PROJET 5

Segmentez des clients d'un site e-commerce

Soutenance de projet

Sommaire

1. Présentation de la problématique :

- Rappel de la problématique
- Interprétation
- Pistes de recherche envisagées

Interprétation de la problématique et pistes de recherche envisagées

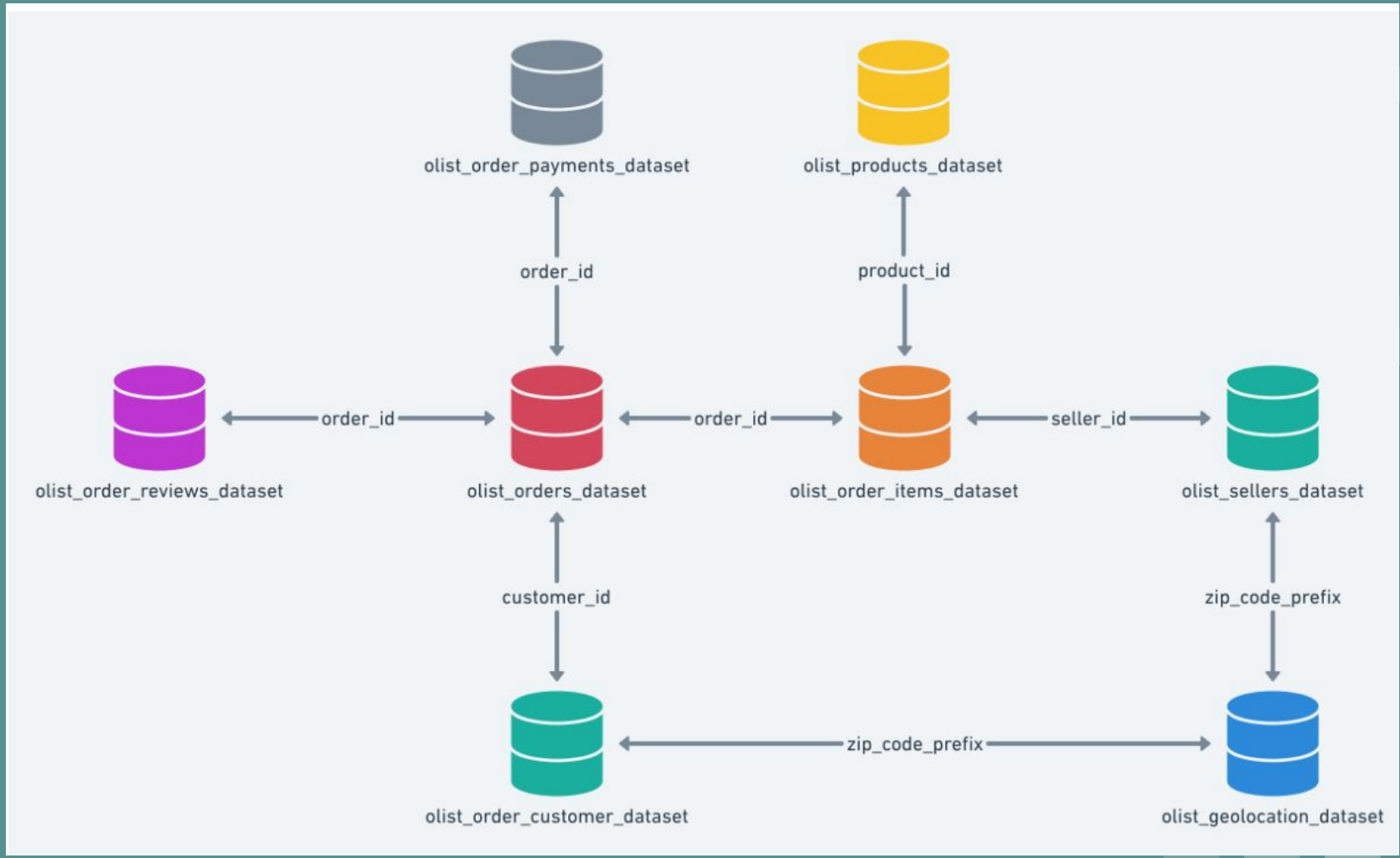
- Exploration des données et choix de features adaptées
- Problème de classification non supervisée
- Les clusters devront être explicables et réutilisables pour des campagnes de communication

2 – PRÉPARATION DU JEU DE DONNÉES

- Cleaning
- Feature engineering
- Exploration

Cleaning

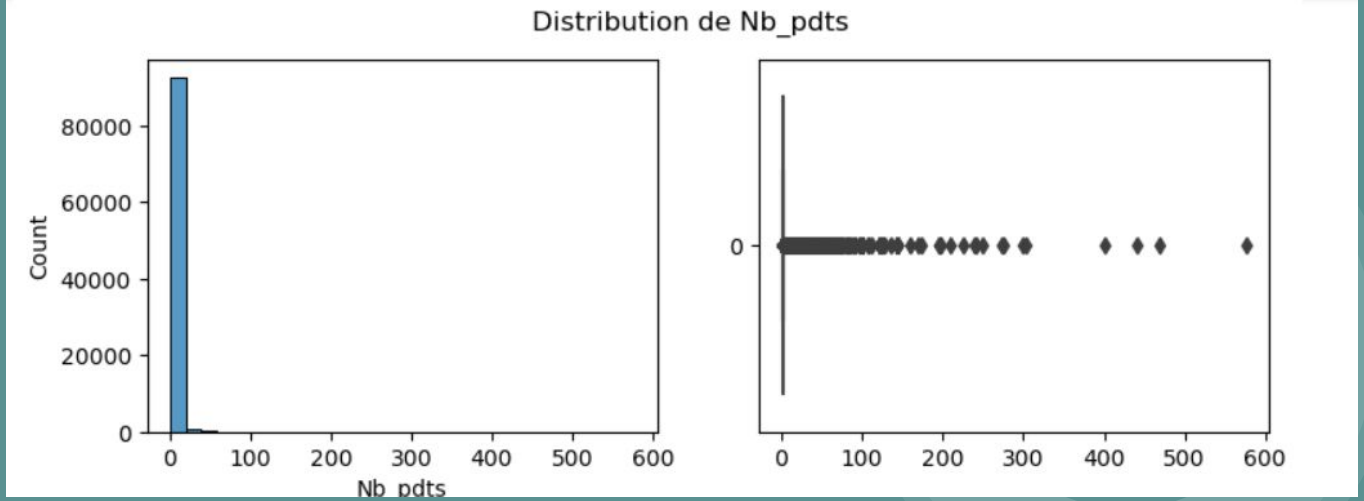
- Données réparties en 9 tables:
  - clients
  - geolocalisation
  - commandes
  - paiements
  - produits
  - vendeurs
- Traduction des catégories de produit

