SEGMENTEZ LES CLIENTS D'UN SITE DE E-COMMERCE

Victoire MOHEBI
Juin 2022

OPENCLASSROOMS

AGENDA

- Problématique & Objectifs
- Mission
- Présentation de jeux de données
- Analyse exploratoire
- Modélisation
- Simulation
- Conclusion
- Recommandation

PROBLÉMATIQUE & OBJECTIF

Problématique :

Olist , un site de e-commerce au Brézil, souhaite obtenir une segmentation de ses clients.

Objectif :

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

MISSION

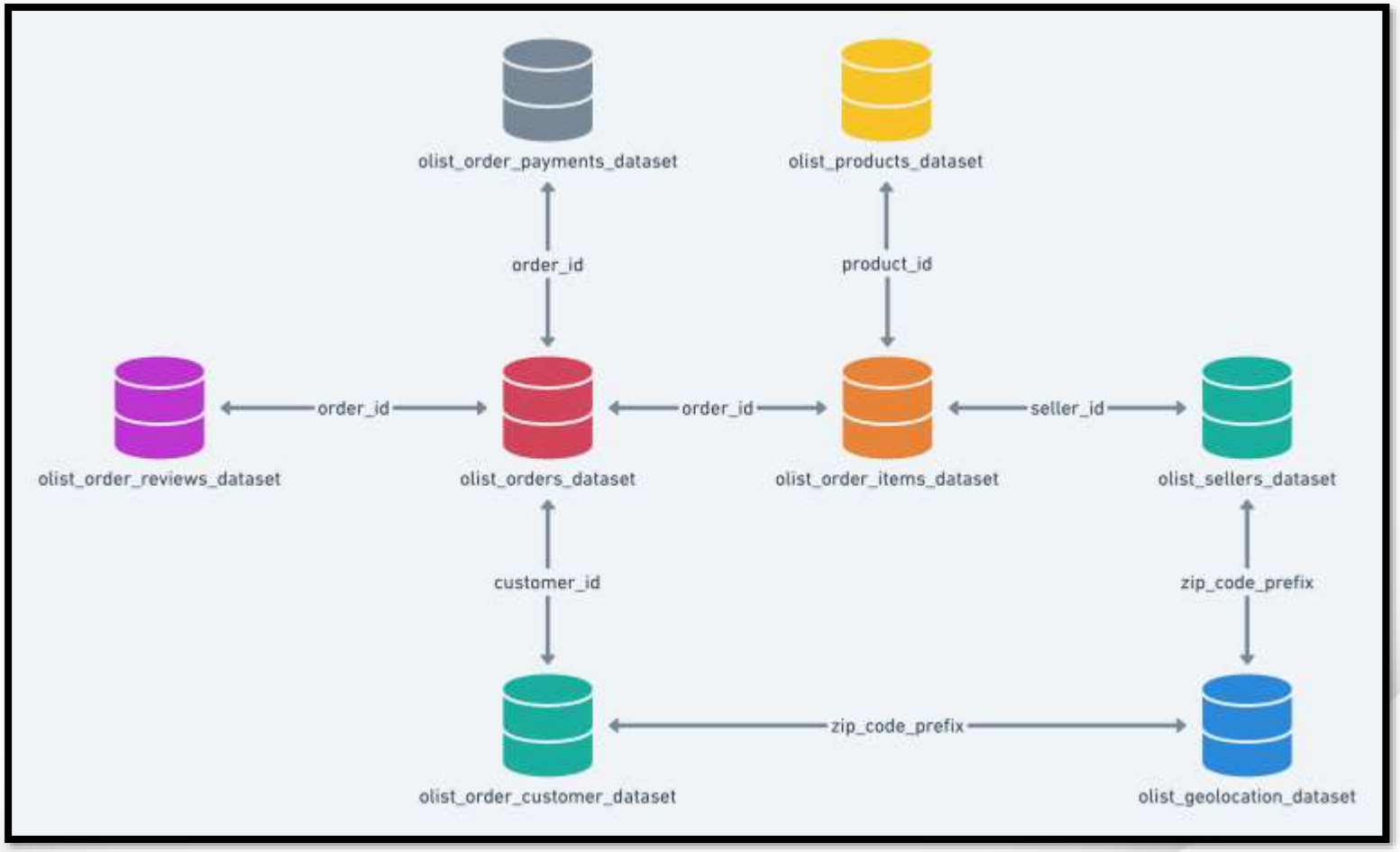
- **Fournir une description actionnable de la segmentation et de sa logique sous-jacente pour une utilisation optimale.**
- **Proposer un contrat de maintenance basée sur une analyse de la stabilité des segments au cours du temps**

PRÉSENTATION DES DONNÉES

- **Sources des données : *Brasilia E-Commerce Public Dataset* by Olis**
- **Neuf tables de dimension différentes**

- **Les informations anonymisées sur l'historique des commandes, les produits achetés, les commentaires de satisfaction, et la localisation des clients**
- **Les donnée complète sur 23 mois**

