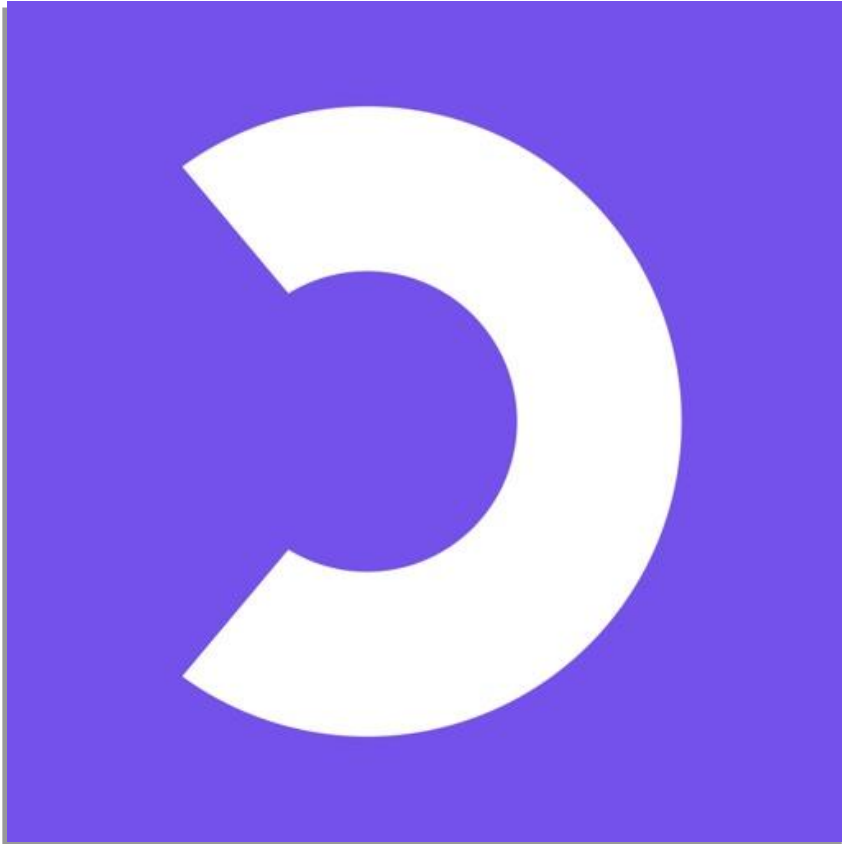
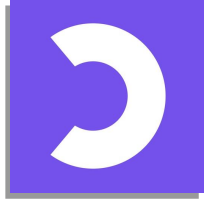


Projet 5 : Segmentez des clients d'un site e-commerce

Eva Rondeau



olist



Présentation

Objectif : segmentation sur l'ensemble des clients afin de comprendre les différents profils et estimation de la fréquence de mise à jour de la segmentation

Jeu de données (9 fichiers) entre 2016 et 2018 :

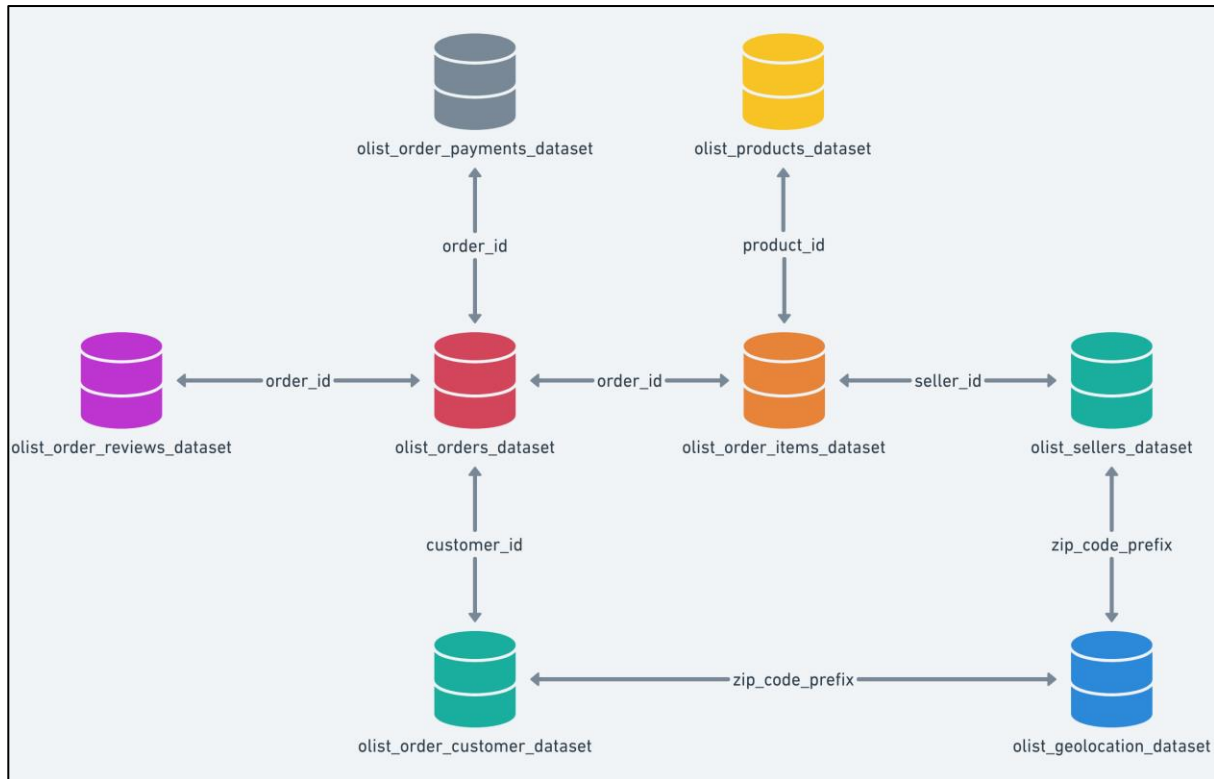
Type de fichier	Informations
Clients	Identifiant, Code Postal, Ville, Etat
Géolocalisation	Code Postal, Coordonnées GPS (latitude, longitude), Ville, Etat
Articles	ID commandes, ID produits, ID vendeurs, Dates, Prix, Coût de fret
Paielements	ID commandes, Mode de paiement, Nombre paiements, Montant
Catégories	Traduction anglaise

Type de fichier	Informations
Avis	ID commandes/avis, Notes, Commentaires (titre, message), Dates enquêtes de satisfaction
Commandes	ID, Statut, Dates (achat, confirmation, livraison réelle et estimée)
Produits	ID, Catégorie, Dimensions
Vendeurs	ID, Code Postal, Ville, Etat










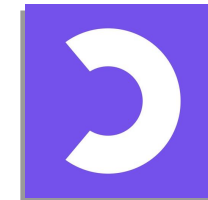
Présentation

1. Regroupement en un seul fichier



2. Feature Engineering

-  • *Nombre d'articles par commande*
 - Mono/Multi-articles
-  • *Délais livraison* : entre horaire estimée et réelle
 - En avance/retard ou à l'heure
-  • *Avis laissé* (variable binaire) : 0 (pas d'avis laissé), 1 (avis laissé)
-  • *Catégories communes*
-  • *Nombre catégories* différentes
-  • *RFM* : Récence (jours), Fréquence (nombre commandes), Montant (somme valeur commandes/client)
-  • *Distance* vendeur/client



