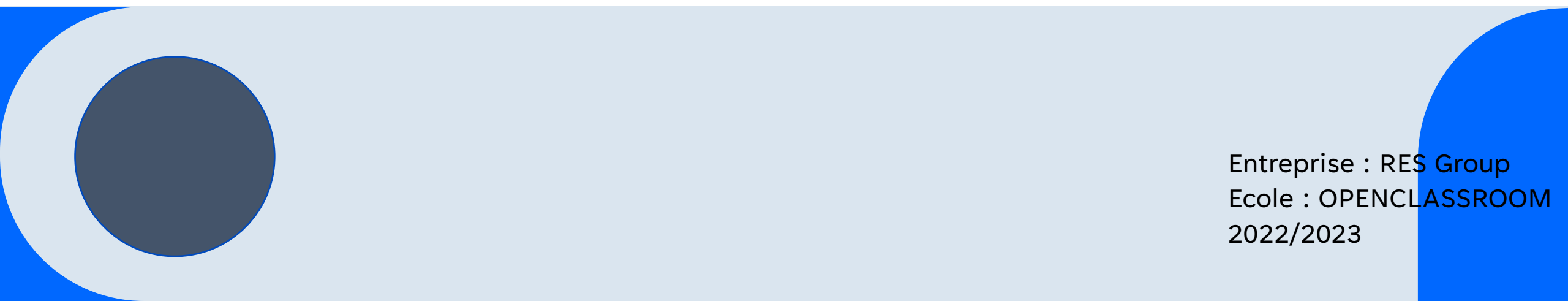




PROJET 5 : SEGMENTER LES CLIENTS D'UN SITE E-COMMERCE

Divine Tulomba



Entreprise : RES Group
Ecole : OPENCLASSROOM
2022/2023



PLAN

Présentation de la problématique

Analyse exploratoire et feature engineering

Pistes de Modélisation

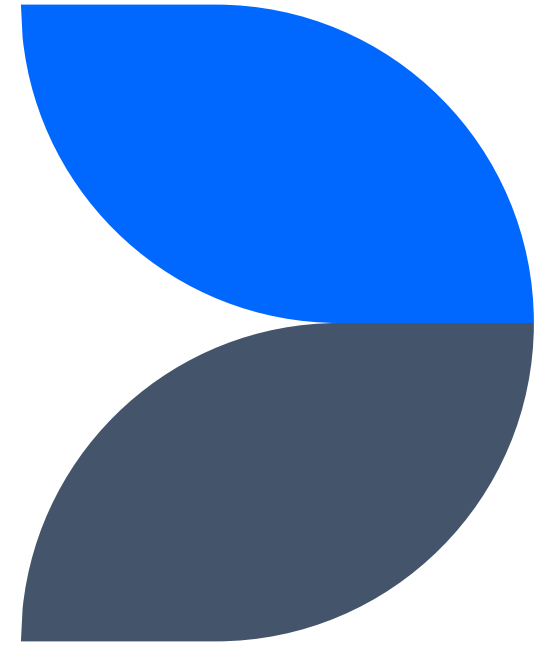
Présentation du modèle final

I - Problématique

Rappel de la problématique

Interprétation

Pistes de recherche envisagées



Présentation et contexte

Olist une entreprise brésilienne qui propose une solution de vente sur les marketplaces en ligne.

❖ Objectifs :

- Fournir aux équipes d'e-commerce une segmentation des clients pour les campagnes de communication
- Comprendre les différents types d'utilisateurs
- Fournir une description actionnable de la segmentation
- Faire une proposition pour la maintenance

