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

Oumeima EL GHARBI

OpenClassrooms – Data Scientist

Soutenance : 08/10/2022

Plan

Introduction

- Problématique
- Présentation du jeu de données

I. Exploration

- Nettoyage
- Feature engineering
- Analyse exploratoire

II. Essais

- 1) CAH
- 2) DBSCAN
- **3) K-Means**
- **4) K-Means / Review Score**
- 5) RFM Score
- 6) Personae

III. Simulation

- Expérience 1 : 9 mois
- Expérience 2 : 3 mois

Conclusion

Introduction

<u>Problématique</u> :

« Olist souhaite que vous fournissiez à ses équipes d'e-commerce une segmentation des clients qu'elles pourront utiliser au quotidien pour leurs campagnes de communication.

Votre objectif est de comprendre les différents types d'utilisateurs grâce à leur comportement et à leurs données personnelles.

Vous devrez fournir à l'équipe marketing une description actionable de votre segmentation et de sa logique sous-jacente pour une utilisation optimale, ainsi qu'une proposition de contrat de maintenance basée sur une analyse de la stabilité des segments au cours du temps.

Votre mission est d'aider les équipes d'Olist à comprendre les différents types d'utilisateurs.»

<u>Implémentation</u> :

Cadre : apprentissage non supervisé

Problème de **clustering**

Modèles de clustering testés :

- Centroid-based Clustering : **K-Means**

- Hierarchical Clustering : **Agglomerative Clustering**

- Density-based Clustering : **DBSCAN**

Evaluation : méthode du « coude », silhouette score, Davies Bouldin score, ARI et matrice de confusion.

I) Exploration

Nettoyage

Exploration

Feature engineering

Exploration
1) Nettoyage

Statistiques générales des jeux de données bruts

```
Duplicated rows ?
The dataset called : dataset_customers has : 0 duplicated rows.
The dataset called : dataset_geolocation has : 261831 duplicated rows.
The dataset called : dataset_orders has : 0 duplicated rows.
The dataset called : dataset_order_items has : 0 duplicated rows.
The dataset called : dataset_order_payments has : 0 duplicated rows.
The dataset called : dataset_order_reviews has : 0 duplicated rows.
The dataset called : dataset_products has : 0 duplicated rows.
The dataset called : dataset_sellers has : 0 duplicated rows.
The dataset called : dataset_product_category_name_translation has : 0 duplicated rows.
```

```
Shape dataset
The dataset called : dataset_customers has a shape : (99441, 5)
The dataset called : dataset_geolocation has a shape : (1000163, 5)
The dataset called : dataset_orders has a shape : (99441, 8)
The dataset called : dataset_order_items has a shape : (112650, 7)
The dataset called : dataset_order_payments has a shape : (103886, 5)
The dataset called : dataset_order_reviews has a shape : (99224, 7)
The dataset called : dataset_products has a shape : (32951, 9)
The dataset called : dataset_sellers has a shape : (3095, 4)
The dataset called : dataset_product_category_name_translation has a shape : (71, 2)
```

Nous allons fusionner les datasets :

- Customers
- Orders
- Order_items

```
Missing values
The dataset called : dataset_customers has : 0.0 % of missing values.
The dataset called : dataset_geolocation has : 0.0 % of missing values.
The dataset called : dataset_orders has : 0.616948743476031 % of missing values.
The dataset called : dataset_order_items has : 0.0 % of missing values.
The dataset called : dataset_order_payments has : 0.0 % of missing values.
The dataset called : dataset_order_reviews has : 21.006294560071872 % of missing values.
The dataset called : dataset_products has : 0.8254681193287002 % of missing values.
The dataset called : dataset_sellers has : 0.0 % of missing values.
The dataset called : dataset_product_category_name_translation has : 0.0 % of missing values.
```

Exploration
2) Feature Engineering

Pour chaque client :

Récence : nombre de jours depuis le dernier achat
Fréquence : nombre total de commandes
Montant : montant total des achats

	Recency	Frequency	Monetary
customer_unique_id			
0000366f3b9a7992bf8c76cfdf3221e2	115	1	129.90
0000b849f77a49e4a4ce2b2a4ca5be3f	118	1	18.90
0000f46a3911fa3c0805444483337064	541	1	69.00
0000f6ccb0745a6a4b88665a16c9f078	325	1	25.99
0004aac84e0df4da2b147fca70cf8255	292	1	180.00

	count	mean	std	min	25%	50%	75%	max
Recency	95420.0	242.600377	153.160320	0.00	118.0	223.0	352.0	728.0
Frequency	95420.0	1.034018	0.211234	1.00	1.0	1.0	1.0	16.0
Monetary	95420.0	142.440198	217.656355	0.85	47.9	89.9	155.0	13440.0

Exploration
2) Feature Engineering

Pour chaque client :

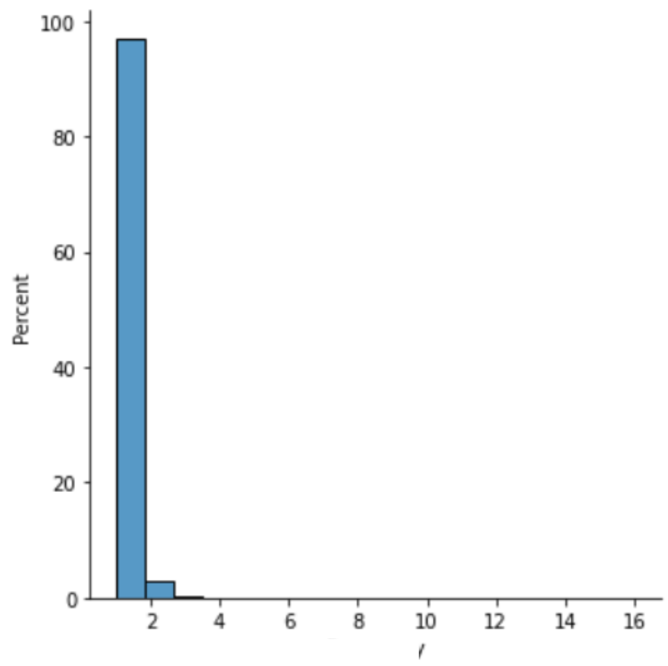
Récence : nombre de jours depuis le dernier achat
Fréquence : nombre total de commandes
Montant : montant total des achats
Review Score : note moyenne sur toutes les commandes

Ajout du dataset order_reviews

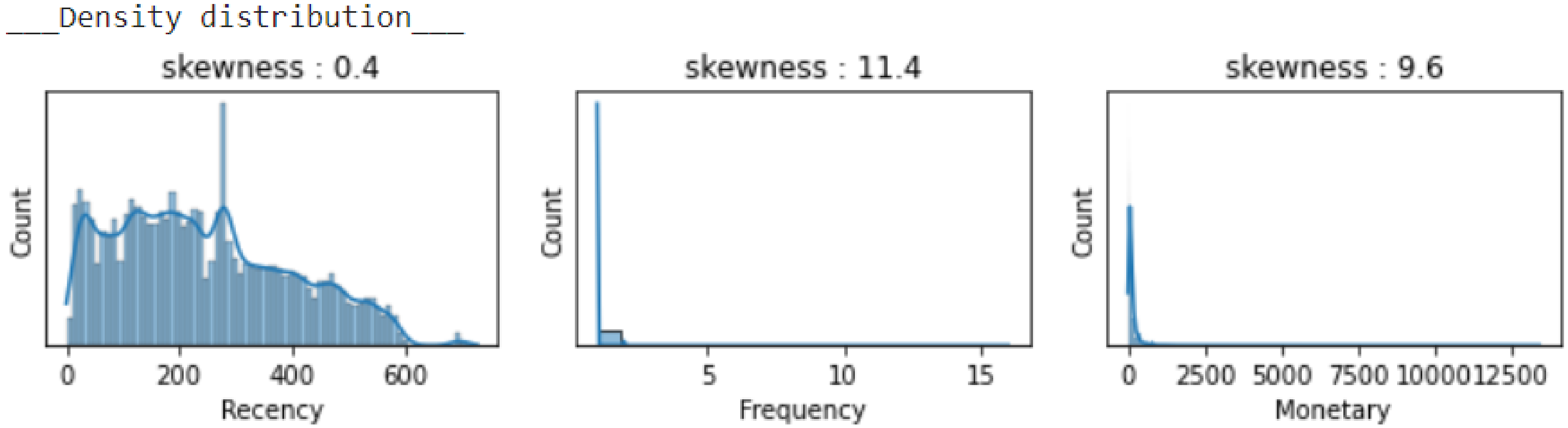
	Recency	Frequency	Monetary	Review Score
customer_unique_id				
0000366f3b9a7992bf8c76cfdf3221e2	115	1	129.90	5.0
0000b849f77a49e4a4ce2b2a4ca5be3f	118	1	18.90	4.0
0000f46a3911fa3c0805444483337064	541	1	69.00	3.0
0000f6ccb0745a6a4b88665a16c9f078	325	1	25.99	4.0
0004aac84e0df4da2b147fca70cf8255	292	1	180.00	5.0

	count	mean	std	min	25%	50%	75%	max
Recency	94721.0	242.442827	153.170660	0.00	118.0	223.0	352.00	728.0
Frequency	94721.0	1.033741	0.210527	1.00	1.0	1.0	1.00	16.0
Monetary	94721.0	142.811254	217.714921	0.85	47.9	89.9	155.96	13440.0
Review Score	94721.0	4.102202	1.326758	1.00	4.0	5.0	5.00	5.0

Exploration
3) Exploration

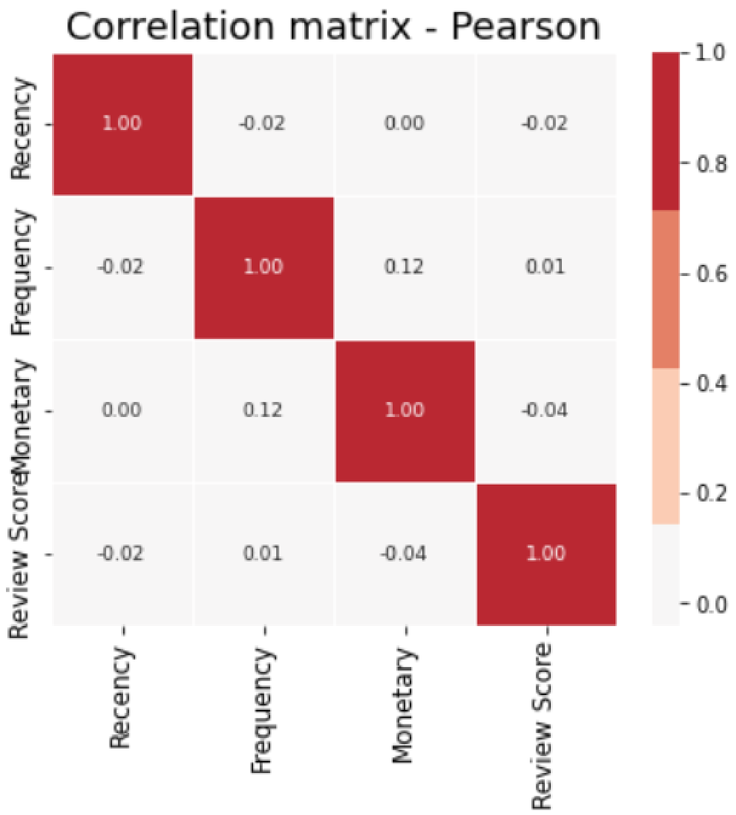


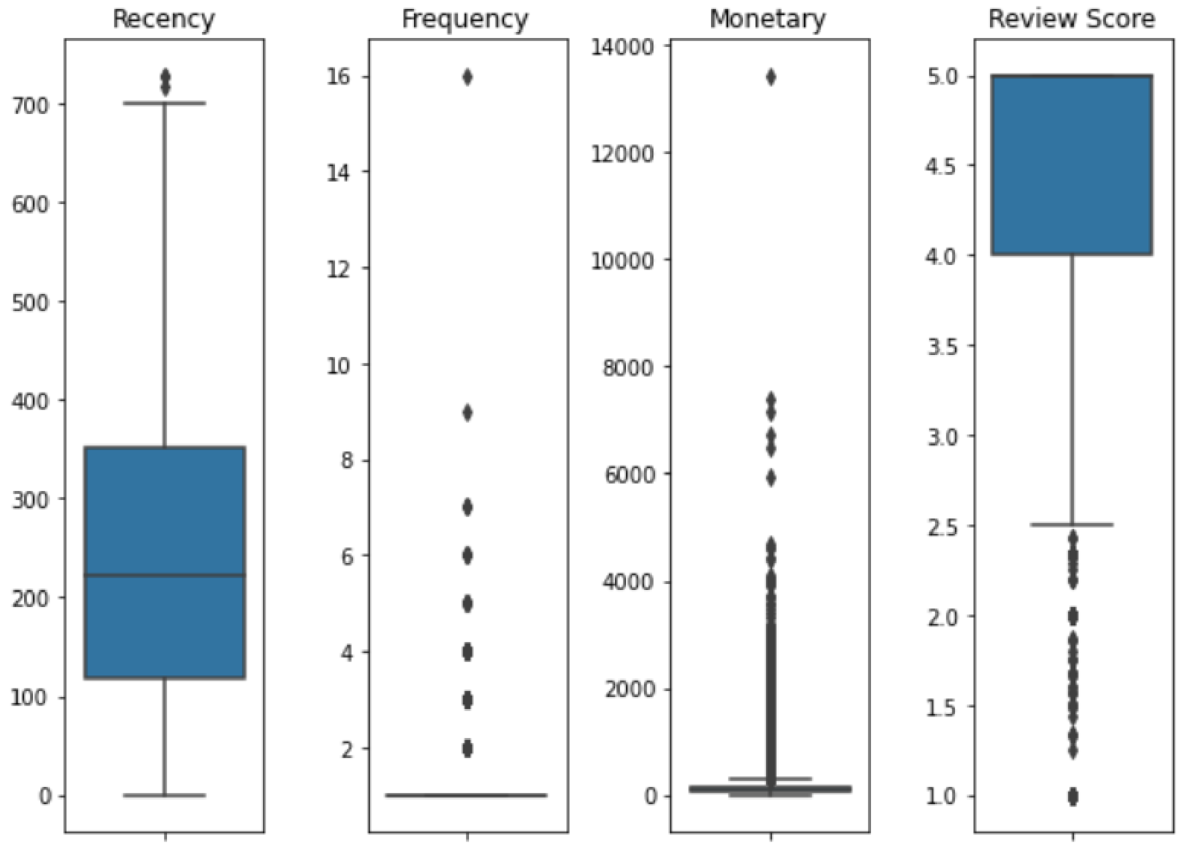
Récence : bonne distribution
Fréquence : seulement 3% des clients ont passé plus d’une commande
Montant : la plupart des commandes ont un montant faible



