



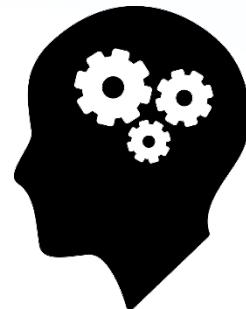
Projet 5 :

Segmentez des clients d'un site e-commerce

Lecerf Defer Amandine

Compétences évaluées

- Adapter les hyperparamètres d'un algorithme non supervisé afin de l'améliorer
- Évaluer les performances d'un modèle d'apprentissage non supervisé
- Transformer les variables pertinentes d'un modèle d'apprentissage non supervisé
- Mettre en place le modèle d'apprentissage non supervisé adapté au problème métier



Plan

- I. Problématique
 - a. Contexte
 - b. Données disponibles
 - c. Interprétation
- II. Préparation du jeu de données
 - a. Nettoyage
 - b. Exploration
- III. Piste de modélisation
 - a. Comparaison d'algorithmes
 - b. Analyse temporelle
- IV. Modèle final sélectionné
- V. Analyse des clusters





Problématique

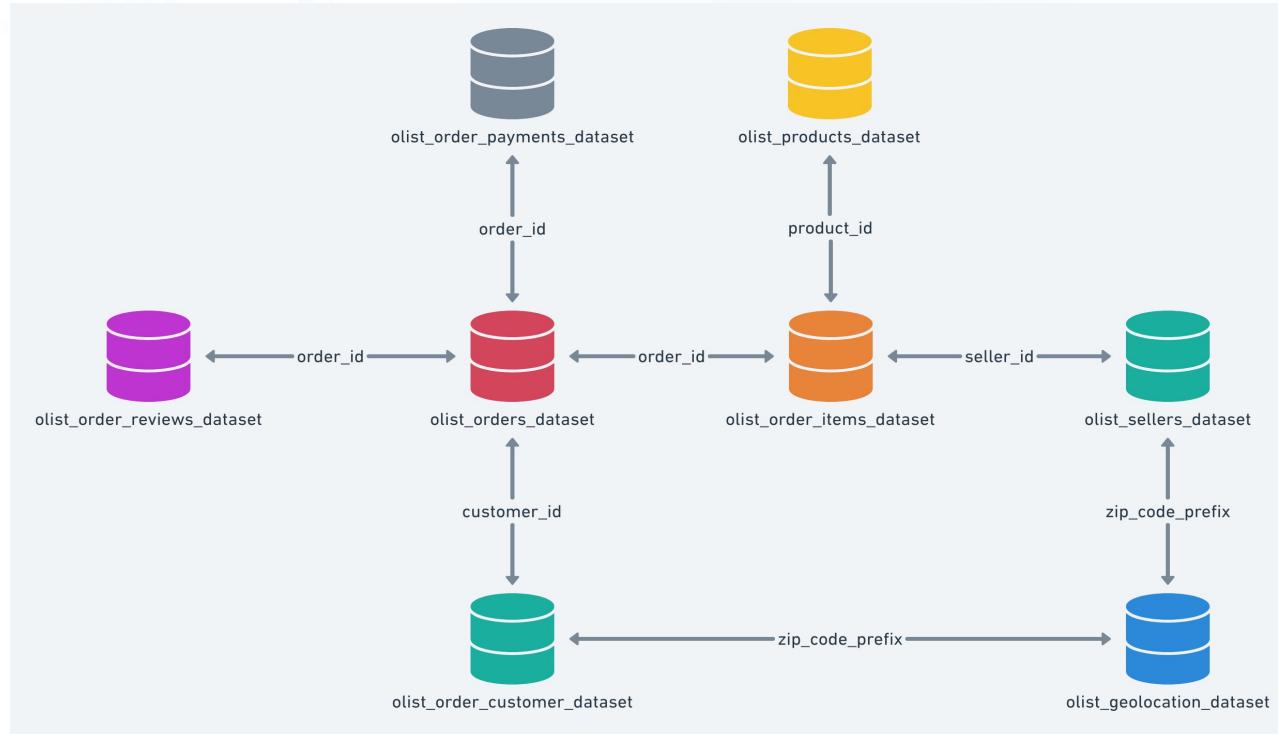
Contexte

- Consultant pour Olist -> une solution de vente sur les marketplaces en ligne (boutiques générale par vendeur)
- Objectifs :
 - Segmentation des clients pour les équipes marketing de *Olist* pour les campagnes de communication
 - Analyser les différents types d'utilisateurs
 - Proposition du contrat de maintenance
 - Fournir une description actionnable de la segmentation

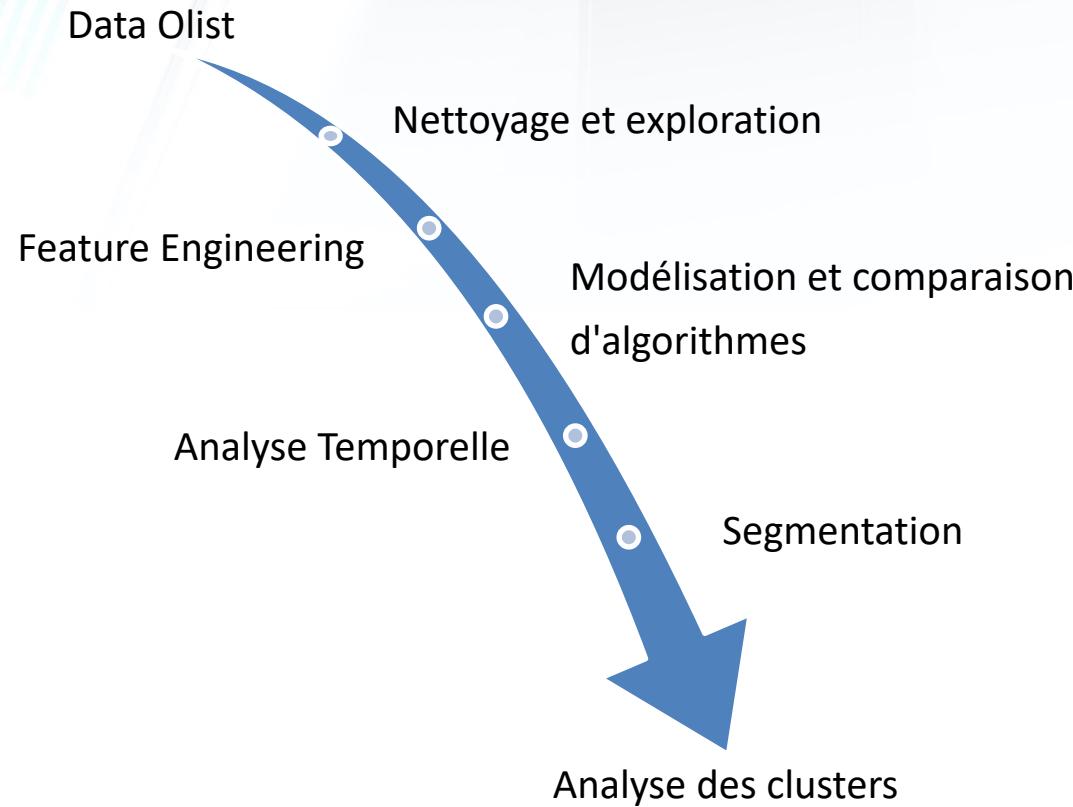
olist

Données Disponibles

- Olist -> base de données anonymisée



Interprétation





Préparation du jeu de données

Nettoyage

- Imputation des données manquantes dans chaque table
- Types de données + regroupements de données en intervalle
- Suppression d'outliers
- Réduction du nombre de produits de 72 à 12
- Assemblage des tables dans un seul dataset unique

Join_dataset | **119 151 lignes** | **48 colonnes**

- Création de nouvelles variables
- Suppression des doublons grâce à l'identifiant unique de chaque client

clean_dataset | **96 096 lignes** | **68 colonnes**

Feature Engineering

- Clients :
 - Id (index), région, distance qui le sépare du vendeur, nombre d'achats, paiement le plus utilisé, catégorie la plus achetée, prix moyen par catégorie d'article, ...
- Achats :
 - Nombre de jours depuis la dernière commande
 - Nombre de jours entre deux commandes, ...
- Commandes :
 - Prix moyen, Volume moyen des produits par commandes, Prix moyens des produits, Frais de livraison moyens, ...



