

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

Jules Prod'homme

Data Scientist



Objectifs

Olist, une plateforme de e-commerce brésilienne, offre des solutions de vente sur des marketplaces en ligne.

Pour améliorer ses campagnes de communication, l'équipe marketing d'Olist cherche à mieux comprendre les différents types d'utilisateurs en se basant sur leur comportement et leurs données personnelles.

Nous devons effectuer une segmentation des clients en regroupant ceux ayant des profils similaires et fournir une description exploitable pour les équipes marketing.

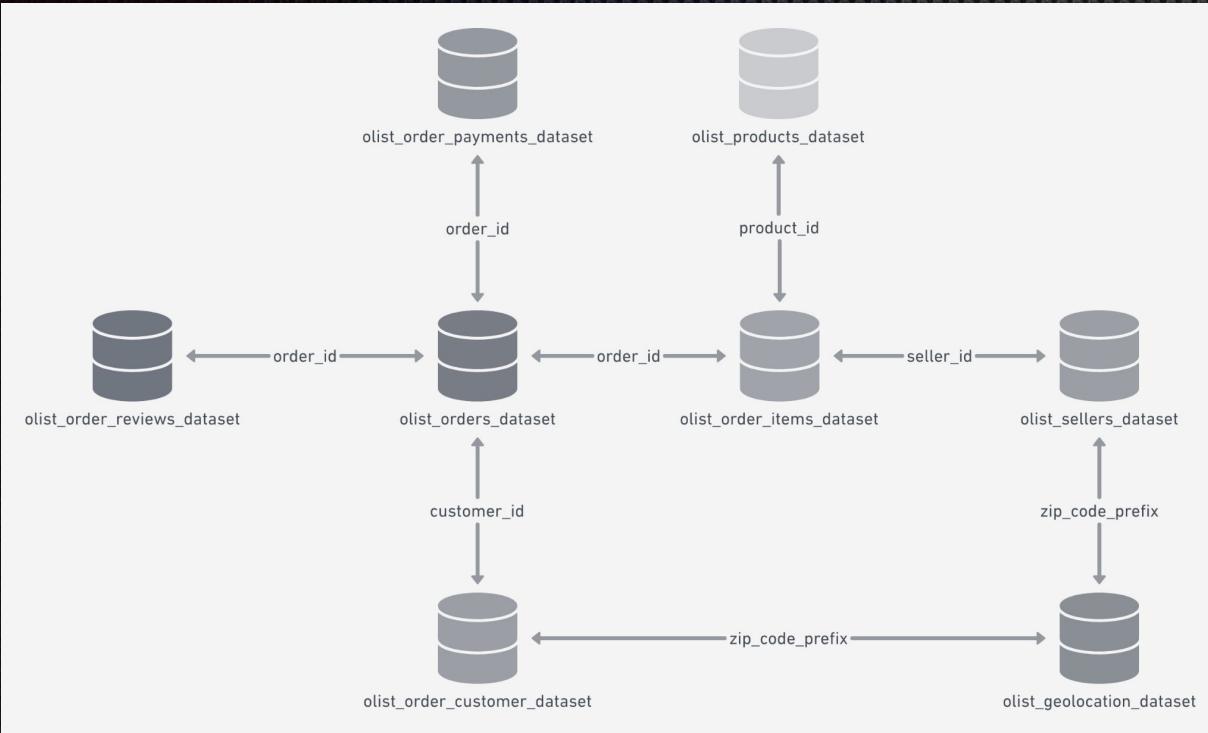
Recommander la fréquence de mise à jour de cette segmentation afin qu'elle reste pertinente (contrat de maintenance).



olist



Base de donnée utilisée



COMMANDES :

- 97% des commandes au statut "livré".

- Historique de octobre

2016 à sept 2018.

PAIEMENTS

- Montant des paiements de 0 à

13.600 reals, en moyenne 154 reals.

- Historique =: oct

2016→sept 2018.

REVIEWS

- 58 % des reviews ont une note

de 5/5, 19% des reviews ont

une note de 4/5. Près de 15%

des reviews avec 1 ou 2.



