Projet 4
Segmentation des clients d’un site e-commerce

Sommaire

Mission

Olist (solution de vente sur les marketplaces en ligne) souhaite fournir à ses équipes d'e-commerce une segmentation des clients pour leurs campagnes de communication.

Fournir à l’équipe marketing une description actionnable de la segmentation
pour une utilisation optimale

La segmentation proposée doit être exploitable et facile d’utilisation pour l’équipe marketing.

Évaluer la fréquence à laquelle la segmentation doit être mise à jour, afin de pouvoir

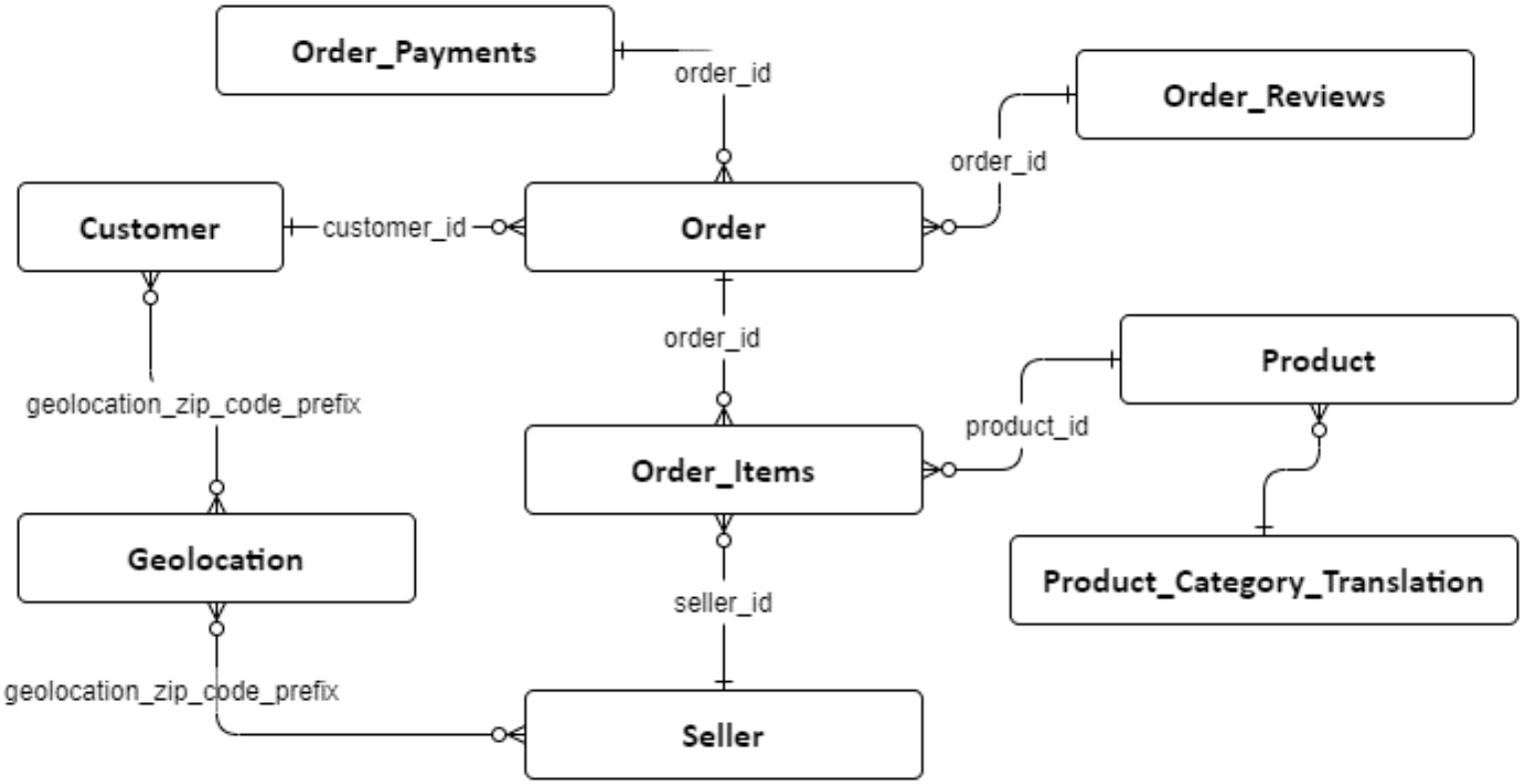
effectuer un devis de contrat de maintenance.

Le code fourni doit respecter la convention PEP8, pour être utilisable par Olist.

Jeu de données

Schéma relationnel de notre base de données

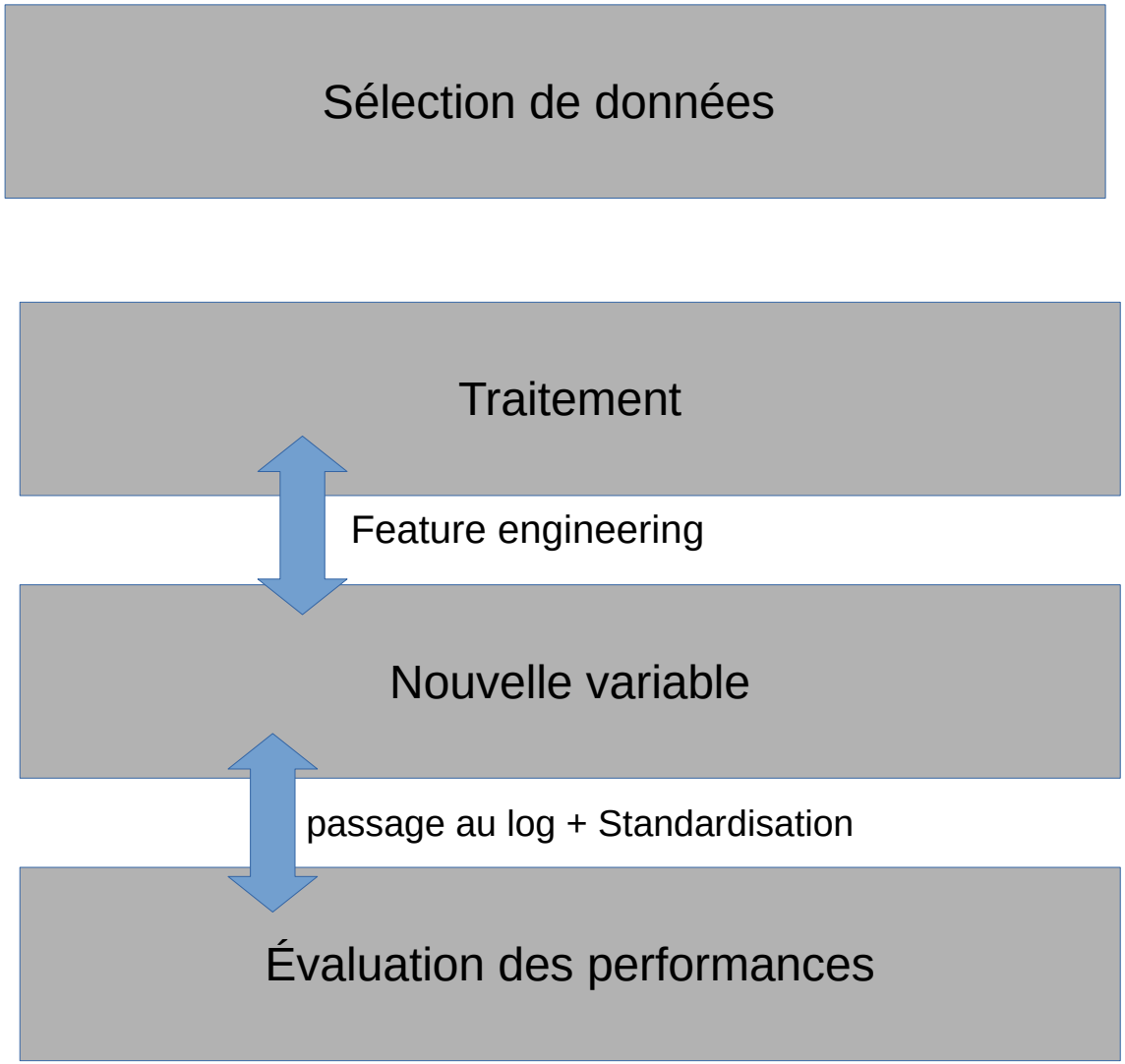
Les données sont organiser afin que les informations ne soit pas redondantes,et respecte une forme normal d une base de données relationnel.



Jeu de données

orders_dataset	Il s'agit de l'ensemble de données de base.					
	Taille:	99441x4	Pct de NaN:	0 %	Doublon:	0
customers_dataset	Informations sur le client et son emplacement.					
	Taille:	99441x5	Pct de NaN:	0 %	Doublon:	0
order_reviews	Informations relatives sur les avis des clients.					
	Taille:	100000x7	Pct de NaN:	20.93 %	Doublon:	0
order_items	Comprend des données sur les articles achetés dans chaque commande.					
	Taille:	112650x7	Pct de NaN:	0 %	Doublon:	0
products_dataset	Contient des données sur les produits vendus par Olist.					
	Taille:	32951x9	Pct de NaN:	0,83 %	Doublon:	0

Analyse exploratoire



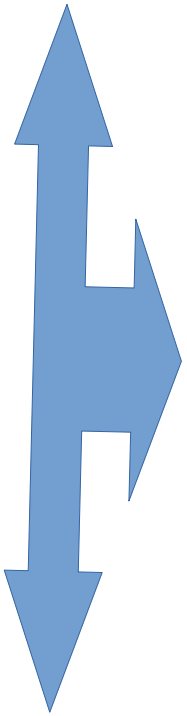
Analyse exploratoire

Le but est de segmenter les **clients**, il est primordial de travailler uniquement avec des features exprimant les données clients.

