Projet 5

Segmentez des clients d'un site de e-commerce

Problématique / Données / Modélisations / Conclusions

Camille BRODIN

Problématique

Missions confiées par Olist

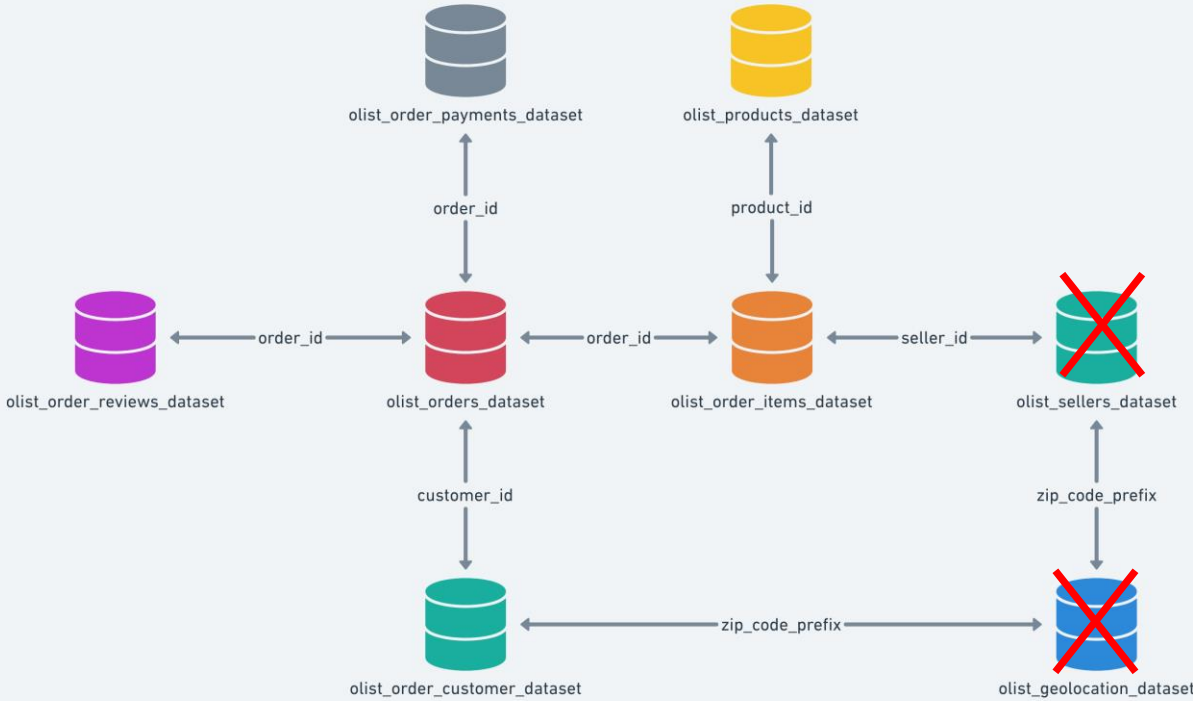
Base de données :

- Historique de commandes, produits achetés, commentaires de satisfaction, numéro de clients depuis janvier 2017.

Trois missions :

- Comprendre les différents types d'utilisateurs grâce à leur comportement et à leurs données personnelles.
- Segmenter ces profils clients.
- Fournir une description actionnable de la segmentation + une proposition de contrat de maintenance (analyse de la stabilité des segments au cours du temps).

Données : Nettoyage



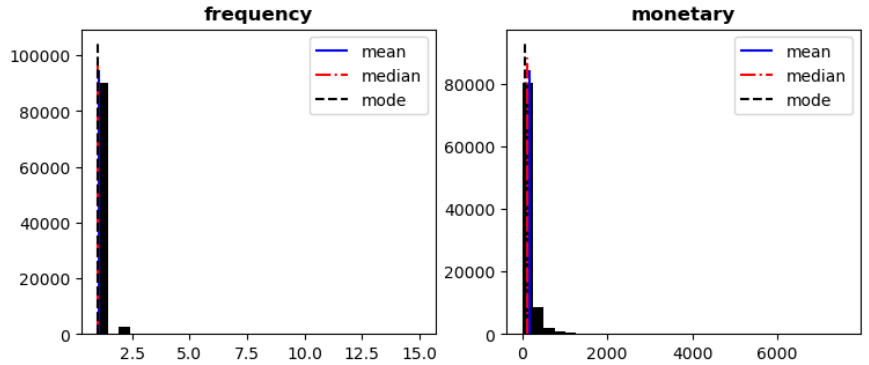
Justifications	Méthodes
0. Merge des datasets et regroupements par clients	df = pd.merge(orders,order_payments, on="order_id") df = pd.merge(df,customers, on="customer_id")..
1. Filtrage projet : commandes non livrées	orders.loc[orders['order_status'] == "delivered", :]
2. Elimination des lignes/cols inexploitables	data.drop(index=data[data[]==0].index) data.drop(columns=[“x”])
3. Outliers : Exploration individualisée et connaissances métiers	products.loc[products[‘weight']<= 40kg] -> table bébé ok orders.drop([13390]) outlier de 13000e pour de la téléphonie fixe df.iloc[i-1:i+2,:] -> 10 meubles 184kg ok cohérent
4. Imputations des données Remplacer par la valeur ‘others’ ou 0	products['product_category'].fillna("others") order_reviews['review'].str.len().fillna(0)

- 99441 lignes et 43 variables **-> 92753 lignes et 16 features sur le jeu de données nettoyé (comprenant les ID)**
- Pour des raisons de praticité, nous n’utiliserons que 20% du dataset, sélectionné au hasard

