Ingénierie IA

Kindi BALDE

Segmentation des clients du site Olist
(un site e-commerce Brésilien)

Présentation des données
Problématique métier

Olist est une société de E-Commerce Brésilien qui vend des produits de tout type et par définition qui met en relation vendeurs et acheteurs.

Les acheteurs se connectent sur le site, font des commandes, et ces dernières sont prises en charge par les vendeurs qui les leurs expédient par la suite.

Olist souhaite fournir à ses équipes d'E-Commerce une segmentation des clients qu'elles pourront utiliser au quotidien pour leurs campagnes de communication.

Pour ce faire, nous nous devons de comprendre les différents types d'utilisateurs grâce à leur comportement et à leurs données personnelles.

Pour au final fournir à l'équipe marketing d'Olist une description actionable de notre segmentation et de sa logique sous-jacente pour une utilisation optimale, ainsi qu'une proposition de contrat de maintenance basée sur une analyse de la stabilité des segments au cours du temps.

Mission

Aider les équipes d'Olist à comprendre les différents types d'utilisateurs. Nous utiliserons donc des méthodes non supervisées pour regrouper l'ensemble des clients de profils similaires. Ces catégories pourront être utilisées par l'équipe marketing pour mieux communiquer.

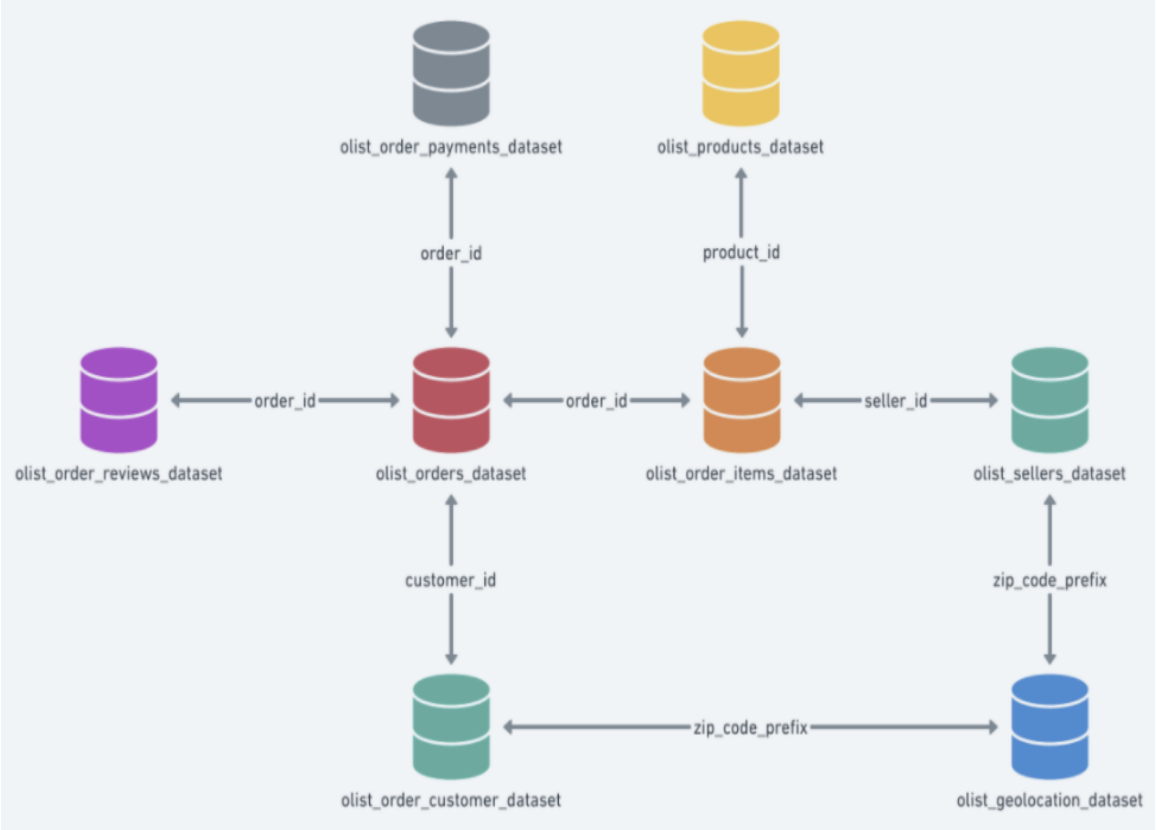
sommaire

1. Garbage in Garbage out
2. EDA et Feature engineering
3. RFM outils marketing
4. Classification avec la méthode des KMeans
5. Proposition de solution de maintenance des Clusters

