MANUEL MARTIN - 01/2023 - PROJET 5

SEGMENTATION CLIENTS

Olist

SEGMENTATION CLIENTS

- Objectifs de la mission
- Analyse exploratoire des données
- Modèles de clustering
- Simulation de maintenance
- Conclusion

OBJECTIFS DE LA MISSION

OBJECTIFS DE LA MISSION

- Réaliser une segmentation clients
  - Exploitable et facile d’utilisation par l’équipe Marketing
  - En terme de commandes et de satisfaction
  - Sur l’ensemble des clients

- Fournir une proposition de contrat de maintenance

ANALYSE EXPLORATOIRE DES DONNÉES

ANALYSE EXPLORATOIRE DES DONNÉES

- Jeu de données :
  - 9 fichiers au format .csv

ANALYSE EXPLORATOIRE DES DONNÉES

- Regroupement des order_id de chaque fichier
- Regroupement des zip_code_prefix dans le fichier « geolocation »
- Regrouper tous les fichiers en un seul via les _id
  - order_id, customer_id, product_id, seller_id, product_category_name et zip_code_prefix
- Suppression de certaines colonnes non pertinentes

« QUAND L'INVISIBLE DEVIENT VISIBLE »

FEATURE ENGINEERING

FEATURE ENGINEERING

- Contruction d’une base client
  - Dernières informations de commandes sur les features textuelles
  - Date de première et dernière commande
  - Moyennes de commande sur les données numériques
  - Nombre total d’articles commandés et de commandes
  - Nombre de commentaires

FEATURE ENGINEERING

- Calcul de nouvelles features
  - Délais entre chaque étape de commande
  - Délai de livraison globale
  - Différence entre livraison estimée et livraison réelle
  - Etc...

« TOUT SEUL ON VA PLUS VITE, MAIS À PLUSIEURS ON VA PLUS LOIN... »

CONNAISSANCES MÉTIER

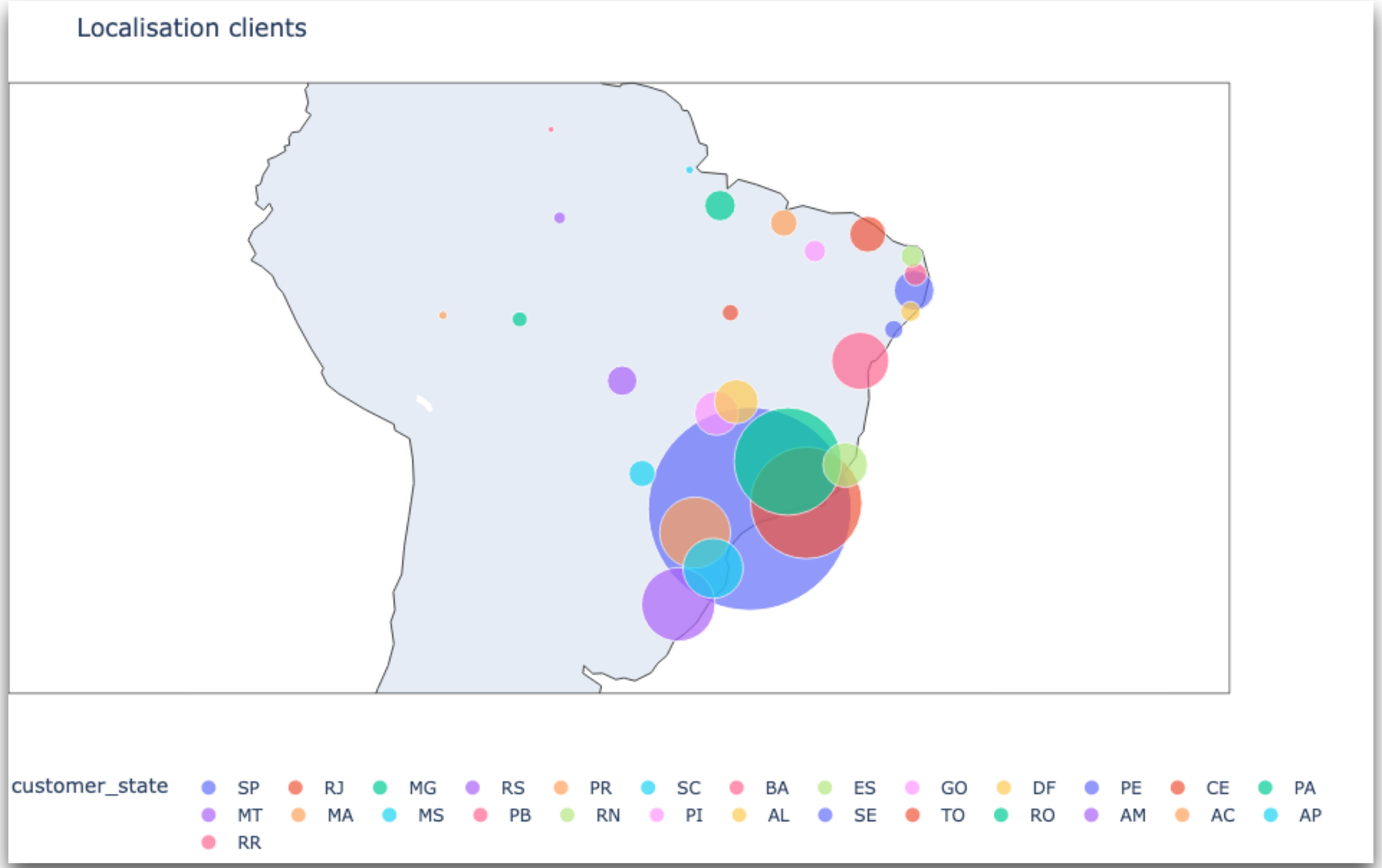
CONNAISSANCES MÉTIER

- RFM
  - Récence : Nb de jours depuis la dernière commande
  - Fréquence : Nb de commandes réalisées sur 2 ans
  - Montant : Moyenne des prix des commandes (coût de la commande + frais de port)

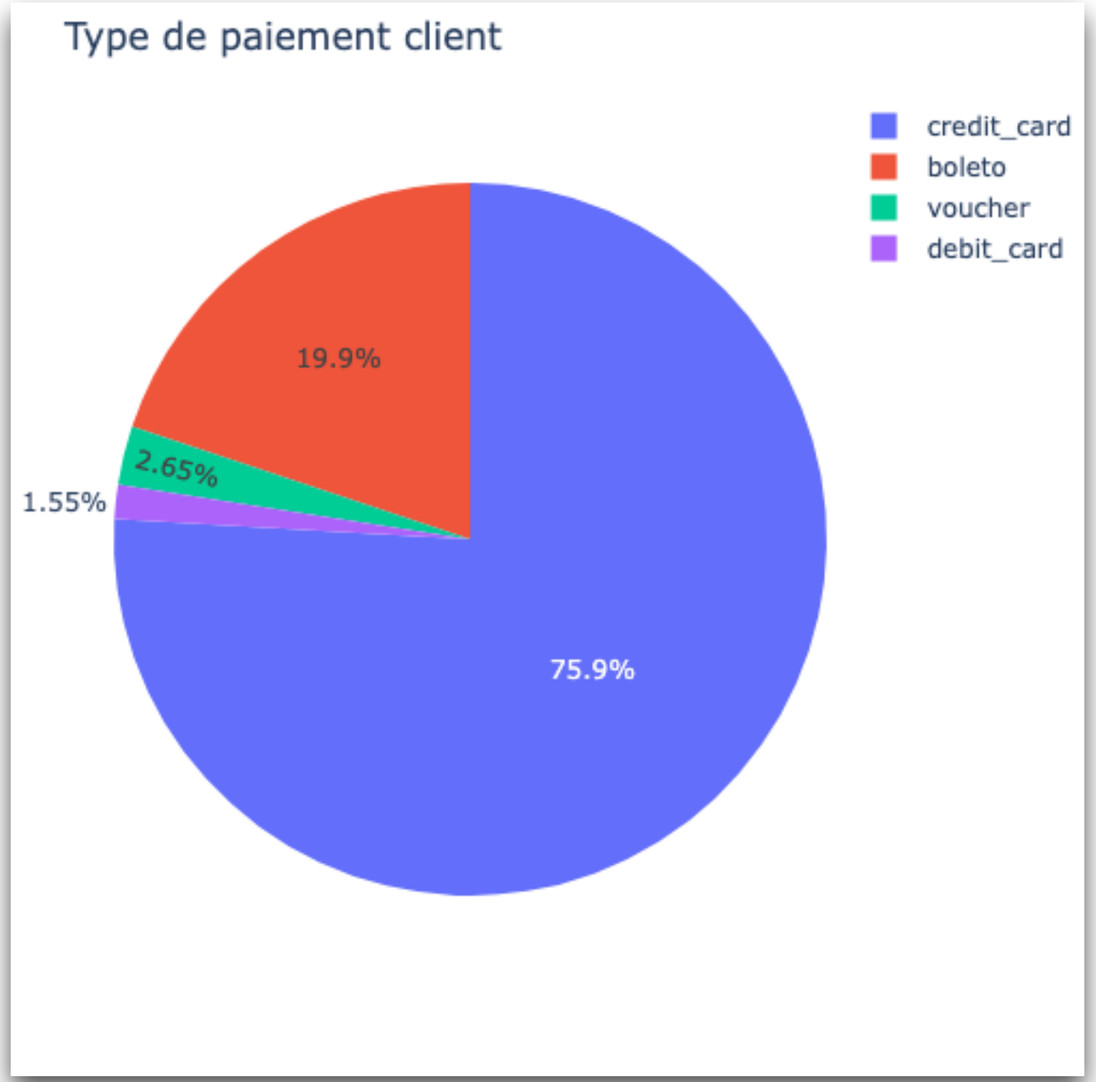
« DES IMAGES PARLENT PLUS QUE DES MOTS »

