Projet 4:

Segmentez des clients d'un site e-commerce

OPENCLASSROOMS

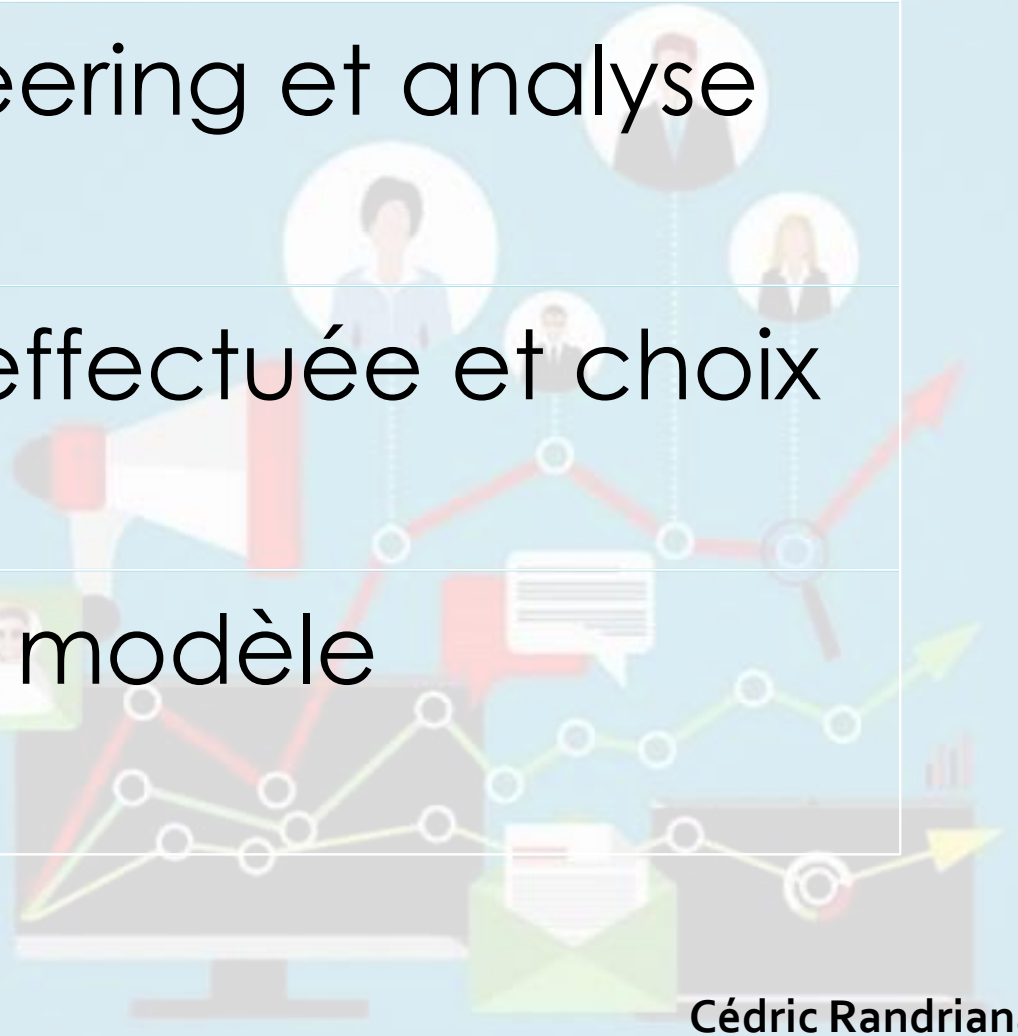
Cédric Randrianarivélo

I. Problématique et Jeu de données

II. Feature Engineering et analyse exploratoire

III. Modélisation effectuée et choix du modèle

IV. Simulation du modèle



I. Problématique et Jeu de données

OPENCLASSROOMS

Contexte

- **Aider** l'entreprise brésilienne « Olist » qui propose une solution de vente sur les marketplaces en ligne.
- **Fournir** à ses équipes une segmentation des clients réutilisable lors de leur campagne de communication.
- **Comprendre** les différents types de clients et étudier leurs comportements par le biais de leurs données.
- **Fournir** une description détaillée de la segmentation et de la logique de segmentation.
- **Proposer** un action de maintenance qui se base sur la stabilité des modèles de segmentation.

Jeu de données

- Une base de données **anonymisée** fournit par « Olist »
- **Informations** sur l'historique de commandes, les produits achetés, les commentaires de satisfaction, et la localisation des clients
- Données depuis **janvier 2017**.

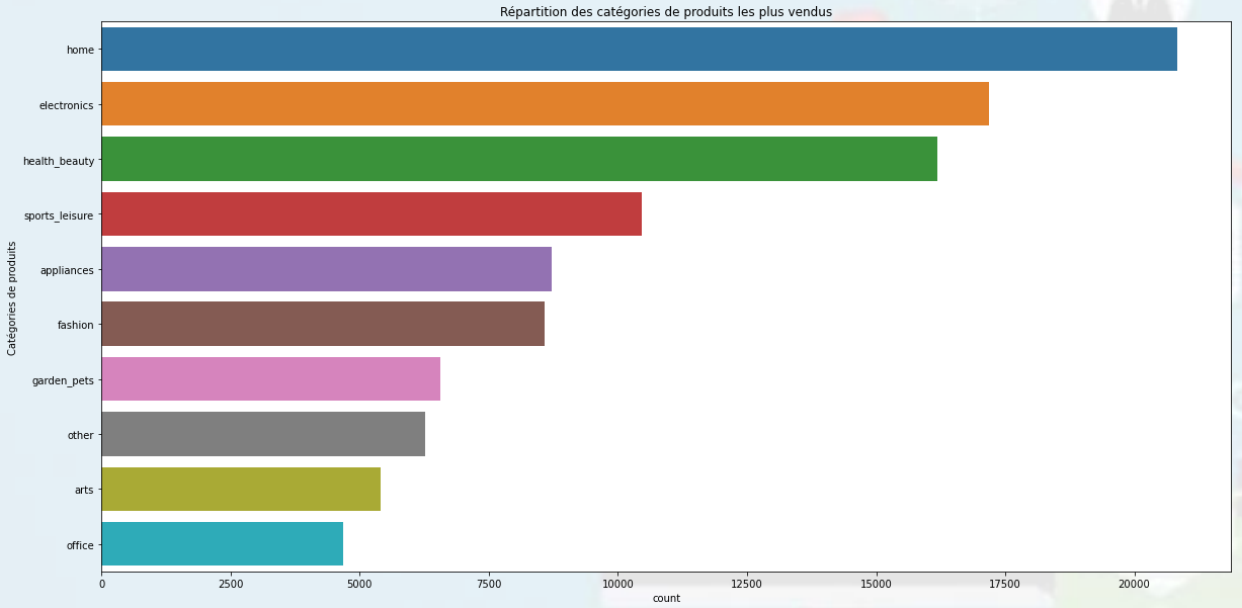
	% du dataset complet	Nombre de valeurs manquantes
order_id	100.00	0
customer_unique_id	100.00	0
payment_value	100.00	1
payment_installments	100.00	1
payment_sequential	100.00	1
customer_id	100.00	0
customer_city	100.00	0
customer_zip_code_prefix	100.00	0
customer_state	100.00	0
order_estimated_delivery_date	100.00	0
order_purchase_timestamp	100.00	0
order_status	100.00	0
geolocation_city	99.93	66
geolocation_zip_code_prefix	99.93	66
geolocation_lat	99.93	66
geolocation_lng	99.93	66
order_approved_at	99.84	160
review_score	99.23	768
price	99.22	775
freight_value	99.22	775
order_delivered_carrier_date	98.21	1783
order_delivered_customer_date	97.02	2965

