OPENCLASSROOMS

Projet 5

Segmentez des clients d'un site e-commerce

Élaborez un modèle de clustering

David Scanu | Février 2025

Parcours **AI Engineer**

⭐ Étapes du projet

- Problématique et Contexte
- Description du jeu de données
- Création des données RFM+S
- Analyse Exploratoire
- Apprentissage non supervisé
- Analyse approfondie des segments
- Plan de maintenance du modèle
- Conclusion

💶 Contexte et problématique métier

📝 Contexte

- **Entreprise :** Olist, entreprise brésilienne de vente sur les marketplaces.
- **Rôle** : accompagner Olist dans leur projet de monter une **équipe Data**.

⚡ Mission

- **Résoudre une urgence**
  - Fournir les requêtes SQL pour le Dashboard Customer Experience.
- **Segmentation client**
  - Explorer et **analyser les données** fournies.
  - **Segmentation des clients** avec des algorithmes d'apprentissage non supervisé (clustering).
  - **Analyse approfondie des segments obtenus** pour une exploitation optimale par l'équipe Marketing.
  - Proposer un **plan de maintenance** basé sur une simulation de fréquence de mise à jour du modèle.

🔎 Description du jeu de données

Origine des Données

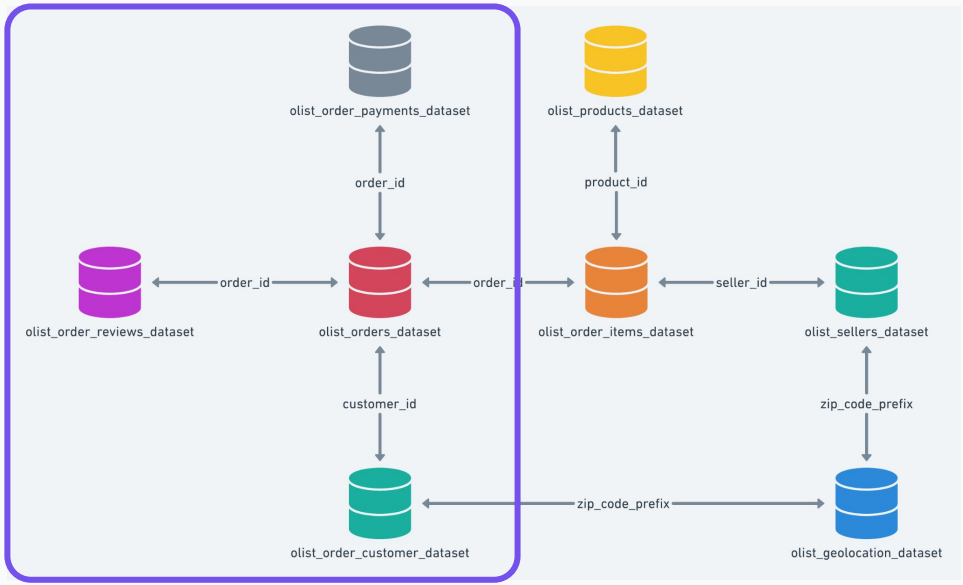
- **Données fournies par Olist**, la plus grande plateforme de vente en ligne au Brésil.

Contenu des Données

- Informations sur **95 831 commandes** réalisées entre 2016 et 2018.
- Inclut des détails sur le statut des commandes, les prix, les paiements, les performances de livraison, et les avis clients.

Tables utilisées

- Commandes (orders)
- Clients (customers)
- Paiements (order_pymts)
- Reviews (order_reviews)



🧪 Création des données RFM+S

Données des commandes

- **Centralisation des informations** concernant les commandes.
- **Jointure des tables** orders, customers, order_pymts et order_reviews.
- **Colonne retenues** : identifiant, date d'achat, ID client, montant, score de satisfaction.
- **Dates des commandes** : entre 03/10/2016 et 29/08/2018.

Données client

- Chaque ligne du tableau représente un **client unique**
- **Agrégation des informations des commandes** pour chaque client.
    - **Identifiant unique du client**
    - **Récence** : Nombre de jours écoulés depuis la dernière commande.
    - **Fréquence** : Nombre total de commandes effectuées par le client.
    - **Montant dépensé** : Somme totale des dépenses réalisées par le client.
    - **Score de satisfaction** : Moyenne des notes attribuées par le client aux commandes passées.

