



Olist

Segmentation des clients d'un site de e-commerce

Présenté par Thiery KAPPE.



Le contexte

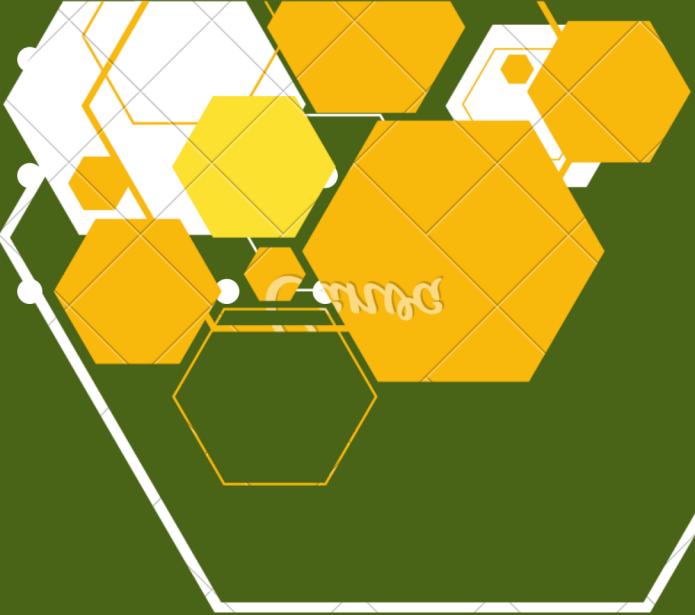
Olist, plate-forme de e-commerce souhaite donner à son équipe Customer Experience un Dashboard contenant les KPIs essentiels pour la compréhension, la visibilité et le pilotage du service client.

Olist souhaite en outre réaliser une segmentation fine de l'ensemble de ses clients afin de l'utiliser dans des campagnes de communication ciblées.

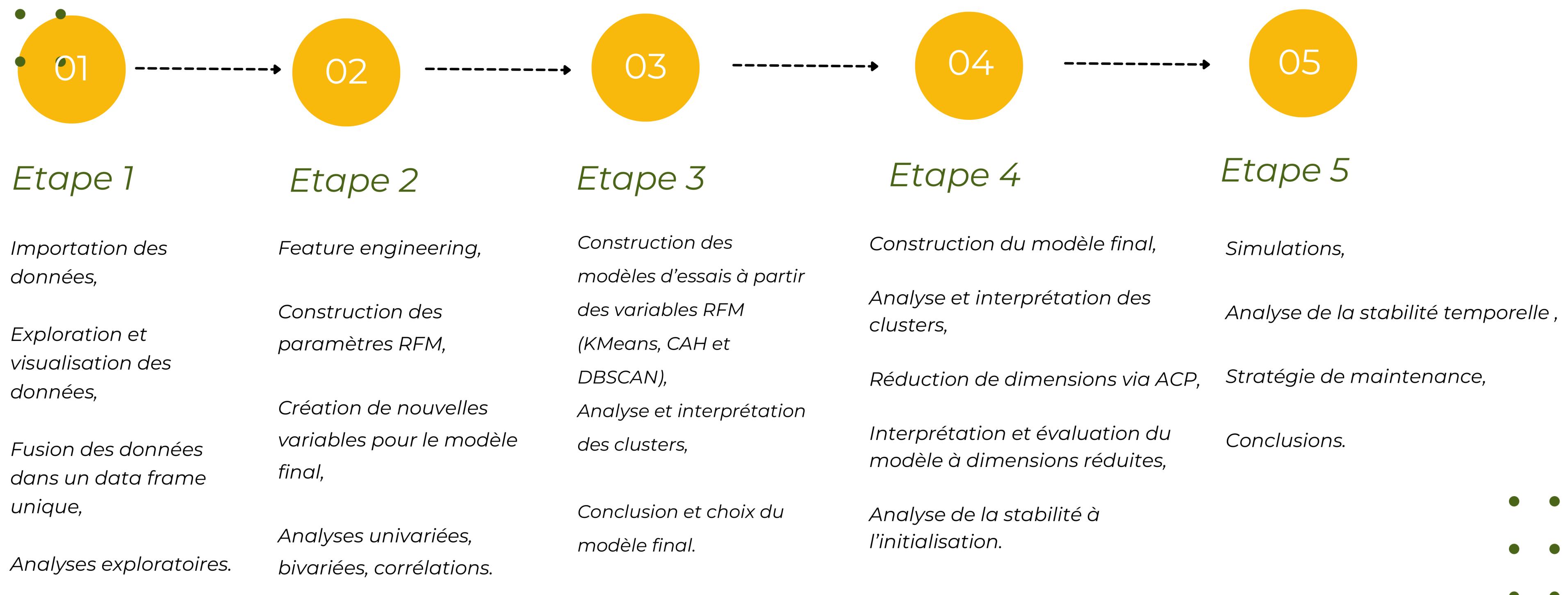
NB : Les travaux tels que présentés ici sont plutôt linéaires afin de faciliter la présentation. Dans la pratique, les étapes étaient plutôt itératives avec des essais et des allers- retours.

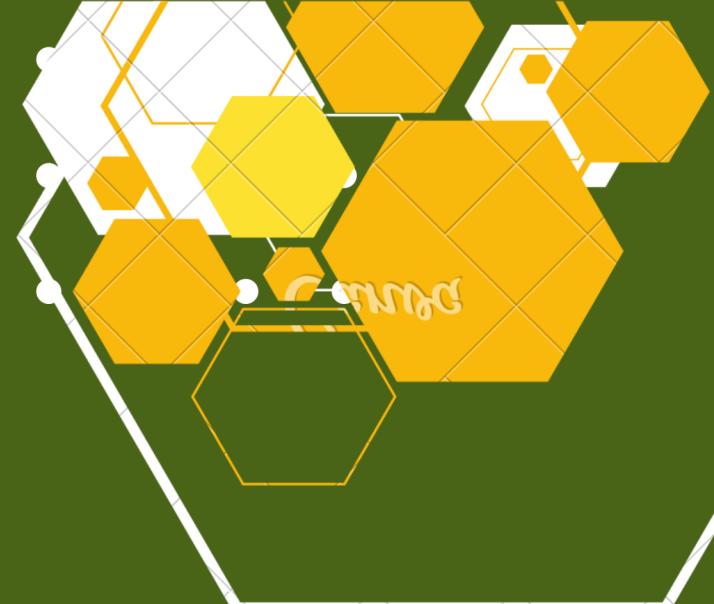
Cette présentation est soutenue par : un script des requêtes SQL, un notebook d'exploration, un notebook de modélisation et un notebook de simulation.





Vue globale des étapes clés

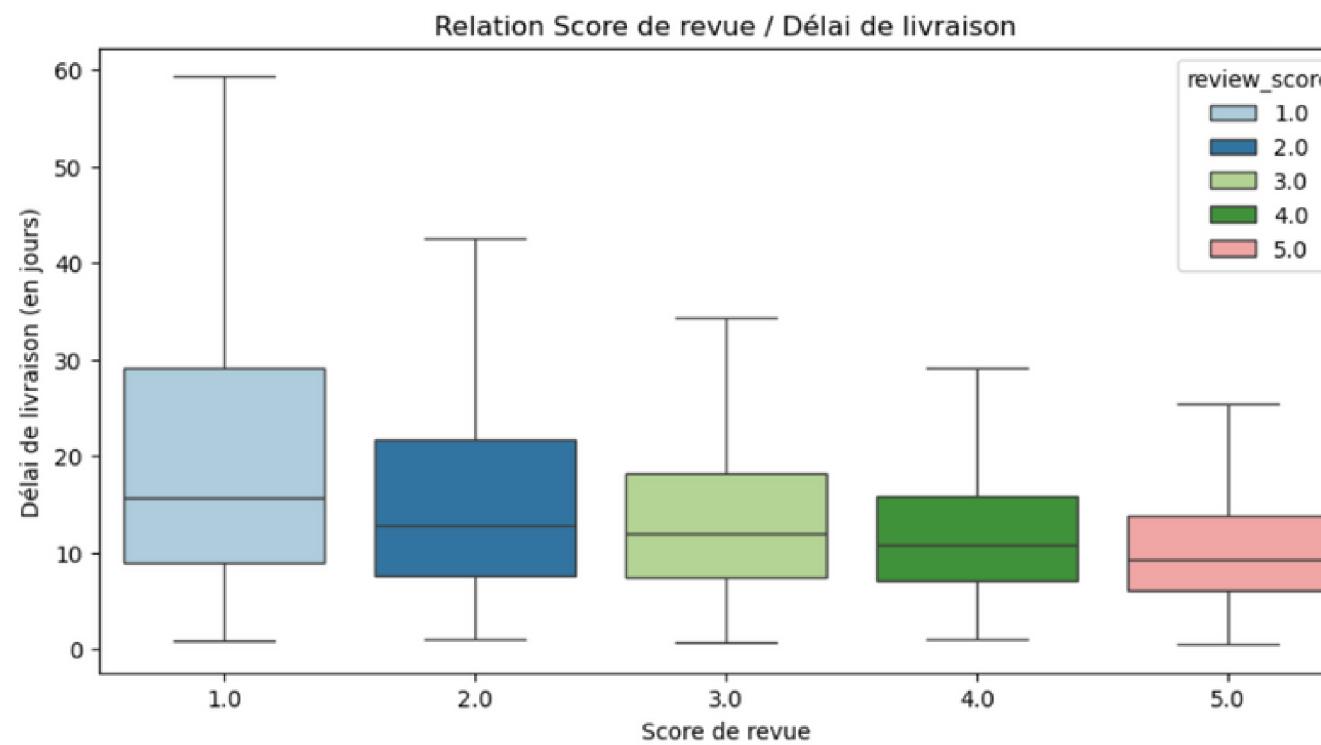
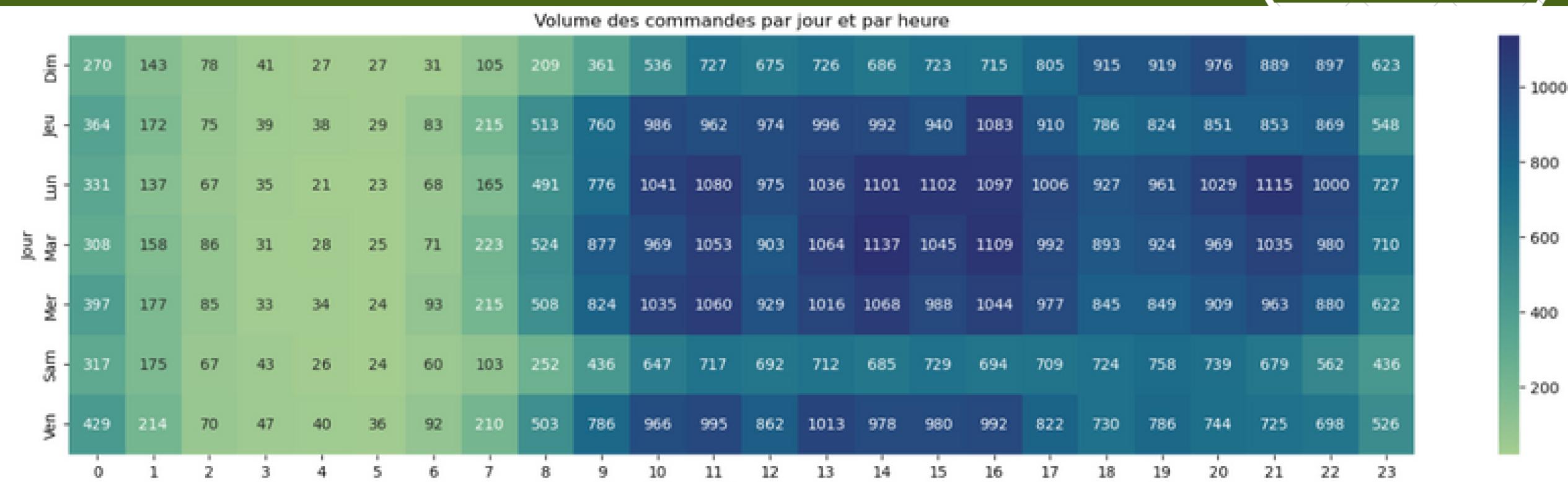




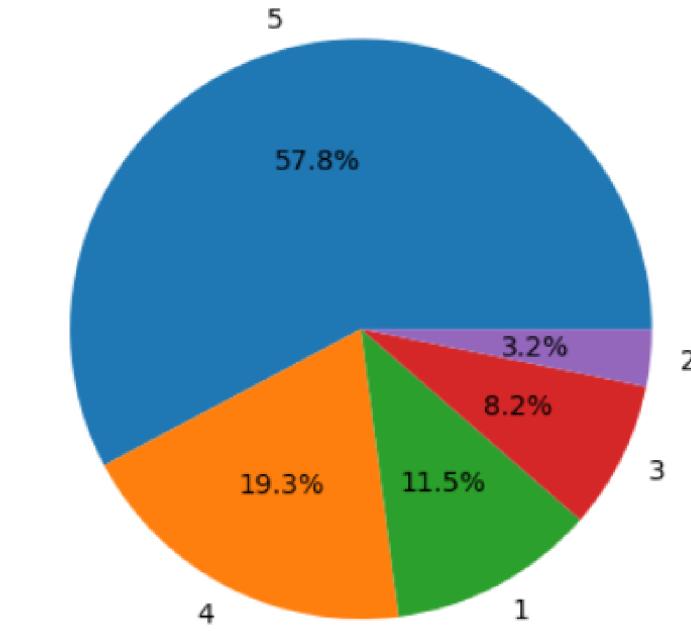
01

*Importation des données,
Exploration et visualisation des données,
Fusion des données dans un data frame unique,
Analyses exploratoires.*

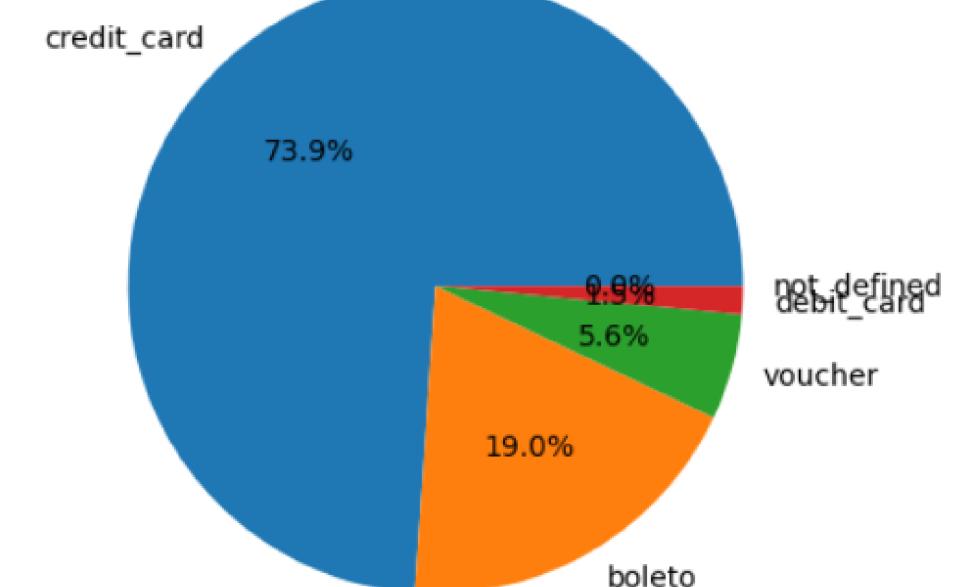
Data frames	lignes	colonnes
customers	99441	5
geolocation	1000163	5
items	112650	7
payments	103886	5
reviews	99224	7
orders	99441	8
products	32951	9
sellers	3095	4
category	71	2
df (données fusionnées)	119143	40



