Segmentez des clients d'un site e-commerce

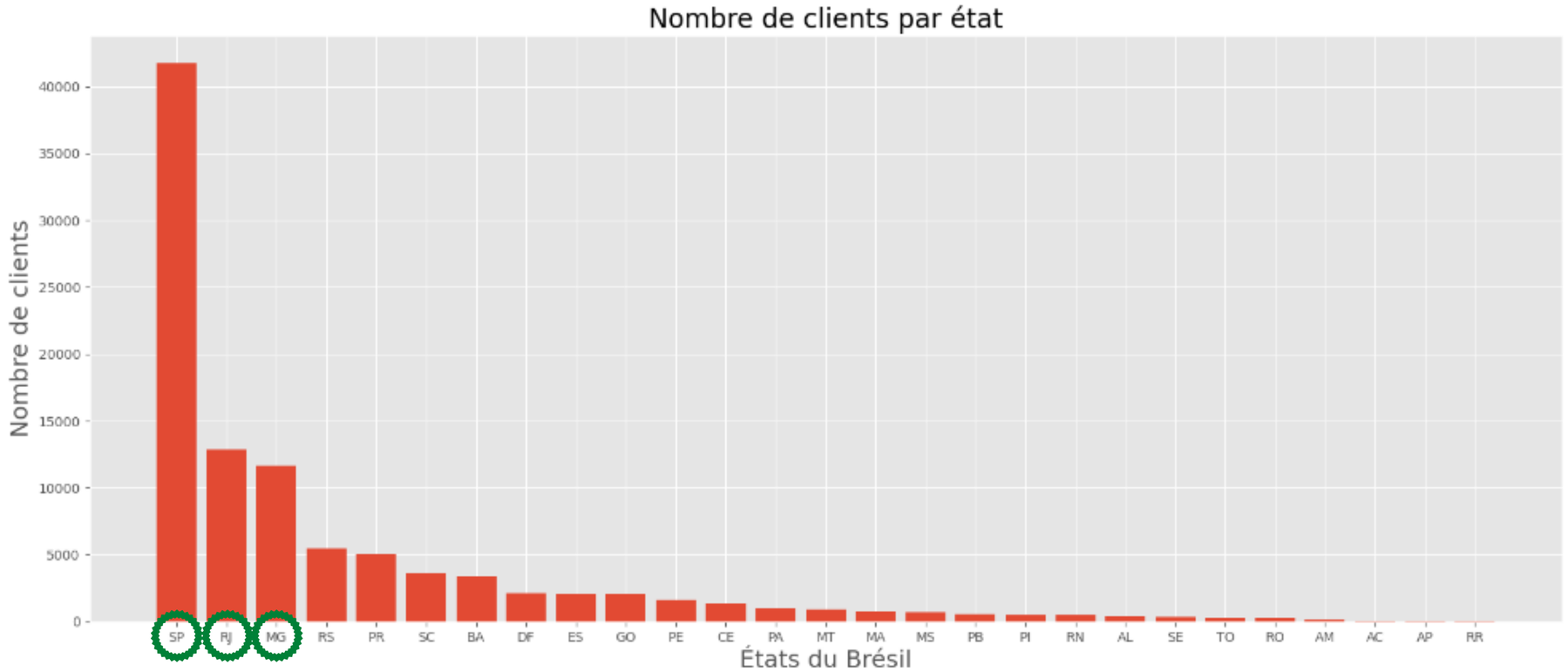
Objectifs

- Comprendre les différents types d'utilisateurs
- Recommander une fréquence de MAJ de la segmentation

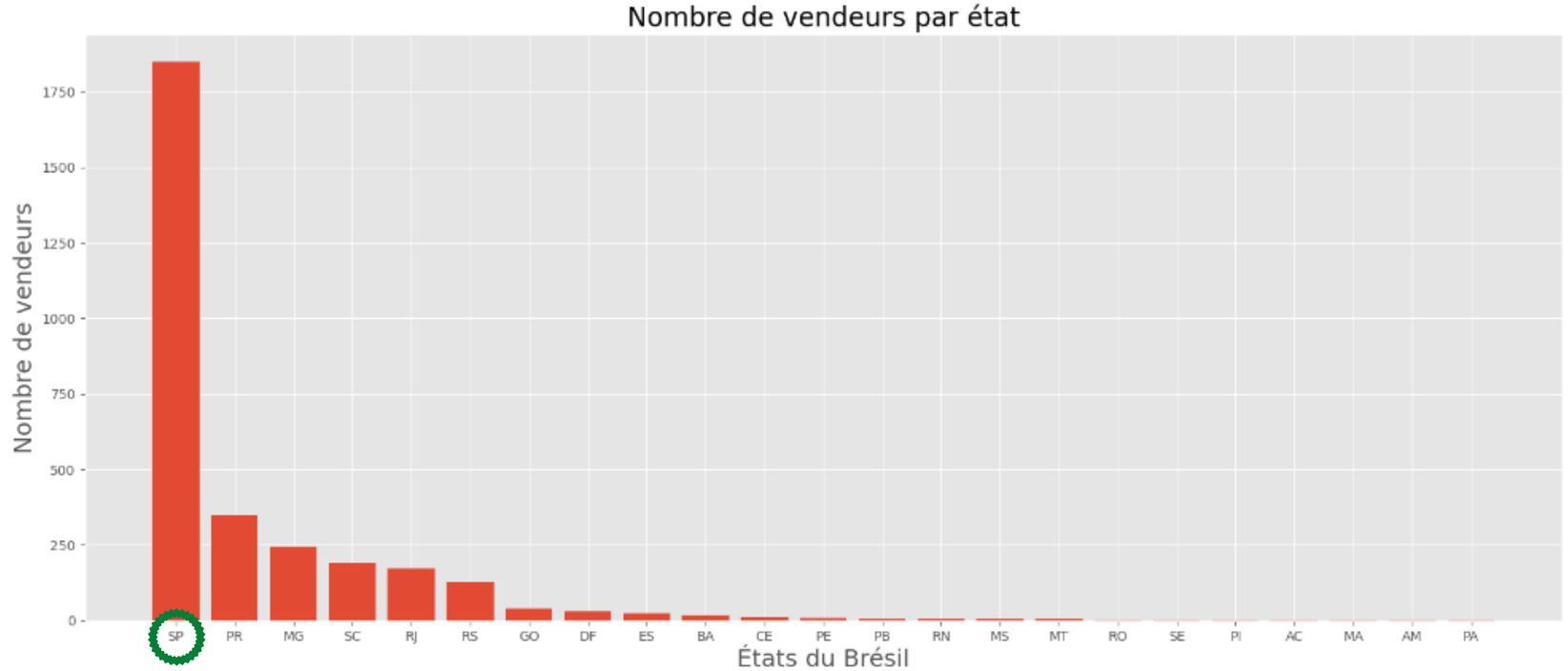
Présentation des 9 jeux de données

www.kaggle.com/datasets/olistbr/brazilian-ecommerce

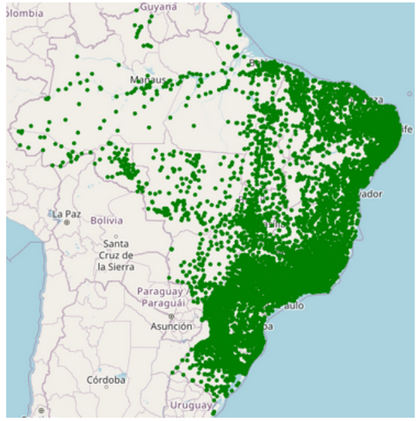
1 - Étude des clients (customers)



2 - Étude des vendeurs (sellers)

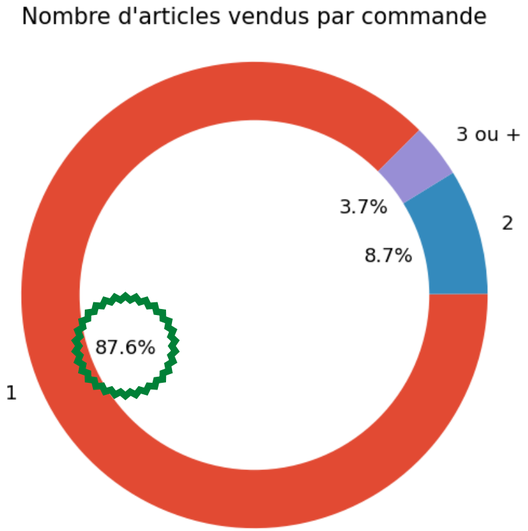


3 - Étude de la géolocalisation (geolocation)

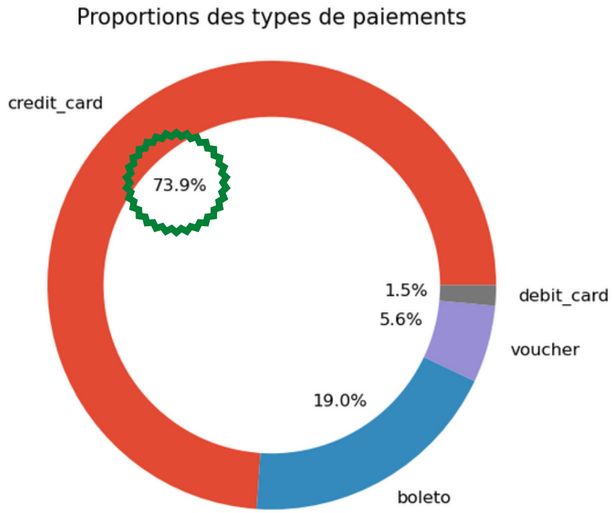


4 - Étude des commandes (orders)

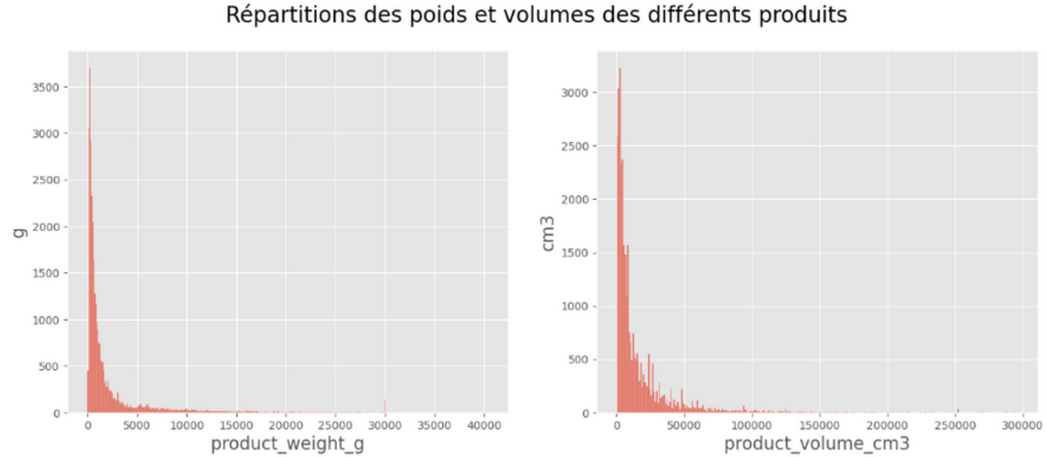
5 - Étude des objets vendus (items)



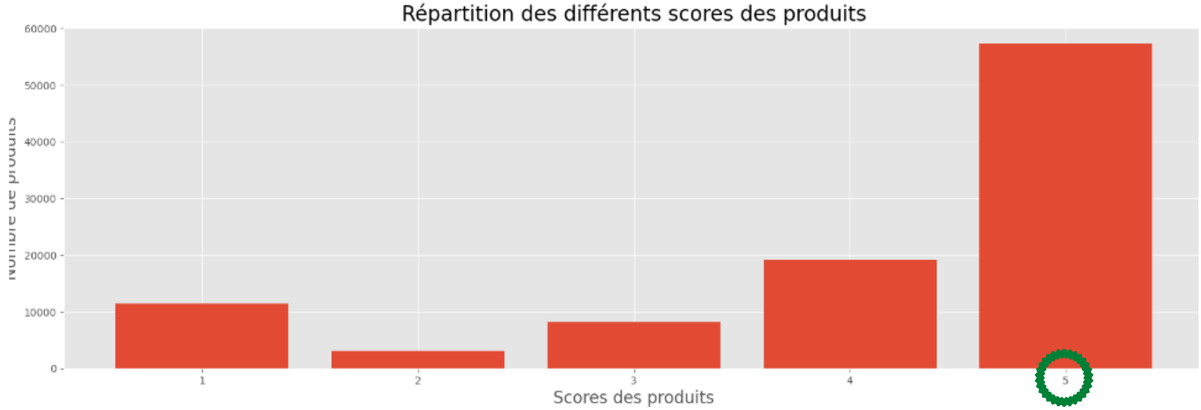
6 - Étude des paiements (payments)



7 - Étude des produits (products)

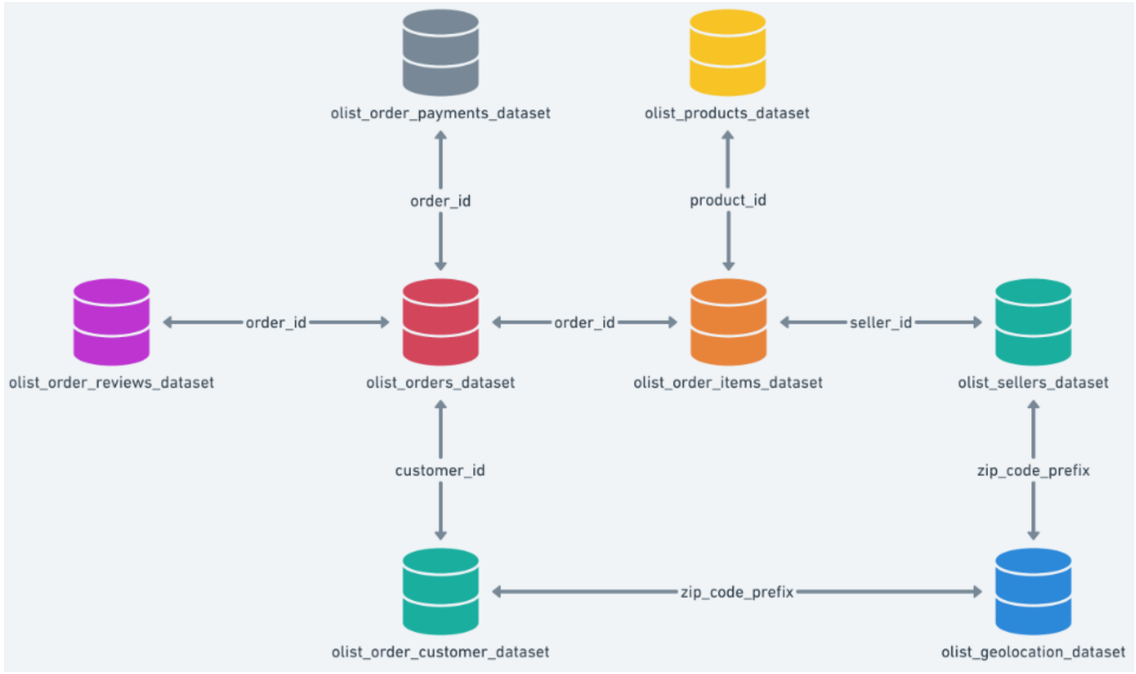


8 - Étude des critiques (reviews)