Commandes (95831)

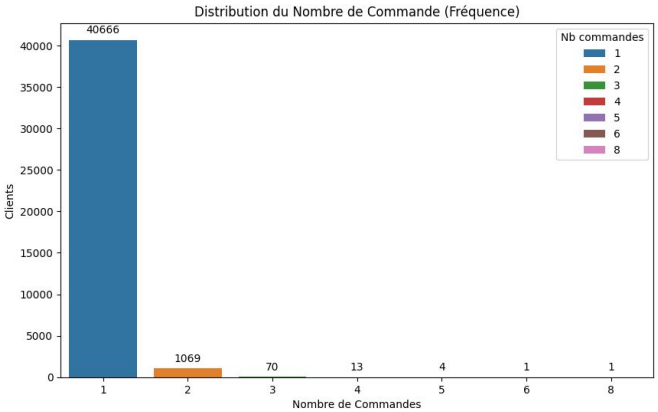
	order_id	order_purchase_...	customer_uniqu...	payment_value	review_score
0	e481f51cbdc5467...	2017-10-02 10:56...	7c396fd4830fd04...	38.71	4.0
1	53cdb2fc8bc7dce...	2018-07-24 20:41...	af07308b275d75...	141.46	4.0
2	47770eb9100c2d...	2018-08-08 08:38...	3a653a41f6f9fc3...	179.12	5.0
3	949d5b44dbf5de...	2017-11-18 19:28...	7c142cf63193a14...	72.20	5.0
4	ad21c59c0840e6...	2018-02-13 21:18...	72632f0f9dd73df...	28.62	5.0

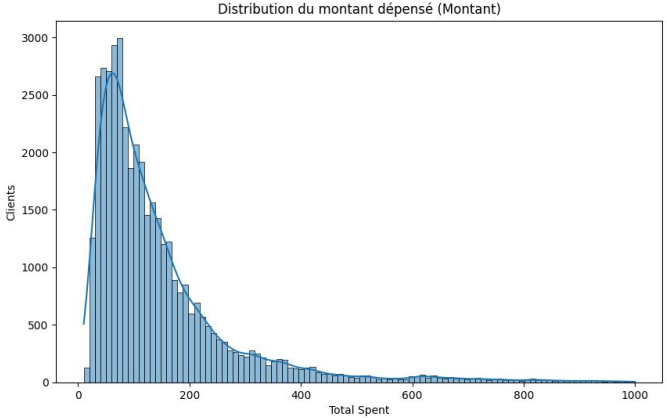
Clients (2017 : 41824)

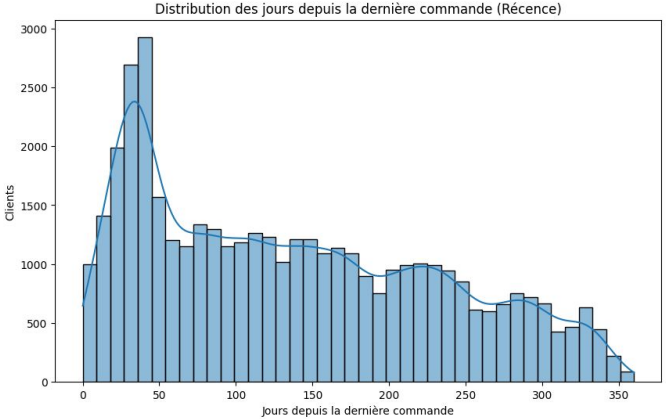
customer_uniqu...	recency	frequency	monetary_value	avg_review_score
0000f46a3911fa...	296	1	86.22	3.0
0000f6ccb0745...	80	1	43.62	4.0
0004aac84e0df...	47	1	196.89	5.0
0005e1862207b...	302	1	150.12	4.0
0006fdc98a402f...	166	1	29.00	3.0
00082cbe03e47...	42	1	126.26	5.0
000a5ad9c4601...	142	1	91.28	4.0
000bfa1d2f1a4...	93	1	46.85	4.5
000c8bdb58a29...	19	1	29.00	5.0
000de6019bb59...	136	1	257.44	2.0

