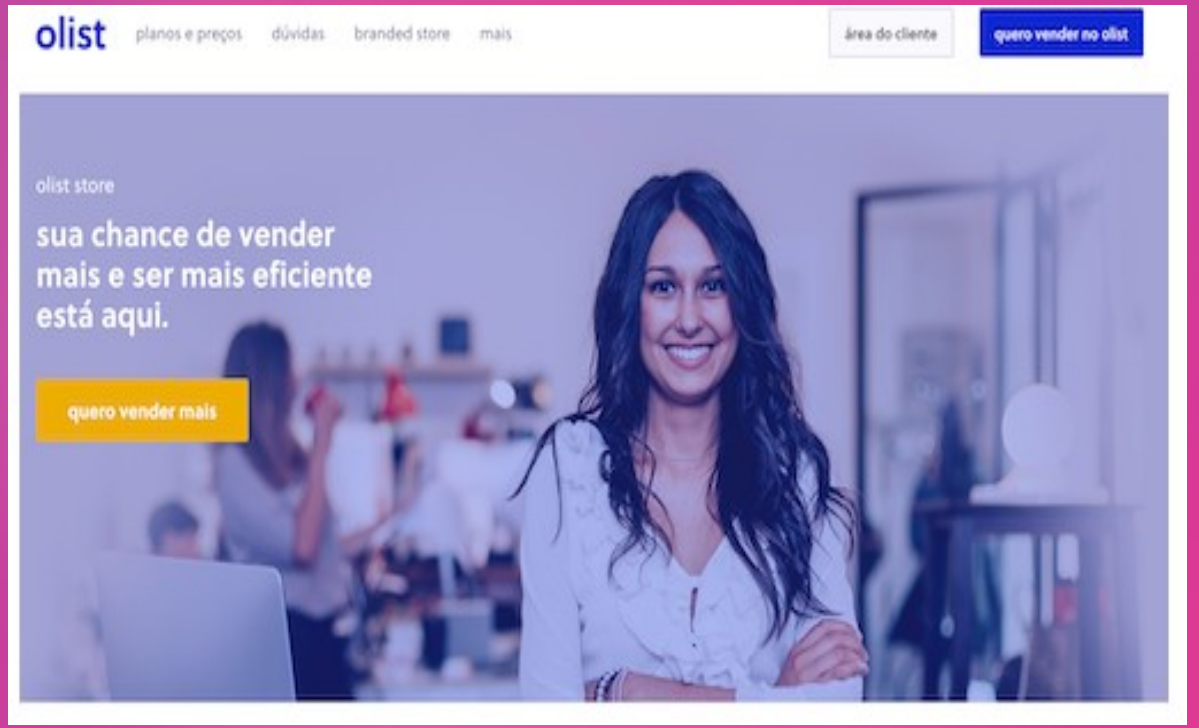
Projet 5 : Segmentez des clients d'un site e-commerce



Abdessalem BOUZAYANI

24/01/2023

Sommaire

Problématique

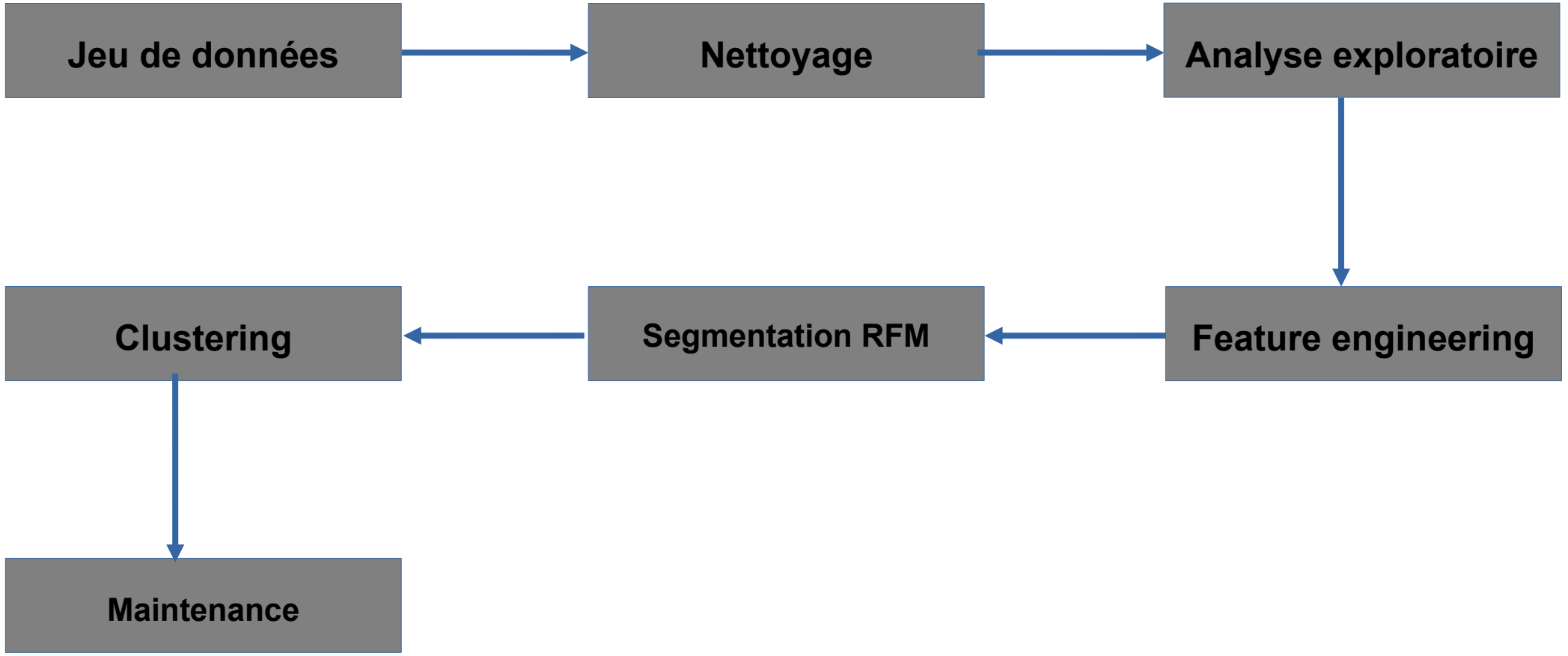
Mission :

- Fournir aux équipes Marketing de l'entreprise Olist(site de e-commerce) une segmentation des clients utilisables dans leurs campagnes de communication

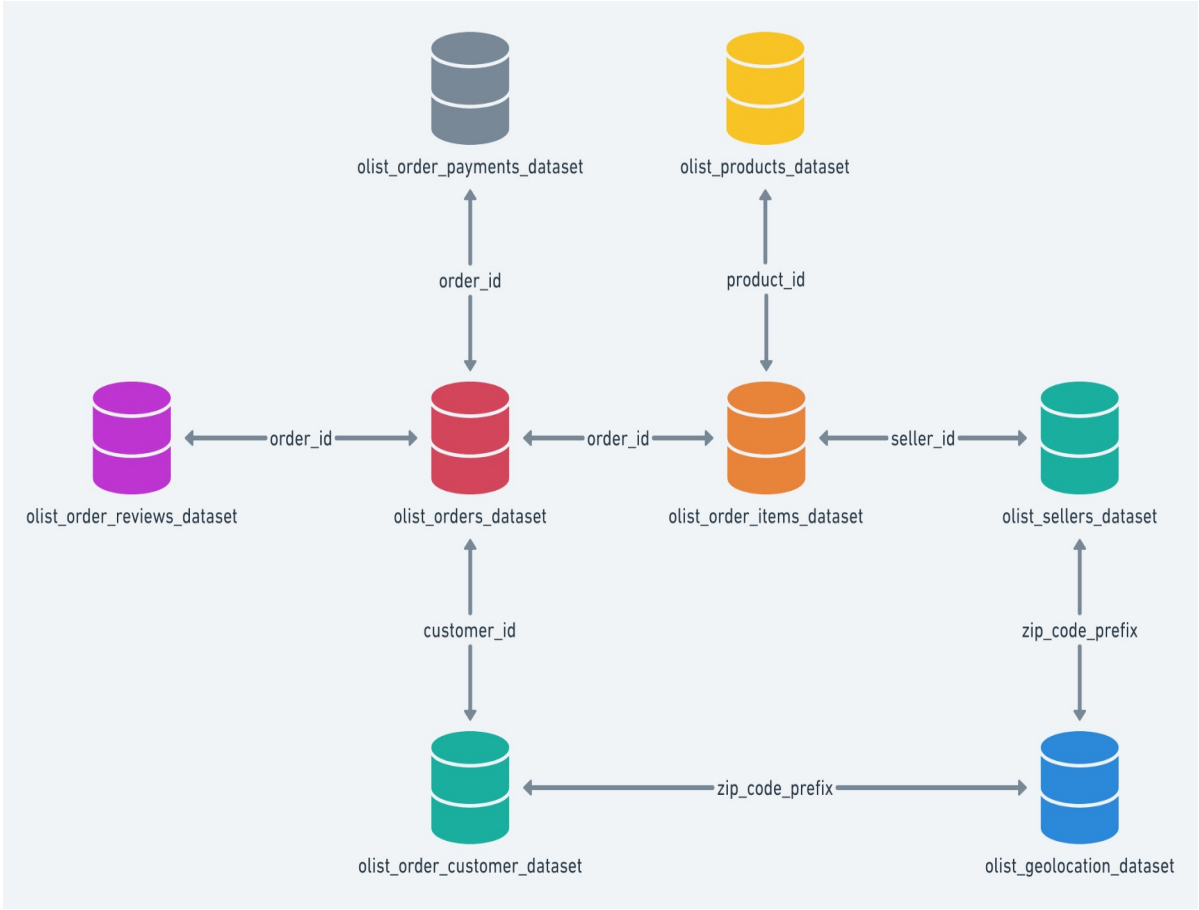
Objectifs :

- Comprendre les différents types d'utilisateurs (comportements, données personnelles)
- Fournir une description actionnable de la segmentation avec une logique sous-jacente pour une optimisation optimale.
- Proposer un contrat de maintenance basé sur une analyse de la stabilité des segments au cours du temps.

