


Segmenter les clients d'un site e-commerce

...

14 avril 2023
Yoann Poupart

Problématique: comprendre et décrire les différents types d'utilisateurs et faire une proposition de contrat de maintenance.

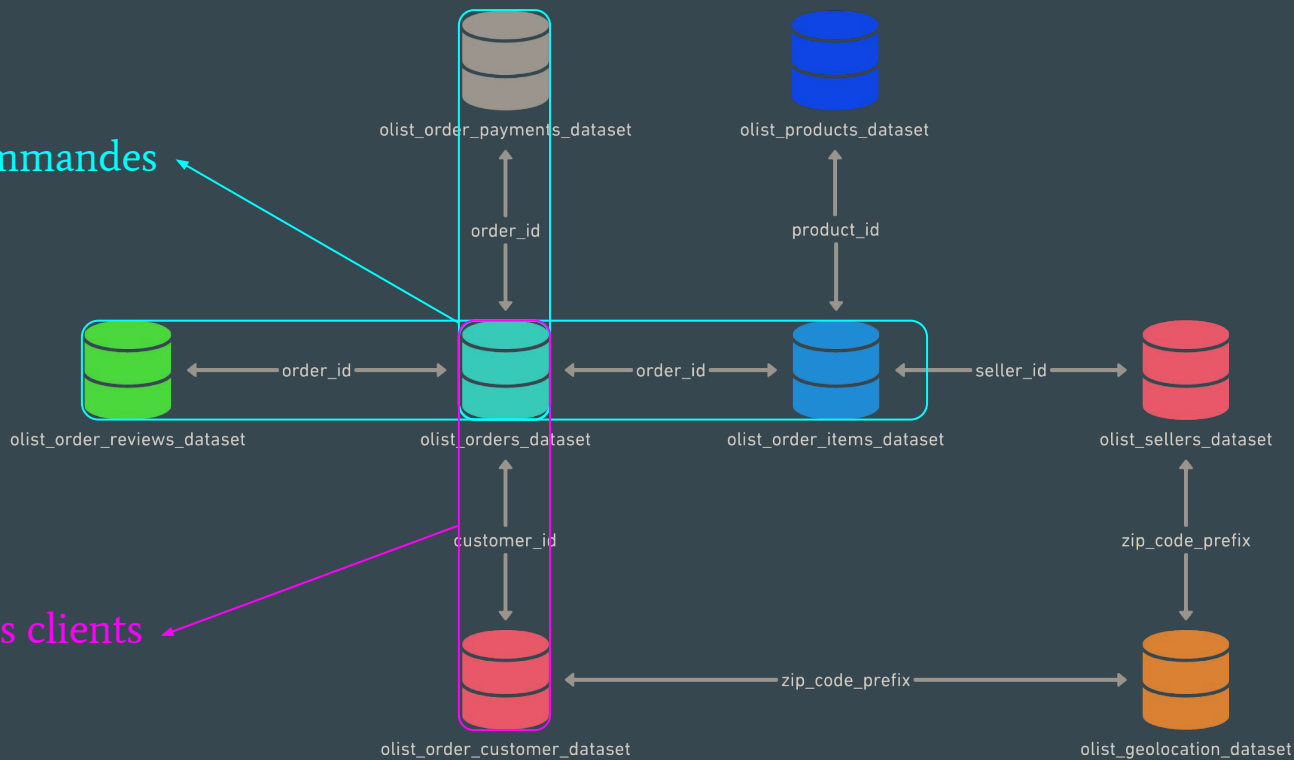
- Feature engineering
- Exploration des données
- Modélisation
- Interprétation
- Maintenance
- Conclusion



Feature engineering

Présentation des données

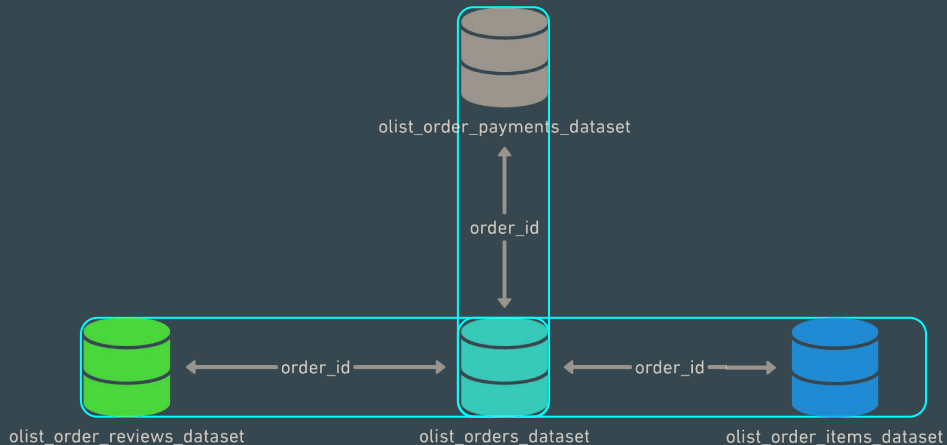
Données commandes



Données commandes

Agrégation des données

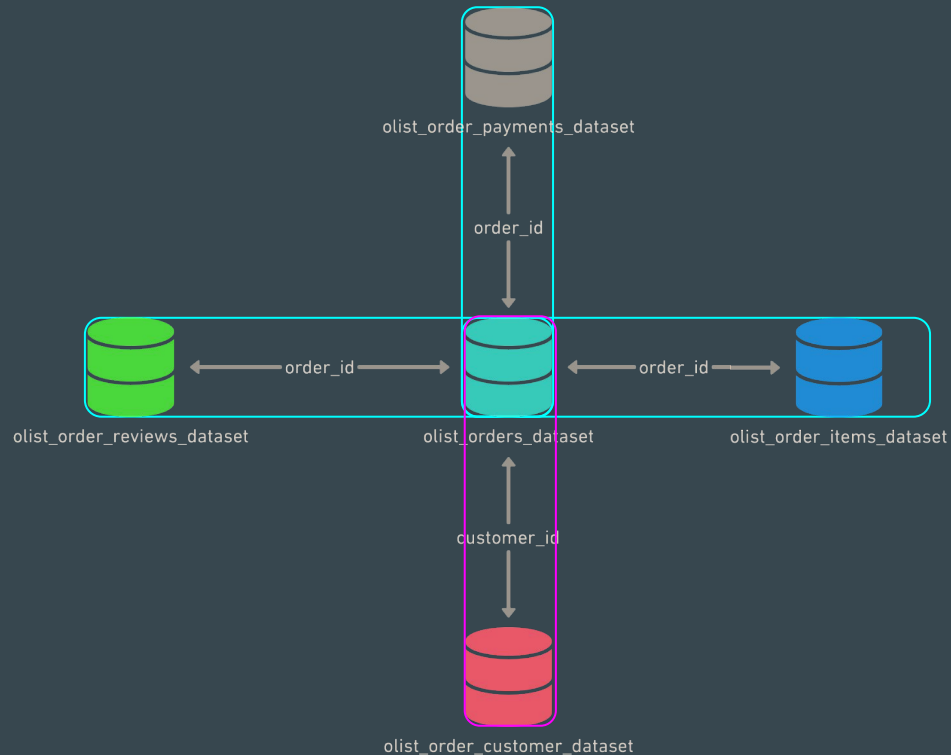
- Articles
 - ➡ Nombre d'articles
- Payement
 - ➡ Montant total
- Commentaires
 - ➡ Plus mauvais avis (satisfaction)



