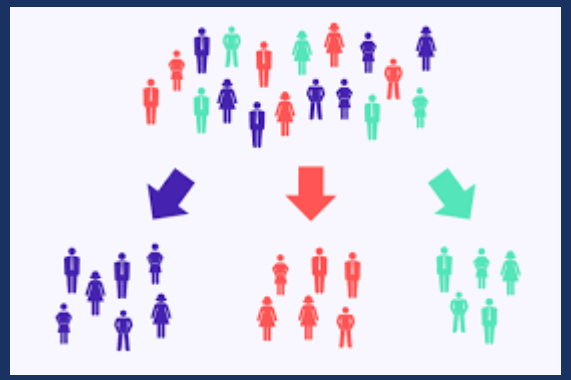
PROJET 4 : SEGMENTEZ DES CLIENTS D'UN SITE DE E-COMMERCE

SOUTENANCE OPENCLASSROOMS, LE XX/05/2022

ERWAN CHESNEAU

olist



PLAN :

I. Contexte
II. Nettoyage et exploration de la base de données
III. Tests de segmentation
IV. Interprétation de la meilleure segmentation
V. Simulation du délai de maintenance
VI. Conclusions
VII. Améliorations à envisager

I. CONTEXTE:

- L'entreprise de e-coomerce Olist souhaite mieux comprendre sa clientèle
- Volonté de segmenter la clientèle
  - Meilleur communication
  - Identification des meilleurs clients
  - Ciblage marketing
- Base de données disponible regroupant un historique des commandes pour chaque client
- Démarches :
  - Explorer les données afin de caractériser la clientèle
  - Effectuer des tests de segmentation en utilisant différents algorithmes
  - Caractériser la segmentation
  - Simuler le délai de mise à jour optimal de la segmentation

II. EXPLORATION DES DONNÉES DISPONIBLES ET NETTOYAGE

- La base de données est séparée en 8 fichiers joints par une variable : (seules les variables utilisées dans le projet sont présentées ici)

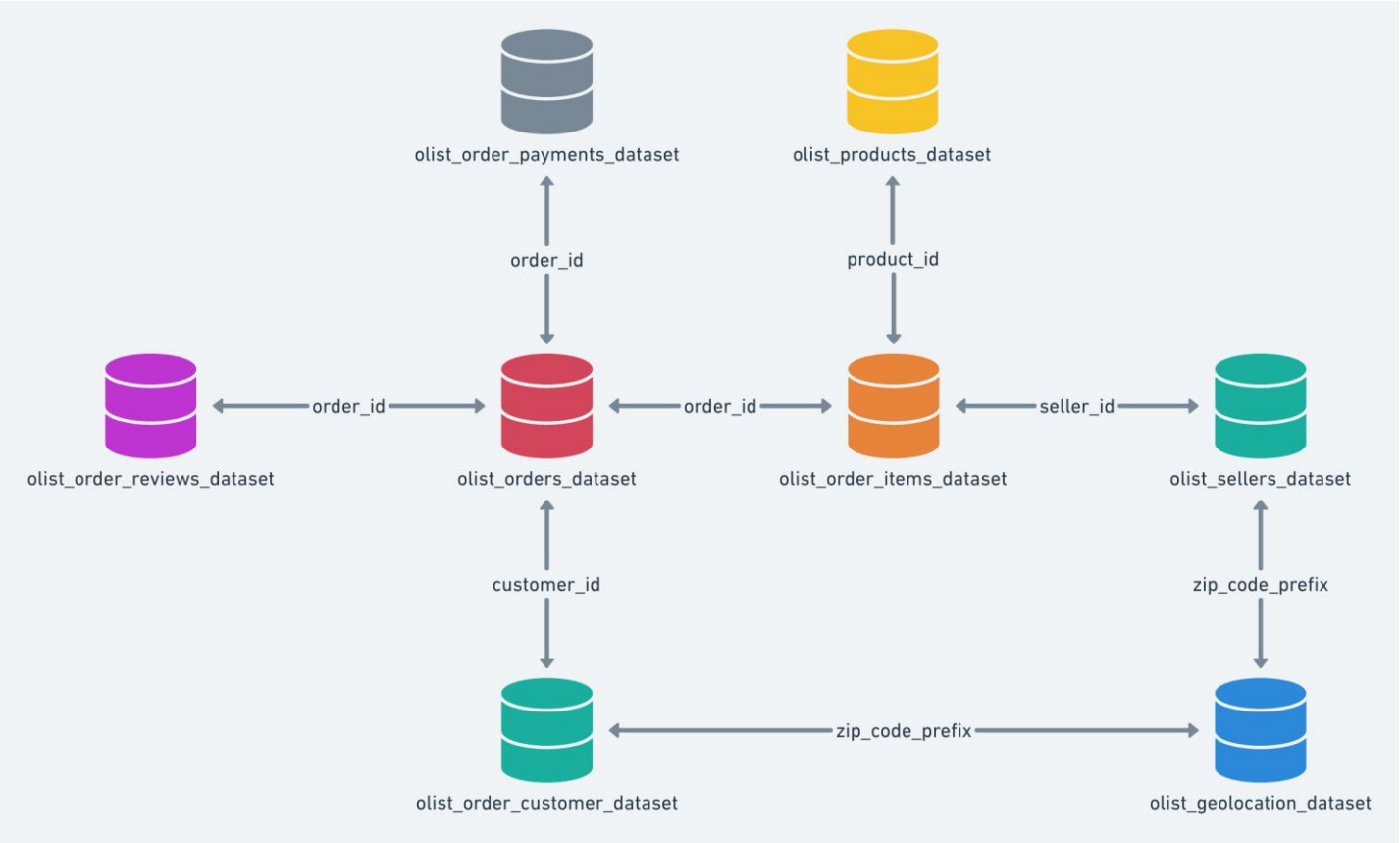
fichiers	contenu	description	nettoyage
olist_customers_dataset	customer_id	ID client / commande	/
	customer_unique_id	ID client unique	/
	customer_zip_code_prefix	Code postal client	int
olist_geolocation_dataset	geolocation_zip_code_prefix	Code postal	int
	geolocation_lat	Laittude	Float [-90, 90]
	geolocation_lng	Longitude	Float [-180, 180]

II. EXPLORATION DES DONNÉES DISPONIBLES ET NETTOYAGE

fichiers	contenu	description	nettoyage
olist_order_items_dataset	order_id	ID commande	/
	product_id	ID produit	/
	price	Prix	Float > 0
olist_order_payments_dataset	order_id	ID commande	/
olist_order_reviews_dataset	review_id	ID de la notation	/
	order_id	ID commande	/
	review_score	Note	Int [0,5]

