

SEGMENTEZ DES CLIENTS D'UN SITE E-COMMERCE



olist
empowering commerce

1

Parcours Data sciences | ABBOUD Marwa | 26 Janvier 2020

Encadrant : Bertrand Beaufiles

Evaluateur :

SOMMAIRE

- PRESENTATION DE LA PROBLÉMATIQUE
- PRESENTATION DES DONNEES:
 - Analyse exploratoire
- FEATURES ENGINEERING:
 - Preprocessing
- SEGMENTATION:
 - Tests des algorithmes de classification non supervisée
 - Sélection du modèle final
 - Analyse des caractéristiques des différents clusters identifiés
 - Observations du comportement des clusters au cours du temps
- CONCLUSION

Présentation de la problématique de Olist

- Olist, solution de vente sur les marketplaces en ligne souhaite comprendre les différents types de clients.

Objectifs de l'étude

- Proposer une solution de segmentation clients.
- Evaluer la fréquence à laquelle la segmentation doit être mise à jour, afin d'établir un contrat de maintenance.

Source de données

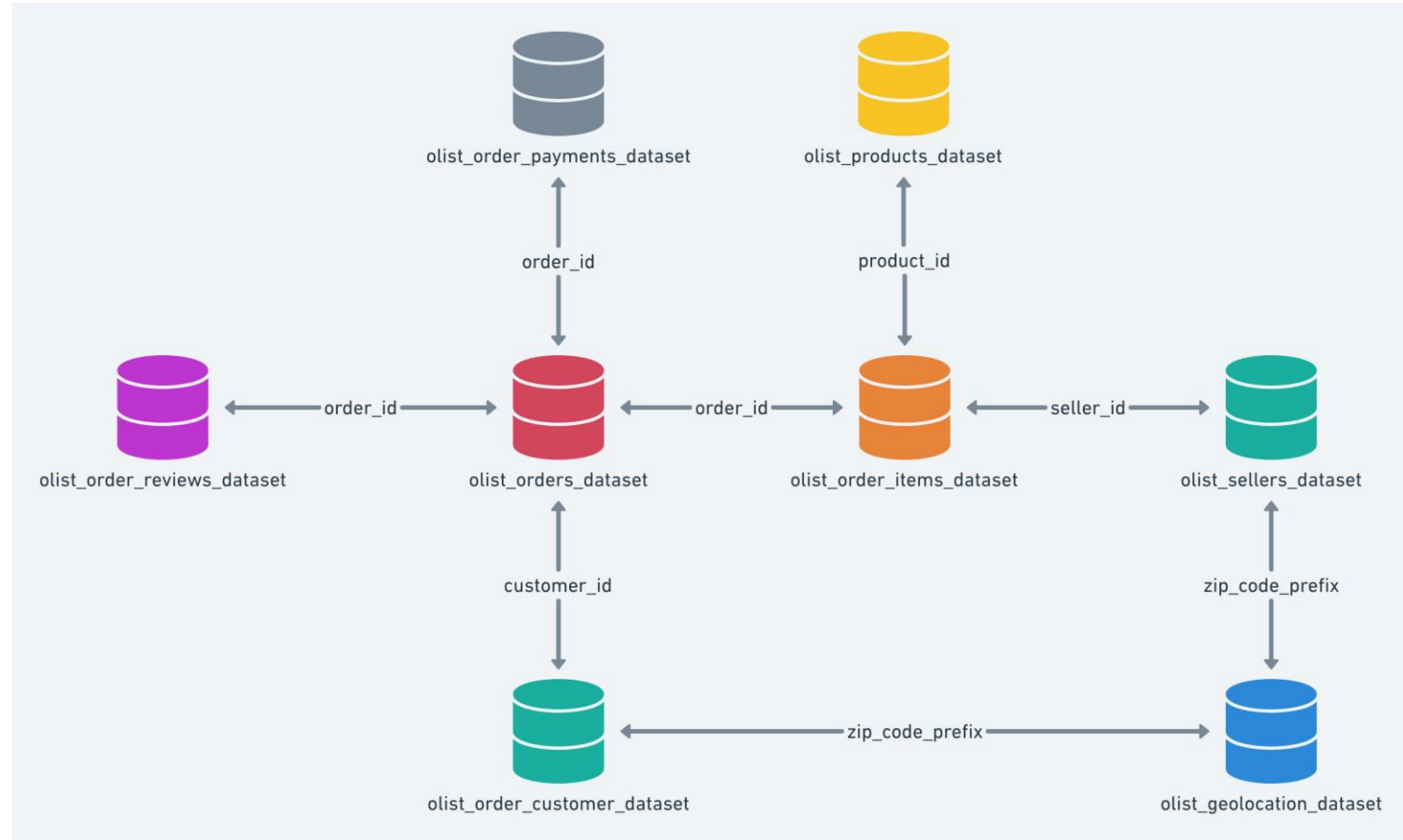
<https://www.kaggle.com/olistbr/brazilian-ecommerce>

PRESENTATION DES DONNEES

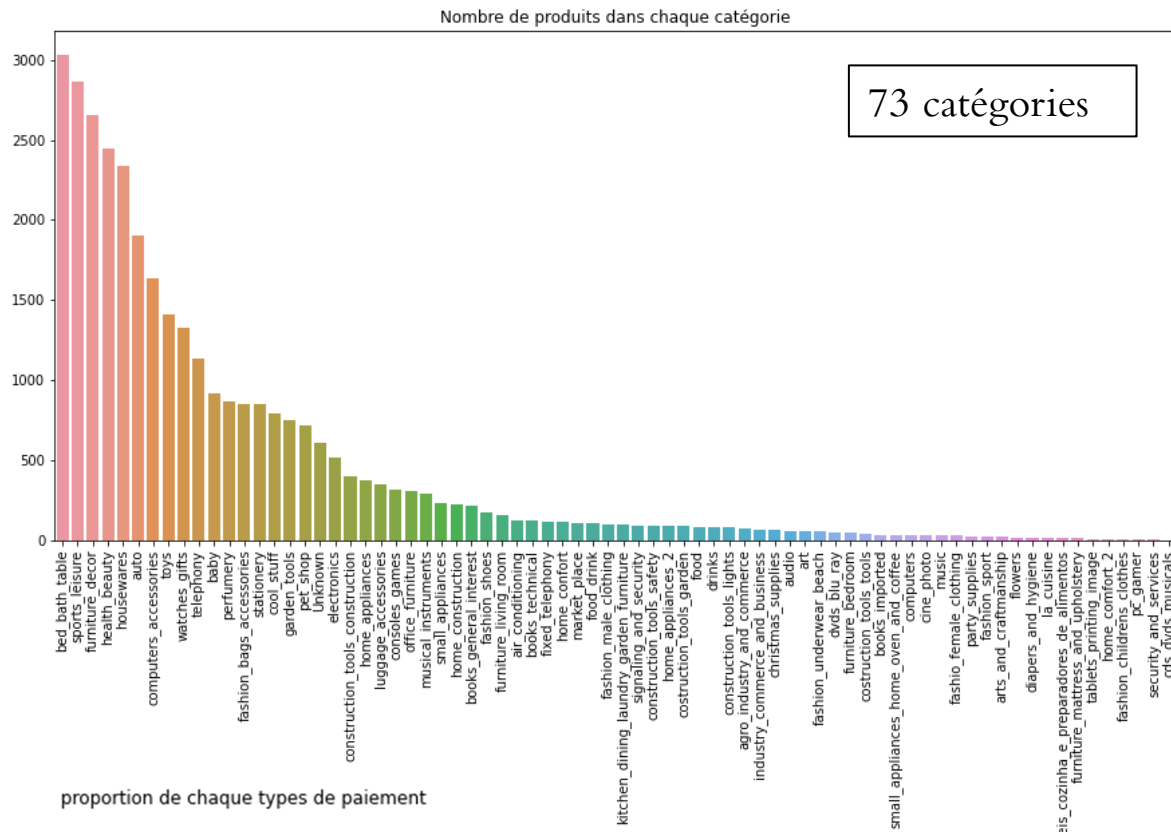
- Base de données composée de 8 Datasets qui sont reliés par des variables clés

