

Formation Data Scientist

Projet 5

« Segmentation des clients d'un site e-commerce »

Soutenance de Projet
Samir HINOJOSA

27 octobre 2021

olist



OPENCLASSROOMS

Plan de la soutenance

1. Mission
2. Présentation du jeu de données
3. Analyse exploratoire
4. Modélisations effectuées
5. Modèle sélectionné
6. Conclusion



1.

Mission



Mission

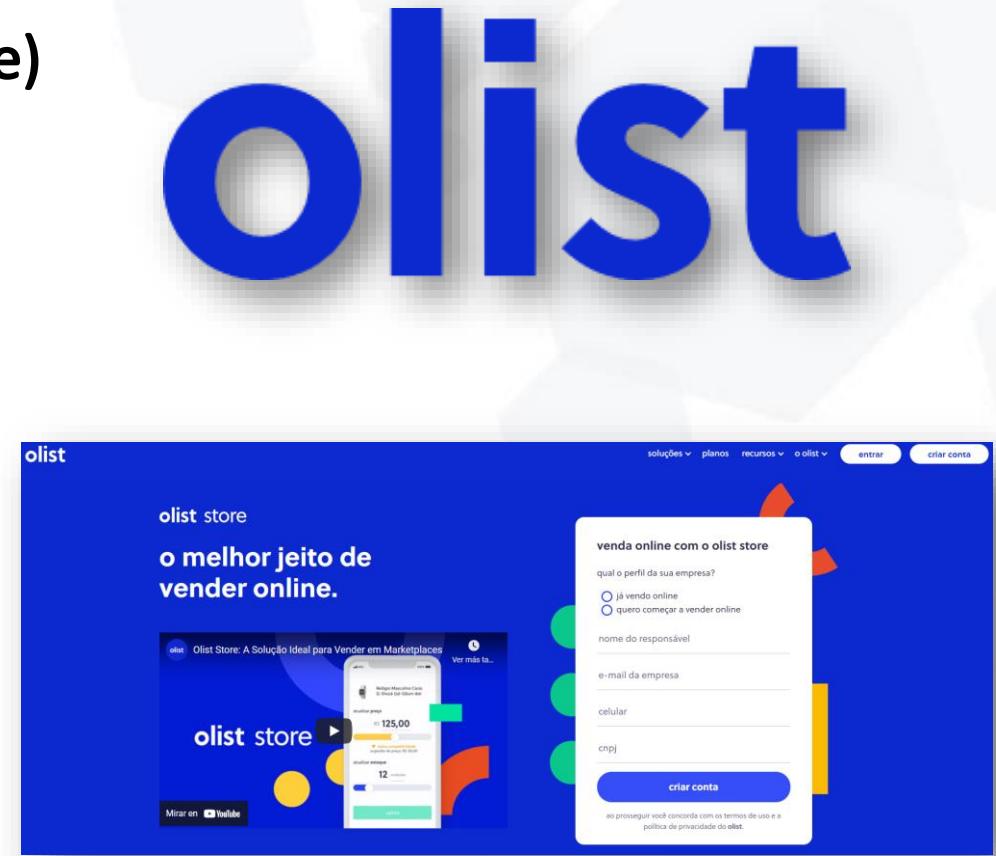
Olist (**solution de vente sur les marketplaces en ligne**) souhaite fournir à ses **équipes d'e-commerce** une **segmentation des clients** pour leurs campagnes de communication.



Fournir à l'équipe marketing une description actionnable de la segmentation pour une utilisation optimale

Prendre en compte :

- La segmentation proposée doit être exploitable et facile d'utilisation pour l'équipe marketing.
- Evaluer la fréquence à laquelle la segmentation doit être mise à jour, afin de pouvoir effectuer un devis de contrat de maintenance.

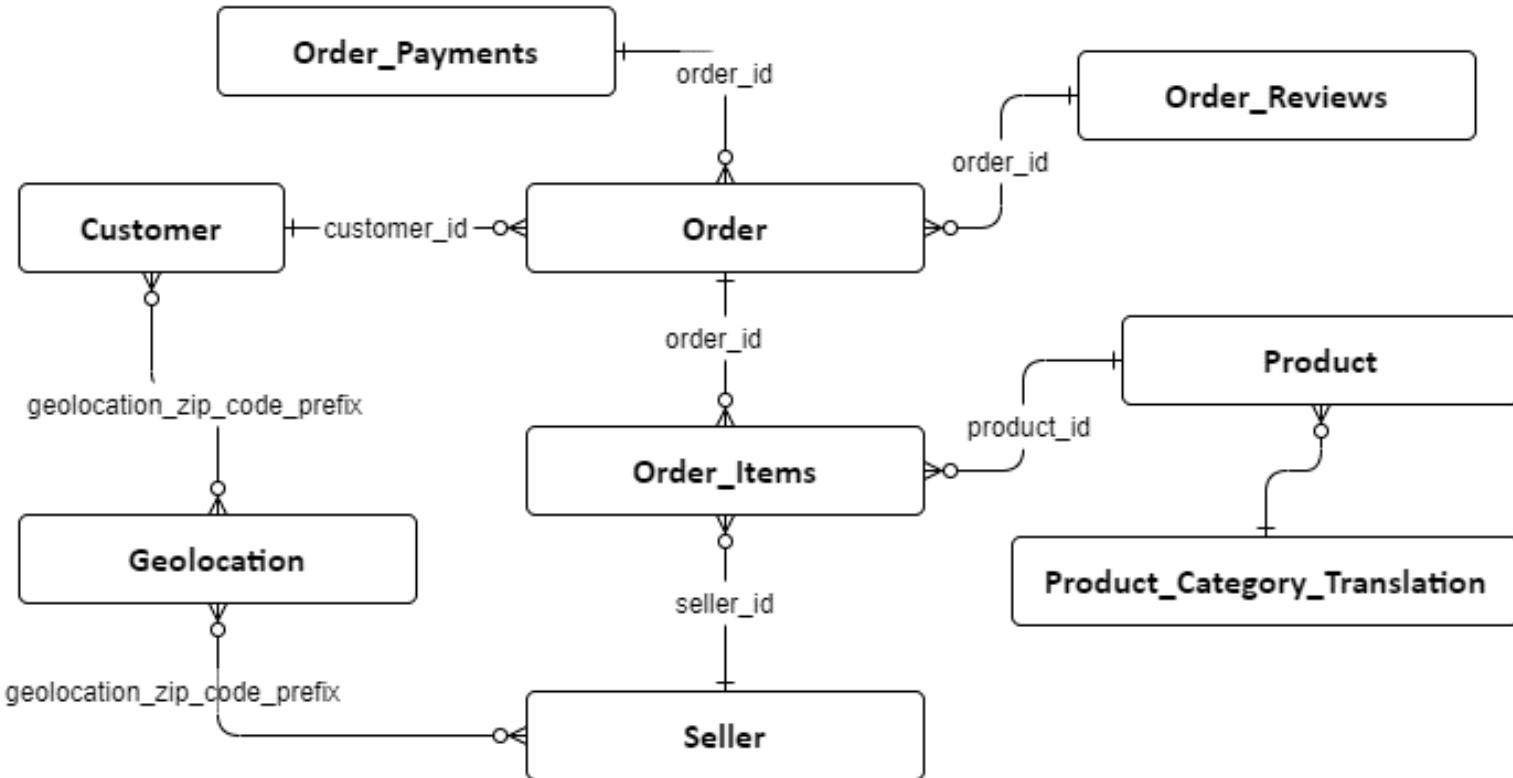


2.

Présentation du jeu de données



Présentation du jeu de données



Il y a des « relations » entre les datasets à travers les colonnes

- **order_id**
- **customer_id**
- **product_id**
- **seller_id**
- **geolocation_zip_code_prefix**

Présentation du jeu de données

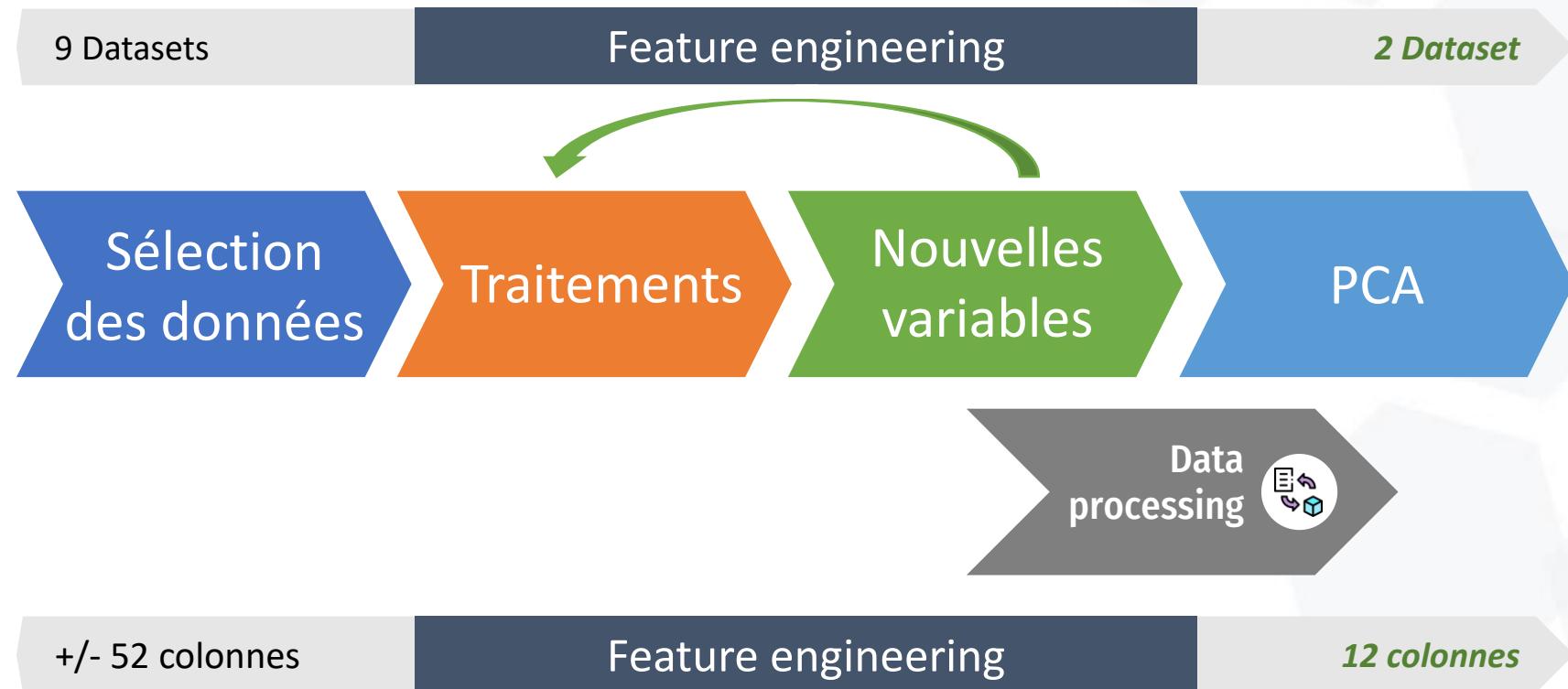
orders_dataset.csv	Il s'agit de l'ensemble de données de base.					
	Taille:	99441x4	Pourcentage de NaN:	0 %	Doublon:	0
customers_dataset.csv	Informations sur le client et son emplacement.					
	Taille:	99441x5	Pourcentage de NaN:	0 %	Doublon:	0
order_reviews_dataset.csv	Informations relatives sur les avis des clients.					
	Taille:	100000x7	Pourcentage de NaN:	20.93 %	Doublon:	0
order_items_dataset.csv	Comprend des données sur les articles achetés dans chaque commande.					
	Taille:	112650x7	Pourcentage de NaN:	0 %	Doublon:	0
products_dataset.csv	Contient des données sur les produits vendus par Olist.					
	Taille:	32951x9	Pourcentage de NaN:	0,83 %	Doublon:	0

3.

Analyse exploratoire



Analyse exploratoire / Feature engineering



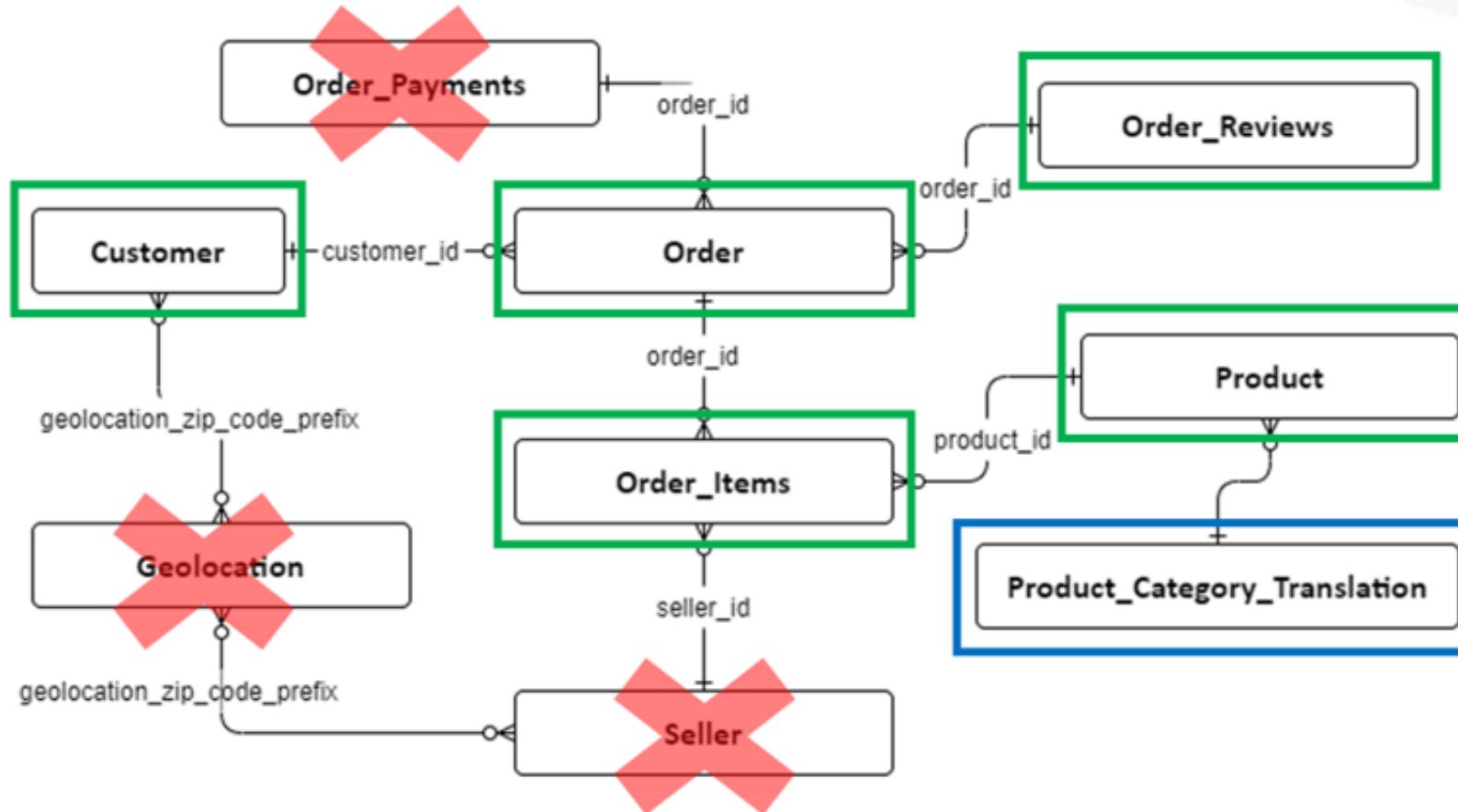
Sélection des données sur la base de customer-focus

 Segmentation comportementale

 Segmentation de la valeur



Sélection des données sur la base de customer-focus



Les colonnes à travailler sont:

- `order_id`
- `customer_id`
- `order_status`
- `order_purchase_timestamp`
- `review_score`
- `product_category`
- `payment_value`
- `product_width_cm`
- `product_height_cm`
- `product_length_cm`
- `product_weight_g`
- `customer_city`
- `customer_state`
- `customer_region`

Aucune variable n'a de distribution normale



Après avoir fait :