Exploration

3) Exploration

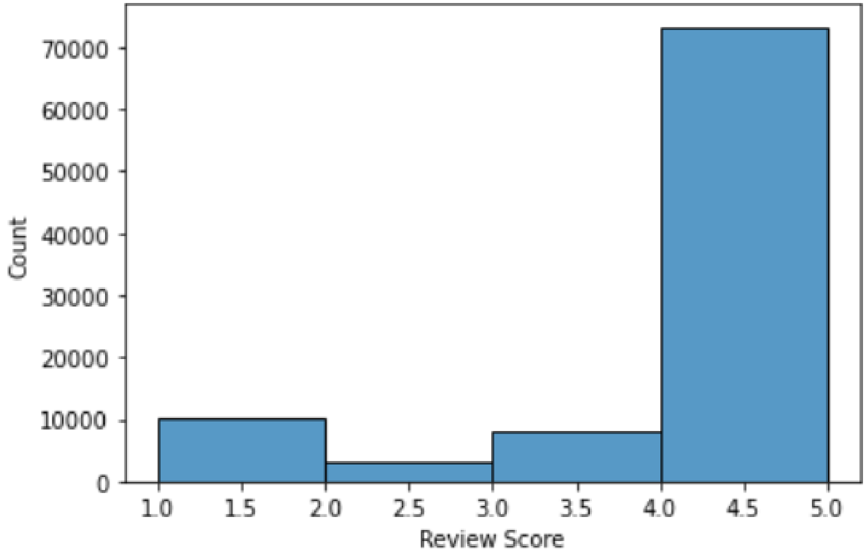


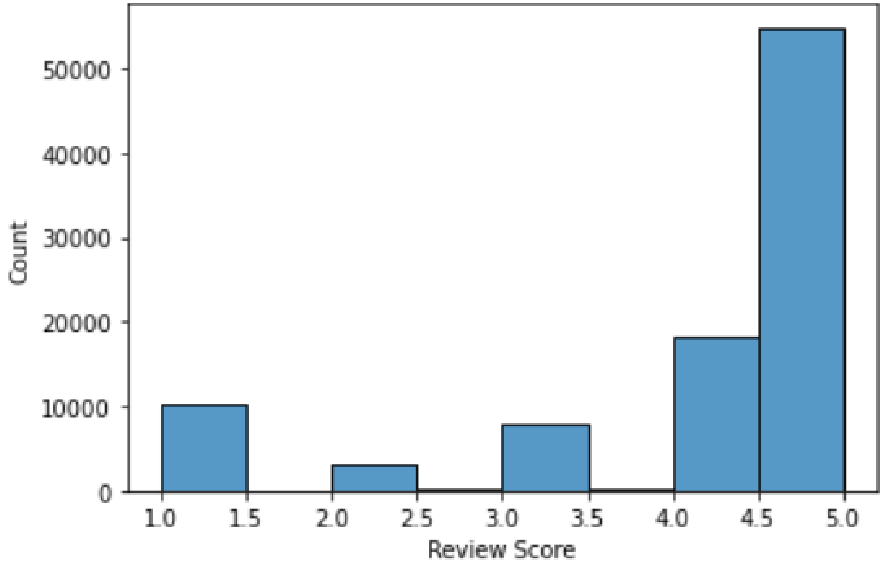


- Pas de corrélation entre les features.
- Récence : distribution assez homogène entre nouveaux clients et anciens clients : cela est dû au fait que les clients ne commandent qu'une fois (dans 97% des cas).
- Review Score : la plupart des clients ont été satisfaits de leur commande

Exploration
3) Exploration

Distribution du Review Score entre par tranche de 1 ou de 0.5





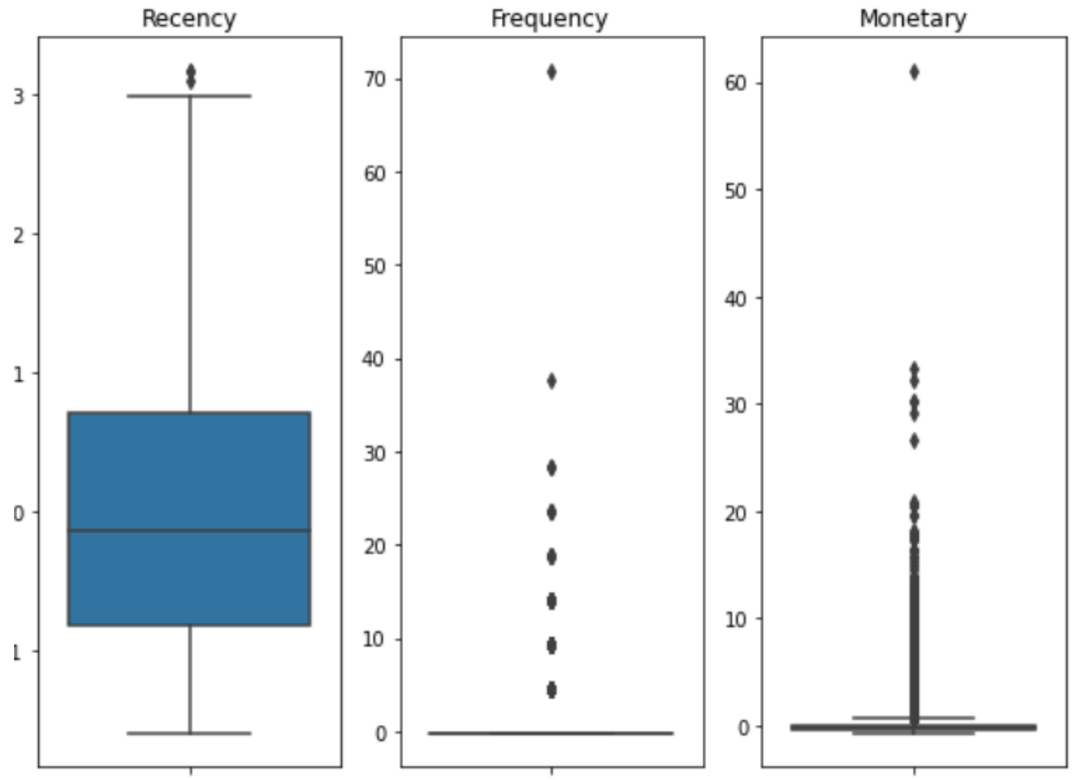
II) Essais

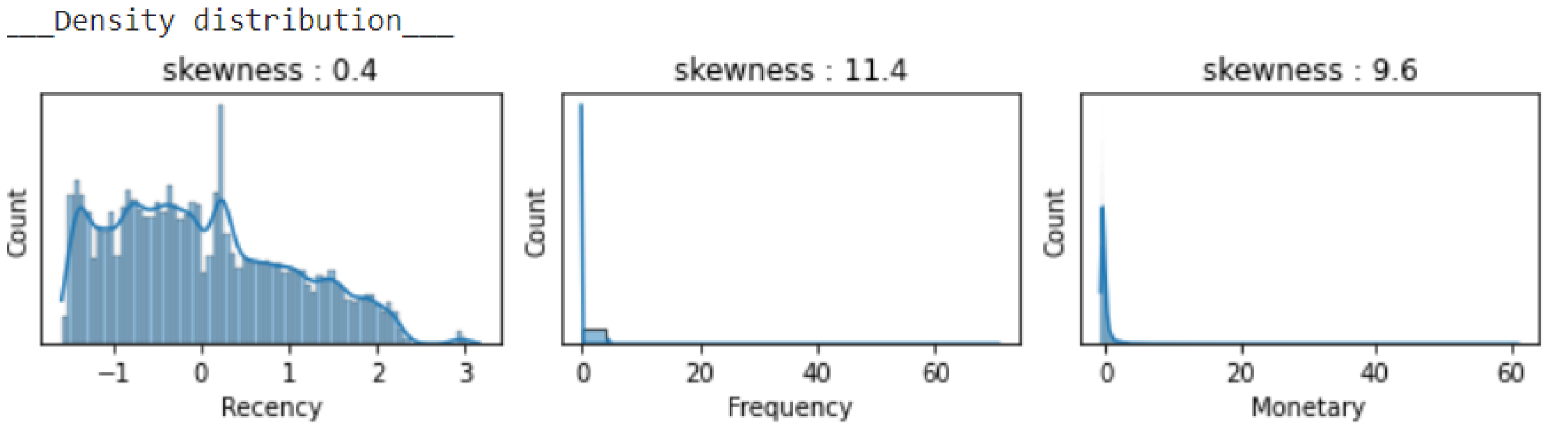
- 1) CAH
- 2) DBSCAN
- 3) K-Means
- 4) K-Means / Review Score
- 5) RFM Score
- 6) Personae

Essais Standardisation

Scale ou Standard Scaler : pour la simulation on utilisera Standard Scaler qui permet de fit/transform.

	Recency	Frequency	Monetary
0	-0.833121	-0.161045	-0.057615
1	-0.813533	-0.161045	-0.567596
2	1.948293	-0.161045	-0.337415
3	0.537999	-0.161045	-0.535021
4	0.322537	-0.161045	0.172566



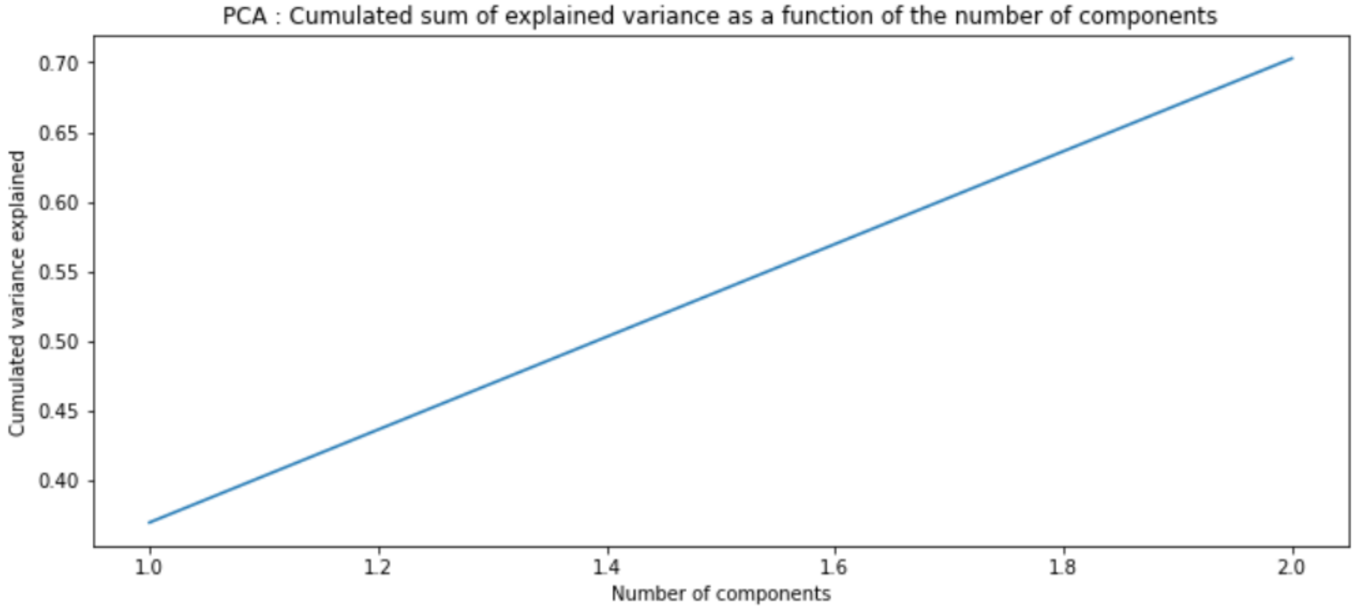


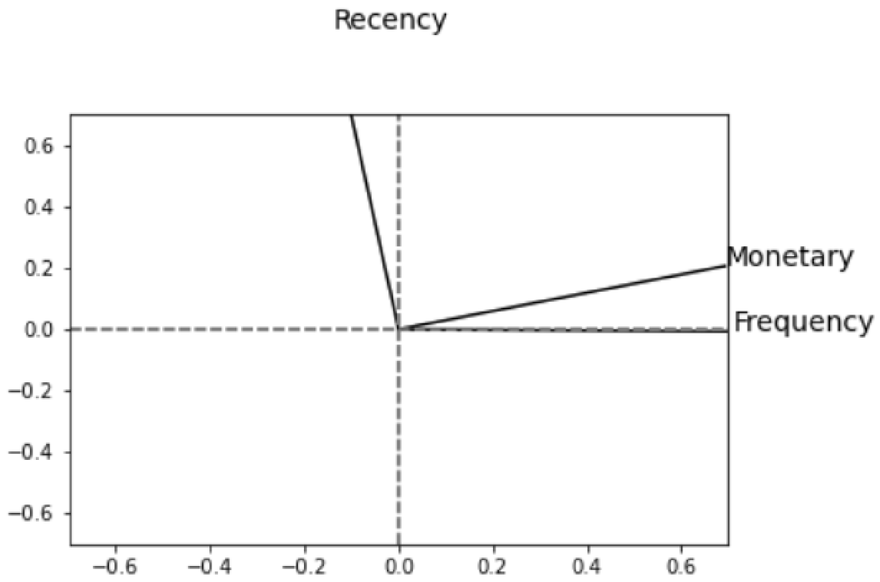
Essais PCA

Les deux premières composantes de la PCA permettent d’expliquer 70% de la variance.