II. Feature Engineering et analyse exploratoire

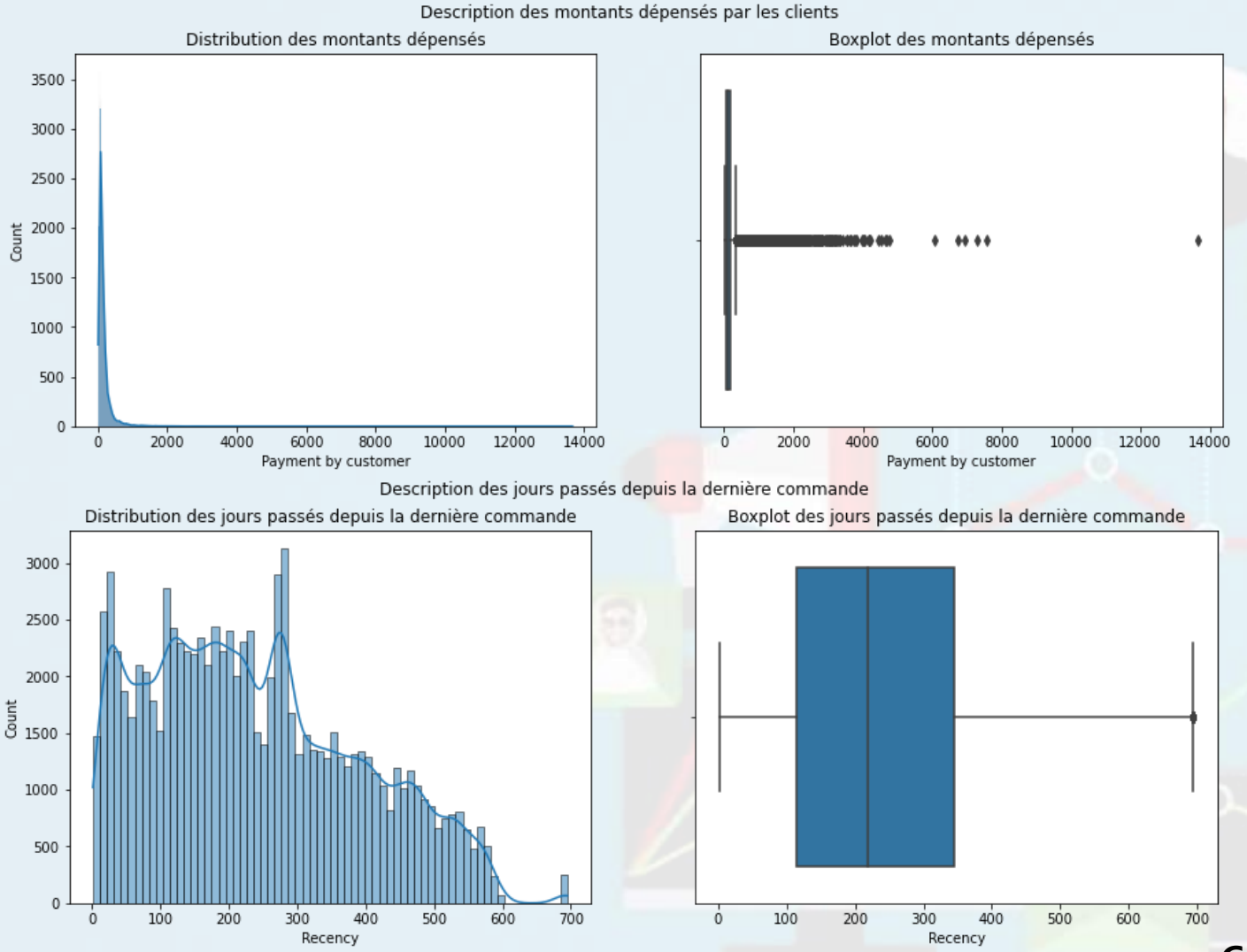
OPENCLASSROOMS

Feature Engineering – Ajout Features

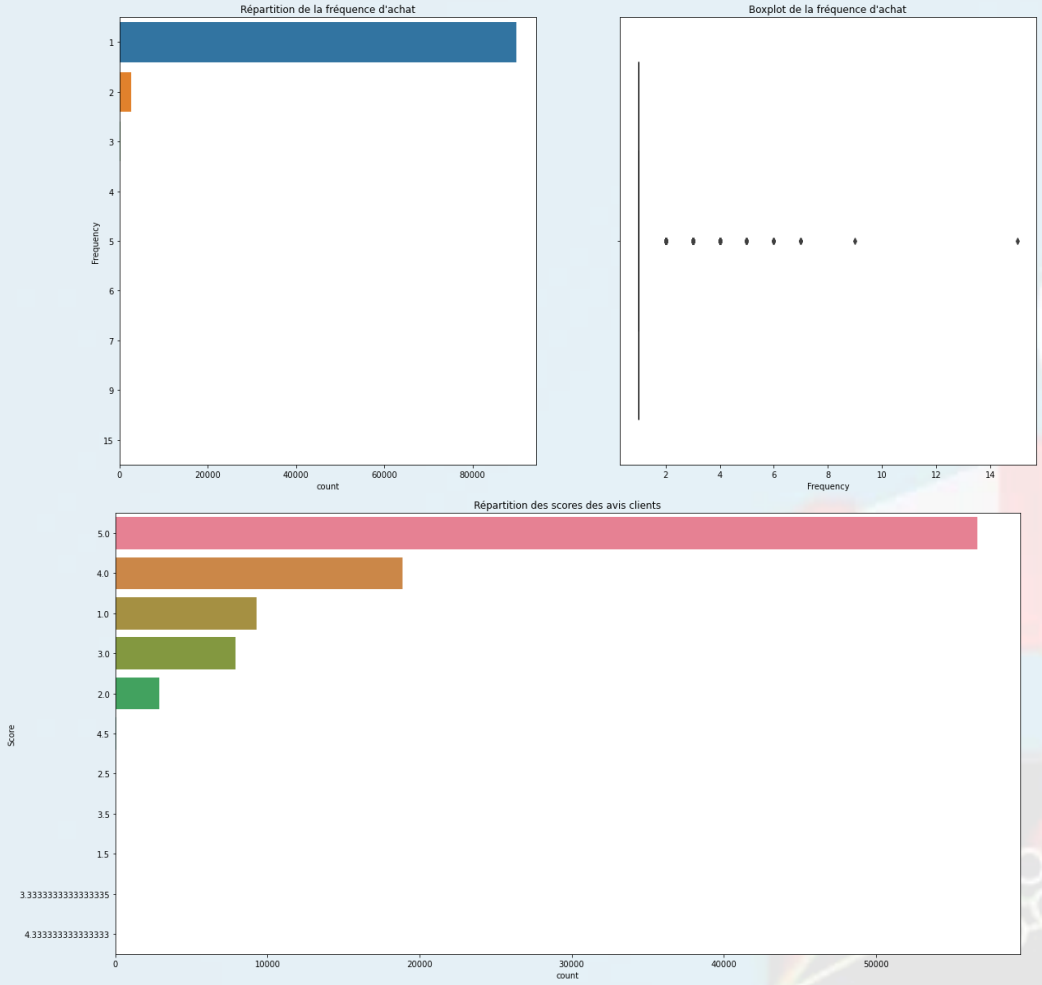
Récence	Fréquence	Montants	Nombres de jours avant la livraison	Nouvelles catégories
• Durée depuis la dernière commande	• Nombre de commandes	• Montant cumulé des commandes	• Nombre de jours avant la livraison de la commande	• Regroupement des produits par catégorie