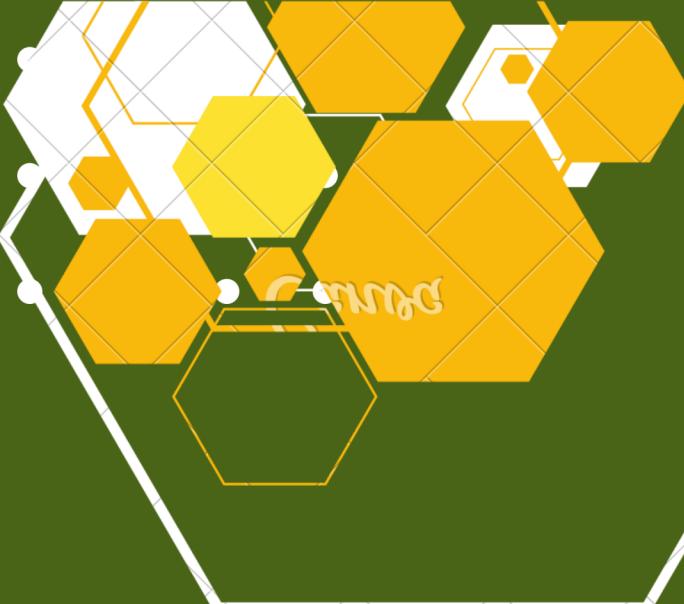
Répartition des clients selon le niveau de satisfaction



Répartition des paiements par type



Feature engineering, Construction des paramètres RFM, Création de nouvelles variables pour le modèle final, Analyses univariées, bivariées, corrélations



• Feature engineering

- => Suppression des doublons
- => Sélection des commandes livrées
- => Conversion des dates au format datetime

• Création de nouvelles variables

Nouvelles variables créées	Description
recency	Nombre de jour depuis le dernier achat du client
frequency	Nombre de commande effectué par le client
monetary	Montant global des commandes du client
freight	Poids des frais de port sur la valeur totale des articles du client
weight	Poids moyen des articles achetés du client
length	longueur moyenne des articles achetés du client
height	hauteur moyenne des articles achetés du client
width	largeur moyenne des articles achetés du client
delivered_delay	Délai de livraison moyen des commdes du client
carrier_delay	Délai de transport moyen des commandes du client
approval_delay	Délai d'approbation moyen des commandes du client
score	Score obtenu par le client à sa dernière commande
credit_card	Paiement par carte de crédit
debit_card	Paiement par carte de débit
voucher	Paiement par bon
small_city	Taille de la ville du client

• Traitement des valeurs manquantes

Variables	Traitement apporté à la valeur manquante identifiée dans la variable
product_weight_g	
product_length_cm	Imputation par la moyenne
product_height_cm	
product_width_cm	
payment_type	Imputation par le mode
review_score	Attribution du score le plus élevé

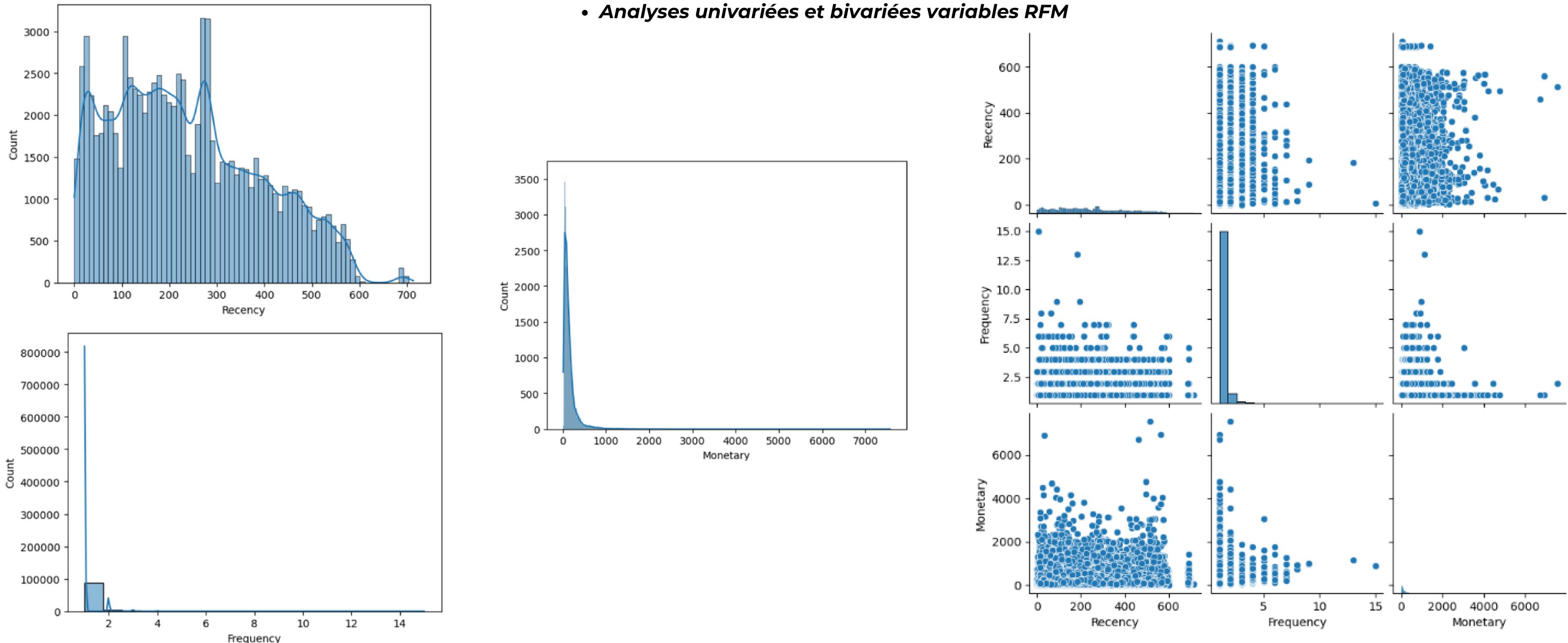
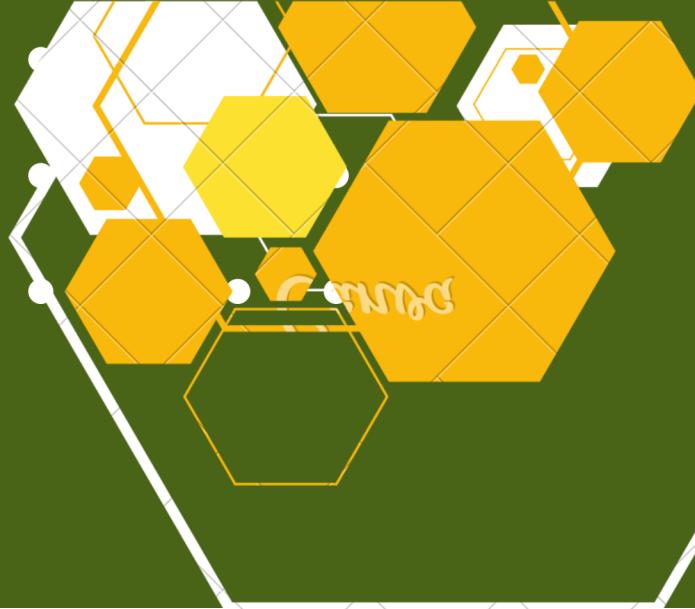
• Encodage des variables catégorielles

Variables catégorielles encodées	Méthodologie utilisée
customer_city	One Hot Encoding avec suppression de la première modalité
payment_type	One Hot Encoding avec suppression de la première modalité

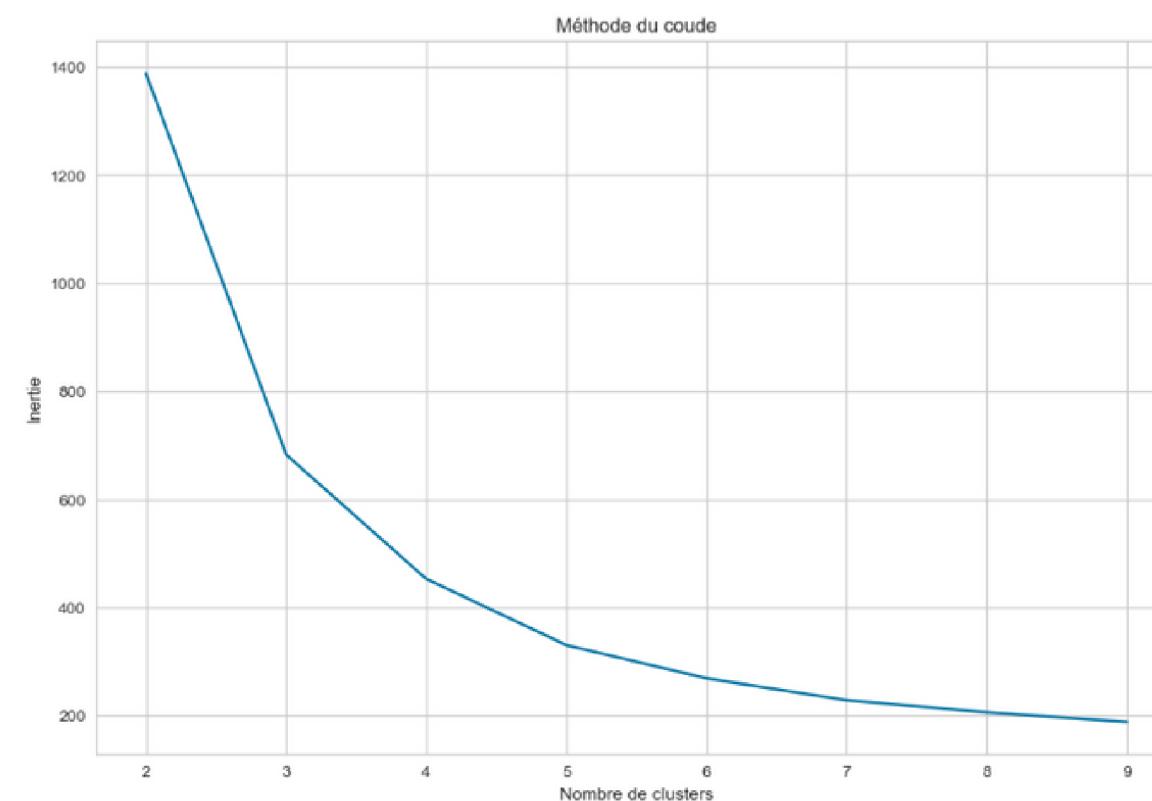
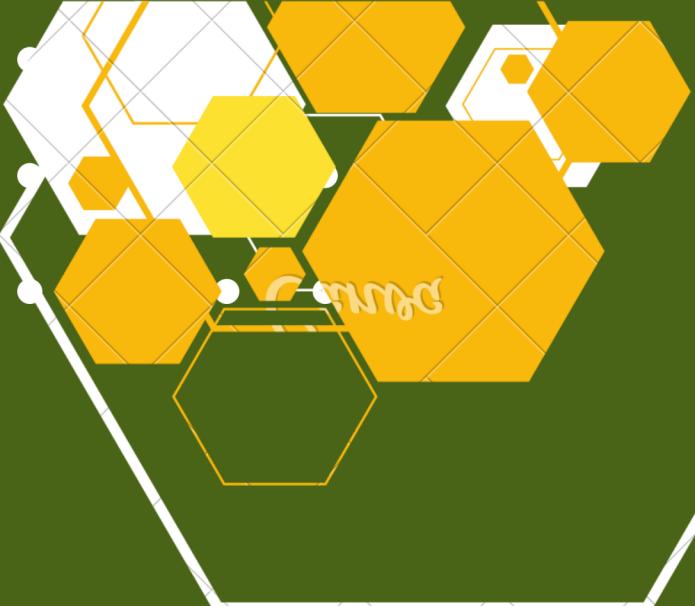
• Récapitulatif des variables RFM

	Recency	Frequency	Monetary
count	93358.000000	93358.000000	93358.000000
mean	236.941773	1.073245	154.354920
std	152.591453	0.328448	206.382393
min	0.000000	1.000000	9.590000
25%	113.000000	1.000000	59.780000
50%	218.000000	1.000000	101.635000
75%	345.000000	1.000000	171.840000
max	713.000000	15.000000	7571.630000

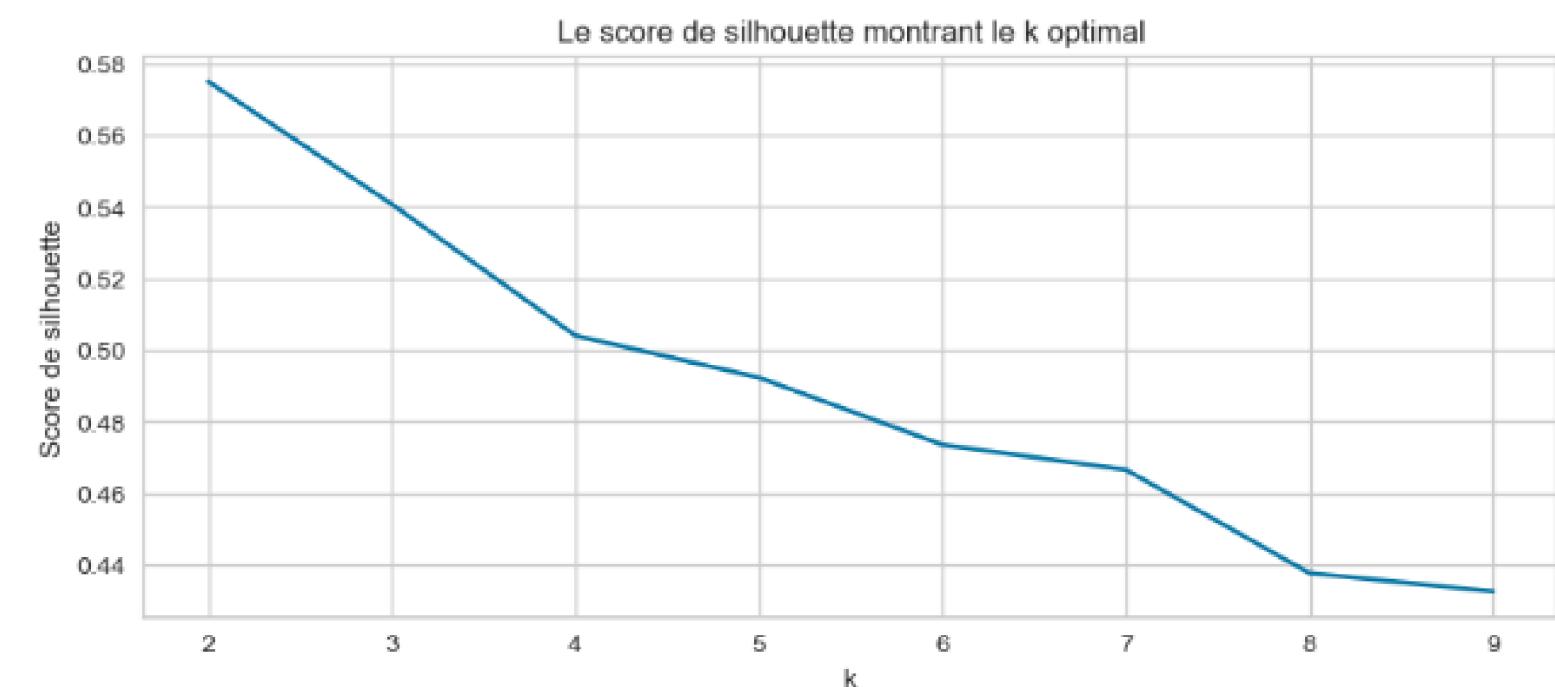
Feature engineering,
Construction des paramètres RFM,
Création de nouvelles variables pour le modèle final,
Analyses univariées, bivariées, corrélations



Construction des modèles d'essais à partir des variables RFM (KMeans, CAH et DBSCAN), Analyse et interprétation des clusters, Conclusion et choix du modèle final.



