SEGMENTATION DES CLIENTS D'UN SITE DE E-COMMERCE

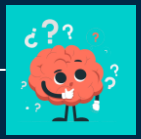
Présenté par Alain KENFACK

20/07/2023

Formation Data scientist : projet 5

olist

Sommaire



Partie : 1

Introduction

Partie : 2

Traitement des données

Partie : 3

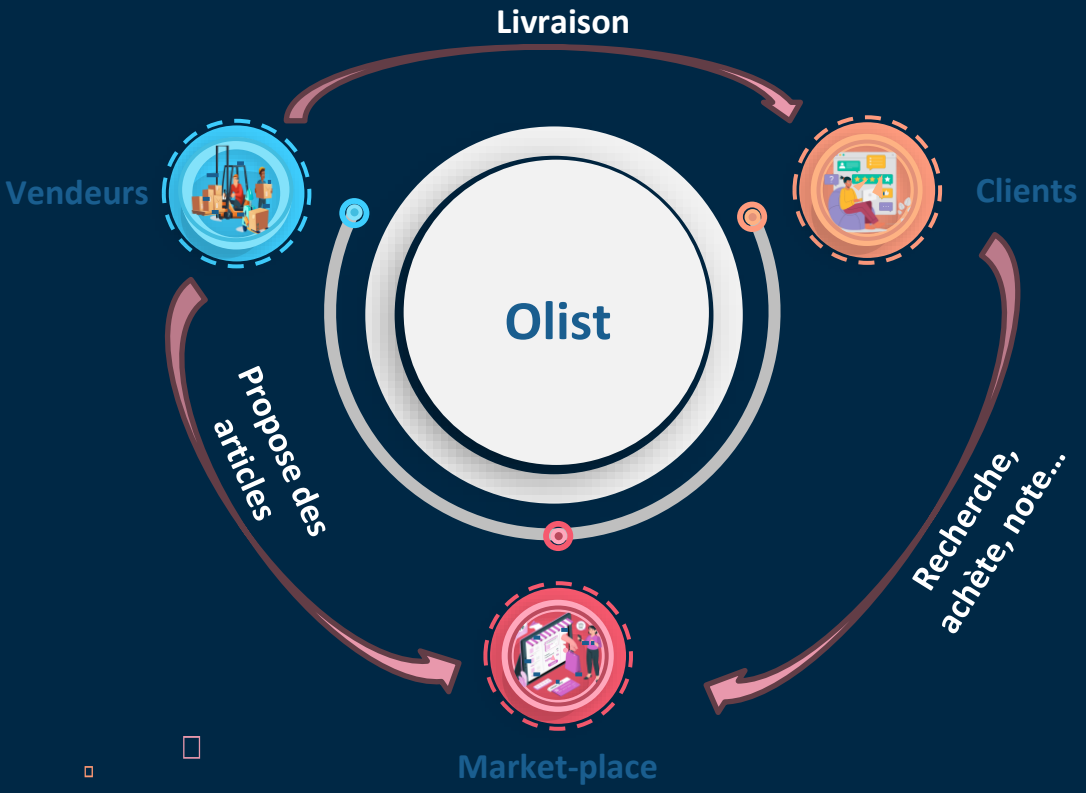
Modélisation

Partie : 4

Conclusion

Partie : 1

Introduction

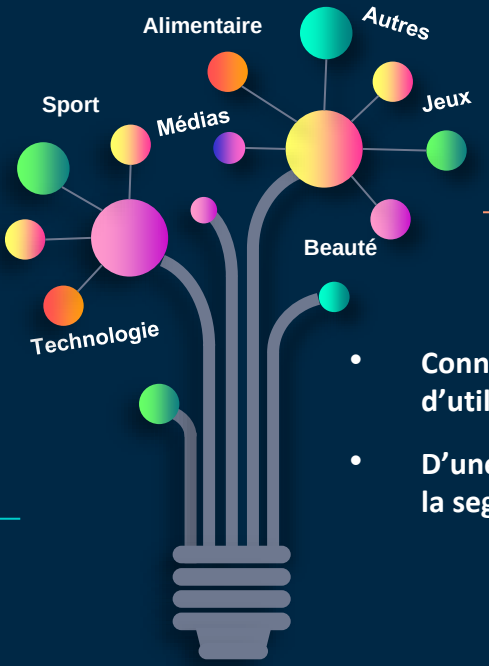


Partie : 1

Problématique

Marketing

Lancer une campagne de communication “ciblée”



- **Connaître les différents profils d’utilisateurs**
- **D’une description actionnable de la segmentation des clients**

Besoin

Partie : 1

Démarche

Traitement des données

- Nettoyage ;
- Analyse ;
- Feature engeneering.

Modélisation

- Segmentation RFM ;
- Segmentations non supervisée ;.

Sélection

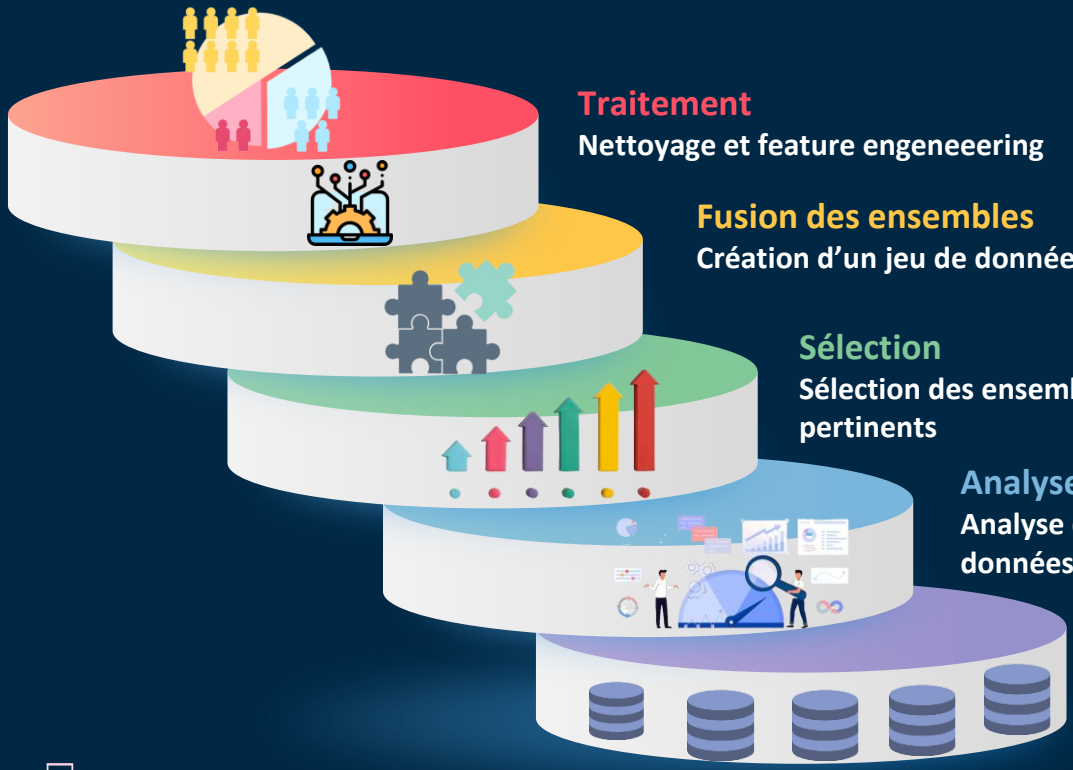
Interprétation des segments et comparaison des modèles ;