→ Comporte 96455 commandes renseignés de 2016 à 2018

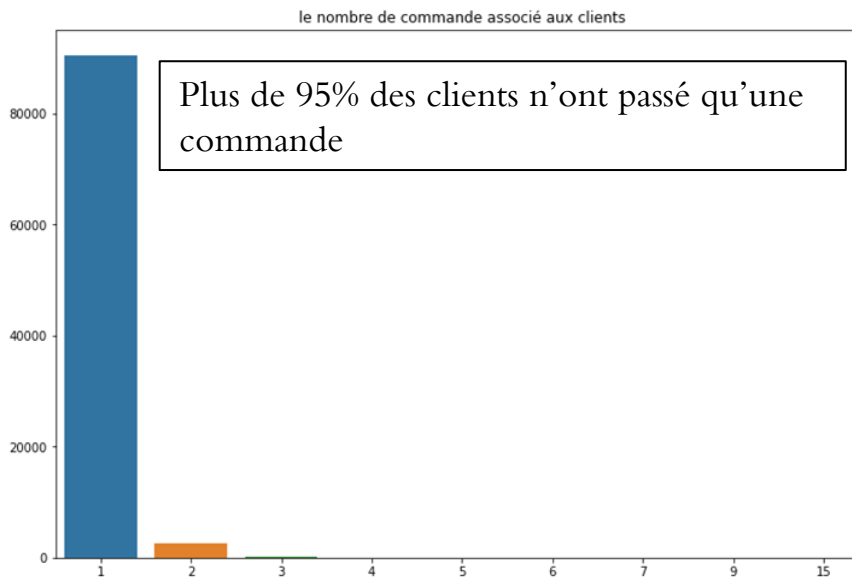
- Clients
- Commandes
- Géolocalisation
- Achats
- Produits
- Vendeurs
- Paiements
- Evaluation des produits



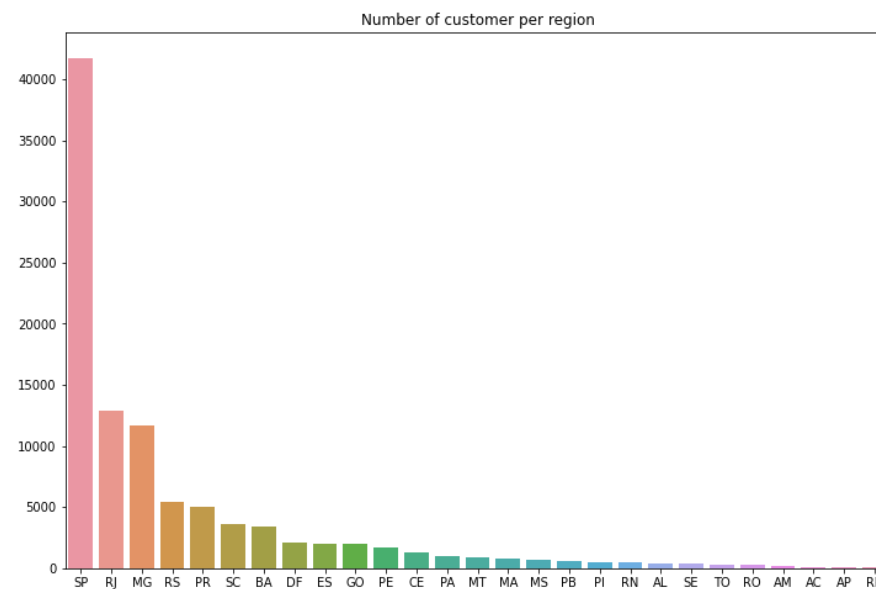
- Dataset supplémentaire pour la traduction des catégories de produits du : Brésilien → Anglais

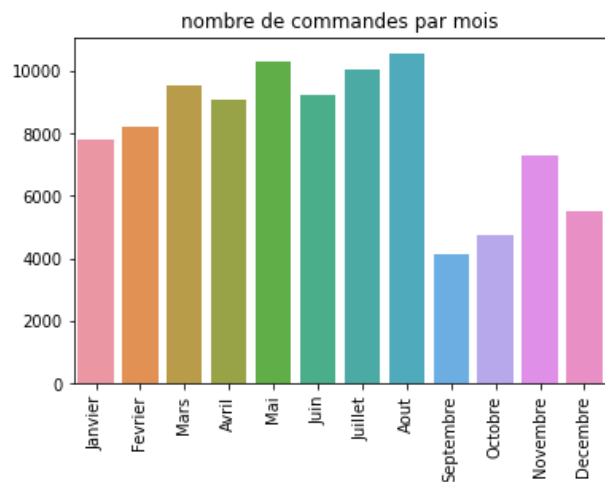
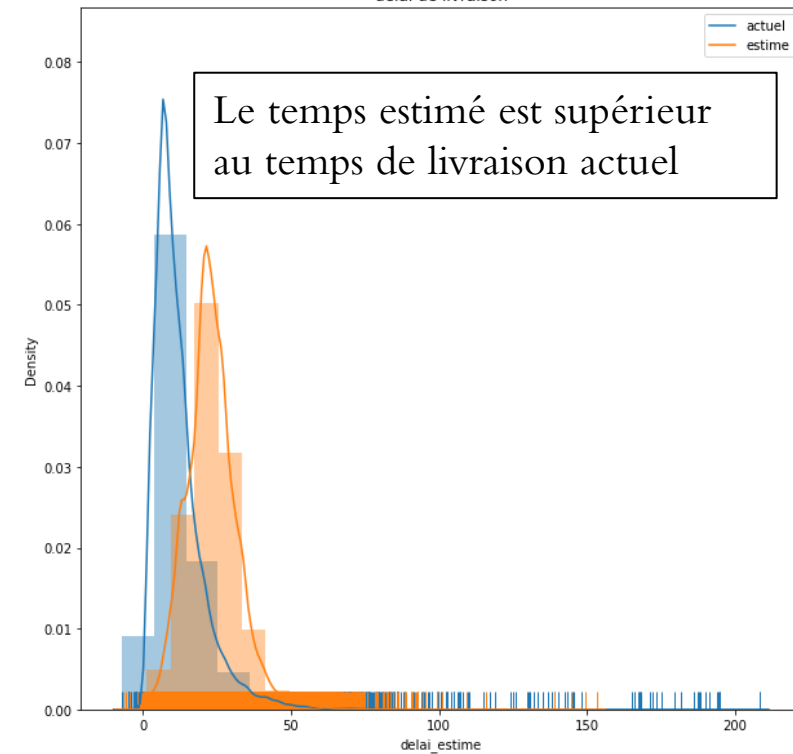


le type de paiement le plus utilise est par carte de crédit

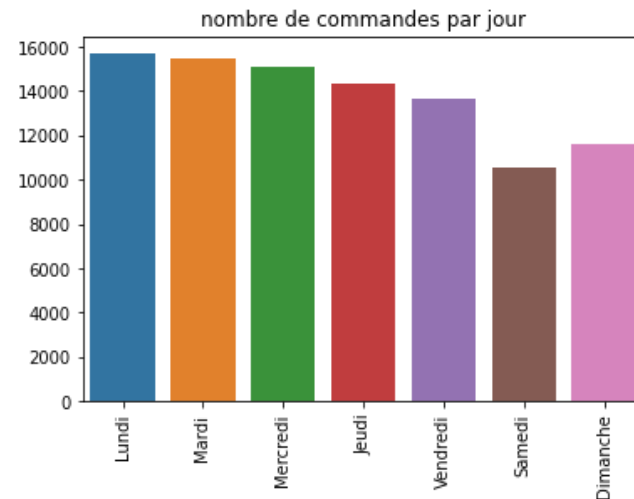


Il y a une prédominance d'acheteurs dans l'état de São Paulo



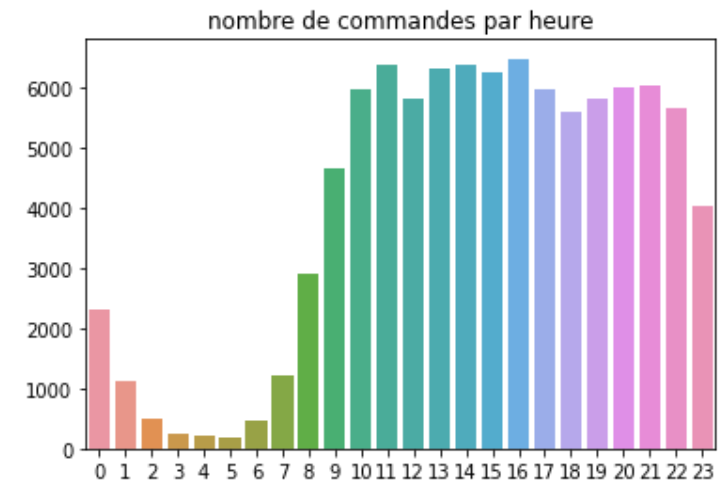
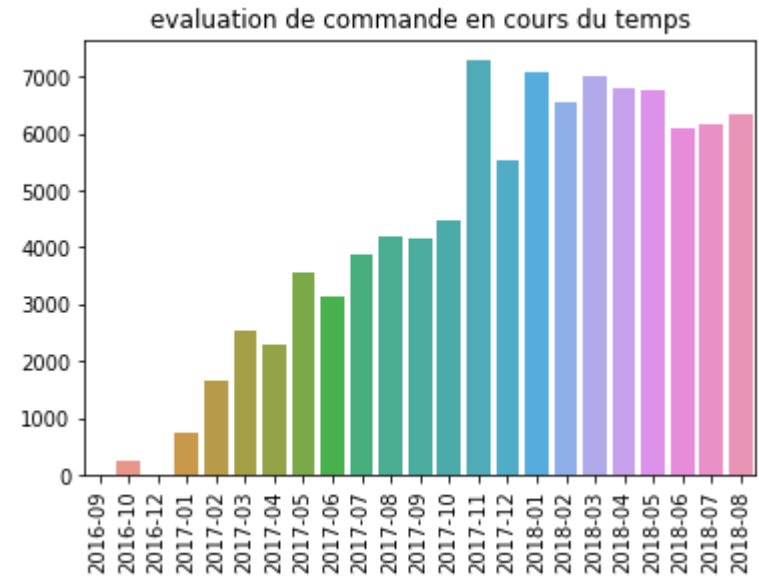