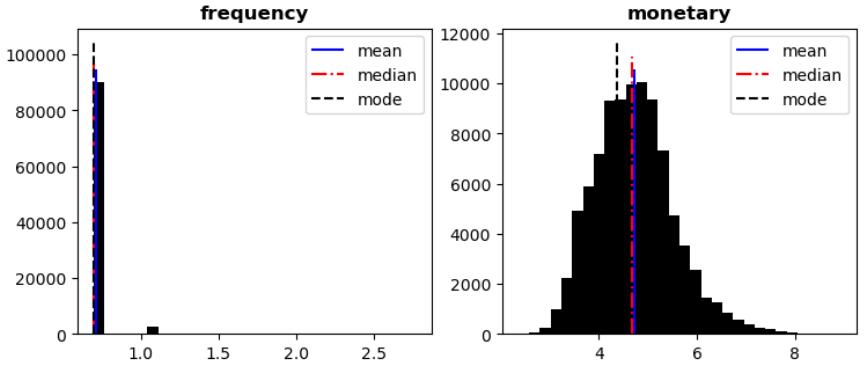
Données : analyses univariées

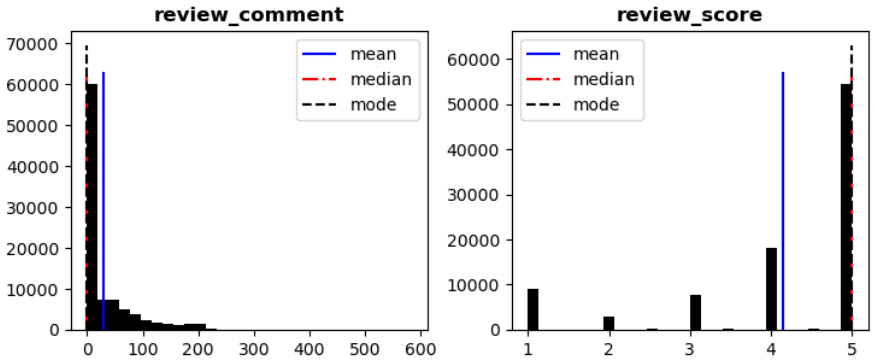
❖ **Variables quantitatives :**

→ **Test de normalité (histogramme, Shapiro-Wilk)**

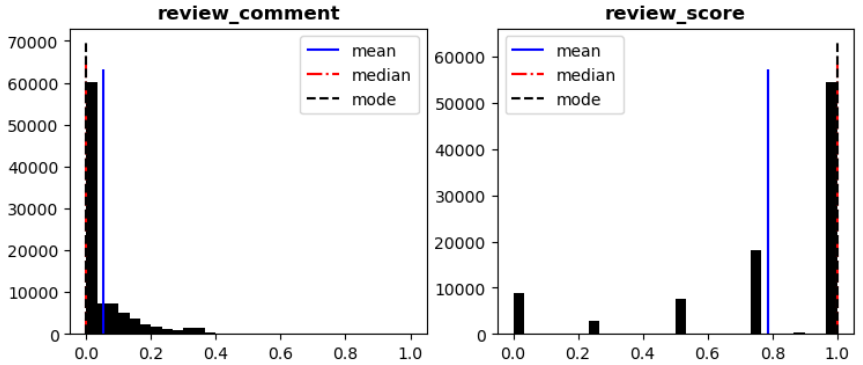


Logarithme





MinMaxscaler

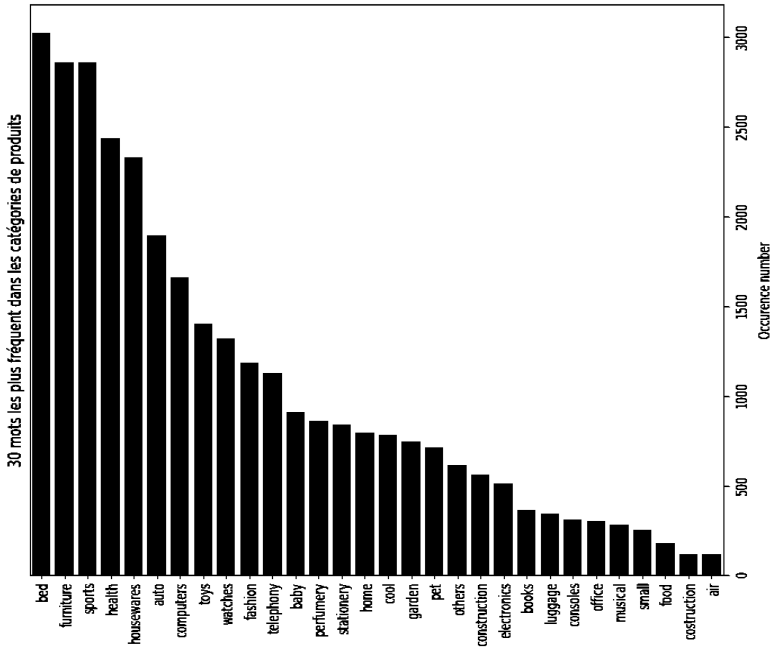


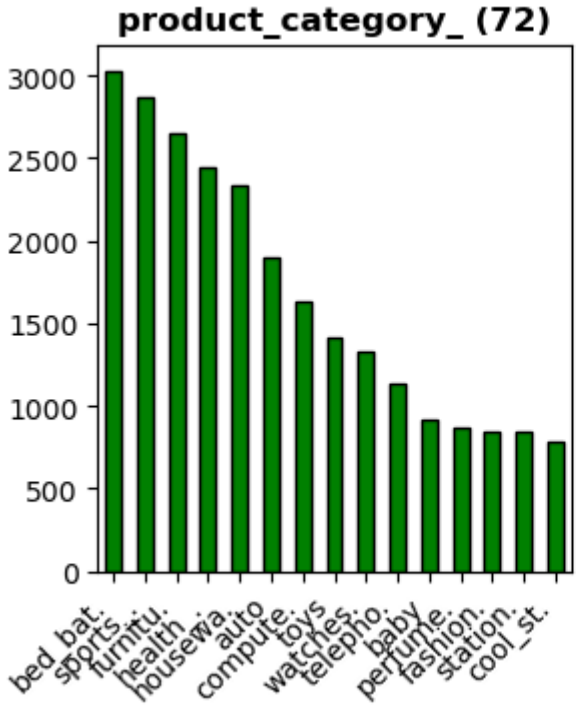
- **Distributions non gaussiennes très rassemblées sur la gauche dans les variables d'études p<0,05.**
- **Nous testons deux conditions sur le df : transformation logarithme népérien +1 (haut) ou scaling MinMax (bas)**

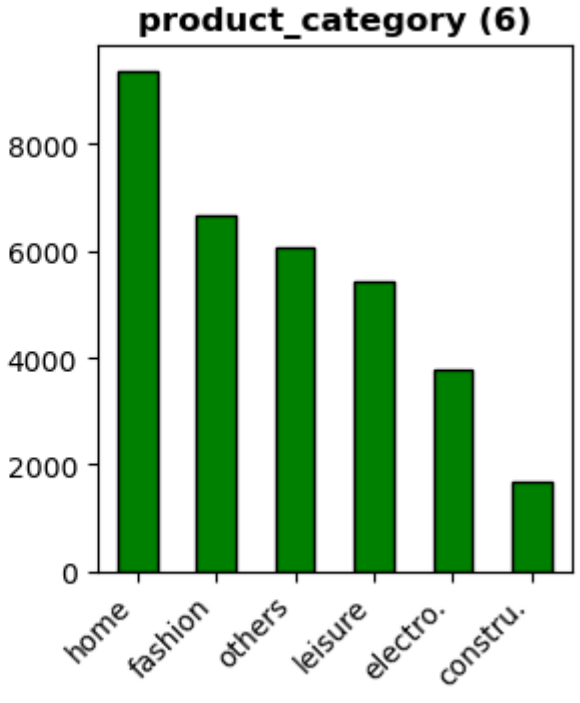
Données : analyses univariées

❖ **Variables qualitatives**

→ **Création de 6 sur-catégories**



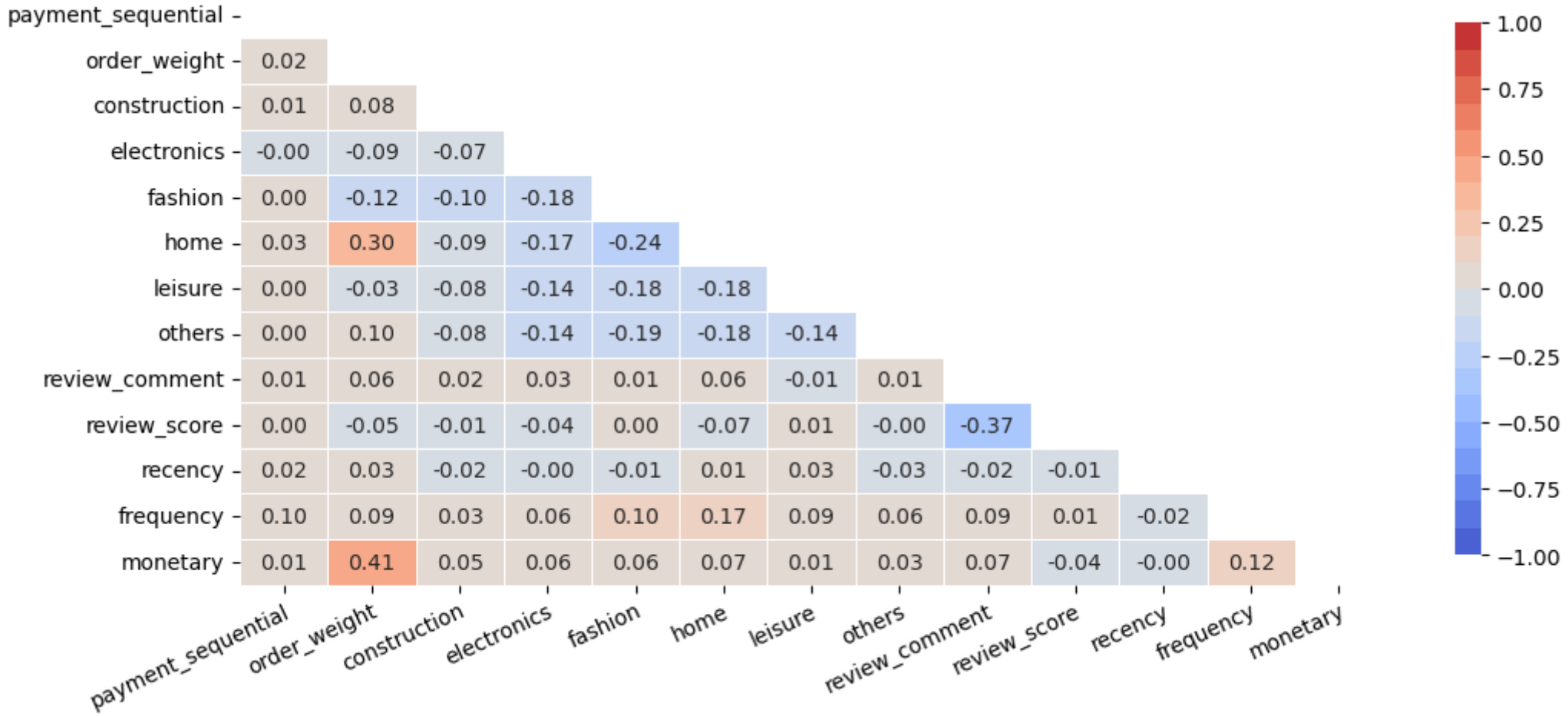




➢ **Création d'une variable catégorie plus large : maison, mode, loisir, électronique, construction**

Données : analyses bivariées

❖ **Corrélation de Pearson**



Données : Sélection et création de variables (feature engineering)

1/ Dataset « segmentation RFM »

- Récence du dernier achat (1)
- Fréquence d'achat (1)
- Montant total dépensé (1)
- RFM catégorie de profil obtenu (1)
- Date de dernière commande (1)