PRÉSENTATION DES DONNÉES

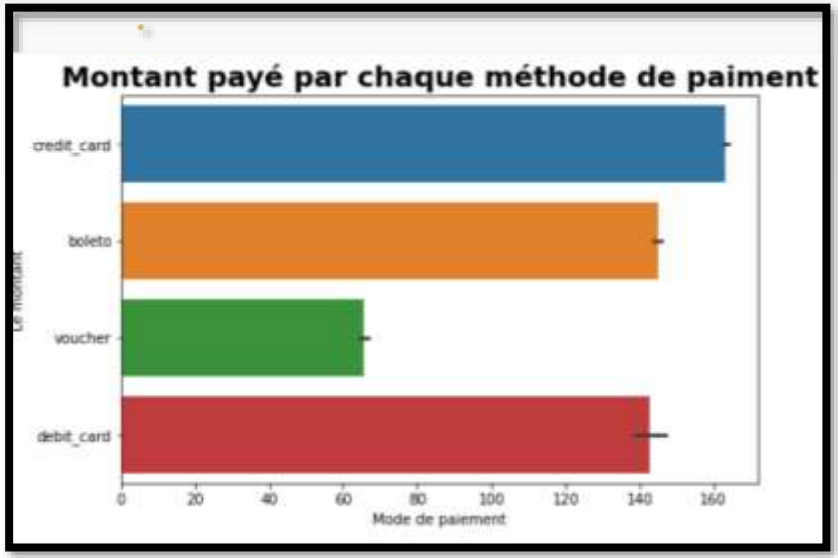


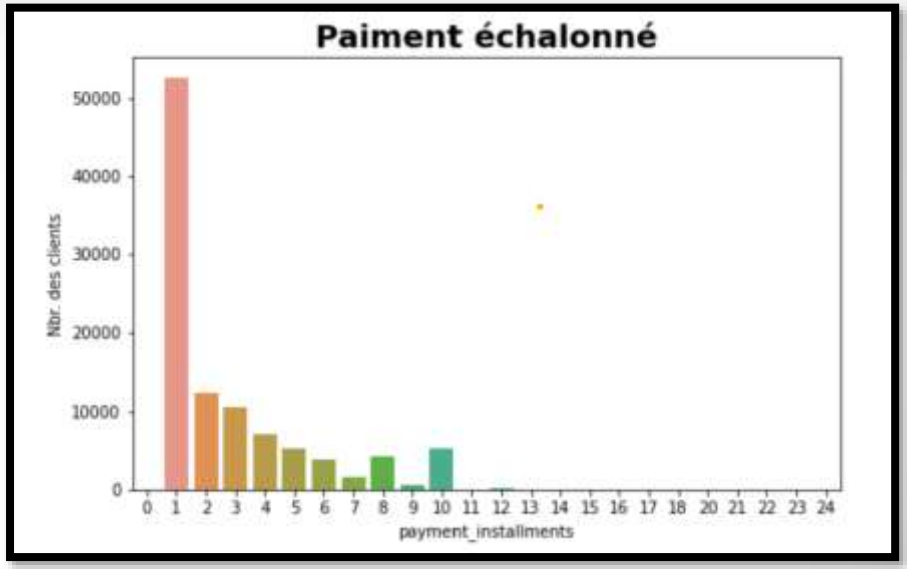
ANALYSE EXPLOIRATOIRE

4

- Les clients sont de 4119 villes et de 27 états en Brézil
- Les vendeurs sont de 23 état en Brézil

ANALYSE EXPLOIRATOIRE





- Les clients ont payé avec quatre modes de paiement
- Ils ont payé en plusieurs fois

ANALYSE EXPLOIRATOIRE

- La note moyenne donnée par les client est élevée
- La majorité des clients sont satisfaits
- Le délai moyen de livraison est de 12 jours

	delivery_days	payment_installments	frequence	Note_Moy_Com	recence	panier_moyen
count	92746.000000	92746.000000	92746.000000	92746.000000	92746.000000	92746.000000
mean	12.060600	2.898307	1.033177	4.153367	237.798245	204.376237
std	9.466231	2.675020	0.208425	1.280530	152.591525	603.916800
min	0.000000	0.000000	1.000000	1.000000	1.000000	9.590000
25%	6.000000	1.000000	1.000000	4.000000	114.000000	63.150000
50%	10.000000	2.000000	1.000000	5.000000	219.000000	110.700000
75%	15.000000	4.000000	1.000000	5.000000	346.000000	195.690000
max	208.000000	24.000000	15.000000	5.000000	695.000000	109312.640000

SEGMENTATION RFM

- **Approche analytique**
- **Approche automatique : méthodes non supervisées pour regrouper ensemble des clients de profils similaires**

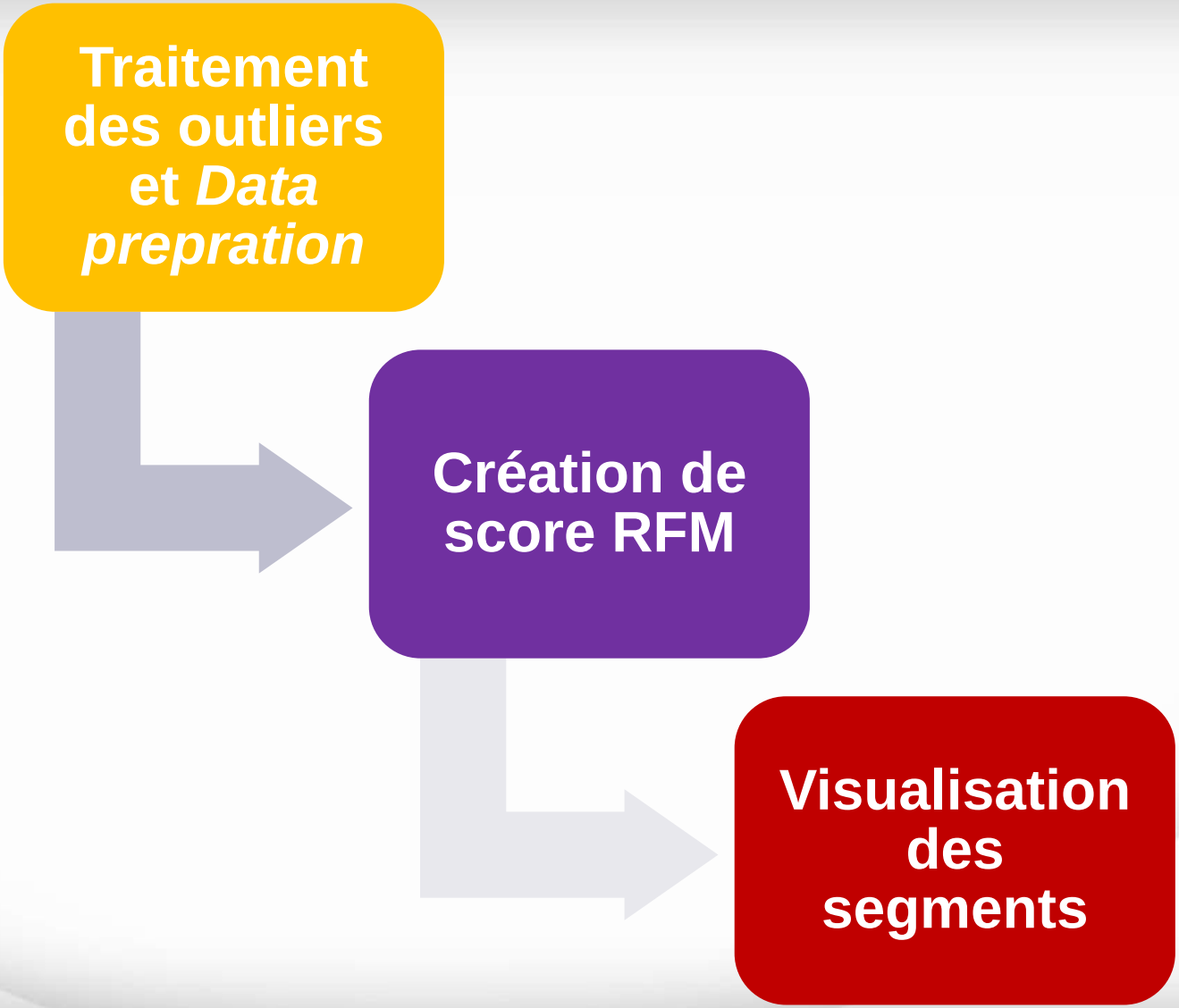
APPROCHE ANALYTIQUE

RFM : Technique de segmentation des clients, basée sur 3 données

- « *Recency* » : la date du dernier achat
- « *Frequency* » : nombre d'achats effectués sur une période précise
- « *Monetary* » : total des achats sur la même période