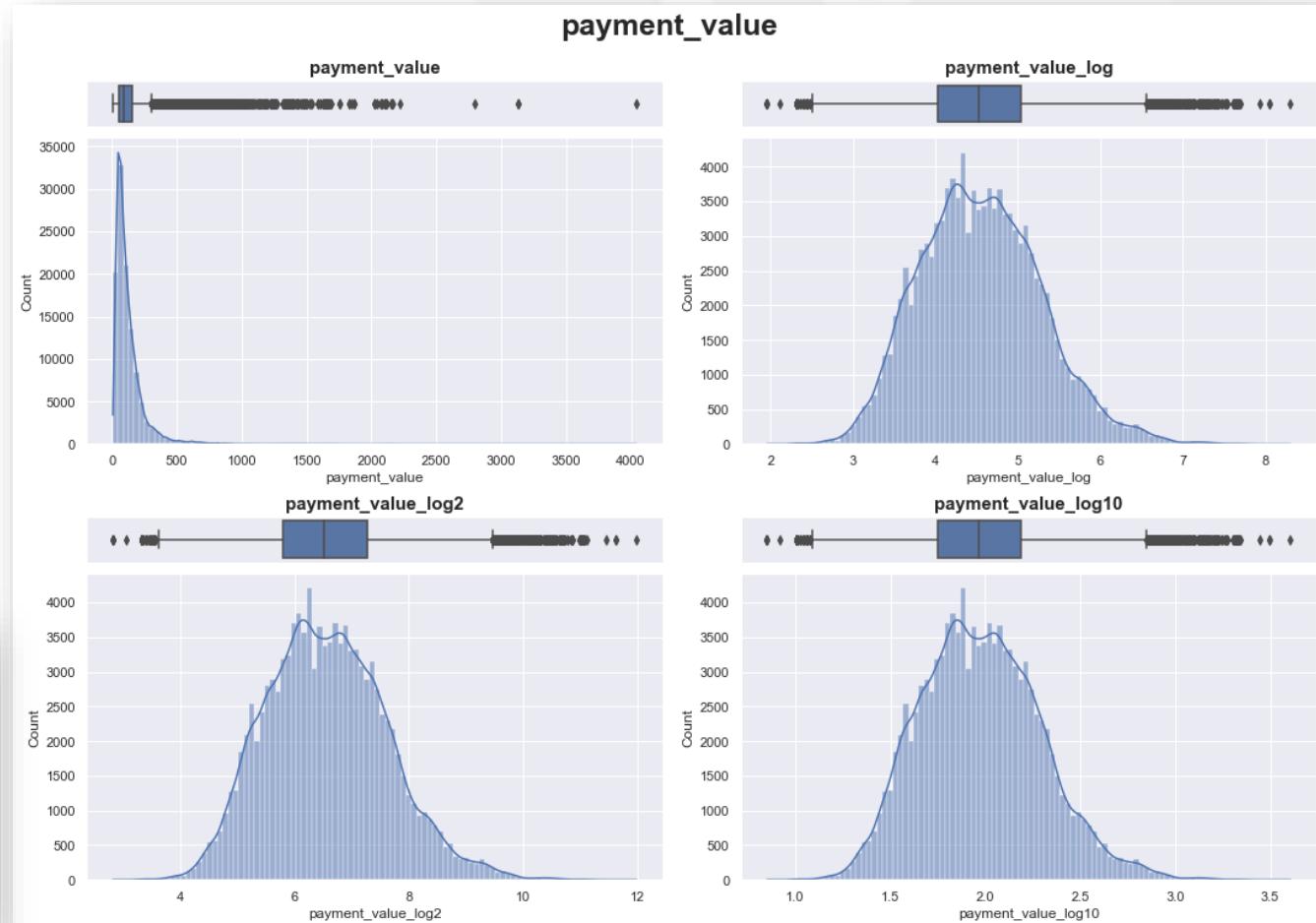
- Le traitement des valeurs aberrantes



Passage au log

- payment_value
- product_length_cm
- product_weight_g
- product_height_cm
- product_width_cm

Variance of payment_value	
-----	variance
payment_value_log	0.534517
payment_value_log2	1.112528
payment_value_log10	0.100816



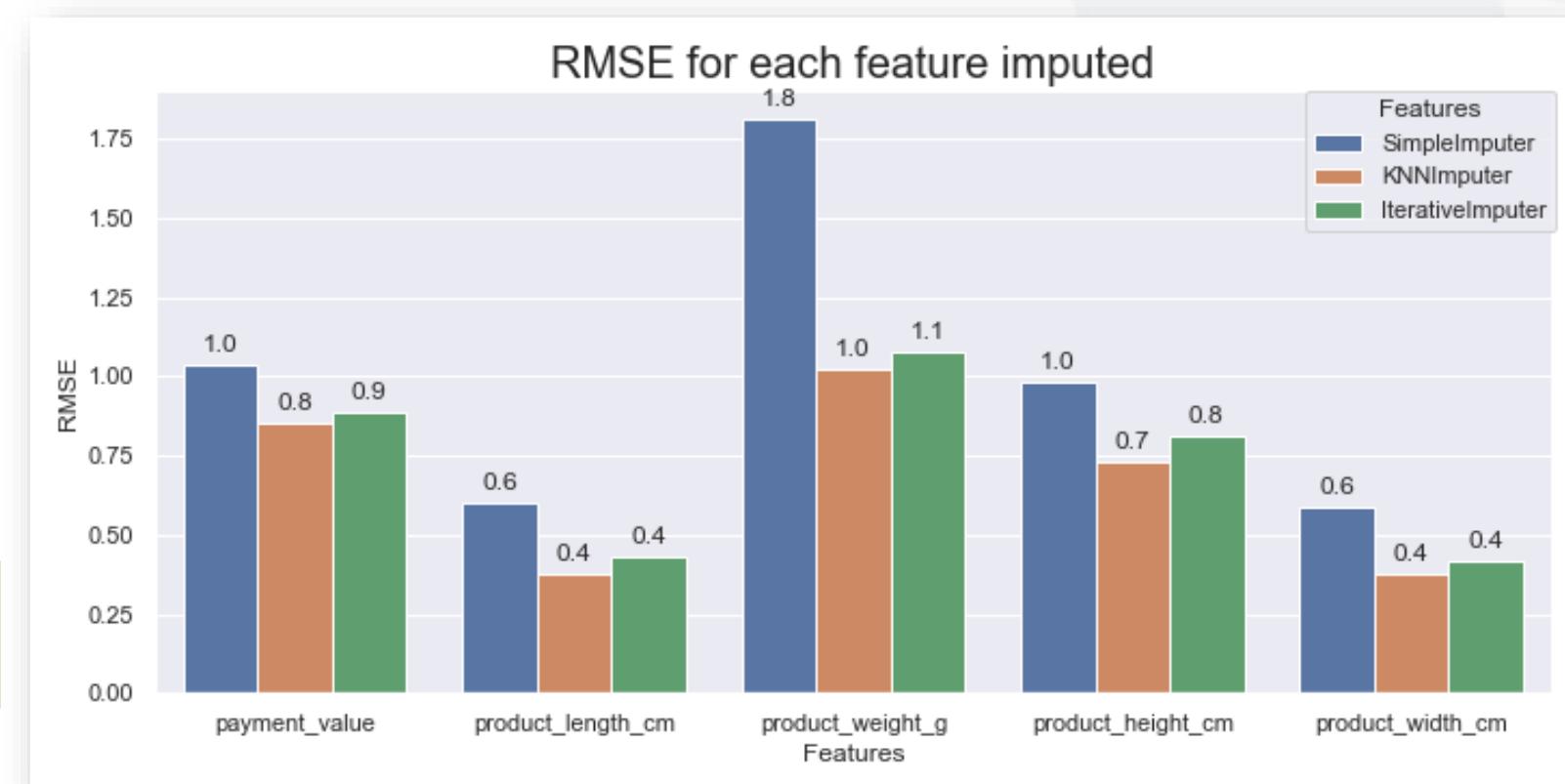
KNN pour les valeurs manquantes



Après avoir calculé **MSE** et **RMSE**, **KNN Imputer** a obtenu le meilleur résultat

À cette phase, la taille du dataset est de 113425 x 14 et 0,83 % de valeurs manquantes

la méthode choisie pour effectuer les imputations sera ***KNN Imputer***



Réduction de catégories



Après avoir fait :

- L'inverse du logarithme
- La suppressions des commandes « unavailable » et « canceled »

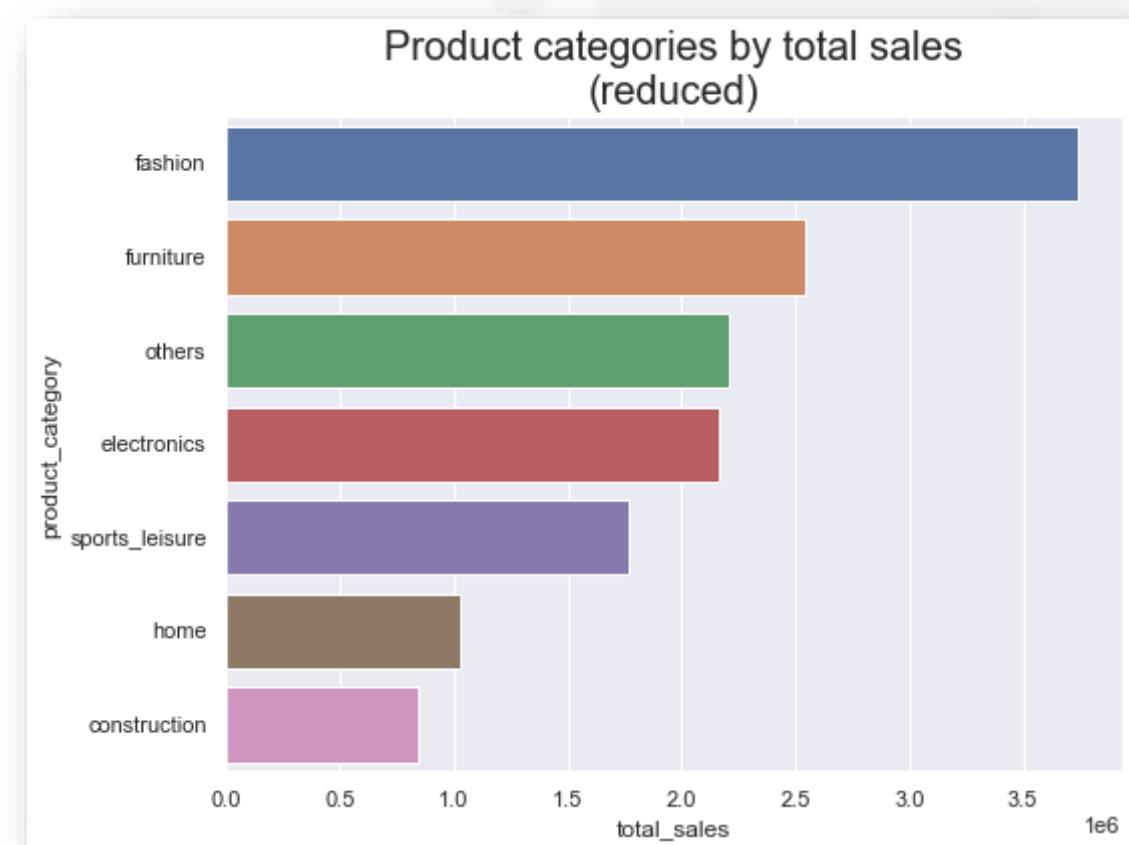
À cette phase, la taille du dataset est de 112109 x 14 avec 98207 clients



product_category

- L'identification des catégories avec le plus de ventes
- Les 30 mots les plus communs dans toutes les catégories
- Regroupement des catégories à travers une phase d'observation
- Transformation de catégories à variables avec pivot_table
- Calcul du poids basé sur le total des achats de chaque client

Depuis 74 catégories à seulement 11



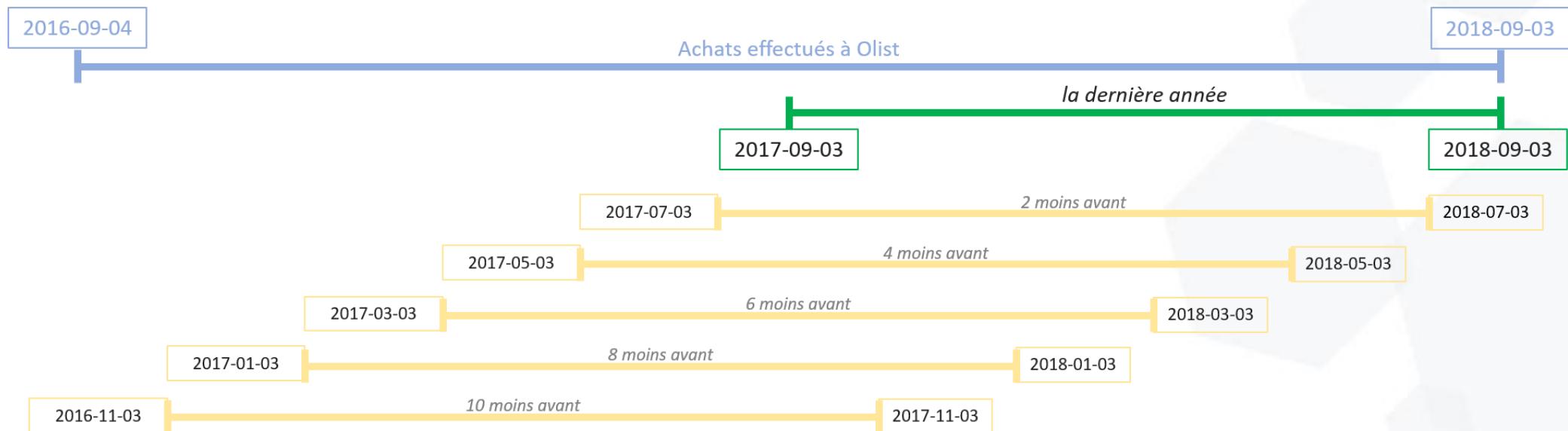
Filtration d'information par dates



La filtration de l'information pour la dernière année

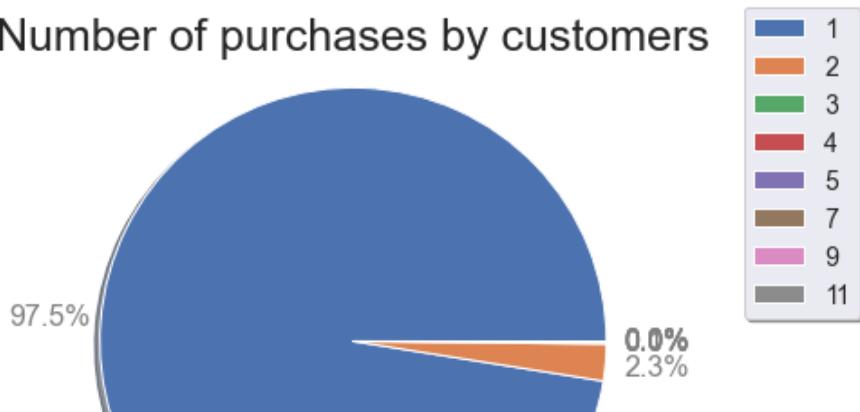


Des datasets décalés de 2 mois par rapport à la dernière année, ont été utilisés pour mesurer la stabilité des clusters

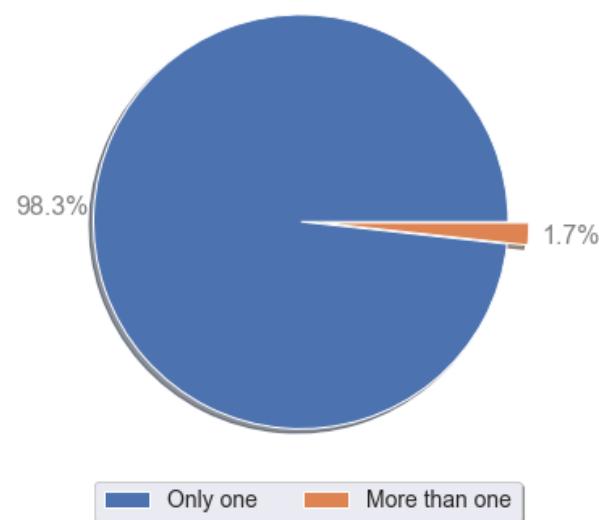


La plupart des clients ont acheté une seule fois dans la dernière année

Number of purchases by customers



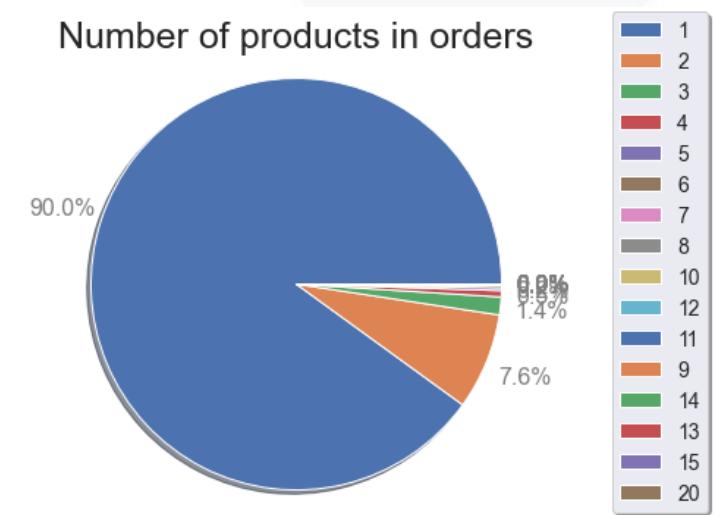
Customers who bought more than one category



