
Segmentez des clients d'un site e-commerce

Projet 4 du parcours Machine Learning Engineer



Sommaire

01

Introduction

02

Exploration et
analyse

03

Feature
engineering

04

Segmentation

05

Maintenance du
modèle

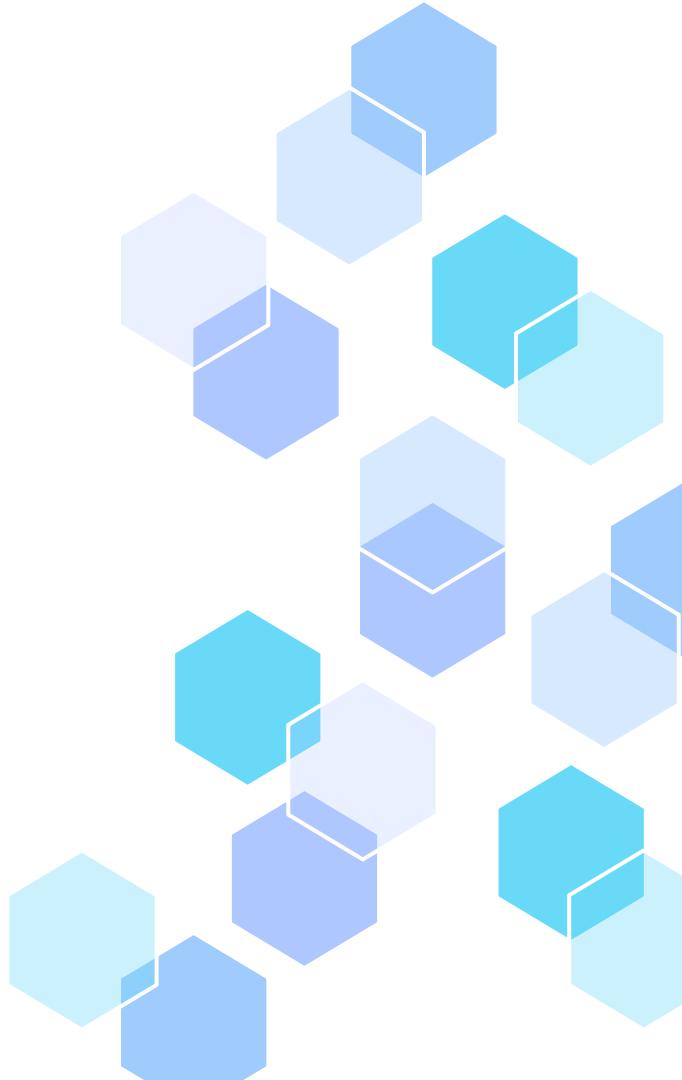
06

Conclusion

Introduction

Cette présentation vise à mettre en lumière les grandes étapes de notre projet pour Olist :

1. Requêtes SQL : Nous avons débuté par la création de la base de données SQL d'Olist. Puis il a fallu extraire les informations demandées
2. Segmentation Client : L'objectif principal est de segmenter les clients d'Olist en clusters distincts, basés sur leurs comportements et leurs caractéristiques, afin de mieux cibler les campagnes marketing.
3. Maintenance du Modèle : Enfin, nous avons abordé la phase de maintenance du modèle, en recommandant une fréquence de mise à jour pour garantir la pertinence et l'efficacité continue de la segmentation client au fil du temps.



