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


OC Data Scientist
Projet 5

Clément Dubois
15/11/2022

# Contexte générale



**Contexte :**

L'e-commerce brésilien **Olist** souhaite fournir à ses équipes marketing une **segmentation des clients** qu'elles pourront utiliser au quotidien pour leurs campagnes de communication.

**Mission :**

- Réaliser une segmentation afin de comprendre les différents types d'utilisateurs grâce à leur comportement et données personnelles.
- Décrire les différents groupes de clients obtenus
- Proposer un contrat de maintenance basé sur la stabilité des segments au cours du temps.

**Données :**

Export de la base de données Olist de 2016 à 2018

# Sommaire

# Partie I : **Présentation des données et analyse exploratoire**



A) Présentation des données
B) Répartition empirique
C) Analyse et prédiction des valeurs manquantes
D) Transformation de variables
E) Analyse de la relation entre les variables

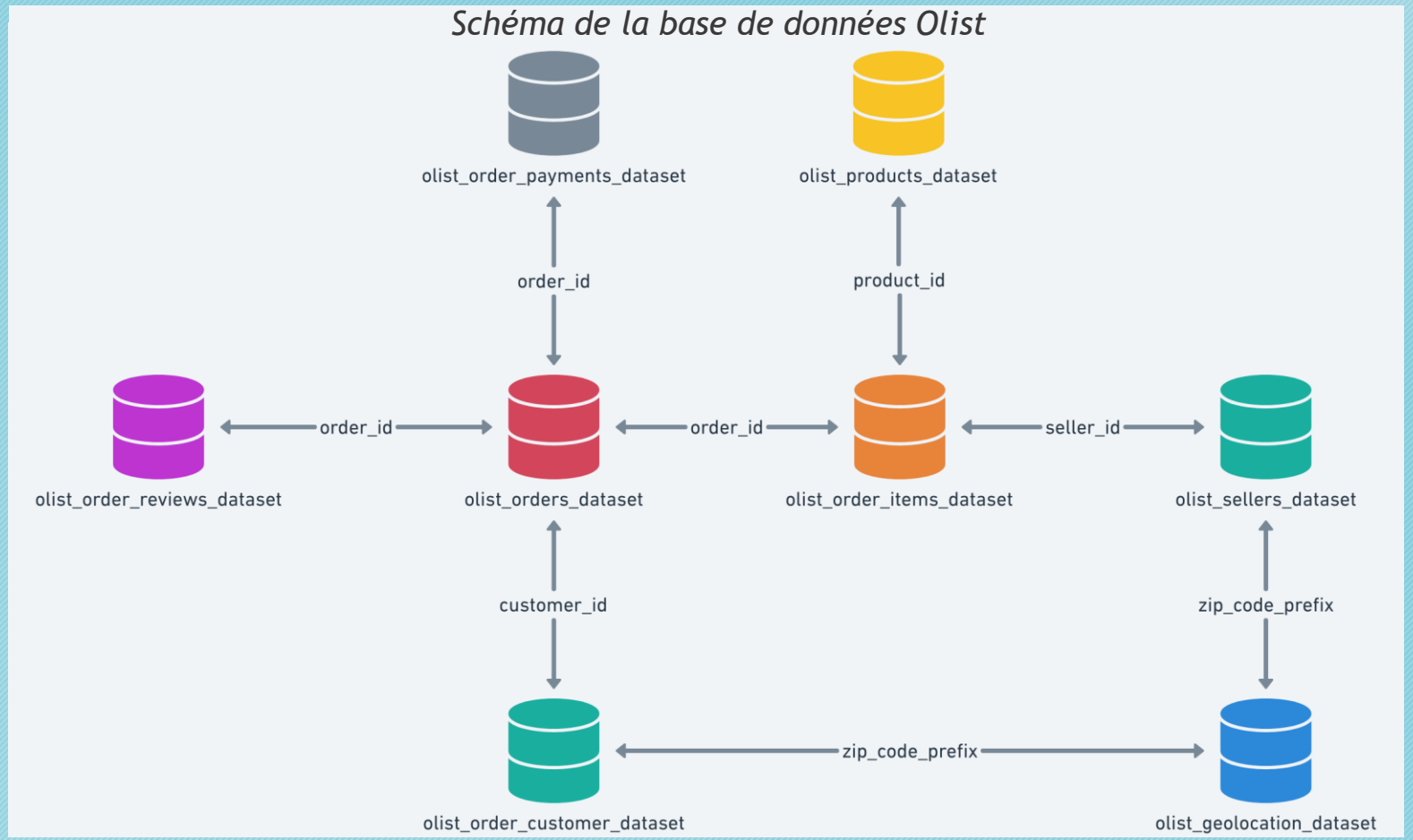
# A) Présentation des données et des variables d'intérêts



- **Nombre de commandes : 98 000**
- **Nombre de clients totaux : 94 990**
- **15 sept 2016 -> 03 sept 2018**

} **3% des clients ont + d'une commande**

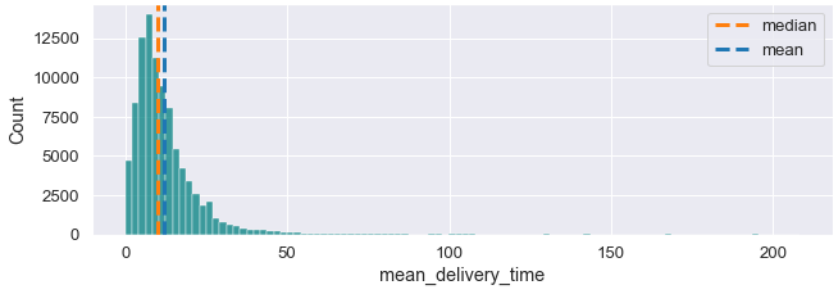
*Schéma de la base de données Olist*



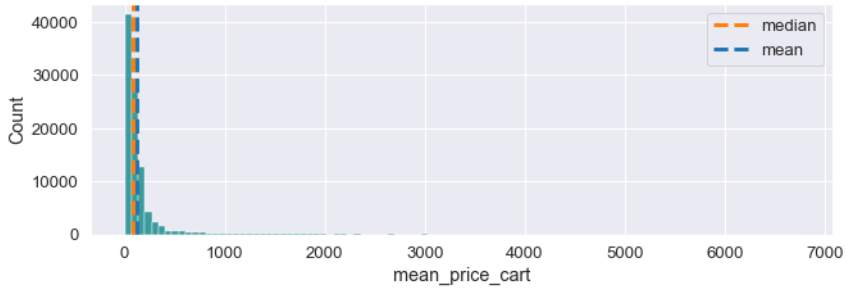

| Variables d'intérêts | Nom des variables (données) |
|---|---|
| **R- Récence = Date de la dernière commande** | recency |
| **F- Fréquence des commande (nb de commande)** | nb_order |
| **M- Montant total dépensé** | total_expanse |
| **Note moyenne** | mean_review_score |
| **Temps de livraison moyen** | mean_delivery_time |
| **Montant du panier moyen** | mean_price_cart |
| **Nombre d'article moyen par commande** | mean_nb_item |
| **Ancienneté** | seniority |

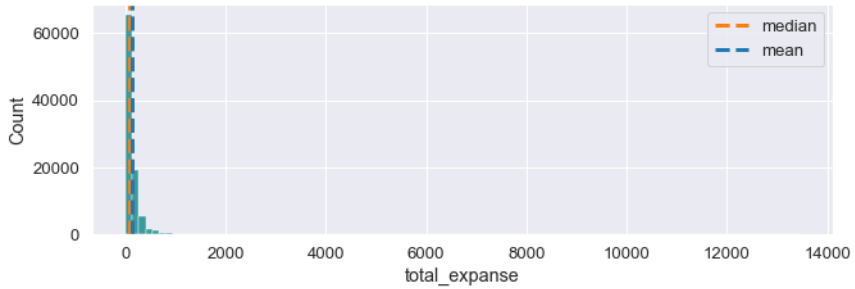
# B) Répartition empirique des variables d'intérêts

