

Segmentez des clients d'un site e-commerce

Agenda

- 1 Contexte
- 2 Approche
- 3 Présentation des jeux de données
- 4 Analyse exploratoire
- 5 Segmentation RFM
- 6 Segmentation Maison
- 7 Comparatif
- 8 Prochaines étapes
- 9 Annexe : l'environnement technique

1) Contexte

Objectif	Olist souhaite disposer d'une segmentation de ses clients qui pourra être utilisée au quotidien pour les campagnes de communications.
Données mises à disposition	Olist fournit une base de données anonymisée comportant des informations sur l'historique de commandes, les produits achetés, les commentaires de satisfaction, et la localisation des clients depuis janvier 2017.
Mission	<p>Assistance aux équipes Olist pour comprendre les différents utilisateurs.</p> <p>Olist souhaite disposer de :</p> <ul style="list-style-type: none">• Une segmentation exploitable et facile d'utilisation par l'équipe marketing• Une fréquence à laquelle la segmentation doit être mise à jour• Respect de la convention PEP 8 pour le code fourni

2) Approche



1) Présentation

Présentation des jeux de données (Taille du jeu de données, nombre de lignes, nombre de colonnes, données manquantes, stratégie de consolidation en un seul fichier,...)

2) Analyse exploratoire



Analyses univariées et multivariées des données numériques, encodage des données catégorielles, ...



3) Machine Learning

Classification RFM standard, via Kmeans.
Classification "Maison" via Kmodes et Kprototypes

3) Présentation des jeux de données (1/3)

Les données sont organisées en différentes tables avec des identifiants communs permettant de faire le lien :

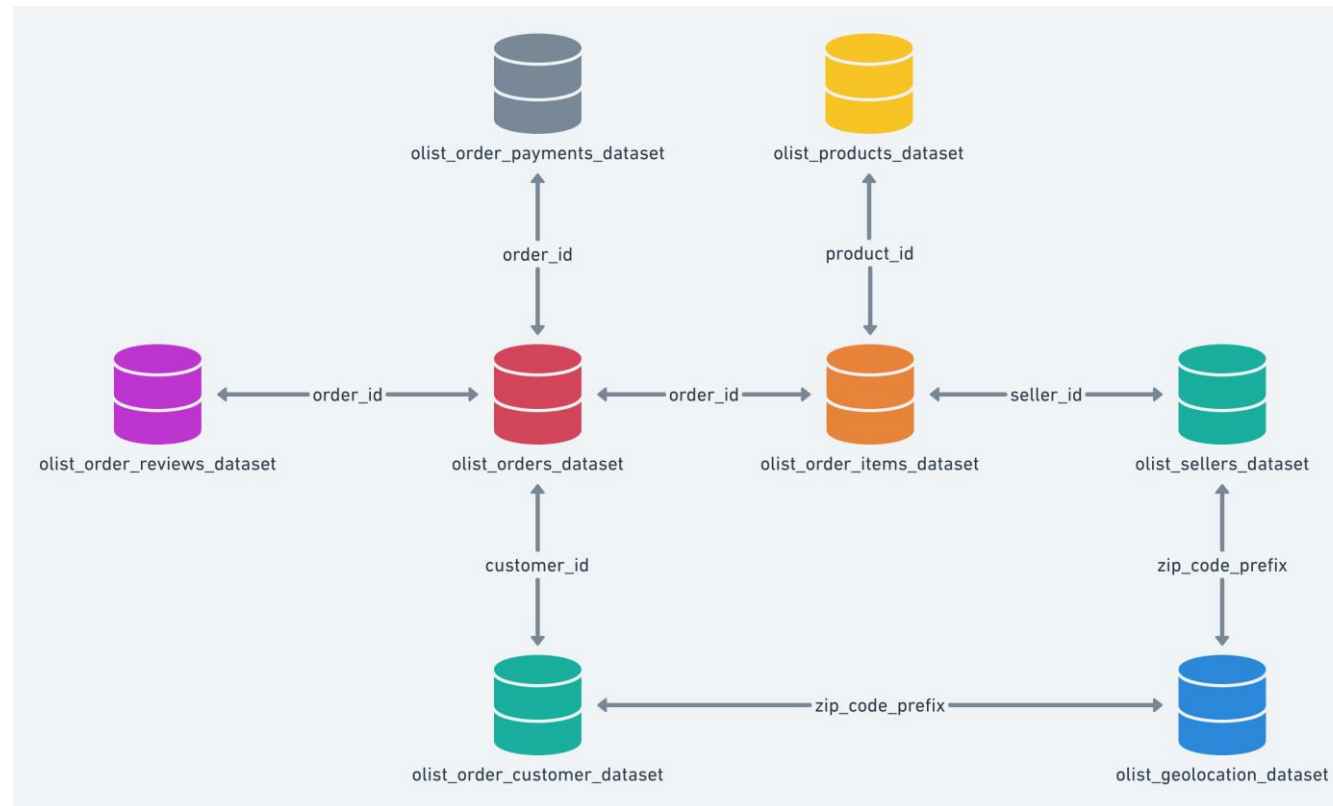


Schéma de bases de données



1) Présentation

2) Analyse exploratoire



3) Machine Learning

3) Présentation des jeux de données (2/3)

La première étape consiste à consolider l'ensemble des tables dans un seul fichier :

	order_id	customer_id	order_status	order_purchase_timestamp	order_approved_at	order_delivered_at
0	e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	2017-10-02 10:56:33	2017-10-02 11:07:15	2017-10
1	e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	2017-10-02 10:56:33	2017-10-02 11:07:15	2017-10
2	e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	2017-10-02 10:56:33	2017-10-02 11:07:15	2017-10
3	128e10d95713541c87cd1a2e48201934	a20e8105f23924cd00833fd87daa0831	delivered	2017-08-15 18:29:31	2017-08-15 20:05:16	2017-08
4	0e7e841ddf8f8f2de2bad69267ecfbcf	26c7ac168e1433912a51b924fbd34d34	delivered	2017-08-02 18:24:47	2017-08-02 18:43:15	2017-08
...
118310	1ab38815794efa43d269d62b98dae815	a0b67404d84a70ef420a7f99ad6b190a	delivered	2018-07-01 10:23:10	2018-07-05 16:17:52	2018-07
118311	b159d0ce7cd881052da94fa165617b05	e0c3bc5ce0836b975d6b2a8ce7bb0e3e	canceled	2017-03-11 19:51:36	2017-03-11 19:51:36	
118312	735dce2d574afe8eb87e80a3d6229c48	d531d01affc2c55769f6b9ed410d8d3c	delivered	2018-07-24 09:46:27	2018-07-24 11:24:27	2018-07
118313	25d2bfa43663a23586afd12f15b542e7	9d8c06734fde9823ace11a4b5929b5a7	delivered	2018-05-22 21:13:21	2018-05-22 21:35:40	2018-05
118314	1565f22aa9452ff278638e87cc895678	56772dfbcbe7df908a284ff0d53adf7d	delivered	2018-05-15 17:41:00	2018-05-16 03:35:29	2018-05