Septembre et Octobre sont des périodes creuses



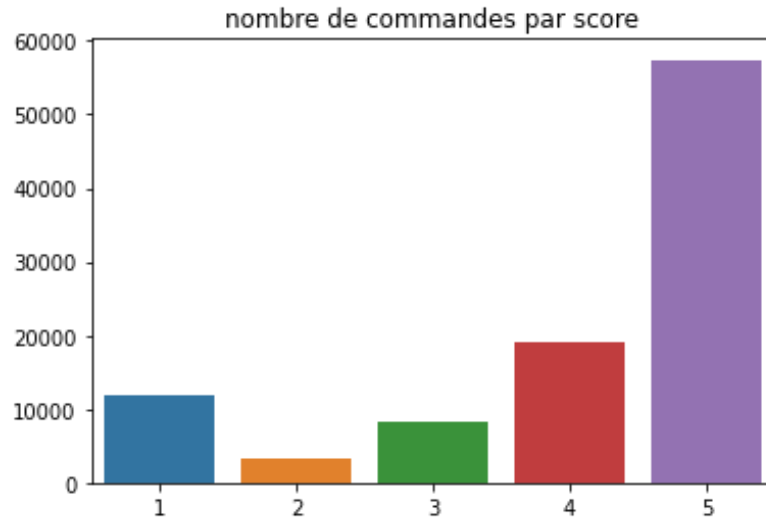
Dans le week-end il y'a moins d'achats effectués que durant la semaine

A partir de fin de 2017 le nombre de commandes à augmenté

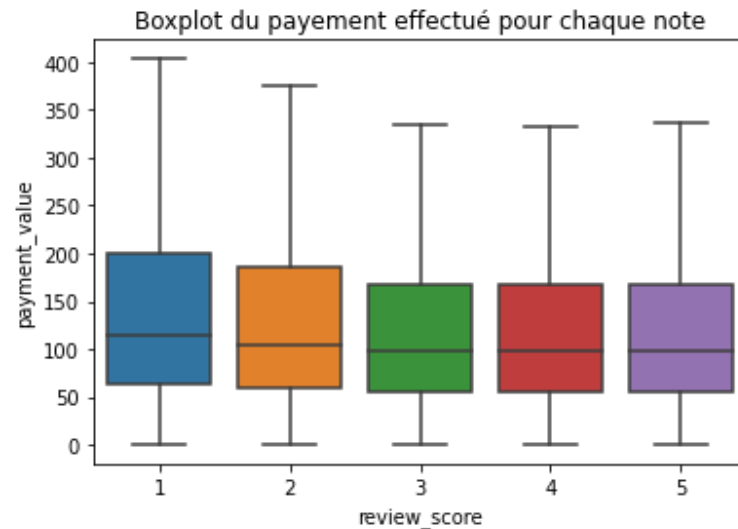


Moins de commande la nuit que dans la journée

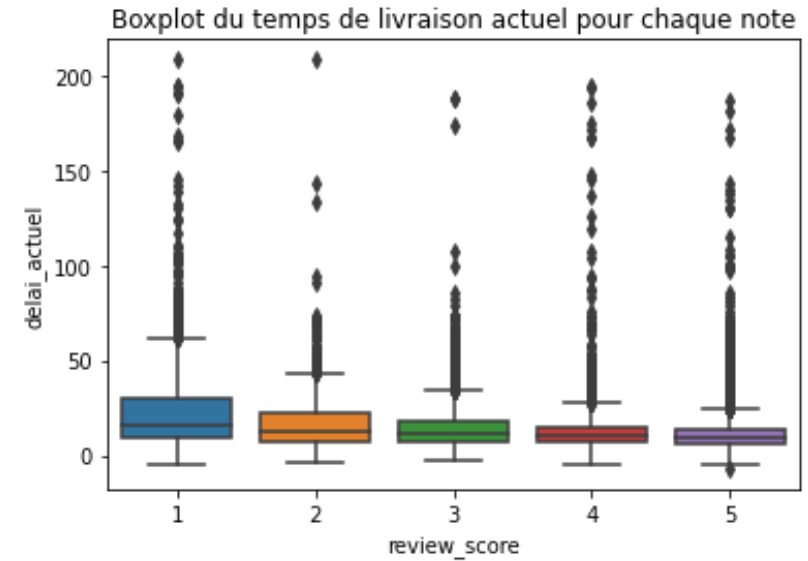
La majorité des produits sont scorées 5



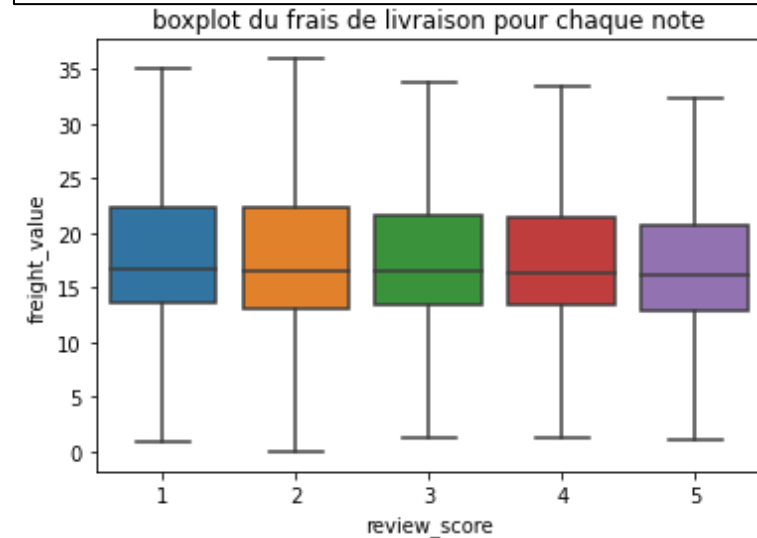
Plus le prix de la transaction est faible
et plus son évaluation est bonne