Maintenance

Etude de la stabilité au fil du temps du meilleur.

Partie : 2

Traitement des données



Traitement
Nettoyage et feature engeneeering

Fusion des ensembles
Création d'un jeu de données unique

Sélection
Sélection des ensembles et colonnes pertinents

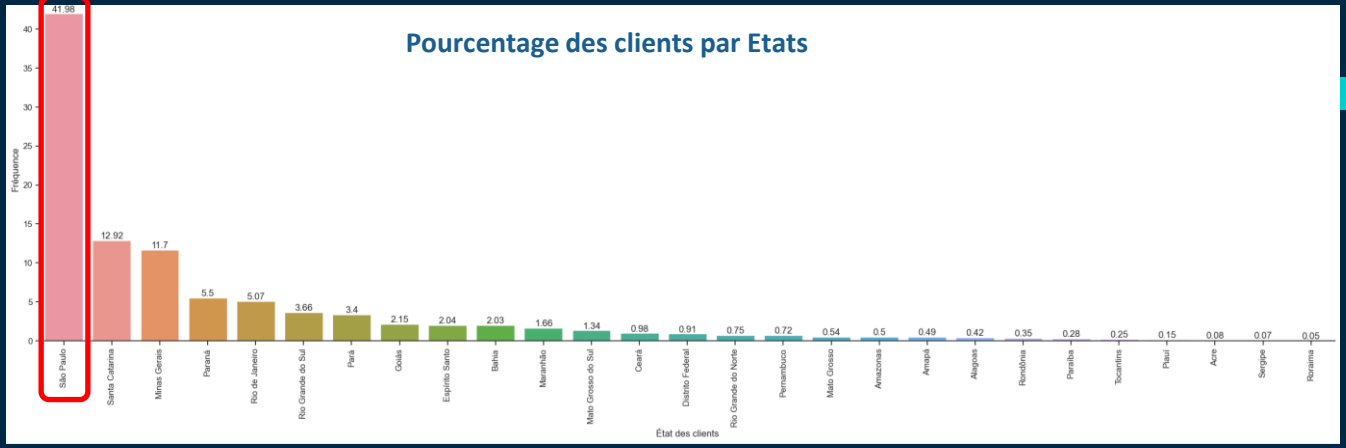
Analyse
Analyse de chaque ensemble de données

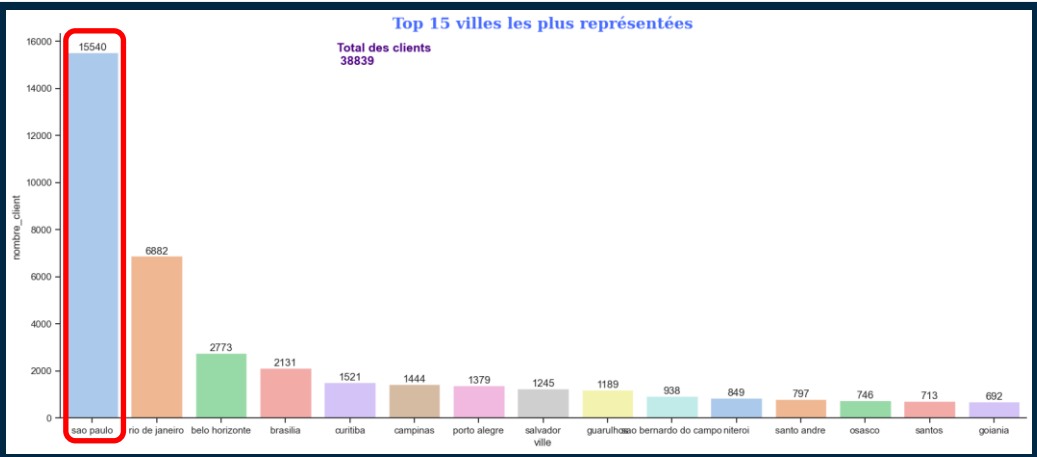
Les données
sont divisées en plusieurs **ensembles.**

