



# Segmentez des clients d'un site e-commerce

Projet 5 - Julien Agneray

# Introduction et contexte



**Présentation d'Olist :** Olist est une plateforme brésilienne de marketplace qui connecte des vendeurs et des clients dans un vaste réseau de commerce en ligne.

**Objectif du projet :** L'objectif de ce projet est de développer une segmentation client efficace afin de mieux comprendre et cibler les différents segments de clients pour des stratégies marketing optimisées.

**Importance de la segmentation :** Une segmentation client bien pensée permet à Olist de personnaliser les offres, d'augmenter l'engagement client, et d'optimiser les efforts marketing.



# Sommaire

1. Exploration des données
2. Segmentation client
3. Choix de la méthode de clustering
4. Poursuite de la segmentation avec la méthode choisie
5. Evolution de la stabilité des clusters
6. Conclusion générale
7. Questions

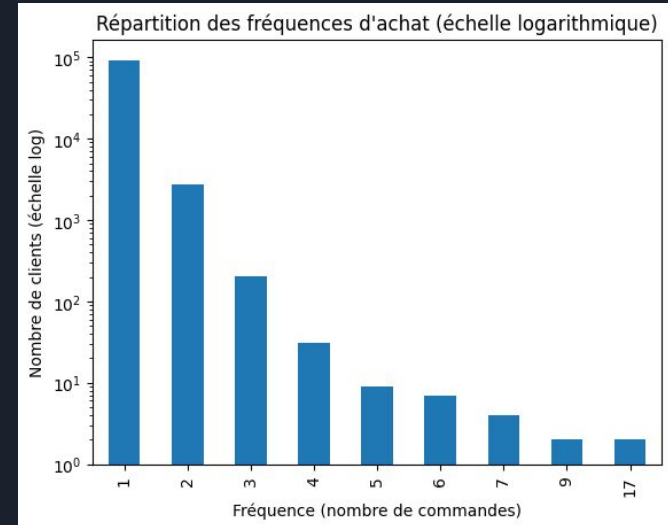
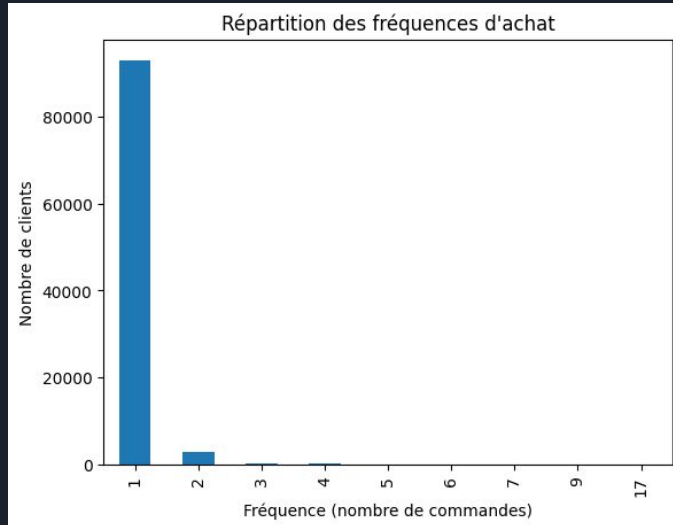
# Exploration des données

Le dataset provient d'Olist et fournit une base de données anonymisée comportant des informations sur l'historique des commandes, les produits achetés, les commentaires de satisfaction, et la localisation des clients depuis janvier 2017.