Extrait du jeu de données consolidé

118315 lignes

39 colonnes

Indicateurs clés du jeu de données



1) Présentation

2) Analyse exploratoire



3) Machine Learning

4) Analyse exploratoire

Analyses univariées sur le dataset fusionné



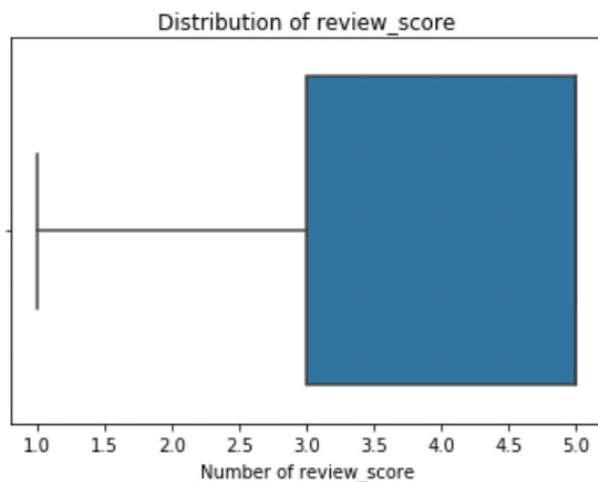
1) Présentation

2) Analyse exploratoire



3) Machine Learning

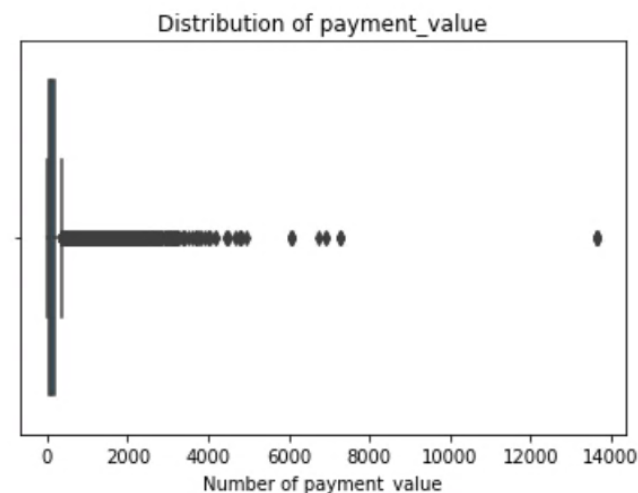
Analyse de review_score



Distribution du review_score

Caractéristiques	Valeur
Nombre de lignes	118 315
Moyenne	4
Ecart type	1,4
Min de valeurs	1
25%	3
50%	5
75%	5
Max	5

Analyse de payment_value

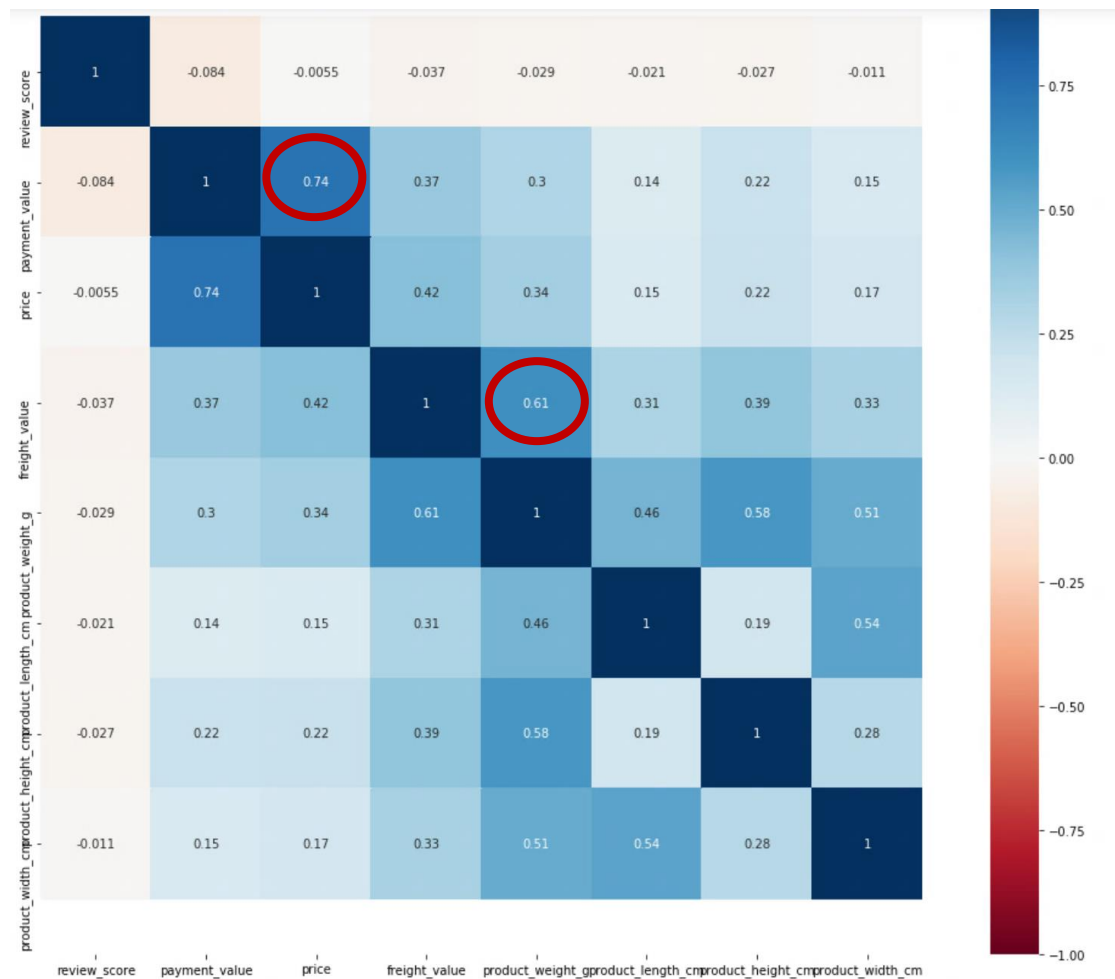


Distribution de payment_value

Caractéristiques	Valeur
Nombre de lignes	118 315
Moyenne	172,57
Ecart type	267,10
Min de valeurs	0
25%	60,85
50%	108,20
75%	189,26
Max	13664,08

4) Analyse exploratoire

Analyse des corrélations



Corrélation entre

- payment value et price
- Freight_value et Product_weight



1) Présentation

2) Analyse exploratoire



3) Machine Learning

4) Analyse exploratoire

Phases relatives à la preparation des données

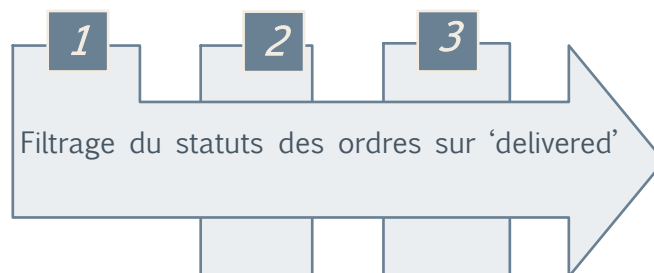


1) Présentation

2) Analyse exploratoire



3) Machine Learning



Les changements opérés sont :

Nom colonne	Action
Order_status	Filtrage sur les valeurs 'delivered'

Aggregation des données par client
(Calcul des éléments suivants : Nombre de commandes, montant des achats, ...)

Le résultat obtenu est :

	number_of_orders	total_price	total_freight	total_payment	first_order_date	last_order_date	recency	time_to_deliver
customer_unique_id								
0000366f3b9a7992bf8c76cfd3221e2	1	129.90	12.00	141.90	2018-05-10 10:56:27	2018-05-10 10:56:27	116	6.0
0000b849f77a49e4a4ce2b2a4ca5be3f	1	18.90	8.29	27.19	2018-05-07 11:11:27	2018-05-07 11:11:27	119	3.0
0000f46a3911fa3c0805444483337064	1	69.00	17.22	86.22	2017-03-10 21:05:03	2017-03-10 21:05:03	542	25.0
0000f6ccb0745a6a4b88665a16c9f078	1	25.99	17.63	43.62	2017-10-12 20:29:41	2017-10-12 20:29:41	326	20.0
0004aac84e0df4da2b147fca70cf8255	1	180.00	16.89	196.89	2017-11-14 19:45:42	2017-11-14 19:45:42	293	13.0