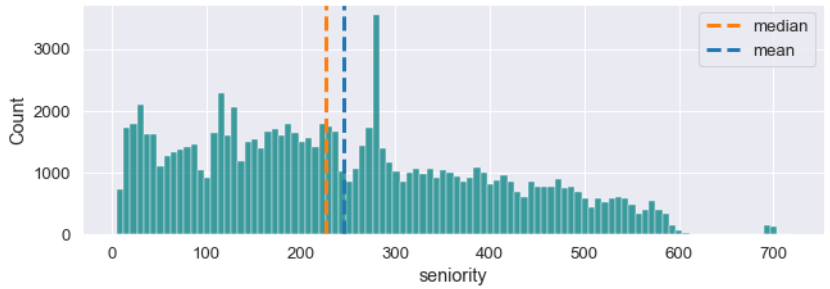


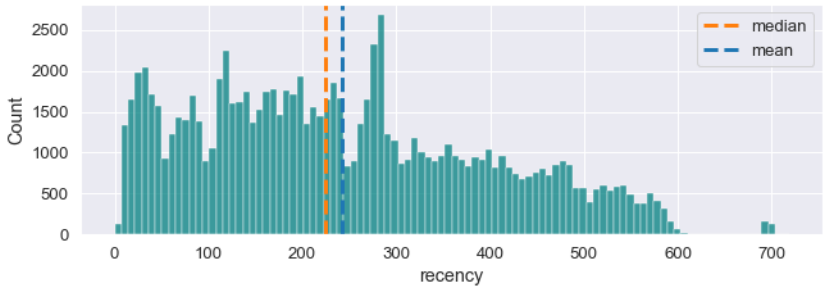


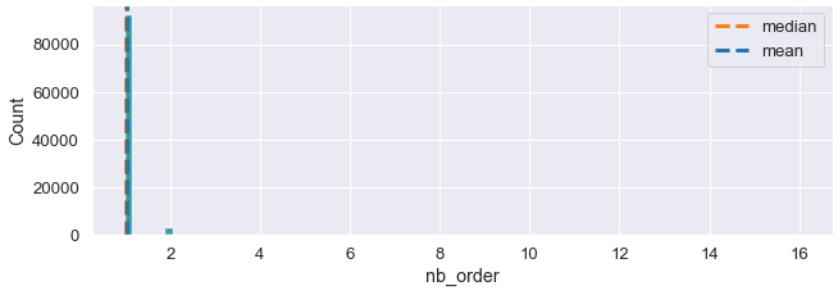


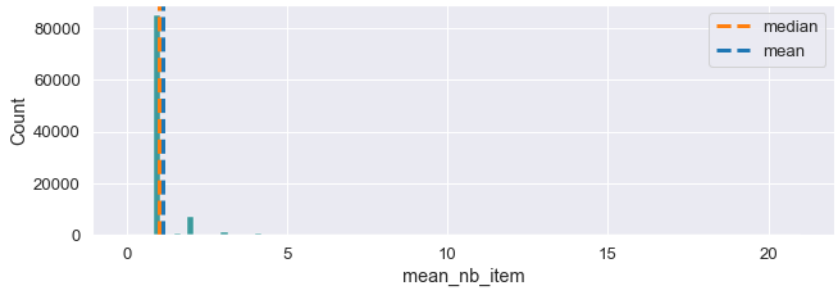


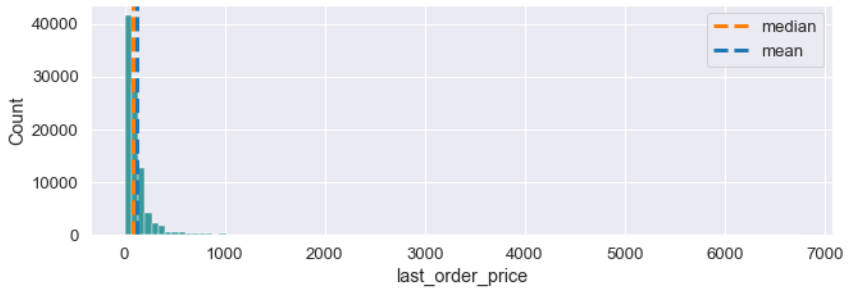



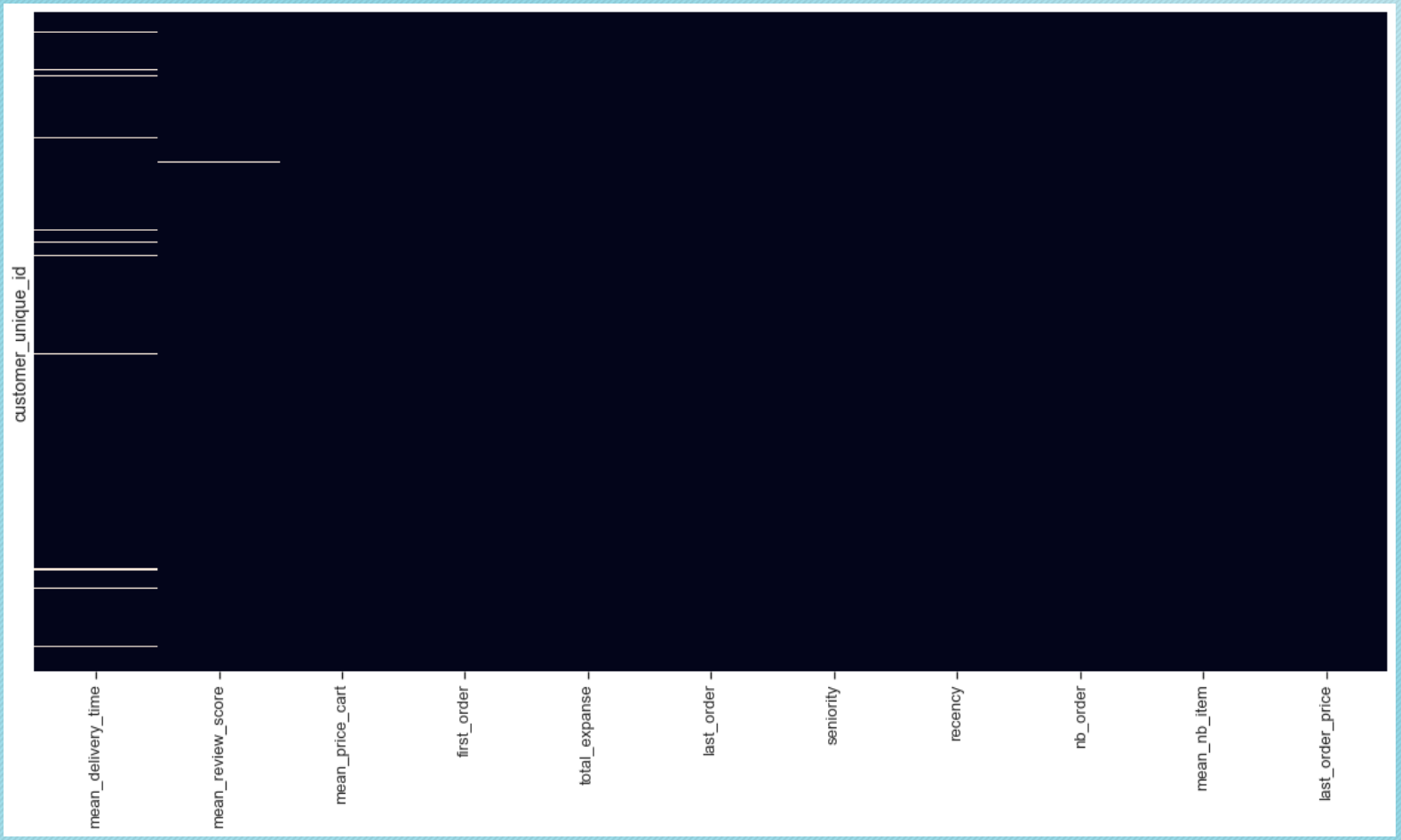



# C) Gestion des valeurs manquantes



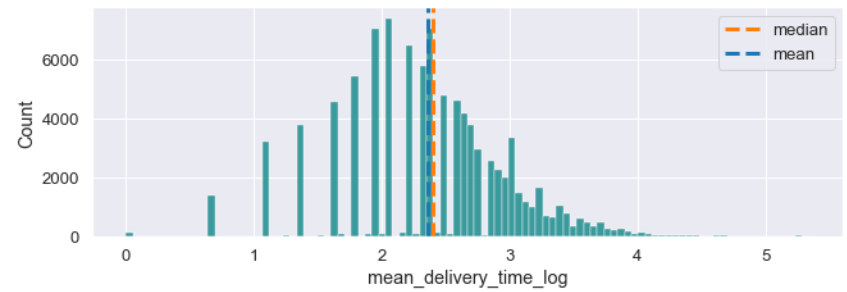


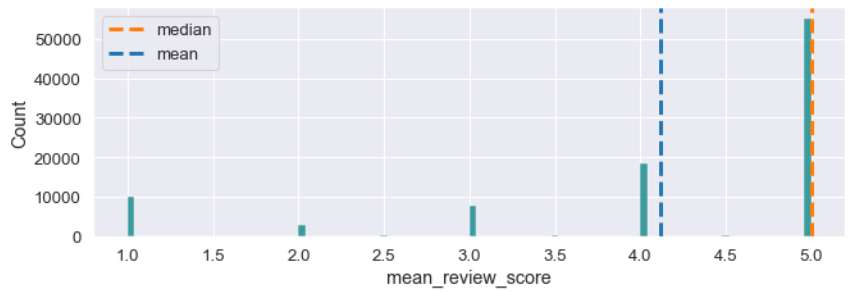

| Variables | Nb de valeurs manquantes | % de valeurs manquantes |
|---|---|---|
| mean_delivery_time | 1653 | 1,74% |
| mean_review_score | 685 | 0,70% |
| mean_price_cart | 7 | < 0,1% |
| first_order | 17 | < 0,1% |
| total_expanse | 0 | 0 |
| last_order | 17 | < 0,1% |
| seniority | 17 | < 0,1% |
| recency | 17 | < 0,1% |
| nb_order | 0 | 0 |
| mean_nb_item | 0 | 0 |
| last_order_price | 7 | < 0,1% |

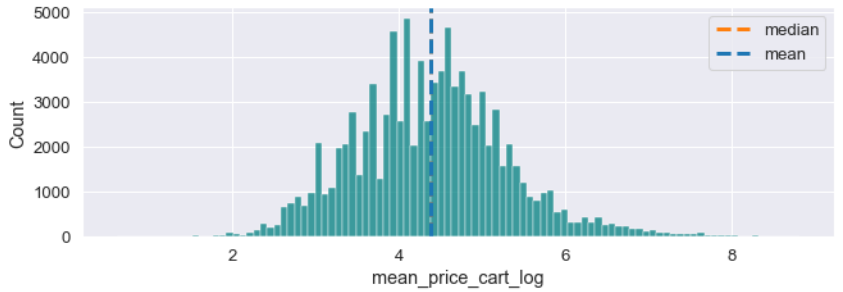
> Remplacement des valeurs manquantes par la médiane (SimpleImputer)

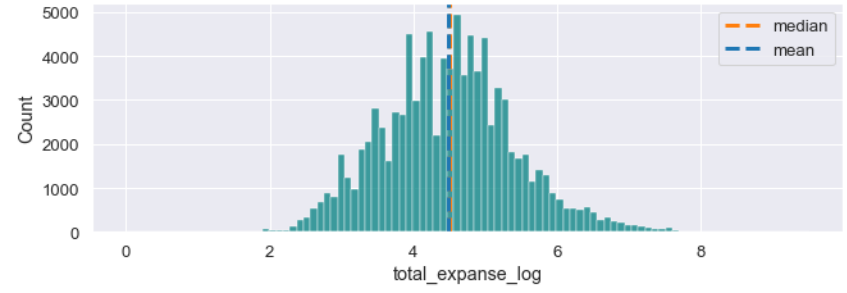
# D) Transformation des variables

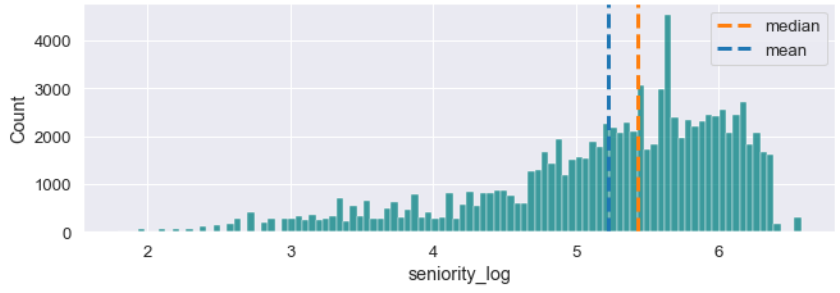


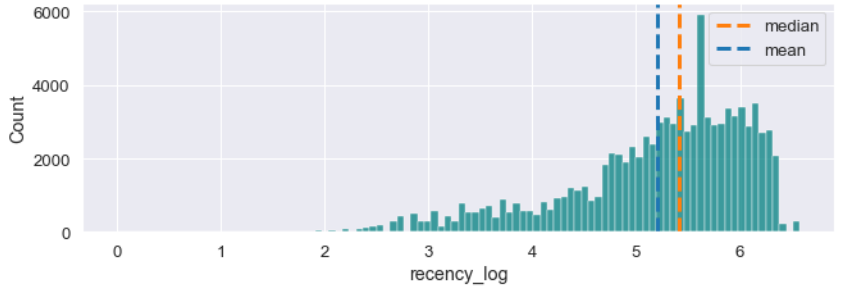


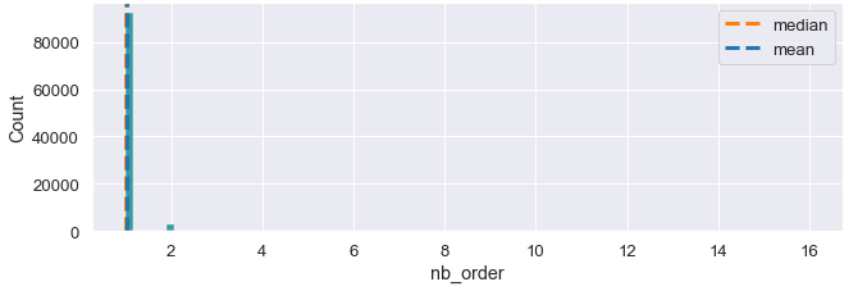



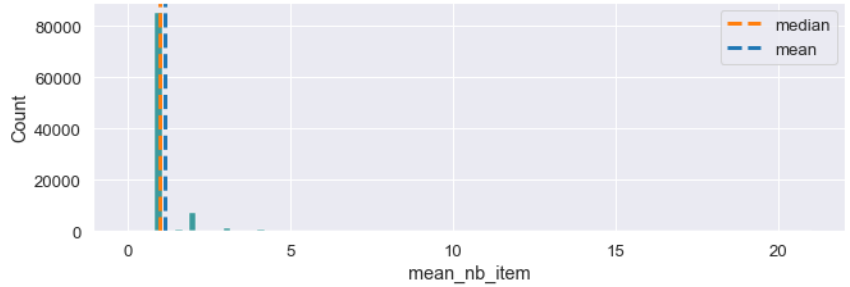



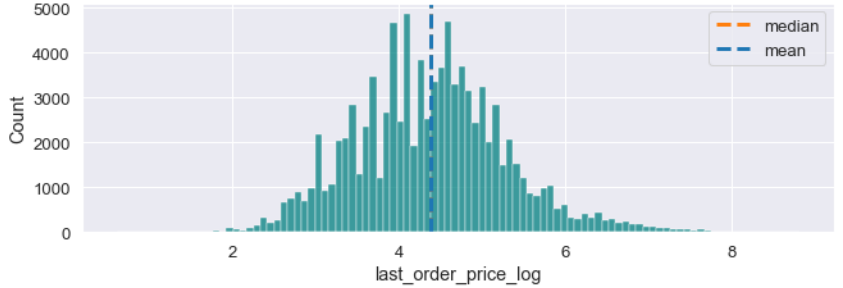



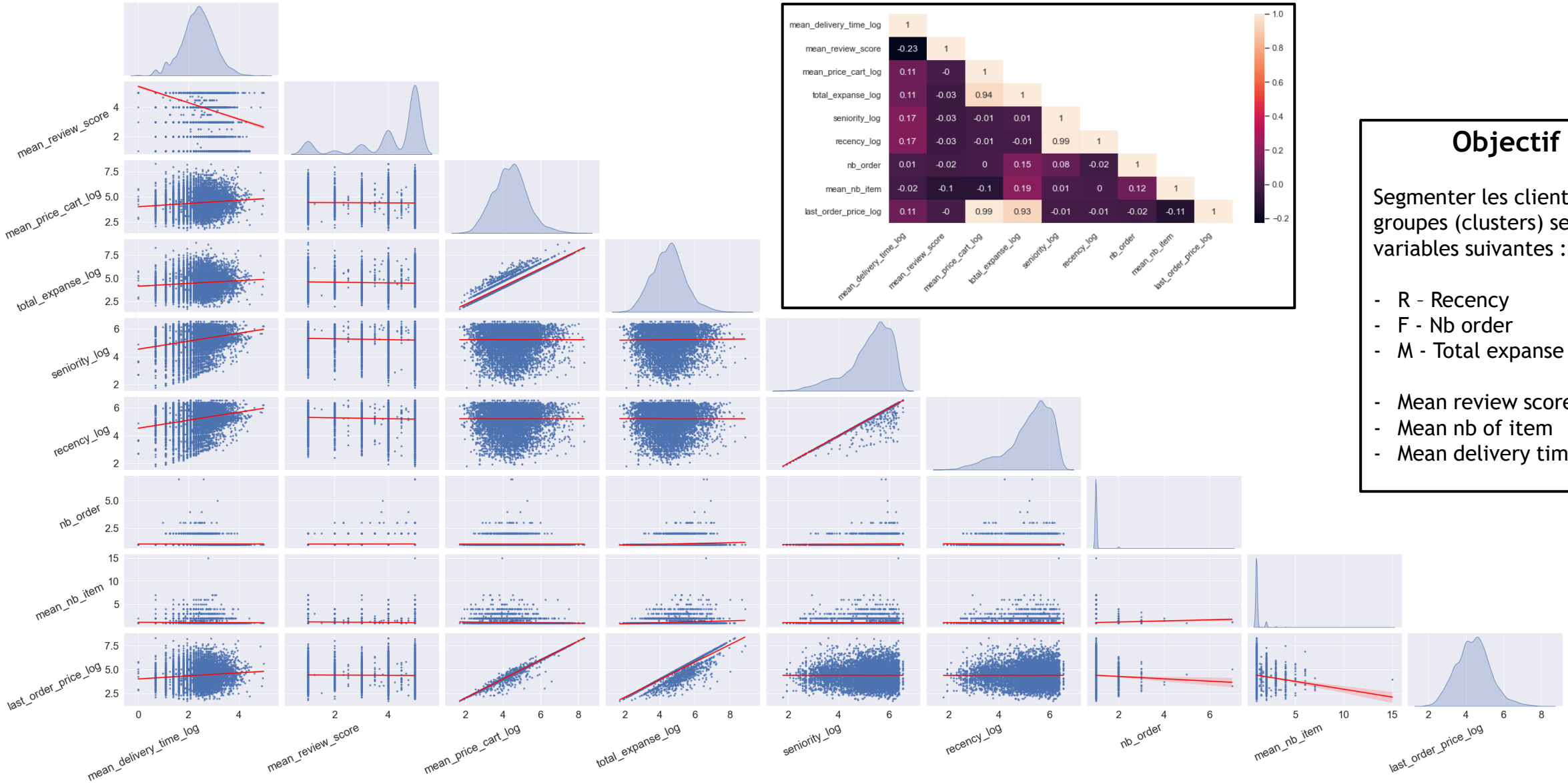




- Certaines distributions s'approchent d'une loi normal après transformation (log+1)
- Les données seront par la suite Standardisées avant le Clustering

# E) Brève analyse de la relation entre les variables









## Objectif

Segmenter les clients en groupes (clusters) selon les variables suivantes :