Données clients

Agrégation des données

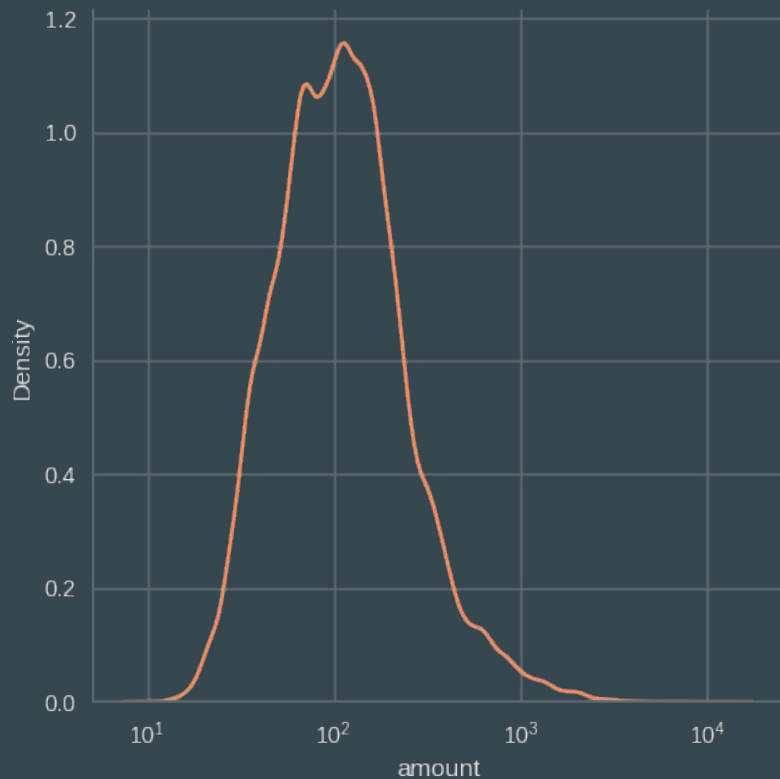
- Date
 - ➔ Récence de l'achat
- Livraison
 - ➔ Retard de la livraison
- Commandes
 - ➔ Nombre de commandes