final_dataset

96 096 lignes

28 colonnes

Exploration Univariée

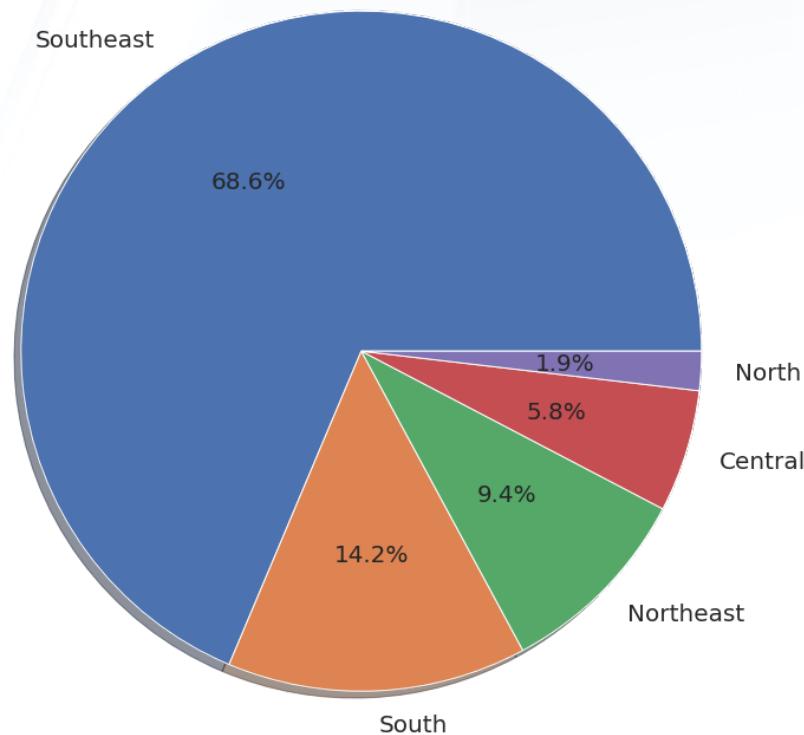


Exploration Univariée

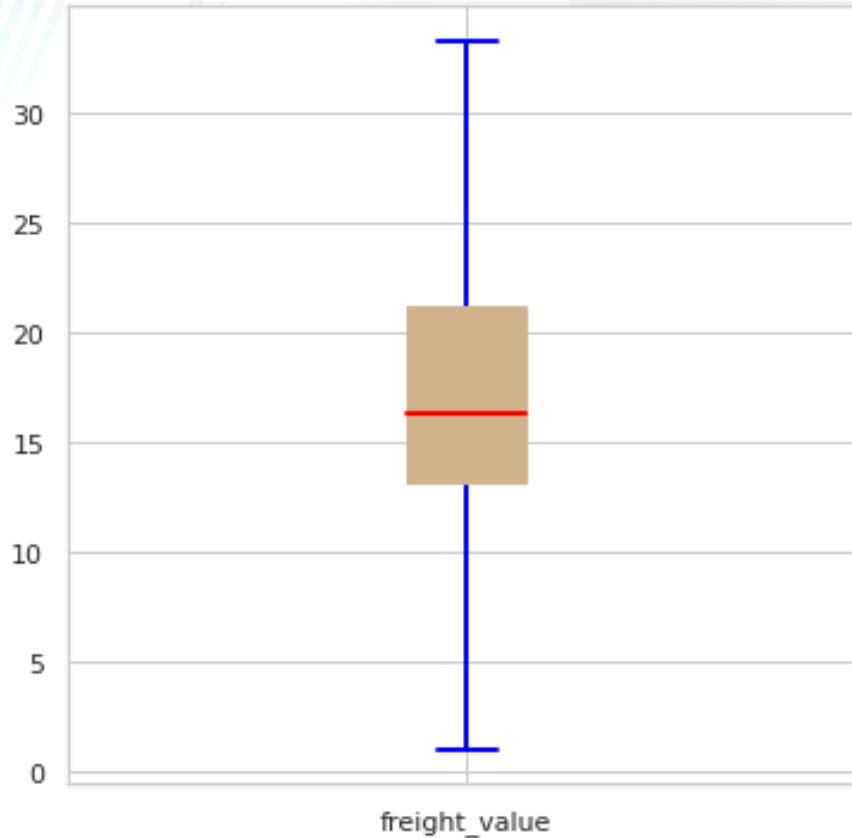


Exploration Univariée

Représentation de la variable customer_region

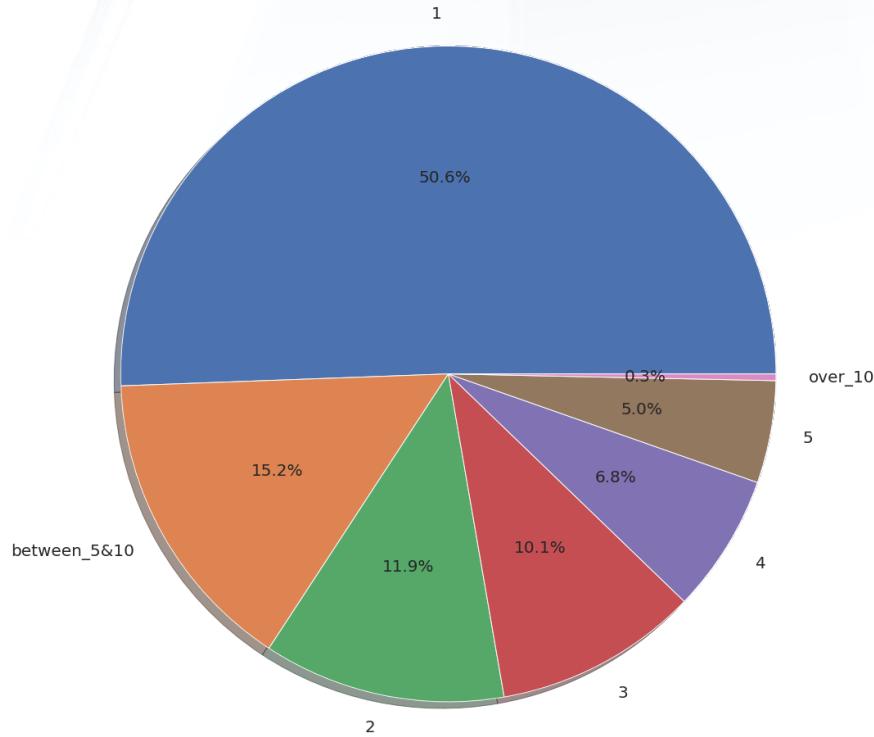


Exploration Univariée

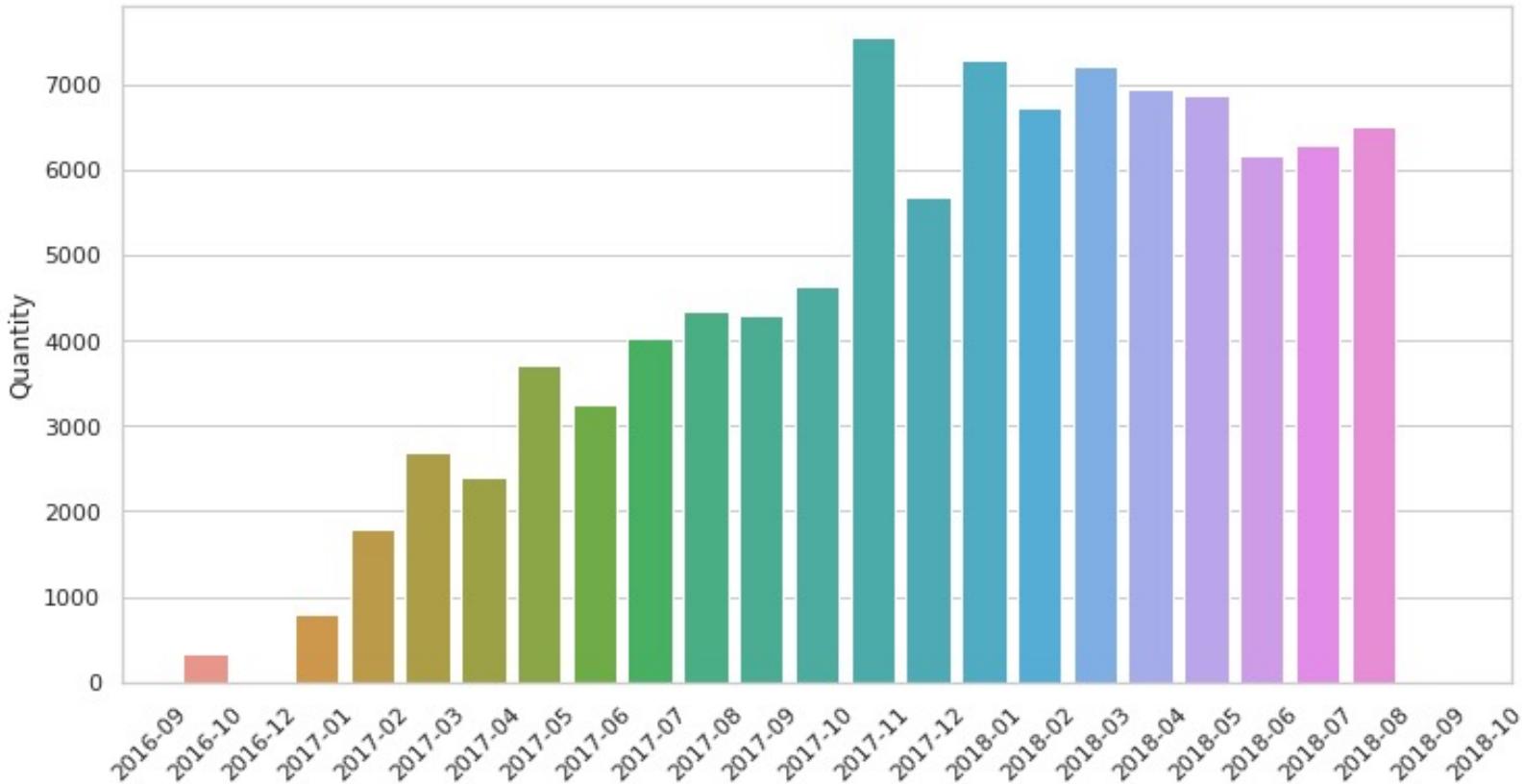


Exploration Univariée

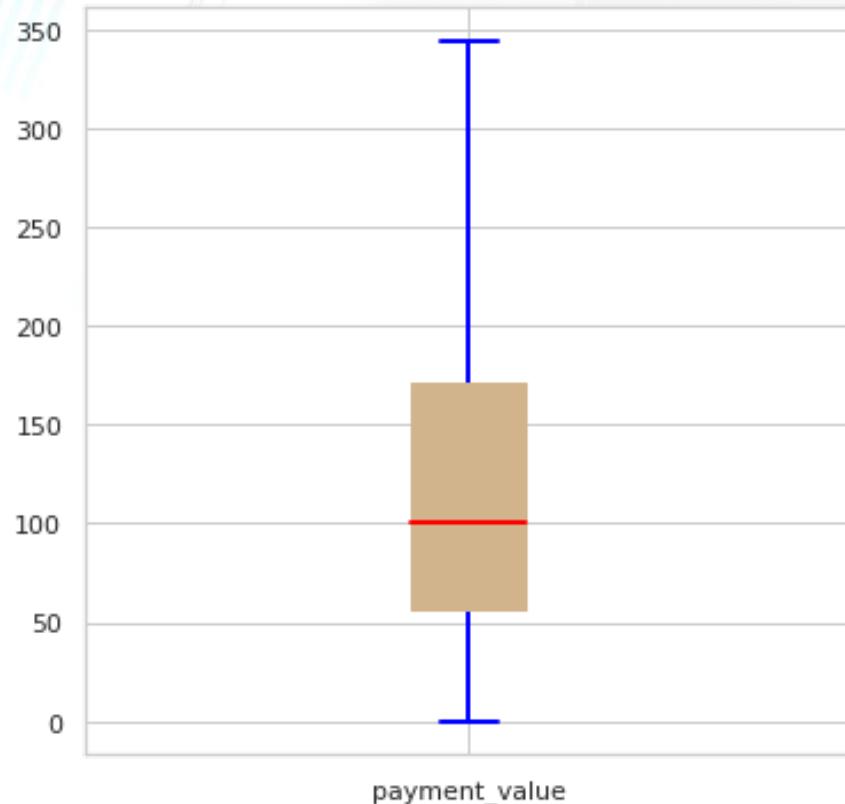
Représentation de la variable payment_installments



Exploration Univariée



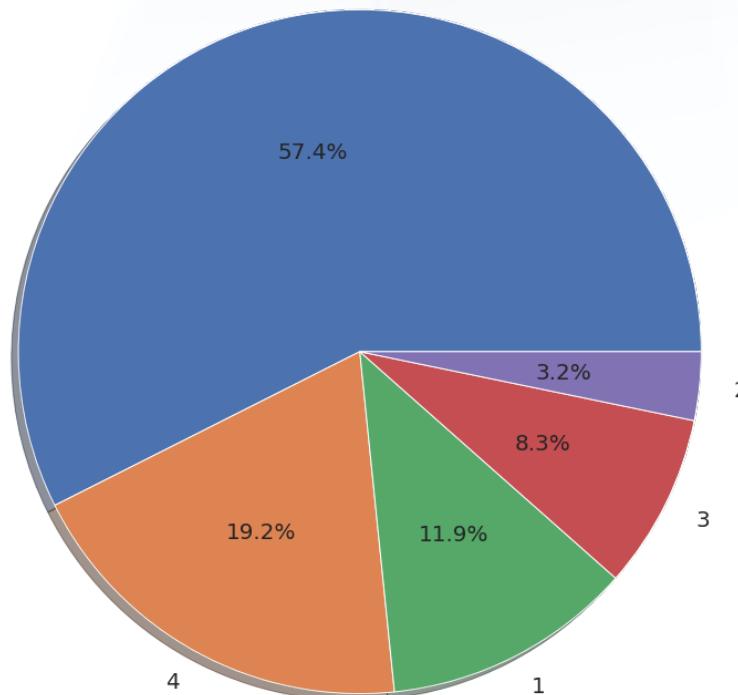
Exploration Univariée



75% : 170 R\$
50% : 100 R\$
25% : 57 R\$

Exploration Univariée

Représentation de la variable review_score



Exploration Multivariée

Date_last_order et date_first_order : corrélation non pertinente
=> la majorité des clients n'a fait qu'une seule commande

	number_order	distance_between_seller_customer	date_last_order	mean_days_between_orders	mean_item_per_order	mean_volume_item_ordered	date_first_order	mean_product_price	mean_freight_value	mean_order_value	mean_review_score	mean_payment_sequential	mean_payment_installment	price_for_Office_equipment_furniture	price_for_clothing_accessories	price_for_electronic设备	price_for_food_drink	price_for_furniture_decoration_home	price_for_garden_pets	price_for_home_appliances	price_for_hygiene_health_wellness	price_for_leisure_homemade	price_for_other	price_for_repairs_constructions																				
number_order	-0.01																																											
distance_between_seller_customer		-0.0017	-0.035																																									
date_last_order			-0.0017																																									
mean_days_between_orders				0.61	-0.0082	0.023																																						
mean_item_per_order					0.014	0.0042	0.015	0.009																																				
mean_volume_item_ordered						-0.0059	-0.025	-0.065	-0.0019	-0.0011																																		
date_first_order							-0.039	-0.035	1	-0.038	0.014	-0.065																																
mean_product_price								-0.03	0.09	0.013	-0.0082	-0.097	0.27	0.013																														
mean_freight_value									-0.011	0.51	0.066	-0.0067	-0.01	0.44	0.067	0.33																												
mean_order_value										-0.028	0.13	0.017	-0.0087	0.19	0.3	0.018	0.89	0.4																										
mean_review_score											0.0011	-0.044	0.021	0.0068	-0.074	-0.0089	0.021	0.017	-0.012	-0.051																								