Les vendeurs ,la géolocalisation et les moyens de paiements ont étaient ignorer

Variables sélectionnées :

payment_sequential : Si plusieurs méthodes de paiement appliquées
payment_installments : Nombre de versements choisis
order_id: Identifiant unique de la commande
order_item_id : Identifiant séquentiel des items d'une même commande (lignes de commande)
product_id : Identifiant unique du produit
seller_id : Identifiant unique du vendeur
price : Prix de la ligne de commande
freight_value : Coût de fret de la ligne (Si plusieurs lignes, le coût de fret est réparti entre toutes les lignes)
order_item_id : est le nombre d'objets dans une même commande
freight_value : est le coût de livraison
shipping_limit_date : correspond à la date d'expédition auprès du transporteur
customer_id : Clé dans le dataset des commandes
customer_unique_id : Identifiant unique du client
customer_state : Etat du client
Distance : distance du client



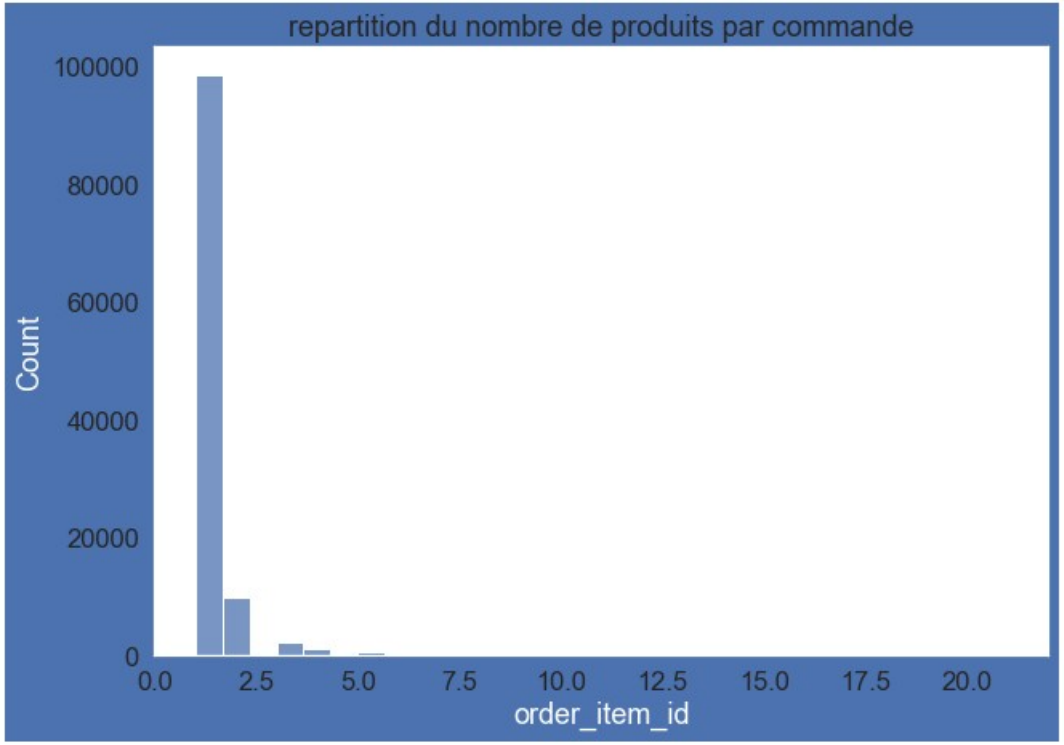
Variables Finales:

customer_unique_id : Identifiant unique du client
total_freight : frais de livraison
mean_payment_installments : moyenne des paiements
mean_review_score : moyenne des notes
delai_dernier_achat_mean : frequence d achat
delai_livraison : délai de livraisons
tot_moy_achats : total des achats
favorite_sale_month : mois favori