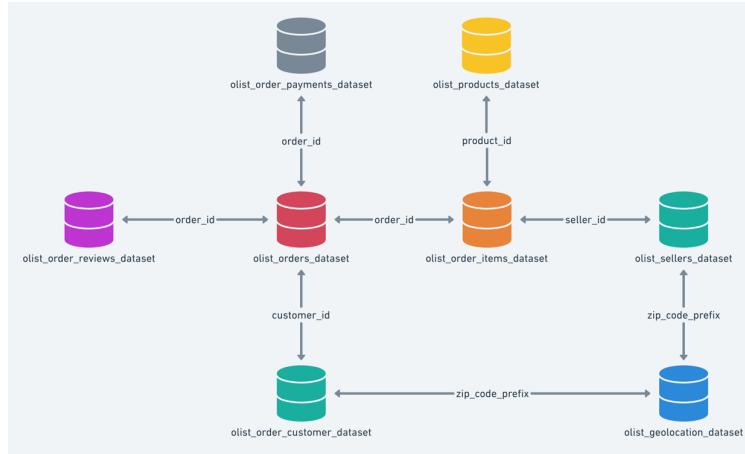
Requêtes SQL

Titre	Requête	Résultat																																																			
Requête 1	En excluant les commandes annulées, quelles sont les commandes récentes de moins de 3 mois que les clients ont reçues avec au moins 3 jours de retard?	321 Commandes																																																			
Requête 2	Qui sont les vendeurs ayant généré un chiffre d'affaires de plus de 100 000 Real sur des commandes livrées via Olist?	<table border="1"><thead><tr><th data-bbox="1288 487 1403 503">seller_id</th><th data-bbox="1403 487 1499 503">total_a...</th><th data-bbox="1499 487 1762 503">total_items</th></tr></thead><tbody><tr><td data-bbox="1288 503 1403 519">Filter...</td><td data-bbox="1403 503 1499 519">Filter...</td><td data-bbox="1499 503 1762 519">Filter...</td></tr><tr><td data-bbox="1288 519 1403 540">1025f0e2d44d7041d6cf58b6550e0bfa</td><td data-bbox="1403 519 1499 540">112211.2</td><td data-bbox="1499 519 1762 540">934</td></tr><tr><td data-bbox="1288 540 1403 560">46dc3b2cc0980fb8ec44634e21d2718e</td><td data-bbox="1403 540 1499 560">120671.5</td><td data-bbox="1499 540 1762 560">504</td></tr><tr><td data-bbox="1288 560 1403 581">4869f7a5dfa277a7dca6462dcf3b52b2</td><td data-bbox="1403 560 1499 581">224921.5</td><td data-bbox="1499 560 1762 581">1135</td></tr><tr><td data-bbox="1288 581 1403 601">4a3ca9315b744ce9f8e9374361493884</td><td data-bbox="1403 581 1499 601">183828.1</td><td data-bbox="1499 581 1762 601">1810</td></tr><tr><td data-bbox="1288 601 1403 622">53243585afdddc2643021fd1853d8905</td><td data-bbox="1403 601 1499 622">203364</td><td data-bbox="1499 601 1762 622">348</td></tr><tr><td data-bbox="1288 622 1403 642">5dccea129747e92ff8ef7a997dc4f8ca</td><td data-bbox="1403 622 1499 642">109116.1</td><td data-bbox="1499 622 1762 642">324</td></tr><tr><td data-bbox="1288 642 1403 663">620c87c171fb2a6dd6e8bb4dec959fc6</td><td data-bbox="1403 642 1499 663">107693.4</td><td data-bbox="1499 642 1762 663">729</td></tr><tr><td data-bbox="1288 663 1403 683">6560211a19b47992c3666cc44a7e94c0</td><td data-bbox="1403 663 1499 683">116660.8</td><td data-bbox="1499 663 1762 683">1905</td></tr><tr><td data-bbox="1288 683 1403 704">7a67c85e85bb2ce8582c35f2203ad736</td><td data-bbox="1403 683 1499 704">139098.8</td><td data-bbox="1499 683 1762 704">1147</td></tr><tr><td data-bbox="1288 704 1403 724">7c67e1448b00f6e969d365cea6b010ab</td><td data-bbox="1403 704 1499 724">140998.6</td><td data-bbox="1499 704 1762 724">995</td></tr><tr><td data-bbox="1288 724 1403 745">7d13fca15225358621be4086e1eb0964</td><td data-bbox="1403 724 1499 745">111212.2</td><td data-bbox="1499 724 1762 745">559</td></tr><tr><td data-bbox="1288 745 1403 765">7e93a43ef30c4f03f38b393420bc753a</td><td data-bbox="1403 745 1499 765">165822.5</td><td data-bbox="1499 745 1762 765">321</td></tr><tr><td data-bbox="1288 765 1403 786">955fee9216a65b617aa5c0531780ce60</td><td data-bbox="1403 765 1499 786">113634.1</td><td data-bbox="1499 765 1762 786">1262</td></tr><tr><td data-bbox="1288 786 1403 806">da8622b14eb17ae2831f4ac5b9dab84a</td><td data-bbox="1403 786 1499 806">141875.4</td><td data-bbox="1499 786 1762 806">1371</td></tr><tr><td data-bbox="1288 806 1403 827">fa1c13f2614d7b5c4749cbc52fecda94</td><td data-bbox="1403 806 1499 827">190917.1</td><td data-bbox="1499 806 1762 827">579</td></tr></tbody></table>	seller_id	total_a...	total_items	Filter...	Filter...	Filter...	1025f0e2d44d7041d6cf58b6550e0bfa	112211.2	934	46dc3b2cc0980fb8ec44634e21d2718e	120671.5	504	4869f7a5dfa277a7dca6462dcf3b52b2	224921.5	1135	4a3ca9315b744ce9f8e9374361493884	183828.1	1810	53243585afdddc2643021fd1853d8905	203364	348	5dccea129747e92ff8ef7a997dc4f8ca	109116.1	324	620c87c171fb2a6dd6e8bb4dec959fc6	107693.4	729	6560211a19b47992c3666cc44a7e94c0	116660.8	1905	7a67c85e85bb2ce8582c35f2203ad736	139098.8	1147	7c67e1448b00f6e969d365cea6b010ab	140998.6	995	7d13fca15225358621be4086e1eb0964	111212.2	559	7e93a43ef30c4f03f38b393420bc753a	165822.5	321	955fee9216a65b617aa5c0531780ce60	113634.1	1262	da8622b14eb17ae2831f4ac5b9dab84a	141875.4	1371	fa1c13f2614d7b5c4749cbc52fecda94	190917.1	579
seller_id	total_a...	total_items																																																			
Filter...	Filter...	Filter...																																																			
1025f0e2d44d7041d6cf58b6550e0bfa	112211.2	934																																																			
46dc3b2cc0980fb8ec44634e21d2718e	120671.5	504																																																			
4869f7a5dfa277a7dca6462dcf3b52b2	224921.5	1135																																																			
4a3ca9315b744ce9f8e9374361493884	183828.1	1810																																																			
53243585afdddc2643021fd1853d8905	203364	348																																																			
5dccea129747e92ff8ef7a997dc4f8ca	109116.1	324																																																			
620c87c171fb2a6dd6e8bb4dec959fc6	107693.4	729																																																			
6560211a19b47992c3666cc44a7e94c0	116660.8	1905																																																			
7a67c85e85bb2ce8582c35f2203ad736	139098.8	1147																																																			
7c67e1448b00f6e969d365cea6b010ab	140998.6	995																																																			
7d13fca15225358621be4086e1eb0964	111212.2	559																																																			
7e93a43ef30c4f03f38b393420bc753a	165822.5	321																																																			
955fee9216a65b617aa5c0531780ce60	113634.1	1262																																																			
da8622b14eb17ae2831f4ac5b9dab84a	141875.4	1371																																																			
fa1c13f2614d7b5c4749cbc52fecda94	190917.1	579																																																			