mean_payment_sequential												0.02	0.015	-0.016	0.022	-0.0068	-0.0053	-0.017	-0.035	-0.0021	-0.12	-0.0019																						
mean_payment_installment													0.019	0.081	-0.051	0.011	0.063	0.14	-0.052	0.33	0.17	0.35	-0.025	-0.069																				
price_for_Office_equipment_furniture														0.0025	-0.011	-0.0043	0.0055	0.007	0.3	-0.0046	0.072	0.14	0.084	-0.014	-0.0059	0.02																		
price_for_clothing_accessories															0.02	0.023	0.026	0.016	-0.01	-0.052	0.025	0.31	0.016	0.27	-0.0014	0.045	0.12	-0.043																
price_for_electronic设备																0.023	0.064	-0.003	0.017	0.051	-0.12	-0.004	0.2	0.028	0.21	-0.023	0.022	-0.02	-0.053	-0.073														
price_for_food_drink																	-0.001	-0.021	0.028	0.00035	-0.0029	-0.029	0.028	-0.055	-0.029	-0.056	0.0073	-0.0078	-0.038	-0.012	-0.017	-0.02												
price_for_furniture_decoration_home																		0.083	-0.023	-0.023	0.035	0.16	0.19	-0.025	0.15	0.12	0.21	-0.035	0.056	0.15	-0.063	-0.093	-0.11	-0.025										
price_for_garden_pets																		0.014	0.01	-0.013	0.021	0.074	0.06	-0.014	0.0095	0.087	0.041	0.0096	0.018	0.0094	-0.036	-0.051	-0.056	-0.014	-0.072									
price_for_home_appliances																		0.026	-0.026	0.026	0.014	0.04	0.16	0.025	0.079	0.095	0.093	0.0063	0.025	0.054	-0.038	-0.055	-0.065	-0.014	-0.079	-0.043								
price_for_hygiene_health_wellness																		0.025	0.057	0.022	0.023	-0.00058	-0.027	0.021	0.34	0.043	0.31	0.011	0.014	0.13	-0.056	-0.079	-0.094	-0.021	-0.12	-0.063	-0.07							
price_for_leisure_homemade																		0.046	0.015	-0.021	0.035	0.02	0.076	-0.023	0.23	0.06	0.21	0.015	0.03	0.031	-0.055	-0.077	-0.092	-0.02	-0.11	-0.063	-0.067	-0.099						
price_for_other																		0.0035	0.023	-0.054	0.0074	-0.017	0.067	-0.055	0.15	0.046	0.13	-0.0018	0.0075	0.047	-0.036	-0.05	-0.059	-0.013	-0.072	-0.04	-0.045	-0.061	-0.063	-0.04				
price_for_repairs_constructions																		0.014	0.011	0.061	0.014	0.022	0.081	0.06	0.15	0.085	0.15	0.0029	0.014	0.045	-0.035	-0.05	-0.058	-0.014	-0.069	-0.037	-0.044	-0.063	-0.061	-0.04				

Corrélation forte :

- Frais d'expédition : distance avec vendeur, volume articles
- Prix moyen d'un produit : frais d'expédition, montant de la commande, échelonnage de paiement
- Prix moyen de la commande : échelonnage de paiement



