olist store

Data scientist
Projet 5

Segmenter des clients d'un site e-commerce

Dolores Valide
Mentor : Adrien Germond

Sommaire

A. Problématique

B. Nettoyage et analyse des données

C. Segmentation

D. Maintenance du modèle

Problématique

Olist, entreprise brésilienne proposant des solutions de ventes sur des marketplace en ligne

- Proposer une segmentation des clients en vue des campagnes publicitaires quotidiennes.
- Proposer un contrat de maintenance de la segmentation

Préparation des données

Etapes du nettoyage

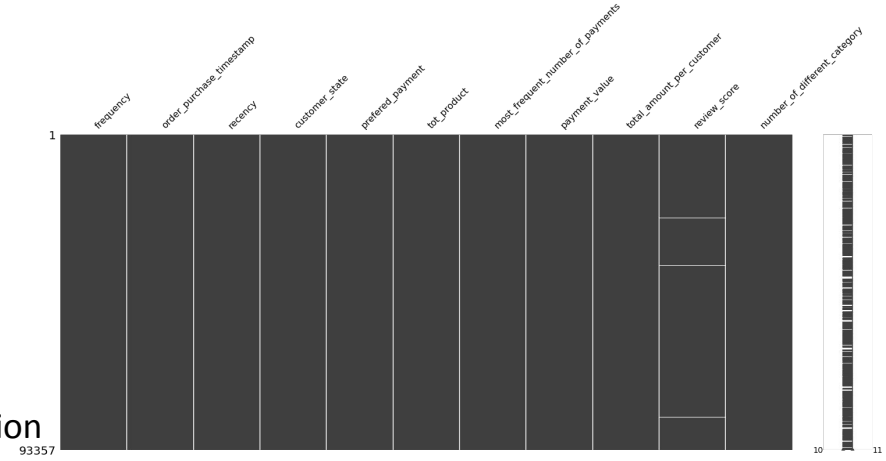
On dispose de 8 datasets

- Suppression des duplicatas et corrections des villes dans **geolocation**
  - 261831 doublons et retrait des accents
- Valeurs manquantes de **orders** et **order_review**
  - Moins de 5% pour orders et 90% et 60% dans orders_review sur les commentaires et leur titre
- Réduction du nombre de catégories
  - De 73 a 14 catégories
- Merge de 7 dataset, jointure à droite
- Conservation des commandes livrées
  - Suppression de 3420 commandes
- Dataset agrégé à la maille client

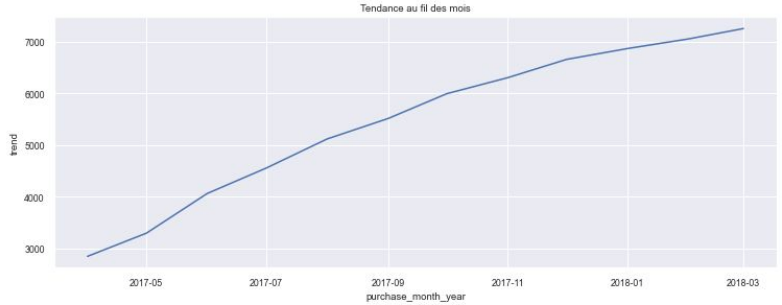
Etapes du feature engineering

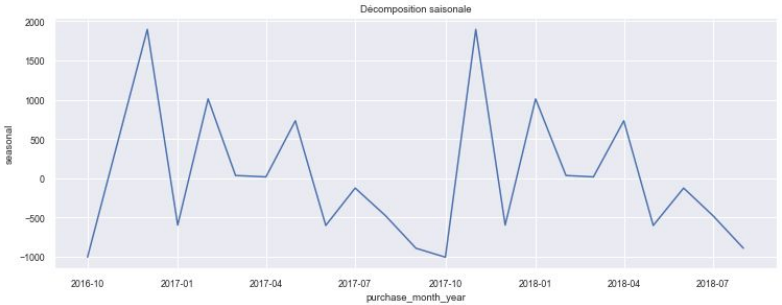
- Nombre de jours depuis le dernier achat
- Nombre de produits par commandes

