SOUTENANCE

SEGMENTATION MARKETING

Cédric Dietzi

Sommaire

- Contexte
- Jeu de données
- Feature engineering
- Evaluation des modèles
- Conclusion

Contexte

La société OLIST souhaite disposer d'un outil de segmentation automatique de ses clients

Utilisé par l'équipe marketing

Simple à interpréter

Le jeu de données

Prêt à dépenser

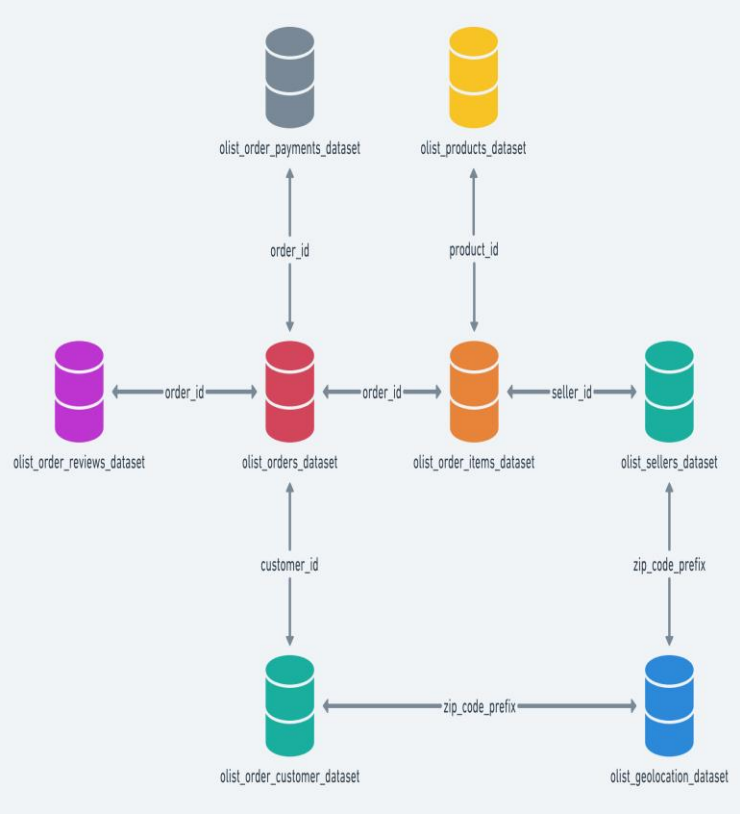
90 000 CLIENTS

8 FICHIERS

50 VARIABLES

INFORMATIONS CRM:
CLIENTS, COMMANDES,
PRODUITS, PRIX, …

Analyse exploratoire des données d'origine

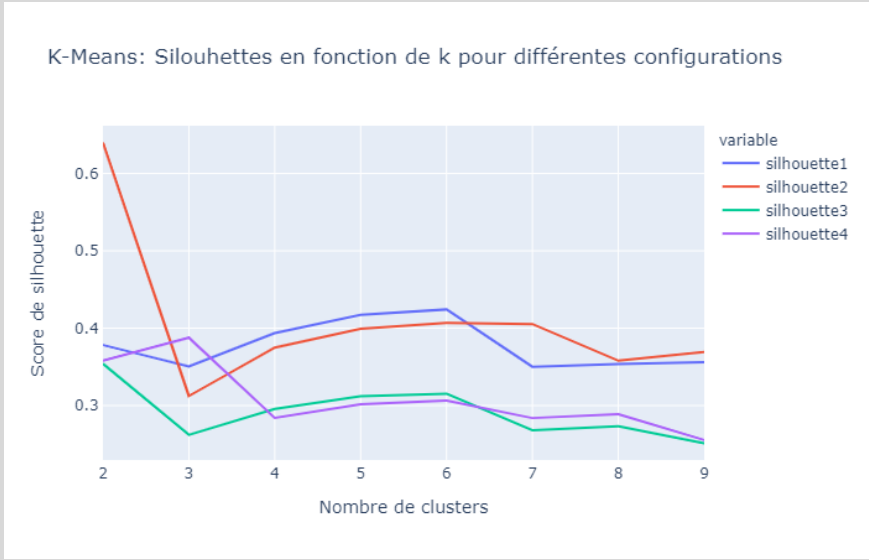


- Exploration des relations entre les données
  - Validation des clés pertinentes pour lier les tables
- Valeurs manquantes: pas impactant
  - Titres et commentaires des revues essentiellement
- Doublons: pas impactant
  - Doublons présents dans la table 'olist_geolocation_dataset'
- Outliers métiers: on n'a pas relevé de valeurs aberrantes sur les données utilisées.