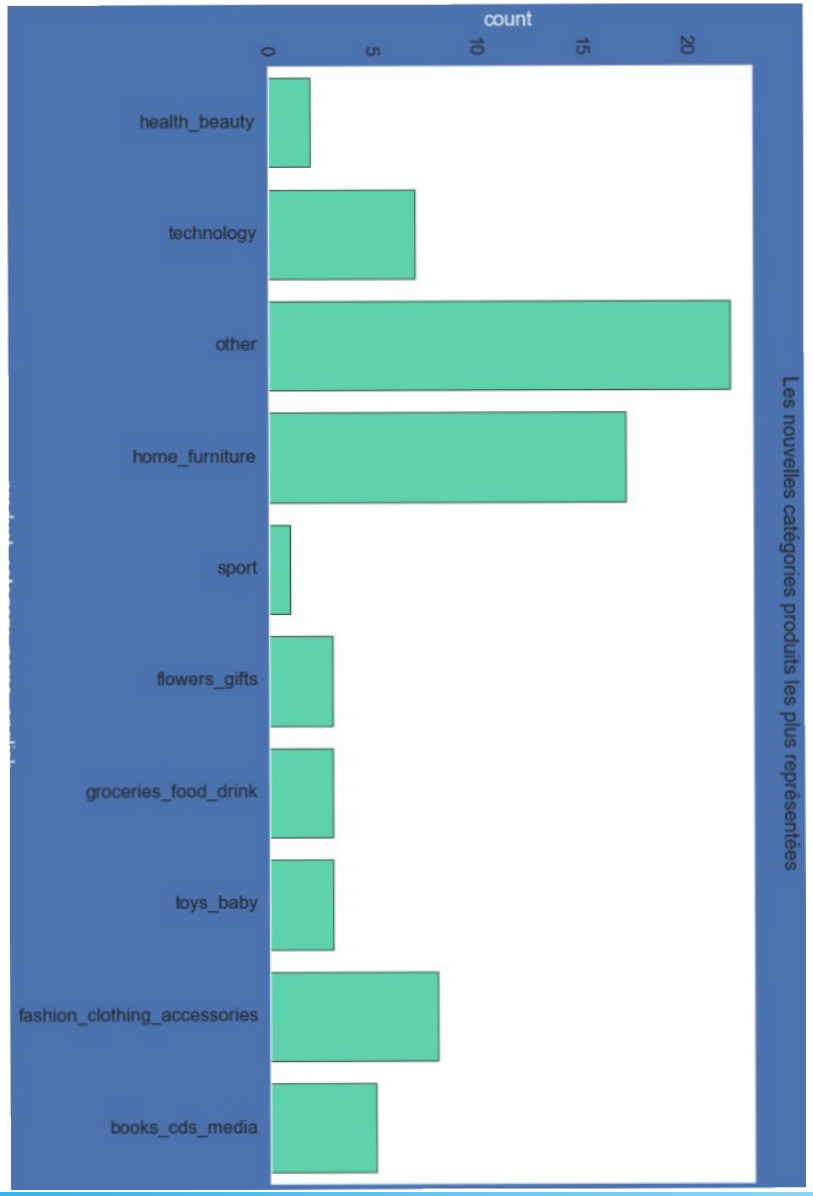
Analyse exploratoire

La plupart des clients ont acheté une seule fois

Avec en moyenne un seul article par commande

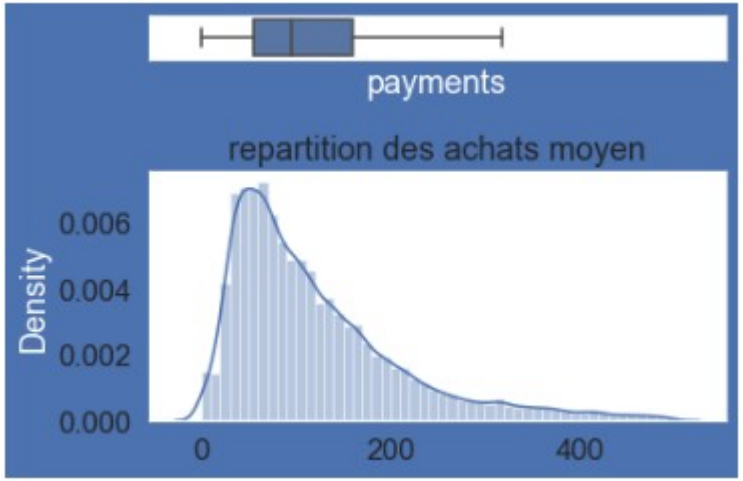


Réduction des catégories



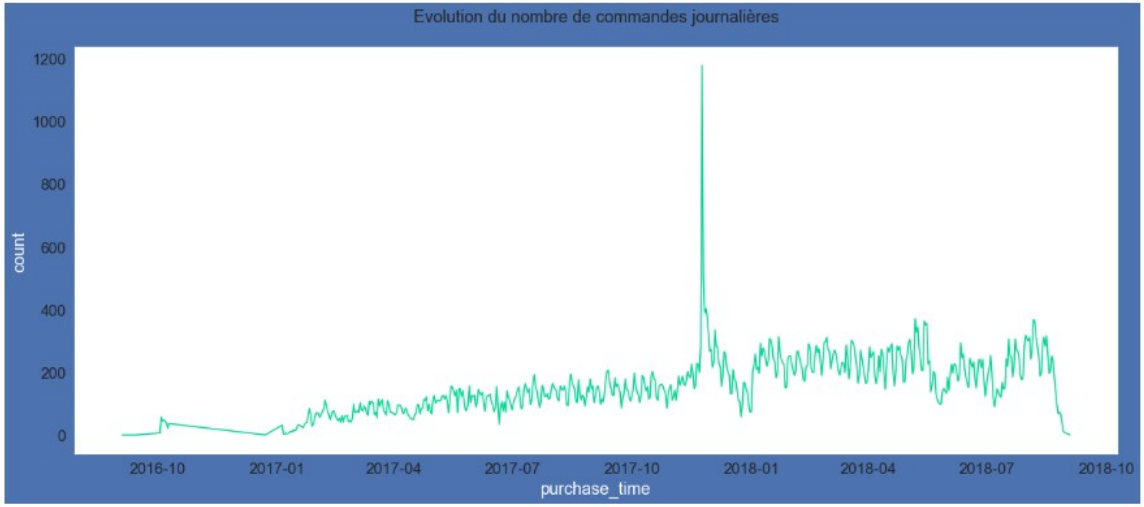
Analyse exploratoire

moyenne des paiements 100

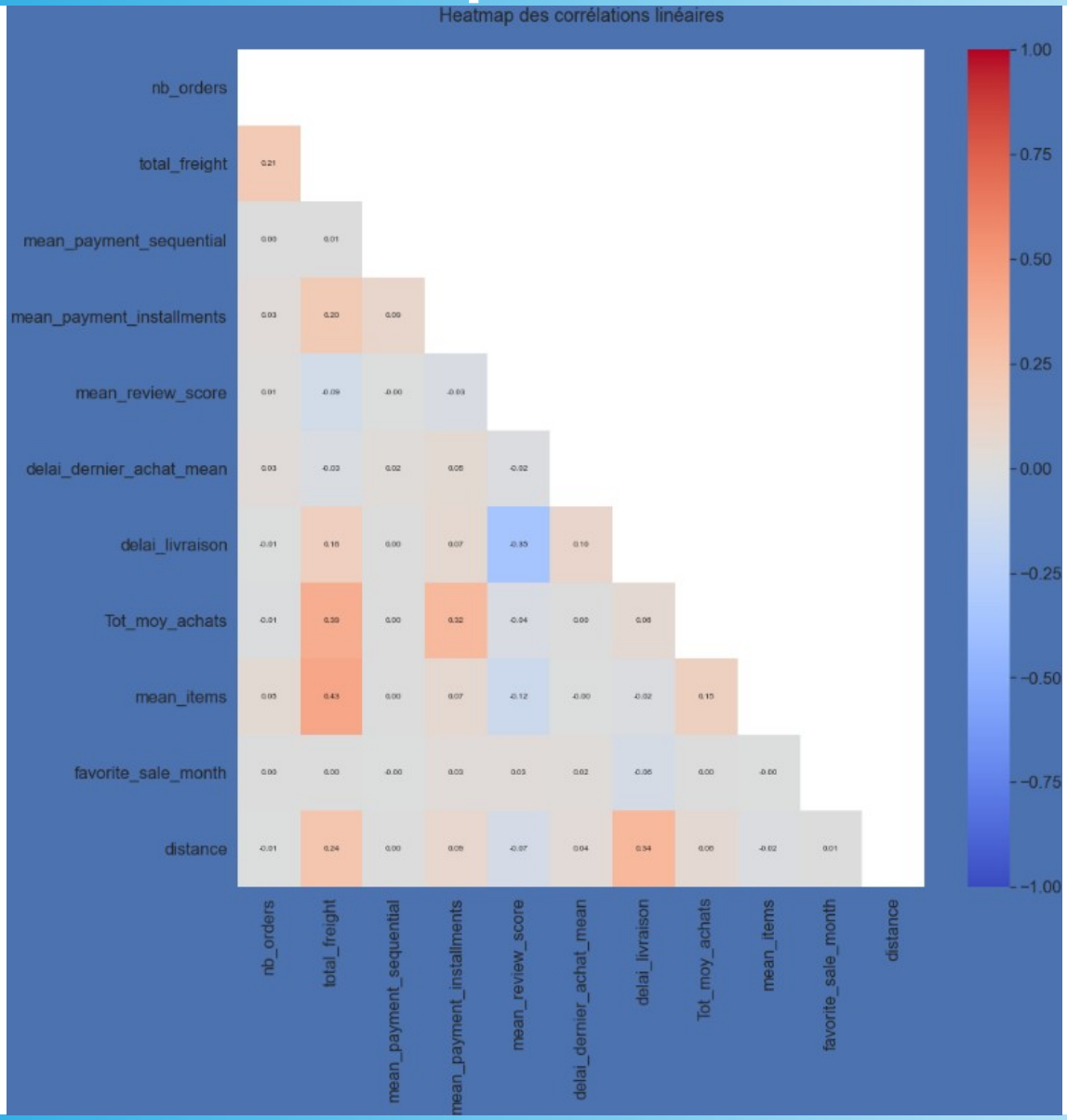


Delai vs delai estimé

Commande dans le temp

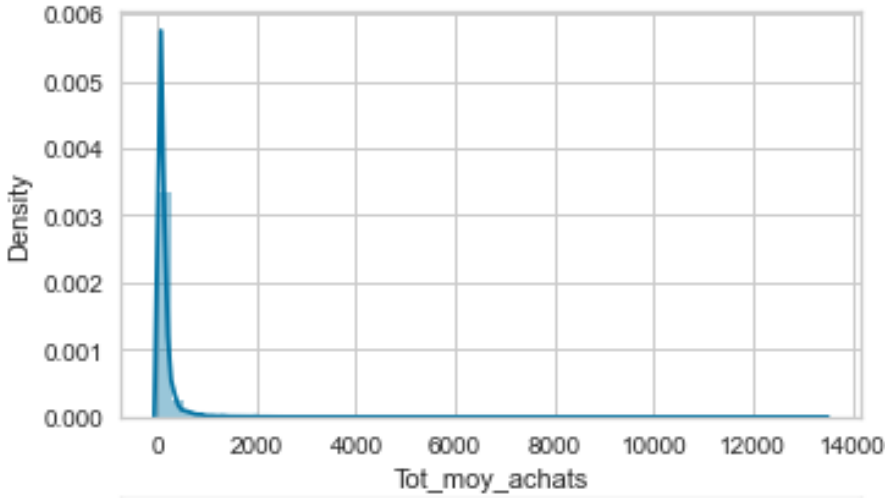