Dataset après agrégation



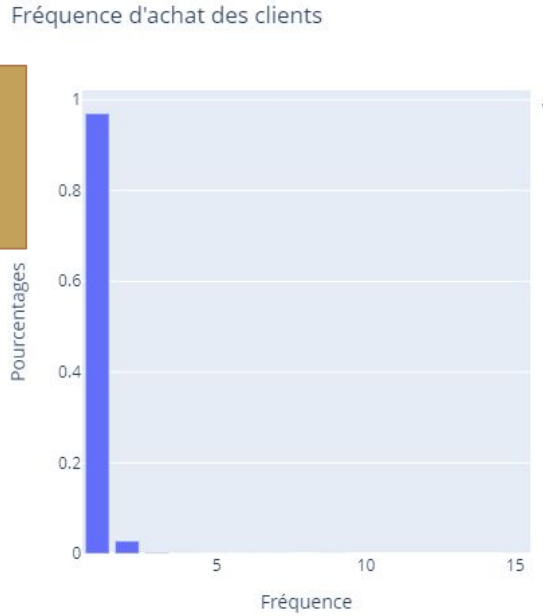
Analyse de la récence et la fréquence





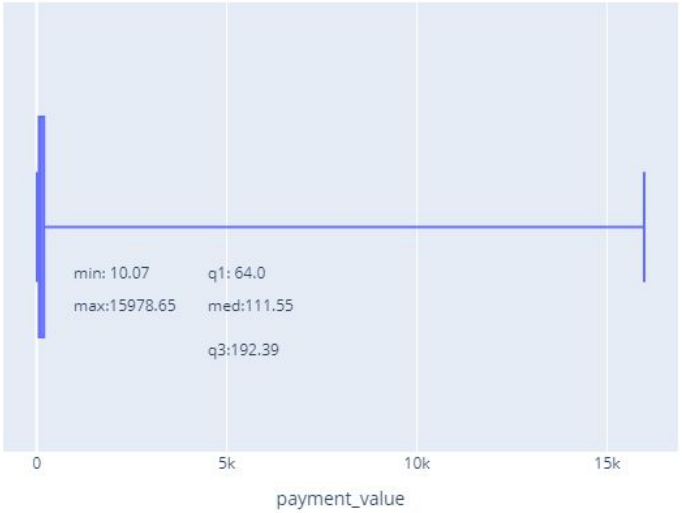
- Constante augmentation
- Cycle avec pic de commandes en novembre-décembre
- Gros creux en septembre

- 3% des clients ont passés commandes plus d'une fois



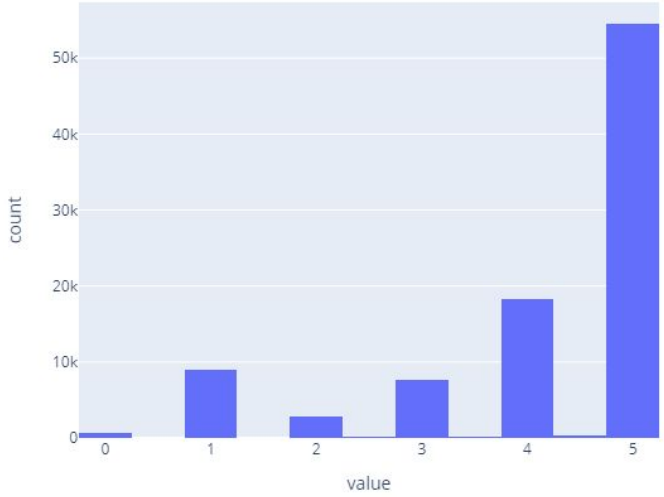
Analyse monétaire et analyse de la satisfaction

Diagramme en boîte des montants dépensés



La médiane des montants totaux dépensés par client est de 111 Reals
Le max atteint 16000 Reals environ

Moyenne des scores donnés



- 58% des clients ont noté 5
- moins de 1% ont noté 0

Les clients sont plutôt satisfaits des services Olist

olist store

Segmentation

Le but est de choisir le bon nombre de variables, le bon algorithme de segmentation et enfin d'interpréter les résultats pour le client

1. Quelle est la démarche de segmentation ?
2. Segmentation RFM :_ Essai Clustering hiérarchique
   _ Essai Dbscan
   _ Essai Kmeans
3. Segmentation à 4 variables