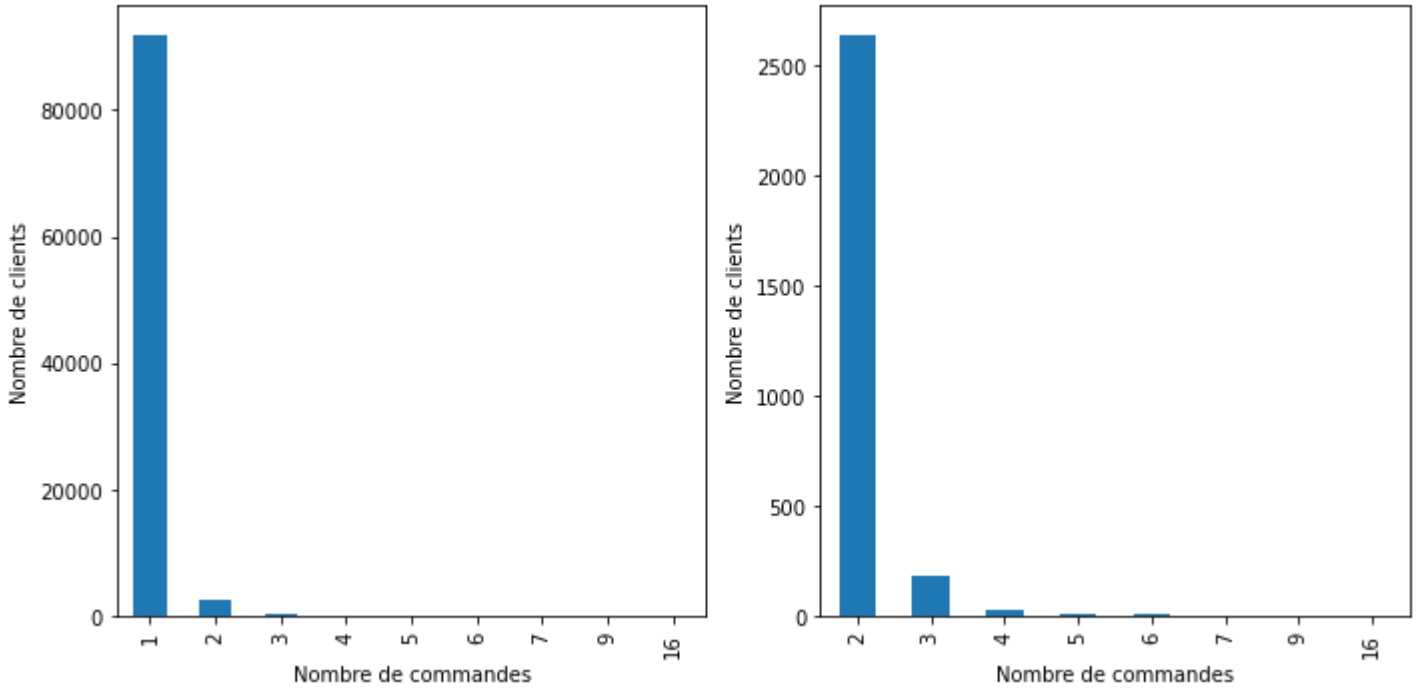
II. NETTOYAGE : REGROUPEMENT DES DONNÉES

- Détections et suppression des doublons
- Création d'une nouvelle base de données pour l'analyse :
  - Info clients :
    - id, unique id, zip code, ville, état, géolocalisation
  - Info commandes :
    - Id, date, id produit, prix
  - Info sur la satisfaction
    - order id, review score
  - Info sur les produits
    - Id, catégories



II. EXPLORATION :

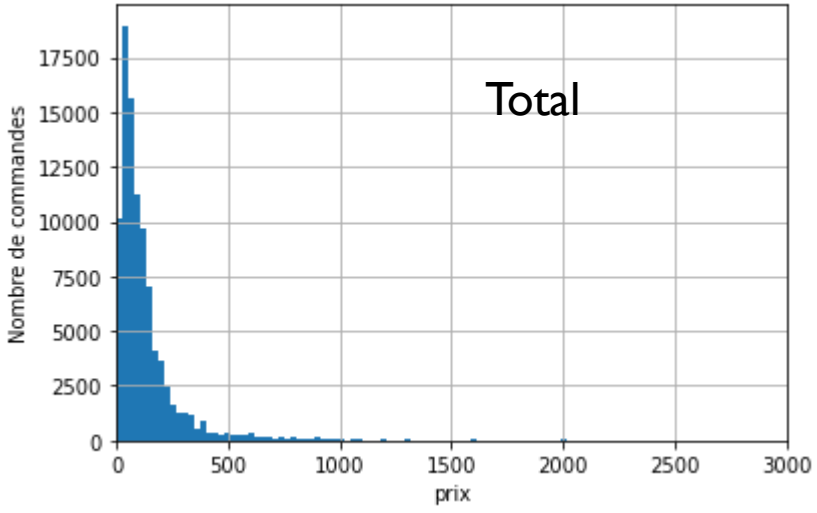
- Nombre de commandes par client :



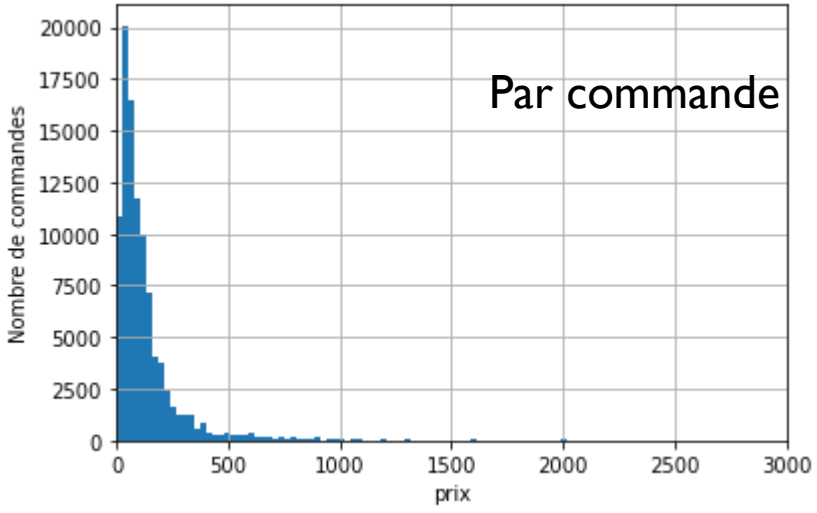
- 3% des clients ont commandé plus de 1 fois.