DÉMARCHE

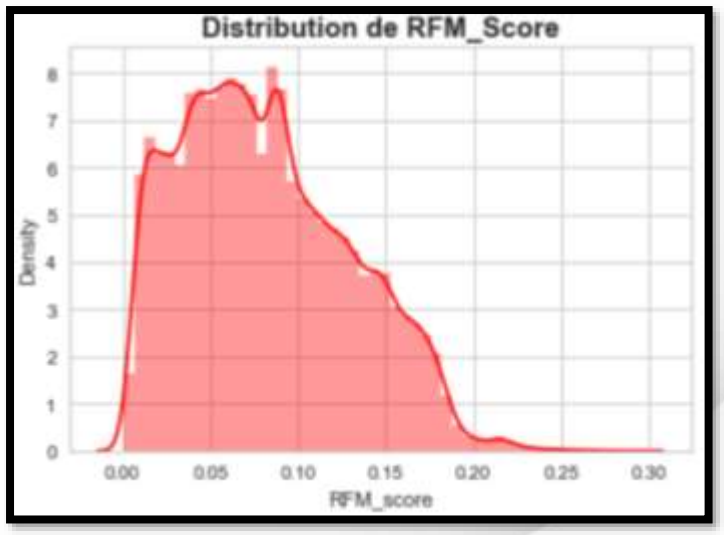
APPROCHE ANALYTIQUE



CRÉATION D'UN SCORE RFM

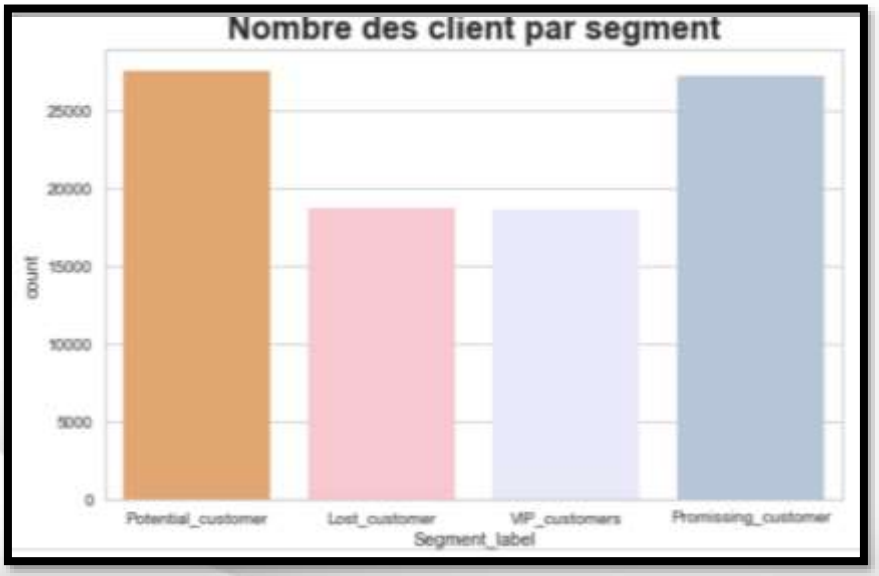
Sur la base du tableau RFM, nous attribuerons trois score 'R', 'F', 'M' à chaque client
- Le moyenne de ces trois scores est le "RFM_score" pour chaque client
- Pour calculer trois score 'R', 'F', 'M', on divise le 'std' de chaque variable par le somme des 'std' de trois variables

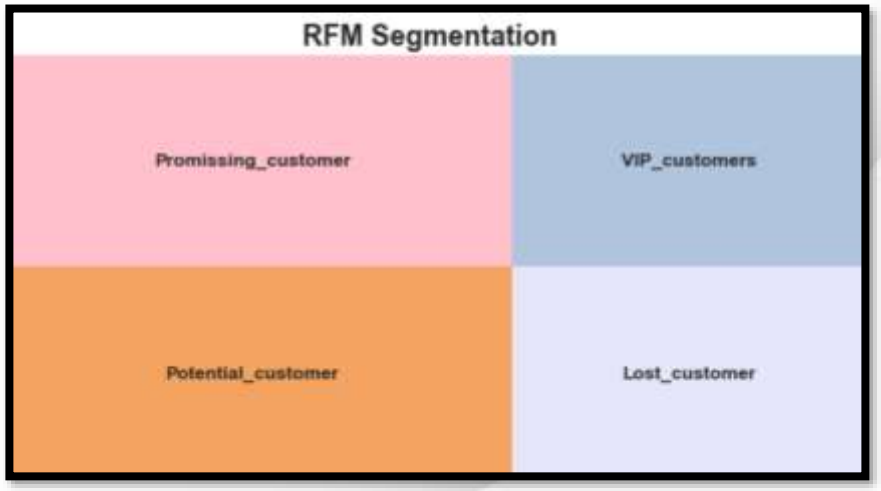
```
std_r = rfm_normalized['recence'].std()
std_f = rfm_normalized['frequence'].std()
std_m = rfm_normalized['panier_moyen'].std()

sum_var = std_r + std_f + std_m

# Crée un poid pour R, F, M
w_r = std_r/ sum_var
w_f = std_f / sum_var
w_m = std_m /sum_var

#Créer un Score R, F et M basé sur le poid
rfm_normalized['R_score'] = w_r * rfm_normalized['recence']
rfm_normalized['F_score'] = w_f * rfm_normalized['frequence']
rfm_normalized['M_score'] = w_m * rfm_normalized['panier_moyen']

# Créer un score final de RFM
rfm_normalized['RFM_score'] =  ((rfm_normalized['R_score'] + rfm_normalized['F_score'] + rfm_normalized['M_score']))/3
```



LABÉLISATION ET VISULISATION

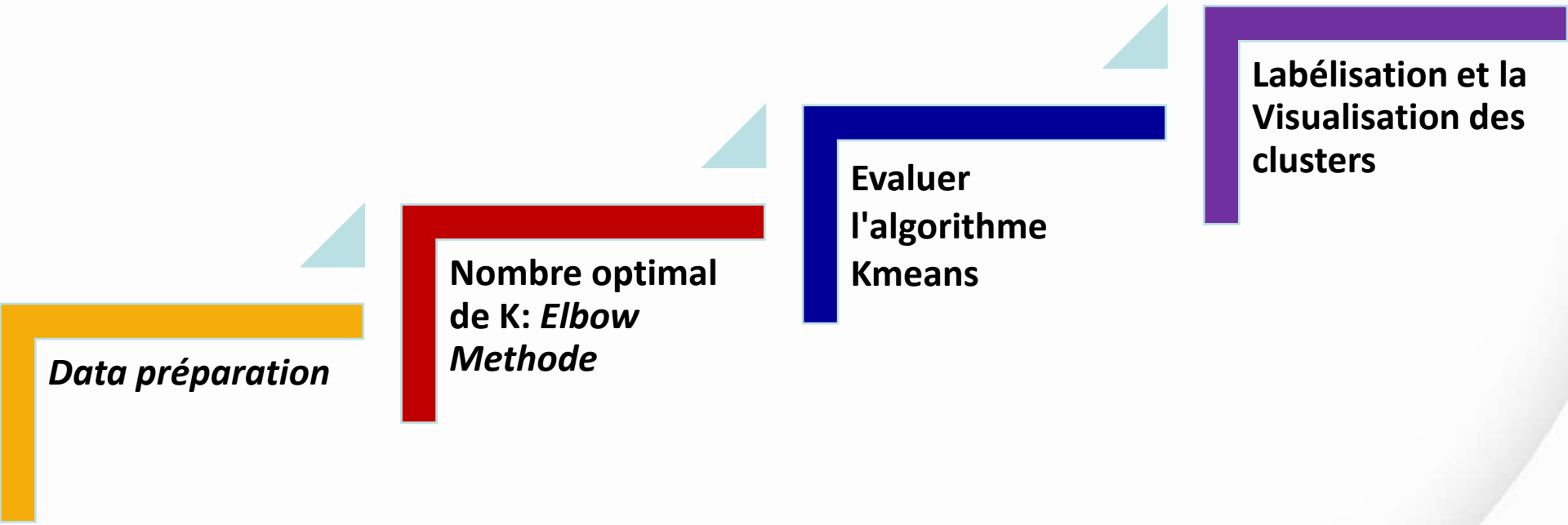
Label	Description
Best Customers	rfm_score >= quantile(0.8)
Potential Customers	Quantile(0.5) <rfm_score < quantile(0.8)
Almost Lost Customers	Quantile(0.2)<rfm_score <=quantile(50)
Lost Customers	rfm_score <= quantile(0.2)