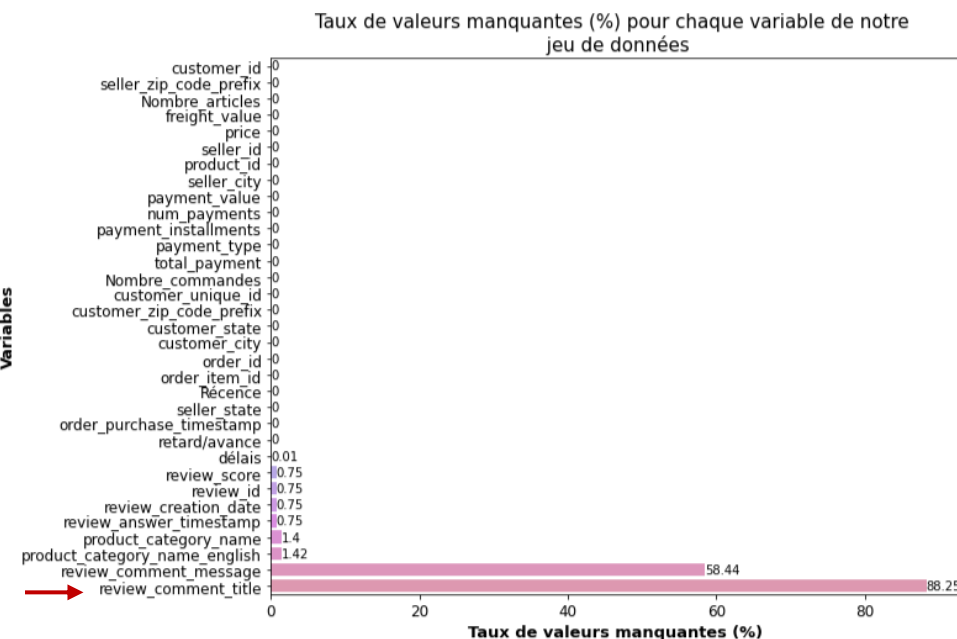
Présentation

3. Traitement des doublons



676 clients en doublons

4. Traitement des valeurs manquantes



Traitement au cas par cas :

- 8 clients NaN pour « Délais »
- 3 clients NaN pour « Type de paiement »

➡ **Suppression de ces clients**

- 858 clients NaN pour « Notes »

➡ **Imputation par la médiane**

- Catégories traduites en anglais

➡ **Traduction et attribution à une catégorie correspondante**

➡ **Conservation du reste (1597 clients)**

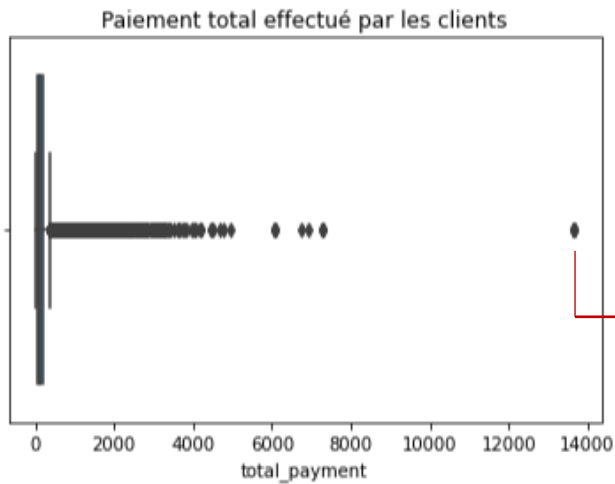
- 67 227 clients NaN pour « avis (messages) »

➡ **Conservation de ces clients**



Présentation

5. Valeurs extrêmes

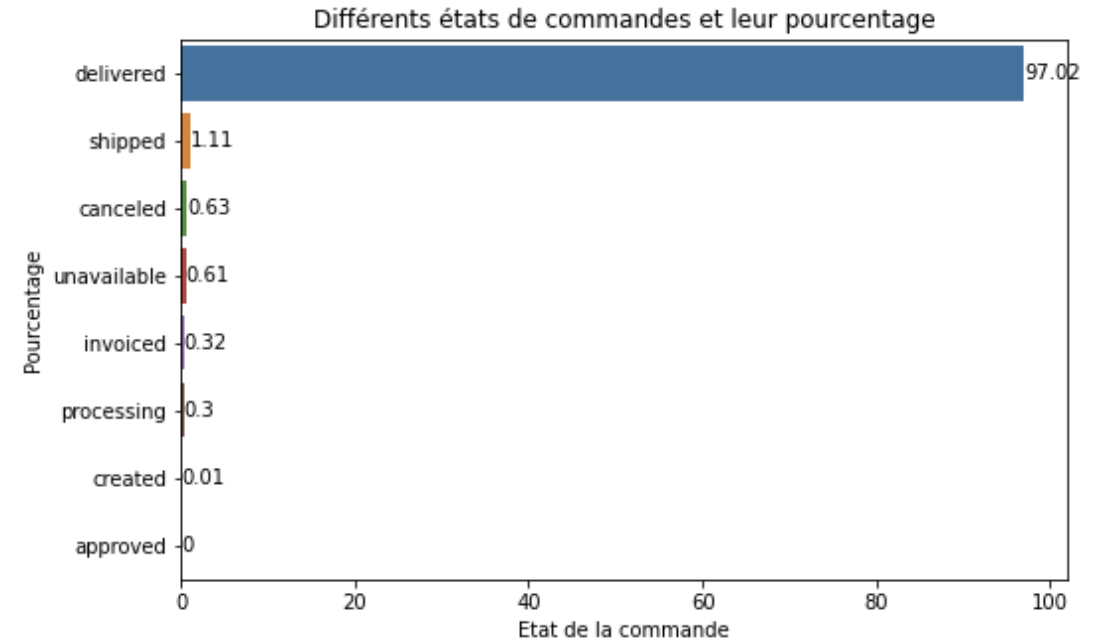


- 1 article commandé
- Article commandé en 8 fois
- Catégorie : « téléphonie fixe »

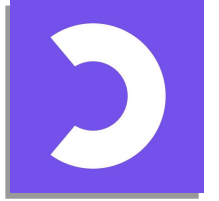
➡ Professionnel (bureau) ?
Commerçant ?

