# Segmentation des clients d'un site e-commerce Olist

Projet N° 4

Irina Maslowski

# Plan

- Problématique et les pistes de recherche envisagées
- Données et leur analyse
- Pistes de modélisation
- Segmentation des clients
- Conclusion

# Problématique et les pistes de recherche envisagées

- **Olist** – marketplace en ligne

https://olist.com/

- **Implantation**: Brésil

- Propose aux petits commerçant

une **plateforme de vente** et des

**solutions d'expédition** des produits aux clients.

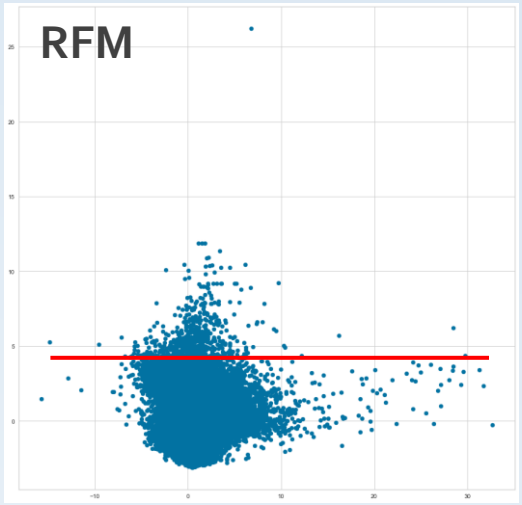
# Problématique et les pistes de recherche envisagées

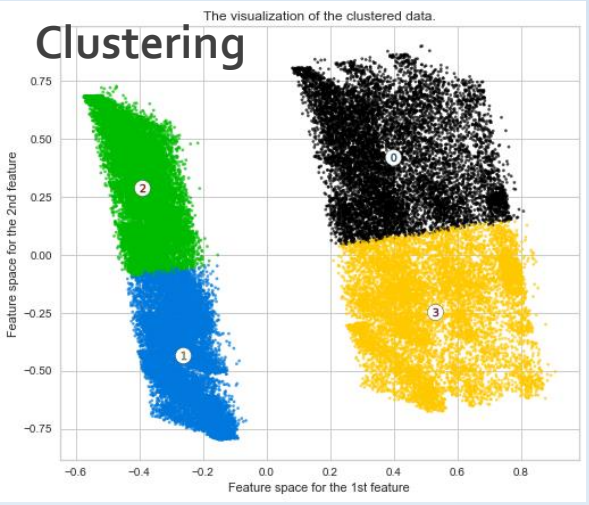
- **Cahier des charges:**
  - Proposer une segmentation des clients utilisable pour les campagnes marketing
  - Fournir à l'équipe marketing une description actionnable de la segmentation
  - Proposer un contrat de maintenance
- **Moyens fournis**:
  - une base de données anonymisée comportant des informations sur l'historique de commandes, les produits achetés, les commentaires de satisfaction, et la localisation des clients

# Problématique et les pistes de recherche envisagées

- Choix du type de segmentation client:
  - « manuelle » : RFM (médiane/quartiles).

    RFM est la segmentation RFM prend en compte la Récence (date de la dernière commande), la Fréquence des commandes et le Montant (de la dernière commande ou sur une période donnée) pour établir des segments de clients homogènes. (https://www.definitions-marketing.com/definition/segmentation-rfm/)
  - non-supervisée: clustering



# Problématique et les pistes de recherche envisagées

- Trame de travail envisagée:
  - Prise-en-main des données (agrégation, nettoyage)
  - Analyse des données
  - Test des algorithmes de clustering
  - Choix d'une solution le mieux adaptée
  - Evaluation des résultats obtenus
  - Description de la segmentation obtenue
  - Proposition d'un contrat de maintenance en fonction de stabilité du modèle

# Plan

- Problématique et les pistes de recherche envisagées
- Données et leur analyse
- Pistes de modélisation
- Segmentation des clients
- Conclusion

# Données et leur analyse

- Données réelles, anonymisées
- Taille: 100k commandes
- Période temporel: 2016 – 2018
- Type d'information: état de commande, prix, payement, dates de livraison, informations sur les produits, géolocalisation des clients et des vendeurs et des avis des clients.