La première composante explique F + M (la fréquence et le montant).

La deuxième composante explique R (la récence).

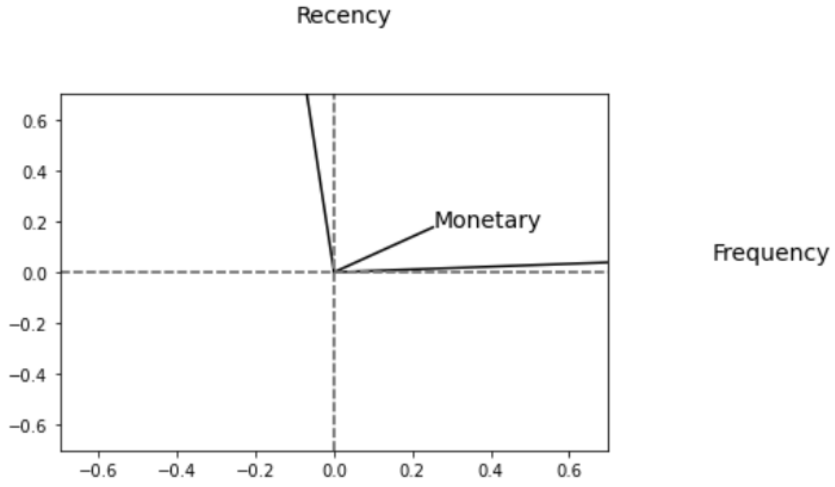


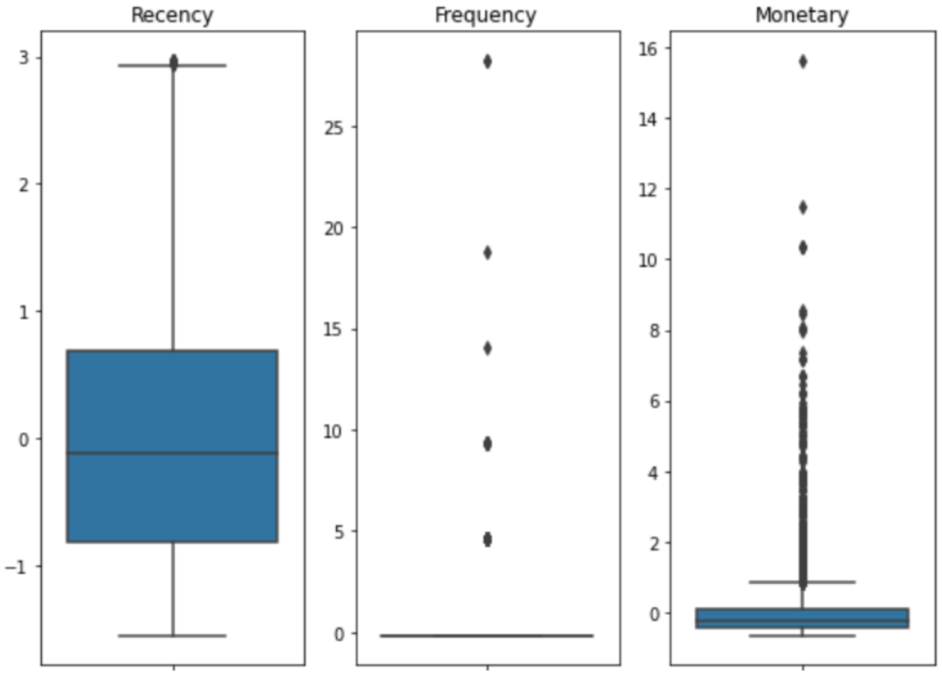


Essais Sample

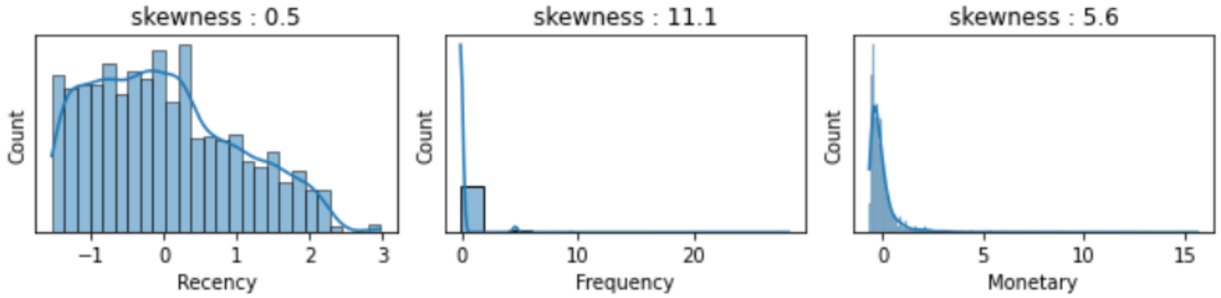
Fléau de la dimensionalité : CAH, DBSCAN : utilisation d'un échantillon de taille 1000.

[0.41106477 0.73319995]





___Density distribution___



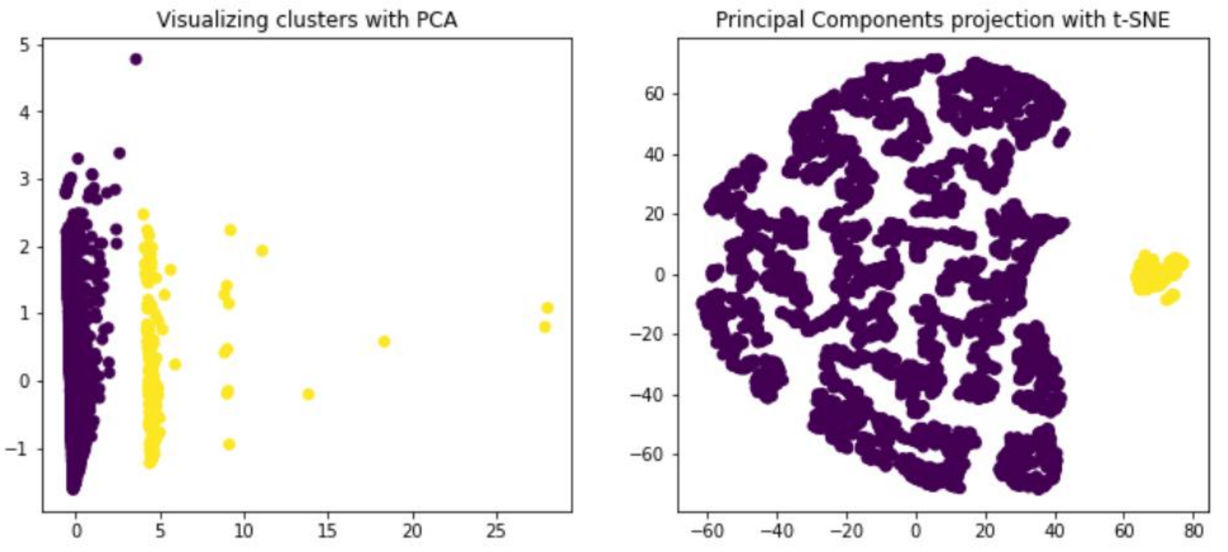
Essais
1) CAH : Clustering Hiérarchique : sample

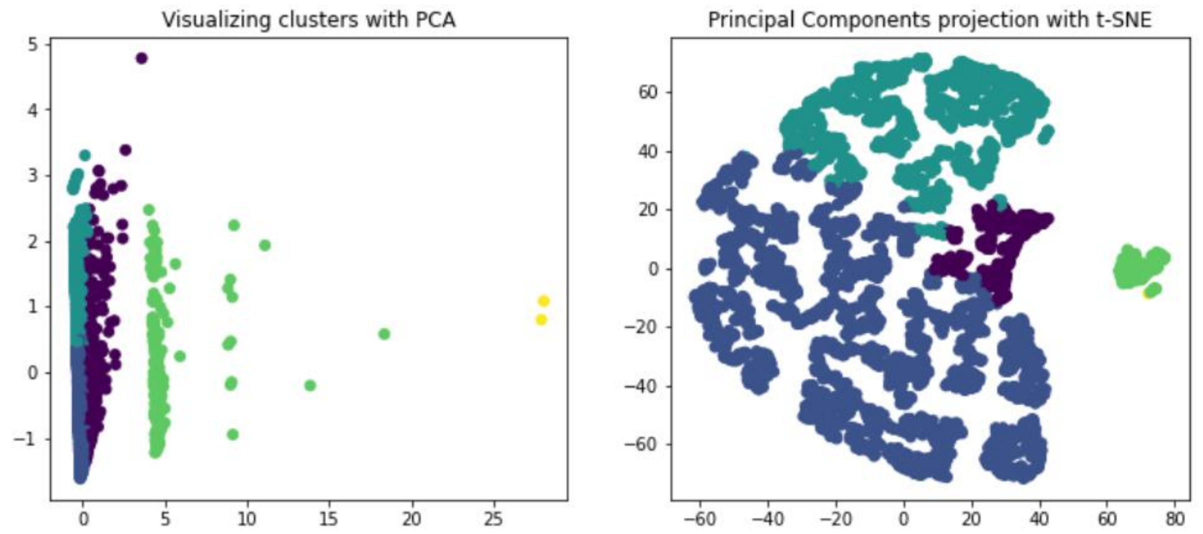
CAH sans paramètres : 2 clusters.

CAH avec n_clusters = 4.

- Clustering similaire à celui du K-Means (cf partie 3).
- CAH n’est pas applicable à notre dataset : dataset trop grand => CAH met trop de temps.

Conclusion : nous n’utiliserons pas CAH



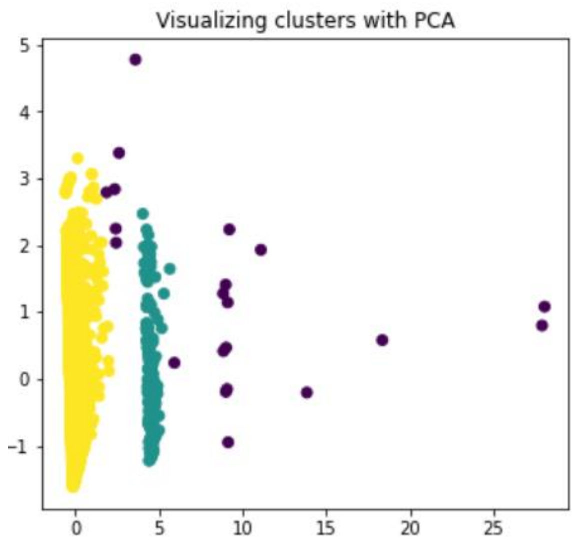


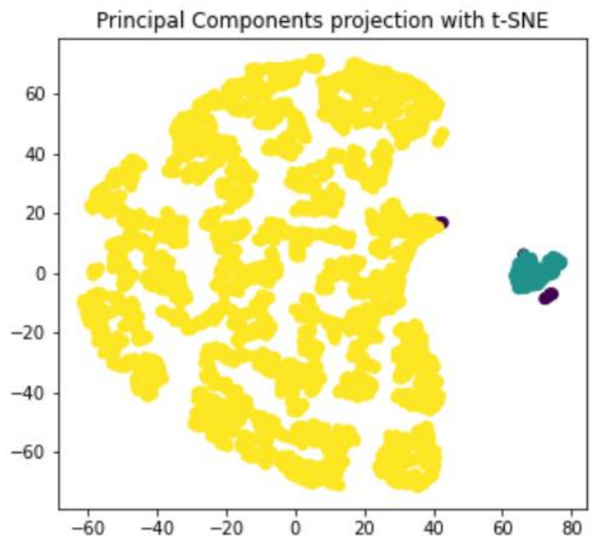
Essais
2) DBSCAN : sample

```
Hyperparameter epsilon =  0.5
Hyperparameter epsilon =  1.0
Hyperparameter epsilon =  1.5
Hyperparameter epsilon =  2.0
Hyperparameter epsilon =  2.5
Hyperparameter epsilon =  3.0
```

DBSCAN : clustering par densité.