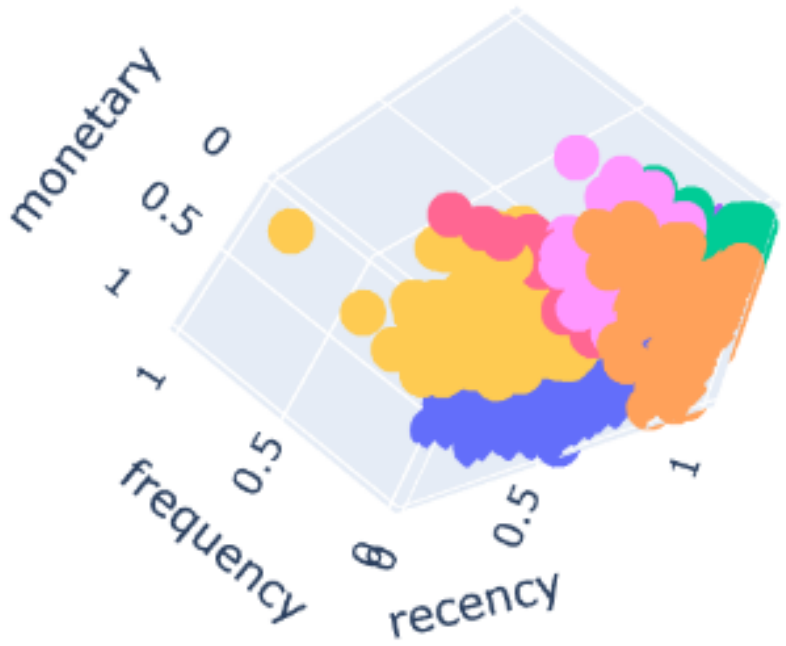
2/ Dataset « complet »

- Catégories de produits (6)
- Satisfaction (1)
- **Méthode de paiement (1)**
- Taille des commandes (1)
- Profil RFM (5)

FEATURES

'customer_unique_id',
'construction',
'electronics',
'fashion',
'home',
'leisure',
'others',
'review_score',
'payment_sequential',
'order_weight',
'recency',
'frequency',
'monetary',
'rfm_score_label'
'order_purchase_timestamp',

Données : Sélection et création de variables (feature engineering)



color

promising	32387
new_customers	18114
lost_customers	7735
hibernating_customers	7431
cannot_lose_them	17492
about_to_sleep	7163
loyal	497
potential_loyalist	324
at_risk	641
champions	969

- **10 profils standards obtenus par score, à titre indicatif.**
- **Modèles de clustering à suivre -> groupes homogènes, moins nombreux => pour une meilleure interprétation business.**

Modélisation : pre-processing et évaluation

ENCODAGE CATEGORIES

- **OneHotEncoder** : catégories produits

TRANSFORMATIONS FEATURES

- **MinMaxScaler**
- **Logarithme népérien (+1)**

VALIDATION CROISEE

- *GridSearchCV* -> **cv=None -> hyperparamètres optimisés**

Metric pour évaluler nos modèles de clustering:

1/ Score de distorsion : somme moyenne des distances quadratiques entre les centres (méthodes du coude, utilisé pour déterminer k)

2/Fit time : tps d'entrainement, plus court mieux c'est.

3/Indice de Rand ajusté (ARI) : mesure la similarité des deux affectations (0:incorrect, +1 parfait).

4/ Coefficient Silhouette: rapport moyen entre la distance intra-cluster et la distance entre les clusters les plus proches (-1:incorrect, +1 très dense).

5/ Indice de Calinski-Harabasz: ratio de la répartition des clusters entre eux et à l'intérieur d'un même cluster (score élevé -> clusters mieux définis).

