OpenClassrooms – Parcours Data Scientist

Projet 5 :
Segmentation des clients d'un site e-commerce

Fanjamalala Rajaonalison

11/2021

Présentation de l'étude

Contexte

Client : 

solution de vente sur les marketplaces en ligne

Mission : fournir aux équipes d'e-commerce une segmentation des clients qu'elles pourront utiliser au quotidien pour leurs campagnes de communication.

⇒ Base de données anonymisée comportant des informations sur l'historique de commandes, les produits achetés, les commentaires de satisfaction, et la localisation des clients.

Problématique

Problématique : Comprendre les différents types d'utilisateurs grâce à leur comportement d'achat et à leurs données personnelles

Objectifs : Segmentation des clients

- **Description actionnable** 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.

Méthodologie

Traitements et Analyses des données

- Mise en forme & Cleaning
- **Exploration**

Essais de différentes segmentation

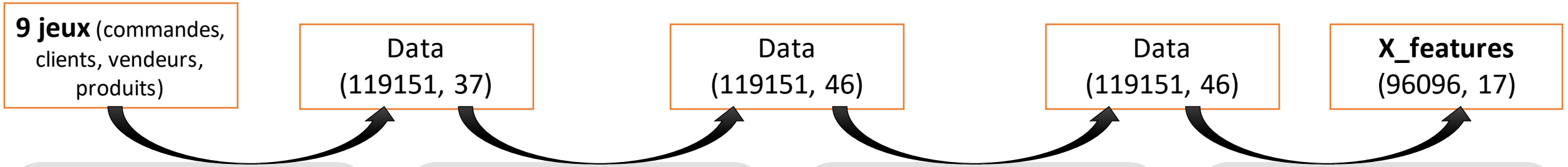
- Segmentation globale
- Segmentation RFM
- Modèles : **DBSCAN, K-MEANS**

Finalisation de l'étude

- **Sélection du modèle final**
- Définition des groupes
- Maintenance du modèle

Traitements
Analyses
des données

Mise en forme & Cleaning



Mise en forme des données

- Renommage des **variables** communes
- Suppression des lignes dupliquées
- Concaténation de tous les jeux

Création / transformation de features

- Paiement : Nombre (par type)
- Transport : Date estimée de livraison, Distance des vendeurs
- Produits : Taille description de catégorie
- Comportements clients : Taille de commentaires, Nombre de jours off

Traitements des valeurs manquantes

- Variable catégorielle → "inconnu"
- Variable numérique → **Médiane de la colonne en** filtrant **les** produits de **même ID** / catégorie

Finalisation des features

- Groupement des données par l'unique_ID des clients
- Calcul de features selon :
  → RFM
  → Paiements
  → Vendeur / Transport
  → Description des produits
  → **Retour clients**

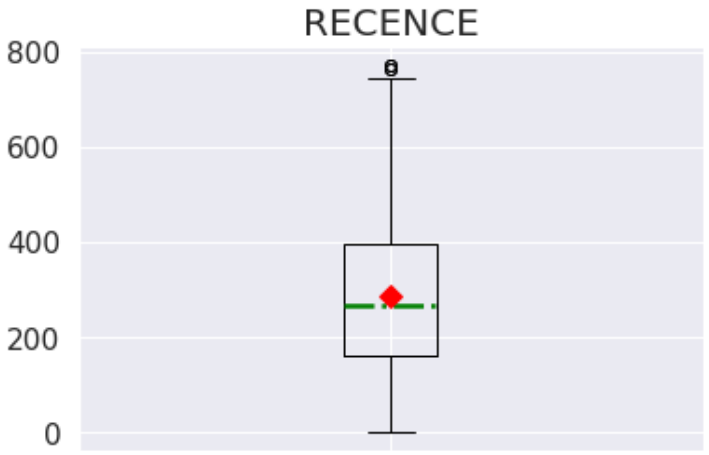
Mise en forme & Cleaning

#	Column	Non-Null Count	Dtype
0	recence	96096 non-null	int64
1	frequence	96096 non-null	int64
2	montant	96096 non-null	float64
3	n_payment	96096 non-null	float64
4	pay_credit_card	96096 non-null	uint8
5	pay_debit_card	96096 non-null	uint8
6	pay_boleto	96096 non-null	uint8
7	pay_voucher	96096 non-null	uint8
8	seller_distance	96096 non-null	float64
9	freight_value	96096 non-null	float64
10	estimated_delivery_days	96096 non-null	int64
11	product_description_lenght	96096 non-null	float64
12	product_category_lenght	96096 non-null	float64
13	product_photos_qty	96096 non-null	float64
14	day_off	96096 non-null	float64
15	review_score	96096 non-null	float64
16	review_comment_lenght	96096 non-null	float64

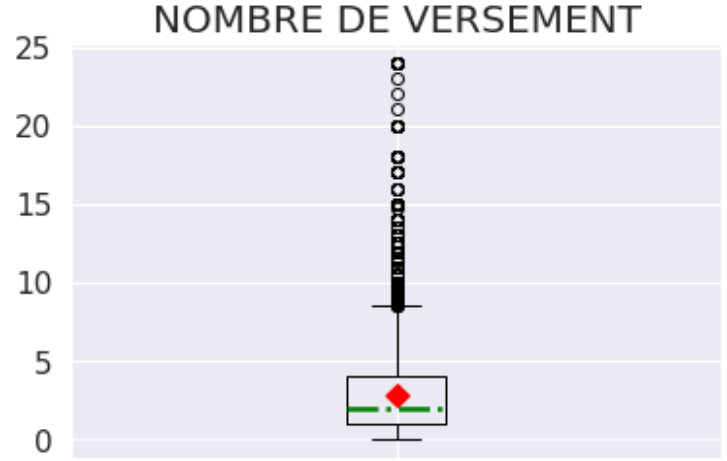
Données : (96096, 17)

- RFM
- Paiements
- Vendeur / Transport
- Description des produits
- Retour clients

