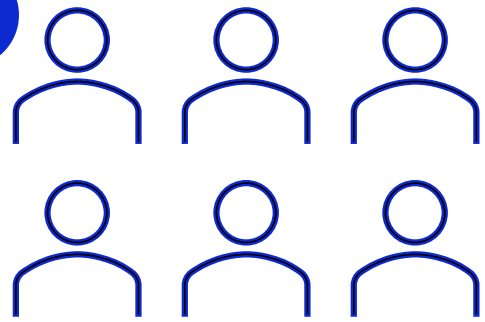
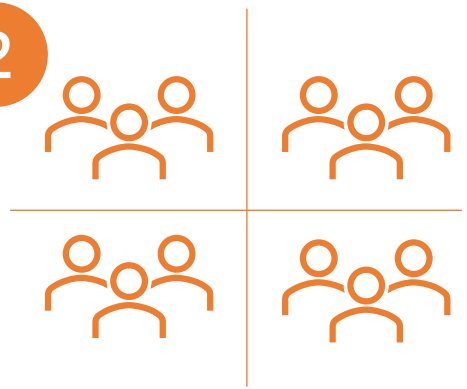
Segmentez des clients d'un site e-commerce

olist
store

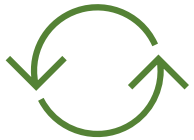
1



2



3



1

Analyse exploratoire

- Montage DataFrame
- Analyse features
- Agrégation par Client
- Analyse des distributions

2

Modélisation / Essais

- Segmentation classique
- Clustering K-means 1
- Clustering K-means 2
- Clustering DBScan
- Modèle retenu

3

Stabilité / Maintenance

- Fréquence = 4 mois
- Fréquence = 3 mois
- Fréquence = 2 mois
- Conclusion & préconisation

1 Analyse exploratoire

Montage DataFrame

Features cibles :
order_id (orders)
customer_unique_id (customers)
order_purchase_timestamp (orders)
payment_value (payments)
review_score (reviews)
nb_articles (items)
nb_diff_products (items)

recency
frequency
monetary

orders
reviews
payments
items
customers
~~products~~
~~sellers~~
~~geolocation~~
~~name_translation~~

orders
reviews
payments
items
customers

1 Analyse exploratoire

payments
items
reviews

par → numéro de commande

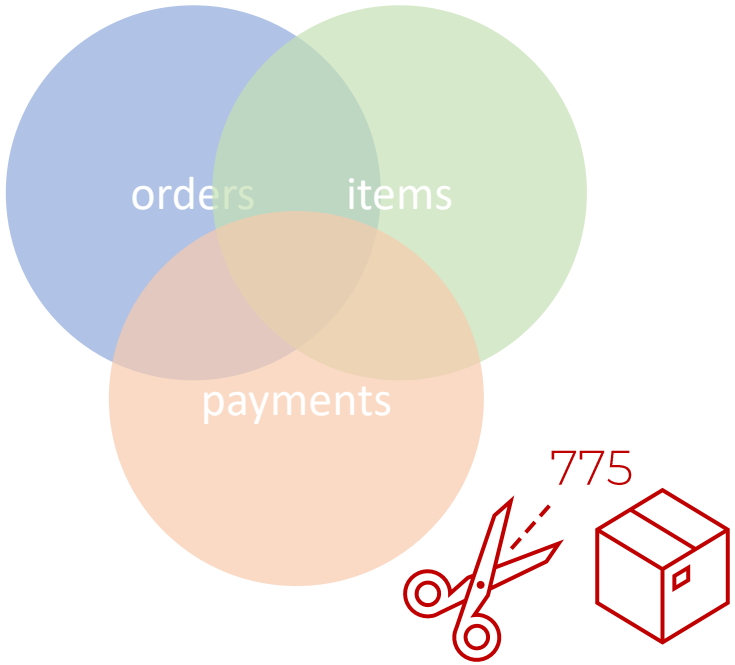
nouvelles features → nb_articles
nb_diff_products

payment_value = price + freight_value (ε = 0,02%)

payments

items

Montage DataFrame



orders payments → customer_id customers → orders reviews → orders items

1 Analyse exploratoire

Analyse features

98665
commandes

95419
clients uniques

7
status commande

15 846 280 R$
chiffre d'affaires

1.03
commandes

1.03
revues

99,5%
delivered

1.04
produits différents

1.14
articles en moyenne

160 R$
en moyenne

4.1 / 5
en moyenne

1 Analyse exploratoire

Agrégation par Client

recency = jour n – jour j

jour 1 — première commande du client

jour j — dernière commande du client

jour n — dernière commande tous clients

03/09/2018

jour 1 + jour 2 + jour j

première commande du client

deuxième commande du client

dernière commande du client

frequency = nombre commandes

satisfaction = moyenne (notes)

nombre notes

monetary = somme CA client

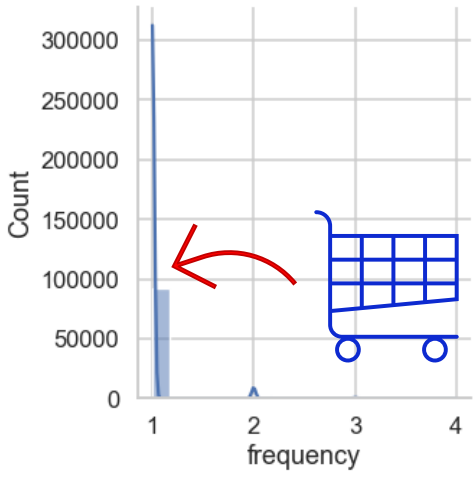
average basket = CA client / nb commandes

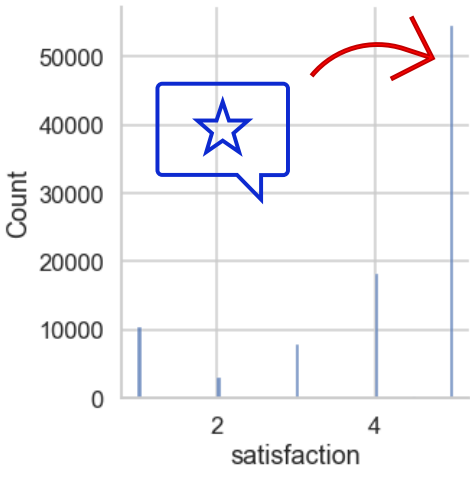
nb produits différents

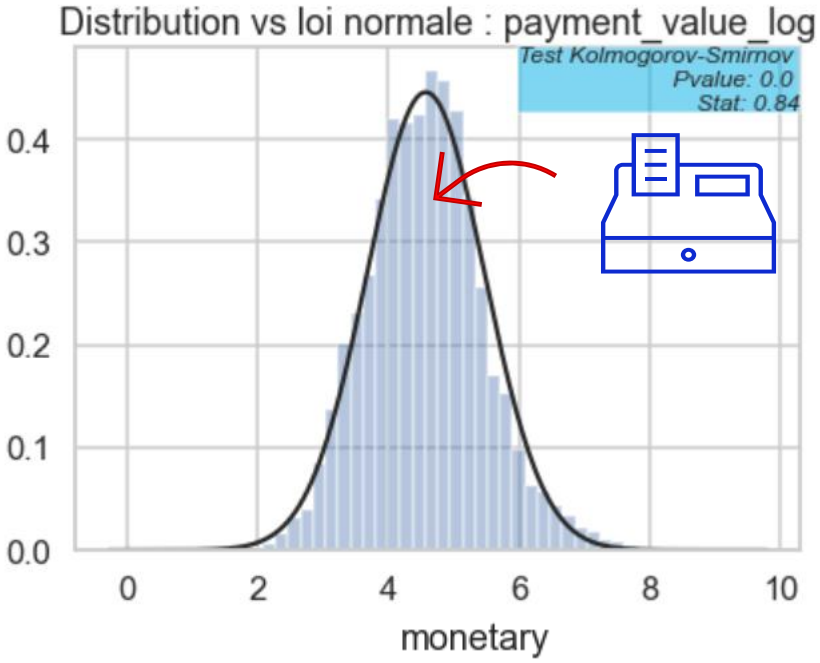
nb moyen d'articles

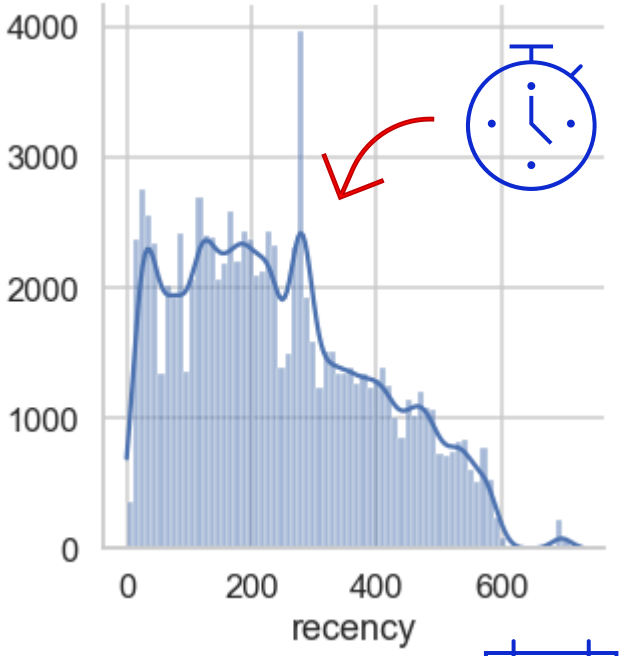
1 Analyse exploratoire

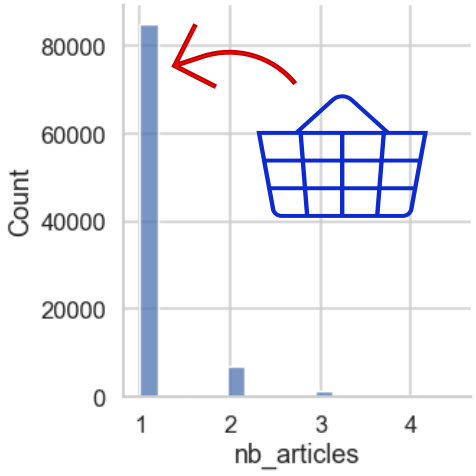
Analyse des distributions

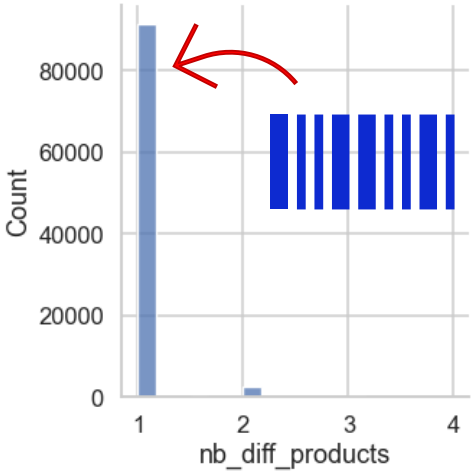


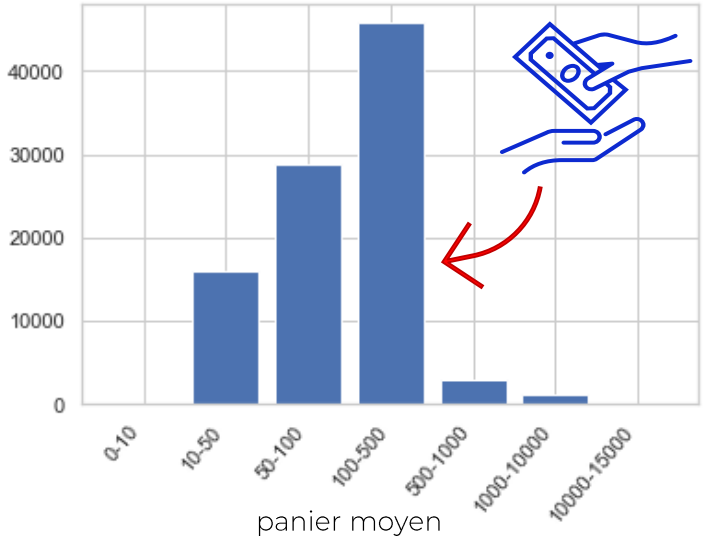


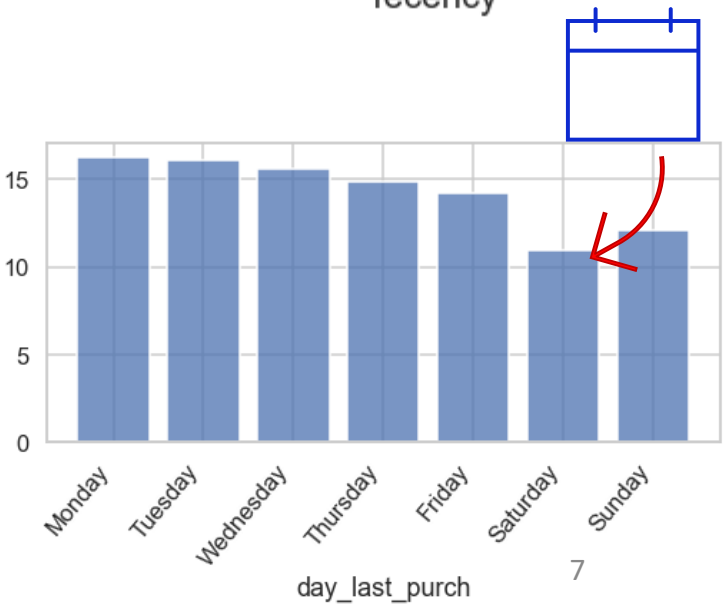






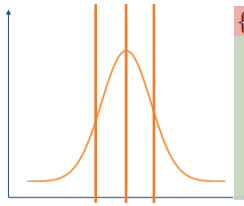






2 Modélisation / Essais

Segmentation classique



```
{'recency': {0.25: 119.0, 0.5: 224.0, 0.75: 353.0},
 'frequency': {0.25: 1.0, 0.5: 1.0, 0.75: 1.0},
 'monetary': {0.25: 63.1, 0.5: 107.85, 0.75: 182.91},
 'satisfaction': {0.25: 4.0, 0.5: 5.0, 0.75: 5.0},
 'nb_articles': {0.25: 1.0, 0.5: 1.0, 0.75: 1.0},
 'nb_diff_products': {0.25: 1.0, 0.5: 1.0, 0.75: 1.0},
 'aver_basket': {0.25: 62.39, 0.5: 105.71, 0.75: 176.98}}
```

