Projet 4: Segmentez des clients d'un site e-commerce

Matthieu MARQUER

Sommaire

Problématique

L’objectif est de fournir une segmentation des clients pour Olist (solution de vente sur marketplaces)

Comprendre les différents types d’utilisateurs grace à leur comportement et à leurs données personnelles.

Fournir à l’équipe Marketing une description actionable de la segmentation

Proposition de contrat de maintenance

Exploration et nettoyage des données

Vérification des types, valeurs manquantes, valeurs uniques par variable.

Renommage des variables geolocation, customers et sellers: “zip_code”, “city” et “state”

Merge des bases de données

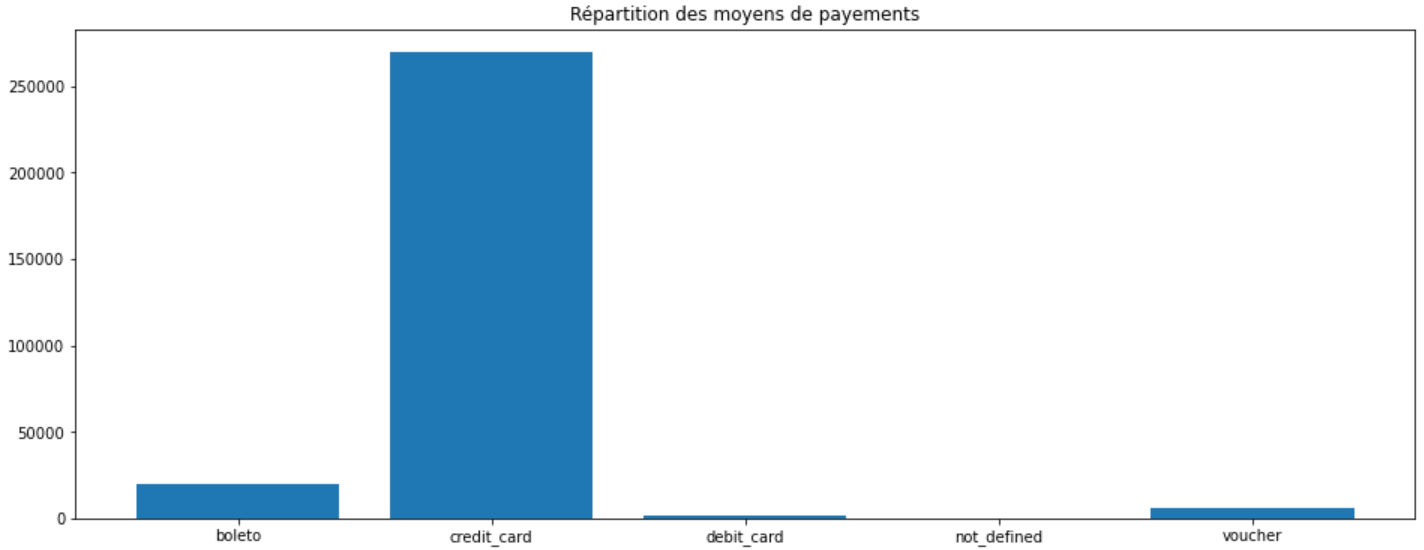
Modification des types en datetime sur orders

Ajout d’un counter sur orders

Création des nouvelles variable:

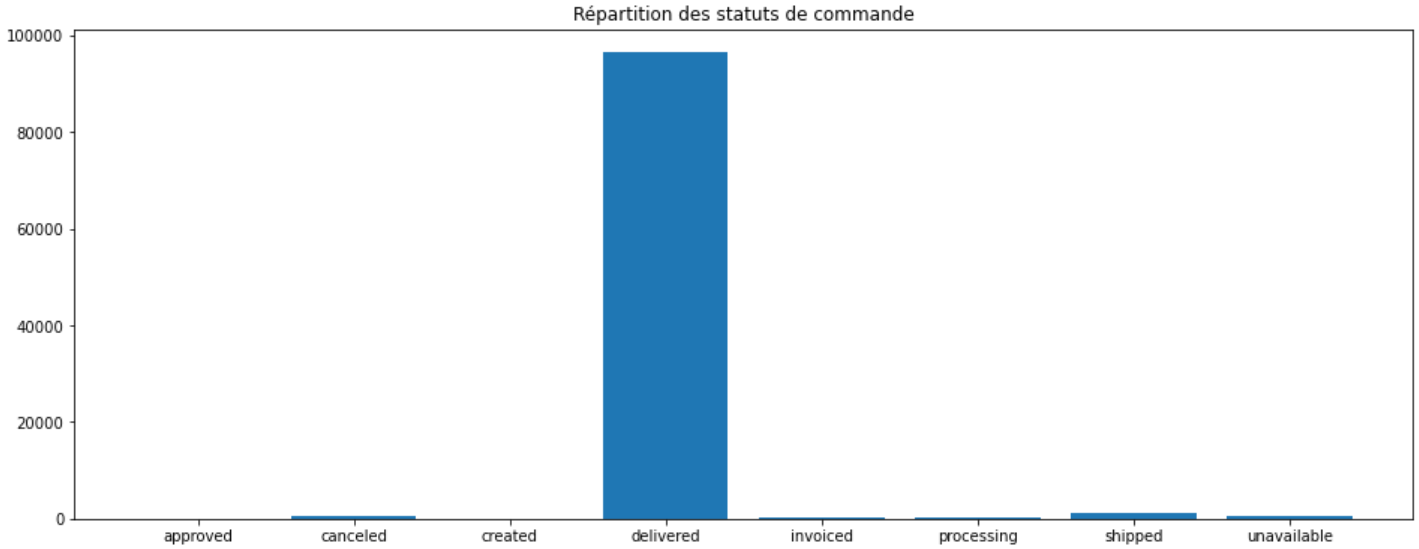
Récence, Frequence, Montant, …

Graphique



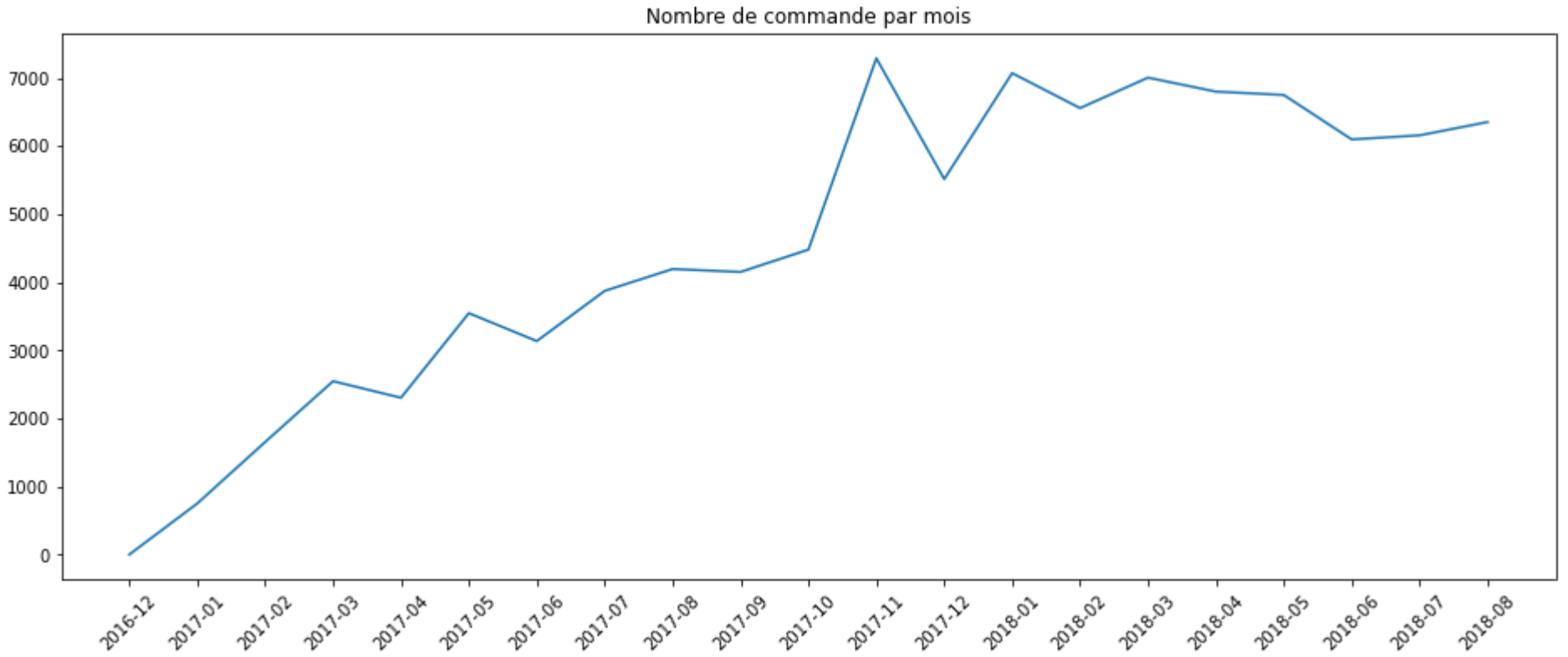
Répartition des moyens de payements