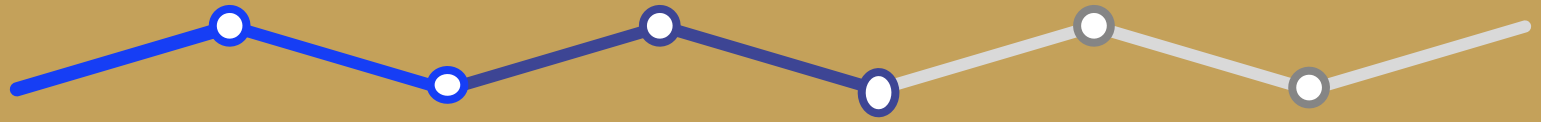
olist store

Processus pour la segmentation

Standardisation des données

Analyse en composantes principales pour réduction de dimension

Application de l'algorithme :
Kmeans
Dbscan
Agglomerative Clustering



Sélection des variables pour la segmentation
Ici on prendra :

- Récence
- Fréquence
- Montant
- Satisfaction

Échantillonnage ou calcul du coefficient de silhouette et méthode du coude pour déterminer le nombre de clusters

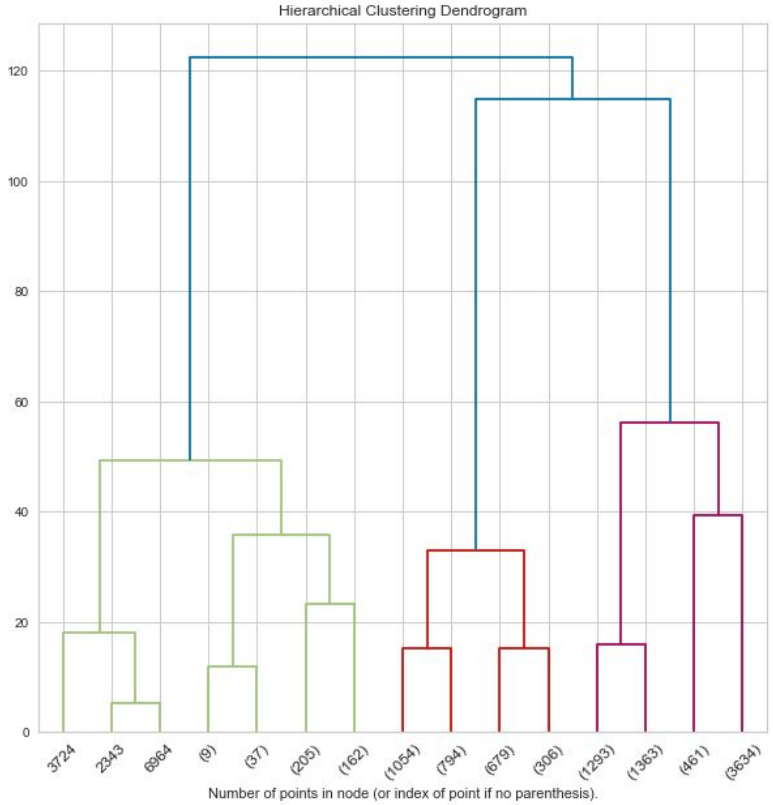
Visualisation et interprétation des clusters

olist store

Clustering hiérarchique sur les variables RFM

- Variable Récence, Fréquence, Monétaire
- Échantillon de 10000 individus

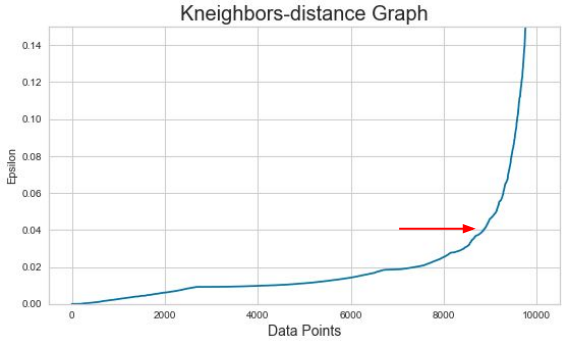
- Model lourd en calcul et qui nécessite un échantillon qui fausse la segmentation.
- Les clusters ne sont pas équilibrés, nous passons de 9 clients à 7000 env.



