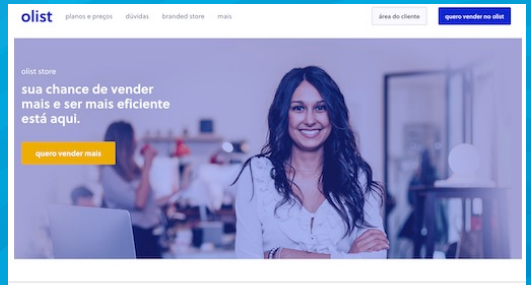


La page d'accueil du site Olist

Segmenter des clients d'un site e-commerce

Soutenance Projet 5 Data Scientist Agnès Regaud août 2023

Sommaire

1.Problématique

Projet de segmentation des clients

- Olist : entreprise brésilienne qui propose une solution de vente sur les marketplaces en ligne
- But : améliorer les campagnes de communication
- Mission 1 : Différencier les bons et moins bons clients en termes de commandes et de satisfaction
- Mission 2 : Déterminer la fréquence de relance de la segmentation

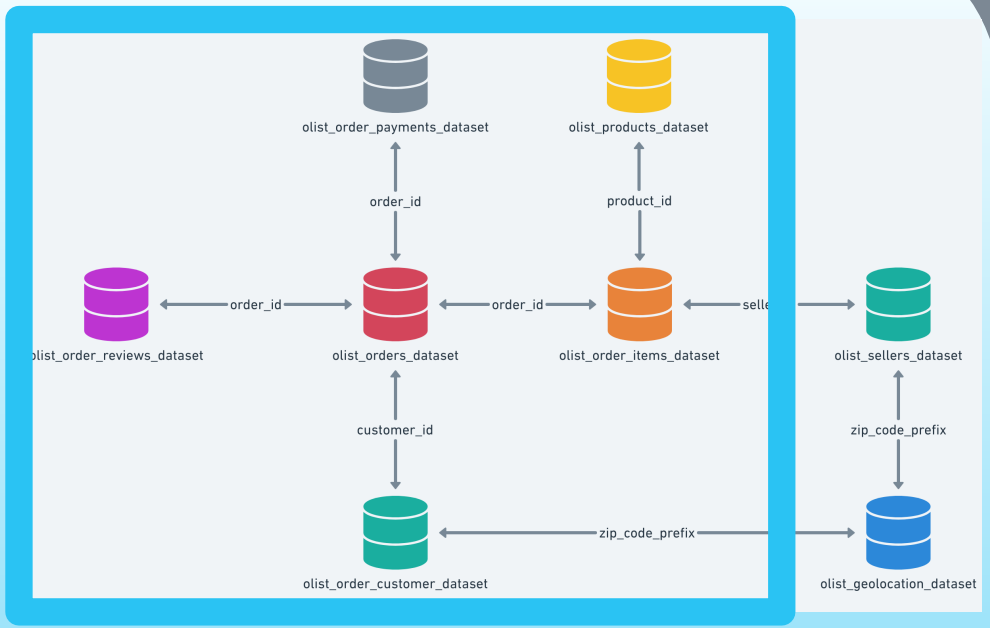
2. Exploration

Etude préliminaire des fichiers de données publiques fournis par Olist

Fichiers sources et présentation

- 9 fichiers dont 7 en rapport avec la problématique
- Environ 100 000 commandes venant de multiples marketplaces au Brésil entre 2016 et 2018
- Données anonymisées, fournies par Olist

	Fichiers	Nombre de lignes	Nombre de colonnes	Valeurs manquantes
0	Clients	99441	5	0
1	Produits	32951	9	2448
2	Commandes	99441	8	4908
3	Paiements	103886	5	0
4	Traduction des catégories de produits	71	2	0
5	Informations sur les commandes	112650	7	0
6	Avis clients	99224	7	145903



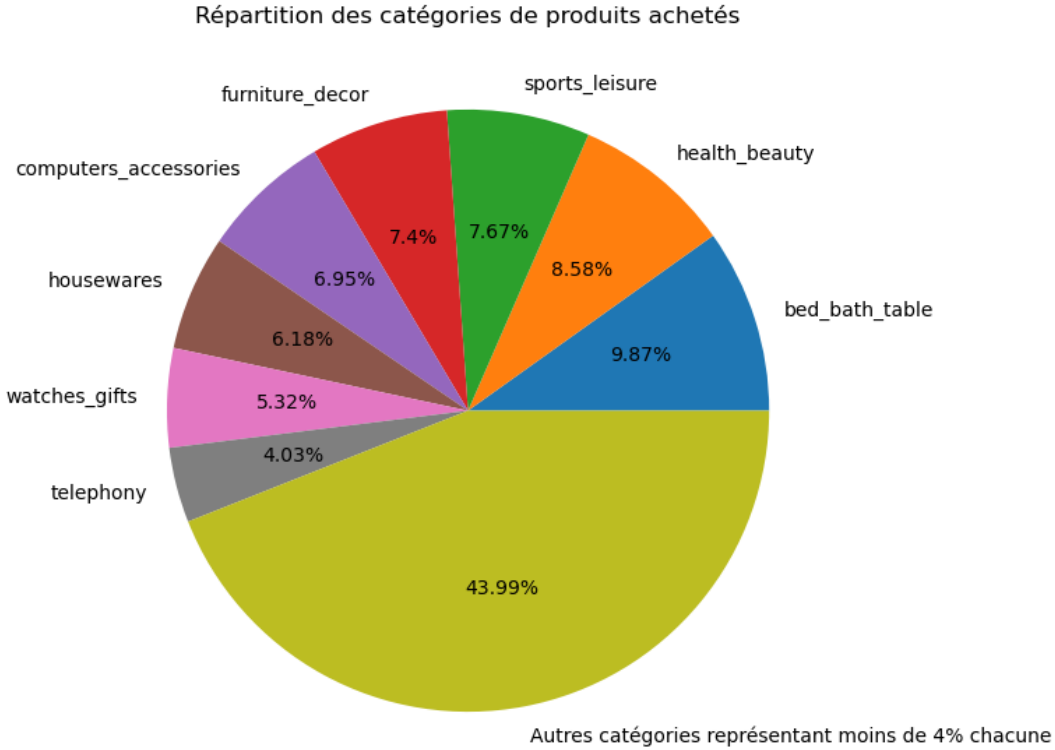
Etude préliminaire des fichiers de données publiques fournis par Olist

Remarques sur les fichiers

- 1 commande sans paiement, le client précise en commentaire que le colis n'est jamais arrivé : commande et client supprimés
- **99440** commandes, pour chacune un numéro de client, différent du numéro de client unique
- **96095** clients différents
- Chaque commande est détaillée par article, avec le type de paiement, si le paiement est fractionné, avec le montant de chaque article, et les frais de transport sont partagés équitablement sur chaque article
- Certains paiements sont faits en « voucher », 9 commandes ont tout ou partie à 0 euros
- Le client donne un avis sur la commande avec une note et un commentaire. Parfois il met un 2e avis pour la même commande, choix de ne prendre que le 2e avis avec sa note.
- Il y a 73 catégories différentes (hors « Inconnue »), et il ne manquait que 2 traductions en anglais qui ont été rajoutées. 1,42% des produits commandés n'ont pas de catégorie

Etude préliminaire des fichiers de données publiques fournis par Olist

Remarque sur les catégories de produit achetés



3. Feature engineering

Etude préliminaire des fichiers de données publiques fournis par Olist

Regroupement des variables pertinentes en seul dataset comportant les variables suivantes :

- Numéro unique du client
- Numéro du client pour chaque commande
- Date de la commande
- Numéro de la commande
- Prix de la commande avec frais de transport
- Avis du client sur la commande

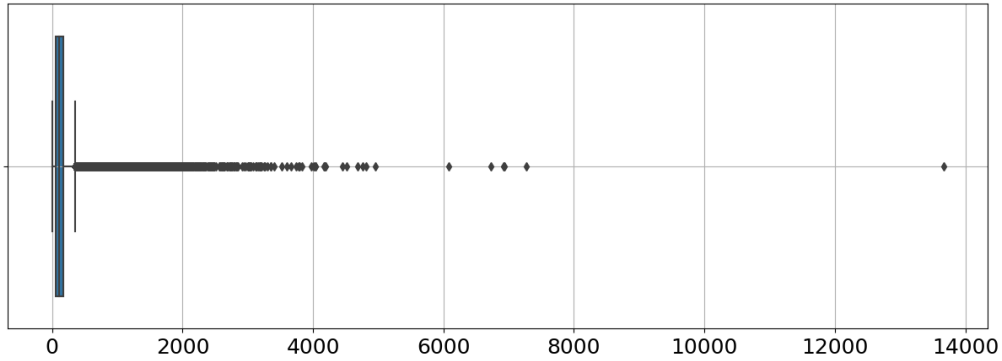
Nombre de lignes du dataset : 99440 (c'est le nombre de commandes)
Nombre de valeurs manquantes :

```
customer_unique_id              0
customer_id                     0
order_purchase_timestamp        0
order_id                        0
payment_value_agreg             0
review_score                  768
dtype: int64
```