1. Garbage in Garbage out

Challenges rencontrés

1. Le regroupement des données est fait suivant une logique bottom-up;
2. Pas de gestion de valeurs nulles à faire;
3. choix d'axes de regroupement pour chaque dataset;
4. choix des tables.



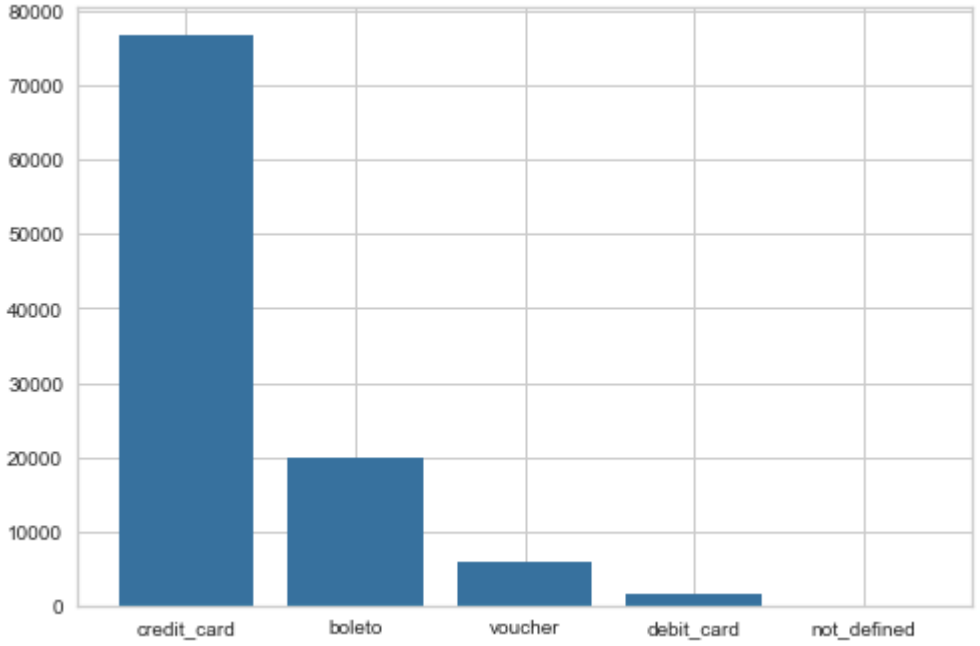
2. EDA and Feature engineering (1/3)

Création de nouvelles variables

1. nombre de type de méthode de paiement (Credit Card, Debit Card, Voucher, Boleto) par commande.
2. le review score moyen par commande.
3. Encodage des variables nominales des cités et des grandes villes de localisation des clients.

Deux tables pour l'étude

1. table clients : le client (customer_unique_id) est la clé primaire.
2. table commandes : le numéro de commande (order_id) est la clé primaire.



