SEGMENTATION DES CLIENTS

OLIST

Marie-France LAROCHE-BARTHET

20/01/2022

INTRODUCTION

PROBLÉMATIQUE

- Client du projet :
  - **OLIST**
  - Entreprise brésilienne
  - Solution de vente sur les marketplaces en ligne
- Enjeux du client :
  - **Comprendre différents types d'utilisateurs**
  - Cibler campagnes de communication
- BUT du projet :
  - Modèle de segmentation **non supervisé**
    - **Segmenter clients**
    - **Caractérisation** des segments
    - **Fréquence de mise à jour du modèle**

ANALYSE EXPLORATOIRE

ANALYSE EXPLORATOIRE

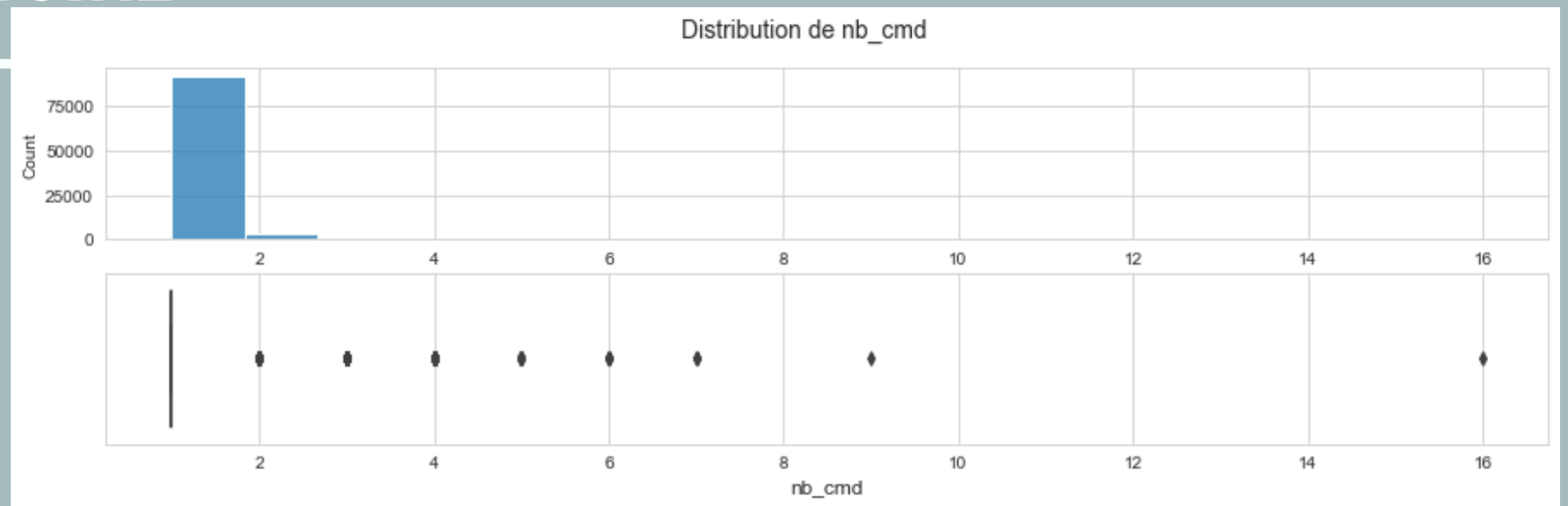
- **Premier aperçu**
  - 9 Datasets
  - **96096** clients uniques
  - **99440** commandes ➡ 3.1% des clients avec au moins 2 commandes
  - Outliers (*price*, *freight_value*, *payment_value*)
  - > 50% *review_score* = 5
- Fusion des datasets : 1 dataset global avec toutes les colonnes
  - **94720** clients uniques **(1.4%)**
  - **97916** commandes **(1.5%)** ➡ 3% des clients avec au moins 2 commandes
  - Données géolocalisation pour client et vendeur
  - Noms des produits en anglais
- Feature Engineering **(94720 x 6)**
  - **5 variables numériques**
    - ***nb_cmd*** : Nombre total de commandes par client
    - ***montant_dep*** : Montant total dépensé par client
    - ***nb_prod*** : Nombre total de produits commandés par client
    - ***note_moy*** : Note moyenne attribuée par client
    - ***nb_jours_last_cmd*** : Nombre de jours écoulés entre dernière commande du client et commande la plus récente des données
  - **1 Variable catégorielle**
    - ***customer_unique_id*** : identifiant unique du client ➡ Servira d'index dans la suite

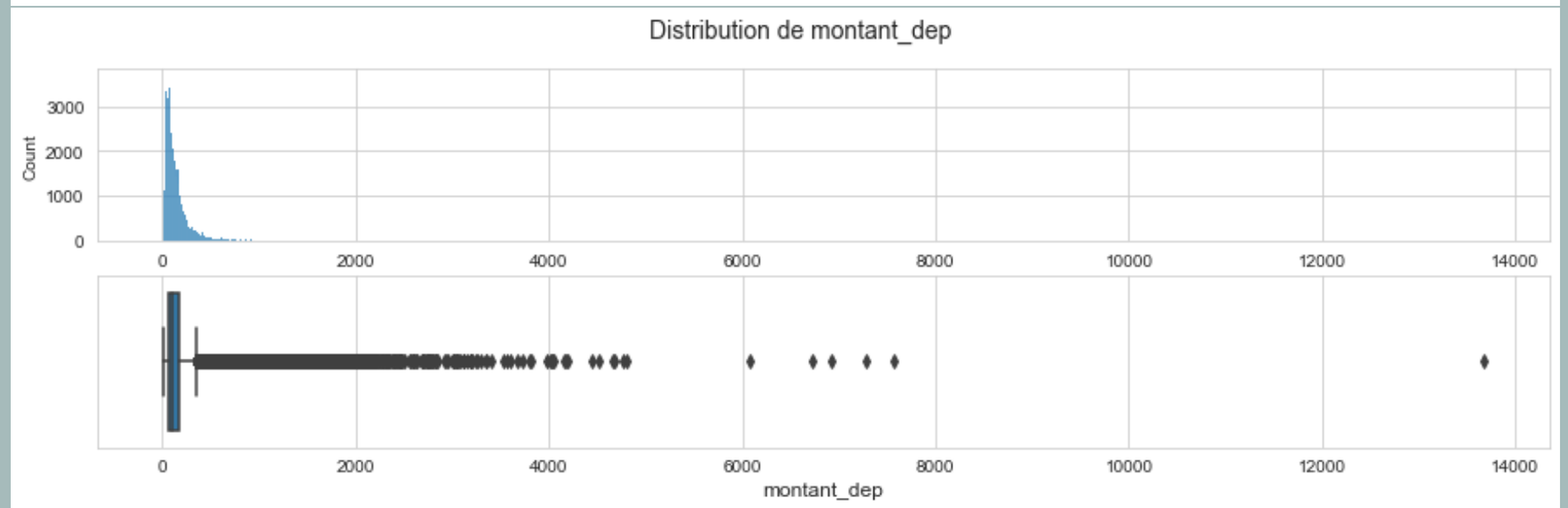
ANALYSE EXPLORATOIRE

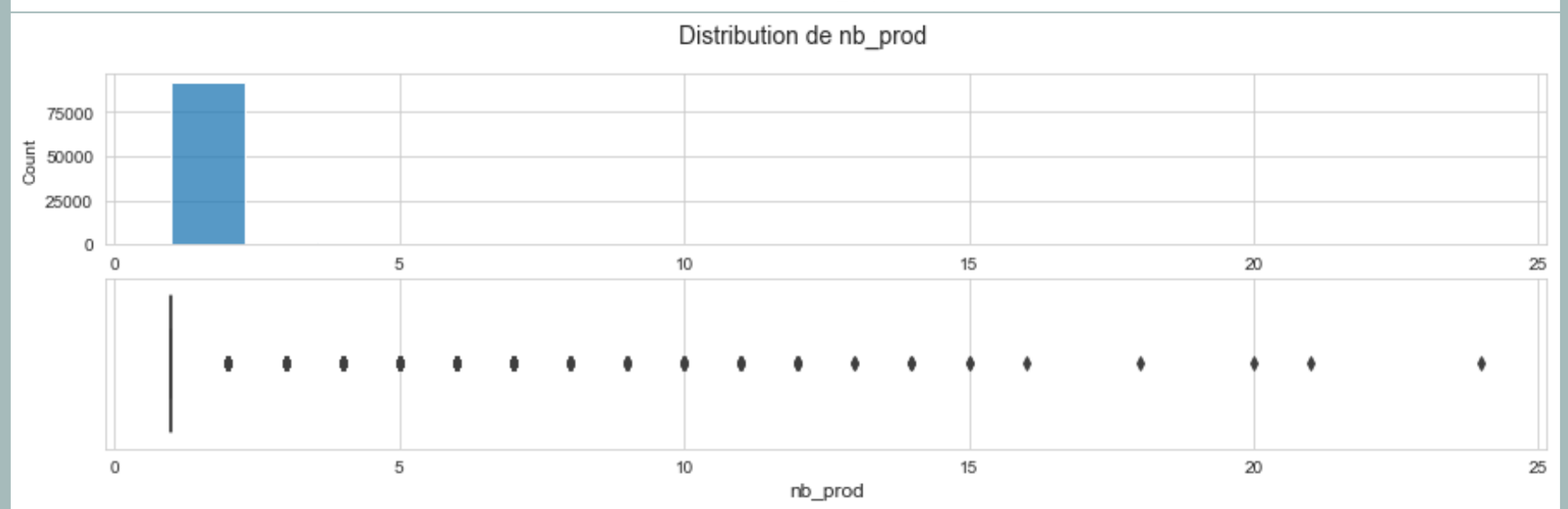
➤ *Distribution des variables*

- ➤ Pas de valeurs impossibles
- ➤ Variables déséquilibrées
- ➤ Présence d'outliers