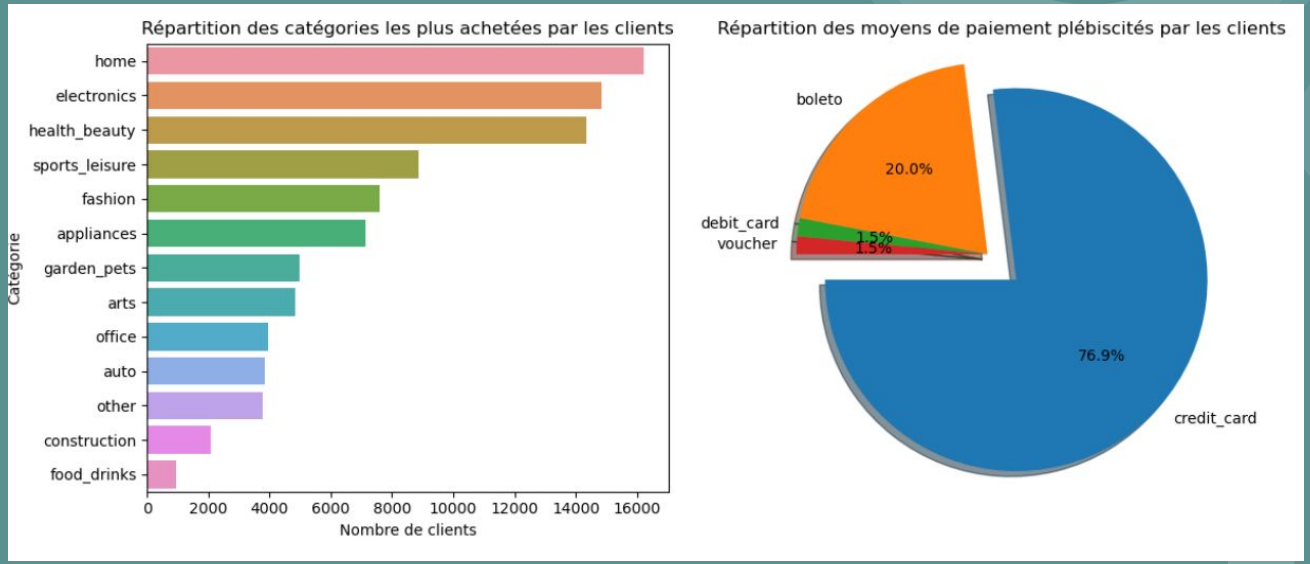

olist_order_payments_dataset
olist_products_dataset
order_id
product_id
olist_order_reviews_dataset
order_id
olist_orders_dataset
order_id
olist_order_items_dataset
seller_id
olist_sellers_dataset
customer_id
zip_code_prefix
olist_order_customer_dataset
zip_code_prefix
olist_geolocation_dataset

Principales étapes du nettoyage

- Types de données
- Réduction du nombre de catégories de produits (de 72 à 12)
- Création de nouvelles features
- Assemblage dans une table unique avec pour index l’id client

Exploration





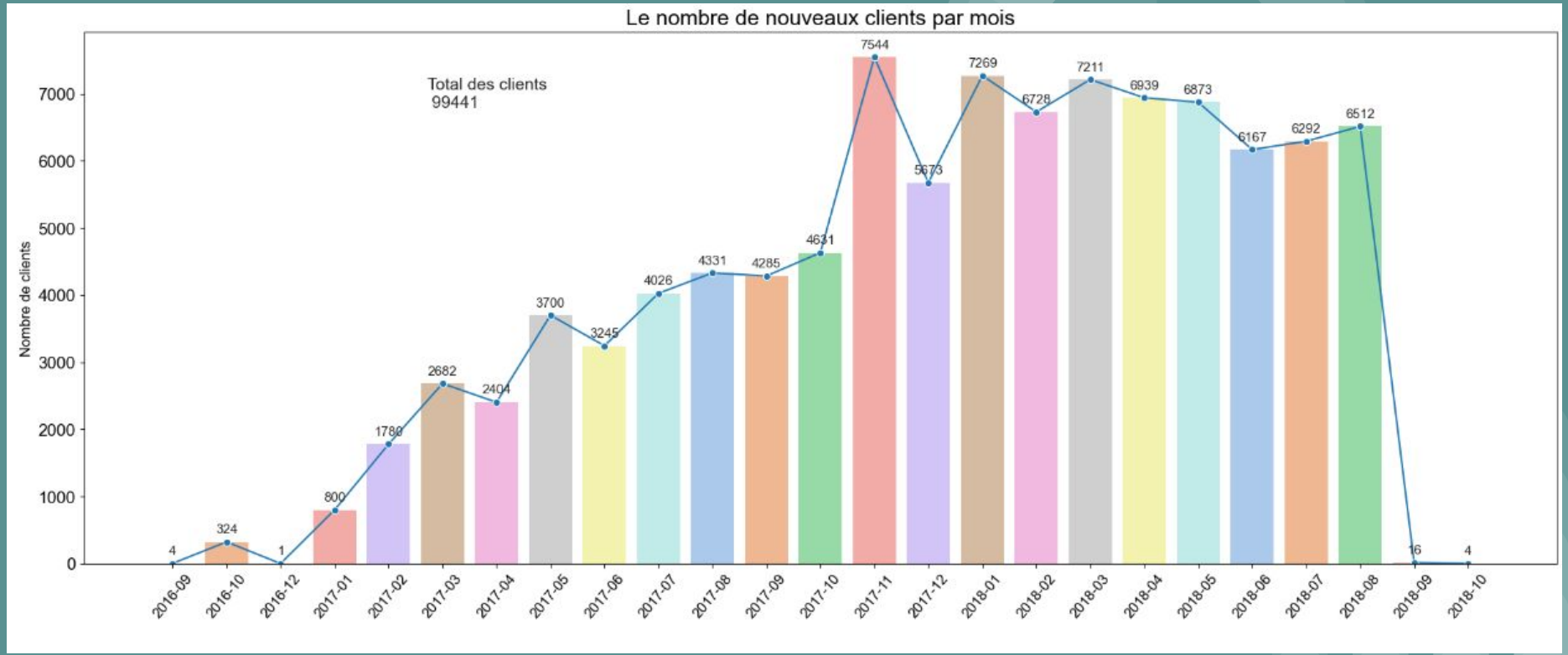
Population sur laquelle nous avons travaillé: 93 396 transactions

```
1  data = pd.read_csv('data_cleaned.csv')
```