<table><tr><td>translation</td></tr><tr><td>123 index</td></tr><tr><td>A-Z product_category_name</td></tr><tr><td>A-Z product_category_name_english</td></tr></table>	translation	123 index	A-Z product_category_name	A-Z product_category_name_english	<table><tr><td>sellers</td></tr><tr><td>123 index</td></tr><tr><td>A-Z seller_id</td></tr><tr><td>123 seller_zip_code_prefix</td></tr><tr><td>A-Z seller_city</td></tr><tr><td>A-Z seller_state</td></tr></table>	sellers	123 index	A-Z seller_id	123 seller_zip_code_prefix	A-Z seller_city	A-Z seller_state	<table><tr><td>customers</td></tr><tr><td>123 index</td></tr><tr><td>A-Z customer_id</td></tr><tr><td>A-Z customer_unique_id</td></tr><tr><td>123 customer_zip_code_prefix</td></tr><tr><td>A-Z customer_city</td></tr><tr><td>A-Z customer_state</td></tr></table>	customers	123 index	A-Z customer_id	A-Z customer_unique_id	123 customer_zip_code_prefix	A-Z customer_city	A-Z customer_state	<table><tr><td>geoloc</td></tr><tr><td>123 index</td></tr><tr><td>123 geolocation_zip_code_prefix</td></tr><tr><td>123 geolocation_lat</td></tr><tr><td>123 geolocation_lng</td></tr><tr><td>A-Z geolocation_city</td></tr><tr><td>A-Z geolocation_state</td></tr></table>	geoloc	123 index	123 geolocation_zip_code_prefix	123 geolocation_lat	123 geolocation_lng	A-Z geolocation_city	A-Z geolocation_state	<table><tr><td>order_pymts</td></tr><tr><td>123 index</td></tr><tr><td>A-Z order_id</td></tr><tr><td>123 payment_sequential</td></tr><tr><td>A-Z payment_type</td></tr><tr><td>123 payment_installments</td></tr><tr><td>123 payment_value</td></tr></table>	order_pymts	123 index	A-Z order_id	123 payment_sequential	A-Z payment_type	123 payment_installments	123 payment_value							
translation																																										
123 index																																										
A-Z product_category_name																																										
A-Z product_category_name_english																																										
sellers																																										
123 index																																										
A-Z seller_id																																										
123 seller_zip_code_prefix																																										
A-Z seller_city																																										
A-Z seller_state																																										
customers																																										
123 index																																										
A-Z customer_id																																										
A-Z customer_unique_id																																										
123 customer_zip_code_prefix																																										
A-Z customer_city																																										
A-Z customer_state																																										
geoloc																																										
123 index																																										
123 geolocation_zip_code_prefix																																										
123 geolocation_lat																																										
123 geolocation_lng																																										
A-Z geolocation_city																																										
A-Z geolocation_state																																										
order_pymts																																										
123 index																																										
A-Z order_id																																										
123 payment_sequential																																										
A-Z payment_type																																										
123 payment_installments																																										
123 payment_value																																										
<table><tr><td>order_items</td></tr><tr><td>123 index</td></tr><tr><td>A-Z order_id</td></tr><tr><td>123 order_item_id</td></tr><tr><td>A-Z product_id</td></tr><tr><td>A-Z seller_id</td></tr><tr><td>A-Z shipping_limit_date</td></tr><tr><td>123 price</td></tr><tr><td>123 freight_value</td></tr></table>	order_items	123 index	A-Z order_id	123 order_item_id	A-Z product_id	A-Z seller_id	A-Z shipping_limit_date	123 price	123 freight_value	<table><tr><td>order_reviews</td></tr><tr><td>123 index</td></tr><tr><td>A-Z review_id</td></tr><tr><td>A-Z order_id</td></tr><tr><td>123 review_score</td></tr><tr><td>A-Z review_comment_title</td></tr><tr><td>A-Z review_comment_message</td></tr><tr><td>A-Z review_creation_date</td></tr><tr><td>A-Z review_answer_timestamp</td></tr></table>	order_reviews	123 index	A-Z review_id	A-Z order_id	123 review_score	A-Z review_comment_title	A-Z review_comment_message	A-Z review_creation_date	A-Z review_answer_timestamp	<table><tr><td>orders</td></tr><tr><td>123 index</td></tr><tr><td>A-Z order_id</td></tr><tr><td>A-Z customer_id</td></tr><tr><td>A-Z order_status</td></tr><tr><td>A-Z order_purchase_timestamp</td></tr><tr><td>A-Z order_approved_at</td></tr><tr><td>A-Z order_delivered_carrier_date</td></tr><tr><td>A-Z order_delivered_customer_date</td></tr><tr><td>A-Z order_estimated_delivery_date</td></tr></table>	orders	123 index	A-Z order_id	A-Z customer_id	A-Z order_status	A-Z order_purchase_timestamp	A-Z order_approved_at	A-Z order_delivered_carrier_date	A-Z order_delivered_customer_date	A-Z order_estimated_delivery_date	<table><tr><td>products</td></tr><tr><td>123 index</td></tr><tr><td>A-Z product_id</td></tr><tr><td>A-Z product_category_name</td></tr><tr><td>123 product_name_lenght</td></tr><tr><td>123 product_description_lenght</td></tr><tr><td>123 product_photos_qty</td></tr><tr><td>123 product_weight_g</td></tr><tr><td>123 product_length_cm</td></tr><tr><td>123 product_height_cm</td></tr><tr><td>123 product_width_cm</td></tr></table>	products	123 index	A-Z product_id	A-Z product_category_name	123 product_name_lenght	123 product_description_lenght	123 product_photos_qty	123 product_weight_g	123 product_length_cm	123 product_height_cm	123 product_width_cm
order_items																																										
123 index																																										
A-Z order_id																																										
123 order_item_id																																										
A-Z product_id																																										
A-Z seller_id																																										
A-Z shipping_limit_date																																										
123 price																																										
123 freight_value																																										
order_reviews																																										
123 index																																										
A-Z review_id																																										
A-Z order_id																																										
123 review_score																																										
A-Z review_comment_title																																										
A-Z review_comment_message																																										
A-Z review_creation_date																																										
A-Z review_answer_timestamp																																										
orders																																										
123 index																																										
A-Z order_id																																										
A-Z customer_id																																										
A-Z order_status																																										
A-Z order_purchase_timestamp																																										
A-Z order_approved_at																																										
A-Z order_delivered_carrier_date																																										
A-Z order_delivered_customer_date																																										
A-Z order_estimated_delivery_date																																										
products																																										
123 index																																										
A-Z product_id																																										
A-Z product_category_name																																										
123 product_name_lenght																																										
123 product_description_lenght																																										
123 product_photos_qty																																										
123 product_weight_g																																										
123 product_length_cm																																										
123 product_height_cm																																										
123 product_width_cm																																										

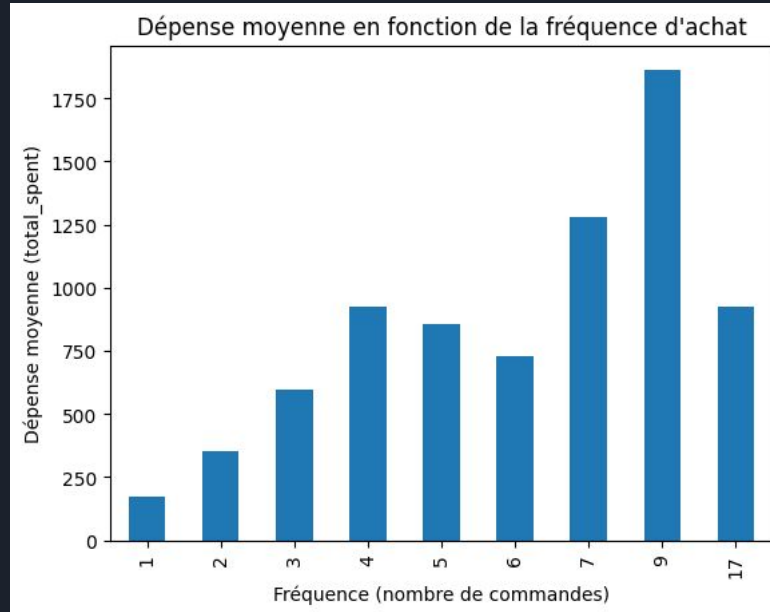
# Analyse exploratoire

Répartition des fréquences d'achat : La plupart des clients réalisent une seule commande, ce qui indique une base large de clients occasionnels avec quelques clients fidèles.



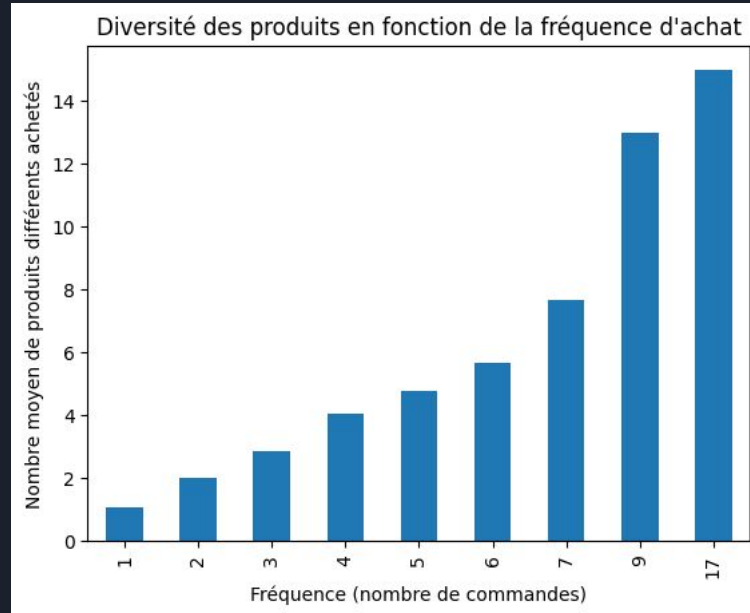
# Analyse exploratoire

Dépense moyenne par fréquence d'achat : Les clients qui achètent fréquemment dépensent plus en moyenne, suggérant un lien entre la fidélité et les revenus générés.