DÉMARCHE

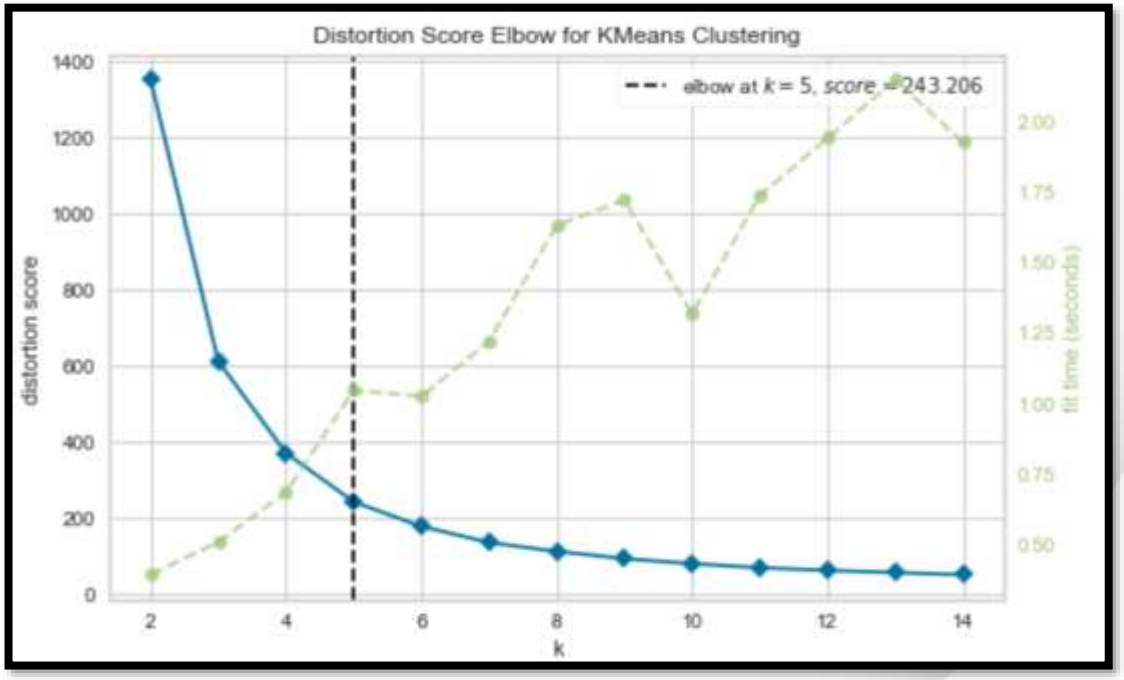
MÉTHODES NON SUPERVISÉES - Kmeans



NOMBRE OPTIMAL DE K

Grâce à la méthode du coude basée sur le score de distortion *(somme moyenne des carrés des distances aux centres)*, une segementation en K=5 clusters serait la meilleure option.

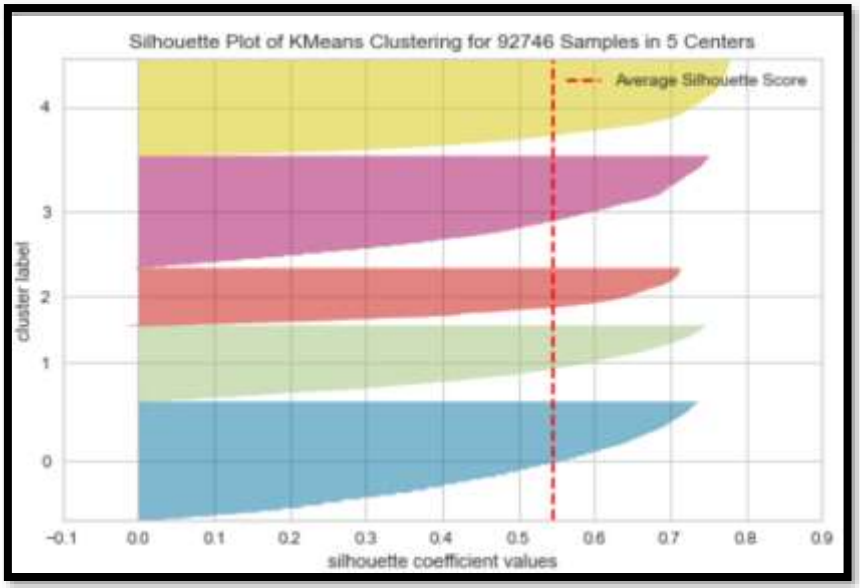
Méthode du coude : déterminer le meilleur K