Plus le produit est livrée rapidement
et plus l'évaluation du produit est bonne

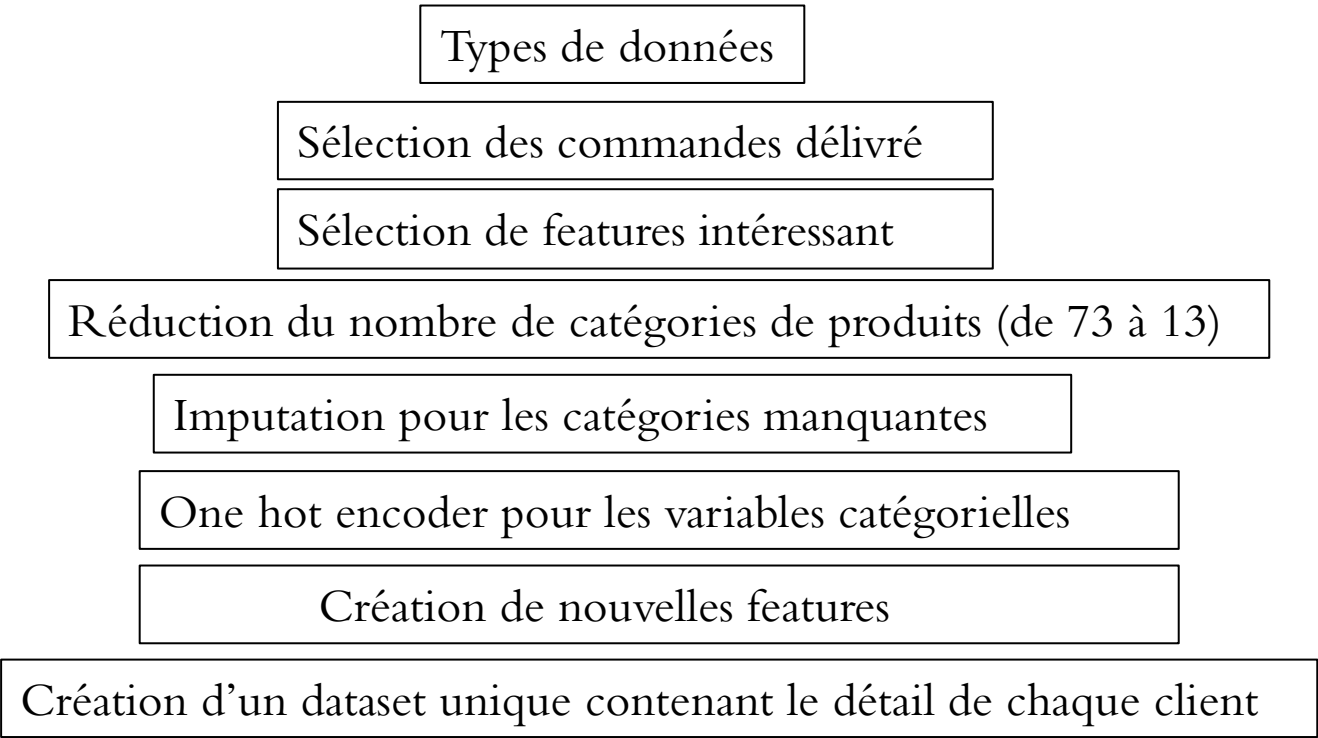


Aucune relation entre les frais de livraison
et l'évaluation du produit



FEATURE ENGINEERING

- Principales étapes du nettoyage:



8 Dataset



1 DATASET de clients

Agrégation de données se fait sur la clé :
« customer_unique_id »

1 ligne = 1 client

Création de features

FEATURE ENGINEERING

Dataset finale

93335 clients

55 features

Aucune valeur manquante

- Choix de features adaptées:

INFOS ACHATS CLIENT

Nombre commandes
Montants total dépenses par clients
Note moyenne
Nombre de reviews
Moyen de paiement

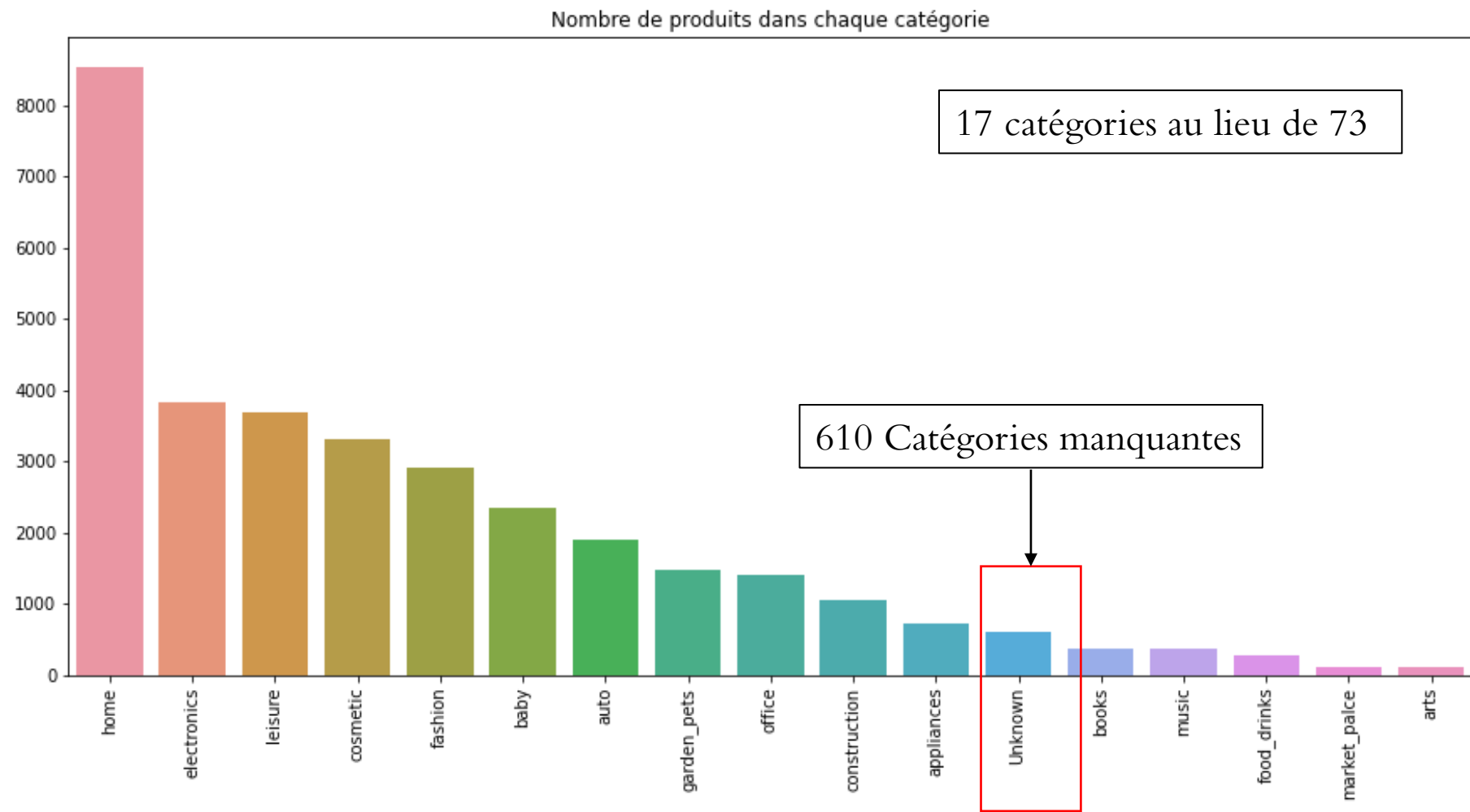
DONNEES TEMPORELLES

Livraison tard ou non
Client de Noël
Temps de livraison actuel et estime
Mois d'achat
Récence du dernier achat