Partie : 2

Analyse des données

Géolocalisation des clients

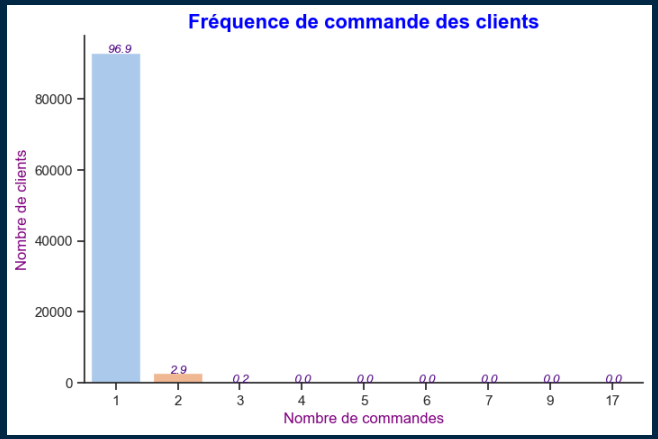


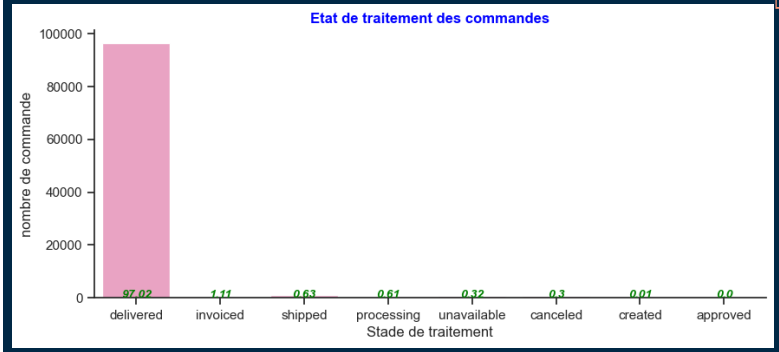


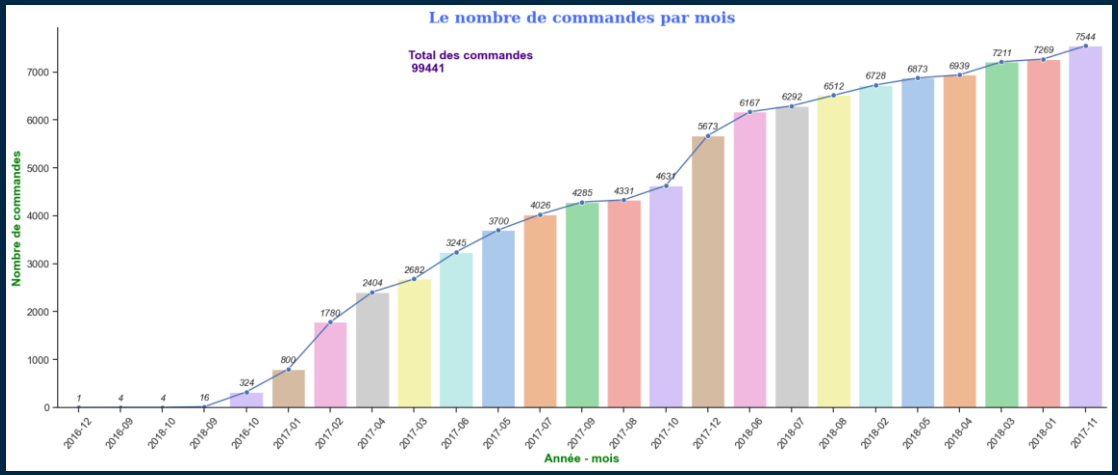
Partie : 2

Analyse des données

Les commandes



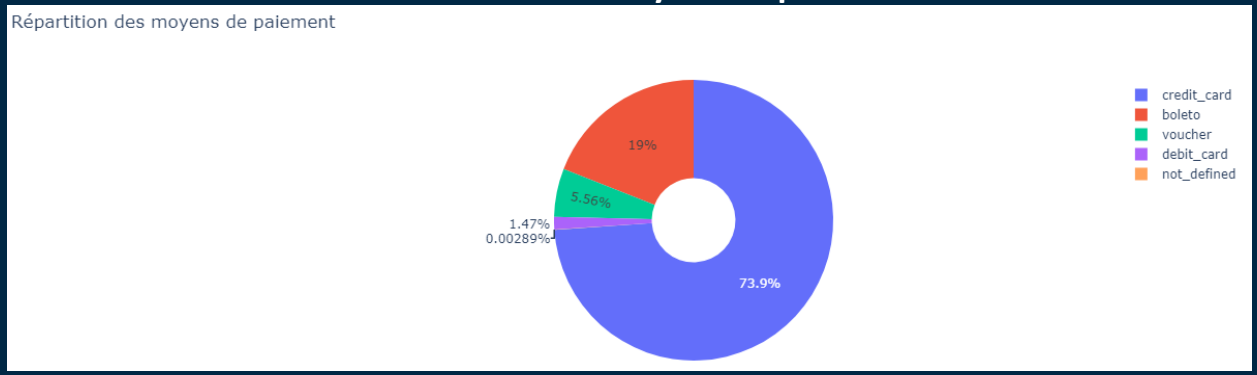


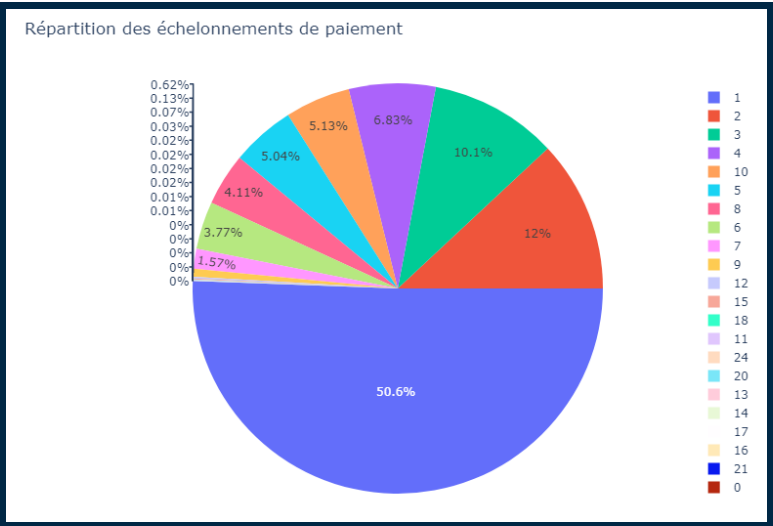


Partie : 2

Analyse des données

Moyens de paiement

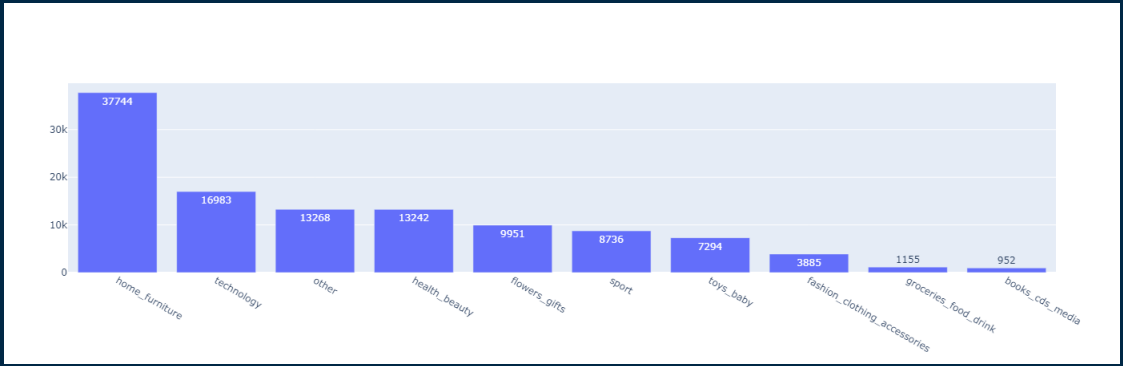


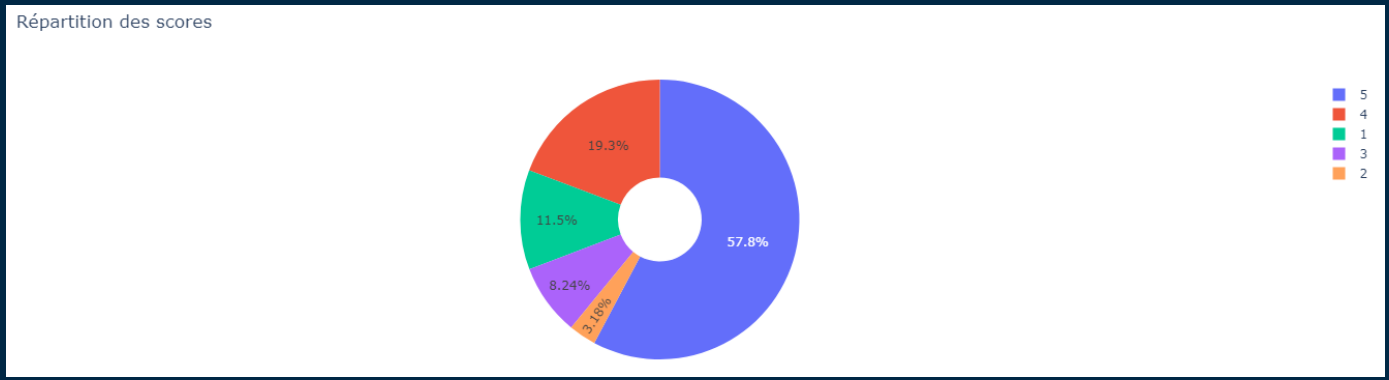


Partie : 2

Analyse des données

Préférences et satisfaction clients





Partie : 2

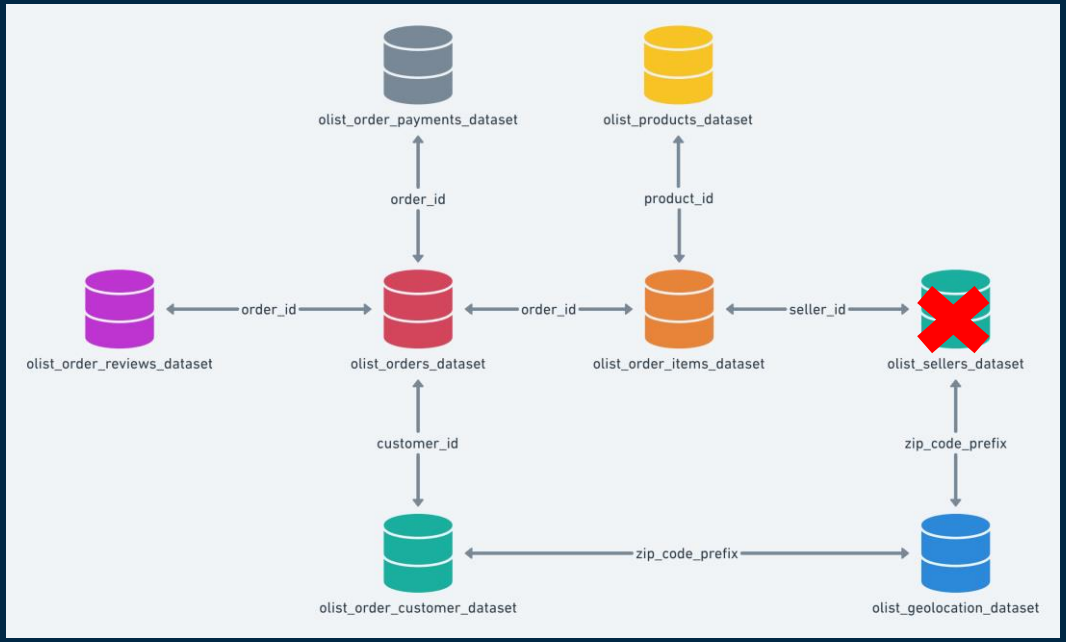
Analyse des données

Nettoyage effectué durant l'analyse

- Noms des Etats : transformation des abréviations en noms complet ;
- Suppression des doublons quand présents ;
- Transformation des variables date du format objet au format « datetime » ;
- Regroupement des données avec la fonction groupby ;
- Suppression des colonnes non pertinentes ;

Partie : 2

Fusion des données



- Certains ensembles de données ont des variables communes comme le décrit la figure ci-dessus.
- J'ai utilisé ces relations pour fusion les données en un seul.
- Le dataset « olist_sellers_dataset » a été écarté, car l'objectif de ce travail demande de se focaliser sur les clients.

Partie : 2

Feature engeneering

Création de nouvelles variables

L'objectif est d'identifier ou créer les variables permettant différencier les clients. 4 groupes de variables ont été créée en fonction des informations disponible dans le jeu de données.