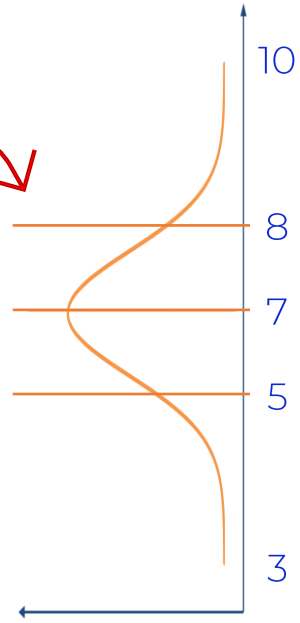
4 ∨ 1

1 ∧ 4

```
'satisfaction': {0.25: 4.0, 0.5: 5.0, 0.75: 5.0}
```

seulement 2 quantiles : 1 ou 2. Peu discriminant.

recency	R	satisfaction	S	monetary	M	RSM Group	RSM Score
116	4	5	2	142	3	423	9
119	4	4	1	27	1	411	6
293	2	5	2	197	4	224	8



Platinum

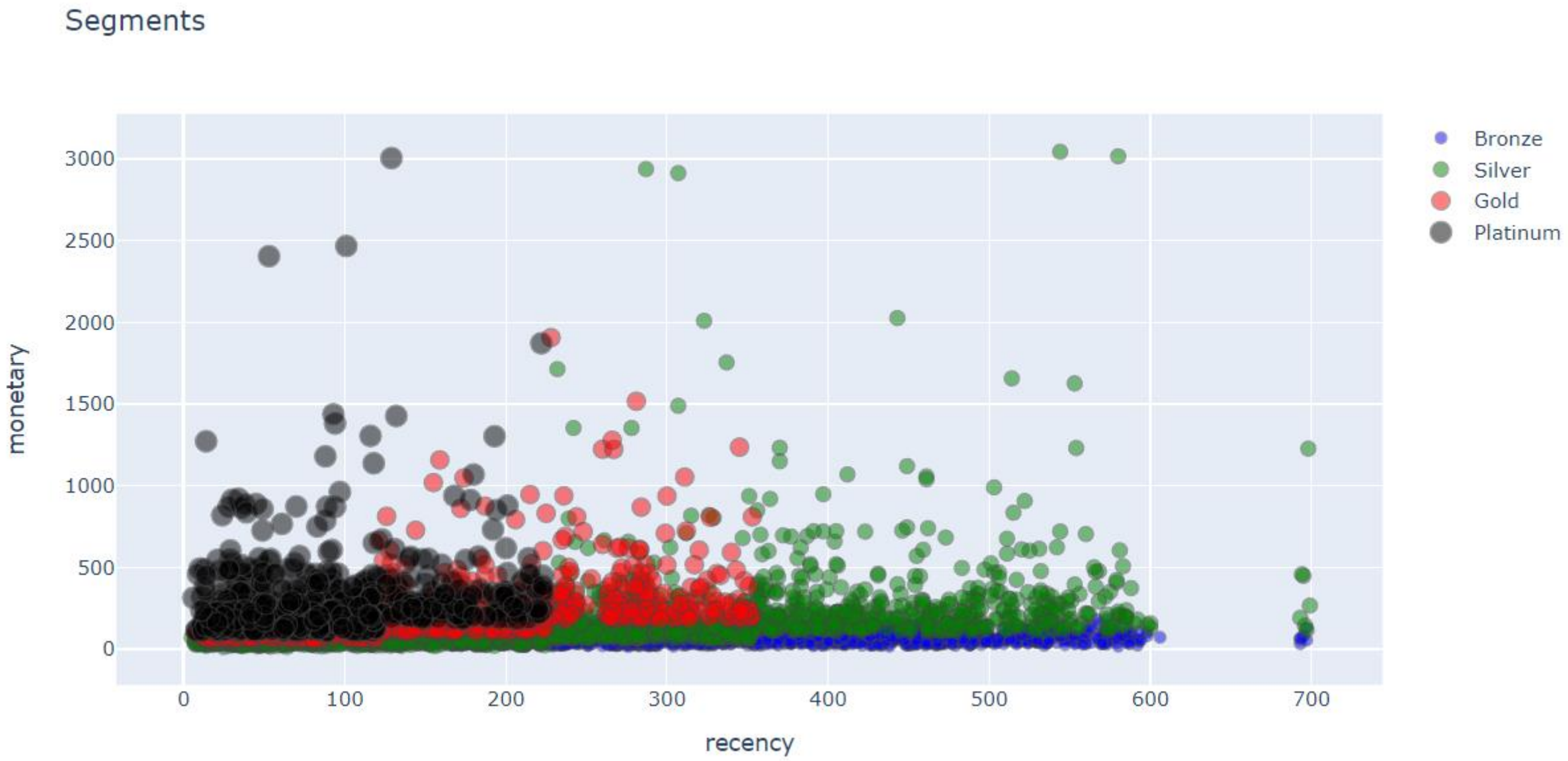
Gold

Silver

Bronze

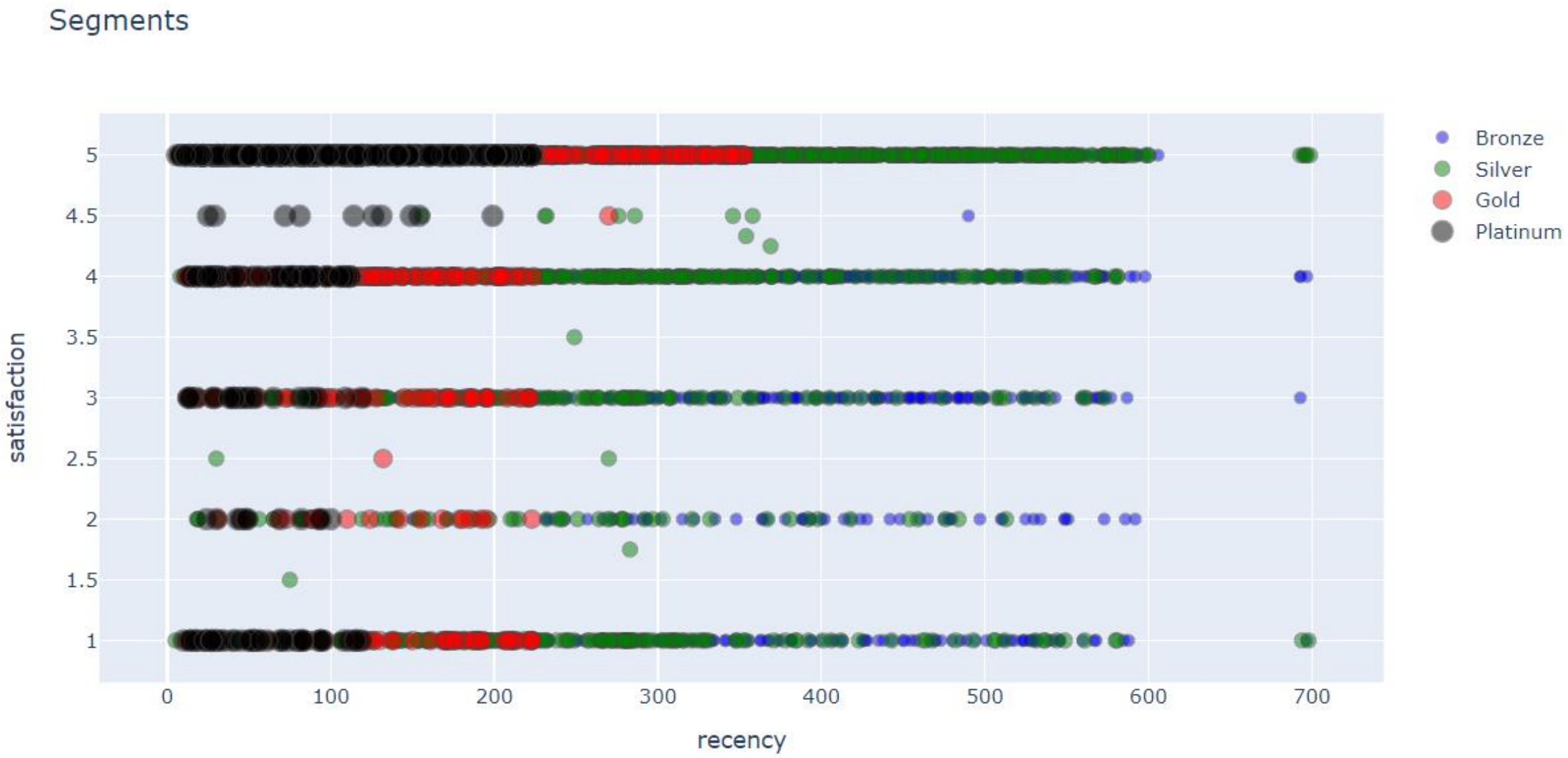
2 Modélisation / Essais

Segmentation classique



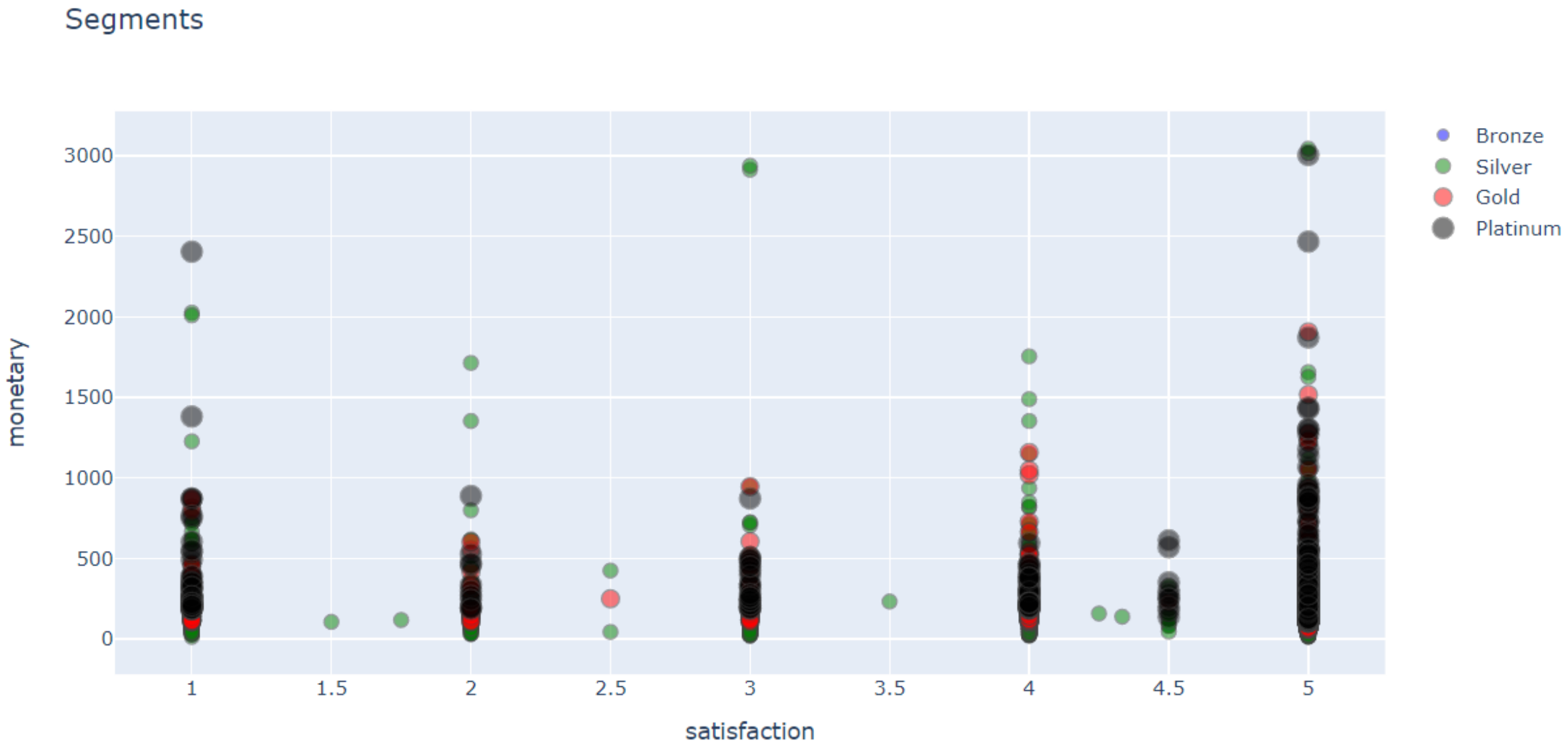
2 Modélisation / Essais

Segmentation classique



2 Modélisation / Essais

Segmentation classique



2 Modélisation / Essais

Segmentation classique

	Nombre clients	Recency (jours)	Monetary (R$)	Satisfaction (note)
Platinum	23 854	88	321	4,6
Gold	23 855	148	225	4,2
Silver	23 855	246	158	4,1
Bronze	23 855	371	66	3,8