Traitement des valeurs manquantes :

- Pour les commandes avec un seul article, on associe une catégorie
- Pour les commandes avec un seul article et une catégorie : on remplace par la moyenne des avis pour cette catégorie
- Pour les commandes restantes (140), on remplace par la moyenne de tous les avis sur toutes les commandes

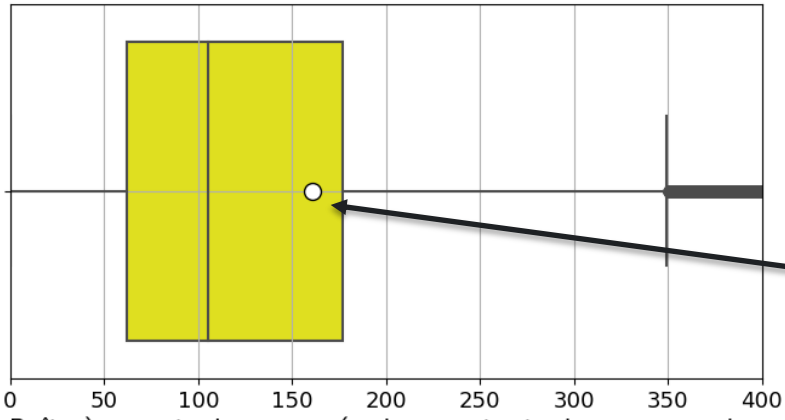
Etude préliminaire des fichiers de données publiques fournis par Olist : outliers



Boîte à moustaches des montants par commande

Remarque : la monnaie brésilienne est le réal (R$), pluriel : réaux
1 R$ ≈ 0,19 euros

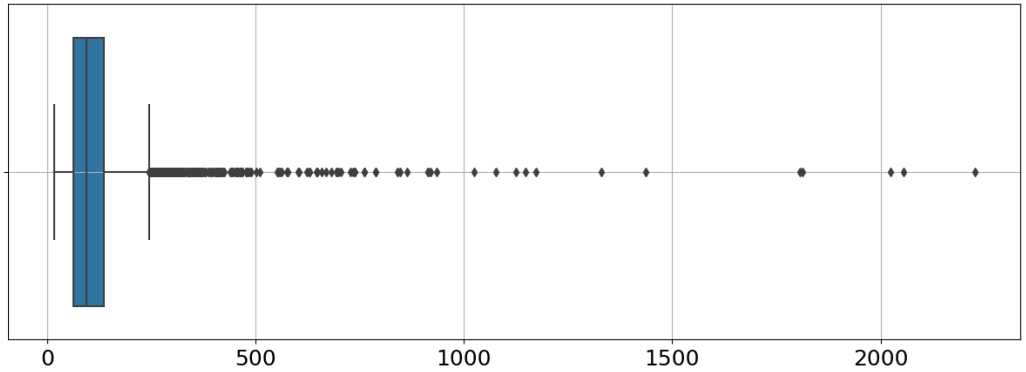
count	99440.000000
mean	160.990267
std	221.951257
min	0.000000
25%	62.010000
50%	105.290000
75%	176.970000
max	13664.080000



Boîte à moustaches zoomée des montants des commandes

Point blanc : moyenne

Etude préliminaire des fichiers de données : outliers catégorie bed_bath_table

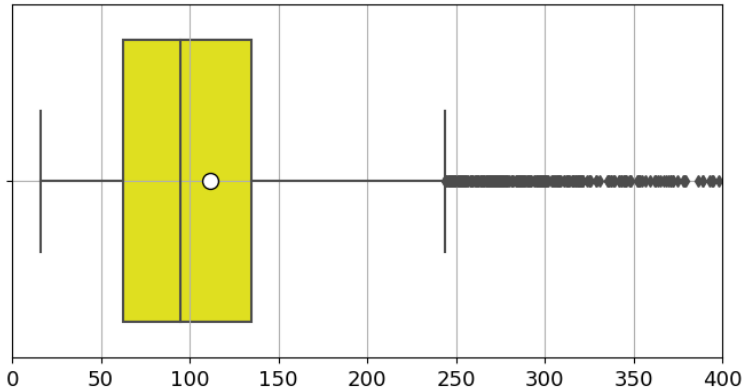


Boîte à moustaches pour les montants de la catégorie bed_bath_table

Remarque : la monnaie brésilienne est le réal (R$), pluriel : réaux
1 R$ ≈ 0,19 euros

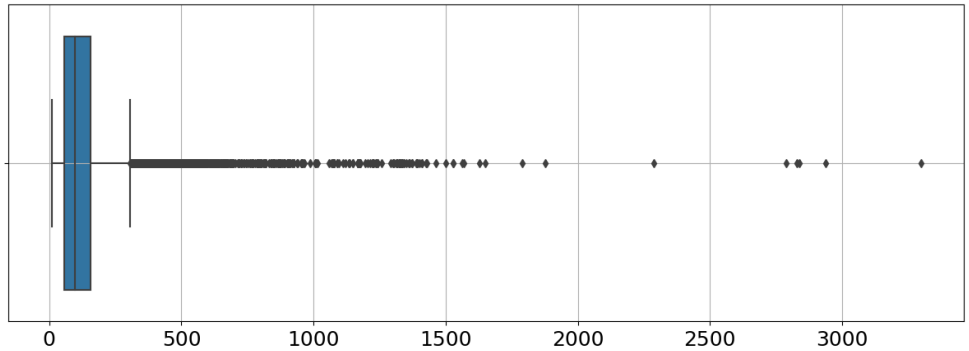
Point blanc : moyenne

count	11115.000000
mean	111.712256
std	87.615626
min	15.890000
25%	62.815000
50%	94.670000
75%	135.160000
max	2225.690000



Boîte à moustaches zoomée pour les montants de la catégorie bed_bath_table

Etude préliminaire des fichiers de données : outliers catégorie health_beauty

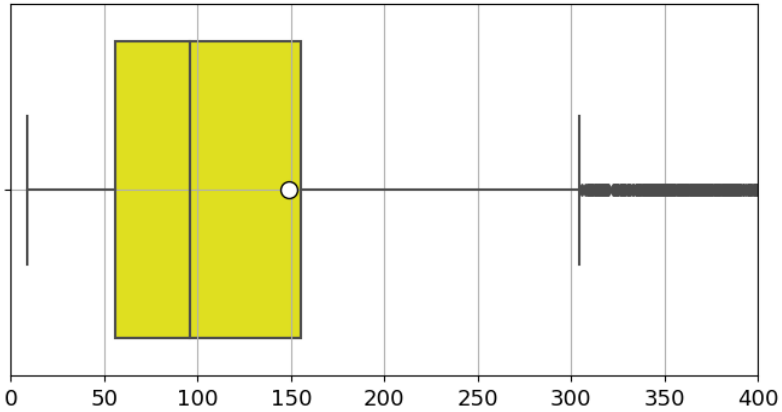


Boîte à moustaches pour les montants de la catégorie health_beauty

Remarque : la monnaie brésilienne est le réal (R$), pluriel : réaux

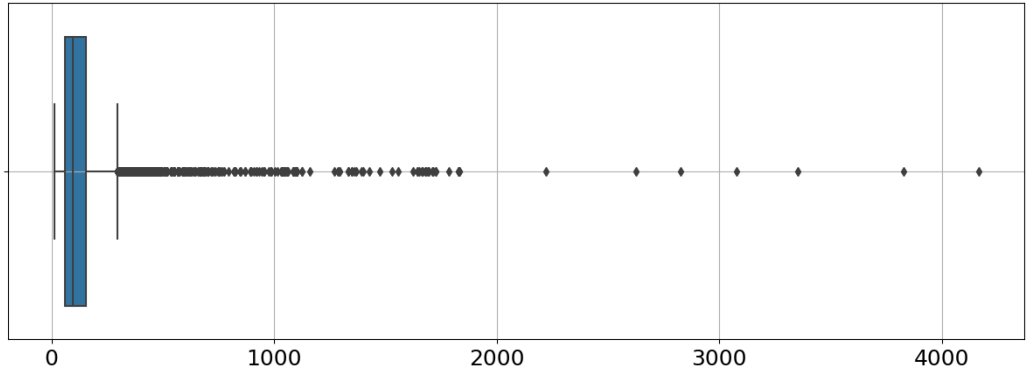
Point blanc : moyenne

count	9667.000000
mean	149.074647
std	187.615017
min	9.090000
25%	56.100000
50%	96.220000
75%	155.450000
max	3297.400000