Exploration

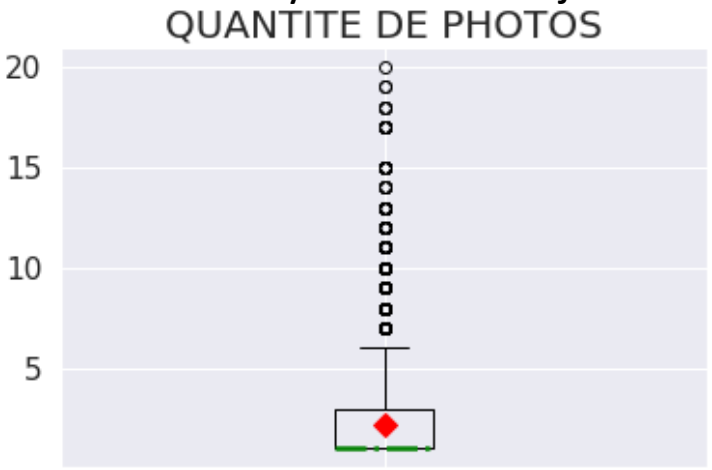


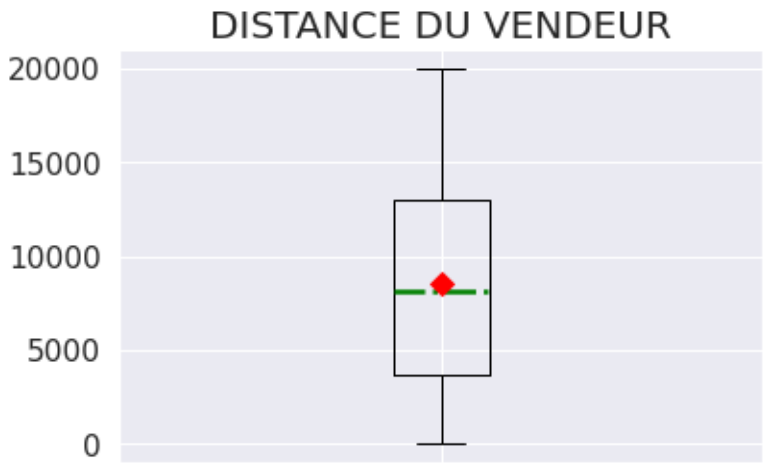
Le nombre de jours passés après la dernières commande varient fortement avec une moyenne de 287 jours.



Les clients paient généralement en 2, 3, 4 fois.

Les clients donnent en général une bonne note.

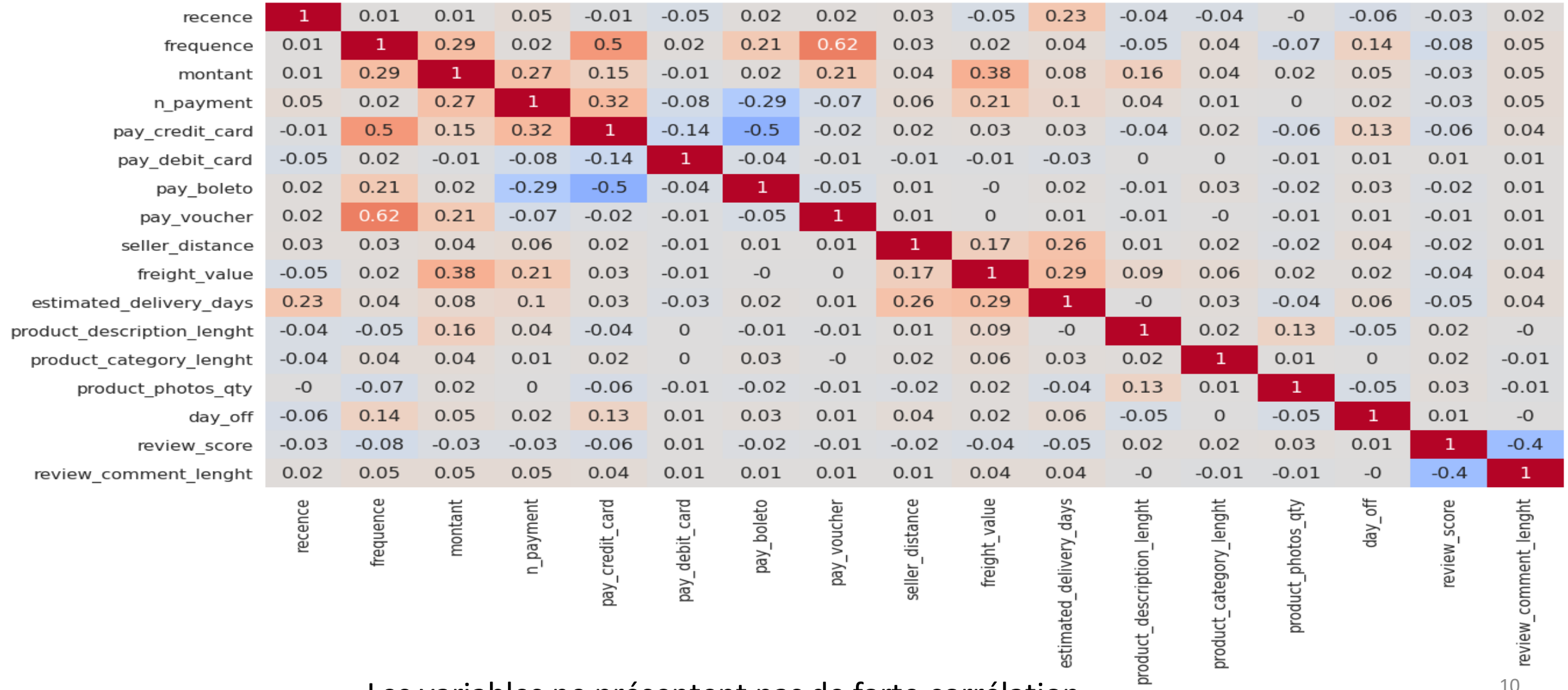




En général, les produits achetés :

- présentent peu de photos
- sont vendus par des vendeurs en moyenne à 10000km des clients

Exploration



Les variables ne présentent pas de forte corrélation

Essais de différentes segmentations

Segmentation globale

Principe :

Segmentation sur l'ensemble des features (17 features)

Essais de 2 modèles :

