Segmentez des clients d'un site e-commerce

PROJET 05/ Openclassrooms

Gulsum Kapanoglu

Dans ce Project..

- ✓Problématique
- ✓Nettoyage des données
- ✓Exploration des données et Feature Engineering
- ✓Modelés d'apprentissage non supervisé
- ✓Simulation pour déterminer la fréquence nécessaire de m-à-j du modèle de segmentation.
- ✓Conclusion

Problématique

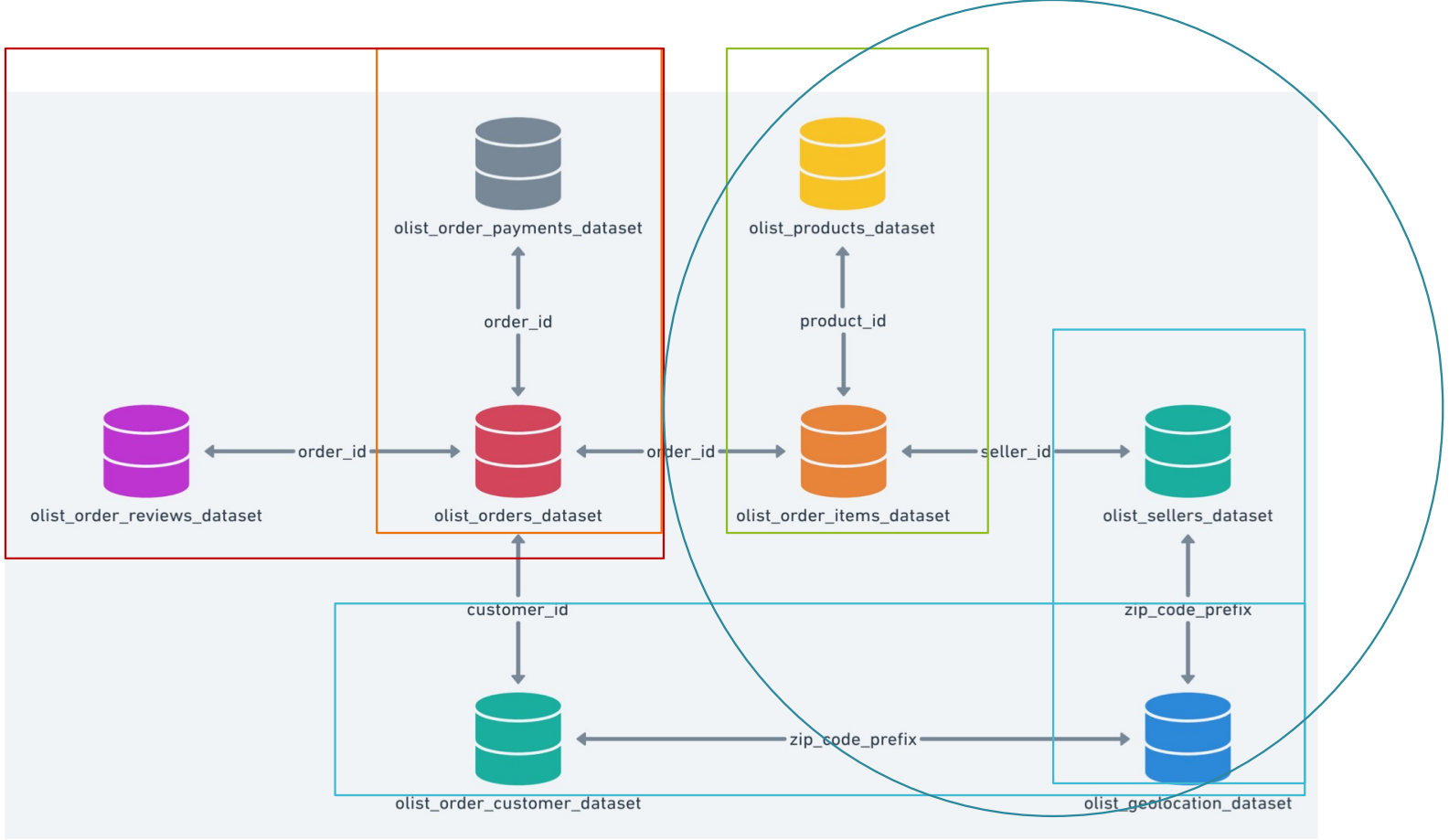
Problématique

Olist voudrais une **segmentation des clients** pour ses équipes d'e-commerce et aussi pour qu'elles pourront utiliser au quotidien pour leurs campagnes de communication.

Nos objectif est de **comprendre les différents types d'utilisateurs** grâce à leur comportement et à leurs données personnelles.

Il voudrai aussi qu'une **proposition de contrat de maintenance** basée sur une analyse de la stabilité des segments au cours du temps.

Données



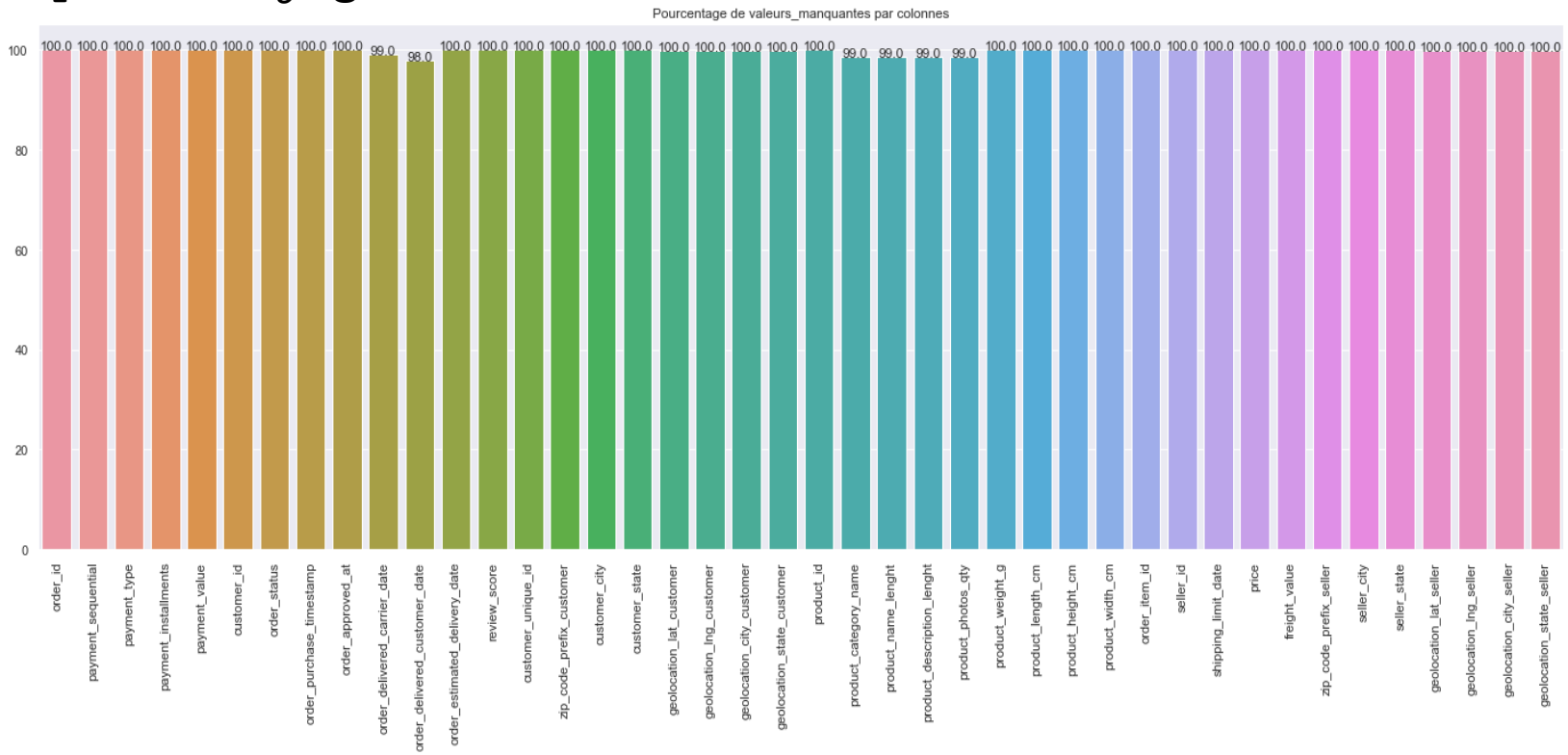
- ✓ Base de données anonymisée
- ✓ Clients: 99441
- ✓ Produits achetés
- ✓ Historique des commandes
- ✓ Commentaires de satisfaction
- ✓ Vendeurs

Nettoyage des données

Nettoyage des données

J'ai détecté et supprimé des données manquantes et infinies

Apres nettoyage:

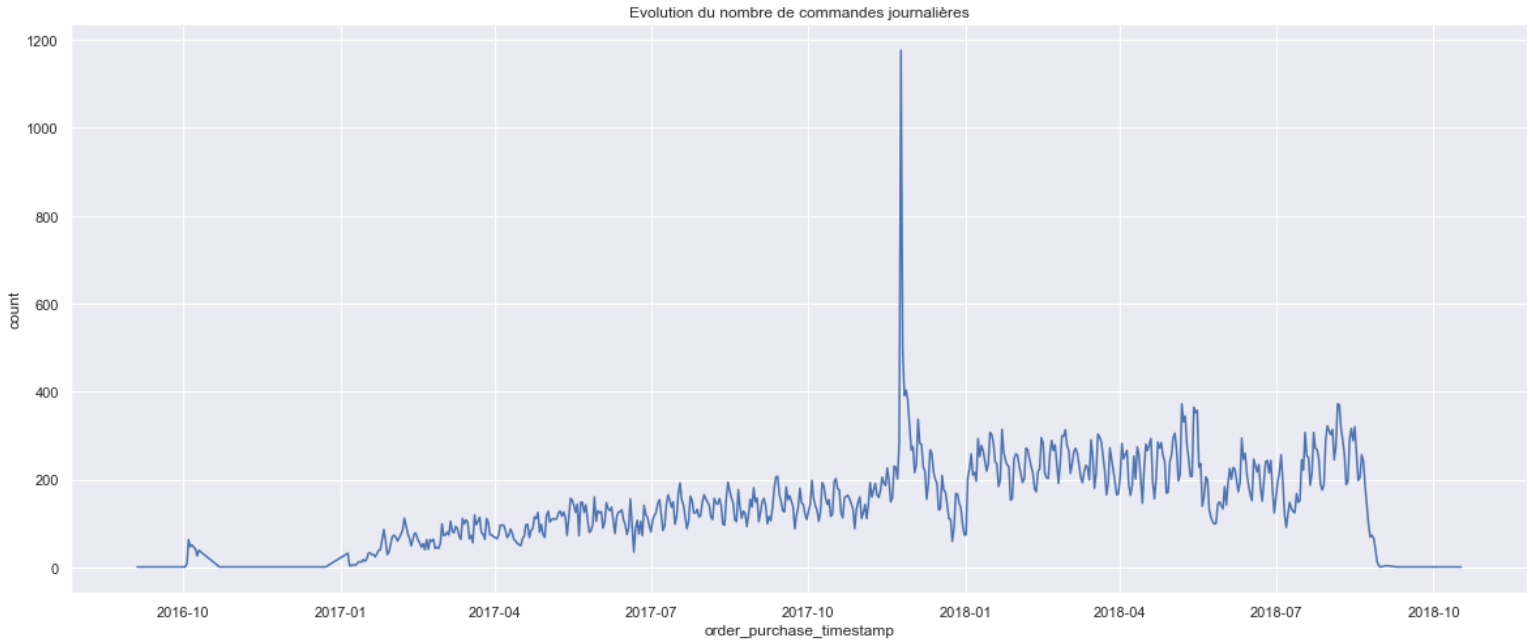


J'ai continué avec le feature engineering.

Exploration des données
&
Feature engineering

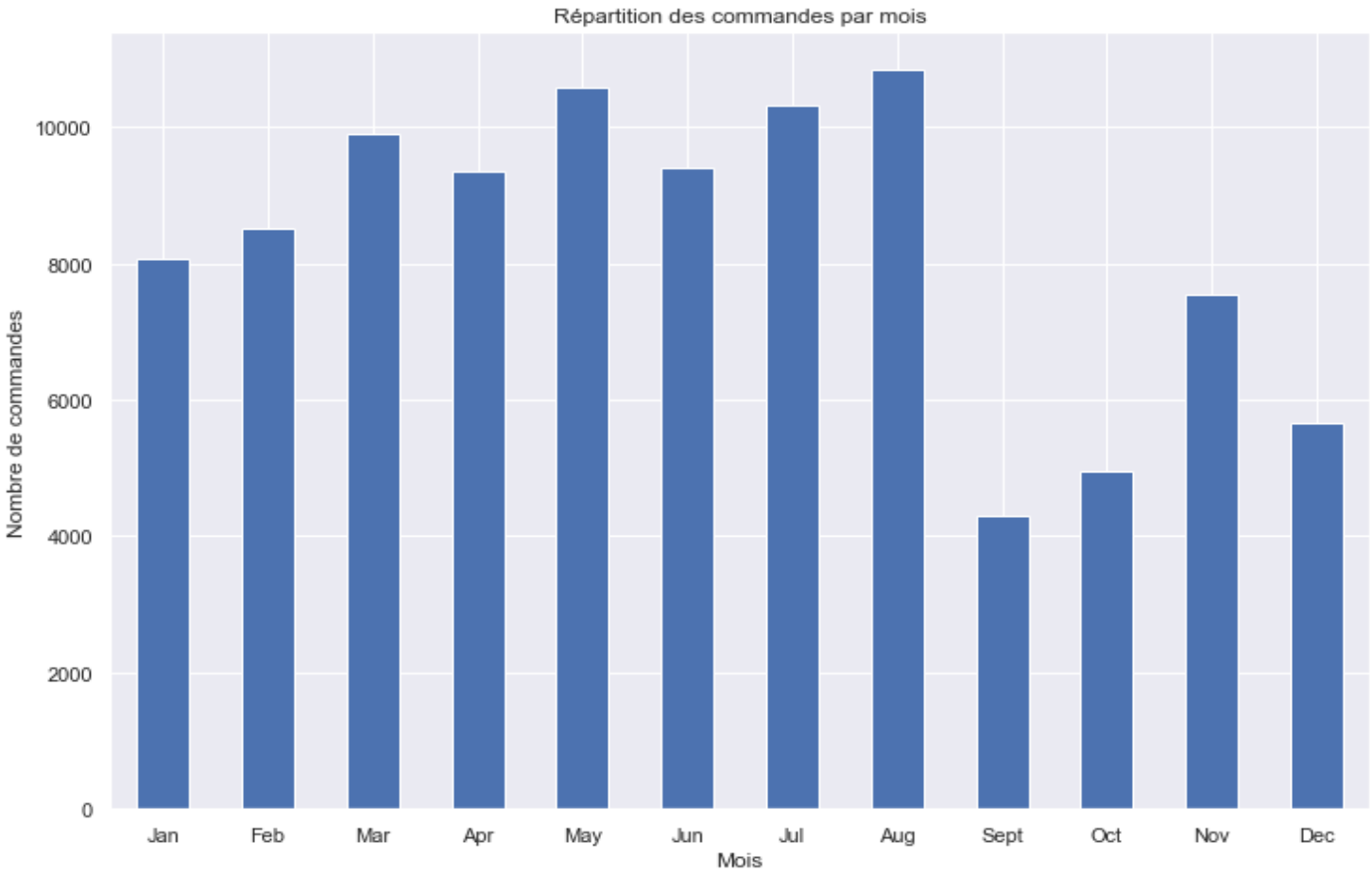
En termes de Commandes

il y a des commandes en cours ou annulées. Ne traitons que ceux qui ont été livrés. Nous ne pouvons pas considérer comme clients ceux qui n'ont pas effectué d'opération d'achat de produit sur le site.



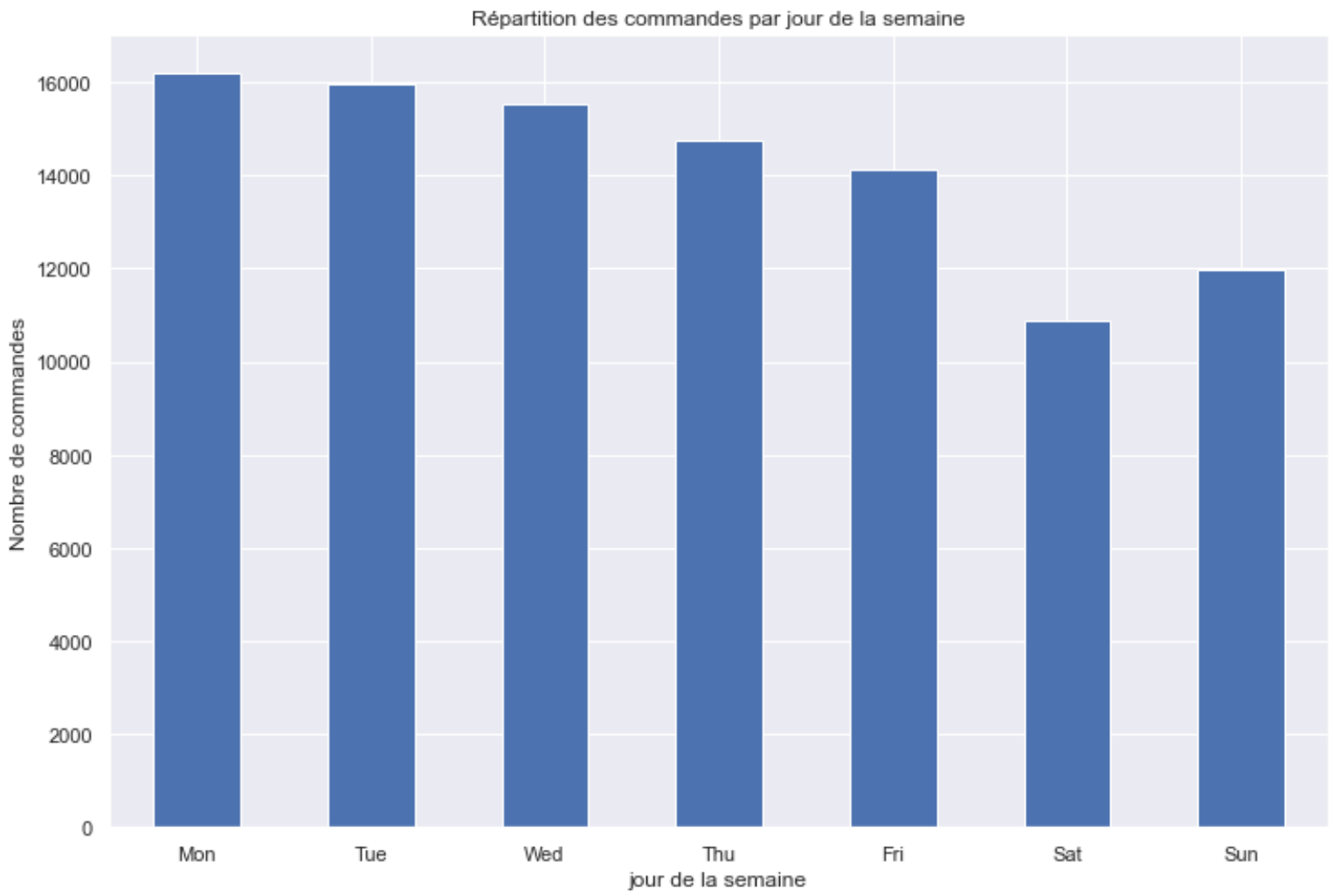
Nous constatons une augmentation de fin d'année en 2017 ; ça doit être les achats de Noël. Et aussi On peut voir la répartition du nombre de commandes par mois, jour de la semaine ou heure de la journée

En termes de Commandes

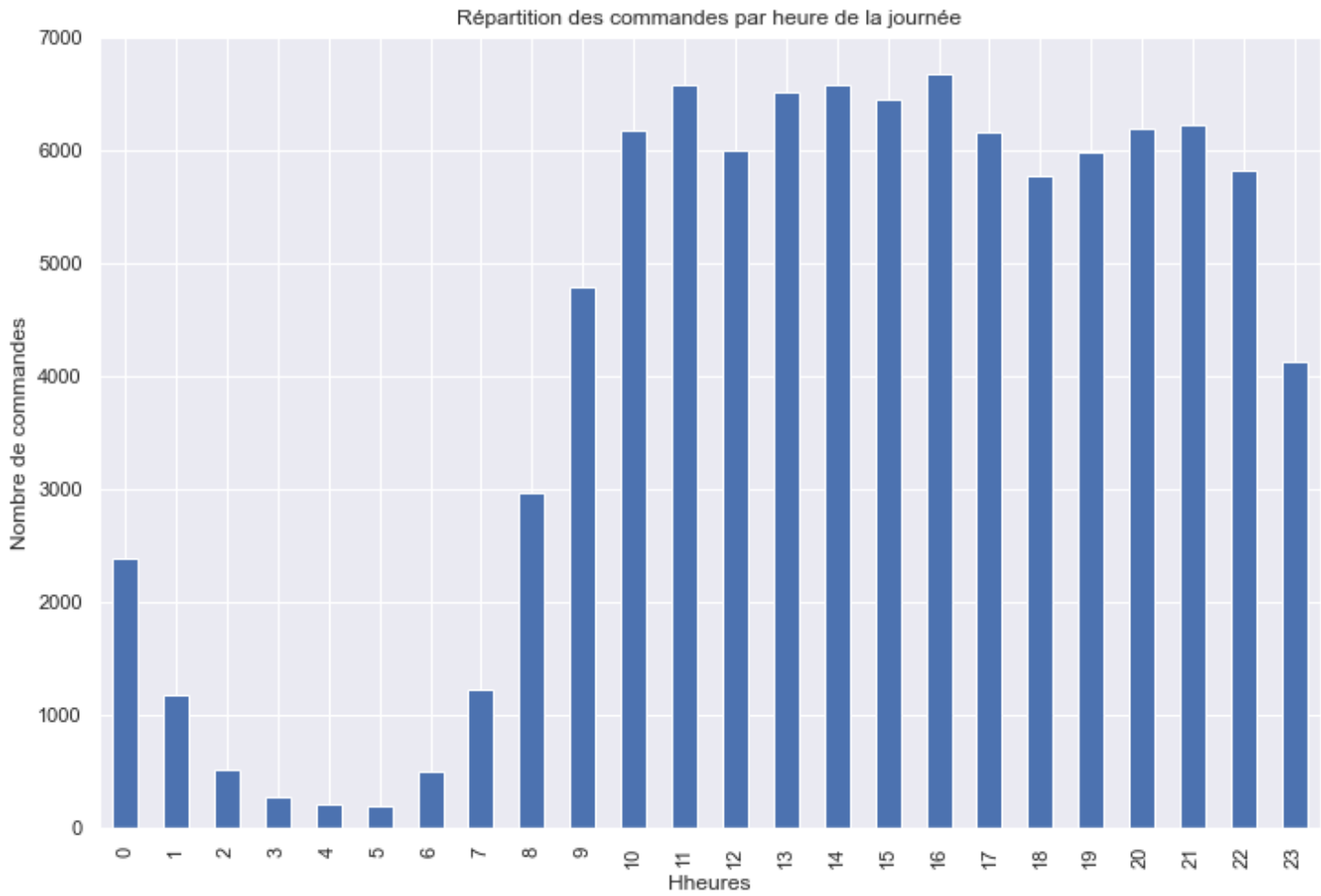


Nous voyons des ventes avec une forte augmentation en août, ceux-ci pourraient être des achats de retour à l'école.