Feuille de route



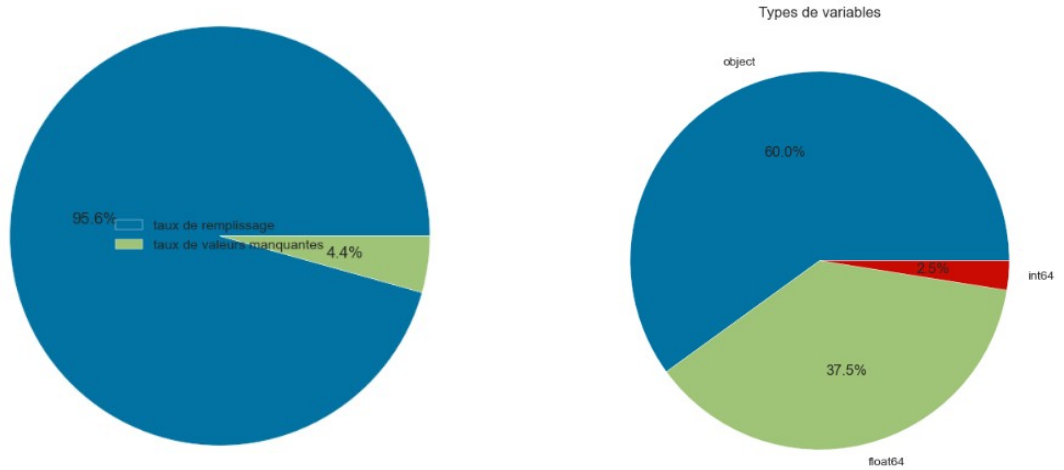
Jeu de données



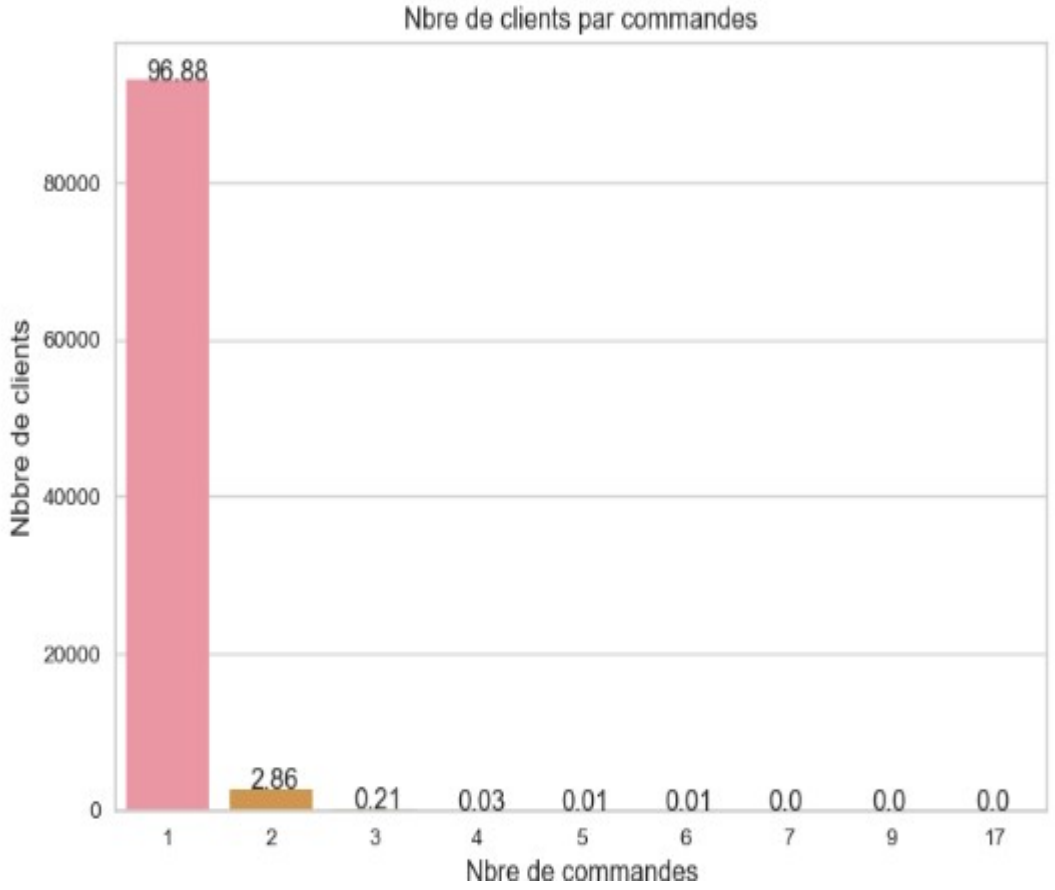
base de données anonymisée comportant des informations sur l'historique de commandes, les produits achetés, les commentaires de satisfaction et la localisation des clients

