

PROJET 5

SEGMENTEZ LES CLIENTS D'UN SITE E-COMMERCE

Formation DATA SCIENTIST - Victor BARBIER



SOMMAIRE

01 CONTEXTE

02 PRÉPARATION
DES DONNÉES

03 PISTES DE
MODÉLISATION

04 ANALYSE
TEMPORELLE

05 CONCLUSION

01

CONTEXTE

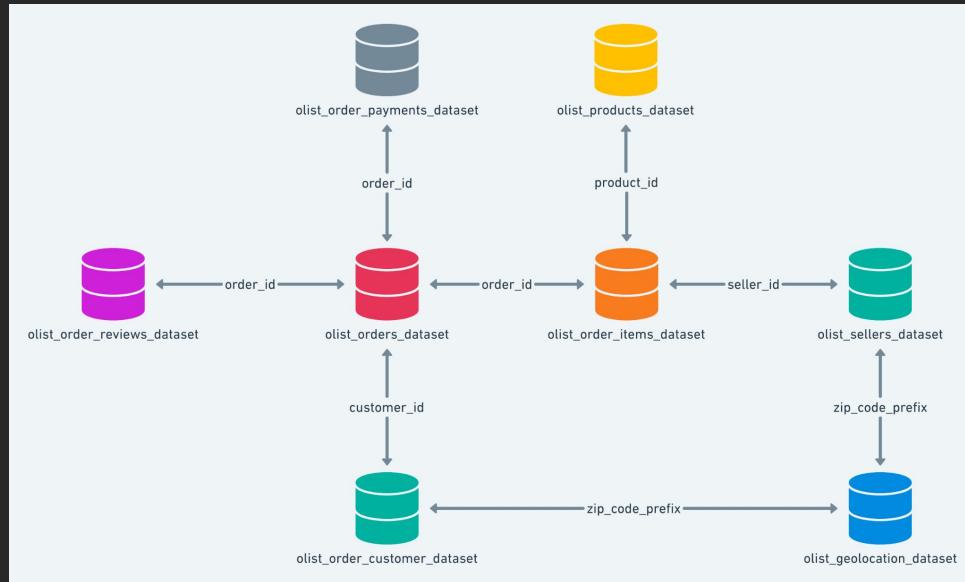


CONTEXTE

ENTREPRISE	MISSION	DONNÉES	MÉTHODOLOGIE
- <i>Olist</i> entreprise brésilienne - Vente sur marketplaces en ligne	- Comprendre les types d'utilisateurs - Proposer et décrire segmentation - Maintenance sur l'analyse des segments	- Historique de commandes depuis 2017 - Commandes, produits, commentaires, localisation ..	- Analyse des données - Algorithmes de clustering non supervisés - Définition des clusters - Analyse temporelle (stabilité)
PROBLÉMATIQUE			
- Segmentation des clients - Campagne marketing			

JEU DE DONNÉES

- Liste des commandes
- 8 datasets différents
- 99441 commandes





02

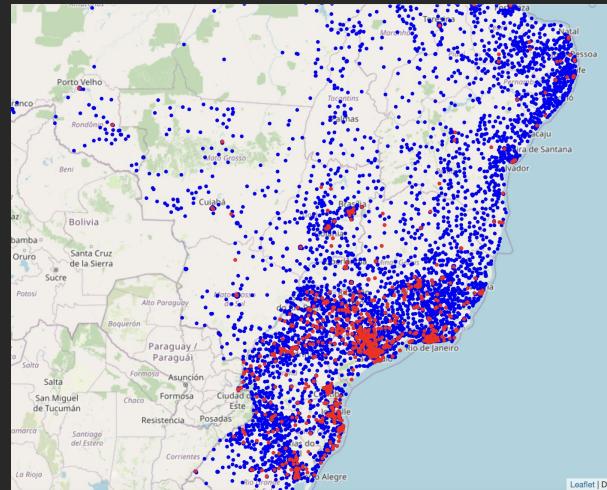
PRÉPARATION DES DONNÉES

ASSEMBLAGE

- Dataframe “global” composés des 8 datasets Olist
- Groupements successifs par “items”, “payments”, “orders” et “customer”
- Dataframe indexé par **unique client** :
 - Nombre de commandes
 - Date de la dernière commande
 - Moyenne de la durée de livraison
 - Paiement moyen / max / total / ecart-type
 - Total paiement par carte / voucher / boleto
 - Nombre de d'échelonnement de paiement moyens
 - Nombre d'items moyens / max par commande
 - Prix par item moyen / max / ecart-type
 - Prix livraison moyen / max
 - Nombre de photos moyen
 - Nombre d'item par categorie
 - Note d'avis moyen
 - Temps d'avis moyen
 - Longitude / latitude client
 - Longitude / latitude vendeurs le plus représenté

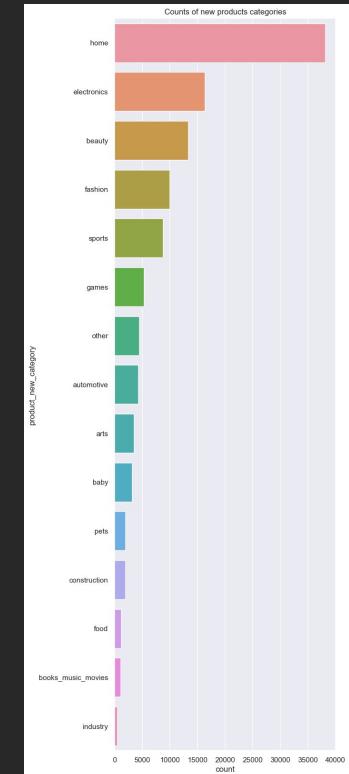
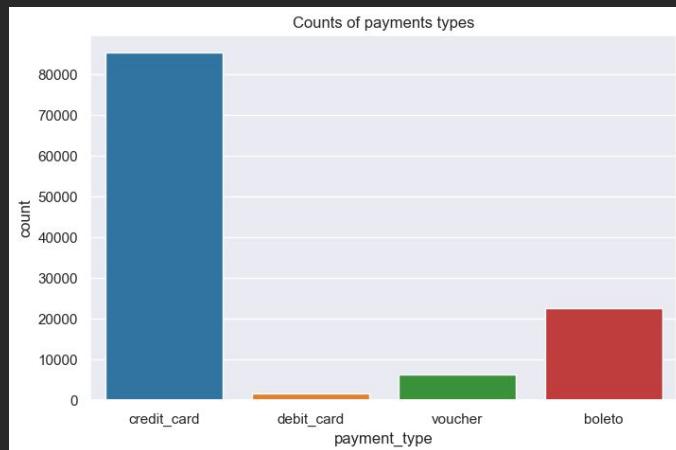
NETTOYAGE