- La structure de nos données ne correspondent pas au DBSCAN => données ne sont pas connectées par densité.
- DBSCAN : problème de dimension => met trop de temps à répondre
- Conclusion : nous n’utiliserons pas DBSCAN



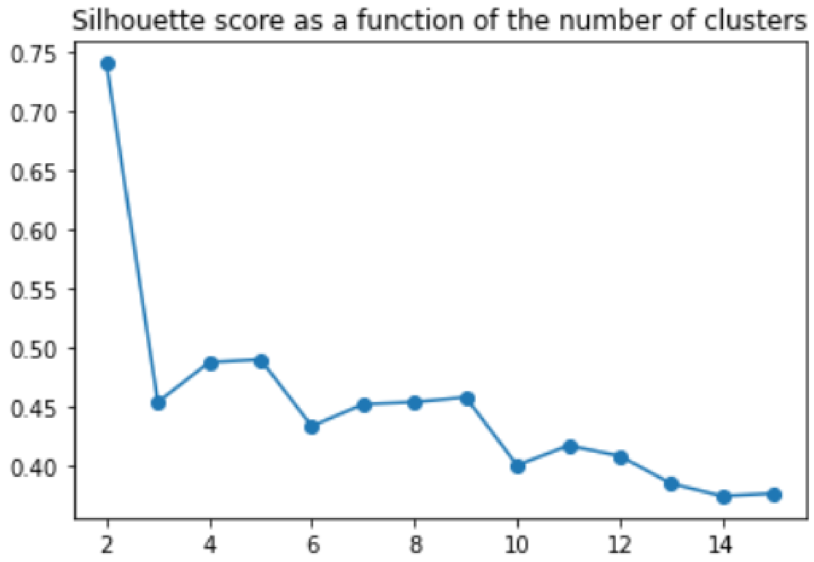


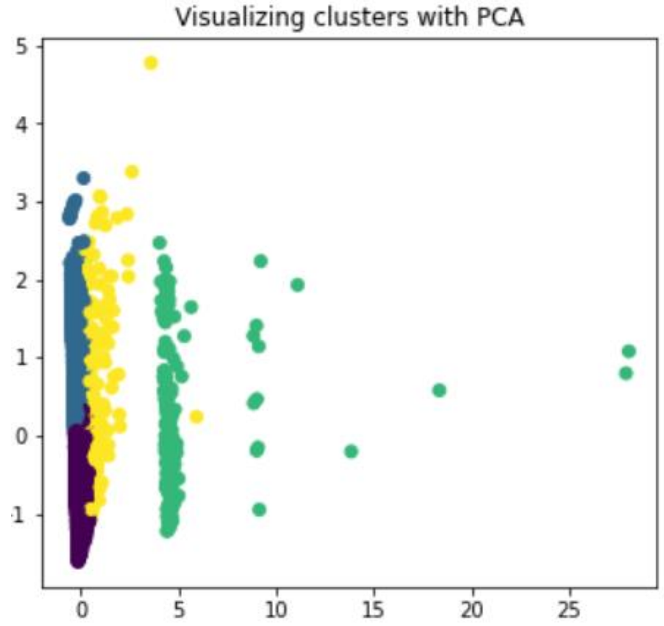
Essais
3) K-Means : sample

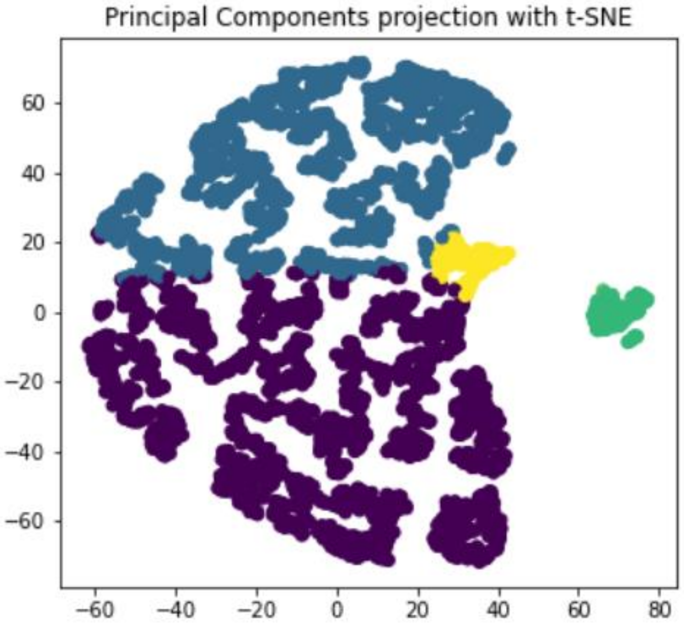
- Utilisation du dataset échantillon pour comparer avec CAH / DBSCAN
- **Silhouette** optimal pour Silhouette proche de 1 : k = 2

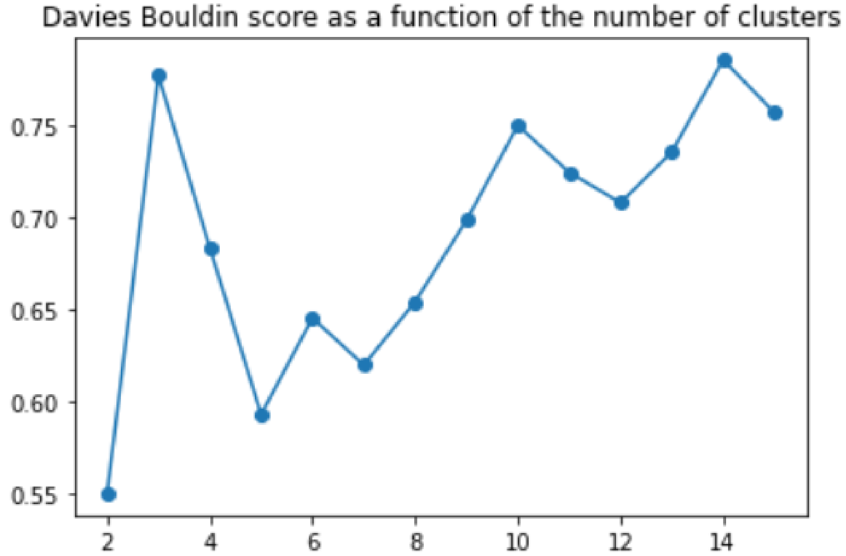
  Besoin métier : clusters entre k = 3 à 5
- **Davies Bouldin** optimal pour DB proche de 0 : k = 2

  Besoin métier : clusters k = 5



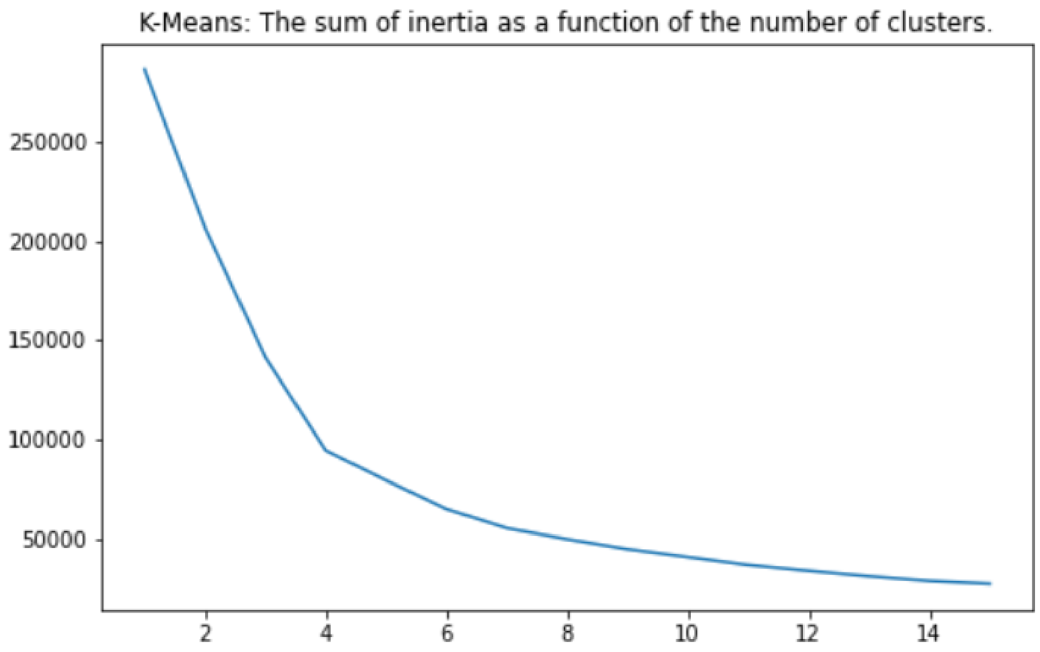


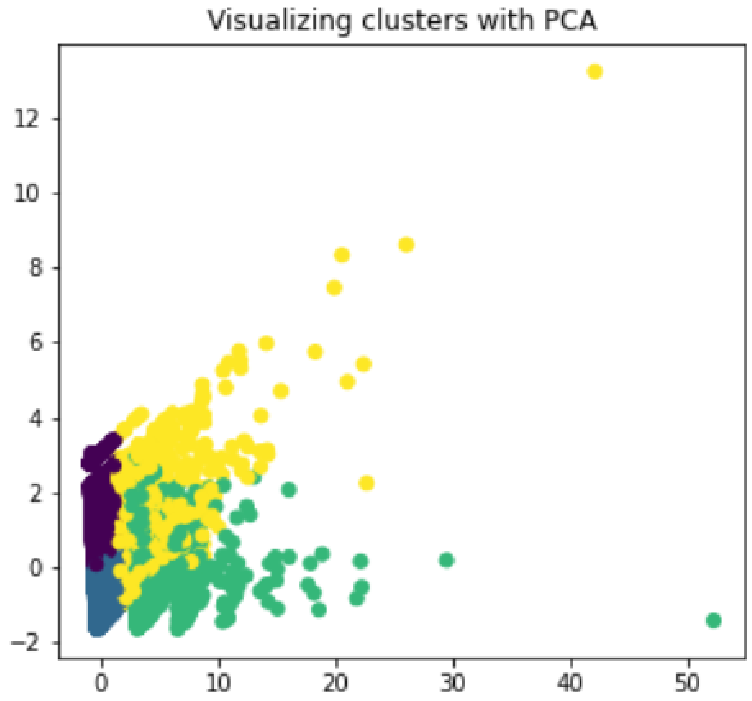




Essais
3) K-Means : complete dataset

- Calcul du coefficient de Silhouette trop long => choix de K avec la méthode du coude / elbow
- K-Means minimise la somme des inerties (variance intra-cluster) : k = 4 ou 5
- Le reste de l’analyse a été réalisée pour K = 4 (besoin métier : pas trop de clusters car plus complexe à interpréter / au moins 4 clusters pour différencier les clients)



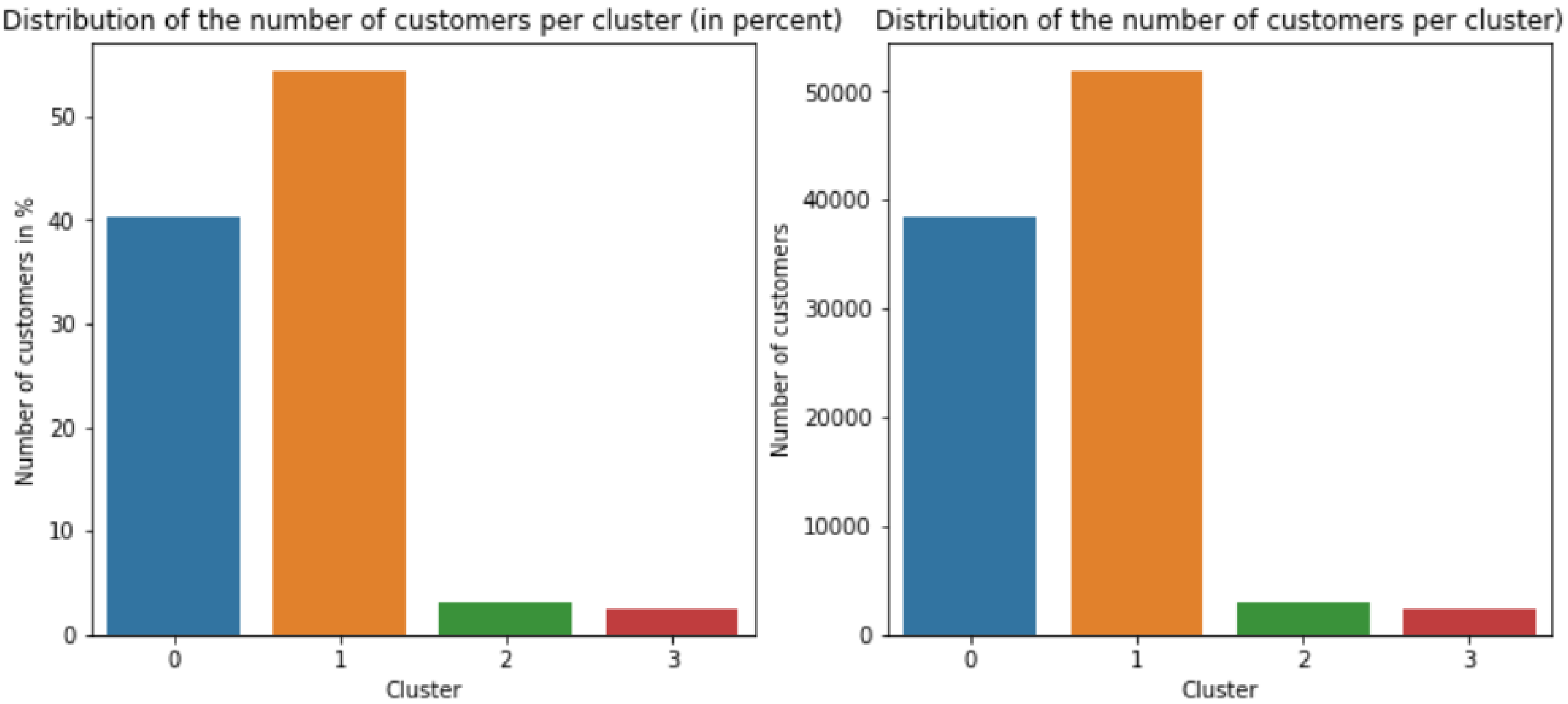


Essais
3) K-Means : analysis

Pour chaque cluster : calcul de la moyenne par feature

=> Les clusters 0 et 1 contiennent plus de clients que les deux autres clusters.

Customer_cluster	Nb customers	Avg Recency	Avg Frequency	Avg Monetary
0	38378	392.691438	1.000000	114.481193
1	51886	132.518810	1.000000	113.596573
2	2883	225.184530	2.114811	243.049823
3	2273	243.351518	1.014078	1145.314571



Essais
3) K-Means : analysis

Diagramme à bâton qui présente la moyenne de chaque feature par cluster.