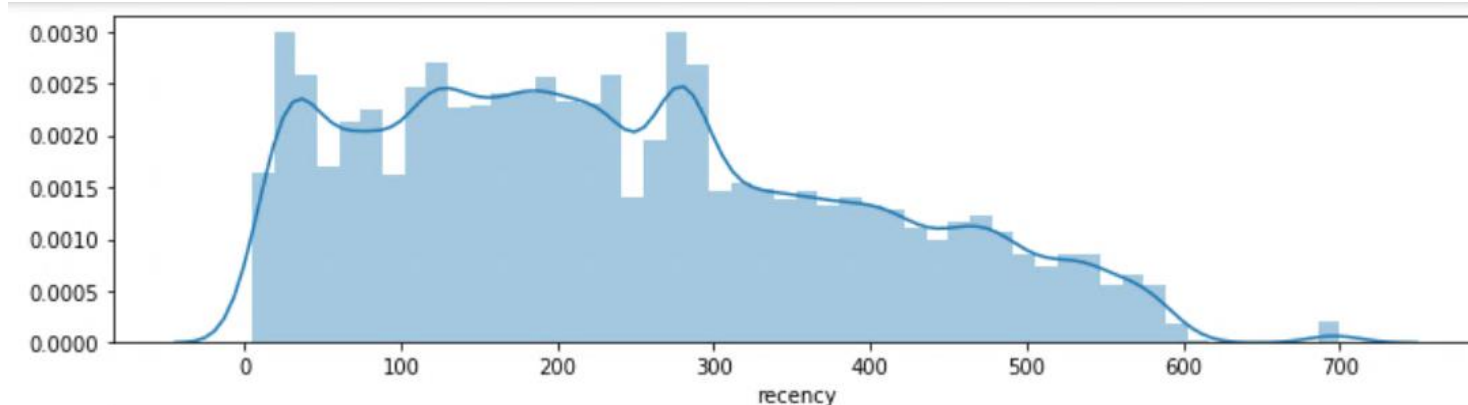
Encodage de type OneHotEncoding pour les colonnes (Product_category_name, review_score, ...)

Les changements suivants sont opérés :

#	DataToEncode	#	DataToEncode	Label1	Label2	Label3
1	Label1	1	Label1	1		
2	Label2	2	Label2		1	
3	Label3	3	Label3			1
4	4			

4) Analyse exploratoire

1) Analyse de Recency



Caractéristiques	Valeur
Nombre de lignes	93 357
Moyenne	242,47
Ecart type	152,58
Min de valeurs	5
25%	119
50%	223
75%	351
Max	700



1) Présentation

2) Analyse exploratoire



3) Machine Learning

4) Analyse exploratoire

Analyses univariées sur le dataset par client



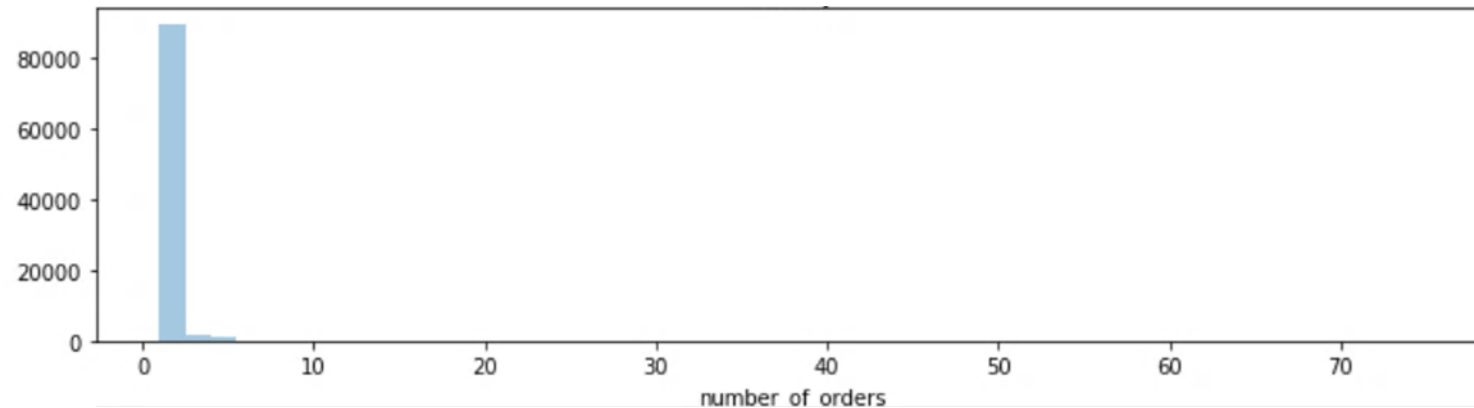
1) Présentation

2) Analyse exploratoire



3) Machine Learning

2) Analyse de Frequency



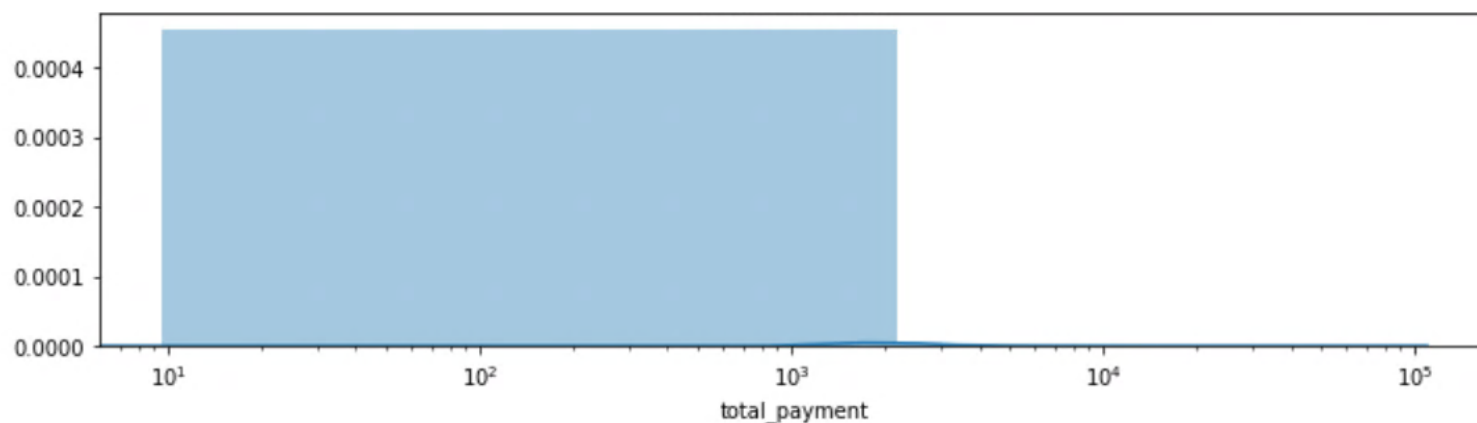
Caractéristiques	Valeur
Nombre de lignes	93 357
Moyenne	1,23
Ecart type	0,849
Min de valeurs	1
25%	1
50%	1
75%	1
Max	75

Les clients font principalement un seul achat sur le site sur la période donnée.

4) Analyse exploratoire

Analyses univariées sur le dataset par client

3) Analyse de Monetary



Caractéristiques	Valeur
Nombre de lignes	93 357
Moyenne	212,98
Ecart type	646,23
Min de valeurs	9,59
25%	63,83
50%	113,14
75%	202,66
Max	109 312