En termes de Commandes

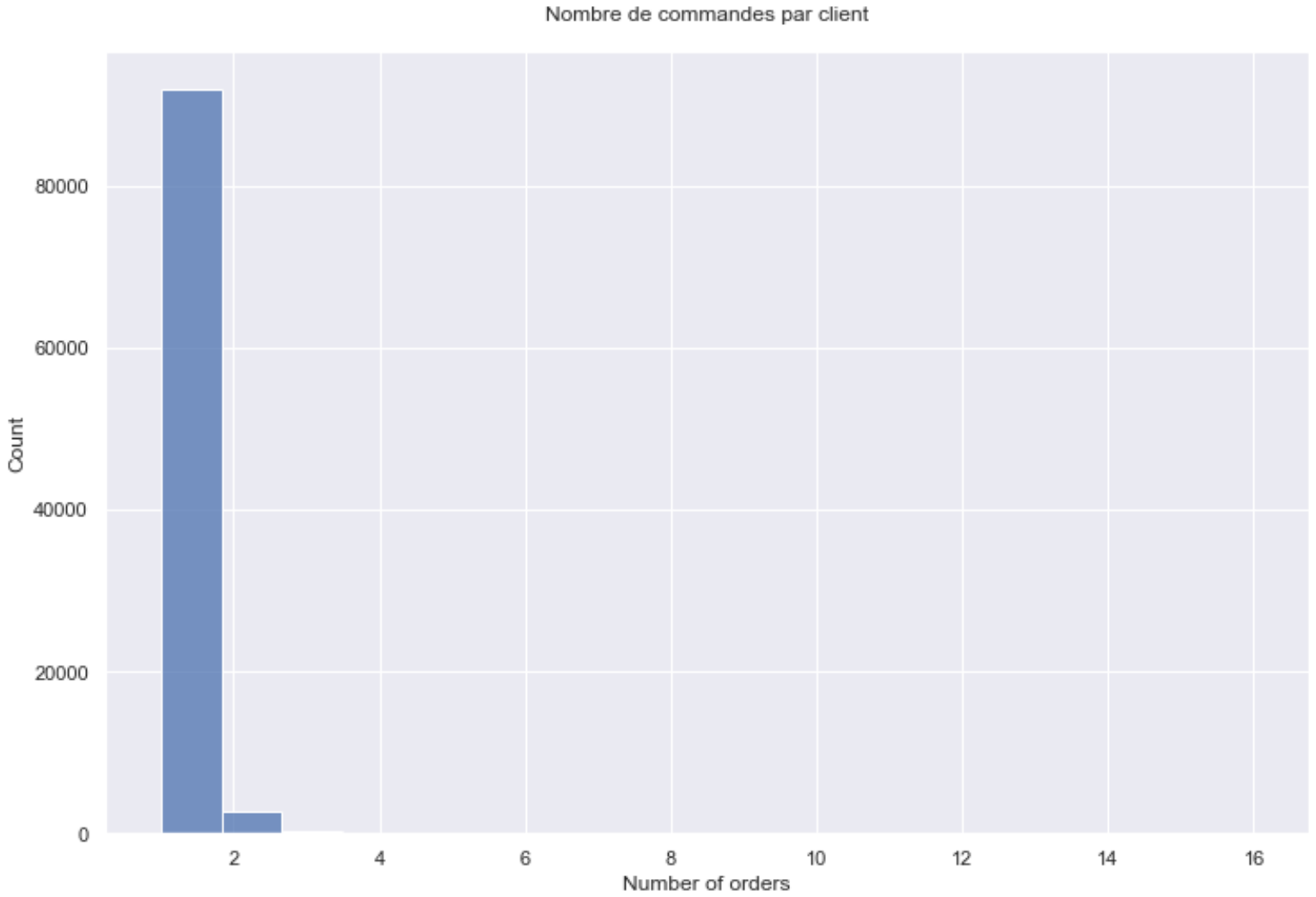


En termes de Commandes



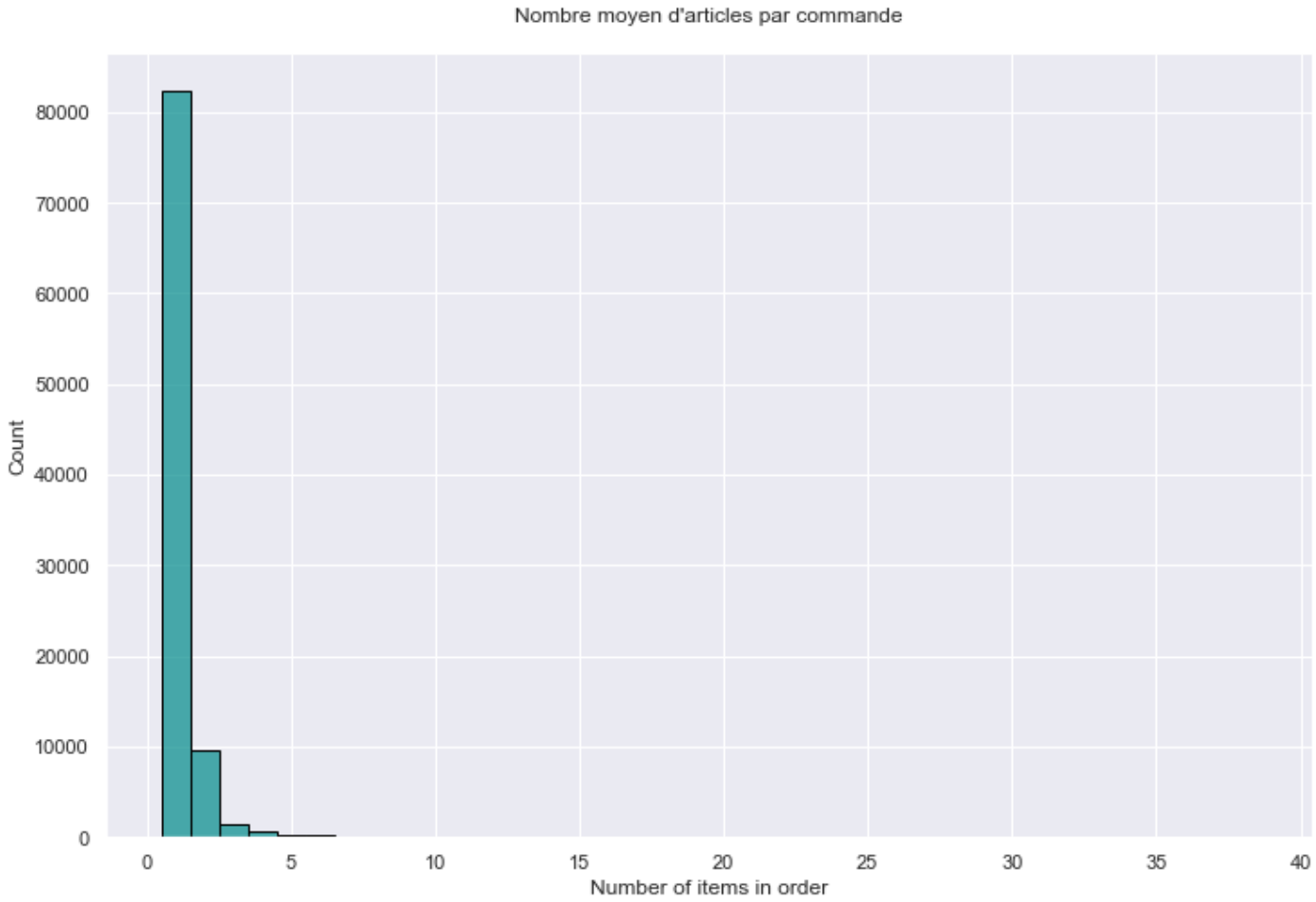
- Pendant la journée, nous constatons que la plupart des commandes sont passées près de la pause déjeuner ou vers la fin de la journée de travail. (11h à 16h)

En termes de Clients



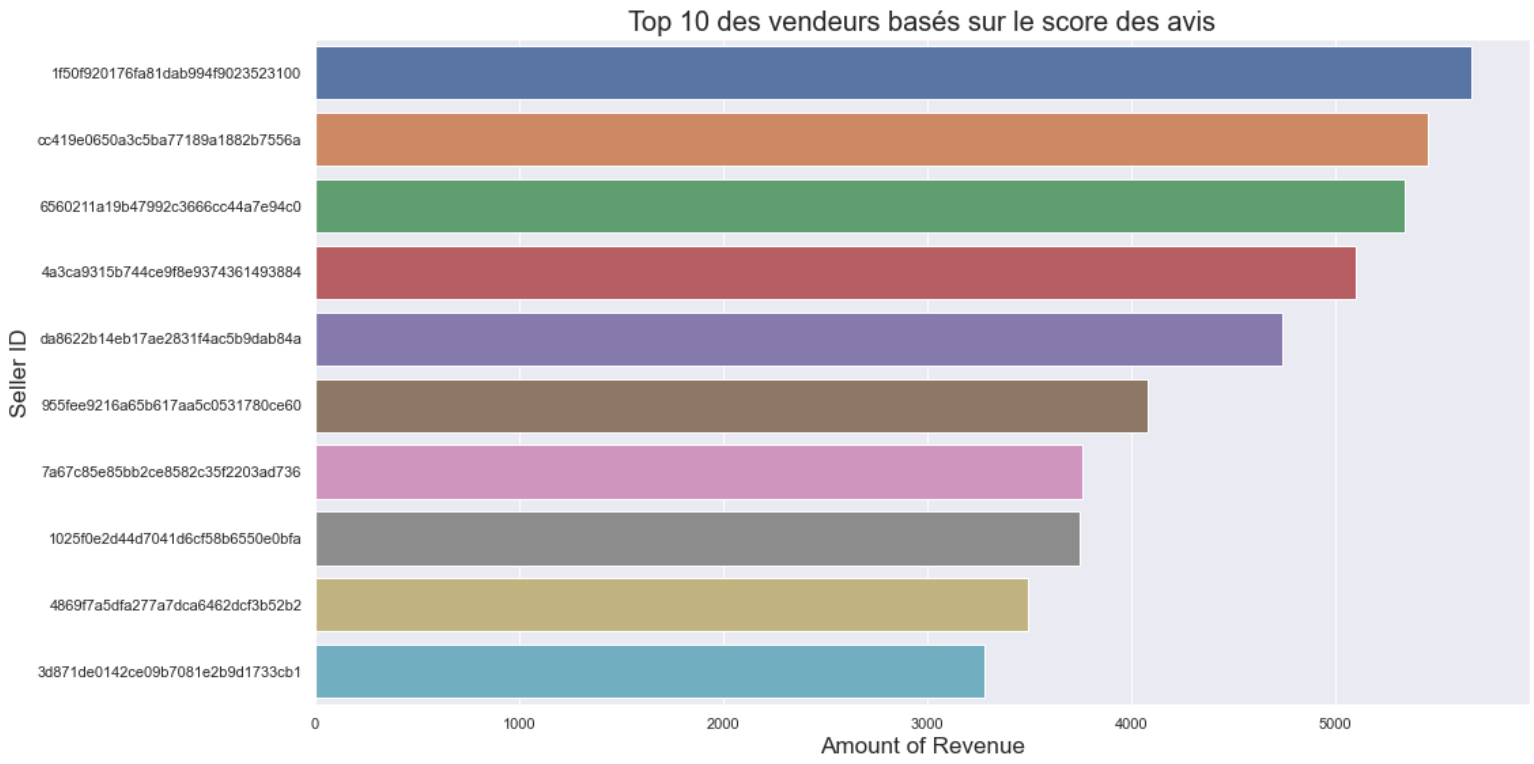
On peut voir que seuls environ 3% des clients ont réalisé plus d'une commande.

En termes de Clients

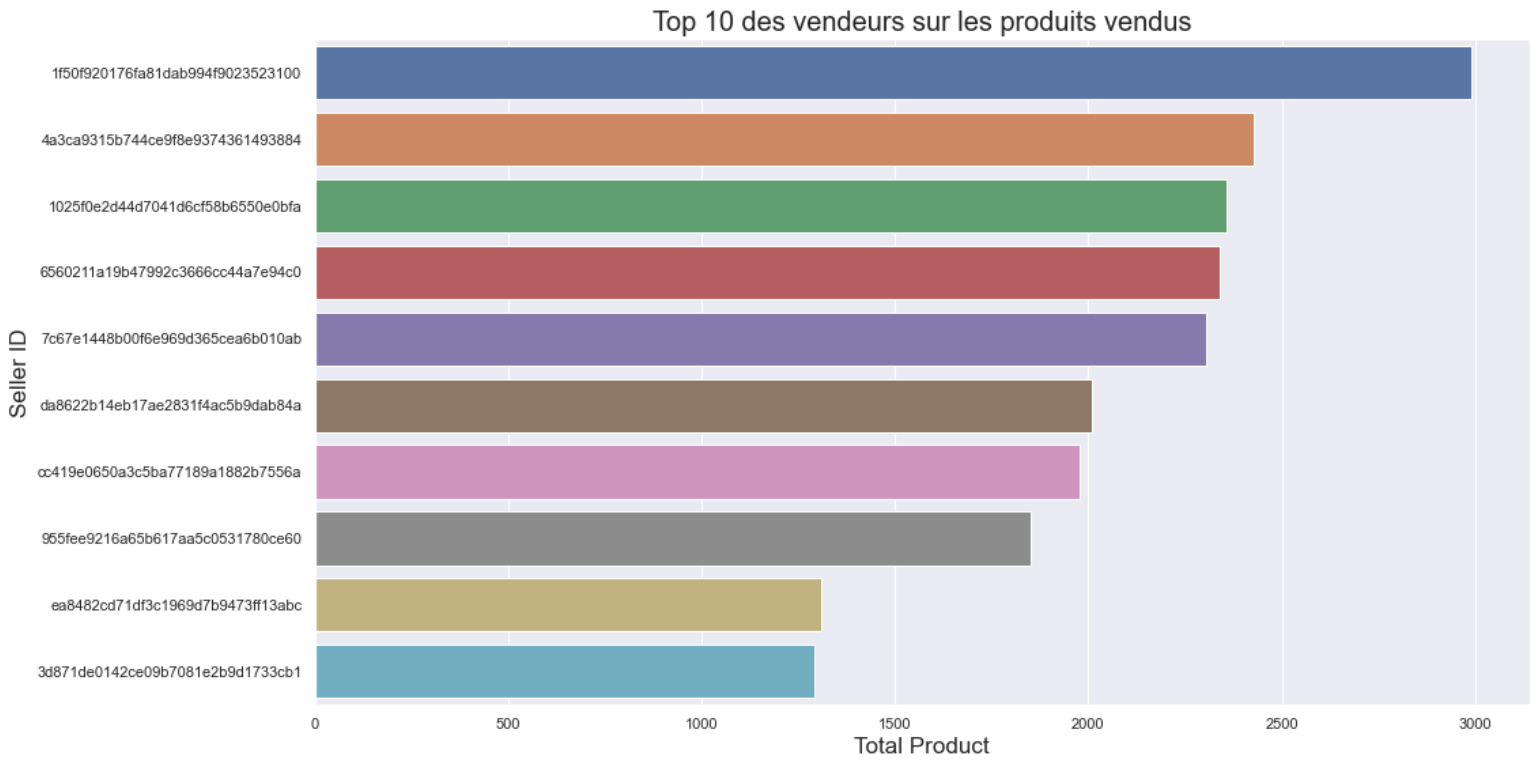


Dans ce graphique, nous voyons le nombre moyen de produits par commande pour chaque client

