Projet P5

Segmentation clients d'un site de e-commerce

OPENCLASSROOMS

Introduction

- Consultant pour Olist -> une solution de vente sur les marketplaces en ligne
- Objectifs :

  - Segmentation des clients pour les équipes marketing de Olist pour les campagnes de communication

  - Analyser les différents types d'utilisateurs

  - Proposition du contrat de maintenance

Données

On a 8 jeux de données :

- **customers** : information sur le clients de chaques commandes
- **geolocation** : information sur localisation
- **orders** : information sur la commandes (date,etc…)
- **orders_items** : info des produits de chaques commandes
- **order_payments** : info sur le paiement
- **order_reviews** : info sur les commentaires de chaque commandes
- **products** : info sur les produits (taille,poids, etc…)
- **sellers** : info sur les vendeurs

Sommaire

1. Nettoyage des données
2. Feature Engineering
3. Analyse exploratoire
4. Modélisation
5. Maintenance
6. Conclusion

Nettoyage des données

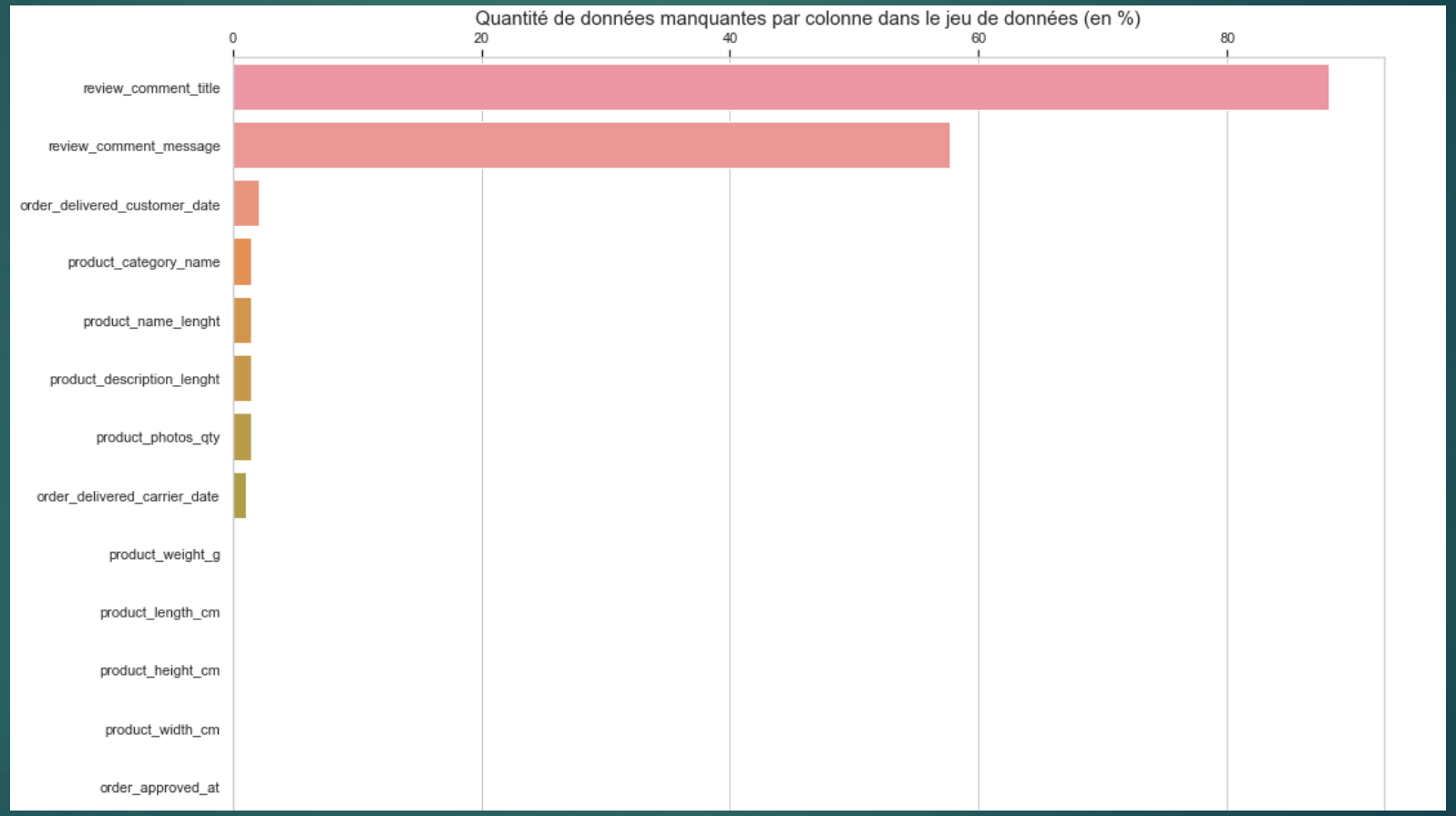
Concaténation des datasets en fonction des commandes passé

Taille datasets : 117329 lignes et 36 colonnes

Une ligne represente une commande passé

Nettoyage des données

Suppression des colonnes inutiles



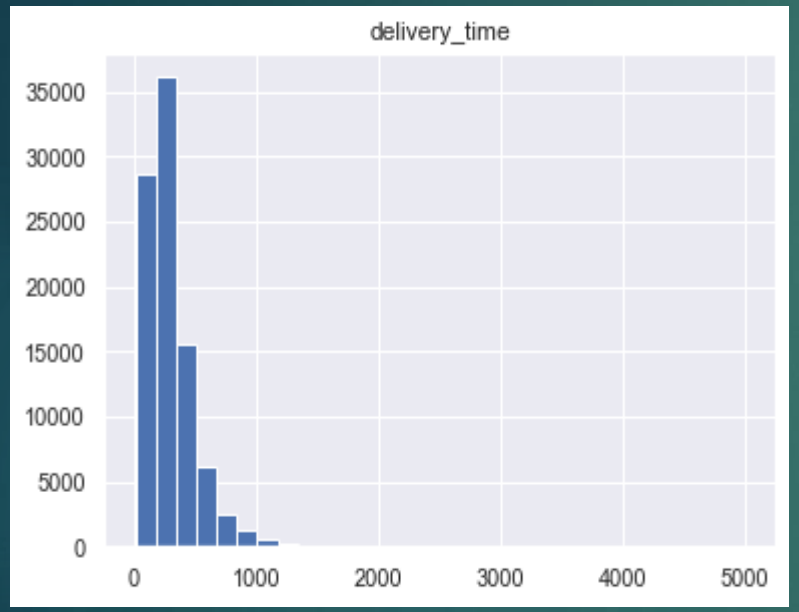
Taille du dataset : 117329 lignes et 24 colonnes

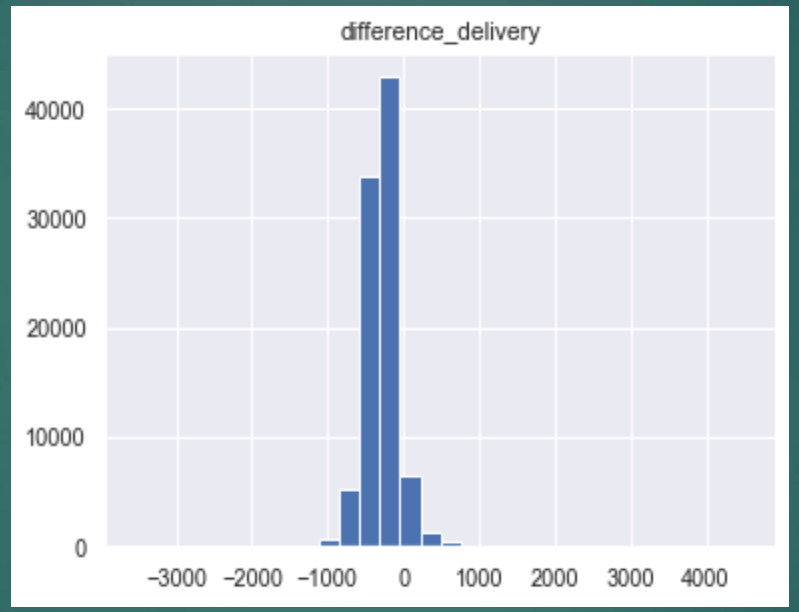
Feature Enginnering

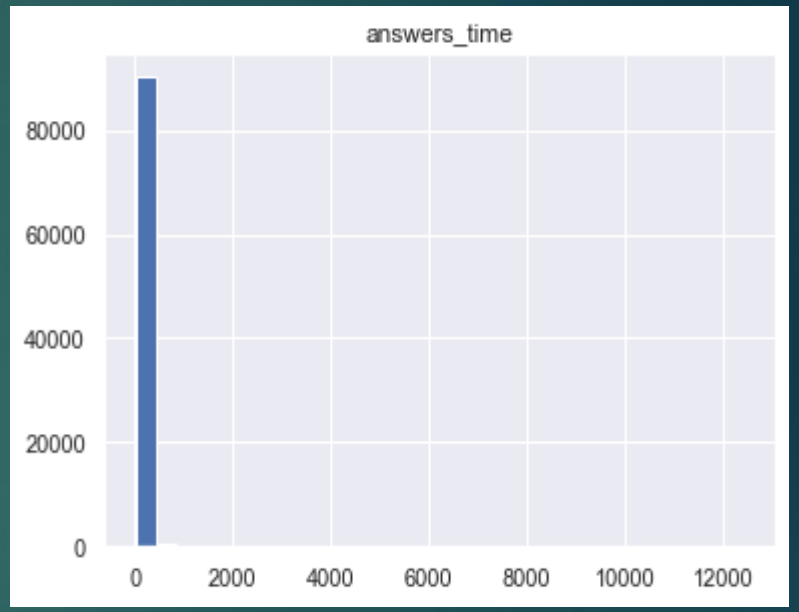
- Regrouper les commandes d'un même clients
- Création de colonnes pour chaque type de paiements
- Création de nouvelles variables :
  - Difference delivery
  - Delivery time
  - Answers time
  - reviews lenght
  - Recency
  - Frequency
  - Monetary
- Regroupements de catégories en 9 supercatégories et création de colonnes pour chacune de ces supercatégories

Taille du Dataset : 91453 lignes et 34 colonnes

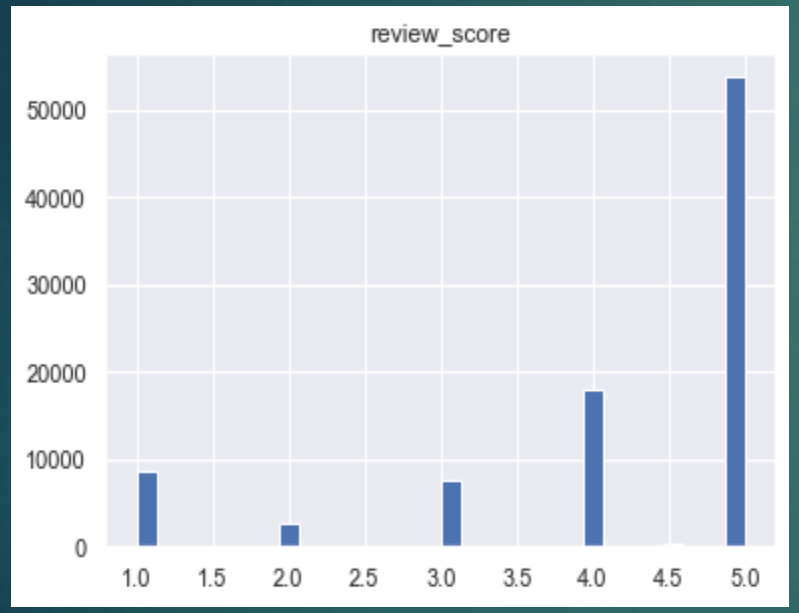
Analyse Exploratoire

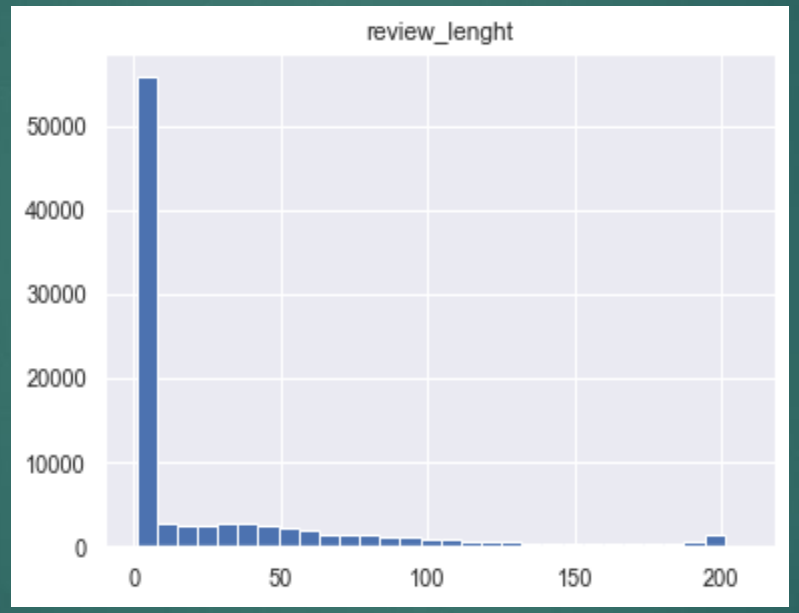


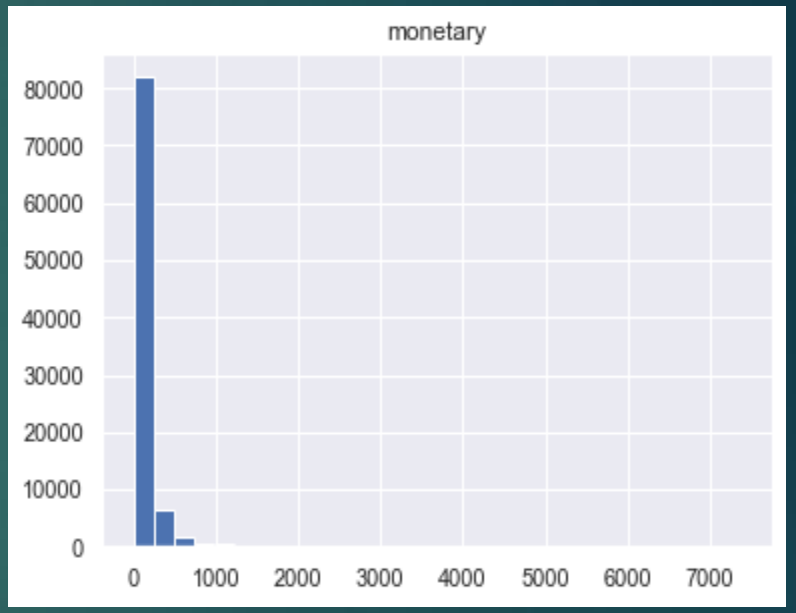




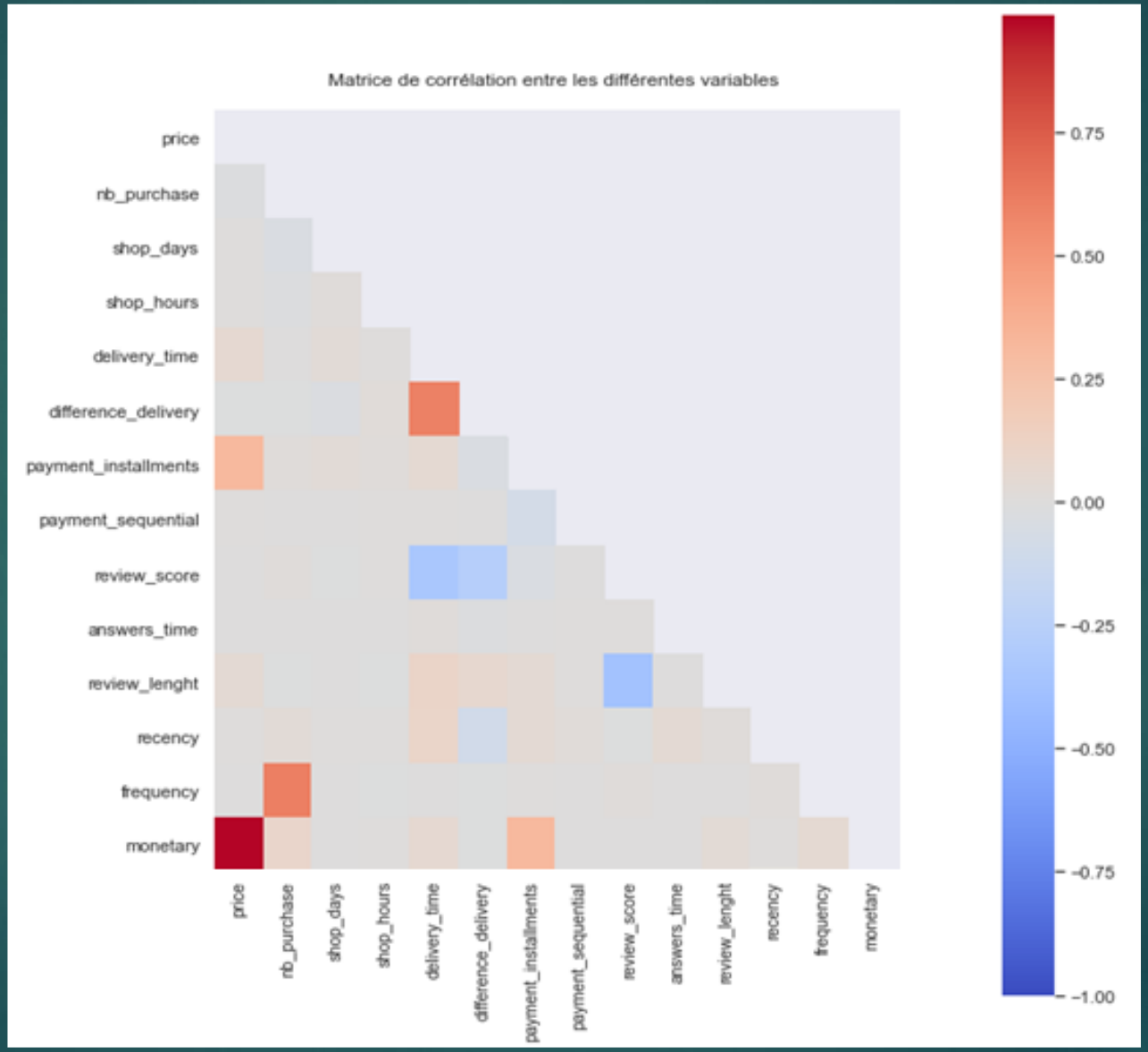
Analyse Exploratoire







Analyse Exploratoire

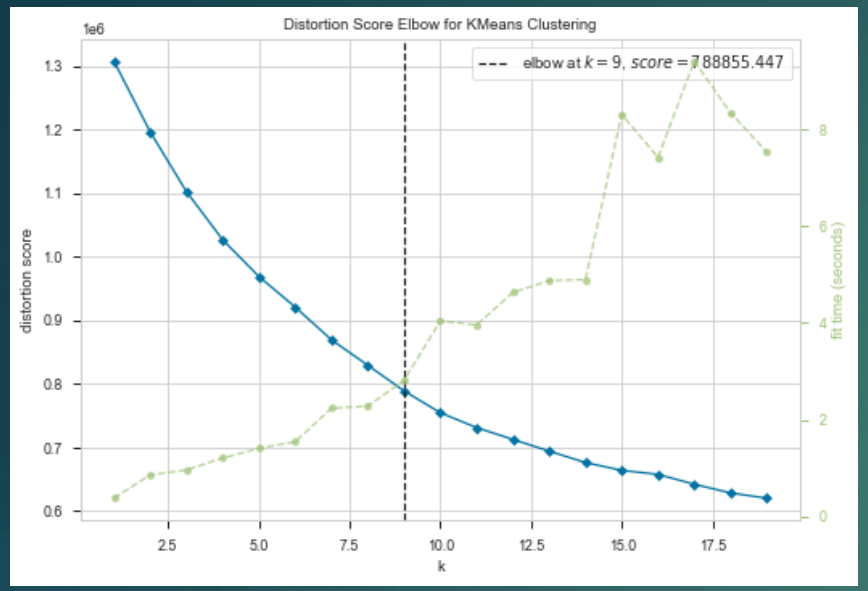


Modélisation

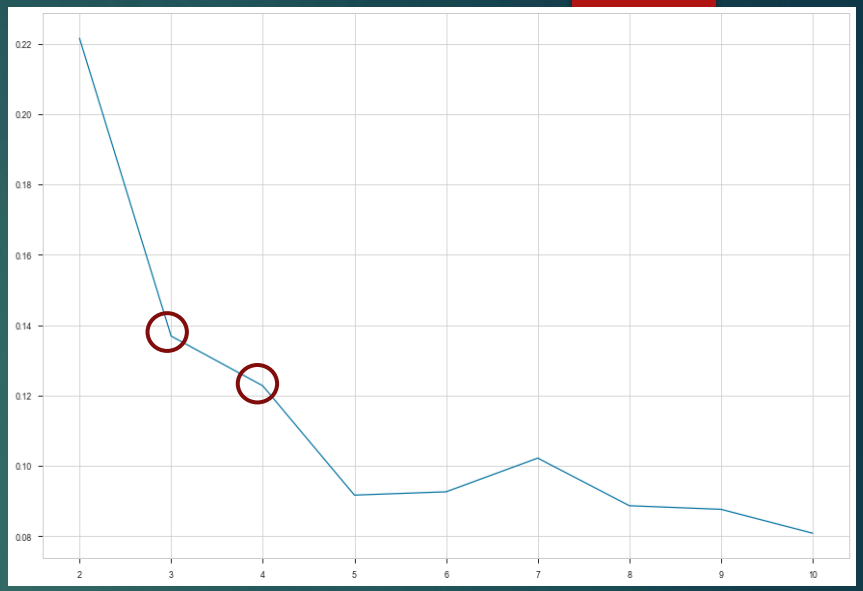
- Normalisation des données numérique
- OneHotEncoder pour la variable catégorielles customer_state
- Réduction de dimension par ACP à 19 variables
- 3 algorithme de clustering utilisé :
  - Kmeans
  - DBScan
  - Clustering hiérarchique
- Comparaison des 3 algorithmes avec :
  - Silhouette
  - indice de David Bouldin
  - distortion (pour Kmeans et CH uniquement)
- Visualisation par t-SNE

Taille dataset : 91453 ligne x 53 colonnes

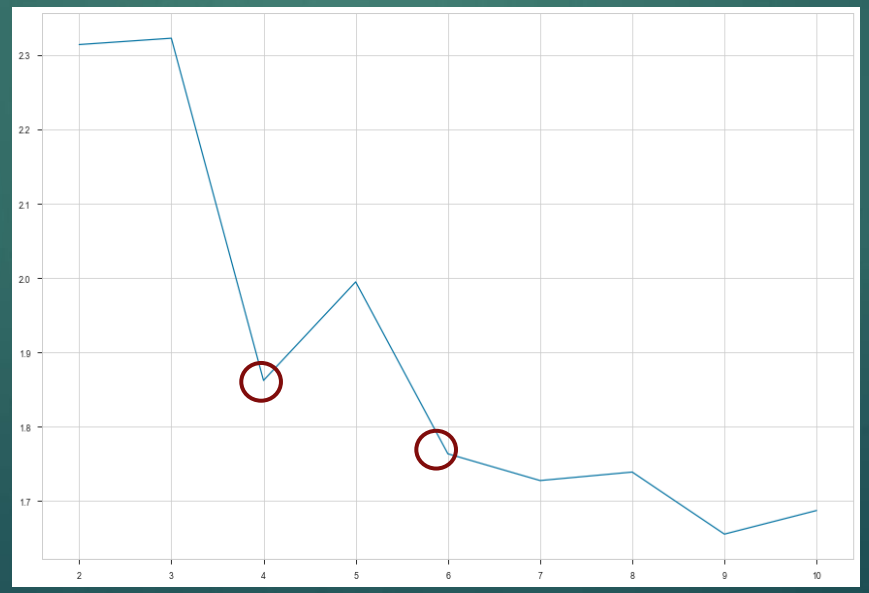
Modelisation avec Kmeans



Distorsion



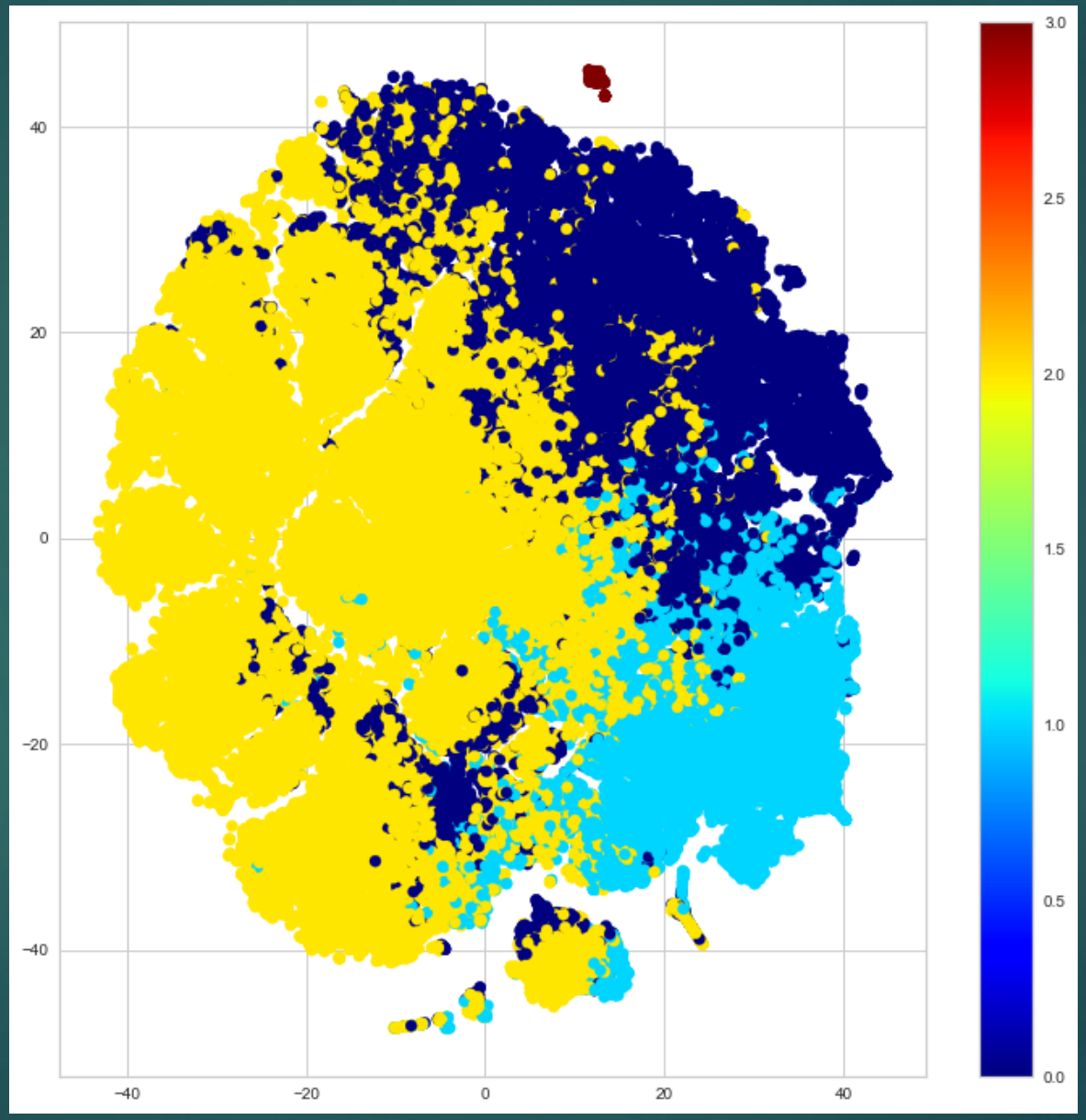
Silhouette



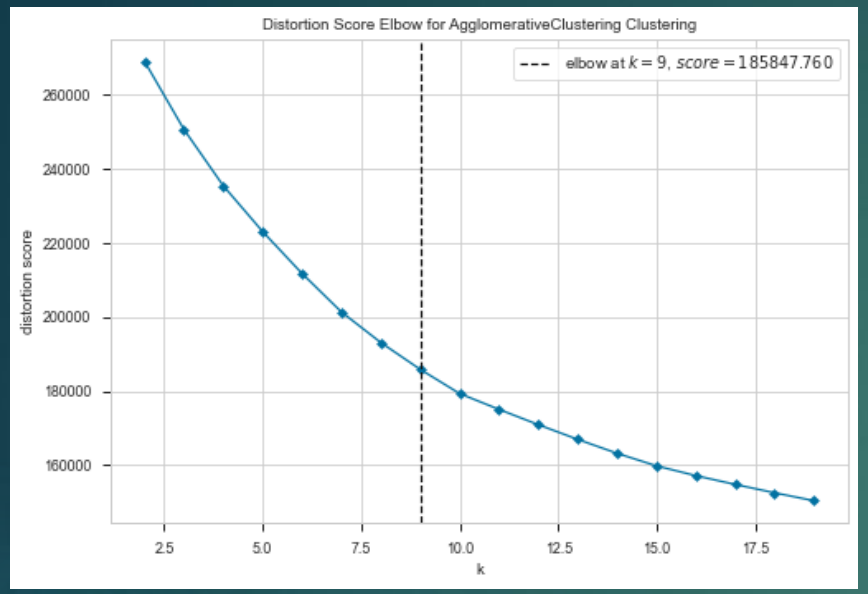
Indice David Bouldin

Nombre de clusters choisi : 4

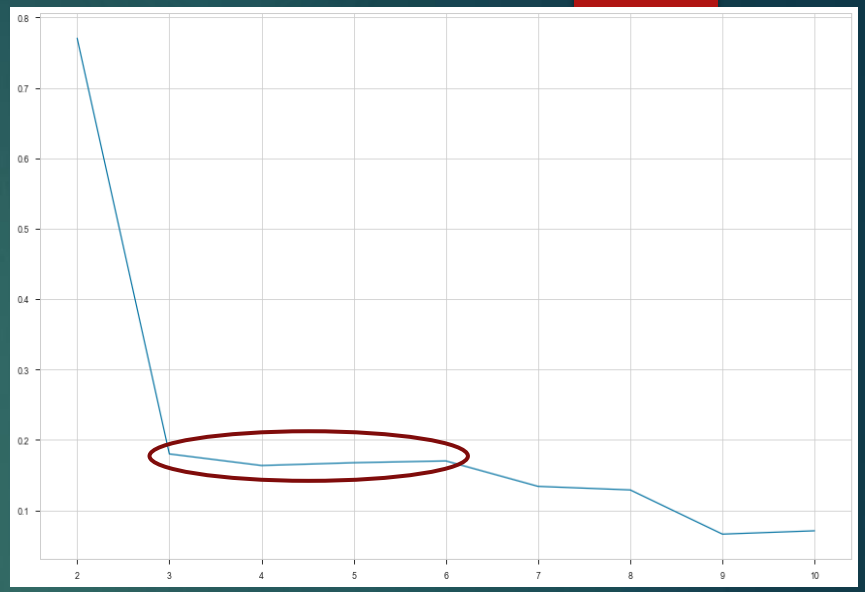
Modelisation avec Kmeans



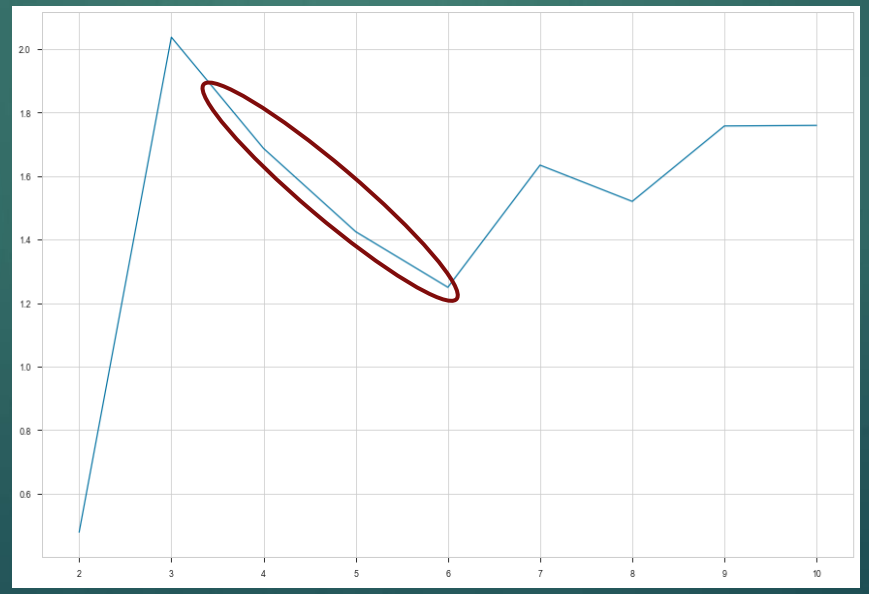
Modelisation avec Clustering Hierachique



Distorsion



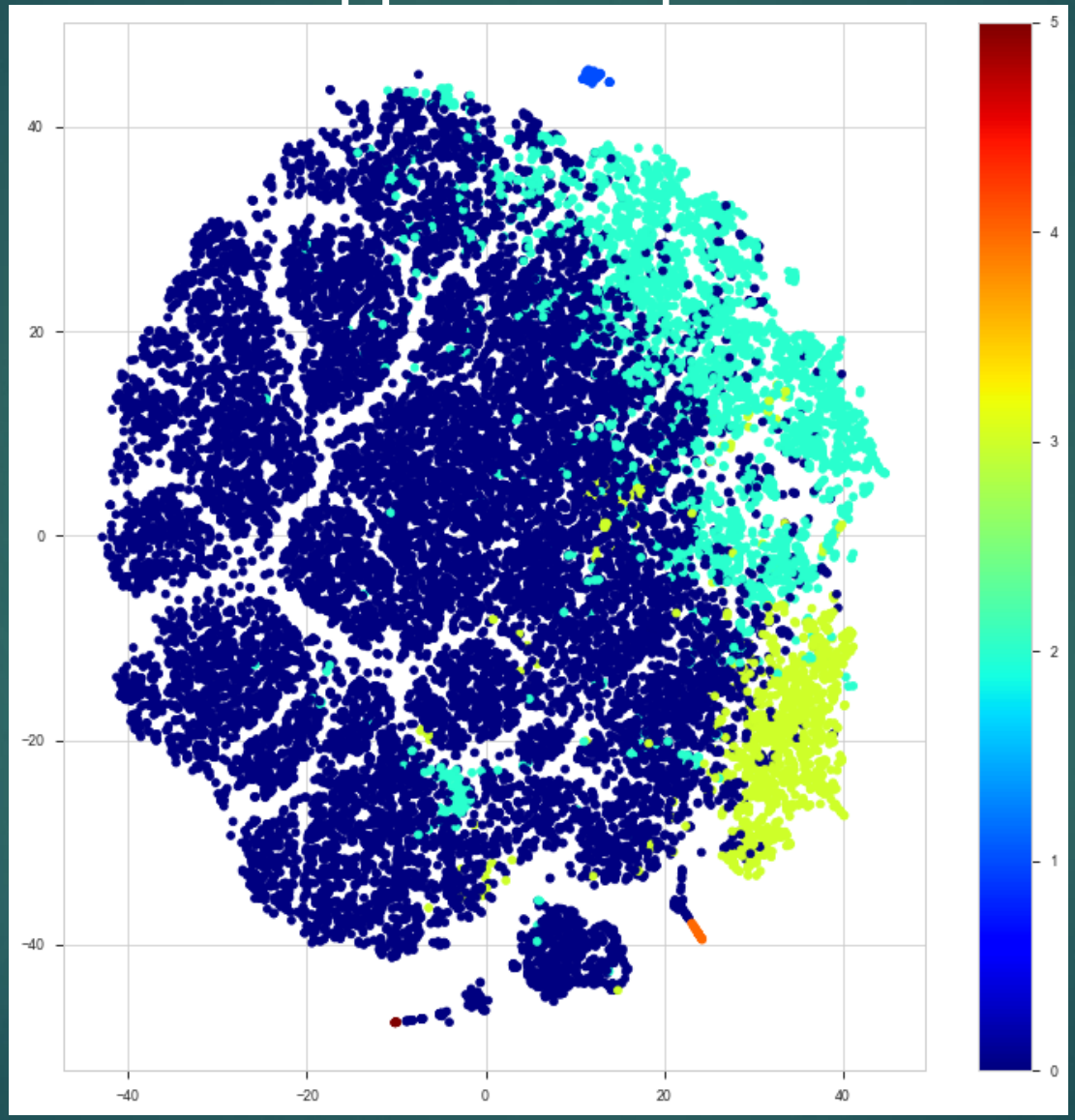
Silhouette



Indice David Bouldin

Nombre de clusters choisi : 6

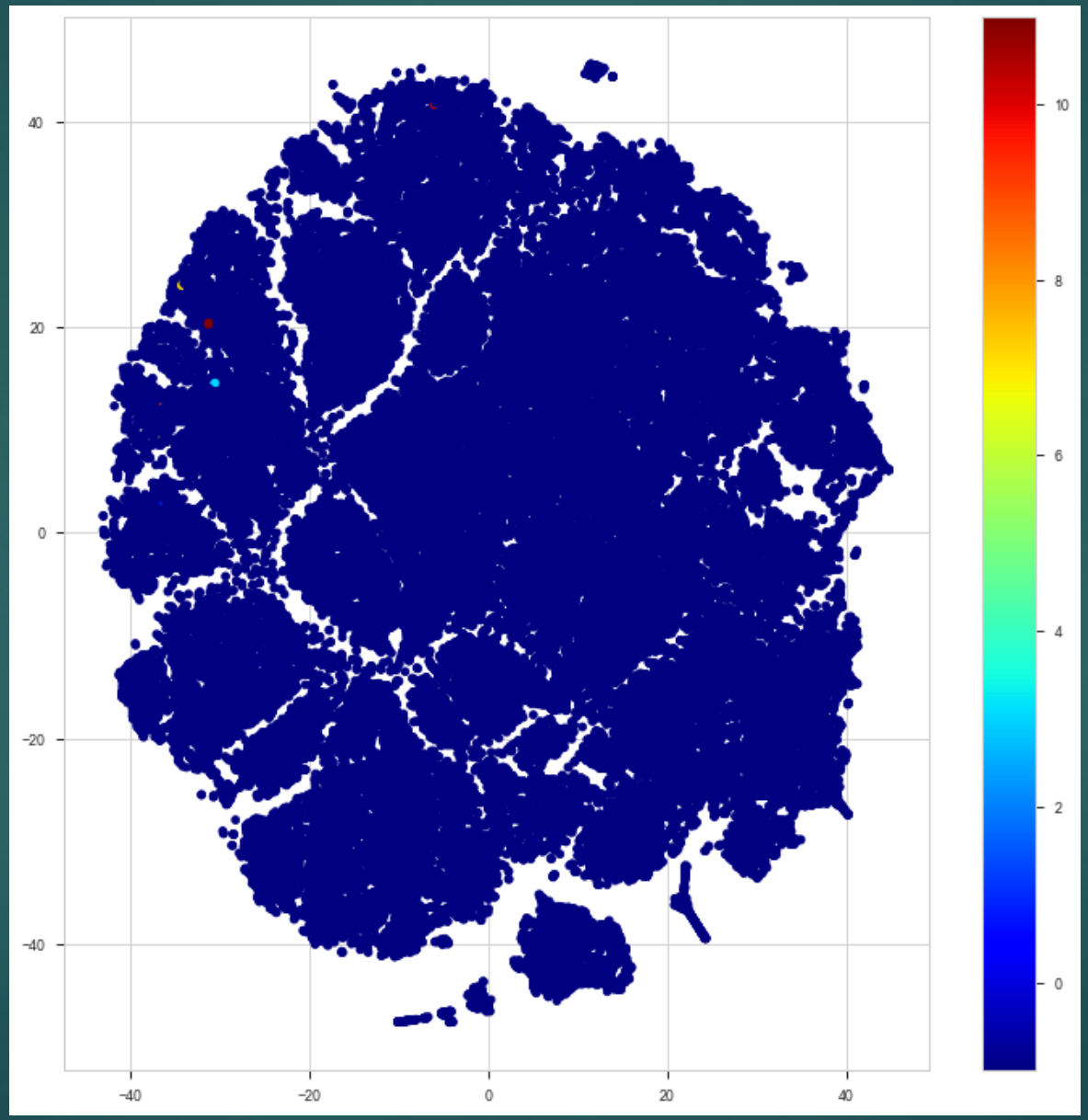
Modelisation avec Clustering Hiérarchique



Taille cluster :
0: 16087
1: 54
2: 2826
3: 1004
4: 20
5: 9

3 cluster avec moins de 100 client

Modélisation avec DBScan



Mauvais clustering

Modélisation

Choix du modèle :

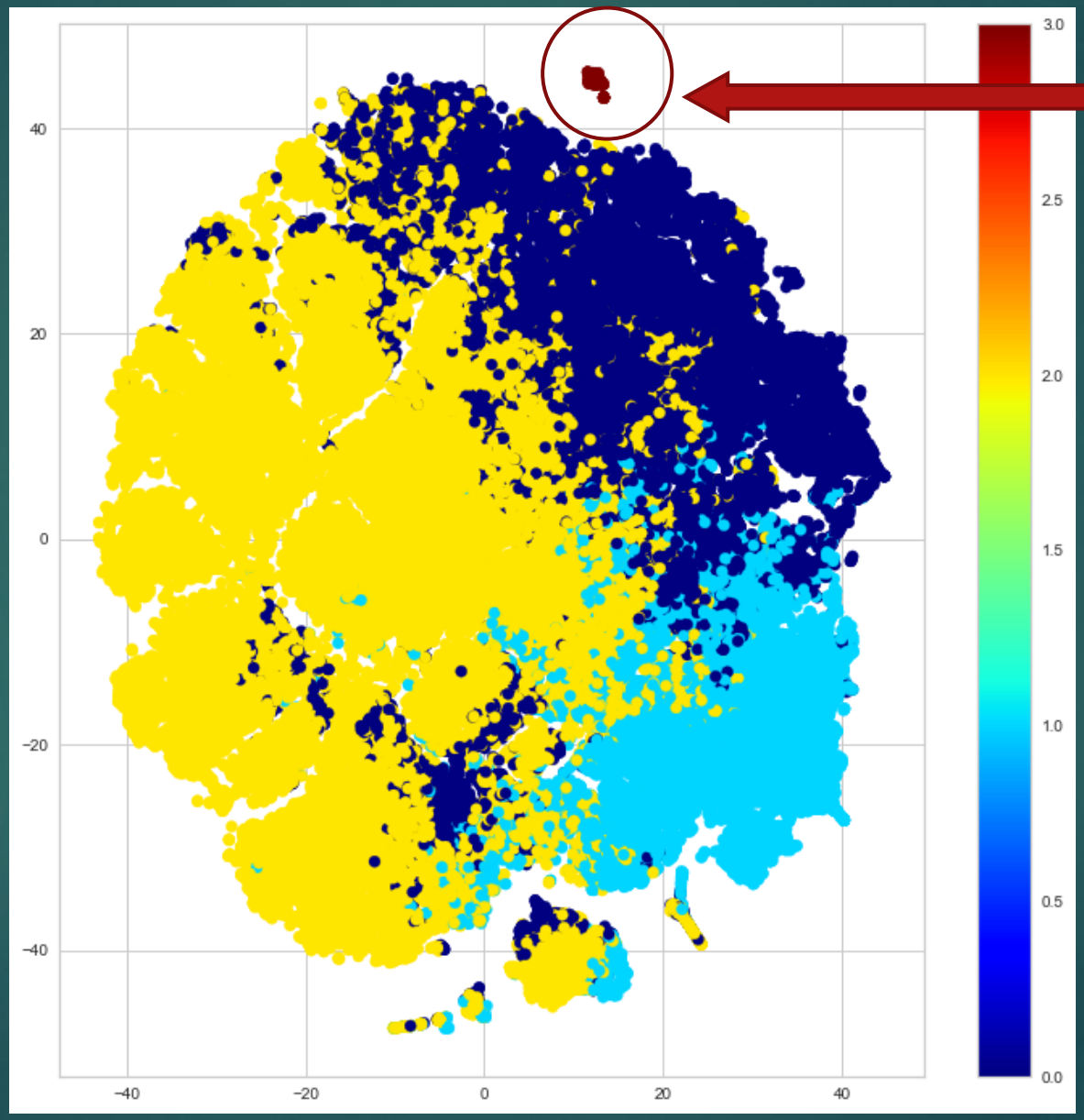
DBScan : Mauvais clustering.

Clustering hiérarchique : 3 clusters ont une taille inferieur à 100

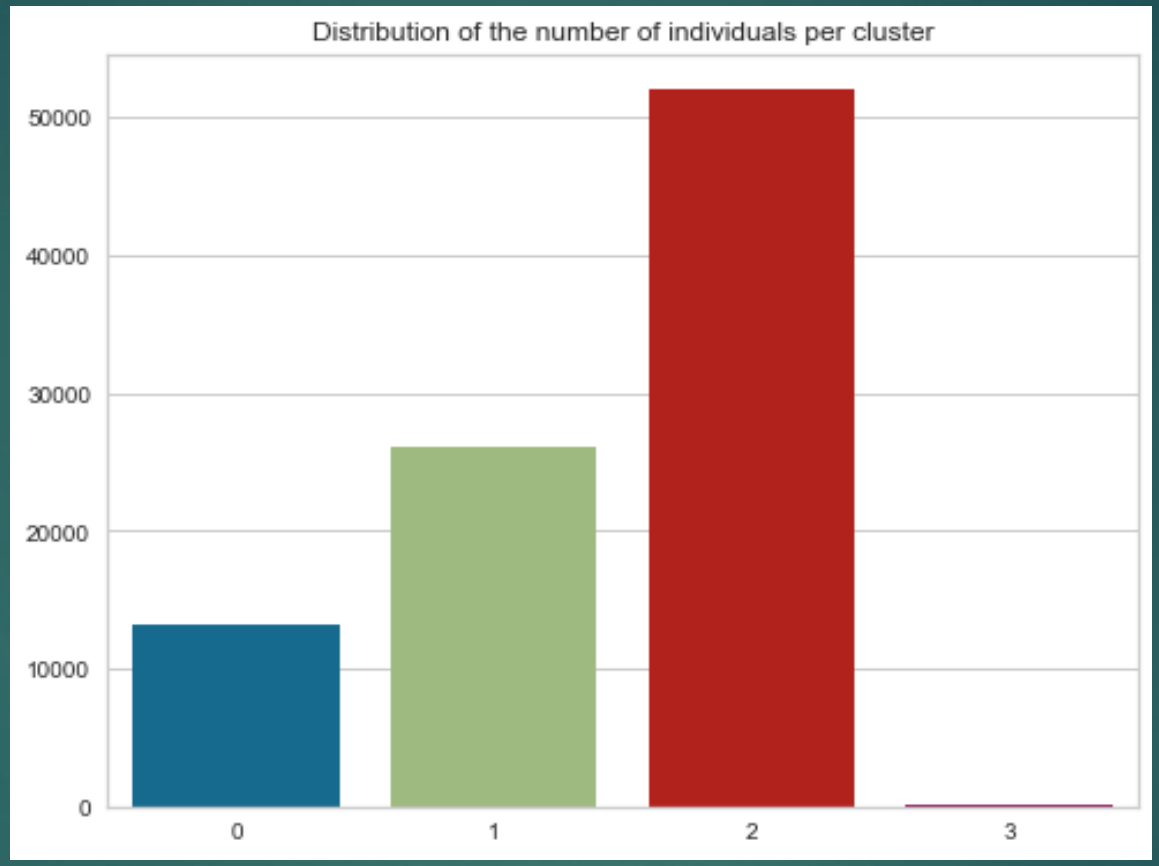
Kmeans : 3 à 4 clusters optimal avec des tailles de clusters relativement correct

On choisit Kmeans mais le nombre de cluster reste à verifier (3 ou 4)

Modelisation



Modélisation : Kmeans

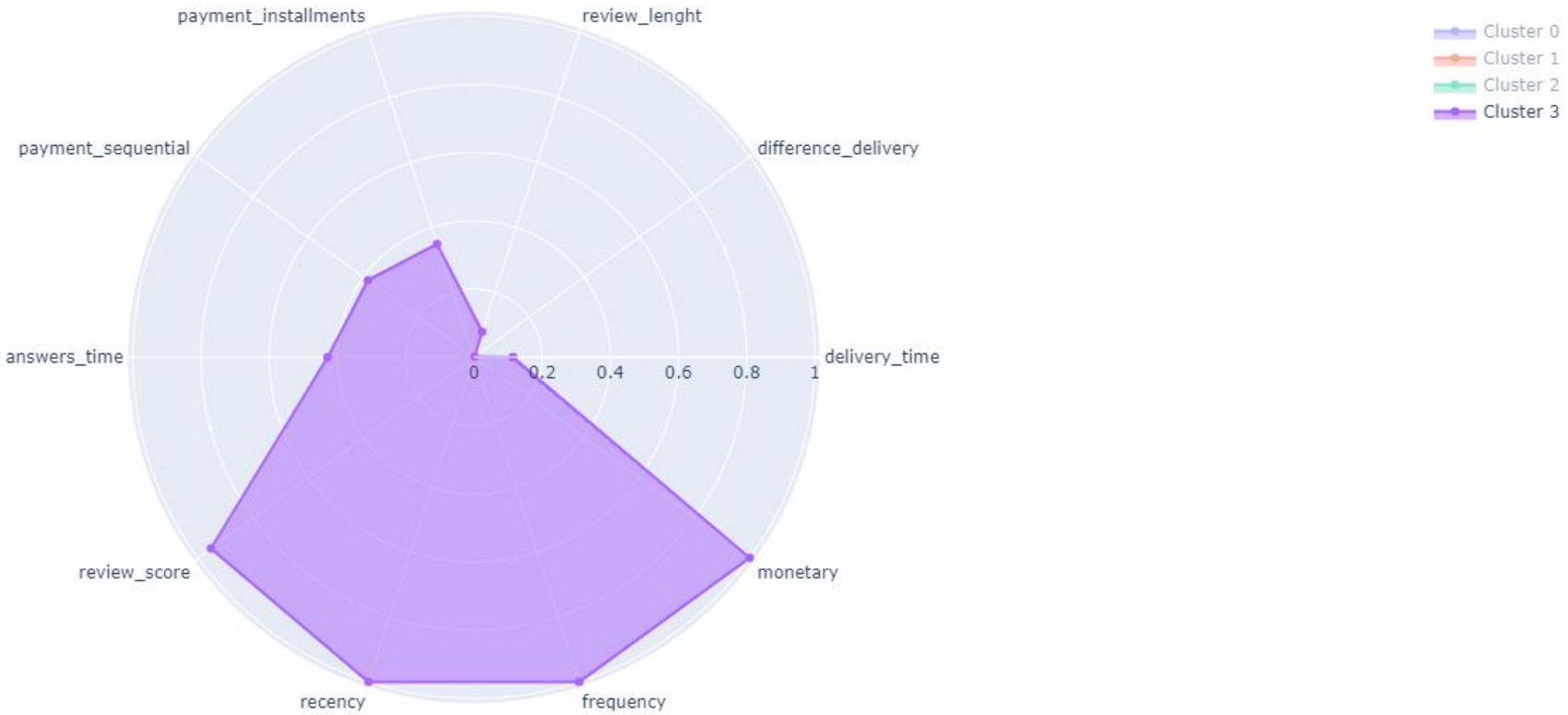


Cluster 3 < 500
On choisit donc n_cluster = 3
Interessant de voir a quoi correspond ce cluster 3 ??

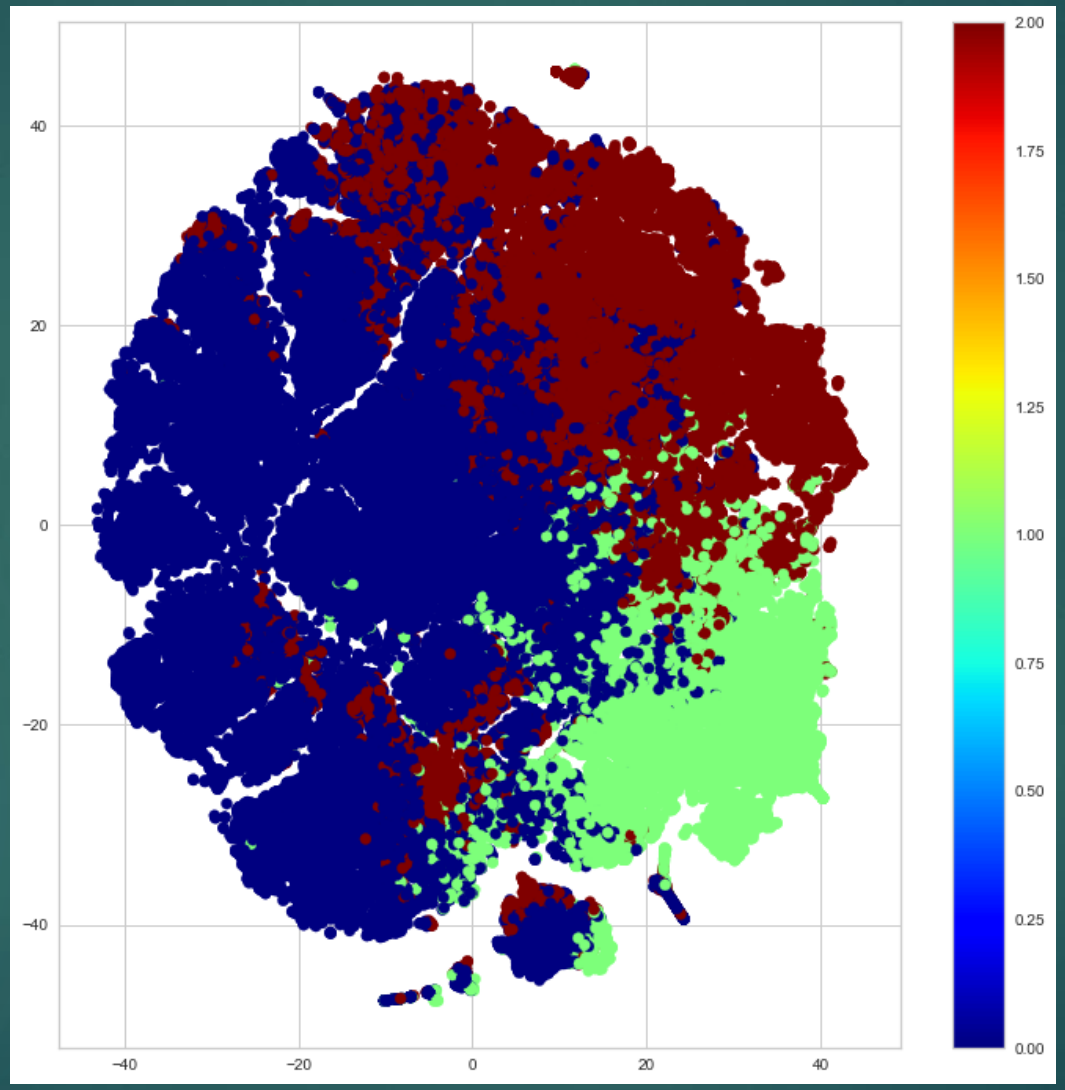
Modélisation : Kmeans

Cluster comparison

n_cluster : 4



Modélisation : Kmeans

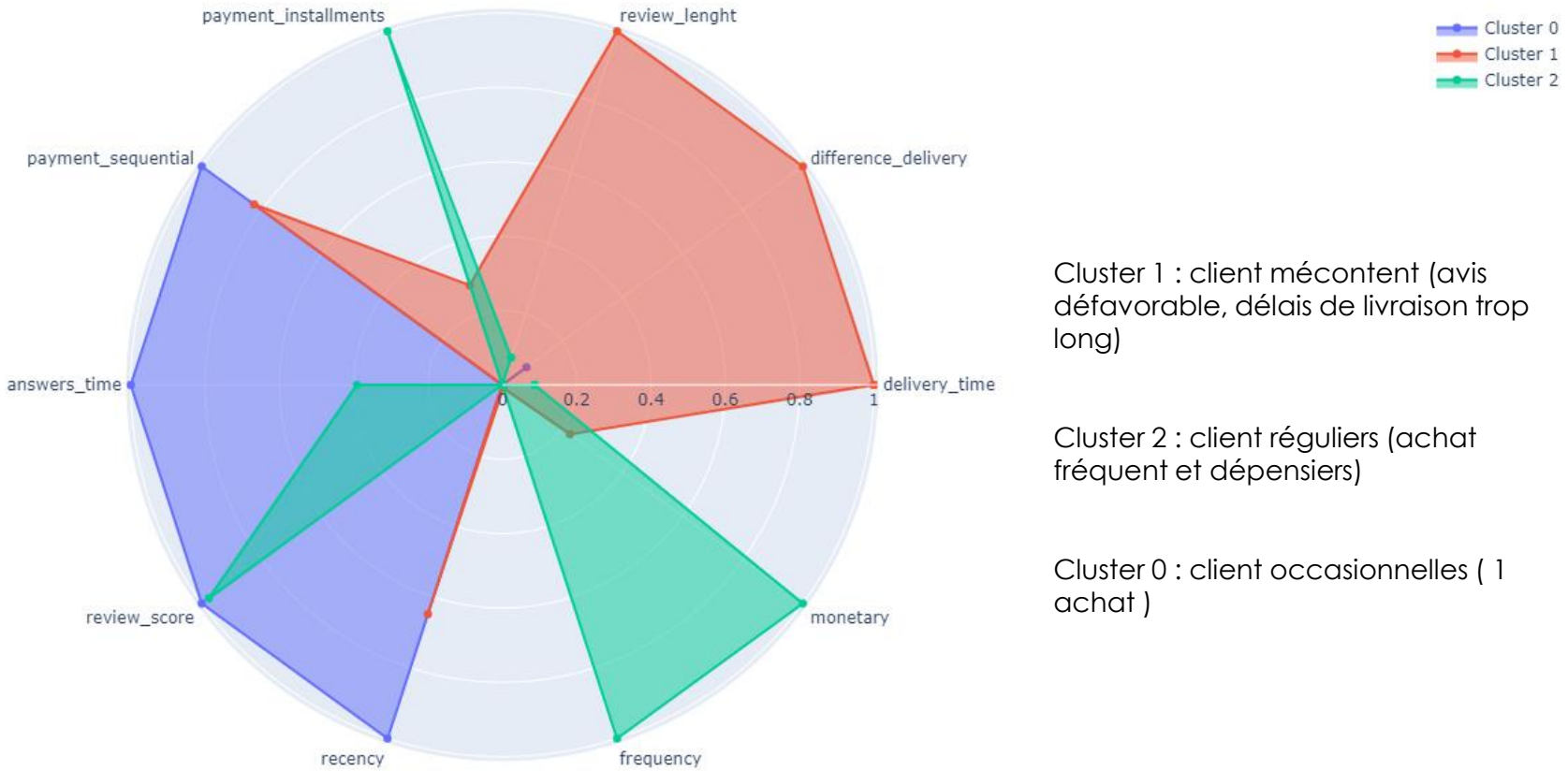


n_cluster : 3

Modélisation : Kmeans

Cluster comparison

n_cluster : 3



Cluster 1 : client mécontent (avis défavorable, délais de livraison trop long)

Cluster 2 : client réguliers (achat fréquent et dépensiers)

Cluster 0 : client occasionnelles (1 achat)

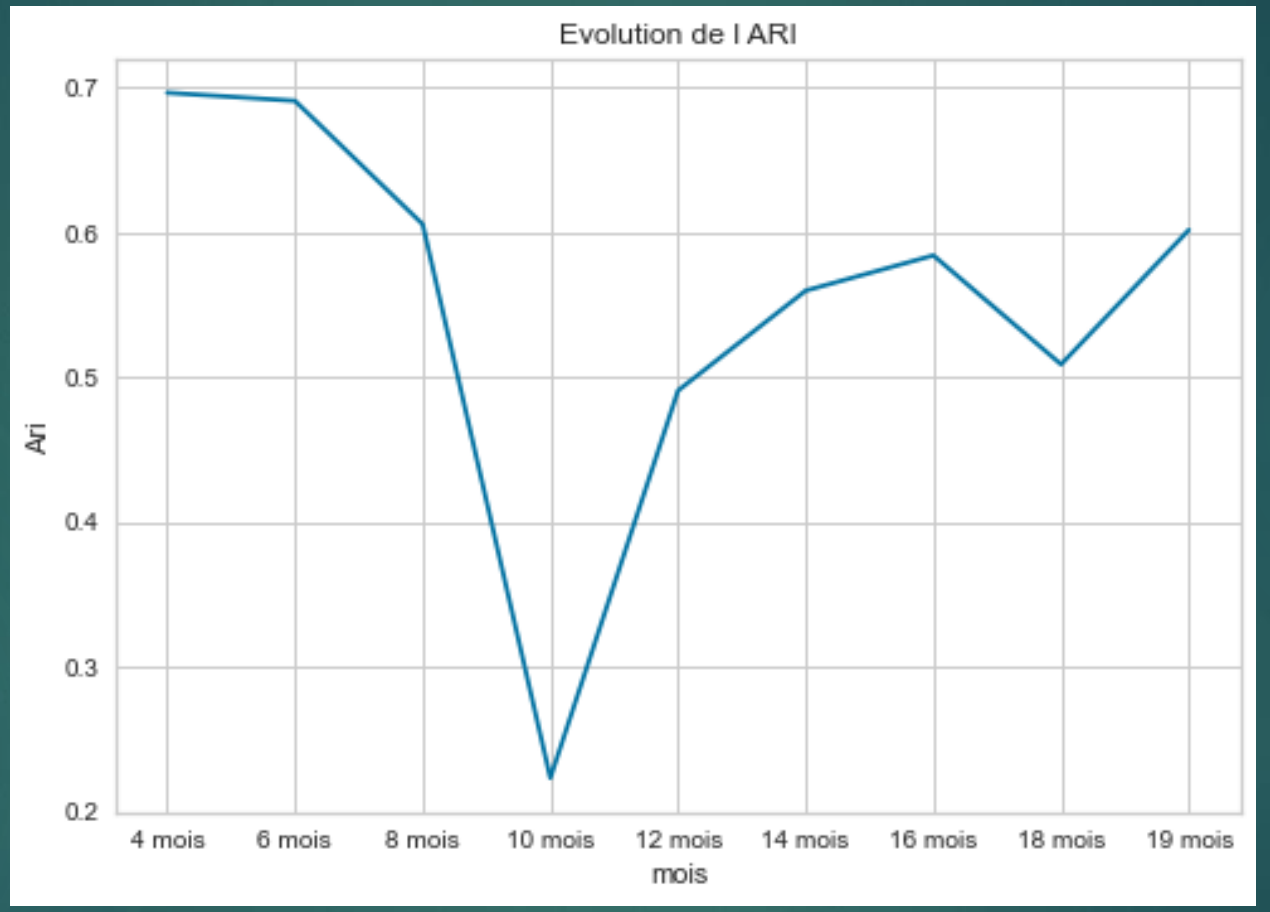
Maintenance du modèle

- On splitte le jeux de données tout les 2 mois

Dataframe 1	0.0 MOIS	248 clients
Dataframe 2	1.0 MOIS	728 clients
Dataframe 3	2.0 MOIS	4006 clients
Dataframe 4	2.0 MOIS	5545 clients
Dataframe 5	2.0 MOIS	6681 clients
Dataframe 6	2.0 MOIS	7868 clients
Dataframe 7	2.0 MOIS	11448 clients
Dataframe 8	2.0 MOIS	11591 clients
Dataframe 9	2.0 MOIS	13389 clients
Dataframe 10	2.0 MOIS	12783 clients
Dataframe 11	2.0 MOIS	12149 clients
Dataframe 12	1.0 MOIS	5017 clients

- On supprime les 2er Dataframe car peu de clients
- On calcule ensuite l'ARI (Adjusted Rand score)

Maintenance du modèle



Chute de l'ARI au bout du 8eme mois
Maintenance à faire tout les 8 mois

Conclusion

Conclusion :

- Meilleur algorithme : K-means avec 3 clusters
- Maintenance du modèle tous les 8 mois
- 3 profils clients :

  - client mécontent (avis défavorable, délais de livraison trop long)
  - client réguliers (achat fréquent et dépensiers)
  - client occasionnelles (1 achat)