9 - Étude des traductions (tanslation)

Concaténation des données



117822 observations

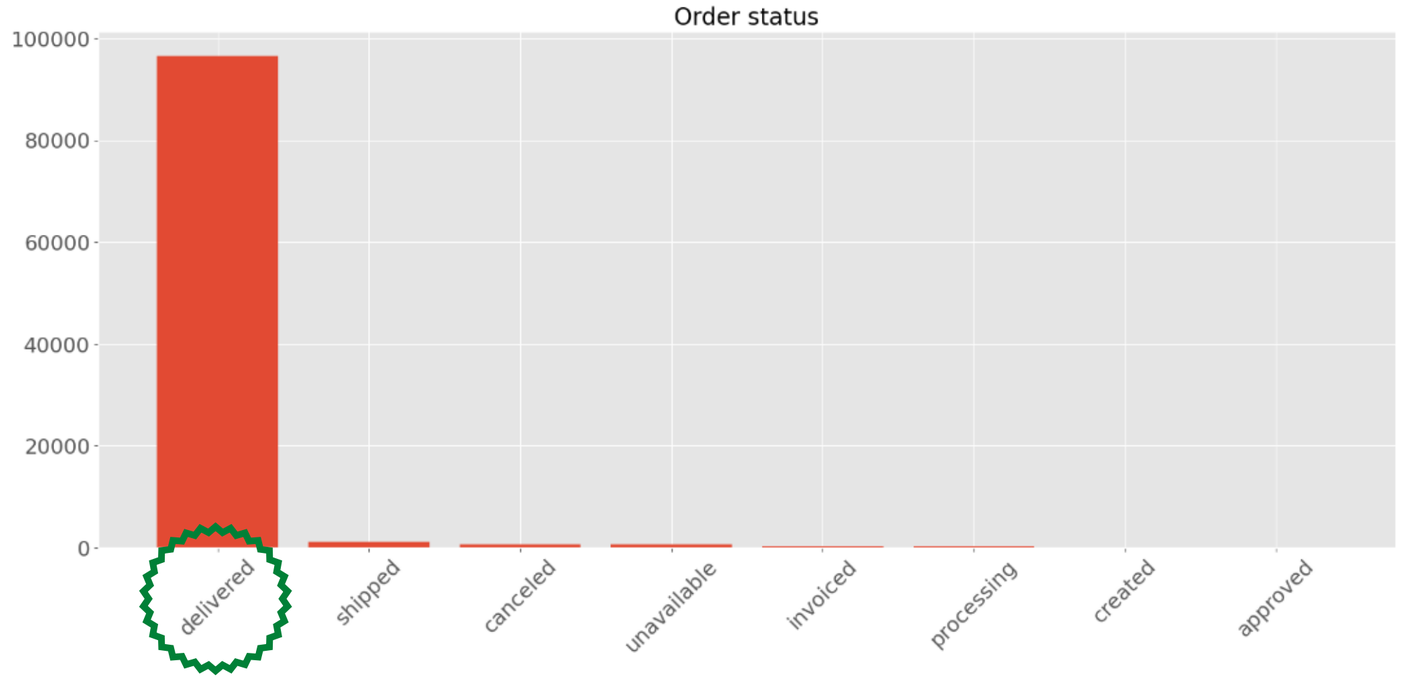
42 paramètres

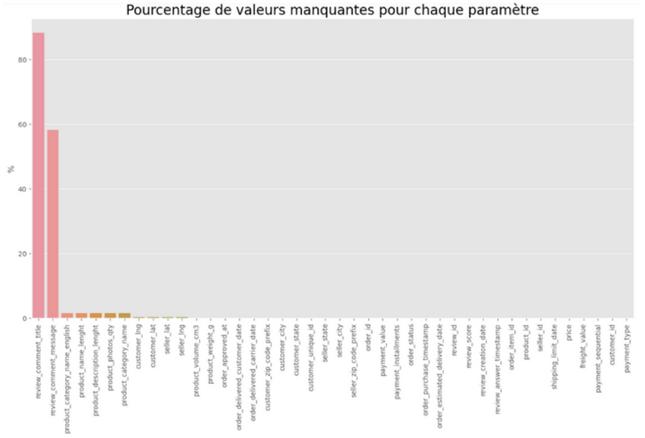
	order_id	customer_id	order_status	order_purchase_timestamp	order_approved_at
0	e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	2017-10-02 10:56:33	2017-10-02 11:07:15
1	e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	2017-10-02 10:56:33	2017-10-02 11:07:15
2	e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	2017-10-02 10:56:33	2017-10-02 11:07:15
3	128e10d95713541c87cd1a2e48201934	a20e8105f23924cd00833fd87daa0831	delivered	2017-08-15 18:29:31	2017-08-15 20:05:16
4	0e7e841ddf8f8f2de2bad69267ecfbcf	26c7ac168e1433912a51b924fbd34d34	delivered	2017-08-02 18:24:47	2017-08-02 18:43:15

Cleaning et Feature Engineering

- Conversion des données temporelles au format datetime64[ns]

- Suppression des observations : order_status ≠ "delivered"





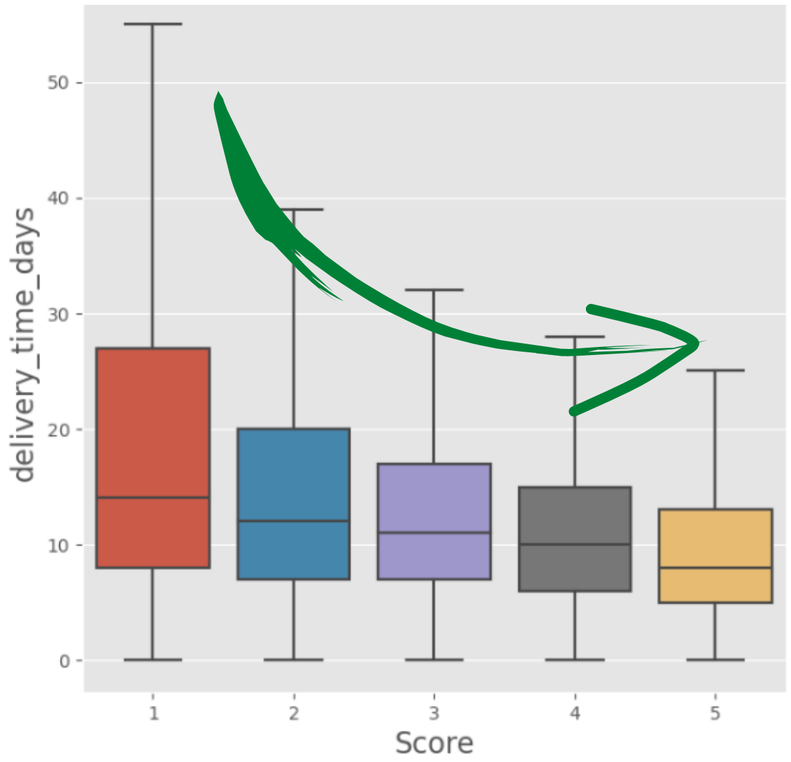
114855 observations
- 3%

40 paramètres

- SimpleImputer sur les données manquantes

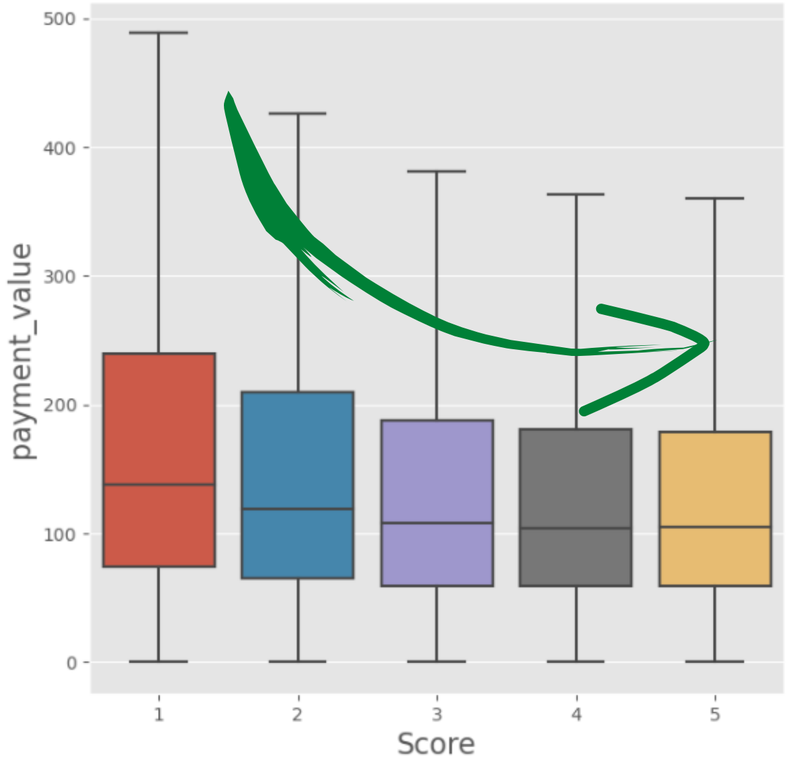
Analyse exploratoire

Scores en fonction des temps de livraison



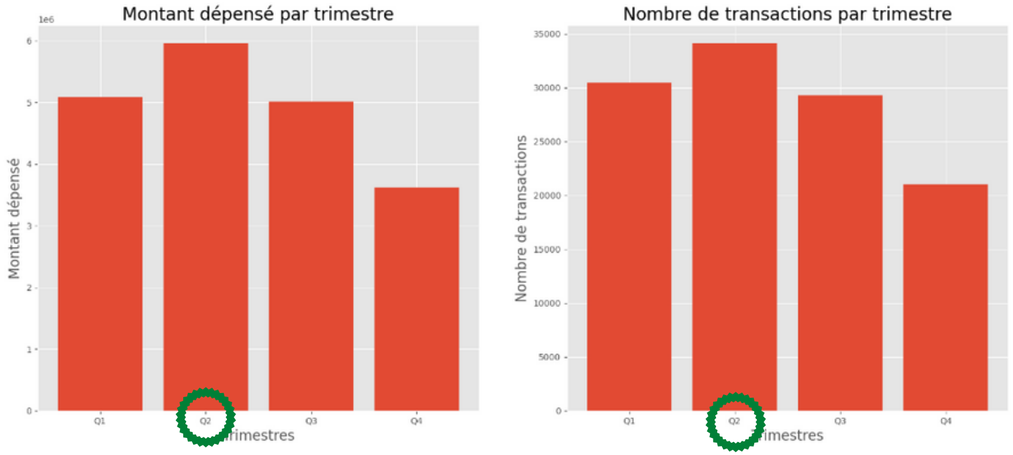
Le score diminue quand le temps de livraison augmente