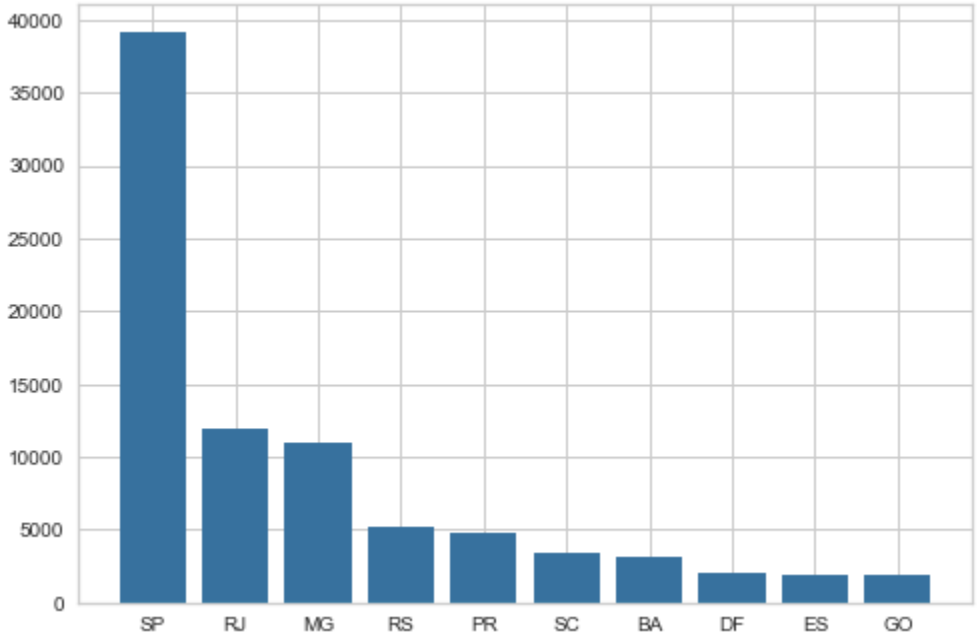
2. EDA and Feature engineering (2/3)

Création de nouvelles variables

1. nombre de types de méthode de paiement (Credit Card, Debit Card, Voucher, Boleto)
2. le review score moyen par commande
3. Encodage des variables nominales des cités et des grandes villes de localisation des clients

Deux tables pour l'étude

1. table clients : le client (customer_unique_id) est la clé primaire.
2. table commandes : le numéro de commande (order_id) est la clé primaire.

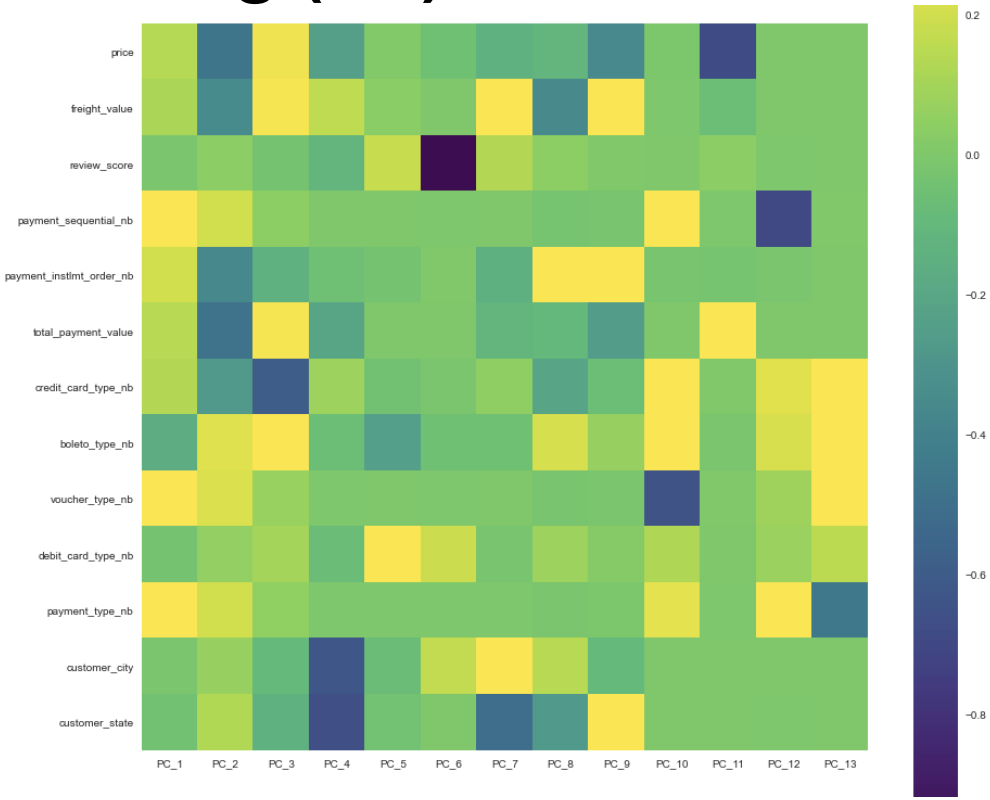


2. EDA and Feature engineering (3/3)

Analyse PCA

Les 6 premiers axes regroupent plus de 80% de l'information globale apportée par les variables.