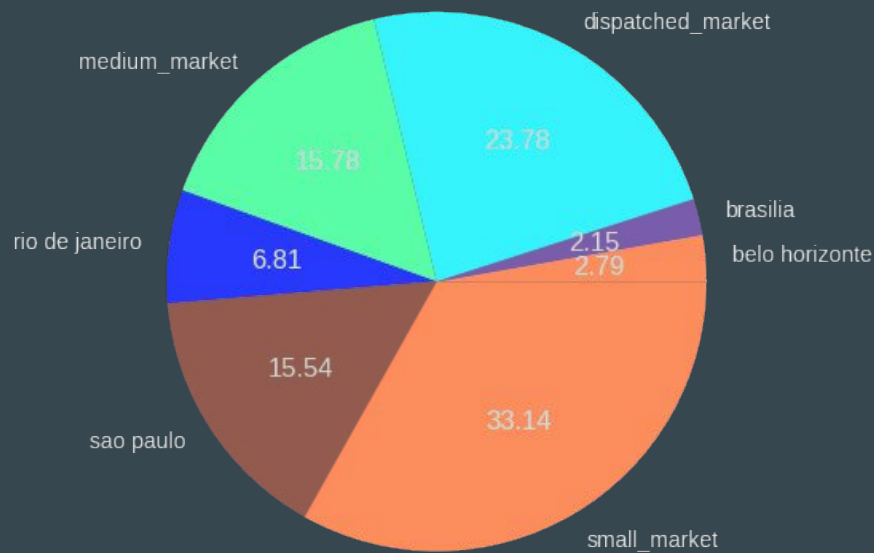


Exploration des données

Analyse univariée

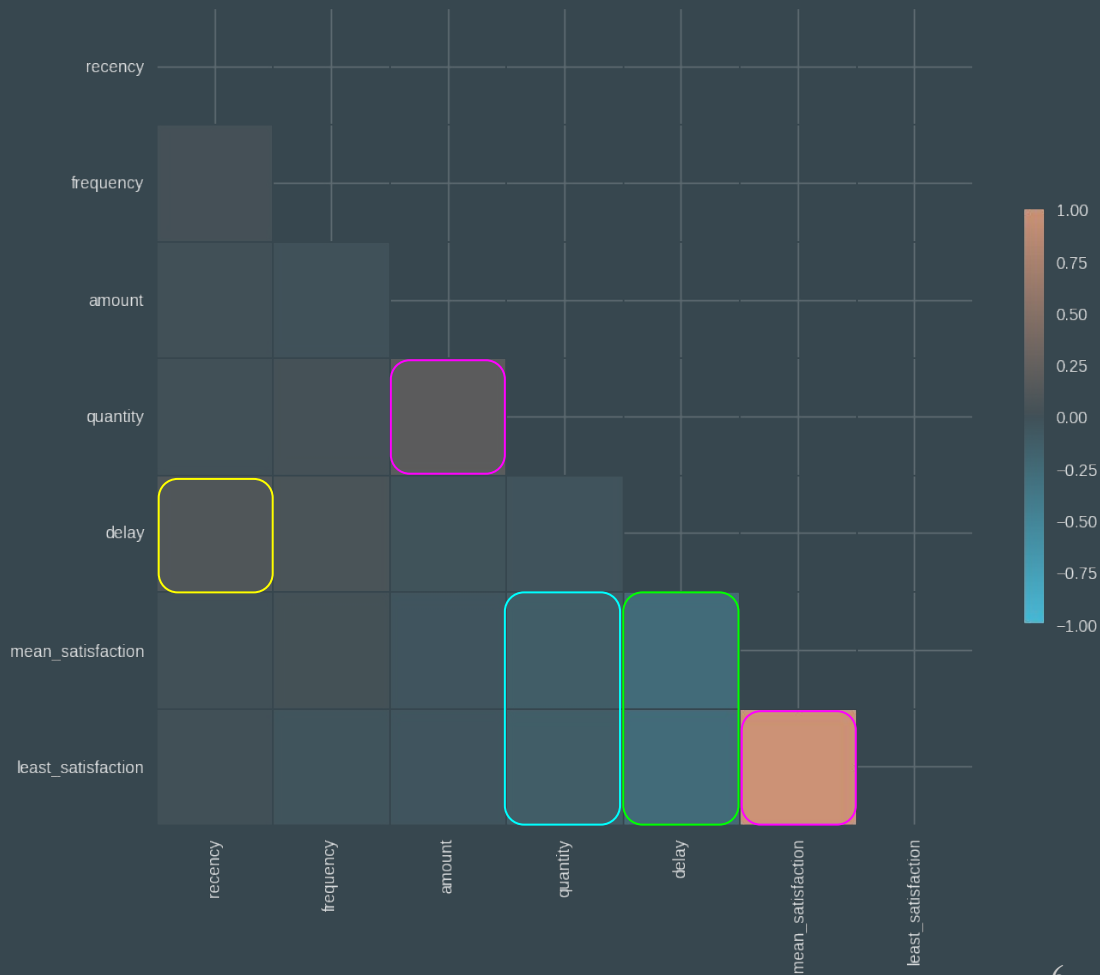


Localisation



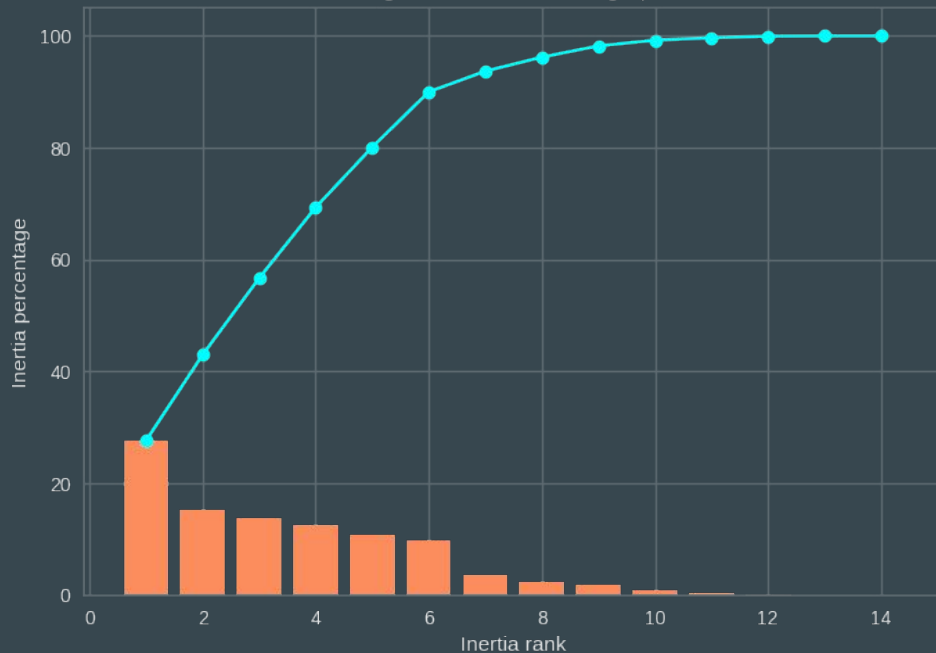
Corrélations

- Corrélations structurelles
- Insatisfaction de la livraison
- Insatisfaction d'un produit
- Dégradation de la livraison ?



Réduction de dimension - PCA

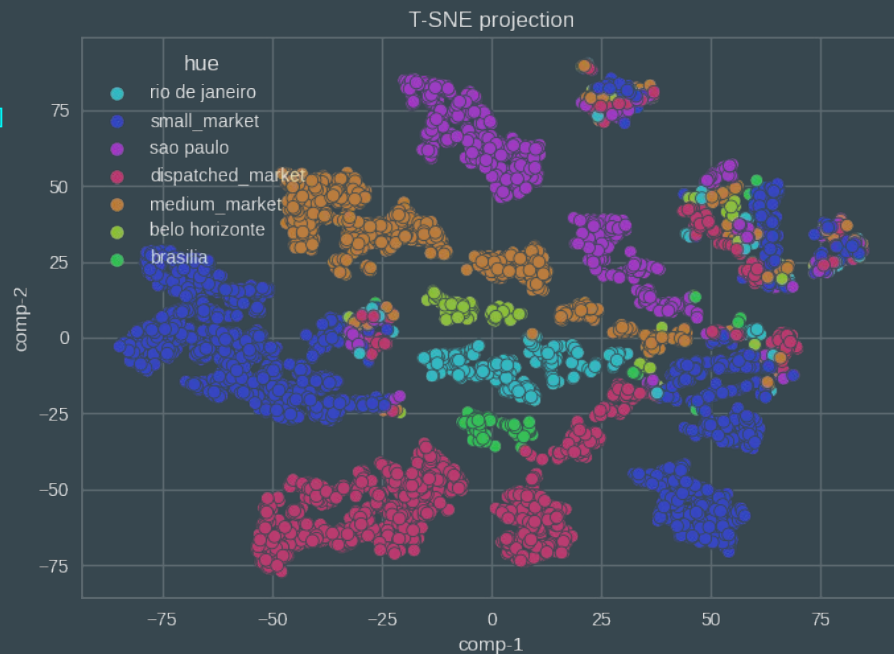
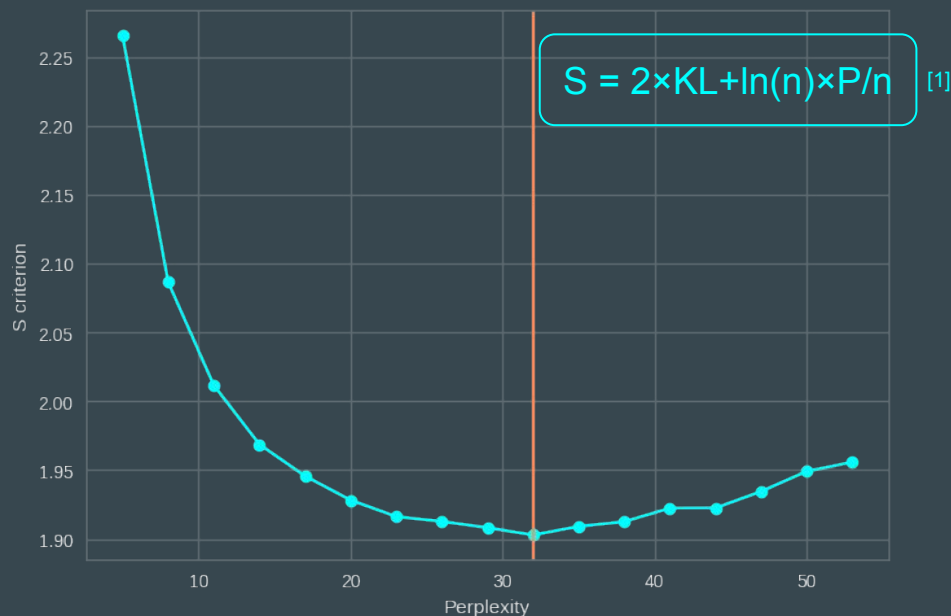
Eigen value cumulative graph



Individuals projection (on F1 & F2)