Scores en fonction du prix de la transaction

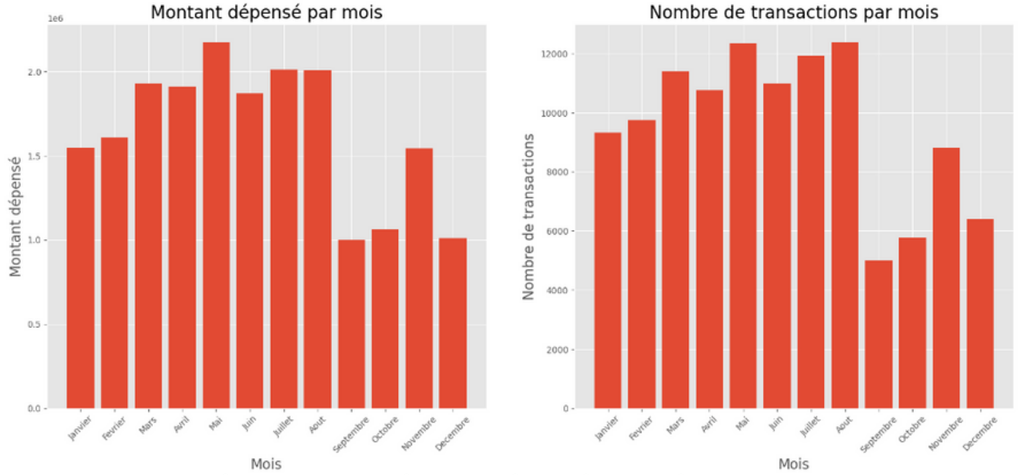


Le score diminue quand le le prix de transaction augmente

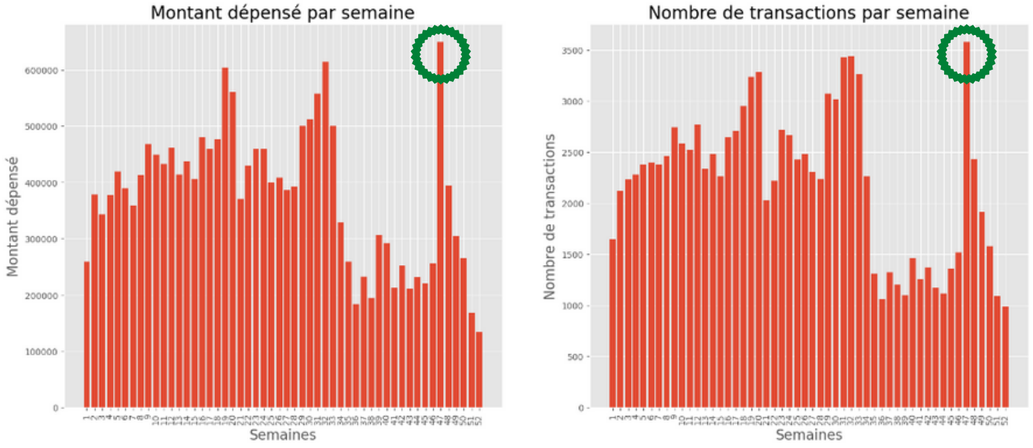
Analyse exploratoire



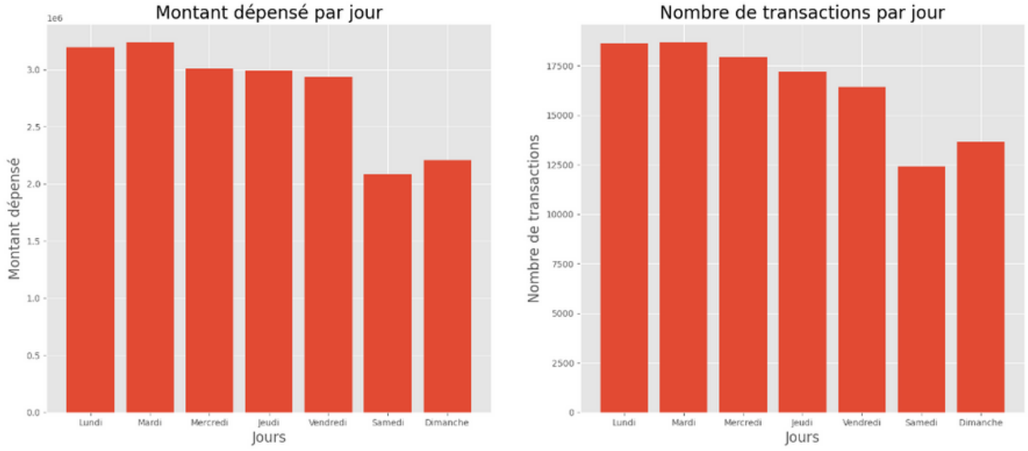
Meilleur trimestre : Q2



Meilleurs mois : Mai, Juin, Juillet, Aout

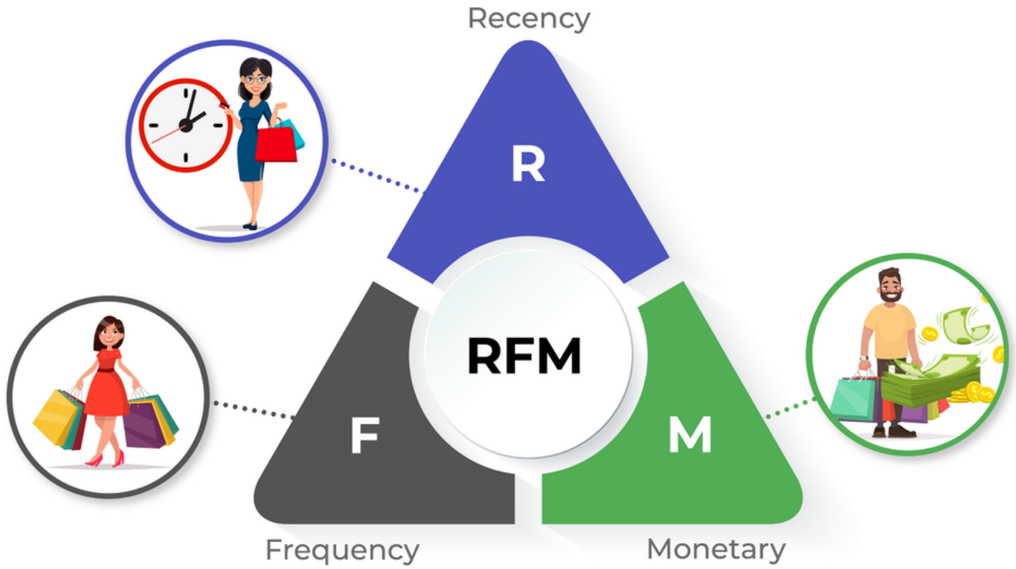


Meilleure semaine : Black Friday



Meilleures ventes en semaine

Présentation de la segmentation RFM



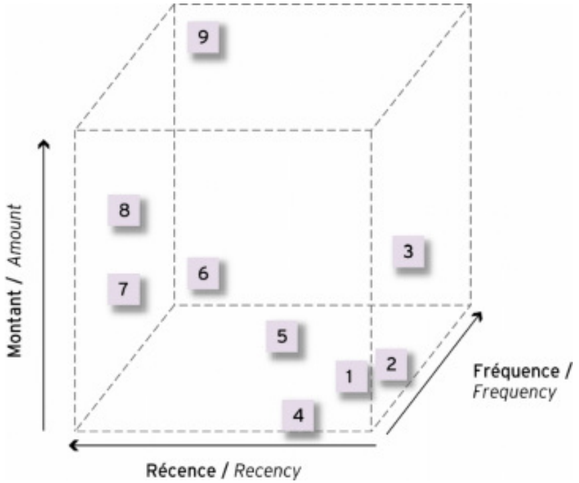
	id	Récence	Fréquence	Montant
0	00012a2ce6f8dcda20d059ce98491703	289	1	114.74
1	000379cdec625522490c315e70c7a9fb	150	1	107.01
2	000419c5494106c306a97b5635748086	181	1	49.40
3	00046a560d407e99b969756e0b10f282	255	1	166.59
4	00050bf6e01e69d5c0fd612f1bcfb69c	347	1	85.23
5	000598caf2ef4117407665ac33275130	19	1	1255.71
6	0005aefbb696d34b3424dccd0a0e9fd0	71	1	147.33
7	00066ccbe787a588c52bd5ff404590e3	205	4	1080.00
8	00072d033fe2e59061ae5c3aff1a2be5	363	1	106.97
9	000bf8121c3412d3057d32371c5d3395	323	2	91.12

DONNÉES → FICHIER CLIENTS

Segmentation par calculs statistiques

Connaissance métiers : La boîte à outils du responsable marketting
Nathalie Van Laethem, Yvelise Lebon, Béatrice Durand-Mégret ©Dunod

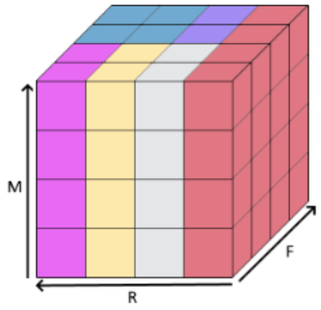
La combinaison des trois critères d'appréciation des clients, montant, récence et fréquence, permet de cibler neuf grands types de clients. 1 = clients perdus depuis longtemps ; 2 = clients non confirmés ; 3 = clients réguliers perdus récemment ; 4 = clients récents à petit CA ; 5 = clients récents à fort CA ; 6 = clients réguliers en décroissance ; 7 = clients réguliers à petit CA ; 8 = clients réguliers en développement ; 9 = très bons clients réguliers.

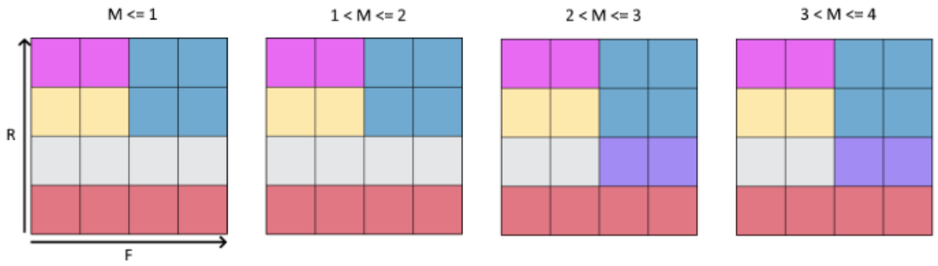



```python
def quartiles_score(x, quartiles, col):

    if x <= quartiles[col][.25]:
        return 1
    elif x <= quartiles[col][.5]:
        return 2
    elif x <= quartiles[col][.75]:
        return 3
    else:
        return 4
```

	R	F	M
0	1	1	3
1	3	1	2
2	2	1	1
3	2	1	3
4	1	1	2
5	4	1	4
6	4	1	3
7	2	4	4
8	1	1	2
9	1	4	2



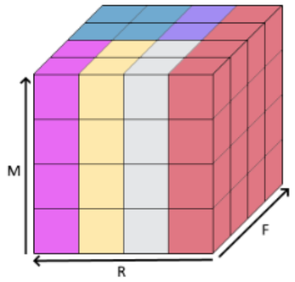


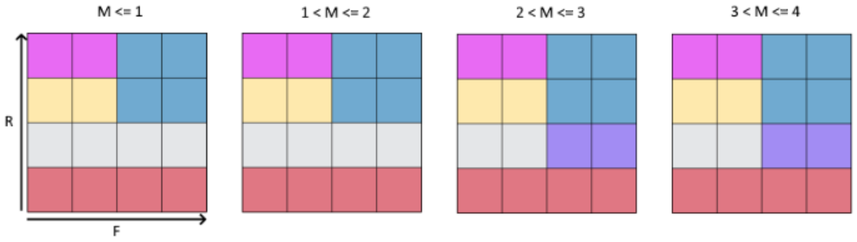
Segmentation par calculs statistiques

```python
def quartiles_score(x, quartiles, col):

    if x <= quartiles[col][.25]:
        return 1
    elif x <= quartiles[col][.5]:
        return 2
    elif x <= quartiles[col][.75]:
        return 3
    else:
        return 4
```

	R	F	M
0	1	1	3
1	3	1	2
2	2	1	1
3	2	1	3
4	1	1	2





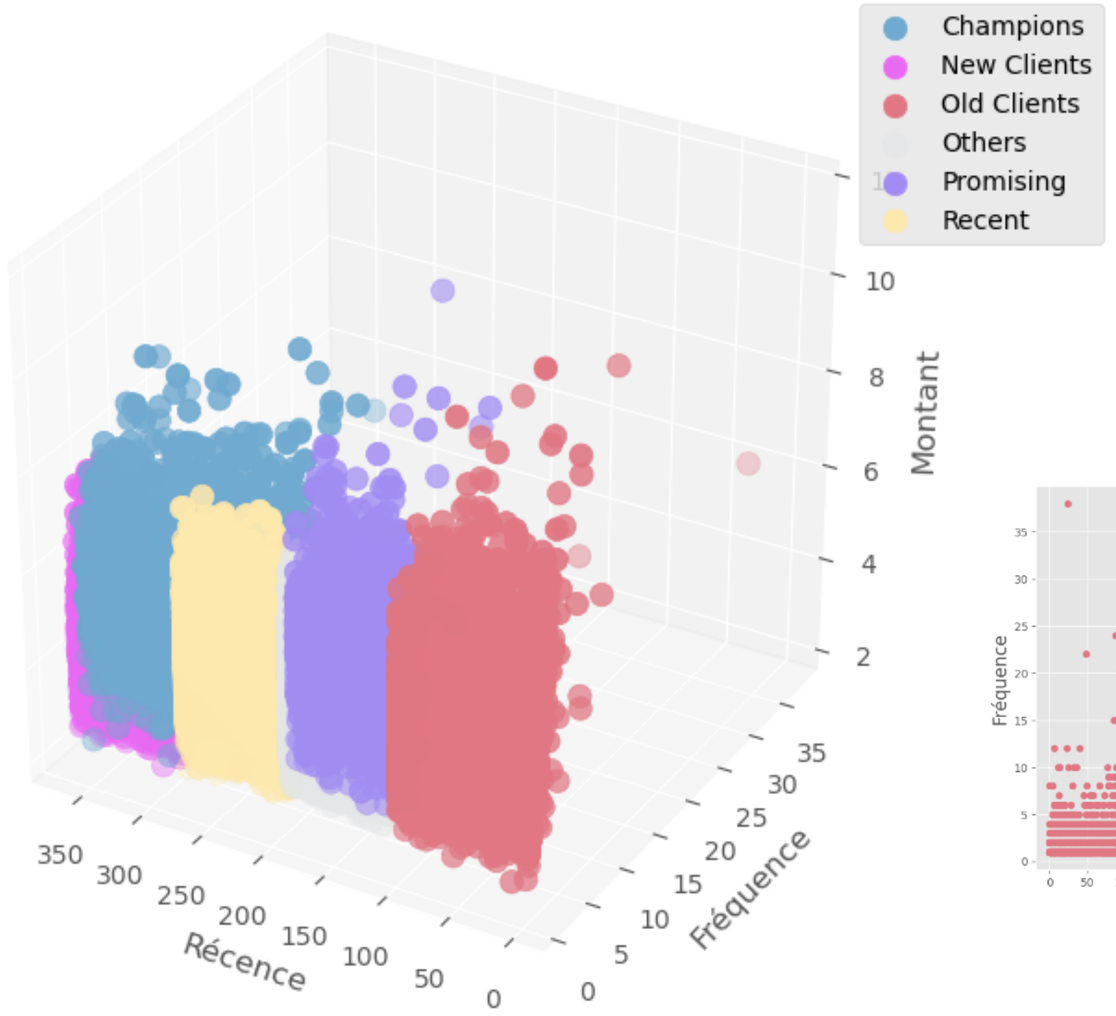
Variable "Récence" inversée

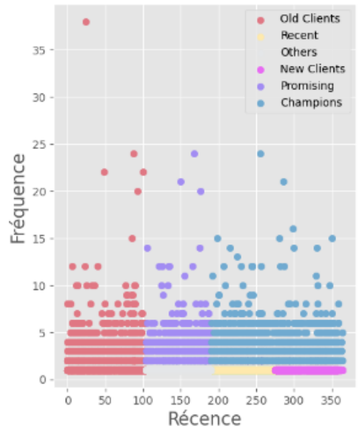
Passage en logarithme de la variable "Montant"