- Fréquence d'achat ;
- Récence ;
- Montant dépensé ;
- Moyen de paiement ;
- Catégorie préféré

Comportementaux

Psychographiques

- Note (satisfaction) ;
- Avis/commentaire

- Localisation ;
- Distance client – Olist ;
- Géolocalisation

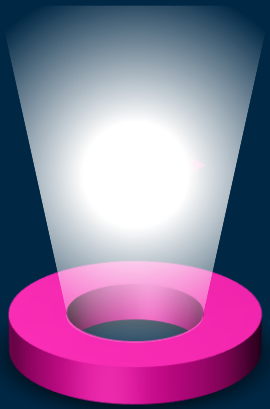
Géographiques

Encodage des variables et création d'un jeu de données avec une ligne par client

Partie : 3

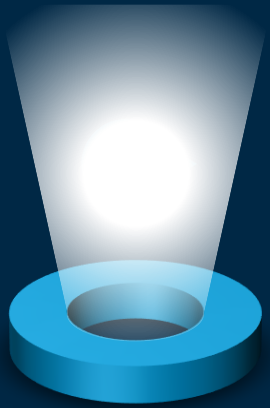
Modélisation

Démarche



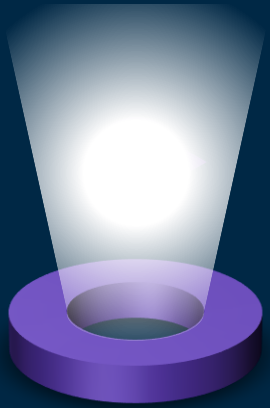
RFM

Segmentation des clients en fonction de 3 variables.



Algorithme non supervisé

K-Means
DBScan
agglomerative



Comparaison

Comparaison des différentes segmentation et sélection du Meilleur modèle.

Maintenance

Déterminer la fréquence nécessaire de mise à jour.

Partie : 3

Segmentation RFM

Procédé :

Discrétisation de ces 3 variables par découpage en quartile

Attribution des groupes en fonction des scores et suivant le modèle proposé par bloomreach

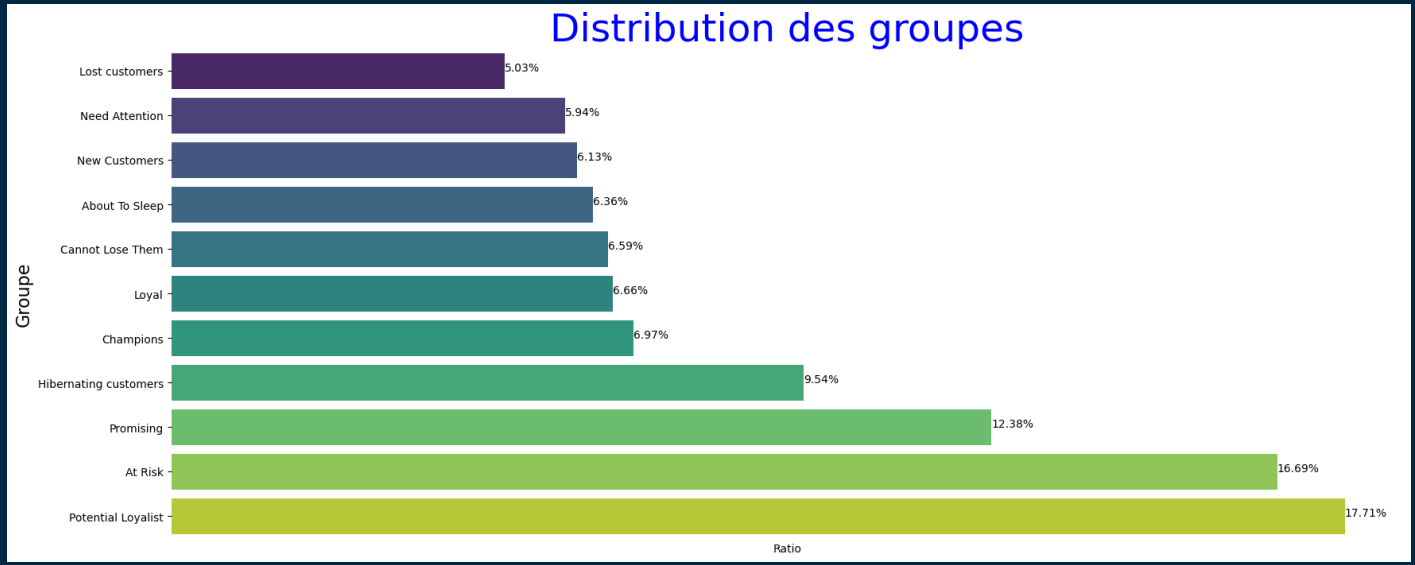
customer_unique_id	recence	frequence	montant	score_recence	score_frequence	score_montant	score	groupe
0000366f3b9a7992bf8c76cfdf3221e2	111	1	141.90	4	1	4	414	Promising
0000b849f77a49e4a4ce2b2a4ca5be3f	114	1	27.19	4	1	1	411	New Customers
0000f46a3911fa3c0805444483337064	536	1	86.22	1	1	2	112	Lost customers
0000f6ccb0745a6a4b88665a16c9f078	320	1	43.62	2	1	1	211	Hibernating customers
0004aac84e0df4da2b147fca70cf8255	287	1	196.89	2	1	4	214	Cannot Lose Them
0004bd2a26a76fe21f786e4fbd80607f	145	1	166.98	4	1	4	414	Promising
00050ab1314c0e55a6ca13cf7181fecf	131	1	35.38	4	1	1	411	New Customers
00053a61a98854899e70ed204dd4bafe	182	2	419.18	3	5	5	355	Loyal
0005e1862207bf6ccc02e4228effd9a0	542	1	150.12	1	1	4	114	Cannot Lose Them
0005ef4cd20d2893f0d9fbd94d3c0d97	169	1	129.76	4	1	3	413	Promising

Combinaison des différents scores pour obtenir le score RFM

Partie : 3

Segmentation RFM

Résultats :



