Soutenance Projet OC P5 DS: Segmentez des clients d'un site e-commerce

Candidat: David CAPELLE
Mentor: Nicolas MICHEL
Evaluateur: D. Lecoeuche

28/10/2021

Formation 100% Pôle Emploi

Plan de la soutenance

- Problématique et présentation de la démarche
- Présentation du nettoyage des données, du feature engineering et de l'analyse exploratoire des données
- Présentation des pistes de modélisation testées
  - Synthèse sur les métriques de qualité du clustering
  - Choix du modèle
  - Signification métier des clusters
- Conclusion, améliorations et limites

Problématique du projet

- Aider les équipes d'Olist à comprendre les différents types d'utilisateurs.
- Utilisation de méthodes non supervisées pour regrouper des clients avec des profils similaires (segmentation client).
- La segmentation doit être facilement exploitable par l'équipe marketing.
- Définir un contrat de maintenance établissant les périodes de mise à jour des segments clientèle.

Démarche pour le choix du modèle de clustering

- Pour chaque modèle de clustering testé :
  - Détermination des paramètres du modèle (nombre clusters,...).
  - Entraînement du modèle avec les paramètres.
  - Visualisation T-SNE 2D et 3D interactif avec les labels des clusters déterminés par le modèle.
- Choix final du meilleur modèle selon les critères :
  - **Techniques** : coefficient de silhouette et distance inter-cluster
  - **Métier** : tailles des clusters, nombre de cluster

  => évaluation de la facilité d'exploitaion / maintenance
- Détermination de la stabilité temporelle des clusters.

Pistes de modélisation

- **Préliminaire :** Analyse RFM avec la prise en compte des catégories de produit (22 variables en entrée).

- **Piste de modélisation (Hypothèse 1)** : modélisation à partir de 22 variables issues du feature engineering, avec les catégories de produit
  - But : segmentation comportementale du client en intégrant la dimension produit

- **NB**: une autre piste est étudiée **(Hypothèse 2)** dans un notebook séparé, avec 12 variables en entrée, sans les catégories produit :
  - But : vérifier si la prise en compte des catégories produit ne fausse pas le clustering(poids des variables produit)
  - Cette analyse est présentée en annexes, mais pas dans la présentation.

Présentation des jeux de données

- **9 jeux de données issues du site e-commerce Olist :**
  - Données client couvrant 23 mois de 2016 à 2018.
  - Contenu :
    - Commandes client (99441)
    - Clients (99441)
    - Moyens de Paiement
    - Avis client
    - Géolocalisation
    - Lignes de commandes
    - Produits
    - Catégories de produit
    - Revendeurs

- Le jeu de données des revendeurs n'est pas utilisé dans ce projet.

Feature engineering - Variables essentielles (1/2)

- Merge des 7 tables principales pour créer un dataset orienté commandes.

- Création d'un dataset orienté client avec 12 nouvelles variables :
  - **nb_orders** : nbre d'achat par client
  - **mean_payment_sequential** : nbre moyen de moyen paiement.
  - **mean_review_score** : score moyen avis client.
  - **mean_payment_installments** : nbre moyen échéances paiement.
  - **mean_delivery_days** : délai moyen de livraison en jour.
  - **favorite_purchase_month** : numéro mois d'achat favori en moyenne.
  - **favorite_purchase_hour** : heure favorite d'achat en moyenne.
  - **mean_nb_items** : nbre d'articles moyen commandés par client.
  - **order_mean_delay** : délai moyen d'achat en jours par client.
  - **freight_ratio** : ratio moyen des frais de livraison sur la dépenses d'achat (en %).
  - **mean_price_order** : dépense moyenne d'achat par client.
  - **harvesine_distance** : éloignement du client du site e-commerce Olist.

Feature engineering - Catégories de produit (2/2)

- **Création de 10 variables Catégories de produit :**
  - Création de regroupements de catégories .
  - Mesure le pourcentage d'achat client d'une catégorie de produit donnée par rapport à l'ensemble des catégories .

- **Regroupement des catégories de produit :**
  - **books_cds_media :** biens culturels
  - **fashion_clothing_accessories** : vêtements, mode
  - **flowers_gifts** : fleurs, cadeaux
  - **groceries_food_drink** : épicerie, boissons
  - **health_beauty** : santé / beauté
  - **home_furniture** : fourniture pour la maison
  - **other** : produits non classés dans les autres catégories
  - **sport** : articles de sport
  - **technology** : biens high-tech
  - **toys_baby** : jouets / produits bébé

Nettoyage des données - Traitement valeurs manquantes

- **Traitement des valeurs quantitatives :**
  - Imputation valeur 1 si présence valeurs NaN pour les variables sur les échéances et moyens de paiement.