➡ Utiliser RobustScaler



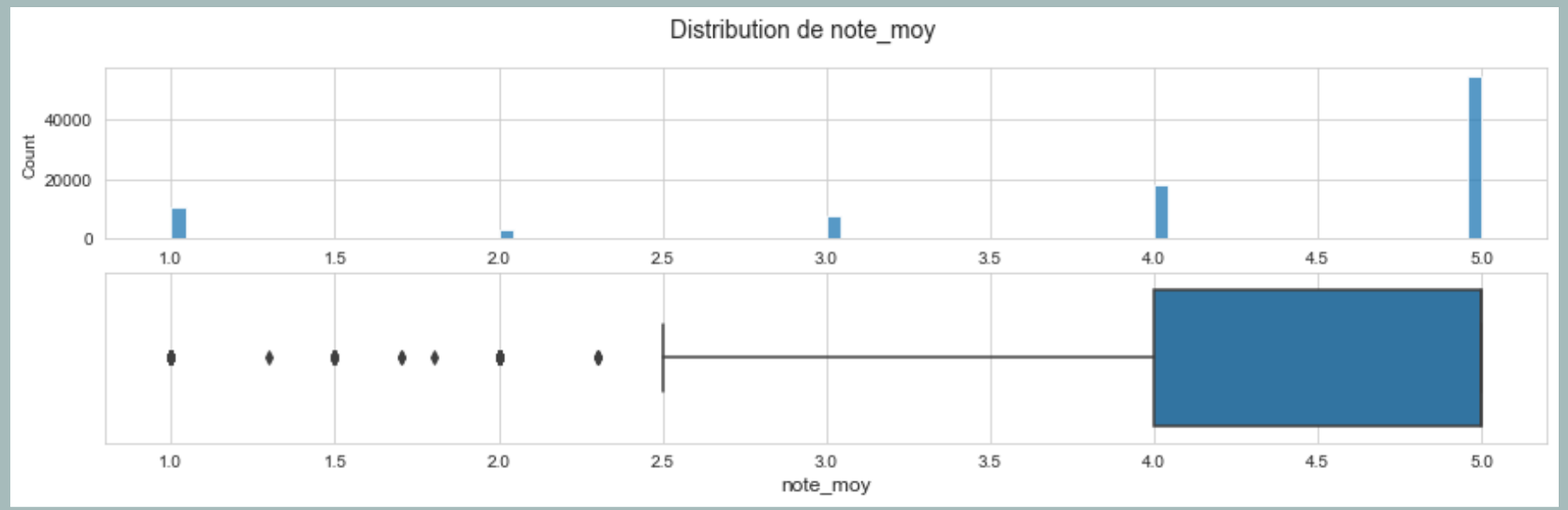




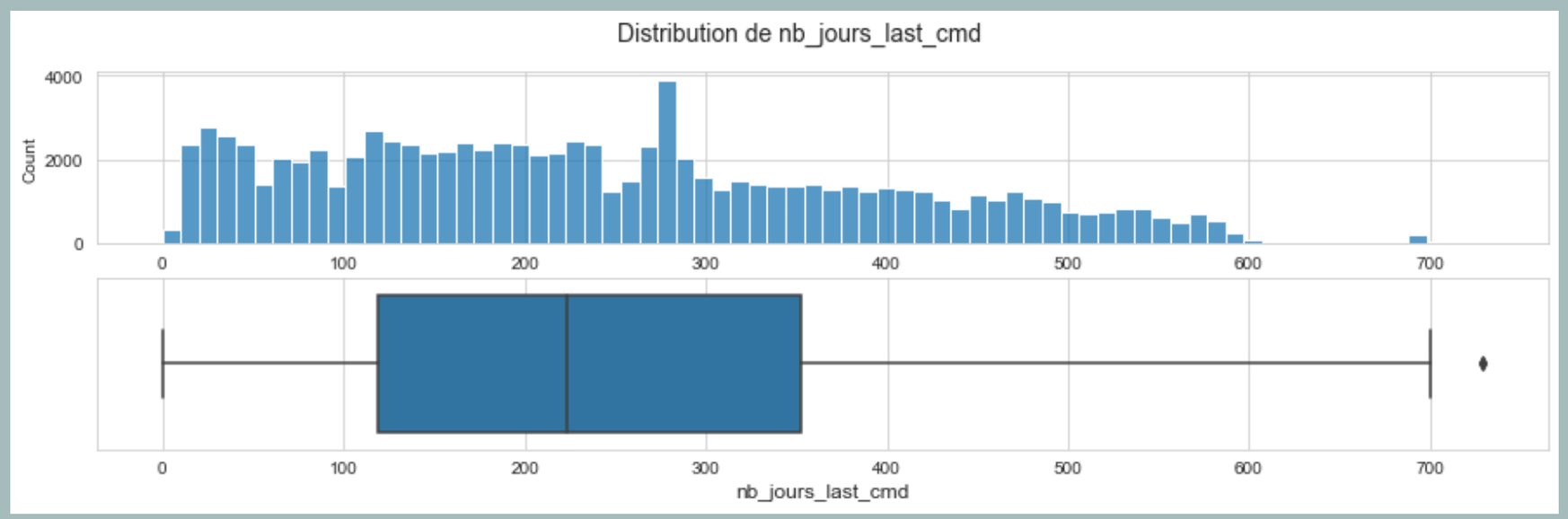
ANALYSE EXPLORATOIRE

- *Distribution des variables*

- Pas de valeurs impossibles
- Variables déséquilibrées
- Présence d'outliers



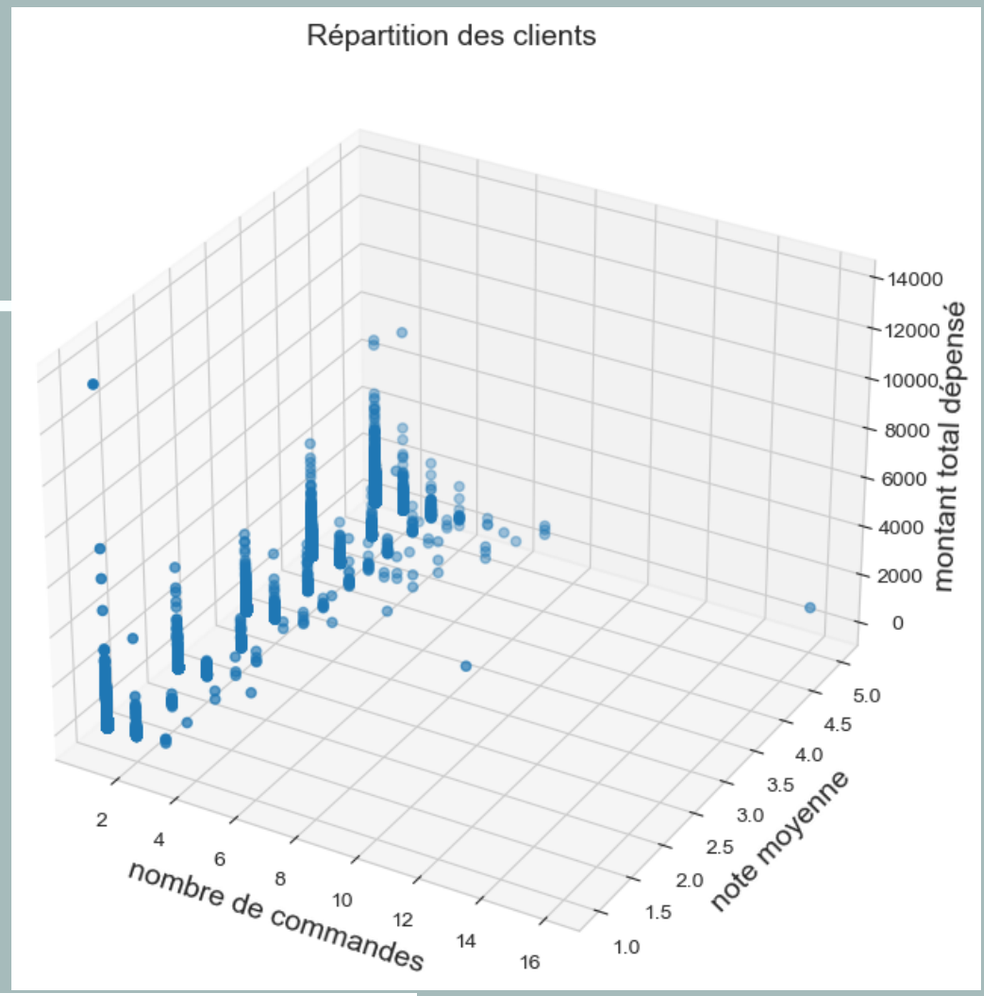
➡ Utiliser RobustScaler

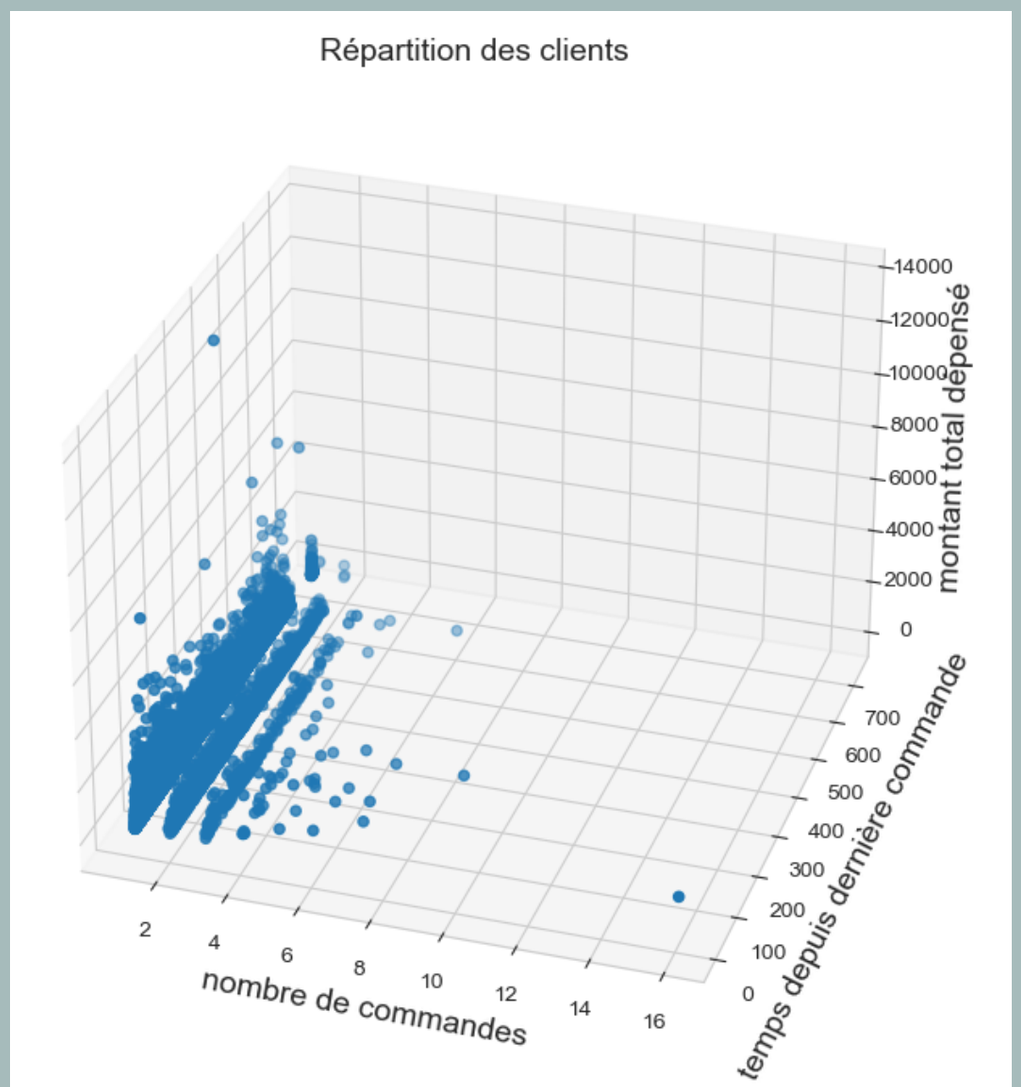


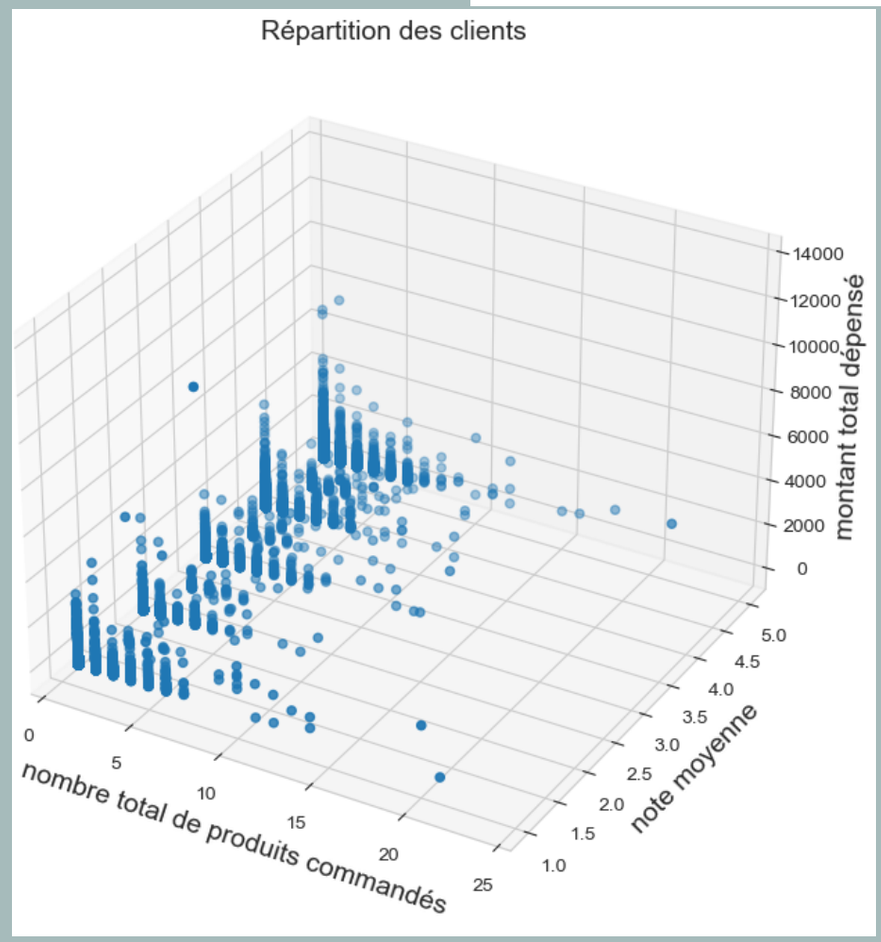
ANALYSE EXPLORATOIRE

➤ *Répartition des clients*

- ➤ Selon 3 variables
- ➤ ***nb_cmd*** et ***note_moy*** semblent séparer clients
- ➤ Doute sur intérêt du ***nb_jours_last_cmd*** et sur ***nb_prod***