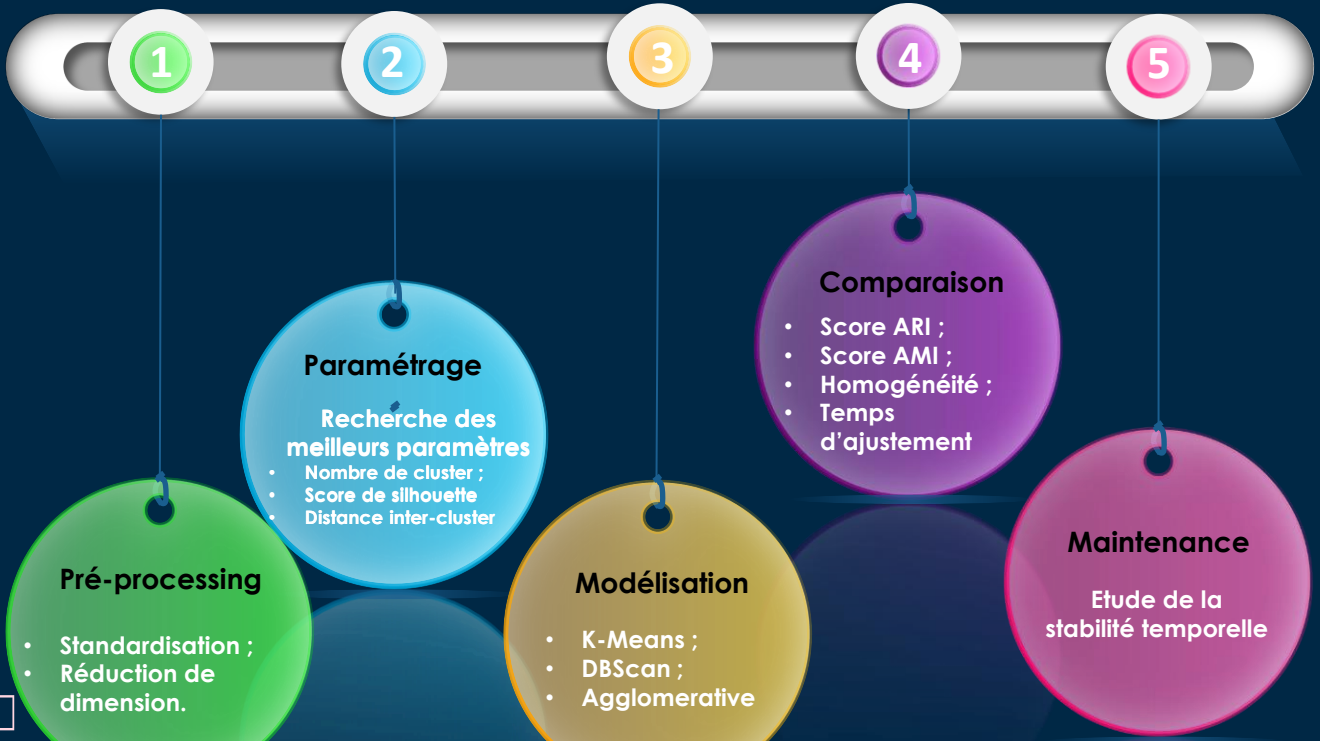
Cette classification est influencée principalement par le montant et la récence d'achat. Plus le montant des achats est élevé et la récence faible, le client est considéré comme loyal ou prometteur.

Mais sachant que la fréquence d'achat est seulement de 1 pour 97% des clients, cette classification est quelque peu biaisée.

Partie : 3

Segmentation non supervisée

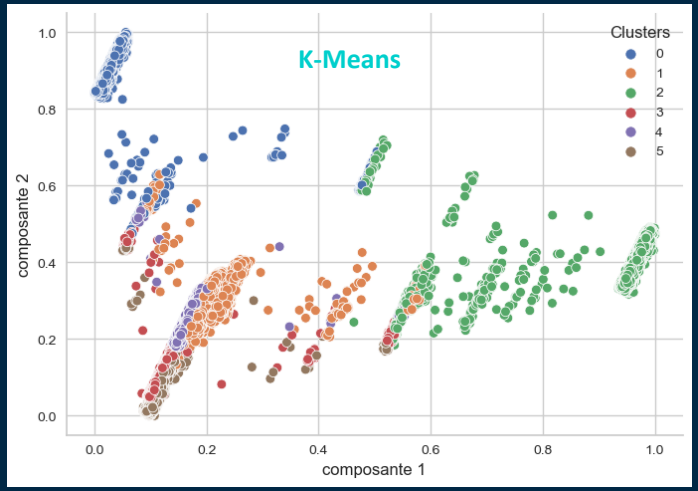
Procédé :