Boîte à moustaches zoomée pour les montants de la catégorie health_beauty

Etude préliminaire des fichiers de données : outliers catégorie sports_leisure

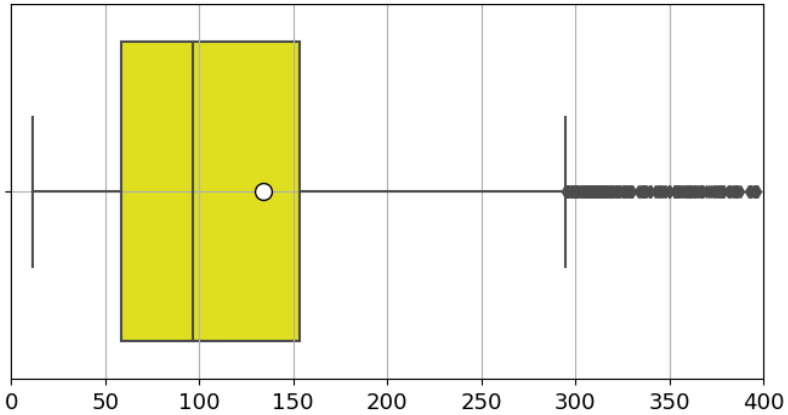


Boîte à moustaches pour les montants de la catégorie sports_leisure

Remarque : la monnaie brésilienne est le réal (R$), pluriel : réaux
1 R$ ≈ 0,19 euros

Point blanc : moyenne

count	8641.000000
mean	133.856785
std	170.653999
min	11.820000
25%	58.620000
50%	96.470000
75%	153.170000
max	4163.510000



Boîte à moustaches zoomée pour les montants de la catégorie sports_leisure

Etude préliminaire des fichiers de données : panier moyen par catégorie : bed_bath_table

	Somme payée pour la catégorie bed_bath_table par commande
count	9417.000000
mean	131.855338
std	111.779143
min	15.890000
25%	70.970000
50%	105.280000
75%	160.470000
max	2225.690000

Remarque : la monnaie brésilienne est le réal (R$), pluriel : réaux
1 R$ ≈ 0,19 euros

Dans une commande il peut y avoir plusieurs articles de même catégorie.
Ici le panier moyen d'articles dans la catégorie bed_bath_table est de 131,86 R$ alors que le prix moyen d'un article de cette catégorie parmi toutes les commandes est de 111,71 R$. Le panier moyen général est de 160,99 R$.

Etude préliminaire des fichiers de données : panier moyen par catégorie : health_beauty

	Somme payée pour la catégorie health_beauty par commande
count	8835.000000
mean	163.113142
std	201.976919
min	9.590000
25%	64.090000
50%	103.590000
75%	174.430000
max	3297.400000

Remarque : la monnaie brésilienne est le réal (R$), pluriel : réaux
1 R$ ≈ 0,19 euros

Dans une commande il peut y avoir plusieurs articles de même catégorie.
Ici le panier moyen d'articles dans la catégorie **health_beauty** est de 163,11 R$ alors que le prix moyen d'un article de cette catégorie parmi toutes les commandes est de 149,07 R$. Le panier moyen général est de 160,99 R$.

Etude préliminaire des fichiers de données : panier moyen par catégorie : sports_leisure

	Somme payée pour la catégorie sports_leisure par commande
count	7720.000000
mean	149.825969
std	186.361093
min	11.820000
25%	65.780000
50%	110.590000
75%	167.755000
max	4163.510000

Remarque : la monnaie brésilienne est le réal (R$), pluriel : réaux
1 R$ ≈ 0,19 euros

Dans une commande il peut y avoir plusieurs articles de même catégorie.
Ici le panier moyen d'articles dans la catégorie sports_leisure est de 149,83 R$ alors que le prix moyen d'un article de cette catégorie parmi toutes les commandes est de 133,86 R$. Le panier moyen général est de 160,99 R$.

4. Pistes de modélisation et modèle final sélectionné

Construction tableau RFM par client

Calcul de la récence : - Date référence de la dernière commande sur tout le dataset : 17/10/2018

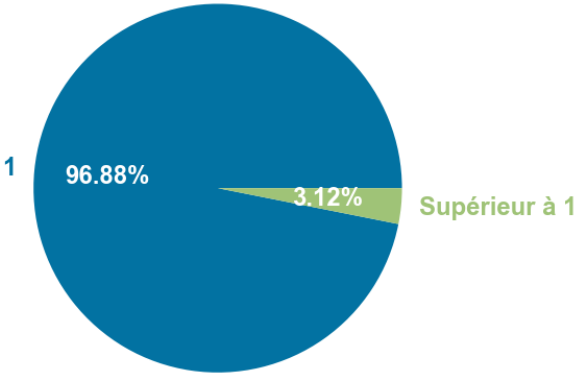
- Par client : calcul du nombre de jours entre sa dernière commande et cette date référence
- Récence max : 772 jours

Calcul de la fréquence :

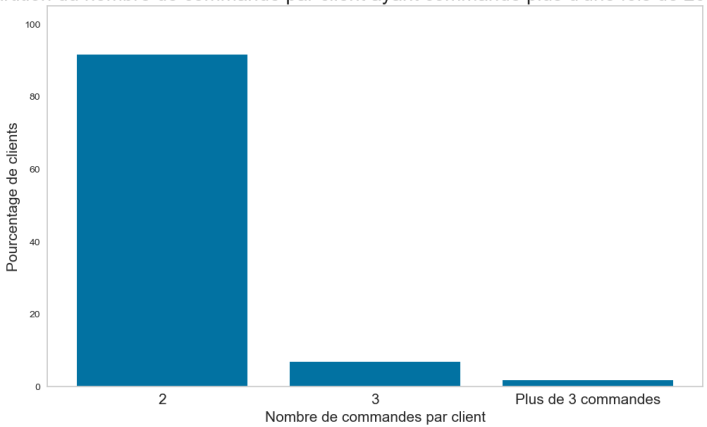
- Calcul du nombre de commandes total du client
- Varie de 1 à 17
- Pourcentage de clients ayant commandé plus d'une fois : 3,12%

	Nombre de commandes	Nombre de clients	Pourcentage
0	1	93098	96.88
1	2	2745	2.86
2	3	203	0.21
3	4	30	0.03
4	5	8	0.01
5	6	6	0.01
6	7	3	0.00
7	9	1	0.00
8	17	1	0.00

Répartition des clients suivant le nombre de commandes



Répartition du nombre de commande par client ayant commandé plus d'une fois de 2016 à 2018



	Nombre de commandes	Pourcentage de client
0	2	91.59
1	3	6.77
2	Plus de 3 commandes	1.63

Construction tableau RFM par client

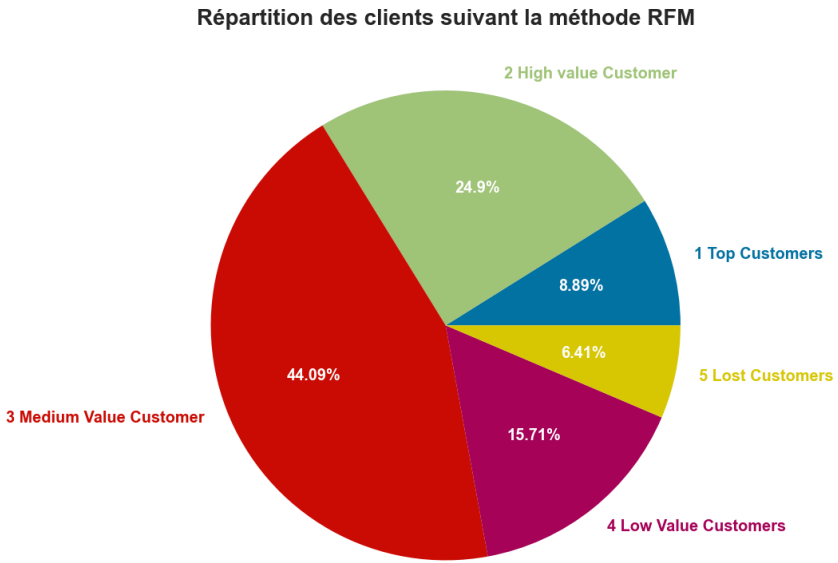
- **Calcul du montant par client** : Montant total payé par client

