



**Rapport de Projet :**  
**Analyse intelligente de la clientèle et des produits**  
**par segmentation, détection d'anomalies et**  
**modélisation temporelle**

Data Mining

Ce travail a été effectué par :

- Yomna Hajji **p2514562**

*(Partie 01 : Segmentation des produits & Détection  
d'anomalies)*

- Oulaiya Gaddari **p2513558**

*(Partie 02 : Segmentation des clients par analyse RFM &  
Détection d'anomalies)*

- Karima Messaoudi **p2415946**

*(Partie 03 : Analyse temporelle & clustering des ventes  
e-commerce)*

# Chapitre 1 : Analyse Produits

## I. Introduction générale

Ce chapitre étudie le comportement des produits pour identifier les dynamiques de vente, les anomalies et les similarités.

Deux volets sont abordés :

- **Détection d'anomalies** : repérage des produits ou transactions atypiques pour fiabiliser les données.
- **Segmentation** : regroupement des articles par clustering non supervisé

Cette double approche améliore la qualité des données et éclaire la structure du portefeuille produit.

### **Partie 1 : Détection d'anomalies:**

## II. Préparation et nettoyage des données

Ce chapitre analyse le comportement des produits à partir du jeu Online Retail II (2009–2011), comptant plus d'un million de transactions. Après un prétraitement (nettoyage, uniformisation, suppression des incohérences), les données ont été agrégées par produit selon quatre indicateurs : **prix moyen, quantité vendue, nombre de transactions et taux de retour**, couvrant **5 616** produits distincts..

StockCode	Description	Prix moyen	Quantité totale	Nb. transactions	Taux de retour
10002	Inflatable political globe	0.98	7613	398	0.018
10002R	Robot pencil sharpener	5.13	4	3	0.000
10080	Groovy cactus inflatable	0.50	315	28	0.000
10109	Bendy colour pencils	0.42	4	1	0.000
10120	Doggy rubber	0.24	650	78	0.013
...	...	...	...	...	...
gift_0001_80	Dotcomgiftshop gift voucher	69.56	0	2	0.500
m	Manual	3.01	5	5	0.000

**Tableau 1-** Exemple de jeu de données agrégé par produit

Le tableau ci-dessus illustre la table finale après agrégation des transactions par produit. Chaque ligne correspond à un article unique, identifié par son StockCode et sa Description.

## III. Détection d'anomalies produits

### 1. Préparation des variables

L'analyse repose sur quatre indicateurs agrégés par produit : `prix_moyen`, `qunatite_totale`, `nb_transactions`, `taux_retour`. Avant d'appliquer le modèle, ces variables ont été **standardisées** afin d'éviter qu'une variable à grande échelle (ex. : `quantite_totale`) domine les autres dans le calcul des distances.

### 2. Méthodologie Isolation Forest

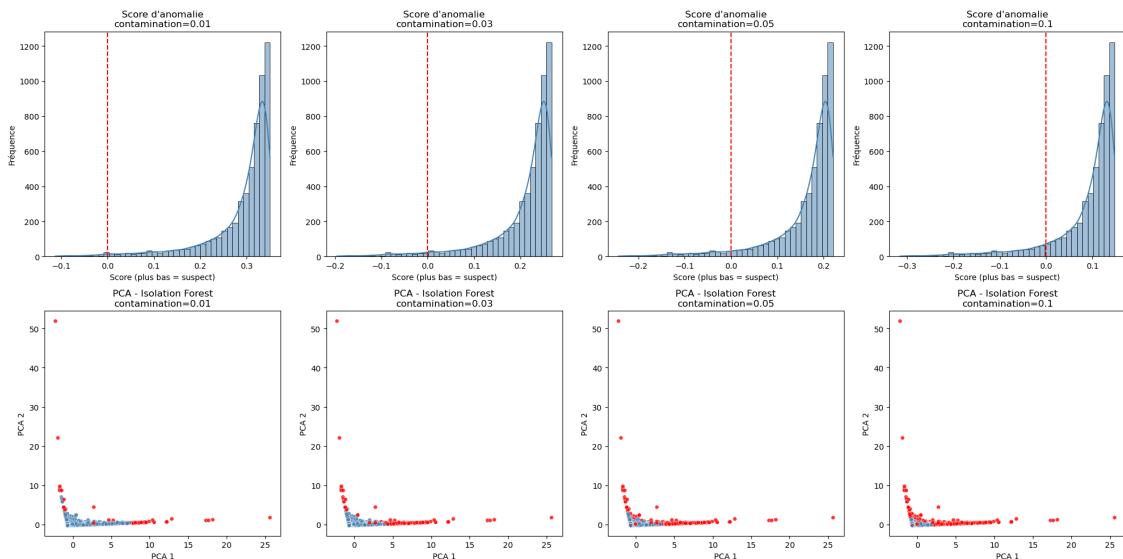
Le modèle Isolation Forest a été choisi pour sa robustesse sur des données multidimensionnelles. Fondé sur l'isolement des observations rares, il identifie efficacement les produits atypiques. Plusieurs taux de contamination (1 %, 3 %, 5 %, 10 %) ont été testés pour mesurer la sensibilité du modèle.

Après standardisation, un score d'anomalie a été calculé et chaque produit classé comme normal ou suspect selon son degré d'isolement.

## 2.1 Visualisation et interprétation des résultats

Les figures suivantes illustrent, pour chaque taux de contamination, deux analyses complémentaires :

- ❖ **Distribution des scores d'anomalie** : l'histogramme montre la répartition des scores, la ligne rouge indiquant le seuil des produits anormaux.
- ❖ **Projection PCA** : les produits normaux (bleu) se concentrent dans une zone dense, tandis que les anomalies (rouge) s'en écartent.



**Figure 1** - Visualisation des résultats du modèle Isolation Forest pour différents niveaux de contamination

## 2.2 Détection finale des anomalies

Le taux de contamination de **3 %** a été retenu comme le plus pertinent : il isole un nombre limité d'anomalies, préserve les produits normaux et offre une séparation claire dans la projection PCA. Avec ce paramètre, le modèle Isolation Forest, entraîné sur les variables normalisées (prix moyen, quantité, transactions, taux de retour), a identifié 169 produits anormaux ( $\approx 3 \%$ ), caractérisés par des prix extrêmes, des volumes atypiques ou des taux de retour élevés.

## 2.3 Analyse comparative entre produits normaux et anomalies

Une analyse statistique des principales variables a été menée pour comparer les produits normaux et les anomalies détectées par le modèle Isolation Forest.

Catégorie	Prix moyen	Quantité totale	Nb. de transactions	Taux de retour
<b>Produits normaux (anomaly_iforest = 0)</b>	<b>4,03</b>	<b>1 462</b>	<b>161</b>	<b>0,018</b>
<b>Produits anormaux (anomaly_iforest = 1)</b>	<b>244,59</b>	<b>17 559</b>	<b>1 105</b>	<b>0,188</b>

**Tableau 2** – Moyennes comparatives entre produits normaux et anormaux

La comparaison révèle des écarts marqués : les produits anormaux affichent un prix moyen 60 fois supérieur, des ventes et transactions jusqu'à 12 et 7 fois plus élevées, et un taux de retour dix fois supérieur. Ces cas peuvent indiquer des produits problématiques ou des erreurs de saisie.

### 3. Méthodologie DBSCAN

#### 3.1 Choix des paramètres du modèle DBSCAN

L'algorithme DBSCAN repose sur deux paramètres clés :  $\epsilon$  (distance maximale entre voisins) et **min\_samples** (nombre minimal de points par cluster). Leur réglage influe sur la distinction entre zones denses et bruit ; une recherche systématique a permis d'en déterminer la combinaison optimale.

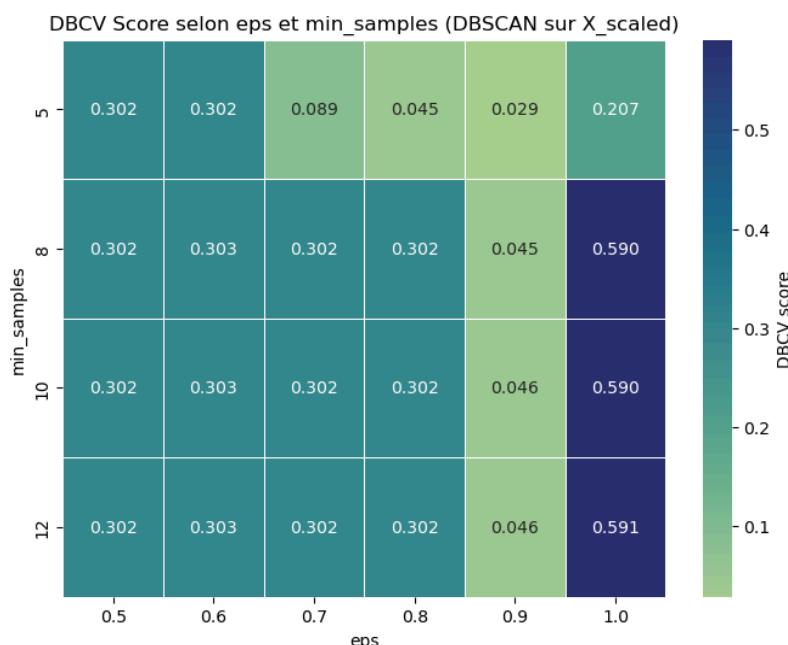
#### 3.2 Méthodologie d'évaluation

Le modèle DBSCAN a été entraîné sur les données standardisées pour diverses combinaisons de ( $\epsilon$ , **min\_samples**). Deux critères ont été évalués : le nombre de clusters et le taux d'anomalies, ainsi que le DBCV Score mesurant la qualité de la structure. Les valeurs testées étaient  $\epsilon \in [0.5-1.0]$  et **min\_samples**  $\in [5, 8, 10, 12]$ .

Eps	Min_samples	Nb. clusters	% anomalies
<b>0.5</b>	5	3	2.03 %
<b>0.7</b>	5	6	1.19 %
<b>0.9</b>	10	3	1.41 %
<b>1.0</b>	8	2	1.01 %
<b>1.0</b>	12	2	<b>1.28 %</b>

**Tableau 3** – Résumé des résultats du modèle DBSCAN selon  $\epsilon$  (eps) et **min\_samples**

#### 3.3 Évaluation par DBCV score



**Figure 2** – Évaluation du modèle DBSCAN à l'aide du DBCV

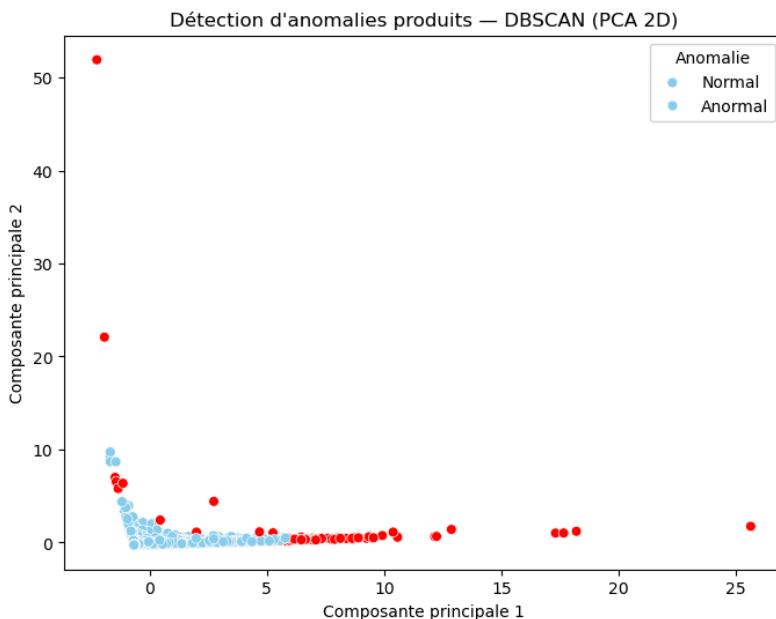
La heatmap illustre les valeurs du DBCV pour différentes combinaisons d'eps et min\_samples du modèle DBSCAN. Un score élevé (proche de 1) indique des clusters denses et bien séparés, tandis qu'un score faible traduit une cohésion limitée.

Le meilleur résultat ( $\approx 0.591$ ), obtenu pour **eps = 1.0** et **min\_samples = 12**, correspond à la configuration la plus stable pour la détection d'anomalies.

### 3.4 Résultats et interprétation du modèle DBSCAN

Avec les paramètres optimaux ( **$\epsilon = 1.0$** , **min\_samples = 12**), le modèle DBSCAN appliqué aux données normalisées a identifié **2 clusters** et **72 anomalies** ( $\approx 1,28\%$  du catalogue), correspondant à des produits isolés hors des zones denses.

#### 3.5.1 Visualisation PCA des anomalies



**Figure 3 – Détection d'anomalies produits avec DBSCAN (projection PCA 2D)**

#### 3.5.2 Analyse comparative entre produits normaux et anomalies

Is_anomaly	Prix_moyen	Quantite_totale	Nb_transactions	Taux_retour
<b>0 (Normal)</b>	4.62	1 600.10	171.78	0.02
<b>1 (Anormal)</b>	523.26	28 588.53	1 511.74	0.09

**Tableau 4 – Moyennes des variables selon le type de produit (DBSCAN)**

Les produits anormaux présentent un prix 130 fois supérieur, des ventes 18 fois plus élevées et un taux de retour 4 à 5 fois supérieur. DBSCAN confirme les tendances d'Isolation Forest, mais avec une proportion plus faible d'anomalies (1,28 % contre 3,01 %), ciblant les cas les plus extrêmes.

## 4. Comparaison de Isolation Forest et DBSCAN

Pour évaluer la cohérence entre Isolation Forest et DBSCAN, les résultats ont été croisés afin d'identifier les anomalies communes, celles propres à chaque modèle, et la proportion globale des produits normaux.

Source_anomalie	Nombre	Pourcentage (%)
Aucune	5 445	96.96
Isolation Forest seul	99	1.76
Les deux	70	1.25
DBSCAN seul	2	0.04

**Tableau 5** – Répartition des anomalies selon leur source de détection

#### 4.1 Interprétation des résultats

Les deux modèles détectent 70 anomalies communes ( $\approx 1,25\%$  du catalogue). DBSCAN, plus sélectif, n'en identifie que 72, tandis qu'Isolation Forest, plus sensible, en détecte 169, dont 99 moins extrêmes.

### 5. Détection combinée : approche hybride Isolation Forest + DBSCAN

#### 5.1 Objectif

Une approche combinée d'Isolation Forest et DBSCAN a été adoptée pour allier la sensibilité du premier (anomalies modérées) et la sélectivité du second (cas extrêmes), offrant une détection plus robuste fondée sur la rareté statistique et l'isolement géométrique.

#### 5.2 Méthodologie

Le score d'Isolation Forest a été inversé et normalisé pour que les valeurs proches de 1 indiquent les produits les plus suspects. Le score DBSCAN, initialement binaire, a été transformé pour être combiné au précédent. Le score final correspond à la moyenne des deux scores, et le 98<sup>e</sup> percentile a été fixé comme seuil, ne conservant que les 2 % de produits les plus suspects comme anomalies finales.

