

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

---

Parcours Data Scientist -  
OpenClassrooms

Soutenance



# Sommaire

I - Problématique

II - Exploration des données

III - Pistes de modélisation

IV - Modèle final

V - Maintenance



# I-Problématique

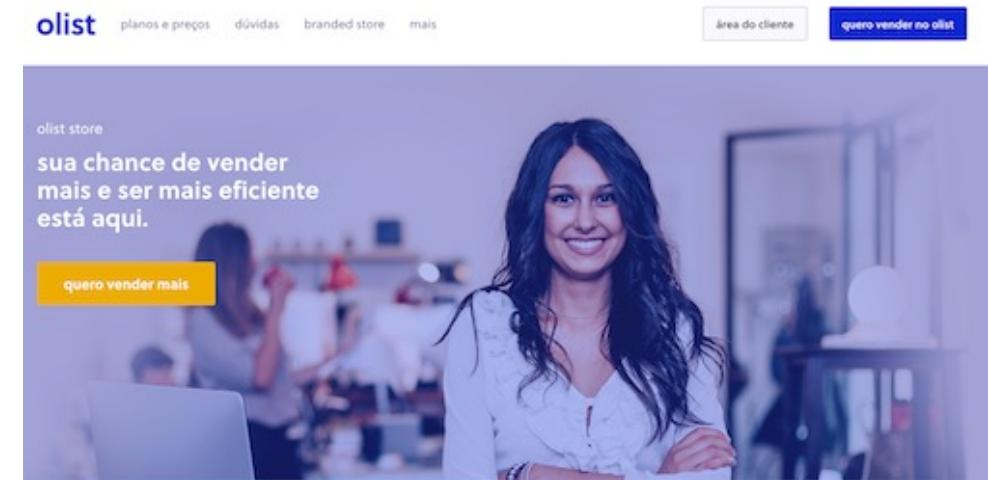
# Problématique

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

## • **Missions :**

- Fournir aux équipes d'e-commerce une segmentation des clients qu'elles pourront utiliser au quotidien pour leurs campagnes de communication.
- Comprendre les différents types d'utilisateurs.
- description actionable de la segmentation et proposition de contrat de maintenance.

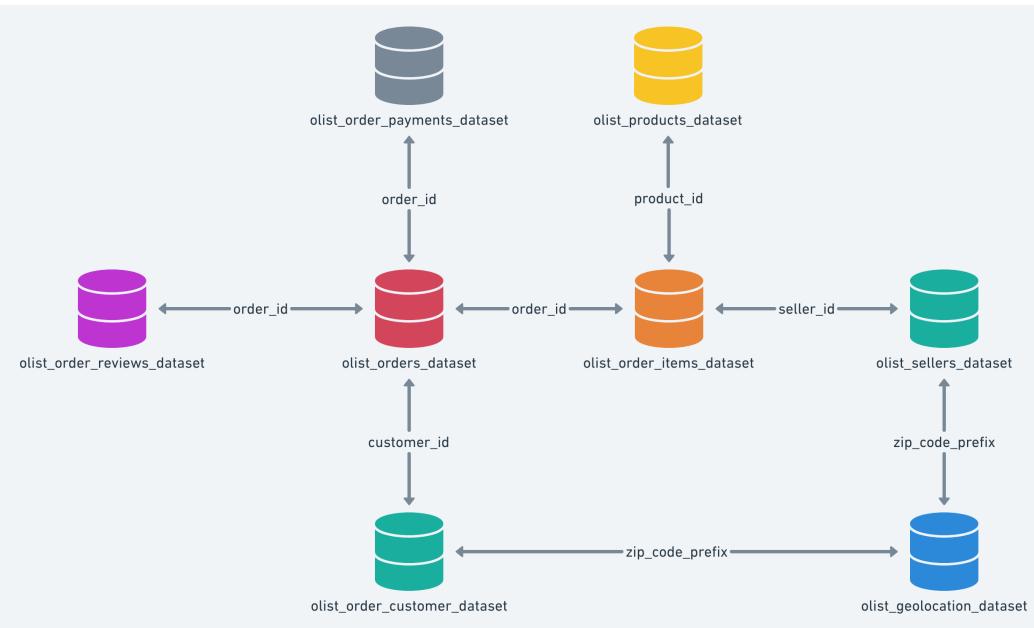
olist





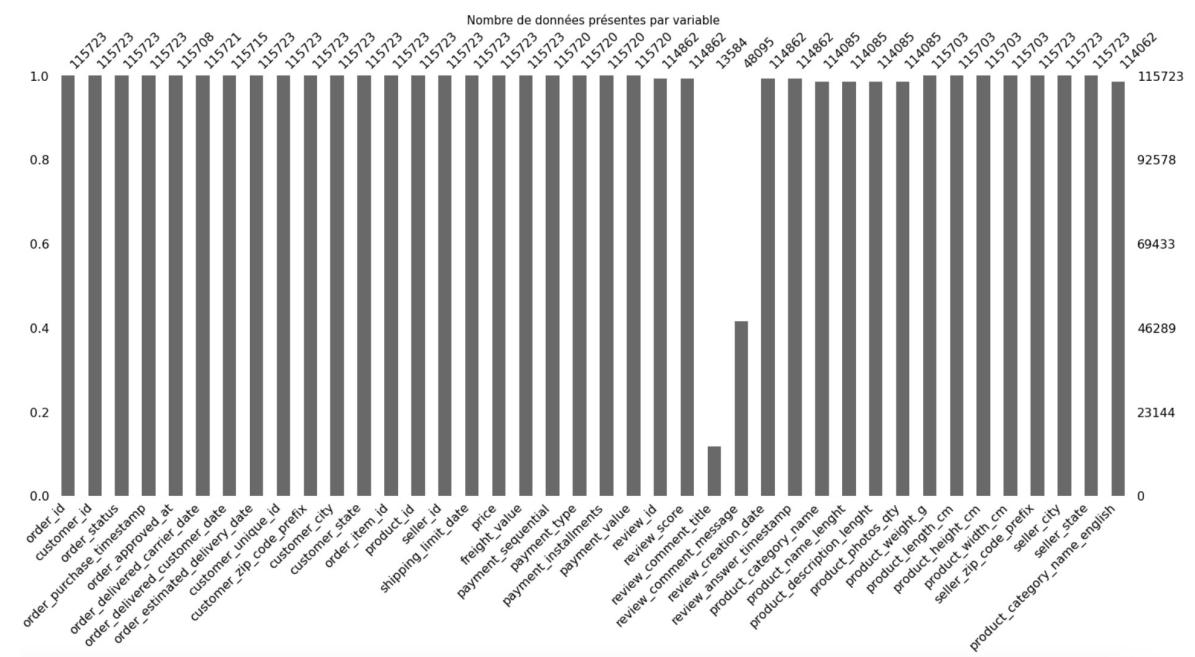
## II - Exploration des données

# Nettoyage des données



- Commandes livrées uniquement
- Peu de données manquantes

- 9 fichiers
- base de données anonymisée comportant des informations sur l'historique de commandes, les produits achetés, les commentaires de satisfaction, et la localisation des clients
- commandes de 2016 à 2018 effectuées sur plusieurs marchés au Brésil



# Feature Engineering



## Fréquence

- Temps depuis le dernier achat
- Temps depuis le premier achat



## Panier

- Nombre de commandes
- Nombre de produits
- Catégorie majoritaire



## Livraison

- Temps de livraison
- Retard de livraison



## Montants

- Panier moyen
- Panier total
- Frais de livraison moyen
- Frais de livraison total
- Prix moyen
- Prix total



