

PROJET 5

Segmenter des clients d'un site e-commerce

Alexandre ROGUES - Mai 2024



Segmentation

Transformations :

Variables catégorielles
Création de nouvelles variables
Transformations mathématiques
Normalisation

Stratégie :

Définir et adapter aux besoins métier
Proposer et justifier le nombre de segments
Ajout de nouveaux clients

Évaluation :

Choix des métriques (silhouette, elbow, etc.)
Forme et stabilité des clusters
Optimisation des hyper-paramètres
Stabilité dans le temps

Qualité du code :

Respect des conventions PEP8
Code commenté et documenté

- *Évaluer les performances des modèles d'apprentissage non supervisé*

- *Sélectionner et entraîner des modèles d'apprentissage non-supervisé*

L'entreprise Olist

olist



Boutique officielle sur les marketplaces -

Stock en ligne -
Multiples canaux de vente -

Collecte de commandes -

Omnicanalité -

Flux financier -

Point de vente -

Livraison bon marché -

Facturation des commandes -

Campagnes promotionnelles -

Attraction de clients -

*“Permettre aux entreprises de gérer facilement leurs opérations
de commerce électronique”*

01

Requêtes SQL urgentes

02

Segmentation clients

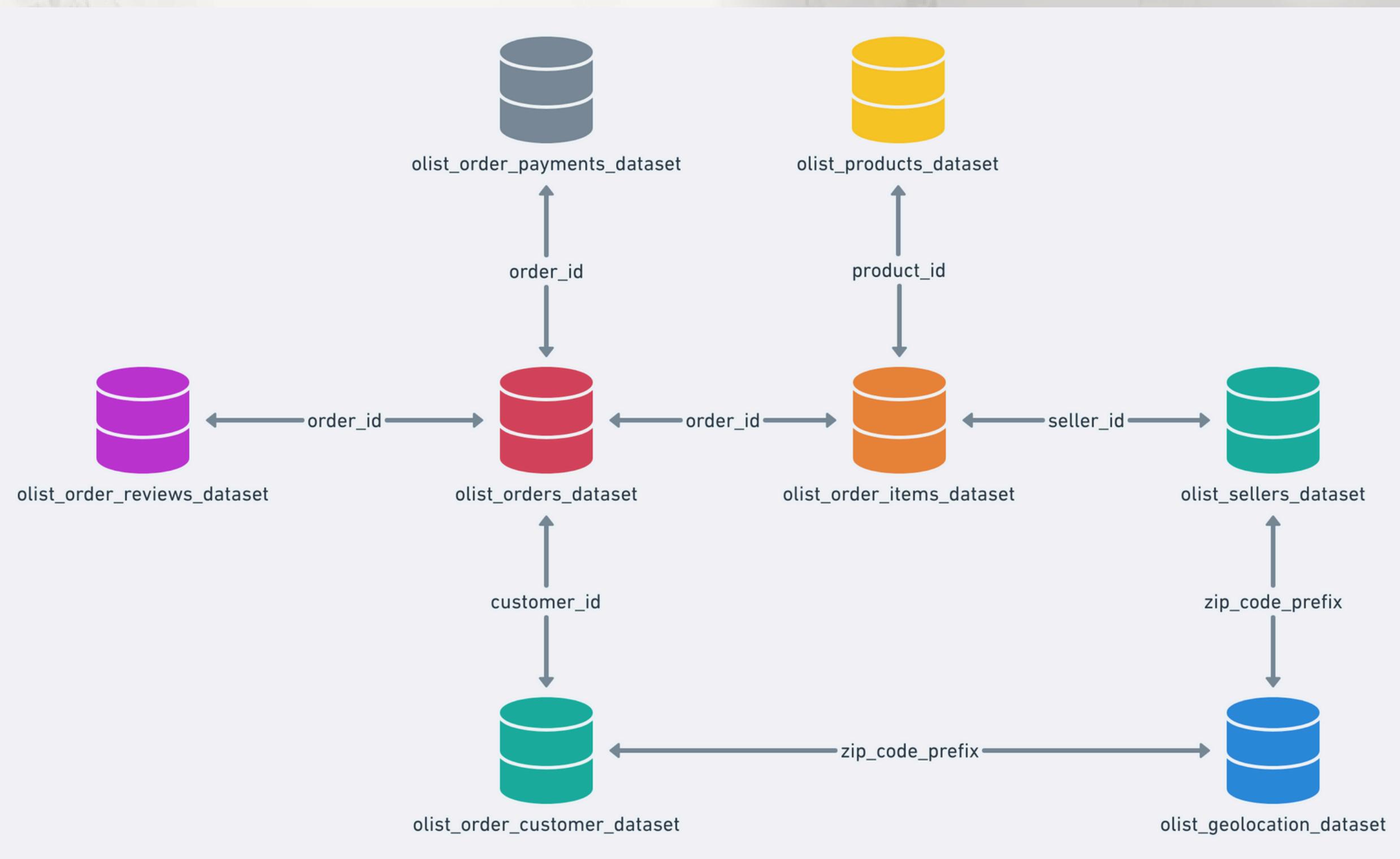
03

Maintenance de la solution

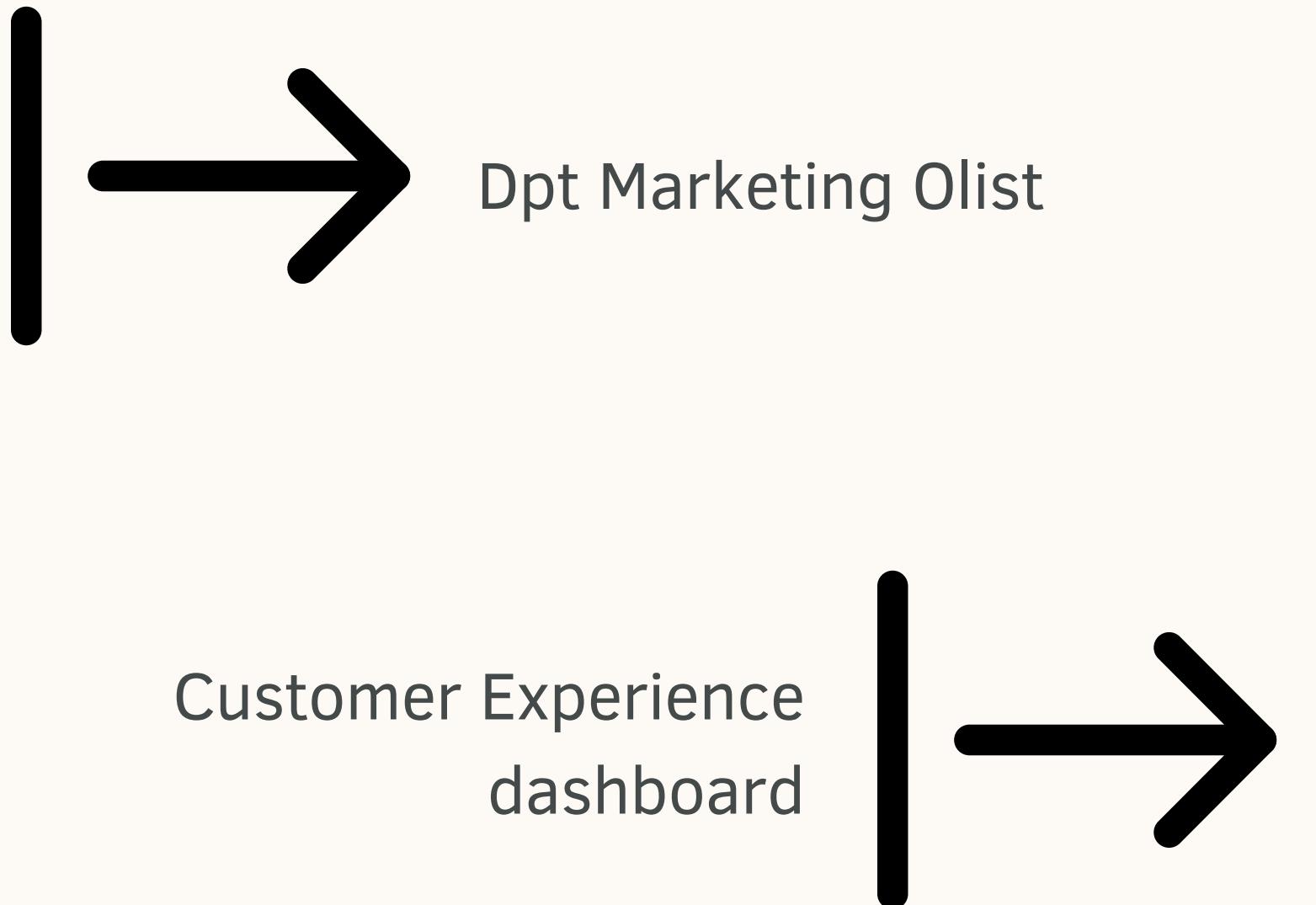


Besoins

Les données



LES REQUETES



Environnement de travail

- module iPython
- Jupyter Lab Notebooks

```
pip install ipython-sql
```

```
# chargement extension SQL ipython  
%load_ext sql  
%sql sqlite:///olist.db
```