II. EXPLORATION :

- Montant dépensé



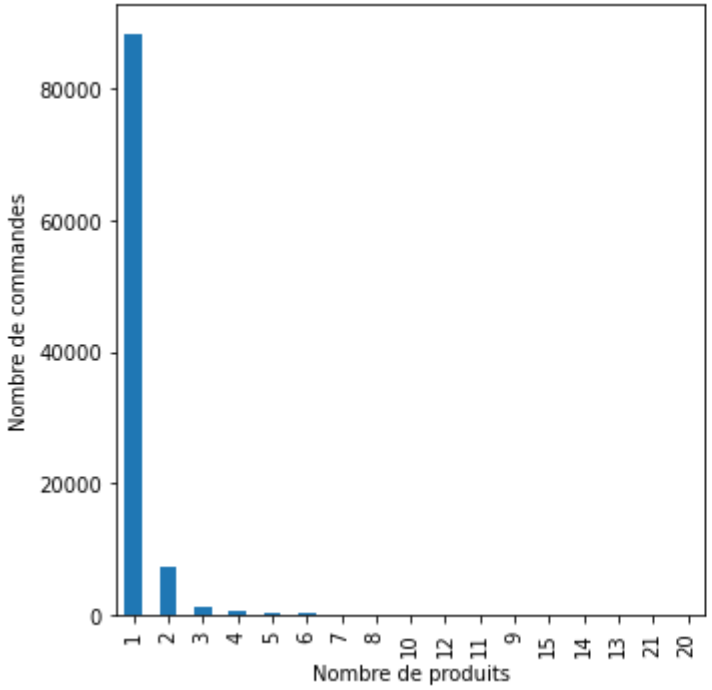
Min	0,85
Max	13440
Moyenne	142,17
Médiane	89,8

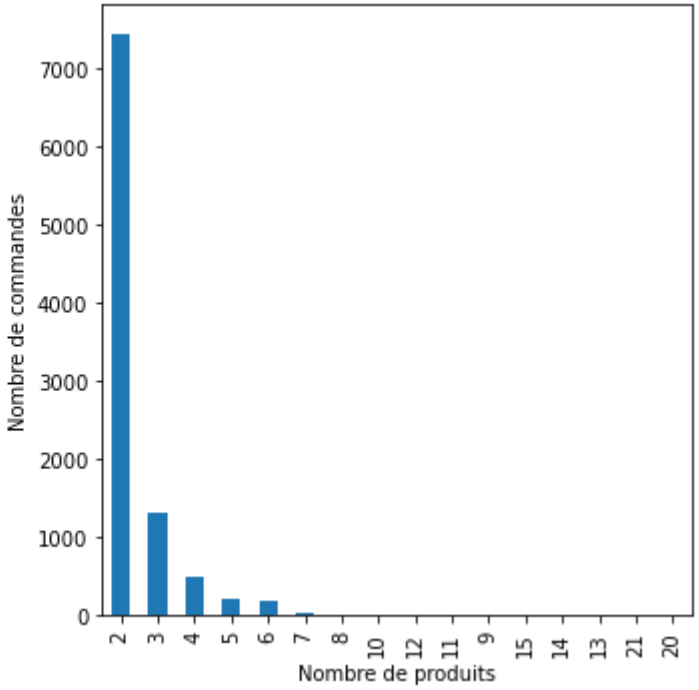


Min	0,85
Max	13440
Moyenne	137,53
Médiane	86,9

II. EXPLORATION :

- Nombre d'articles par commande

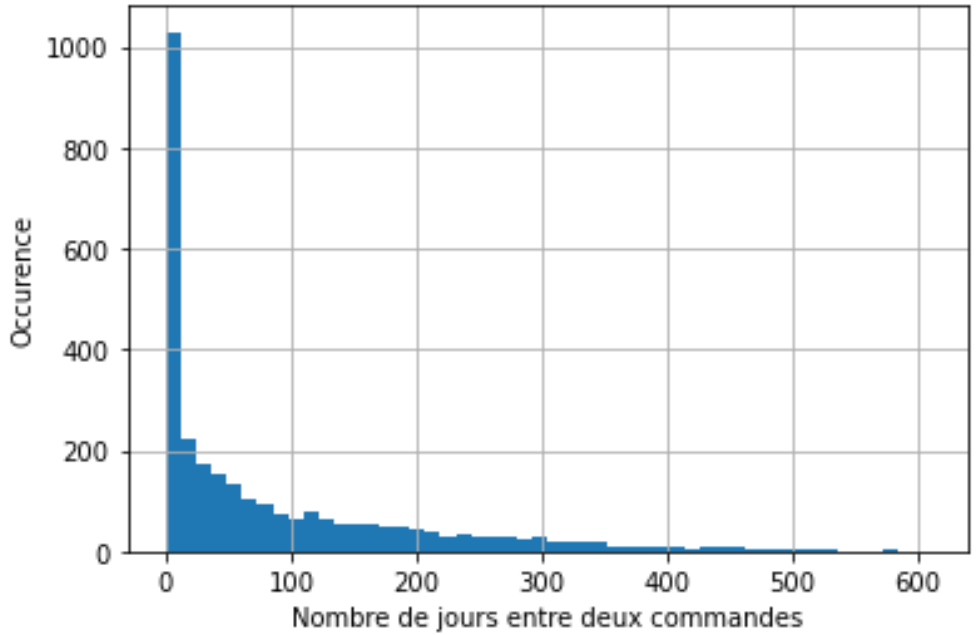




Min	1
Max	21
Moyenne	1,14
Médiane	1

II. EXPLORATION :

- Délai entre deux commandes



Min	1sec
Max	608j 23h29min
Moyenne	86j 16h20min
Médiane	39j 14h36min

- Date de la première commande : 04/09/2016
- Date de la dernière commande : 03/09/2018

III. TEST DE DIFFÉRENTS MODÈLES : RFM

- Sélection des clients ayant réalisés plus de 1 commande (limiter la taille de la matrice pour l'AgglomerativeClustering)
- Création de nouvelles variables :
  - Recency : nombre de jours depuis la dernière commande
  - Frequency : nombre de commande sur la période
  - Monetary : montant dépensé