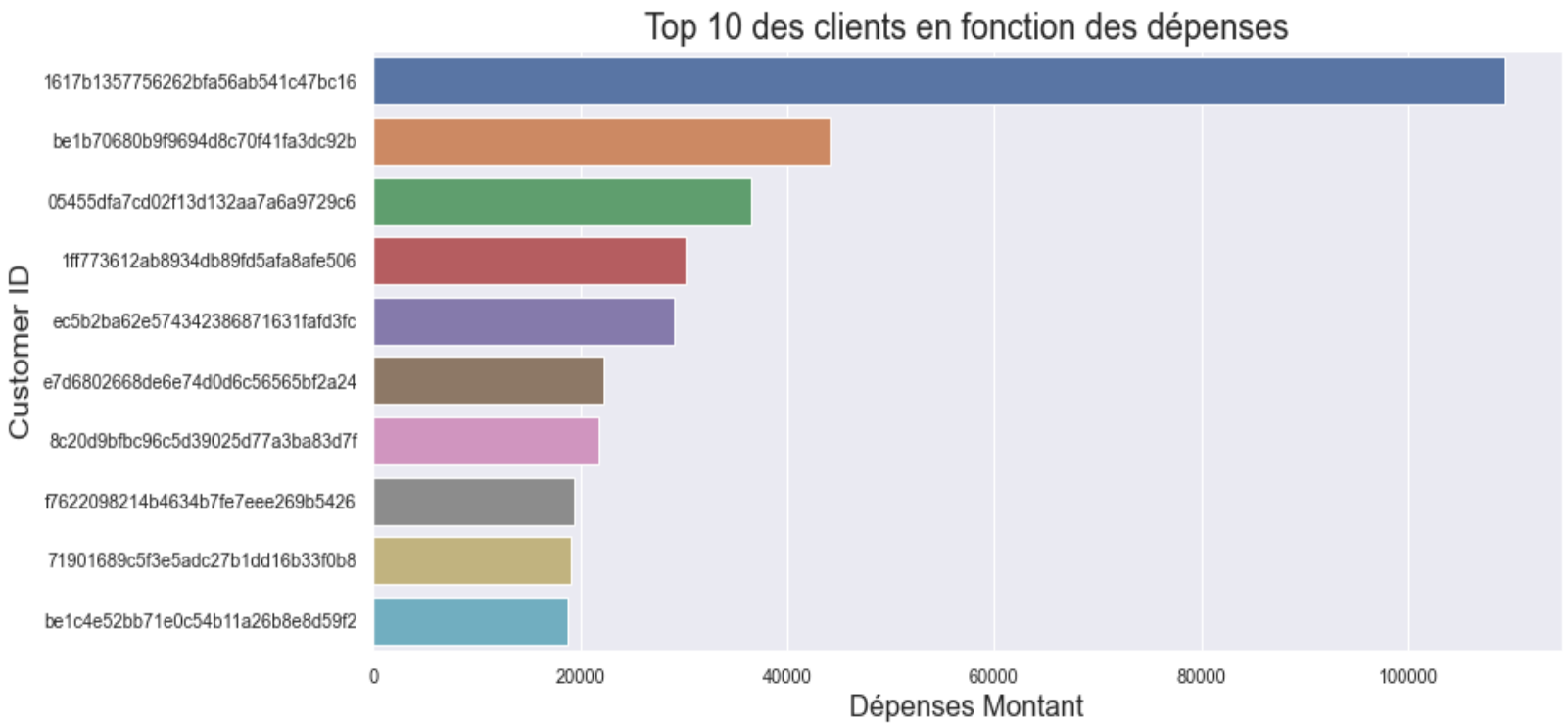
Top 10 Meilleurs acheteurs



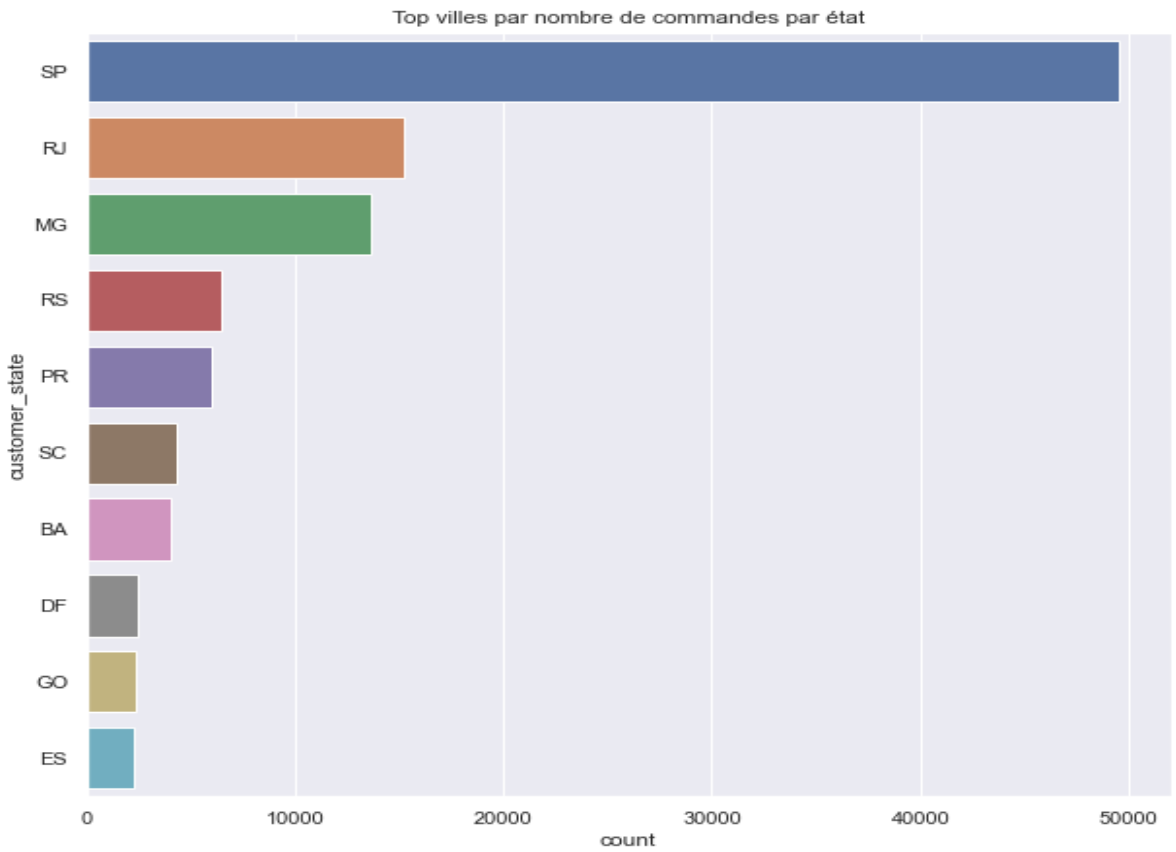
Top 10 Meilleurs acheteurs



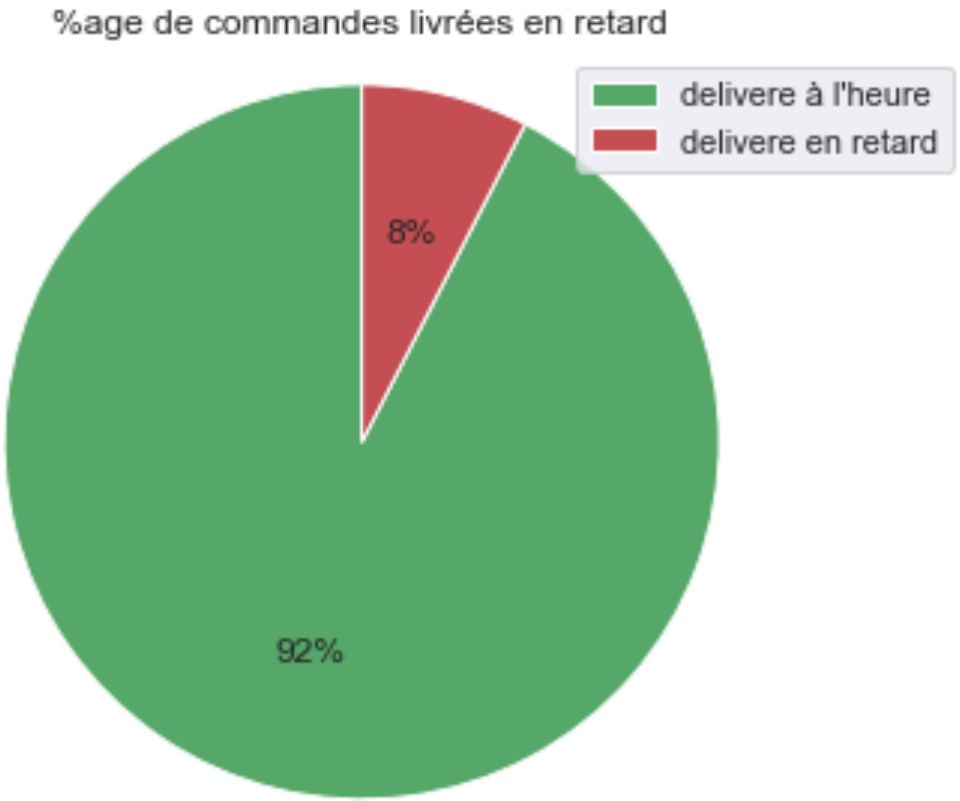
Top 10 Meilleurs acheteurs



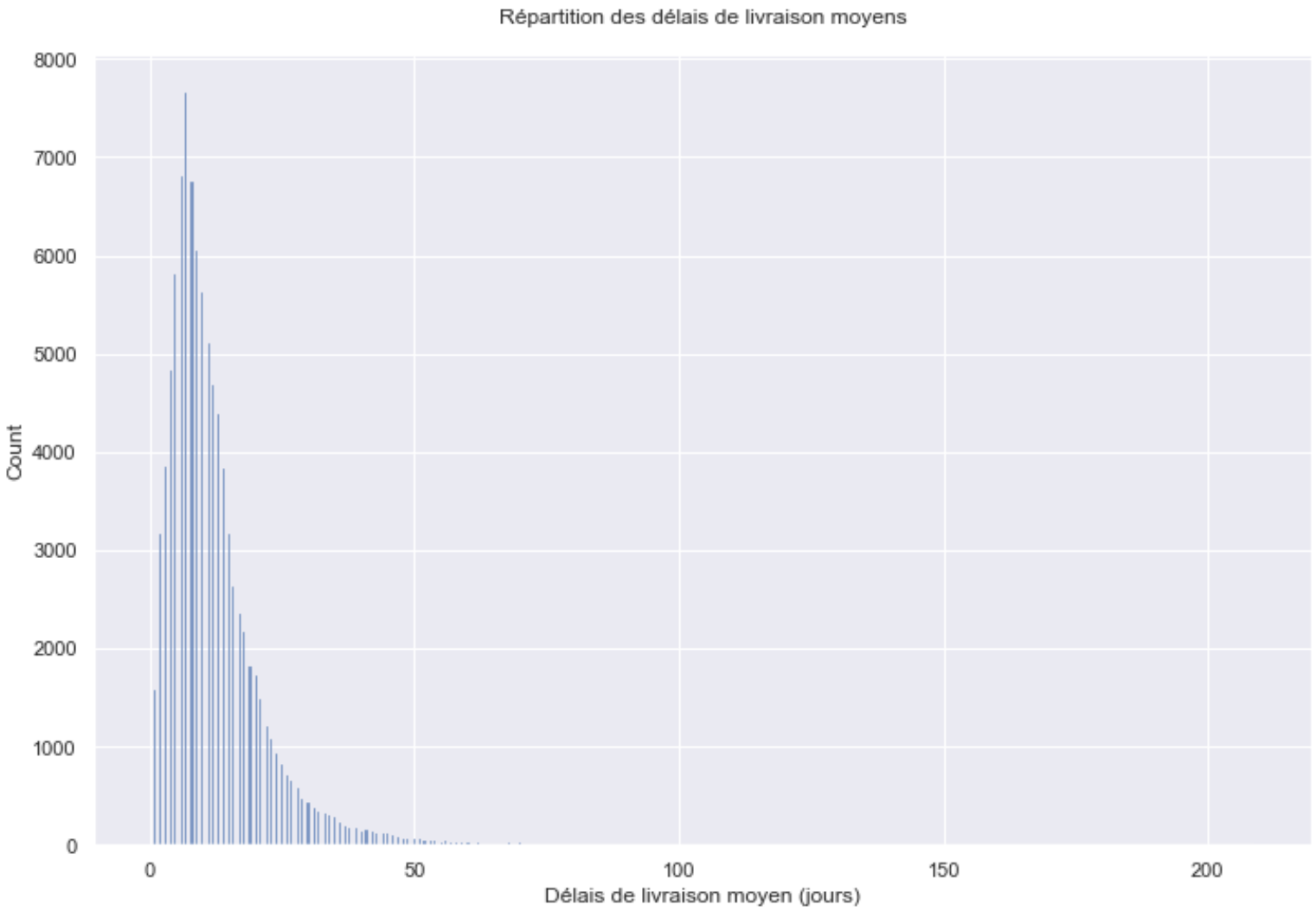
En termes de Commande et Son Temps



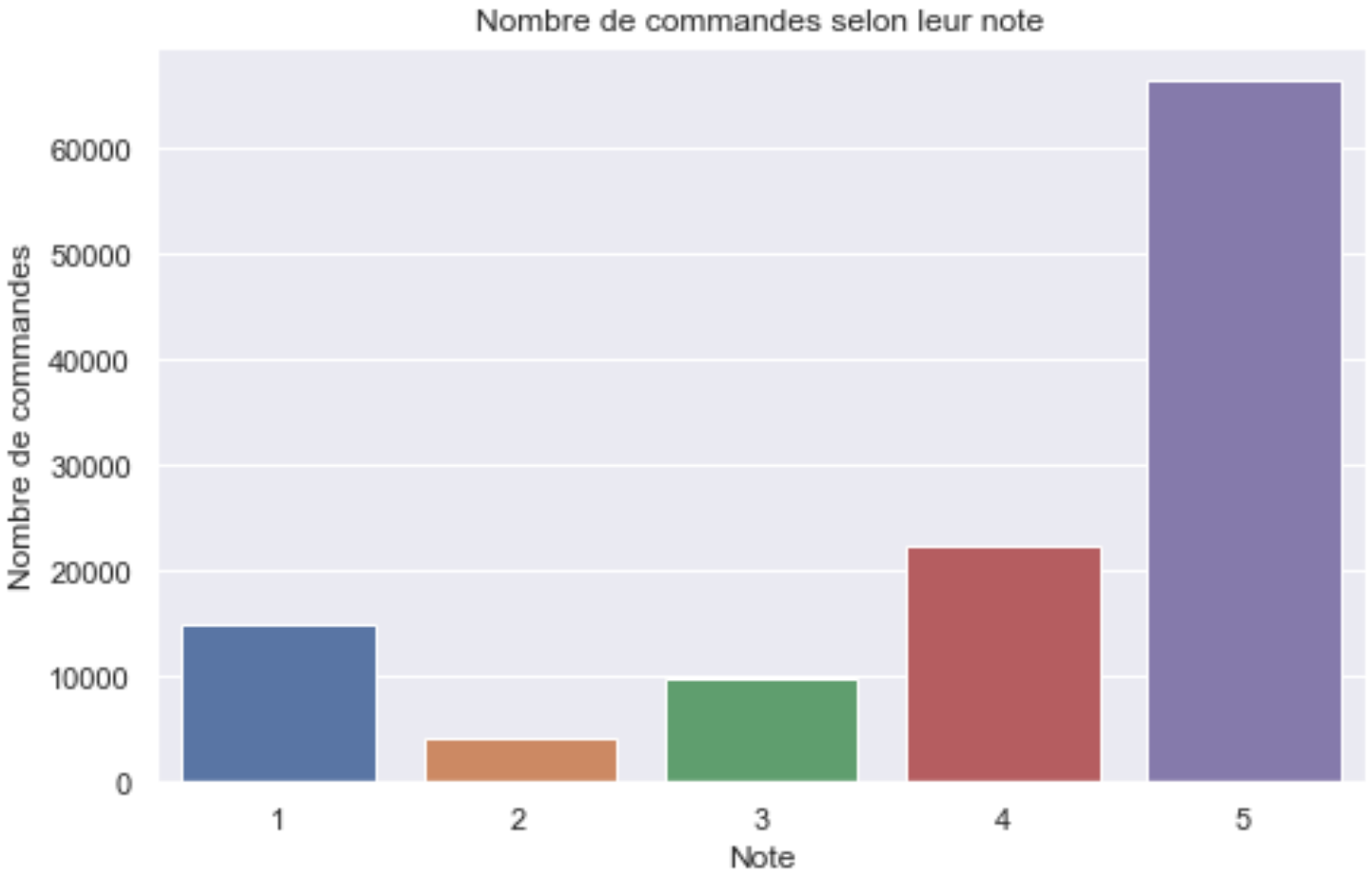
En termes de Livraison et Review



En termes de Livraison et Review

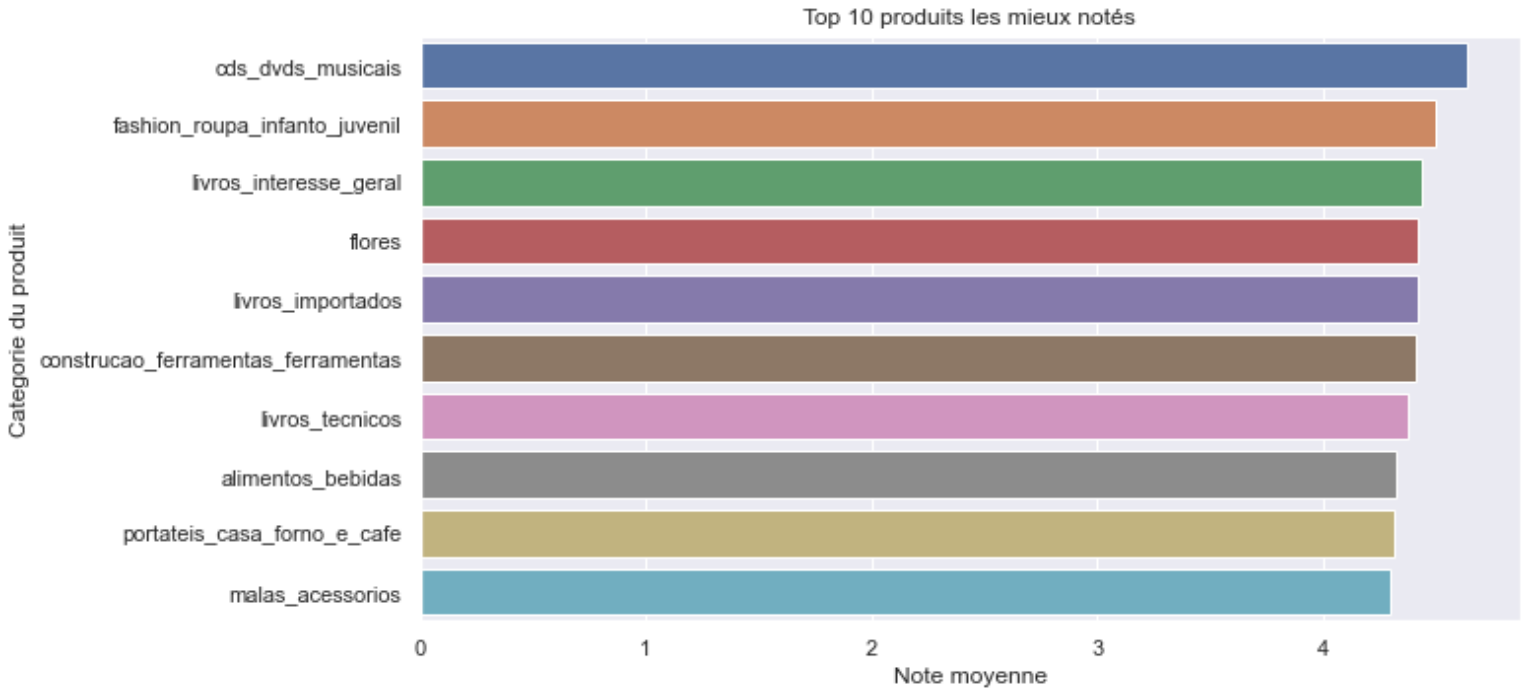


En termes de Livraison et Review

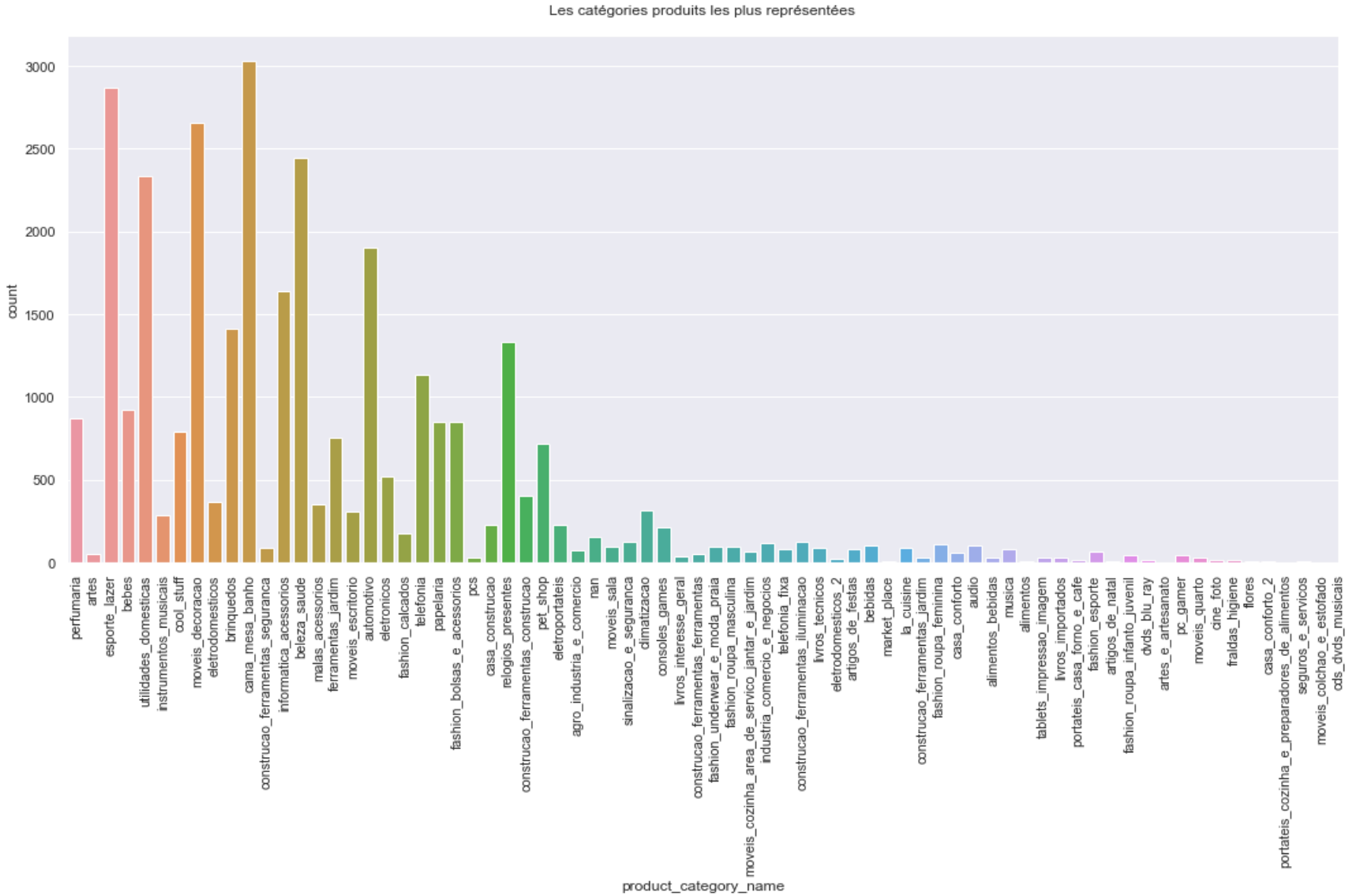


- Nous constatons que les clients qui achètent les produits sont généralement satisfaits

