Segmentez des clients d'un site e-commerce

Yonss JOSE

Parcours: Data Scientist

30/11/2021

Résumé du projet

Olist souhaite avoir une segmentation clients pour leurs campagnes de communication

Objetifs

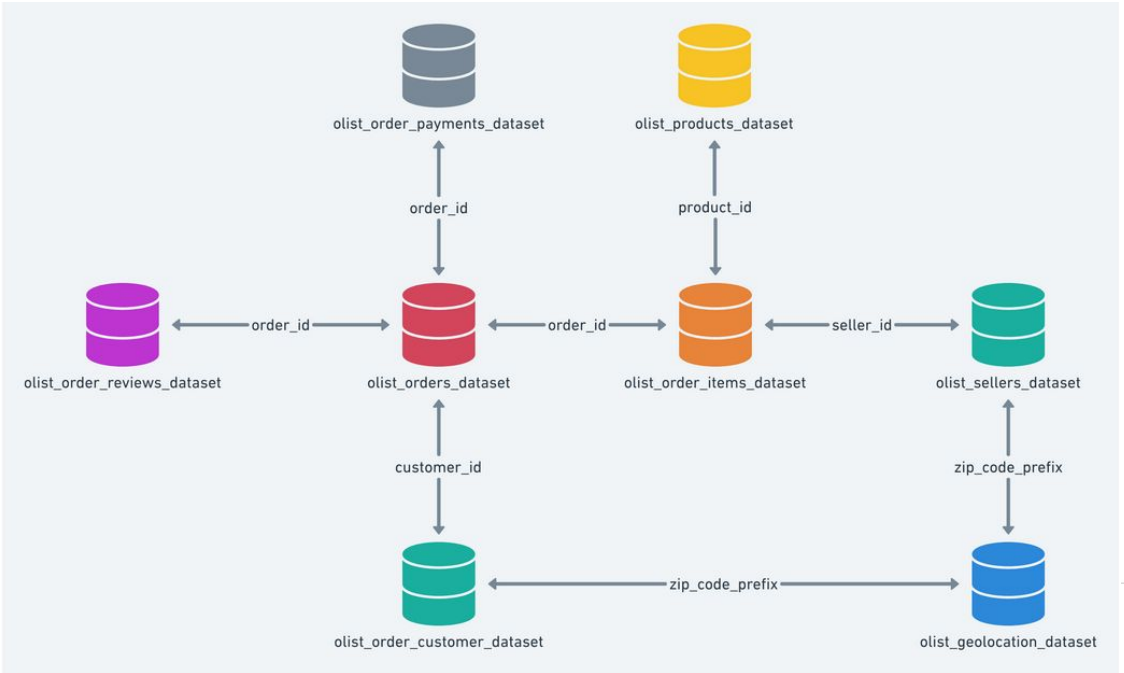
- Comprendre les différents types d'utilisateurs grâce à leur comportement et à leurs données personnelles

Demande du client

- La segmentation proposée doit être exploitable et facile d'utilisation pour l'équipe marketing
- Evaluer la fréquence à laquelle la segmentation doit être mise à jour, afin de pouvoir effectuer un devis de contrat de maintenance
- Le code doit respecter la convention PEP8, pour être utilisable par Olist

Delivrables

- Description actionable de la segmentation et de sa logique sous-jacente pour une utilisation optimale
- Proposition de contrat de maintenance basée sur une analyse de la stabilité des segments au cours du temps



Matériel

Données retenue

Taille de tableaux: (45569,14)