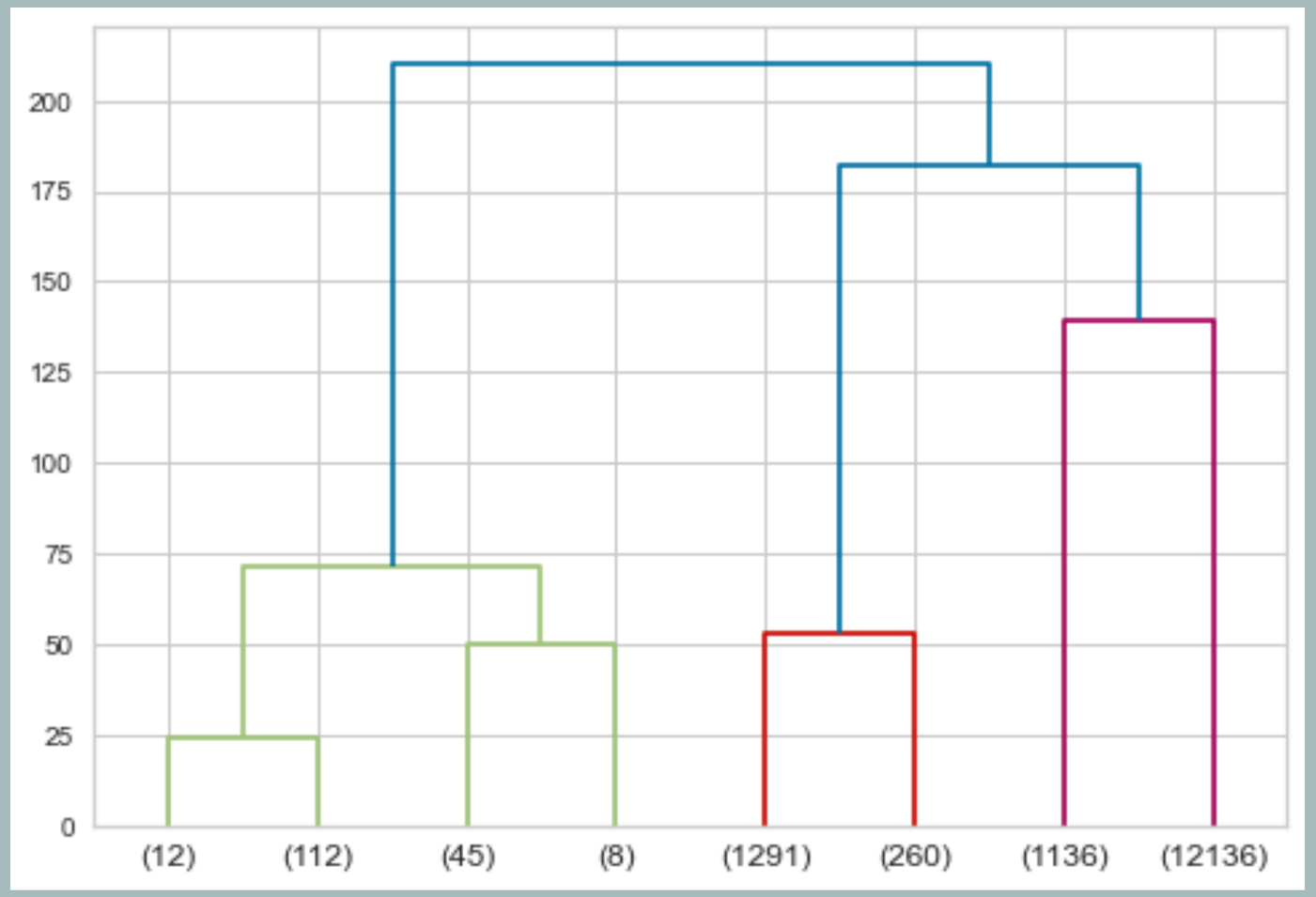


➡ Utilisation PCA ?

MODÉLISATIONS EFFECTUÉES

ESTIMATION DU NOMBRE DE SEGMENTS

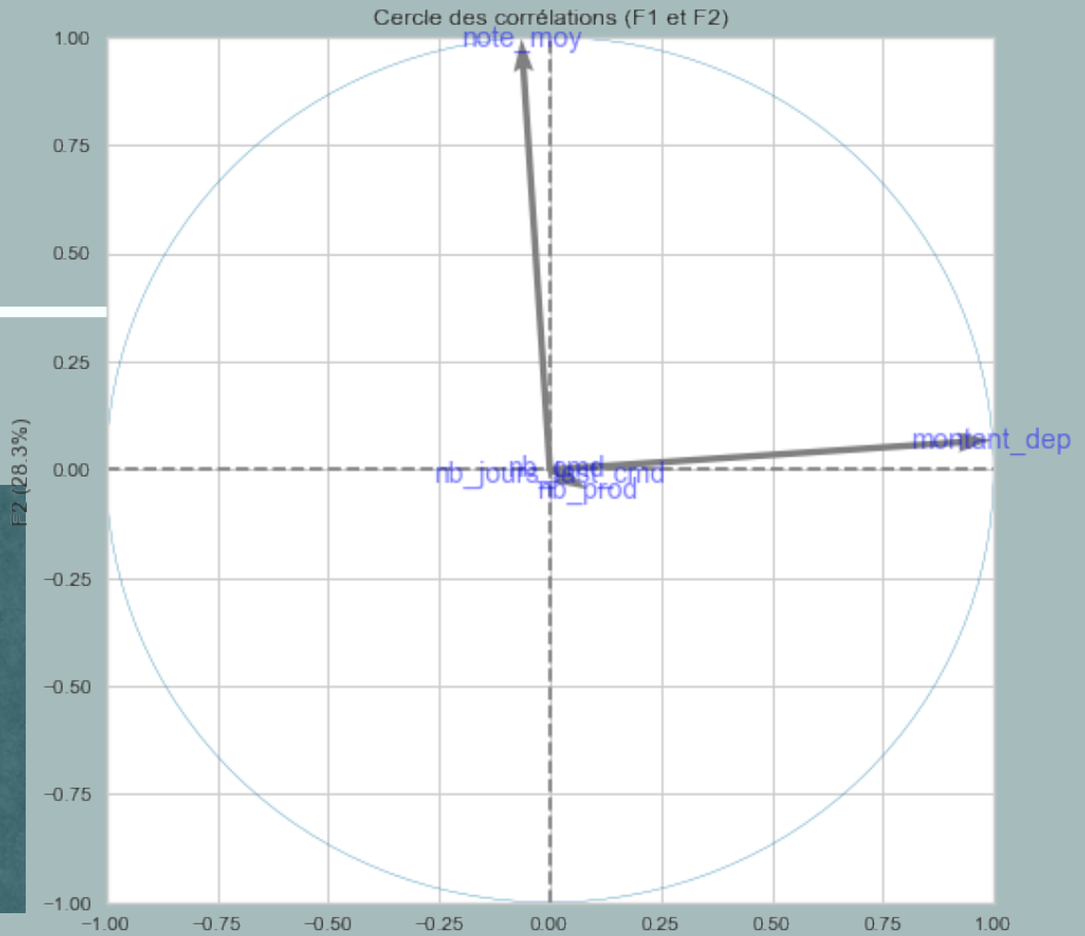
- Echantillon **réduit** **(15000 x 5)** pour raison performance
- **Dendrogramme** (méthode de Ward)
- 3 ou 4 clusters



INFORMATIONS APPORTÉES PAR PCA

➤ *Cercles des corrélations*

- ➤ Projection des 5 variables selon composantes PCA
- ➤ 93.6% de la variance expliquée avec 3 composantes
- ➤ 3 variables se détachent :
  - ***montant_dep***
  - ***note_moy***
  - ***nb_jours_last_cmd***