- DBSCAN
- K-MEANS

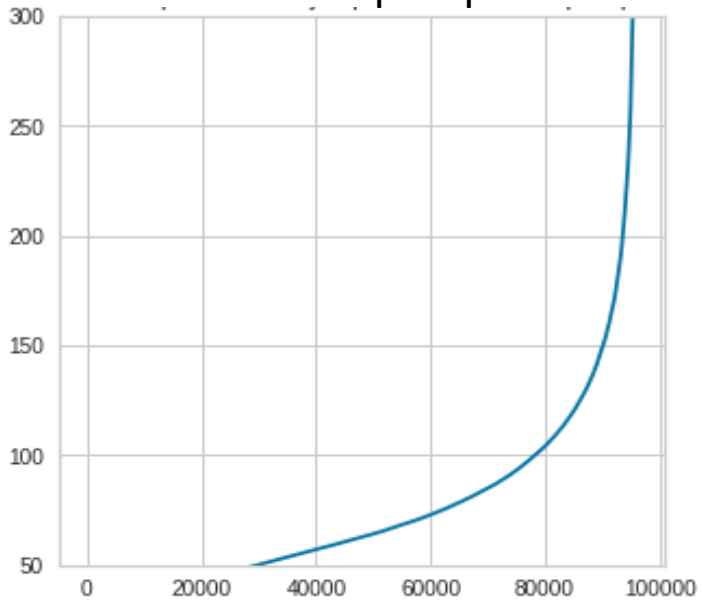
Segmentation globale

1. DBSCAN

a. Estimation des paramètres : eps & min_sample

- min_sample : ~2 fois la taille du jeu ⇒ **34**

- eps : Points triés par distance jusqu'au 34e voisin le plus proche



⇒ **125**

Segmentation globale

1. DBSCAN

b. Implémentation du modèle

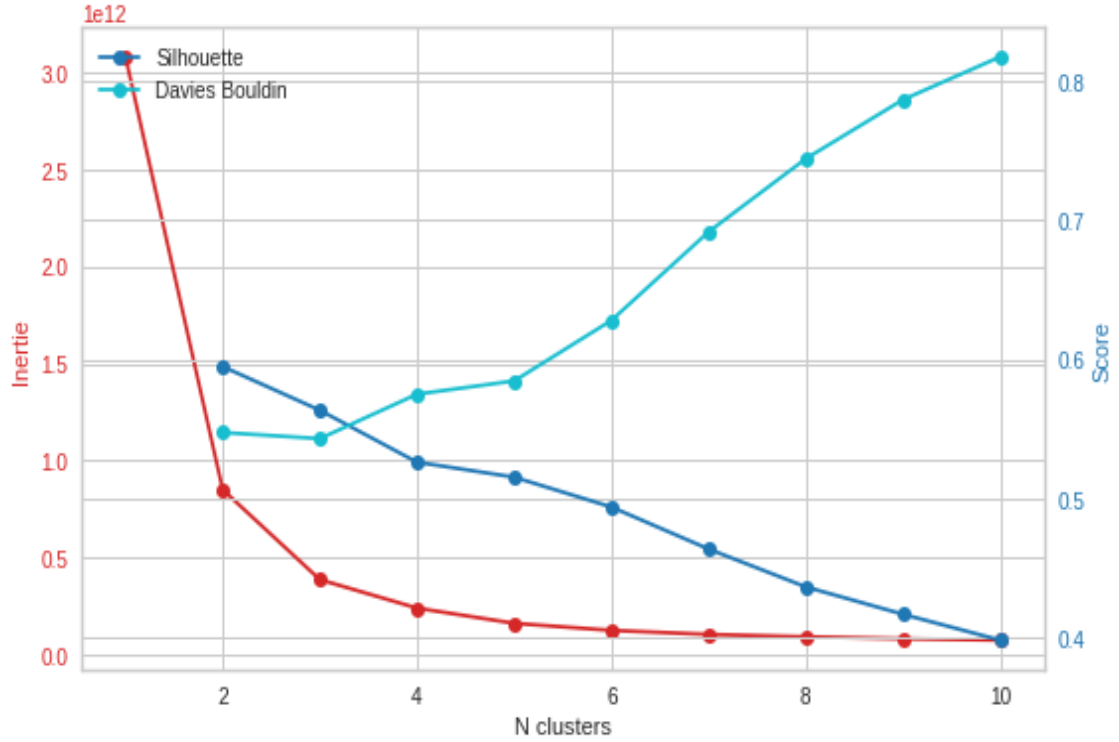
- Nombre de clusters : 14 ⇒ Assez grand mais acceptable
- Silhouette Coefficient: -0.752
- Score Davies Bouldin: 2.426
- ⇒ Scores pas bons

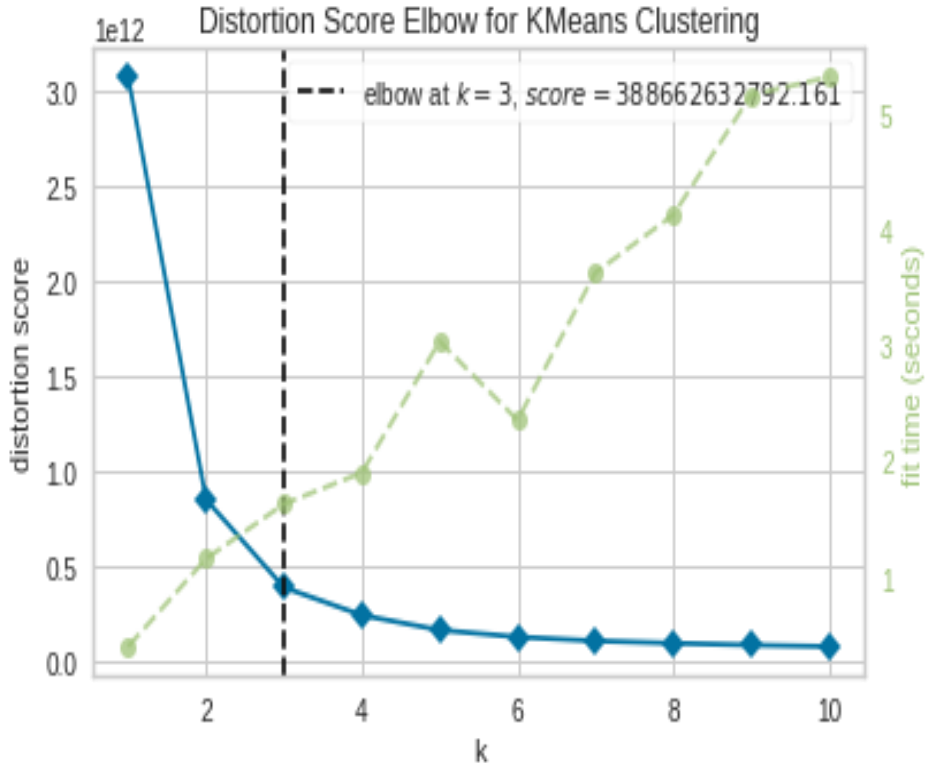
⇒ Modèle pas pertinent pour les données

Segmentation globale

2. K-MEANS

a. Estimation du nombre de cluster





⇒ N_clusters = 3

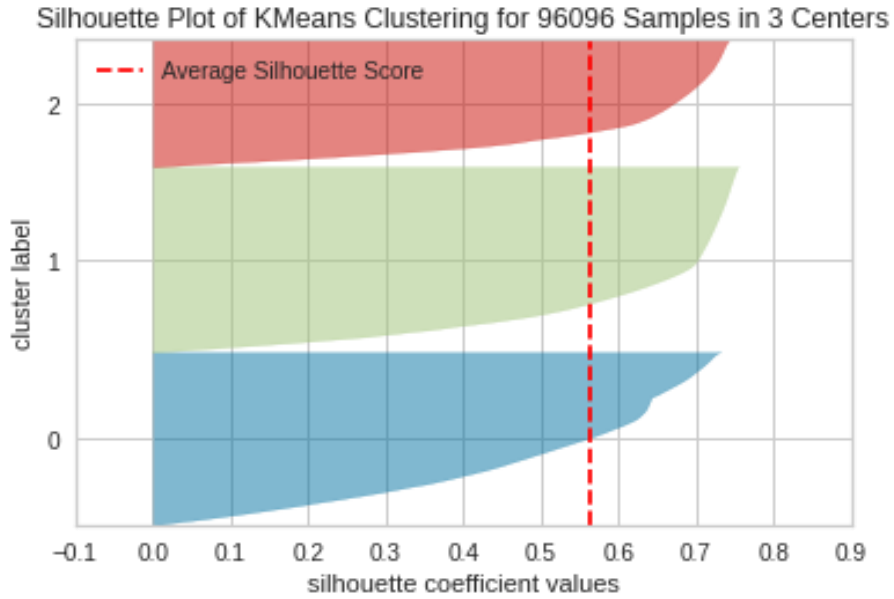
Segmentation globale

2. K-MEANS

b. Implémentation du modèle

- Nombre de clusters : 3

- Silhouette Coefficient: 0.564 ⇒ score assez bon



- clusters avec un silhouette score > silhouette moyen
  ⇒ chaque groupe bien éloigné des autres

- épaisseurs du tracé des silhouette des clusters similaires
  ⇒ taille des clusters assez similaires entre eux.

