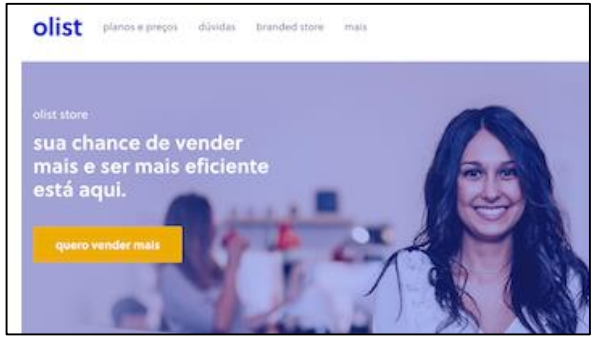
Projet 4 - Segmentez des clients d'un site e-commerce

Aurélien Corroyer-Dulmont, PhD

Ingénieur imagerie médicale

Centre Baclesse
L'excellence pour vaincre votre cancer

Rappel de l’appel à projet



- Problématique :
  - La société OLIST, société d'e-commerce souhaite segmenter ses clients pour mener des campagnes de communications
  - Comprendre les différents types d’utilisateurs grâce à leur comportement et à leurs données personnelles
- Objectif :
  - Fournir à l’équipe marketing une description actionnable de la segmentation
- Données :
  - Base de données comportant des informations sur l’historique de commandes, les produits achetés, les commentaires de satisfaction et la localisation des clients

Nettoyage des données

- Construction d'une base de donnée incluant toutes les informations :
  - *"customers, geolocation,*
  - *items payments,*
  - *reviews,*
  - *orders products,*
  - *sellers,*
  - *category_name"*

Data base unique

customer_unique_id	customer_state	payment_type	review_score_mean	Time_after_last_purchase_days	nb_of_product	moy_achats	Main_prod_category	Tot_achats	Customer_score
0000366f3b9a7992bf8c76cfdf3221e2	SP	credit_card	5.0	160.273507	1	129.90	home	141.90	8
0000b849f77a49e4a4ce2b2a4ca5be3f	SP	credit_card	4.0	163.263090	1	18.90	health_beauty	27.19	8
0000f46a3911fa3c0805444483337064	SC	credit_card	3.0	585.850868	1	69.00	office	86.22	7
0000f6ccb0745a6a4b88665a16c9f078	PA	credit_card	4.0	369.875428	1	25.99	electronics	43.62	9
0004aac84e0df4da2b147fca70cf8255	SP	credit_card	5.0	336.905972	1	180.00	electronics	196.89	11
...	...	...	...	...	...	...	...	...	...
fffcf5a5ff07b0908bd4e2dbc735a684	PE	credit_card	5.0	495.853958	4	1570.00	health_beauty	4134.84	6
fffea47cd6d3cc0a88bd621562a9d061	BA	credit_card	4.0	310.890532	1	64.89	health_beauty	84.58	11
ffff371b4d645b6ecea244b27531430a	MT	credit_card	5.0	617.070162	1	89.90	auto	112.46	7
ffff5962728ec6157033ef9805bacc48	ES	credit_card	5.0	168.092095	1	115.00	fashion	133.69	8
ffffd2657e2aad2907e67c3e9daecbeb	PR	credit_card	5.0	532.883021	1	56.99	health_beauty	71.56	7

- Décompte des variables présentant un nombre de données manquantes trop important :
  - suppression de ces variables si *NaN* > 50 %
- Normalisation des données

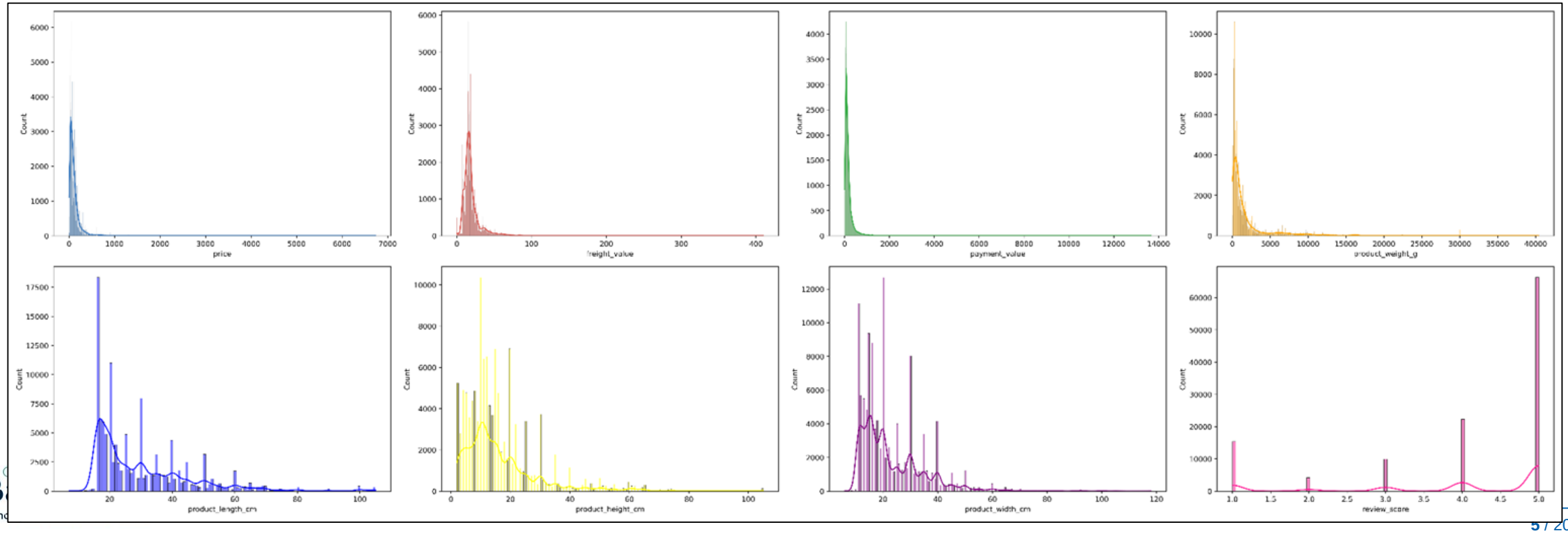
Features engineering

- Informations crées :
  - *« Temps depuis dernier achat »*
    - En utilisant les variables “*order_purchase_timestamp*” et la valeur max de celle-ci
  - *« Délai de livraison »*
    - En utilisant les variables “*order_delivered_customer_date" et "order_purchase_timestamp*”
  - *« Délai de réponse aux commentaires »*
    - En utilisant les variables “*review_answer_timestamp" et "review_creation_date"*
  - *« Nombre de produits, montant moyen et montant total acheté par client »*
    - En utilisant les variables “*customer_id"", “"product_id" et “price”*
  - *« Catégorisation des produits plus succincte afin d’être plus spécifique »*
  - *« Catégorie la plus achetée »*
    - En utilisant les variables “*customer_id"" et “product_category_name_english"*