Tableau RFM avec avis du moyen du client, calcul du RFM_score et RFM_segment :

	customer_unique_id	Recency	Frequency	Monetary	R_rank_norm	F_rank_norm	M_rank_norm	RFM_Score	avis moyen du client	RFM_segment
0	87ab9fec999db8bd5774917de3cdf01c	0	1	89.71	100.0	48.44	41.25	3.16	1.0	143
1	262e1f1e26e92e86375f86840b4ffd63	0	2	444.06	100.0	98.31	94.41	4.88	5.0	111
2	af5454198a97379394cacf676e1e96cb	13	3	592.65	100.0	99.84	96.39	4.94	1.0	111
3	634420a0ea42302205032ed44ac7fccc	16	2	160.76	100.0	98.31	69.17	4.46	2.0	112
4	9bb92bebd4cb7511e1a02d5e50bc4655	18	1	137.03	100.0	48.44	61.44	3.50	1.0	142
...	...	...	...	...	...	...	...	...	...	...
96090	2f64e403852e6893ae37485d5fcacdaf	744	1	39.09	0.0	48.44	9.37	0.96	4.0	414
96091	0eb1ee9dba87f5b36b4613a65074337c	744	1	109.34	0.0	48.44	50.56	1.65	1.0	412
96092	009b0127b727ab0ba422f6d9604487c7	764	1	40.95	0.0	48.44	10.45	0.98	1.0	414
96093	4854e9b3feff728c13ee5fc7d1547e92	772	1	75.06	0.0	48.44	33.04	1.36	1.0	413
96094	b7d76e111c89f7ebf14761390f0f7d17	772	1	136.23	0.0	48.44	61.14	1.83	1.0	412

96095 rows × 9 columns

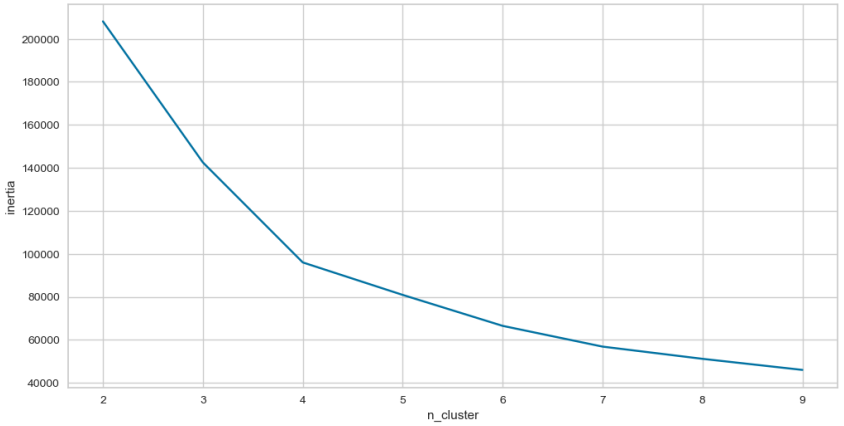
Clustering simple

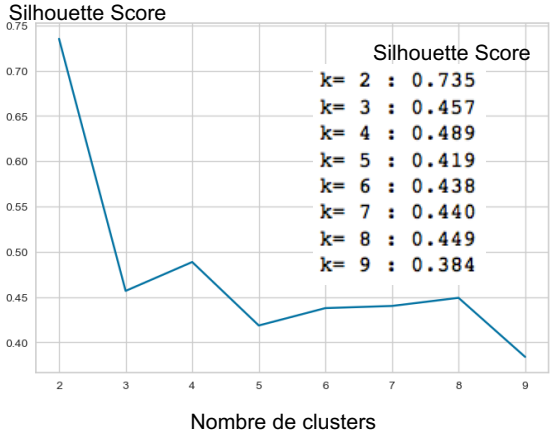


Score RFM sur 5 : S	Cluster
3,5 < S ≤ 5	1 Top Customer
2,8 < S ≤ 3,5	2 High value Customer
1,9 < S ≤ 2,8	3 Medium Value Customer
1,4 < S ≤ 1,9	4 Low value Customer
S ≤ 1,4	5 Lost Customers

I Clustering k-means
1) RFM sans rang

Sans review_score



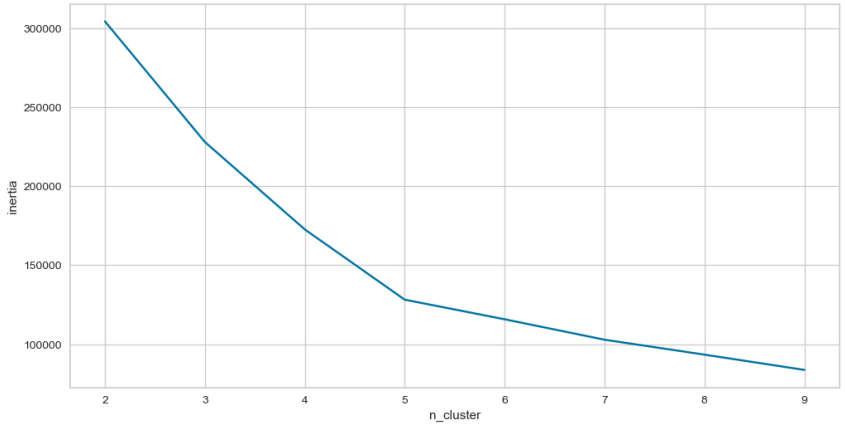


Proportion de chaque cluster pour k-means 4 et 5 clusters :

0	0.541714
1	0.402258
2	0.030824
3	0.025204

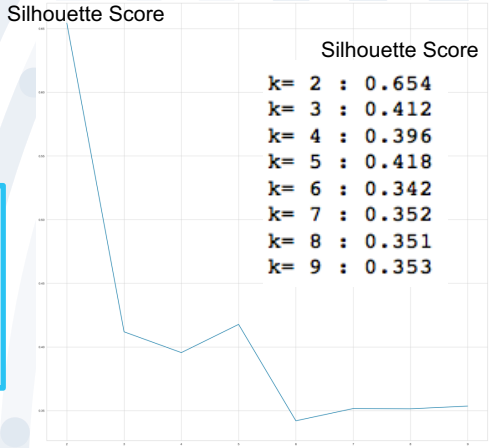
2	0.362391
1	0.358770
3	0.224538
0	0.030824
4	0.023477

Avec review_score



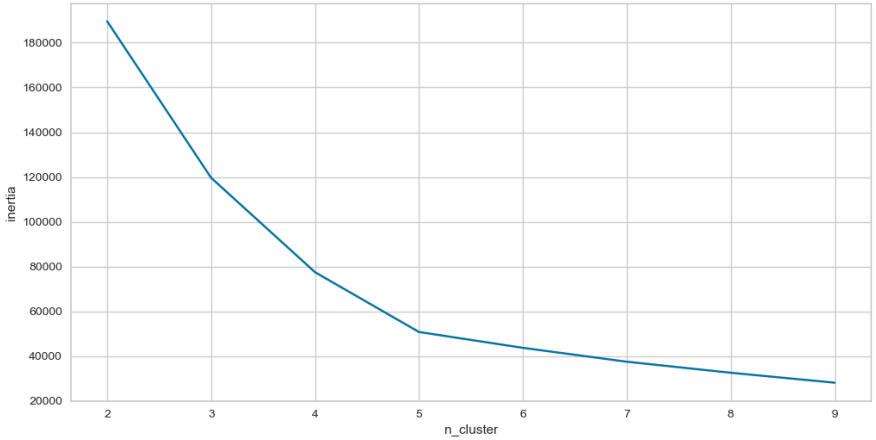
Proportion de chaque cluster pour k-means 5 clusters :

2	0.441636
1	0.332629
3	0.173901
0	0.030824
4	0.021010

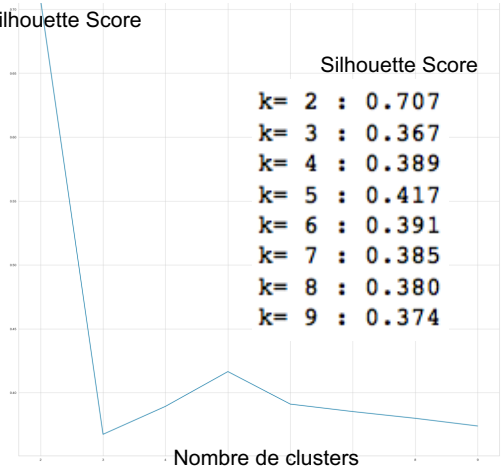


Clustering k-means
2) RFM avec rang

Sans review_score



Silhouette Score



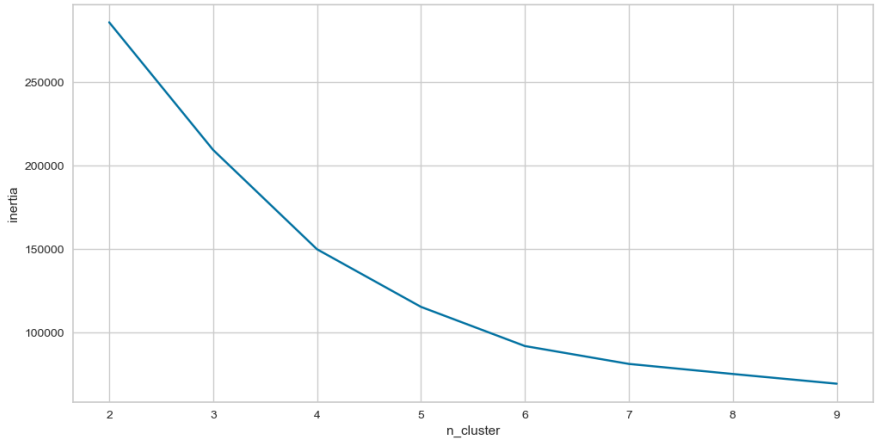
Silhouette Score

k= 2 : 0.707
k= 3 : 0.367
k= 4 : 0.389
k= 5 : 0.417
k= 6 : 0.391
k= 7 : 0.385
k= 8 : 0.380
k= 9 : 0.374