- Une analyse des corrélations montre que les variables relatives au type de paiement se positionnent sur le premier axe, PC_1.
- L'axe PC_3, est l'axe du price, du freight value, du total payment value, et curieusement du boleto_type_nb.
- PC_2 ressemble un peu au PC_1



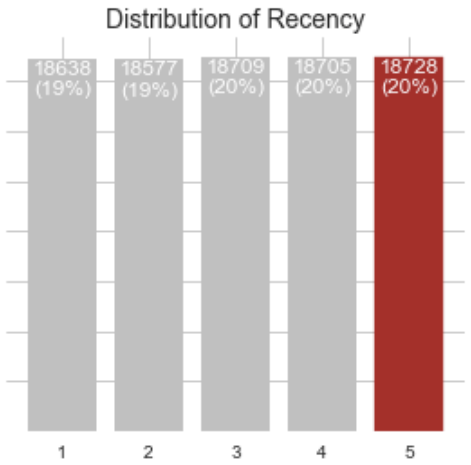
3. RFM - Outils marketing (1 / 2)

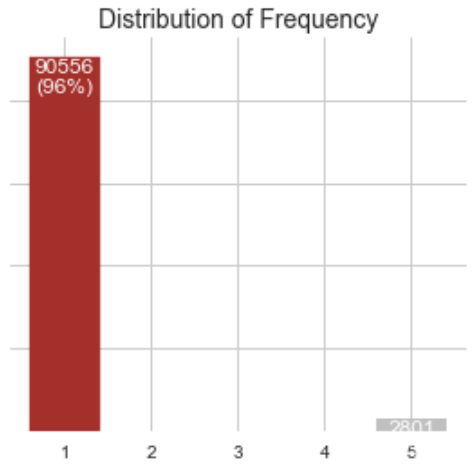
RFM est un outil de segmentation de clients qui nous permet à partir de trois indicateurs de construire une stratégie marketing. Ces trois indicateurs sont :

- Recency : sur une période d'étude donnée, à quand date la dernière fois que le client nous a rendu visite?
- Frequency: sur cette même période, combien de fois est-il venu?
- Monetary : combien a-t-il dépensé sur cette période?

Nous avons mené cette étude sur toute la période avec la recommandation du mentor.

Comme le montre le graphique ci-contre, 96% de nos clients ne sont venus qu'une seule fois durant ces deux ans d'études; 20% des clients sont venues récemment nous rendre visite.





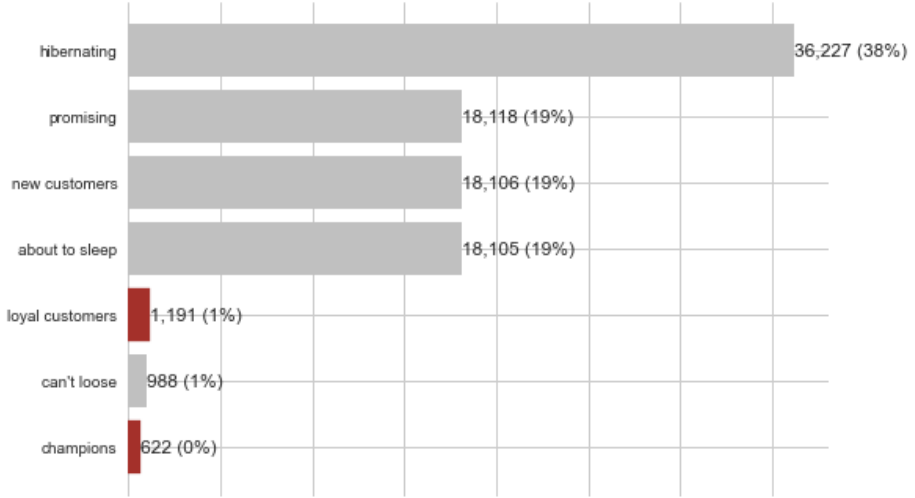
3. RFM - Outils marketing (2/2)

