

PROJET 5 – « SEGMENTEZ DES CLIENTS D'UN SITE E-COMMERCE »

Soutenance de projet
22 Janvier 2020



Sommaire

- I. Présentation de la problématique
- II. Préparation des données et exploration
- III. Pistes de modélisations
- IV. Présentation du modèle final


I - PROBLÉMATIQUE

Rappel de la problématique

Interprétation

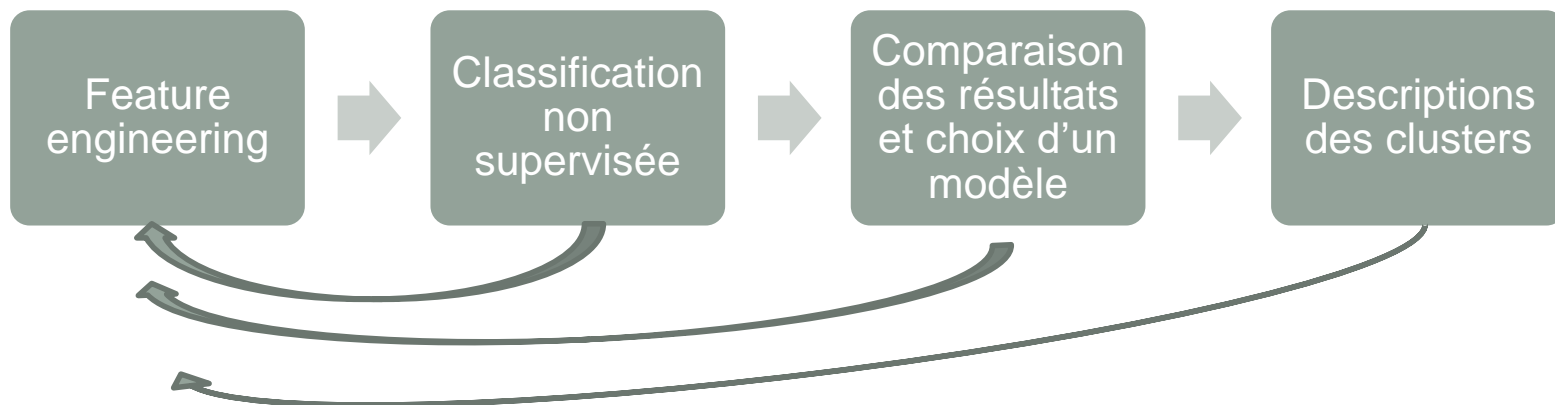
Pistes de recherche envisagées

Présentation de la problématique

- Mission de consultant pour **Olist**, solution de vente sur marketplaces en ligne 
- Objectifs:
 - Fournir aux équipes d'e-commerce une segmentation des clients pour les campagnes de communication
 - Comprendre les différents types d'utilisateurs
 - Fournir une description actionable de la segmentation
 - Faire une proposition de contrat de maintenance

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 expliquables et réutilisables pour des campagnes de communication



II – 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 produits

Principales étapes du nettoyage

- Imputation pour les informations manquantes
- Types de données
- Réduction du nombre de catégories de produits (de 72 à 12)
- Suppression des outliers univariés et multivariés
- Création de nouvelles features
- Assemblage dans une table unique avec pour index l'id client

Feature engineering

- **Intuitions : Indicateurs généraux par client**

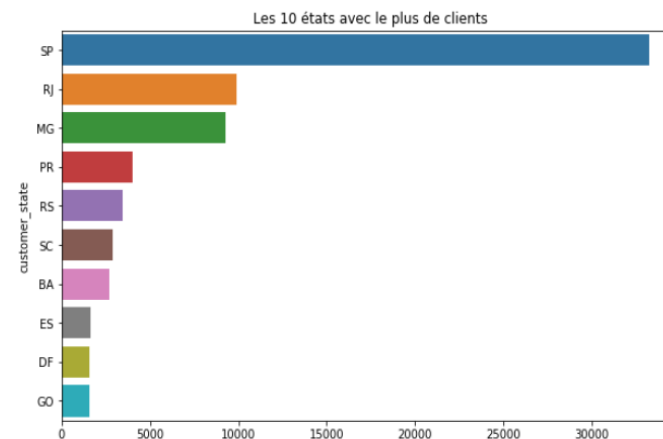
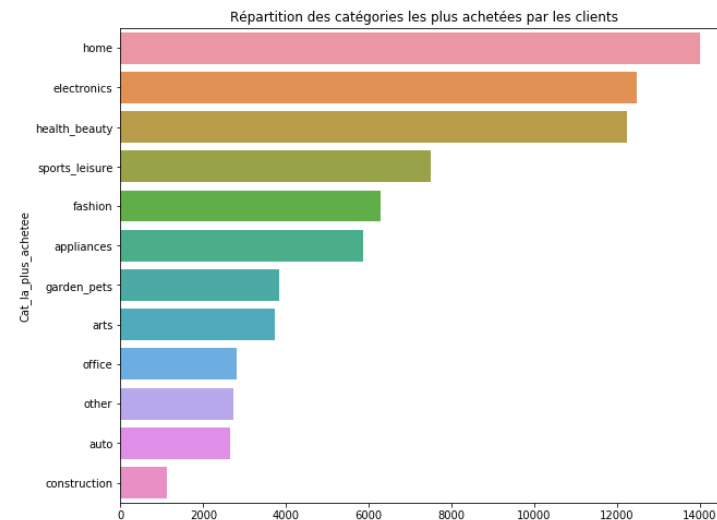
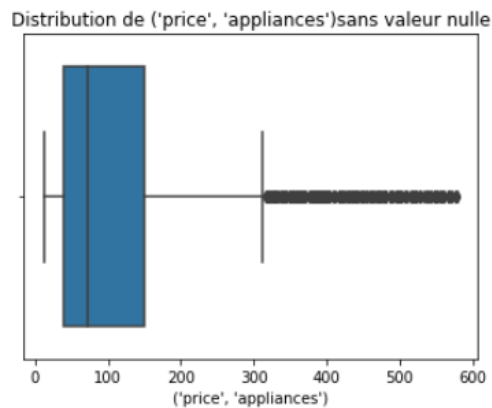
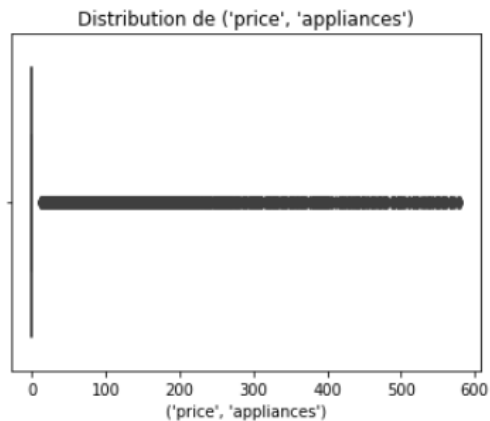
- Nombre d'achats
- Catégorie la plus achetée
- Dépense par catégorie
- Date du dernier achat (Récence)
- Fréquence d'achat
- Panier moyen
- Jour avec le plus de commandes,
- Facilités de paiement
- Moyen de paiement
- Note moyenne
- Etc.