Proportion de chaque cluster pour k-means 5 clusters :

2	0.253072
0	0.242874
1	0.240481
4	0.232385
3	0.031188

Avec review_score

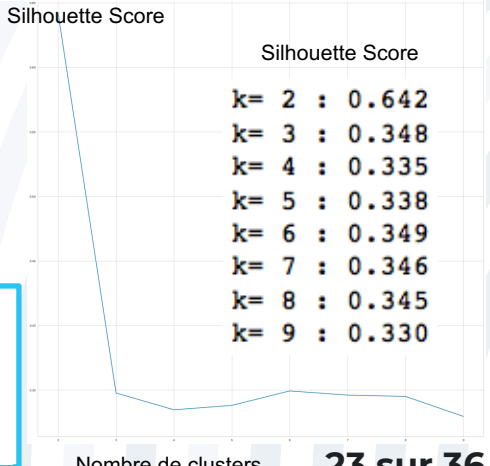


Proportion de chaque cluster pour k-means 4 et 5 clusters :

3	0.405443
0	0.394901
2	0.168469
1	0.031188

1	0.315136
0	0.264603
3	0.238733
4	0.150341
2	0.031188

Silhouette Score



Silhouette Score

k= 2 : 0.642
k= 3 : 0.348
k= 4 : 0.335
k= 5 : 0.338
k= 6 : 0.349
k= 7 : 0.346
k= 8 : 0.345
k= 9 : 0.330

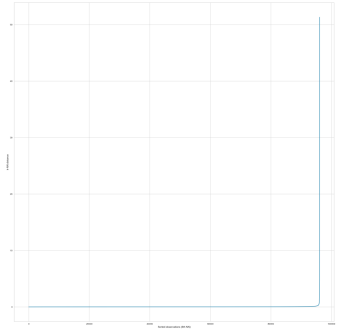
II DBScan

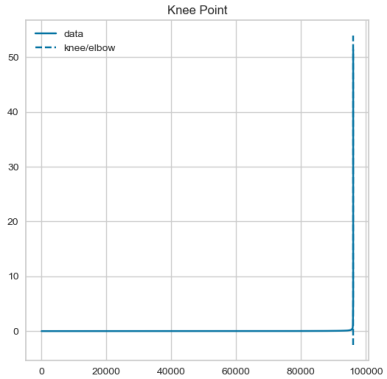
1) RFM sans review_score

Sans rang

Méthode : recherche des meilleurs hyper-paramètres eps et min_samples :

- ❖ eps à déterminer avec NearestNeighbors et KneeLocator
- ❖ min_samples = 2 x nombre de variables (ici 2x3 = 6)





Ici y = 2,11 = eps

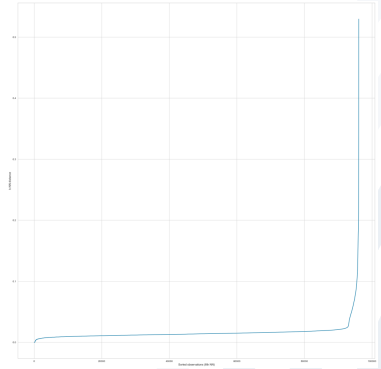
Silhouette score = 0,707

```
Estimated number of clusters: 5
Estimated number of noise points: 30
```

Proportion des 5 clusters :

0	0.968739
1	0.028513
2	0.002081
-1	0.000312
3	0.000281
4	0.000073

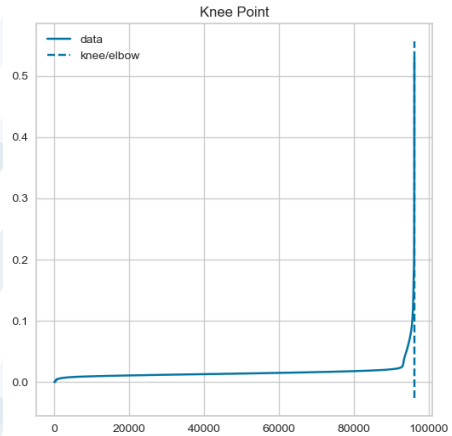
Avec rang



Ici y = 0,39 = eps

Proportion des 2 clusters :

0	0.968812
1	0.031188



Silhouette score = 0,707

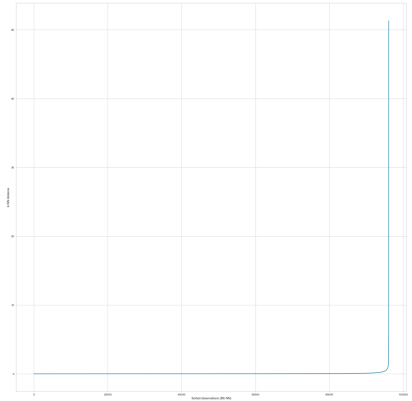
```
Estimated number of clusters: 2
Estimated number of noise points: 0
```

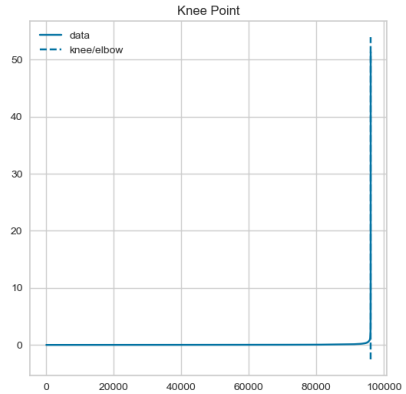
DBScan
2) RFM avec review_score

Sans rang

Méthode : recherche des meilleurs hyper-paramètres eps et min_samples :

- ❖ eps à déterminer avec NearestNeighbors et KneeLocator
- ❖ min_samples = 2 x nombre de variables (ici 2x4 = 8)





Ici y = 2,02 = eps

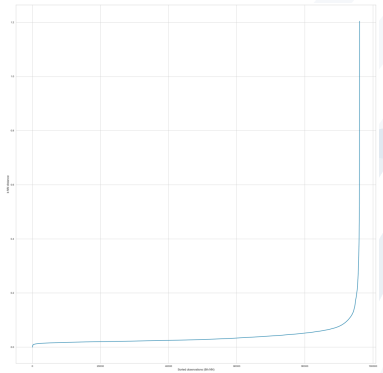
Silhouette score = 0,617

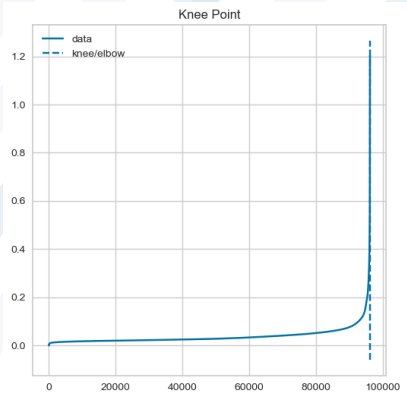
```
Estimated number of clusters: 4
Estimated number of noise points: 49
```

Proportion des 4 clusters :

0	0.968646
1	0.028503
2	0.002071
-1	0.000510
3	0.000271

Avec rang





Proportion des 10 clusters :

7	0.553702
5	0.194360
0	0.110807
6	0.079307
9	0.030636
1	0.014808
4	0.009366
8	0.003205
3	0.001956
2	0.001727
-1	0.000125

Ici y = 2,02 = eps

Silhouette score = 0,093

```
Estimated number of clusters: 10
Estimated number of noise points: 12
```

III Agglomerativ clustering

Méthode :

- ❖ RFM avec review_score
- ❖ Sampling de 10000 lignes

Sans rang

Proportion pour 4 clusters :

0	0.8399
3	0.1096
1	0.0313
2	0.0192

Proportion pour 5 clusters :

0	0.6937
4	0.1462
3	0.1096
1	0.0313
2	0.0192

Proportion pour 6 clusters :