Requêtes SQL

Titre	Requête	Résultat																		
Requête 3	Qui sont les nouveaux vendeurs (moins de 3 mois d'ancienneté) qui sont déjà très engagés avec la plateforme (ayant déjà vendu plus de 30 produits)?	<table><thead><tr><th>seller_id</th><th>total_amount_...</th><th>total_items_sold</th></tr></thead><tbody><tr><td>240b9776d844d37535668549a39...</td><td>13332.089999999997</td><td>35</td></tr><tr><td>81f89e42267213cb94da7ddc3016...</td><td>3522</td><td>46</td></tr><tr><td>d13e50eaa47b4cbe9eb81465865d...</td><td>6987.14999999999</td><td>67</td></tr></tbody></table>	seller_id	total_amount_...	total_items_sold	240b9776d844d37535668549a39...	13332.089999999997	35	81f89e42267213cb94da7ddc3016...	3522	46	d13e50eaa47b4cbe9eb81465865d...	6987.14999999999	67						
seller_id	total_amount_...	total_items_sold																		
240b9776d844d37535668549a39...	13332.089999999997	35																		
81f89e42267213cb94da7ddc3016...	3522	46																		
d13e50eaa47b4cbe9eb81465865d...	6987.14999999999	67																		
Requête 4	Quels sont les 5 codes postaux, enregistrant plus de 30 commandes, avec le pire review score moyen sur les 12 derniers mois ?	<table><thead><tr><th>customer_zip_...</th><th>avg_review_sc...</th><th>nb_reviews</th></tr></thead><tbody><tr><td>22753</td><td>2.8085</td><td>47</td></tr><tr><td>22770</td><td>3.1351</td><td>37</td></tr><tr><td>22793</td><td>3.2333</td><td>90</td></tr><tr><td>21321</td><td>3.2778</td><td>36</td></tr><tr><td>22780</td><td>3.3514</td><td>37</td></tr></tbody></table>	customer_zip_...	avg_review_sc...	nb_reviews	22753	2.8085	47	22770	3.1351	37	22793	3.2333	90	21321	3.2778	36	22780	3.3514	37
customer_zip_...	avg_review_sc...	nb_reviews																		
22753	2.8085	47																		
22770	3.1351	37																		
22793	3.2333	90																		
21321	3.2778	36																		
22780	3.3514	37																		

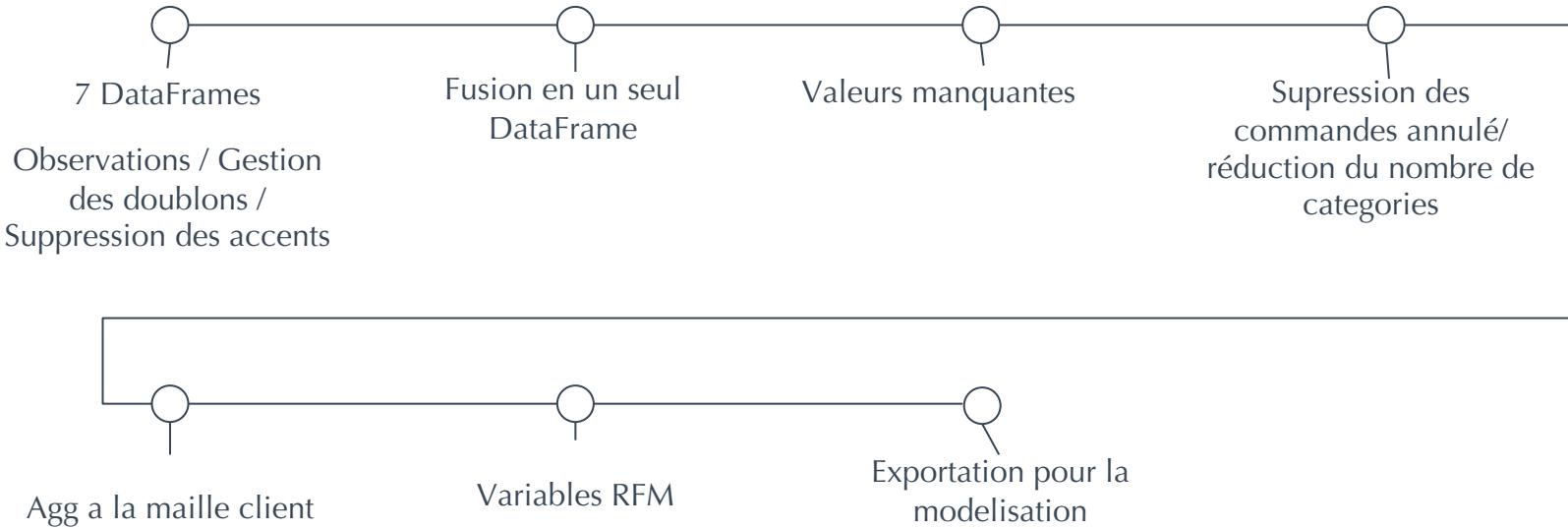
Présentation des données



Nous avons 8 jeux de données :

- Order_Payments : montant du paiement et moyen de paiement
- Orders : Détails des dates d'achat et de livraison, statut de la livraison
- Order_customer : ID du client, localisation du client
- Order_reviews : La date, le commentaire et la note créés par un client pour une commande
- Order_items : ID de la commande, du produit et du vendeur, ainsi que le prix et les frais de livraison pour un item
- Products : Liste des produits, leur catégorie, leurs dimensions et leur poids
- Geolocation : Données géographiques pour un code postal : latitude, longitude, ville, état.