Feature engineering

1. Segmentation RFM

La segmentation RFM, une méthode courante en segmentation comportementale, évalue la valeur des clients en les regroupant en segments homogènes. Elle se base sur trois critères :

- Récence : le nombre de jours écoulés depuis le dernier achat.
- Fréquence : le nombre d'achats effectués par le client sur une période donnée.
- Montant : le total des dépenses du client sur cette période.

2. Intégration du Score reviews

Les scores de revue sont agrégés en calculant la moyenne des évaluations attribuées par un client à ses différentes commandes.

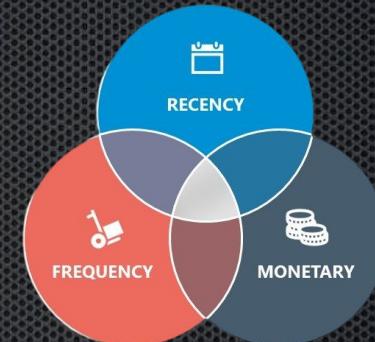
3. Intégration des Régions des clients

Après l'intégration du score de revue, les régions des clients seront également intégrées pour enrichir l'analyse. On utilise le code postal pour déterminer si les clients proviennent de zones urbaines ou rurales

Mise en place :

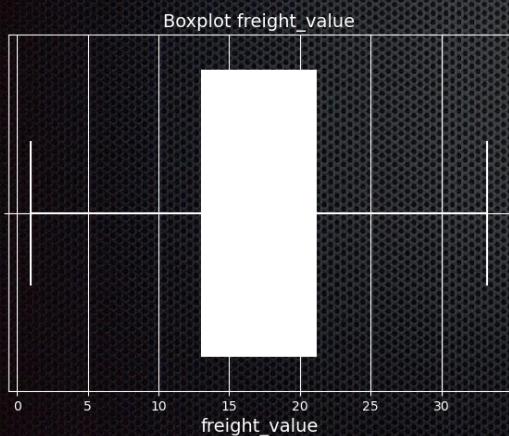
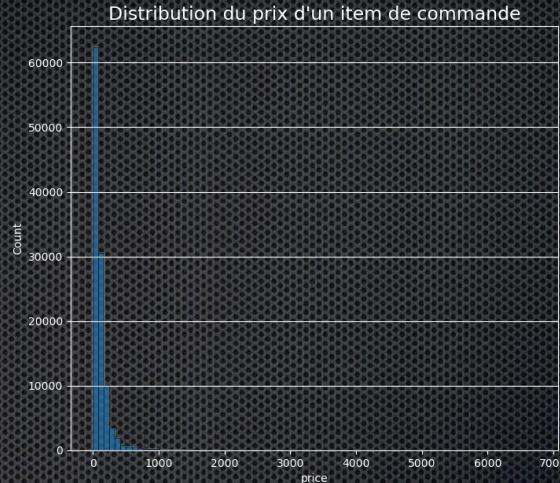
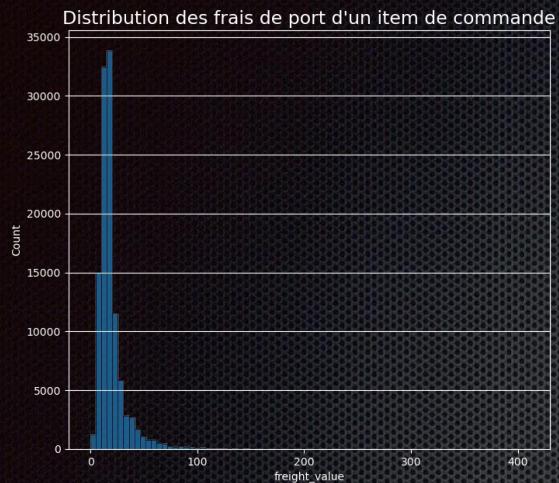
- L'analyse couvre toute la période disponible, de septembre 2016 à octobre 2018.
- Pour le calcul de la récence, la date est située au lendemain de la dernière commande effectuée.
- Pour la provenance des clients on se base sur la nomenclature des codes postaux brésiliens

Un fichier contenant les données de 92 593 clients uniques, avec les caractéristiques RFM, le Score de Revue intégré et l'information sur les régions des clients.