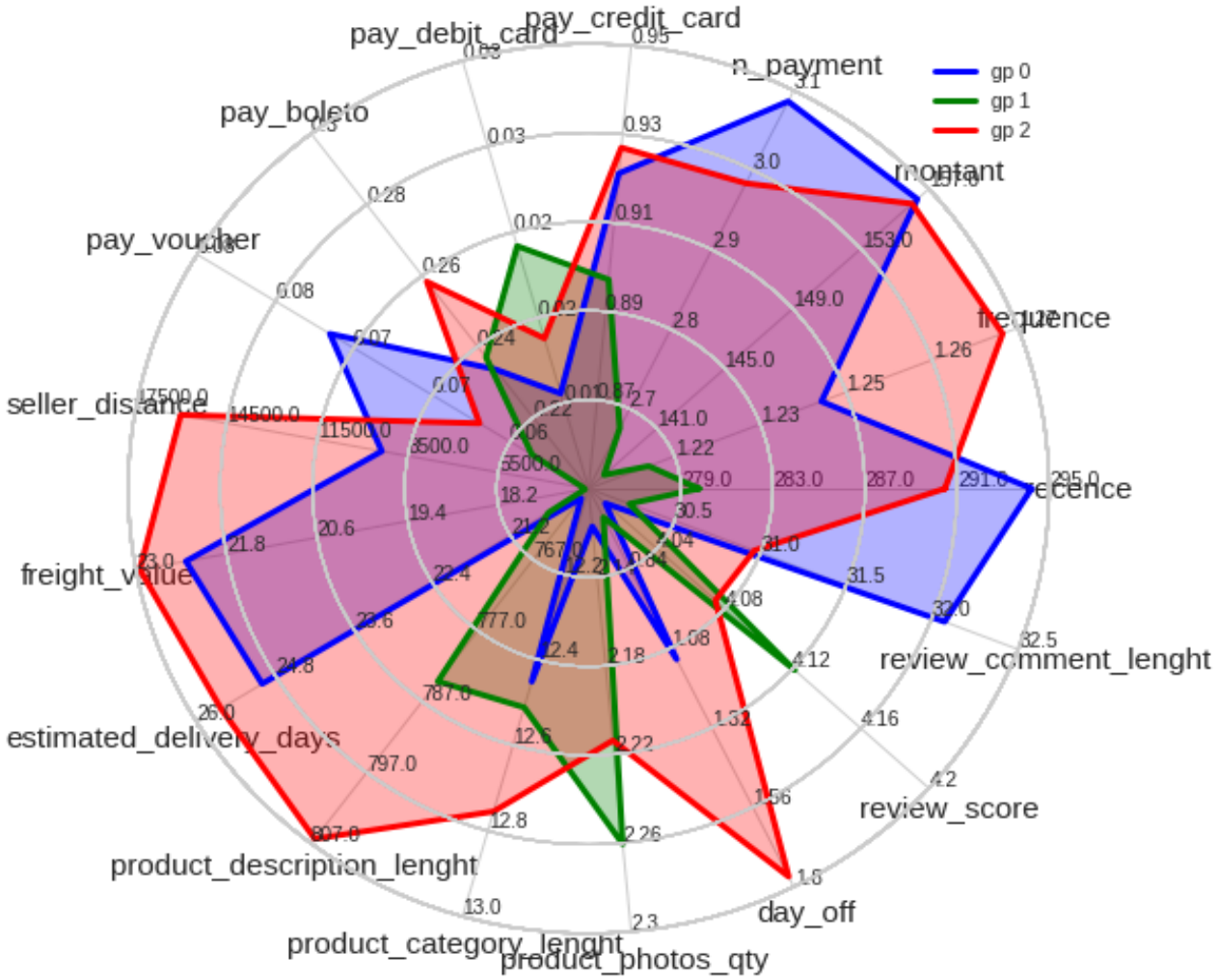
- Score Davies Bouldin: 0.543 ⇒ score assez bon

⇒ Modèle pertinent pour les données

Segmentation globale

2. K-MEANS

c. Interpretation



Comportements différents pour les 3 groupes obtenus

⇒ **Bonne segmentation des clients**

Segmentation RFM

Principe :

Segmentation sur 3 features :

- Récence : date de la dernière commande (en nombre de jours)
- Fréquence : fréquence de commandes
- Montant : montant total de tous les commandes passées

Essais de 2 modèles :

- DBSCAN
- K-MEANS

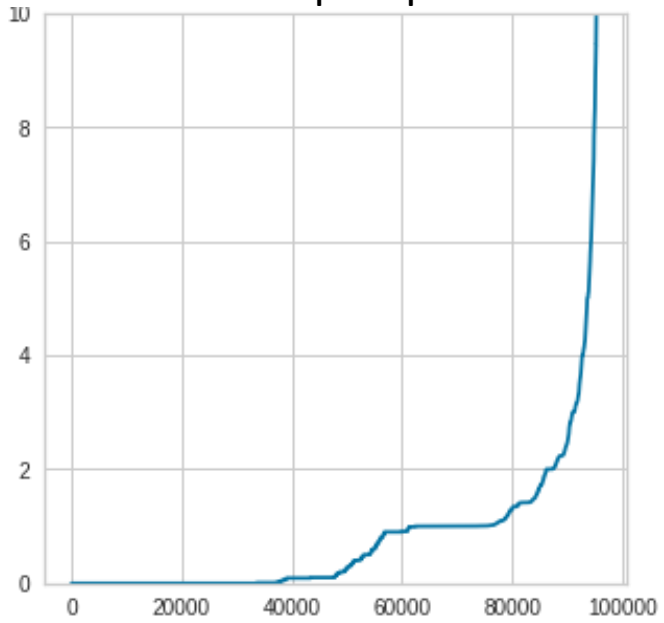
Segmentation RFM

1. DBSCAN

a. Estimation des paramètres : eps & min_sample

- min_sample : ~2 fois la taille du jeu ⇒ 6

- eps : Points triés par distance jusqu'au 6e voisin le plus proche



⇒ 1.25

Segmentation RFM

1. DBSCAN

b. Implémentation du modèle

- Nombre de clusters : 110 ⇒ Trop grand
- Silhouette Coefficient: -0.660
- Score Davies Bouldin: 1.784