Commandes récentes avec au moins 3 jours de retard

order_id	customer_id	order_purchase_timestamp	order_estimated_delivery_date	order_delivered_customer_date
7e708aed151d6a8601ce8f2ea712bf4	033fab69968b0d69099d64423831a236	2018-06-02 18:37:14	2018-07-13 00:00:00	2018-10-17 13:22:46
450cb96c63e1e5b49d34f223f67976d2	27ae7c8a8fc20ce80d96f01b6f19961b	2018-05-21 06:48:46	2018-06-27 00:00:00	2018-10-11 16:41:14
b2997e1d7061605e9285496c58b1d1fb	9e83d47684eb1a58b1c31830f5de10ac	2018-07-30 09:08:06	2018-08-14 00:00:00	2018-10-02 00:18:50
a2b4be96b53022618030c17ed437604d	ffa87b4246c4848711afb512bd51f161	2018-07-22 09:54:03	2018-08-17 00:00:00	2018-09-27 02:24:33
7d09831e67caa193da82cfea3bee7aa5	1409b2945191b7aff1975ba2ce9918c5	2018-08-05 17:11:44	2018-08-20 00:00:00	2018-09-25 00:47:25
f23681a0ffdb8051c674707c7e912ef	7930549f156eea2b01b0fc2fd323063	2018-07-15 02:11:15	2018-08-06 00:00:00	2018-09-21 23:46:29
1e7d25f611e794f9614dd3e10a8596e7	8be45a1114ff0e79615f7b8189aec7df	2018-08-01 19:43:06	2018-08-23 00:00:00	2018-09-21 15:55:02
4af2fb154881f350d8696f7f7a7f80d3	7c71fa0871e272a25eccc52af09595	2018-07-23 10:22:26	2018-08-13 00:00:00	2018-09-20 16:08:33
1b3190b2dfa9d789e1f14c05b647a14a	d306426abe5ca15e5b645e4462dc7b	2018-02-23 14:57:35	2018-03-15 00:00:00	2018-09-19 23:24:07
4505ac3759da6b9c7d79a80d29ab3bb	a3587bee339b45240b5a327d933509b	2018-08-06 14:32:27	2018-08-17 00:00:00	2018-09-19 16:44:44
ae213a9f84c777fb7a31e8c0f09fd30c	cac4524e1e6714ef3fd324fc0f86cfb	2018-07-03 22:13:48	2018-07-31 00:00:00	2018-09-19 15:46:39
0096668e5b0b8e9657a6f7209a4e58b4	b6fbaad4eb3a0794111683dca122610	2018-07-11 13:33:21	2018-07-19 00:00:00	2018-09-17 14:42:46
238652e39c5fd89a8fd44776f532501	fc041ede47154c40f55455e20c1a1954	2018-07-25 09:15:28	2018-08-10 00:00:00	2018-09-17 14:37:09
84db939b1ab9686533e9a06a0354beb6	c0789eee49fe7d5d93d5e412c14181ce	2018-07-20 13:14:25	2018-08-16 00:00:00	2018-09-17 13:49:04
6325af88a0611fc357055cb87dce11e	bfc2488928300e9b18e0d96637b8404	2018-07-18 13:40:31	2018-08-21 00:00:00	2018-09-17 13:48:28
84869ba3df14629b57ca40c491a842e6	8000d8c2201ad0577d5f459c6325ccdc	2018-07-27 13:50:48	2018-08-13 00:00:00	2018-09-14 17:11:19
c005c973843746a08a6ea826a4ce0c0	71ac72b29860fdac58666426bbe6b4ba	2018-07-29 18:24:11	2018-08-21 00:00:00	2018-09-13 21:22:27
1f0e3e7a13d98443333a705e4ea42148	31a8654d467ab5c0d784a273272c460f	2018-08-15 07:19		
6d0940a8f5fba47562b14cd97df6da	548692bdc6e3683ff306ac9d8418d6	2018-08-10 00:17		
7797e37c568b84182c813b9b2492b384	0411ba9ee6c7697c60b2e05345557958	2018-06-25 16:44		
3794be706c3088573f54b492768e7689	67a2903f301a0840437ec01ee05d292	2018-08-06 17:19		
83329d0539a9b3f0870f5db119bee0d6	c888151a54dd6f97054f6501441e421	2018-08-14 17:49		
fd5771079a027230000ab7992a902f0c	8216dbb58369d7e43239401a4a9e4f0b	2018-08-21 09:52		
72652c482c51119a0a338edba23e0027	8c89ab30610f6af8d8724259d0c5aa80	2018-08-09 10:00		
e10c080e6f7f8136087836cadcd26199	6f58e369991106fe684c3c436450894	2018-08-02 13:38		

total_deliveries

367

Vendeurs Performants (avec plus de 100 000 Reals de chiffre d'affaires.)

seller_id	seller_city	seller_state	total_revenue
4869f7a5dfa277a7dca6462dcf3b52b2	guariba	SP	247007.06000000023
7c67e1448b00f6e969d365cea6b010ab	itaquaquecetuba	SP	237806.68999999983
4a3ca9315b744ce9f8e9374361493884	ibitinga	SP	231220.4300000005
53243585a1d6dc2643021fd1853d8905	lauro de freitas	BA	230797.02000000037
fa1c13f2614d7b5c4749cbc52fecda94	sumare	SP	200833.49999999985
da8622b14eb17ae2831f4ac5b9dab84a	piracicaba	SP	184706.7799999995
7e93a43ef30c4f03f38b393420bc753a	barueri	SP	171973.55000000013
1025f0e2d44d7041d6cf58b6550e0bfa	sao paulo	SP	171924.96000000002
7a67c85e85bb2ce8582c35f2203ad736	sao paulo	SP	160278.52000000016
955fee9216a65b617aa5c0531780ce60	sao paulo	SP	156606.47999999937
6560211a19b47992c3666cc44a7e94c0	sao paulo	SP	148050.94000000067
1f50f920176fa81dab994f9023523100	sao jose do rio preto	SP	141712.94000000056
46dc3b2cc0980fb8ec44634e21d2718e	rio de janeiro	RJ	134162.8799999997
a1043baf4d71df536d0c462352beb48	ilicinea	MG	130412.74000000012
620c87c171fb2a6dd6e8bb4dec959fc6	petropolis	RJ	126278.18000000024
cc419e0650a3c5ba77189a1882b7556a	santo andre	SP	125936.98999999925
5dceca129747e92ff8ef7a997dc4f8ca	santa barbara d'oeste	SP	124702.92000000001
7d13fc1a5225358621be4086e1eb0964	ribeirao preto	SP	120934.46999999962
3d871de0142ce09b7081e2b9d1733cb1	campo limpo paulista	SP	115515.65000000027

Nouveaux Vendeurs (moins de 3 mois) Engagés (ayant vendu plus de 30 produits.)

seller_id	first_sale_date	total_products_sold
240b9776d844d37535668549a396af32	2018-07-17 13:48:59	36
81f89e42267213cb94da7ddc301651da	2018-08-08 12:45:12	52
d13e50eaa47b4cbe9eb81465865d8cfc	2018-08-04 09:09:37	69

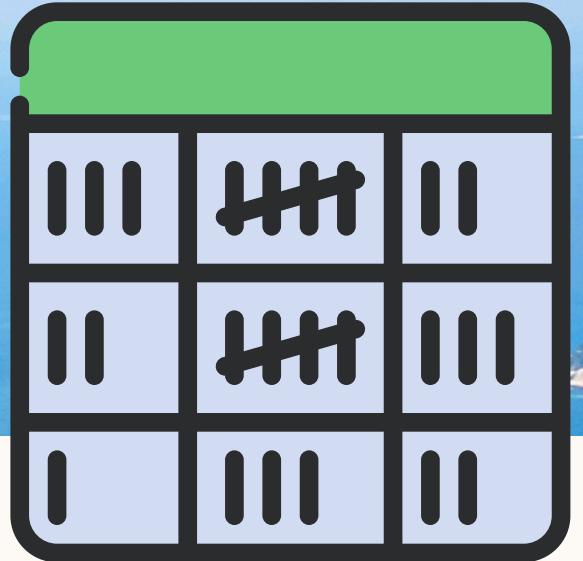
Zones avec Mauvais Scores (Pire score moyen et plus de 30 commandes.)

postal_code	average_review_score	total_orders
22753	2.8085106382978724	47
22770	3.135135135135	37
22793	3.2333333333333334	90
21321	3.2777777777777777	36
22780	3.3513513513513	37

EXPLORATION

sélection & transformation

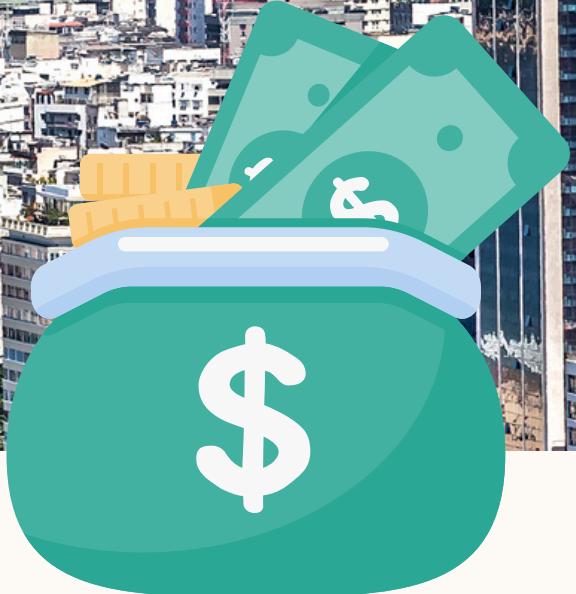




FREQUENCES

- **Agrégation** : Compte le nombre de commandes par client.
- **Filtrage** : Inclut uniquement les commandes avec le statut 'delivered'.
- **Jointure** : Relie les tables orders et customers via customer_id.
- **Groupement** : Groupement par customer_unique_id pour compter les commandes par client, puis groupement par nombre de commandes.
- **Ordre** : Trie les résultats par nombre de commandes.

nbre_de_commandes	effectif
1	90557
2	2573
3	181
4	28
5	9
6	5
7	3
9	1
15	1



MONTANTS

Sélection et comparaison

- **payment_value** de **order_pymts**
- **price + freight_value** (**prix**) de **order_items**

	order_id	payment_value	prix
	b81ef226f3fe1789b1e8b2acac839d17	99.33	99.33
	a9810da82917af2d9aefd1278f1dcfa0	24.39	24.39
	25e8ea4e93396b6fa0d3dd708e76c1bd	65.71	65.71000000000001
	ba78997921bbcdc1373bb41e913ab953	107.78	107.78
	42fdf880ba16b47b59251dd489d4441a	128.45	128.45



RECENCE

- Extraction dynamique de la date de la dernière commande
- Durée en jours écoulés depuis la commande

	order_id	order_purchase_timestamp	days_since_order
	e481f51cbdc54678b7cc49136f2d6af7	2017-10-02 10:56:33	380.2734375
	53cdb2fc8bc7dce0b6741e2150273451	2018-07-24 20:41:37	84.86714120395482
	47770eb9100c2d0c44946d9cf07ec65d	2018-08-08 08:38:49	70.3690856480971
	949d5b44dbf5de918fe9c16f97b45f8a	2017-11-18 19:28:06	332.9181944443844
	ad21c59c0840e6cb83a9ceb5573f8159	2018-02-13 21:18:39	245.84142361115664

Statut possible d'une commande

AUTRES VARIABLES

Nombre de produits par commandes

number_of_products	frequency
1	88863
2	7516
3	1322
4	505
5	204
6	198
7	22
8	8
9	3
10	8
11	4
12	5
13	1
14	2
15	2
20	2
21	1

Emplacement vendeurs

COUNT(DISTINCT seller_state)
23

Moyens de paiements

payment_type
credit_card
boleto
voucher
debit_card
not_defined

order_status
delivered
invoiced
shipped
processing
unavailable
canceled
created
approved