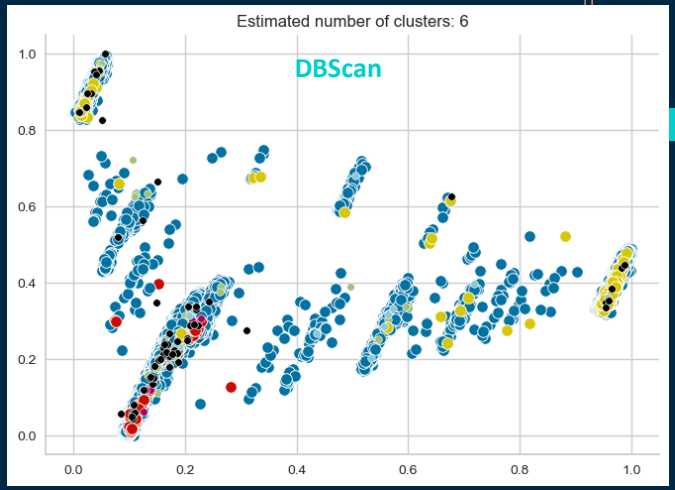


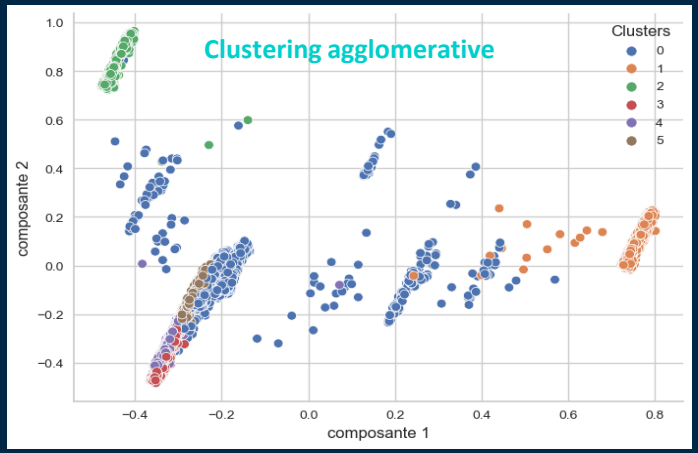
Partie : 3

Segmentation non supervisée

Visualisation des clusters sur le premier plan factoriel





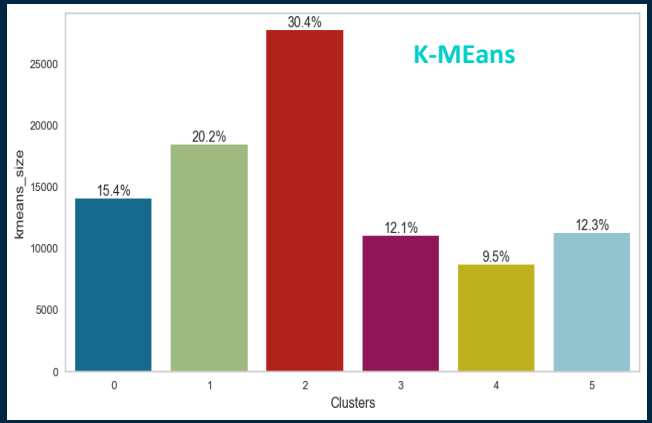


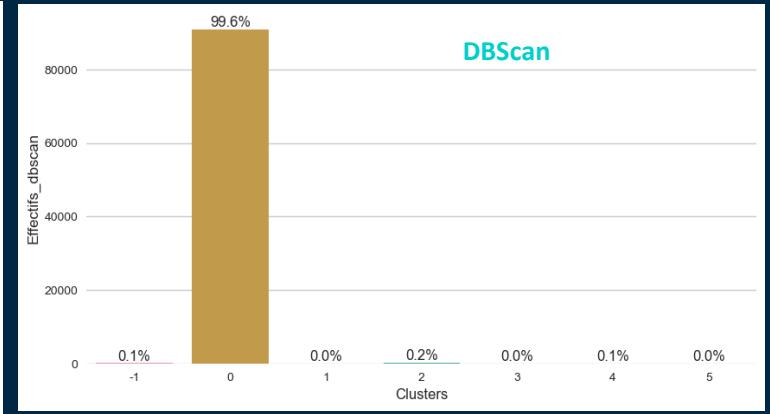
- Les clusters présentent beaucoup de similitude.
- DBScan à tendance à regrouper tous les clients au sein d'un même cluster ;
- K-Means et agglomerative produisent des clusters cohérents.

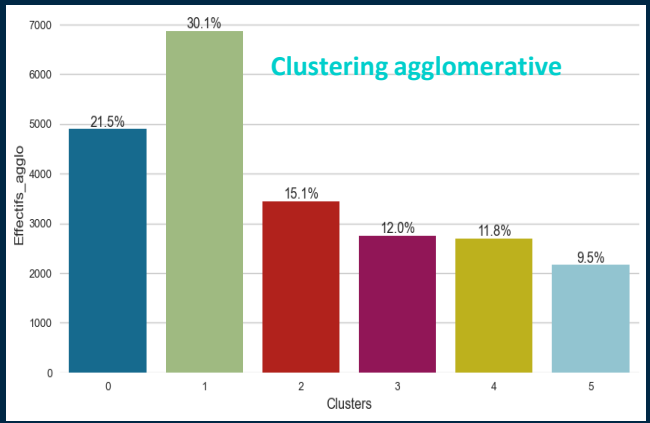
Partie : 3

Segmentation non supervisée

Répartitions des effectifs dans les clusters







Partie : 3

Segmentation non supervisée

Résultats : Métriques

Métriques du clustering agglomératif

	Temps d'ajustement	Homogénéité	ARI	AMI	Inertie approximative
0	41.882491	0.170183	0.0	-2.234586e-11	4.402152

Métriques du clustering avec KMeans

	Iteration	Temps d'ajustement	Inertia	Homogénéité	ARI	AMI
10	Moyenne	0.093655	16333.331766	0.866199	0.801926	0.880782

- **Avec le clustering agglomérative et le K-Means, on obtient des clusters de taille équivalente ;**
- **Les métriques sont à l'avantage du K-Means avec qui on obtient des meilleurs scores ;**
- **Par conséquent, des 3 algorithmes d'apprentissage non supervisé utilisés, le K-Means est celui qui permet le mieux de segmenter les clients de ce jeu de données.**