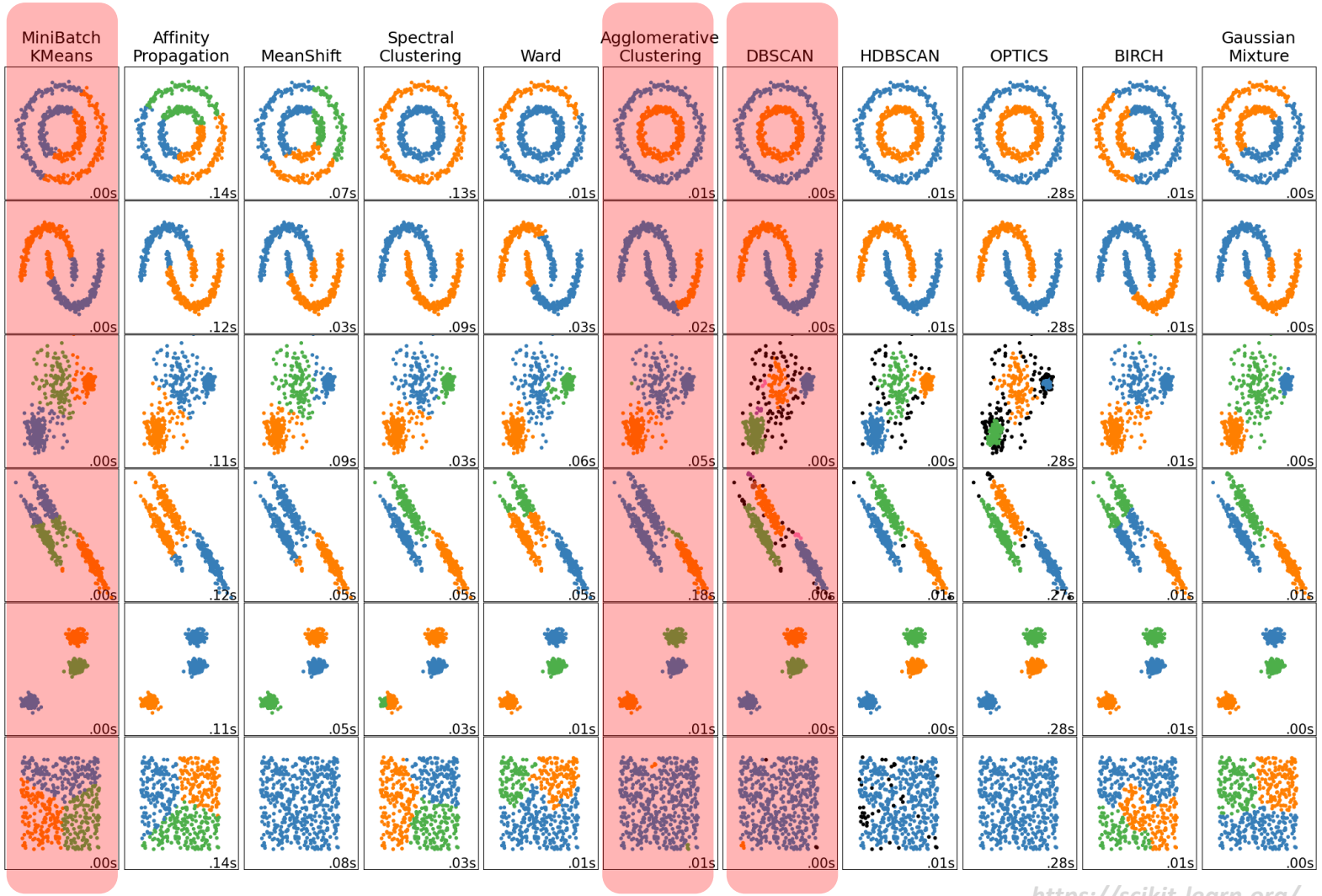
Modélisation

Problématique

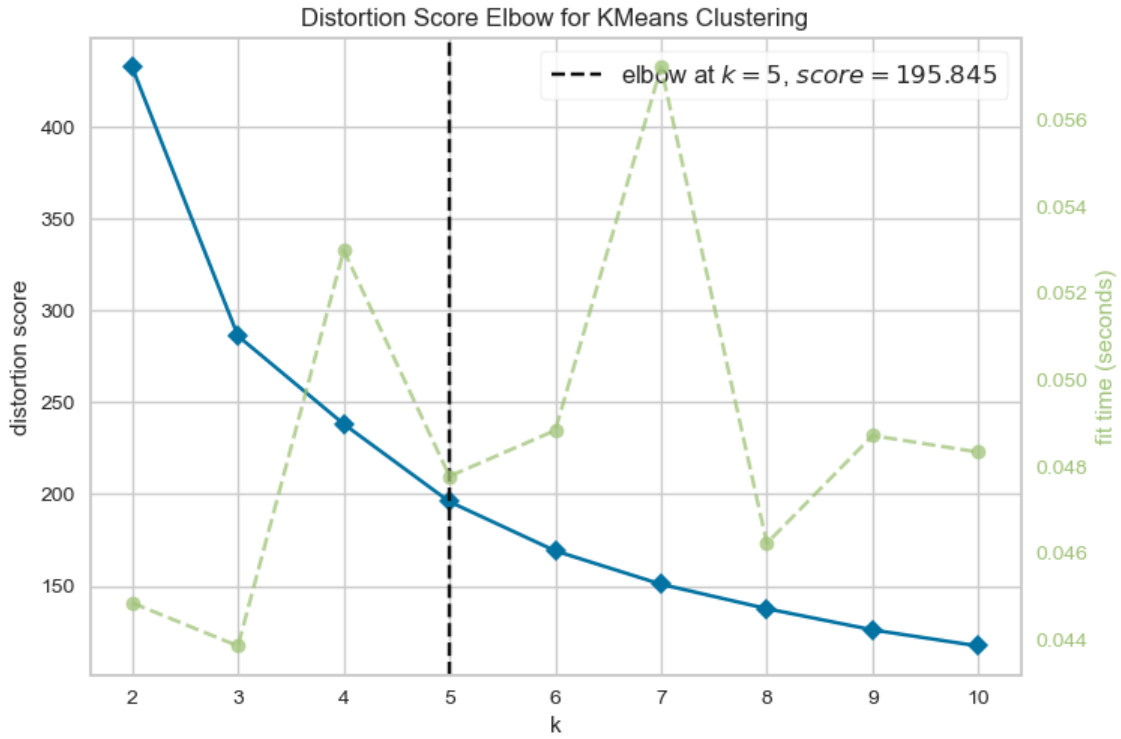
Données

Modélisation

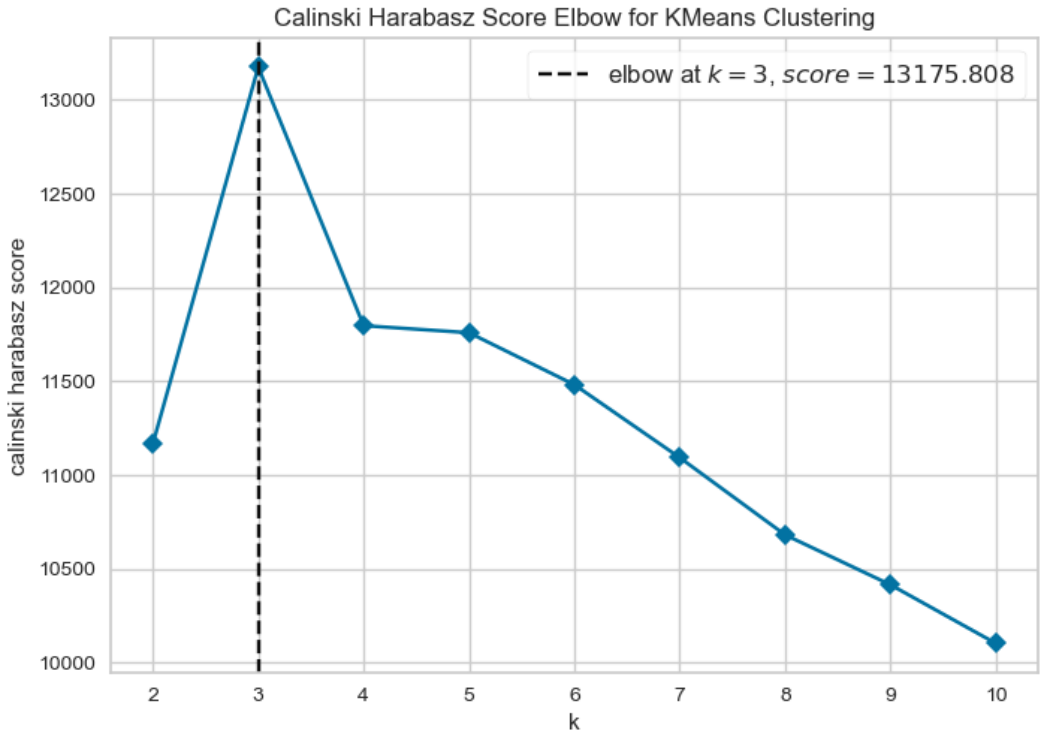
Conclusions



Modélisation 1 : kmeans avec le dataset condensé RFM

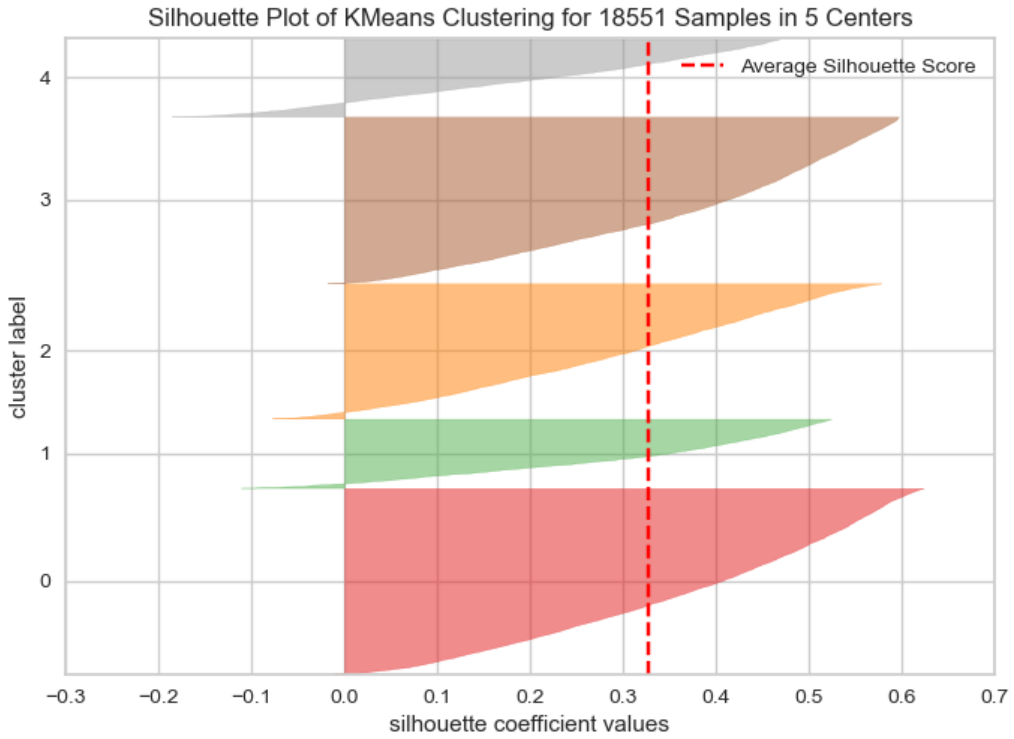


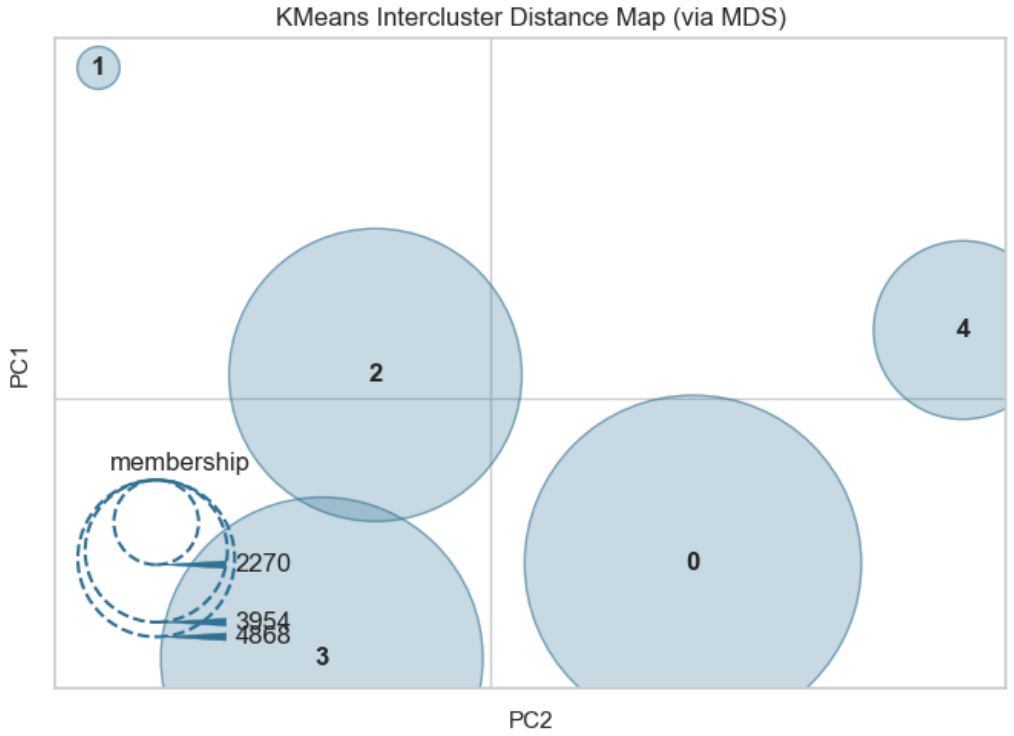
➢ **Score de distorsion -> k = 5**



➢ **Indice de Calinski-Harabasz-> k = 3**

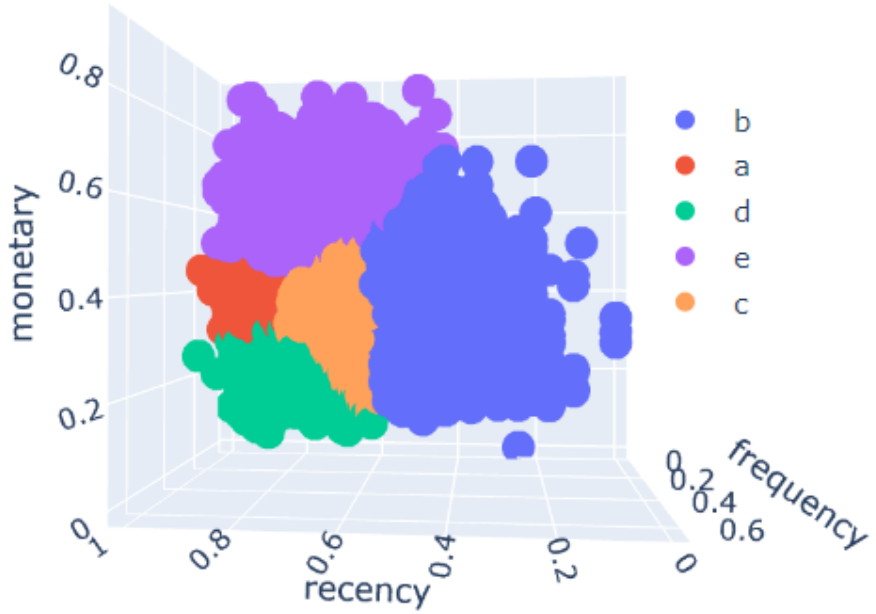
Modélisation 1 : kmeans avec le dataset condensé RFM