Analyse exploratoire



Analyse exploratoire



III. Modélisation effectuée et choix du modèle

OPENCLASSROOMS

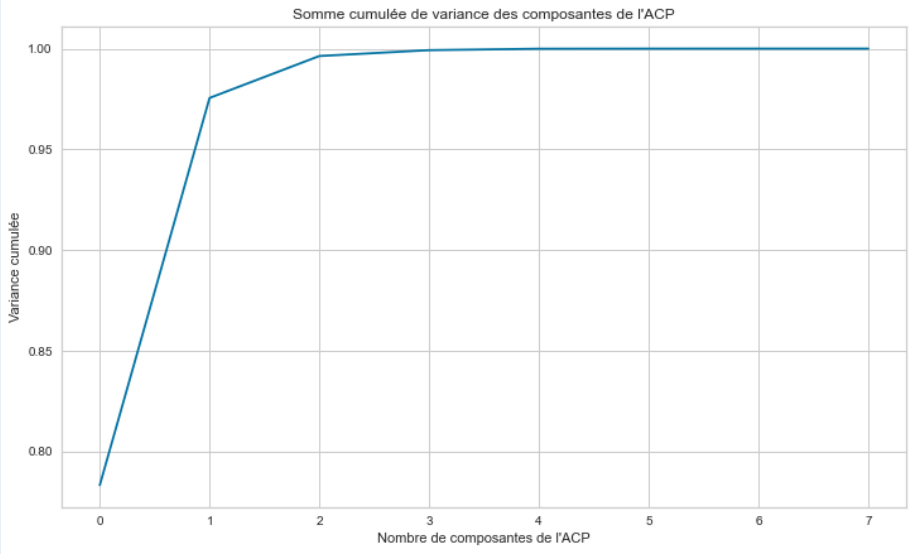
Test des modèles

Modèles utilisés:

- K-means
- DBSCAN
- Agglomerative

Composition testée

- Avec 3, 4 et 5 features.
- 4 et 5 clusters
- ACP pour déterminer le nombre optimal de composantes