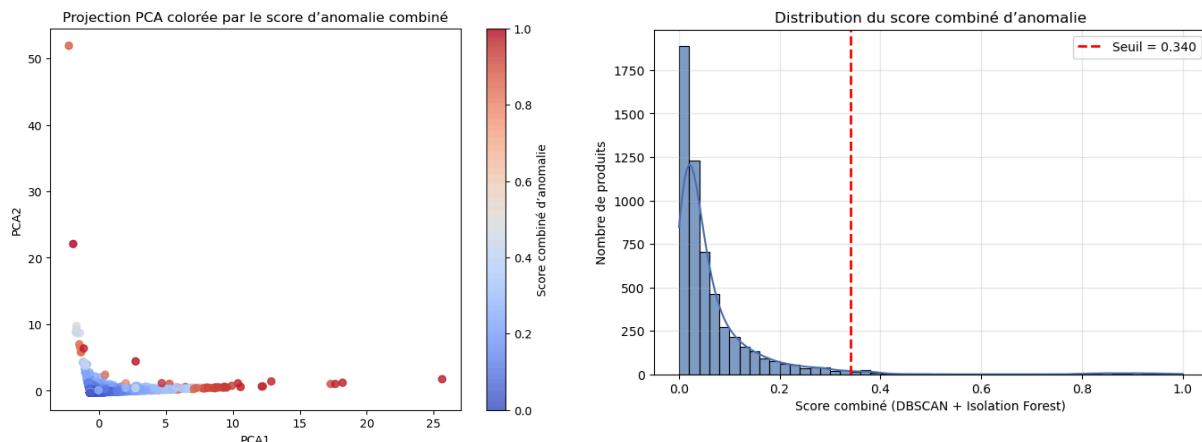
#### 5.3 Résultats de la détection combinée

Type de produit	Nombre	Pourcentage
Normaux (is_anomaly_vote = 0)	5 503	97,99 %
Anormaux (is_anomaly_vote = 1)	113	2,01 %

**Tableau 6** – Résultats globaux de la détection combinée Isolation Forest + DBSCAN

Le seuil optimal retenu est 0,340, correspondant à la limite supérieure du 98<sup>e</sup> percentile du score combiné. La détection combinée identifie 113 produits anormaux sur 5 616, soit environ 2 % du catalogue.

### 5.3.1 Distribution du score combiné et Visualisation PCA



**Figure 4 – Distribution du score combiné d'anomalie et visualisation PCA (DBSCAN + Isolation Forest)**

### 5.4 Analyse qualitative des produits anormaux

Les produits atypiques se répartissent en plusieurs catégories :

- ❖ **Produits internes ou administratifs** (ex. *amazon fee, discount*), liés à des ajustements comptables plutôt qu'à des ventes réelles.
- ❖ **Produits à volumes extrêmes** (ex. *white hanging heart t-light holder*), vendus en quantités massives, souvent saisonniers.
- ❖ **Produits à taux de retour élevé**, signalant des problèmes de qualité ou de logistique.
- ❖ **Produits incohérents**, présentant des valeurs erronées (prix nuls, quantités négatives).

Globalement, les anomalies détectées se divisent entre anomalies métiers légitimes et anomalies techniques liées à la qualité des données.

## 6. Conclusion

La détection d'anomalies a permis d'isoler environ 2 % de produits suspects au comportement atypique (prix élevés, volumes incohérents, taux de retour excessifs). L'approche combinée Isolation Forest–DBSCAN a renforcé la fiabilité en alliant sensibilité et sélectivité, identifiant 113 produits à exclure pour garantir une base de données propre et représentative avant la phase de segmentation. Ces produits ont été exclus afin d'assurer une segmentation basée sur des données propres et représentatives du catalogue.

## **Partie 2 :Segmentation des produits :**

### **I. Contexte**

Après avoir identifié et exclu les produits présentant des comportements anormaux, l'analyse peut désormais se concentrer sur la segmentation des produits.

Cette étape vise à regrouper les articles selon leurs caractéristiques communes (prix, volume vendu, fréquence d'achat, revenu généré) afin de mettre en évidence des profils types de produits.

L'objectif est de mieux comprendre la structure du portefeuille et d'identifier des groupes homogènes facilitant l'interprétation commerciale

### **II. Préparation et nettoyage de données**

#### **1. Description du jeu de donnée**

Après suppression des anomalies, le jeu de données comprend 290 613 transactions et 4 932 produits uniques. Pour la segmentation, les ventes ont été agrégées par produit afin de calculer quatre indicateurs clés : **quantité totale, chiffre d'affaires, prix moyen et nombre de factures distinctes** décrivant la popularité, le volume, le revenu et le positionnement prix de chaque article.

L'agrégation a produit un nouveau jeu de données où chaque ligne représente un produit unique avec ses statistiques synthétiques :

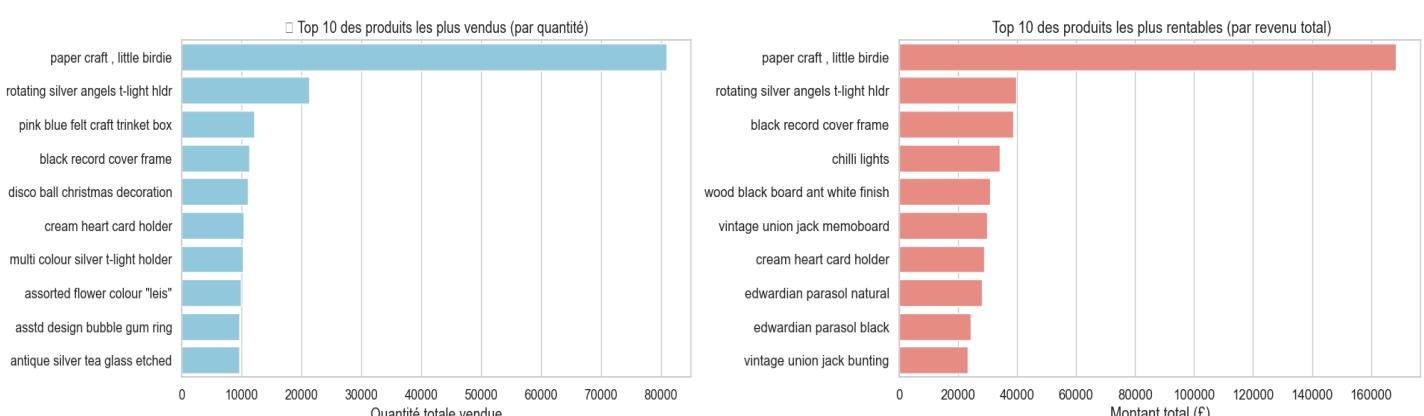
Description	Quantity	TotalPrice (€)	Price moyen (€)	NumInvoices
<b>10 colour spaceboy pen</b>	5 579	4 634.51	0.85	202
<b>11 pc ceramic tea set polkadot</b>	3	14.85	4.95	1
<b>12 ass zinc christmas decorations</b>	241	506.10	2.10	24
<b>12 coloured party balloons</b>	2 056	1 276.40	0.65	100
<b>12 daisy pegs in wood box</b>	393	648.45	1.65	72

**Tableau 7-** Extrait du tableau complet des statistiques produits

### **III. Analyse exploratoire des produits**

Une analyse exploratoire du catalogue agrégé a été réalisée pour identifier les produits dominants en volume et rentabilité, et mieux comprendre les écarts de performance.

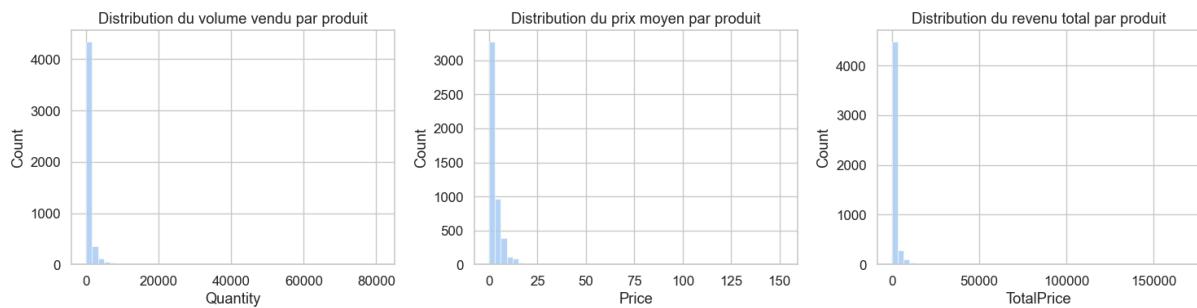
#### **1. Produits les plus vendus et les plus rentables**



**Figure 5 – Top 10 des produits les plus vendus (par quantité totale) et les plus rentables**

Les ventes sont fortement concentrées sur quelques produits, dont *Paper craft, little birdie* (>80 000 unités), illustrant la longue traîne du e-commerce. En revenu, la hiérarchie reste similaire, dominée par ces articles ainsi que par des produits décoratifs (*Black record cover frame, Chilli lights*) et quelques articles premium (*Edwardian parasol*).

## 2. Analyse des distributions des variables produits



**Figure 6 – Distributions des principales variables par produit**

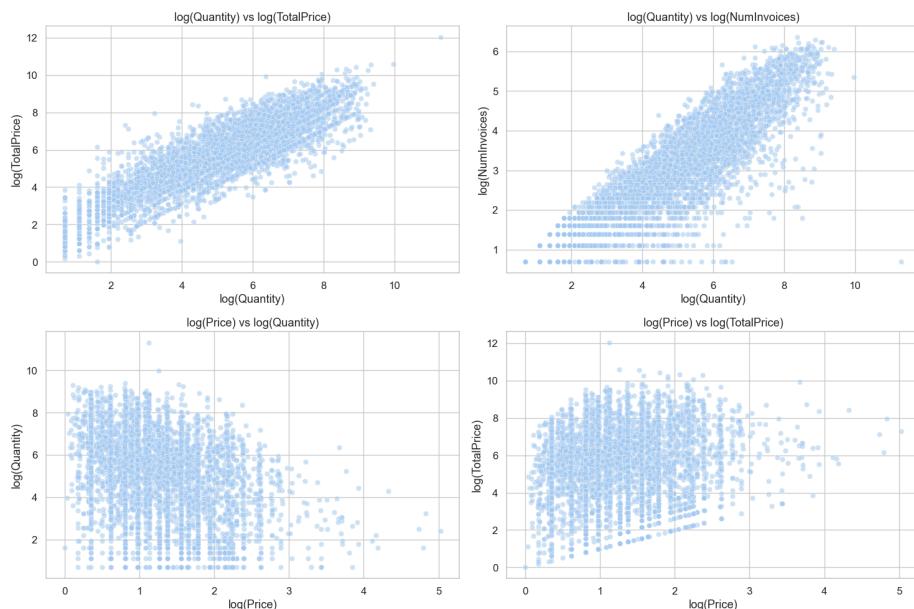
Le graphique montre la distribution des indicateurs Quantity, Price et TotalPrice, révélant une forte asymétrie à droite : la majorité des produits affichent de faibles ventes et prix, tandis qu'une minorité présente des valeurs très élevées (best-sellers ou premium).

Cette dispersion justifie la normalisation des variables avant le clustering.

## 3. Analyse des relations entre variables

Afin d'examiner les corrélations entre les principales variables sans être biaisé par les valeurs extrêmes, une transformation logarithmique a été appliquée sur les indicateurs Quantity, Price, TotalPrice et NumInvoices.

Les graphiques ci-dessous présentent les relations entre ces variables deux à deux.

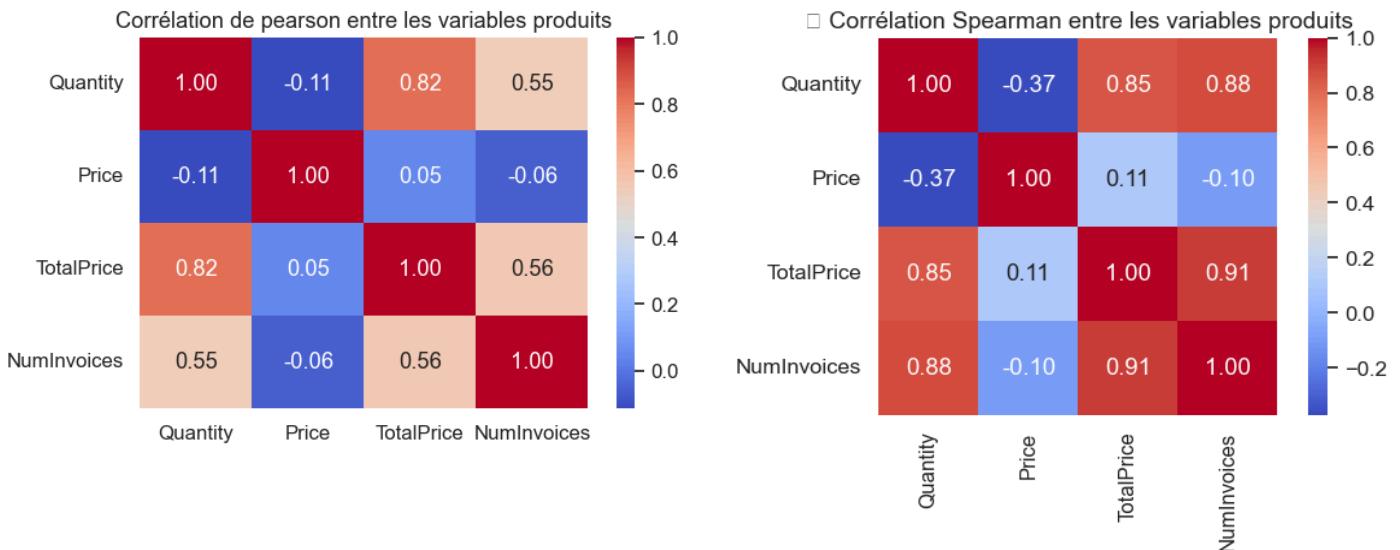


**Figure 7 – Relations entre variables après transformation logarithmique**

L'analyse des relations log-transformées met en évidence plusieurs tendances :

- ❖ Une forte corrélation positive entre *Quantity* et *TotalPrice*
- ❖ une corrélation positive entre *Quantity* et *NumInvoices* (produits populaires)
- ❖ Une corrélation négative faible entre *Price* et *Quantity*
- ❖ Enfin, *Price* et *TotalPrice* présentent une corrélation modérée

#### 4. Analyse de corrélation entre les variables produits



**Figure 8 – Matrices de corrélation de Pearson et de Spearman entre les variables produits**

Les corrélations entre *Quantity*, *Price*, *TotalPrice* et *NumInvoices* ont été évaluées à l'aide des coefficients de Pearson (relation linéaire) et Spearman (relation monotone).

#### 5.1 Corrélation de Pearson

Les corrélations de Pearson révèlent des liens forts entre *Quantity*, *TotalPrice* ( $r = 0.82$ ) et *NumInvoices* ( $r \approx 0.55$ ), confirmant que les produits les plus vendus génèrent plus de revenus et d'achats. En revanche, *Price* est faiblement corrélé, montrant que le prix moyen influence peu le volume ou la fréquence d'achat.

#### 5.2 Corrélation de Spearman

La corrélation de Spearman, mieux adaptée aux distributions non linéaires et asymétriques, confirme les tendances observées. Les variables *Quantity*, *TotalPrice* et *NumInvoices* présentent des corrélations très fortes entre elles ( $\rho \approx 0.85\text{--}0.91$ ), révélant une même dimension liée à la popularité et à la performance commerciale des produits.

En revanche, *Price* montre une corrélation faible et négative avec *Quantity* ( $\rho = -0.37$ ). Le test de significativité ( $p \approx 0.00000 < 0.05$ ) indique que cette relation est hautement significative, excluant une coïncidence aléatoire.

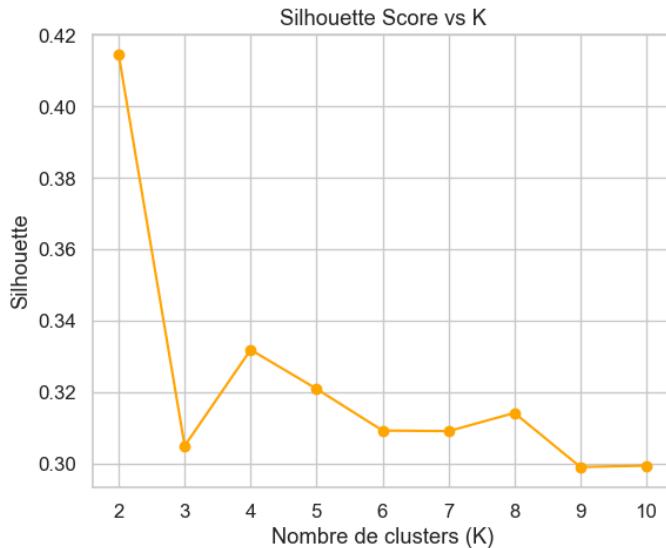
Ainsi, on peut conclure qu'il existe une relation monotone négative réelle entre le prix et la quantité vendue : à mesure que le prix augmente, les ventes diminuent, illustrant la loi classique de la demande dans le comportement des clients.