À cette phase, la taille du dataset est de 85870 x 14

- 73075 clients uniques
- 75081 commandes uniques

Number of products in orders



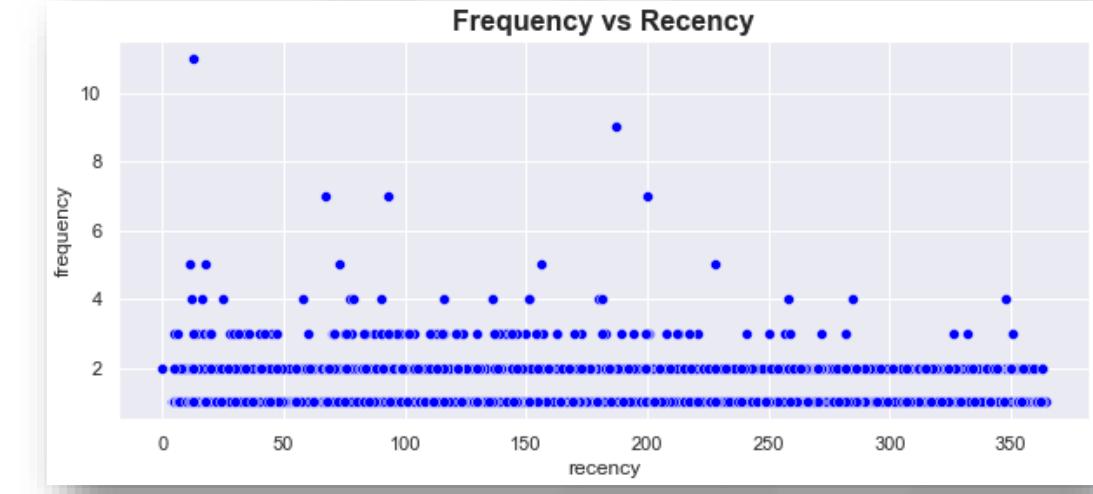
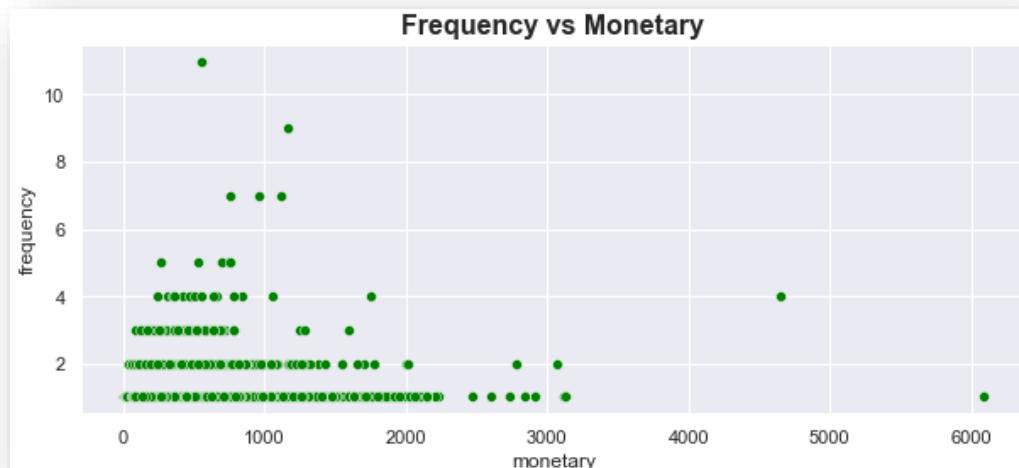
Segmentation RFM

Faible somme d'argent et fréquence



Frequency vs Monetary:

les achats réguliers sont pour une petite somme d'argent



Frequency vs Recency:

la fréquence d'achat est faible par rapport à récemment

Segmentation RFM



Des segments selon *Exponea*¹

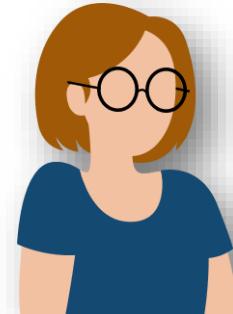
Des segments	Scores (RFM)
Champion	555, 554, 544, 545, 454, 455, 445
Fidèle	543, 444, 435, 355, 354, 345, 344, 335
Loyal potentiel	553, 551, 552, 541, 542, 533,
Nouveaux clients	512, 511, 422, 421 412, 411, 311
Prometteur	525, 524, 523, 522, 521, 515, 514
Besoin d'attention	535, 534, 443, 434, 343, 334, 325, 324
Sur le point de dormir	331, 321, 312, 221, 213, 231, 241, 251
À risque	255, 254, 245, 244, 253, 252, 243, 242
On ne peut pas les perdre	155, 154, 144, 214, 215, 115, 114, 113
En hibernation	332, 322, 231, 241, 251, 233, 232
Perdus	111, 112, 121, 131, 141, 151



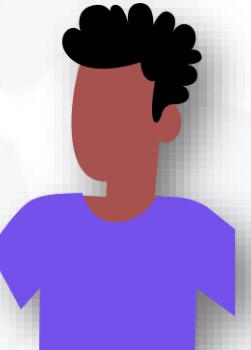
Champion



Loyal potentiel



Nouveaux clients

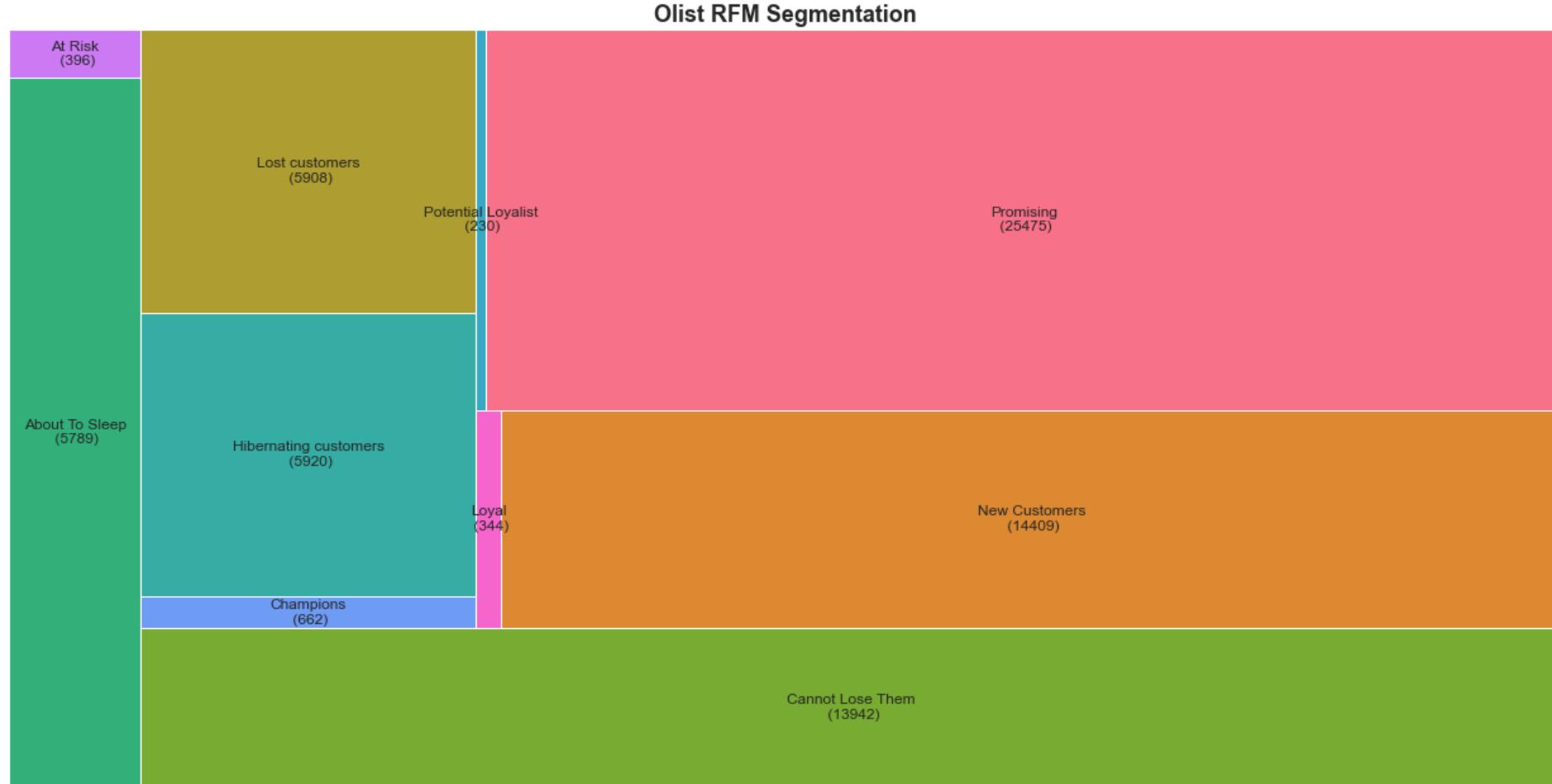


On ne peut pas les perdre

1 - <https://docs.exponea.com/docs/rfm-segmentation>

Nouvelles variables

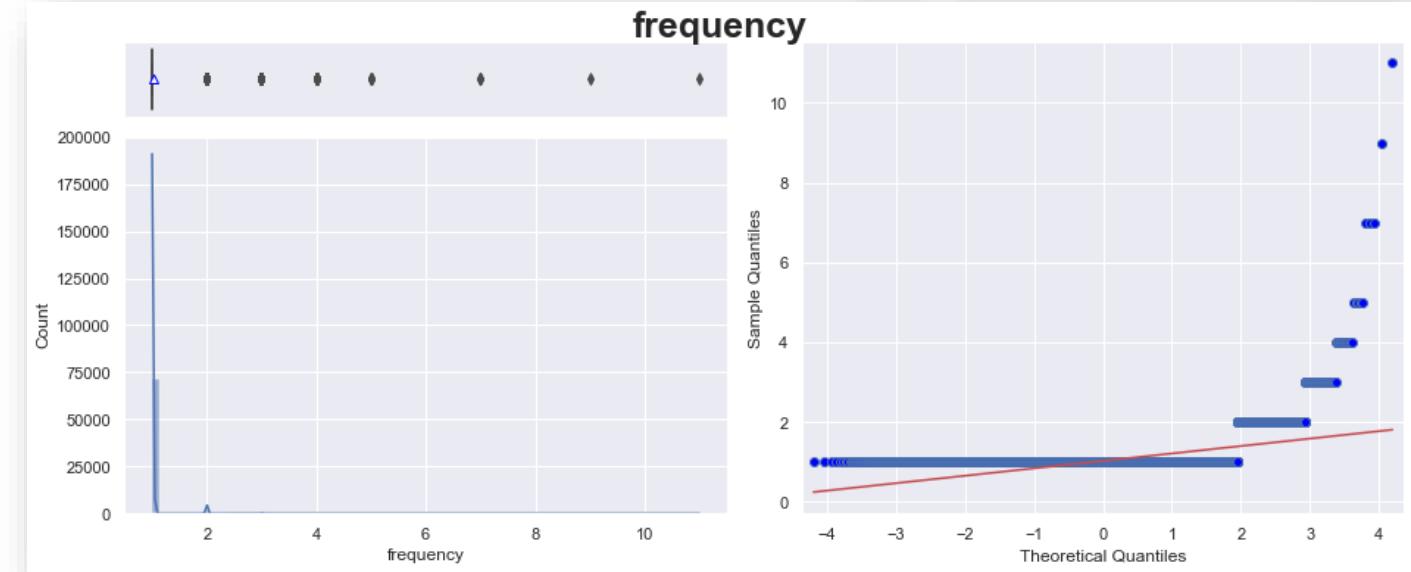
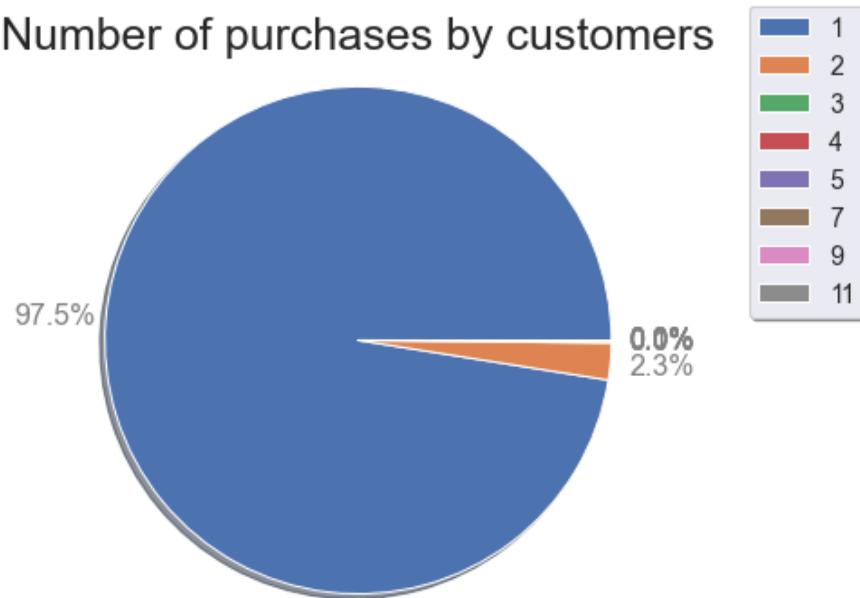
Prometteur et Nouveaux clients sont la plupart de clients dans la segmentation



Sur base : <https://docs.exponea.com/docs/rfm-segmentation>

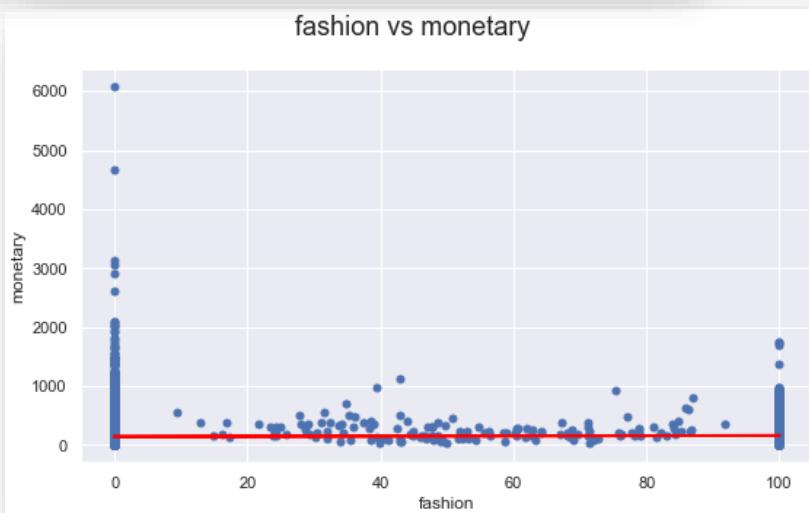
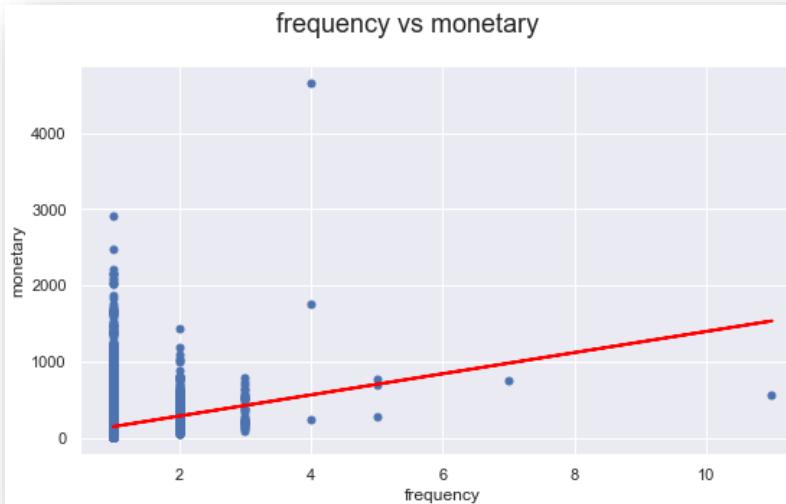
La plupart des clients ont acheté une seule fois