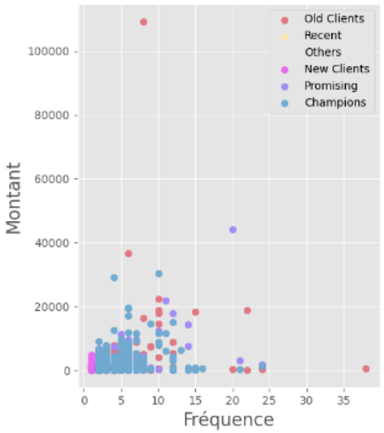
	Récence					Fréquence					Montant				
	count	min	mean	median	max	count	min	mean	median	max	count	min	mean	median	max
clusters_stats															
Champions	4605	191	272.885559	268.0	364	4605	2	2.504452	2.0	24	4605	2.260721	5.781645	5.759406	10.315134
New Clients	16165	276	320.379709	324.0	364	16165	1	1.000000	1.0	1	16165	2.387845	4.646168	4.611152	8.451434
Old Clients	18689	0	59.616887	66.0	103	18689	1	1.213816	1.0	38	18689	2.309561	4.797941	4.698387	11.601967
Others	15986	104	150.474353	151.0	190	15986	1	1.000000	1.0	1	15986	2.623218	4.607099	4.585274	8.249784
Promising	2366	104	151.149620	152.0	190	2366	2	2.467878	2.0	24	2366	2.795450	5.713270	5.704715	10.693035
Recent	16234	191	231.885549	232.0	275	16234	1	1.000000	1.0	1	16234	2.516890	4.657131	4.621437	8.336932

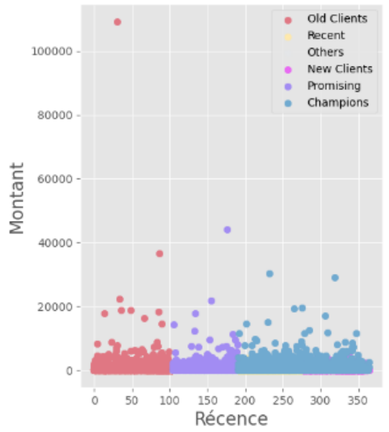
Segmentation par calculs statistiques

Représentation 3D des différents individus

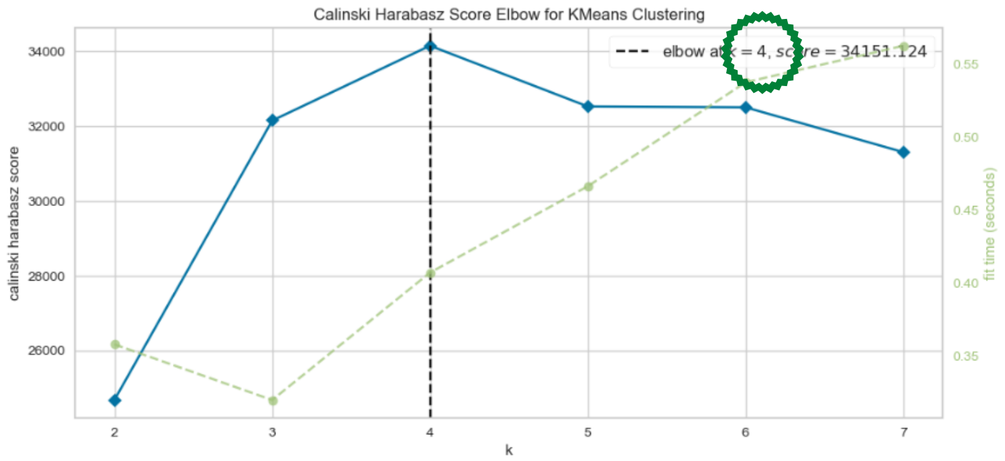


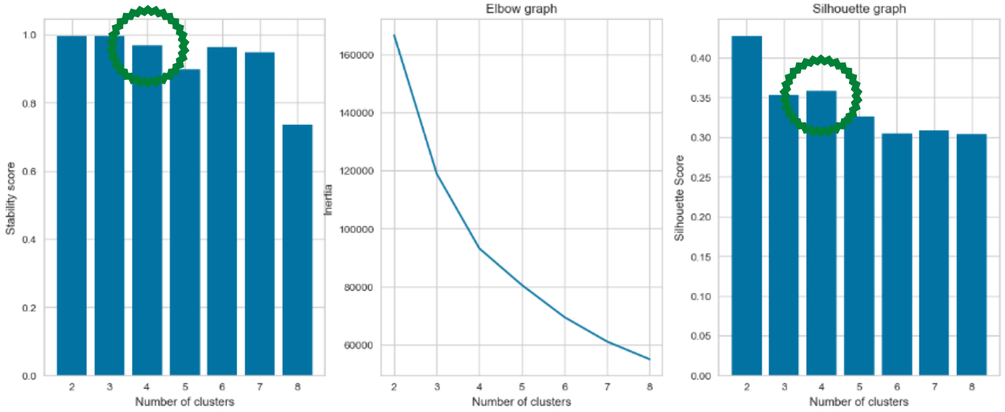


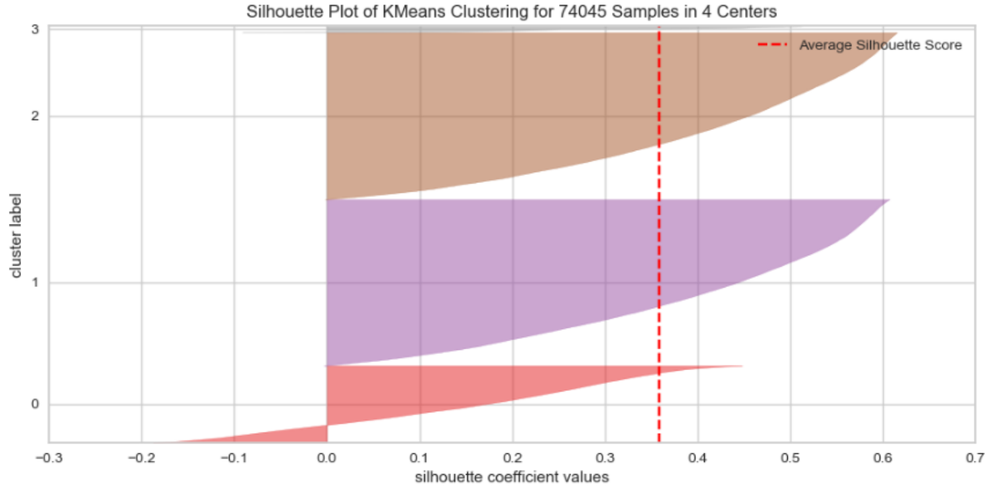




Segmentation RFM par algorithme non supervisé KMEANS

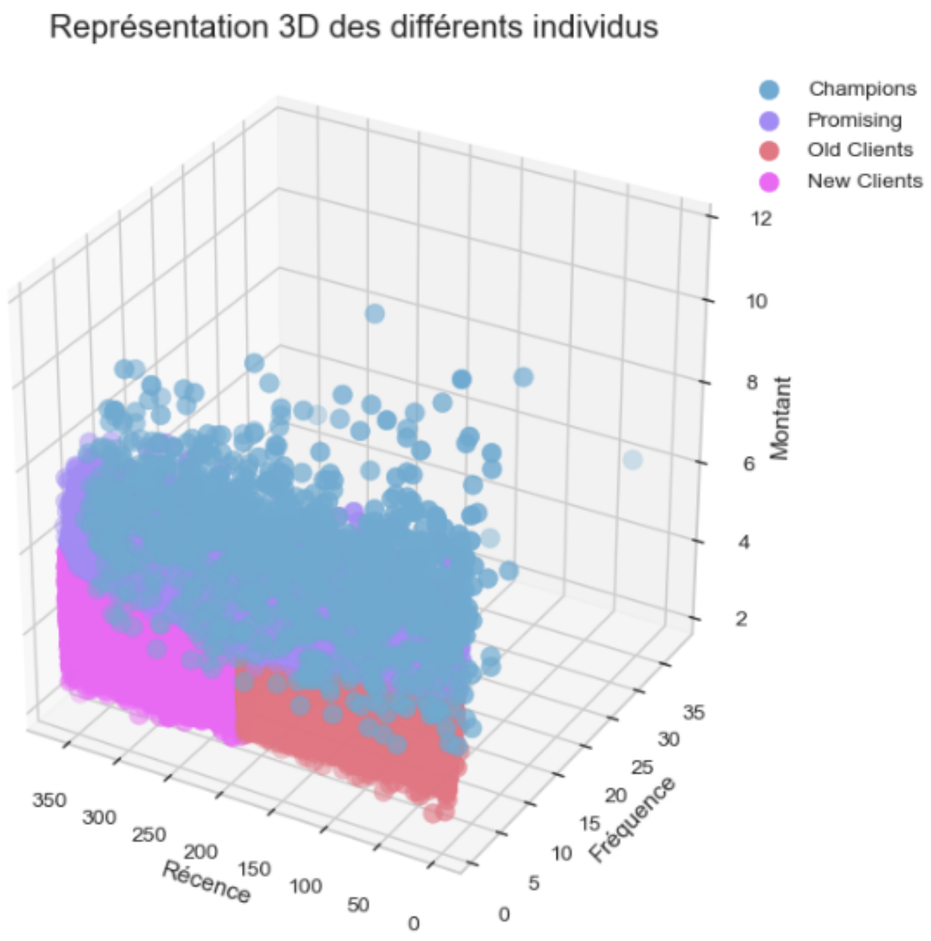






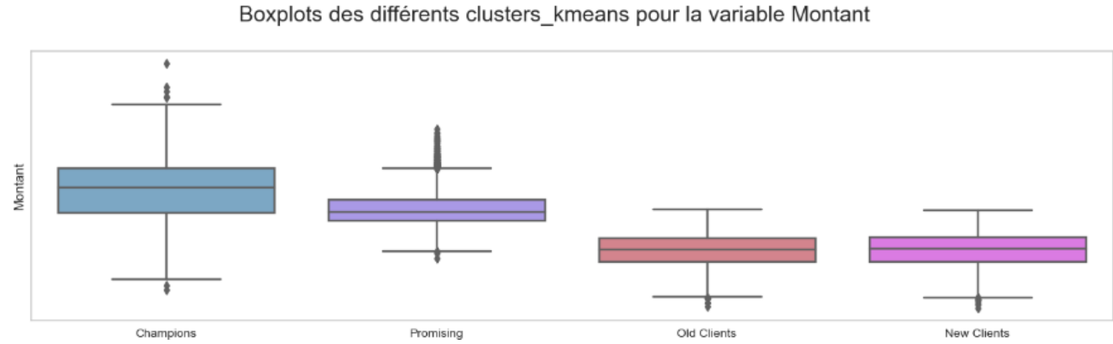
	Récence				Fréquence				Montant			
	count	min	median	max	count	min	median	max	count	min	median	max
clusters_kmeans												
0	29663	190	280.0	364	29663	1	1.0	4	29663	2.260721	4.510750	5.999557
1	29550	0	99.0	192	29550	1	1.0	3	29550	2.309561	4.468491	6.021533
2	13715	0	190.0	364	13715	1	1.0	3	13715	4.145671	5.928072	9.092795
3	1117	0	183.0	364	1117	4	4.0	38	1117	2.962692	6.867100	11.601967

Segmentation RFM par algorithme non supervisé KMEANS

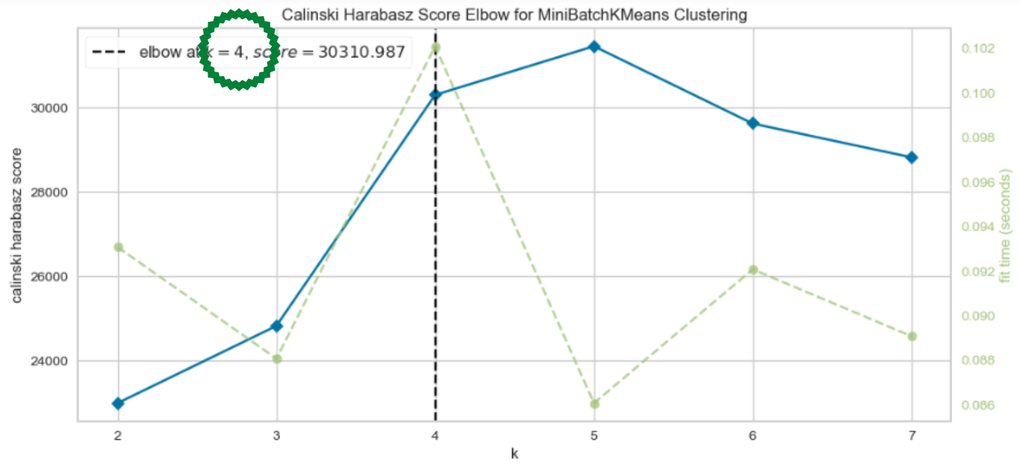


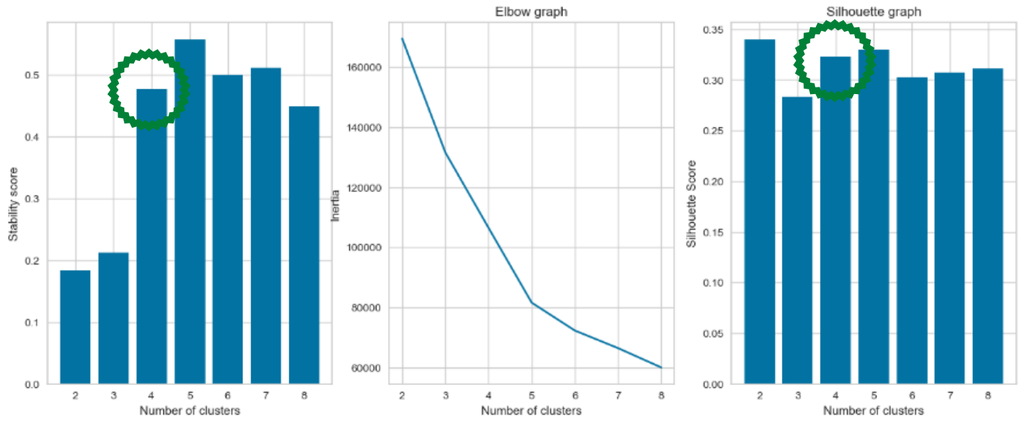
	Récence				Fréquence				Montant			
	count	min	median	max	count	min	median	max	count	min	median	max
clusters_kmeans												
0	29663	190	280.0	364	29663	1	1.0	4	29663	2.260721	4.510750	5.999557
1	29550	0	99.0	192	29550	1	1.0	3	29550	2.309561	4.468491	6.021533
2	13715	0	190.0	364	13715	1	1.0	3	13715	4.145671	5.928072	9.092795
3	1117	0	183.0	364	1117	4	4.0	38	1117	2.962692	6.867100	11.601967

