


The background features abstract geometric shapes in various shades of blue. On the left, a solid light blue triangle points towards the center. On the right, a complex arrangement of overlapping triangles in different blue tones (light, medium, and dark) creates a dynamic, layered effect. The central text is positioned in the white space between these blue elements.

Segmentation d'un site de e-commerce

Sommaire :

- ▶ Rappel de la problématique
- ▶ Analyse exploratoire et Feature Engineering
- ▶ Modélisations
 - ▶ Kmeans
 - ▶ DBScan
- ▶ Modèle retenu
- ▶ Maintenance et fréquence de mise à jour



Rappel de la problématique

Rappel de la problématique :



- ▶ Olist, marketplaces en ligne
- ▶ Segmentation de clients
- ▶ Comprendre les différents types d'utilisateurs, définir des profils
- ▶ Mise en place d'algorithmes non supervisés de clustering
 - ▶ Kmeans
 - ▶ Dbscan
 - ▶ Hierarchique

Descriptif du jeu de données :



Payments

Type, Montant,
Plusieurs fois



Products

Catégories, Taille/poids,
Description, Photos



Reviews

Notes
Commentaires



Items

Détails produit



Orders

Infos commande,
date commande, date livraison



Customers

Villes
Etats



Sellers

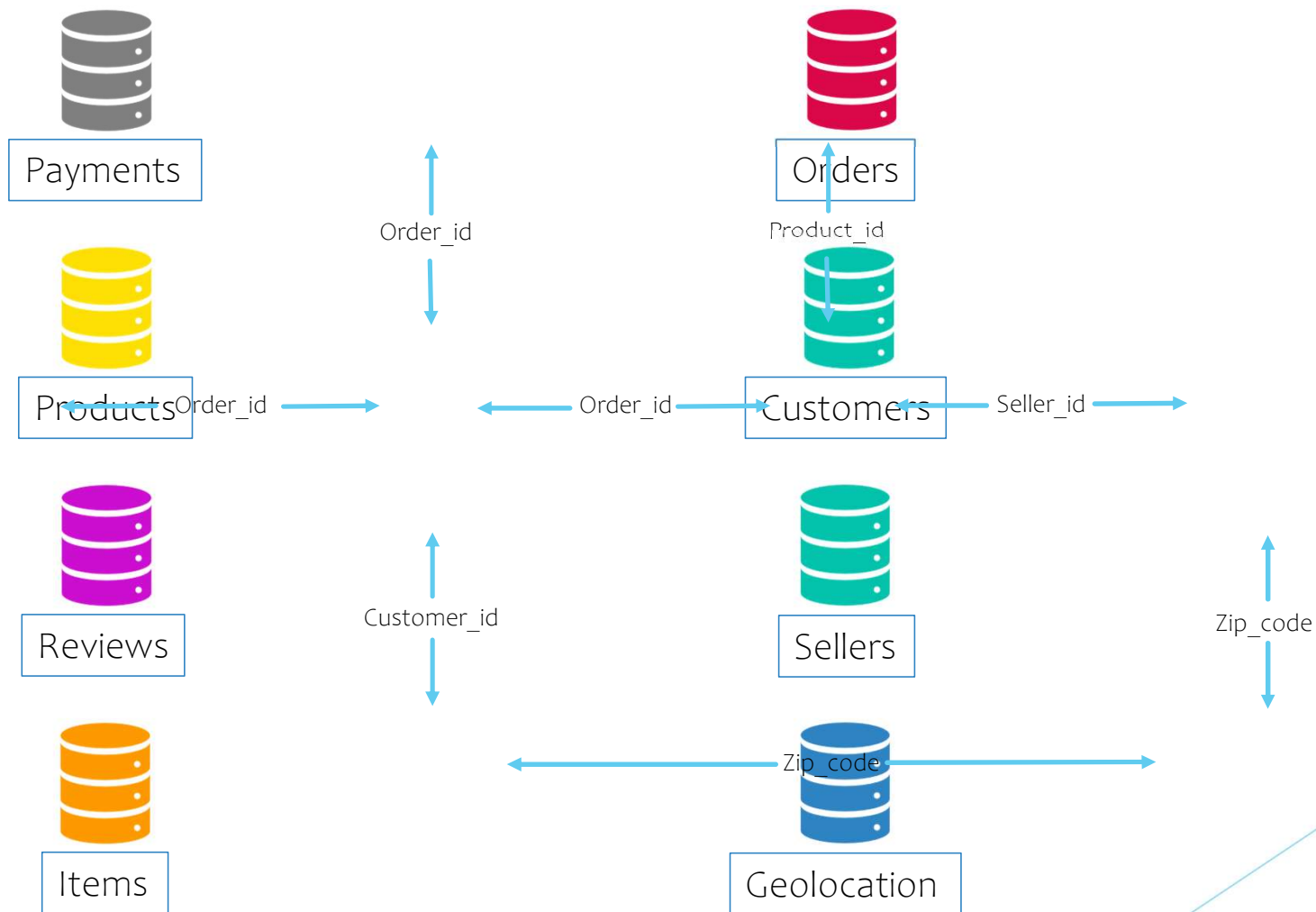
Villes
Etats

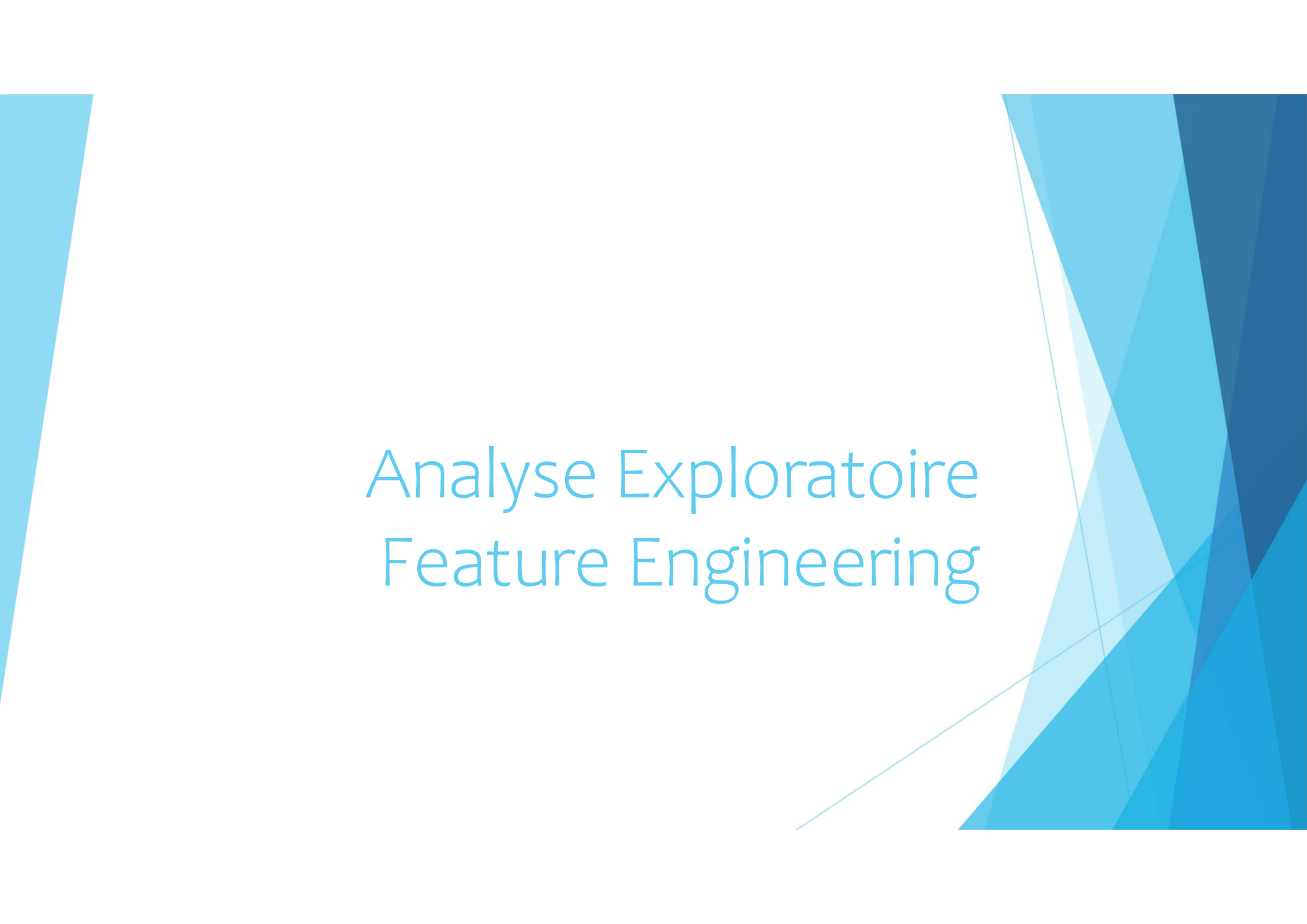


Geolocation

Latitude
Longitude

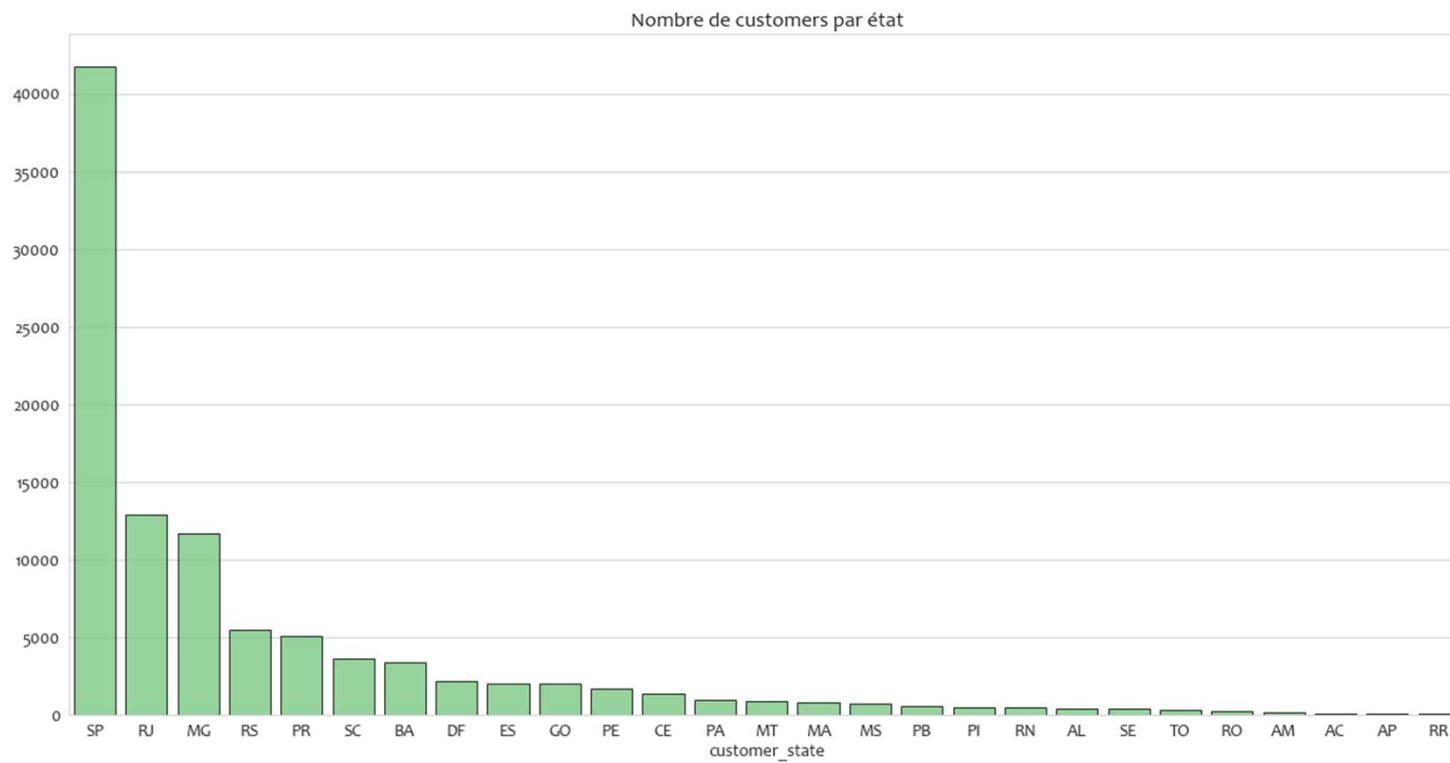
Descriptif du jeu de données :



The background features abstract geometric shapes in various shades of blue. On the left, a solid light blue triangle points towards the center. On the right, a complex arrangement of overlapping triangles in different blue tones (light, medium, and dark) creates a dynamic, layered effect. The central text is positioned on a white background that tapers towards the right, where it meets the blue geometric shapes.

Analyse Exploratoire Feature Engineering

Analyse Exploratoire : Localisation

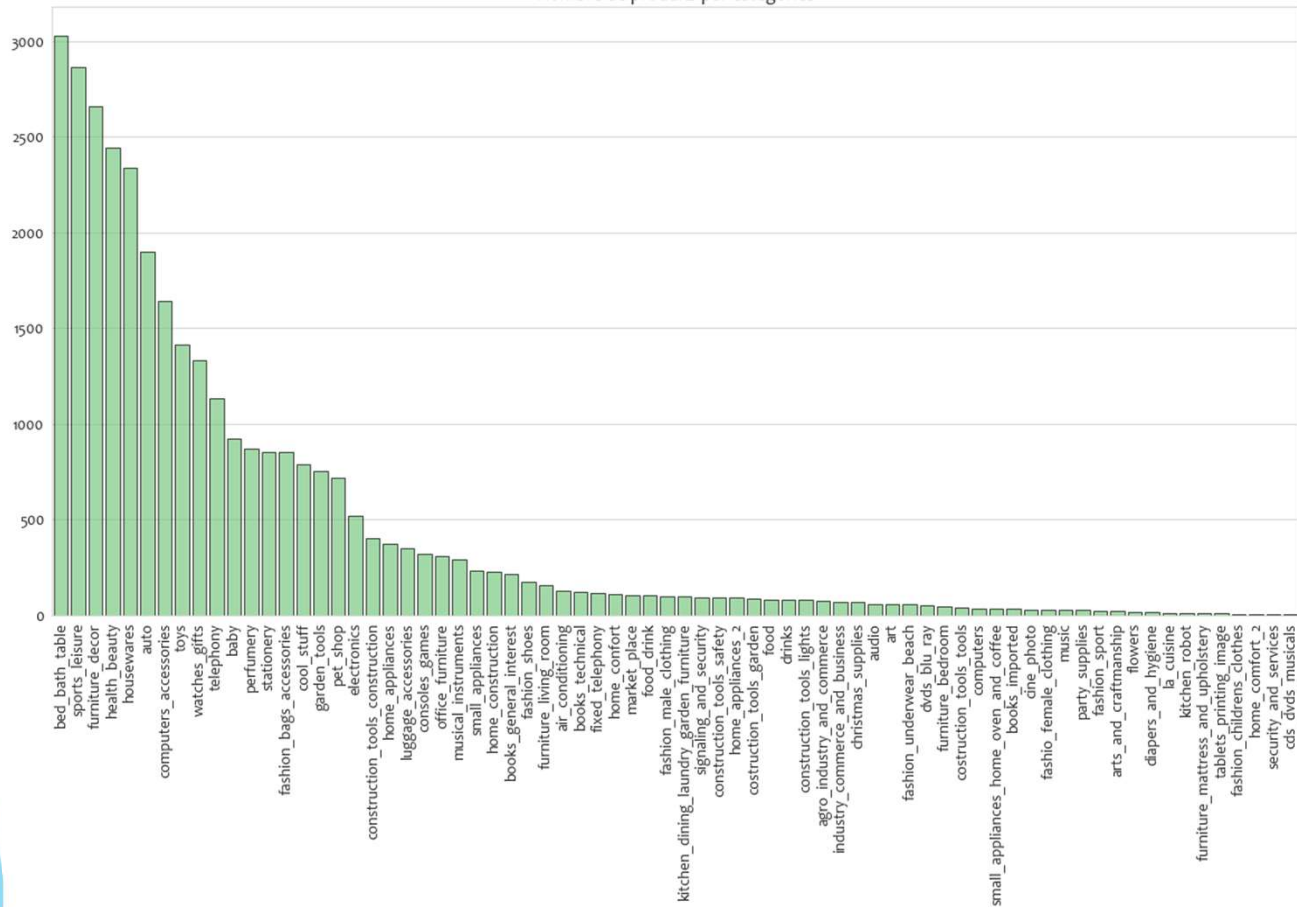


96 000 clients.