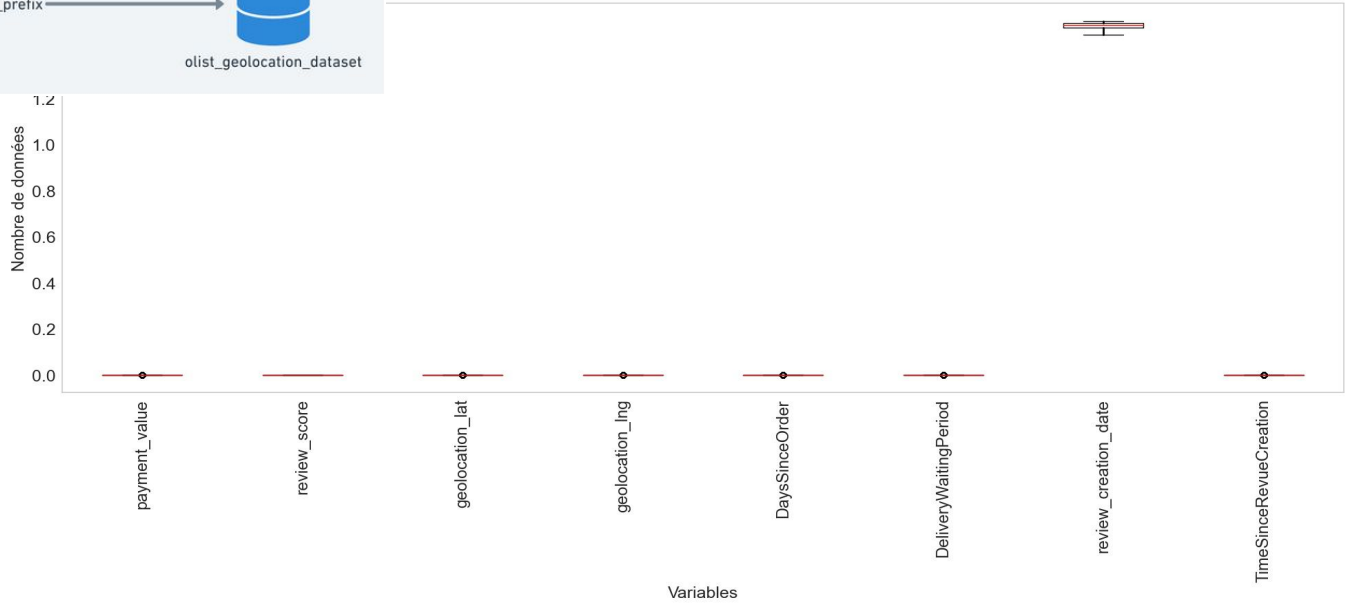
Valeurs manquantes (VM): Non

VM imputé: Non

Doublons: Non

Outliers: écart interquartile

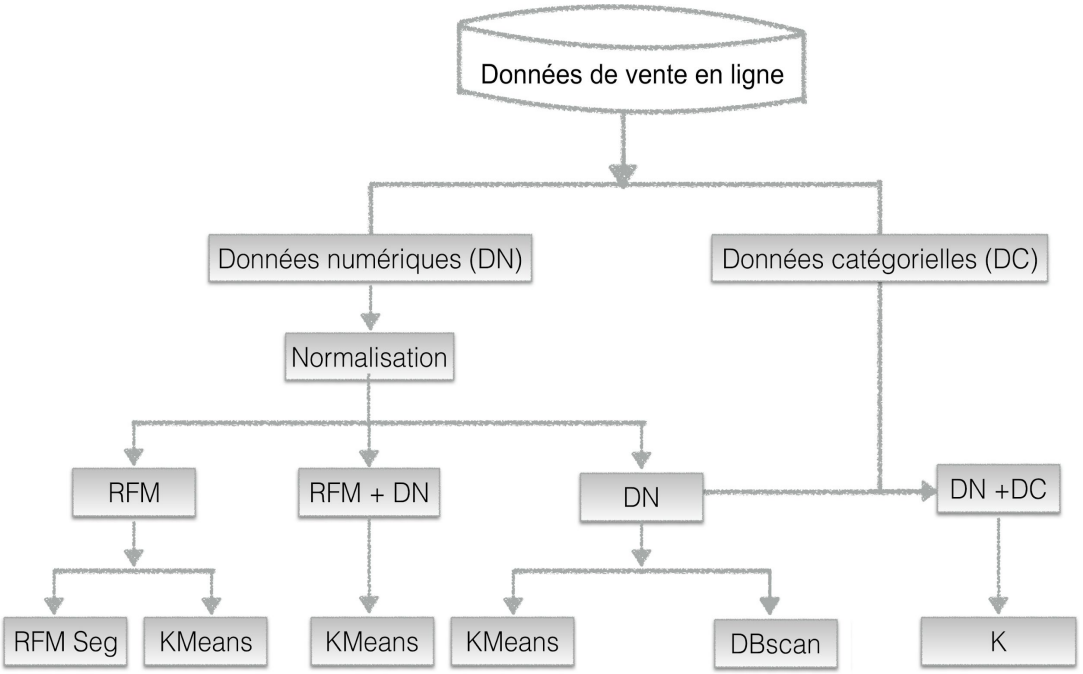
Extraction de dates: Oui



Logiciel : JupyterLab

Méthode

Pour soutenir notre client dans son projet de segmentation de la clientèle, nous avons appliqué la méthode suivante :



Pour cela, nous avons commencé par:

Nettoyer les données et faire l'ingénierie des caractéristiques

Sélectionner les informations pertinentes pour le projet

Avoir un aperçu rapide des données

Avoir un aperçu du comportement des consommateurs Brésilien à Olist

ANALYSE EXPLORATOIRE

Segmentation RFM

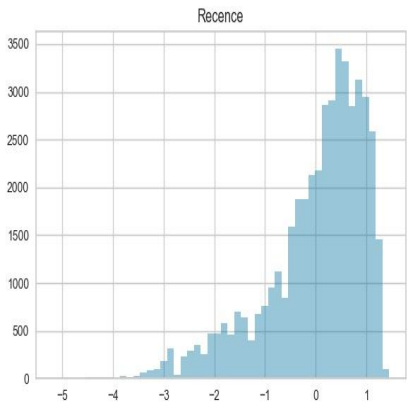
Estimation de Récence , Fréquence et Montant