- Filtrage commande *delivered* (97%)
- Suppression des duplcats (5472)
- Suppression mauvaises coordonnées longitude latitude (155)
- Suppression NaN (0.13%)



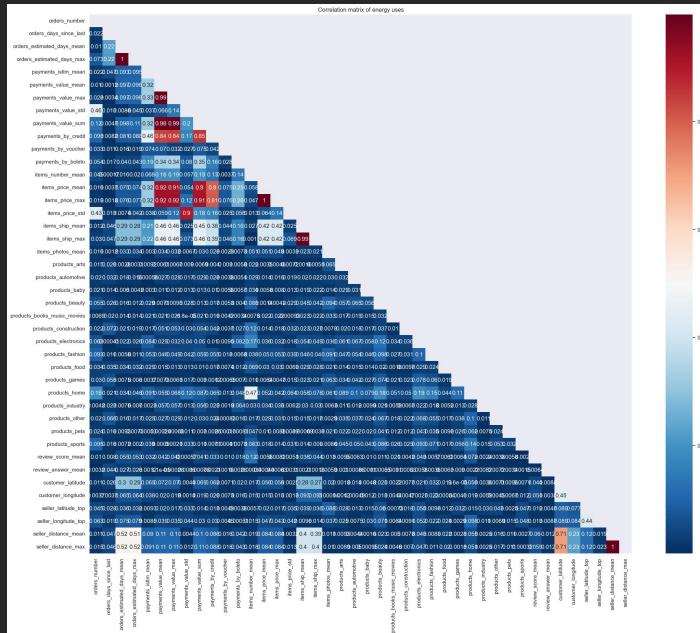
NOUVELLES VARIABLES

- **seller_distance** : Distance harvesine entre client et vendeur
- **product_XXX** : Encodage de nouvelles catégories de produits
- **payment_by_XXX** : Encodage des type de paiement