- R – Recency
- F - Nb order
- M - Total expanse

- Mean review score
- Mean nb of item
- Mean delivery time

# Partie II : **Comparaison des méthodes de Clustering**

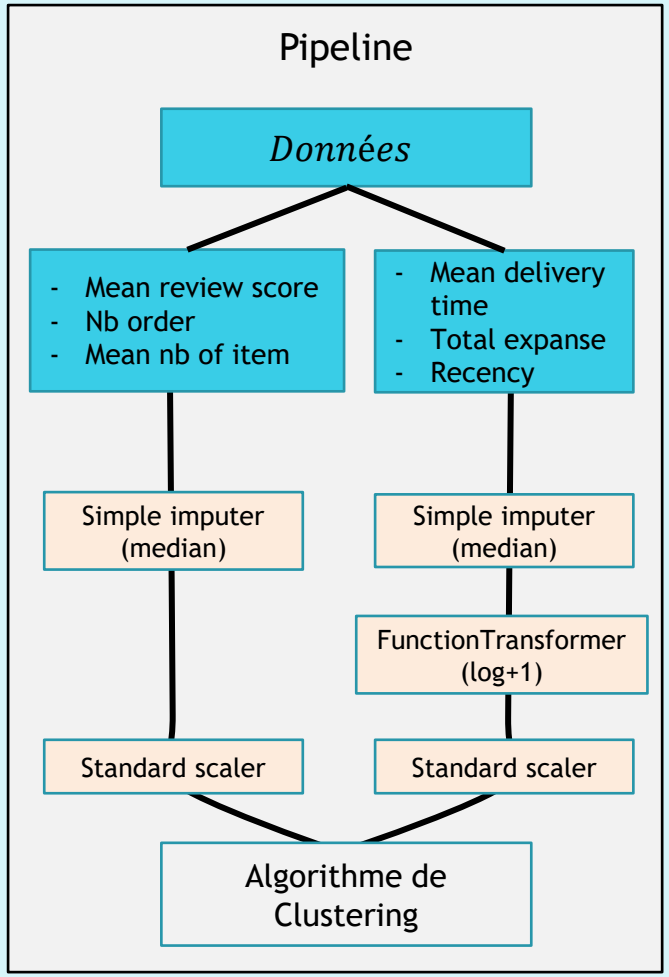


A) Démarche de modélisation
B) Algorithme des kmeans
C) Classification Ascendante hiérarchique
D) DBSCAN
E) Analyse de stabilité des clusters

# A) Présentation de la démarche de modélisation



## Etapes de la modélisation :




A) Détermination des meilleurs hyper-paramètres

B) Entrainement des modèles et prédictions des clusters

C) Comparaison de la **stabilité** des clusters entre les différents algorithmes **au cours du temps.** (26 dernières semaines)

## Echantillonnage des données

Propriétés :

- Chaque client doit être présent du début à la fin de l'expérience de modélisation (premier achat > 26 semaines)

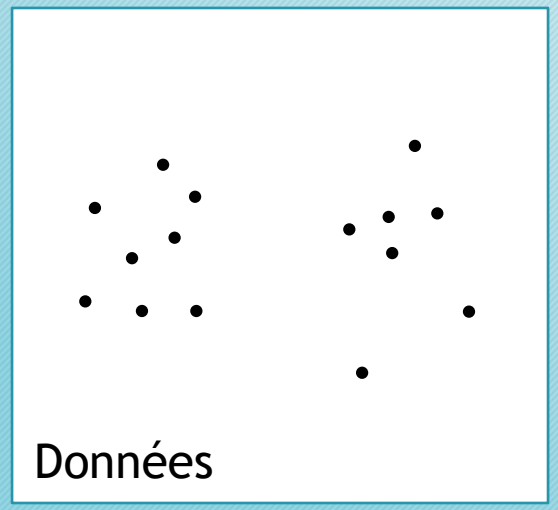
- Le jeu de donnée doit évoluer (nb commande final > 1)

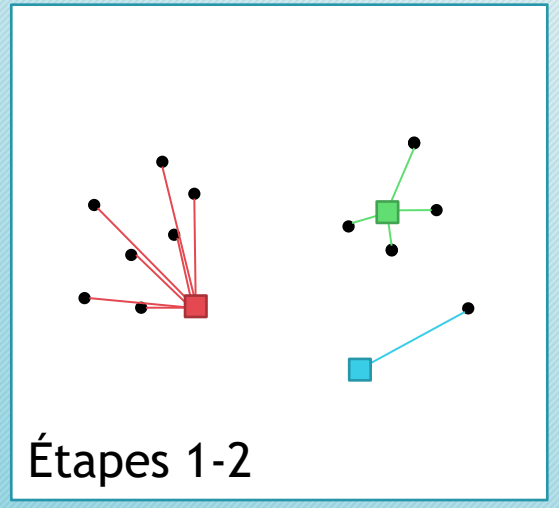
# B) Modèle 1 : Algorithme des kmeans

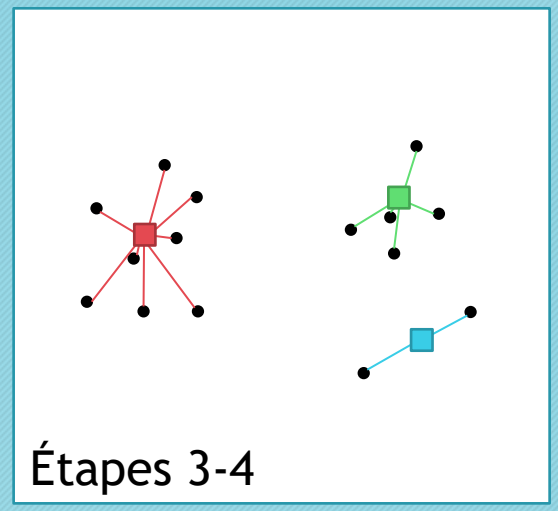


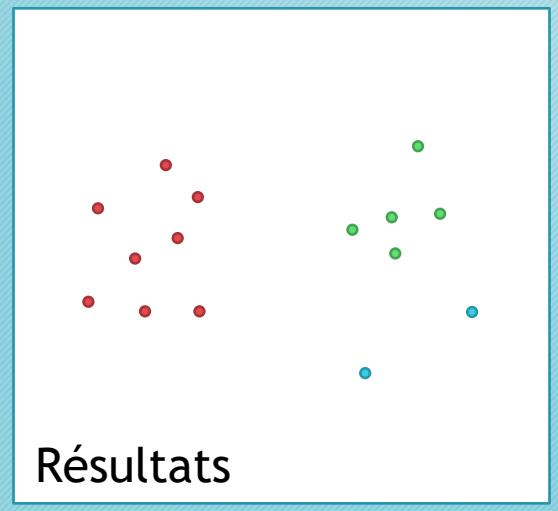
**Principe** : Déterminer k groupes homogènes et compacts en minimisant l'inertie intra-classe.

1) Placement de *k* centroïdes au hasard (ou observation)
2) Attribution d'un groupe aux observations les plus proches des centroïdes
3) Calculer le centre de gravité du groupe
4) Y Déplacer le centroïde > Etape 2


Données


Étapes 1-2


Étapes 3-4

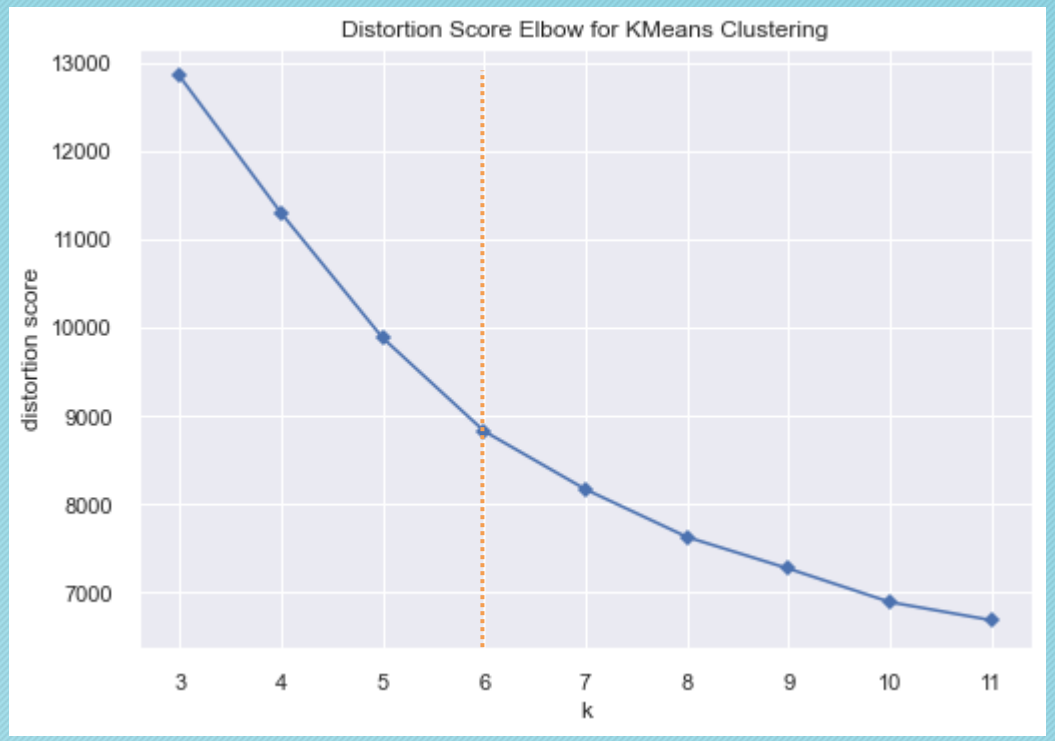

Résultats

# B) Modèle 1 : Algorithme des kmeans



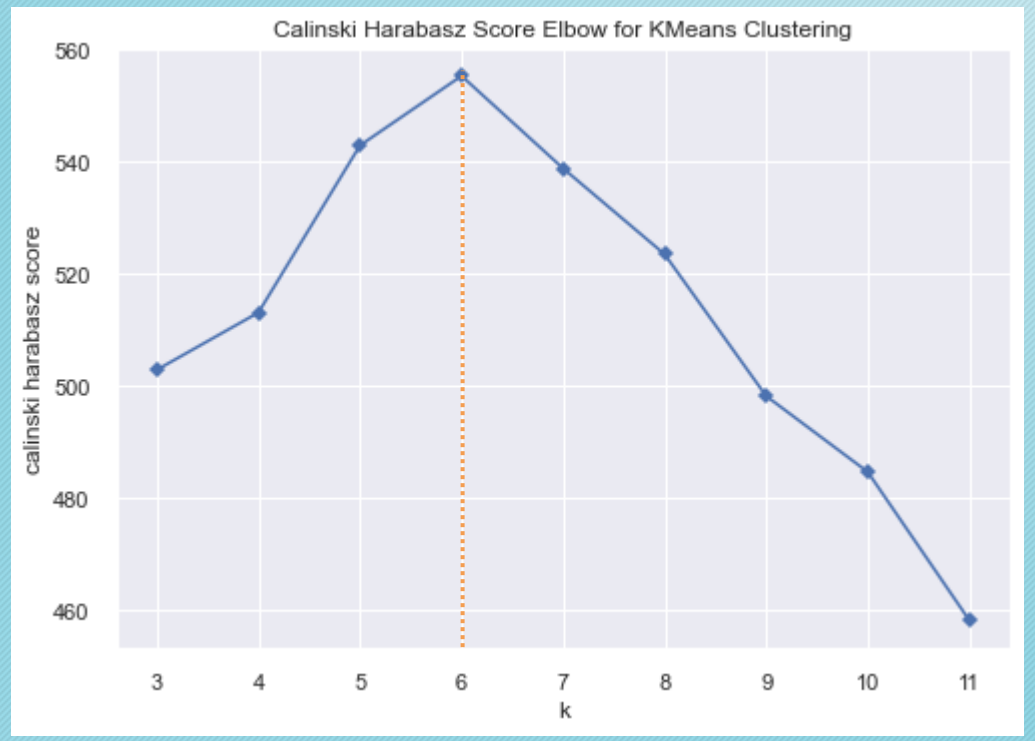
## Hyper paramètre : k > Le nombre de clusters

**Méthode du coude :** Repérer le nombre de clusters où la variance intra-classe ne diminue plus significativement




**Score de Calinski Harabasz** ($\frac{variance\ inter-groupe}{variance\ intra-groupe}$) :

- Score entre 0 et +∞
- Plus le score est haut, plus les clusters sont denses et bien séparés.