RFM est un outil de segmentation de clients qui nous permet à partir de trois indicateurs de construire une stratégie marketing. Ces trois indicateurs sont :

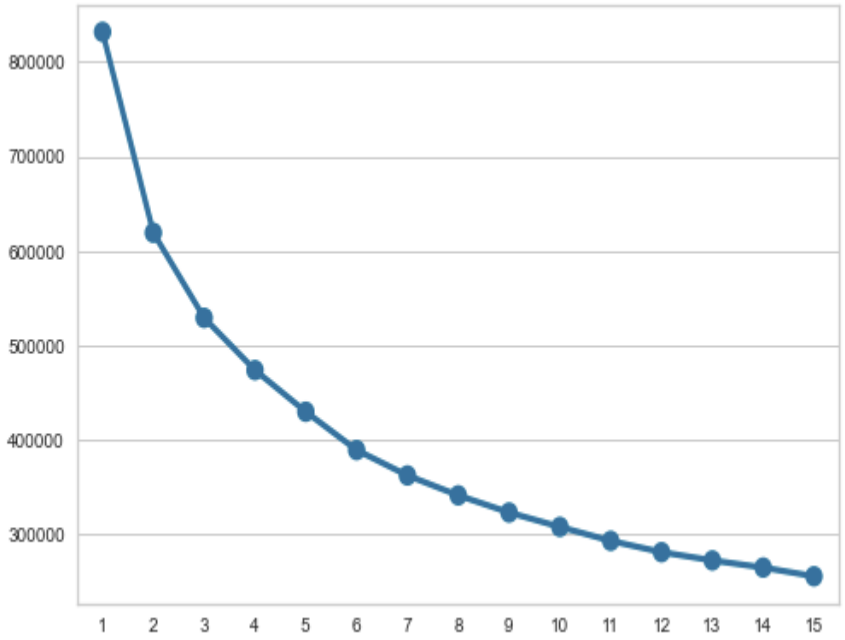
- Recency : sur une période d'étude donnée, à quand date la dernière fois le client nous a rendu visite?
- Frequency: sur cette même période, combien est-il venu?
- Monetary : combien a-t-il dépensé?

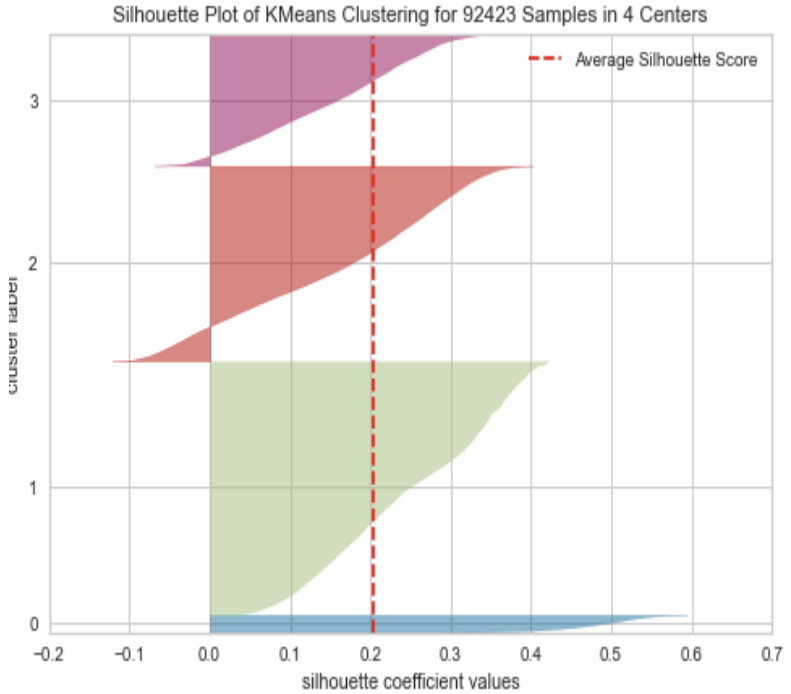
Nous avons mené cette étude sur toute la période, sur recommandation du mentor.