Pistes de recherche envisagées

Exploration des données et choix de features adaptées

Problème de classification non supervisée avec des modèles spécifiques

Comparaison des résultats et choix d'un modèle

Descriptions des clusters, puis mis en place du score ARI

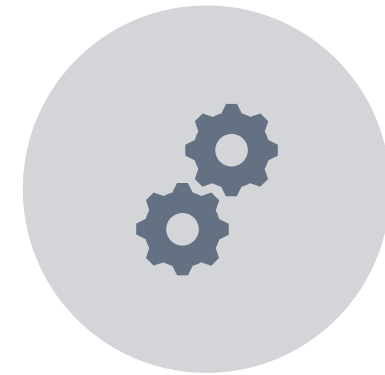
Préparation du jeu de données



PRÉSENTATION DE LA
BASE



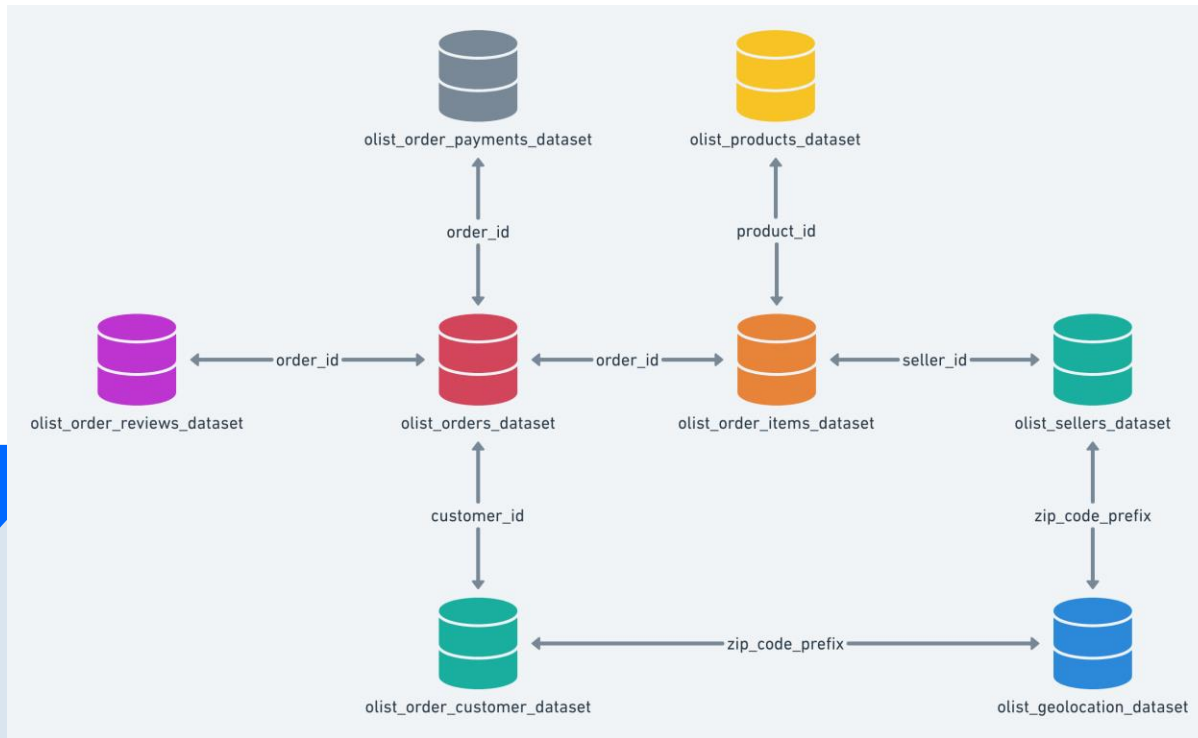
EXPLORATION &
ANALYSE



FEATURE
ENGINEERING

Présentation de la base de données

Données réparties en 9 tables :



Répartition de la base de données

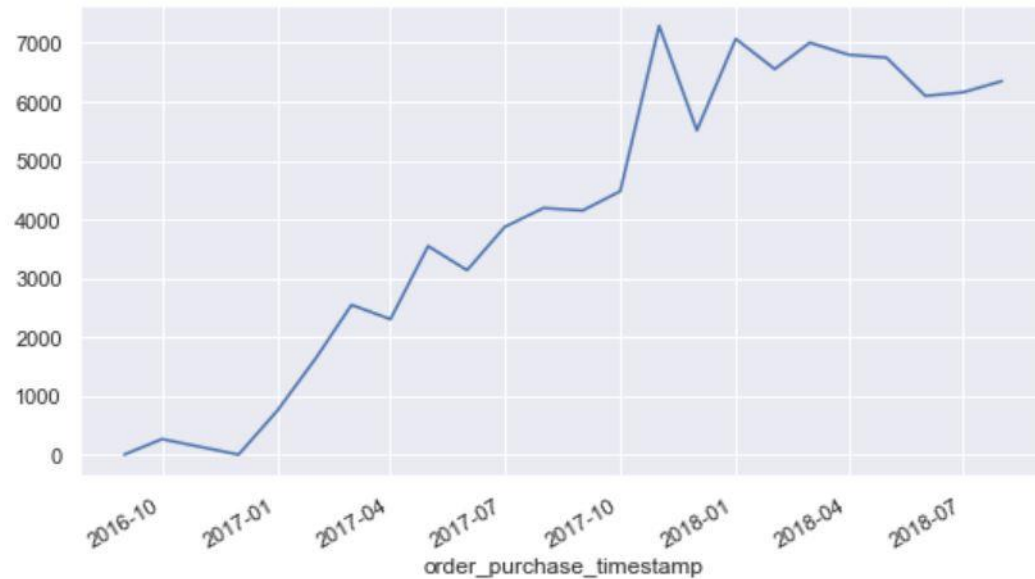
tables clients : customers , Geolocation