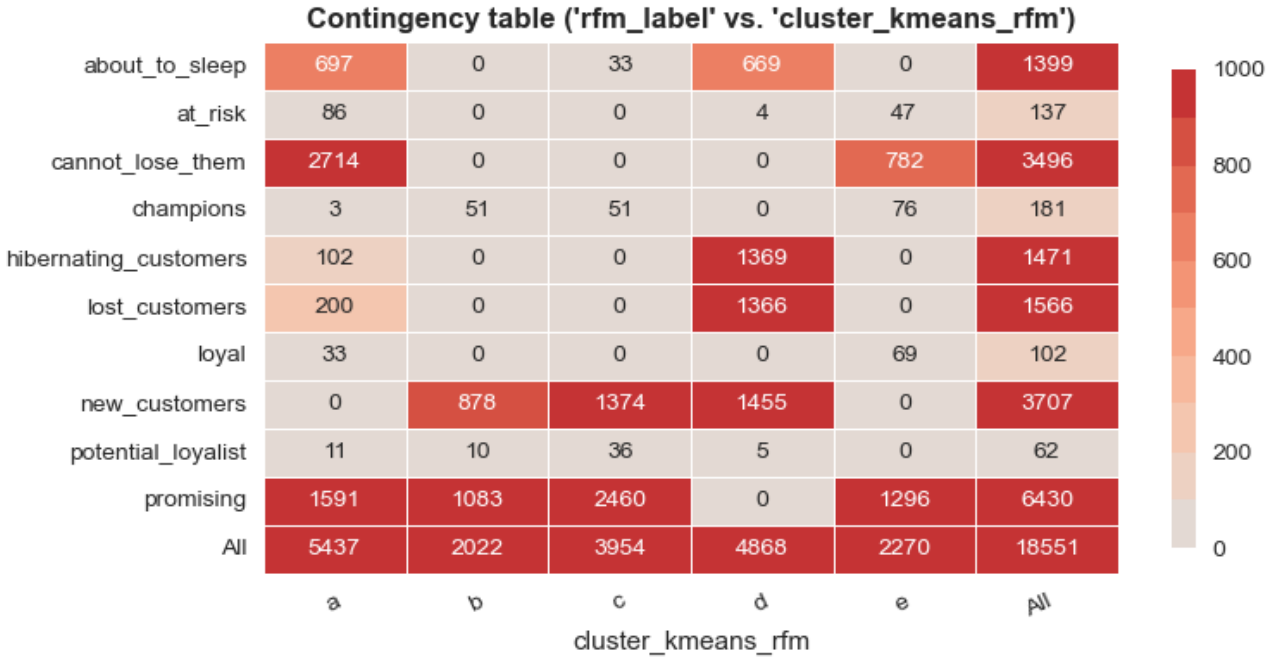




➢ ***Forme des clusters*** : on utilise le coefficient de silhouette : 0,31 ; clusters assez denses et assez bien séparés.

Modélisation 1 : Caractérisation des clusters

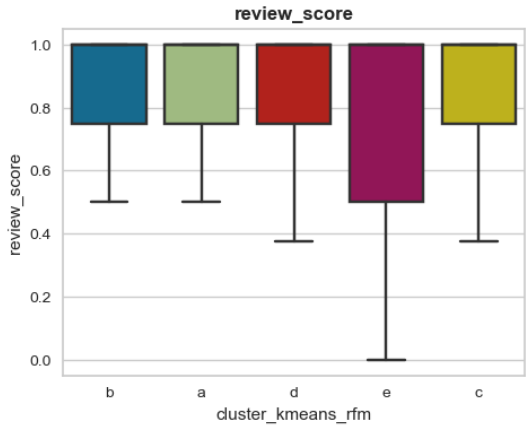


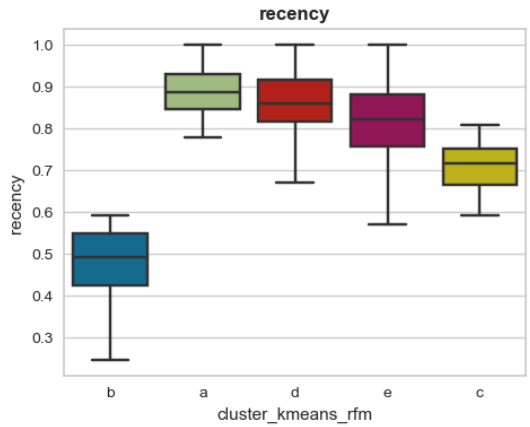


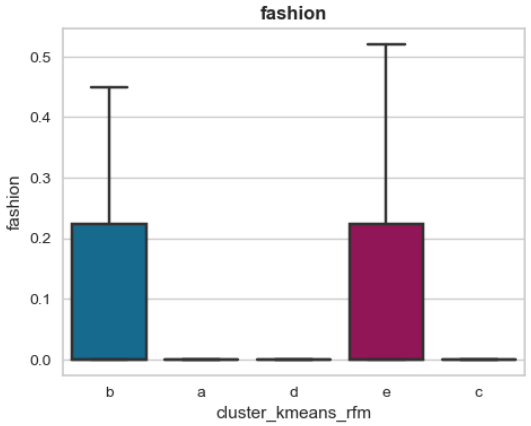
➢ ***Compatibilité des résultats avec des connaissances spécifiques au domaine*** :

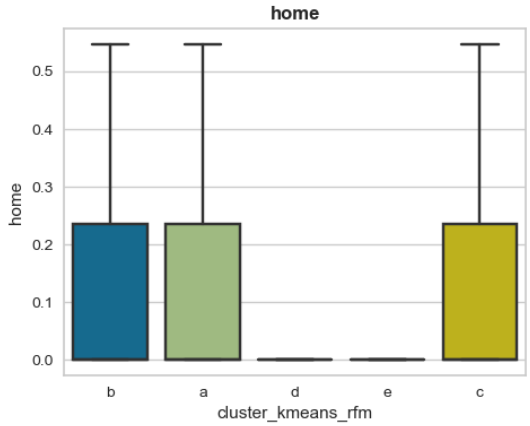
(a) "new_promising_need_stimulation“ = rouge
(b) "loyal_asleep“ = bleu
(c) "new_promising_loyal_in_futur_high_potential“ = orange
(d) "already_lost_no_potential« = vert
(e) "best_champions« = violet

Modélisation 1 : Caractérisation des clusters









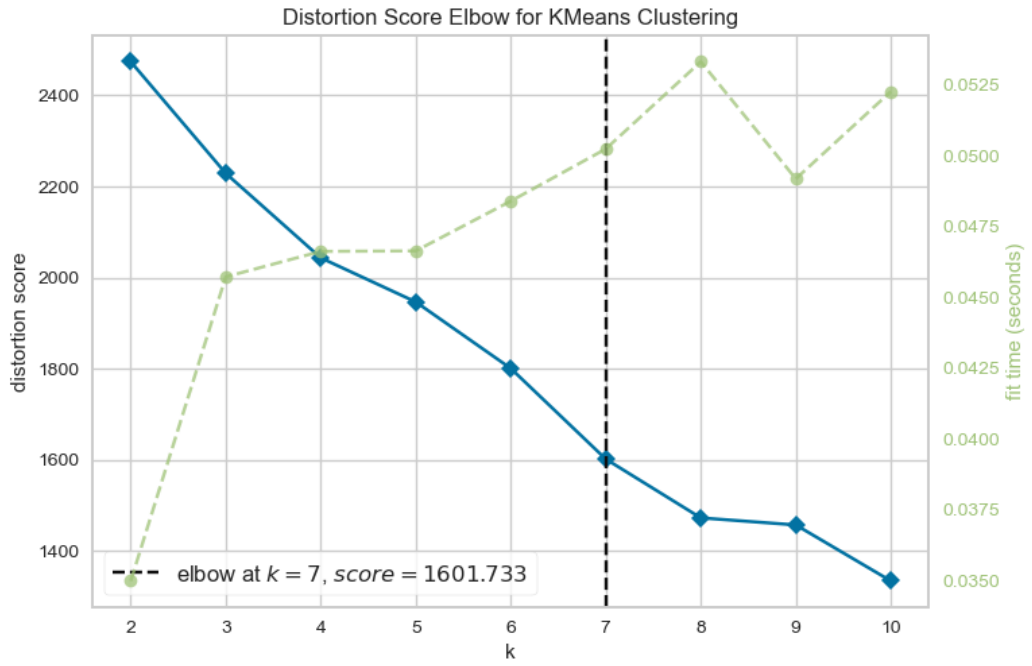
- ***Compatibilité des résultats avec des connaissances spécifiques au domaine (suite)***