Comme le montre le graphique ci-contre, à peu près 1% de nos clients sont des "loyal customers"; 622 clients parmi eux sont des champions; 38% sont en état d'hibernation.

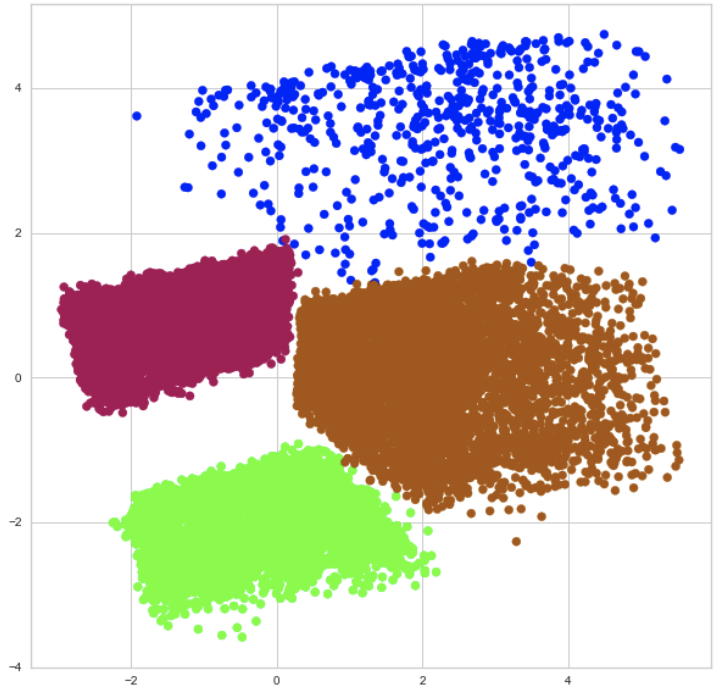


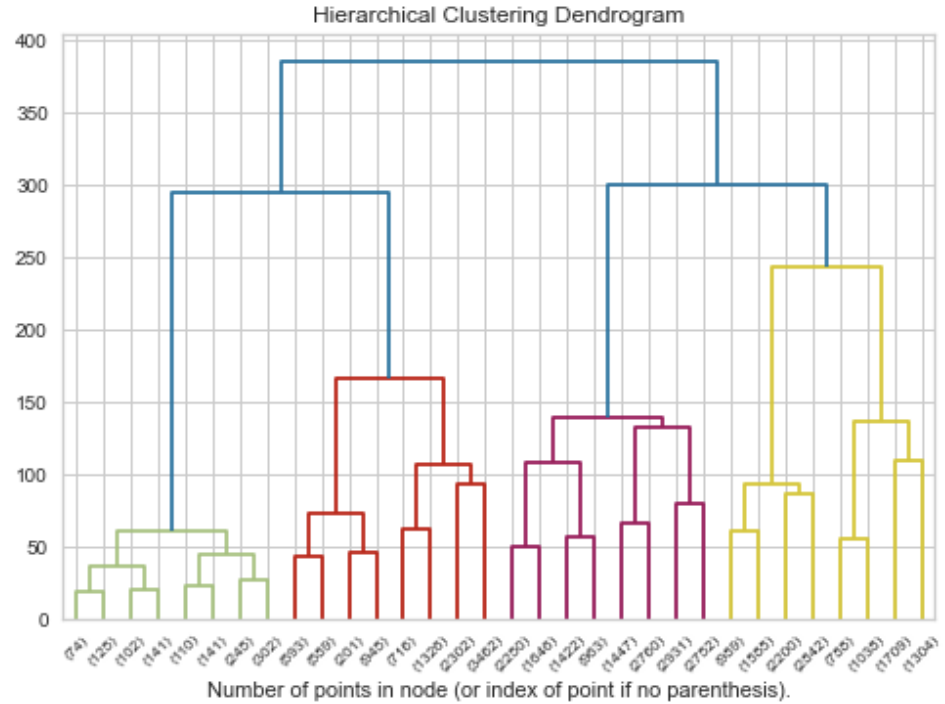
4. Classification avec la méthode des KMeans (1/3)