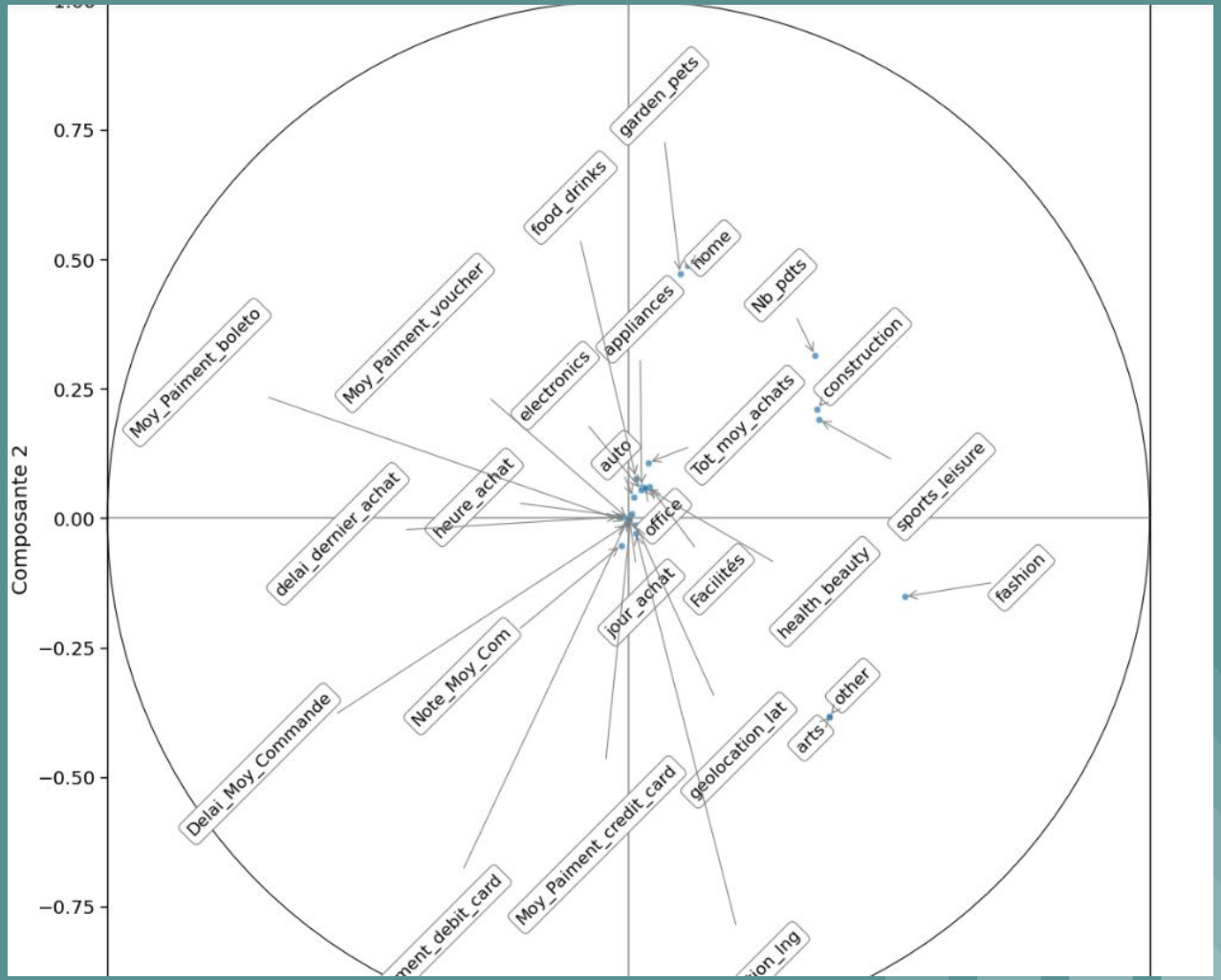
```
1  data.shape
```

```
(93396, 30)
```