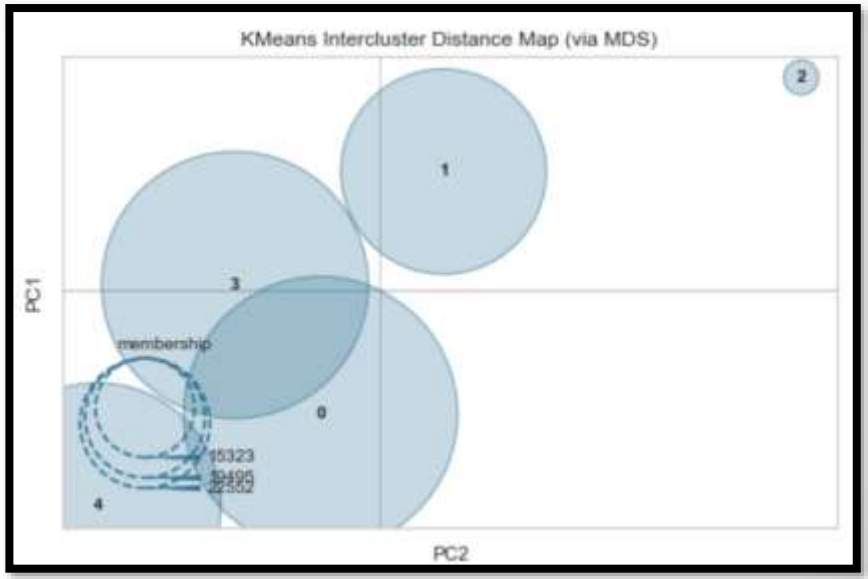
EVALUER LA PERFORMACE DE KMEANS

MESURE DE FORME

Coefficient de silhouette



Forme des clusters



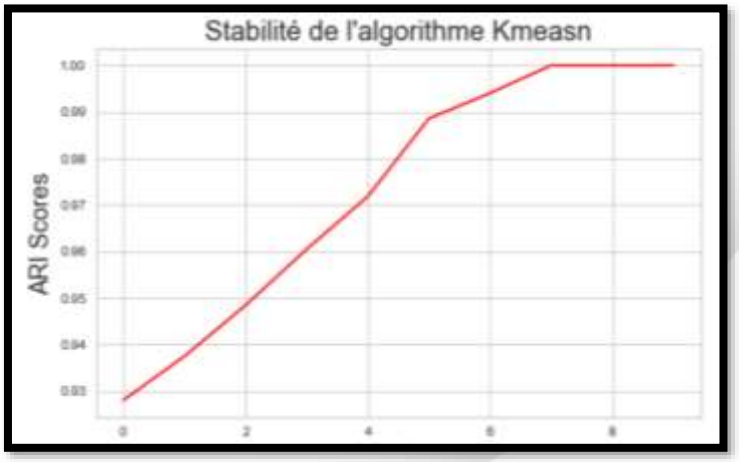
- ***SilhouetteVisualizer*** **pour visualiser coefficient de silhouette pour un échantillonage de chaque cluster.**
- **Les clusters sont relativement bien répartis et les séparations sont claires**

La visualisation en 2D avec InterclusterDistance montre que les clusters sont bien séparés

EVALUER LA STABILITÉ DE KMEANS

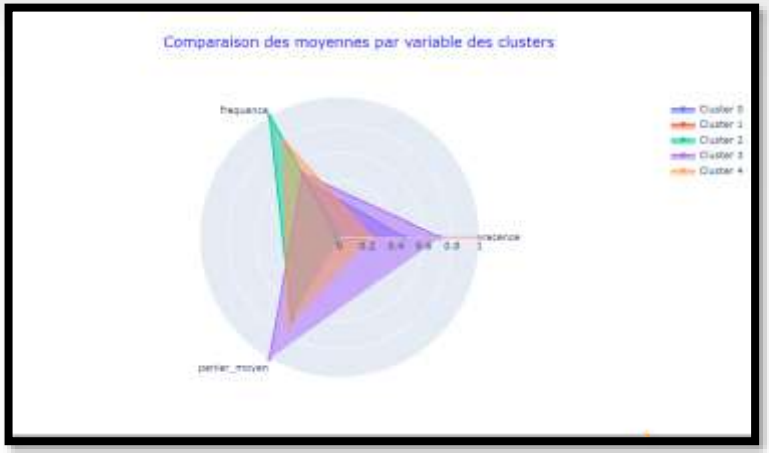
- Entrainer plusieurs fois le modèle sans fixer le RandomState
- Label initiaux: les clusters calculés dans le dernier modèle
- Comparer les ARI de chaque itération
- Les différentes itérations montrent des un score ARI proche de 1.
- On peut donc déduire que la stabilité à l'initialisation du modèle K-Means est bonne.

```
iterations = 10
centroids = None
label_initieux  = model_best.labels_
labels_pred = []
ari= []

for i in range(iterations):
    kmeans = KMeans(
        max_iter=1,
        init=(centroids if centroids is not None else 'k-means++'),
        n_clusters = 5)
    kmeans.fit(rfm_kmeans)
    centroids = kmeans.cluster_centers_

    pred = kmeans.predict(rfm_kmeans)
    labels_pred.append(pred)
    ari.append(adjusted_rand_score(model_best.labels_, labels_pred[i]))
```



IMPORTANCE DE CHAQUE FEAUTURE

Le radar chart montre que le clustering s'est basé principalement sur la variable récence



Valeur moyenne de R-F-M par clusters

	kmeans_label	recence_moy	frequence_moy	panier_moyen_moy
0	0	261.642	1.031	200.465
1	1	511.327	1.023	200.229
2	2	48.779	1.040	205.797
3	3	379.899	1.032	208.437
4	4	153.436	1.036	206.248

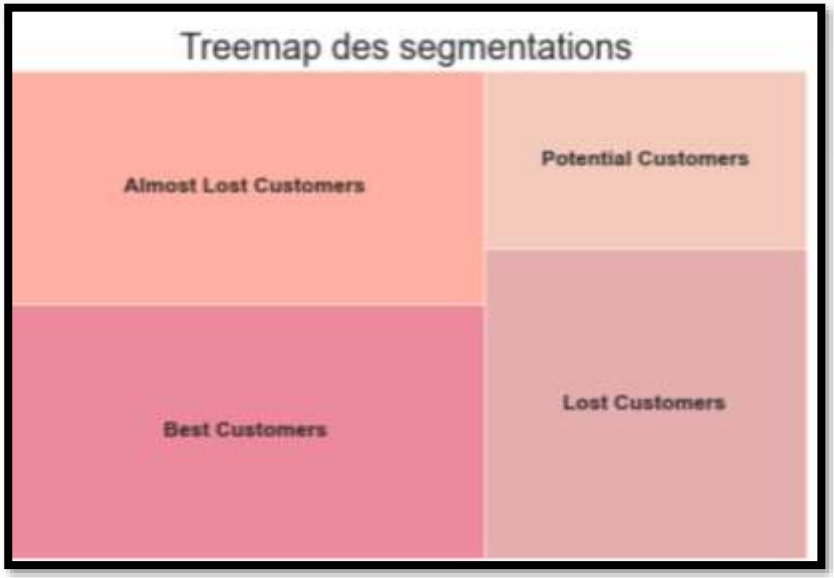
LABÉLISATION DES CLUSTERS

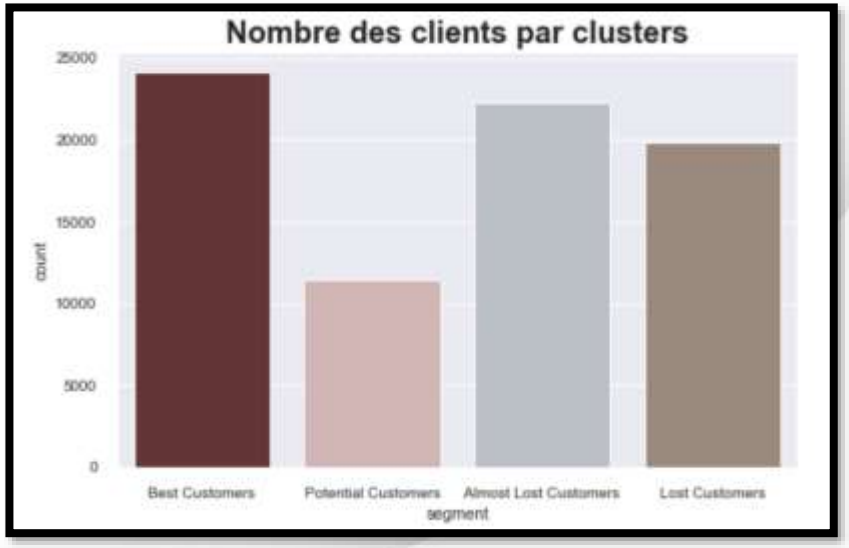
Après avoir identifier les composantes métier de chaque cluster, on peut regrouper les client en 4 catégorie :

