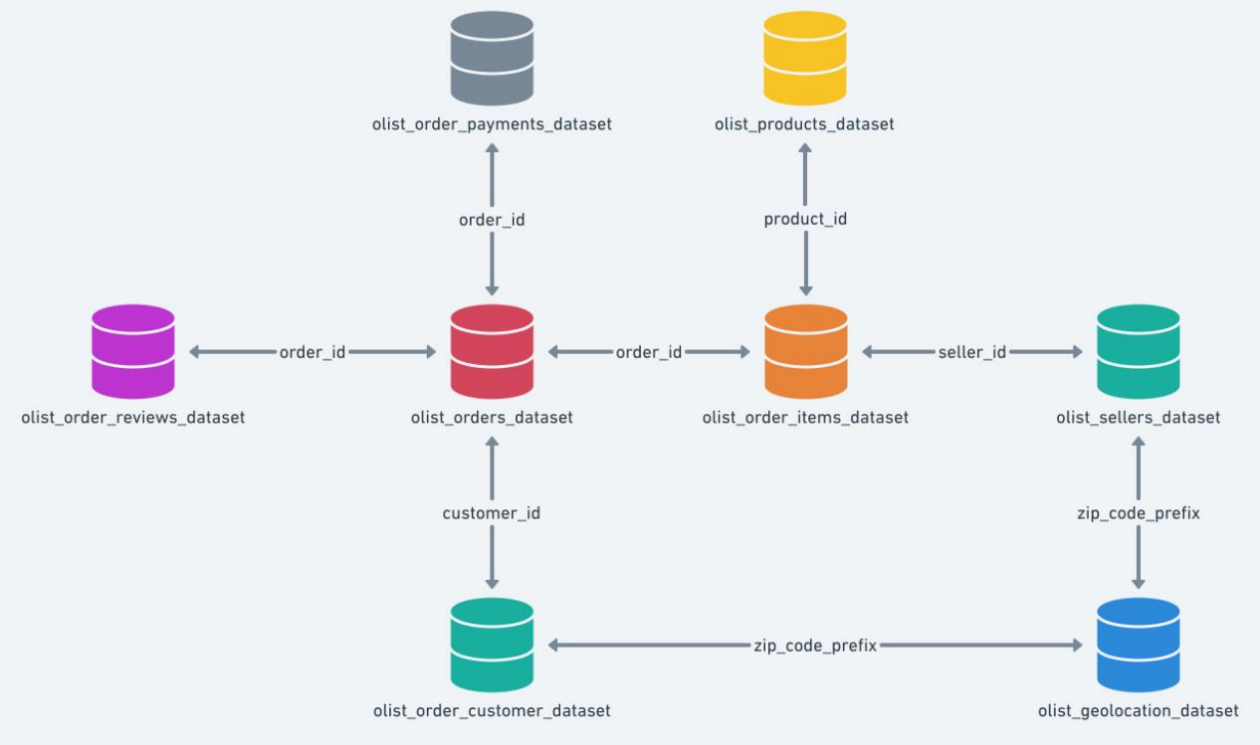
Réalisé par: Maidoumi Mohamed

10/12/2021

PROJET 5: Segmentation des clients de site de E-Commerce Olist.

Sommaire

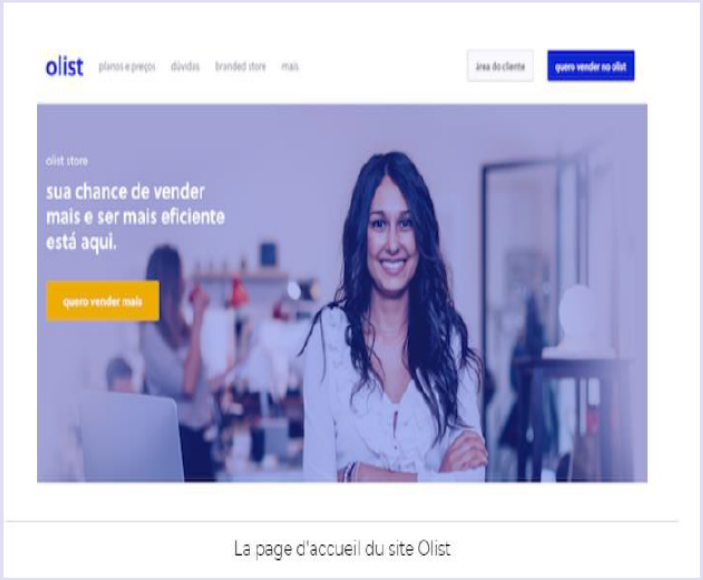
I- Chargement des données et description.



Architecture de la base de données:

Le projet et la mission

- **OLIST**: Solution de vente sur les marketplaces en ligne.
- **Segmenter les clients** pour campagnes de communication.
- **Méthodes non supervisée** des clients.
- **Comprendre** les différents types d'utilisateurs.



La page d'accueil du site Olist

OPENCLASSROOMS

Jeu de donnée

- une base de donnée anonymisée fourni par OLIST
- 9 fichiers Excel de données structurées: (116276 Clients, 41 Variables)

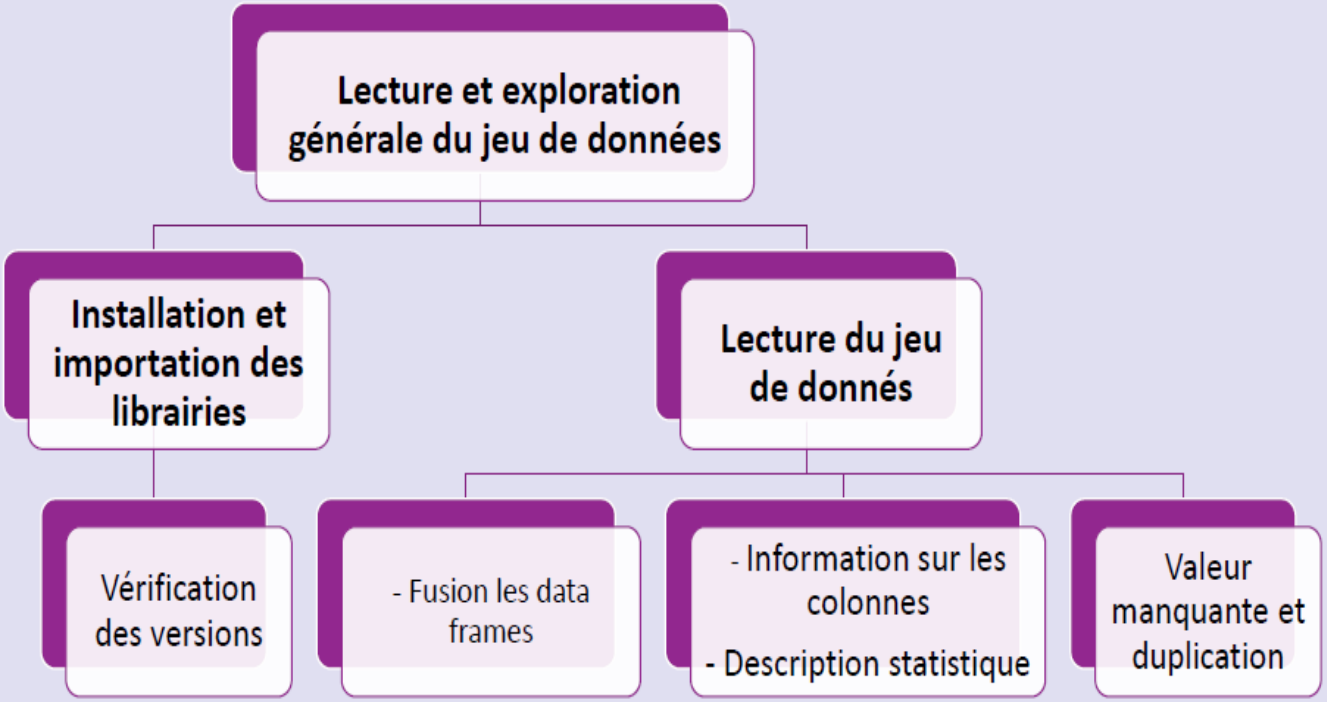
Avantages

Aspect business et Marketing :

- Améliorer la connaissance des clients
- Renseigner sur leur comportement
- Proposer des produits, inciter aux achats..