tables de commandes: orders, order_items, order_payments, order_reviews

tables de produits : products, Categories_en

Analyse exploratoire

Evolution des achats entre 2016 et 2018



Ventes en constante évolution depuis 2016

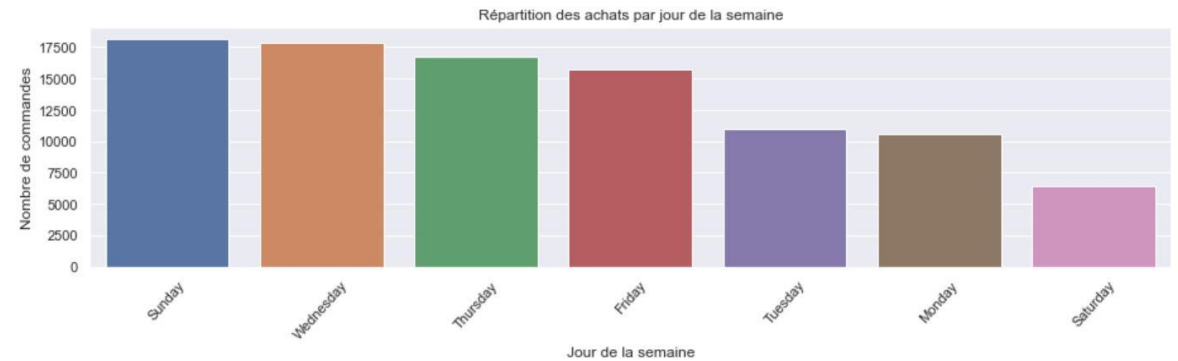
Pic de vente en 2017

Analyse exploratoire

Pic de vente en novembre 2017

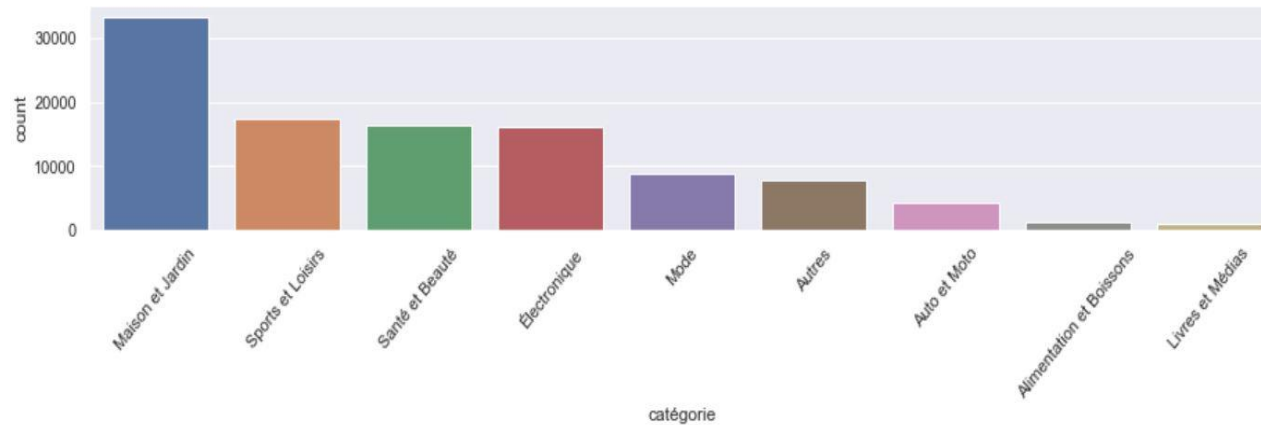


Les clients achètent le plus le mercredi et le dimanche

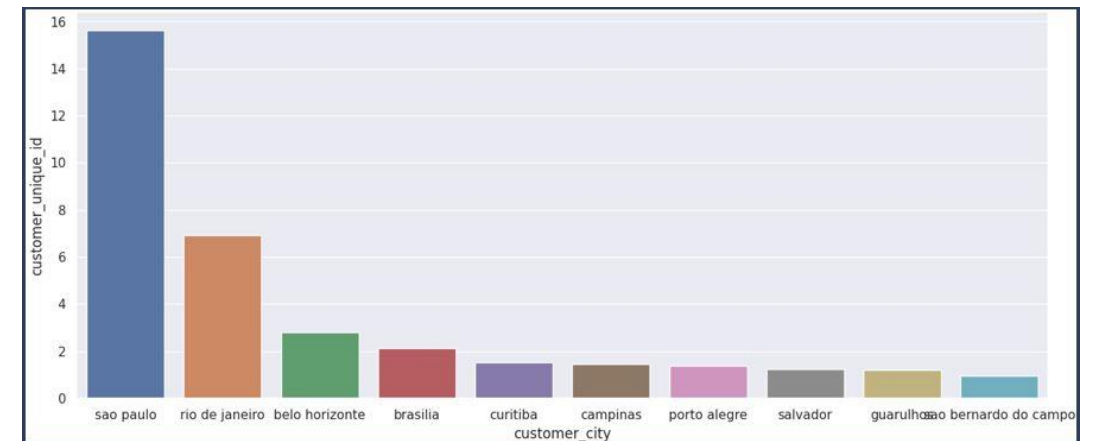


Analyse exploratoire

Catégories de produits les plus commandés

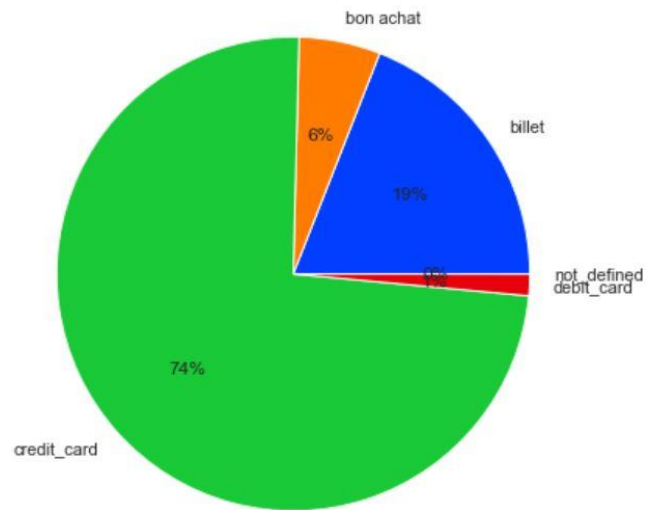


Villes où se trouvent les clients

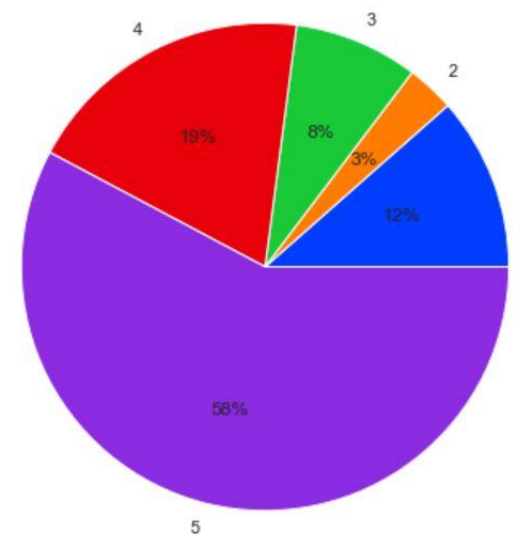


Analyse exploratoire

Répartition des moyens de paiement



Répartition des notes données par les clients



Feature engineering

Nous avons créé des nouvelles variables en vue de créer la segmentation RFM

Variables	Description
customer_unique_id	Identifiant client
Recency	temps écoulé depuis le dernier achat du client
Frequency	le nombre d'achats effectués par le client
monetary	le montant total dépensé par le client
Note_com	Note donné par le client
Frais_livraison	Frais de livraison moyen
Temps moyen de livraison	Temps moyen de livraison de la commande