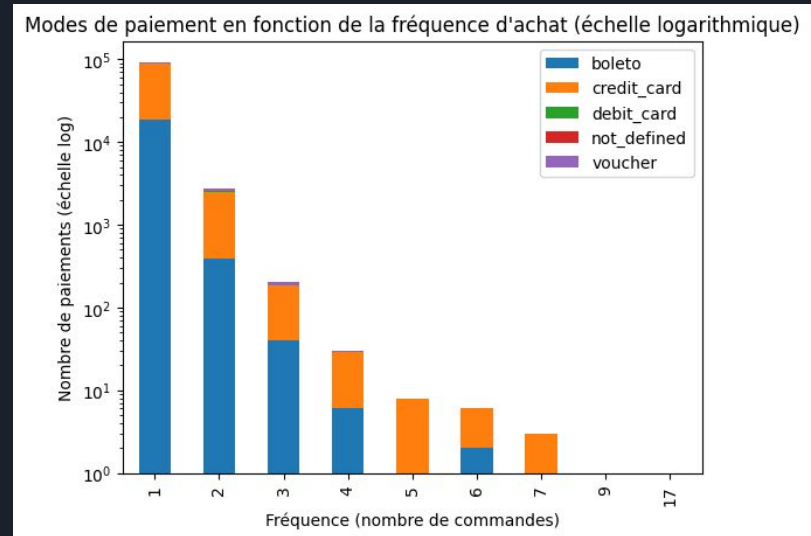
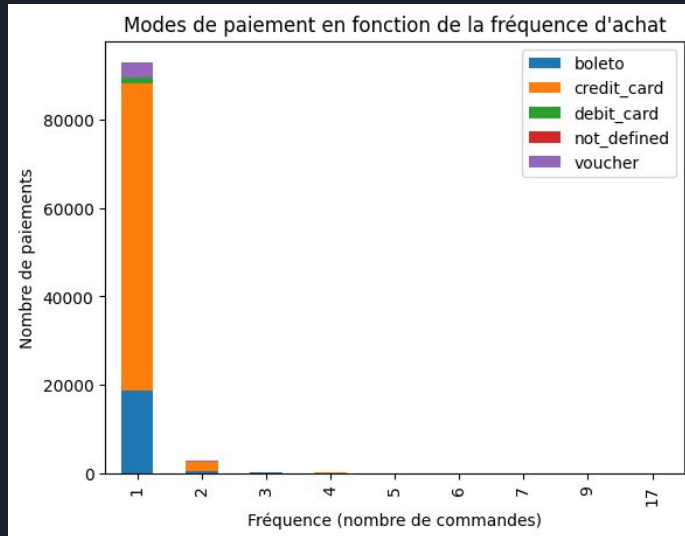
# Analyse exploratoire

Diversité des produits par fréquence d'achat : Les clients réguliers achètent une plus grande diversité de produits, ce qui indique qu'ils explorent davantage l'offre d'Olist.



# Analyse exploratoire

**Modes de paiement et fréquence d'achat :** La carte de crédit est le mode de paiement favori. Le boleto, un moyen de paiement très populaire au Brésil, arrive en seconde position. Il s'agit d'un ticket de paiement que les clients peuvent régler en ligne ou en espèces dans les points de vente.







# Segmentation client

## Introduction

**Objectif de la modélisation :** L'objectif de cette modélisation est de segmenter les clients d'Olist en groupes distincts afin de mieux comprendre leurs comportements et de cibler efficacement les stratégies marketing.

- **K-Means :** "Algorithme de partitionnement basé sur la distance, idéal pour identifier des groupes homogènes de clients."
- **DBSCAN :** "Algorithme de clustering basé sur la densité, capable de détecter les clusters denses et d'identifier les outliers."
- **Agglomerative Clustering :** "Méthode de clustering hiérarchique qui regroupe progressivement les clients en fonction de leurs similitudes."

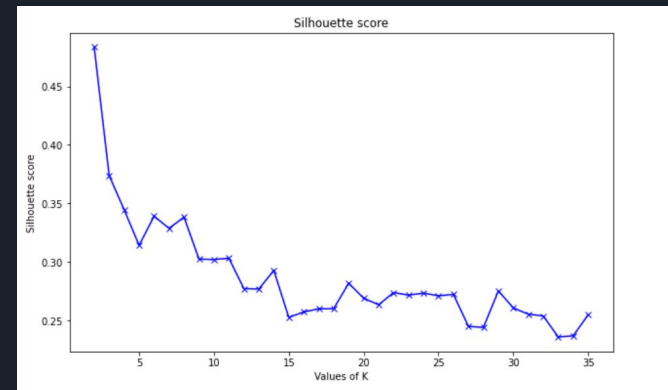
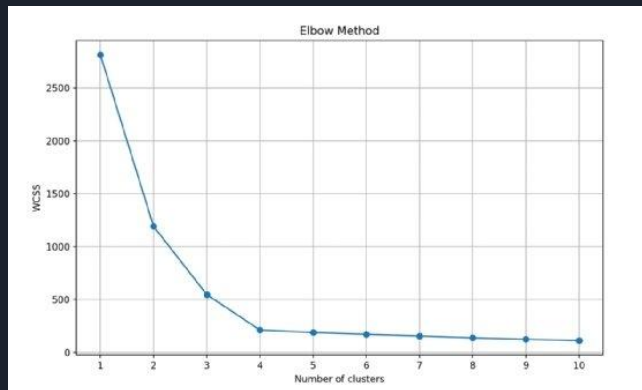
**Processus de sélection :** En explorant ces différentes approches, j'ai pu comparer les résultats, tester la cohérence des clusters, et sélectionner la méthode qui apporte les segments les plus significatifs pour Olist.