MODELISATION VIA KMeans



Caractéristiques des centroïdes

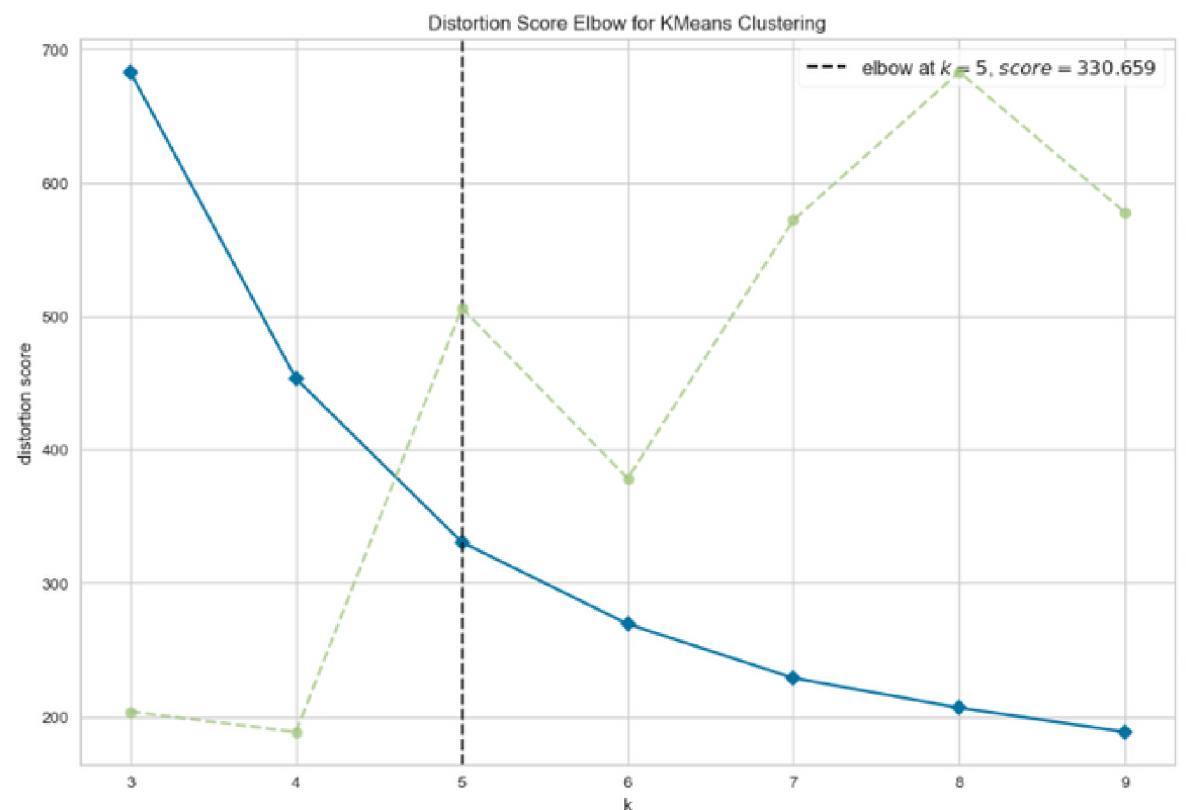
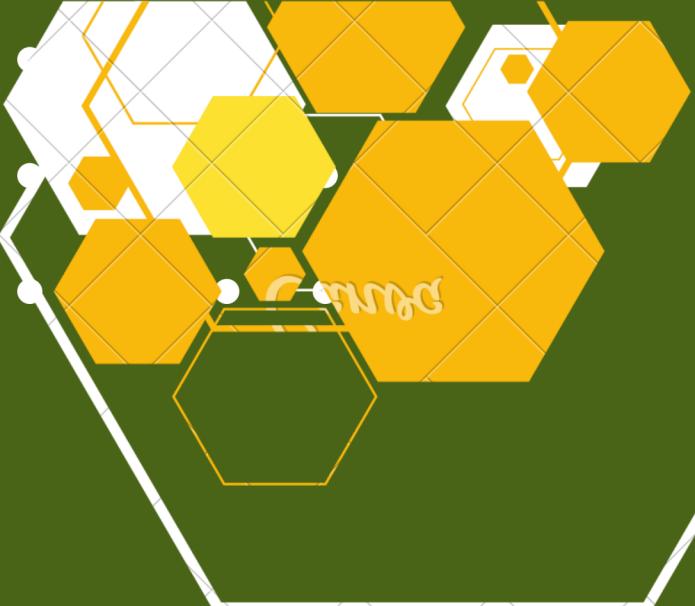
Recency Frequency Monetary cluster_label

cluster_label	Recency	Frequency	Monetary	cluster_label
0	182.797829	1.071785	150.268375	0.0
1	482.547901	1.062865	154.105147	1.0
2	61.339699	1.080682	159.170138	2.0
3	313.034770	1.074195	154.047119	3.0

Interprétation des clusters

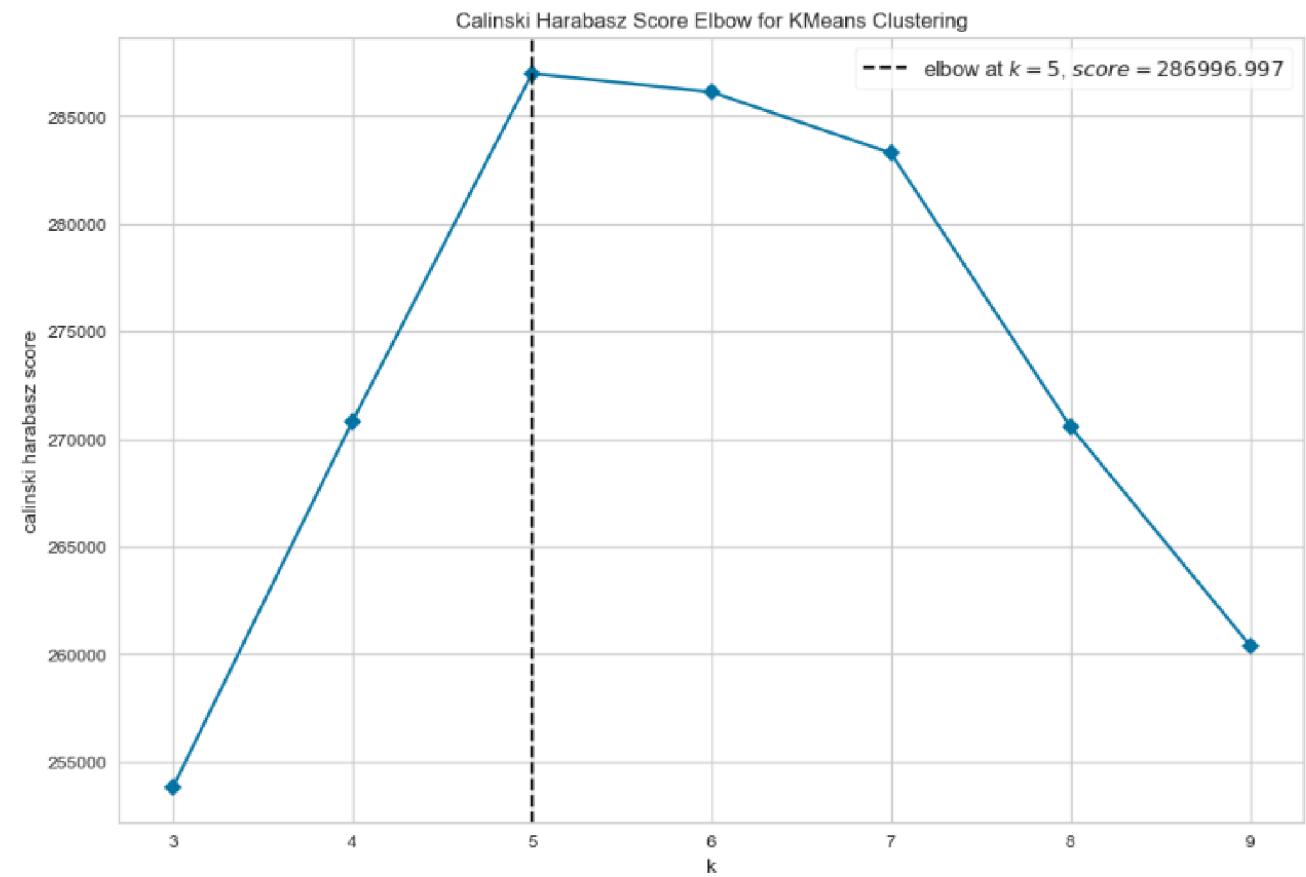
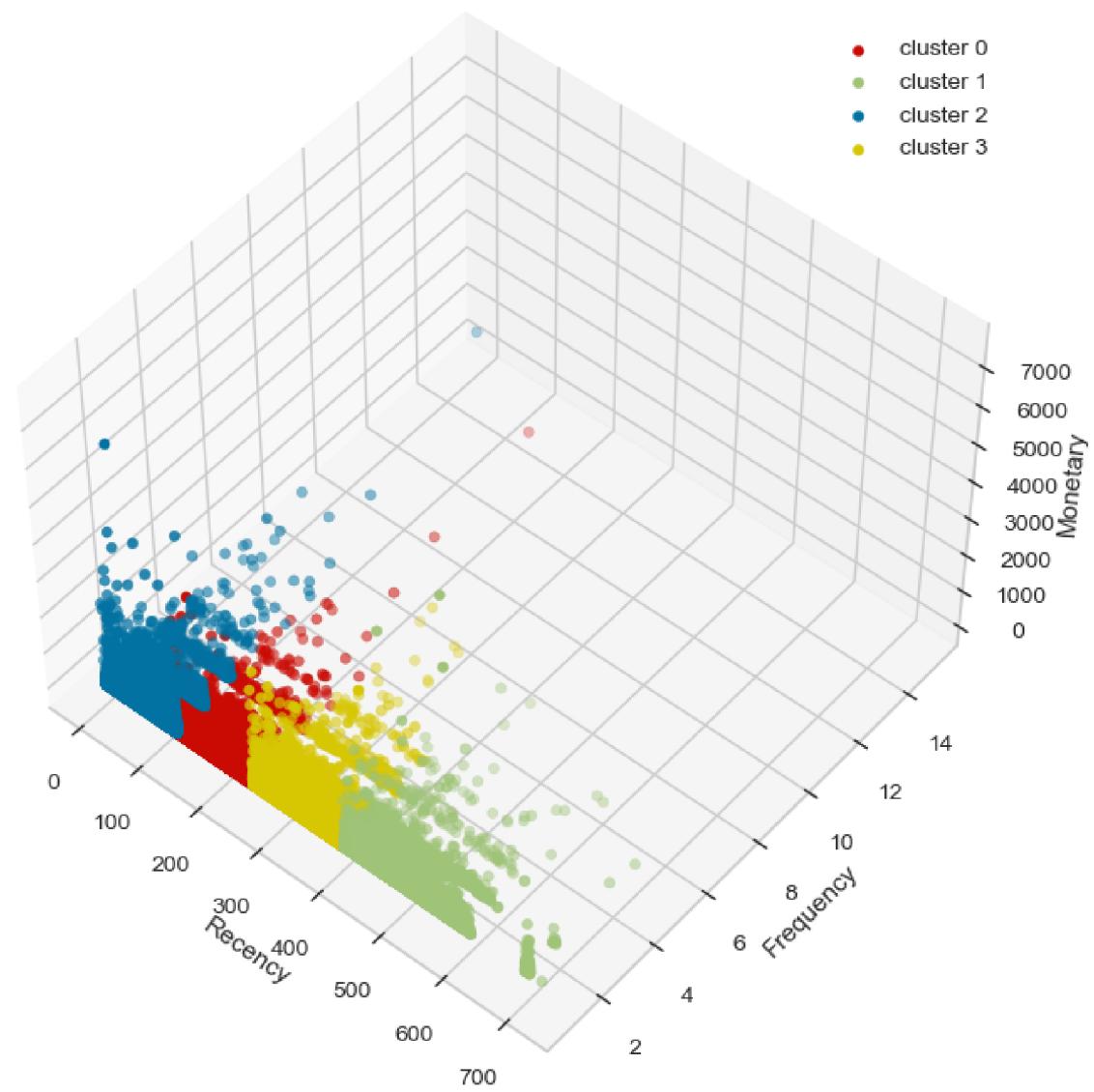
- **Cluster 0 : Clients à réactiver**
- **Cluster 1 : Clients perdus**
- **Cluster 2 : Clients fidèles à forte capacité d'achat**
- **Cluster 3 : Clients à risque**

Construction des modèles d'essais à partir des variables
RFM (KMeans, CAH et DBSCAN),
Analyse et interprétation des clusters,
Conclusion et choix du modèle final.



MODELISATION VIA KMeans

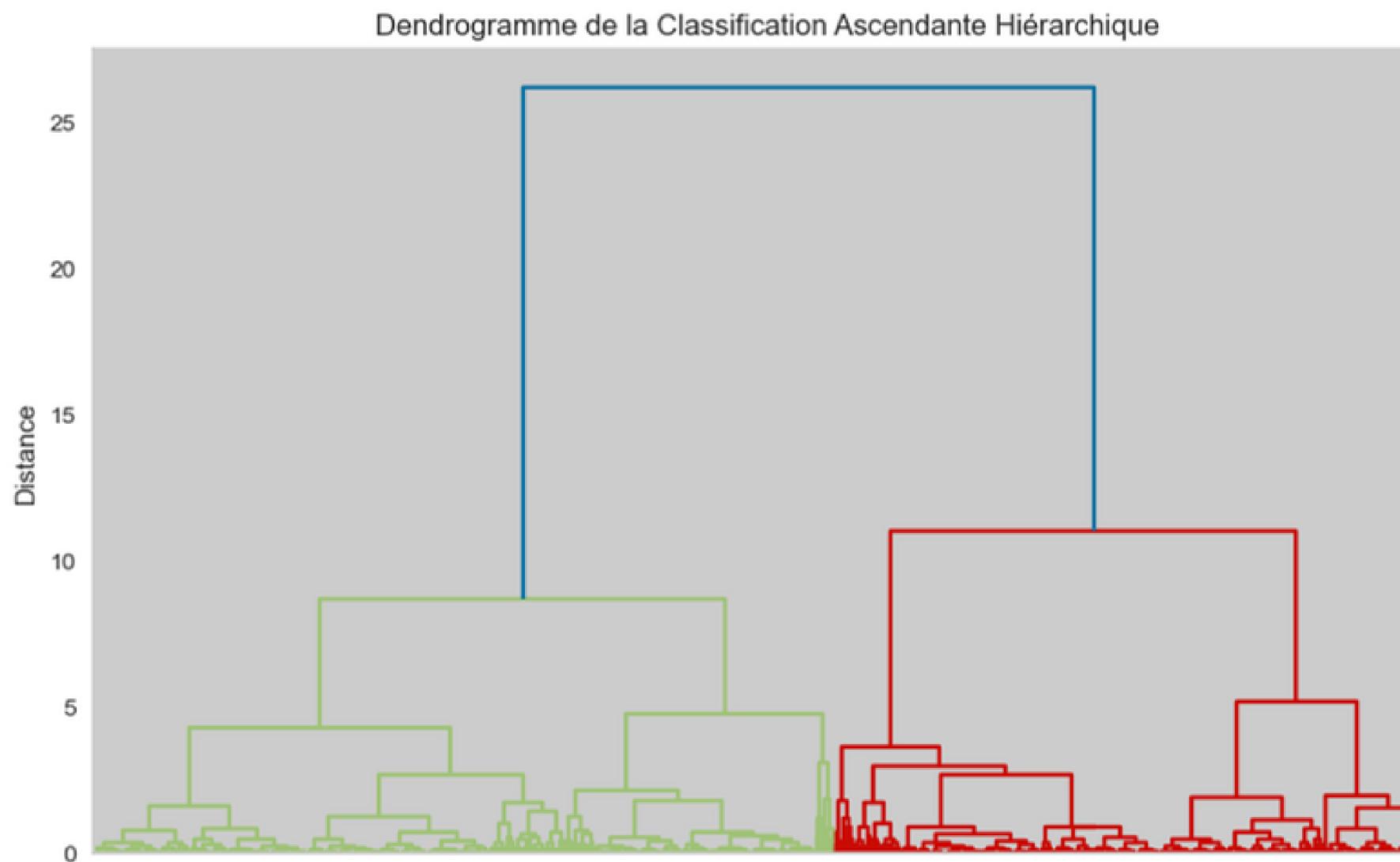
- **Evaluation et visualisation**



Construction des modèles d'essais à partir des variables
RFM (KMeans, CAH et DBSCAN),
Analyse et interprétation des clusters,
Conclusion et choix du modèle final.



