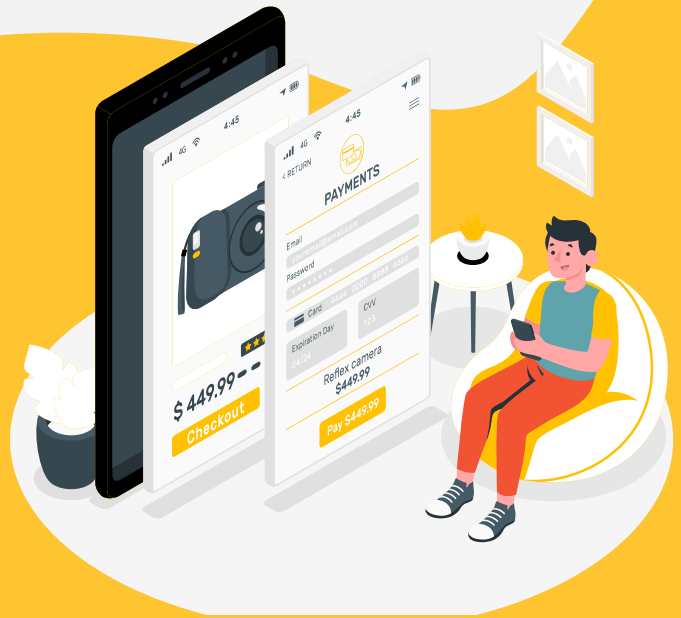
Parcours « Data Scientist »

PROJET 5:

Segmentation des clients d'un site e-commerce

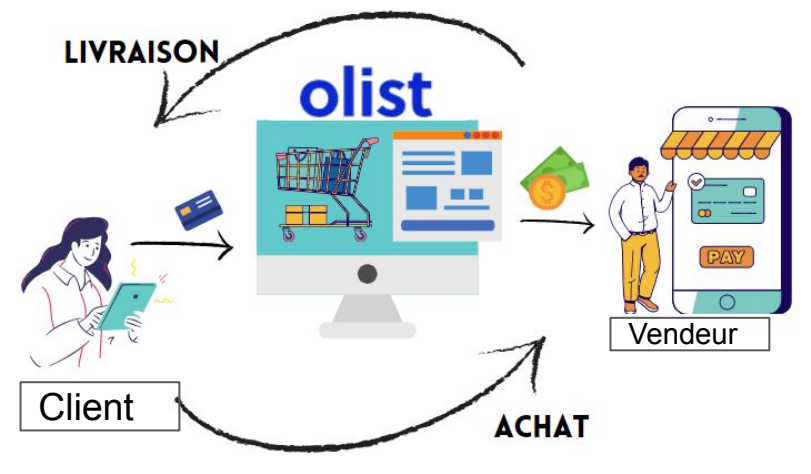


Etudiant : Fatma Aidi
Mentor : Kezhan Shi

Evaluateur :Mohammed Sedki
Date :15/04/2021

TABLE

L’entreprise



-**Market-place** en ligne au Brésil

-Mission: Segmentation des clients pour mener des actions marketing ciblées dans l’objectif de :

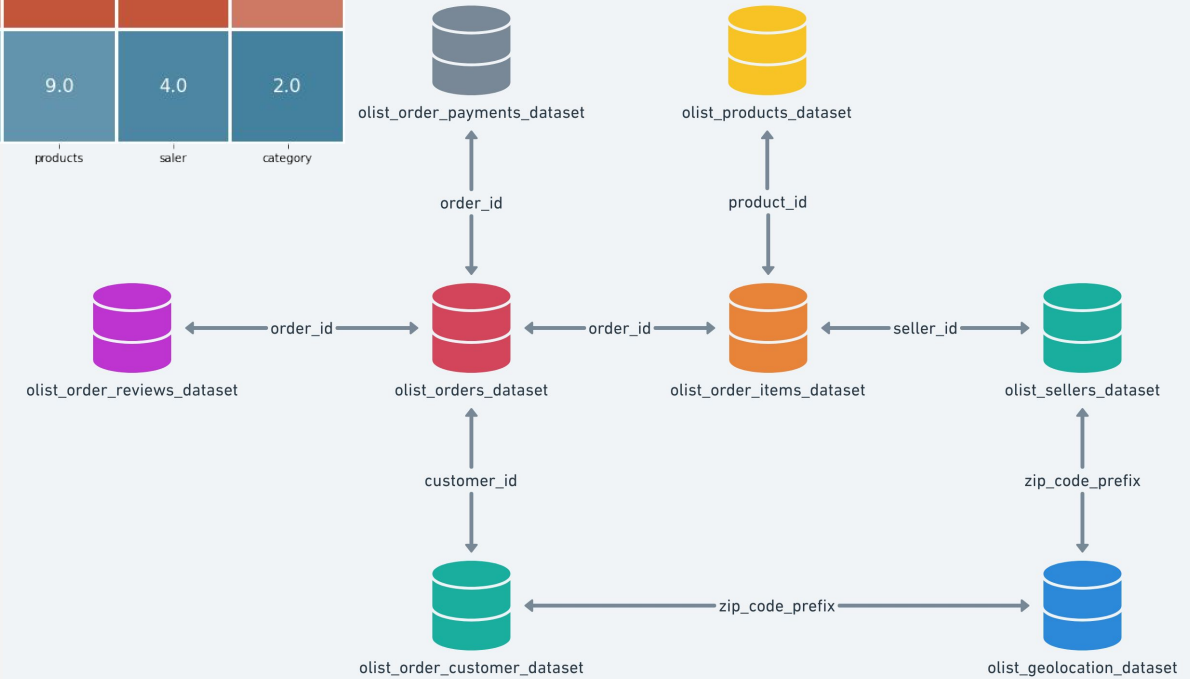
- Ramener plus de clients
- Fidéliser les clients
- Booster les ventes

-Cahier des charges :

- Résultat actionnable d’une segmentation,
- Proposition de contrat de maintenance basée sur une analyse de la stabilité des segments au cours du temps.

2-Préparation des données et analyse

Jeux de données



-**8** bases de données

-les **tailles** varient entre :

- 2 et 9 **variables**
- 71 et 112650 **observations**

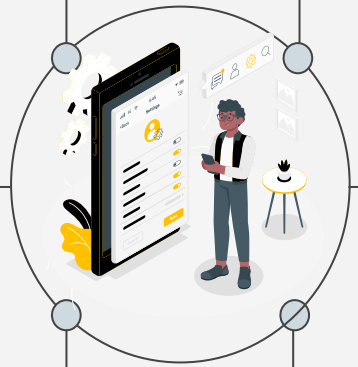
Profil client recherché

Qu'est-ce qu'il achète?

Quand il achète?

A qui il achète?

Categorie de produits



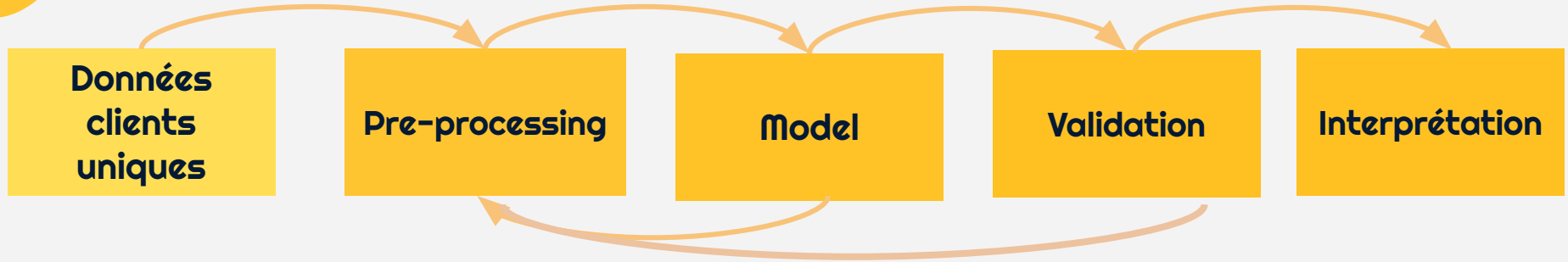
Fréquence d'achat

Nbr de commandes

Combien il paie?

Comment il paie?

Démarche



-Jointure et **agrégation** des tables par client unique

-Features engineering

-**Analyse exploratoire**

-**Features selection**

-Transformation et **normalisation** des variables

-Standardisation et mise à l'échelle

-Centroïde: K-means

-Densité: OPTICS

-Regroupement hiérarchique: AgglomerativeClustering

-Stabilité

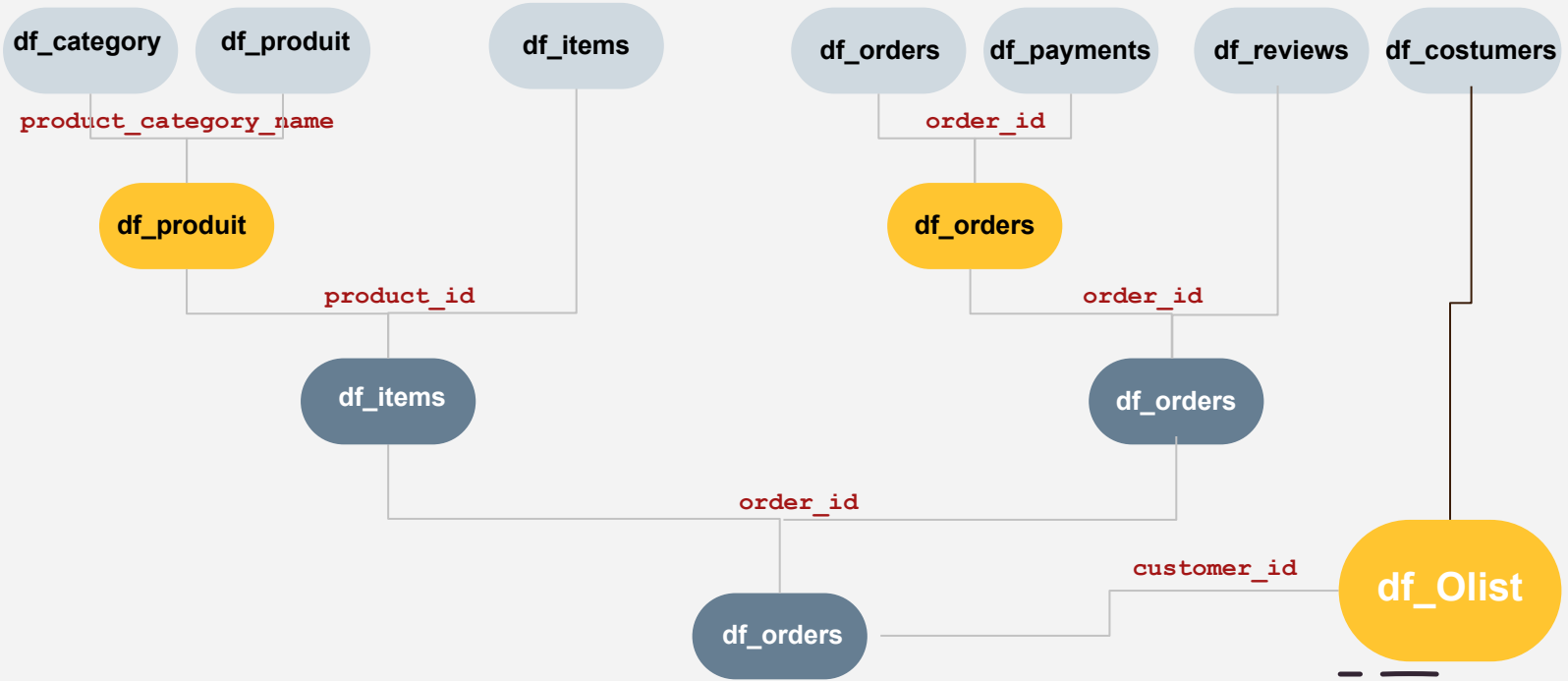
-Homogénéité des clusters
-Interprétabilité du résultat
-Evolution dans le temps