# Données et leur analyse

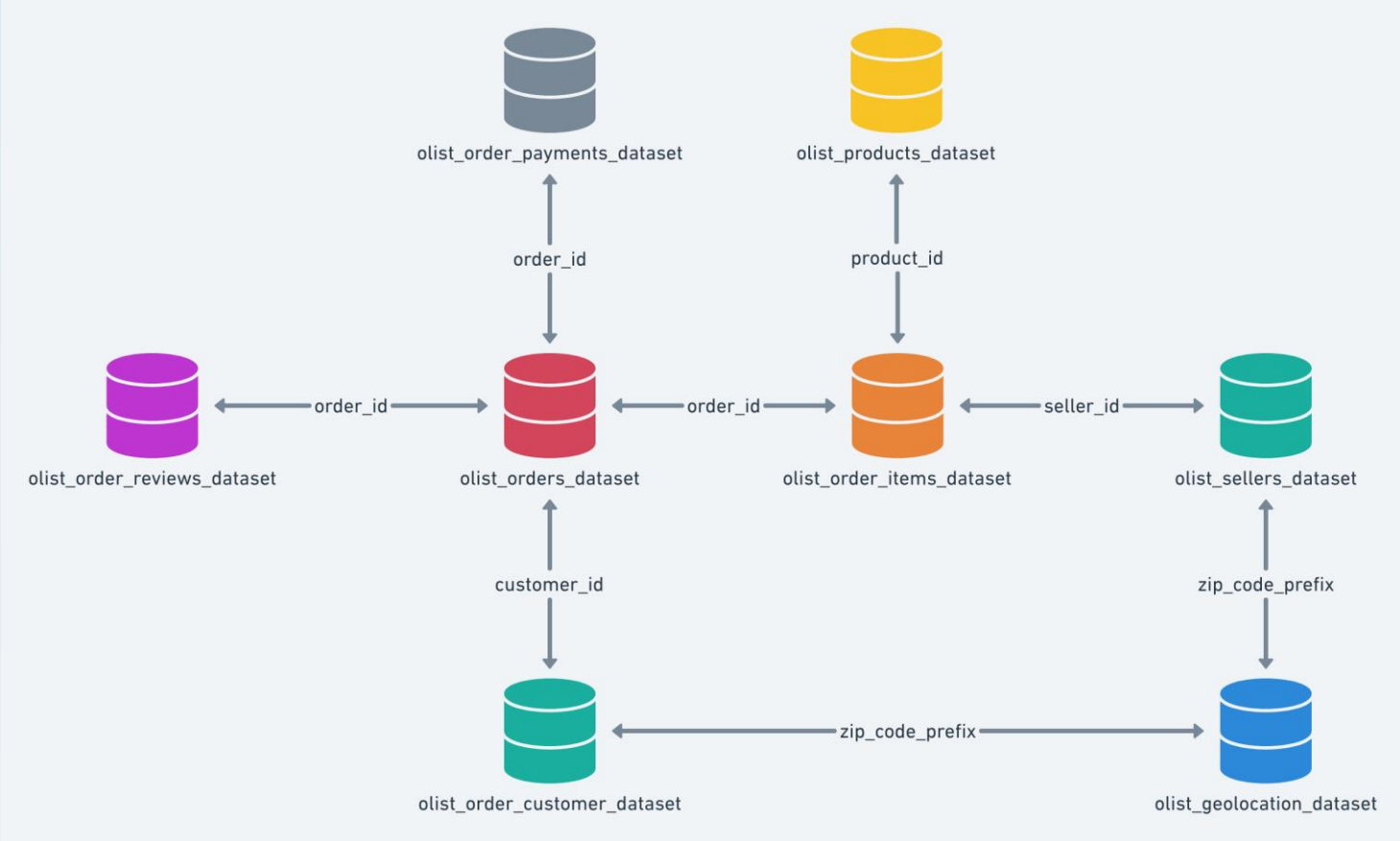
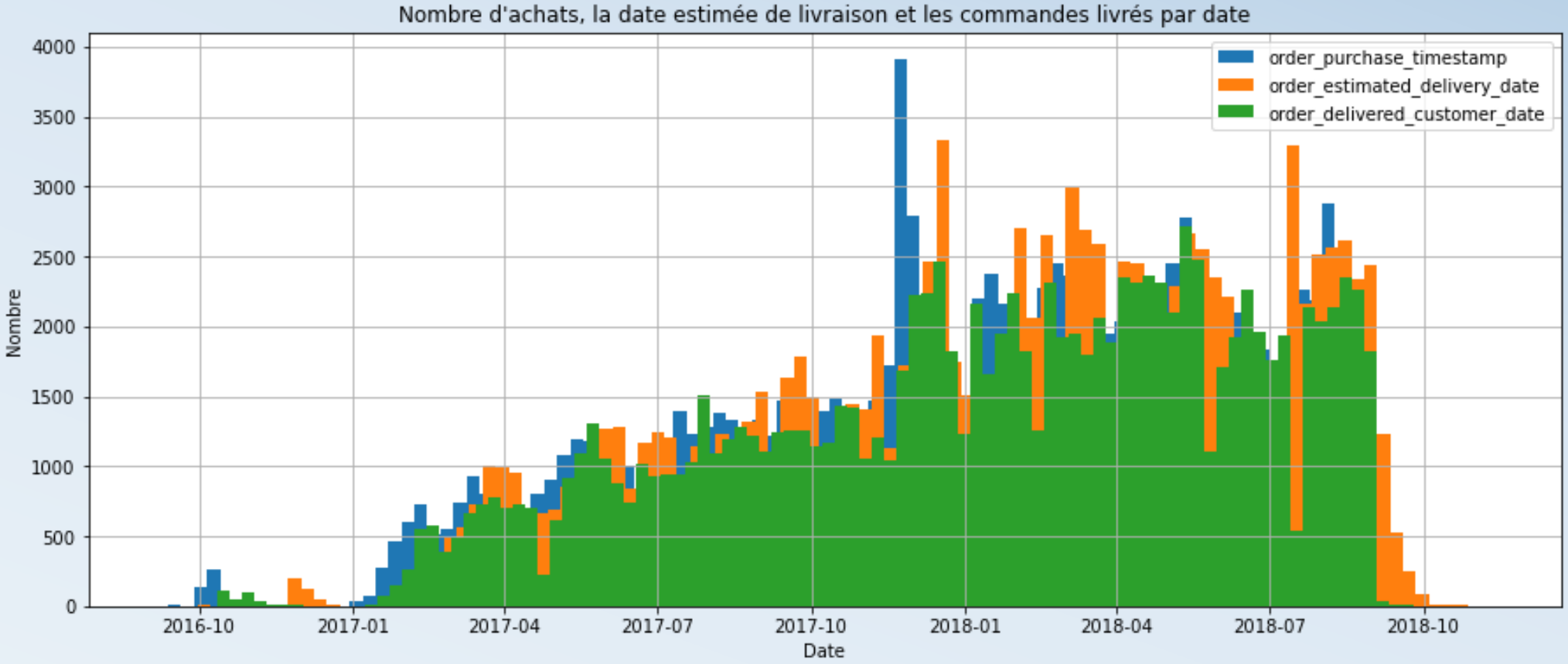



Schéma de la base de donnée

# Données et leur analyse

- Étapes de travail avec les données:
  - Assemblage de tous les dataframes
  - 5 nouvelles variables crées:
    - Temps depuis la commande
    - Distance entre le client et le vendeur
    - Durée de livraison
    - Volume des produits
    - Livraison à temps