Répartition des statuts de commande



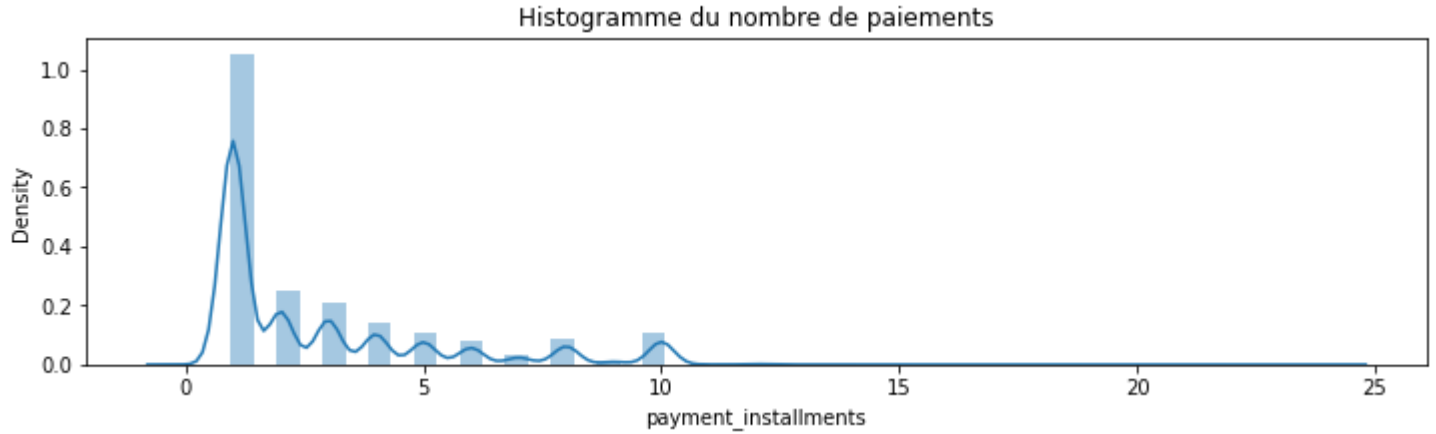
Graphique

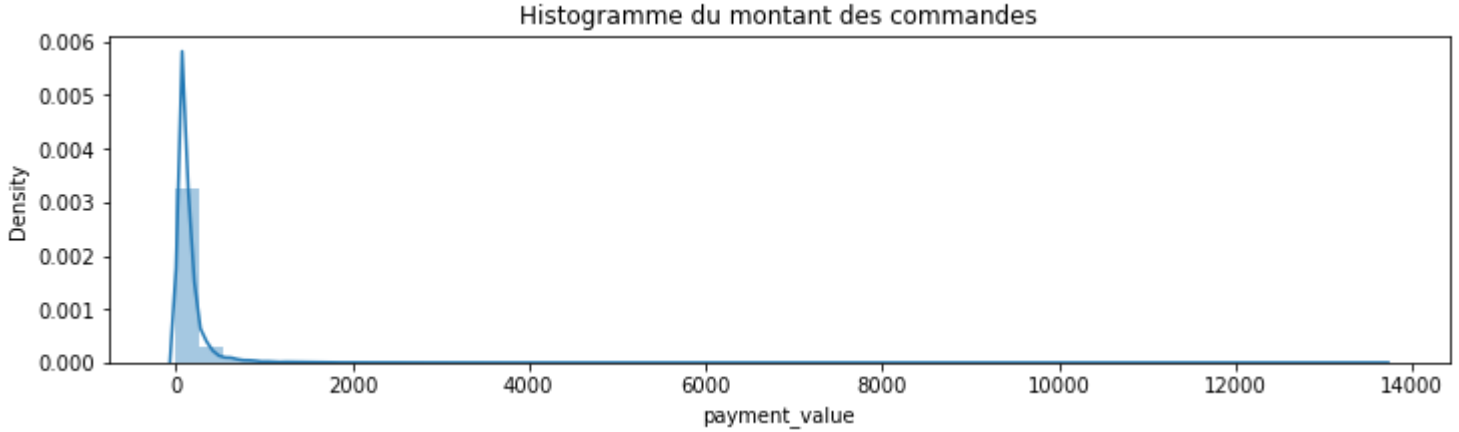