-Profil client

- Comparaison des modeles

-Proposition contrat de maintenance

Jointure et agrégation des tables

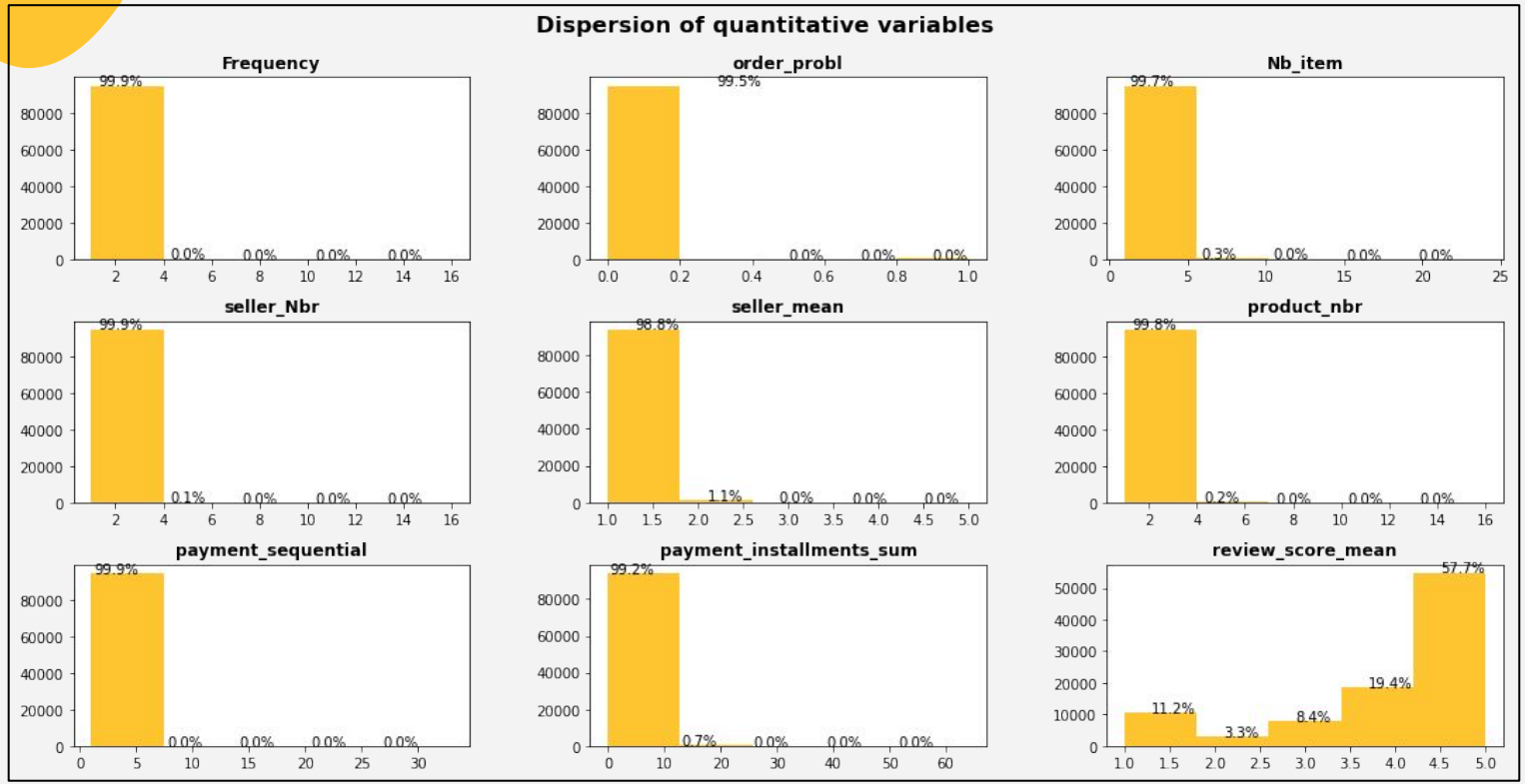


Features engineering

RFM	COMMANDE	PAIEMENT	PRODUIT	SATISFACTION
Recency	Product_nbr (nombre produit/ commande)	payment_sequential	density_product	review_len_message
Monetary	Nb_item (nombre d'article acheté)	payment_type_credit_card	product_weight_g	review_score_mean
Frequency	order_probl	payment_installments	Volume_product_mean	**Vendeur**
Frequency_age	price	nbr_payment_type	catg	seller_Nbr
age	finish_order			

Analyse Exploratoire

-Distributions positivement biaisées (non normales)
-**94%** ont commandé 1 fois
-90% ont acheté un seul article
-**70%** sont satisfaits
-**97%** ont passé une commande chez un seul vendeur

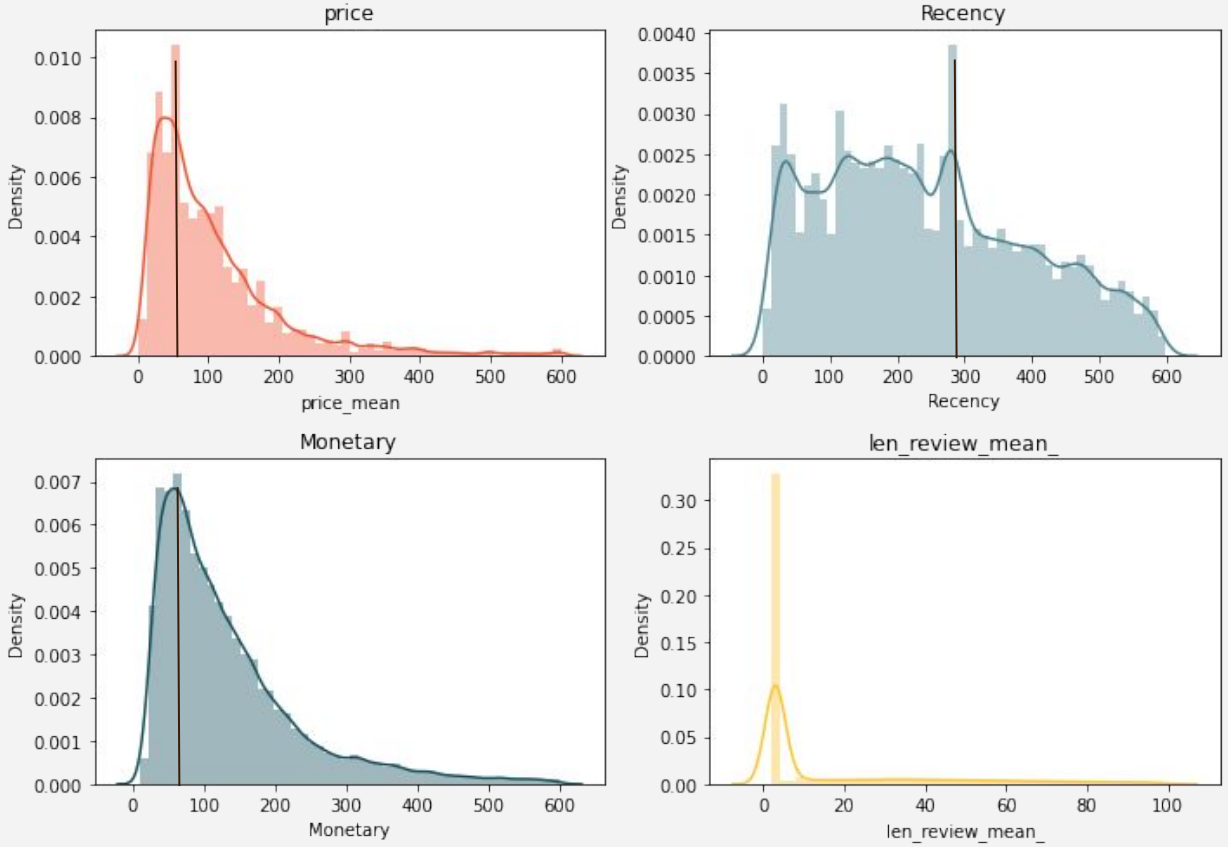


Analyse Exploratoire

Distributions **positivement biaisées** avec **pic**

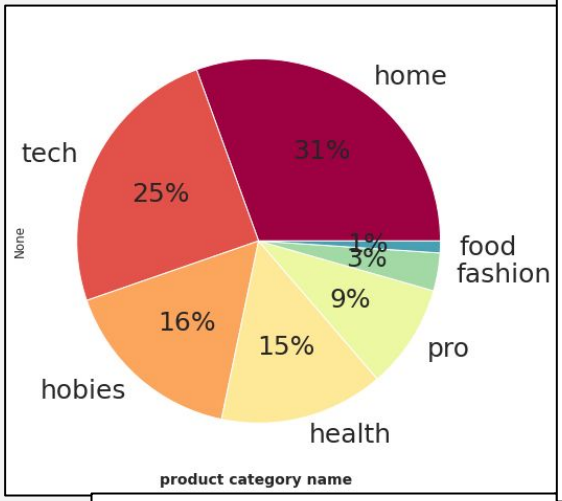
- **60** Réals prix moyen des articles achetés par client.
- **70** Réals dépense par client.
- **5** mots par commentaire

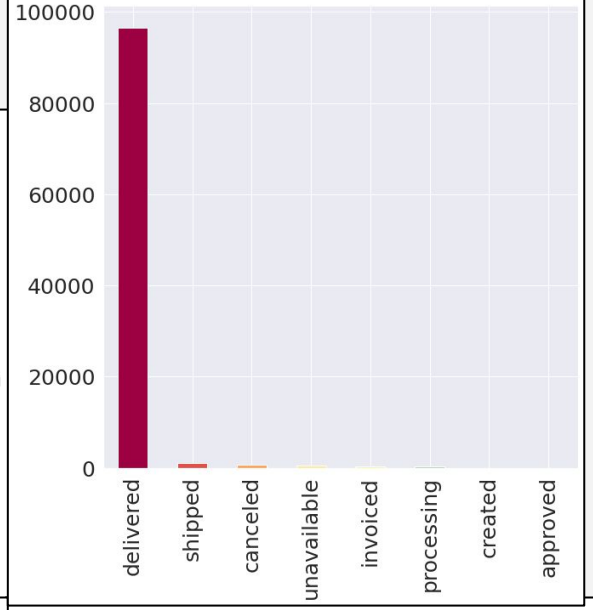
- **300** jours depuis la dernière commande

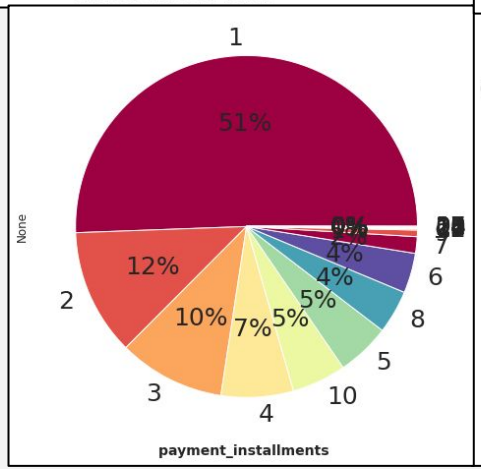


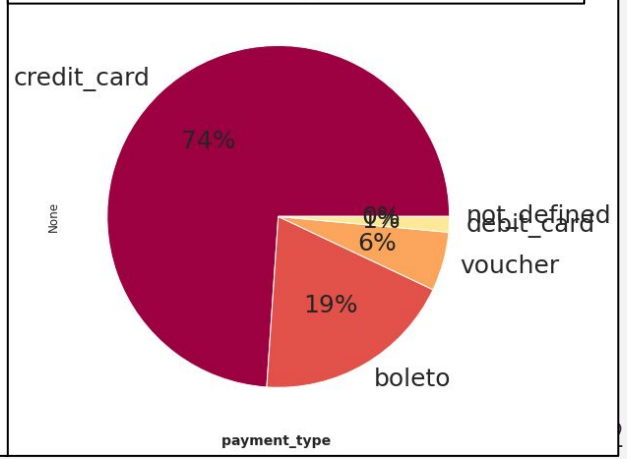
Analyse Exploratoire

- Paiement par **carte bancaire** majoritaire (plus de **74** %)
- Les **Catégories** maison et électronique représentent plus **50**% des articles vendus
- **95**% des commandes sont **sans problèmes de livraison**
- **50**% des clients paient en une seule fois.









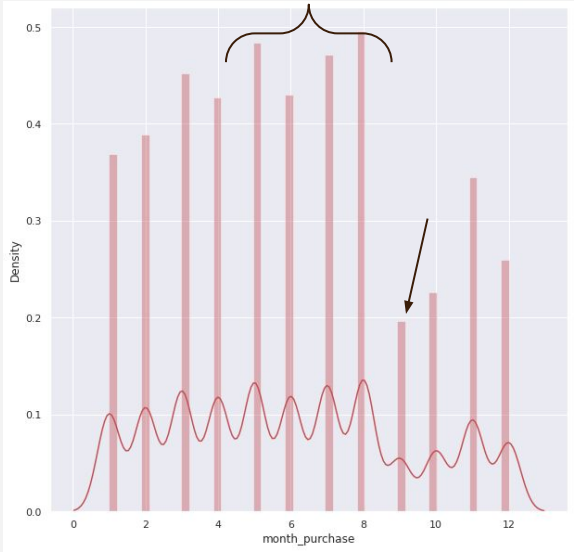
Analyse Exploratoire

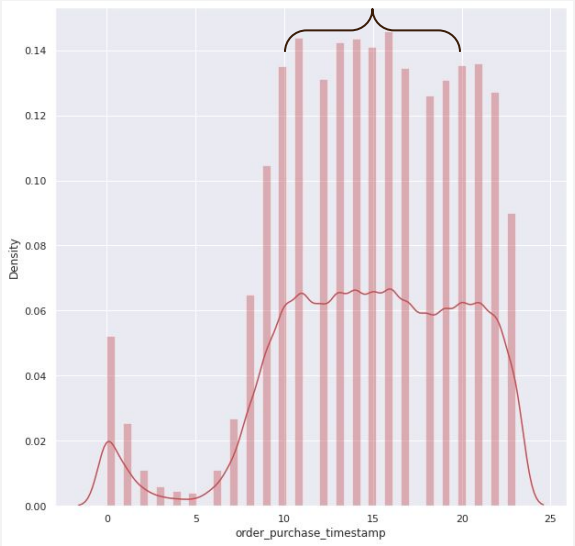
-**Opération d'achat** dure en moyenne 10 jrs

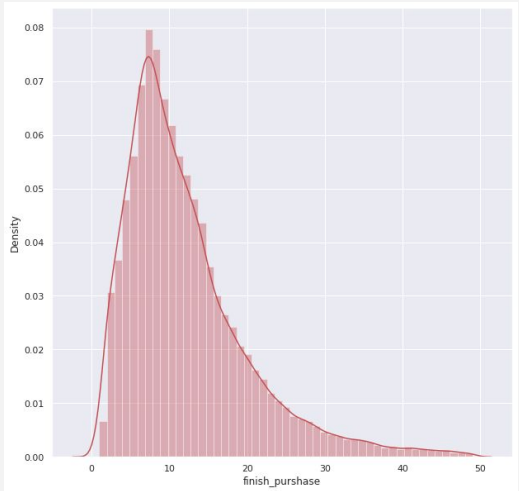
-Augmentation de nombre de commandes en **2018**