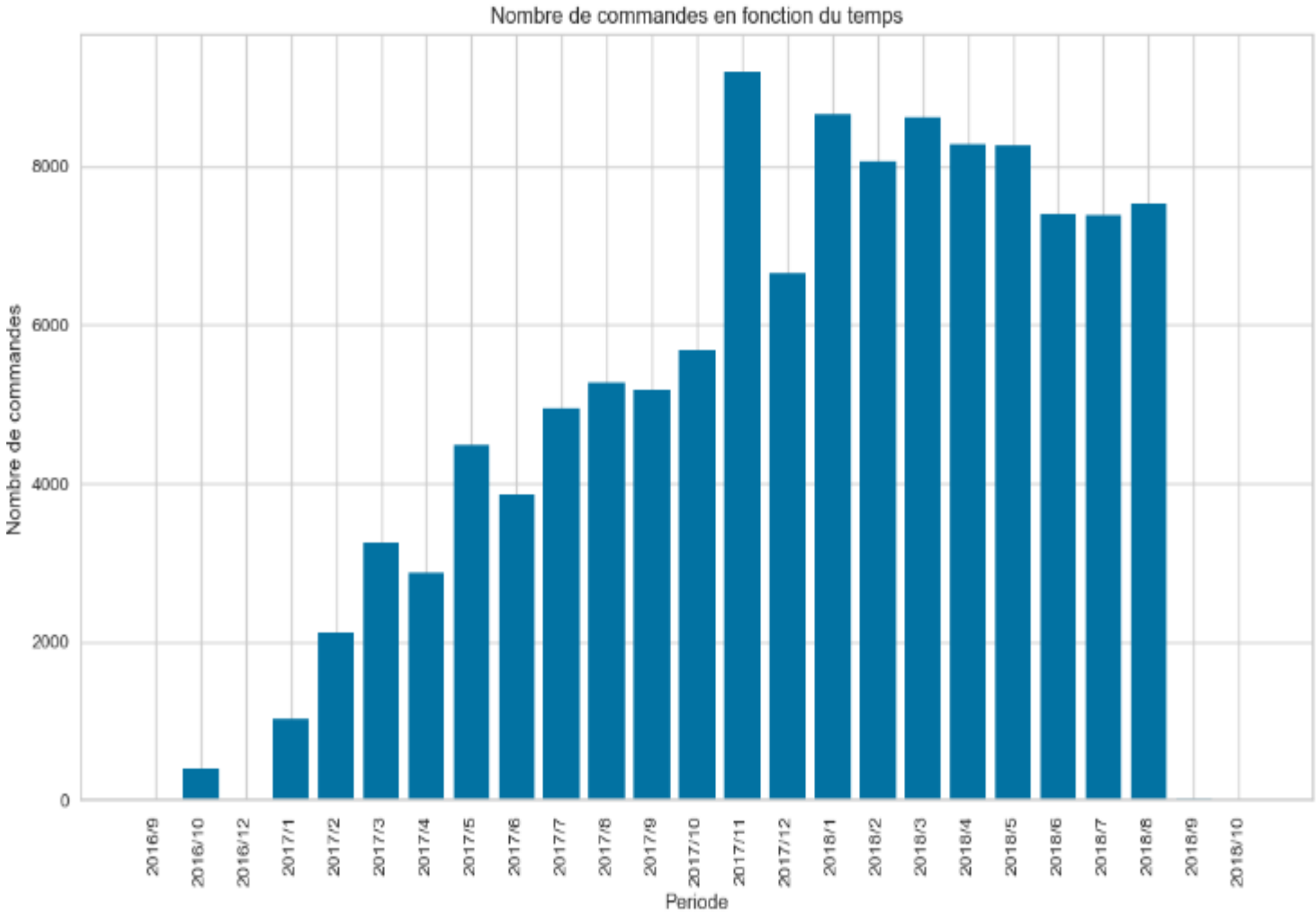
40 colonnes

119143 lignes

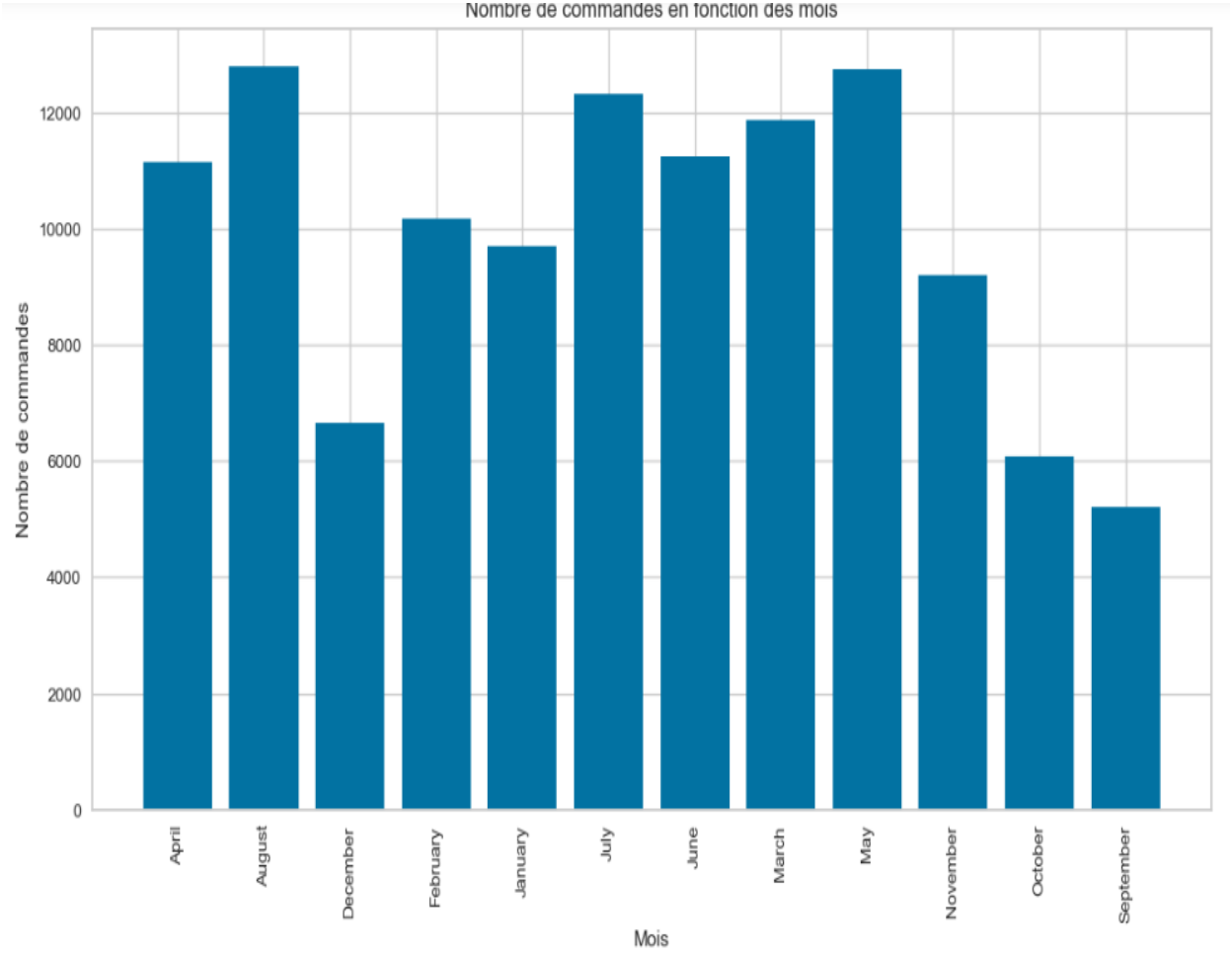


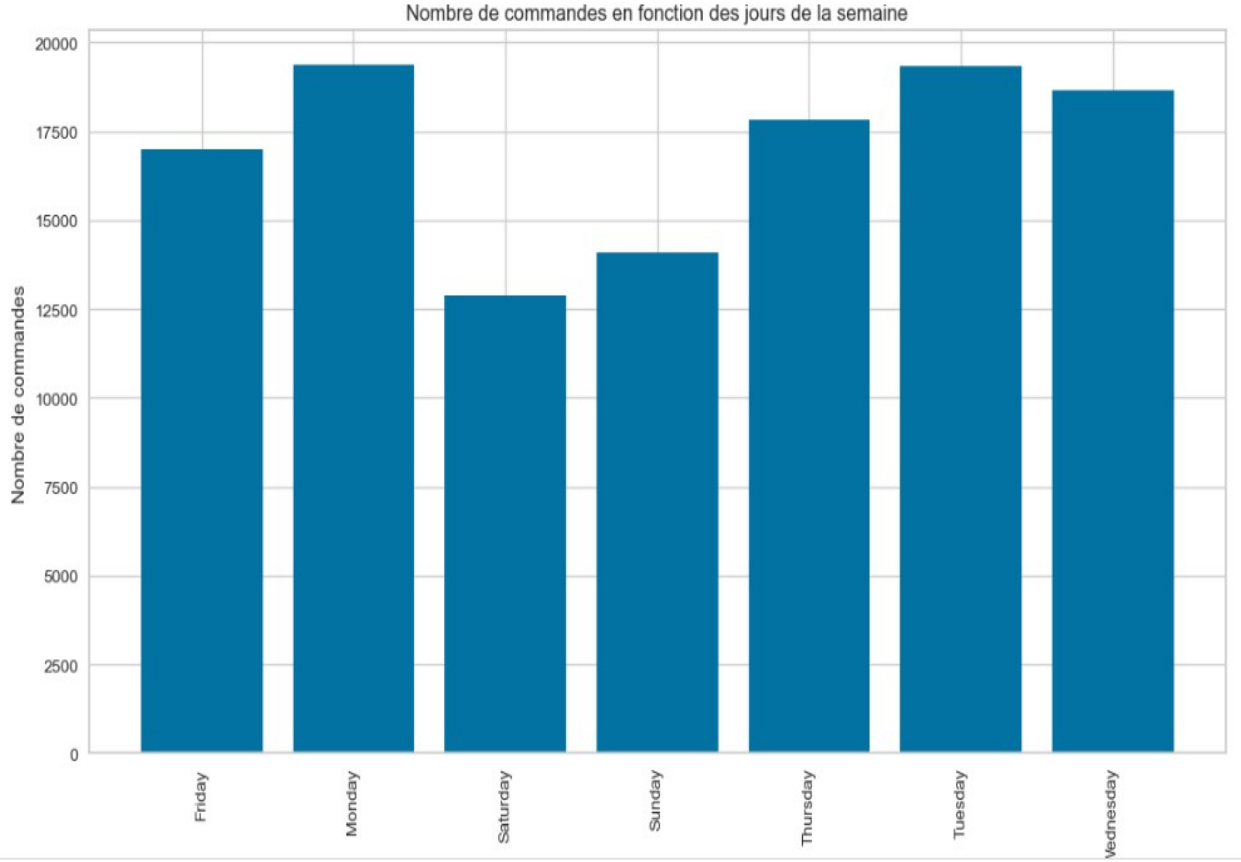
Analyse exploratoire



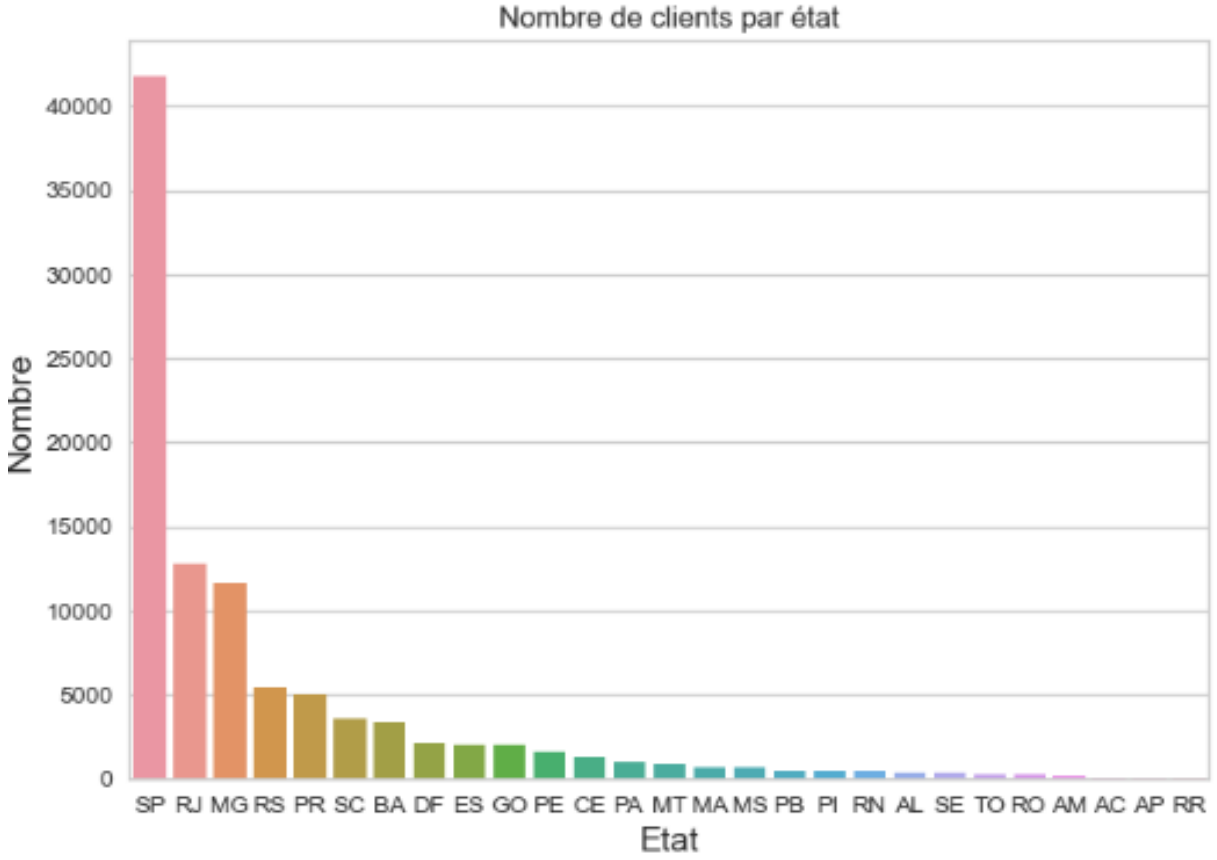


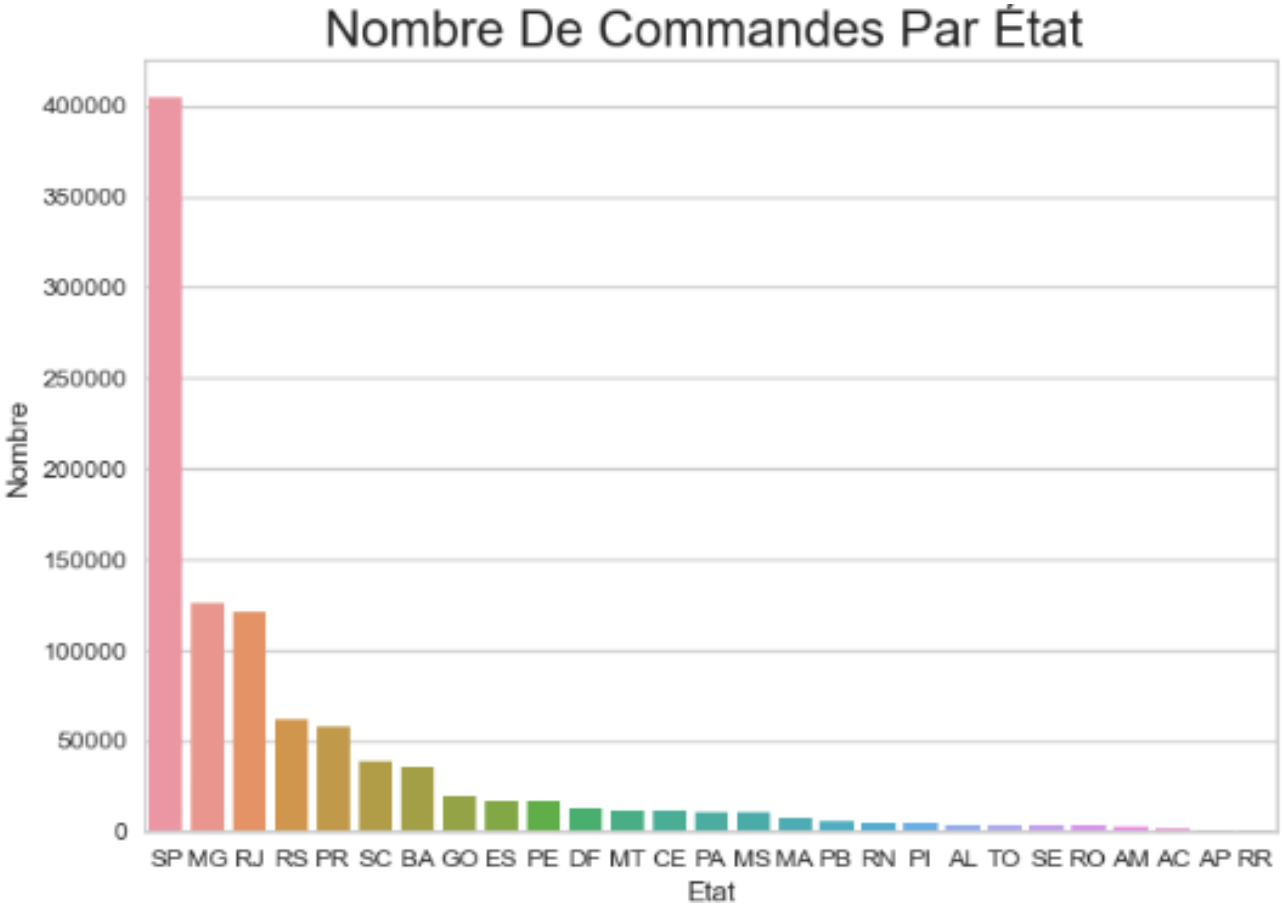
Analyse exploratoire



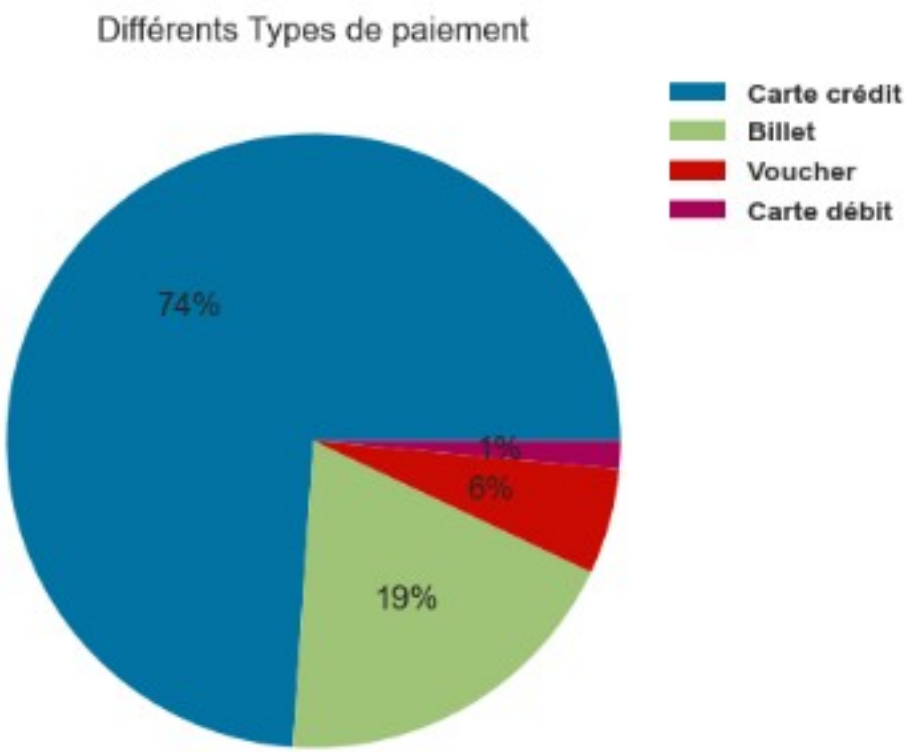


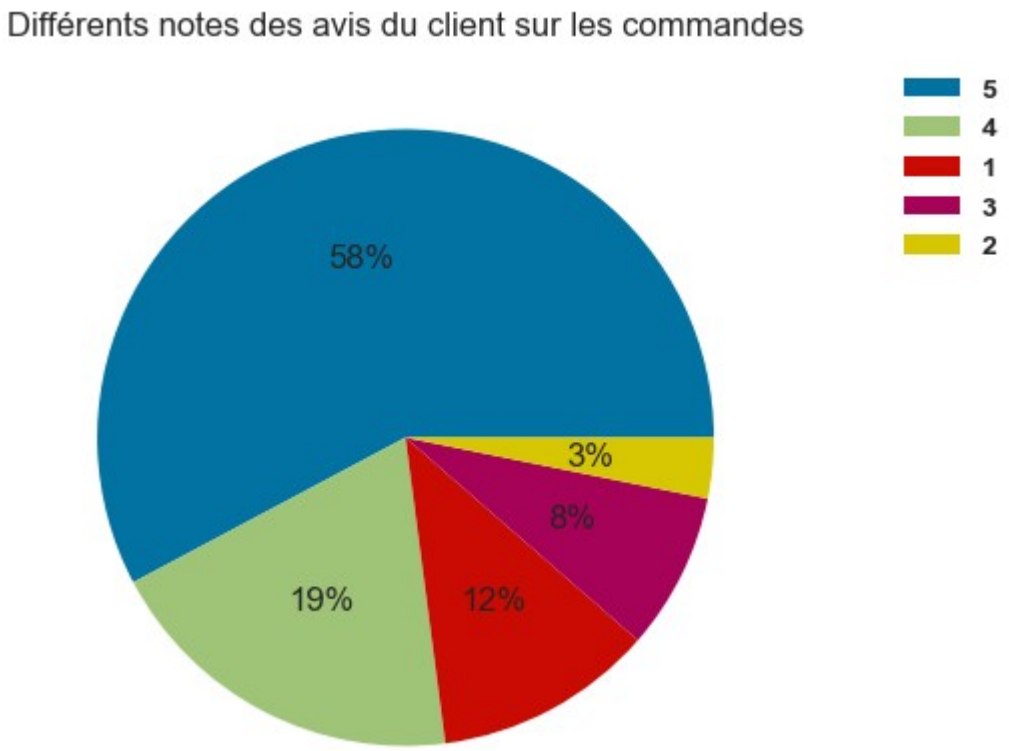
Analyse exploratoire





Analyse exploratoire



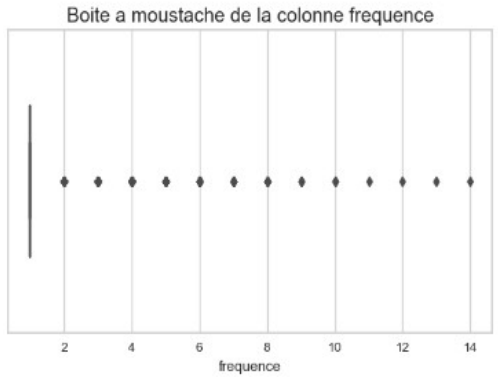


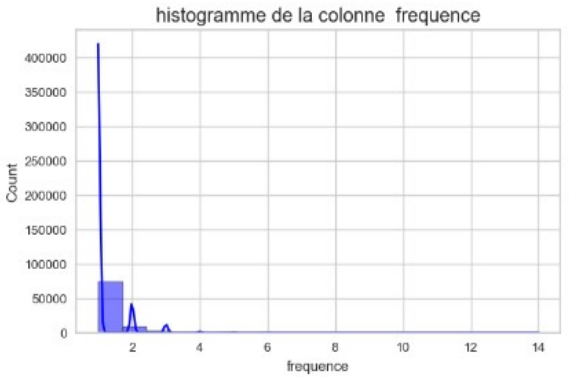
Feature engineering

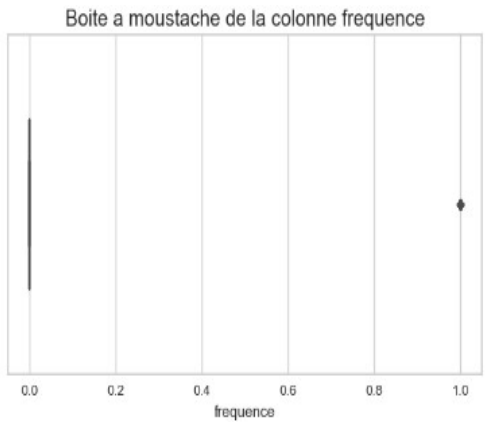
Variables	Transformations
product_category_name_english (71 valeurs)	product_category_name (5 valeurs)
customer_state (27 valeurs)	customer_state (5 valeurs)