  => Taille du jeu de données : (110197 lignes, 23 variables) Proportion NaN : 0,06 %

- **Traitement des valeurs date :**
  - Suppression des valeurs NaN pour les dates de livraison client.

  => Taille du jeu de données : (110189 lignes, 23 variables) Proportion NaN : 0.06 %

- **Traitement des valeurs qualitatives :**
  - Imputation de la valeur « other » si présence valeurs NaN dans les catégories de produit.

  => Taille du jeu de données : (110189 lignes, 23 variables) Proportion NaN : 0.06 %

Nettoyage des données - Traitement des outliers

- **Traitement des outliers - valeurs aberrantes :**
  - Remplacement des valeurs 0 par 1 pour la variable moyenne du nombre d'échéances de paiement.

- **Traitement des outliers - analyse valeurs atypiques :**
  - Suppression des valeurs atypiques par la méthode inter-quartile pour les variables « mean_delivery_days » et « mean_price_order »

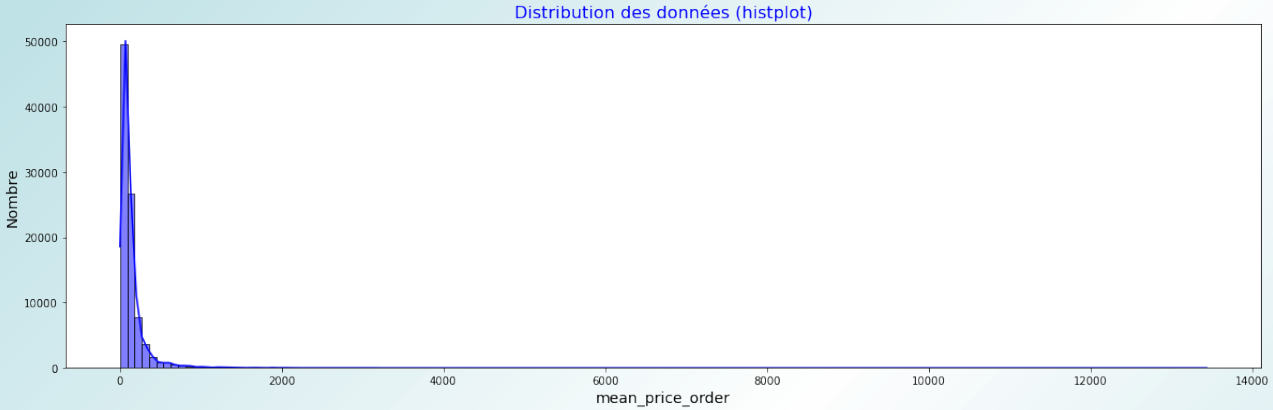
=> Après feature engineering et agrégation données par client , taille du jeu de données : (92234 lignes, 30 variables) Proportion de NaN : 0 %

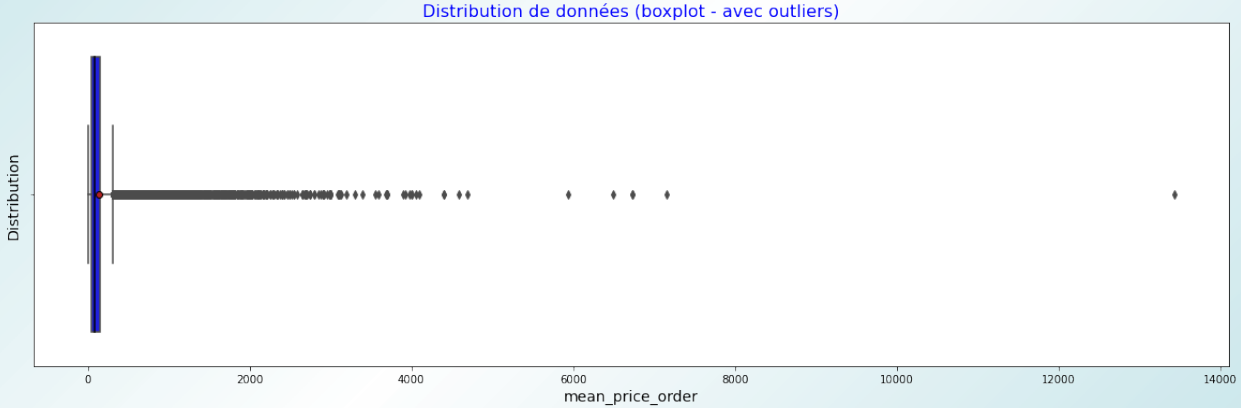
Analyse exploratoire des données
Analyse univariée – Principaux enseignements