1) Présentation

2) Analyse exploratoire



3) Machine Learning

5) Segmentation RFM standard

Approche : 1) Affectation d'une valeur à chaque domaine

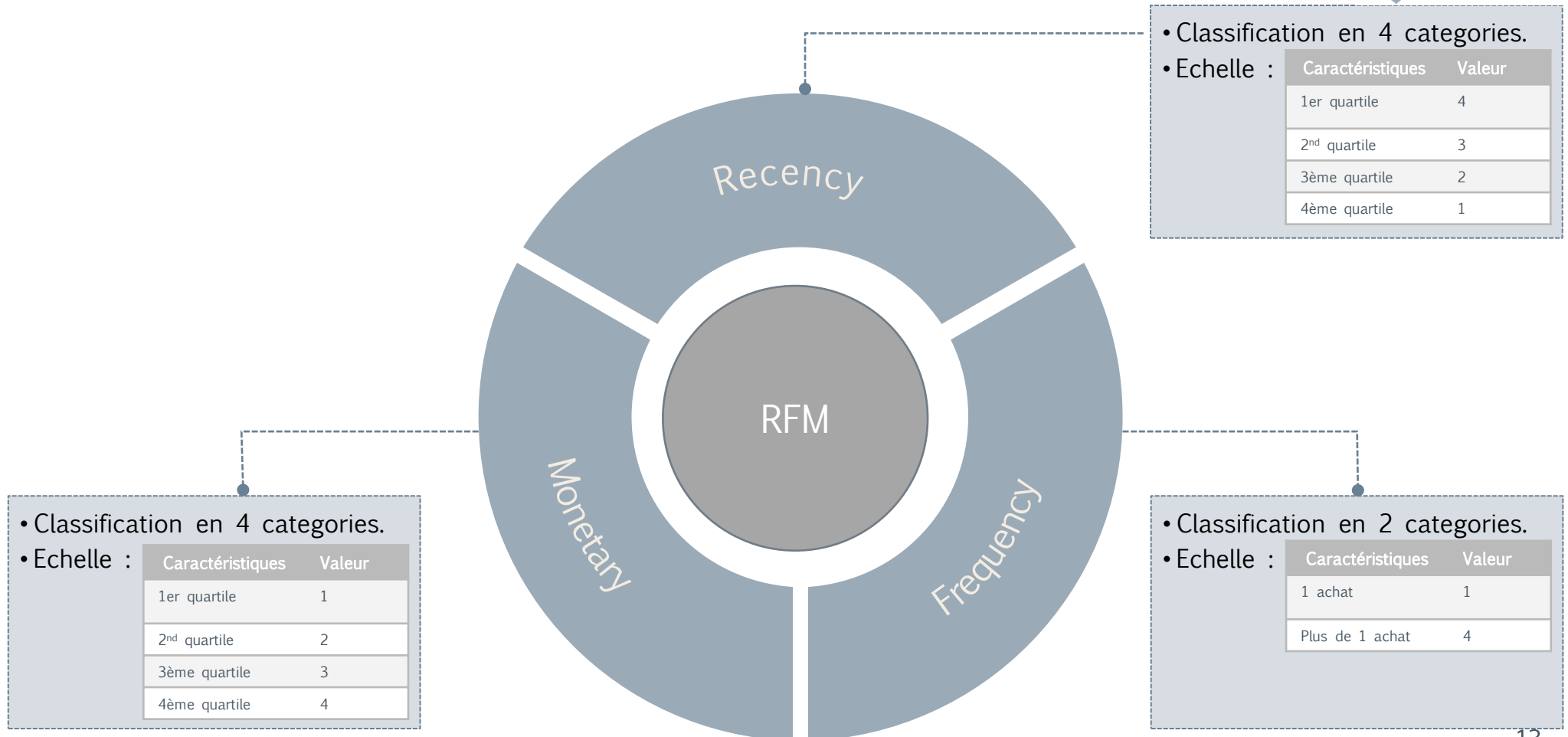


1) Présentation

2) Analyse exploratoire



3) Machine Learning





5) Segmentation RFM standard

Approche : 2) Calcul du score RFM

Score RFM Client = Valeur Recency Client + Valeur Frequency Client + Valeur Monetary Client

#	Segment	Score RFM	Recency Mean	Frequency Mean	Monetary Mean	Client Number
1	Can't Loose Them	≥ 9	184,1	2,3	592,8	15 614
2	Champions	8	136,5	1,2	239,0	10 106
3	Loyal	7	165,9	1,1	186,2	14 633
4	Needs Attention	6	366,3	1,0	64,8	11 184
5	Potential	5	222,5	1,0	153,1	19 911
6	Promising	4	296,8	1,0	90,3	16 210
7	Require Activation	≤ 3	459,0	1,0	43,8	5 699

Environ 50% des clients sont catégorisés dans les 3 premiers segments (Can't Loose Them, Champions, Loyal)

5) Segmentation RFM standard

Représentation graphique

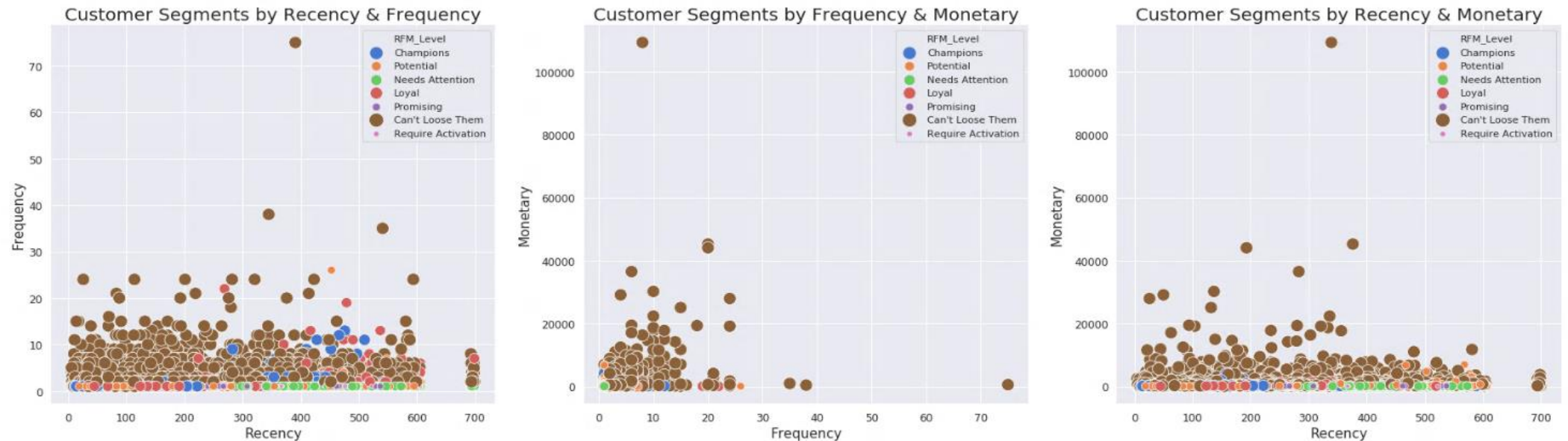


1) Présentation

2) Analyse exploratoire



3) Machine Learning



- La catégorie Can't Loose them a la fréquence des achats la plus élevée ainsi que les montants dépensés.