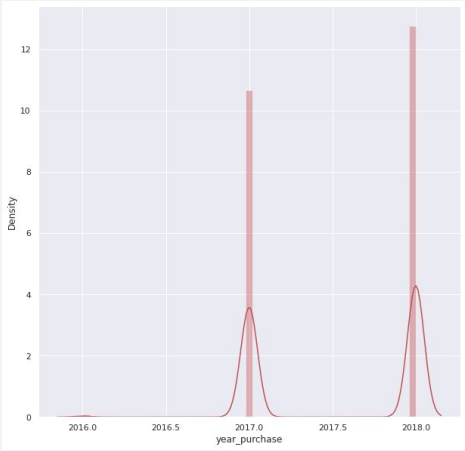
-Les meilleurs **mois** sont de Mai à Août

-La majorité des ventes se font entre 12h et 19h





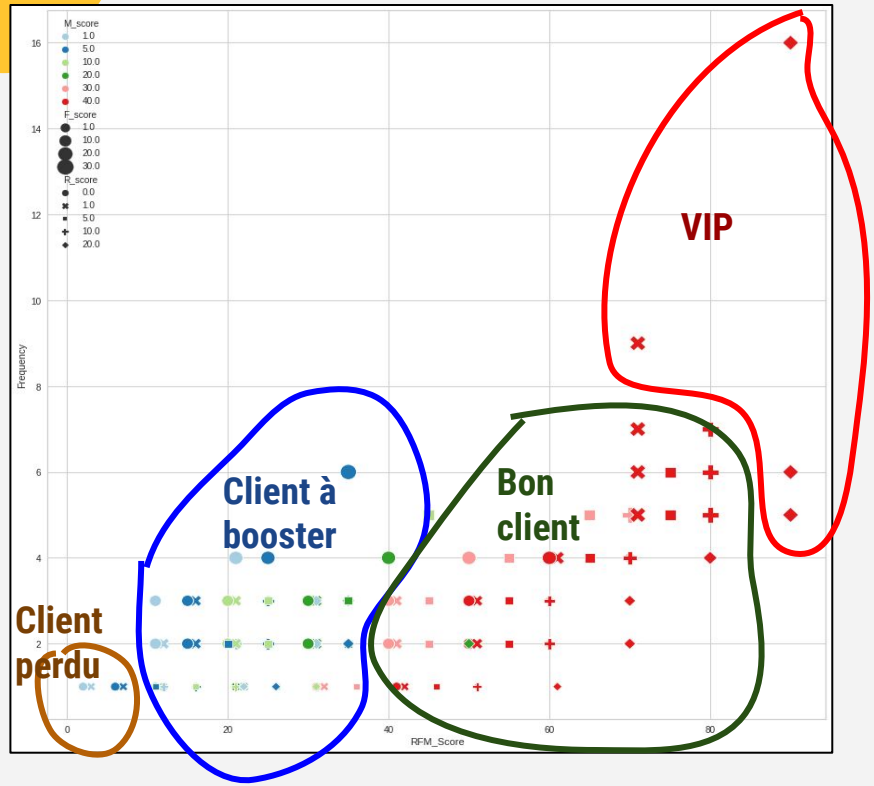




Récapitulatif

- **96% des commandes contiennent un seul article**
- **2%** des clients ont acheté chez plusieurs vendeurs
- **94**% de nos clients ont acheté sur le site une **seule fois**.
- **95**% ont payé avec un **seule mode de paiement**
- **50**% ont choisi un paiement en **plusieur fois**.

Segmentation RFM



RÈGLE pour RFM SCORE

Recency/JR	FREQUENCE	MONTANT
0-60 : 20pt	1 : 1pt	0-100 : 1pt
60-120 : 10pt	2-3 : 10pt	100-200 : 5pt
120-180: 5pt	4 : 20pt	200-300 : 10pt
180-240: 1pt	>=5 : 30pt	300-400 : 20pt
>=240 : 0pt		400-500 : 30 pt
		>=500 : 40 pt

-Client **Haut_VIP** RFMS>80
-Client **bon** 40<RFMS<80
-Client **à booster** 10<RFMS<40
-Client **perdu** RFMS<10

Pour la suite on isole les **VIP** que l'équipe marketing traite à part. Ils sont **fidèles,** achètent au moins **5 fois** et sont dépensiers: **panier moyen élevé**

RFM signifie Récence, Fréquence et Valeur monétaire, chacun correspondant à une caractéristique d'un client . Ces mesures sont des indicateurs importants du comportement d'un client:

- **Récence**: à quand remonte la dernière fois qu'ils ont acheté?
-Fréquence: à quelle fréquence et pendant combien de temps ont-ils acheté?
-**Valeur monétaire / ventes**: combien ont-ils acheté

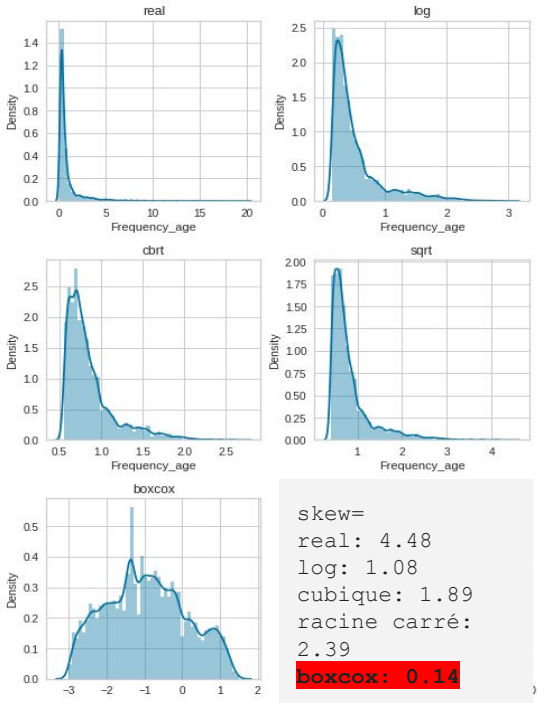
3-MODEL NON SUPERVISE: K-MEANS

K-MEANS : Pre-processing

-**Choix K-means:** C'est l'un des algorithmes de clustering les plus répandus.Il s'appuie sur la mesure de distance pour créer les clusters en minimisant la somme des carrés des distance entre un point et la moyenne des points de son cluster(centroïde).

-PowerTransformer pour stabiliser la variance et rendre les distributions plus normalisées. le résultat de calcul d'asymétrie des données est presque nul.

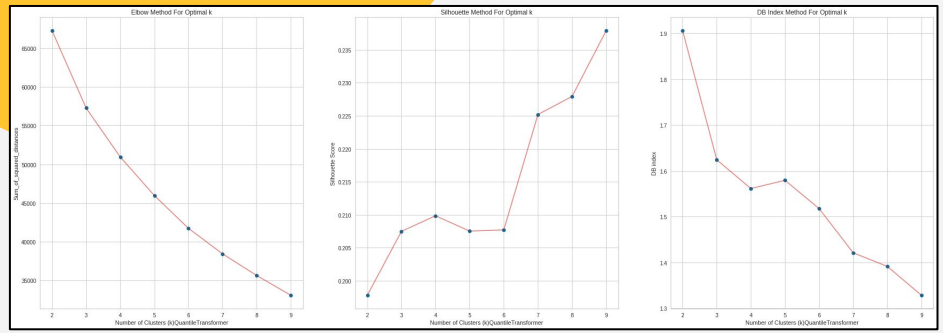
-**Transformation et Mise à l'échelle** :MinMaxScaler ,StandardScaler, QuantileTransformer



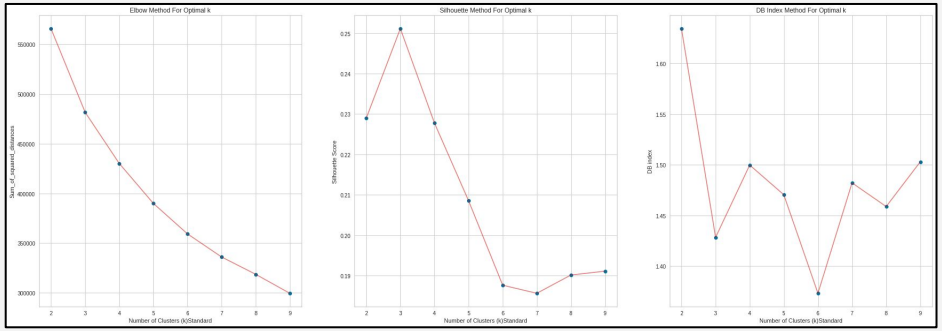
Features selections: 8	
Recency	La période depuis la dernière commande
Frequency_age	Taux d'achat par jour
Monetary	Le montant d'achat
age	La période depuis la première commande
payment_installments_sum	nombre de tranches de paiement
review_score_mean	Le score moyen donné par le client
len_review_mean	La longueur moyenne du commentaire
Nb_item	Nombre d'article acheté

K-MEANS :Nombre de clusters

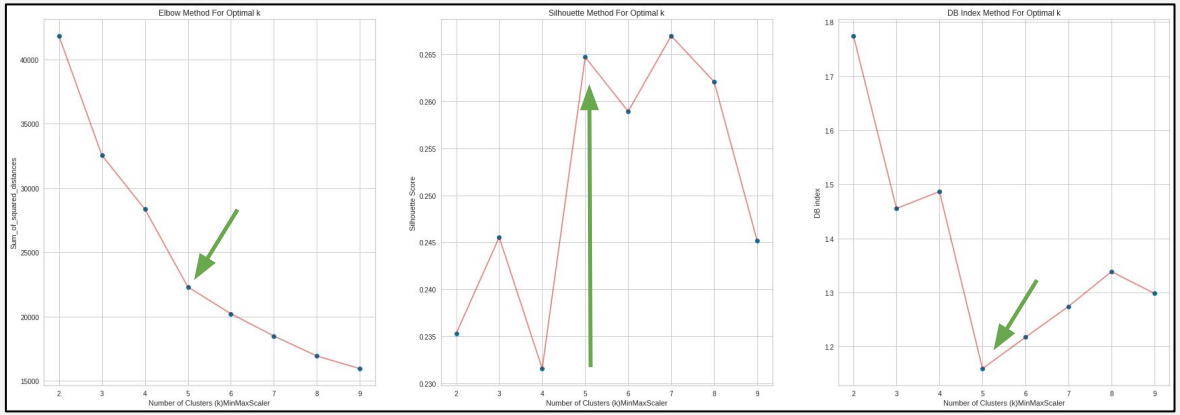
QuantileTransformer nbr:7,9,5



StandardScaler:6,3,5



MinMaxScaler nbr:5



-**Indice Silhouette** (à maximiser):mesure la **cohésion** et la **séparation**: calcul de la différence entre la distance «intra-classe» et la distance au centroïde le plus proche d'un autre cluster.

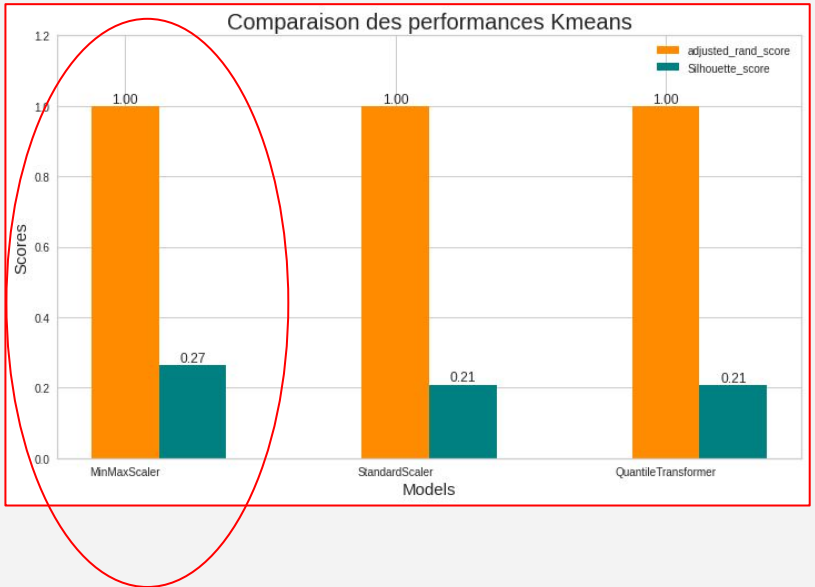
-Distortion (coude):mesure la **variance** «intra-classe» : Moyenne des sommes des distances quadratiques des points au centroïde le plus proche.

-**Davies_Bouldin** (à minimiser) il mesure la **l'homogénéité** et la **séparation**:est basé sur le rapport entre les distances «intra-classe» et «inter-classe»

Stabilité des clusters

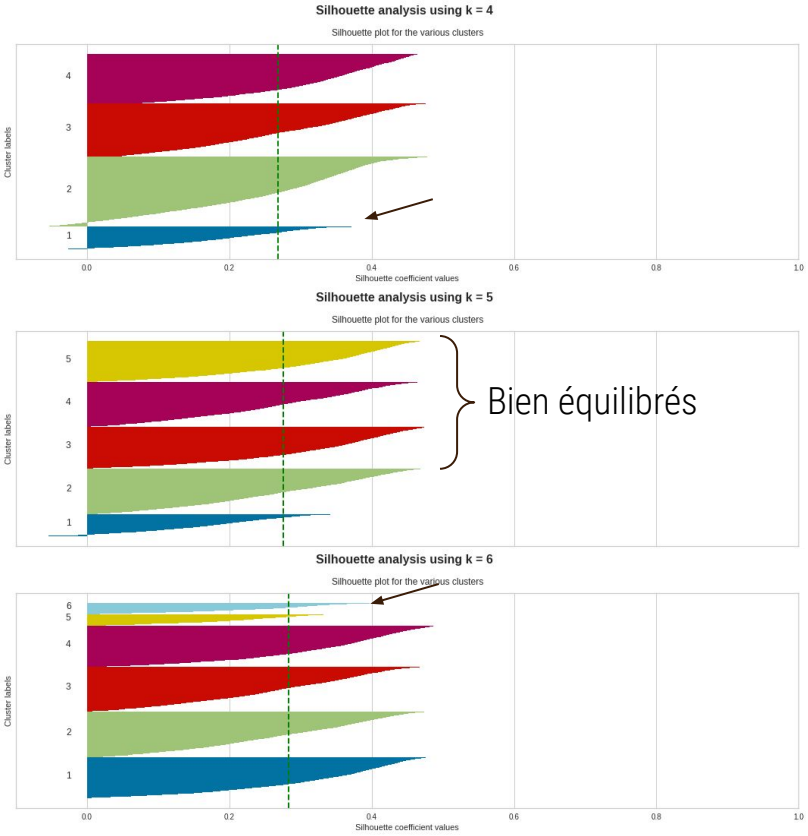
- ARI: adjusted rand score: est une mesure de similarité entre deux classifications, un indice qui utilise les étiquettes externes (réelles). Dans notre cas, on l'utilise pour voir si, à l'initialisation, les clients changent de cluster ou non.
- les 3 modèles sont stables mais on retient le **MinMaxScaler** car il donne un meilleur score de silhouette à 0.27.