Jeu final

- 75000 lignes
- 31 colonnes

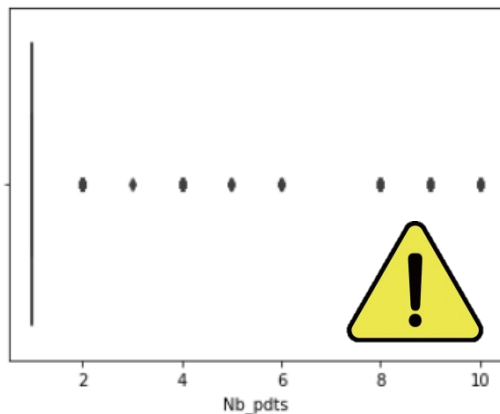
- Log transformation pour certaines features créées

Exploration

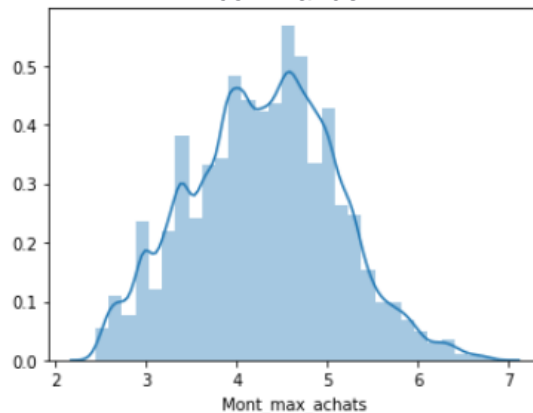


Exploration

Nombre de produits achetés par client



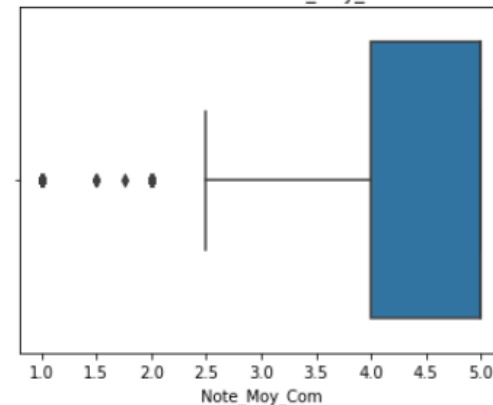
Distribution du montant maximum dépensé par commande



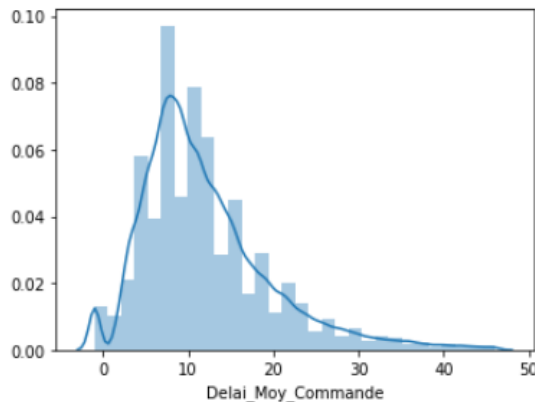
Autres informations:

- 2 années d'historique
- 100 000 clients

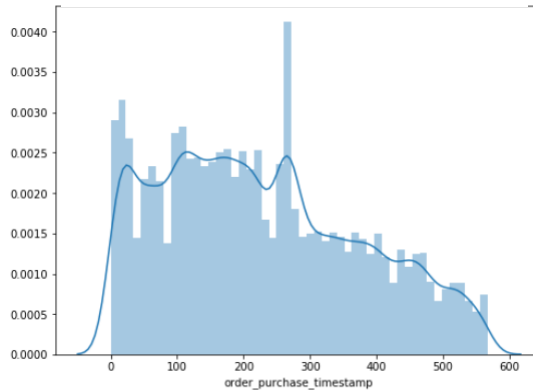
Distribution de Note_Moy_Com



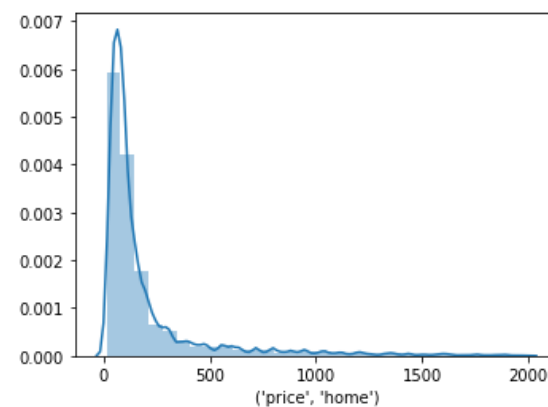
Délai moyen de commande en jours



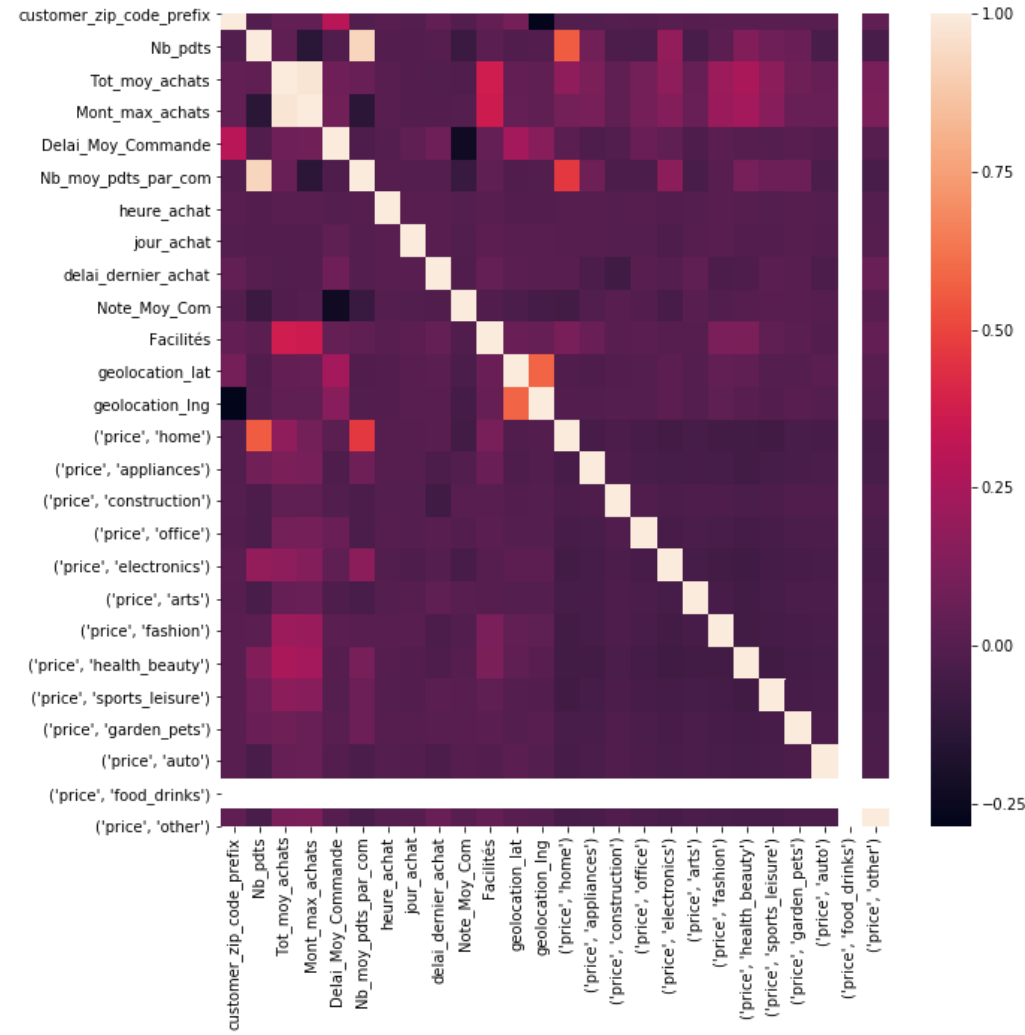
Nombre de jours écoulés depuis la dernière commande



Distribution des dépenses dans la catégorie « home »



Exploration : Corrélations



• Forte corrélation entre :

- Total moyen achats et montant max achats
- Nombre moyens de produits par commande et nombre de produits achetés
- Nombre de produits achetés et dépense dans la catégorie home
- Nombre moyen de produits achetés et dépense dans la catégorie home
- Latitude et longitude

⇒ Non pertinence de certaines features dues au jeu de données comprenant plus de 90 % de clients avec un seul achat