# Segmentation client

## Choix du nombre optimal de clusters

La méthode du coude consiste à observer la variation de l'inertie (ou SSE) par rapport au nombre de clusters. Le point où la réduction de l'inertie devient moins significative est appelé le coude et suggère un bon compromis entre cohésion et nombre de clusters.

Le silhouette score varie entre -1 et 1 et mesure la similarité d'un point avec les autres points de son cluster par rapport aux points des clusters voisins. Un score proche de 1 indique que les clusters sont bien séparés et cohérents.

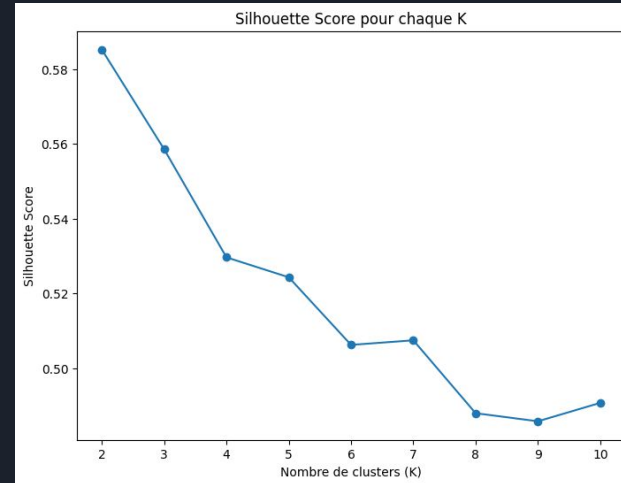
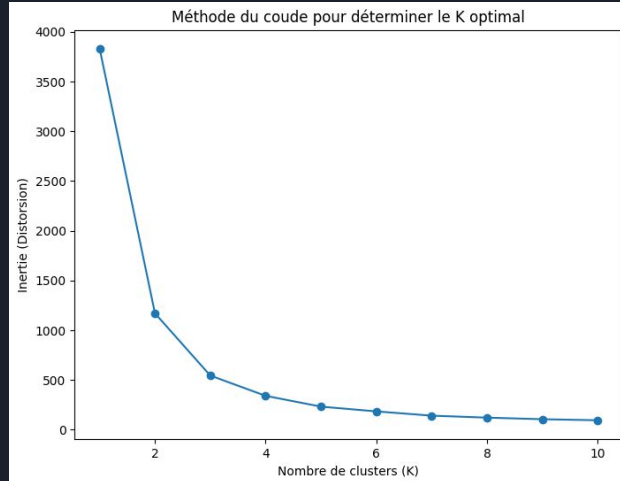


# Segmentation client

## Segmentation par K-Means avec les variables RFM

Les variables RFM permettent de capturer les comportements d'achat des clients en fonction de la récence des achats, leur fréquence, et les montants dépensés.

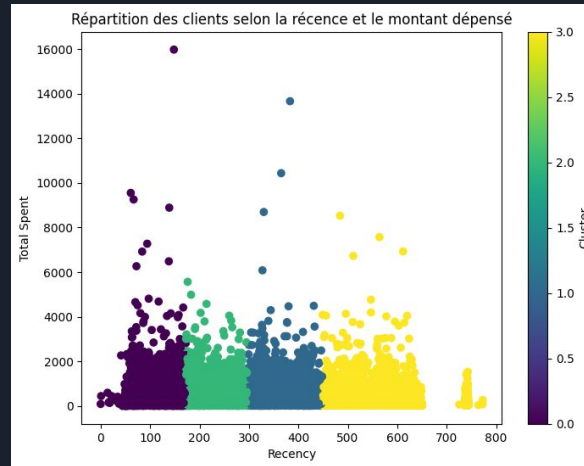
Le coude a été identifié à 4 clusters, ce qui est confirmé par un silhouette score stable et satisfaisant.



# Segmentation client

## Segmentation par K-Means avec les variables RFM

Le graphique illustre un bon clustering, avec des clusters bien distincts en termes de récence et de montant dépensé. Chaque groupe est bien séparé, ce qui démontre que le K-Means a efficacement segmenté les clients en fonction de leurs comportements d'achat. Cette séparation claire entre les clusters confirme la pertinence de cette méthode pour identifier des segments clients distincts et exploitables.





# Segmentation client

## Segmentation par K-Means avec les variables RFM

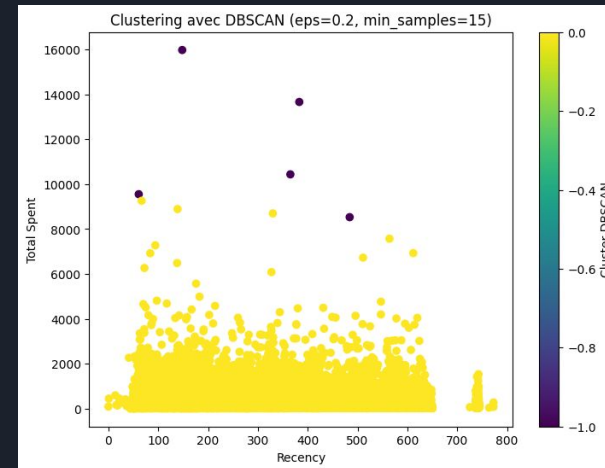
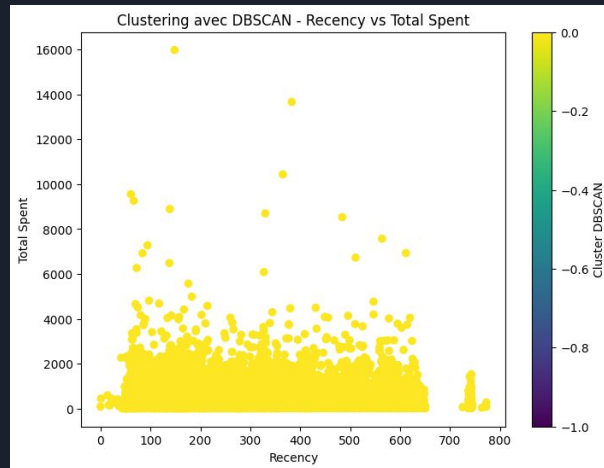
Cluster	Caractéristique principale	Récence moyenne (jours)	Fréquence moyenne	Montant moyen (€)	Score de satisfaction moyen	Nombre de clients
0	Clients les plus récents, satisfaction élevée, dépense la plus élevée	110.88	1.04	183.65	4.24	25,874
1	Clients plutôt anciens, satisfaction moyenne, dépense élevée	363.40	1.03	178.81	4.07	25,171
2	Clients plutôt récents, satisfaction la plus basse, dépense la plus modeste	232.64	1.04	172.16	3.95	27,811
3	Clients très anciens, satisfaction moyenne, dépense modérée	534.16	1.03	176.52	4.09	17,239

# Segmentation client

## Segmentation par DBSCAN

DBSCAN identifie des clusters basés sur la densité des données et permet de détecter les outliers, ce qui est utile pour repérer des comportements atypiques chez certains clients.