Nombre de clients

NombreDeClients
99441

NombreDeClientsUniques
96096

DATAFRAME DE TRAVAIL

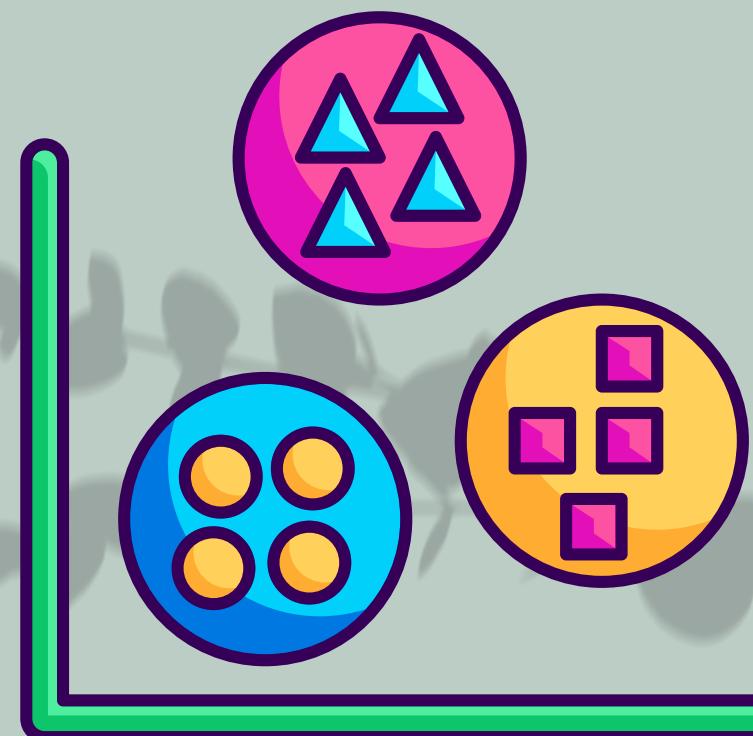
- Requête SQL
- Export vers dataframe Pandas
- Traitement - Sélection
- Nettoyage Outliers / Montant

	customer_unique_id	frequence	montant	moyenne_review_score	recence
0	0000366f3b9a7992bf8c76cfdf3221e2	1	141.90	5.0	160.273507
1	0000b849f77a49e4a4ce2b2a4ca5be3f	1	27.19	4.0	163.263090
2	0000f46a3911fa3c0805444483337064	1	86.22	3.0	585.850868
3	0000f6ccb0745a6a4b88665a16c9f078	1	43.62	4.0	369.875428
4	0004aac84e0df4da2b147fca70cf8255	1	196.89	5.0	336.905972
...
92748	ffffbf87b7a1a6fa8b03f081c5f51a201	1	167.32	5.0	293.787234
92750	fffea47cd6d3cc0a88bd621562a9d061	1	84.58	4.0	310.890532
92751	ffff371b4d645b6ecea244b27531430a	1	112.46	5.0	617.070162
92752	ffff5962728ec6157033ef9805bacc48	1	133.69	5.0	168.092095
92753	fffffd2657e2aad2907e67c3e9daecbeb	1	71.56	5.0	532.883021

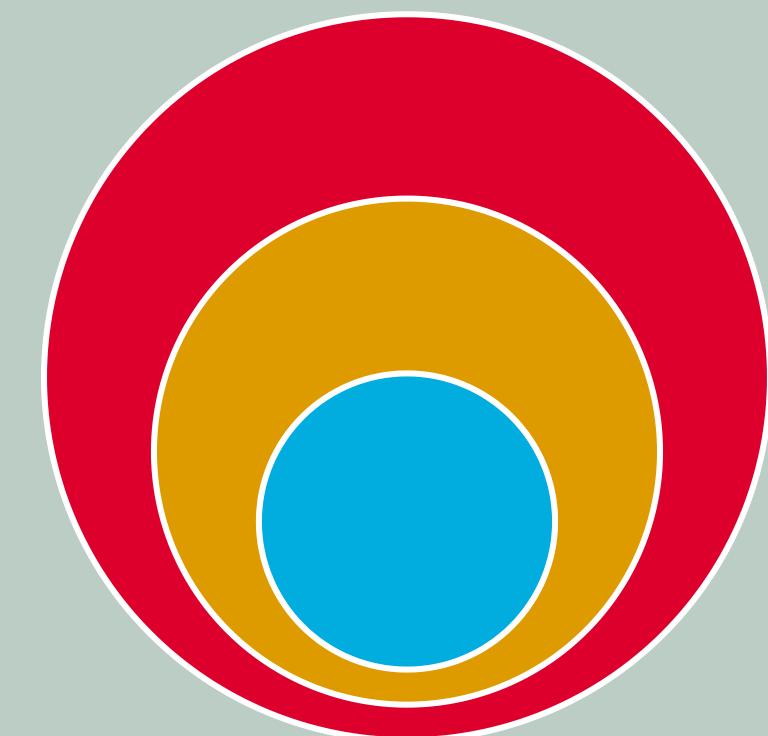
85368 rows × 5 columns

SEGMENTATION

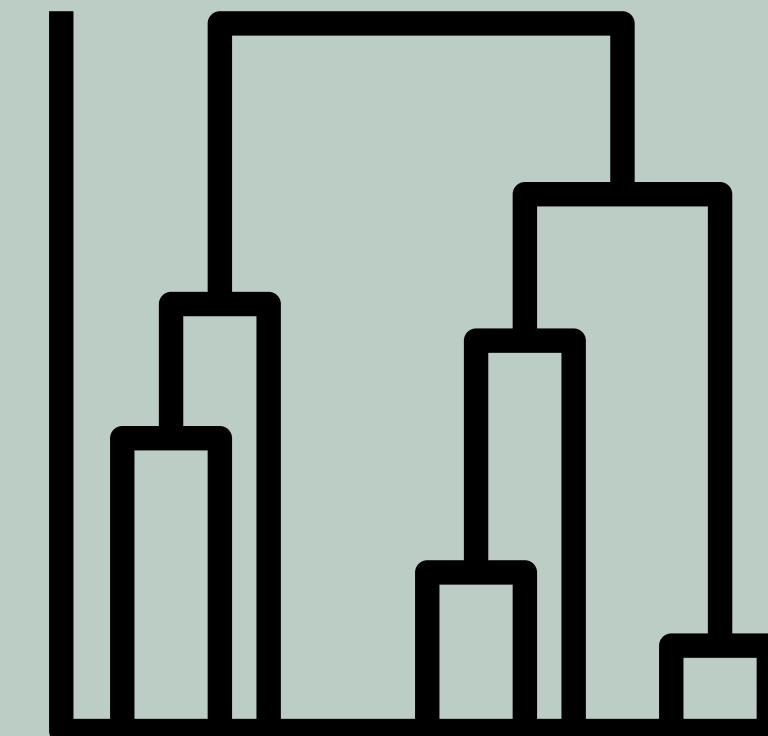
Essai et optimisation de différents modèles



K-means

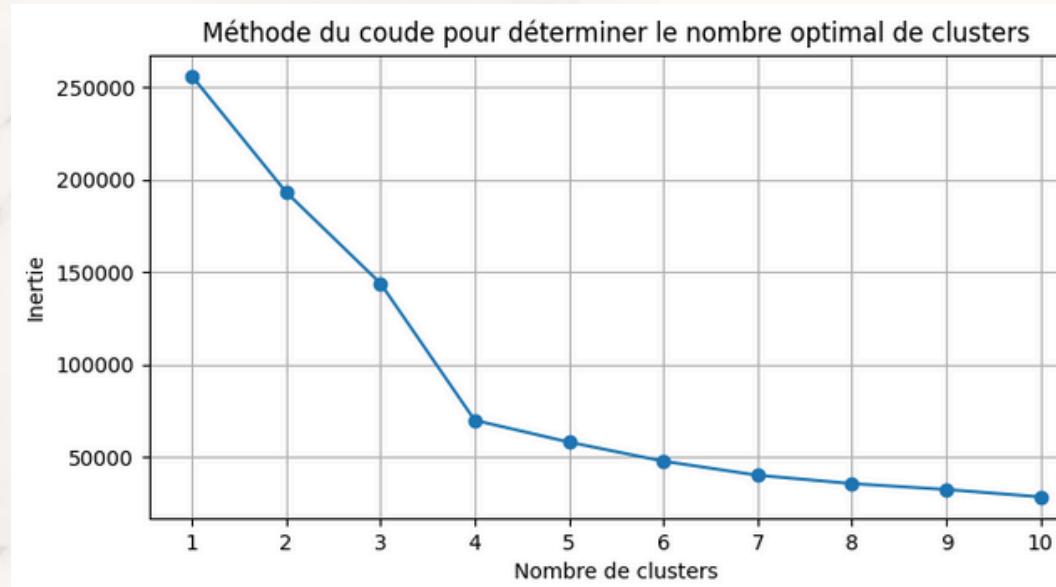


DBScan

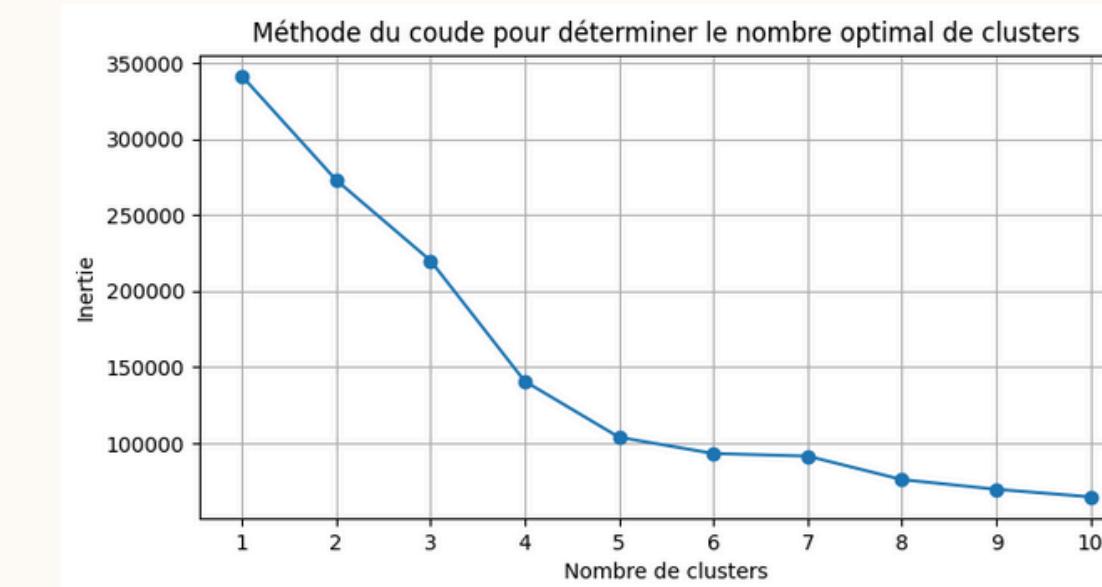


Agglomerative

Optimisation des hyperparamètres



k-means 3 features

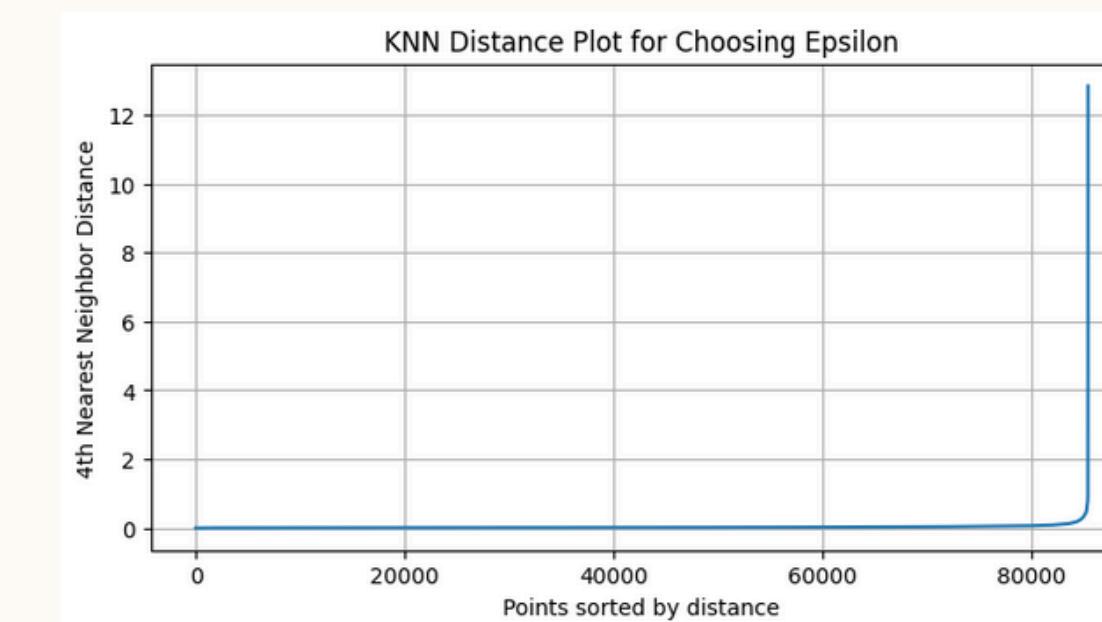


k-means 4 features

k-means : n_cluster, init seed

DBScan : epsilon / min_samples

+ Normalisation



Cohésion : Mesure de la similarité des points dans le même cluster.