Nombre de commande par mois



Graphique

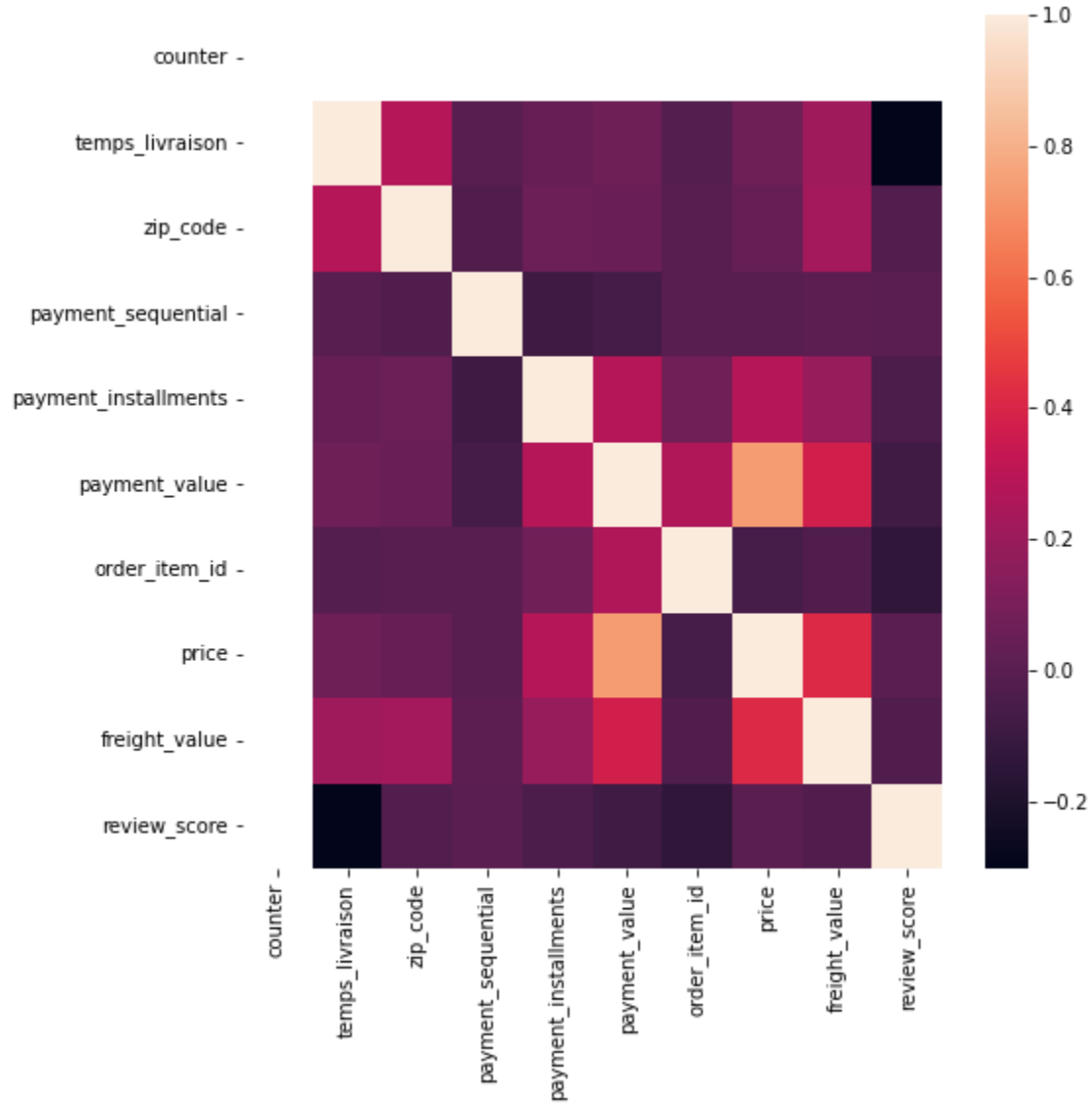
Histogramme du nombre de paiements et du montant des commandes