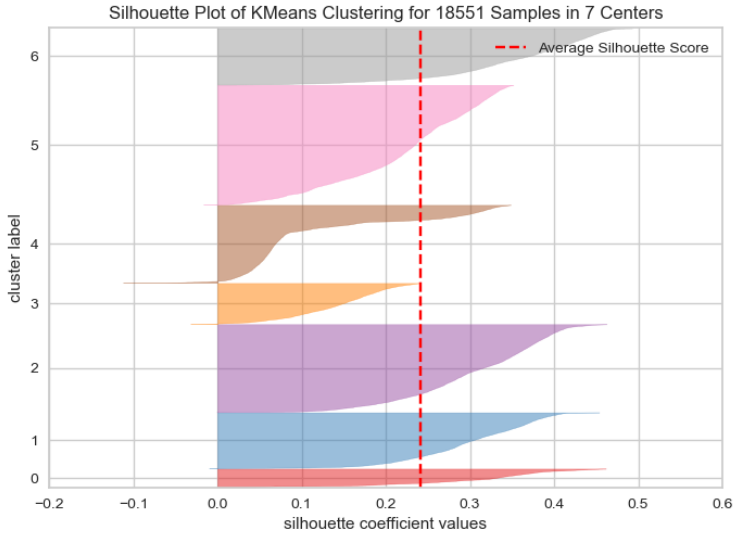
- ***Stabilité de l'algorithme***: on utilise l'Indice de Rand ajusté (ARI) : mesure la similarité de deux affectations (0 : incorrect, +1 parfait).

ARI moyen après 10 initialisations = 0,9861

Excellente stabilité de l'algorithme à l'initialisation

Modélisation 2 : kmeans avec le dataset complet



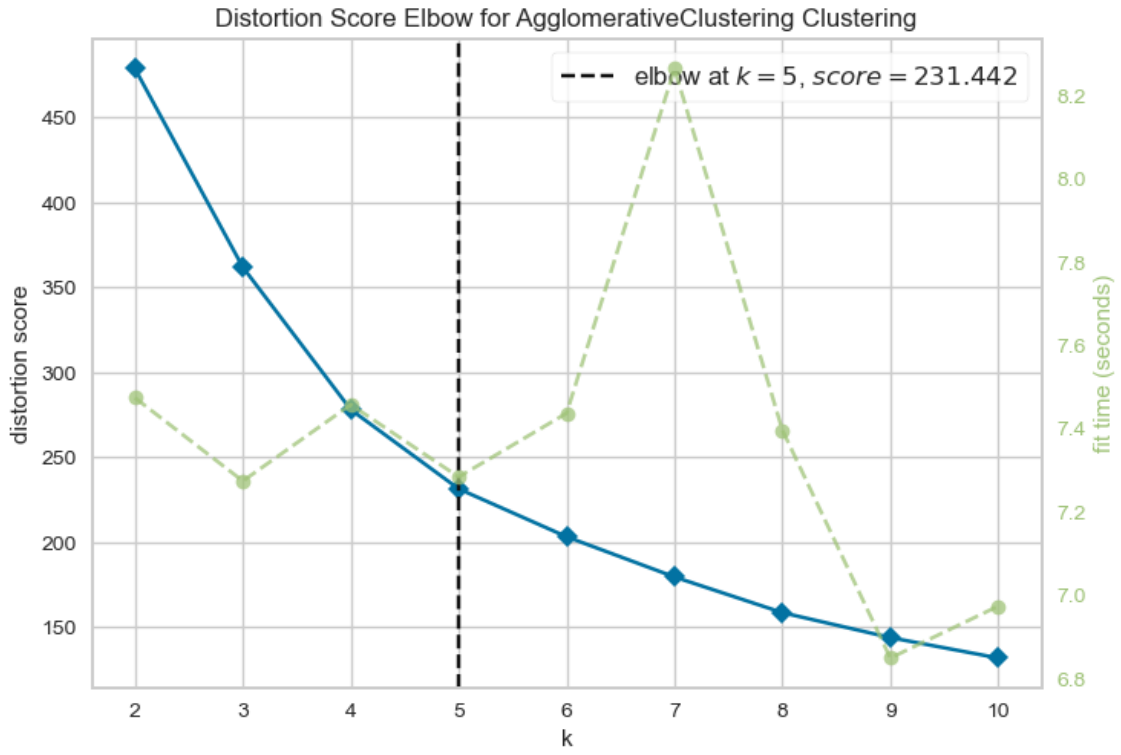


	a	b	c	d	e	f	g	All
about_to_sleep	45	173	260	154	242	336	189	1399
at_risk	11	12	37	3	17	31	26	137
cannot_lose_them	184	390	745	283	550	884	460	3496
champions	10	27	54	7	24	29	30	181
hibernating_customers	43	155	232	124	289	409	219	1471
lost_customers	55	167	313	103	265	460	203	1566
loyal	11	9	26	7	19	16	14	102
new_customers	102	483	619	293	686	1066	458	3707
potential_loyalist	2	14	16	2	13	9	6	62
promising	289	841	1263	683	1052	1615	687	6430
All	752	2271	3565	1659	3157	4855	2292	18551

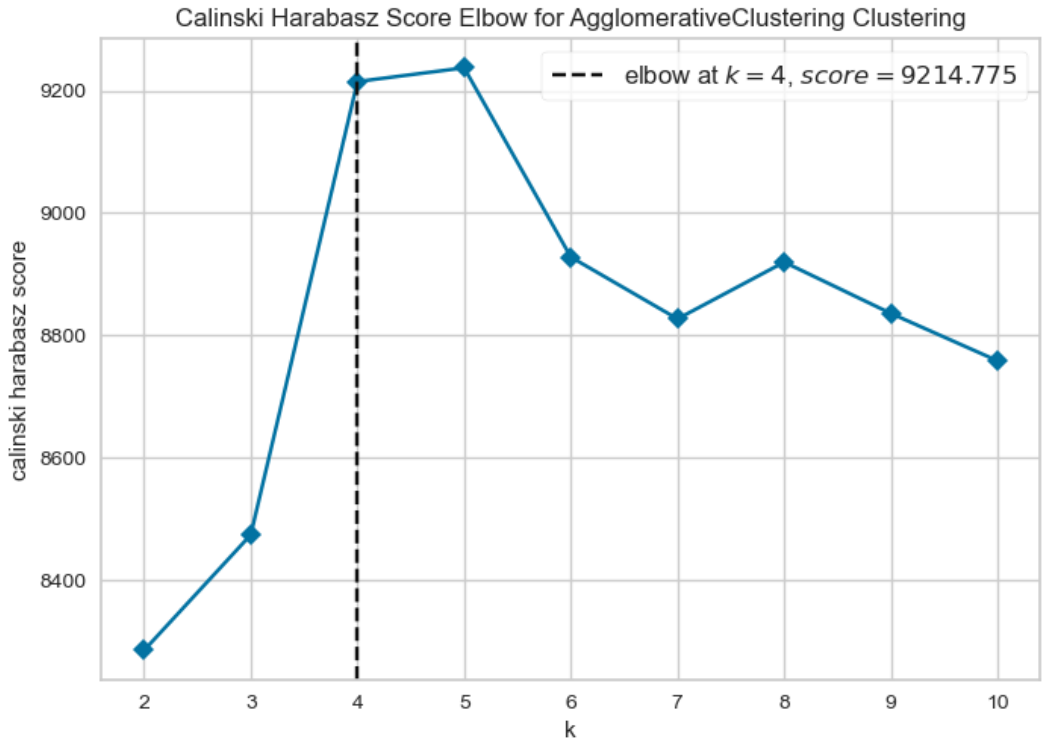
cluster_kmeans_total

- ***Forme des clusters*** : clusters assez peu denses et assez bien séparés.
- ***Incompatible avec les connaissances spécifiques au domaine.***