- ***Best Customers***
- ***Potential Customers***
- ***Almost lost Customers***
- ***Lost Customers***

Label	Description
Best Customers	**Ils ont acheté le plus récemment, et qui dépensent le plus**
Potential Customers	**Ils n'ont pas acheté récemment mais qui dépensent beaucoup. Ils sont les client à fidéliser**
Almost Lost Customers	**Ils n'ont pas acheté depuis un certain temps mais qui ont dépensé beaucoup**
Lost Customers	**Ils n'ont pas acheté depuis un certain et n'ont pas dépensé beaucoup**

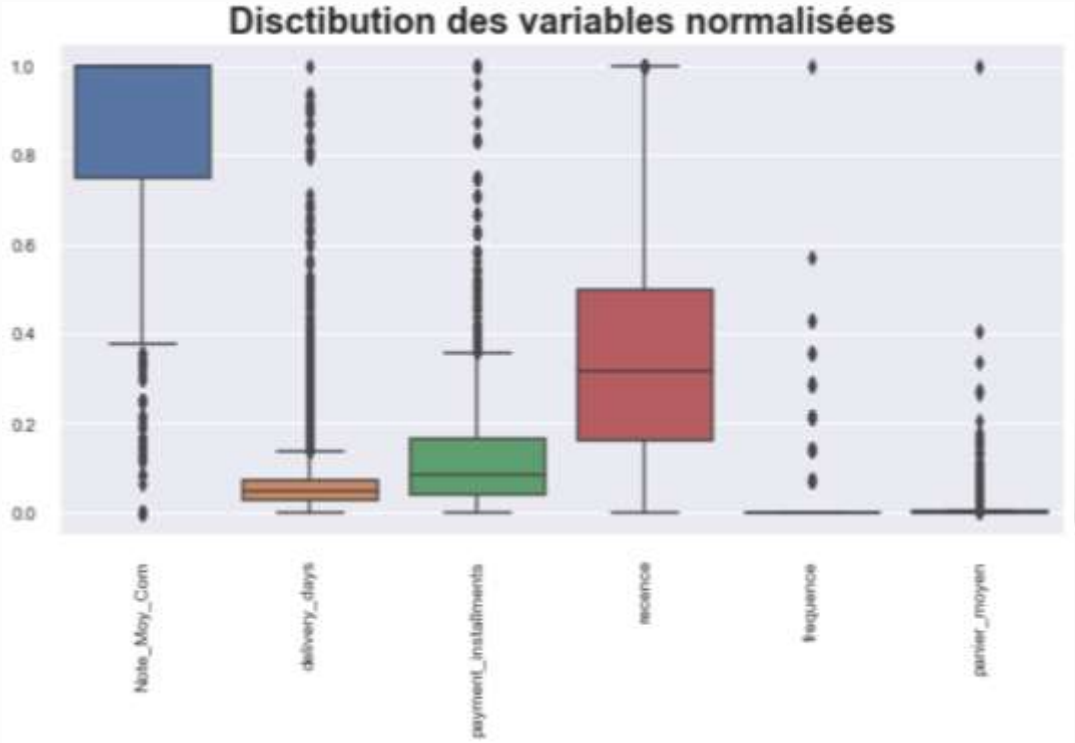
VISULISATION DES CLUSTERS





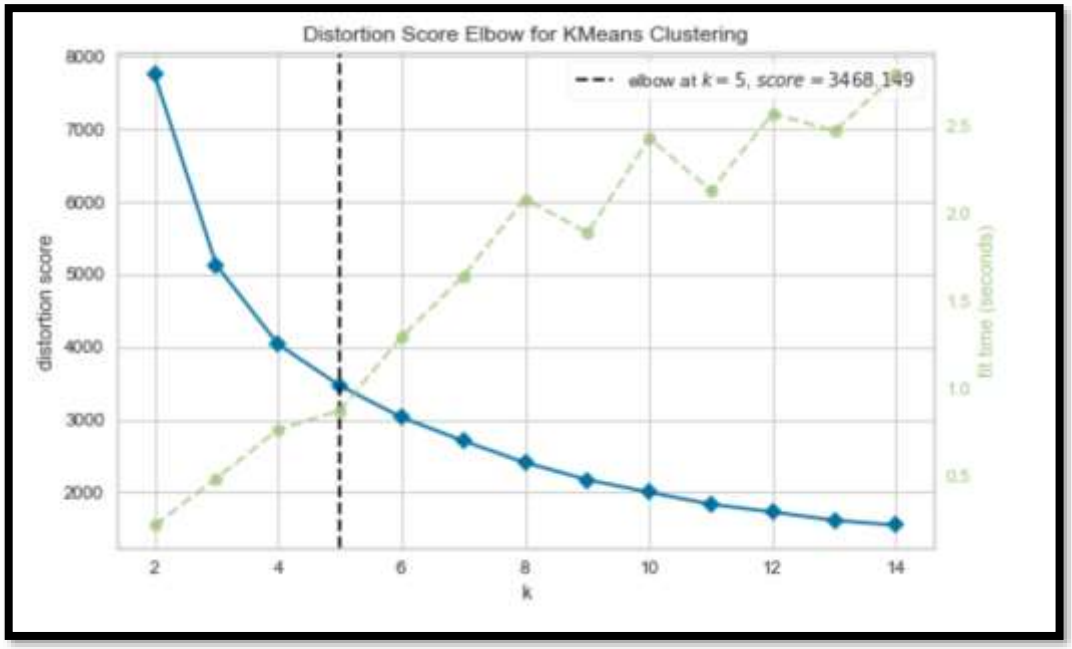
MODELISATION SUR PLUS DE VARIABLES

- **Choisir plusieurs variables pertinentes pour améliorer le modèle**



NOMBRE OPTIMAL DE K

Méthode du coude : déterminer le meilleur K

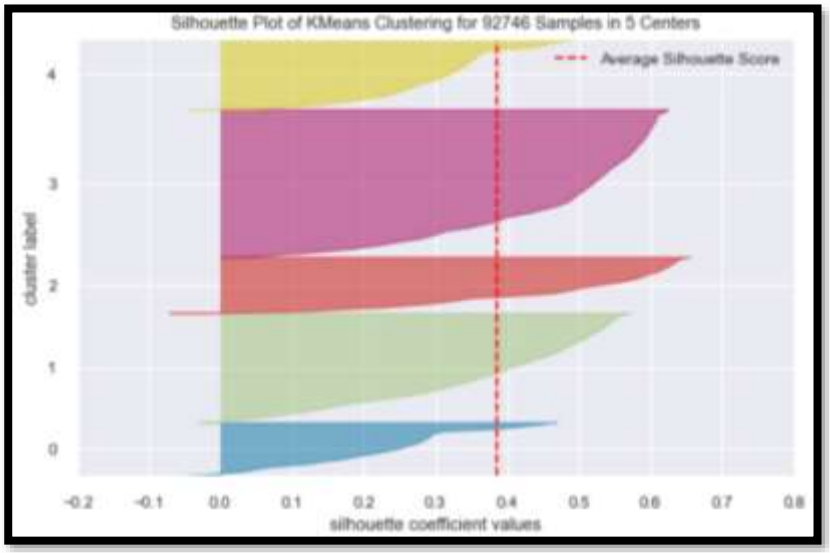


Grâce à la méthode du coude un clustering en K=5 clusters serait la meilleure option.