➡ **Suppression de ce client**

6. Commandes livrées



➡ Conservation des clients ayant reçu leur commande (97%)

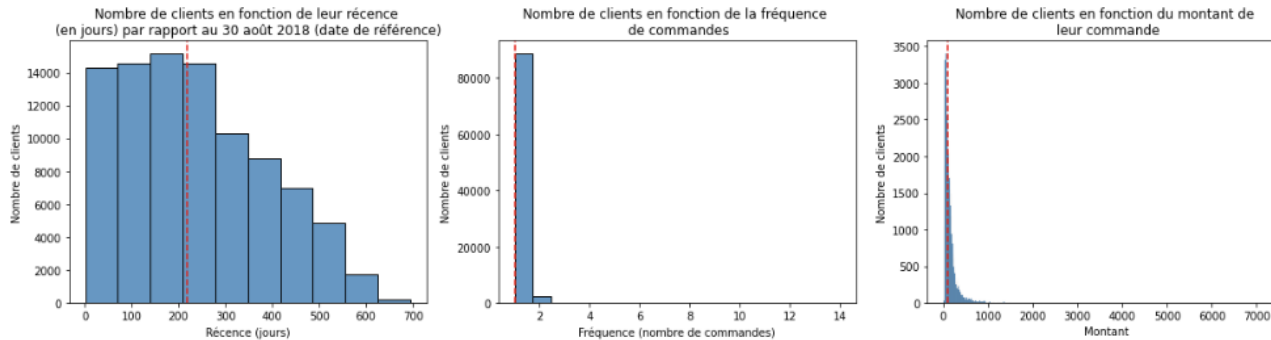


Analyse exploratoire

Variables RFM (Récence, Fréquence et Montant)

- Analysent la valeur d'un client
- Segmentation des clients
- Conception de stratégies de marketing plus ciblées

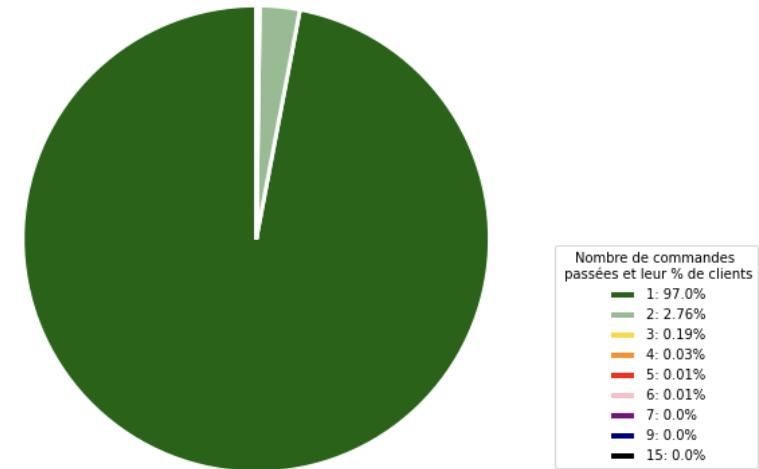
1. Analyse univariée



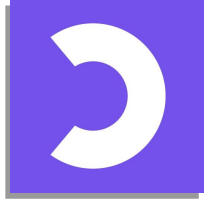
50% des clients ont :

- Commandé il y a moins de 218 jours
- Fréquence de 1 commande
- Montant inférieur à 105,38 R\$

Répartition sur un diagramme en camembert du % de clients selon le nombre de commandes passées.



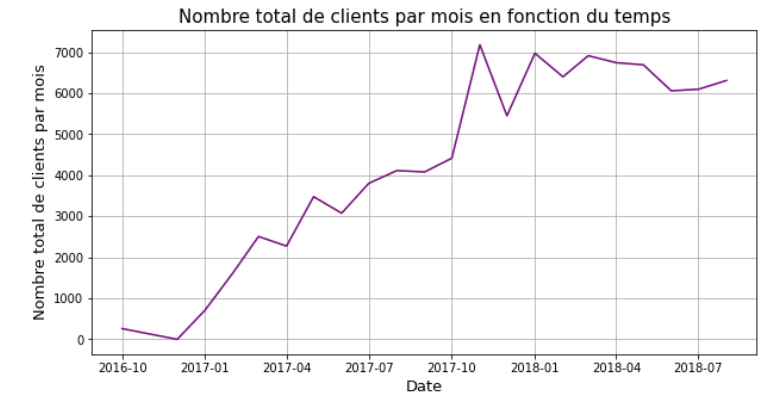
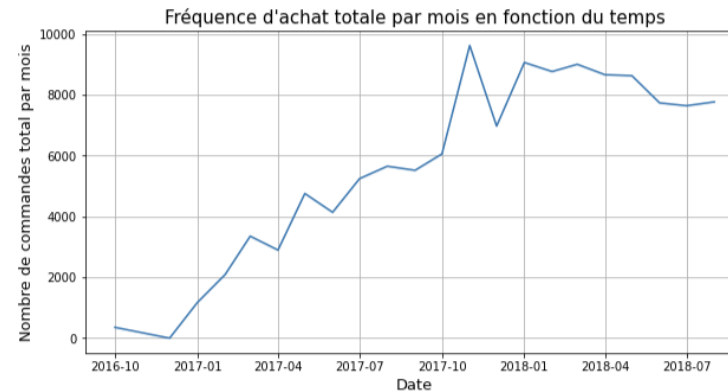
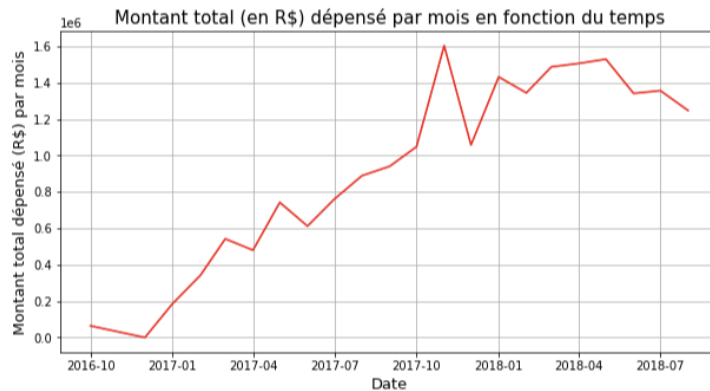
- 3% des clients ont commandé plus d'1 fois



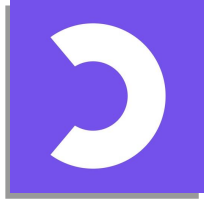
Analyse exploratoire

Variables RFM (Récence, Fréquence et Montant)

2. Analyse bivariable



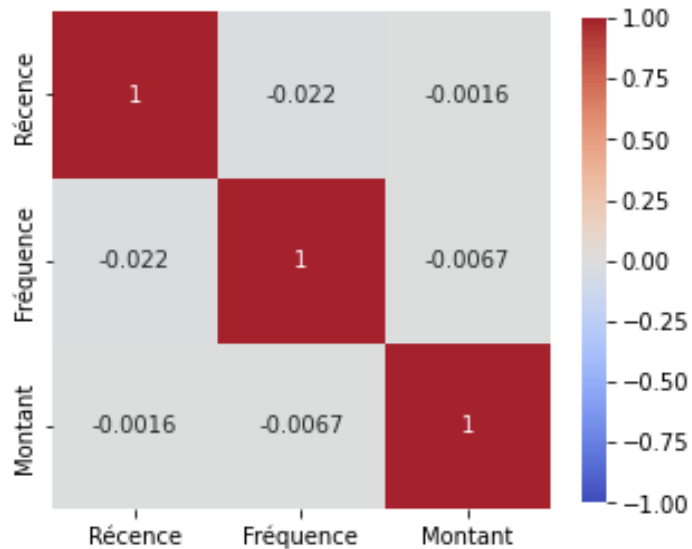
Augmentation du montant total dépensé et de la fréquence totale d'achat par mois en fonction du temps liées à l'augmentation du nombre de clients.