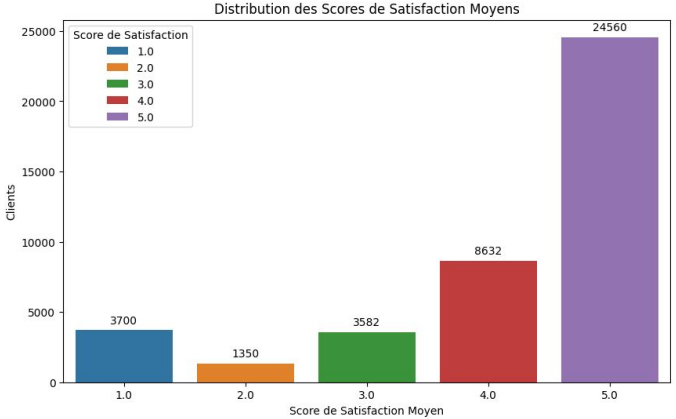
Analyse Exploratoire

OPENCLASSROOMS









🤖 Apprentissage non supervisé

Stratégie

- Segmentation client **de référence**
  - Données de l'**année 2017**
  - **41 824 clients / 43 100 commandes**

Traitement

- Standardisation / Normalisation

Algorithmes de clustering testés

- K-means
- DBSCAN

Choix du nombre de clusters (k)

- Méthode du coude
- Coefficient de silhouette
- Faisabilité métier

🔵 Pas assez de clusters (k trop petit)

- Groupes sont trop larges et hétérogènes
- Perte d'informations
- Peu exploitable pour le marketing

🔴 Trop de clusters (k trop grand)

- Risque de sur-ajustement (*overfitting*)
- difficile à interpréter et à utiliser en pratique.

🟢 Nombre optimal de clusters (bon équilibre)

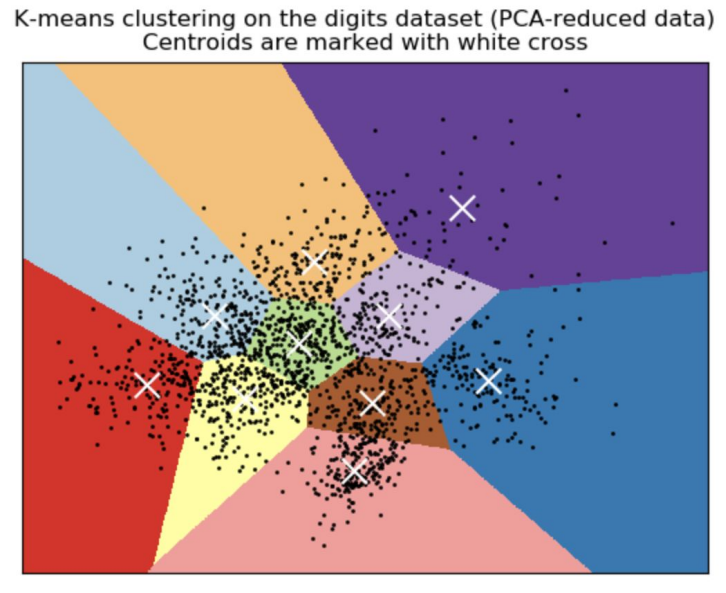
- Groupes homogènes et exploitables.
- Segmentation est claire et utile pour l'entreprise.
- Différences entre segments sont significatives.

Algorithmes de clustering : K-means

Algorithme de regroupement (clustering) qui divise un ensemble de données en k clusters en minimisant l'inertie (somme des distances intra-cluster).

Principe

- Choisir k centroids initiaux
- Assigner chaque point au centroid le plus proche
- Recalculer les centroids comme la moyenne des points assignés
- Répéter jusqu'à convergence (centroids ne bougent plus)



🦾 Méthode du coude (Elbow Method)

Déterminer le nombre optimal de clusters (k)