Number of purchases by customers

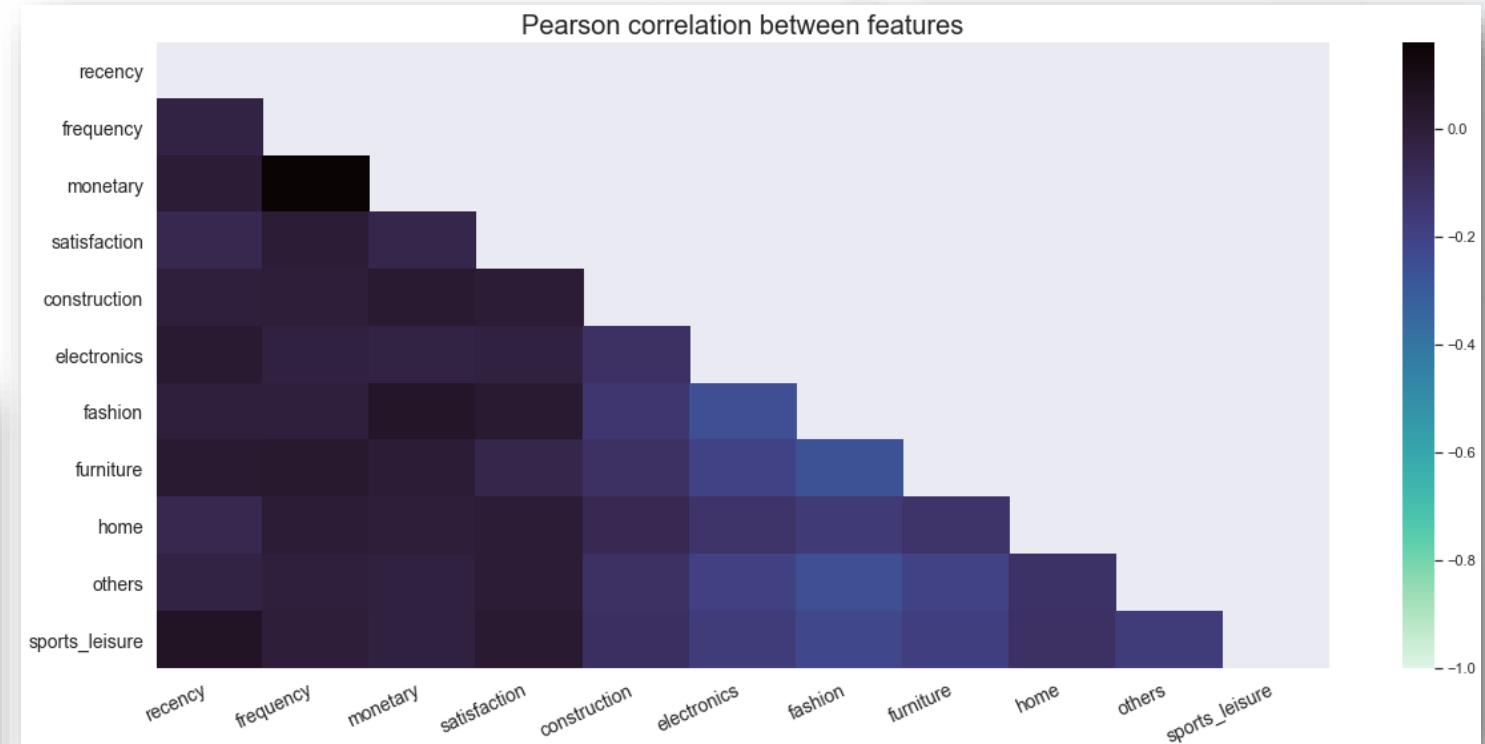


Si le client a acheté plus d'une fois, la valeur sera « True », sinon, la valeur sera « False ».

Les corrélations entre les variables



Les corrélations entre les variables sont faibles



Utilisation des données sans transformation

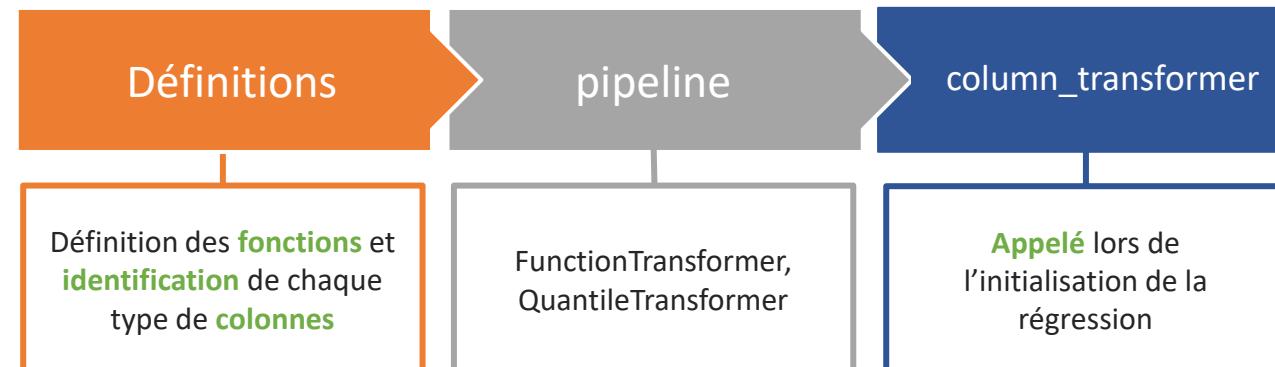
À cette phase, la taille du dataset est de 73074 x 11

Indicateurs RFM			Satisfaction et Catégories							
recency	frequency	monetary	satisfaction	fashion	furniture	others	electronics	sports_leisure	home	construction
0.0	0.001001	0.082583	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
0.0	0.002002	0.285953	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
0.0	0.008509	0.005189	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
0.0	0.015516	0.902495	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
0.0	0.015516	0.630678	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0



Variables exclues

- Étiquette RFM

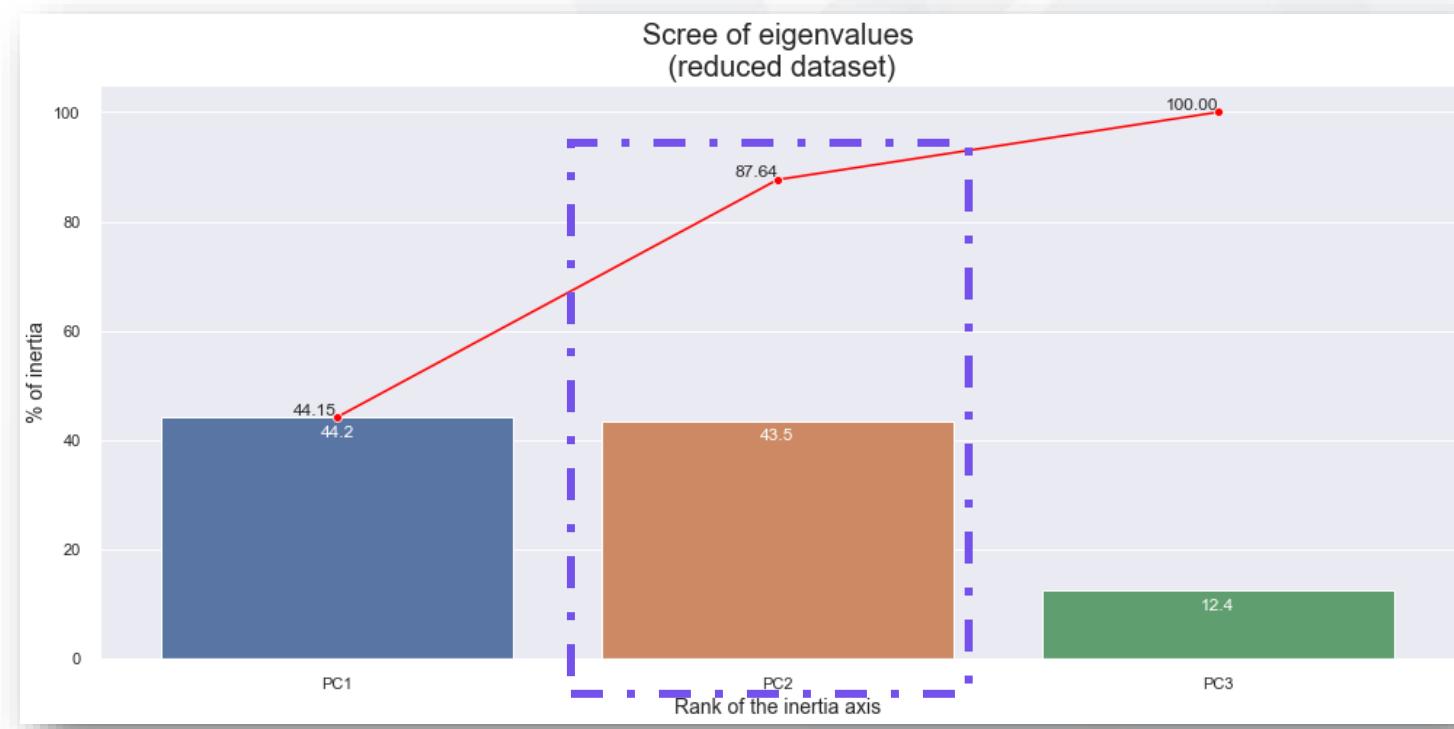
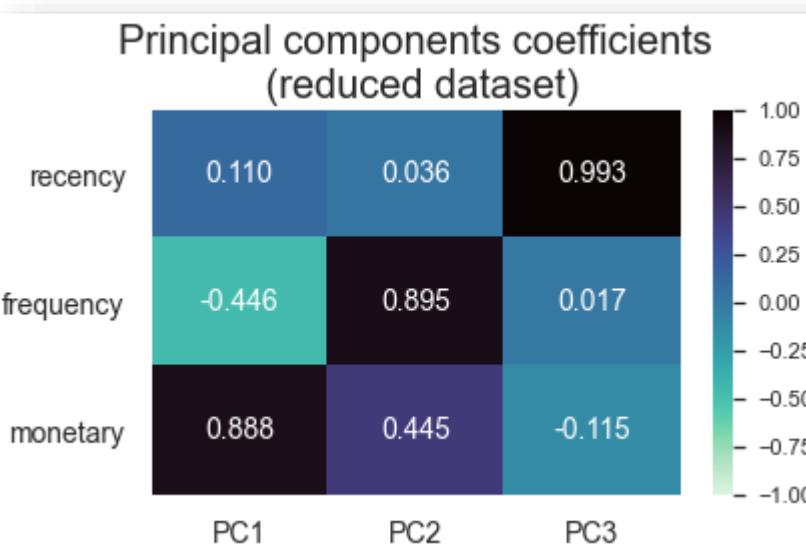


Analyse des composantes principales

PCA

 Seulement les indicateurs RFM

 PC2 décrit jusqu'à **87,64 %** de la variance des données.

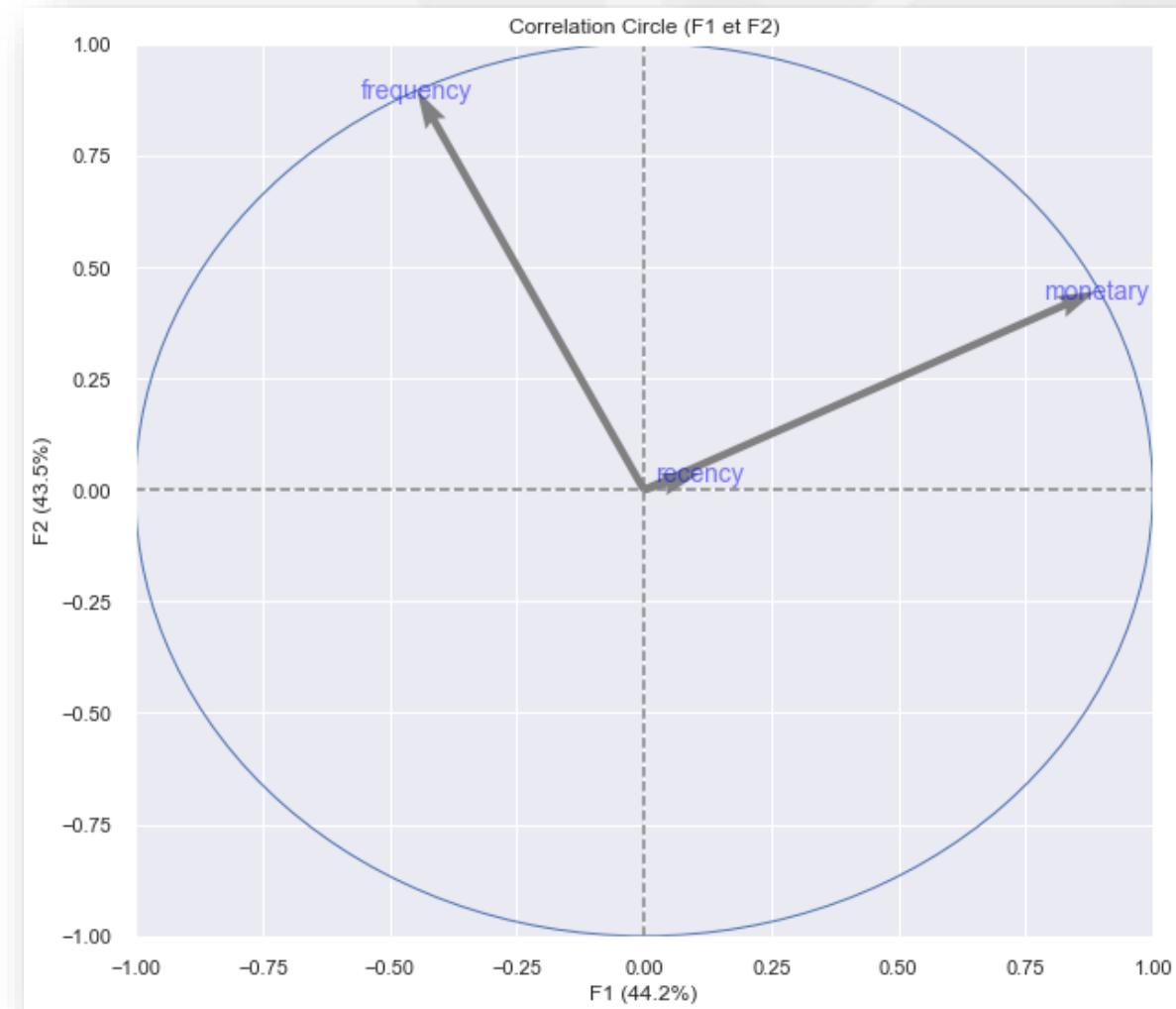
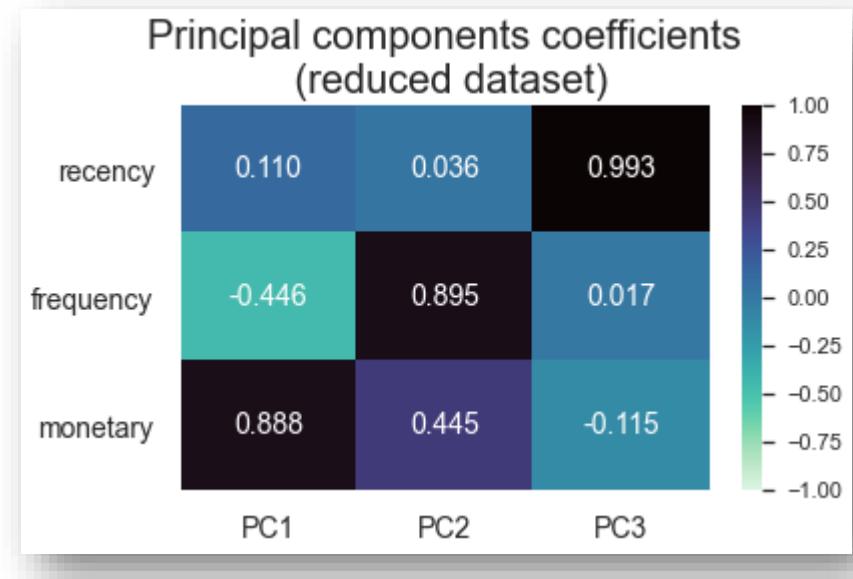


Analyse des composantes principales

PCA



Seulement les indicateurs RFM



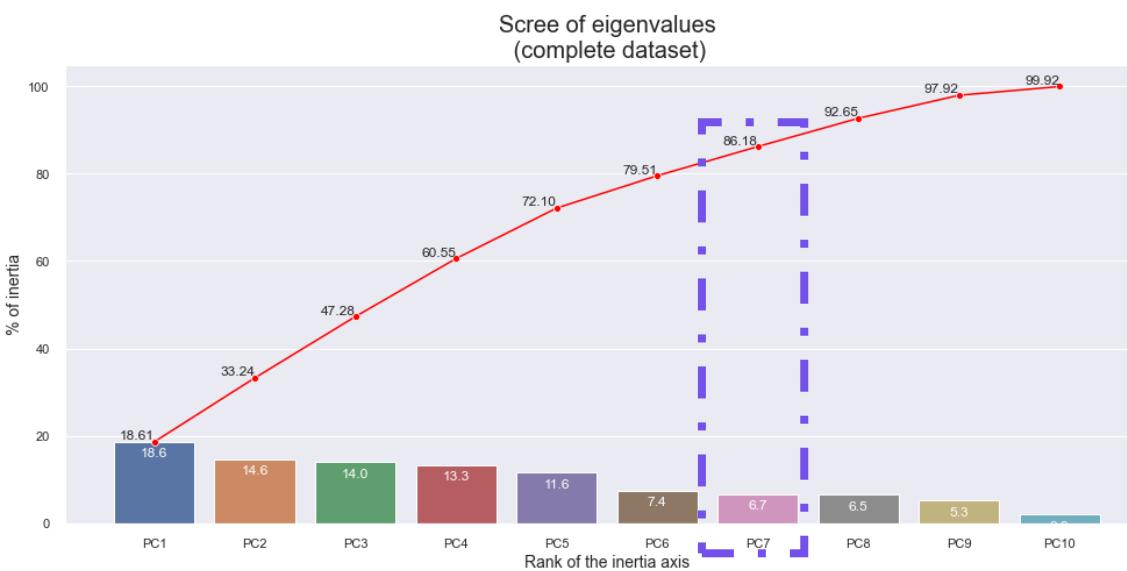
Analyse des composantes principales



Les indicateurs RFM plus les catégories



PC7 décrit jusqu'à **86,18 %** de la variance des données.