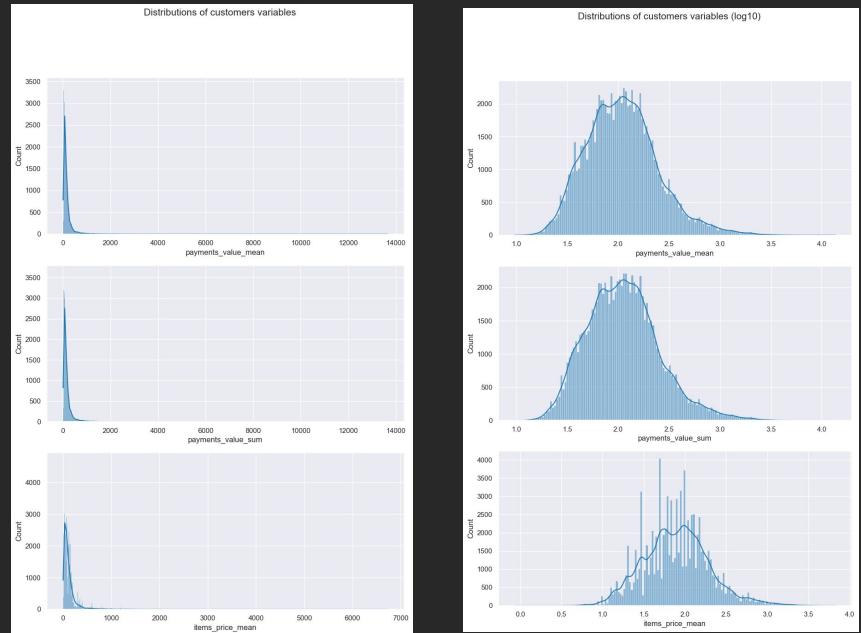


TRANSFORMATION

Suppression de 5 variables fortement corrélées



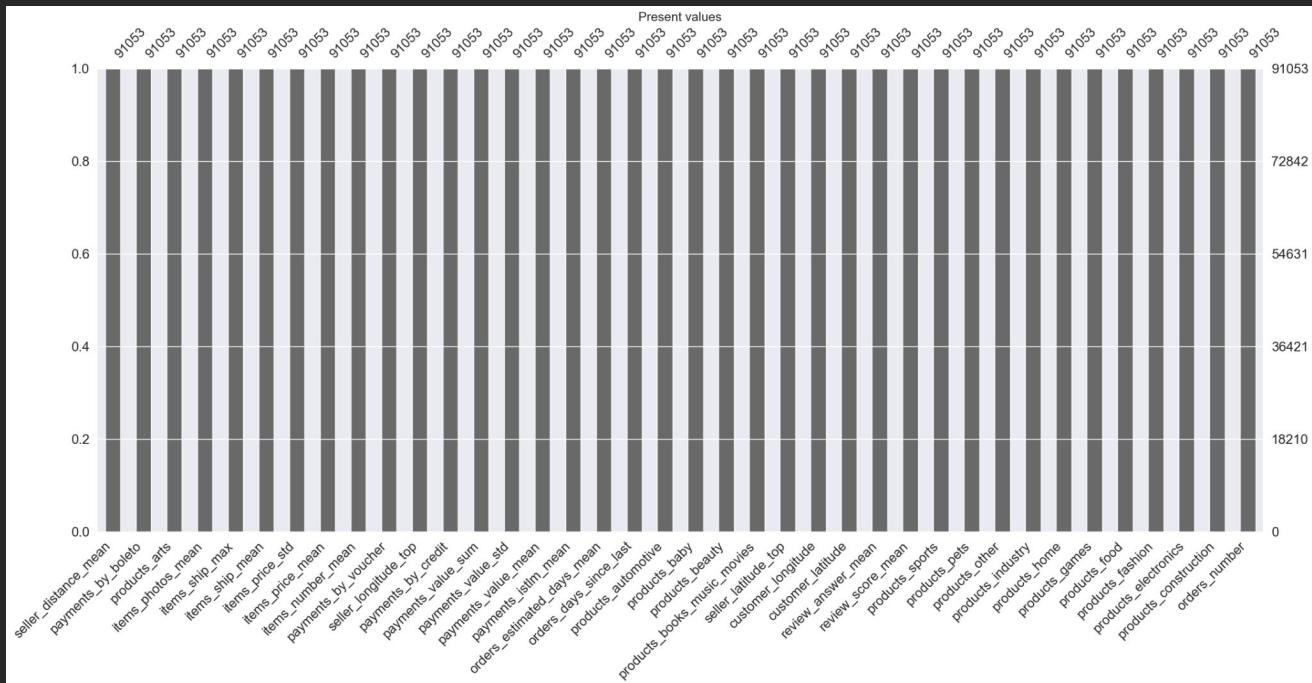
Transformation de 3 variables qui présentent une asymétrie (\log_{10})



JEU FINAL

91053 clients

37 variables



03

PISTES DE MODELISATION



SEGMENTATION RFM

RFM

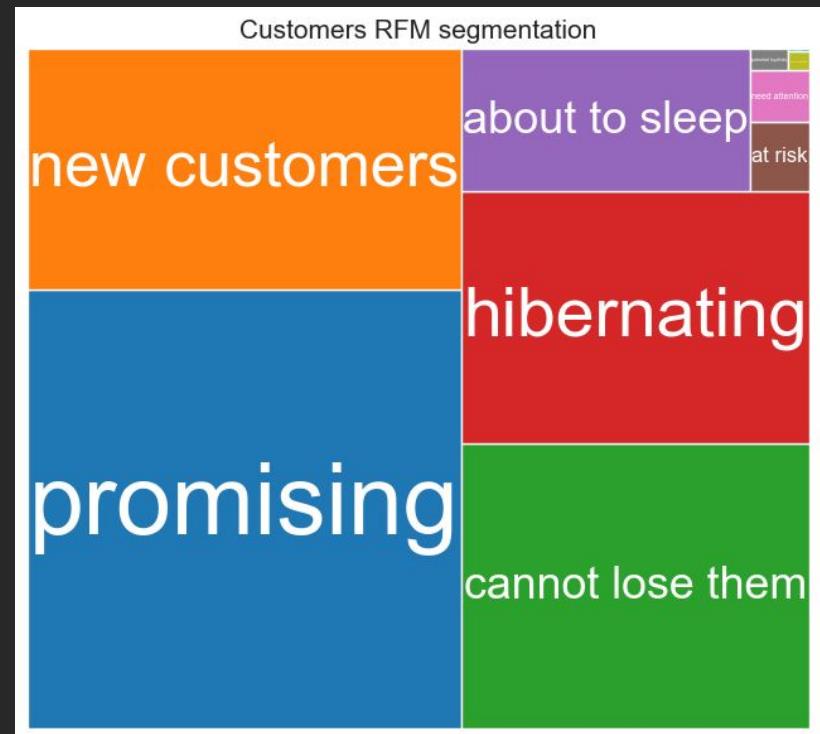
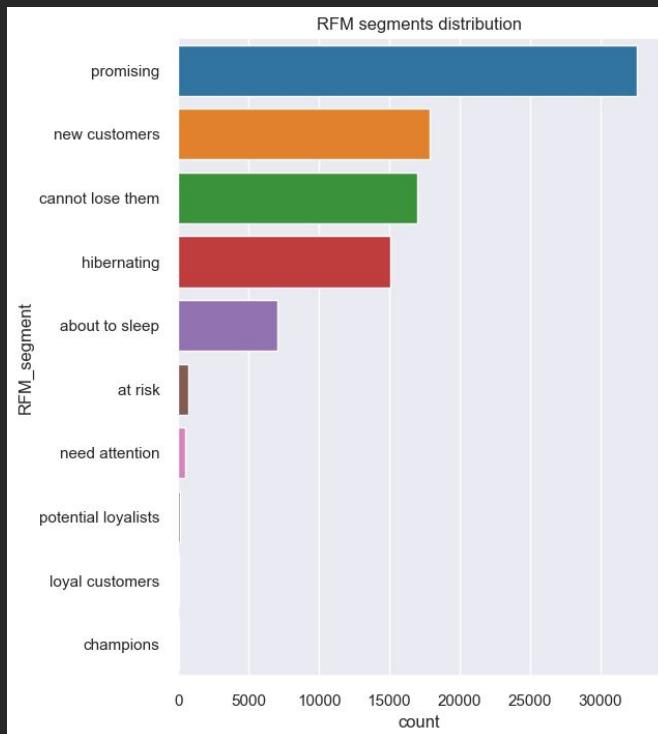
- **Recency** : date dernière commande
- **Frequency** : nombre de commandes
- **Monetary** : valeurs totales des commandes
- Notes de 1 à 5
- Segments de clients en fonctions des notes