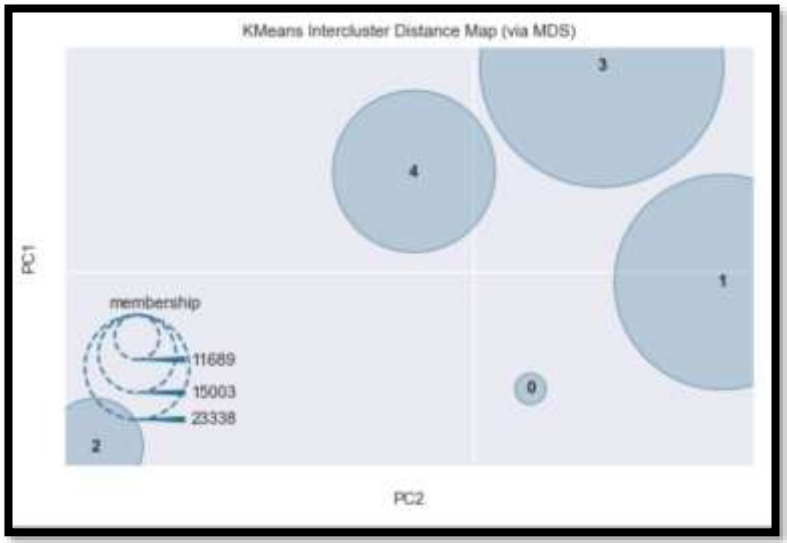
EVALUER LA PERFORMACE DE KMEANS

MESURE DE FORME

Coefficient de silhouette



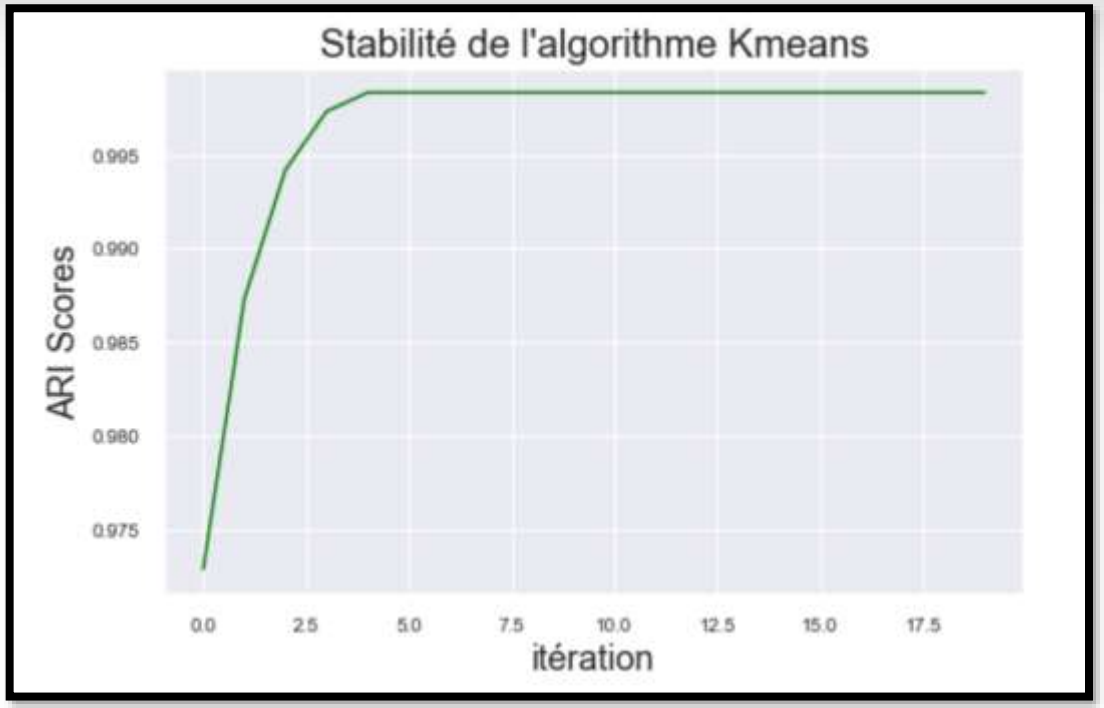
Forme des clusters



- *La* visualisation de coefficient de silhouette pour un échantillonage de chaque cluster montre que les
- Les clusters sont relativement bien répartis, avec cependant quelques erreurs

La visualisation en 2D avec *InterclusterDistance* montre que les clusters sont bien séparés

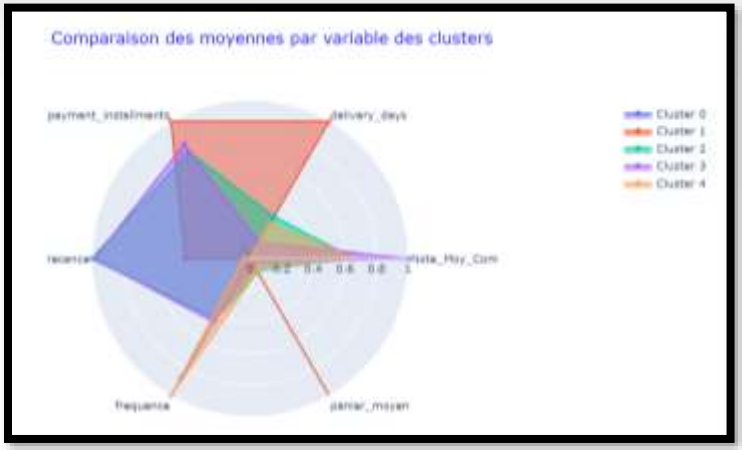
EVALUER LA STABILITÉ DE KMEANS



- **Les différentes itérations montrent un score ARI proche de 1.**
- **On peut donc déduire que la stabilité à l'initialisation du modèle K-Means est bonne.**