120 clients ont acheté
de plusieurs lieux.

Analyse exploratoire : Catégories de produits

Nombre de produits par catégories



Réduction du nombre de catégories.

Regroupement par thème.

Aménagement maison

Electronique

Santé/ beauté

Sport

Mode

Jouet et bébés

Voiture

Construction

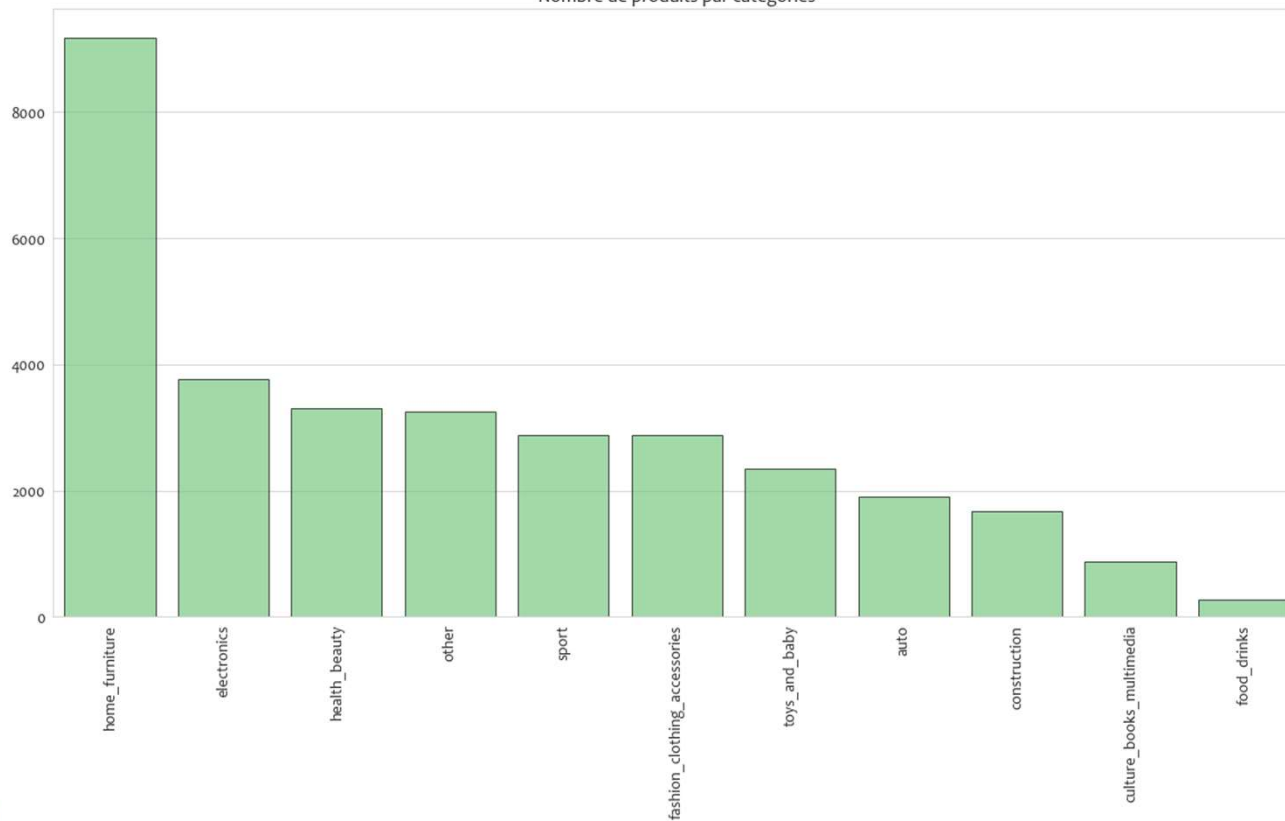
Culture/livres/multimedia

Nourriture/boisson

Autre

Analyse exploratoire : Catégories de produits

Nombre de produits par catégories

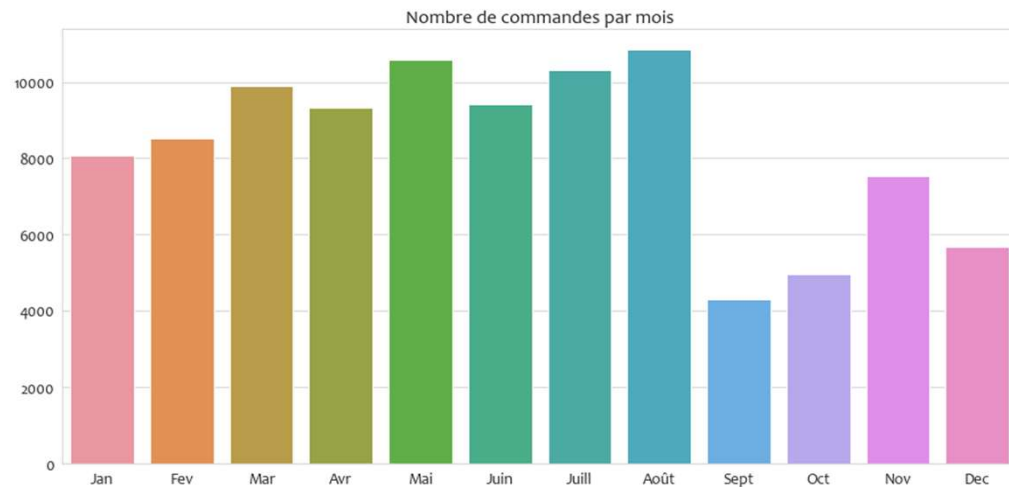


Réduction du nombre de catégories.

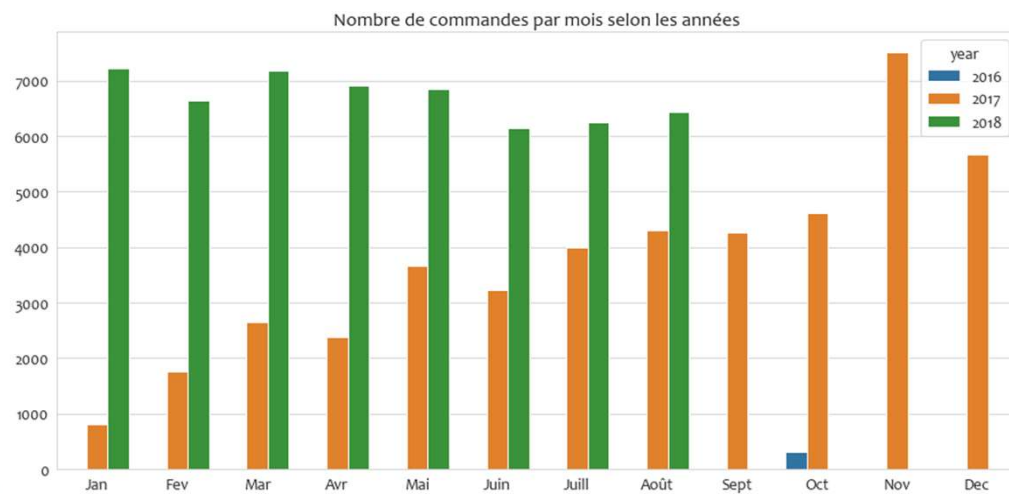
Regroupement par thème.

Aménagement maison
Electronique
Santé/ beauté
Sport
Mode
Jouet et bébés
Voiture
Construction
Culture/livres/multimedia
Nourriture/boisson
Autre

Analyse exploratoire : Dates de commande



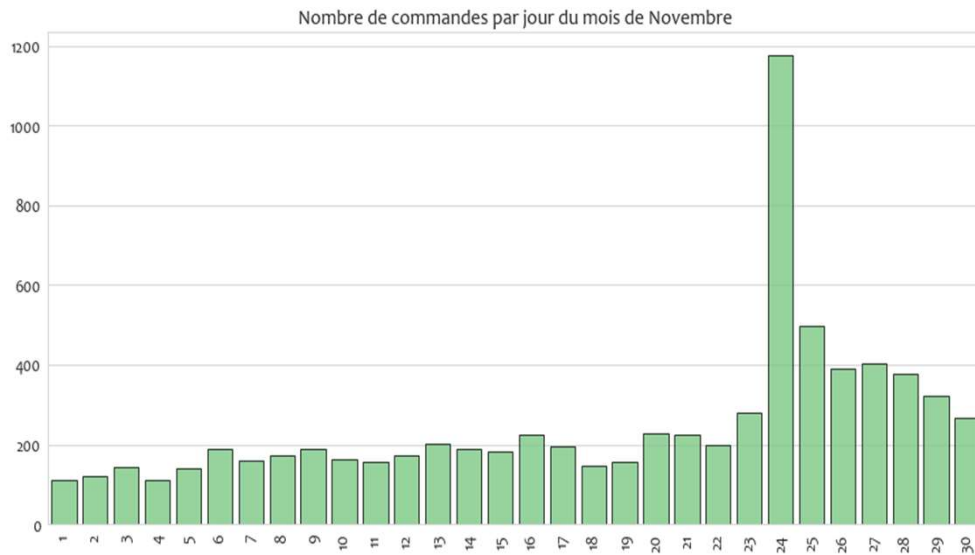
Septembre à octobre : une seule année.



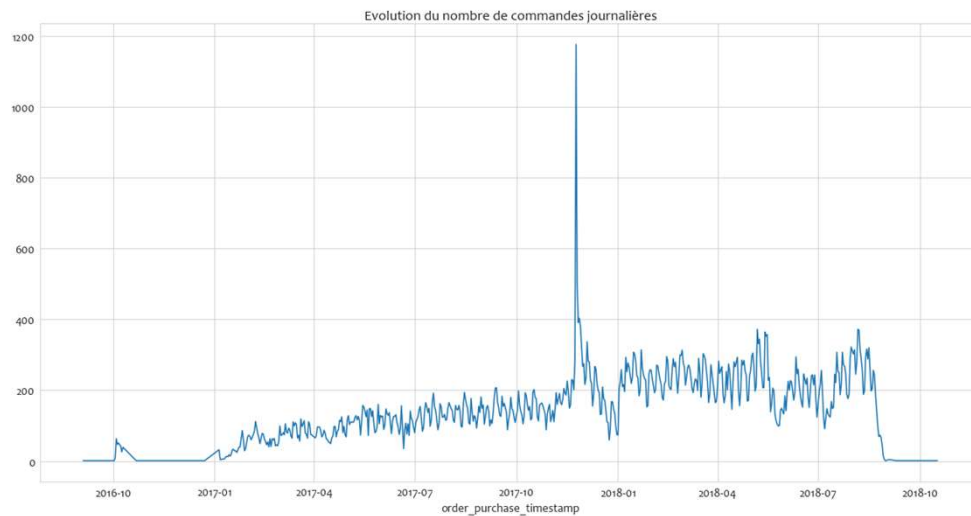
Croissance sur année 2017.

Pic en novembre.

Analyse exploratoire : Dates de commande



Pic de commandes le 24 Novembre.
Commandes plus élevées sur la fin
du mois : BlackFriday.

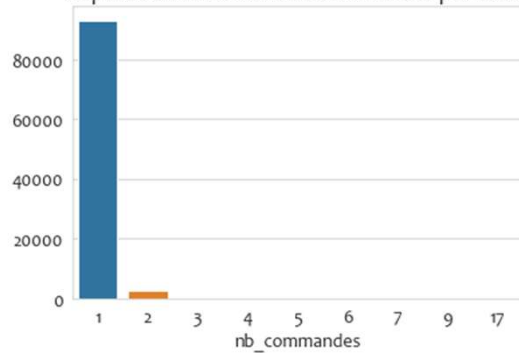


Un seul pic aussi élevé sur l'année.

Légère baisse en été.

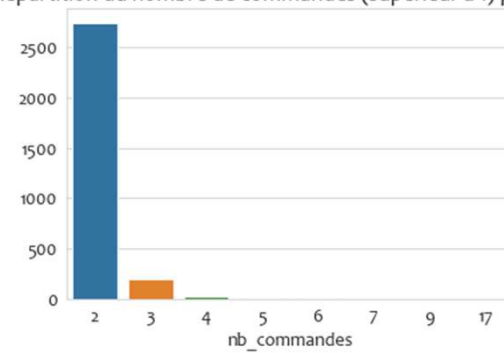
Analyse exploratoire : Nombre de commandes

Répartition du nombre de commandes par client



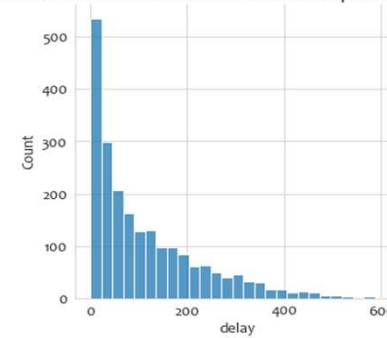
Majoritairement 1 seule commande par client

Répartition du nombre de commandes (supérieur à 1) par client



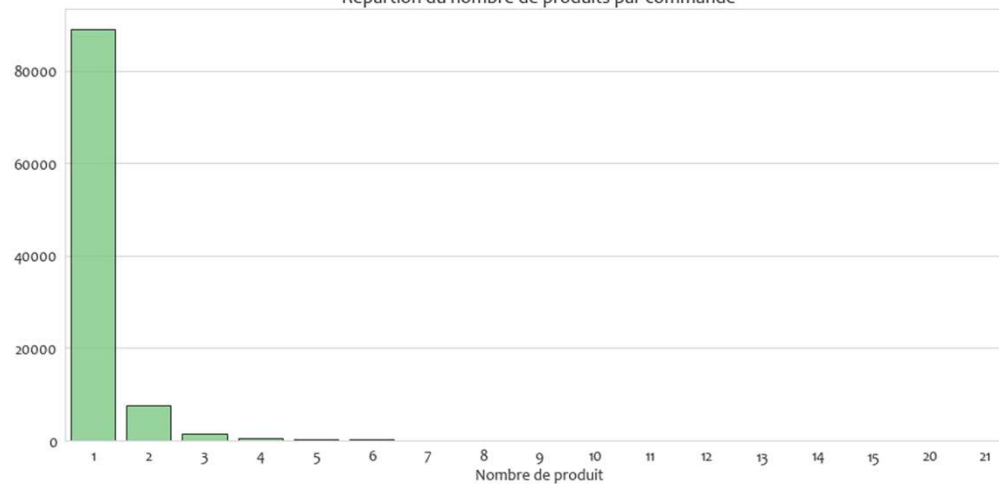
Parmi les clients à commandes multiples majoritairement 2 commandes

Distribution des délais entre 2 commandes pour chaque client



Délai entre commandes inférieur à 6 mois

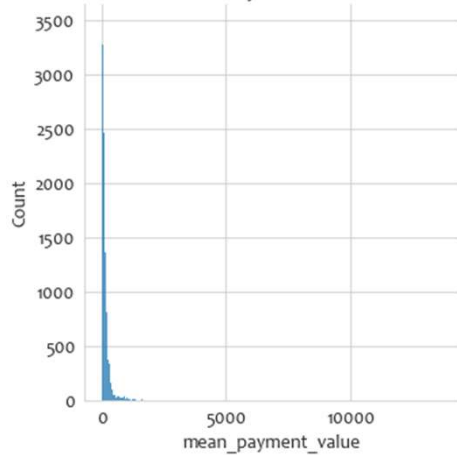
Répartition du nombre de produits par commande