MODELISATION VIA CAH (Traitement réalisé sur un échantillon de 10000 clients)

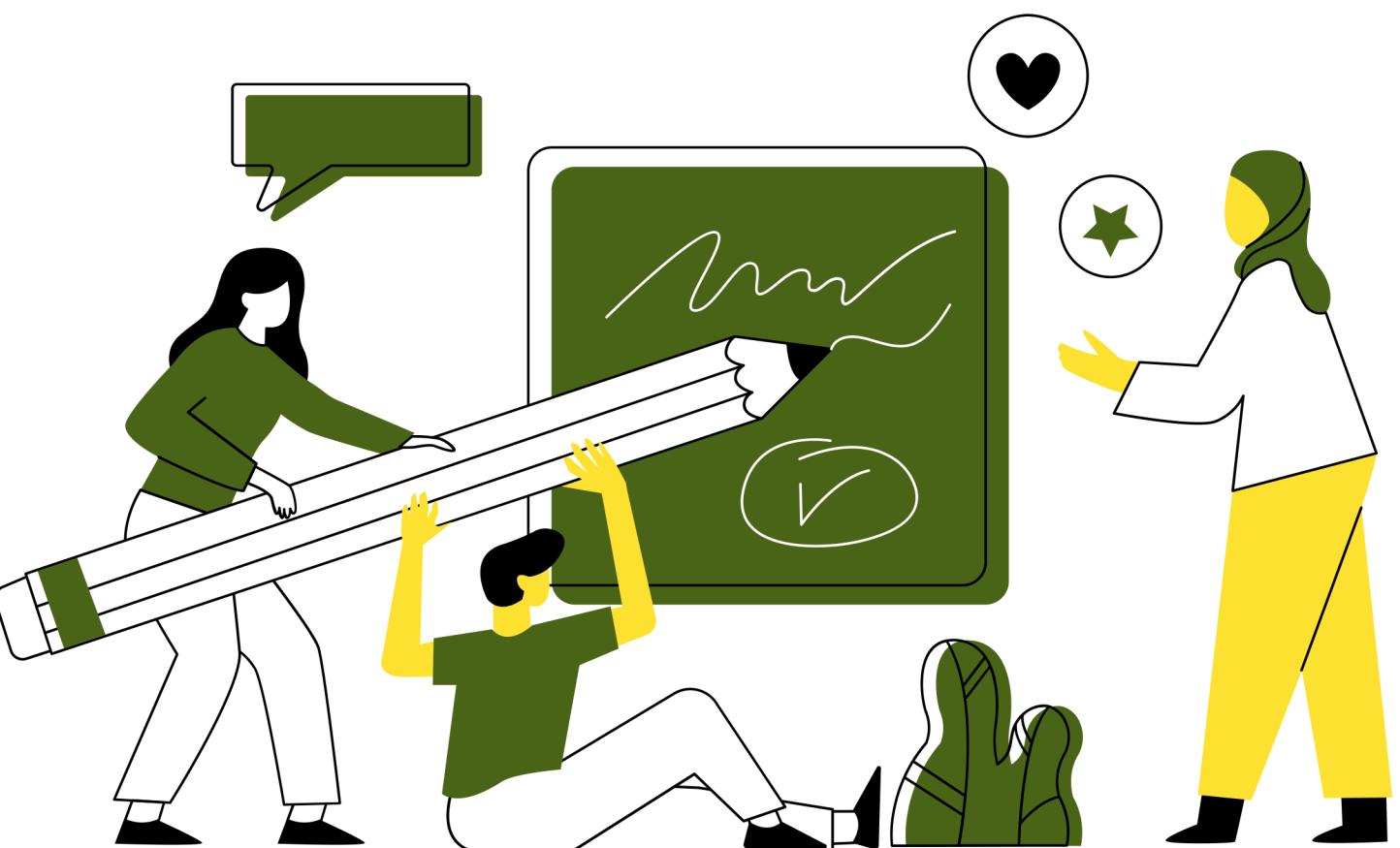
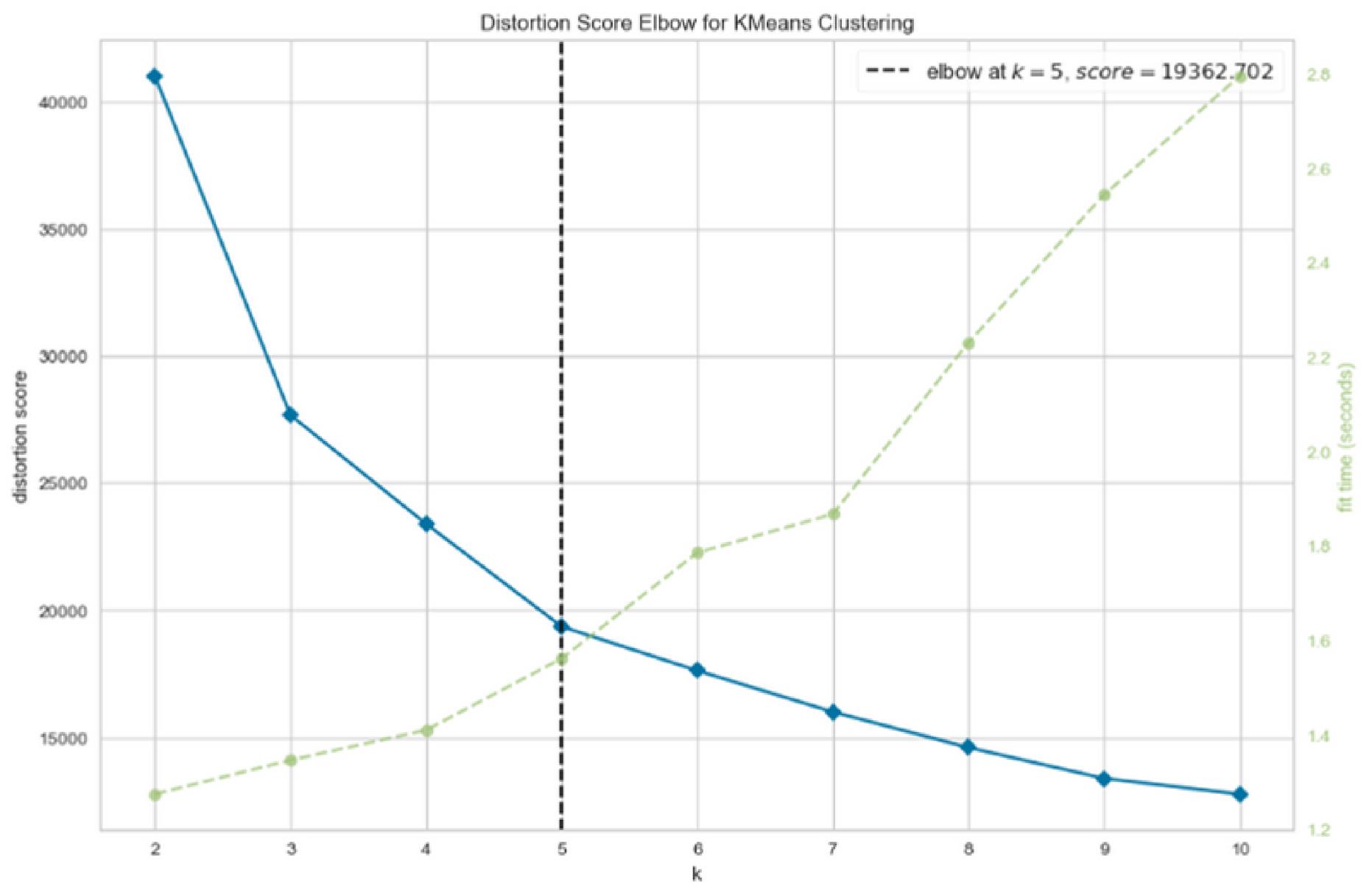
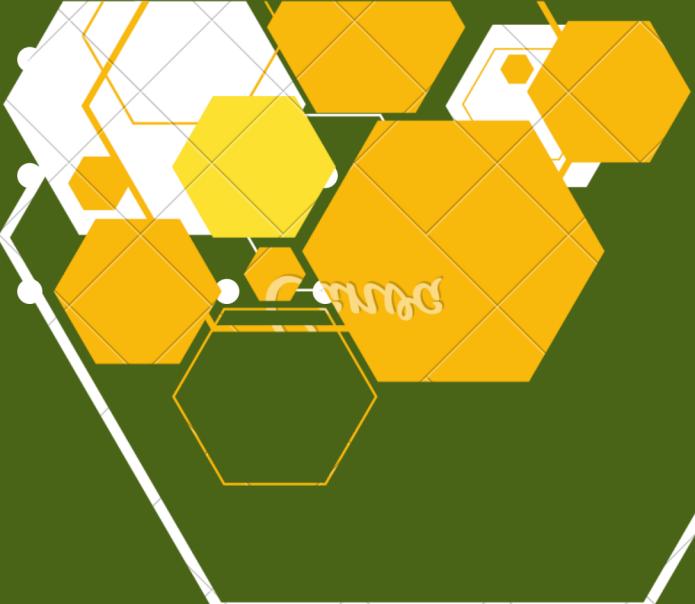


MODELISATION VIA DBSCAN (Traitement réalisé sur un échantillon de 10000 clients)

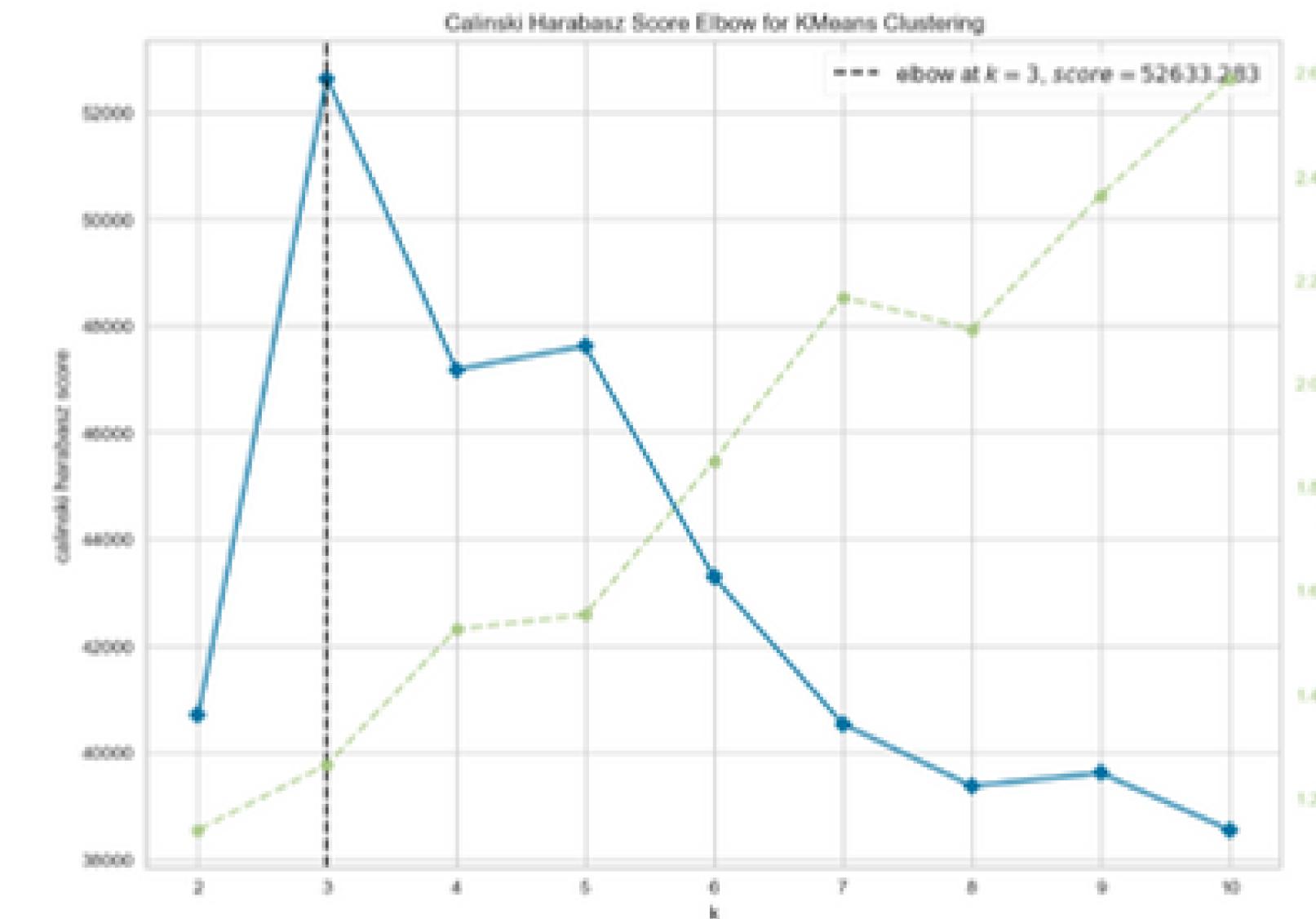
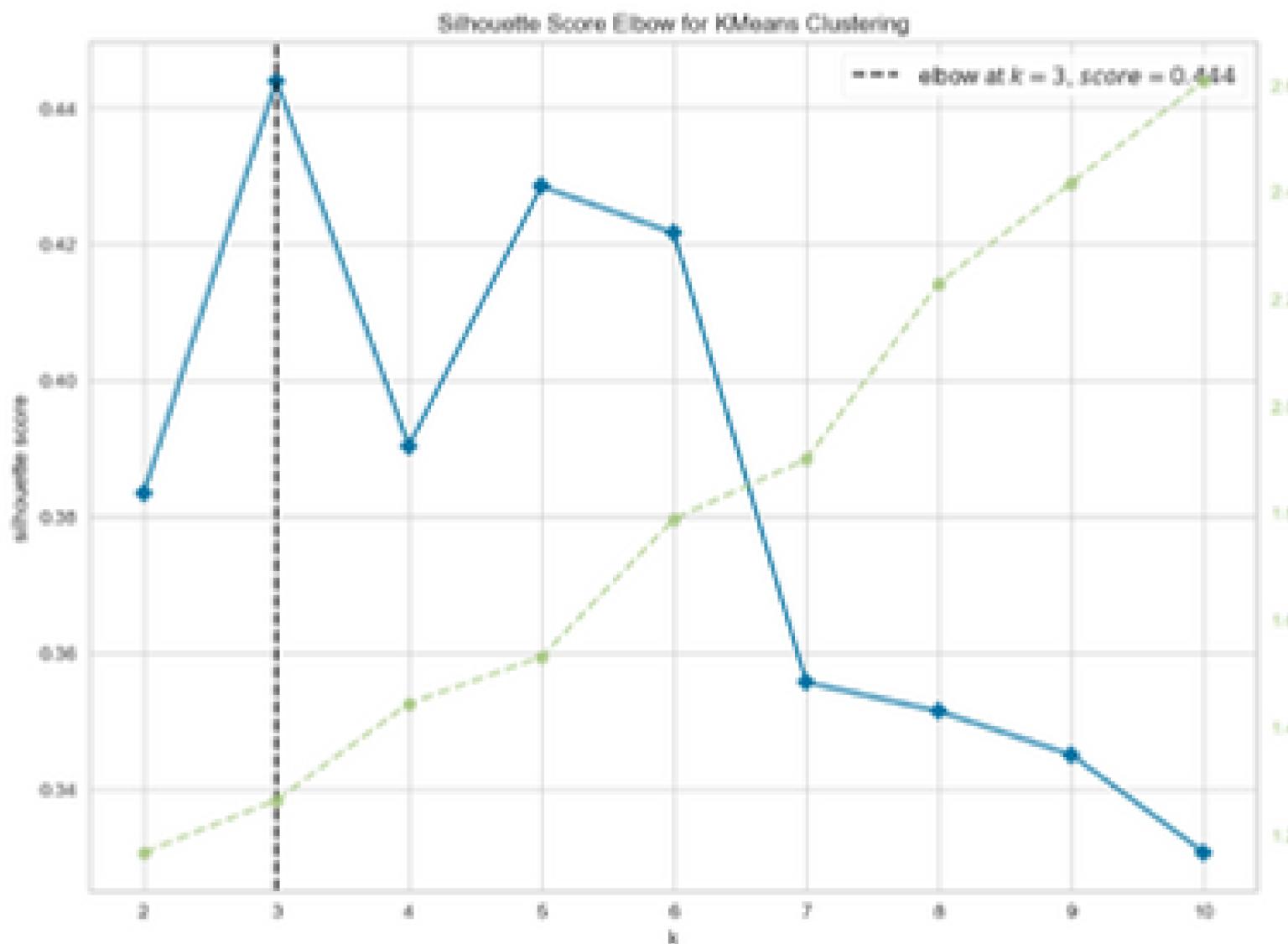
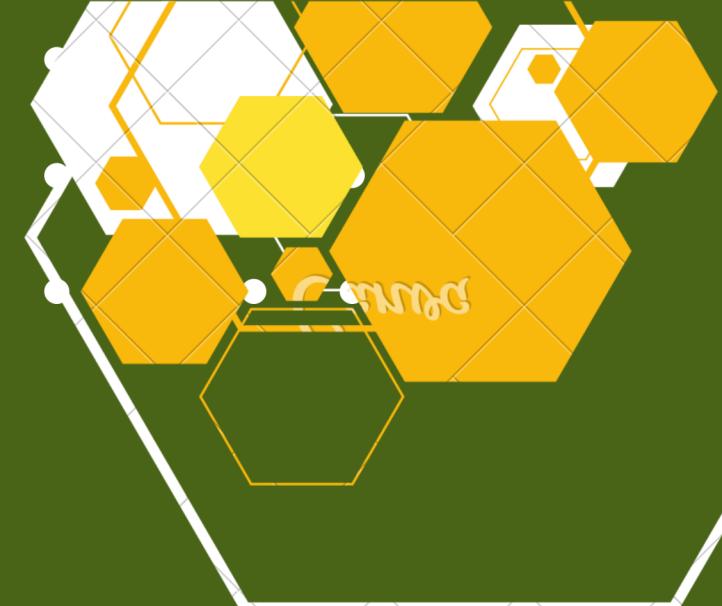
- Sur 10000 clients un seul cluster formé composé de 9824 clients et 176 considérés comme du bruit et non affectés à un cluster