Séparation : Mesure de la différence entre les points de différents clusters.

Échelle : Le score varie de -1 à 1.

Interprétation :

1 : Les clusters sont bien séparés et distincts.

0 : Les clusters se chevauchent.

-1 : Les points sont mal assignés, plus proches des points d'un autre cluster.

k-means 3 features - 4 clusters : score correct ≈ 0.4

k-means 4 features - 5 clusters : score correct ≈ 0.4

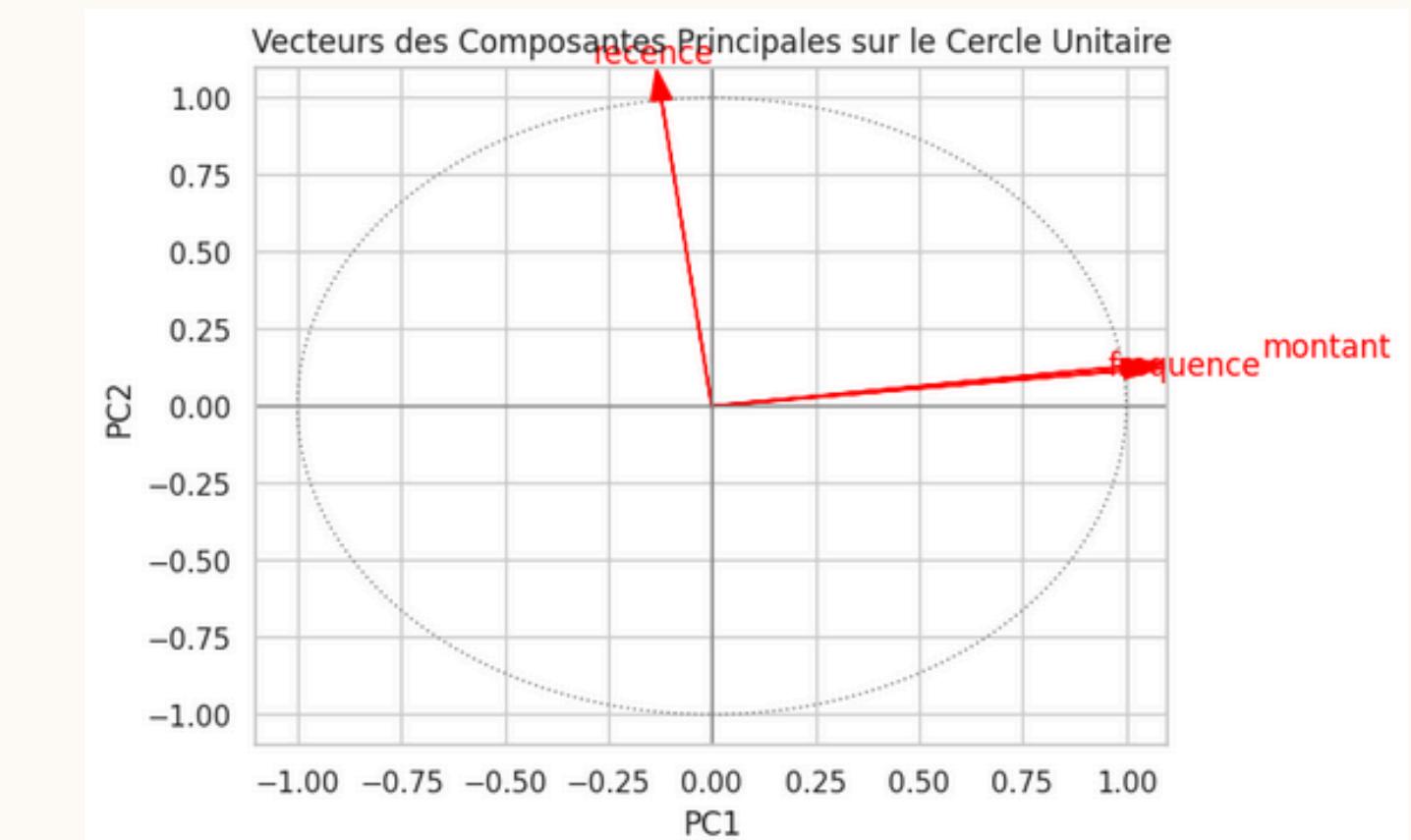
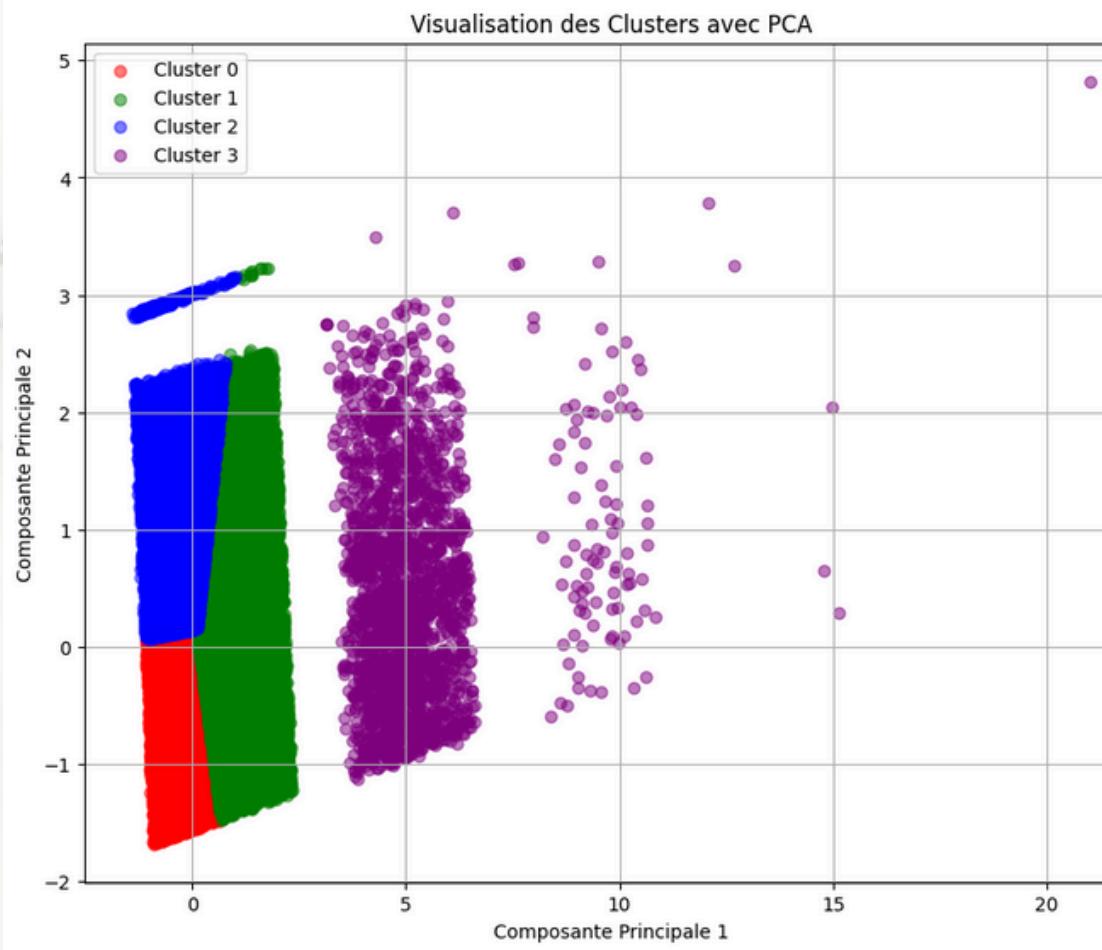
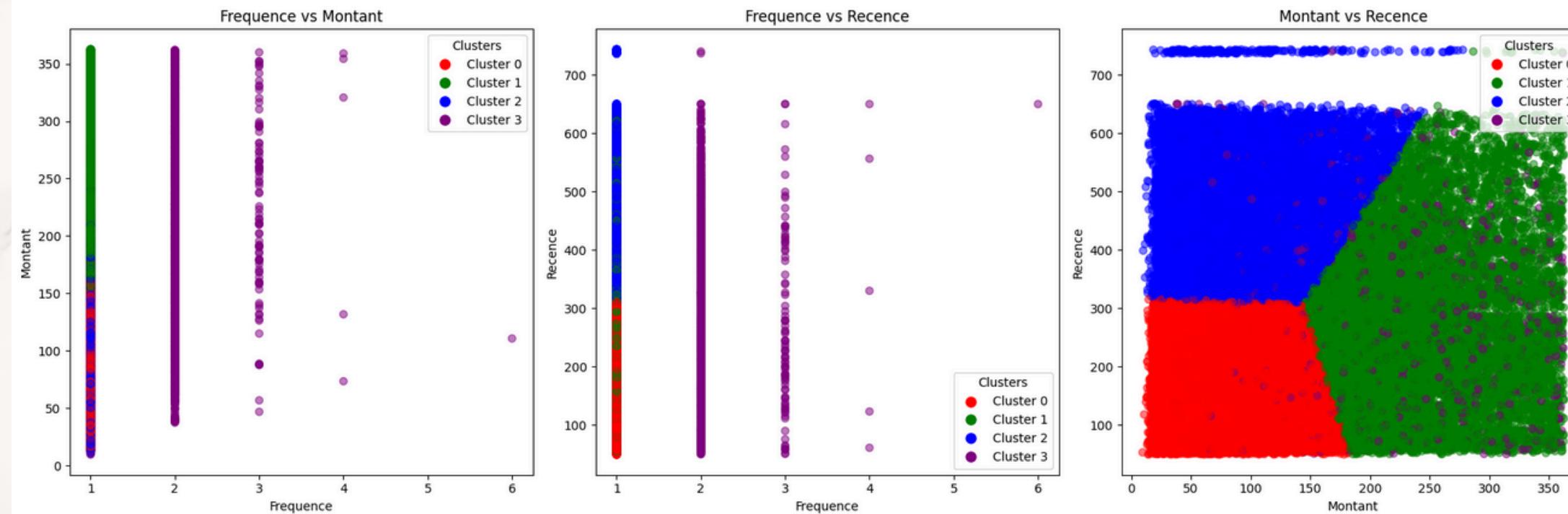
Agglomerative : score moyen ≈ 0.2