Modélisation 3 : Agglomerative clustering sur le dataset rfm

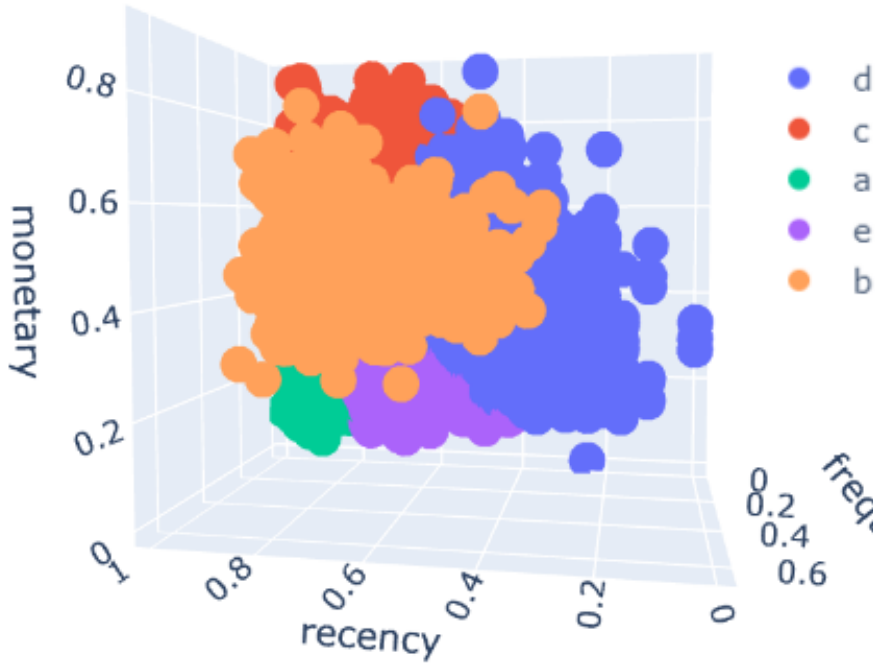


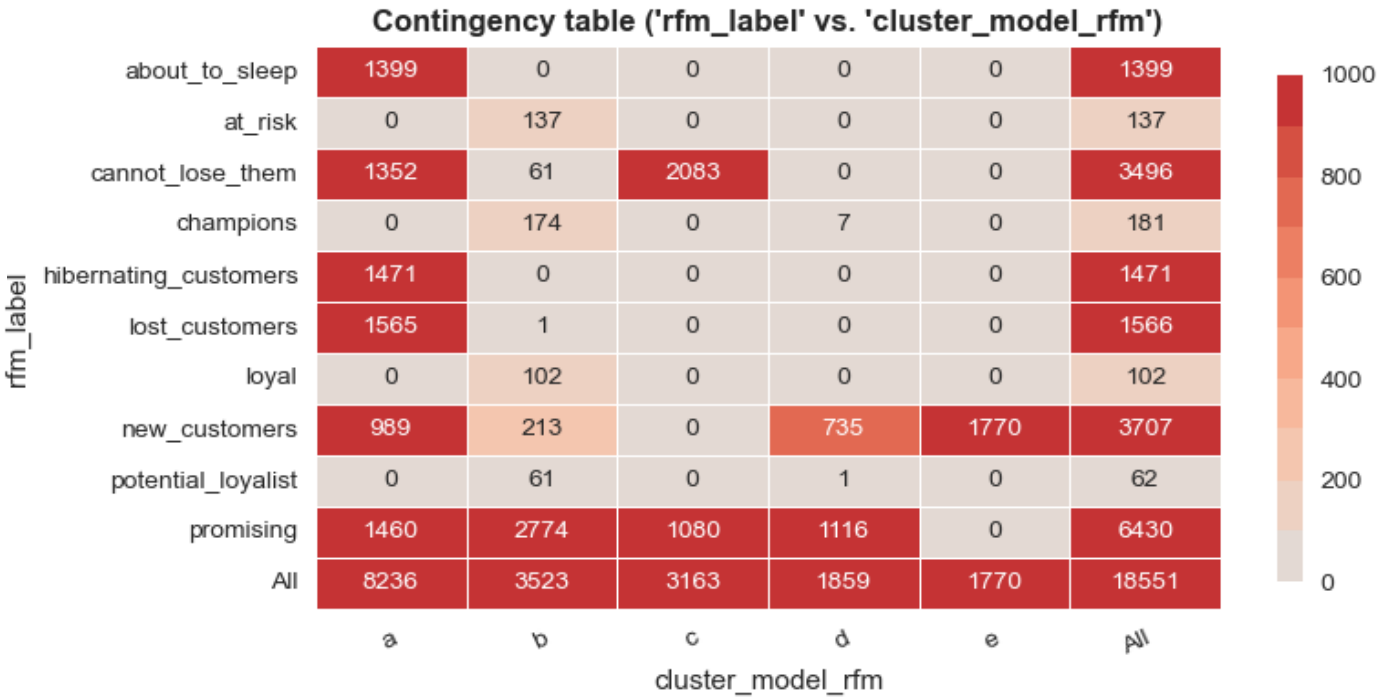
➢ **Score de distorsion -> k = 5**



➢ **Indice de Calinski-Harabasz-> k = 4**

Modélisation 3 : Agglomerative clustering sur le dataset rfm

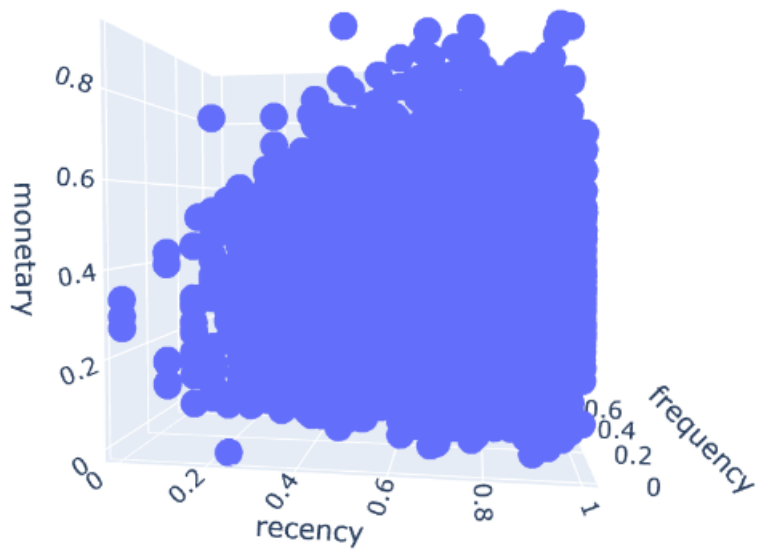


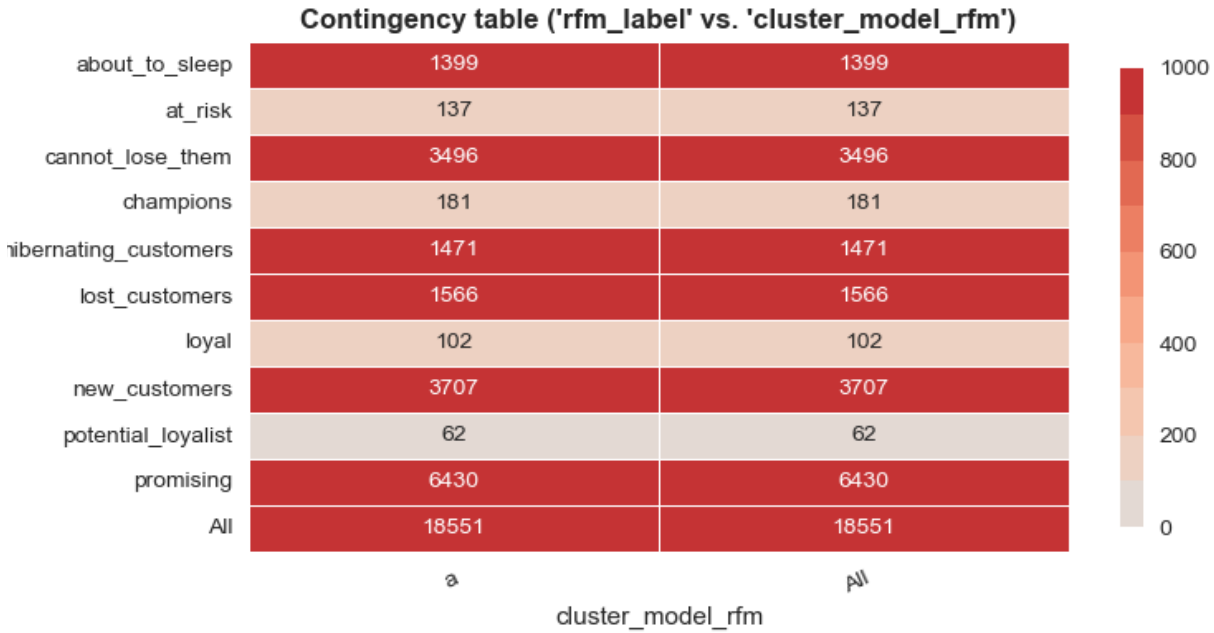


➢ ***Compatibilité des résultats avec des connaissances spécifiques au domaine*** :

(a) “already_lost_no_potential“
(b) “champions"
(c) “futur loyalist“
(d) et (e) environ égaux « new customers »