VISUALISATIONS

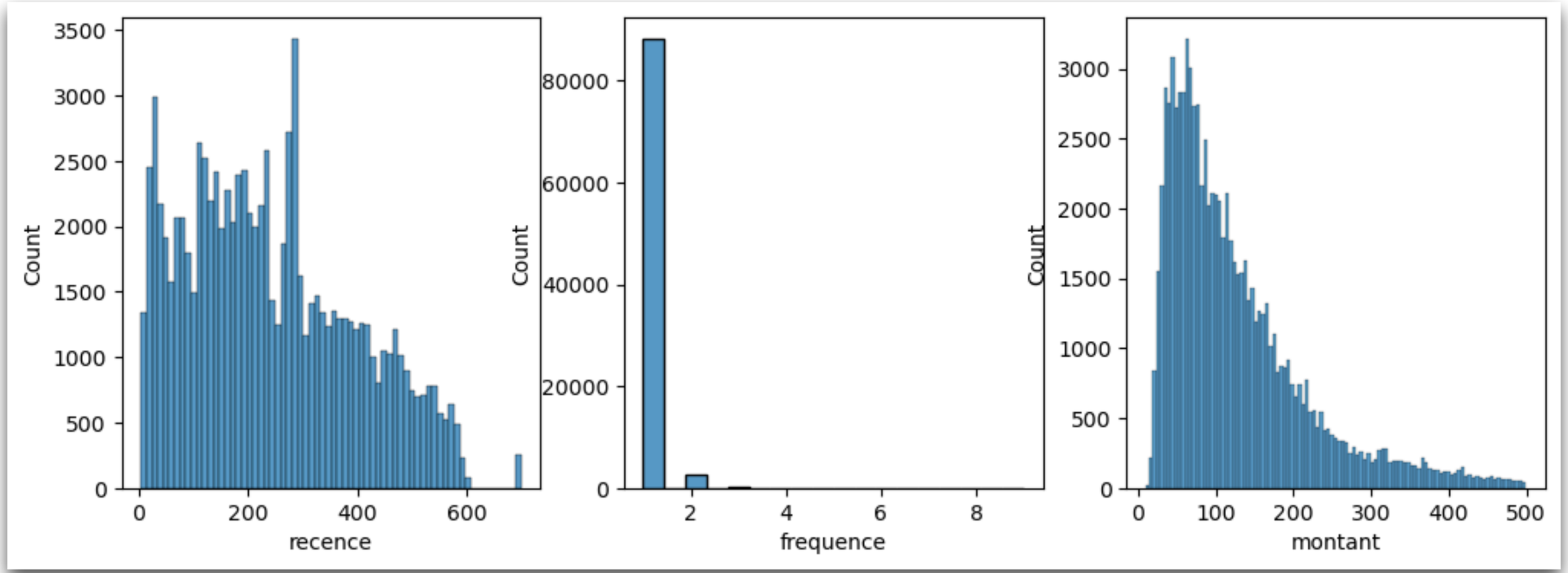
VISUALISATIONS



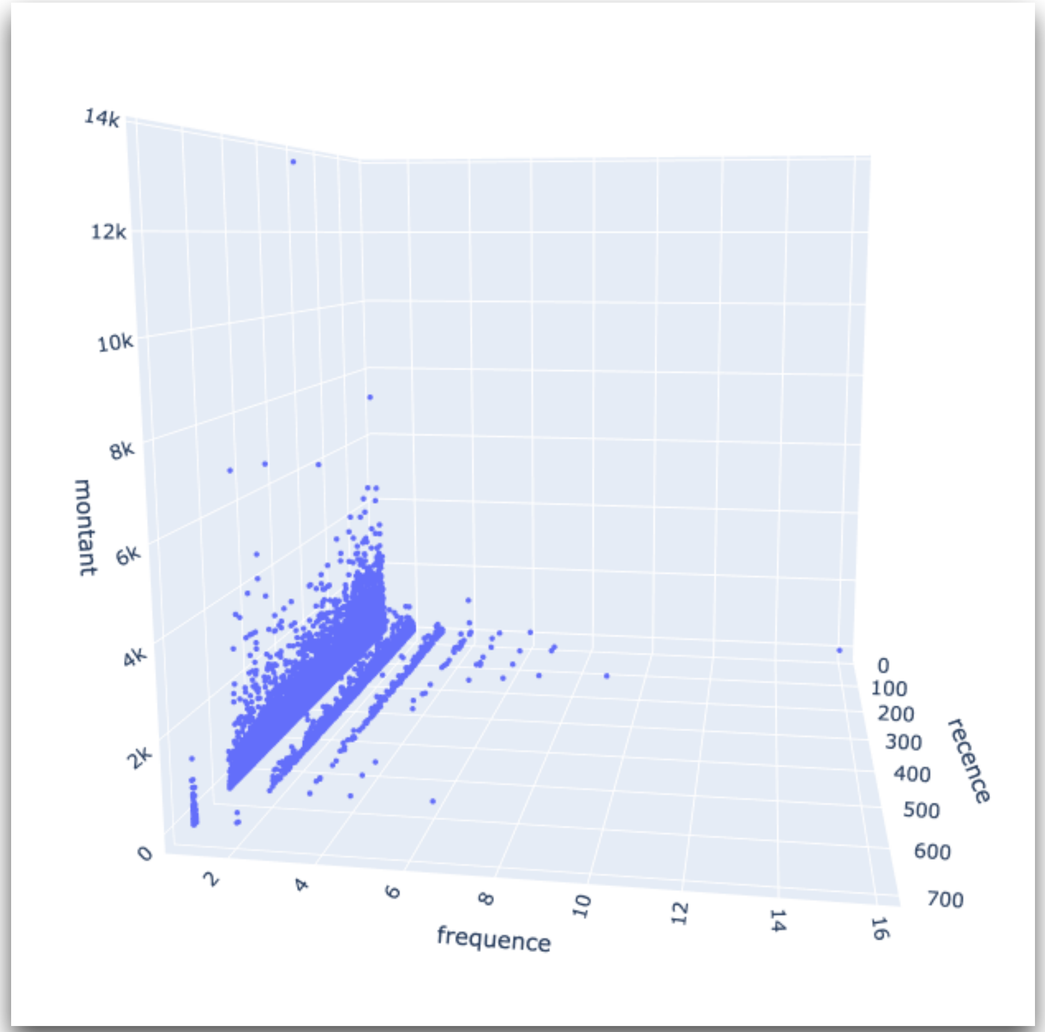
VISUALISATIONS



VISUALISATIONS



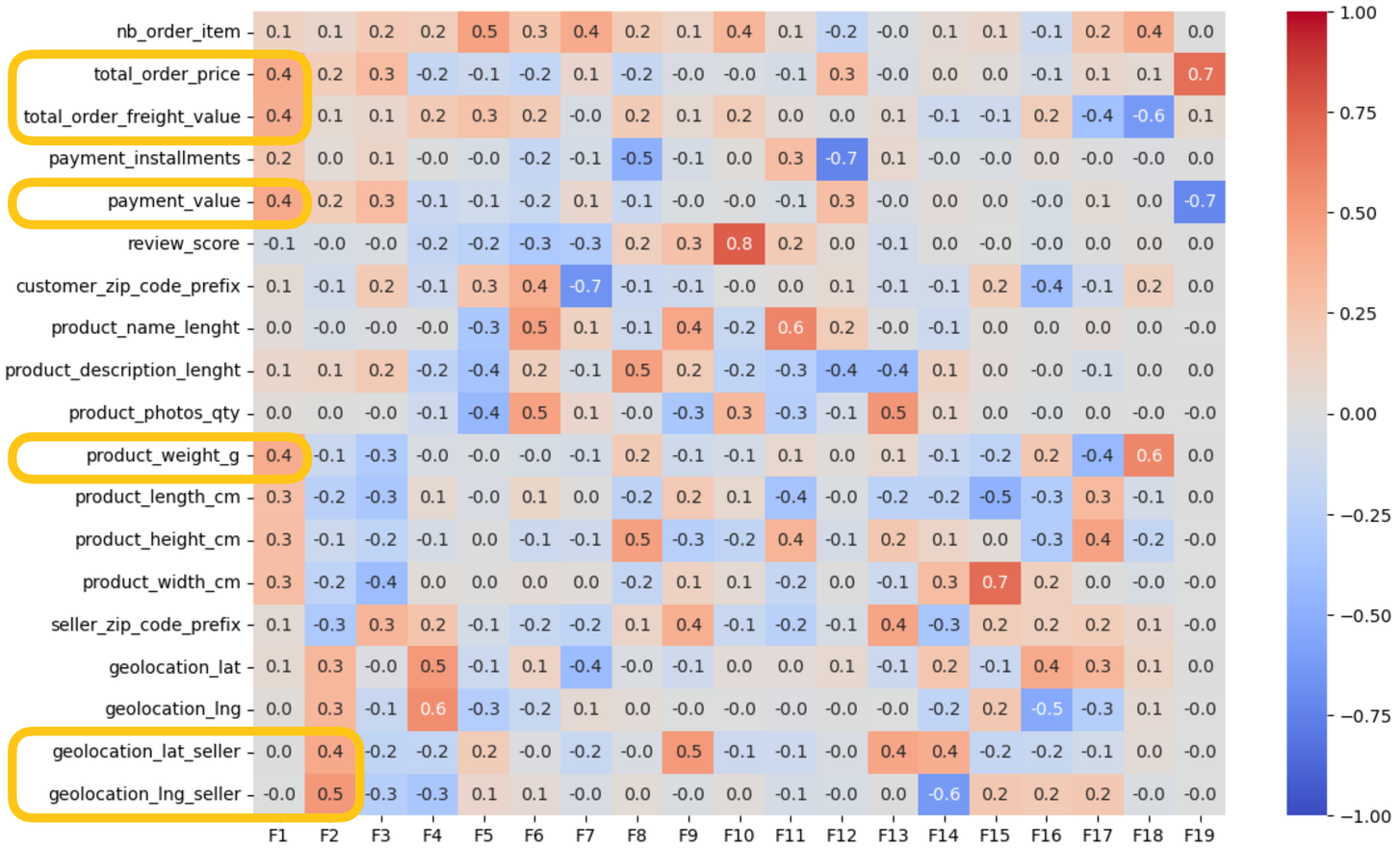
VISUALISATIONS



MODÈLES DE CLUSTERING

MODÈLES DE CLUSTERING

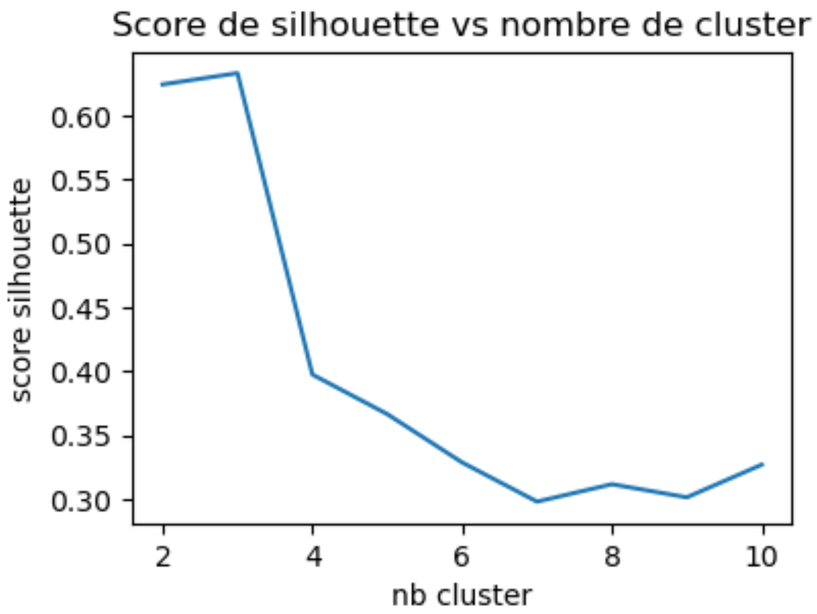
- Sans feature engineering
  - Données de base
- Feature les plus représentatives
  - ACP
    - Plan F1-F2

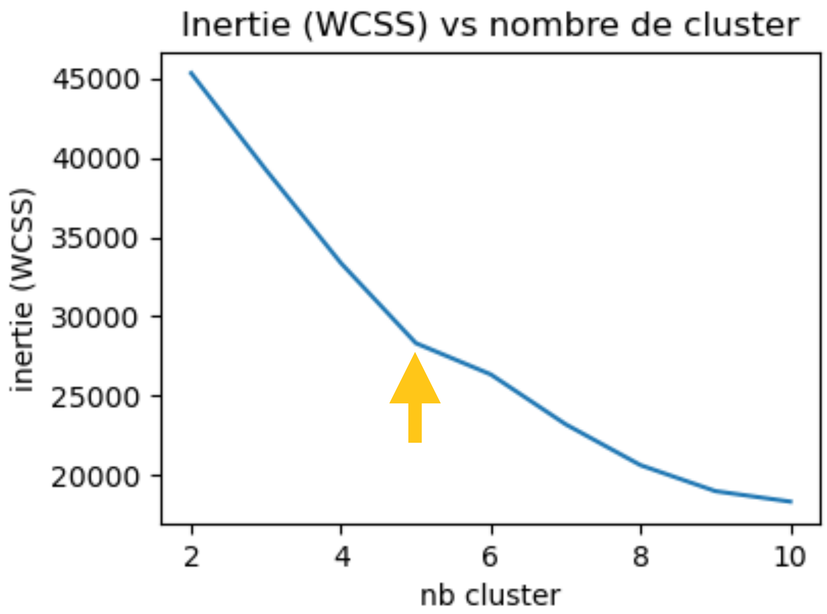


MODÈLES DE CLUSTERING

- K-Means
  - 5 clusters

K-Means

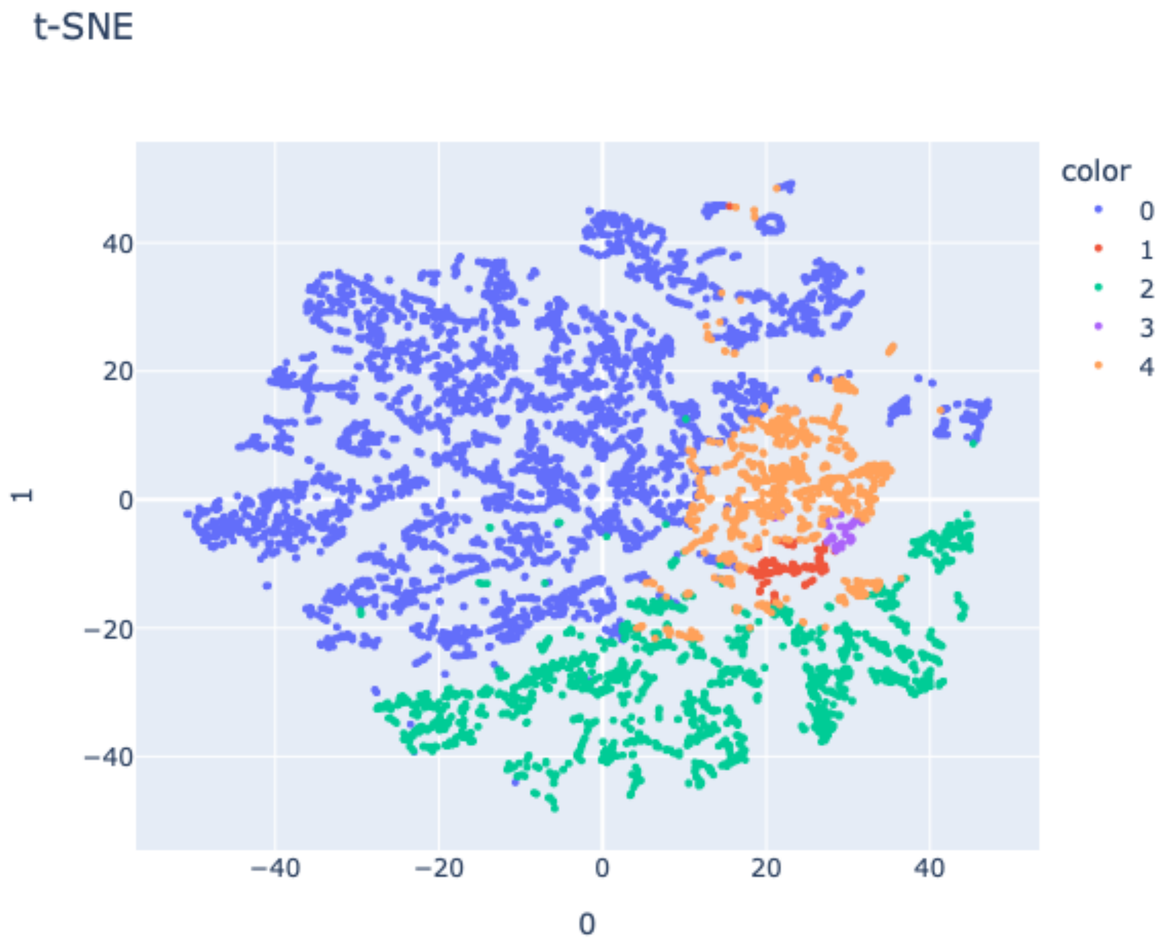




MODÈLES DE CLUSTERING

- T-SNE :
  - Clusters de tailles très faibles (cluster 4 contient 450 individus)