++ récents / ++ acheteurs

+ récents / + acheteurs

- récents / - acheteurs

+ anciens / - acheteurs

2 Modélisation / Essais

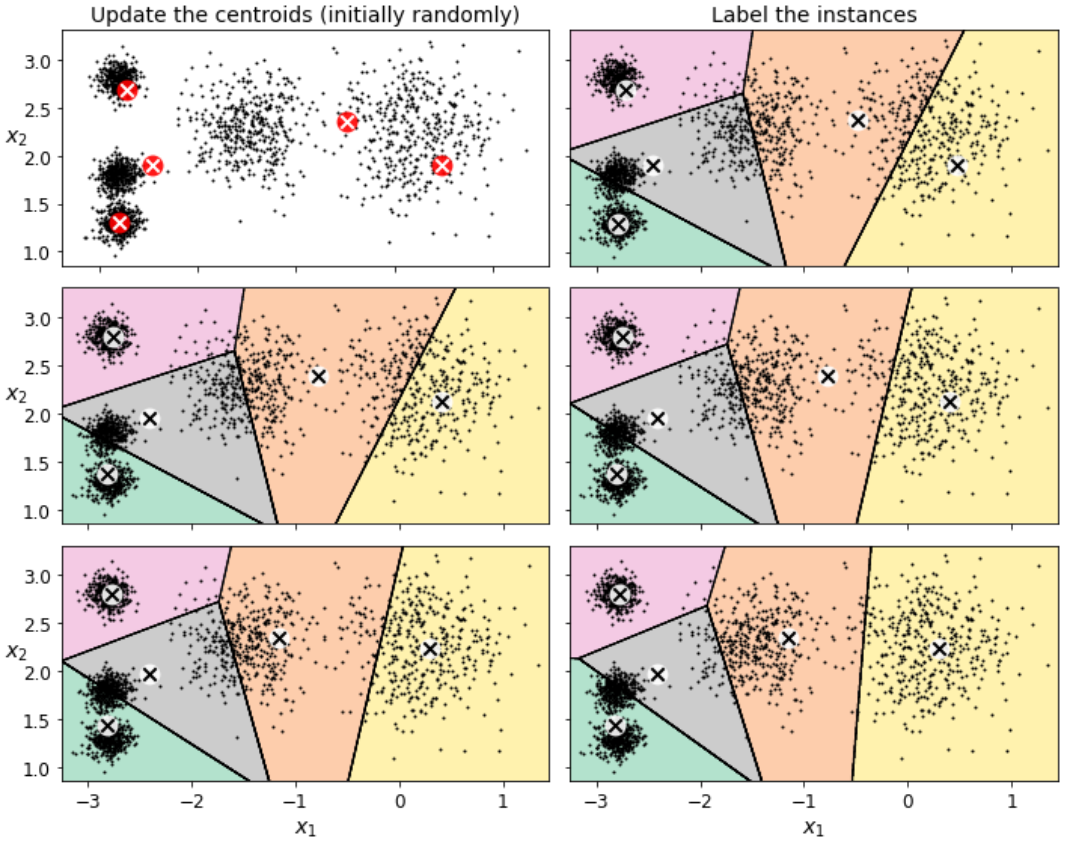
Clustering K-means 1

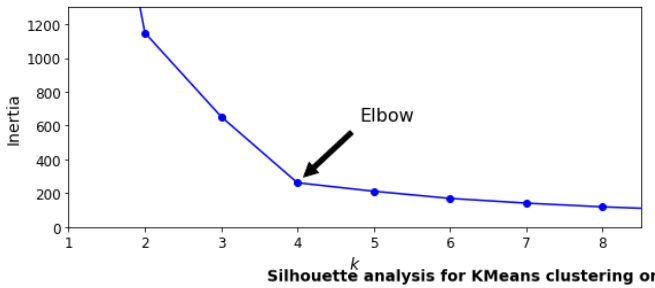
R A M S

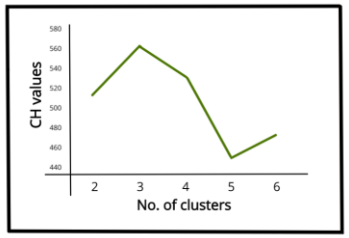
première approche
- Inertie

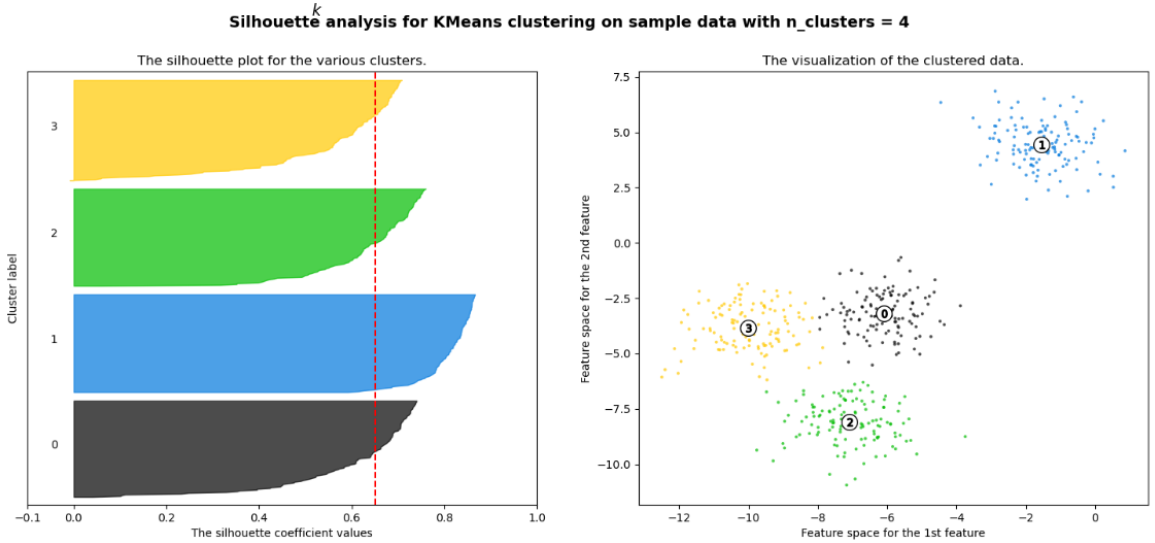
pré-sélection 2 candidats k
- Silhouette score

choix final du k
- Calinski Harabasz (interne)









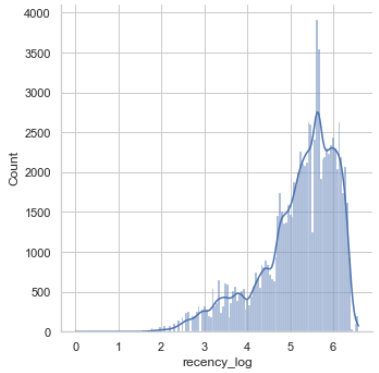
2 Modélisation / Essais

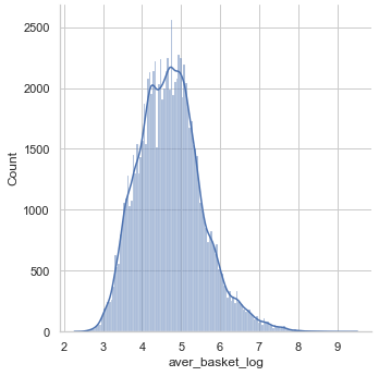
Clustering K-means 1

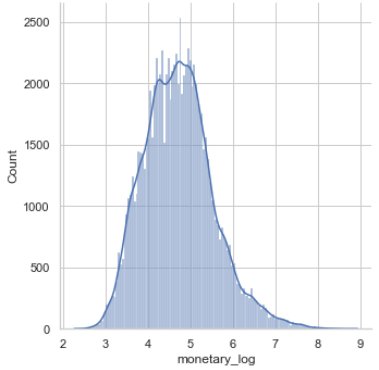
linéariser les cibles pour régression
- Log Transformation (RAM)

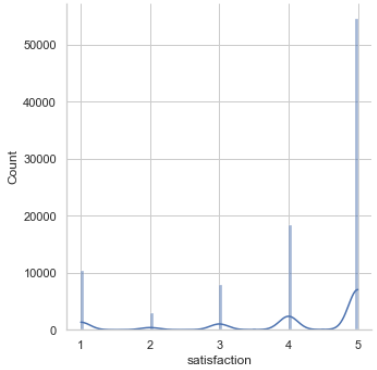
standardiser les données
- Centrage et réduction (StandardScaler)

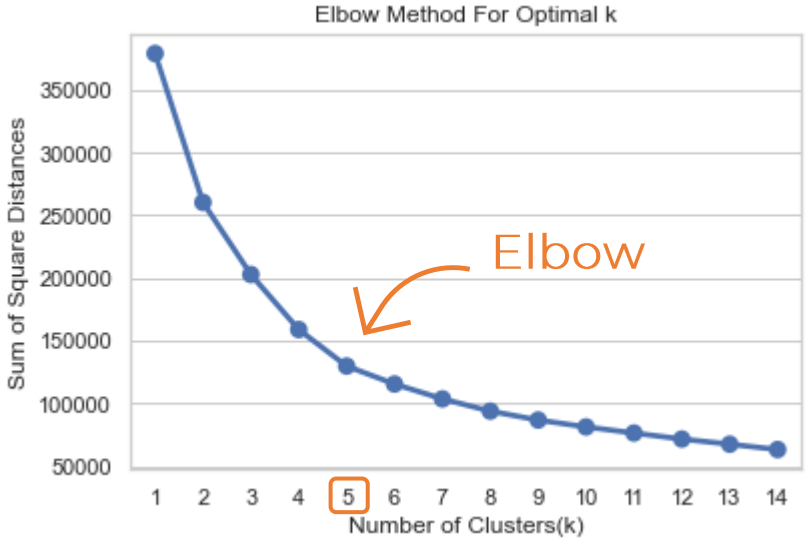
construction du modèle – choix du k
- Inertie