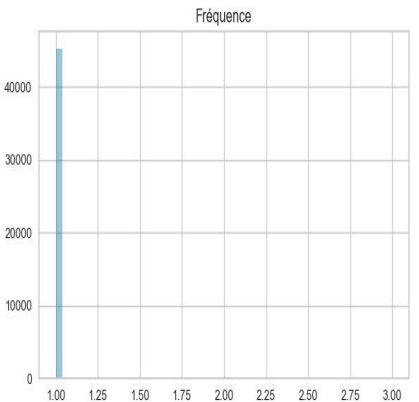
R - Récence - Nombre de jours depuis la dernière transaction du client

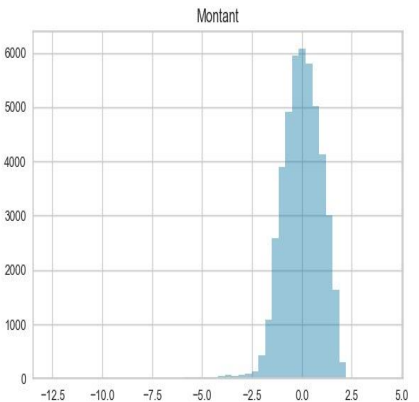
F – Fréquence - Nombre de transactions

M - Montant - Total des achats du client sur une période donnée

Quantile	R score	F score	M score	RFM score	RFM Segment
<=20%	5	1	1	>10	Champions
>20% & <= 40%	4	2	2	10>= & >8	Loyal
>40% & <= 60%	3	3	3	8>= & >6	Need attention
>60% & <= 80%	2	4	4	6>= & >4	At risk
>80%	1	5	5	<=4	Lost







Segmentation des clients

- Utiliser les quantiles pour générer des limites de coupure

- Créer des intervalles basés sur les points de coupure

- Utiliser les intervalles pour attribuer un score

Segmentation RFM

Stratégie commerciale à adopter

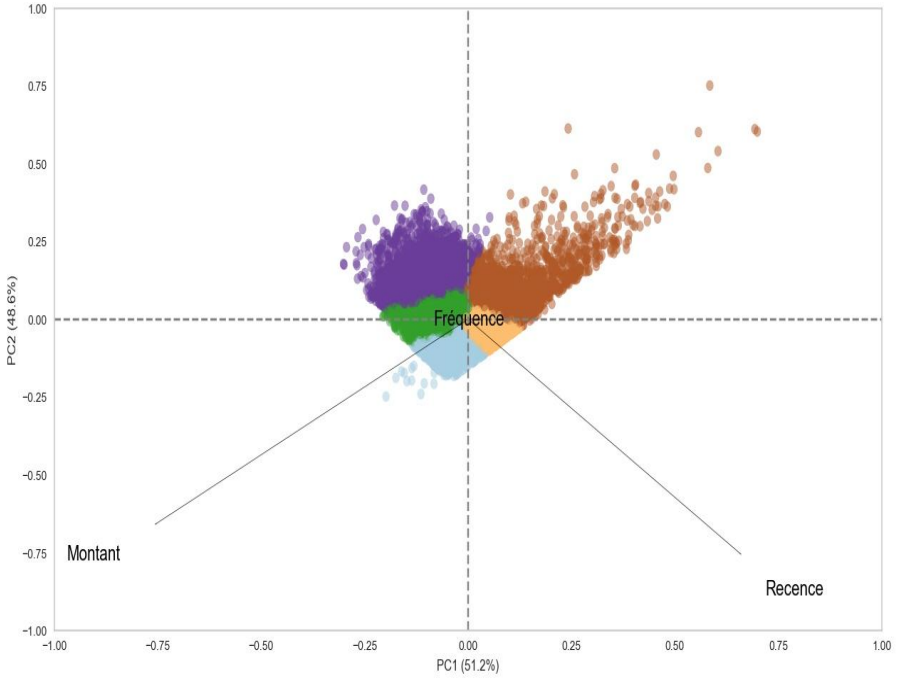
Les 3 segments RFM le plus important:

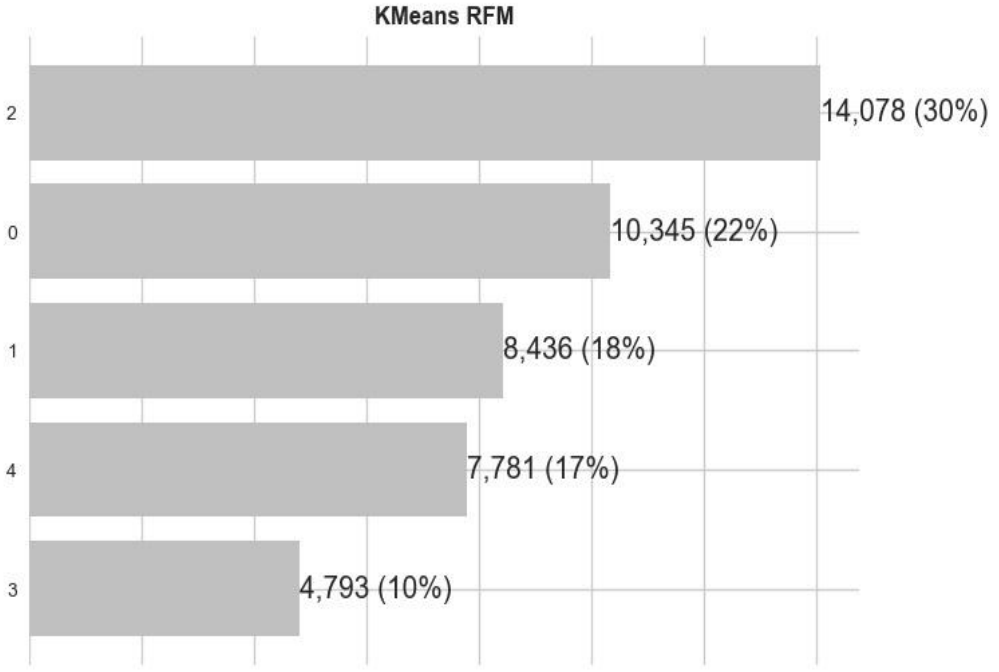
- Need attention (36%)
- At risk (25%)
- Loyal (24%)