K-Means Choix du nombre de clusters et de features

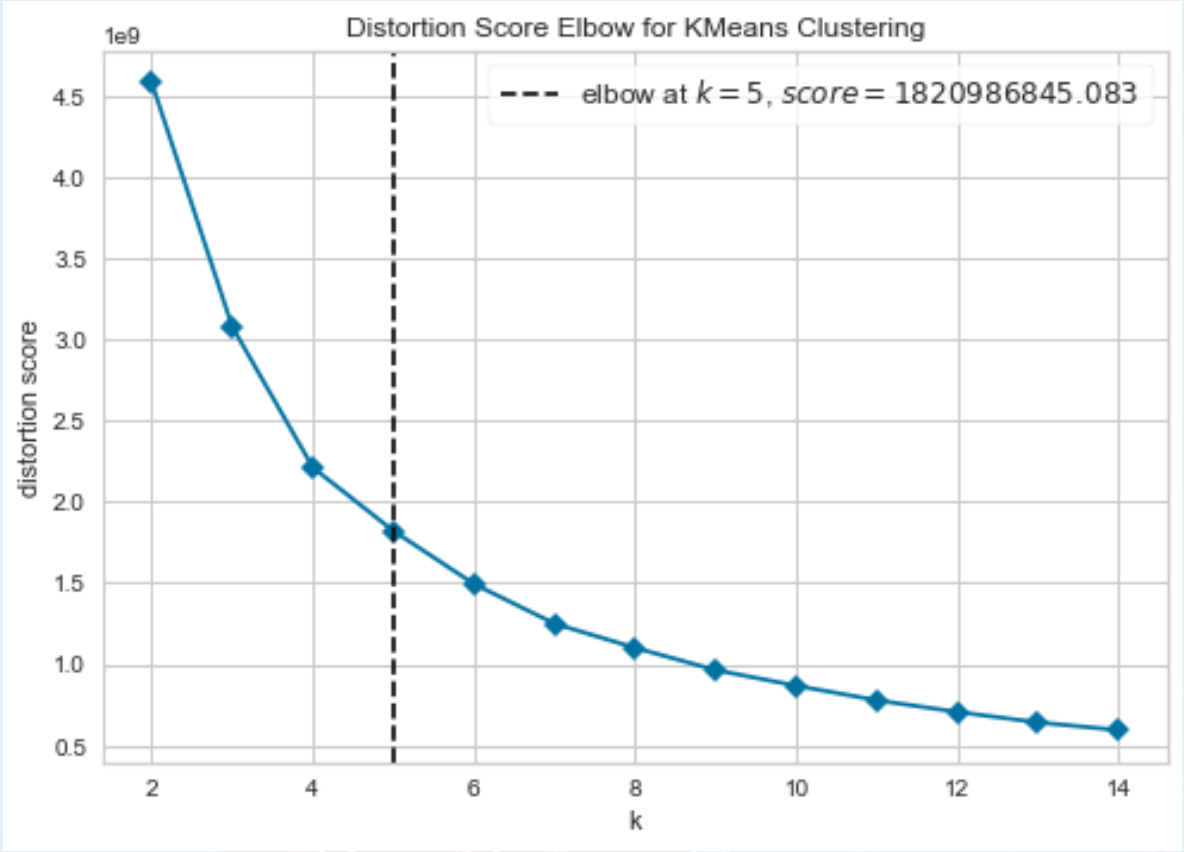
On détermine le nombre de clusters théorique :

- Méthode du coude : 5 clusters

En pratique :

- Une Classe à 300 clients
- Classe avec trop peu de clients

Théoriquement, on décide de partir sur 4 clusters ! On valide nos résultats graphiquement.



K-Means – Résultats

Compositions essayées :

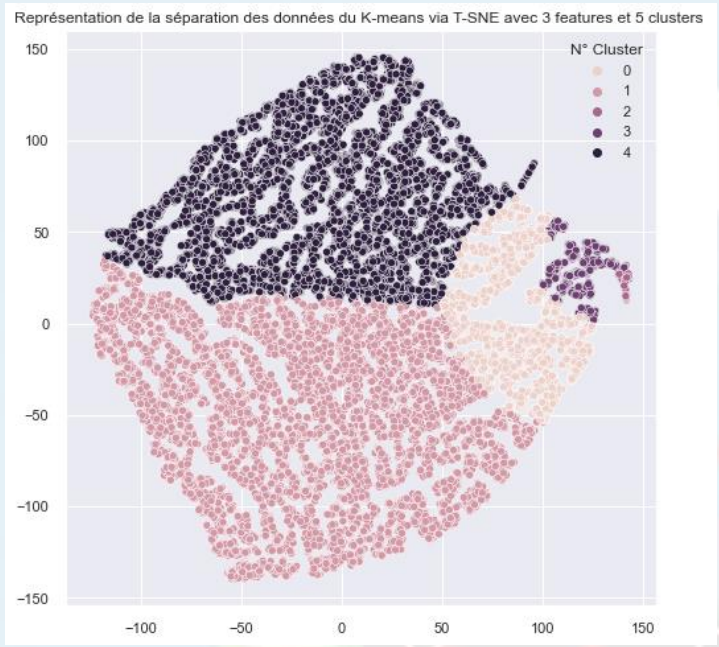
- 3 features et 5 clusters
- 3 features et 4 clusters

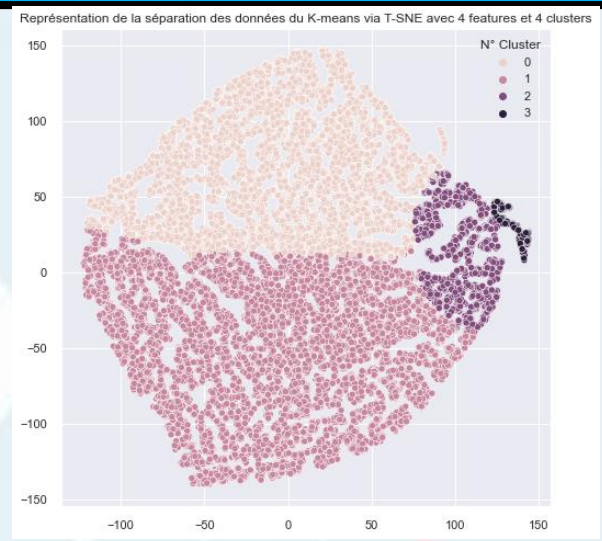
On confirme graphiquement nos résultats obtenus et on part sur 4 clusters

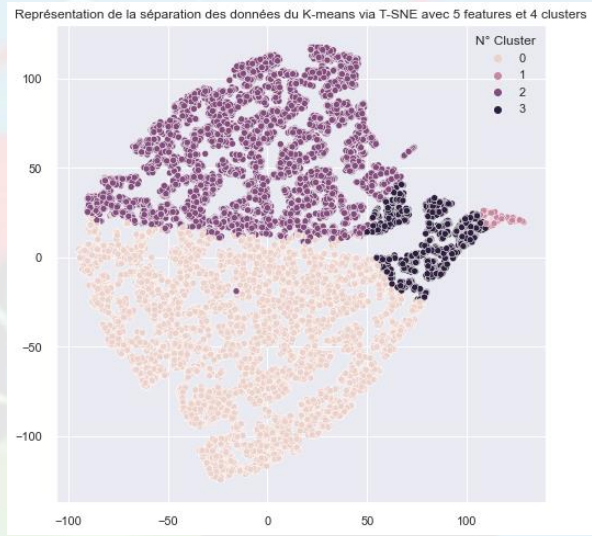
Compositions essayées :

- 4 features et 4 clusters
- 5 features et 4 clusters

Modèle à 4 features et 4 clusters retenus. Car « le review score » et « le nombre de jours » avant livraison sont très corrélés. Et on souhaite avoir autant de clusters que de features.