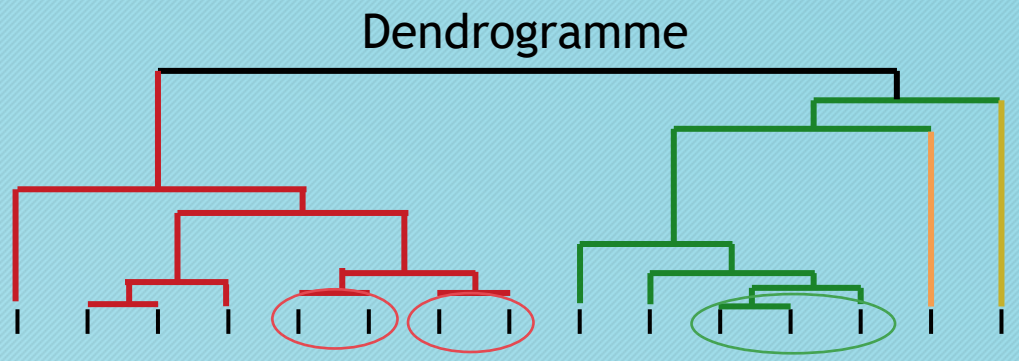


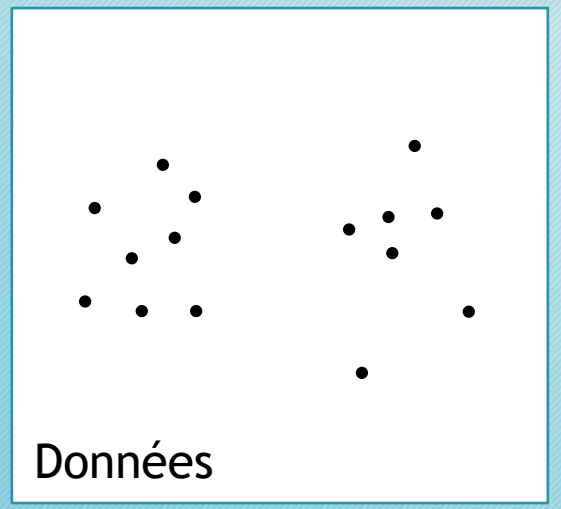
# C) Modèle 2 : Classification Ascendante Hiérarchique

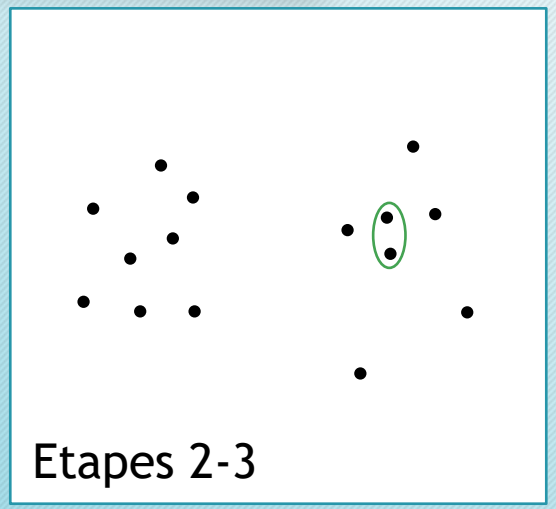


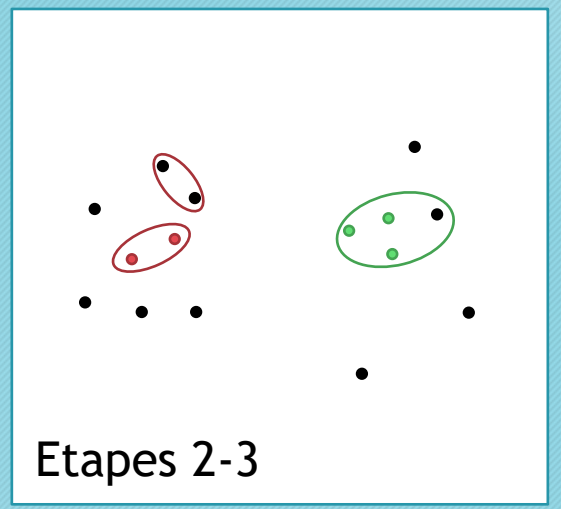
**Principe** : Rassembler les observations deux à deux selon un critère de distance

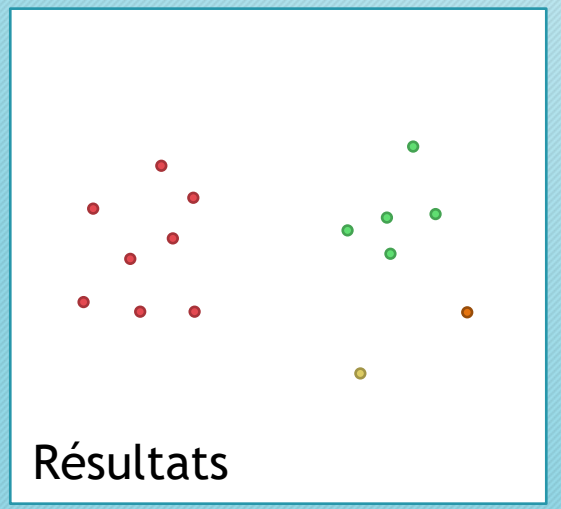
1) Attribution d'un groupe à chaque observation
2) Calcule de la distance entre chaque groupe.
3) Agrégation des deux groupes les plus proches (méthode de Ward) > Etape 2.





Données

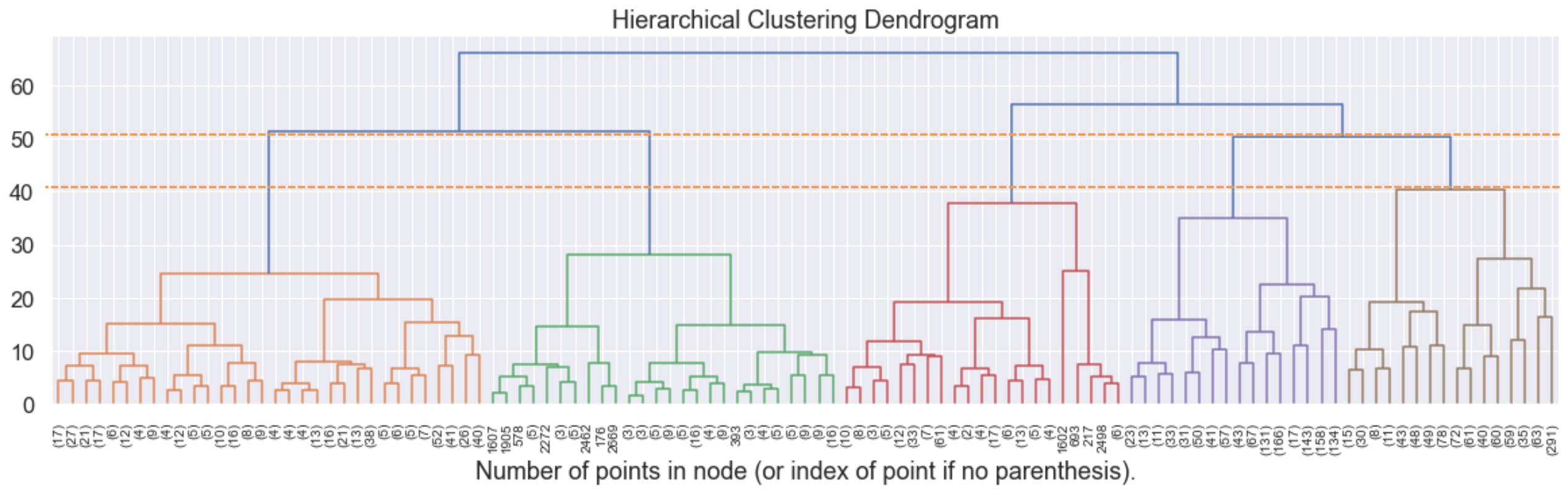

Etapes 2-3


Etapes 2-3


Résultats

# C) Modèle 2 : Classification Ascendante Hiérarchique





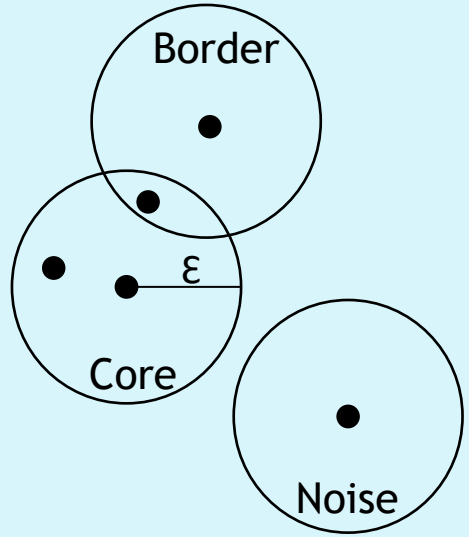

- **Le modèle de Classification Ascendante Hierarchique distingue 5 clusters dans nos données**

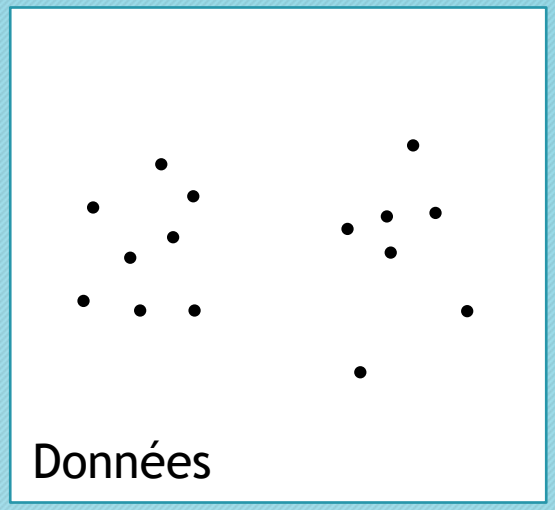
# D) Modèle 3 : Density Based Spatial Clustering of Applications with Noise (DBSCAN)

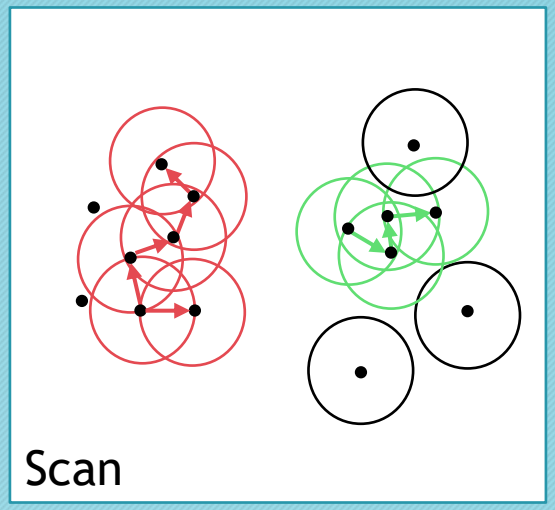


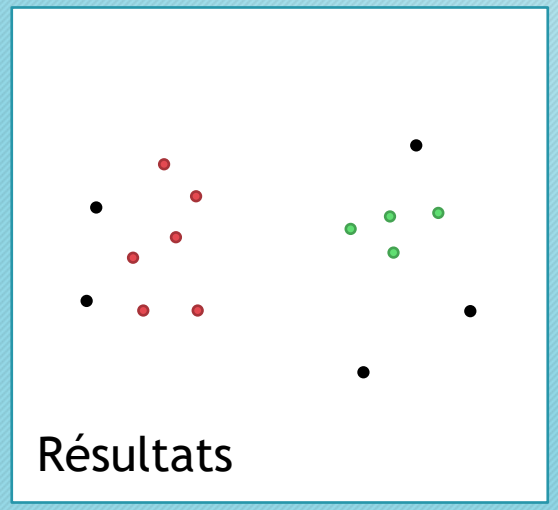
**Principe** : Déterminer des espaces à haute densité de points (clusters) séparés par des espaces à faible densité.

1) Calcul du nombre de voisins dans le rayon de voisinage (Epsilon) d'un premier point.
2) - Si nombre de voisins dans le rayon > seuil (minPts) : le point est un 'core' point.
   - Si le point est dans le voisinage d'un core point : point de « border »
   - Sinon, le point est considéré comme un 'noise' point
3) Même analyse sur le point suivant.



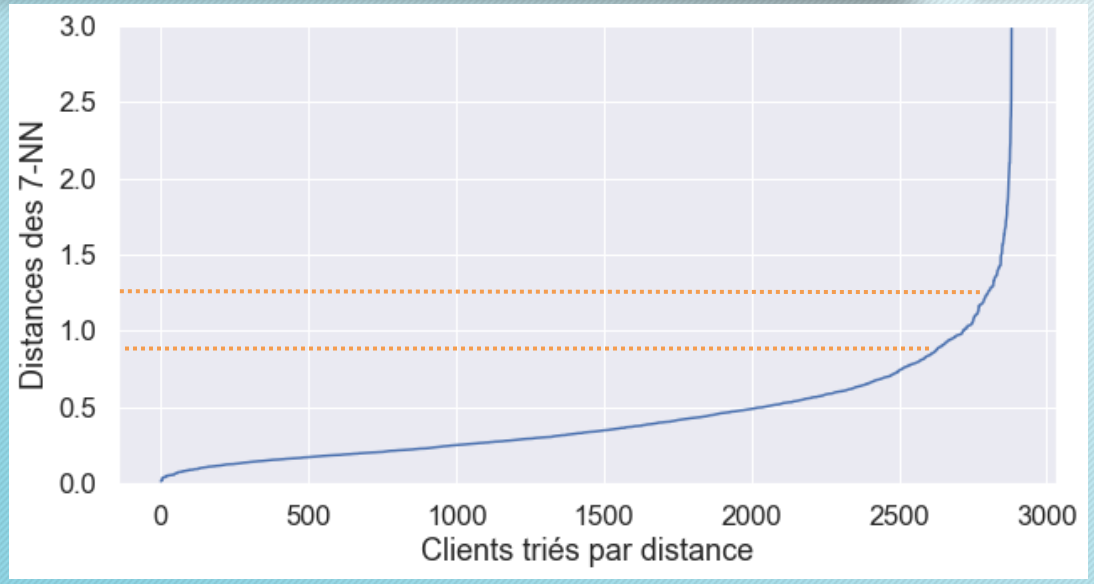



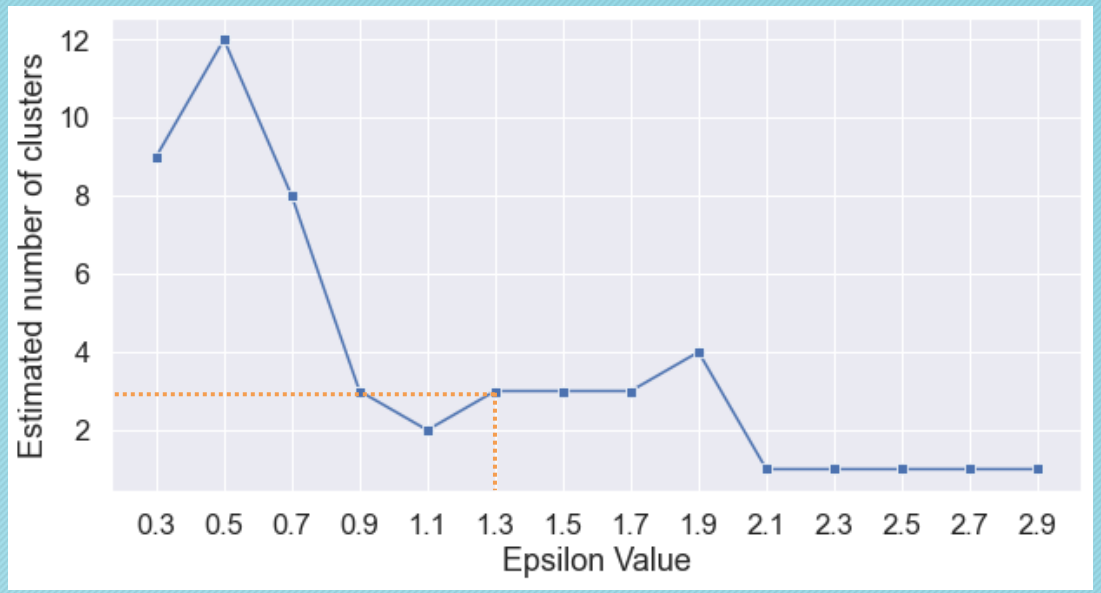
# D) Modèle 3 : Density Based Spatial Clustering of Applications with Noise (DBSCAN)

