Segmenter des clients d'un site e-commerce

Contexte & Enjeux du Projet Olist

Présentation d'Olist

- Plateforme e-commerce brésilienne (marketplace B2B/B2C)
- Met en relation vendeurs et clients à l’échelle nationale
- Forte croissance, structuration d’une équipe Data

Mission confiée

- Accompagner la création de l’équipe Data Science
- Premier cas d’usage : segmentation des clients Olist
- Objectif : fournir une segmentation exploitable au quotidien par le marketing

Objectifs opérationnels

- Comprendre les différents profils clients via leur comportement d’achat et leur satisfaction
- Proposer des segments actionnables pour personnaliser les campagnes de communication
- Recommander une fréquence de mise à jour pour garantir la pertinence de la segmentation dans le temps

Plan

- Présentation des données
- Préparation des données
- Modélisation et choix de l’algorithme
- Présentation du meilleur modèle
- Profils clients et actions marketing ciblées
- Recommandation de fréquence de mise à jour

Présentation des données Olist

Table clients

- customers : informations clients et localisation

Tables commandes

- orders, order_items, order_payments, order_reviews : informations sur les commandes, articles achetés, paiements, avis clients

Tables produits

- products, translation : détails produits et catégories

Tables vendeurs et géolocalisation

- sellers, geolocation : informations vendeurs et localisation géographique

Chiffres clés

- Période couverte : 2016 – 2018
- Nombre de commandes : 99 441
- Nombre de clients : 96 096
- Nombre de produits : 32 951
- Nombre de vendeurs : 3 095

Préparation des données

Nettoyage, fusion et agrégation

- Nettoyage : gestion des types, valeurs manquantes, suppression des incohérences
- Fusion des principales tables (clients, commandes, produits, paiements, avis, géolocalisation)
- Agrégation au niveau client pour obtenir une table synthétique
- Résultat : un DataFrame de 93 103 clients, avec 33 variables quantitatives et qualitatives

Point méthodologique

- Ces 33 variables constituent la base initiale pour l'analyse exploratoire.
- Elles ont ensuite été analysées, ajustées, regroupées ou supprimées selon leur pertinence statistique et métier lors des étapes suivantes.

Analyse univariée & ajustements des variables

Exploration des variables quantitatives

- Outliers détectés (frais_livraison_moyens, retards de livraison) → création de variables relatives (part_frais_livraison) ou de classes qualitatives (catégories de délais de livraison).

Gestion des valeurs manquantes & enrichissement

- Notes clients manquantes remplacées par la moyenne globale pour conserver tous les clients.
- Ajout d'une variable d'engagement (a_donne_une_note) pour distinguer clients engagés et passifs.

Variables peu discriminantes

- Suppression ou regroupement des variables trop homogènes (ex : total_articles_achetes, articles_moyens_par_commande). sauf si elles apportent une information clé pour la segmentation (ex : nb_commandes pour la fréquence d'achat)
- Regroupement des modalités pour les variables déséquilibrées (catégories produits, moyens de paiement, régions).

Conclusion

Affinement du jeu de données : passage de 33 à 29 variables, toutes analysées et ajustées pour maximiser la pertinence du clustering.

Analyse bivariée & réduction de la redondance

Corrélations entre variables numériques

- Test de normalité (Kolmogorov-Smirnov) : toutes les variables numériques sont non normales
- Corrélations de Spearman utilisées pour mesurer les relations monotones
- Résultat : très fortes corrélations entre plusieurs variables (jusqu'à 0.99), indiquant une forte redondance d'information

Décision : suppression des variables fortement corrélées pour éviter la surpondération dans le clustering

Conclusion

Affinement du jeu de données : Passage de 29 variables à 23 variables après suppression des redondances

Analyse multivariée & sélection finale des variables

Démarche

Étape 1 : Analyse de la redondance géographique

Vérification de la redondance entre les coordonnées géographiques et la région :

→ Suppression des coordonnées, la région étant suffisante pour capter l'information spatiale.

Étape 2 : Réduction de dimension via ACP

Application de l'Analyse en Composantes Principales pour identifier les axes les plus discriminants et éliminer les variables peu contributives ou redondantes.

Résultat

- Affinement progressif du jeu de données :
- Passage de 23 à 21 variables après l'analyse géographique
- Puis de 21 à 15 variables après l'ACP

Jeu de données final :

Un ensemble restreint de variables quantitatives, sélectionnées pour leur pouvoir discriminant et leur pertinence pour la segmentation.

Modélisation

Préparation des données pour la modélisation

Standardisation des variables quantitatives

- Éviter que les variables à grande échelle dominent le calcul des distances.
- Garantir une contribution équitable de chaque variable dans le clustering.
- Préserver la cohérence et l'interprétabilité des résultats.

Comment ?

- Toutes les variables quantitatives ont été standardisées (moyenne = 0, écart-type = 1).
- Cette étape est essentielle pour les algorithmes basés sur la distance, comme K-Means.

Encodage des variables catégorielles

- Les variables catégorielles ont été transformées via One-Hot Encoding afin de pouvoir être prises en compte dans les algorithmes de clustering.

Choix des algorithmes

K-Means

Fonctionnement :
→ Crée des clusters sphériques autour de centroïdes (centres virtuels).
→ Objectif : minimiser la distance entre les points et leur centroïde.