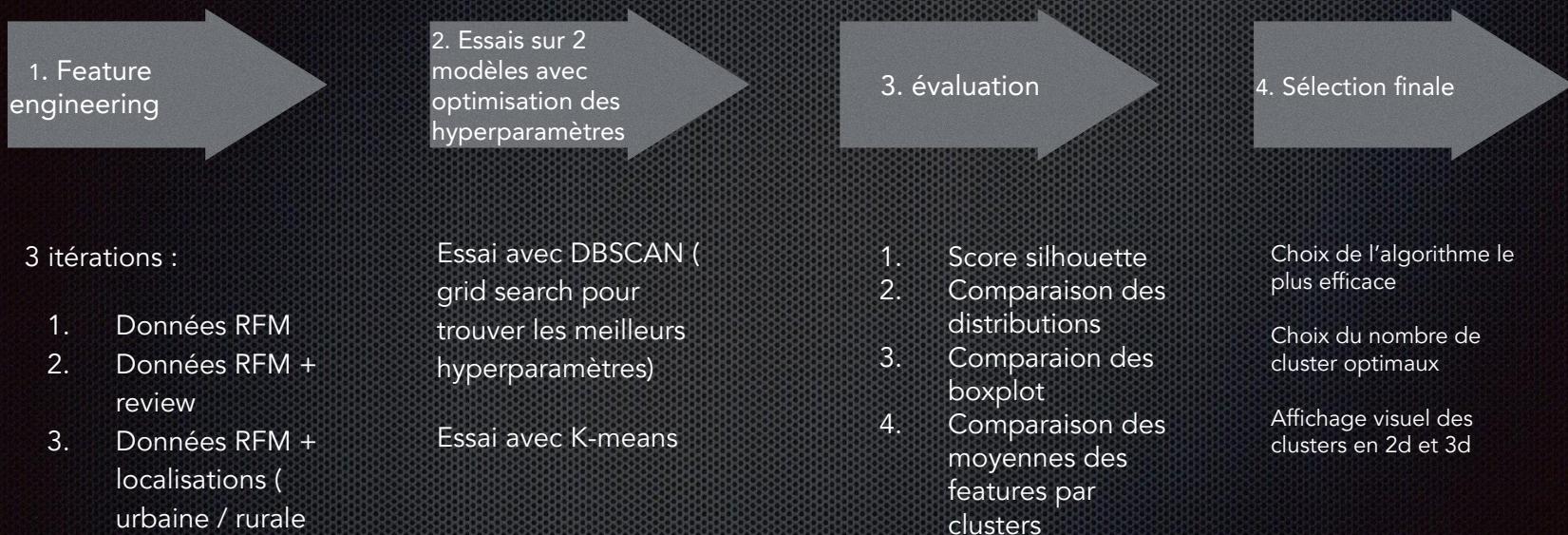
Distribution des commandes



Approche des modélisations



Itérations entre feature engineering et entraînement des modèles.



Algorithme K-means



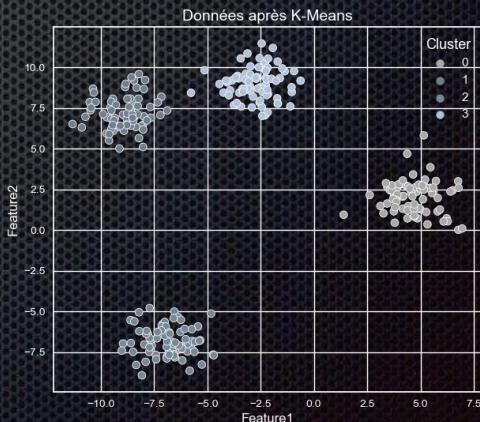
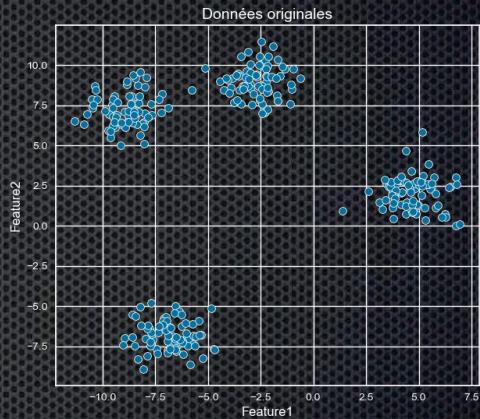
K-Means est un algorithme de machine learning non supervisé, souvent utilisé pour le clustering.