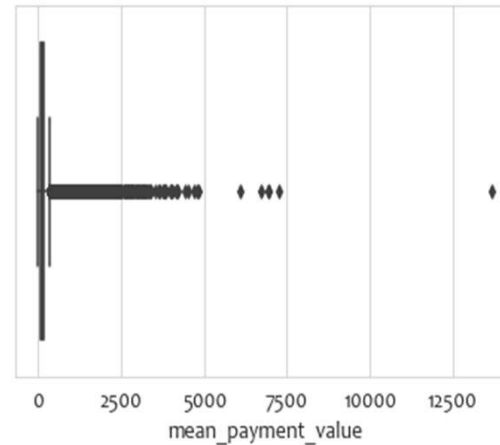
1 produit par commande

Analyse exploratoire : Montant moyen

Distribution de la valeur moyenne des commandes par client

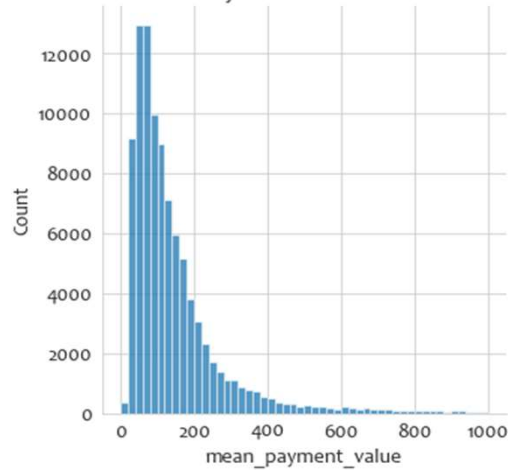


Visualisation d'outliers



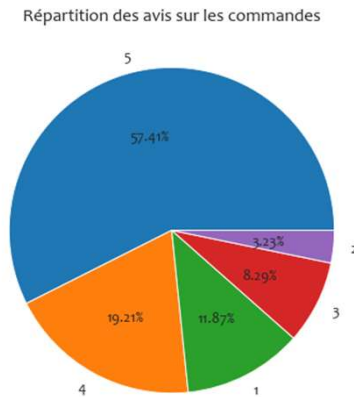
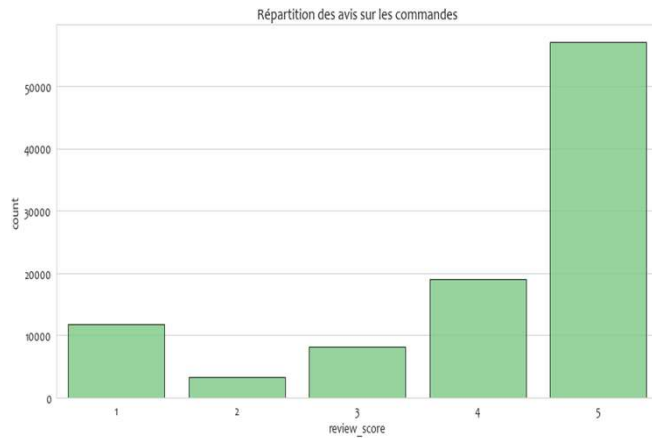
Détection d'outliers :
au dessus de 5 000\$
pour une commande.

Distribution de la valeur moyenne des commandes < à 1000\$ par client



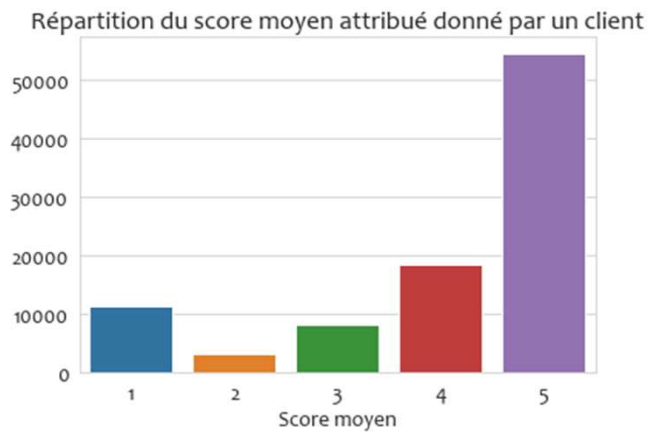
La majorité des paiements
sont inférieurs à 200\$

Analyse exploratoire : Reviews



Avis plutôt positifs: 4 ou 5/5.

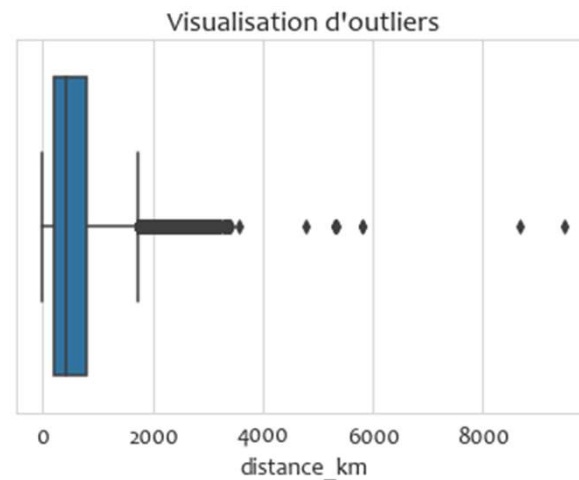
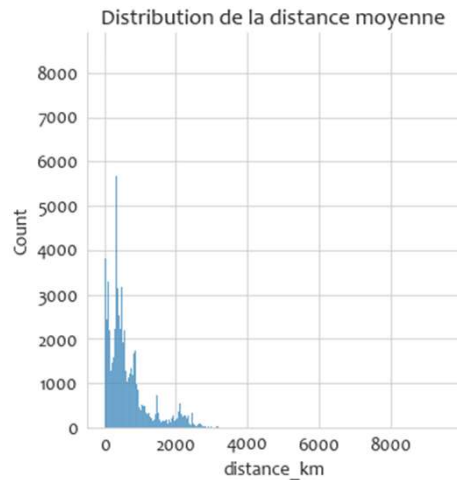
1/5 troisième note la plus présente avec 12%.



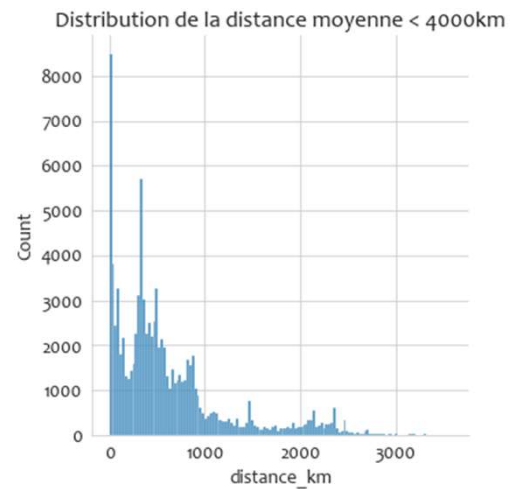
La plupart des clients ont 1 seule commande.

La distribution est presque identique.

Analyse exploratoire : Distances



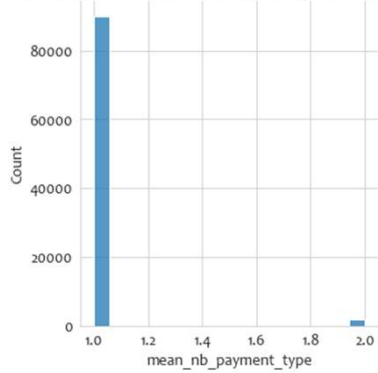
Détection d'outliers :
au dessus de 4 000 km



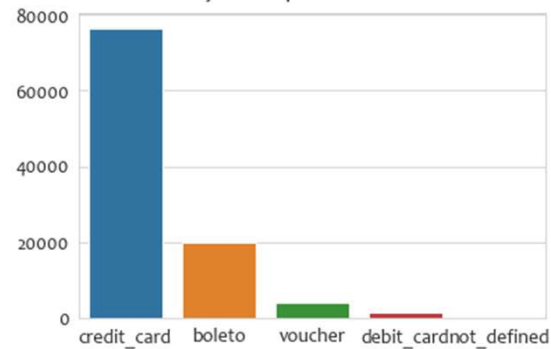
La majorité des distances sont
inférieures à 1 000 km.

Analyse exploratoire : Paiements

Distribution du nombre de type de paiement moyen

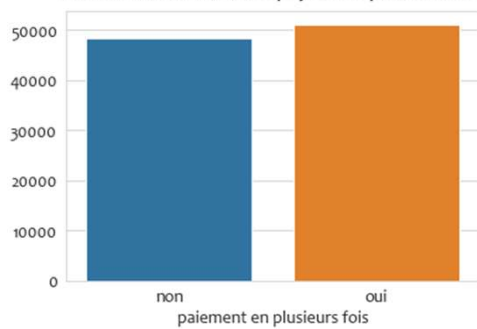


Moyens de paiement utilisés

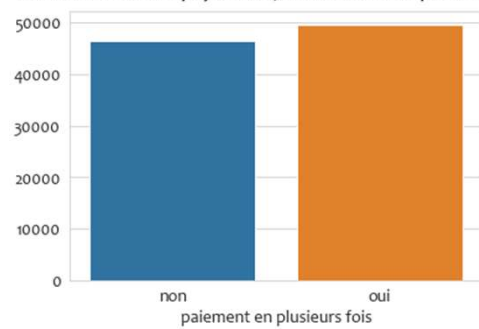


Majoritairement 1 seul type de paiement:
Credit Card

Nombre de commandes payées en plusieurs fois



Nombre de clients payant majoritairement en plusieurs fois



Environ la moitié des commandes
sont payées en plusieurs fois sur
un des types de paiement.

Majoritairement 1 commande par client:
résultat presque identique

Feature Engineering :

- ▶ Rassembler les informations à l'échelle des clients en jouant avec les données
- ▶ Achat au black Friday ?
- ▶ Plusieurs commandes ? Délai entre commandes
- ▶ Paiement moyen
- ▶ Note moyenne
- ▶ Distance
- ▶ Paiements en plusieurs fois/types
- ▶ Types de paiement
- ▶ Catégories de produits

Feature Engineering :

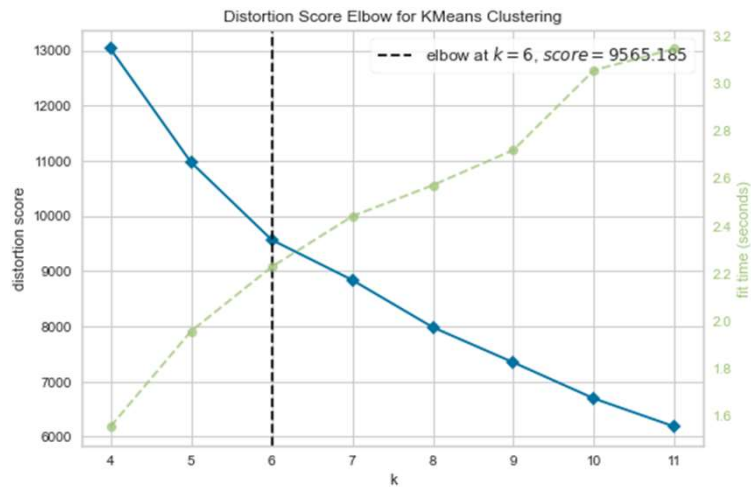
- ▶ Binarisation
 - ▶ BlackFriday
 - ▶ Plusieurs commandes binarisé
 - ▶ Paiement en plusieurs fois : si 50%+ des commandes en plusieurs fois
- ▶ Type de paiement : compter le fois d'utilisation par moyen
- ▶ Compter nombre produit par catégorie
- ▶ MinMaxScaler
 - ▶ La plupart des valeurs sont entre 0 et 1
 - ▶ Réduire l'impact des variables paiement moyen et distance

The background features abstract geometric shapes in various shades of blue. On the left, a solid light blue triangle points upwards. On the right, a complex arrangement of overlapping triangles in different blue tones (light, medium, and dark) creates a dynamic, layered effect. The word 'Modélisation' is centered in the white space between these blue elements.

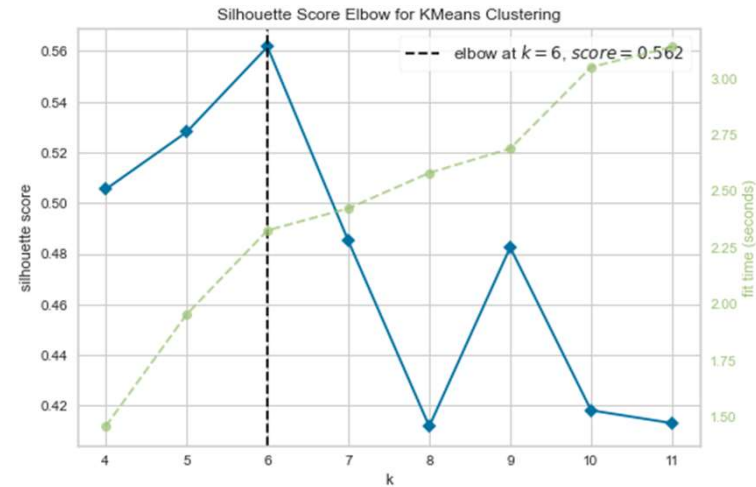
Modélisation

Clustering KMeans

Kmeans KElbow method :



Somme des distances aux centroïdes.

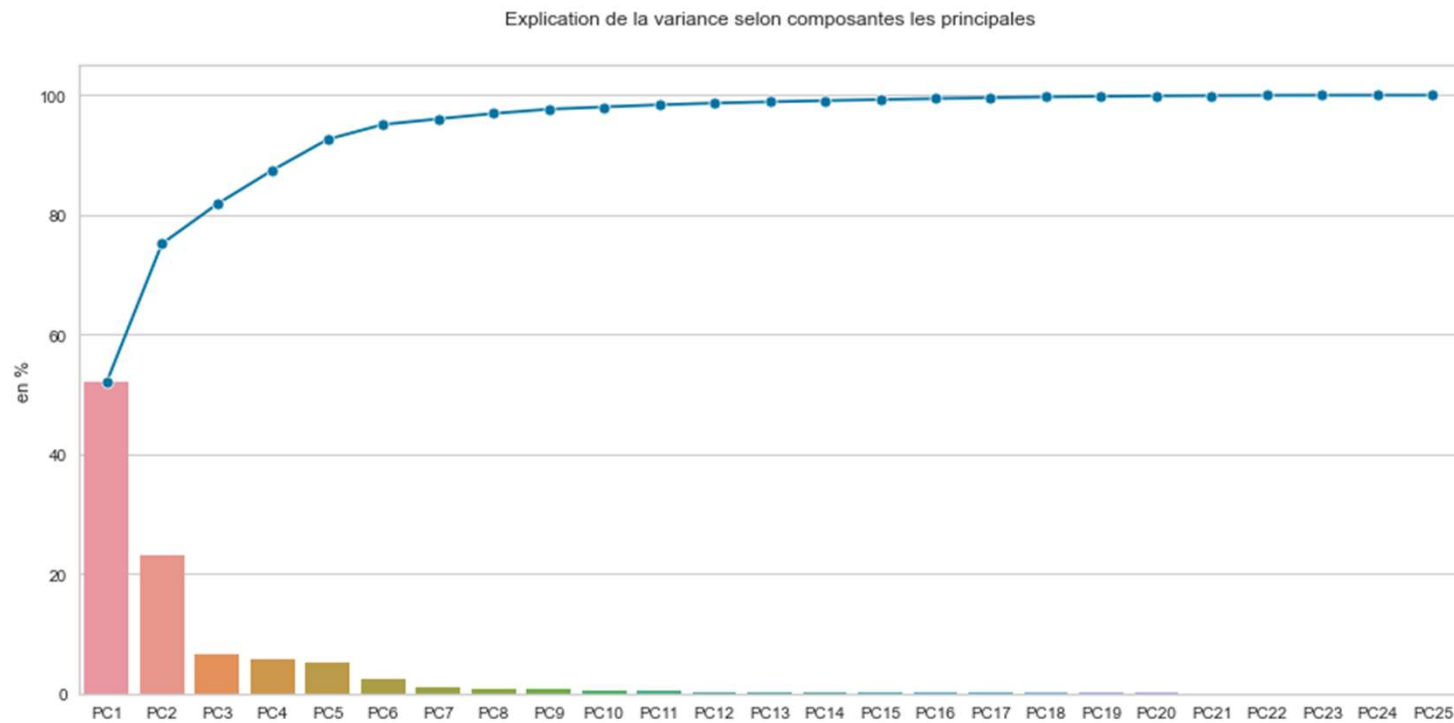


Différence entre distance moyenne au même cluster et distance moyenne autre clusters avec rapport au max.