Points forts :
✓Rapide, idéal pour grands volumes
✓Résultats stables et interprétables
✓Adapté à des groupes équilibrés

Paramètre clé : nombre de clusters (déterminé par méthode du coude/silhouette).

DBSCAN

Fonctionnement :
→ Identifie des zones denses de points, ignore les zones peu denses (bruit).
→ Pas de centroïdes : les clusters peuvent avoir des formes arbitraires.

Points forts :
✓Détecte les outliers
✓Pas besoin de spécifier le nombre de clusters
✓Gère bien les clusters non sphériques

Paramètres clés : distance epsilon (eps) et min_samples.

Démarche K-Means

Détermination du nombre optimal de clusters

- À chaque itération, utilisation de la méthode du coude et du coefficient de silhouette pour identifier la valeur optimale de k

Approche itérative

- À chaque étape :
  - Analyse des variables les moins discriminantes.
  - Modification (découpage en classes) ou suppression progressive de ces variables.
  - Nouvelle recherche du k optimal adaptée à la nouvelle configuration des variables.
  - Test de la stabilité à l'initialisation : plusieurs lancements de K-Means pour vérifier la robustesse des clusters face à l'aléa de départ.
  - Comparaison systématique des scores (silhouette, inertie…) pour évaluer la qualité des clusters obtenus.

Objectif

- Améliorer la qualité et l'interprétabilité des clusters.
- Arriver à une segmentation finale exploitable.