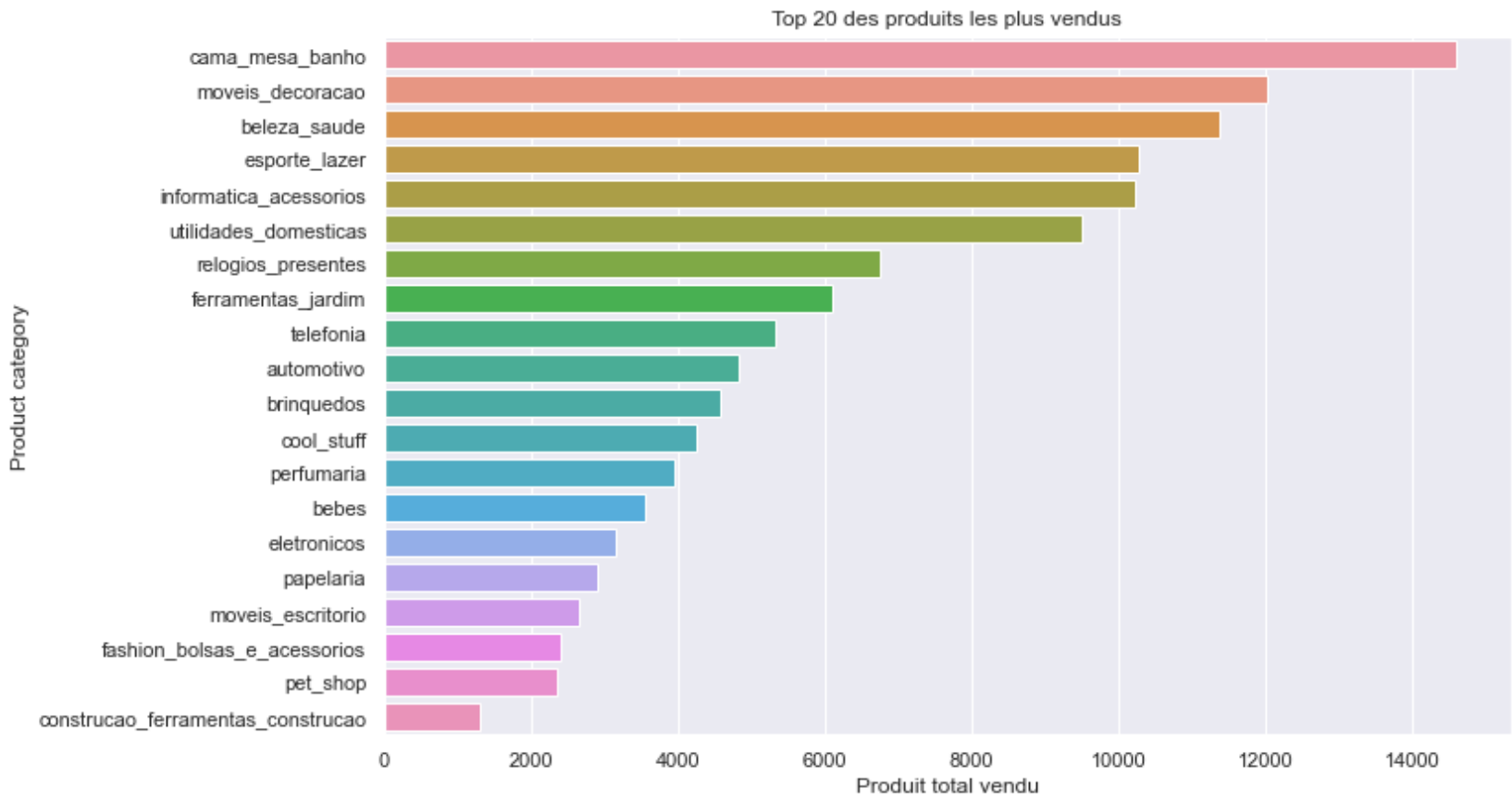
En termes de Livraison et Review



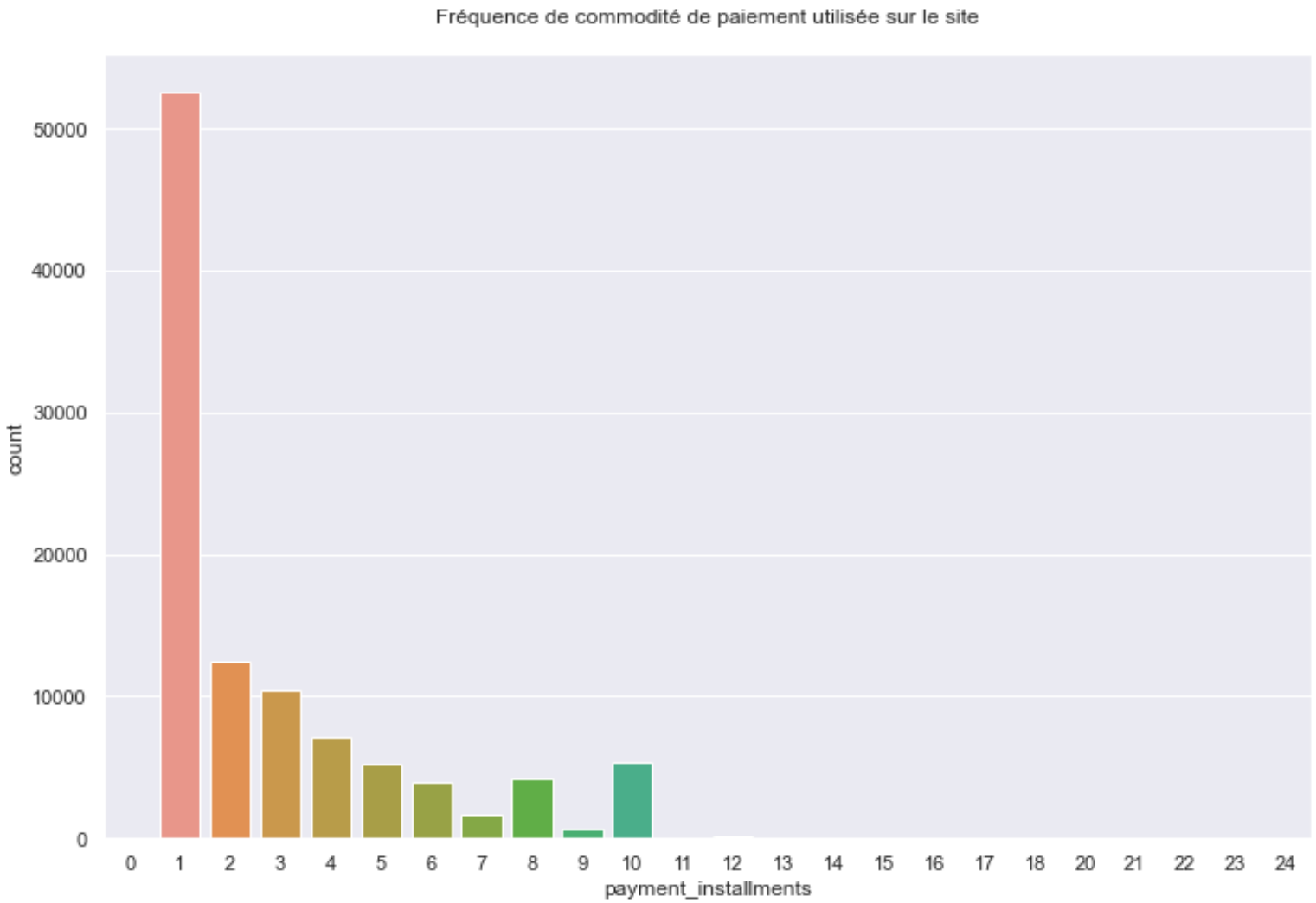
En termes de Produits



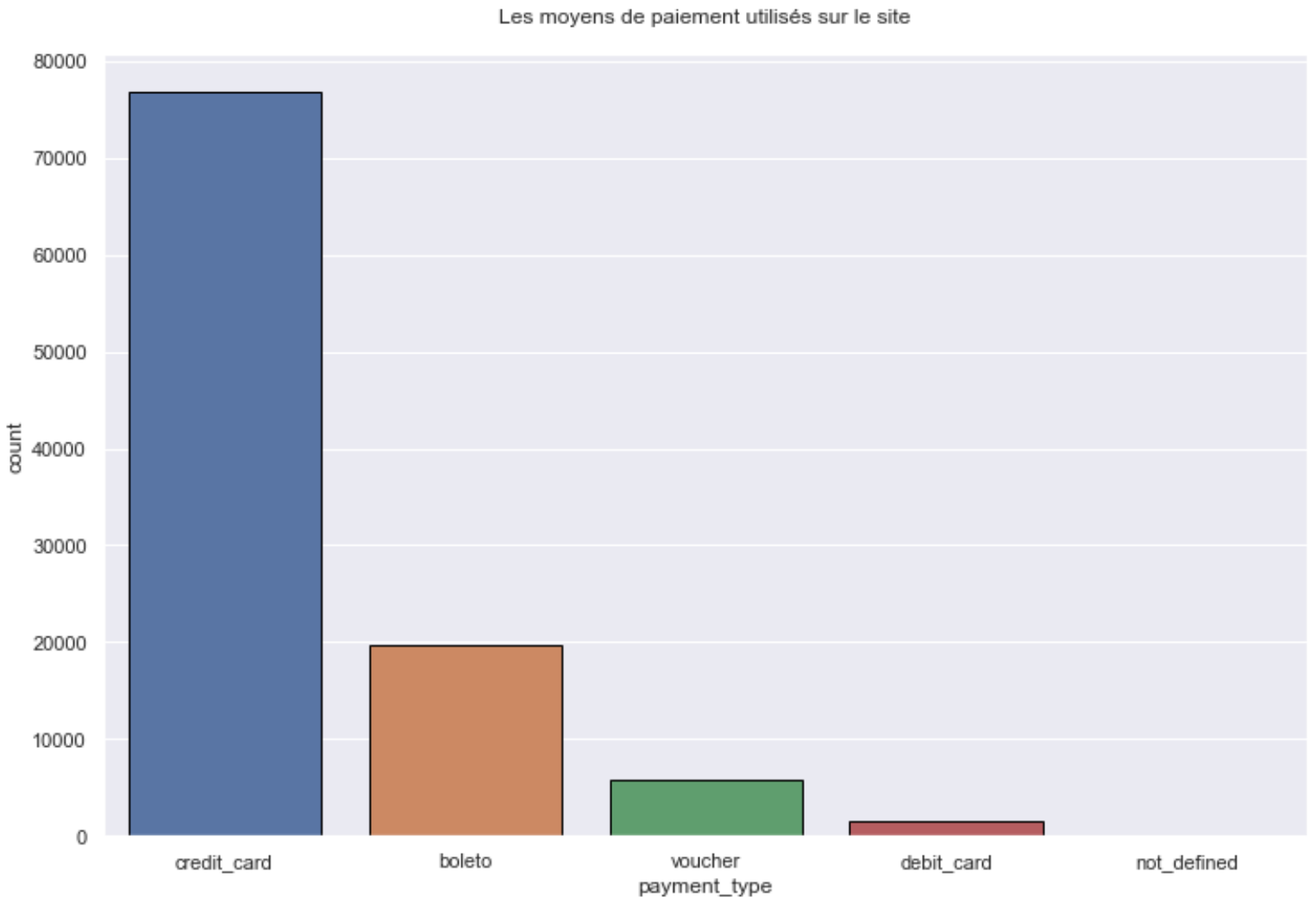
En termes de Produits



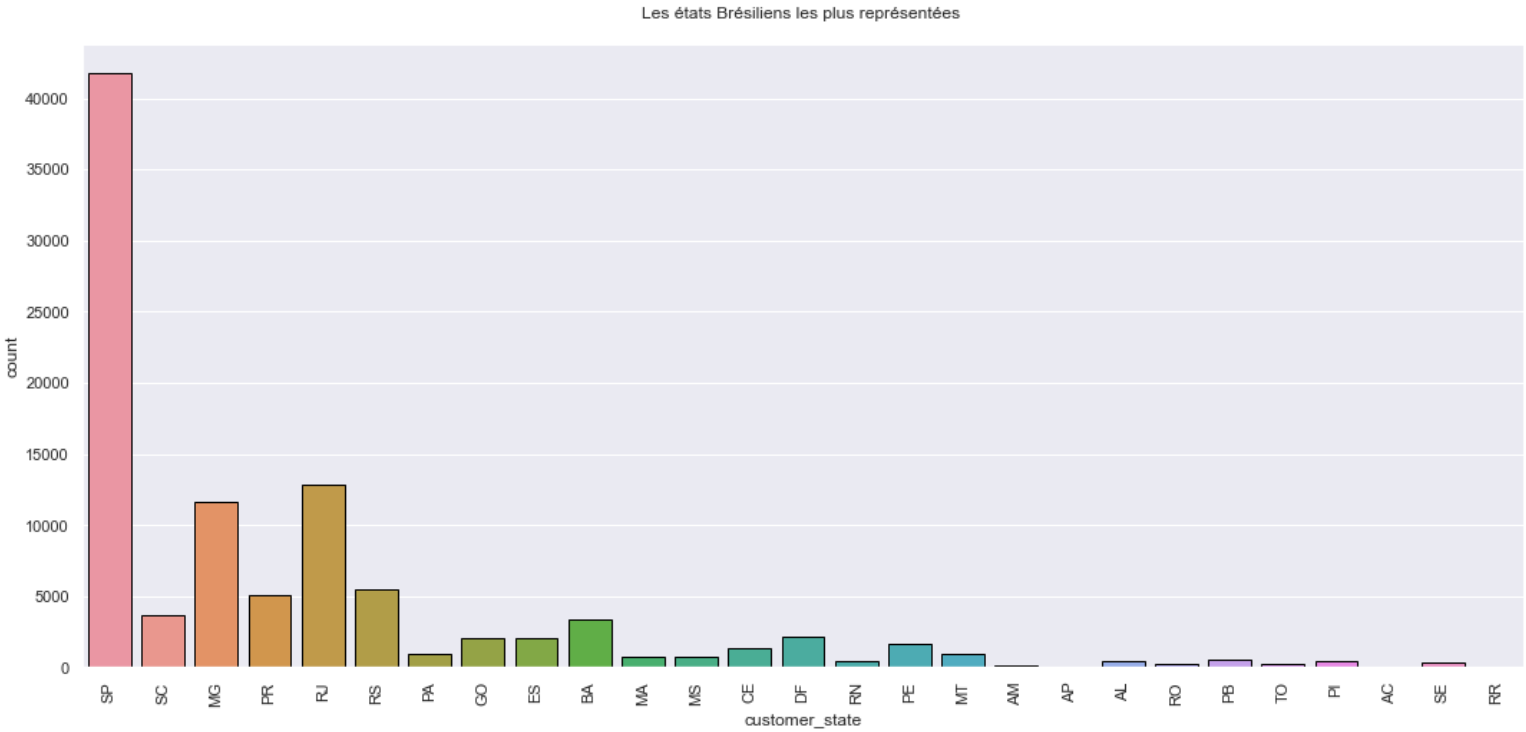
En termes de Produits



En termes de Produits



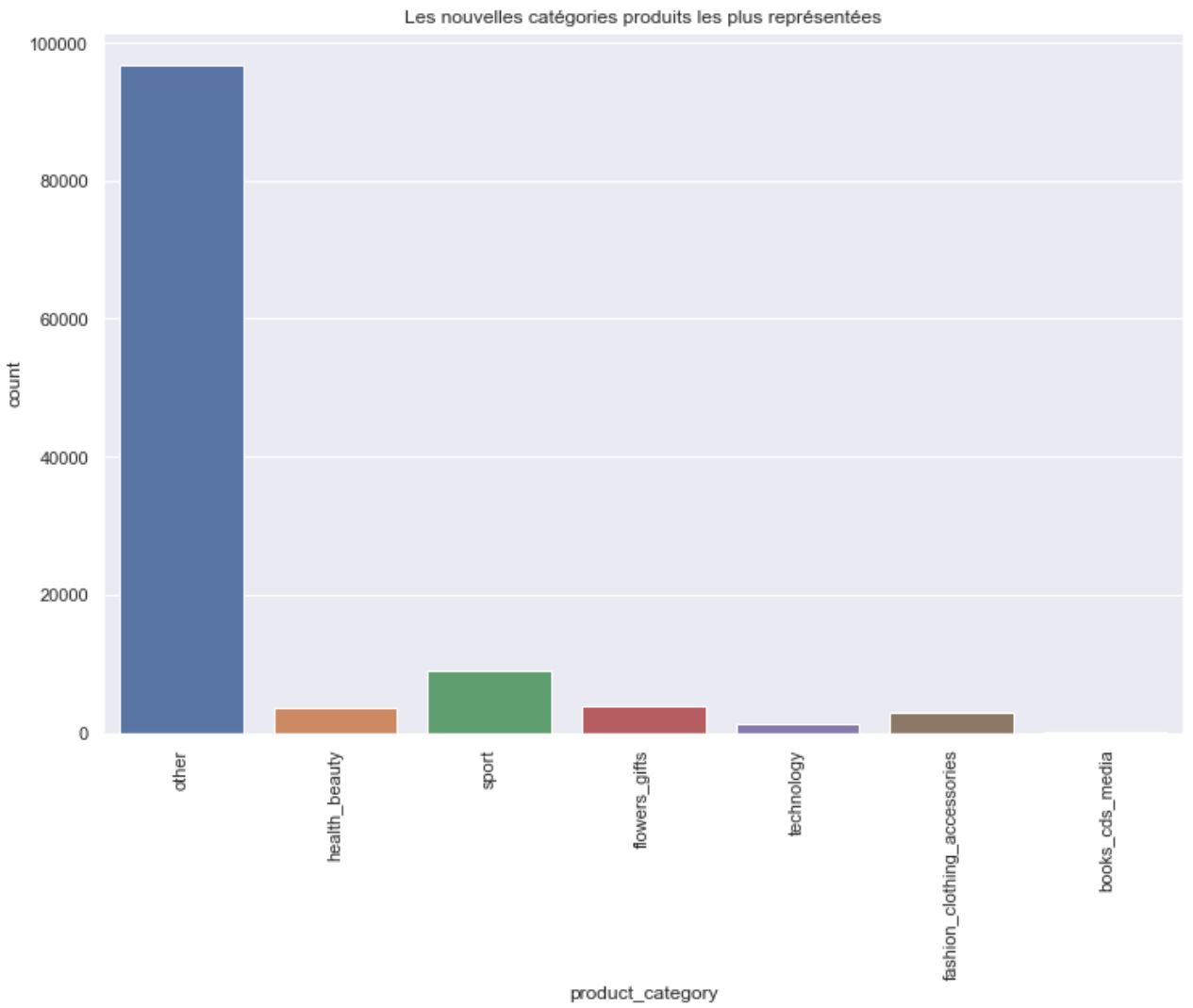
En termes de Produits



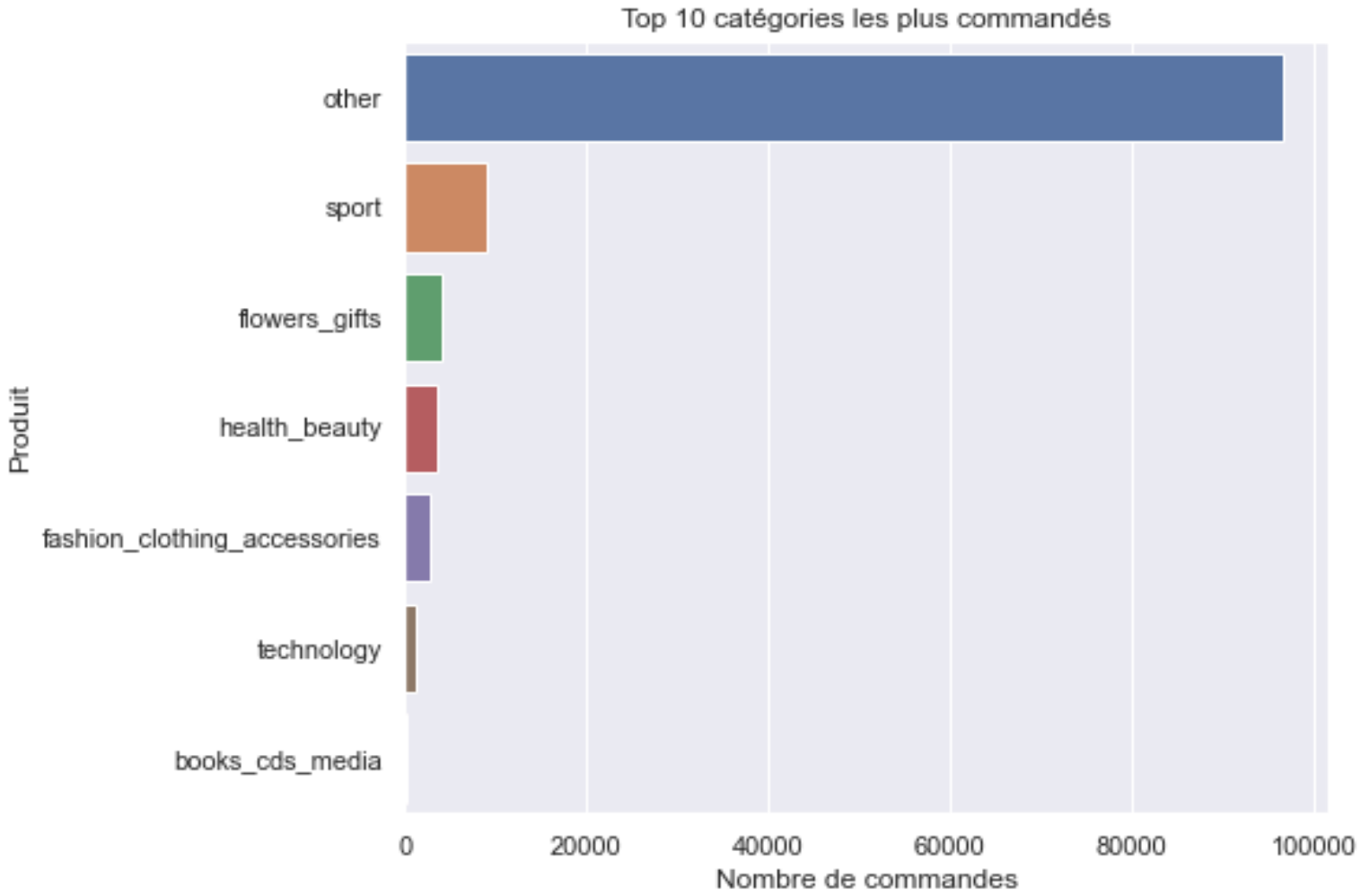
En termes de Produits et Ses catégories

- Il y a 72 catégories au total. Ceux-ci peuvent causer des problèmes dans le encodage des données. Pour un examen facile, nous rassemblerons les catégories sous les rubriques les plus larges.
- Fashion, clothing and accessories Health and Beauty Toys and baby equipment Books, CDs and other physical media Groceries, food and drink Technology (including phones and computers), Home and furniture Flowers and gifts Other et nous ajoutons la catégorie sport.

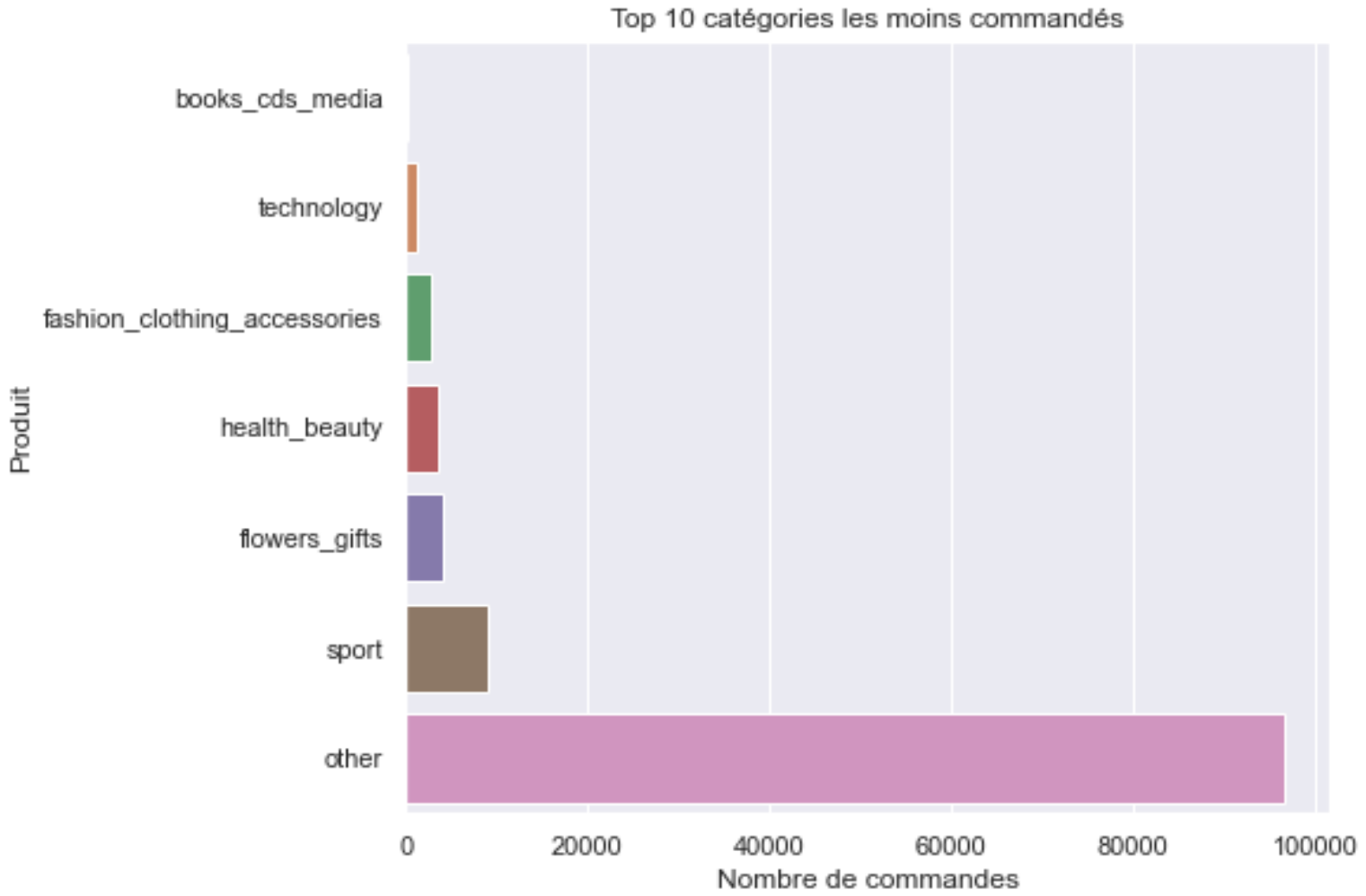
En termes de Produits et Ses catégories



En termes de Produits et Ses catégories



En termes de Produits et Ses catégories



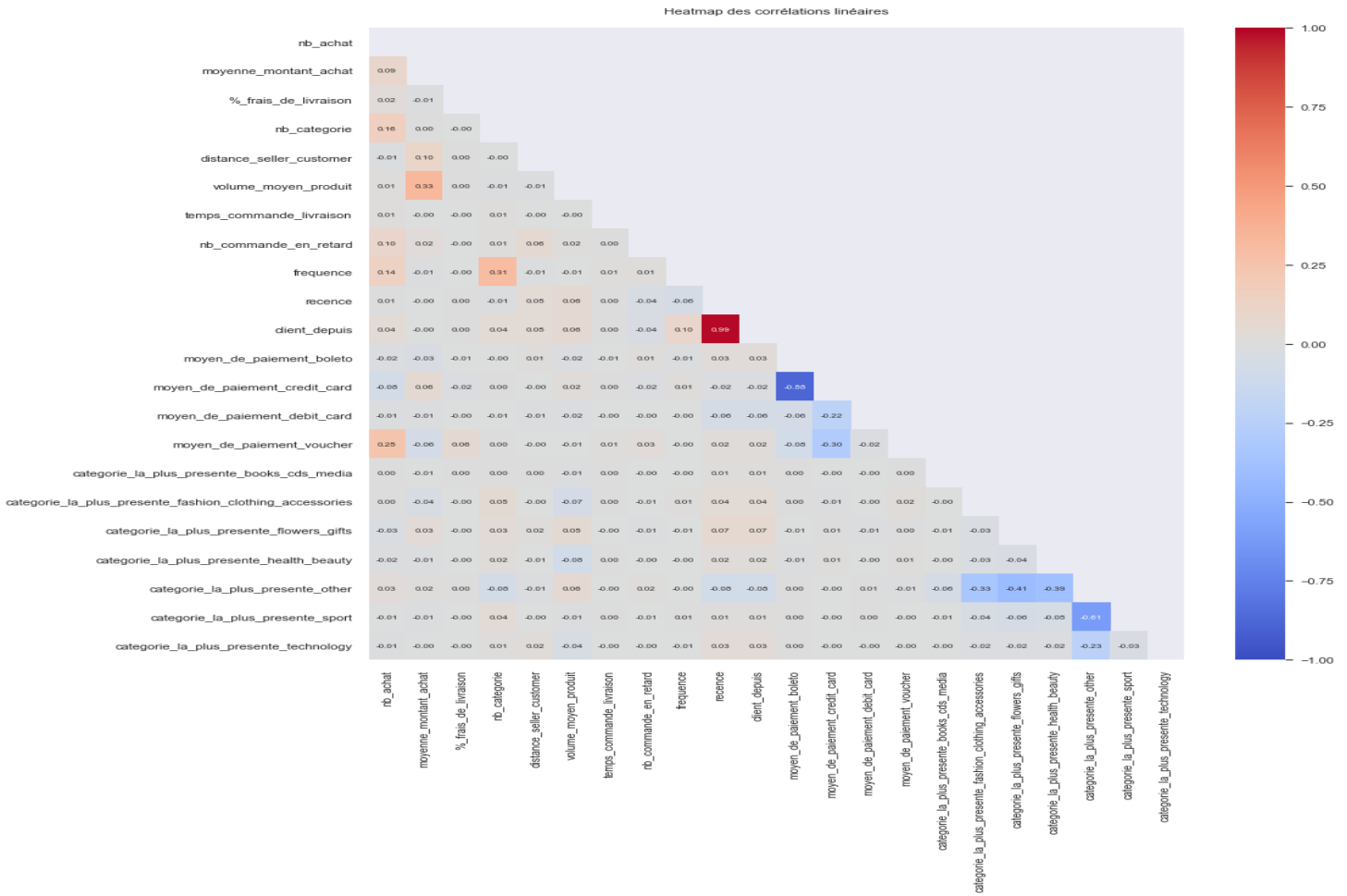
Modelés d’apprentissage non supervisé

Modelés d’apprentissage

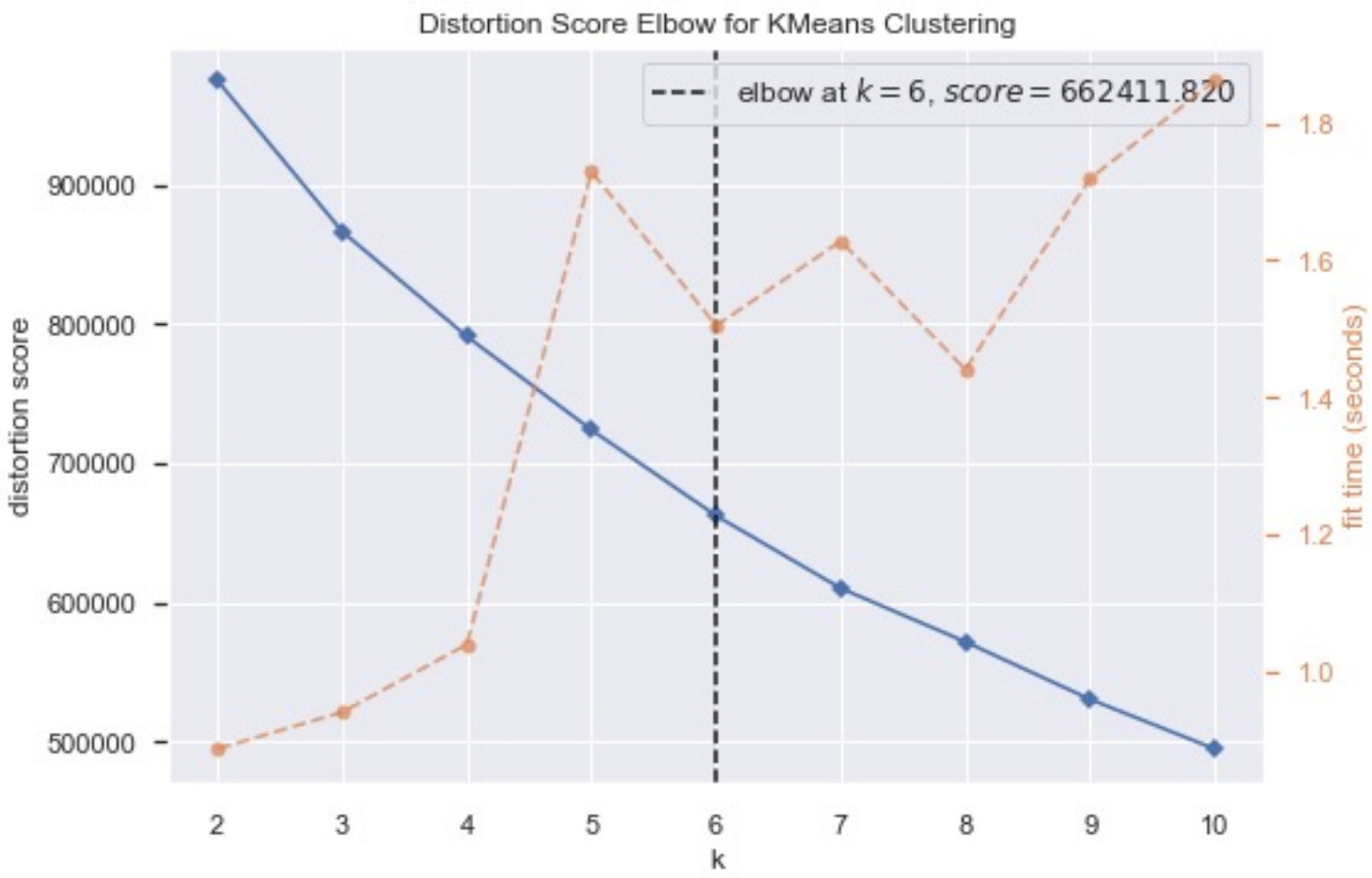
- ✓ Algorithme K-means
- ✓ Agglomerative clustering
- ✓ Clustering par DBSCAN
- ✓ Segmentation RFM