Présentation du dataset final

	customer_unique_id	recency	frequency	monetary	note_com	frais_livraison	temps_moyen_livraison	Alimentation et Boissons	Auto et Moto	Autres	Livres et Médias	Maison et Jardin	Mode	Santé et Beauté	Sports et Loisirs	Électronique
7019	13467e882eb3a701826435ee4424f2bd	273	1	134.83	5.0	17.53	NaN	0	1	0	0	0	0	0	0	0
8516	175378436e2978be55b8f4316bce4811	62	1	54.97	5.0	9.07	NaN	0	0	0	1	0	0	0	0	0
10132	1bd06a0c0df8b23dacfd3725d2dc0bb9	58	1	158.07	5.0	19.07	NaN	0	0	0	0	0	1	0	0	0
17095	2f17c5b324ad603491521b279a9ff4de	70	1	354.24	5.0	25.24	NaN	0	0	1	0	0	0	0	0	0
21667	3bc508d482a402715be4d5cf4020cc81	58	1	158.07	5.0	19.07	NaN	0	0	0	0	0	1	0	0	0
74122	cce5e8188bf42ffb3bb5b18ff58f5965	82	1	120.12	1.0	9.13	NaN	0	0	0	0	0	0	0	0	1
77971	d77cf4be2654aa70ef150f8bfec076a6	460	1	194.00	5.0	15.00	NaN	0	0	0	0	0	0	0	1	0
85362	ebf7e0d43a78c81991a4c59c145c75db	58	1	204.62	5.0	15.63	NaN	0	0	0	0	0	0	0	1	0

Présentation du dataset final

	customer_unique_id	recency	frequency	monetary	note_com	frais_livraison	temps_moyen_livraison	Alimentation et Boissons	Auto et Moto	Autres	Livres et Médias	Maison et Jardin	Mode	Santé et Beauté	Sports et Loisirs	Électronique
7019	13467e882eb3a701826435ee4424f2bd	273	1	134.83	5.0	17.53	NaN	0	1	0	0	0	0	0	0	0
8516	175378436e2978be55b8f4316bce4811	62	1	54.97	5.0	9.07	NaN	0	0	0	1	0	0	0	0	0
10132	1bd06a0c0df8b23dacfd3725d2dc0bb9	58	1	158.07	5.0	19.07	NaN	0	0	0	0	0	1	0	0	0
17095	2f17c5b324ad603491521b279a9ff4de	70	1	354.24	5.0	25.24	NaN	0	0	1	0	0	0	0	0	0
21667	3bc508d482a402715be4d5cf4020cc81	58	1	158.07	5.0	19.07	NaN	0	0	0	0	0	1	0	0	0
74122	cce5e8188bf42ffb3bb5b18ff58f5965	82	1	120.12	1.0	9.13	NaN	0	0	0	0	0	0	0	0	1
77971	d77cf4be2654aa70ef150f8bfec076a6	460	1	194.00	5.0	15.00	NaN	0	0	0	0	0	0	0	1	0
85362	ebf7e0d43a78c81991a4c59c145c75db	58	1	204.62	5.0	15.63	NaN	0	0	0	0	0	0	0	1	0



	customer_unique_id	recency	frequency	monetary
	0000366f3b9a7992bf8c76cfd3221e2	111	1	141.90
	0000b849f77a49e4a4ce2b2a4ca5be3f	114	1	27.19
	0000f46a3911fa3c0805444483337064	536	1	86.22
	0000f6ccb0745a6a4b88665a16c9f078	320	1	43.62
	0004aac84e0df4da2b147fca70cf8255	287	1	196.89

	fffcf5a5ff07b0908bd4e2dbc735a684	446	2	4134.84

Présentation du dataset final

	customer_unique_id	recency	frequency	monetary	note_com	frais_livraison	temps_moyen_livraison	Alimentation et Boissons	Auto et Moto	Autres	Livres et Médias	Maison et Jardin	Mode	Santé et Beauté	Sports et Loisirs	Électronique
7019	13467e882eb3a701826435ee4424f2bd	273	1	134.83	5.0	17.53	NaN	0	1	0	0	0	0	0	0	0
8516	175378436e2978be55b8f4316bce4811	62	1	54.97	5.0	9.07	NaN	0	0	0	1	0	0	0	0	0
10132	1bd06a0c0df8b23dacfd3725d2dc0bb9	58	1	158.07	5.0	19.07	NaN	0	0	0	0	0	1	0	0	0
17095	2f17c5b324ad603491521b279a9ff4de	70	1	354.24	5.0	25.24	NaN	0	0	1	0	0	0	0	0	0
21667	3bc508d482a402715be4d5cf4020cc81	58	1	158.07	5.0	19.07	NaN	0	0	0	0	0	1	0	0	0
74122	cce5e8188bf42ffb3bb5b18ff58f5965	82	1	120.12	1.0	9.13	NaN	0	0	0	0	0	0	0	0	1
77971	d77cf4be2654aa70ef150f8bfec076a6	460	1	194.00	5.0	15.00	NaN	0	0	0	0	0	0	0	1	0
85362	ebf7e0d43a78c81991a4c59c145c75db	58	1	204.62	5.0	15.63	NaN	0	0	0	0	0	0	0	1	0

customer_unique_id	recency	frequency	monetary
0000366f3b9a7992bf8c76cfd3221e2	111	1	141.90
0000b849f77a49e4a4ce2b2a4ca5be3f	114	1	27.19
0000f46a3911fa3c0805444483337064	536	1	86.22
0000f6ccb0745a6a4b88665a16c9f078	320	1	43.62
0004aac84e0df4da2b147fca70cf8255	287	1	196.89
...
fffcf5a5ff07b0908bd4e2dbc735a684	446	2	4134.84

customer_unique_id	recency	frequency	monetary	note_com	frais_livraison	temps_moyen_livraison
0000366f3b9a7992bf8c76cfd3221e2	111	1	141.90	5.0	12.00	6.0
0000b849f77a49e4a4ce2b2a4ca5be3f	114	1	27.19	4.0	8.29	3.0
0000f46a3911fa3c0805444483337064	536	1	86.22	3.0	17.22	25.0
0000f6ccb0745a6a4b88665a16c9f078	320	1	43.62	4.0	17.63	20.0
0004aac84e0df4da2b147fca70cf8255	287	1	196.89	5.0	16.89	13.0
...
fffcf5a5ff07b0908bd4e2dbc735a684	446	2	4134.84	5.0	248.71	27.0