- Score de stabilité (ARI) = 0,60

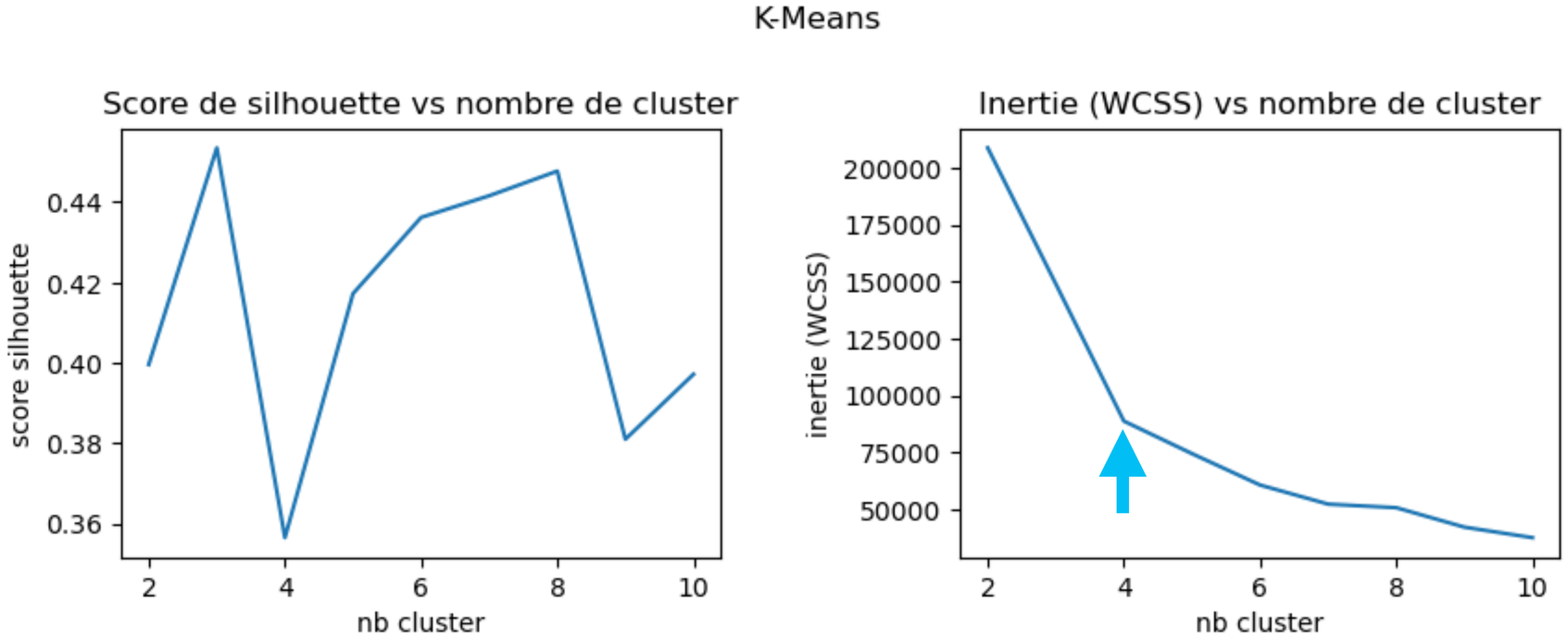


« ON COMPTE SUR VOUS »

RFM

RFM

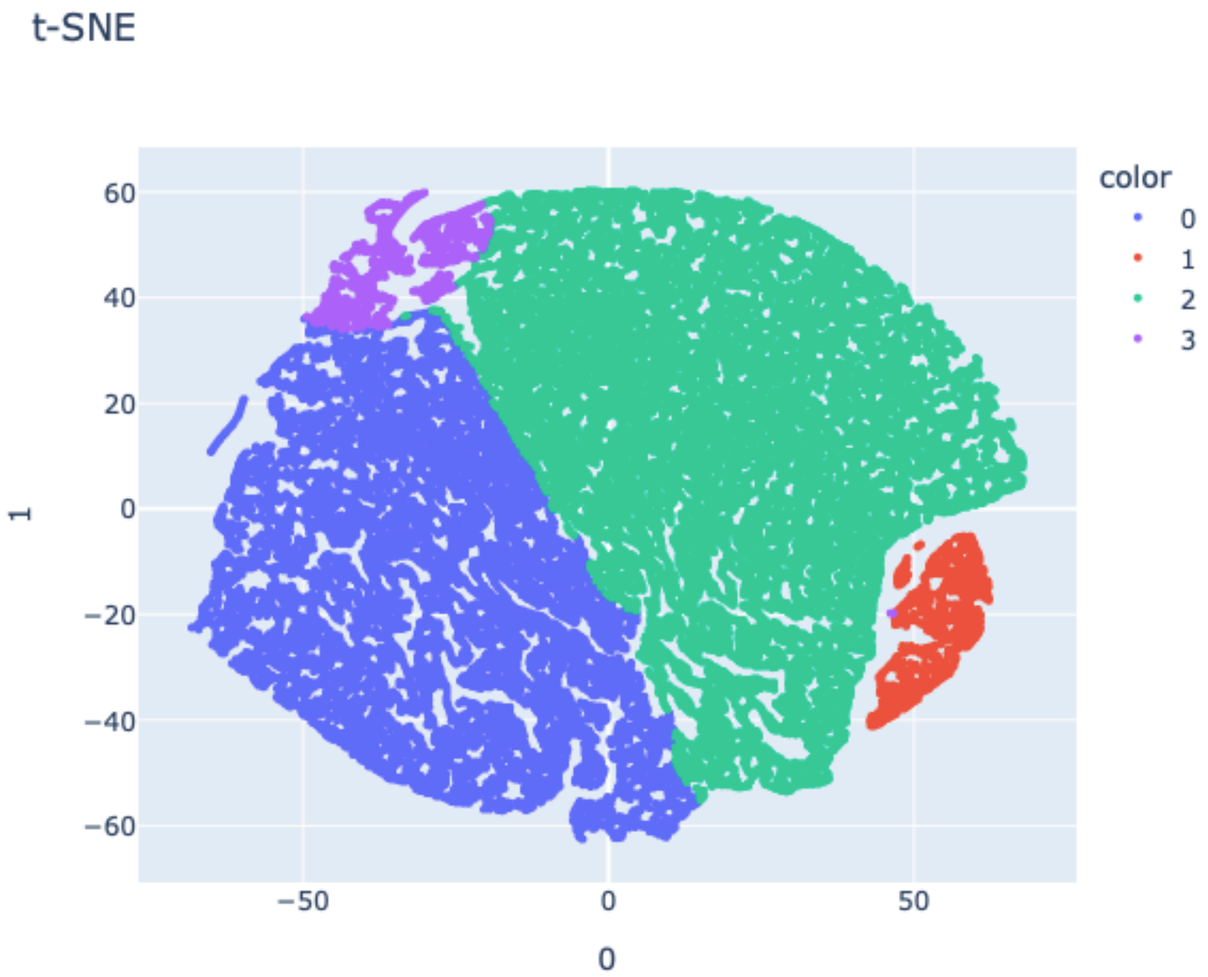
- K-Means
  - 4 clusters



RFM

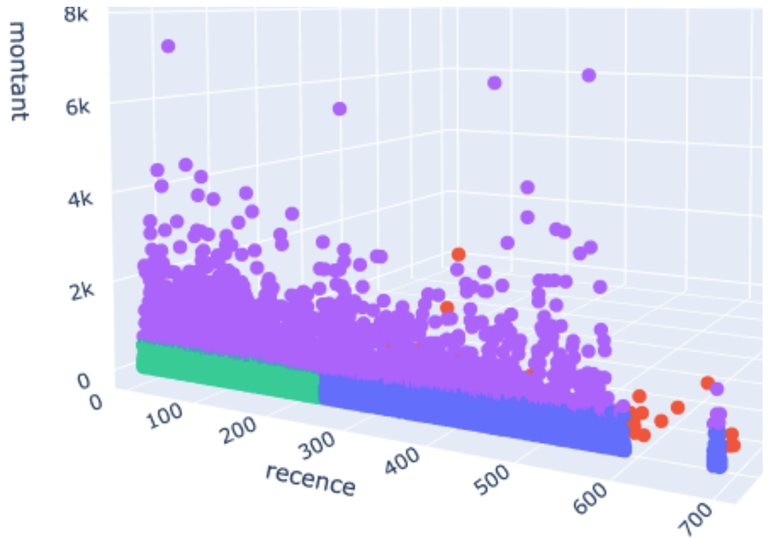
- T-SNE :
  - 2 tailles de clusters
    - Cluster 0 = 36307
    - Cluster 1 = 2871
    - Cluster 2 = 49139
    - Cluster 3 = 2678