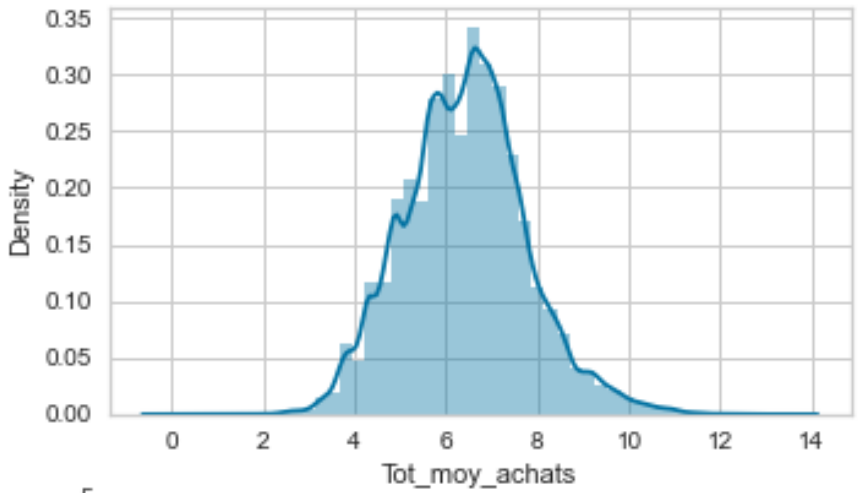
Corrélation après traitement



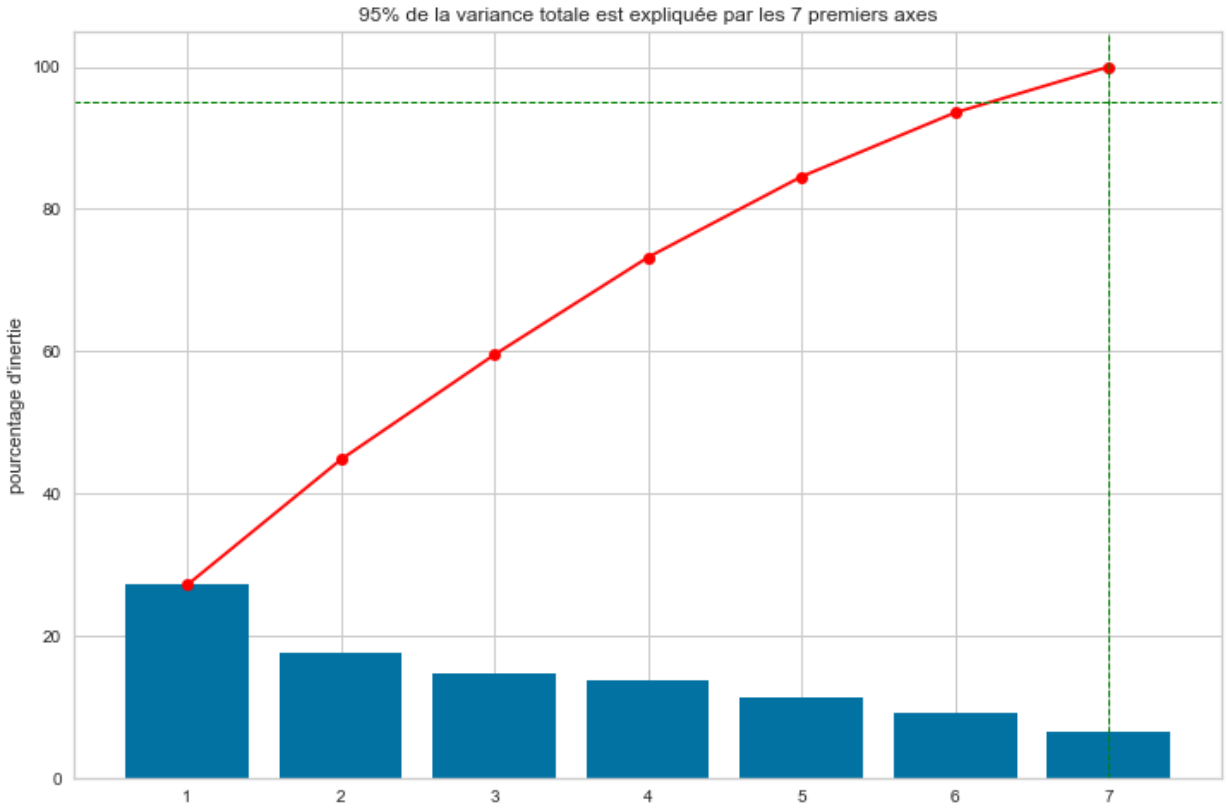
Passage au log

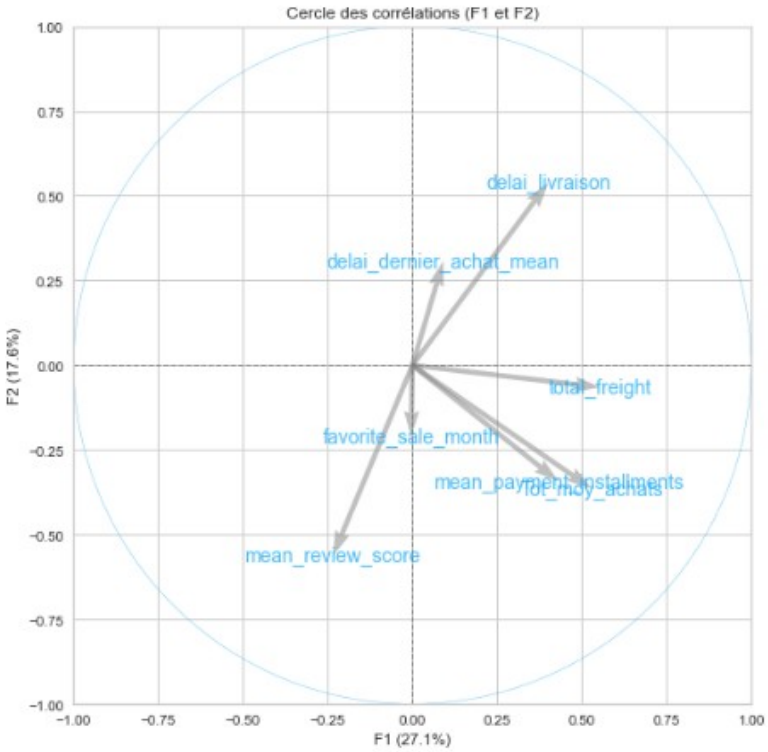
Aucune variable n'a de distribution normale





ACP

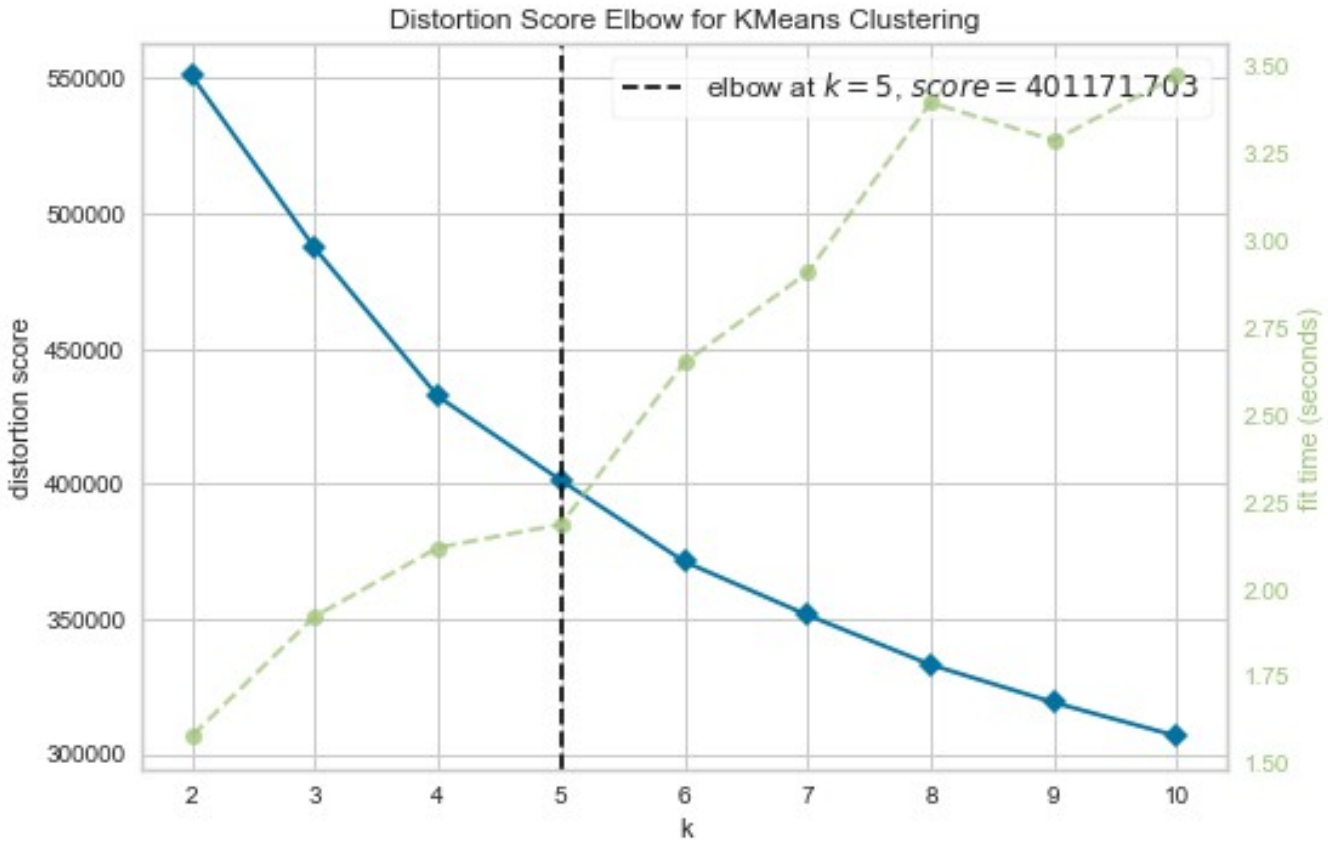




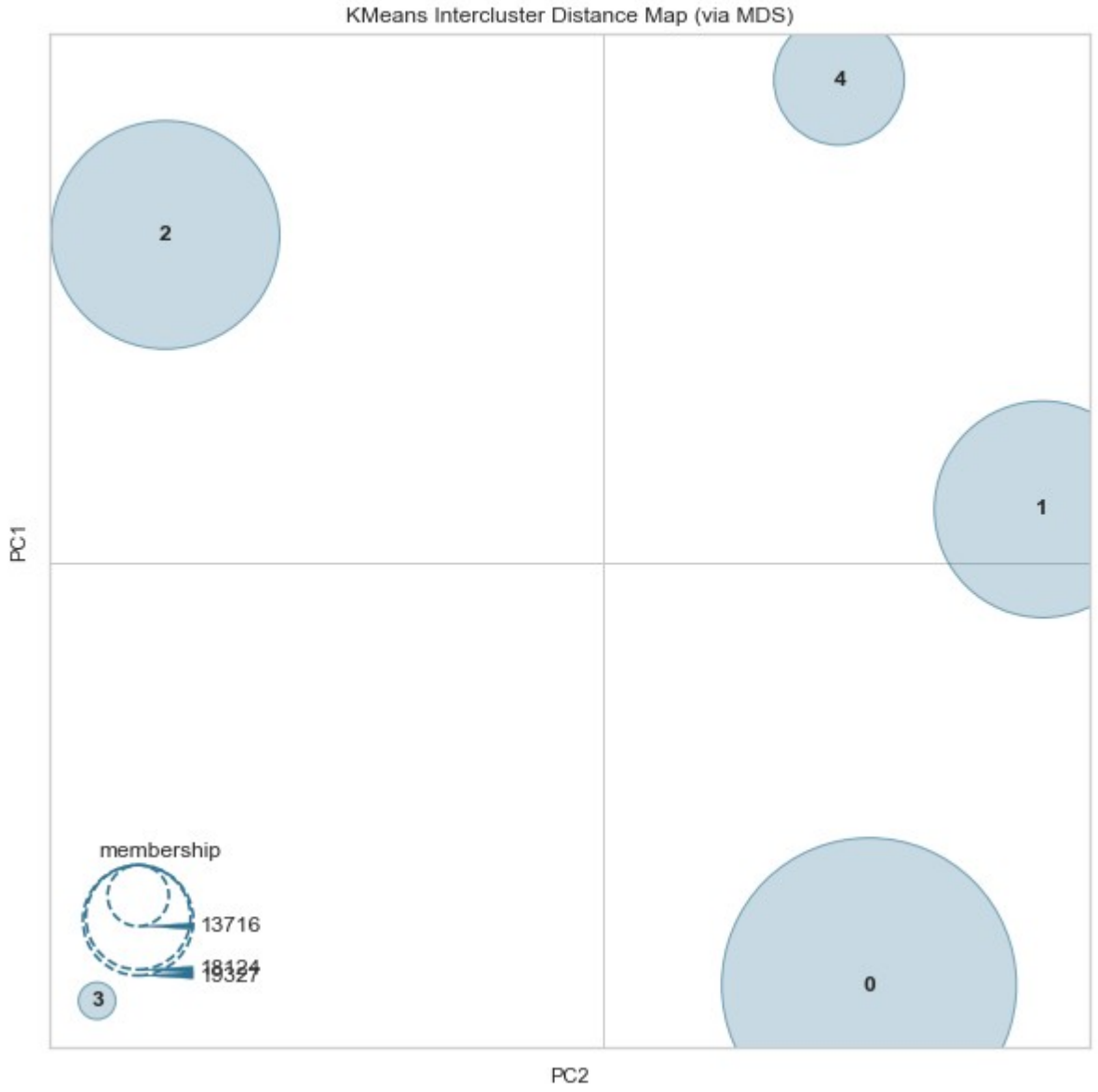
Nous allons pouvoir réduire notre dataset en conservant 95 % de la variance sur les 7 premiers axes