Le nombre de nouveaux clients par mois



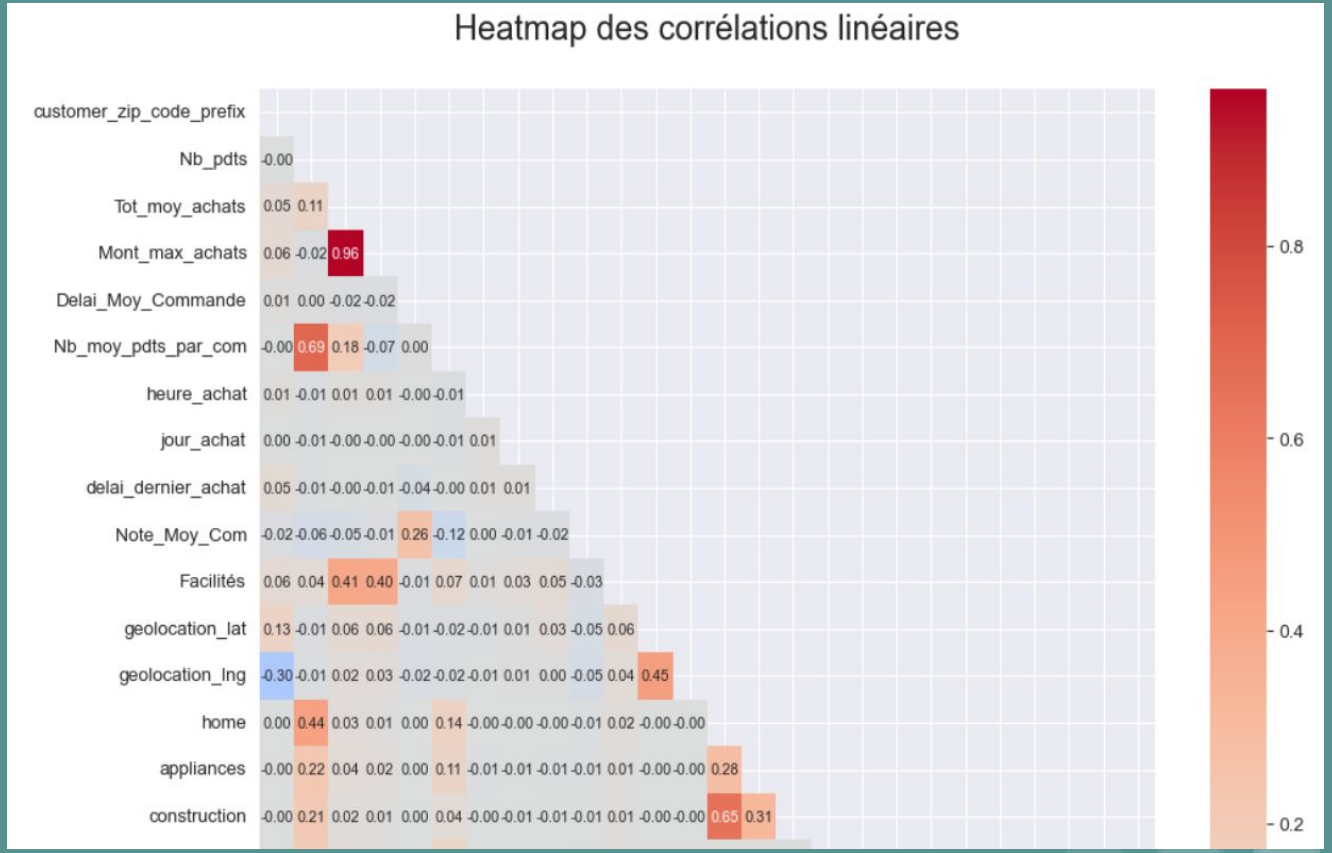
Plan factoriel pour les 2 premières composantes de l'ACP



Commentaires

1. La première ligne du résultat montre une corrélation très forte de 0.96203 entre les variables "Tot_moy_achats" (Total des achats moyens) et "Mont_max_achats" (Montant maximal des achats).

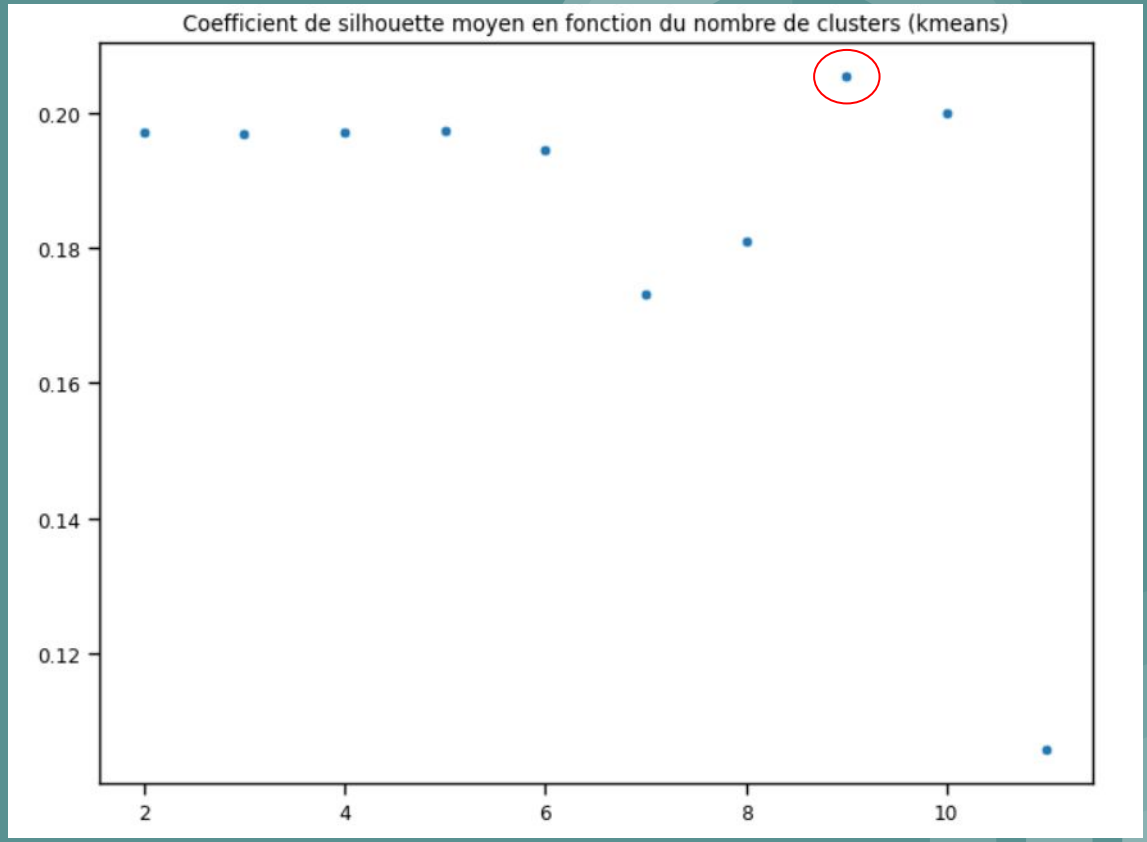
2. La deuxième ligne du résultat montre une corrélation de 0.69426 entre les variables "Nb_pdts" (Nombre de produits) et "Nb_moy_pdts_par_com" (Nombre moyen de produits par commande). Cette corrélation indique une relation positive modérée entre ces deux variables.