customer_unique_id, order_id	nbre_commande
customer_unique_id, order_id	nbre_produit
customer_unique_id, review_score	score_moyen
order_purchase_timestamp, order_delivered_customer_date	delai_livraison
order_purchase_timestamp	recence
order_id	frequence
payment_value	montant
Variables catégorielles	One hot encoding
Variables continues	Normalisation, transformation log

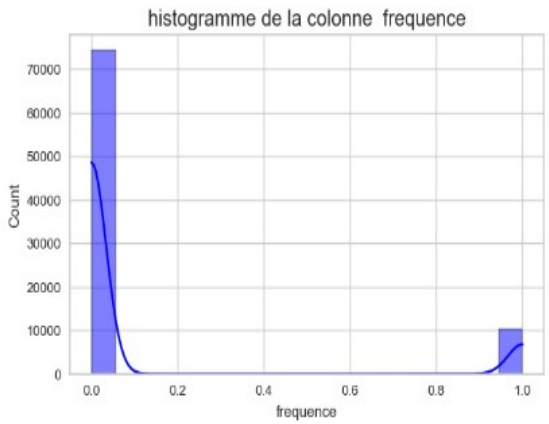
Feature engineering

Transformation dichotomique de la variable ‘frequence’

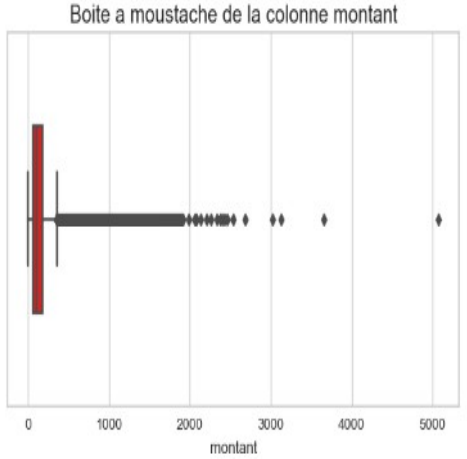


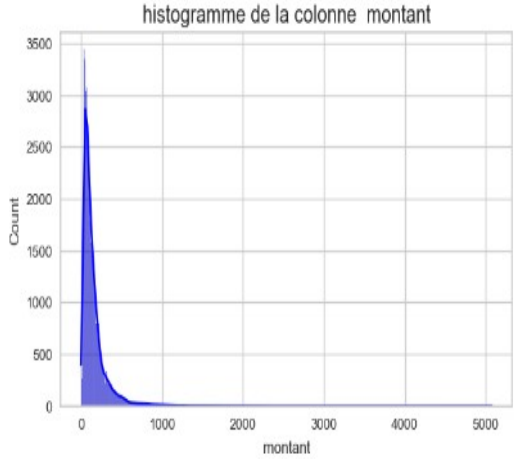


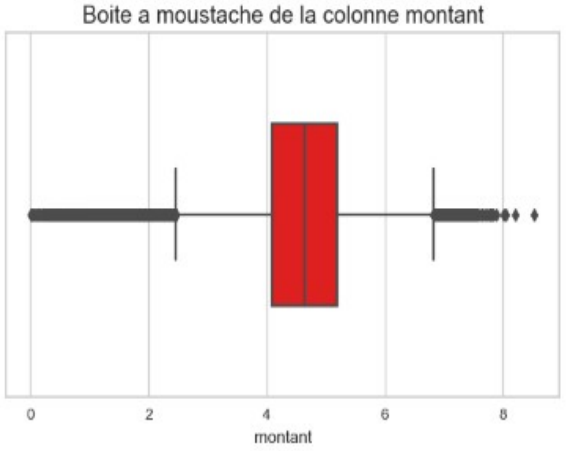


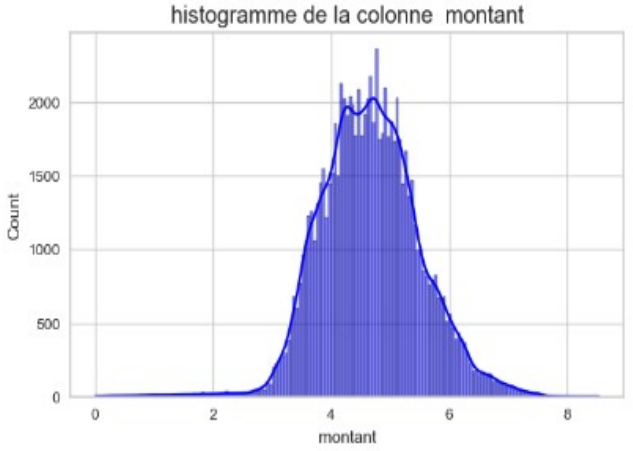


Transformation logarithmique de la variable ‘montant’