Modélisation

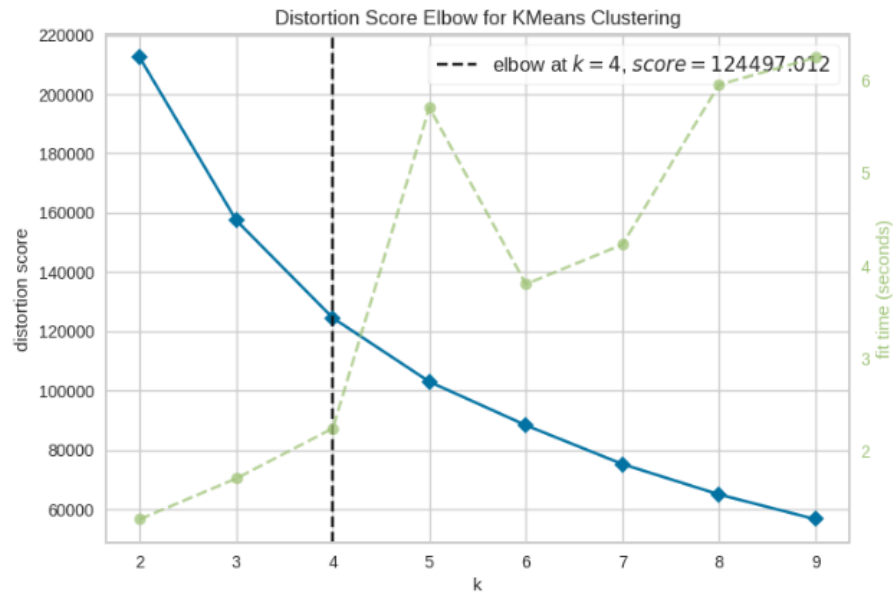
K-means

Agglomerative
clustering

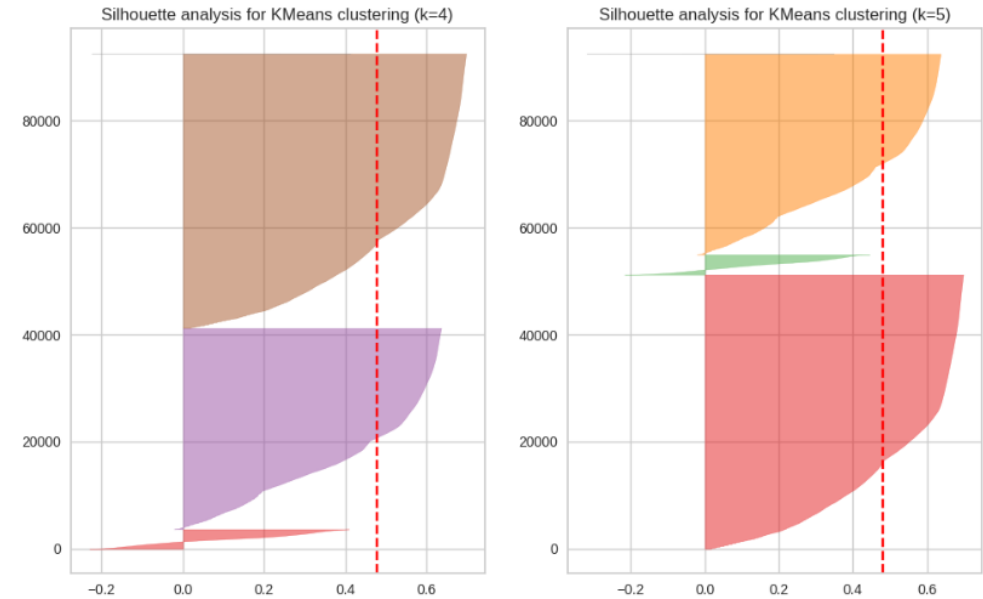
Dbscan

Kmeans : segment 1

Distortion Score Elbow

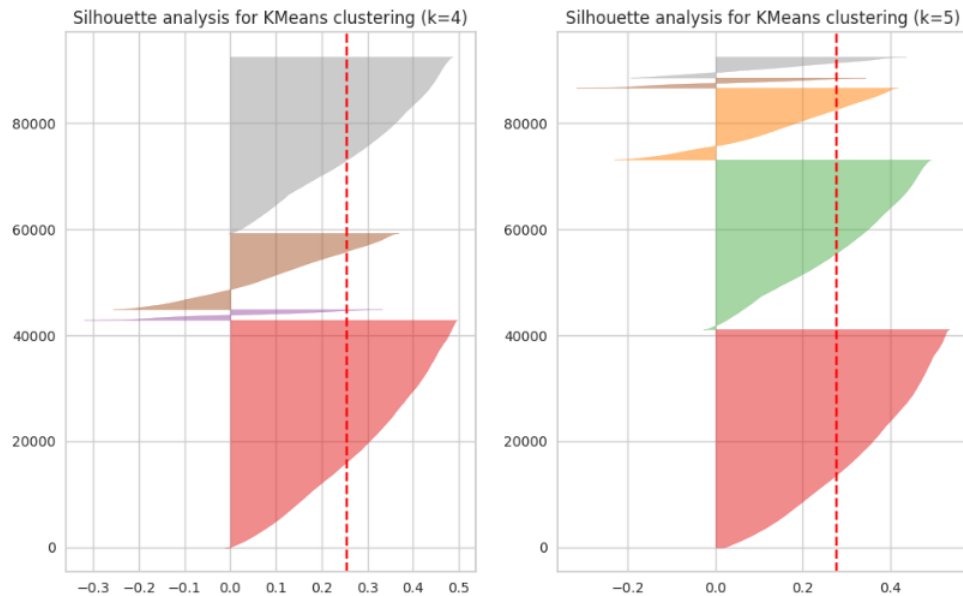


Score silhouette

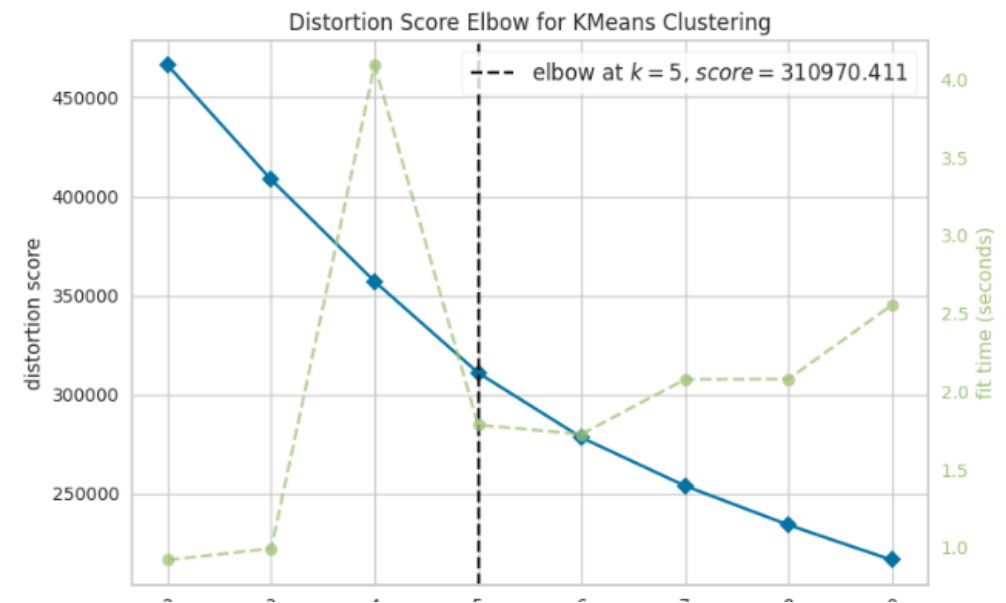


Kmeans : segment 2

Silhouette Score

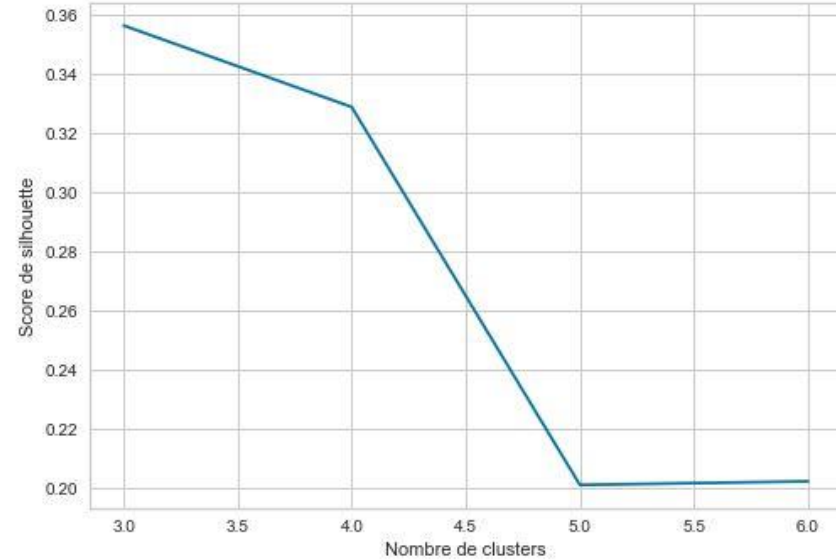


Distortion Score Elbow

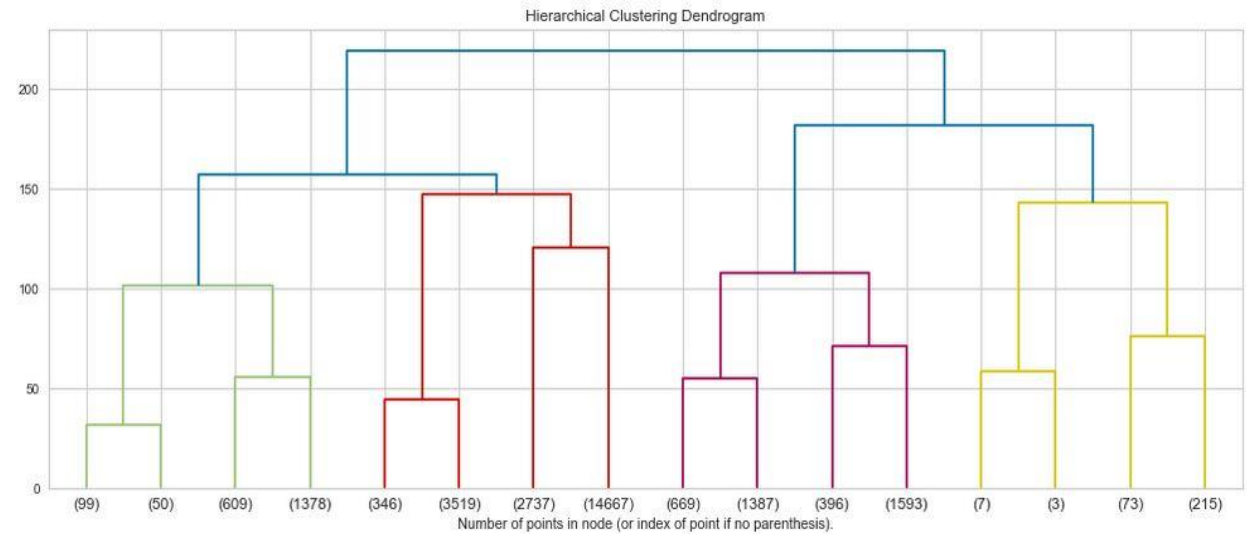


Agglomerative clustering : Segment 2

Score silhouette



Dendrogram



Dbscan : Segment 2

	eps	min sample	nb cluster	silhouette score
7	1.2	25	4	0.137127