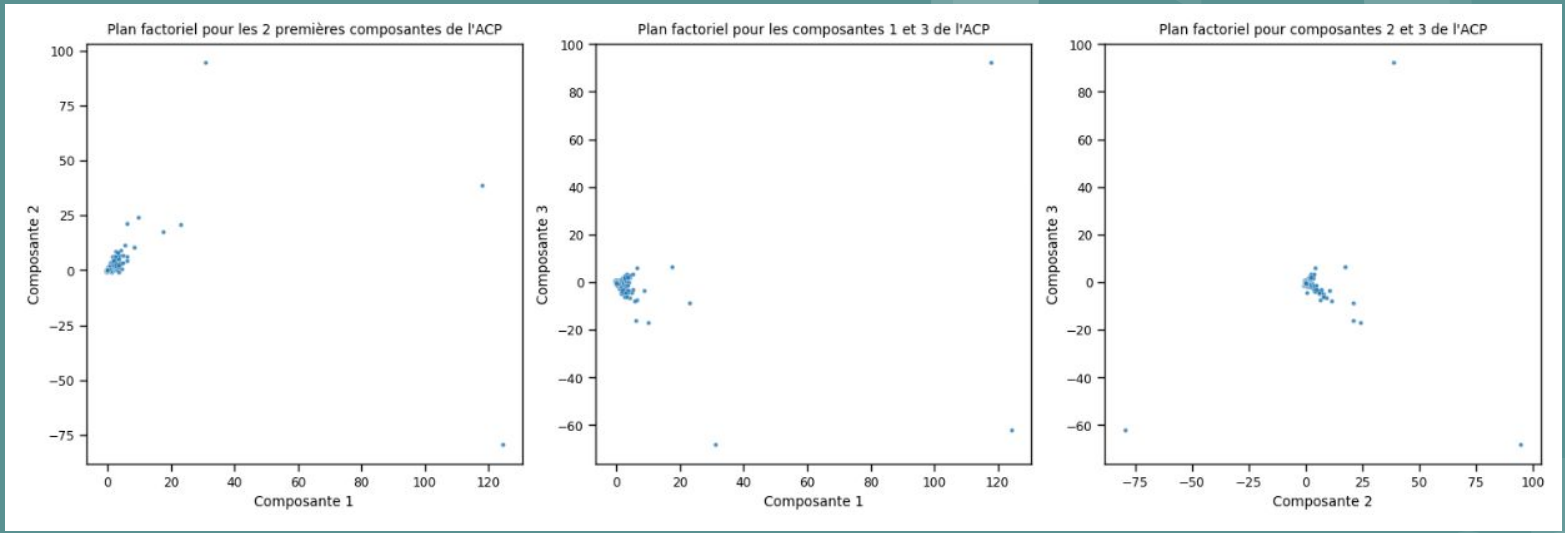
3. PISTES DE MODÉLISATIONS

Kmeans : détermination optimum du nombre de clusters

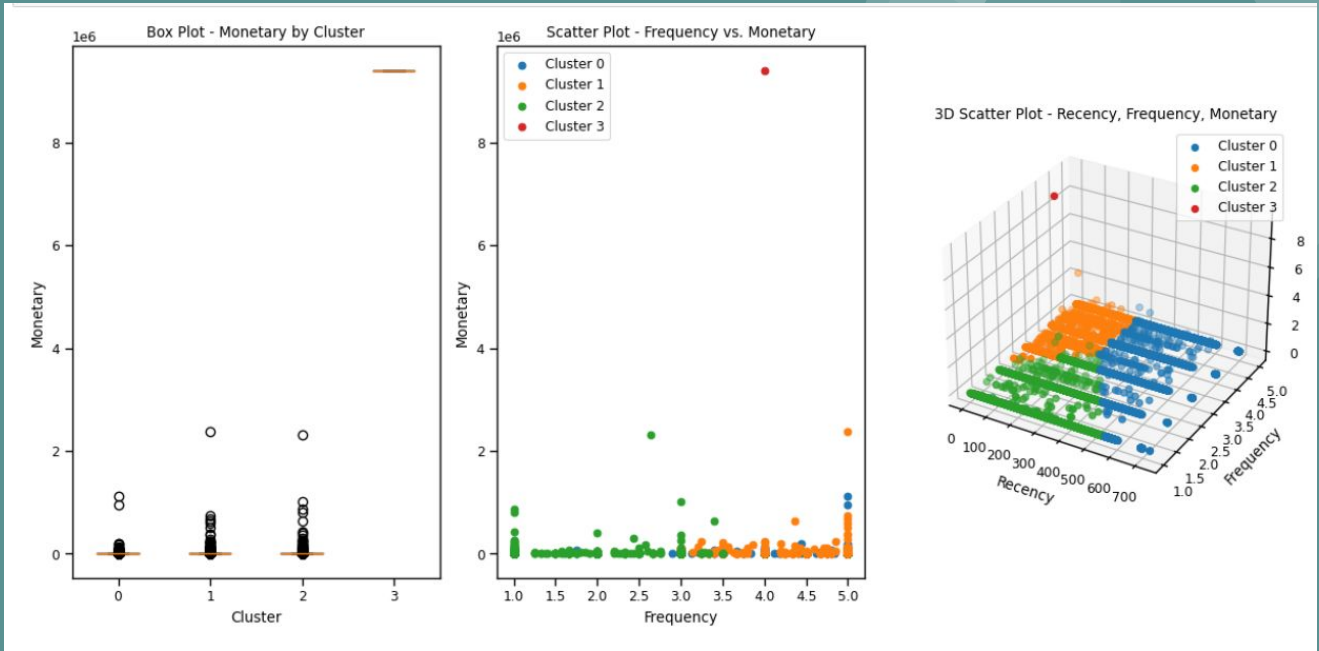