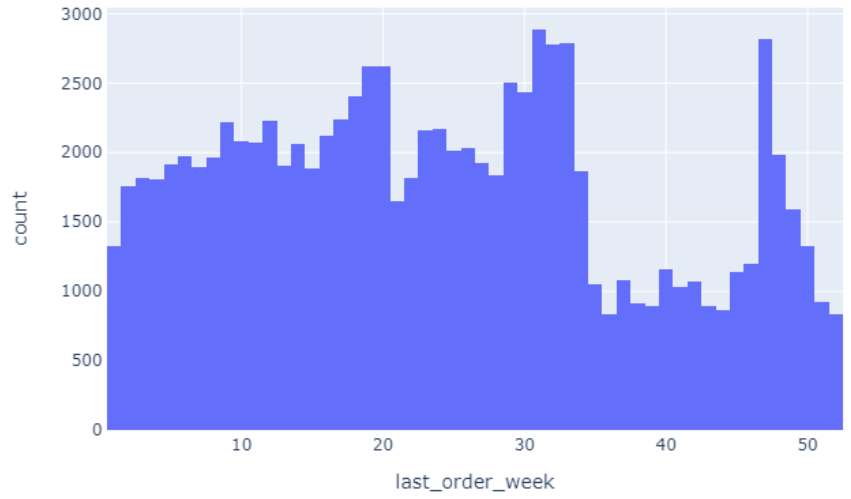
Analyse exploratoire des données construites

- Variables construites pour chaque client:
  - **Nombre de commandes**
  - Temps écoulé depuis la dernière commande
  - Temps écoulé depuis la première commande
  - **N° de semaine de la dernière commande**
  - Durée de la commande à la livraison
  - Ecart entre la date de livraison estimée et réelle
  - **Panier moyen**
  - **Score moyen**