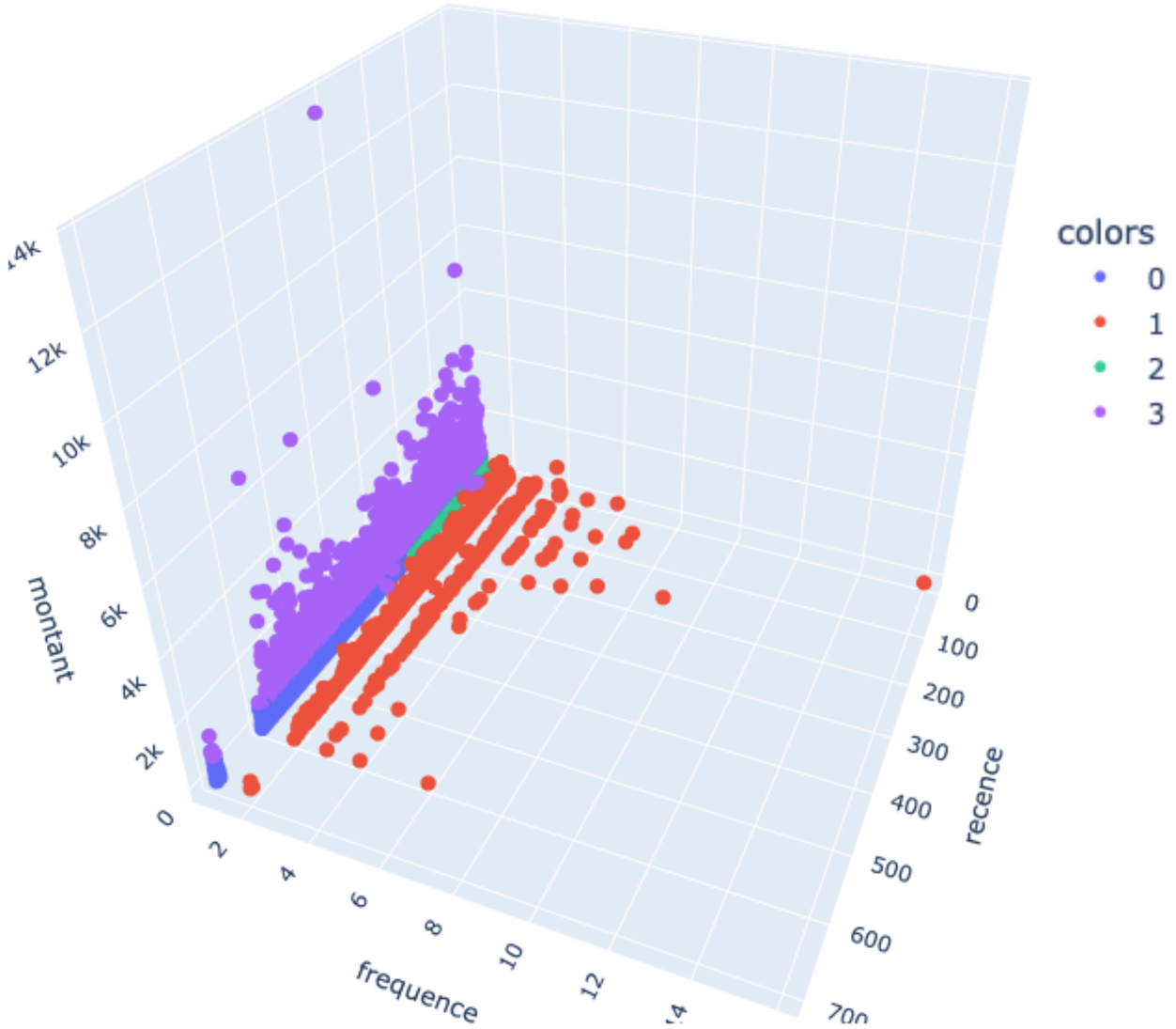
- Score de stabilité (ARI) = 0,93



RFM

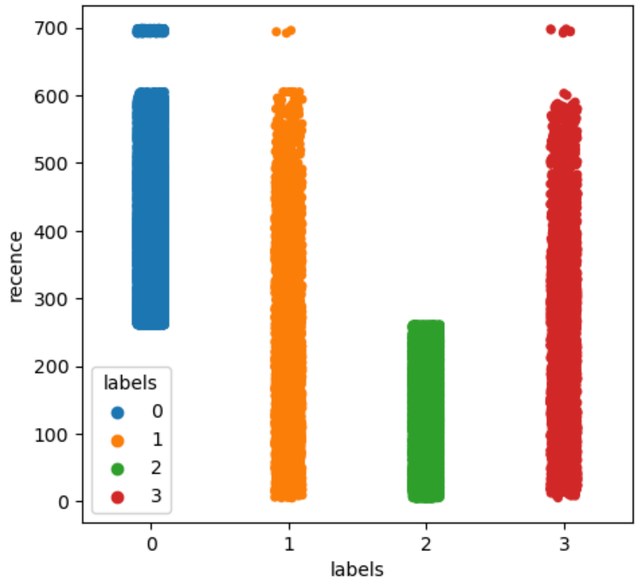
- Vue 3D
  - RFM

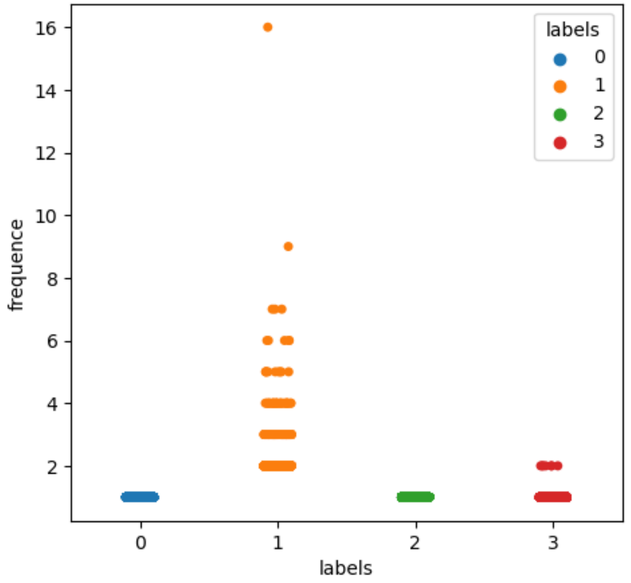


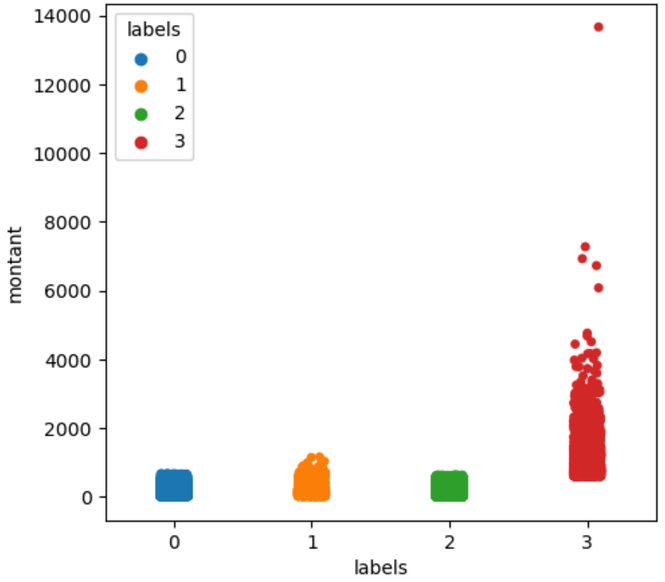


RFM

- Interprétation
  - Cluster 0 :
    - Clients n’ayant pas achetés récemment
  - Cluster 1 :
    - Clients ayant achetés plus qu’une fois
  - Cluster 2 :
    - Clients ayant achetés récemment
  - Cluster 3 :
    - Clients ayant dépensés d’avantage





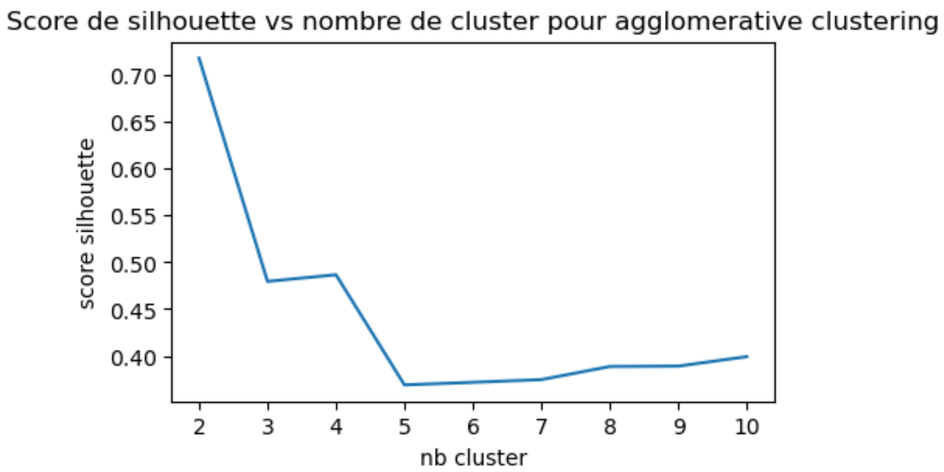


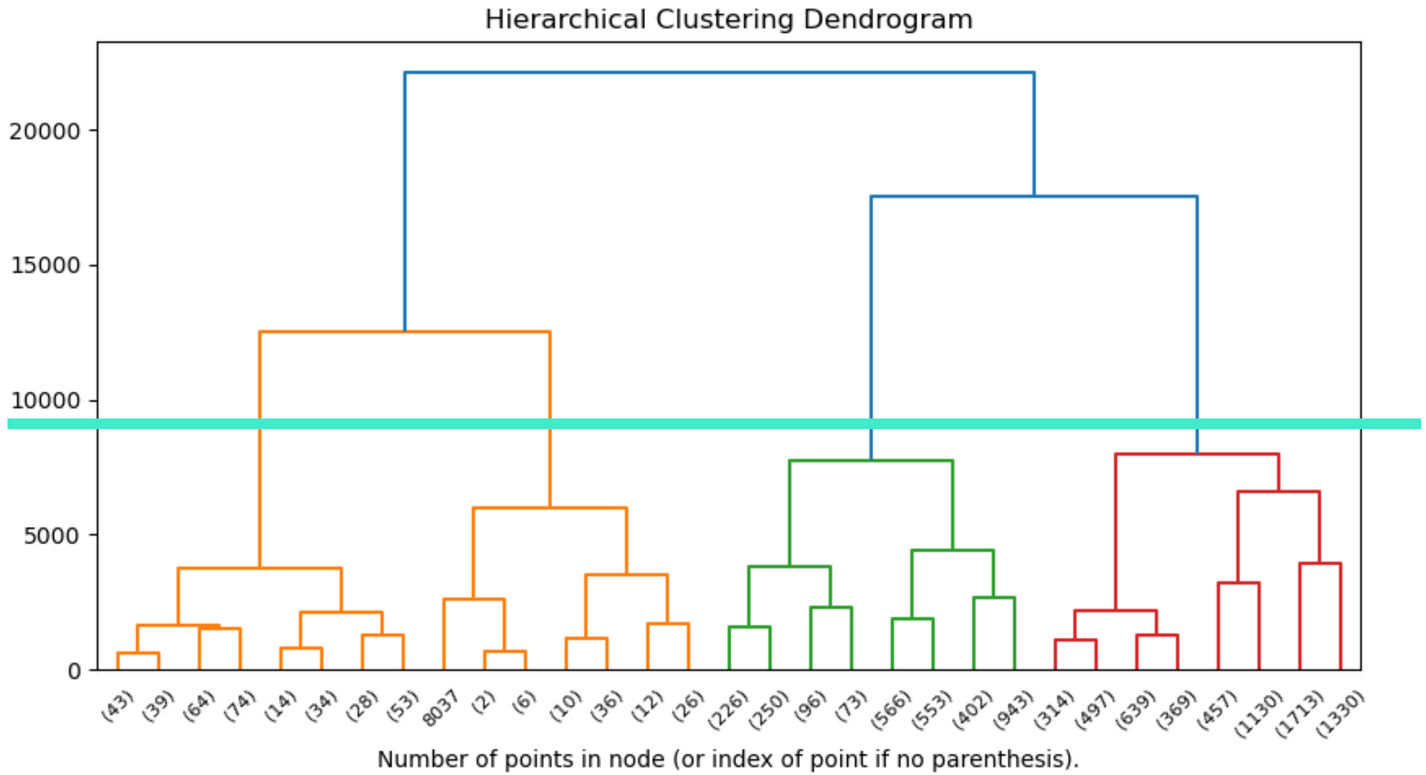
« CHANGEMENT DE JOUEUR »

CLASSIFICATION ASCENDANTE HIÉRARCHIQUE

CLASSIFICATION ASCENDANTE HIÉRARCHIQUE

- CAH
  - 4 clusters

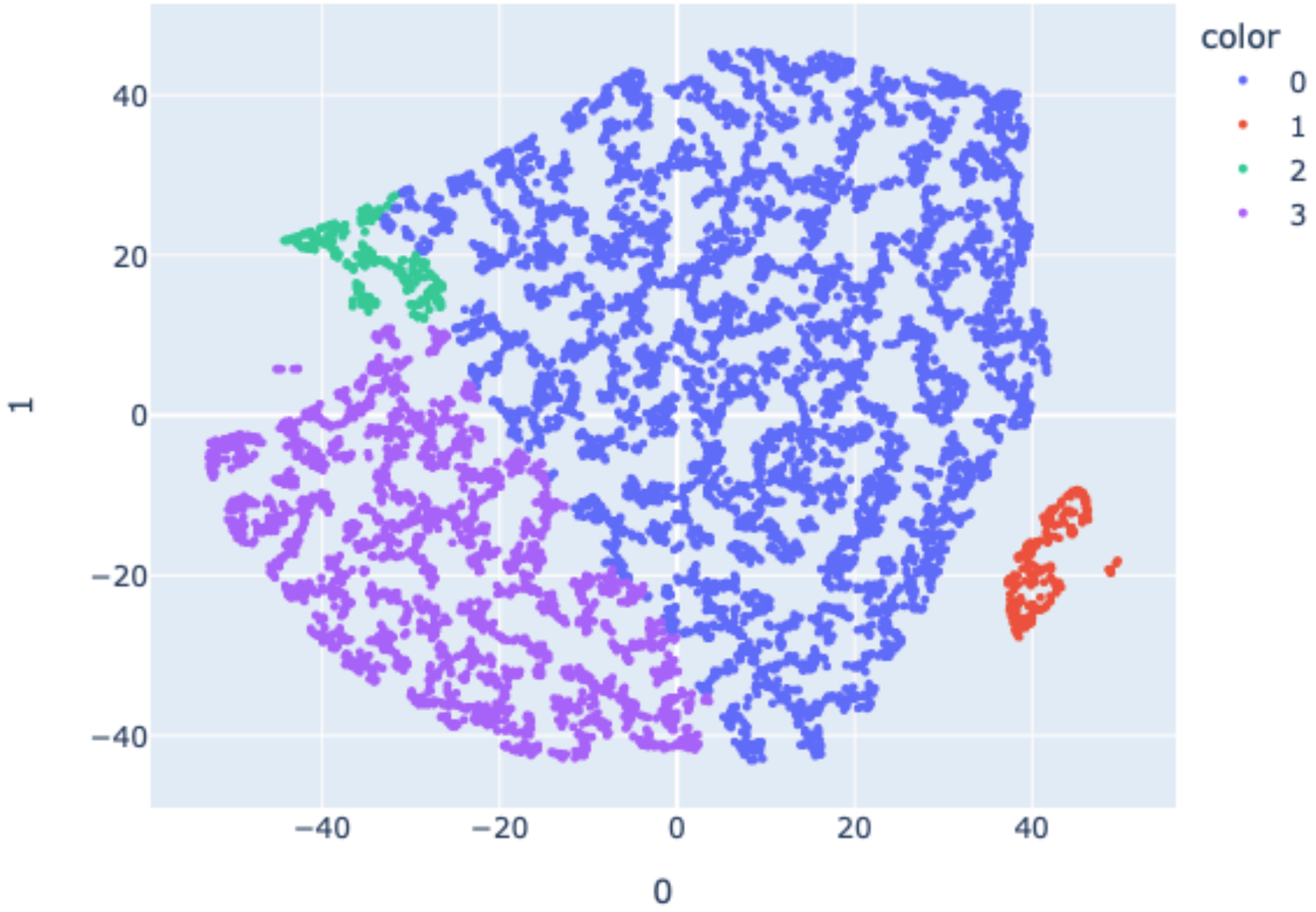




CLASSIFICATION ASCENDANTE HIÉRARCHIQUE

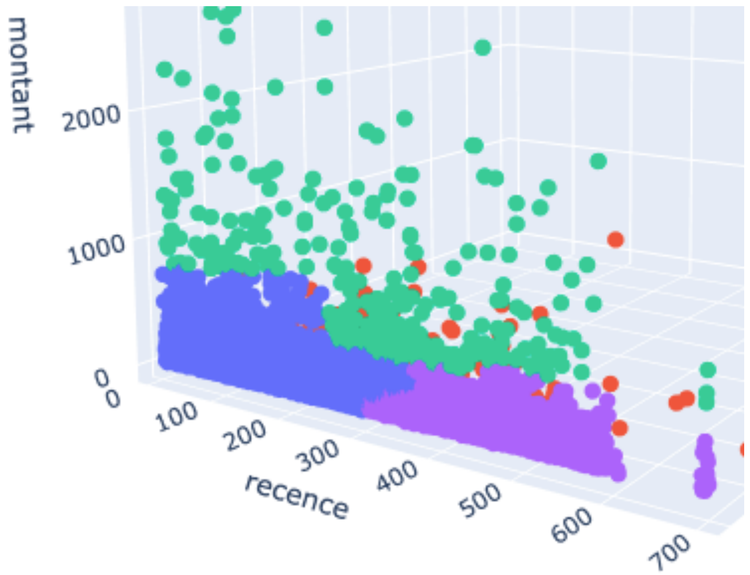
- T-SNE :
  - 3 tailles de clusters
    - Cluster 0 = 66635
    - Cluster 1 = 3133
    - Cluster 2 = 2991
    - Cluster 3 = 27252