SEGMENTATION RFM

SEGMENT	RFM SCORE	COMPORTEMENT	STRATEGIE POTENTIELLE
Champion	555, 554, 544, 545, 454, 455, 445.	Commande récente. Achats fréquents et fortes dépenses.	Récompenses, fidélité.
Loyal Customer	543, 444, 435, 355, 354, 345, 344, 335.	Commandes fréquentes, fortes dépenses.	Ventes incitatives, offrir des objets coûteux.
Potential Loyalist	553, 551, 552, 541, 542, 533, 532, 531, 452, 451, 442, 441, 431, 453, 433, 432, 423, 353, 352, 351, 342, 341, 333, 323.	Commandes récentes, moyennes dépenses.	Programme de fidélité, proposer des nouveaux produits
New Customer	512, 511, 422, 421, 412, 411, 311.	Commande récente	Marketing de bienvenue.
Promising	525, 524, 523, 522, 521, 515, 514, 513, 425, 424, 413, 414, 415, 315, 314, 313.	Commande récente, faible dépense.	Offrir des discounts
Need Attention	535, 534, 443, 434, 343, 334, 325, 324.	Commandes ancienne, dépense moyenne	Offres limitées pour réactiver le client
Cannot Lose Them	155, 154, 144, 214, 215, 115, 114, 113 .	Grosses commandes fréquentes et anciennes	Marketing nouveau produits, ne pas laisser à la concurrence
About To Sleep	331, 321, 312, 221, 213.	En dessous des moyennes	Reconnection et produits populaires
At Risk	255, 254, 245, 244, 253, 252, 243, 242, 235, 234, 225, 224, 153, 152, 145, 143, 142, 135, 134, 133, 125, 124.	Grosses commandes très anciennes	Marketing personnalisé pour ne pas perdre le client
Hibernating	332, 322, 231, 241, 251, 233, 232, 223, 222, 132, 123, 122, 212, 211.	Commandes anciennes et peu fréquentes	Marketing pour créer de la valeur

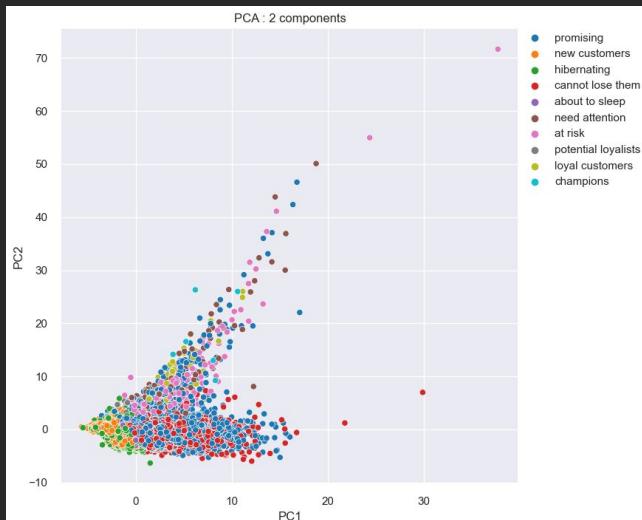
SEGMENTATION RFM



RÉDUCTION DE DIMENSION

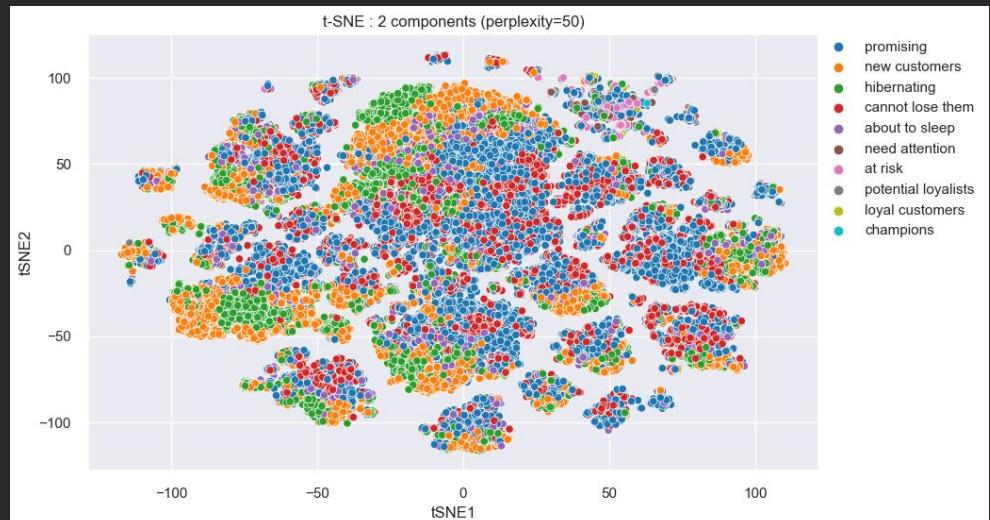
PCA

2 composantes
18.9% variance cumulée



t-SNE

2 composantes
Test de différentes valeurs de perplexité



CLUSTERING

Agglomerative clustering

- Clustering hiérarchique
- Approche ascendante
- Stratégie de fusion *Ward*

HDBSCAN

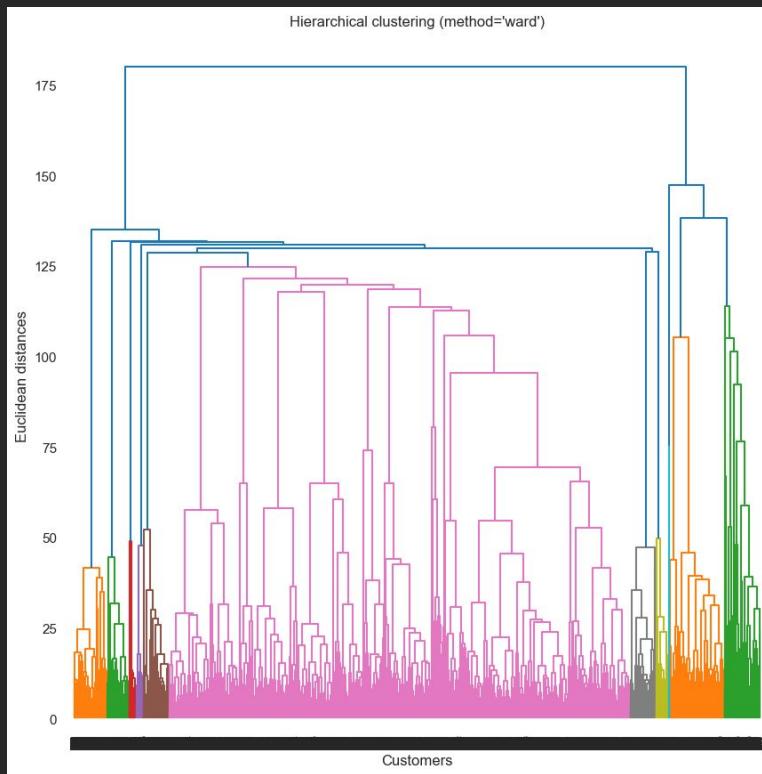
- Clustering hiérarchique et de densité
- Nombre de clusters pas nécessaires