Modèle K means

Un cluster avec K = 5 est retenu, grâce a la méthode du coude

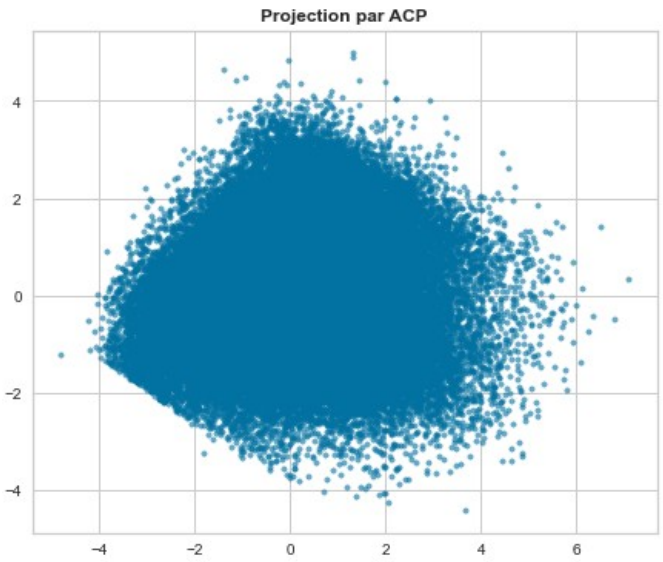


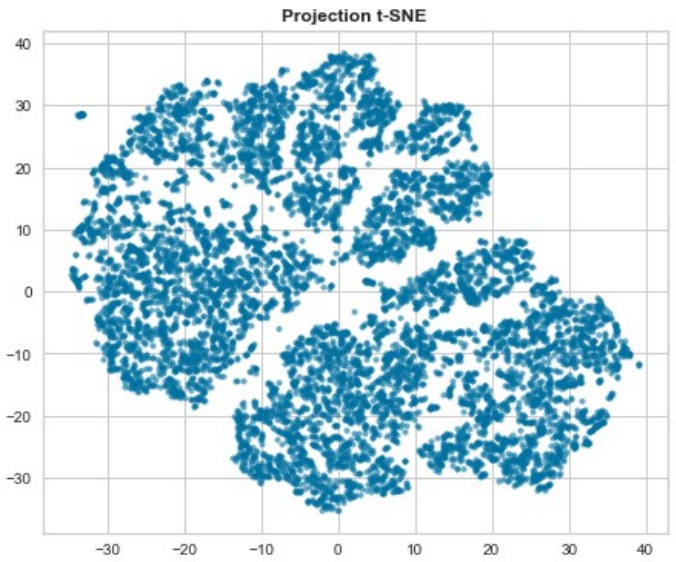
Séparation inter clusters

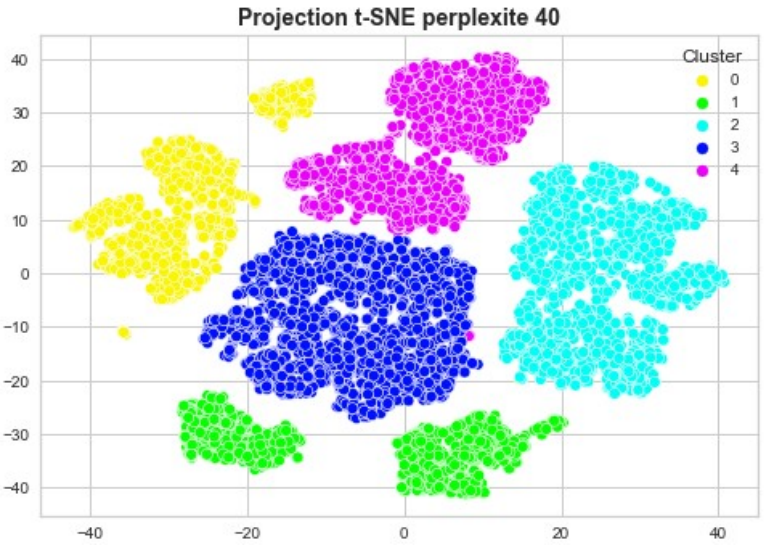


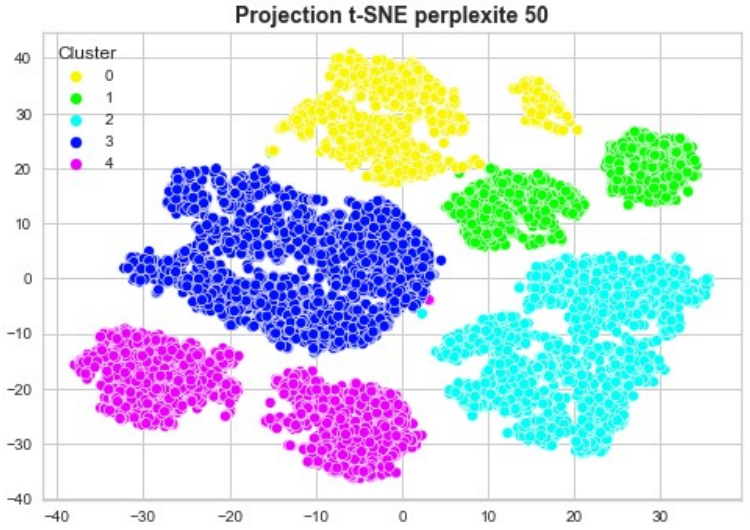
Les distances inter clusters sont homogènes
La distances semble suffisante entre les clusters

TSNE





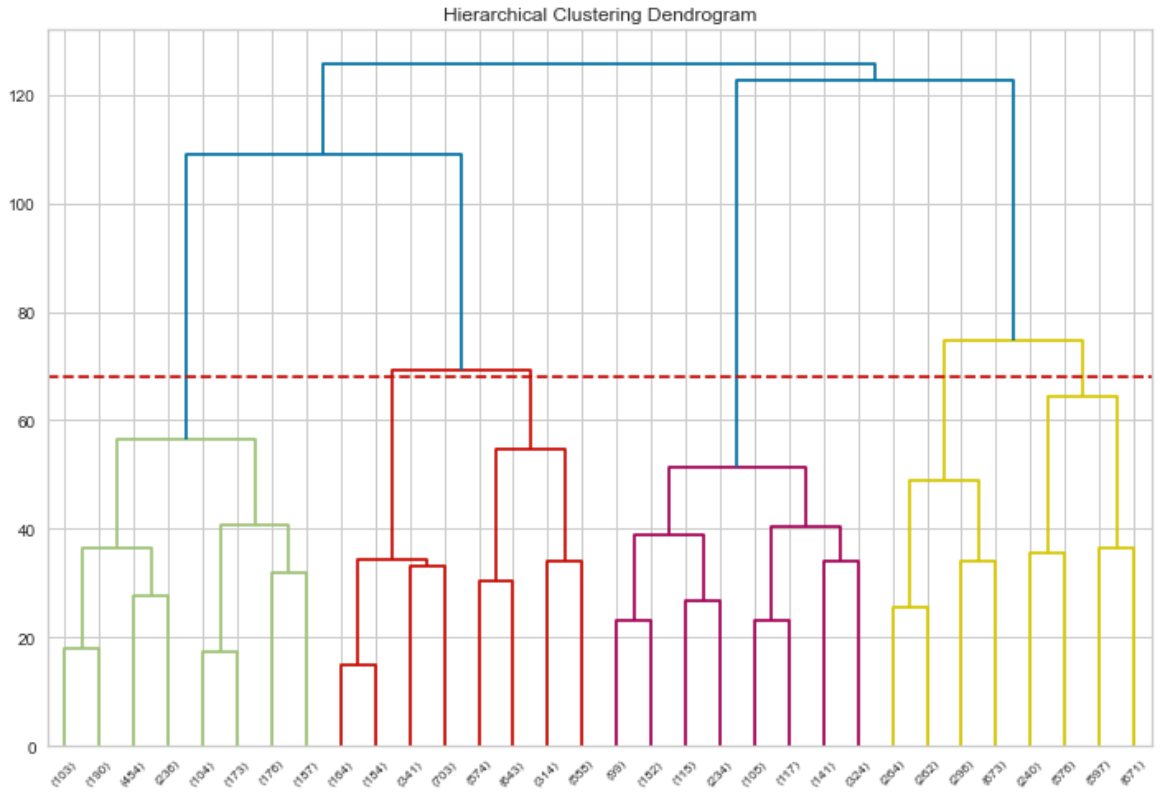




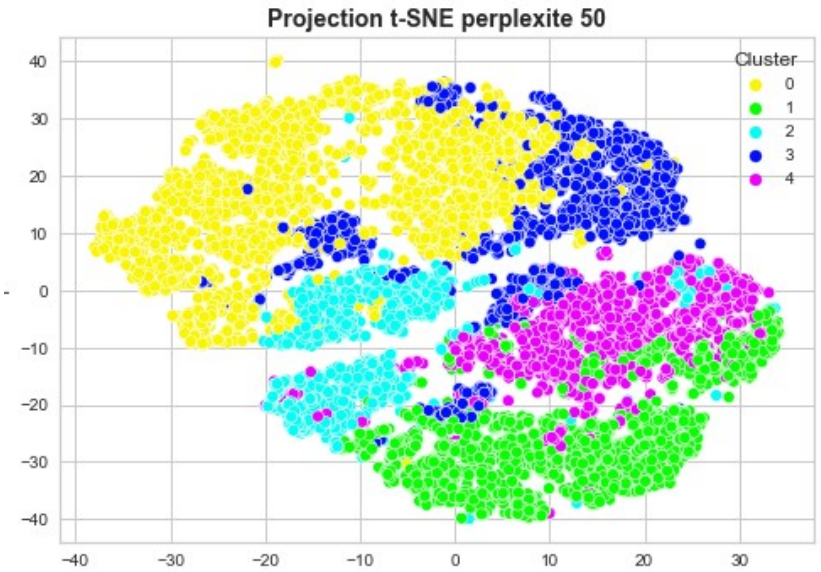
Projection de nos clusters Kmeans sur TSNE