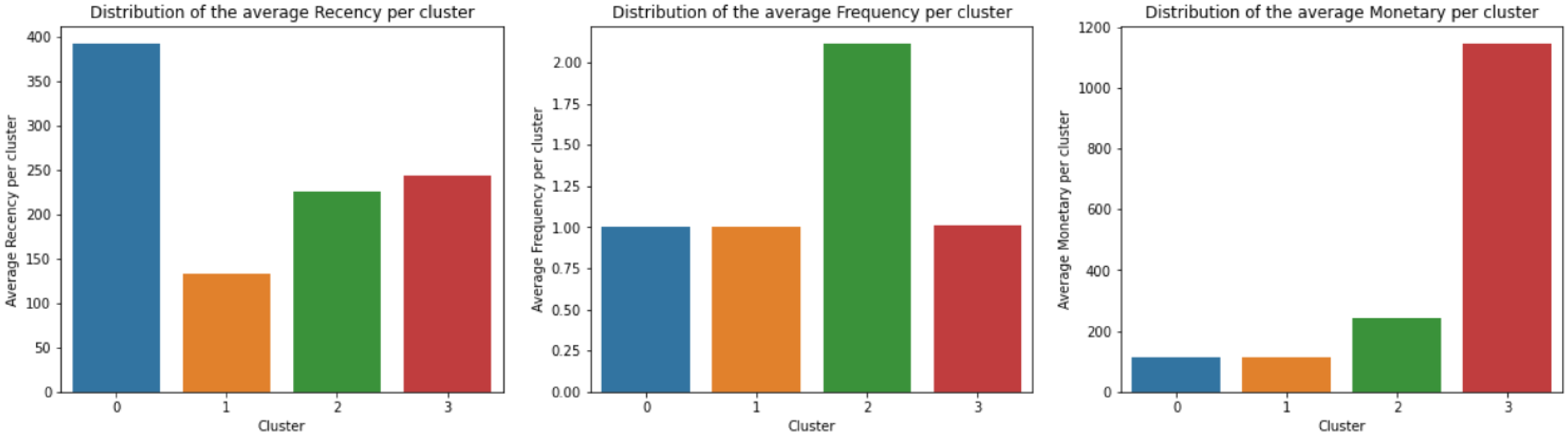
Cluster 0 : clients perdus

Cluster 1 : nouveaux clients

Cluster 2 : clients fidèles

Cluster 3 : clients qui dépensent beaucoup

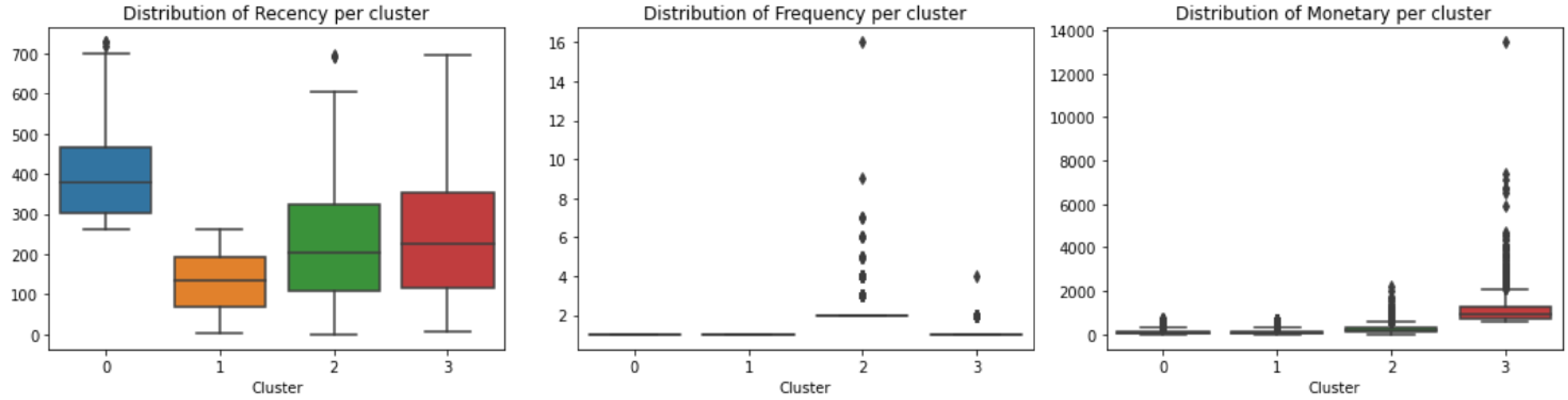
Customer_cluster	Nb customers	Avg Recency	Avg Frequency	Avg Monetary
0	38378	392.691438	1.000000	114.481193
1	51886	132.518810	1.000000	113.596573
2	2883	225.184530	2.114811	243.049823
3	2273	243.351518	1.014078	1145.314571



Essais
3) K-Means : analysis

	Nb customers	Avg Recency	Avg Frequency	Avg Monetary
Customer_cluster				
0	38378	392.691438	1.000000	114.481193
1	51886	132.518810	1.000000	113.596573
2	2883	225.184530	2.114811	243.049823
3	2273	243.351518	1.014078	1145.314571

Pour chaque feature, analyse de la distribution par cluster

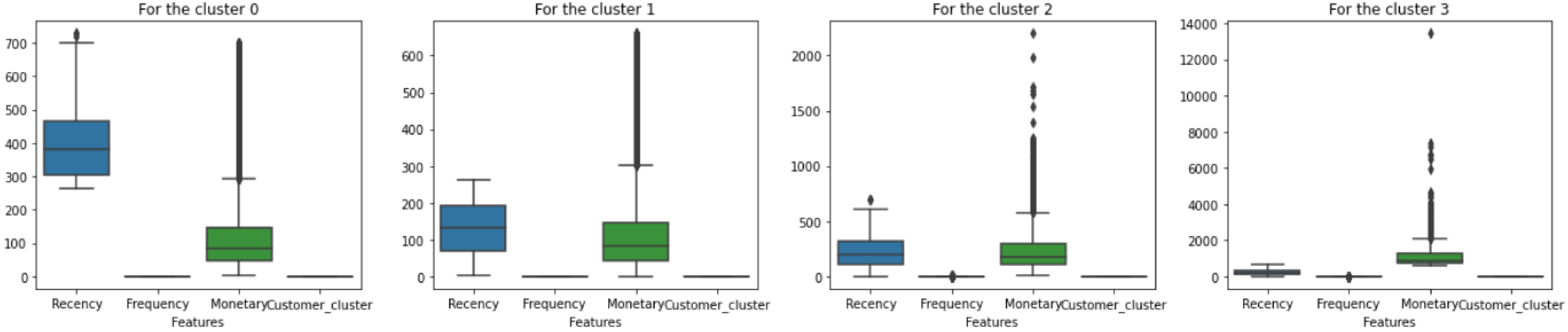


Essais
3) K-Means : analysis

	Nb customers	Avg Recency	Avg Frequency	Avg Monetary
Customer_cluster				
0	38378	392.691438	1.000000	114.481193
1	51886	132.518810	1.000000	113.596573
2	2883	225.184530	2.114811	243.049823
3	2273	243.351518	1.014078	1145.314571

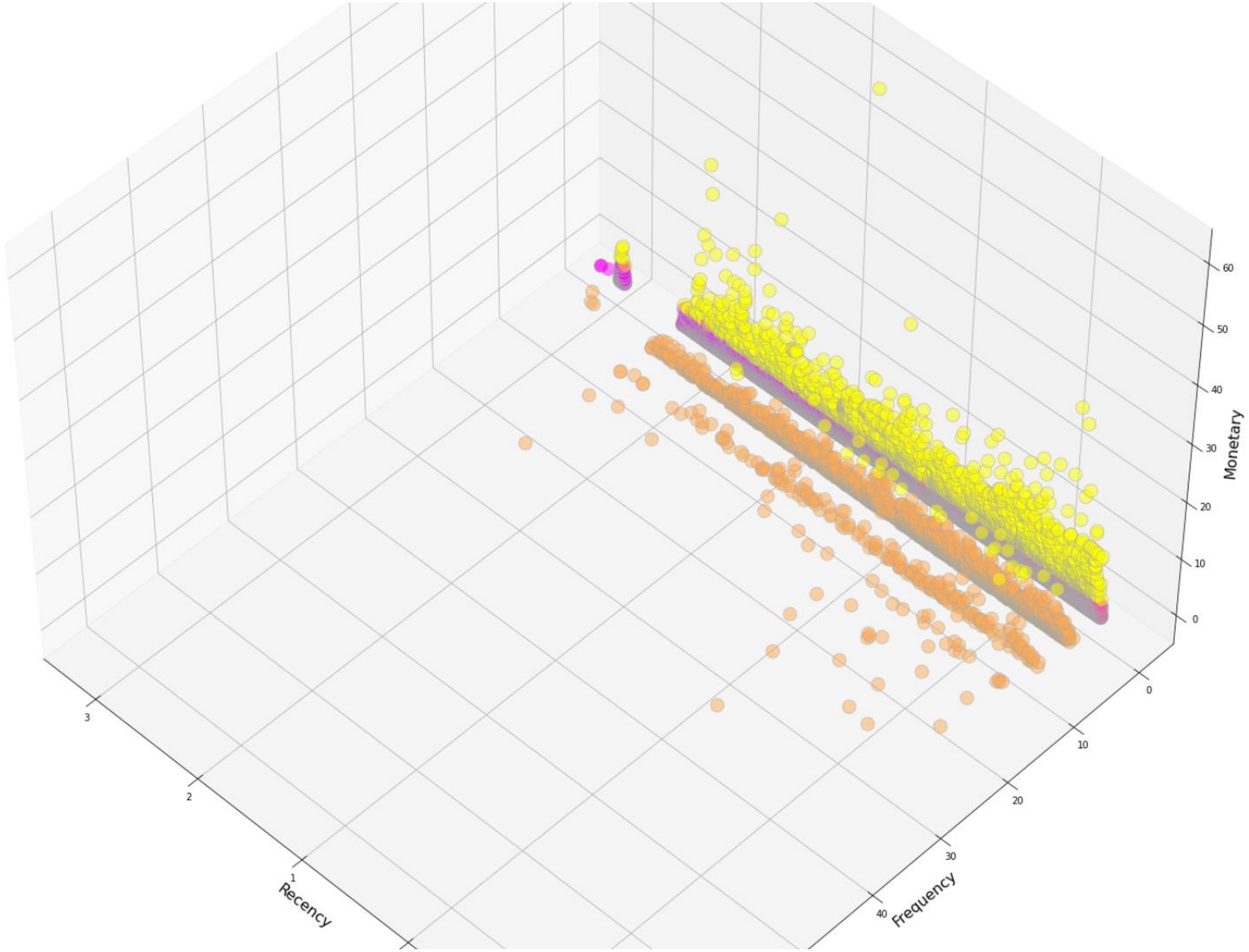
Pour chaque cluster, distribution de ses features

Différente échelle pour chaque cluster



Essais
3) K-Means : analysis

Représentation 3D : pas évident de bien visualiser les différents clients / clusters



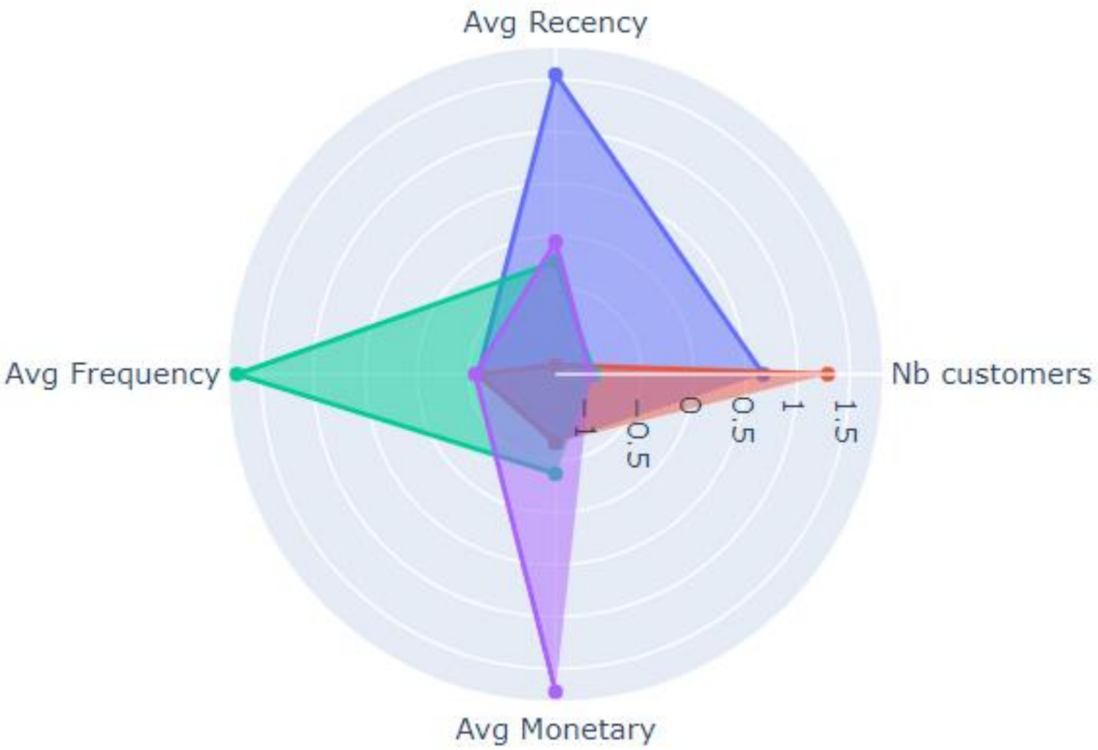
Essais
3) K-Means : analysis

Cluster 0 : clients perdus

Cluster 1 : nouveaux clients

Cluster 2 : clients fidèles

Cluster 3 : clients qui dépensent beaucoup



This is the analysis of our segmentation :

- Cluster 0 (blue) : **lost customers**, customers that didn't buy recently nor ordered more than once and didn't make expensive purchases.
- Cluster 1 (red) : **new customers**, customers who have made a purchase recently.
- Cluster 2 (green) : **loyal customers**, customers that ordered more than once even though they didn't make expensive orders.
- Cluster 3 (purple) : **royal customers**, customers that made expensive purchases.