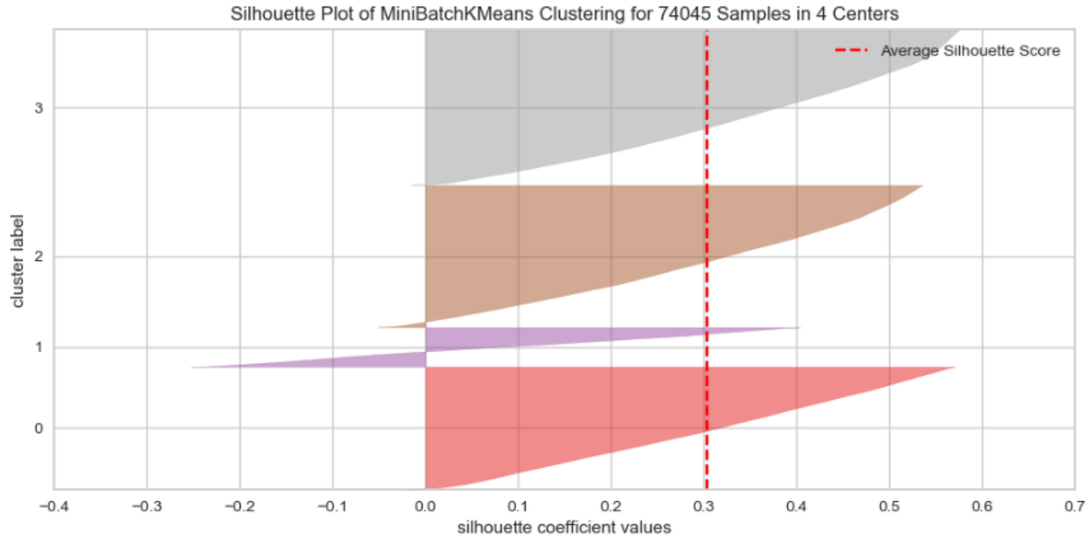
	Récence				Fréquence				Montant			
	count	min	median	max	count	min	median	max	count	min	median	max
clusters_kmeans												
New Clients	29663	190	280.0	364	29663	1	1.0	4	29663	2.260721	4.510750	5.999557
Old Clients	29550	0	99.0	192	29550	1	1.0	3	29550	2.309561	4.468491	6.021533
Promising	13715	0	190.0	364	13715	1	1.0	3	13715	4.145671	5.928072	9.092795
Champions	1117	0	183.0	364	1117	4	4.0	38	1117	2.962692	6.867100	11.601967



Segmentation RFM par algorithme non supervisé
MINI BATCH KMEANS

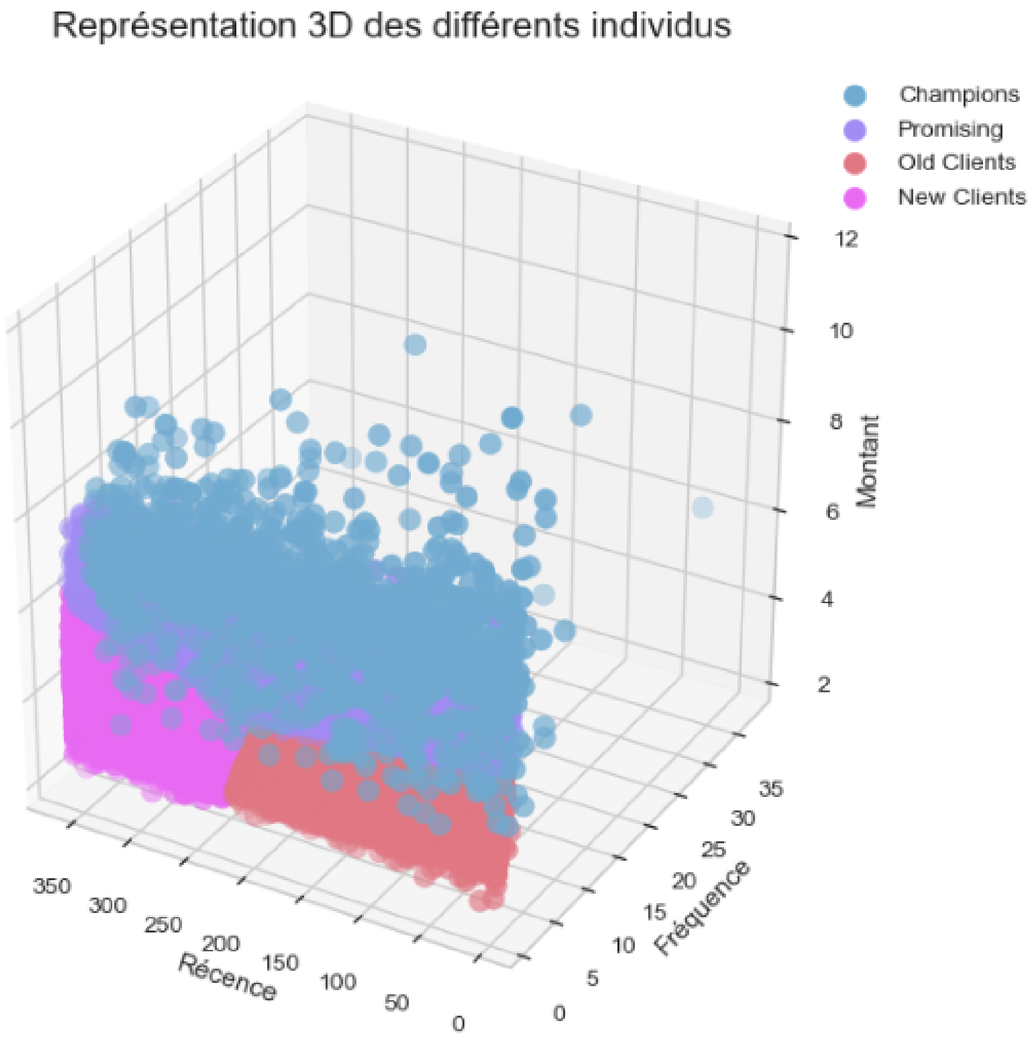




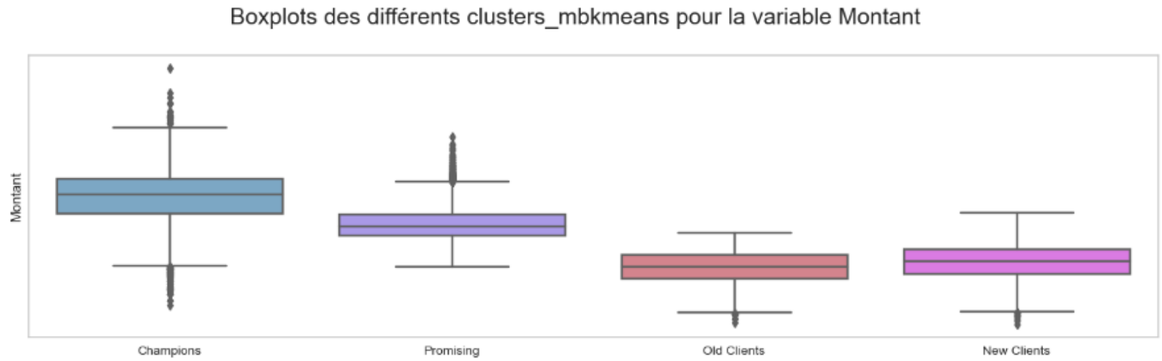


	Récence				Fréquence				Montant			
	count	min	median	max	count	min	median	max	count	min	median	max
clusters_mbkmeans												
New Clients	28131	192	287.0	364	28131	1	1.0	3	28131	2.260721	4.560696	6.301721
Old Clients	29141	0	105.0	223	29141	1	1.0	3	29141	2.309561	4.352083	5.590614
Champions	1477	0	196.0	364	1477	3	4.0	38	1477	2.962692	6.978698	11.601967
Promising	15296	0	174.0	364	15296	1	1.0	3	15296	4.334542	5.810003	9.092795

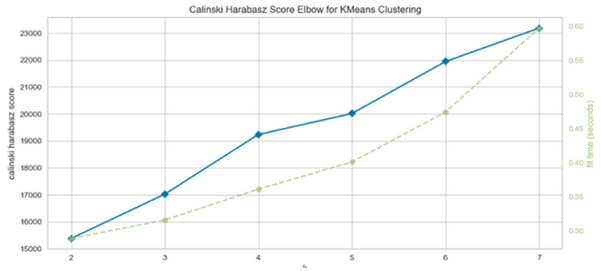
Segmentation RFM par algorithme non supervisé
MINI BATCH KMEANS



	Récence				Fréquence				Montant			
	count	min	median	max	count	min	median	max	count	min	median	max
clusters_mbkmeans												
New Clients	28131	192	287.0	364	28131	1	1.0	3	28131	2.260721	4.560696	6.301721
Old Clients	29141	0	105.0	223	29141	1	1.0	3	29141	2.309561	4.352083	5.590614
Champions	1477	0	196.0	364	1477	3	4.0	38	1477	2.962692	6.978698	11.601967
Promising	15296	0	174.0	364	15296	1	1.0	3	15296	4.334542	5.810003	9.092795

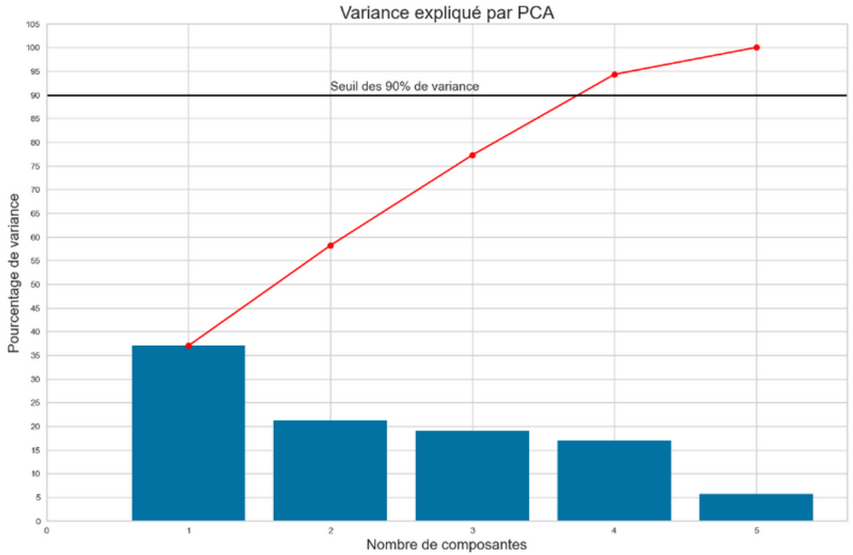


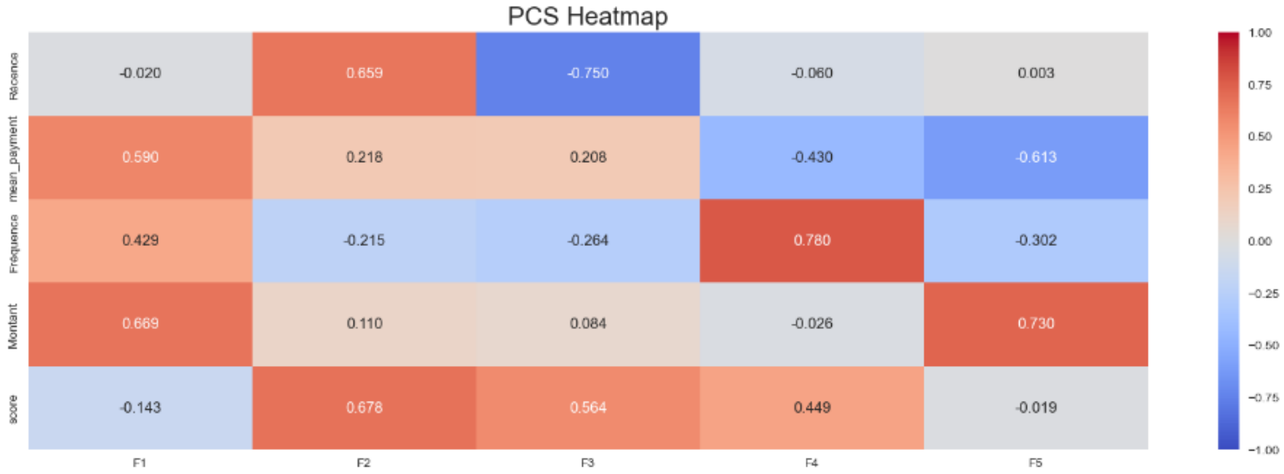
Segmentation améliorée avec ajout des paramètres Scores & Montants moyens



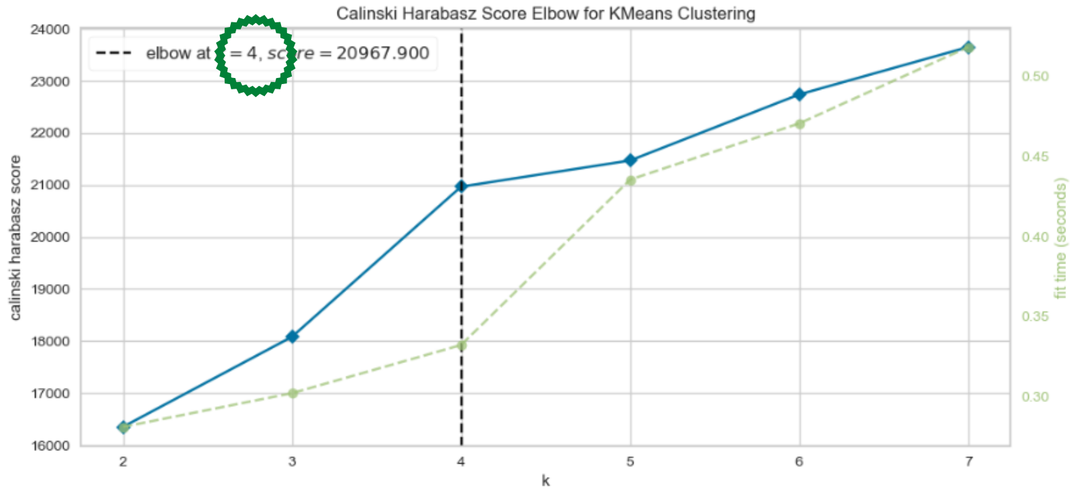
	Récence					Fréquence					Montant				
	count	min	mean	median	max	count	min	mean	median	max	count	min	mean	median	max
clusters_kmeansplus															
0	72307	0	190.190756	190.0	364	72307	1	1.160109	1.0	38	72307	2.260721	4.718592	4.680741	7.119247
1	26	13	176.269231	179.5	347	26	4	9.769231	10.0	22	26	9.339657	9.769345	9.762372	10.693035
2	1	30	30.000000	30.0	30	1	8	8.000000	8.0	8	1	11.601967	11.601967	11.601967	11.601967
3	1711	0	187.348334	198.0	364	1711	1	2.510812	2.0	24	1711	6.798253	7.432009	7.303379	9.186022