Le coefficient de silhouette est maximal pour 09 clusters



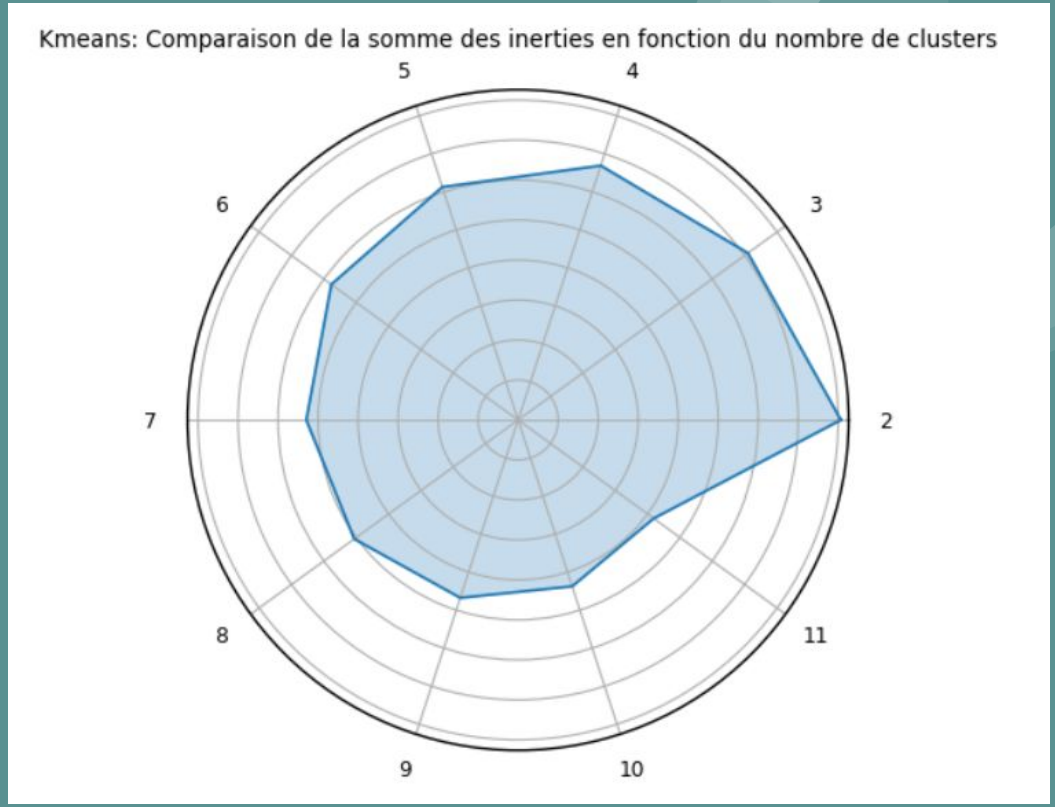
Plan factoriel pour les 2 premières composantes de l'ACP