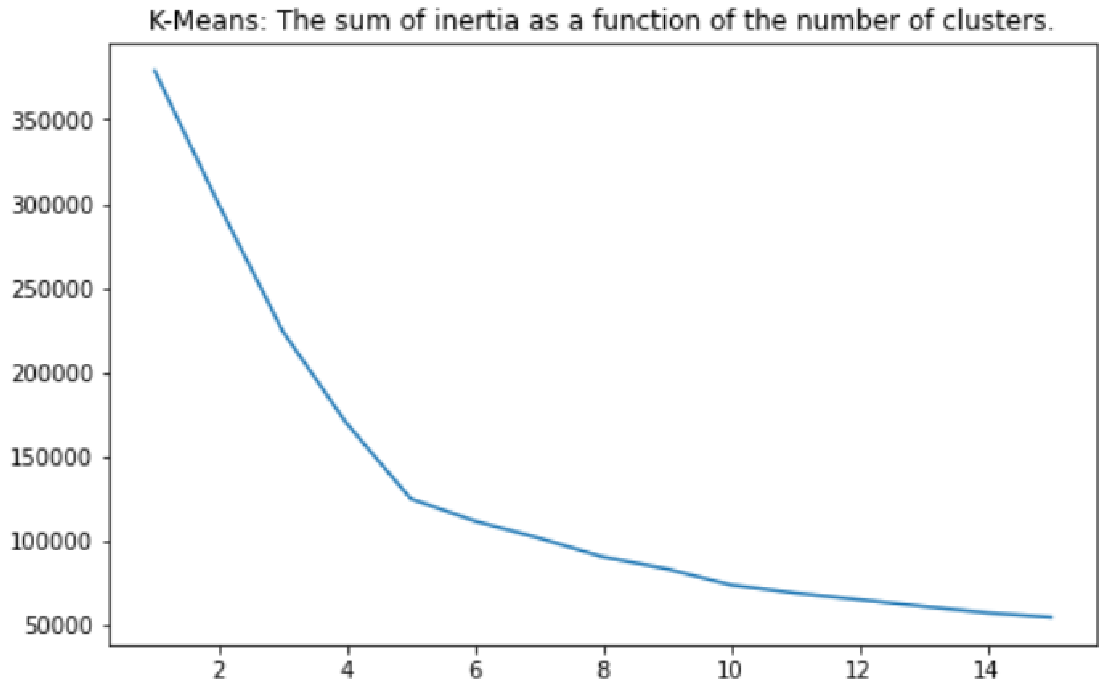
Essais
4) K-Means / Review Score

Ajout du Review Score moyen par client.

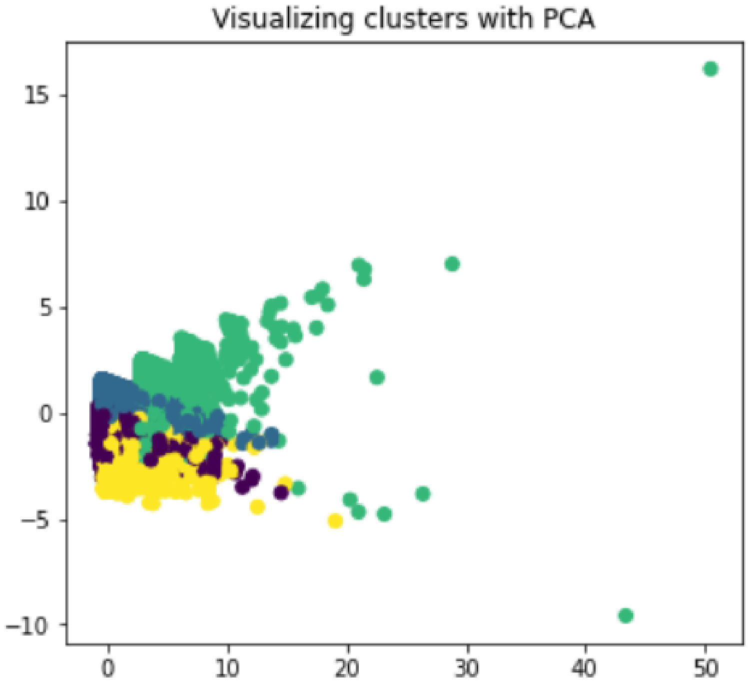
Coude pour K = 5

Choix de K = 4 pour faciliter l'interprétation des clusters

	Recency	Frequency	Monetary	Review Score
0	-0.832036	-0.160271	-0.059304	0.676689
1	-0.812450	-0.160271	-0.569148	-0.077032
2	1.949190	-0.160271	-0.339029	-0.830752
3	0.538991	-0.160271	-0.536582	-0.077032
4	0.323544	-0.160271	0.170815	0.676689



Essais
4) K-Means / Review Score

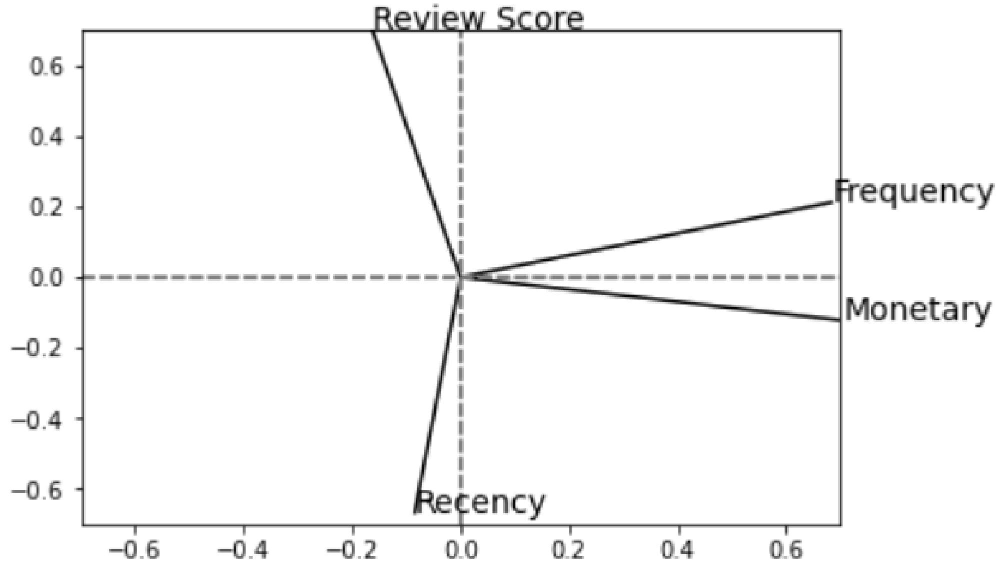


PCA :

- La première composante explique F + M (fréquence et montant)

- La deuxième composante explique la Récence et le Review Score

```
[0.28102799 0.53782495]
The PCA with 2 components explains 53.800000000000004 % of the variance.
```



Essais
4) K-Means / Review Score : analysis

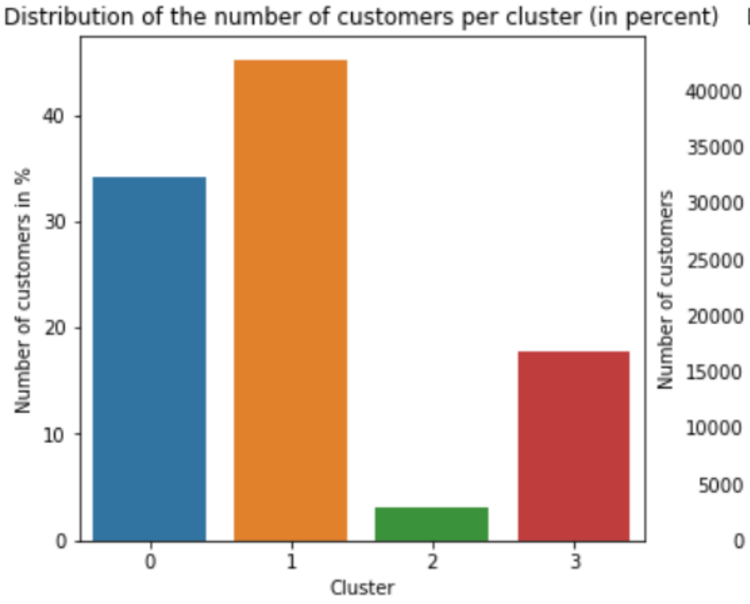
Cluster 0 : clients perdus

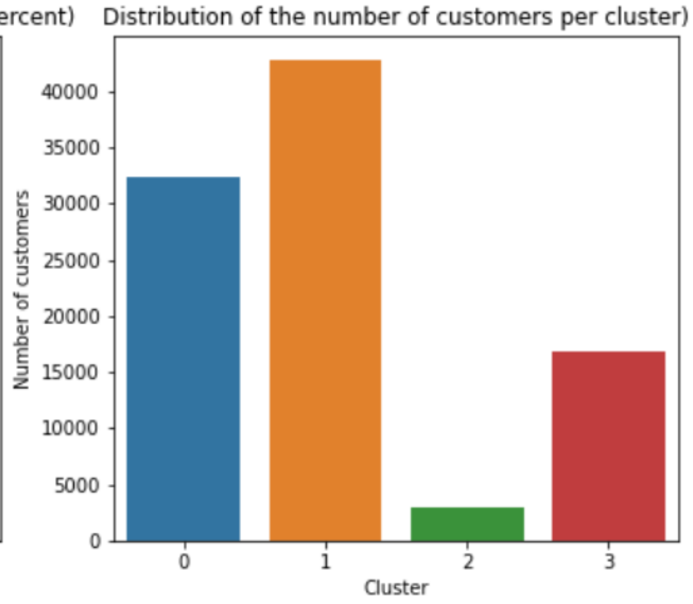
Cluster 1 : nouveaux clients

Cluster 2 : moins de clients => clients fidèles et qui dépensent beaucoup

Cluster 3 : clients mécontents

Customer_cluster	Nb customers	Avg Recency	Avg Frequency	Avg Monetary	Avg Review Score
0	32299	397.363107	1.000000	136.156248	4.632930
1	42776	126.224518	1.000000	131.160897	4.674245
2	2874	225.671190	2.112039	292.661889	4.145321
3	16772	243.384033	1.000000	159.662849	1.613791





Essais
4) K-Means / Review Score : analysis

Diagramme à bâton qui présente la moyenne de chaque feature par cluster.

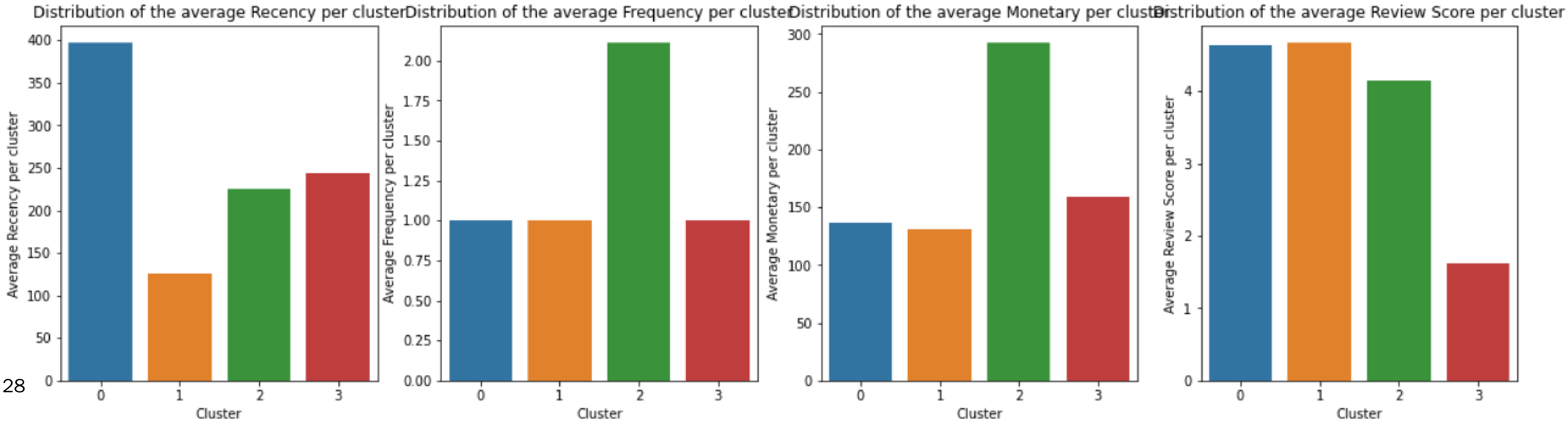
Cluster 0 : clients perdus

Cluster 1 : nouveaux clients

Cluster 2 : clients fidèles et dépensent beaucoup

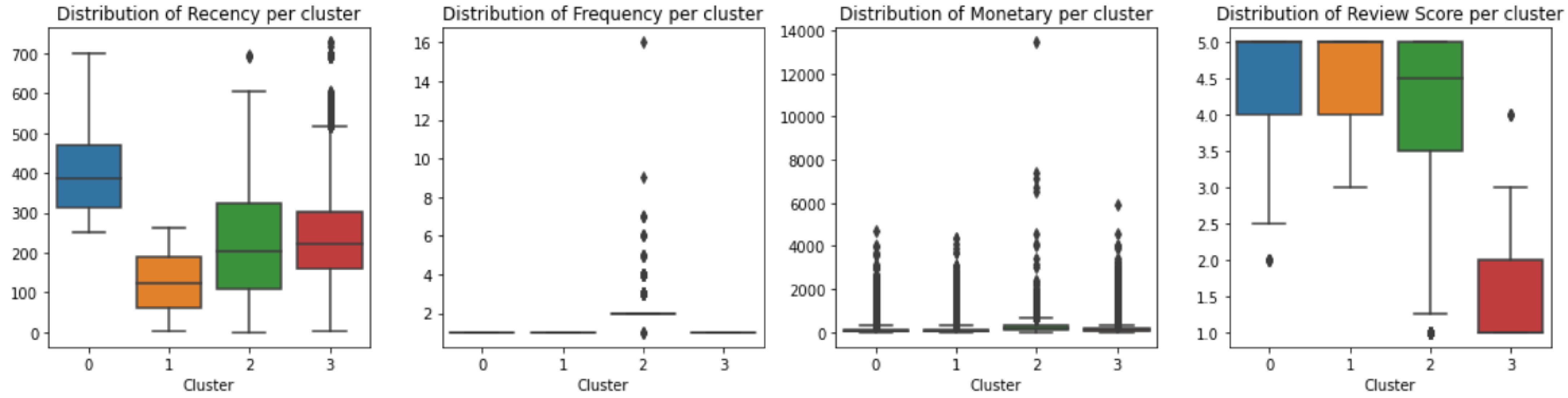
Cluster 3 : clients mécontents

Customer_cluster	Nb customers	Avg Recency	Avg Frequency	Avg Monetary	Avg Review Score
0	32299	397.363107	1.000000	136.156248	4.632930
1	42776	126.224518	1.000000	131.160897	4.674245
2	2874	225.671190	2.112039	292.661889	4.145321
3	16772	243.384033	1.000000	159.662849	1.613791