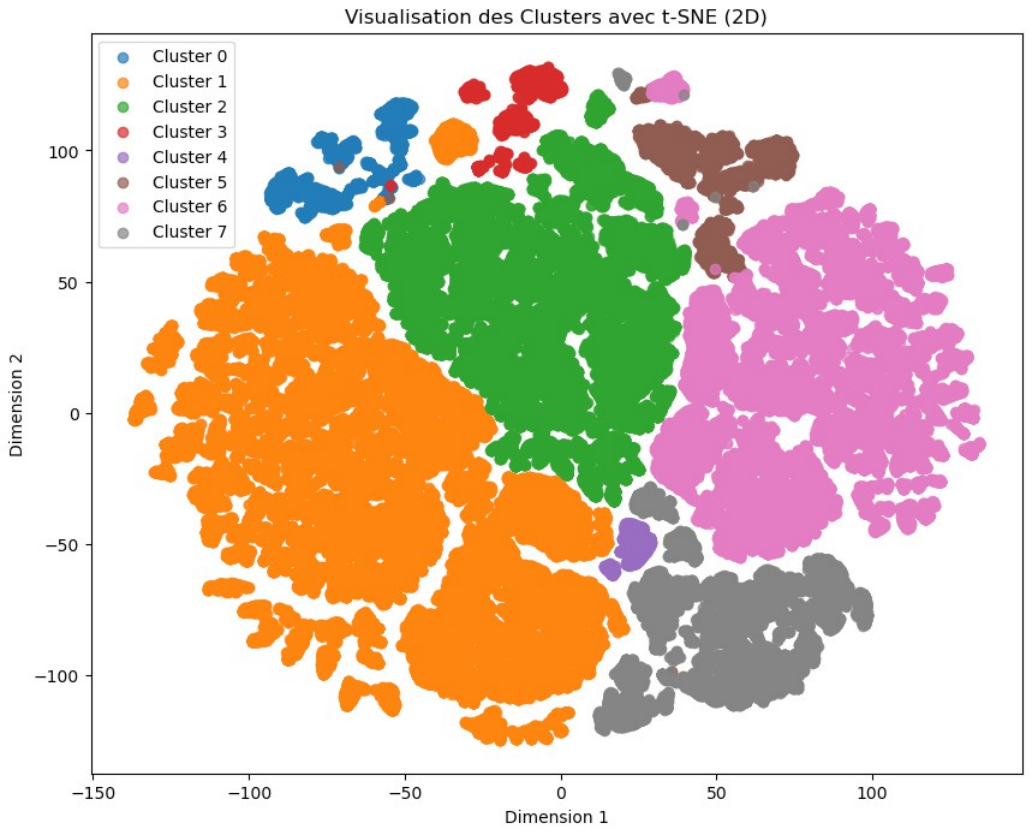
Démarche DBSCAN

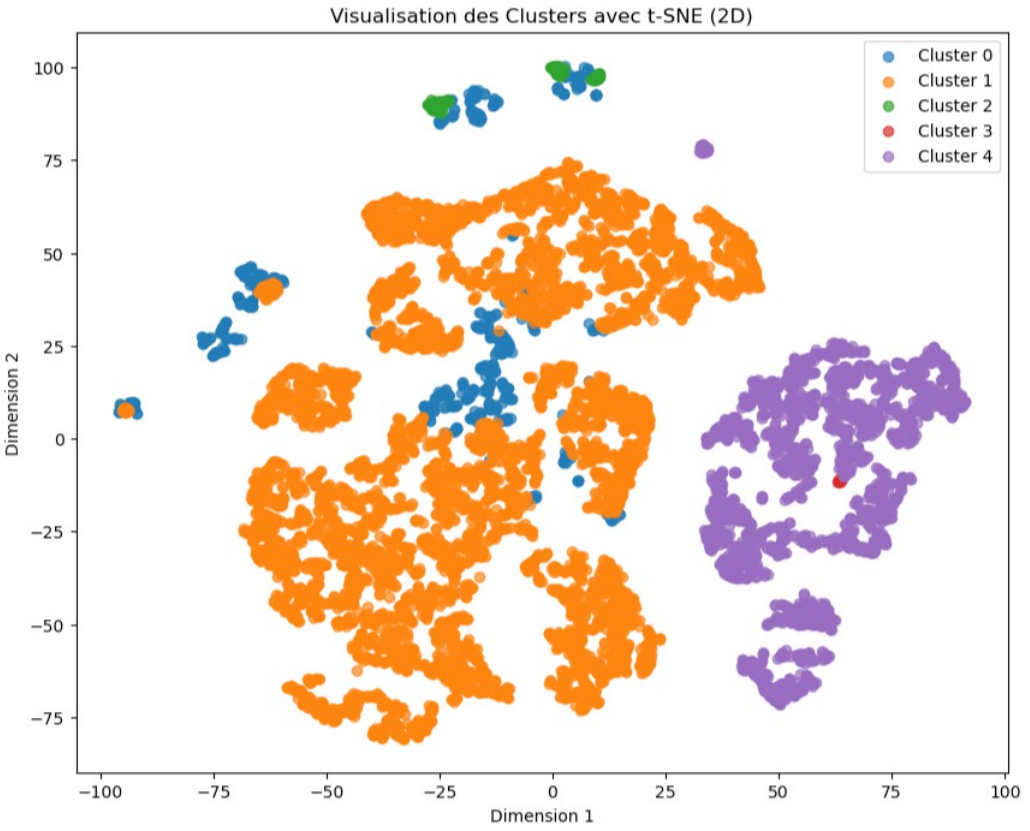
- Algorithme basé sur la densité, détecte automatiquement les clusters et les outliers, sans besoin de fixer le nombre de groupes à l'avance.
- Utilisation des mêmes variables que pour le meilleur K-Means pour garantir la cohérence de la comparaison.
- Travail sur un sous-échantillon de 10 % des données (9 310 clients) pour accélérer les calculs.
- Optimisation des paramètres :
  - Recherche systématique (grid search) de la meilleure combinaison de epsilon (ε) et min_samples.
  - ε (epsilon) : distance maximale pour être voisins.
  - min_samples : nombre minimal de points pour former un cluster.
  - Sélection des valeurs optimales à l'aide d'un k-distance plot et de l'évaluation des scores de qualité.
- Stabilité des clusters testée et validée.

Comparaison synthétique des modèles

Modèle	Cluster	Silhouette	Davies-Bouldin	Calinski-Harabasz	Inertie
K-Means	8	0.440	0.975	21 564	461 701
DBSCAN	5*	0.412	1.464	742	-

* (après regroupement, initialement 10 + bruit)