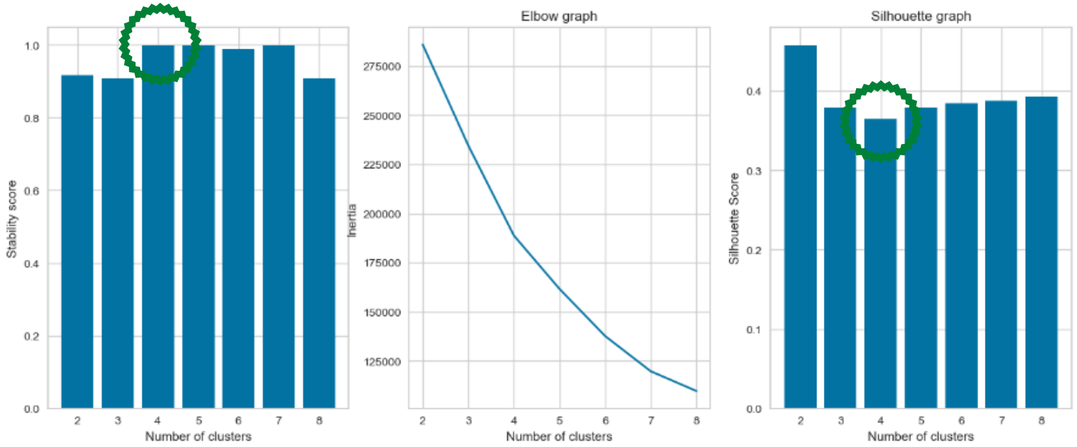
Réduction des dimensions par PCA

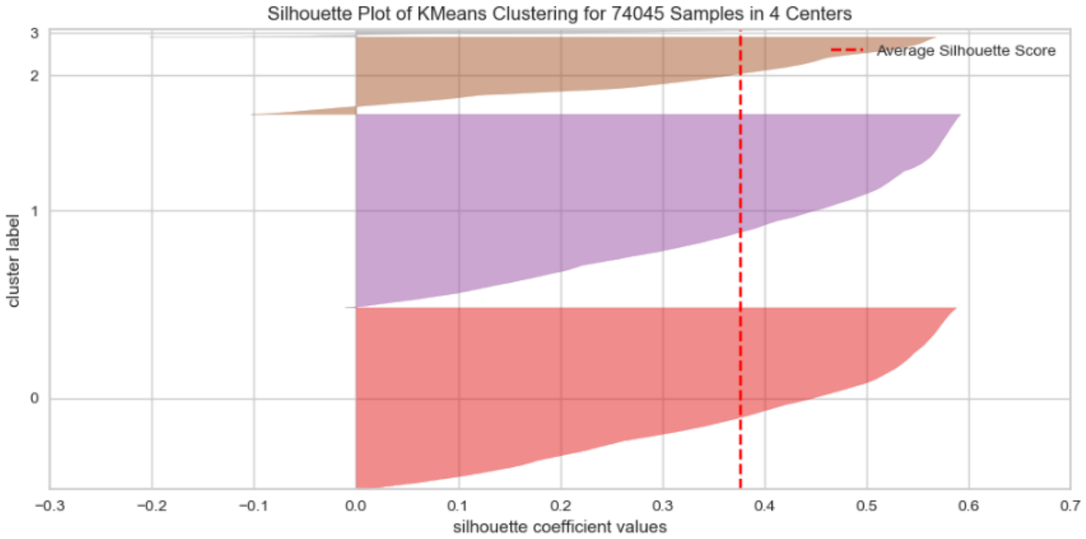




Segmentation améliorée avec ajout des paramètres Scores & Montants moyens

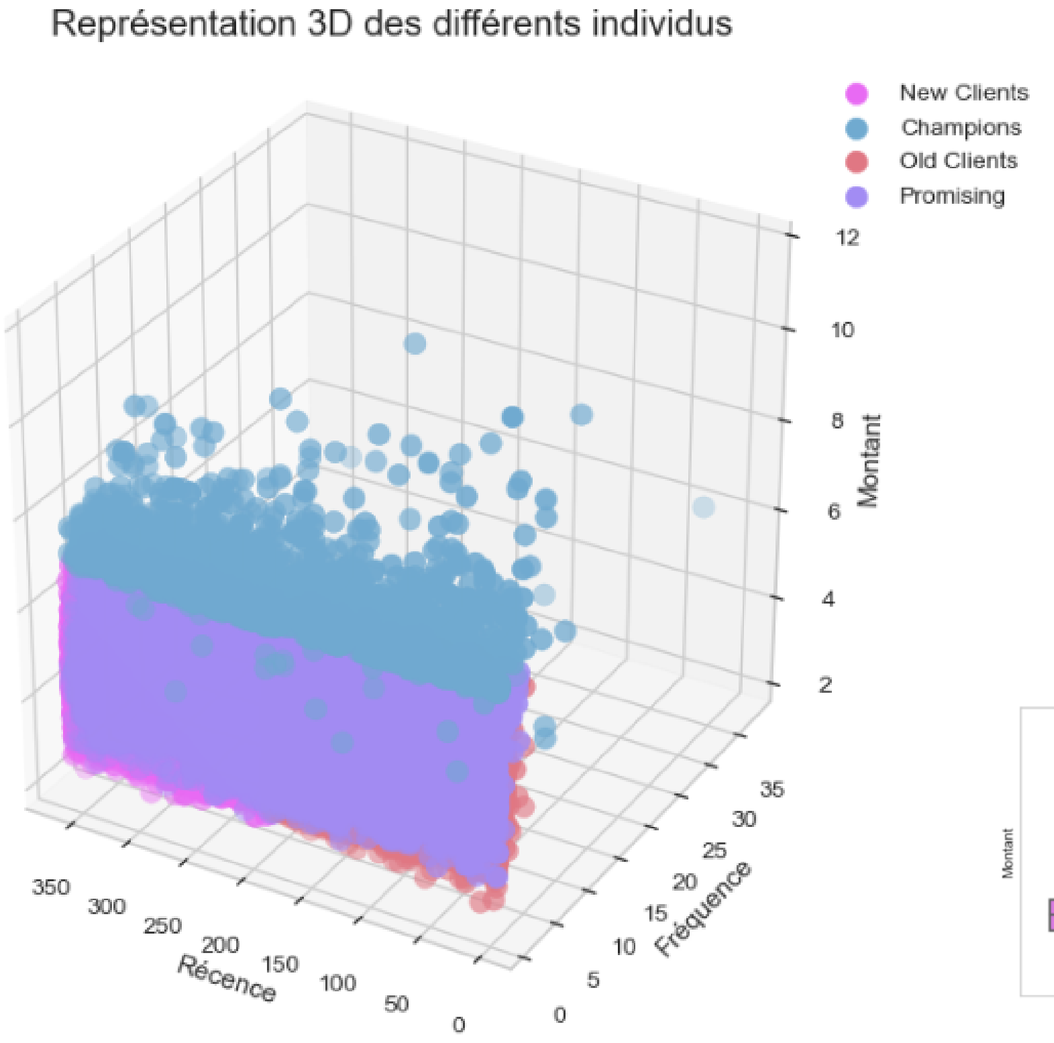




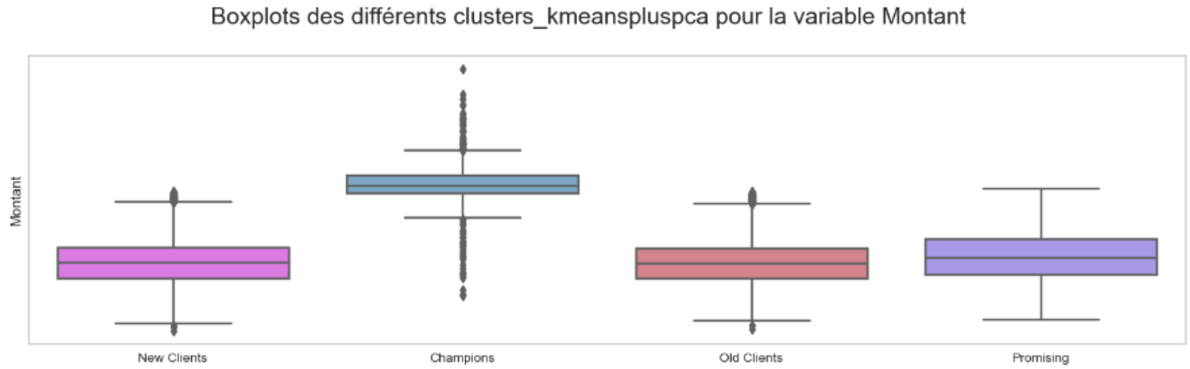


	Récence				Fréquence				Montant			
	count	min	median	max	count	min	median	max	count	min	median	max
clusters_kmeansplus												
Old Clients	28972	0	97.0	190	28972	1	1.0	7	28972	2.309561	4.633466	7.213503
Promising	12323	0	180.0	364	12323	1	1.0	8	12323	2.631169	4.853045	7.322200
New Clients	31291	185	279.0	364	31291	1	1.0	7	31291	2.260721	4.683149	7.216651
Champions	1459	0	202.0	364	1459	1	2.0	38	1459	3.524594	7.413391	11.601967

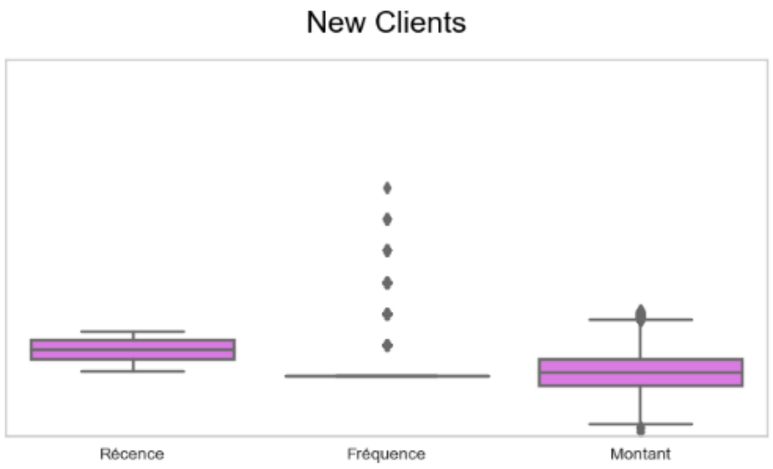
Segmentation améliorée avec ajout de paramètres

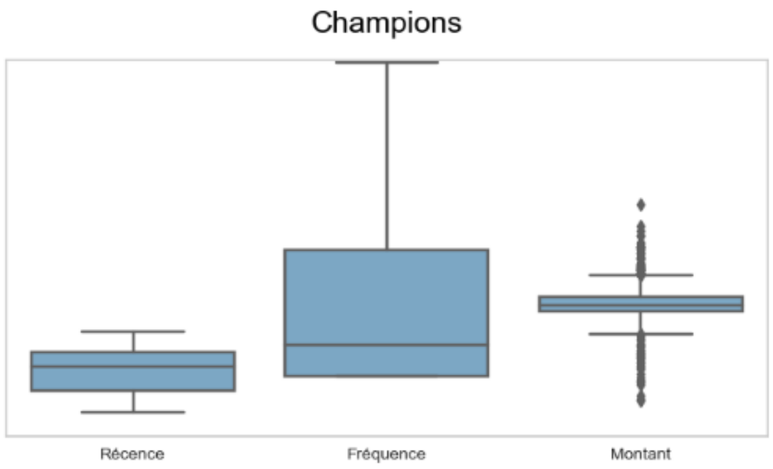


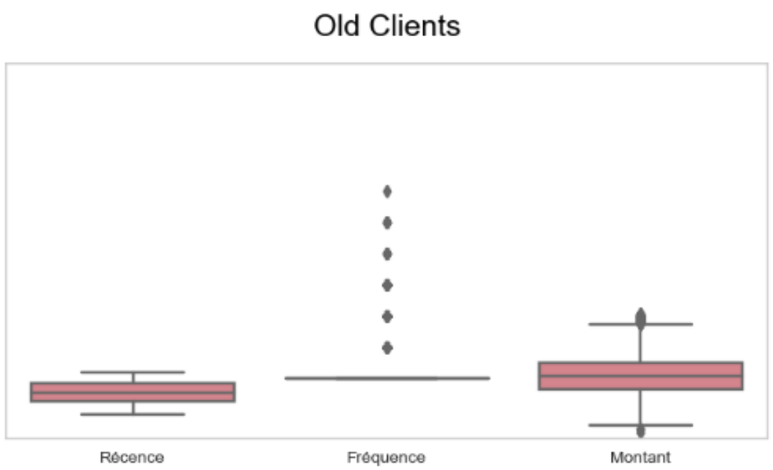
	Récence				Fréquence				Montant			
	count	min	median	max	count	min	median	max	count	min	median	max
clusters_kmeansplus												
Old Clients	28972	0	97.0	190	28972	1	1.0	7	28972	2.309561	4.633466	7.213503
Promising	12323	0	180.0	364	12323	1	1.0	8	12323	2.631169	4.853045	7.322200
New Clients	31291	185	279.0	364	31291	1	1.0	7	31291	2.260721	4.683149	7.216651
Champions	1459	0	202.0	364	1459	1	2.0	38	1459	3.524594	7.413391	11.601967