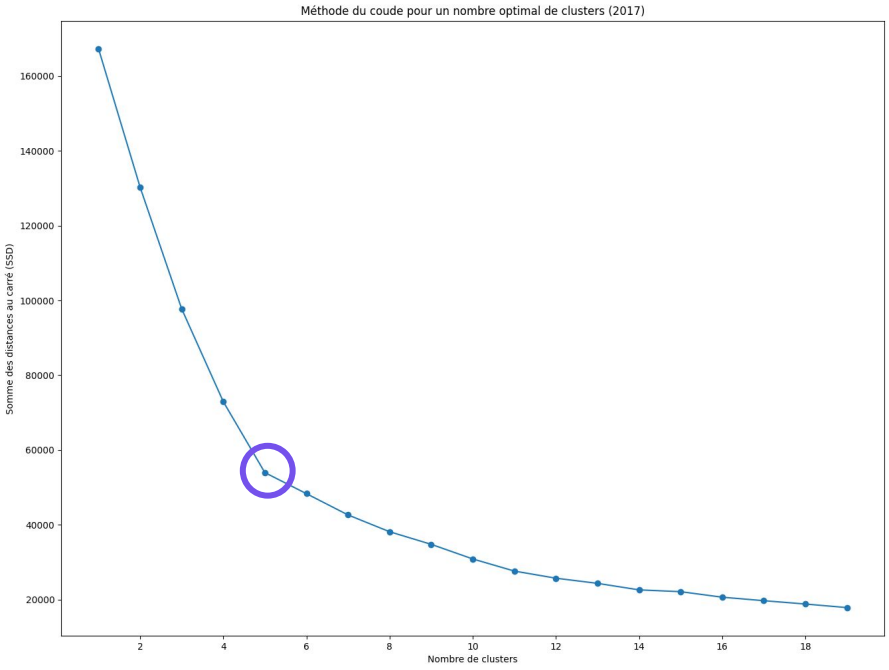
- Analyse de l'inertie intra-cluster
- Somme des distances des points à leur centre de cluster (WCSS, SSD)
- En fonction du nombre de clusters (k)

Recherche

- Point de coude sur la courbe
- Equilibre précision et simplicité

Doit être compléter par

- Coefficient de silhouette



👤 Coefficient de silhouette

Mesure la qualité du regroupement des clusters

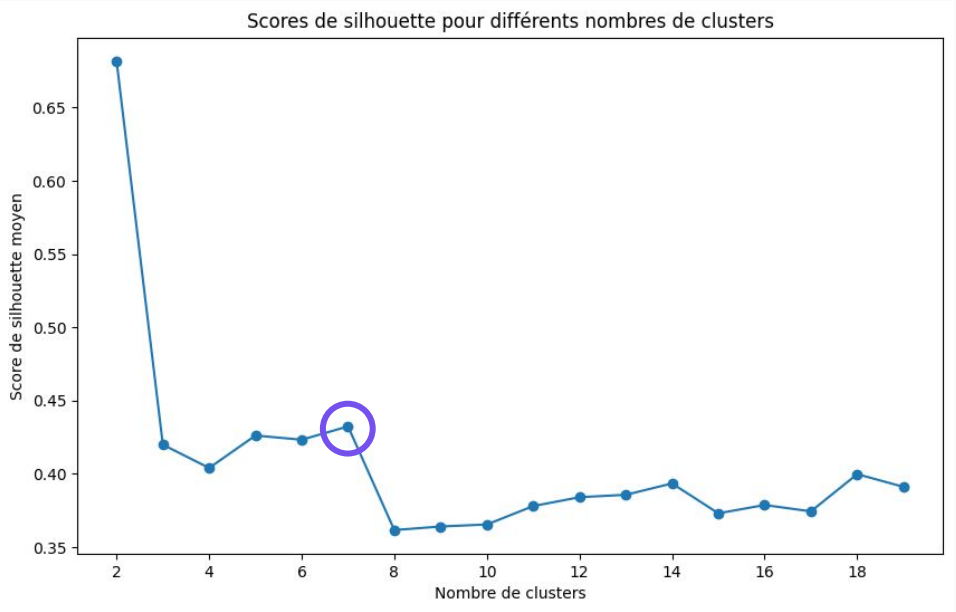
- Cohésion intra-cluster
- Séparation inter-cluster

Interprétation des valeurs

- Compris entre -1 et 1
- Doit être **élevé** pour indiquer un bon clustering

Nombre de clusters choisis

- **n_clusters = 7**
- Score de silhouette moyen est **0.4324**



📊 Analyse approfondie des segments

Visualisation des clusters

customer_unique_id	recency	frequency	monetary	avg_review_score	cluster
0000f46a3911fa3c0805444483337064	296	1	86.22	3.0	2
0000f6ccb0745a6a4b88665a16c9f078	80	1	43.62	4.0	0
0004aac84e0df4da2b147fca70cf8255	47	1	196.89	5.0	0
0005e1862207bf6ccc02e4228effd9a0	302	1	150.12	4.0	1
0006fdc98a402fceb4eb0ee528f6a8d4	166	1	29.00	3.0	2