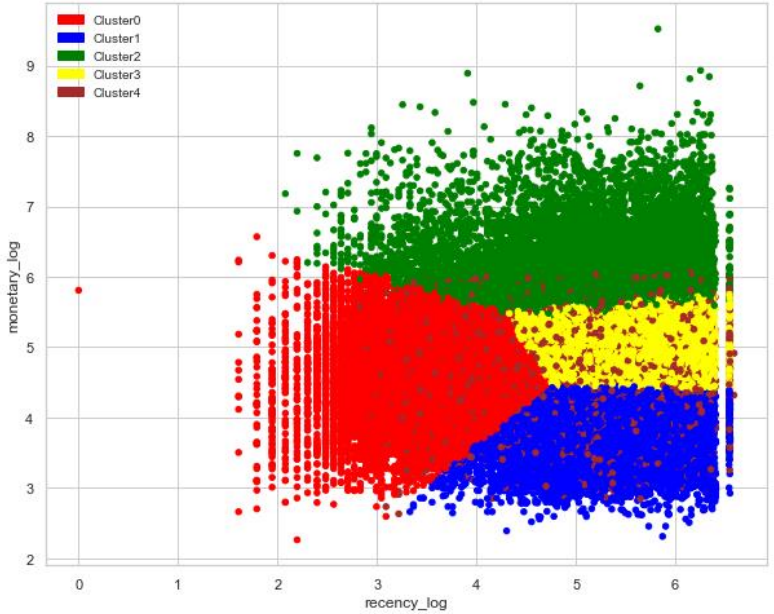


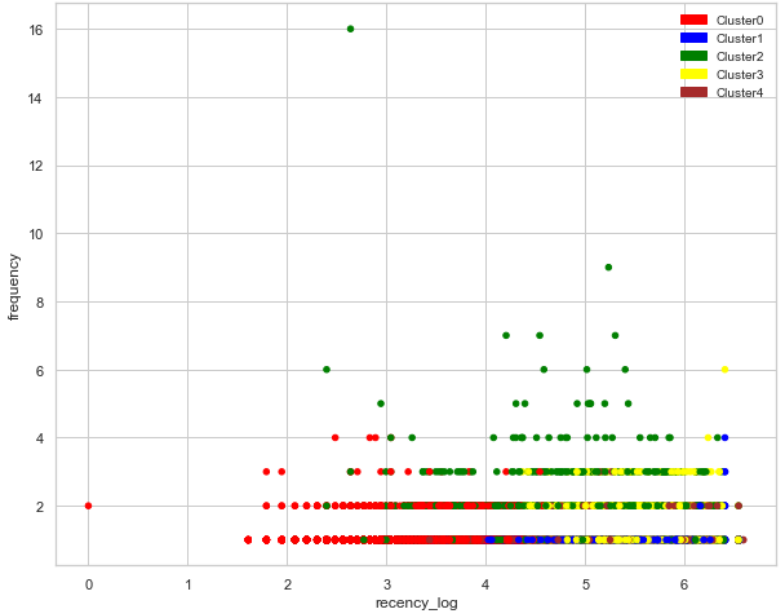


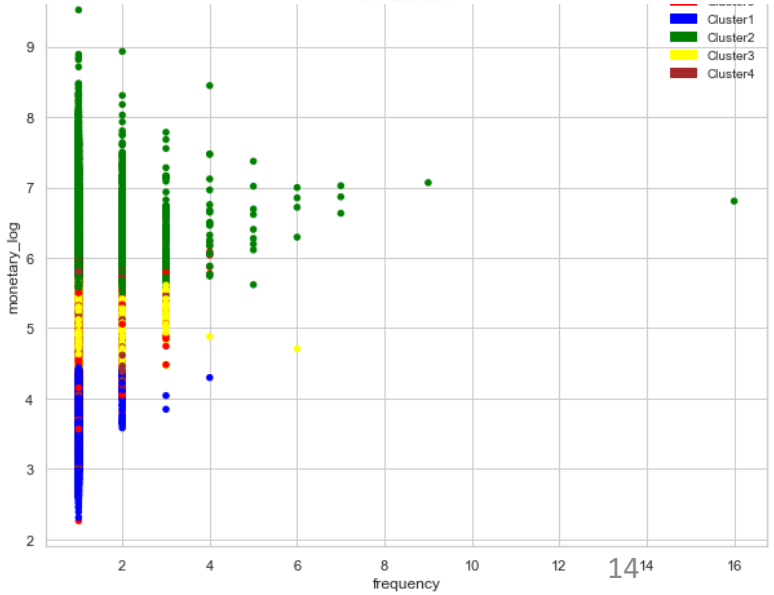












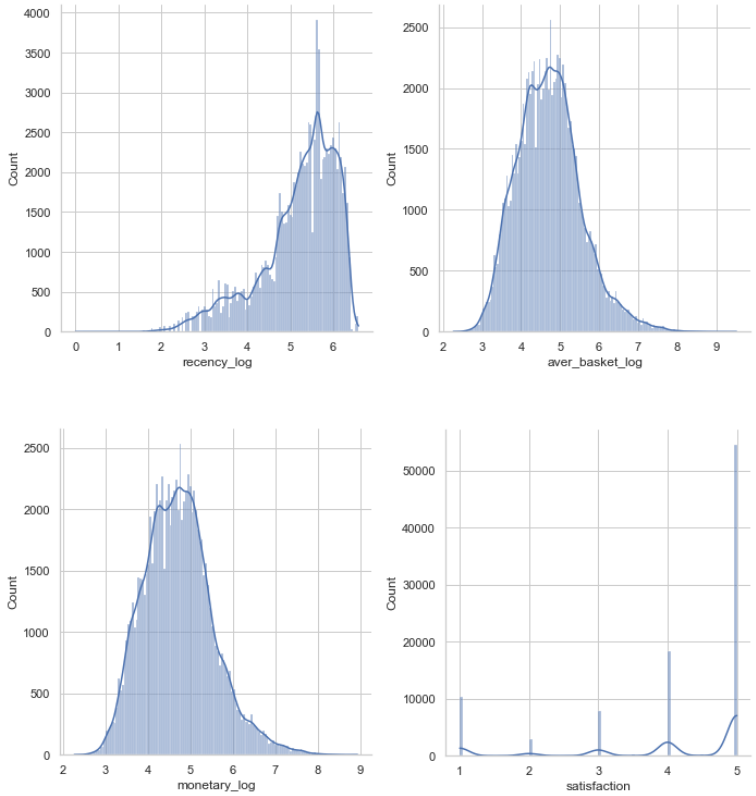
2 Modélisation / Essais

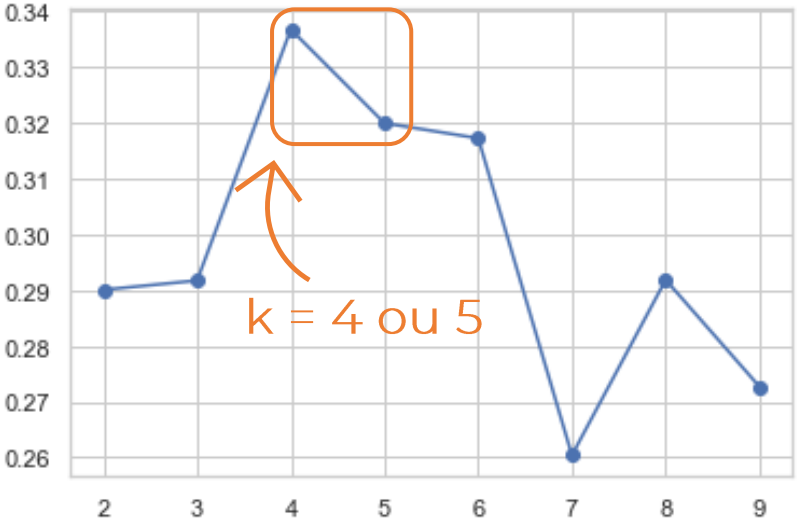
Clustering K-means 1

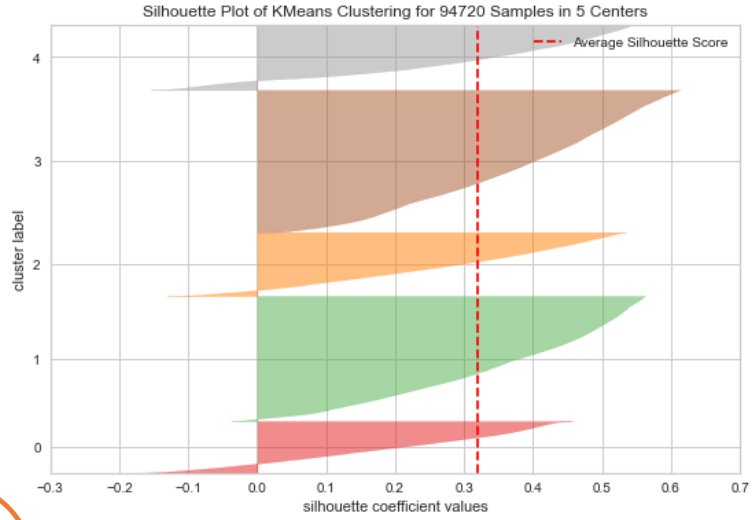
linéariser les cibles pour régression
- Log Transformation (RAM)

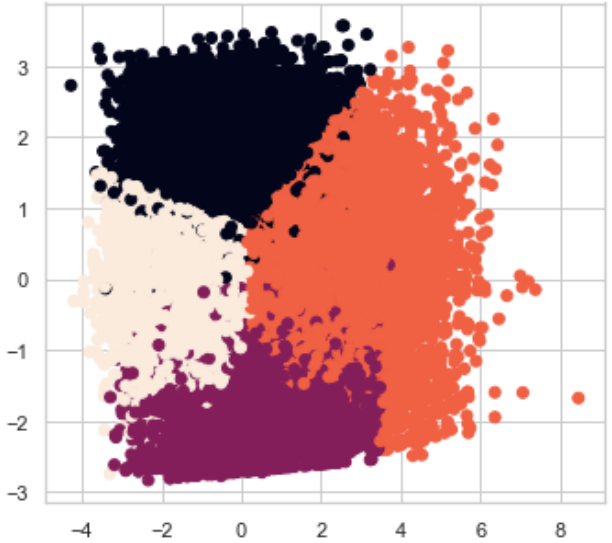
standardiser les données
- Centrage et réduction (StandardScaler)

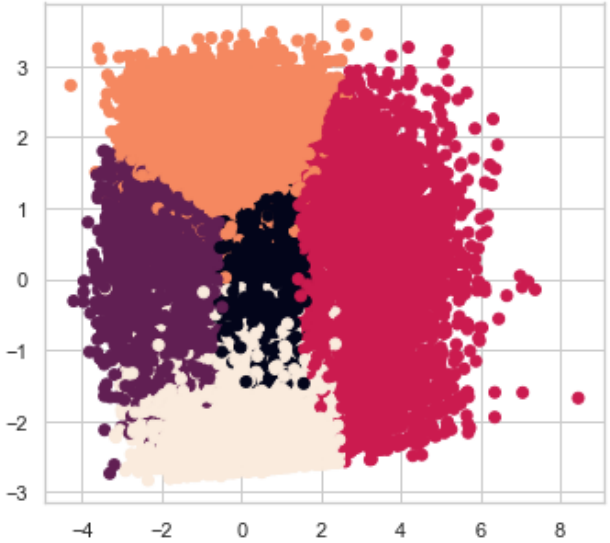
construction du modèle – choix du k
- Silhouette Score