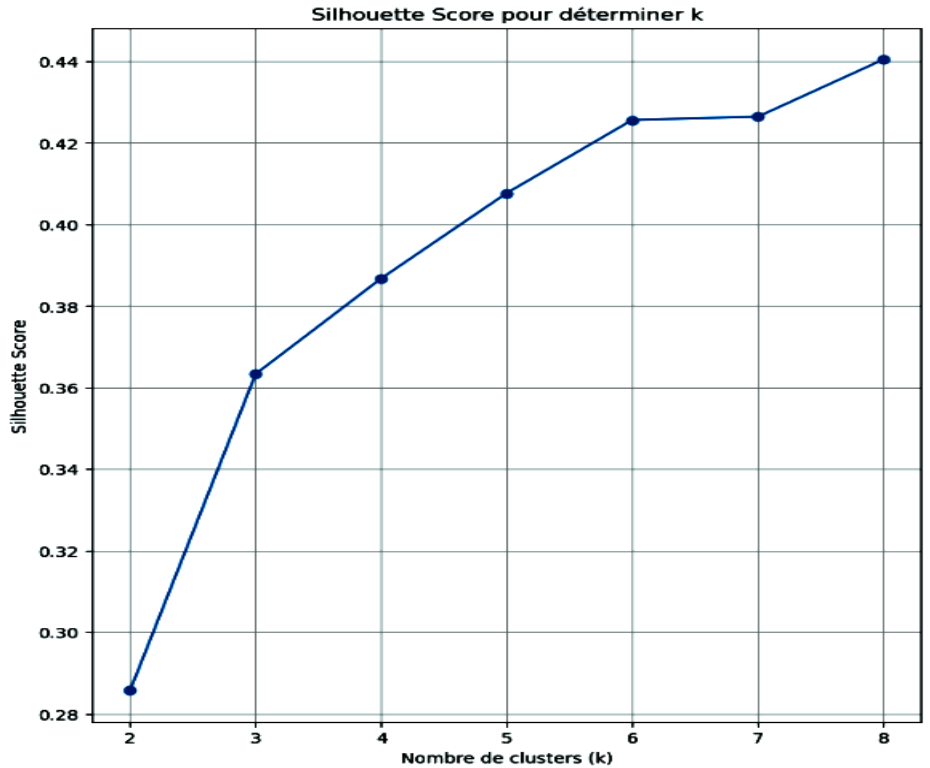


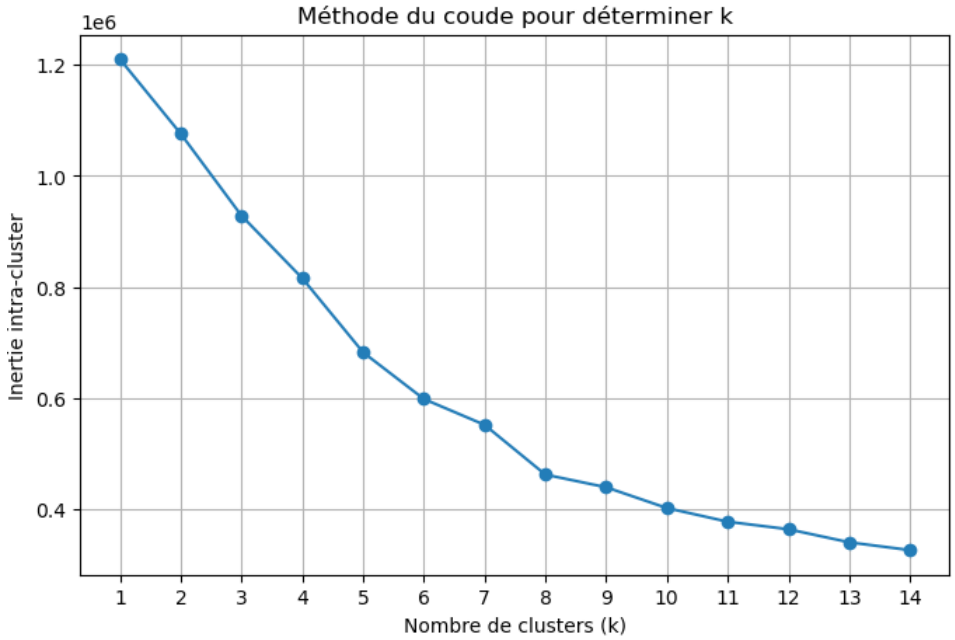
Variables retenues pour la segmentation

Variable	Intérêt pour la segmentation
nb_commandes	Fréquence d'achat, fidélité
total_depense	Valeur client, potentiel
recence_en_jours_cat	Récence de l'activité (nouveaux vs anciens clients)
note_moyenne_client	Satisfaction globale
a_donne_une_note	Engagement client (feedback)
total_retards_livraison	Expérience logistique, source potentielle d'insatisfaction
nb_paiements_total	Intensité des transactions
diversite_max_paiements	Souplesse et diversité des moyens de paiement
diversite_categories	Appétence pour la variété de produits
poids_moyen_commandes	Type de commandes (petit vs gros achats)
frais_livraison_moyens	Sensibilité aux coûts logistiques

Focus sur le modèle retenu

Le score silhouette mesure la séparation entre les clusters : plus il est élevé, plus les groupes sont bien séparés.