## IV. Clustering des produits

### 1. K-Means

#### 1.1 Choix du nombre de clusters (K)

Pour déterminer le nombre optimal de segments, plusieurs modèles K-Means ont été testés avec  $K$  variant de 2 à 10. Le score de silhouette a servi d'indicateur de performance : un score élevé indique des clusters bien séparés, tandis qu'un score faible ou négatif traduit une mauvaise segmentation.



**Figure 9 – Évolution du score de silhouette selon le nombre de clusters**

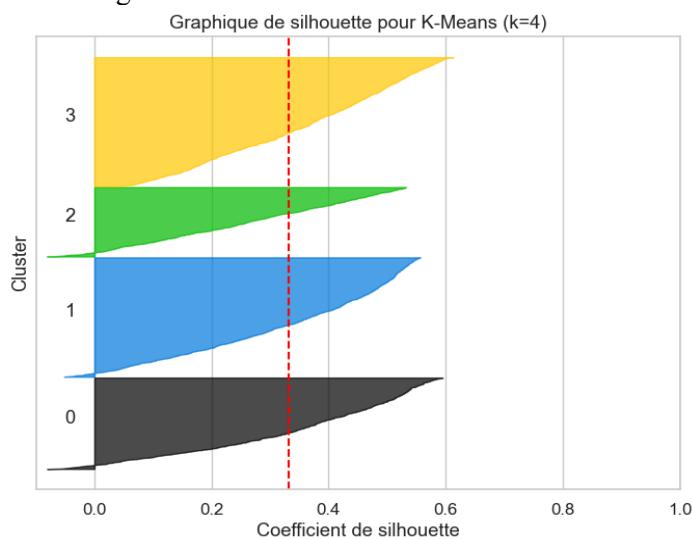
Le meilleur score de silhouette ( $\approx 0.41$ ) est obtenu pour  $K = 2$ , mais la segmentation est trop grossière. Le compromis optimal se situe à  $K = 4$  (score  $\approx 0.33$ ), offrant un bon équilibre entre cohérence statistique et pertinence métier.

Le choix de  $K = 4$  permet ainsi d'identifier clairement quatre profils produits :

- ❖ **des produits stars**, très vendus et générant le plus fort chiffre d'affaires,
- ❖ **des produits premium**, à prix élevés mais volumes plus faibles,
- ❖ **des produits de rotation moyenne**, à prix bas et quantité élevé.
- ❖ **des produits dormants**, à faible activité commerciale.

#### 1.2 Analyse du graphique de silhouette (K-Means, k = 4)

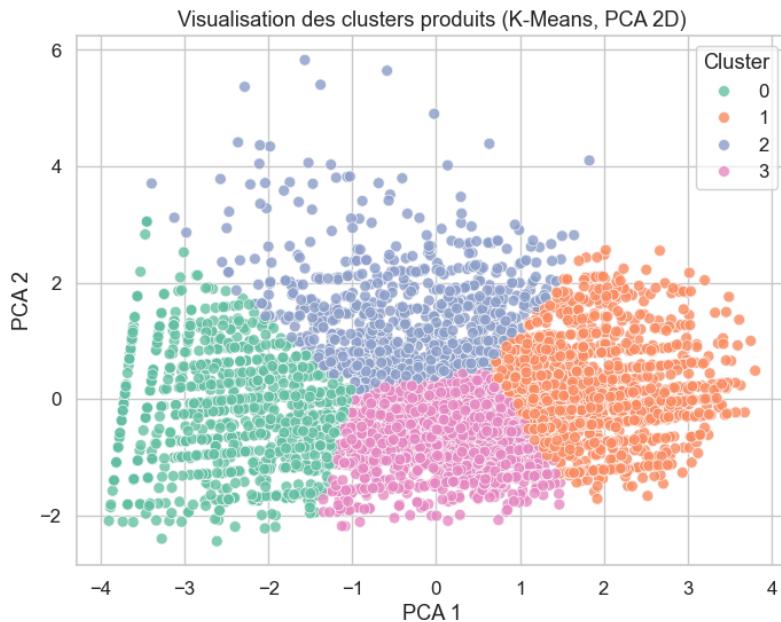
Le score de silhouette moyen obtenu pour le modèle K-Means à  $k = 4$  est de 0.332, ce qui indique une cohésion interne correcte et une séparation modérée entre les groupes. Bien qu'il ne s'agisse pas d'un score élevé ( $\geq 0.5$ ), cette valeur reste tout à fait acceptable dans le cadre d'une segmentation sur des données économiques et hétérogènes.



On observe que : Les clusters 1 et 3 présentent les meilleurs coefficients de silhouette, indiquant des groupes bien séparés. Les clusters 1 et 3 sont bien séparés, tandis que les clusters 0 et 2 sont plus dispersés mais cohérents. Le modèle K-Means ( $K = 4$ ) offre ainsi une segmentation équilibrée, reflétant la structure commerciale et la diversité des produits.

### 1.3 Visualisation des clusters produits

La PCA projette les produits selon leurs caractéristiques normalisées ; ses deux premières composantes résument l'essentiel de la variance et illustrent la segmentation K-Means ( $K = 4$ ).



**Figure 11** – Représentation des 4 clusters produits dans l'espace PCA

La figure montre une séparation claire des quatre clusters K-Means, confirmant un partitionnement de qualité. Les groupes sont distincts avec peu de chevauchement. Comme K-Means segmente mieux les structures circulaires, la forme quasi circulaire des clusters ici renforce la cohérence du modèle.

### 1.4 Profil et interprétation des clusters produits

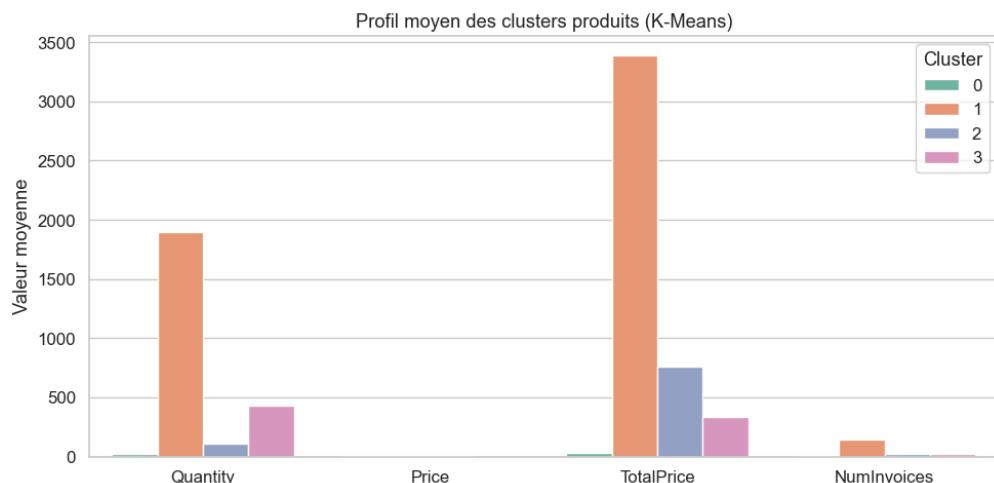
Le tableau ci-dessous présente les moyennes des principales variables quantitatives pour chaque groupe identifié par le modèle K-Means ( $k = 4$ ) :

Cluster	Quantity	Price (£)	TotalPrice (£)	NumInvoices
<b>0</b>	19.38	3.25	36.40	3.52
<b>1</b>	1893.41	2.61	3384.71	141.57
<b>2</b>	113.06	9.29	759.34	25.33
<b>3</b>	427.18	1.34	336.20	26.67

**Tableau 8** - profil moyen des produits par cluster

- ❖ **Cluster 0:** Produits dormants / peu demandés : faibles volumes et revenus modestes, achetés de manière ponctuelle.
- ❖ **Cluster 1:** Produits stars / best-sellers : fortes quantités vendues, prix accessibles, présents dans un grand nombre de factures , ils génèrent la majorité du chiffre d'affaires.
- ❖ **Cluster 2:** Produits premium : prix unitaires élevés, volumes limités mais revenu moyen par transaction important , produits de niche ou à forte valeur perçue.
- ❖ **Cluster 3:** Produits à rotation moyenne : prix très bas, ventes régulières mais revenus modérés, articles de base ou consommables.

Le graphique illustre les profils moyens des clusters selon les variables clés, montrant des différences nettes et confirmant la pertinence de la segmentation.



**Figure 12-** Profil moyen des clusters produits selon les principales variables quantitatives

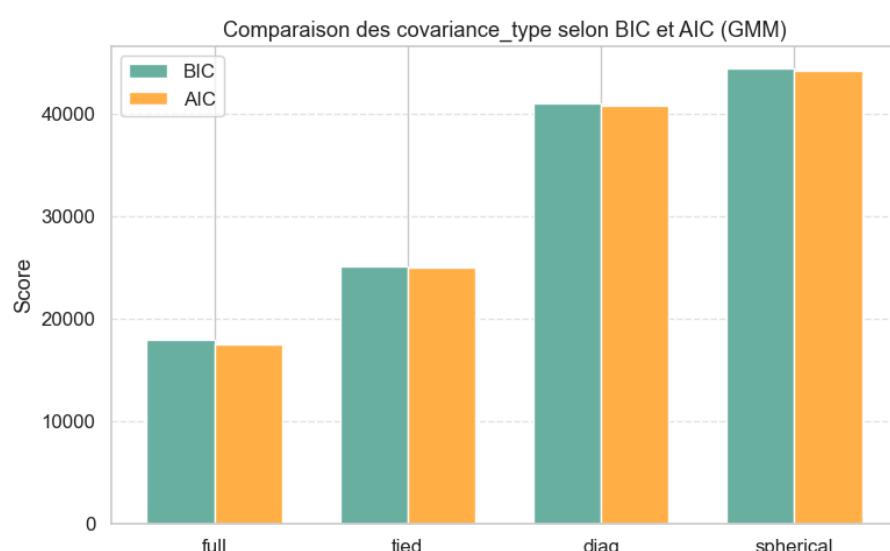
## 2. GMM

Afin d'évaluer la pertinence du modèle GMM pour la segmentation des produits, plusieurs types de matrices de covariance ont été testés : full, tied, diag et spherical.

Deux approches complémentaires ont été utilisées pour comparer les modèles :

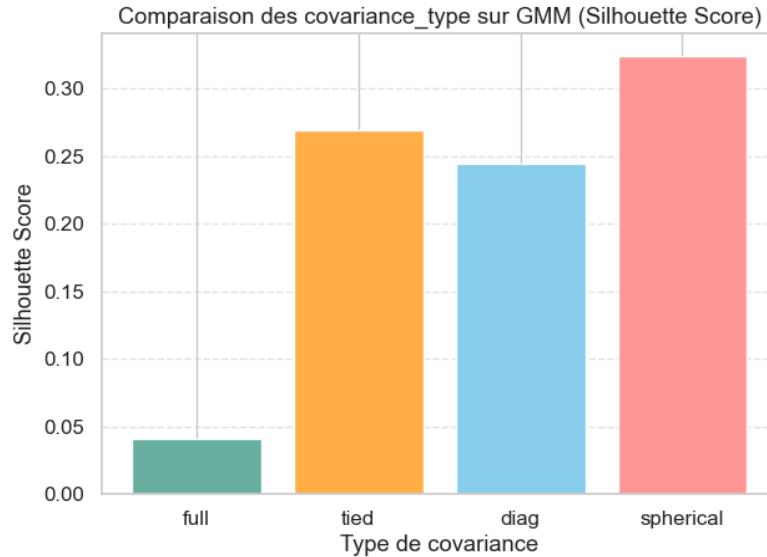
- ❖ les critères d'information **AIC** et **BIC**, qui mesurent la qualité d'ajustement statistique,
- ❖ et le **Silhouette Score**, qui évalue la cohérence des clusters obtenus.

### 2.1 Analyse des critères AIC et BIC



Les résultats montrent que le modèle covariance\_type='full' obtient les valeurs AIC et BIC les plus faibles, indiquant le meilleur compromis entre précision du modèle et complexité. Les autres configurations (tied, diag, spherical) affichent des scores plus élevés, traduisant un ajustement moins optimal. D'un point de vue purement statistique, le modèle à covariance complète ("full") est le plus performant.

## 2.2 Analyse du Silhouette Score



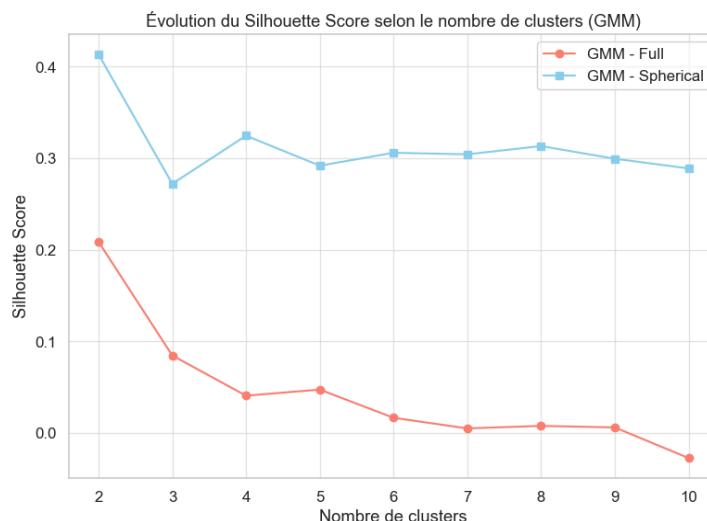
En revanche, le Silhouette Score, qui mesure la qualité de la segmentation dans l'espace des données, est le plus élevé pour le modèle covariance\_type='spherical' ( $\approx 0.32$ ). Ce type de covariance favorise des clusters plus compacts et mieux séparés, suggérant une meilleure interprétabilité des groupes formés. Sur le plan de la cohérence des clusters, la configuration "spherical" fournit la segmentation la plus claire.

## 2.3 Interprétation générale

Le modèle "full" offre la meilleure précision statistique (AIC/BIC), tandis que le "spherical" procure une segmentation plus claire visuellement. Le choix dépend donc de l'objectif : rigueur statistique ou lisibilité opérationnelle.

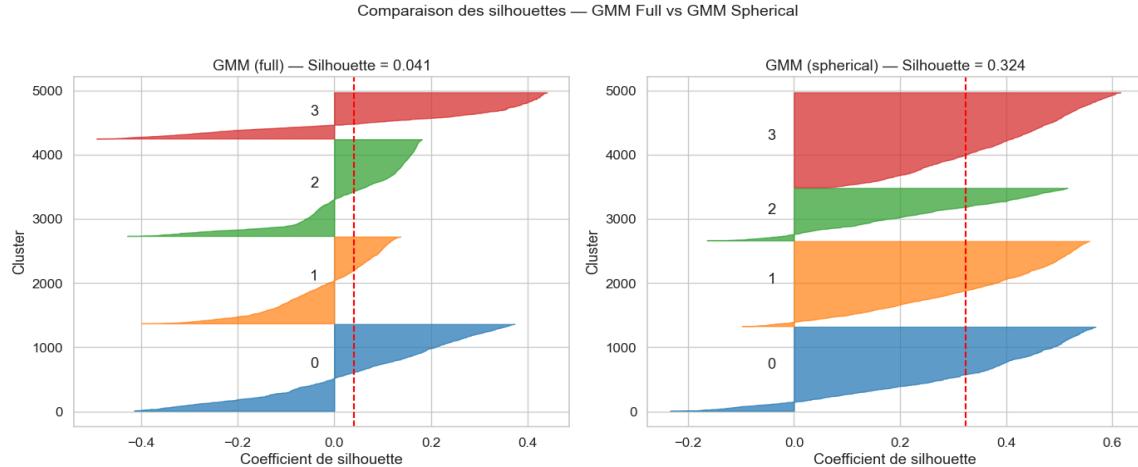
## 2.4 Analyse des modèles GMM: Full VS spherical

### 2.4.1 Évolution du Silhouette Score



Le Silhouette Score montre la supériorité du GMM Spherical, avec un score maximal de 0.41 pour 2 à 4 clusters, signe d'une meilleure séparation. Le GMM Full, en revanche, affiche des scores faibles ( $\approx 0.04$ ), traduisant un chevauchement important entre les groupes.