Compris entre -1 et 1, 1 étant le meilleur.

6 clusters

Analyse en composantes principales :

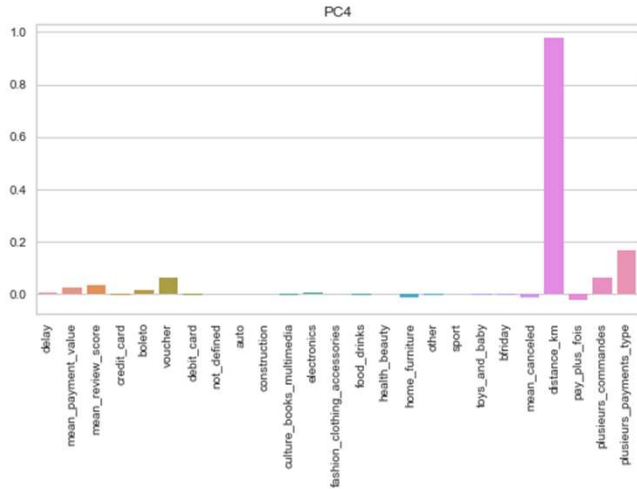
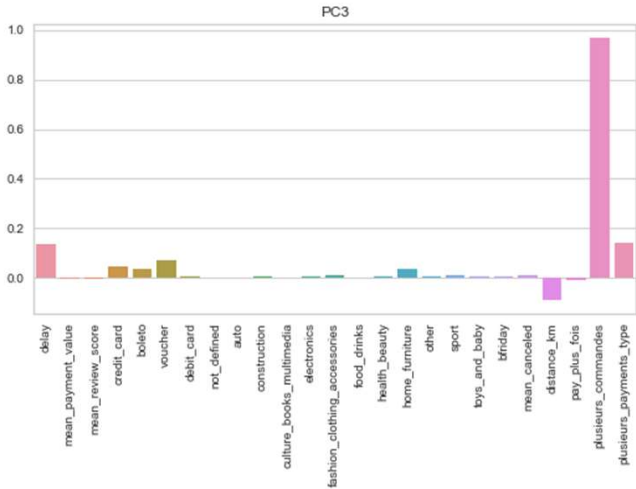
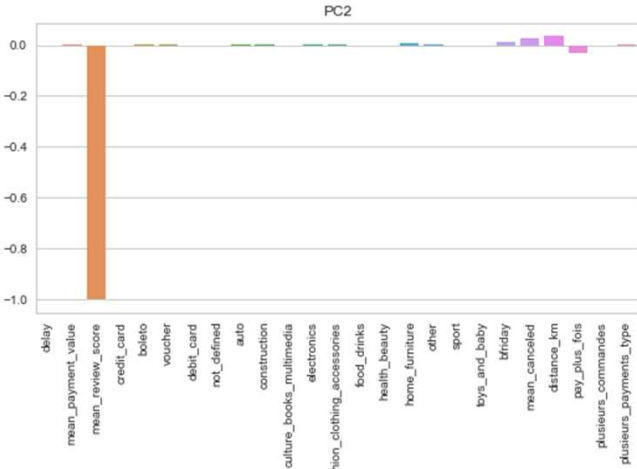
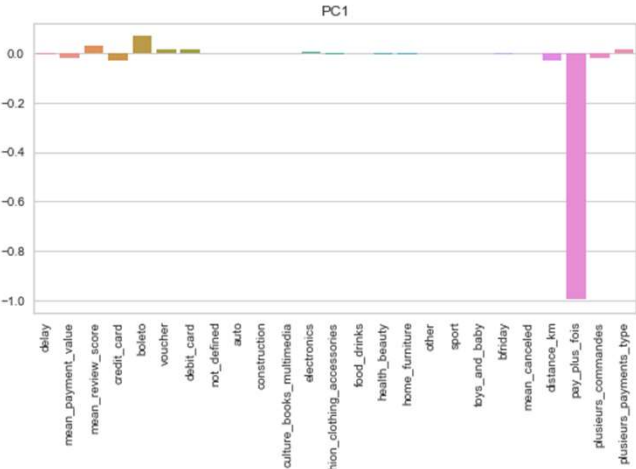


Au départ 25 features

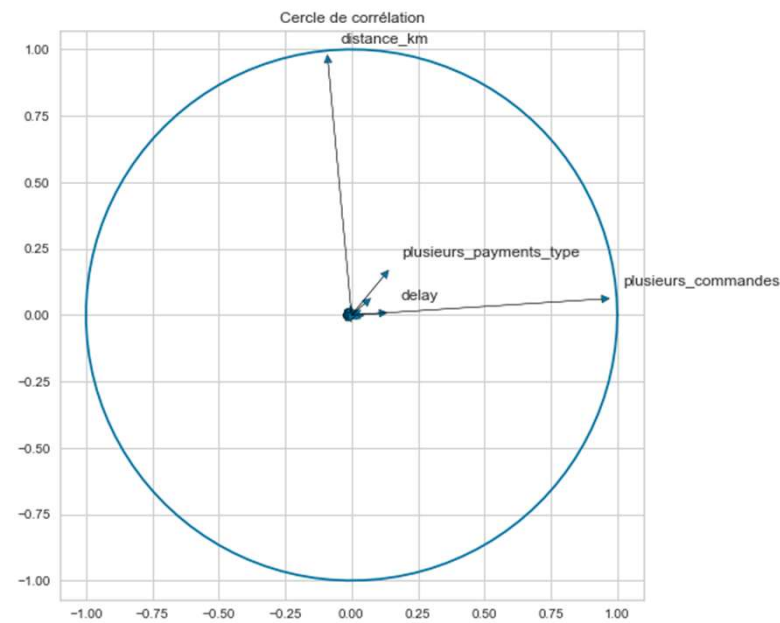
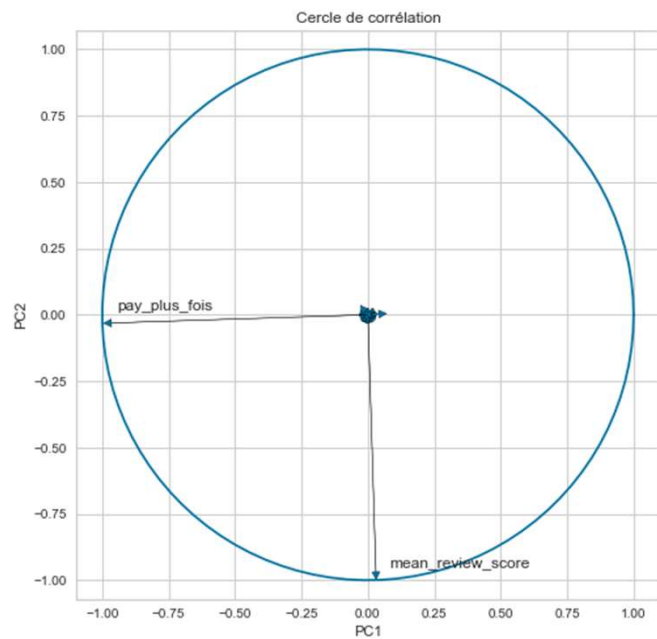
87% de variance expliqué par
les 4 premières composantes
principales

Analyse en composantes principales :

Features des composantes principales de la PCA

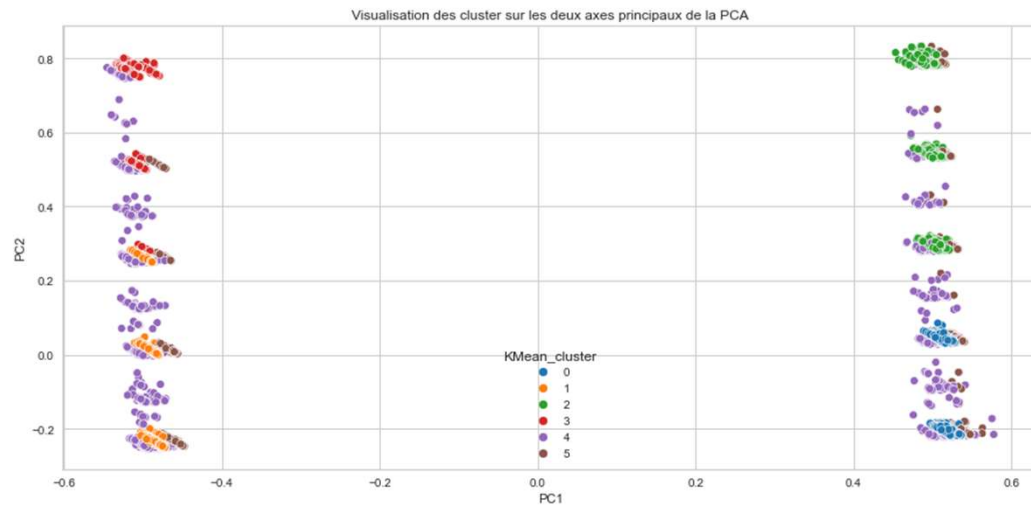


Analyse en composantes principales :

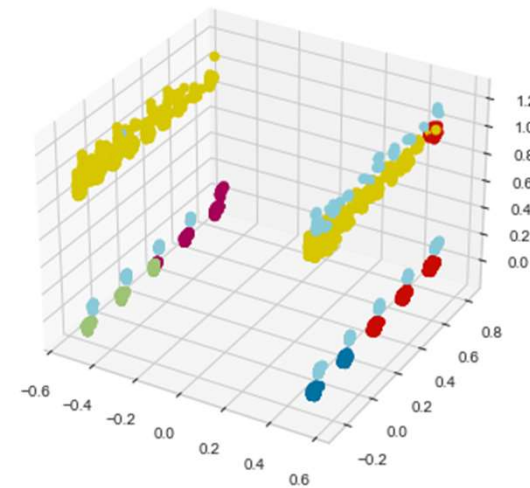


Features non corrélées linéairement

PCA : Représentation des clusters



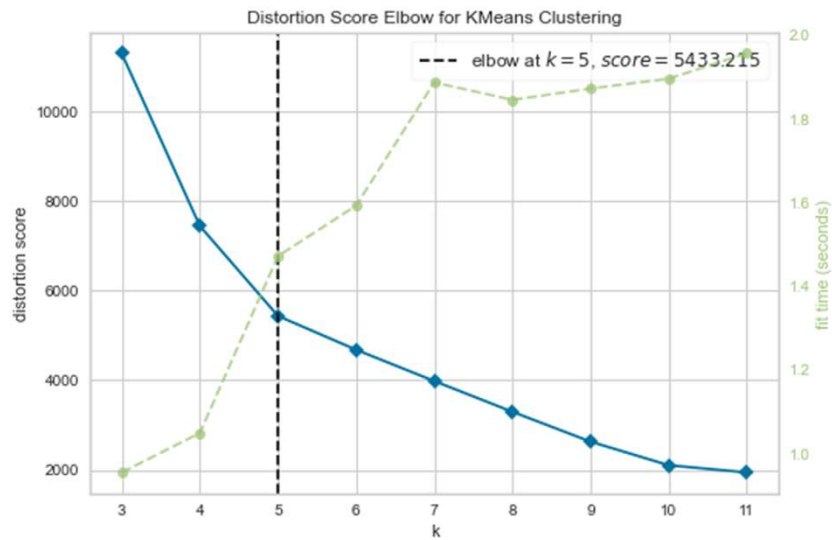
Visualisation des cluster sur les trois axes principaux de la PCA



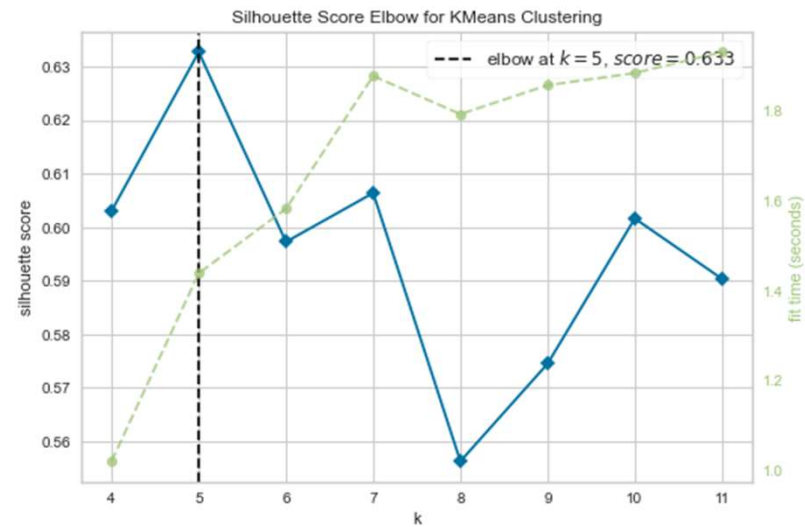
Quelques clusters bien définis mais un cluster qui a l'air de l'être un peu moins.

Possibilité d'entraînement du modèle après la PCA sur les 4 variables principales ?

Entraînement après réduction dimensionnelle :