DBScan : score mauvais ≈ -0.02

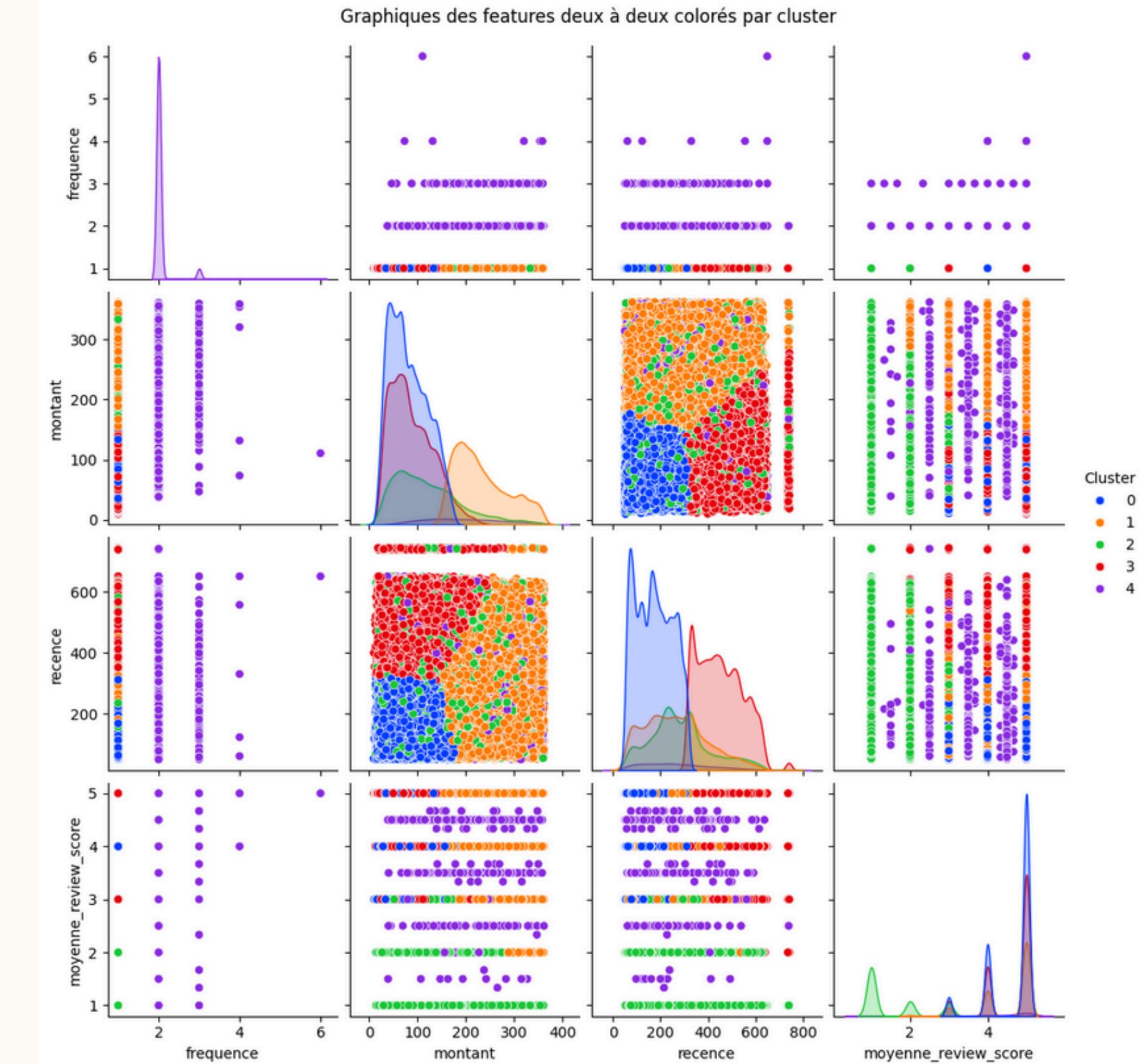
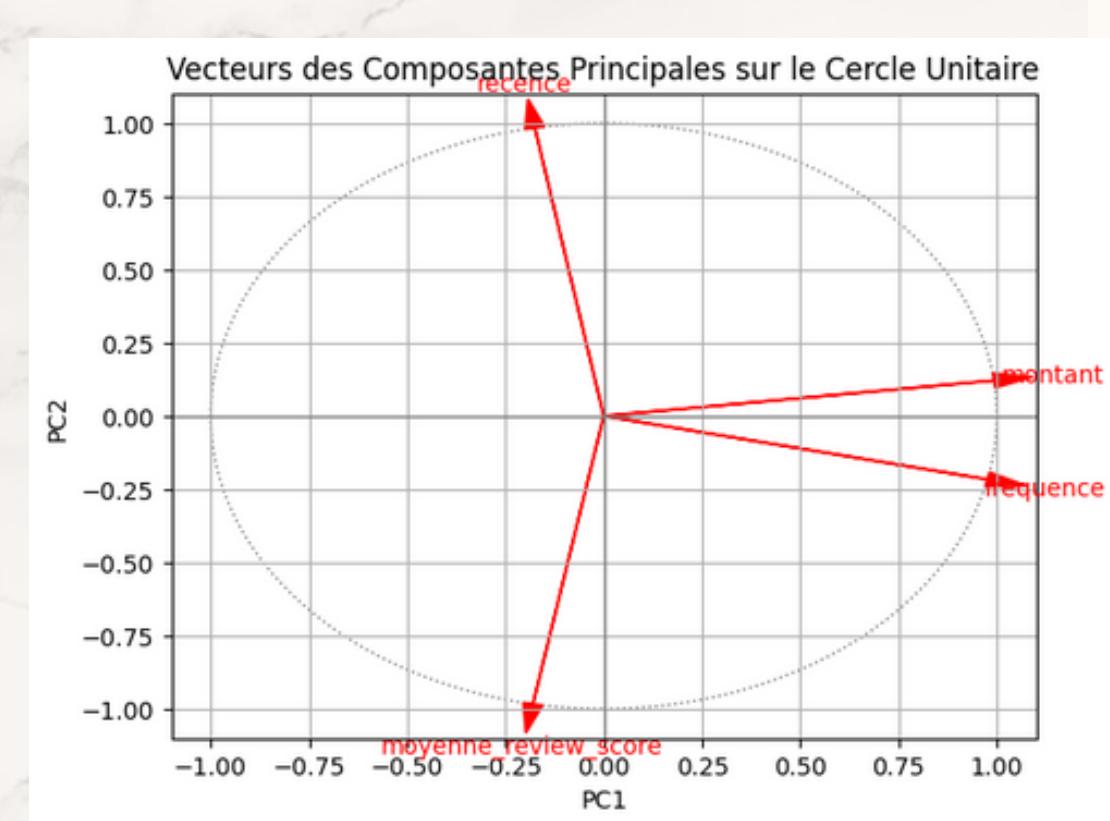
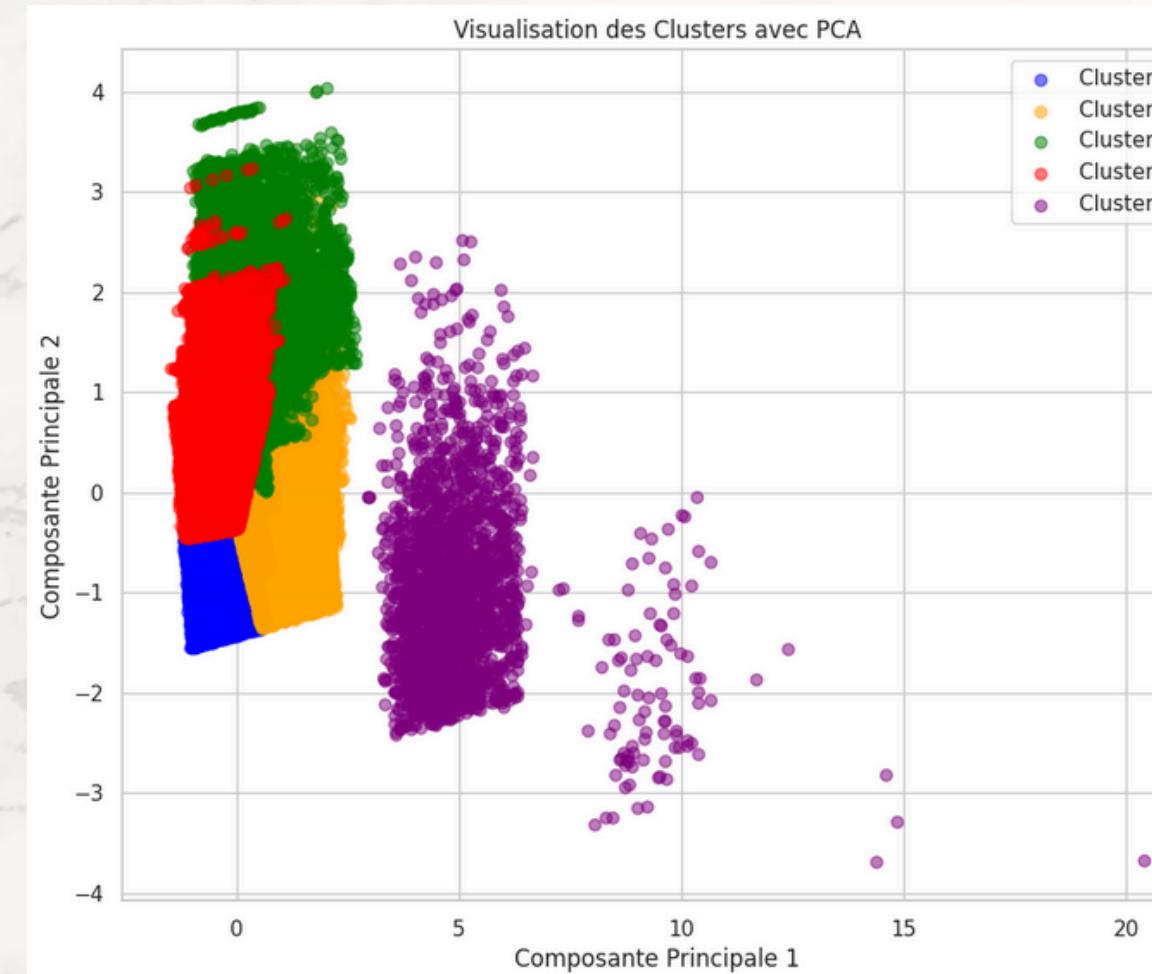
QUALIFICATION DES CLUSTERS