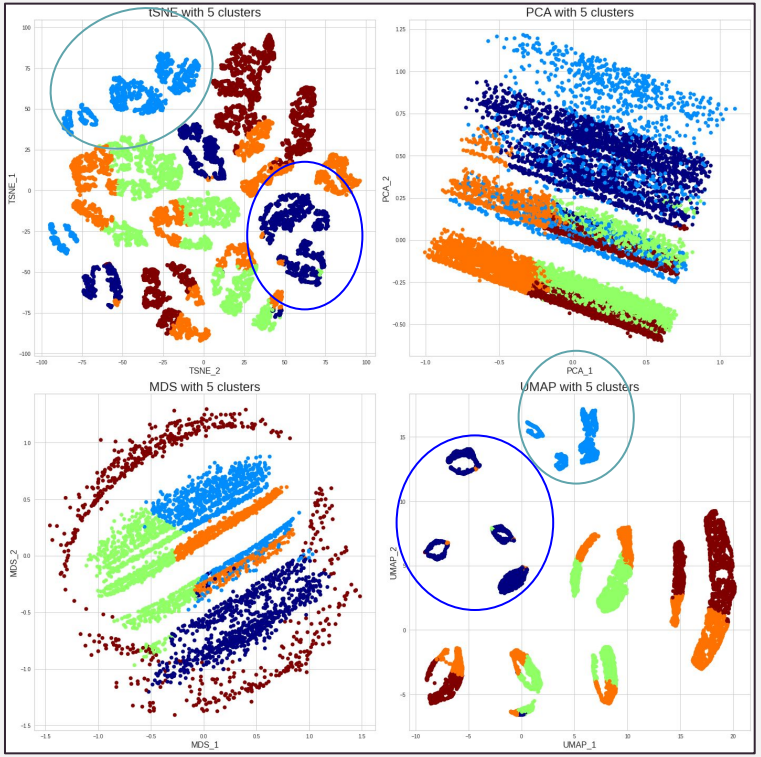
```python
def score_ari(df,model, nb_itr=6):
    # Fitting the model
    model.fit(df)
    labels_true=model.labels_
    # Calculate the ARI scores
    ARI_scores = []
    # Iterating
    for i in range(0,nb_itr):
        # Fitting the model
        model.fit(df)
        labels_predict= model.labels_
        # Compute the ARI score with labels_true
        ARI_score= adjusted_rand_score(labels_predict,labels_true)
        ARI_scores.append(ARI_score)
    #Compute the mean of ARI scores
    ARI_mean = statistics.mean(ARI_scores)
    return ARI_mean
```



K-MEANS :Visualisation

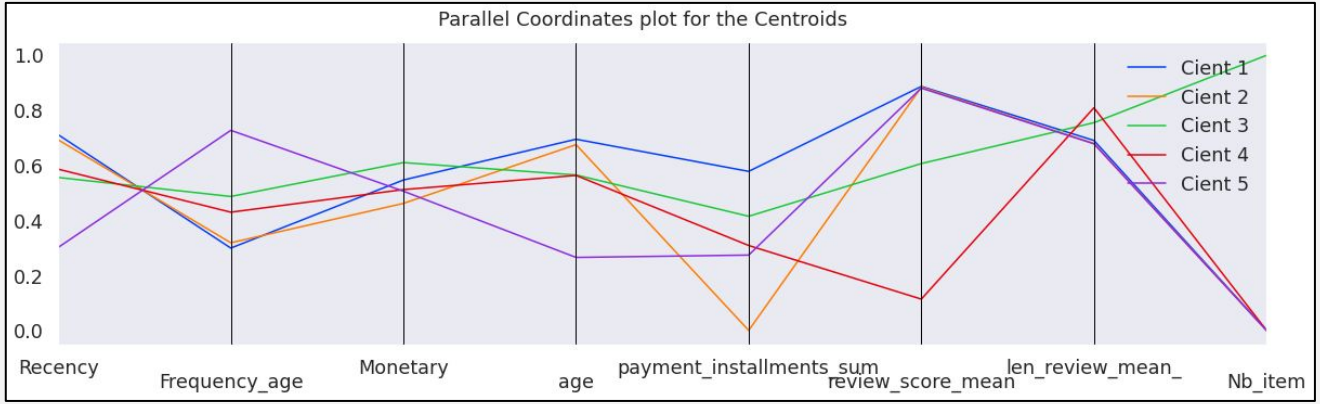
- La projection 2D de tSNE , MDS ,PCA et UMAP montrent que au moins 2 clusters sont bien regroupés
- 5 est le nombre de cluster optimal

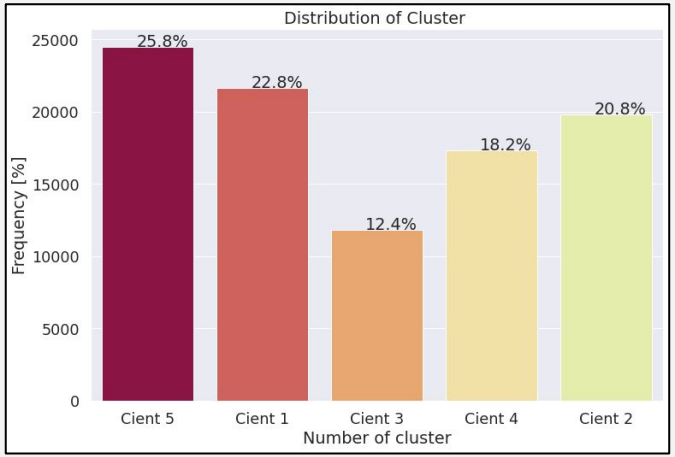


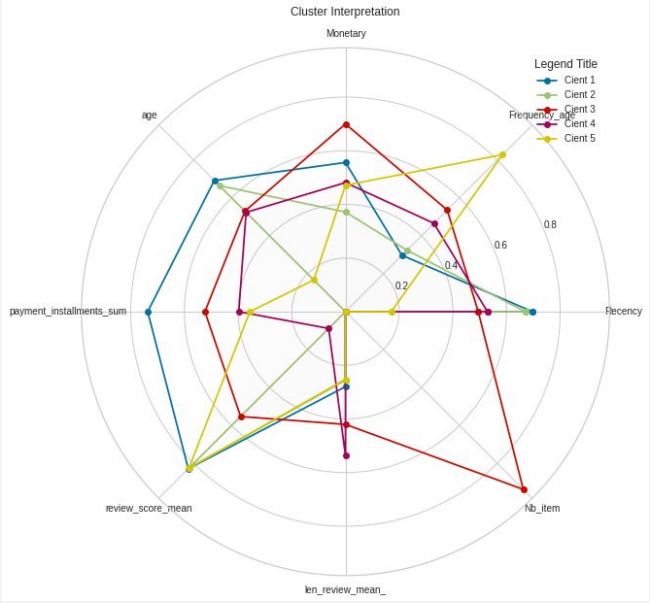


INTERPRETATION

- **Clusters sont :**
  - Homogènes
  - Équilibrés
  - Lisibles
  - Interprétables