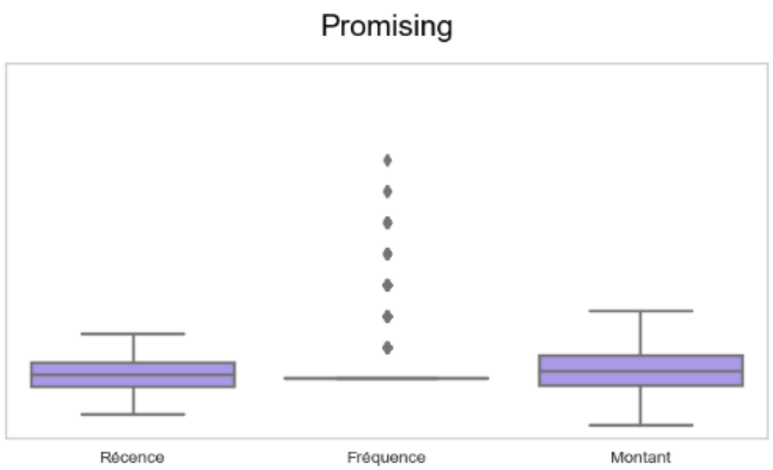


Segmentation améliorée avec ajout de paramètres

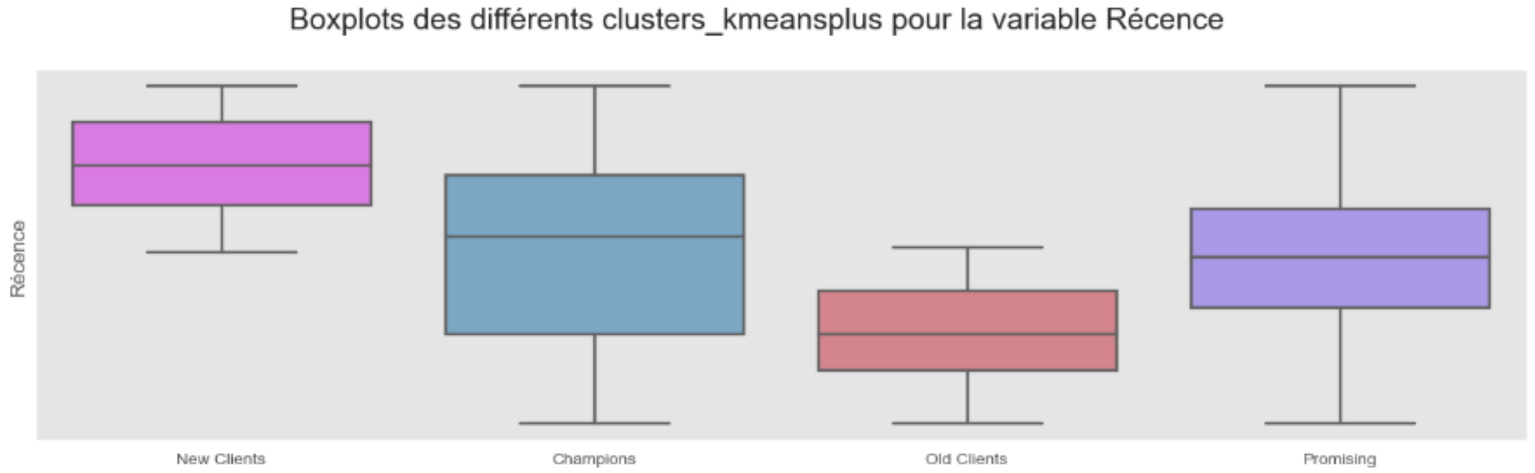


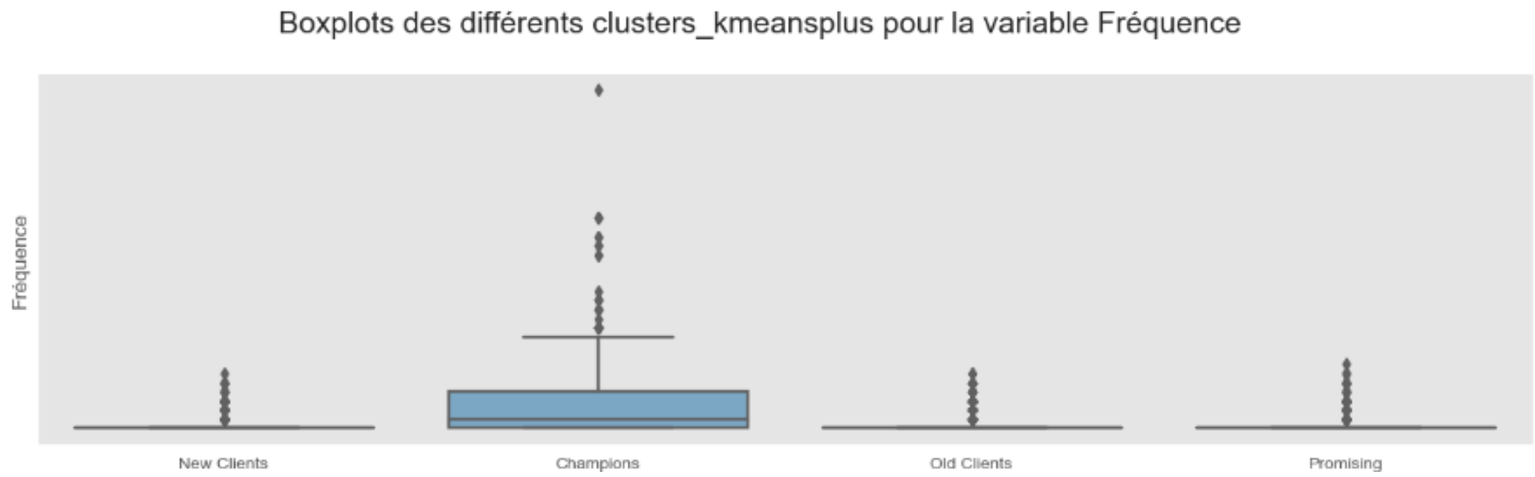




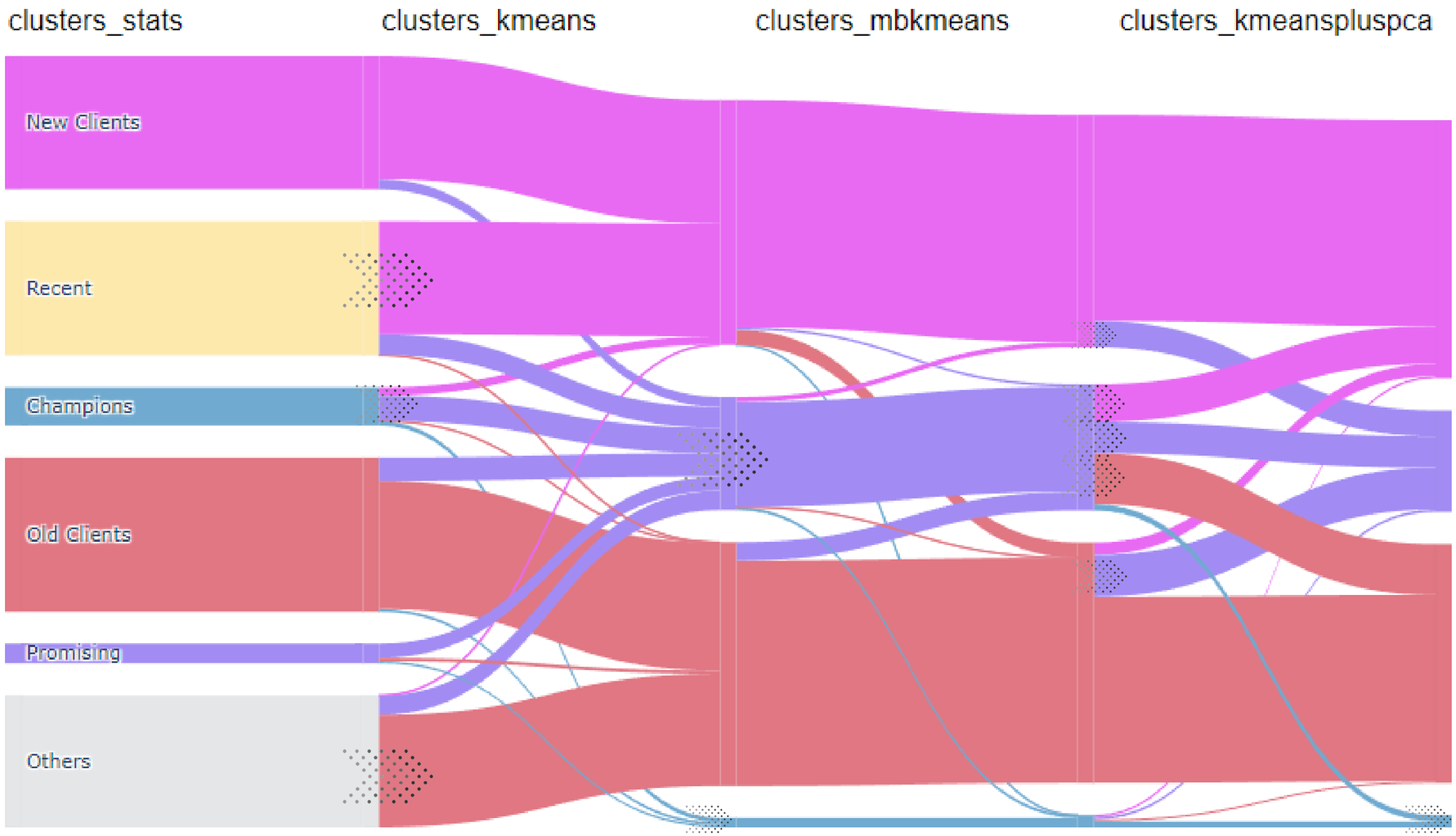


Segmentation améliorée avec ajout de paramètres

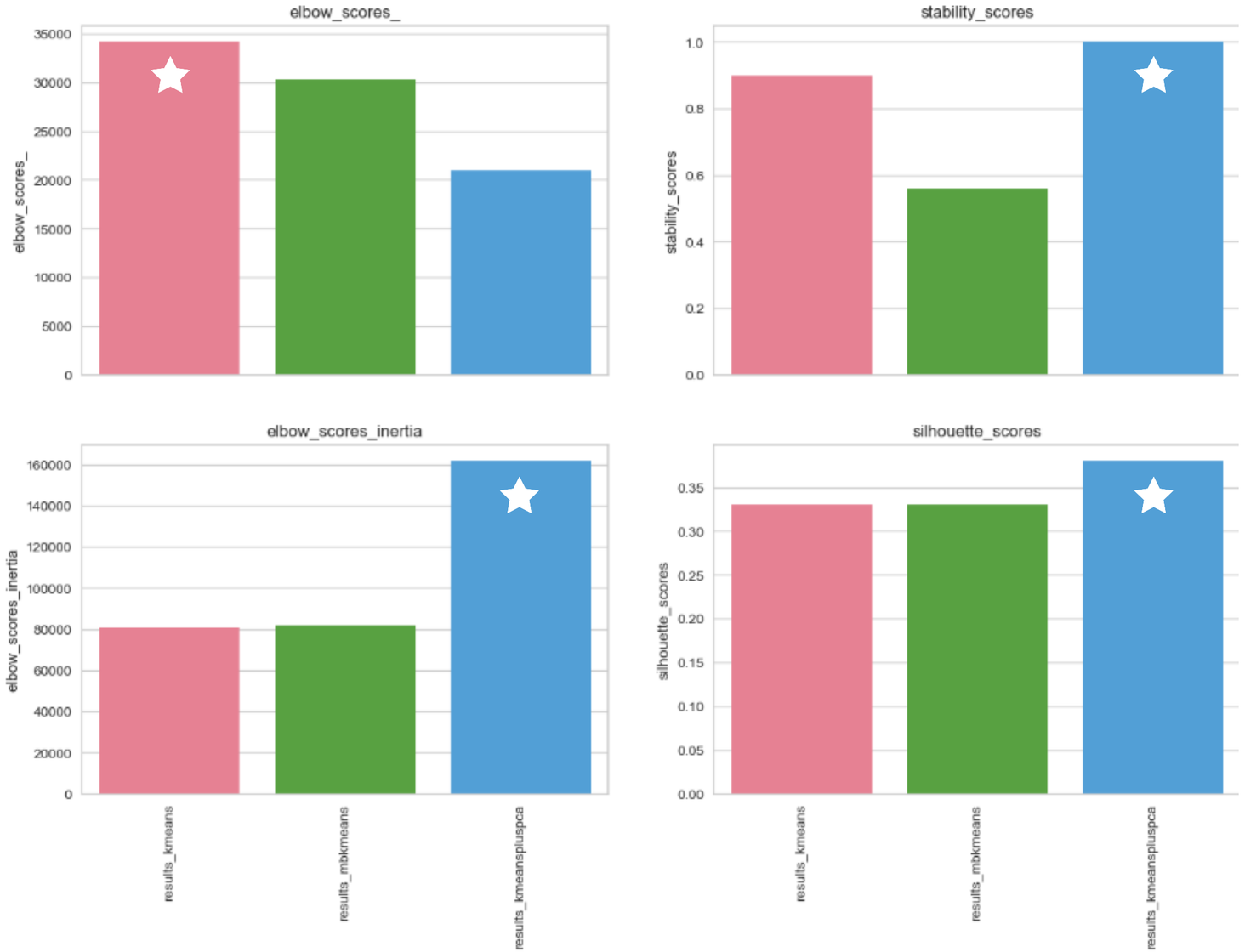




Comparaison des 4 segmentations



Comparaison des 4 segmentations



KMeans

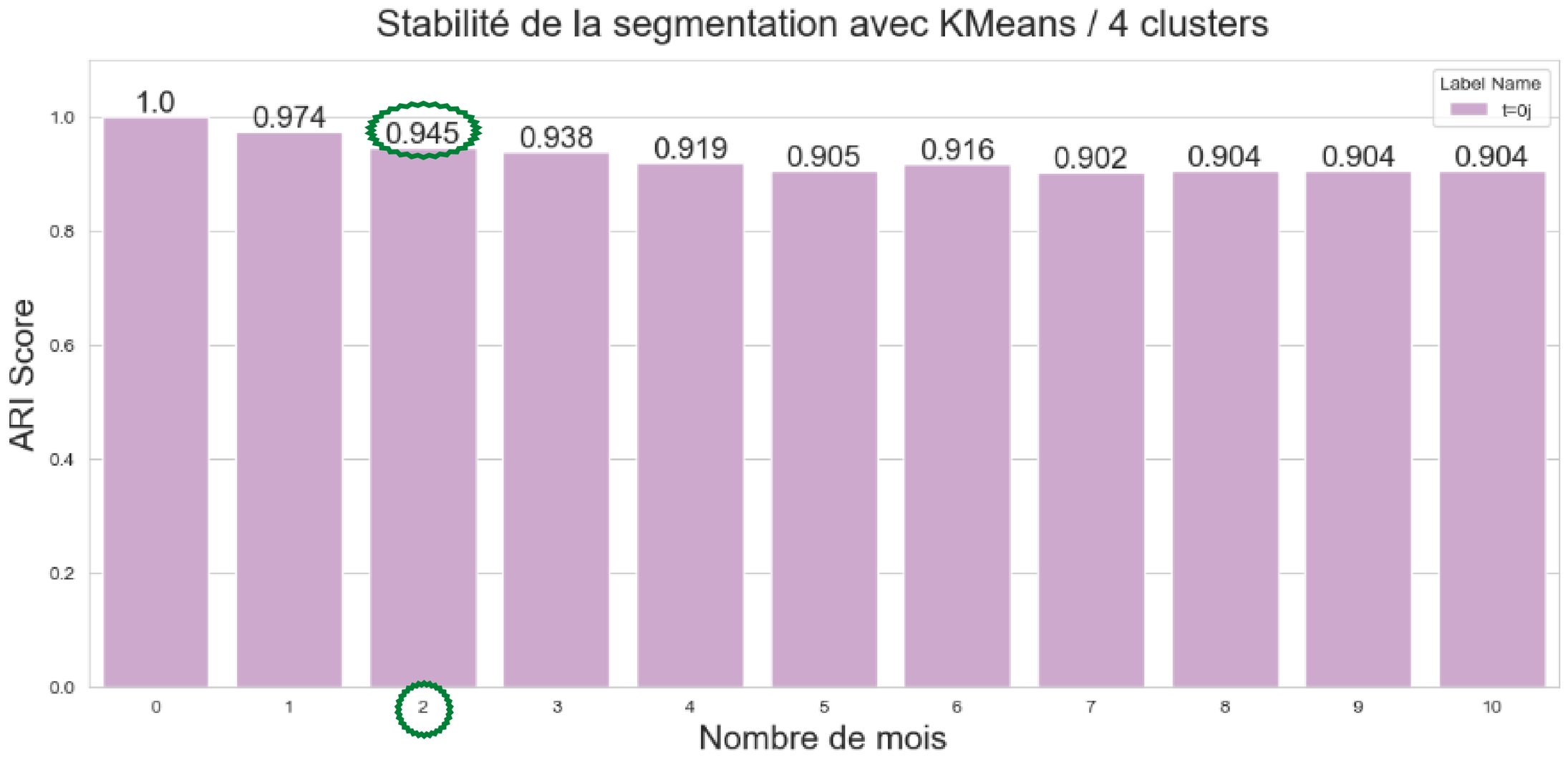
5 paramètres

4 clusters

PCA

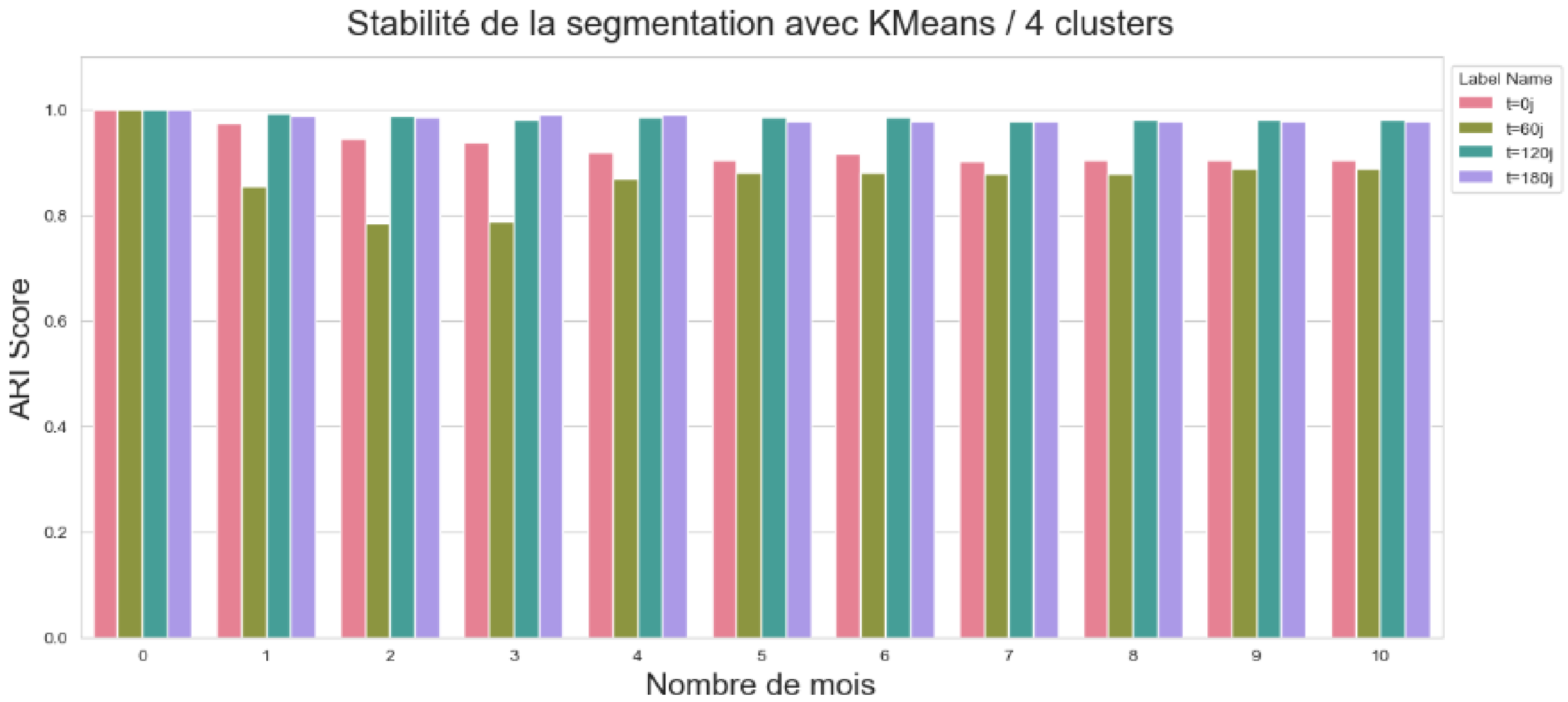
Seuil : ARI < 0.95

Fréquence de MAJ de la segmentation



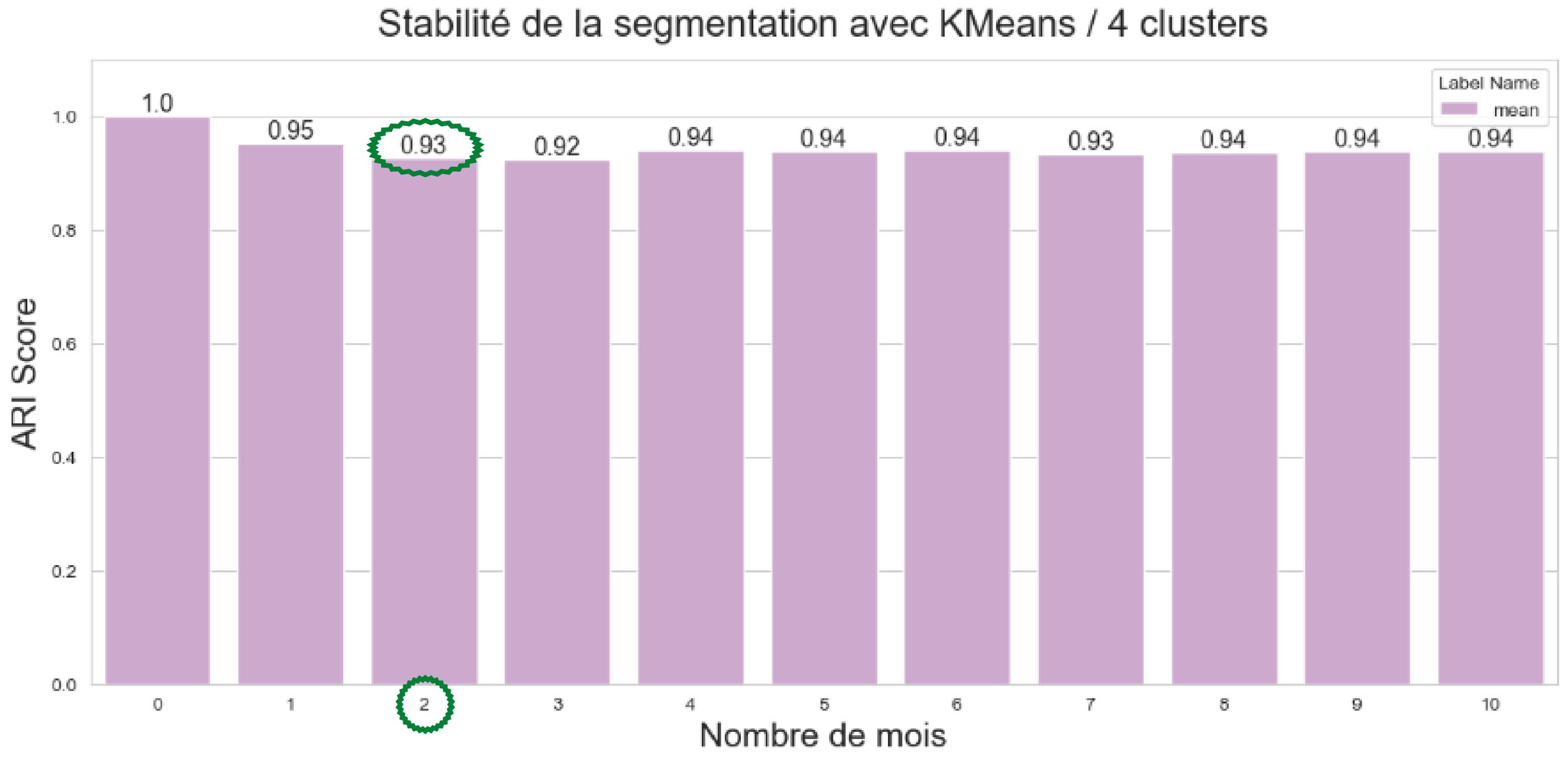