Centre Baclesse
L'excellence pour vaincre votre cancer

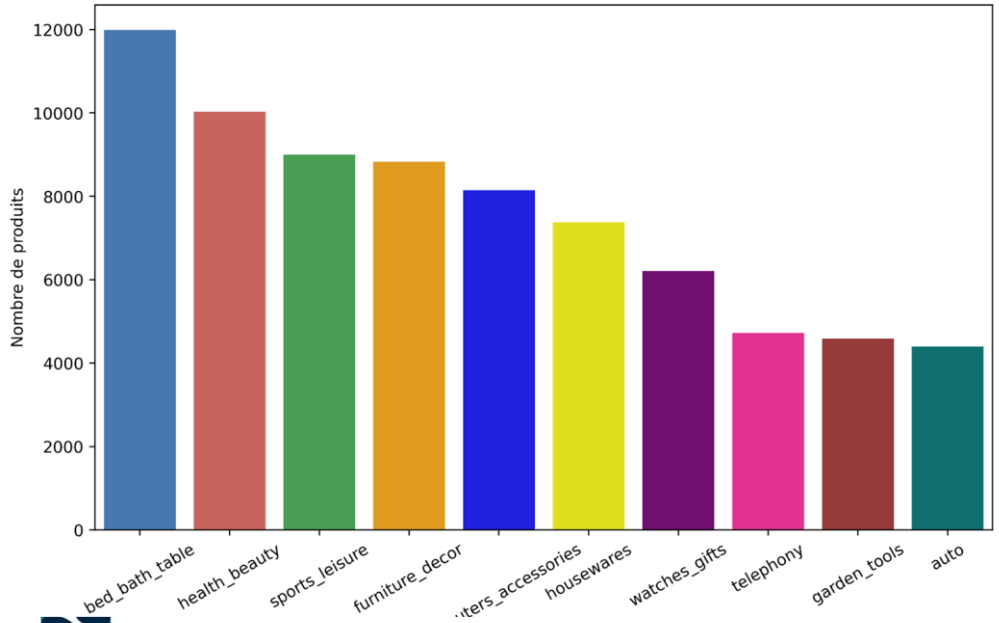
Exploration des données

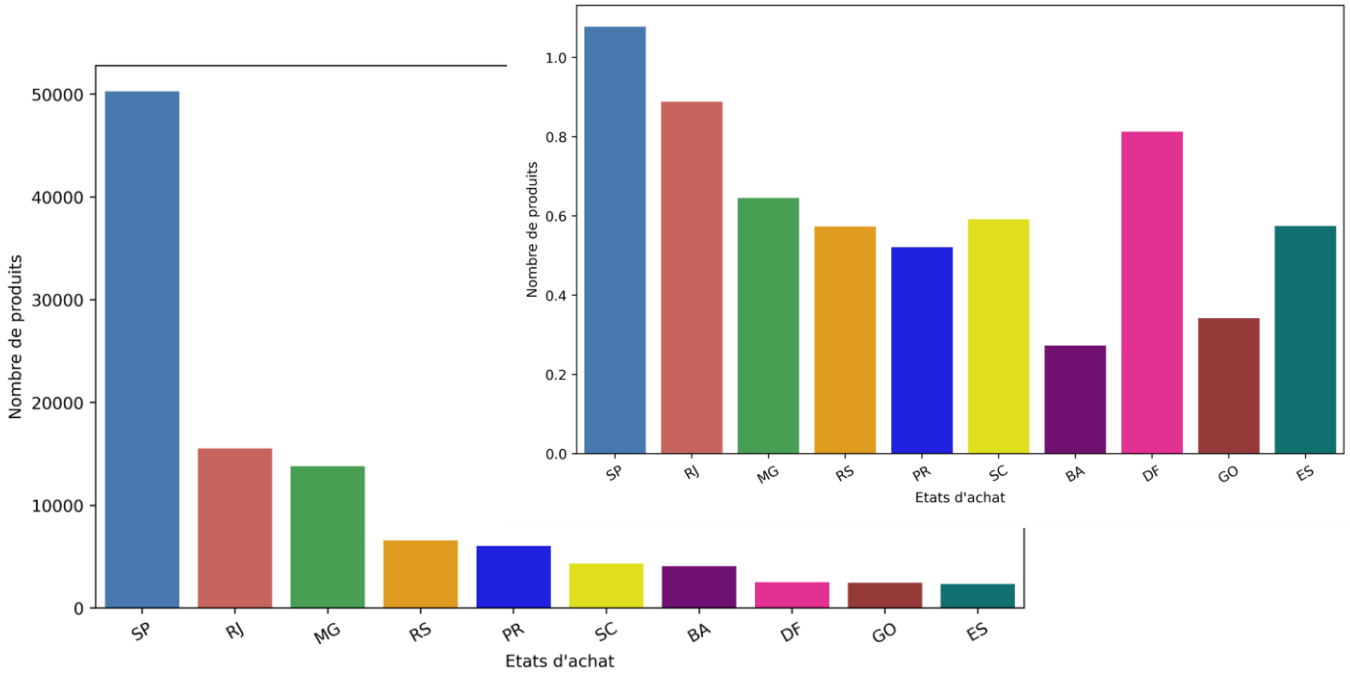
- Exploration globale des variables d'intérêt
  - Les valeurs qui pourraient sembler aberrantes sont bien cohérentes avec la logique, il n'y a pas lieu de faire une sélection des valeurs aberrantes,
  - Seul **3%** des clients sont revenus sur le site, **97% sont des commandes uniques**



Exploration des données

- Exploration globale des variables d'intérêt
  - Les produits de chambre/salle de bain et de beauté sont les plus achetées,
  - L'état de Sao polo et de Rio de Janeiro sont les plus gros acheteurs