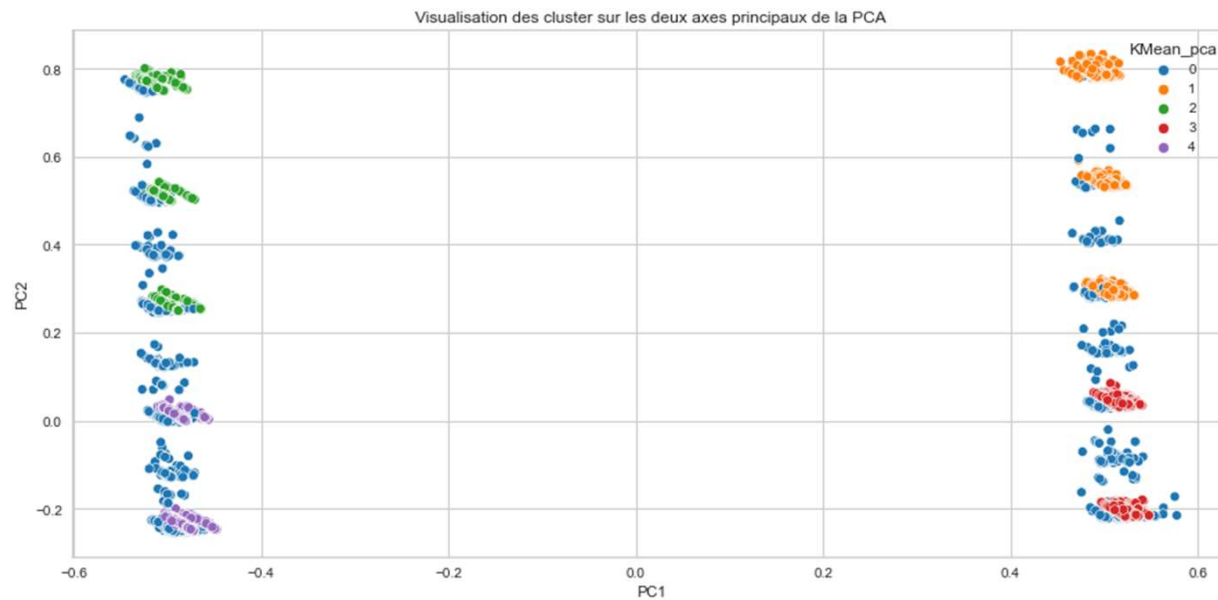


Nouveau nombre de cluster : 5

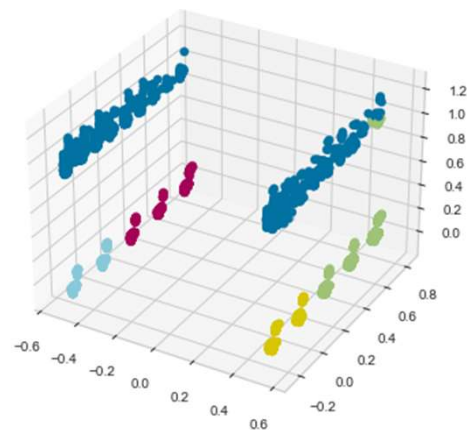


meilleure silhouette : 0,562 vs 0,633

Après PCA : Représentation des clusters

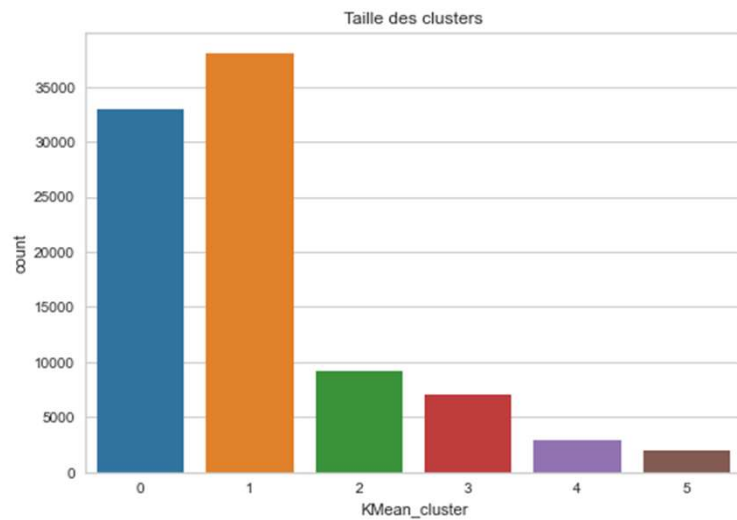


Visualisation des cluster sur les trois axes principaux de la PCA

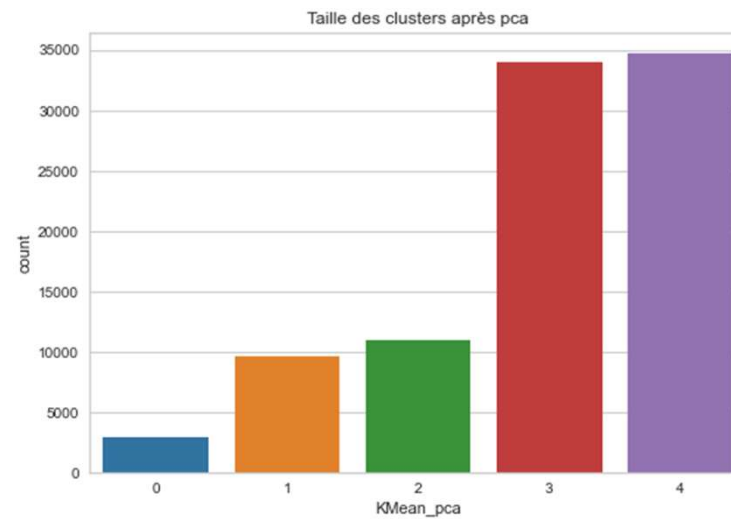


Clusters mieux séparés

Taille des clusters avant et après PCA :

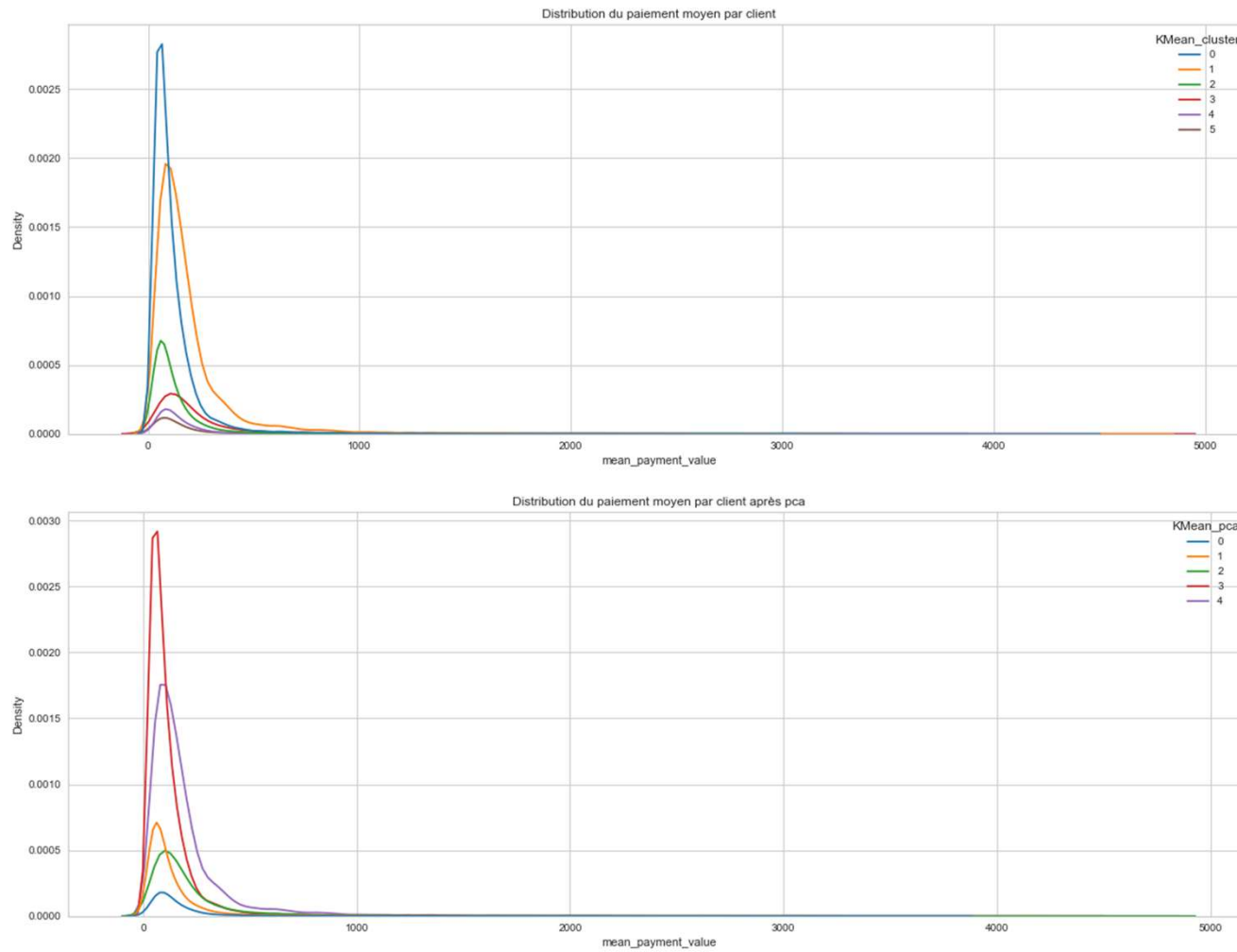


2 clusters principaux : 75% des points.



Le numéro du cluster n'a pas d'importance.

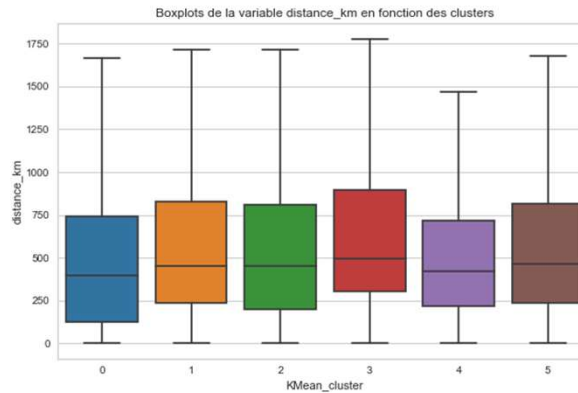
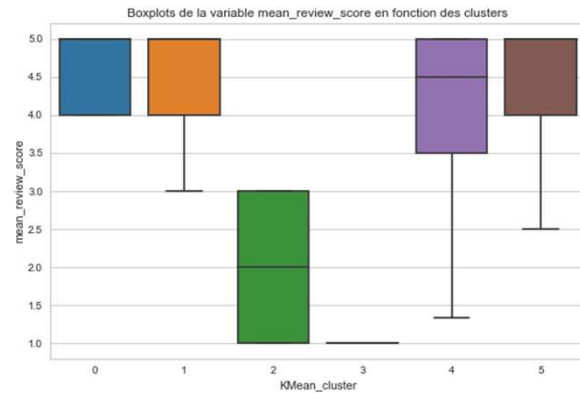
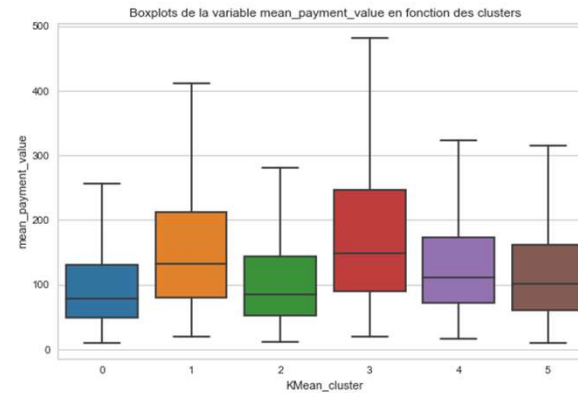
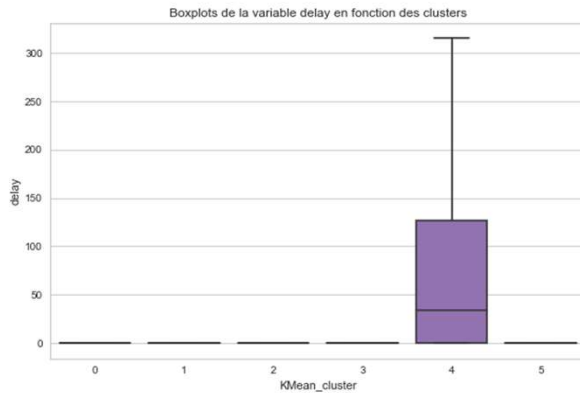
Distribution du paiement moyen avant et après PCA :



Distribution
'semblables'

Boxplots par cluster avant PCA :

Boxplot des variables ['delay', 'mean_payment_value', 'mean_review_score', 'distance_km'] avec les clusters de KMean_cluster



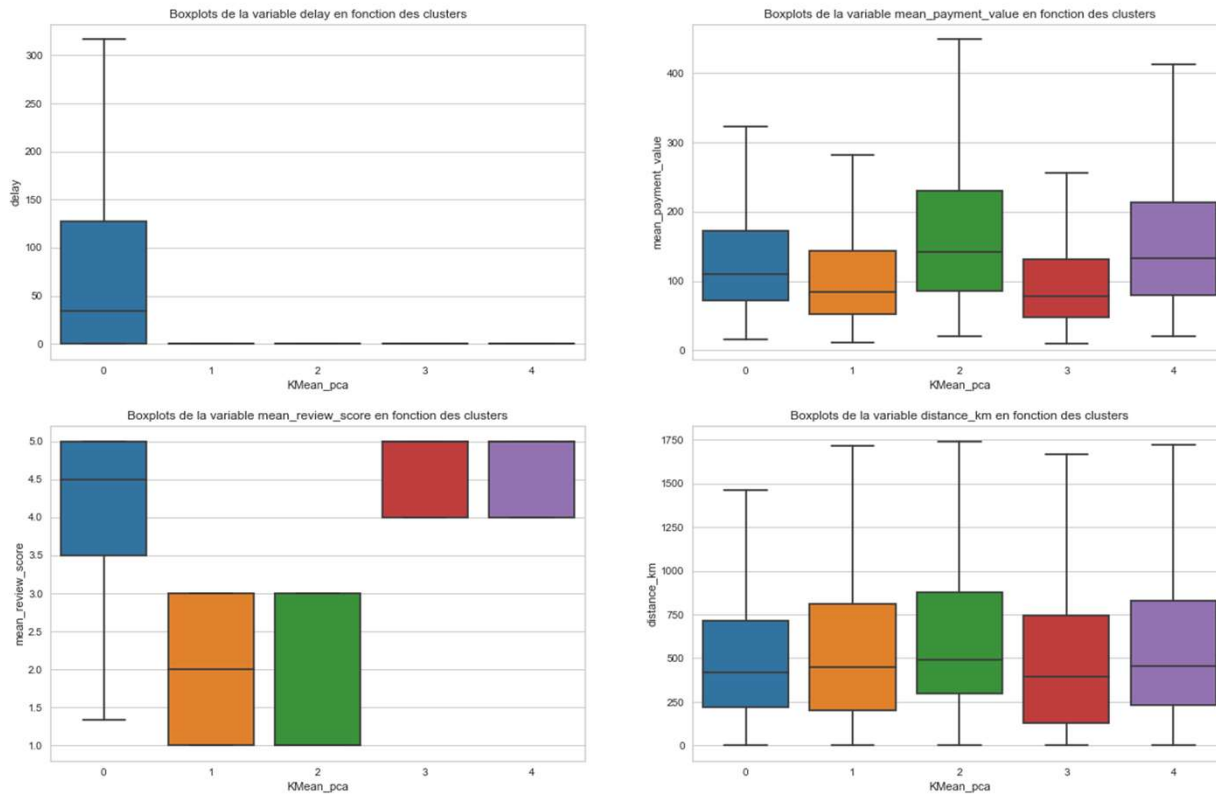
Plusieurs commandes
cluster 4

Clusters 1 et 4 :
paiement moyen plus élevé

Avis :
Cluster 3 très mécontent
Cluster 2 plutôt insatisfait
Cluster 1, 4 et 5 mélangés
Cluster 0 satisfait

Boxplots par cluster après PCA :

Boxplot des variables ['delay', 'mean_payment_value', 'mean_review_score', 'distance_km'] avec les clusters de KMean_pca

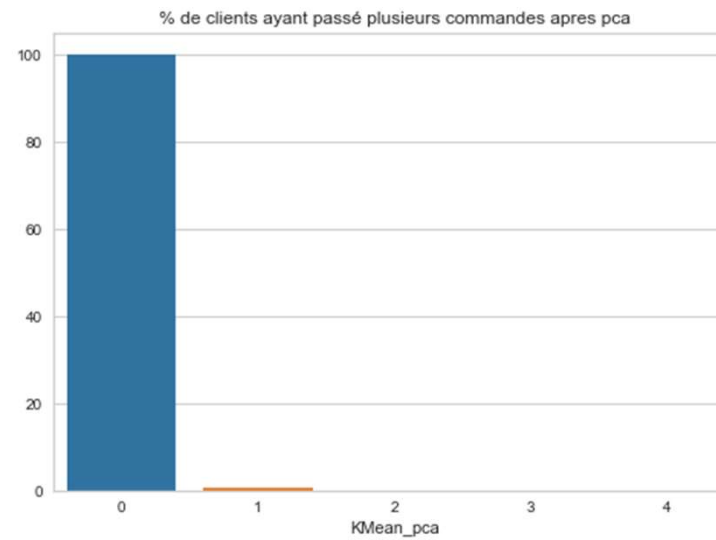
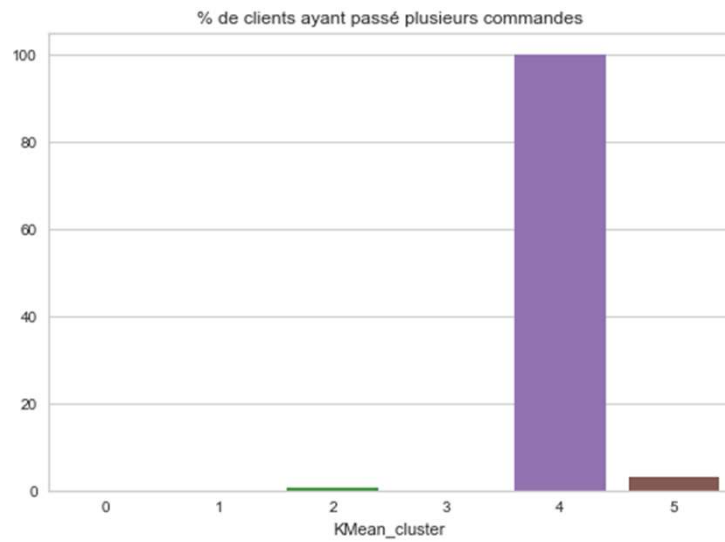