➔ Création du score R, F et M : entre 1 et 5

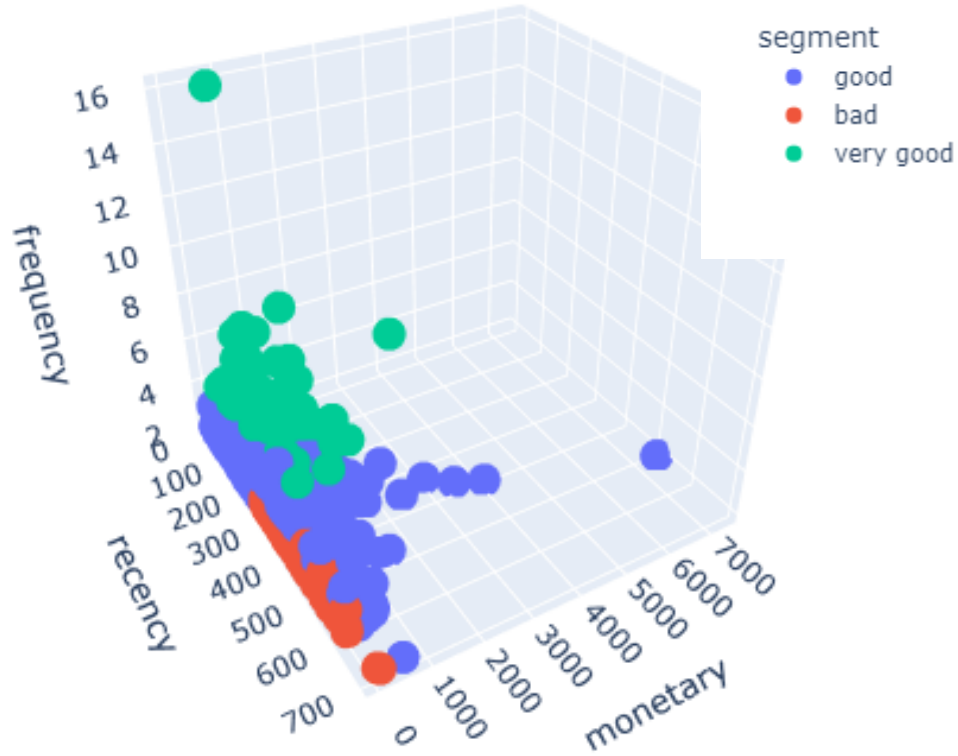
  - R : à partir des quantiles [20%, 40%, 60%, 80%], 1 pour les clients dont la durée entre deux commandes est faible (dernier quantile) 5 pour les clients dont la durée entre deux commandes est longue (1er quantile)
  - M : à partir des quantiles [20%, 40%, 60%, 80%], 1 pour les clients dépensant le plus (dernier quantile), 5 pour les clients dépensant le moins (1er quantile)
  - F : quantiles ne fonctionne pas: F = nb_commande -1 si nb_commande <=6, 5 sinon.

- Score RFM : somme de R, F et M

III. TEST DE DIFFÉRENTS MODÈLES : RFM

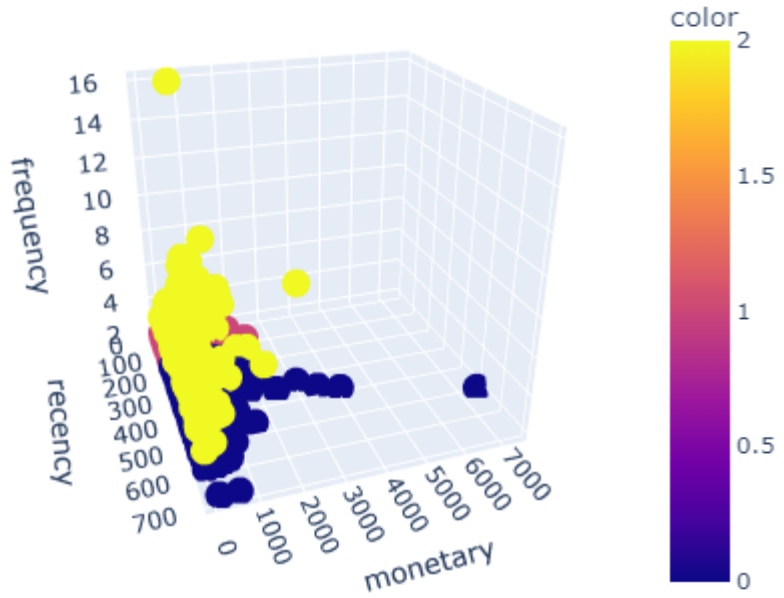
- Segmentation à partir du score RFM:
  - Mauvais client : RFM < 5 : 13,1%
  - Bon client : 5 > RFM < 10 : 72,3%
  - Très bon client : RFM > 10 : 14,5%

- Interprétation :
  - Mauvais client : dépense peu, peu souvent et il y a longtemps
  - Très bon client : commande régulièrement
  - Bon : les autres

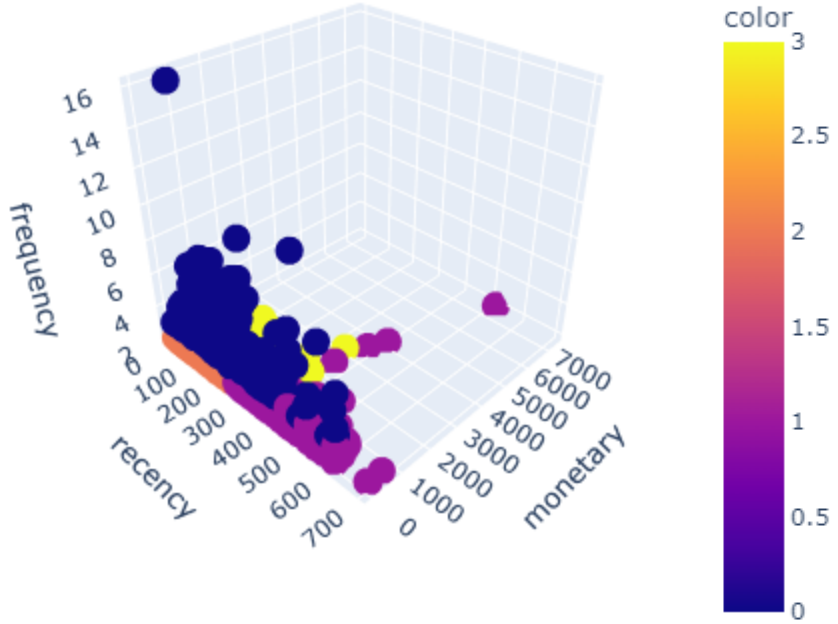


III. TEST DE DIFFÉRENTS MODÈLES : RFM

- Segmentation à l'aide d'un KMEANS :