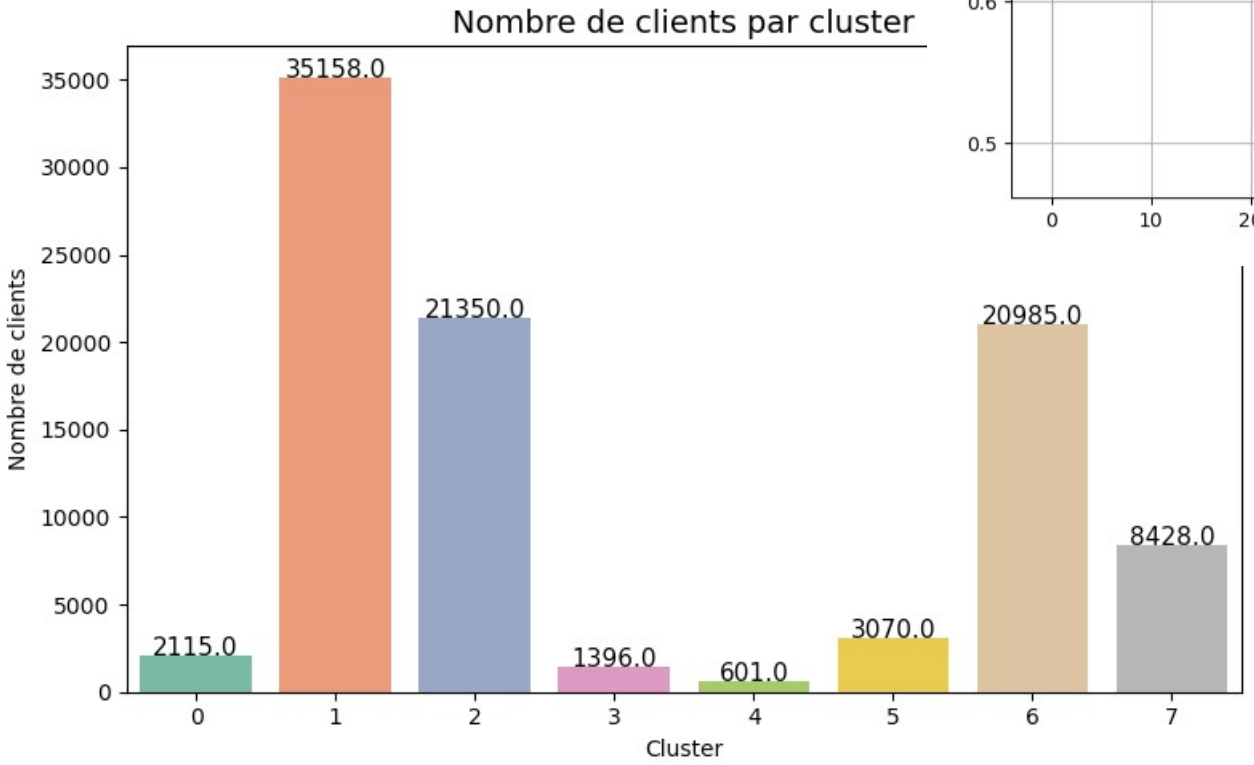


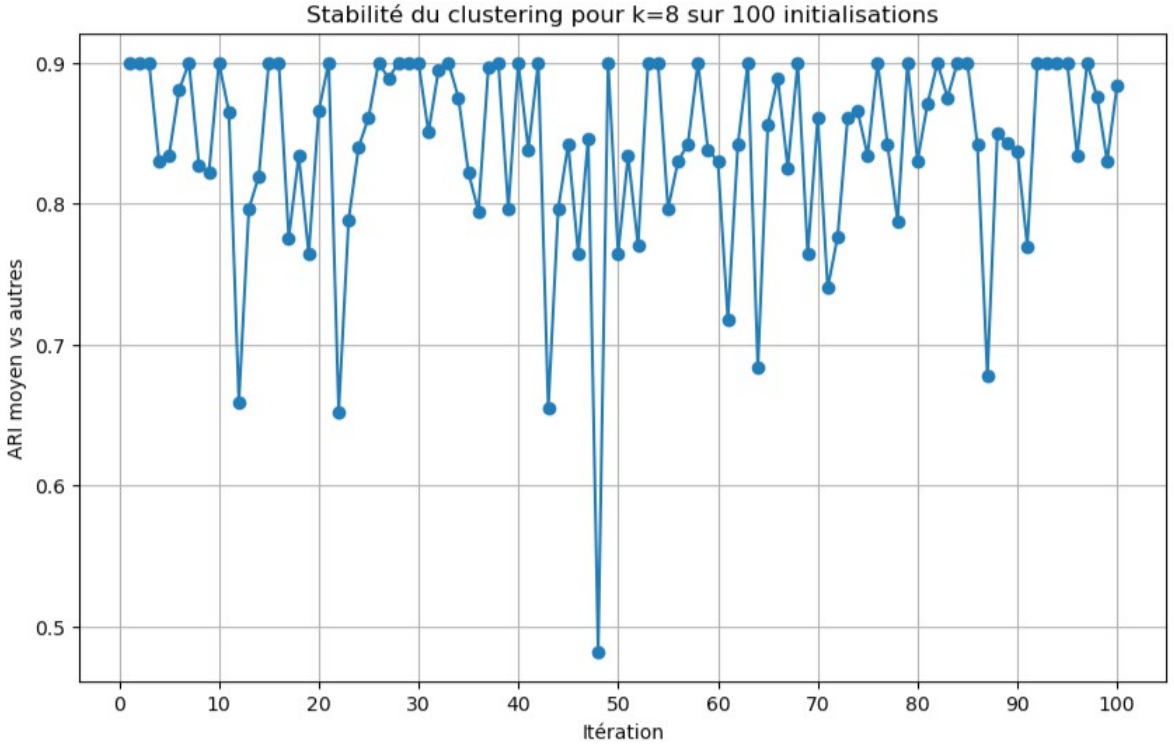
L'inertie mesure la compacité des clusters : on cherche le point où l'ajout d'un cluster n'apporte plus de gain significatif ('coude').

Le choix de 8 clusters maximise la séparation et la compacité, tout en restant interprétable pour le marketing.

Répartition des clients et stabilité des clusters (k=8)

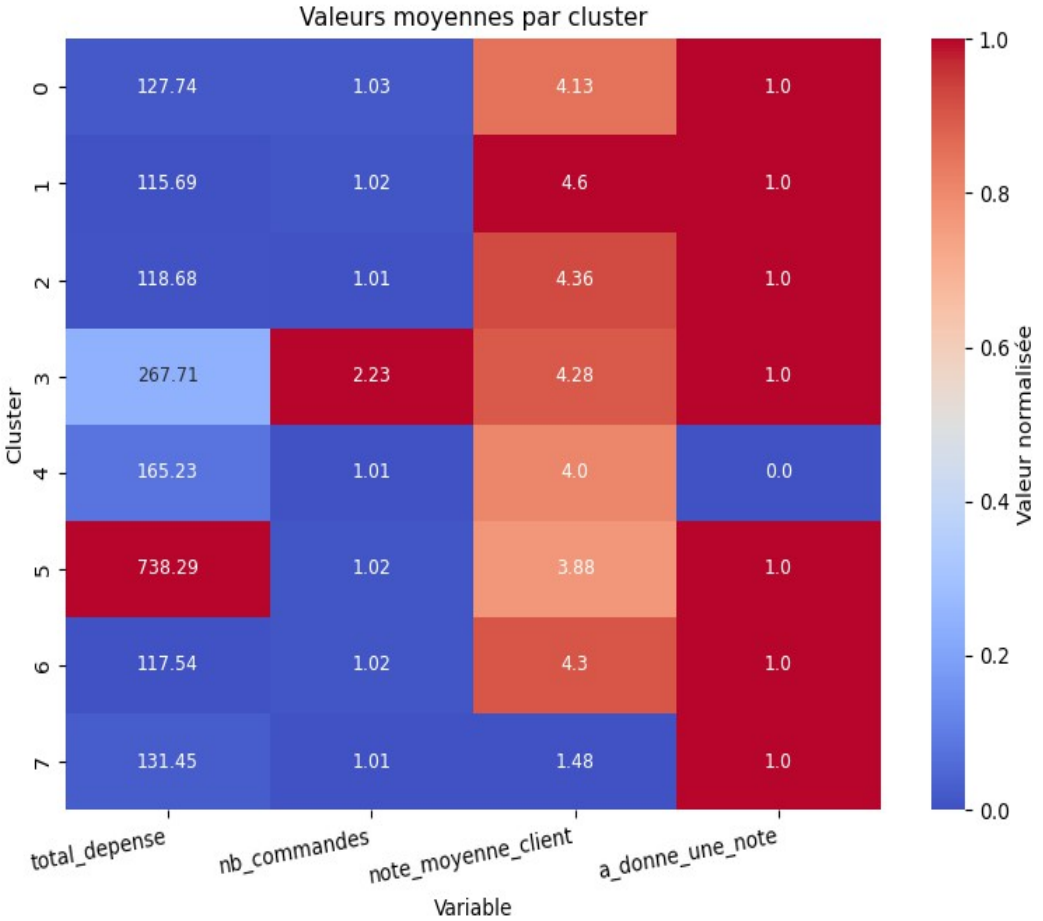
Les effectifs sont suffisamment répartis pour permettre des actions marketing ciblées sur chaque segment.