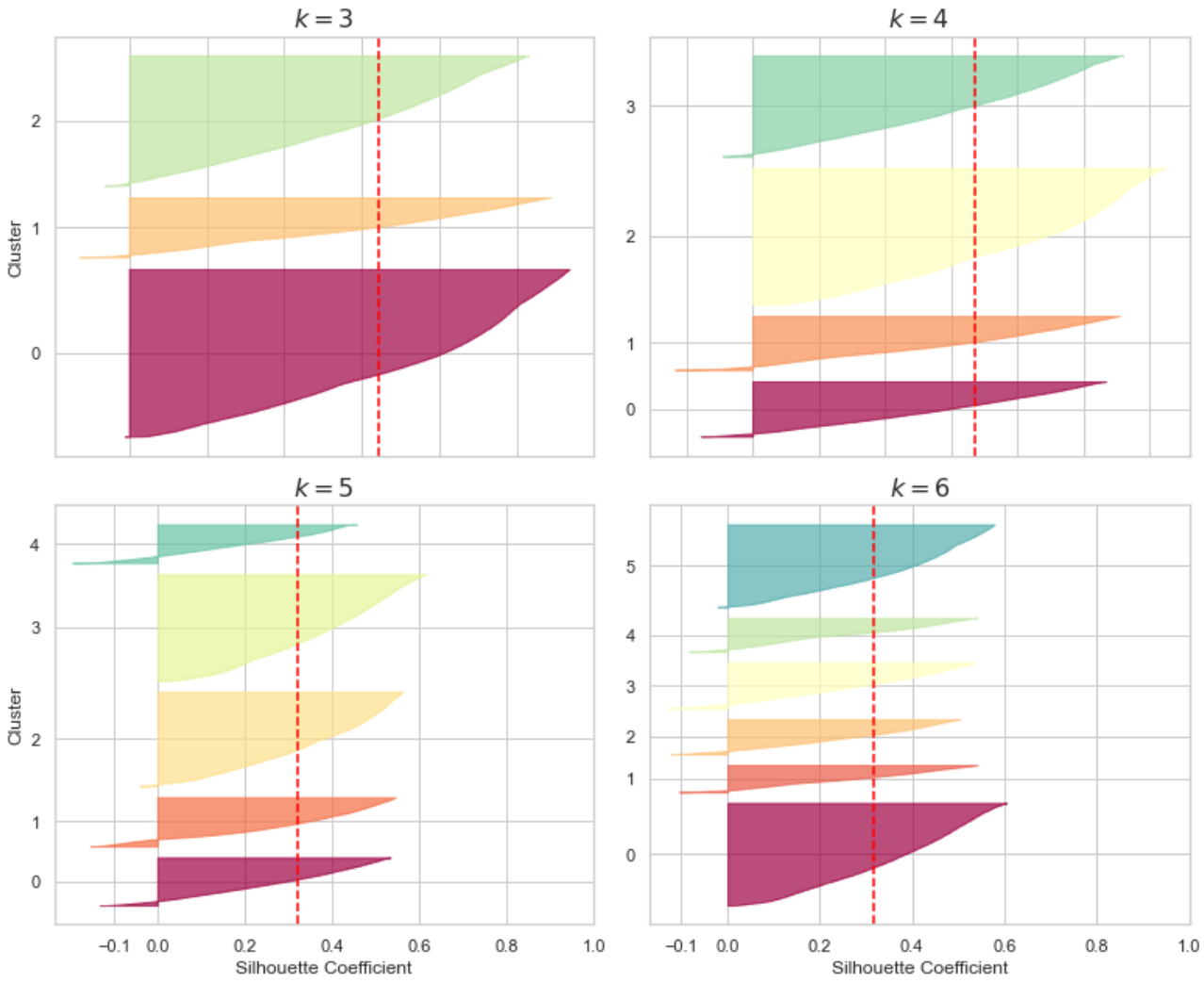


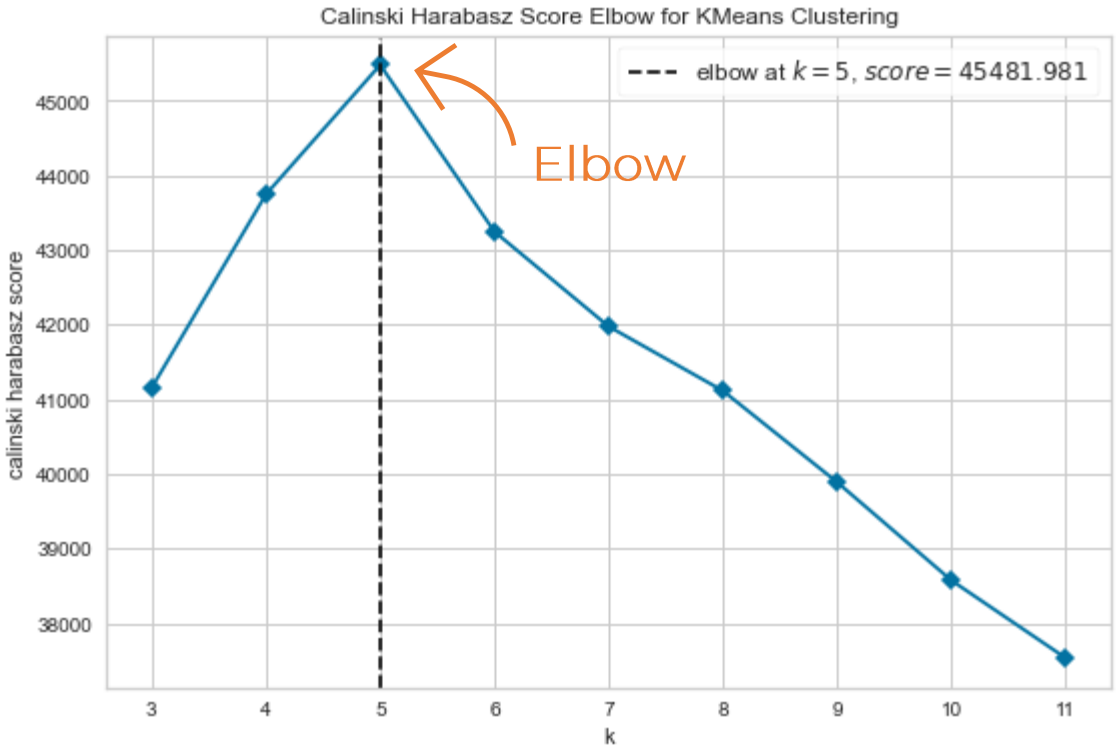




2 Modélisation / Essais

Clustering K-means 1

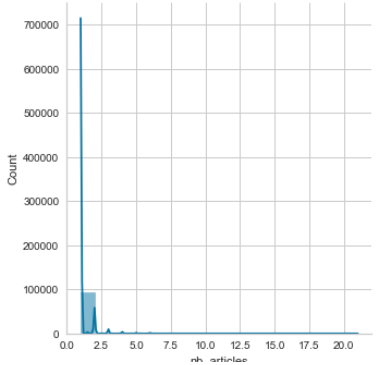


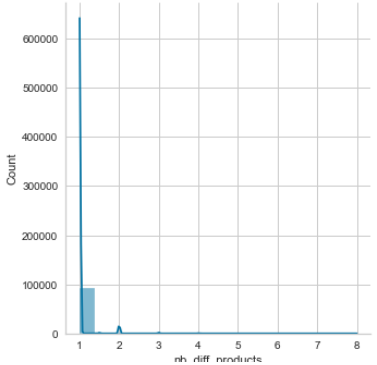


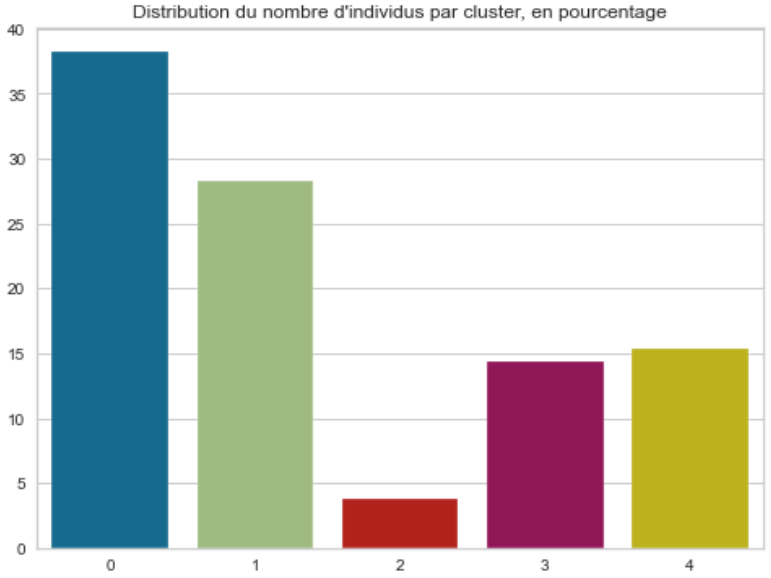
2 Modélisation / Essais

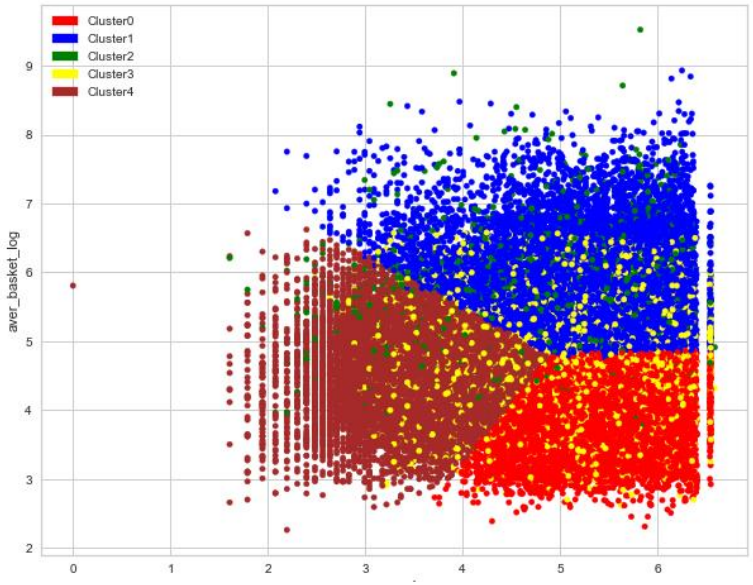
Clustering K-means 2

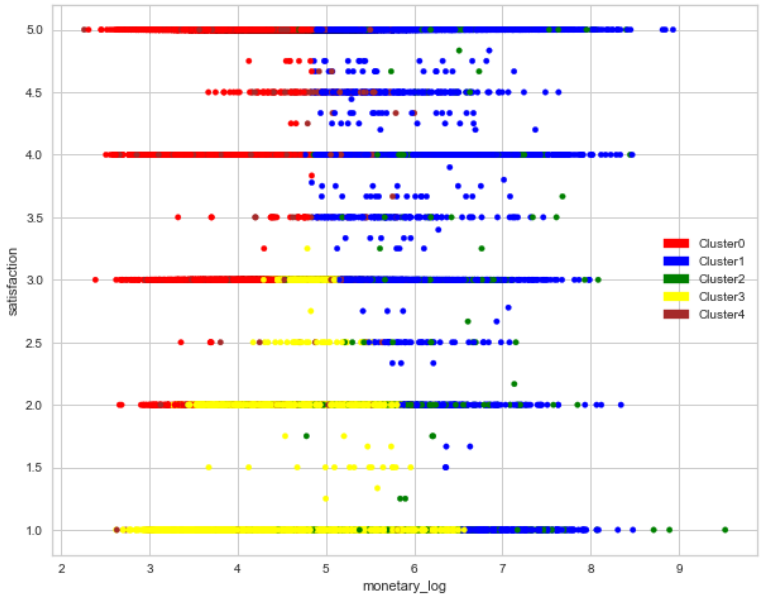
R A M S + NA D

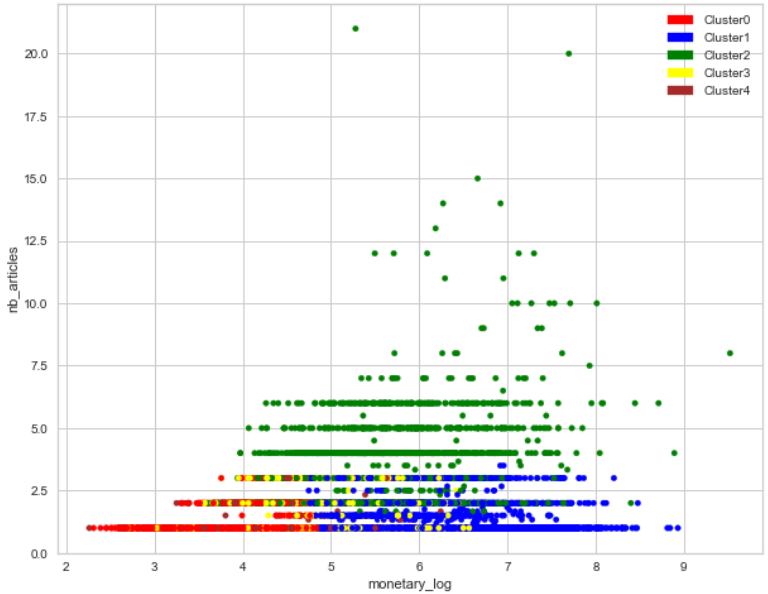


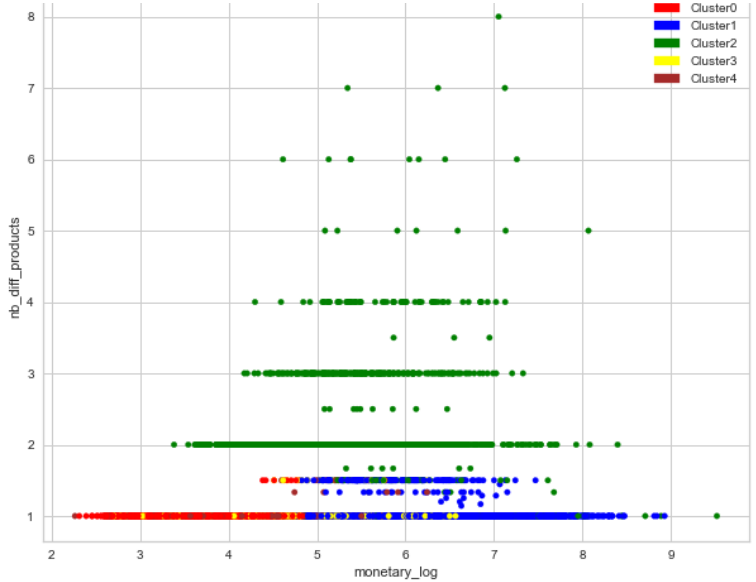






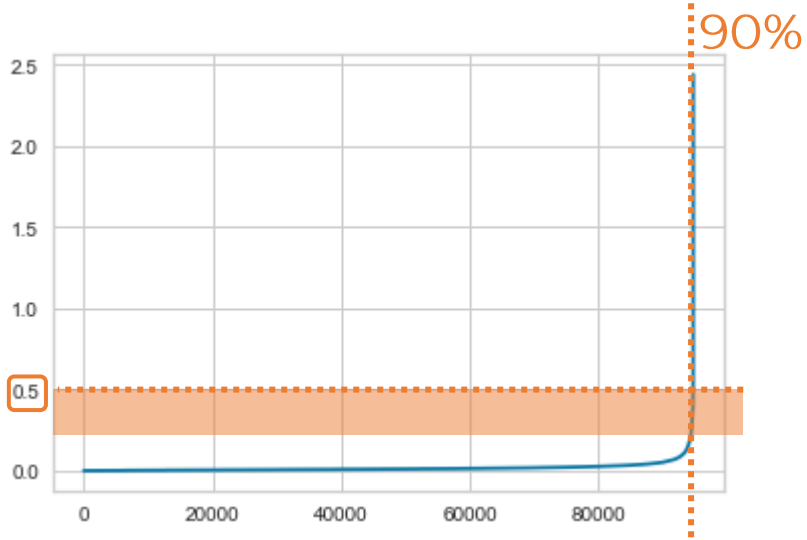


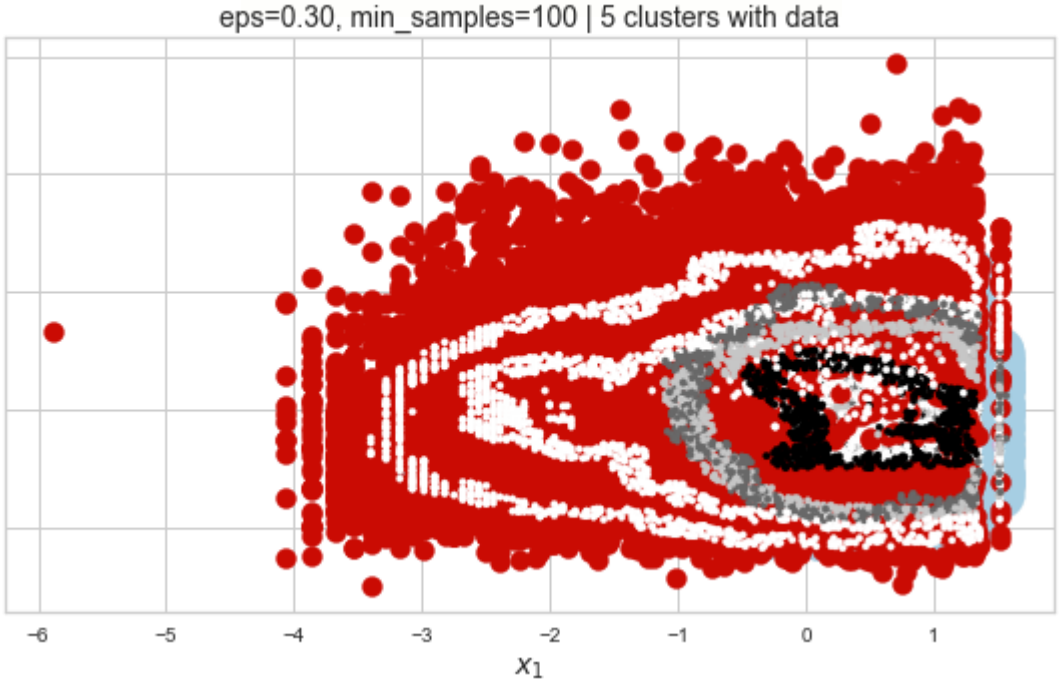


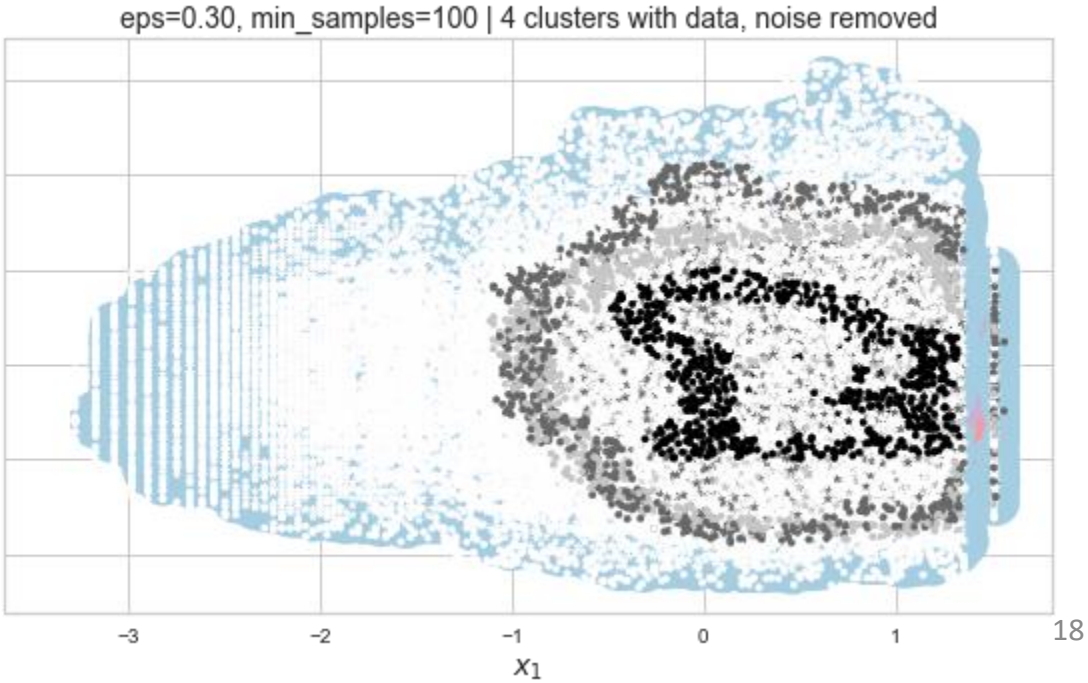


2 Modélisation / Essais

Clustering DBScan



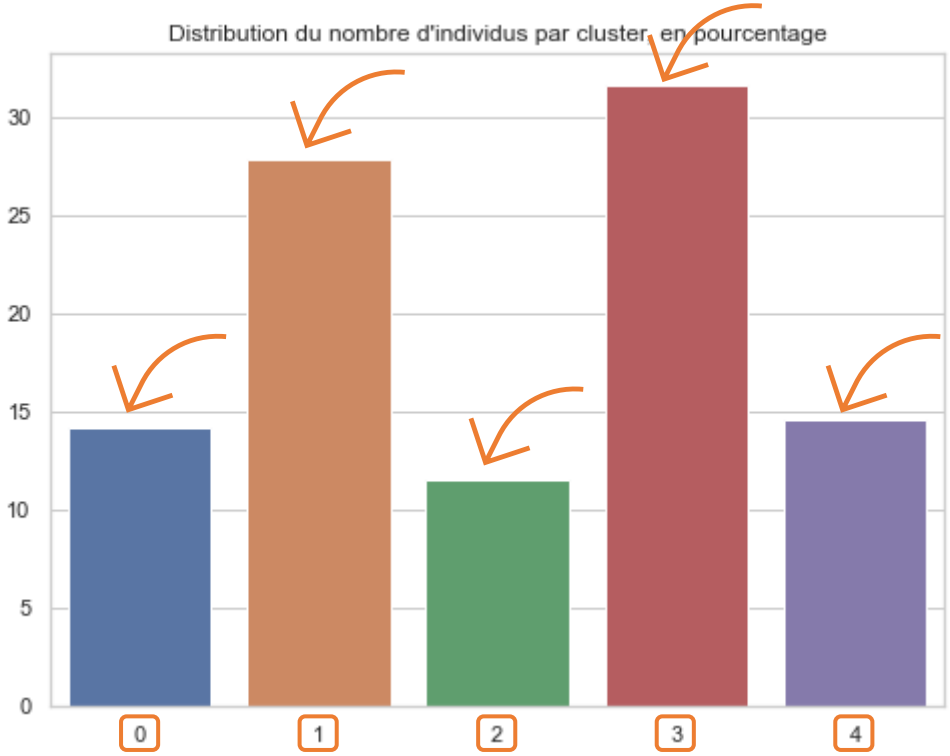


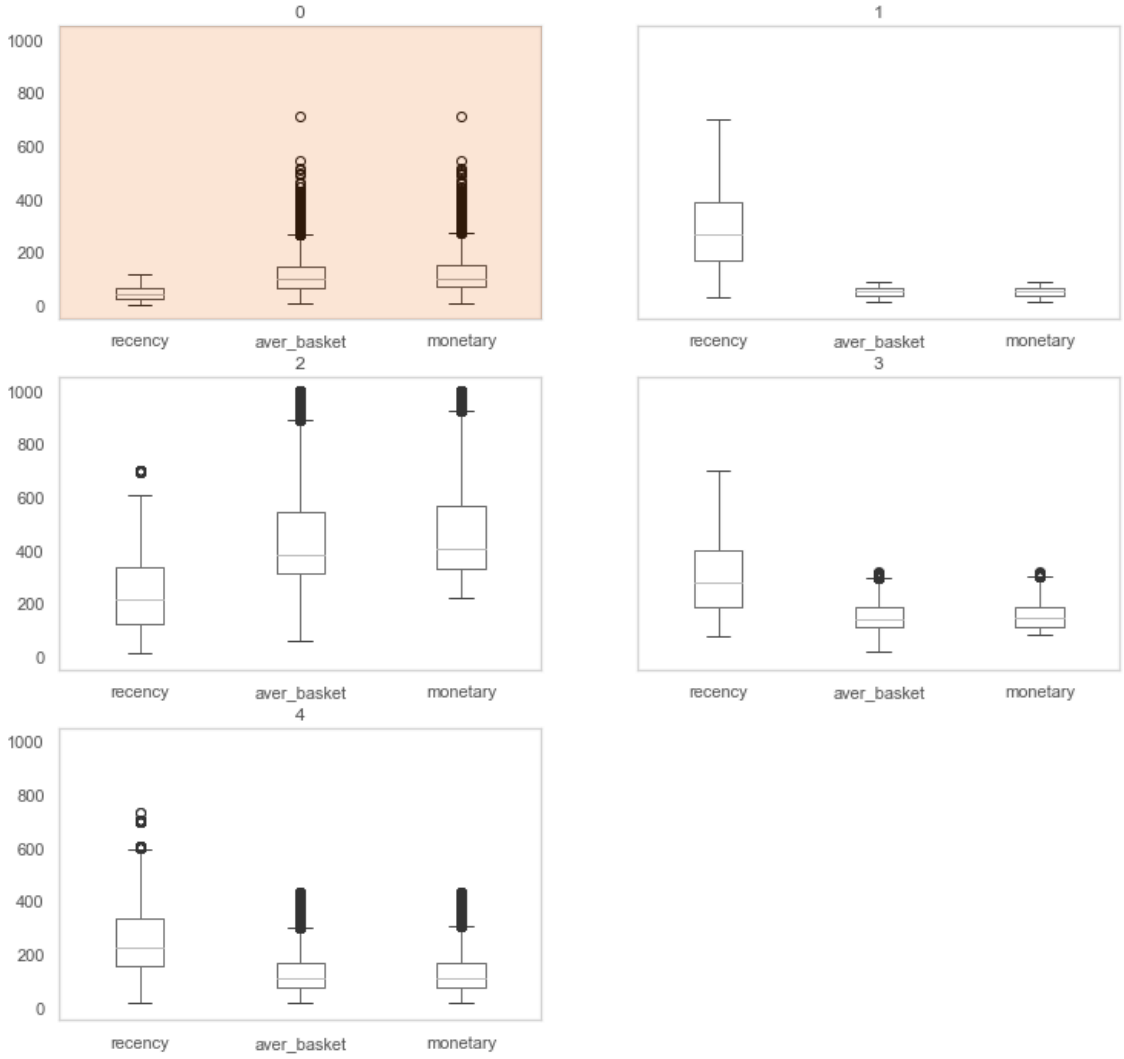


2 Modélisation / Essais

Modèle retenu

Modèle retenu : K-means
Nombre de clusters : 5
Métrique : Calinski Harabasz





3 Stabilité / Maintenance

Méthode retenue