Réduction de dimension - t-SNE





Modélisation

Contraintes

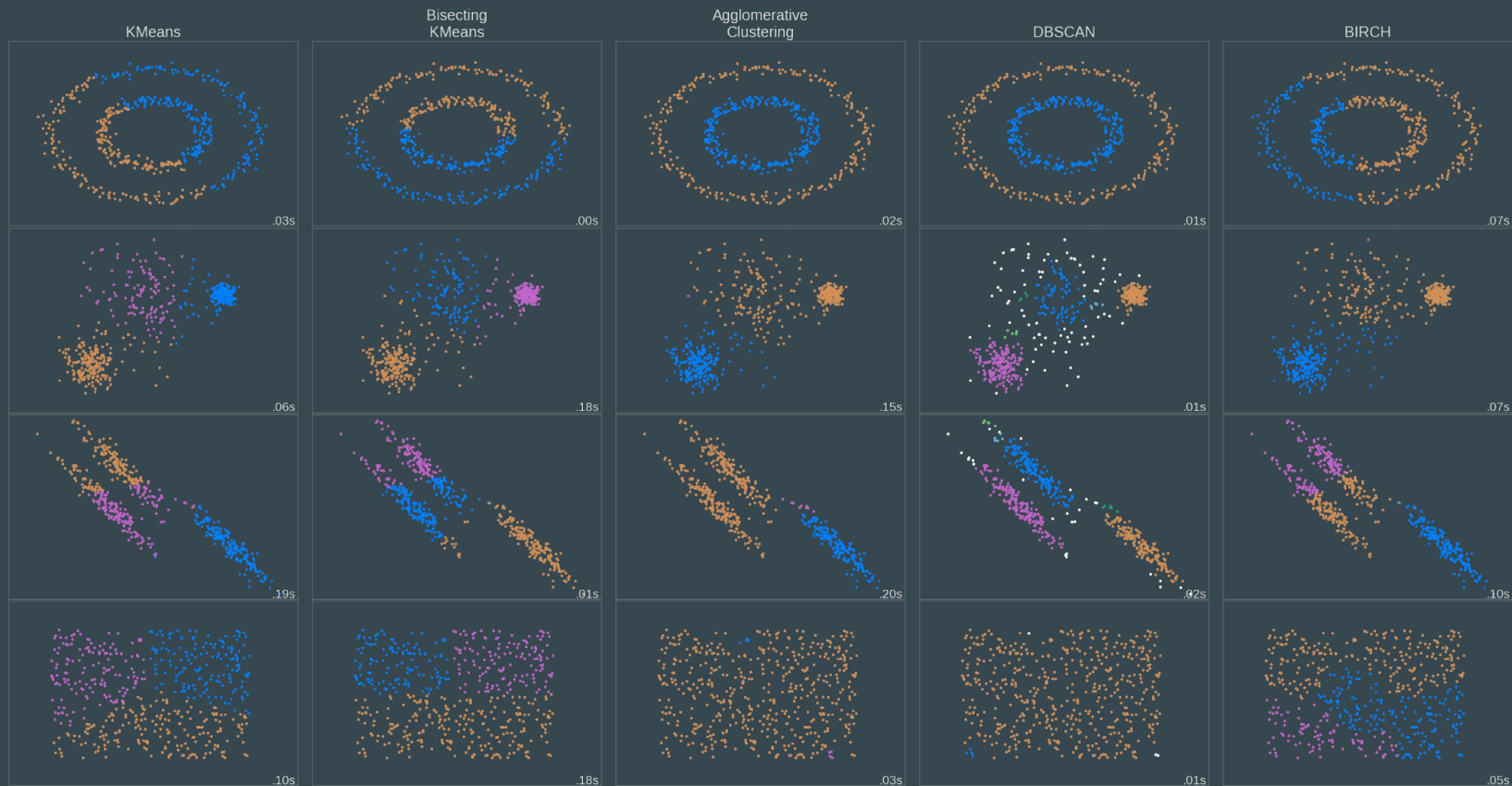
Entraînement d'un modèle

- Stabilité
 - ➡ Modélisation stable
- Quantité des données
 - ➡ Temps d'entraînement
 - ➡ Segmentation pertinente

Segmentation actionnable

- Feature engineering
 - ➡ Explicabilité des relations
- **Forme des clusters**
 - ➡ **Clusters identifiables (nombre)**
 - ➡ **Clusters représentatifs (taille)**

Comparaison initiale



RFM - Filtrage

Aglomerative Clustering

- Taille des données
⇒ Trop lent à entraîner

Birch

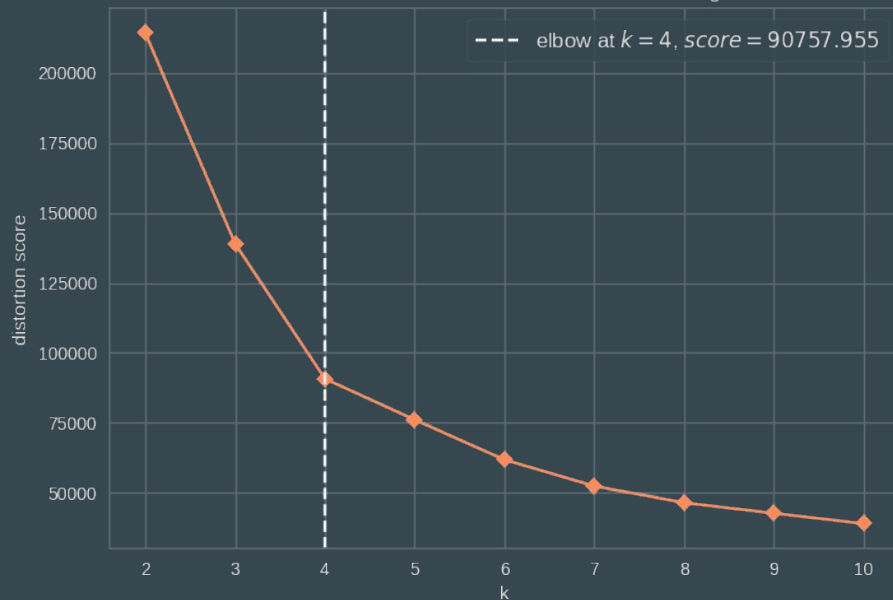
- Taille des clusters
⇒ Clusters trop éparées

DBSCAN

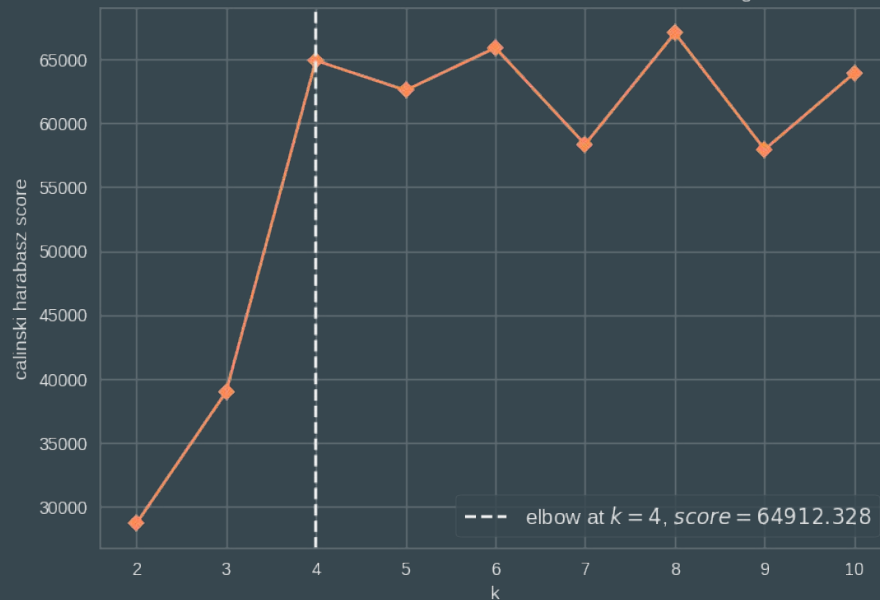
- Taille des données
⇒ Trop lent à entraîner
- Taille des clusters
⇒ Clusters trop éparées
- Nombre de clusters
⇒ Trop de mini-clusters