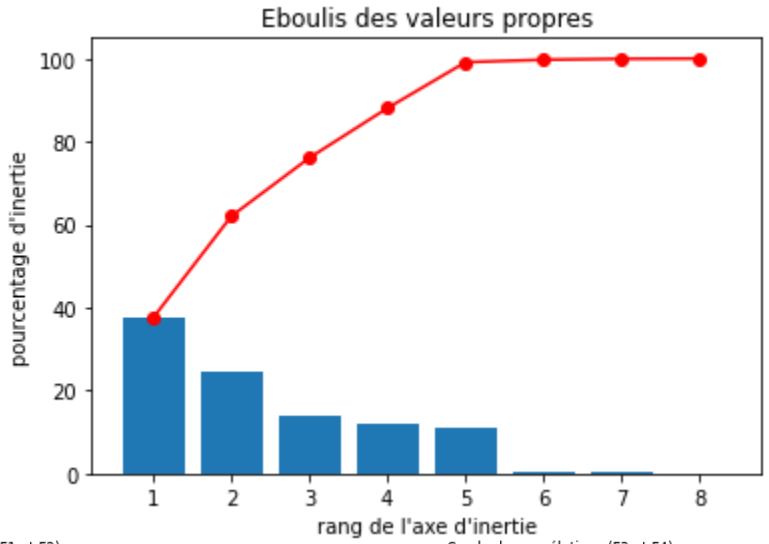
- Cluster 0 : clients commandant peu et dont la dernière commande date de plus de 200j
- Cluster 1 : clients commandant peu mais avec une dernière commande récente
- Cluster 2 : clients commandant fréquemment

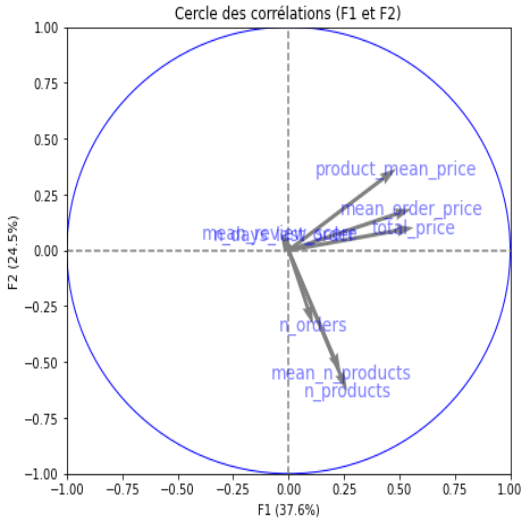


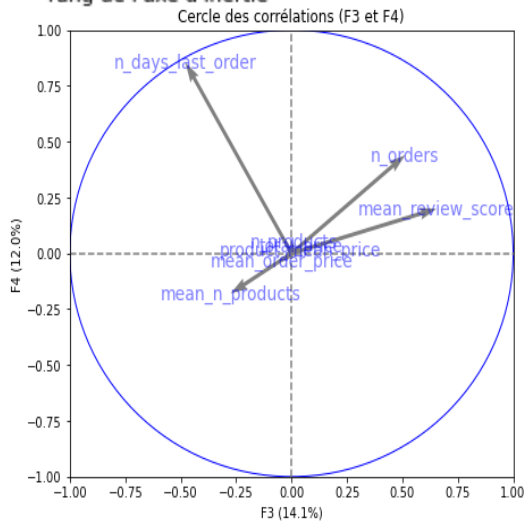
- Cluster 0 : clients avec des montants de commandes très faibles
- Cluster 1 : clients avec une faible fréquence de commande et la dernière datant de plus de 200jours
- Cluster 2 : clients commandant peu et dépensant peu
- Cluster 3 : clients commandant peu, mais avec un montant important

III. TEST DE DIFFÉRENTS MODÈLES :

- Création de nouvelles variables pour caractériser les clients :
  - Nombre de commandes
  - Temps depuis a dernière commande
  - Montant total dépensé
  - Prix moyen des articles
  - Nombre d'articles acheté
  - Prix moyen par commande
  - Nombre d'article moyen par commande
  - Satisfaction moyenne
- ACP pour visualiser les résultats
  - F1 montant dépensé
  - F2 fidélité
  - F3 : nombre de commande et satisfaction