# Données et leur analyse

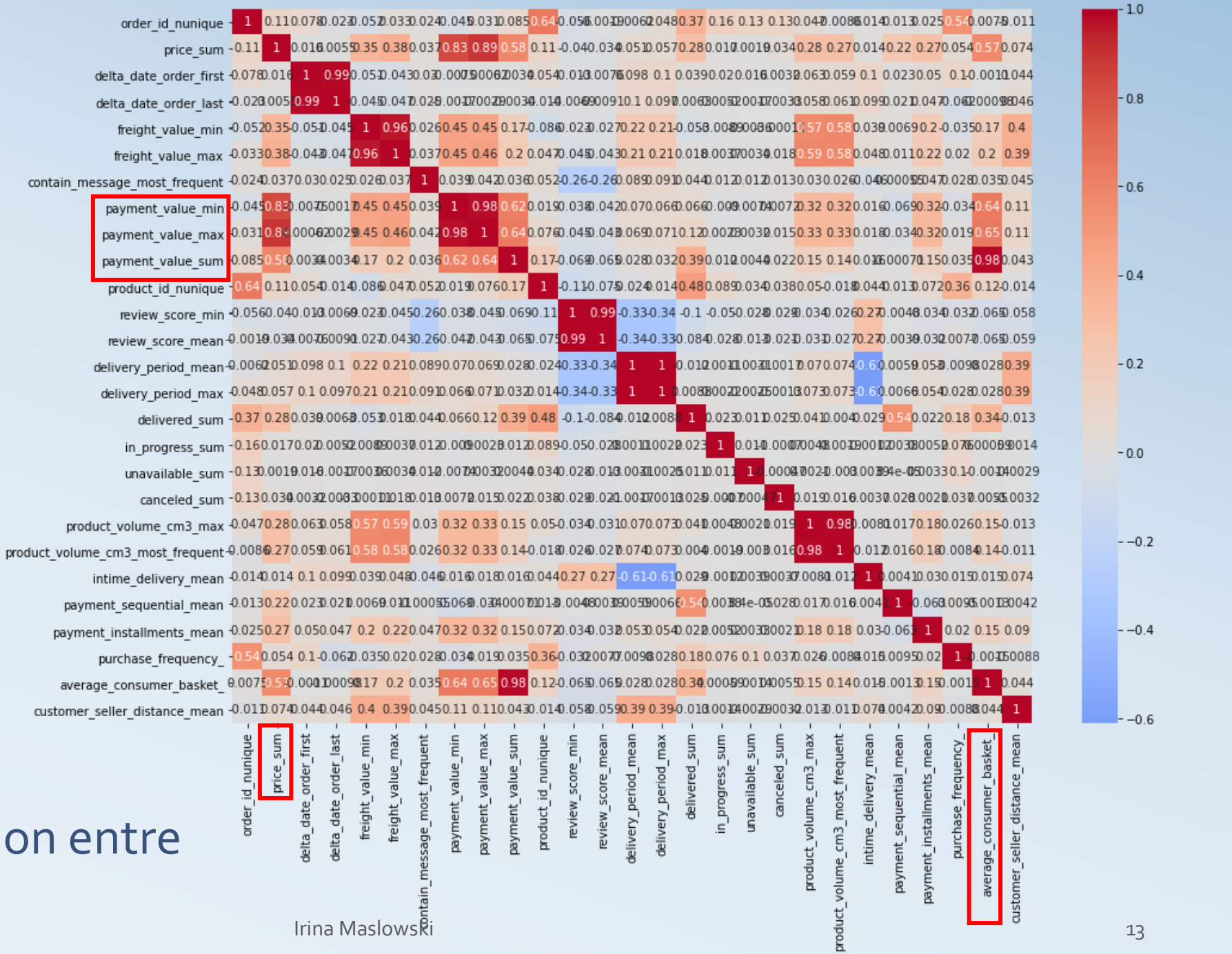



Evolution de l’activité de marketplace dans le temps

# Données et leur analyse

- Étapes de travail avec les données:
  - Agrégation des données par client
  - Création de 3 nouvelles variables:
    - Fréquence d'achats
    - Panier moyen de client
    - Catégories de produits plus générales
  - Trie des variables en fonction de leur pertinence (trop corrélées, les variables qualitatives)

# Données et leur analyse




## Heatmap de corrélation entre les variables

# Données et leur analyse

| 13 variables | 7 variables |
|---|---|
| nombre d'achats | nombre d'achats |
| temps depuis la première commande | temps depuis la première commande |
| temps depuis la dernière commande | temps depuis la dernière commande |
| note minimal de l'avis client | note minimal de l'avis client |
| livraison à temps (moyenne) | livraison à temps (moyenne) |
| temps passé entre deux commandes (moyenne) | temps passé entre deux commandes (moyenne) |
| prix maximal de livraison | somme dépensée |
| durée maximale de livraison | |
| taille le plus fréquent de produit | |
| nombre de payements séquentiels en moyenne | |
| distance client – vendeur | |
| panier moyen | |
| nombre d'étalement des payements en moyenne | |

## Deux dataframes pour les tests

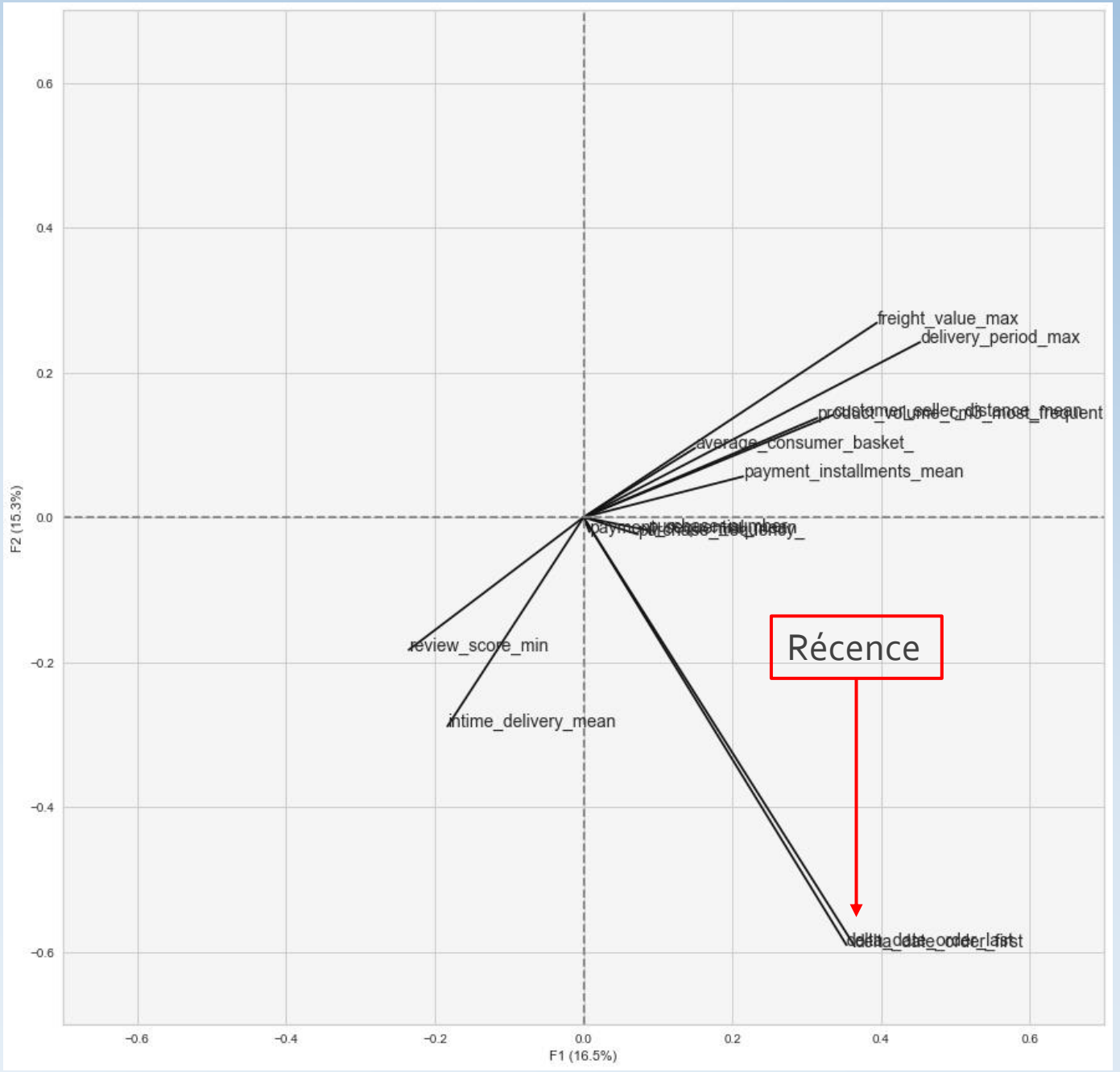
# Plan

- Problématique et les pistes de recherche envisagées
- Données et leur analyse
- Pistes de modélisation
- Segmentation des clients
- Conclusion

# Pistes de modélisation

- Analyse en composantes principales (ACP)
- Normalisation des données
- Algorithmes de clustering: k-means, classification ascendante hiérarchique (CAH) et DBSCAN
- Moyens de visualisation: librairie python « yellowbrick », ACP et TSNE