DBSCAN a été testé avec deux configurations différentes pour explorer les regroupements possibles dans les données en ajustant les paramètres.





# Segmentation client

## Segmentation par DBSCAN

Dans le **premier essai**, DBSCAN a regroupé la majorité des clients dans **un seul cluster**, suggérant des comportements d'achat globalement similaires.

Pour le **deuxième essai**, j'ai ajusté les paramètres afin de mieux distinguer les variations de comportements. Cette approche a permis d'identifier un **petit nombre de clients aux comportements atypiques**.

Dans le contexte de nos données, **DBSCAN ne semble pas être une méthode pertinente** pour segmenter les clients. L'algorithme ne parvient pas à identifier des regroupements naturels dans les données, et la majorité des points sont regroupés dans un seul cluster.

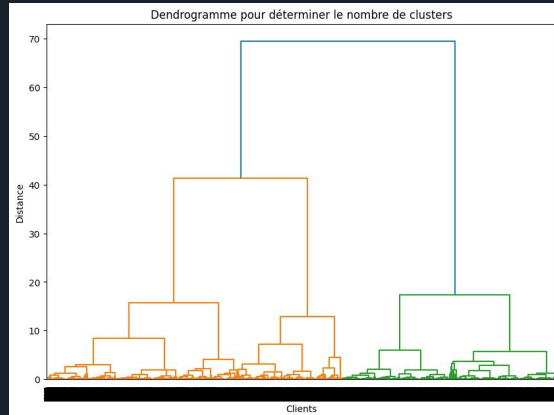
Cela peut s'expliquer par la **structure des données RFM**, qui ne présente peut-être pas la densité requise pour que DBSCAN puisse identifier des clusters significatifs.

# Segmentation client

## Segmentation par Agglomerative Clustering

L'Agglomerative Clustering crée une hiérarchie de regroupements en fusionnant progressivement les clients selon leurs similarités, ce qui permet de visualiser les relations entre eux sous forme de dendrogramme.

Le dendrogramme montre la hiérarchie des regroupements possibles. J'ai choisi de couper à trois clusters, car ce point de coupure minimise la distance entre les regroupements tout en préservant des groupes distincts et significatifs.

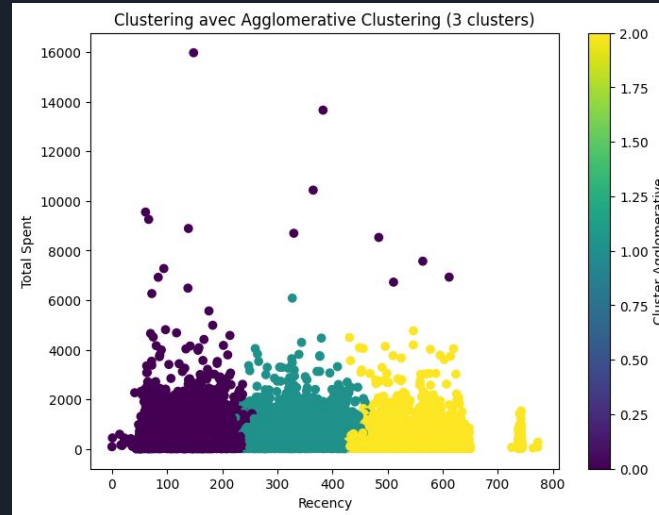




# Segmentation client

## Segmentation par Agglomerative Clustering

Agglomerative Clustering a permis d'identifier 3 clusters en capturant les relations hiérarchiques. Le temps de calcul s'est avéré raisonnable sur les 96 095 lignes, montrant que cette méthode est praticable dans notre contexte. Bien qu'efficace, elle reste plus coûteuse en ressources comparée à K-Means, et serait moins adaptée si le dataset devenait plus volumineux.





# Choix de la méthode de clustering

Après avoir comparé **K-Means**, **DBSCAN**, et **Agglomerative Clustering**, K-Means a été retenu pour la segmentation client, en raison des avantages suivants :

1. **K-Means est efficace et rapide**, particulièrement adapté pour traiter de grands volumes de données comme celui de notre projet. Sa simplicité d'implémentation et sa faible demande en ressources le rendent **facilement reproductible et scalable**.
2. **DBSCAN n'a pas réussi à segmenter les clients** de manière utile dans ce contexte, probablement en raison de la distribution uniforme des données RFM, qui ne révèle pas de densités de clusters exploitables.
3. **Agglomerative Clustering offre une bonne visualisation** des relations hiérarchiques, mais à un **coût en ressources plus élevé** sans apporter de segmentation plus exploitable que K-Means. Pour des projets nécessitant rapidité et efficacité, K-Means reste plus approprié.