4. Classification avec la méthode des KMeans (2/3)





4. Classification avec la méthode des KMeans (3/3)

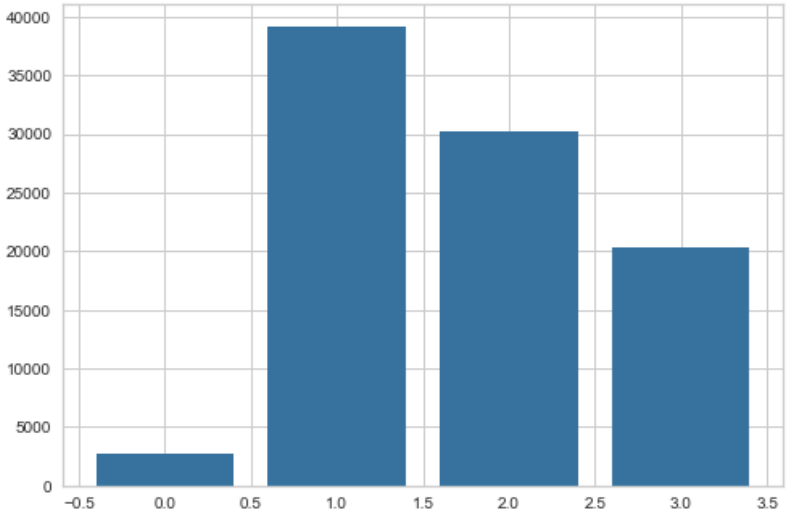
cluster	CustomerNumberPerCluster	ProductPrice	CustomerTotalExpense	RecencyMean	FrequencyMean	MonetaryMean
0	20342	1601.66	6093.23	2.87	1.0	2.31
1	39193	1400.34	4808.20	3.02	1.0	2.09
2	30112	4150.10	17273.70	3.05	1.0	4.55
3	2776	3708.55	16564.56	3.15	5.0	4.14

1. La classe labellisée 0 est la classe des clients qui ont une dépense correcte et qui ne sont venus qu'une seule fois.

1. La classe labellisée 1 est la classe des moins dépensiers. Ils ne sont venus qu'une seule fois sur la période d'étude. Ils sont plus nombreux et représente plus de 42% des clients.

1. La classe labellisée 2, même s'ils ne sont venus qu'une seule fois, sont ceux qui dépensent le plus et ont tendance à aller vers des produits très coûteux.

1. La classe labellisée 3 représente la classe des 3%, de clients qui viennent souvent et qui dépensent beaucoup.



5. Maintenance des Clusters

En partant de la classification que nous avons mise en place dans l'étape précédente, nous allons construire une analyse de stabilité des classes.

Pour ce faire, nous avons utilisés plusieurs scores, notamment le ARI, le Adjusted Rand Index, qui permet de comparer la correspondance (ou la similarité) entre deux classes distinctes.

En prenant notre classification par la méthode des KMeans sur toute les données comme étant notre "ground truth", nous avons découpé les données en deux parties (avant 2008 et après 2008). sur la période avant 2008, nous avons reconstruit le clustering et nous nous sommes assurés que notre ARI est bien égale à 1 (en comparant le nouveau clustering et l'ancien jusqu'à la période 2018). Nous avons ensuite ajouté de façon séquentielle des données mensuelles de janvier à Août 2018. Nous avons constaté une détérioration rapide des indicateurs comme montré sur le graphe et tableau ci-contre.

Nous préconisons une mise à jour de la classification tous les mois.

	V_SCORE	ADJ_RAND_SCORE	MUTUAL_INF_SCORE
1	0.6737	0.6546	0.6737
2	0.5590	0.4799	0.5589
3	0.4938	0.3724	0.4938
4	0.4611	0.3163	0.4611
5	0.4443	0.2884	0.4442
6	0.4371	0.2787	0.4371
7	0.4358	0.2784	0.4358
8	0.4369	0.2840	0.4369

Le Adjusted Rand Score pour les 8 derniers mois de l'année 2018