Analyse exploratoire

Variables RFM (Récence, Fréquence et Montant)

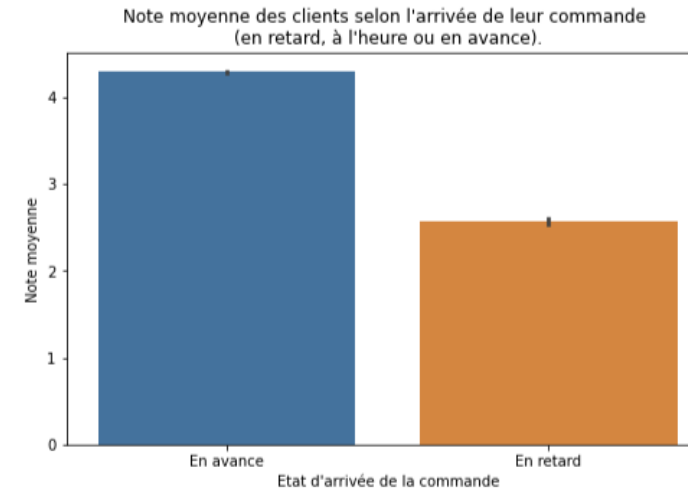
3. Analyse multivariée



Pas de corrélation observée

Autres variables

➤ Notes :



- En avance : note moyenne = 4,3
- En retard : note moyenne = 2,6

➡ Test Mann-Whitney : les **2 groupes sont significativement différents** au risque de 5%

Analyse exploratoire

Autres variables

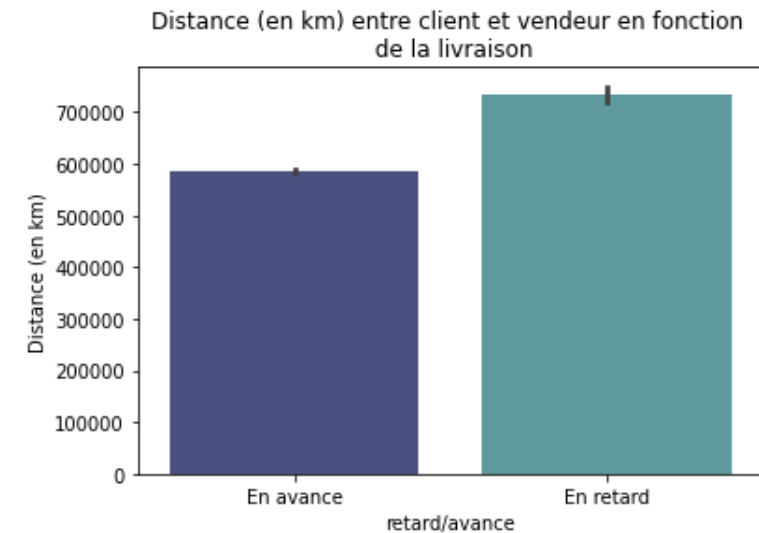
- Avis laissé :



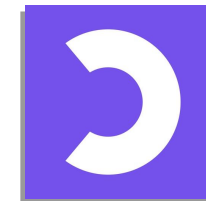
- Avis non laissé : note moyenne = 4,4
- Avis laissé : note moyenne = 3,8

➡ Test Mann-Whitney : les 2 groupes sont significativement différents au risque de 5%

- Distance client/vendeur :



➡ Test Mann-Whitney : les 2 groupes sont significativement différents au risque de 5%



Analyse exploratoire

Cartographies

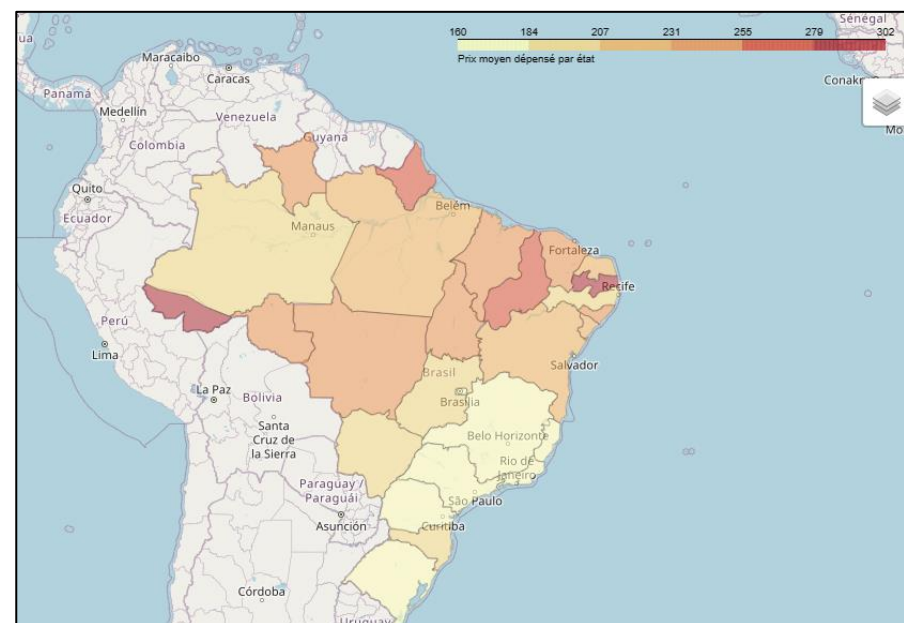
➤ Nombre de clients



Etats comptant le plus de clients :

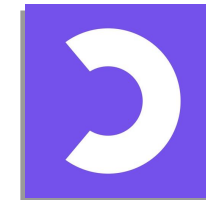
1. Sao Paulo
2. Rio de Janeiro
3. Minas Gerais

➤ Dépenses moyennes



Etats les plus dépensiers :