Plusieurs commandes
cluster 0

Clusters 2 et 4 :
paiement moyen plus élevé
1 et 3 légèrement plus bas

Avis :
Clusters 1 et 2 plutôt insatisfaits
Cluster 0 mélangé
Clusters 3 et 4 satisfaits
Perte cluster 'très mécontents'

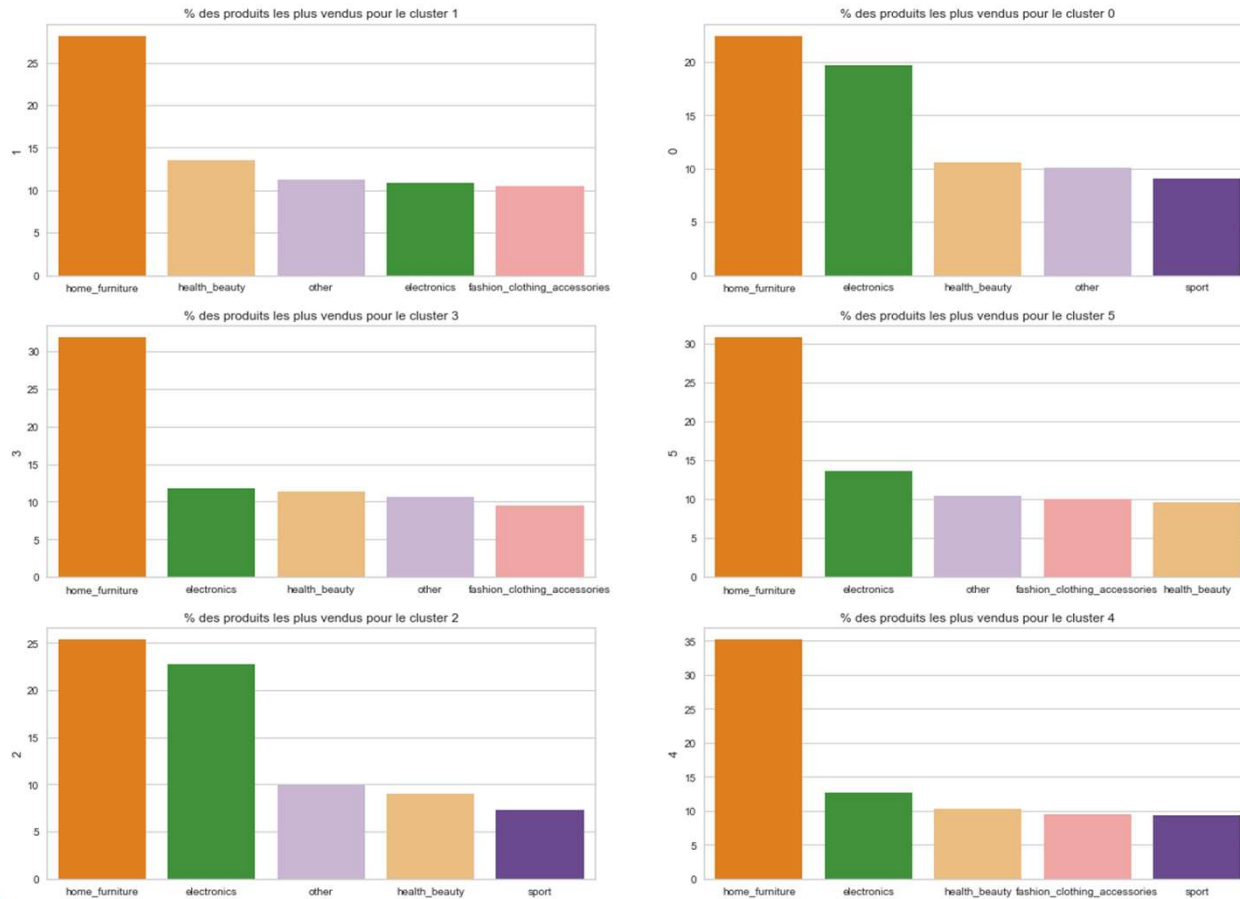
Plusieurs commandes avant et après PCA :



Majoritairement répartis dans un seul cluster à part quelques exceptions.

Top produit avant pca :

% du top 5 des produits achetés pour KMean_cluster



Home furniture :
toujours élevé

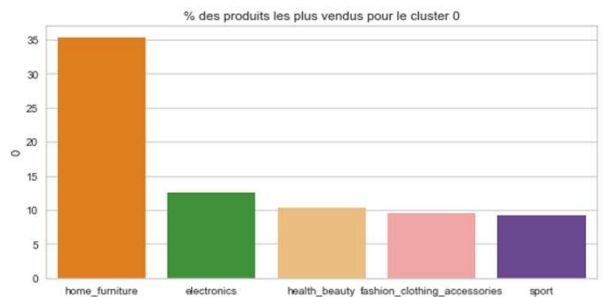
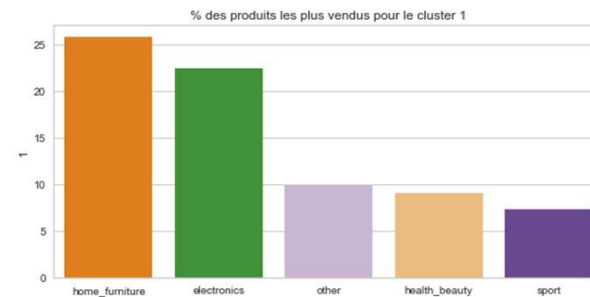
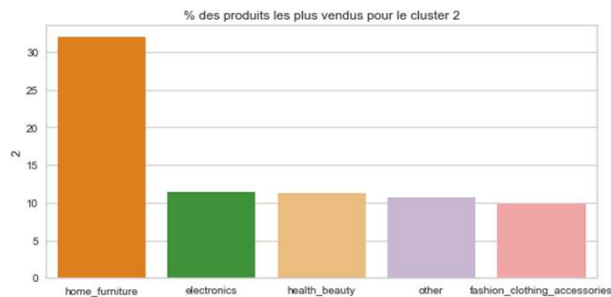
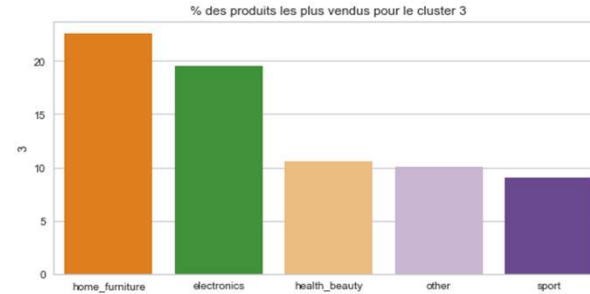
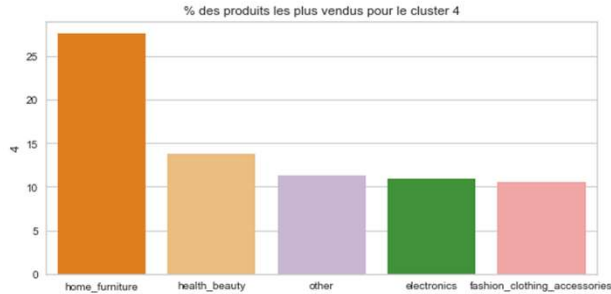
2 clusters privilégiant
aussi l'électronique

3 clusters avec la
catégorie sport

4 clusters avec la
catégorie fashion

Top produits après PCA :

% du top 5 des produits achetés pour KMean_pca



On garde à peu près les mêmes distributions par cluster

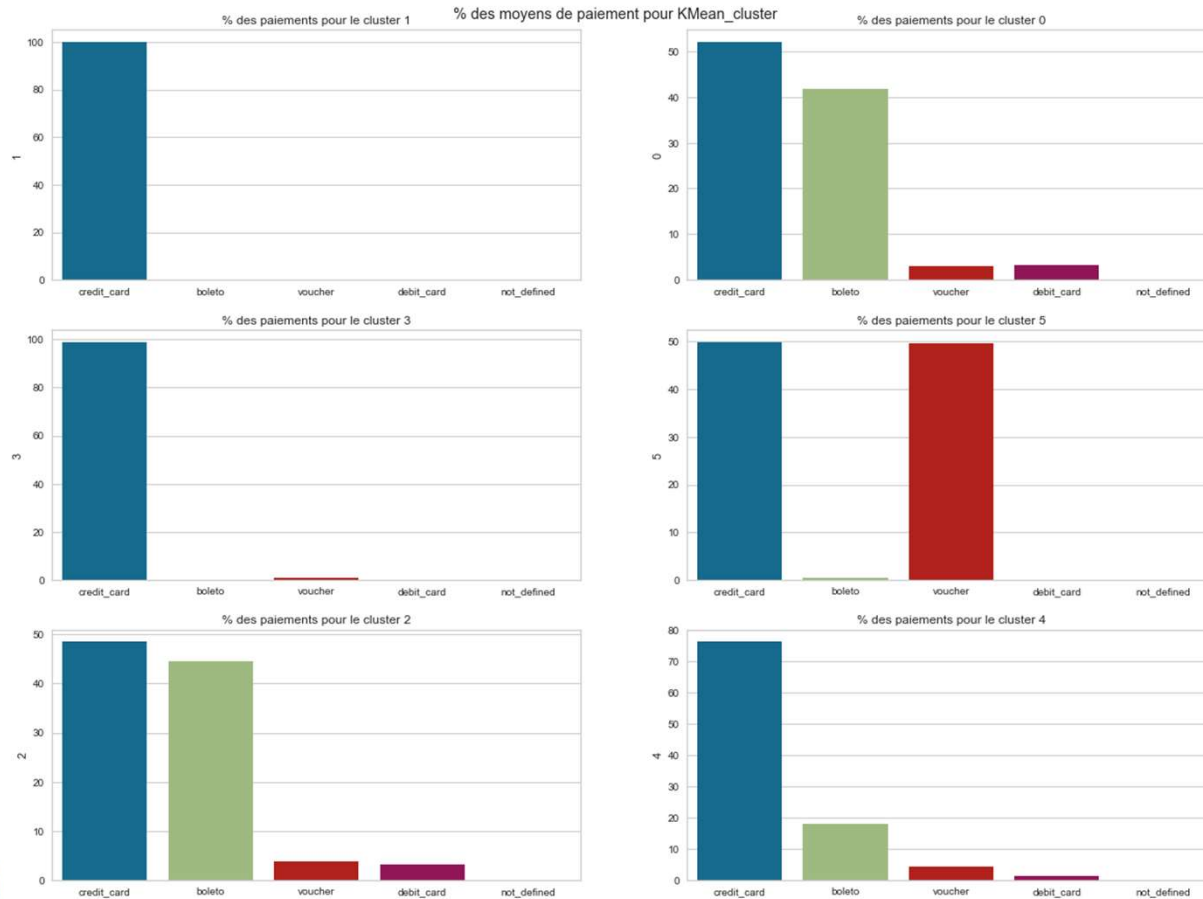
Home furniture :
toujours élevé

2 clusters privilégiant
aussi l'électronique

3 clusters avec la
catégorie sport

3 clusters avec la
catégorie fashion

Types de paiement avant PCA :

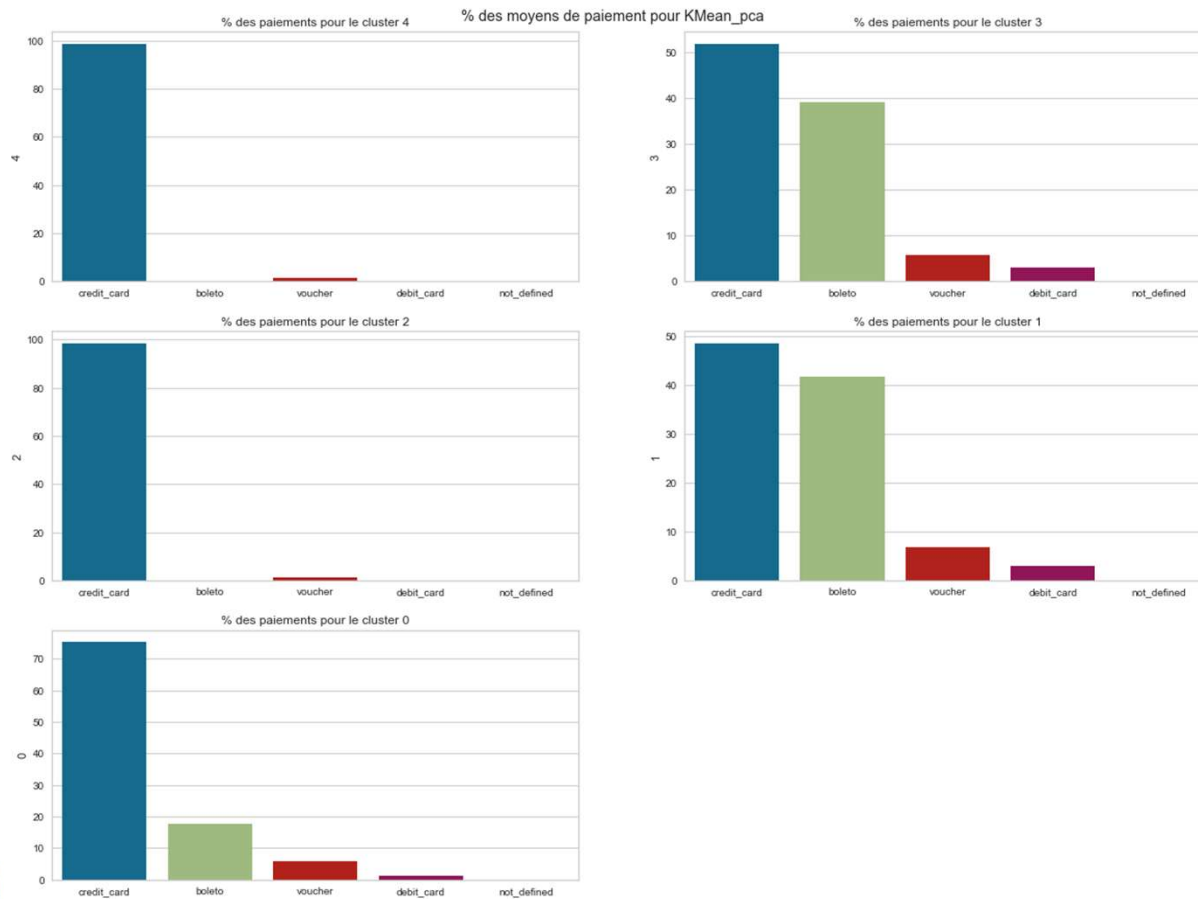


2 clusters uniquement
Credit Card

2 clusters Credit Card &
Boleto

1 cluster Credit Card & Voucher

Types de paiement après PCA :