Piste de modélisation

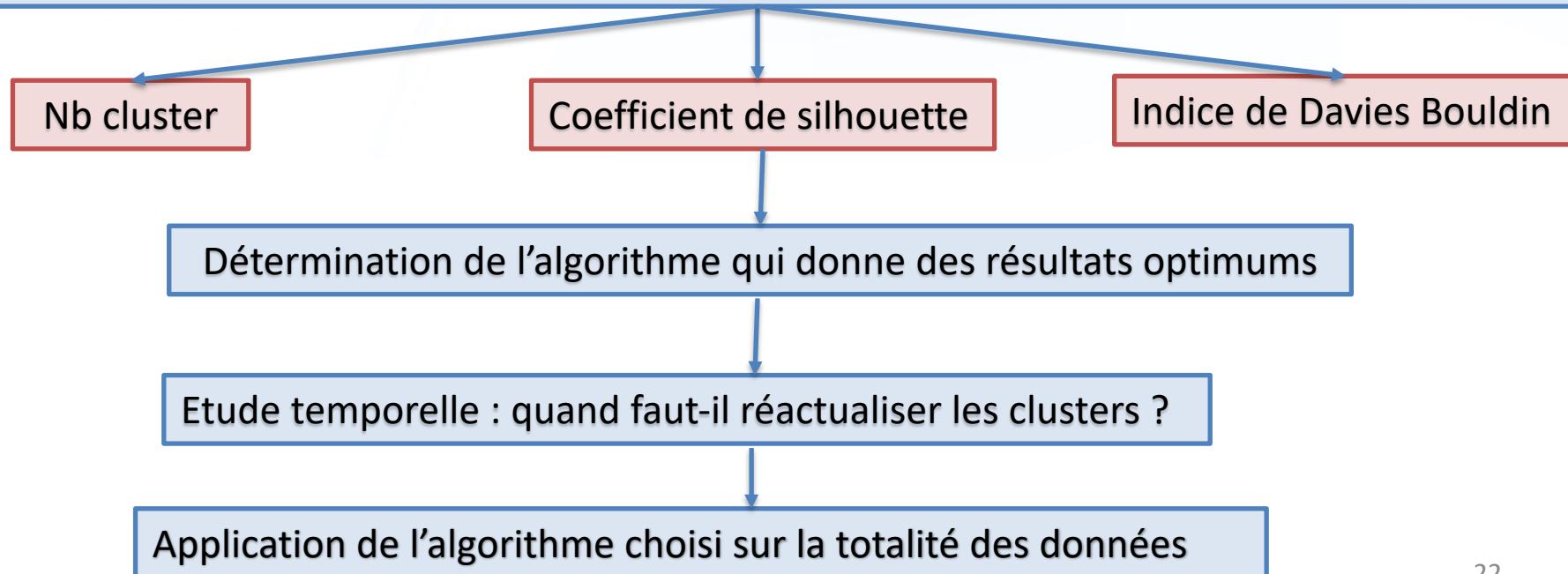
Comparaison d'algorithmes

Data Processing

- Séparation des données :
 - Selon des périodes de 3 mois pour l'étude temporelle = 7 échantillons
- Préparation des 7 échantillons :
 - Normalisation des données numériques
 - OneHotEncoder pour :
 - customer_region
 - most_payment_used
 - category_most_purchased
- Réduction de dimension par ACP :
 - Réduction à 19 features avec 93% de variance expliquée

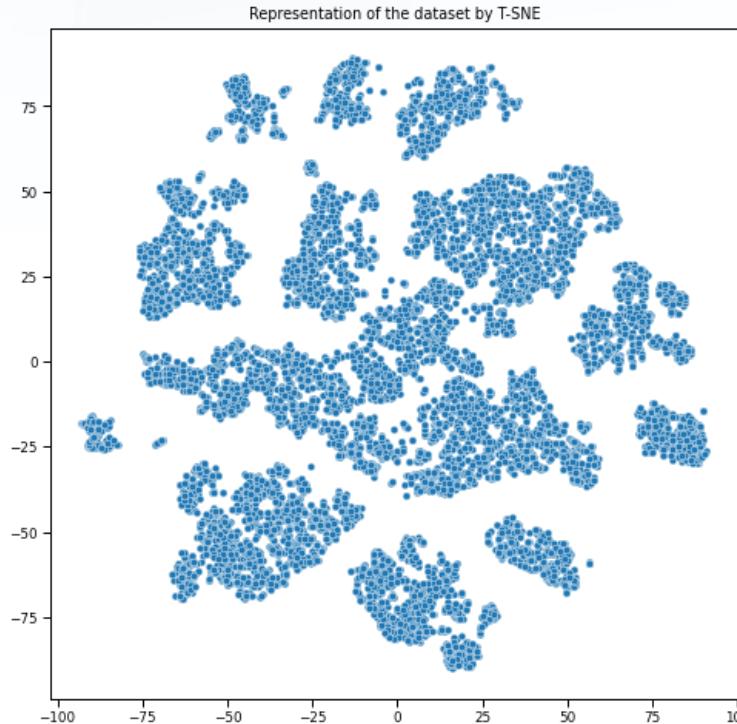
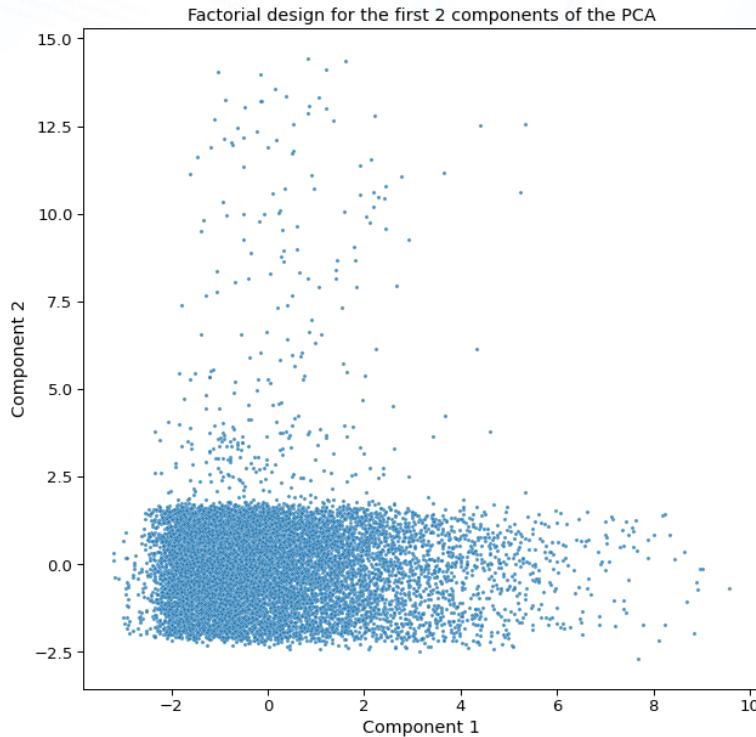
Méthode envisagée

Comparaison d'algorithmes * sur un échantillon de données (6 premiers mois)
* K-means, Clustering hiérarchique, DBScan



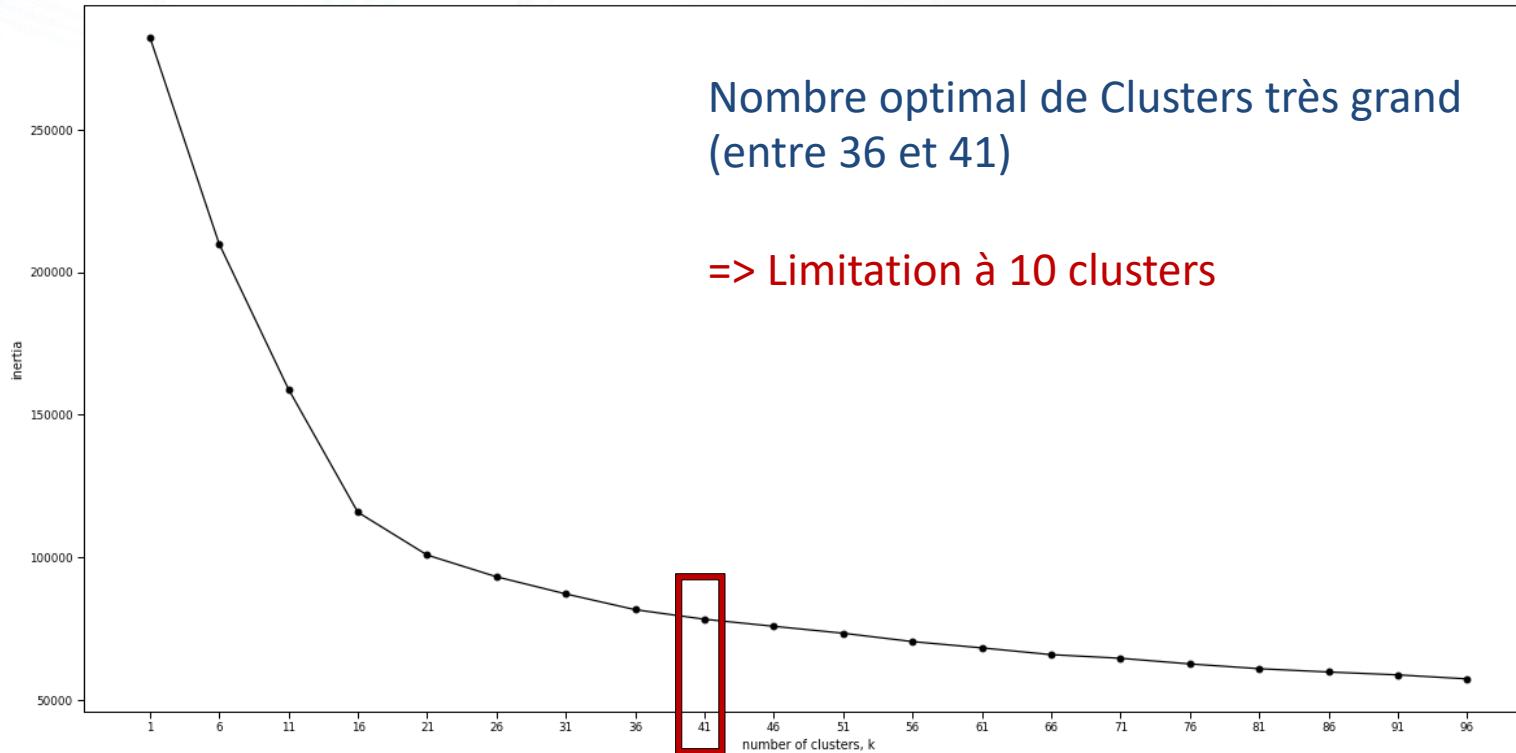
Visualisation

- Visualisation des données sur les 6 premiers mois.