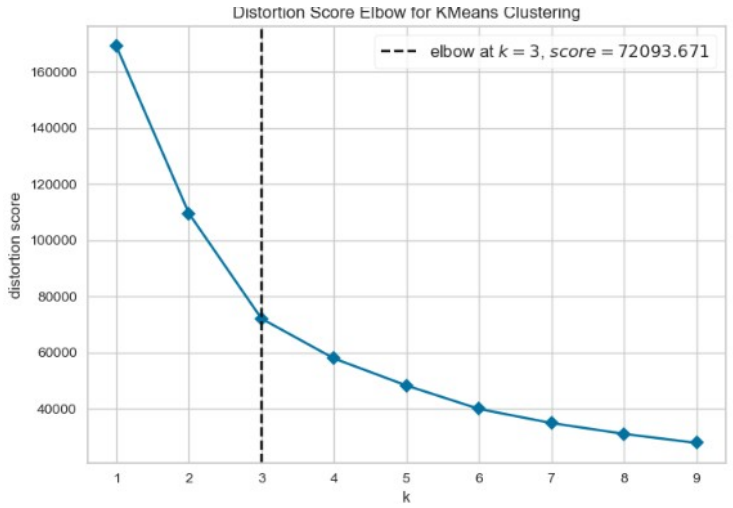


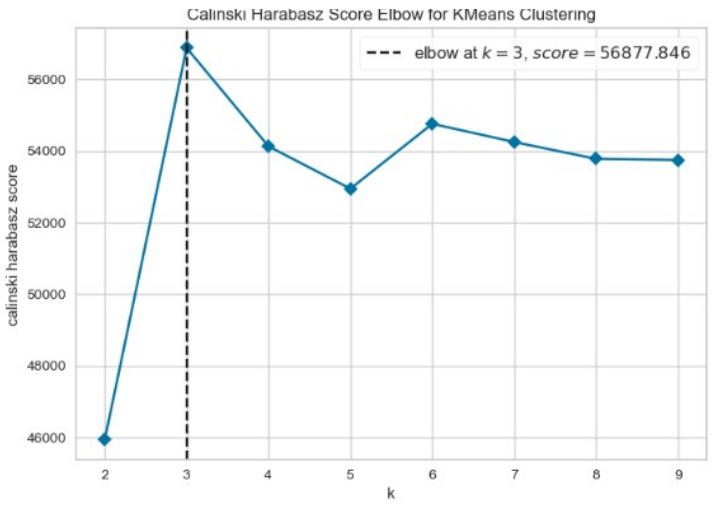


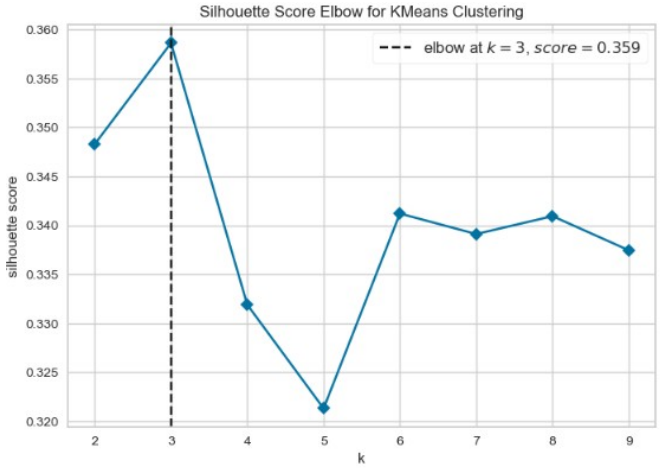


Segmentation RFM

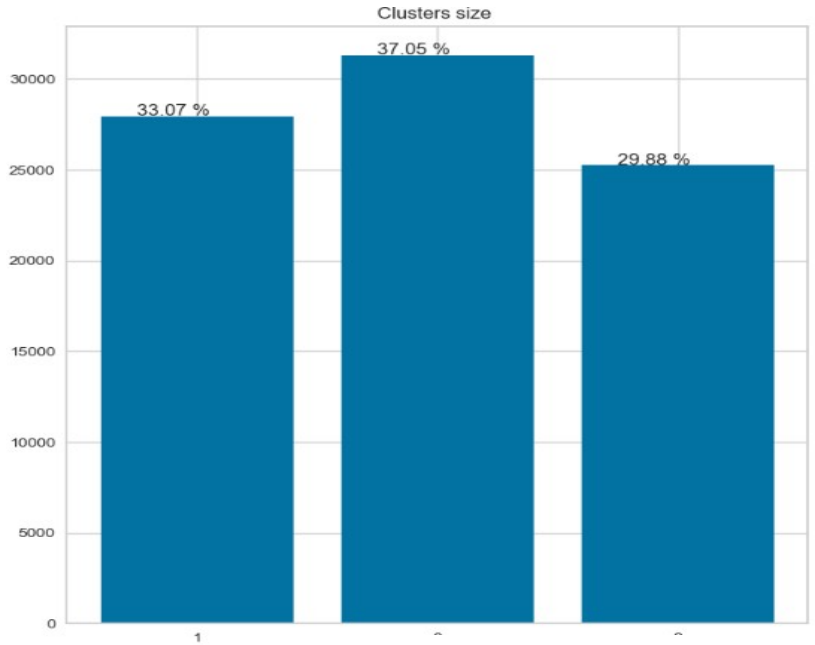
- **R** : nombre de jour écoulé depuis le dernier achat
- **F** : nombre d'achat effectué
- **M** : somme totale dépensée