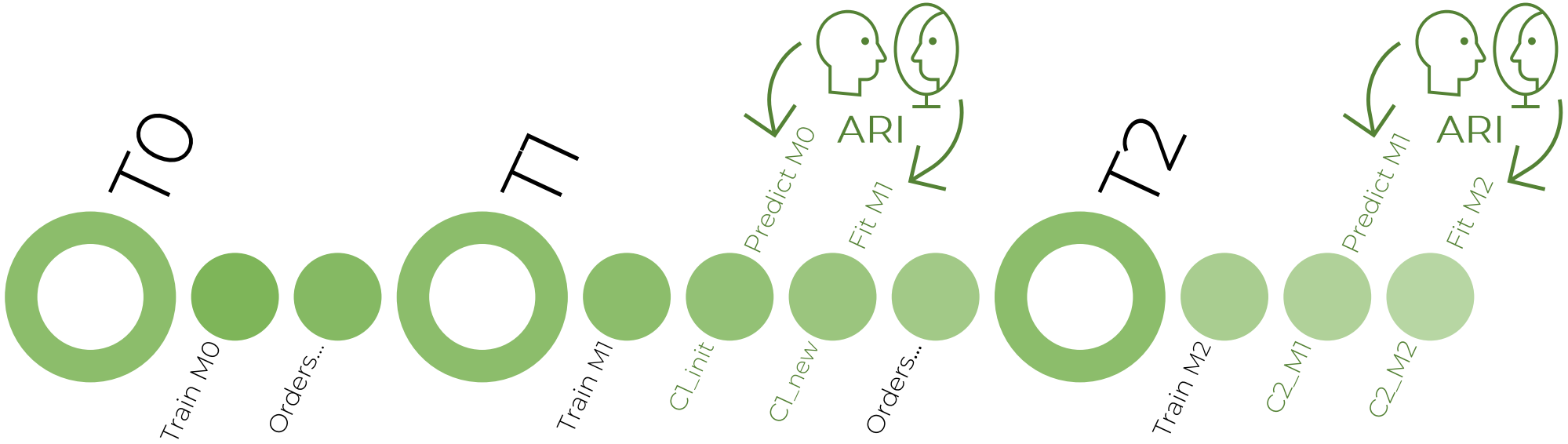
Modèle retenu : K-means
Nombre de clusters : 5
Métrique : Calinski Harabasz
Features : RAMS

order_purchase_timestamp	order_purchase_timestamp	
2016	3	2
	4	304
2017	1	5122
	2	9192
	3	12404
	4	17487
2018	1	20918
	2	19831
	3	12656

Fréquence départ 4 mois

date plus récente = 03/09/2018

dates	4 mois	3 mois	2 mois
T0	31/12/2017	28/02/2018	30/04/2018
T1	30/04/2018	31/05/2018	30/06/2018
T2	31/08/2018	31/08/2018	31/08/2018



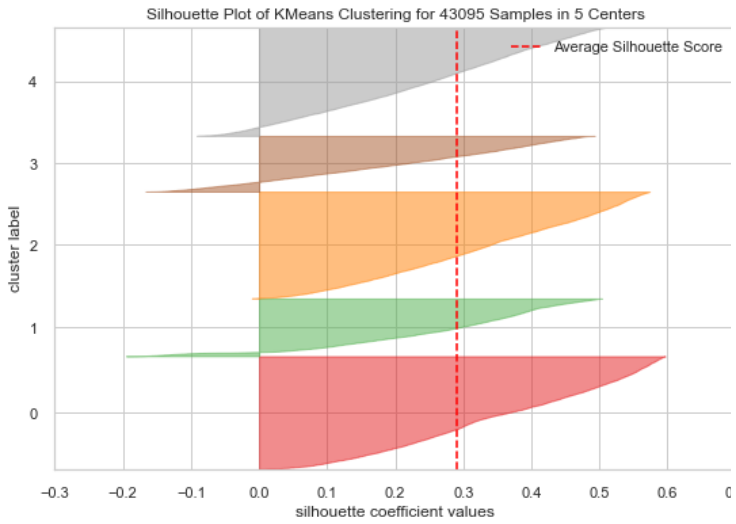
3 Stabilité / Maintenance

Fréquence = 4 mois

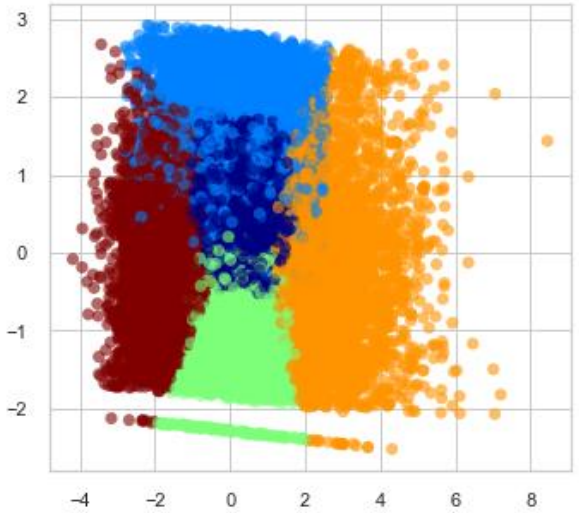
T0 | 31/12/2017
T1 | 30/04/2018

à T1

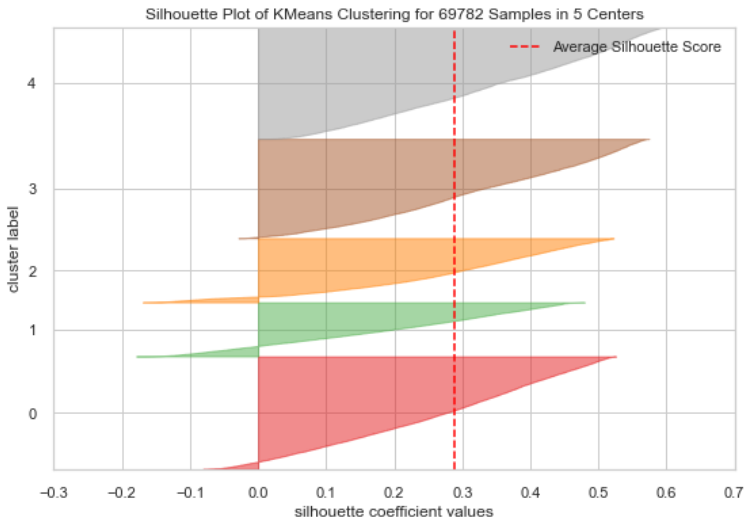
à T0 | C0 sur modèle M0
43 095 clients uniques



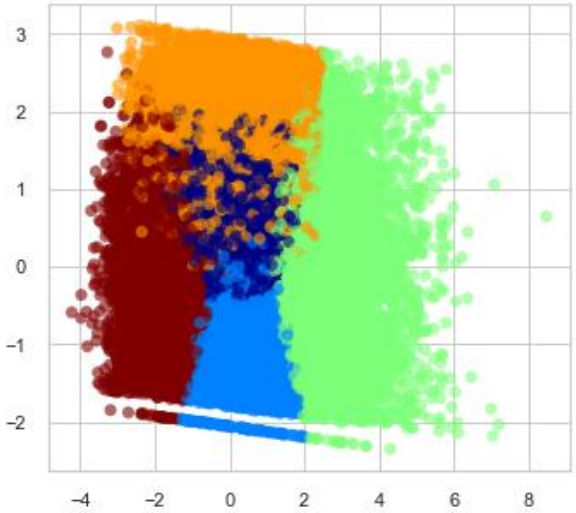
silhouette score = 0.29



à T1 | C1_new base M1
69 782 clients uniques



silhouette score = 0.289



C1_init
(predict M0)

C1_new
(fit M1)

Adjusted Rand Index
0.8 (ok >= 0.8)