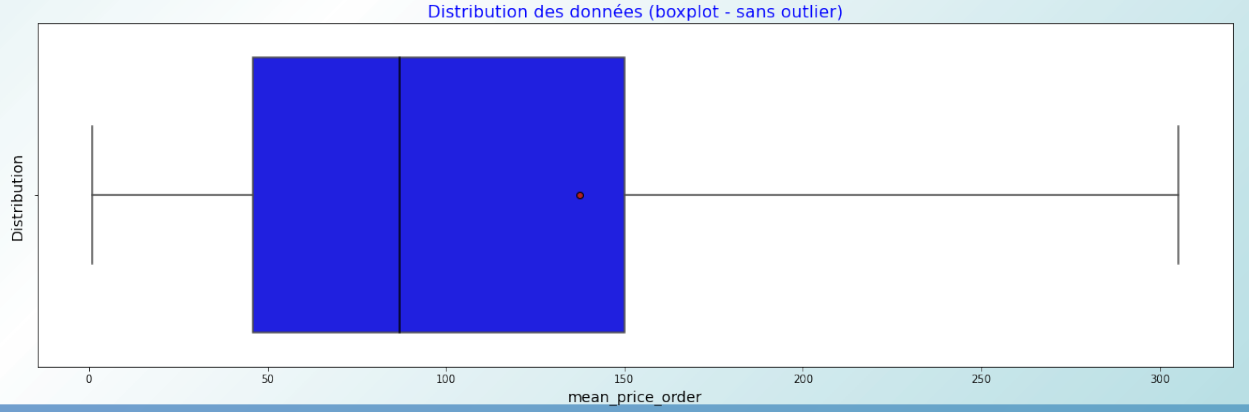
- **Analyse des variables:**
  - Les clients commandent 1 seul article en général.
  - Les clients dépensent assez peu entre 60 et 180 real en moyenne.
  - Paiement avec un seul moyen de paiement (carte crédit).
  - Les clients sont assez satisfaits (score entre 4 et 5).
  - Les clients paient en 1 jusqu'à 4 échéances de paiement.
  - En moyenne, le délai de livraison est de 10 jours.
  - Panier moyen constitué d'un seule catégorie de produit.
  - Les clients achètent rarement plus de 1 article.

Analyse exploratoire des données

Analyse univariée – Distribution variable « mean_price_order »