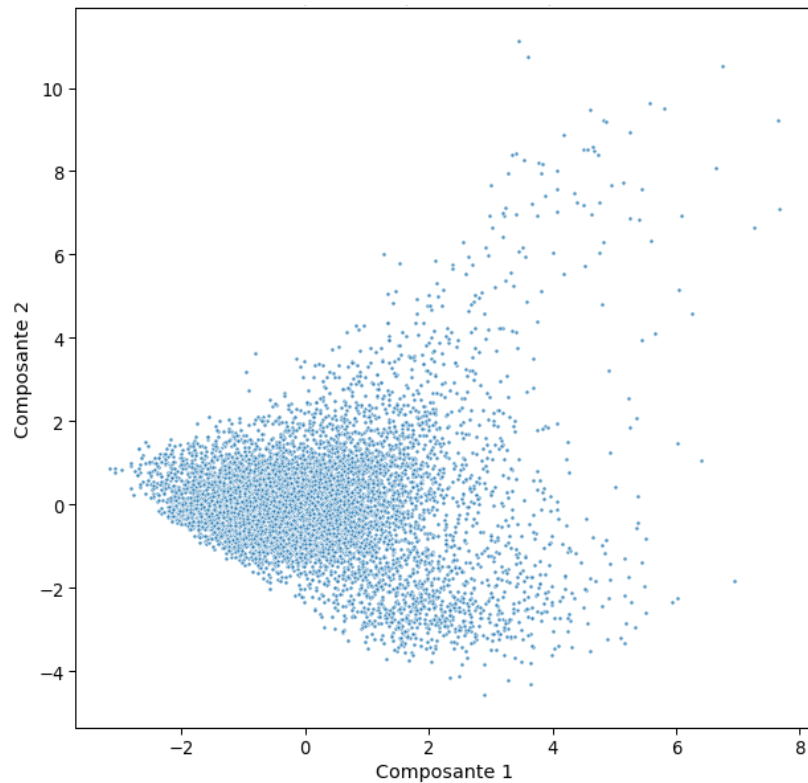
Préparation des données : Finalisation

- **Suppression des features non adaptées:**
 - Montant maximum d'une commande
 - Nombre moyen de produits par commande
- **Préparation**
 - One hot encoder (moyens de paiement)
 - StandardScaler
- **Réduction de dimension par ACP**
 - Réduction à 19 features avec 94 % de variance

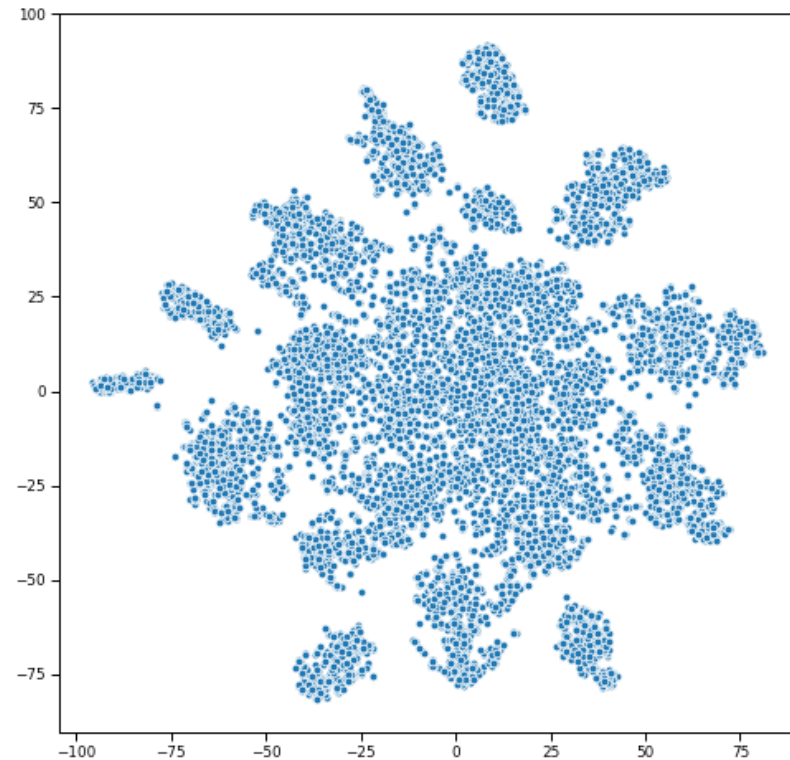
III – PISTES DE MODÉLISATIONS

Intuition : Visualisation

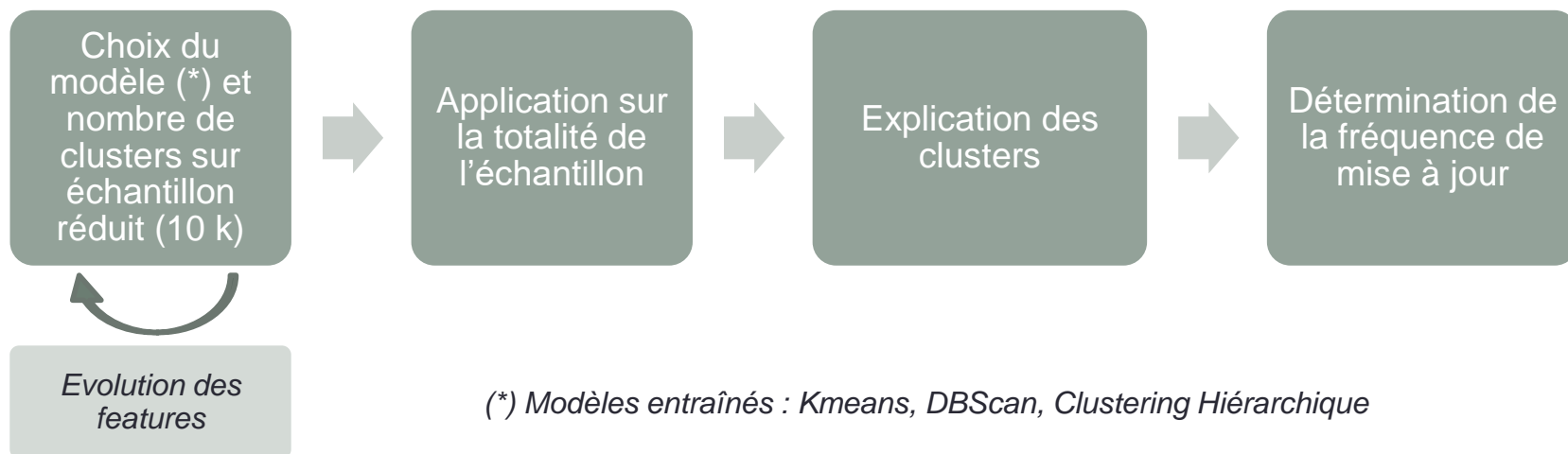
Projection des données sur les 2 premières composantes de l'ACP (20 % de variance)