## 2.4.2 Interprétation des silhouettes

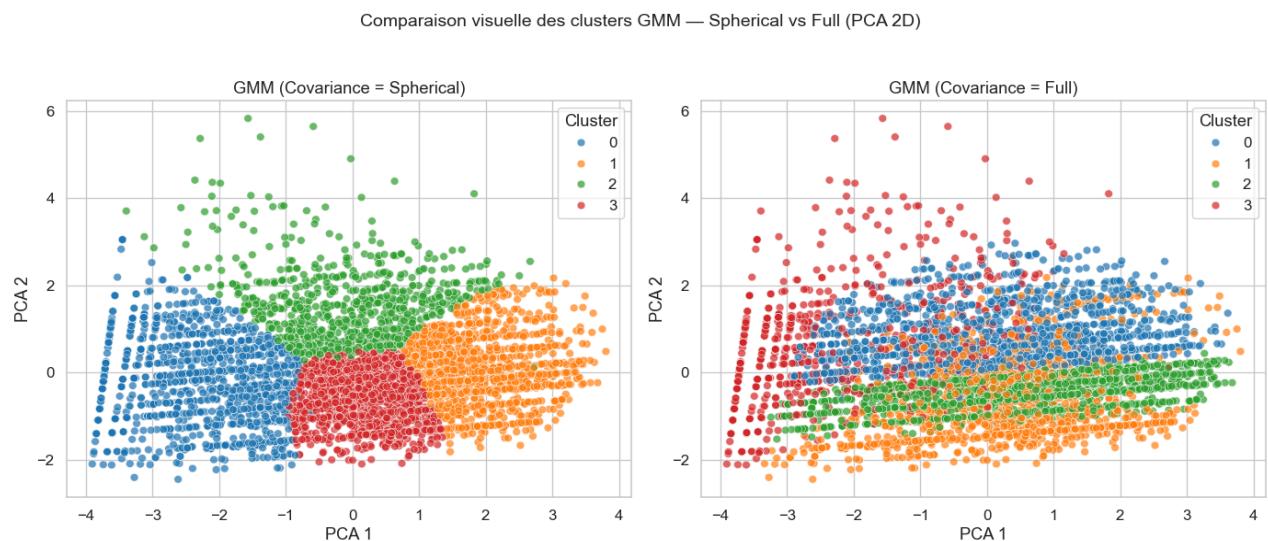


**Figure 16-** Comparaison des silhouettes GMM full VS spherical

L'analyse des silhouettes (graphique 2) confirme ces observations :

- ❖ Le **GMM Full** montre des silhouettes très dispersées autour de 0, ce qui traduit des frontières floues et une forte interférence entre clusters.
- ❖ Le **GMM Spherical**, au contraire, présente des silhouettes globalement positives ( $\approx 0.32$ ), suggérant des clusters bien définis et homogènes.

## 2.4.3 Visualisation PCA



**Figure 17-** Visualisation PCA GMM full VS spherical

La PCA montre que le GMM Spherical produit des clusters plus compacts et bien séparés, tandis que le GMM Full présente un fort chevauchement.

Avec un Silhouette Score de 0.324 contre 0.041, le modèle spherical s'avère plus stable, cohérent et interprétable, offrant une meilleure distinction entre les groupes de produits

## 2.5 Interprétation des clusters GMM (spherical)

Le GMM (spherical) a identifié quatre profils produits distincts à partir des variables de vente, révélant des comportements commerciaux différenciés.

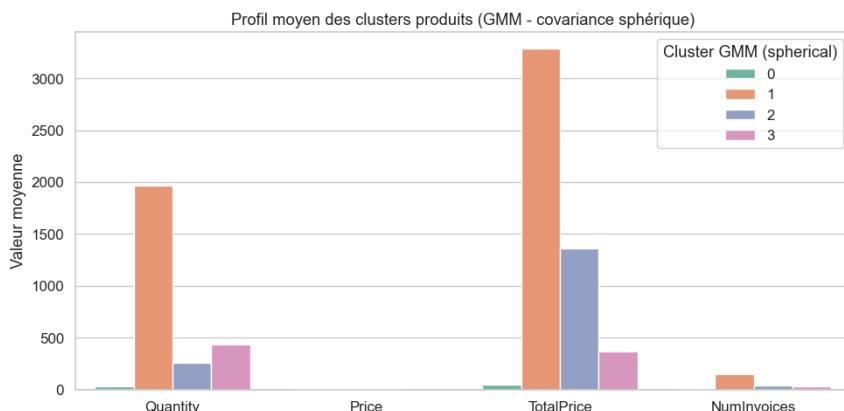
Cluster_GMM_Spherical	Quantity	Price	TotalPrice	NumInvoices
0	29.03	3.20	47.69	4.21
1	1 969.02	2.28	3 283.39	145.13
2	262.28	9.64	1 364.20	36.96
3	436.21	1.41	366.90	28.64

**Tableau 9-** Moyennes des indicateurs de vente par cluster (GMM Spherical)

0 – *Produits dormants* | 1 – *Produits stars* | 2 – *Produits premium* | 3 – *Produits à rotation*

## 2.6 Interprétation du profil moyen des clusters (GMM Spherical)

Le graphique ci-dessus illustre la moyenne des principales variables de vente pour chaque cluster identifié par le modèle GMM à covariance sphérique



## 3. Evaluation comparative du clustering K-Means VS GMM

Pour évaluer les modèles de clustering, en plus du Silhouette Score, on utilise souvent Calinski–Harabasz et Davies–Bouldin (surtout quand on compare des méthodes comme GMM et K-Means).

Modèle	Silhouette	Calinski–Harabasz	Davies–Bouldin
K-Means	<b>0.3318</b>	<b>3704.11</b>	0.941
GMM (spherical)	0.3245	3606.08	<b>0.938</b>

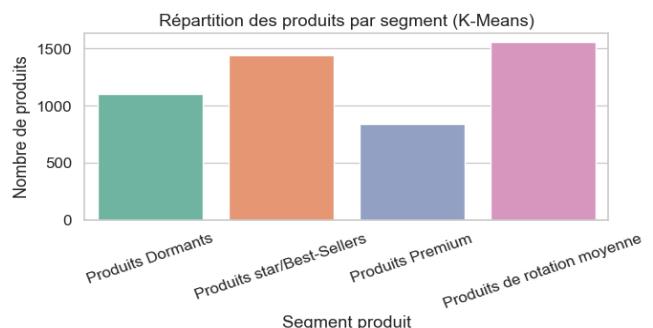
**Tableau 10-** Comparaison de K-Means et GMM (spherical)

Les indices Silhouette, Calinski–Harabasz et Davies–Bouldin montrent des performances similaires entre K-Means et GMM (spherical). K-Means présente une meilleure séparation (silhouette 0.33 vs 0.32) et une plus grande homogénéité intra-cluster, tandis que GMM modélise plus finement les zones de chevauchement. En somme, les deux modèles sont complémentaires, mais K-Means a été retenu pour sa stabilité, sa simplicité et sa robustesse sur des clusters sphériques.

#### 4. Segmentation finale des produits avec K-Means

Le modèle K-Means ( $K = 4$ ) a segmenté les produits en quatre groupes distincts selon leurs caractéristiques commerciales normalisées.

Segment produit	Taille	Part du total
<b>Produits Dormants</b>	1 101	19,6 %
<b>Produits Star / Best-Sellers</b>	1 439	25,6 %
<b>Produits Premium</b>	836	14,9 %
<b>Produits de rotation moyenne</b>	1 556	27,7 %



**Tableau 11-** Répartition des produits par segment

Le graphique montre une répartition équilibrée des segments produits : les best-sellers et produits de rotation moyenne dominent le portefeuille, les dormants (environ 20 %) offrent un potentiel d'optimisation, et les produits premium, plus rares, sont orientés valeur et marge plutôt que volume.

#### 5. Analyse qualitative du clustering

Le modèle K-Means ( $k=4$ ) a permis de distinguer quatre segments produits aux comportements commerciaux bien différenciés :

- ❖ **Produits Dormants (19,6 %)** : Faible volume et peu de ventes. Ces articles souffrent d'un manque d'attractivité et immobilisent du stock inutilement.
- ❖ **Produits Star / Best-Sellers (25,6 %)** : Forte demande et chiffre d'affaires élevé. Ils constituent le moteur principal des ventes et de la rentabilité.
- ❖ **Produits Premium (14,9 %)** : Articles à prix élevé mais à faible rotation. Ils ciblent une clientèle spécifique à forte valeur ajoutée.
- ❖ **Produits de rotation moyenne (27,7 %)** : Produits équilibrés en termes de ventes et de volume, garants d'une stabilité commerciale.

##### 5.1 Recommendations

Les **recommandations** visent à adapter la stratégie à chaque segment :

- ❖ **Dormants** : réduire les stocks, promouvoir ou retirer les articles non rentables.
- ❖ **Stars** : sécuriser l'approvisionnement, renforcer le marketing et développer des variantes.
- ❖ **Premium** : valoriser par des campagnes ciblées pour maximiser la marge et l'image
- ❖ **Rotation moyenne** : suivre la performance, ajuster les prix et repérer les futurs best-sellers.

#### V. Conclusion générale

Ce chapitre a permis d'analyser en profondeur le comportement des produits du dataset *Online Retail II*, en alliant rigueur technique et vision métier.

La détection d'anomalies a d'abord permis d'écartier les enregistrements atypiques (erreurs, retours, incohérences) afin d'assurer la fiabilité des données. Ensuite, la segmentation non supervisée (K-Means et GMM) a mis en évidence quatre profils produits — *stars*, *premium*, *rotation moyenne* et *dormants* — offrant une lecture claire du portefeuille et des pistes d'action stratégique sur les stocks, les prix et le marketing.

## Chapitre2 :

### Segmentation des clients par analyse RFM et détection d'anomalies

#### I. Introduction générale

La segmentation des clients est une étape essentielle dans la mise en place d'une stratégie de marketing ciblée et efficace. Elle permet de regrouper les clients en sous-groupes homogènes selon leurs comportements d'achat, afin de mieux comprendre leurs besoins et d'adapter les offres et les actions commerciales.

#### II. Préparation et traitement des données

Les données issues d'un fichier Excel regroupant l'historique des transactions clients ont été inspectées et harmonisées par un renommage standardisé des colonnes. L'analyse exploratoire a mis en évidence des valeurs manquantes, des incohérences de types, des retours produits, des doublons et la nécessité de définir la période couverte. Le nettoyage a permis de corriger ces problèmes en supprimant les lignes incomplètes, en traitant les retours, en éliminant les doublons et en uniformisant les libellés. Le jeu de données obtenu est propre et prêt pour le calcul des indicateurs RFM et la segmentation client.

Invoice	StockCode	Description	Quantity	InvoiceDate	Price	CustomerID	Country	IsReturn	LineAmount
0	489434	85048 15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	2009-12-01 07:45:00	6.95	13085.0	United Kingdom	False	83.4
1	489434	79323P PINK CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom	False	81.0
2	489434	79323W WHITE CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom	False	81.0
3	489434	22041 RECORD FRAME 7" SINGLE SIZE	48	2009-12-01 07:45:00	2.10	13085.0	United Kingdom	False	100.8
4	489434	21232 STRAWBERRY CERAMIC TRINKET BOX	24	2009-12-01 07:45:00	1.25	13085.0	United Kingdom	False	30.0

#### III. Analyse RFM (Récence, Fréquence, Montant)

##### 1. Calcul des métriques RFM :

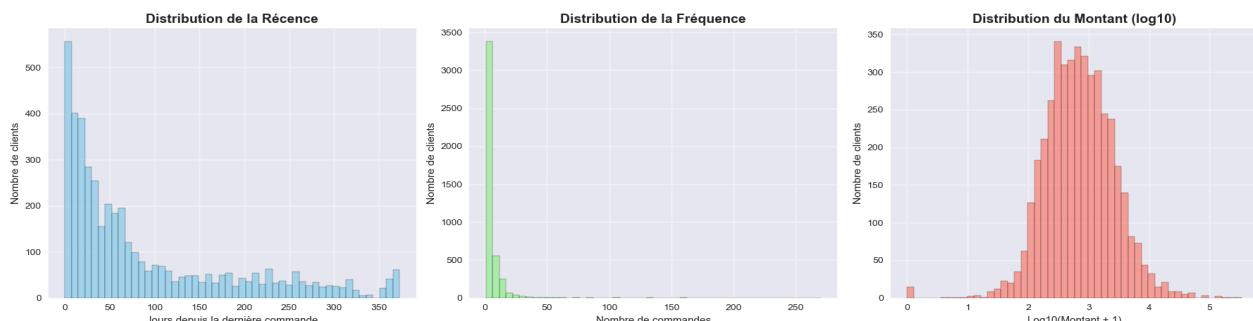
Pour chaque client, trois indicateurs comportementaux ont été calculés à partir des données de transactions nettoyées :

- **Récence (Recency)** : nombre de jours écoulés depuis la dernière commande du client.
- **Fréquence (Frequency)** : nombre total de commandes distinctes passées par le client
- **Montant (Monetary)** : somme totale dépensée par le client.

Ces métriques ont été agrégées au niveau du client en groupant les transactions par *CustomerID*. Ce calcul a permis de construire une table RFM contenant, pour chaque client, ses valeurs de récence, de fréquence et de montant, ainsi qu'un résumé statistique.

##### 2. Analyse exploratoire des métriques RFM :

Les distributions des trois composantes RFM sont représentées sur la figure ci-dessous :



➤ La distribution de la récence montre qu'un grand nombre de clients ont acheté très récemment, tandis qu'une longue traîne correspond à des clients inactifs depuis longtemps.

⇒ Plus la récence est faible, plus le client est fidèle et engagé.

➤ La distribution de la fréquence est très asymétrique : la majorité des clients n'ont commandé qu'une ou deux fois, alors qu'une minorité a passé de nombreuses commandes.

⇒ *Les bons clients sont rares mais essentiels, ce sont eux qui assurent la régularité du chiffre d'affaires.*

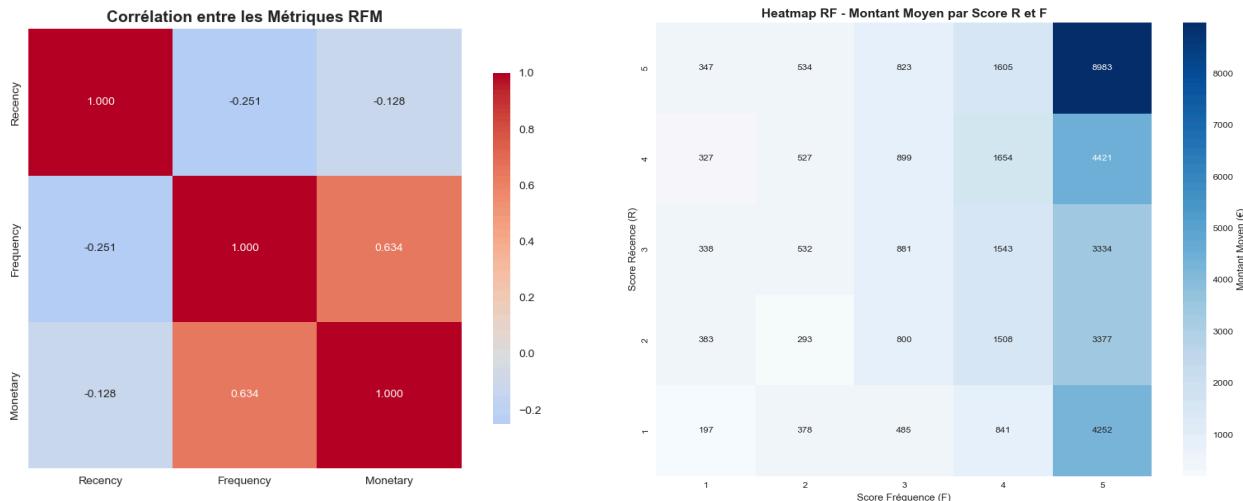
➤ La distribution du montant total (après log) devient plus symétrique. Cela révèle que la dépense varie fortement d'un client à l'autre, certains achètent énormément tandis que beaucoup dépensent peu.