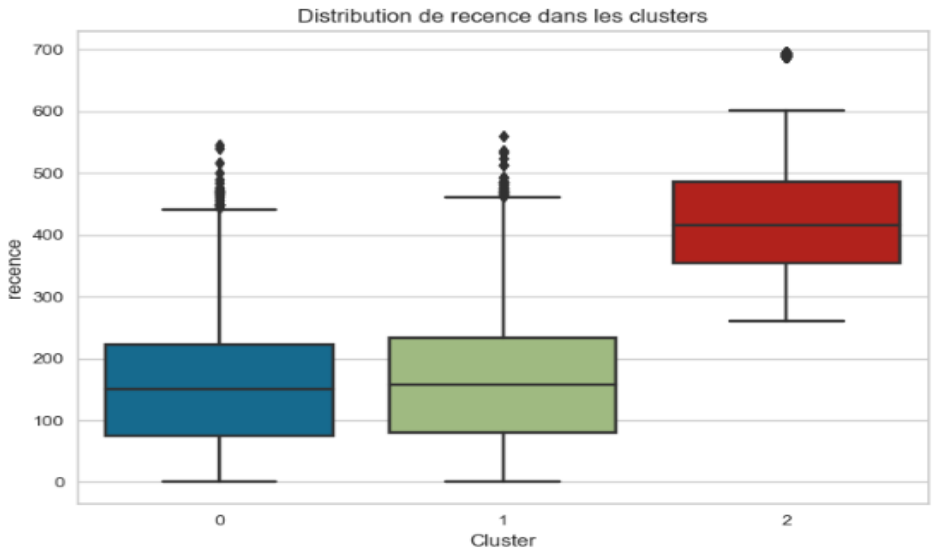


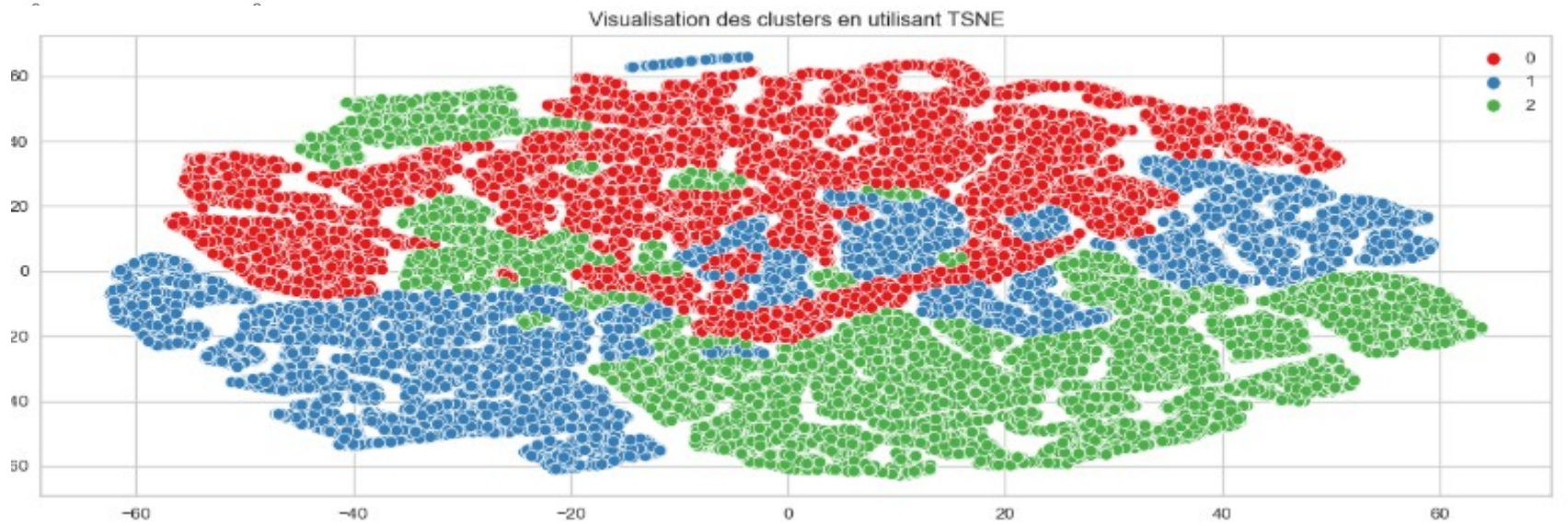




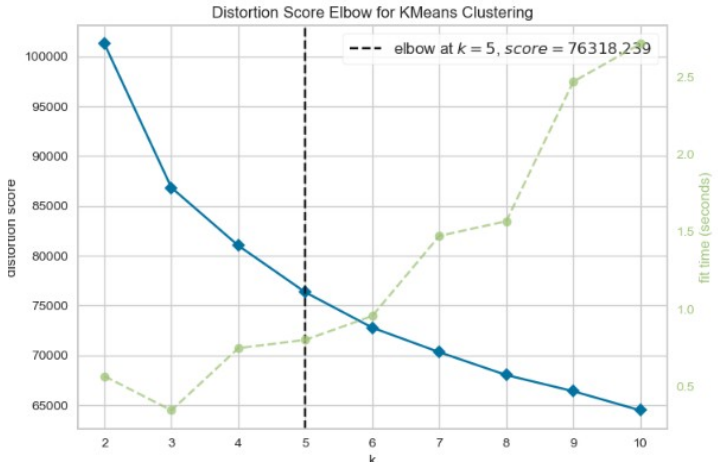
Segmentation RFM

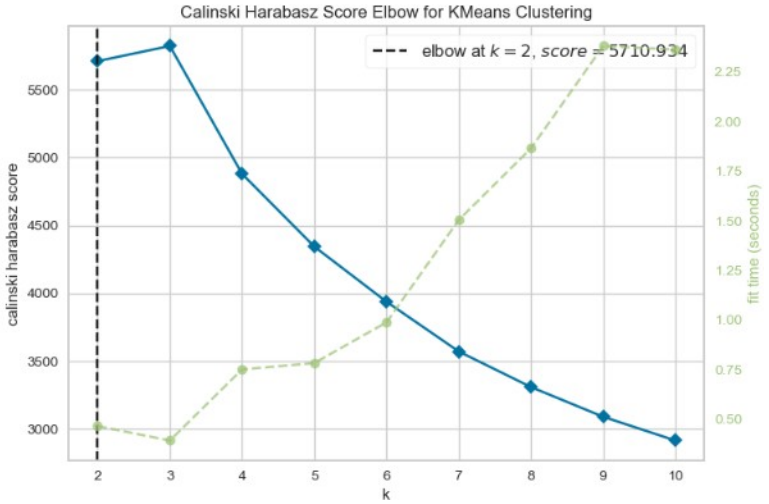


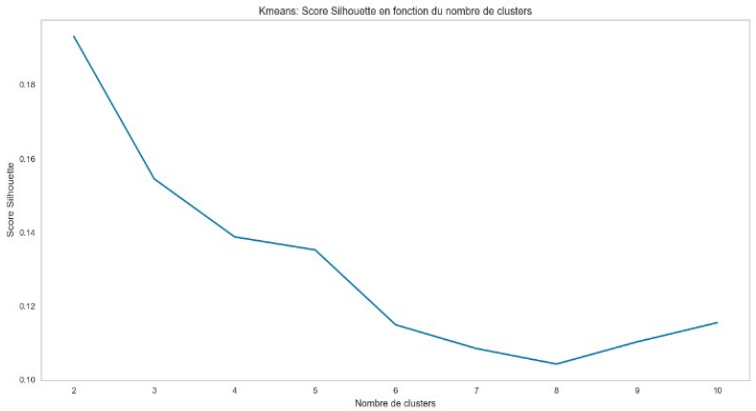


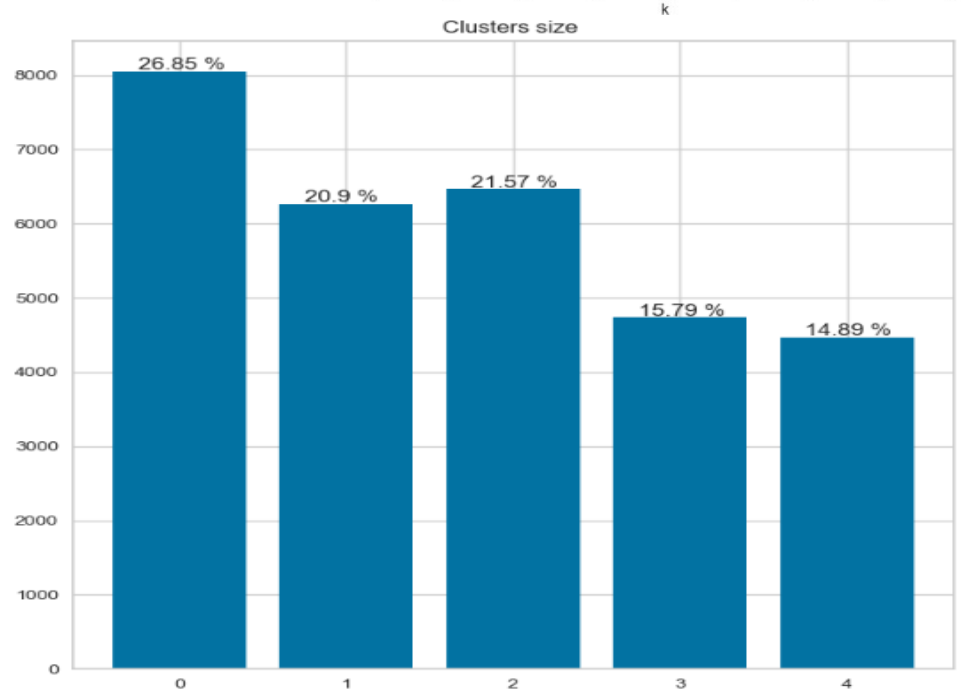


Kmeans


Distortion Score Elbow for KMeans Clustering
elbow at k = 5, score = 76318.239
distortion score
fit time (seconds)
k

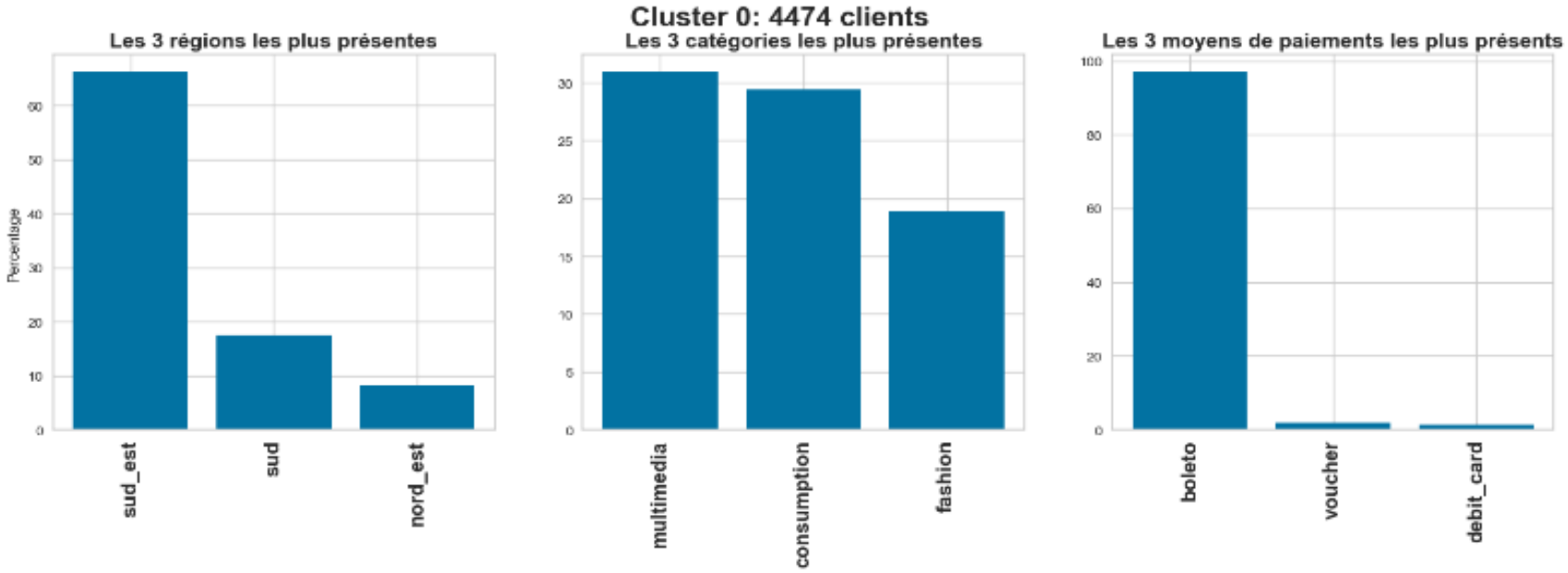

Calinski Harabasz Score Elbow for KMeans Clustering
elbow at k = 2, score = 5710.934
calinski harabasz score
fit time (seconds)
k


Kmeans: Score Silhouette en fonction du nombre de clusters
Score Silhouette
Nombre de clusters

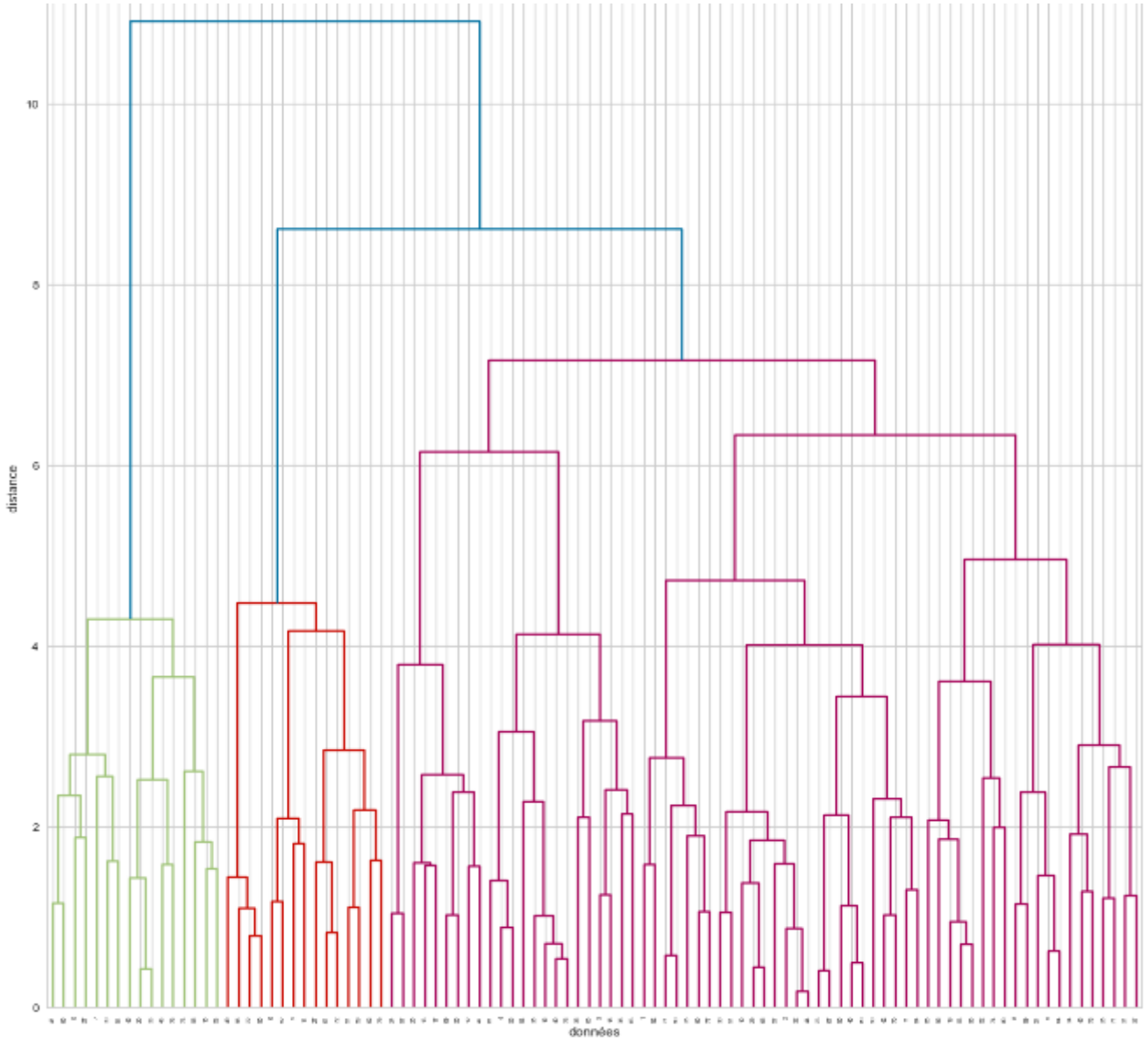

Clusters size
26.85 %
20.9 %
21.57 %
15.79 %
14.89 %

Kmeans

	recence				frequence				montant				score_moyen				delai_livraison			
	min	mean	median	max	min	mean	median	max	min	mean	median	max	min	mean	median	max	min	mean	median	max
Cluster																				
0	46	280.5	267.0	694	1	1.1	1.0	6	6.9	108.3	87.7	383.0	1.0	4.2	5.0	5.0	1	13.3	12.0	189
1	0	36.1	32.0	109	1	1.3	1.0	14	0.7	162.9	117.9	1917.6	1.0	4.3	5.0	5.0	0	7.3	6.0	88
2	39	270.1	248.0	694	1	1.1	1.0	4	4.5	129.1	118.6	771.6	1.0	3.8	4.0	5.0	3	19.7	17.0	195
3	30	279.1	266.0	694	1	1.1	1.0	5	0.0	76.6	67.8	215.0	1.0	4.4	5.0	5.0	0	7.7	7.0	50
4	20	256.2	231.0	694	1	2.2	2.0	13	150.1	508.5	393.1	5076.0	1.0	3.9	5.0	5.0	0	11.1	9.0	124