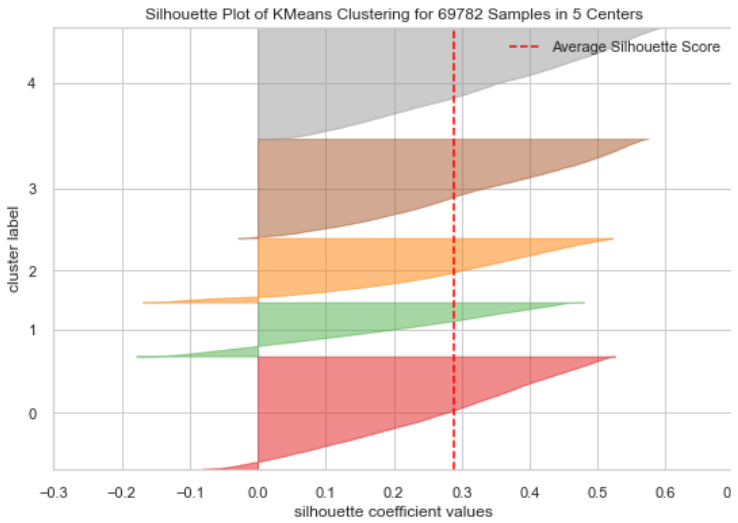
On vérifie à T2

3 Stabilité / Maintenance

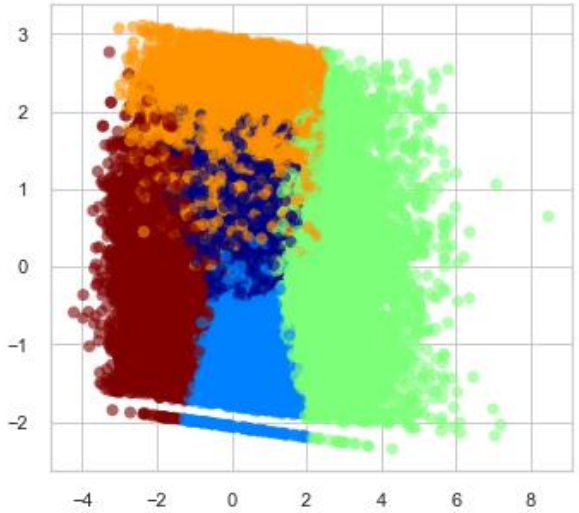
Fréquence = 4 mois à T2

T1 | 30/04/2018
T2 | 31/08/2018

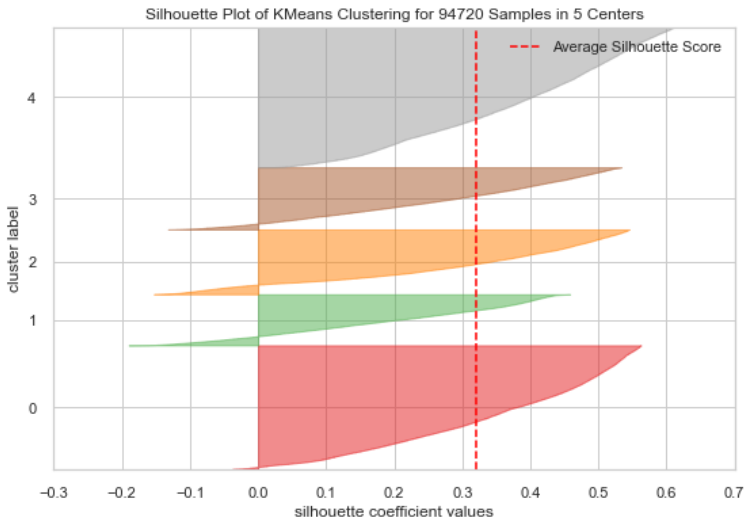
à T1 | C1_new base M1
69 782 clients uniques



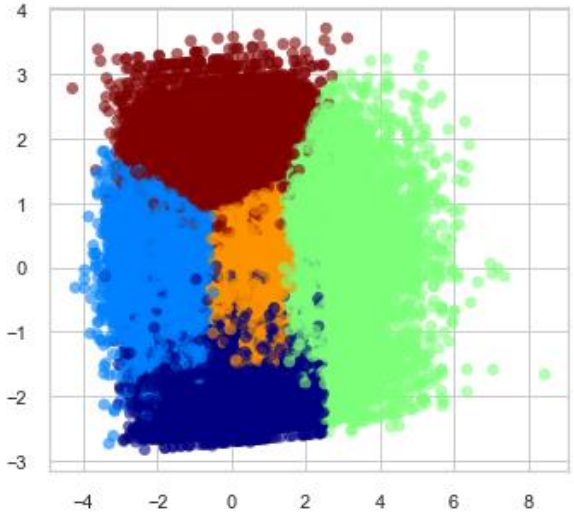
silhouette score = 0.289



à T2 | C2_M2 base M2
94 720 clients uniques



silhouette score = 0.321



C2_M1 (predict M1)

C2_new (fit M2)

Adjusted Rand Index
0.7 (ko < 0.8)

On diminue la fréquence…

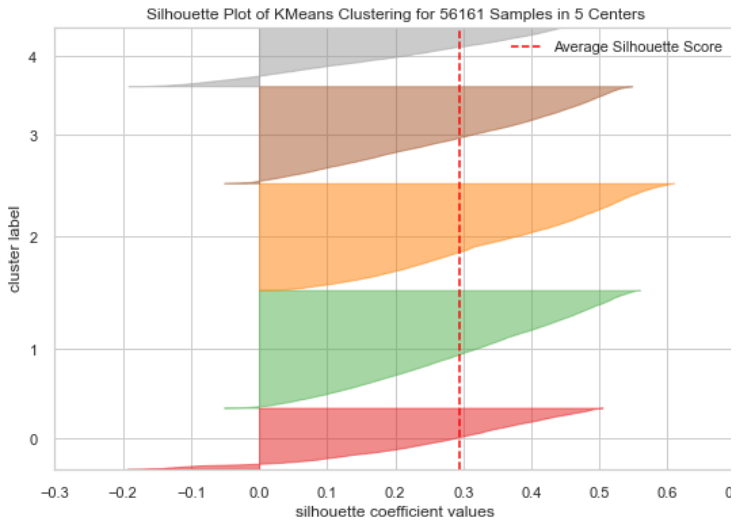
3 Stabilité / Maintenance

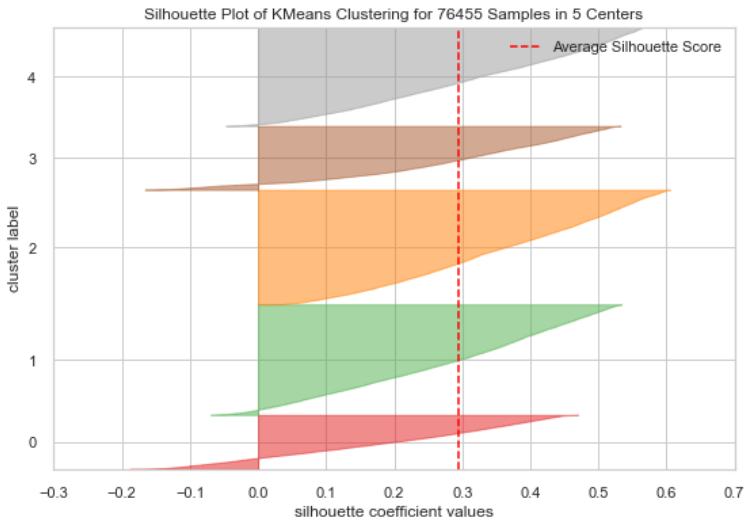
Fréquence = 3 mois

à T0 | C0 sur modèle M0
56 161 clients uniques

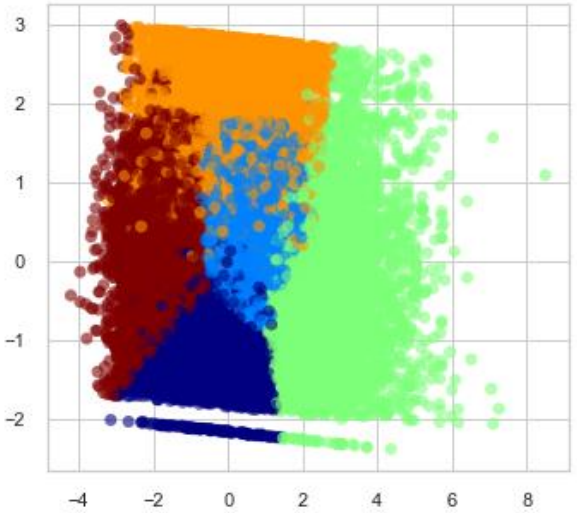
à T1 | C1_new base M1
76 455 clients uniques

T0 | 28/02/2018
T1 | 31/05/2018

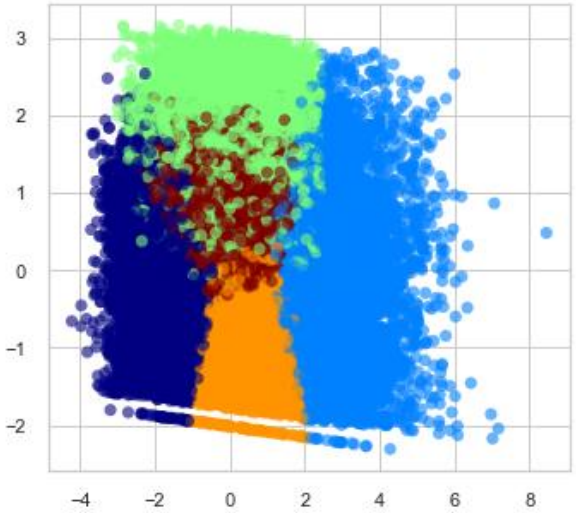




silhouette score = 0.294



silhouette score = 0.295



C1_init
(predict M0)

C1_new
(fit M1)

Adjusted Rand Index
0.48 (ko < 0.8)

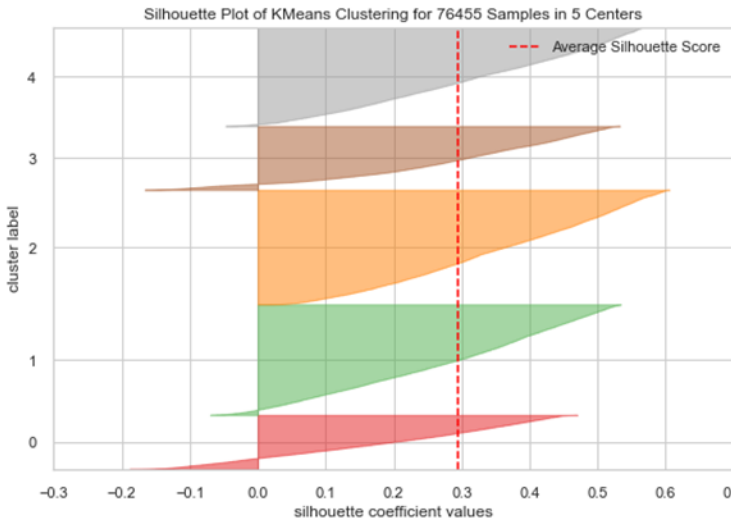
On vérifie à T2

3 Stabilité / Maintenance

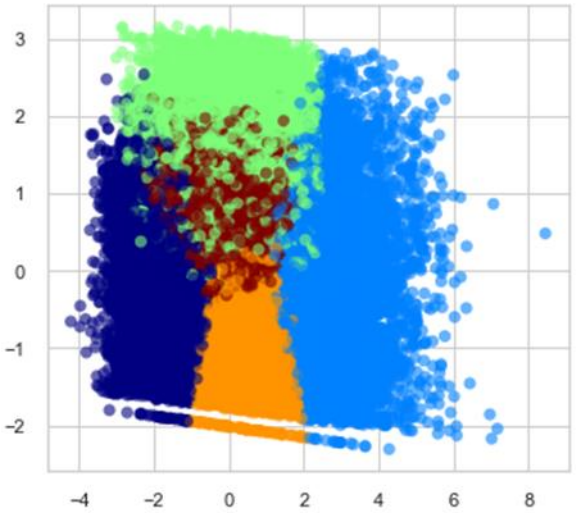
Fréquence = 3 mois à T2

T1 | 31/05/2018
T2 | 31/08/2018

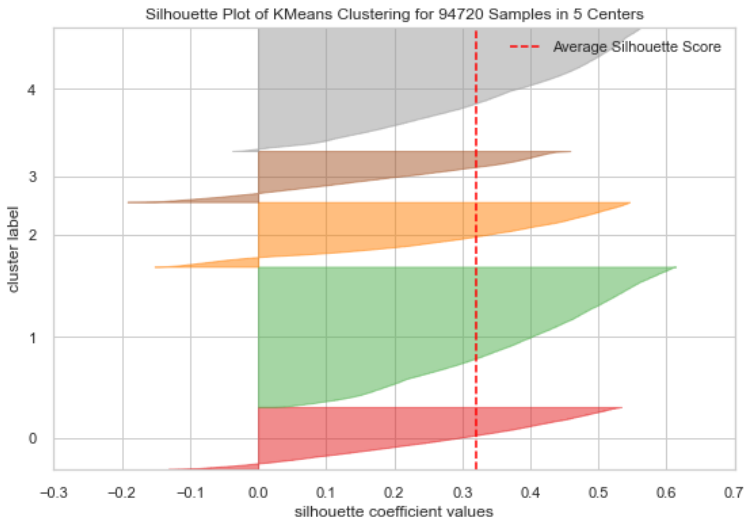
à T1 | C1_new base M1
76 455 clients uniques



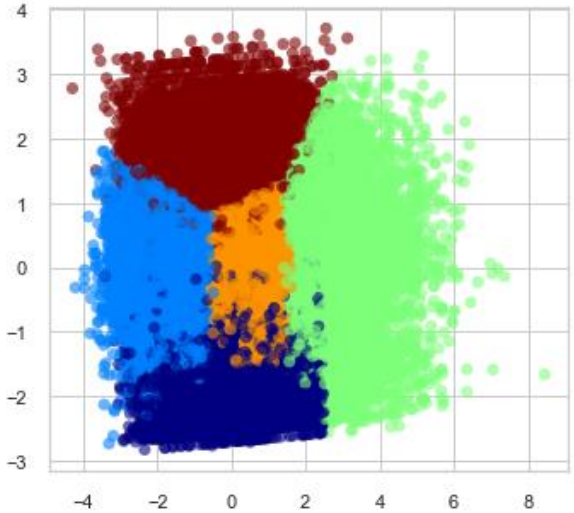
silhouette score = 0.295



à T2 | C2_M2 base M2
94 720 clients uniques



silhouette score = 0.321



C2_M1 (predict M1)

C2_new (fit M2)