Heatmap

Peu de correlation sauf entre price et payment_value



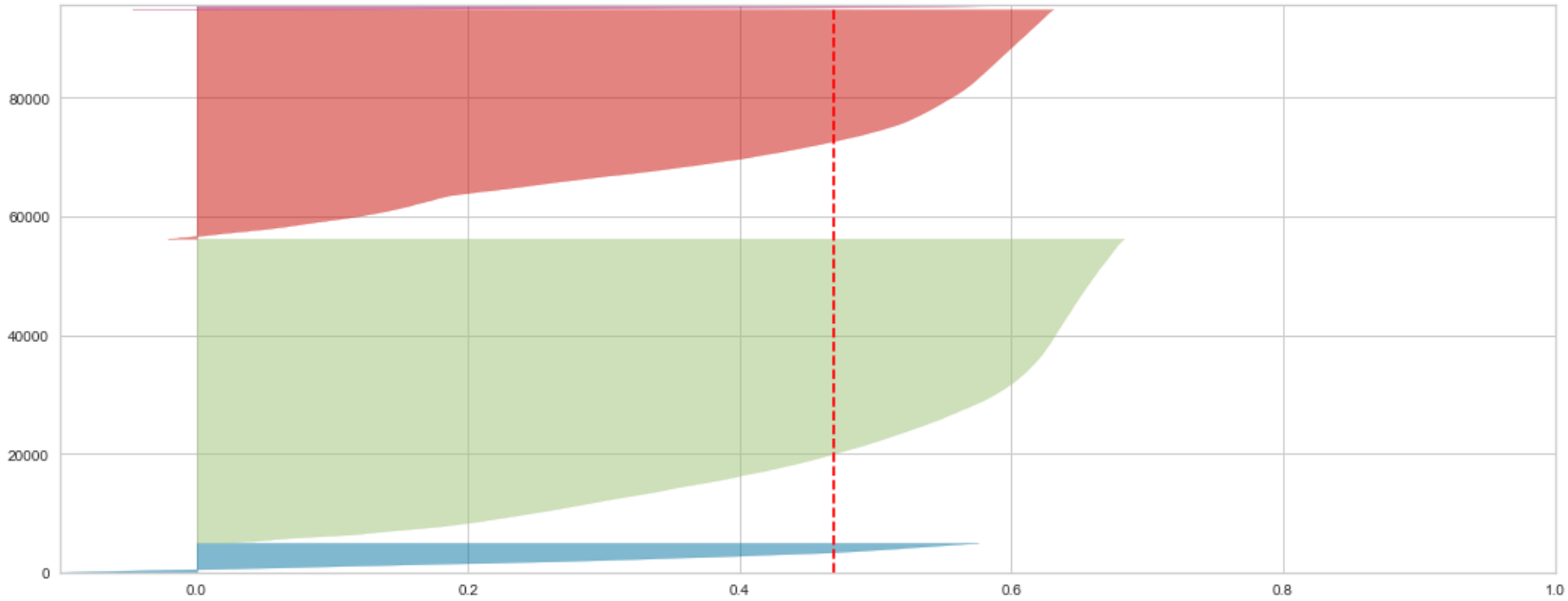
RFM

Création de la segmentation des catégories de clients via Regex (expression régulière)