K-means

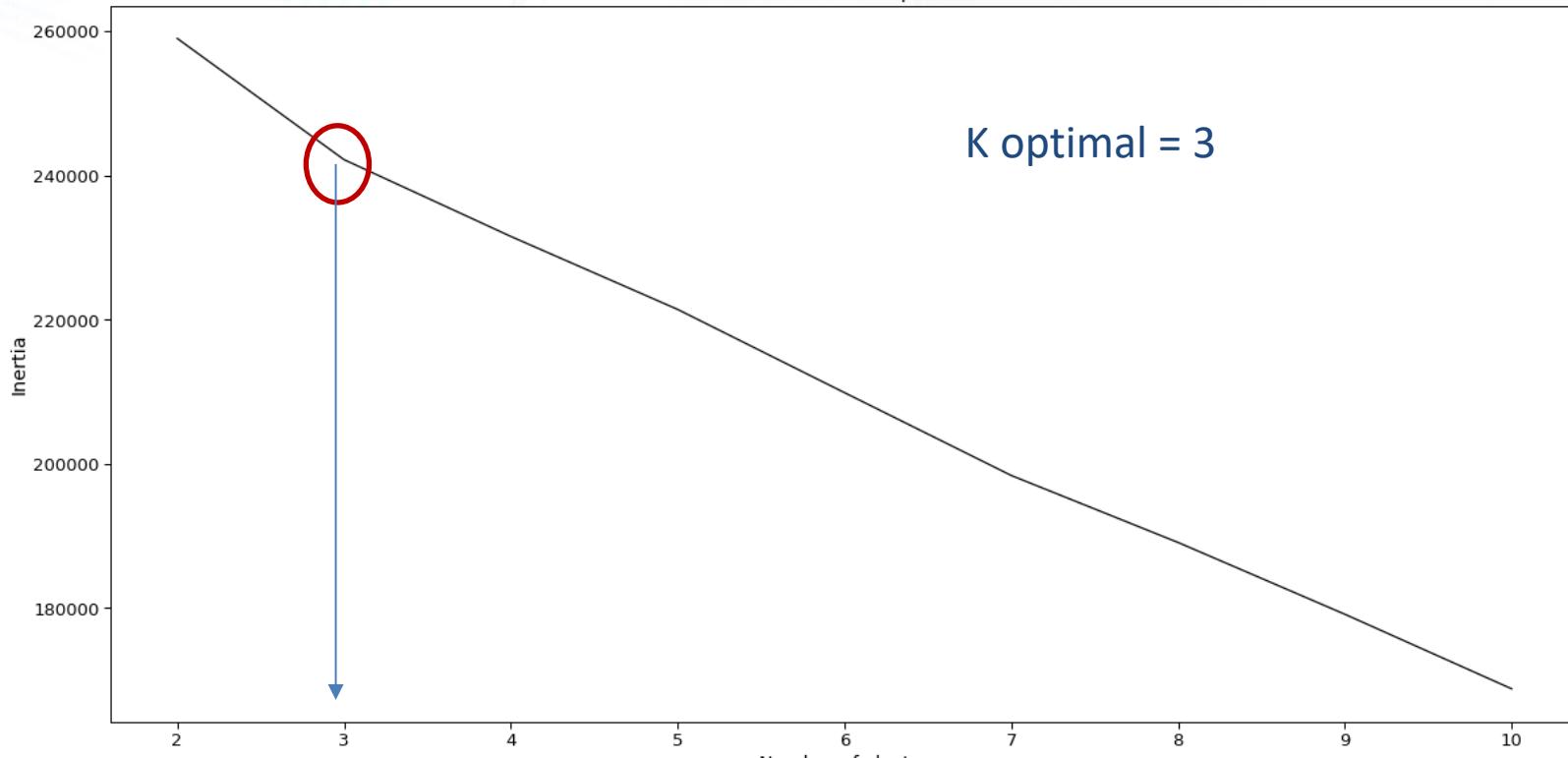
- Détermination du nombre de Clusters : La méthode du coude



K-means

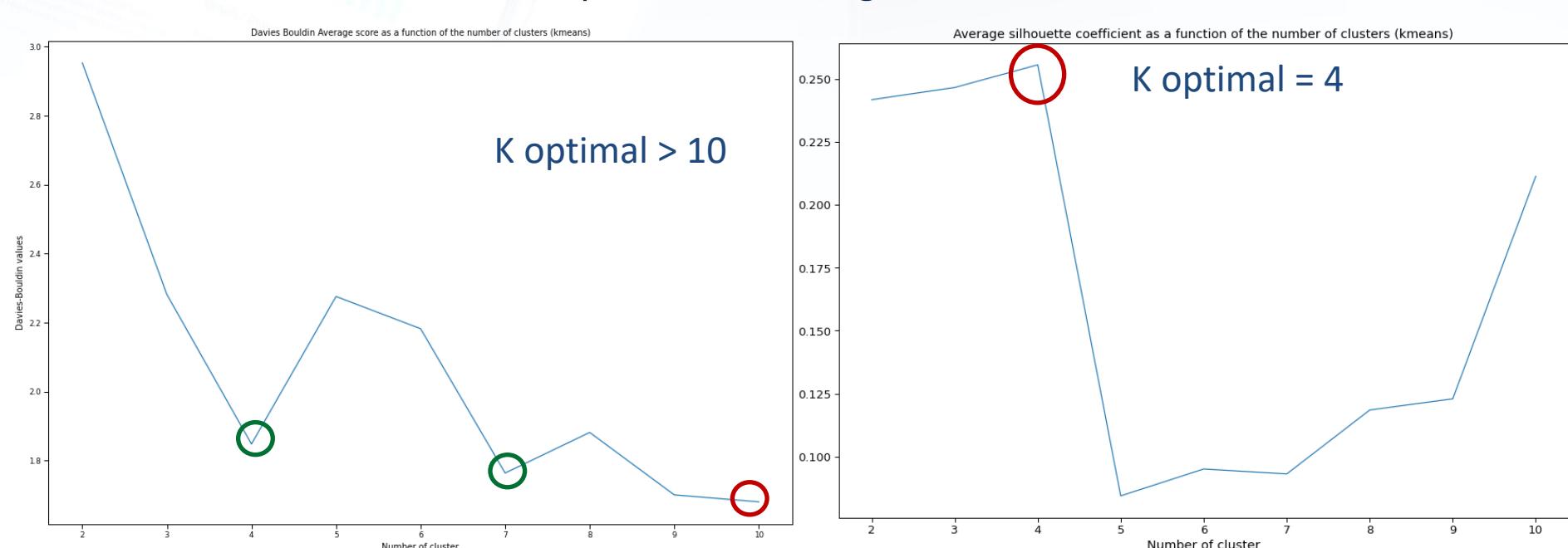
- Détermination du nombre de Clusters : La méthode du coude