RFM - KMeans

Distortion Score Elbow for KMeans Clustering

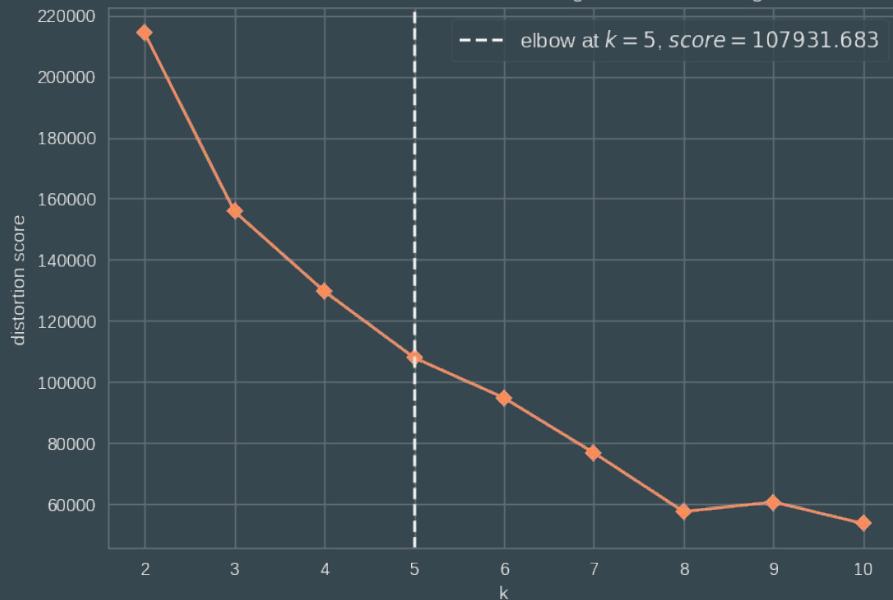


Calinski Harabasz Score Elbow for KMeans Clustering

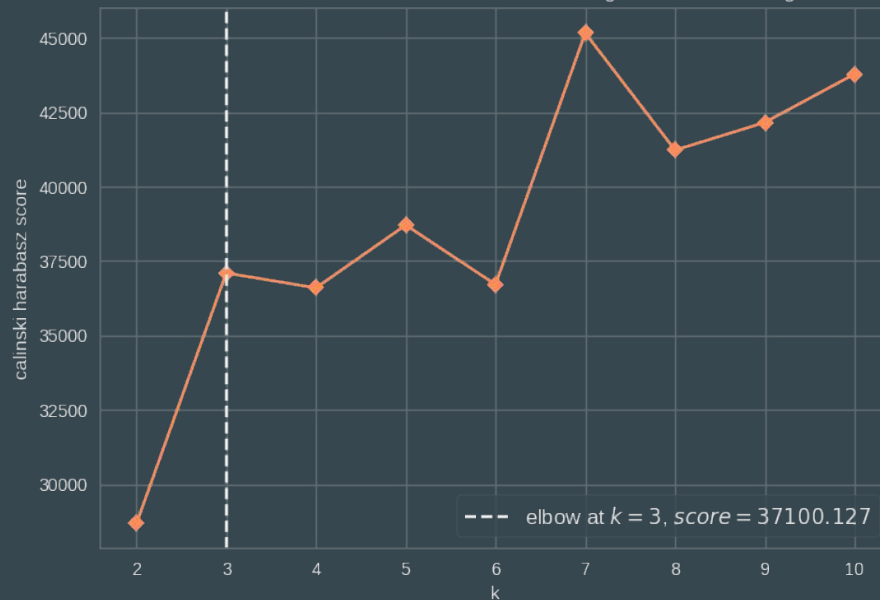


RFM - Bisecting KMeans

Distortion Score Elbow for BisectingKMeans Clustering

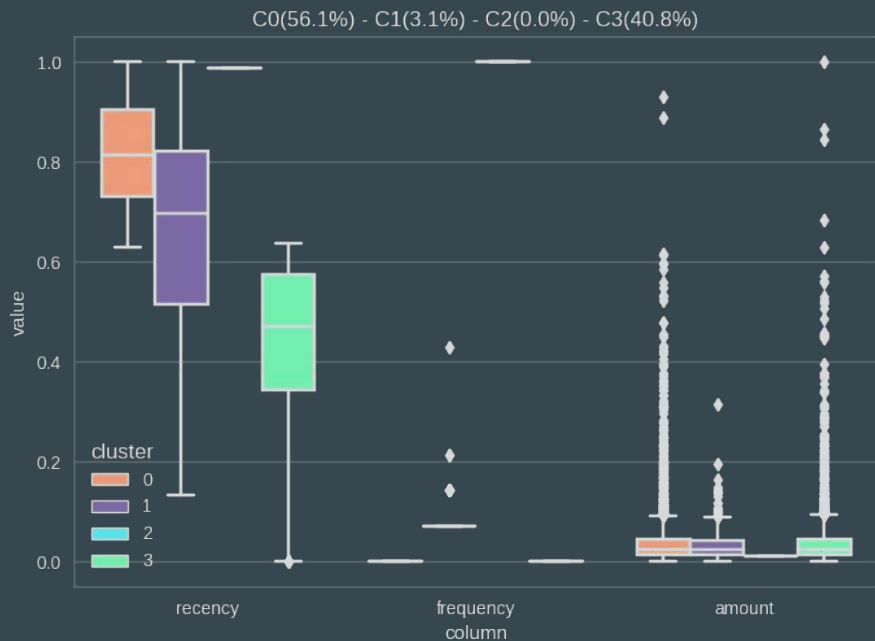


Calinski Harabasz Score Elbow for BisectingKMeans Clustering

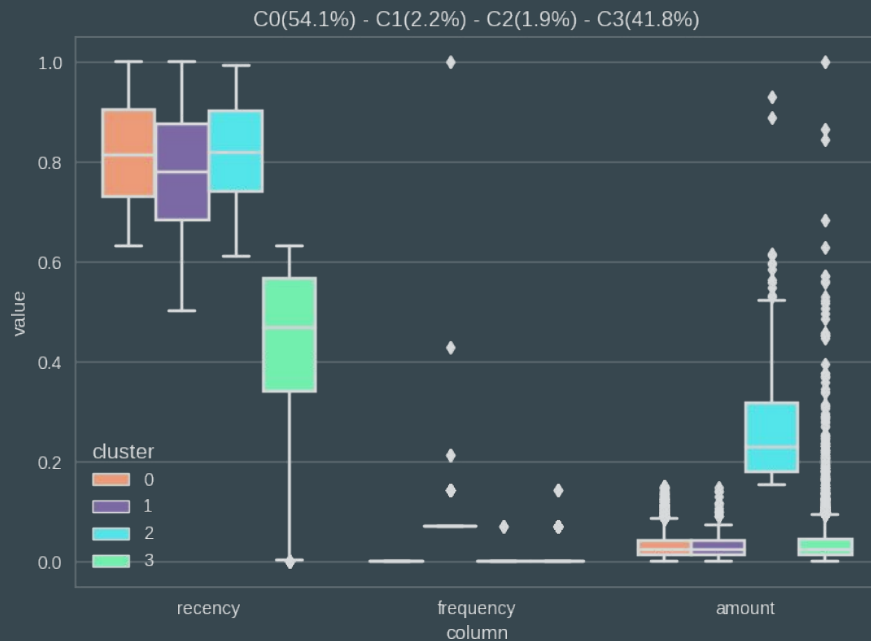


RFM

KMeans k=4

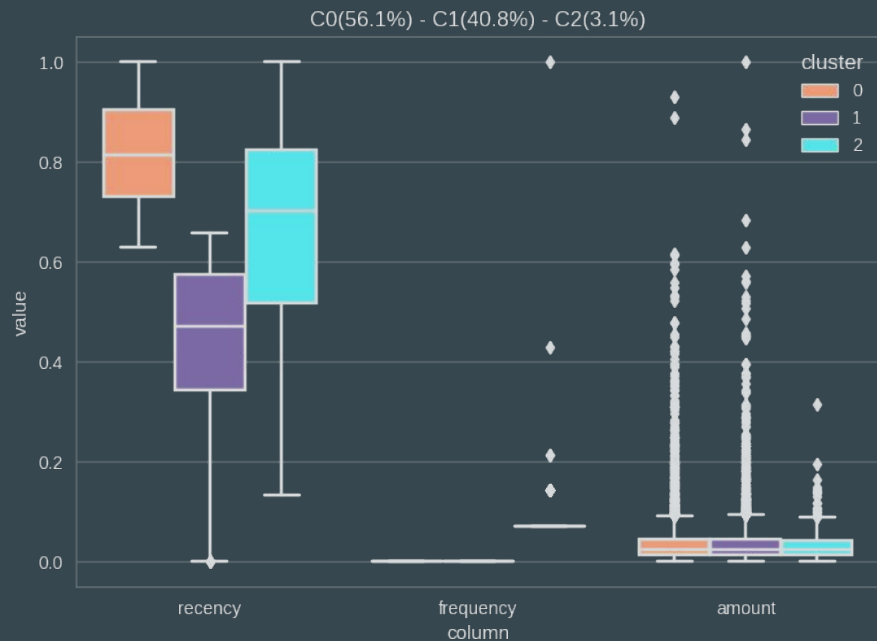


Bisecting KMeans k=4

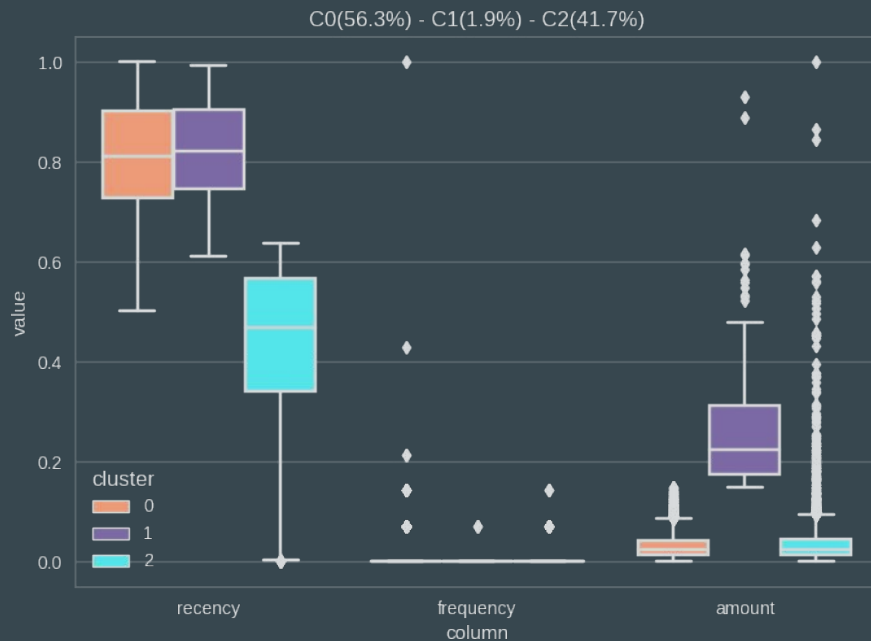


RFM

KMeans k=3



Bisecting KMeans k=3



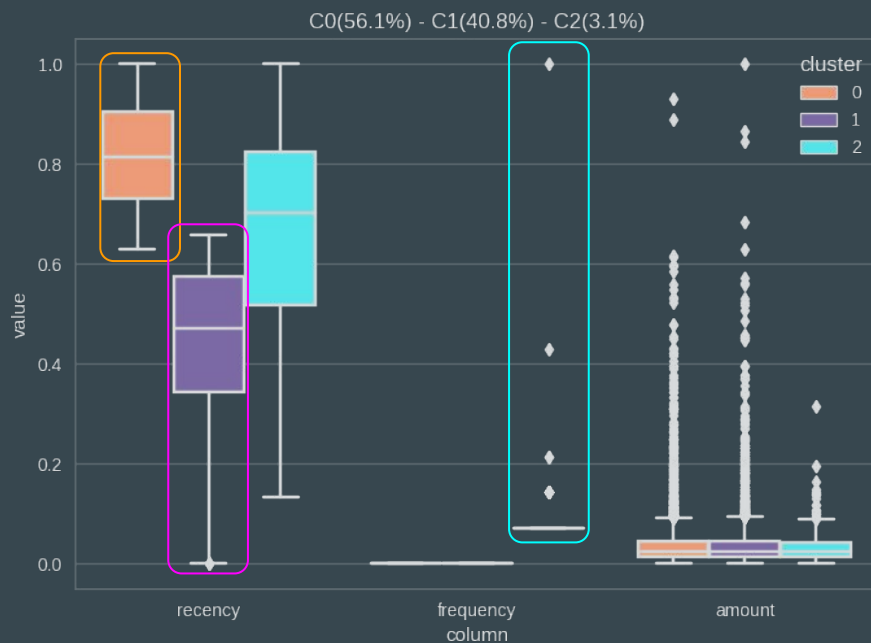


Interprétation

RFM

- Nouveaux clients
- Anciens clients
- Clients récurrents
- Montant pas utilisé ici

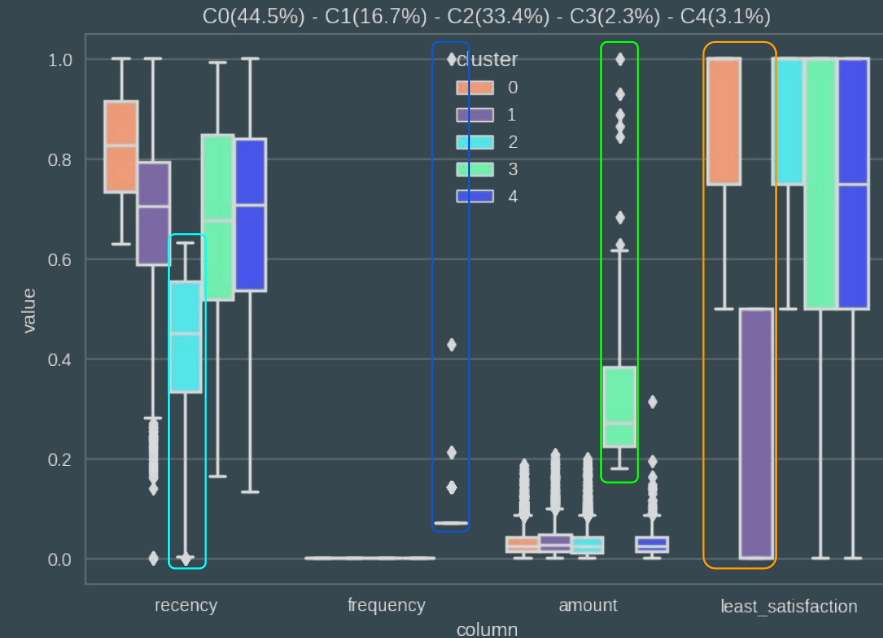
KMeans k=3



Satisfaction

- Partition satisfaction
- Anciens clients
- Clients dépensiers
- Clients récurrents

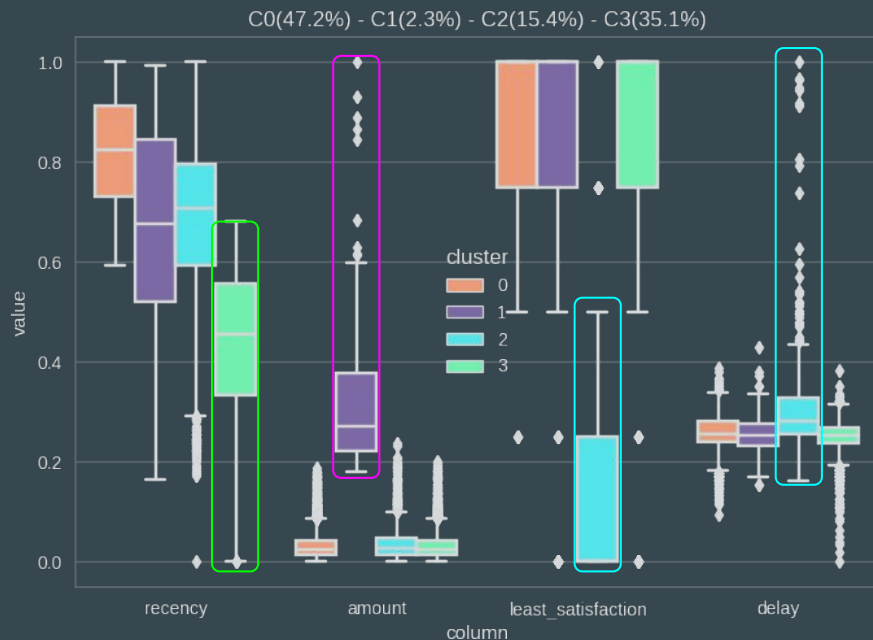
KMeans k=5



Satisfaction - délai

- Reste des clients
- Clients dépensiers
- Clients insatisfaits / délais plus longs
- Ancients clients

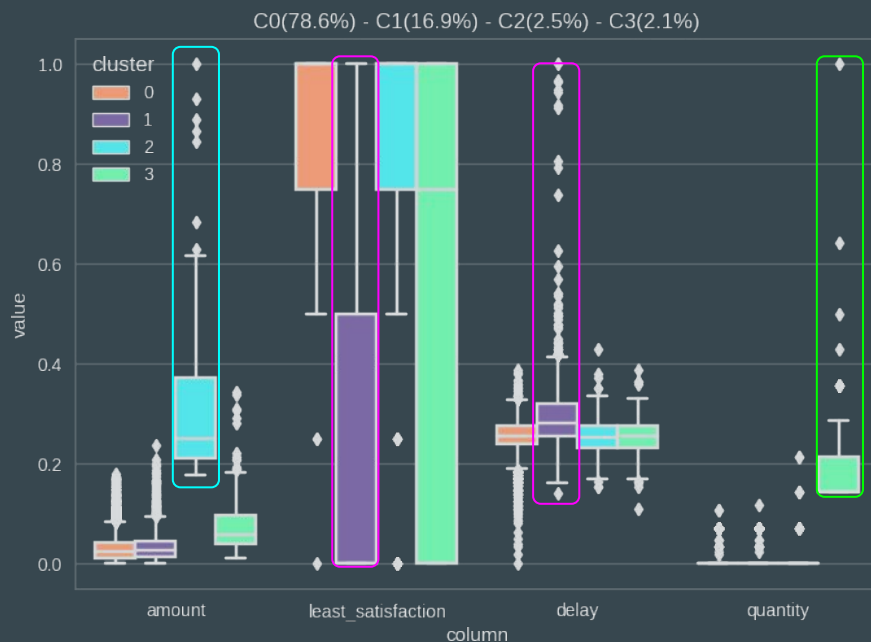
KMeans k=4



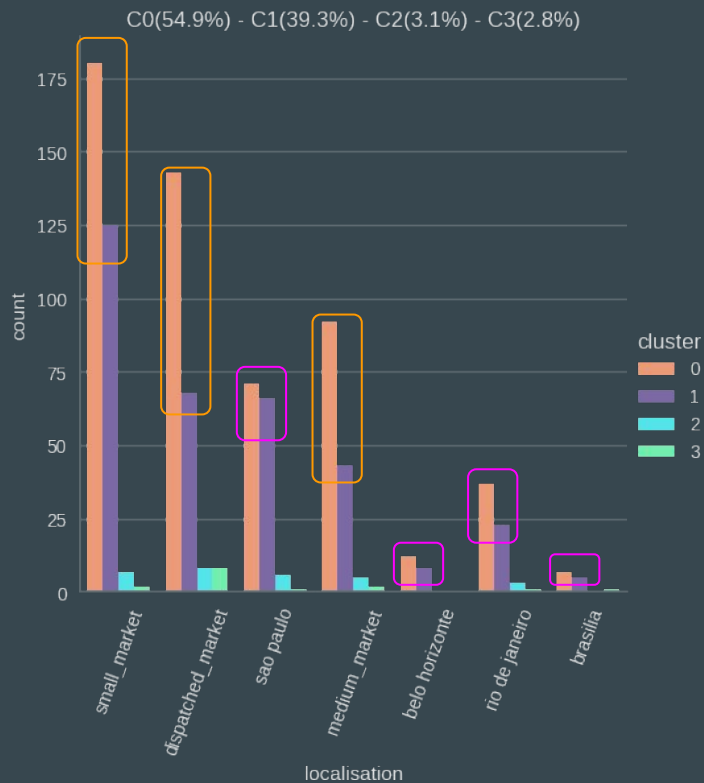
MSDQ

- Reste des clients
- Clients insatisfaits / délais plus longs
- Clients dépensiers
- Clients multi-achats

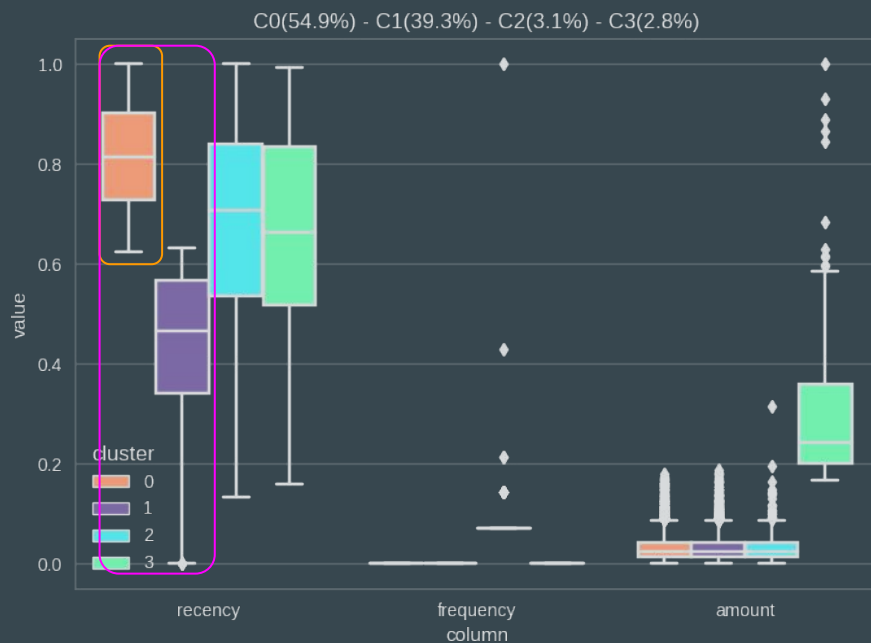
KMeans k=4



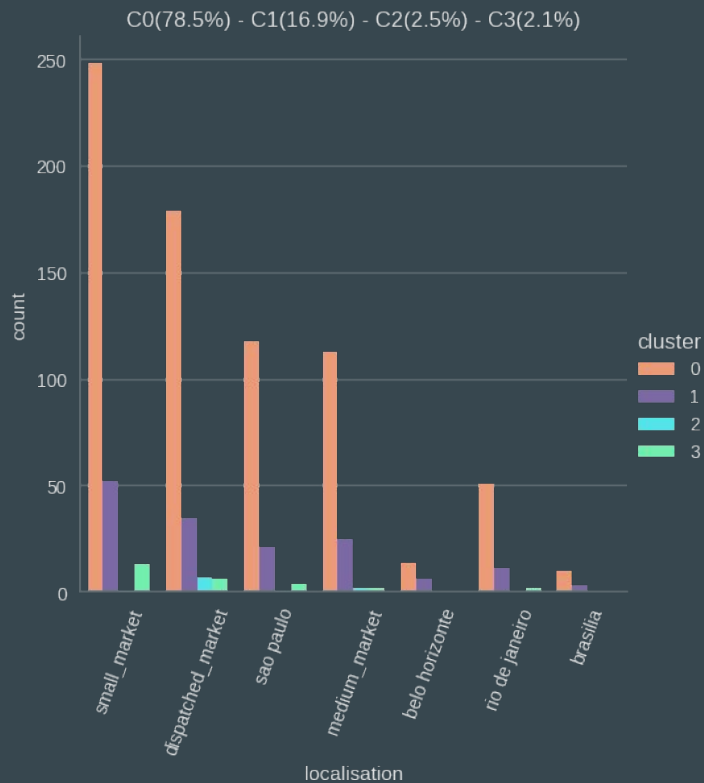
RFM - Localisation



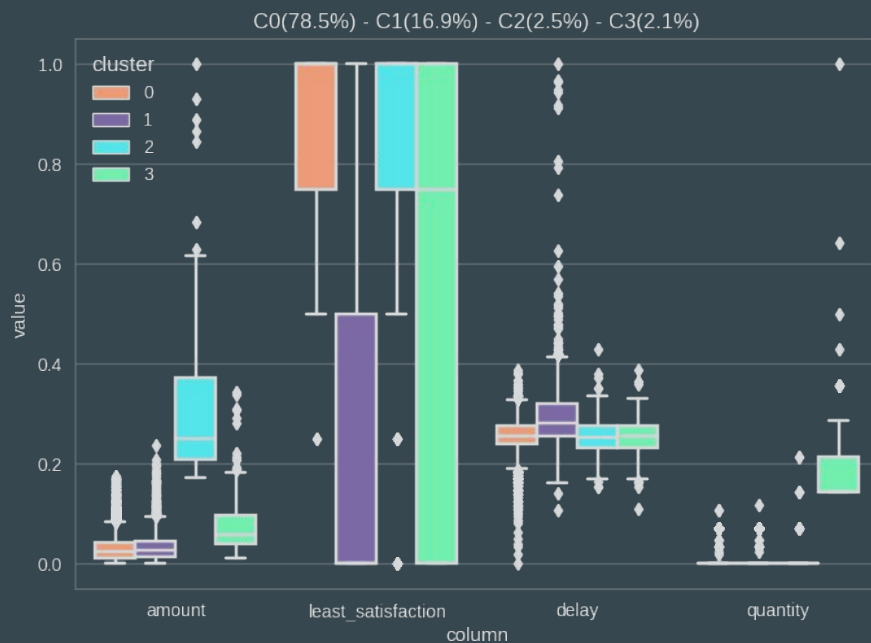
KMeans k=4



MSDQ - Localisation



KMeans k=4



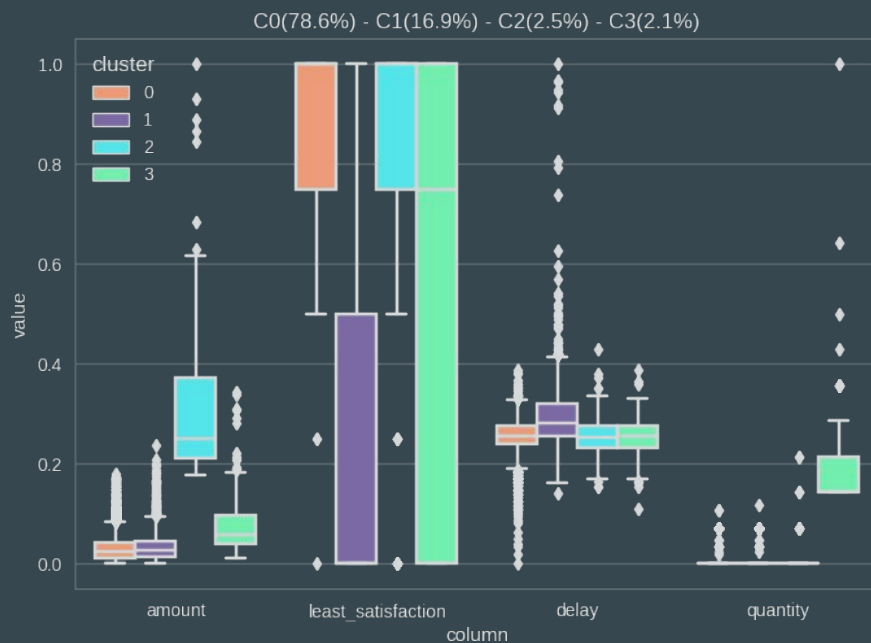


Maintenance

Modèle final

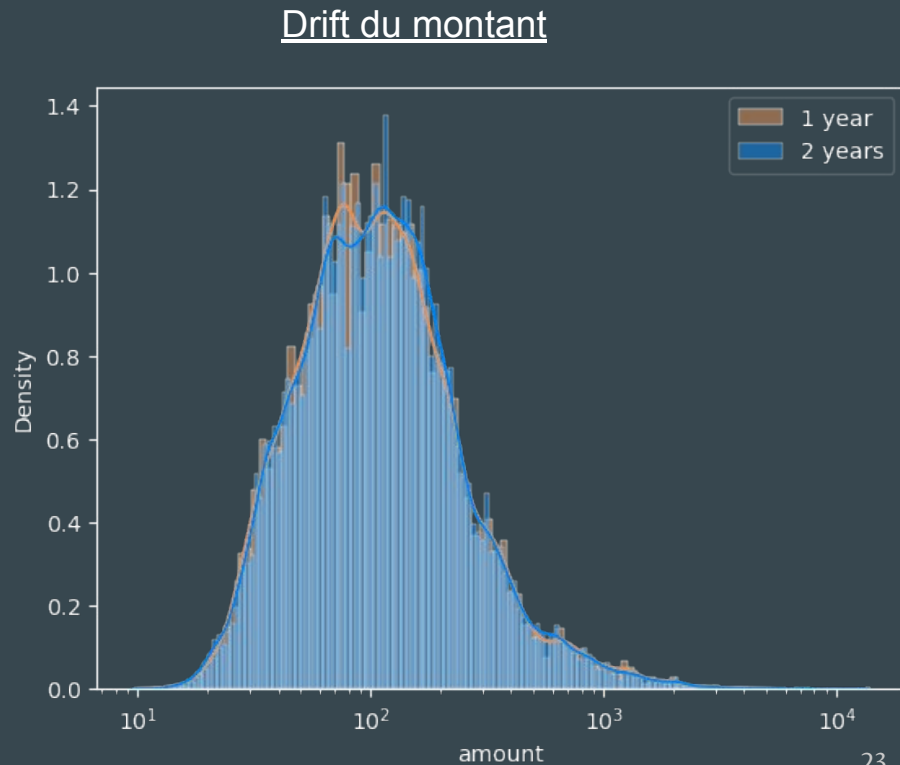
- Clusters identifiables
- Différenciation marketing
- Robuste au data drift

MSDQ - KMeans k=4



Data drift

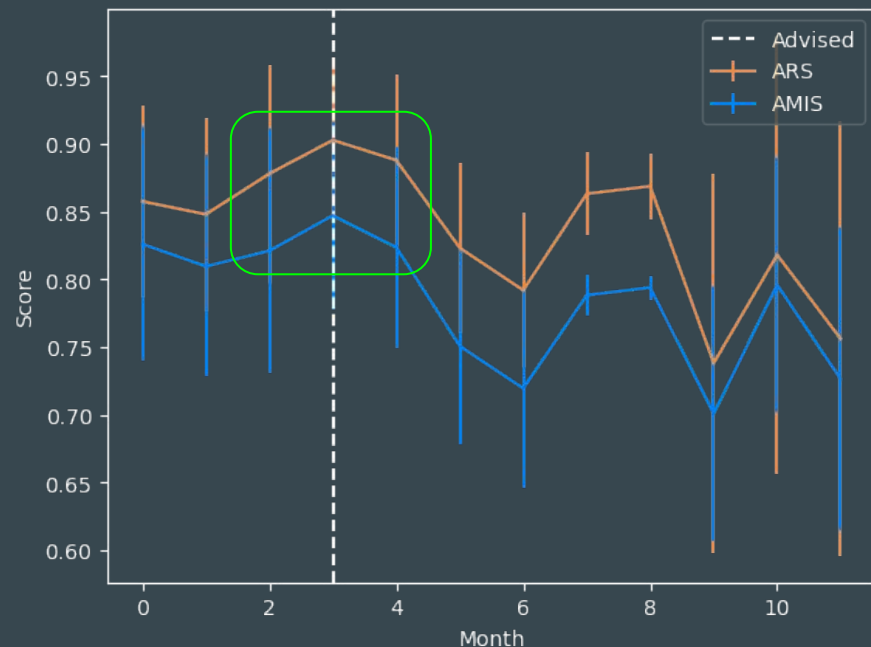
- Drift “naturel”
- Impact des tendances
- Impact des crises



Recommandation

- Scores ARI et AMI
- Plage de maintenance possible
- Robustesse face au drift
- Instabilité notable

Courbe de maintenance





Conclusion

Conclusion générale

Analyse rétrospective

- Segmentation
 - ➔ Actionnable mais déterminée par les variables retenues
- Maintenance
 - ➔ Fortement dépendante des variables retenues
- Algorithmes
 - ➔ Pousser les comparaisons

Axes d'amélioration

- Feature engineering
 - ➔ Ajouter un encoding numérique des vendeurs/villes
 - ➔ Segmenter selon les habitudes d'achats (différents produits)
- Maintenance
 - ➔ Analyser la segmentation dans le temps

Merci de votre attention.

...

Des questions ?