Modèle Hierarchical Clustering

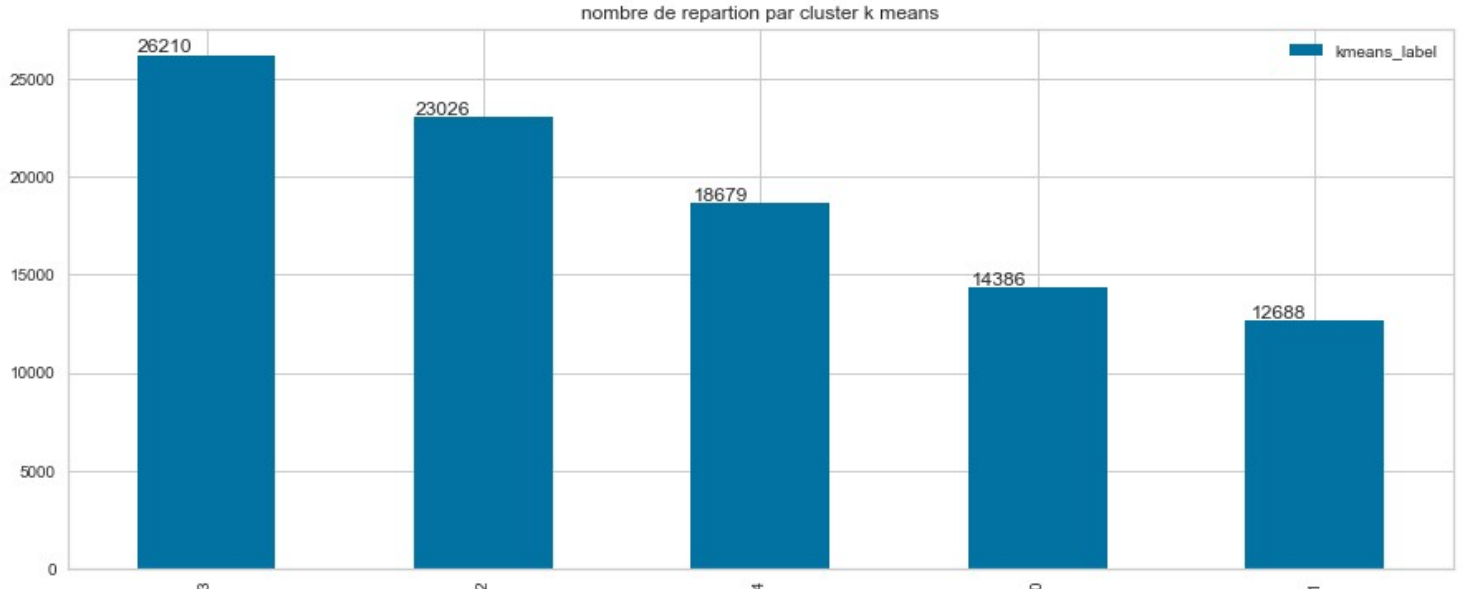
AgglomerativeClustering

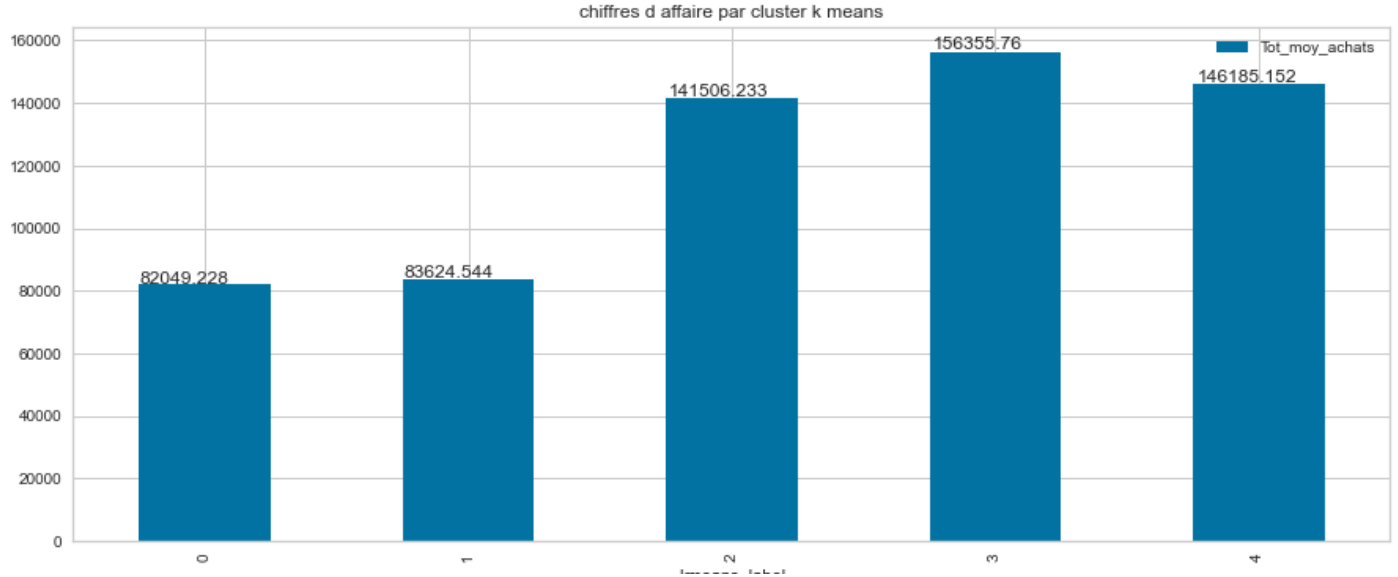


Projection de nos clusters hiérarchique sur TSNE



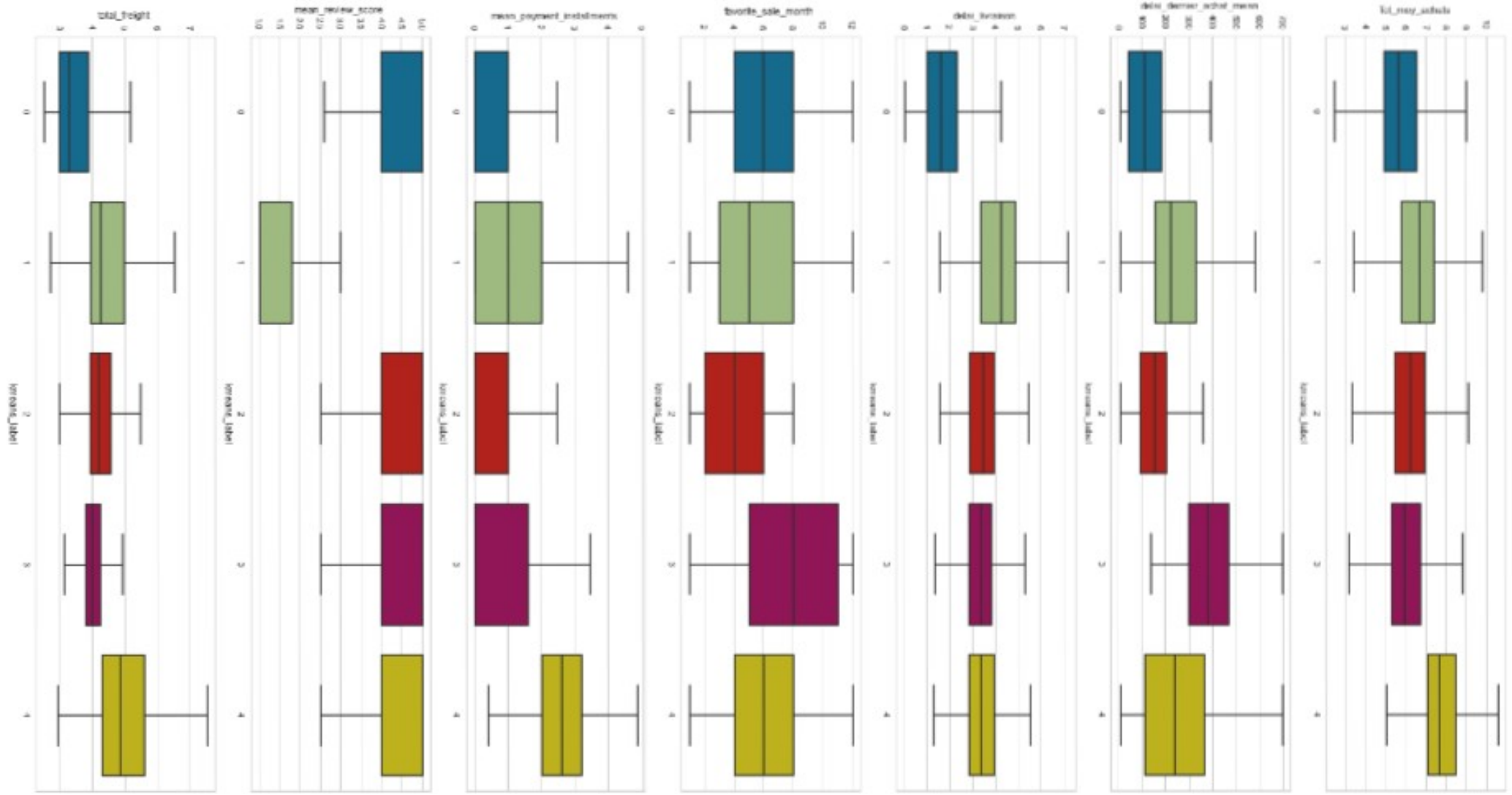
Analyse métier de nos clusters





Analyse métier de nos clusters

Frais transport | Note | Nb paiement | Mois favoris | Delai livraison | Delai achat | Montant