5) Segmentation RFM standard

Stabilité dans le temps sur les mois de Juin / Juillet / Août 2018

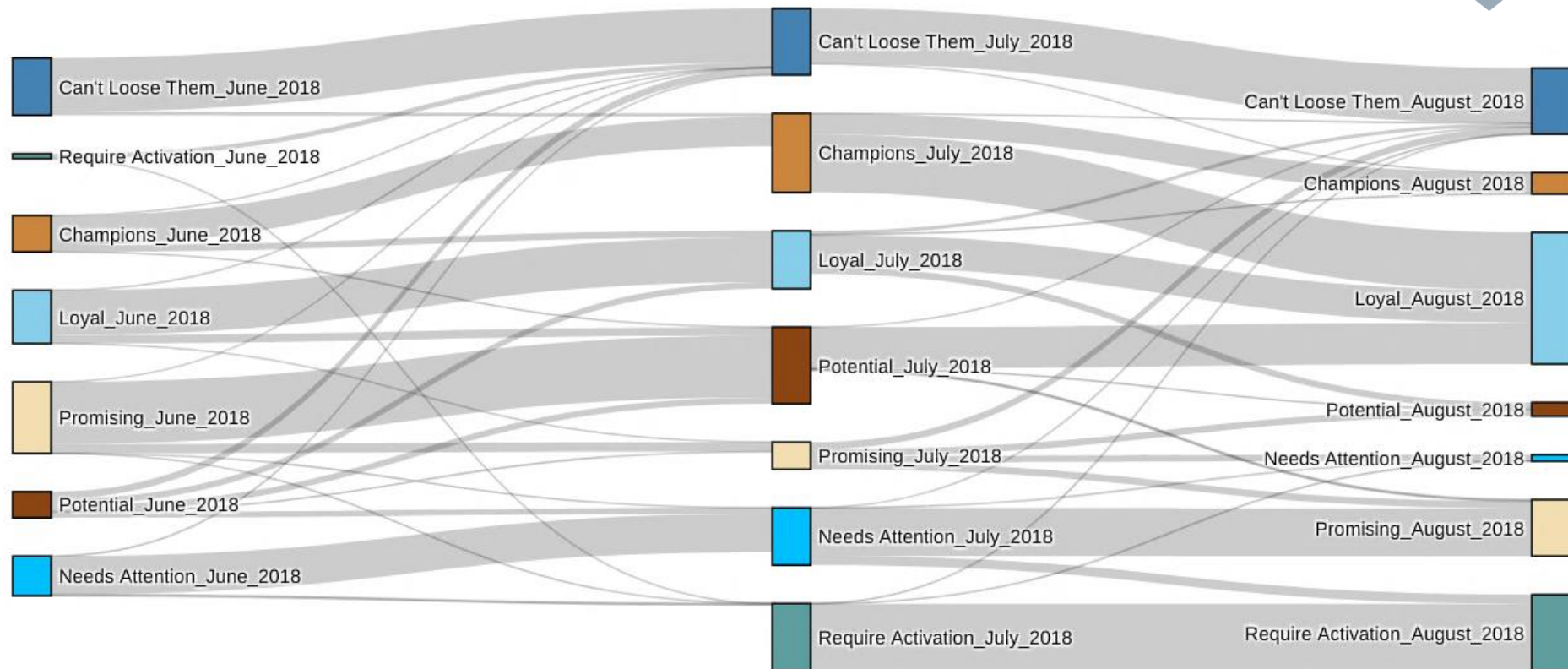


1) Présentation

2) Analyse exploratoire



3) Machine Learning



- Les segments (Can't Loose Them, Champions, Loyal) augmentent dans la globalité
- Le segment "Require Activation" augmente sur la période

5) Segmentation RFM standard

Analyse des nouveaux clients



1) Présentation

2) Analyse exploratoire



3) Machine Learning



5) Segmentation RFM standard

Suivi des clients



1) Présentation

2) Analyse exploratoire



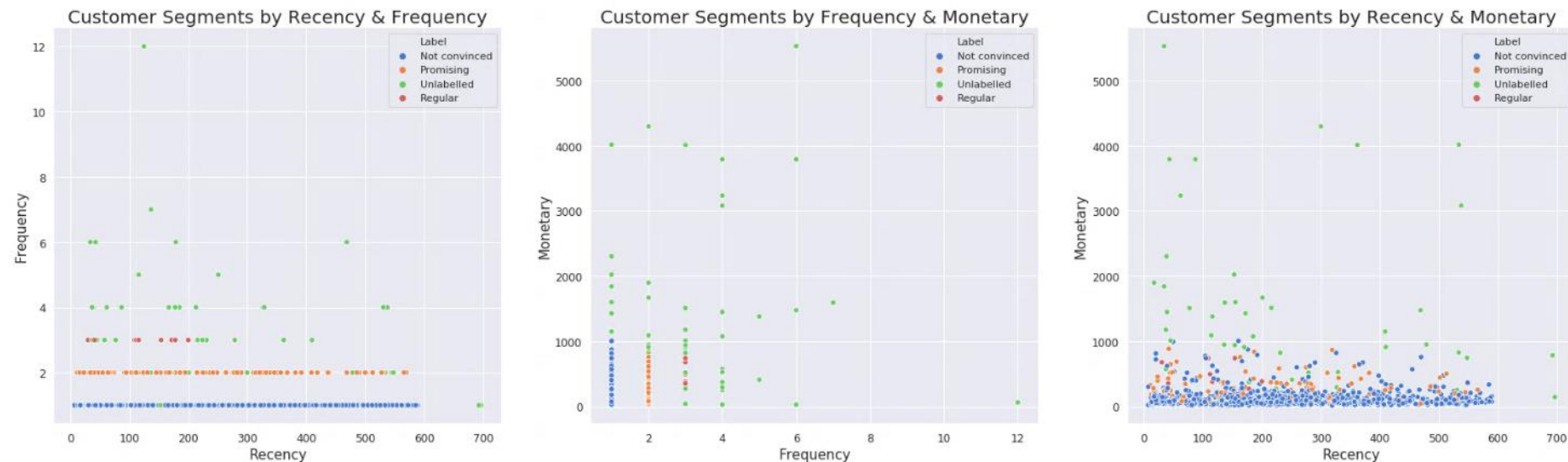
3) Machine Learning

#	Segment	Action Plan
1	Can't Loose Them	To thank their loyalty, offering them a discount on product categories that they have never purchased.
2	Champions	To thank their loyalty, offering them a discount on product categories that they have never purchased.
3	Loyal	To thank their loyalty, offering them a discount on product categories that they have never purchased.
4	Needs Attention	Made some initial purchase but have not seen them since. Was it a bad customer experience? Or product-market fit? Let's spend some resource build our brand awareness with them through a survey.
5	Potential	High potential to enter our loyal customer segments, why not throw in some gifts on their next purchase to show that you value them!
6	Promising	Showing promising signs with quantity and value of their purchase but it has been a while since they last bought sometime from you. Let's target them with their wishlist items and a limited time offer discount.
7	Require Activation	Poorest performers of our RFM model. They might have gone with our competitors for now and will require a different activation strategy to win them back such as lower prices than competitors on the categories of products that they have purchased.



5) Segmentation RFM via DBScan

Restriction : Limitation aux 1000 premières lignes car l'algorithme est gourmand en mémoire (calcul le carré du nombre d'observations)



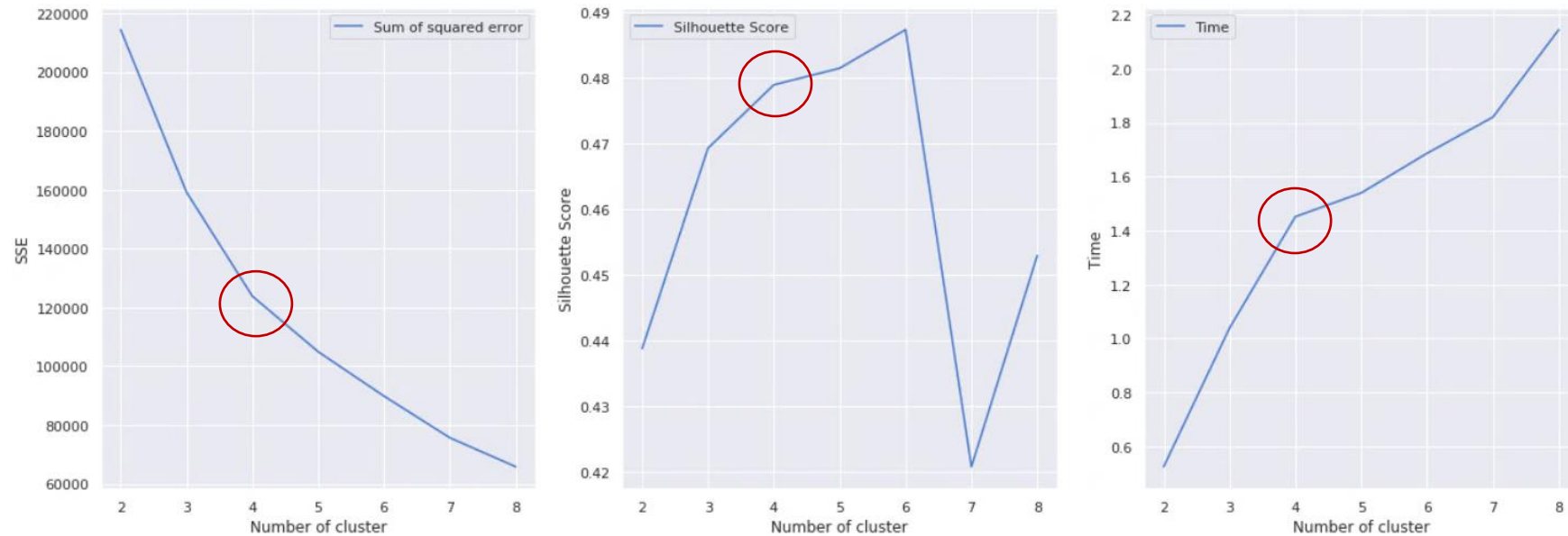
De part le nombre de calculs trop importants pour cet algorithme, ce dernier est écarté.