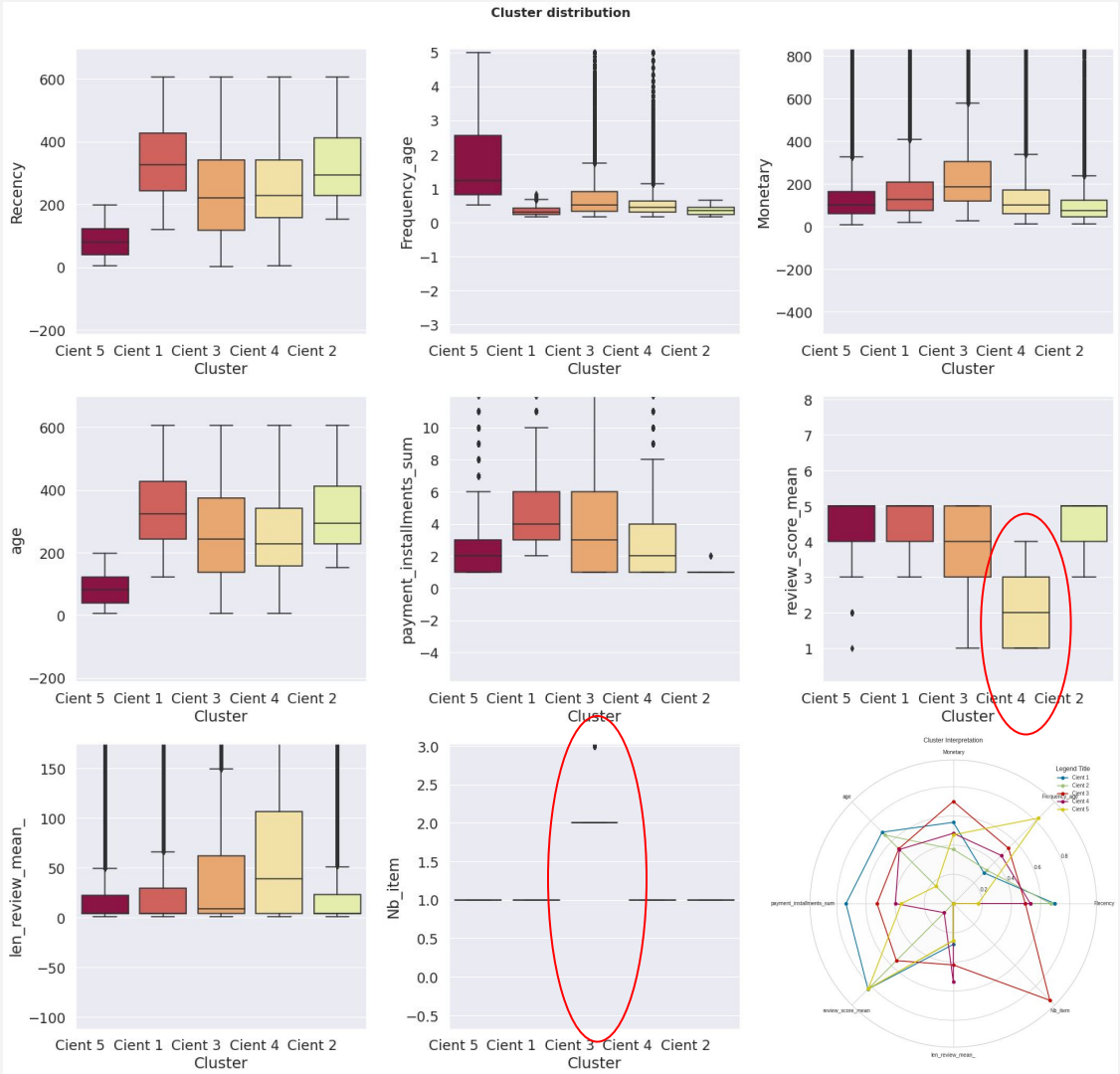






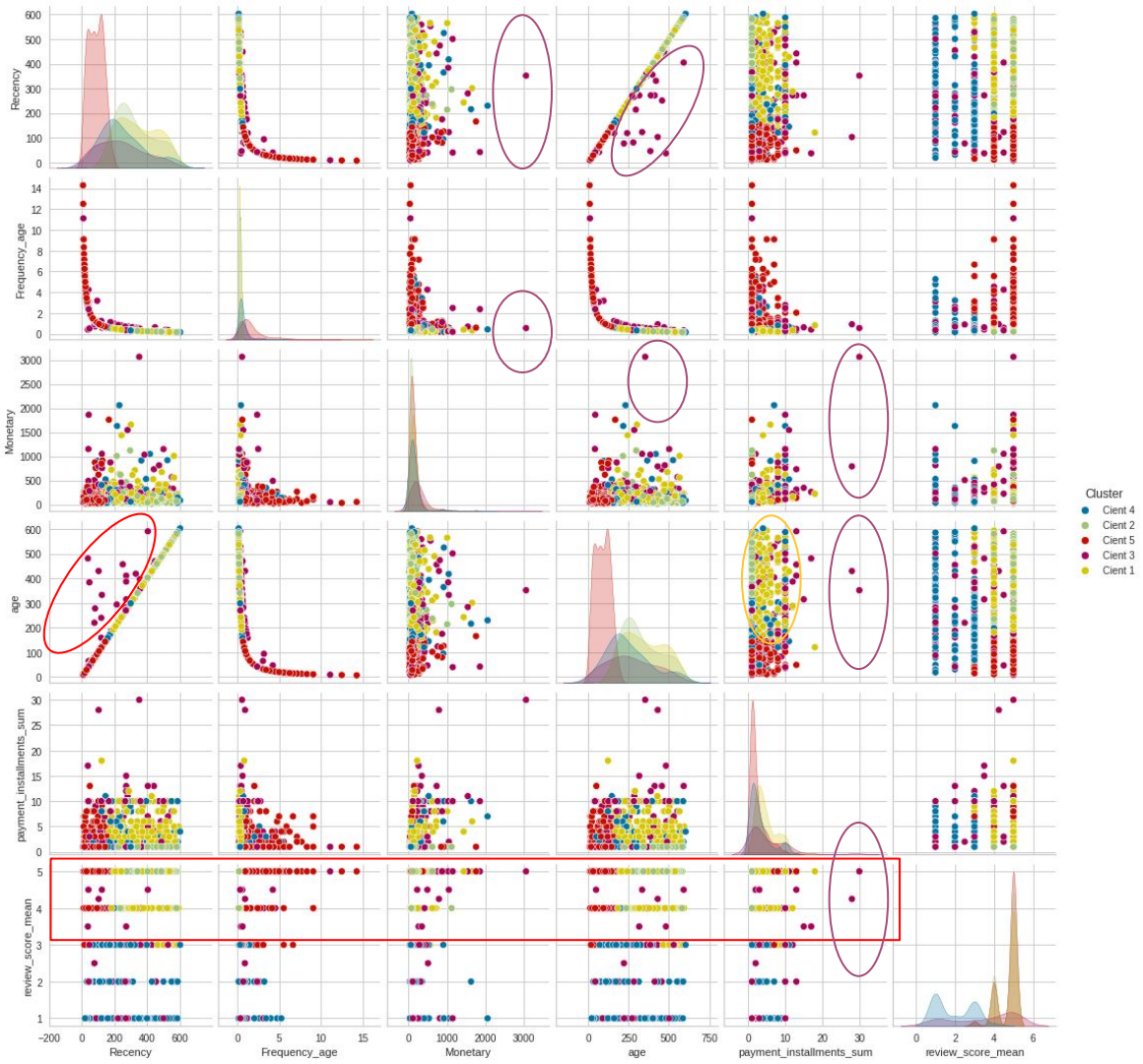
INTERPRETATION

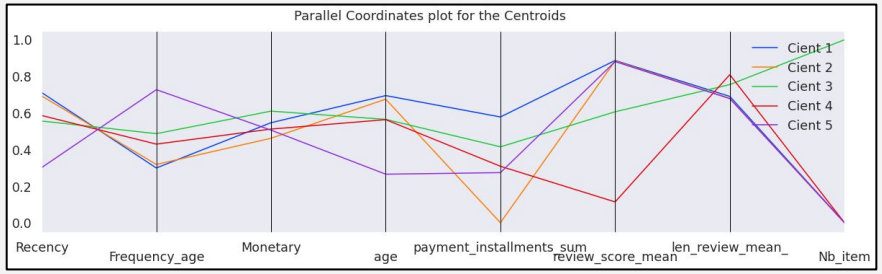
- **Client 2, 1 et 5** sont satisfaits
- **Client 5** est nouveau.
- **Client 2** paye en seule fois.
- **Client 4 et 3** sont les moins satisfaits.
-**Client 3** est le plus dépensier et qui a acheté au moins 2 articles.



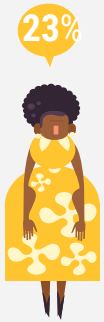
INTERPRETATION

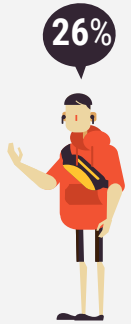
- **Recency /age** on arrive à séparer les clients qui ont acheté plusieur fois (**6**% du nombre total) quand age=Recency.
- Plus Monetary est grand plus le paiement échelonné.
- **Client 3** represente le client outlier .
-**Client 1** est l'ancien ou il n'a pas acheté depuis longtemps.





PROFIL CLIENT K-means





Ancien et aléatoire

Les clients **anciens** qui n'ont pas fait d'achat **depuis longtemps,** ils ont fait des achats pour des montants **moyens (200 Real).**

Ancien économe

Les clients **anciens** qui n'ont pas fait d'achats **depuis longtemps**, font des achats pour des **montants faibles** et il paient en **une seule fois**.

Dépensier

Les clients plutôt **anciens** :qui achètent **différents** articles, pour des montants **élevés** (>500 real)et ont besoin d'un paiement **échelonné** (>5 fois).

Insatisfait

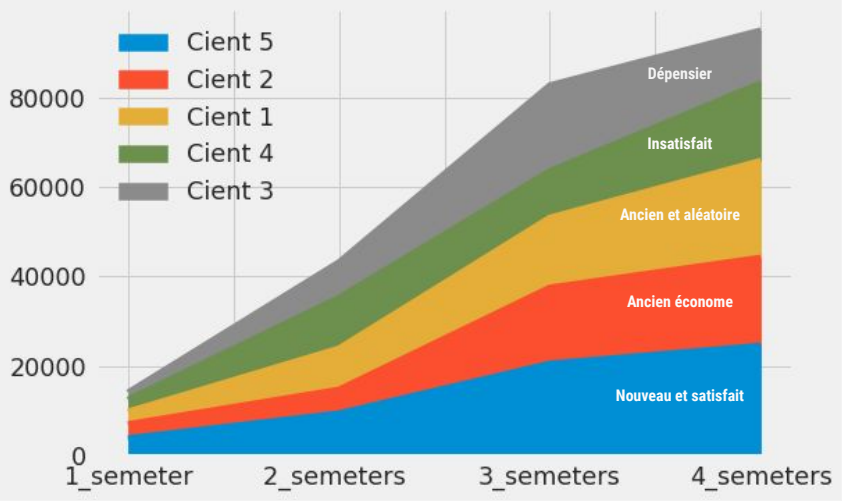
Les clients qui ont fait des achats **récemment**, ils achètent **rarement**, font des achats pour des montants **faibles** à **moyens**. Ils sont **mécontents** (ils laissent des commentaires assez longs)

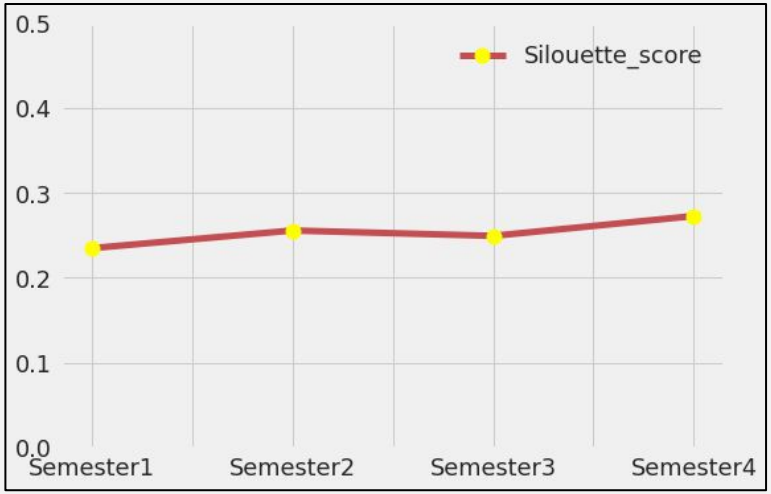
Nouveau et satisfait

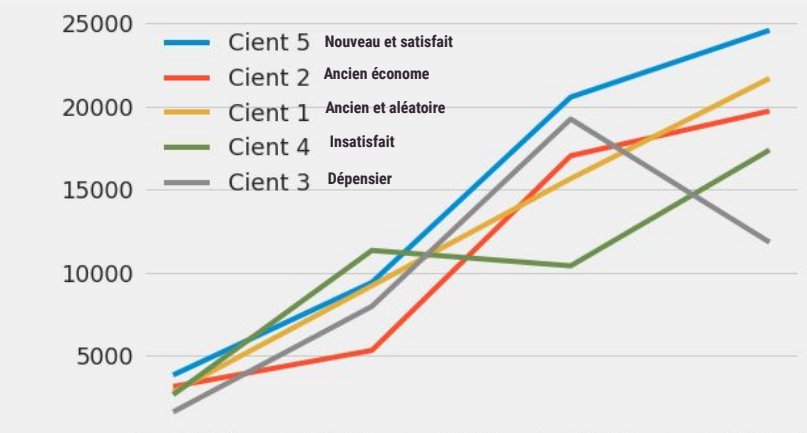
Les clients **nouveaux** : font des achats **récemment** pour des montants **faibles** à **moyens** et ils sont satisfaits de leur achat.

Stabilité des clusters dans le temps

- Diviser la data en 4 temps et faire la classification à chaque fois et recaler les noms clusters (donner aléatoirement par K-means).
-Stabilité dans le temps du **silhouette_score**
-Baisse du **nombre de clients dépensiers** (client 3) sur le 4ème semestre.
- Hausse du **nombre de nouveaux clients** (client 5) sur le 4ème semestre.

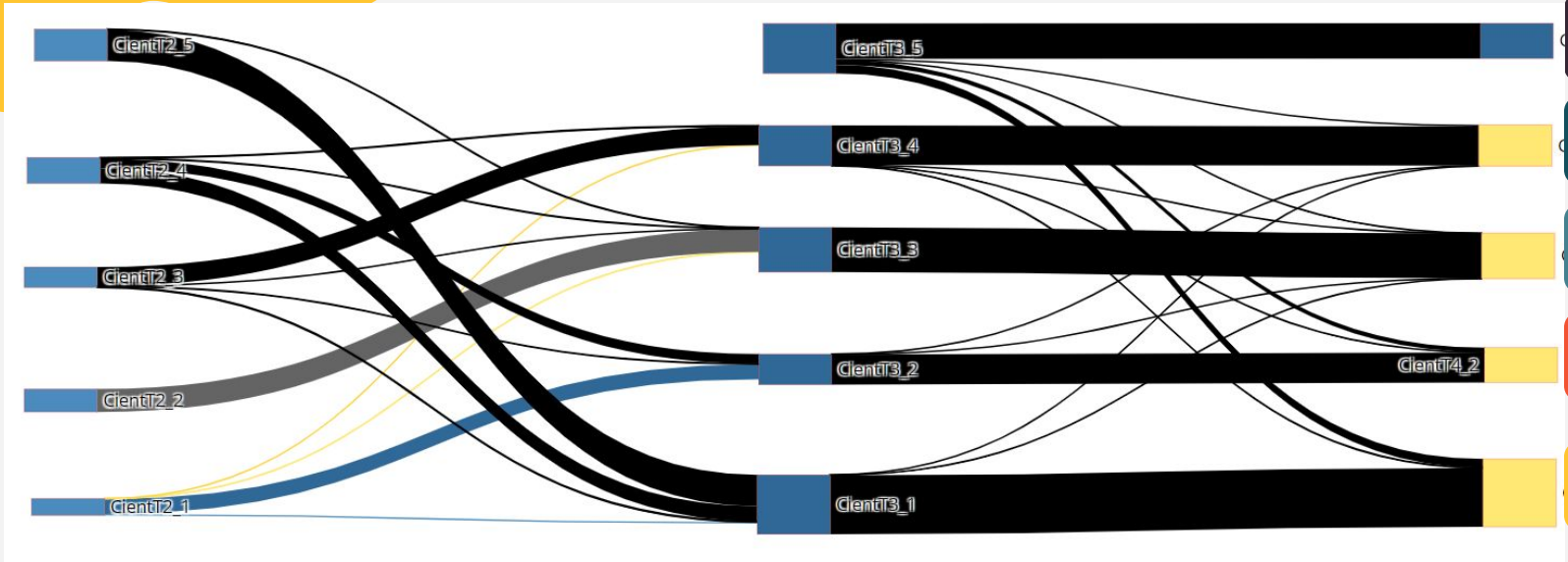






	Cient 5	Cient 2	Cient 1	Cient 4	Cient 3
1 semeter	3840	3145	2808	2657	1619
2 semeters	9404	5318	9207	11330	7966
3 semeters	20541	17016	15644	10402	19226
4 semeters	24560	19706	21667	17363	11820

Contrat de maintenance



Nouveau et satisfait

Insatisfait

Dépensier

Ancien économe

Ancien et aléatoire

-On fait la classification sur chaque base de données de chaque période (condition du futur contrat de maintenance)

-Diagramme de **Sankly**:

- Les clients **migrent** d'un groupe à un autre chaque semestre.
- un **contrat de maintenance** est nécessaire tous les **6 mois**

4- OPTICS: Ordering Points To Identify the Clustering Structure

OPTICS : Pre-processing

-**Choix OPTICS:** s'appuie sur la densité estimée des clusters pour faire la classification. On travaille sur un échantillon de 30 000 clients.

.

-PowerTransformer pour normaliser des distributions.

-**Transformation et Mise à l'échelle** :MinMaxScaler ,StandardScaler, QuantileTransformer.