Kmeans(non centrée Reduit): Silhouette

Sur l’ensemble des données

Score Silhouette: 0.46

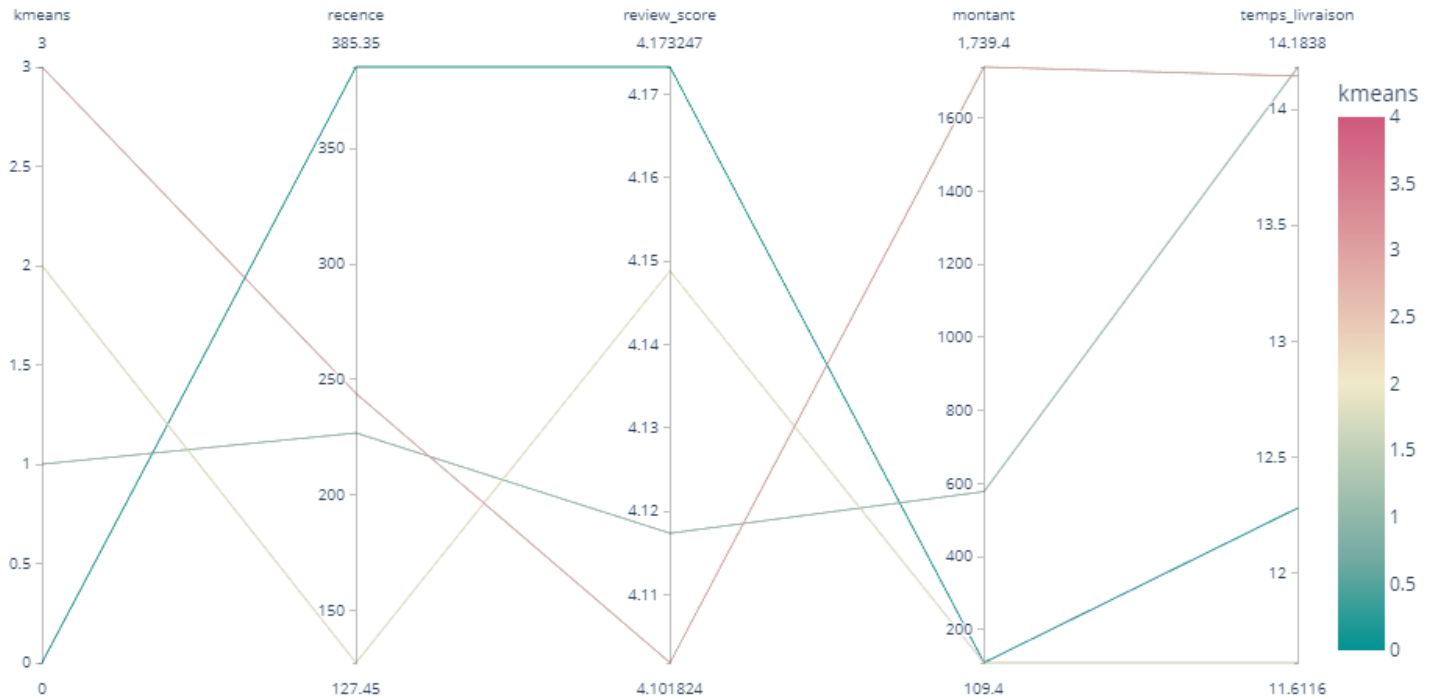


Kmeans (non centrée Reduit): Parallel plot

Parallel plot sur l'ensemble des données suivant les résultats du kmeans

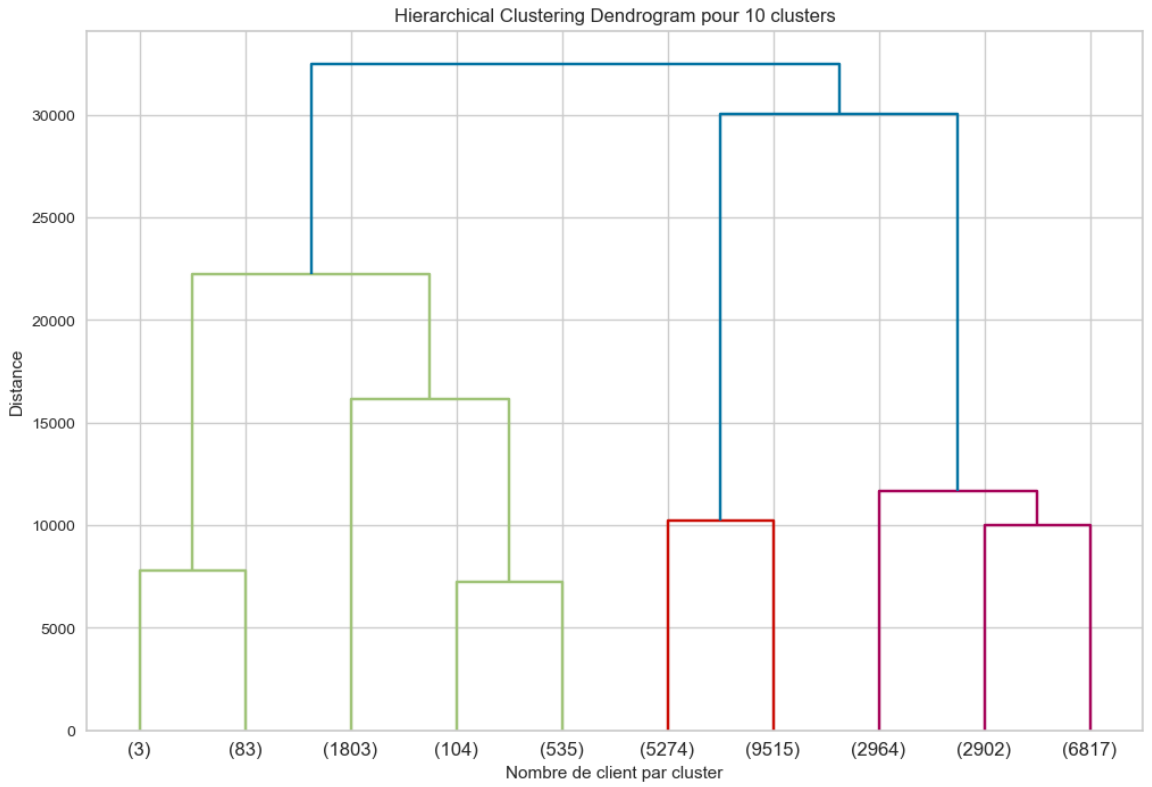
4 clusters:

Clusters 0 et 2: Dépense faibles
Cluster 1: Dépense moyenne
Cluster 3: Dépense forte