OPENCLASSROOMS

Nettoyage des données

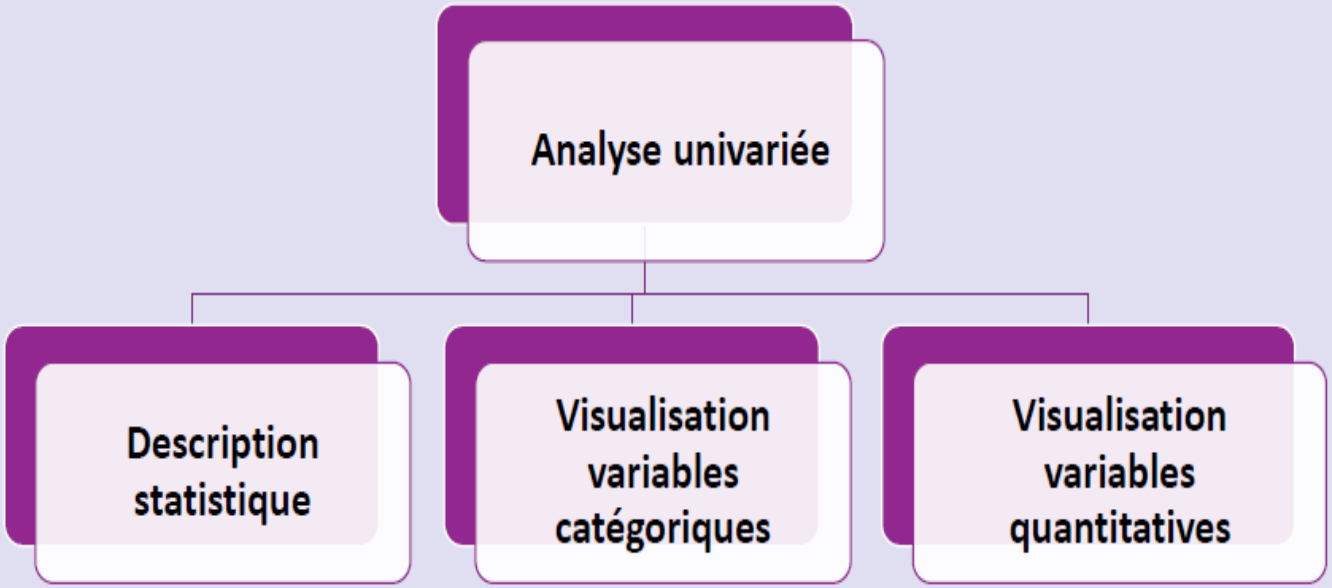


Analyse Exploratoire

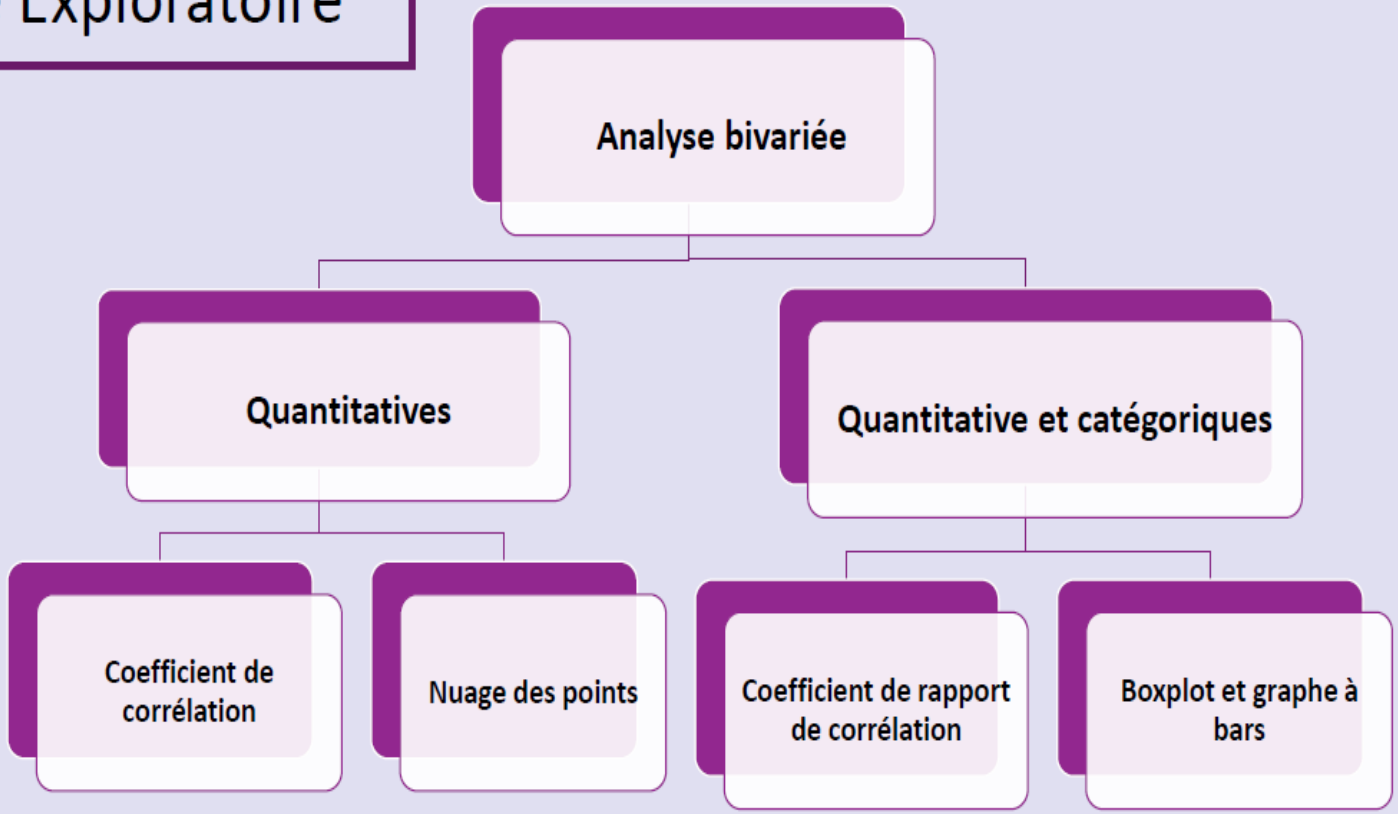
1. Analyse univariée

2. Analyse bivariée

Analyse Exploratoire



Analyse Exploratoire



Corrélation entre les variables créer et sélectionnées:

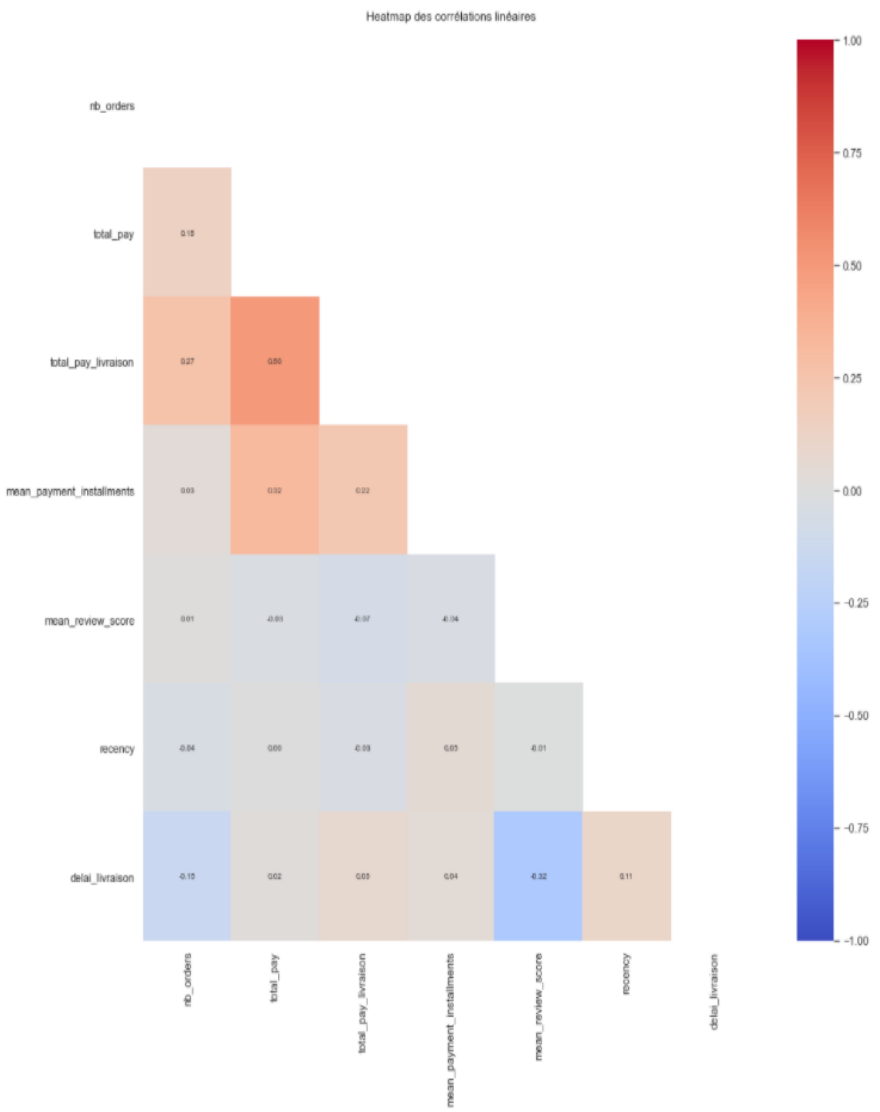
Le total payé pour livraison par clients est corrélé avec le total payé et total livraison et les echeance.

On remarque que le score de review est inversement corrélé au délai de livraison.

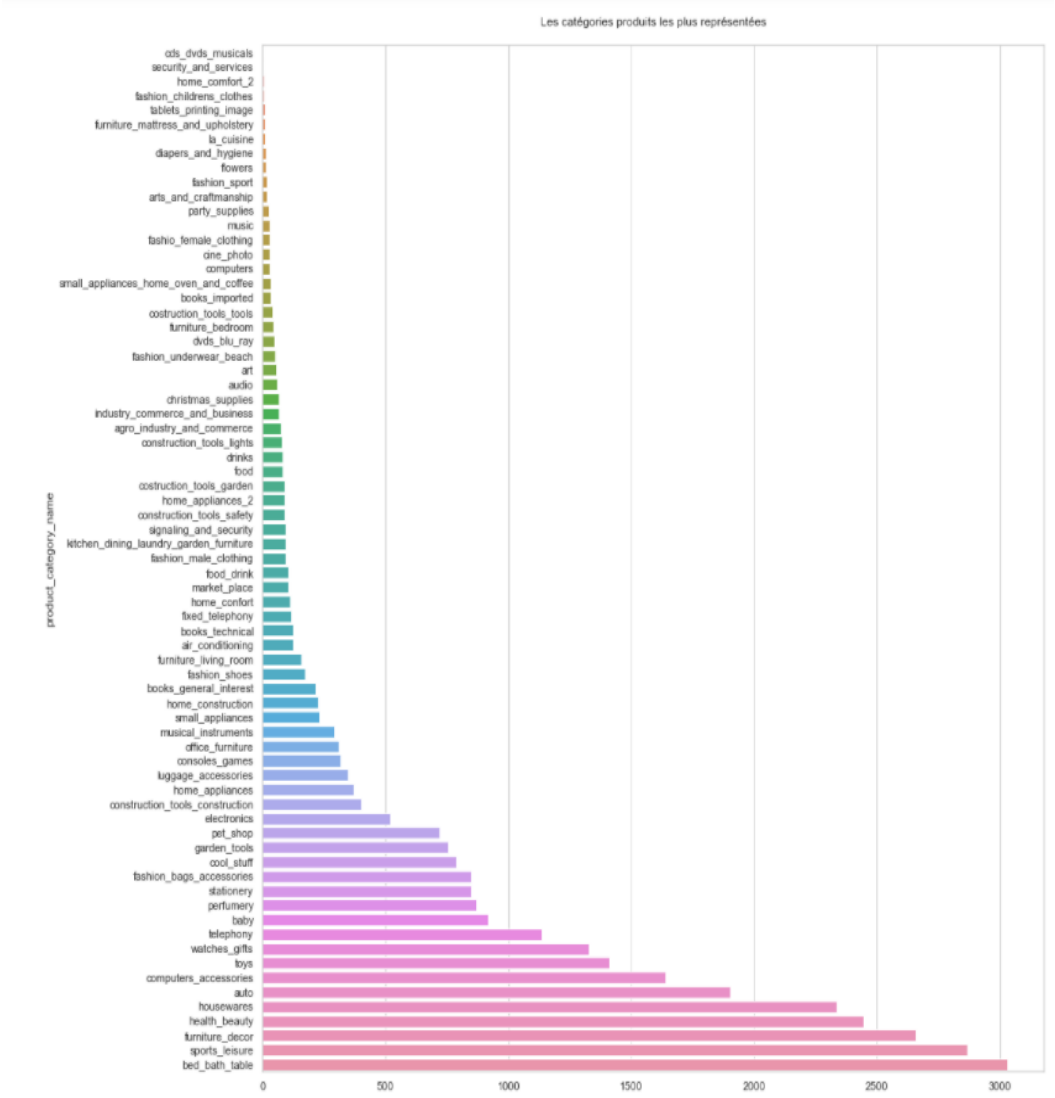
Le nombre de moyens de payement est positivement corrélé au montant totale payé sur la plateforme.