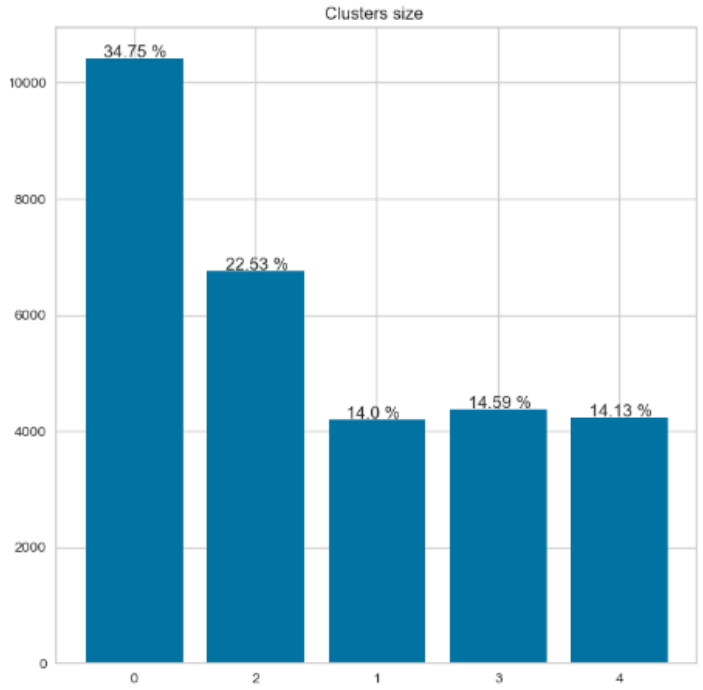
Clustering Hiérarchique

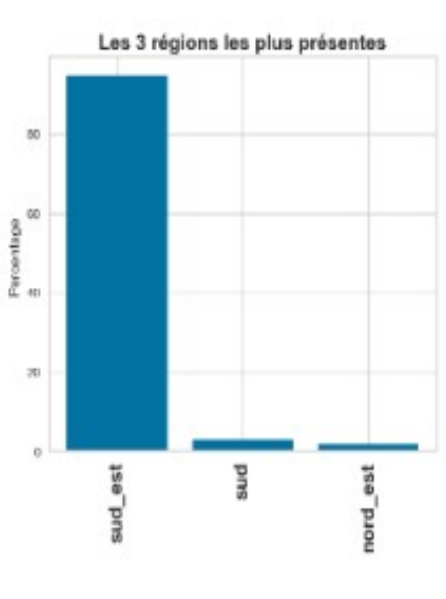


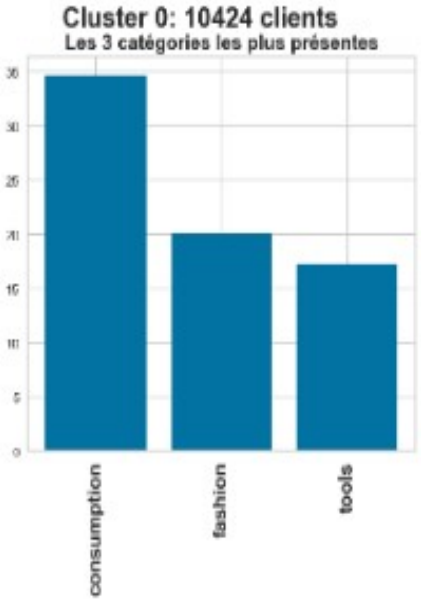
n_clusters	coef_silh	davies_bouldin	calinski_harabasz
2	0.208068	1.766434	4417.544227
3	0.113245	2.149770	4047.284454
4	0.102126	2.219944	3531.949834
5	0.100065	2.191752	3198.953492
6	0.078324	2.121315	2861.875329
7	0.078308	2.093554	2645.394516
8	0.074269	2.065837	2481.429038
9	0.081165	1.937038	2334.957952

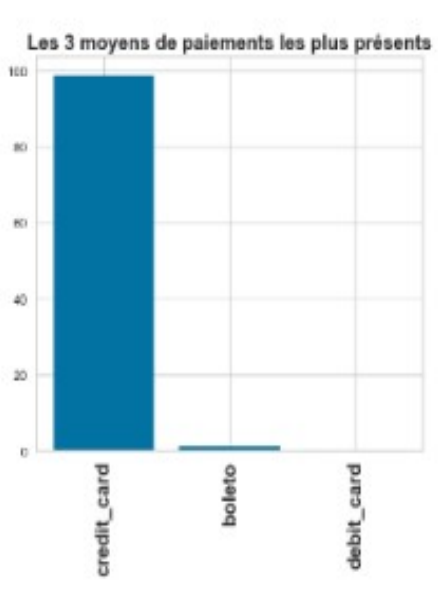
Clustering Hiérarchique

	recence				frequence				montant				score_moyen				delai_livraison			
	min	mean	median	max	min	mean	median	max	min	mean	median	max	min	mean	median	max	min	mean	median	max
Cluster																				
0	8	263.7	249.0	694	1	1.1	1.0	5	0.0	110.2	84.6	1562.7	1.0	4.2	5.0	5.0	1	11.0	9.0	136
1	11	281.7	268.0	694	1	1.1	1.0	9	5.4	129.5	104.1	1483.1	1.0	4.1	5.0	5.0	1	17.3	15.0	195
2	13	251.8	225.0	694	1	2.0	2.0	13	29.6	443.2	336.3	5076.0	1.0	3.9	5.0	5.0	0	11.1	9.0	124
3	21	279.8	266.0	694	1	1.2	1.0	6	3.8	111.0	80.8	1070.3	1.0	4.2	5.0	5.0	1	12.7	11.0	108
4	0	39.3	29.0	275	1	1.3	1.0	14	6.6	185.7	117.8	2459.5	1.0	4.3	5.0	5.0	1	8.1	7.0	88



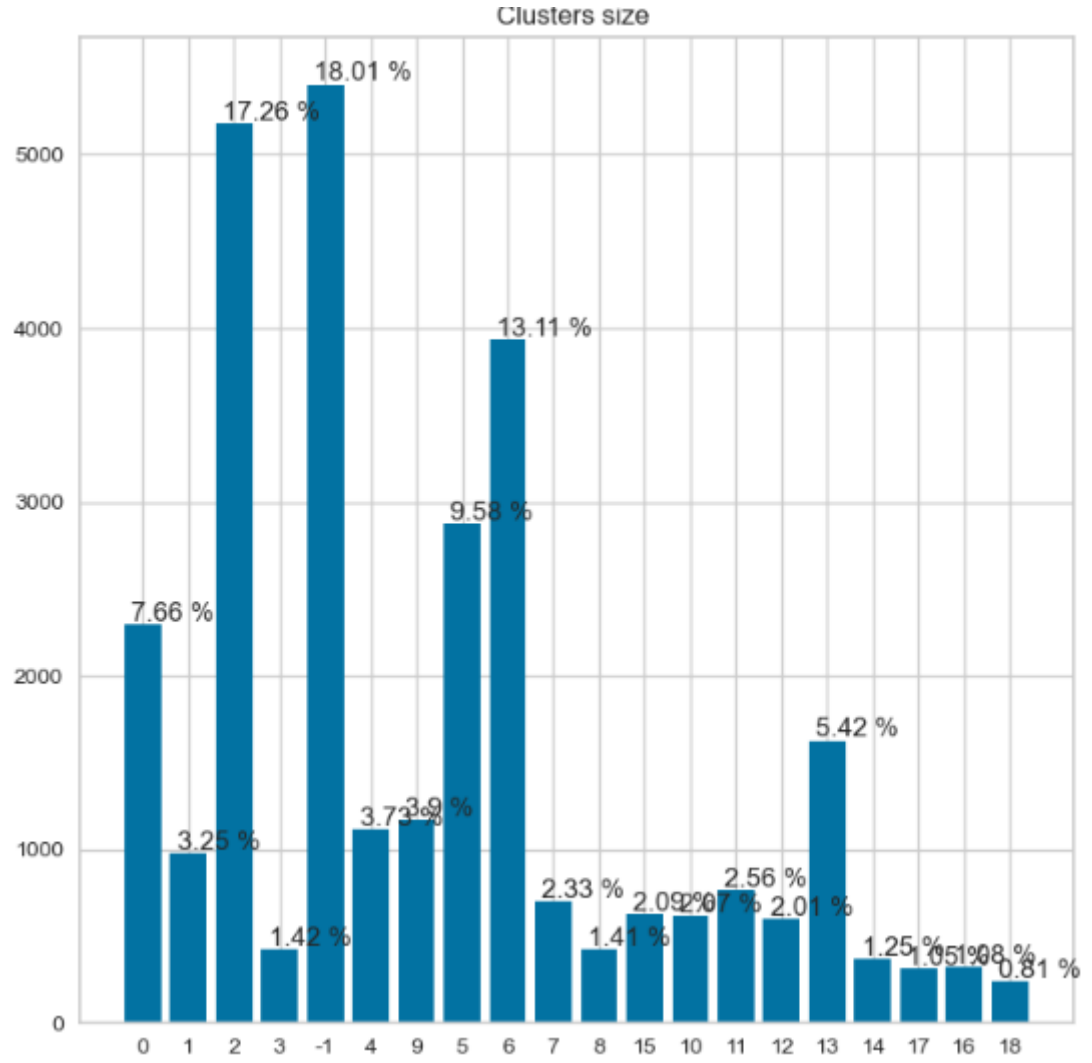






DBSCAN

ebsilon	coef_silh	davies_bouldin	calinski_harabasz
0.5	-0.013076	2.167868	924.299413
1.0	0.193035	2.228448	1712.878807
1.5	0.683651	1.690495	1074.796026
2.0	0.792322	1.088870	1597.360114
2.5	0.820351	1.135780	1200.796307
3.0	0.839821	1.172890	913.908346
4.0	0.860818	0.791321	522.340016



DBSCAN

	recence				frequence				montant				score_moyen				delai_livraison			
	min	mean	median	max	min	mean	median	max	min	mean	median	max	min	mean	median	max	min	mean	median	max
Cluster																				
-1	0	218.9	200.0	694	1	1.5	1.0	14	0.0	257.8	129.0	5076.0	1.0	4.0	5.0	5.0	1	14.8	13.0	195
0	4	235.2	218.0	693	1	1.5	1.0	9	4.8	225.1	129.6	2130.0	1.0	4.0	5.0	5.0	1	10.7	9.0	136
1	10	250.0	226.0	693	1	1.3	1.0	5	11.2	197.4	130.2	1206.5	1.0	4.2	5.0	5.0	1	12.8	11.0	74
2	1	228.3	210.0	693	1	1.3	1.0	8	2.8	174.8	115.4	1917.0	1.0	4.2	5.0	5.0	0	9.1	7.0	145
3	21	251.4	253.0	694	1	1.2	1.0	4	11.8	165.6	116.9	793.8	1.0	4.2	5.0	5.0	1	13.6	12.0	55
4	11	242.0	223.0	693	1	1.3	1.0	5	14.4	175.6	99.9	1566.6	1.0	4.0	5.0	5.0	1	11.4	10.0	68
5	5	254.8	234.0	694	1	1.5	1.0	9	3.4	207.5	120.4	2409.3	1.0	4.2	5.0	5.0	1	9.8	8.0	69
6	2	219.5	199.0	693	1	1.3	1.0	9	1.8	176.2	110.4	2372.8	1.0	4.1	5.0	5.0	0	10.3	8.0	87
7	21	266.0	252.0	693	1	1.4	1.0	6	13.4	163.8	98.7	1174.4	1.0	4.2	5.0	5.0	1	10.9	9.0	75
8	29	266.8	259.0	574	1	1.4	1.0	6	22.7	163.7	111.0	724.7	1.0	4.0	5.0	5.0	1	12.2	10.0	66
9	8	237.5	223.0	694	1	1.3	1.0	5	13.4	162.1	98.2	1336.6	1.0	4.2	5.0	5.0	1	10.1	8.0	73
10	19	265.7	258.5	691	1	1.4	1.0	5	13.1	213.5	134.4	1306.4	1.0	4.2	5.0	5.0	1	12.9	11.0	50
11	15	231.6	207.0	589	1	1.2	1.0	4	5.9	163.7	104.7	1473.1	1.0	4.1	5.0	5.0	1	13.5	11.0	77
12	19	241.7	222.0	691	1	1.2	1.0	4	8.8	161.7	120.1	1083.9	1.0	3.9	5.0	5.0	3	19.0	16.0	86
13	6	215.2	186.0	694	1	1.4	1.0	7	5.2	219.2	137.2	1905.6	1.0	4.1	5.0	5.0	1	9.7	8.0	61
14	30	257.5	236.5	575	1	1.3	1.0	5	15.6	159.0	123.2	734.0	1.0	4.0	5.0	5.0	2	14.3	12.0	54
15	19	243.9	225.5	691	1	1.2	1.0	4	14.8	185.8	140.2	1368.4	1.0	4.1	5.0	5.0	2	18.1	16.0	76
16	35	257.2	226.0	585	1	1.3	1.0	3	17.0	144.4	111.4	665.2	1.0	4.3	5.0	5.0	2	11.4	10.0	37
17	40	302.4	295.0	694	1	1.3	1.0	6	18.6	189.0	136.4	936.7	1.0	3.9	5.0	5.0	6	20.0	17.0	65
18	46	243.2	229.5	571	1	1.1	1.0	3	23.2	119.8	99.9	420.8	1.0	4.2	5.0	5.0	3	14.8	14.0	47