Algorithme K-means

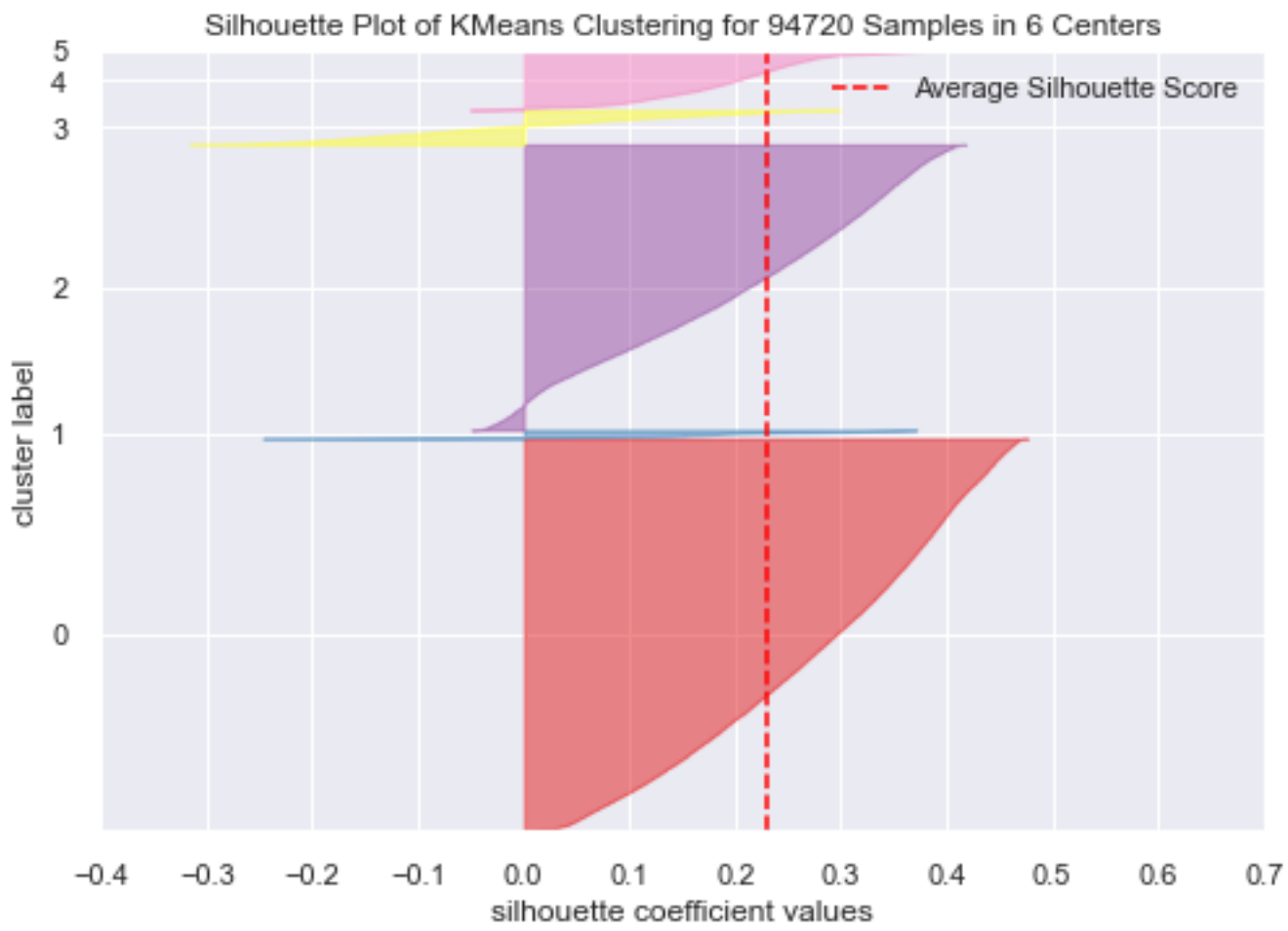
- Scaler les donnés pour preprocessing avec StandardScaler,
- Pour les données numériques et un get dummies pour les données catégorielles.
- Corrélations entre certaines variable



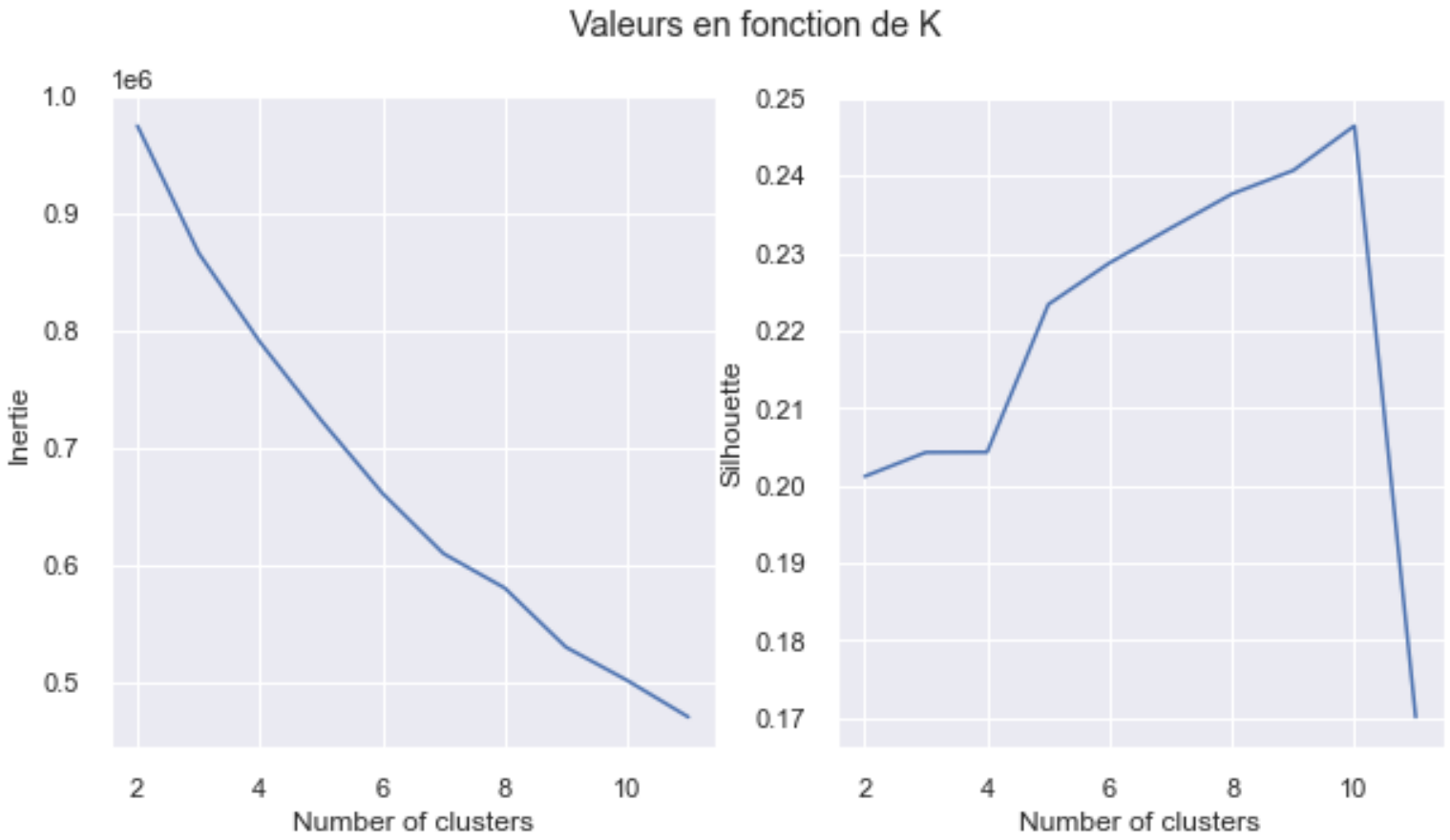
Méthode du coude : Détermination du meilleur K



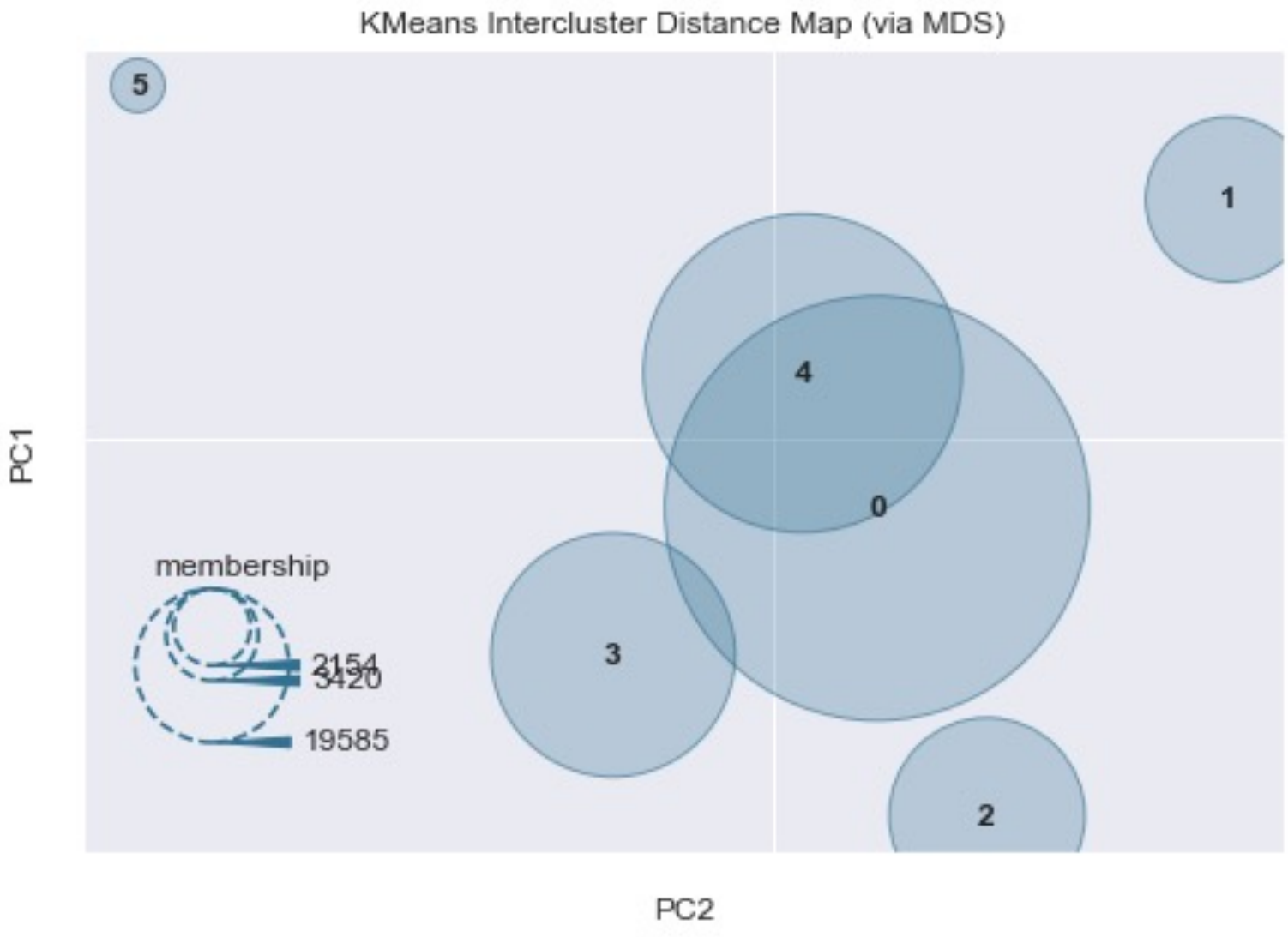
Coefficient de silhouette



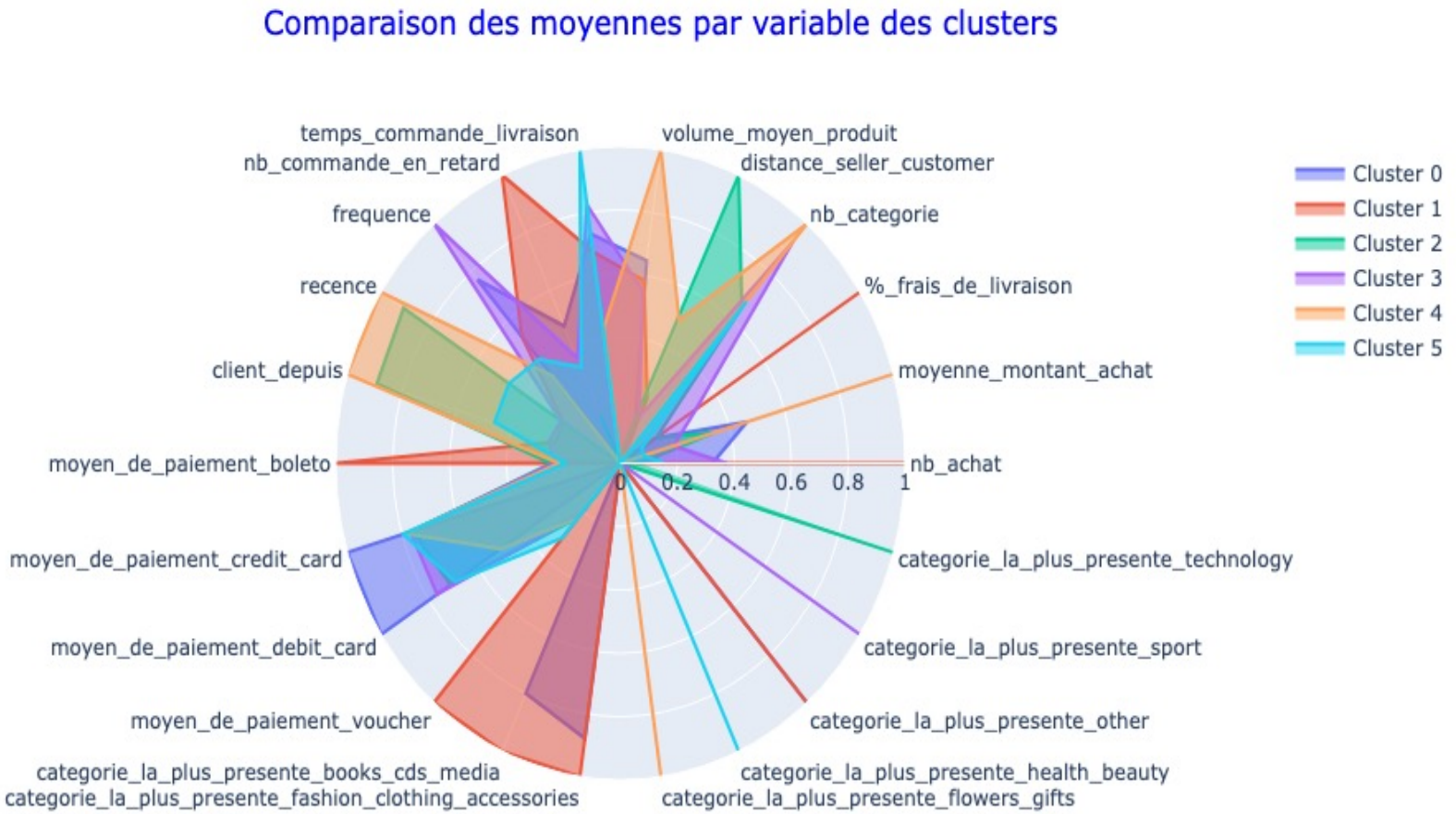
Silhouette score



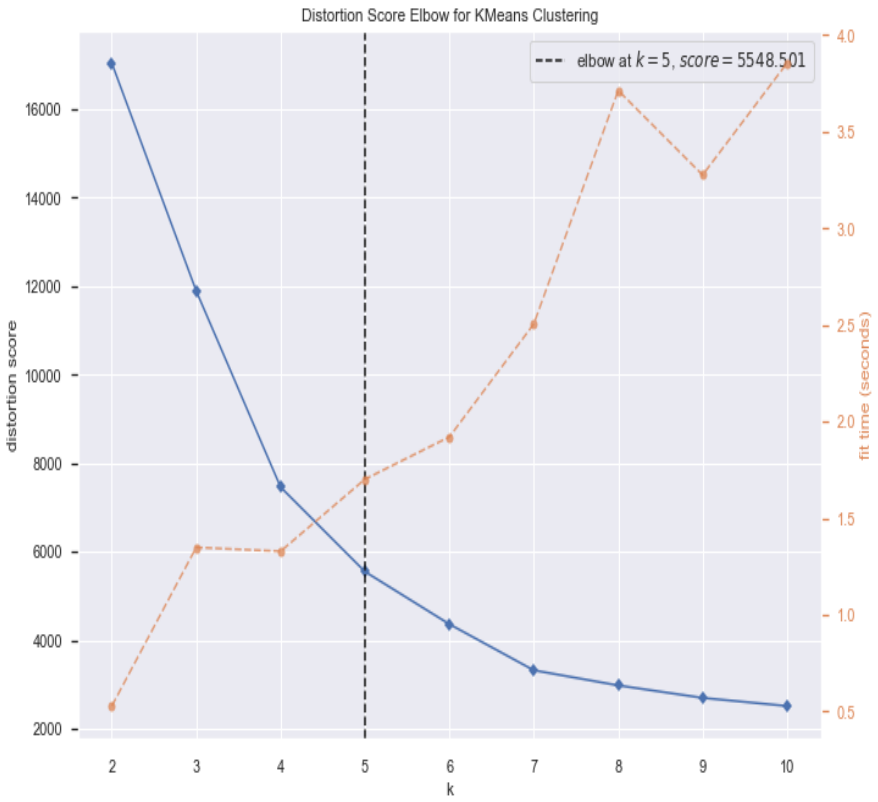
Distances intercluster

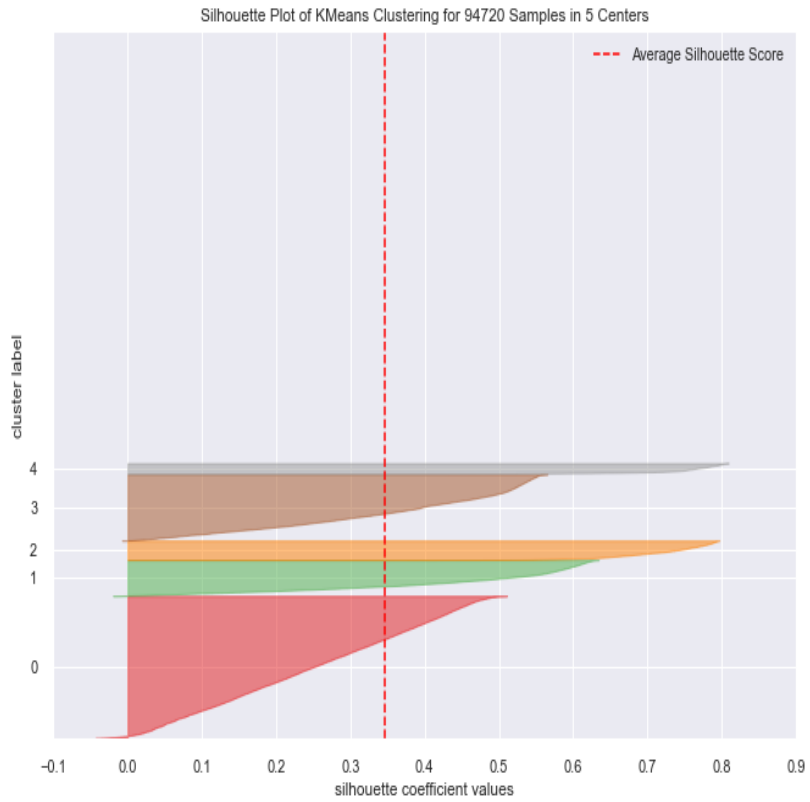


Radar plot

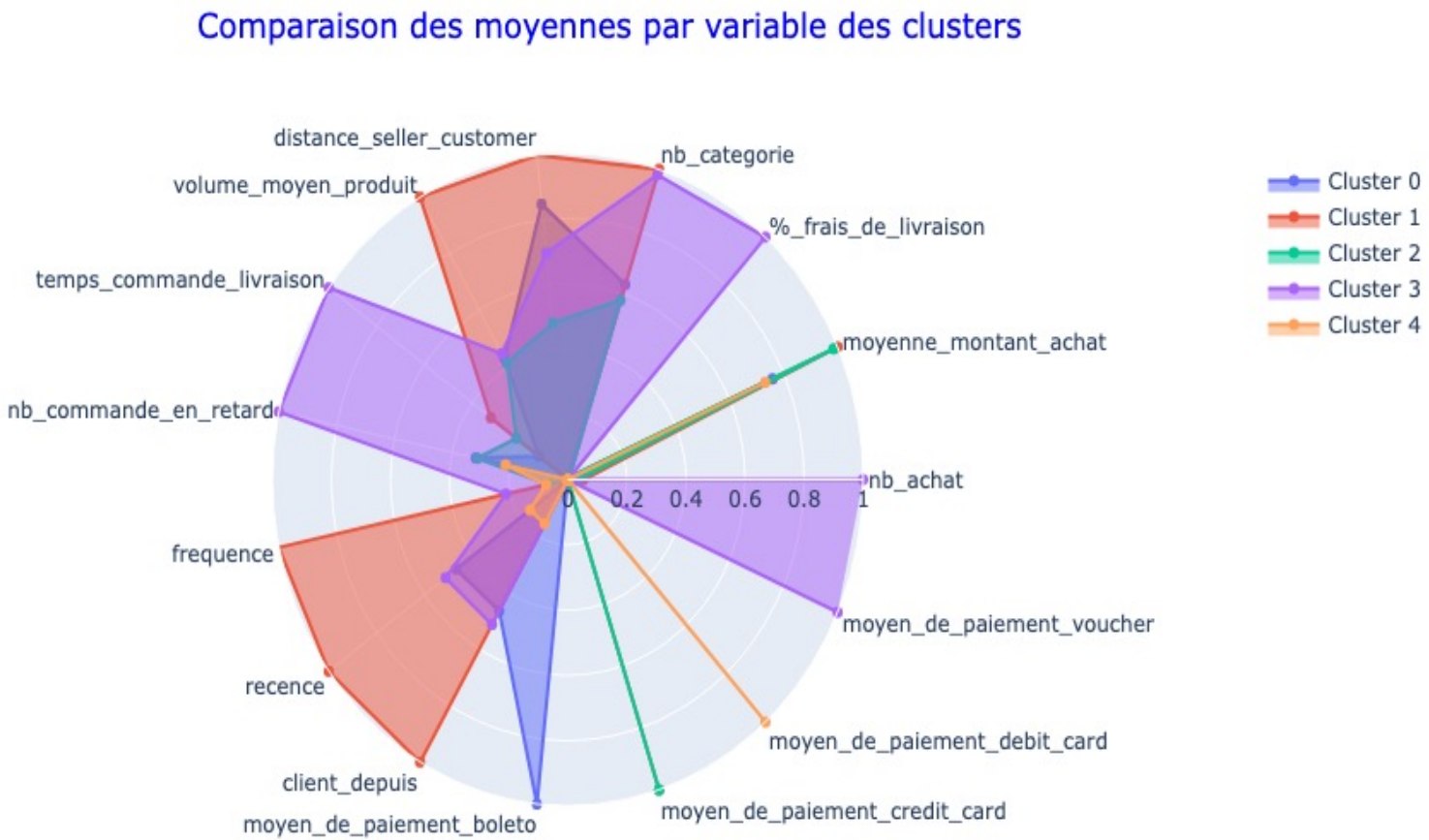

Comparaison des moyennes par variable des clusters
temps_commande_livraison
volume_moyen_produit
nb_commande_en_retard
distance_seller_customer
frequence
nb_categorie
recence
%_frais_de_livraison
client_depuis
moyenne_montant_achat
moyen_de_paiement_boleto
nb_achat
0
0.2
0.4
0.6
0.8
1
moyen_de_paiement_credit_card
categorie_la_plus_presente_technology
moyen_de_paiement_debit_card
categorie_la_plus_presente_sport
moyen_de_paiement_voucher
categorie_la_plus_presente_other
categorie_la_plus_presente_books_cds_media
categorie_la_plus_presente_health_beauty
categorie_la_plus_presente_fashion_clothing_accessories
categorie_la_plus_presente_flowers_gifts
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5