Adjusted Rand Index
0.77 (ko < 0.8)

On diminue la fréquence…

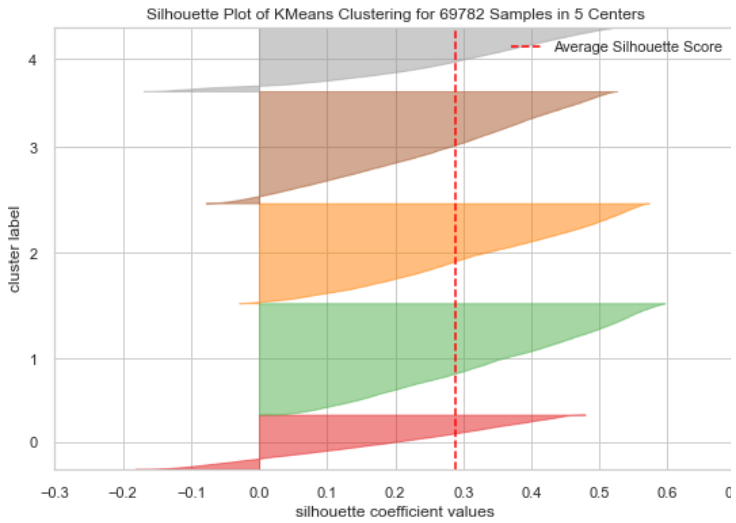
3 Stabilité / Maintenance

Fréquence = 2 mois

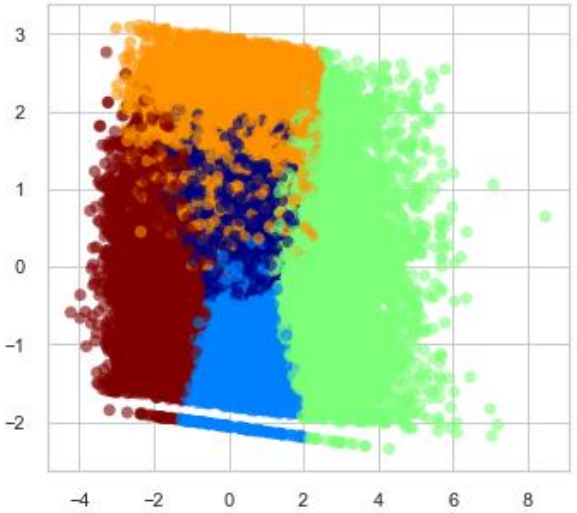
T0 | 30/04/2018
T1 | 30/06/2018

à T1

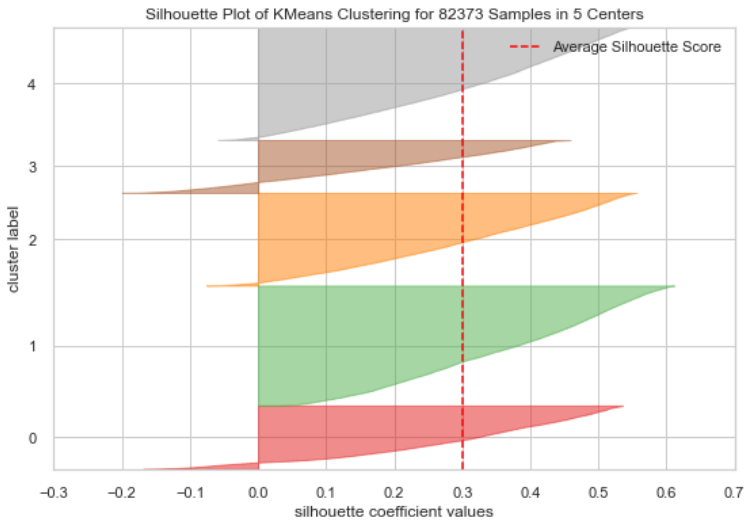
à T0 | C0 sur modèle M0
69 782 clients uniques



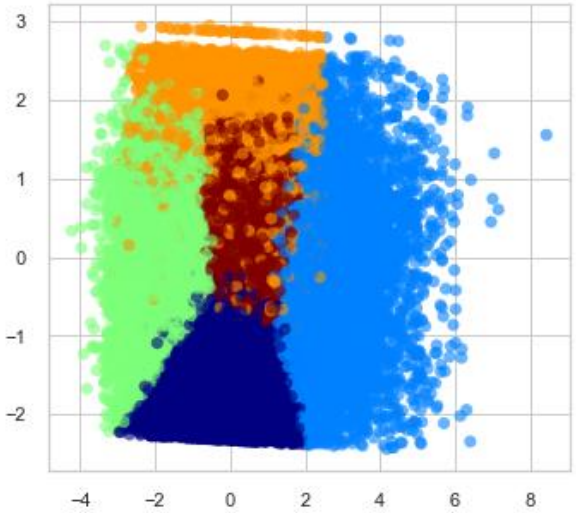
silhouette score = 0.289



à T1 | C1_new base M1
82 373 clients uniques



silhouette score = 0.301



C1_init
(predict M0)

C1_new
(fit M1)

Adjusted Rand Index
0.83 (ok >= 0.8)

On vérifie à T2

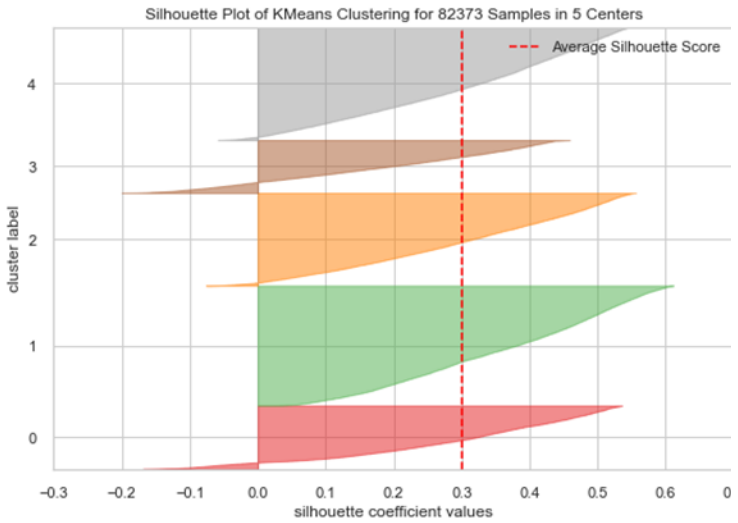
3 Stabilité / Maintenance

Fréquence = 2 mois

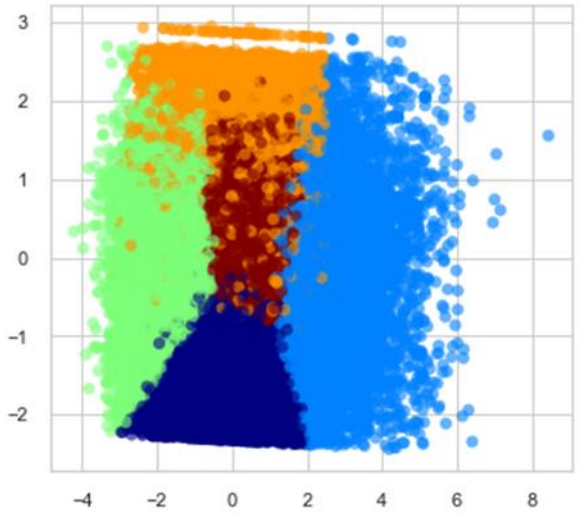
T1 | 30/04/2018
T2 | 31/08/2018

à T2

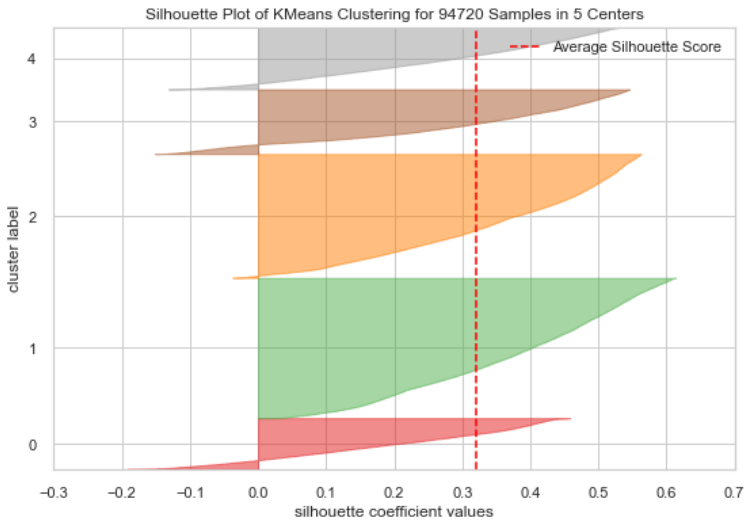
à T1 | C1_new base M1
82 373 clients uniques



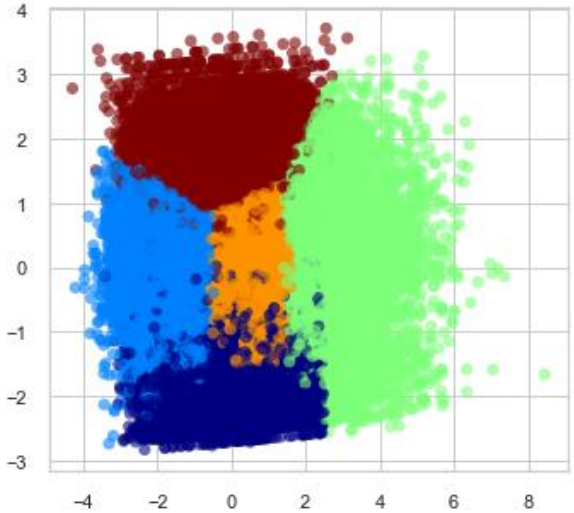
silhouette score = 0.301



à T2 | C2_M2 base M2
94 720 clients uniques



silhouette score = 0.321



C2_M1 (predict M1)

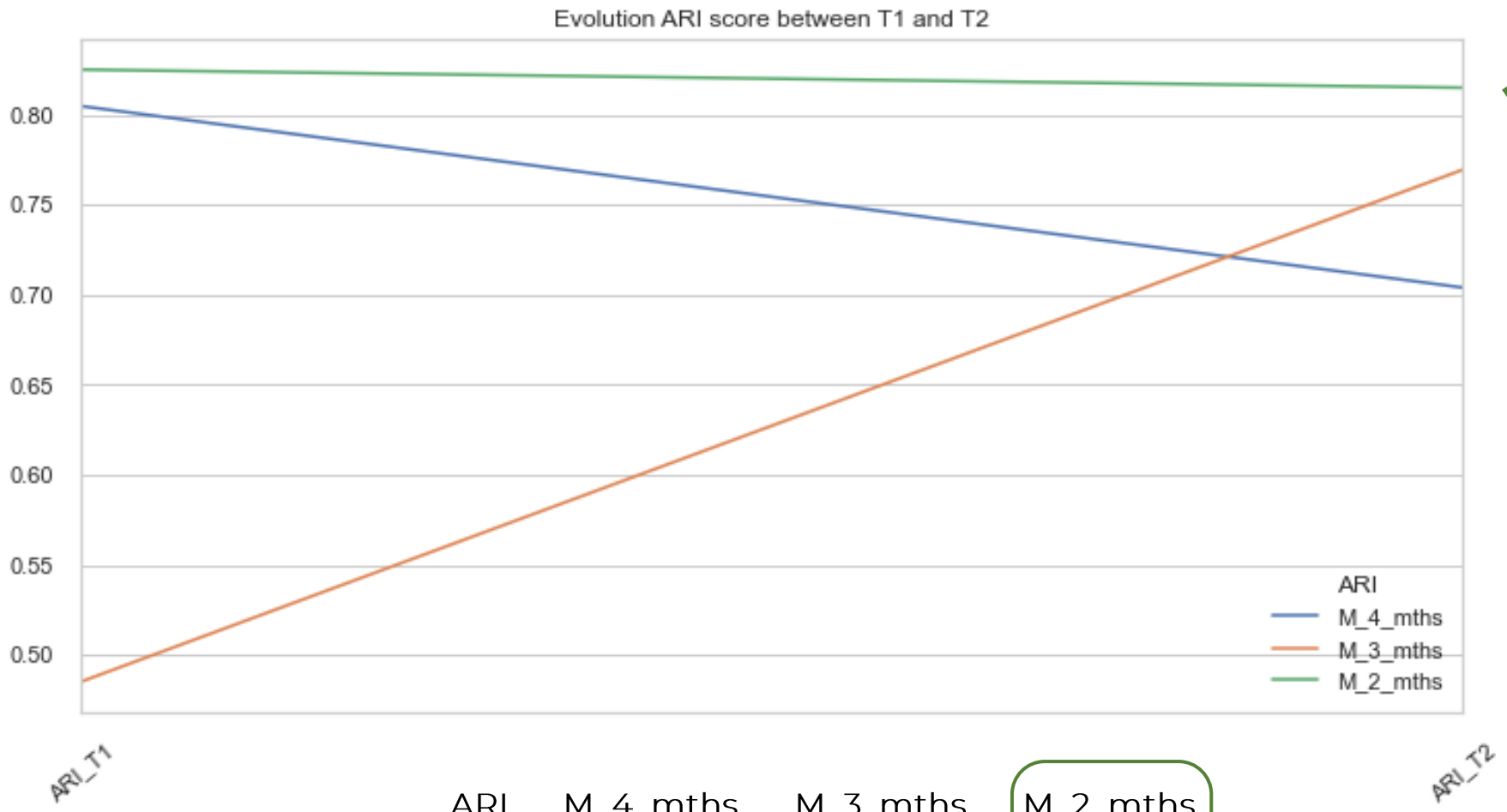
C2_new (fit M2)

Adjusted Rand Index
0.82 (ok >= 0.8)

Fréquence = pertinente

3 Stabilité / Maintenance

Conclusion & préconisation



préconisation

ARI	M_4_mths	M_3_mths	M_2_mths
ARI_T1	0.80	0.48	0.83
ARI_T2	0.70	0.77	0.82

préconisation

Questions ?

Merci !