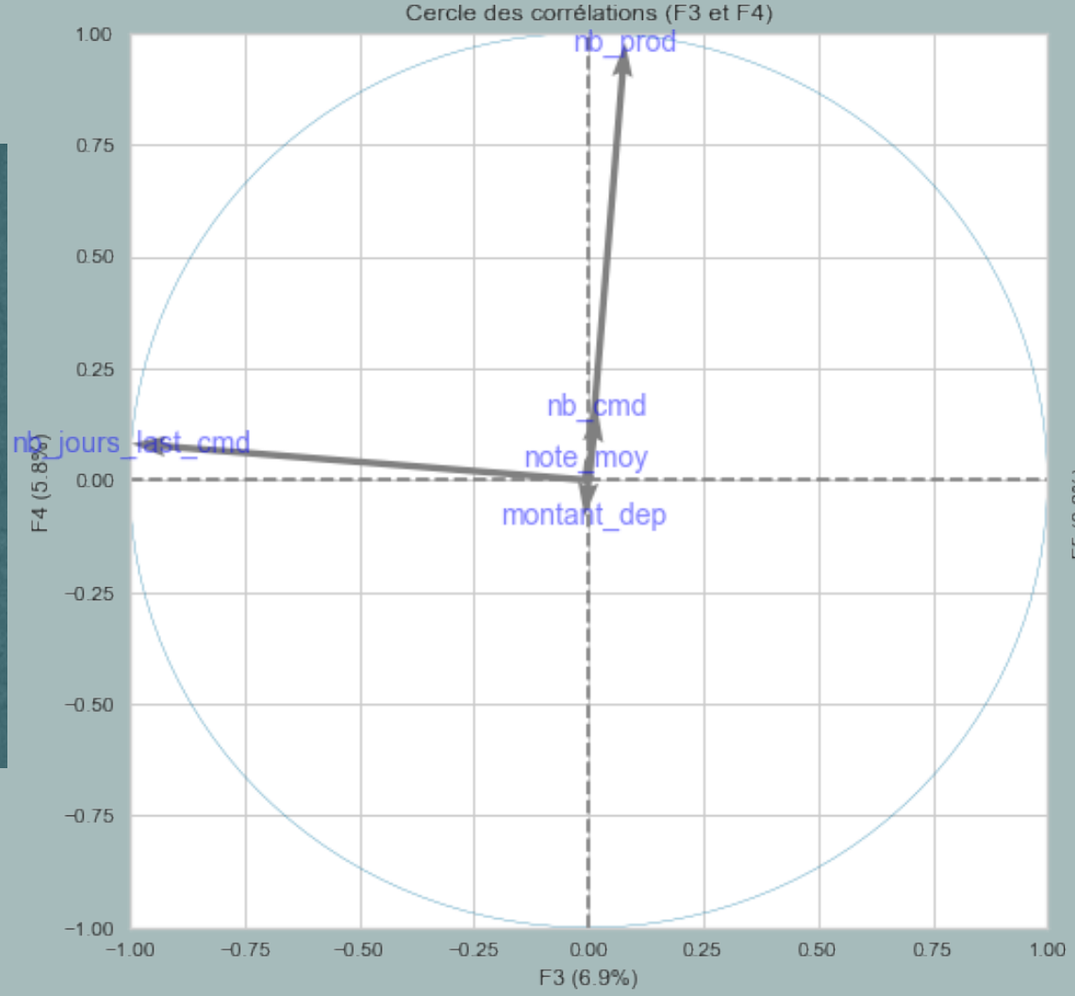


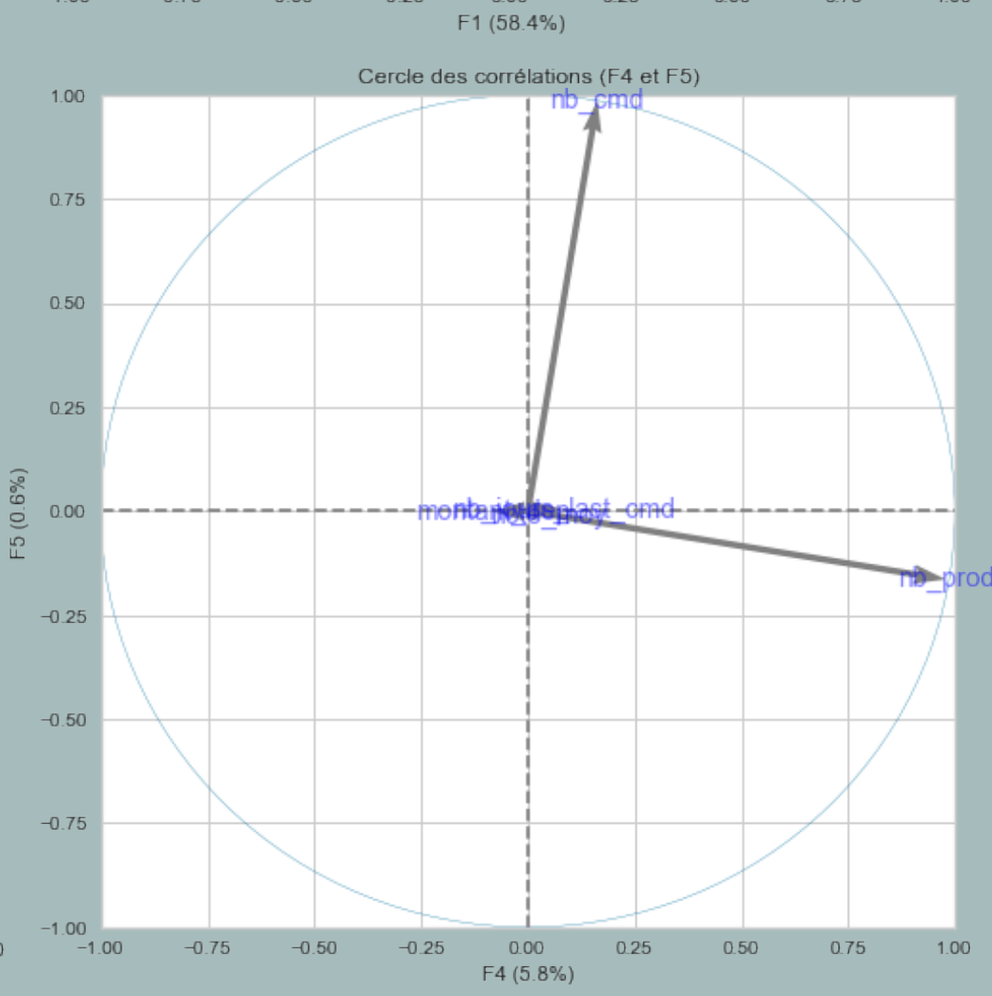
➡ Comparaison de 3 jeux de données

✓ 5 variables

✓ 5 variables sur 3 axes PCA

✓ 3 variables





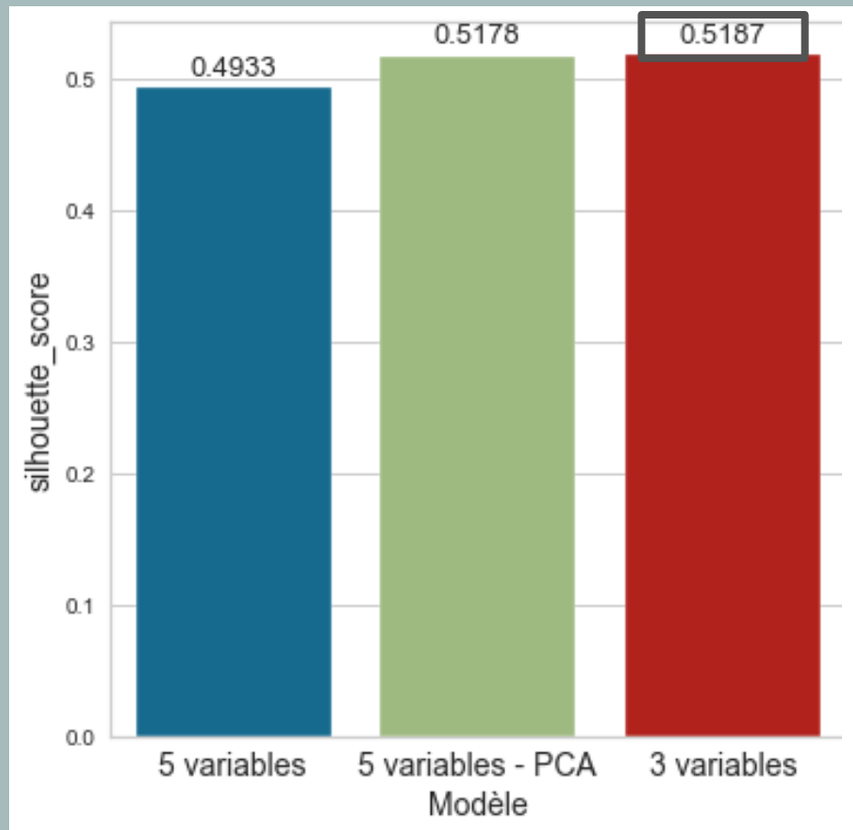
MODÈLES TESTÉS

- **DBSCAN** ➡ Non concluant
  - Raison performance, test en changeant un hyperparamètre à la fois
  - Trop de clusters
  - Clusters très déséquilibrés
- **AGGLOMERATIVE CLUSTERING** ➡ Non concluant
  - Impossible à faire fonctionner sur l'ensemble des données
- **KMEANS** ➡ Concluant pour n = 4 clusters
  - Test pour un nombre de clusters entre 2 et 10
  - Test sur 3 jeux de données
    - ➡ 5 variables ***nb_cmd, montant_dep, nb_prod, note_moy, nb_jours_last_cmd***
    - ➡ 5 variables projetées sur 3 axes PCA
    - ➡ 3 variables ***montant_dep, note_moy, nb_jours_last_cmd***
  - Métriques utilisées
    - ➡ **Score de Silhouette**
    - ➡ **Indice de Davies-Bouldin**
    - ➡ **Score de Calinski-Harabasz**
  - « Elbow method » pour aide au choix du nombre de clusters