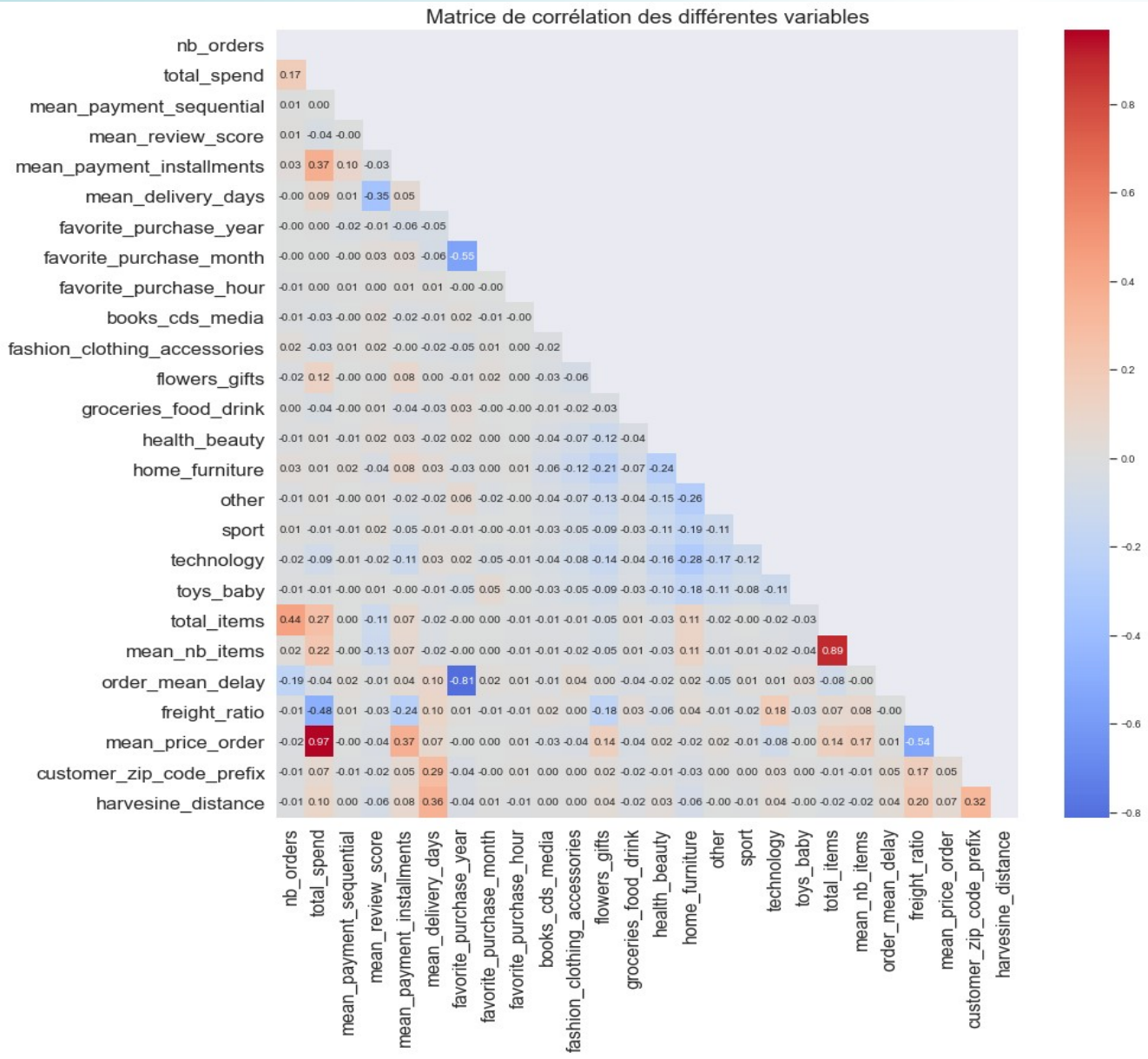






Analyse exploratoire des données

Analyse multivariée – Matrice des corrélations



Piste de modélisation – Etape préliminaire
Pré-requis et Analyse RFM

- Pré-requis : échantillonnage du dataset sur une période de 3 mois, soit 21 % des données d'origine. *Contraintes de ressources système (CPU et mémoire)*.

- Calcul des scores R, F, et M (note de 1 à 4).

- Affectation des segments RFM par sous-ensemble des scores RFM.

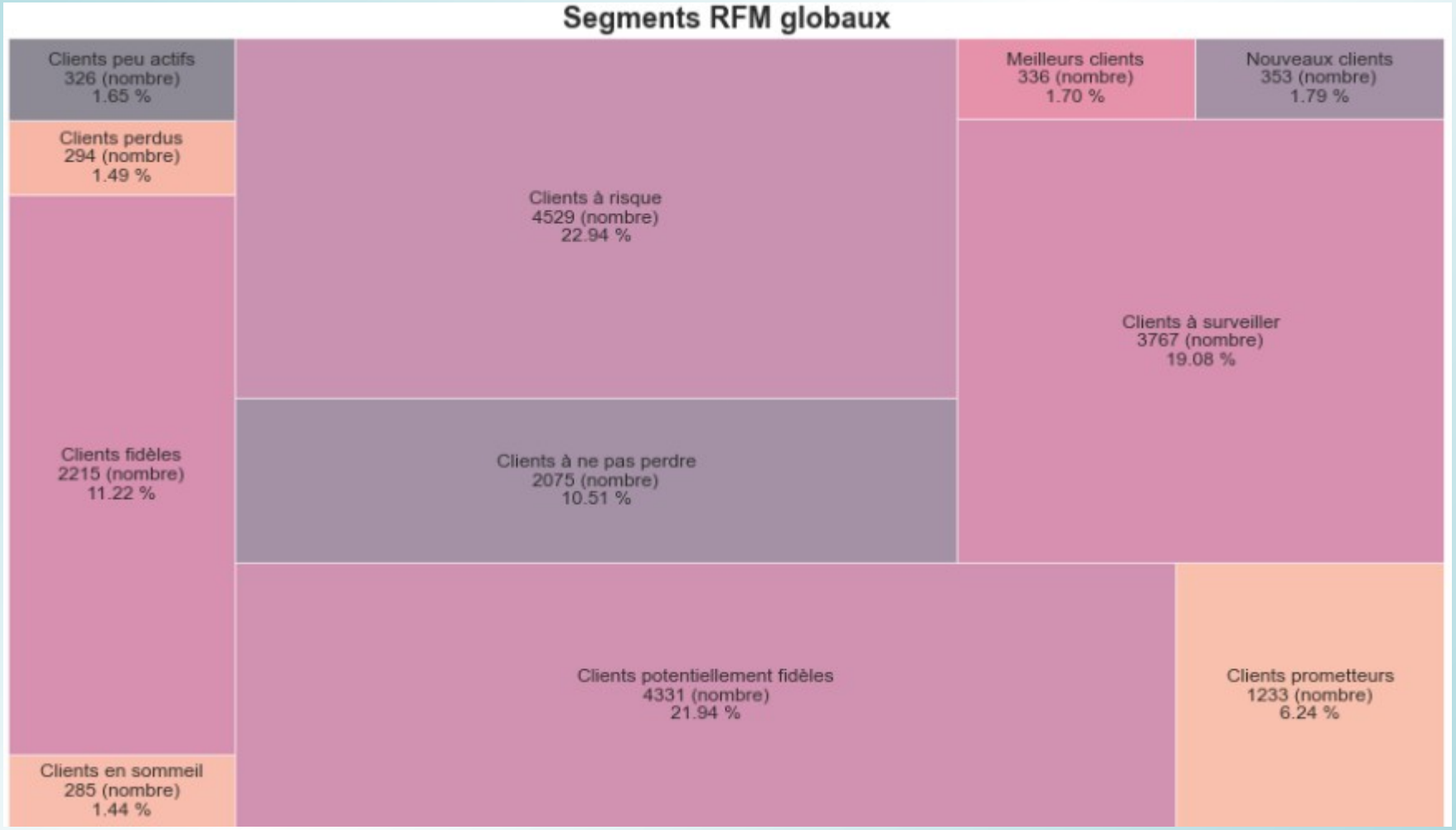
- Analyse de la qualité de la segmentation clientèle proposée par RFM par visualisation T-SNE 2D et 3D avec labels segments RFM.