CATEGORIES

Catégorie de produits

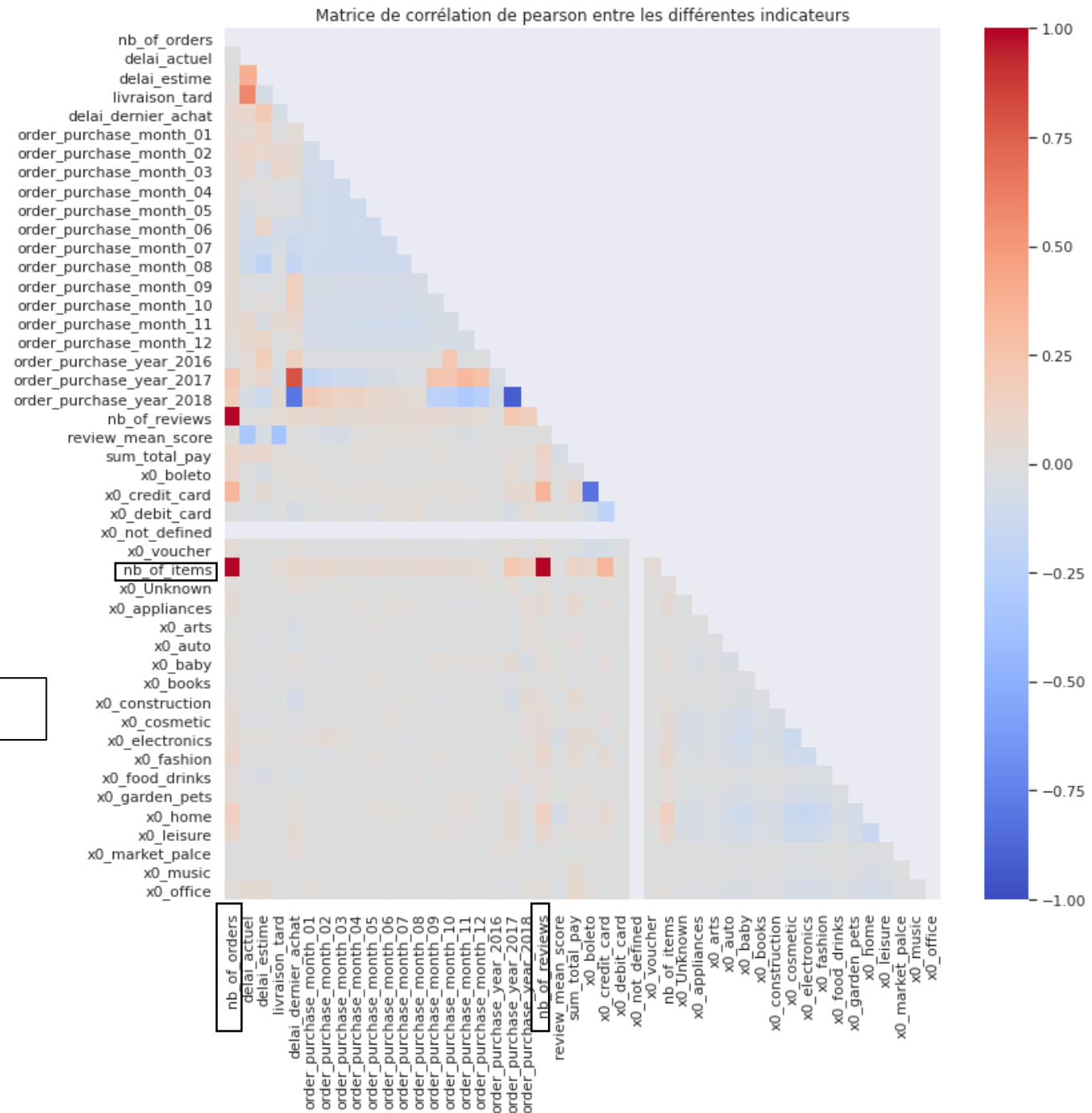
FEATURE ENGINEERING



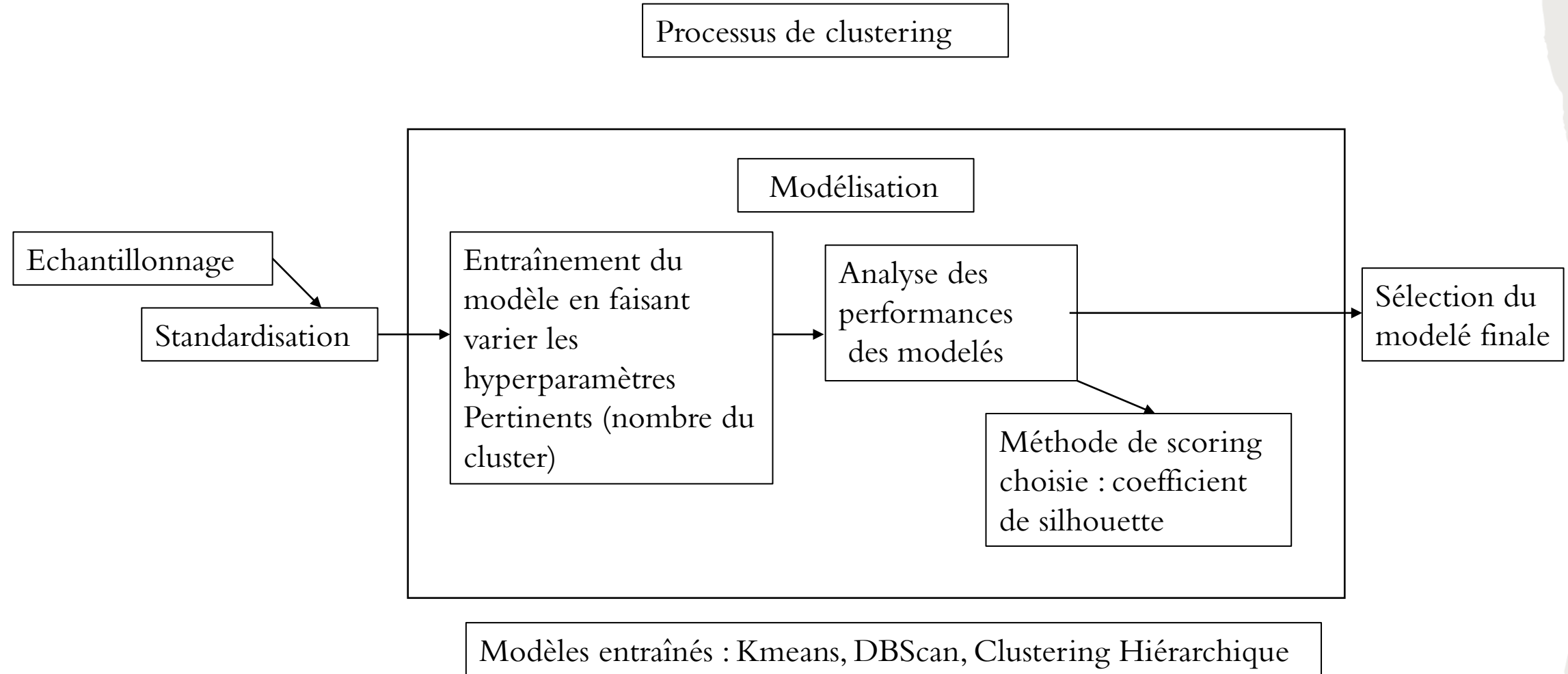
FEATURE ENGINEERING

- Forte corrélation entre :
 - Nombre de produits achetés et nombre d'ordres.
 - Nombre d'ordres et nombre de reviews.
 - Livraison tard et délai actuel de livraison.

Suppression des variables trop corrélées



TESTS DES ALGORITHMES DE CLASSIFICATION NON SUPERVISÉE



Pour le modelé de : Kmeans

- But : Détermination optimum du nombre de clusters

Critères de sélection

Calcul inertie par la methode d'elbow

Maximisation du coefficient de Silhouette

Minimisation de score Davis bouldin

Répartition des clients par cluster (en évitant d'avoir un cluster qui représente 90% des clients)

Nombre de clusters cohérent par rapport à la problématique marketing (10 max)

Visualisation des clusters 2D (ACP) et TSNE

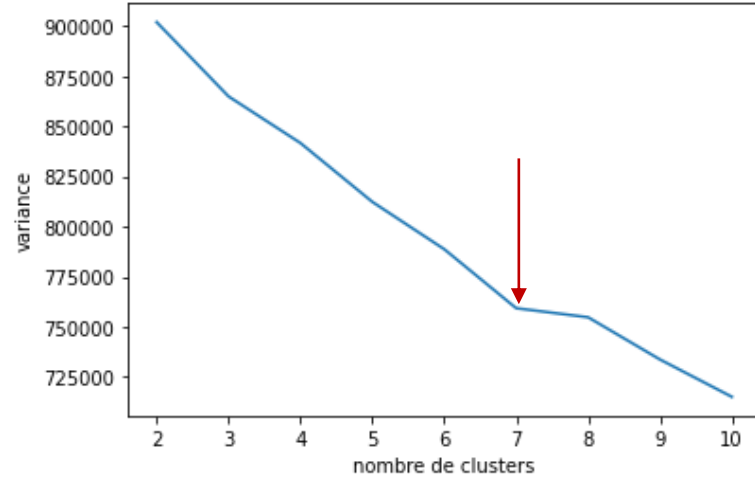


Vérification de la stabilité de clustering sur plusieurs itérations

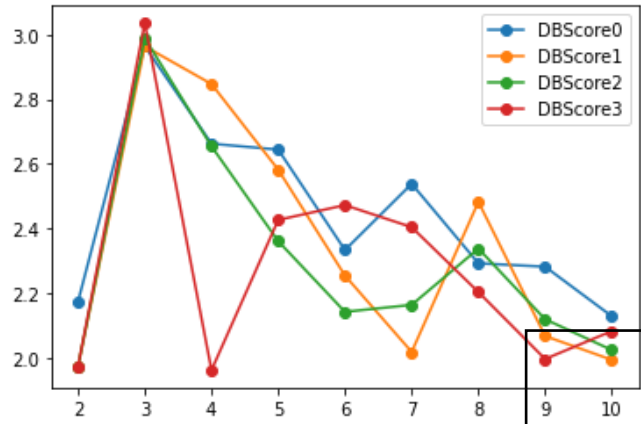
Pour le modelé de : Kmeans

- Entraînement de modèles avec 2 à 10 clusters

Comparaison de la somme des inerties en fonction du nombre de clusters



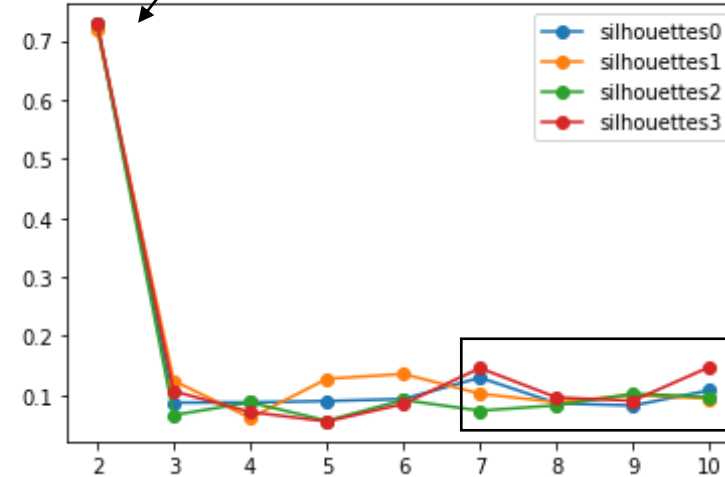
coefficient de Davies-Bouldin en fonction du nombre de clusters