le nombre d echeance est positivement corrélé avec total payé et total livraison

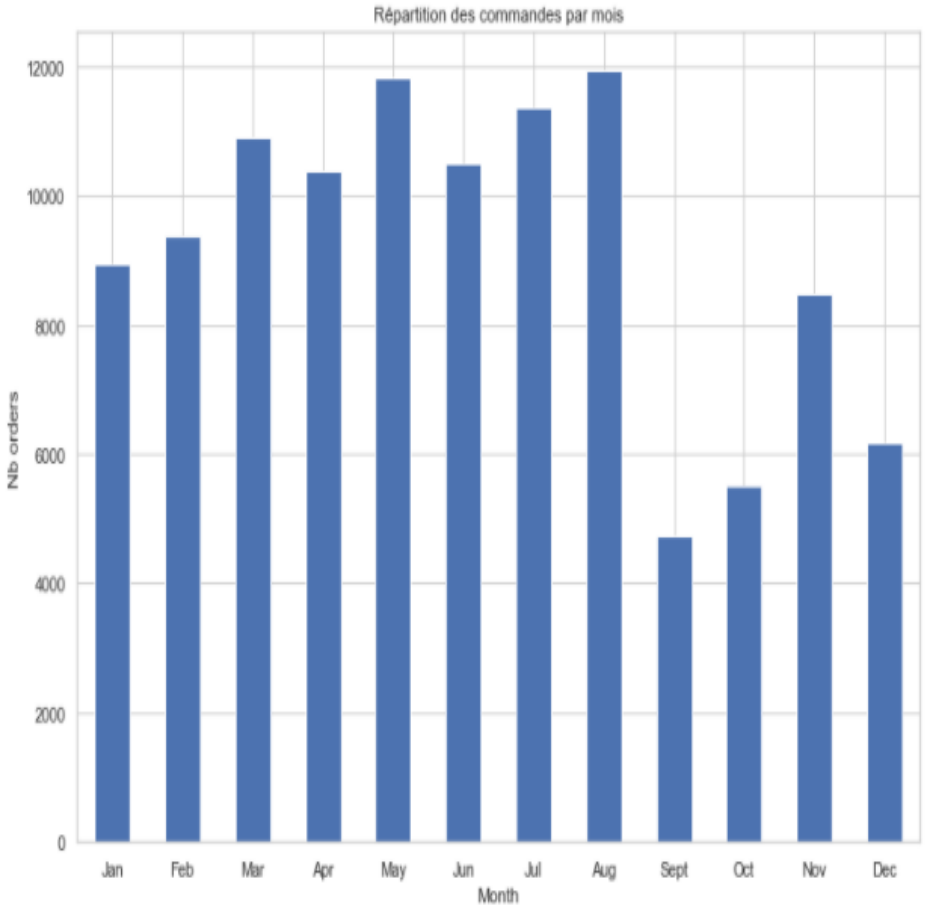
nombre de commande est negativement corrélé à la période entre les commandes passées (donc positivement corrélé à la frequence des commandes passées).



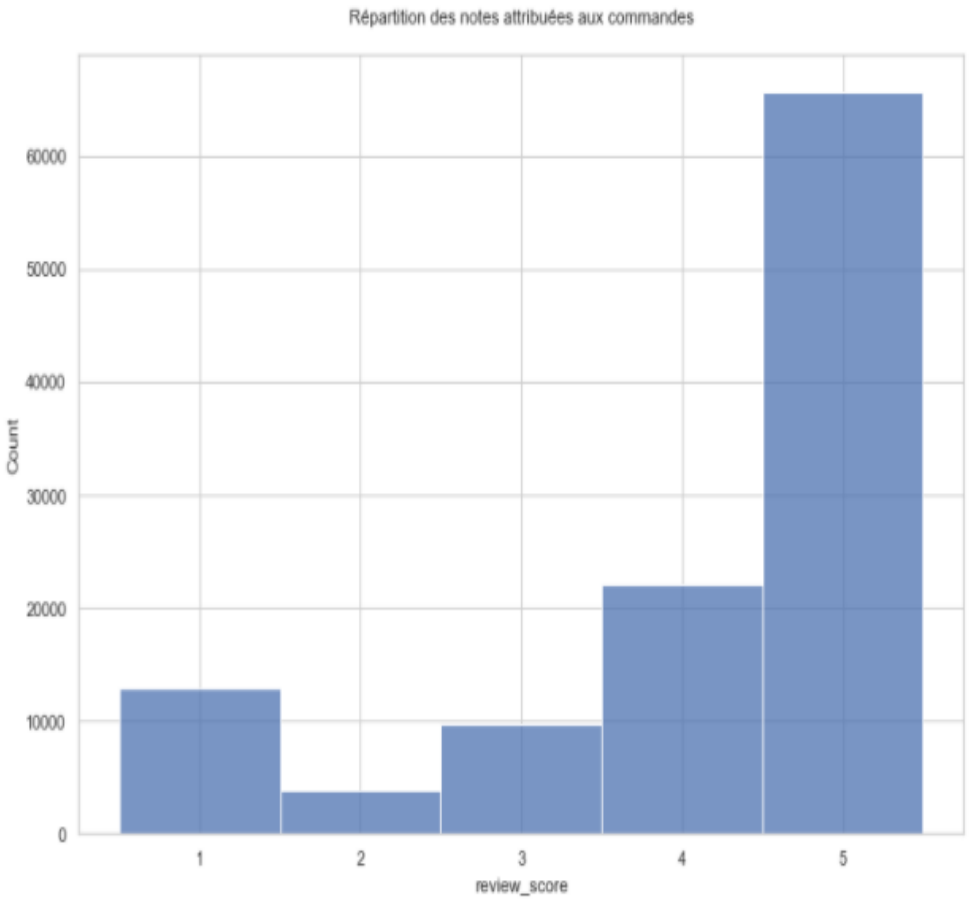
Présence de Catégories de produits dans la base :



Répartition des commandes par mois dans la base de données:

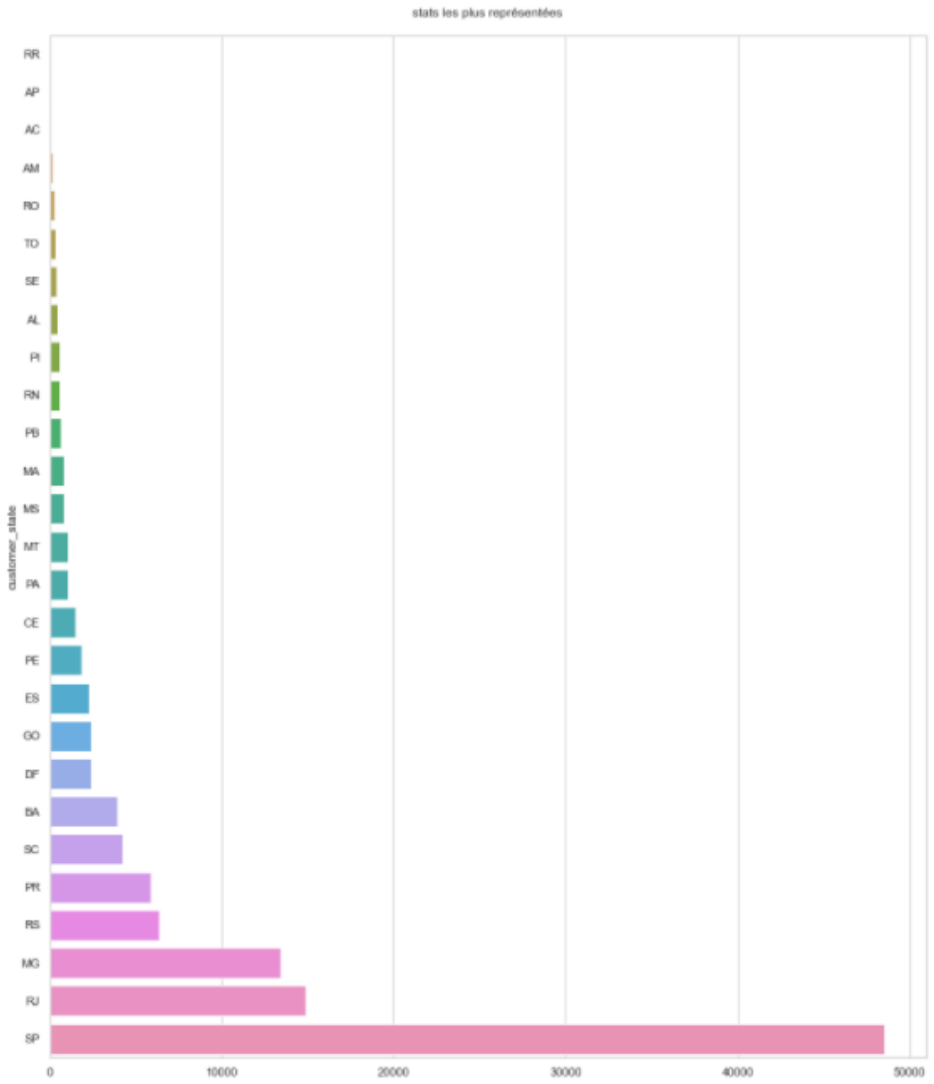


Répartition des notes attribués aux commandes dans la base mergée:



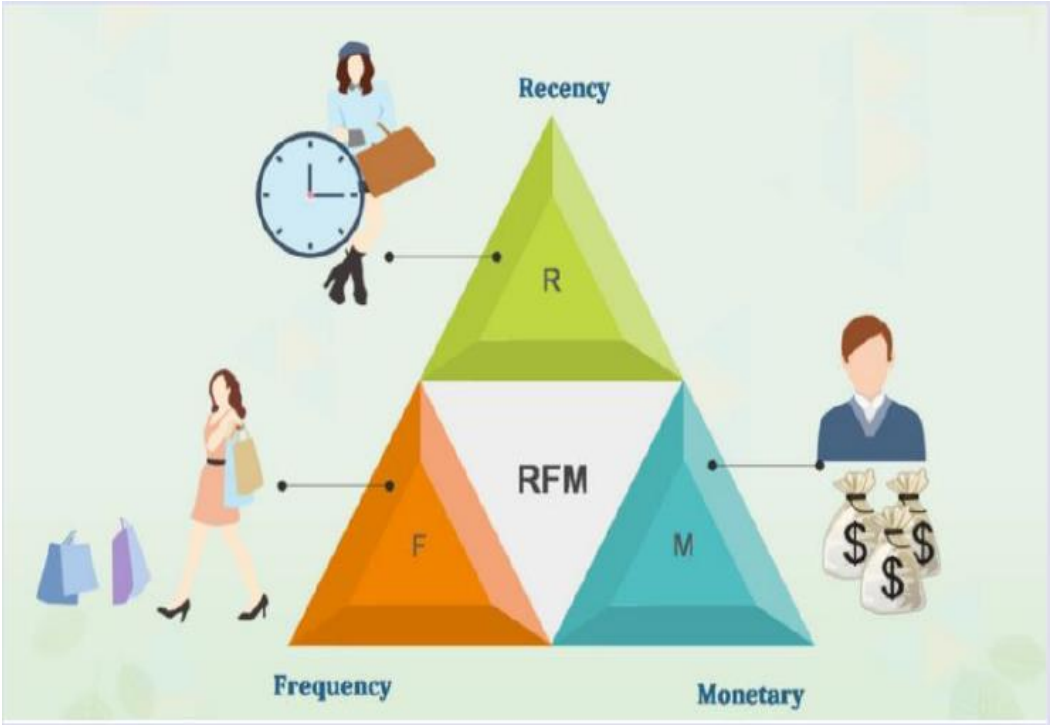
Etats des clients présent dans la base de données:

Les Etats Sao Palo et Rio de Janeiro et Minas Gerais sont les plus présents.



III- PRÉPARATION DU JEU DE DONNÉES/Feature engineering.

Méthodologie de RFM:



Nombre de commande passées.

Delai de livraison: Différence entre la date de commande et la date de livraison en jours.

Rencency : Différence entre la dernière date de mise à jour de la base et la date de la dernière commande par client.

Nombre de mode de payement utilisés.

Note moyenne attribuée par le client sur l'ensemble de ses commandes.

Délai de livraison moyen par client.

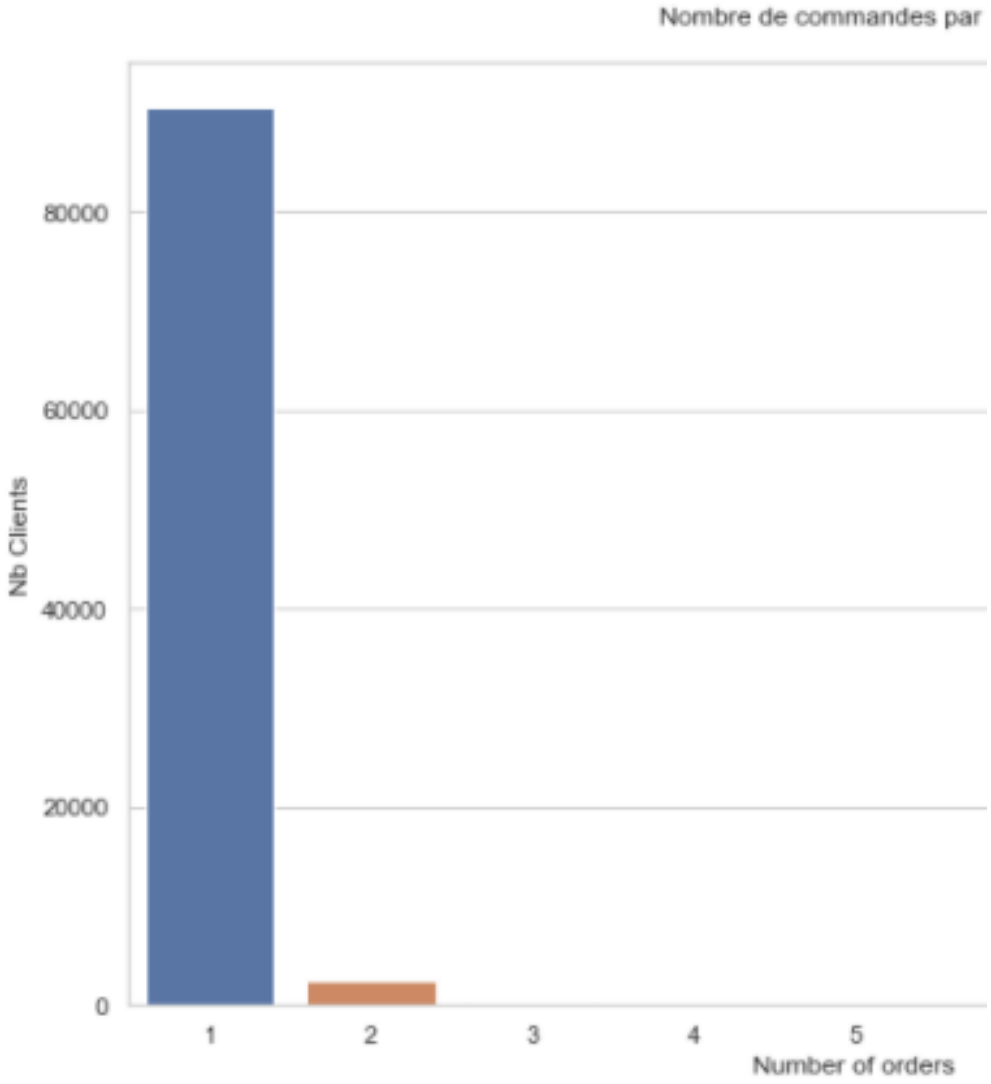
Total payé par client.

Total payement de livraison par client.

Variables crées:

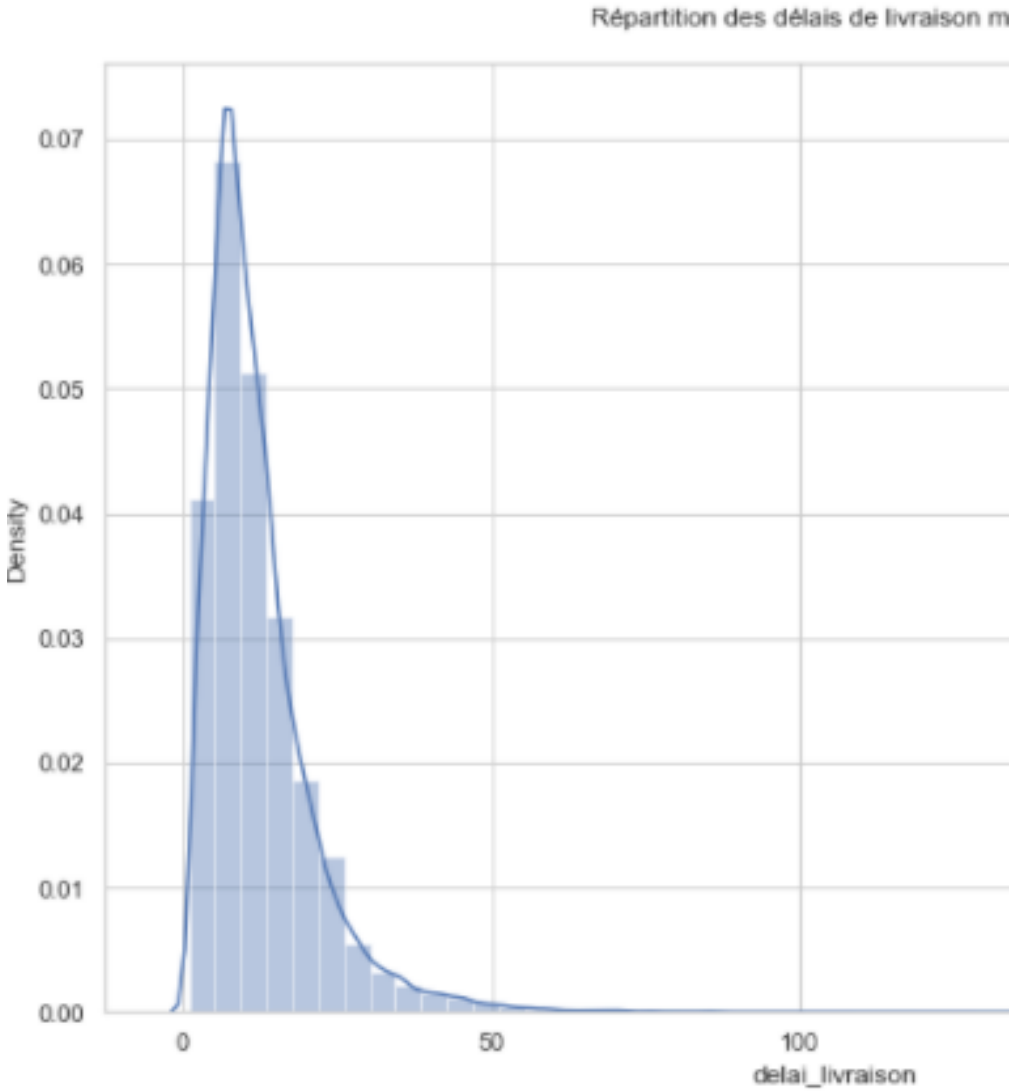
Nombre de commandes par client dans la base de données:

La majeure partie des clients n'ont passé qu'une seule commande. Il sera donc compliqué d'établir un classement de leur catégorie produit préférée



Répartition des délais de livraison dans la base de données:

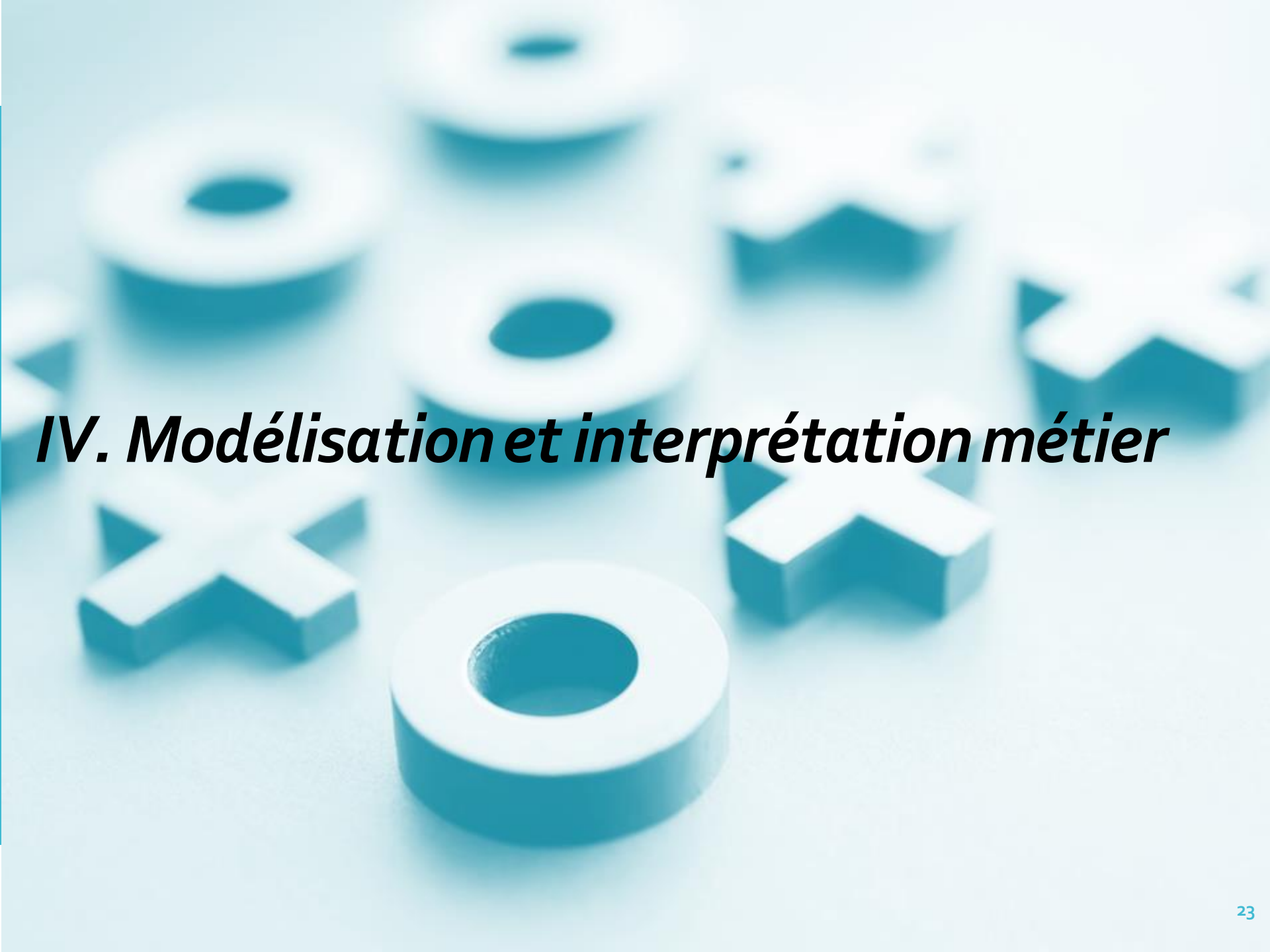
Délai de livraison suit une lois normale asymétrie avec une moyenne de 10 jours.



Variable sélectionnée pour la modélisation:

```
In [11]:  1  data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 112639 entries, 871766c5855e863f6eccc05f988b23cb to cd76a
Data columns (total 16 columns):
 #   Column                    Non-Null Count   Dtype
---  ------                    --------------   -----
 0   order_id                  112639 non-null  object
 1   order_item_id             112639 non-null  int64
 2   product_id                112639 non-null  object
 3   price                     112639 non-null  float64
 4   freight_value             112639 non-null  float64
 5   customer_id               112639 non-null  object
 6   order_status              112639 non-null  object
 7   order_purchase_timestamp  112639 non-null  object
 8   sum_payment_installments  112636 non-null  float64
 9   payment_type              112639 non-null  int32
 10  review_score              111792 non-null  float64
 11  customer_zip_code_prefix  112639 non-null  int64
 12  customer_city             112639 non-null  object
 13  product_category_name     111038 non-null  object
 14  delai_livraison           112631 non-null  float64
 15  product_category          112639 non-null  int32
dtypes: float64(5), int32(2), int64(2), object(7)
memory usage: 13.7+ MB
```



IV. Modélisation et interprétation métier

Catégorisation des variables non numériques.

Créer une pipeline qui effectue:

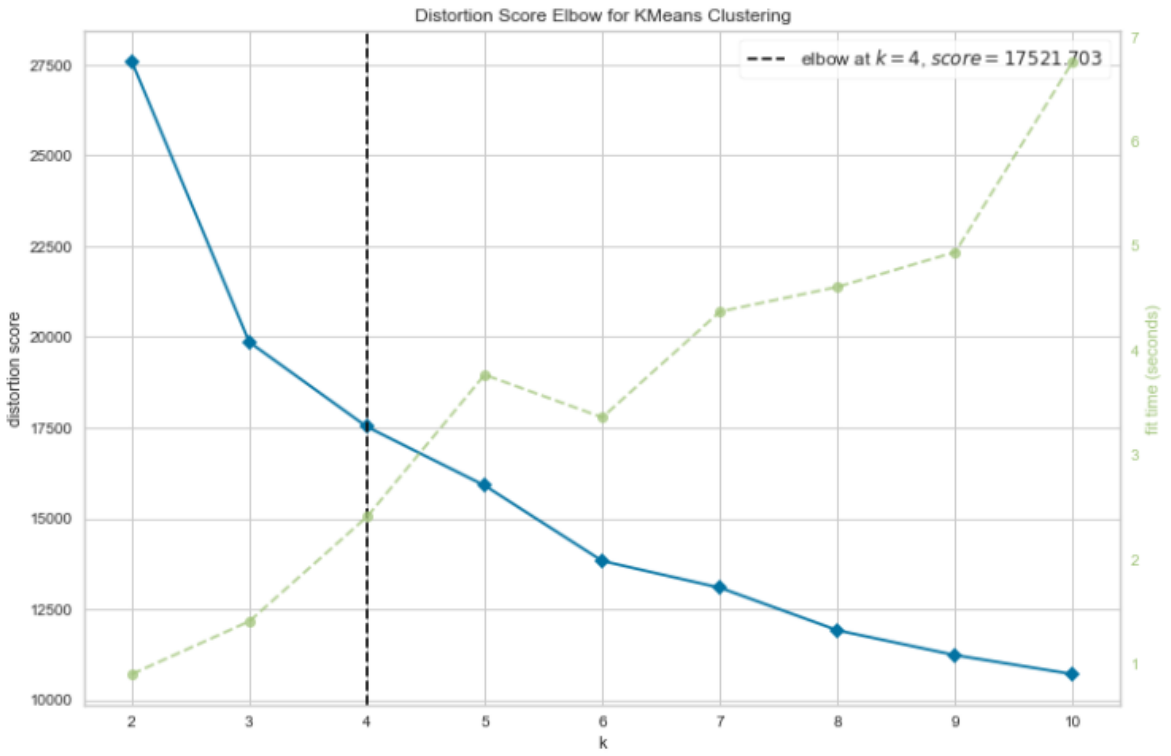
- Standarisation des variables.
- Clustering Kmeans-DBscan.

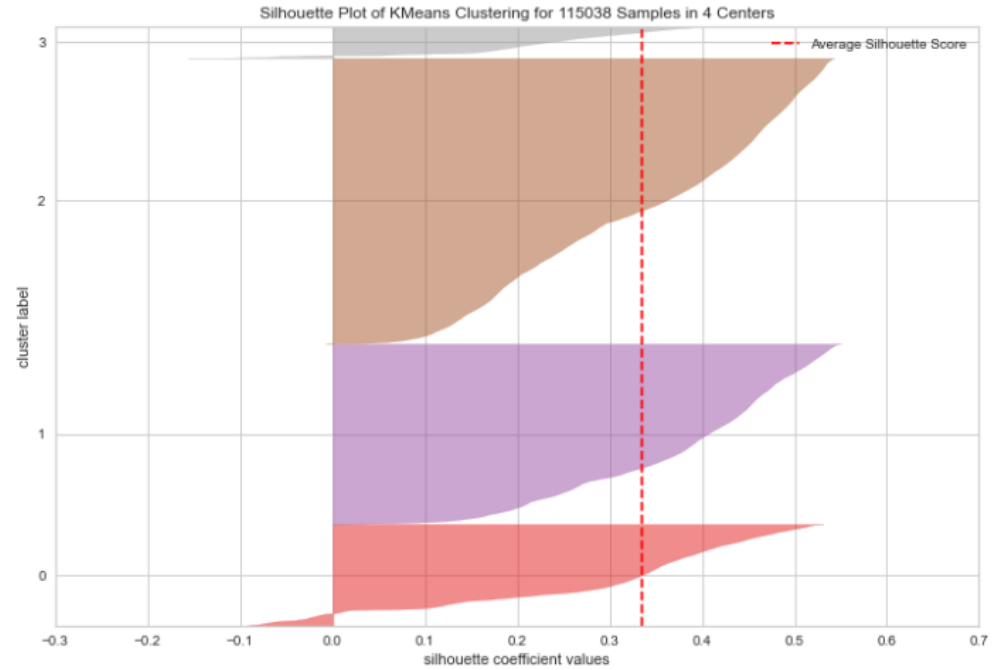
Modéliser avec une PCA.