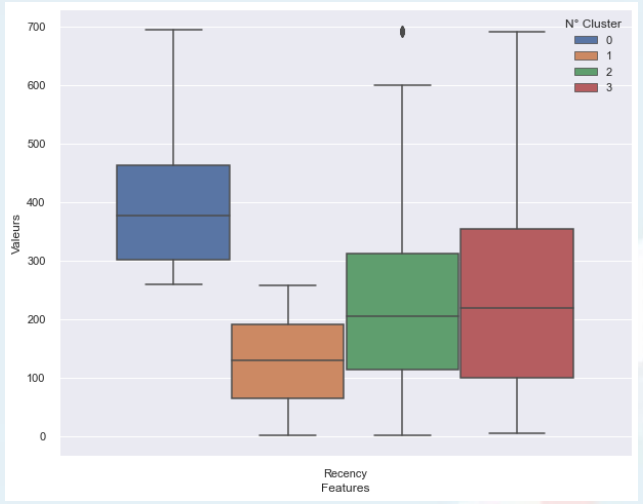


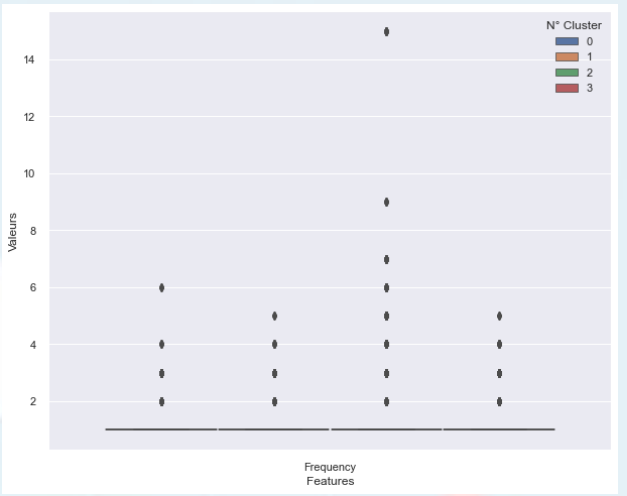


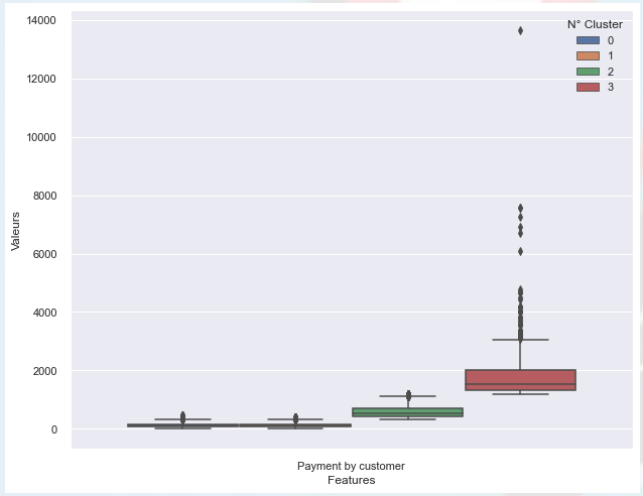
K-Means – Résultats

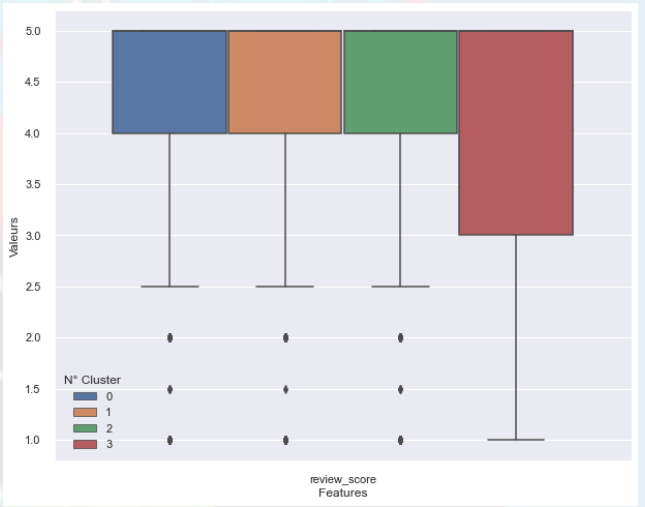
Résultats 4 features et K = 4 :

- *Cluster 0* : Forte récence et faible paiement cumulé
- *Cluster 1* : Faible récence faible paiement cumulé
- *Cluster 2* : Fréquence la plus haute
- *Cluster 3* : Paiement cumulé le plus élevé et « review score » le plus dispersé.