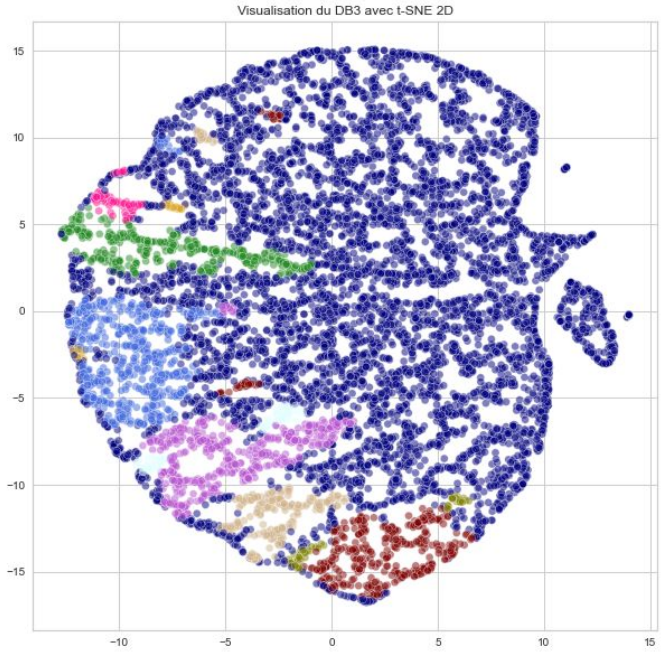
Clustering à densité sur les variables RFM

On travail encore avec un échantillon de 10 000 individus

Détermination du nombre optimal pour epsilon, la distance entre deux points d'un même cluster.



- Paramètres : epsilon = 0.04; min_samples = 20
- On trouve près de 22 clusters
- Sur 10 000 clients 7271 env sont non classés

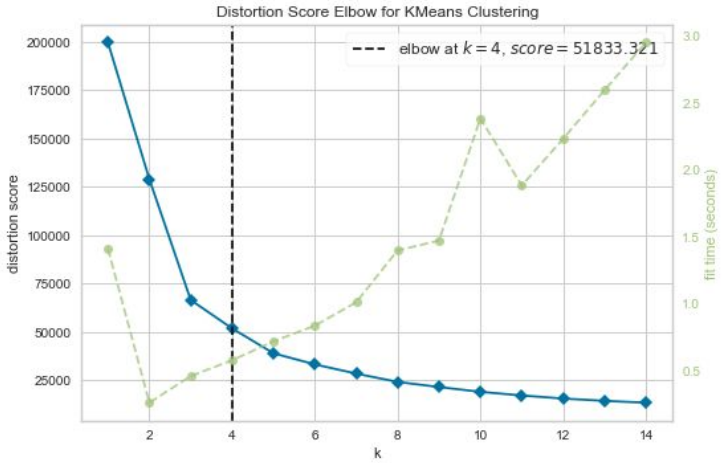


Model non utilisable pour la segmentation. Il n'est pas pertinent

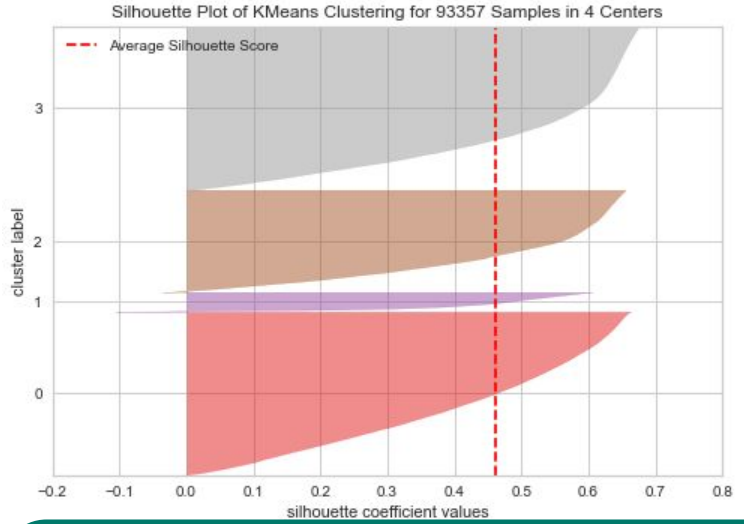
olist store

Clustering non supervisé sur les variables RFM

Détermination du nombre de clusters



Le nombre de clusters à calculer est 4



Nous obtenons 1 cluster très petit et 3 homogènes en termes de taille et densité.

Pour le petit cluster nous atteignons une valeur négative de -0.1, il y a un possible mauvais choix de cluster.

Visualisation et interprétation des clusters du Kmeans

Les composantes principales de l'ACP sont utilisées pour visualiser nos résultats

On obtient 4 clusters plutôt bien séparés.