Essais
4) K-Means / Review Score : analysis

Customer_cluster	Nb customers	Avg Recency	Avg Frequency	Avg Monetary	Avg Review Score
0	32299	397.363107	1.000000	136.156248	4.632930
1	42776	126.224518	1.000000	131.160897	4.674245
2	2874	225.671190	2.112039	292.661889	4.145321
3	16772	243.384033	1.000000	159.662849	1.613791



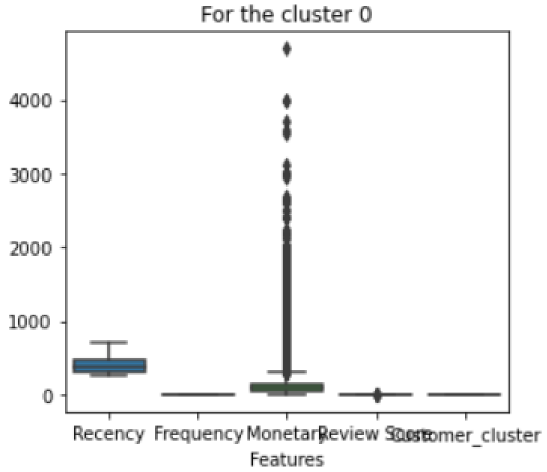
Essais
4) K-Means / Review Score : analysis

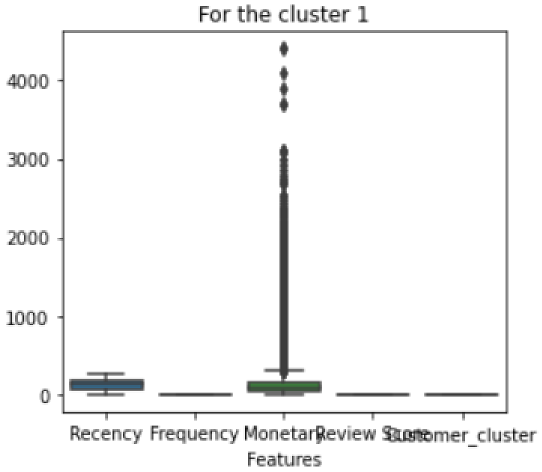
Echelle différente pour chaque cluster

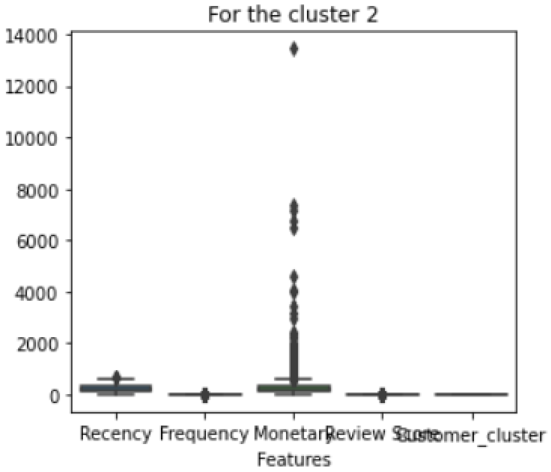
Distributions assez similaires entre les clusters

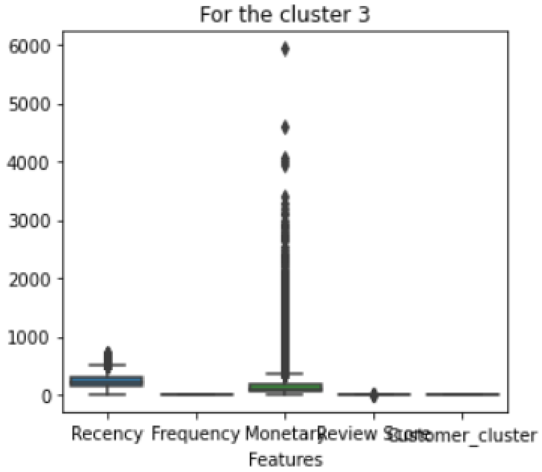
Cluster 2 : grands ordres de grandeurs => top customers

Customer_cluster	Nb customers	Avg Recency	Avg Frequency	Avg Monetary	Avg Review Score
0	32299	397.363107	1.000000	136.156248	4.632930
1	42776	126.224518	1.000000	131.160897	4.674245
2	2874	225.671190	2.112039	292.661889	4.145321
3	16772	243.384033	1.000000	159.662849	1.613791









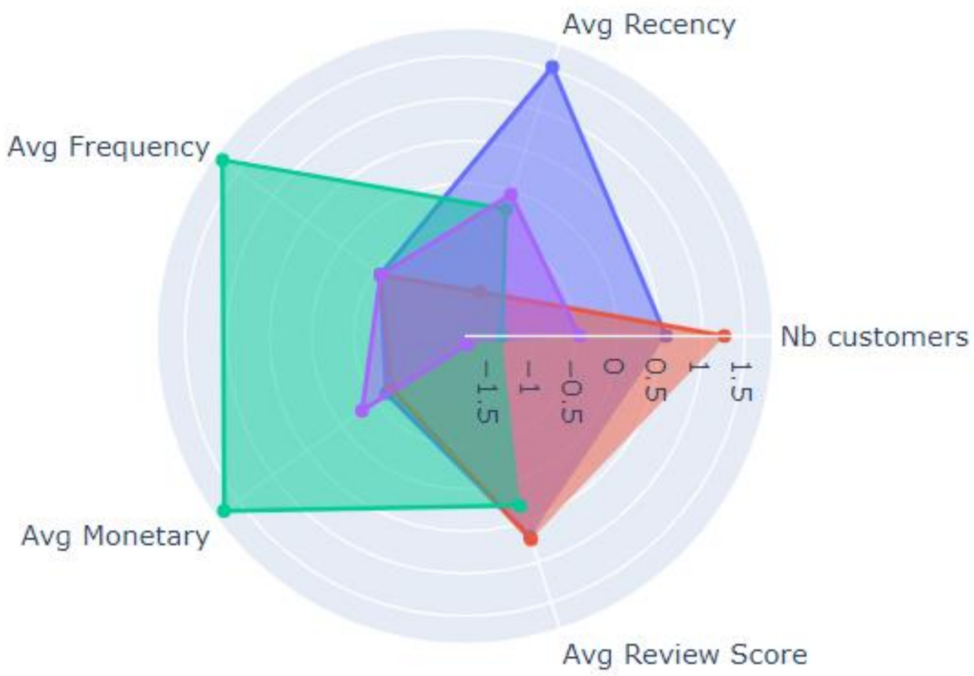
Essais
4) K-Means / Review Score : analysis

Cluster 0 : clients perdus

Cluster 1 : nouveaux clients

Cluster 2 : clients « royaux » = fidèles et dépensent beaucoup

Cluster 3 : clients mécontents et assez récents



By adding the review score we can see that we have an interesting cluster : the cluster 3 corresponds to unsatisfied clients !

Adding the review score made our segments more explicit.

This is the analysis of our segmentation :

- Cluster 0 (blue) : **lost customers**, customers that didn't buy recently nor ordered more than once and didn't make expensive purchases.
- Cluster 1 (red) : **new customers**, customers who have made a purchase recently.
- Cluster 2 (green) : **royal customers**, customers that ordered more than once and made expensive orders. They are mostly satisfied by the service offered.
- Cluster 3 (purple) : **new and unhappy customers**, customers gave a low review score.

Essais
5) RFM Score

Clustering par feature

4 clusters pour chaque feature => 64 combinaisons de clusters

	Recency							
	count	mean	std	min	25%	50%	75%	max
Recency_cluster								
0	16723.0	490.362973	58.790423	406.0	444.0	481.0	532.0	728.0
1	25022.0	319.903285	44.061560	255.0	281.0	313.0	357.0	405.0
2	27735.0	188.320606	35.263820	128.0	159.0	188.0	219.0	254.0
3	25940.0	66.341403	36.097045	0.0	33.0	66.0	100.0	127.0

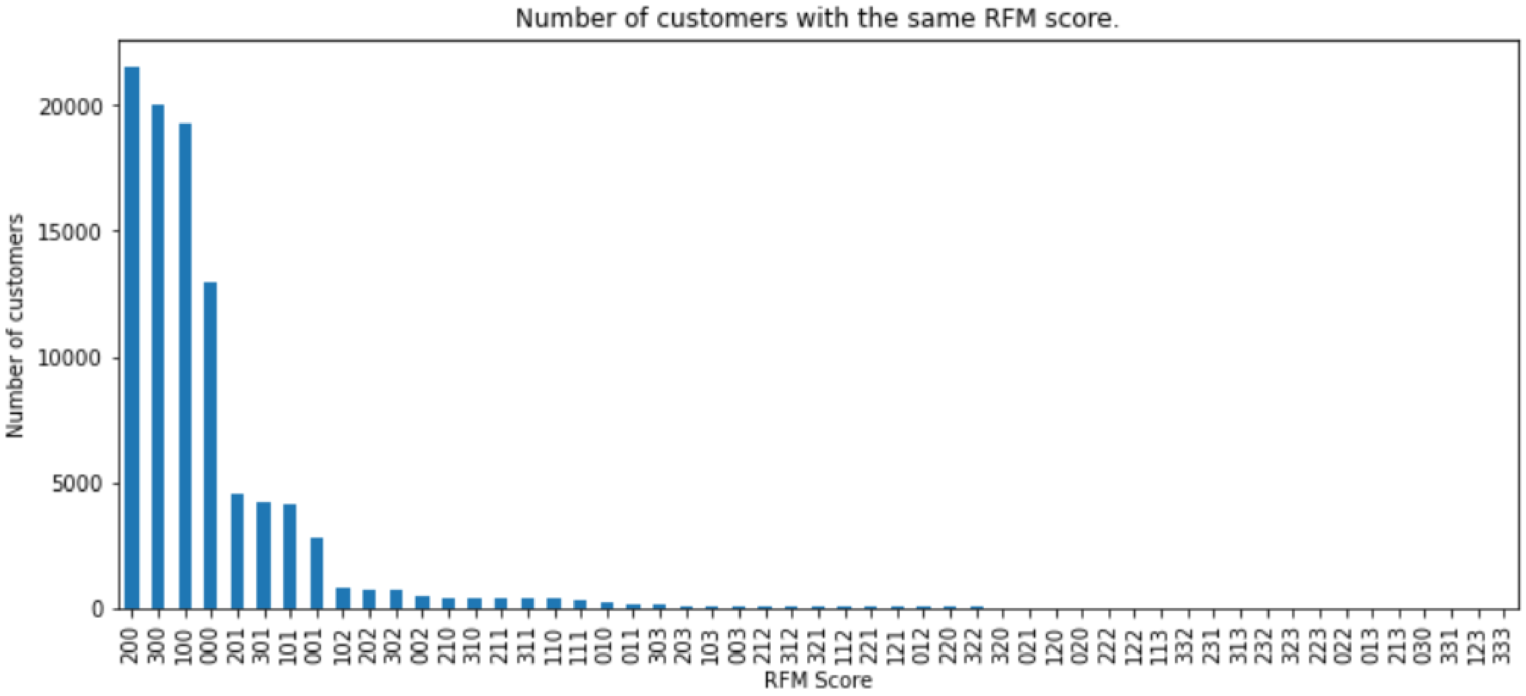
	Frequency							
	count	mean	std	min	25%	50%	75%	max
Frequency_cluster								
0	92507.0	1.000000	0.000000	1.0	1.0	1.0	1.0	1.0
1	2673.0	2.000000	0.000000	2.0	2.0	2.0	2.0	2.0
2	221.0	3.131222	0.338409	3.0	3.0	3.0	3.0	4.0
3	19.0	6.368421	2.564946	5.0	5.0	6.0	6.5	16.0

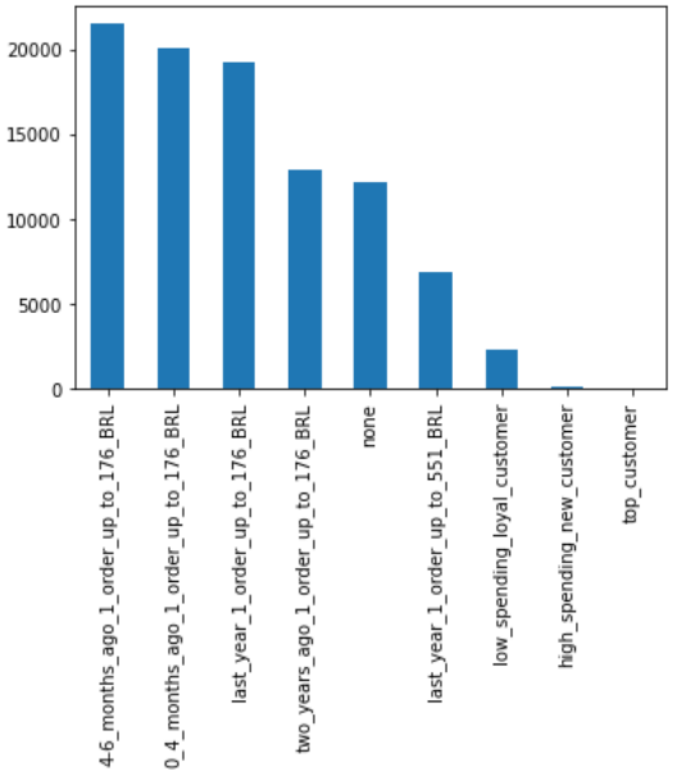
	Monetary							
	count	mean	std	min	25%	50%	75%	max
Monetary_cluster								
0	75253.0	75.527705	43.316084	0.85	39.89	69.0	109.0	175.66
1	16892.0	275.783227	91.733765	175.79	199.90	249.0	329.0	550.90
2	2885.0	825.691875	241.741323	550.99	629.00	750.0	960.0	1520.88
3	390.0	2223.861590	1007.485998	1534.90	1699.99	1980.0	2300.0	13440.00