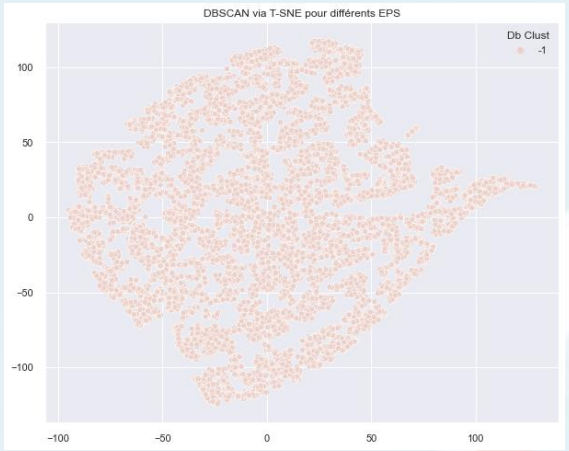


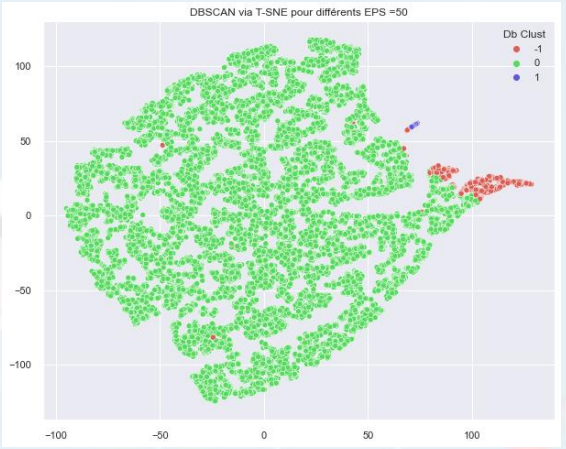


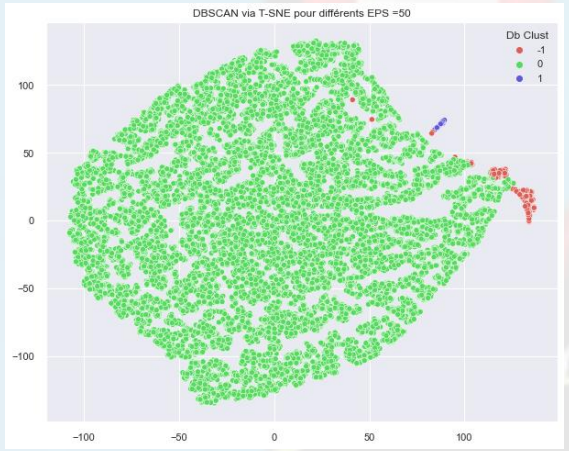


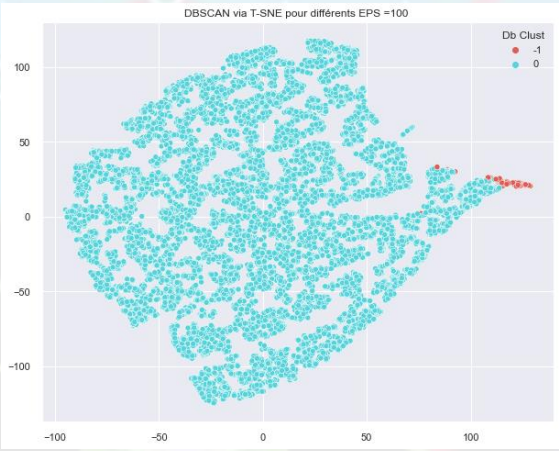
DBSCAN – Plusieurs Epsilon

- Test Epsilon = 1, 50, 80 et 80
- Une classe représentant plus **97%** des données en général
- Se base sur la densité
- Non pertinent







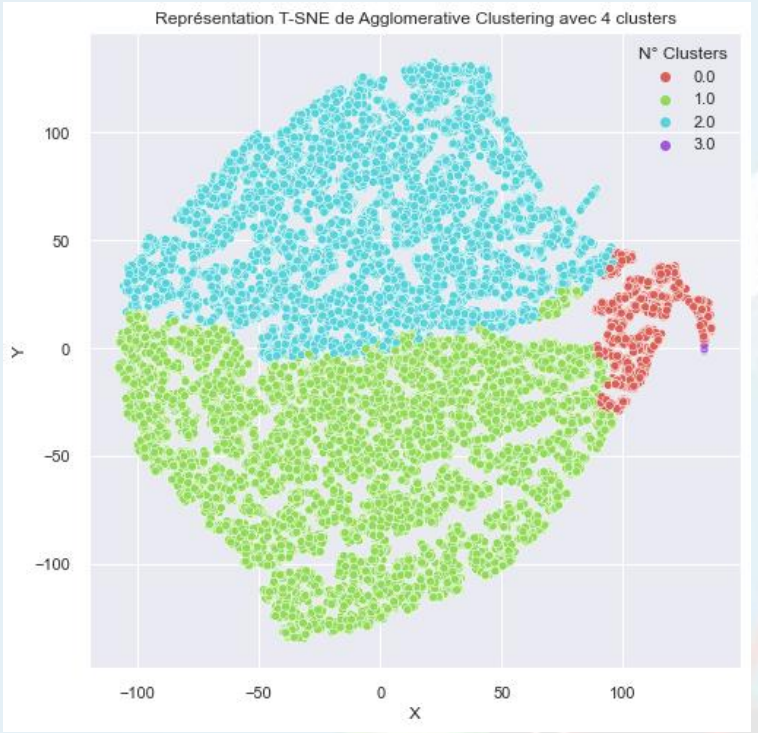


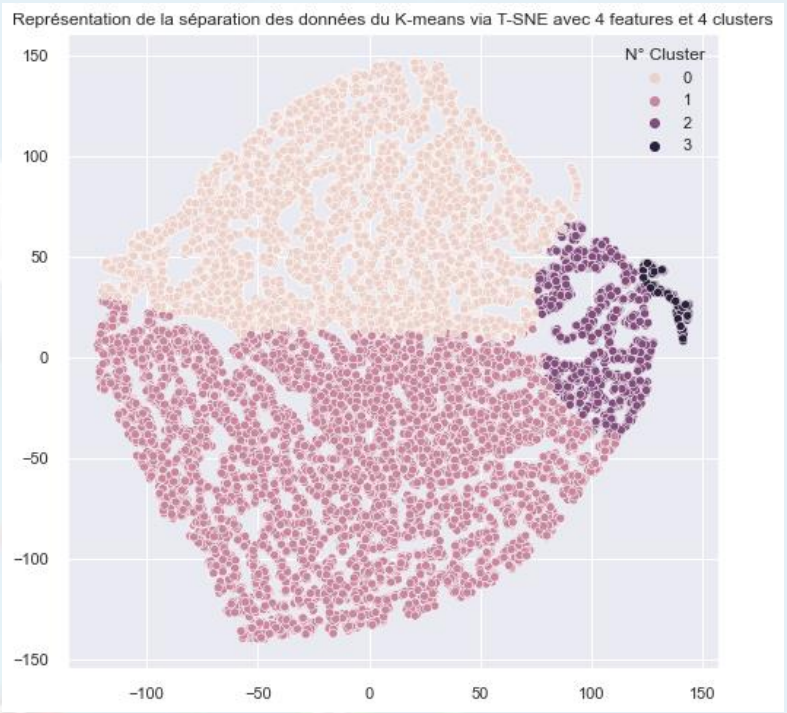
OPENCLASSROOMS

AgglomerativeClustering – Résultats

	K-means	AgglomerativeClustering
Temps fit	>1 s	132.5 s
Nombre éléments	100 000	40 000