Centre Baclesse
L'excellence pour vaincre votre cancer

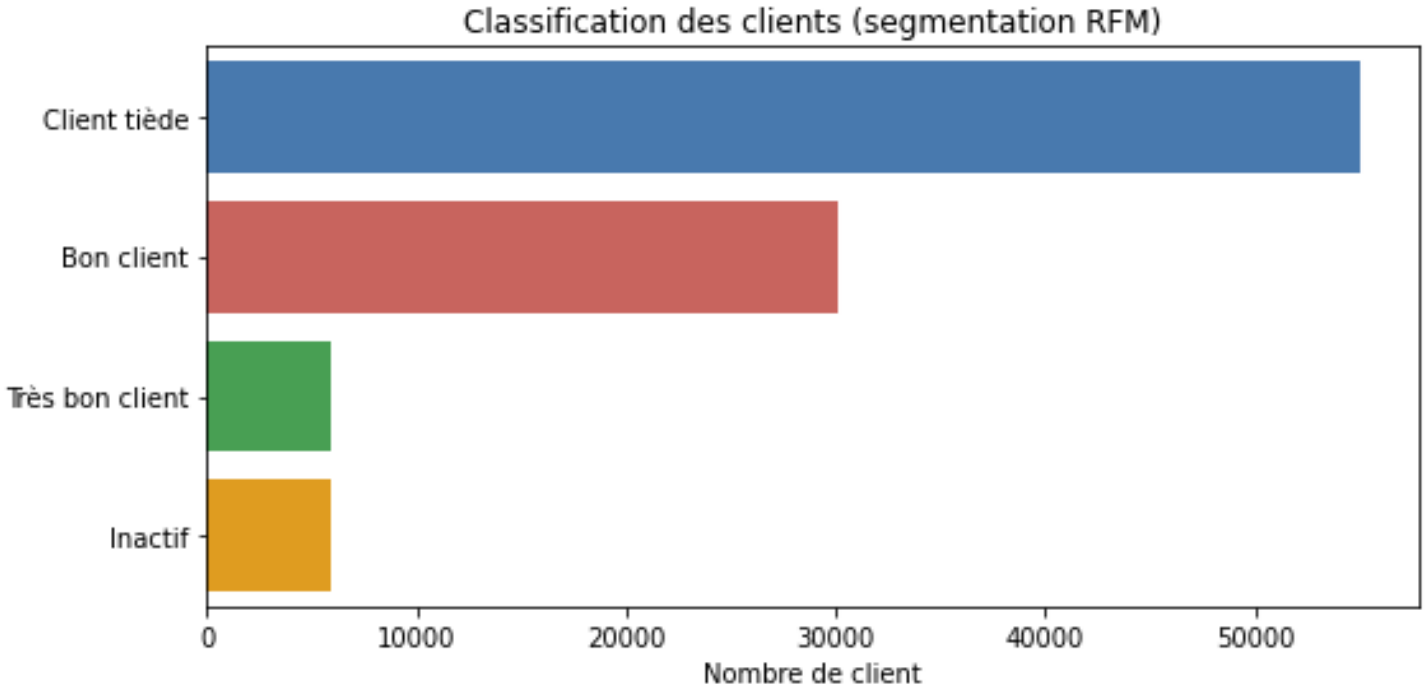
Modélisation

- Modélisation “RFM” :

Segment	Score RFM	Profil	Action
Très bons clients	RFM 13 à 15	Ambassadeurs de la marque, ils consolident l'image et le CA. Ils ont potentiellement une plus forte prévision de commandes.	Action de fidélisation : cartes premium, invitations...
Bons clients	RFM 7 à 12	Achats récents et fréquents. La confiance existe et doit être consolidée. Taux de renouvellement potentiel de commandes inférieur à celui des très bons clients.	Action de développement de la fréquence d'achat ou du montant par commande, couplé avec des actions de fidélisation, offre couplée...
Clients tièdes	RFM 1 à 6	Clients encore volatils hésitant à acheter dans un court délai à la concurrence. La certitude qu'ils rachètent prochainement est faible.	Action de fidélisation par des offres spécifiques et récurrentes : réductions, promotion sur certaines offres...
Nouveaux clients	RFM 8	Premier achat, l'analyse doit se poursuivre lors des prochaines consommations suite au réengagement.	Action de fidélisation (à la fin de leur engagement) et de développement de leurs achats.
Inactifs	RFM 0	Aucune consommation durant la dernière année.	Action de réactivation du client ou abandon.

Source:“https://www,e-marketing,fr/Thematique/academie-1078/fiche-outils-10154/scoring-RFM-306775,htm”

- Recency « Temps depuis dernier achat »
- Frequency « Nombre de produits »
- Monetary « Montant total »



Centre Baclesse
L'excellence pour vaincre votre cancer

Modélisation

- Choix des modèles de classification étudiés :
  - Comprendre le comportement des variables :
    - Principal Components Analysis()
  - Classification des clients :
    - **KMeans()**
    - **DBScan()**
    - **Cluster hiérarchique()**

Centre Baclesse
L'excellence pour vaincre votre cancer

Modélisation

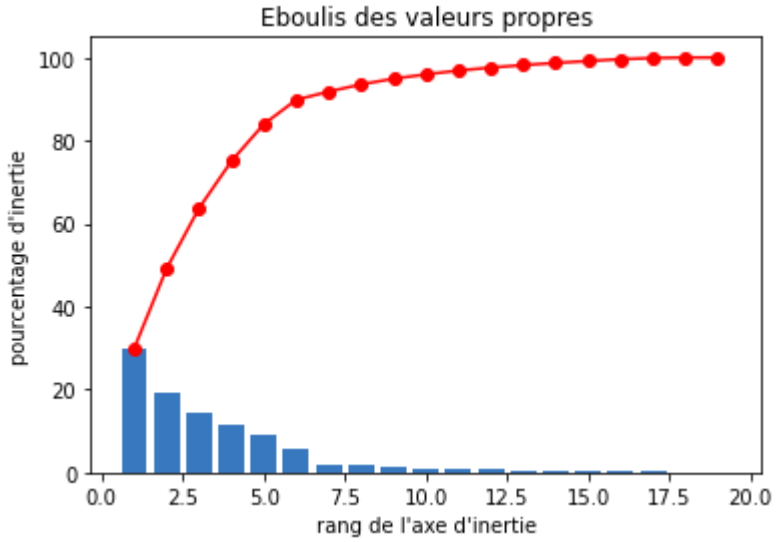
- Features utilisés pour l'entrainement des modèles :

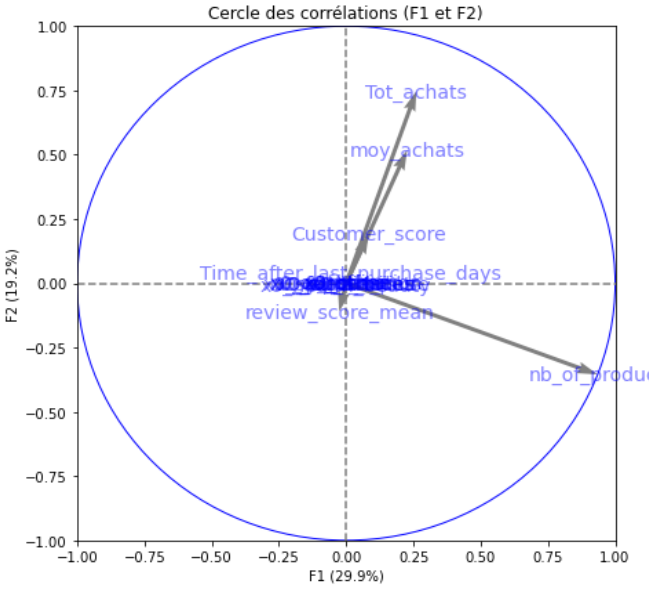
  - Features **quantitatifs** :
    - review_score_mean
    - Time_after_last_purchase_days
    - nb_of_product
    - moy_achats
    - Tot_achats
  - Features **catégoriels** :
    - customer_state
    - payment_type
    - Main_prod_category
    - RFM

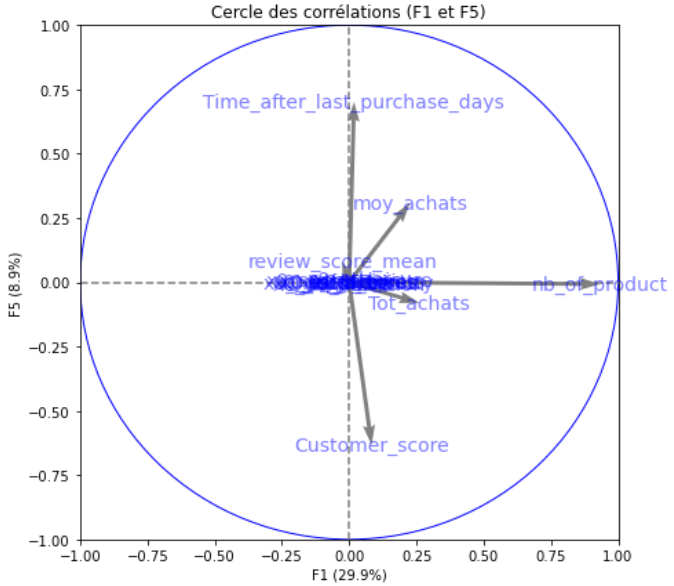
Centre Baclesse
L'excellence pour vaincre votre cancer