5) Segmentation RFM via KMeans

Approche : faire évoluer le nombre de cluster entre 2 et 8 afin d'analyser les métriques suivants :

- Sum of Squared Error
- Silhouette Score
- Time

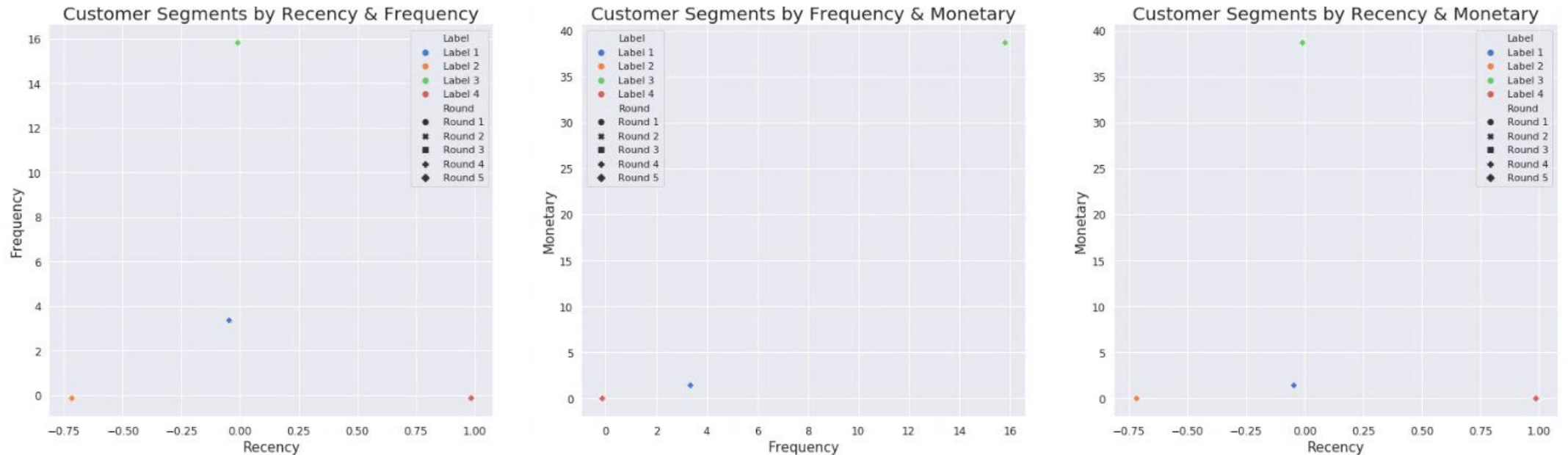


4 Clusters est un bon compromis parmi les métriques sélectionné.

5) Segmentation RFM via KMeans

Analyse de la stabilité des centroids

Approche : faire évoluer le paramètre random_state dans les valeurs 1, 5, 10, 15, 20



Les centroides se superposent, ils sont donc stables.



1) Présentation

2) Analyse exploratoire

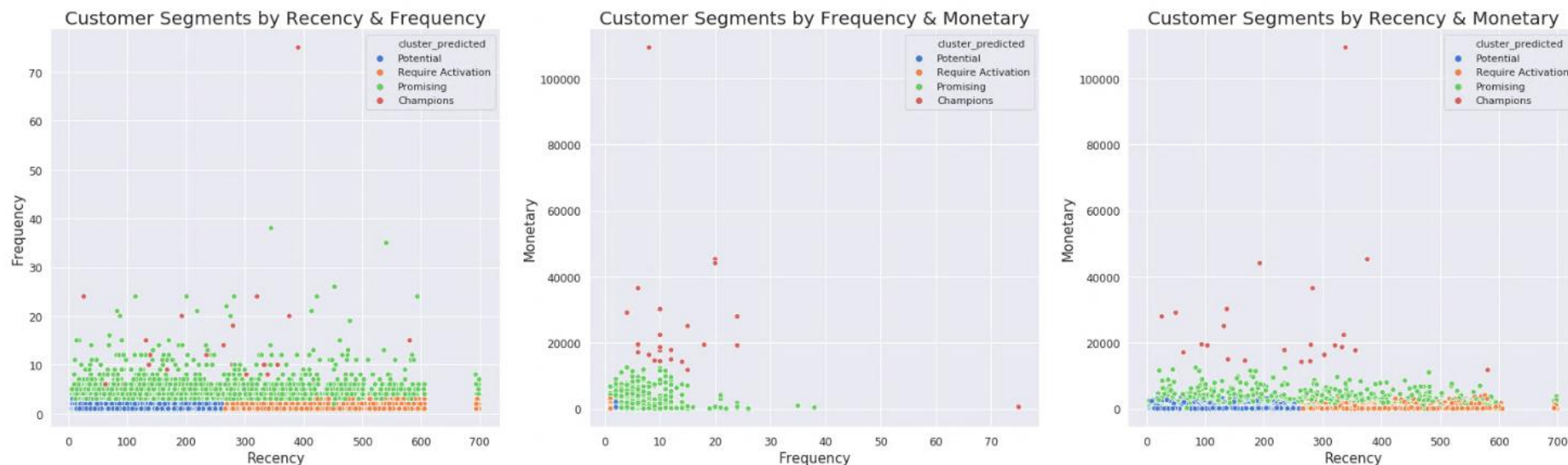


3) Machine Learning



5) Segmentation RFM via KMeans

Résultats pour 4 Clusters



#	Segment	Recency Mean	Frequency Mean	Monetary Mean	Client Number
1	Champions	241,0	14,7	25 200,6	24
2	Promising	235,7	4,1	1094,6	3 713
3	Potential	132,5	1,1	169,5	51 608
4	Require Activation	392,4	1.1	170,1	38 012

5) Segmentation RFM via KMeans

Plan d'actions



1) Présentation

2) Analyse exploratoire



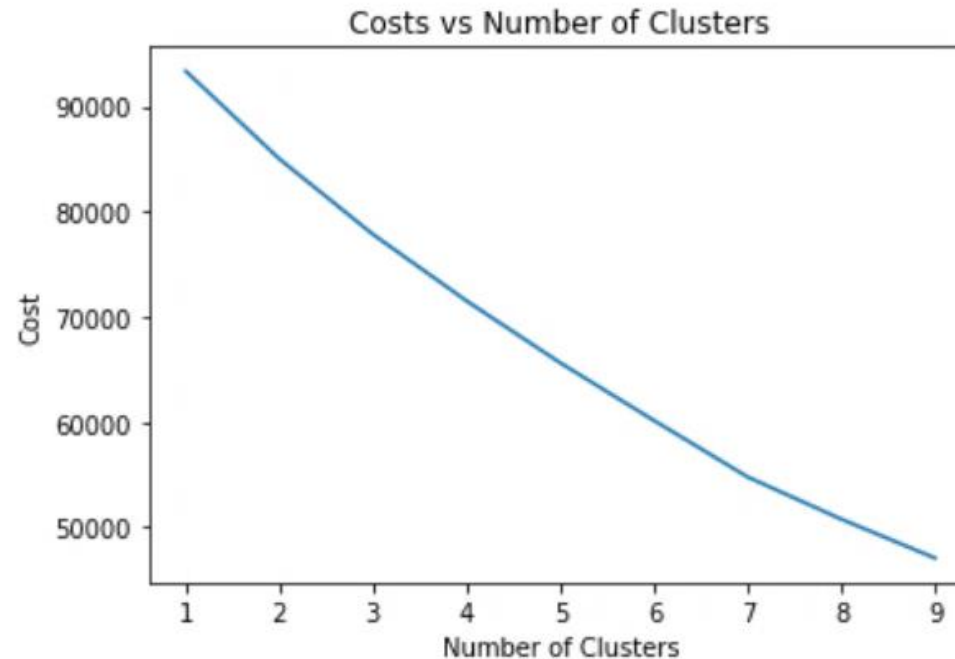
3) Machine Learning

#	Segment	Action Plan
1	Champions	To thank their loyalty, offering them a discount on product categories that they have never purchased.
2	Promising	Showing promising signs with quantity and value of their purchase but it has been a while since they last bought sometime from you. Let's target them with their wishlist items and a limited time offer discount.
3	Potential	High potential to enter our loyal customer segments, why not throw in some gifts on their next purchase to show that you value them!
4	Require Activation	Poorest performers of our RFM model. They might have gone with our competitors for now and will require a different activation strategy to win them back such as lower prices than competitors on the categories of products that they have purchased.