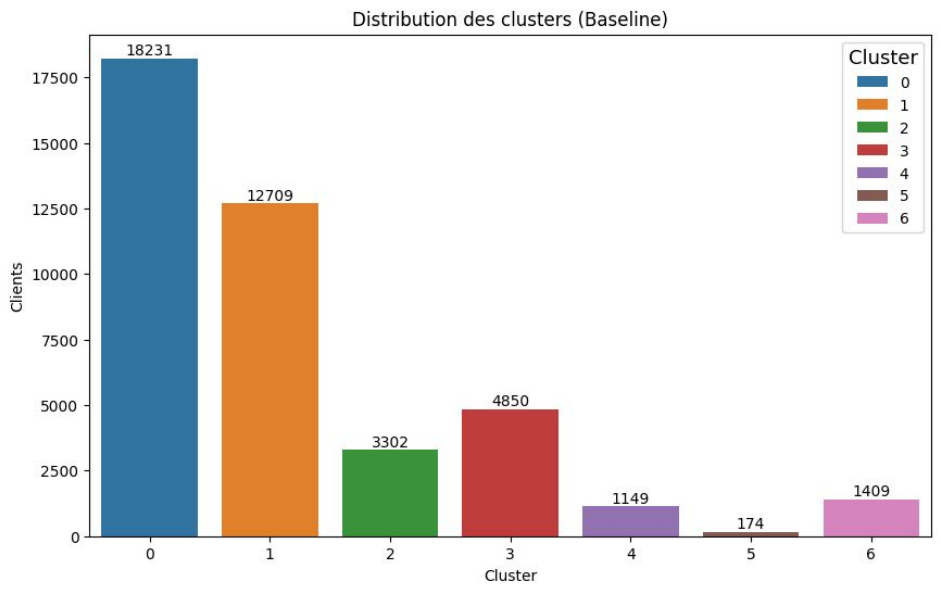
Indices de qualité des clusters

- **Inertie**: 42620.24 (élevé)
- **Score Silhouette** : 0.43
- **Indice de Davies-Bouldin (DBI)** : 0.83
- **Indice de Calinski-Harabasz** : 20387.68

Interprétation

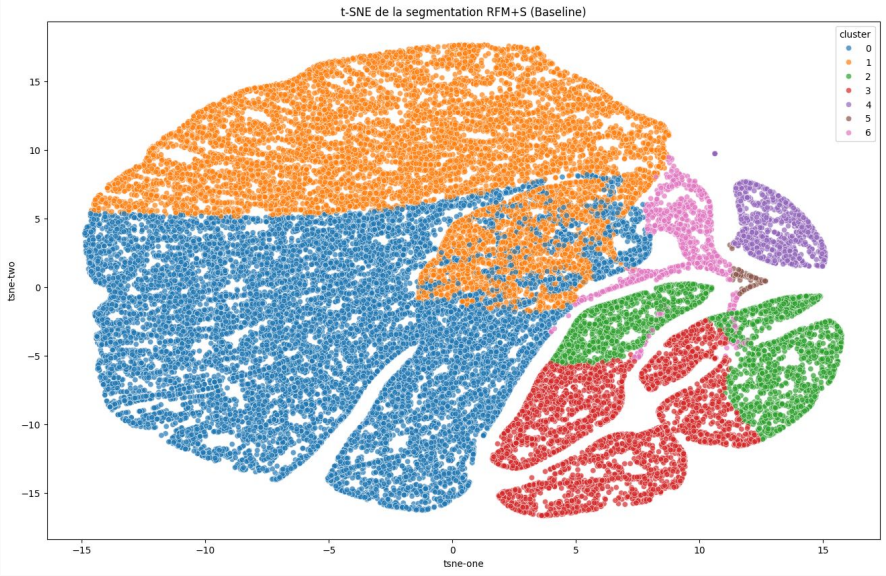
- Clusters sont relativement bien définis (Silhouette = 0.43)
- Bonne séparation entre clusters (DBI = 0.76 et CHI élevé)