Démarche de modélisation

Modélisation sans PCA: La méthode du coude et exploration du K optimal:

Grâce à la méthode du coude basée sur le score de distortion (somme moyenne des carrés des distances aux centres), une segementation en K=4 clusters serait la meilleure option

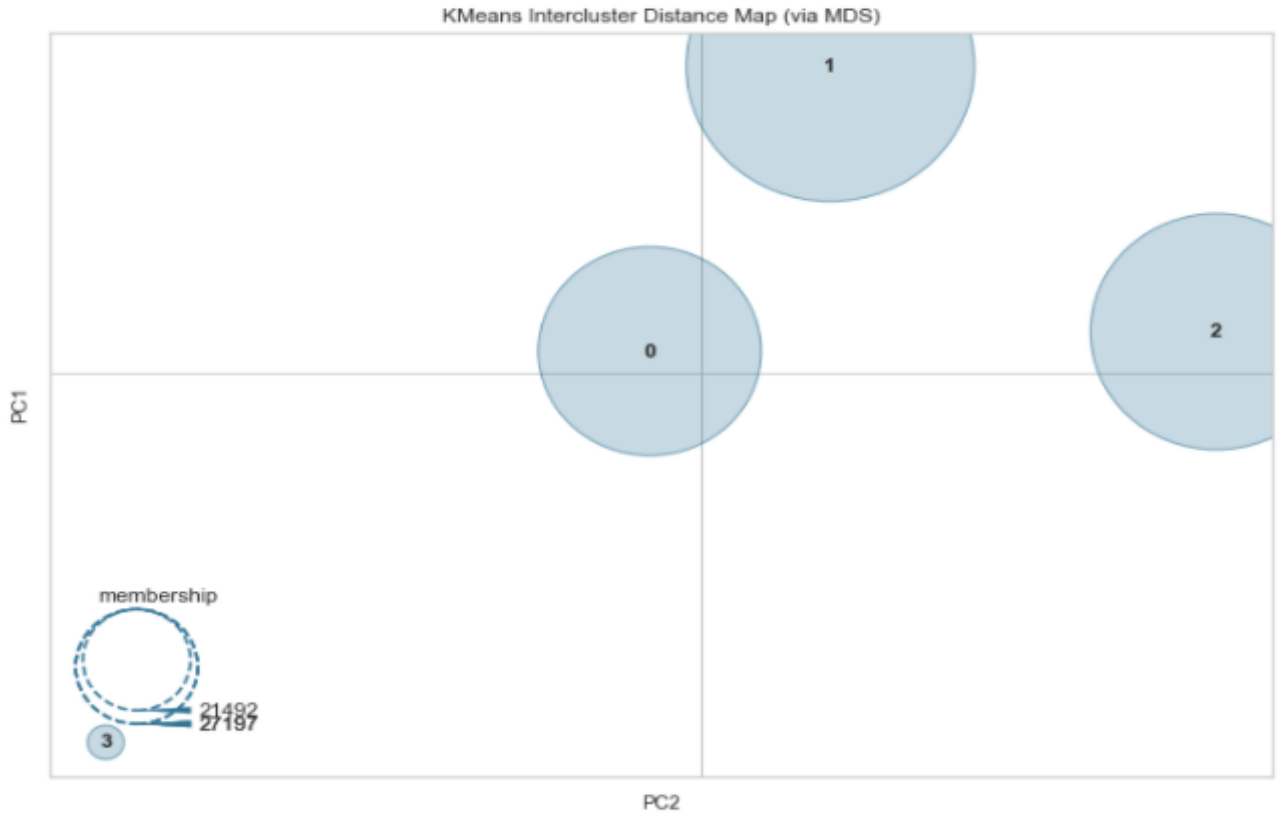




Coefficiant de Silhouette

Le score de chaque échantillon est calculé en faisant la moyenne du coefficient de silhouette (différence entre la distance moyenne intra-cluster et la distance moyenne du cluster le plus proche pour chaque échantillon), normalisée par la valeur maximale. Cela nous donne un score entre -1 et 1, qui nous permet de déterminer si la séparation est efficace ou si les points sont assignés au mauvais cluster.

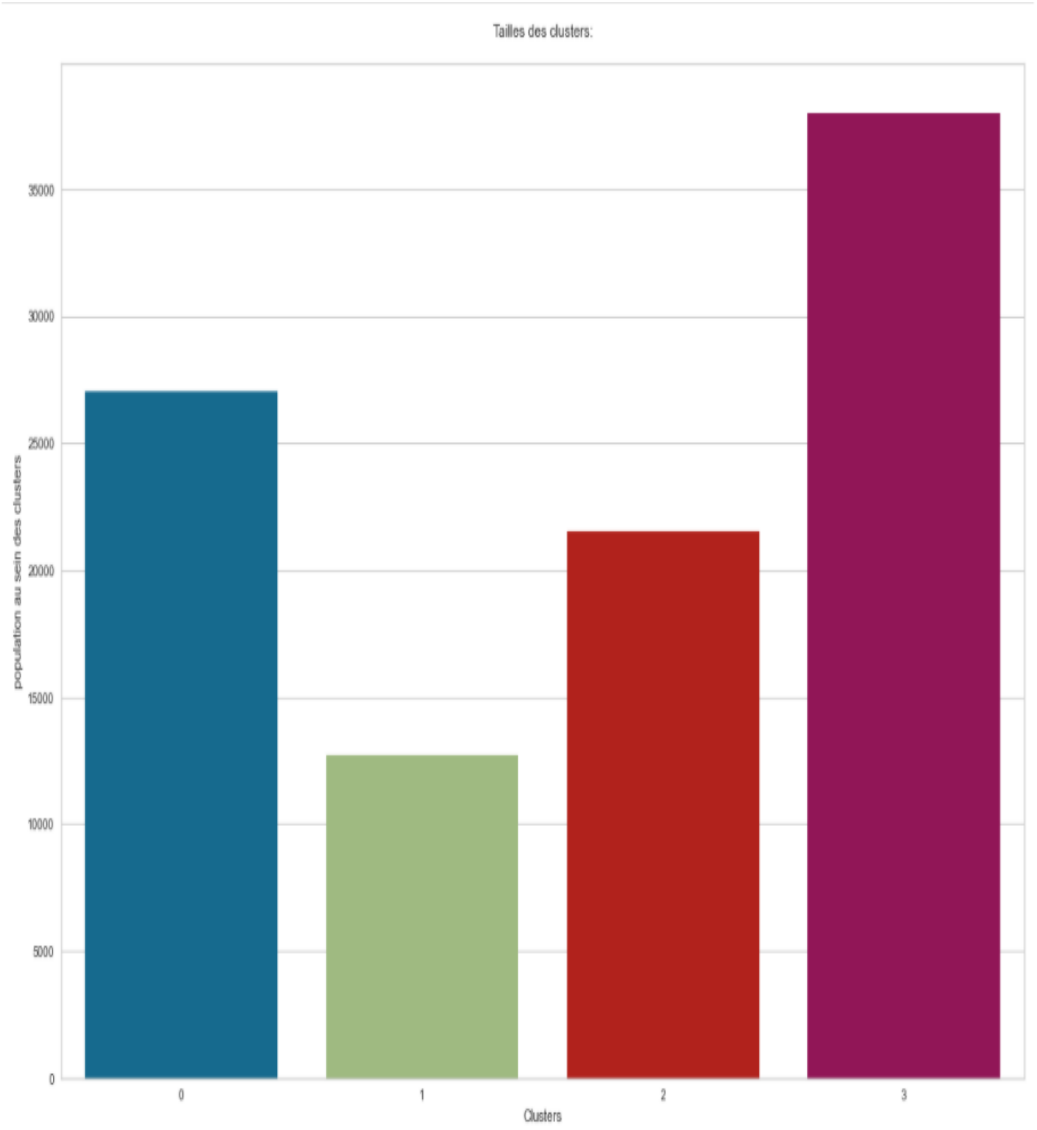
Ici, les clusters semblent relativement bien répartis et les séparations sont claires avec cependant quelques erreurs sur l'un des clusters.



Distance interclasses sur les deux prmière composantes principales:

Les clusters semblent très bien séparés sur le prmier plan factorielle (réduction de dimension MDS).

Tailles des clusters:



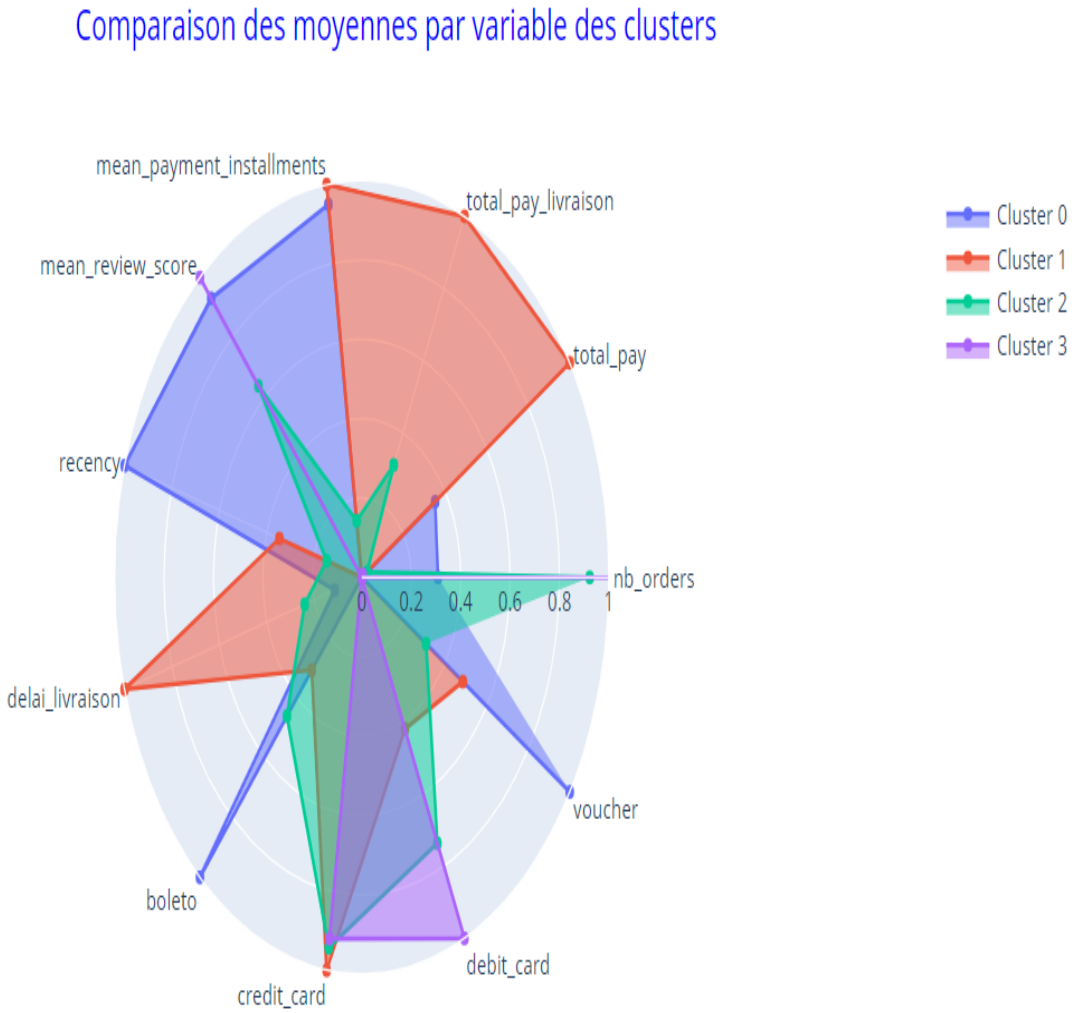
Interprétation métier des clusters

Cluster 0: les clients qui attribut une bonne note et qui achète avec un montant grand, utilisent plusieurs mode de payements- tendance a espacer leurs achats. * mode de payement: voucher.