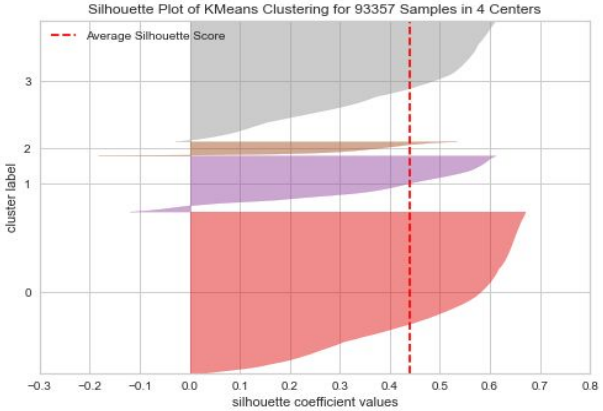
	recency	frequency	payment_value	
	mean	mean	mean	count
Cluster				
0	87.0	1.0	148.0	34542
1	221.0	2.0	809.0	4100
2	255.0	1.0	152.0	33779
3	458.0	1.0	159.0	20936

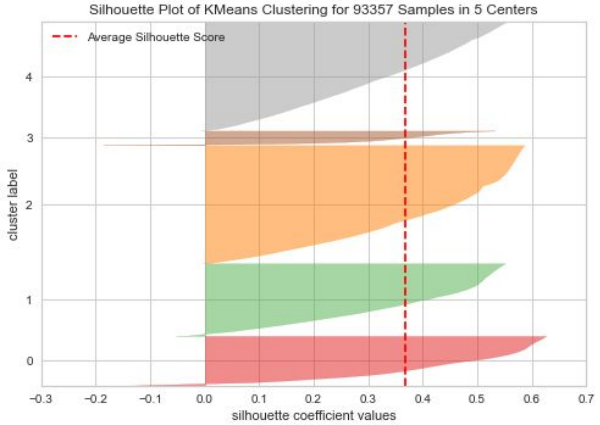
- 0 : Clients récents et unique
- 1 : Clients réguliers (multiples commandes) avec un panier total à 800
- 2 : Clients uniques un peu anciens avec un panier total
- 3 : Clients très anciens(15 mois)

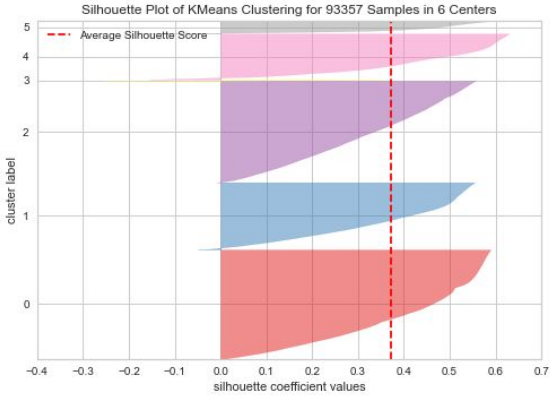
olist store

Segmentation K-means avec 4 variables

On ajoute la satisfaction client au clustering.







Ici 5 clusters sont nécessaires, en considérant les 4 variables en entrée et l'aspect des silhouettes plutôt homogène en longueur.

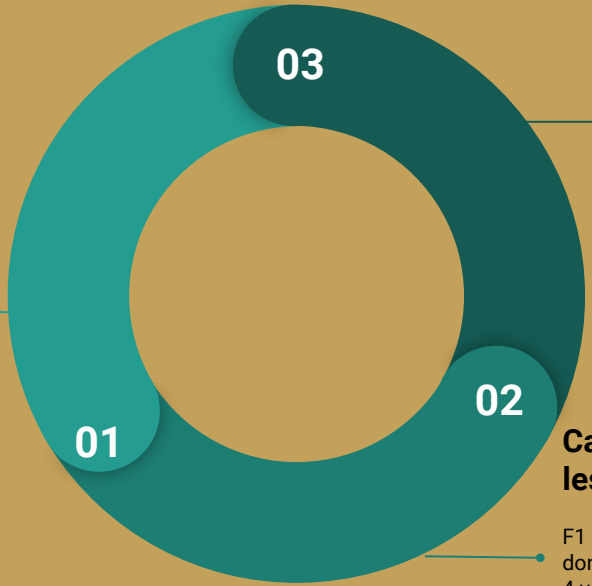
Visualisation et interprétation des clusters

	recency	frequency	monetary		satisfaction	Categorie de clusters
	mean	mean	mean	count	mean	
Cluster						
0	239.0	1.0	1768.0	1320	4.0	Clients satisfaits à relancer rapidement
1	234.0	1.0	181.0	15291	2.0	Clients insatisfaits à récupérer
2	122.0	1.0	144.0	42157	5.0	Nouveaux clients, à fidéliser
3	392.0	1.0	149.0	31824	5.0	Clients perdus
4	219.0	2.0	351.0	2765	4.0	Clients fidèles à récompenser