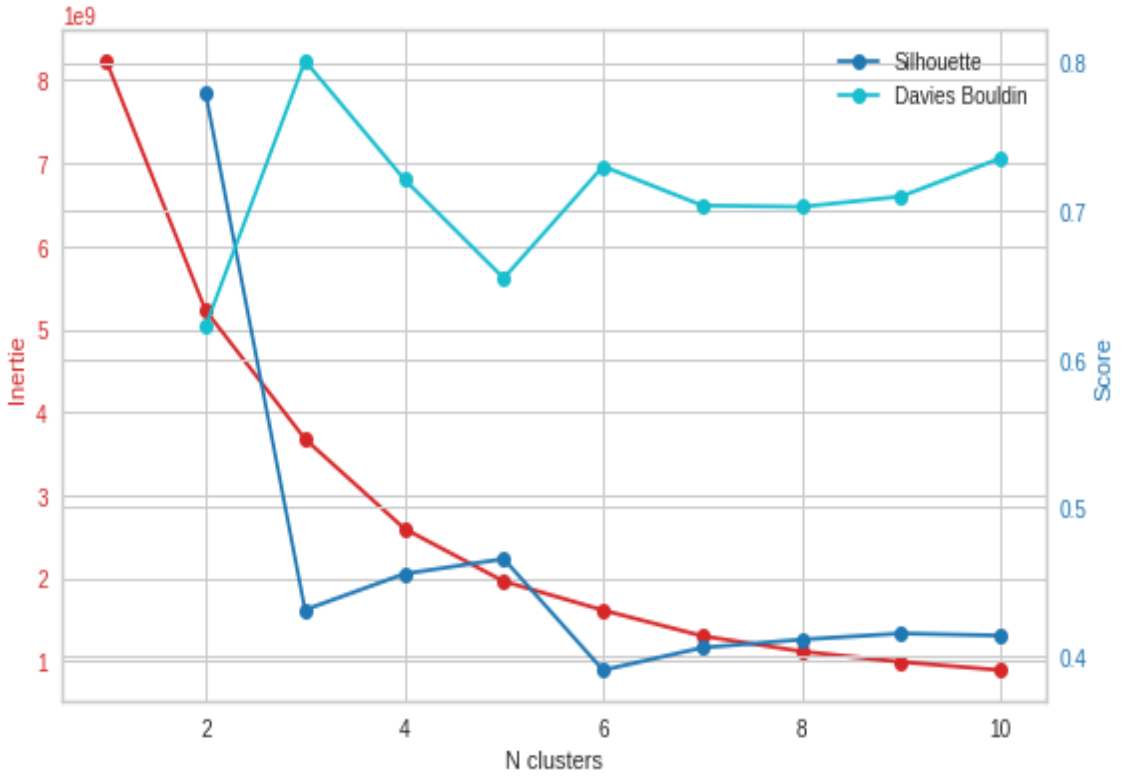
⇒ Scores pas bon

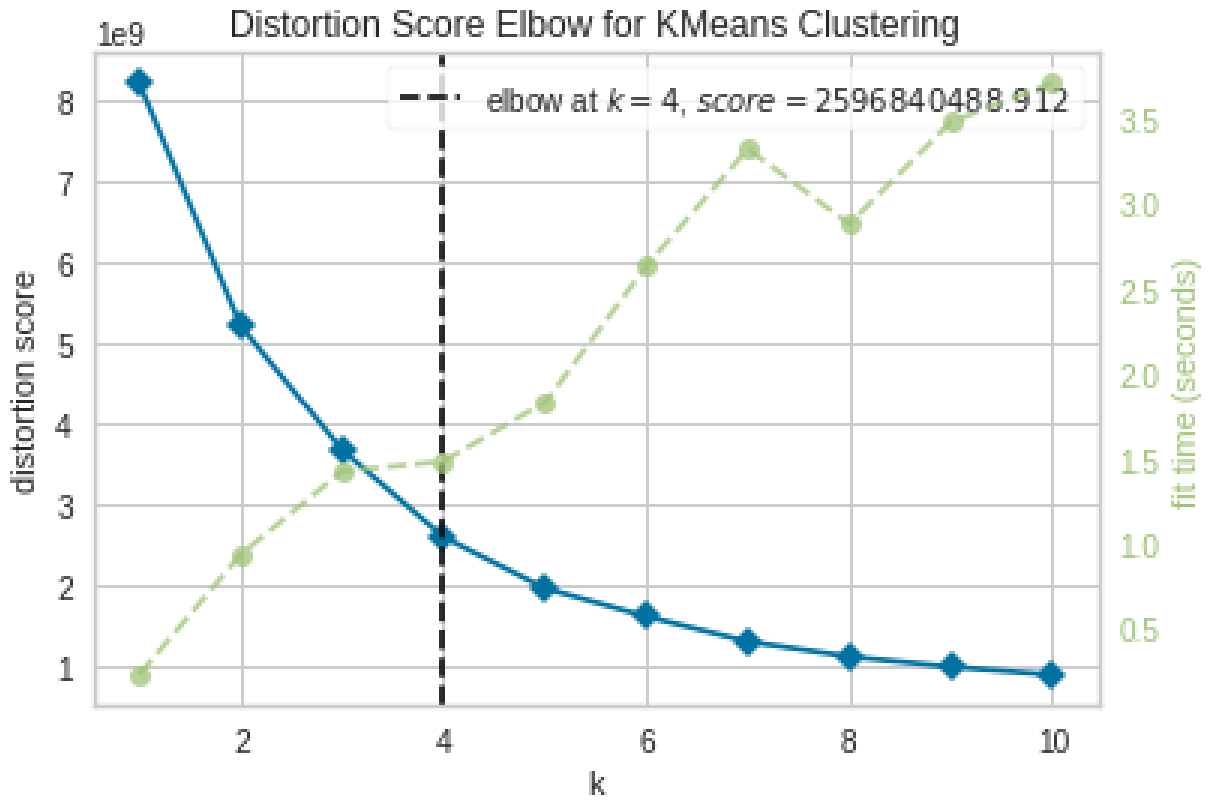
⇒ Modèle pas pertinent pour les données

Segmentation RFM

2. K-MEANS

a. Estimation du nombre de cluster





⇒ N_clusters = 5 (meilleurs scores)