Zone plus stable

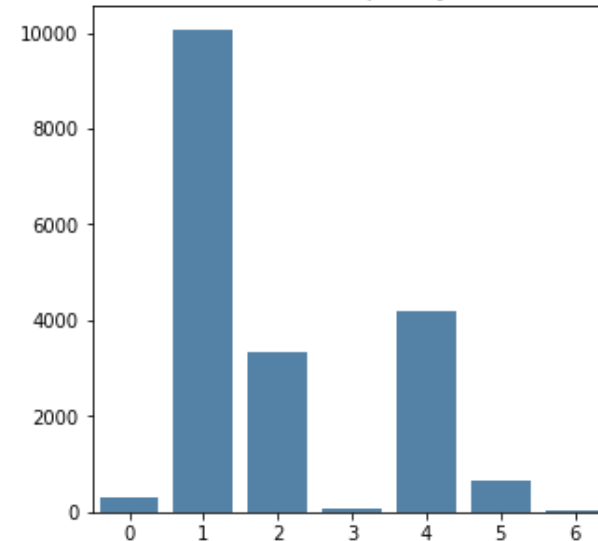
Silhouette intéressante mais clusters non pertinents

coefficient de silhouette en fonction du nombre de clusters



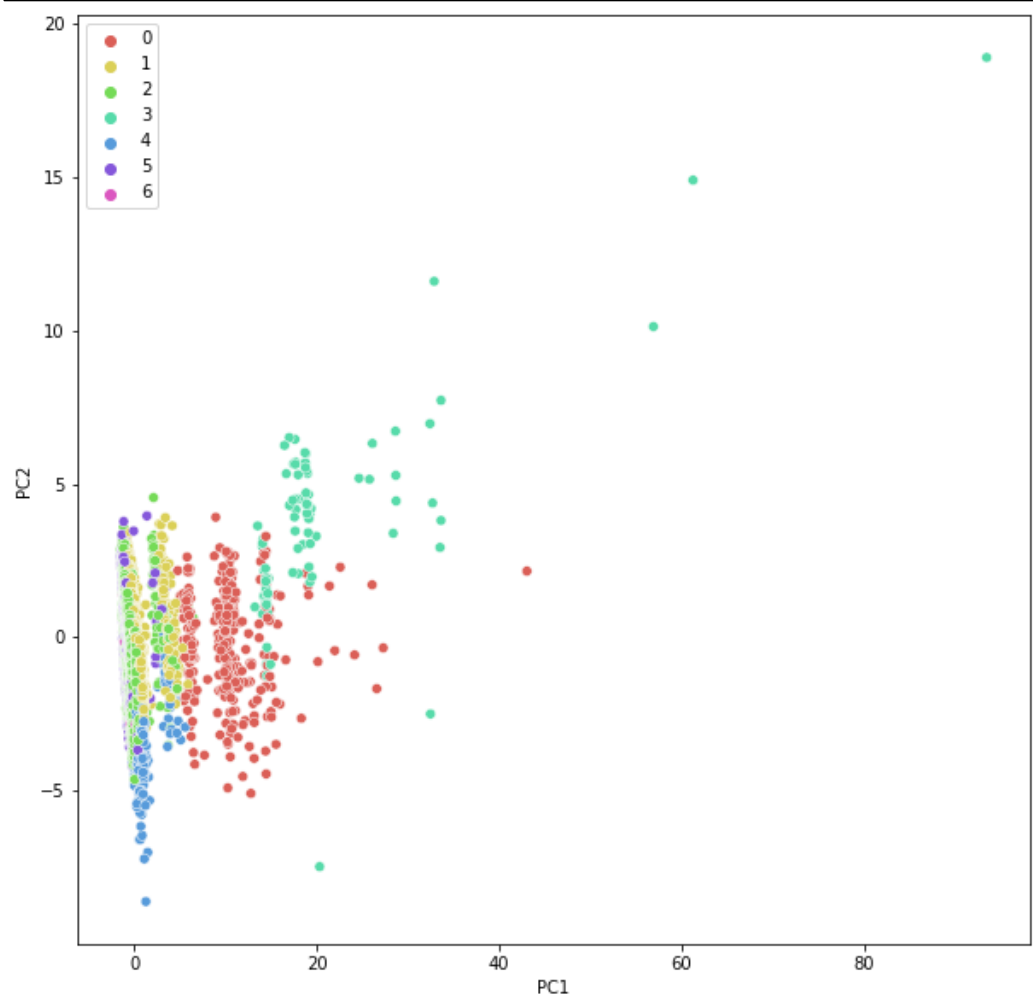
Zone plus stable

Number of sample by cluster