Essais
5) RFM Score

Combinaison de cluster => RFM Score

Interprétation assez difficile





customer_unique_id	Recency	Frequency	Monetary	Recency_cluster	Frequency_cluster	Monetary_cluster	RFM_score
0000366f3b9a7992bf8c76cfdf3221e2	115	1	129.90	3	0	0	300
0000b849f77a49e4a4ce2b2a4ca5be3f	118	1	18.90	3	0	0	300
0000f46a3911fa3c0805444483337064	541	1	69.00	0	0	0	000
0000f6ccb0745a6a4b88665a16c9f078	325	1	25.99	1	0	0	100
0004aac84e0df4da2b147fca70cf8255	292	1	180.00	1	0	1	101

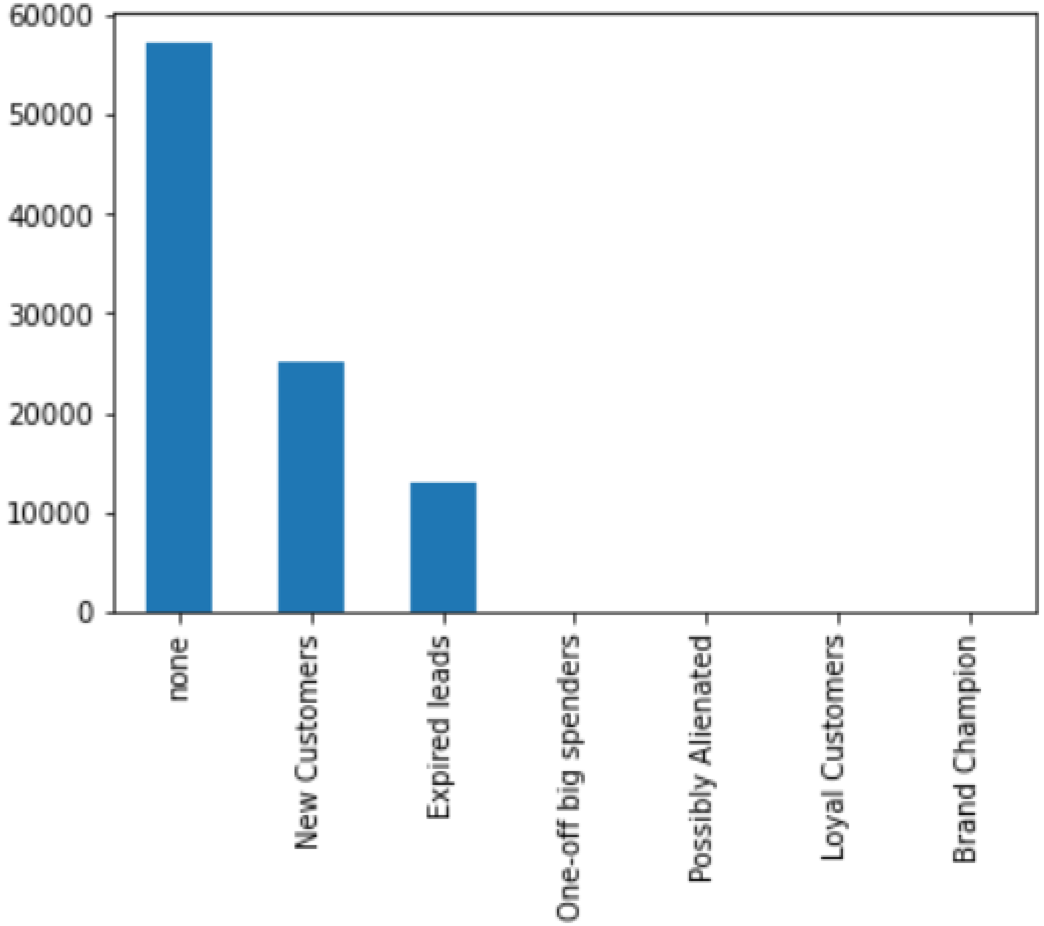
Essais
6) Personae

Clustering par feature : 4 clusters pour chaque feature

Combinaison de cluster => attribution de profils clients (personae)

Interprétation assez difficile : exemple avec 6 Personae

Mais n’explique pas la majorité des clients



III) Simulation

- Expérience 1 : 9 mois / 15 jours
- Expérience 2 : 3 mois / 7 jours

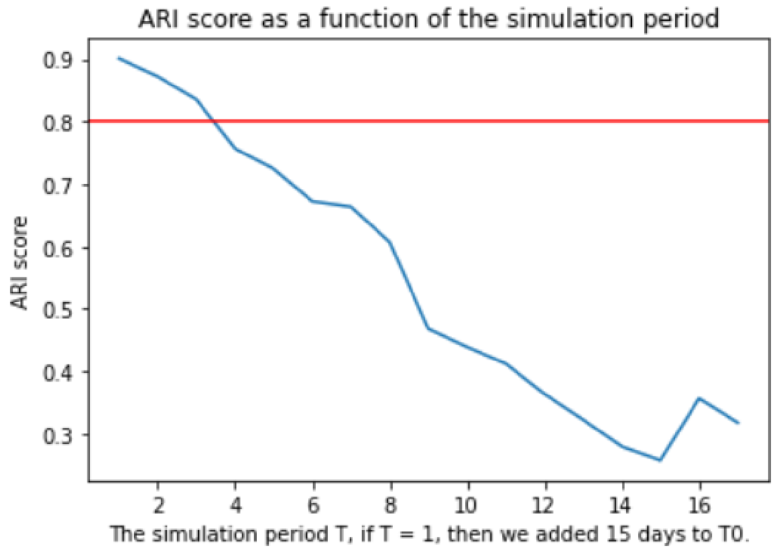
- Simulation sur le dataset RFM (3 features) avec un clustering K-Means (K = 4)

Simulation
1) Expérience 1 : 9 mois / 15 jours

```
Step : 1 Maximum order purchase date : 2017-12-22 09:06:57
Verification of the filter : 2017-12-22 09:06:20
This dataset has 42415 unique clients


Step : 2 Maximum order purchase date : 2018-01-06 09:06:57
Verification of the filter : 2018-01-06 09:03:41
This dataset has 44483 unique clients


Step : 3 Maximum order purchase date : 2018-01-21 09:06:57
Verification of the filter : 2018-01-21 09:03:17
This dataset has 48074 unique clients


Step : 4 Maximum order purchase date : 2018-02-05 09:06:57
Verification of the filter : 2018-02-05 09:06:12
This dataset has 51346 unique clients


Step : 5 Maximum order purchase date : 2018-02-20 09:06:57
Verification of the filter : 2018-02-20 09:05:49
This dataset has 54669 unique clients


Step : 6 Maximum order purchase date : 2018-03-07 09:06:57
Verification of the filter : 2018-03-07 09:02:41
This dataset has 58414 unique clients
```

Baisse significative du ARI score à partir de T = 4

Soit 2 mois



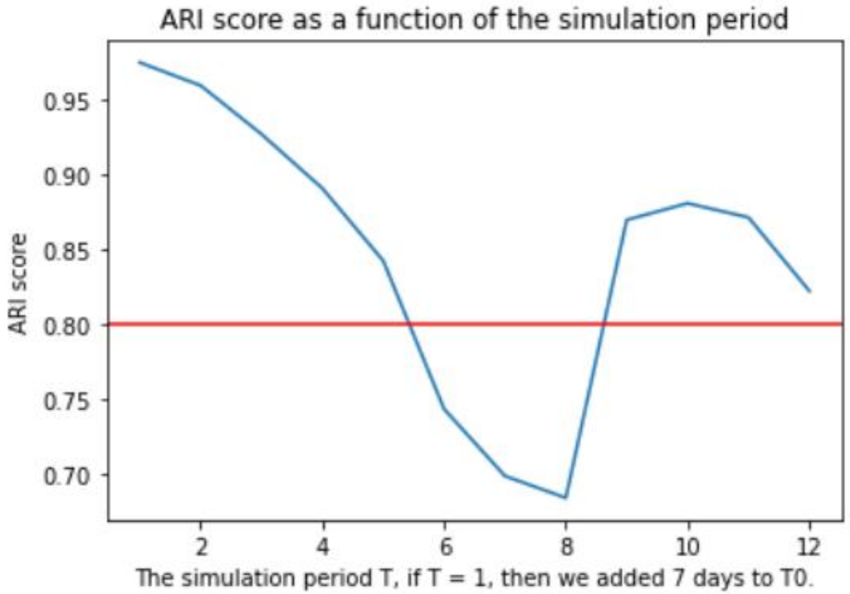
Simulation
2) Expérience 2 : 3 mois / 7 jours

```
Step : 1 Maximum order purchase date : 2018-06-11 09:06:57
Verification of the filter : 2018-06-11 08:58:16
This dataset has 79092 unique clients

Step : 2 Maximum order purchase date : 2018-06-18 09:06:57
Verification of the filter : 2018-06-18 08:59:45
This dataset has 80547 unique clients

Step : 3 Maximum order purchase date : 2018-06-25 09:06:57
Verification of the filter : 2018-06-25 09:05:04
This dataset has 81949 unique clients

Step : 4 Maximum order purchase date : 2018-07-02 09:06:57
Verification of the filter : 2018-07-02 09:01:31
This dataset has 83297 unique clients

Step : 5 Maximum order purchase date : 2018-07-09 09:06:57
Verification of the filter : 2018-07-09 09:06:51
This dataset has 84460 unique clients
```

```
For T = 5
(85421, 3)
We make a new clustering using that fits the new dataset.
KMeans(n_clusters=4, random_state=0)
We predict a clustering using the clustering at T0 for the new dataset.
ARI for T = 5
```

	T	ARI
0	1	0.974220
1	2	0.958896
2	3	0.926436
3	4	0.890263
4	5	0.841963

	pred_0	pred_1	pred_2	pred_3
true_0	3707	0	28699	5
true_1	0	2572	0	0
true_2	48076	0	0	0
true_3	51	0	0	2311



Baisse significative du ARI score à partir de T = 5

➔ Contrat de maintenance pour faire la segmentation toutes les 4/5 semaines.

Simulation
Contrat de Maintenance

Deux simulations faites pour déterminer la période de renouvellement de la segmentation : toutes les 4 à 5 semaines soit une maintenance de la segmentation tous les mois.

We did two simulations.

For the first simulation, we predicted the initial clustering in December 2017 soit 9 months before the end of the dataset. We simulated the clustering prediction every 15 days.

After two months, we saw the ARI score decreased and get under 0.8.

So, for the second simulation, we predicted the initial clustering in June 2017 so three months before the end of the dataset. We ran the simulation every week.

- NB : In fact, we tried a simulation two months (nb_periods = 8) prior to the end date of the dataset and a simulation three months before (nb_periods = 12). It was difficult to conclude with a simulation period of two months, so we used three months.

We saw that the ARI score decreases significantly at the 5th week.

Thus, we will recommend to our client that the maintenance should be done every month (every 4 or 5 weeks).

Conclusion

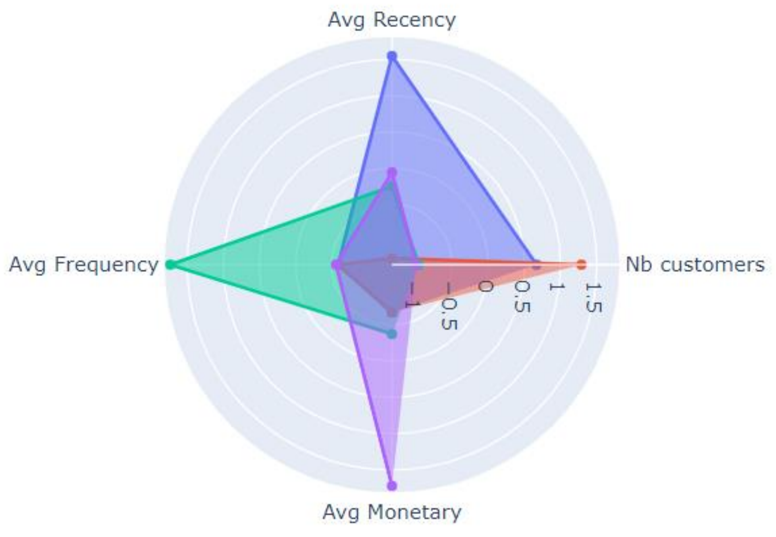
- <u>Segmentation avec un clustering K-Means (K = 4)</u>

Cluster 0 : clients perdus

Cluster 1 : nouveaux clients

Cluster 2 : clients fidèles

Cluster 3 : clients qui dépensent beaucoup



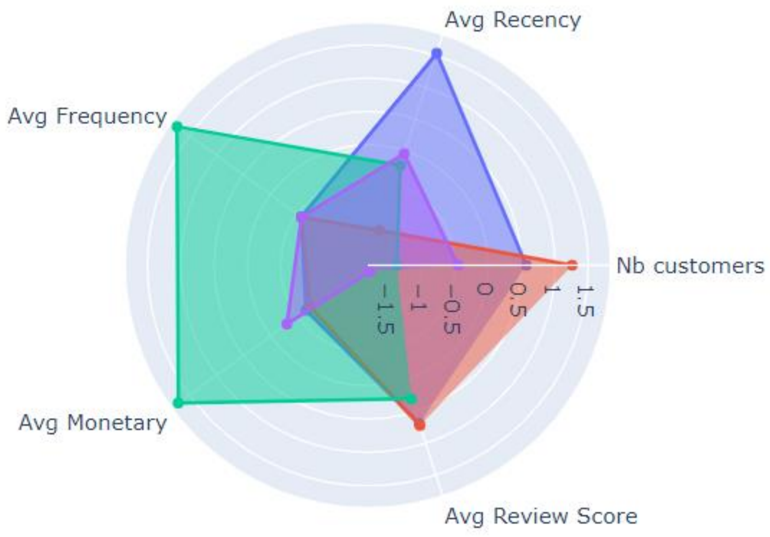
- L'ajout du **Review Score** permet de déterminer un cluster intéressant : les clients mécontents

Cluster 0 : clients perdus

Cluster 1 : nouveaux clients

Cluster 2 : clients « royaux » = fidèles et dépensent beaucoup

Cluster 3 : clients mécontents et assez récents



<u>- Maintenance : une fois par mois</u>

Merci !