=> L'analyse RFM ne permet pas de distinguer des regroupements distincts par rapport aux segments RFM.

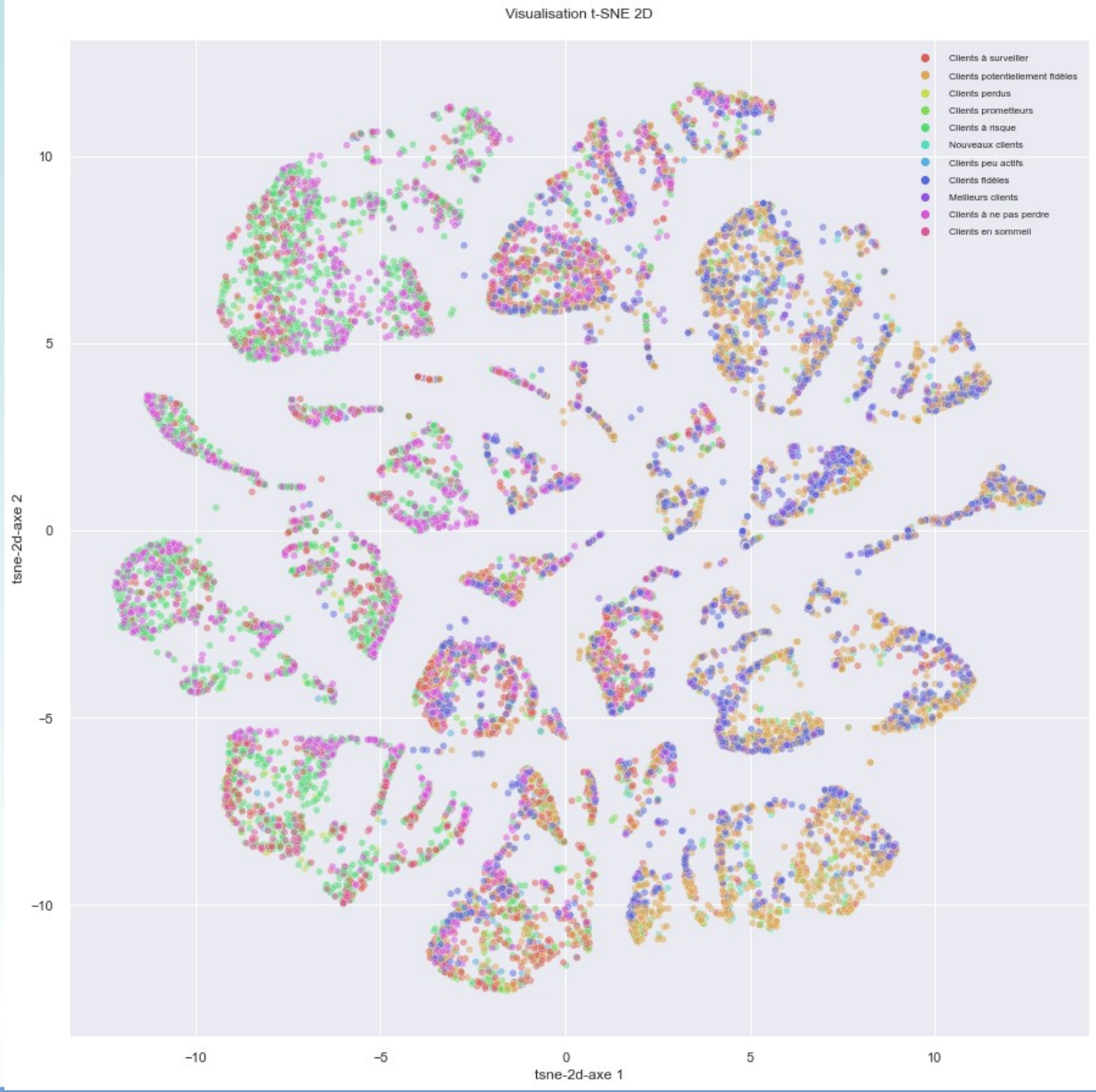
=> Tentative de détermination de clusters plus précis par clustering non supervisé.