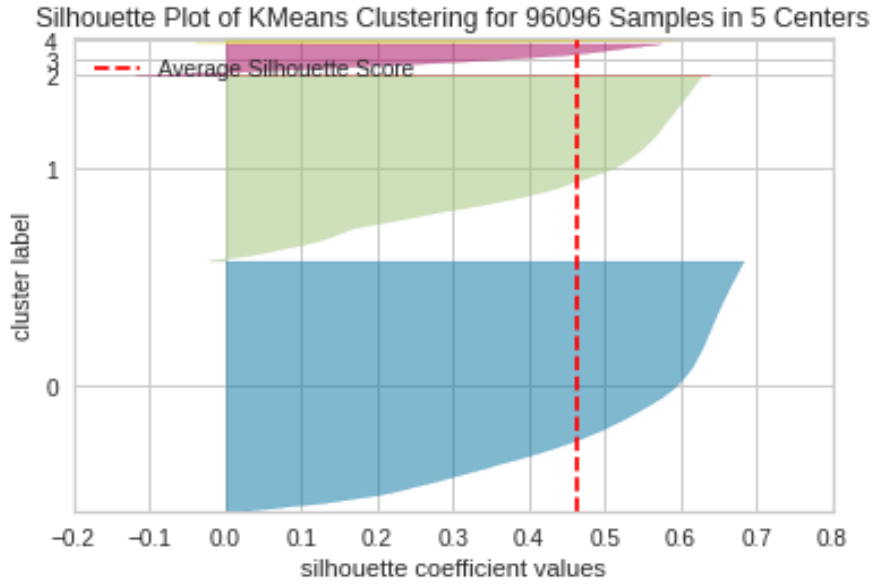
Segmentation RFM

2. K-MEANS

b. Implémentation du modèle

- Nombre de clusters : 5

- Silhouette Coefficient: 0.465 ⇒ score moyennement bon



- clusters avec un silhouette score > silhouette moyen
  ⇒ chaque groupe bien éloigné des autres

- épaisseurs du tracé des silhouette des clusters non similaires
  ⇒ tailles des clusters pas similaire entre eux.

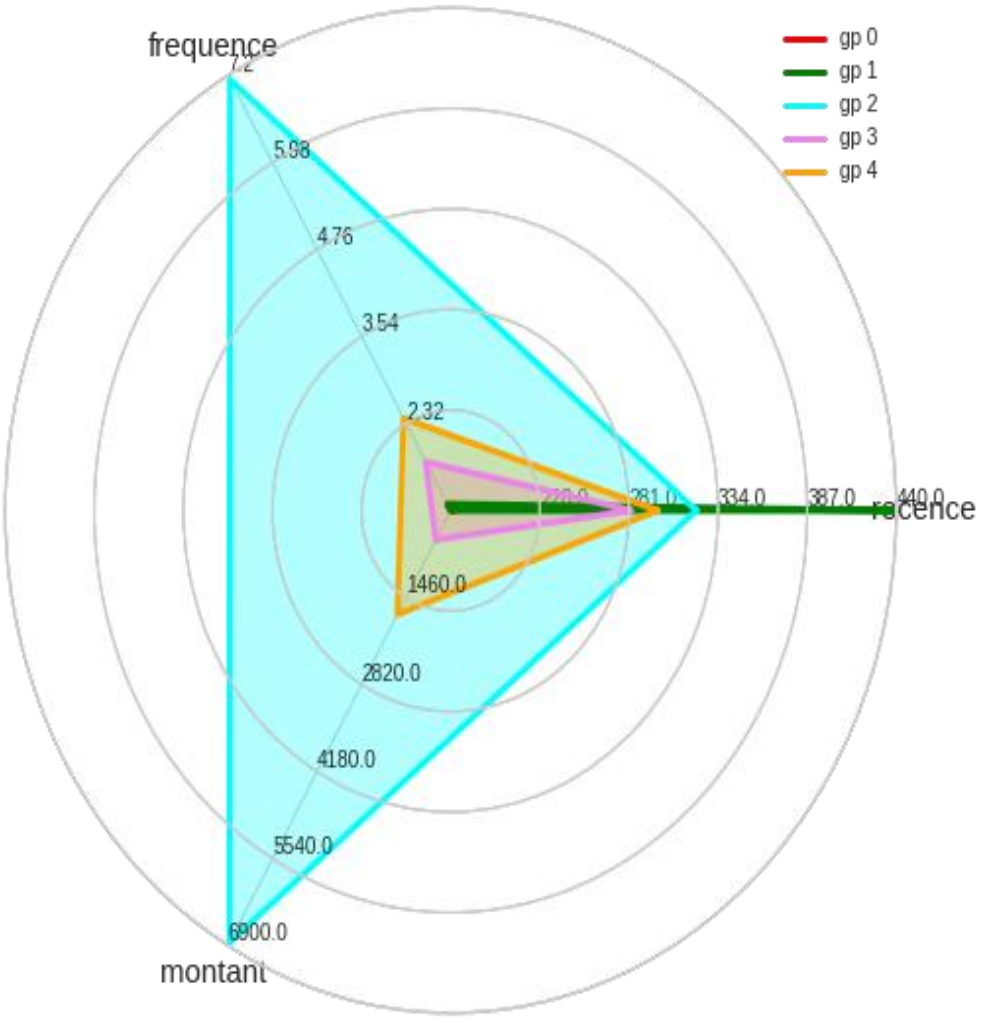
- Score Davies Bouldin: 0.655 ⇒ score moyennement bon

