



# SEGMENTER DES CLIENTS D'UN SITE E-COMMERCE

PROJET DATA SCIENCE



# SOMMAIRE

1

Introduction

2

Préparation des  
données

3

Segmentation  
Clients

4

Maintenance

5

Conclusion



1

# INTRODUCTION



## SEGMENTER DES CLIENTS D'UN SITE E-COMMERCE

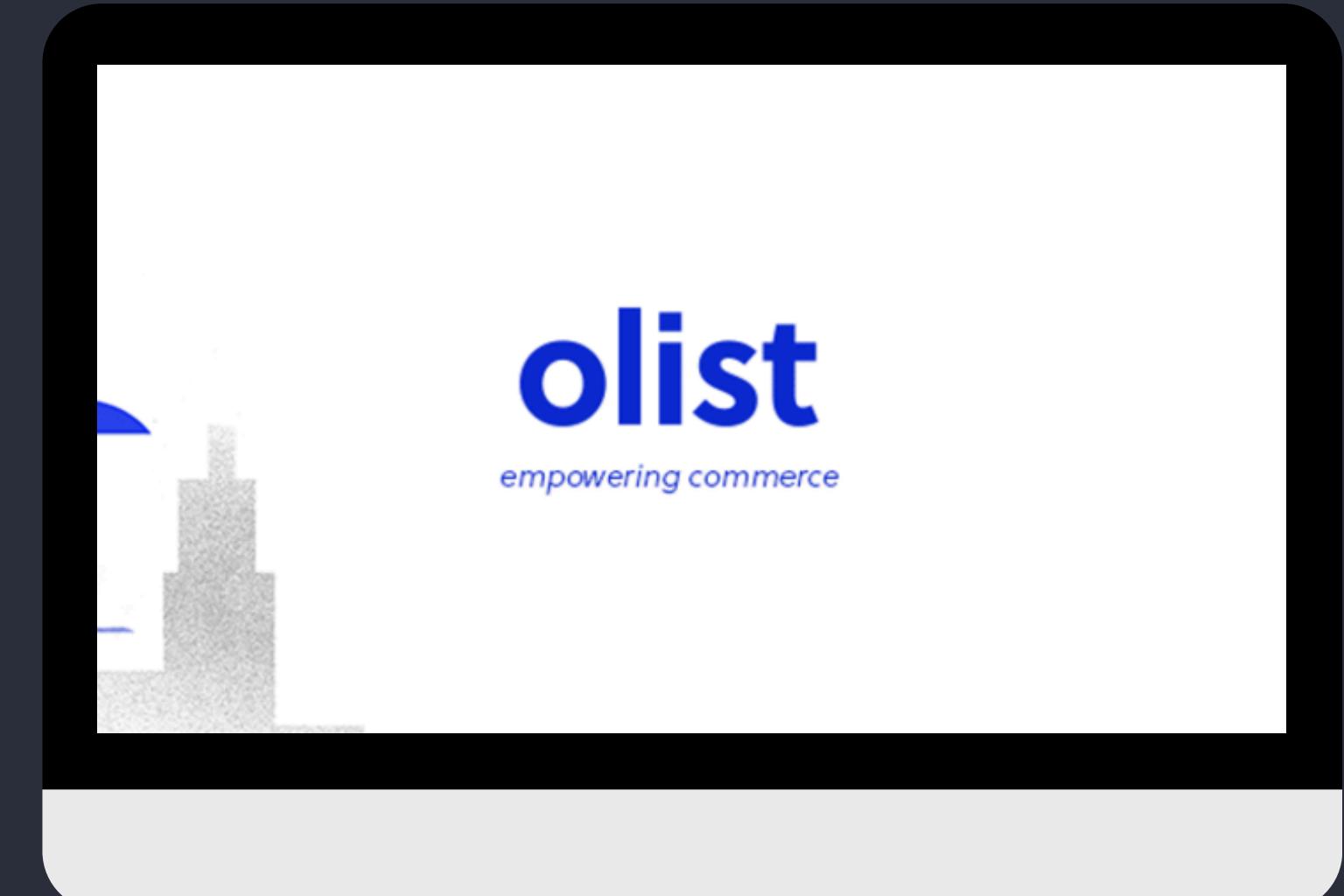
# INTRODUCTION

Olist est une entreprise brésilienne offrant une solution pour vendre sur les marketplaces en ligne.

Afin de mettre au point des campagnes de communication pertinentes, Olist souhaite comprendre les différents types d'utilisateurs de son site.

Dans ce contexte, le projet suivant s'intéresse au comportement et aux données personnelles de ces utilisateurs.

L'**objectif** est de fournir une segmentation des clients d'Olist destinée à l'équipe Marketing afin d'améliorer la communication de l'entreprise.





2

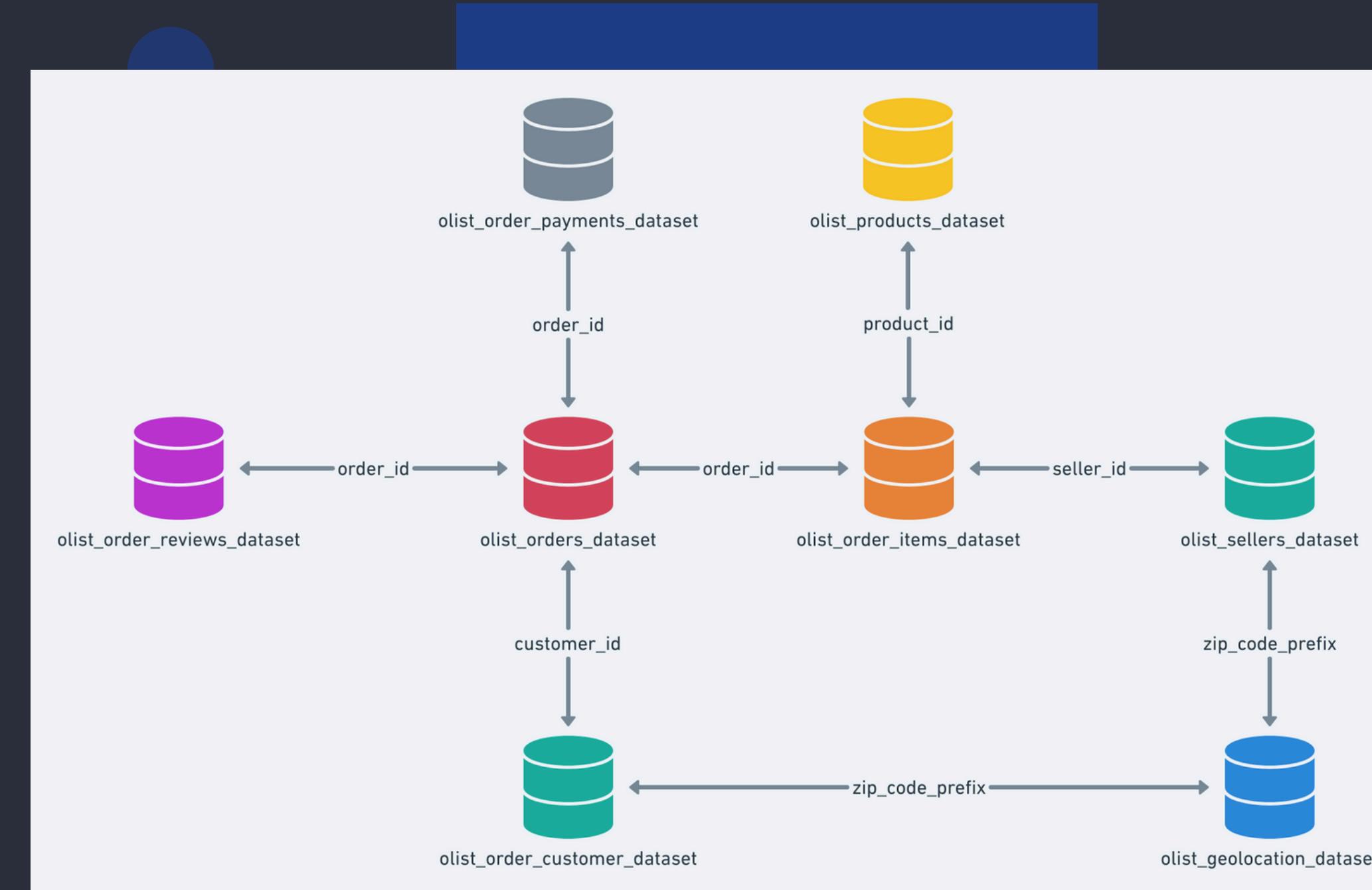
# PRÉPARATION DES DONNÉES



## PRÉSENTATION DES DONNÉES

# SCHÉMA DE LA BASE DE DONNÉES

La base de données représentée ci-contre possède 9 tables parmi lesquelles on retrouve des informations concernant : les utilisateurs, les vendeurs, les commandes, les produits, les paiements, les avis laissés et les positions géographiques.



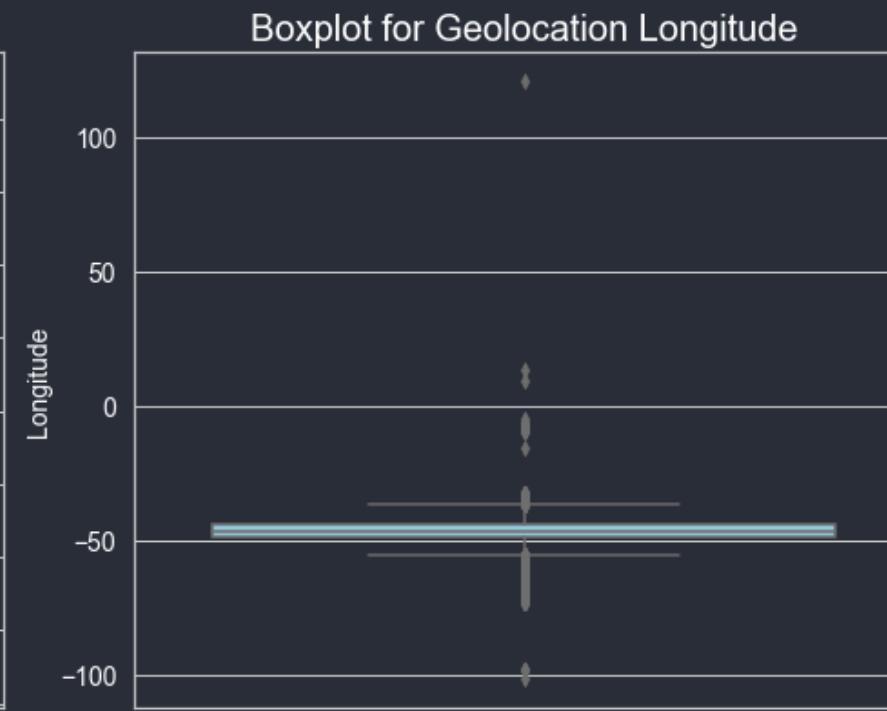
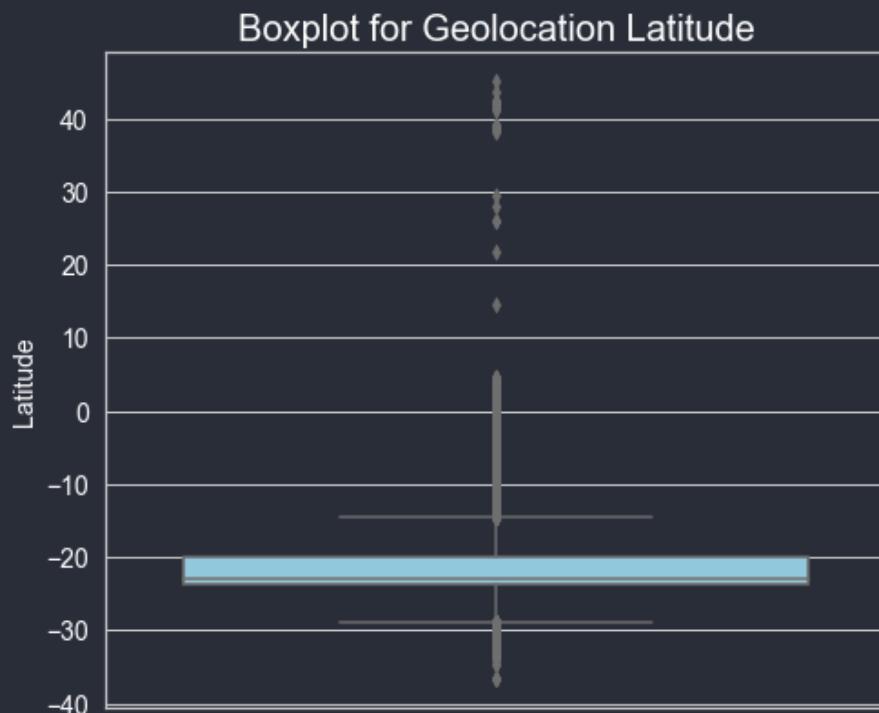


## PRÉSENTATION DES DONNÉES

# LES LOCALISATIONS

Le jeu de données concernant les localisations fourni des renseignements géographiques en faisant les liens entre code postale, nom de la ville, coordonnées géographique (latitude, longitude) et État du Brésil associé.

On observe des outliers pour les variables relatives aux coordonnées géographique.



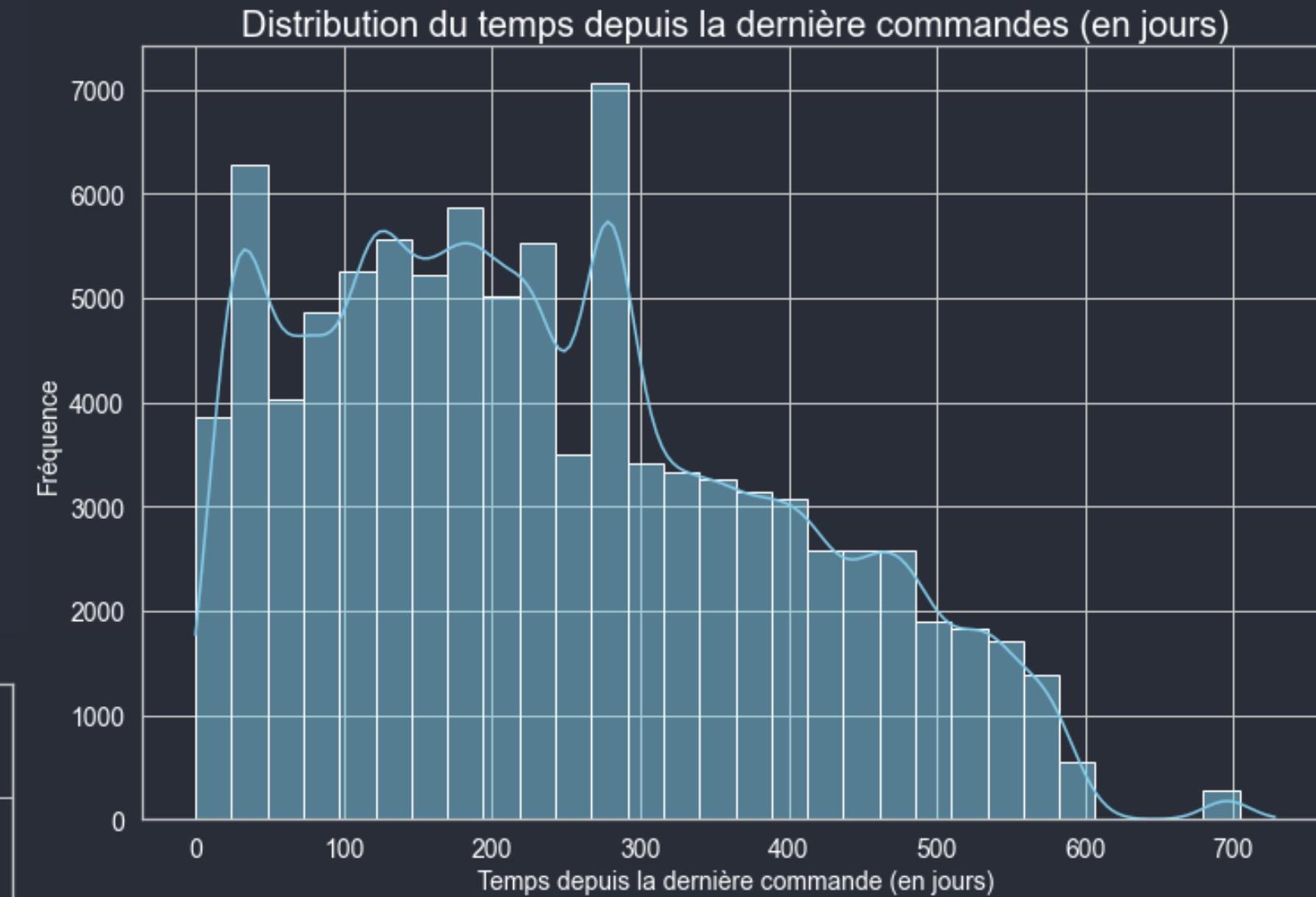
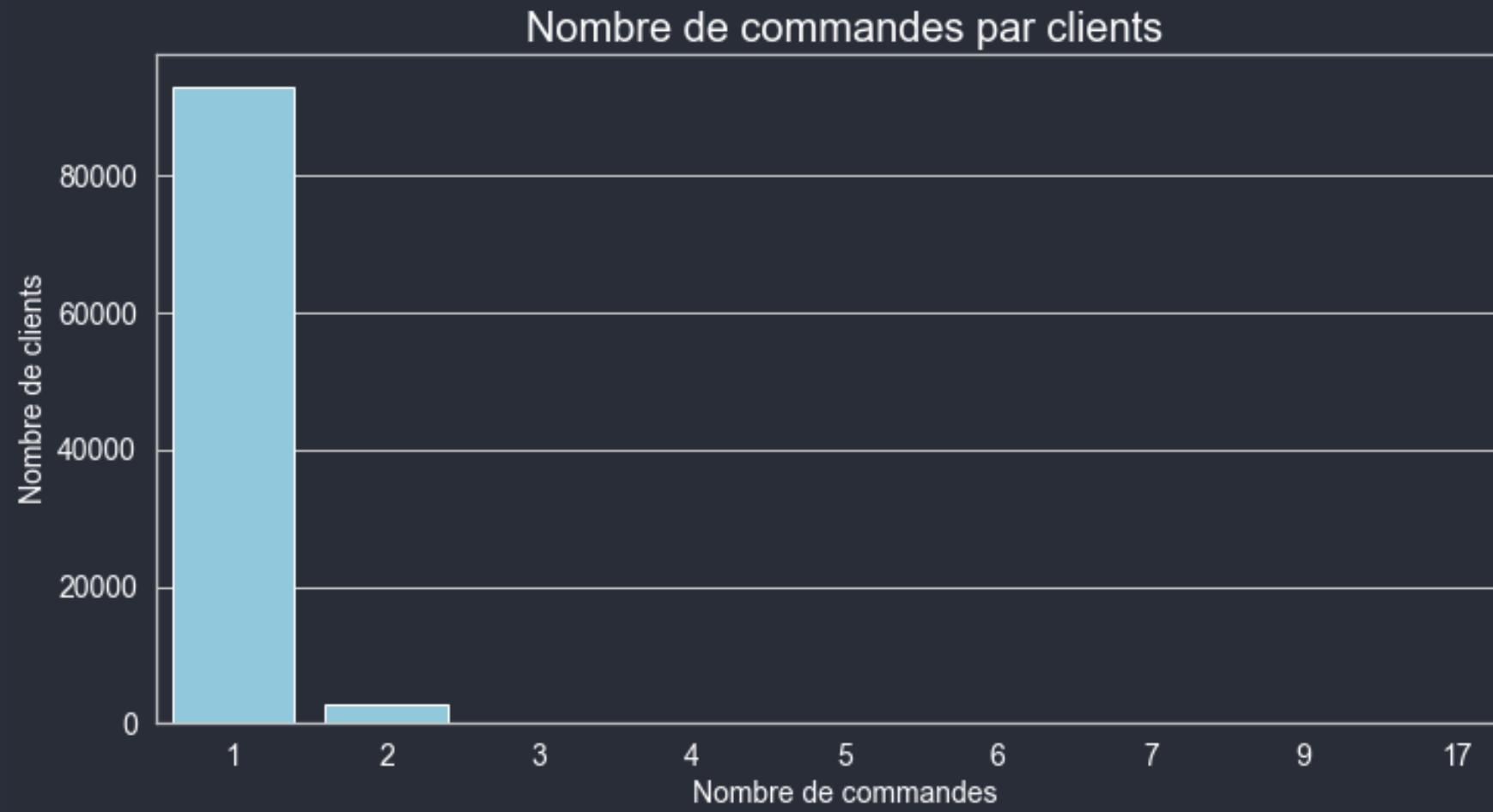
Parmi les 1000163 lignes du jeu de données il en a 31 qui ne sont pas correctement localisées.  
Les coordonnées ont été corrigées à l'aide des variables relatives aux villes et aux États.



## PRÉSENTATION DES DONNÉES

# LES UTILISATEURS

On constate que la plupart des clients n'ont passé qu'une seule commande sur le site.



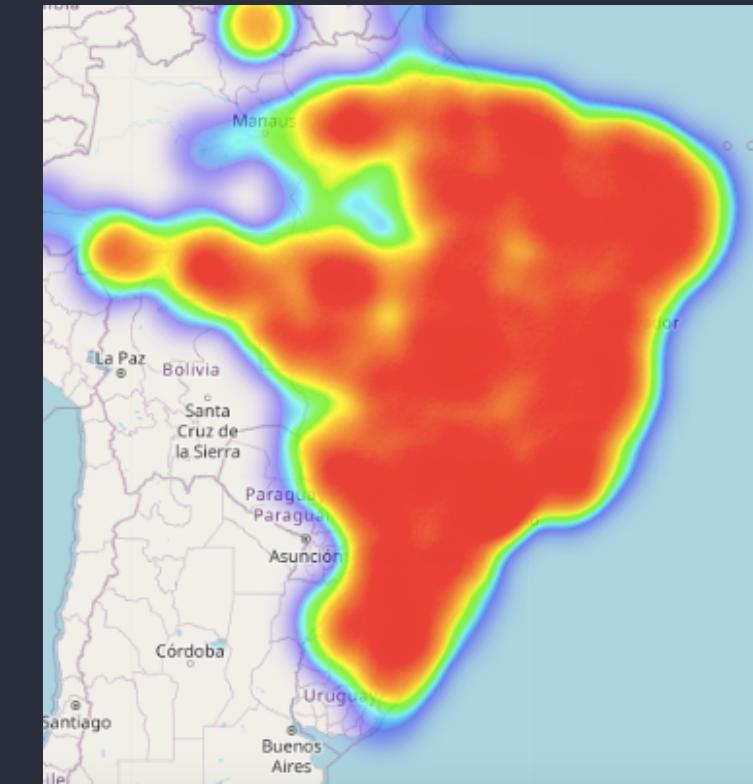
On constate aussi qu'une proportion significative de clients a commandé récemment, ce qui montre une activité régulière sur le site. Cependant, une partie des clients semble inactive depuis plus de 400 jours, avec une diminution progressive des fréquences après environ 350 jours.



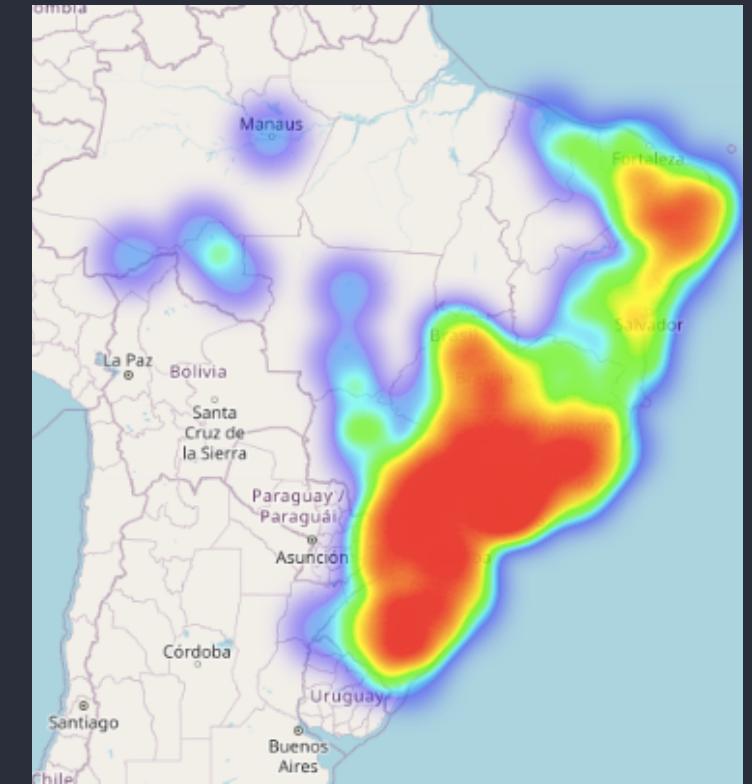
## PRÉSENTATION DES DONNÉES

# LES COMMANDES

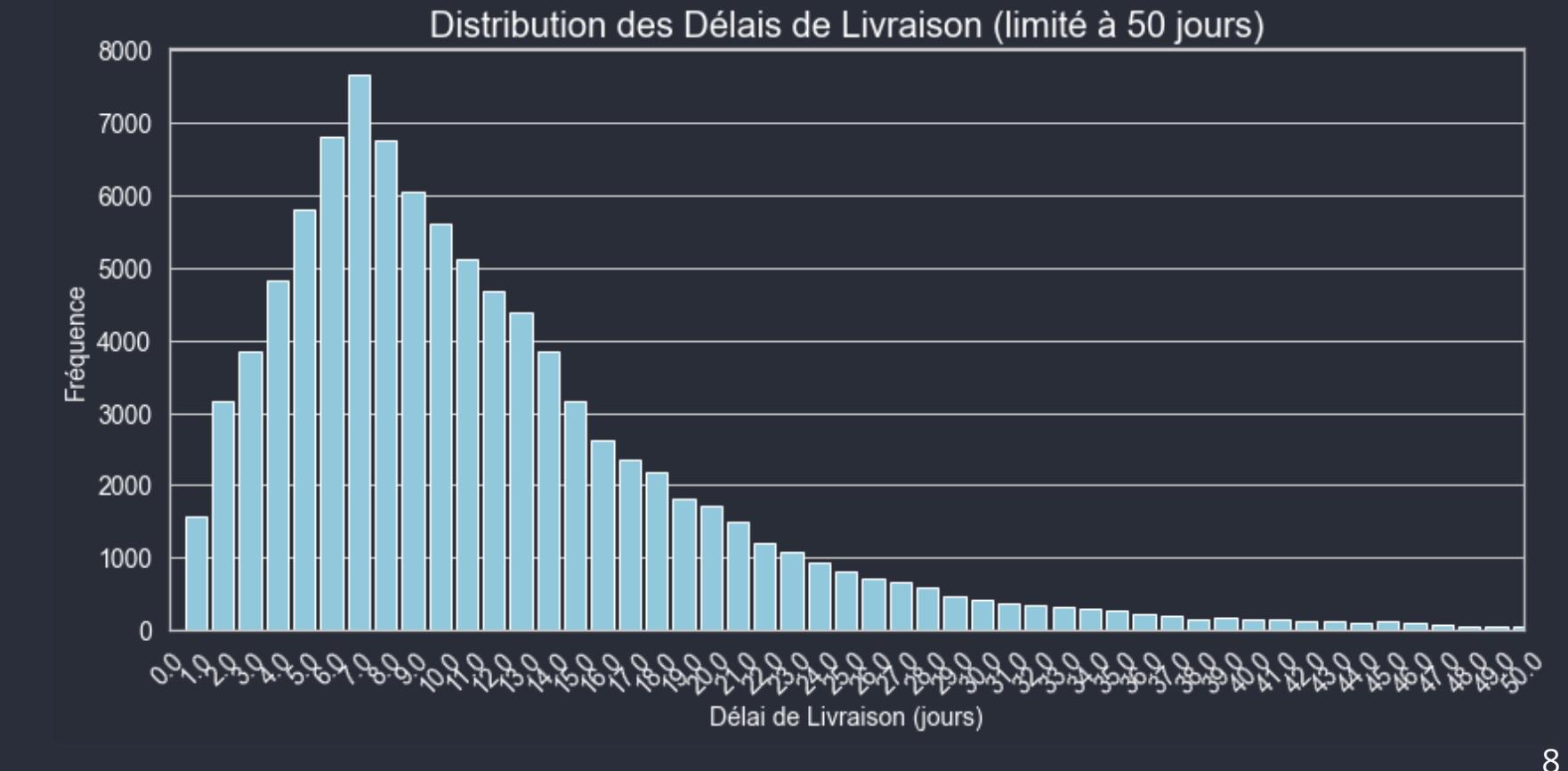
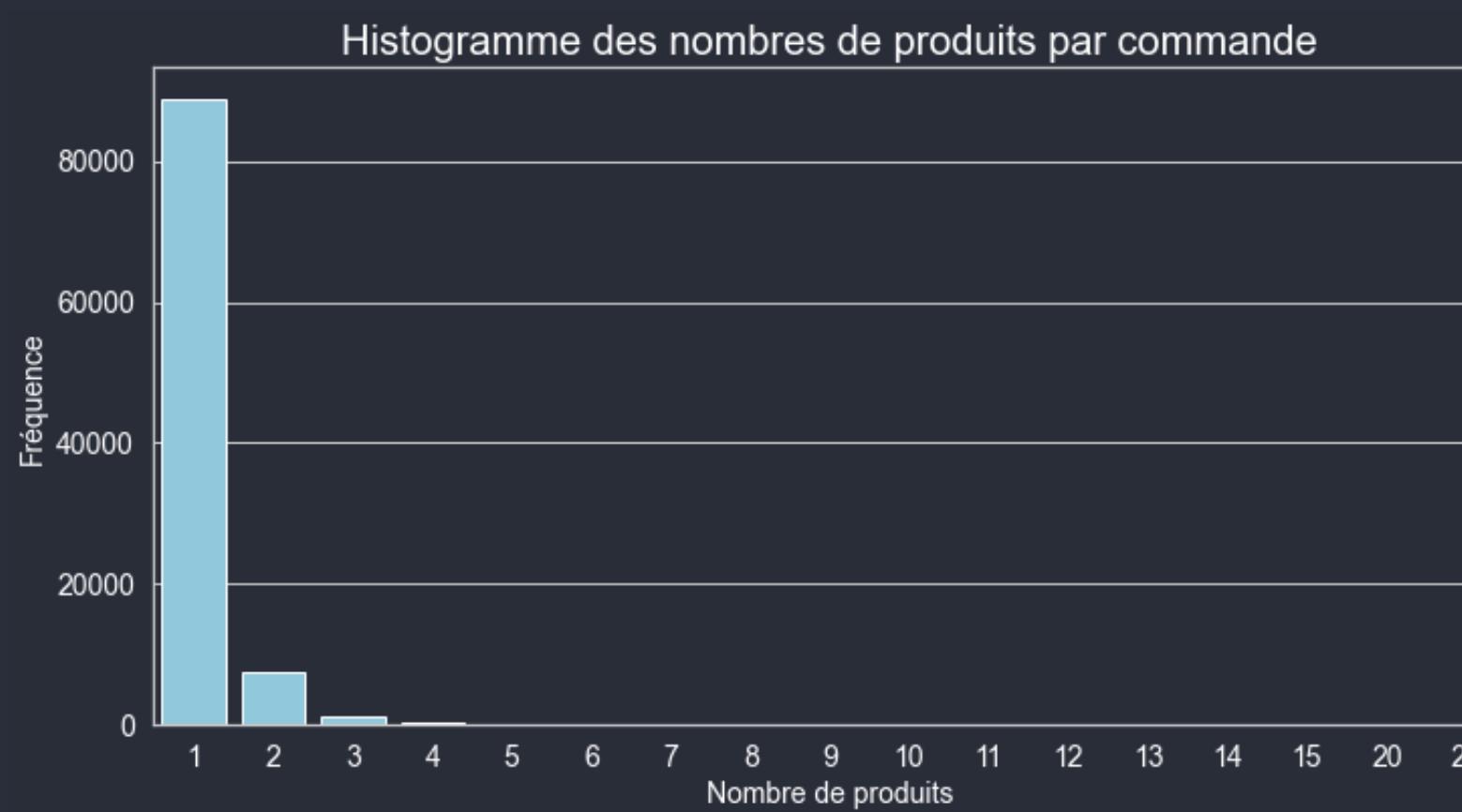
Les commandes passées contiennent en grande majorité un seul produit. Il semble aussi que majorité des clients reçoivent leurs commandes dans un délai raisonnable



Localisation des utilisateurs



Localisation des vendeurs

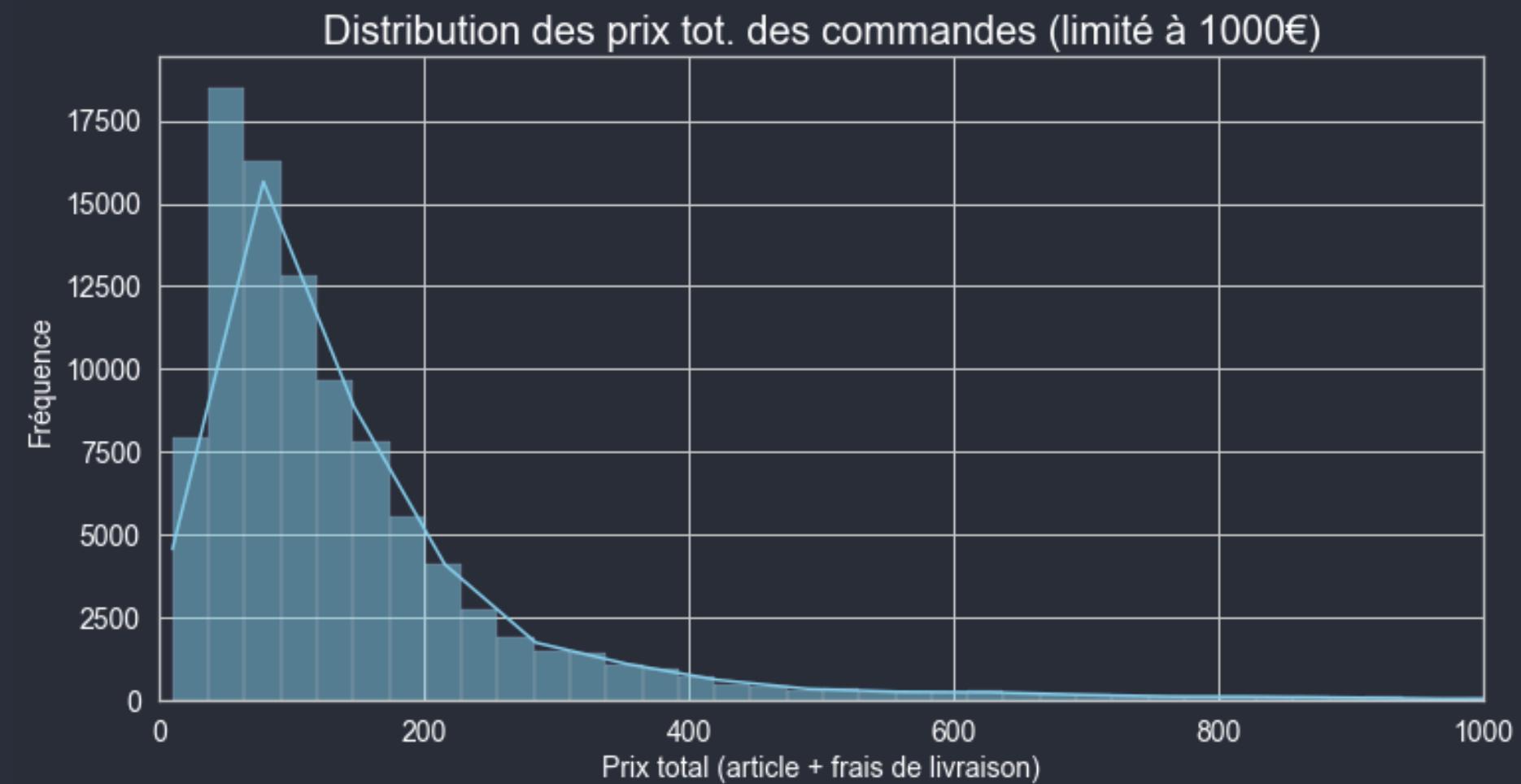
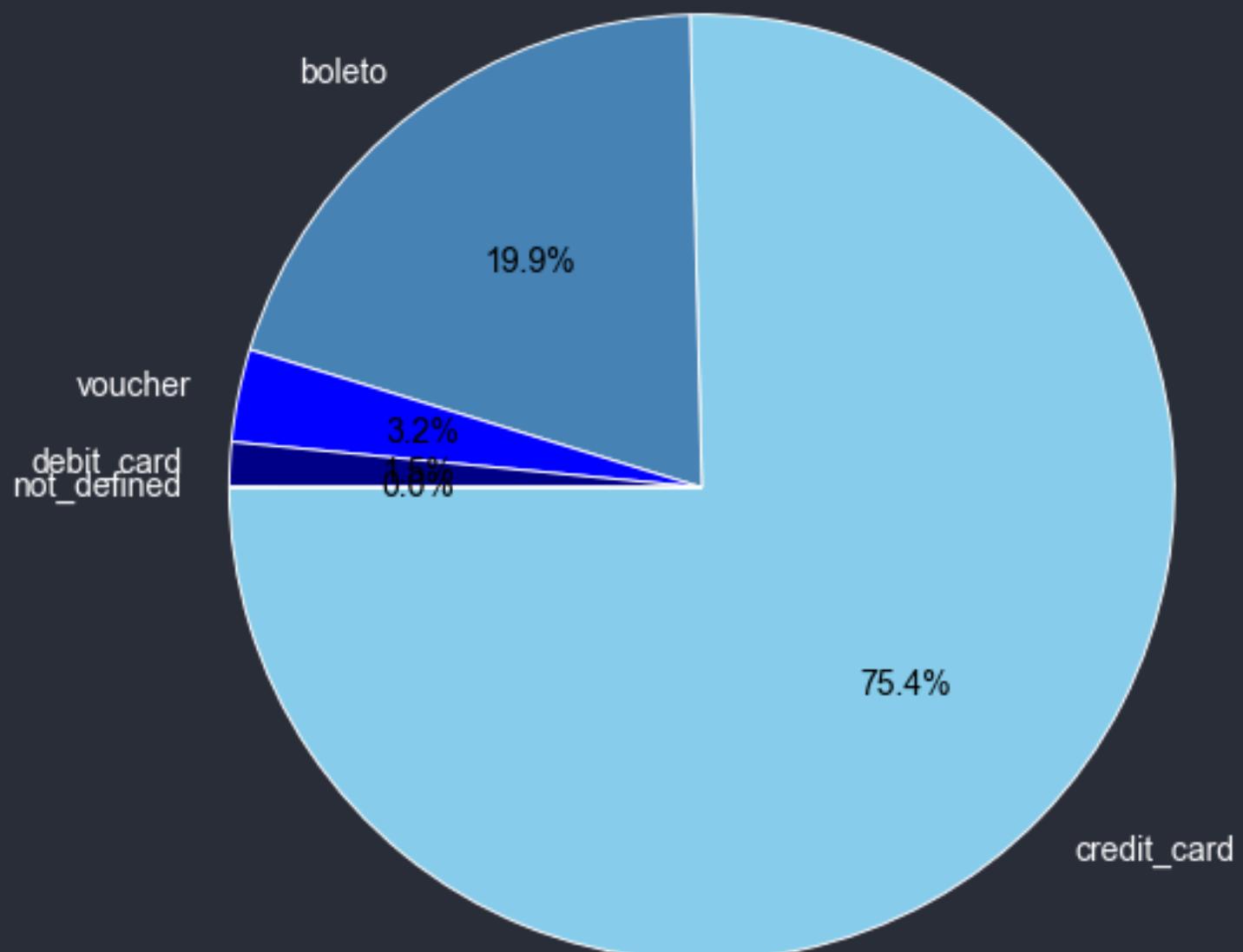




## PRÉSENTATION DES DONNÉES

# LES PAIEMENTS

Répartition des types de paiement principaux



On constate une forte concentration de commandes à faible montant.

En effet, la majorité des commandes se situent entre 0 et 200 €, avec un pic très marqué autour de cette plage.

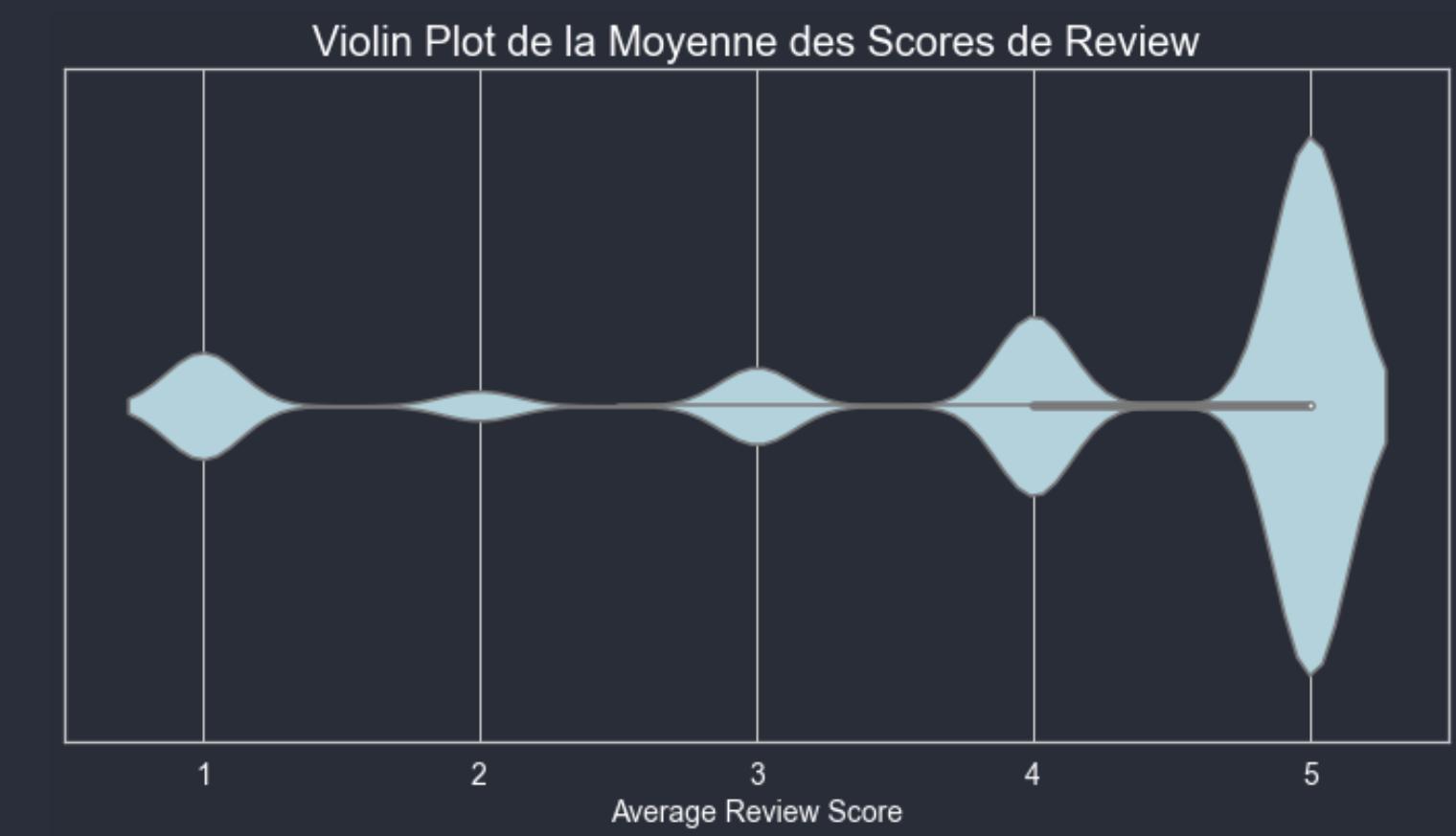
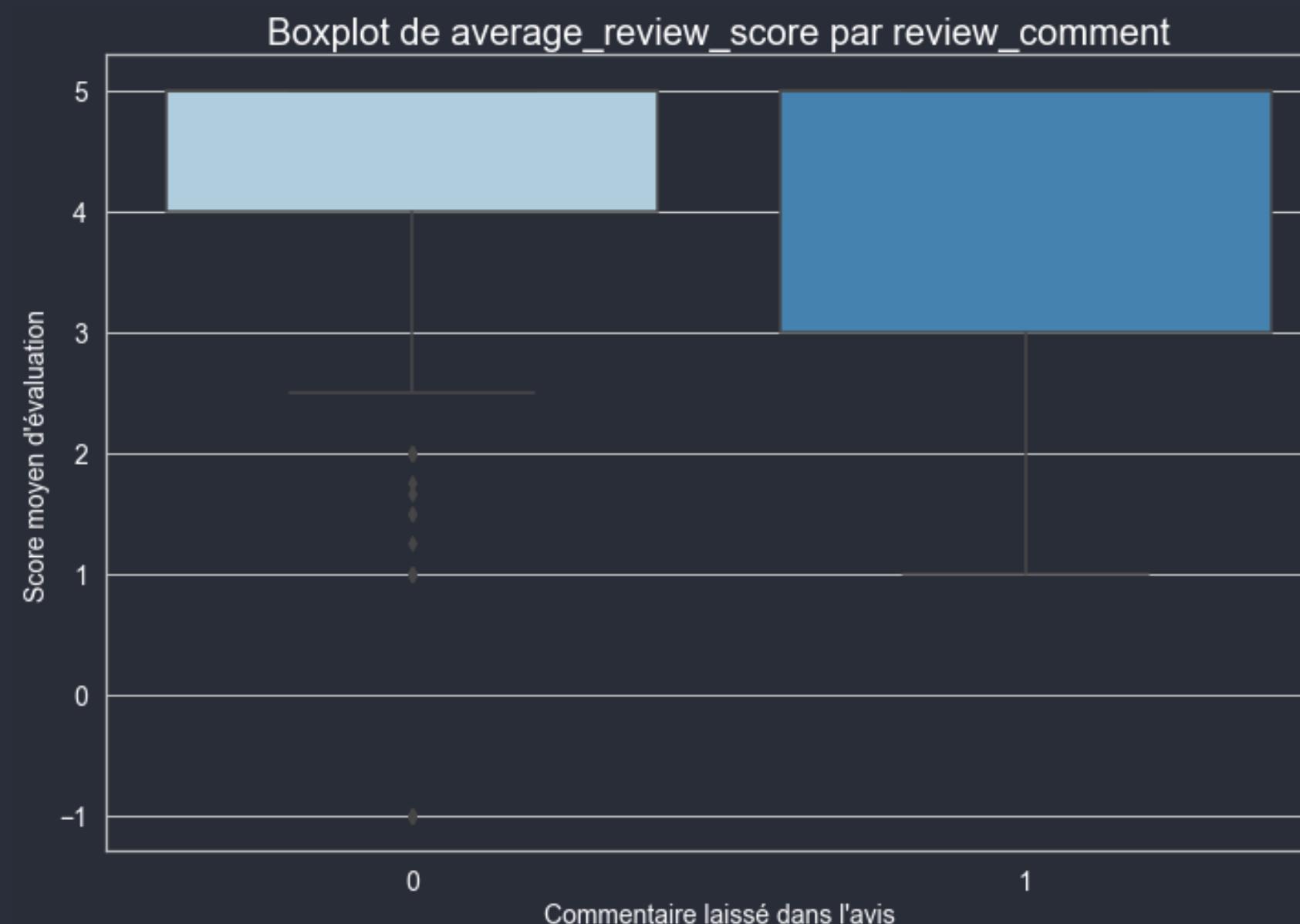
Une queue allongée à droite montre l'existence de commandes plus coûteuses.

On retrouve plusieurs moyens de paiement avec en grande majorité les cartes bancaires puis, avec 20%, boleto qui est un moyen de paiement via des coupons au Brésil.



PRÉSENTATION DES DONNÉES

# LES AVIS



On constate davantage de notes positives que négatives.

Les clients qui donnent une note mais ne laissent pas de commentaire semblent généralement plus satisfaits (note moyenne plus élevée).

Les clients qui laissent un commentaire donnent en moyenne une note légèrement inférieure, suggérant qu'ils sont plus enclins à exprimer une opinion détaillée, qu'elle soit positive ou négative.



Data Science

PRÉPARATION DES DONNÉES

# FEATURES ENGINEERING

## Fusion des jeux de données

Dataset avec une ligne par client.

## Nettoyage du jeu de données

Suppression des commandes annulées, indisponibles, etc  
Gestion des valeurs manquantes.  
Suppression de certaines variables.

## Création de variables

Prix total dépensé, mode de paiement préféré du client, nombre de commande passée, délais de livraison, temps depuis la dernière commande, etc...

**93347**  
lignes

**18**  
colonnes

**0%**  
valeurs  
manquantes

Matrice de corrélation des clients																																	
customer_zip_code_prefix	number_of_orders	total_price	time_since_last_order	number_of_products	average_number_products_per_order	average_price_per_order	average_payment_installments	average_delivery_time	average_products_weight_g	average_product_volume	number_of_review	average_review_score	review_comment	number_of_sellers	delayed	customer_zip_code_prefix	number_of_orders	total_price	time_since_last_order	number_of_products	average_number_products_per_order	average_price_per_order	average_payment_installments	average_delivery_time	average_products_weight_g	average_product_volume	number_of_review	average_review_score	review_comment	number_of_sellers	delayed		
1	-0.0061	0.058	0.048	-0.011	0.0081	0.059	0.056	0.27	0.002	-0.0012	0.00260	0.0190	0.0027	-0.006	0.028	0.0061	1	0.13	-0.021	0.7	0.027	-0.011	0.0260	0.000550	0.00630	0.00610	0.89	0.00550	0.041	0.86	-0.018		
number_of_orders	1	0.13	-0.021	0.7	0.027	-0.011	0.0260	0.000550	0.00630	0.00610	0.89	0.00550	0.041	0.86	-0.018	0.058	1	0.13	-0.0048	0.15	0.082	0.98	0.33	0.07	0.37	0.33	0.1	-0.042	0.044	0.14	0.018		
total_price	0.058	1	-0.0048	0.15	0.082	0.98	0.33	0.07	0.37	0.33	0.1	-0.042	0.044	0.14	0.018	0.048	-0.0210	0.0048	1	-0.0150	0.00240	0.00130	0.05	0.097	0.037	0.061	-0.012	-0.012	0.023	-0.03	-0.03		
time_since_last_order	0.048	-0.0210	0.0048	1	-0.0150	0.00240	0.00130	0.05	0.097	0.037	0.061	-0.012	-0.012	0.023	-0.03	0.011	0.7	0.15	-0.015	1	0.72	0.048	0.069	-0.021	-0.015	-0.014	0.62	-0.069	0.065	0.82	-0.027		
number_of_products	-0.011	0.7	0.15	-0.015	1	0.72	0.048	0.069	-0.021	-0.015	-0.014	0.62	-0.069	0.065	0.82	-0.027	0.00810	0.027	0.082	-0.0024	0.72	1	0.079	0.074	-0.029	-0.016	-0.013	0.022	-0.11	0.055	0.32	-0.022	
average_number_products_per_order	0.00810	0.027	0.082	-0.0024	0.72	1	0.079	0.074	-0.029	-0.016	-0.013	0.022	-0.11	0.055	0.32	-0.022	0.059	-0.011	0.98	-0.00130	0.048	0.079	1	0.33	0.071	0.38	0.34	-0.016	-0.044	0.039	0.018	0.021	
average_price_per_order	0.059	-0.011	0.98	-0.00130	0.048	0.079	1	0.33	0.071	0.38	0.34	-0.016	-0.044	0.039	0.018	0.021	0.056	0.026	0.33	0.05	0.069	0.074	0.33	1	0.053	0.2	0.18	0.023	-0.034	0.047	0.054	0.013	
average_payment_installments	0.056	0.026	0.33	0.05	0.069	0.074	0.33	1	0.053	0.1	0.082	0.073	-0.014	-0.33	0.081	-0.021	0.6	0.27	0.000550	0.07	0.097	-0.021	0.029	0.071	0.053	1	0.082	0.073	-0.014	-0.33	0.081	-0.021	0.6
average_delivery_time	0.27	0.000550	0.07	0.097	-0.021	0.029	0.071	0.053	1	0.082	0.073	-0.014	-0.33	0.081	-0.021	0.6	0.002	-0.00630	0.37	0.037	-0.015	0.016	0.38	0.2	0.082	0.073	-0.014	-0.33	0.081	-0.021	0.6		
average_products_weight_g	0.002	-0.00630	0.37	0.037	-0.015	0.016	0.38	0.2	0.082	1	0.81	-0.01	-0.01	1	0.099	0.054	0.76	-0.033	0.1	-0.21	-0.051	-0.39	0.03	-0.00850	0.026	0.0120	0.00610	0.33	0.061	-0.028			
average_product_volume	0.00120	0.00610	0.33	0.061	-0.014	0.013	0.34	0.18	0.073	0.81	1	-0.01	-0.01	1	0.099	0.054	0.76	-0.033	0.1	-0.21	-0.051	-0.39	0.03	-0.00850	0.026	0.0120	0.00610	0.33	0.061	-0.028			
number_of_review	0.0026	0.89	0.1	-0.012	0.62	0.022	-0.016	0.023	-0.014	-0.01	-0.01	1	0.099	0.054	0.76	-0.033	0.0190	0.00550	-0.042	-0.012	-0.069	-0.11	-0.044	-0.034	-0.33	-0.031	-0.028	0.099	1	-0.21	-0.051	-0.39	
average_review_score	-0.0190	0.00550	-0.042	-0.012	-0.069	-0.11	-0.044	-0.034	-0.33	-0.031	-0.028	0.099	1	-0.21	-0.051	-0.39	0.00270	0.041	0.044	0.023	0.065	0.055	0.039	0.047	0.081	0.03	0.024	0.054	-0.21	1	0.063	0.094	-0.028
review_comment	0.00270	0.041	0.044	0.023	0.065	0.055	0.039	0.047	0.081	0.03	0.024	0.054	-0.21	1	0.063	0.094	-0.028	0.00660	0.76	-0.051	0.063	1	-0.28										
number_of_sellers	-0.006	0.86	0.14	-0.03	0.82	0.32	0.018	0.054	-0.021	0.00850	0.00660	0.76	-0.051	0.063	1	-0.28	0.028	-0.018	0.018	-0.03	-0.027	-0.022	0.021	0.013	0.6	0.026	0.022	-0.033	-0.39	0.094	-0.028	1	
delayed	0.028	-0.018	0.018	-0.03	-0.027	-0.022	0.021	0.013	0.6	0.026	0.022	-0.033	-0.39	0.094	-0.028	1	0.028	0.018	0.018	-0.03	-0.027	-0.022	0.021	0.013	0.6	0.026	0.022	-0.033	-0.39	0.094	-0.028	1	



3

# MODÉLISATION



## SEGMENTATION RFM

# MÉTHODE RFM

La méthode RFM est une méthode de segmentation qui prend en compte :

- la Récence,
- la Fréquence,
- le montant

pour établir des segments de clients homogènes.



Étapes :





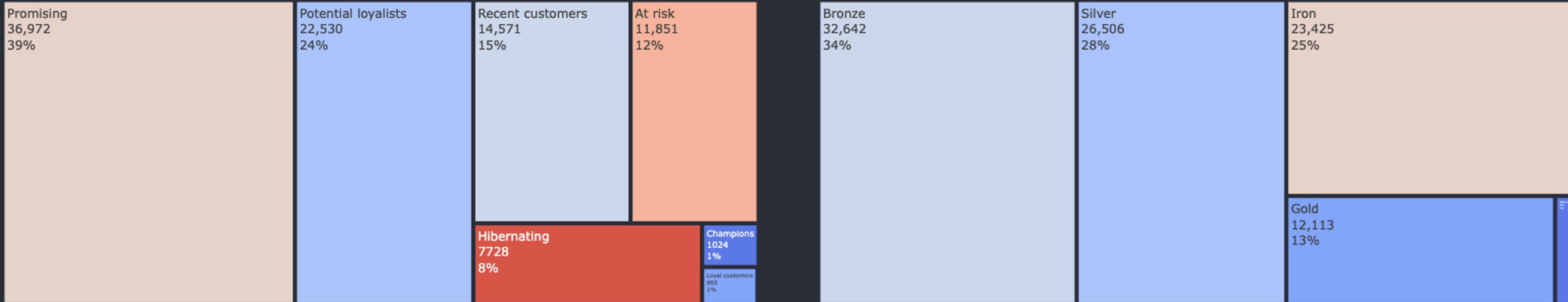
## SEGMENTATION RFM

Segment client	Interprétation
Champions	Les meilleurs clients, ils achètent et dépensent beaucoup et ont effectué leur dernier achat récemment.
Loyal customers	Clients récents et fréquents.
Potential loyalists	Des clients récents, mais qui ont déjà beaucoup dépensé.
Recent customers	Clients récents, qui n'ont effectué que quelques achats.
Promising	Clients qui achètent fréquemment et dépensent beaucoup, mais qui ont effectué leur dernier achat il y a quelque temps.
Need attention	Clients récents avec des dépenses supérieures à la moyenne, mais une faible fréquence.
At risk	Clients qui achètent fréquemment, mais qui n'ont effectué aucun achat depuis longtemps.
Can't loose them	Des clients qui ont beaucoup acheté, mais qui sont inactifs depuis longtemps.
Hibernating	Clients peu fréquents et peu dépensiers qui n'ont pas acheté depuis longtemps.



## SEGMENTATION RFM

## RFM Segments



## Treemap des Score RFM

Les segments ont été obtenu par concaténation des scores RFM individuels.

Les scores ont été obtenu par somme des scores RFM individuels.

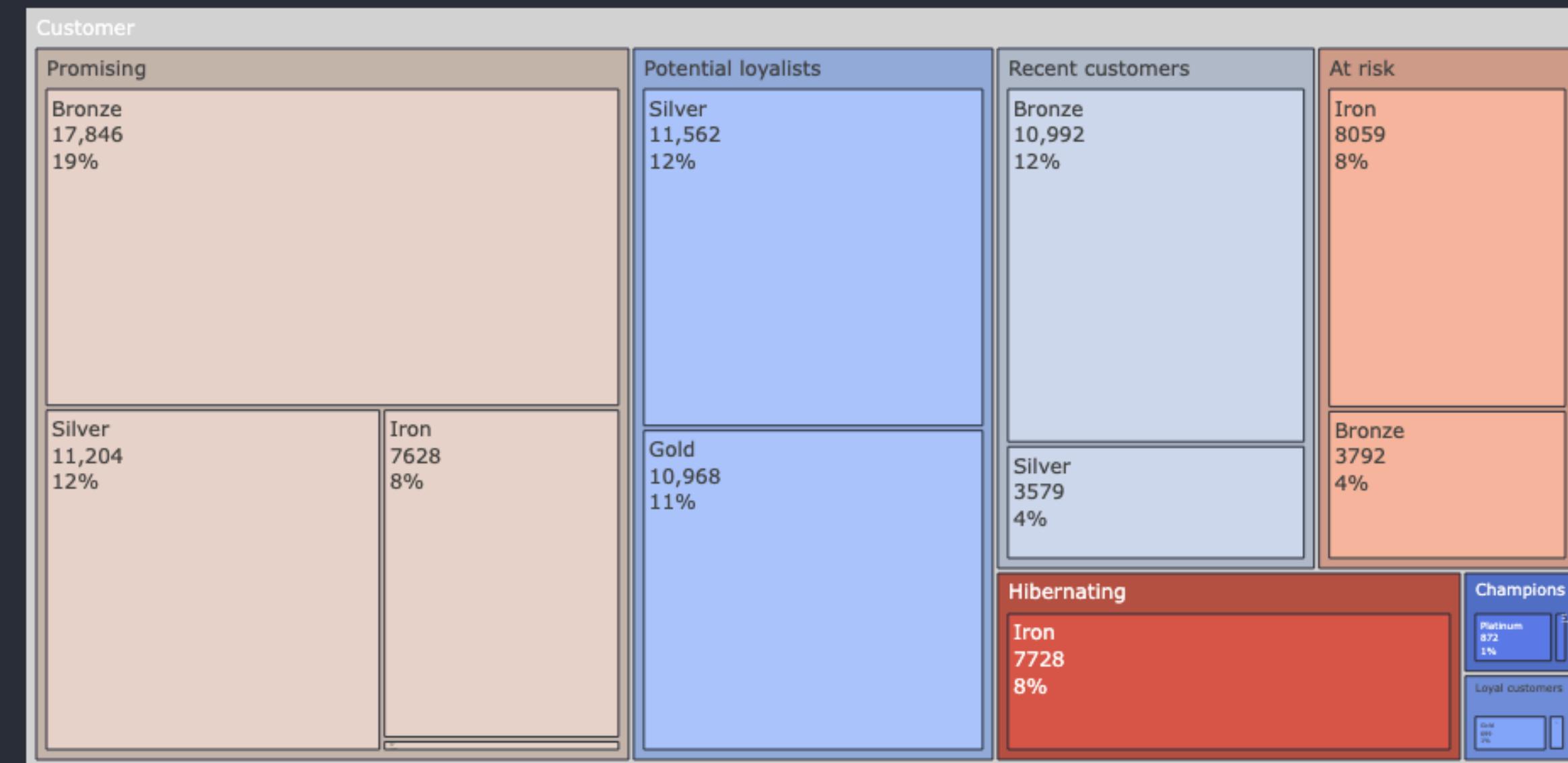
	customer_unique_id	Recency	Frequency	MonetaryValue	R	F	M	RFM_Segment	RFM_Score	Segment	Score
0	0000366f3b9a7992bf8c76cfdf3221e2	-0.753603	-0.172331	0.292922	4	3	4	434	11	Potential loyalists	Silver
1	0000b849f77a49e4a4ce2b2a4ca5be3f	-0.727855	-0.172331	-1.540226	4	3	1	431	8	Recent customers	Bronze
2	0000f46a3911fa3c0805444483337064	1.648389	-0.172331	-0.264655	1	3	2	132	6	At risk	Iron
3	0000f6ccb0745a6a4b88665a16c9f078	0.629383	-0.172331	-1.021608	2	3	1	231	6	Hibernating	Iron
4	0004aac84e0df4da2b147fca70cf8255	0.448534	-0.172331	0.660602	2	3	4	234	9	Promising	Bronze



SEGMENTATION RFM

# MÉTHODE RFM

Treemap des Segmentations clients

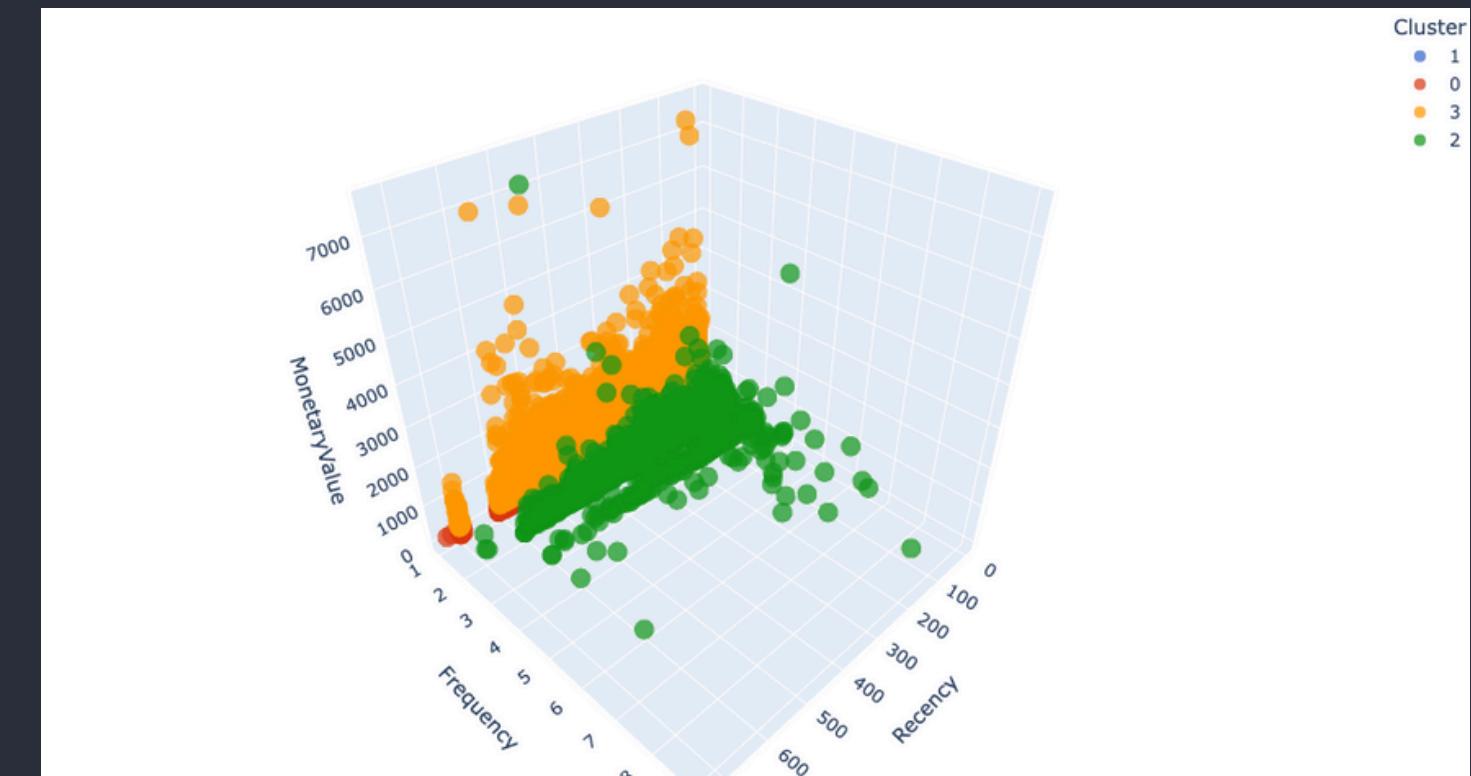




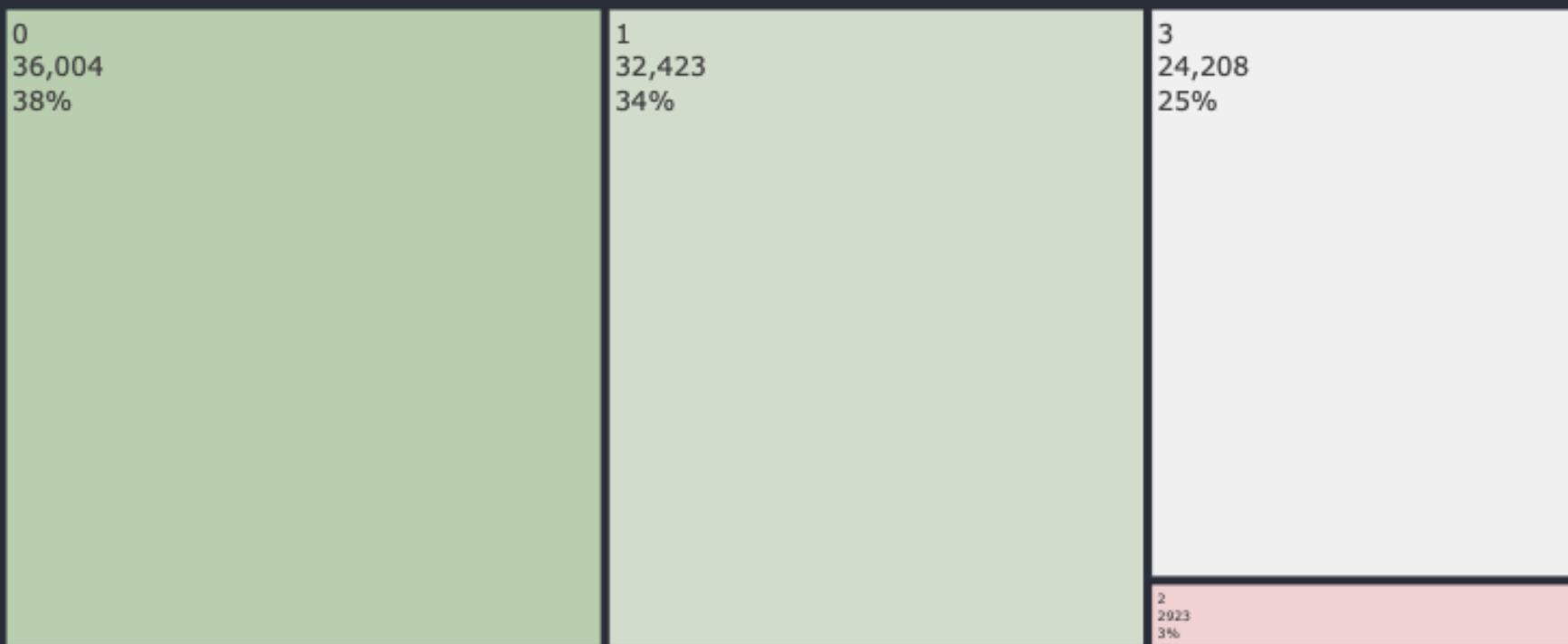
Data Science

K-MEANS SUR DONNÉES RFM

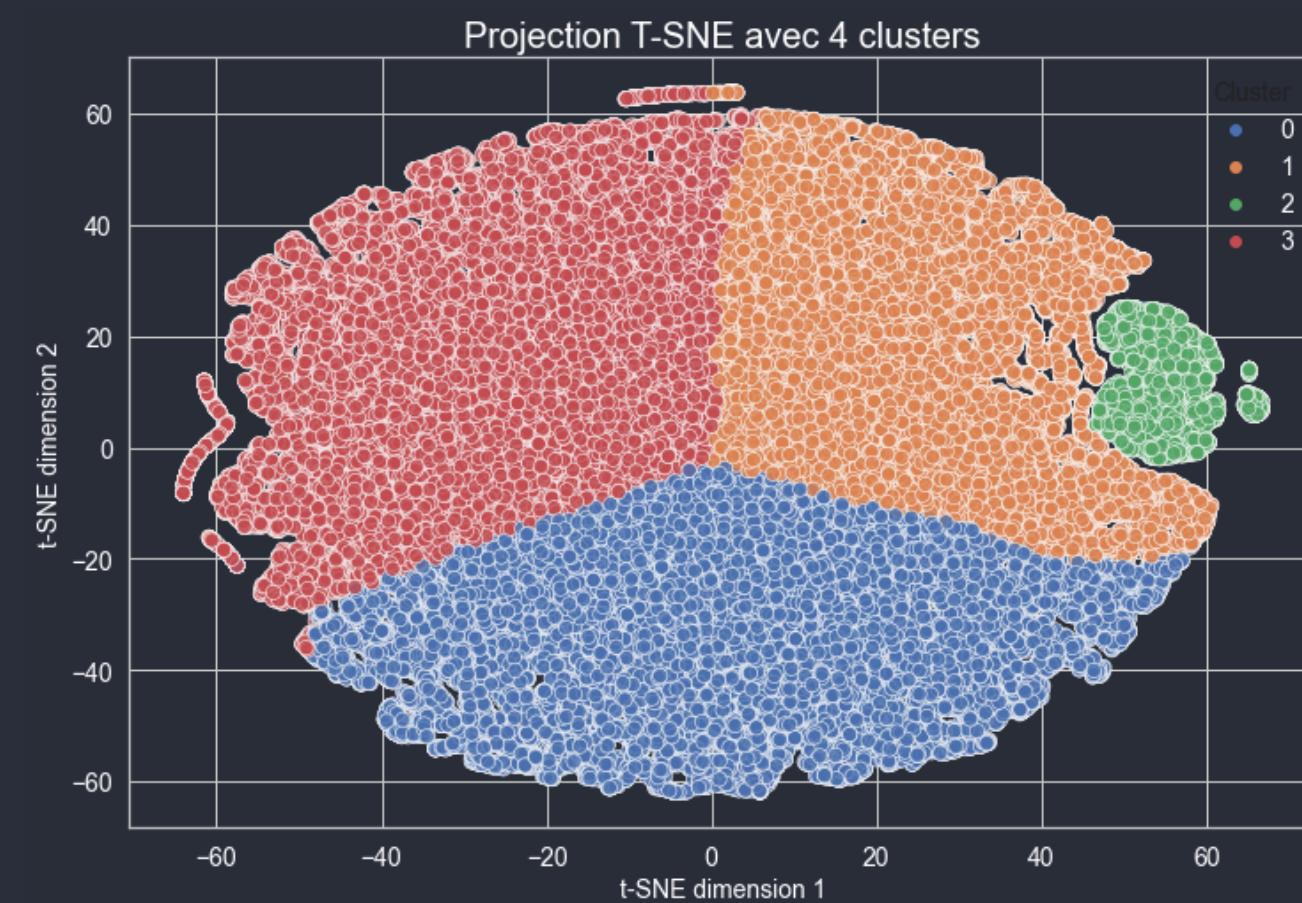
# K-MEANS



Distribution cluster

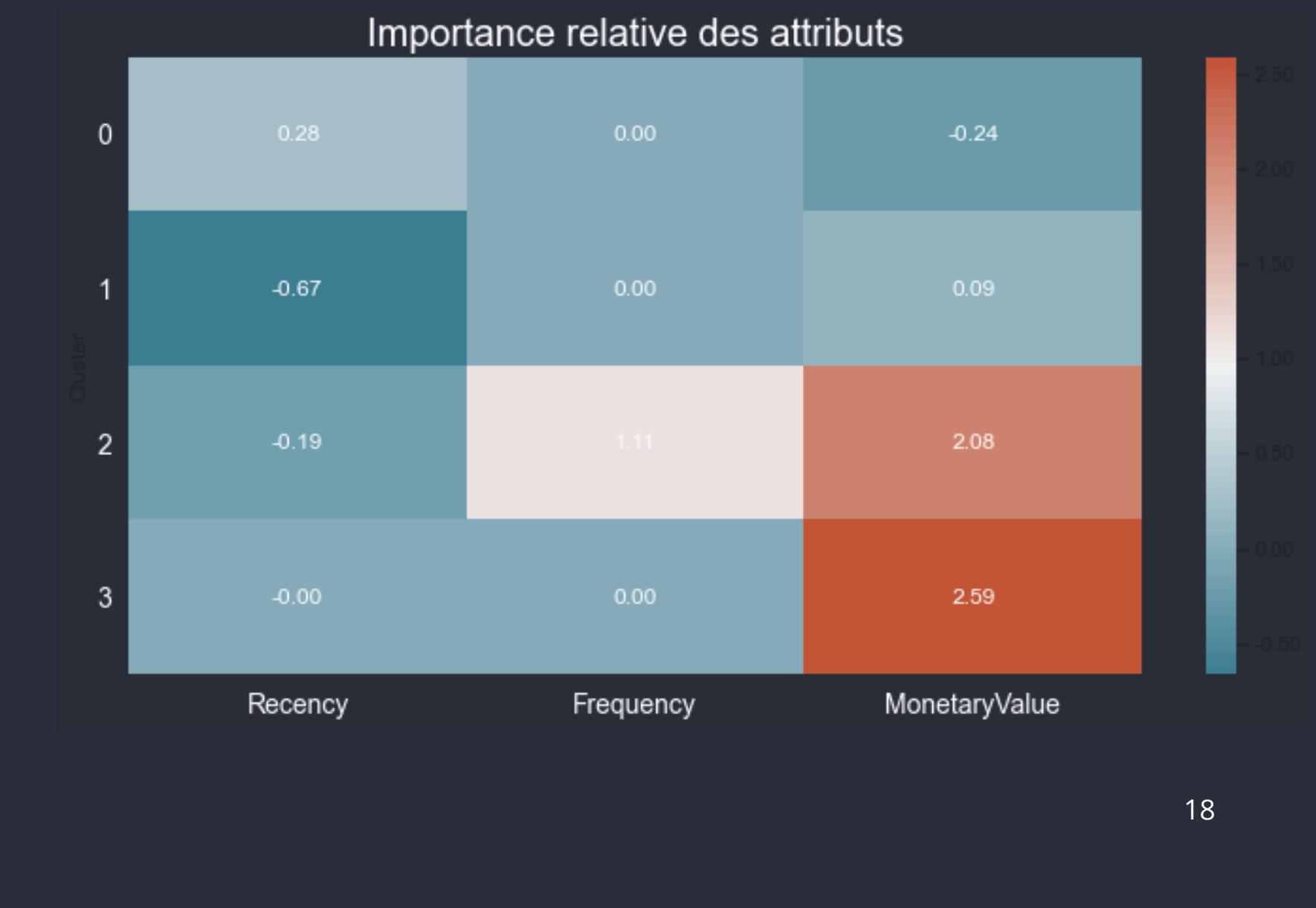
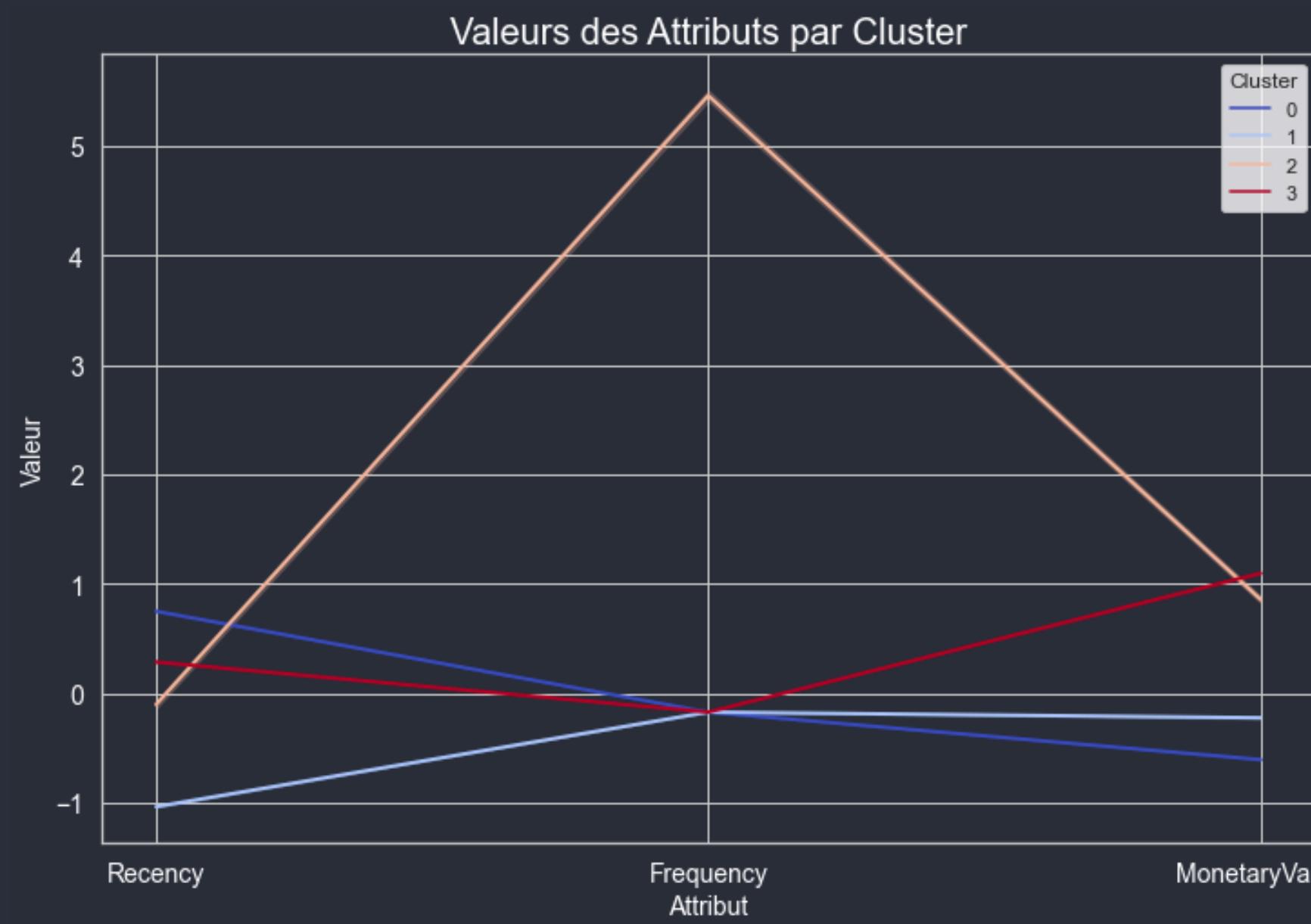


score silhouette = 0.3586



K-MEANS SUR DONNÉES RFM

# K-MEANS

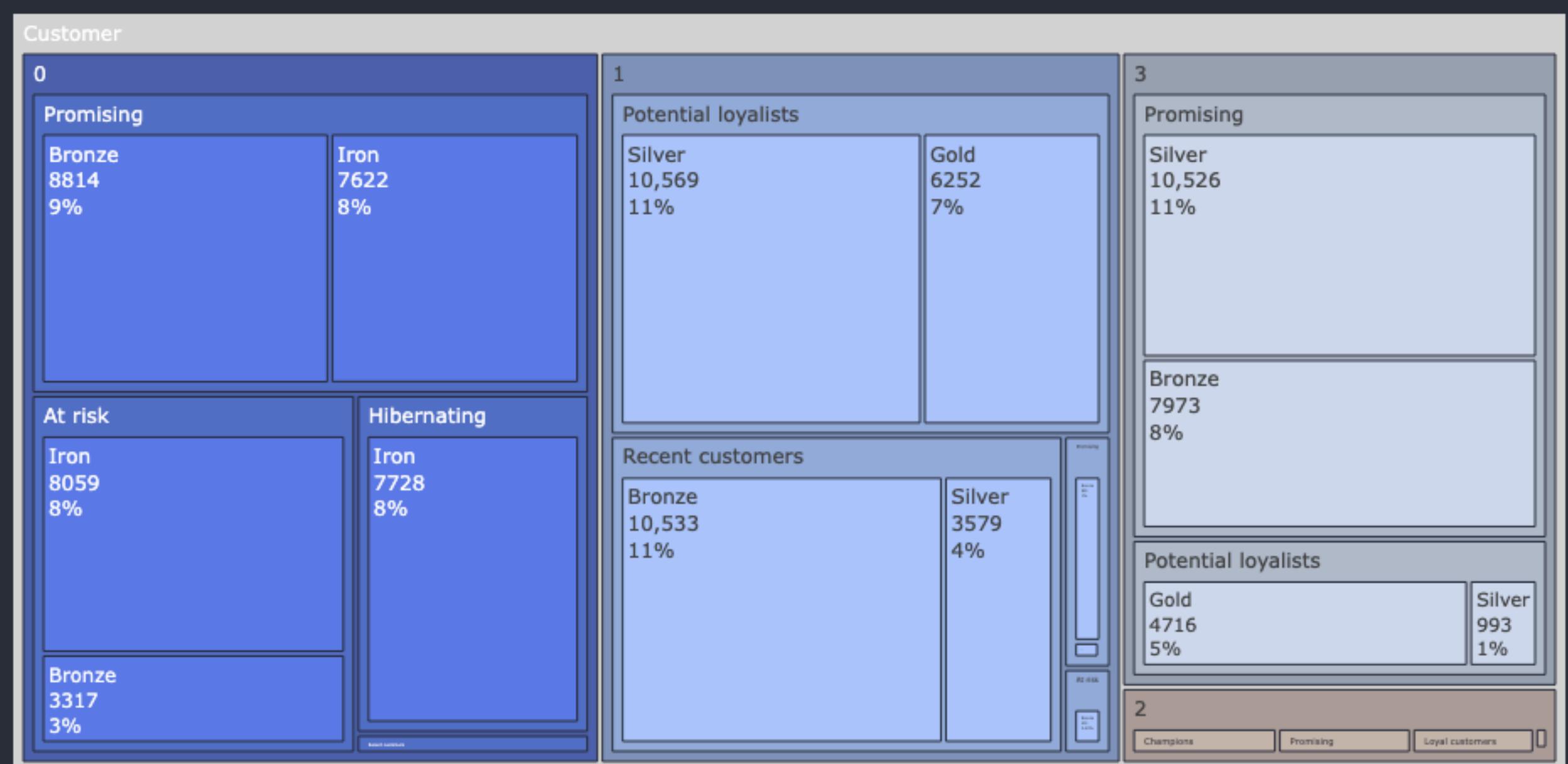




## K-MEANS SUR DONNÉES RFM

# K-MEANS

Treemap des Segmentations clients



## K-MEANS SUR DONNÉES RFM

# K-MEANS :

# INTERPRÉTATION

Cluster	Type de clients	%	Label RFM	Interprétation	Actions à mener
0	Clients à risque	38%	46% Promising; 32% At Risk; 21% Hibernating	Clients qui ont effectué leur dernière transaction il y a longtemps et qui ont effectué peu d'achats. Il pourrait donc s'agir du groupe des clients à risques/perdus.	envoyer des e-mails personnalisés pour renouer ; proposer d'autres produits pertinents et des remises spéciales ; relancer l'intérêt avec une campagne de sensibilisation
1	Nouveaux clients	34%	52% Potential loyalist; 44% Recent Customers	Clients ayant effectué des transactions récemment et ayant une fréquence d'achat plus faible, avec un faible montant de dépenses. Selon la segmentation RFM, la moitié d'entre eux sont des potentiels clients fidèles.	leur donner un succès rapide ; commencer à construire une relation
2	Clients Loyaux	3%	35% Champions; 32% Promising; 30% Loyal customers	Les clients les plus fréquents ayant le montant dépensé le plus élevé et ayant effectué une commande récemment.	demander des avis ; les récompenser
3	Clients prometteurs	25%	76% Promising; 24% Potential loyalist	Clients qui achètent fréquemment et dépensent beaucoup, mais qui ont effectué leur dernier achat il y a quelque temps.	faire des offres à durée limitée ; offrir des échantillons de produits gratuits



## CLUSTERING

# AJOUT DE VARIABLES

Nous allons désormais inclure davantage de variables comme par exemple :

- “delayed” : qui vaut 0 ou 1 si la commande est arrivée dans les temps estimés ou non
- “average\_number\_products\_per\_order” : le nombre moyen de produit par commande
- etc...



## Étapes :



## CLUSTERING

N°	Modèle	Variables	Nombre de clusters	Score silhouette	Notes
1	k-means	'number_of_orders', 'total_price', 'time_since_last_order', 'average_number_products_per_order'	3	0.6873	Le score silhouette est bon mais d'un point de vue métier le nombre de cluster semble trop faible pour permettre une segmentation assez précise des utilisateurs.
2	k-means	'number_of_orders', 'total_price', 'average_number_products_per_order', 'delayed'	5	0.5664	Assez bon score silhouette. D'un point de vue métier, l'interprétation des clusters est à améliorer car l'absence de variable relative au temps ne permet pas d'identifier les nouveaux clients, les clients absents, etc.
3	k-means	'number_of_orders', 'total_price', 'time_since_last_order', 'average_number_products_per_order', 'average_delivery_time', 'average_review_score', 'delayed'	4	0.5301	Assez bon score silhouette. D'un point de vue métier nous avons 4 clusters qui sont facilement interprétables.
4	k-means	'total_price', 'delayed', 'average_number_products_per_order', 'time_since_last_order', 'number_of_orders'	4	0.6060	Bon score silhouette. D'un point de vue métier nous avons 4 clusters qui sont assez facilement interprétables.
5	DBSCAN	idem que pour le modèle 3	11	0.4205	Les hyper-paramètres du modèle ont été optimisés. Le score silhouette est le moins bon.

## CLUSTERING

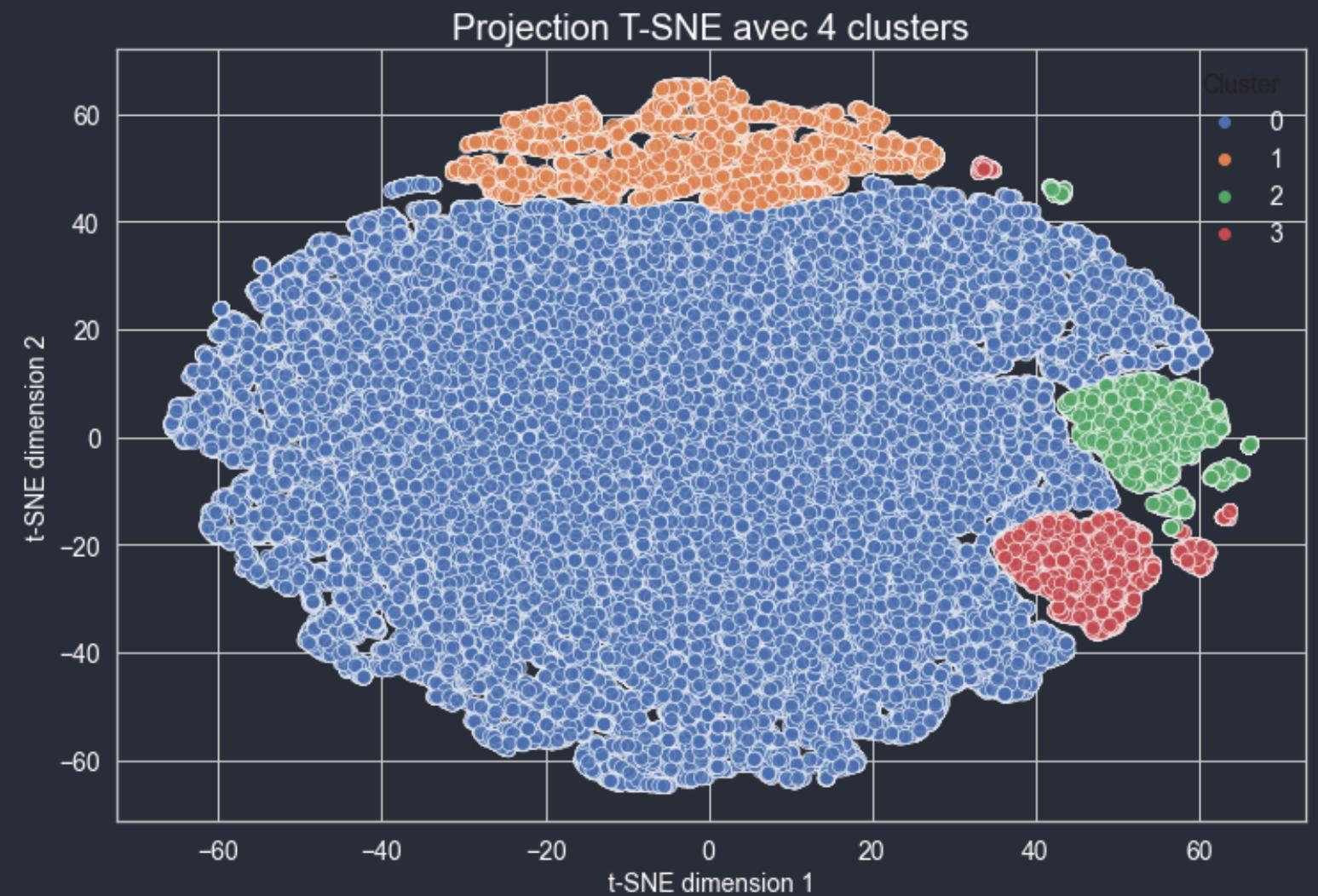
N°	Modèle	Variables	Nombre de clusters	Score silhouette	Notes
1	k-means	'number_of_orders', 'total_price', 'time_since_last_order', 'average_number_products_per_order'	3	0.6873	Le score silhouette est bon mais d'un point de vue métier le nombre de cluster semble trop faible pour permettre une segmentation assez précise des utilisateurs.
2	k-means	'number_of_orders', 'total_price', 'average_number_products_per_order', 'delayed'	5	0.5664	Assez bon score silhouette. D'un point de vue métier, l'interprétation des clusters est à améliorer car l'absence de variable relative au temps ne permet pas d'identifier les nouveaux clients, les clients absents, etc.
3	k-means	'number_of_orders', 'total_price', 'time_since_last_order', 'average_number_products_per_order', 'average_delivery_time', 'average_review_score', 'delayed'	4	0.5301	Assez bon score silhouette. D'un point de vue métier nous avons 4 clusters qui sont facilement interprétables.
4	k-means	'total_price', 'delayed', 'average_number_products_per_order', 'time_since_last_order', 'number_of_orders'	4	0.6060	Bon score silhouette. D'un point de vue métier nous avons 4 clusters qui sont assez facilement interprétables.
5	DBSCAN	idem que pour le modèle 3	11	0.4205	Les hyper-paramètres du modèle ont été optimisés. Le score silhouette est le moins bon.



CLUSTERING

# MODÈLE SÉLECTIONNÉ

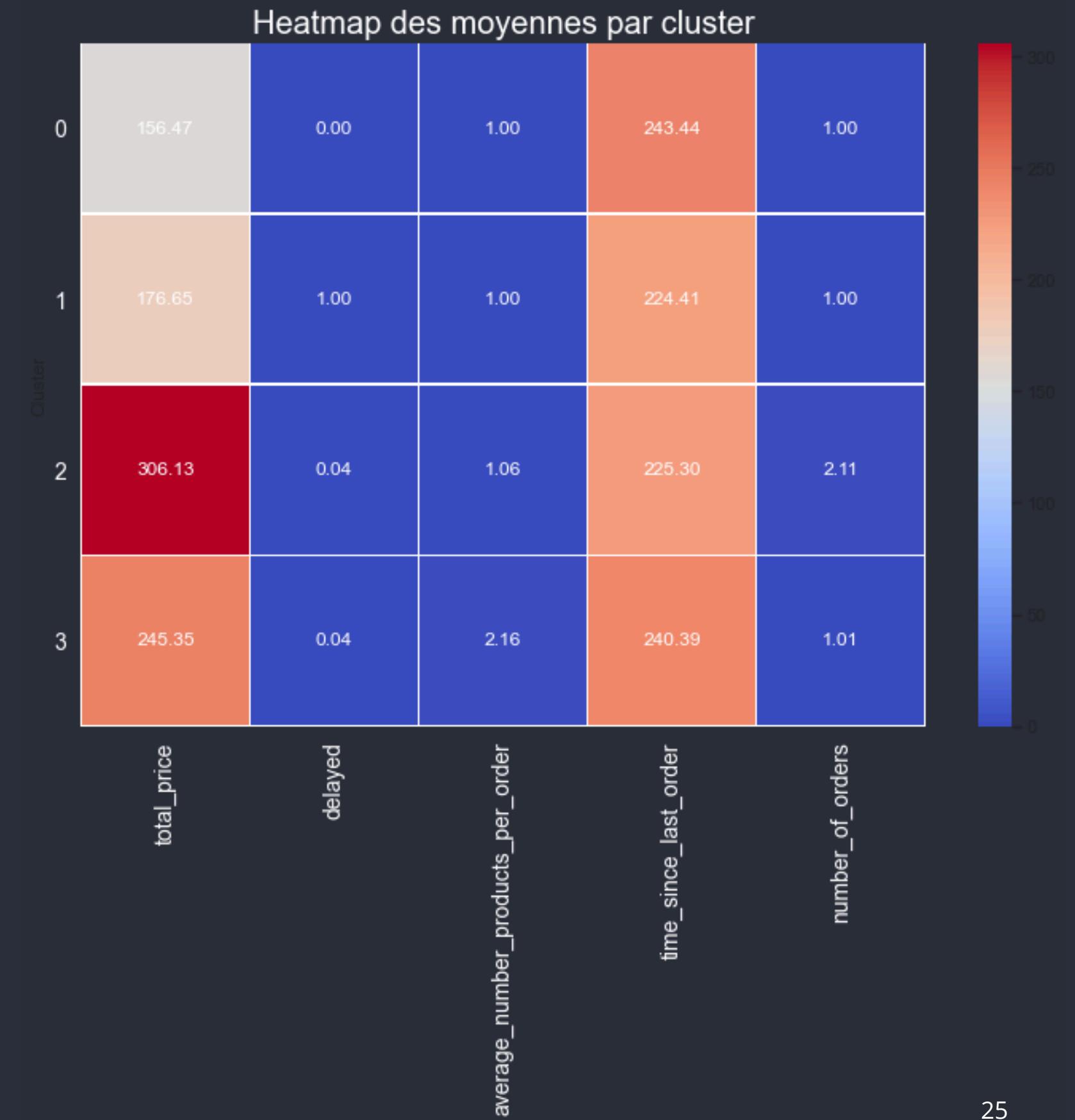
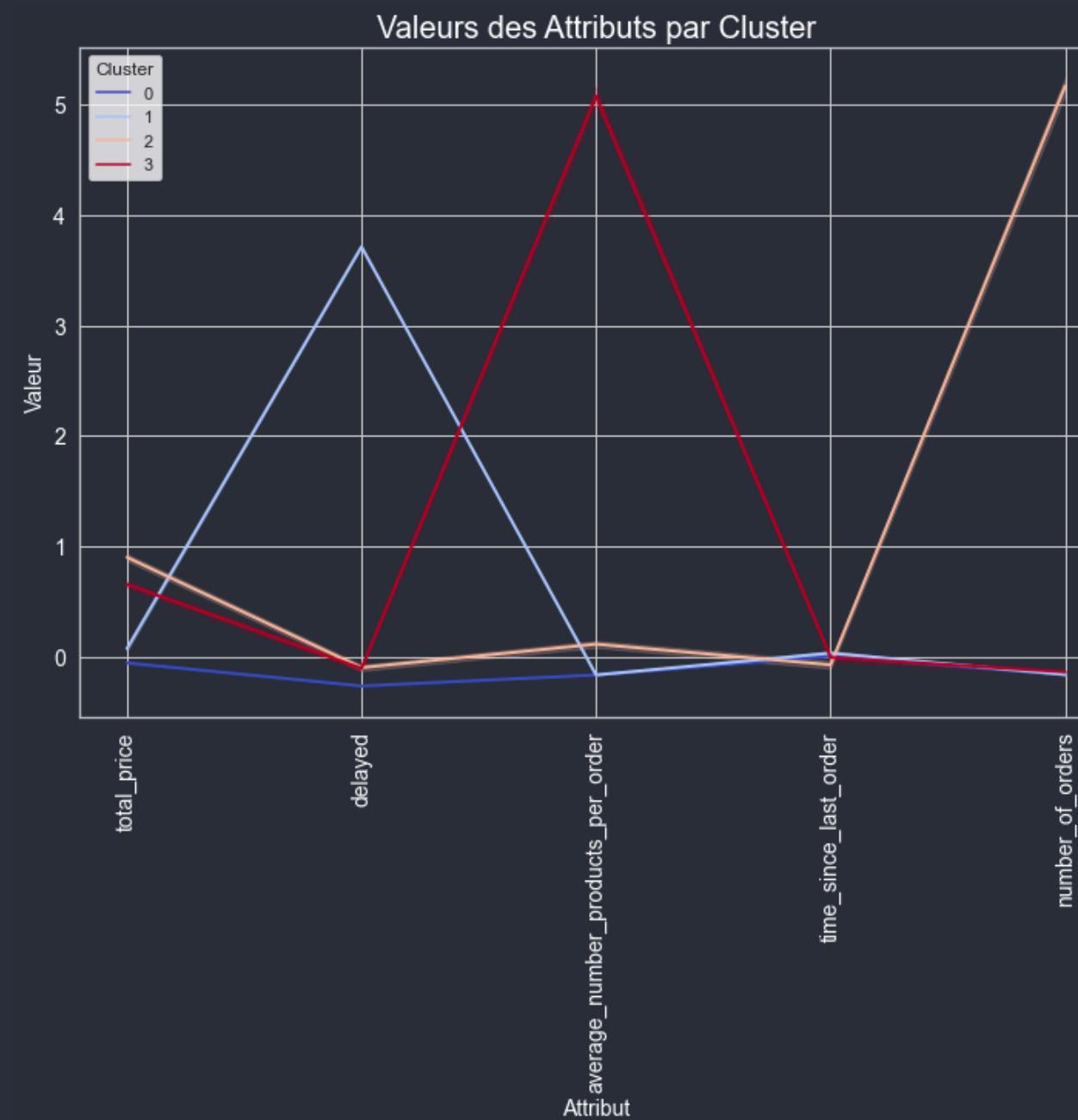
Distribution cluster





CLUSTERING

# MODÈLE SÉLECTIONNÉ





## CLUSTERING

# MODÈLE SÉLECTIONNÉ : INTERPRÉTATION

Cluster	Type de clients	%	Interprétation	Actions à mener
0	Clients à ne pas perdre	87%	Clients qui ont peu commandé et dépensé peu d'argent	faire des offres à durée limitée; proposer d'autres produits pertinents et des remises spéciales ; relancer l'intérêt avec une campagne de sensibilisation
1	Clients à risque	7%	Clients qui ont commandé une seule fois, qui ont donc peu dépensé et dont la commande a été livré en retard ce qui peut tendre vers une mauvaise expérience de leur part.	envoyer des e-mails personnalisés pour renouer ; offrir des échantillons de produits gratuits ; proposer d'autres produits pertinents et des remises spéciales ; relancer l'intérêt avec une campagne de sensibilisation ; faire des offres à durée limitée
2	Clients loyaux	3%	Clients qui ont commandé récemment, qui ont dépensé le plus et qui sont revenu plusieurs fois	demander des avis ; les récompenser
3	Clients prometteurs	3%	Clients qui ont beaucoup dépensé (grosse commande) mais qui n'ont pas passé un grand nombre de commande	faire des offres à durée limitée ; offrir des échantillons de produits gratuits, relancer l'intérêt avec une campagne de sensibilisation



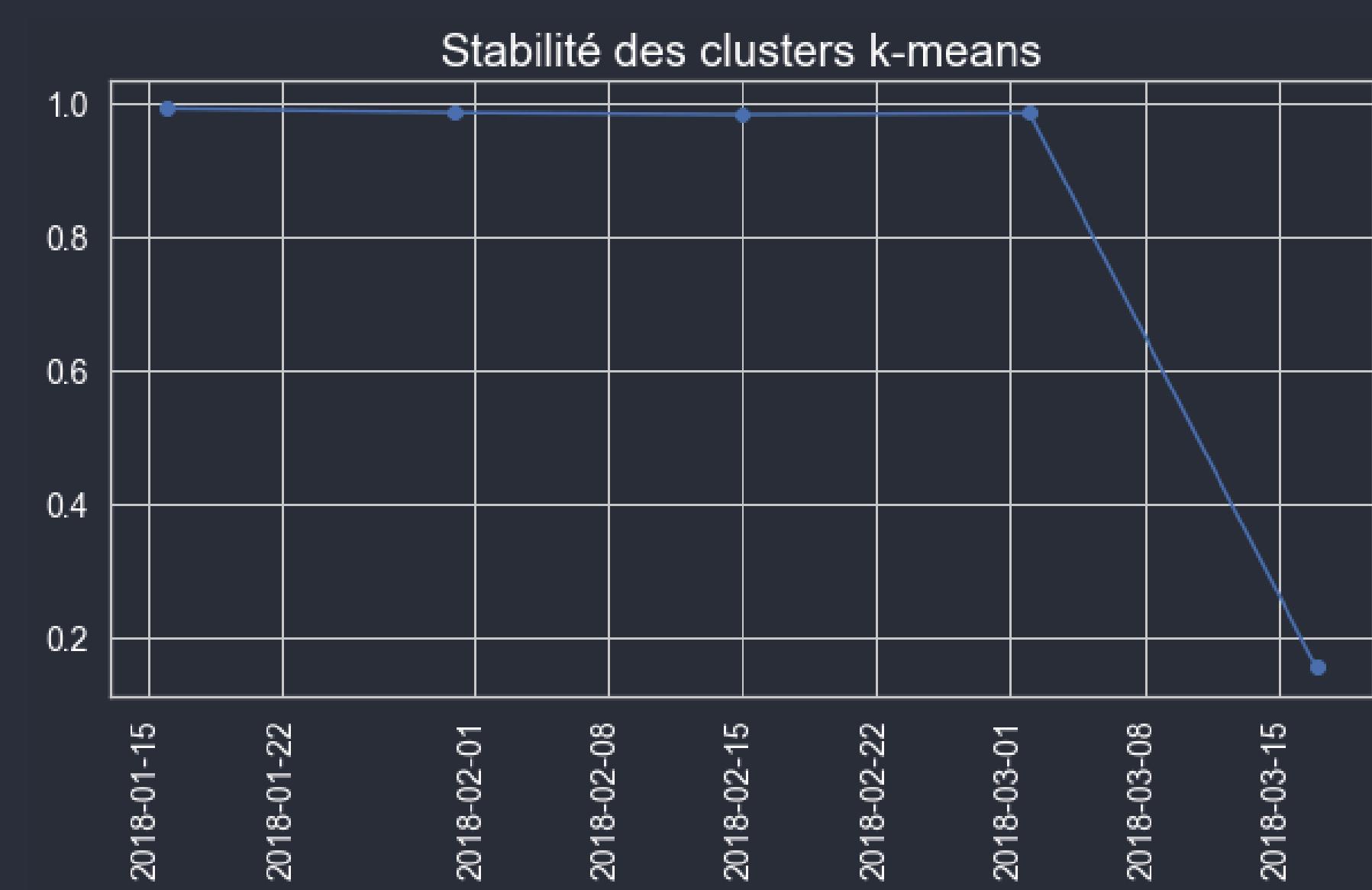
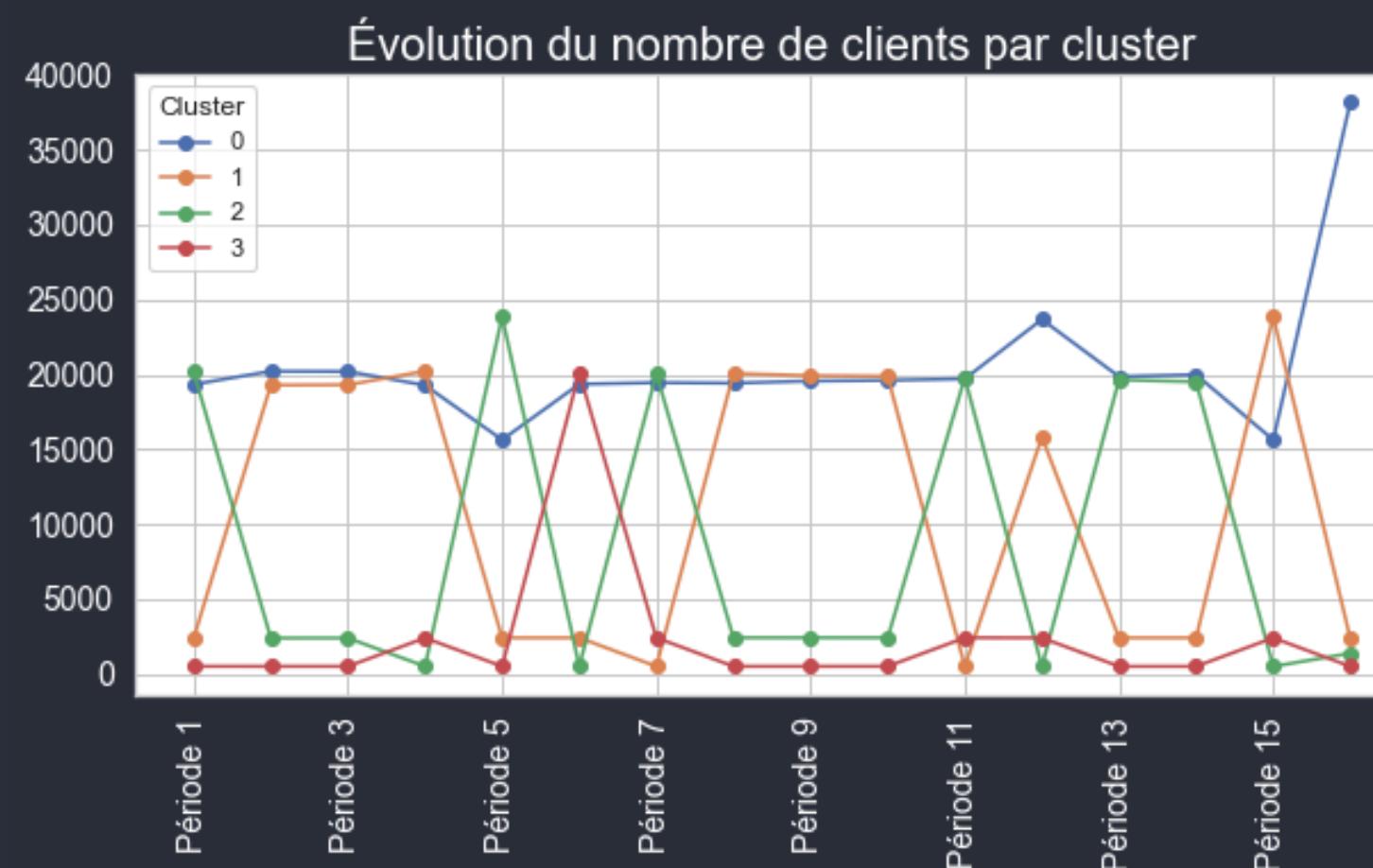
4

# MAINTENANCE



MAINTENANCE

# SIMULATION



On évalue la stabilité des clusters créés à l'aide de l'algorithme k-means. Pour cela on calcule le score Adjusted Rand Index (ARI), qui mesure la similarité entre les clusters obtenus à différents moments dans le temps (période de 15 jours). On créer un seuil à 0.7.

Finalement, au vu de l'évolution bi-mensuelle des clusters, de l'évolution du score ARI, la mise à jour devrait être faite au bout de 2 mois / 2 mois et demi pour garantir une stabilité optimale.



5

# CONCLUSION



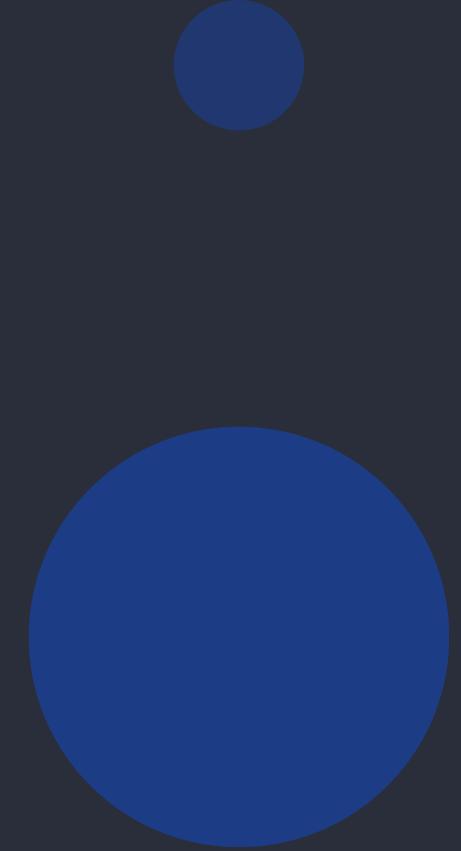
SEGMENTER DES CLIENTS D'UN SITE DE  
E-COMMERCE

# CONCLUSION

Ce projet de machine learning non supervisé a permis d'expérimenter plusieurs méthodes de segmentation, notamment à l'aide de divers algorithmes de clustering. Nous avons sélectionné le modèle le plus pertinent, interprété les différents clusters et analysé la stabilité du modèle dans le temps.

## Pistes d'amélioration :

- apporter de nouvelles données, ici nos données sont biaisées (presque tous les clients ne commandent qu'une seule fois)
- ajouter des variables relatives aux utilisateurs (ex : sexe, âge, profession, etc..)
- explorer d'autres algorithmes de clustering
- échanger avec l'équipe marketing pour la pertinence des variables du modèle et la précision de la segmentation souhaitée.



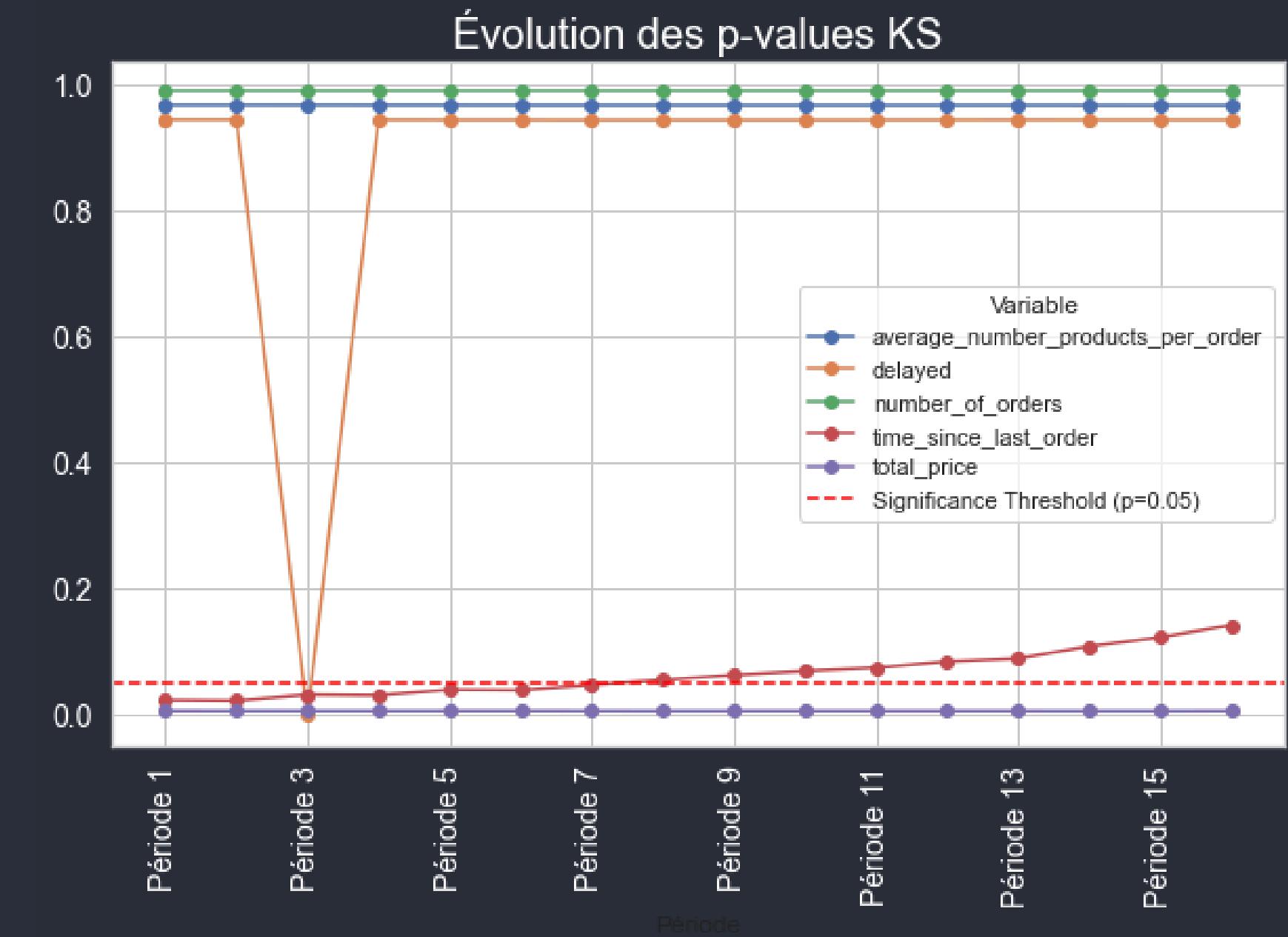
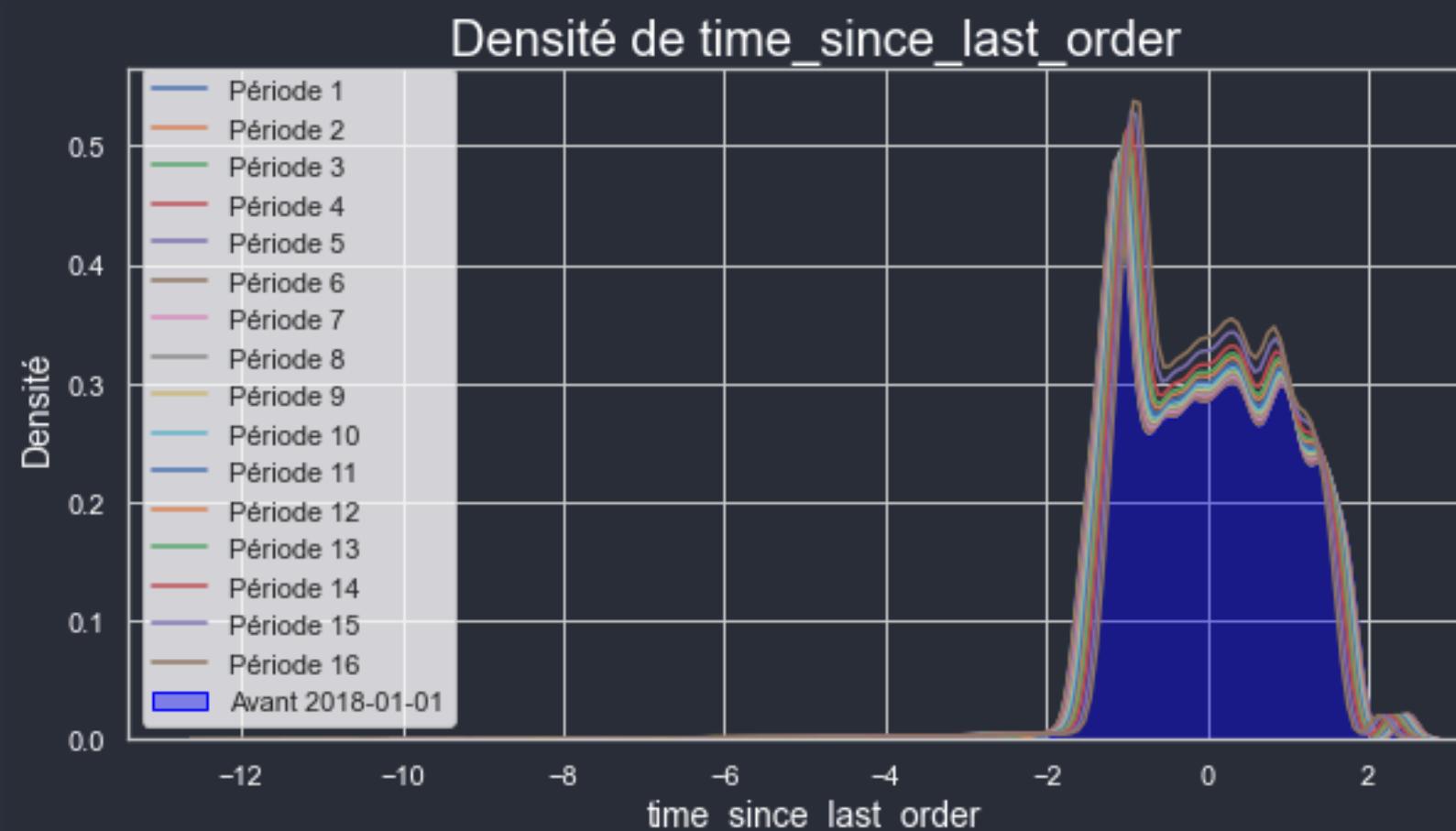
# ANNEXES



MAINTENANCE

# SIMULATION

Nous avons observé la distribution de nos variables numériques dans les données utilisées pour l'entraînement initial et les prédictions ultérieures.



Aux vues des résultats, la stratégie pourrait être de réentraîner le modèle à des intervalles réguliers, par exemple tous les 1 à 2 mois.