Objectif : Diviser les données en (k) groupes tout en minimisant l'inertie, c'est-à-dire la somme des carrés des distances au sein de chaque cluster.

Initialisation des centroïdes : Utilisation de la méthode k-means++ pour choisir des points initialement éloignés les uns des autres, plutôt qu'une initialisation totalement aléatoire.

Il est présent de nombreux avantages :

- Simplicité : Facile à comprendre et à mettre en œuvre.
- Convergence : L'algorithme converge toujours et de manière rapide.
- Clusters convexes : Efficace pour identifier des clusters de formes convexes.
- Grande échelle : Convient bien aux grands ensembles de données.



Algorithme DBSCAN



DBSCAN, est un algorithme de clustering par densité largement utilisé.

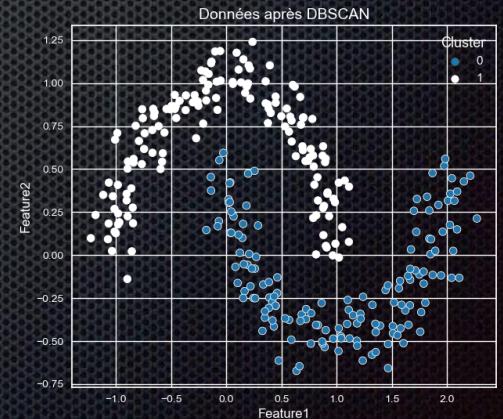
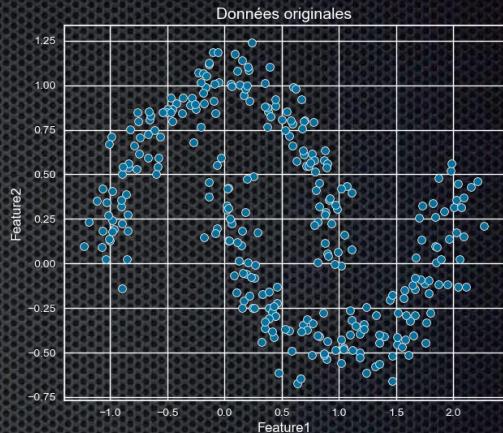
Contrairement à d'autres méthodes de clustering, DBSCAN ne nécessite pas de spécifier le nombre de clusters à l'avance, ce qui le rend très flexible.

Hyperparamètres :

- **Min_samples** : Ce paramètre spécifie le nombre minimal de points dans le voisinage d'un point donné pour qu'il soit considéré comme faisant partie d'une région dense.
- **Epsilon** : C'est la distance maximale entre deux points pour qu'ils soient considérés comme faisant partie du même cluster.

Avantages :

- Déterministe : Contrairement à certains algorithmes de clustering, DBSCAN produit toujours le même résultat pour un ensemble de données donné.
- Flexibilité des clusters : Les clusters produits par DBSCAN ne sont pas restreints à des formes géométriques spécifiques. Cela signifie que les clusters peuvent être de formes arbitraires, ce qui est particulièrement utile pour les ensembles de données contenant des clusters de formes complexes ou non convexe.
- DBSCAN est particulièrement utile pour détecter des clusters dans des ensembles de données où les clusters ont des formes et des densités variées. Son approche basée sur la densité lui permet de gérer efficacement les données avec du bruit et de détecter des clusters de formes non conventionnelles.



Présentation des résultats



De tous les essais c'est le clustering par K-means qui a donné les meilleurs résultats avec 5 clusters et les features RFM + score review.

Nous allons présenter les différentes étapes du clustering ainsi que les méthodes d'évaluation :