0	0.4952
5	0.1985
4	0.1462
3	0.1096
1	0.0313
2	0.0192

Avec rang

Proportion pour 5 clusters avec rang :

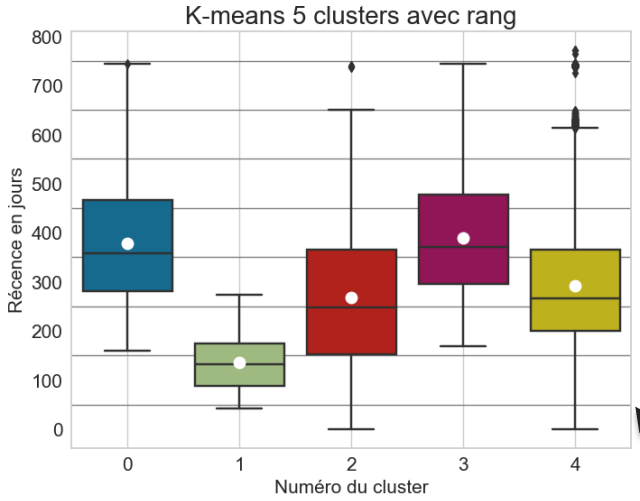
0	0.4136
1	0.2339
2	0.1851
4	0.1355
3	0.0319

Modèle choisi

Modèle choisi : K-means 5 clusters avec rang et review_score :

Résultats sur la récence par cluster pour ce modèle :

	0	1	2	3	4
count	25427.000000	30283.000000	2997.000000	22941.000000	14447.000000
mean	378.514335	135.136743	268.206874	388.763785	291.424586
std	118.295540	52.825204	145.352456	117.209964	142.294966
min	159.000000	41.000000	0.000000	168.000000	0.000000
25%	281.000000	87.000000	152.000000	295.000000	199.500000
50%	358.000000	132.000000	248.000000	371.000000	266.000000
75%	466.000000	175.000000	366.000000	477.000000	365.000000
max	744.000000	274.000000	740.000000	743.000000	772.000000

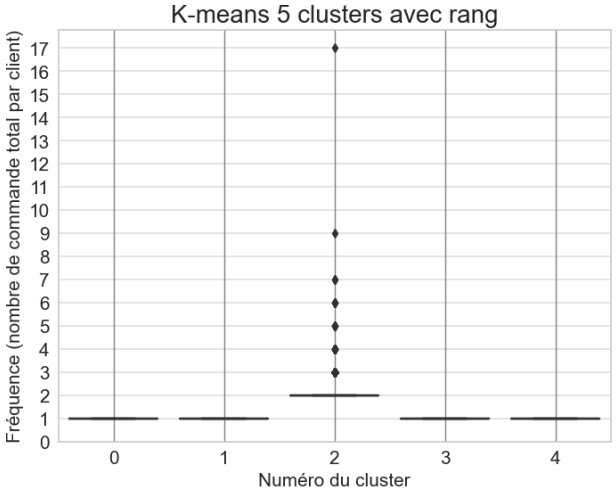


Cluster 0 : clients qui n'ont pas commandé depuis très longtemps
Cluster 1 : clients qui ont commandé récemment
Cluster 2 : clients qui n'ont pas commandé depuis longtemps
Cluster 3 : clients qui n'ont pas commandé depuis très longtemps
Cluster 4 : clients qui n'ont pas commandé depuis longtemps

Point blanc : moyenne

Résultats sur la fréquence par cluster pour ce modèle :

	0	1	2	3	4
count	25427.0	30283.0	2997.000000	22941.0	14447.0
mean	1.0	1.0	2.116116	1.0	1.0
std	0.0	0.0	0.516610	0.0	0.0
min	1.0	1.0	2.000000	1.0	1.0
25%	1.0	1.0	2.000000	1.0	1.0
50%	1.0	1.0	2.000000	1.0	1.0
75%	1.0	1.0	2.000000	1.0	1.0
max	1.0	1.0	17.000000	1.0	1.0



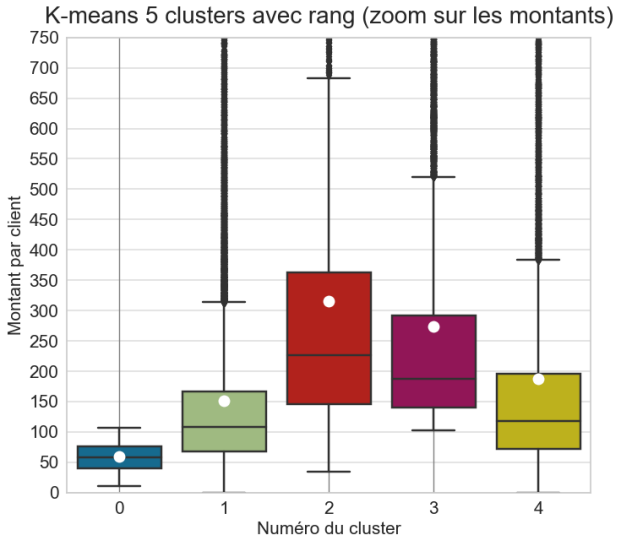
Cluster 0 : clients qui ont commandé une seule fois
Cluster 1 : clients qui ont commandé une seule fois
Cluster 2 : clients qui ont commandé au moins 2 fois
Cluster 3 : clients qui ont commandé une seule fois
Cluster 4 : clients qui ont commandé une seule fois

Modèle choisi : K-means 5 clusters avec rang et review_score :

Résultats sur le montant par cluster pour ce modèle :

Moyenne = 160,99

	0	1	2	3	4
count	25427.000000	30283.000000	2997.000000	22941.000000	14447.000000
mean	59.071472	150.960451	314.989226	273.579836	187.935816
std	22.166345	195.334188	369.591581	277.100493	284.784460
min	10.070000	0.000000	34.970000	102.900000	0.000000
25%	40.380000	68.220000	146.000000	140.080000	71.715000
50%	57.600000	108.290000	225.840000	187.530000	118.170000
75%	76.020000	166.350000	362.780000	292.340000	196.420000
max	106.870000	6922.210000	9553.020000	6929.310000	13664.080000

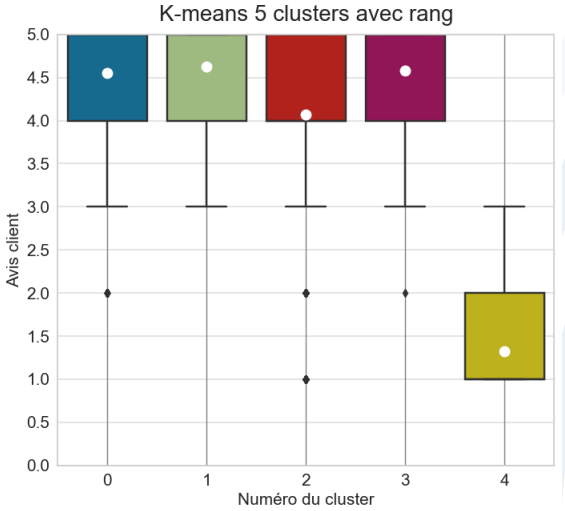


Cluster 0 : clients qui ne dépensent pas beaucoup en moyenne
Cluster 1 : clients qui dépensent un montant dans la moyenne général
Cluster 2 : clients qui dépensent 2 fois plus que la moyenne
Cluster 3 : clients qui dépensent plus que la moyenne, avec des montants assez hauts
Cluster 4 : clients qui dépensent un peu plus que la moyenne, avec des montants qui vont du mini (0) au maxi (13 664)

Point blanc : moyenne

Résultats sur l'avis client moyen par cluster pour ce modèle :

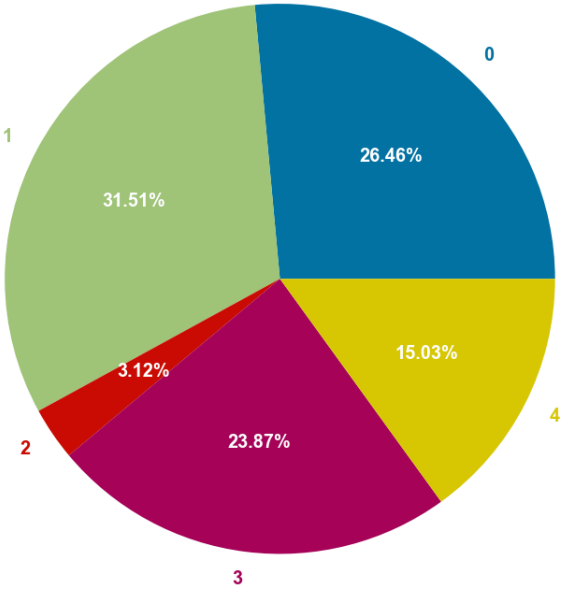
	0	1	2	3	4
count	25427.000000	30283.00000	2997.000000	22941.000000	14447.000000
mean	4.555944	4.62365	4.071738	4.581840	1.327127
std	0.676352	0.61897	1.158583	0.645337	0.590314
min	2.000000	3.00000	1.000000	2.000000	1.000000
25%	4.000000	4.00000	4.000000	4.000000	1.000000
50%	5.000000	5.00000	4.000000	5.000000	1.000000
75%	5.000000	5.00000	5.000000	5.000000	2.000000
max	5.000000	5.00000	5.000000	5.000000	3.000000