Comprendre le comportement des variables :

- Analyse en composantes principales :





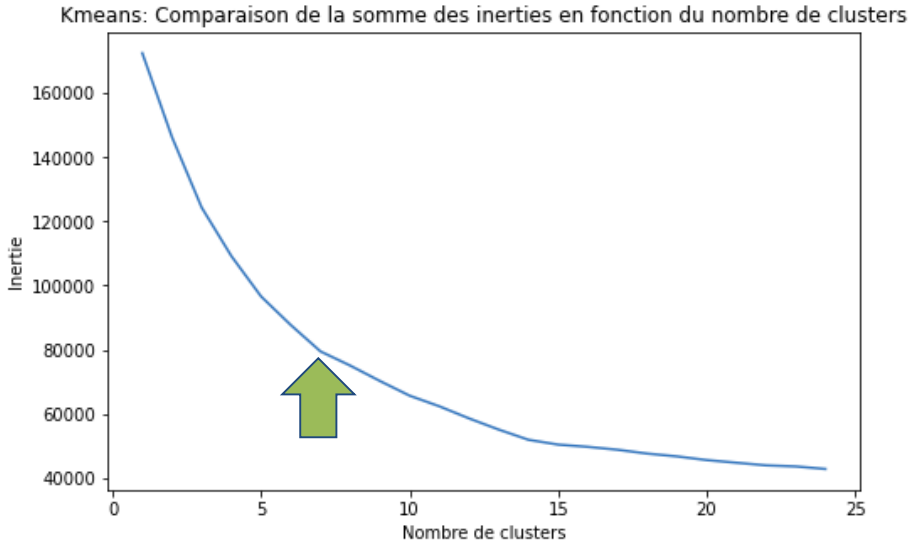


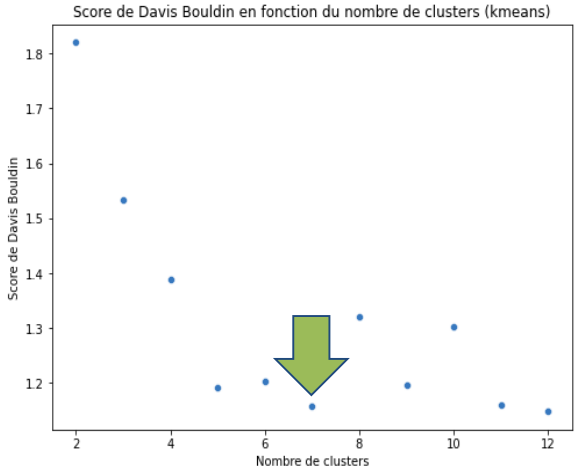
La PCA avec 7 composantes expliques assez bien les différents caractéristiques du jeu de donnée

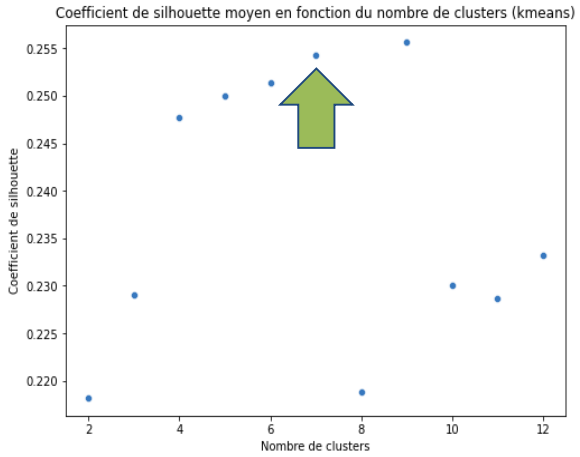
Centre Baclesse
L'excellence pour vaincre votre cancer

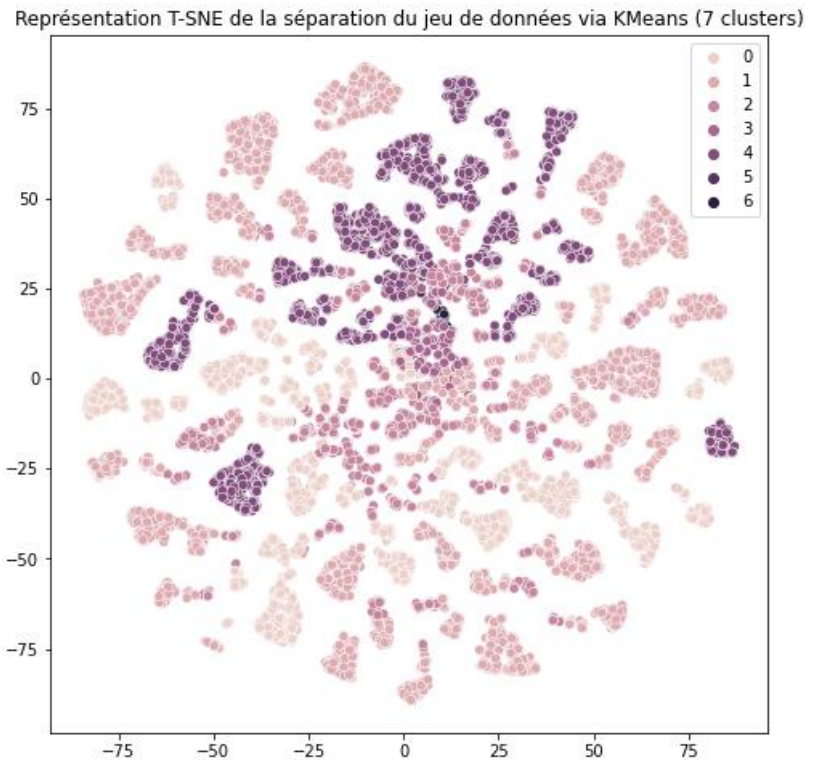
Classification des clients

- Classification KMeans :