Modélisation 4 : algorithme DBSCAN sur le dataset rfm





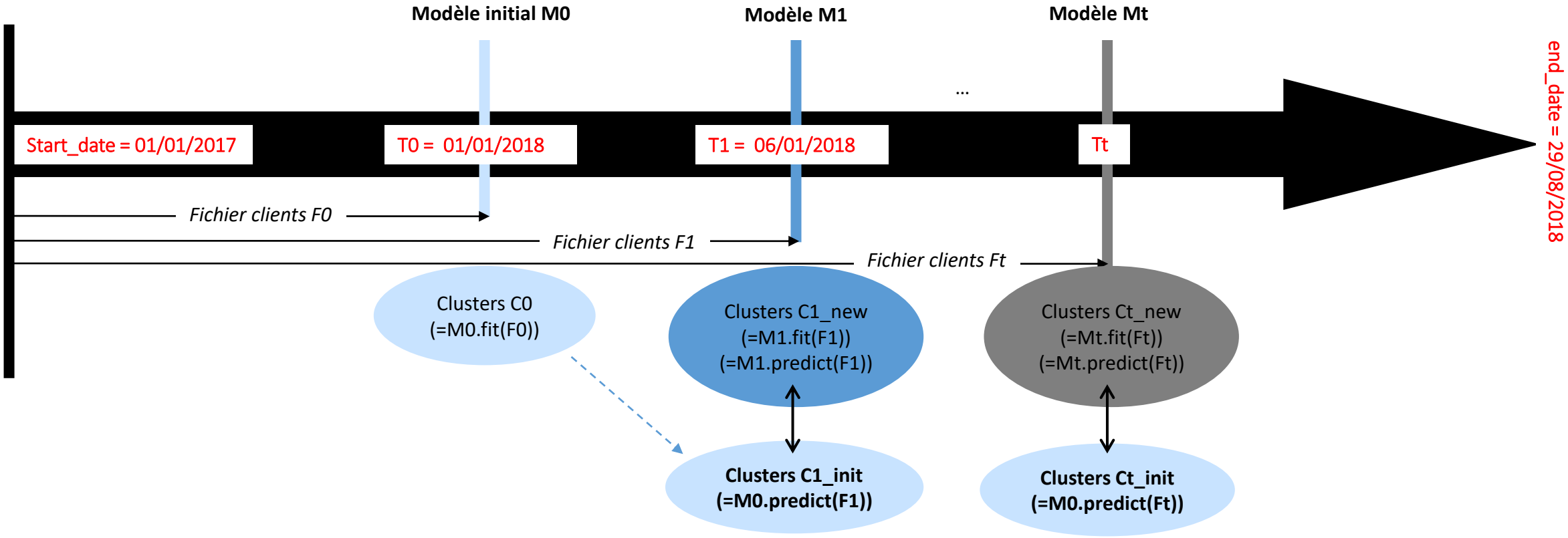
➢ **Un seul cluster en sortie, cet algorithme n'est pas adapté à notre problématique métier.**

Modélisation : modèle sélectionné = kmeans rfm

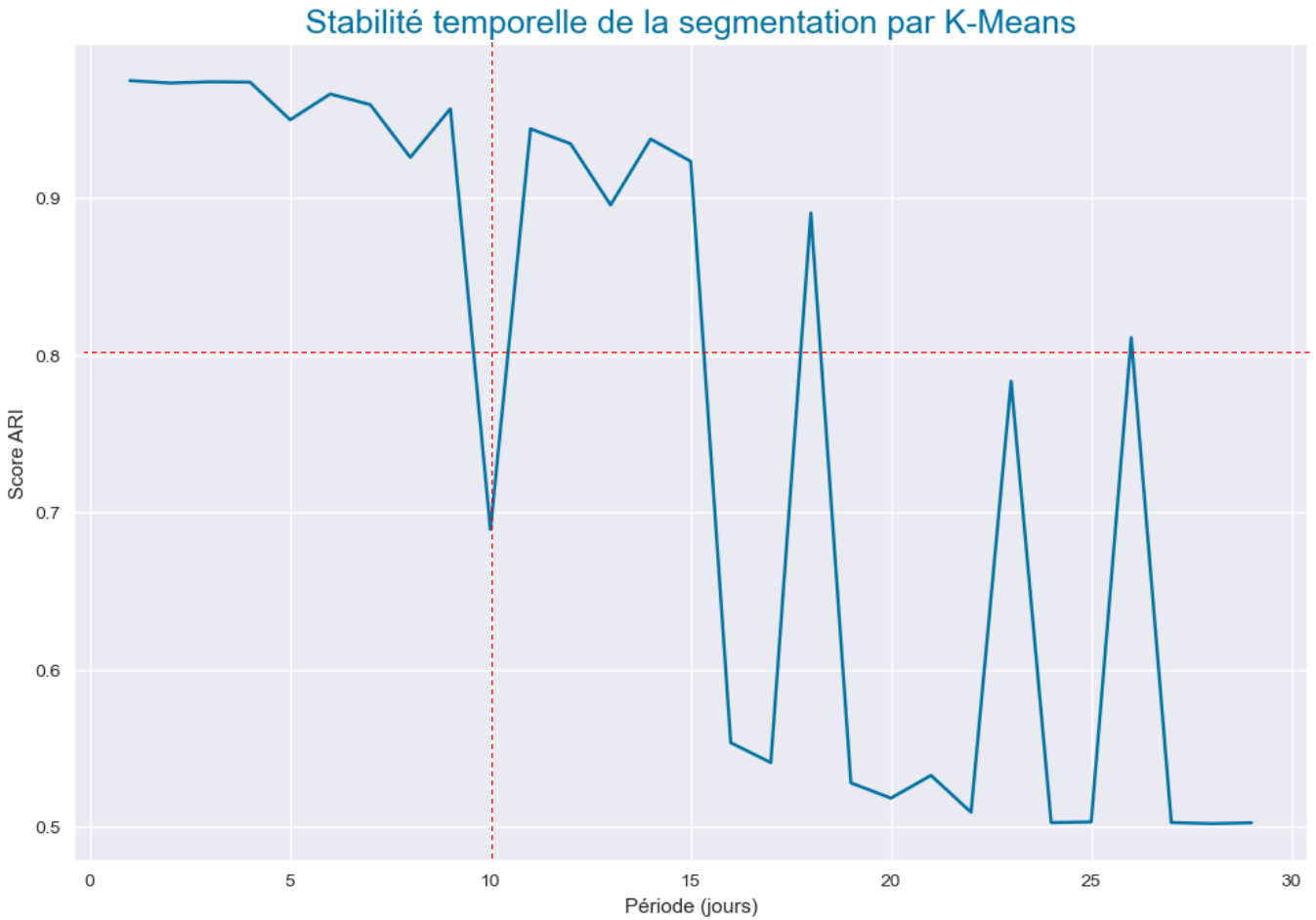
Caractéristique modèle	Kmeans (rfm)	Kmeans (complet)	Agglomerative clustering	DBSCAN
Nombre clusters	5	7	5	NA
Temps de calcul	0,05 sec	0,05 sec	7,9 sec	NA
Distorsion score	195	1601	240	NA
Calinski-harabasz	11750	4800	9250	NA
Silhouette coef	0,32	0,24	NA	NA

Modélisation : Evaluation de la stabilité dans le temps

Objectif = déterminer la fréquence à laquelle la segmentation doit être mise à jour pour rester pertinente, afin de pouvoir effectuer un devis de contrat de maintenance.



Modélisation : mise à jour conseillée = 10 jours



- ✓ Lorsque l'ARI passe en dessous de 0.8, il est pertinent de reproposer un entraînement de modèle au client.
- ✓ Fréquence à laquelle la segmentation doit être mise à jour : tous les 10 jours.
- ❖ **Délais de maintenance très court.**

Conclusion générale

Trois missions remplies :

- Comprendre les différents types d'utilisateurs grâce à leur comportement et à leurs données personnelles.
- Segmenter ces profils clients.
- Fournir une description actionnable de la segmentation + une proposition de contrat de maintenance -> **tous les 10 jours.**

- ✓ Adapter les hyperparamètres d'un algorithme non supervisé afin de l'améliorer.
- ✓ Évaluer les performances d'un modèle d'apprentissage non supervisé.
- ✓ Transformer les variables pertinentes d'un modèle d'apprentissage non supervisé.
- ✓ Mettre en place le modèle d'apprentissage non supervisé adapté au problème métier.

Merci pour votre attention

ANNEXES

Segment	Activity	Actionable Tip
Champions	Bought recently, order often and spend the most.	Reward them. Can be early adopters of new products. Will promote your brand. Most likely to send referrals.
Loyal	Orders regularly. Responsive to promotions.	Upsell higher value products. Ask for reviews.
Potential Loyalists	Recent customers who spent good amounts.	Offer membership / loyalty program. Keep them engaged. Offer personalized recommendations.
New Customers	Bought most recently.	Provide on-boarding support, give them early access, start building relationship.
Promising	Potential loyalist a few months ago. Spends frequently and a good amount. But the last purchase was several weeks ago.	Offer coupons. Bring them back to the platform and keep them engaged. Offer personalized recommendations.
Need attention	Core customers whose last purchase happened more than one month ago.	Make limited time offers. Offer personalized recommendations.
About to sleep	Made their last purchase a long time ago but in the last 4 weeks either visited the site or opened an email.	Make subject lines of emails very personalized. Revive their interest by a specific discount on a specific product.
Cannot Lose Them	Made the largest orders, and often. But haven’t returned for a long time.	Win them back via renewals or newer products, don’t lose them to competition. Talk to them if necessary. Spend time on highest possible personalization.
At Risk	Similar to 'Cannot Lose Them' but with smaller monetary and frequency value.	Provide helpful resources on the site. Send personalized emails.
Hibernating customers	Customers who made smaller and infrequent purchases before but haven't purchased anything in a long time.	Include them in your standard email communication but regularly check if they don't flag your content as spam. Do not overspend on this segment.
Lost	Made last purchase long time ago and didn’t engage at all in the last 4 weeks.	Revive interest with reach out campaign. Ignore otherwise.