Features selections: 8	
price_mean	Le prix moyen des produits
order_item_mean	Nombre moyen des produit par commande
product_nbr	Le nombre de différent produit par commande
seller_Nbr	La période depuis la première commande
Frequency_age	Taux d'achat par jour
payment_installments_sum	nombre de tranches de paiement
age	La période depuis la première commande
item_Monetary	Panier moyen par client(montant/nbr de commande)

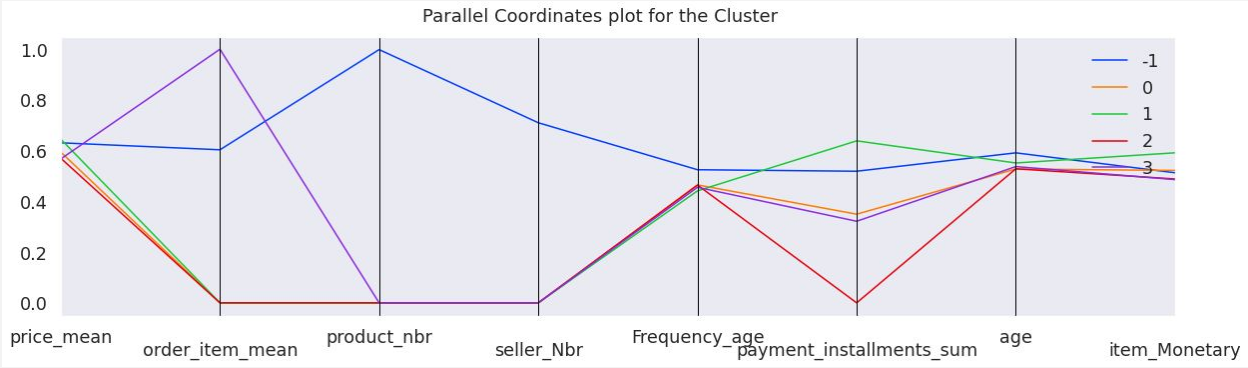
OPTIC+ MinMaxScaler

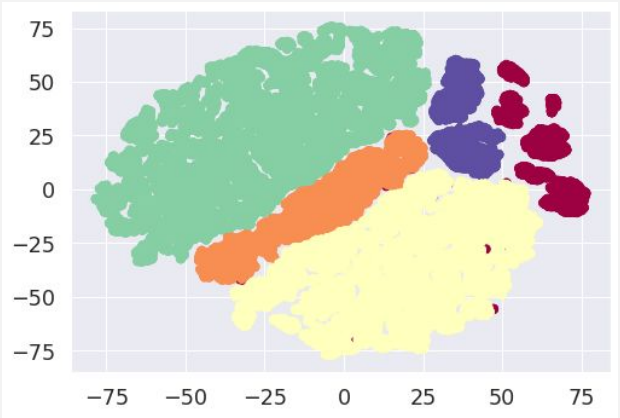
-Nombre de clusters :5

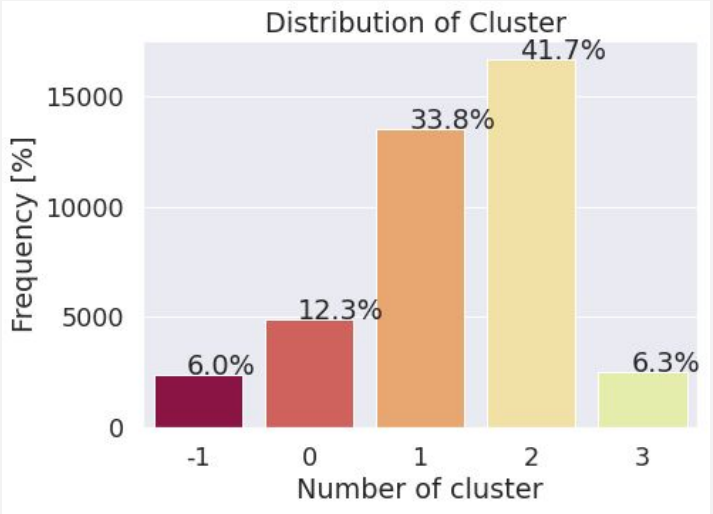
-Client 2 représente les clients qui payent en une seul fois.

-Client -1 représente les clients qui achètent plusieurs produits différents et chez des vendeurs différents.

-On a 6% de bruit:-Clusters déséquilibrés mais interprétables.





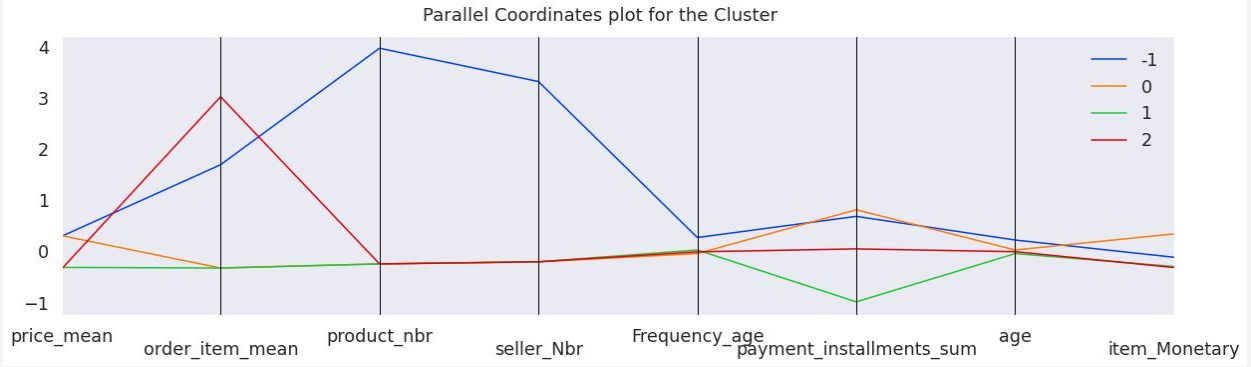


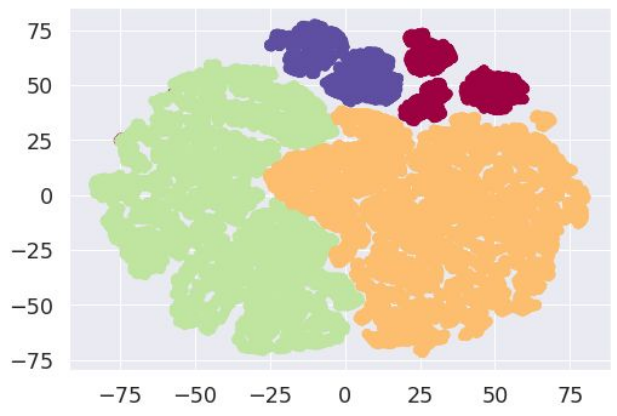
OPTIC+ StandardScaler

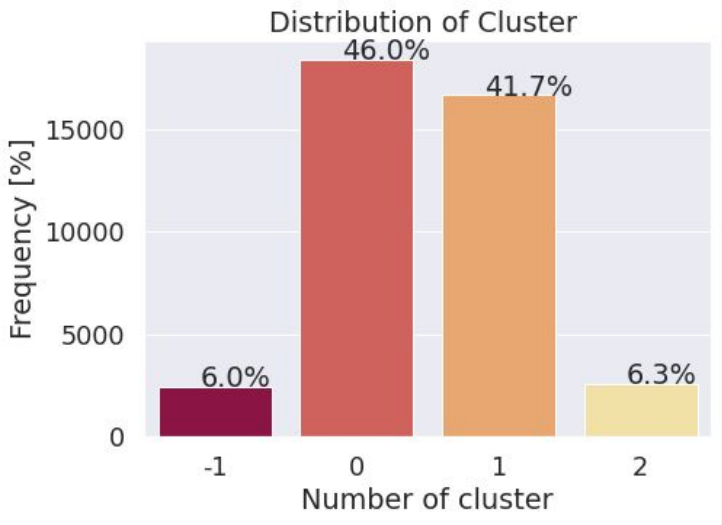
-Nombre de cluster :4

-Client (-1) représente les clients qui achètent chez plusieurs vendeurs.

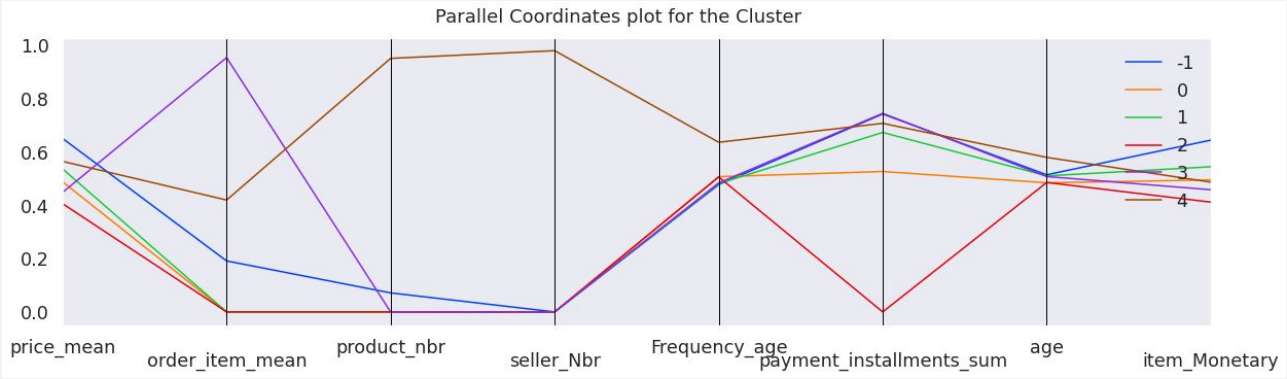
-On a 6% de bruit:-Clusters déséquilibrés mais interprétables.







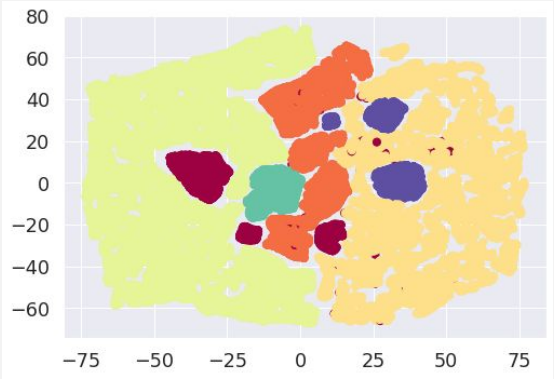
OPTICS+ QuantileTransformer

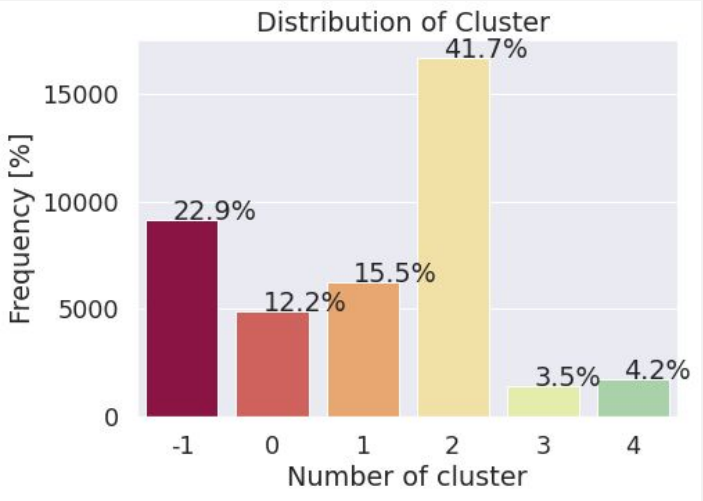


-Nombre de cluster :6

-Client 4 représente les clients qui achètent chez plusieurs vendeurs.

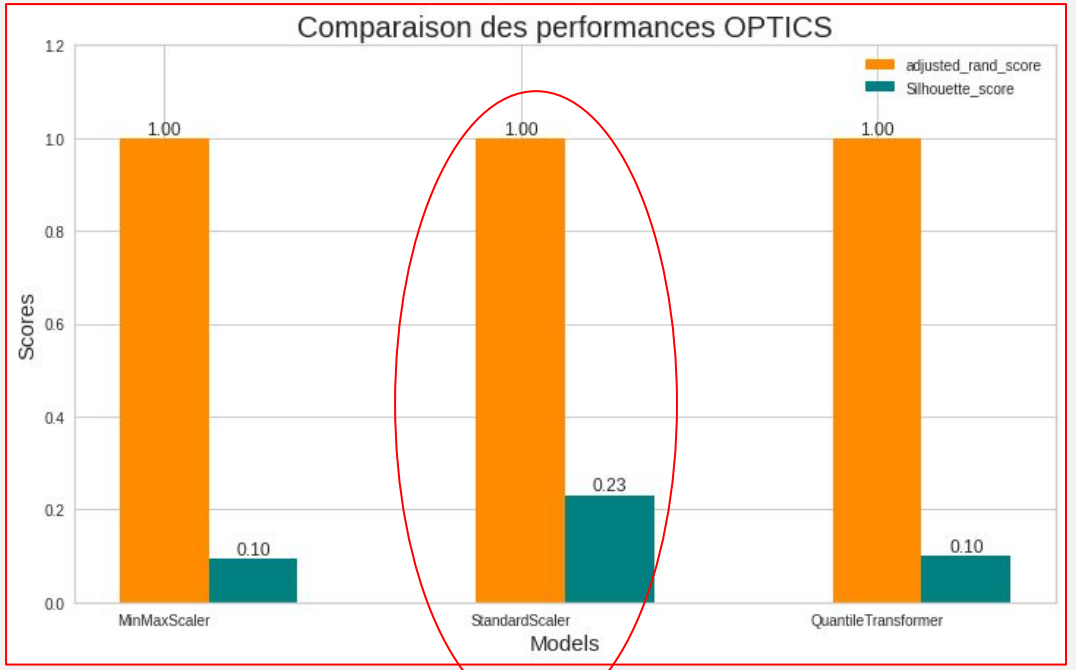
-On a 23% des données sont des bruits:-Clusters déséquilibrés mais interprétables.