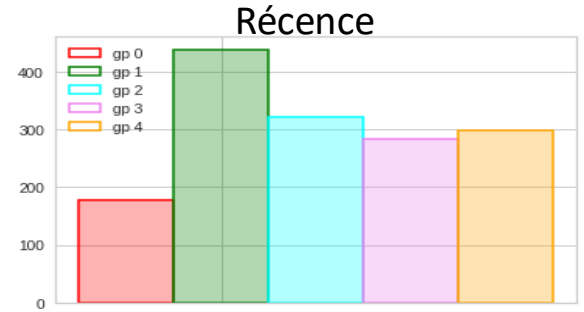
⇒ Modèle pertinent pour les données

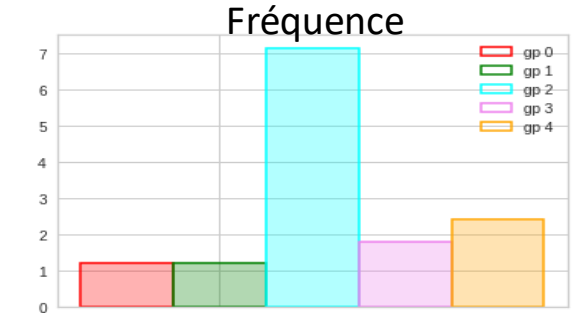
Segmentation RFM

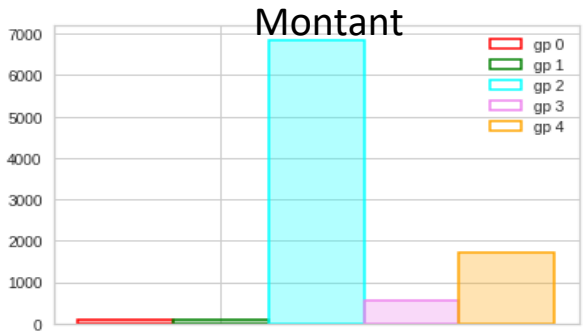
2. K-MEANS

c. Interpretation







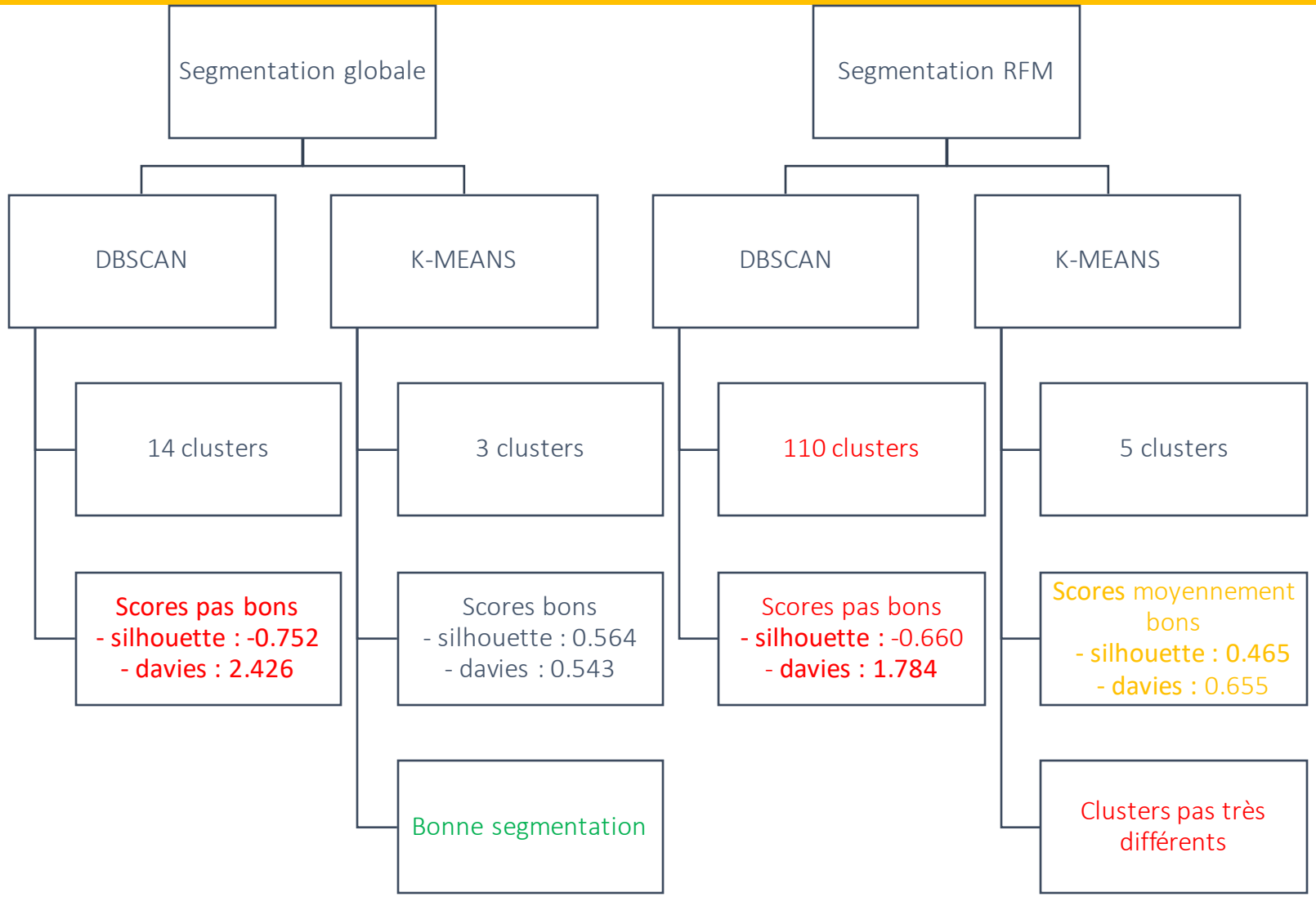


- Comportements assez similaires pour les groupes 0 et 1 (seule différence la date de dernière commande)
- Comportements assez similaires pour les groupes 3 et 4 (seule différence la date de dernière commande)