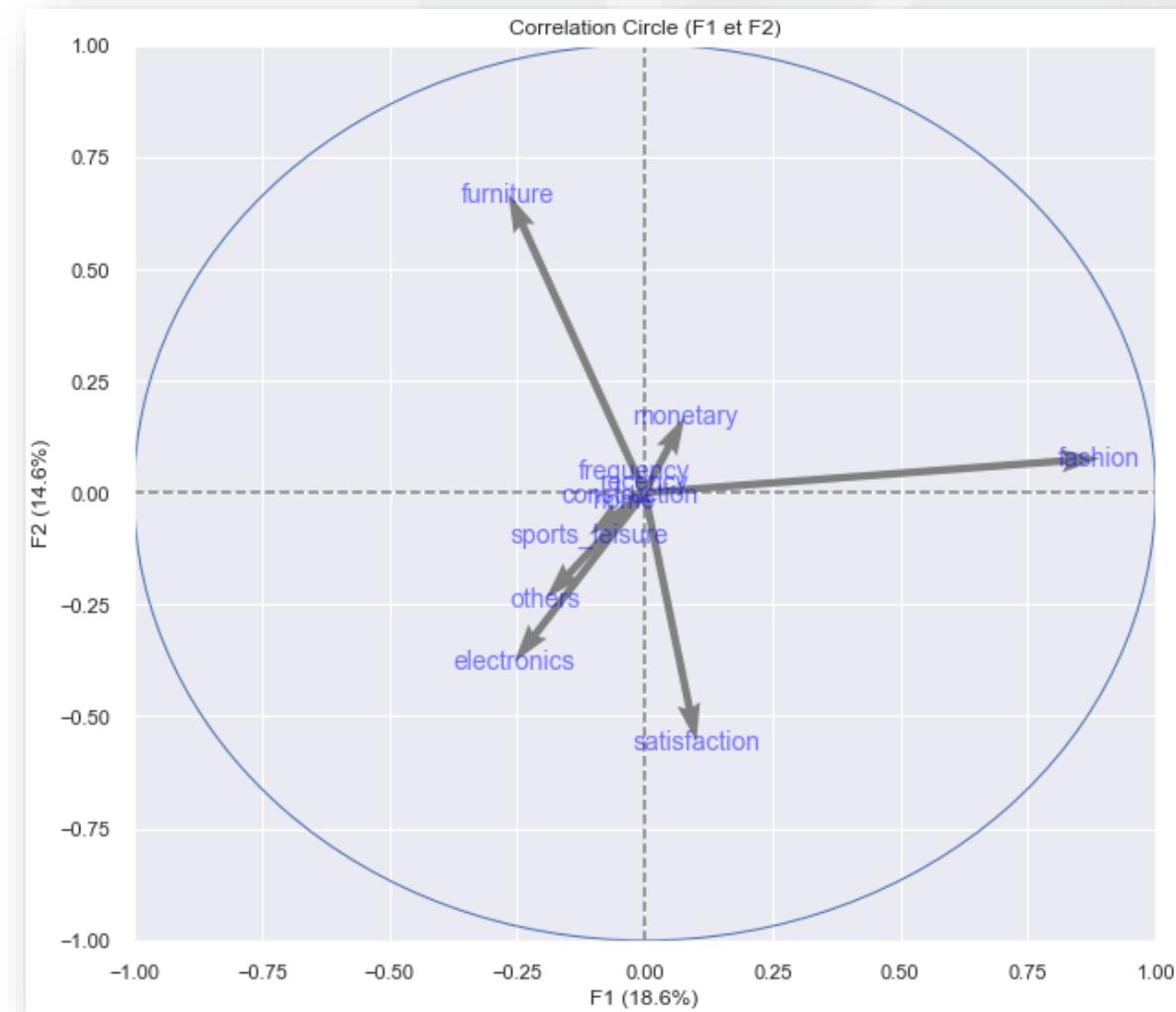
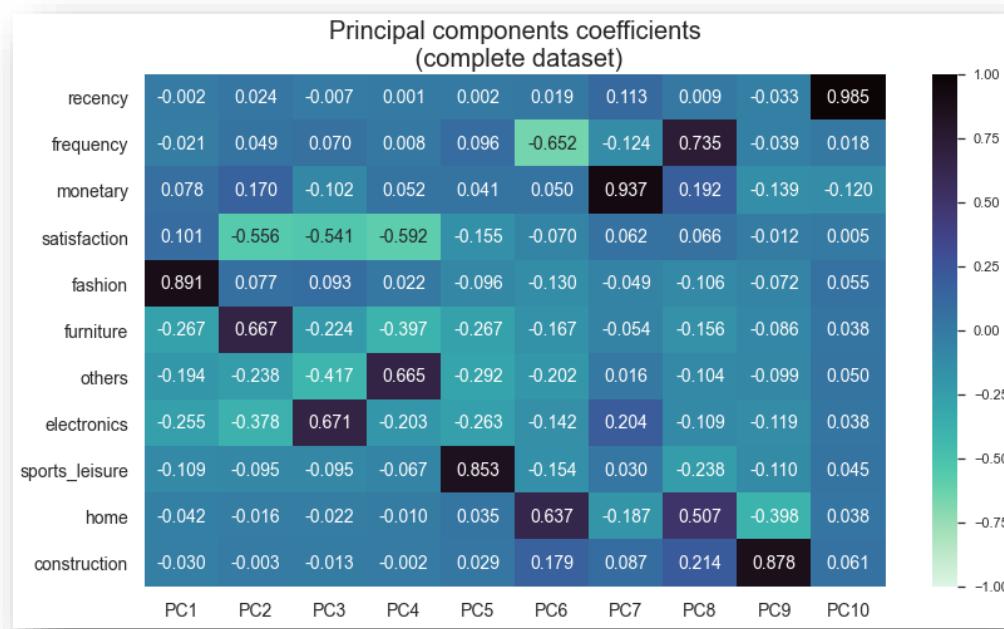


	Principal components coefficients (complete dataset)									
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
recency	-0.002	0.024	-0.007	0.001	0.002	0.019	0.113	0.009	-0.033	0.985
frequency	-0.021	0.049	0.070	0.008	0.096	-0.652	-0.124	0.735	-0.039	0.018
monetary	0.078	0.170	-0.102	0.052	0.041	0.050	0.937	0.192	-0.139	-0.120
satisfaction	0.101	-0.556	-0.541	-0.592	-0.155	-0.070	0.062	0.066	-0.012	0.005
fashion	0.891	0.077	0.093	0.022	-0.096	-0.130	-0.049	-0.106	-0.072	0.055
furniture	-0.267	0.667	-0.224	-0.397	-0.267	-0.167	-0.054	-0.156	-0.086	0.038
others	-0.194	-0.238	-0.417	0.665	-0.292	-0.202	0.016	-0.104	-0.099	0.050
electronics	-0.255	-0.378	0.671	-0.203	-0.263	-0.142	0.204	-0.109	-0.119	0.038
sports_leisure	-0.109	-0.095	-0.095	-0.067	0.853	-0.154	0.030	-0.238	-0.110	0.045
home	-0.042	-0.016	-0.022	-0.010	0.035	0.637	-0.187	0.507	-0.398	0.038
construction	-0.030	-0.003	-0.013	-0.002	0.029	0.179	0.087	0.214	0.878	0.061

Analyse des composantes principales



Les indicateurs RFM plus les catégories



4.

Modélisations effectuées

KMeans

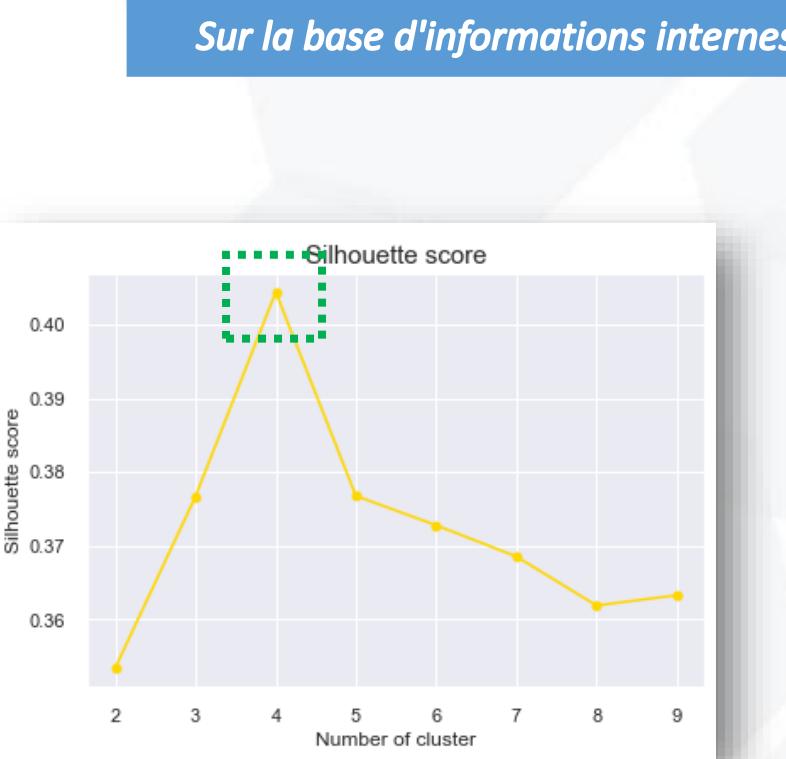
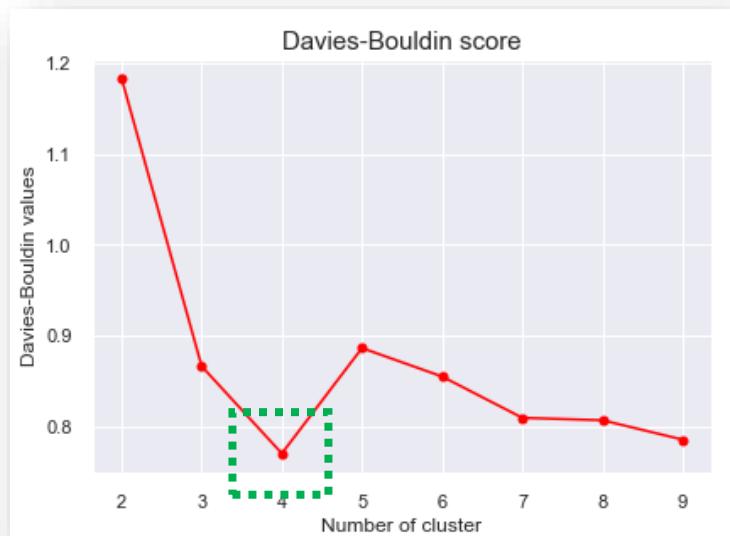
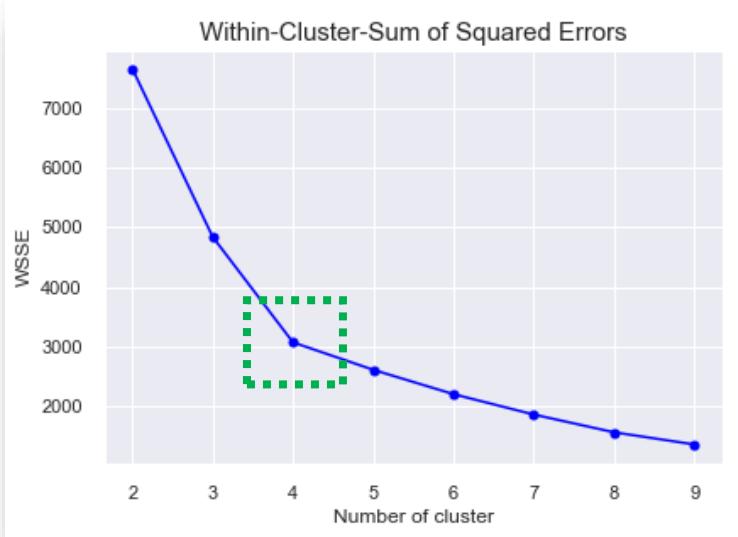


	PC1	PC2	PC3
recency	0.097	0.043	0.994
frequency	-0.542	0.840	0.017
monetary	0.835	0.541	-0.105

4 clusters pour les indicateurs RFM



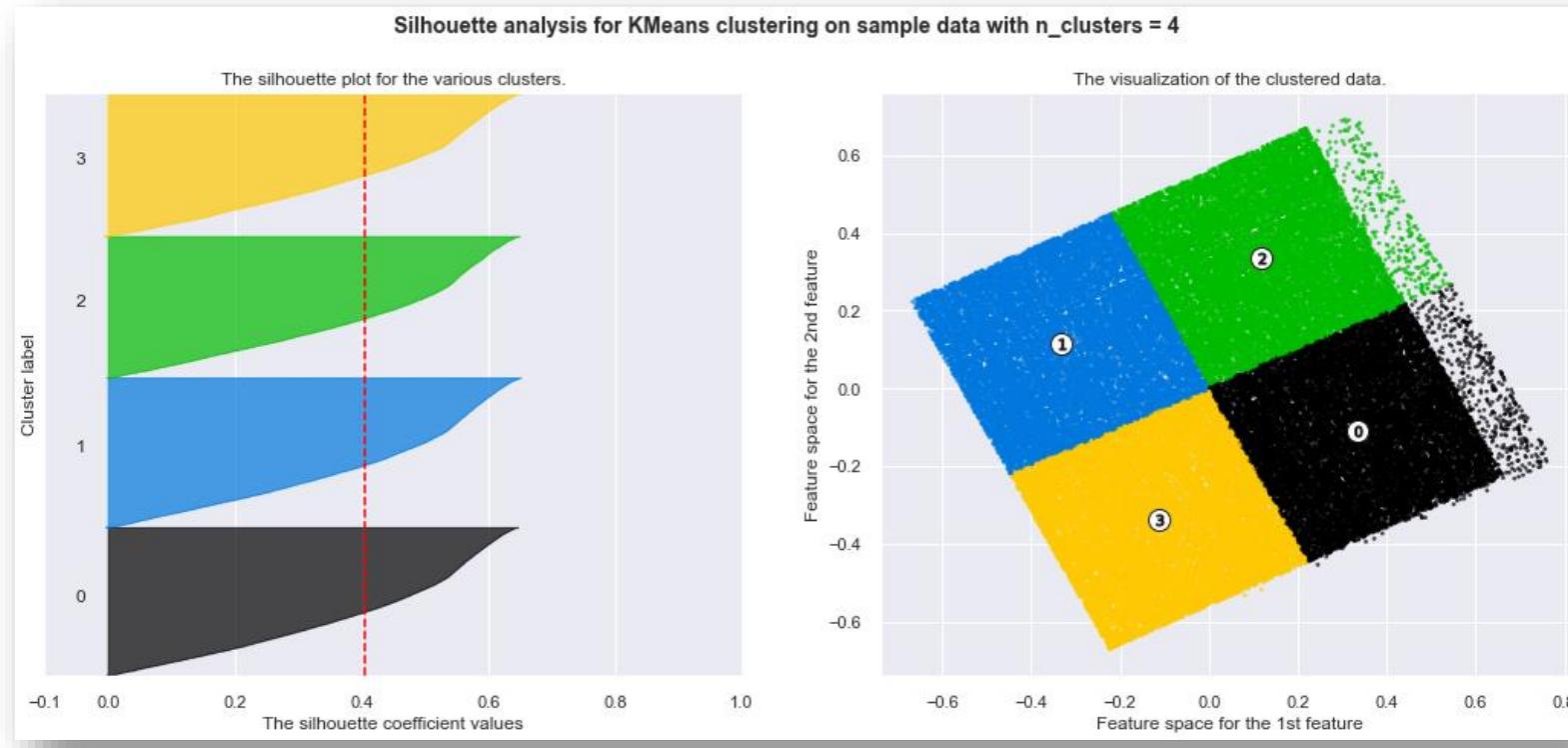
Trois métriques utilisées pour l'identification des clusters