Analyse RFM

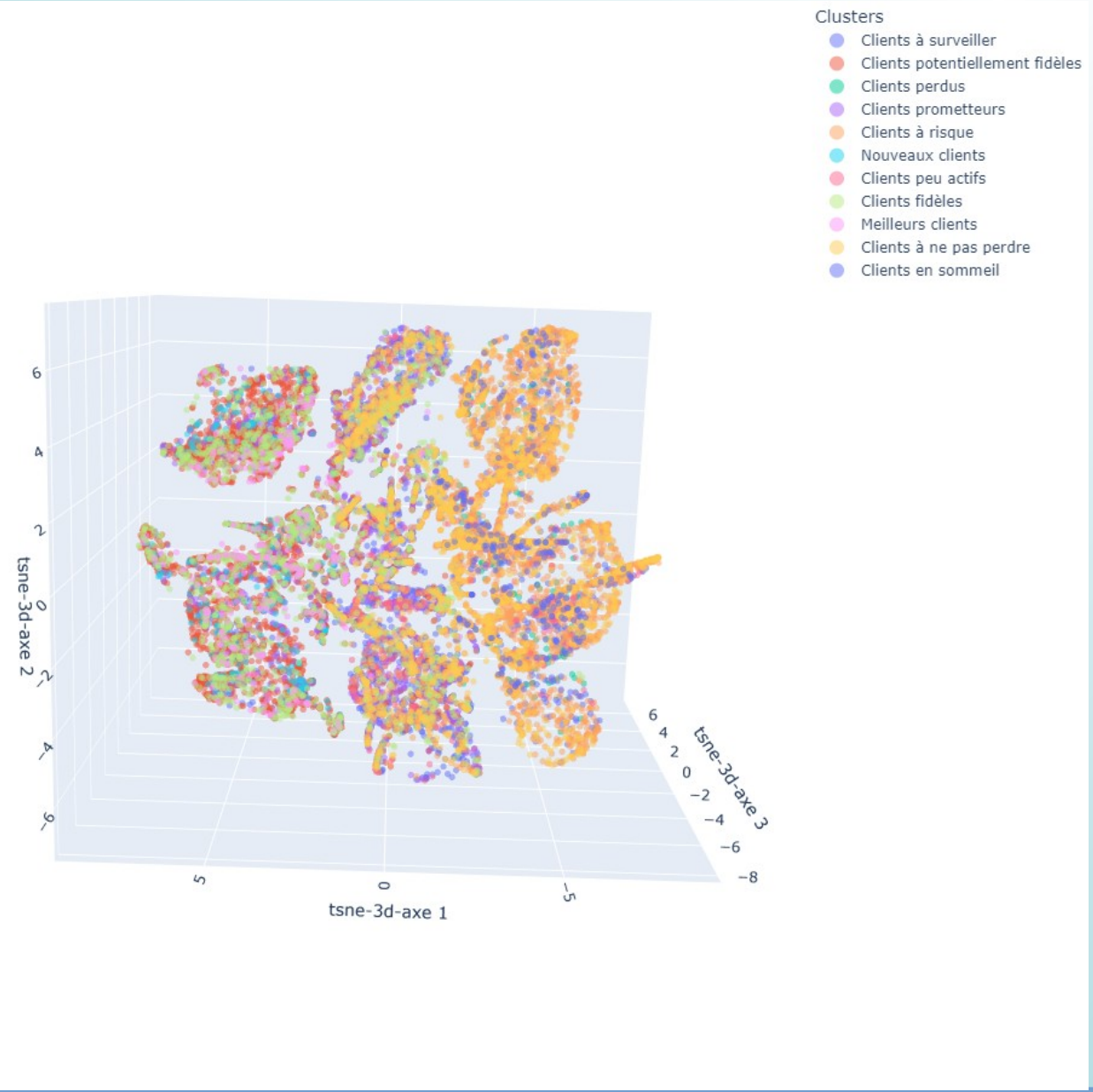
Décomposition des segments RFM



Analyse RFM - Visualisation T-SNE 2D



Analyse RFM - Visualisation T-SNE 3D



Piste de modélisation – Hypothèse 1 - Principes

Méthode de clustering K-means

- Modélisation la plus simple avec 22 variables en entrée (ajout catégories de produit).