KMeans

- Flat clustering
- Nombre de clusters nécessaires
- Calcul des centroides

$$\sum_{i=0}^n \min_{\mu_j \in C} (\|x_i - \mu_j\|^2)$$

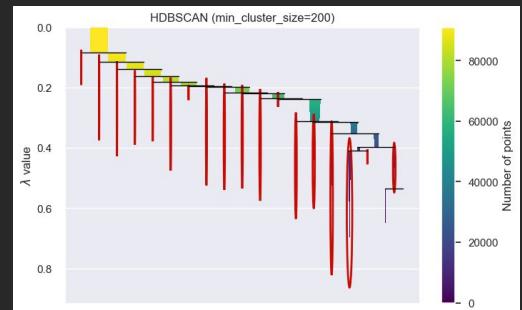
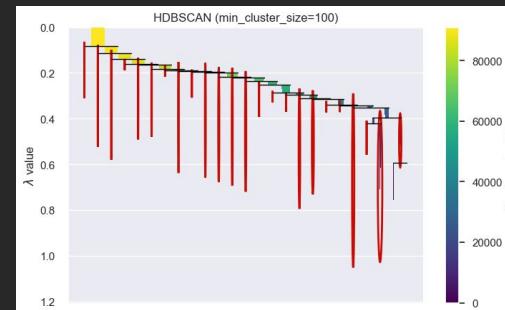
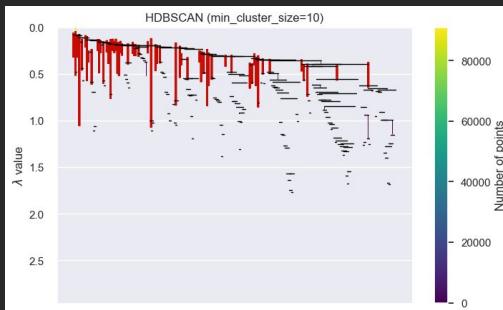
AGGLOMERATIVE CLUSTERING



- Apprentissage sur un échantillon du dataset : problèmes de mémoire
- Test différentes valeurs de nombre de clusters (4 à 8)
- Prédictions des labels
- Calcul des métriques

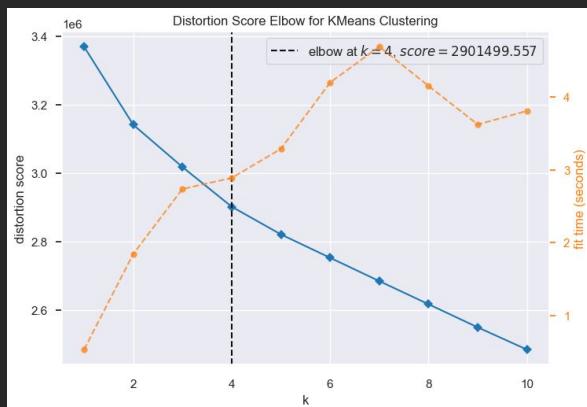
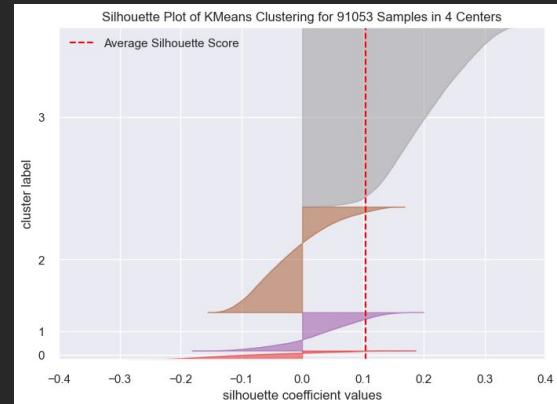
HDBSCAN

- Apprentissage sur le jeu complet (standardisé)
- Test de différentes valeurs de min_cluster_size (10, 20, 100, 200)
- Prédictions des labels
- Calcul des métriques



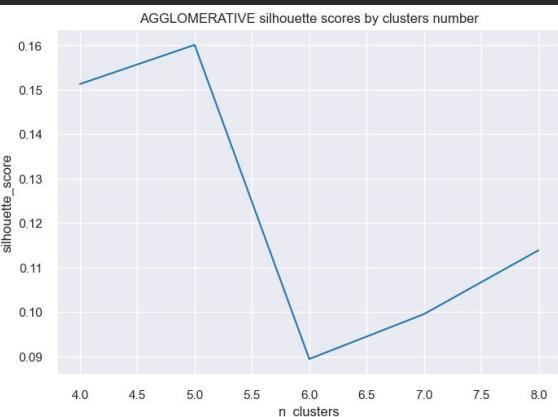
KMEANS

- Apprentissage sur le jeu complet (standardisé)
- Elbow method : nombre de clusters optimal (4)
- Test avec différent nombre de clusters (4 à 8)
- Prédiction des labels
- Calcul des métriques