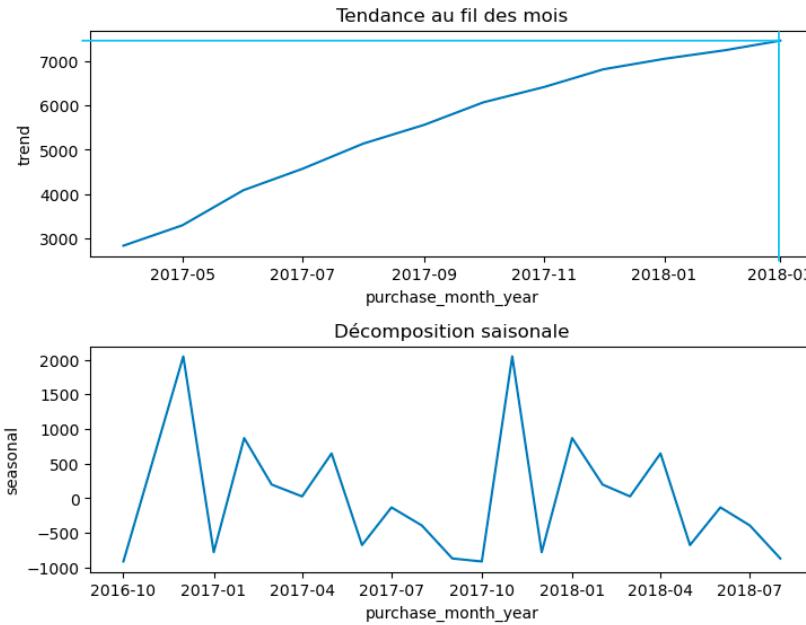
Préparation des données





Exploration

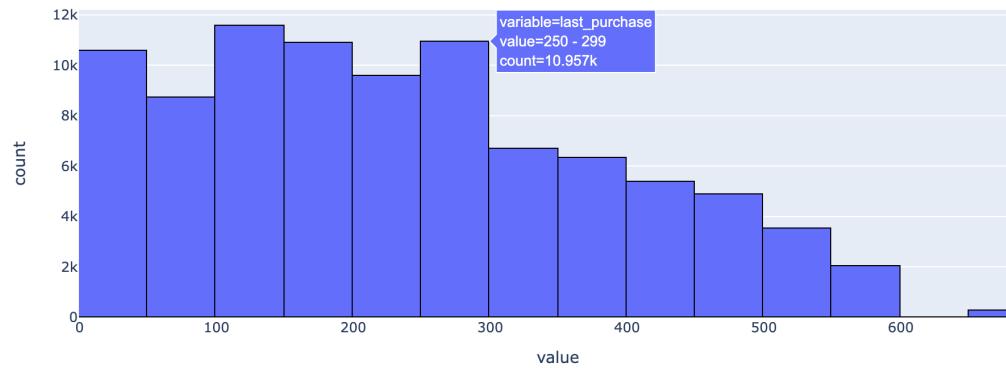
Tendance mensuelle et saisonnière



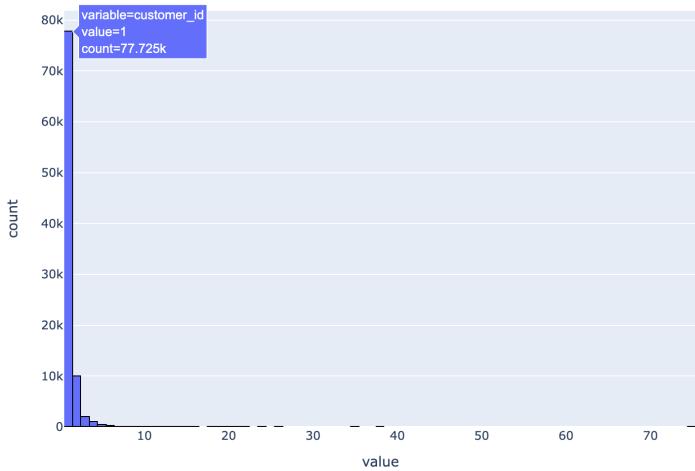
La tendance est à l'augmentation constante. On peut observer deux pics de commandes en décembre 2016 et en novembre 2017. Les périodes creuses se situent fin août pour 2017 et 2018. On peut également observer des tendances similaires d'une année à l'autre.

Observation de la récence et de la fréquence

Histogramme de la récence (jours)



Histogramme de la fréquence d'achat

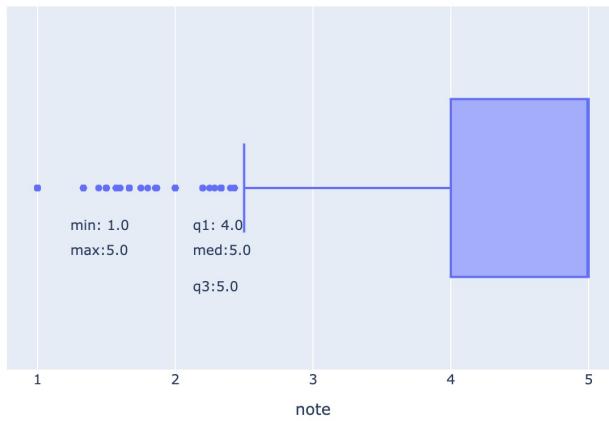


Sur le premier histogramme, nous observons la récence d'achat des clients. On observe que la majorité des clients ont effectué leur dernier achat il y a moins d'un an, indiquant une activité récente et continue.