- Sélection des variables construites

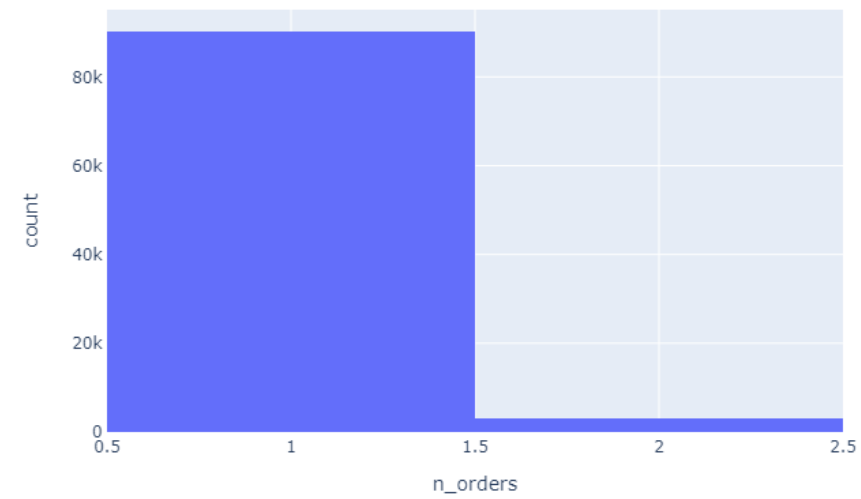


N° de semaine de la dernière commande

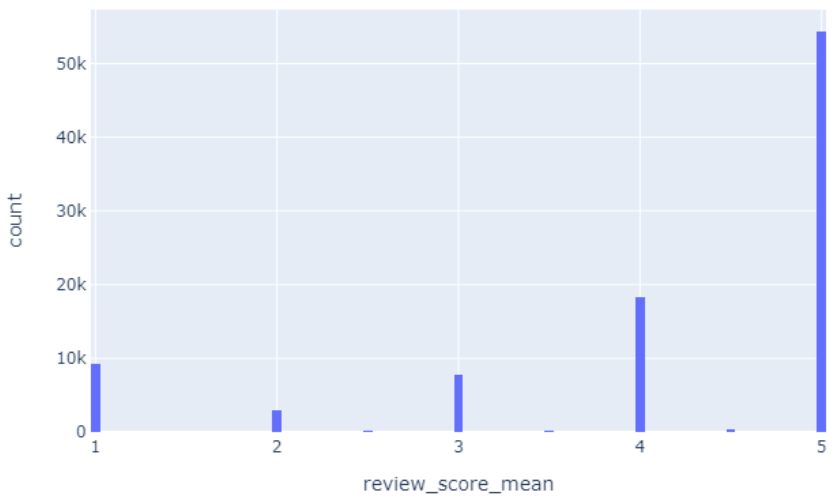


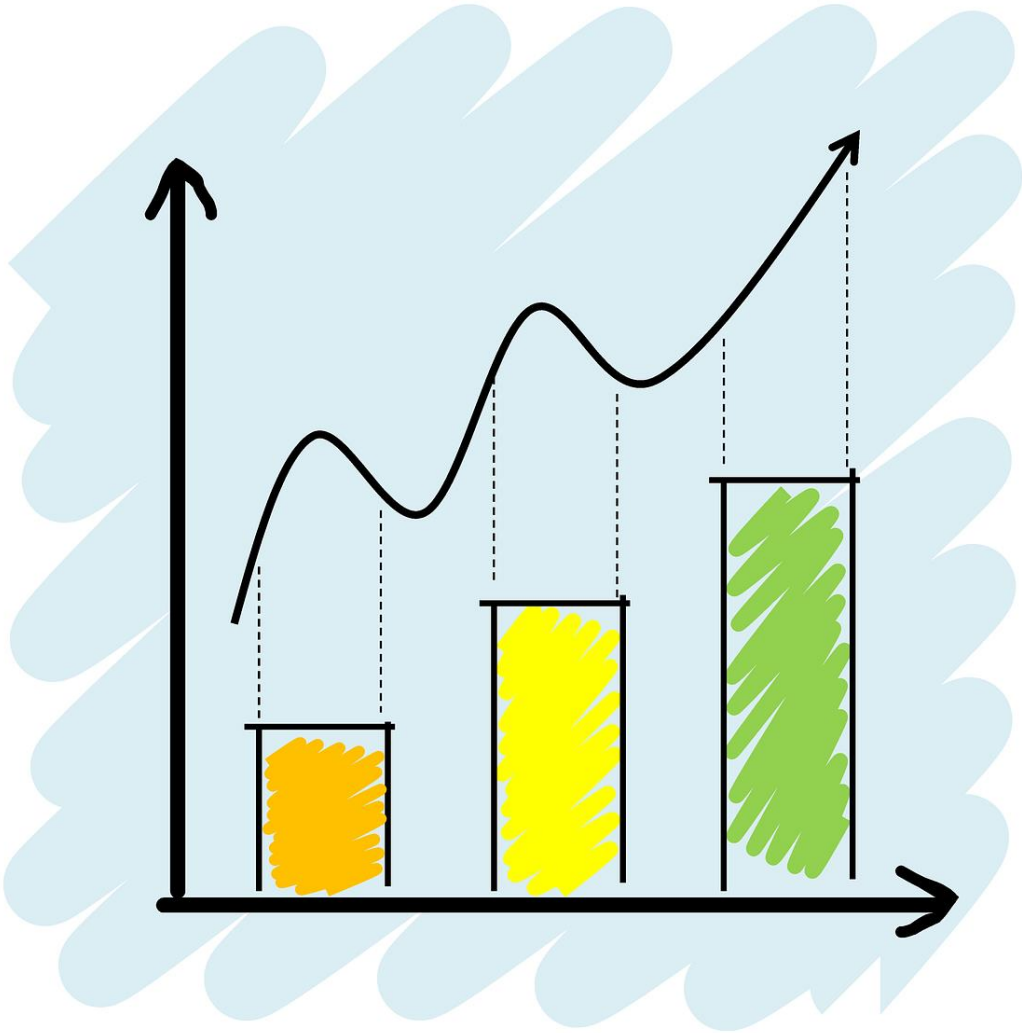
Panier moyen

Nombre de commandes