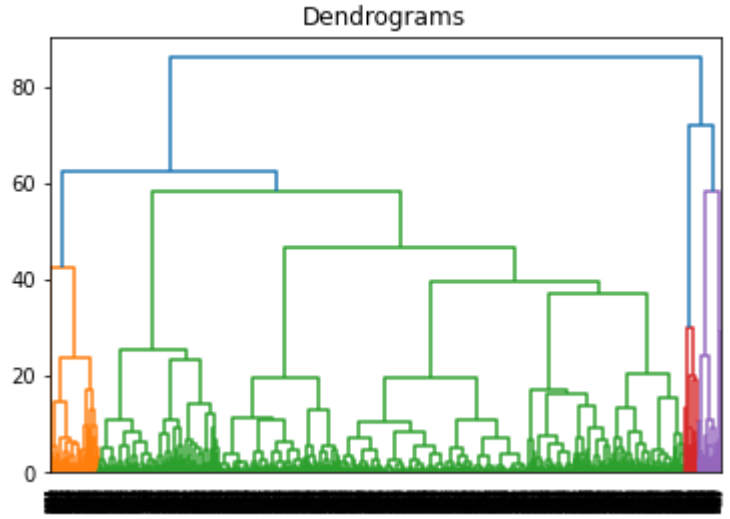


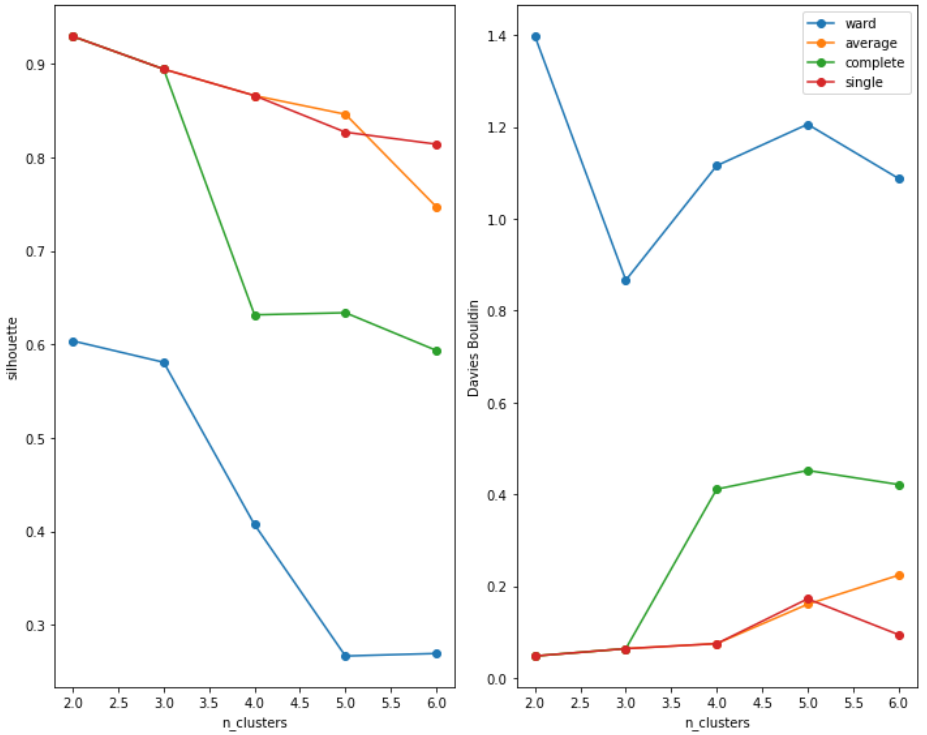




III. TEST DE DIFFÉRENTS MODÈLES : AGGLOMERATION

- Dendrogramme : 4 ou 3 clusters





- Le DB score et le coefficient de silhouettes indique un nombre de clusters optimal à 3

III. TEST DE DIFFÉRENTS MODÈLES : AGGLOMERATION

- 4 clusters permet d'obtenir également une segmentation interpréta
- Interprétation :
  - F1 : proportionnel au montant dépensé
  - F2 : inversement proportionnel au nombre de produits / commandes
  - F3 : proportionnel au nombre de commande et à la satisfaction
  - Cluster 0 : 87,5%

Client moyen = proche de 0 selon tous les axes

  - Cluster 1: 3,2%

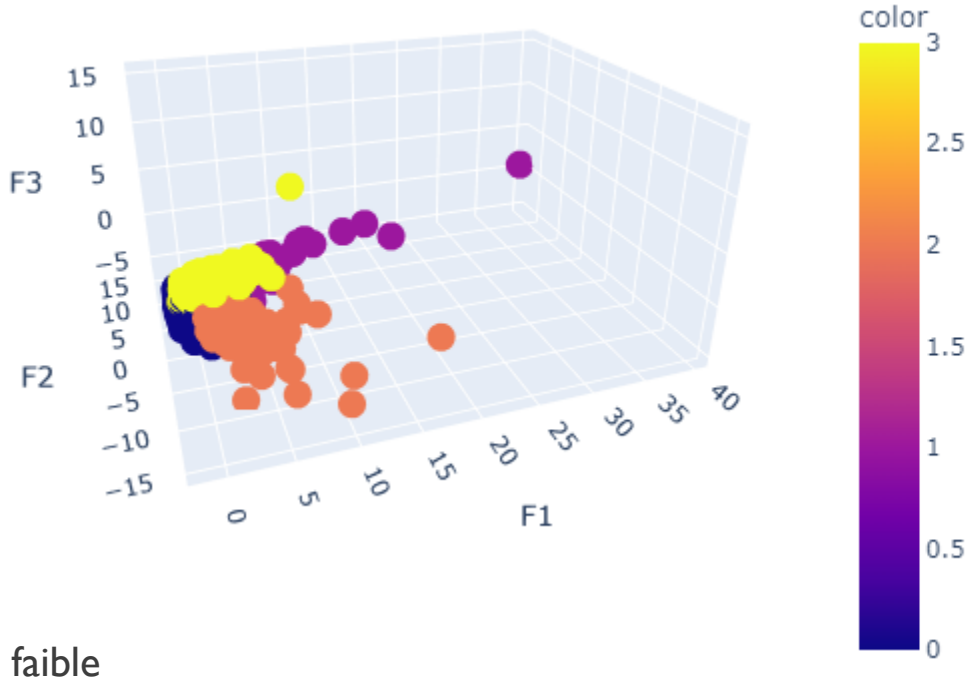
Client commandant beaucoup de produit mais avec peu de commande ou une faible satisfaction

  - Cluster 2 : 2,0%

Peu de produits commandés, peu de commande ou peu satisfait

  - Cluster 3 : 7,3%

Peu de produits commandés, mais beaucoup de commande ou grande satisfaction



III. TEST DE DIFFÉRENTS MODÈLES : KMEANS

- Calcul du coefficient de silhouette et de l'indice de Davies-Bouldin
- Cluster 0 : 3,5%