- Résultats Similaires au K-Means
- Temps de fitting Long

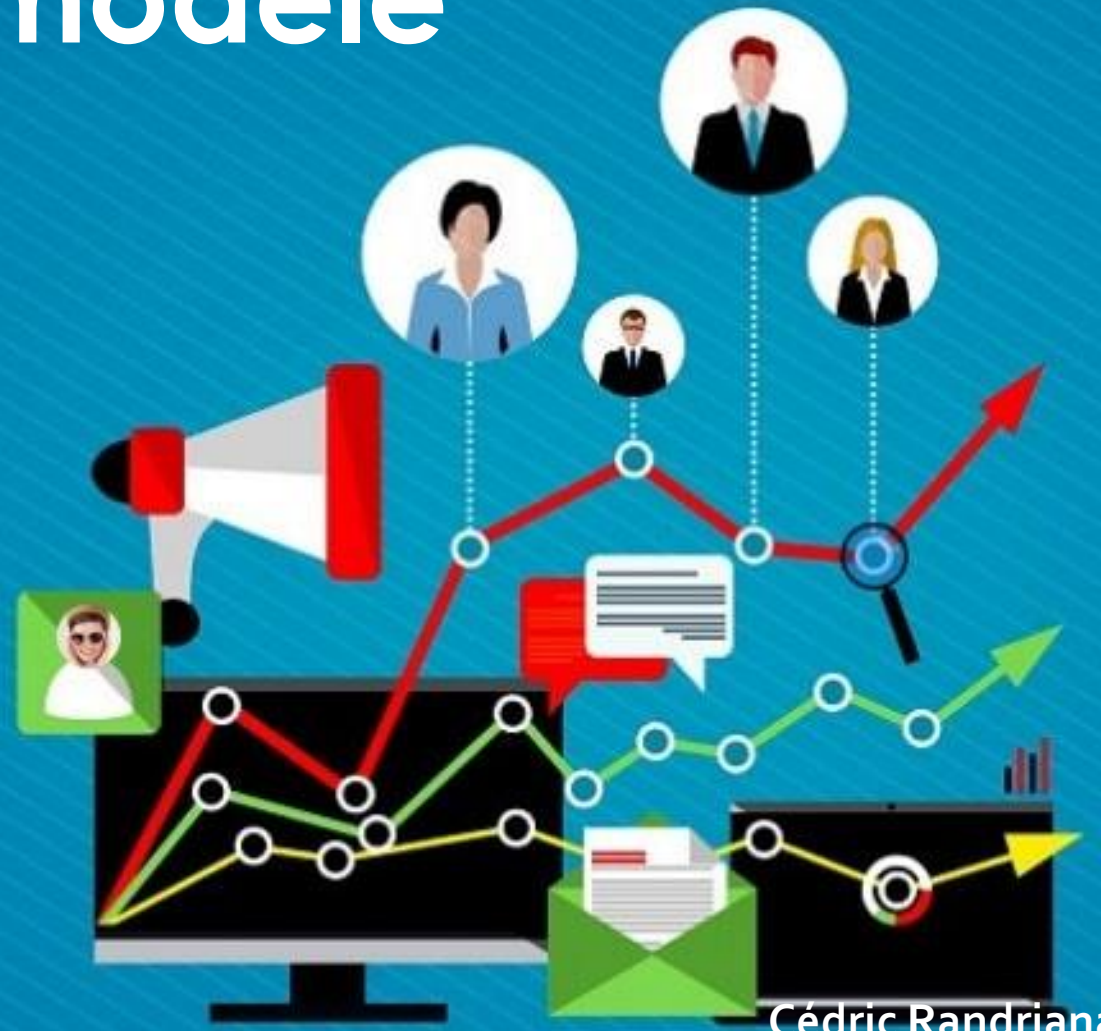




Paramètres

Modèle Choisi :	K-means
	Temps faible de fitting
	Regroupe les clients par leur caractéristiques et non densité
Paramètres :	4 clusters
	4 features (Review Score, Payment by customers, Recency et Frequency)