silhouette_score

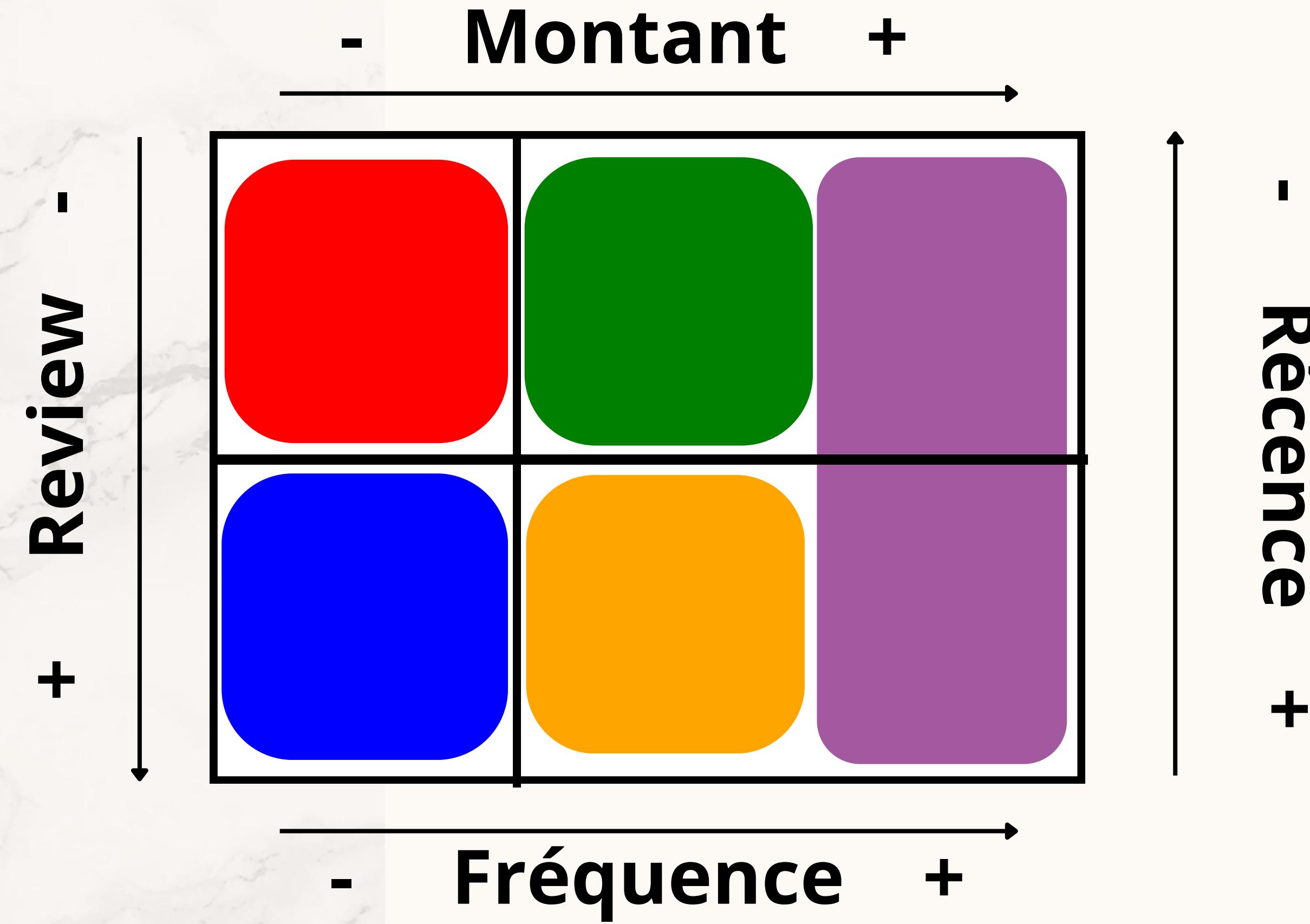
Visualisation des résultats - kmeans RFM



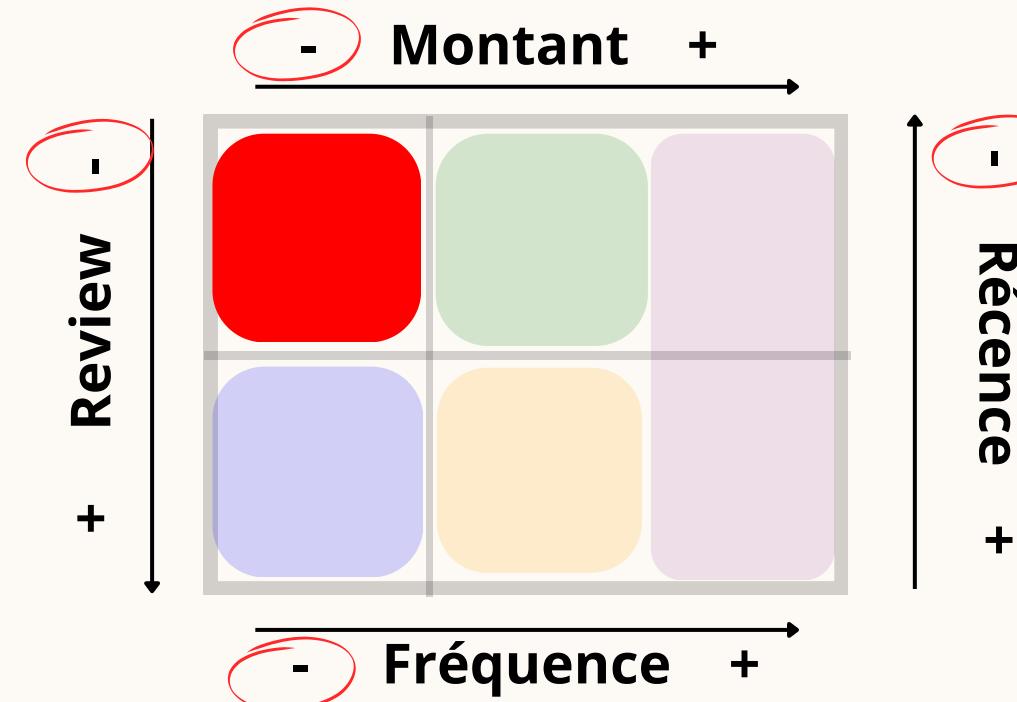
Visualisation des résultats - kmeans RFM+Review



Interprétation métier



Interprétation métier



Cluster Rouge

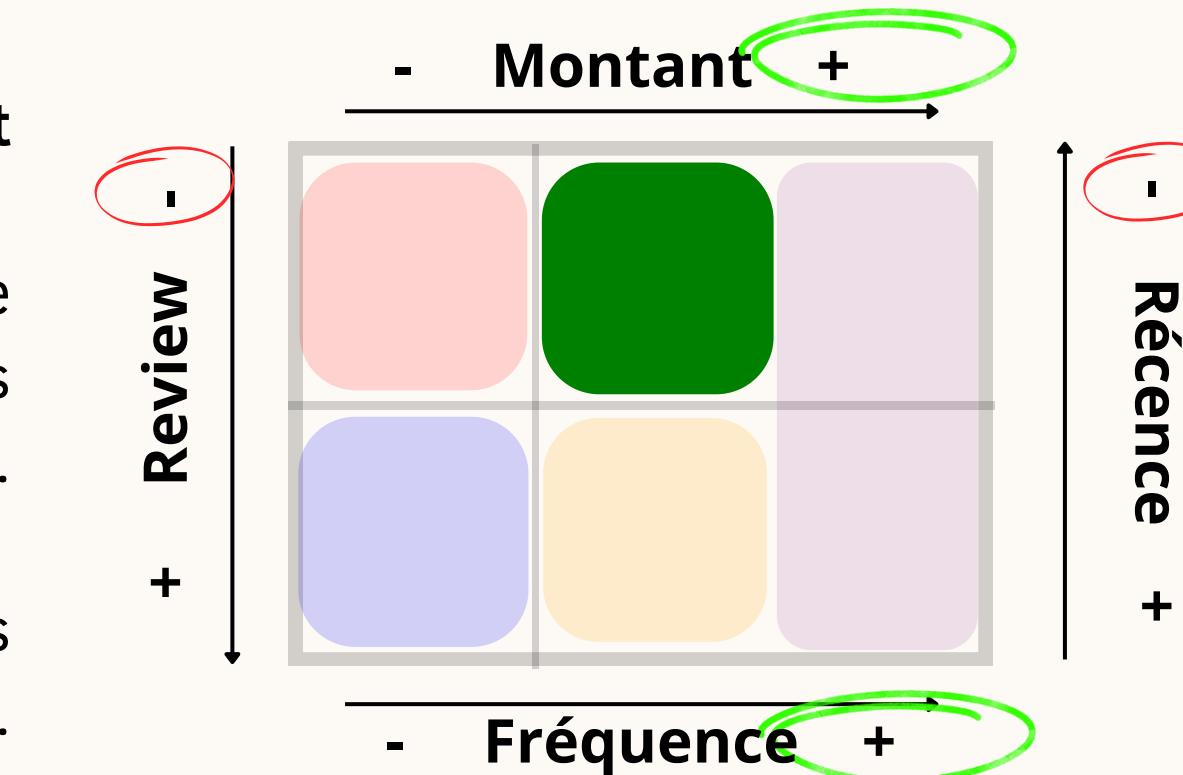
Clients insatisfaits ou désengagés, achetant rarement, dépensant peu, et laissant des avis négatifs.

Nécessitent des interventions pour améliorer leur expérience et potentiellement les réengager.

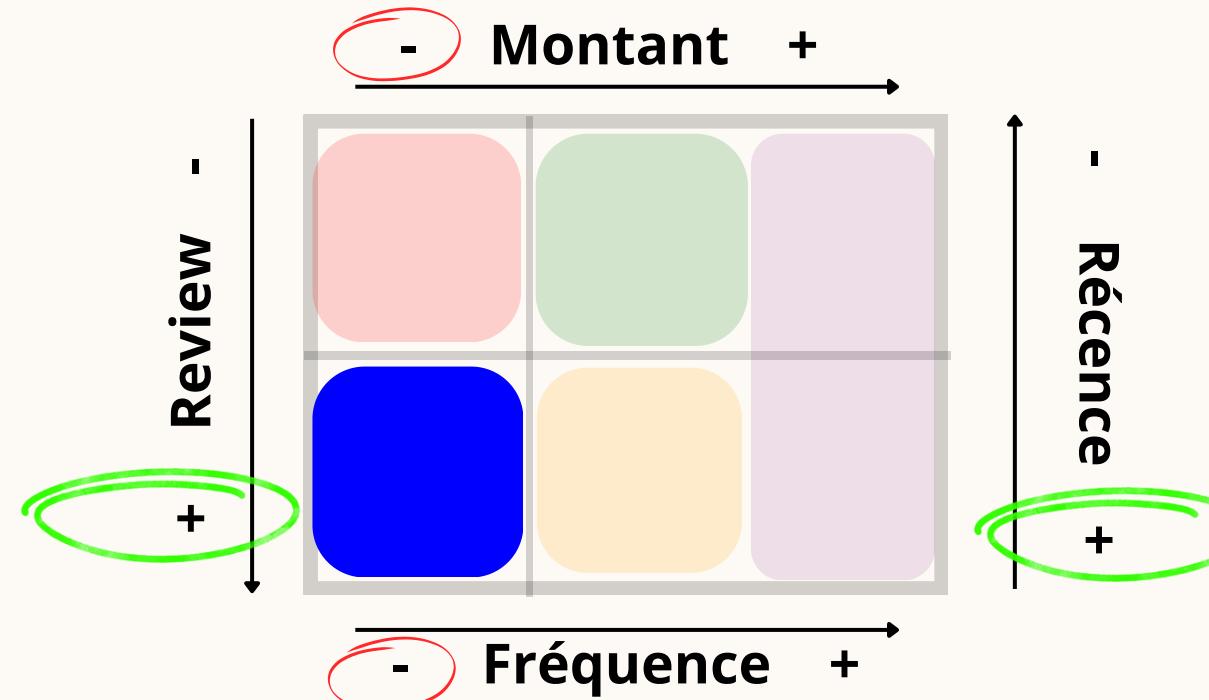
Cluster Vert

Clients fidèles avec des dépenses et une fréquence d'achat élevées, mais avec des avis négatifs et des achats non récents.