Elbow Method For Optimal k



K-means

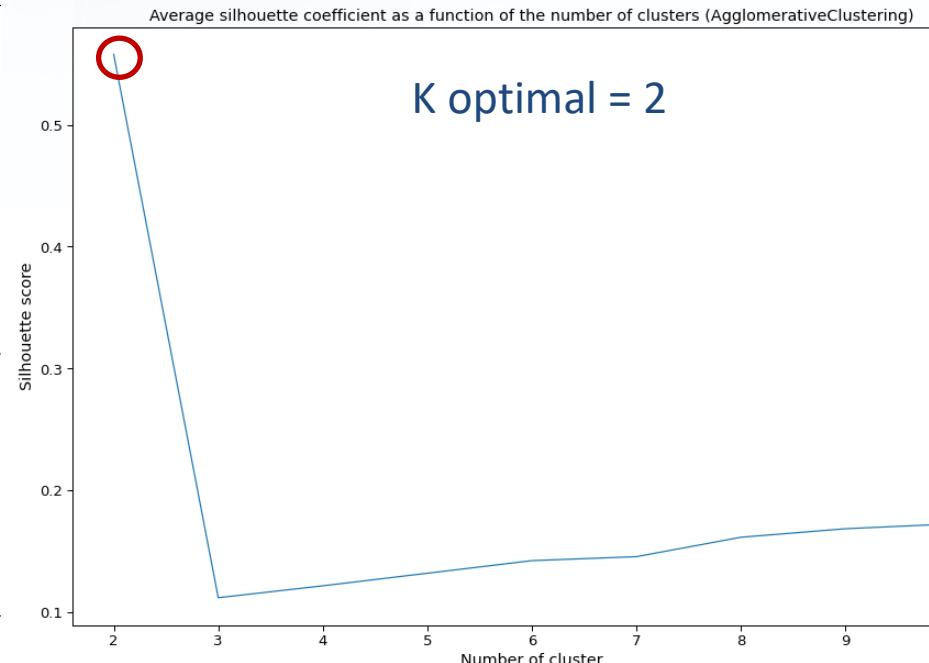
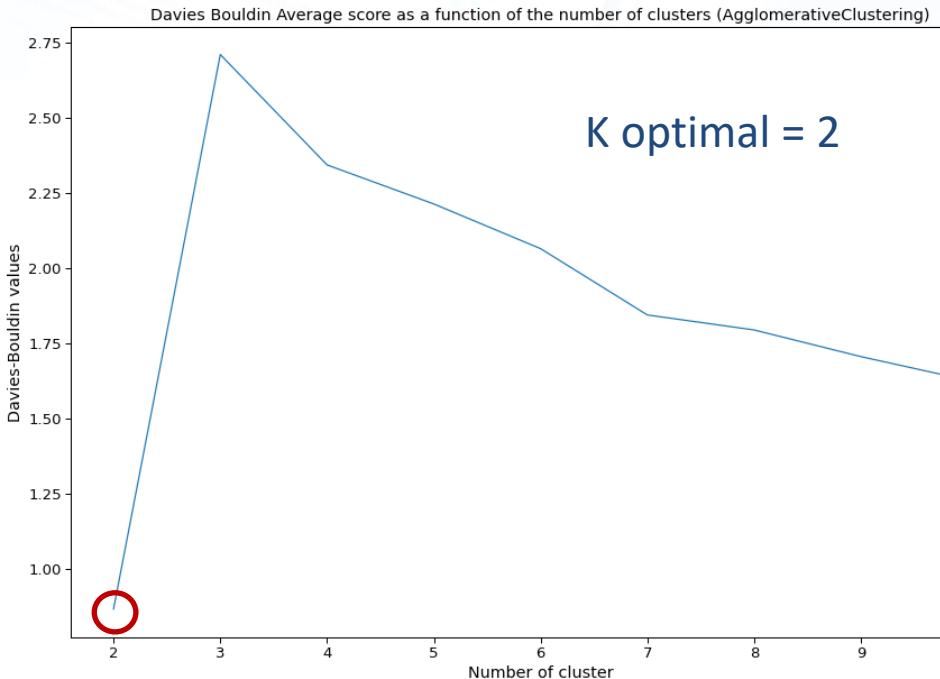
- Nombre de Clusters et qualité clustering



Ici : K optimal = 4

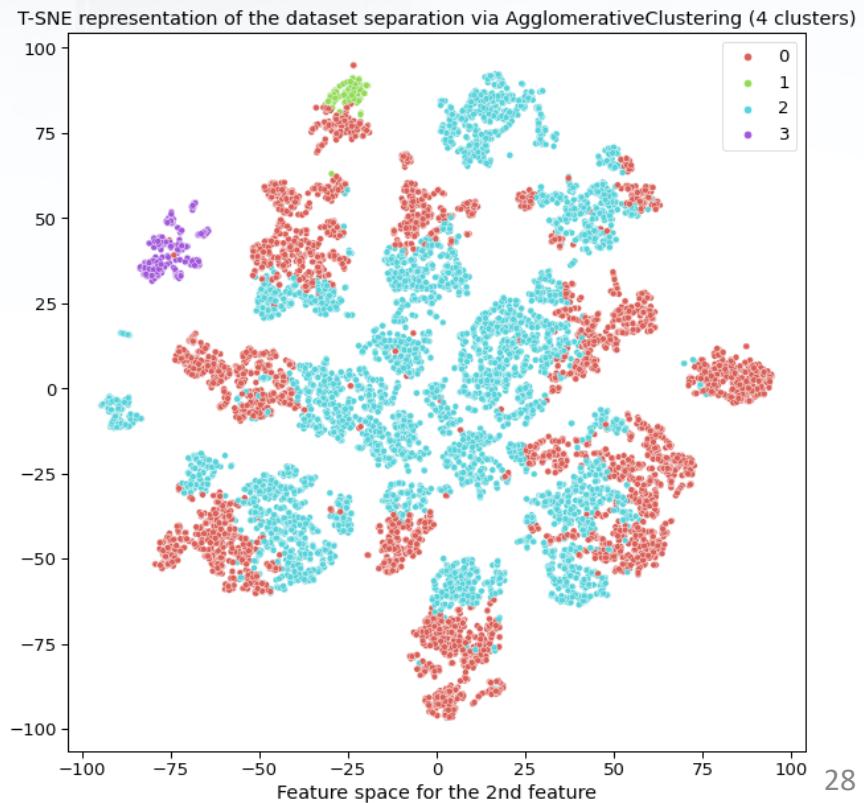
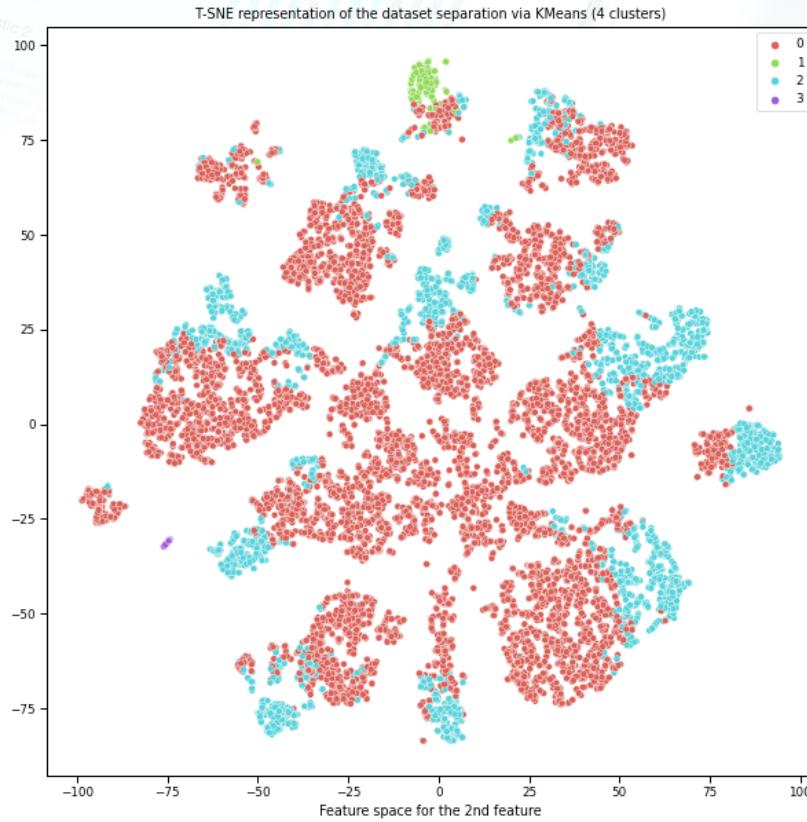
Clustering Hiérarchique

- Nombre de Clusters et qualité clustering



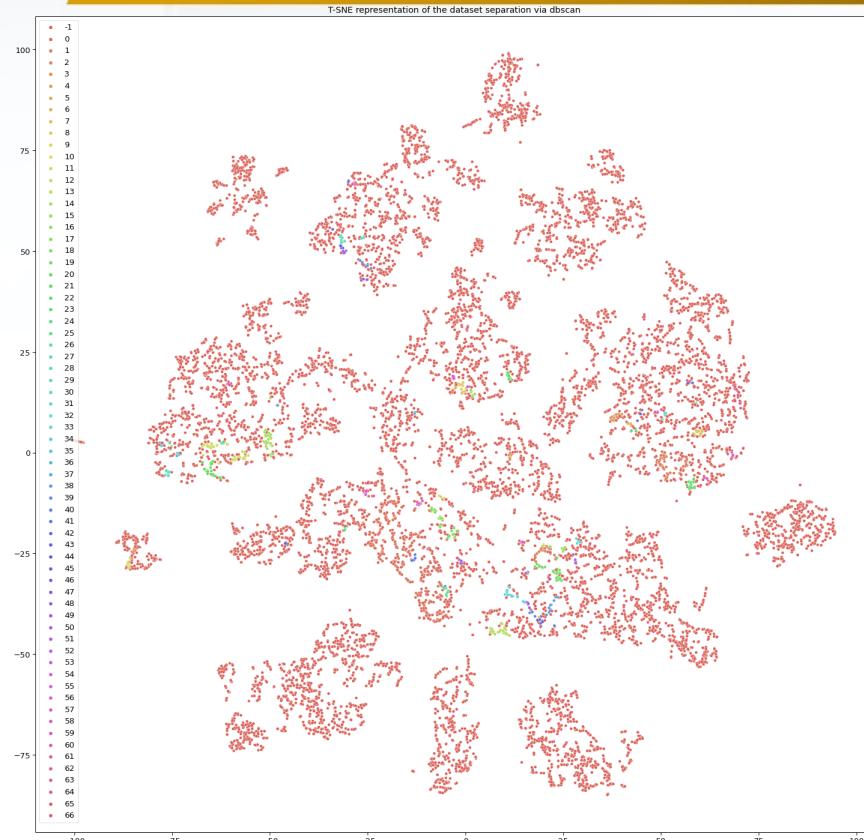
Ici : K optimal = 2 mais comparaison avec k=4

Comparaison visuelle



DBScan

- Exemple d'exécution :
 - Epsilon = 0.5
 - Min_samples = 5
- Choix de l'intervalle de recherche de clusters = NON
- Nombre de clusters optimal : 68