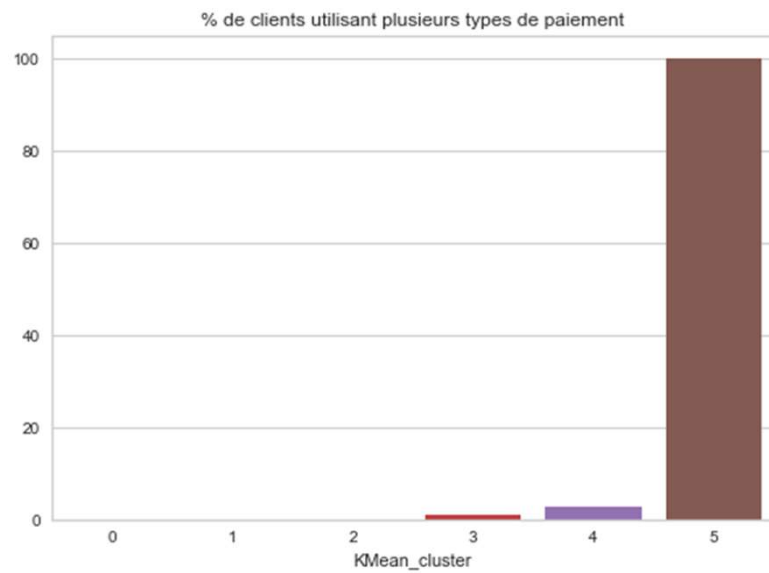


2 clusters uniquement
Credit Card

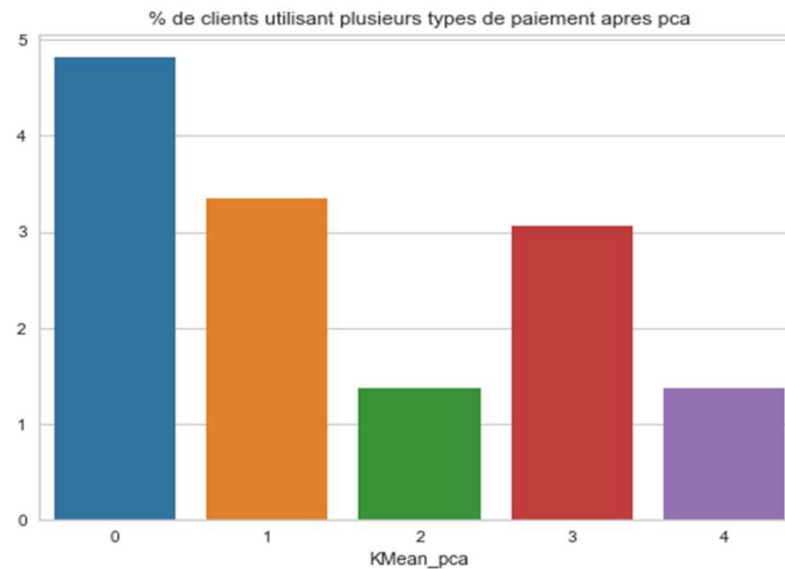
2 clusters Credit Card &
Boleto

Perte du cluster 5 :
Credit Card & Voucher

Types de paiement avant et après PCA :

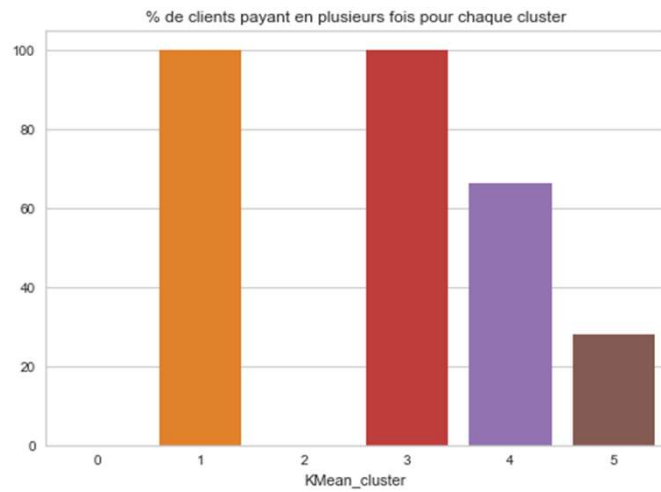


Cluster 5 clients utilisant plusieurs types de paiement homogénéisé dans les autres clusters

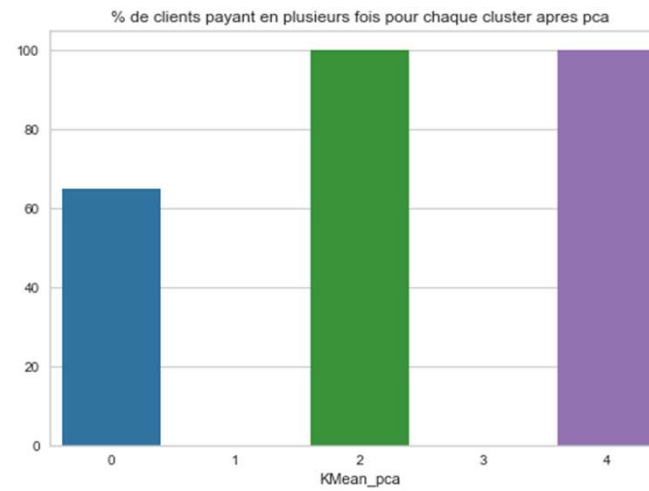


Information perdue

Payement plusieurs fois avant et après PCA :

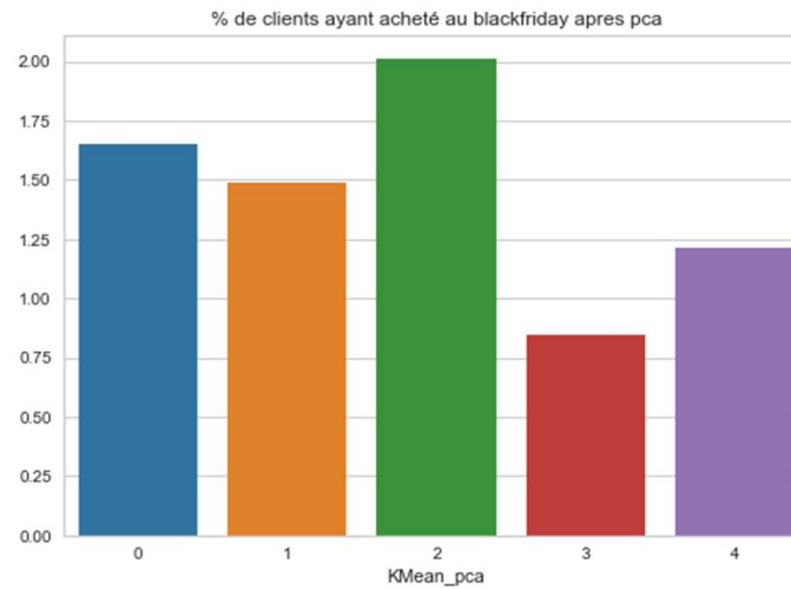
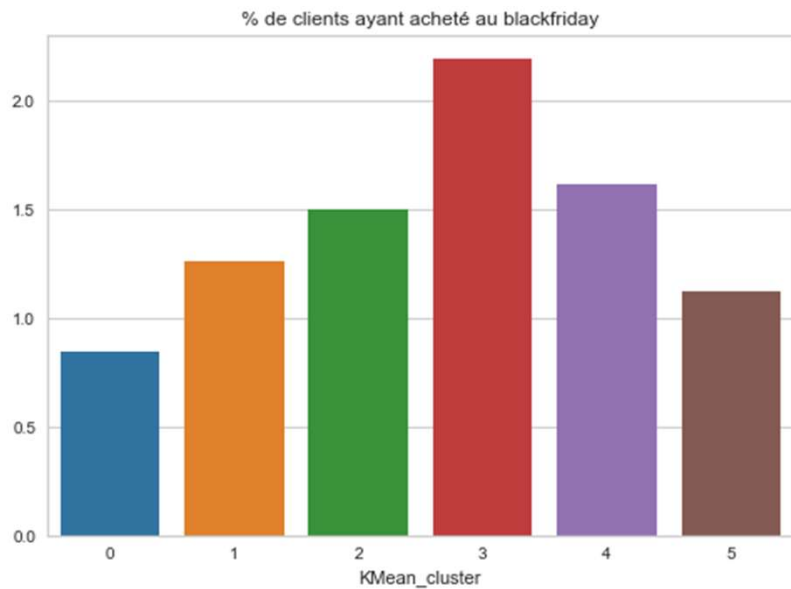


2 clusters qui paient uniquement en plusieurs fois



Clients qui paient en plusieurs fois repartis plus distinctement (3v4 clusters)

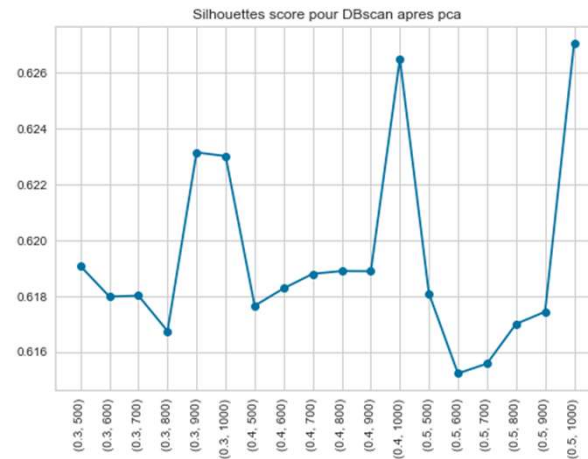
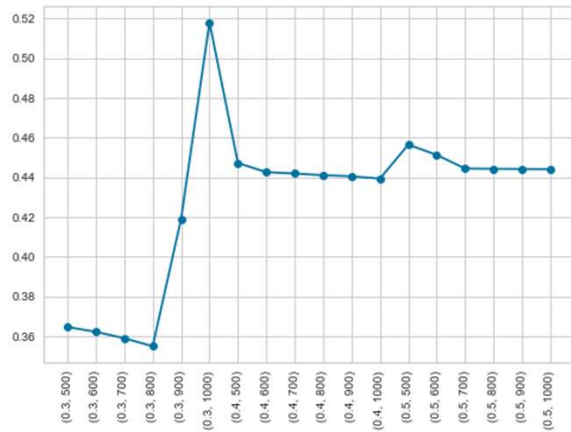
Achats pendant le BlackFriday avant et après PCA :



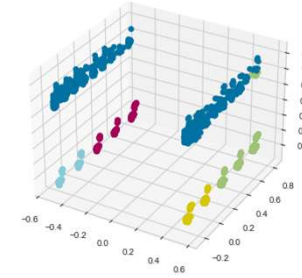
Le pic d'achats pendant le BlackFriday n'est pas impactant sur notre clustering : les gens sont repartis dans les différents clusters.

Clustering DBscan

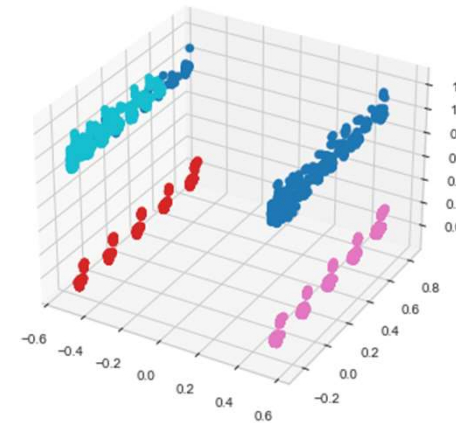
Silhouette et représentation :



Visualisation des cluster sur les trois axes principaux de la PCA



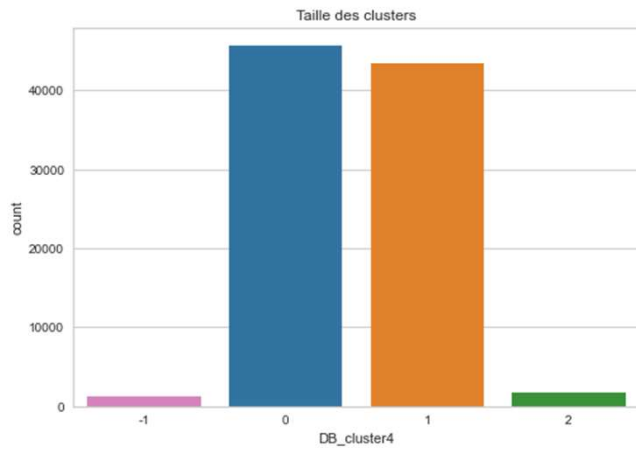
Visualisation des cluster sur les trois axes principaux de la PCA



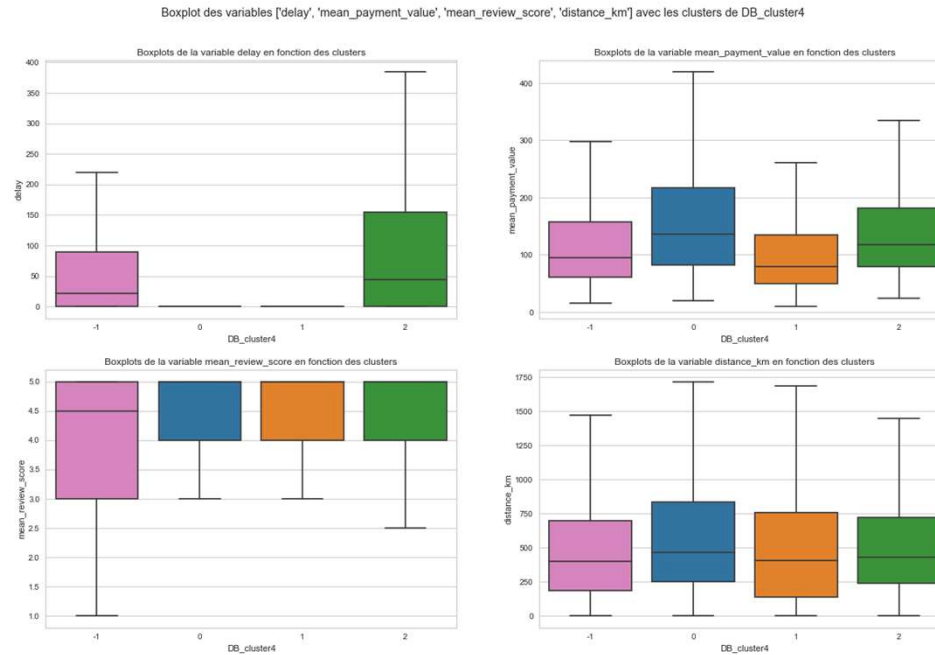
Meilleures silhouettes lors de l'entraînement après PCA

3 clusters + un cluster 'bruit'

Taille des clusters et boxplots :



Toujours 2 majoritaires

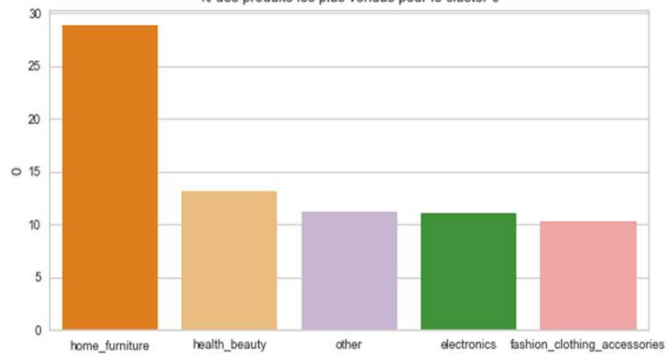


2 clusters avec plusieurs commandes
Avis moins bien définis
2 clusters avec paiement un peu plus élevés