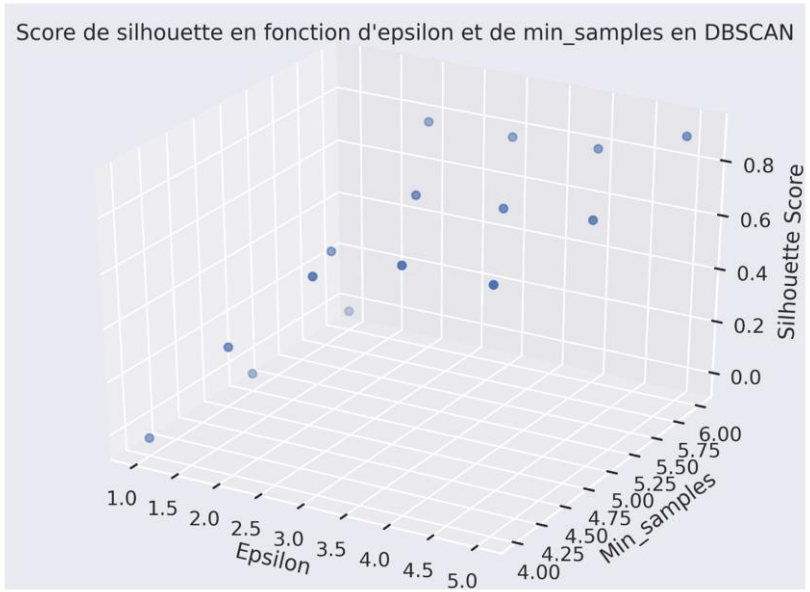


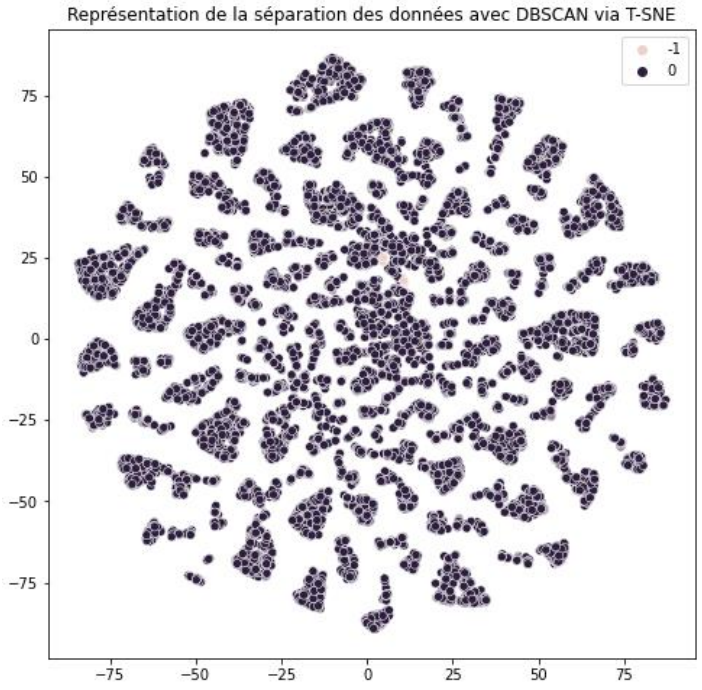


Le nombre de cluster idéal pour le KMeans semble être de 7

Classification des clients

- Classification DBSCAN :



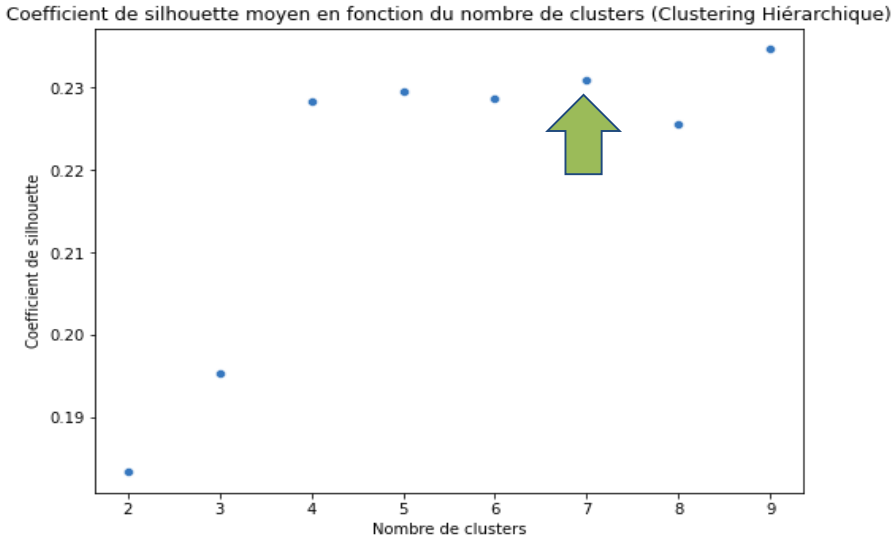


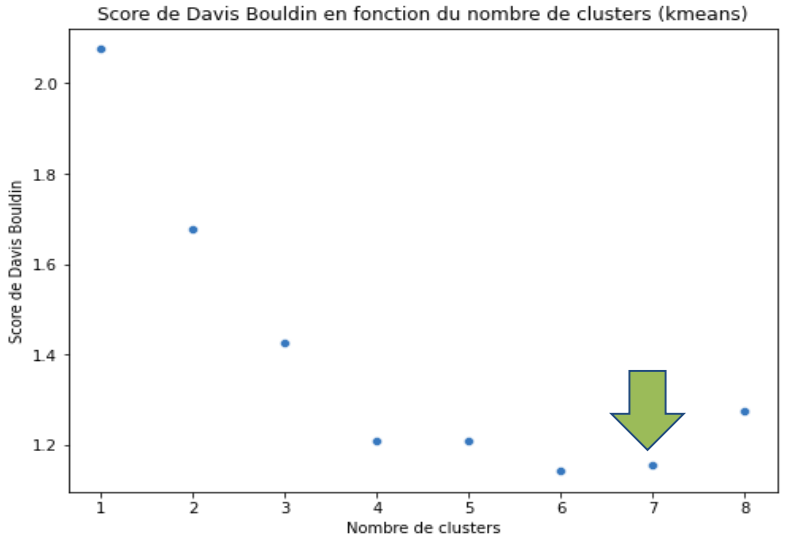
Le DBSCAN ne semble pas être adapté aux données, car malgrès un score de silhouette très intéressant il ne propose qu'une séparation par deux clusters, la densité des données doit être trop similaire,

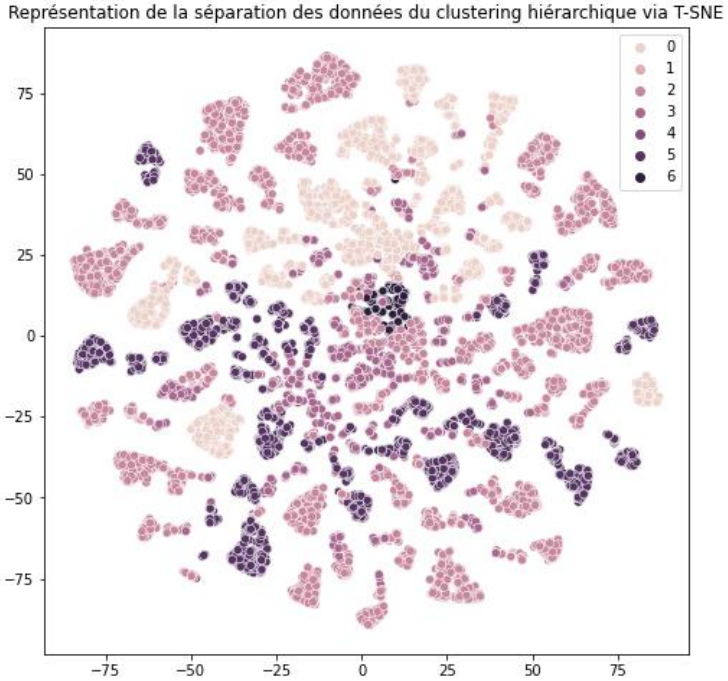
Centre Baclesse
L'excellence pour vaincre votre cancer