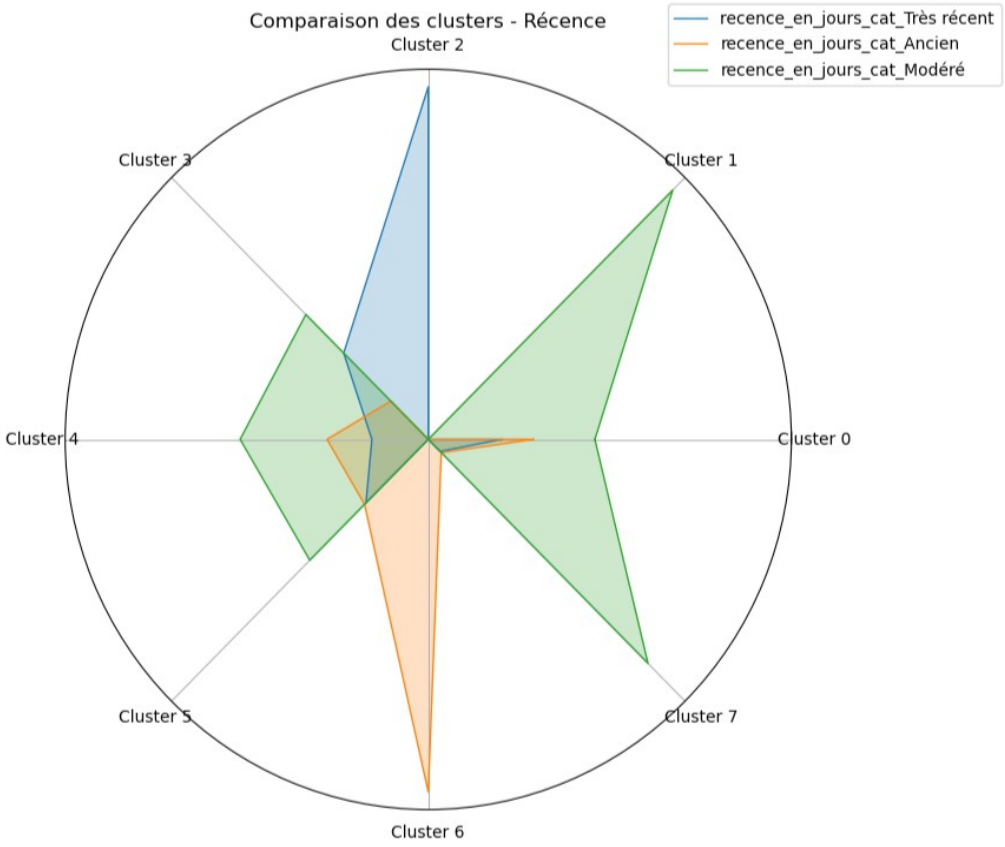




L'ARI moyen de 0,84 indique une bonne stabilité des clusters à l'initialisation.

Différenciation des clusters sur les variables clés





Les clusters présentent des profils différenciés sur les variables clés de valeur, récence et satisfaction, permettant une segmentation marketing fine et exploitable.