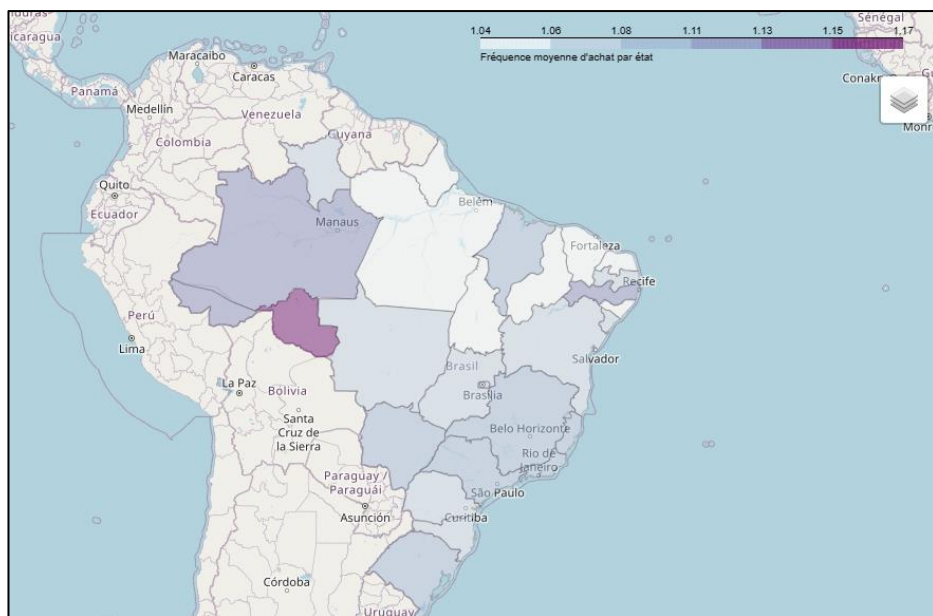
1. Acre
2. Paraíba



Analyse exploratoire

Cartographies

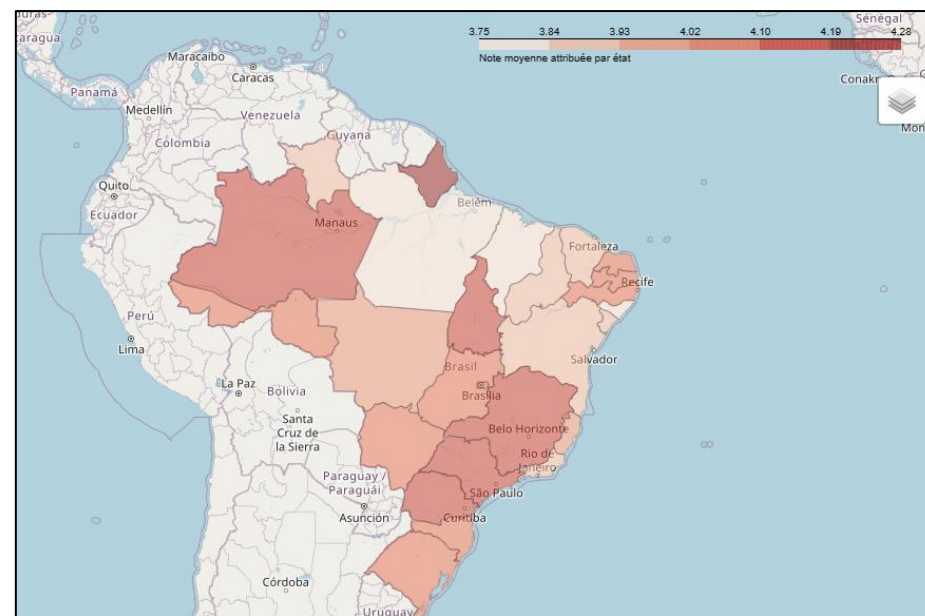
➤ Fréquences moyennes d'achat



Etat dont les achats sont les plus fréquents :

1. Rondonia

➤ Notes moyennes



Etat dont la note moyenne est la plus élevée :

1. Amapa

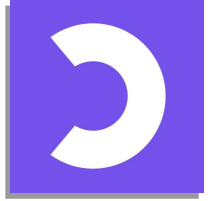
2. Paraba

3. Sao Paulo

4. Minas Gerais

5. Amazonas

6. Tocantins



Modélisation

Différents modèles à tester :

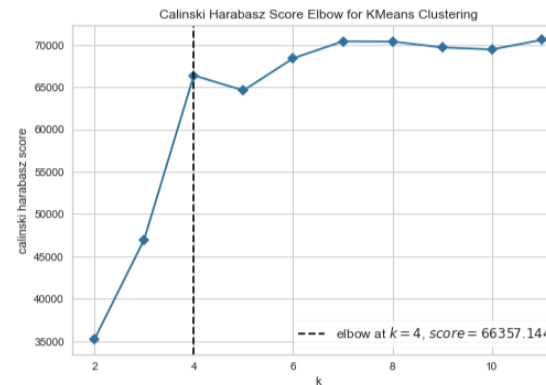
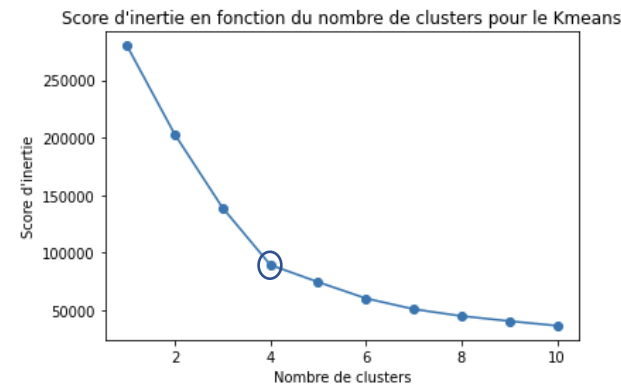
- Temps de calcul
- Segmentation client (groupes homogènes et distincts)
- Interprétabilité (groupes facilement identifiables)
- Stabilité dans le temps

Métriques calculées :

- **Score silhouette** (entre -1 et 1) : évalue si un point appartient au bon cluster.
- **Score Davies-Bouldin** : mesure la distance entre les centres de chaque cluster
- **Score Calinski-Harabasz** : mesure la variance inter et intraclusters

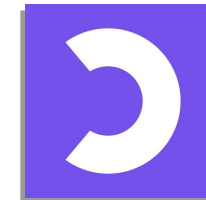
Variables RFM (Récence, Fréquence et Montant)

➤ Kmeans

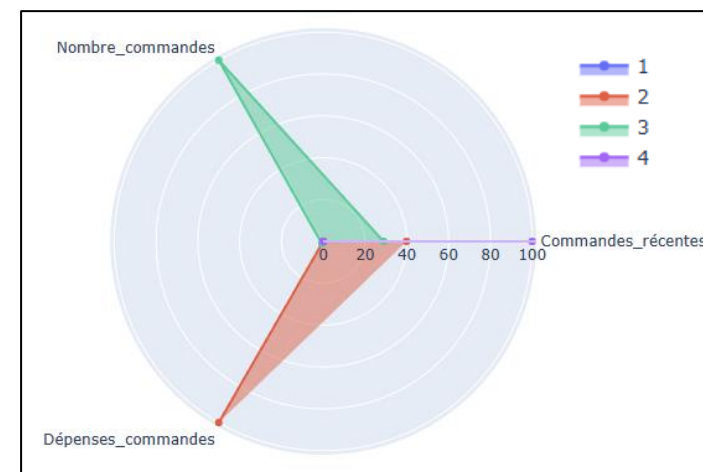
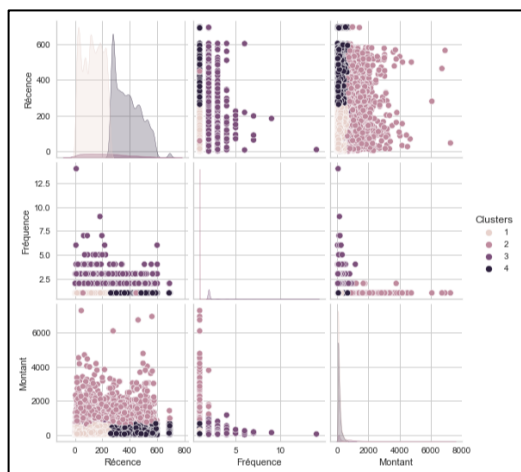
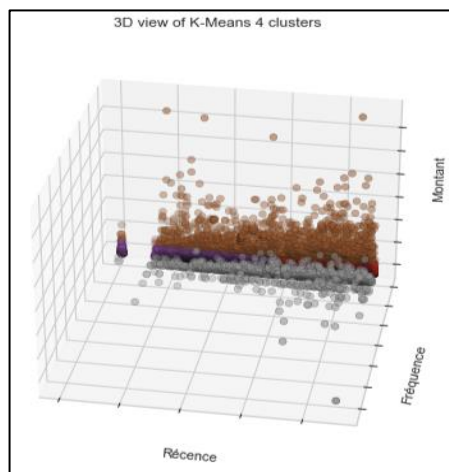


➡ Nombre optimal de clusters = 4

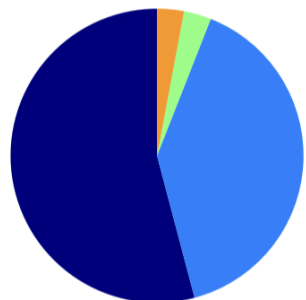
Modélisation

Variables RFM (Récence, Fréquence et Montant)