Distribution des clusters

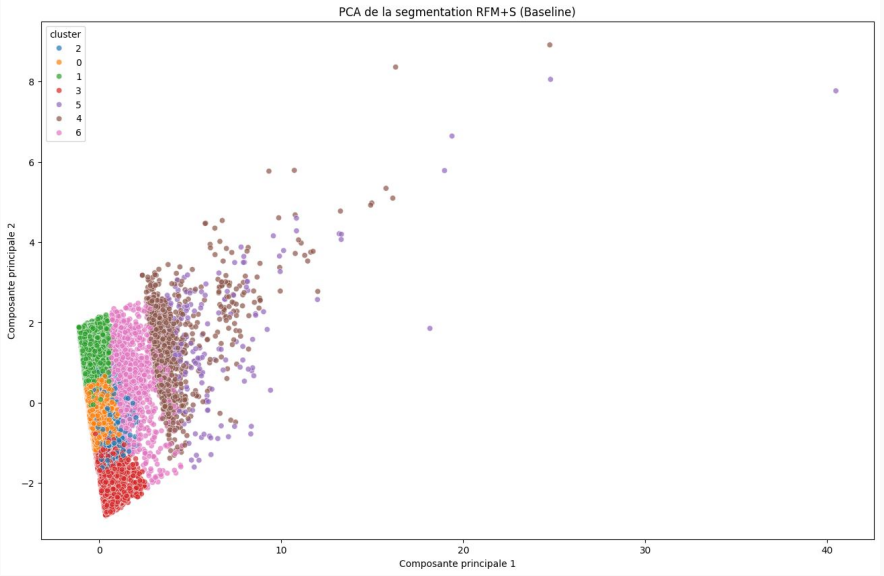


👀 Visualisation par réduction de dimensions

Visualisation t-SNE



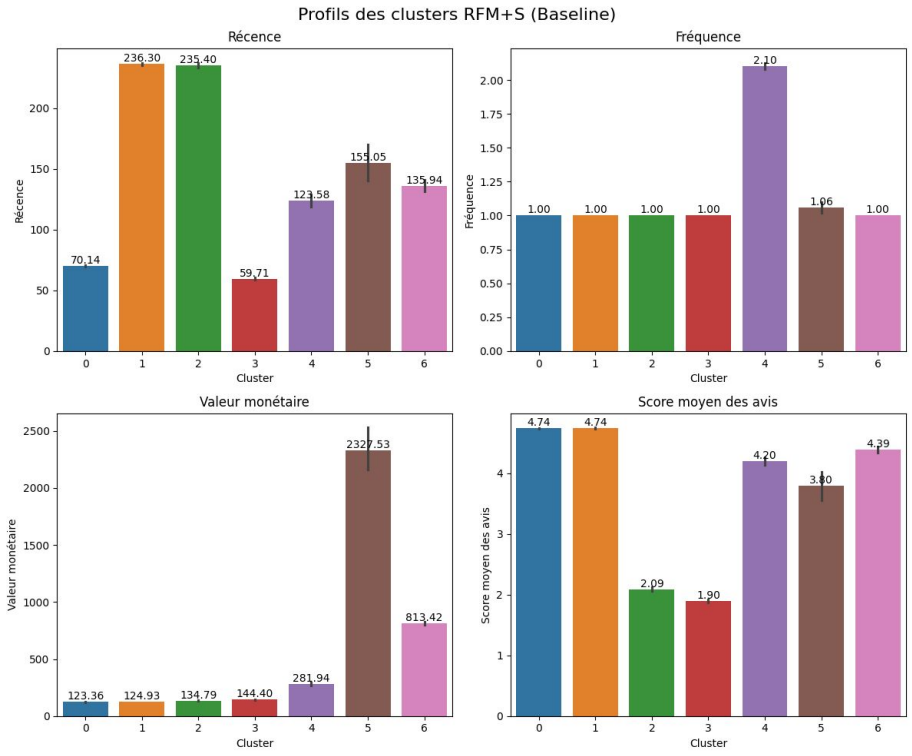
Visualisation PCA



📊 Statistiques descriptives des clusters

Moyennes et écarts-types des variables pour chaque segment

- **Moyenne** : tendance centrale des variables pour chaque cluster, permettant d'identifier les valeurs typiques
- **écarts-type** : variabilité au sein de chaque cluster, indiquant si les comportements des clients sont homogènes ou diversifiés