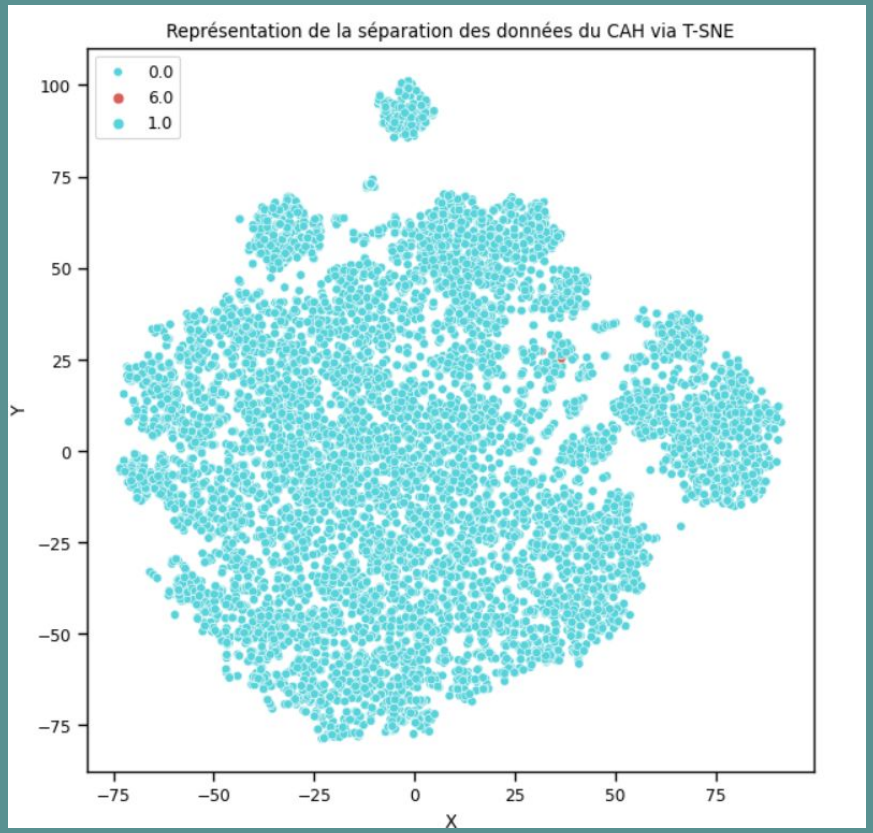
RFM Recency, Frequency, Monetary



Kmeans : représentation graphique

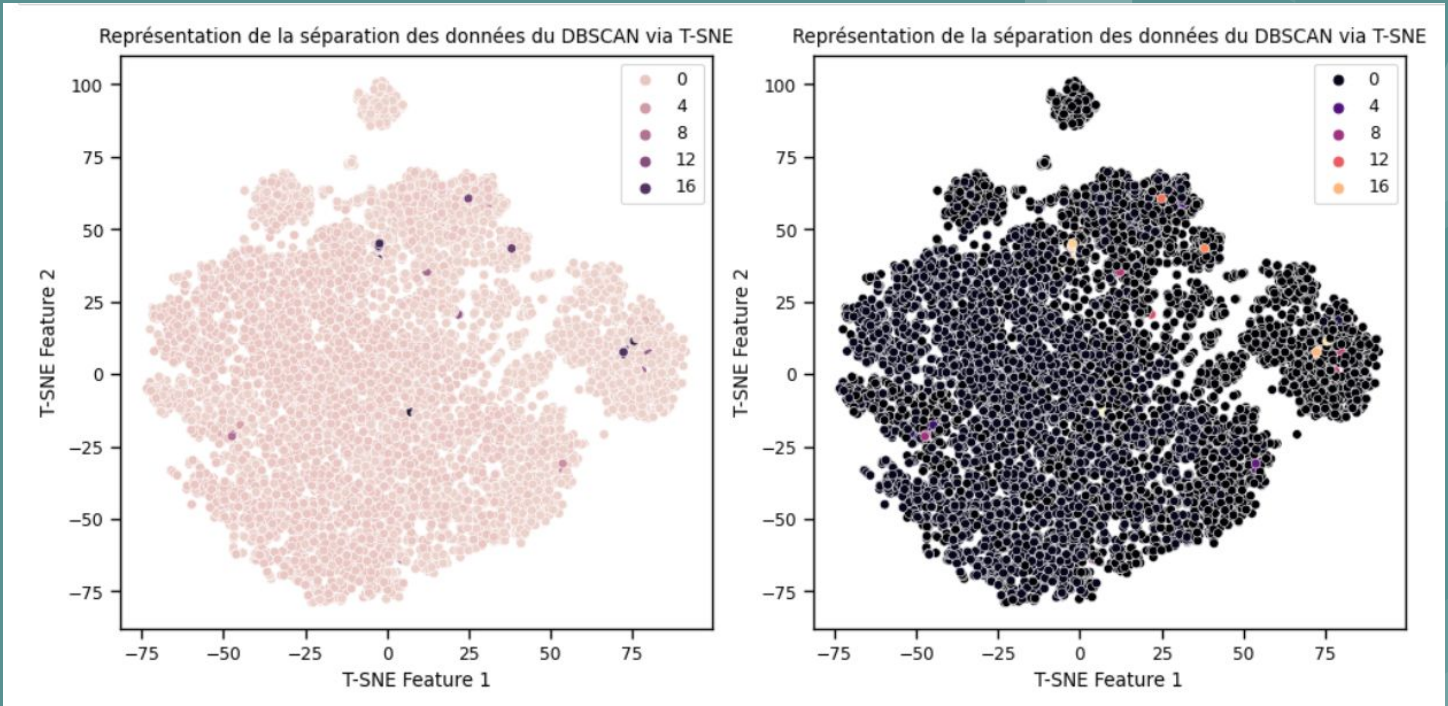


Clustering hiérarchique



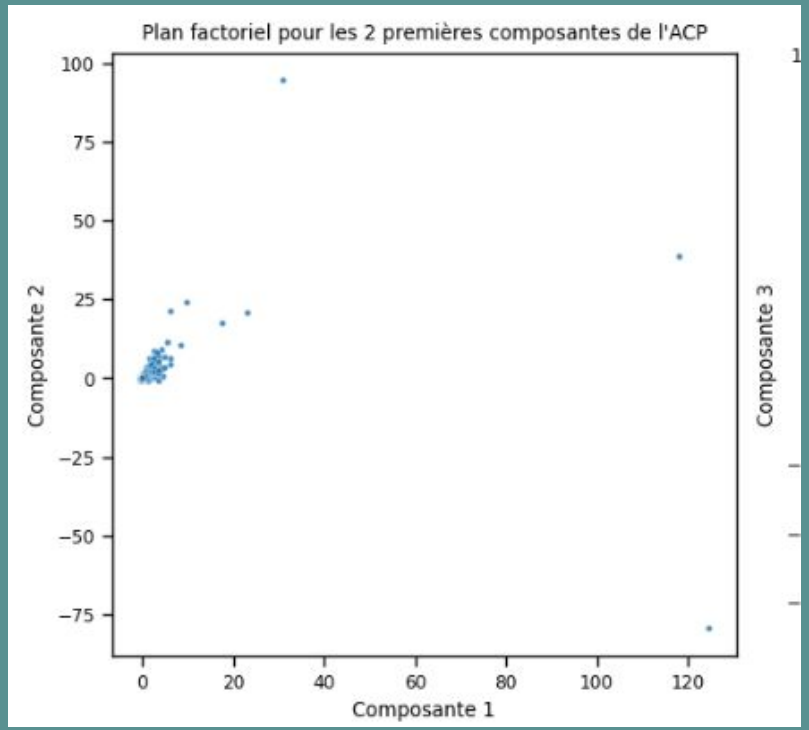
DBScan

- Exemple
- Epsilon = 1
  Min_samples = 5



04. PRÉSENTATION DU MODÈLE FINAL

Kmeans sur intégralité de l'échantillon



- Stabilité du silhouette score

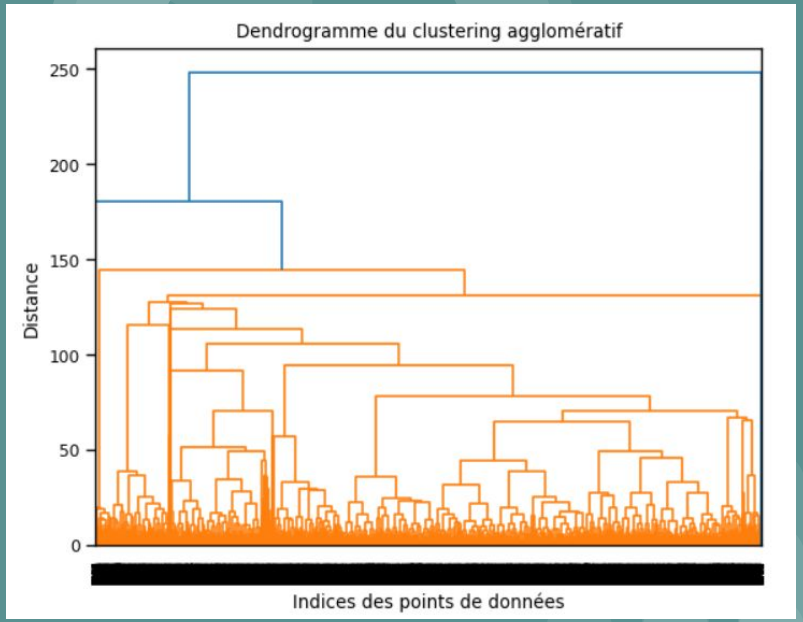
```
Boucle 2 - Silhouette score : 0.1970610921621437
Boucle 3 - Silhouette score : 0.19694119608375946
Boucle 4 - Silhouette score : 0.1970332090941206
Boucle 5 - Silhouette score : 0.19740581811745672
Boucle 6 - Silhouette score : 0.1944219573679476
Boucle 7 - Silhouette score : 0.17323989083960037
Boucle 8 - Silhouette score : 0.18088261429830854
Boucle 9 - Silhouette score : 0.20532644440123715
Boucle 10 - Silhouette score : 0.19989978591176072
Boucle 11 - Silhouette score : 0.10563584761257852
```

Clusters identifiés et actions

	Catégorie	Nombre de clients	Note_Moy_Com	Delai_Moy_Commande	Tot_moy_achats
0	home	16213	4.15712	-411.18703	4.35963
1	electronics	14828	4.22282	-390.46121	4.39442
2	health_beauty	14351	4.08512	-373.86340	4.46826
3	sports_leisure	8873	4.09174	-305.92086	4.73427
4	fashion	7597	4.03356	-374.39343	4.10117
5	appliances	7125	4.09147	-406.08197	4.70044
6	garden_pets	4987	4.26620	-386.40394	3.93364
7	arts	4849	4.17759	-279.18416	4.48607
8	office	3952	4.15795	-358.91183	4.49162
9	auto	3828	3.99628	-302.64166	4.47559
10	other	3765	4.04005	-299.77259	4.68028
11	construction	2092	4.17548	-325.19797	4.77639
12	food_drinks	936	4.18811	-390.01356	4.52303

Clustering hiérarchique