Profils clients et recommandations marketing

Cluster	Nombre de clients	Profil client	Variable(s) discriminante(s)	Action marketing recommandée
0	2115	Satisfaits, récence moyenne, **faibles dépenses**	**Diversité des paiements élevée**	Offres multi-paiements, augmenter panier
1	35158	**Très satisfaits**, récence moyenne, **faibles dépenses**	Note moyenne très élevée	Cibler avec des offres simples et peu coûteuses.
2	21350	**Très récents, très satisfaits**, **faibles dépenses**	Récence très forte, très satisfaits	Inciter au 2e achat avec des offres
3	1396	Récents/moyens, **forte dépense, nombre de commandes élevée**	Nb de commandes élevé, diversité catégories	Fidélisation premium via recommandations personnalisées et avantages exclusifs.
4	601	Anciens/moyens, dépense modérée, **pas de note**	Pas de note donnée, **livraison parfois en retard**	Réactivation via campagnes ciblées et amélioration de l'expérience logistique.
5	3070	**Légèrement insatisfaits**, **très forte dépense**	Dépense très élevée, frais élevés	Résolution des problèmes d'insatisfaction et valorisation via offres VIP ou premium.
6	20985	**Anciens**, satisfaits, **faible dépense**	Ancienneté, livraison fiable	Réactivation avec des offres spéciales pour les inciter à revenir plus souvent.
7	8428	**Insatisfaits**, récence moyenne, **faible dépense**	Note très faible, **retards livraison**	Compensation, améliorer logistique

Maintenance du modèle

Pourquoi mettre à jour la segmentation ?

- Évolution des comportements clients

  Les habitudes d'achat, attentes et besoins changent dans le temps.

- Nouveaux clients et données

  L'arrivée régulière de nouveaux clients modifie la composition des segments.

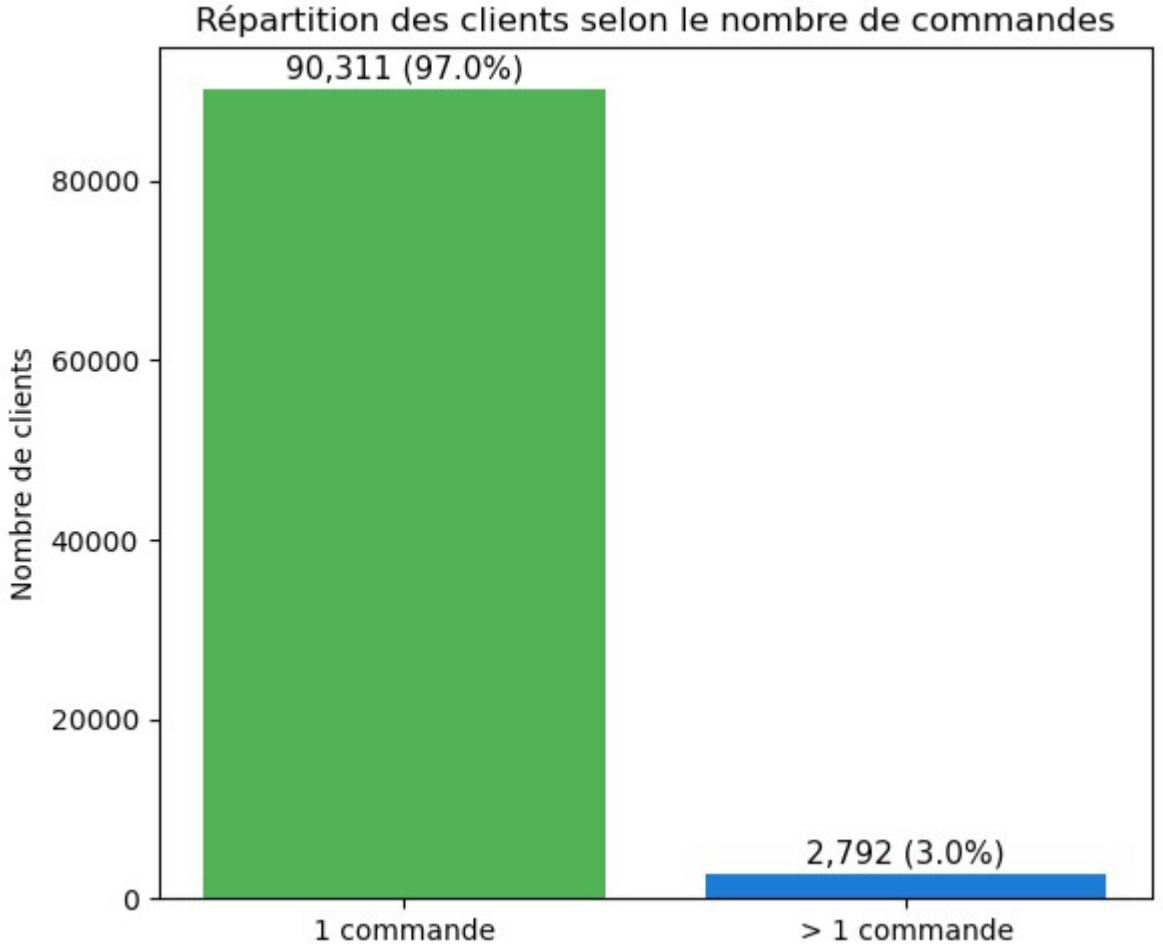
- Changements dans l'offre ou le marché

  Nouveaux produits, promotions, concurrence, contexte économique.

- Maintien de la pertinence marketing

  Une segmentation figée peut devenir obsolète et moins efficace.

Comment déterminer la fréquence de mise à jour ?



- 97% des clients ne commandent qu'une seule fois

→ La base client se renouvelle fortement chaque mois.

- Les nouveaux clients façonnent la dynamique de la clientèle

→ Leur comportement influence la structure des segments.