6) Segmentation “Maison” via KModes

Principe : segmentation catégorielle des habitudes de consommation

Approche : faire évoluer le nombre de cluster entre 2 et 8 afin d’analyser le ratio Cost vs Nombre de clusters



Absence de “vrai” coude, 8 est choisi afin de minimiser la valeur Cost



1) Présentation

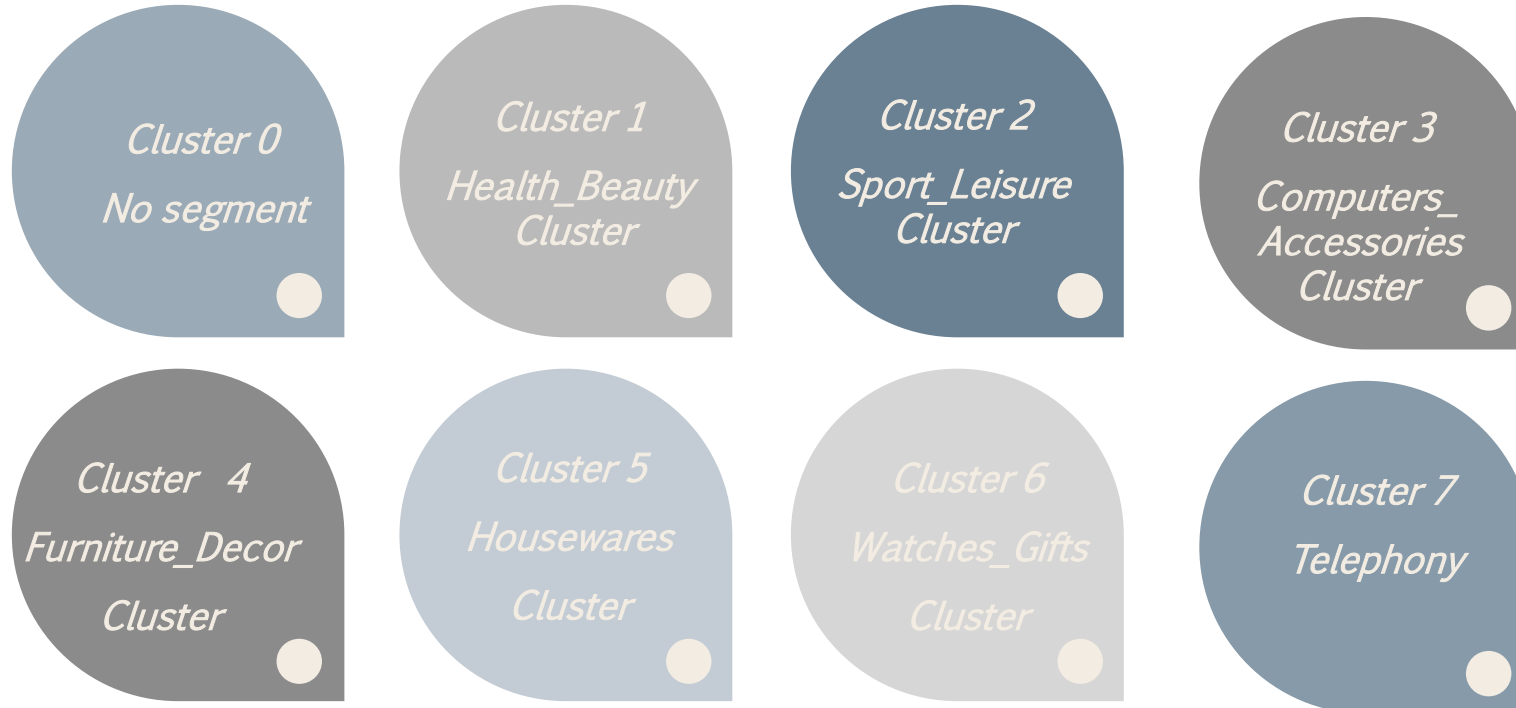
2) Analyse exploratoire



3) Machine Learning

6) Segmentation “Maison” via KModes

Résultats



1) Présentation

2) Analyse exploratoire



3) Machine Learning

6) Segmentation “Maison” via KModes

Résultats complete avec la RFM

#	Segment	Recency Mean	Frequency Mean	Monetary Mean	Client Number
1	No Segment	247,0	1,2	216,7	50 770
2	Health Beauty	220,7	1,2	194,0	8 352
3	Sports Leisure	250	1,2	188,4	7 170
4	Computers Accessoires	236,6	1,3	247,6	6 325
5	Furniture Décor	264,4	1,4	238,4	5 890
6	Housewares	228,8	1,3	193	5 472
7	Watches Gifts	207,4	1,2	261,1	5 368
8	Telephony	258,1	1,2	120,4	4 010



1) Présentation

2) Analyse exploratoire

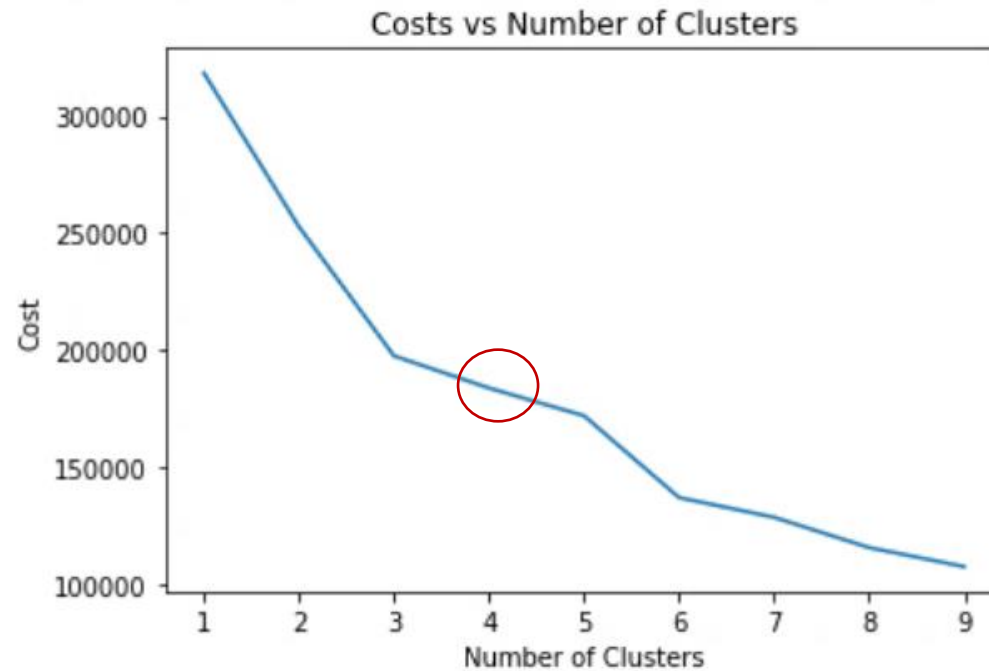


3) Machine Learning

L'algorithme dans ce cas se contente de mettre en avant les catégories où les produits sont les plus vendus car le nombre d'achat par client est proche de 1. De plus, le nombre de clients non catégorisé est important (environ 50% des clients). Pour ces raisons, cet algorithme n'est pas pertinent dans ce contexte.

6) Segmentation “Maison” via KPrototypes

Principe : segmentation catégorielle des reviews et segmentation numérique des paramètres RFM
Approche : faire évoluer le nombre de cluster entre 1 et 10 afin d’analyser le ratio Cost vs Nombre de clusters



La valeur théorique elbow serait 3 cependant, 4 clusters sont choisis car le 1er correspond aux individus non catégorisés.