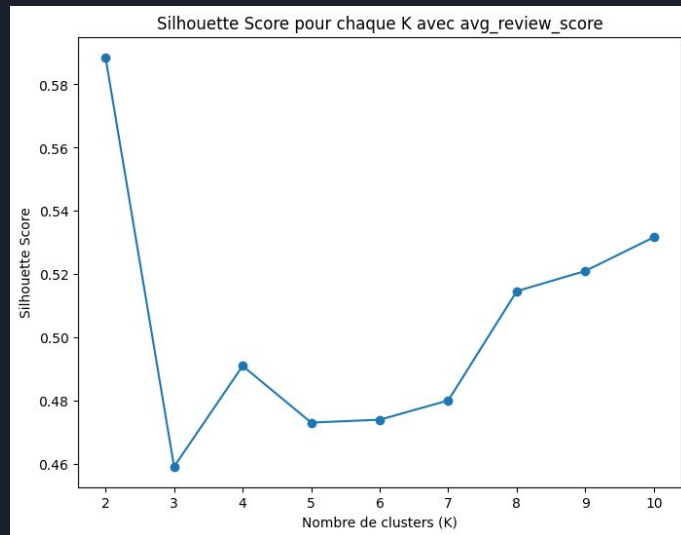
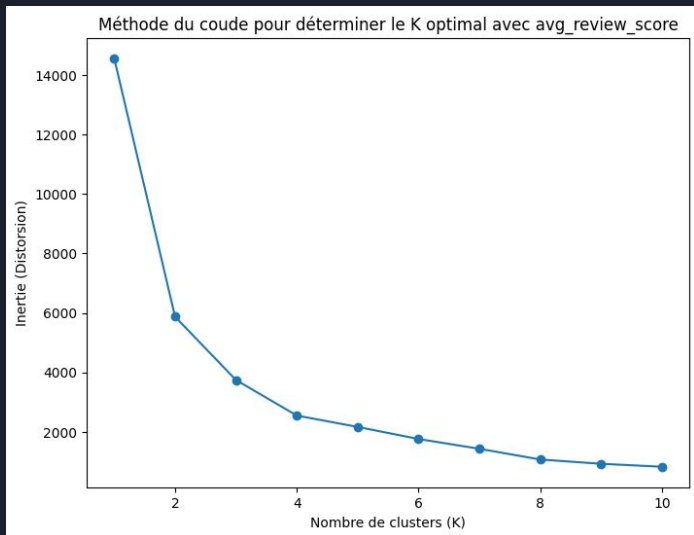
En résumé, **K-Means a été choisi** pour sa **rapidité** et son **adaptabilité** aux grandes bases de données, offrant une segmentation utilisable sans nécessiter de ressources excessives.

# Poursuite de la segmentation

## K-Means avec RFM + Satisfaction

**Objectif :** Intégrer la satisfaction client dans la segmentation pour capturer une dimension qualitative, en plus des comportements d'achat.

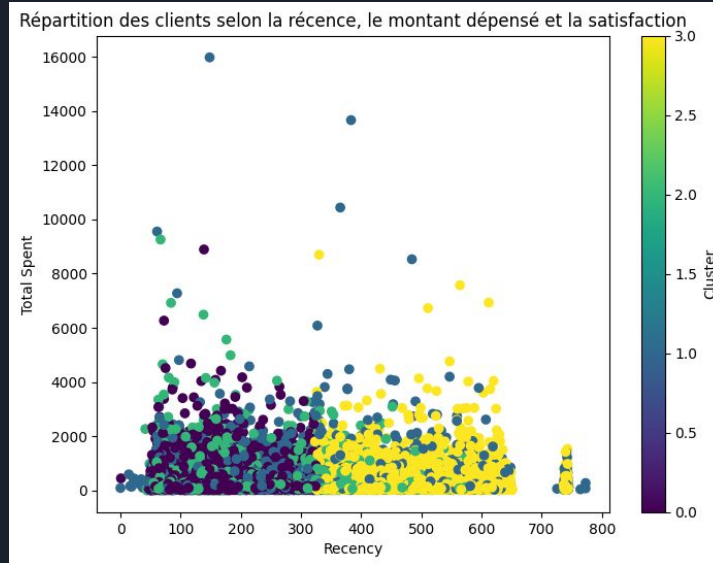
J'ai utilisé la méthode du coude et le Silhouette Score pour choisir le nombre de 4 clusters.



# Poursuite de la segmentation

## K-Means avec RFM + Satisfaction

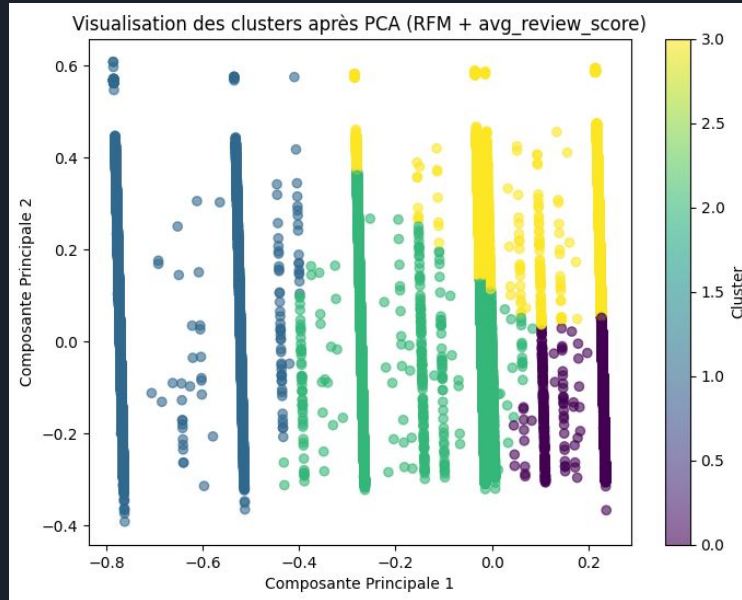
Le graphique montre la répartition des clients selon leur récence et leur montant dépensé, avec les clusters représentés par les différentes couleurs. On observe ainsi des tendances distinctes selon les clusters, notamment parmi les clients qui dépensent le plus ou ceux qui reviennent plus fréquemment.



# Poursuite de la segmentation

## K-Means avec RFM + Satisfaction

La PCA simplifie la visualisation en réduisant les dimensions, tout en préservant les différences entre les clusters.