Classification des clients

- Clustering hiérarchique :







Le nombre de cluster idéal pour le Clustering hiérarchique semble être de 7

Centre Baclesse
L'excellence pour vaincre votre cancer

Classification des clients

- Résultats des classifications (KMeans) :

customer_unique_id	customer_state
e7902094a37e9f2e51796	CE
259287bcd5b5a766adf2a	SP
376177c2b40499c881c84	BA
8c6886e88a58e1148c7ea	MG
98c42e572c43e516b6552	RJ
...	...
3c0dac3b48c8dd46259e	SP
7df51447510ede80822ec	RJ
eea3452bb0ff82082da8f8	RJ
af66702aa8acdd82ffb55	SE
b6ab611d0cdd89184f39f	GO

...

Clustering_hiérarchique	Clustering_Kmeans
2	1
5	0
5	0
2	1
5	0
...	...
2	1
2	1
2	1
2	1
0	4

	review_score_mean	Time_after_last_purchase_days	nb_of_product	moy_achats	Tot_achats
Label 1	4,09	285,78	2,29	150,83	239,01
Label 2	4,08	285,98	2,25	142,60	228,39
Label 3	4,10	286,81	2,50	144,00	213,92
Label 4	4,05	287,67	1,91	141,94	216,94
Label 5	4,12	283,08	2,32	145,17	231,30
Label 6	4,08	263,65	3,11	93,97	225,62
Label 7	4,20	354,96	2,89	172,18	333,50

Création d'une database avec le n° de clustering par client

Classification des clients

- Résultats des classifications (Clustering hiérarchique) :

customer_unique_id	customer_state
e7902094a37e9f2e51796	CE
259287bcd5b5a766adf2a	SP
376177c2b40499c881c84	BA
8c6886e88a58e1148c7ea	MG
98c42e572c43e516b6552	RJ
...	...
3c0dac3b48c8dd46259e	SP
7df51447510ede80822ec	RJ
eea3452bb0ff82082da8f8	RJ
af66702aa8acdd82ffb55	SE
b6ab611d0cdd89184f39f	GO

...

Clustering_hiérarchique	Clustering_Kmeans
2	1
5	0
5	0
2	1
5	0
...	...
2	1
2	1
2	1
2	1
0	4

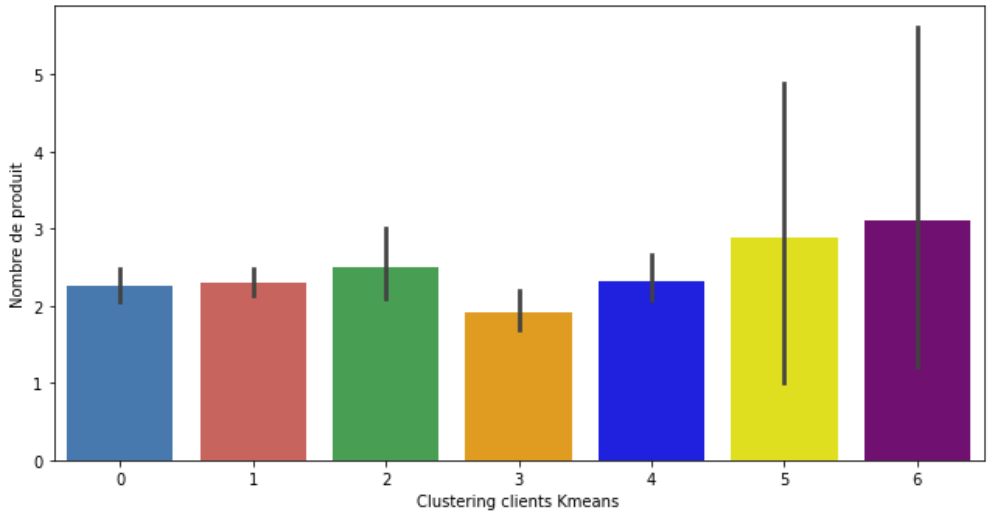
	Review_score	Time_after_last_purchase_days	Nb_of_product	moy_achats	Tot_achats
Label 1	4,09	286,13	2,30	150,75	238,34
Label 2	4,08	286,43	2,29	142,85	230,11
Label 3	4,10	284,56	2,26	143,21	211,90
Label 4	4,04	293,31	1,82	149,25	239,75
Label 5	4,12	283,22	2,41	144,07	228,21
Label 6	4,15	274,11	3,32	84,46	222,09
Label 7	3,97	303,42	2,00	103,75	218,22

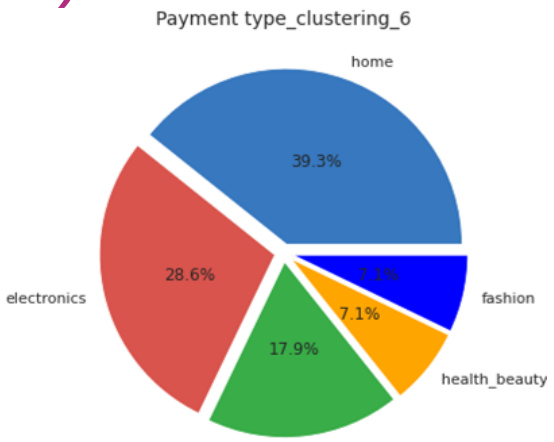
Création d'une database avec le n° de clustering par client

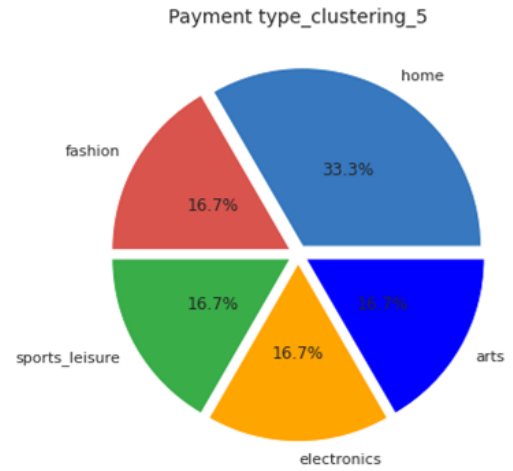
Centre Baclesse
L'excellence pour vaincre votre cancer