⇒ Segmentation des clients moyennement bonne

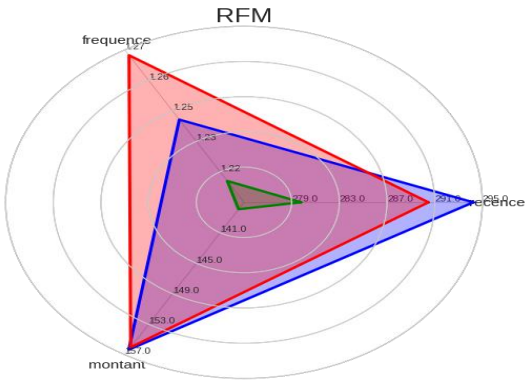
Finalisation de l'étude

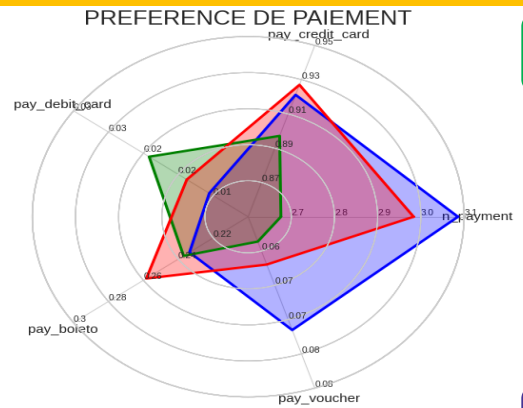
Choix de la Segmentation

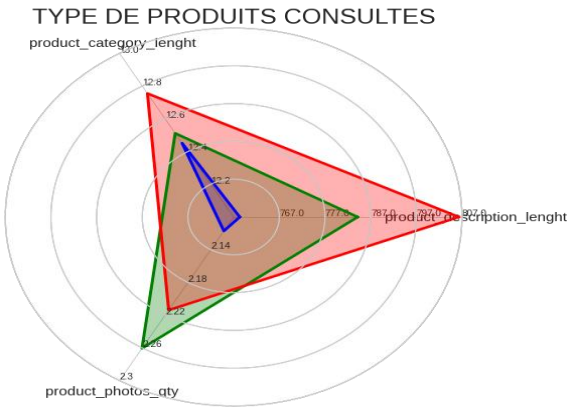


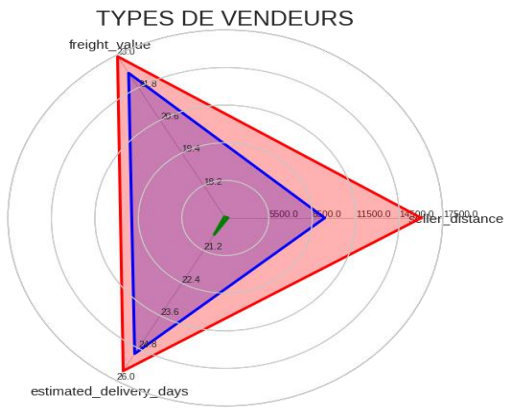
⇒ Segmentation globale avec K-MEANS

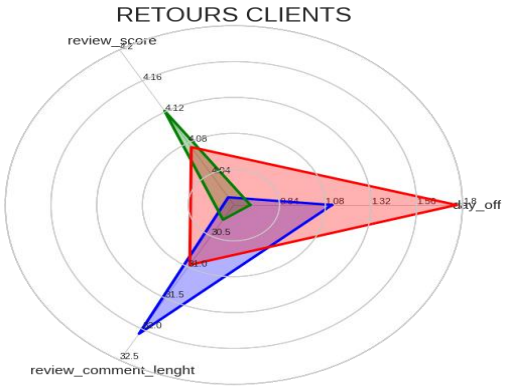
Segmentation globale avec K-MEANS











PETITS CLIENTS

- **Comportement RFM :** peu d'achats, petits montants, dernière commande pas trop longtemps.
- **Préférence de paiement :** en peu de fois et plutôt par carte de crédit et/ou carte de débit.
- **Type de produits consultés :** Qté d'informations (description, catégorie) moyennes et beaucoup de photos.
- **Type de vendeurs fréquentés :** vendeurs proches, avec très peu ou pas de frais de livraison et peu de temps de livraison.
- **Retour :** bonnes notes, très peu de commentaires.

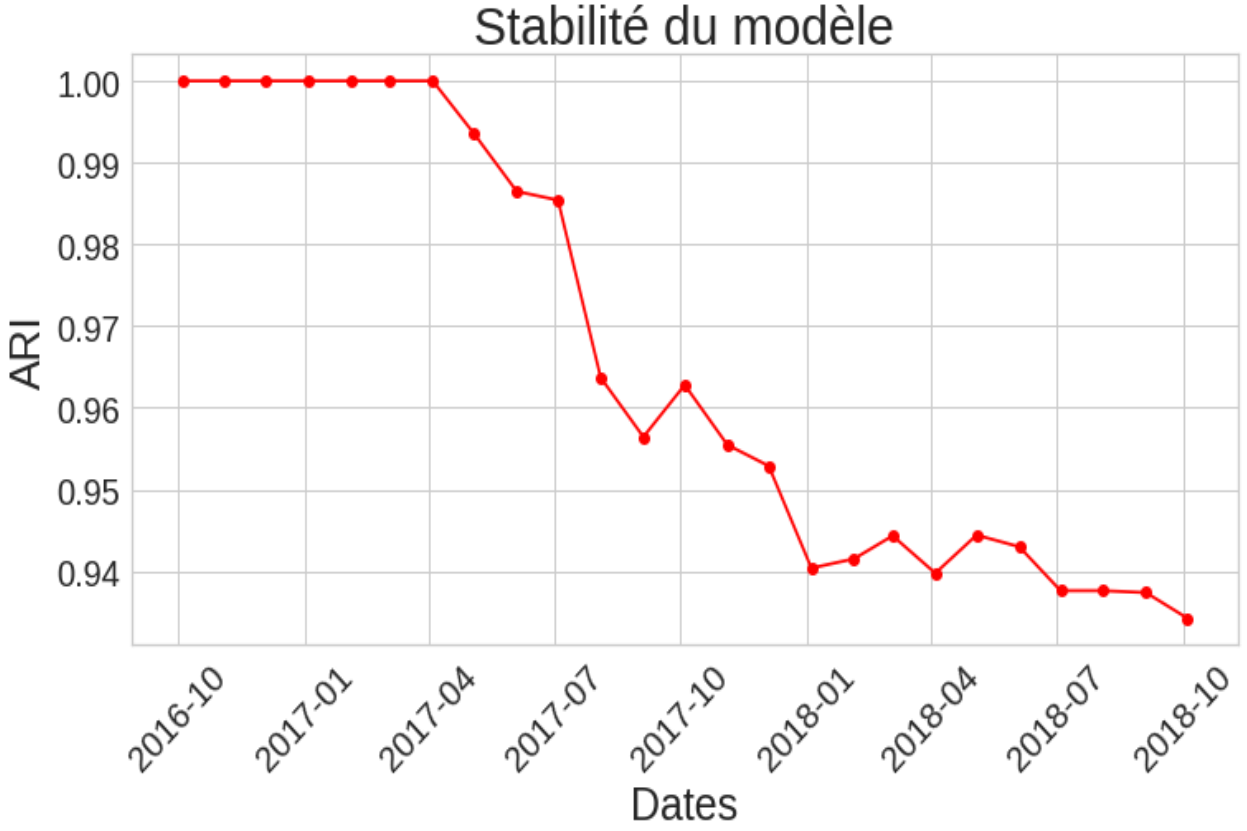
MOYENS CLIENTS

- **Comportement RFM :** achats à fréquence moyenne, montants élevés, dernière commande depuis un grand moment.
- **Préférence de paiement :** en plusieurs fois et par carte de crédit et/ou par voucher.
- **Type de produits consultés :** peu de description, autres d'informations (catégorie, photo) moyennes.
- **Type de vendeurs fréquentés :** vendeurs moyennement loins, avec des frais et temps de livraison assez chers.
- **Retour :** notes moyennes, beaucoup de commentaires.

GROS CLIENTS

- **Comportement RFM :** achats très fréquents, montants élevés, dernière commande depuis un certain temps.
- **Préférence de paiement :** versements moins nombreux et paiement par carte de crédit et/ou par billets.
- **Type de produits consultés :** bonne décription, catégories bien renseignés, avec Qté moyenne de photos.
- **Type de vendeurs fréquentés :** vendeurs plus loins, donc avec des frais et temps de livraison plus conséquents.
- **Retour :** notes moyennes, peu de commentaires.

Maintenance du modèle



- Modèle de base : K-MEANS (n=3) fitté sur des données de référence (6 premiers mois)
- Ajout de données par mois
- Modèle : K-MEANS (n=3) fitté sur les nouvelles données
- Evaluation :
  ⇒ ARI (groupes base % groupes nouv.)

⇒ Modèle stable sur la durée des données

⇒ Tendance négative su fur et à mesure du temps

Améliorations

- Essais d'autres features (selon les besoins, en concertation avec eq. marketing)
- Essais de segmentation avec d'autres features (*ex :* RFM + notes)
- Essais d'autres modèles de segmentation
- Essais d'autres évaluateurs (*ex :* Calinski_Harabasz score)
- Correction aux cas où modèle n'est plus stable :
  ⇒ vérifier minutieusement les données à partir de la date,
  ⇒ réaliser un autre modèle à partir de cette date.

MERCI

Questions

Incompréhensions