Choix du meilleur algorithme

Algorithme	Nbre clusters	Silouette	Davies bouldin	Calinski harabasz	Complexité	Interprétation des clusters
KMEANS	5	0,13	1,98	1446	faible	facile
CAH	5	0,10	2,19	1067	forte	facile
DBSCAN	19	0,08	2,07	146	forte	difficile

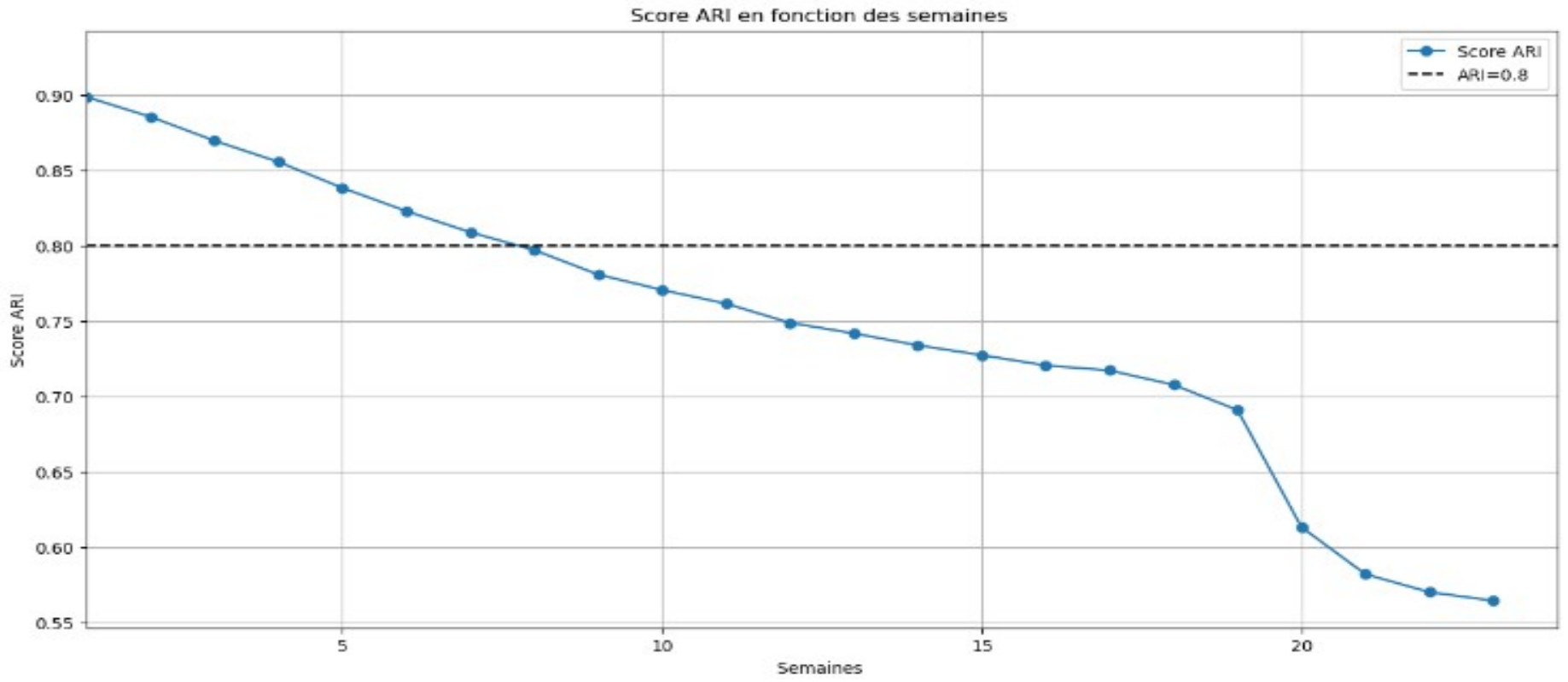
Chaque métrique est une moyenne de 10 répétitions
avec un échantillon de 10000 observations

Maintenance

- Choisir une période d'étude [T0 TN]
- CO=M0.fit(F0) (T0)
- T1 = T0+n jours
- C1_fit=M1.fit(F1) (T1)
- C1_predict=M0.predict(F1) (T1)
- Calculer ARI(C1_fit, C1_predict)
- Répéter jusqu'à TN
- Choisir T_maintenance (ARI<seuil)

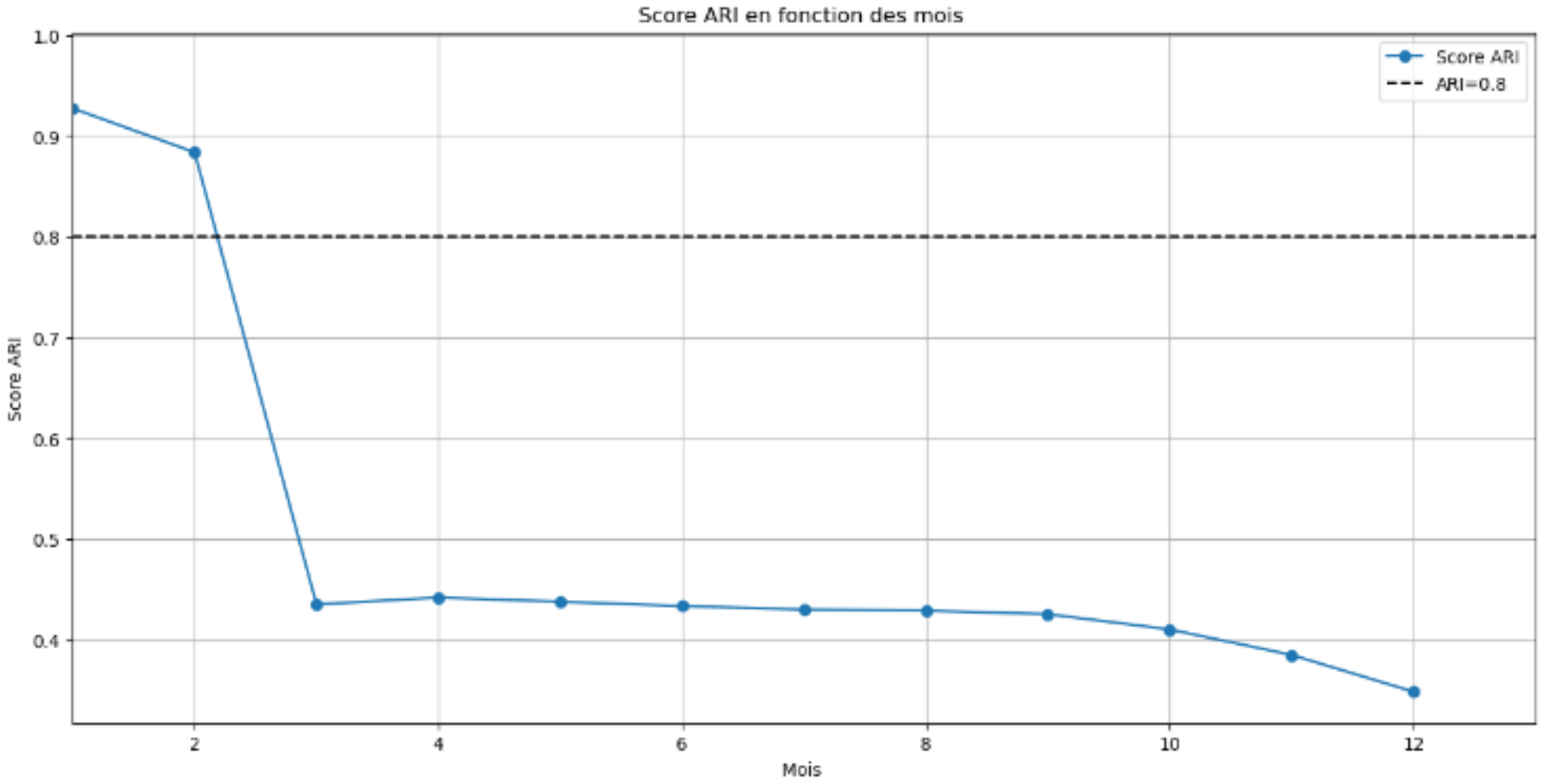
Maintenance

- **Période : 2018-03-18……..2018-08-29**
- **Fréquence : chaque semaine**



Maintenance

- **Période : 2017-06-30……..2018-08-29**
- **Fréquence : chaque mois**



Conclusion

- **Segmentation RFM**
  - **Simple à mettre en œuvre, rapide, marketing traditionnel.**
  - **3 variables seulement prise en compte, tout refaire pour l'ajout de nouveaux clients.**
- **Algorithmes de clustering : Kmeans, Agglomérative, Dbscan**
  - **Différentes métriques pour étudier la qualité du clustering**
  - **Modèle retenu : Kmeans avec 5 clusters interprétables**

Conclusion

Type	%	Caractéristiques	Stratégie
Perdus	15	>>R, <<F, <>M, >N, >L	Faire une étude Offrez des remises Recommandation de produits populaires
Nouveaux	16	<<R, <>F, <>M, >N, <<L	Envoi d'offres Proposez des produits pertinents et de bonnes affaires
Infidèles	21	<>R, <<F, <>M, <<N, >>L	Recommandations de produits en fonction de leur comportement Montrez l'importance d'acheter avec l'entreprise
A risque	27	>>R, <<F, <<M, >>N, <L	Proposer des produits susceptibles de les intéresser Proposer des réductions pour ces clients afin qu'ils se sentent valorisés
Fidèles	21	>R, >>F, >>M, <>N, <>L	Offrez des récompenses Offrez des remises

<< : plus petit
>> : plus grand
<> : intermédiaire
< : faible
> : grande

R : récence
F : fréquence
M : montant
N : note
L : délai de livraison

Perspectives

- **Jeu de données**

- **Nécessite plus de données :**
  - **démographiques (âge, profession, sexe, nombre d'enfants..)**
  - **psychographiques (avis sur le produit, centre d'intérêt…)**
- **Biaisés :**
  - **96% des clients avec une seule commande**
  - **Notes toutes très positives.**

- **Appliquer la segmentation des quantiles utilisée dans l'outil d'analyse RFM de PUTLER**

Merci de votre attention

Contact : bouzaieni@gmail.com