Clustering sans catégorie

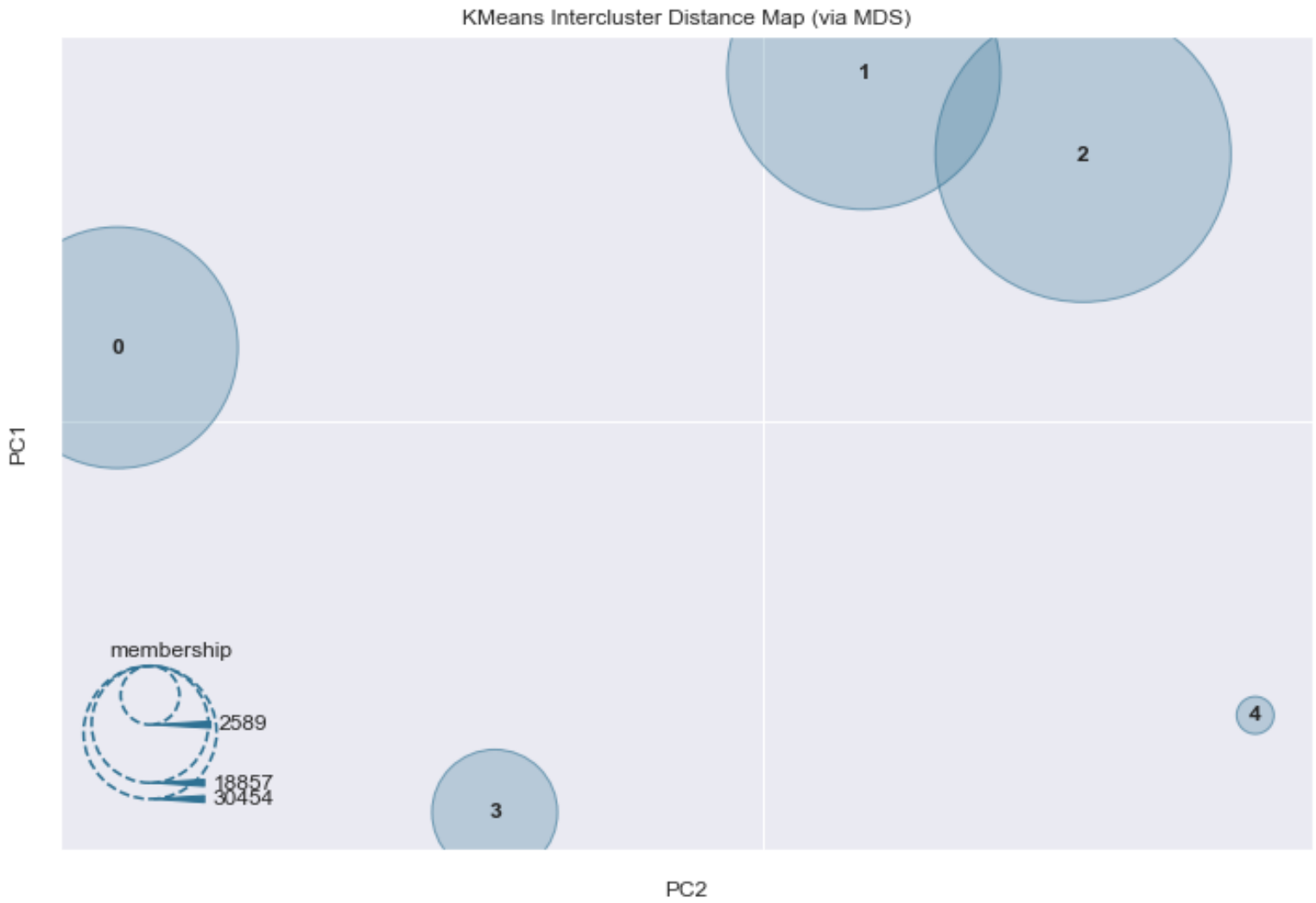




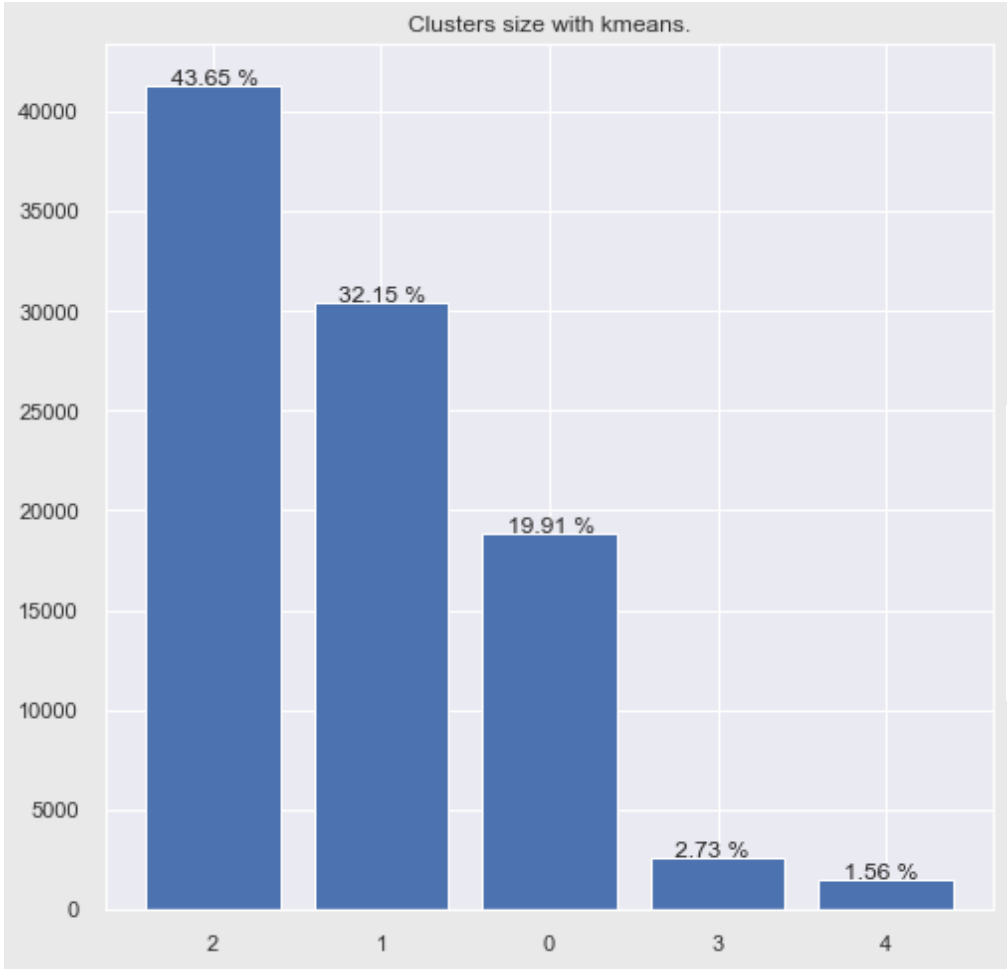
Plot radars sans catégorie



Intercluster Distance



Cluster size

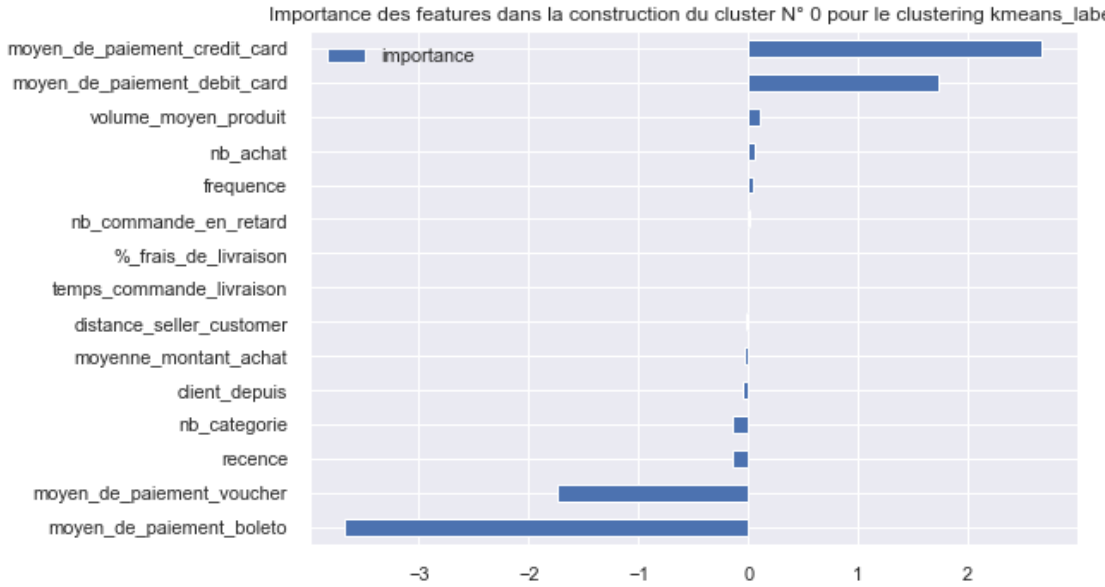


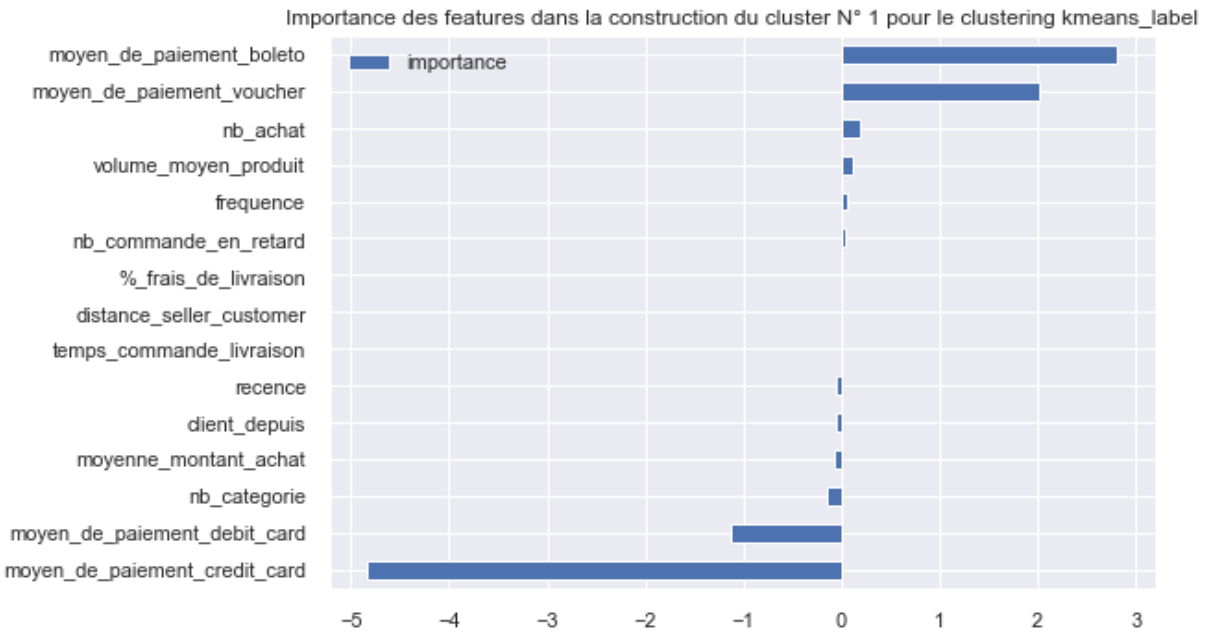

```
from collections import Counter
count = Counter(kmeans_labels_new)

count
```

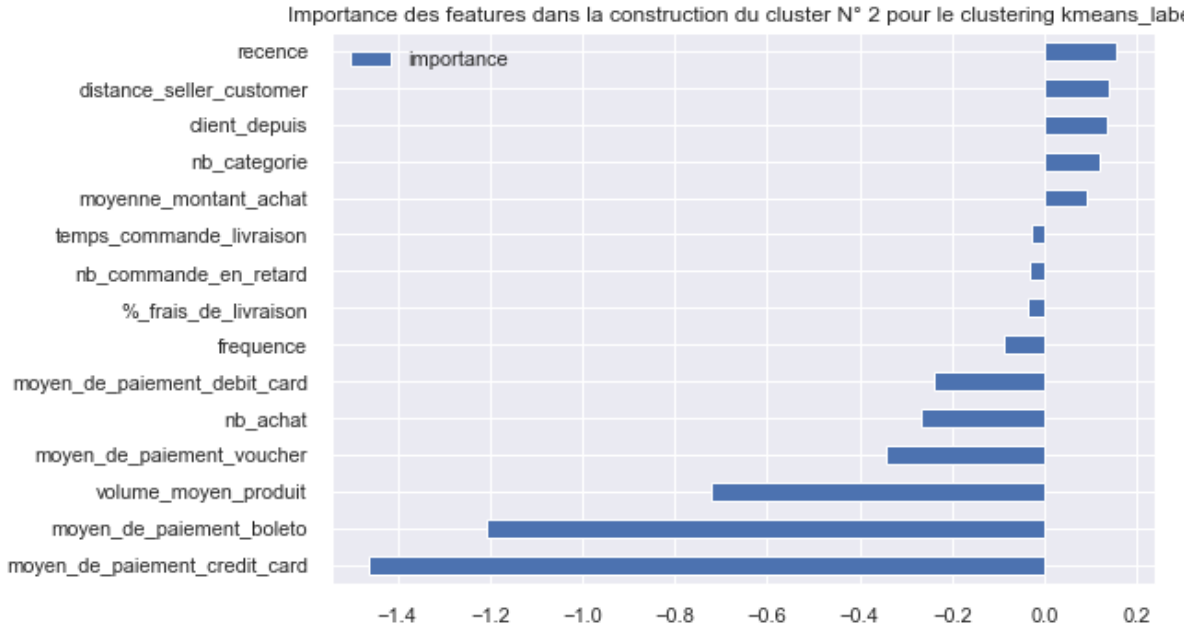
```
Counter({2: 41346, 1: 30454, 0: 18857, 3: 2589, 4: 1474})
```

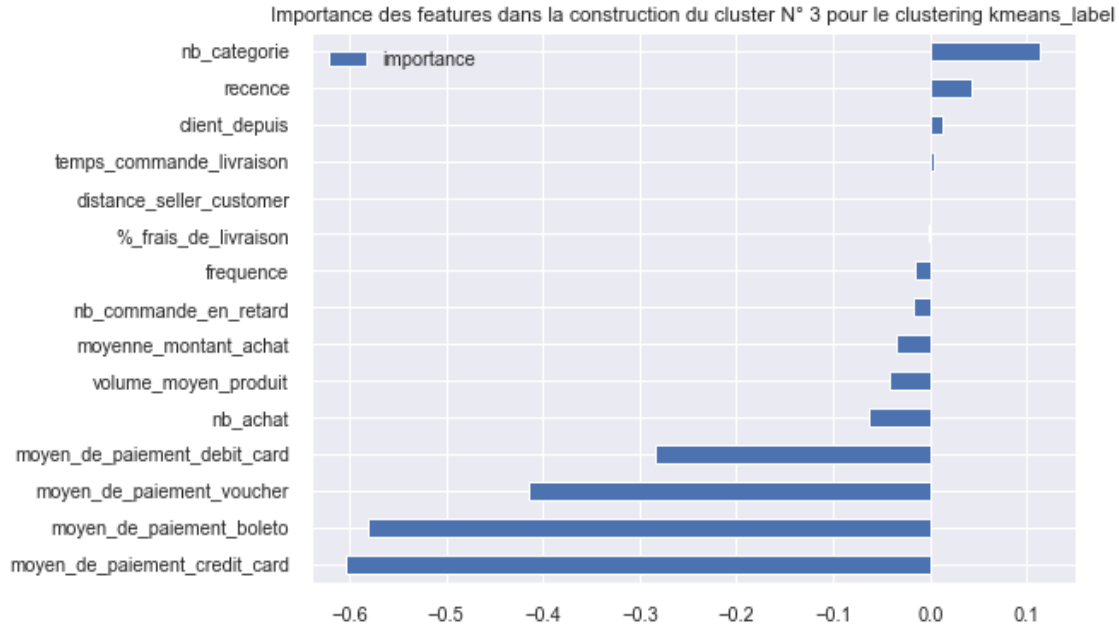
Feature Importance Sans Catégorie



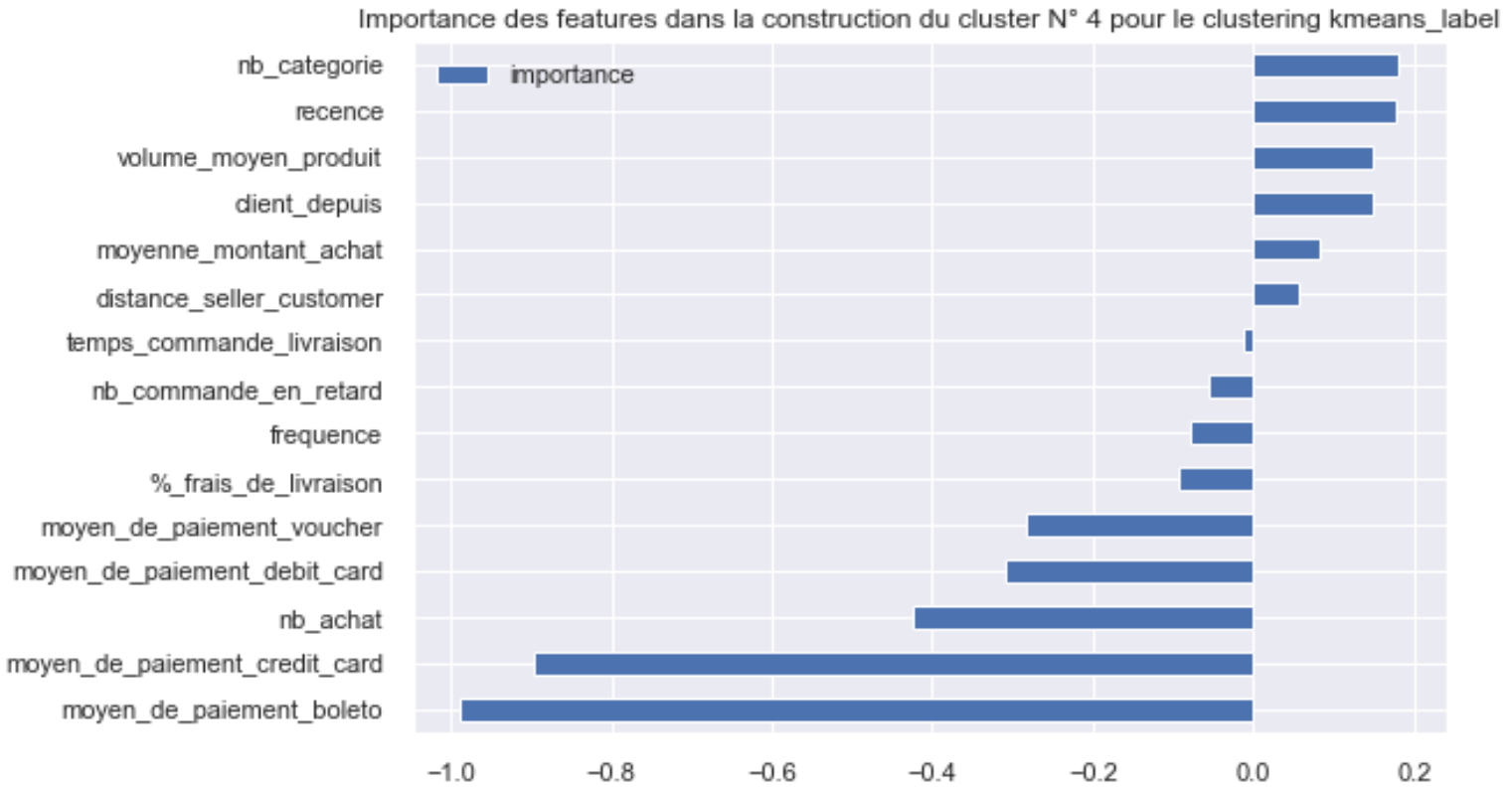


Feature Importance Sans Catégorie

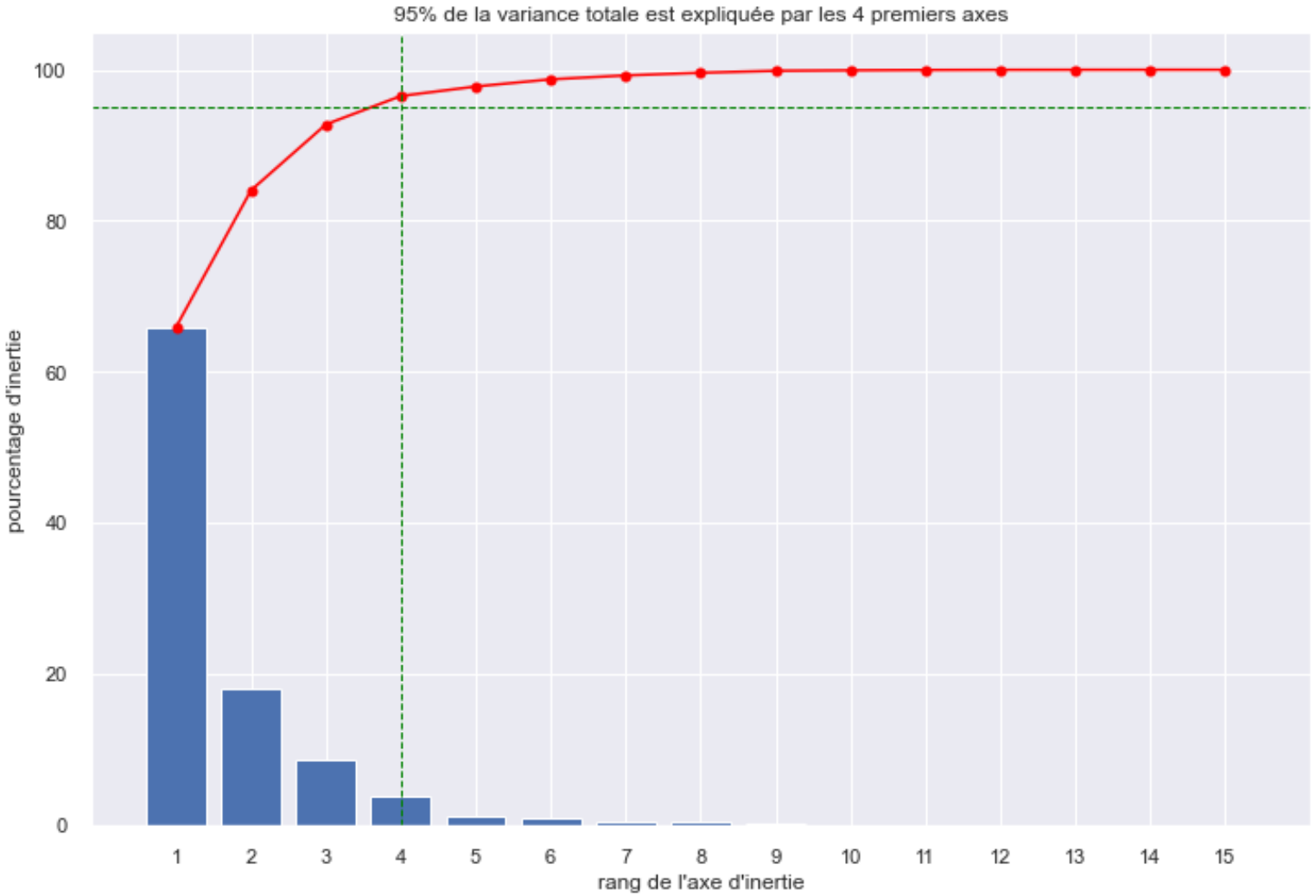




Feature Importance Sans Catégorie

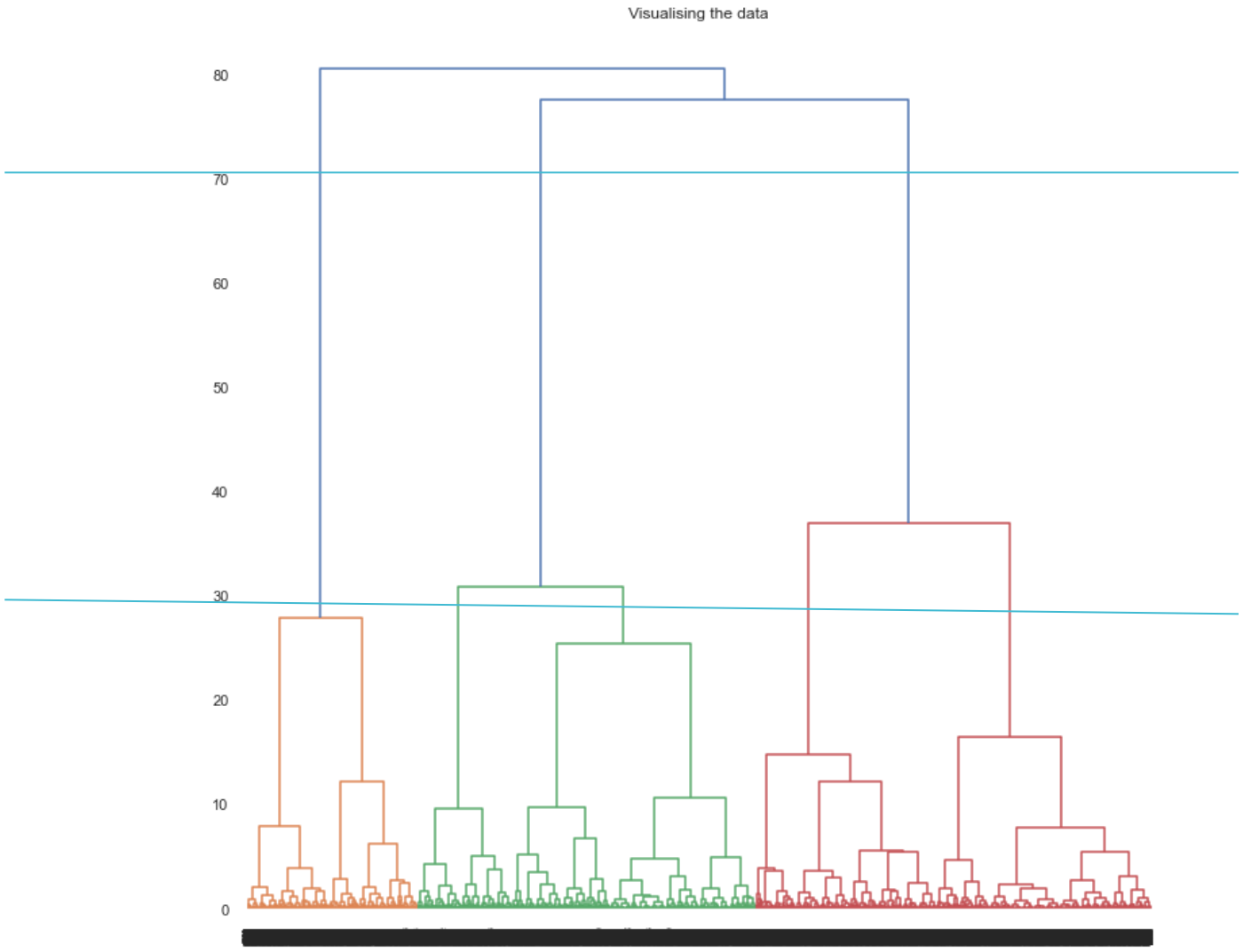


Réduction dimensionnelle - PCA



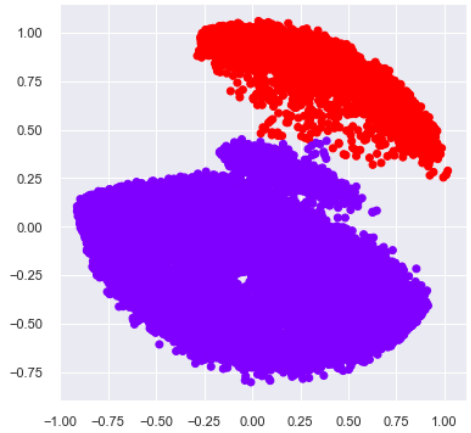
Il faut donc conserver 4 axes principaux pour expliquer la variance à 95%.