⇒ *La valeur client est très concentrée sur une petite fraction de gros acheteurs.*

### 3. Corrélations entre les métriques RFM :

La matrice de corrélation met en évidence trois relations importantes :

- **Récence vs Fréquence** : corrélation négative modérée. Les clients qui achètent souvent ont tendance à être revenus récemment. Cela traduit une fidélité active.
- **Fréquence vs Montant** : corrélation positive relativement forte. Les clients les plus fréquents sont aussi ceux qui dépensent le plus au total.
- **Récence vs Montant** : corrélation légèrement négative. Les clients ayant dépensé beaucoup ont globalement acheté plus récemment, mais la relation est moins marquée. On observe donc aussi des gros clients qui ne sont plus revenus récemment.

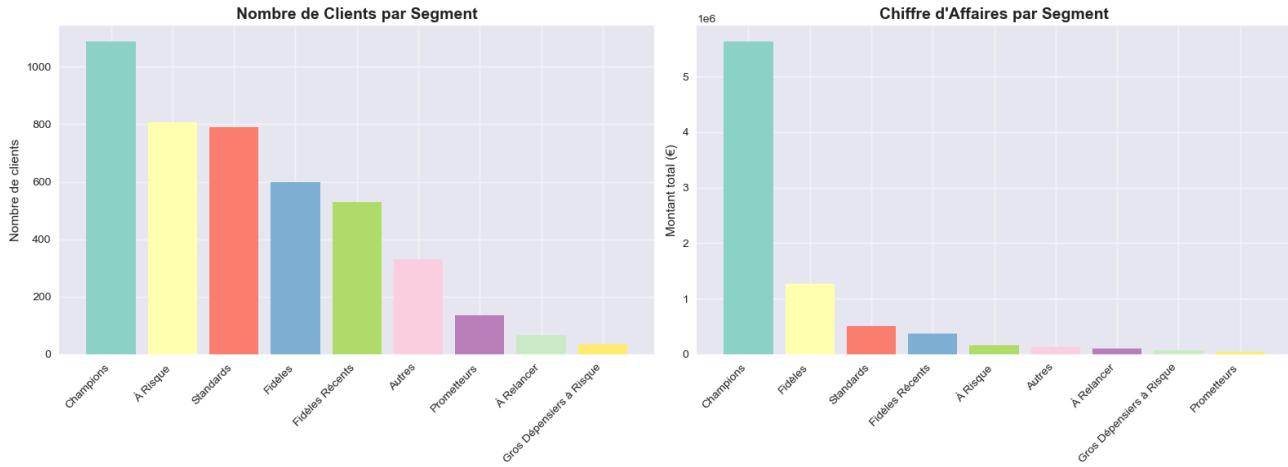


### 4. Scoring RFM et segmentation heuristique :

Pour rendre ces indicateurs exploitables en marketing, chaque client a été transformé en un profil RFM discret :

1. Les métriques continues R, F et M ont été converties en **scores sur une échelle de 1 à 5** à l'aide de coupures par quantiles :
  - **R\_Score** : score inversé de récence (un client récent reçoit un score élevé).
  - **F\_Score** : plus la fréquence est élevée, plus le score est élevé.
  - **M\_Score** : plus le montant total est élevé, plus le score est élevé.
2. Des scores combinés ont été dérivés :
  - **RF\_Score** : concaténation de R\_Score et F\_Score, utile pour capturer l'engagement.
  - **RFM\_Sum** : somme des trois scores, qui résume la valeur globale du client.
3. À partir des scores, une **segmentation marketing heuristique** a été appliquée. Chaque client a été affecté à un segment interprétable, par exemple :
  - **Champions** : clients très récents et très fréquents (forte activité, forte valeur).
  - **Fidèles Récents** : clients récents et réguliers, en phase d'engagement actif.
  - **Fidèles** : clients à forte fréquence mais moins récents, historiquement bons.
  - **Prometteurs** : nouveaux clients récents avec un potentiel de développement.
  - **Gros Dépensiers à Risque** : clients historiquement importants mais qui n'ont pas acheté récemment.
  - **À Relancer** : clients autrefois actifs (fréquence élevée) mais inactifs récemment.
  - **Standards / Autres** : clients moyens ou occasionnels.

Cette typologie servira ensuite de “vérité de référence” pour évaluer la cohérence des clusters obtenus par les algorithmes de segmentation non supervisée.



Un pairplot des scores R\_Score, F\_Score et M\_Score coloré par segment met en évidence la structure des segments dans l'espace RFM (voir figure ci-dessous) :

- Les *Champions* sont concentrés dans les zones de scores élevés sur les trois axes, ce qui confirme leur forte activité récente, leur fréquence d'achat élevée et leur contribution financière.
- Les segments *À Risque* et *Gros Dépenseurs à Risque* se situent dans des zones à faible récence (ils ne sont pas revenus récemment) et parfois faible fréquence actuelle, suggérant une perte potentielle de clients de grande valeur.
- Les *Fidèles* et *Prometteurs* occupent des régions intermédiaires, cohérentes avec des clients actifs mais à des stades différents du cycle de vie.



## IV. Méthodes de Clustering

### 1. Préparation des données :

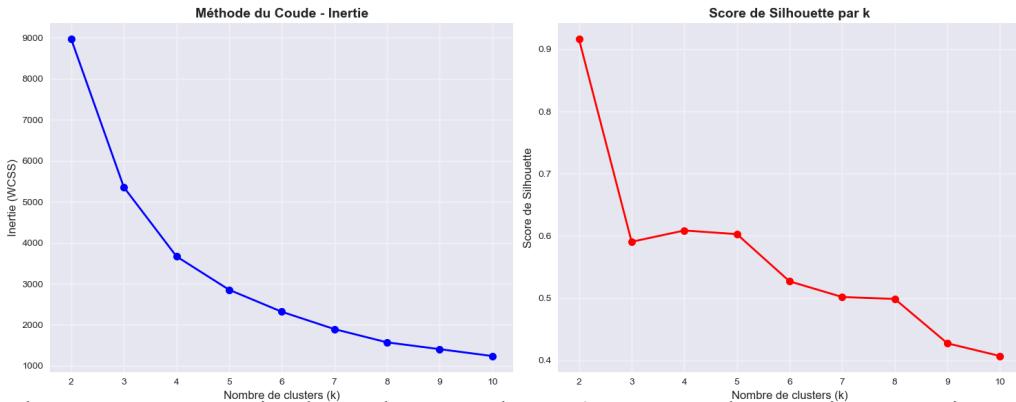
- Avant d'appliquer les algorithmes de clustering, les trois variables (**R**, **F** et **M**) présentent en effet des ordres de grandeur différents, ce qui impose une **normalisation**.
- Les valeurs ont été centrées et réduites à l'aide du **StandardScaler** (moyenne 0, écart-type 1), produisant trois variables normalisées : *Recency\_scaled*, *Frequency\_scaled* et *Monetary\_scaled*.
- Cette étape assure que chaque composante contribue équitablement aux distances calculées par les algorithmes.

### 2. Clustering par K-Means :

#### a- Sélection du nombre optimal de clusters :

Pour déterminer le nombre optimal de clusters (k), deux approches ont été employées :

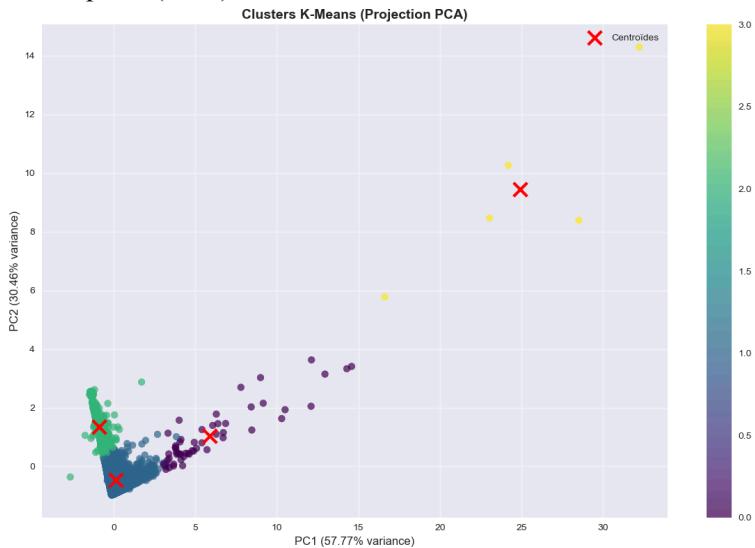
- la **méthode du coude**, qui observe la diminution de l'inertie intra-cluster (WCSS).
- le **score de silhouette**, qui mesure la compacité et la séparation entre les clusters.



Les résultats montrent un point de coude autour de  $k = 4$ , correspondant aussi à un maximum local du score de silhouette ( $\sim 0.6$ ). Ce compromis a été retenu comme configuration finale pour K-Means.

#### b- K-Means Final avec $k$ Optimal ( $k = 4$ ) :

- Les clusters ont ensuite été projetés sur un plan bidimensionnel à l'aide d'une Analyse en Composantes Principales (PCA) afin d'en faciliter la visualisation.



- Le graphique PCA montre que les clients se répartissent en quatre groupes bien distincts :
  - un grand ensemble de clients moyens/actifs (dominant).
  - un groupe de clients fidèles légèrement différent.
  - un groupe de clients dormants plus dispersés.
  - un petit groupe VIP très éloigné du reste, représentant les clients les plus précieux.

Cette séparation nette confirme la pertinence du modèle K-Means, puisque les clusters sont visuellement bien différenciés et les centroïdes isolés les uns des autres.

#### c- Évaluation des performances du modèle K-Means :

La qualité du partitionnement obtenu avec **K-Means** a été évaluée à l'aide de trois métriques internes : **Silhouette** mesure la cohésion et la séparation des clusters. Plus elle est élevée, meilleure est la segmentation, **Calinski-Harabasz** compare la dispersion entre et au sein des clusters. Une valeur élevée indique des groupes bien distincts, et **Davies-Bouldin** mesure la similarité moyenne entre clusters. Plus le score est faible, plus les clusters sont compacts et séparés.

Les valeurs obtenues pour le modèle K-Means (avec  $k = 4$ ) sont les suivantes :

- **Silhouette (0.609)** : Les clusters sont globalement bien séparés, avec peu de recouvrement entre groupes.
- **Calinski-Harabasz(3768.1)** : Forte dispersion entre les clusters, ce qui confirme une bonne séparation structurelle.
- **Davies-Bouldin (0.607)** : Les clusters sont compacts et peu similaires entre eux, signe d'une segmentation nette

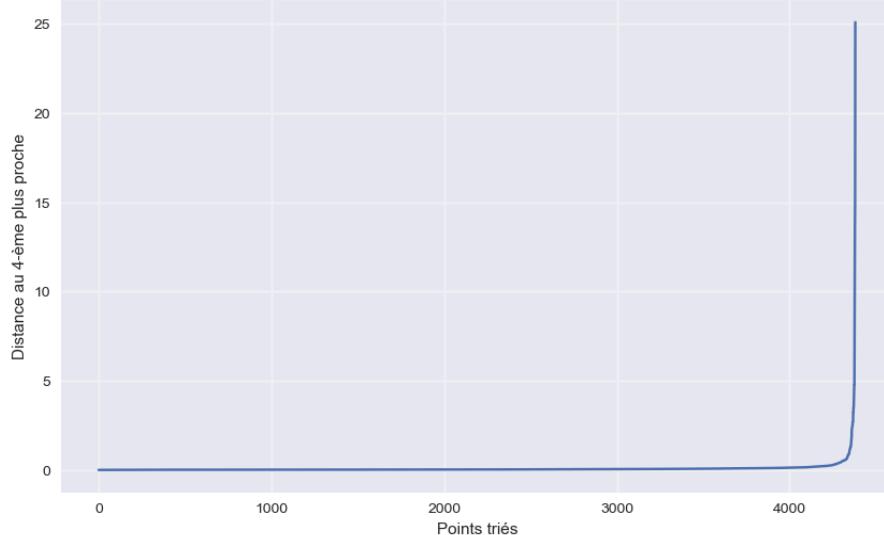
### 3. Clustering par DBSCAN:

#### a- Sélection des paramètres :

DBSCAN nécessite deux hyperparamètres : **eps** (rayon de voisinage), et **min\_samples** (nombre minimal de points pour former un cluster).

Le choix du paramètre **eps** s'appuie sur la méthode du k-distance plot, où le point d'infexion indique le seuil idéal de densité

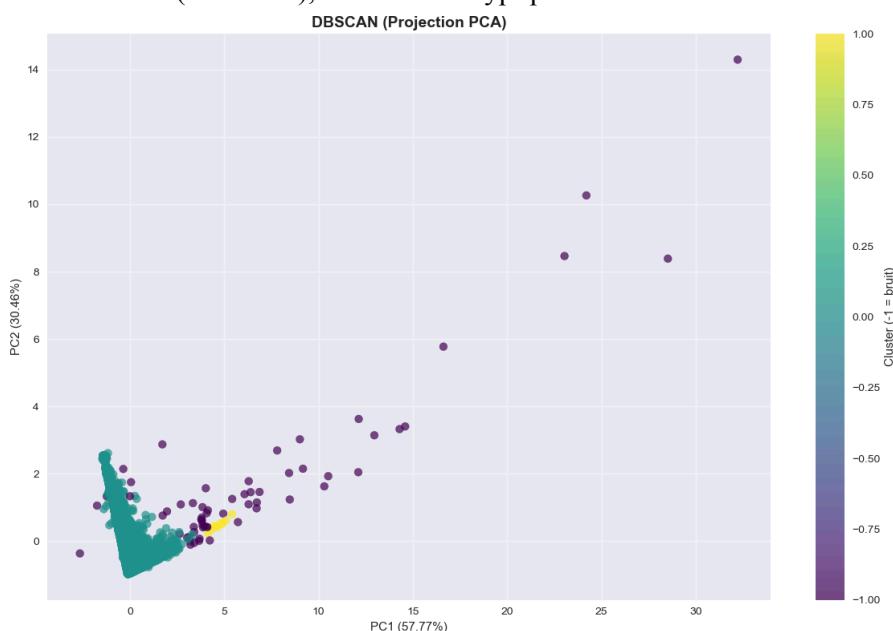
k-distance plot (k=4)



- La courbe montre une rupture nette suggérant une valeur de  $\text{eps} \approx 0.5$ . Ce paramètre a été retenu avec  $\text{min\_samples} = 4$  (règle pratique : dimension + 1).

#### b- DBSCAN avec **eps** et **min\_samples** Optimal :

- L'algorithme **DBSCAN** a permis d'identifier un nombre limité de clusters, tout en classant une partie des clients comme "bruit" (label = -1), c'est-à-dire atypiques ou isolés.