Segment	Count	Strategie
At risk	11580	Essayez de les attirer avec des promotions limitées
Champions	3043	Communication personnalisée, offre d'un programme de fidélité
Lost	2838	Ne pas perdre de temps et d'argent pour les gagner
Loyal	11295	Proposer un programme de fidélité
Need attention	16677	Recommander des produits et offrir des promotions

Kmeans avec les données RFM

Nombre de clusters - 5





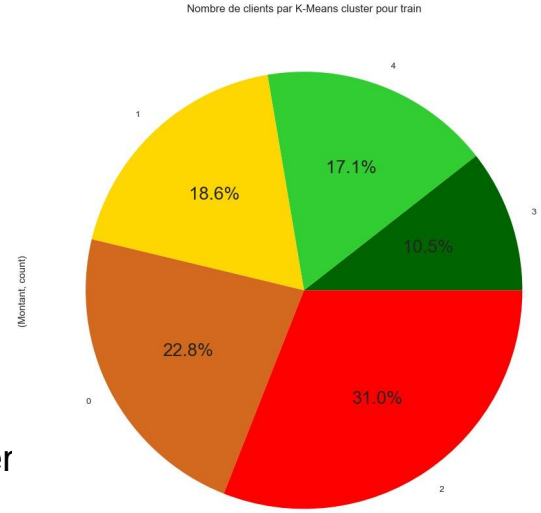
RFM vs Kmeans avec données RFM

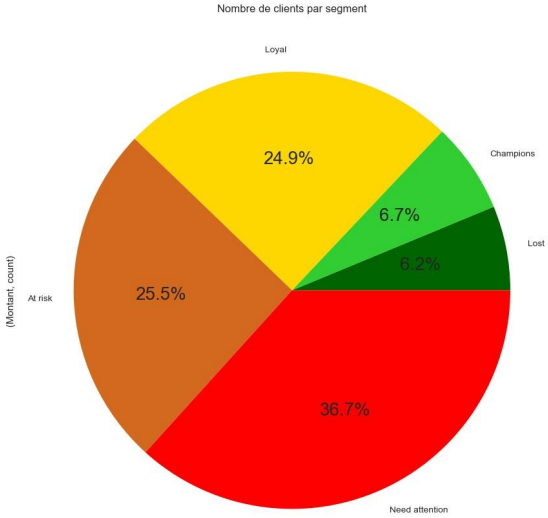
Score silhouette est :

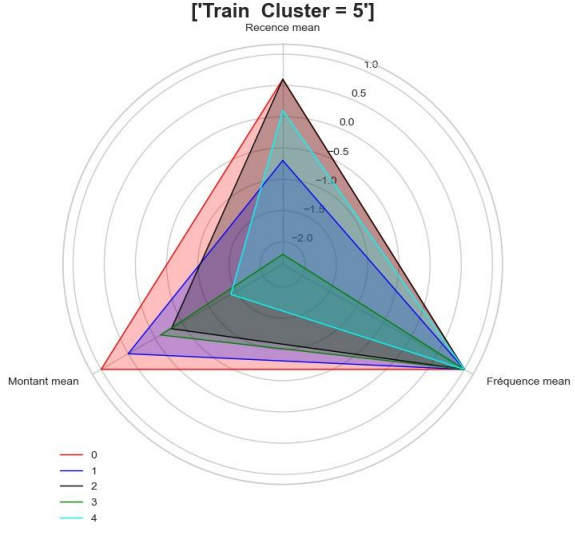
- Élevé dans le K-Means- Capture au mieux la segmentation
- Négatif dans- Classification est faible