CONCLUSION : Compte tenu de la difficulté d'application sur l'ensemble des clients, CAH et DBSCAN ne seront pas retenus comme modèle à construire pour la modélisation finale. Nous retenons donc KMeans.

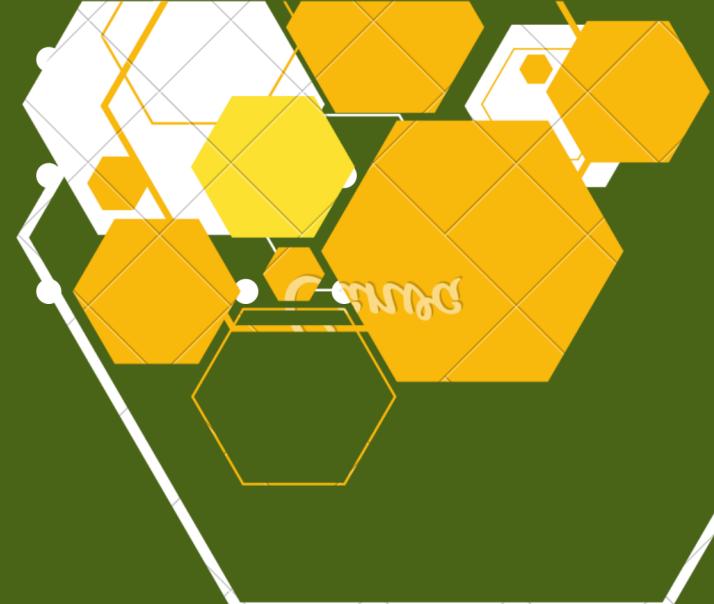
*Construction du modèle final,
Analyse et interprétation des clusters,
Réduction de dimensions via ACP,
Interprétation et évaluation du modèle à dimensions réduites,
Analyse de la stabilité à l'initialisation.*



*Construction du modèle final,
Analyse et interprétation des clusters,
Réduction de dimensions via ACP,
Interprétation et évaluation du modèle à dimensions réduites,
Analyse de la stabilité à l'initialisation.*



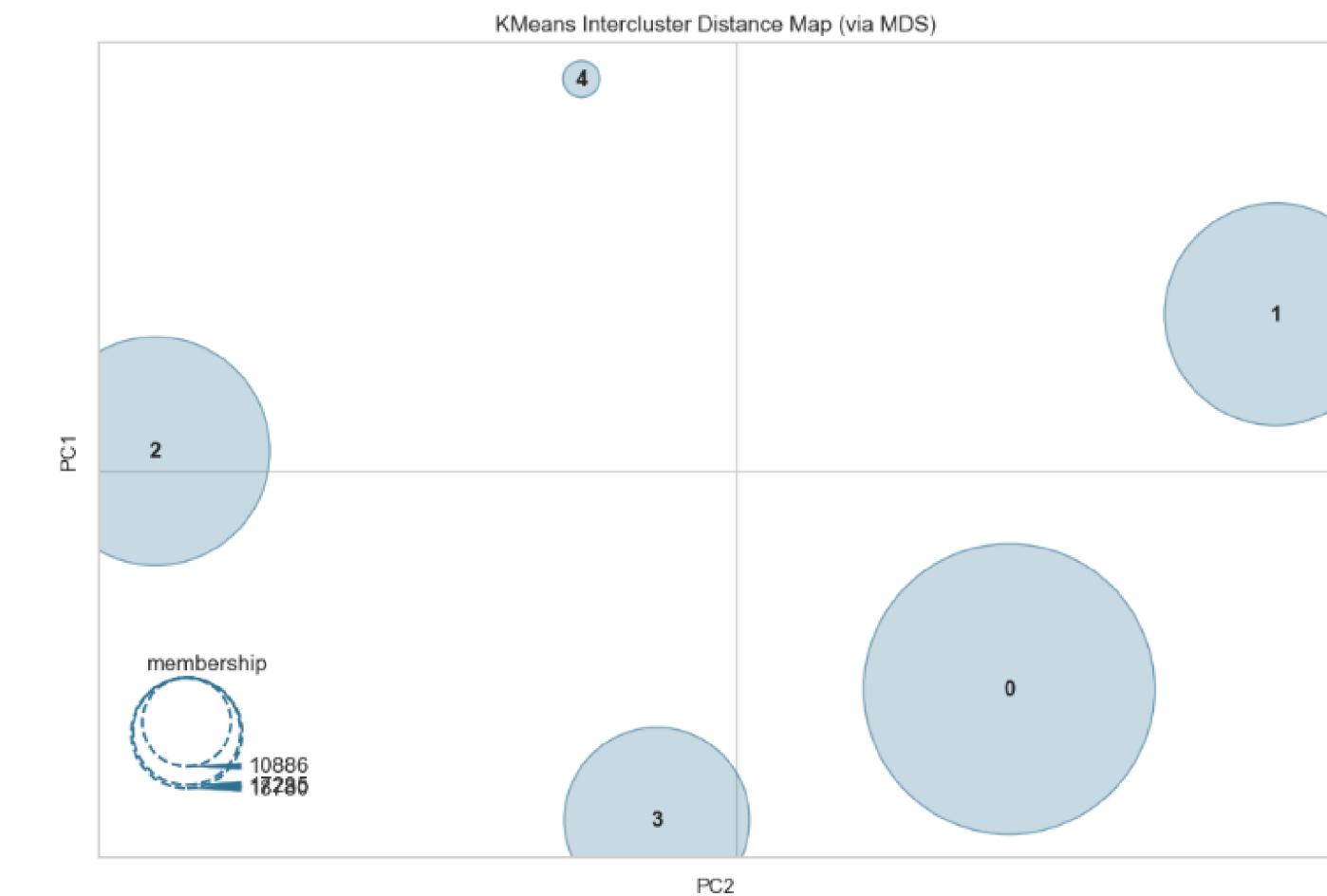
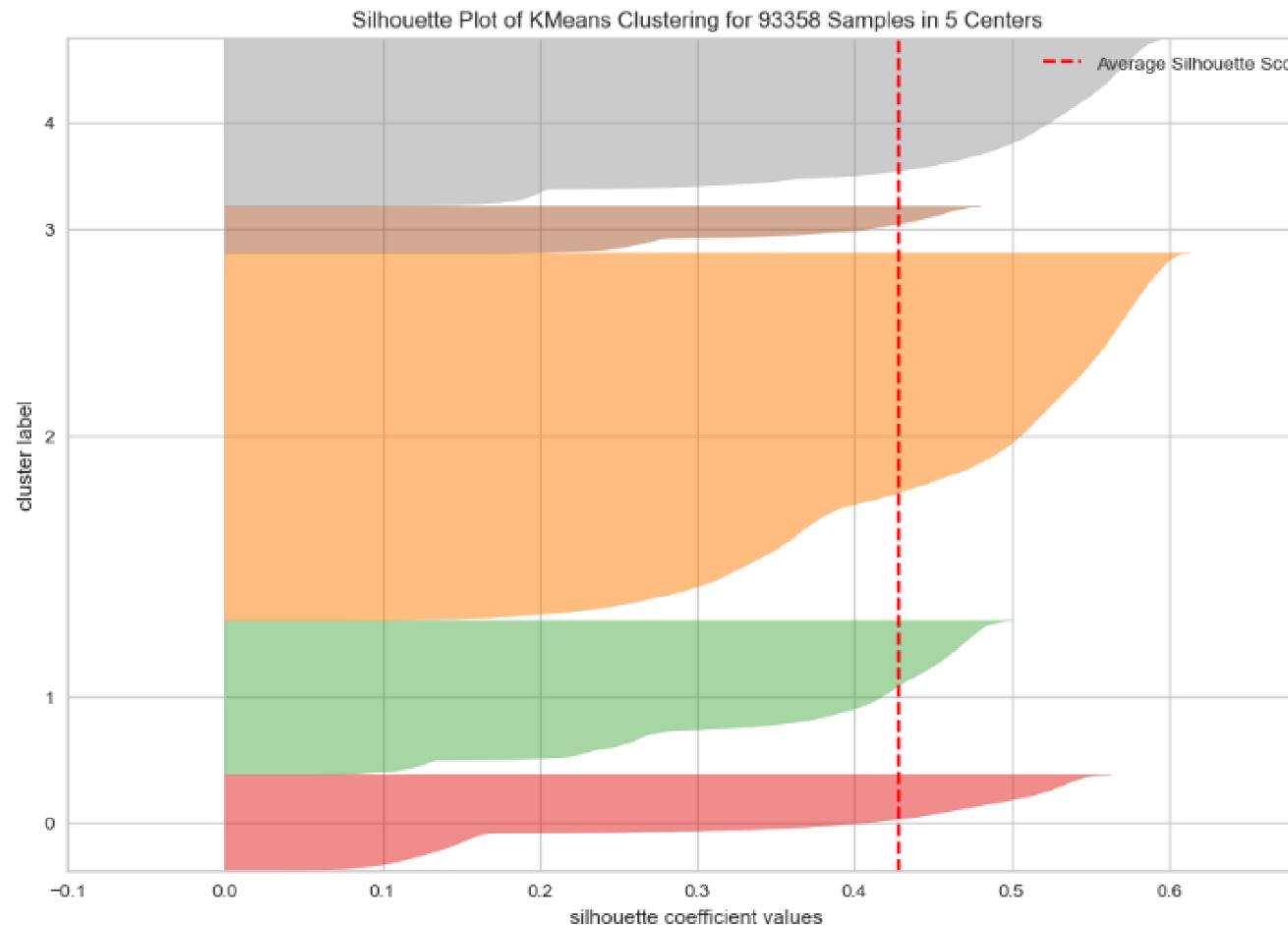
*Construction du modèle final,
Analyse et interprétation des clusters,
Réduction de dimensions via ACP,
Interprétation et évaluation du modèle à dimensions réduites,
Analyse de la stabilité à l'initialisation.*



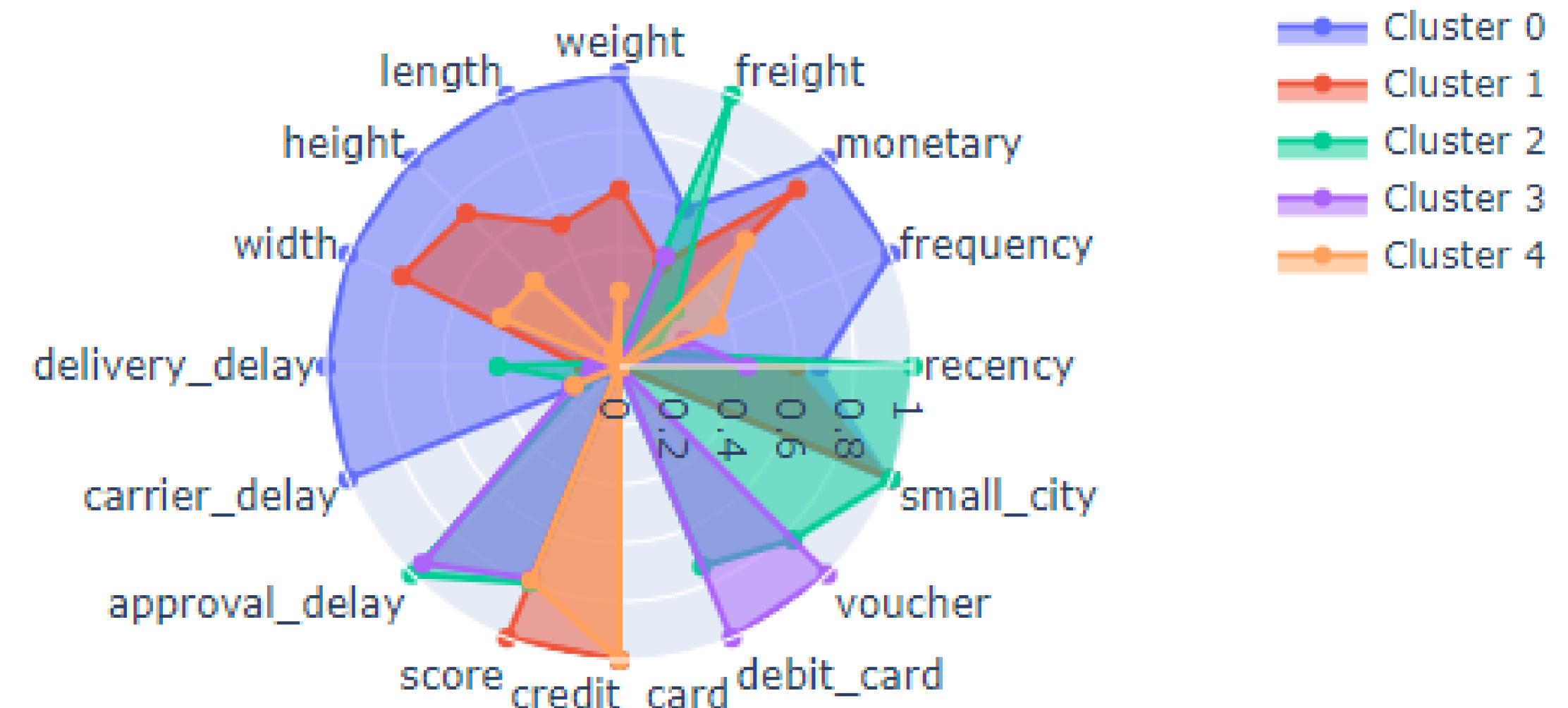
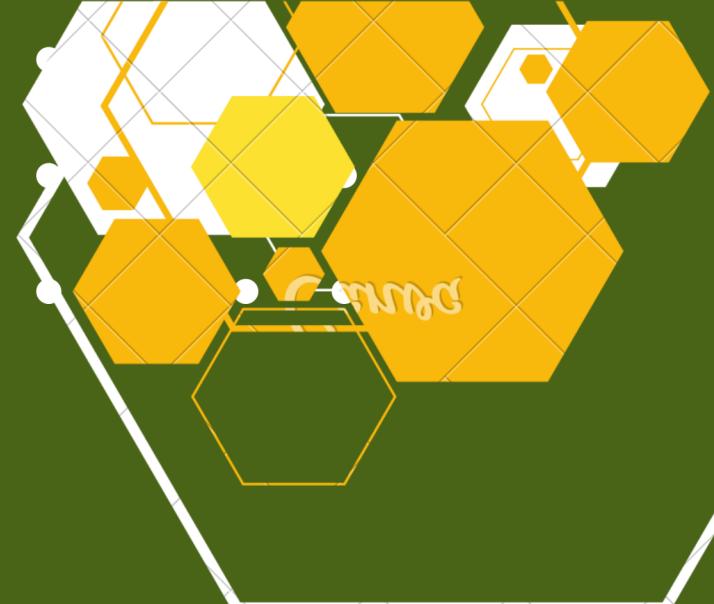
- **Caractéristiques des centroïdes**