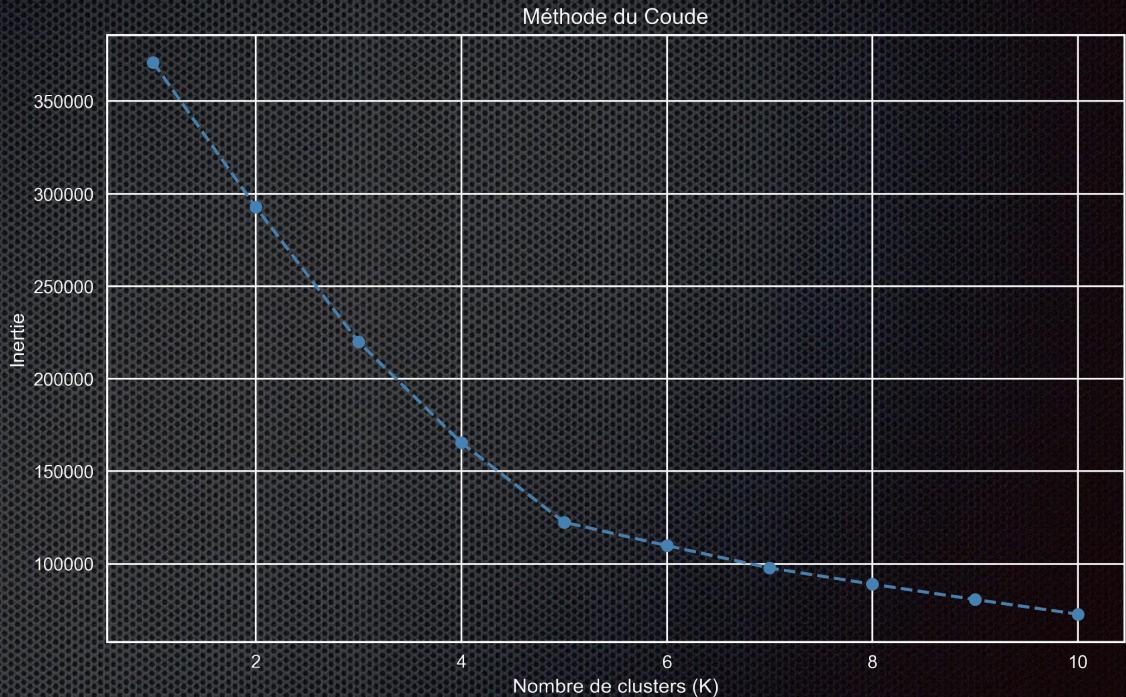
- Choix du nombre de clusters via méthode du coude
- Affichage du score de silhouette et de leurs clusters
- Représentations en 2d et 3d par ACP
- Etudes des histogrammes par clusters
- Analyse comparatives de moyennes des features par clusters

Choix du nombre de clusters par méthode du coude



La méthode du coude est utilisée pour déterminer le nombre optimal de clusters (K) dans le clustering K-means. Voici comment elle fonctionne :

1. Calcul de l'inertie : On exécute K-means pour différentes valeurs de K et on calcule l'inertie (somme des distances au carré entre chaque point et son centre de cluster).
2. Tracé de la courbe : On trace l'inertie en fonction de K . L'inertie diminue avec l'augmentation de K .
3. Identification du coude : Le "coude" de la courbe est le point où l'inertie ne diminue plus significativement avec l'augmentation de K . Ce point suggère le nombre optimal de clusters.