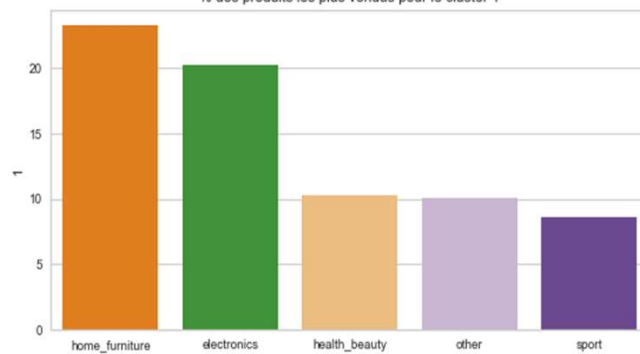
Top produits

% du top 5 des produits achetés pour DB_cluster4

% des produits les plus vendus pour le cluster 0

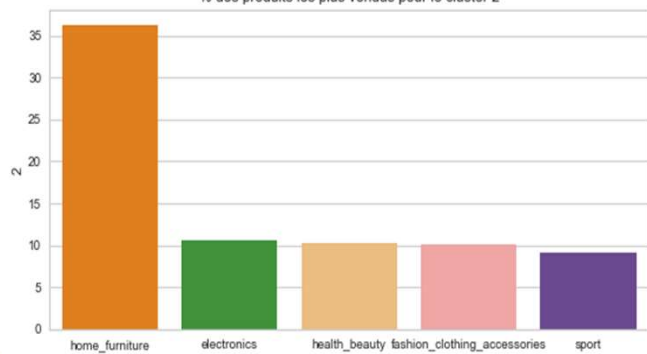


% des produits les plus vendus pour le cluster 1

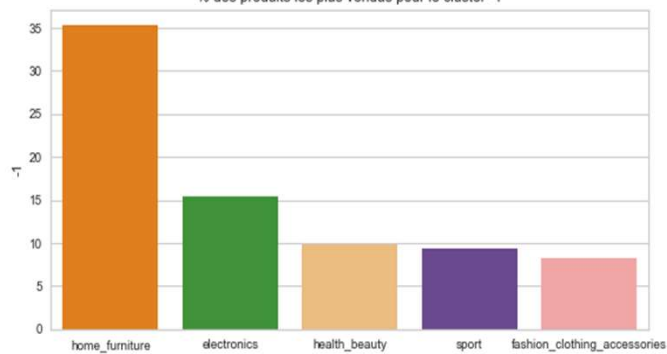


Home furniture
toujours élevé

% des produits les plus vendus pour le cluster 2

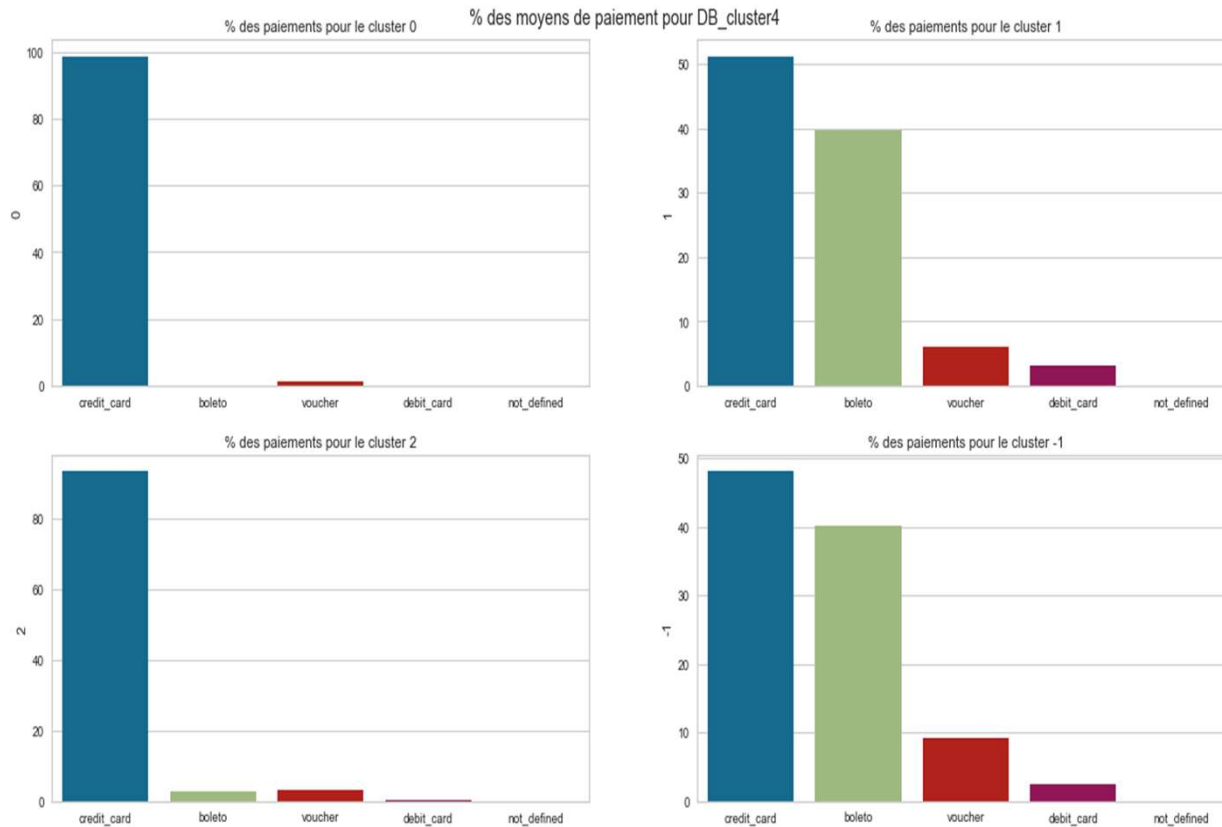


% des produits les plus vendus pour le cluster -1



Plus qu'un seul
cluster privilégiant
aussi l'électronique

Types de paiement :

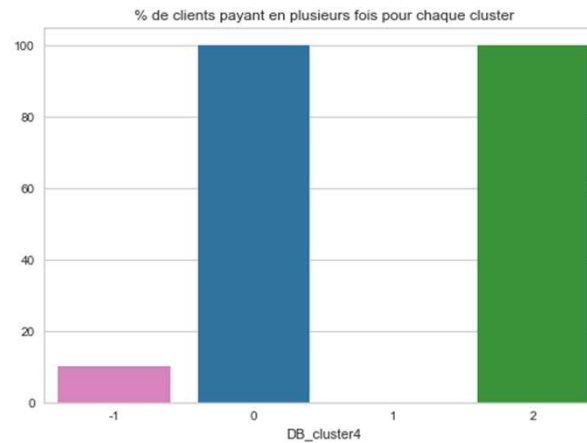
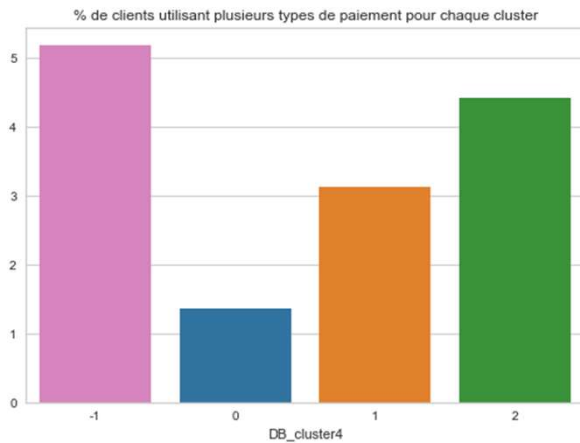


2 clusters uniquement
Credit Card

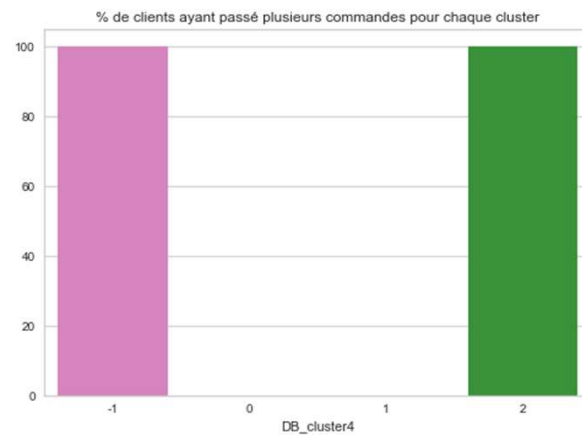
2 clusters Credit Card &
Boleto

Voucher légèrement plus
élevés

Informations sur les paiements :



2 clusters paient
en plusieurs fois

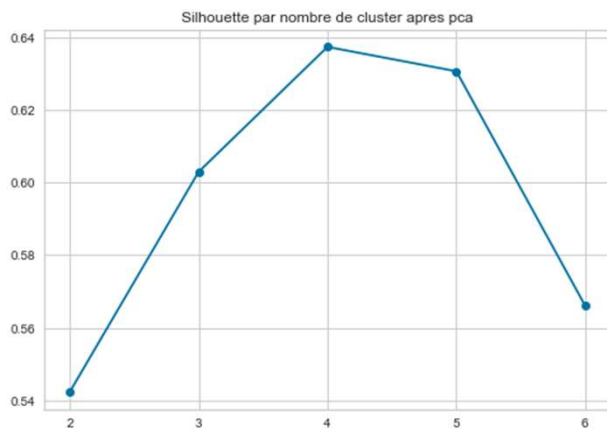


1 cluster + cluster 'bruit'
(2 clusters) ont passé
plusieurs commandes

Clustering hiérarchique

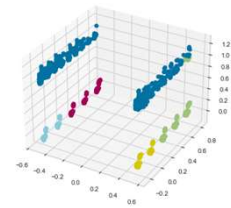
Silhouette et représentation :

Echantillonnage de 50% de la population

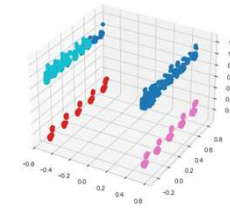


4 clusters

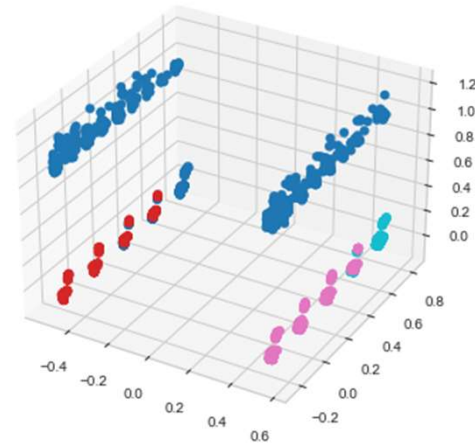
Visualisation des cluster sur les trois axes principaux de la PCA



Visualisation des cluster sur les trois axes principaux de la PCA



Visualisation des cluster sur les trois axes principaux de la PCA



Clusters moins bien définis

The background features abstract geometric shapes in various shades of blue. On the left, a solid light blue triangle points upwards. On the right, a complex arrangement of overlapping triangles in different blue tones (light, medium, and dark) creates a dynamic, layered effect. The central text is positioned in the white space between these blue elements.

Clustering retenu

Modele retenu :

- ▶ kmeans après PCA
- ▶ Profils d'utilisateurs :

Le fidèle

Montant moyen : 148\$
Distance : 560 km
Plusieurs commandes
Paye en plusieurs fois
4,1/5
5% des clients

Le déçu

Montant moyen : 130\$
Distance : 607 km
Commande unique
Paye en une fois
1,9/5
10% des clients

L'exigeant

Montant moyen : 212\$
Distance : 700 km
Commande unique
Paye en plusieurs fois
1,8/5
10% des clients

L'économe

Montant moyen : 116\$
Distance : 542 km
Commande unique
Paye en une fois
4,7/5
35% des clients

Le satisfait

Montant moyen : 195\$
Distance : 634 km
Commande unique
Paye en plusieurs fois
4,7/5
40% des clients

Actions à entreprendre :

Fidéliser clients satisfaits

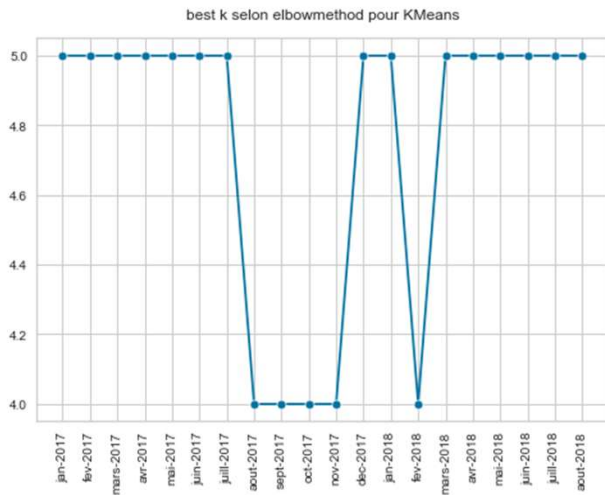
Rechercher l'explication des clients déçus : délai, commentaires, type de produits...

Clustering sur les axes principaux de la PCA : pertinence (enlever paiement en plusieurs fois ?)
satisfaction,
commandes multiples,
paiements en plusieurs fois
distance

The background features abstract geometric shapes in various shades of blue. On the left, a light blue trapezoidal shape points towards the center. On the right, a complex arrangement of overlapping triangles and polygons in different blue tones creates a dynamic, layered effect. The central text is positioned on a white background that tapers towards the right, where it meets the blue geometric shapes.

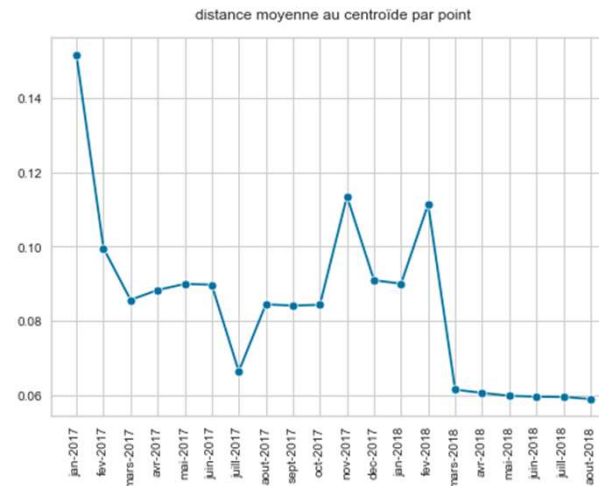
Maintenance et fréquence de mise à jour

Kelbow method et distances :



Alternance entre 4 et 5 clusters

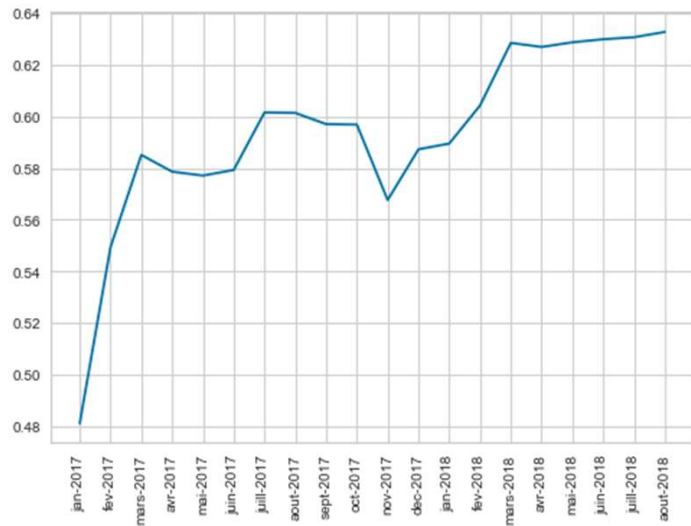
Shift au deuxième semestre et au mois de février



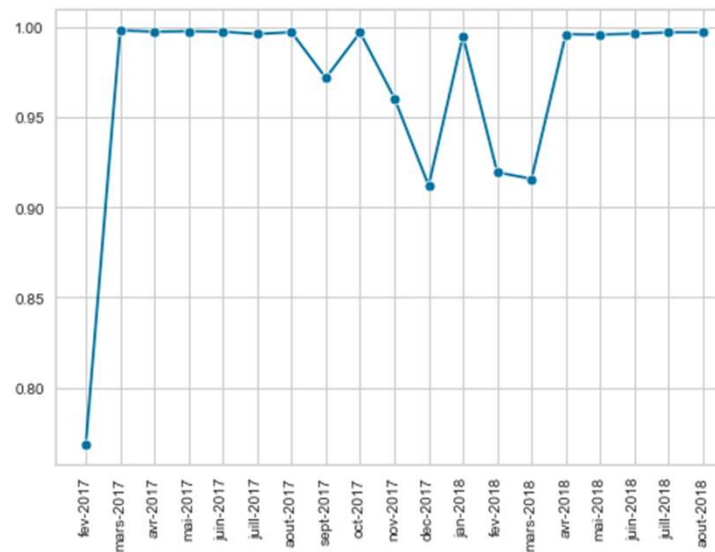
Pics aux mois de novembre et février

Maintenance Kmeans à 5 clusters :

Silhouette_score en fonction des mois pour KMeans à 5 clusters



ari_score en fonction des mois pour KMeans à 5 clusters



Amélioration silhouette avec données

Baisse au mois de novembre

Amélioration deuxième semestre et en février

Proche de 1 → Clusters stables

Chute en Novembre et février

Maintenance trimestrielle