Cluster 0 : clients plutôt satisfaits en moyenne mais pouvant être assez insatisfaits
Cluster 1 : clients très satisfaits en moyenne mais pouvant être un peu insatisfaits
Cluster 2 : clients assez satisfaits en moyenne mais pouvant être très insatisfaits
Cluster 3 : clients plutôt satisfaits en moyenne mais pouvant être assez insatisfaits
Cluster 4 : clients très insatisfaits

Point blanc : moyenne

Résumé des clusters :



Cluster 0 : anciens clients pas dépensiers : clients qui n'ont pas commandé depuis très longtemps, qui ont commandé une seule fois, qui ne dépensent pas beaucoup en moyenne, plutôt satisfaits en moyenne mais pouvant être assez insatisfaits

Cluster 1 : clients récents dans la moyenne : clients qui ont commandé récemment, qui ont commandé une seule fois, qui dépensent un montant dans la moyenne général, très satisfaits en moyenne mais pouvant être un peu insatisfaits

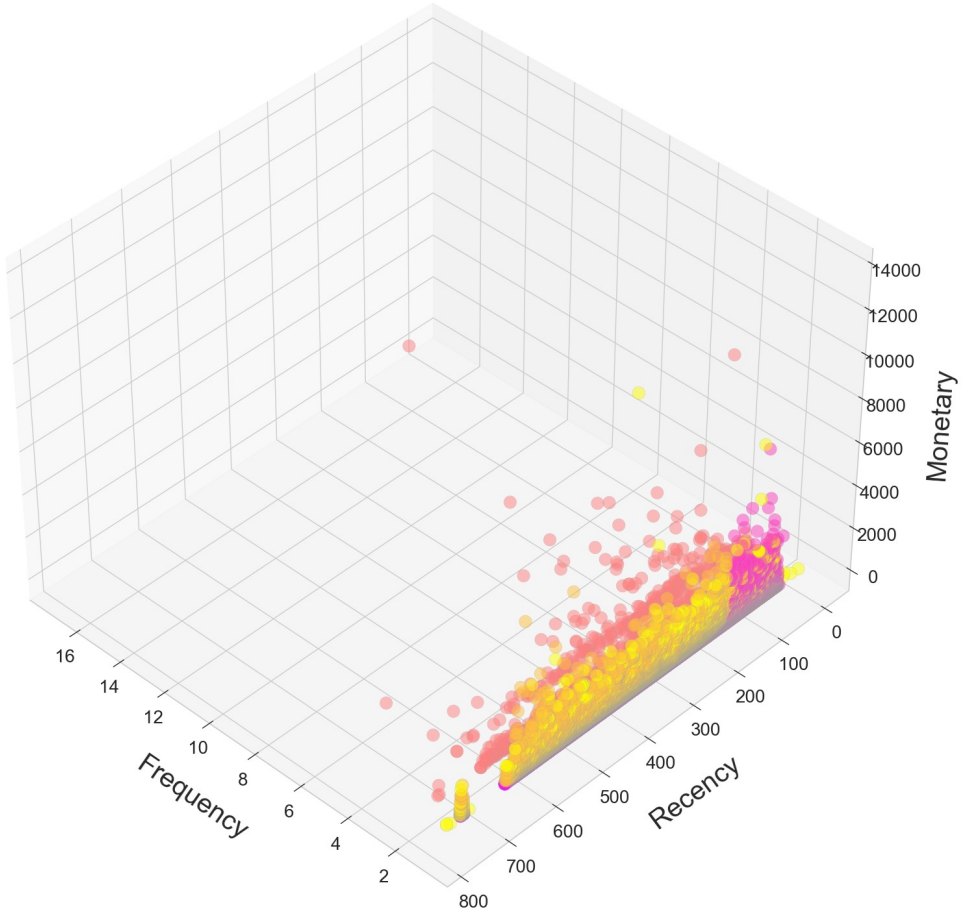
Cluster 2 : clients dépensiers qui ont commandé plusieurs fois : clients qui n'ont pas commandé depuis longtemps, qui ont commandé au moins 2 fois, qui dépensent 2 fois plus que la moyenne, assez satisfaits en moyenne mais pouvant être très insatisfaits

Cluster 3 : anciens clients dépensiers : clients qui n'ont pas commandé depuis très longtemps, qui ont commandé une seule fois, qui dépensent plus que la moyenne, avec des montants assez hauts, plutôt satisfaits en moyenne mais pouvant être assez insatisfaits

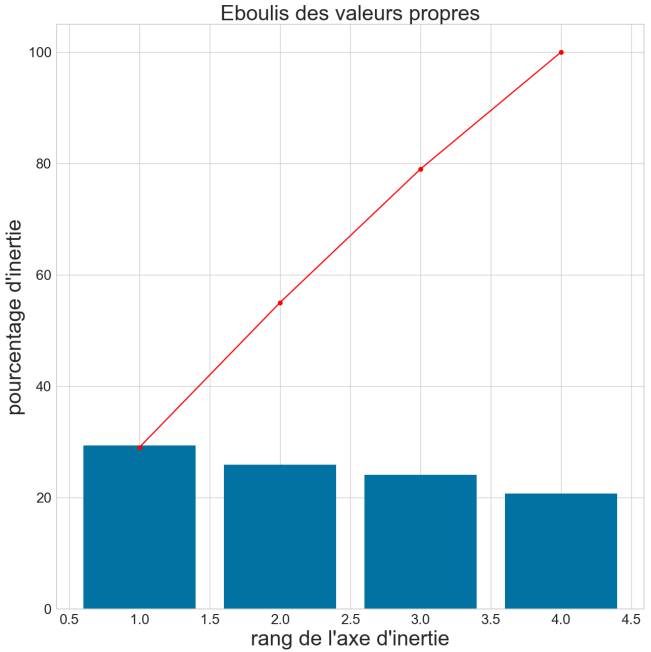
Cluster 4 : Clients insatisfaits : clients qui n'ont pas commandé depuis longtemps, qui ont commandé une seule fois, qui dépensent un peu plus que la moyenne, avec des montants qui vont du mini (0) au maxi (13664), mais très insatisfaits

Visualisation clusters :

En 3D avec RFM

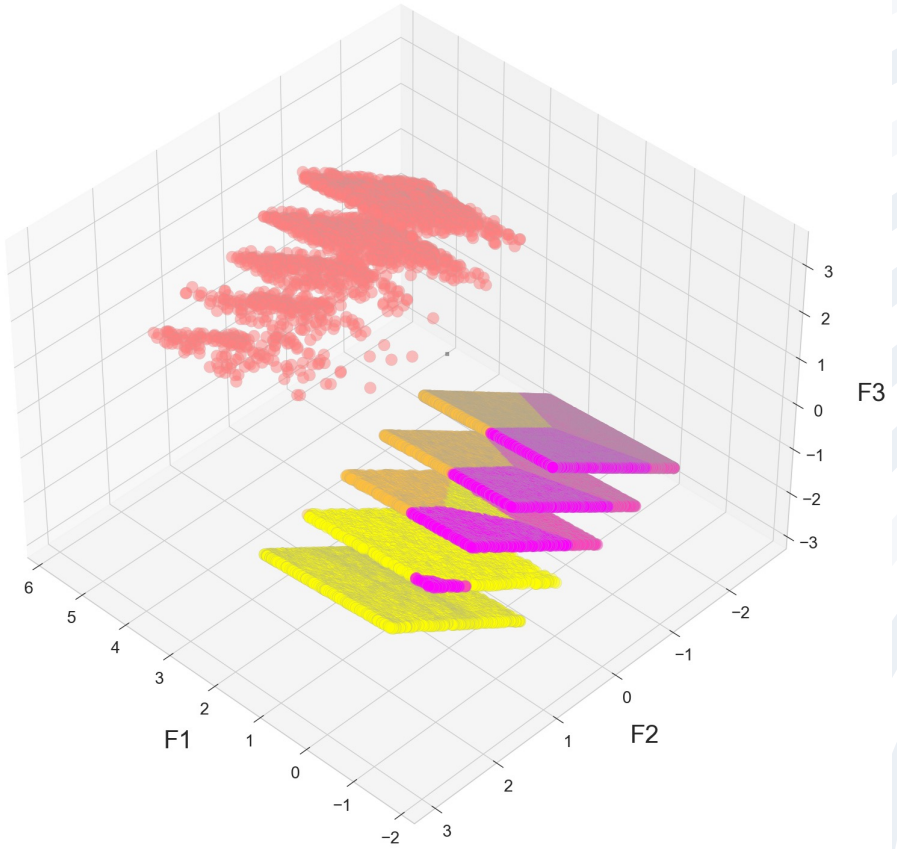


Visualisation clusters :