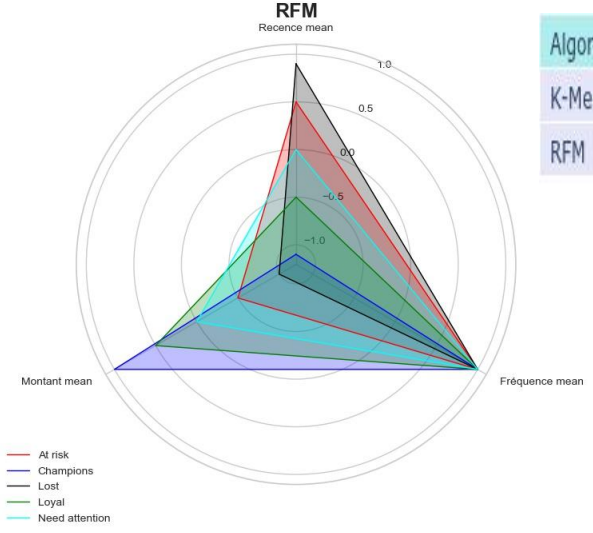
Score Davies-Bouldin est :

-Minimum dans le K-Means - Meilleur regrouper
- Élevé dans RFM - Classification est faible









Algorithms	Davies Bouldin	Silhouette Score
K-Means RFM	0.98	0.33
RFM	2.79	-0.07

Ces résultats mettent en évidence les limitations de la segmentation RFM.

Pour avoir plus de certitude dans la classification obtenue, nous allons tester différent jeux de données

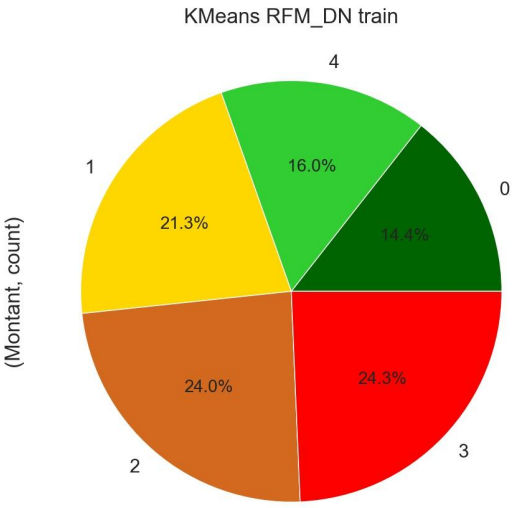
Kmeans avec RFM additionnées aux données brutes (RFM+DN) vs Kmeans avec les données brutes seules (DN)

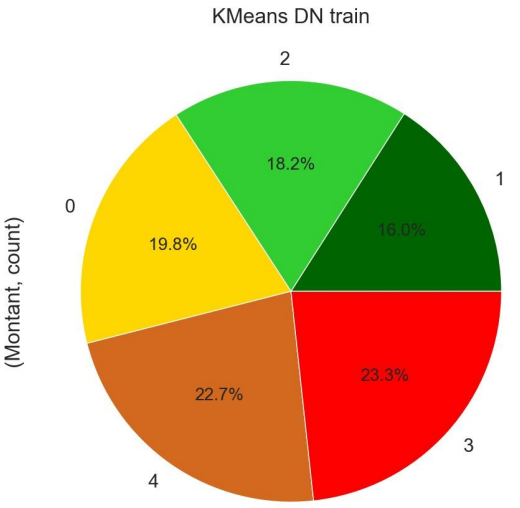
Score silhouette est :

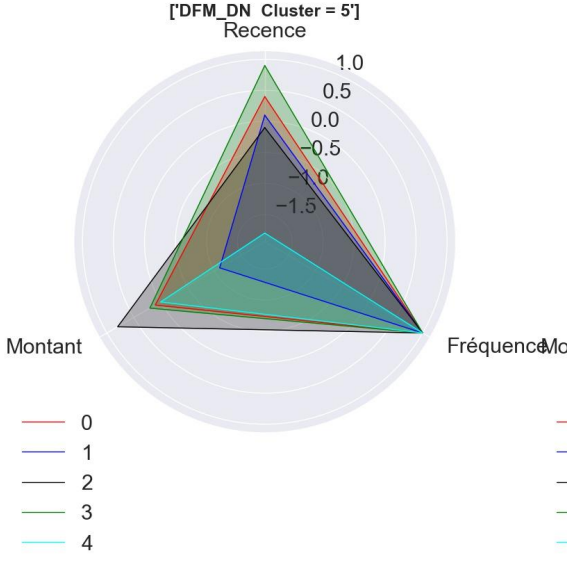
- Élevé dans RFM+DN - Capture au mieux la segmentation

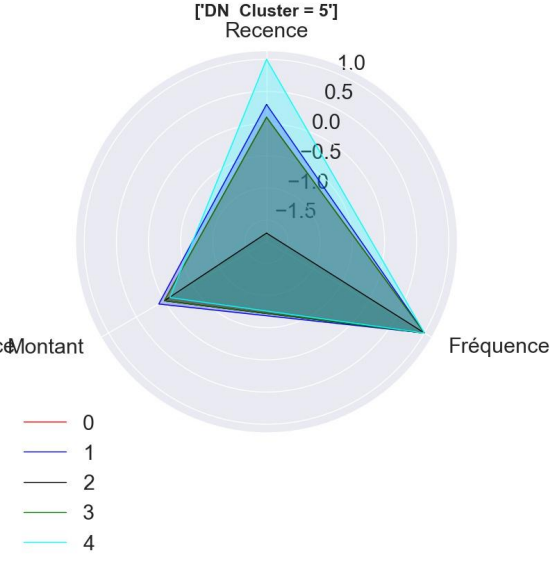
Score Davies-Bouldin est :