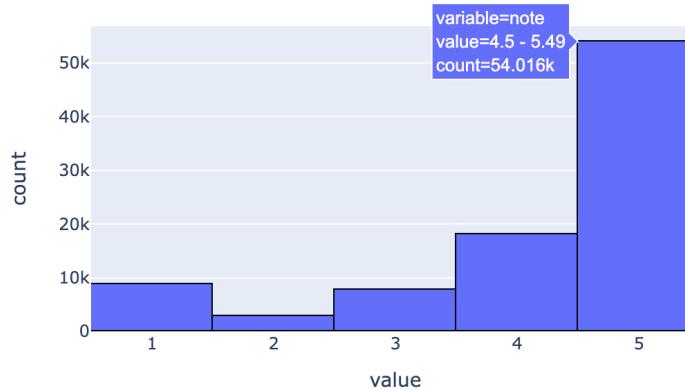
Le deuxième histogramme illustre la fréquence d'achat, nous constatons que sur un total de 91,459 clients, 77,725 clients ont effectué un seul achat. Cela représente environ 85%.

Observation de la note moyenne

Boxplot de la note moyenne

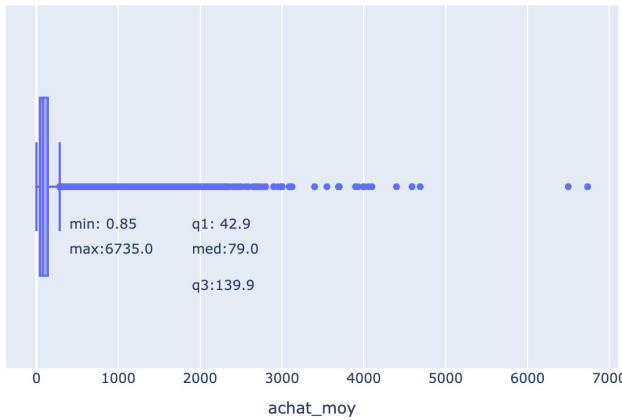


Histogramme de la note moyenne



Observation des dépenses

Boxplot du panier moyen



Bien que la plupart des clients aient des dépenses moyennes relativement modestes, on remarque également la présence de transactions à des valeurs très élevées, comme en témoigne le maximum de 6735 Real.

Ces observations indiquent une diversité dans les habitudes de dépenses des clients, avec une majorité préférant des achats à faible coût tandis qu'une minorité investissent dans des achats plus importants.



Modélisation

Données sélectionnées pour les essais

Recency	Nombres de jours depuis la dernière visite
Frequency	Le nombre de fois où un client a effectué des achats.
Monetary	Dépense moyenne par client.
Note	Note moyenne laissée par un client.
Payment_type	Le moyen de paiement le plus utilisé par un client.
Customer_state	
Product_category	La catégorie où le client a fait le plus d'achats.

Algorithmes testés sur les variables RFM : Les variables numériques ont été standardisées avec StandardScaler(), tandis que les variables catégorielles ont été encodées avec OneHotEncoder().

- K-means
- DBScan
- Clustering hiérarchique

Evaluation des modèles

Algorithme	Choix du nombre de cluster	Evaluation	Visualisation
K-Means	Méthode du coude, score silhouette	Score silhouette, Indice de Davies-Bouldin	Scatter plot avec réduction de dimension avec T-SNE
DB scan	Déterminer epsilon grâce à la méthode de densité atteignable	Score silhouette, Indice de Davies-Bouldin	Scatter plot avec réduction de dimension avec T-SNE
Clustering hiérarchique	Méthode du coude, score silhouette	Score silhouette, Indice de Davies-Bouldin	Dendrogramme

Le score de silhouette évalue la qualité d'un clustering en mesurant la similitude intra-cluster et la dissimilitude inter-cluster.

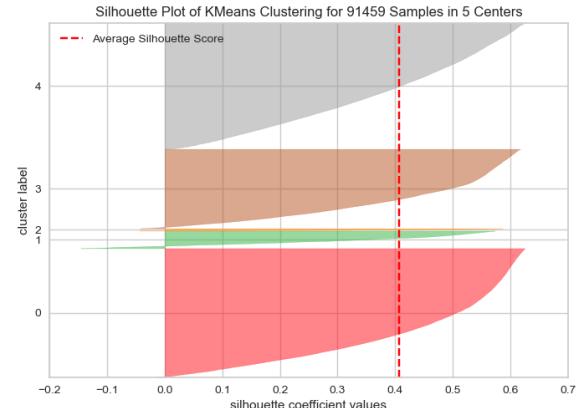
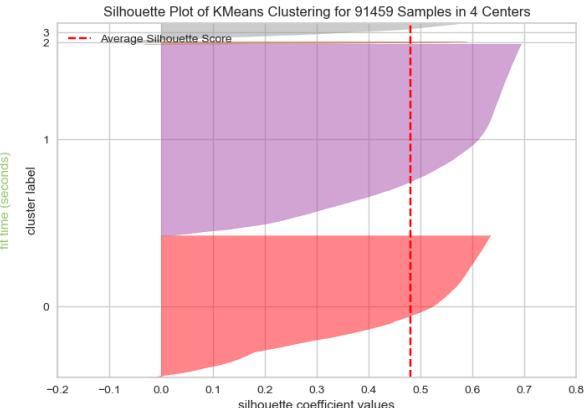
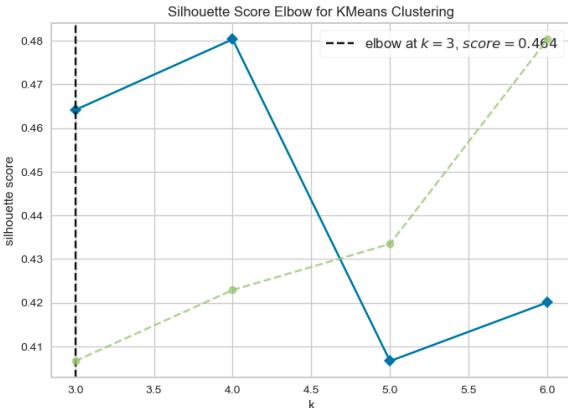
L'indice de Davies-Bouldin évalue la qualité d'un clustering en mesurant la similitude entre les clusters voisins et la séparation entre eux.