## Localisation

- État du client
- Ville du client
- État du vendeur
- Ville du vendeur



## Satisfaction

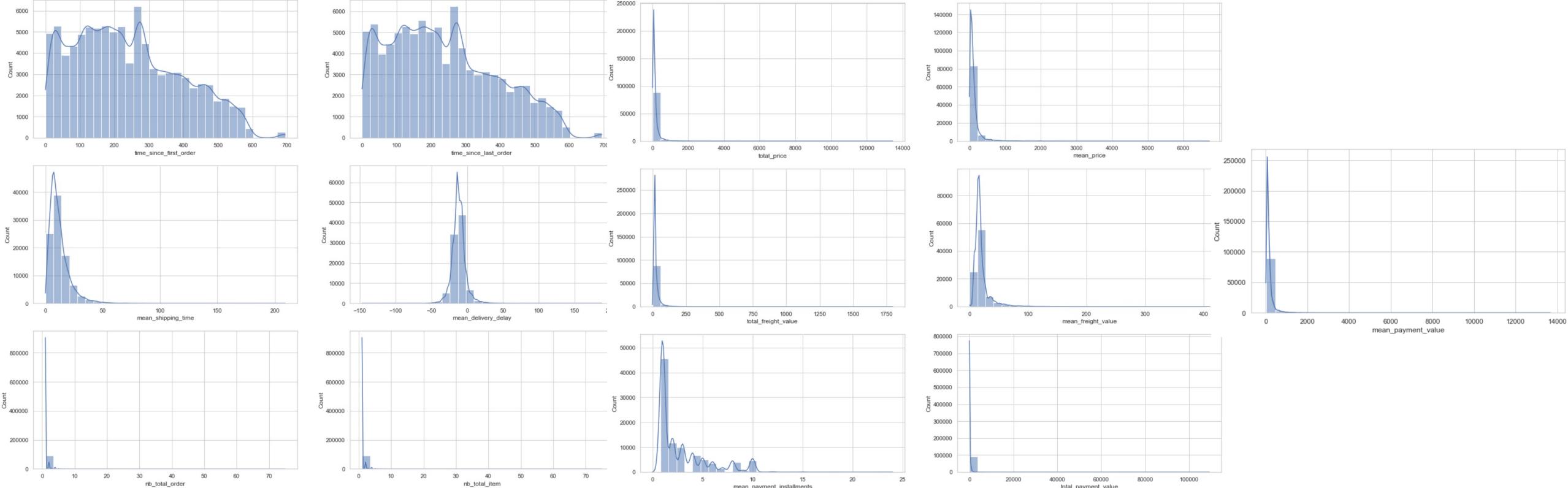
- Note moyenne attribuée



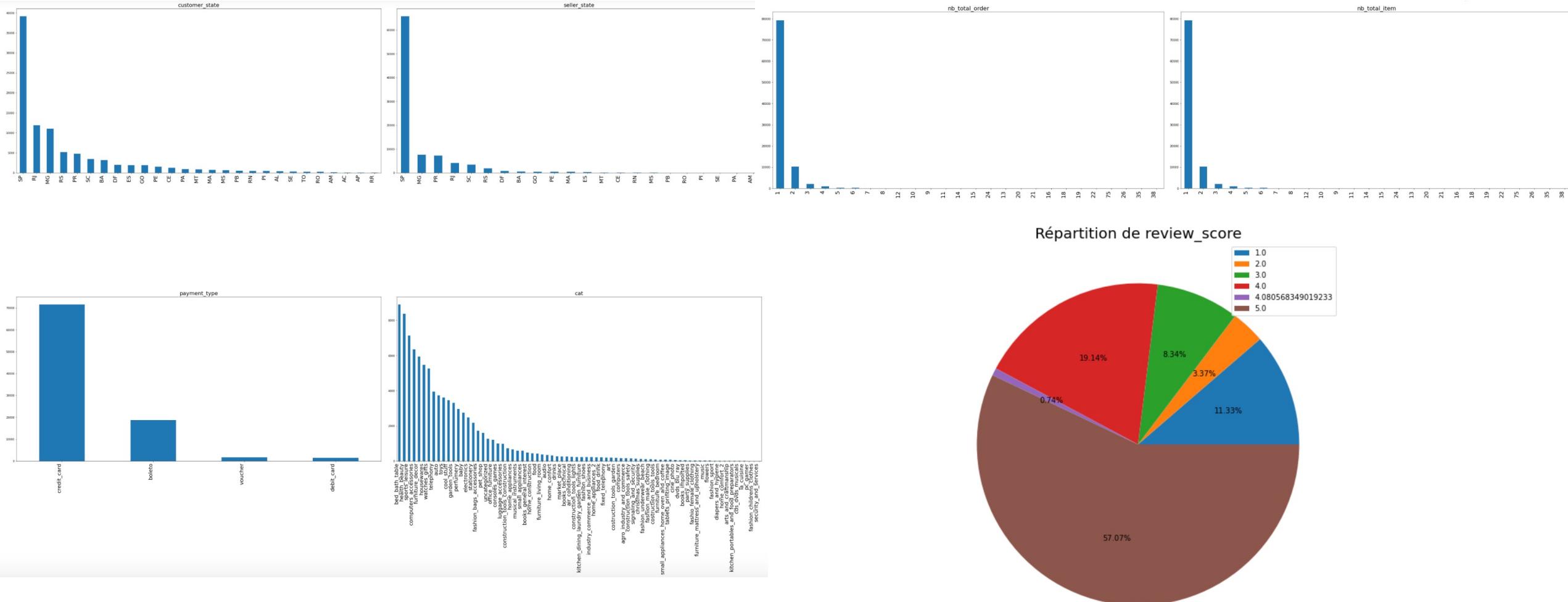
## Paiement

- Moyen de paiement
- Nombre d'échéances

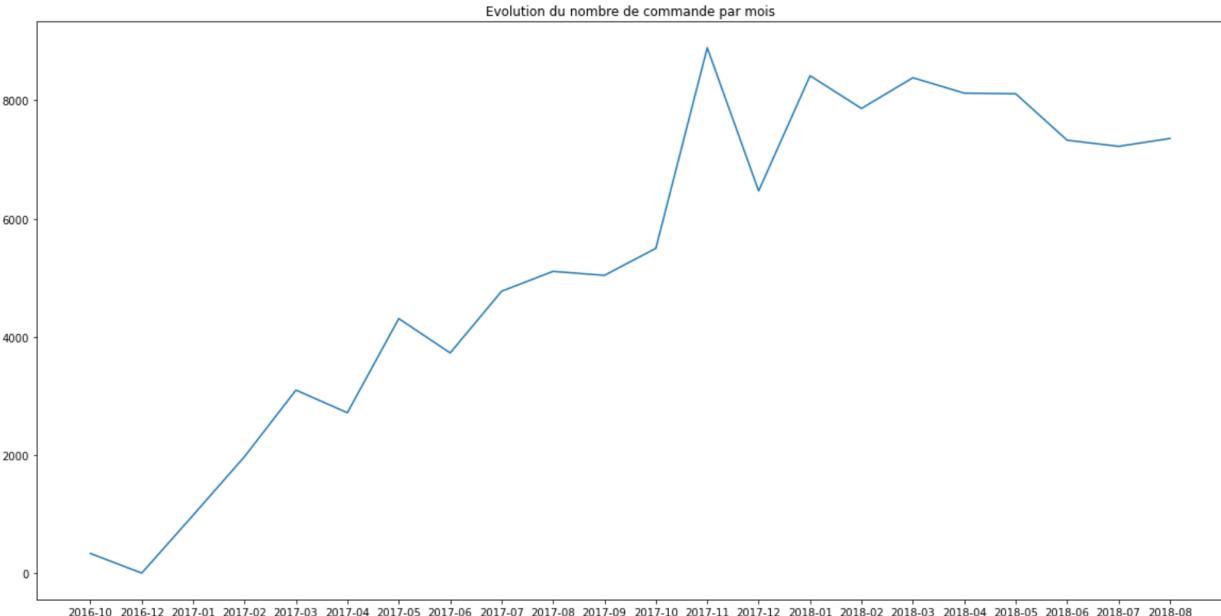
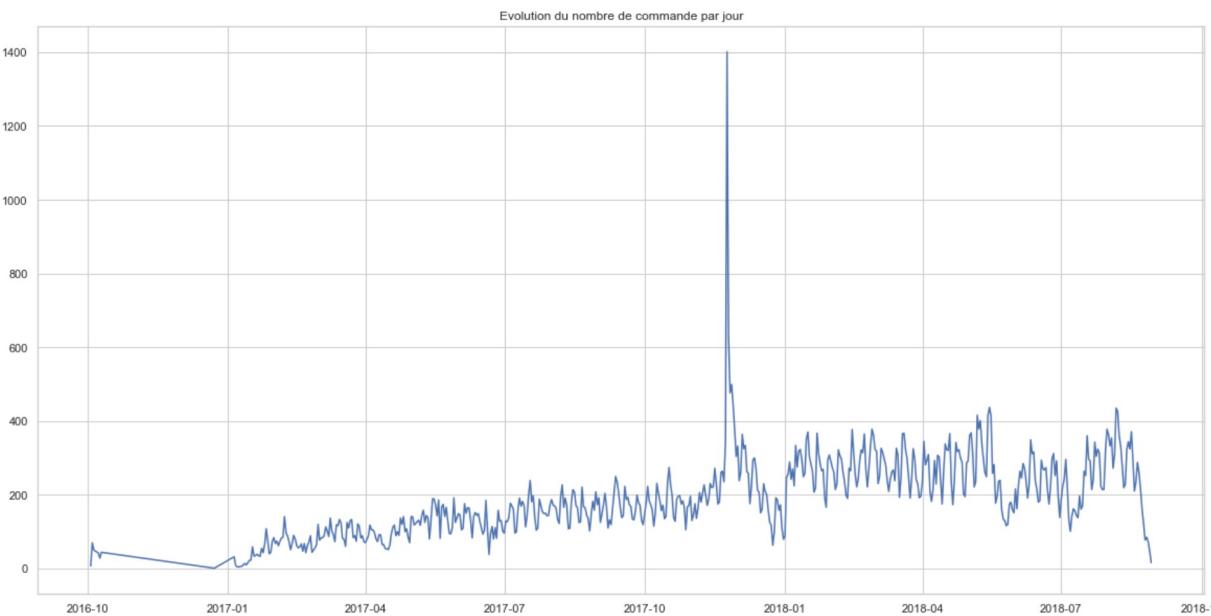
# Analyse exploratoire