IMPORTANCE DE CHAQUE FEAUTURE

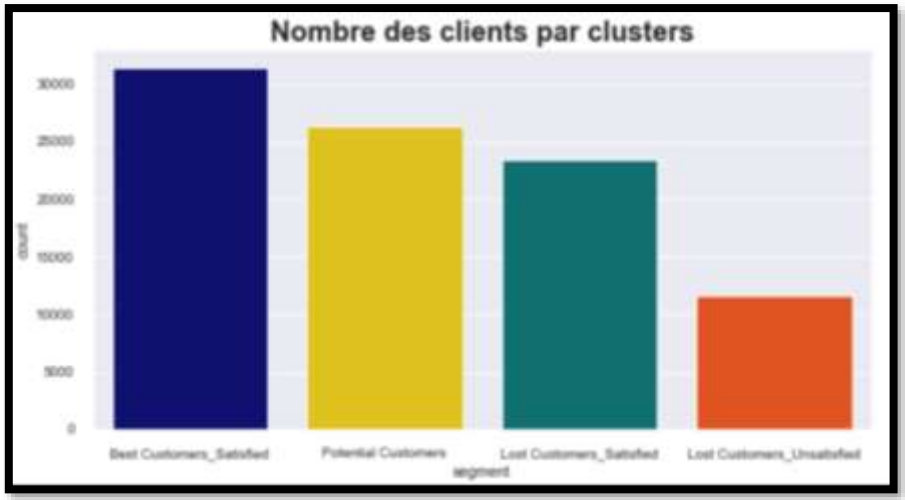
- Le radar chart montre que le clustering s'est basé principalement sur la variable « note moyenne »
- Le regroupement des clients est fortement liée à leur satisfaction

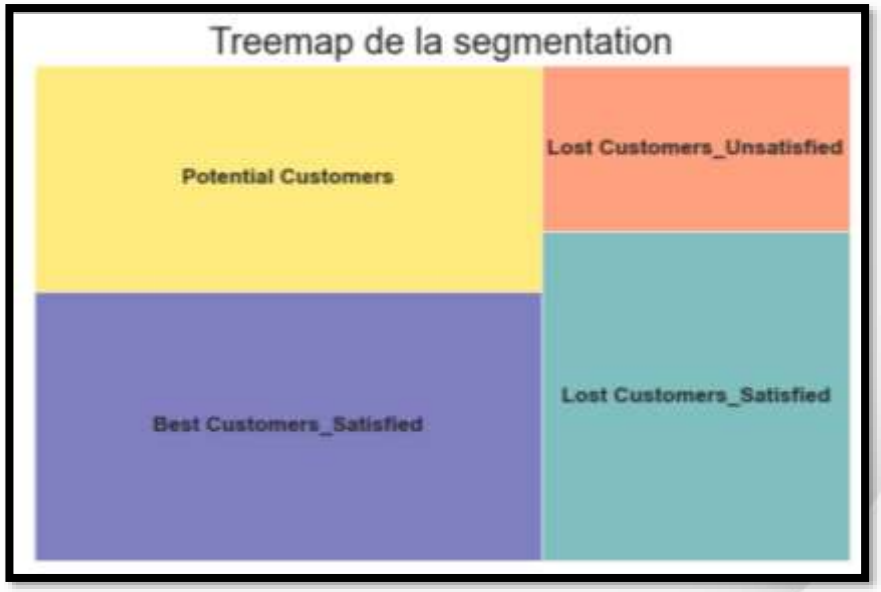


Valeur moyenne des variables par clusters

labels	score_mean	deliveray_days_mean	payment_installments_mean	recence_mean	frequence_mean	panier_moyen_mean
0	3.703	12.194	2.687	133.753	1.040	194.832
1	4.998	10.872	3.075	387.889	1.030	189.823
2	1.237	19.854	3.152	233.269	1.021	310.169
3	4.997	9.724	2.720	122.395	1.038	183.409
4	3.676	12.782	3.048	392.942	1.030	196.009

LABÉLISATION ET VISUALISATION

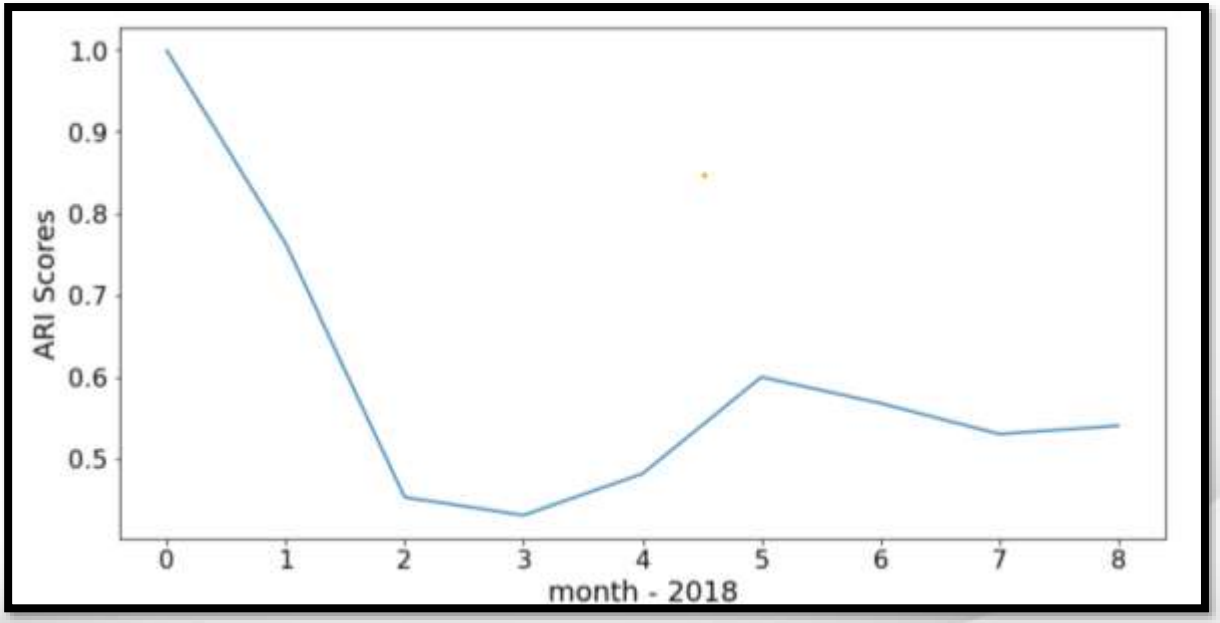




CONTRAT DE MAINTENANCE

- Tester la stabilité temporelle de l'algorithme de segmentation client
- À quel moment les clients changent les clusters ?
- Pour cela il faut itérer le K-Means sur toute la période et calculer le score ARI par rapport avec la période initiale
- Une forte inflexion après 2 mois

Stabilité temporelle de la segmentation par K-Means



CONCLUSION

- La difficulté de la segmentation RFM est liée à la variance faible de ‘F’ (la majorité des clients ont effectué un seul achat)
- Les méthode non-suppervisé nous permettent également de regrouper les clients

- Ces méthode pose des problèmes d’interprétation
- Regroupement des clients avec la méthode non-suppervisé est fortement liées leur satisfaction des clients

Recommandation et perspective

i. **Prévoir la maintenance du programme de segmentation tous les 2 ou 3 mois**
ii. **Re-tester cette stabilité temporelle au fil du temps. Il sera donc nécessaire de redéfinir les segments clients à chaque maintenance.**
iii. **Réaliser une analyse sur les données plus réelles (fréquence plus représentative) pour une meilleure visualisation des difficultés en entreprise**

MERCI DE VOTRE ATTENTION !