T-SNE est utilisé pour la réduction de dimension et la visualisation des données dans des espaces de faible dimension.

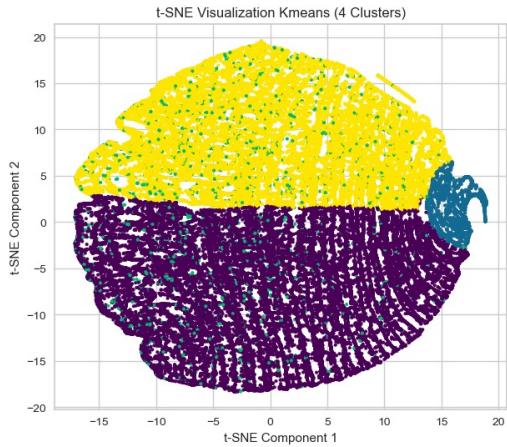
K-means

Le k-means est un algorithme de clustering qui sépare un ensemble de données en différents groupes en fonction de leur similarité. L'objectif est de minimiser la distance entre les points dans un même cluster tout en maximisant la distance entre les différents clusters.

Définir le nombre de cluster avec les features RFM



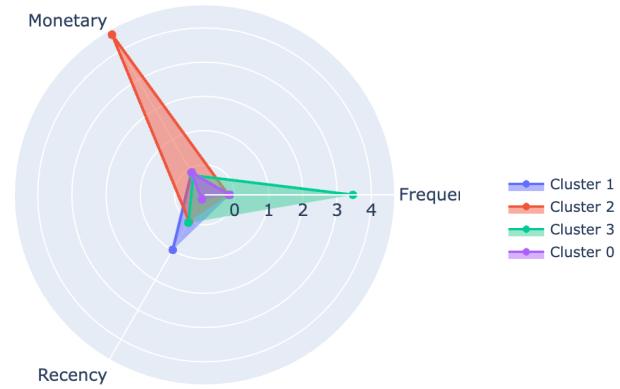
K-means sur les données RFM



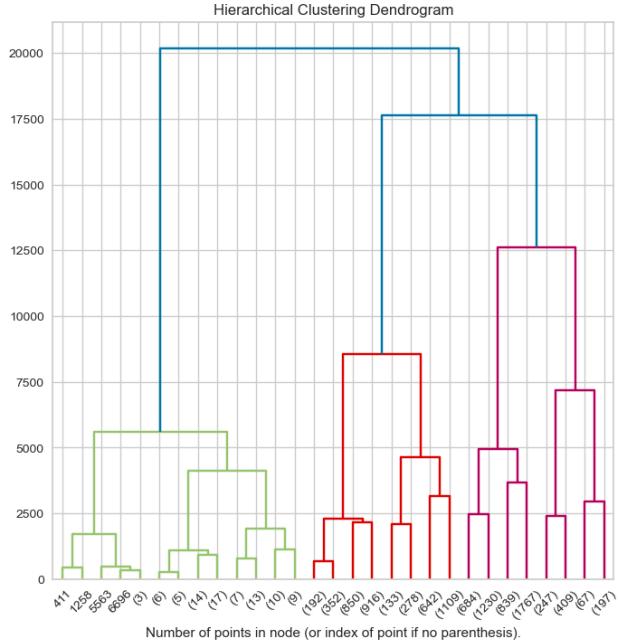
cluster	Frequency	Recency	Monetary	
	mean	mean	mean	count
0	1.11	126.93	102.75	49221
1	1.12	386.87	102.70	36097
2	1.08	240.28	979.95	2471
3	4.16	247.49	83.89	3670

Coefficient silhouette : 0.2682138013312619

Coefficient de Davies-Bouldin : 1.6081947092492823



Clustering hiérarchique



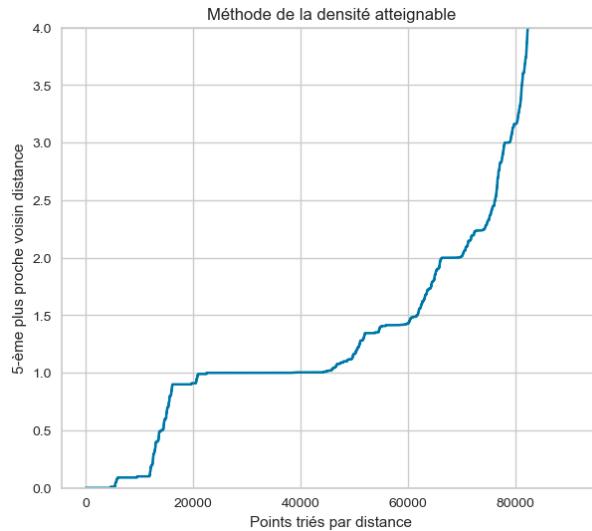
Bien que le clustering hiérarchique agglomératif puisse fournir des informations précieuses sur la structure des données, les contraintes de performance en termes de temps d'exécution le rendent moins adapté à notre cas d'utilisation. Un échantillon de taille 40000 nécessite 81 minutes de traitement.

Silhouette score: 0.45548131521334745
Davies-Bouldin index: 0.7793515560285316

DB_Scan

Un algorithme de clustering basé sur la densité, permettant d'identifier des clusters dans des ensembles de données présentant des formes complexes, tout en ayant la capacité de détecter des points isolés considérés comme du bruit.