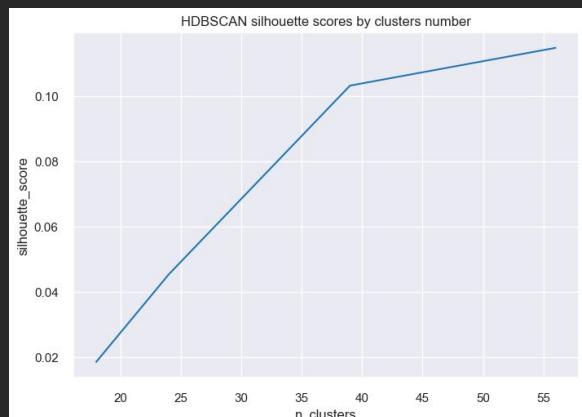
SILHOUETTE SCORE

$$s = \frac{b - a}{\max(a, b)}$$



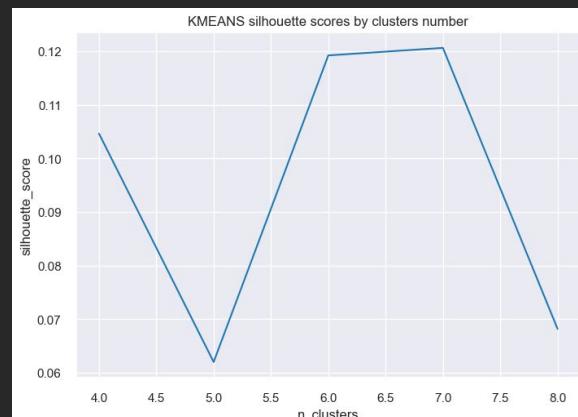
AGGLOMERATIVE

5 clusters



HDBSCAN

56 clusters

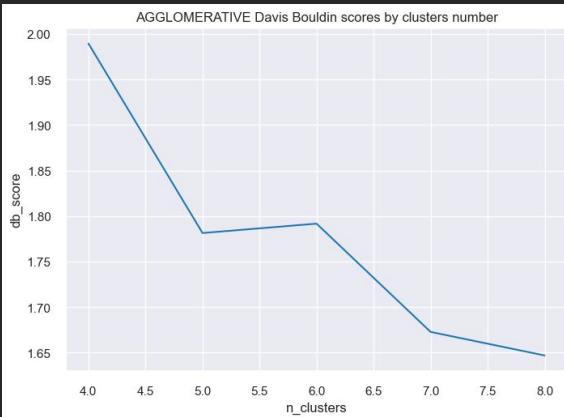


KMEANS

6 ou 7 clusters

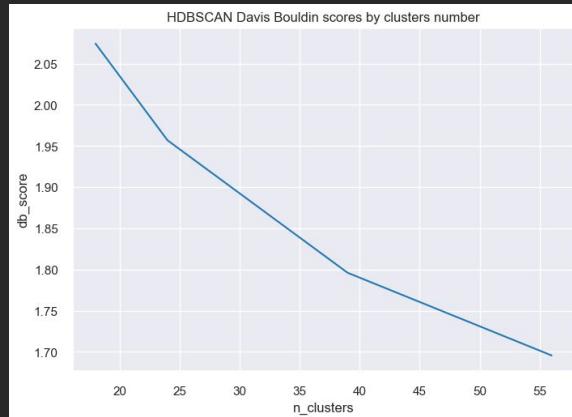
DAVIES BOULDIN SCORE

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{i \neq j} R_{ij}$$



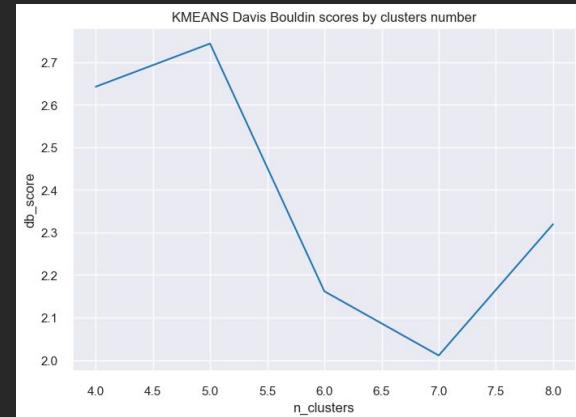
AGGLOMERATIVE

8 clusters



HDBSCAN