Agglomérative clustering

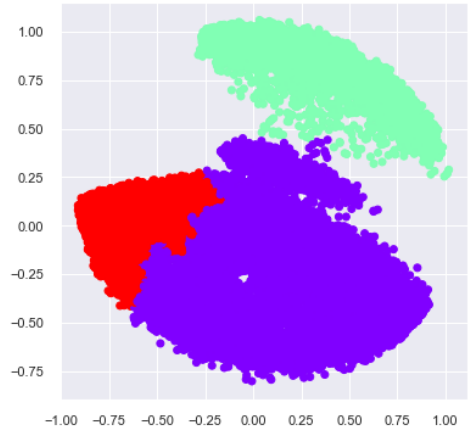


Construire et visualiser

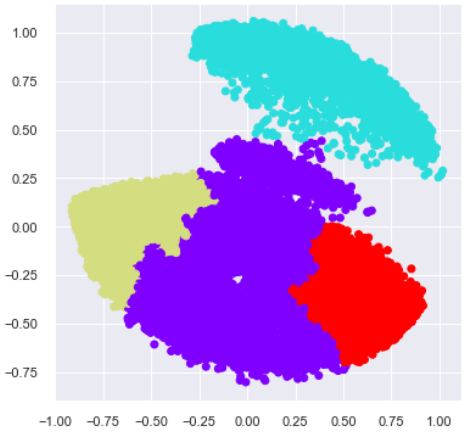
les différents modèles de clustering pour différentes valeurs de k



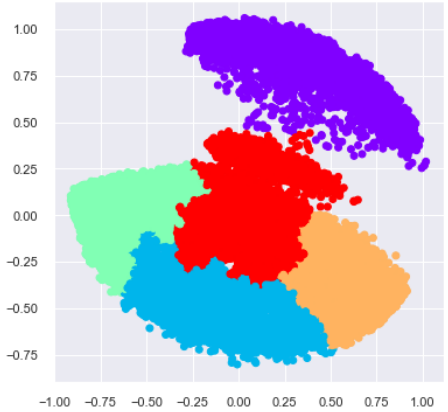
k=2



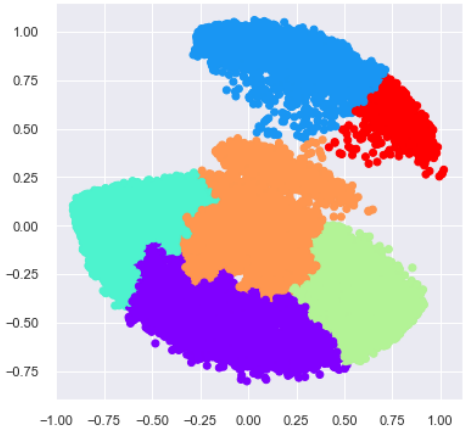
k=3



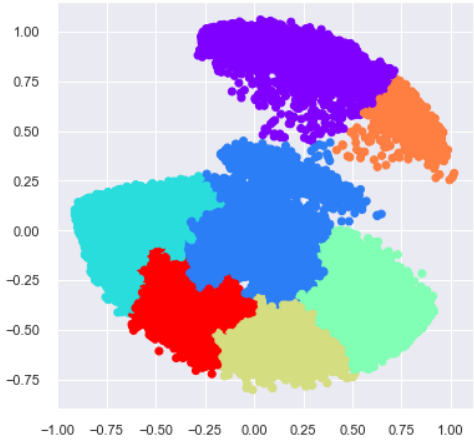
k=4



k=5

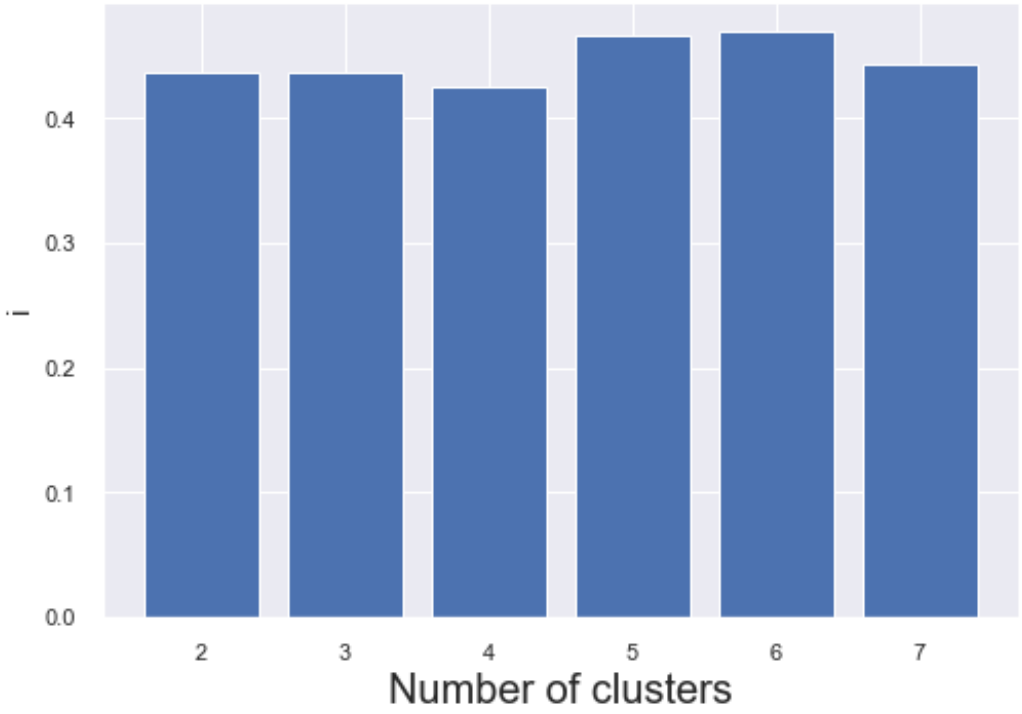


k=6



k=7

Silhouette scores




```
y_test['cluster'].value_counts()
```

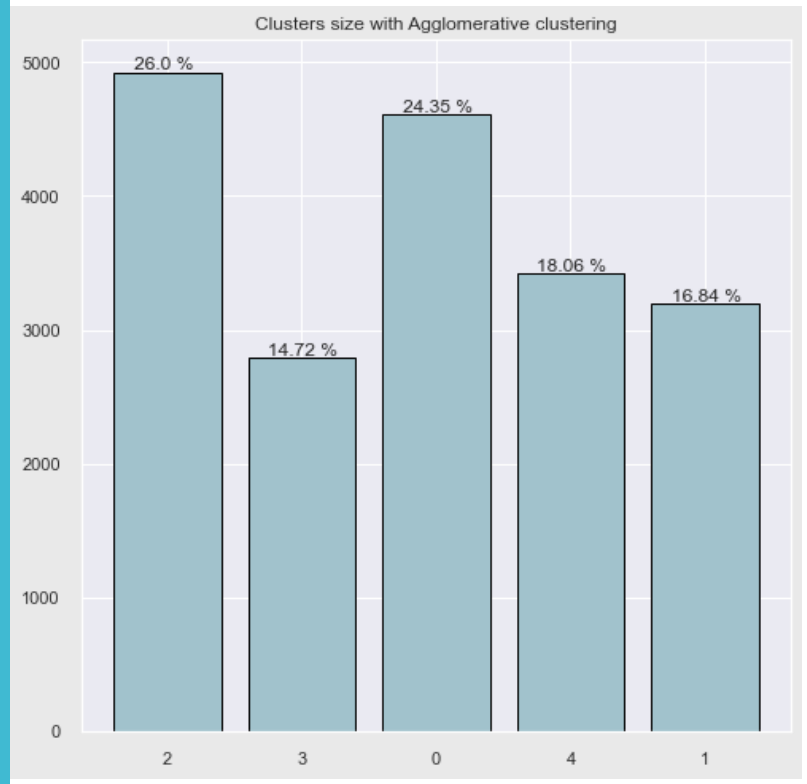
```
2    4926
4    3733
3    3422
0    3192
1    2790
5     881
Name: cluster, dtype: int64
```

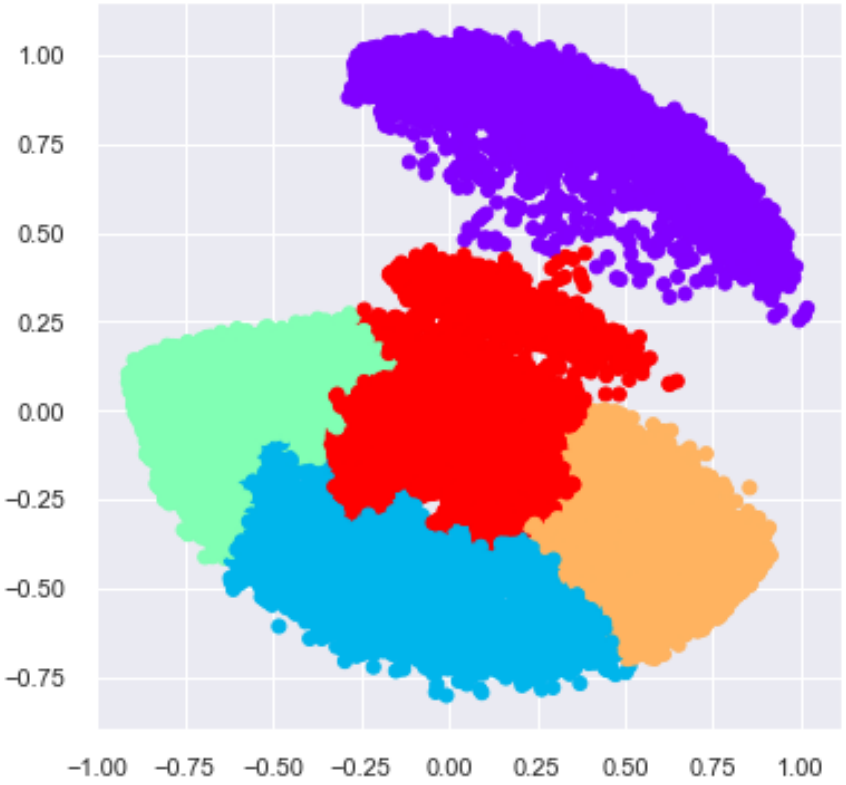
```
y_test['cluster'].value_counts()
```

```
2    4926
0    4614
4    3422
1    3192
3    2790
Name: cluster, dtype: int64
```

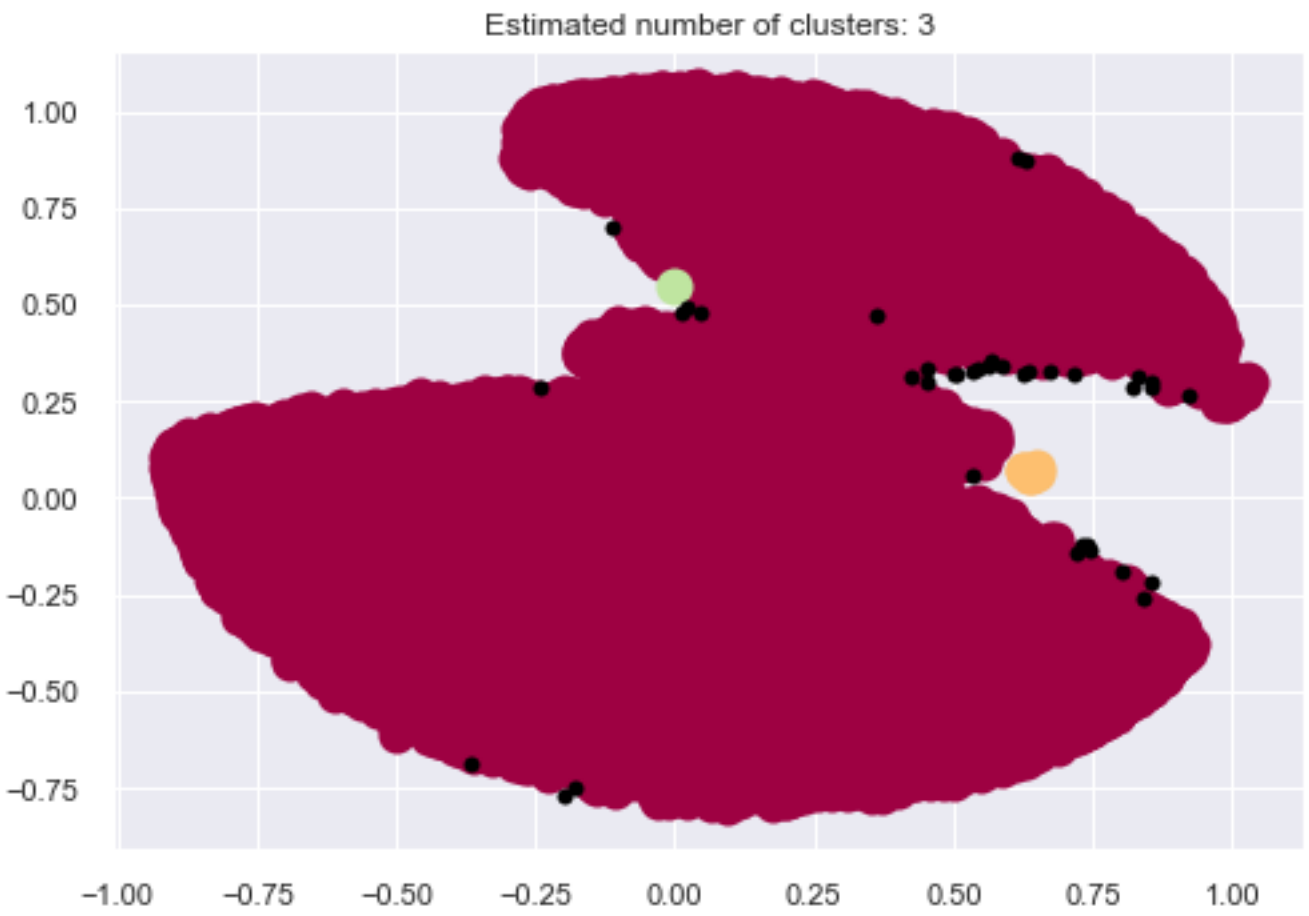
Bien que silhouette score 6 dit bon, On voit que les 5 clusters sont plus homogènement répartis. Donc, le nombre optimal des clusters est bien 5.

Cluster size pour Agglomérative clustering



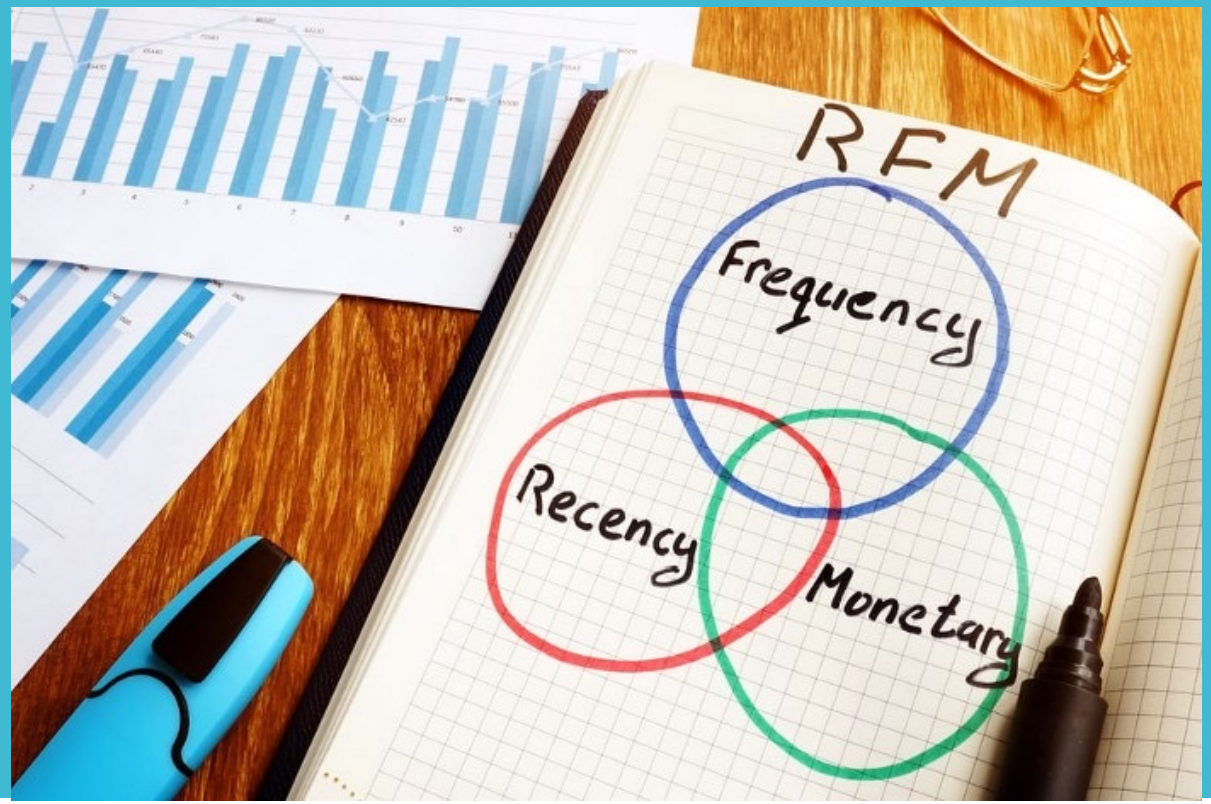


Clustering par DBSCAN



Mauvaise segmentation avec cette méthode

Segmentation de RFM



Segmentation RFM

- La segmentation RFM ou méthode RFM est une méthode de segmentation principalement développée à l'origine pour les actions de marketing direct et qui s'applique désormais également aux acteurs du e-commerce et du commerce traditionnel.

La segmentation RFM prend en compte:

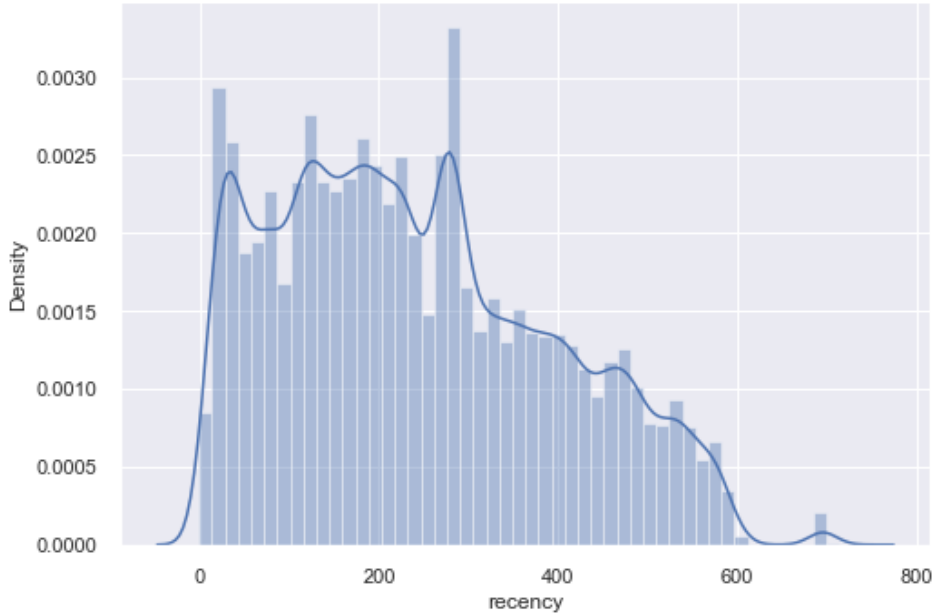
- La Récence (date de la dernière commande)
- La Fréquence des commandes
- Le Montant (de la dernière commande ou sur une période donnée) pour établir des segments de clients homogènes.