Déterminer la valeur de epsilon



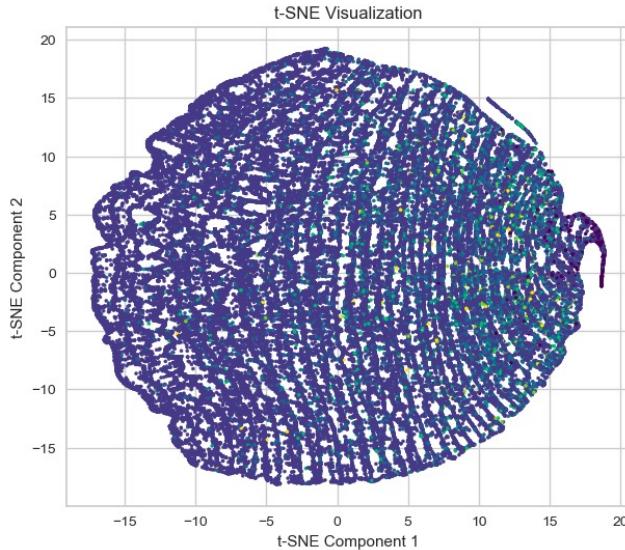
Paramètres importants:

Epsilon: distance max entre deux points pour qu'ils soient considérés comme voisins.

Min_samples: Nombre minimum de points dans le voisinage d'un point pour qu'il soit considéré comme un point noyau.

```
Meilleurs paramètres : eps = 1 , min_samples = 50  
Meilleur score de silhouette : -0.324813393570755
```

DB_Scan sur les données RFM



cluster	Frequency		Recency	Monetary	count
	mean	mean	mean	mean	
-1	6.53	277.43	937.34	455	
0	1.00	235.64	127.07	77619	
1	2.00	245.47	93.26	9950	
2	4.00	266.02	79.02	916	
3	3.00	243.18	83.26	1920	
4	5.00	244.76	79.27	323	
5	6.00	239.30	65.42	276	

Coefficient de silhouette : -0.1330300147149712
Coefficient de Davies-Bouldin : 27.050070927249056

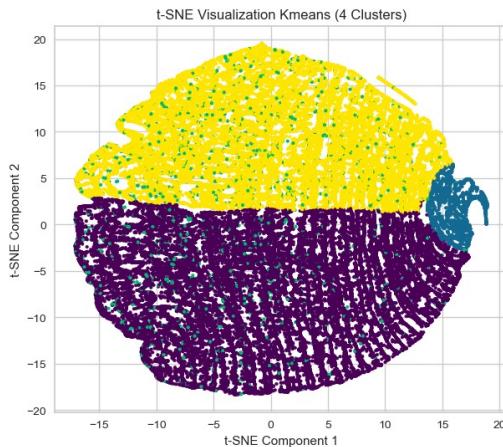
Après avoir évalué les performances de l'algorithme DBSCAN, nous avons décidé de ne pas l'utiliser pour notre cas. Les principales raisons sont sa lenteur, qui rend son utilisation inefficace sur notre ensemble de données, et le fait que les clusters formés par DBSCAN sont beaucoup moins précis que ceux obtenus avec l'algorithme KMeans. De plus, le score de silhouette est très bas (< 0) et le coefficient Davies-Bouldin est très élevé.

Exploration de scenarios avec K-means

Après l'exploration de différents algorithmes, nous avons opté pour K-Means. Désormais, nous allons explorer plusieurs scénarios d'application de cet algorithme, en analysant différentes configurations et en évaluant leur performance.

Variables	Nbr k	Score silhouette	Coefficient Davies-Bouldin	
RFM	4	0.2682	1.6081	
RFM	5	0.2633	4.7127	
RFM + Note	4	0.1821	1.7669	
RFM + Note	5	0.1600	3.6525	
RFM + Note + Payment-type	4	0.1818	1.7695	

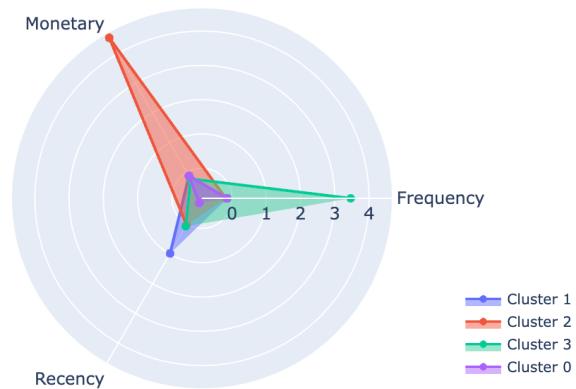
K-means sur les données RFM



	Frequency	Recency	Monetary	
cluster	mean	mean	mean	count
0	1.11	126.93	102.75	49221
1	1.12	386.87	102.70	36097
2	1.08	240.28	979.95	2471
3	4.16	247.49	83.89	3670

Coefficient silhouette : 0.2682138013312619

Coefficient de Davies-Bouldin : 1.6081947092492823



Cluster 0 : Clients assez peu fréquents, dépensant modestement, et récents. (Clientèle à fidéliser)