# Poursuite de la segmentation

## K-Means avec RFM + Satisfaction

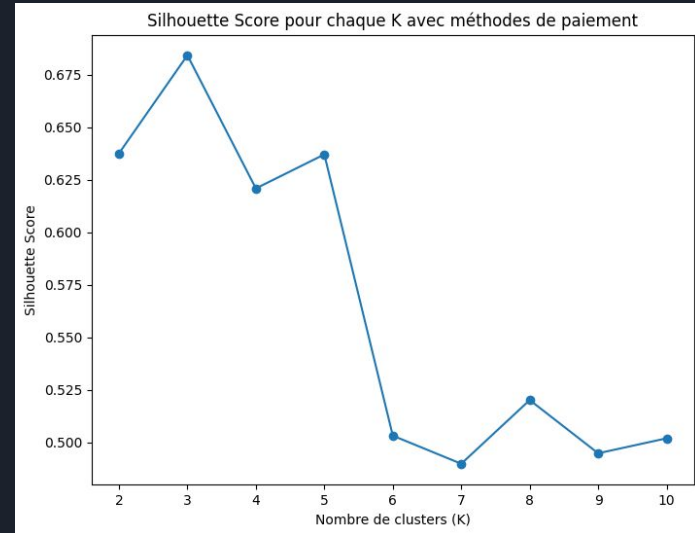
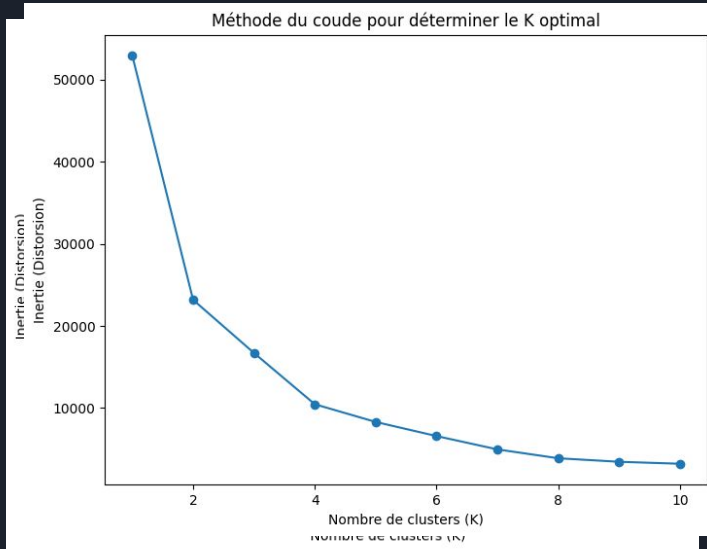
Cluster	Caractéristique principale	Récence moyenne (jours)	Fréquence moyenne	Montant moyen (€)	Score de satisfaction moyen	Nombre de clients
0	Clients récents, très satisfaits, dépense modeste	182.78	1.04	167.03	4.99	34,314
1	Clients anciens, satisfaction très basse, dépense élevée	294.94	1.03	223.36	1.22	13,916
2	Clients moyennement récents, satisfaction moyenne, dépense modérée	239.95	1.04	172.37	3.65	21,825
3	Clients très anciens, satisfaction élevée, dépense modérée	463.94	1.03	172.10	4.78	26,040

# Poursuite de la segmentation

## K-Means avec RFM + Satisfaction + Méthodes de paiement

**Objectif :** Intégrer les méthodes de paiement dans la segmentation pour mieux comprendre les préférences de paiement.

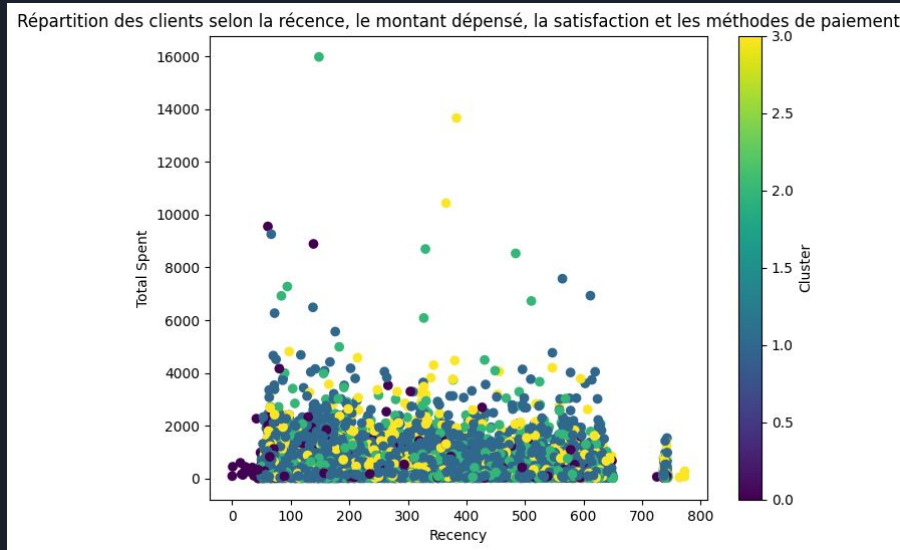
J'ai utilisé la méthode du coude et le Silhouette Score pour choisir le nombre de 4 clusters.



# Poursuite de la segmentation

## K-Means avec RFM + Satisfaction + Méthodes de paiement

Le graphique montre la segmentation des clients selon la récence, le montant dépensé, la satisfaction et les méthodes de paiement, révélant des comportements distincts dans chaque cluster.

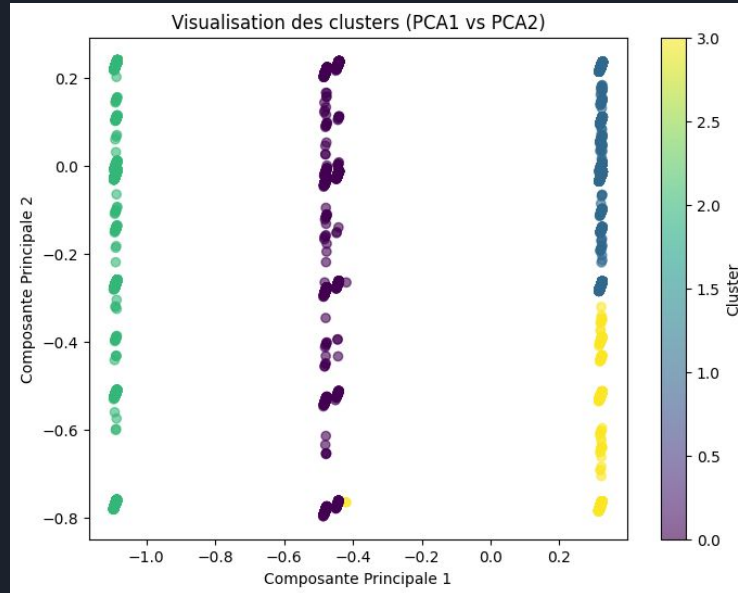




# Poursuite de la segmentation

## K-Means avec RFM + Satisfaction + Méthodes de paiement

La PCA montre des clusters bien distincts, confirmant la pertinence des segments identifiés.