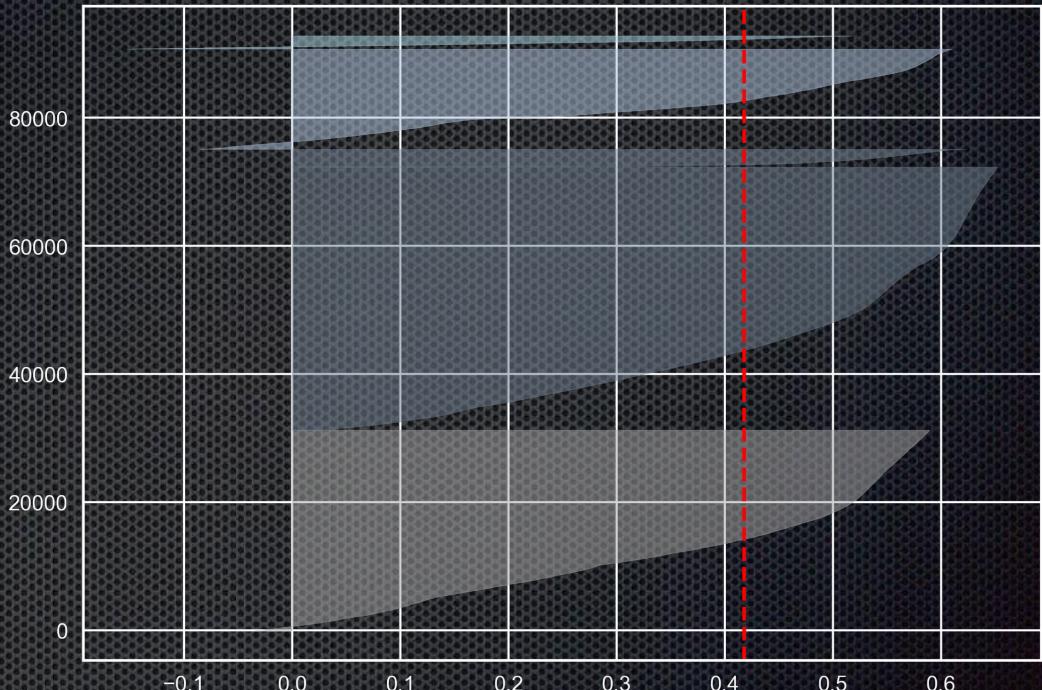


Visualisation score silhouette



Yellowbrick fournit des visualisations interactives qui permettent une compréhension plus intuitive des données et des résultats du clustering.

On peut voir l'épaisseur des clusters ainsi que leur score silhouette moyen en ordonnée (ligne pointillée rouge) ce qui nous permet de constater rapidement de la qualité des clusters



Boxplots pour chaque features par cluster

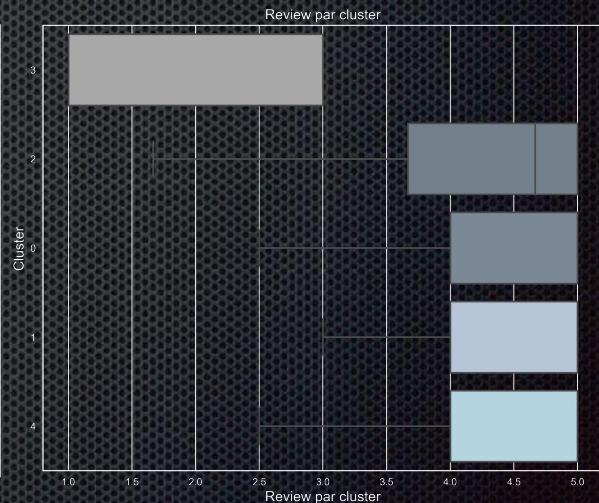
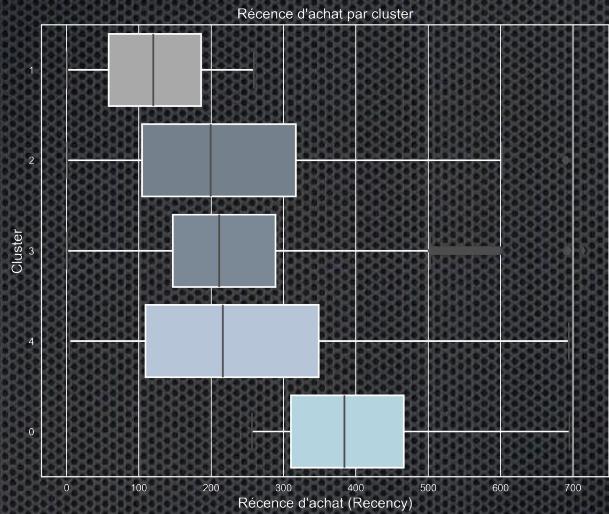
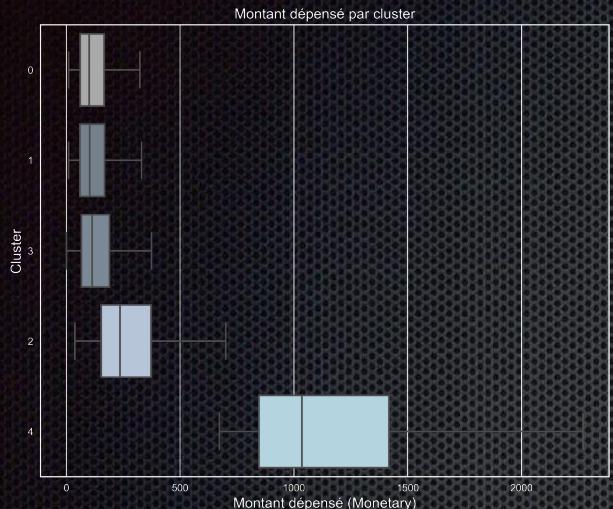


Tableau des moyennes par clusters



Les moyennes dépassant le seuil d'erreur de 5% (qui ne sont pas statistiquement suffisamment différentes) ont été surlignées en bleu.