🎯 Identification des segments clés et stratégies adaptées

Cluster	Profil client	Nombre de clients	Récence	Fréquence	Montant dépensé	Score moyen des avis	Stratégie
0	🌟 Nouveaux Clients Satisfaits	18 231	70,14	1,00	123,36	4,74	Programme de fidélité et email de bienvenue personnalisé
1	😢 Clients Inactifs Satisfaits	12 709	236,30	1,00	124,93	4,74	Promotions spéciales et rappels de produits populaires
2	❌ Clients Inactifs et Insatisfaits	3 302	235,40	1,00	134,79	2,09	Identifier les raisons de l'insatisfaction et proposer des solutions ciblées
3	🚨 Nouveaux Clients Déçus	4 850	59,71	1,00	144,40	1,90	Analyser le parcours d'achat et améliorer l'expérience client
4	🚀 Clients à Potentiel Élevé	1 149	123,58	2,10	281,94	4,20	Programmes de fidélité et recommandations personnalisées
5	🎩 Clients Premium	174	155,05	1,06	2327,53	3,80	Offres VIP et service client premium
6	💰 Clients Dépensiers	1 409	135,94	1,00	813,42	4,39	Offres spéciales et recommandations basées sur l'historique d'achat

Algorithmes de clustering : DBSCAN

DBSCAN

Algorithme de clustering qui regroupe des points proches les uns des autres en fonction de la densité.

- N'exige pas de spécifier le nombre de clusters à l'avance.
- Utile pour les données avec des formes de clusters arbitraires.
- Identifier les points de bruit.

Principaux paramètres de DBSCAN

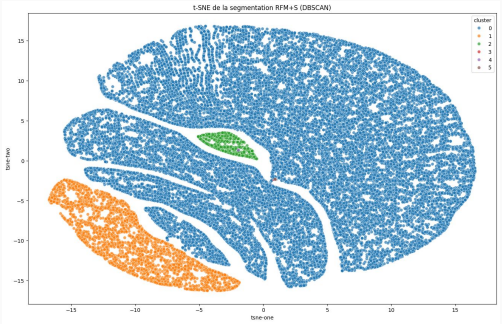
- **eps** : Distance maximale entre deux points pour qu'ils soient considérés comme voisins.
- **min_samples** : Nombre minimum de points dans un voisinage pour qu'un point soit considéré comme un noyau de cluster.

Difficultés

- Choix du nombre clusters.

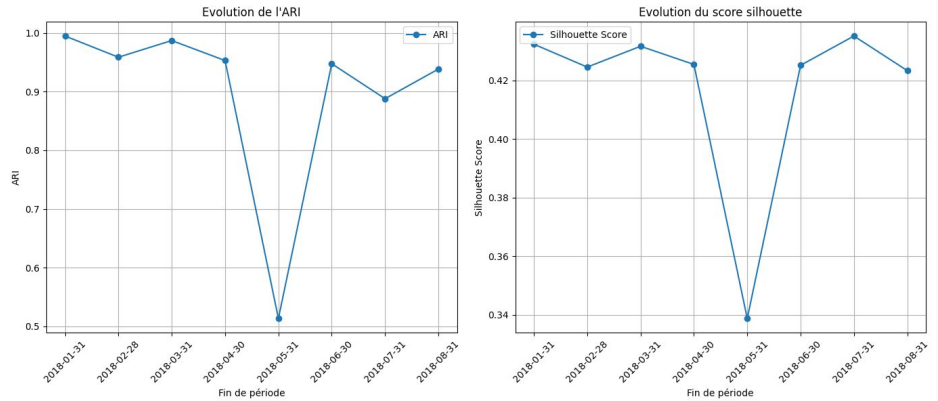
Résultats

- Clusters peu interprétables et moins intéressants.
- Profils client très similaires.
- Certains clusters contiennent très peu de clients.
- Métriques de qualité des clusters : bonnes mais ne surpassent pas celles obtenues avec K-Means.