-Minimum dans RFM+DN - Meilleur regroupement









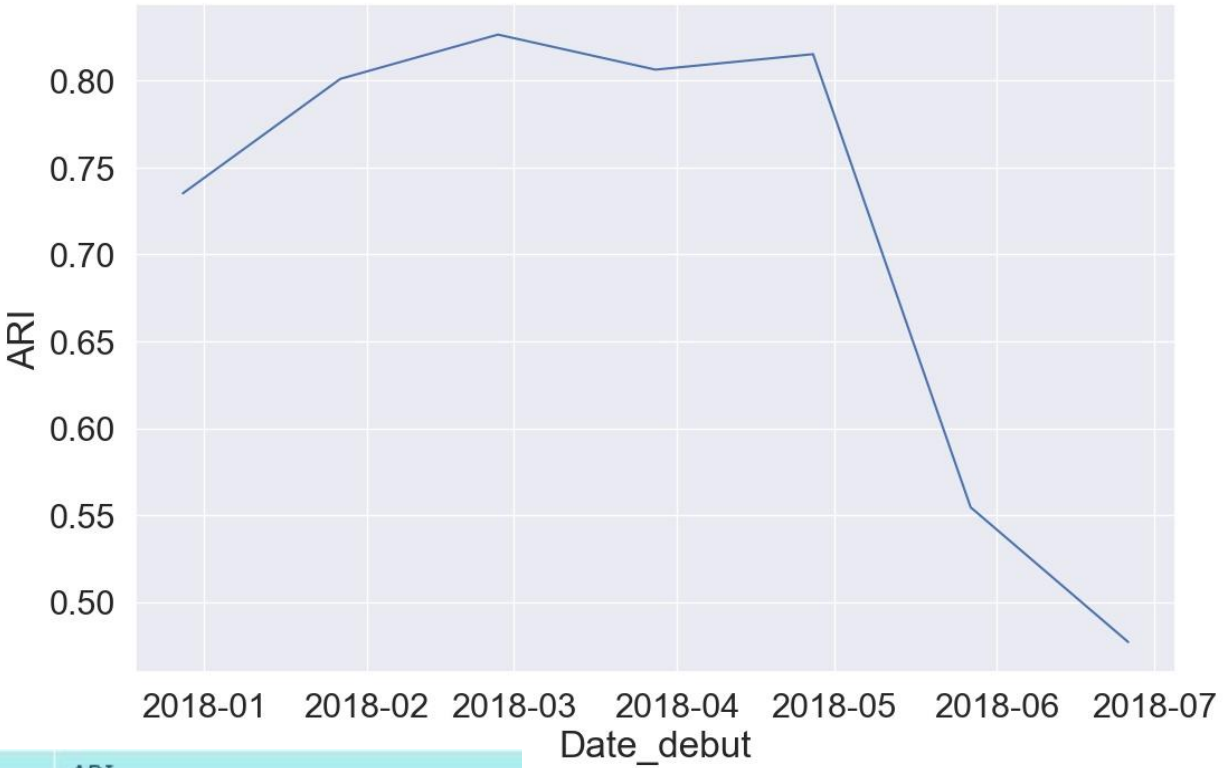
Algorithms	Davies Bouldin	Silhouette Score
K-Means RFM + DN	1.46	0.18
K-Means DN	1.69	0.16

On utilisera les données RFM+DN par la suite

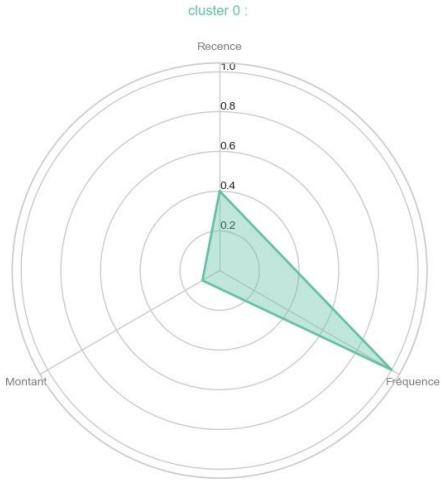
Fréquence de maintenance

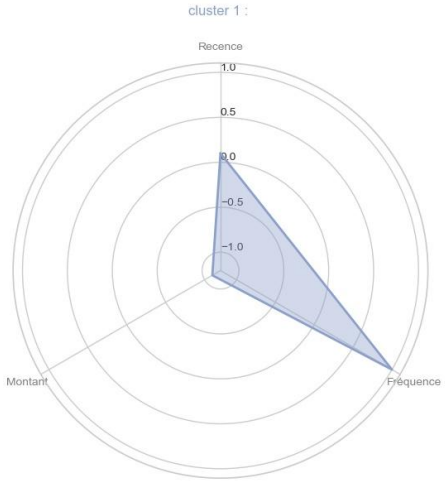
On remarque une baisse significative de la stabilité du modèle à partir d'avril 2018.