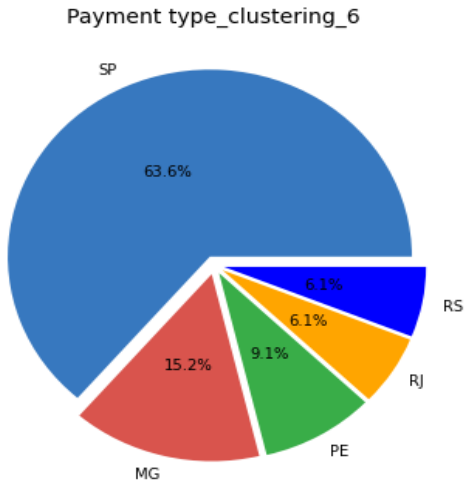
Classification des clients

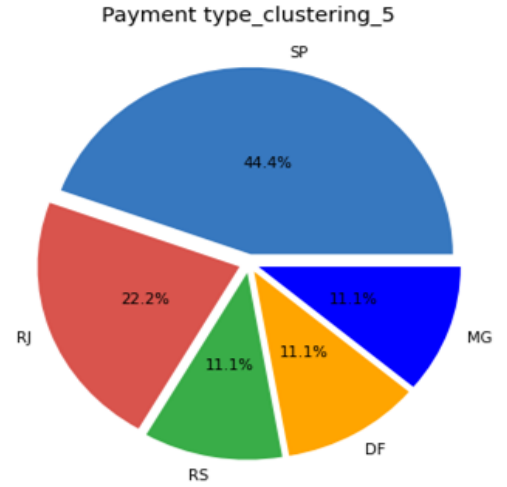
- Résultats des classifications (KMeans) :











La segmentation par KMeans semble bien discriminer des populations de client

Classification des clients

- Résultats des classifications (KMeans) :

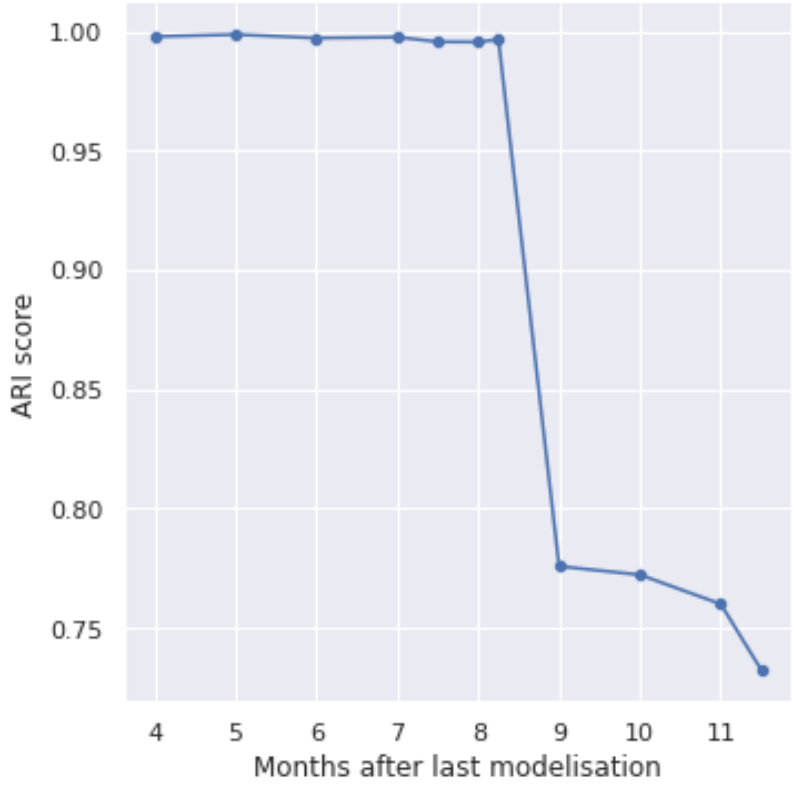
	Nb de client	type de client (1)	type de client (2)	Classe client	recommandation pour ce client	catégorie la plus acheté	région préférentielle	RFM score	score avis	montant moyen	nb produit	montant total	temps depuis dernier achat
Label 1	21956	achète un panier moyen assez élevé	Client moyen / bon	**	Action de fidélisation par des offres ciblées	home, élec, health	SP, RJ, MG	5,93	4,09	150,83	2,29	239,01	285,78
Label 2	43390	Plus grosse cohorte de client	Client moyen	**	Actions de développement du montant de la commande. Offres ciblées “santé”	home, health, élec	SP, RJ, MG	5,91	4,08	142,60	2,25	228,39	285,98
Label 3	15801	achète peu en montant total	Mauvais client	*	Offres ciblées “santé”	home, health, élec	SP, RJ, MG	5,95	4,10	144,00	2,50	213,92	286,81
Label 4	578	achète en moins et montant faible, score avis moins bon que les autres	Mauvais client	*	représente faible effectif, l'oublier	home, élec, health	SP, RJ, MG	5,93	4,05	141,94	1,91	216,94	287,67
Label 5	5029	score avis meilleur que les autres	Client moyen	**	Offres ciblées “home”	home, élec, health	SP, RJ, MG	5,95	4,12	145,17	2,32	231,30	283,08
Label 6	38	achète beaucoup en qté mais peu au total et souvent	Client moyen	**	pub pour articles de sport	home, sport, élec	SP, RJ, RS	5,82	4,08	93,97	3,11	225,62	263,65
Label 7	9	achète beaucoup, note bien mais a acheté il y a longtemps	Très bon client	***	Action de fidélisation, carte premium	home, élec, appliances	SP, MG, PE	6,00	4,20	172,18	2,89	333,50	354,96

Délai de maintenance

- Simulation efficacité des modèles en fonction du temps :

  - Données disponibles sur 24 mois
  - Segmentation des périodes par mois
  - Etude de la performance du modèle en soustrayant un mois à chaque itération

  => Utilisation du score ARI (*Adjusted Rand Index*)



Une maintenance tous les **8 mois** semble optimale, en effet au delà le score ARI diminue drastiquement

Centre Baclesse
L'excellence pour vaincre votre cancer

CONCLUSION

- Rappel de la problématique :
  - Segmenter des clients d'une société de e-commerce pour mener des campagnes de communications
- Résultats :
  - Une classification par RFM permet une première segmentation des clients pour optimiser les modèles de segmentation
  - Les modèles KMeans et clustering hiérarchique donnent des résultats intéressants
  - La segmentation par **KMeans** est cependant meilleure et semble permettre de proposer des campagnes de communications spécifiques pour 7 catégories de client
  - Pour une meilleure efficacité du modèle et de prise en compte des nouveaux clients, une maintenance tous les 8 mois est préférable