- La projection PCA montre une majorité de points appartenant à un cluster principal, avec quelques petits groupes dispersés et plusieurs points isolés. DBSCAN s'avère donc utile pour **identifier les comportements anormaux** (clients atypiques ou extrêmes), mais il segmente moins finement les comportements réguliers que K-Means.

#### c- Évaluation des performances du modèle DBSCAN :

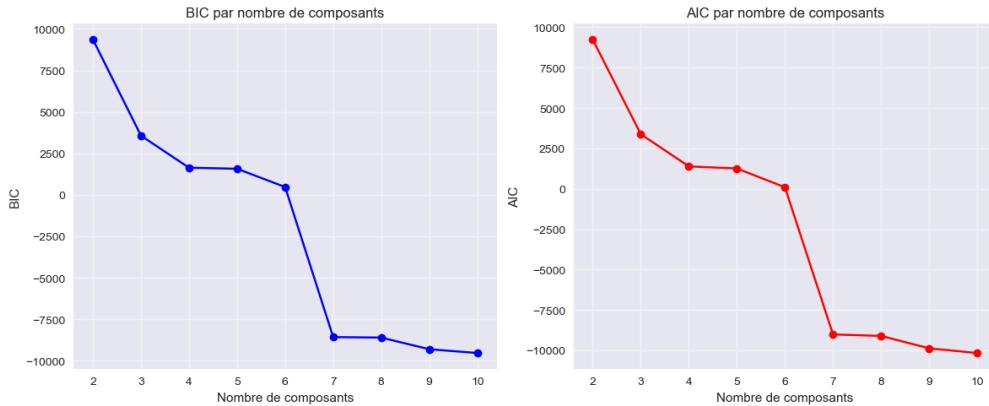
- Silhouette: 0.7486851247375772
- Calinski-Harabasz: 190.03733752371986
- Davies-Bouldin: 0.25909975409629254

- Les scores obtenus confirment ces observations : Les métriques disent que les clusters que DBSCAN accepte sont très propres et bien distincts. Mais DBSCAN couvre moins de monde : il est plus sélectif.

#### 4. Clustering par Modèle de Mélange Gaussien (GMM) :

##### a- Sélection du nombre de composants :

- Le Modèle de Mélange Gaussien (GMM) suppose que les données proviennent d'un mélange de plusieurs distributions gaussiennes. Le nombre de composantes (ou clusters) à modéliser doit donc être choisi avec soin.
- Pour cela, deux critères d'information ont été utilisés :
  - **BIC** : mesure la qualité d'un modèle tout en pénalisant sa complexité. Plus l'AIC est bas, meilleur est le compromis entre précision et simplicité.
  - **AIC** : même idée que AIC, mais pénalise plus fortement les modèles trop complexes. Plus le BIC est bas, meilleur est le modèle.
- Le nombre optimal de composantes correspond au point où ces valeurs atteignent leur minimum.



- L'analyse montre une diminution progressive des valeurs du BIC et de l'AIC jusqu'à environ **10 composantes**, ce qui a été retenu comme **nombre optimal** pour le modèle GMM.

##### b- GMM Optimal (10 composantes) :

- Chaque client a été assigné à un cluster selon la probabilité d'appartenance la plus élevée.
- La distribution des effectifs par cluster met en évidence des groupes de tailles variables, certains très denses et d'autres beaucoup plus restreints, révélant des sous-populations spécifiques dans les comportements clients.



- La projection PCA montre une structure complexe : plusieurs petits groupes coexistent, traduisant la capacité du GMM à capturer des **variations fines de densité** au sein des données, contrairement à K-Means qui impose des frontières nettes.
- Cette approche probabiliste permet donc d'identifier des segments plus nuancés, potentiellement liés à des comportements intermédiaires ou transitoires.

### c- Évaluation des performances du modèle GMM:

- Silhouette: -0.014
- Calinski-Harabasz: 1046.78
- Davies-Bouldin: 1.995

Les résultats indiquent des valeurs légèrement inférieures à celles de K-Means, ce qui suggère que le GMM, bien que plus flexible, produit des frontières plus floues entre clusters.

Cependant, il reste performant pour modéliser des comportements clients complexes où les appartenances peuvent se chevaucher.

### 5. Comparaison des méthodes de clustering

- **DBSCAN** obtient le meilleur score de silhouette et le plus faible indice Davies–Bouldin. Cela signifie que les groupes “denses” qu’il détecte sont très cohérents et bien séparés des autres. En pratique, DBSCAN isole très clairement certains comportements typiques et rejette le reste comme bruit.
- **K-Means** obtient une bonne séparation globale (Calinski–Harabasz élevé) et des clusters compacts. Cela confirme que la structure en 4 groupes trouvée par K-Means est stable et exploitable pour une segmentation marketing classique.
- **GMM** est moins performant sur ces métriques internes. Les clusters se chevauchent plus, ce qui est logique car GMM autorise des appartenances probabilistes et des frontières floues.

### 6. Conclusion

L’analyse RFM a d’abord mis en évidence des tendances fortes : une majorité de clients récents et peu fréquents, contrastant avec une minorité de clients très fidèles et à forte valeur.

Trois méthodes de clustering ont ensuite été comparées :

- **K-Means**, qui a produit une segmentation claire et équilibrée en quatre groupes principaux.
- **DBSCAN**, qui s’est révélé particulièrement performant pour isoler les comportements atypiques et identifier des clients “bruit” ou marginaux ;
- **GMM**, qui a offert une représentation probabiliste plus fine, permettant de capturer des sous-groupes intermédiaires et de mieux reproduire la logique des segments RFM.

Les résultats montrent que **K-Means** constitue le meilleur compromis entre lisibilité et performance, tandis que **DBSCAN** complète efficacement l’analyse en détectant les anomalies, et que **GMM** fournit une compréhension plus nuancée des comportements clients.

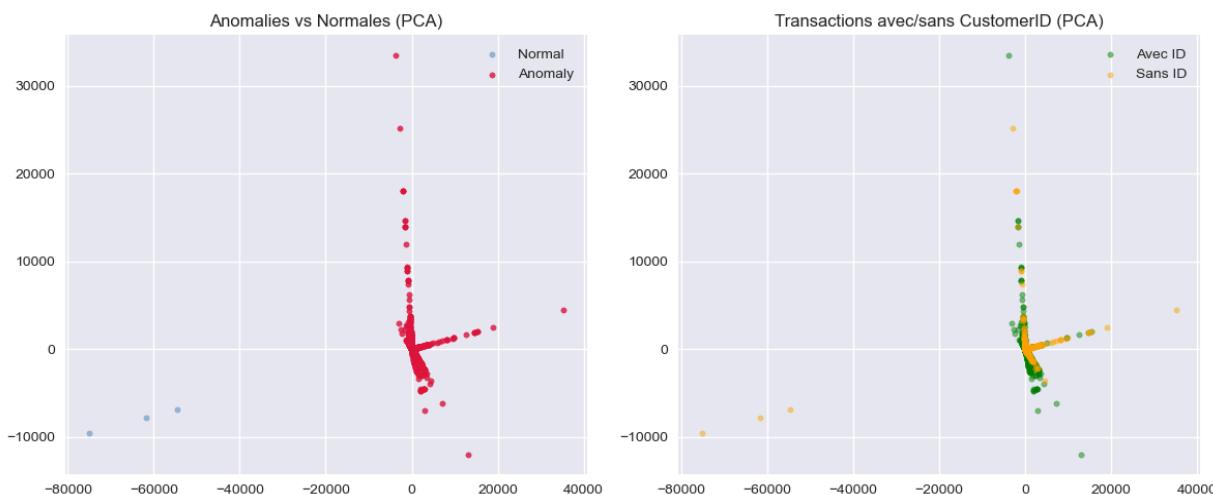
## V. Détection d’anomalies

### A. ISOLATION FOREST

L’algorithme Isolation Forest a été utilisé pour détecter les transactions atypiques dans le jeu de données nettoyé. Il repose sur le principe d’isolation : les observations inhabituelles sont isolées plus rapidement que les points normaux lorsqu’on partitionne l’espace des données.

#### 1. Mise en œuvre et identification des anomalies :

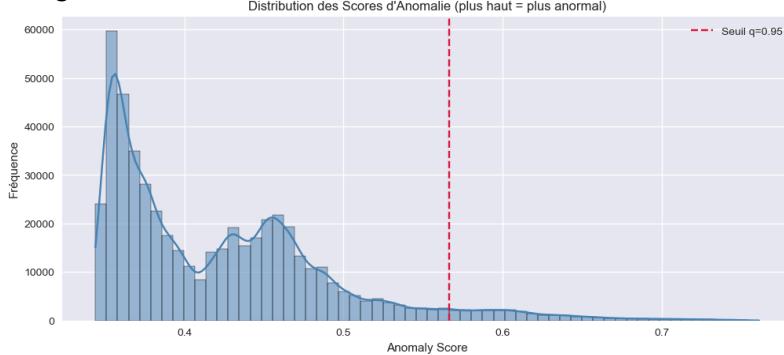
- Le modèle a été entraîné avec un taux de contamination de 8 %, correspondant à la proportion attendue d’anomalies dans l’échantillon.



- Sur la projection PCA, les **points rouges** représentent les anomalies identifiées, concentrées autour de zones extrêmes du nuage principal.
- On observe que la plupart des anomalies se distinguent nettement des transactions normales.
- Au total, **environ 8 % des transactions ont été détectées comme anormales**, confirmant une distribution réaliste pour ce type d'analyse.

#### 2. Distribution des scores et seuil de décision:

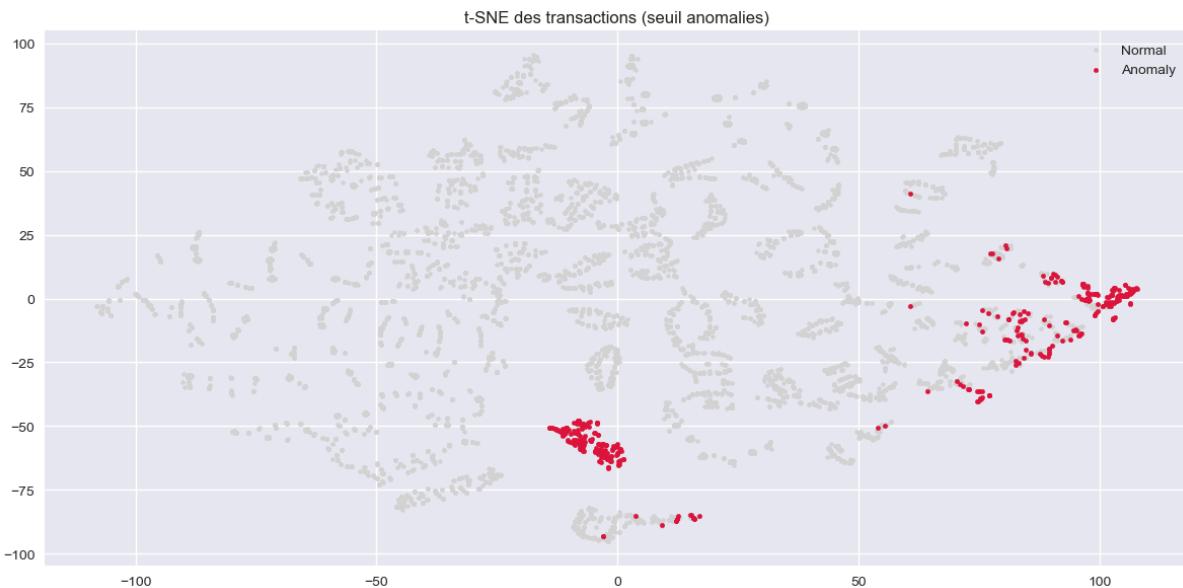
- Pour raffiner la détection, les scores d'anomalie ont été analysés afin de définir un seuil optimal basé sur les quantiles (ici **q = 0,95**). Les transactions dont le score dépasse ce seuil sont considérées comme des anomalies significatives.



- La figure montre une distribution asymétrique : la majorité des transactions ont un score faible, tandis qu'une minorité présente des valeurs très élevées, typiques de comportements rares ou extrêmes.
- Le seuil choisi (ligne rouge) isole environ **5 % des cas les plus anormaux**.

#### 3. Représentation t-SNE et distribution des anomalies:

- Pour mieux visualiser la séparation entre transactions normales et anormales, une réduction de dimension par **t-SNE** a été effectuée.
- Cette méthode met en évidence la structure locale des données.



- Les **points rouges** identifient des groupes distincts d'anomalies répartis dans l'espace.
- Certains forment des sous-groupes cohérents, indiquant des types spécifiques d'anomalies (par exemple, de très grosses commandes ou des retours massifs).
- D'autres points isolés peuvent correspondre à des erreurs individuelles.

#### 4. Interprétation et conclusion :

L'application de **Isolation Forest** a permis d'isoler efficacement les transactions atypiques.

Les résultats montrent que :

- Les anomalies concernent principalement des **quantités ou montants extrêmes**,
- Beaucoup d'entre elles sont liées à des **retours** ou à des **commandes sans identifiant client**,
- Certaines zones géographiques concentrent davantage ces comportements.

Sur le plan métier, ces anomalies peuvent refléter :

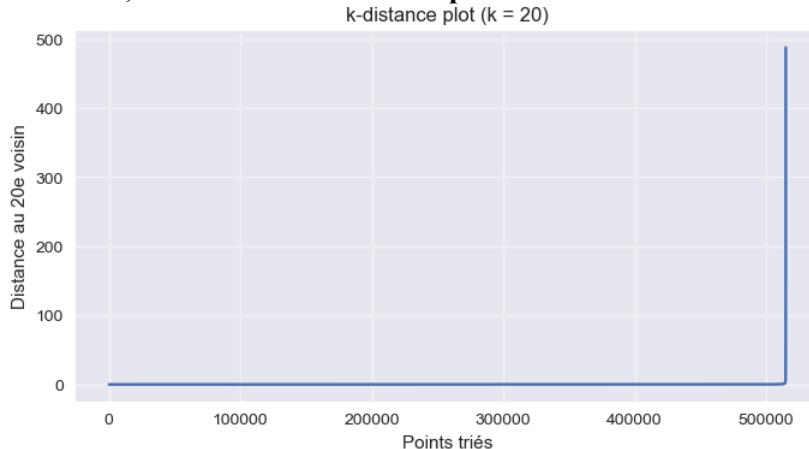
- des erreurs de facturation ou d'enregistrement,
- des clients irréguliers à fort volume, à surveiller pour éviter les pertes,
- ou des signaux précoce de comportements à risque ou de fraude.

## B. DBSCAN

- L'algorithme **DBSCAN** repose sur le principe de densité locale pour regrouper les points : les zones de forte densité sont considérées comme des clusters, tandis que les points isolés, éloignés de tout groupe dense, sont qualifiés de bruit.
- Cette approche est particulièrement adaptée à la détection d'anomalies, car elle ne suppose pas de distribution spécifique des données et identifie naturellement les comportements rares.

### 1. Choix des paramètres et estimation du rayon de voisinage :

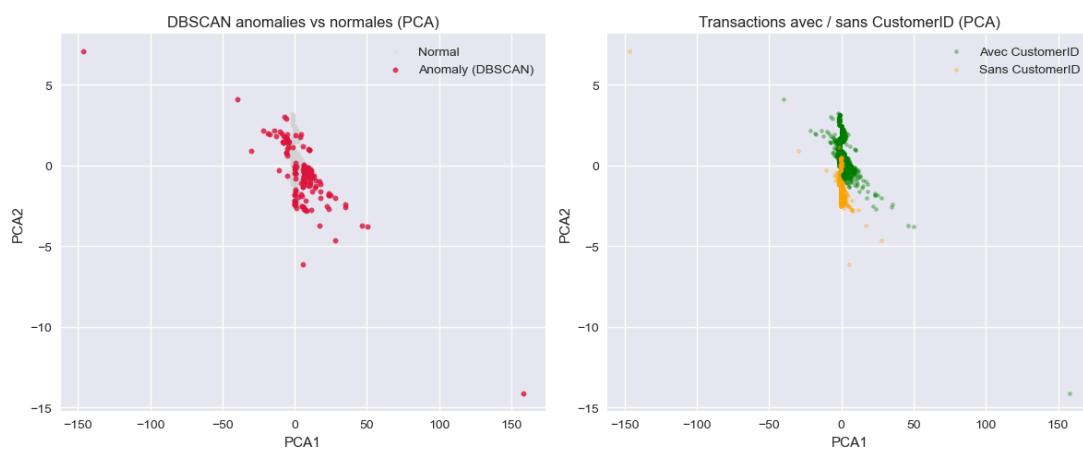
- Le paramètre principal de DBSCAN est le rayon de voisinage (*eps*), qui définit la distance maximale entre deux points pour qu'ils appartiennent à la même région dense.
- Pour le déterminer, la méthode du **k-distance plot** a été utilisée.



