



# SEGMENTATION DES CLIENTS

Juillet 2022

# SOMMAIRE



Rappel des objectifs



Présentation du jeu de données



Analyse exploratoire et feature engineering



Présentation des essais



Simulation



Conclusion

## RAPPELS DES OBJECTIFS

- ☐ Fournir aux équipes d'e-commerce une segmentation des clients afin de les aider pour les campagnes de communication.
- ☐ Comprendre les différents types d'utilisateurs à partir des données.
- ☐ Fournir à l'équipe marketing une description actionnable de la segmentation.
- ☐ Fournir à l'équipe marketing une proposition de contrat de maintenance.

# SOMMAIRE



Rappel des objectifs



Présentation du jeu de données



Analyse exploratoire et feature engineering



Présentation des essais

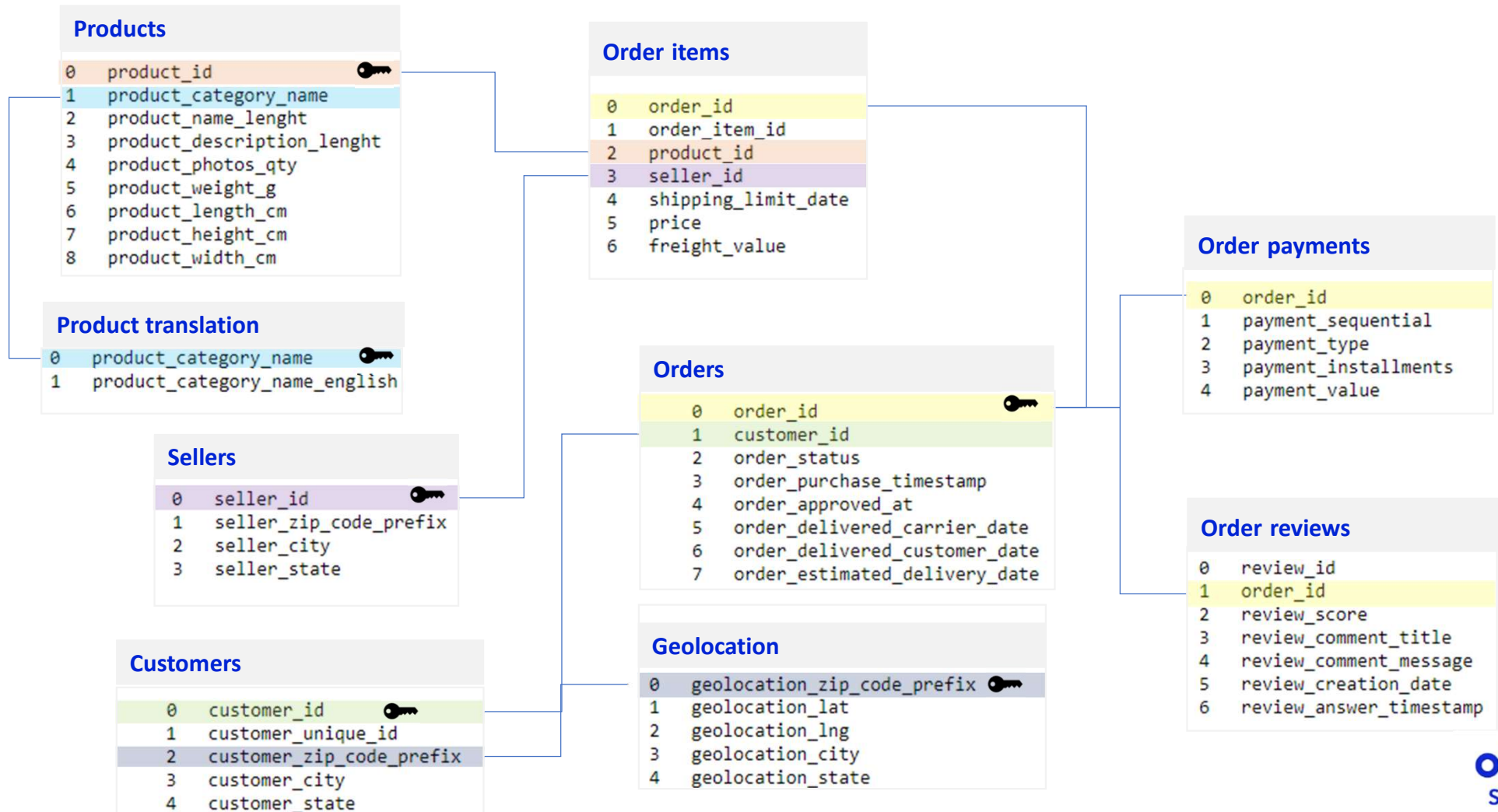


Simulation

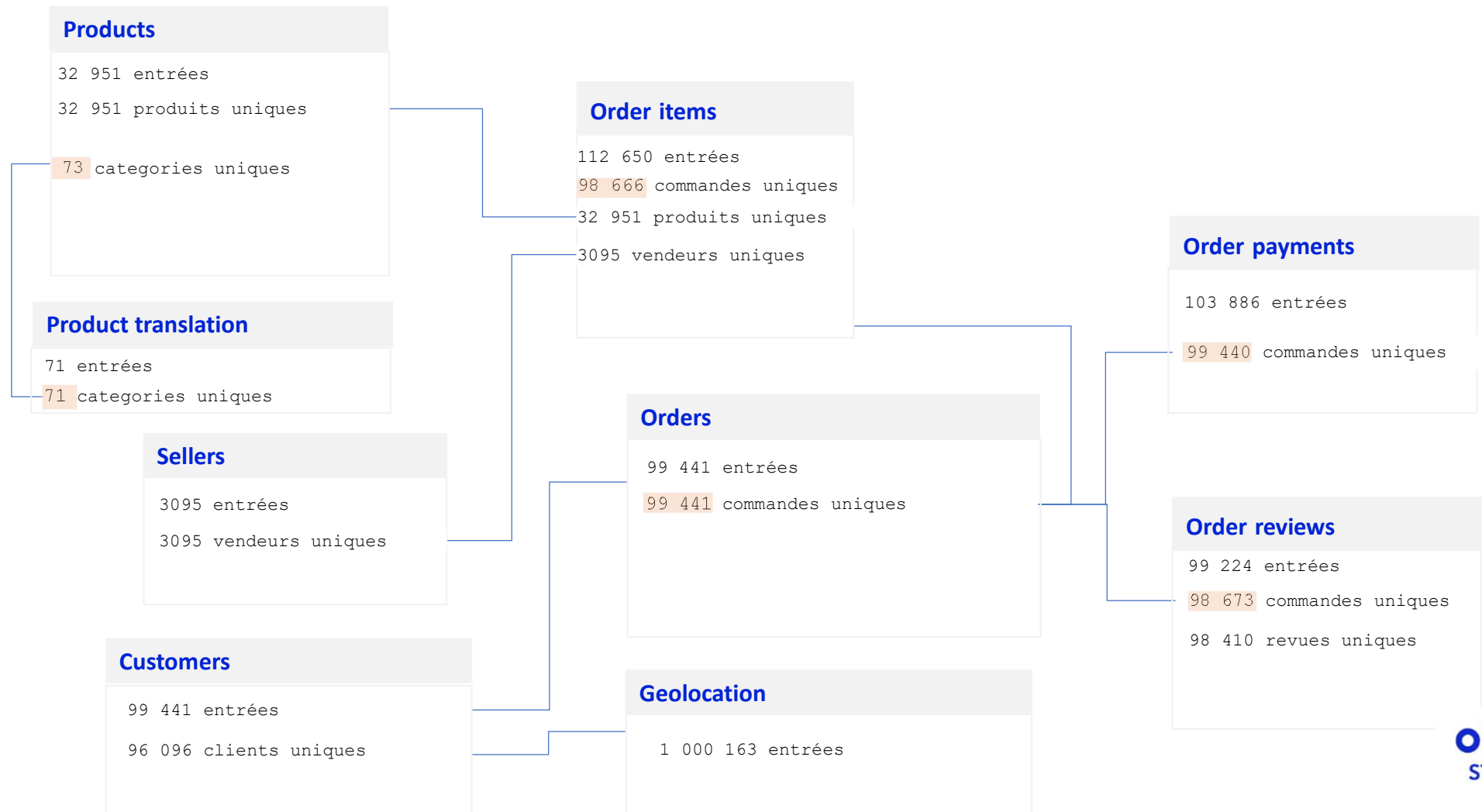


Conclusion

# SCHEMA RELATIONNEL DES TABLES FOURNIES



# NOMBRE D'ENTREES ET IDENTIFICATION DES ECARTS ENTRE LES TABLES



# ANALYSE ET RECONCILIATION DES ECARTS



Nbre de commandes dans Orders qui ne sont pas dans Order\_payments : 1  
Nbre de commandes dans Order\_payments qui ne sont pas dans Orders : 0

Commande non payée



Nbre de commandes dans Orders qui ne sont pas dans Order\_reviews : 768  
Nbre de commandes dans Order\_reviews pas dans Orders : 0

Commandes non  
commentées ou notées



Nbre de commandes dans Orders qui ne sont pas dans Order\_items : 775  
Nbre de commandes dans Order\_items qui ne sont pas dans Orders : 0



Montant total des commandes	15,843,553.24
Montant total des paiements	<u>16,008,872.12</u>
Ecart	165,318.88

Montant des 775 commandes en écart	162,591.95
Autres écarts	<u>2,726.93</u>
	165,318.88

1% des paiements  
correspondent à des  
commandes en anomalie

# SOMMAIRE



Rappel des objectifs



Présentation du jeu de données



Analyse exploratoire et feature engineering



Présentation des essais



Simulation

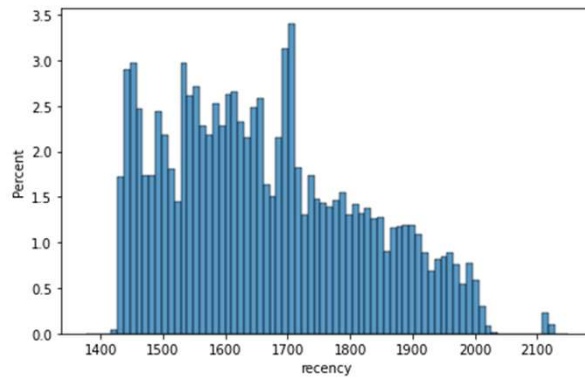


Conclusion



# CREATION DES VARIABLES RFM

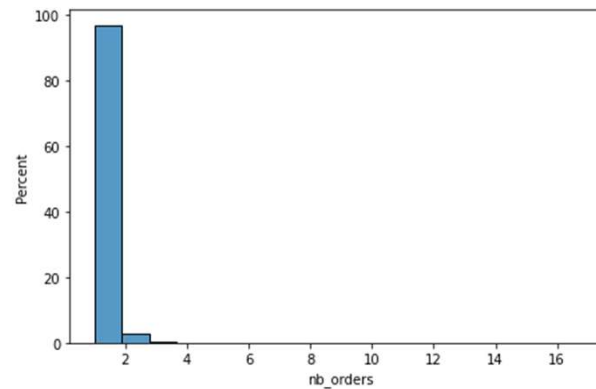
## Récence



Nombre de jours depuis le dernier achat.

recency	
count	96096.000000
mean	287.735691
std	153.414676
min	0.000000
25%	163.000000
50%	268.000000
75%	397.000000
max	772.000000

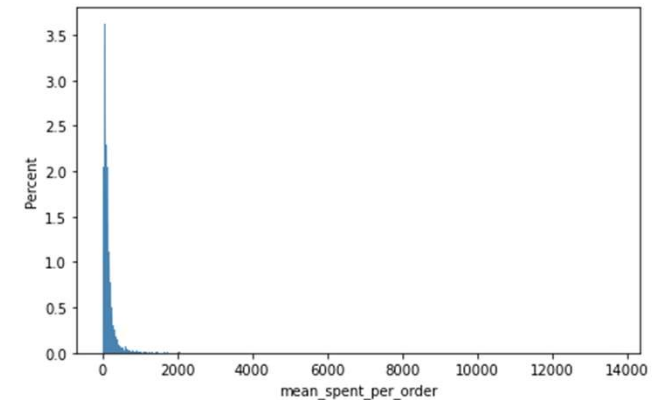
## Fréquence



Nombre de commandes sur la période.

nb_orders	
count	96096.000000
mean	1.034809
std	0.214384
min	1.000000
25%	1.000000
50%	1.000000
75%	1.000000
max	17.000000

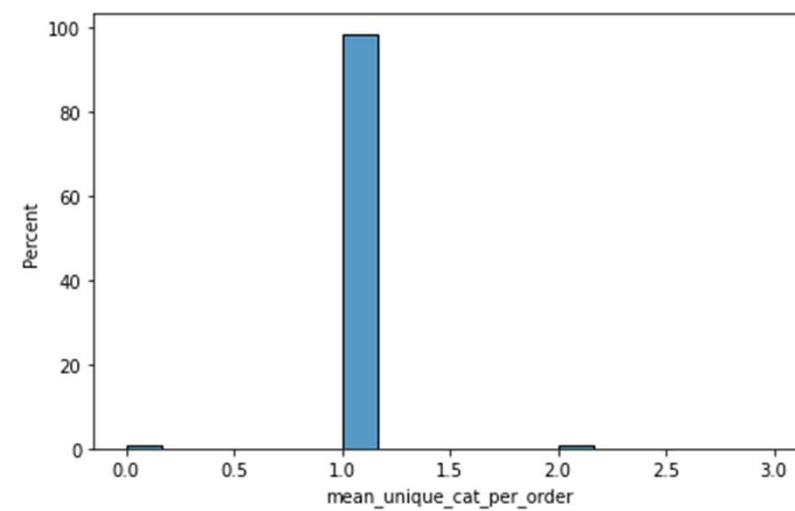
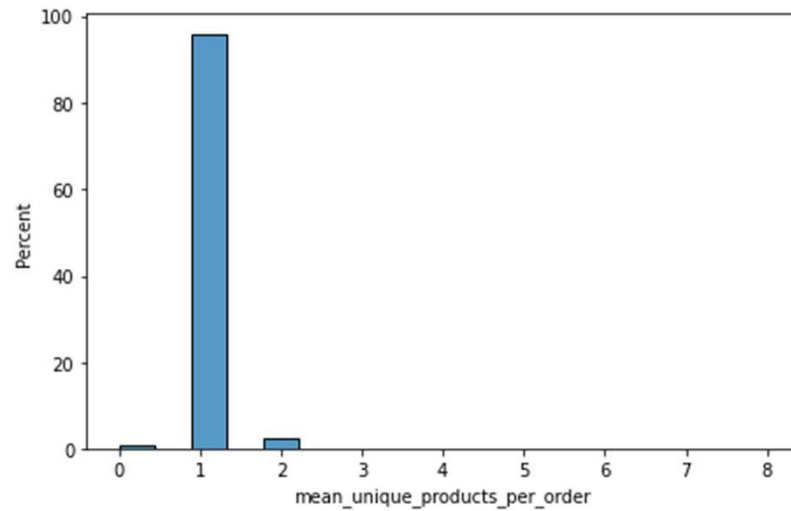
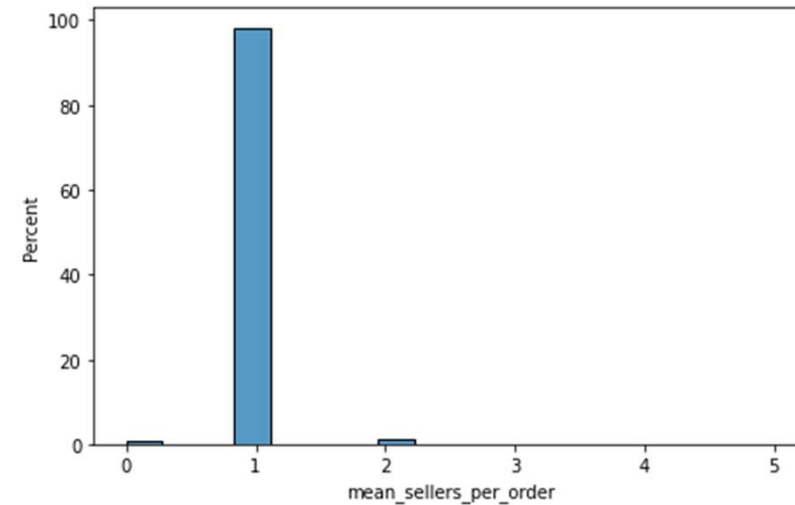
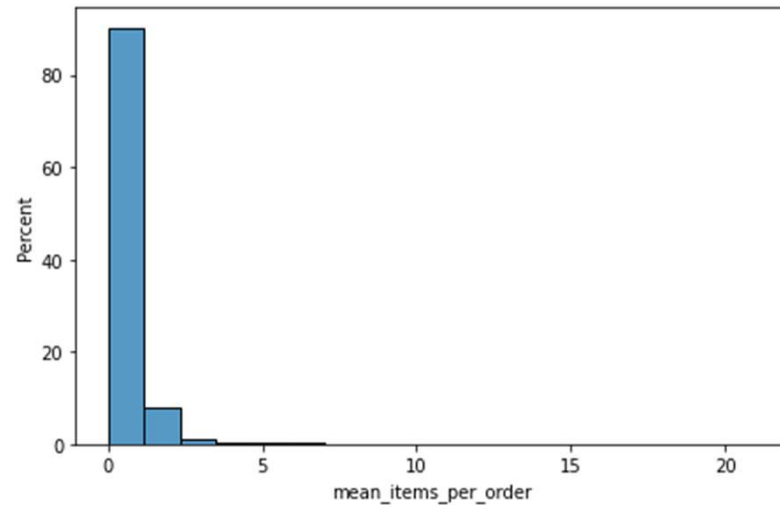
## Montant



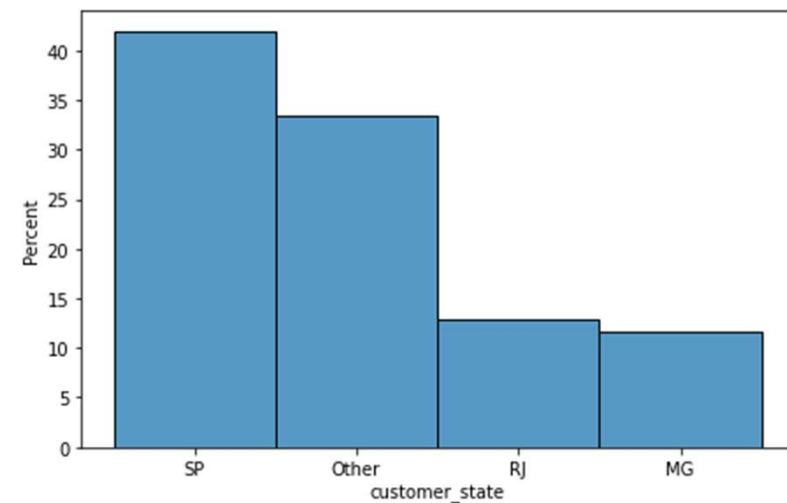
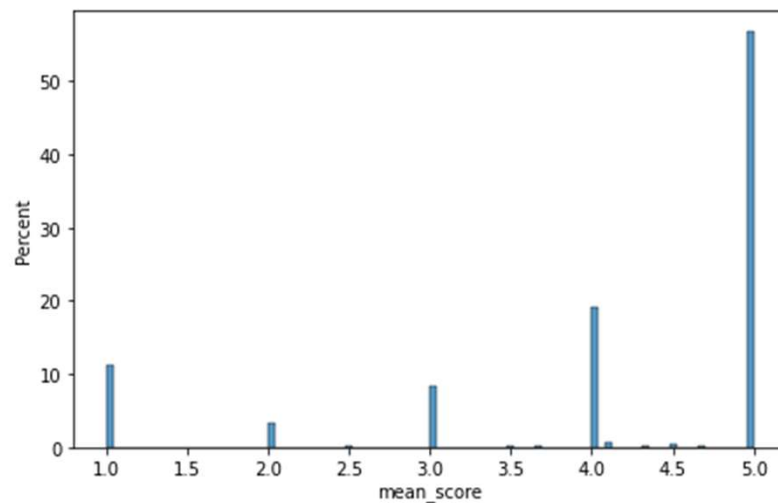
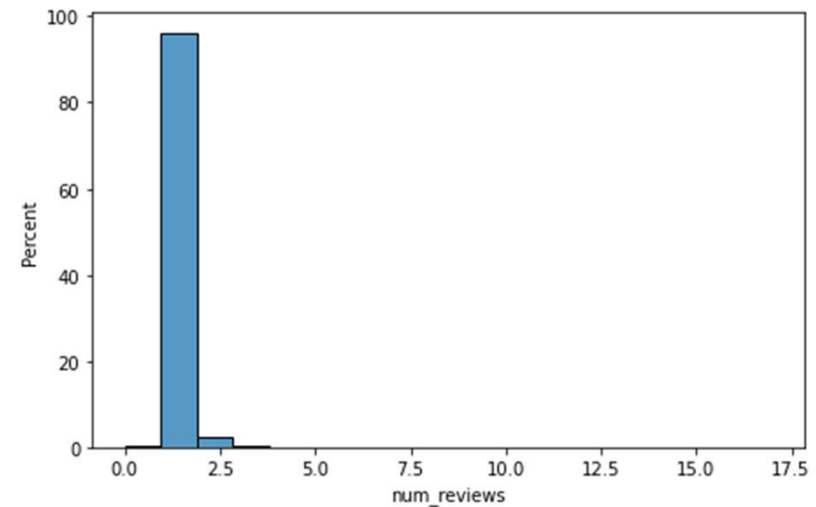
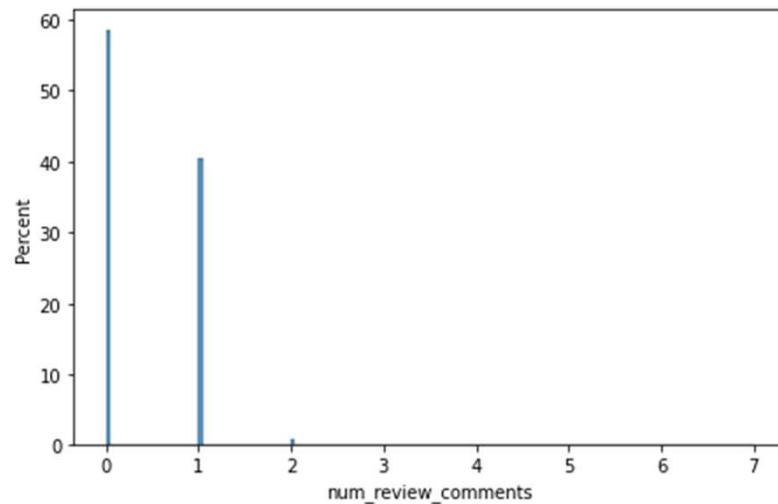
Montant moyen dépensé par commande.

mean_spent_per_order	
count	96096.000000
mean	161.373125
std	222.291655
min	0.000000
25%	62.440000
50%	105.810000
75%	177.160000
max	13664.080000

## AUTRES VARIABLES RELATIVES AUX COMMANDES



## VARIABLES RELATIVES A LA SATISFACTION ET LOCALISATION DES CLIENTS



# VARIABLES RETENUES POUR LA MODELISATION FINALE

nb_orders	1	-0.0089	0.025	0.017	0.027	0.018	0.17	0.89	0.0063	-0.025	0.075
mean_spent_per_order	-0.0089	1	0.18	0.031	0.063	0.015	0.038	-0.015	-0.049	0.00093	-0.00034
mean_items_per_order	-0.025	0.18	1	0.32	0.49	0.27	0.049	0.017	-0.085	-0.0097	-0.0078
mean_sellers_per_order	-0.017	0.031	0.32	1	0.66	0.75	0.02	0.018	0.0037	-0.049	-0.047
mean_unique_products_per_order	-0.027	0.063	0.49	0.66	1	0.56	0.034	0.025	-0.033	-0.02	-0.018
mean_unique_cat_per_order	-0.018	0.015	0.27	0.75	0.56	1	-0.0006	0.021	0.056	-0.043	-0.04
num_review_comments	0.17	0.038	0.049	0.02	0.034	-0.0006	1	0.19	-0.25	0.024	0.042
num_reviews	0.89	-0.015	0.017	0.018	0.025	0.021	0.19	1	0.005	-0.016	0.063
mean_score	-0.0063	-0.049	-0.085	0.0037	-0.033	0.056	-0.25	0.005	1	-0.026	-0.024
recency	-0.025	0.00093	-0.0097	-0.049	-0.02	-0.043	0.024	-0.016	-0.026	1	0.99
age	0.075	-0.00034	-0.0078	-0.047	-0.018	-0.04	0.042	0.063	-0.024	0.99	1

Nous n'avons pas retenu les variables fortement corrélées aux variables mean\_items\_per\_order et mean\_unique\_cat\_per\_order.

Nous n'avons pas retenu de variable de localisation dans le modèle final.

## Variables retenues

- 1 nb\_orders
- 2 mean\_spent\_per\_order
- 3 mean\_items\_per\_order
- 4 mean\_unique\_cat\_per\_order
- 5 num\_review\_comments
- 6 mean\_score
- 7 recency

# SOMMAIRE



Rappel des objectifs



Présentation du jeu de données



Analyse exploratoire et feature engineering



Présentation des essais



Simulation



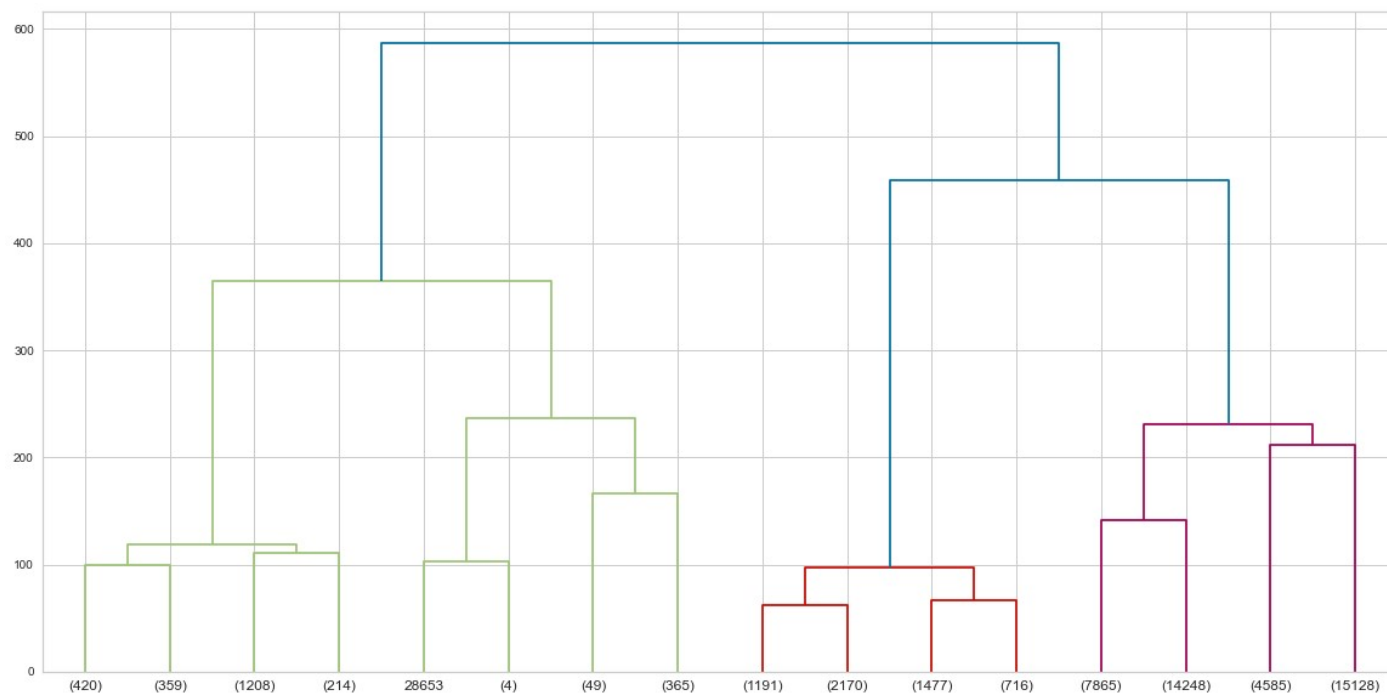
Conclusion

# CLUSTERING HIERARCHIQUE

Algorithme non retenu



Nécessité de travailler sur un échantillon car trop gourmand en ressources.



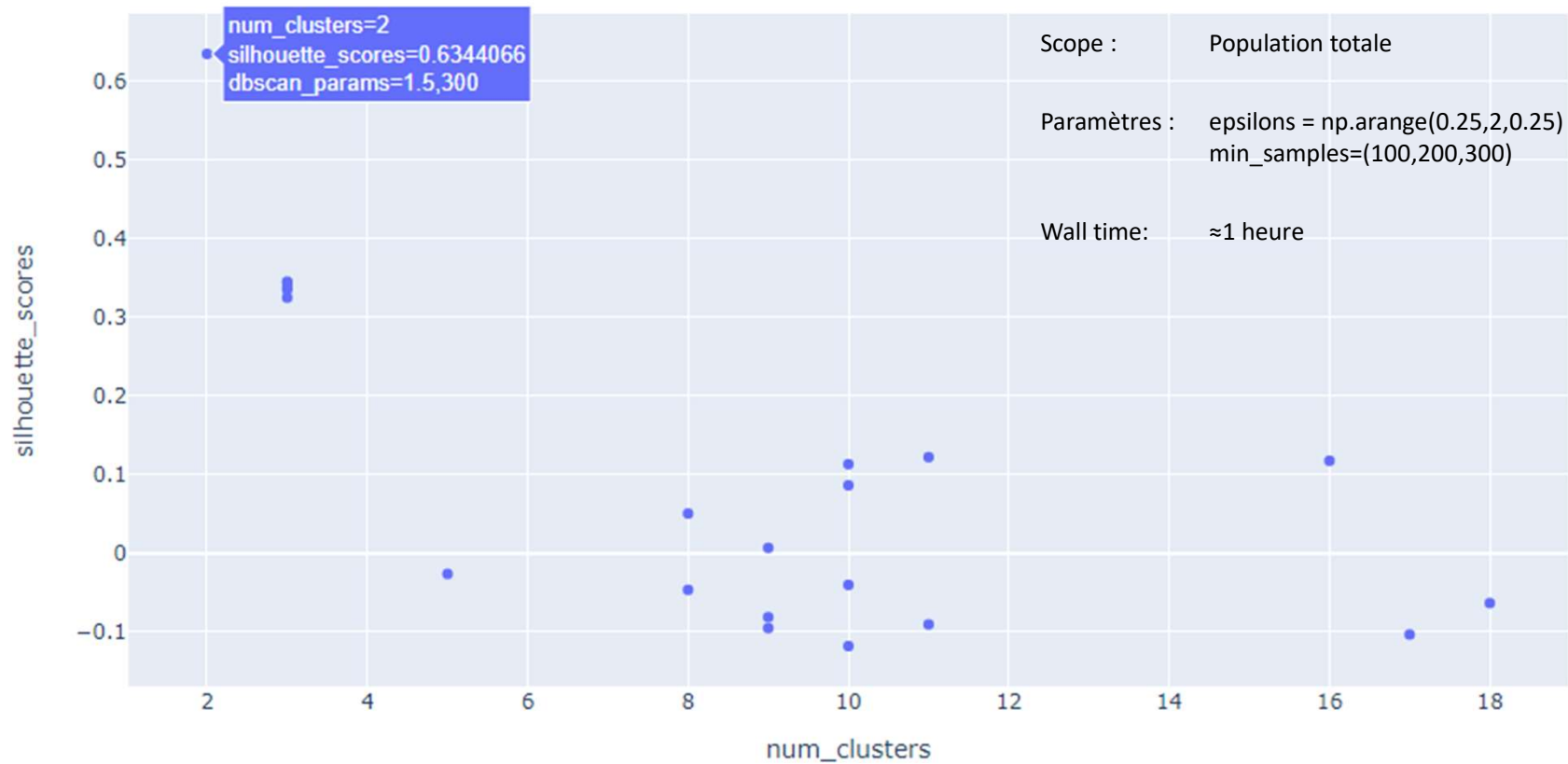
Scope	Echantillon
Taille	50 000
Nombre de clusters évalués	2-10
Meilleur nbre de cluster – dispersion score	4
Meilleur nbre de cluster – silhouette score	2
Fit time	≈3min

# DBSCAN

Algorithme non retenu



Temps de calcul très élevé

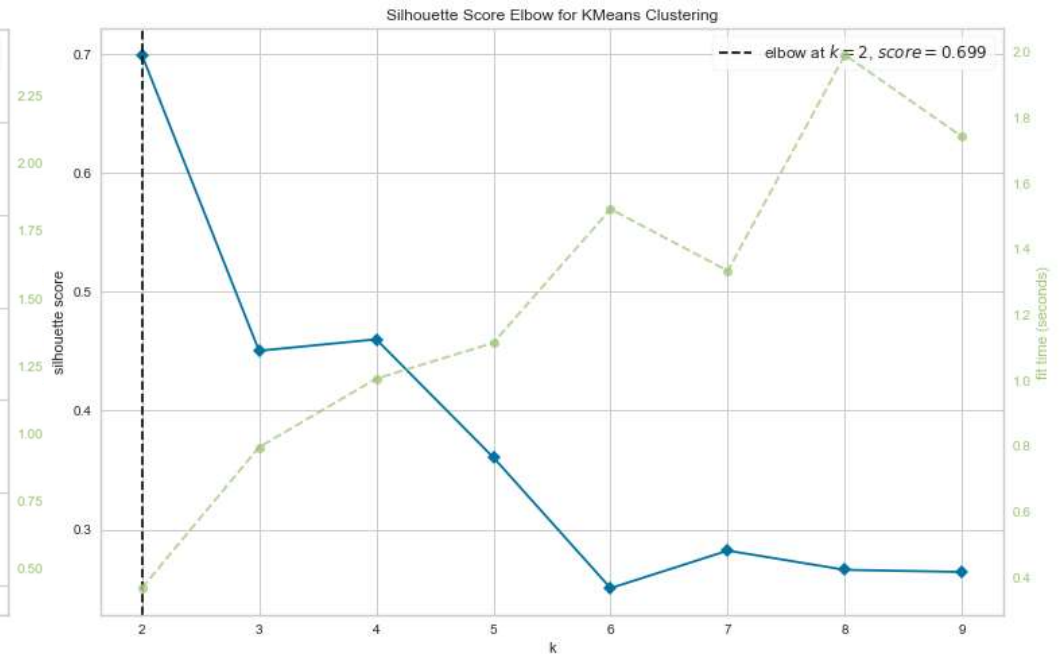
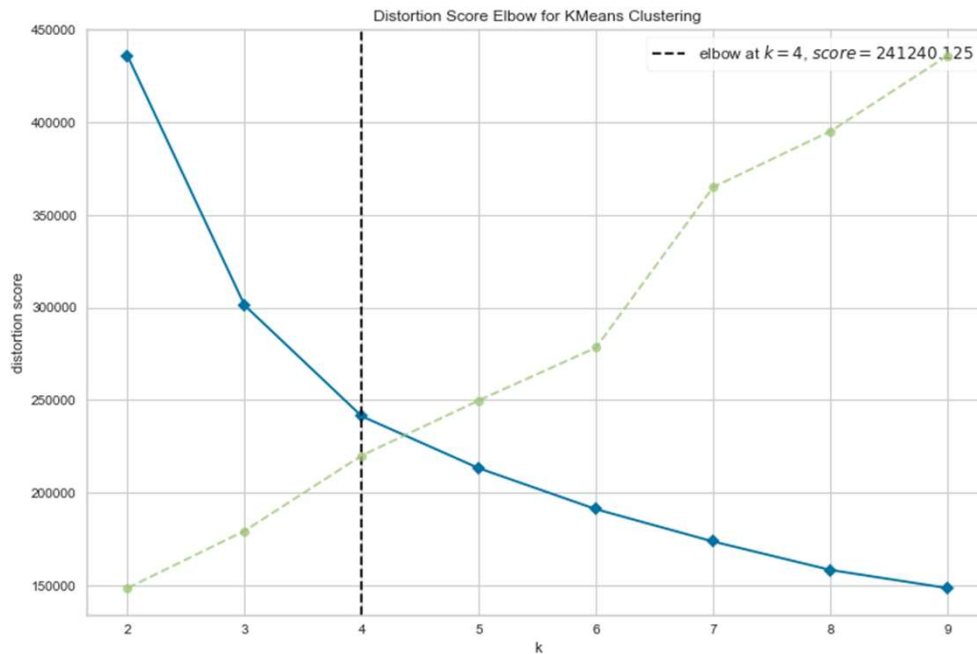


# KMEANS

Algorithme retenu

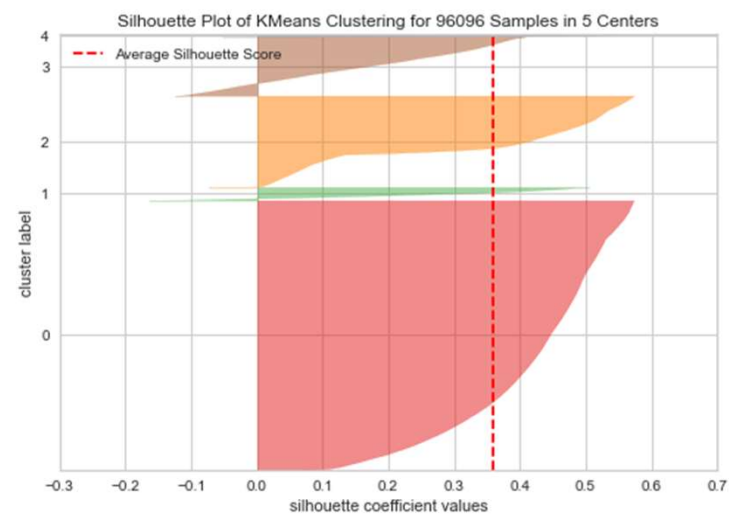
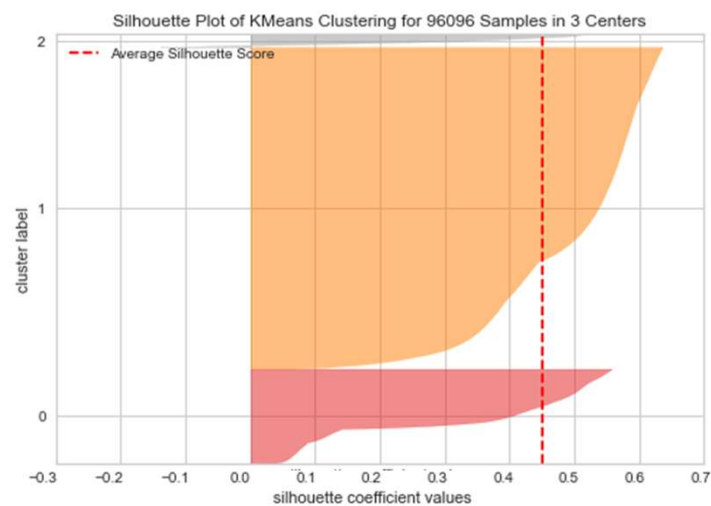
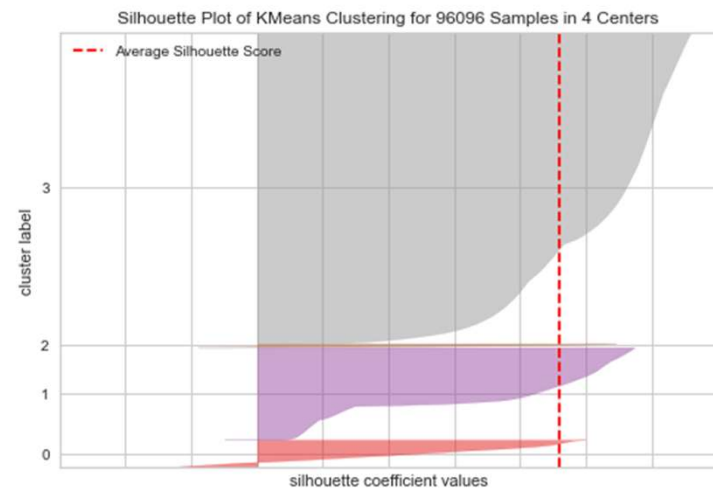
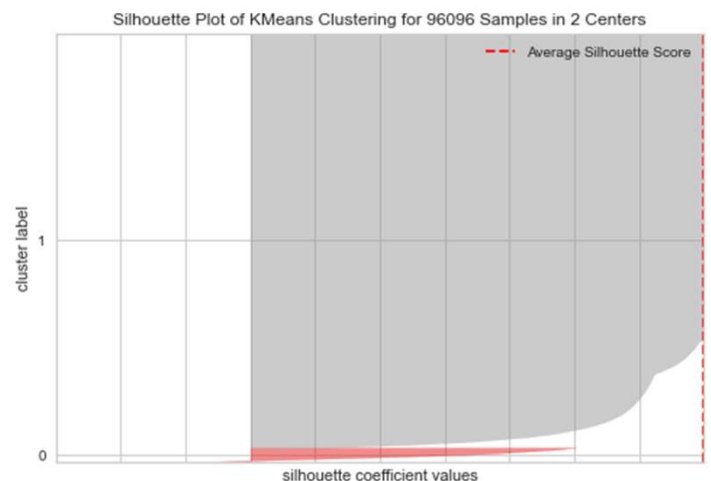


Scope : population totale  
fit time: < 1 seconde





# KMEANS - ANALYSE DE LA DENSITE DES CLUSTERS



# KMEANS - ANALYSES DES CLUSTERS

Analyse des centroïdes

nb_orders	1.0	1.0	1.0	1.0
mean_spent_per_order	114.2	1814.9	579.6	129.1
mean_items_per_order	1.1	1.5	1.4	1.2
mean_unique_cat_per_order	1.0	1.0	1.0	1.0
num_review_comments	0.4	0.5	0.4	0.7
mean_score	4.7	3.8	4.2	1.9
recency	285.2	297.3	289.0	295.4
	0	1	2	3

Clients très satisfaits mais qui dépensent peu

Clients à fort pouvoir d'achats dont la satisfaction peut être améliorée

Clients très satisfaits et qui ont un potentiel de dépenses élevées

Clients à fort potentiel de nuisance : mécontents ils dépensent peu et se manifestent au niveau des commentaires

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
	Clients à stimuler	Clients à fidéliser	Clients à développer	Clients à neutraliser
Population du cluster	68 709	795	6 163	20 429

# SOMMAIRE



Rappel des objectifs



Présentation du jeu de données



Analyse exploratoire et feature engineering



Présentation des essais



Simulation de maintenance



Conclusion

# METHODOLOGIE DE SIMULATION DE MAINTENANCE

Afin d'évaluer les besoins en maintenance, nous avons simulé une évolution de la base client dans le temps et utilisé l'adjusted rand score (ARI) pour comparer si une nouvelle segmentation de la base client nous donnerait des résultats similaires à la segmentation initialement effectuée pour Olist.

Nous avons procédé aux étapes suivantes :

I. Constitution de la base clients initiale à partir des clients dont l'ancienneté est supérieure à un an

II. Segmentation des clients de la base initiale

III. Evolution de la base clients sur des périodes de un à douze mois par ajouts des nouveaux clients de la période

IV. Après chaque période, deux segmentations de la nouvelle base client :

1. à partir de la segmentation obtenue à l'étape II.
2. Une segmentation directe

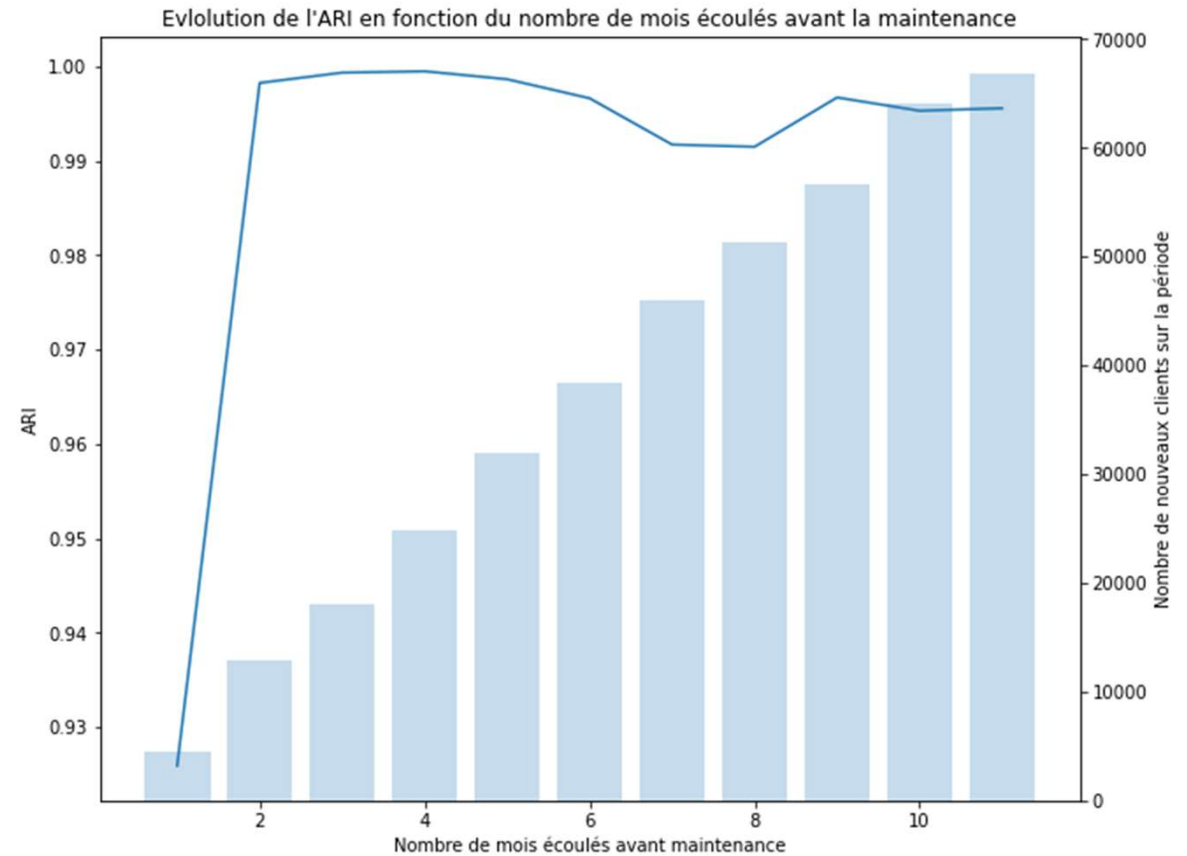
V. Calcul de l'ARI après chaque période.

VI. Analyse de l'évolution de l'ARI

# RESULTATS DE LA SIMULATION

Segmentation des clients anciens de plus d'un an

nb_orders -	1.0	1.1	1.1	1.0
mean_spent_per_order -	750.0	121.5	130.5	2230.7
mean_items_per_order -	1.4	1.1	1.1	1.6
mean_unique_cat_per_order -	1.0	1.0	1.0	0.9
num_review_comments -	0.5	0.4	0.7	0.5
mean_score -	4.2	4.7	1.9	3.7
recency -	117.0	109.4	120.7	122.2
	0	1	2	3



# SOMMAIRE



Rappel des objectifs



Présentation du jeu de données



Analyse exploratoire et feature engineering



Présentation des essais



Simulation



Conclusion

# CONCLUSION



## Choix de l'algorithme

K-Means



## Résultat de la segmentation

	Groupe 1	Groupe 2	Groupe 3	Groupe 4
Action à mener	Clients à stimuler	Clients à fidéliser	Clients à développer	Clients à neutraliser
Population	68 709	795	6 163	20 429



## Recommandation de maintenance

6 mois

La périodicité sera à réévaluer si suite aux campagnes de marketing il y a des modifications significatives dans la structure de la base clients.