kmeans_label	recency	frequency	monetary	freight	weight	length	height	width	delivery_delay	carrier_delay	approval_delay	score	credit_card	debit_card	voucher	small_city
0	239.256384	1.117858	169.049324	0.318700	2323.309679	30.727821	17.032460	23.374066	18.303558	3.684787	0.202888	1.914595	1.0	0.000000	0.000000	1.0
1	237.807530	1.060200	162.698344	0.300182	2167.179673	30.241916	16.724980	23.205057	11.577177	2.593872	0.184216	4.749102	1.0	0.000000	0.000000	1.0
2	244.983000	1.067363	136.872992	0.357293	1943.202900	29.708603	15.859528	22.544369	13.796387	2.738025	1.208295	4.171976	0.0	0.059472	0.105251	1.0
3	234.780840	1.074130	125.055307	0.302739	1929.501521	29.753849	15.922638	22.515000	11.357566	2.788709	1.154009	4.130584	0.0	0.080783	0.126022	0.0
4	226.926624	1.081150	151.786620	0.265453	2030.575598	29.763873	16.341495	22.889961	10.622912	2.776160	0.179508	4.154313	1.0	0.000000	0.000000	0.0

- **Visualisations**

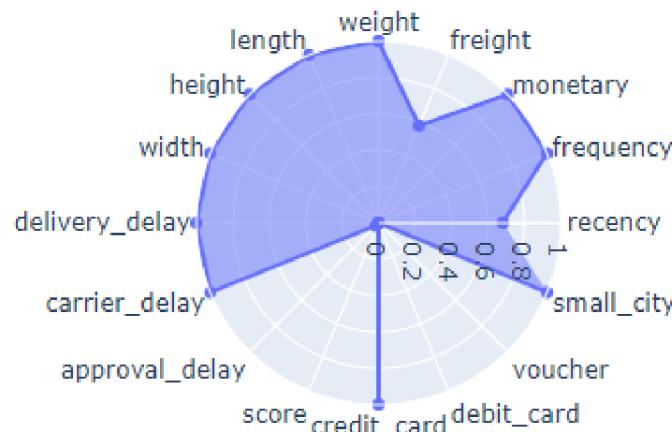
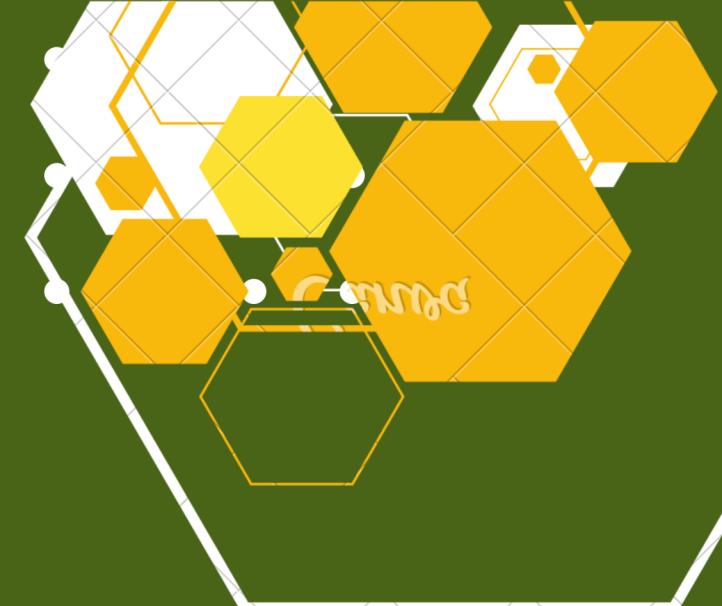


*Construction du modèle final,
Analyse et interprétation des clusters,
Réduction de dimensions via ACP,
Interprétation et évaluation du modèle à dimensions réduites,
Analyse de la stabilité à l'initialisation.*



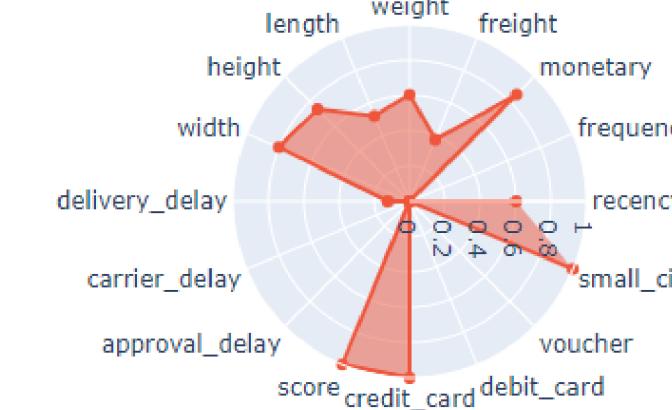
Comparaison des moyennes par variable des clusters

Construction du modèle final, Analyse et interprétation des clusters, Réduction de dimensions via ACP, Interprétation et évaluation du modèle à dimensions réduites, Analyse de la stabilité à l'initialisation.



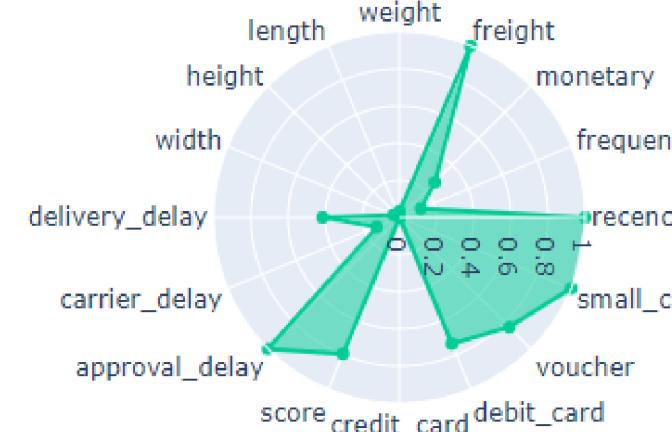
Cluster 0

Ce sont les clients ayant le panier moyen le plus élevé, ils résident dans de petites villes. Ils achètent des articles volumineux, ont une bonne fréquence d'achat et paient des frais de port considérables. Leurs délais de livraison et de transport sont les plus élevés. Ils règlent par carte de crédit et sont très insatisfaits de leur expérience d'achat.



Cluster 1

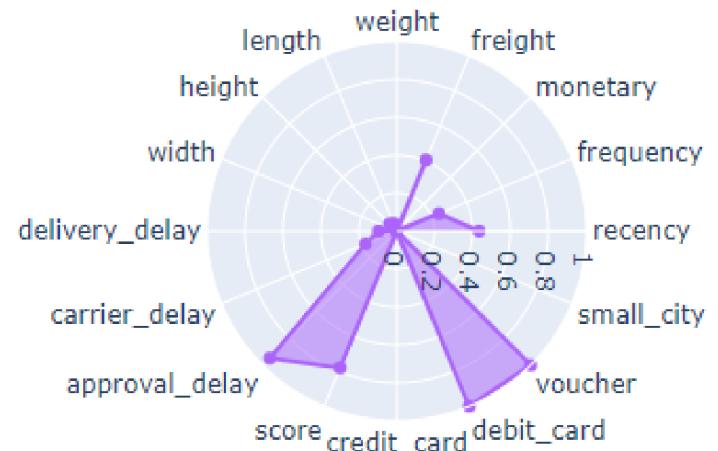
Ce sont les clients ayant un panier moyen élevé, résidant dans de petites villes. Ils règlent par carte de crédit et sont très satisfaits de leur expérience d'achat.



Cluster 2

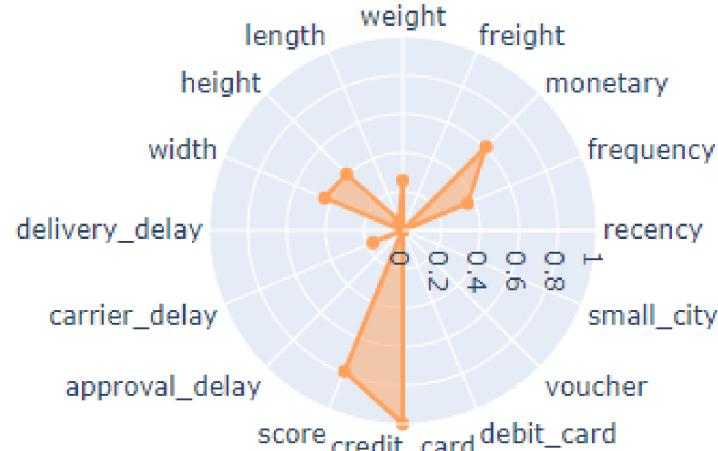
Ce sont les clients qui paient les frais de port les plus élevés. Leurs achats sont les plus récents et les délais d'approbation sont les plus élevés. Ils résident dans de petites villes, règlent par carte de débit ou par bons. Ils sont satisfaits de leur expérience d'achat.

*Construction du modèle final,
Analyse et interprétation des clusters,
Réduction de dimensions via ACP,
Interprétation et évaluation du modèle à dimensions réduites,
Analyse de la stabilité à l'initialisation.*



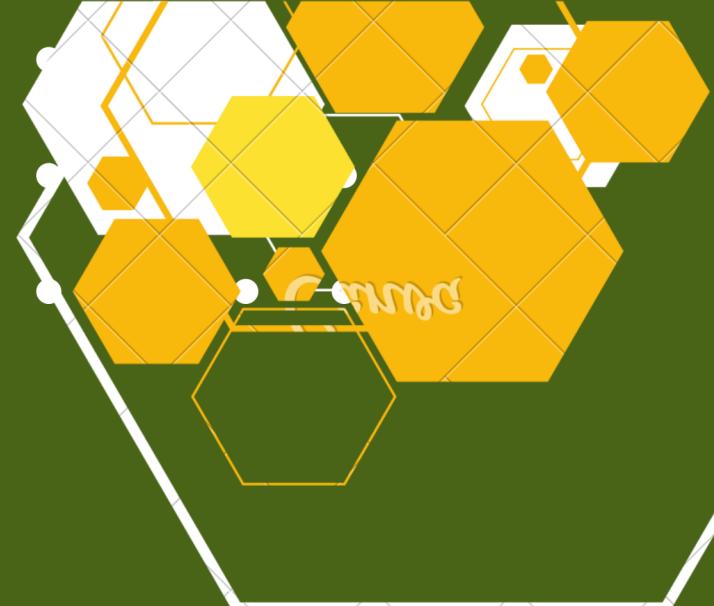
Cluster 3

Ce sont les clients qui ont un panier moyen faible, ils résident dans de grandes villes. Leurs achats ne sont pas fréquents. Ils règlent soit par bons ou carte de débit, ont un délai d'approbation élevé et sont satisfaits de leur expérience d'achat.