	R_rank_norm	F_rank_norm	M_rank_norm	avis moyen du client
F1	0.11	0.67	0.70	-0.21
F2	-0.70	-0.16	0.05	-0.70
F3	-0.71	0.27	0.04	0.65
F4	0.02	0.67	-0.71	-0.22

En 3D : ACP avec rank RFM et review-score

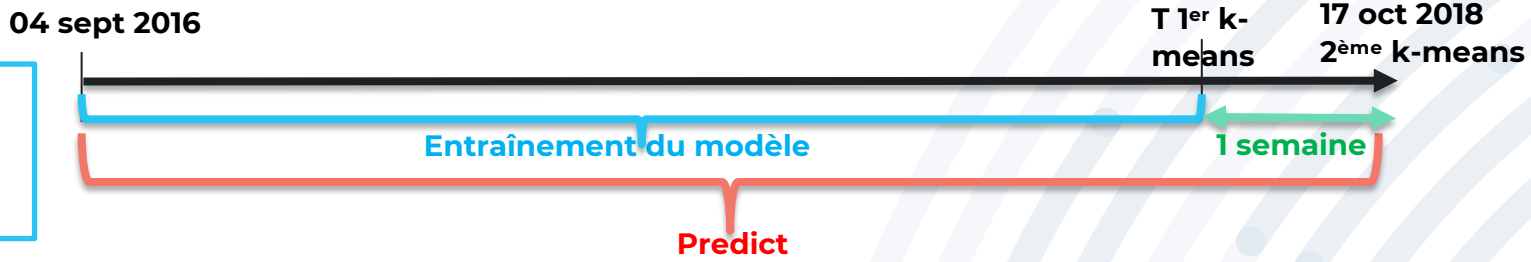


5. Présentation de la simulation pour définir le délai de maintenance du modèle (contrat de maintenance)

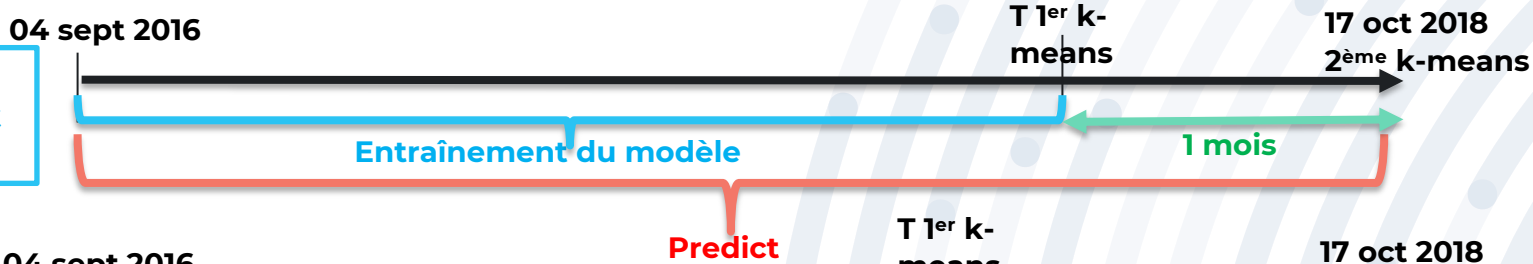
Méthode :

- K-means 5 clusters avec rangs RFM et review_score
- Détermination de la date de début $T_{début}$ et de fin T_{fin} de toutes les commandes
- Découpage de la période de commande en 2 parties : le k-means est entraîné sur la première partie : de $T_{début}$ jusqu'à une date T qui varie, avec création du RFM et du review-score au préalable (données normalisées) et on applique le predict sur tout le dataset, comme si de T à T_{fin} il y avait de nouveaux clients,
- Raisonnement à rebours : date T qui est la date du 1er k-means, et date T_{fin} qui est la date du 2ème k-means,
- Ensuite on compare les résultats entre le k-means sur tout le dataset avec le predict précédent

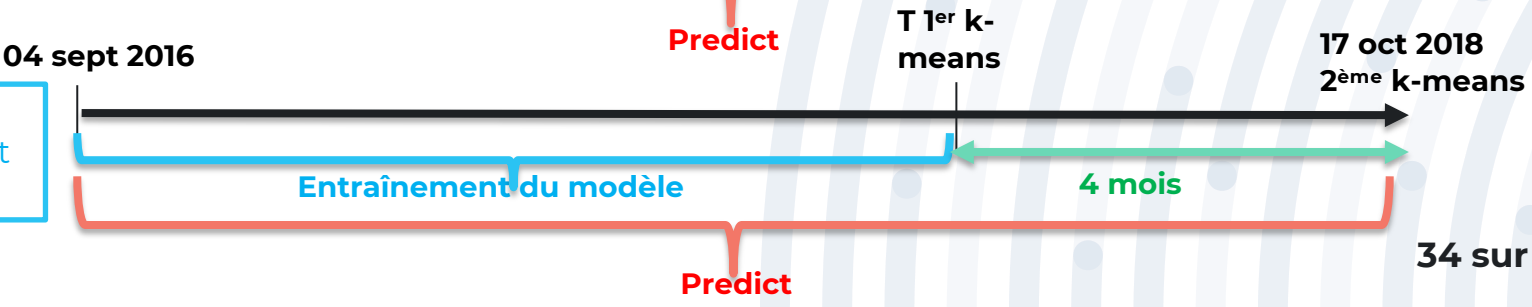
Calcul ARI : 1 semaine entre 1er et 2ème k-means :



Calcul ARI : 1 mois entre 1er et 2ème k-means :

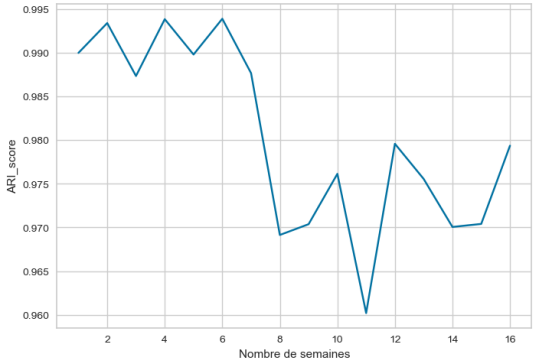


Calcul ARI : 4 mois entre 1er et 2ème k-means :



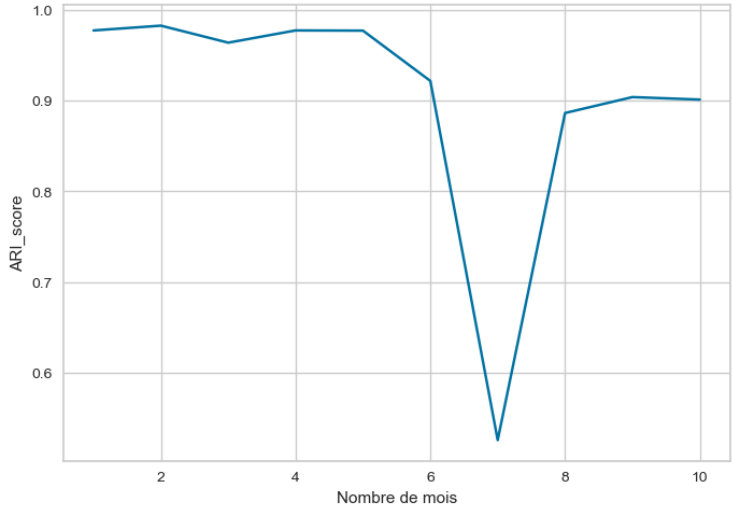
Résultats :

❖ Sur 16 semaines : on n'est pas venu depuis 1 semaine, 2 semaine, …. 16 semaines :



❖ Sur 8 mois : on n'est pas venu depuis 1 mois, 2 mois, …, 10 mois :

On observe ici une chute importante de l'ARI-score qui indique un changement de profil des commandes, il se peut qu'il y ait eu des promotions sur le site par exemple, à voir avec l'entreprise (T = 17 mars 2018) alors qu'il y a chaque mois environ 7000 commandes



Conclusion :

On peut proposer de refaire un k-means complet **tous les 6 mois.**

SOURCE

- https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce?select=olist_orders_dataset.csv