Comparaison

	Modele	nb_cluster	min_Davies_Bouldin	max_silhouette
0	Kmeans	4	1.679764	0.255669
1	hierarchical clustering	4	2.342708	0.121591
2	DBscan	68	1.885589	-0.441045

}

Nb_cluster = 4

DB[1] > DB[0]

Sil[1] < Sil[0]

Clustering hiérarchique :
NON

Nb_cluster = 68

Trop de Clusters pour l'équipe
marketing

Trop de Clusters Unitaire

DBScan : NON

Algorithme choisi pour la suite des
Analyses:

- K-means
- Nb_Cluster = 4

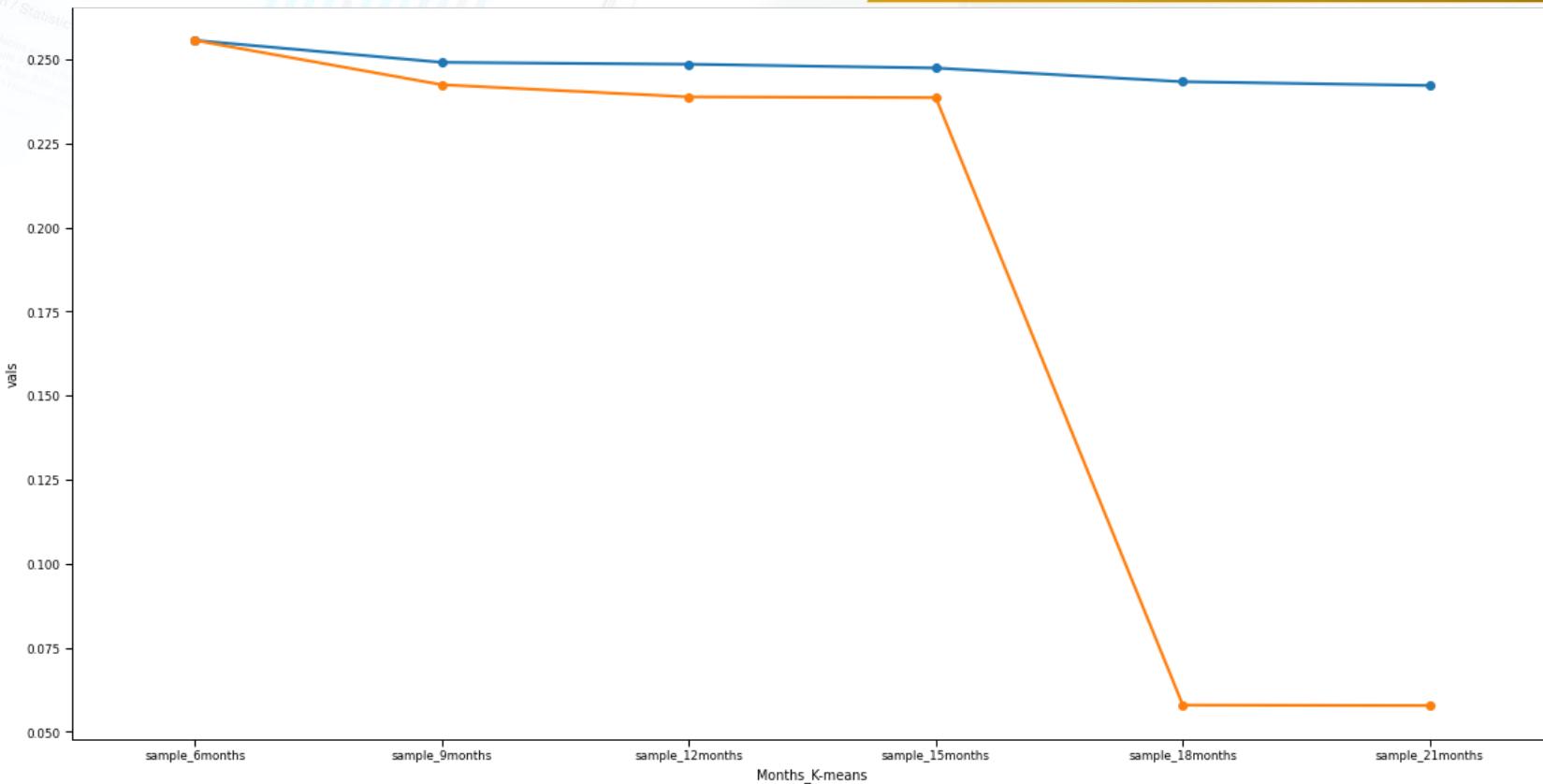
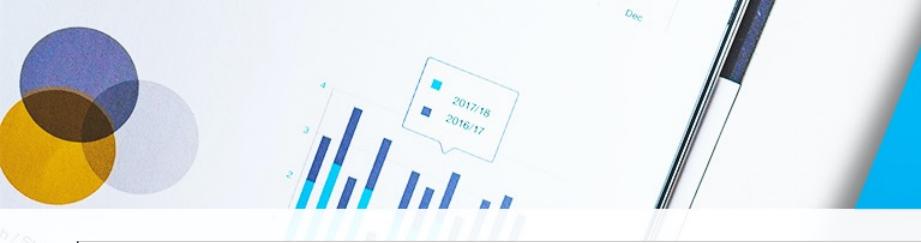
Piste de modélisation

Analyse temporelle

Analyse Temporelle

- Stabilité des Clusters après ajout de données
- Calcul Clusters vs Prediction Clusters :
 - Calcul clusters coûteux
 - Prédiction clusters : fiable ?
 - Au bout de combien de temps la prédiction n'est plus assez fiable ce qui entraîne le recalcul des clusters ?
 - Recherche après ajout de données tous les 3 mois

Analyse Temporelle



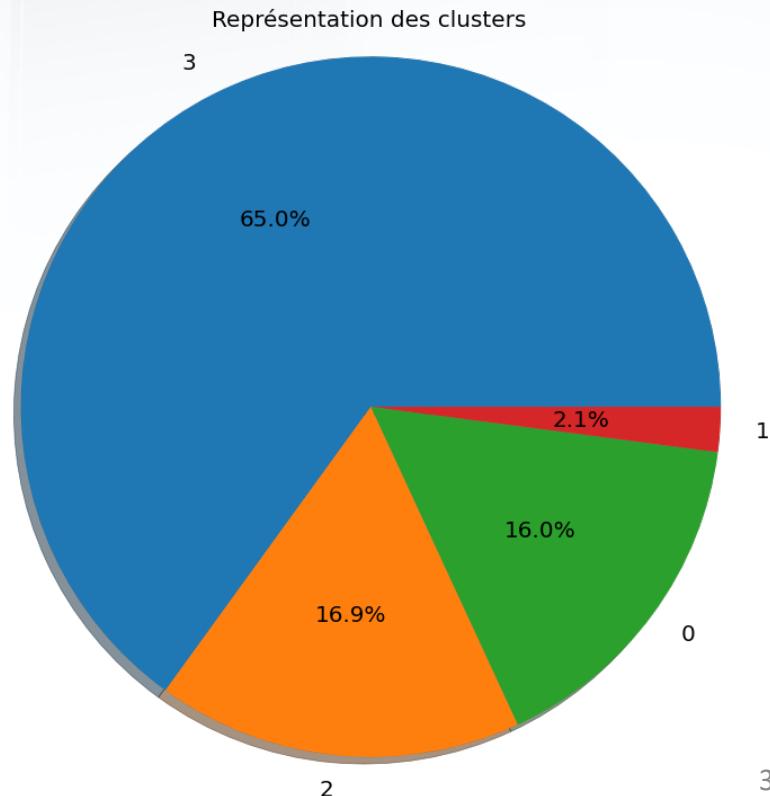
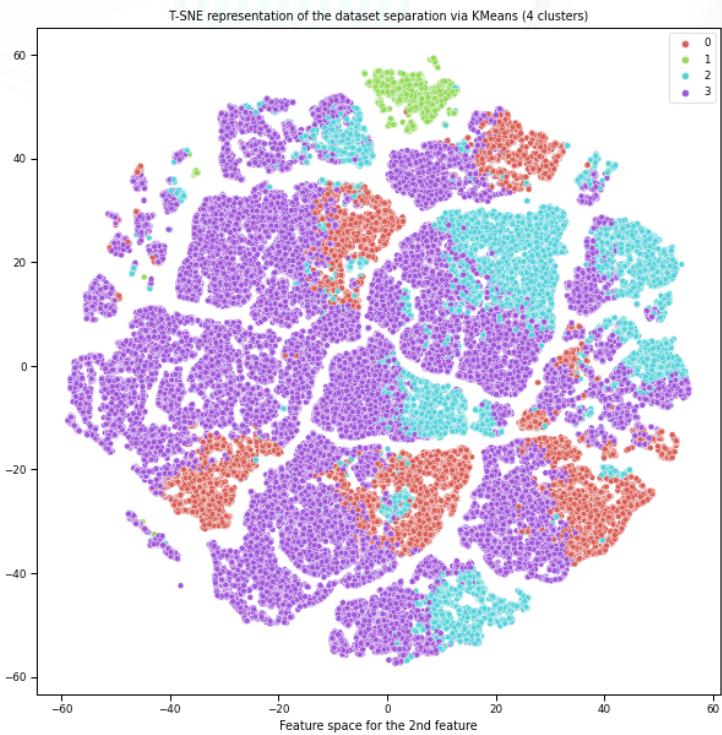
ce