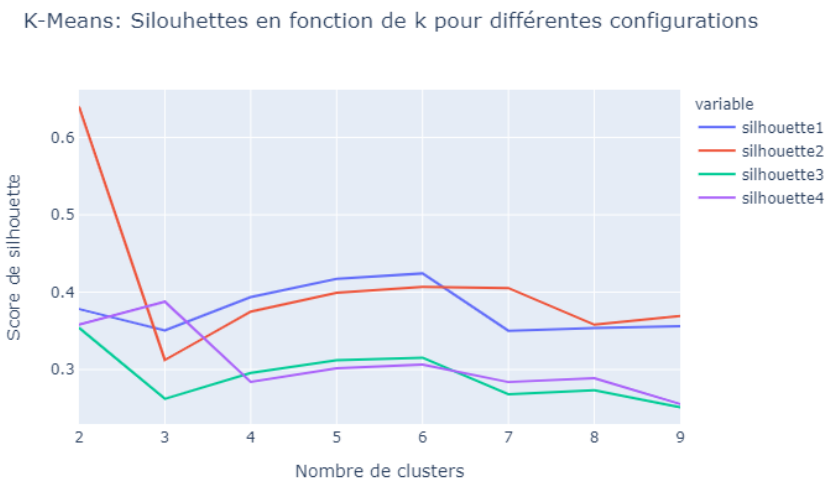
Scores

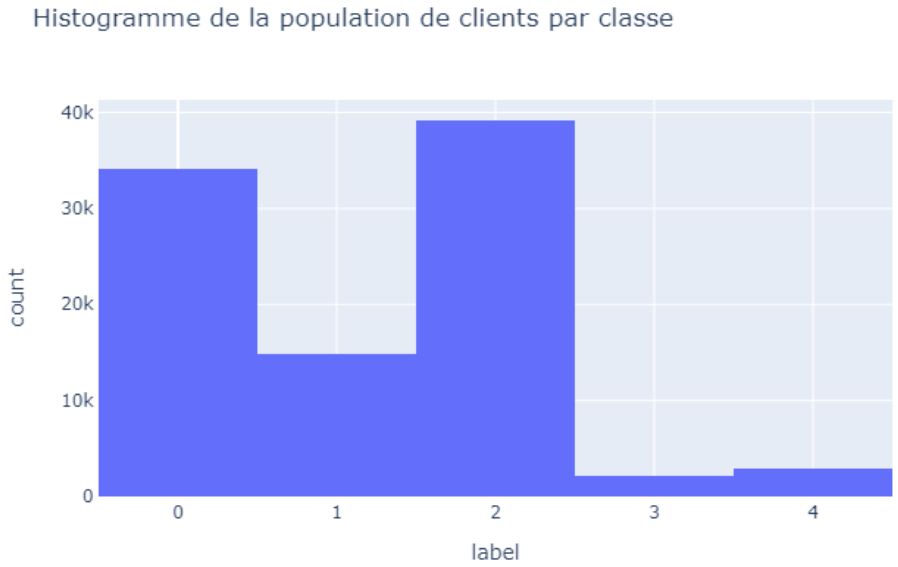




Evaluation des modèles

K-Means
DB-Scan
Agglomerative clustering

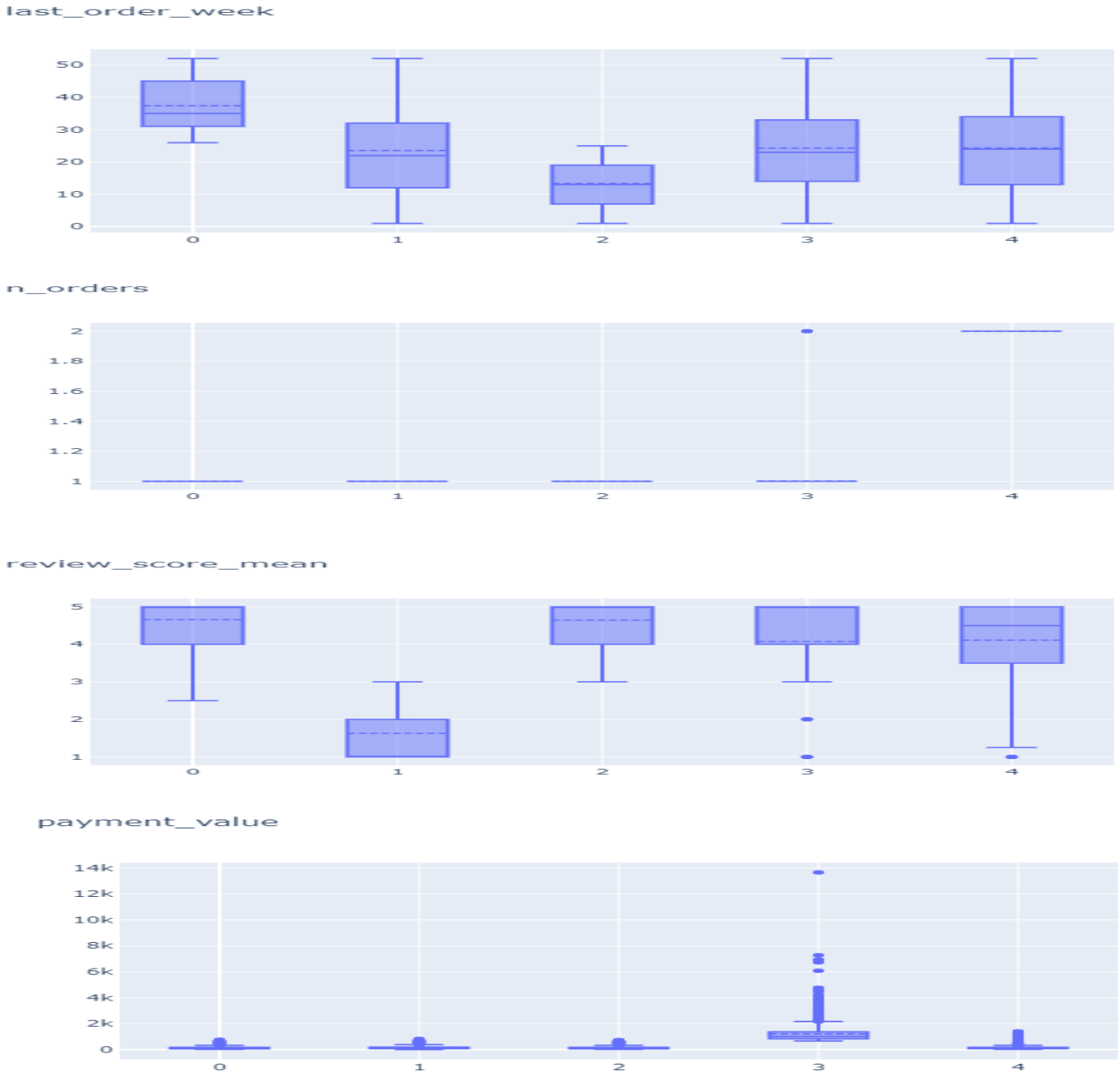




K-Means

Choix du nombre de classes:

1) Score de silhouette élevé
2) Répartition de population pertinente



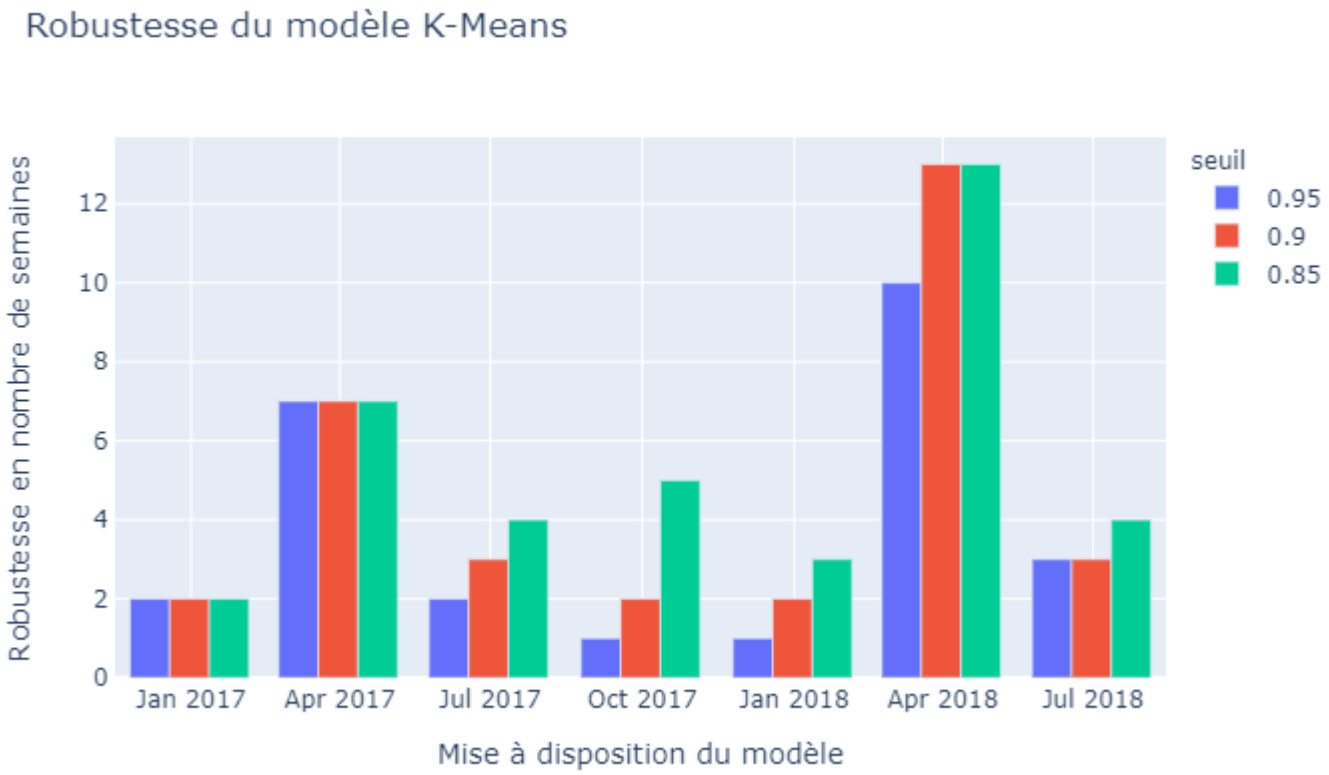
K-Means

Interprétation des classes:

La majorité des clients sont des clients satisfaits qui font un seul achat de faible valeur en début (42%) ou en fin d'année (37%).

Certains clients se distinguent par le fait qu'ils:

- sont insatisfaits (16%)
- ou, commandent plusieurs fois (3%)
- ou, ont un panier moyen plus élevé (2%)