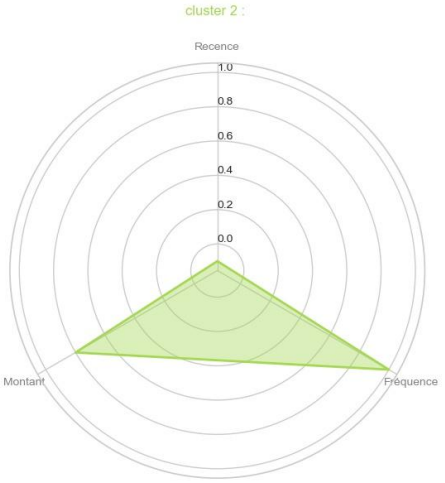
On envisage de faire une maintenance de l'algorithme tous les 3 mois, à partir de Avril 2018.

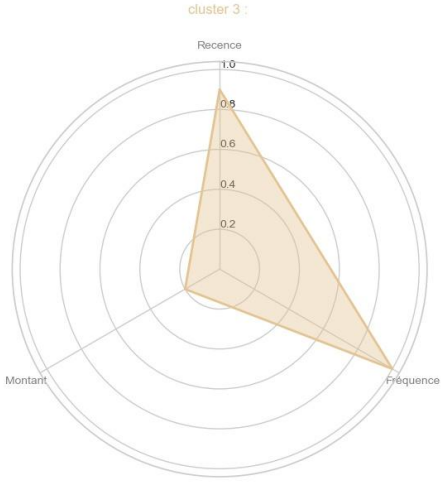


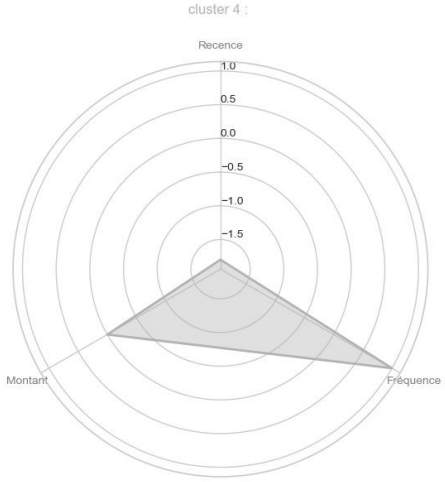
Date_debut	Date_Fin	ARI
2018-06-26T00:00:00	2018-03-28T00:00:00	0.4769605149643045
2018-05-27T00:00:00	2018-02-26T00:00:00	0.5545189897831131
2018-04-27T00:00:00	2018-01-27T00:00:00	0.8152710300231891
2018-03-28T00:00:00	2017-12-28T00:00:00	0.8063519212718807
2018-02-26T00:00:00	2017-11-28T00:00:00	0.8265227763079813
2018-01-27T00:00:00	2017-10-29T00:00:00	0.8010650981427119
2017-12-28T00:00:00	2017-09-29T00:00:00	0.7351569040833176











Conclusion

Cluster 3 : Premier cluster ayant le plus grand nombre de clients. La récence la plus élevée mais la valeur du paiement la plus faible - **Lost**

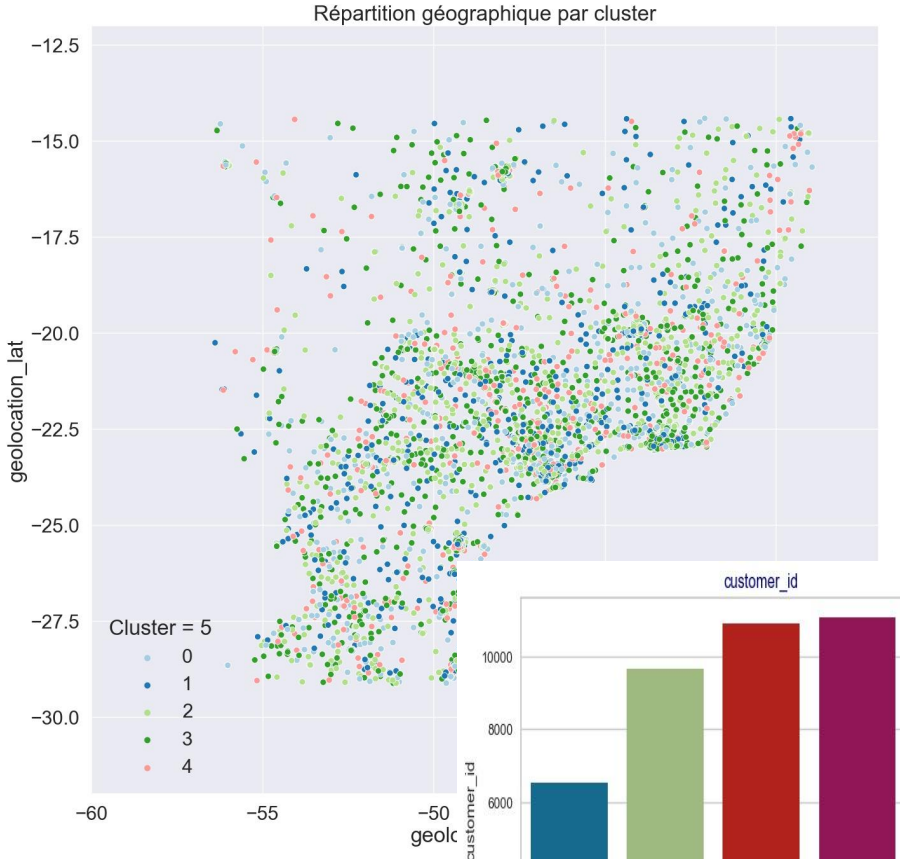
Cluster 2 : Deuxième cluster ayant le plus grand nombre de clients et la plus grande valeur de paiement - **Champions**