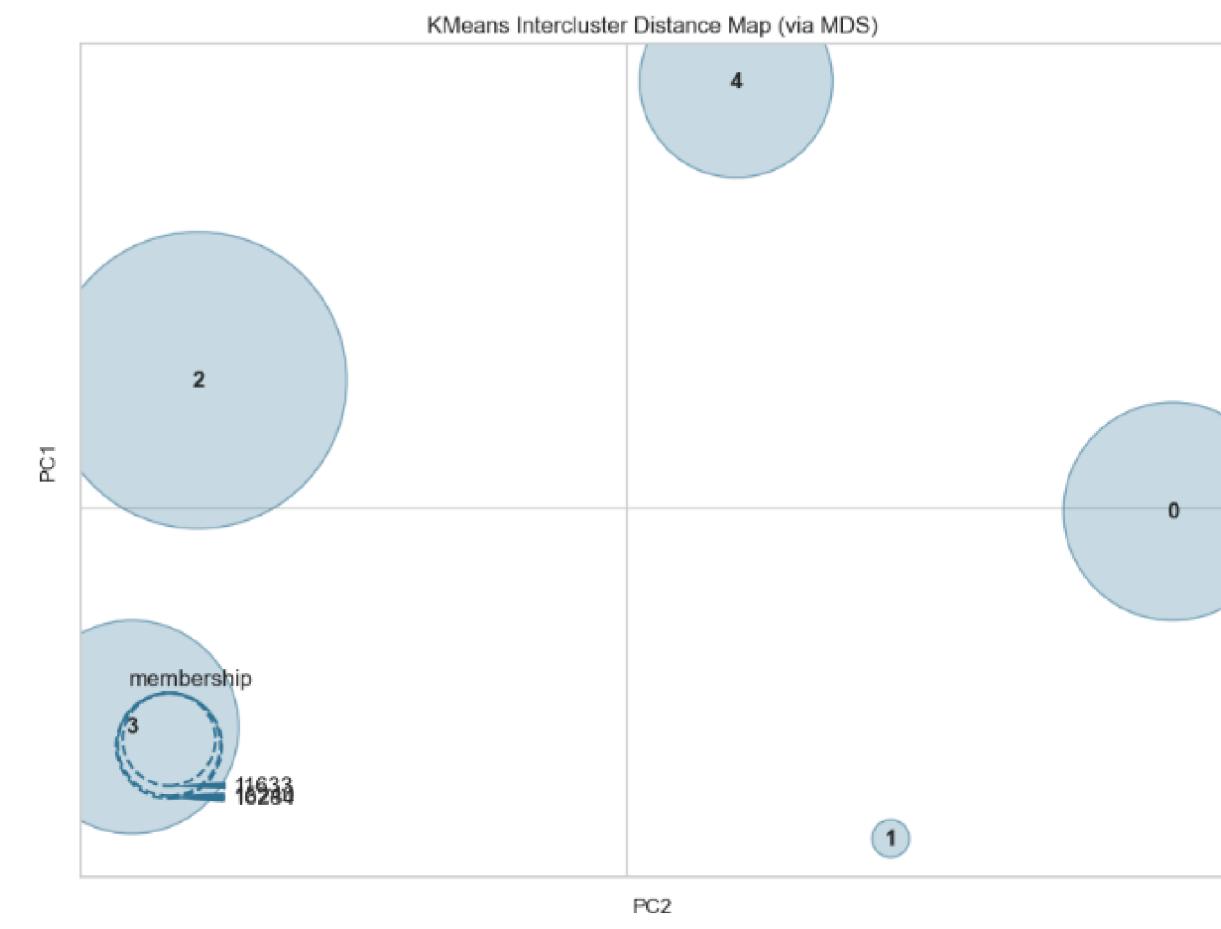
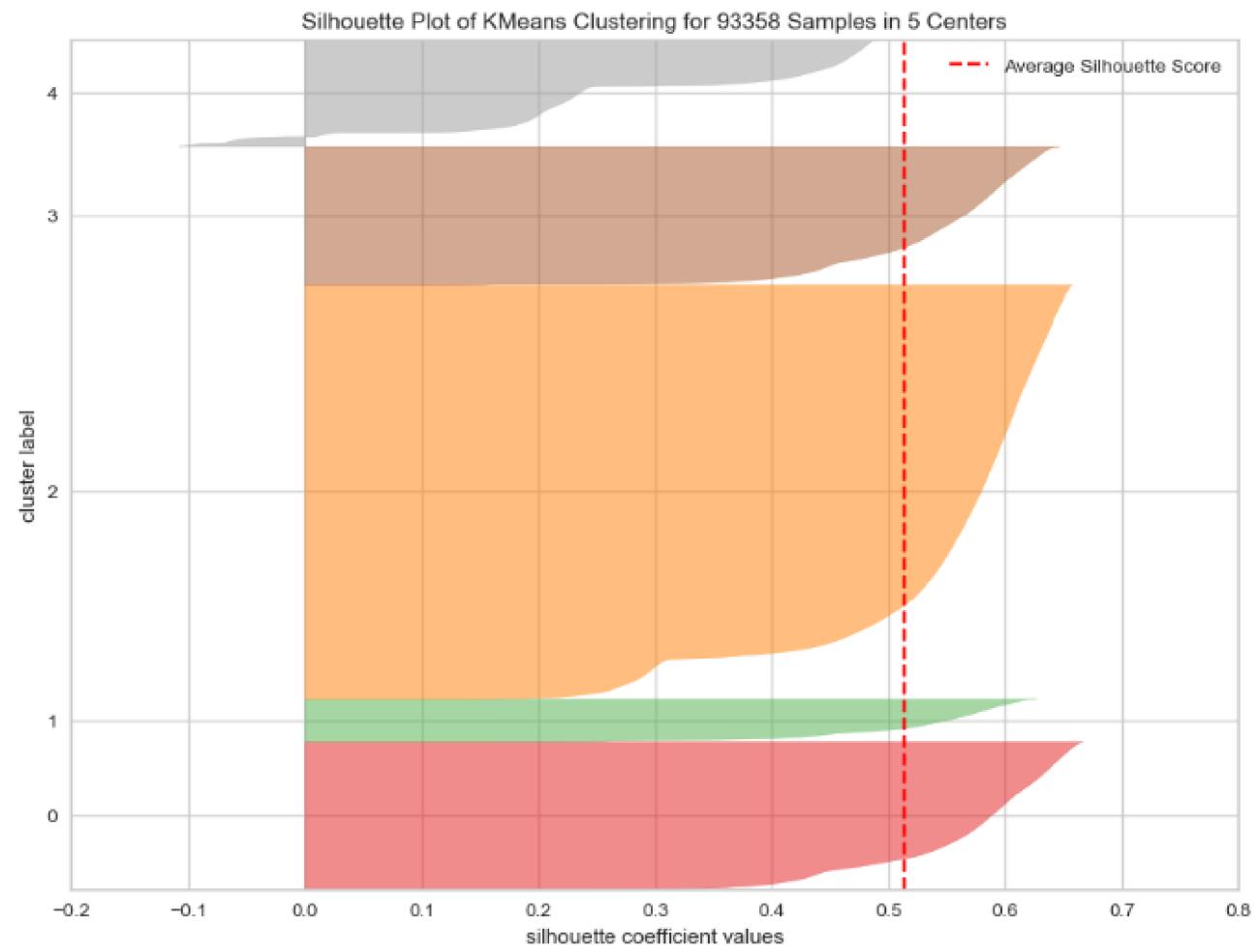
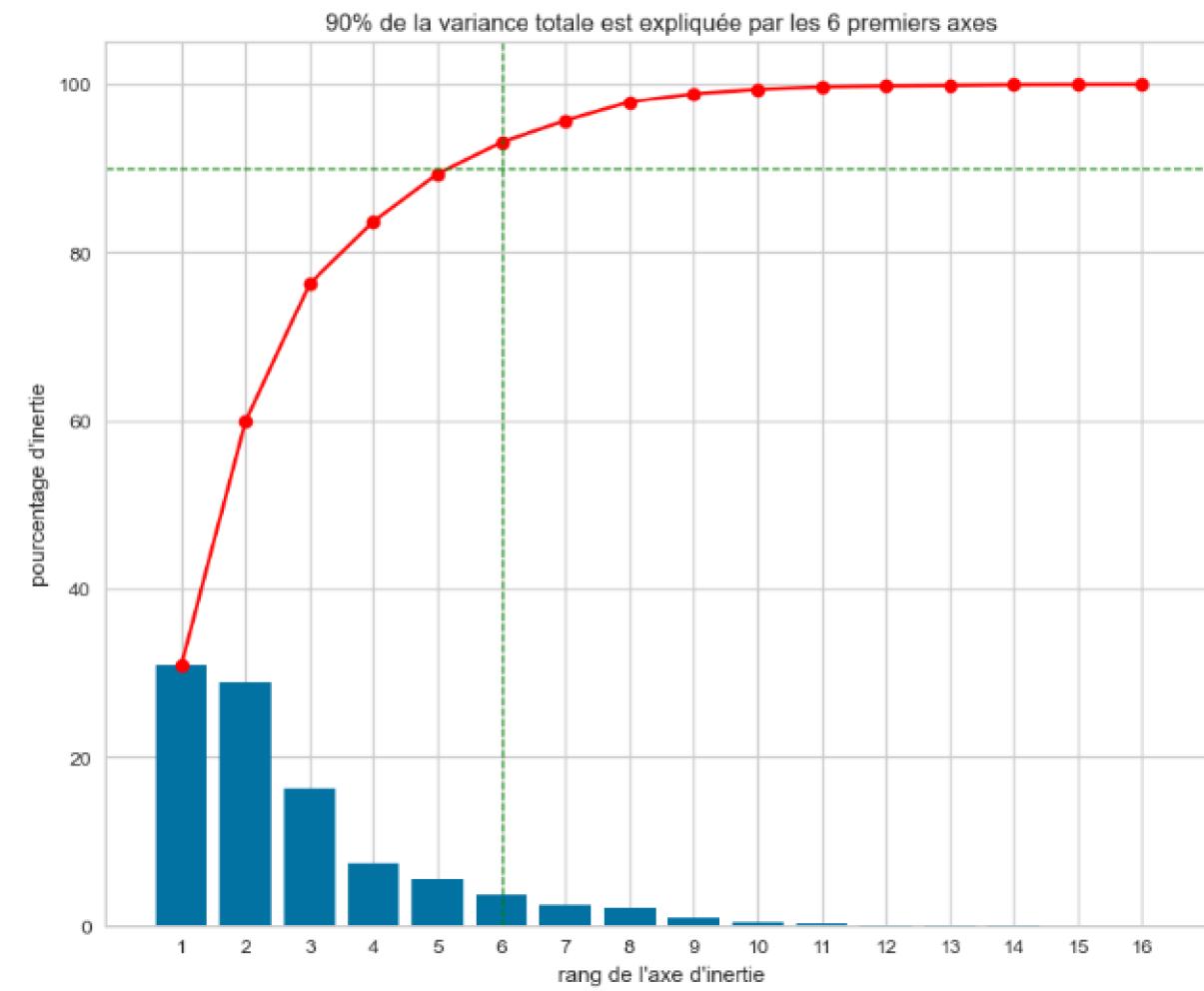


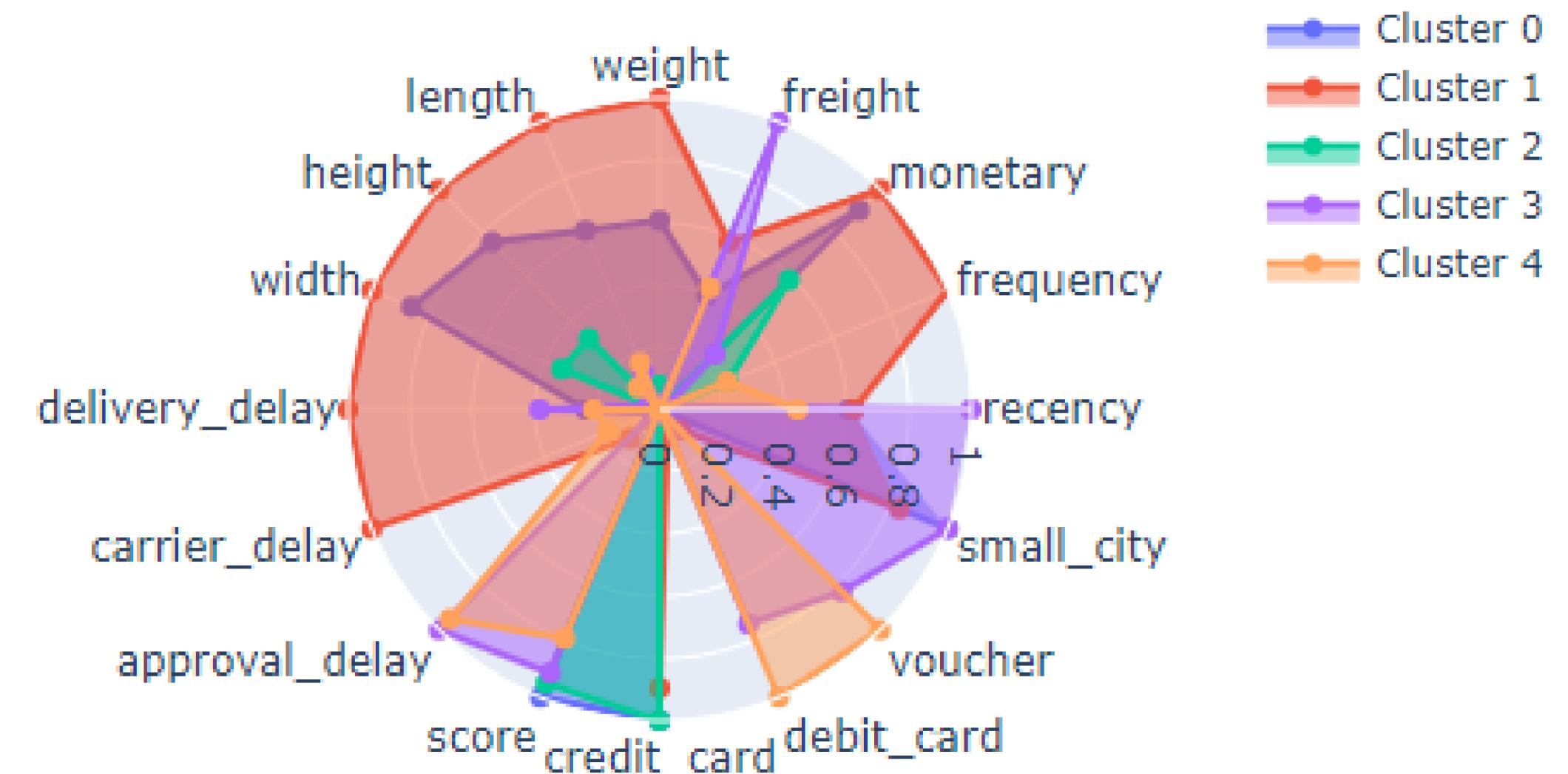
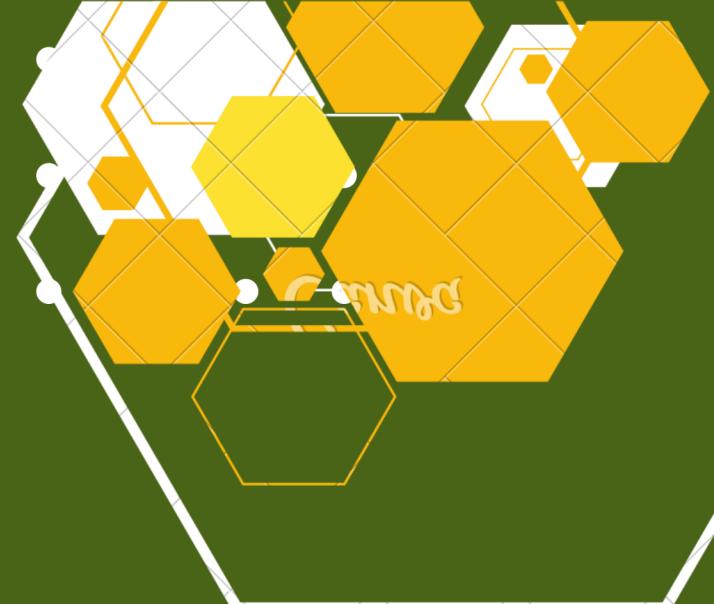
Cluster 4

Ce sont les clients dont le dernier achat est le plus lointain, ils paient les frais de port les moins élevés. Leur panier moyen est assez élevé, ils résident dans de grandes villes et règlent exclusivement par carte de crédit. Ils sont satisfaits de leur expérience d'achat.



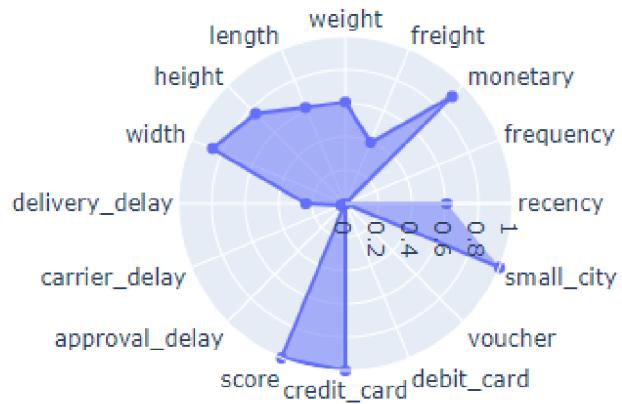
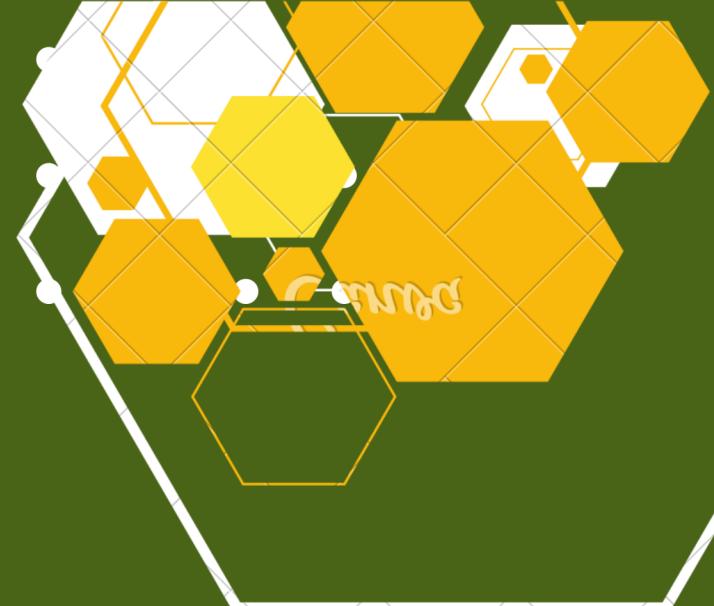
CLUSTERING APRES REDUCTION DES DIMENSIONS VIA L'ACP





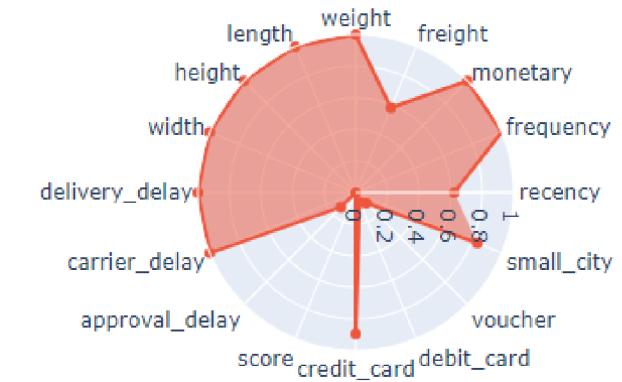
Comparaison des moyennes par variable des clusters

Construction du modèle final, Analyse et interprétation des clusters, Réduction de dimensions via ACP, Interprétation et évaluation du modèle à dimensions réduites, Analyse de la stabilité à l'initialisation.