- Dans ce cas, le coude observé suggère un **eps** ≈ 2.0 avec **min\_samples** = 20, garantissant un compromis entre sensibilité et robustesse.

### 2. Application du modèle et visualisation PCA :

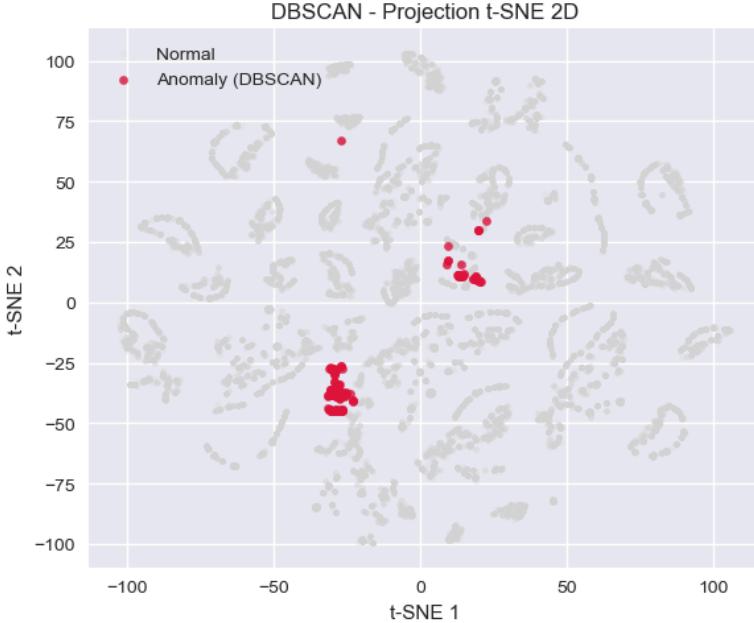
- DBSCAN a été appliqué sur un échantillon de 50 000 transactions normalisées.
- Chaque point a reçu un label de cluster (entier positif) ou -1 pour les anomalies (bruit).
- La proportion de points étiquetés "bruit" correspond directement au **taux d'anomalies détectées**.



- La projection en deux dimensions (PCA) met en évidence les **points rouges** considérés comme anomalies.
- Ces observations se situent en périphérie du nuage principal, loin des régions denses correspondant aux transactions normales (gris clair).
- Le graphique de droite montre que certaines de ces anomalies concernent davantage des **transactions sans identifiant client**, souvent sources de comportements atypiques (retours, erreurs ou commandes incomplètes).
- Le modèle DBSCAN a détecté un taux d'anomalies modéré, cohérent avec les résultats d'Isolation Forest.

### 3. Visualisation avancée avec t-SNE :

- Afin d'obtenir une représentation plus fidèle des structures locales et des zones de densité, une réduction de dimension t-SNE a été appliquée sur les mêmes données.
- Cette méthode permet de mieux visualiser les regroupements complexes et la répartition des anomalies.



- La figure montre clairement que les anomalies (points rouges) se trouvent **en marge des zones normales** (gris clair).
- Elles apparaissent dans plusieurs sous-régions distinctes, ce qui confirme que les comportements atypiques ne se limitent pas à un seul type de transaction, mais couvrent différents profils :
  - transactions isolées à très fortes valeurs,
  - commandes incomplètes ou sans identifiant client,
  - ou encore des anomalies liées à des erreurs de saisie.

### 4. Interprétation générale :

L'application de DBSCAN à la détection d'anomalies a permis de :

- repérer automatiquement des **zones de faible densité** correspondant à des comportements rares
- identifier des **transactions hors du comportement standard**,
- et confirmer la cohérence avec les résultats d'**Isolation Forest** (les deux méthodes repèrent souvent les mêmes points atypiques).

Sur le plan métier, ces anomalies peuvent être interprétées comme :

- des **transactions inhabituelles mais légitimes** (gros volumes ou remises exceptionnelles),
- des **données erronées** (quantités négatives, prix aberrants, absence de client ID),
- ou des **signaux faibles de fraude ou de risque client**.

## **VI. Conclusion Générale**

Cette partie du projet visait à segmenter les clients selon leurs comportements d'achat et à détecter les transactions atypiques. L'analyse **RFM** a permis d'établir une segmentation de référence fondée sur la récence, la fréquence et le montant des achats. Différents algorithmes de clustering (**K-Means**, **DBSCAN**, **GMM**) ont ensuite été appliqués : K-Means offre une segmentation claire, GMM capte les comportements intermédiaires et DBSCAN repère les profils marginaux. En parallèle, **Isolation Forest** et **DBSCAN** ont permis d'identifier des anomalies liées à des montants, quantités ou retours inhabituels, révélant ainsi à la fois des signaux de risque et des opportunités commerciales.

## Chapitre 3: Analyse temporelle & clustering des ventes

### I. Agrégation temporelle (séries mensuelles):

Pour obtenir une vue d'ensemble de l'activité, on a regroupé les ventes par mois et calculé plusieurs indicateurs : le revenu total, la quantité vendue, le nombre de factures, le nombre de clients distincts et le panier moyen.

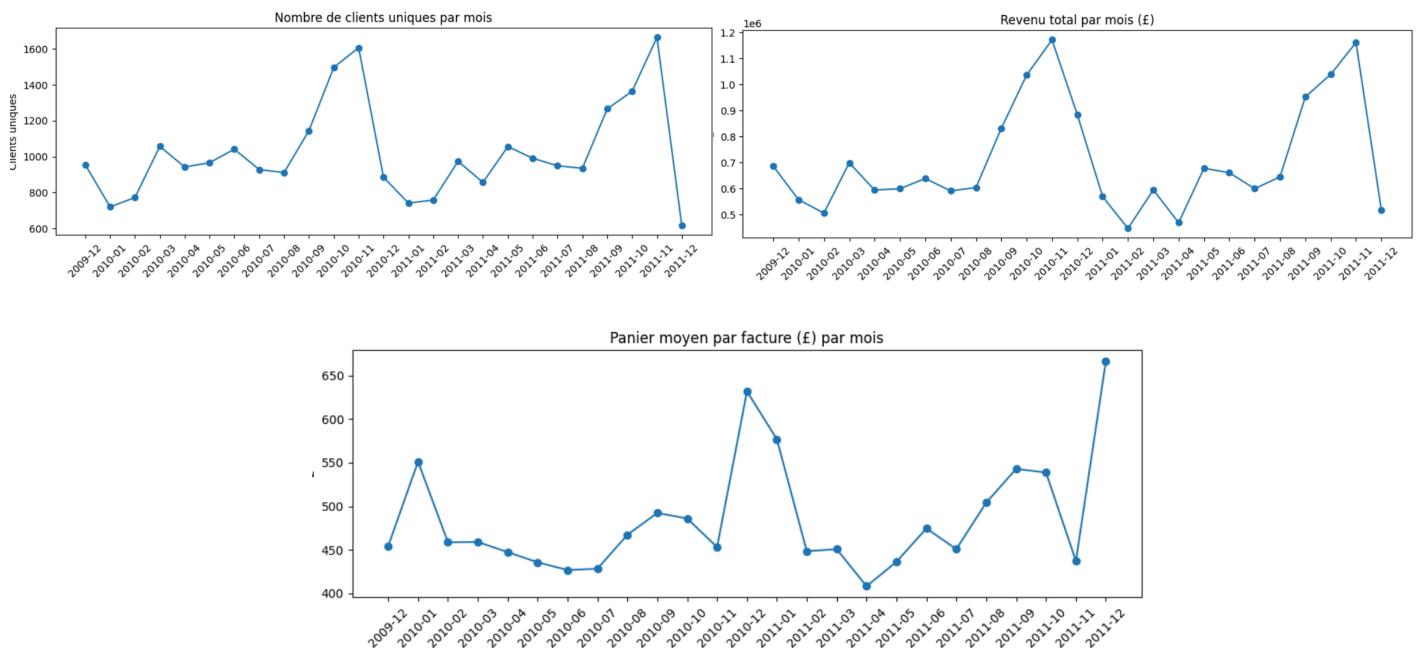
InvoiceMonth	revenue	quantity	nb_invoices	nb_customers	avg_basket
2009-12	686654.160	400153	1512	955	454.136349
2010-01	557319.062	370921	1011	720	551.255254
2010-02	506371.066	372761	1104	772	458.669444
2010-03	699608.991	503466	1524	1057	459.061018
2010-04	594609.192	352025	1329	942	447.410980

En agrégeant les transactions au niveau mensuel, nous observons une forte saisonnalité des ventes. Les revenus atteignent des pics majeurs en octobre-novembre 2010 puis à nouveau en octobre-novembre 2011, suivis d'un ralentissement en début d'année.

Le nombre de clients uniques augmente sur ces mêmes périodes. Cela correspond à une dynamique d'achat saisonnière (cadeaux, décorations de Noël,...).

En parallèle, le panier moyen par facture dépasse parfois 600€ sur certains mois, ce qui indique l'existence de gros paniers (probablement clients revendeurs / grossistes). L'entreprise semble donc reposer sur deux segments distincts :

- un socle de gros acheteurs avec des paniers très élevés,
- une vague saisonnière d'acheteurs



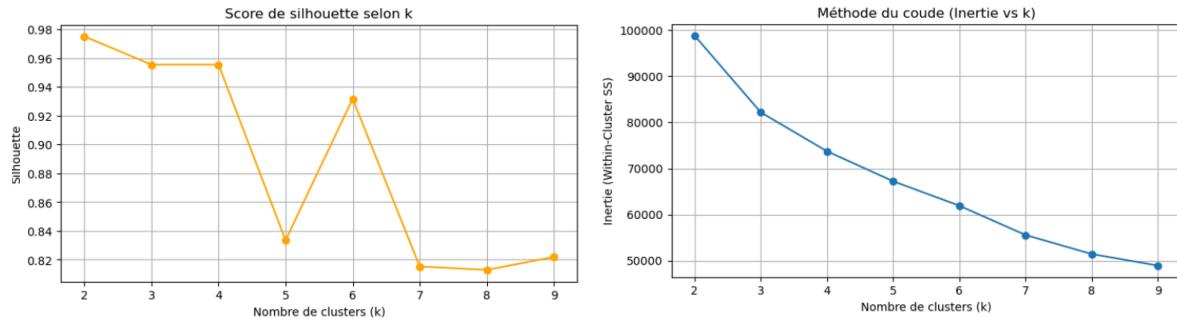


- **Normalisation des données** : on standardise chaque feature (chaque colonne/mois) pour donner à tous les mois le même poids.

### 3. Choix du nombre optimal de clusters (KMeans)

L'objectif ici est de déterminer le nombre optimal de groupes de clients ( $k$ ) à l'aide de l'algorithme KMeans, en mesurant la qualité de la segmentation selon deux critères :

- Inertie intra-cluster (méthode du coude) : mesure la compacité des clusters.
- Score de silhouette : évalue la séparation moyenne entre clusters.



Après comparaison des valeurs de  $k$  entre 2 et 9, nous retenons  $k = 4$ .  $k = 4$  offre encore une excellente séparation entre groupes (silhouette  $\approx 0.96$ ) tout en permettant une segmentation plus riche que  $k = 2$  (qui sépare seulement "gros clients" vs "autres"). Au-delà de  $k = 4$ , l'amélioration d'inertie devient marginale et la silhouette se dégrade, ce qui suggère de la sur-segmentation.

### 4. Comparaison de trois méthodes de clustering

Kmeans Avant transformation log (k=4)	Clustering Apres transformation log (k=2)	Clustering hiérarchique agglomératif, Après transformation log (ward)
cluster 0 taille =5861 cluster 1 taille =1 cluster 2 taille =15 cluster 3 taille =1	cluster 0 taille =816 cluster 1 taille =5062	cluster 0 taille=1042 cluster 1 taille=4836

*Interprétation technique :*

Après avoir construit des profils clients mensuels, on a appliqué plusieurs approches de clustering.

- ❖ **KMeans sans transformation** : L'algorithme KMeans repose sur la **distance euclidienne** et cherche à minimiser la variance interne des clusters.

Lorsqu'il est appliqué directement sur les montants de dépense , la forte **asymétrie des valeurs** (quelques clients dépensent des milliers de £, la majorité quelques dizaines) entraîne :

un **effet d'échelle** : les clients les plus dépensiers dominent la variance totale,

une **structure non sphérique** : les distances entre clients sont écrasées par les outliers.

Résultat : KMeans forme un grand cluster “moyen” contenant presque tout le monde, et plusieurs micro-clusters isolant les clients extrêmes.

- ❖ **KMeans après transformation  $\log(1 + \text{dépense})$**  : L'application de  $\log(1 + \text{dépense})$  réduit la dispersion extrême et **rétablit des distances comparables** entre clients.

En supprimant l'effet d'échelle, KMeans capture désormais la **forme des séries temporelles** plutôt que leur amplitude.

Après standardisation, chaque client est traité comme un vecteur de 25 points de dépense “centré-réduit”.

Résultat : Deux groupes principaux apparaissent (~816 vs ~5062).

- ❖ **Clustering hiérarchique agglomératif (Ward), sur features log-transformées**: Le clustering hiérarchique, notamment avec la **méthode de Ward**, construit des groupes en fusionnant successivement les plus proches selon la variance intra-cluster.

Sur des données déjà log-transformées et standardisées Ward retrouve la même structure que KMeans, ce qui confirme la **stabilité topologique** de la séparation à deux groupes.

### *Interpretation business:*



#### ❖ KMeans sans transformation :

**Cluster 0 (bleu)** : la courbe est très basse et plate (autour de ~0 à quelques dizaines seulement) mais leur contribution au chiffre d'affaires est quasiment toujours entre 70% et 90% du chiffre d'affaires total du mois. Ce sont les clients "classiques", beaucoup de petits achats, pas de gros pics, présents presque tout le temps.

**Cluster 2 (vert)** : dépenses intermédiaires , répétées plusieurs mois, avec des petits pics récurrents, pas juste un one-shot. Leur contribution au chiffre d'affaires monte parfois à 15-20%. Ce sont des clients importants, qui achètent en gros, de manière répétée.

**Cluster 1 (orange)** : montant moyen par client extrêmement élevé, pas tous les mois mais plutôt quelques pics.

On peut déduire que c'est un méga-clients qui passent des commandes énormes à certains moments.

**Cluster 3 (rouge)** : grosses dépenses aussi (10k-40k £) On peut dire que c'est des clients "gros mais irréguliers". Leur contribution au chiffre d'affaires reste en dessous de ~10% sauf quelques pics, parfois proches de zéro.

on conclut que cet algorithme n'a pas réussi à capturer les clients saisons des autres clients régulier, c'était plutôt orienté détection outliers

#### ❖ KMeans avec transformation log :

Avec cette représentation log-transformée, nous obtenons deux groupes de taille réaliste : environ 816 clients (cluster 0) et 5062 clients (cluster 1)

**Cluster 0** : clients à forte dépense moyenne par mois, présents sur de nombreux mois. Ce cluster domine le chiffre d'affaires hors saison (environ 60–75% du CA mensuel). On peut déduire que ce sont des clients réguliers à haute valeur, probablement des acheteurs B2B / revendeurs.