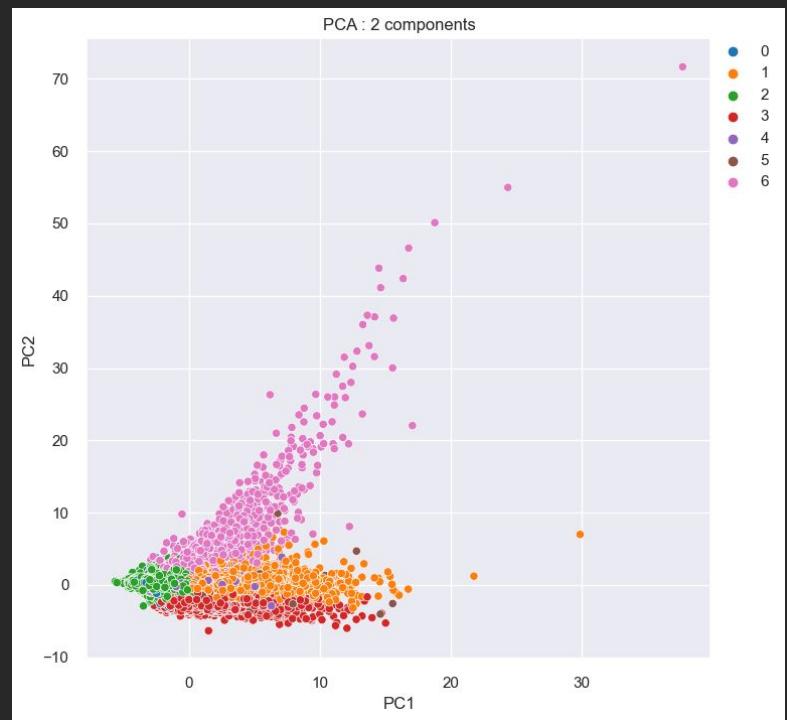
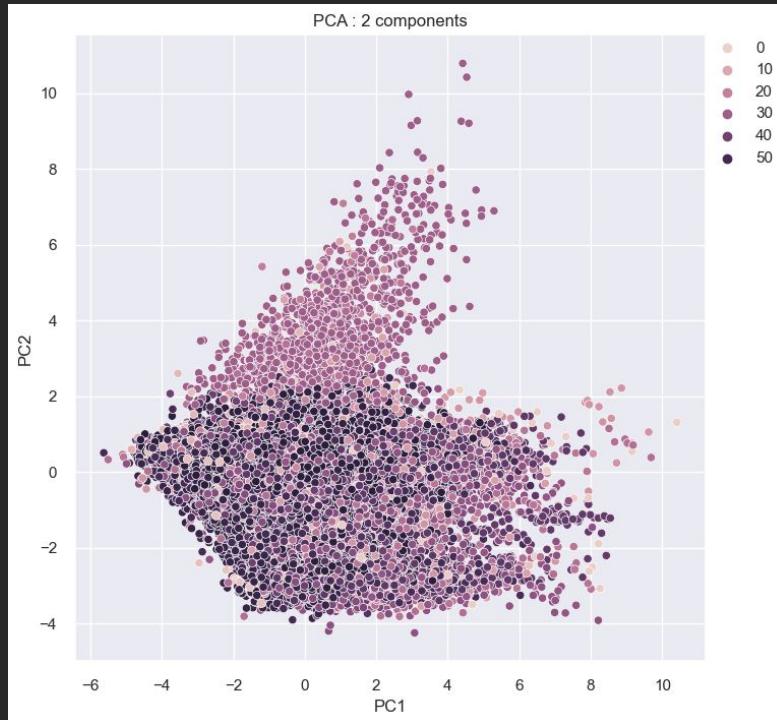
56 clusters



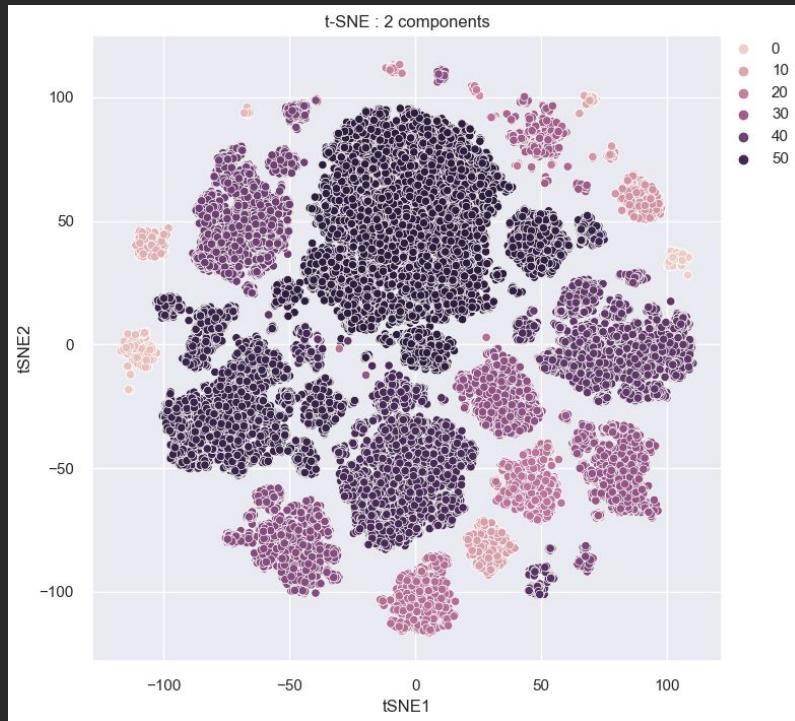
KMEANS

7 clusters

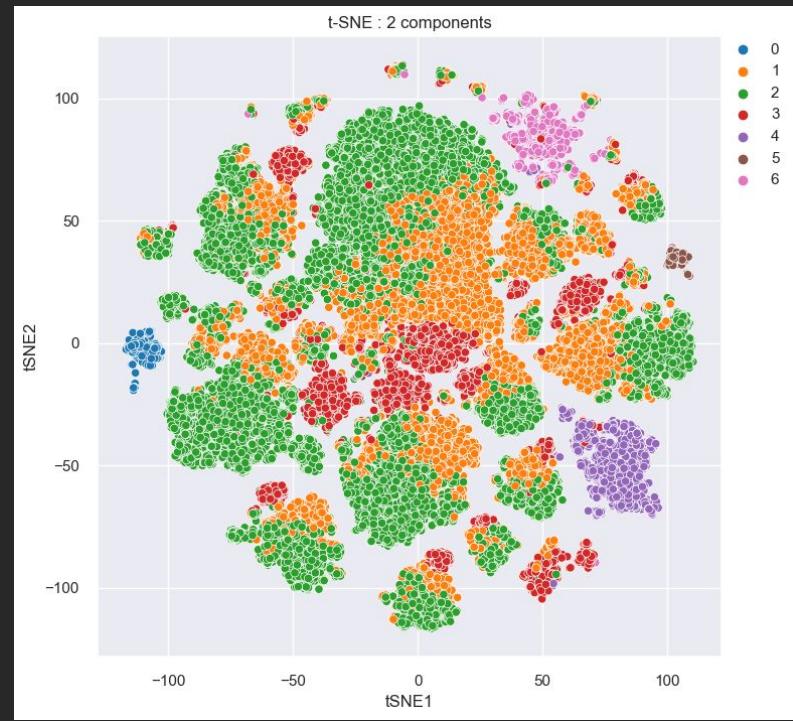
VISUALISATION : PCA



VISUALISATION : T-SNE



HDBSCAN 56 clusters



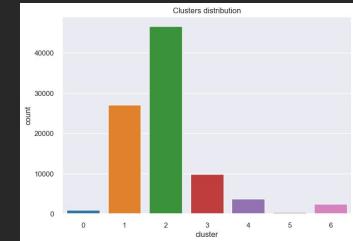
KMEANS 7 clusters

CHOIX DU MODÈLE

- Agglomerative clustering → problème de mémoire
- HDBSCAN : 56 clusters → trop pour une analyse marketing
- KMeans : 7 clusters → ok pour analyse marketing