Représentation des données via t-SNE

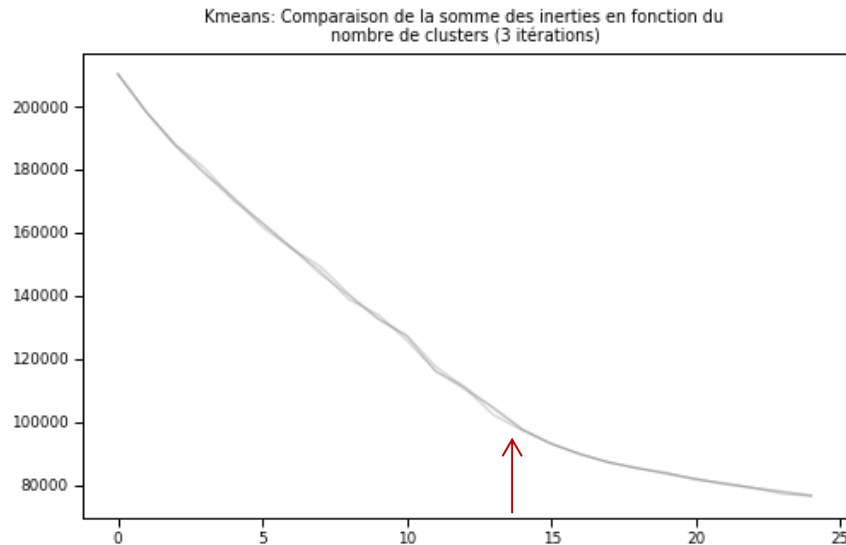


Processus de clustering



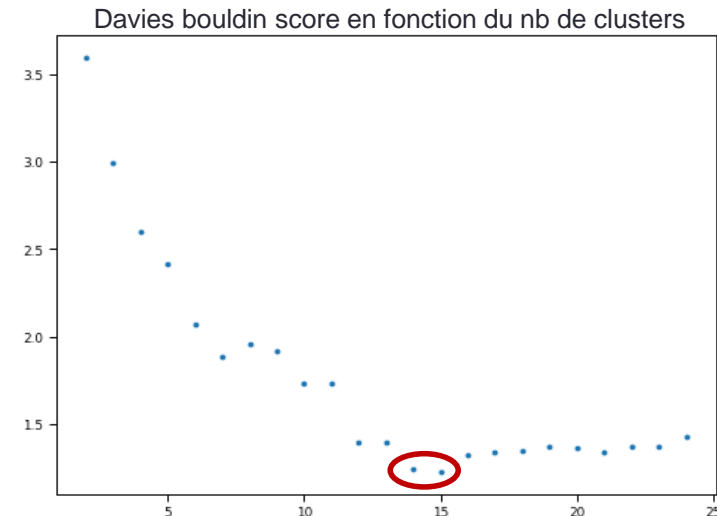
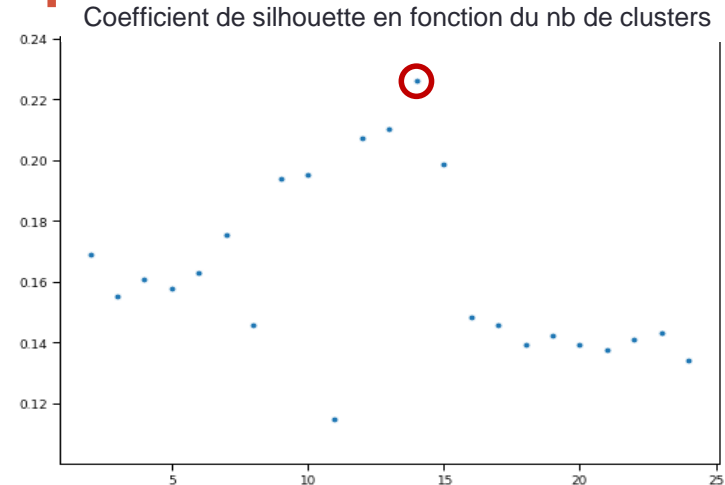
Kmeans : détermination optimum du nombre de clusters