les trois métriques ont donné des résultats similaires dans 4 clusters

Analyse de Silhouette

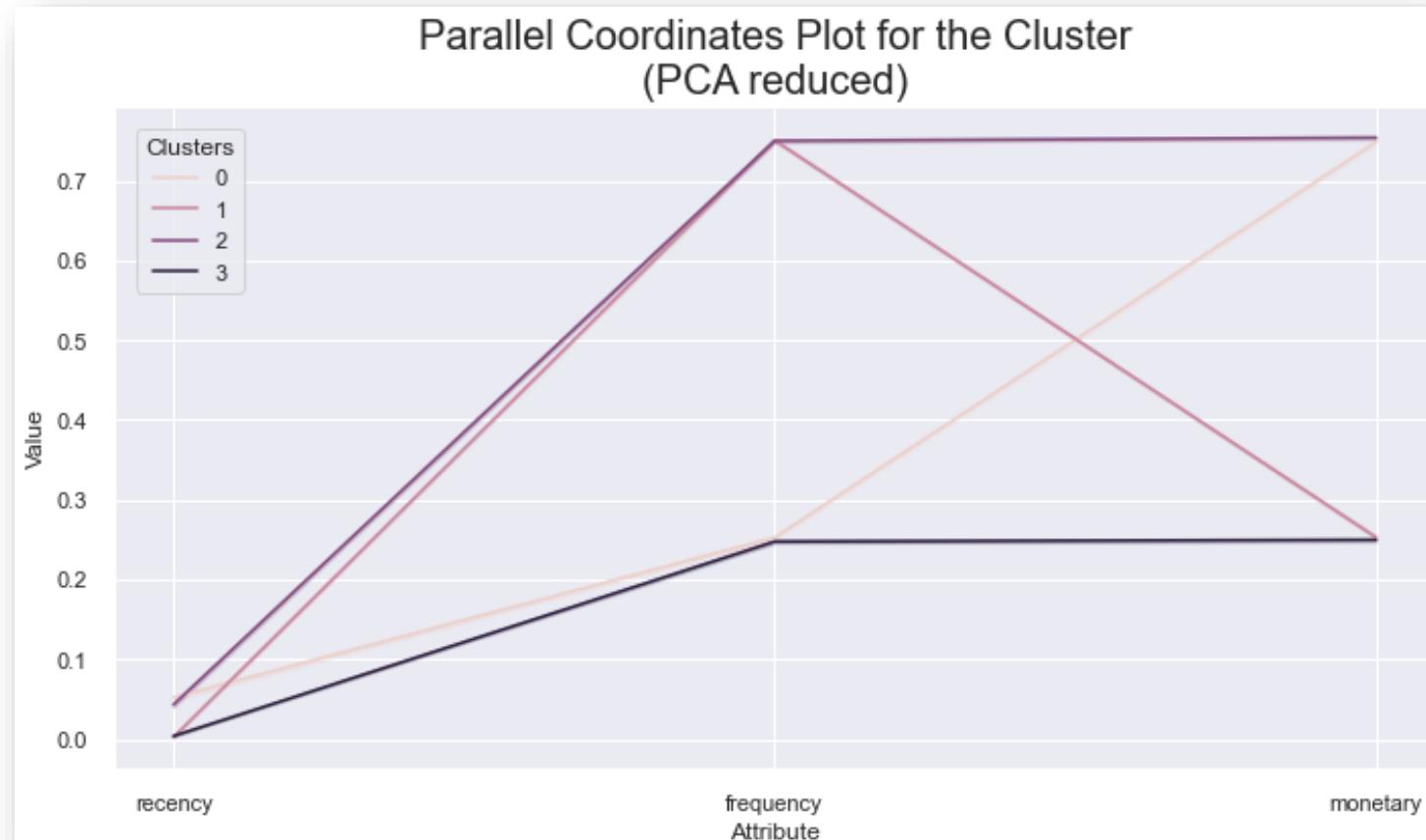
Pour n_clusters = 4. La silhouette_score moyenne est de : 0.404



```
n_clusters = 2 The silhouette_score is : 0.35352040765439713  
n_clusters = 3 The silhouette_score is : 0.37674029464077297  
n_clusters = 4 The silhouette_score is : 0.40445401159149613  
n_clusters = 5 The silhouette_score is : 0.37678773346336986  
n_clusters = 6 The silhouette_score is : 0.3728151050570926  
n_clusters = 7 The silhouette_score is : 0.3685782522275606  
n_clusters = 8 The silhouette_score is : 0.3619232638470083  
n_clusters = 9 The silhouette_score is : 0.3633075725034612
```

Les valeurs pour chacun des clusters étaient comprises entre 0.35 et 0.40

Signification des clusters



Cluster 0 : Qui n'a pas acheté récemment, une seule fois et pour une bonne somme d'argent.



Cluster 1 : Qui a acheté récemment, plus d'une fois et pour une somme modique d'argent.

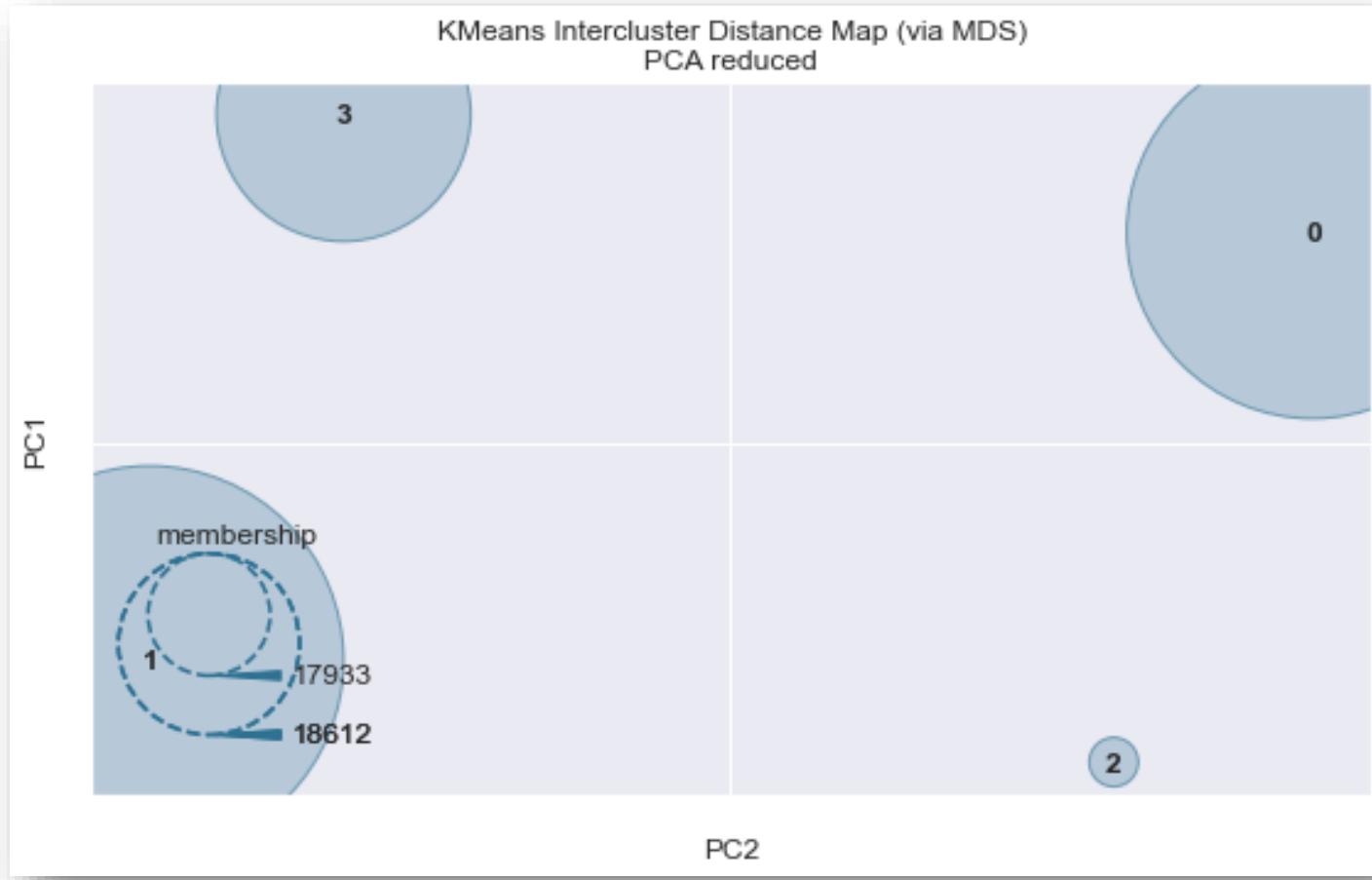


Cluster 2 : Qui n'a pas acheté récemment, plus d'une fois et pour une bonne somme d'argent



Cluster 3 : Qui a acheté récemment, une seule fois et pour une somme modique

Distance entre les clusters

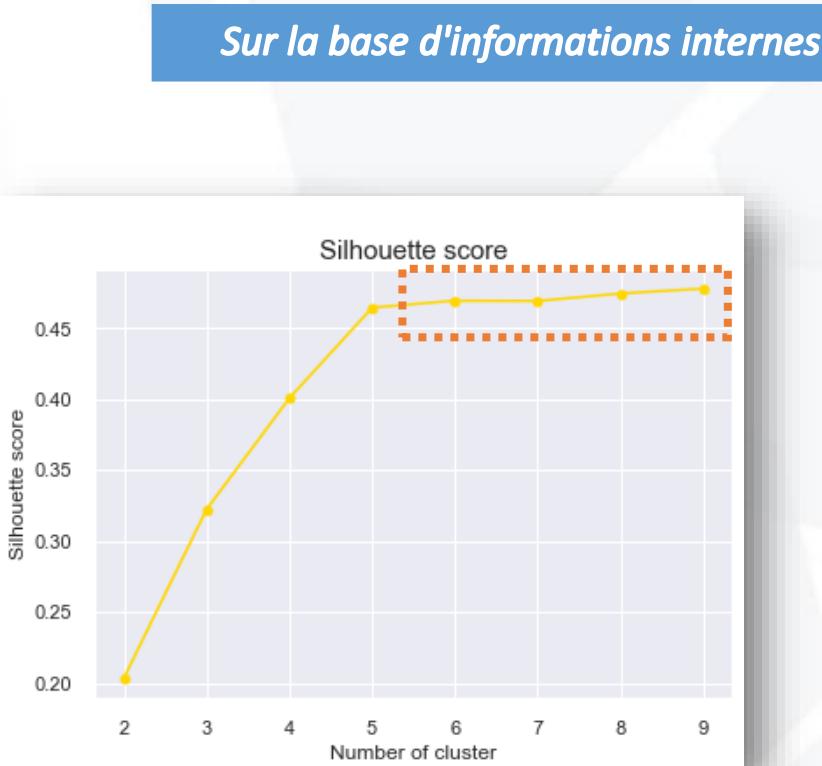
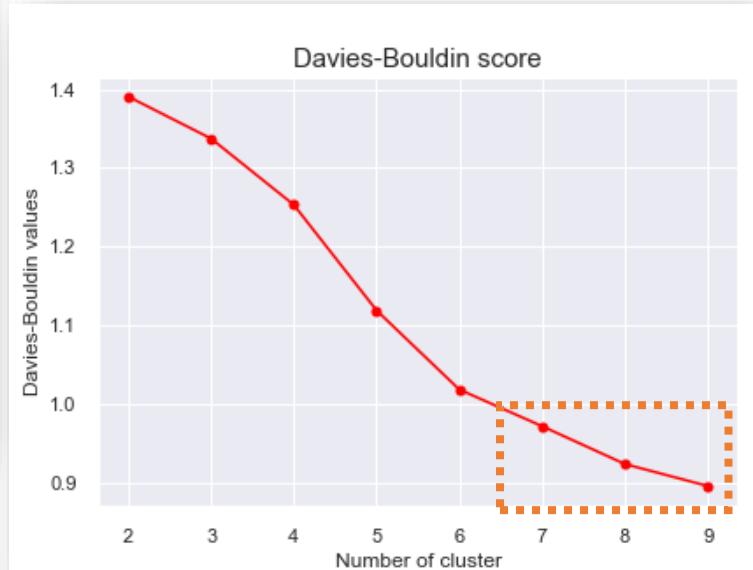
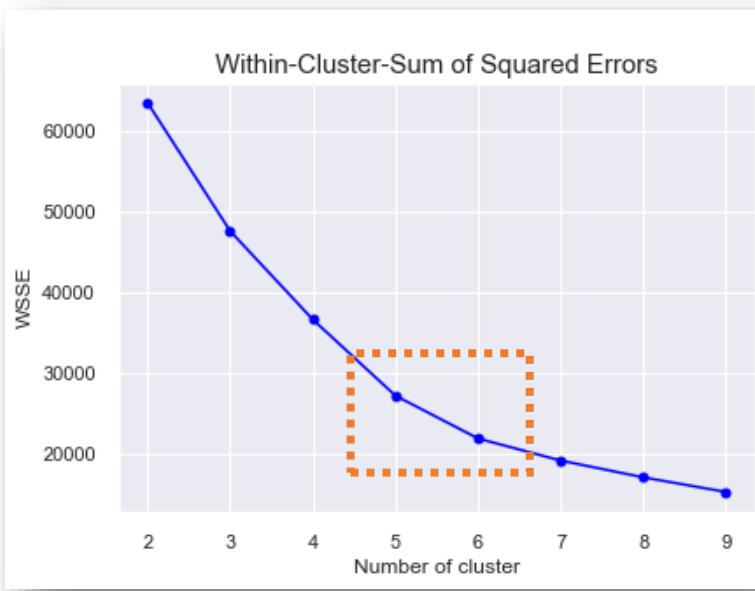