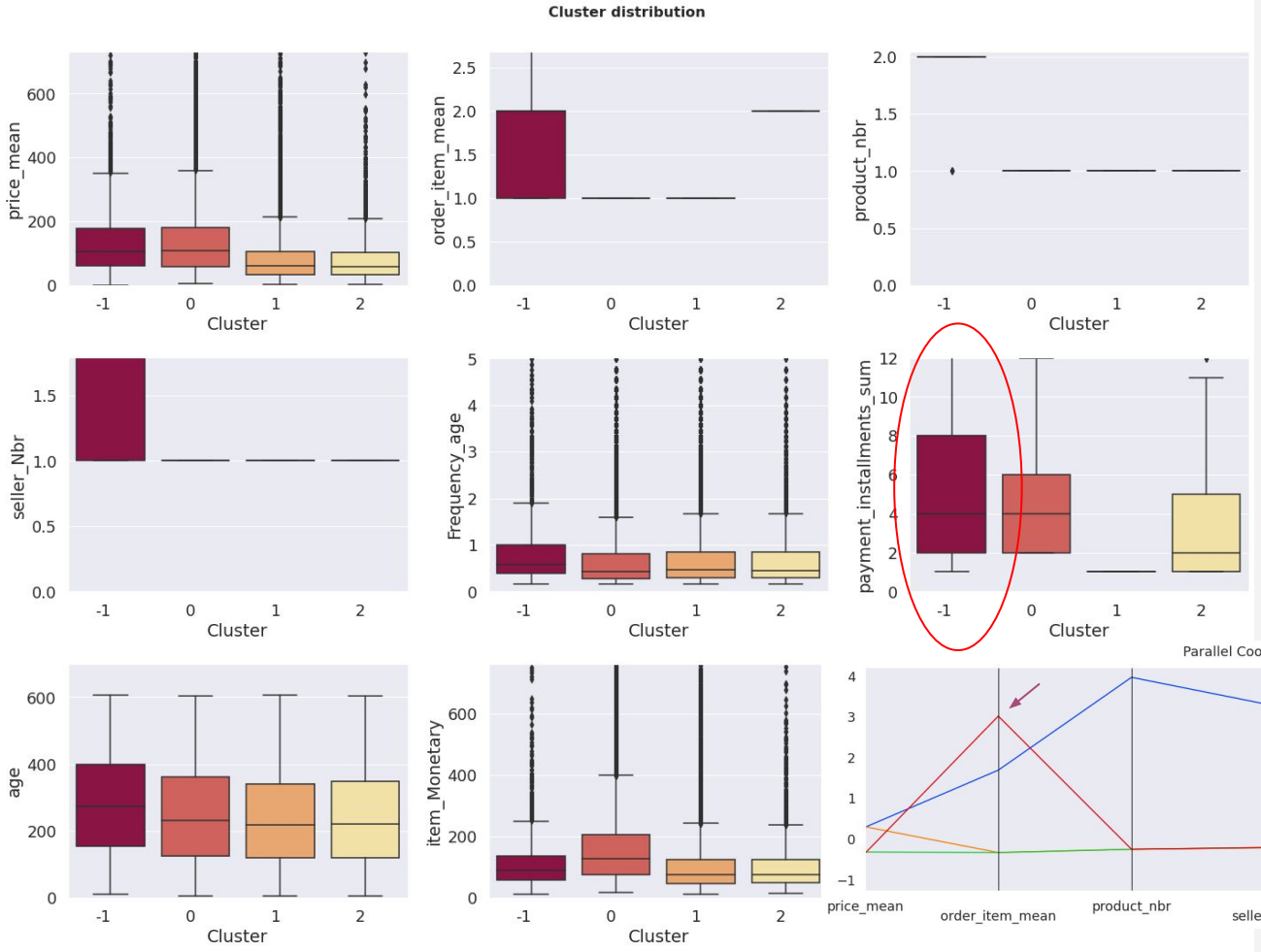


Stabilité des clusters

- les 3 modèles sont **stables** mais on retient le **StandardScaler** car il donne un meilleur score de silhouette à 0.23.

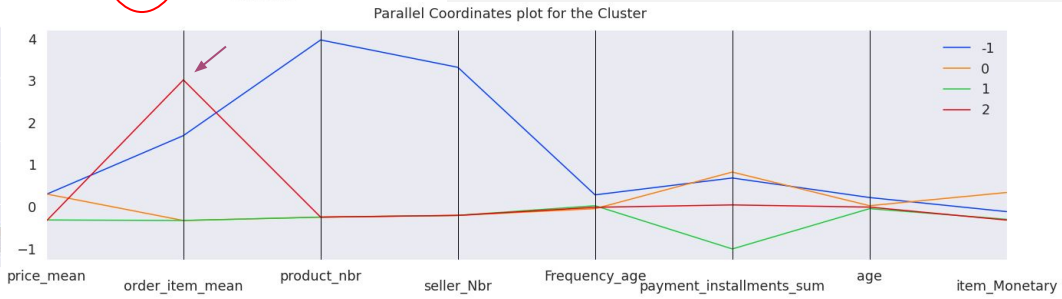


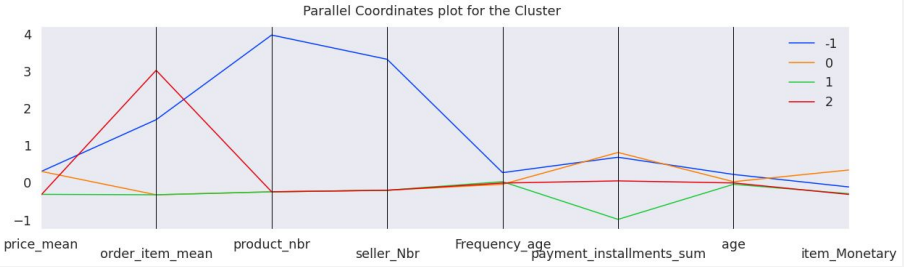
INTERPRETATION



-**Le client -1** achètent chez plusieurs vendeurs et des articles différents

- **Le client 2** achètent le même article en quantité





PROFIL CLIENT OPTICS

6%

Fidèle

Les clients **anciens** qui achètent **régulièrement**, chez **plusieurs vendeurs, et** plusieurs produits **différents (product_nbr)** avec une valeur de **panier moyennement élevée** et ont besoin d'un paiement **échelonné**.

46%

Panier moyen

Les clients qui font un seul achat d'un seul produit avec un prix relativement **élevé (100 Réal)** et ils payent en plusieur fois.

42%

Econome

Les clients qui font des achats d'articles **pas chers** et payent en une **seule fois**.

6%

Fidèle à un article

Les clients qui achètent le même article, en quantité, pour des montants relativement **faibles à moyens. I**ls ont besoin de payer en **plusieur fois**

4-MODEL: Hierarchial Clustering

Algorithmes de clustering hiérarchique : Pre-processing

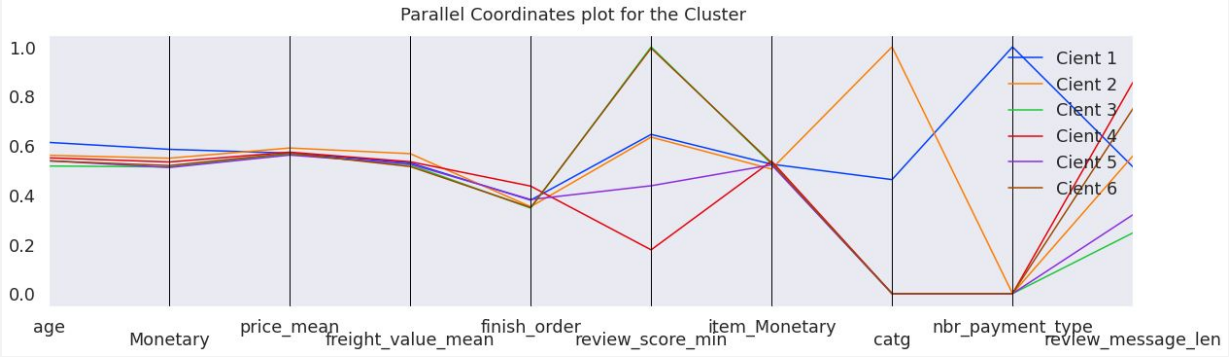
-**Choix AgglomerativeClustering:** Approche ascendante qui commence par de nombreux petits clusters et les fusionne pour créer de plus grands clusters.

-PowerTransformer pour normaliser des distributions.

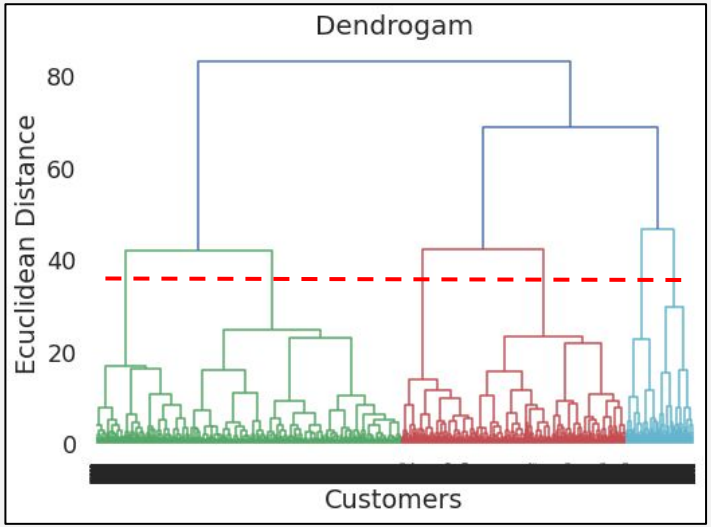
-**Transformation et Mise à l'échelle** :MinMaxScaler ,StandardScaler, QuantileTransformer

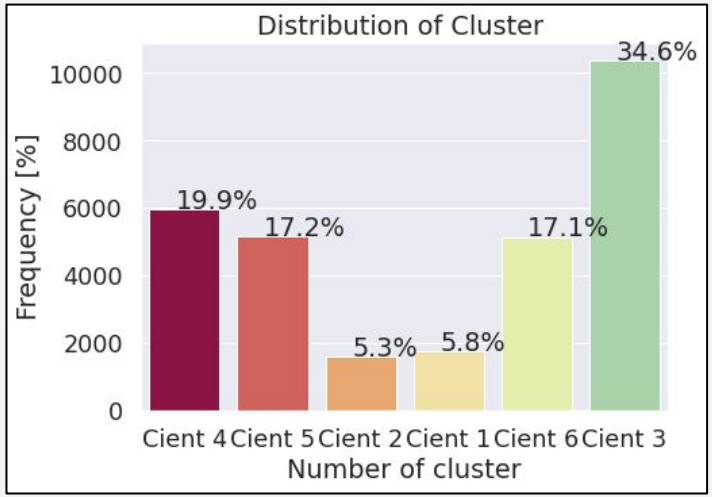
Features selections: 10	
age	La période depuis la première commande
Monetary	Le montant d'achat
price_mean	Le prix moyen des produits
freight_value_mean	Poids moyen des produits achetés
finish_order	La durée pour clôturer une commande
review_score_min	Le score minimum donné par le client
item_Monetary	Panier moyen par client(montant/nbr de commande)
catg	Le nombre des catégories des produits commandés
nbr_payment_type	Le nombre de types de paiements utilisés par le client
review_message_len	La longueur moyenne du commentaire /message

CAH+ MinMaxScaler



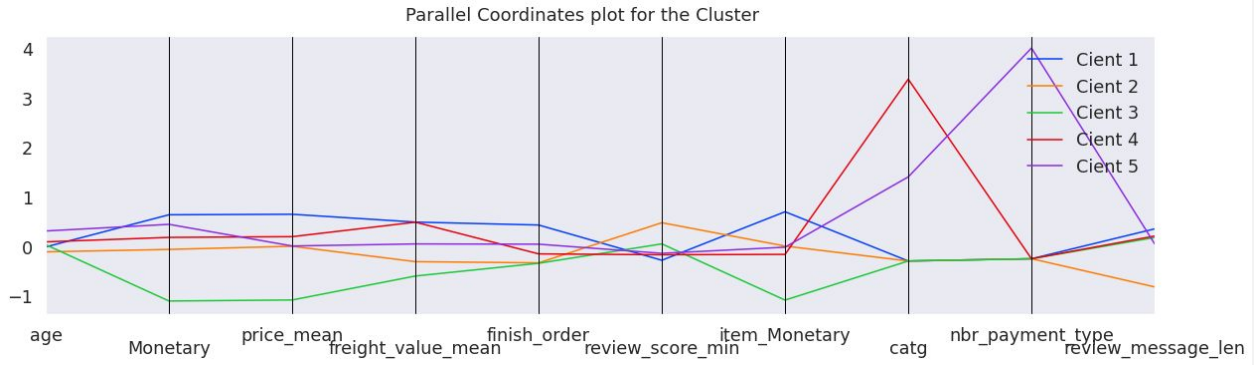
- Nombre de cluster **6**
- **Non** interprétable
- Non équilibrés