- Standardisation des données.

- Détermination du nombre de clusters :
  - Par la méthode du coude en visuel **=> K entre 6 et 7**.
  - Par la méthode du coude basée sur le score de distorsion

  => K = 8

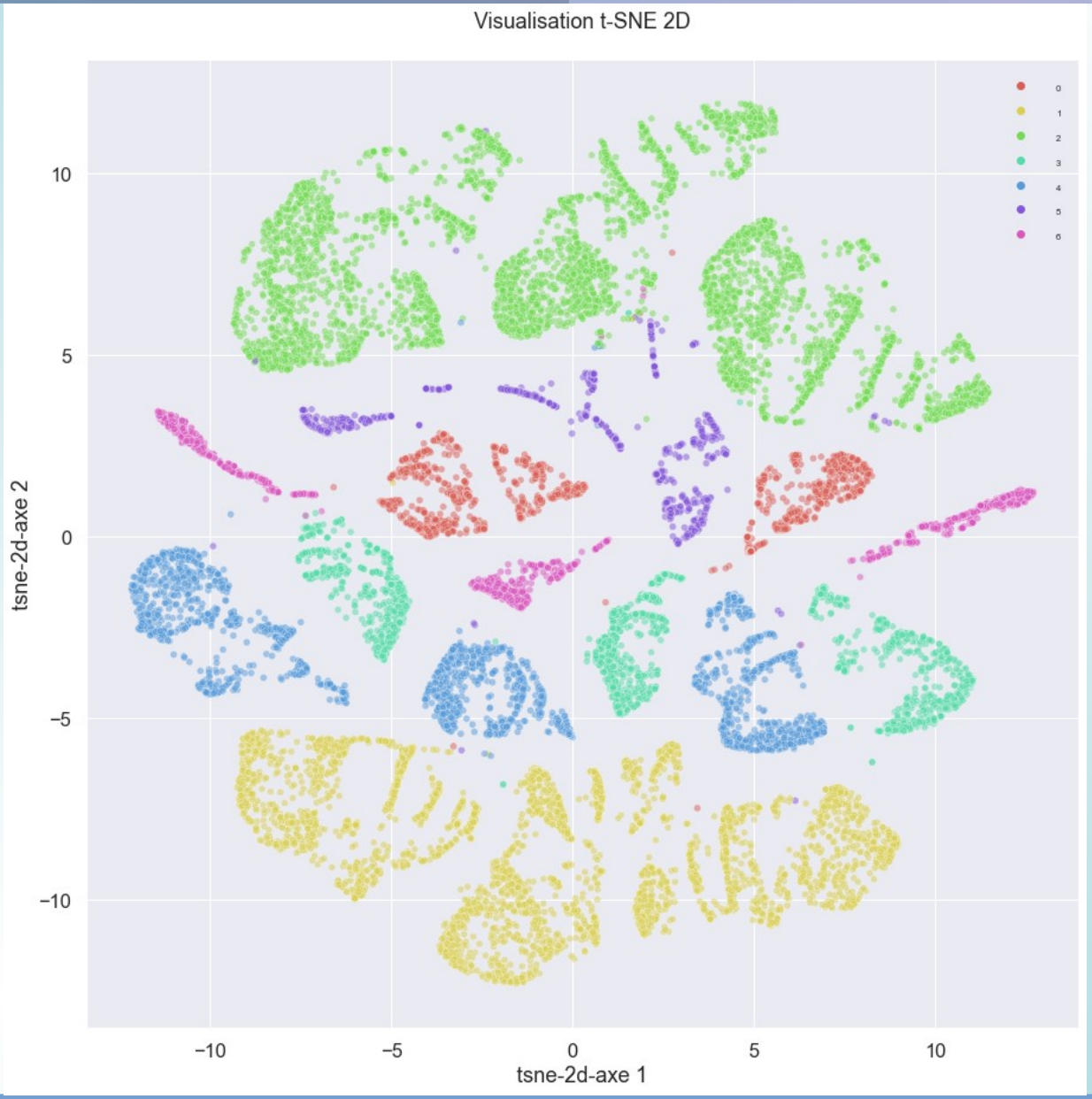
  - En trouvant le score de silhouette optimal