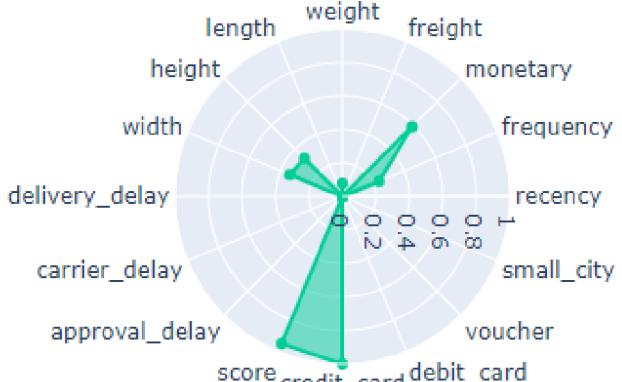
Cluster 0

Ce sont les clients ayant un panier moyen élevé, résidant dans de petites villes. Ils règlent par carte de crédit et sont très satisfaits de leur expérience d'achat.



Cluster 1

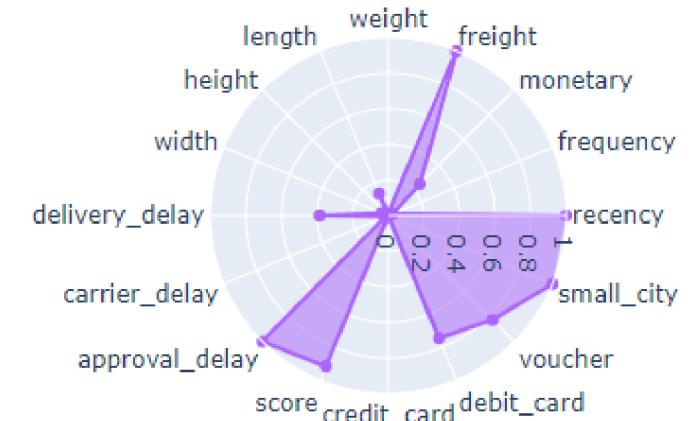
Ce sont les clients ayant le panier moyen le plus élevé, ils résident dans de petites villes. Ils achètent des articles volumineux, ont une très bonne fréquence d'achat et paient des frais de port considérables. Leurs délais de livraison et de transport sont les plus élevés. Ils règlent principalement par carte de crédit et sont très insatisfaits de leur expérience d'achat.



Cluster 2

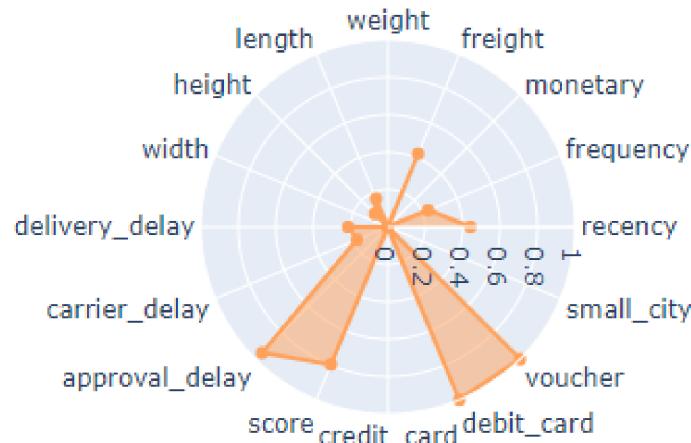
Ce sont les clients dont le dernier achat est le plus lointain, ils paient les frais de port les moins élevés. Leur panier moyen est assez élevé, ils résident dans de grandes villes et règlent exclusivement par carte de crédit. Ils sont très satisfaits de leur expérience d'achat.

*Construction du modèle final,
Analyse et interprétation des clusters,
Réduction de dimensions via ACP,
Interprétation et évaluation du modèle à dimensions réduites,
Analyse de la stabilité à l'initialisation.*



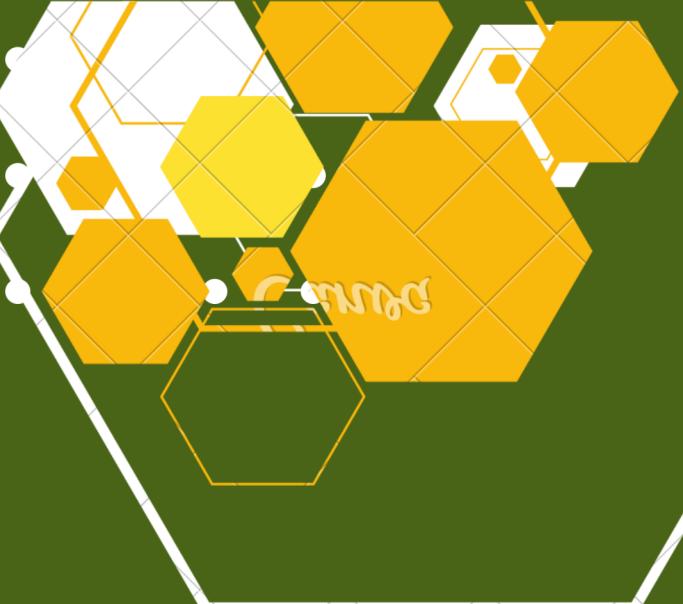
Cluster 3

Ce sont les clients qui paient les frais de port les plus élevés. Leurs achats sont les plus récents et les délais d'approbation sont les plus élevés. Ils résident dans de petites villes, règlent par carte de débit ou par bons. Ils sont satisfaits de leur expérience d'achat.



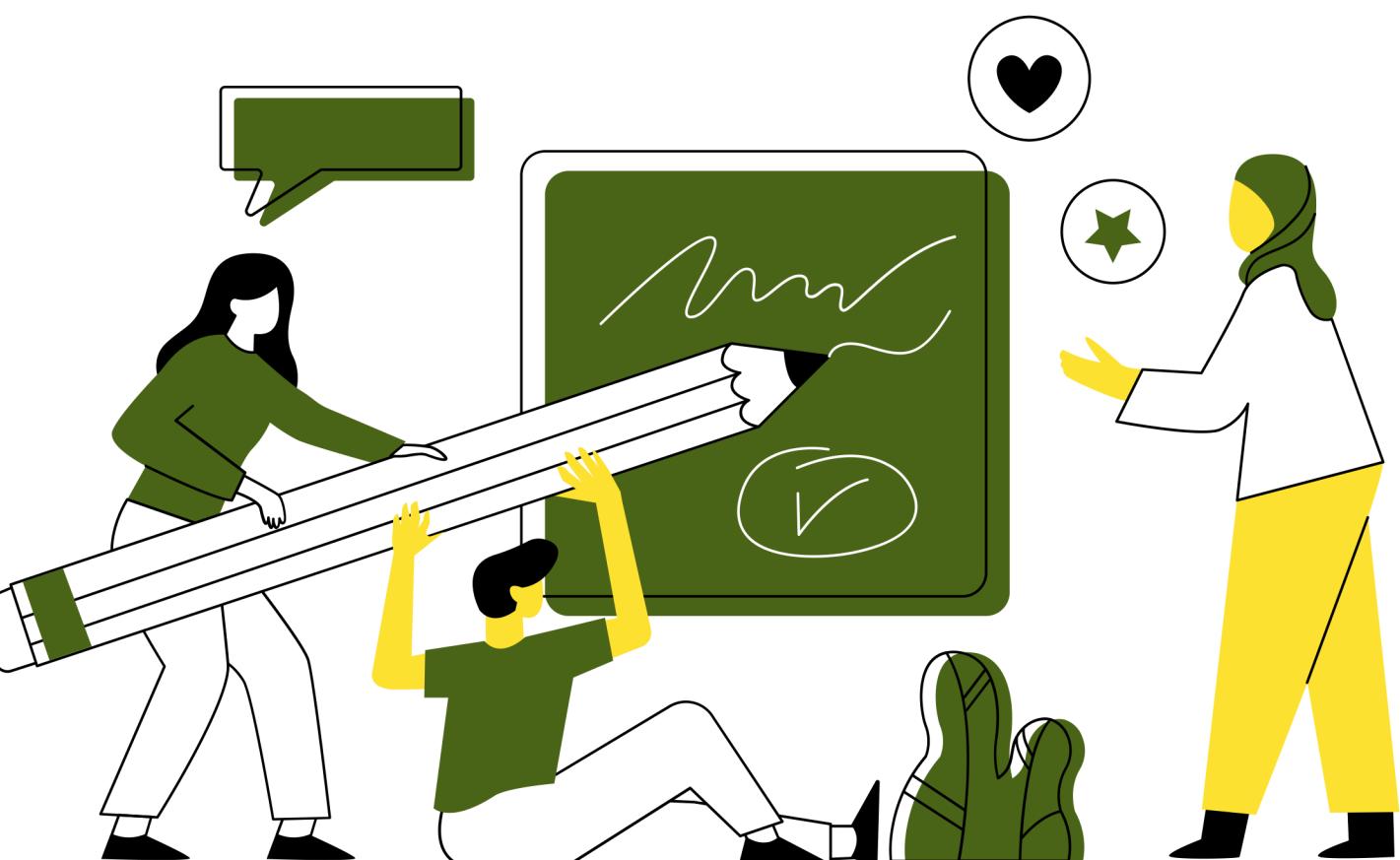
Cluster 4

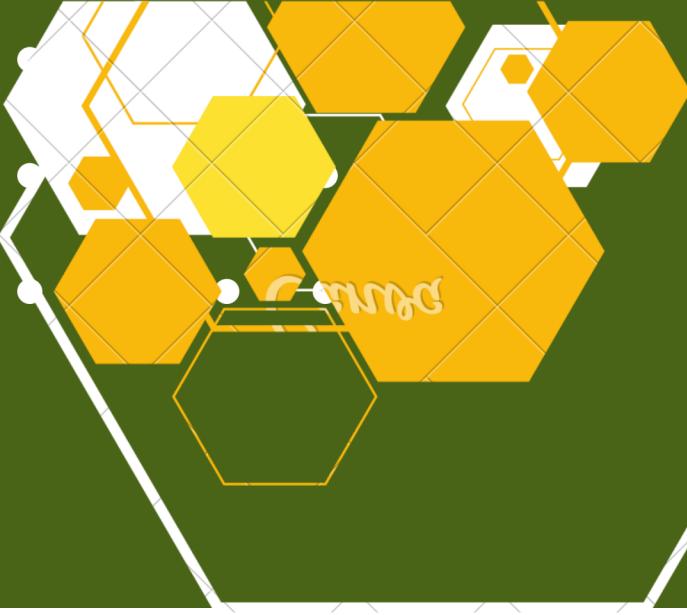
Ce sont les clients qui ont un panier moyen le plus faible, ils résident dans de grandes villes. Leurs achats ne sont pas fréquents. Ils règlent soit par bons ou carte de débit, ont un délai d'approbation élevé et sont satisfaits de leur expérience d'achat.



Scores de stabilité à l'initialisation

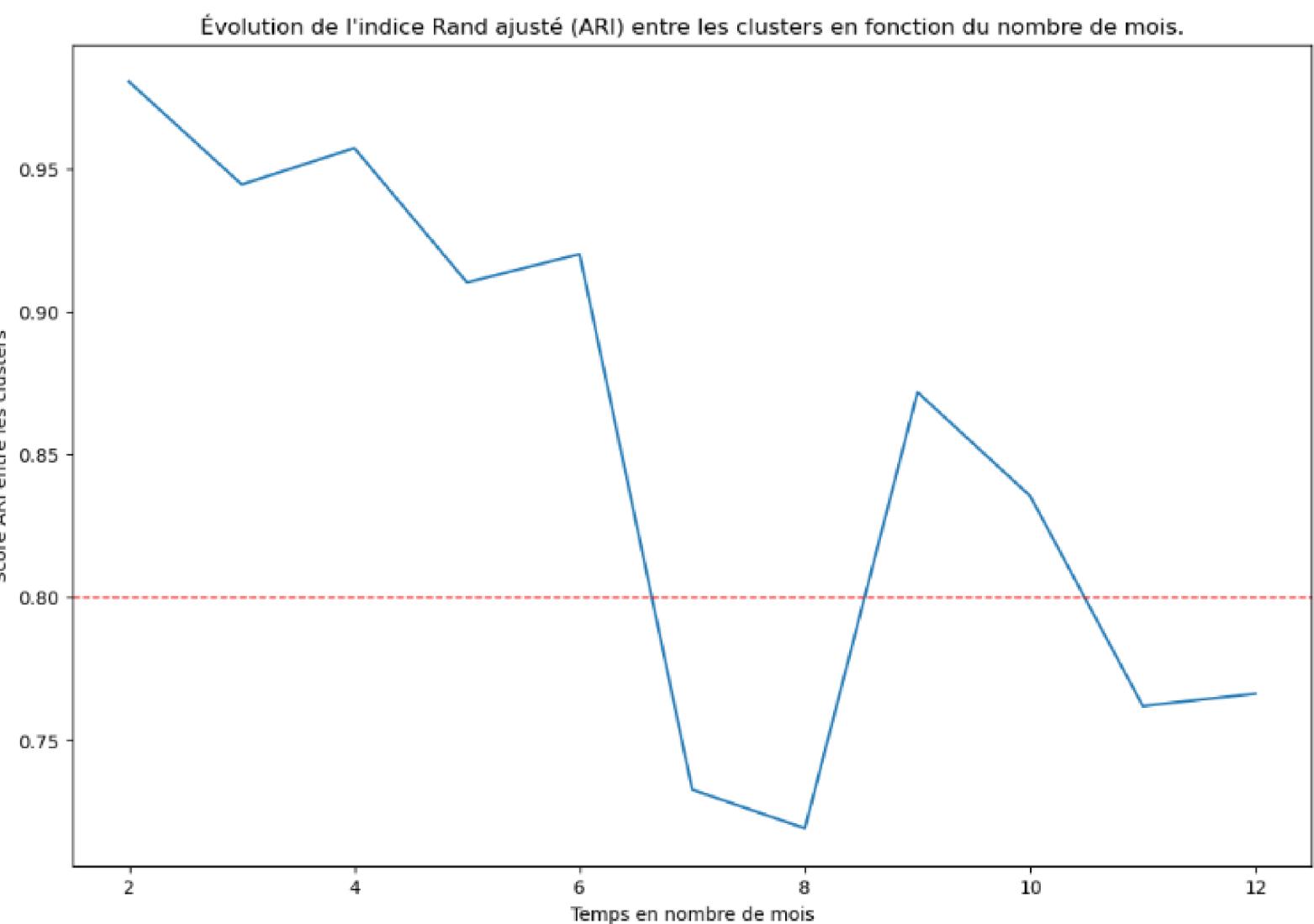
Iteration	FitTime	Inertia	Homo	ARI	AMI
Iter 0	0.249s	8461	0.893	0.878	0.911
Iter 1	0.254s	8859	0.871	0.922	0.874
Iter 2	0.237s	9202	0.829	0.702	0.794
Iter 3	0.230s	8458	1.000	1.000	1.000
Iter 4	0.221s	8869	0.870	0.921	0.873
Iter 5	0.238s	8477	0.863	0.865	0.880
Iter 6	0.256s	9130	0.789	0.614	0.749
Iter 7	0.261s	9098	0.778	0.609	0.738
Iter 8	0.243s	8859	0.871	0.922	0.874
Iter 9	0.238s	9128	0.777	0.608	0.738





==> Approche mise en place

- L'ensemble des données couvrent environ 24 mois,
- Initialisation d'une période de référence T_0 correspondant à (12 mois d'achats) et entraînement d'une modèle KMeans à 5 clusters,
- Définition d'une variable d'incrémentation pour chaque période supplémentaire (un mois supplémentaire d'achats)
- $T_n = T_0 + n$ mois
- Prédictions pour chaque période supplémentaire sur la base du modèle entraîné à T_0 .
- Entraînement du modèle à T_n et comparaison via le score ARI entre T_0 et T_n avec un seuil minimum fixé à 0,8.



==> Conclusion

IL est nécessaire d'effectuer une maintenance du modèle initial après environ 6,5 mois.

22/22

Merci