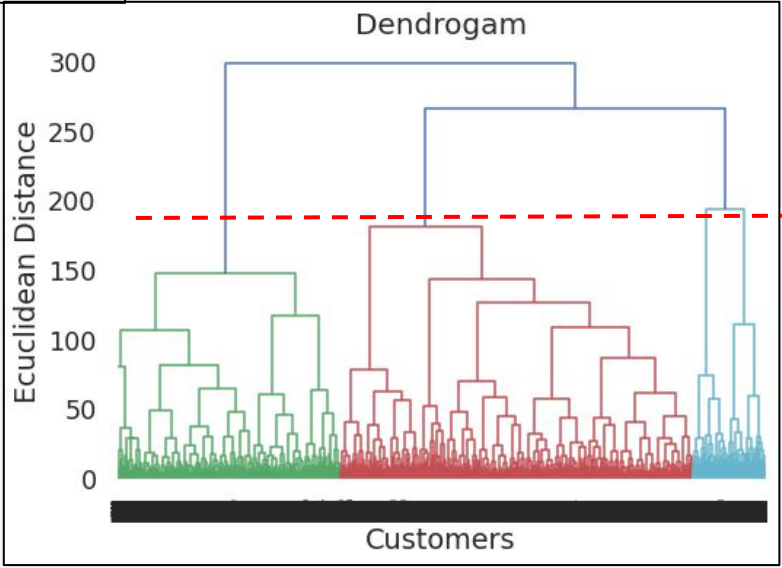


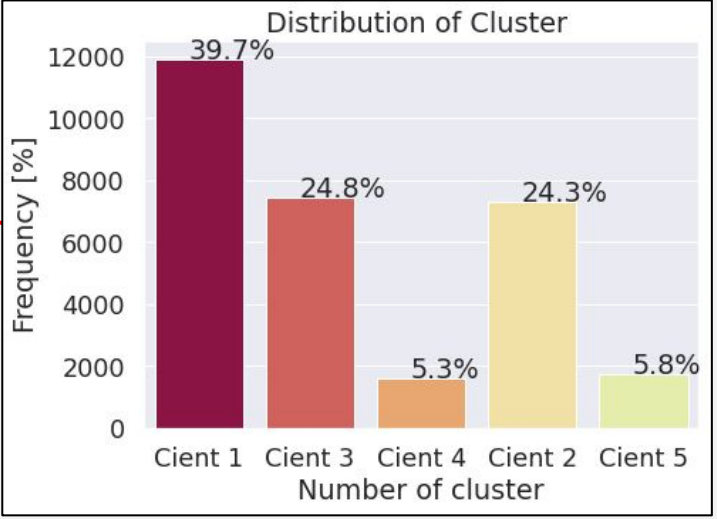


CAH+ StandardScaler

- Nombre de cluster **5**
- interprétable
- Non équilibré en nombre de client par cluster

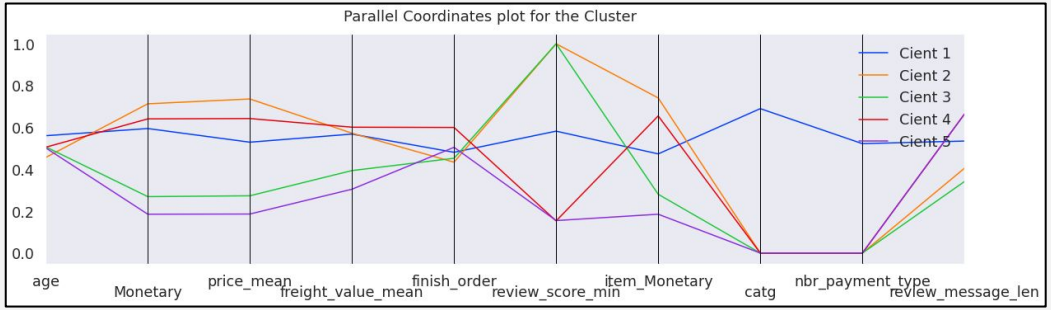


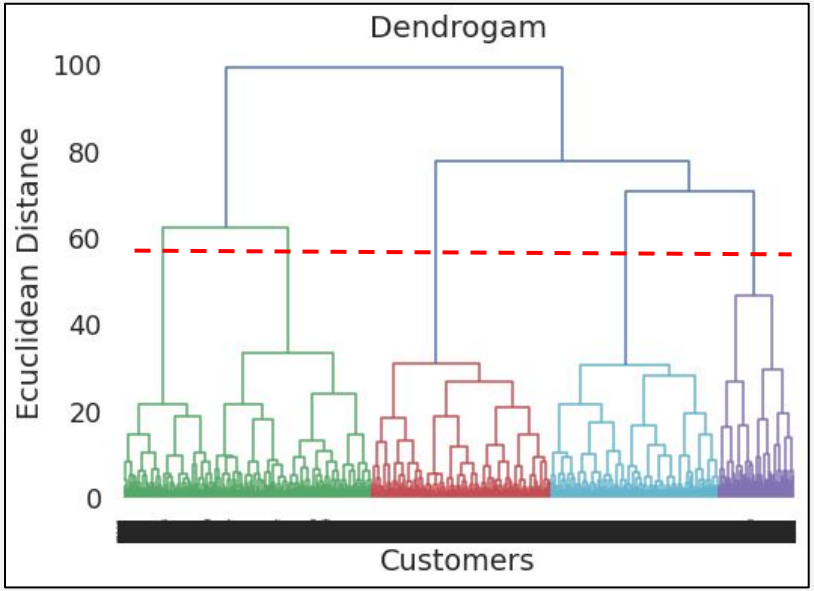


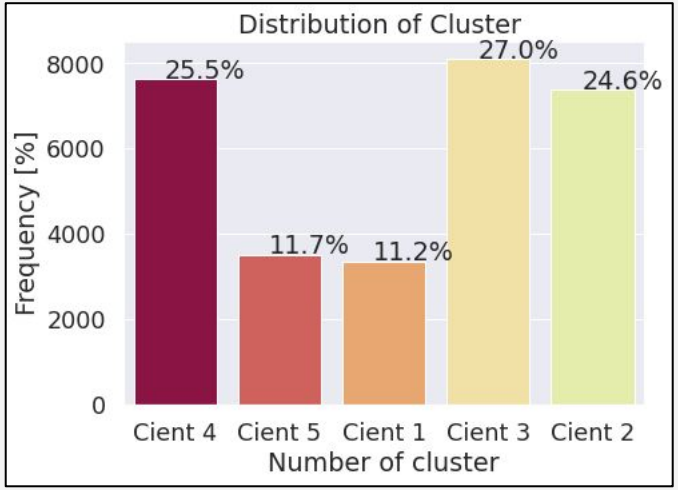


CAH+ QuantileTransformer

- Nombre de cluster **5**
- **interprétable**
-équilibré

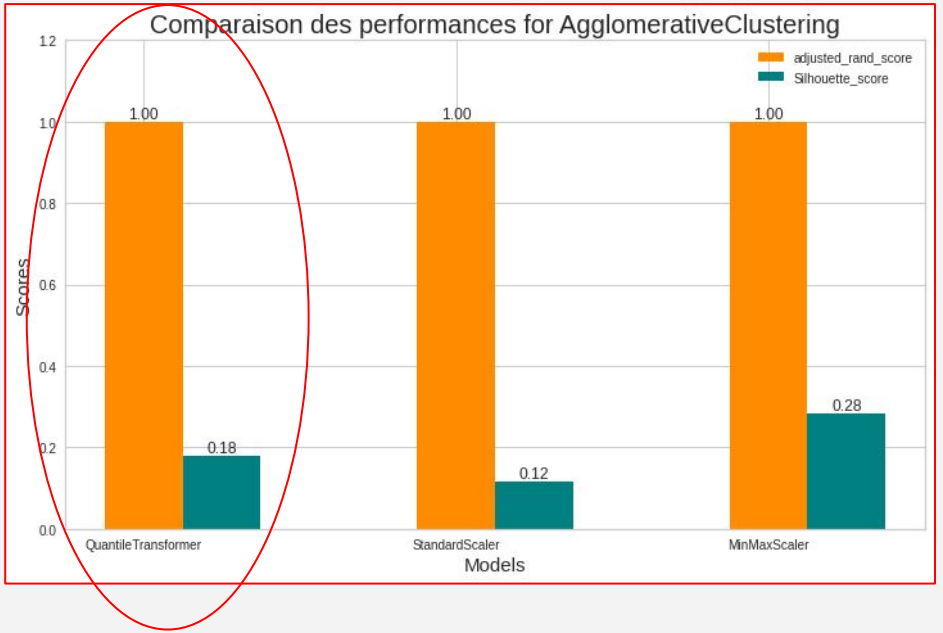


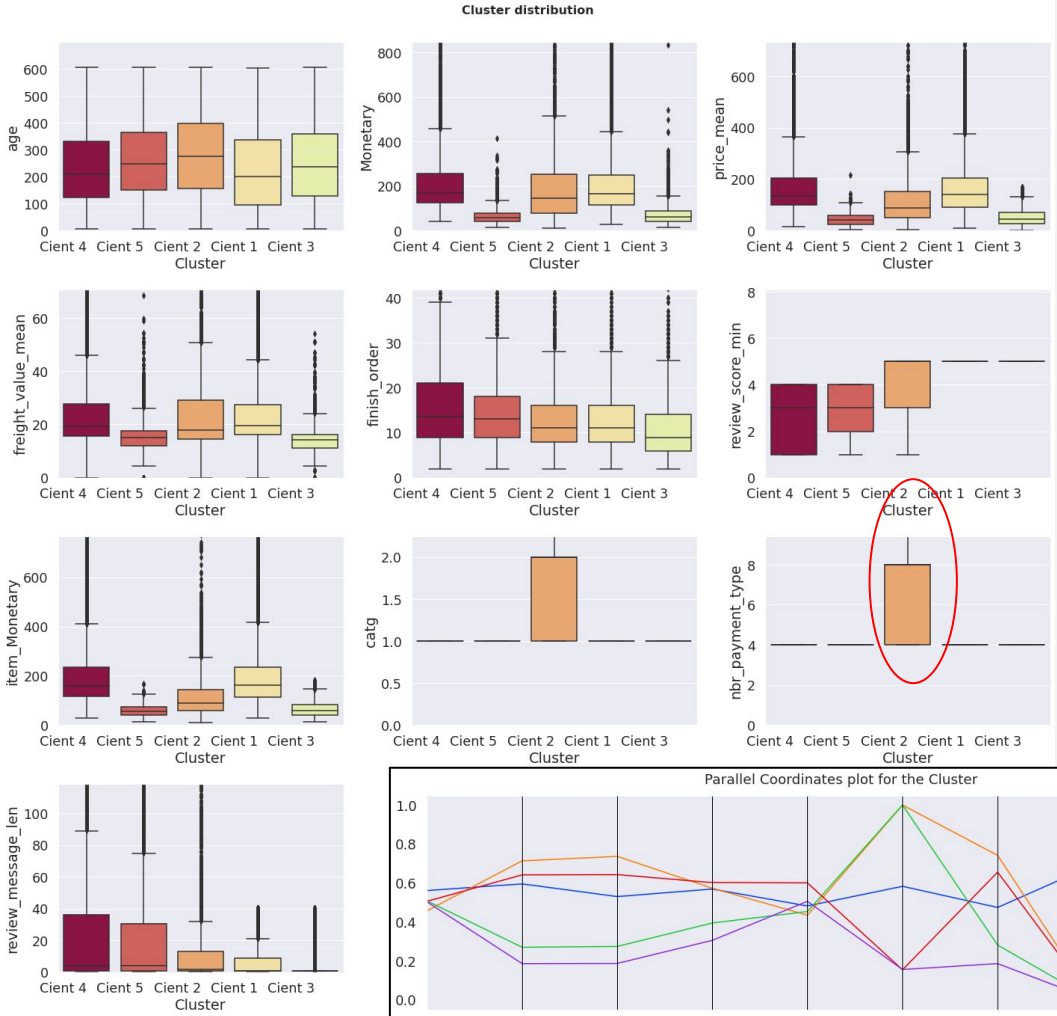




Stabilité des clusters

- les 3 modèles sont **stables** mais on retient le **QuantileTransformer** car il donne un meilleur score de silhouette à 0.18 et le mieu interprétable





INTERPRETATION

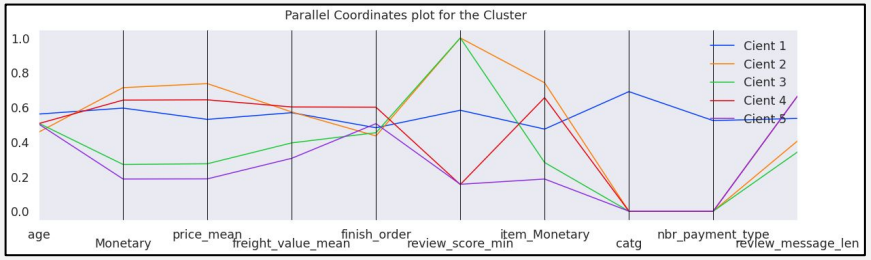
-Le client 4: sa commande a pris beaucoup de temps à être livré.

- Le client 5 et 3: les plus économes.

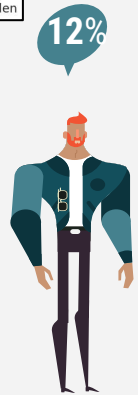
-Le client 3: achète plusieurs catg d'articles

-Le client 2 le plus anciens

-Le client 4 et 1 sont les plus dépensiers



PROFIL CLIENT CAH –



Fidèle

Les clients **anciens et réguliers** qui utilisent plusieurs **types de paiement** et achètent différentes **catégories** de produits.

Nouveau dépensier

Les clients **nouveaux** : font des achats pour des montants **élevés** et ils sont **contents**.

Econome Satisfait

Les clients plutôt **anciens** qui font des achats pour des montants **faibles** à **moyens** et ils sont **contents.**

Dépensier Insatisfait

Les clients qui sont plutôt anciens ,font des achats pour des montants relativement **élevés ,** achètent des produits assez volumineux et ils sont **mécontents** car le délais de la **livraison** était **long**.

Econome Insatisfait

Les clients plutôt **anciens**,font des achats pour des **montants faibles,** ils sont insatisfaits de leurs achats et ils ont laissé un commentaire assez long .

Comparaison

	Kmeans	OPTICS	CAH
Nombre de clusters	5	4	5
Taille de la dataframe	full	40 000	30 000
Stabilité	Bon	Bon	Bon
Interpretabilité	oui	oui	oui
Temps de calcul	rapide	long	long
Equilibre des clusters	Oui	Non	Non

Conclusions

Contrat de maintenance à **renouveler** tous les 6 mois .

Revoir les résultats avec l'équipe Marketing pour choisir le modèle qui répond aux mieux à leurs attentes.

Pistes d'amélioration:

Manque de précision de la part de l'équipe de marketing pour nous guider sur la partie de features selection .

Manque de plus de données sur les clients comme l'age , sexe..

Essayer d'autres modèles de classification