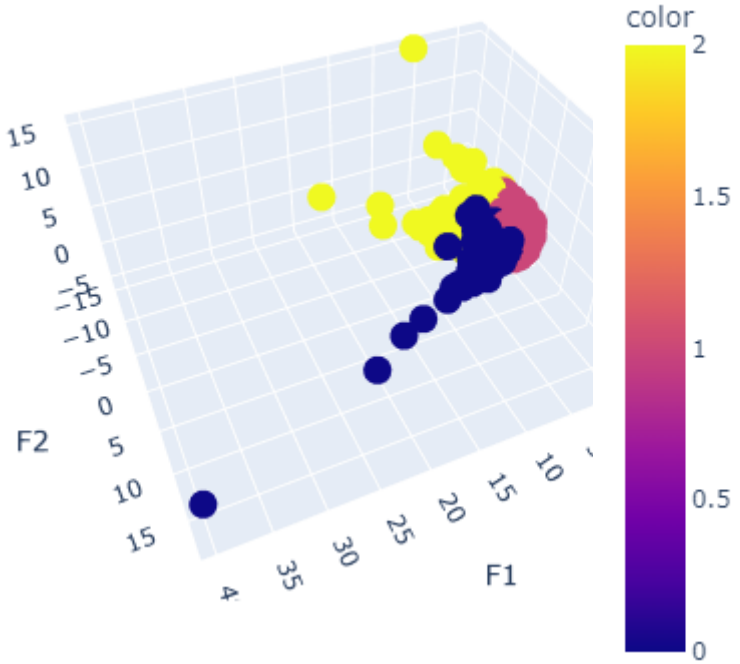
Peu de produit commandé

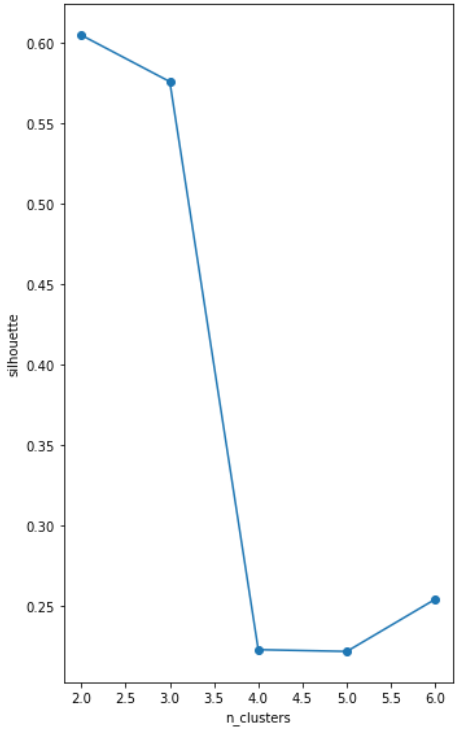
- Cluster 1: 93,2%

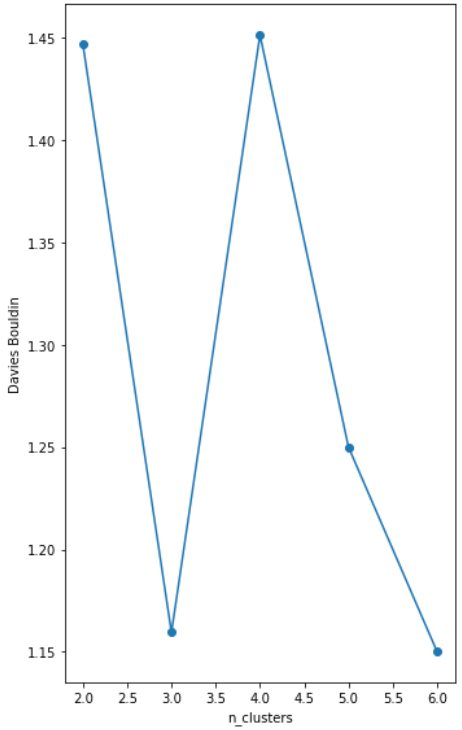
Client moyen

- Cluster 2: 3,3%

Beaucoup de produit commandé

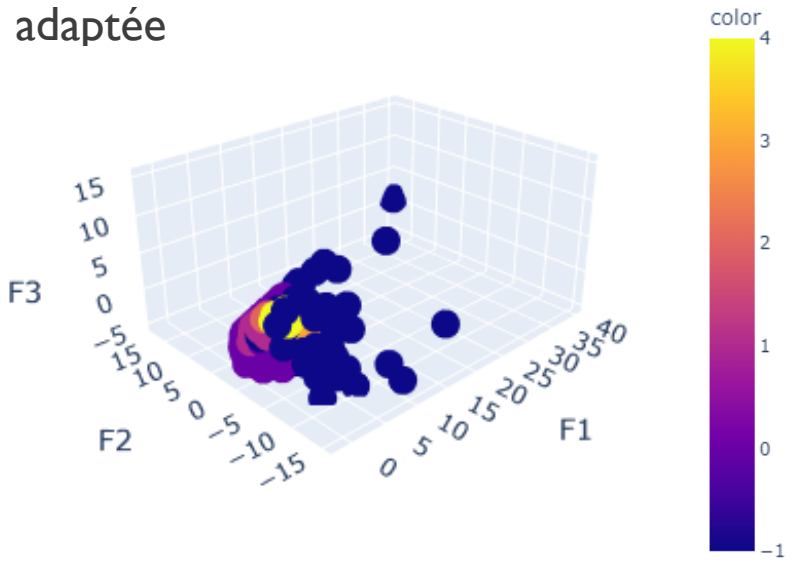


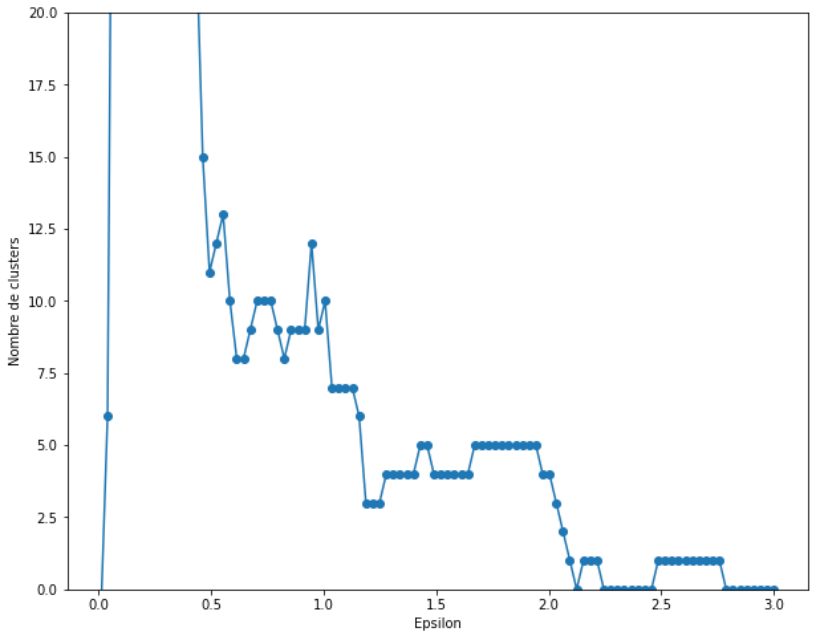




III. TEST DE DIFFÉRENTS MODÈLES : DBSCAN

- Effet du epsilon :
  - Plus epsilon est grand, plus le nombre de cluster est faible
  - Variation « anarchique » : difficile de déterminer une valeur optimale
- Faible séparation entre clusters
- Majorité dans le cluster « bruit »
- Méthode non adaptée





IV. INTERPRÉTATIONS DE LA SEGMENTATION : AGGLOMERATIVECLUSTERING

Params	Cluster 0	Cluster 1	Cluster 2	Cluster 3	All
N_orders	2.0	2.0	2.27	3.44	2.11
N_days_last_order	229	225	170	204	226
Total_price	205	1212	733	387	261
N_products	2.3	2.18	8.8	4.0	2.5
Review_score	4.1	4.0	3.68	4.5	4.1

- Cluster 0 : client moyen, dépensant peu
- Cluster 1 : client dépensant beaucoup
- Cluster 2 : client ayant commandé récemment, dépensant beaucoup (quantité + montant)
- Cluster 3 : client commandant régulièrement, un nombre important d'articles et plutôt satisfait.