La distance semble suffisante entre les clusters

Les indicateurs RFM plus les catégories



Trois métriques utilisées pour l'identification des clusters

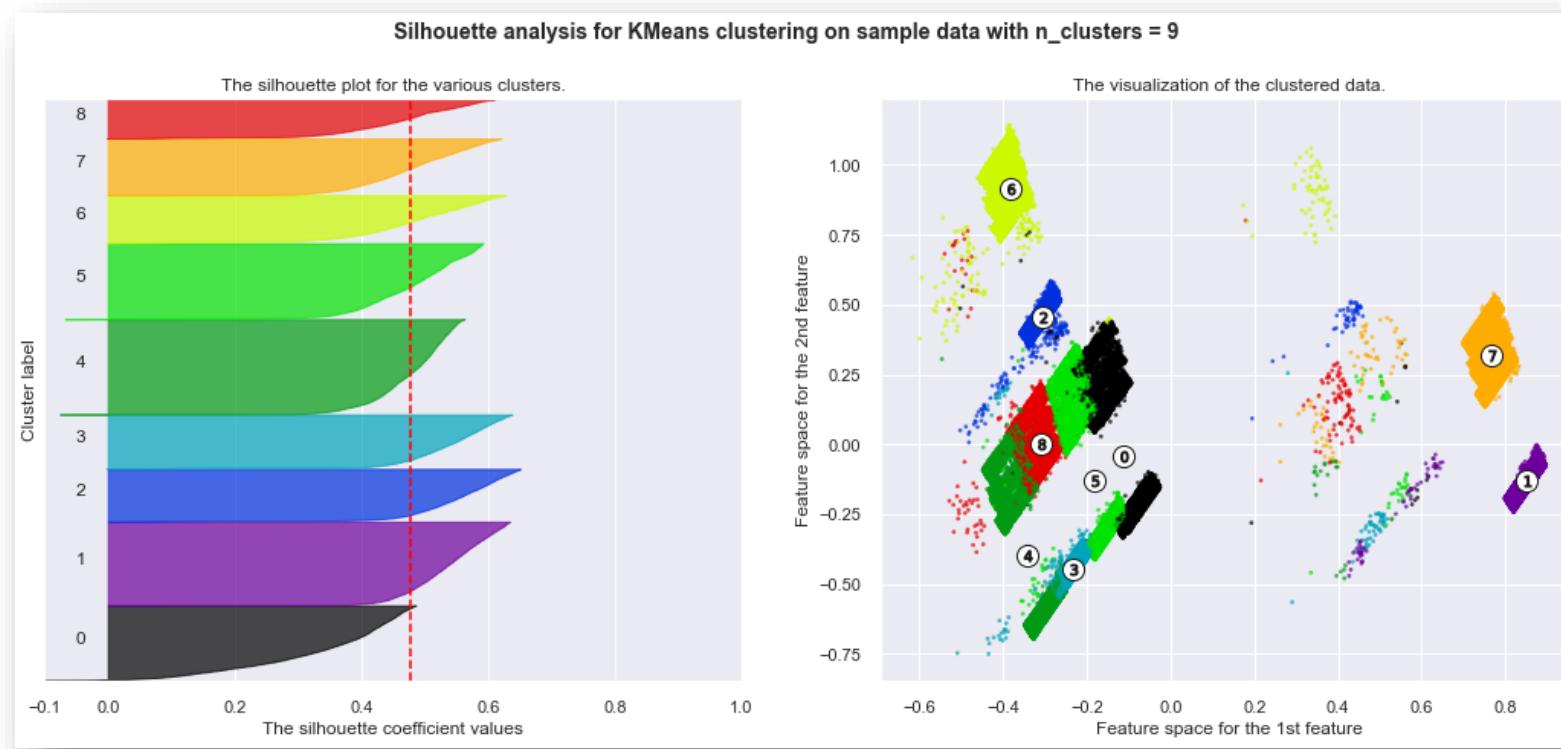


Dans ce cas-là, ce n'est pas très clair par rapport au RFM seulement

	Principal components coefficients (complete dataset)									
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
recency	-0.002	0.024	-0.007	0.001	0.002	0.019	0.113	0.009	-0.033	0.985
frequency	-0.021	0.049	0.070	0.008	0.096	-0.652	-0.124	0.735	-0.039	0.018
monetary	0.078	0.170	-0.102	0.052	0.041	0.050	0.937	0.192	-0.139	-0.120
satisfaction	0.101	-0.556	-0.541	-0.592	-0.155	-0.070	0.062	0.086	-0.012	0.005
fashion	0.891	0.077	0.093	0.022	-0.096	-0.130	-0.049	-0.106	-0.072	0.055
furniture	-0.267	0.667	-0.224	-0.397	-0.267	-0.167	-0.054	-0.156	-0.086	0.038
others	-0.194	-0.238	-0.417	0.655	-0.292	-0.202	0.016	-0.104	-0.099	0.050
electronics	-0.255	-0.378	0.671	-0.203	-0.263	-0.142	0.204	-0.109	-0.119	0.038
sports_leisure	-0.109	-0.095	-0.095	-0.067	0.853	-0.154	0.030	-0.238	-0.110	0.045
home	-0.042	-0.016	-0.022	-0.010	0.035	0.637	-0.187	0.507	-0.398	0.038
construction	-0.030	-0.003	-0.013	-0.002	0.029	0.179	0.087	0.214	0.678	0.061

Analyse de Silhouette pour approfondir

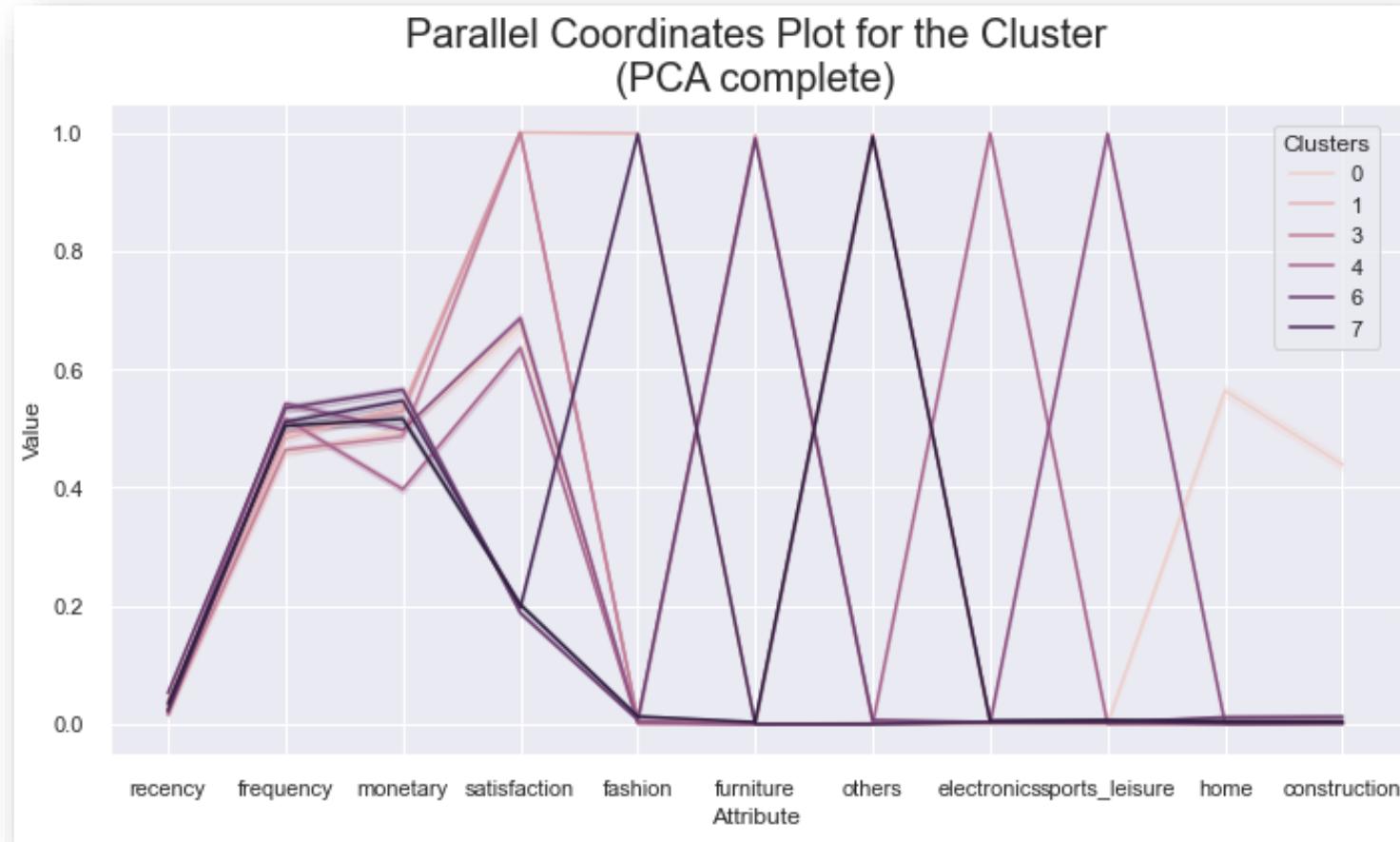
Pour n_clusters = 9. La silhouette_score moyenne est de : 0.478



```
n_clusters = 2 The silhouette_score is : 0.20388522732084619  
n_clusters = 3 The silhouette_score is : 0.32272570091548125  
n_clusters = 4 The silhouette_score is : 0.40096180909595214  
n_clusters = 5 The silhouette_score is : 0.4646766527291683  
n_clusters = 6 The silhouette_score is : 0.46955771592323126  
n_clusters = 7 The silhouette_score is : 0.4693509987403661  
n clusters = 8 The silhouette score is : 0.47458832183501604  
n clusters = 9 The silhouette score is : 0.4780515708079847
```

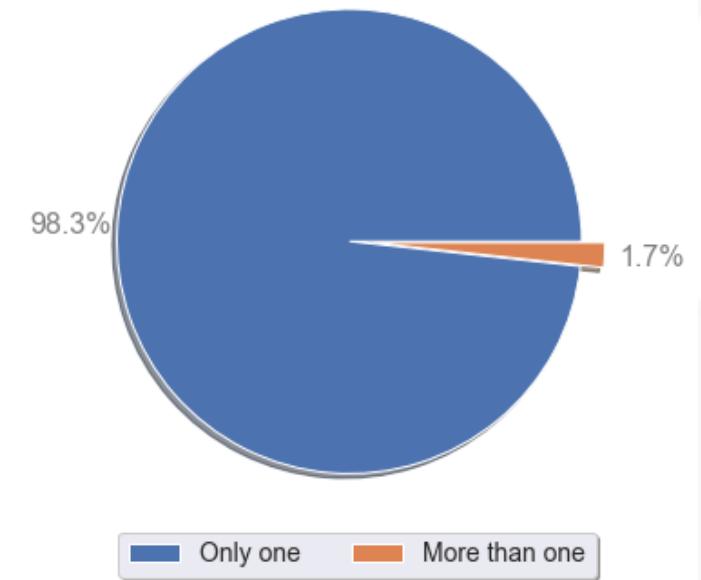
Les valeurs entre les clusters 6 et 9 étaient comprises entre 0.469 et 0.478

Signification des clusters

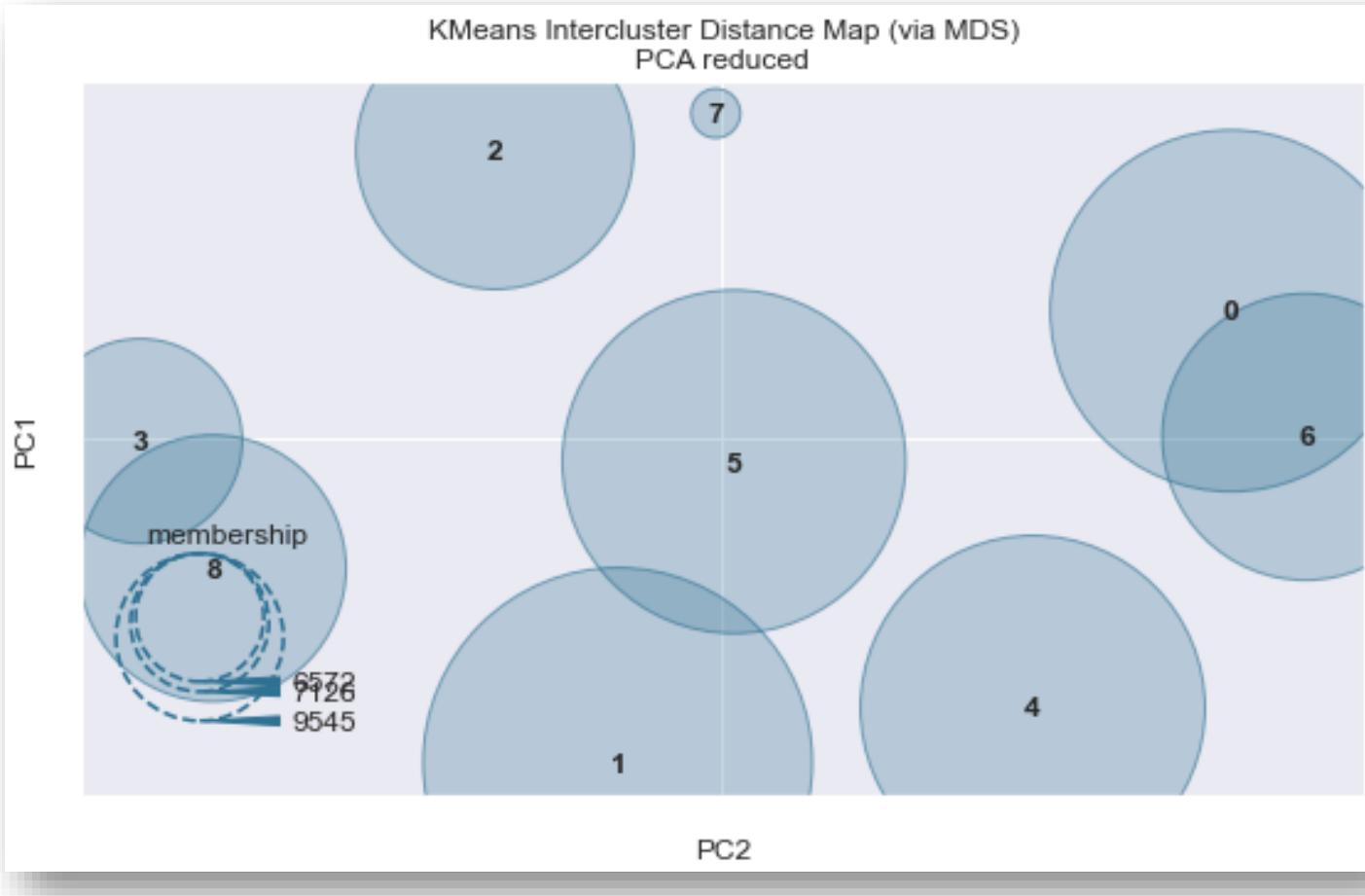


Il faut prendre en compte que :

Customers who bought more than one category



Distance entre les clusters



Dans ce cas-là, la distance ne semble pas suffisante entre les clusters

5.

Modèle sélectionné

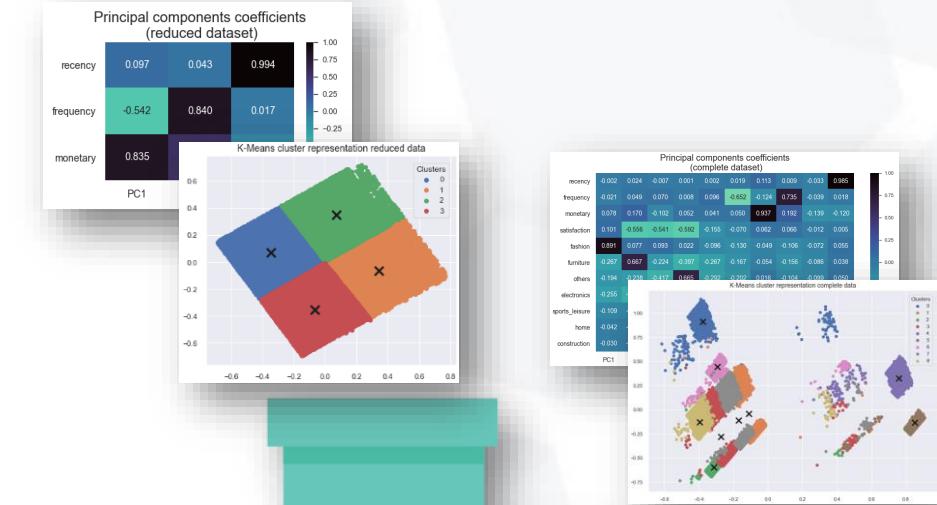


4 clusters pour la segmentation

RFM	vs	RFM + les catégories
4		# Clusters
0,479s		Time
3076		Inertia
72548,351		calinski-harabasz
0.770		davies-bouldin
0.404		Silhouette

Informations internes

RFM a donné le meilleur résultat



1

2

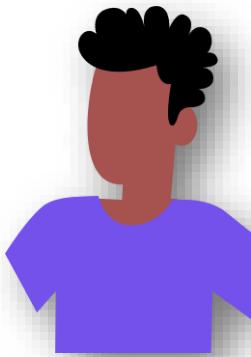
Il n'y a pas d'informations externes pour mesurer les résultats