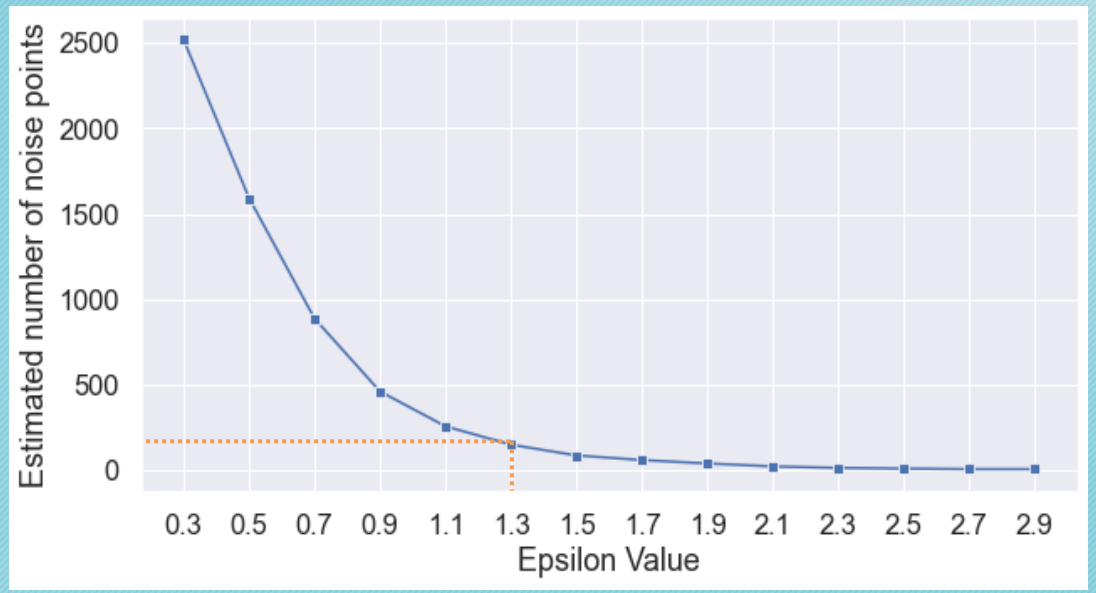


**Calibrage des hypers paramètres :**

- **minPts =** Nb de dimensions + 1
- **epsilon :** Distances des voisins les plus proches / le nombre de cluster souhaitées / le nombre d'outliers (noise points)



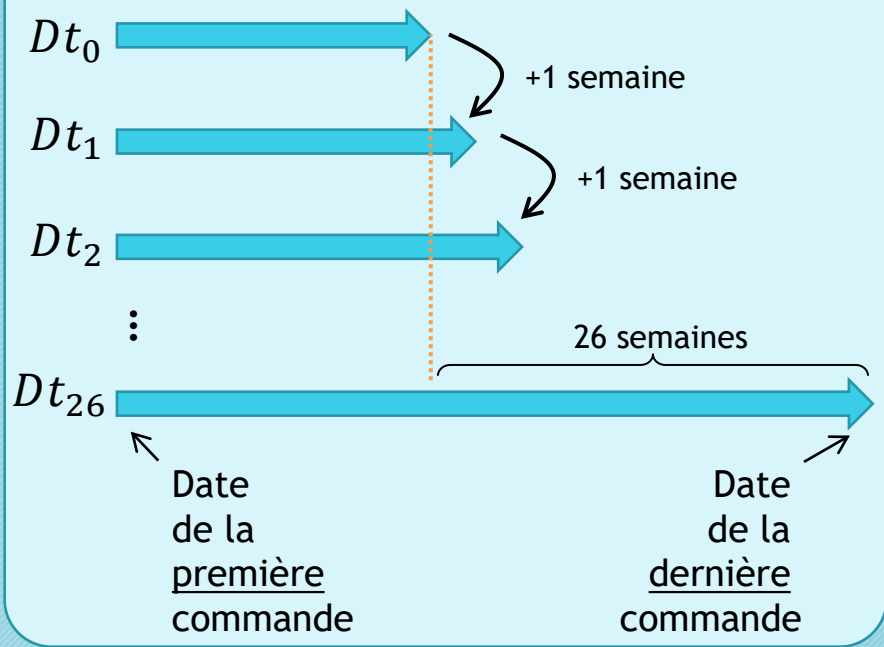
# E) Analyse de stabilité des clusters : Méthode



## 1) Echantillonnage des Données

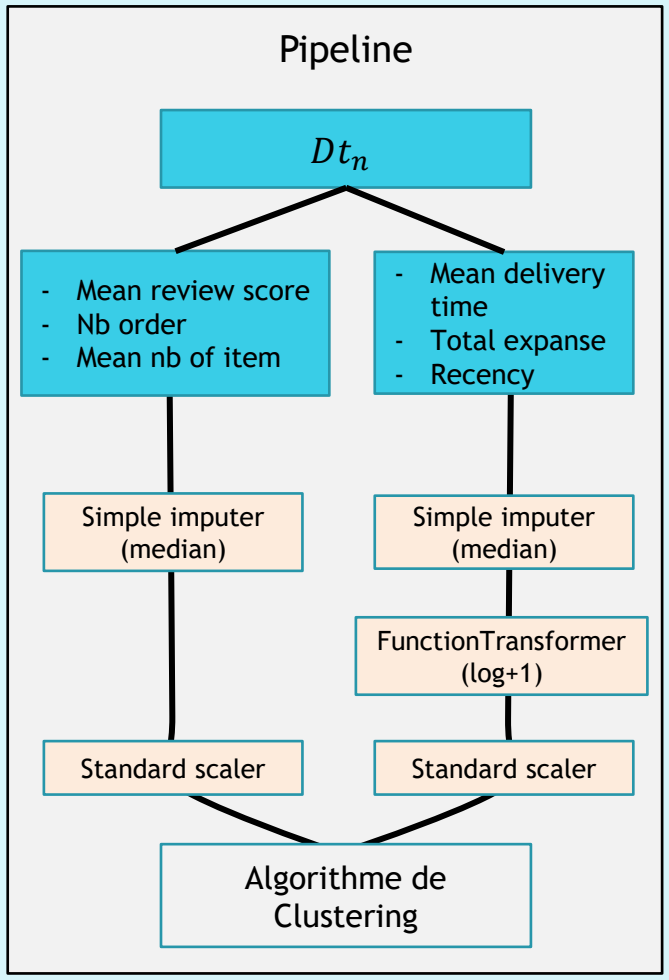
Propriétés :

- Chaque client doit être présent du début à la fin de l'expérience (premier achat > 26 semaines)
- Le jeu de donnée doit évoluer (nb commande final > 1)




## 2) Modélisation

A) **Entraînement (Dtn)**
B) **Prédiction des clusters (Dtn)**




## 3) Stabilité

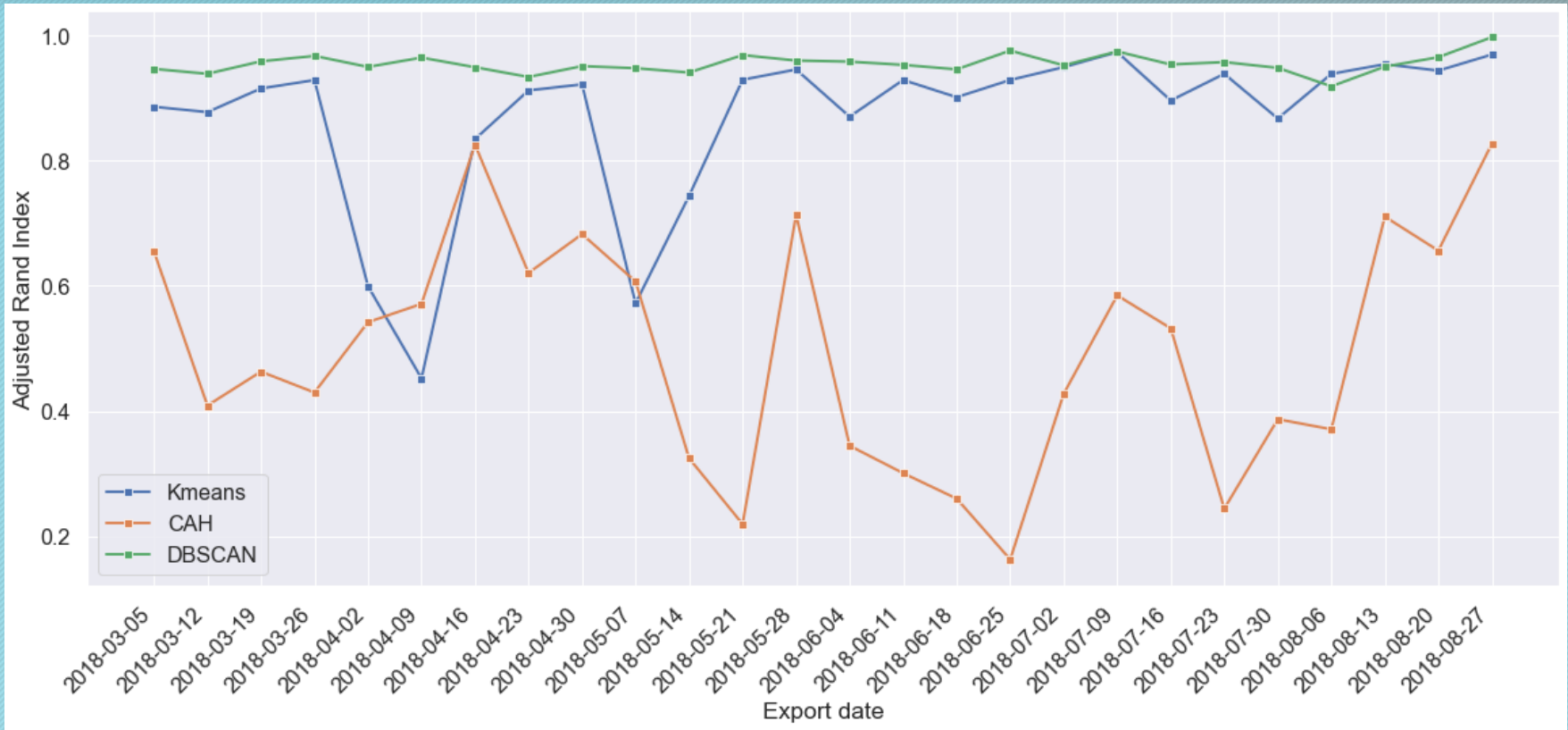
**Comparaison de la stabilité des clusters entre tn et tn+1 jusqu'à t26**

Adjusted Rand Index (ARI) :
Mesure de similarité entre deux segmentations par comparaison des assignations de groupe au niveau individuel.

ARI = 0 > les groupes sont aléatoires
ARI = 1 > Les deux groupes sont identiques

# E) Analyse de stabilité des clusters : Résultats






| Algorithme | ARI moyen |
|---|---|
| Kmeans | 0.87 |
| CAH | 0.5 |
| DBSCAN | 0.96 |

- **Objectif : Segmenter les clients en groupes compacts sans outlier (noise) :**
  DBSCAN > + stable et + approprié (/!\ impossible de prédire de nouvelles données sans ré-entrainer le modèle)

- **Si l'objectif est de classifier tous les clients :**
  Kmeans est le plus approprié (+ Stable et + rapide)