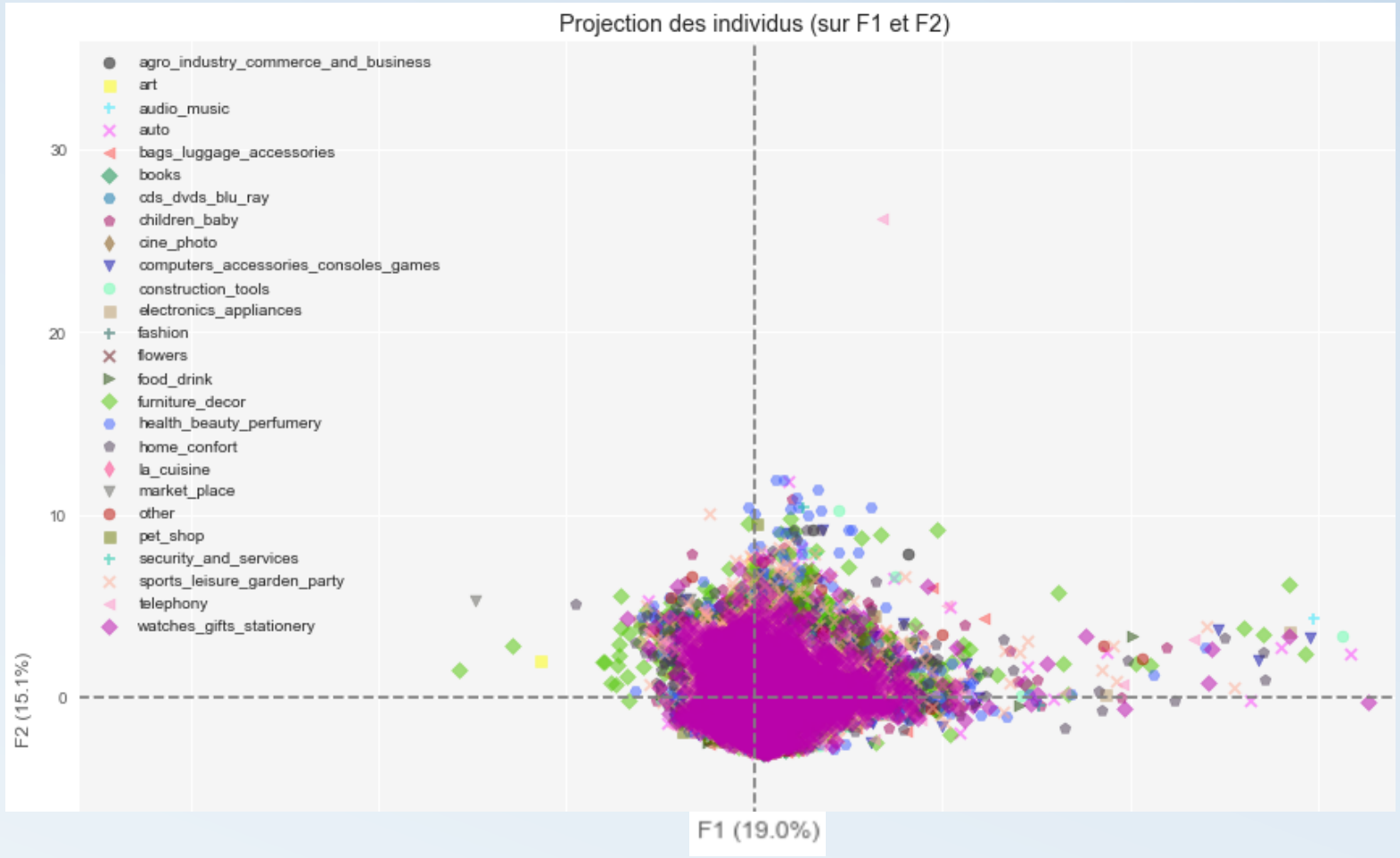
# Pistes de modélisation

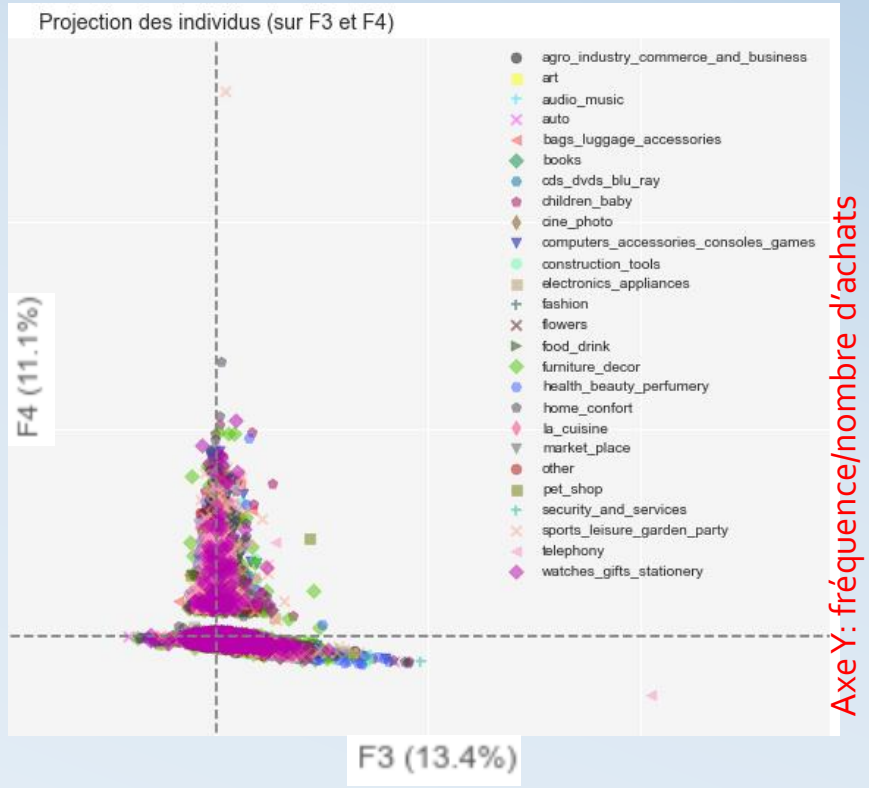



Deux premiers plans ACP pour les 13 variables

# Pistes de modélisation

**Projection des 4 premiers plans de l'ACP.**
Variable illustrative: catégories générales les plus fréquentes des produits




Axe Y: fréquence/nombre d'achats

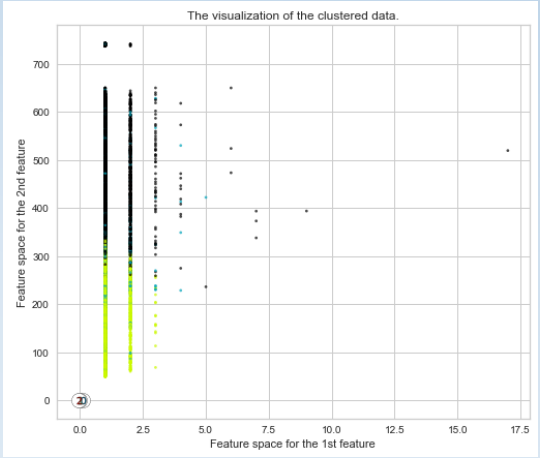
# Pistes de modélisation

- Normalisation des données

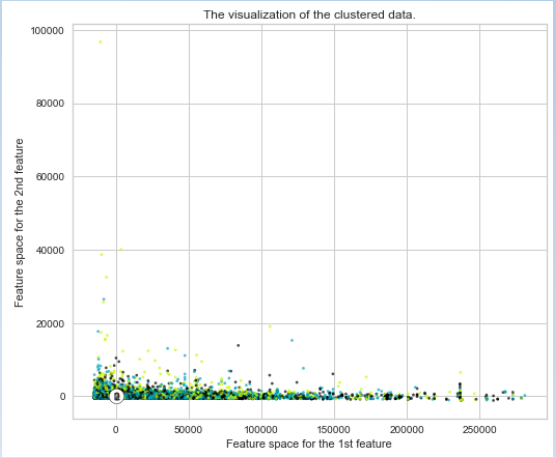
  Les types de standardisation testés avec k-means:

  - StandardScaler
  - MaxAbsScaler
  - MinMaxScaler
  - RobustScaler
  - QuantileTransformer

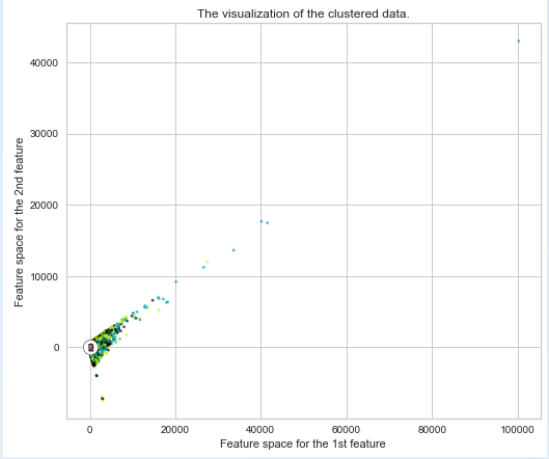
StandardScaler



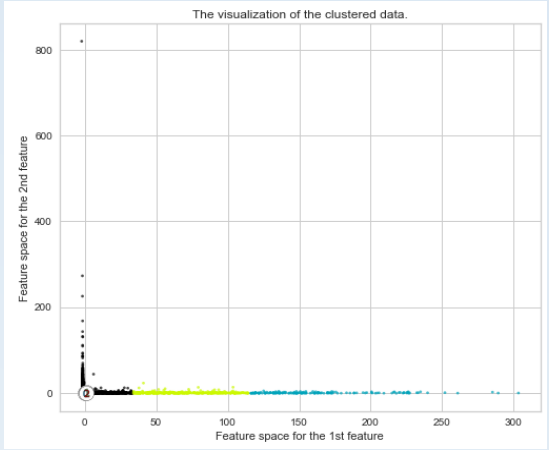

MaxAbsScaler



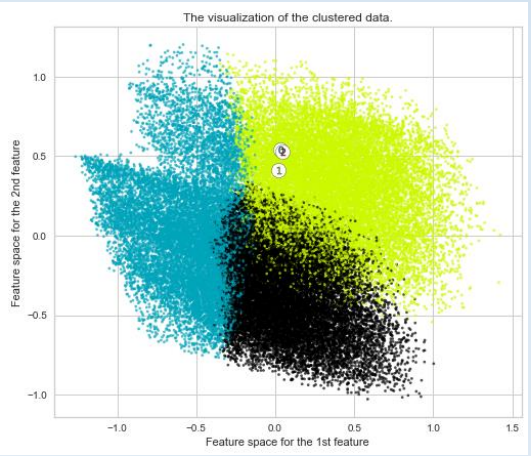

MinMaxScaler




RobustScaler




QuantileTransformer



# Pistes de modélisation

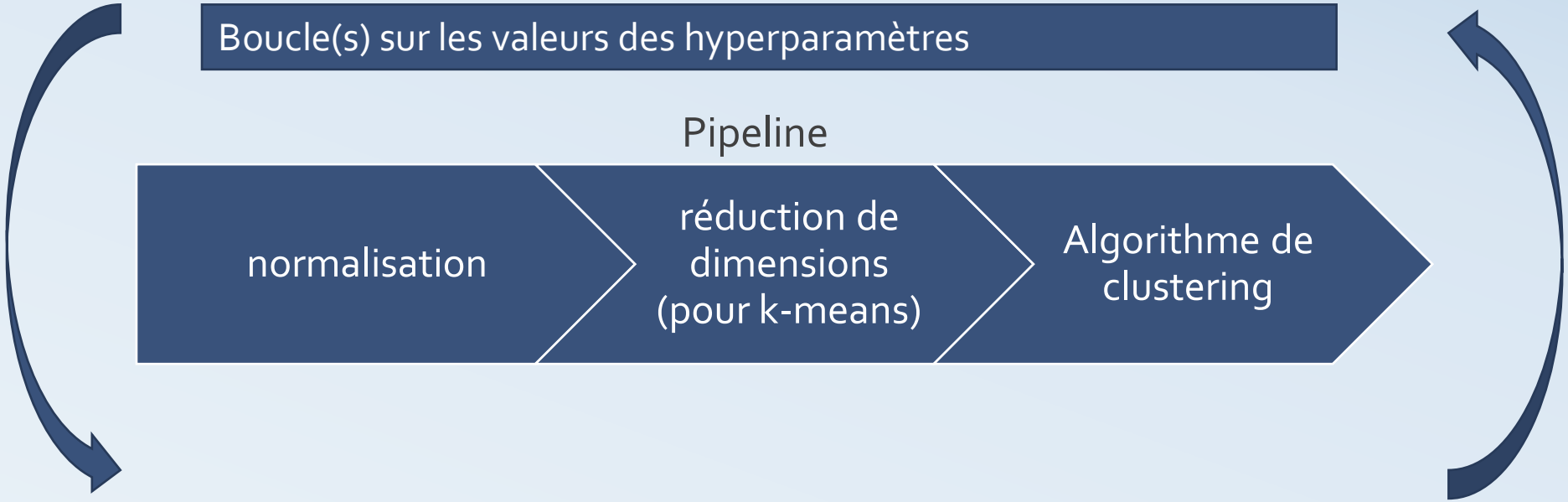
- Optimisation des hyperparamètres

| Algorithme | Hyperparamètres |
|---|---|
| K-means | k, le nombre de clusters<br>n, le nombre de composantes ACP |
| CAH* | k, le nombre de clusters |
| DBSCAN | eps, epsilone<br>ms, min_samples |

*CAH est utilisé avec les résultats de K-means (k=1 000) en entrée

# Pistes de modélisation

- Optimisation des hyperparamètres



# Pistes de modélisation

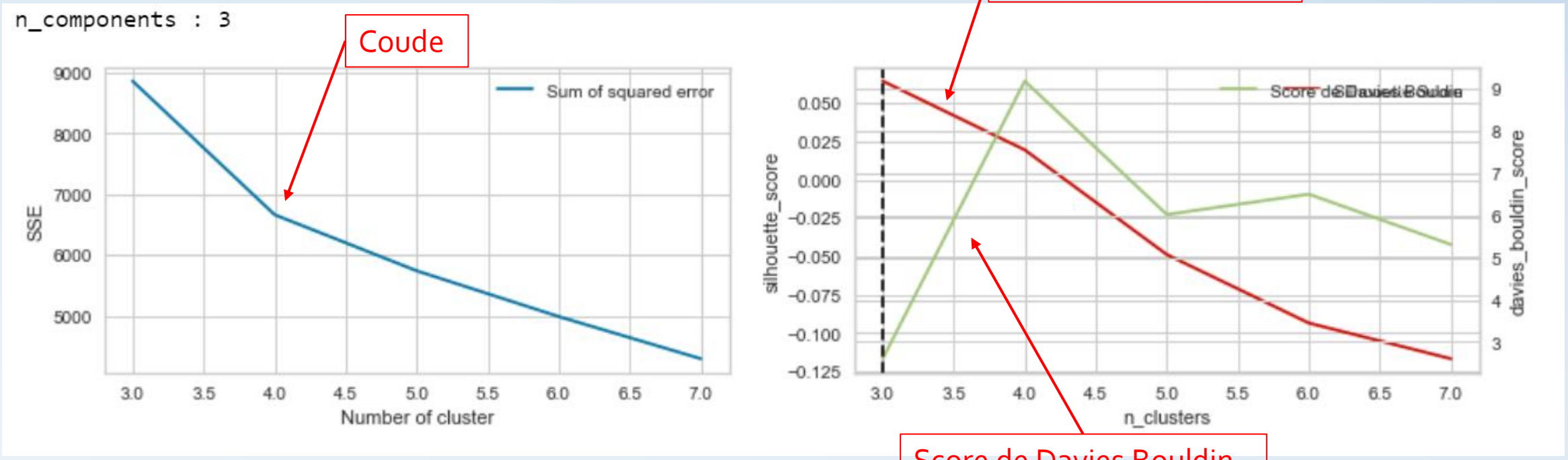
- Moyens d'évaluation d'algorithmes:

| | Coude | Score de silhouette | Score de Davis Bouldin |
|---|---|---|---|
| K-means | Pour l'inertie (somme des carrés des erreurs) | oui | oui |
| CAH* | non | oui | oui |
| DBSCAN | Pour le choix de meilleur epsilone | oui | oui |

*CAH est utilisé avec les résultats de K-means (k=1 000) en entrée

# Pistes de modélisation

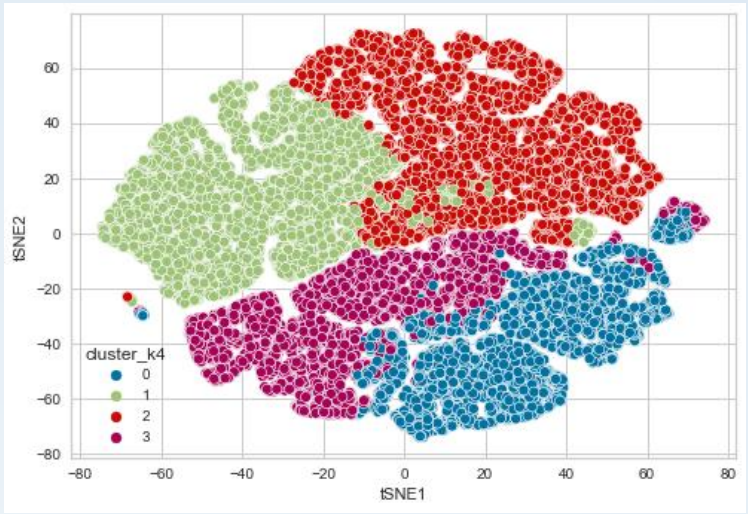
- Optimisation des hyperparamètres de k-means sur le dataframe de 7 variables



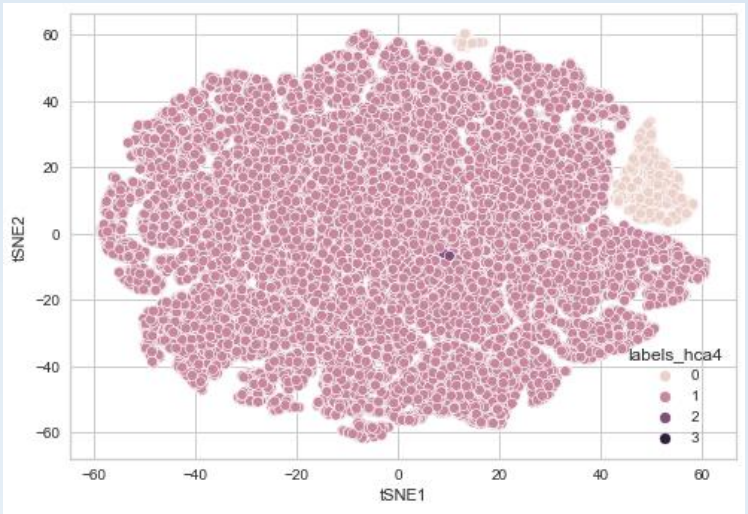
# Pistes de modélisation

- Meilleurs segmentations obtenues pour chaque algorithme sur le dataframe de 7 variables :