Analyse métier de nos clusters

Groupe 1 : Courts délais de livraison, commandant vers le début d'année pour des montants faibles.
Ils paient avec 1 type de moyen de paiement et avec un nombre faible d'échéances.
Les avis de ces clients sont très bons. Petit montant

Groupe 2 : Ce sont des clients mécontents (les avis sont mauvais).
Les délais de livraison sont très importants et les frais de port élevés.
Règle avec plusieurs moyens de paiements , montant moyen, achat vers le début d année

Groupe 3 : Clients de fin d'année.
Ils règlent avec un moyens de paiement pour des montants faibles.
les délais de livraison sont long. Les avis de ces clients sont très bons.
Petit écart avec le dernier achat effectuer

Groupe 4 : Client très satisfait , avec de grand délai de livraisons et de frais de port
sur des petits achats, les délais entre deux commandes dont très grand, client de fin d année

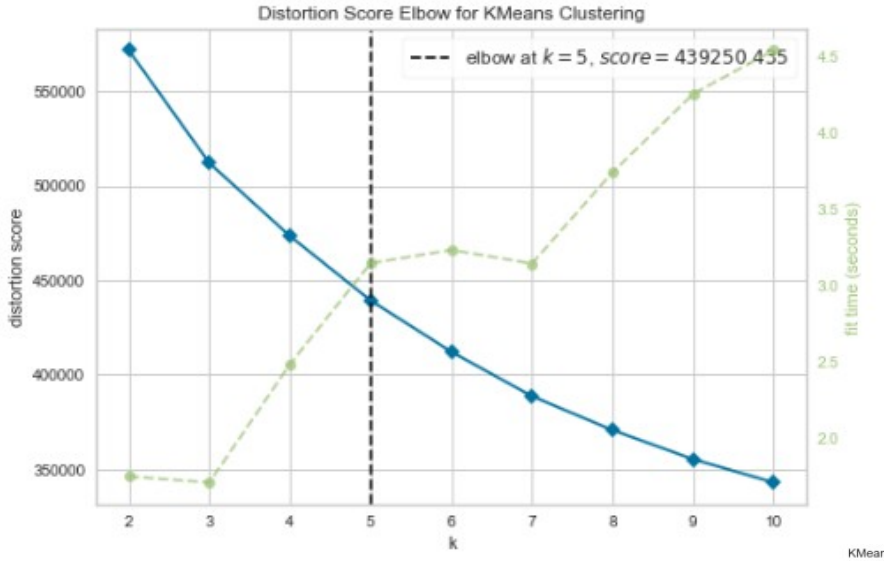
Groupe 5 : Regroupe les clients qui utilisent plusieurs moyens de paiement
et un nombre important d'échéances. Ils ont tendance à espacer les délais entre 2 commandes.
Les avis de ces clients sont également très bons.
Cette catégorie est celle ou les montant sont les plus élevées

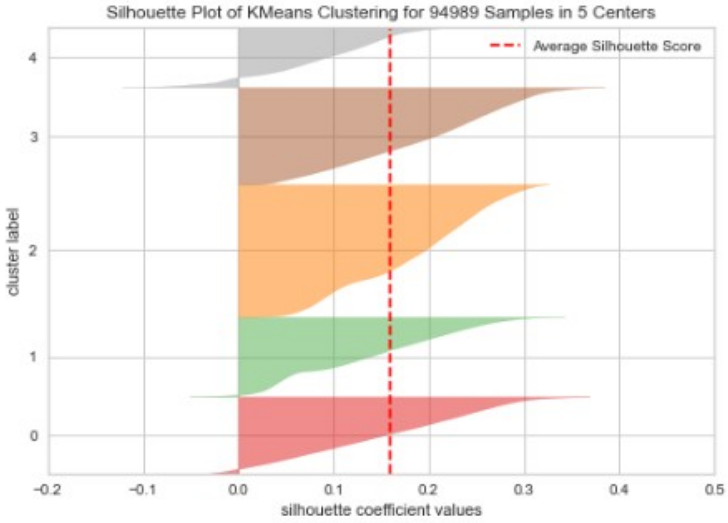
Stabilité du K means

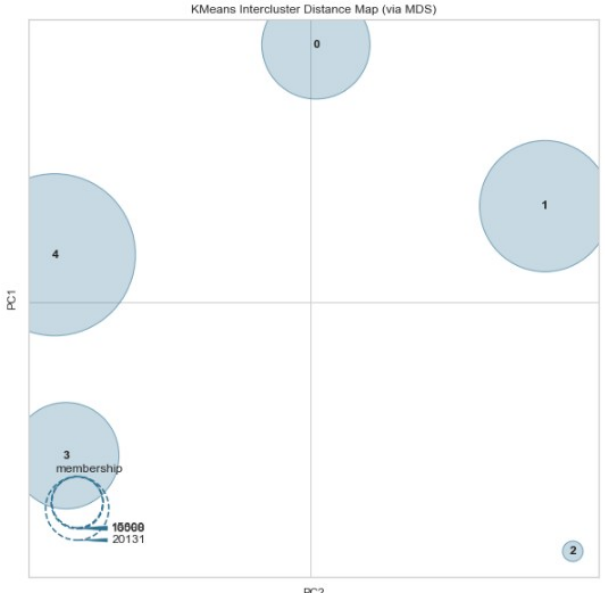
moyenne adjusted_mutual_info_score : 0.9954036968144797
moyenne adjusted_rand_score : 0.9979715192195708
moyenne homogeneity_score : 0.995853729839451

	iteration	FitTime	Inertia	homogeneity_score	adjusted_rand_score	adjusted_mutual_info_score
0	iter 0	2.751431703567505	439079.755692	0.997682	0.999085	0.997700
0	iter 1	2.3610599040985107	439081.133094	0.994012	0.997292	0.994040
0	iter 2	2.4884681701660156	439081.807047	0.993452	0.996990	0.993481
0	iter 3	2.613607168197632	439080.395174	0.996230	0.998397	0.996244
0	iter 4	2.3470041751861572	439080.005602	0.996685	0.998628	0.996709
0	iter 5	2.2880606651306152	439079.781581	0.996445	0.998505	0.996474
0	iter 6	2.41557240486145	439081.380449	0.993842	0.997197	0.993869
0	iter 7	2.2675344944000244	439080.471070	0.996412	0.998495	0.996423
0	iter 8	2.4876163005828857	439082.961431	0.994357	0.997448	0.994340
0	iter 9	2.682260751724243	439081.017822	0.994738	0.997679	0.994756

Kmeans sur les 7 axes ACP







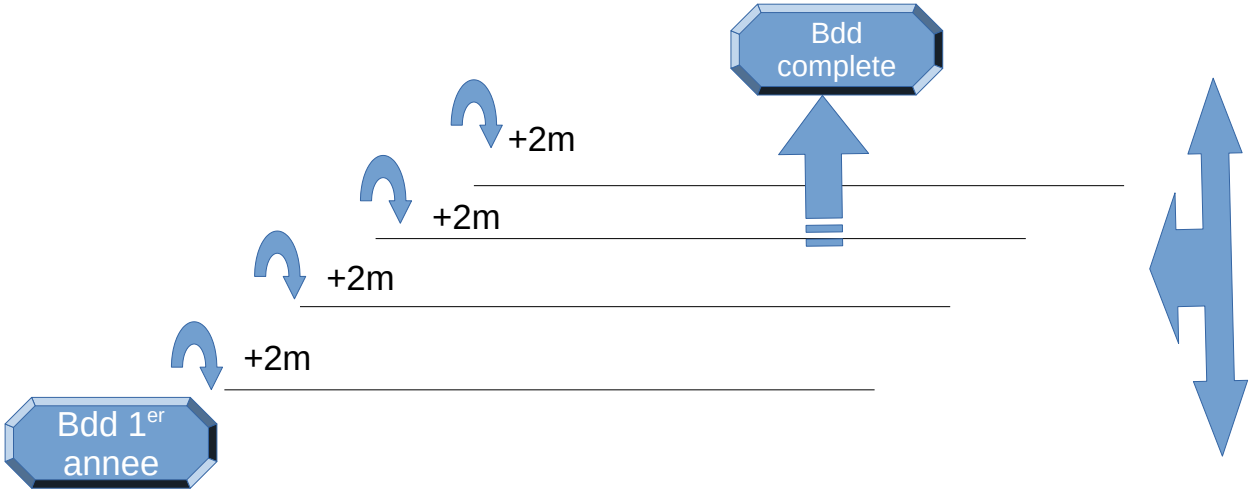
Stabilité temporelle de la segmentation

Nous allons comparer **deux modèles kmeans** :