# Partie III : Clustering par kmeans des données totales



A) Choix du nombre de clusters

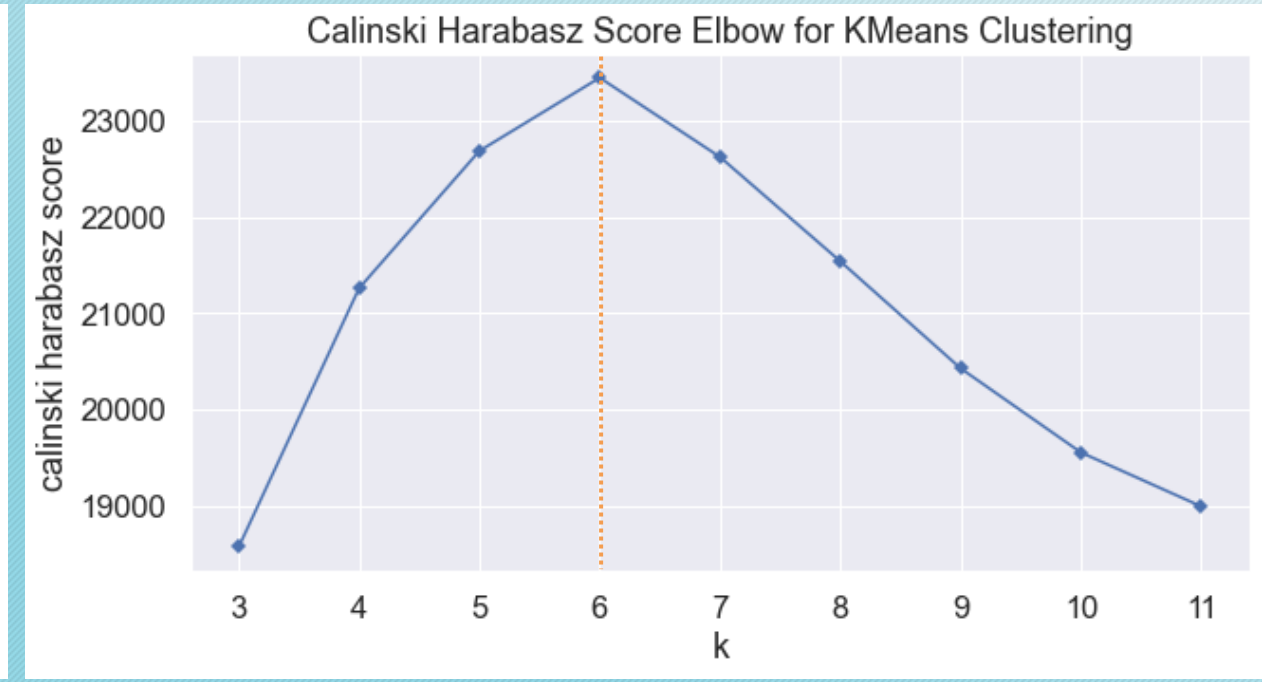
B) Analyse de stabilité

C) Répartition des effectifs

D) Descriptions des clusters

# A) Choix du nombre de clusters

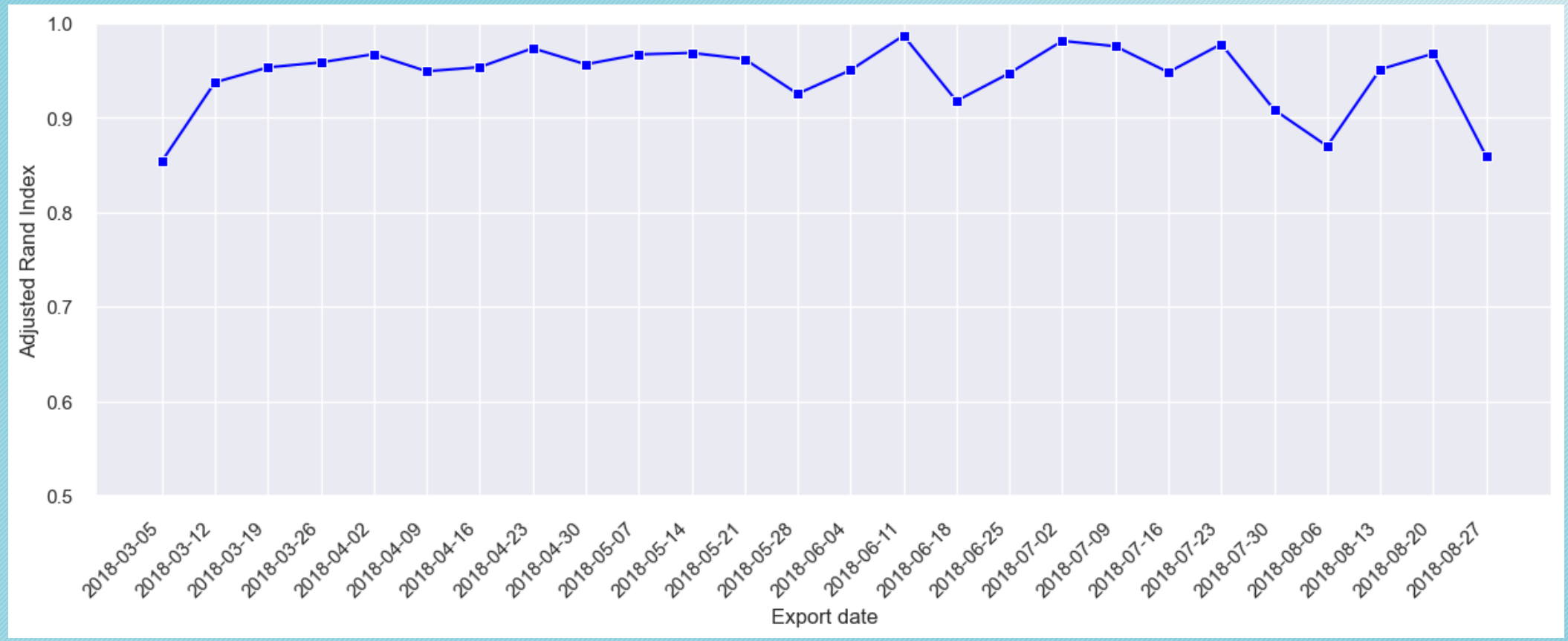



- **6 Clusters semblent être la valeur optimale sur les données totales**

# B) Stabilité des clusters des données totales

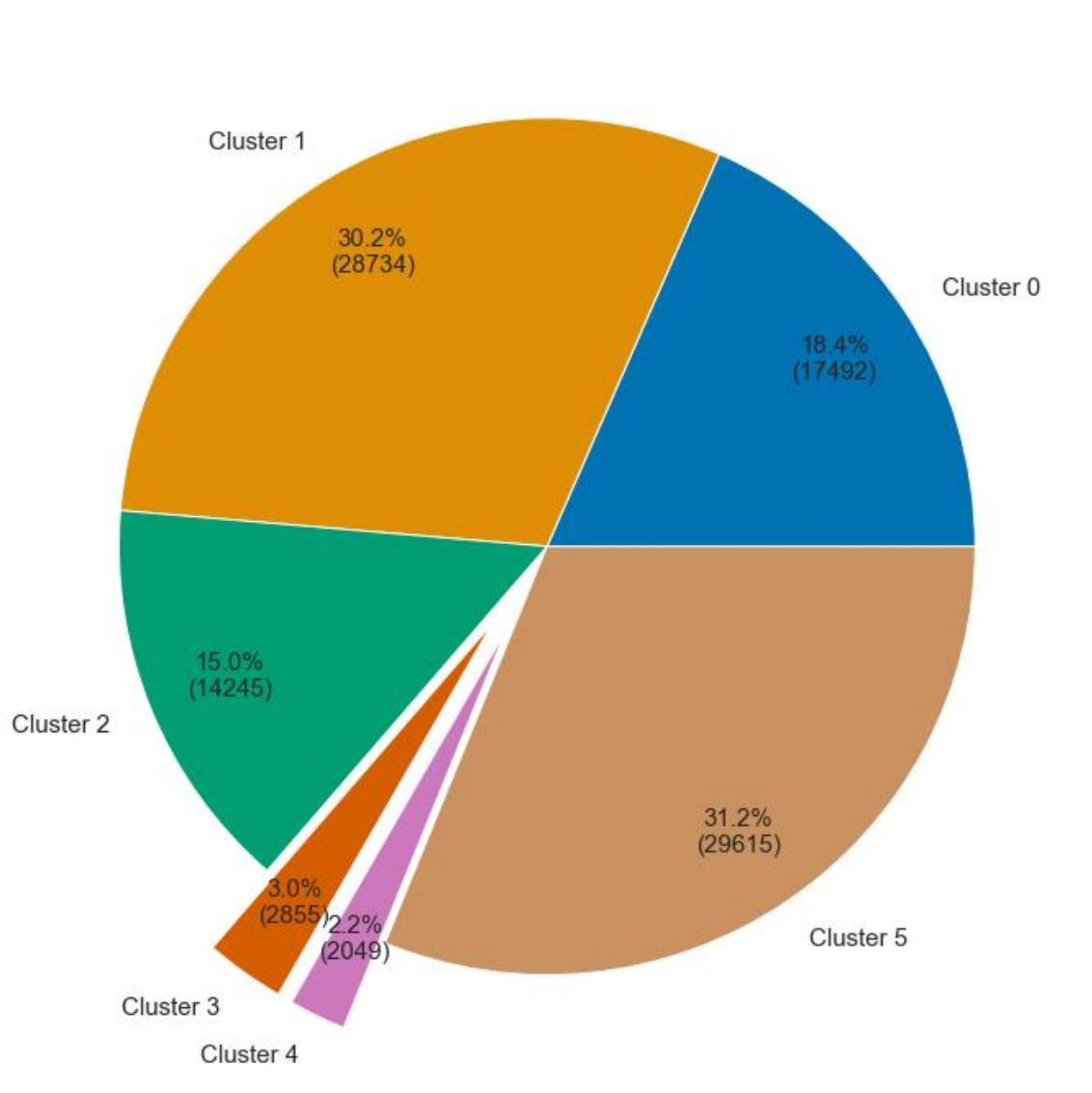





- **Les clusters issus du kmeans sont stables dans le temps. (ARI moyen : 0,95)**

# C) Description des clusters : Répartition des effectifs

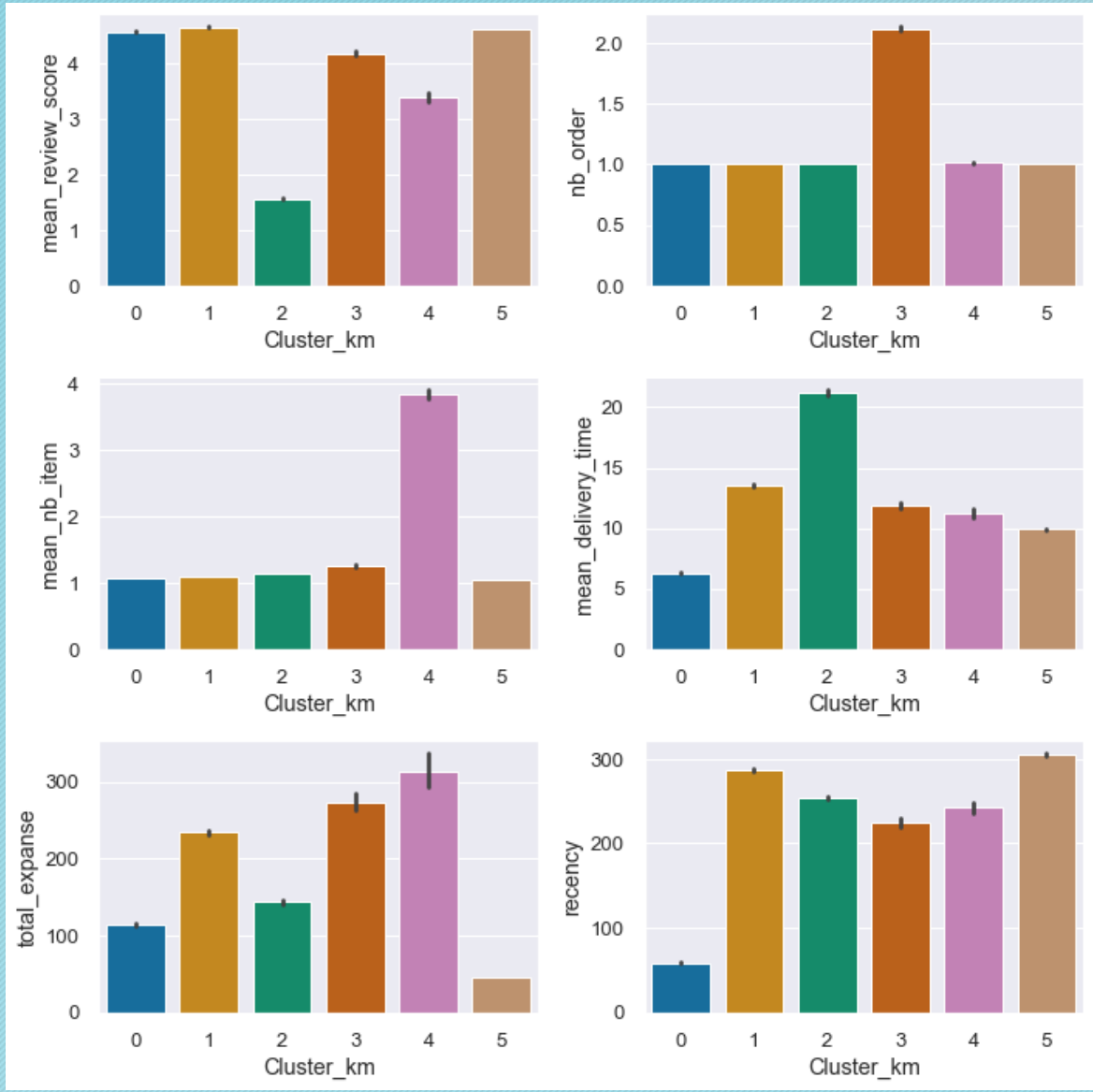





- Cluster 1 et 5 : Représentent 60% des clients
- Cluster 3 et 4 : Représentent 5% des clients

# D) Description des clusters : Valeur moyenne des différents indicateurs par cluster






Cluster 0 : Achats récents et livraisons rapides

Cluster 1 : Achats anciens, dépenses moyennes, clients satisfaits

Cluster 2 : Clients mécontents, probablement lié au temps de livraison élevé

**Cluster 3 : Clients fidèles (plus d'une commande)**

**Cluster 4 : Achats de nombreux produits, dépense totale élevée**

Cluster 5 : Achats très ancien, très faible dépense.