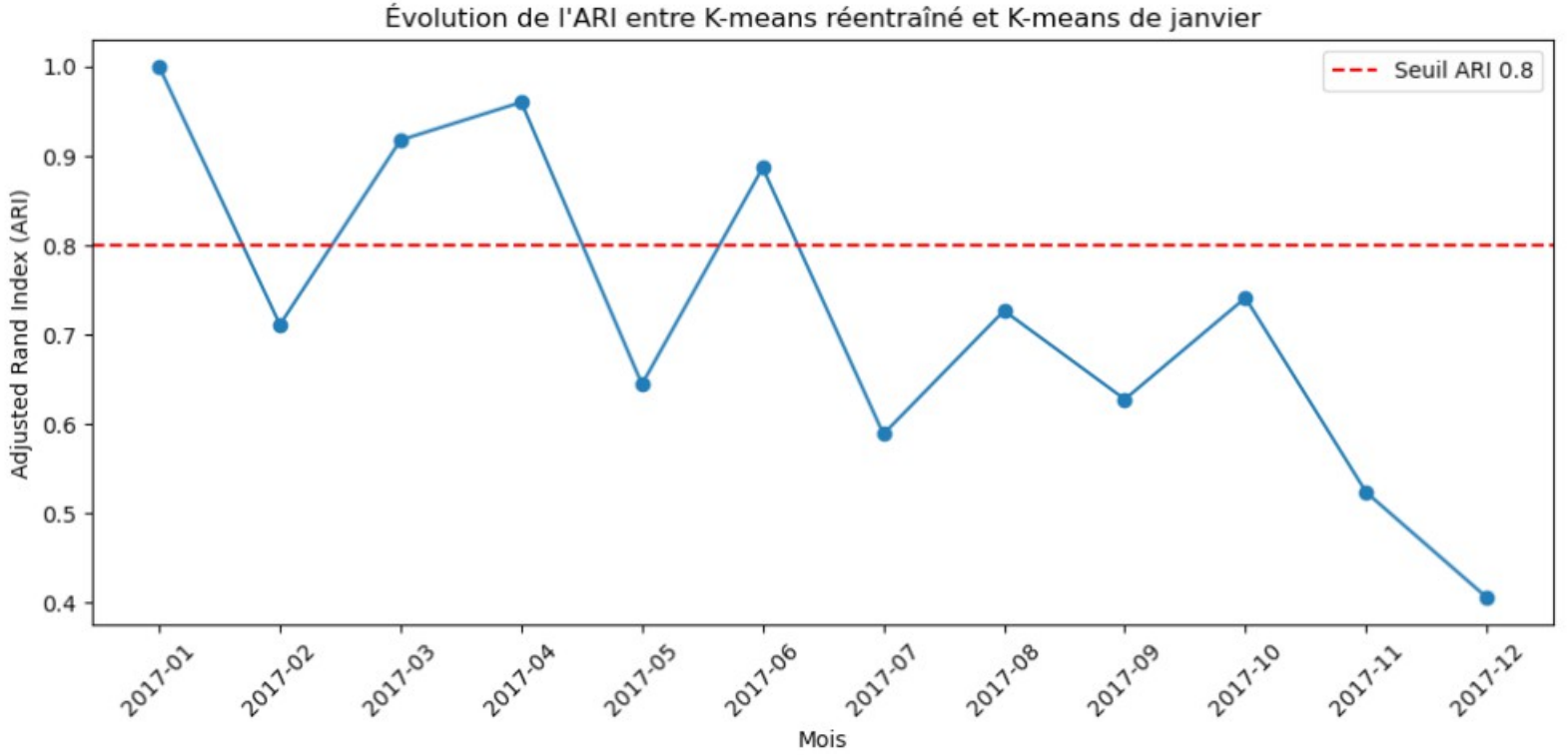
- Analyser uniquement les anciens clients masquerait les évolutions réelles

→ Cumuler plusieurs mois diluerait les changements récents.

Donc, pour déterminer la fréquence de mise à jour :
On analyse chaque mois si la segmentation reste adaptée aux nouveaux clients.

Fréquence optimale de mise à jour : analyse de stabilité (ARI)

- L'ARI est un indicateur qui mesure la similarité entre deux segmentations.
- Il varie entre 0 (aucune similarité) et 1 (identique).
- Plus l'ARI est élevé, plus la segmentation reste stable dans le temps.



Fréquence de mise à jour recommandée : mensuelle

Stratégie d'ajout d'un nouveau client

Étapes principales :

- Collecte & préparation : Récupérer les données du nouveau client et appliquer les mêmes traitements que pour les clients existants.
- Attribution du cluster : Utiliser le modèle K-means entraîné pour prédire le cluster du client.

À retenir :

- L’ajout de nouveaux clients peut se faire en continu, même si la réévaluation globale du clustering est mensuelle.
- Cela permet d’actualiser la segmentation en temps réel et de garder une vision toujours pertinente.

Limitation actuelle :

- Certaines étapes restent manuelles, mais une automatisation est envisageable à terme.

Conclusion

- Les clusters sont globalement bien séparés, ce qui confirme la qualité de la segmentation.
- Quelques chevauchements mineurs existent, reflétant des profils clients proches ou des comportements partagés.
- Le score de silhouette de 0,44 indique une séparation correcte, mais montre aussi que certains groupes restent proches : la segmentation est utile, mais perfectible.
- Elle pourra être améliorée ultérieurement en intégrant de nouvelles données ou en testant d’autres méthodes.