1) Présentation

2) Analyse exploratoire



3) Machine Learning

6) Segmentation “Maison” via KPrototypes

Résultats



1) Présentation



2) Analyse exploratoire



3) Machine Learning

#	Segment	Recency Mean	Frequency Mean	Monetary Mean	Review Score 1	Review Score 2	Review Score 3	Review Score 4	Review Score 5	Client Number
1	Champions	226,2	14,7	25 785,6	14	0	2	3	4	23
2	Promoters	247,7	4,9	1504,5	474	128	191	368	1031	2 192
3	Potential	164,6	1,2	177,5	4 288	1 342	3 717	0	39 013	48 360
4	Require Activation	330,3	1,1	173,2	4 648	1 440	3 846	17 999	14 849	42 782

La catégorie “Champions” qui a procédé à la quantité la plus importante d’achats semble mécontente, il est nécessaire d’investiguer.

6) Segmentation “Maison” via KPrototypes

Suivi des clients



1) Présentation
















2) Analyse exploratoire



3) Machine Learning

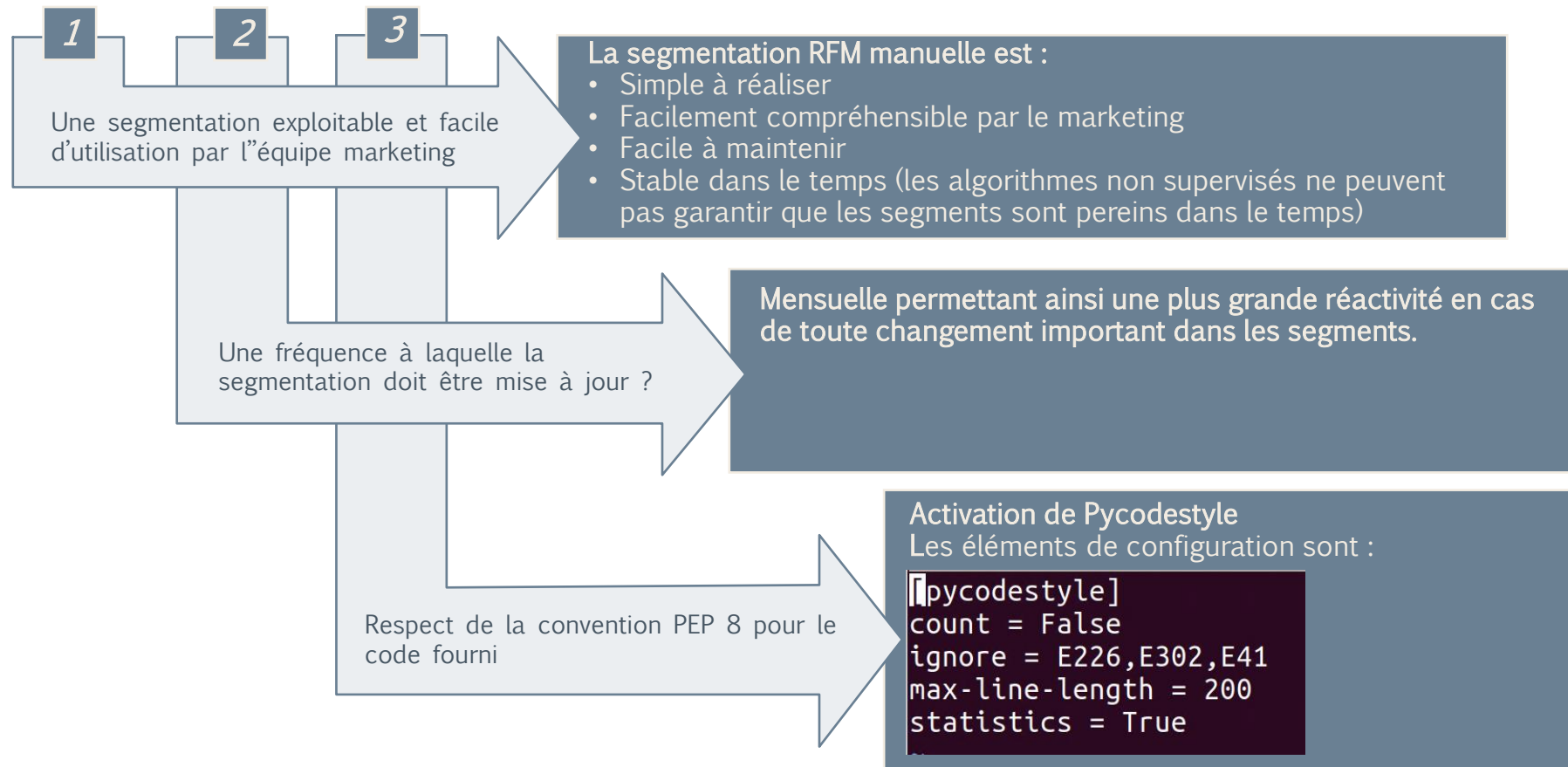
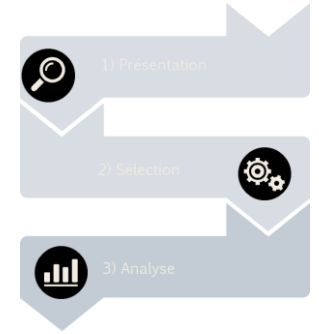
#	Segment	Action Plan
1	Champions	To thank their loyalty, offering them a discount on product categories that they have never purchased and launch a survey to understand why their reviews were bad.
2	Promoters	To thank their loyalty, offering them a discount on product categories that they have never purchased.
3	Potential	High potential to enter our loyal customer segments, why not throw in some gifts on their next purchase to show that you value them!
4	Require Activation	Poorest performers of our RFM model. They might have gone with our competitors for now and will require a different activation strategy to win them back such as lower prices than competitors on the categories of products that they have purchased.

7) Comparatif

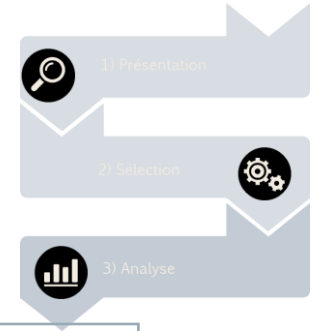
Critère	RFM Standard	RFM DBScan	RFM Kmeans	Kmodes	KPrototypes
Intérêt Métier	 <ul style="list-style-type: none"> Simple à comprendre Connue par les équipes marketing Bonne répartition des clients 	 <ul style="list-style-type: none"> De part son fonctionnement très gourmand, uniquement environ 1% des individus ont pu être traités. 	 <ul style="list-style-type: none"> Segments pertinents Besoin de vulgariser le monde de fonctionnement Mauvaise répartition des clients dans les segments 	 <ul style="list-style-type: none"> A cause de la nature des données où il y a principalement 1 achat par client, la segmentation par produit n'est pas pertinente. Un grand nombre de clients ne sont pas segmentés. 	 <ul style="list-style-type: none"> Segments difficilement compréhensibles Besoin de vulgariser le monde de fonctionnement Mauvaise répartition des clients dans les segments
Maintenance	 <ul style="list-style-type: none"> Simple, integration des nouvelles données Garantie que les segments resteront les mêmes 		 <ul style="list-style-type: none"> Simple, integration des nouvelles données Pas de garantie que les segments resteront les mêmes. 		 <ul style="list-style-type: none"> Simple, integration des nouvelles données Pas de garantie que les segments resteront les mêmes.
Stabilité	 <ul style="list-style-type: none"> L'analyse menée à montrer une certaine stabilité 		 <ul style="list-style-type: none"> Les centroides sont stables 		
Résultats					

Légende :  Excellent  Good  Satisfactory  Poor  Very poor

8) Proposition



9) Prochaines étapes



Mise en œuvre d'un
modèle prédictif
pour les nouveaux
clients

- Compléter la RFM avec un modèle prédictif d'évolution des nouveaux clients sous 6 mois (Gradient Boosting, Random Forest...) afin de pouvoir "anticiper" leur segment et prendre ainsi les mesures adaptées en amont.

Mise en œuvre d'un
modèle prédictif
pour les prospects

- Mettre en oeuvre un modèle prédictif d'évolution des prospects vers le statut de client.

10) Environnement technique