- Entraînement de modèles avec 1 à 25 clusters

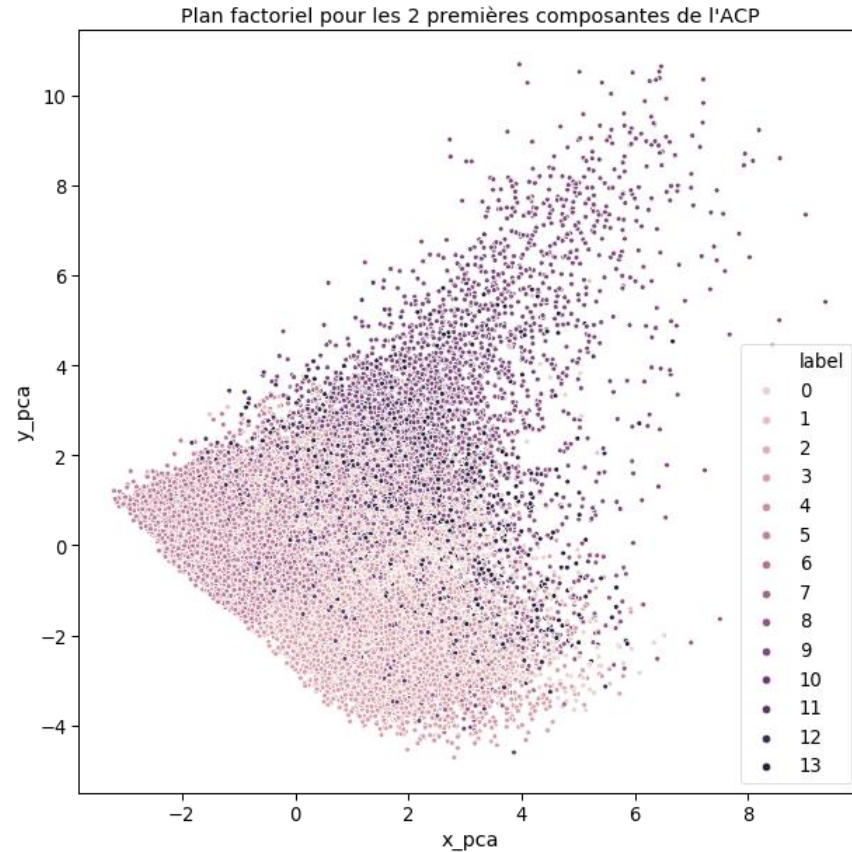
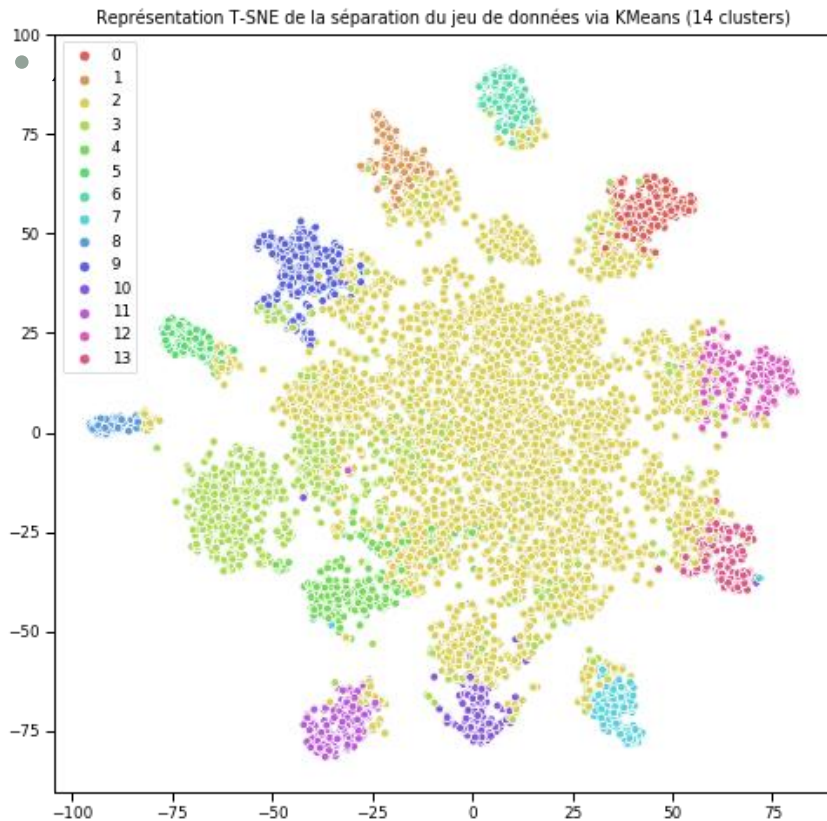


« Méthode du coude »