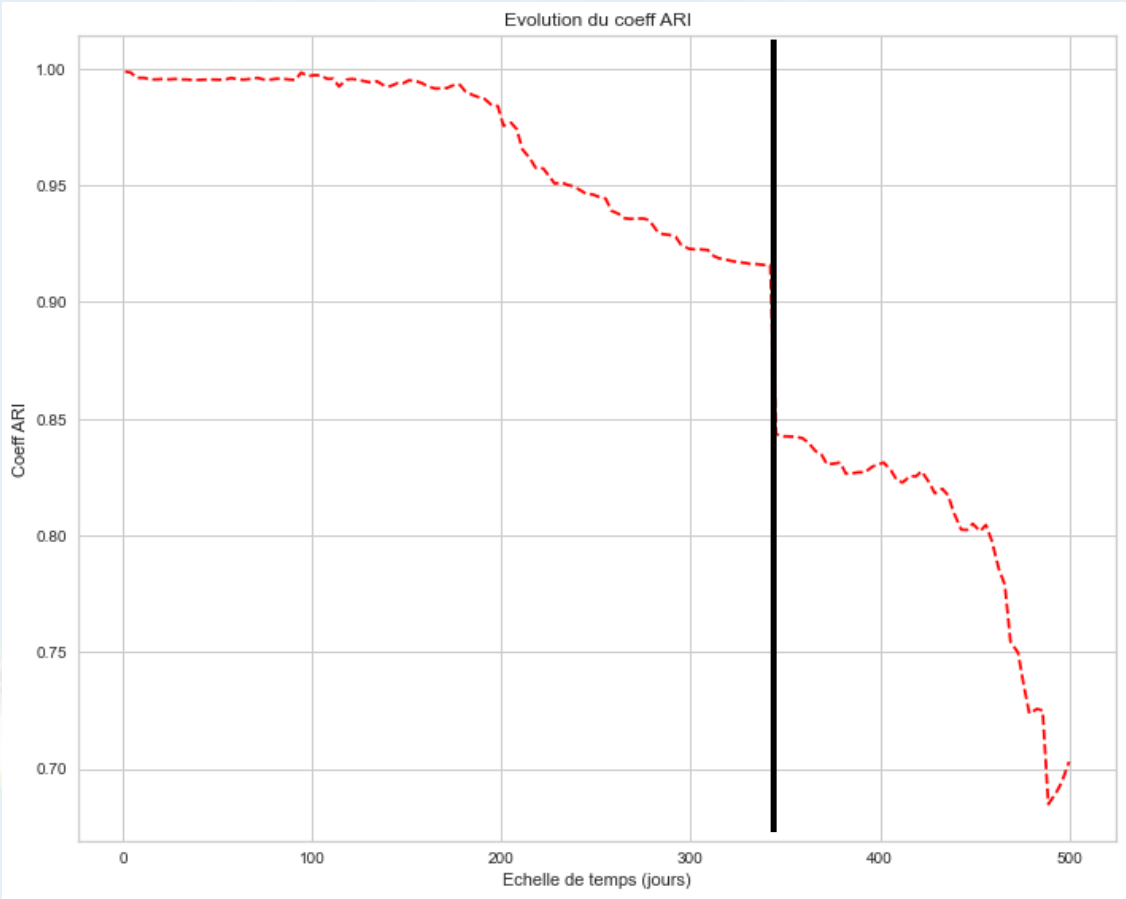
IV. Simulation du modèle



OPENCLASSROOMS

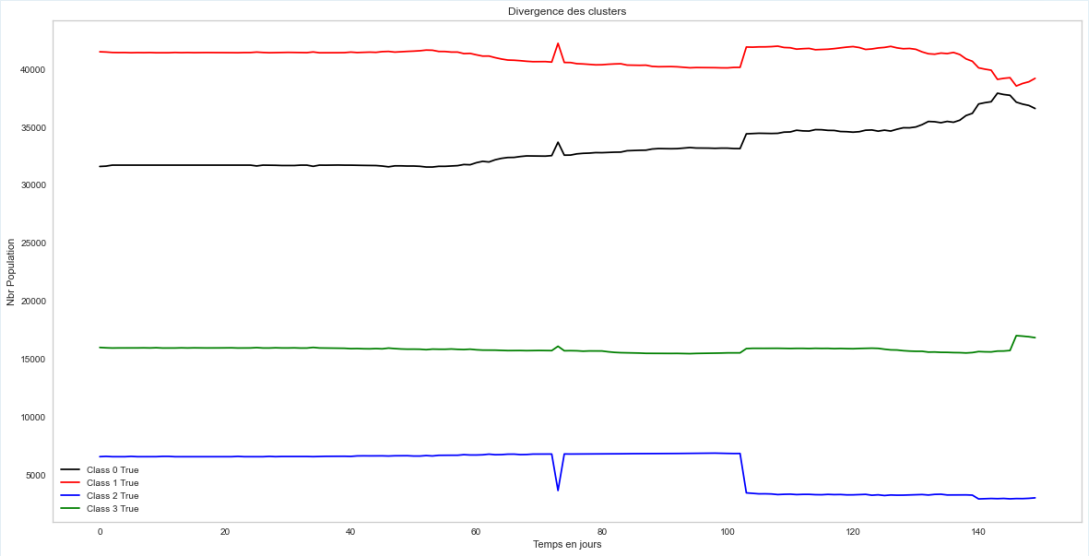
Stabilité

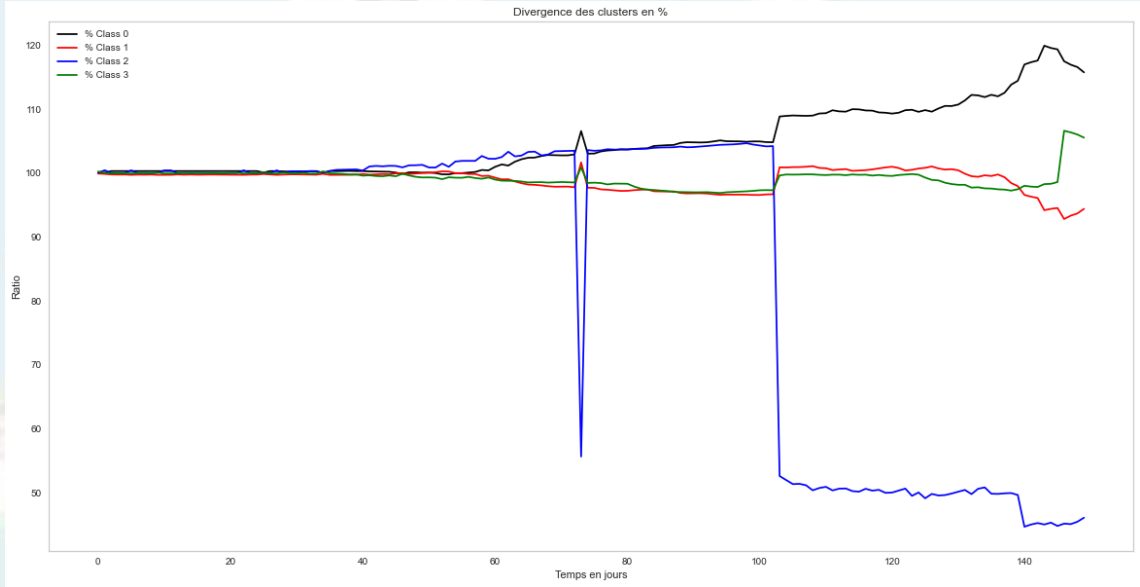
- 342 jours, le coeff ARI passe en dessous de 0,9.
- Il faut partir sur une mise à jours de modèle environ tous les 340 jours.



Divergence Clusters

- Le cluster avec le plus faible nombre de clients diminue le plus (cluster les clients avec les clients qui dépense le plus)
- Le cluster avec les paiements cumulés faible augmente





Conclusion

Choix du modèles pour la segmentation

- Sélection du **modèle K-means.**
- **Temps faible** de fitting très nettement inférieur à l'agglomerative clustering
- La création de classe est **plus précise** que le DBSCAN (se basant sur la densité) qui se base sur leurs caractéristiques.
  - Regroupe les clients par leur caractéristiques et non densité.

Paramètres :

- \- 4 features (Review Score, Payment by customers, Recency et Frequency

- 4 clusters comprenant :
  - Un cluster avec une forte récence/ Faible paiement
  - Un cluster faible récence et faible paiement cumulé
  - Un cluster avec une fréquence d'achat haute
  - Un cluster avec le paiement cumulé élevé et des review scores plus mauvais.

N° Cluster	Recency	Frequency	Payment by customer	review_score
0	388.50	1.05	121.05	4.18
1	128.06	1.06	121.07	4.16
2	222.55	1.33	583.54	4.03
3	235.12	1.25	1800.82	3.96

Stabilité :

- Cluster instable au bout de 342 jours
- Contrat de maintenance tous les 340 jours

Merci pour votre attention !
Question ?

OPENCLASSROOMS