un grand montant et géographiquement sont très éloigné et le temps de livraison très important et ils ne sont pas satisfait de leurs achats, Commande pas très récente

Visualisation avec radar plot:

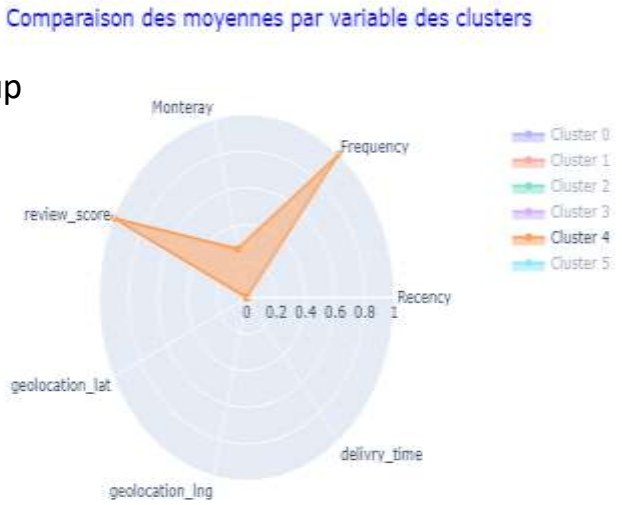
Clusters 2:
clients moyennement satisfait, cours temps de Livraison et proche géographiquement



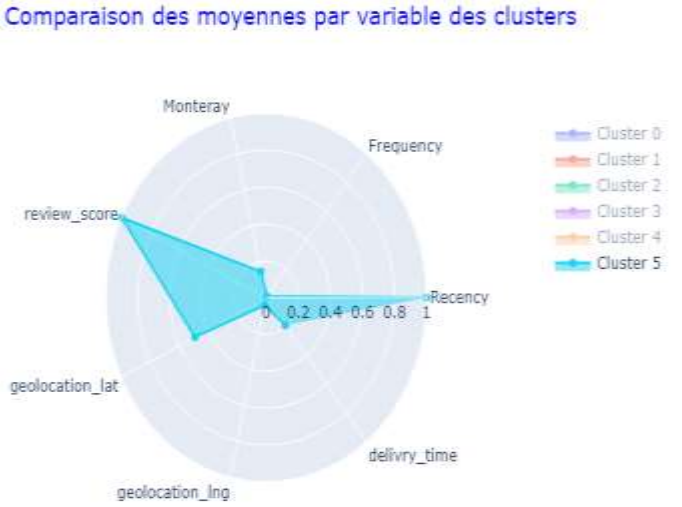
Clusters3:
Clients qui passe des commandes moyenne, et géographiquement sont très éloigné et le temps de livraison très important et bien satisfait de leurs achats,



Clusters4:
Clients qui commande beaucoup très satisfait de
leurs achats et
Proche géographiquement
Livraison rapide, commande non récente



Clusters 5 :
clients très satisfait, qui réalise des commande très récente , le temps de livraison très rapide et très proche géographiquement , Le montant pour les commande est faible



CONTRAT DE MAINTENANCE
STABILITÉ TEMPORELLE DES CLUSTERS AVEC KMEANS

Contrat de maintenance:

L'objectif est de vérifier à quel moment les clients changent de clusters.
On propose un contrat de maintenance basé sur une analyse de la stabilité des segments au fil du temps.
* Je réalise la stabilité avec les données RFM+review_score.

Méthodologie:

- Définition d'un modèle de clustering sur les commande plus ancienne(04/09/2016 à 31/08/2017)

calcul de ARI:

- **Phase initiale:**

Période de référence : septembre 2017

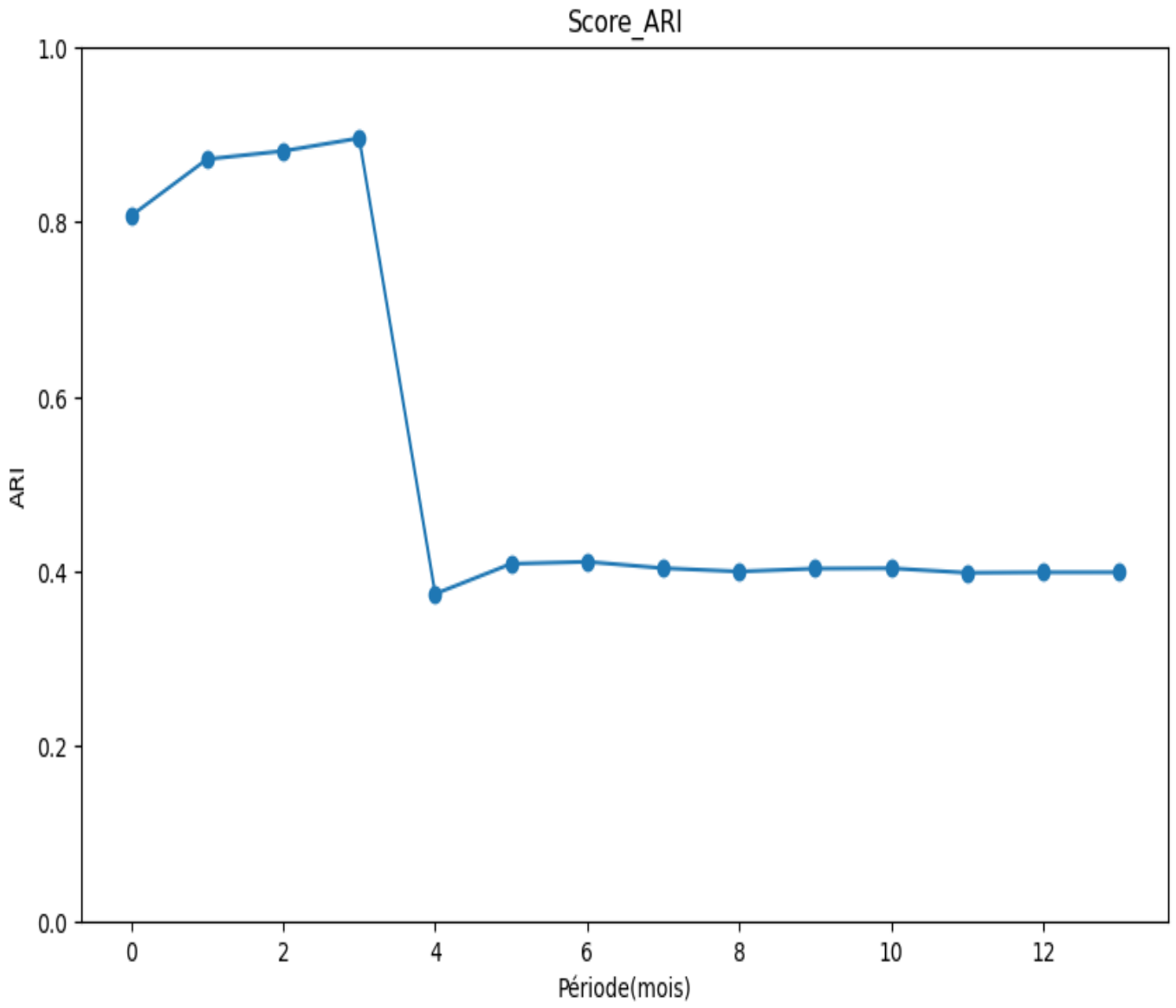
- **Période de Glissement:**

Glissement consiste à décaler la période d'analyse sur 12 mois à partir de la phase initiale .
Définitions des clusters à nouveau à partir de ces glissés.

- **Prédiction et comparaison:**

Prédiction sur ancienne modèle (04 /09/2016 à 31/septembre 2017).
Comparaison les deux clusters avec l'ARI.

Évolution de la Similarité ARI et Maintenance de la Segmentation:



la courbe des score ARI commence à subit une inflexion ,nette à partir de 3 mois.

La mesure de similarité ARI :

- révèle une diminution significative des scores, passant de 0,93 à 0,33. Cela indique que les clients changent de clusters

Conséquence :

- Il est important de noter que ce changement dans la similarité ARI suggère une évolution des préférences et des comportements des clients au fil du temps.

Maintenance de la segmentation :

- Pour maintenir une segmentation précise et pertinente, il est recommandé de prévoir la maintenance de la segmentation tous les 3 mois. Cela permettra d'ajuster les clusters en fonction des nouvelles tendances et des changements de comportement des clients.

Conclusion:

- L'utilisation de l'algorithme K-means pour la segmentation des clients sur les données Olist a été une étape importante pour mieux comprendre et servir nos clients. En exploitant les informations fournies par cet algorithme, nous pourrons mettre en œuvre des stratégies de marketing plus ciblées, renforcer la satisfaction client.

- Nous avons intégré avec succès différentes données pour la segmentation de nos clients. Cependant, une exception subsiste concernant la catégorie des produits. Avec 73 catégories distinctes, il devient crucial de repenser notre approche pour affiner nos modèles.

- L'analyse des scores ARI souligne l'importance de surveiller et de mettre à jour régulièrement la segmentation client.

La maintenance tous les 3 mois garantira une meilleure adéquation entre les clusters et les besoins changeants des clients.

olist

empowering commerce

MERCI POUR VOTRE ATTENTION