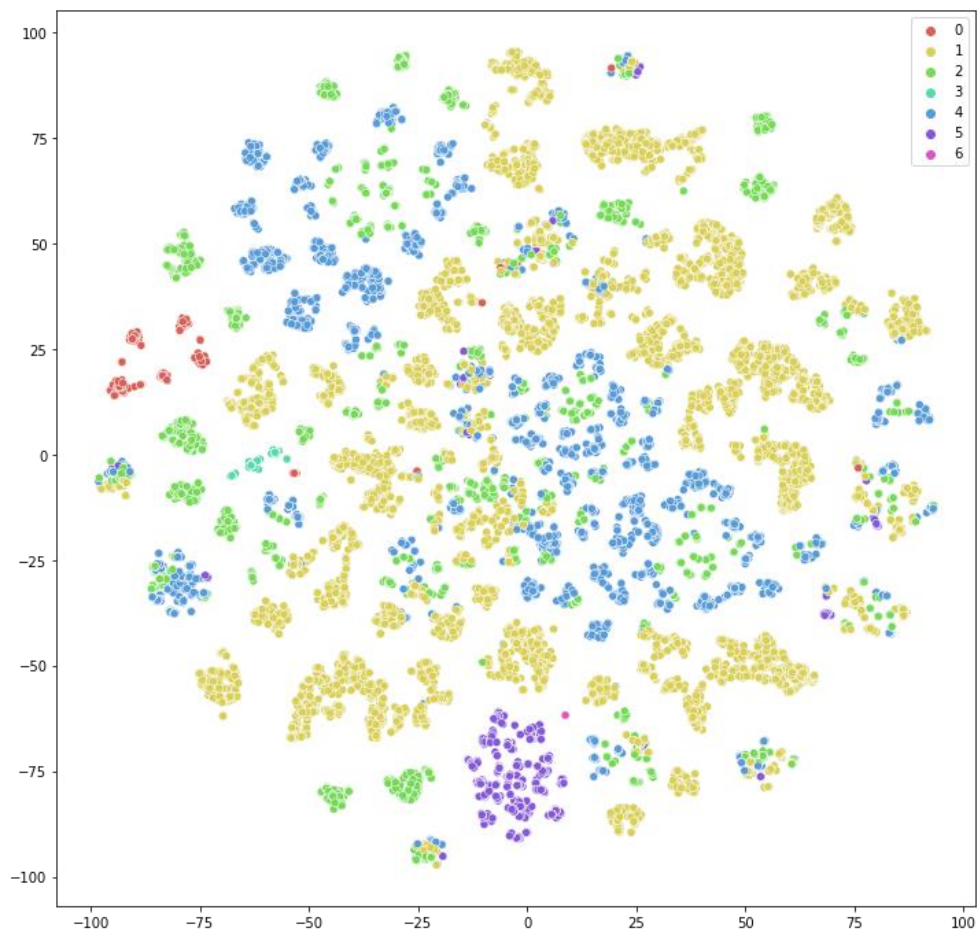


Visualisation du données pour n= 7 clusters

Projection des données sur les 2 premières composantes de l'ACP

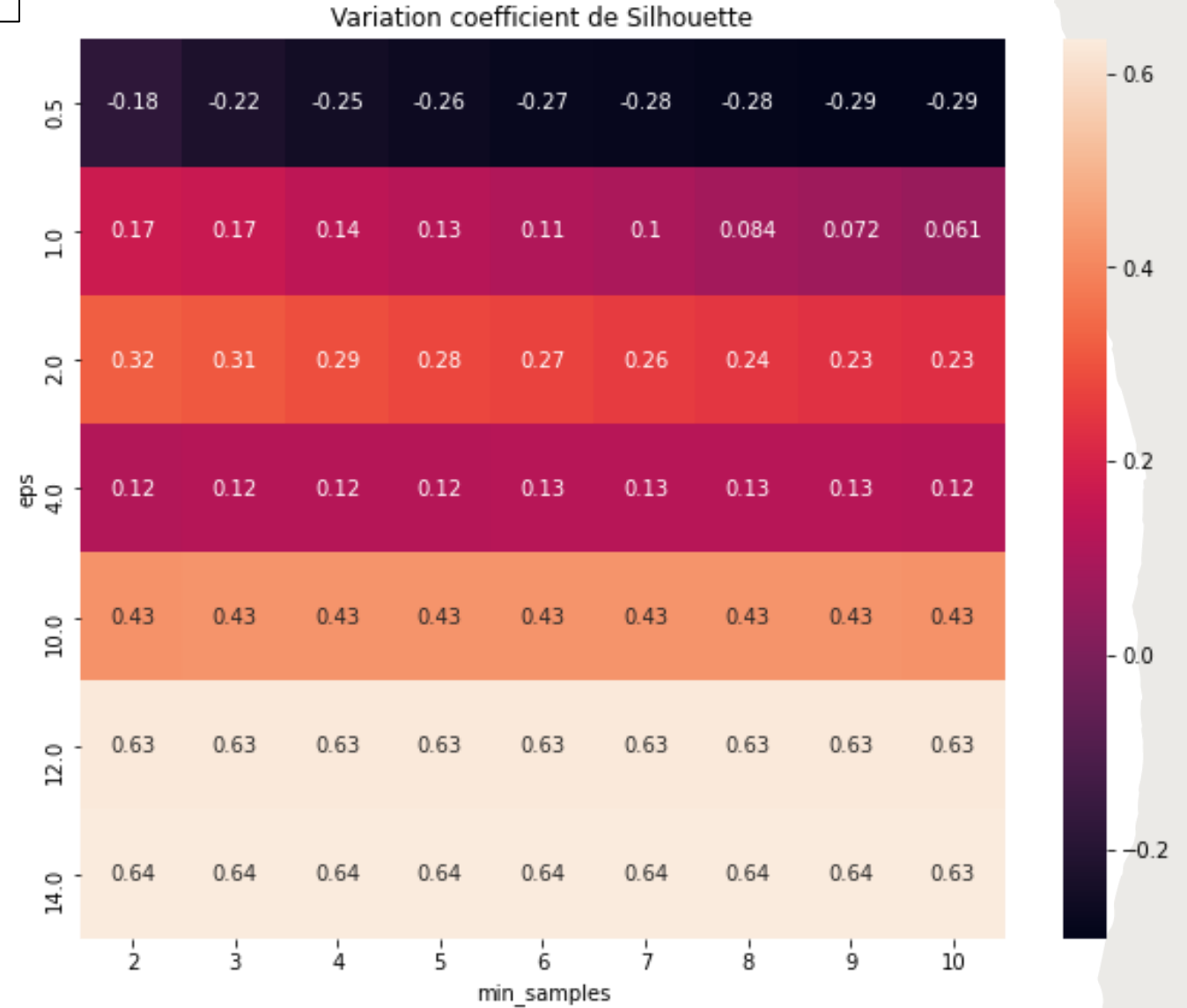
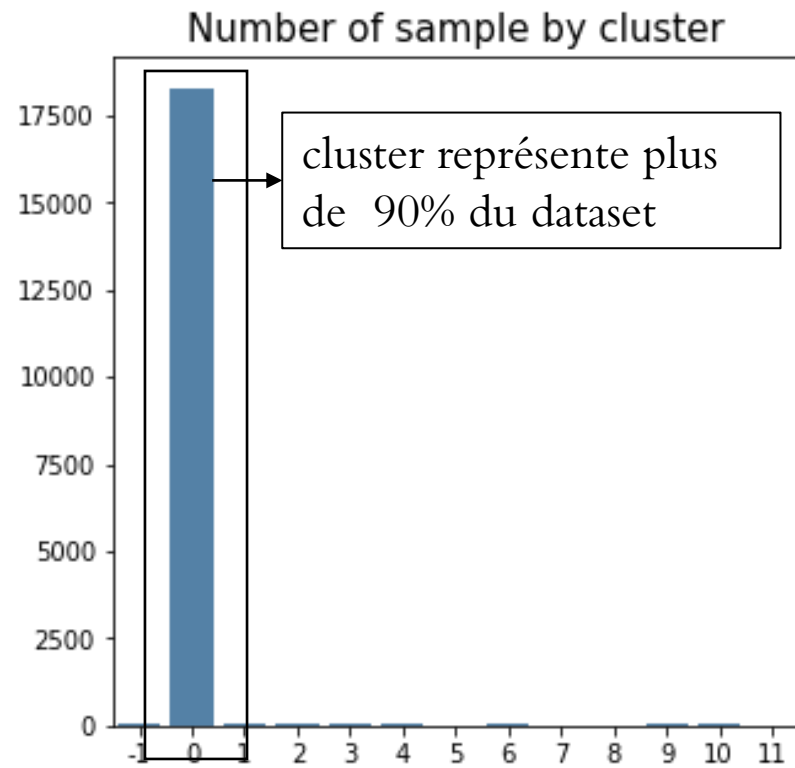