Cluster	Récence	Fréquence	Montant	Review
0	394	1	134	5
1	123	1	133	5
2	220	2	306	4
3	231	1	151	5
4	235	1	125	4

Analyse des moyennes



Pour chaque caractéristique, les moyennes entre les clusters ont été comparées à l'aide du test de Student pour évaluer la significativité des différences observées.

Les p-valeurs ont été calculées pour chaque comparaison afin de déterminer s'il existait une différence significative entre les valeurs moyennes des clusters.

À l'exception d'une seule caractéristique, pour laquelle les p-valeurs étaient supérieures au seuil de 5% d'erreur, toutes les autres p-valeurs étaient inférieures à ce seuil. En conséquence, il a été conclu que le clustering était viable et que les clusters étaient statistiquement significatifs dans leur différenciation des données

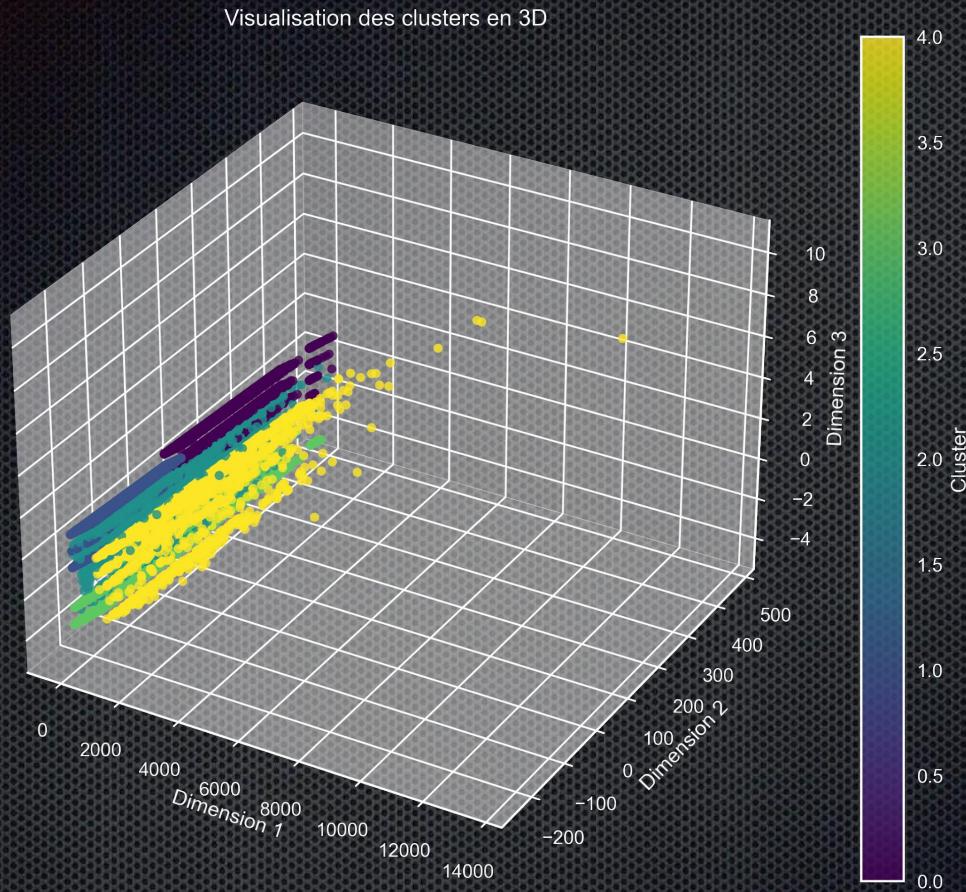
Comparaison de moyennes sur le feature montant