CAH

Score Silhouette: 0.39 avec 3 clusters
Calculé sur 30 000 lignes

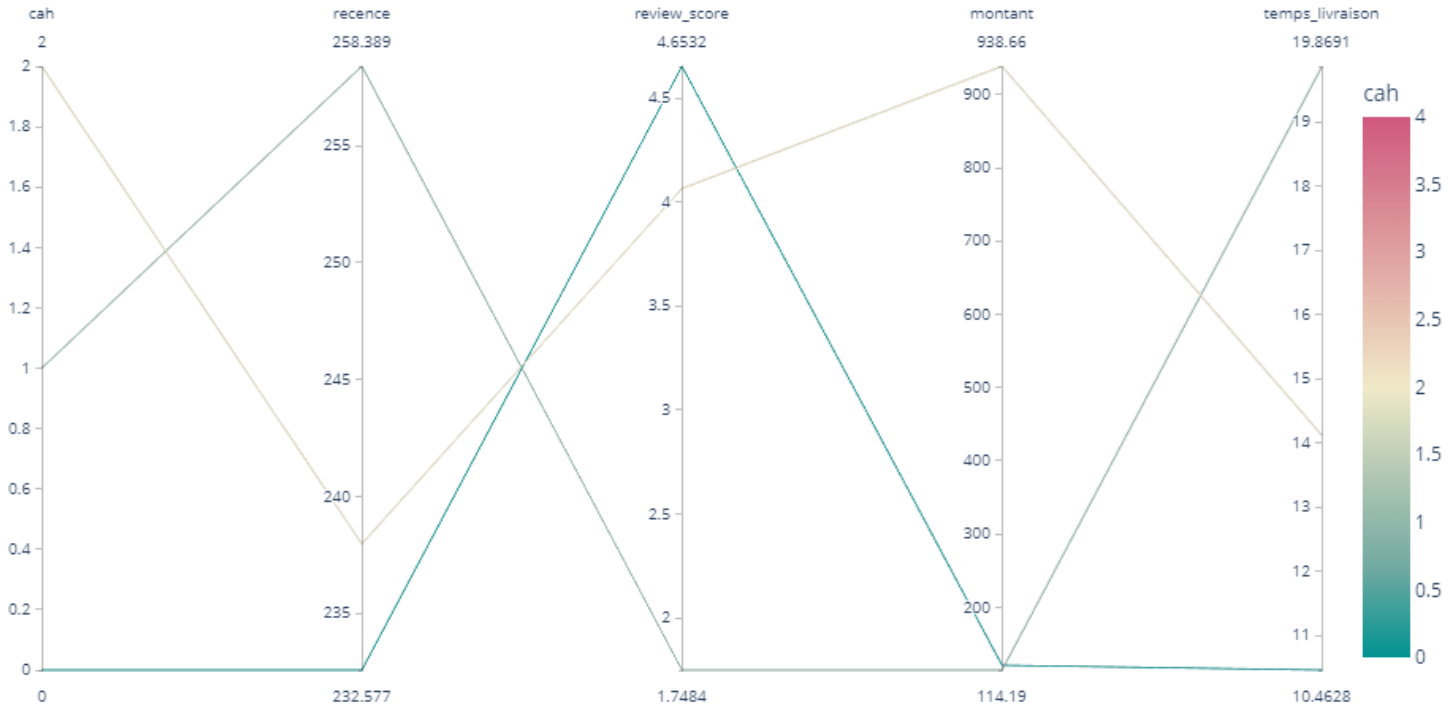


CAH: Parallel plot

Sur 30 000 lignes

3 clusters:

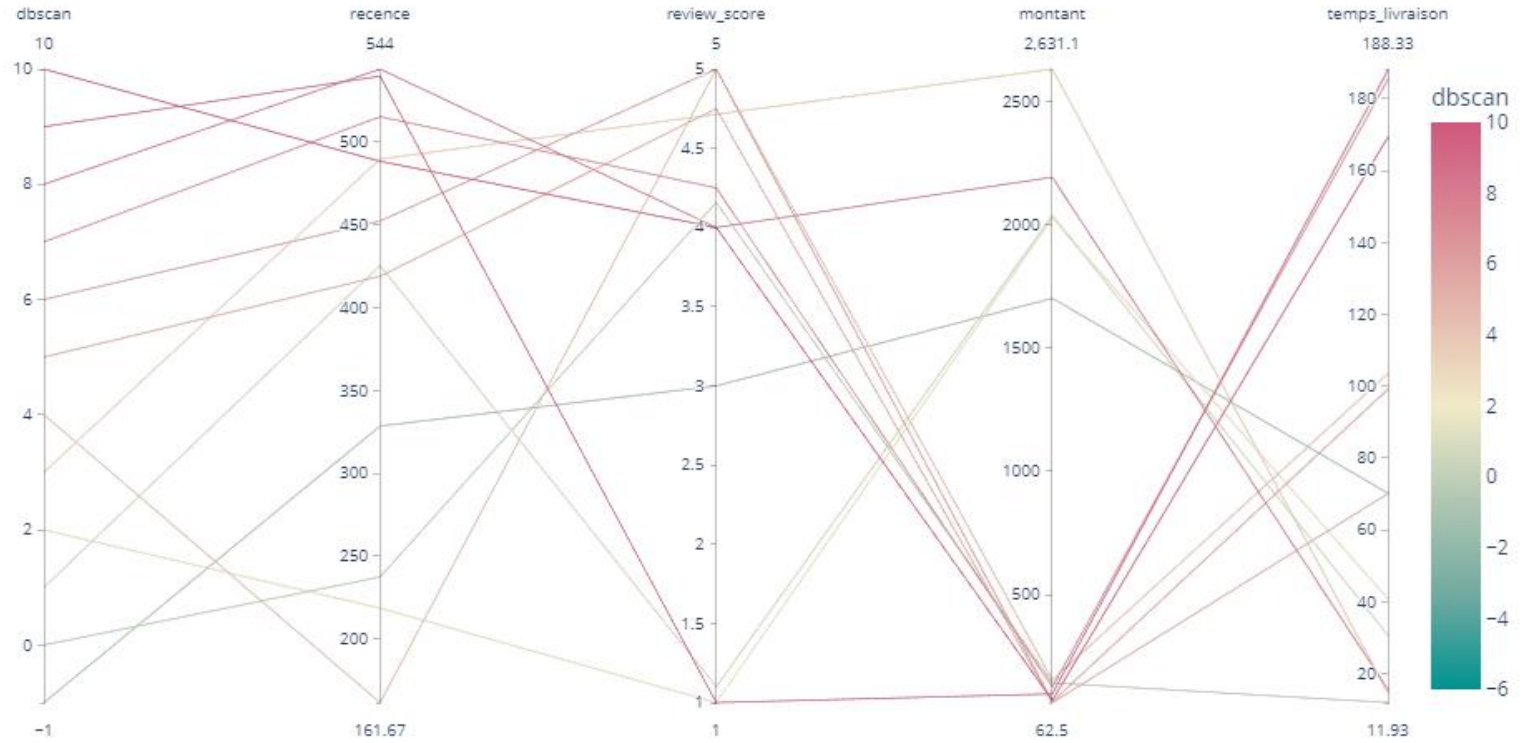
Clusters 0 et 1: Dépense faibles
Cluster 2: Dépense forte