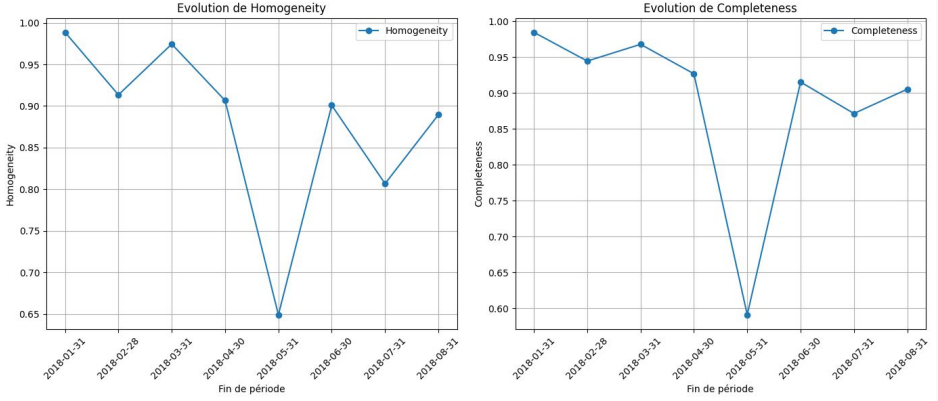


📉 Analyse de la stabilité des segments

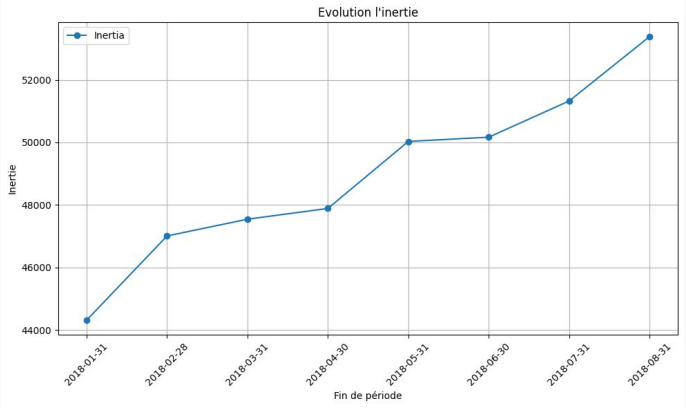
Fin de période	Inertie	Silhouette	ARI	Homogeneity	Completeness	Delta Inertie	Delta Silhouette
2017 (baseline)	42620.2368	0.4323					
2018-01-31	44312.0628	0.4323	0.9948	0.9885	0.9845	0.0397	0.0002
2018-02-28	47007.8855	0.4245	0.9587	0.9134	0.9445	0.1029	0.0181
2018-03-31	47544.2227	0.4316	0.9871	0.9744	0.9678	0.1155	0.0017
2018-04-30	47888.4922	0.4254	0.9530	0.9066	0.9268	0.1236	0.0161
2018-05-31	**50031.7873**	**0.3387**	**0.5135**	**0.6492**	**0.5912**	**0.1739**	**0.2166**
2018-06-30	50169.0351	0.4251	0.9476	0.9008	0.9153	0.1771	0.0169
2018-07-31	51326.7586	0.4351	0.8880	0.8067	0.8715	0.2043	0.0063
2018-08-31	53388.6962	0.4233	0.9385	0.8895	0.9053	0.2527	0.0208


Evolution de l'ARI
ARI
Fin de période

Evolution du score silhouette
Silhouette Score
Fin de période


Evolution de Homogeneity
Homogeneity
Fin de période

Evolution de Completeness
Completeness
Fin de période


Evolution l'inertie
Inertia
Fin de période

🚧 Plan de maintenance du modèle

Déterminer la meilleure fréquence d'entraînement

- **Exécutant le code de surveillance** sur une fenêtre glissante (par exemple, tous les mois)
- **Analyse l'évolution des métriques** de dérive
- **Compléter ces métriques par des informations commerciales** (par ex : campagnes saisonnières, changements de produits)

Fréquence Recommandée

- **Tous les 4 mois**

Justification

Stabilité et Dérive des Clusters

- Clusters relativement stables sur des périodes de 4 mois
- Dérive significative observée après 4 mois (en mai 2018)

Équilibre entre Stabilité et Réactivité

- Maintenir un équilibre entre la stabilité des clusters
- Réactivité aux changements dans les données
- Minimiser les coûts et les efforts associés à des mises à jour trop fréquentes

⚡ Conclusion

Résultats

- Identification des segments clés
- Analyse approfondie des segments
- Recommandations marketing ciblées
- Plan de maintenance du modèle

Technologies utilisées

- SQL
- Python
- scikit-learn, pandas, seaborn

Merci 🙏

Avez-vous des questions ?

🗂️ Annexes

- Dépôt GitHub

Documentation

- Brazilian E-Commerce Public Dataset by Olist
- Yellowbrick — Clustering Evaluation Examples
- KMeans — scikit-learn 1.6.1 documentation
- Drift Tracking: A Complete Guide | TrueFoundry