Représentation des données via T-SNE



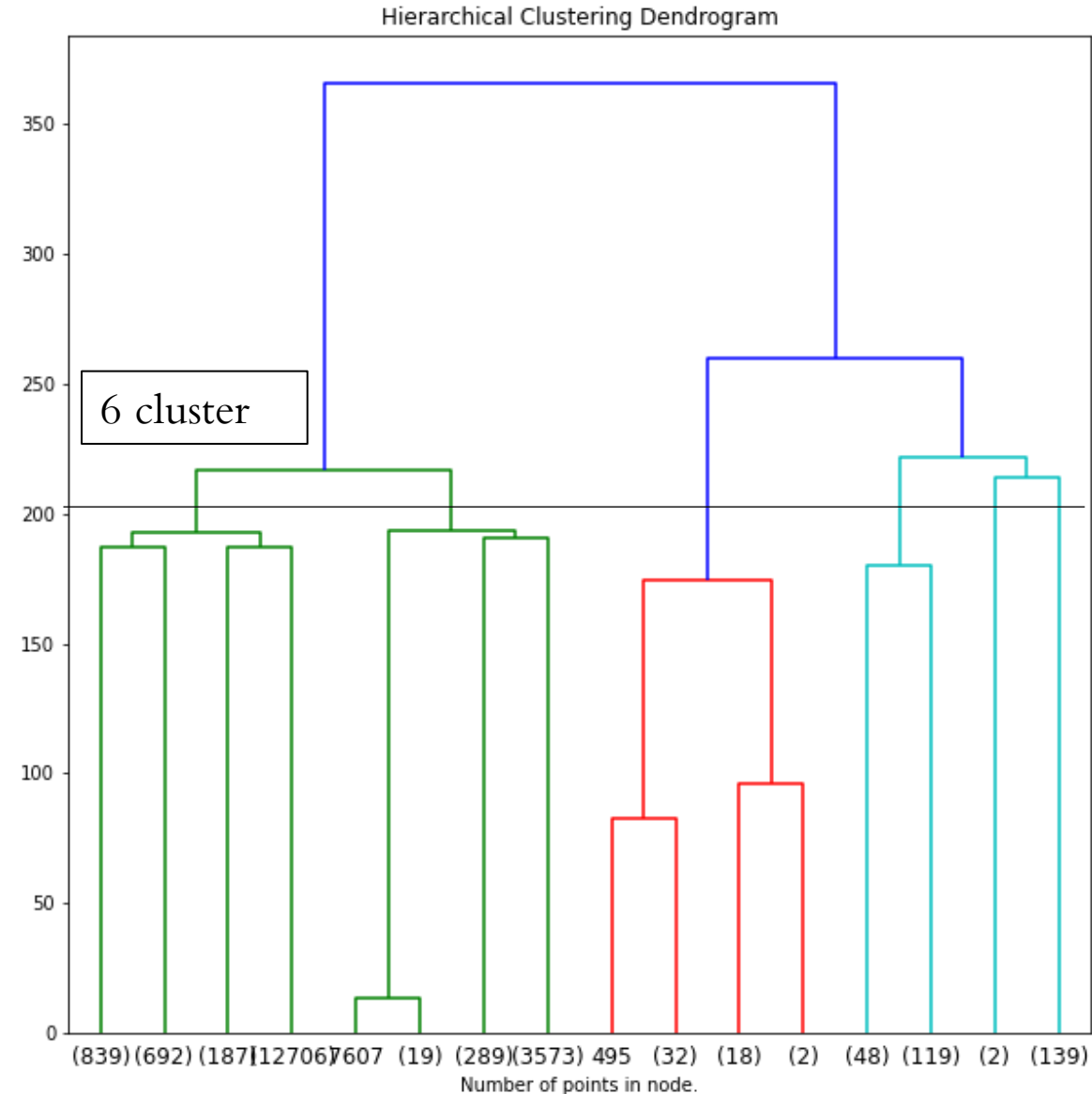
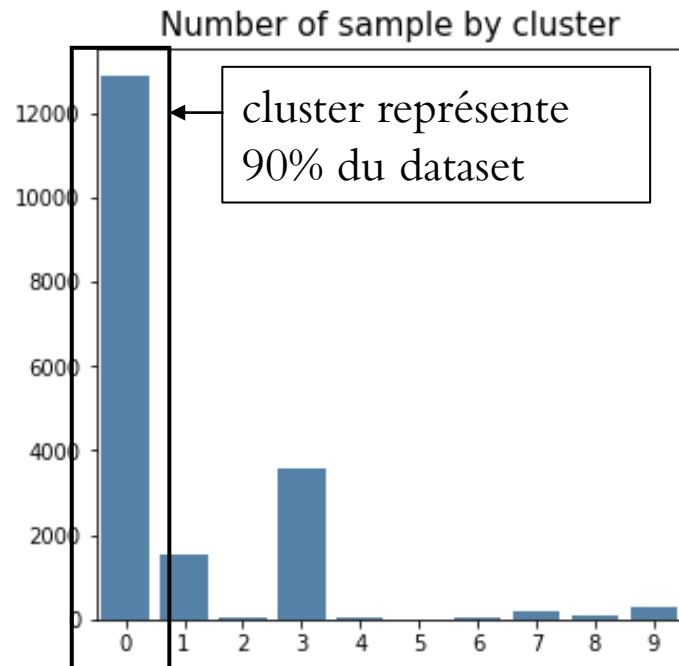
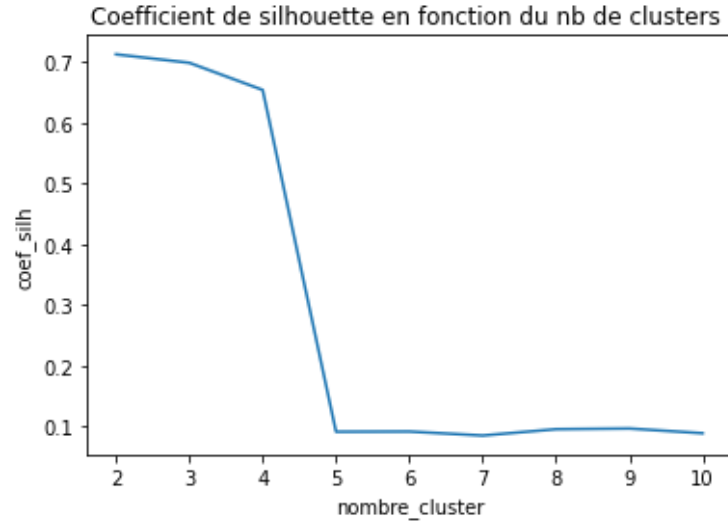
DBScan

DBSCAN(eps, min_samples)



Clustering Hiérarchique

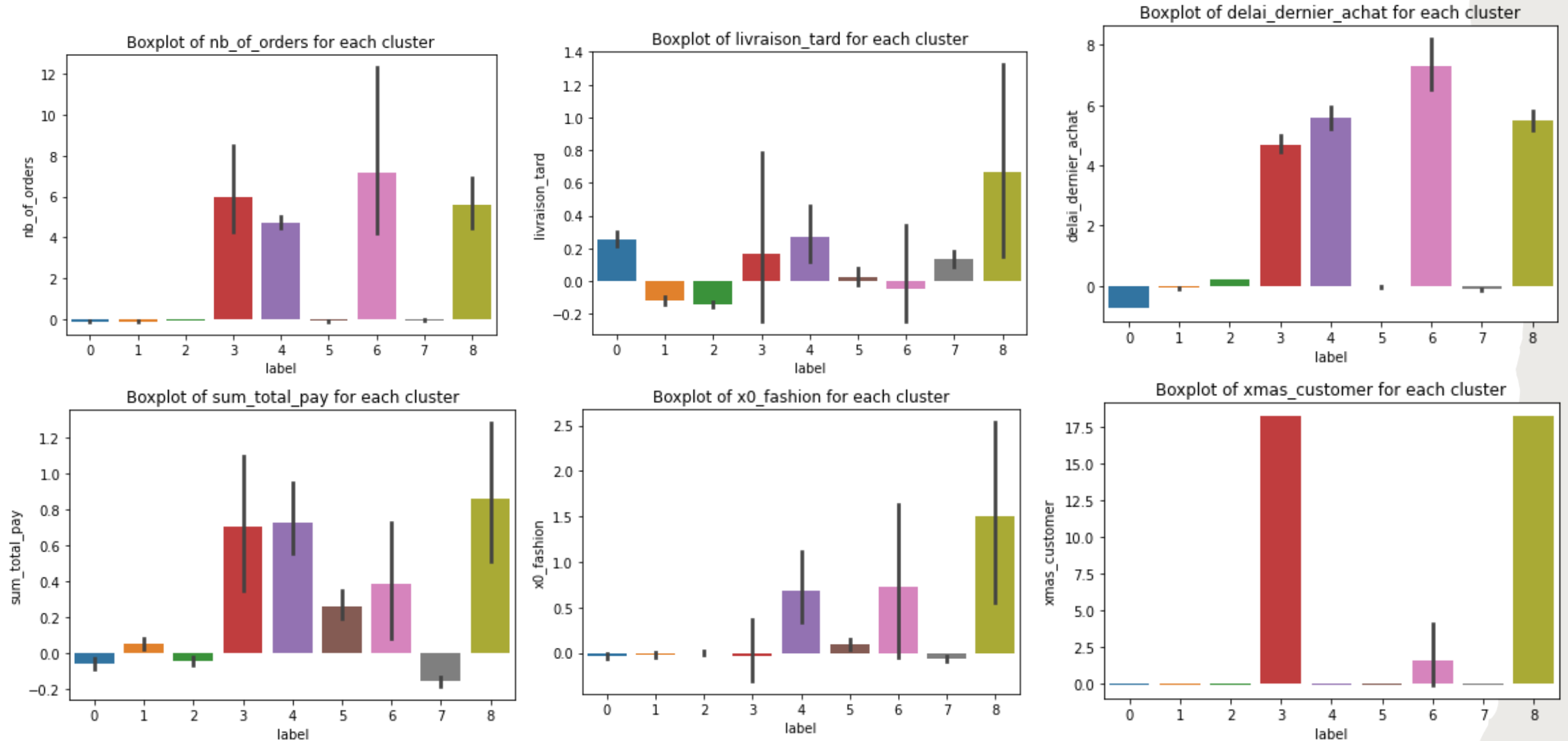
```
AgglomerativeClustering(n_clusters=4, linkage="ward", affinity="euclidean")
```



MODÈLE FINAL SÉLECTIONNÉ →

Kmeans avec cluster : 9

Analyse des caractéristiques des différents clusters identifiés



Observations du comportement des clusters au cours du temps

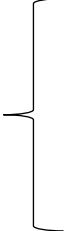
A faire

Observ cluster characteristics for both predicted and new fitted clusters

Score de similarité: 'Adjusted Rand Score'

les clusters évoluent chaque 3 mois : il faut actualiser sur une période de 3 mois

Conclusion

- A partir de 8 datasets :
 - Création de features catégorielles permettant d'expliquer la segmentation des clusters (catégories de produits, dépenses..)
 - Création d'un dataset listant les détails de clients
 - Mise en application des modèles de classification non supervisée 
 - Kmeans
 - Clustering hiérarchique
 - DBScan
 - Modèle finale sélectionnée : Kmeans avec un nombre de clusters optimal 14
 - Clustériser nos utilisateurs via un algorithme de kmeans qui nous permet de différencier différents types d'utilisateurs.
 - Un intervalle de temps de 3 mois pour la maintenance de la segmentation déterminé par l'analyse des données fournies

Merci pour votre attention