➤ Kmeans



Profils de clients selon les différents clusters



- Clients peu dépensiers mais récents, 54.2%
- Clients anciens, peu fréquents, faibles dépenses, 39.8%
- Clients avec de bonnes dépenses et fréquents, 3.0%
- Clients très dépensiers, 3.0%

Clients récents
54,2 %

Clients fidèles
3 %

Clients dépensiers
3 %

Anciens clients
39,8 %



Modélisation

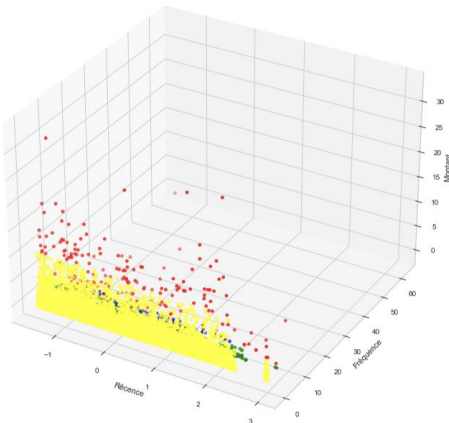
Variables RFM (Récence, Fréquence et Montant)

➤ DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

2 paramètres :

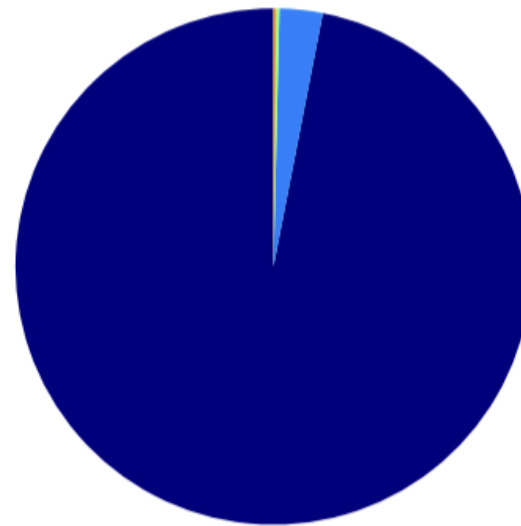
- Epsilon-voisinage (eps) : distance
- Nb min_points : nb de points minimum

➔ **Point appartenant à un cluster si au moins min_points sont à une distance inférieure à eps.**



Taux de bruit : 0,23 %

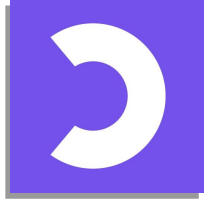
Profil de clients selon les différents clusters



- Clients anciens, dépenses les plus élevées, 96.9%
- Clients plus fréquents, 2.7%
- Bruit, 0.2%
- Clients les plus fréquents et les plus récents, 0.2%

Remarques :

- Clusters très **hétérogènes** (97% des clients appartenant au même cluster)
- Peu de groupes
- Difficilement nettement **identifiables**
- Faire ressortir les « bons » des « moins bons » clients



Modélisation

Variables RFM (Récence, Fréquence et Montant)

➤ CAH (Classification Ascendante Hiérarchique)

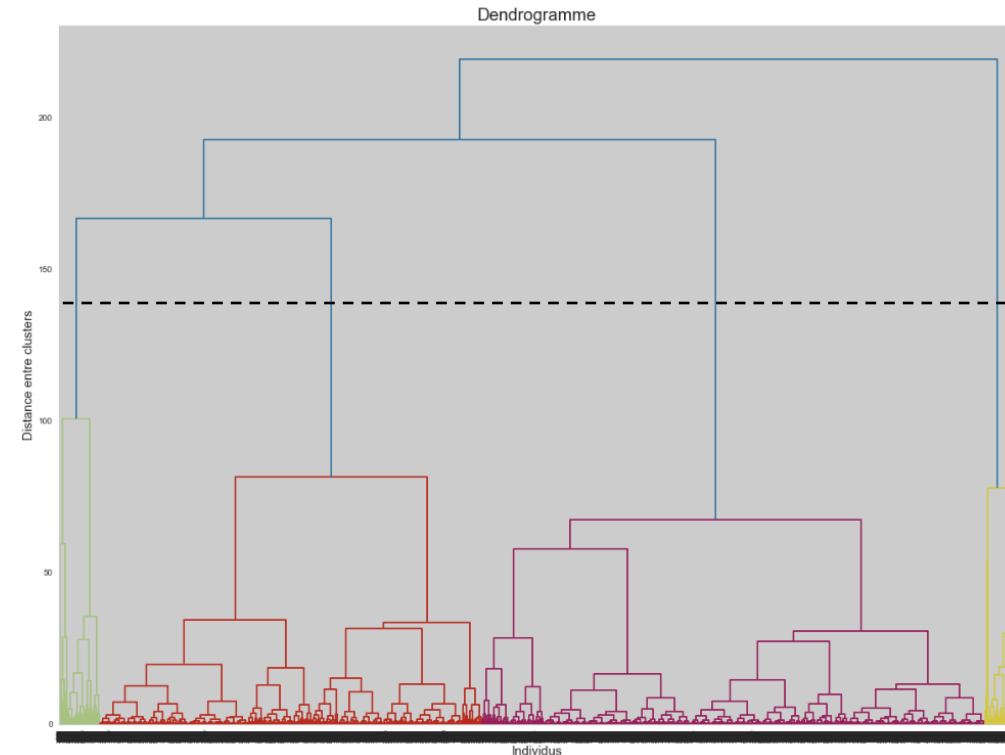
- Sélection d'un échantillon aléatoire : 30% de notre jeu de données
- Nombre clusters sélectionné : 4

Remarques :

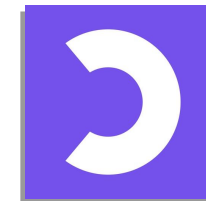
- Nécessite une échantillonnage :
 - Réduction du jeu de données
 - Perte d'information
 - Mauvaise représentativité
- Temps de calcul très long

➤ Résumé scores calculés

	Silhouette_score	Davies_bouldin_score	Calinski_harabasz_score	Time
Kmeans	0.486856	0.669867	66558.674708	99.187338
DBSCAN	0.706420	1.234585	14846.132632	607.415320
CAH	0.479362	0.715880	18387.541526	178.809870



Méthode du **Kmeans** retenue : rapidité, qualité clustering, interprétation



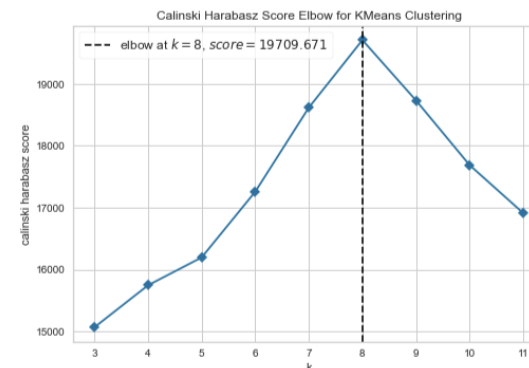
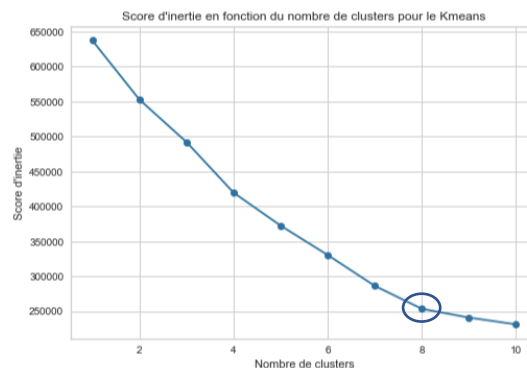
Modélisation