Optimum retenu : 14 clusters

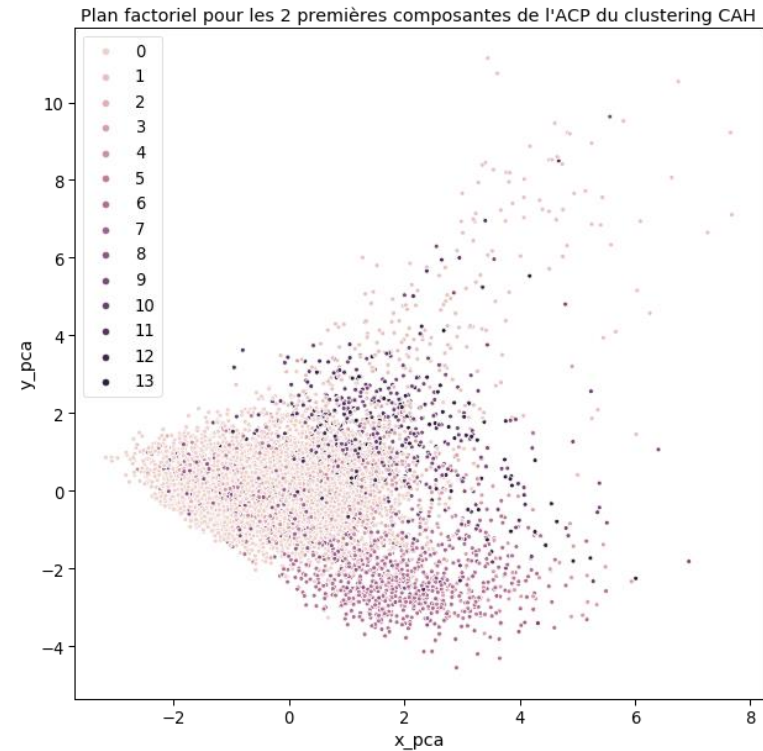
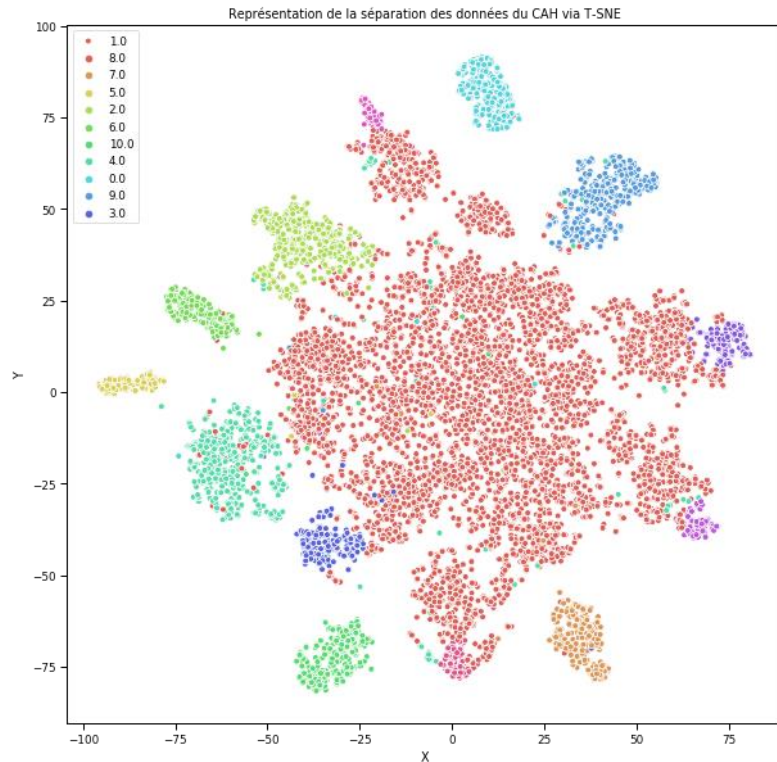