DBscan: Parallel plot

Sur l’ensemble des données
Score Silhouette: 0.82
avec 11 clusters
(plus le bruit en -1)

Trop de clusters ayant un montant assez similaire



Stabilité dans le temps

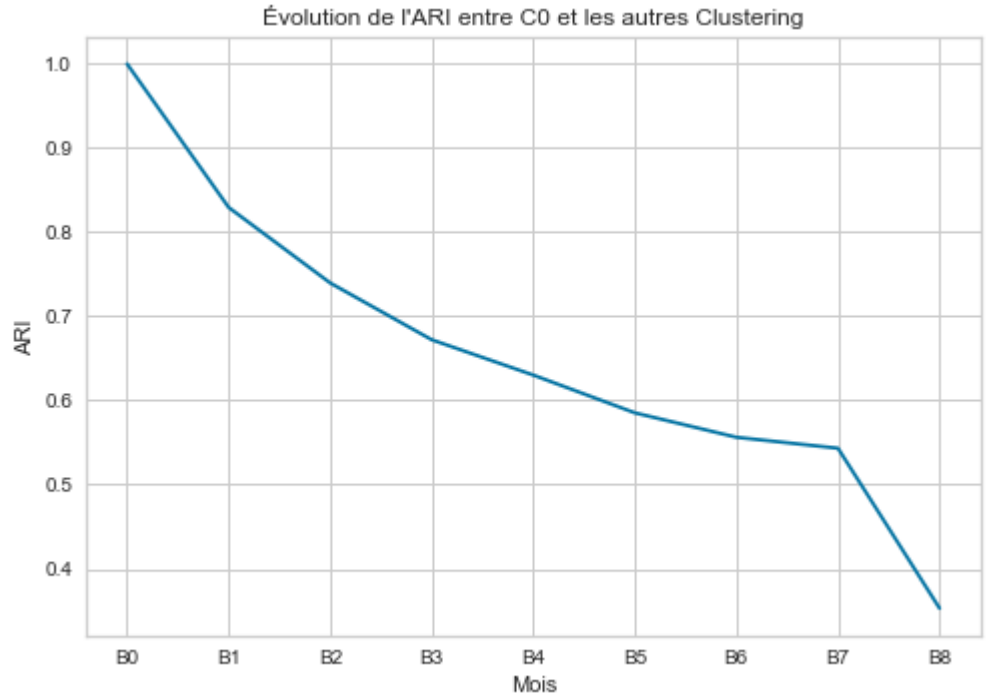
Sur l’ensemble des données

Création de nouveau dataframe B0 à B7 avec:
B0 base d’un an, puis B1 = B0 + 1 mois, … Bn
Vérification des dataframes

Apprentissage des clustering C0 à C7

Segmentation B0 à Bn par rapport à C0 à Cn

Tableau de comparaison des segmentations



Conclusion

1. K-means est le modèle le plus approprié de par son résultat et sa vitesse de calcul.

2. Au niveau de la stabilité dans le temps, nous pourrions recalculer le model tous les 2 mois, mais cela reste à définir après avoir revu le point 3.

3. Il serait important de revoir la partie lié au achat client, car les clients non fait qu'un seul achat chacun.

4. L'ajout d'une variable "utilisation_coupon" suite à un achat après une campagne de mailing pourrait améliorer nos résultats.

Matthieu MARQUER