Important à récupérer en adressant leurs préoccupations spécifiques.



Interprétation métier



Cluster Bleu

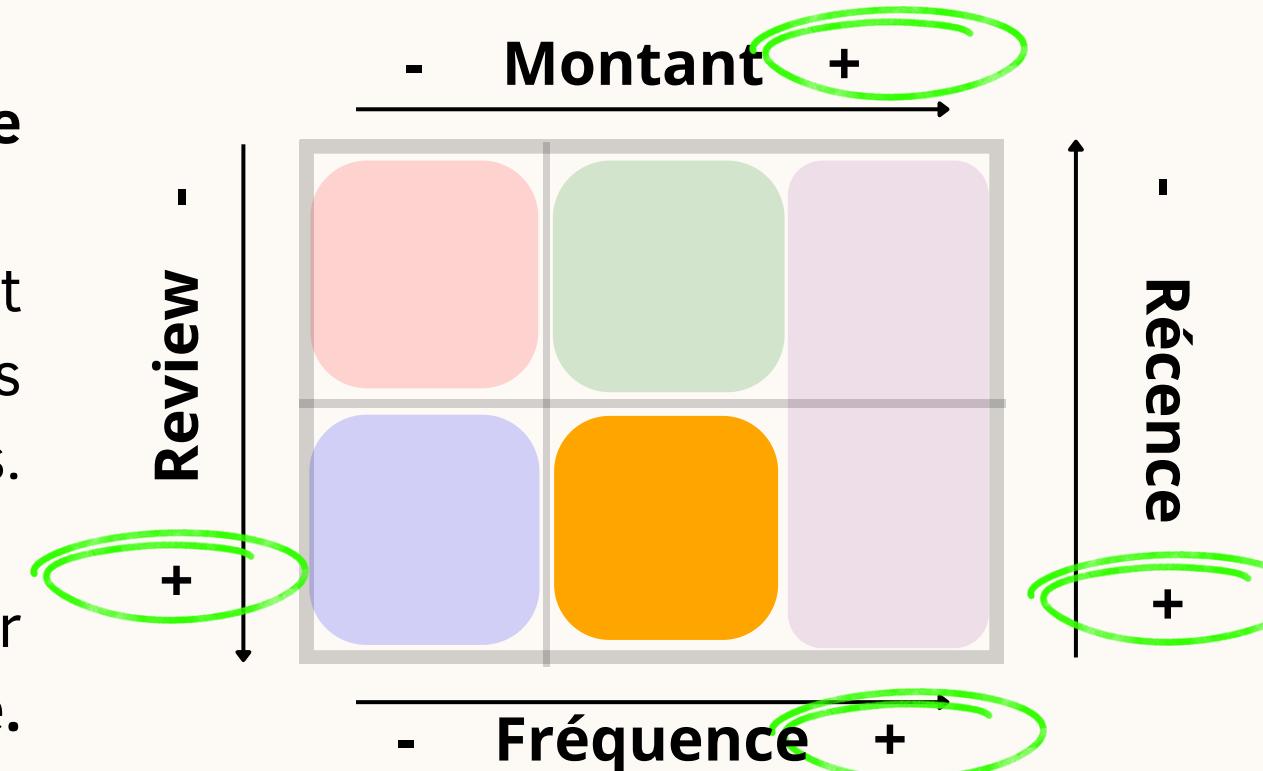
Nouveaux clients ou clients potentiels avec un faible panier moyen, achetant rarement mais récemment, et laissant des avis positifs.

Nécessitent des stratégies pour augmenter leur fréquence d'achat et leur montant moyen.

Cluster Orange

Clients idéaux, très satisfaits et fidèles, dépensant beaucoup, achetant fréquemment, avec des achats récents et des avis positifs.

Devraient être valorisés et récompensés pour leur fidélité.

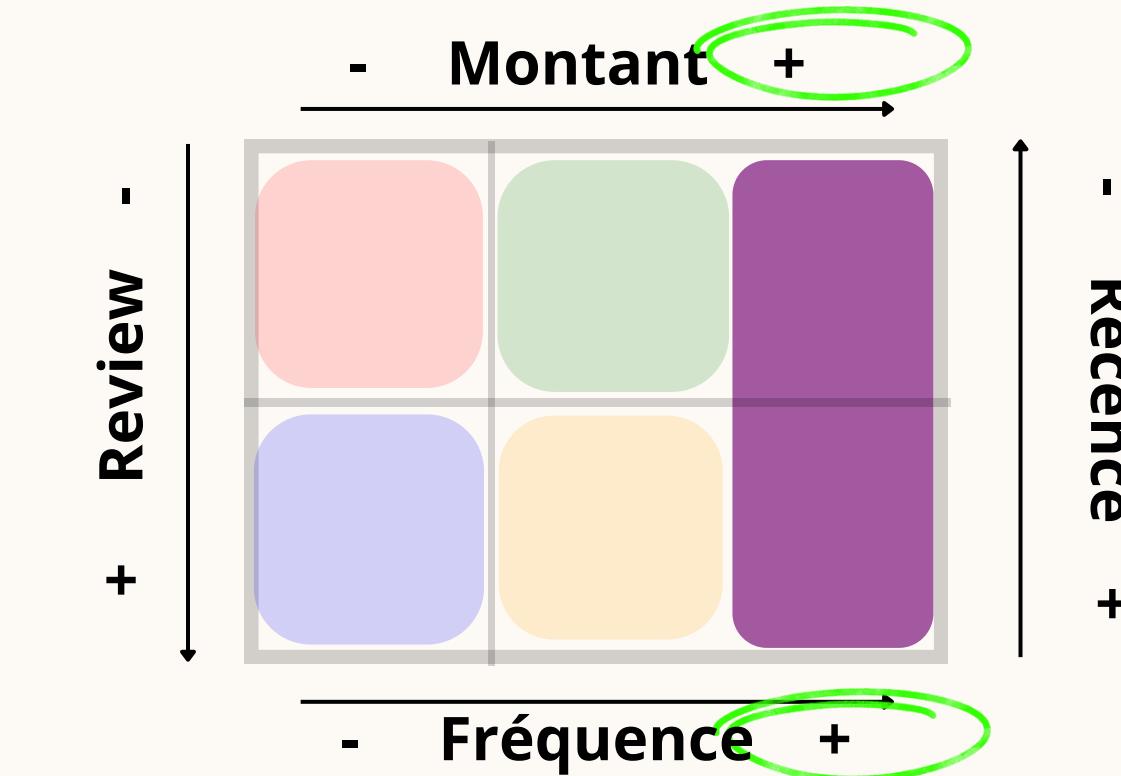


Interprétation métier

Cluster Violet

Clients réguliers avec des dépenses et une fréquence d'achat élevées, mais avec des avis et une récence variés.

Nécessitent une attention particulière pour comprendre et améliorer leur satisfaction.