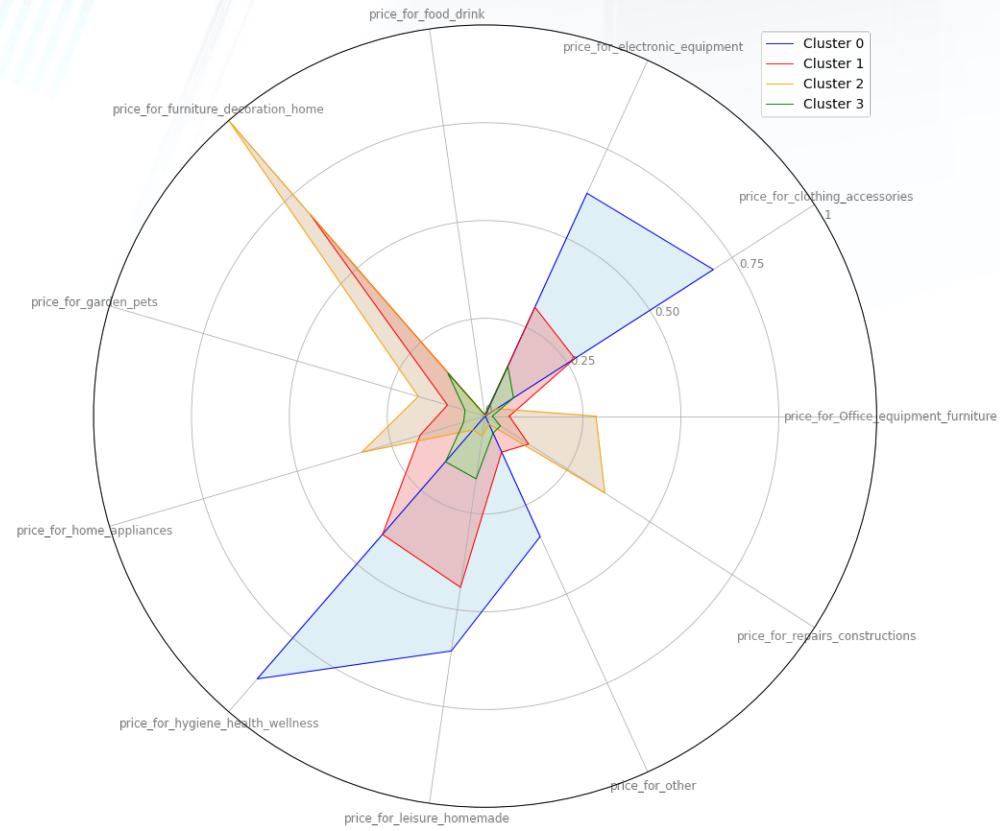
Modèle final sur la totalité des données

K-means, nb_clusters = 4

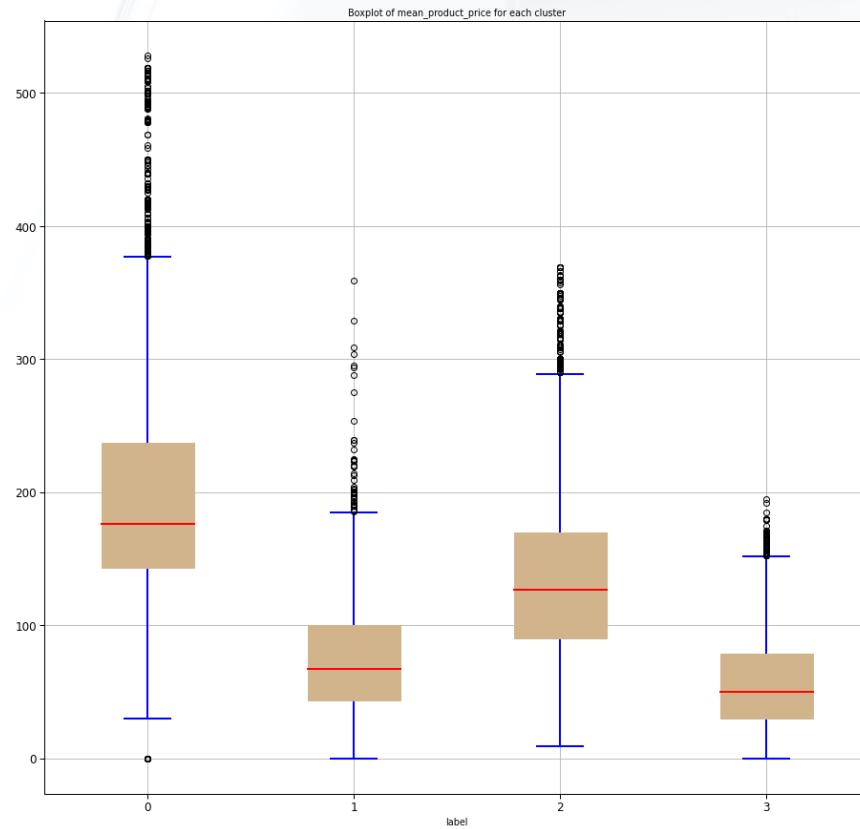
K-means 4 clusters



Analyses

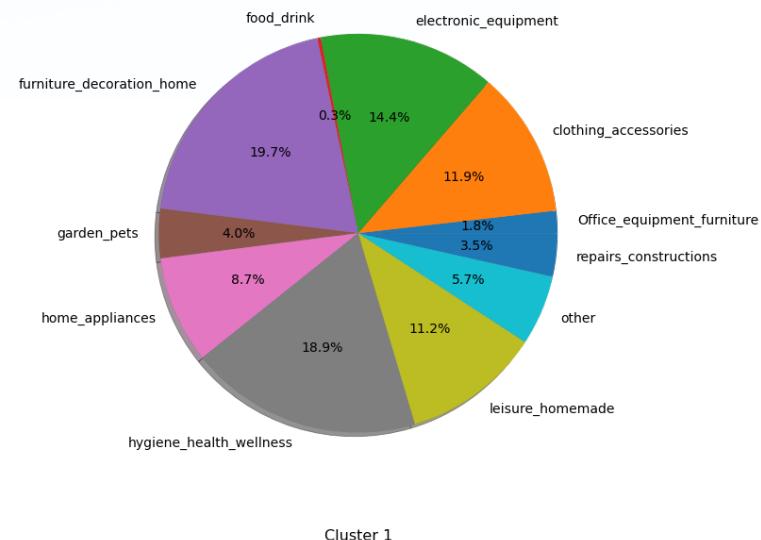
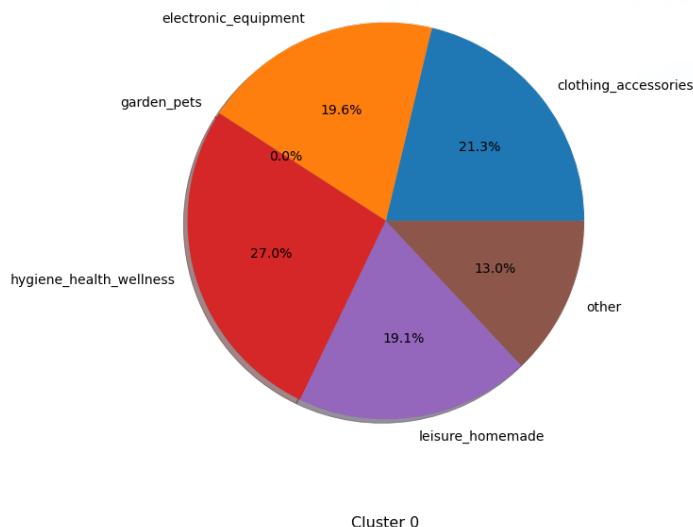


Analyses



Analyses

Distribution of category most purchased according to clusters



Analyse



Cluster	% clients	Nb_commandes	Éloignement avec vendeur	Prix Produits	Volumes produits	Notes	Echelonnages paiement	Cat la plus achetée	Cat : le + de dépenses
0	16 %	1	modéré / fort	élevé	modéré	4.05	x4	Hygiène, vêtement, loisir	Équipement électronique, vêtement, hygiène
1	2.1%	2 ancienne commande	faible	faible	modéré	4.09	X2-3	Décoration, hygiène, Alimentaire	Pas de distinction
2	16.9%	1	faible	modéré	élevé	3.9	X4	Décoration, hygiène, électronique	Fourniture, construction, électroménager
3	65%	1	fort /modéré	faible	faible	4.1	X1-2	Décoration, fourniture, construction	Pas de distinction

Conclusion

- Conclusion :
 - Meilleur algorithme non supervisé : K-means avec 4 clusters
 - Maintenance du modèle tous les ans
 - Création de 4 profils clients :
 - achats coûteux beauté/loisir
 - achats pour la maison et l'alimentaire
 - achats de décoration et produits électroniques volumineux
 - achats peu coûteux liés à la maison
 - Proposer un ciblage par groupe de catégories achetées (publicités, réductions)

Conclusion

- Améliorations possibles :
 - Segmenter le plus grand cluster qui contient plus de 60% des clients
 - Refaire l'étude en gardant les outliers (éviter la suppression de grosses commandes)
 - Au lieu de prendre les 6 premiers (contraintes techniques et de performance) mois prendre :
 - un échantillon aléatoire
 - la totalité du fichier
 - Imputation par la moyenne et non par 0
 - Affiner clustering en optimisant les divers paramètres
 - Ajout de nouvelles features (genre, âge)

Fin de la présentation



Merci pour
votre attention