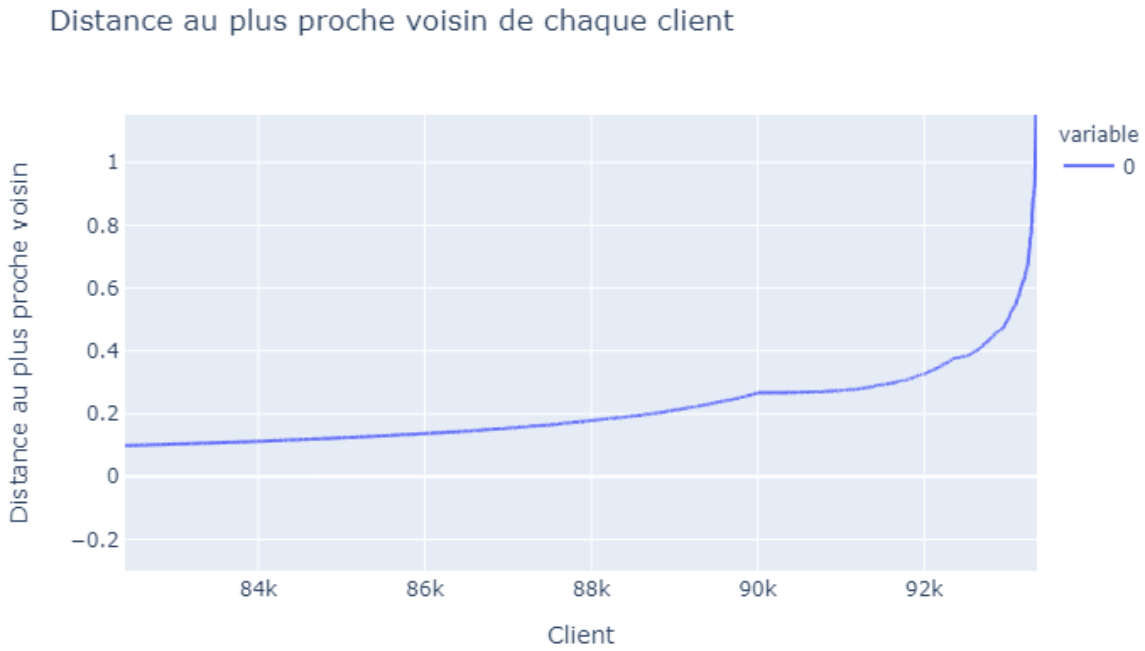


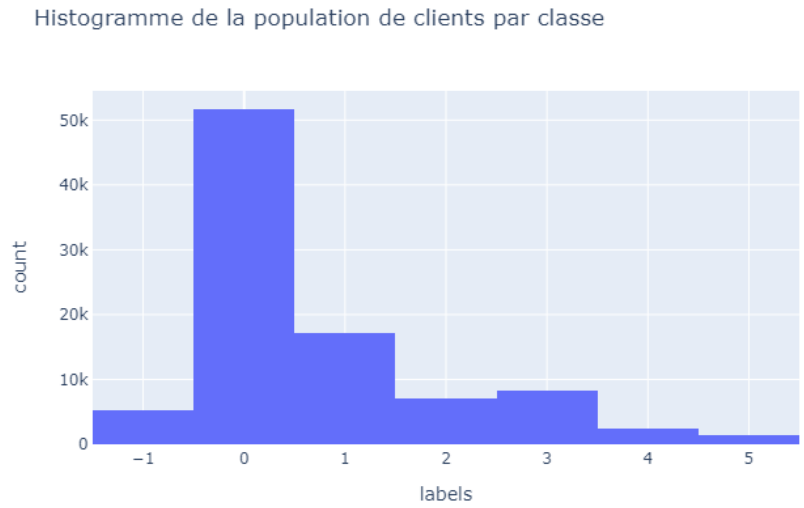
K-Means

Robustesse du modèle de classification:

La robustesse du modèle a été mesurée par le 'Adjusted Rand Score Index'

En mettant le modèle à jour toutes les trois semaines on garantit une bonne qualité de classification avec un score entre 0,8 et 0,9 .