Variables RFM + autres variables

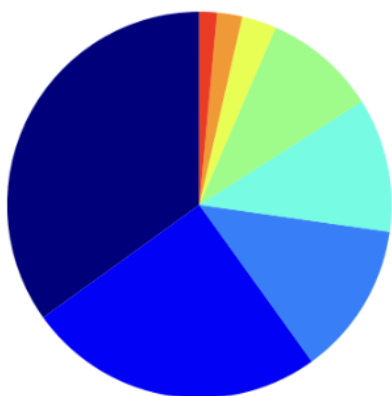
Variables :

- RFM (Récence, Fréquence, Montant)
- Note moyenne (satisfaction)
- Nombre d'articles
- Nombre de paiements
- Distance client/vendeur

➤ Kmeans



Profil de clients selon les différents clusters



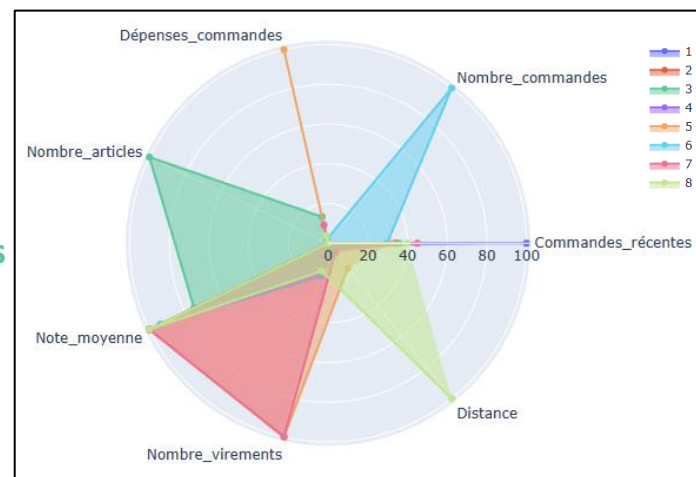
- Clients les plus récents, très satisfaits, faibles dépenses, faibles distances des vendeurs, 35.0%
- Anciens clients très satisfaits, dépenses faibles, 25.0%
- Clients non satisfaits, dépenses faibles, 12.8%
- Clients très satisfaits, nombreux virements, 11.2%
- Clients très satisfaits, à grandes distances des vendeurs, 9.5%
- Clients assez satisfaits, dépenses faibles, plusieurs commandes effectuées, 3.0%
- Clients assez satisfaits, bonnes dépenses, plusieurs articles par commande, 2.1%
- Clients très satisfaits, très dépensiers, nombreux virements, 1.4%

Anciens clients, faibles dépenses
25 %

Clients insatisfaits
12,8 %

Clients achetant plusieurs articles
2,1 %

Clients récents, moins distants des vendeurs
35 %

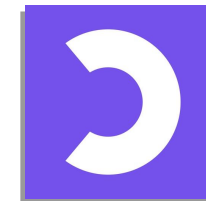


Clients dépensiers et en versements
1,4 %

Clients fidèles
3 %

Clients en versements
11,2 %

Clients distants des vendeurs
9,5 %



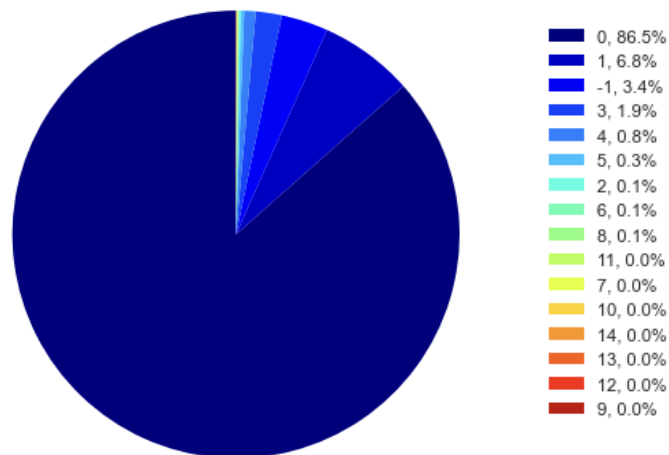
Modélisation

Variables RFM + autres variables

➤ DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

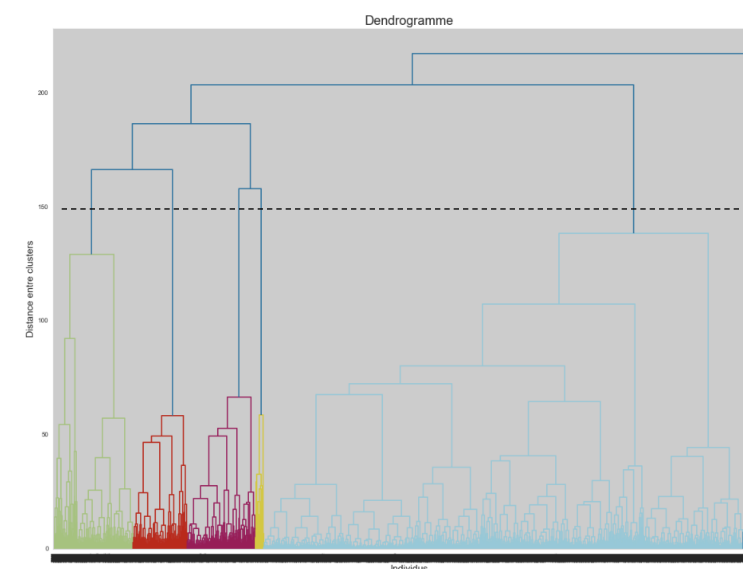
➤ CAH (Classification Ascendante Hiérarchique)

Profil de clients selon les différents clusters



Remarques :

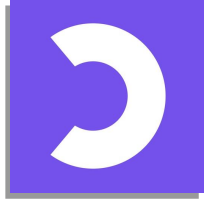
- 16 clusters identifiés
- 1 cluster correspondant au bruit (3,4 %)
- Clusters toujours hétérogènes
- Groupe contenant 86,5 % des clients



➤ Résumé scores calculés

	Silhouette_score	Davies_bouldin_score	Calinski_harabasz_score	Time
Kmeans	0.270539	1.045519	19959.000841	99.531030
DBSCAN	0.161984	1.610259	1005.732687	316.654359
CAH	0.316744	1.220026	4437.852706	270.200598