Cluster 4 : Ce groupe a le plus grand nombre de clients et la plus faible récence - **Loyal**

Cluster 1 : Troisième cluster ayant le plus grand nombre de clients, mais la valeur du paiement la plus faible – **Need attention**

Cluster 0 : Le plus petit nombre de clients se trouve dans ce groupe, qui se situe au-dessus de la moyenne en termes de récence et de valeur de paiement – **At risk**

Pas de patterns géographiques



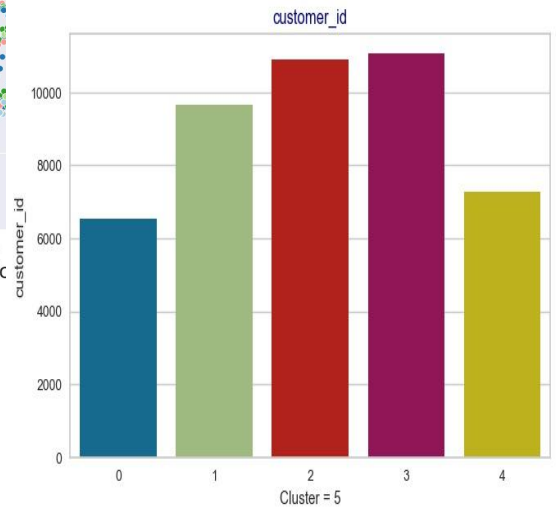
Cluster 3 - Lost

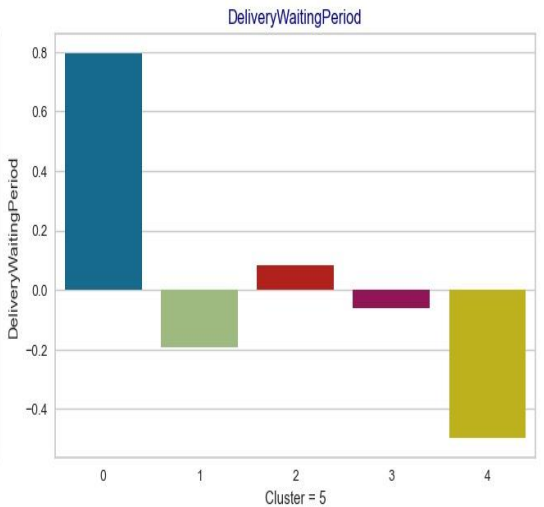
Cluster 2 - Champions – Attention à ne pas augmenter le temps de livraison

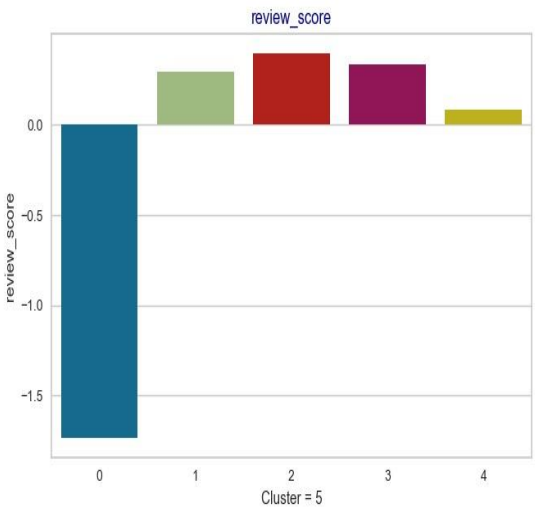
Cluster 4 - Loyal – Attention : **clients pas contents**

Cluster 1 - Need attention

Cluster 0- At risk – Clients pas contents. Améliorer le temps de livraison







- Segmentation RFM ne semble pas adapté pour faire la répartition des clients pour Olist
- Segmentation peut être amélioré avec la méthode des kmeans
    - K-Prototypes a un temps de calcule élevé le élevé en comparaison avec Kmeans
    - Dbscan considère que 3183 données sont des bruits
- On envisage de faire une maintenance de l'algorithme tous les 3 mois, à partir d’Avril 2018