Kmeans : représentation graphique



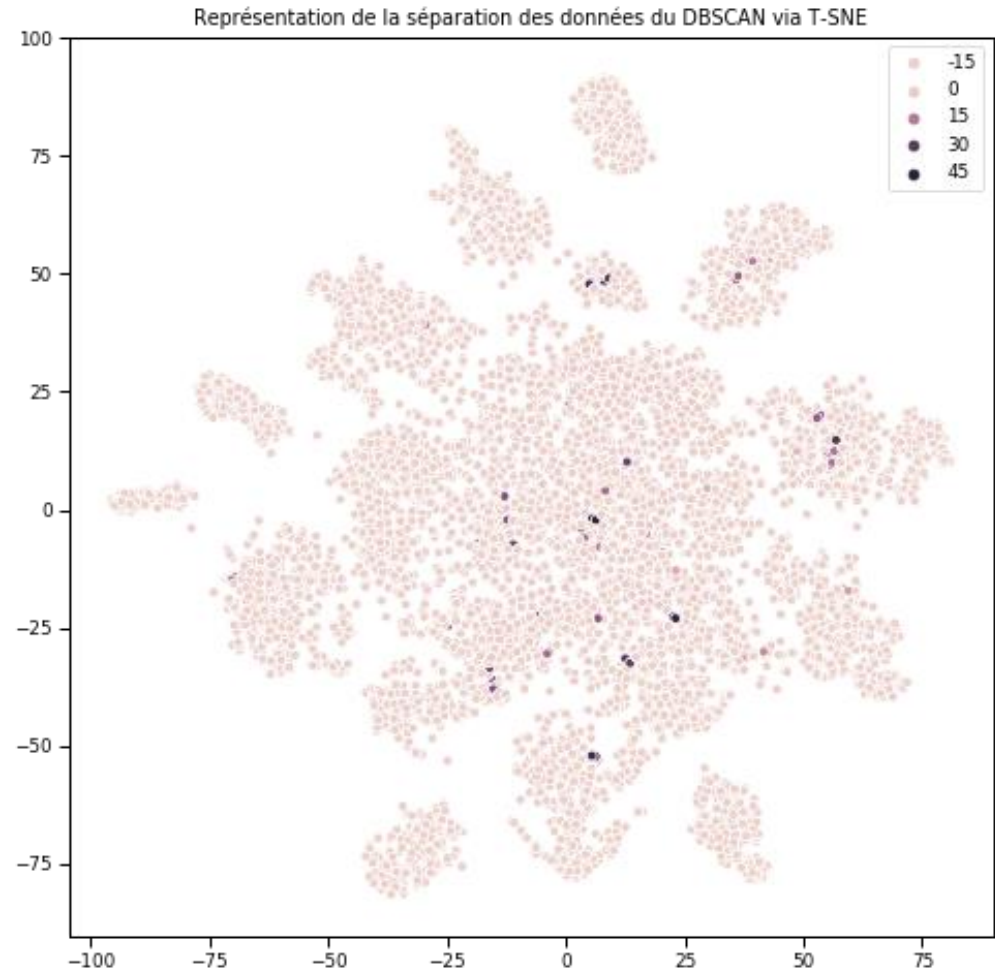
Clustering hiérarchique

- Avec 14 clusters



DBScan

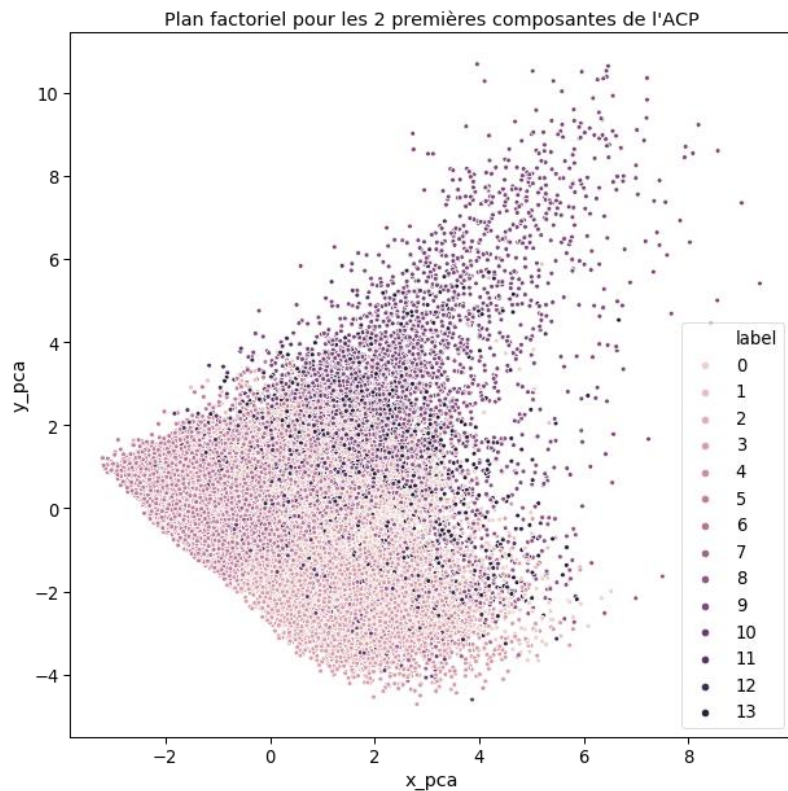
- Exemple ci-contre:
 - Epsilon = 1
 - Min_samples = 5



IV – PRÉSENTATION DU MODÈLE FINAL

ainsi que des améliorations effectuées.

Kmeans sur intégralité de l'échantillon



- Stabilité du silhouette score
 - ...
 - 13 clusters : 0.229
 - **14 clusters : 0.238**
 - 15 clusters : 0.149
 - ...

Clusters identifiés et actions

Catégorie	Catégorie la plus achetée	Autre Caractéristique et action potentielle	Nombre de clients	Note moyenne	Montant moyen dépensé par commande (log)	Montant dépensé par article par client (log)	Ancienneté moyenne dernière transaction (jours)
5	NA	ont le moins recours aux facilités de paiement	43412	4,3	4,0	71	238
3	NA	les plus longs délais de traitement de commande plus mauvaises notes en moyenne	8659	3,4	4,2	88	254
0	fashion		2773	4,1	5,2	243	218
8	health / beauty		2703	4,2	5,5	369	236
1	sport leisure		2492	4,2	5,1	256	246
2	arts		2185	4,2	4,6	113	266
4	office		2078	4,1	4,8	128	244
9	home	catégorie qui achète le plus de produits catégorie qui dépense le plus	1895	3,7	4,9	775	238
11	appliances		1877	4,1	5,1	247	214
7	other	Transactions les plus anciennes	1856	4,2	4,9	143	295
12	garden pets		1824	4,2	4,8	175	245
13	electronics	2e catégorie qui dépense le plus	1433	3,8	5,4	477	253
10	auto		1327	4,2	4,8	125	220
6	construction	commandes les plus récentes	717	4,2	4,6	110	153

Actions :

- Ciblage de clients par catégorie
- Ciblage de services pour certaines catégories (facilités de paiement, livraison express, ...)

Clusters identifiés : améliorations?

- Subdivision du cluster 5 : non concluant (silhouette score : 0.11)
- Suppression de certaines features : pas d'amélioration du clustering constatée
- Proposition de clustering « manuel »:
 - clients insatisfaits (note moyenne de 1/5)
 - clients avec un très long délai de livraison (> 1 mois)
 - 4600 clients qui achètent le plus et sous clustering par application d'un kmeans :
 - Identification des meilleurs clients par catégorie
 - Identification des catégories complémentaires pour cibler la publicité

Contrat de maintenance

- Identification de la période de maintenance:
 - Réduction du jeu de données sur la dimension « durée » (exemple : 3 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
 - Un cluster avec 50 % des clients, peu intelligibles
- 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

MERCI DE VOTRE
ATTENTION