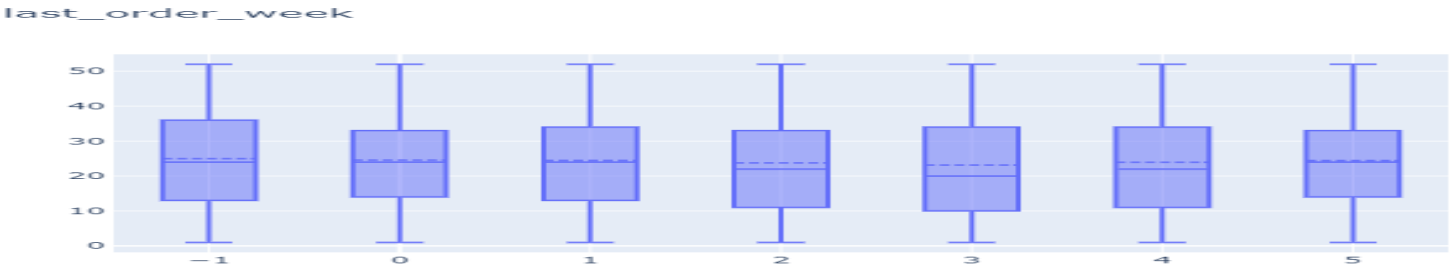
DB-Scan

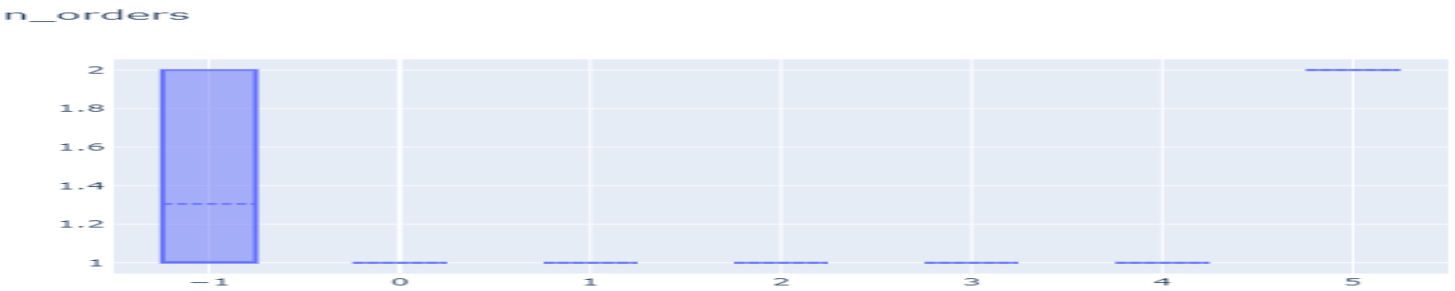
Choix de la résolution (‘epsilon’):

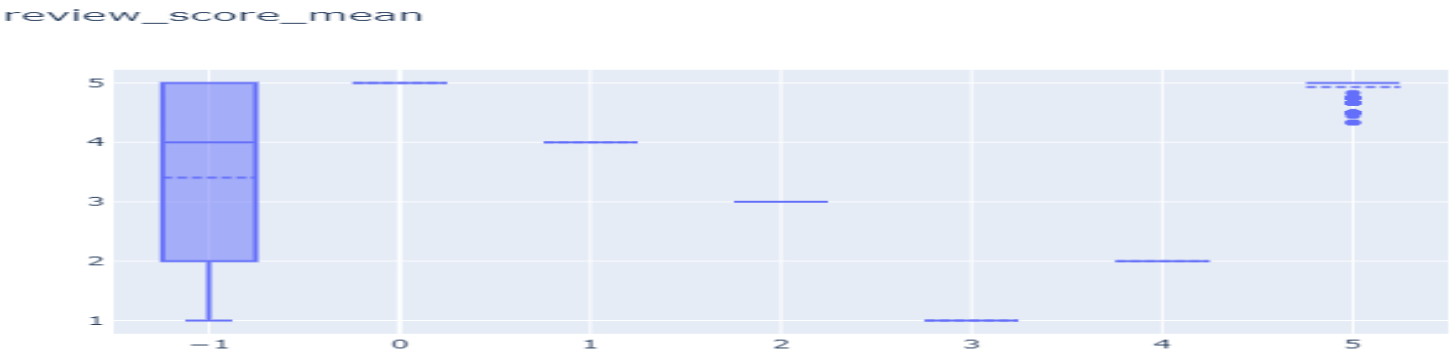
On prend une valeur au niveau du coude de telle sorte à obtenir un nombre de classes suffisant.

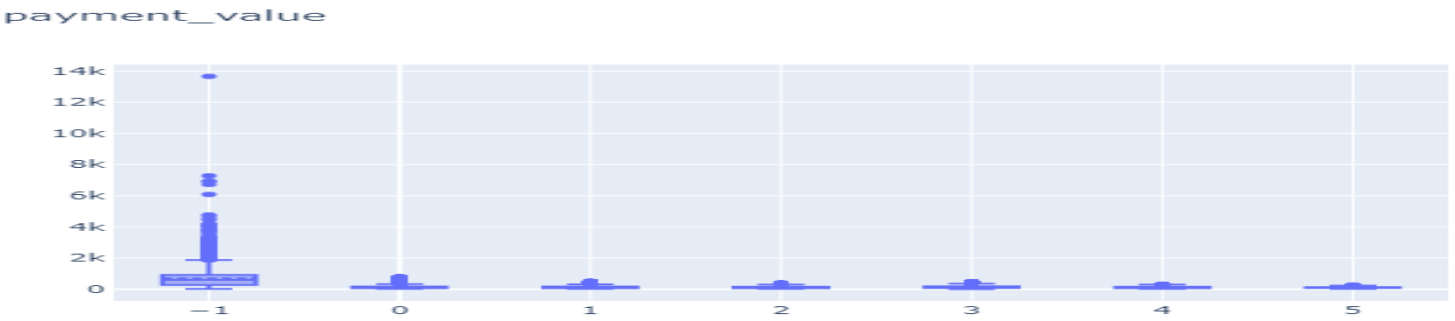
Choix de la densité minimale (‘MinPts’):

Par essai-erreurs pour avoir un nombre de classes pas trop grand tout en gérant les points considérés comme du bruit.







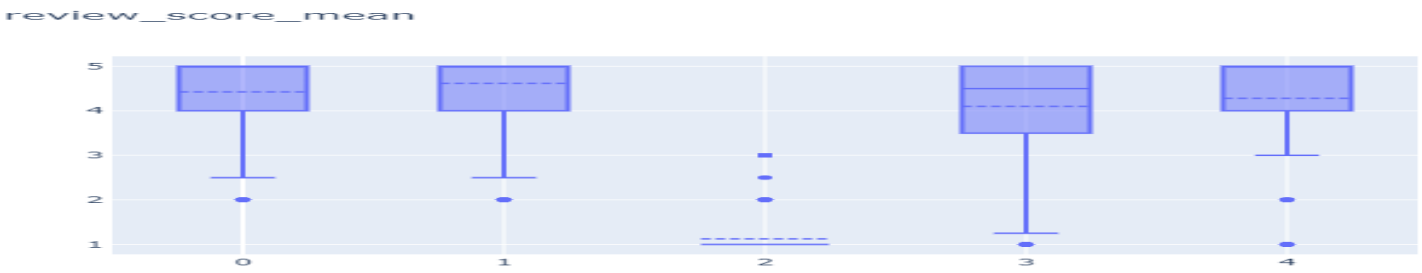


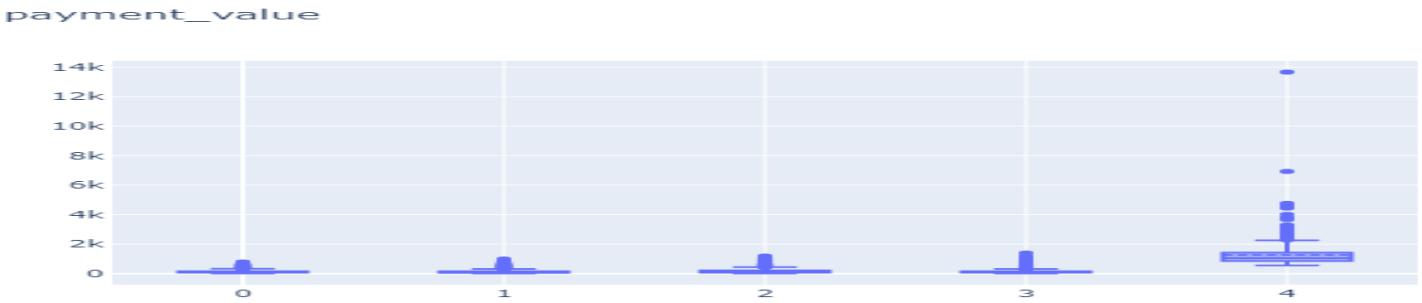
DB-Scan

DB Scan gère mal les jeux de données dans des espaces de densité variables.

C'est le cas ici. DB-Scan détecte un cluster par score et le cluster des clients ayant acheté plusieurs fois.

L'algorithme n'est pas adapté aux données.





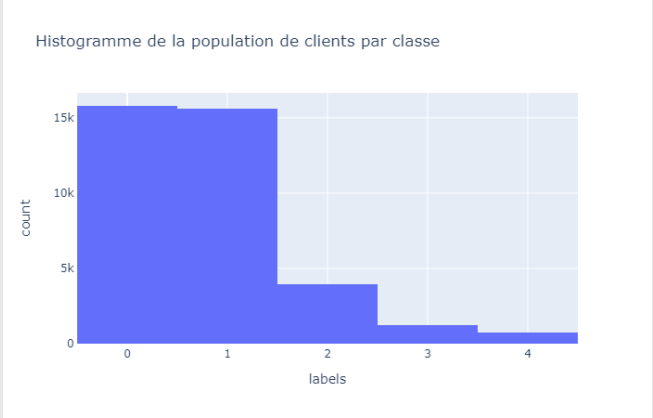
Agglomerative clustering

Classification proche des résultats K-Means.

Métrique utilisée: Ward

5 classes.

Lent => On préfère K-Means.



Conclusion

- Nous fournissons un modèle de clustering des données par l'algorithme du K-Means.
- Les autres algorithmes testés sont inadaptés aux données ou plus lents.
- Le clustering est effectué sur des variables métiers construites. Elles décrivent des aspects temporels, de fréquence d'achat, de montant payé et de satisfaction.
- En segmentant les clients en 5 classes, on obtient un bon compromis entre qualité du clustering et interprétabilité.
- Le clustering a date indique que la majorité des clients sont des clients satisfaits qui font un seul achat de faible valeur en début (42%) ou en fin d'année (37%). Certains clients se distinguent par le fait qu'ils: sont insatisfaits (16%), commandent plusieurs fois (3%), ont un panier moyen plus élevé (2%).
- Les clusters sont robustes sur des périodes de trois semaines, après quoi, le modèle doit être mis à jour.