	Cluster1	Cluster2	Mean1	Mean2	T-score_monetary	p-value
0	1.0	0.0	133.0	134.0	-1.0	0.15
1	1.0	3.0	133.0	151.0	-17.0	0.00
2	1.0	2.0	133.0	306.0	-73.0	0.00
3	1.0	4.0	133.0	1250.0	-267.0	0.00
4	0.0	3.0	134.0	151.0	-15.0	0.00
5	0.0	2.0	134.0	306.0	-68.0	0.00
6	0.0	4.0	134.0	1250.0	-240.0	0.00
7	3.0	2.0	151.0	306.0	-50.0	0.00
8	3.0	4.0	151.0	1250.0	-177.0	0.00
9	2.0	4.0	306.0	1250.0	-66.0	0.00

Comparaison de moyennes sur le feature récence

	Cluster1	Cluster2	Mean1	Mean2	T-score Recency	p-value
0	1.0	0.0	123.0	394.0	-432.0	0.00
1	1.0	3.0	123.0	231.0	-126.0	0.00
2	1.0	2.0	123.0	220.0	-62.0	0.00
3	1.0	4.0	123.0	235.0	-63.0	0.00
4	0.0	3.0	394.0	231.0	156.0	0.00
5	0.0	2.0	394.0	220.0	87.0	0.00
6	0.0	4.0	394.0	235.0	69.0	0.00
7	3.0	2.0	231.0	220.0	4.0	0.00
8	3.0	4.0	231.0	235.0	-2.0	0.13
9	2.0	4.0	220.0	235.0	-4.0	0.00

Visualisation des cluster en 3d par ACP



Interprétation des résultats



Cluster 0 :

- Récence : Environ 394 jours depuis le dernier achat.
- Fréquence : En moyenne, chaque client a effectué un seul achat.
- Monétaire : La somme totale dépensée par les clients dans ce cluster est d'environ 134 reals.
- Score de revue : Les clients ont laissé un score de revue parfait de 5.

Interprétation : "Récents ou occasionnels avec une faible dépense, mais très satisfaits"

Cluster 1 :

Récence : Environ 123 jours depuis le dernier achat.

- Fréquence : En moyenne, chaque client a effectué un seul achat.
- Monétaire : La somme totale dépensée par les clients dans ce cluster est d'environ 133 reals.
- Score de revue : Les clients ont laissé un score de revue parfait de 5.

Interprétation : "Récents avec une faible dépense, mais très satisfaits"

Cluster 2 :

- Récence : Environ 220 jours depuis le dernier achat.
- Fréquence : En moyenne, chaque client a effectué deux achats.
- Monétaire : La somme totale dépensée par les clients dans ce cluster est d'environ 306 reals.
- Score de revue : Les clients ont laissé un score de revue de 4.

Interprétation : "Actifs et dépensiers, globalement satisfaits"

Cluster 3 :

- Récence : Environ 231 jours depuis le dernier achat.
- Fréquence : En moyenne, chaque client a effectué un seul achat.
- Monétaire : La somme totale dépensée par les clients dans ce cluster est d'environ 151 reals.
- Score de revue : Les clients ont laissé un score de revue de 2.

Interprétation : "Inactifs ou occasionnels avec une dépense modérée, mais insatisfaits"

Cluster 4 :

- Récence : Environ 235 jours depuis le dernier achat.
- Fréquence : En moyenne, chaque client a effectué un seul achat.
- Monétaire : La somme totale dépensée par les clients dans ce cluster est d'environ 1250 reals.
- Score de revue : Les clients ont laissé un score de revue de 4.

Interprétation : "Inactifs ou occasionnels, mais avec une dépense significative et globalement satisfaits"

Contrat de maintenance



Objectif : Évaluer la stabilité des segments au fil du temps et définir la fréquence optimale de mise à jour du modèle de segmentation pour établir un devis de contrat de maintenance.

Méthodologie :

- L'historique des commandes du jeu de données fourni par Olist couvre environ 24 mois. Nous avons instancié un algorithme k-means similaire à celui utilisé pour le clustering des clients sur une période initiale de 12 mois, puis avons itéré sur les 12 mois suivants, par intervalles d'un mois, 2 mois, et trimestre

- Notre objectif était de déterminer à quel moment la prédiction obtenue avec le modèle de clustering initial devenait obsolète, indiquant ainsi la nécessité d'entraîner un nouveau modèle de clustering.

- Pour comparer deux clusterings successifs, nous avons utilisé le score ARI. Nous avons effectué des essais en évaluant cette similarité à des intervalles mensuels, bimensuels et trimestriels pour déterminer la périodicité optimale de mise à jour du modèle.