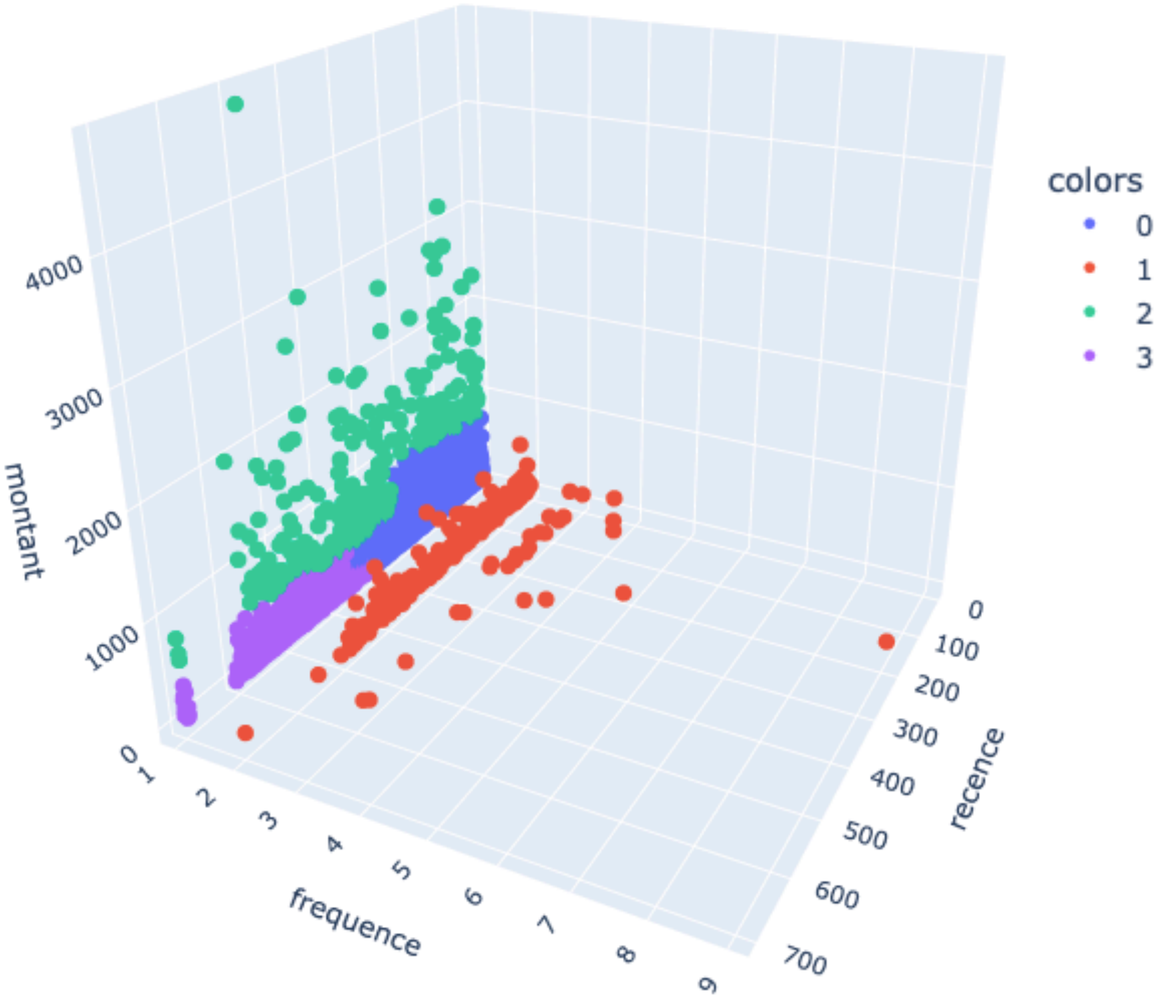
- Score de stabilité (ARI) entre K-Means et CAH = 0,58



CLASSIFICATION ASCENDANTE HIÉRARCHIQUE

- Vue 3D
  - RFM



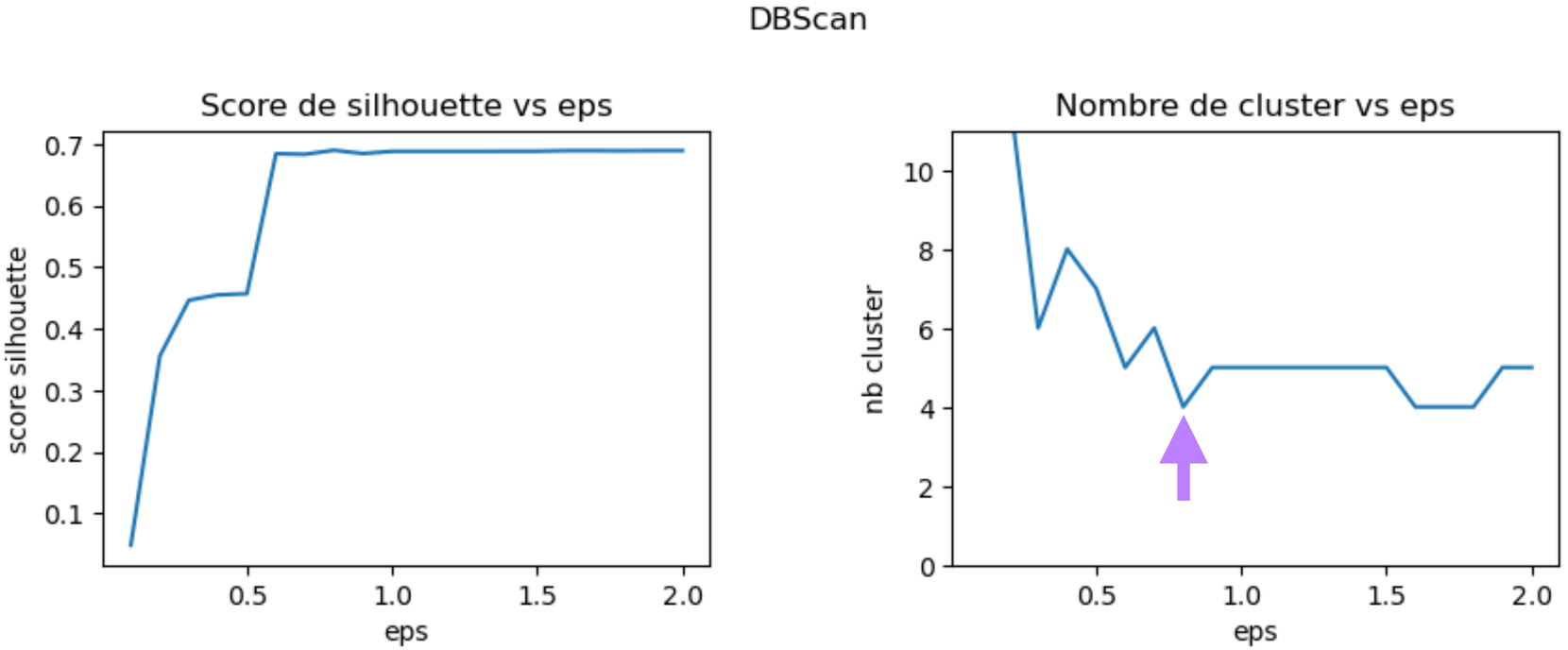


« CHANGEMENT DE JOUEUR… ENCORE »

DBSCAN

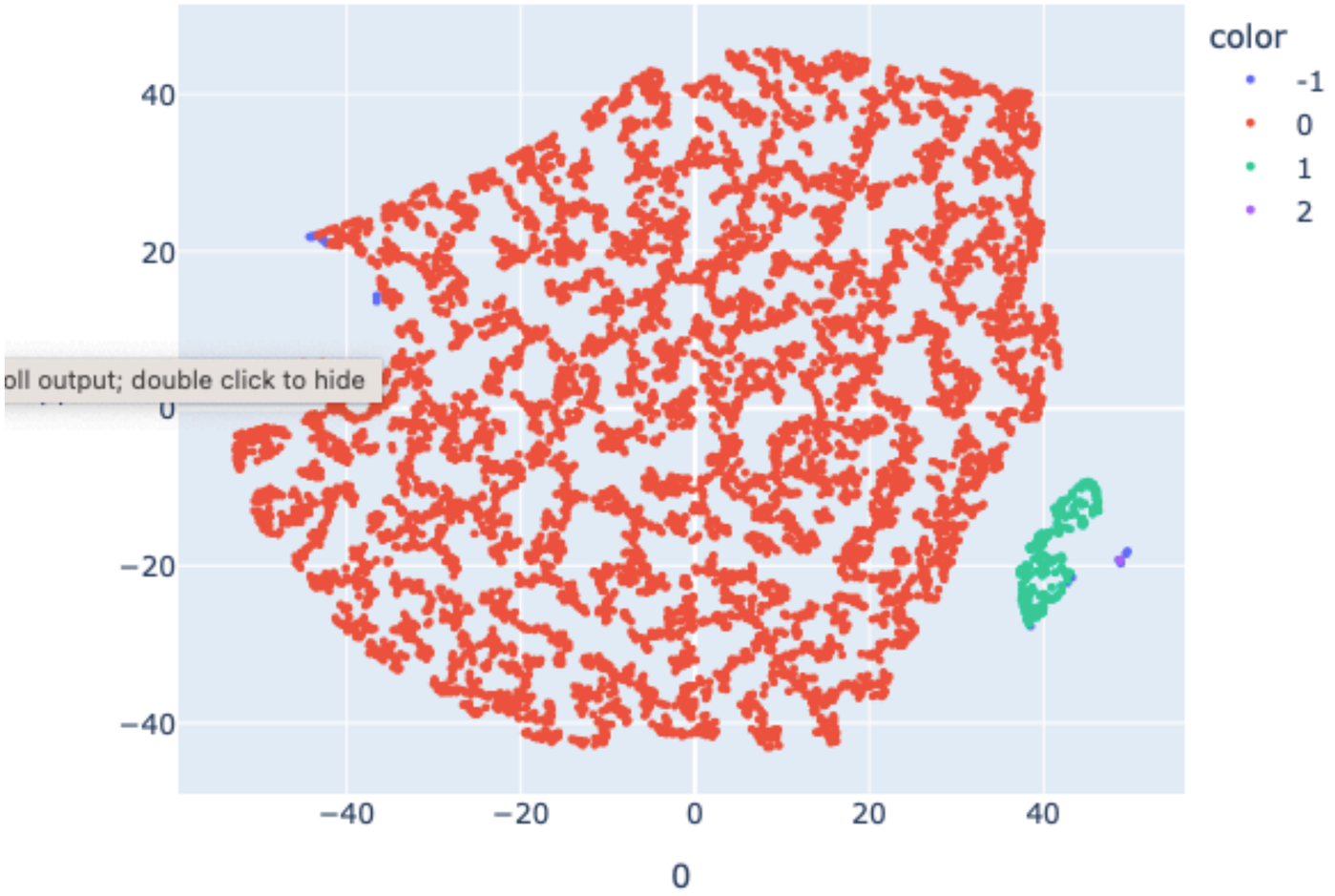
DBSCAN

- DBScan
  - 4 clusters
    - eps = 0,8



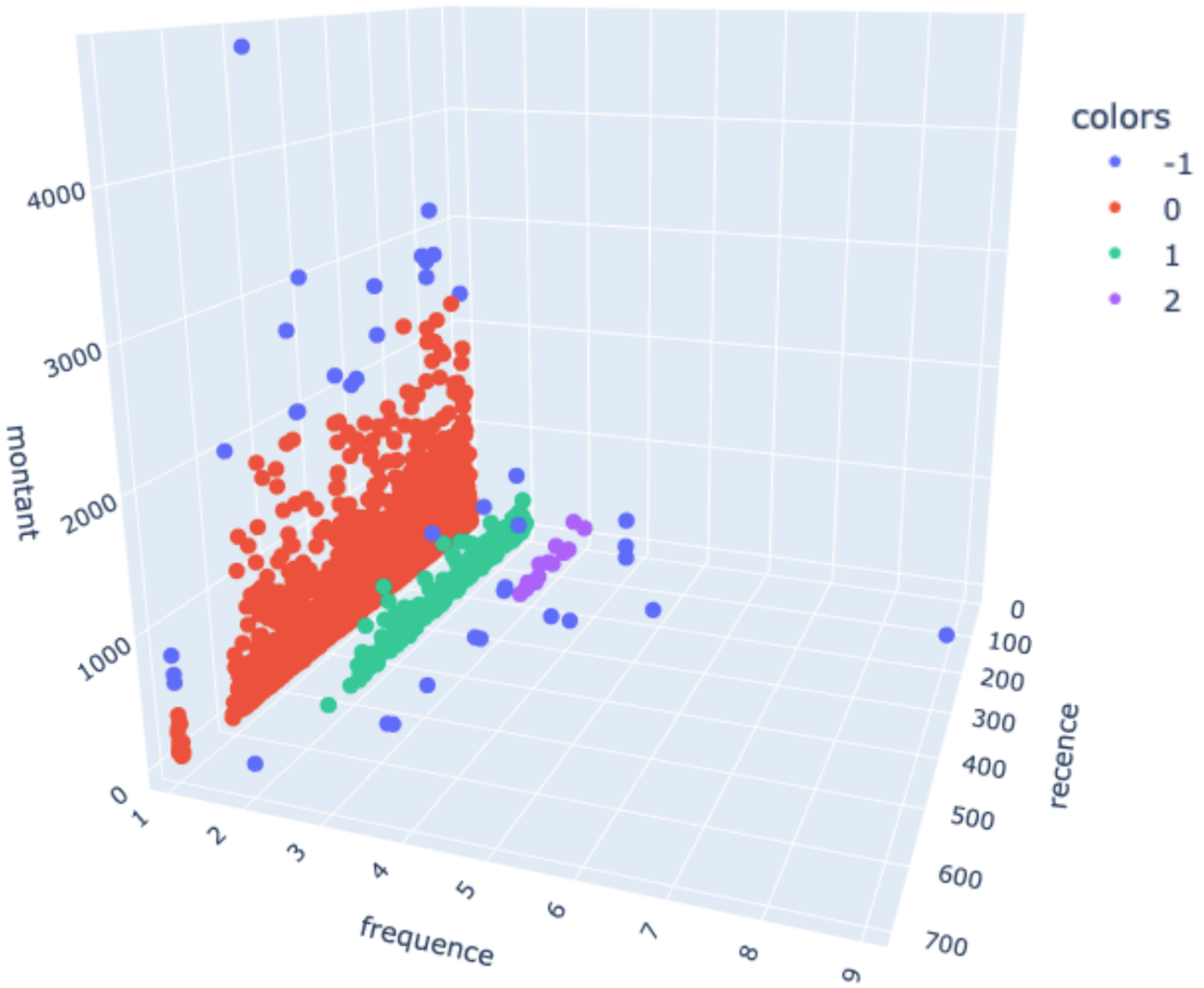
DBSCAN

- T-SNE :


color
-1
0
1
2
40
20
0
−20
−40
−40
−20
0
20
40
0
oll output; double click to hide

DBSCAN

- Vue 3D
  - RFM

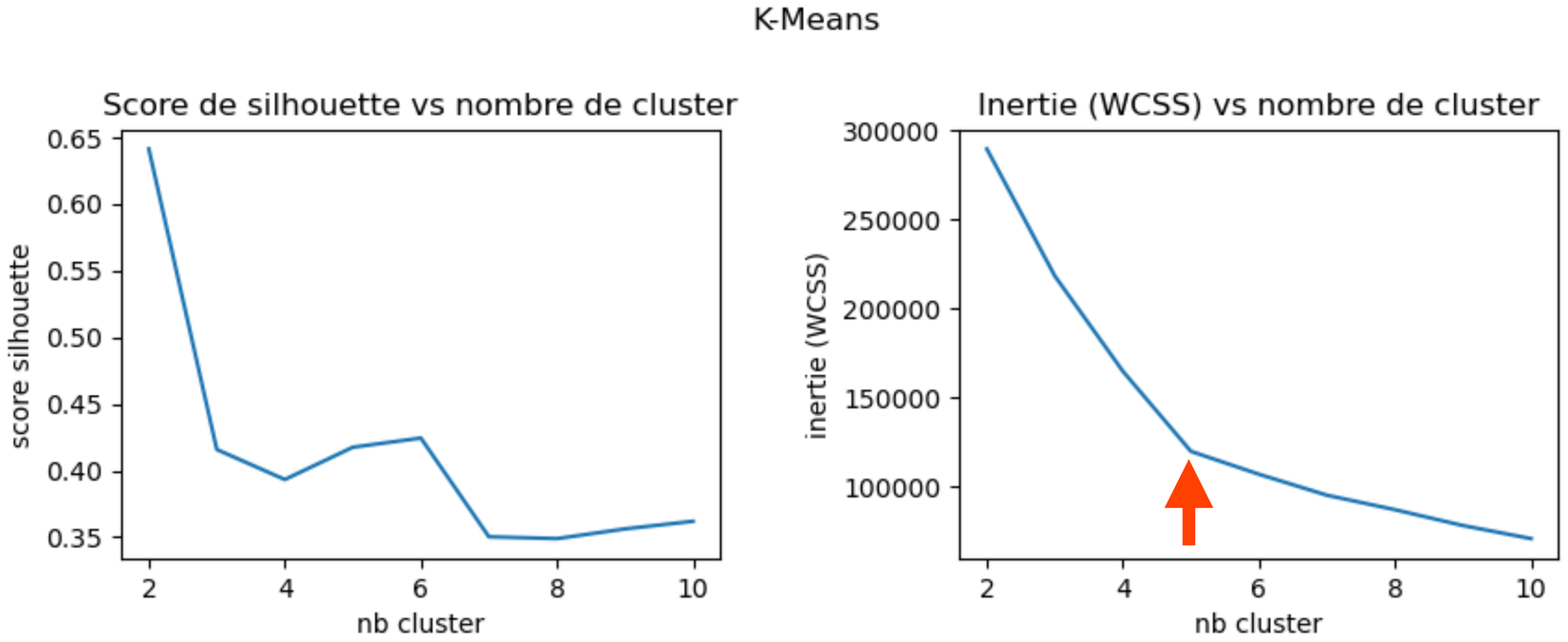


« LE PETIT NOUVEAU »

RFM + REVIEW SCORE

RFM + REVIEW SCORE

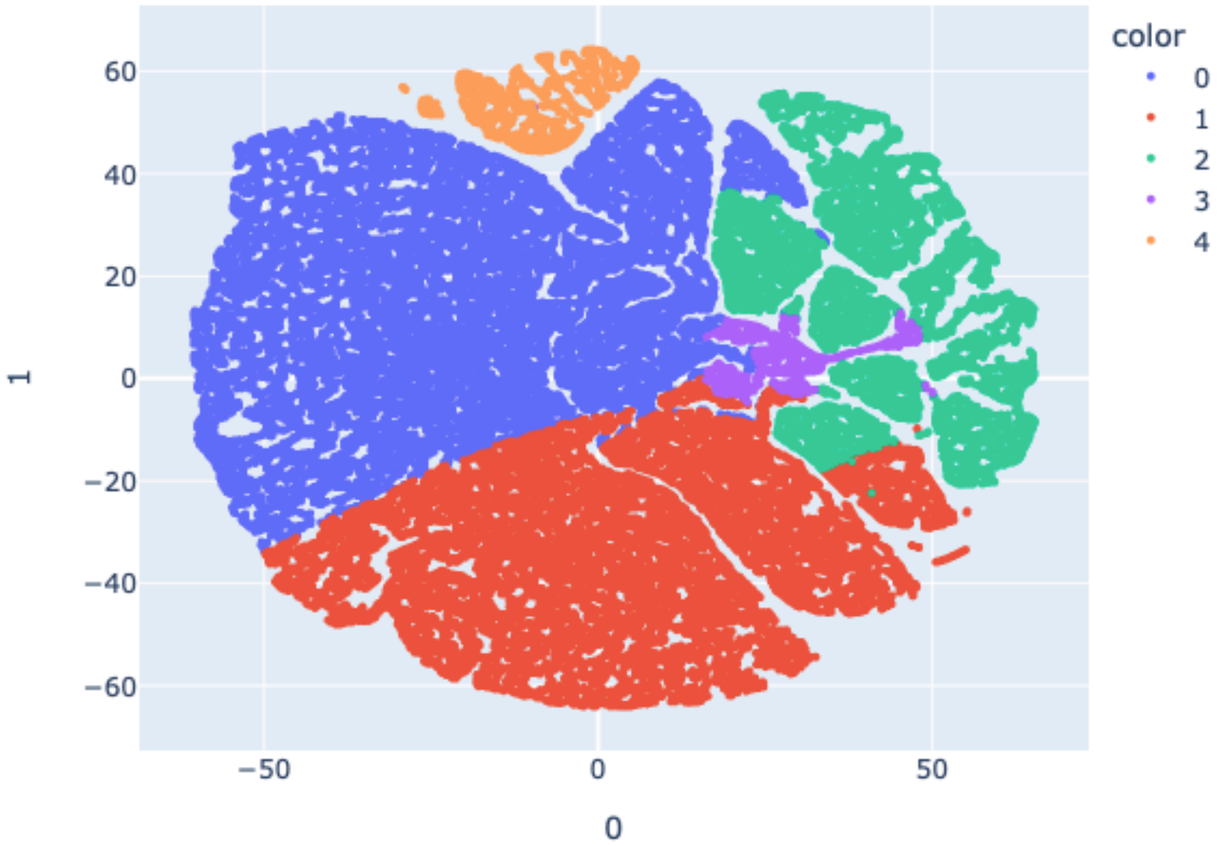
- K-Means
  - 5 clusters



RFM + REVIEW SCORE

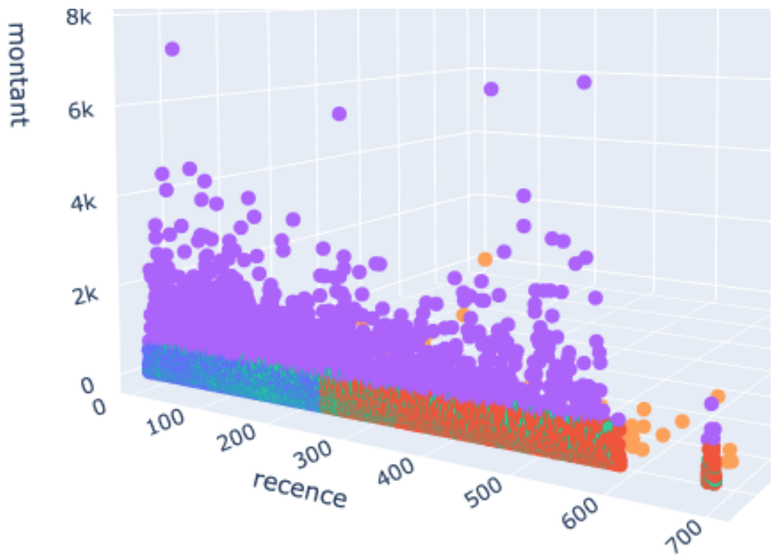
- T-SNE :
  - 2 tailles de clusters
    - Cluster 0 = 40206
    - Cluster 1 = 30524
    - Cluster 2 = 15261
    - Cluster 3 = 2133
    - Cluster 4 = 2871

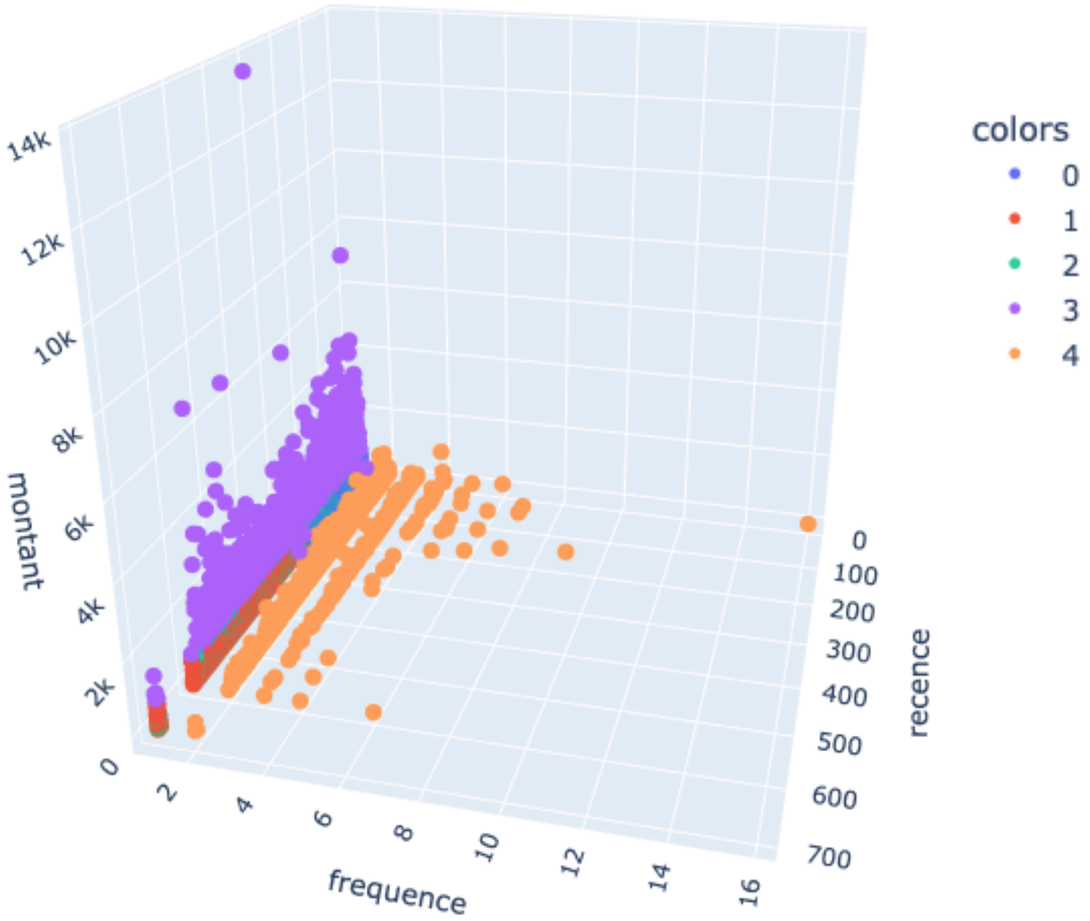
- Score de stabilité (ARI) = 0,97



RFM + REVIEW SCORE

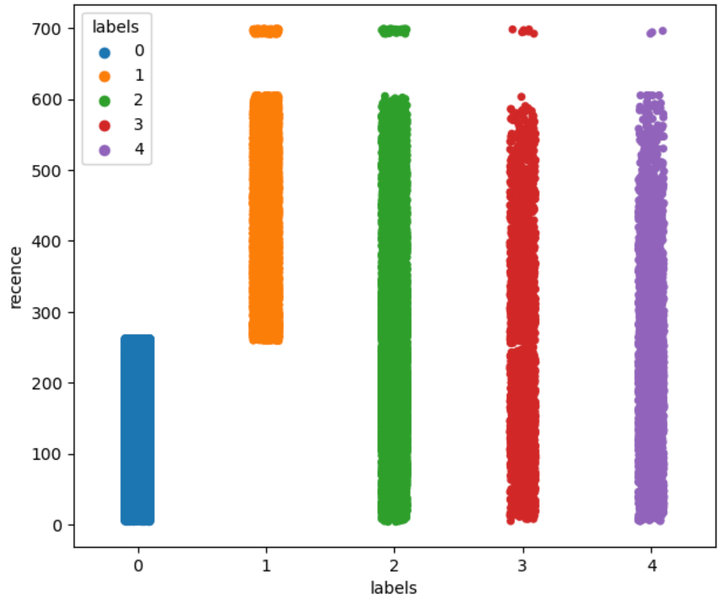
- Vue 3D
  - RFM

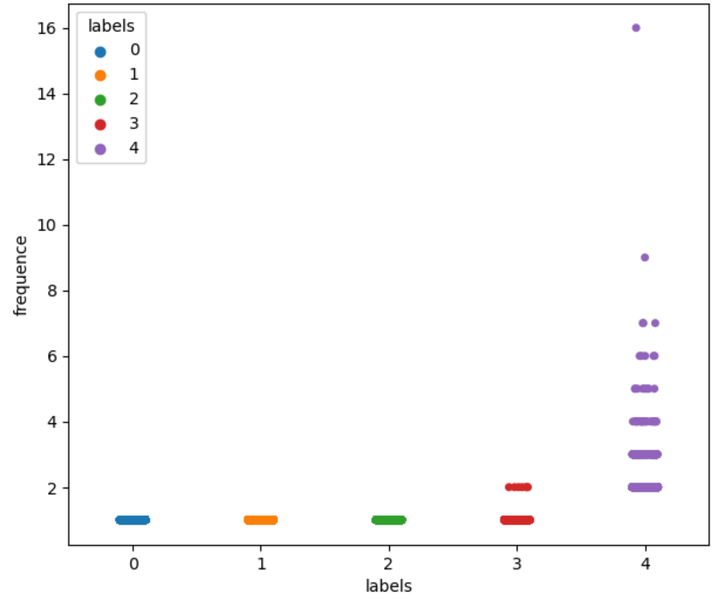


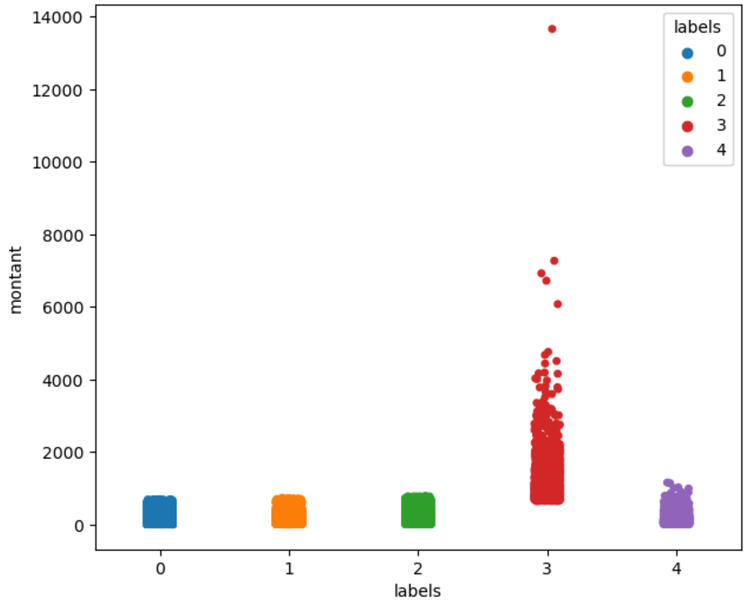


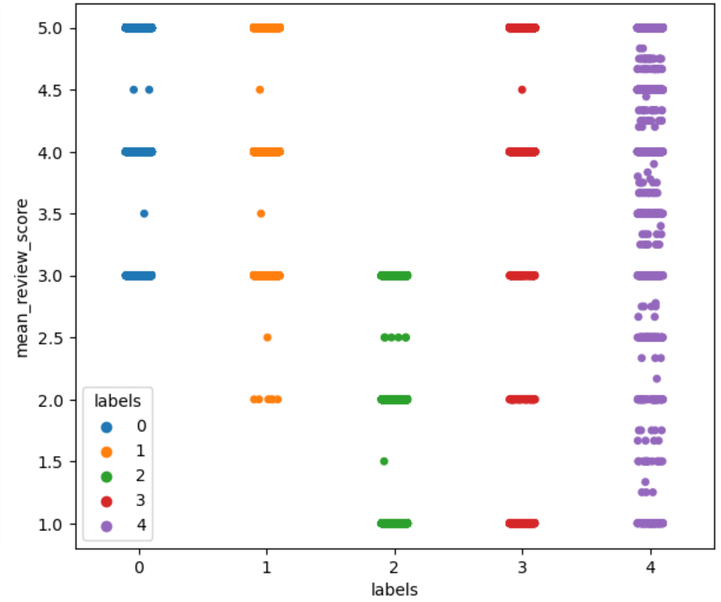
RFM + REVIEW SCORE

- Interprétation
  - Cluster 0 :
    - Clients ayant achetés récemment et ayant mis une bonne note
  - Cluster 1 :
    - Clients n'ayant pas achetés récemment
  - Cluster 2 :
    - Clients ayant mis une mauvaise note
  - Cluster 3 :
    - Clients ayant dépensés d'avantage
  - Cluster 4 :
    - Clients ayant achetés plus qu'une fois



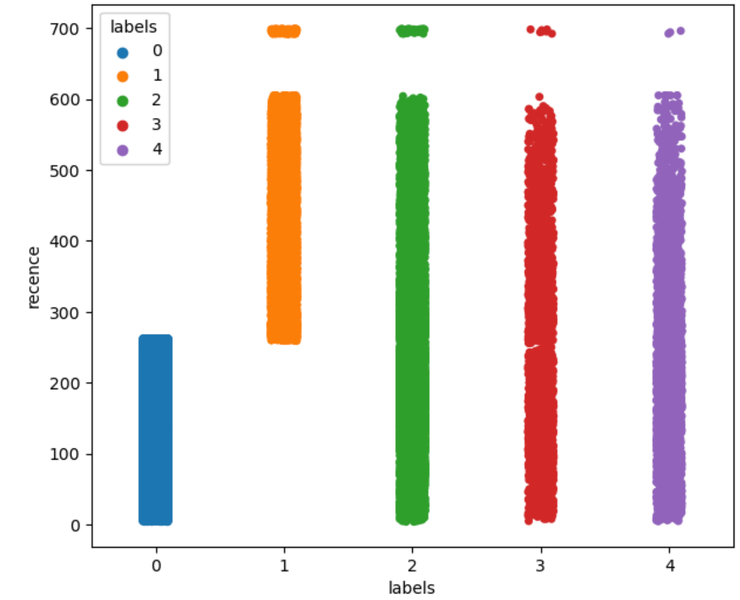


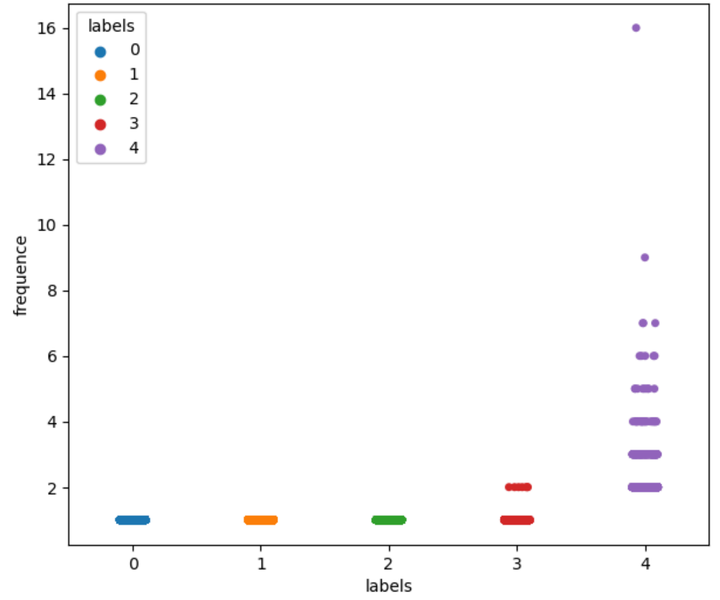


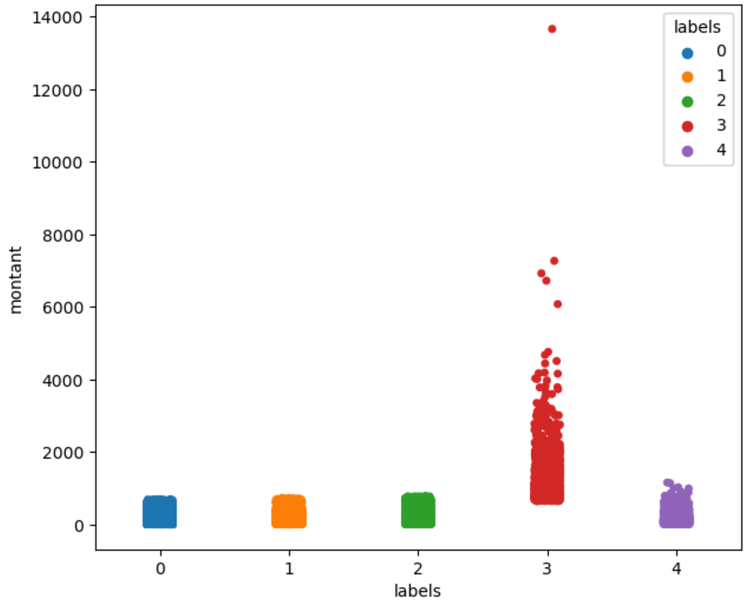


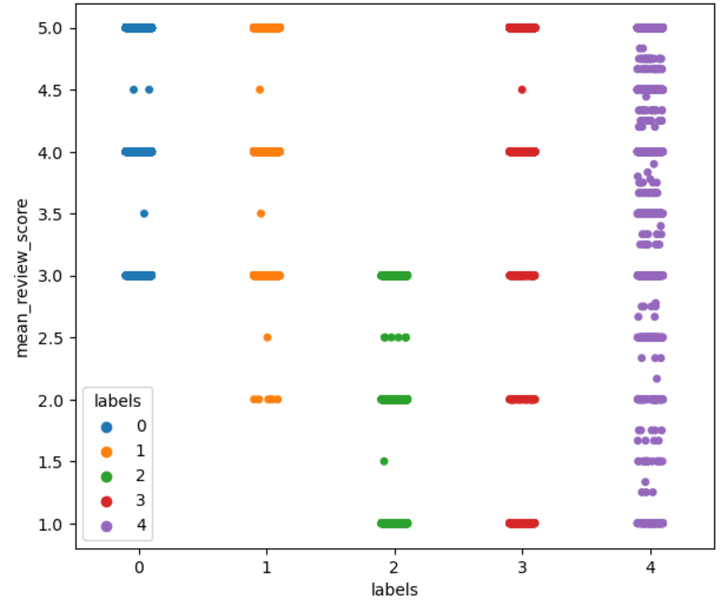
RFM + REVIEW SCORE

- Interprétation
  - Cluster 0 :
    - Nouveaux clients contents
  - Cluster 1 :
    - Anciens clients
  - Cluster 2 :
    - Clients mécontents
  - Cluster 3 :
    - Clients dépensiers
  - Cluster 4 :
    - Clients fréquents









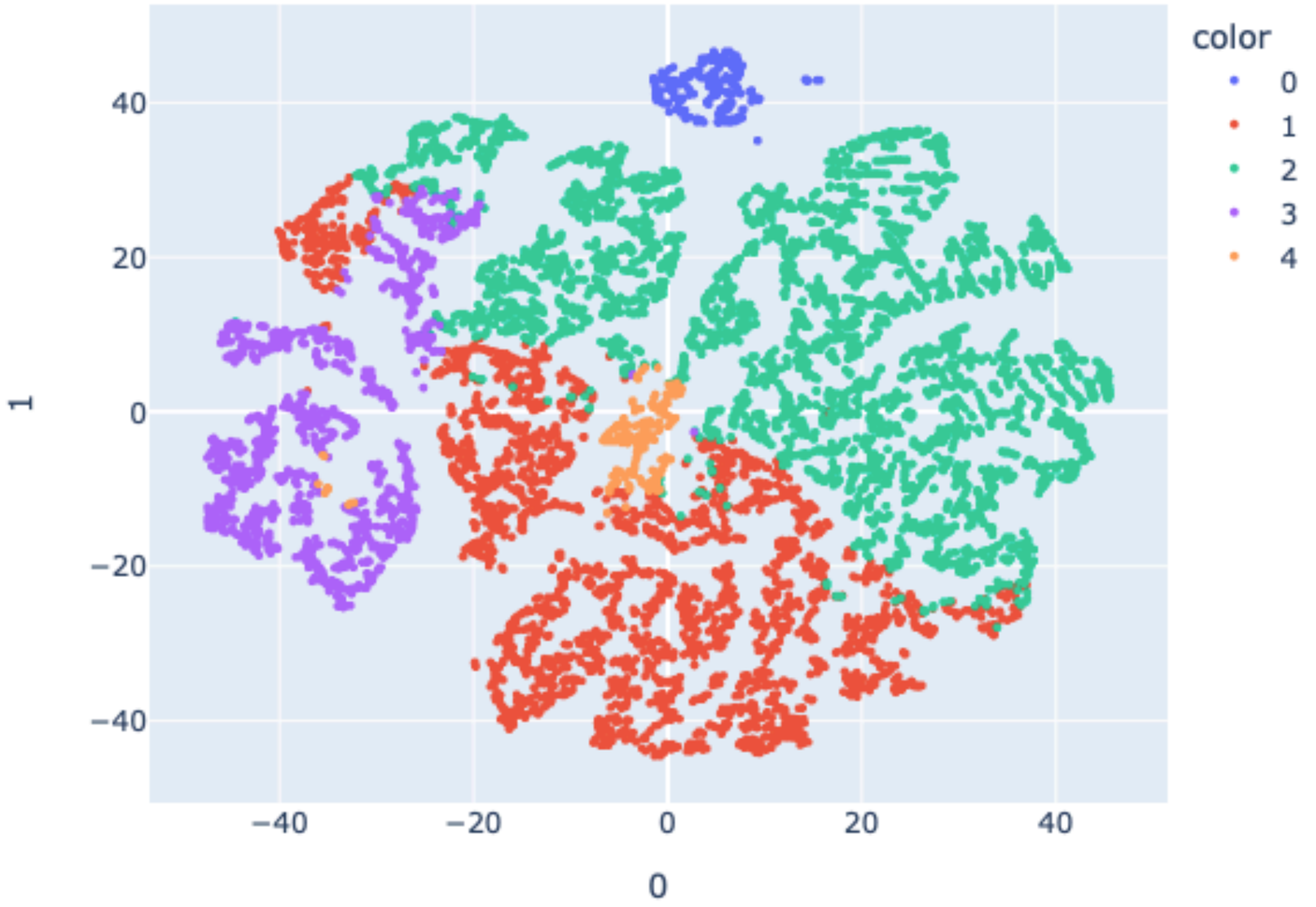
« QUELQU'UN D'AUTRE ? »

ESSAIS AVEC D'AUTRES FEATURES

ESSAIS AVEC D'AUTRES FEATURES

- K-Means
- T-SNE

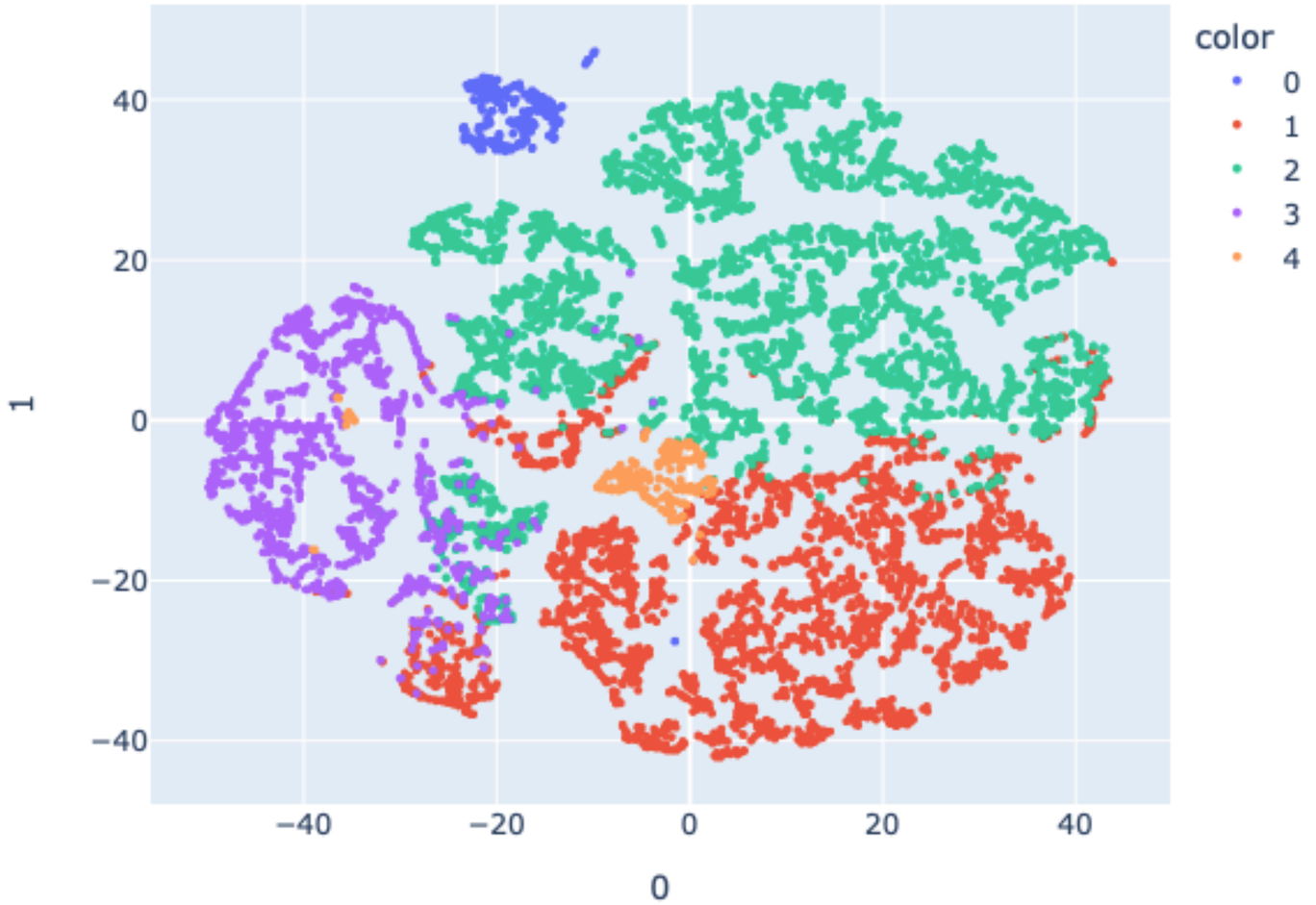
- Délai de livraison
  - 5 clusters



ESSAIS AVEC D'AUTRES FEATURES

- K-Means
- T-SNE

- Différence entre livraisons estimée et réelle
  - 5 clusters

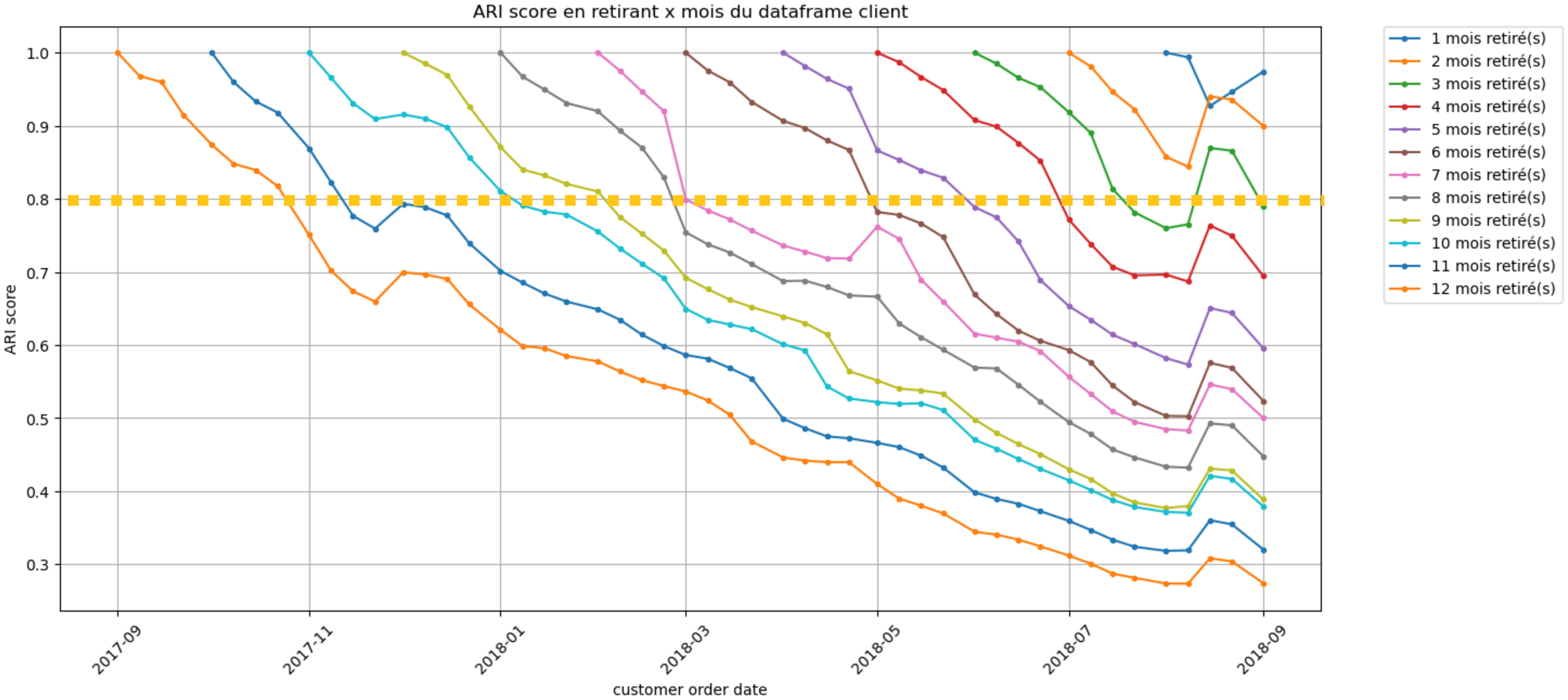


SIMULATION DE MAINTENANCE

SIMULATION DE MAINTENANCE

- Prérequis
  - Utiliser au maximum la base de données fournie
  - Affiner avec précision le besoin de mise à jour (unité en semaine)
  - Utiliser le meilleur modèle
    - K-Means, features « RFM et review score »
    - 5 clusters

SIMULATION DE MAINTENANCE



SIMULATION DE MAINTENANCE

- Limite à 80%
  - Entre 4 et 8 semaines
  - Moyenne à 6,7

CONCLUSION

CONCLUSION

- *Réaliser une segmentation clients*
  - *Exploitable et facile d'utilisation par l'équipe Marketing*
  - *En terme de commandes et de satisfaction*
  - *Sur l'ensemble des clients*

- **Segmentation K-Means**
  - **5 clusters simples**
  - **Utilisation des features « RFM + Review Score »**
  - **Mapping de tous les clients**

CONCLUSION

- *Fournir une proposition de contrat de maintenance*

- **Maintenance toutes les 6 semaines**
  - **Avec reprise de la nouvelle base de données**
  - **Tests de dérive du clustering (nb de clusters, ARI)**


MERCI