Comment l'ajout de clients
impacte le modèle

MAINTENANCE DU MODÈLE

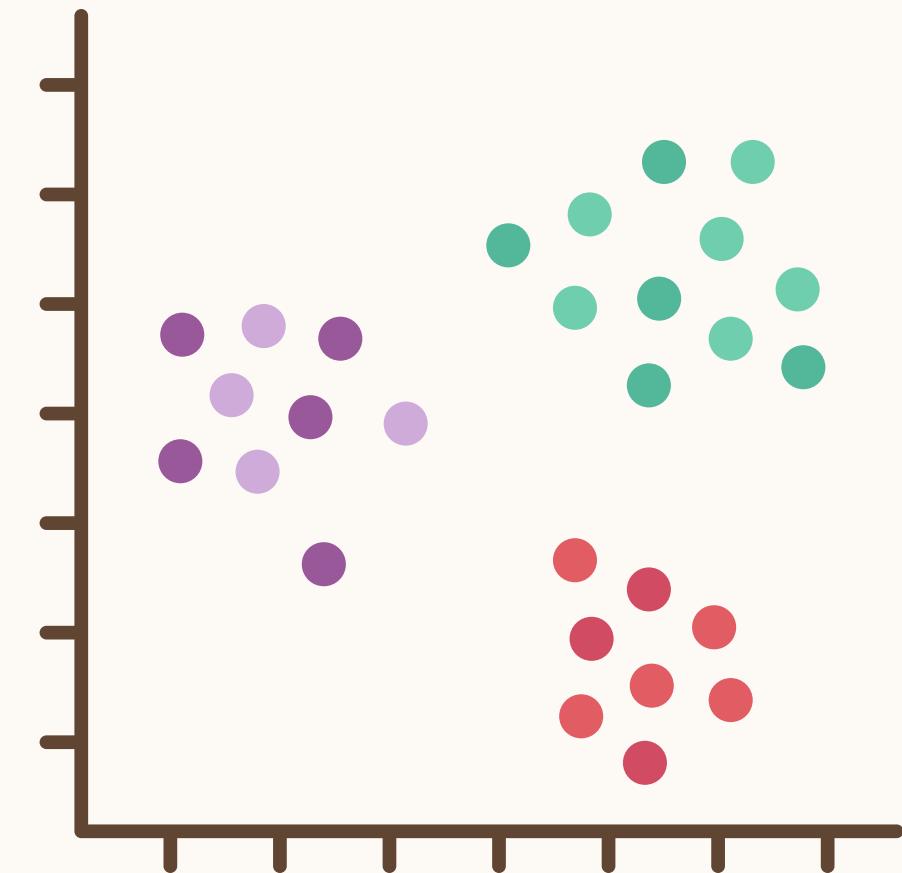
étude de la stabilité dans le temps

Principe

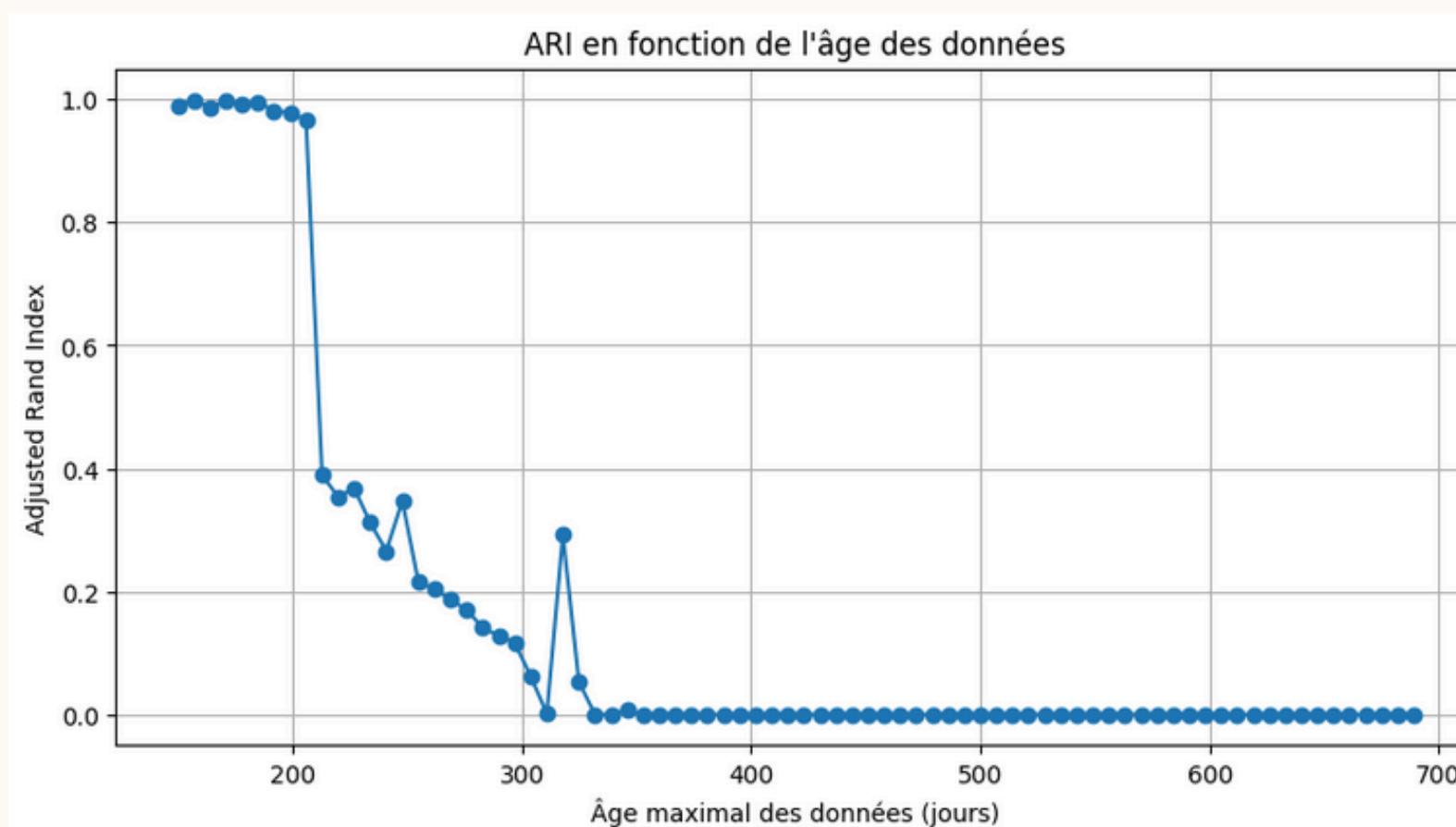
- Clustering Initial : Chaque client appartient à un cluster
- Évolution : Ajout de nouveaux clients, recalcul des clusters

L'ARI évalue :

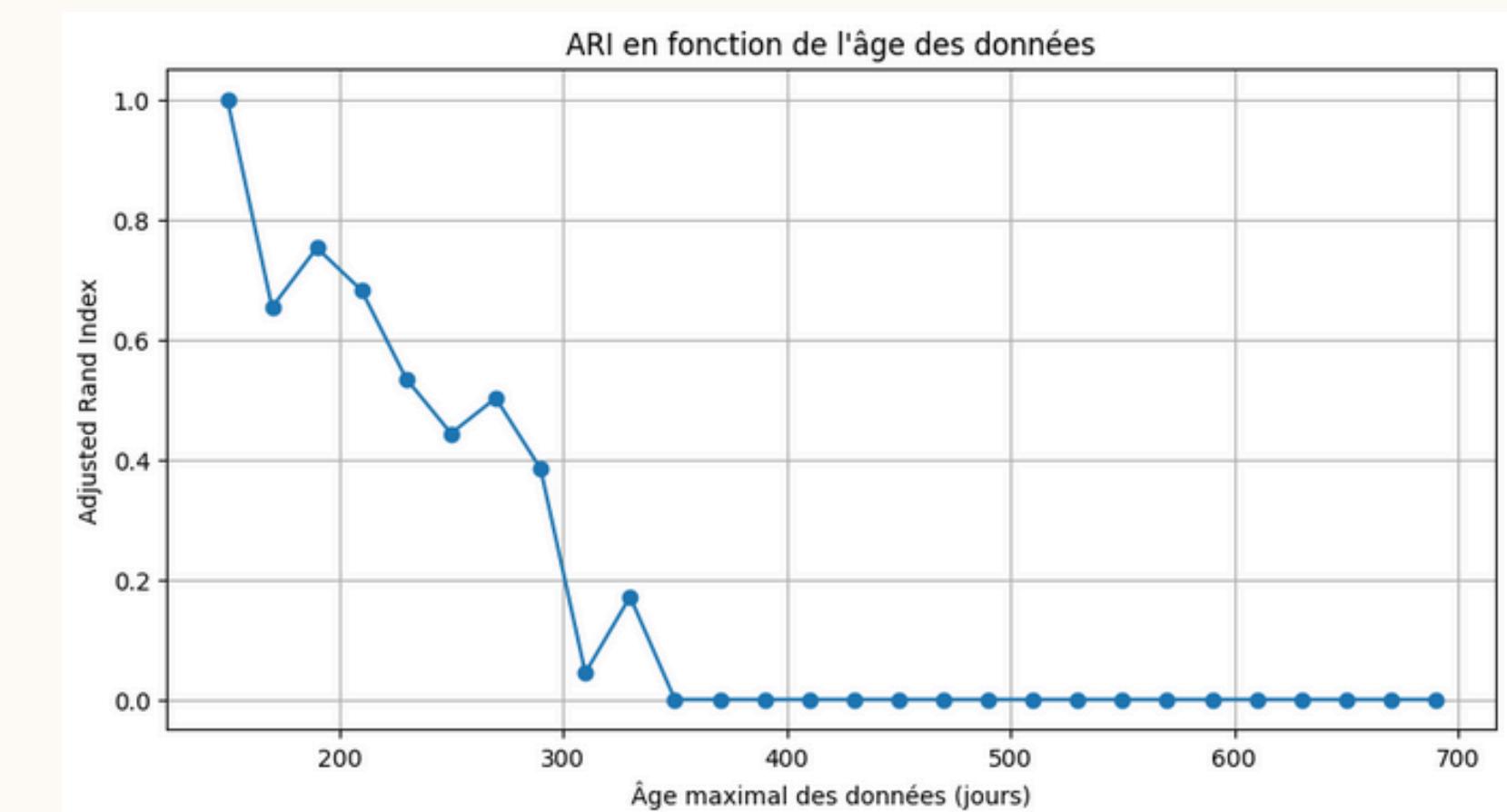
1. Stabilité Intra-Cluster :
 - Clients initialement dans le même cluster doivent rester ensemble
 2. Stabilité Inter-Cluster :
 - Clients initialement dans des clusters différents doivent rester séparés
- ARI = 1 : Clusters stables
 - ARI bas : Changements significatifs dans les clusters



Illustration

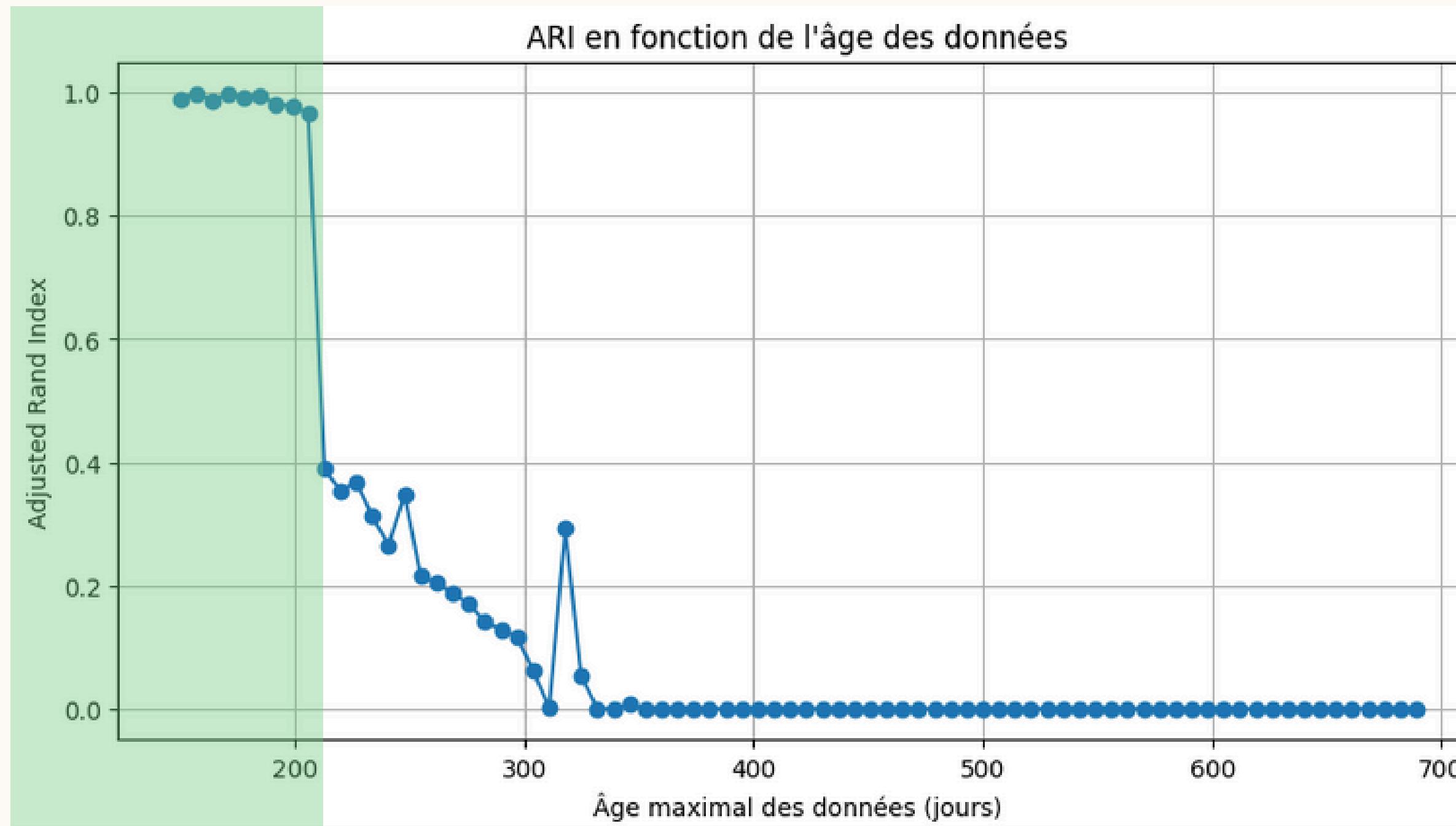


ARI sur Kmeans 4 clusters



ARI sur Kmeans 5 clusters

ARI en fonction de l'âge des données



60 jours de stabilité

Merci