**Cluster 1** : clients à faible dépense moyenne (30–100 £), dont l'activité se concentre fortement autour d'octobre-novembre-décembre. En fin d'année, ce cluster représente jusqu'à ~50-60% du chiffre d'affaires mensuel. On peut le considérer comme groupe des acheteurs occasionnels attirés par la période de Noël (cadeaux, décorations).

#### ❖ Clustering hiérarchique agglomératif (Ward), sur features log-transformées:

Les profils temporels moyens par cluster montrent à nouveau deux comportements :

**Cluster 0** : de clients à forte dépense mensuelle, actifs quasiment toute l'année, qui contribue surtout au chiffre d'affaires hors saison ;

**Cluster 1** : deq clients à faible dépense individuelle mais très nombreux, qui devient dominant pendant les mois d'octobre à décembre.

Ces résultats confirment ceux obtenus avec KMeans (après transformation logarithmique) : notre base de clients se compose de deux segments stratégiques : des clients réguliers à forte valeur et des clients saisonniers Noël en volume.

## 5. Comparaison des clustering DTW et KShape :

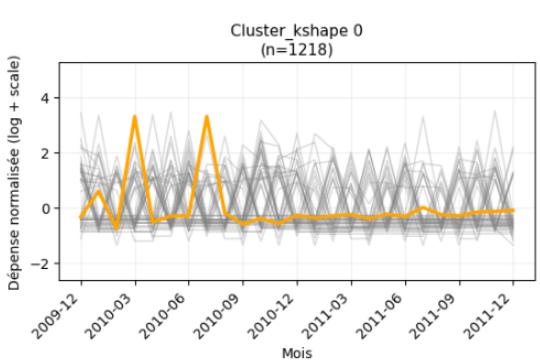
Les premiers clusters (KMeans brut, KMeans log, Agglomératif Ward) utilisent tous la même logique de base : on prend chaque client comme un vecteur de 25 valeurs (dépense mensuelle), et on mesure la distance entre ces vecteurs de manière standard (euclidienne).

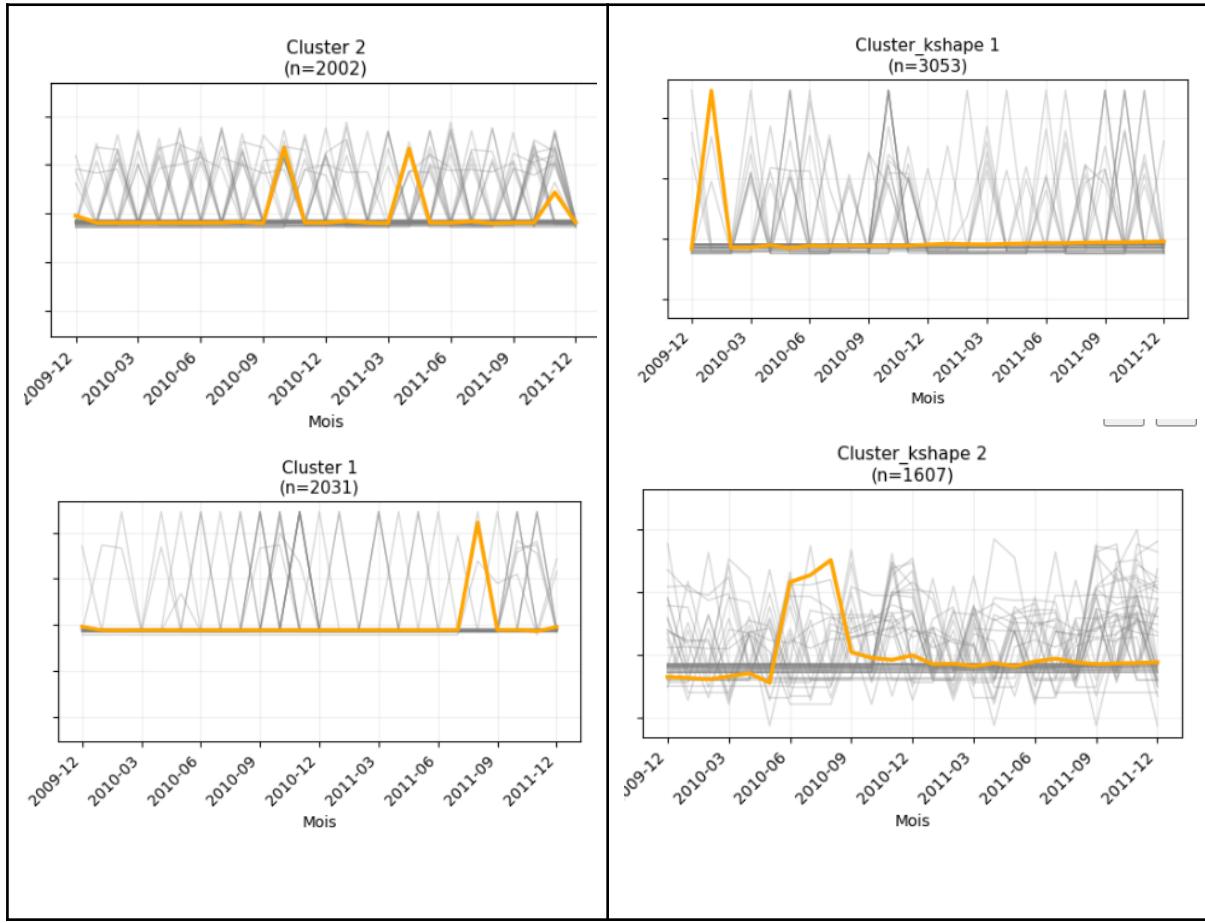
Problème technique :

- La distance euclidienne suppose que deux clients sont “semblables” si, mois par mois, ils dépensent les mêmes montants aux mêmes dates.
- Donc si Client A achète surtout en novembre et Client B achète surtout en décembre, ils seront considérés différents, même si leur comportement est identique à un simple décalage temporel près.
- Or, du point de vue business, ces deux profils sont équivalents : ce sont des acheteurs saisonniers de fin d'année, simplement déphasés dans le temps.
- Le profil d'évolution temporelle est donc mal capturé par les approches euclidiennes classiques.

Pour surmonter cette limite, nous avons introduit deux algorithmes spécialement conçus pour les données temporelles séquentielles : DTW (Dynamic Time Warping) et KShape.

- Ils ne comparent plus les clients point par point, mais cherchent à identifier des formes similaires dans le temps.
- Ils se concentrent sur la dynamique (la manière dont les dépenses évoluent dans l'année) plutôt que sur le niveau absolu des montants dépensés.

DTW	KShape
Taille des clusters (DTW) : cluster_dtw 0 taille = 1845 cluster_dtw 1 taille = 2031 cluster_dtw 2 taille = 2002	Taille des clusters (Kshape) : cluster_Kshape 0 taille = 1218 cluster_Kshape 1 taille = 3053 cluster_Kshape 2 taille = 1607
	



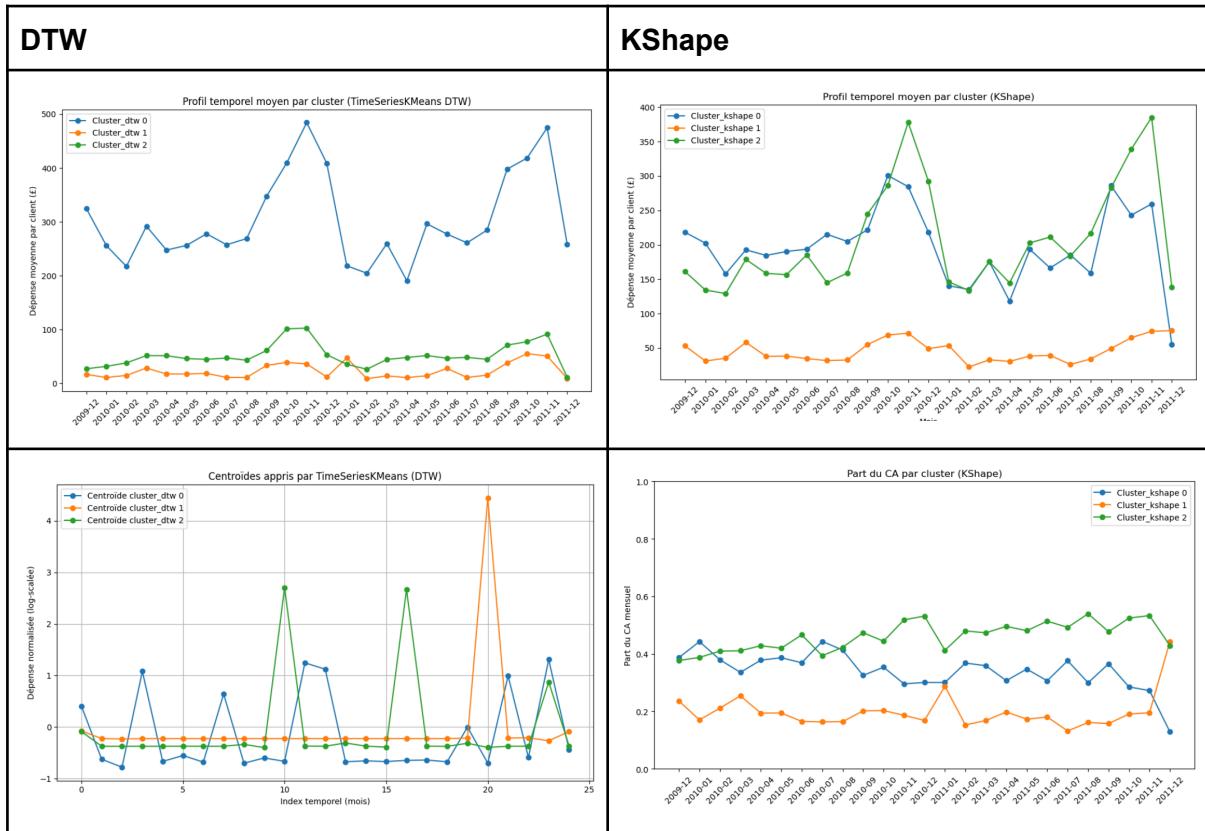
### **Interprétation technique :**

Les deux méthodes produisent trois clusters interprétables, mais elles ne mettent pas exactement les mêmes profils.

- **DTW (Dynamic Time Warping):** aligne les séries temporelles dans le temps. Deux clients sont considérés comme similaires même si leurs pics d'achat ne tombent pas exactement le même mois.
- **KShape :** compare les séries sans réaligner le temps, mais en normalisant l'amplitude et en utilisant une corrélation de forme.

KShape segmente plutôt selon le calendrier réel plutôt que DTW reste plus flexible, aligne dynamiquement les séries et reconnaît des comportements similaires **même décalés dans le temps**.

## Interprétation business :



- **DTW (Dynamic Time Warping)**: clients réguliers à forte valeur (achats toute l'année), clients “one-shot” (un achat massif à un seul moment), clients saisonniers (pics récurrents autour des saisons comme Noël ou été).
- **KShape** : clients inactifs presque tout le temps , clients concentrant leurs achats tous au même moment précis de l'année, clients actifs surtout au début de la période observée, avec quelques pics répétés.

Ces deux lectures sont complémentaires : DTW est utile pour comprendre le rôle business des clients (récurrent, saisonnier, opportuniste). KShape est utile pour cibler des groupes de clients à des périodes précises du calendrier.

## II. Embedding et clustering des produits:

L'objectif est d'analyser les produits afin de comprendre comment ils se comportent dans le temps et entre eux. Il s'agit de regrouper les produits selon leur proximité sémantique, leur co-achat et leur dynamique saisonnière.



Shape des features temporelles produits : (4631, 4)

StockCode	total_revenue	peak_month	peak_ratio	concentration
10002	6761.52	2010-05	2.676471	0.200770
10080	124.61	2011-11	3.132895	0.411161
10109	1.68	2009-12	1.000000	1.000000
10120	139.44	2009-12	3.192771	0.186194
10125	1601.80	2011-09	2.451617	0.092284

⚠️ Produits les plus saisonniers (concentration élevée) :

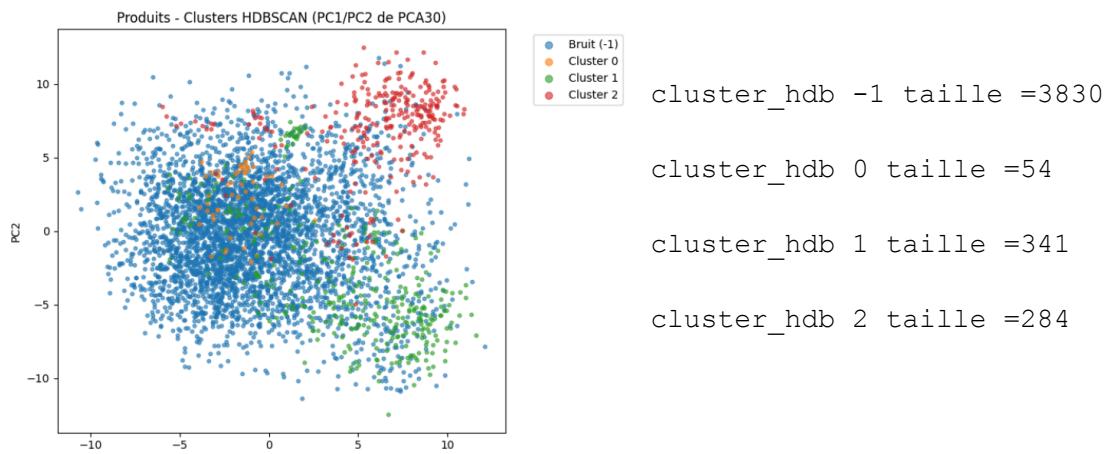
- TEST002 | This is a test product. | mois pic = 2009-12
- 47556 | SET/2 SPOTTY + ROSE TEA TOWELS | mois pic = 2009-12
- 72487 | ROSES WHITE ROUND CANDLE | mois pic = 2009-12
- 21480 | PINK STRIPE HOT WATER BOTTLE | mois pic = 2009-12
- 48179 | DOOR MAT SWEET HOME | mois pic = 2009-12
- 48175 | DOOR MAT CAMOUFLAGE | mois pic = 2009-12
- 20891 | LARGE GLASS ROSE SCENTED CANDLE | mois pic = 2010-01
- 20890 | SMALL GLASS ROSE SCENTED CANDLE | mois pic = 2010-01
- 47569 | ENGLISH ROSE DESIGN SHOPPING BAG | mois pic = 2009-12
- 20887 | BOX OF 6 PEBBLE CANDLES | mois pic = 2009-12

#### Etape 4 : Fusion des embeddings et clustering

Les trois types d'informations ont ensuite été combinés : la sémantique du produit (embeddings texte), les relations d'achat (Word2Vec), et les caractéristiques temporelles (features saisonnières).

Après normalisation, une réduction de dimension par PCA (30 composantes) a été appliquée pour stabiliser les distances, puis un clustering non paramétrique HDBSCAN a été utilisé.

Ce choix d'algorithme permet d'identifier des groupes de taille variable et de détecter automatiquement le "bruit"



Malheureusement on a pas pu avancer plus sur cette partie et expérimenter d'autres algorithmes de clustering plus avancés.

#### Conclusion

Cette analyse a mis en évidence la forte saisonnalité des ventes, avec des pics marqués en fin d'année. Le clustering client a révélé deux profils principaux : des clients réguliers à forte valeur et des acheteurs saisonniers concentrés sur la période des fêtes. Les approches DTW et KShape ont confirmé ces tendances en identifiant des comportements similaires, même décalés dans le temps. Enfin, l'exploration des embeddings produits (textuels, co-achat et temporels) a ouvert la voie à une segmentation plus riche, intégrant dynamique d'achat, sémantique et relations entre produits.