Centre Baclesse
L'excellence pour vaincre votre cancer

Projet 4 - Segmentez des clients d'un site e-commerce

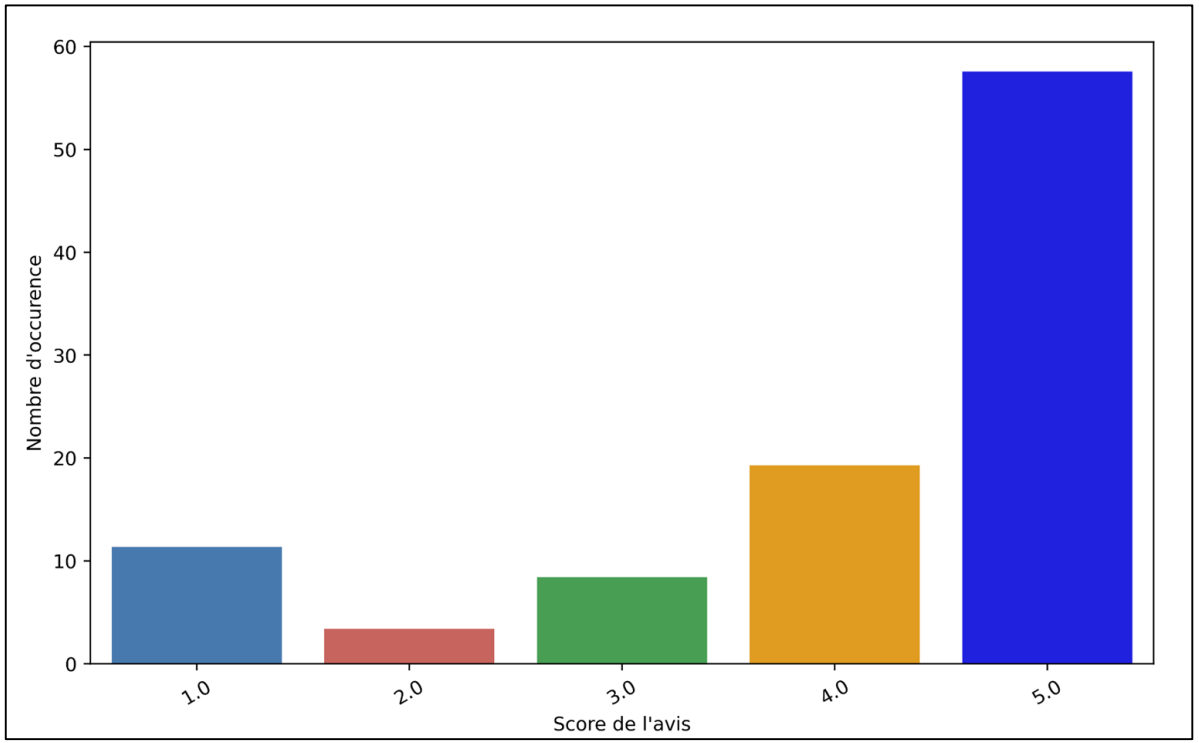
Aurélien Corroyer-Dulmont, PhD

Ingénieur imagerie médicale

Centre Baclesse
L'excellence pour vaincre votre cancer

Exploration des données

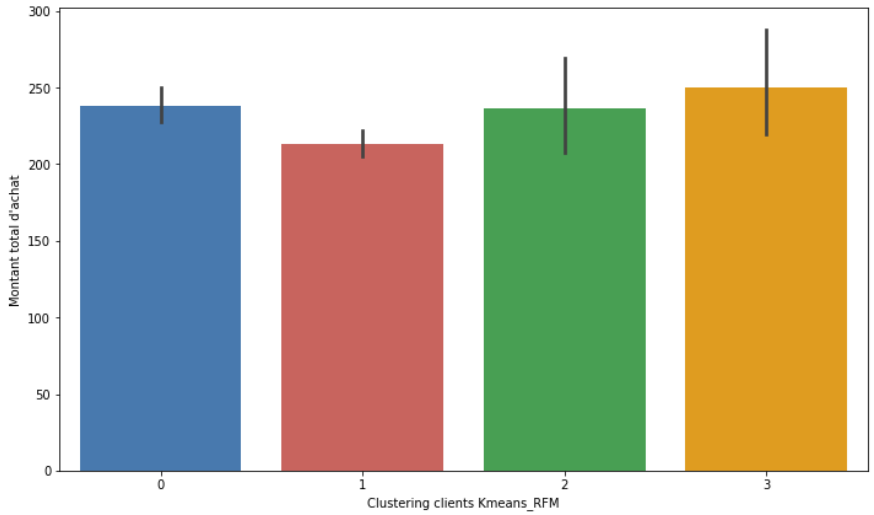
- Exploration globale des variables d'intérêt
  - Il y a globalement beaucoup plus de vote à 5 (presque 60%) en revanche il y a plus de 10% des votes qui sont pour le plus faible score, c'est un point important,

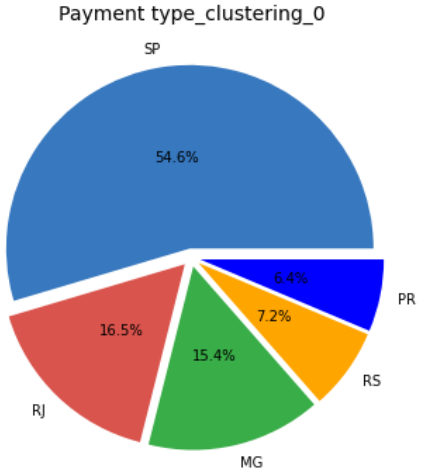


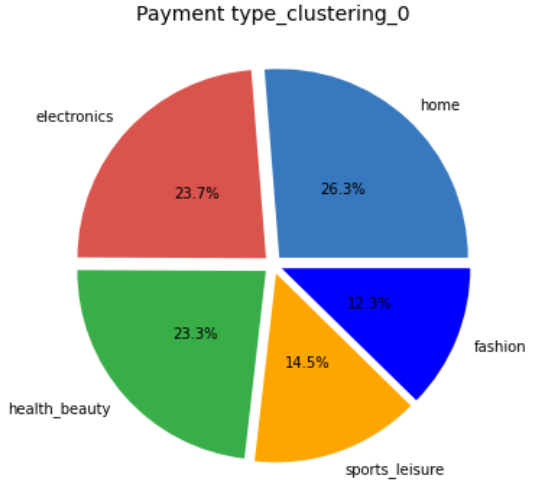
Centre Baclesse
L'excellence pour vaincre votre cancer

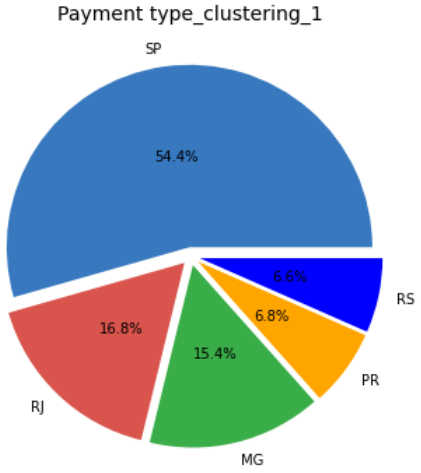
Classification des clients

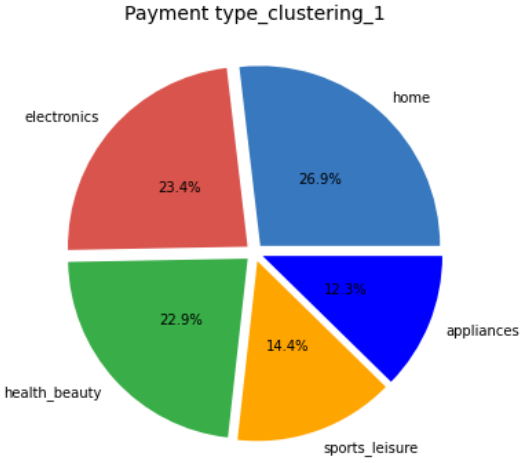
- Classification KMeans **utilisant uniquement les données RFM** :











Centre Baclesse
L'excellence pour vaincre votre cancer