Fréquence de MAJ de la segmentation

Seuil : ARI < 0.95

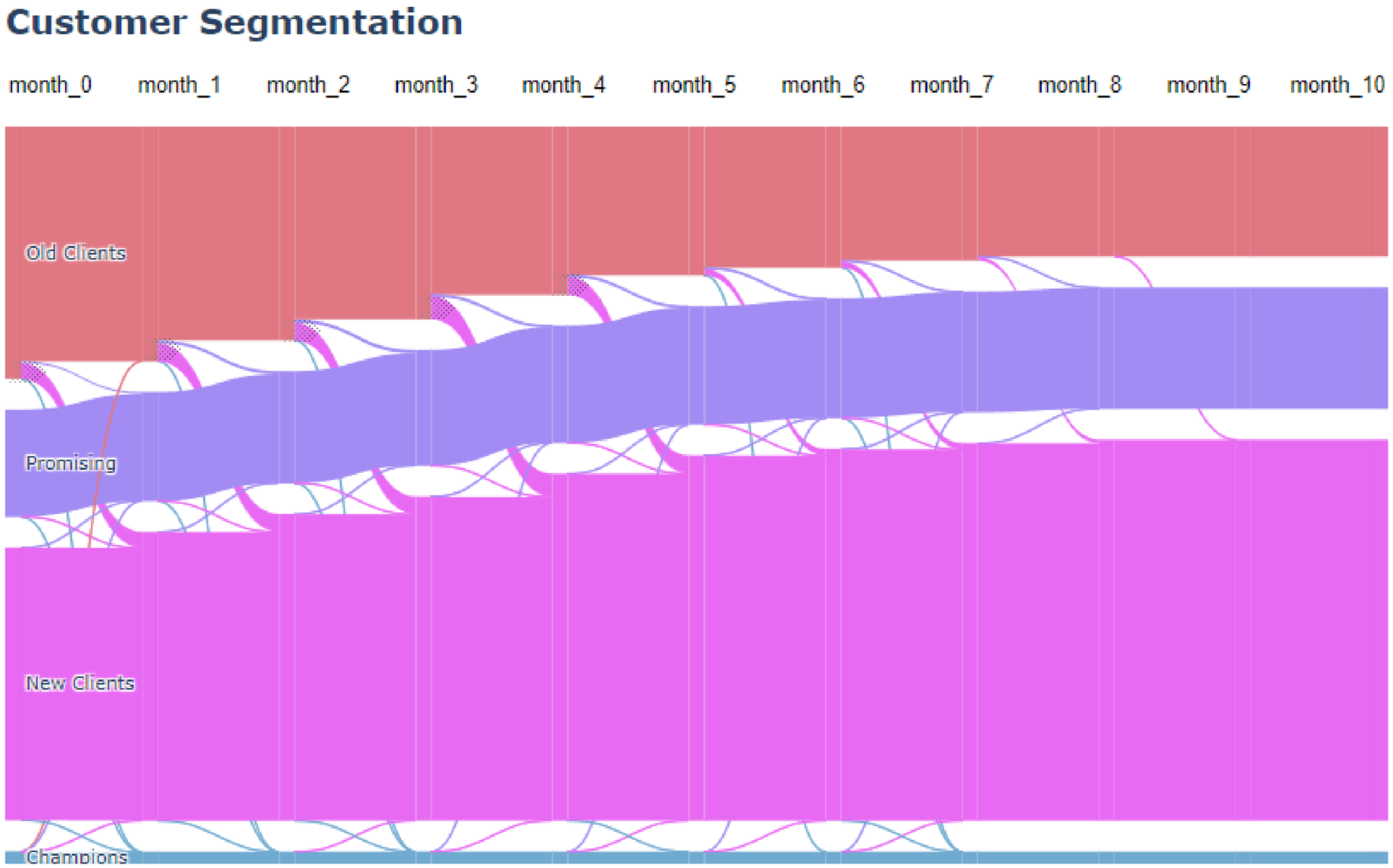


Seuil : ARI < 0.95

Fréquence de MAJ de la segmentation



Fréquence de MAJ de la segmentation



Rappel des objectifs

- Comprendre les différents types d'utilisateurs
- Recommander une fréquence de MAJ de la segmentation

Conclusions

Segmentation recommandée : 4 clusters d'utilisateurs

Champions : excellents acheteurs fidèles
New clients : nouveaux clients à convertir en promising !
Promising : clients à très forts potentiels
Old clients : clients anciens qui tendent à disparaître

Fréquence de MAJ de la segmentation : 2 mois