Cluster 1: delai de livraison grand, total_payé petit, un review score defavorable * bed_bath_table et moins

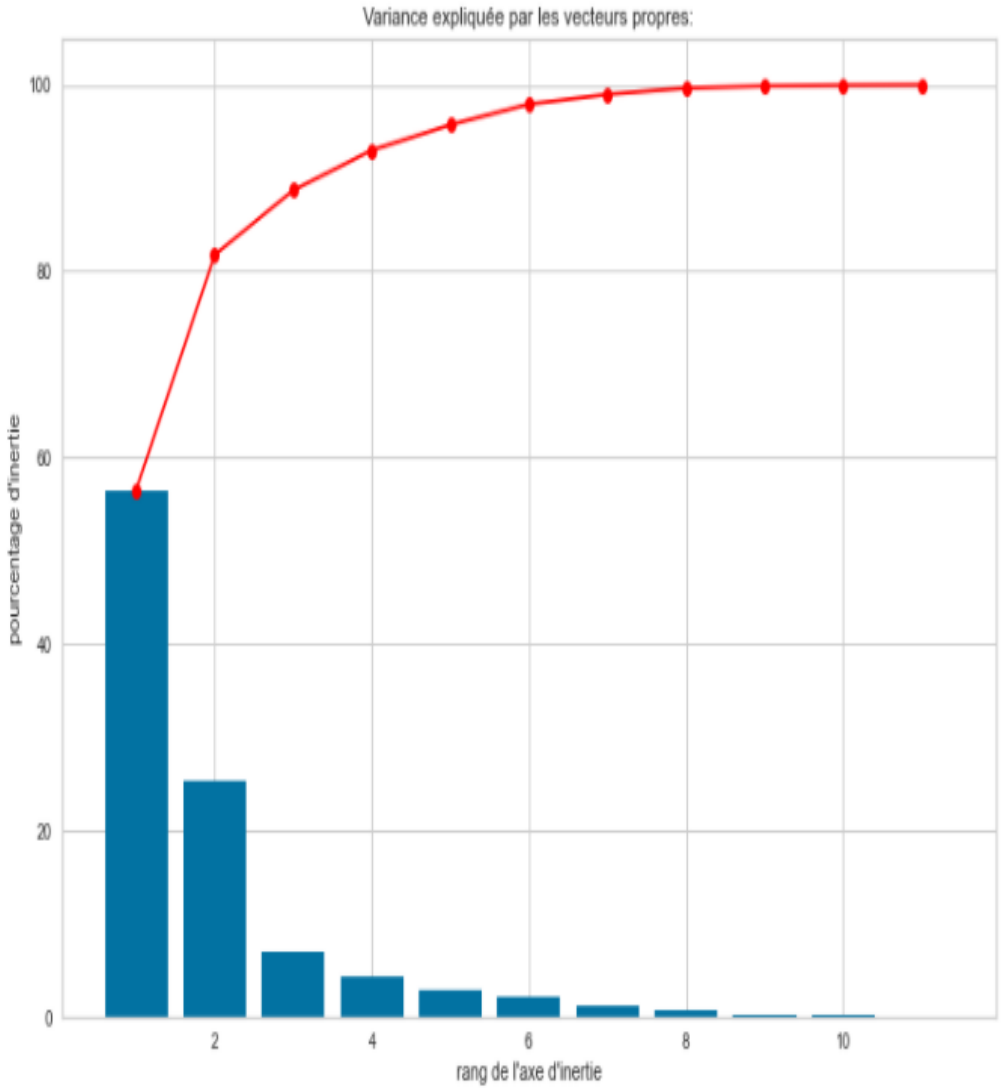
Cluster 2: un grand nombre de commande passé ils ont tendance a espacé leurs commandes, un total payé assez elevé, plusieurs methodes de payements, plusieurs echeances , un bonne review score

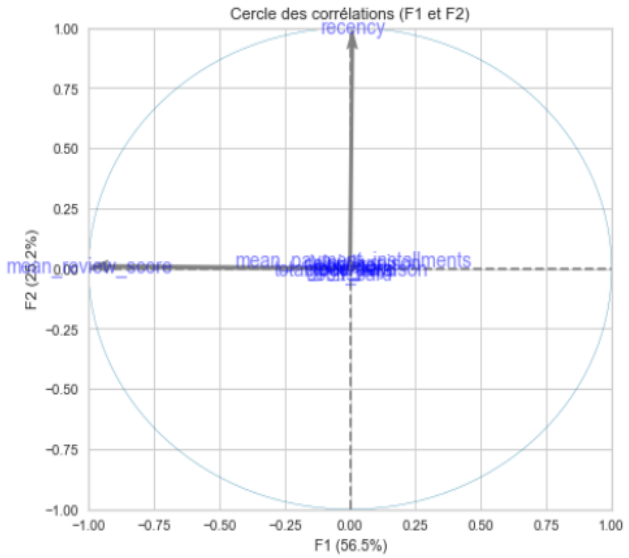
Cluster 3: Regroupe les clients qui utilisent plusieurs moyens de paiement et un nombre important d'échéances. Ils ont tendence à espacer les délais entre 2 commandes. Les avis de ces clients sont également très bons.

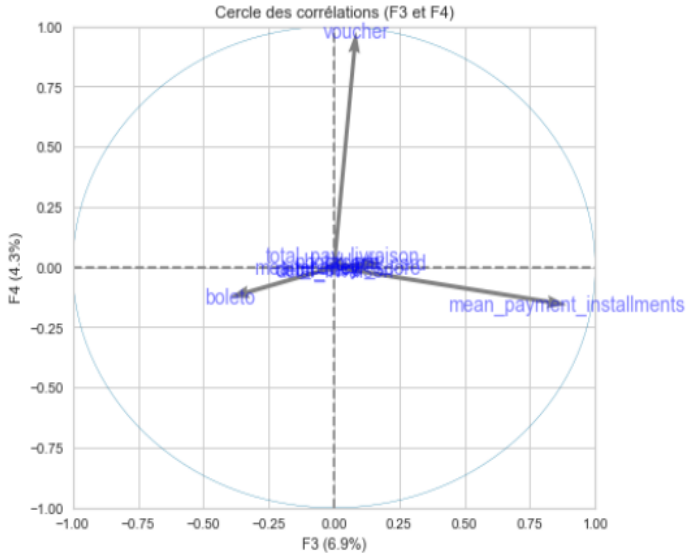


Réduction dimentionnelle :Variance expliquées par les vecteurs propre.

4 variables peuvent expliquer 95% de l inertie







Cercle de corrélation:

On peut ainsi voir parfaitement les variables qui contribuent le plus à chaque axe:

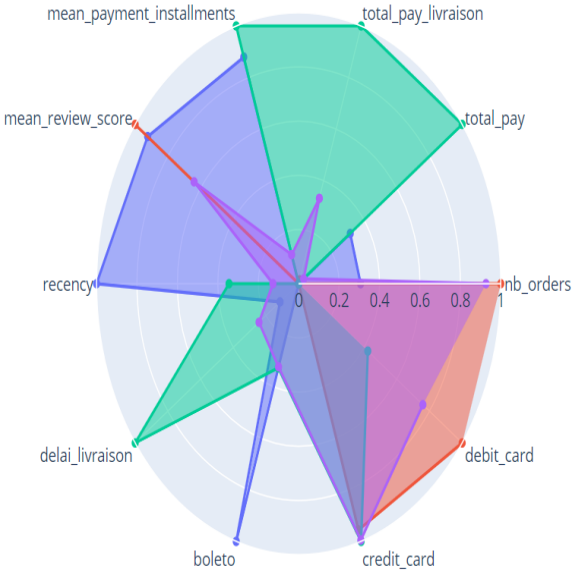
* F1 qui represente principalement la récence des commandes.

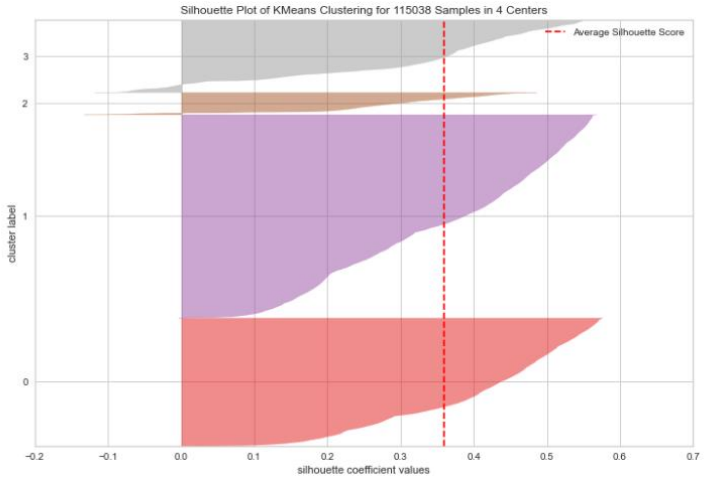
* F2 représentera principalement les review score.

* F3 : voucher.

* F4: positivement le nombre de mode payement utilisé, et négativement par le mode de payement boleto.