# Poursuite de la segmentation

## K-Means avec RFM + Satisfaction + Méthodes de paiement

Cluster	Caractéristique principale	Récence moyenne (jours)	Fréquence moyenne	Montant moyen (€)	Score de satisfaction moyen	Méthode de paiement	Nombre de clients
0	Très petit cluster, satisfaction élevée, dépense modérée	277.30	1.05	155.65	4.06	70.9% Voucher, 29.1% Carte de débit	5,186
1	Très gros cluster, très satisfaits, dépense moyenne	285.23	1.04	176.30	4.57	100% Carte de crédit	61,399
2	Cluster moyen, assez satisfaits, dépense modérée	298.30	1.03	159.85	4.08	100% Boleto	19,039
3	Petit cluster, très insatisfaits, dépense élevée	292.64	1.03	229.99	1.23	100% Carte de crédit	10,471



# Poursuite de la segmentation

## Conclusion

L'analyse **K-Means** a été menée en utilisant **trois ensembles de variables** : les données **RFM** seules, **RFM combinées** avec la **satisfaction client**, et **RFM avec les méthodes de paiement**. Les résultats montrent que la segmentation RFM, basée uniquement sur la récence, la fréquence et le montant dépensé, offre la meilleure cohérence et des clusters bien différenciés.

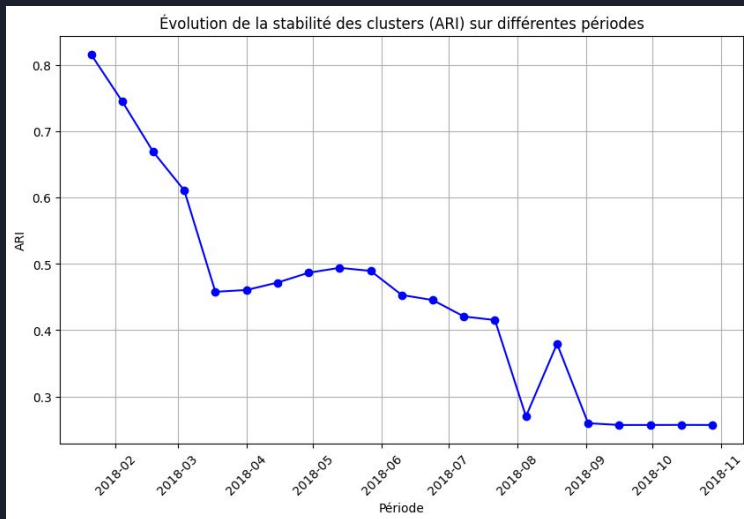
Bien que l'ajout de **variables supplémentaires** ait apporté de nouvelles perspectives, ces informations ont parfois **complexifié la segmentation** sans fournir de distinctions claires supplémentaires. La **segmentation RFM reste donc plus directement exploitable** pour cibler les comportements d'achat, avec des segments clairs et alignés sur les objectifs marketing.

Sur cette base, j'aborde maintenant l'évolution de la stabilité des clusters dans le temps, pour vérifier la durabilité de cette segmentation et affiner la fréquence de mise à jour nécessaire.

# Evolution de la stabilité des clusters

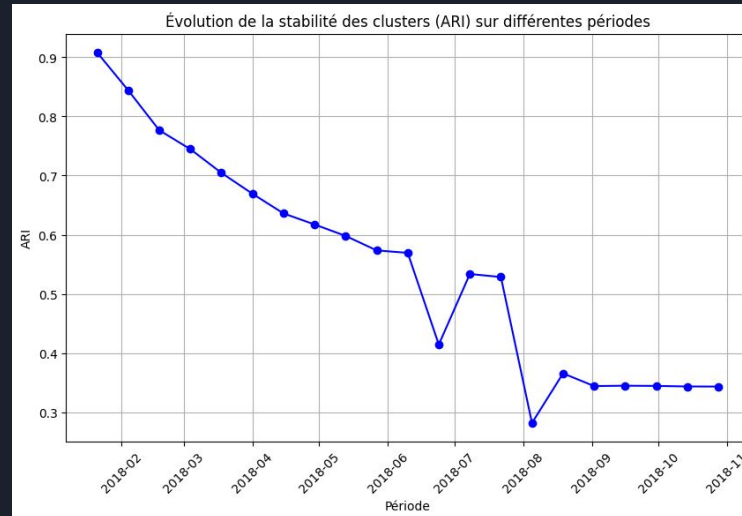
## MinMaxScaler

En normalisant les données avec **MinMaxScaler**, nous avons observé une **dégradation** progressive de la **stabilité des clusters** sur plusieurs semaines. L'Adjusted Rand Index (ARI) diminue au fil du temps, indiquant que les comportements clients évoluent et que **les clusters deviennent moins cohérents**.



# Evolution de la stabilité des clusters StandardScaler

En normalisant les données avec **StandardScaler**, l'évolution de l'ARI montre une **stabilité plus élevée** des clusters au fil du temps. La décroissance de l'ARI est moins marquée, indiquant que cette méthode de normalisation aide à **maintenir des segments plus cohérents** malgré les variations dans les comportements clients.





# Evolution de la stabilité des clusters

## Conclusion sur la simulation

L'analyse de la stabilité des clusters a permis d'identifier les meilleures pratiques pour maintenir une segmentation client efficace dans le temps. Les principaux enseignements incluent :

- **StandardScaler pour la stabilité** : Les résultats montrent que le StandardScaler offre une meilleure stabilité des clusters, avec un ARI qui diminue plus progressivement que le MinMaxScaler. Cela permet de conserver une cohérence dans la segmentation.
- **Fréquence de mise à jour** : Pour garantir une segmentation pertinente, je recommande de réentraîner le modèle tous les 1 mois et demi, afin de maintenir l'ARI au-dessus de 0.8. Cela assurera une réactivité aux changements dans les comportements clients.

En suivant ces recommandations, il sera possible de maintenir une segmentation dynamique, adaptée aux évolutions des comportements d'achat, pour une meilleure efficacité des stratégies marketing.



# Conclusion générale

Ce projet a permis de **développer une segmentation client détaillée** pour Olist, en testant différentes méthodes de clustering. Les analyses ont mis en lumière les comportements d'achat et les préférences des clients, révélant **des segments distincts et exploitables**. Le choix de K-Means s'est avéré le plus adapté pour notre volume de données, en offrant un bon équilibre entre **performance et simplicité**.

La simulation de la stabilité des clusters a souligné **l'importance de la normalisation** avec le **StandardScaler** pour maintenir des segments cohérents dans le temps. **En réentraînant le modèle tous les 1 mois et demi**, Olist pourra suivre les évolutions des comportements clients et adapter ses stratégies marketing.

En conclusion, cette segmentation propose une base solide pour mieux cibler les actions marketing et renforcer la fidélité client à long terme.

Avez-vous des questions?