6	1.1	25	4	0.135392
3	0.8	25	3	0.117826
5	1.0	25	5	0.115156
4	0.9	25	6	0.105448
2	0.7	25	7	0.049069
0	0.5	25	8	0.031685
1	0.6	25	9	0.030140

Clustering de qualité insuffisante.

	recency	frequency	montant	cluster	note_com	frais_livraison	temps_moyen_livraison
cluster							
-1	241.301038	3.412918	1666.301753	867	3.194642	49.157285	20.201845
0	236.873873	1.000000	141.072247	23294	4.218297	19.552327	11.877050
1	236.217376	2.000000	303.513640	2912	3.902816	17.389682	10.804258
2	238.627219	4.000000	508.882722	169	4.573964	15.221642	9.171598
3	236.136364	3.000000	473.509802	506	3.969697	16.220823	9.871542

4 clusters

Silhouette = 0.14

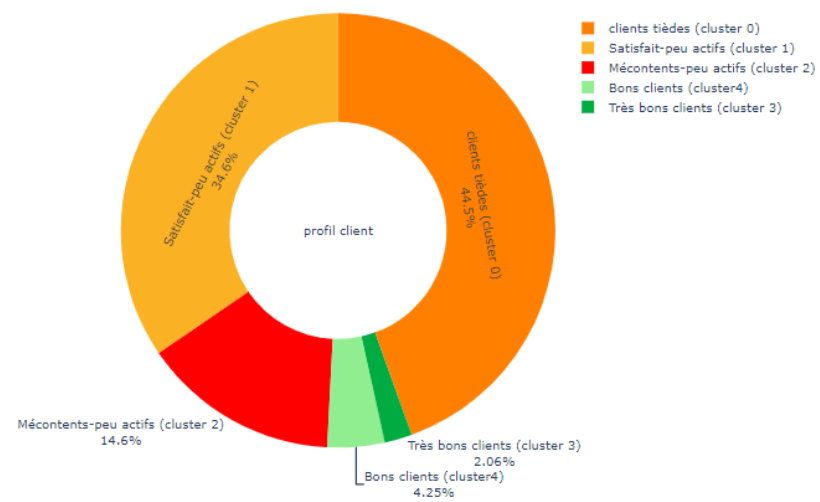
Modélisation et choix du modèle final

Modèle	Score silhouette	Nombre de clusters	Avantage/inconvénient	Retenu
K-means		5	Temps d'exécution modéré. Score silhouette correct. Séparation et interprétabilité des clusters satisfaisants.	Oui
Agglomerative clustering		4	Temps d'exécution trop long et entraînement possible qu'avec maximum 30% du dataset. Score silhouette pas meilleur que celui du kmeans.	Non
DBSCAN		5	Temps d'exécution trop long. Incapacité de séparer efficacement les clusters.	Non