⇒ Choix KMeans avec 7 clusters

ANALYSE DES CLUSTERS



CLUSTER	DESCRIPTION	STRATEGIE POTENTIELLE
Cluster 0	Commandes anciennes, faibles prix d'items, faible satisfaction, uniquement produits musiques et films	Essayer de les faire revenir en basant le marketing sur la catégorie musiques et films
Cluster 1	Deuxième plus grand groupe, grands paiements, grands prix d'items, grands échallonnages	Proposer des produits à fortes valeur avec des possibilités de paiement étalées
Cluster 2	Plus grand groupe, faibles paiements, faibles prix d'items, faible satisfaction, faible prix de livraison	Essayer de redonner confiance à la plateforme, offres spéciales, pas de frais de livraison,
Cluster 3	Grands temps de livraison, grands prix de livraison, grandes distances avec le vendeurs, paiements moyens	Proposer des produits "exotiques", offres spéciales d'autres villes
Cluster 4	Moyen sur tout, commandes anciennes, majorité de produits "other"	Mauvais clients : essayer de les faire revenir sinon abandonner
Cluster 5	Plus petit groupe, grands paiements, grand nombre de photos, grande satisfaction, produits "industry"	Proposer du marketing sur des produits "industry" uniquement. (Potentiellement des professionnels). Mettre en avant les articles qui contiennent des photos.
Cluster 6	Grands nombre de commandes, grands paiements, grand nombre d'items	Clients loyaux : fidéliser le client, proposer des nouveaux produits

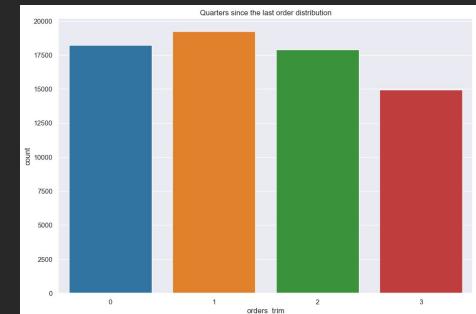
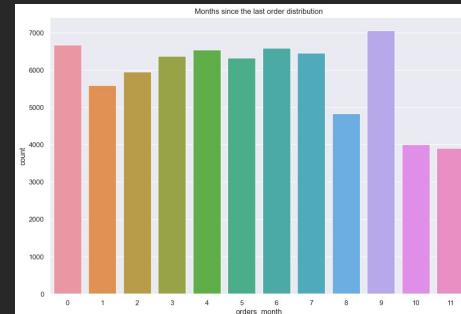
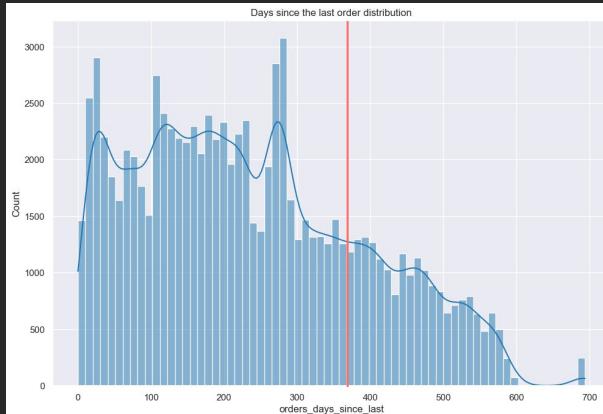


04

ANALYSE TEMPORELLE

ANALYSE TEMPORELLE

- Choix d'étudier sur 1 an seulement
- Utilisation du modèle KMeans
(7 clusters)
- Création de nouvelles variables de périodes :
 - Basé sur la date de dernière commande
 - Mois et trimestres



CLUSTERS ÉVOLUTION

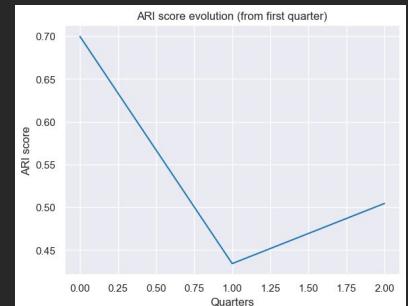
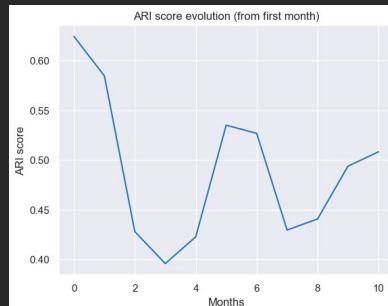


Évolution mois par mois

- Apprentissage sur le jeu complet
- Visualisation du nombre d'individu par cluster au cours des mois

ARI SCORE

- Apprentissage du modèle sur la période la plus ancienne et prédiction sur les périodes suivantes
- Apprentissage et prédiction du modèle sur les périodes suivantes
- Comparaisons des prédictions par ARI score



Stabilisation jusqu'à 4 mois

05

CONCLUSION



CONCLUSION

Conclusion

- Clustering non supervisés avec différents algorithmes
- Plusieurs types de segmentation
- Visualisation grâce à la réduction de dimension
- Analyse de métriques pour clustering
- Choix de 7 clusters et analyse de leurs spécificités
- Analyse de la stabilité dans le temps (environ 4 mois)

Améliorations

- Jeu de donnée très faible
- Plus de détails sur les clients
- Meilleurs choix de variables
- Plus de recherche sur l'analyse des clusters finaux
- Retirer outliers (groupe -1)

MERCI
DES QUESTIONS ?