Méthode
du **Kmeans**
retenue



Maintenance

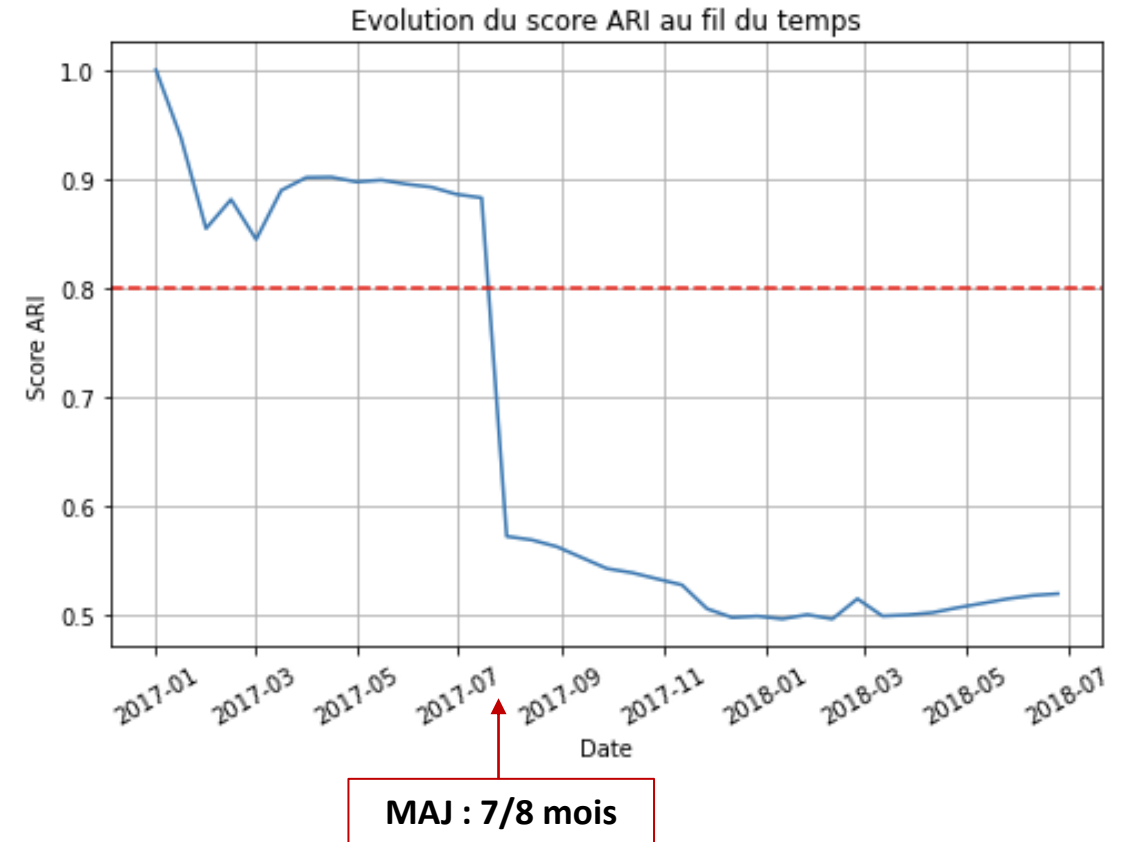
- Modèle de segmentation choisi : **Kmeans**
- Etude de la stabilité du modèle au cours du temps
- Fréquence de mise à jour de la segmentation client

Calcul score **ARI** (Adjusted Rand Index) :

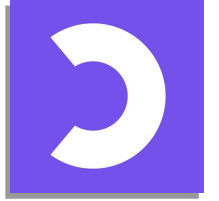
- Mesure de similitude entre 2 résultats de clustering
- Entre -1 et 1 (similitude totale)

Dans notre cas :

- Période de temps : 1 an et demi, pas de 15 jours
- T0 = 01-01-2017
- Modèle M0
- Calcul score ARI entre prédictions modèle M0 et prédiction modèle M0 + 15j, ...

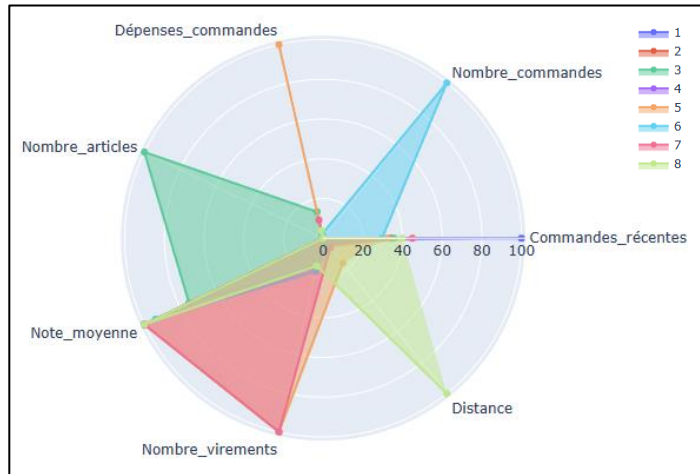


➡ **Mesure de similitude entre clusters prédits par notre modèle et évaluation de la tendance des clients à rester dans le même cluster au fil du temps**



Conclusion

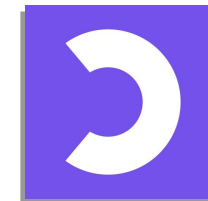
112,5 R\$



Lit, bain, table	Beauté et Hygiène	Meubles et décors	Agro, industrie et commerce
<p>Anciens clients, faibles dépenses (25%) 112,5 R\$</p> <p>Clients insatisfaits (12,8%) 132,9 R\$</p> <p>Clients fidèles (3%) 143,6 R\$</p> <p>Clients en versements (11,2%) 250,4 R\$</p> <p>Clients récents, moins distants des vendeurs (34,9%) 112,5 R\$</p>	<p>Clients distants des vendeurs (9,5%) 168,9 R\$</p>	<p>Clients achetant plusieurs articles (2,1%) 334,3 R\$</p>	<p>Clients dépensiers et en versements (1,4%) 1425,6 R\$</p>

Stratégies à mettre en place :

Clients récents (34,9%)	Anciens clients (25%)	Clients insatisfaits (12,8%)	Clients en versements (11,2%)	Clients distants (9,5%)	Clients fidèles (2,9%)	Plusieurs articles (2,1%)	Clients dépensiers (1,5%)
<ul style="list-style-type: none"> Amenés à revenir (voir catégories produits) Vendeurs locaux 	<ul style="list-style-type: none"> Produits similaires (newsletter) 	<ul style="list-style-type: none"> Traitement au cas par cas (enquêtes de satisfaction) Proposer une solution (proximité) 	<ul style="list-style-type: none"> Options de paiements en plusieurs fois (gestion budget) 	<ul style="list-style-type: none"> Produits similaires à proximité Offres de livraison plus rapide ou moins chères 	<ul style="list-style-type: none"> Offres spéciales 	<ul style="list-style-type: none"> Réductions seuil nombre d'articles 	<ul style="list-style-type: none"> Offres spéciales



Conclusion

Scores variables RFM :

	Silhouette_score	Davies_bouldin_score	Calinski_harabasz_score	Time
Kmeans	0.486856	0.669867	66558.674708	99.187338
DBSCAN	0.706420	1.234585	14846.132632	607.415320
CAH	0.479362	0.715880	18387.541526	178.809870

Scores variables RFM + autres variables :

	Silhouette_score	Davies_bouldin_score	Calinski_harabasz_score	Time
Kmeans	0.270539	1.045519	19959.000841	99.531030
DBSCAN	0.161984	1.610259	1005.732687	316.654359
CAH	0.316744	1.220026	4437.852706	270.200598

- Contribution au choix du modèle de segmentation

D'un point de vue métier :

- Ajout de variables -> affinement des profils de clients