Partie : 3

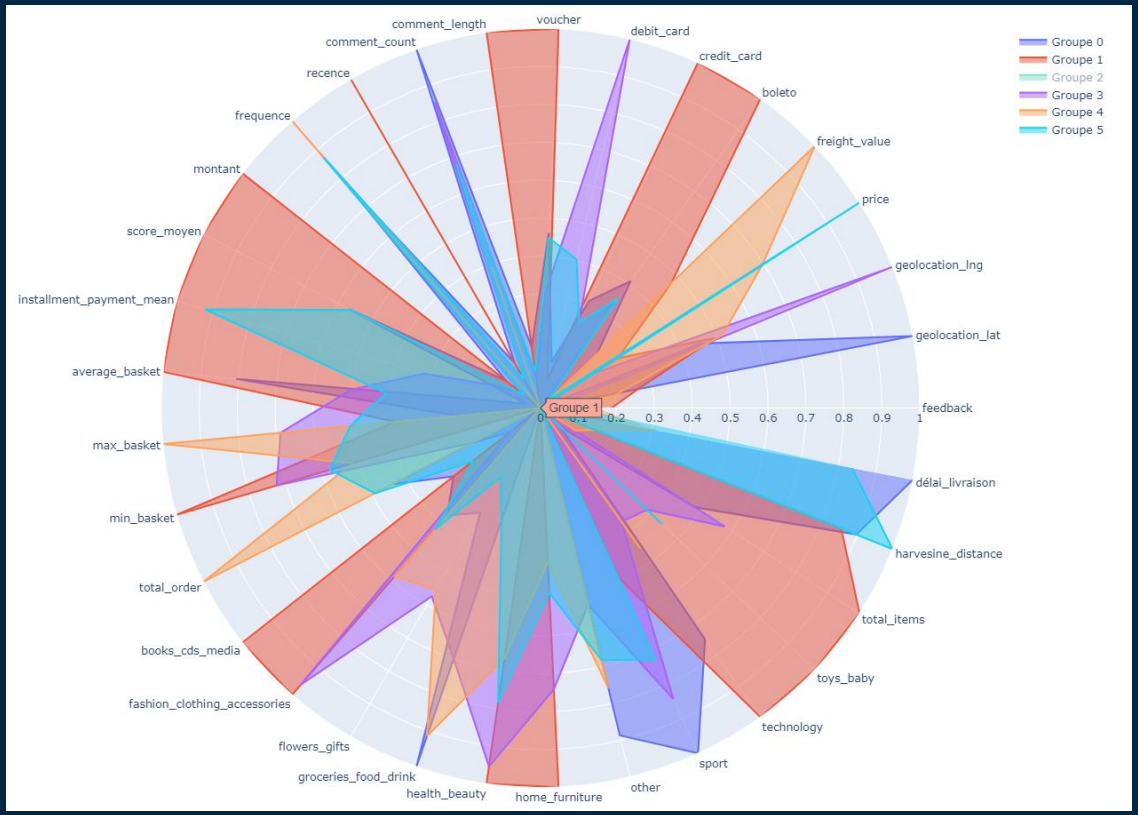
K-Means vs RFM

	Avantages	Limites
RFM	• Simple à mettre en œuvre ; • Rapide ; • Marketing traditionnel.	• Peu flexible ; • Tout refaire pour l'ajout de nouveaux clients ; • Ne tient pas compte d'autres facteurs importants
K-Means	• Plus précise ; • Prends compte plusieurs facteurs ; • Flexible.	Complexe.

Partie : 3

Segmentation non supervisée

Analyse des clusters K-Means



Partie : 3

Interprétation des clusters K-Means

Cluster 0

- Clients peu satisfaits ;
- Résident loin de Olist ;
- Catégorie : technologie, sport et flowers ;
- Ancienneté relative, dépensent peu

30,4 % des clients

Cluster 1

- Nouveaux clients ;
- Attentifs à leurs santés et bien-être ;
- Montant des dépenses faible;
- Ne paient pas les frais de livraison ;
- Résident loin de Olist.

12,3 % des clients

Cluster 2

- Nouveaux clients à portefeuille important ;
- Fréquence d'achat moyen supérieur à 1 ;
- Catégorie : santé-beauté et « home-furniture » ;
- Réside proche de Olist , mais paient cher les frais de livraison.

9,6 % des clients

Cluster 3

- Nouveau Client, fréquence d'achat supérieur à 1 ;
- Résident loin de Olist, mais délai de livraison relativement court ;
- Sont intéressés par la mode et la technologie ;
- Paient en plusieurs fois et dépensent peu

7,9 % des clients

Cluster 4

- Clients anciens avec un fort pouvoir d'achat ;
- Dépensent sur les jouets, le sport et la beauté ;
- Résident près de Olist et les coûts de livraison sont faibles ;
- Clients entièrement satisfaits.

24,4 % des clients

Cluster 5

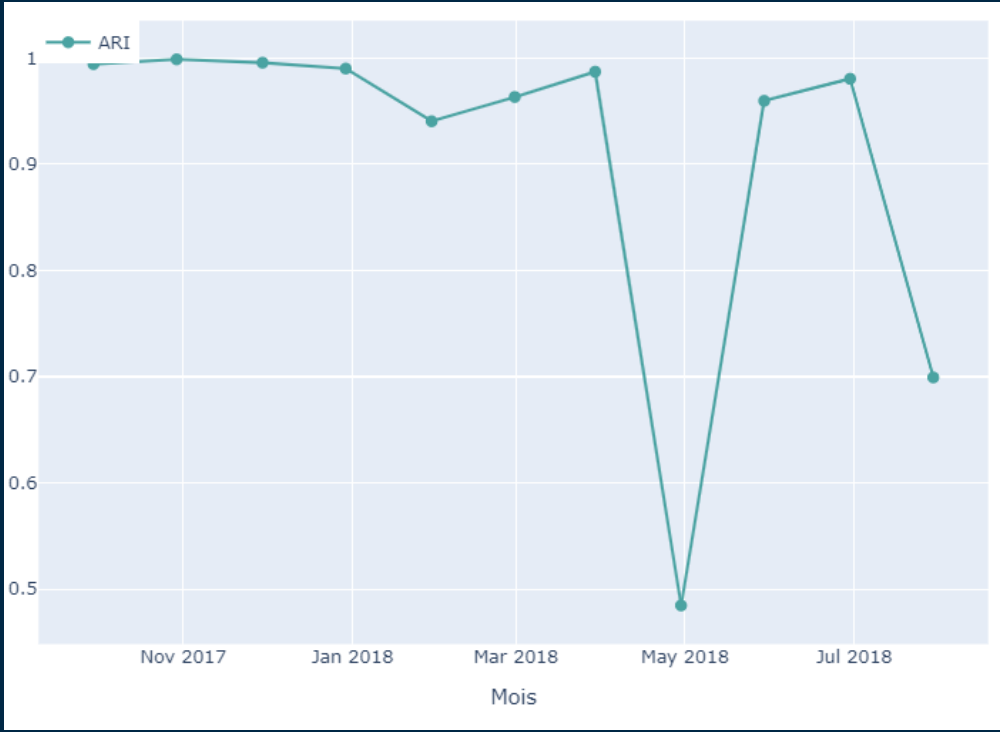
- Clients relativement anciens;
- Accro de sport et technologie et dépensent aussi dans l'alimentaire ;
- Faible pouvoir d'achat et niveau de satisfaction moyen.

15,4 % des clients

Partie : 3

Segmentation RFM

Stabilité temporelle du K-MEans



- **Les 7 premiers mois, le score ARI varie très peu et reste supérieur à 0,94.**
- **Au-delà de 7 mois, le score chute fortement pour passer en dessous de 0,5. Une maintenance est à prévoir tous les 7 mois.**

Partie : 3

Conclusion

- Pour ce projet, nous avons apporté une solution au problème marketing de Olist en procédant en la segmentation de leur clientèle sur la base des données fournit ;
- La segmentation a été faite via l'algorithme d'apprentissage non supervisé K-Means ;
- Grâce à l'algorithme, nous avons obtenu 6 clusters avec un grand nombre de facteur y compris les facteurs classique de segmentation RFM ;
- Toutefois, un programme de maintenance est à prévoir chaque semestre, car la stabilité de la segmentation diminue après 6 mois.

Merci pour votre attention