# Partie IV : Contrat de maintenance



A) Méthode de détermination de la période de maintenance

B) Résultats

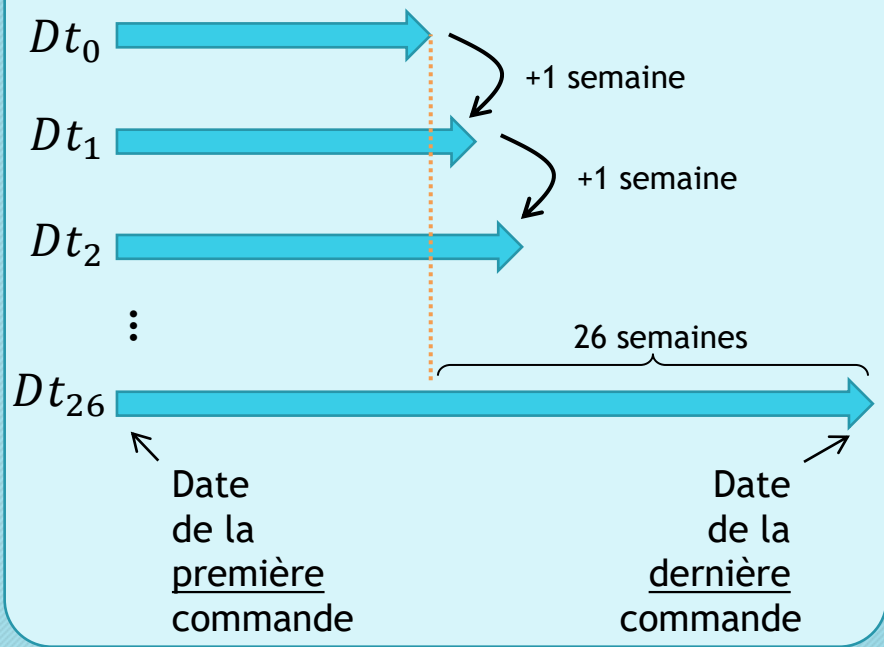
# A) Détermination de la periode de maintenance : Méthode



## 1) Echantillonnage des Données

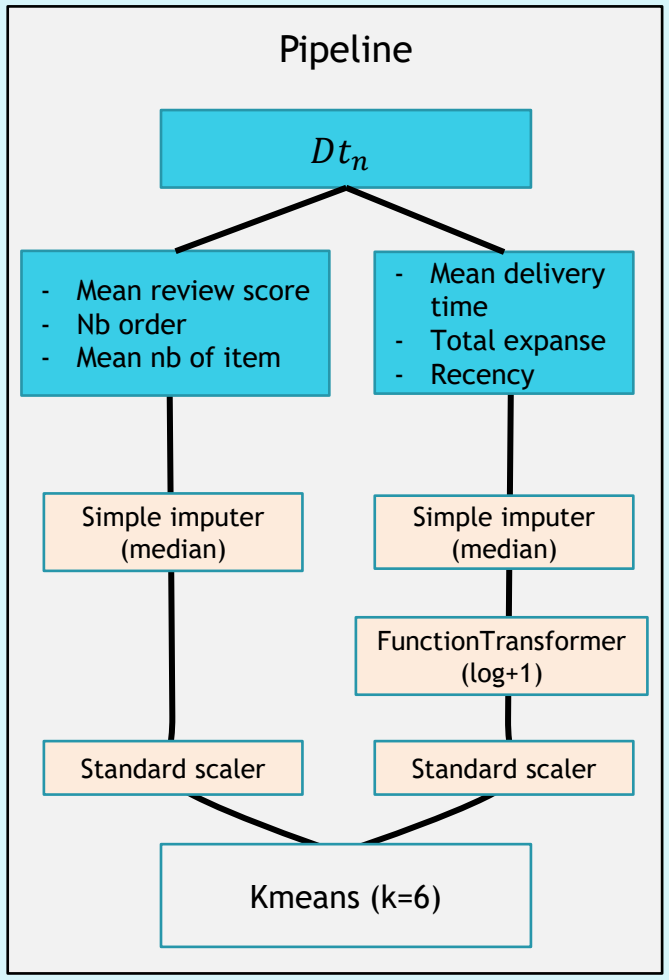
Propriétés :

- Chaque client doit être présent du début à la fin de l'expérience (premier achat > 26 semaines)
- ~~Le jeu de donnée doit évoluer (nb commande final > 1)~~




## 2) Modélisation

A) **Entraînement (Dt0)**
B) **Prédiction des clusters (Dtn)**




## 3) Stabilité

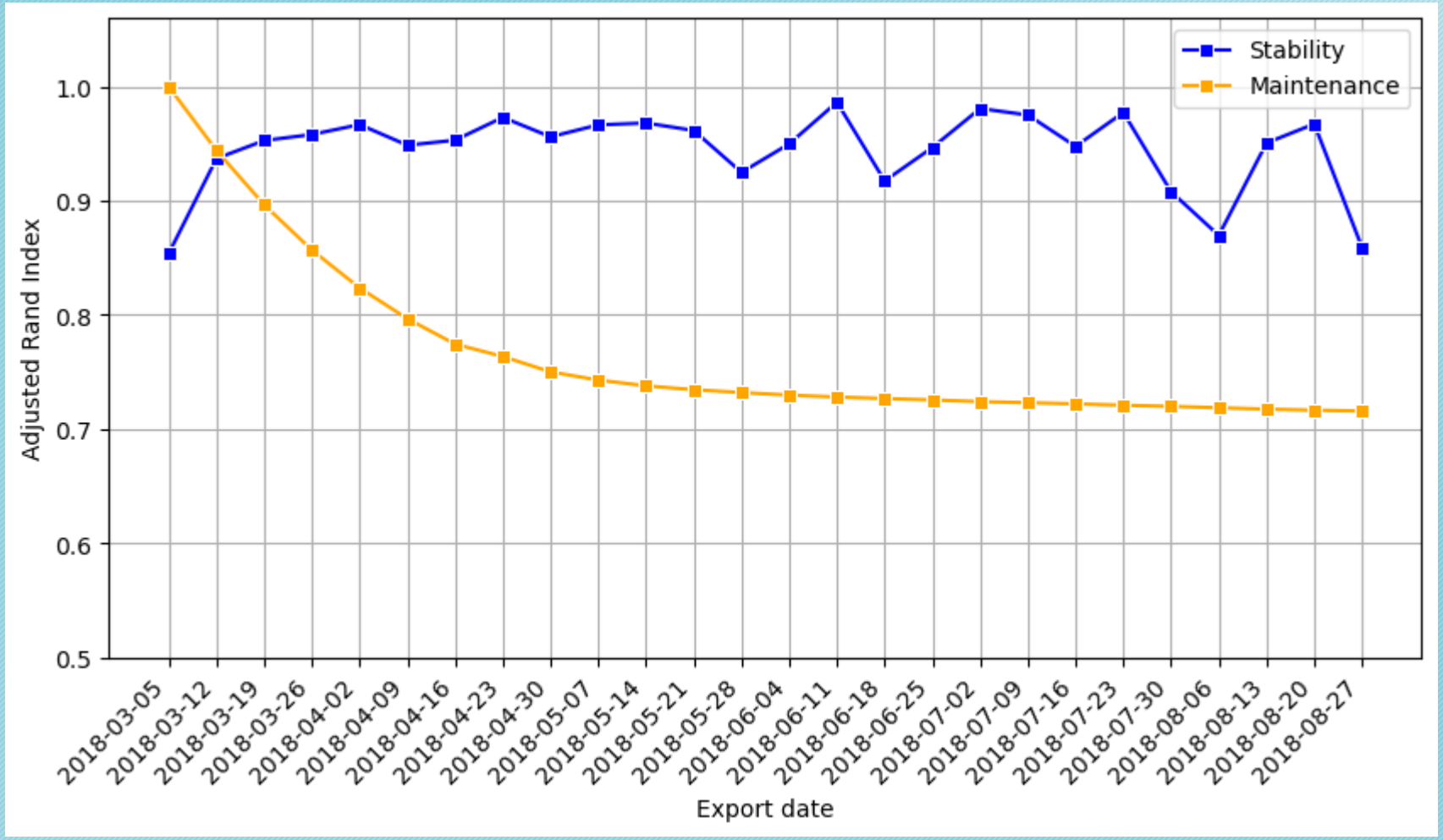
**Comparaison de la stabilité des clusters entre t0 et tn jusqu'à t26**

Adjusted Rand Index (ARI) :
Mesure de similarité entre deux segmentations par comparaison des assignations de groupe au niveau individuel.

ARI = 0 > les groupes sont aléatoires
ARI = 1 > Les deux groupes sont identiques

# B) Détermination de la periode de maintenance : Résultats





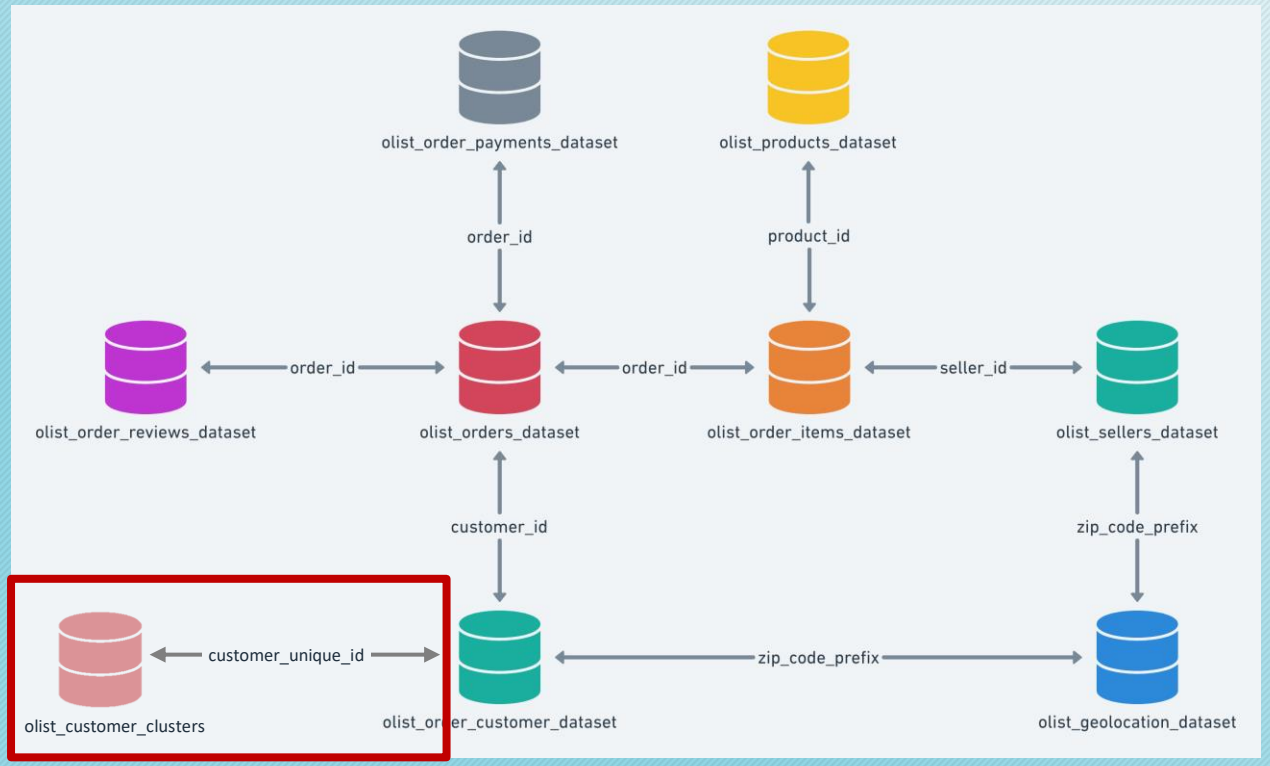

Maintenance :

- Diminution les 6 premières semaines (ARI = 0,77)
- Stabilisation au dessus d'un ARI de 0,7.

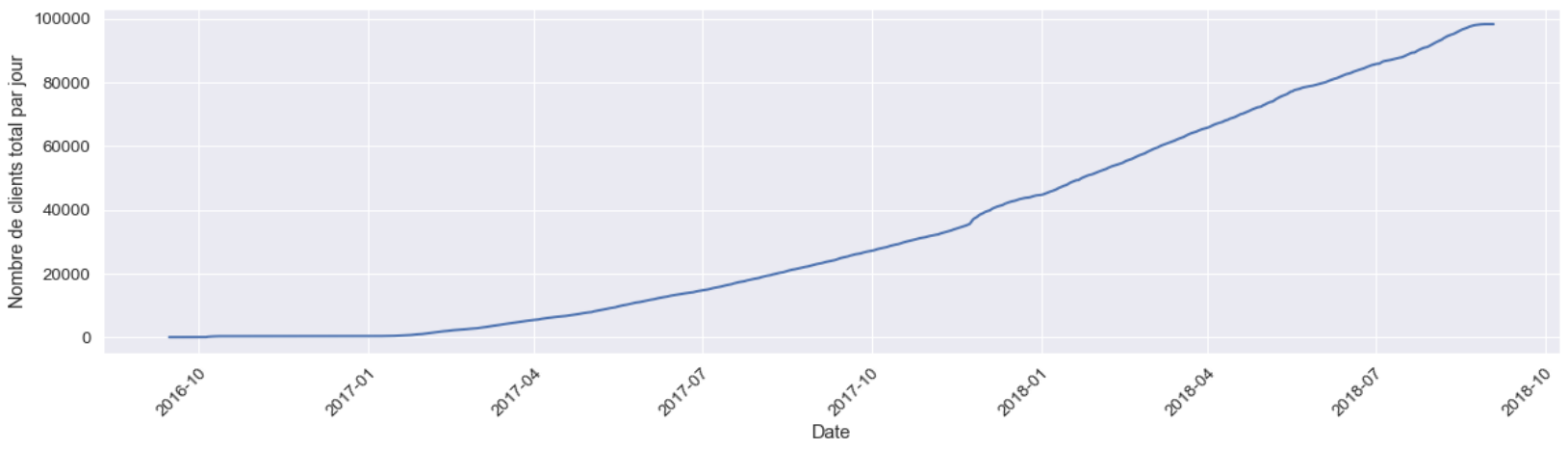
# Conclusions



- Le jeu de données permet une segmentation des clients par différentes méthodes
- La méthode des kmeans + rapide et efficace
- Clusters stables dans le temps
- Les meilleurs clients appartiennent aux clusters 3 (clients fidèles) et 4 (forte dépense)
- Top clients : ~5% des clients
- Le modèle doit être ré-entrainé toutes les 6 semaines (contrat de maintenance).
- Utilisation du code par Olist : code formaté en PEP8 par le plugin autopep8