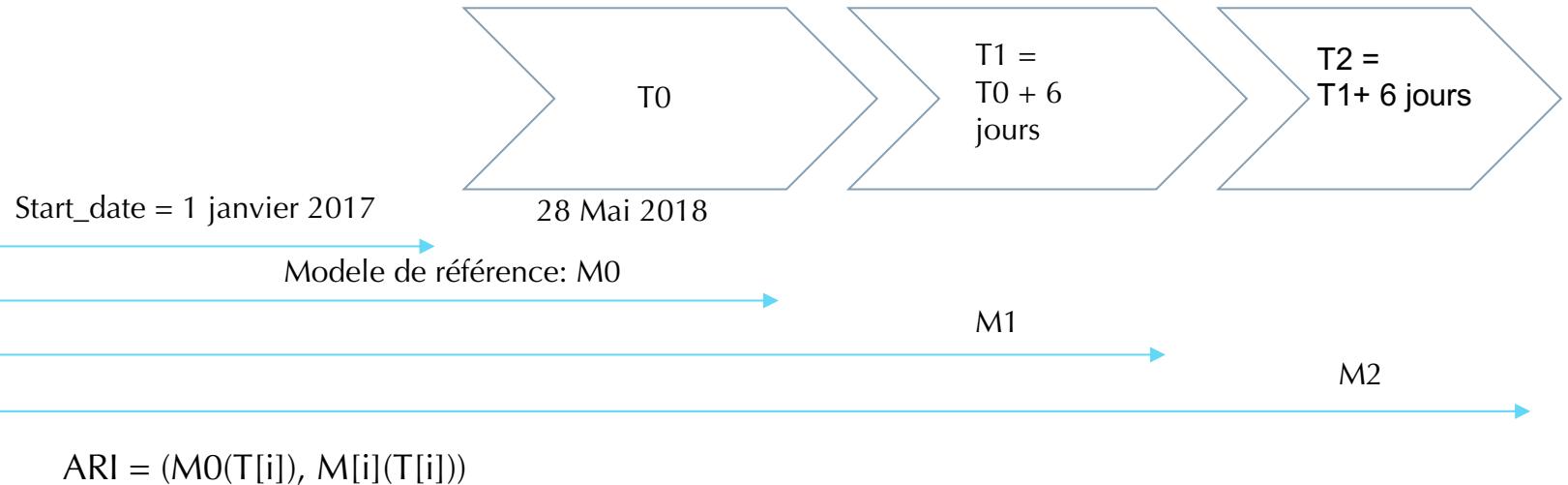
Cluster 1 : Clients assez peu fréquents, et en moyenne n'étant pas venus depuis plus d'un an. (Clientèle à récupérer)

Cluster 2 : Clients assez peu actifs, mais dépensant beaucoup. (Clientèle à récompenser/fidéliser)

Cluster 3 : Clients très actifs, mais dépensant modestement. (Clients loyaux)

Maintenance du modèle

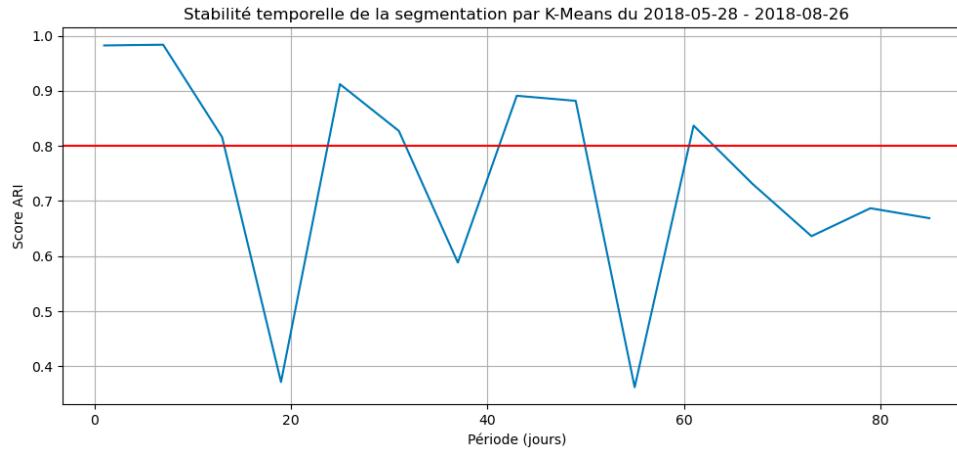
Stabilité du modèle



Le score Adjusted Rand Index (ARI) est une mesure de similarité utilisée pour évaluer à quel point les regroupements de données entre différentes périodes sont cohérents.

Stabilité du modèle

Un ARI proche de 1 indique une forte similitude entre les regroupements, ce qui suggère une stabilité des structures sous-jacentes des données au fil du temps. En revanche, un score plus bas peut indiquer des différences significatives entre les regroupements.



La diminution du score ARI en dessous de 0,8 après 12 jours souligne l'importance d'une réévaluation fréquente des méthodes de modélisation.

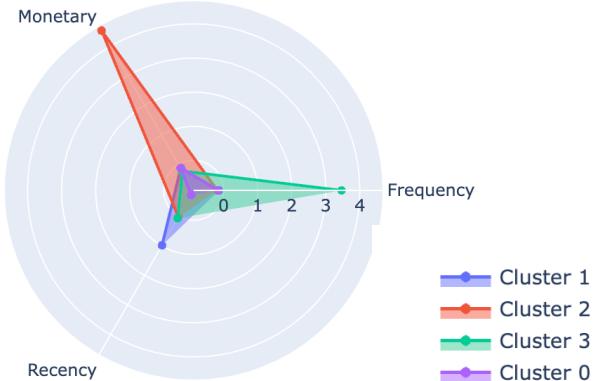
Conclusion

Modèle recommandé : K-means avec les données

RFM et k=4

Maintenance du modèle : Le modèle devra être actualisé tous les 12 jours.

Résultats de la segmentation



Cluster 0 : Clientèle à fidéliser

Cluster 1 : Clientèle perdu à récupérer

Cluster 2 : Clientèle à récompenser/fidéliser

Cluster 3 : Clients loyaux

Merci!

Avez vous des questions?

CREDITS: This presentation template was created by [Slidesgo](#), and includes icons by [Flaticon](#), and infographics & images by [Frepik](#)