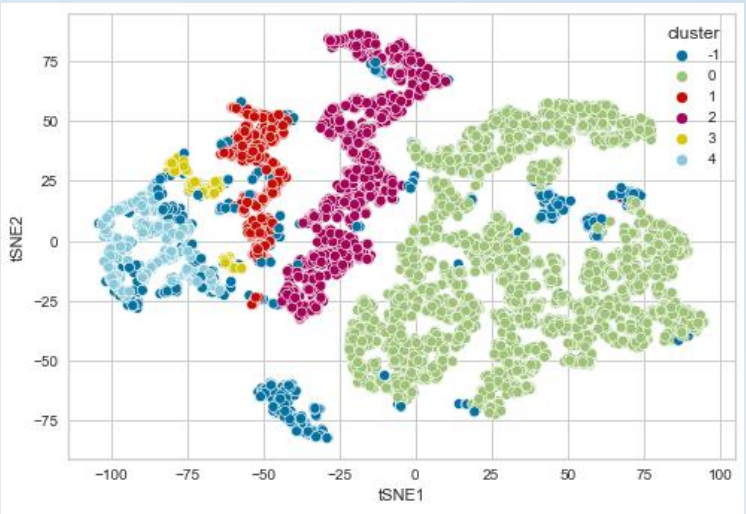
K-means : n_composants = 3, k = 4




K-means (k= 1 000) + CAH : k = 4




DBSCAN: eps = 0,6; min_samples=12



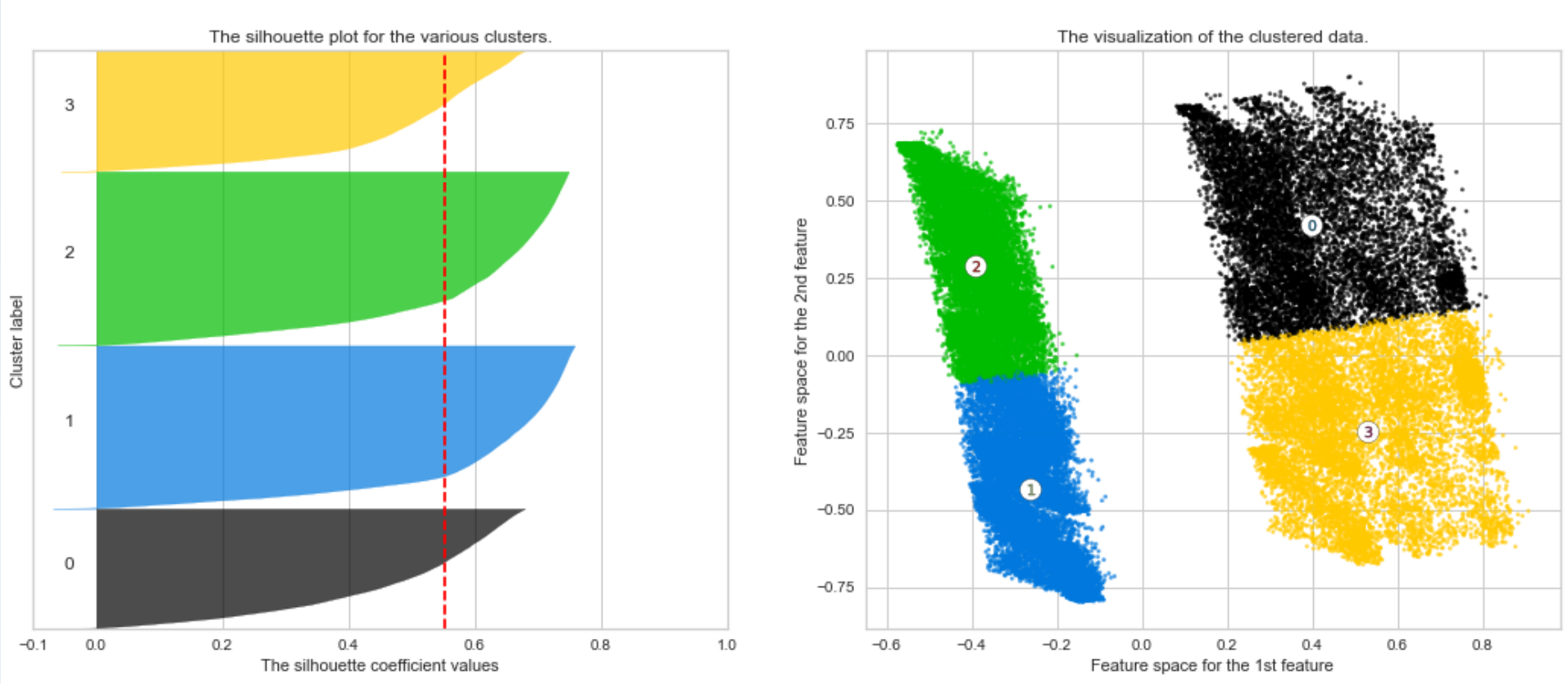
# Plan

- Problématique et les pistes de recherche envisagées
- Données et leur analyse
- Pistes de modélisation
- Segmentation des clients
- Conclusion

# Segmentation des clients

## Segmentation choisie:

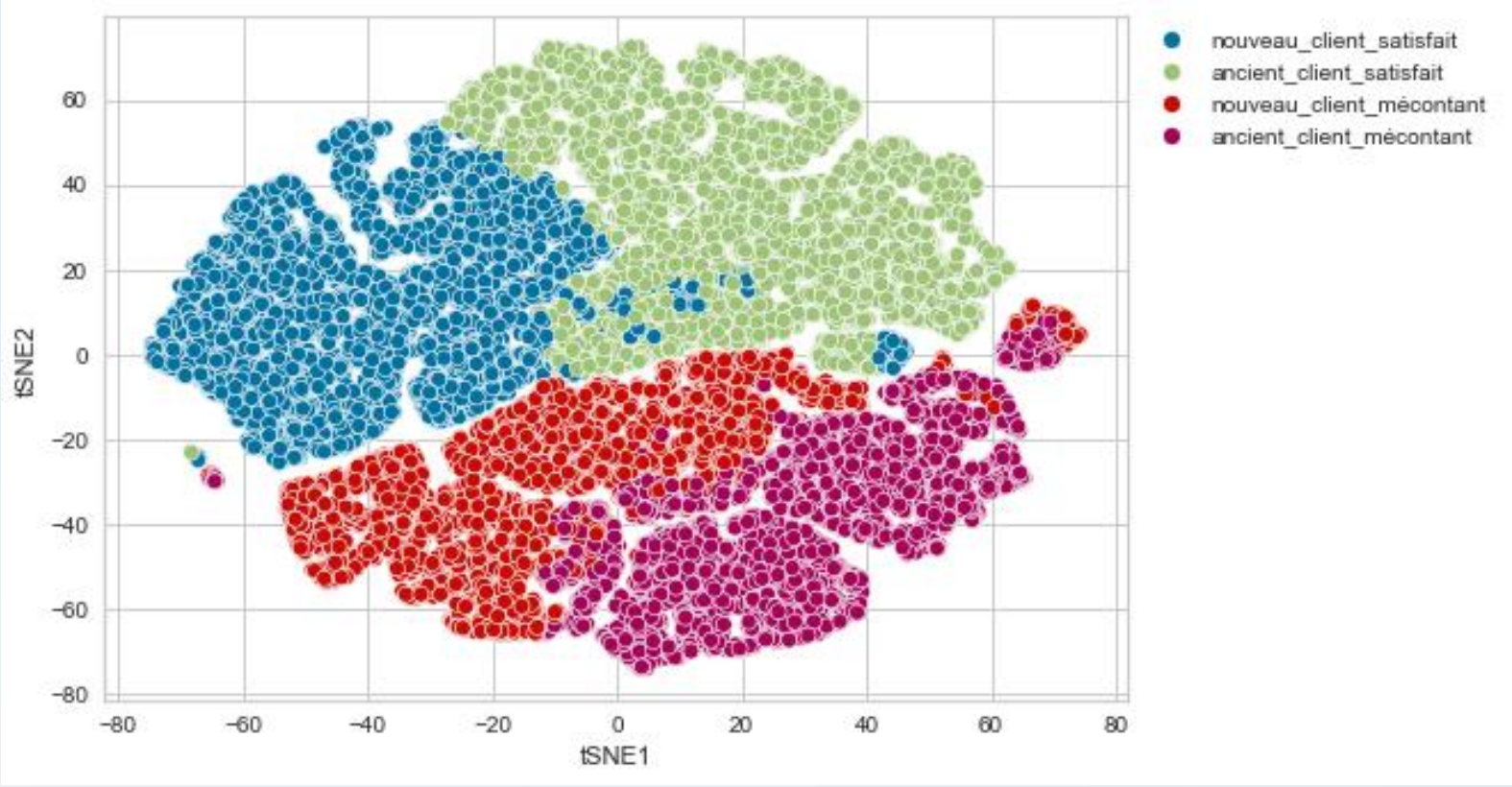
K-means, nombre de composants ACP = 3, nombre de clusters = 4




```
n_clusters = 4 The average silhouette_score is : 0.5532158035021276
```

# Segmentation des clients

- 4 groupes de clients:



# Segmentation des clients

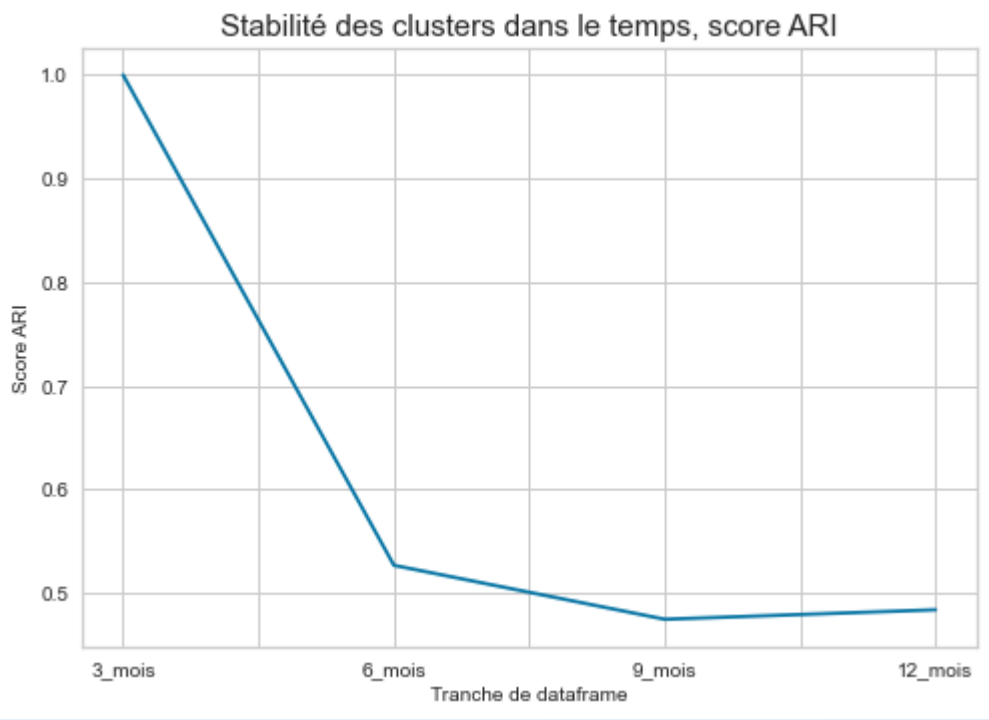
| | Nouveau client satisfait | Ancien client satisfait | Nouveau client mécontent | Ancien client mécontent |
|---|---|---|---|---|
| Récence | nouveau | ancien | nouveau | ancien |
| Fréquence d'achats moyenne | 0 | si revient, revient rapidement | si revient, revient rapidement | si revient, revient sous un délais court |
| Montant, R$ | 208 | 200 | 252 | 249 |
| Nombre moyen d'achats | 1 | 1 | 1 | 1 |
| Livraison | normal | rapide | longue ou retardée | normal |
| Satisfaction | très satisfait | très satisfait | mécontent | mécontent |

?

Piste d'étude: analyse des commentaires client

# Segmentation des clients

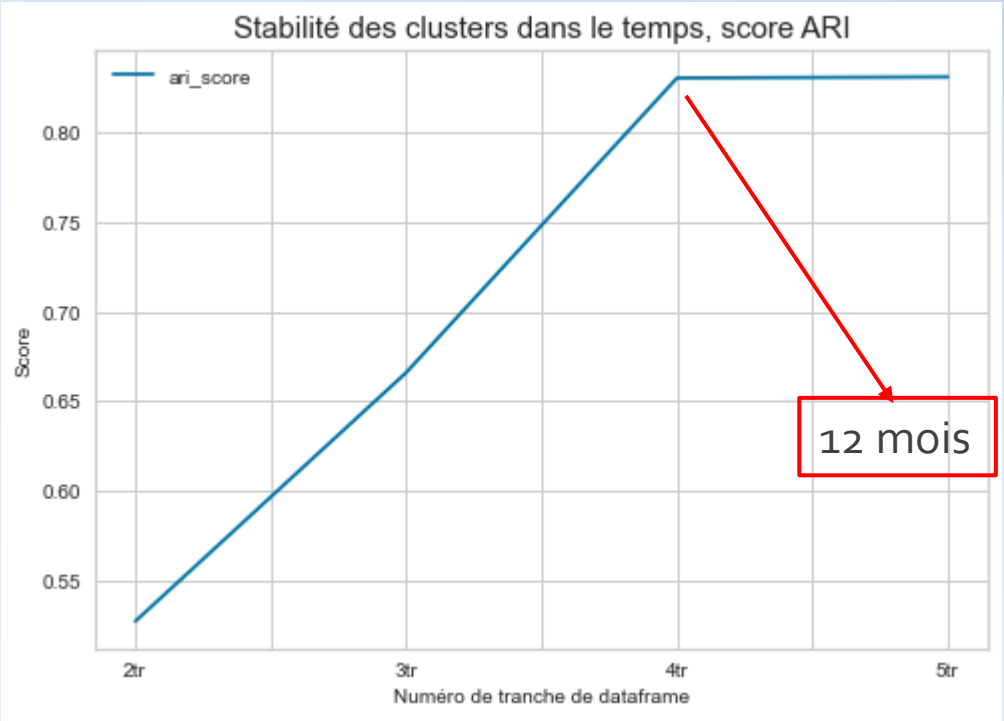
- Fréquence de mise à jour
  - Si la taille initiale de données est de 3 mois




**Mise à jour nécessaire chaque 3 mois**

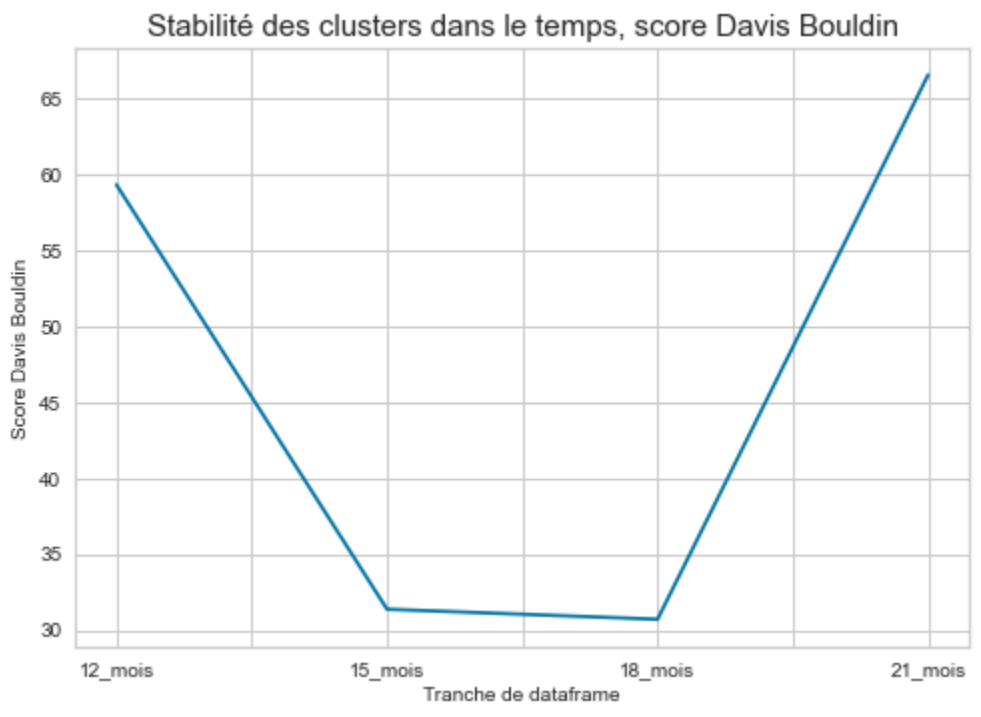
# Segmentation des clients

- Fréquence de mise à jour
  - Quelle taille initiale de données permet la segmentation plus durables dans le temps?



# Segmentation des clients

- Fréquence de mise à jour
  - Si la taille initiale de données est de 12 mois




**Mise à jour nécessaire chaque 6 mois**

# Plan

- Problématique et les pistes de recherche envisagées
- Données et leur analyse
- Pistes de modélisation
- Segmentation des clients
- Conclusion

# Conclusion

- Segmentation non-supervisée des clients d'un site e-commerce
- Données avec beaucoup de dimensions, mais peu de variabilité : la majorité des clients ne font qu'un achat
- Recherche approfondie de composition optimale des briques de pipeline et des hyperparamètres
- Segmentation des clients:
  - 4 clusters, départagés en fonction de la récence et la satisfaction des clients
  - à mettre à jour chaque 6 mois