IV. INTERPRÉTATIONS DE LA SEGMENTATION : KMEANS

Params	Cluster 0	Cluster 1	Cluster 2	All
N_orders	1.0	1.0	1.2	1.0
N_days_last_order	245	243	236	243
Total_price	862.8	107	355	142
N_products	1.2	1.1	5.3	1.2
Review_score	4	4.1	3.35	4.1

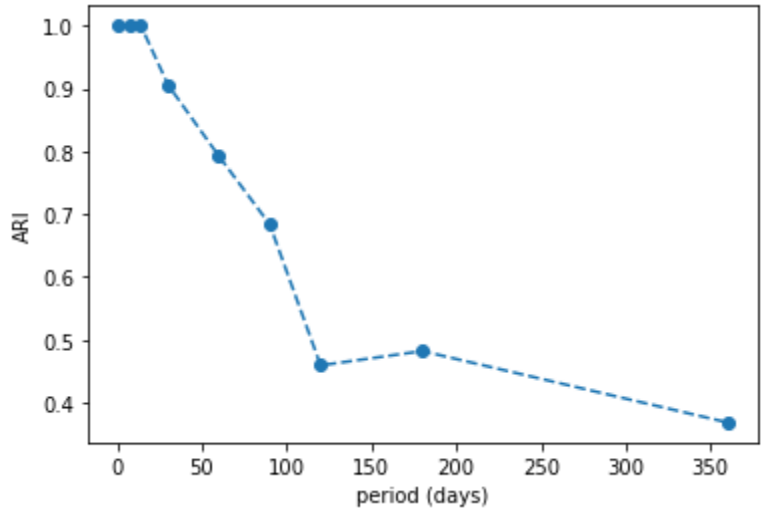
- Cluster 0 : Client dépensant beaucoup
- Cluster 1 : client moyen, dépensant peu
- Cluster 2 : client ayant commandé plus régulièrement, plus de produits, pour un montant plus élevé

IV. MISE A JOUR DU MODELE

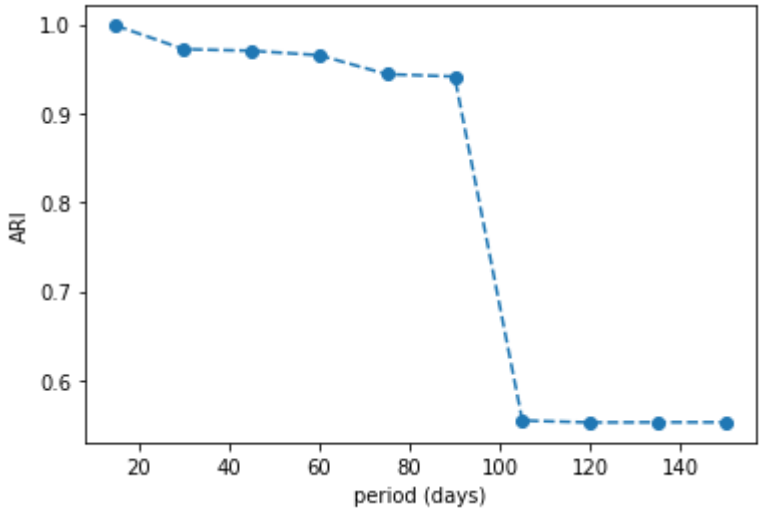
- Démarches :
  1. Sélection d'une période initiale pour construire la base de données
  2. Normalisation des données
  3. Création d'un classifieur
  4. Création d'une nouvelle base de données sur une période de temps incrémentée
  5. Segmentation de ces données avec le classifieur initiale
  6. Création d'un nouveau classifieur
  7. Segmentation des données avec ce nouveau classifieur
  8. Comparaison des segmentations grâce au calcul de l'indice de rang ajusté

IV. MISE A JOUR DU MODELE

- Période initiale : 09/2016 ➔11/2016 (90j)



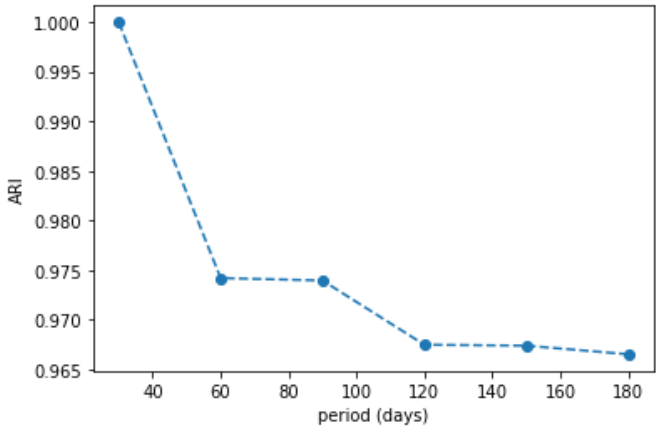
- Période initiale : 02/2018 ➔05/2018 (90j)



- Il est nécessaire de mettre à jour la segmentation tous les 3 mois

V. CONCLUSIONS

- La base de données est de qualité et facilement transformable
- La majorité des clients commandent qu'une seule fois
- La segmentation RFM permet d'obtenir un segmentation rapide et facilement interprétable
- La segmentation AgglomerativeClustering permet de créer 4 clusters interprétables
  - Mais elle est très couteuse en mémoire
- La segmentation KMEANS permet de créer 3 groupes de clients interprétables.
- Dans tous les cas on arrive à détecter le client moyen et les bon clients.
- Une mise à jour de la base de données est nécessaire tous les 3 mois
- Une mise à jour plus longue est possible si on accepte de prendre la totalité de la base de données



V. AMÉLIORATIONS

- Retour vers les experts commerces pour discuter de l'interprétation de la segmentation
- Envisager des options de big data pour utiliser l'algorithme agglomerativeclustering sur toute la base de données
- Prendre en compte les préférences des produits pour effectuer un ciblage commercial précis
- Identifier les vendeurs les plus performants pour les mettre en avant

MERCI POUR VOTRE ATTENTION !