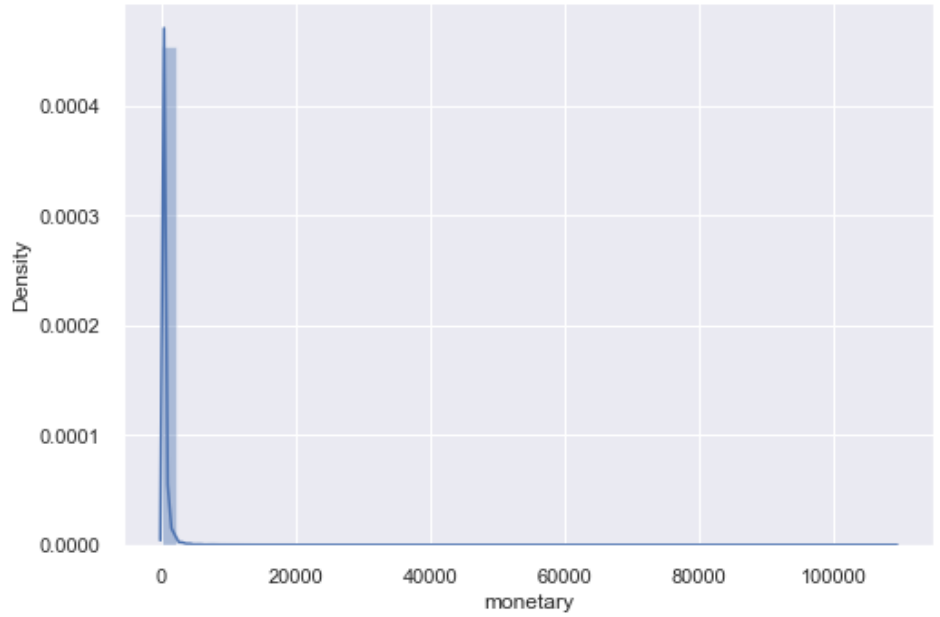
Segmentation RFM

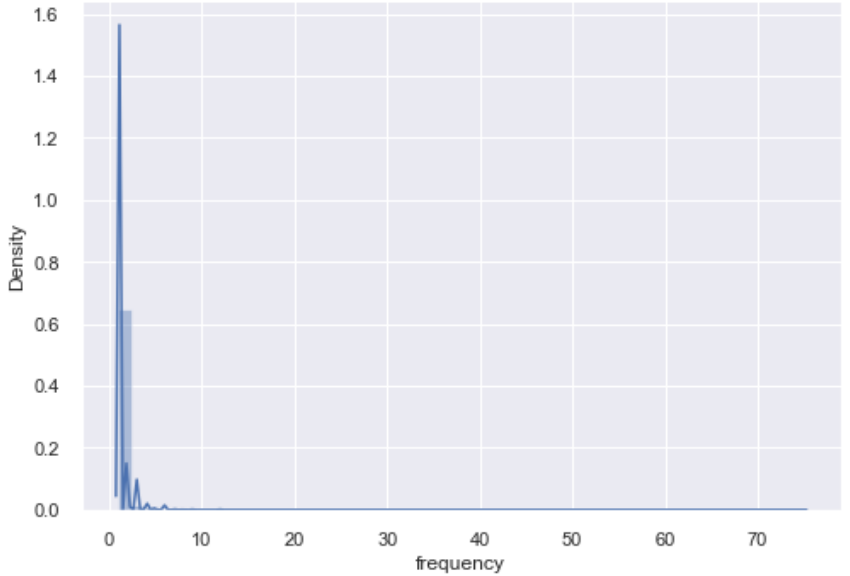
Par exemple:

- Quels sont vos meilleurs clients ?
- Quels clients sont sur le point de baratter ?
- Qui a le potentiel d'être converti en clients plus rentables
- Quels clients sont perdus/inactifs ?
- Quels clients est-il essentiel de fidéliser ?
- Qui sont vos fidèles clients ?
- Quel groupe de clients est le plus susceptible de répondre à votre campagne actuelle ?

Segmentation RFM





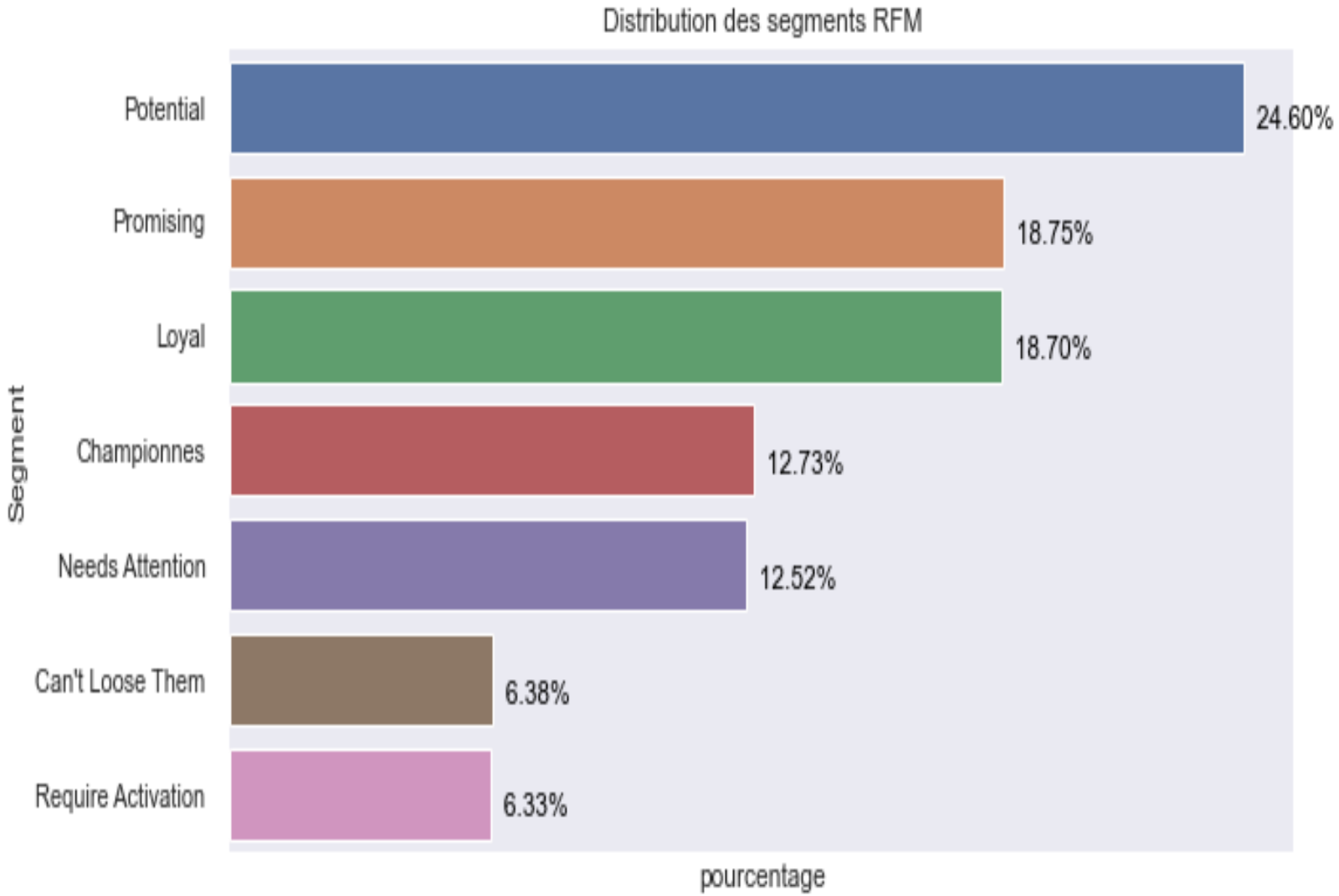


Calcul d'un RFM score

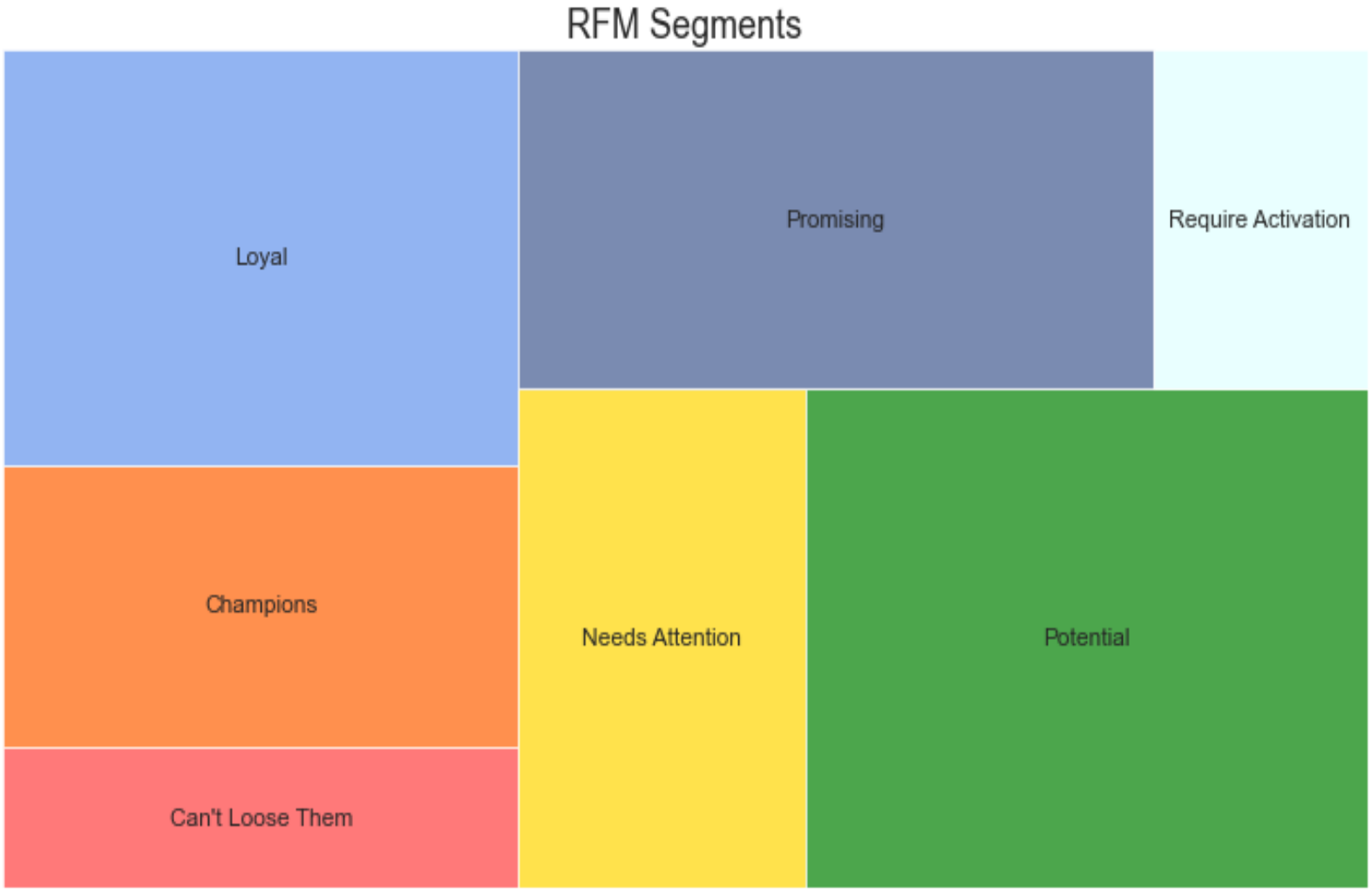
RFM score	RFM level
>= 9	Can't loose them
>= 8 & < 9	Champions
>= 7 & < 8	Loyal
>= 6 & < 7	Potential
>= 5 & < 6	Promising
>= 4 & < 5	Needs Attention
Others	Require Activation

Segment

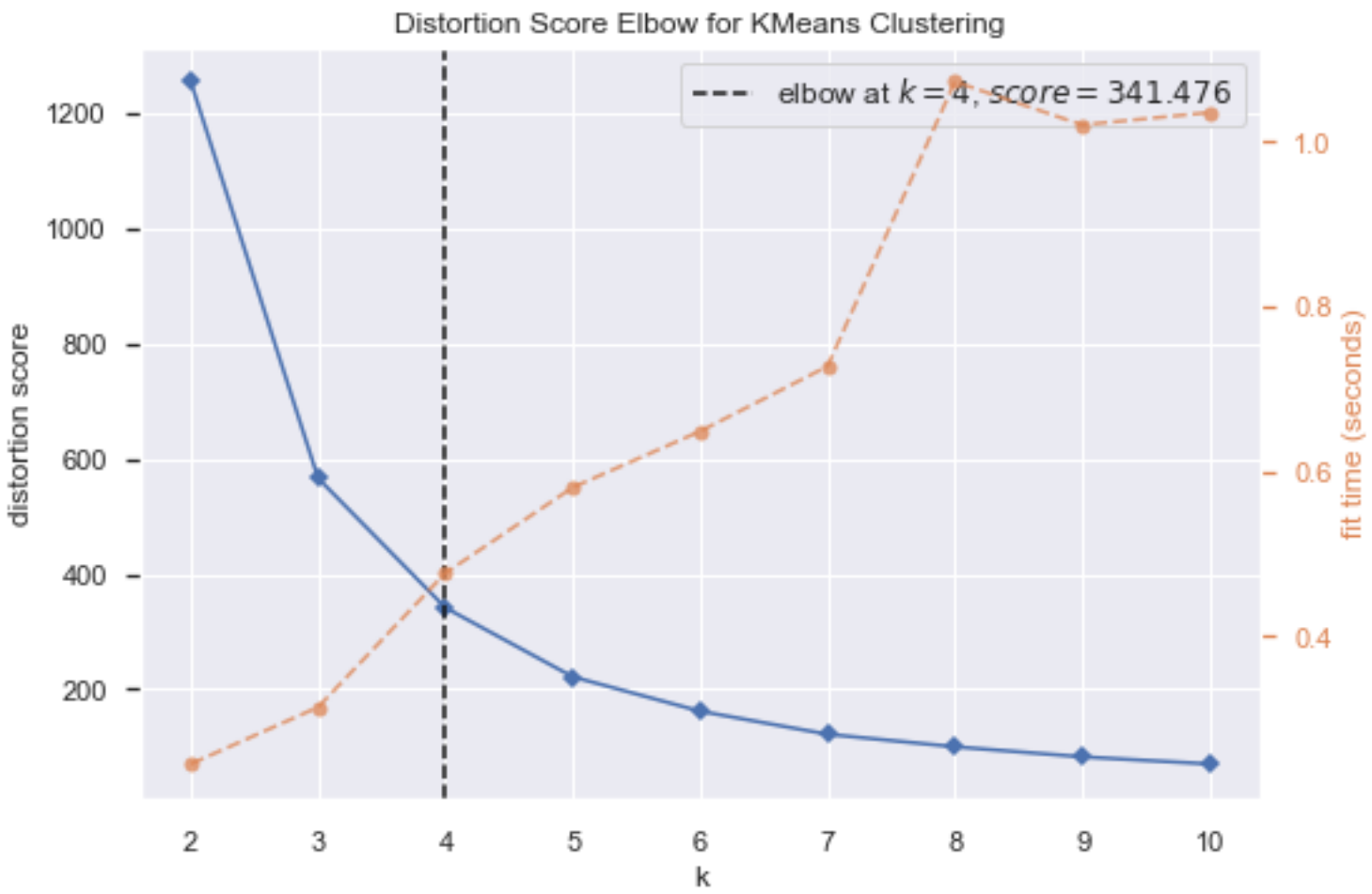
RFM level pour chaque client



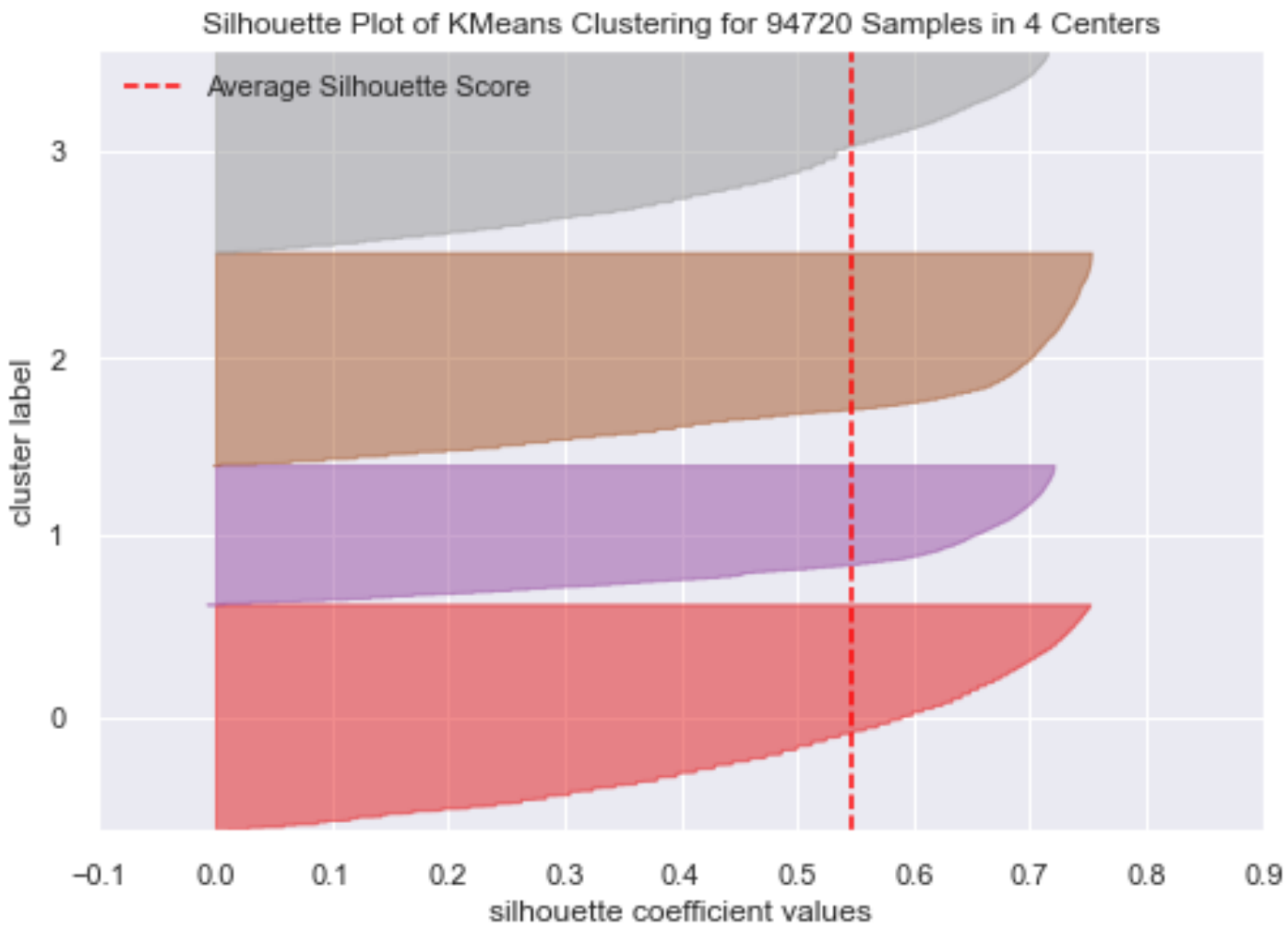
Segmentation



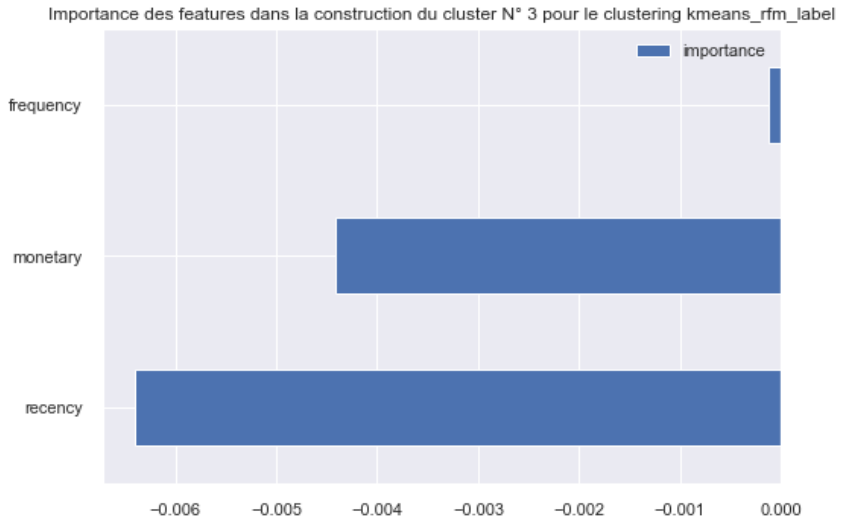
KMeans avec RFM

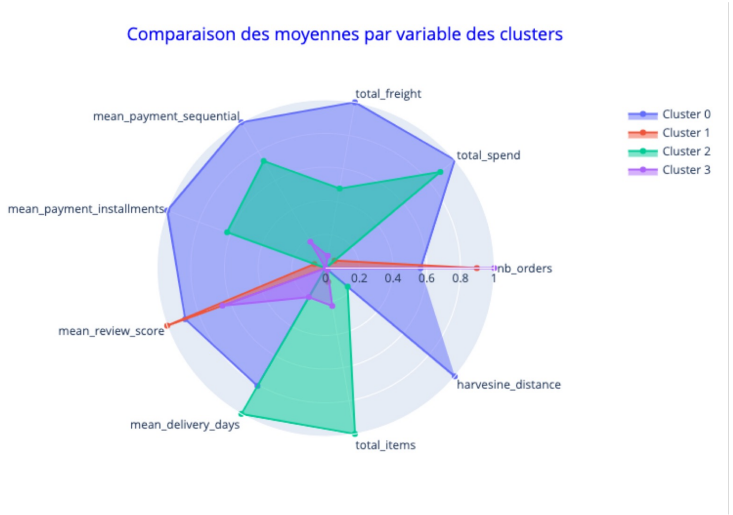


KMeans avec RFM



Importance pour RFM et Choix de modélisation final





Profils des clients

Groupe 1:

- ✓ Clients géographiquement éloignés avec des délais de livraison longs
- ✓ Passer des commandes pour des volumes élevés et des frais d'expédition élevés. Les avis de ces clients sont bons.

Groupe 2:

- ✓ Des clients géographiquement proches avec un grand nombre de commandes mais peu de produits et de faibles quantités.
- ✓ Les critiques sont très bonnes.

Groupe 3:

- ✓ Clients relativement proches, ils commandent beaucoup d'articles pour des montants élevés.
- ✓ Des clients insatisfaits..

Groupe 4:

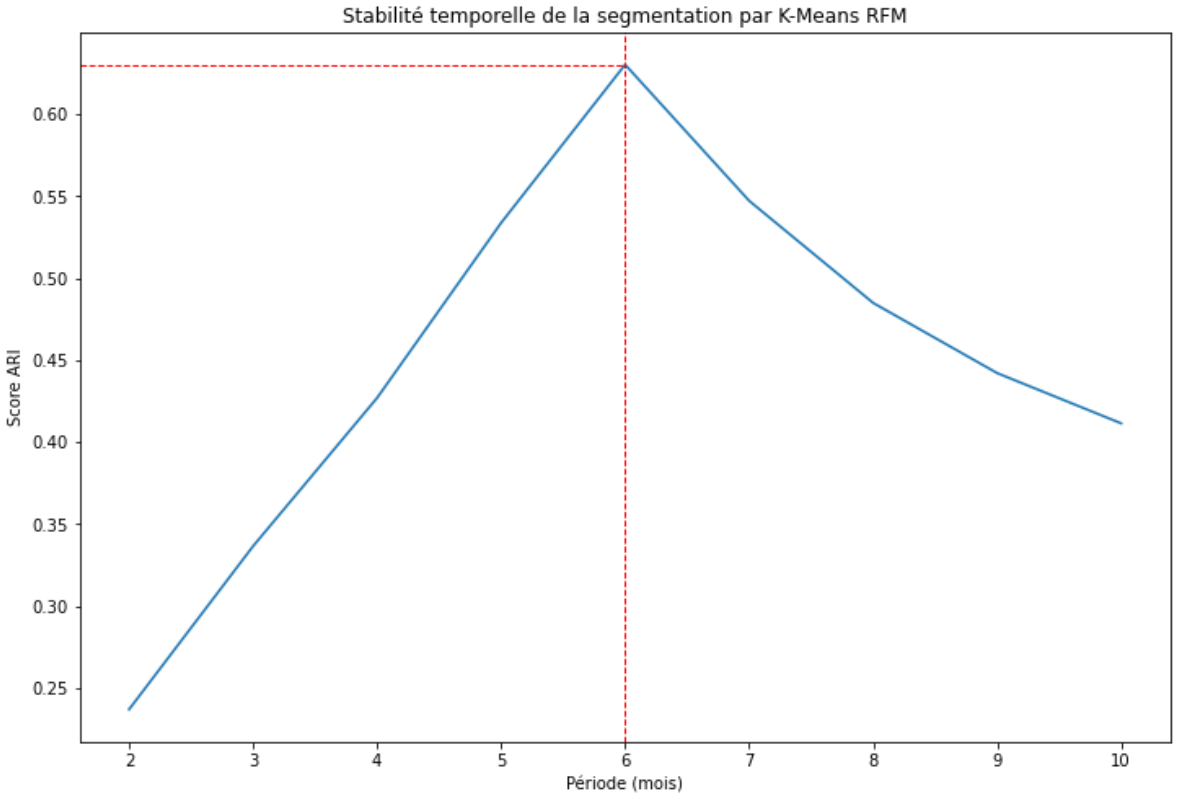
- ✓ Clients qui dépensent peu avec peu d'articles.
- ✓ Leurs avis sont moyennement bons

Maintenance

Stabilité Temporelle de la segmentation

- Dans le but d'établir un contrat de maintenance de l'algorithme de segmentation client, nous devons tester sa stabilité dans le temps.
- Pour déterminer le moment où les clients changent de cluster, nous allons itérer le K-Means sur toute la période des commandes (23 mois) avec des deltas de 2 mois et calculer le score ARI.

Stabilité Temporelle de la segmentation



Prévoir la maintenance du programme de segmentation tous les 6 mois

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

Conclusion

- Segmentation RFM avec un K-means à renouveler tous les 6 mois.
- Utilise Random-state paramètre est très importante

Merci!