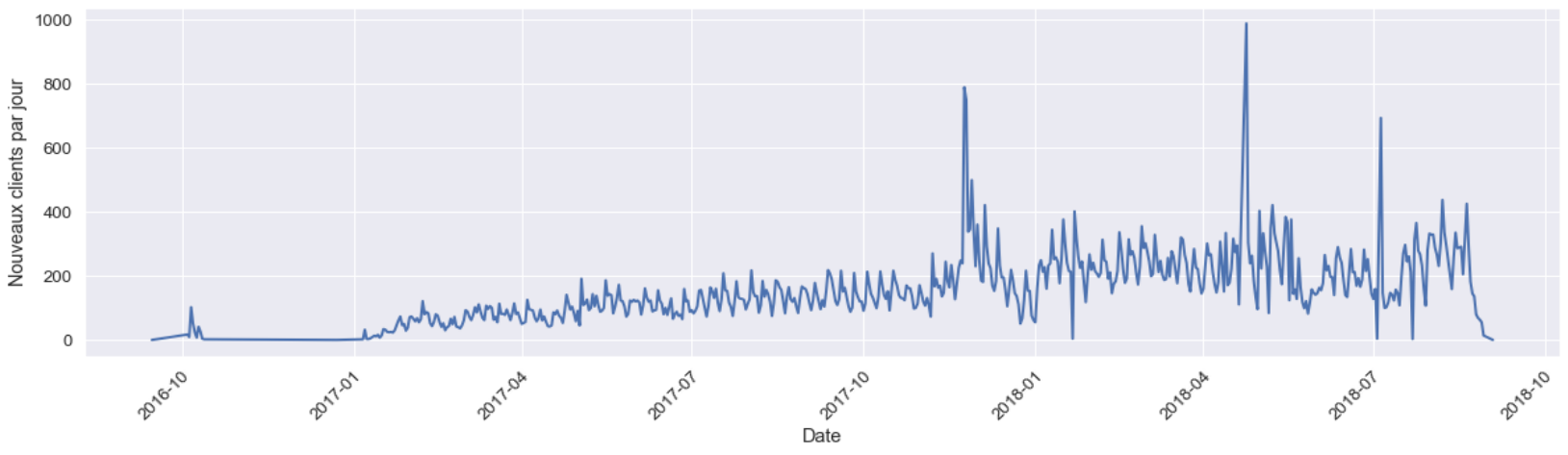


# Analyse du nombre de clients total et par jour




**3 pics d'activités :**

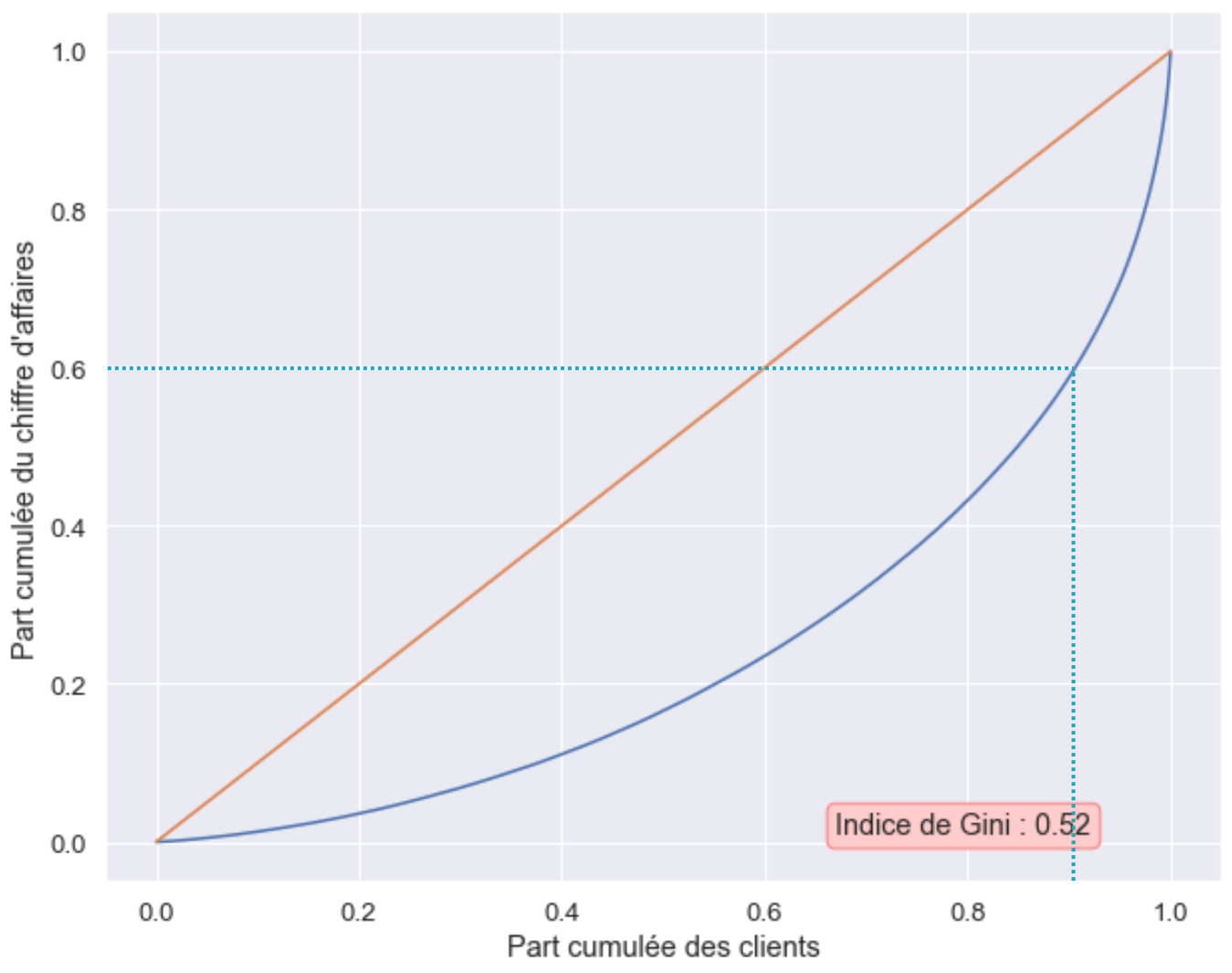
- **24-25 Novembre 2017**
- **24 avril 2018**
- **05 Juillet 2018**




- **Le nb de commandes**
- **Le nb de clients**
- **Le chiffre d'affaires**

} **Même dynamique**

# Répartition du CA en fonction des clients



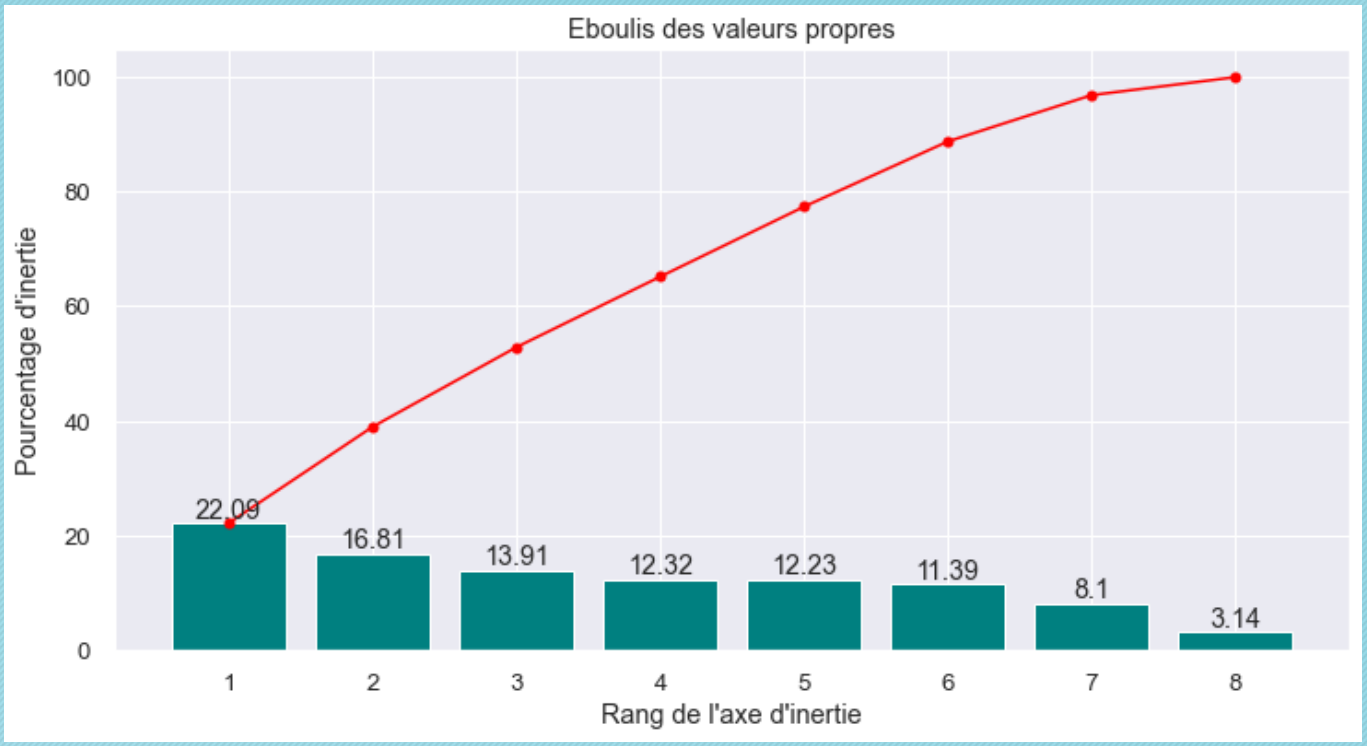

**L'indice de Gini : varie de 0 (répartition égale) à 1 (répartition inégale).**

- **Indice de Gini : 0,52**
- **10% des clients partagent 40% du chiffre d'affaires totale**

# Visualisation par Analyse en Composantes Principales

**Principe** : Résumer l'information qui est contenue dans de nombreuses variables en un certain nombre d'axes synthétiques (Composantes principales) en gardant le plus d'information possible.

**Détermination du nombre d'axes d'intérêts :**




**Critère de Kayser :** on ne garde que les composantes > (100/*p*)% où *p* est le nombre de variables.

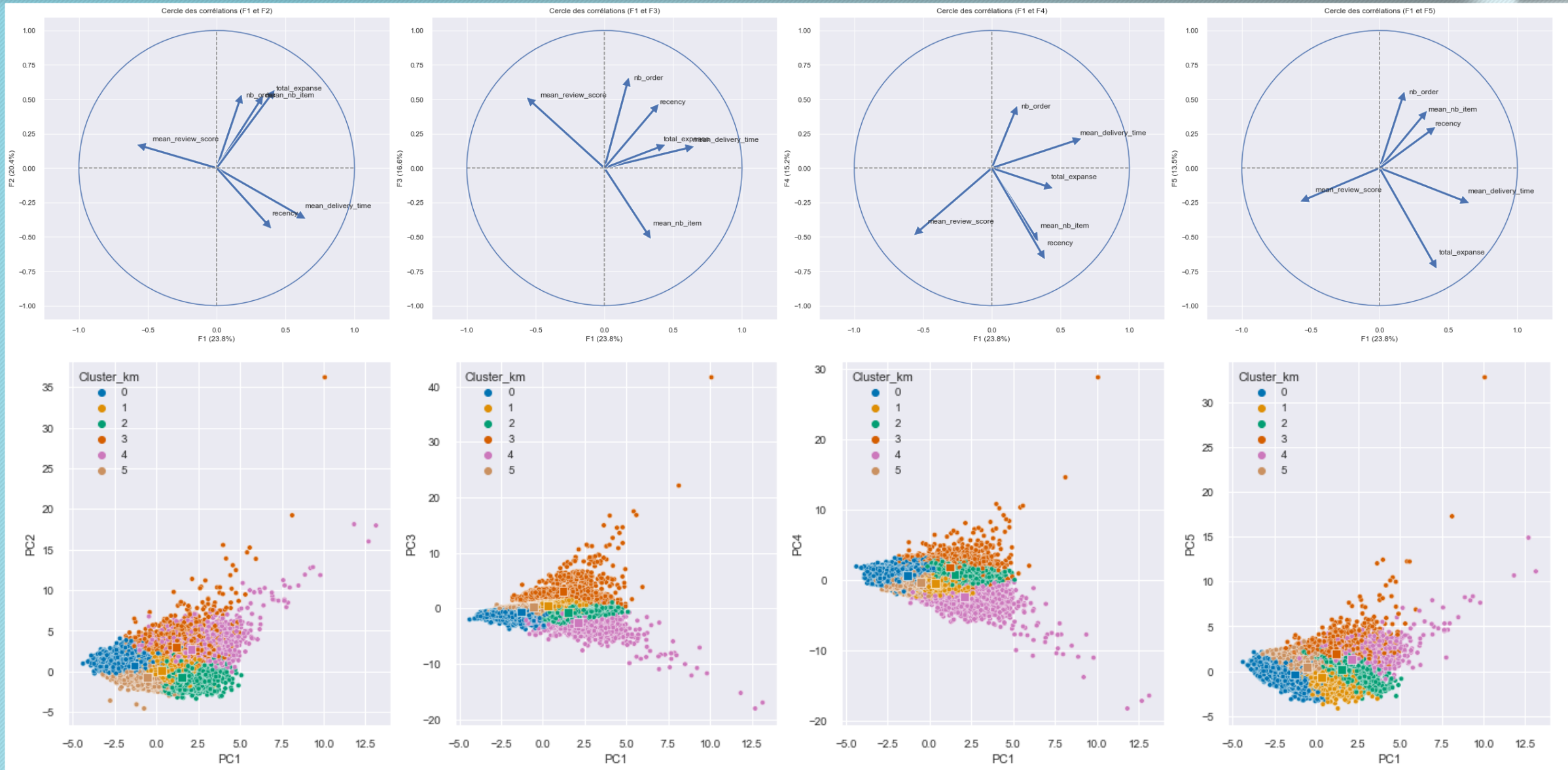
- (100/*p*)% = 12,5 % > Axe 1 à 3

**Méthode du coude :** méthode visuelle

- Axe 1 à 6

Nous analyserons les **3 premières dimensions**, expliquant 61% de la variance totale.

# Visualisation par Analyse en Composantes Principales



# Visualisation par t-SNE