Comparaison des moyennes par variable des clusters

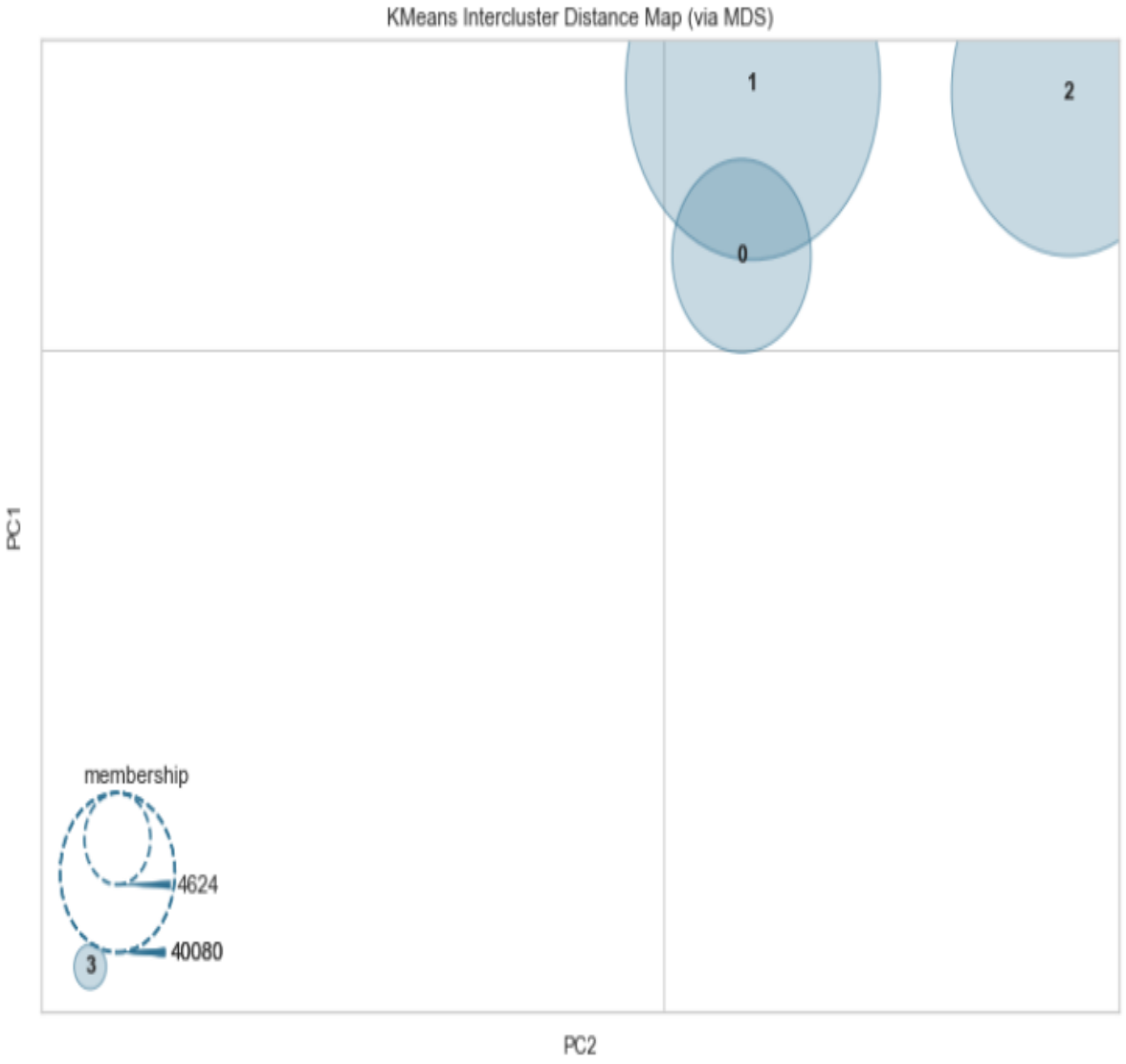




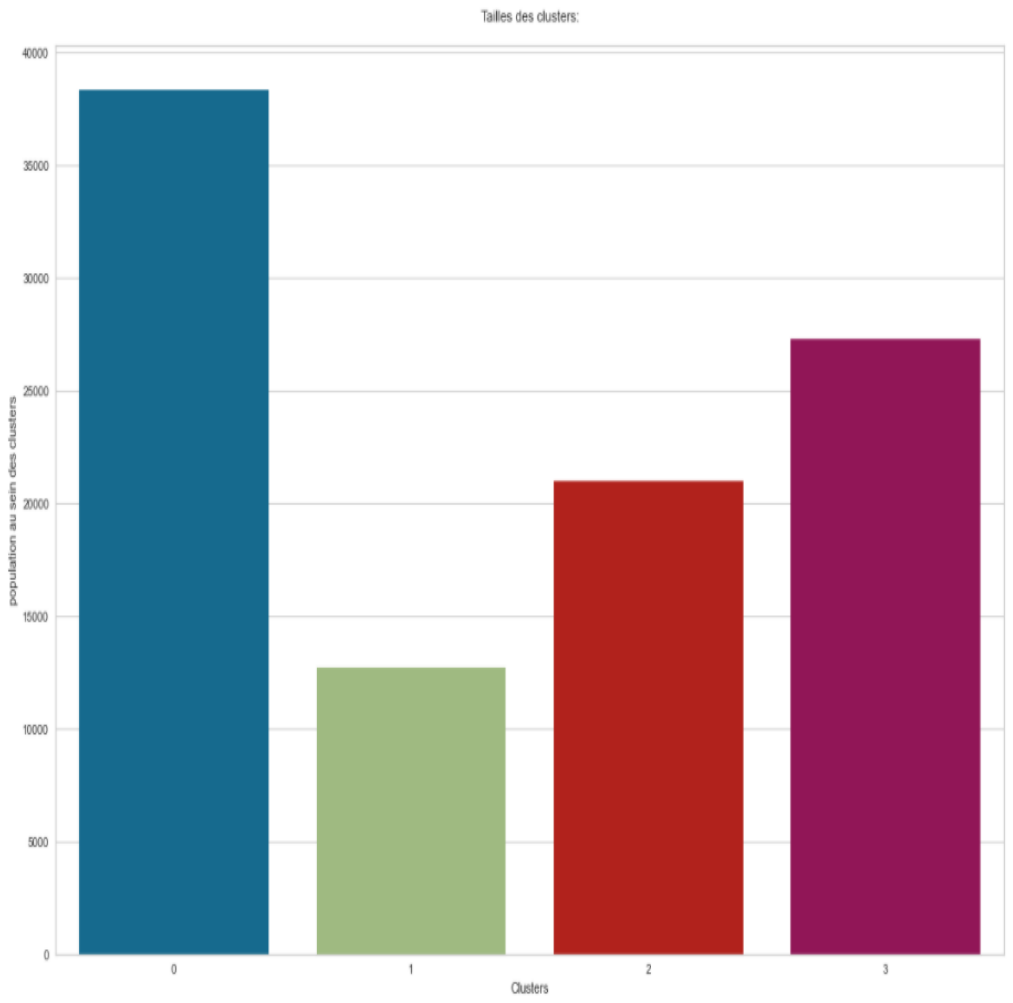
K-Means après réduction de dimenssions

Distance inter-classes avec PCA:

Les Clusters sont moins séparés par rapport à la version originale.



Tailles des clusters après PCA

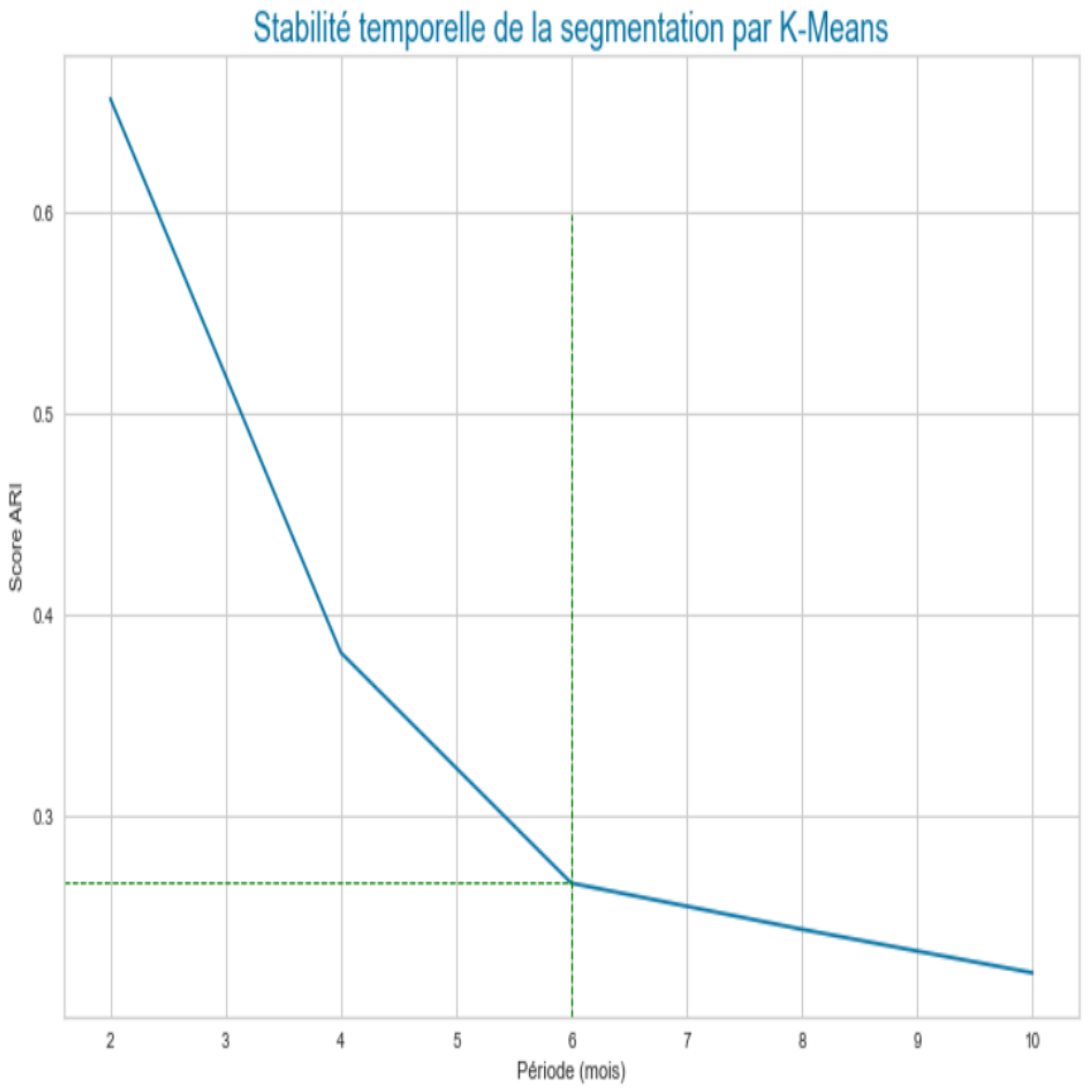


Stabilité temporelle du modèle

Le score ARI en fonction du score ARI entre les data périodisés et le data initial :

scores ARI obtenus sur les itérations par période de 2 mois, on remarque une forte inflexion après 6 mois sur les clients initiaux.

Il faudra donc prévoir la maintenance du programme de segmentation tous les 6 mois dans un premier temps puis re-tester cette stabilité temporelle au fil du temps afin de l'affiner. Il sera donc nécessaire de redéfinir les segments clients à chaque maintenance.



Conclusion: