Pour info, sur la slide 5 il y a une animation qui cache le texte en dessous donc si possible plutôt regarder le ppt en mode diaporama pour voir le texte sur cette slide.

Segmentation clients

olist

Sommaire

1. Rappel de la problématique
2. Présentation de l’analyse exploratoire et des manipulations sur les données
3. Pistes de modélisation effectuées et modèle retenu
4. Analyse du délai de maintenance du modèle
5. Conclusion

Rappel de la problématique

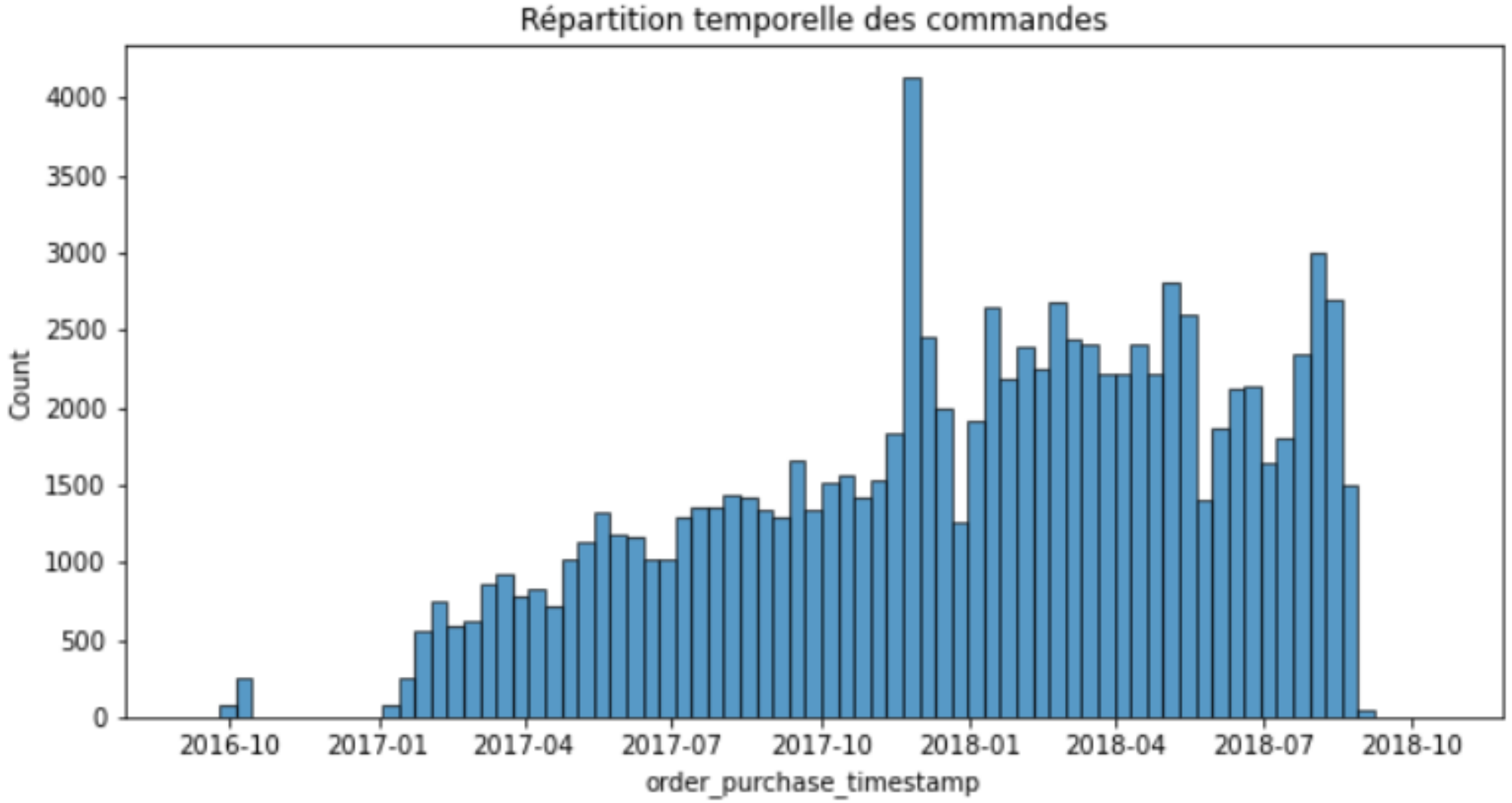
- **Olist**, entreprise brésilienne qui propose une solution de vente sur les marketplaces en ligne.

- Demande de **l'équipe marketing**.

- *Objectif :* créer une **segmentation clients** actionnable pour optimiser les campagnes de communication.

- Proposer un **contrat de maintenance** de cette segmentation.

Analyse exploratoire



(d

Manipulations sur les données

- Création d'un pivot table pour obtenir un détail par client.
- Variables retenues pour les modèles :
  - Nombre de commandes par client
  - La récence de la dernière commande par client
  - La moyenne du montant payé par commande par client
  - La moyenne des échéances de paiement par client
  - La moyenne des notes de satisfaction par client
  - La moyenne des % de 'vouchers' utilisés pour payer une commande par client

Manipulations sur les données

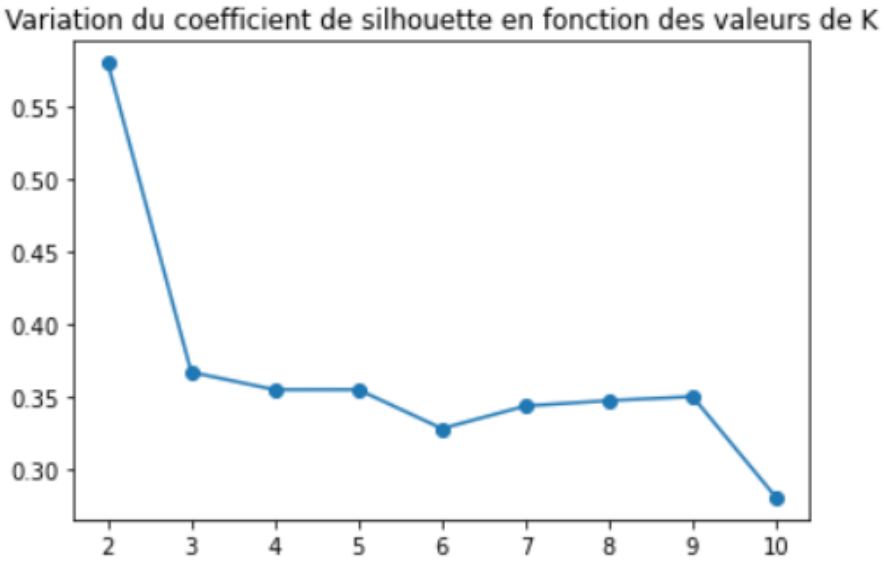
- Ordre de grandeur des différentes variables retenues :

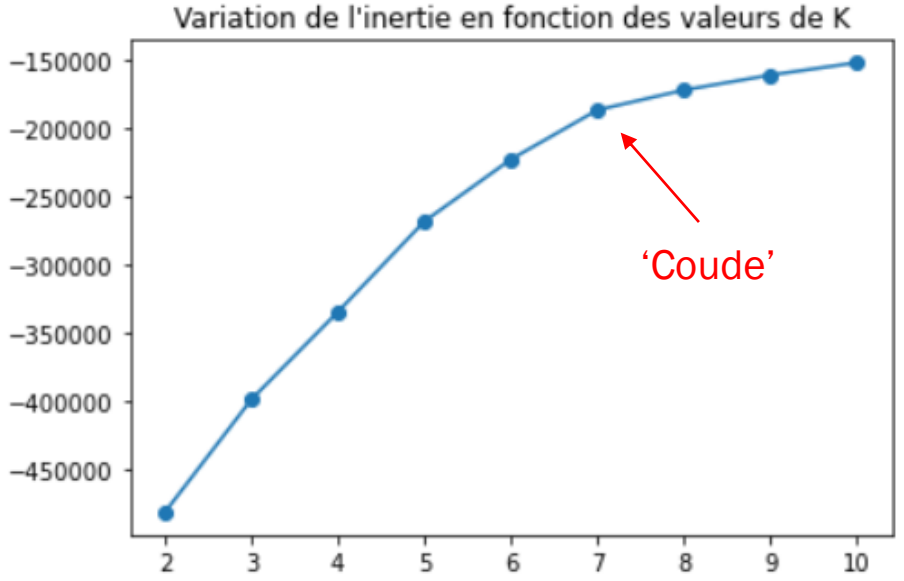
	mean_payment_installments	number_of_orders	last_order_purchase_recency_in_months	mean_ratio_voucher	mean_review_score	mean_order_value
count	95377.000000	95377.000000	95377.000000	95377.000000	95377.000000	95377.000000
mean	2.915372	1.040293	9.463576	0.030167	4.085066	161.140262
std	2.691178	0.254925	5.040478	0.159668	1.341513	221.010681
min	1.000000	1.000000	0.000000	0.000000	1.000000	9.590000
25%	1.000000	1.000000	5.377375	0.000000	4.000000	62.410000
50%	2.000000	1.000000	8.830819	0.000000	5.000000	105.740000
75%	4.000000	1.000000	13.052159	0.000000	5.000000	176.990000
max	24.000000	17.000000	25.391691	1.000000	5.000000	13664.080000

Modélisation : k-means

- Variables uniquement numériques : utilisation d'un StandardScaler pour la normalisation.

- Choix de la valeur de k grâce au coefficient de silhouette et à l'inertie :





→ Valeur retenue k=7

Modélisation : k-means

- Clustering obtenu avec k=7 :

kmeans_labels	count	mean_last_order_purchase_recency_in_months	mean_review_score	mean_ratio_voucher	mean_payment_installments	mean_order_value	mean_number_of_orders
0	12149	9.751611	4.292370	0.002687	8.284386	258.503092	1.000082
1	35733	5.619686	4.662357	0.001383	1.860297	118.528441	1.000000
2	14384	9.558620	1.535943	0.001763	2.444869	142.637125	1.000000
3	2930	9.998658	4.035836	0.895242	1.146075	114.706766	1.020819
4	1510	9.514848	3.832119	0.006131	6.449669	1413.842394	1.007285
5	2962	8.874368	4.116405	0.031066	3.340810	143.060039	2.272789
6	25709	14.620809	4.627796	0.001769	2.052939	118.507713	1.000000

Cluster 0 → clients d'un niveau économique plus faible car ils paient en 8 fois des montants beaucoup plus faibles que par exemple le groupe 4 dont les clients paient en 6 fois.

Cluster 1 → clients qui ont passé leur dernière commande récemment et qui sont très satisfaits.

Cluster 2 → clients mécontents.

Cluster 3 → clients qui utilisent quasiment à 100% des vouchers pour payer.

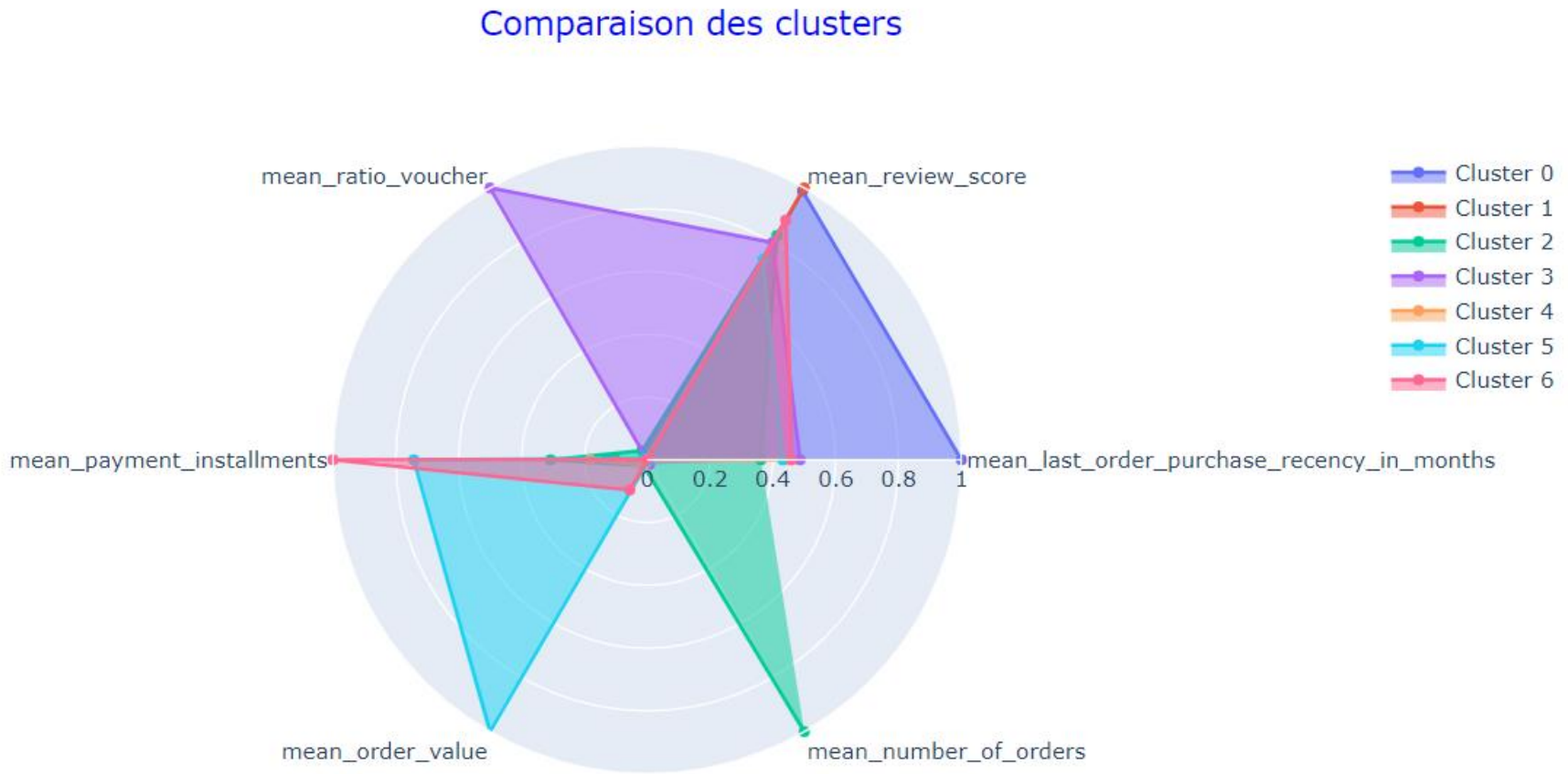
Cluster 4 → clients qui ont passé une commande d'un montant élevé et qui sont assez satisfaits.

Cluster 5 → clients qui ont passé plusieurs commandes.

Cluster 6 → clients qui ont passé leur dernière commande il y a longtemps et qui sont très satisfaits.

Modélisation : k-means

- Illustration sur un radar plot :



Modélisation : k-means

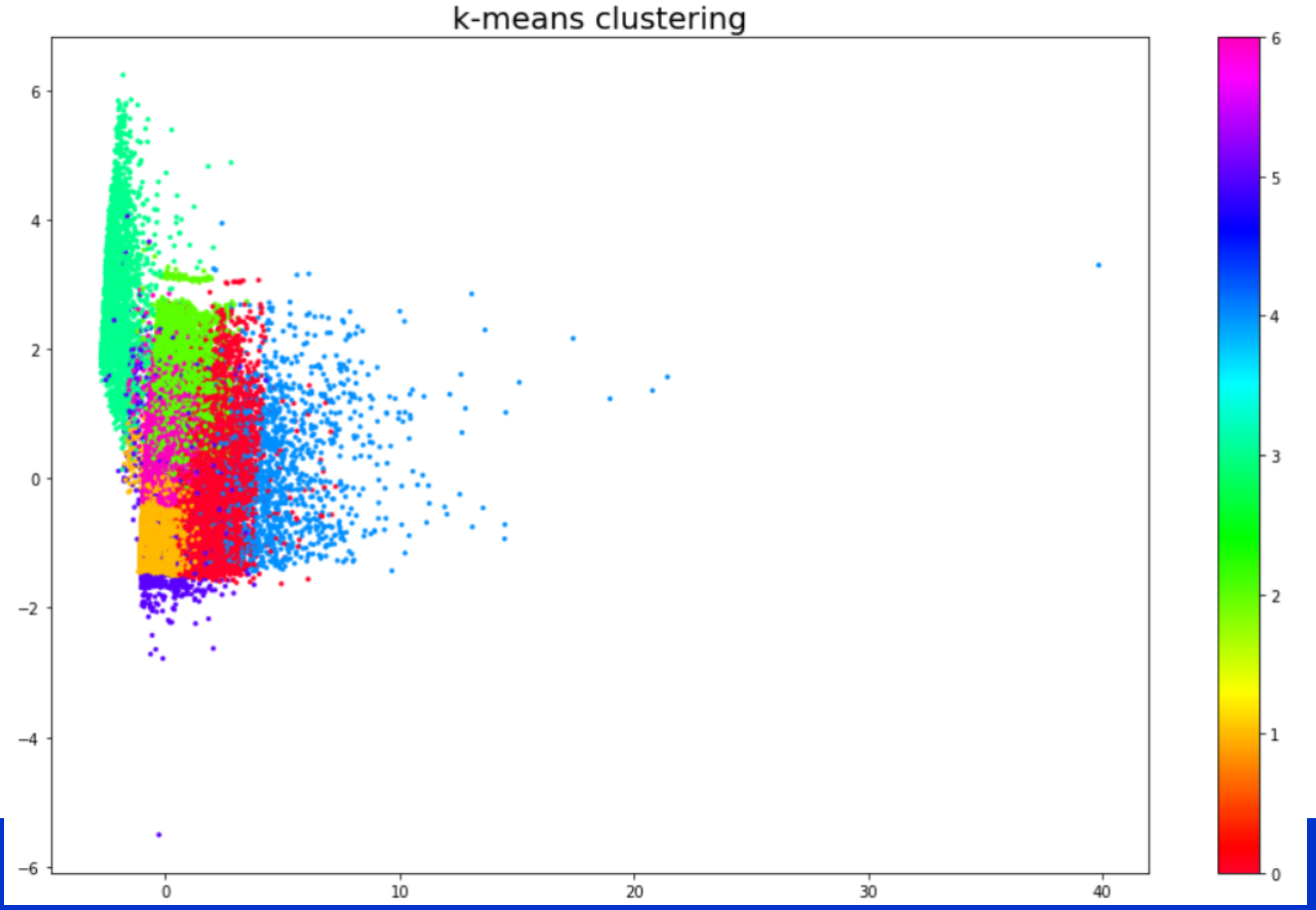
- Analyse de la stabilité à l'initialisation du k-means → on compare les clusters obtenus précédemment avec 10 nouvelles itérations du k-means grâce à l'indice de Rand ajusté (ARI).

```
Itération 1 : ARI = 0.995
Itération 2 : ARI = 0.716
Itération 3 : ARI = 0.999
Itération 4 : ARI = 0.999
Itération 5 : ARI = 0.997
Itération 6 : ARI = 0.998
Itération 7 : ARI = 0.998
Itération 8 : ARI = 0.956
Itération 9 : ARI = 0.958
Itération 10 : ARI = 0.999
```

→ ARI proche de 1 pour quasiment toutes les itérations donc la stabilité à l'initialisation du k-means est bonne.

Modélisation : k-means

- Réalisation d'une analyse en composantes principales (PCA) pour permettre d'afficher les clusters en 2 dimensions :



Modélisation : DBSCAN

- Echantillonnage du jeu de données en ne prenant que 20% pour des questions de temps de calcul.

- StandardScaler également utilisé pour normaliser les données.

- Application du DBSCAN avec valeurs des hyperparamètres par défaut :

  - 32 clusters retenus par l'algo dont la majorité contiennent moins de 25 clients.

  - Coefficient de silhouette égal à -0,016

→ Différentes valeurs des hyperparamètres 'eps' et 'min_samples' testées pour essayer d'améliorer le clustering.

Modélisation : DBSCAN

	eps	min_samples	number_of_clusters	silhouette_score
0	0.5	5.0	32.0	-0.016089
1	0.5	10.0	15.0	-0.005176
2	0.5	20.0	8.0	0.031040
3	0.5	50.0	7.0	0.000620
4	0.6	5.0	28.0	0.000316
5	0.6	10.0	12.0	0.068183
6	0.6	20.0	9.0	0.016060
7	0.6	50.0	8.0	-0.018355
8	0.8	5.0	9.0	0.282222
9	0.8	10.0	8.0	0.401169
10	0.8	20.0	3.0	0.423027
11	0.8	50.0	3.0	0.405879

```
dbscan_10_labels
 0    0.917274
-1    0.033761
 1    0.025583
 2    0.019030
 3    0.002307
 4    0.000629
 6    0.000629
 5    0.000524
 7    0.000262
```

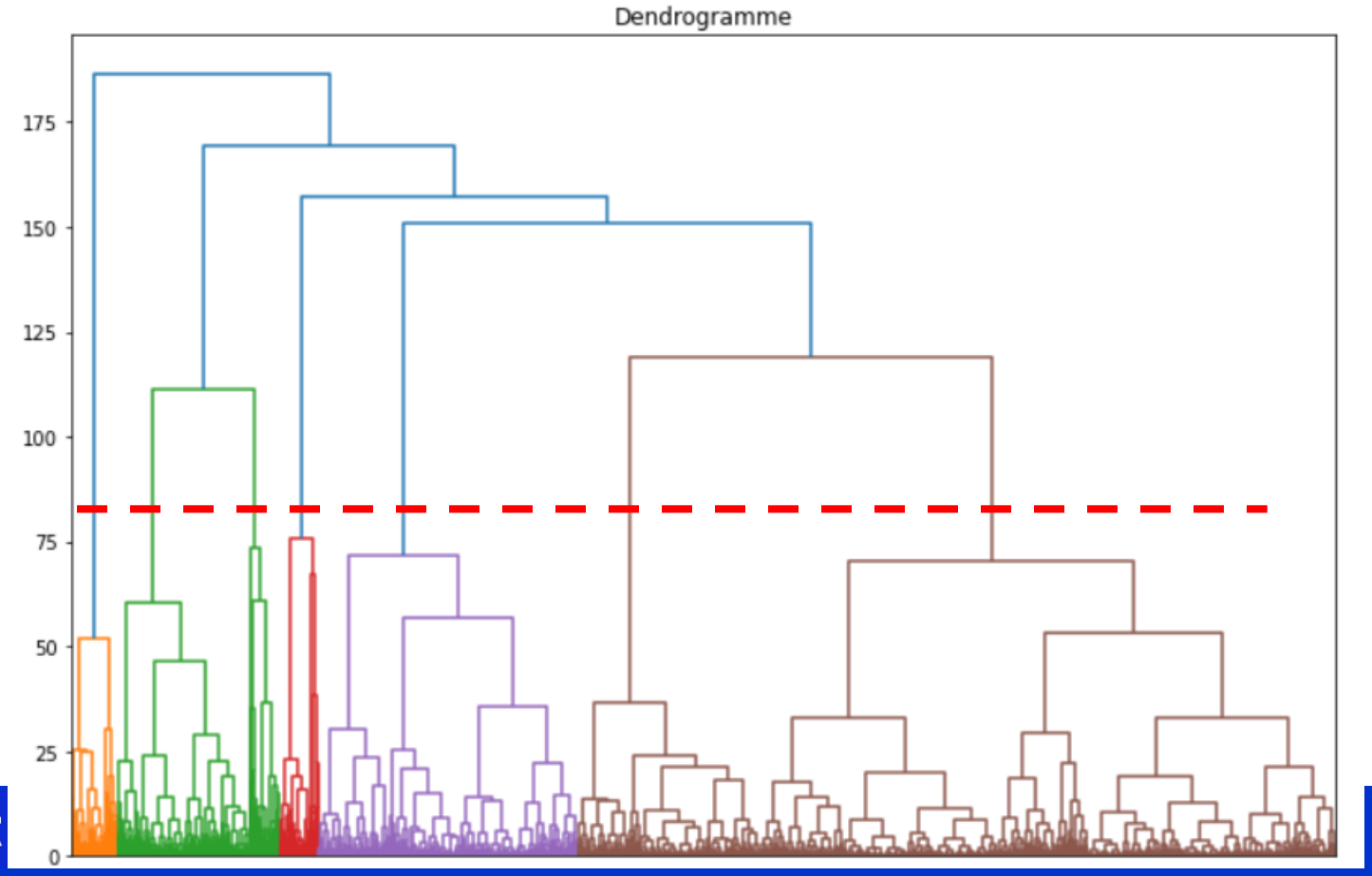
```
dbscan_20_labels
 0    0.910826
-1    0.051009
 1    0.021547
 2    0.016619
```

→Les clusters sont très déséquilibrés que ce soit avec 8 ou 3 clusters.

→L'algo DBSCAN ne semble pas approprié pour ce jeu de données.

Modélisation : Clustering hiérarchique

- Echantillonnage du jeu de données en ne prenant que 20% pour des questions de temps de calcul.

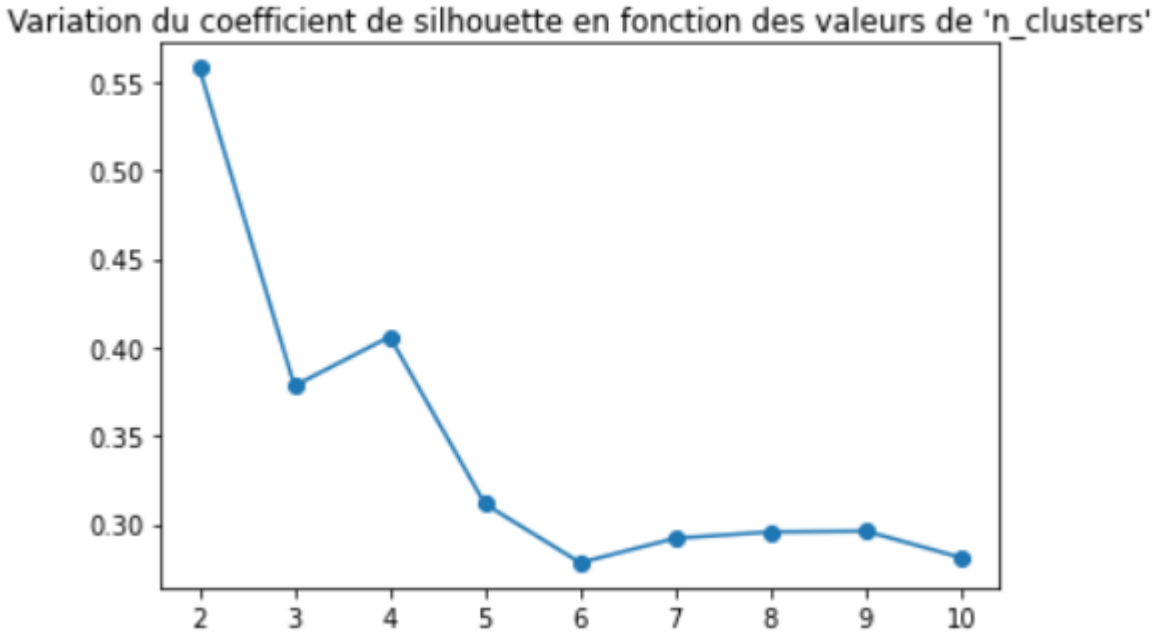


→Couper vers l'ordonnée 80 semble être une bonne coupe.

→On obtiendrait 7 clusters.

Modélisation : Clustering hiérarchique

- Analyse théorique par l'intermédiaire du coefficient de silhouette :



→On analyse les profils des clusters avec 2, 4 et 7 clusters.

Modélisation : Clustering hiérarchique

- Proportion de clients par cluster :
  - Pour 2 clusters :

0	0.963722
1	0.036278

→Très déséquilibré.

  - Pour 4 clusters :

0	0.804509
2	0.129279
3	0.036278
1	0.029934

→Assez déséquilibré également.

  - Pour 7 clusters :

2	0.454364
4	0.205033
5	0.145111
6	0.105269
3	0.036278
0	0.029934
1	0.024010

→Mieux équilibré et on remarque en analysant les profils des clusters que l'on retrouve le même clustering que celui obtenu avec le k-means et k=7.

Modèle retenu

- D'un point de vue métier, la segmentation en ***7 groupes*** semble être la plus pertinente dans la perspective de faire ensuite des ***campagnes marketing ciblées par rapport au niveau économique des clients, de leur satisfaction ou de leur appétence pour des bons de réduction*** par exemple.

- ***Modèle retenu :*** k-means avec k=7.

Délai de maintenance du modèle

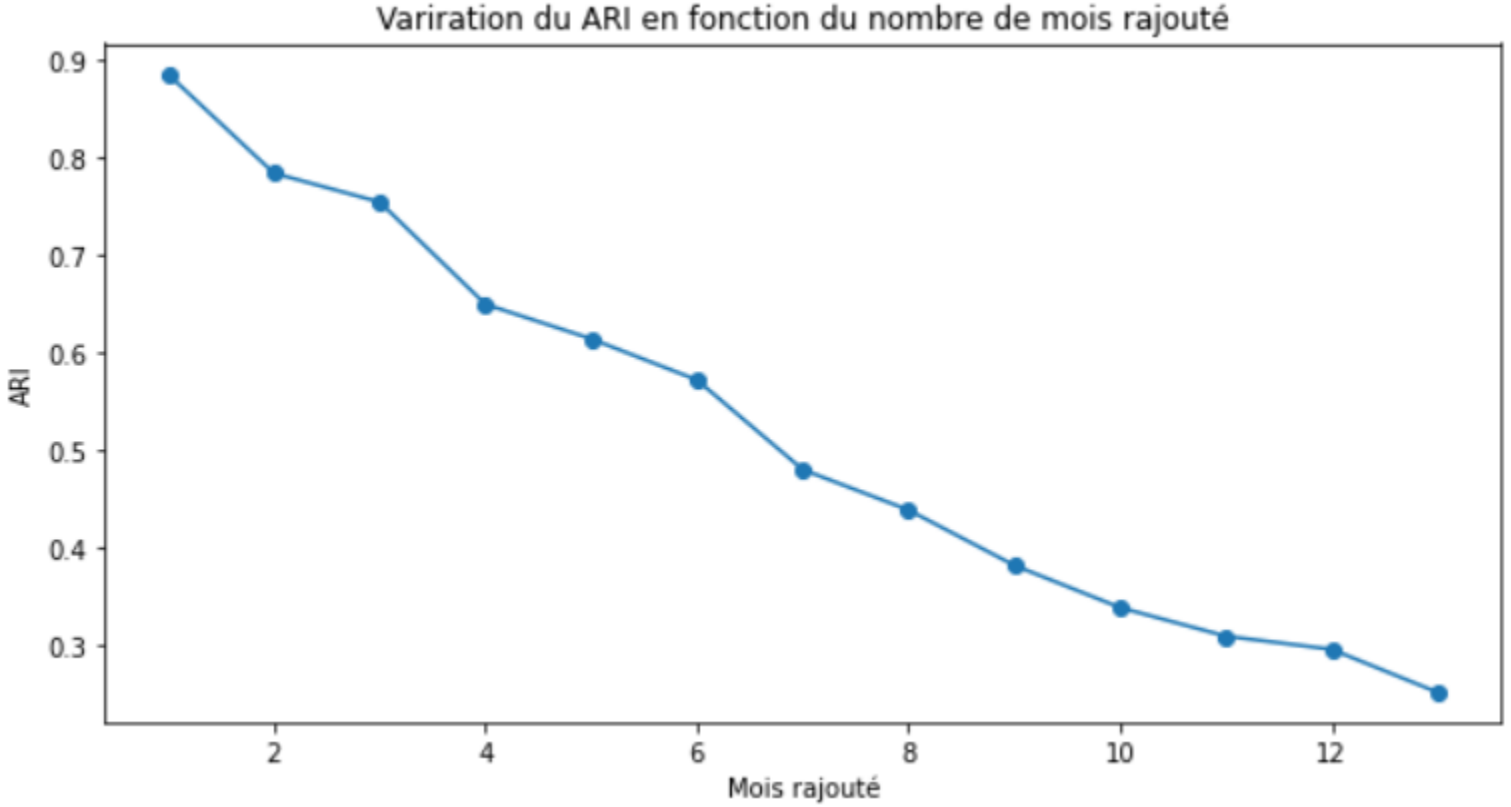
- Analyse de la stabilité temporelle de la segmentation retenue.

- ***Méthode utilisée :***

  - Entrainement du k-means avec k=7 sur les 12 premiers mois de commandes.
  - A chaque ajout d'un nouveau mois de commande, comparaison des résultats grâce au Adjusted Rand Index (ARI) de la prédiction du k-means initialement entrainé et d'un nouveau k-means réentrainé sur la nouvelle base de données.
  - Affichage des résultats sur un graphique et recherche du point d'inflexion du ARI.

Délai de maintenance du modèle

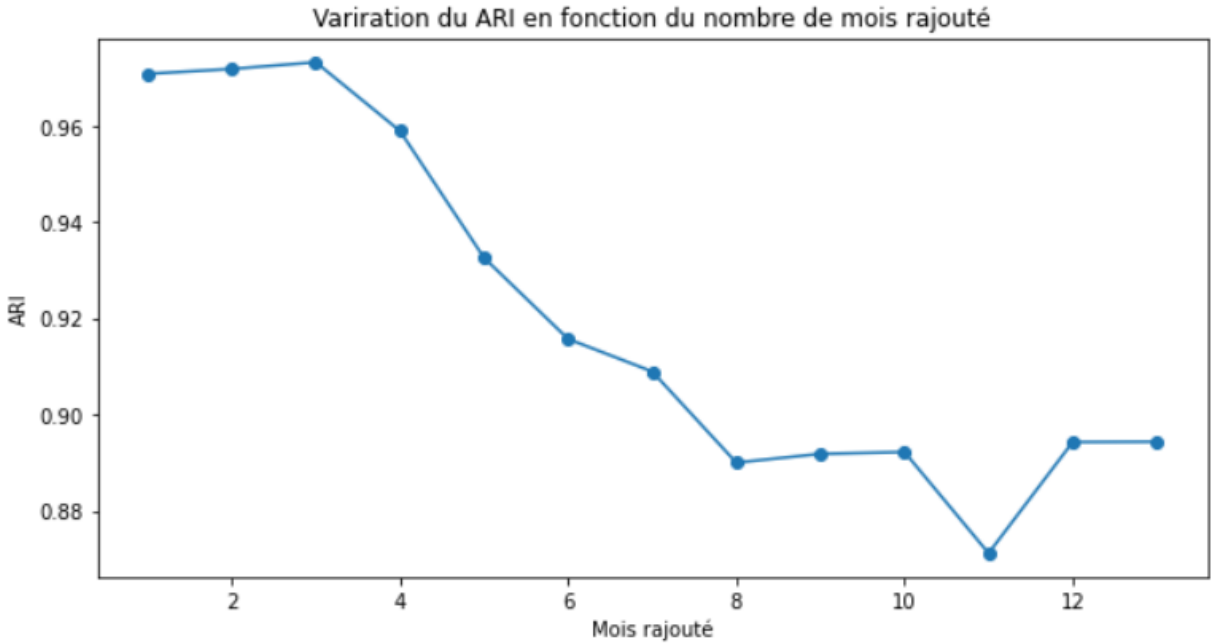
- Résultats :



→ Diminution continue du ARI dans le temps. Donc la 1ière maintenance devrait intervenir à la fin du 1er mois.

Délai de maintenance du modèle

- Résultats en utilisant un nouveau StandardScaler après chaque ajout de mois avant l'utilisation du k-means initial :



→ ARI plus stable avec de plus grandes valeurs. Dans ce cas la 1ière maintenance devrait plutôt avoir lieu à la fin du 3ième mois.

Conclusion

- *Modèle retenu :* k-means avec k=7.
- Segmentation en ***7 groupes.***
- Campagnes marketing ciblées notamment par rapport au ***niveau économique des clients, de leur satisfaction ou de leur appétence pour des bons de réduction.***
- *Délai de maintenance :* tous les mois.
- Tous les 3 mois s'il est possible pour l'entreprise de réextraire facilement chaque mois la nouvelle base de données clients et que l'on applique à chaque fois un nouveau StandardScaler aux données avant d'utiliser le modèle.