- L'un s'étant entraîner sur la première année
- L'autre s'étant entraîner sur la totalité

Methodologie :

Création d un dataset avec uniquement la première année ,
Ajout par tranche de deux mois les nouveaux clients au données du dataset.
A chaque itération , nous réalisons une clusterisation sur ce dataset ,
Que nous comparons aux prédictions des deux modelés avec nos métriques



Metriques :
homogeneity_score
adjusted_rand_score
adjusted_mutual_info_score

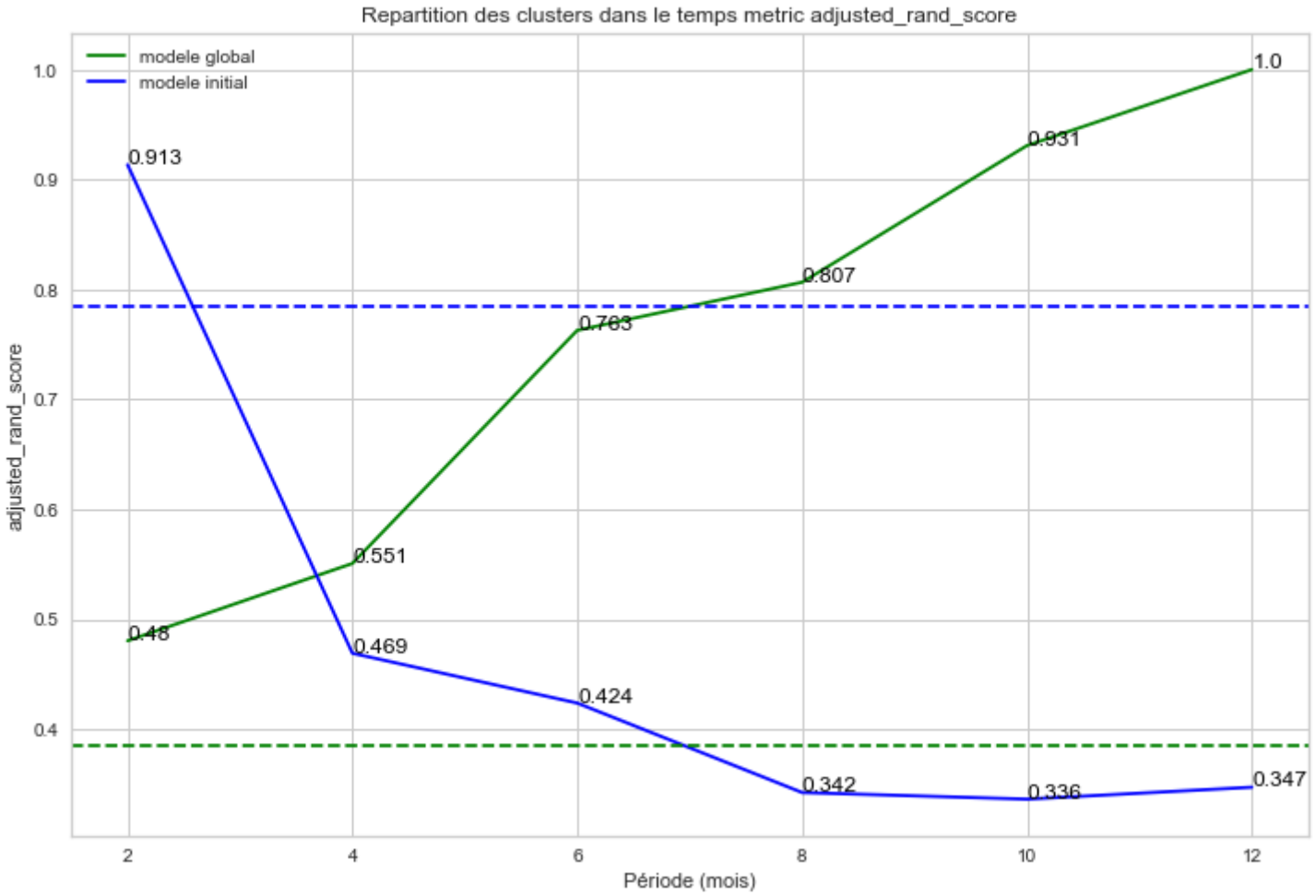
- Kmeans sur la periode
- Comparaisons sur modele 1
- Comparaisons sur modele 2

Stabilité temporelle de la segmentation

Metriques :
homogeneity_score
adjusted_rand_score
adjusted_mutual_info_score

Une mise à jour des clusters

- Tous les **4** à **6 mois** est recommandé

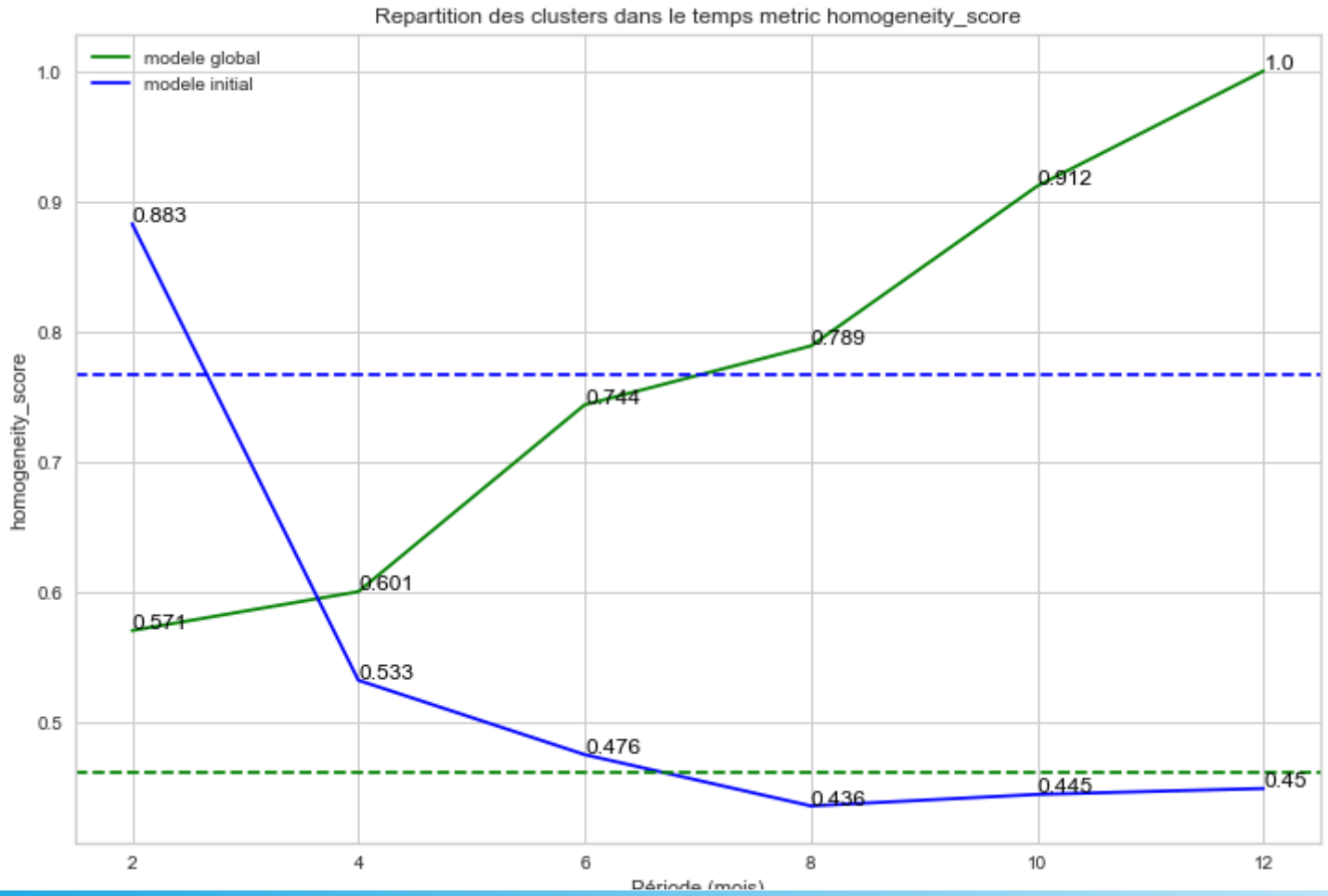


Stabilité temporelle de la segmentation

Metriques :
homogeneity_score
adjusted_rand_score
adjusted_mutual_info_score

Une mise à jour des clusters

- Tous les **4** à **6 mois** est recommandé



Conclusion

Il est nécessaire de prendre en compte plus de variable pour identifier chaque type de clients selon ses comportements d'achat. (anniversaire, article en promotion...)

La clusterisation nous à permis d identifier différents groupes , en particulier les meilleurs clients

Difficulté d exploiter certaine variables, comme les nombres d achat (majoritairement à un),ou le nombre de produit par commande(majoritairement à un aussi),
Un dataset plus représentatif avec plus de fréquence d achat, permettrais une meilleur segmentation des clients

Les catégories d achat n ont pas pus être exploiter , approfondir d autres méthodes d encoding de variable catégorielle

L analyse est fortement liée à la satisfaction client

Les algorithmes de clustering permettent de regrouper les données
mais posent un problème d interprétation (ex : attribuer des rangs au groupe)

Questions/Réponses

Thank you!