regroupement agglomératif en utilisant la méthode de liaison de War



Clusters identifiés :

- Subdivision du cluster 5 : non concluant (silhouette score : 0.11)

- Suppression de certaines features : pas d’amélioration du clustering constatée

Contrat de maintenance

- **Identification de la période de maintenance:**

```
==========
#interval = pd.DateOffset(months=24)
**interval = pd.DateOffset(months=1) **

Fecha: 2016-10-04 00:00:00, ARI: 0.3119266055045872
Fecha: 2016-11-04 00:00:00, ARI: 0.22627221191010818
Fecha: 2016-12-04 00:00:00, ARI: 0.22627221191010818
Fecha: 2017-01-04 00:00:00, ARI: 0.22627221191010818
Fecha: 2017-02-04 00:00:00, ARI: 0.15085426208521896
```

- **Réduction du jeu de données sur la dimension « durée » (exemple : 3/4 mois)**

- **Vérification de la stabilité du nombre de clusters, du coefficient de silhouette et des valeurs des features • Compromis identifié : 3 mois**

- **Nombre de clusters optimal sur Kmeans : 14**
- **Coefficient de silhouette stable • Conservation des caractéristiques principales des clusters (catégories les plus dépensières, notes, etc.) • Variation à la marge de certaines valeurs de features**

Conclusion

- Mise en application des algorithmes de classification non supervisée et application à un problème métier

- Limites du clustering proposé

- Pas ou peu d'apport des algorithmes
- Opportunités d'amélioration du clustering
- Nouvelles features / clients ayant acheté plusieurs articles
- Caractérisation dans le détail des produits des champs textuels

Données plus précises sur les clients (à anonymiser) : âge, sexe 25

MERCI DE VOTRE ATTENTION