Description du cluster

Cluster 0



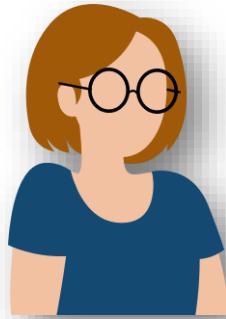
Cluster 1



Cluster 2

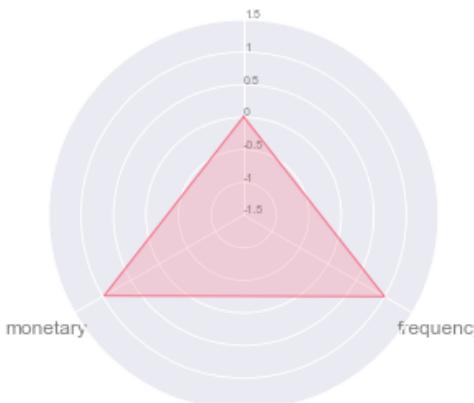


Cluster 3



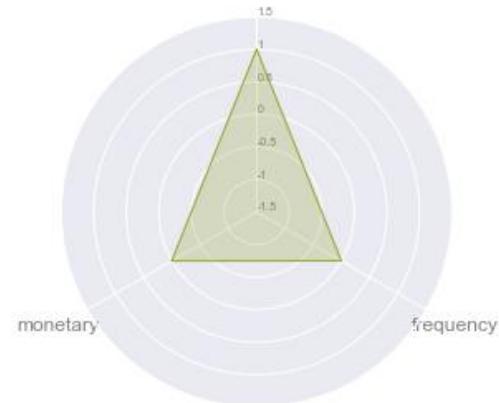
Cluster 0

recency



Cluster 1

recency



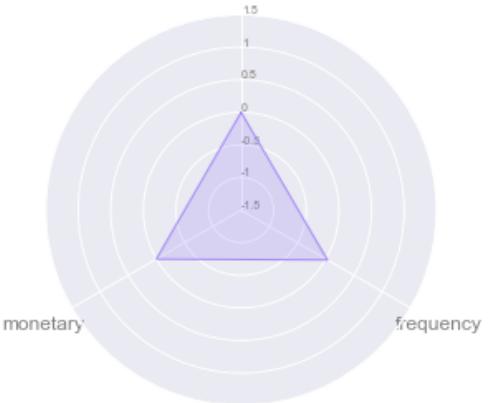
Cluster 2

recency



Cluster 3

recency



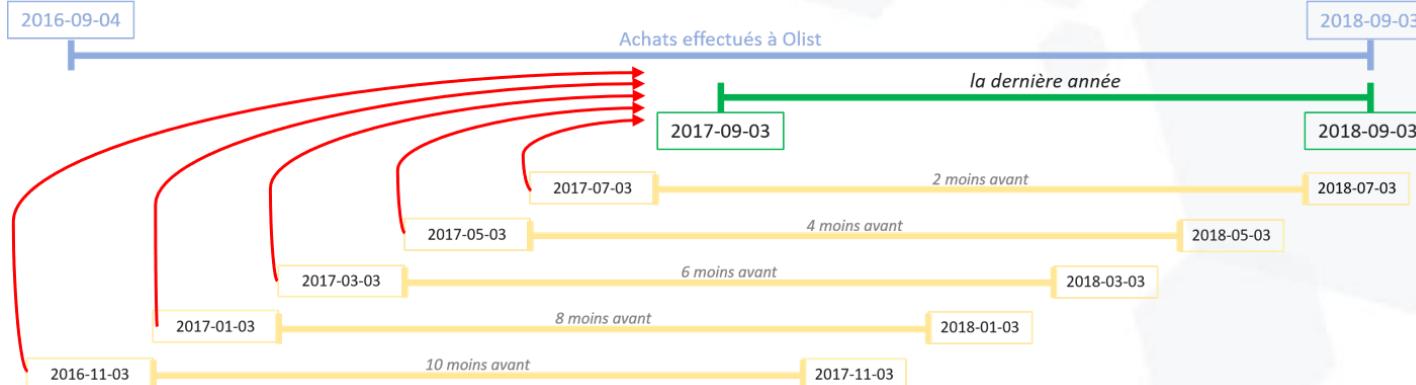
La fréquence doit être mise à jour tous les 4 mois



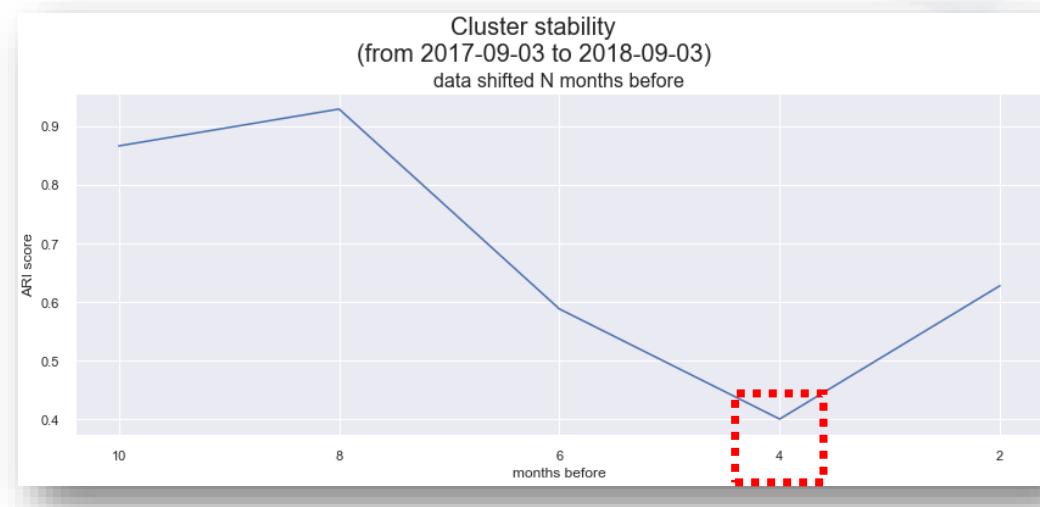
Clients communs pour chaque dataset par rapport à la dernière année

- 61176 (*2 mois*)
- 48869 (*4 mois*)
- 35232 (*6 mois*)
- 21687 (*8 mois*)
- 9008 (*10 mois*)

Des clients ont beaucoup changé lors des deux dernières années



À cette phase, la taille du dataset est de 73074 x 12

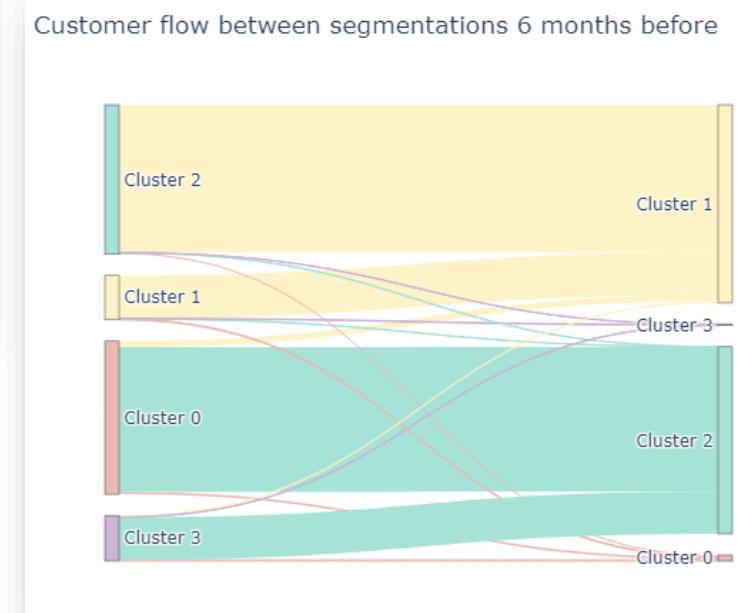
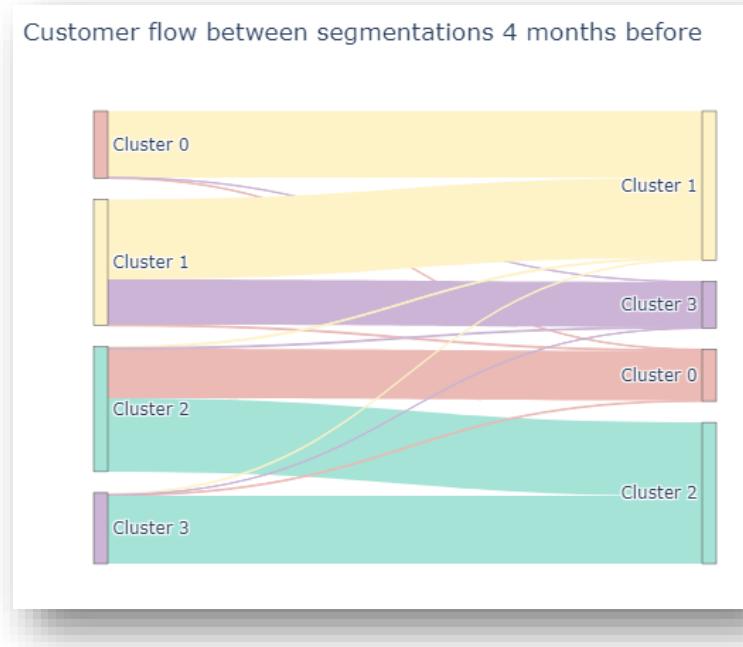


Le pourcentage de clients qui ont changé de segment est de % 60,4

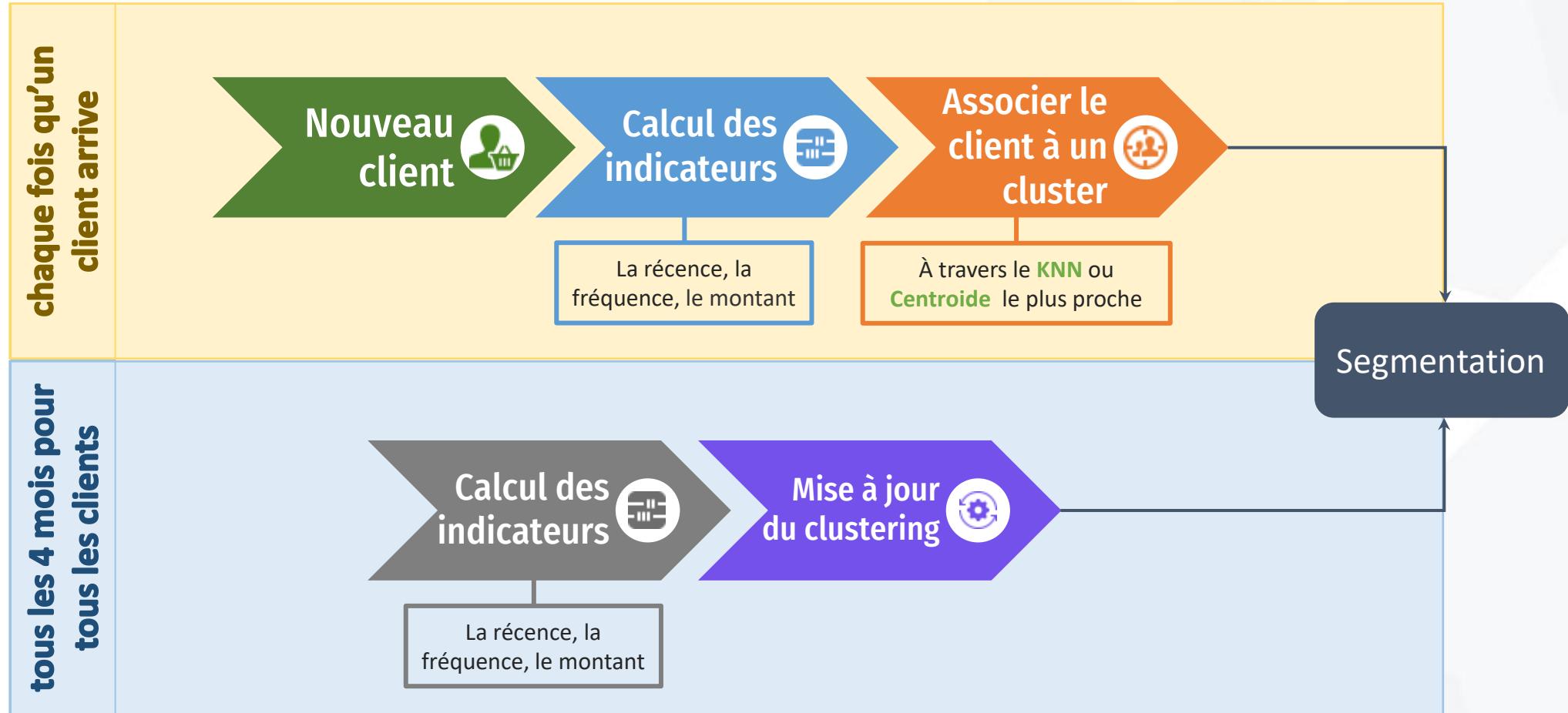


Clients communs pour chaque dataset par rapport à la dernière année

- 61176 (*2 mois*) % 17,04
- 48869 (*4 mois*) % 60,4
- 35232 (*6 mois*) % 88,4
- 21687 (*8 mois*) % 98,76
- 9008 (*10 mois*) % 51,02



La stratégie d'ajout de nouveaux clients



6.

Conclusions



Les conclusions de la mission

Basé sur la segmentation du comportement et de la valeur



Il est nécessaire de prendre en compte plus de variable pour identifier chaque type de clients selon ses comportements d'achat. *Par exemple, si c'est une promotion, l'anniversaire du client.*

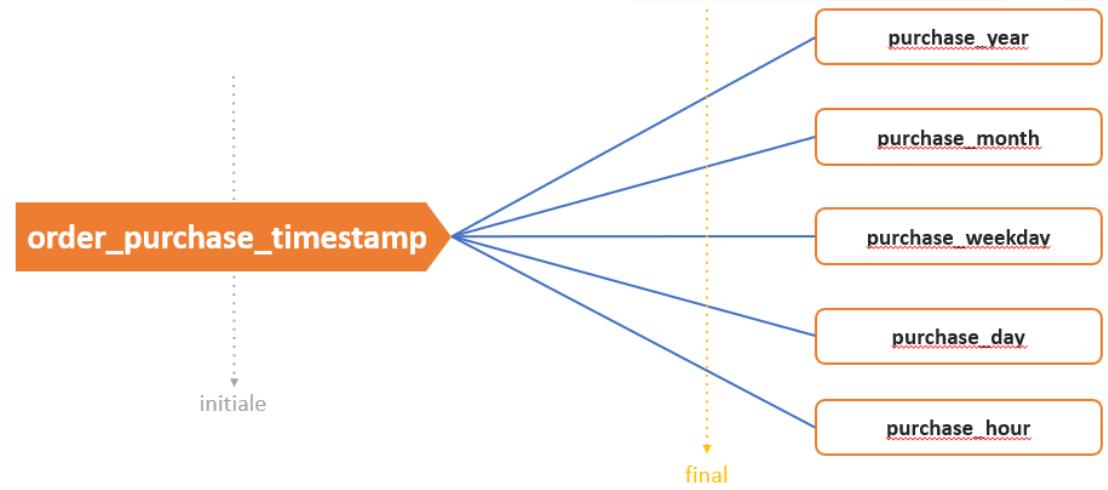


Il faut utiliser la variable « `order_purchase_timestamp` » ainsi que les catégories pour savoir s'il y a des clients qui achètent uniquement à certaines périodes de l'année.

Par exemple, des achats pour la fête de Noé.



Explorez d'autres types de réduction de dimension tels que UMAP, ISOMAP, etc., et aussi d'autres types d'algorithme de clustering tels que DBSCAN, MiniBatchKMeans, etc.



Avez-vous des questions ?



MERCI

Soutenance de Projet
Samir HINOJOSA

27 octobre 2021



OPENCLASSROOMS

Annexes

Présentation du jeu de données

sellers_dataset.csv	comprend des données sur les vendeurs qui ont exécuté les commandes passées sur Olist				
	Taille:	3095x4	Pourcentage de NaN:	0 %	Doublon:
order_payments_dataset.csv	Informations sur les options de paiement des commandes.				
	Taille:	1103886x5	Pourcentage de NaN:	0 %	Doublon:
geolocation_dataset.csv	contient des informations sur les codes postaux brésiliens et ses coordonnées lat/lng.				
	Taille:	1000163x5	Pourcentage de NaN:	20.93 %	Doublon:
product_category_name_translation.csv	Traduit le product category name en anglais.				
	Taille:	71x2	Pourcentage de NaN:	0 %	Doublon:

Liste de catégories des produits à travailler

Catégories de produits	Nouvelles catégories
fashion_bags_accessories, fashio_female_clothing, fashion_sport, fashion_shoes, fashion_male_clothing, fashion_underwear_beach, fashion_childrens_clothes, cool_stuff, art, arts_and_craftmanship, health_beauty, perfumery, watches_gifts	fashion
kitchen_dining_laundry_garden_furniture, furniture_decor, office_furniture, furniture_bedroom, furniture_living_room, furniture_mattress_and_upholstery, bed_bath_table	furniture
home_appliances, home_appliances_2, home_confort, home_comfort_2, air_conditioning, housewares, flowers	home
electronics, audio, tablets_printing_image, telephony, fixed_telephony, small_appliances, small_appliances_home_oven_and_coffee, computers_accessories, computers, pc_gamer, consoles_games, dvds_blu_ray, portateis_cozinha_e_preparadores_de_alimentos	electronics
construction_tools_construction, construction_tools_lights, construction_tools_safety, costruction_tools_garden, costruction_tools_tools, garden_tools, home_construction	construction
sports_leisure, musical_instruments, toys, cine_photo, cds_dvds_musicals, music, books_general_interest, books_imported, books_technical	sports_leisure
christmas_supplies, stationery, party_supplies, auto, luggage_accessories, signaling_and_security, agro_industry_and_commerce, security_and_services, market_place, pet_shop, industry_commerce_and_business, baby, diapers_and_hygiene, drinks, food, food_drink, la_cuisine, unknown	Others