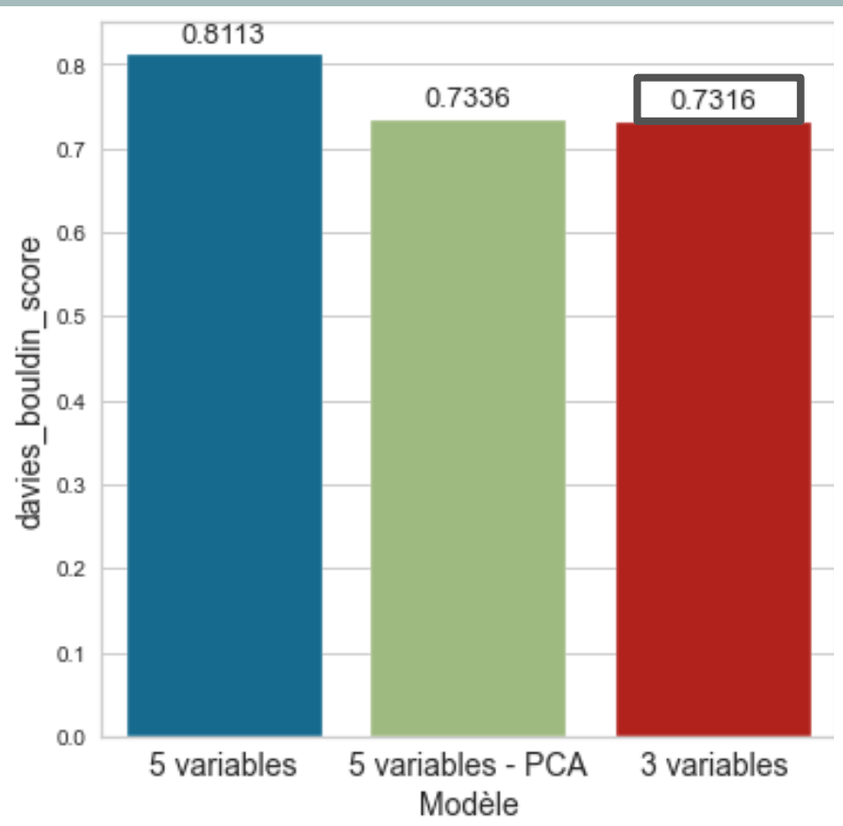
KMEANS 4 CLUSTERS

➤ *Comparaison des métriques*

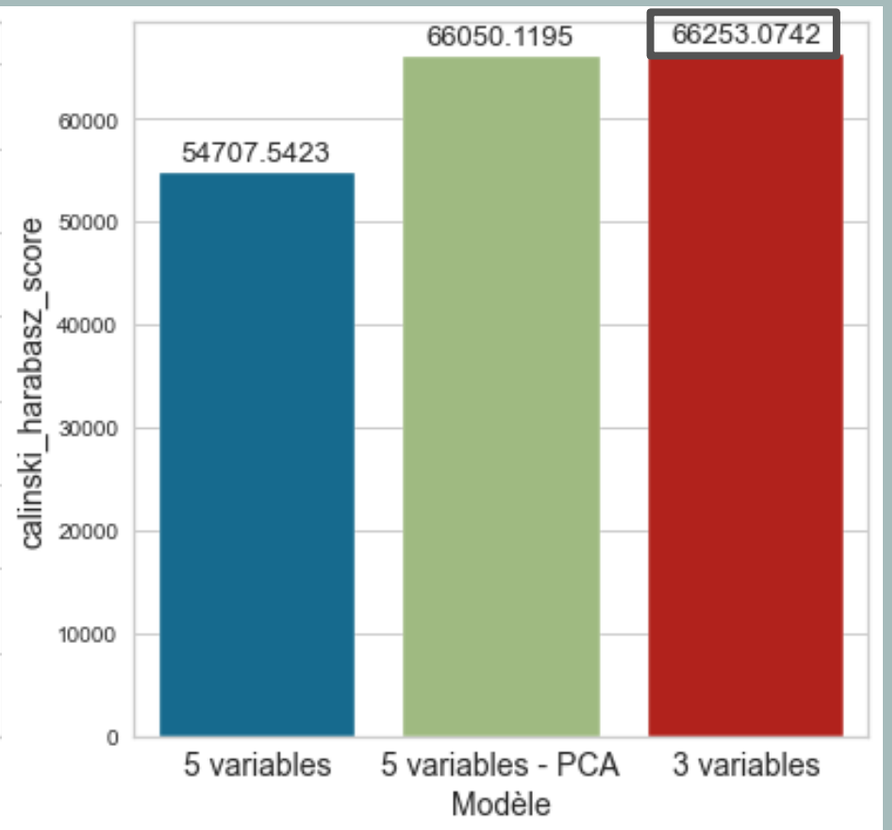
Higher is better



Lower is better



Higher is better

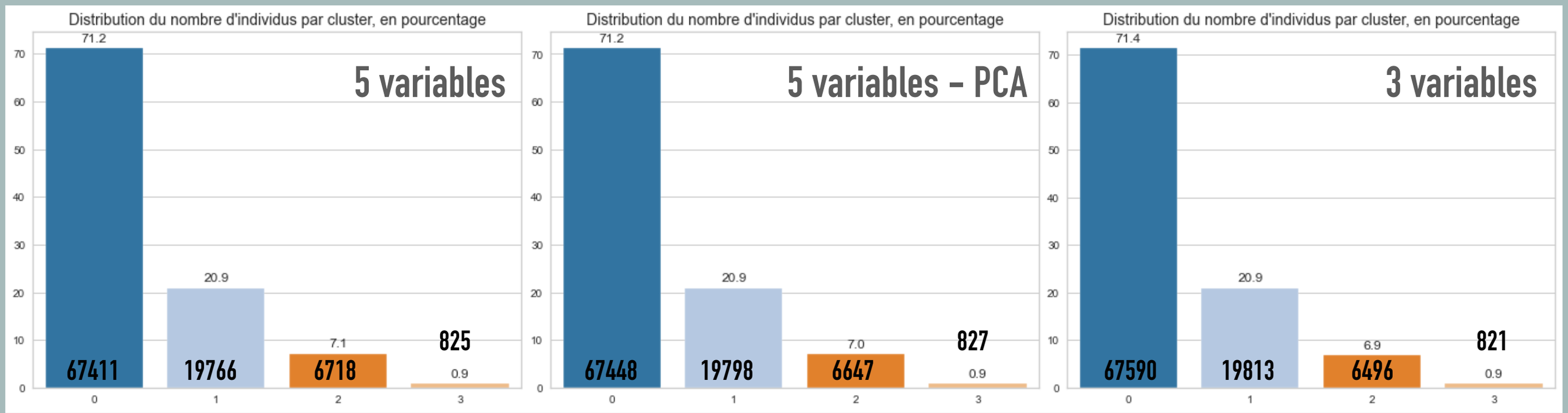


➤ Amélioration des scores en « réduisant » à 3 variables
- Inertie plus faible également

➡ 3 variables

KMEANS 4 CLUSTERS

➤ *Comparaison de la répartition des clients*



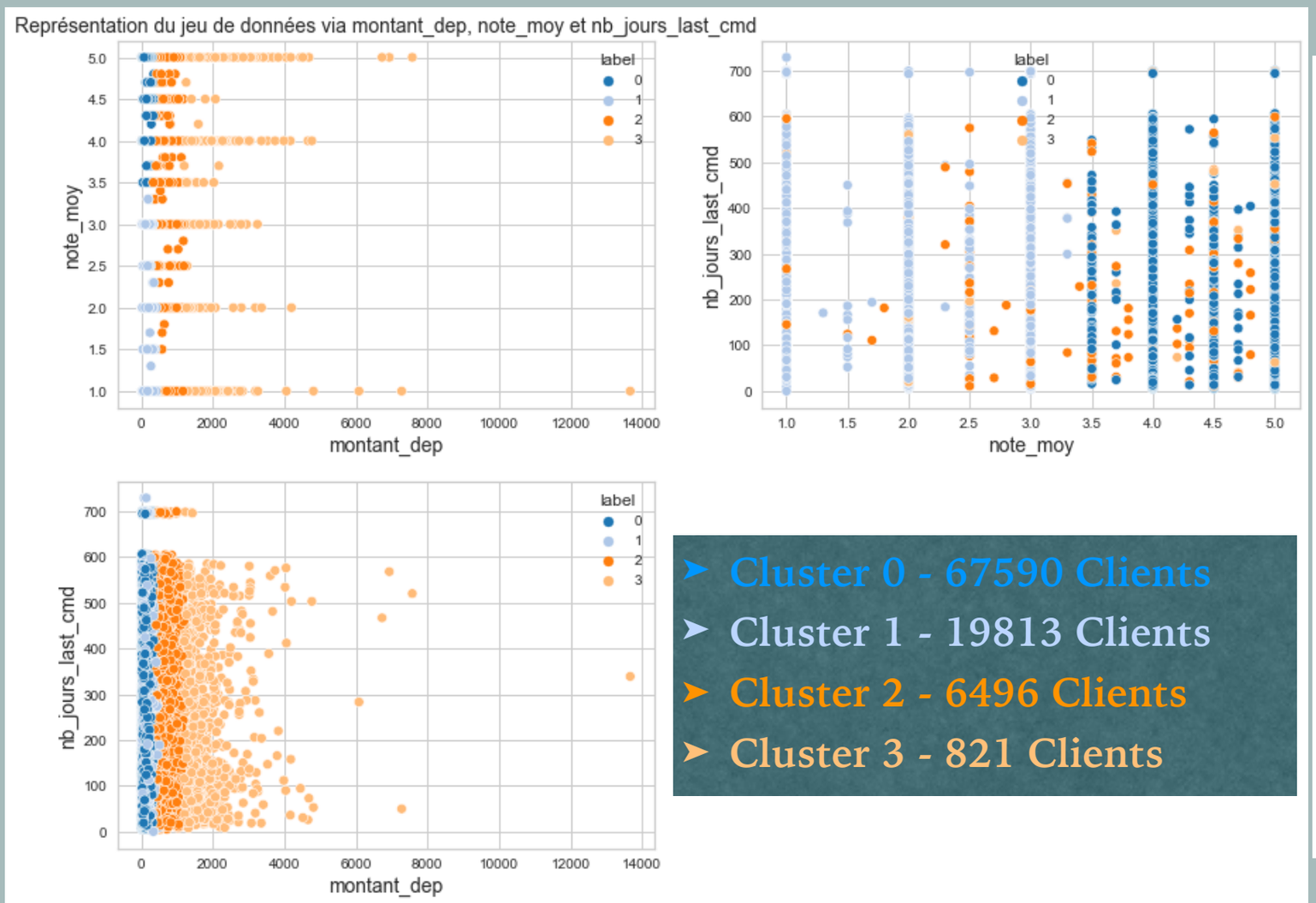
➤ Répartition quasi similaire
- Plus de clients dans les clusters 0 et 1 avec 3 variables

➡ 3 variables

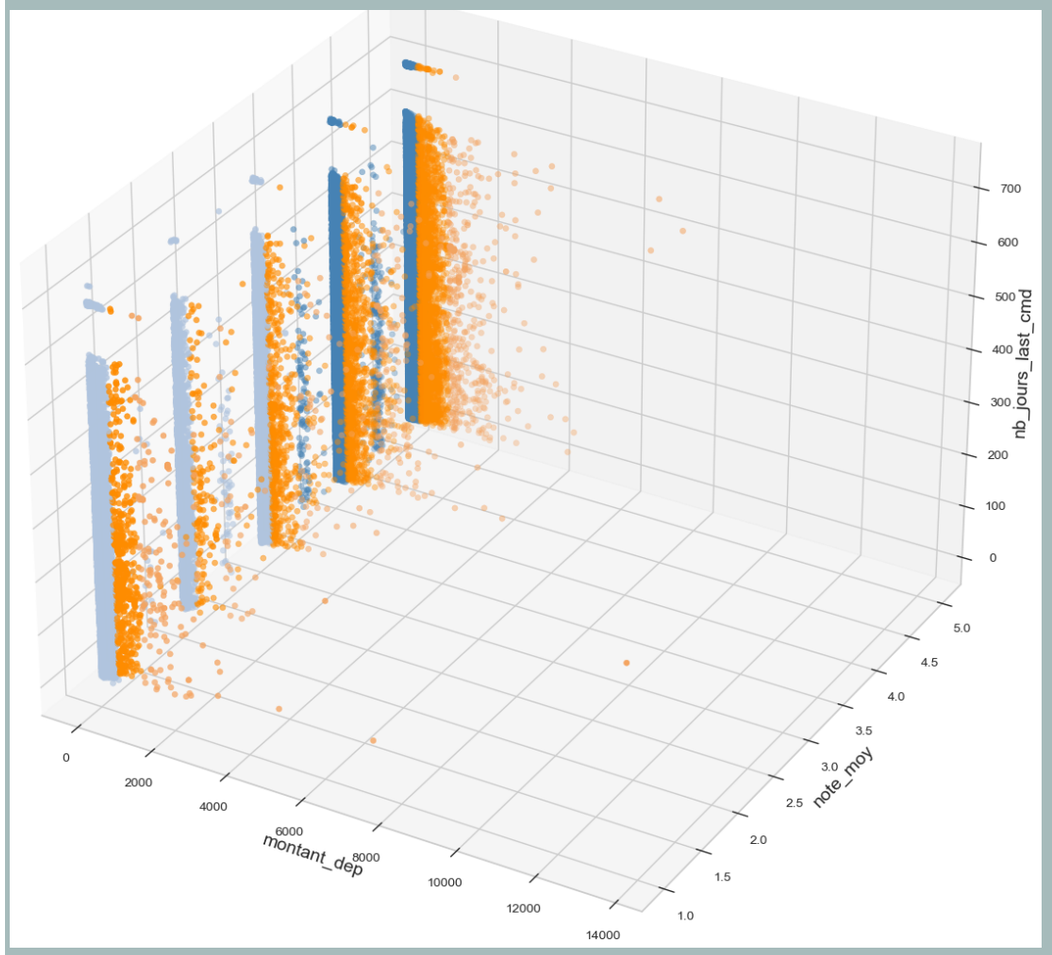
MODÈLE FINAL

KMEANS 4 CLUSTERS – 3 VARIABLES

➤ *Répartition des clients*

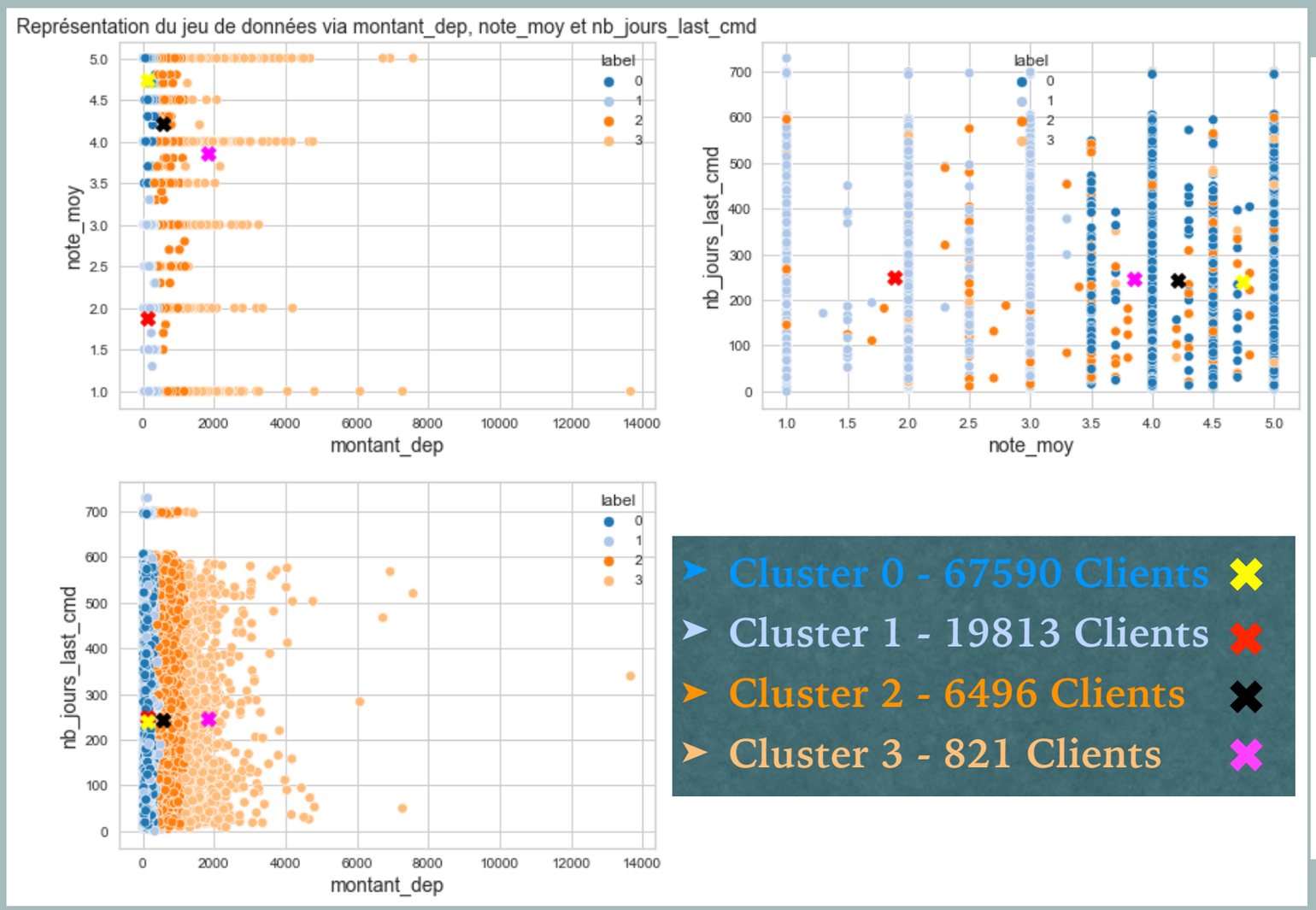


- Cluster 0 - 67590 Clients
- Cluster 1 - 19813 Clients
- Cluster 2 - 6496 Clients
- Cluster 3 - 821 Clients

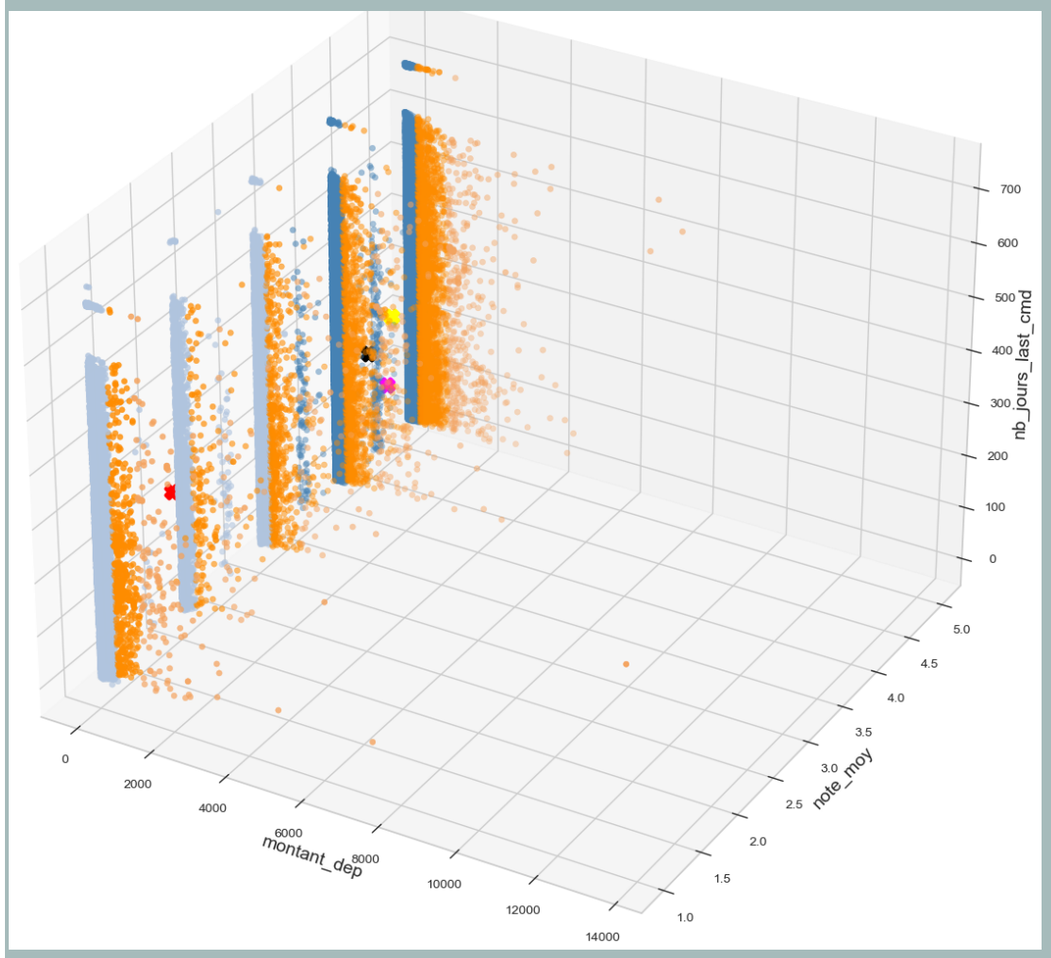


KMEANS 4 CLUSTERS – 3 VARIABLES

➤ *Répartition des clients avec centroids*

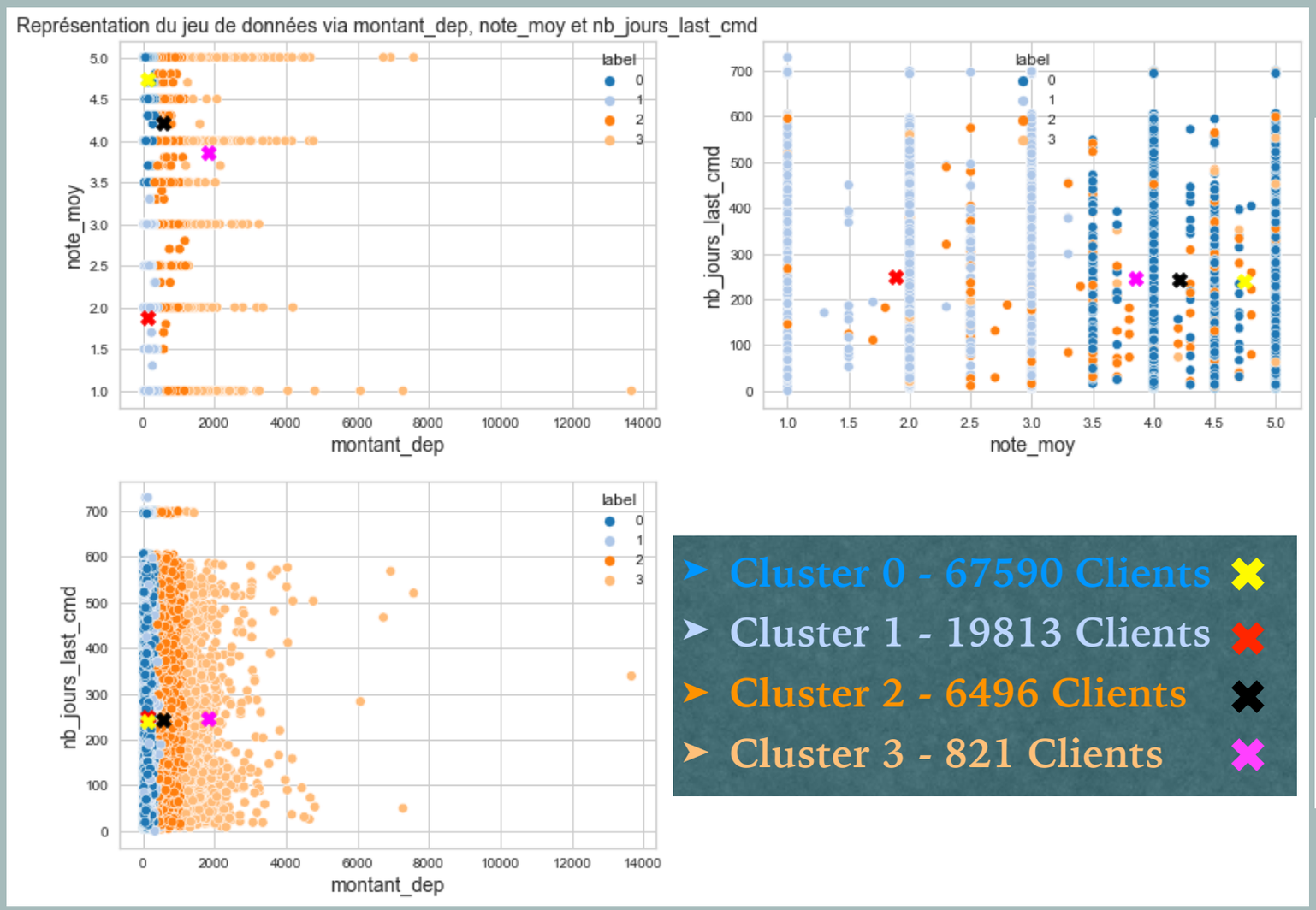


- Cluster 0 - 67590 Clients
- Cluster 1 - 19813 Clients
- Cluster 2 - 6496 Clients
- Cluster 3 - 821 Clients



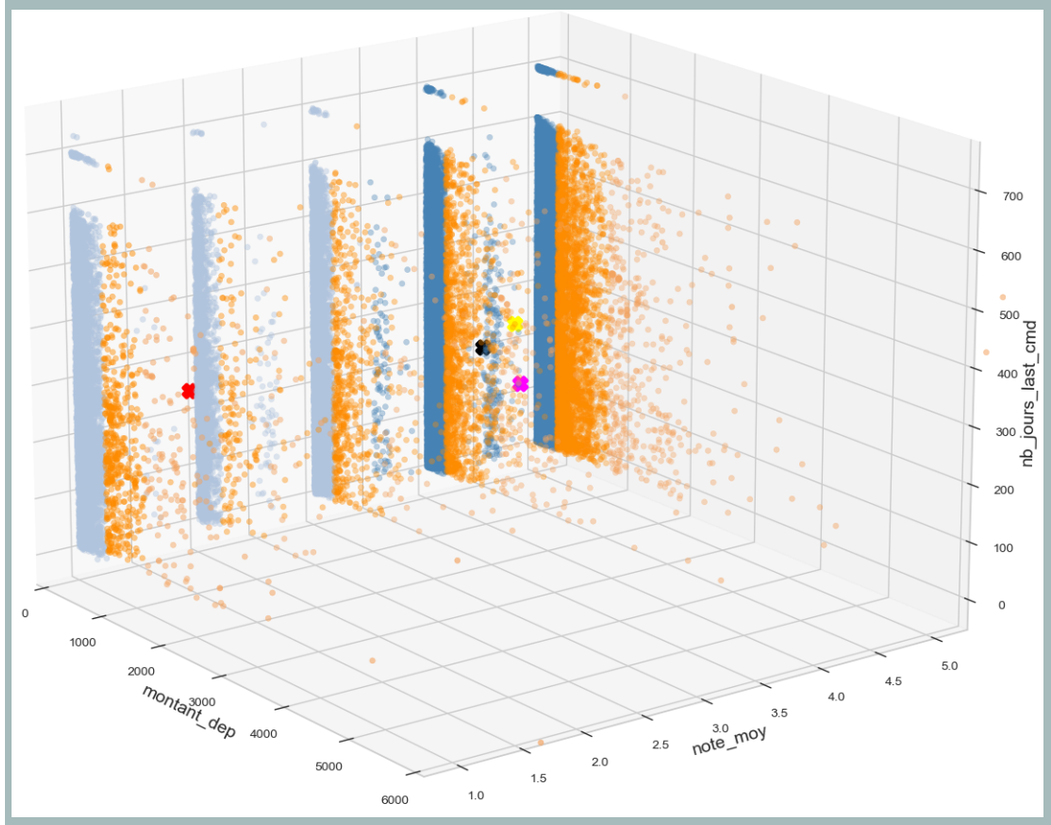
KMEANS 4 CLUSTERS – 3 VARIABLES

➤ *Répartition des clients avec centroids*



- Cluster 0 - 67590 Clients ✖
- Cluster 1 - 19813 Clients ✖
- Cluster 2 - 6496 Clients ✖
- Cluster 3 - 821 Clients ✖

ZOOM (*montant_dep* <6000)



KMEANS 4 CLUSTERS – 3 VARIABLES

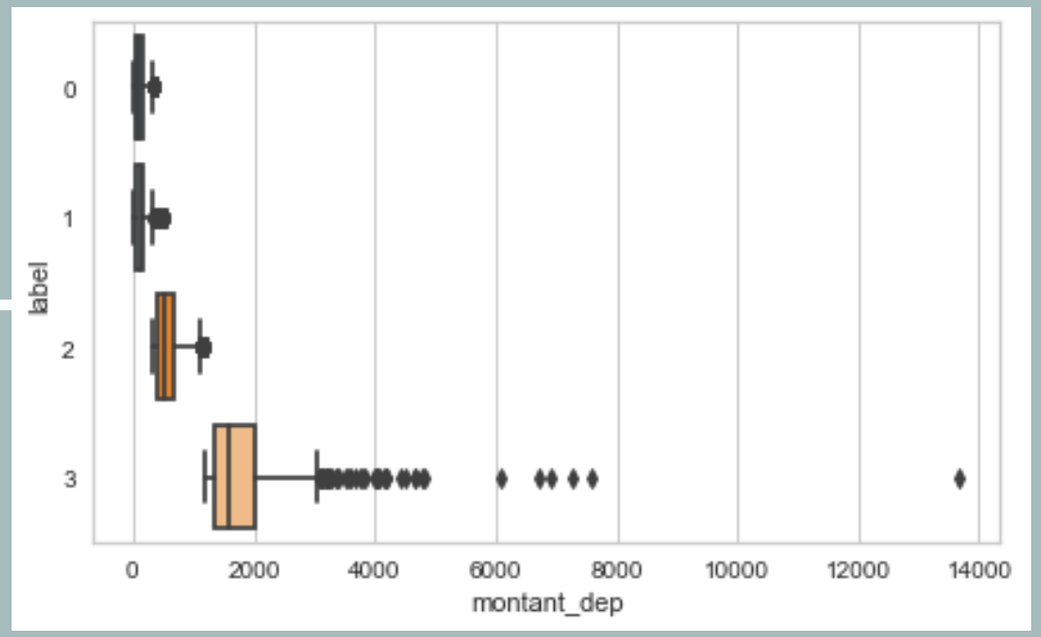
➤ *Caractéristiques des clusters*

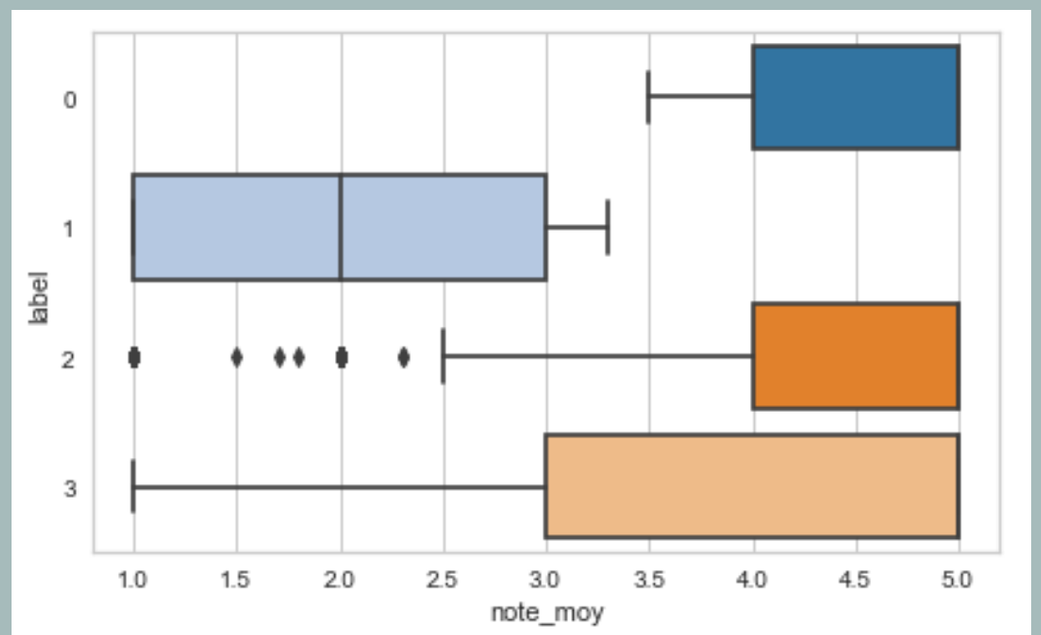
➤ **Cluster 0 - 67590 Clients**
- **Montant dépensé faible** (< 360 euros ; moy ~ 120 euros)
- **Les plus satisfaits** (moy 4.7)
- 2.5% avec au moins 2 commandes
- 9.4% avec au moins 2 produits commandés

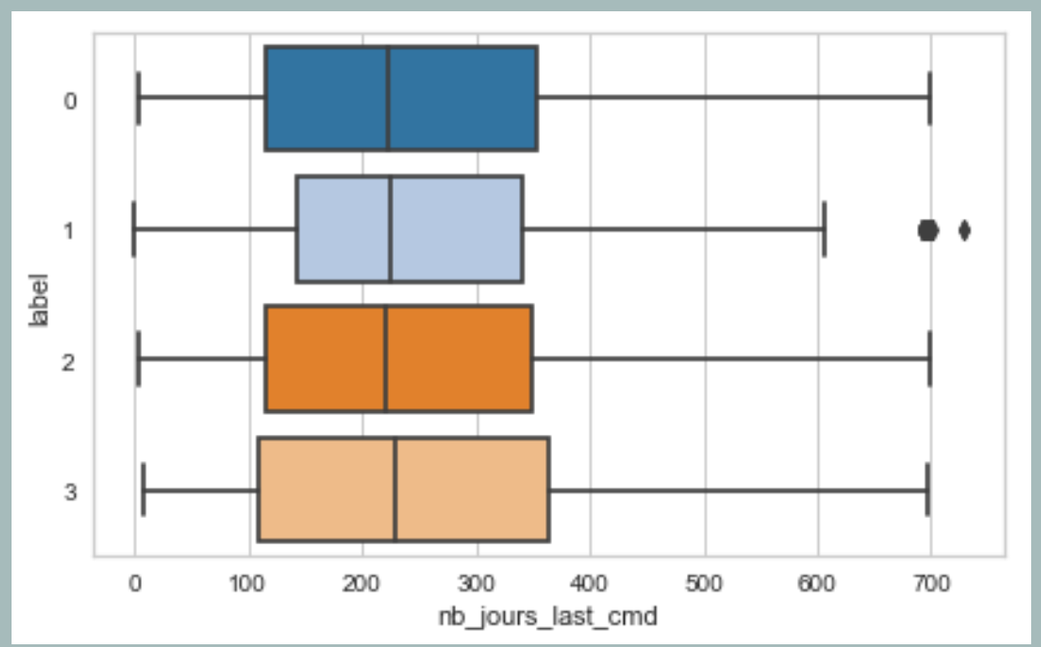
➤ **Cluster 1 - 19813 Clients**
- **Montant dépensé faible** (< 510 euros ; moy ~ 130 euros)
- **Les moins satisfaits** (moy 1.9)
- 2.4% avec au moins 2 commandes
- 17.1% avec au moins 2 produits commandés

➤ **Cluster 2 - 6496 Clients**
- **Montant dépensé moyen** (330 euros < < 1200 euros)
- Satisfaction globalement bonne (moy 4.2)
- **10.1%** avec **au moins 2 commandes**
- **28%** avec au moins 2 produits commandés

➤ **Cluster 3 - 821 Clients**
- **Montant dépensé élevé** (> ~ 1200 euros)
- Satisfaction assez bonne (moy 3.8)
- **7.3%** avec **au moins 2 commandes**
- **23.8%** avec **au moins 2 produits commandés**







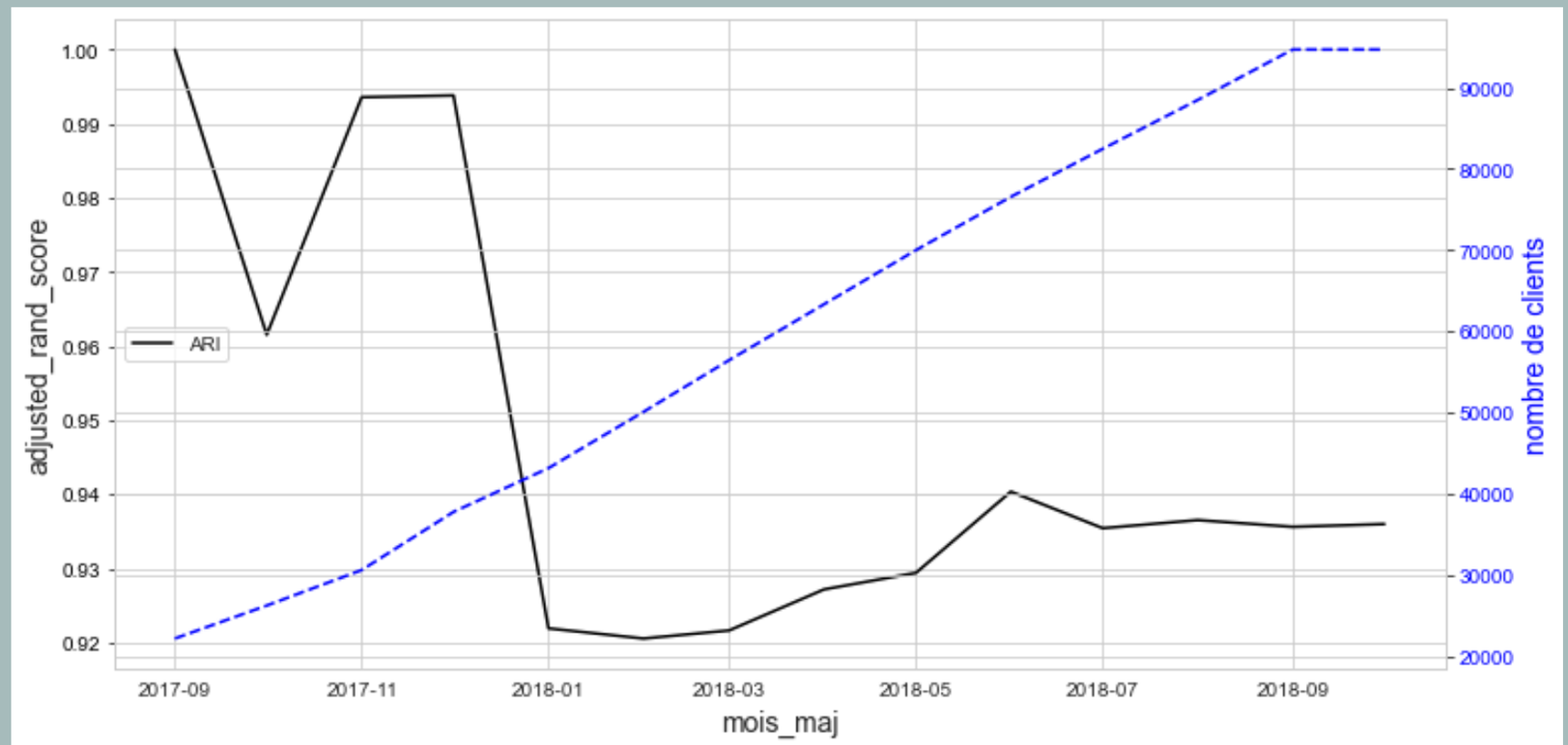
DELAI DE MAINTENANCE

MÉTHODE

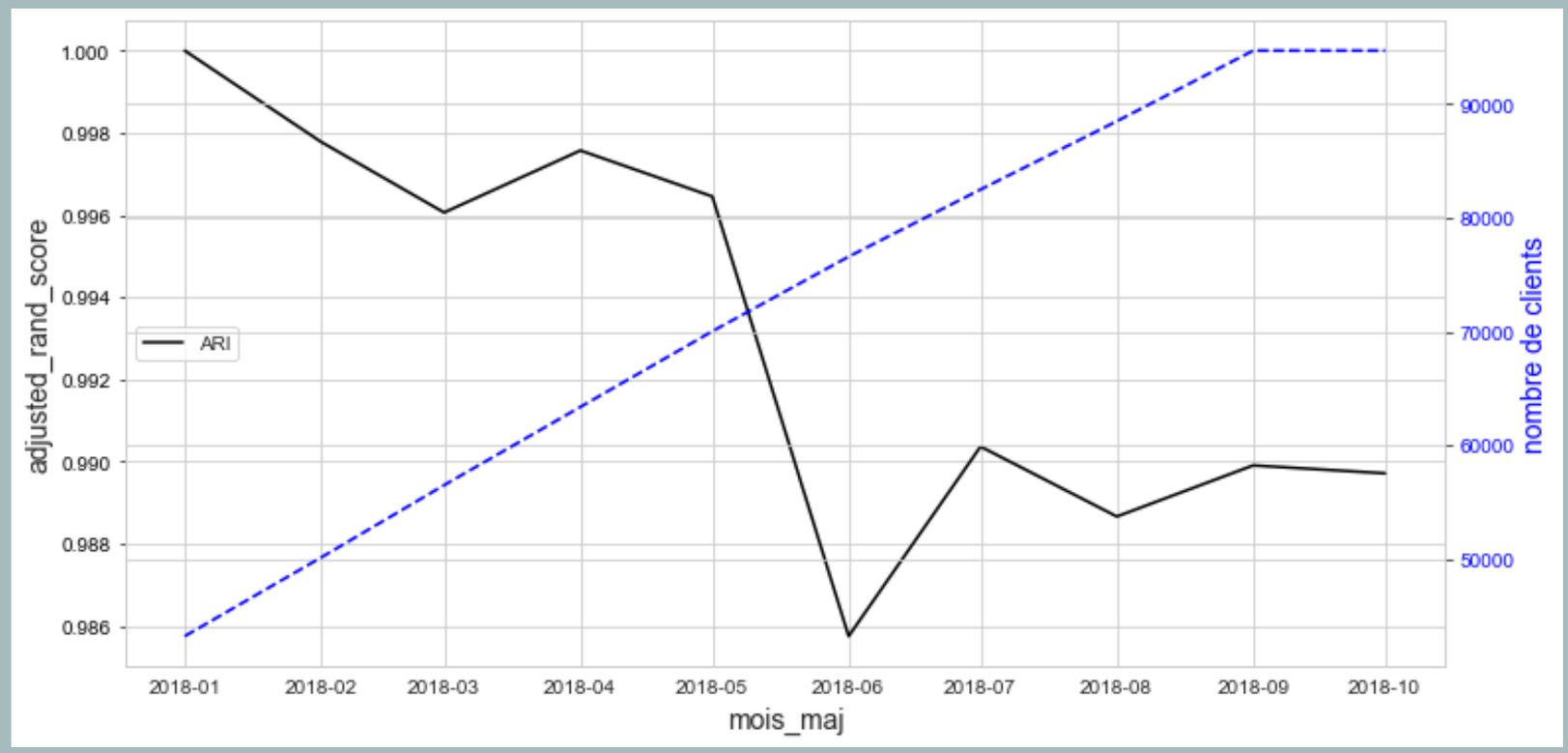
- Utilisation du **modèle précédent**
  - Non supervisé
  - **Base de la segmentation**
    - ➡ Modèle classification (supervisé)
- **Modèle de base**
  - Entrainé sur **période de base donnée**
- Ajout de nouveaux clients
  - Par **période régulière : tous les mois**
  - Entrainement du **modèle** sur nouvelle période
    - ➡ Prédiction **« vraie »** des clusters
  - Prédiction du **modèle de base** sur nouvelle période
  - **Comparaison des prédictions**
    - ➡ Via **ARI** (Adjusted Rand Index)
    - ➡ Si la valeur **diminue trop**, **mise à jour** du modèle **nécessaire**

RÉSULTATS

- **Période de base : Sept 2016-Août 2017**
  - Valeurs ARI assez élevées
  - Evolution quasi linéaire du nb de clients
  - Chute au bout de **4 mois**
    - ➡ Mise à jour



- **Période de base : Sept 2016-Déc 2017**
  - Valeurs ARI très élevées
  - Evolution linéaire du nb de clients
  - Chute (relative) au bout de **5 mois**
    - ➡ Mise à jour



CONCLUSION

CONCLUSION

- Modèle pour segmenter clients de chez OLIST
  - **KMeans**
  - **4 clusters** de clients
  - 3 variables
  - Robuste
- Fréquence de maintenance
  - 4 à 5 mois
    - Valeur seuil
    - **Ou** post Noël
- **Piste d'amélioration** :
  - Inclusion d'autres variables
    - Localisation
    - Moyen de paiement
    - Catégorie de produits

➡ Affiner le « gros » cluster

MERCI

QUESTIONS