0: Clients unique satisfaits commencent à être anciens
1: Clients mécontents unique (7 mois)
2: Clients récents, satisfaits
3: Clients presque perdus(+ 13 mois) , satisfaits du service
4: Clients réguliers avec dépense totale élevé et plutôt satisfaits

olist store

Processus pour la détermination de l'intervalle de maintenance du modèle



Modèle 0

Features 0 = F0: 12 premiers mois de données
4 variables : R-F-M-S
ACP
Kmeans

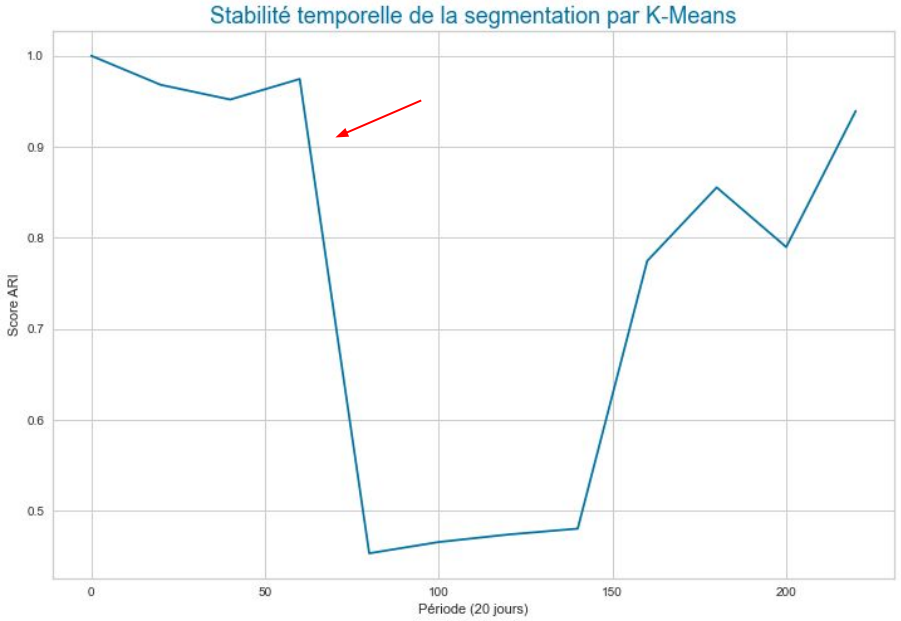
Calcul de segmentation tous les 20 jours

F1 : 12 mois + intervalles de 20 jours de données et ainsi de suite
4 variables : R-F-M-S
ACP du **modèle 0 sur F1**
Kmeans avec ACP **de M0 et M1**

Comparaison : indice de Rand

On compare les résultats de M0 sur F1 avec M1 sur F1
Et ainsi de suite

Quel est le délai de stabilité de la segmentation?



Au bout de 60 jours, nous devrions proposer au client une nouvelle segmentation .

olist store

Conclusion

Proposition à faire au client :

- ★ Segmentation avec le modèle K-means à l'aide d'une ACP basée sur 5 critères
- ★ Actualisation **bimensuelle** du modèle

Pistes d'améliorations :

- Nous pouvons augmenter le nombre de variables, en prenant en compte les catégories achetées.
- calculer le panier moyen par commande afin d'avoir une variables plus représentative par commandes

Proposition marketing :

- Envoyer un bon de réduction sur la prochaine commande aux **nouveaux clients**(10%) et **clients satisfaits à relancer rapidement** (20 %)
- Proposer une réduction sur les catégories achetées le plus souvent aux **clients fidèles**
- Proposer un produit équivalent ou les frais de livraison offerts aux **clients insatisfaits**.

olist store

Merci

Questions - Réponses