# Analyse exploratoire



# Analyse exploratoire





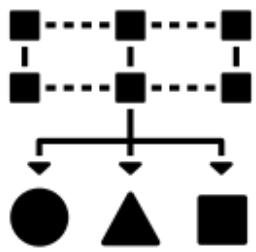
### III – Pistes de modélisation

# Processus



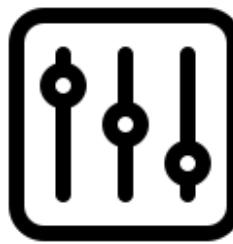
## Preprocessing

*MinMaxScaler*  
*OneHotEncoder*



## Classification non-supervisée

*Kmeans*  
*DBSCAN*  
*CAH (Classification Ascendante Hierarchique)*



## Choix des paramètres

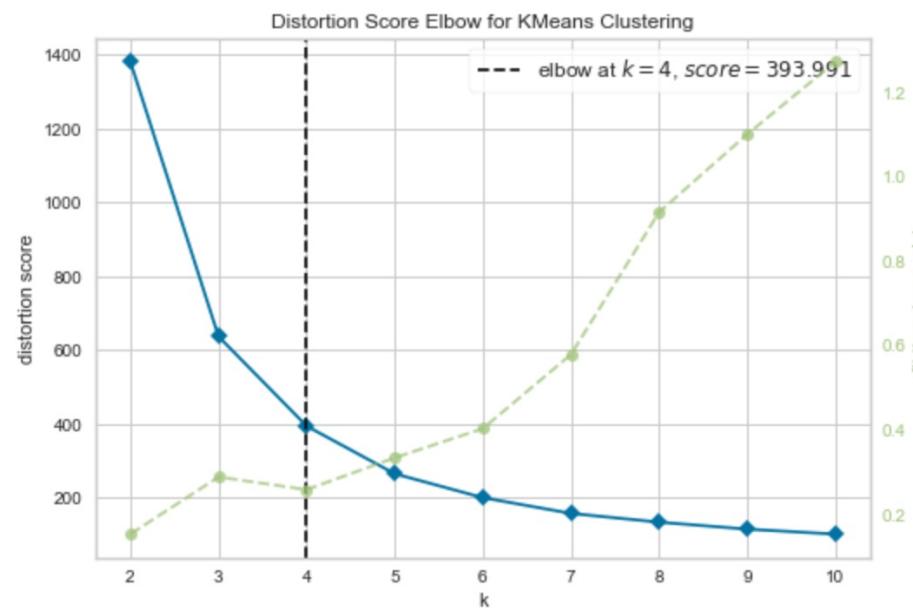


## Analyse des clusters

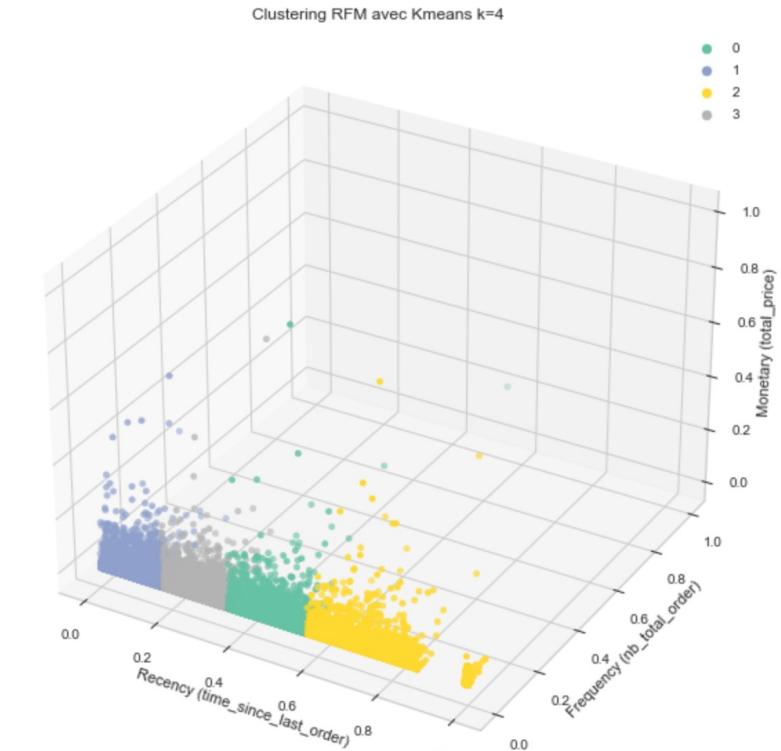
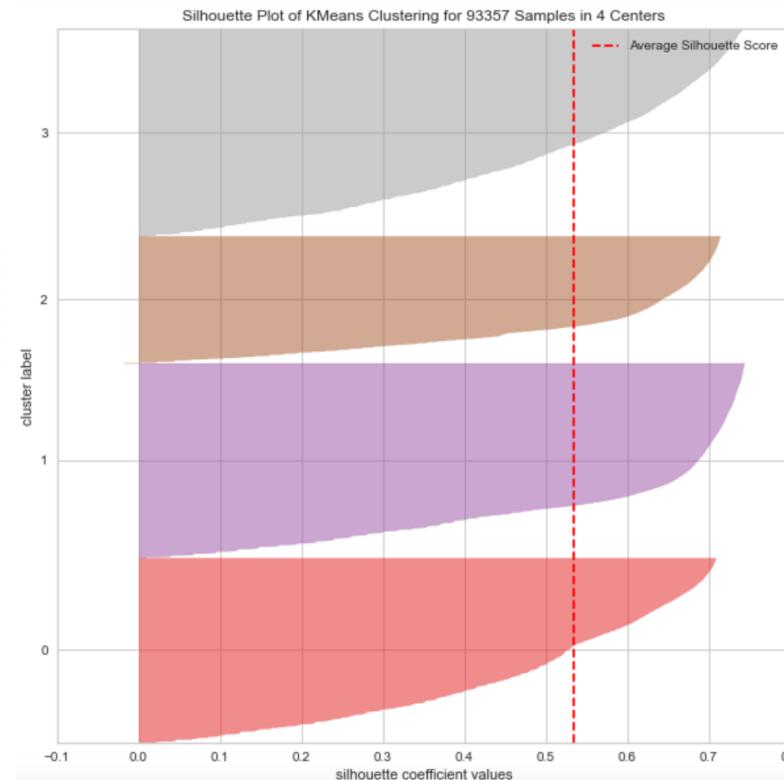
# Segmentation RFM : Kmeans

**R**ecency  
**F**requency  
**M**onetary

- **Silhouette Score** : similarité d'une observation avec son propre cluster par rapport aux autres clusters.
- **Davies Bouldin Score** : mesure de similarité moyenne de chaque cluster avec son cluster le plus similaire.
- **Calinski Harabasz Score** : rapport de la somme de la dispersion inter-cluster et de la somme de la dispersion intra-cluster pour tous les clusters.

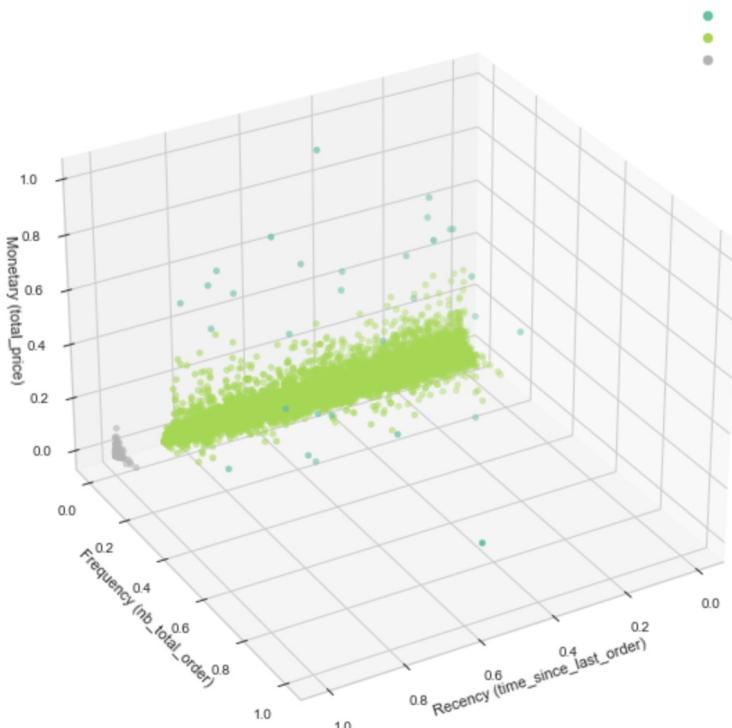


Modèle kmeans  $k=4$   
Silhouette Score : 0.5336532966419448  
Davies Bouldin Score : 0.5502103932434423  
Calinski Harabasz Score : 328732.63378289214



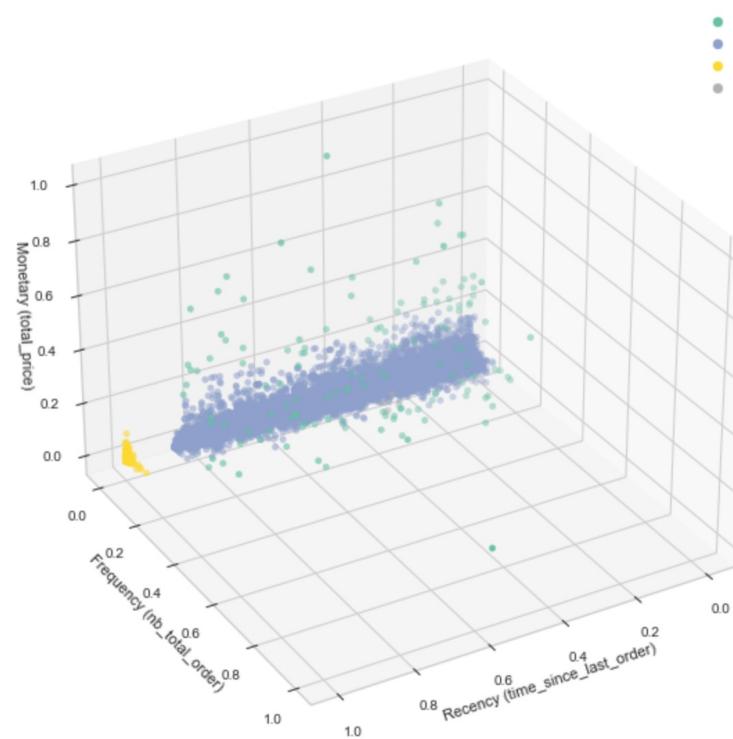
# Segmentation RFM : DBSCAN

Clustering RFM avec DBSCAN (eps:0.1, min\_samples=6)



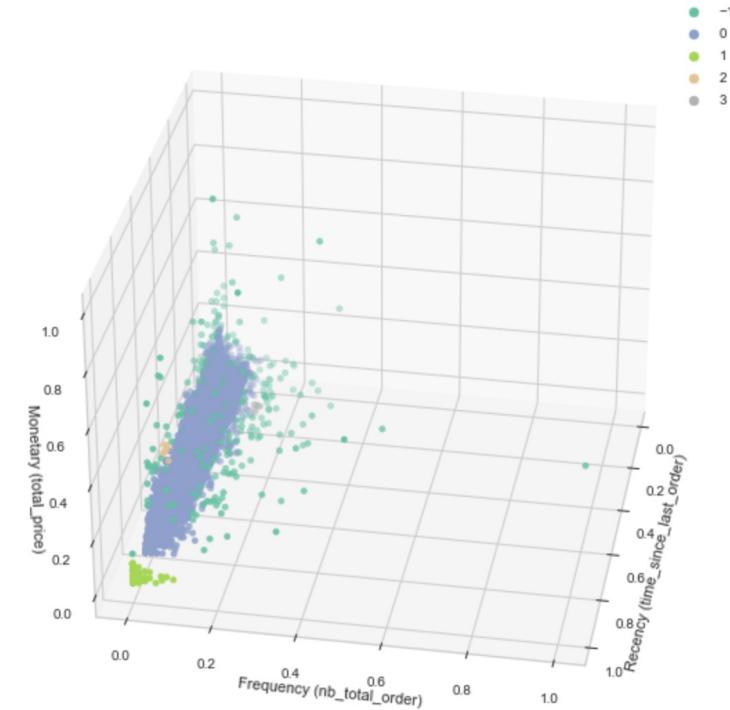
Paramètres du modèle : epsilon : 0.1, min\_samples : 6.  
Nombre de clusters avec le modèle DBSCAN : 2  
Nombre de points considérés comme du bruit: 31  
Silhouette Score : 0.4876415604531913  
Davies Bouldin Score : 1.0054772953103024  
Calinski Harabasz Score : 1207.8292213149625

Clustering RFM avec DBSCAN (eps:0.04, min\_samples=6)



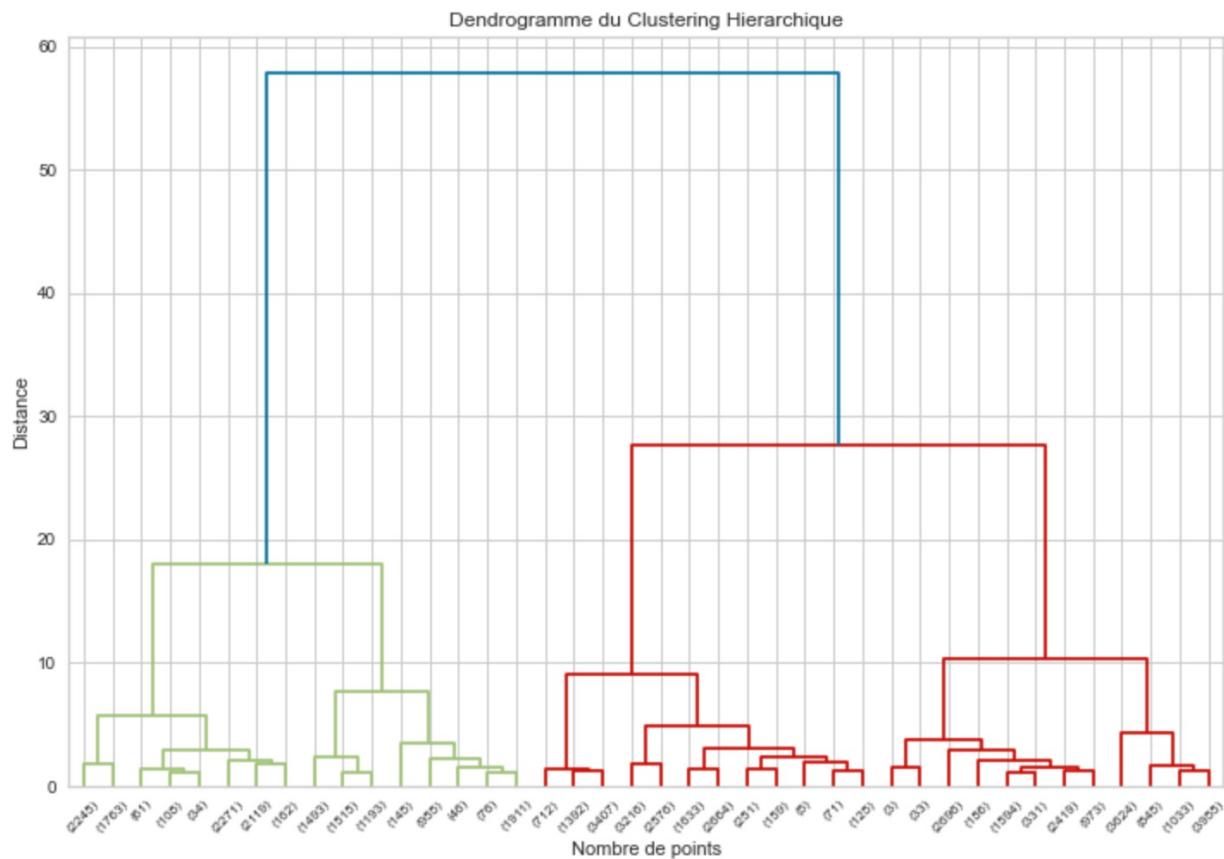
Paramètres du modèle : epsilon : 0.04, min\_samples : 6.  
Nombre de clusters avec le modèle DBSCAN : 3  
Nombre de points considérés comme du bruit: 131  
Silhouette Score : -0.0499551510803133  
Davies Bouldin Score : 1.6064371395065011  
Calinski Harabasz Score : 819.2254768624456

Clustering RFM avec DBSCAN (eps:0.04, min\_samples=6)



Paramètres du modèle : epsilon : 0.03, min\_samples : 6.  
Nombre de clusters avec le modèle DBSCAN : 4  
Nombre de points considérés comme du bruit: 217  
Silhouette Score : -0.17824029183999743  
Davies Bouldin Score : 1.5521972301136293  
Calinski Harabasz Score : 623.1058652413222

# Segmentation RFM : CAH



Paramètres du modèle CAH: k=2.

Silhouette Score : 0.5836469986245769

Davies Bouldin Score : 0.5501983100929154

Calinski Harabasz Score : 105043.33081484096

---

Paramètres du modèle CAH: k=3.

Silhouette Score : 0.5254445608364084

Davies Bouldin Score : 0.5698238844178025

Calinski Harabasz Score : 124059.79588981035

Paramètres du modèle CAH: k=4.

Silhouette Score : 0.5118661755150412

Davies Bouldin Score : 0.5825808492069495

Calinski Harabasz Score : 147138.55790322978

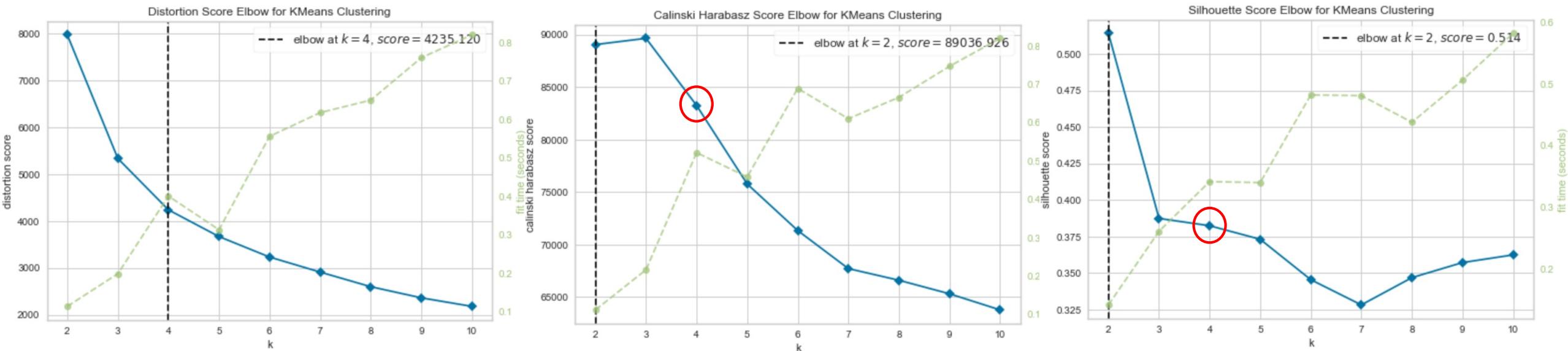
# Autres segmentations

Nom	Description	Nombre de clusters	Silhouette Score	Davies Bouldin Score	Calinski Harabasz Score	Avantages	Inconvénients
Segmentation RFM + score	Ajout score moyen attribué par le client	4	0.48	0.77	127 716.77	Simple à interpréter	Segmentation un peu trop simpliste
Segmentation RFM + score + livraison	Ajout du temps de livraison et du retard de livraison	4	0.46	0.81	117 208.52	Simple à interpréter	Retard de livraison ayant un score négatif sont considérées comme un retard
Segmentation avec les catégories	Ajout des catégories et d'autres variables	12	0.29	2.13	8 528.08	Prend en considération toutes les variables	Trop de clusters ; segmente en fonction de la catégorie de produit au dépend des autres variables



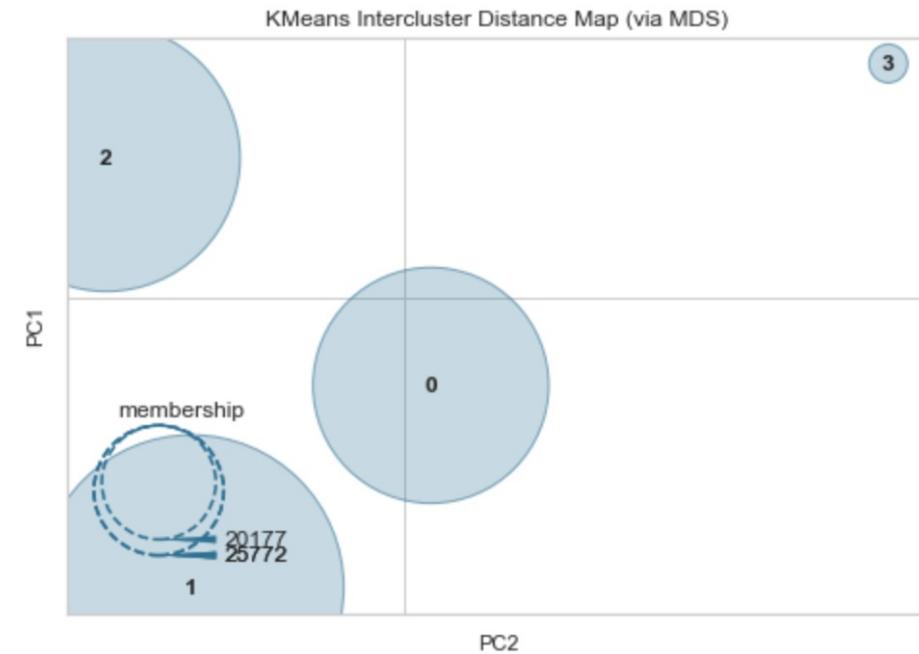
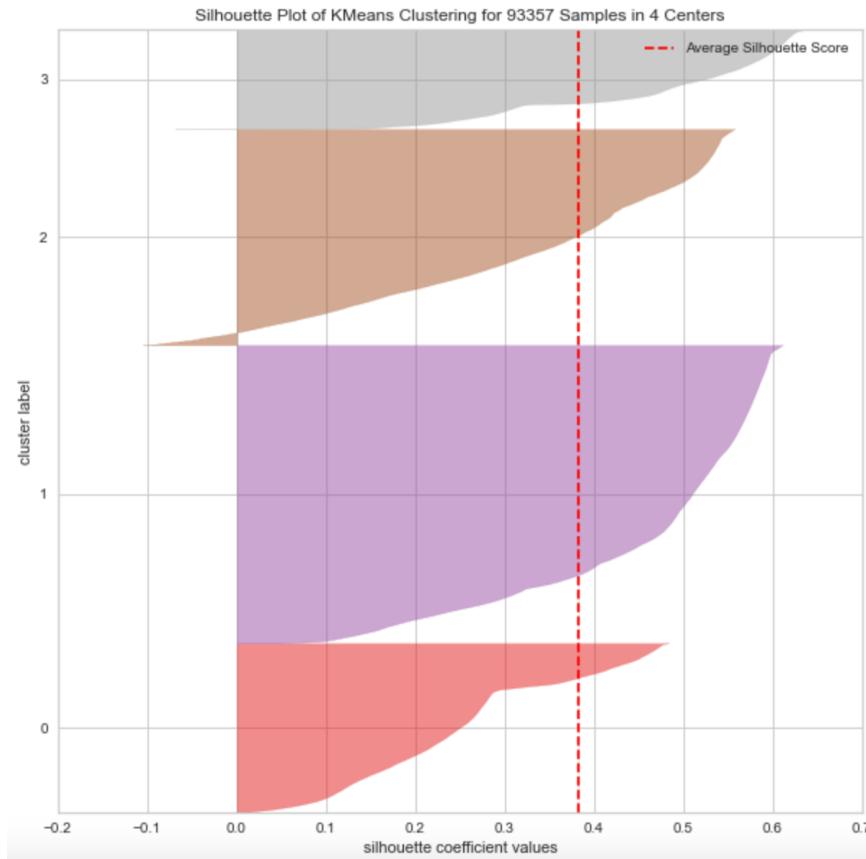
## IV - Modèle final

# Optimisation du nombre de clusters



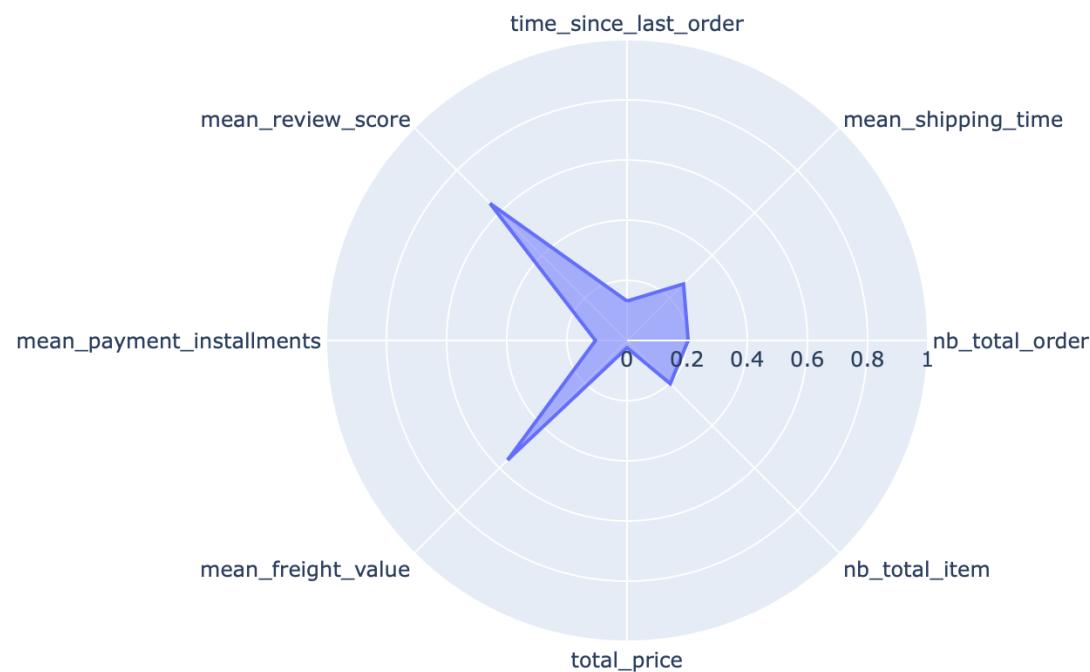
Scores pour le modèle kmeans (k=4):  
Silhouette Score : 0.3822852286421737  
Davies Bouldin Score : 0.9653031788087589  
Calinski Harabasz Score : 83242.24580047058

# Visualisation des clusters

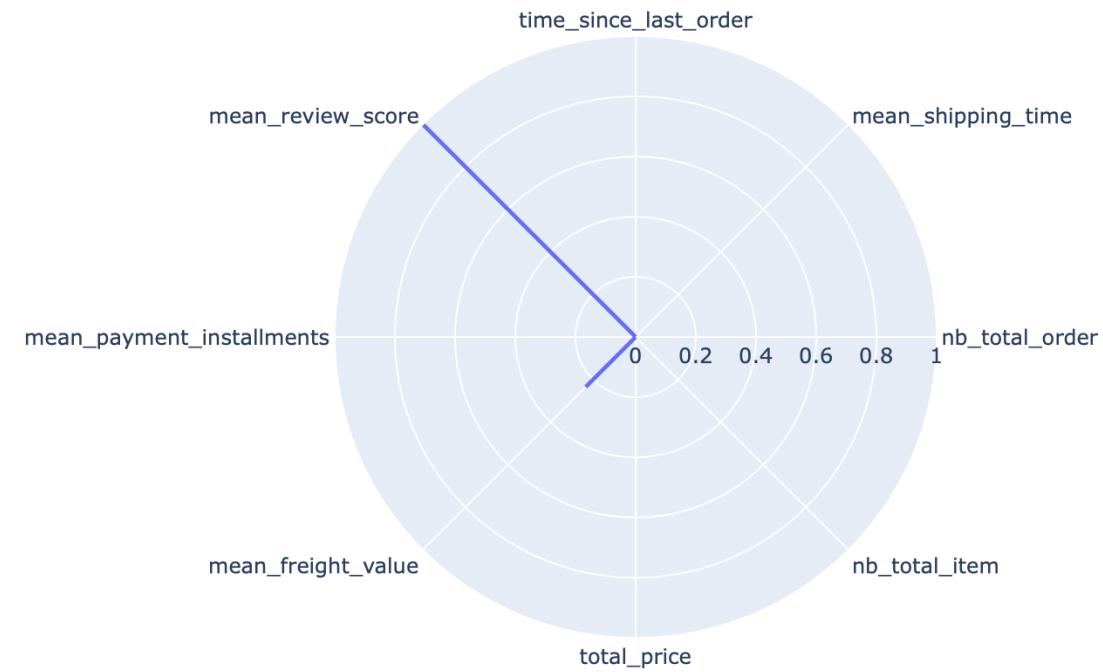


# Interprétation des clusters

Cluster 0

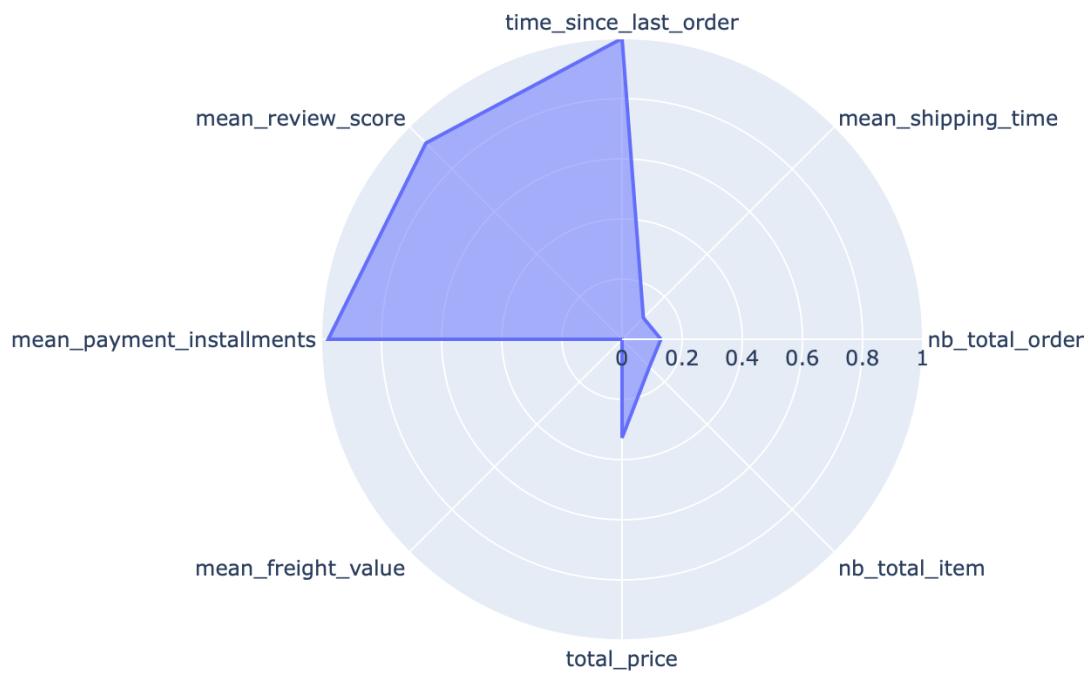


Cluster 1

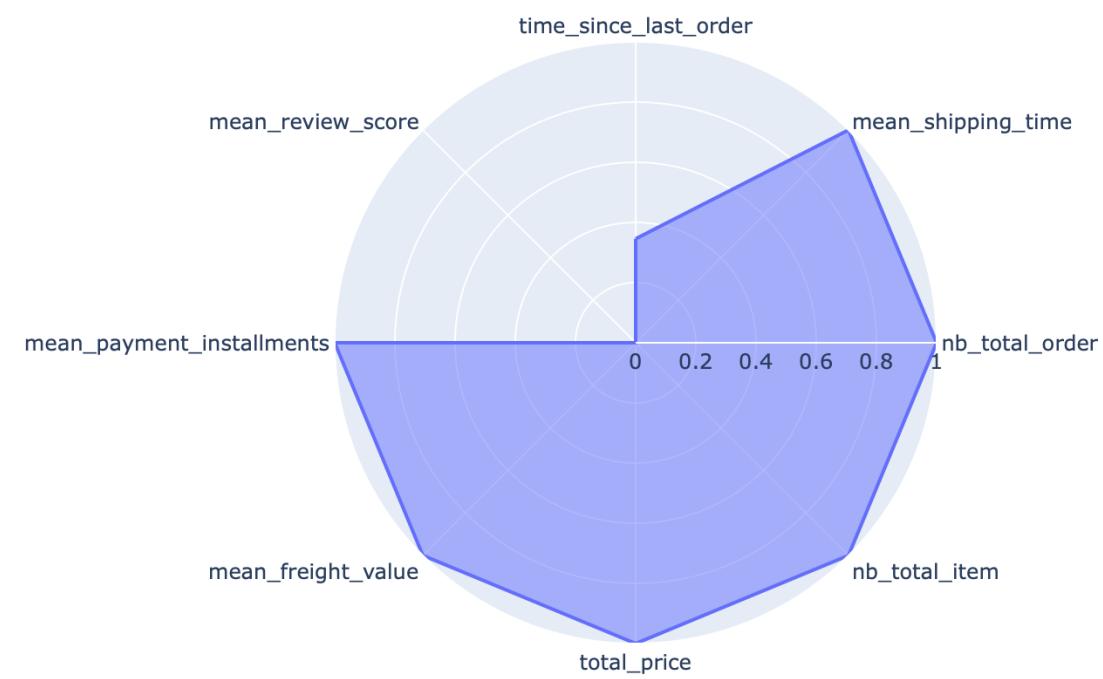


# Interprétation des clusters

Cluster 2



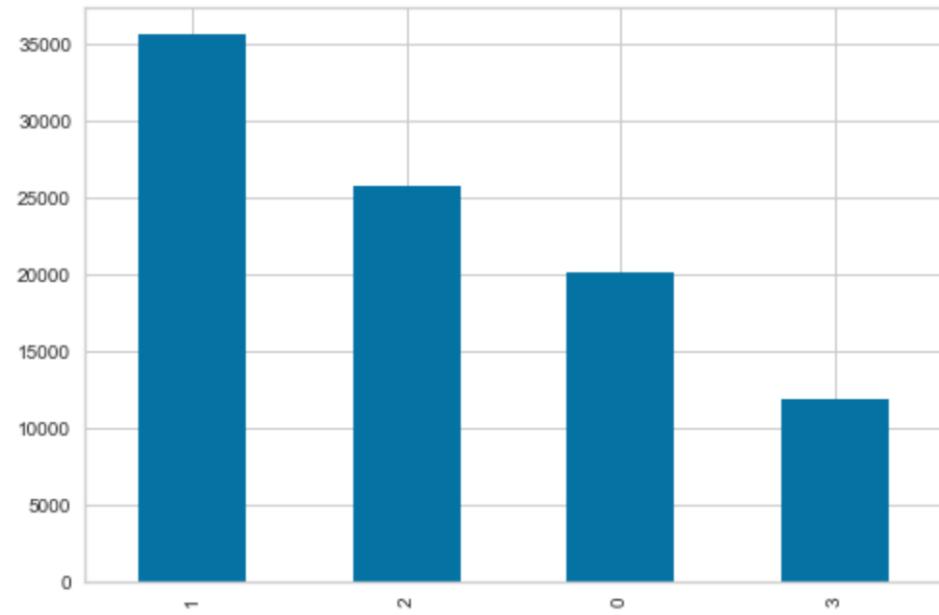
Cluster 3



# Récapitulatif des clusters

cluster	time_since_last_order	mean_shipping_time	nb_total_order	nb_total_item	total_price	mean_freight_value	mean_payment_installments	mean_review_score
0	0.131111	0.265677	0.203127	0.203127	0.024149	0.561505	0.105070	0.644243
1	0.000000	0.000000	0.000000	0.000000	0.000000	0.233962	0.000000	1.000000
2	1.000000	0.100375	0.129243	0.129243	0.327792	0.000000	0.977022	0.922357
3	0.345613	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000

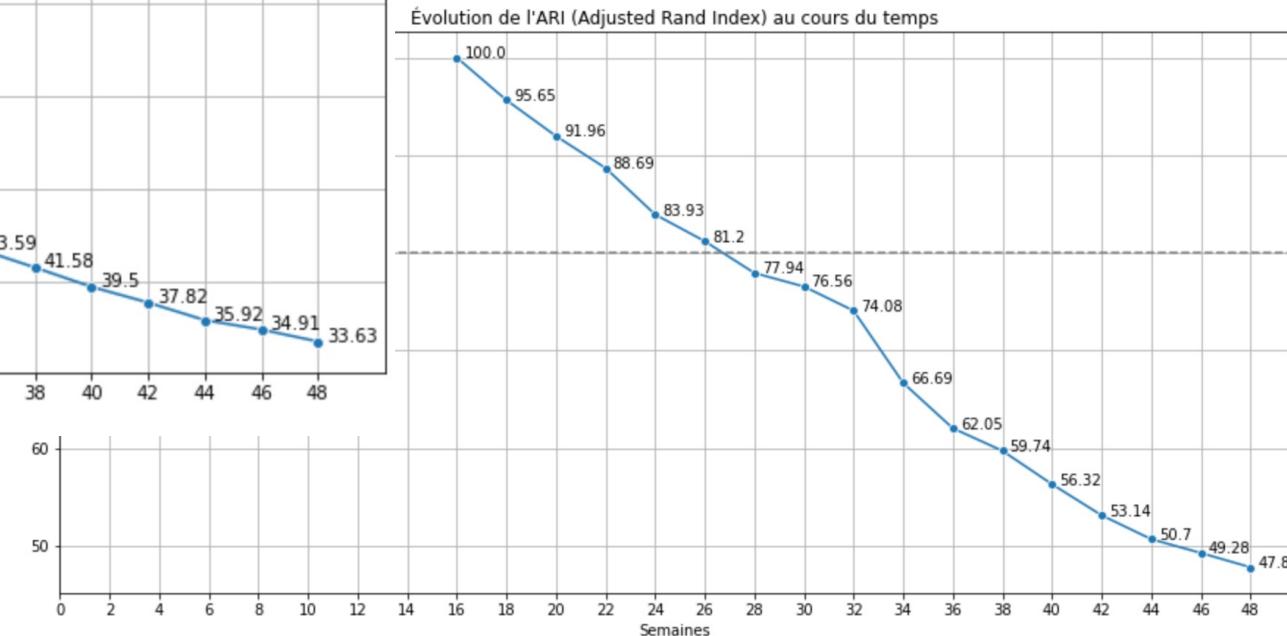
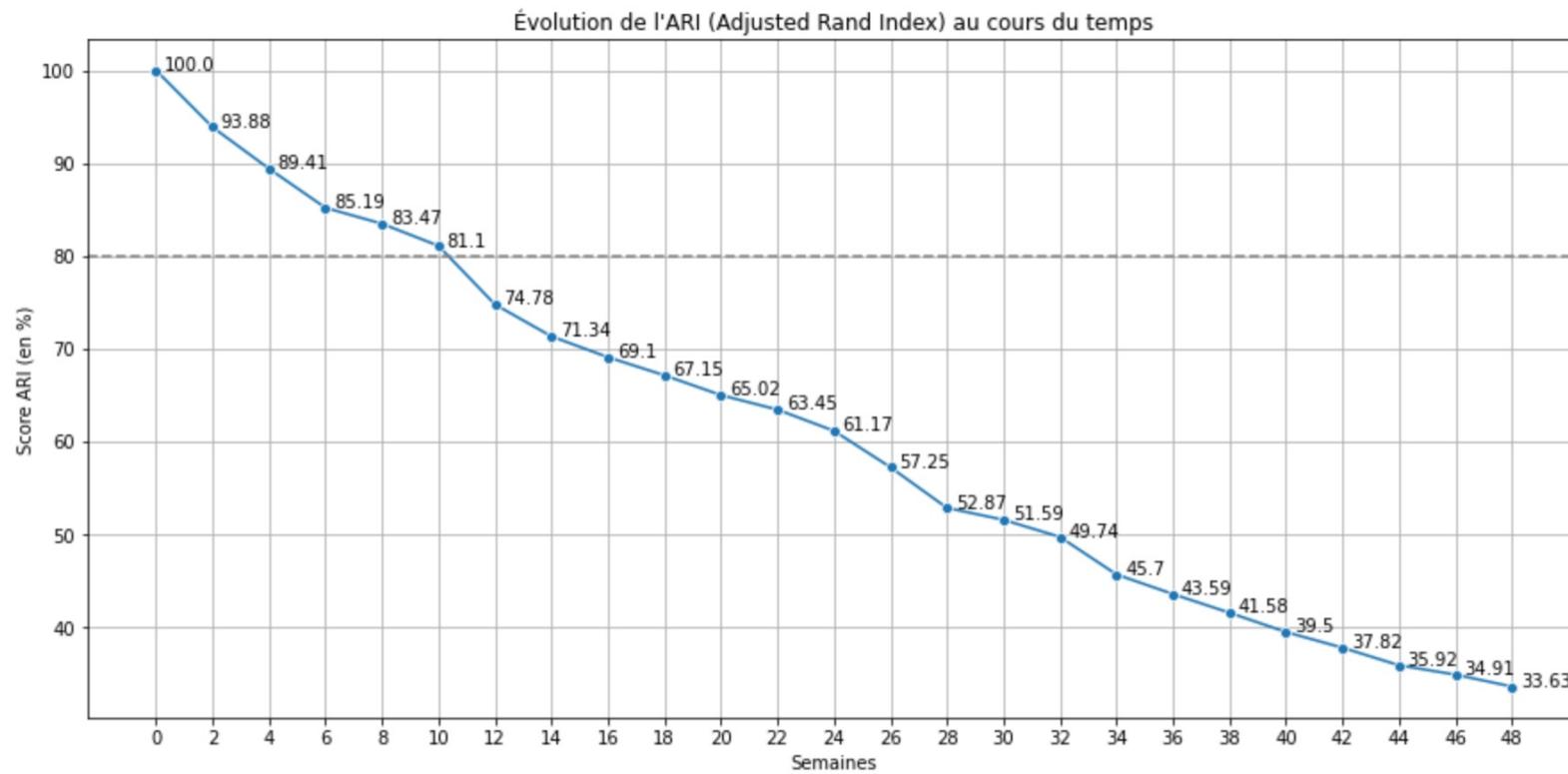
cluster	time_since_last_order	mean_shipping_time	nb_total_order	nb_total_item	total_price	mean_freight_value	mean_payment_installments	mean_review_score
0	175.652525	12.626059	1.238836	1.238836	142.566661	20.634463	2.764486	3.662927
1	138.638428	10.035949	1.193575	1.193575	141.827300	19.977367	2.719053	4.996697
2	420.950101	11.014512	1.222373	1.222373	151.863314	19.508007	3.141527	4.705604
3	236.208960	19.785038	1.416399	1.416399	172.444314	21.514145	3.151463	1.247592





# V - Maintenance

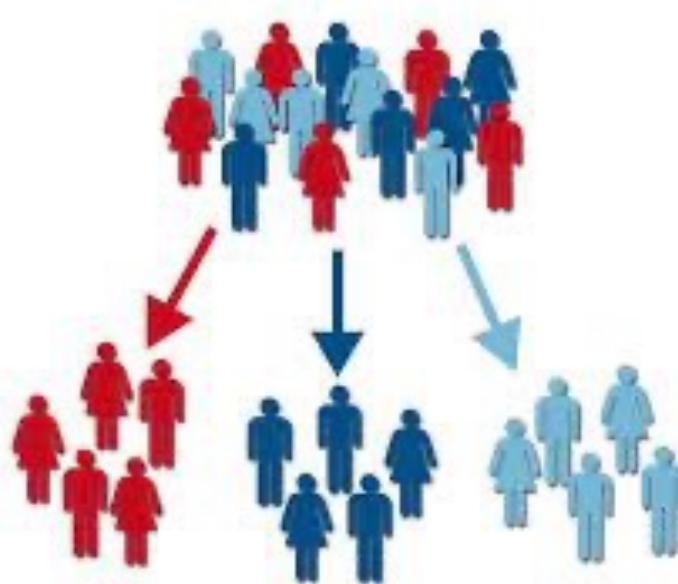
# Stabilité du clustering en fonction du temps



# Conclusion

## Variables d'entrée :

- le temps depuis la dernière commande (recence)
- le nombre total de commande (frequence)
- le nombre total d'article
- le prix total du panier (monetary)
- le montant moyen de frais de port
- le temps de livraison
- le nombre moyen d'échéances de paiement mis en place
- le score moyen attribué par les clients



- Kmeans avec  $k = 4$
- 4 clusters interprétables
- Mise à jour tous les 2.5 mois



Merci de votre attention