Caractérisation des clusters

Algorithme utilisé : K-means

Caractérisation des clients



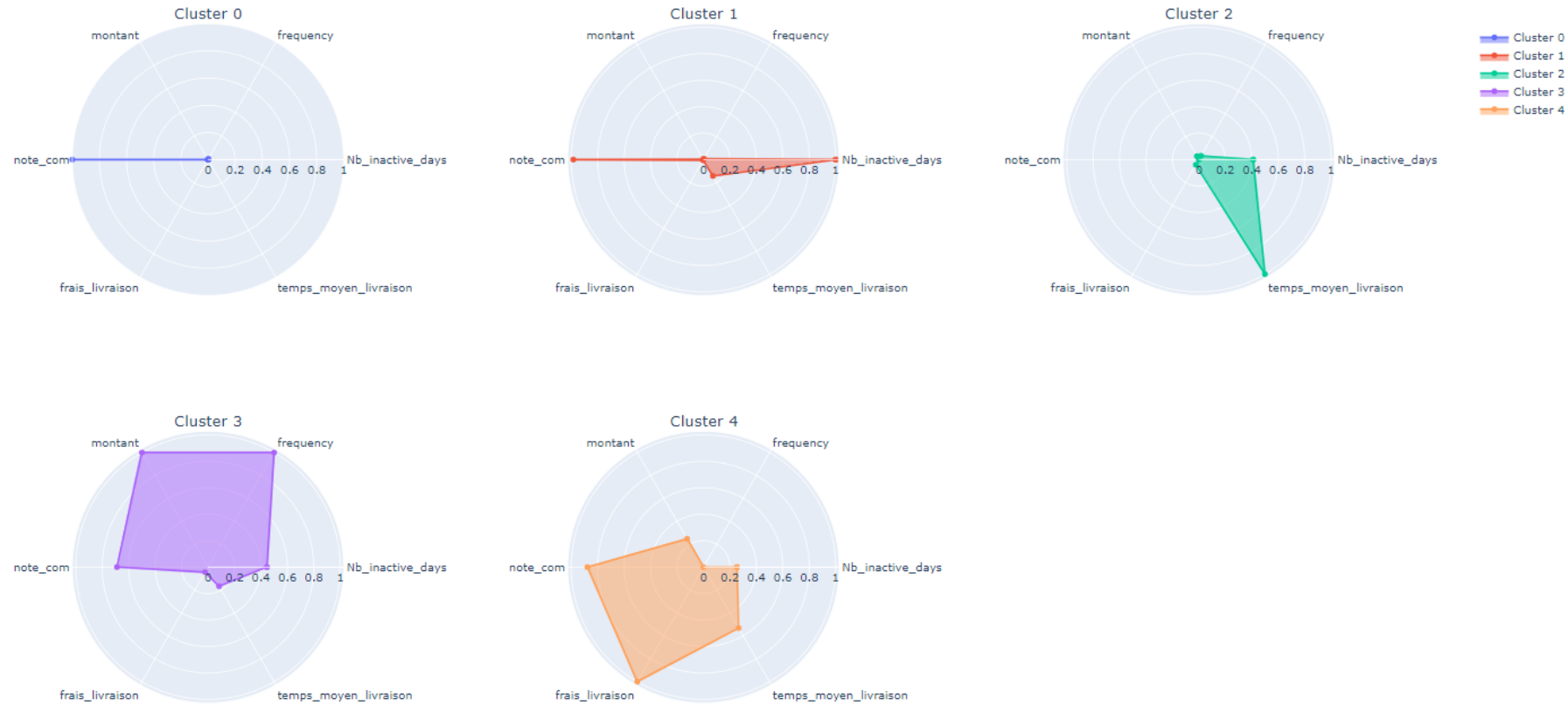
Distribution des clusters

	cluster	recency	frequency	montant	note_com	frais_livraison	temps_moyen_livraison	cluster count
0	0	122.984321	1.137898	151.859692	4.643550	17.172908	9.253361	41139
1	1	390.372743	1.141570	157.590317	4.596661	17.503821	11.053061	31963
2	2	232.949439	1.247269	201.144655	1.644665	19.818049	21.863596	13548
3	3	242.124673	5.227344	1790.877444	3.709360	19.716619	11.360398	1909
4	4	190.794103	1.111337	556.403592	4.275758	74.640649	15.960346	3934

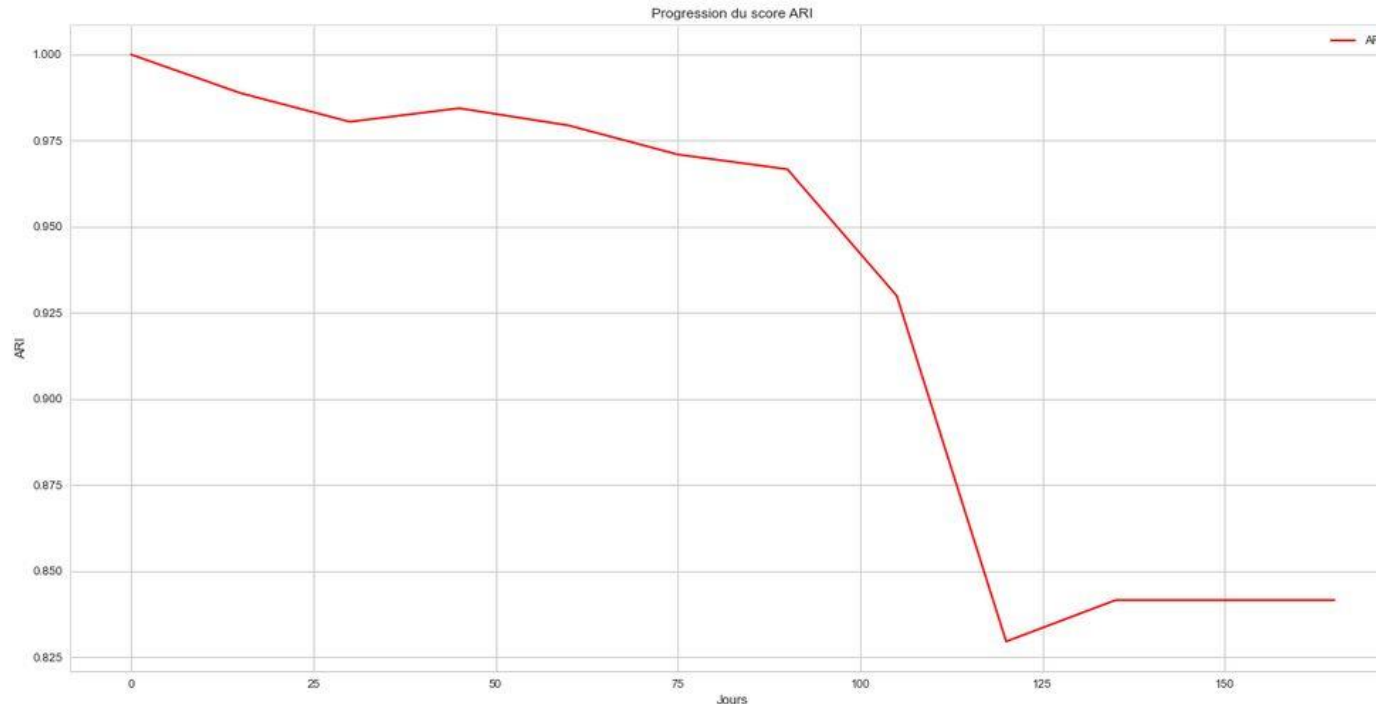
Cluster 0 :
Cluster 1 :
Cluster 2 :

Caractérisation des clusters

Radar plot des clusters



Maintenance



- ❖ Calcul du delta de temps
- ❖ Mis en place du score ARI étape par étape
- ❖ Interprétation du résultat



**Conclusion et quelques points
d'ouvertures**

Merci pour votre attention

Divine Tulomba
divinetlmb@gmail.com