  => K = 7 pour un score de silhouette moyen = 0.4109

Choix du nombre de clusters = 7

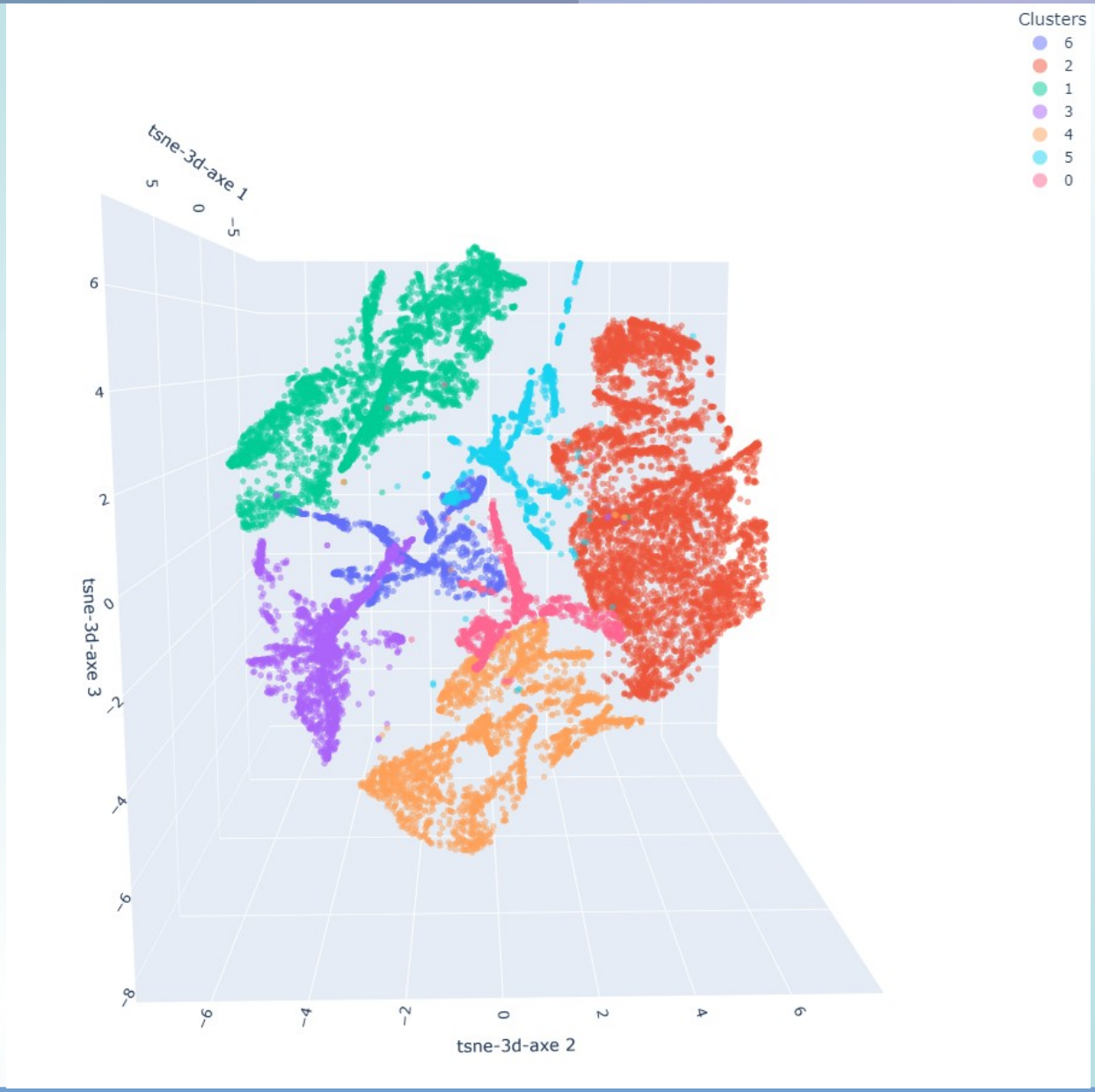
Hypothèse 1 - Visualisation T-SNE 2D

Modélisation de clustering K-means



Hypothèse 1 – Visualisation T-SNE 3D

Méthode de clustering K-means



Piste de modélisation – Hypothèse 1 - Principes

Méthode de clustering DBSCAN

- Modélisation la plus simple avec 22 variables en entrée.

- Standardisation des données.

- Détermination des paramètres :
  - Pour min_samples, application de la règle 2*dimensions

  => min_samples = 44

  - Pour epsilon (eps), application de la méthode des plus proches voisins

  => eps environ 0,3

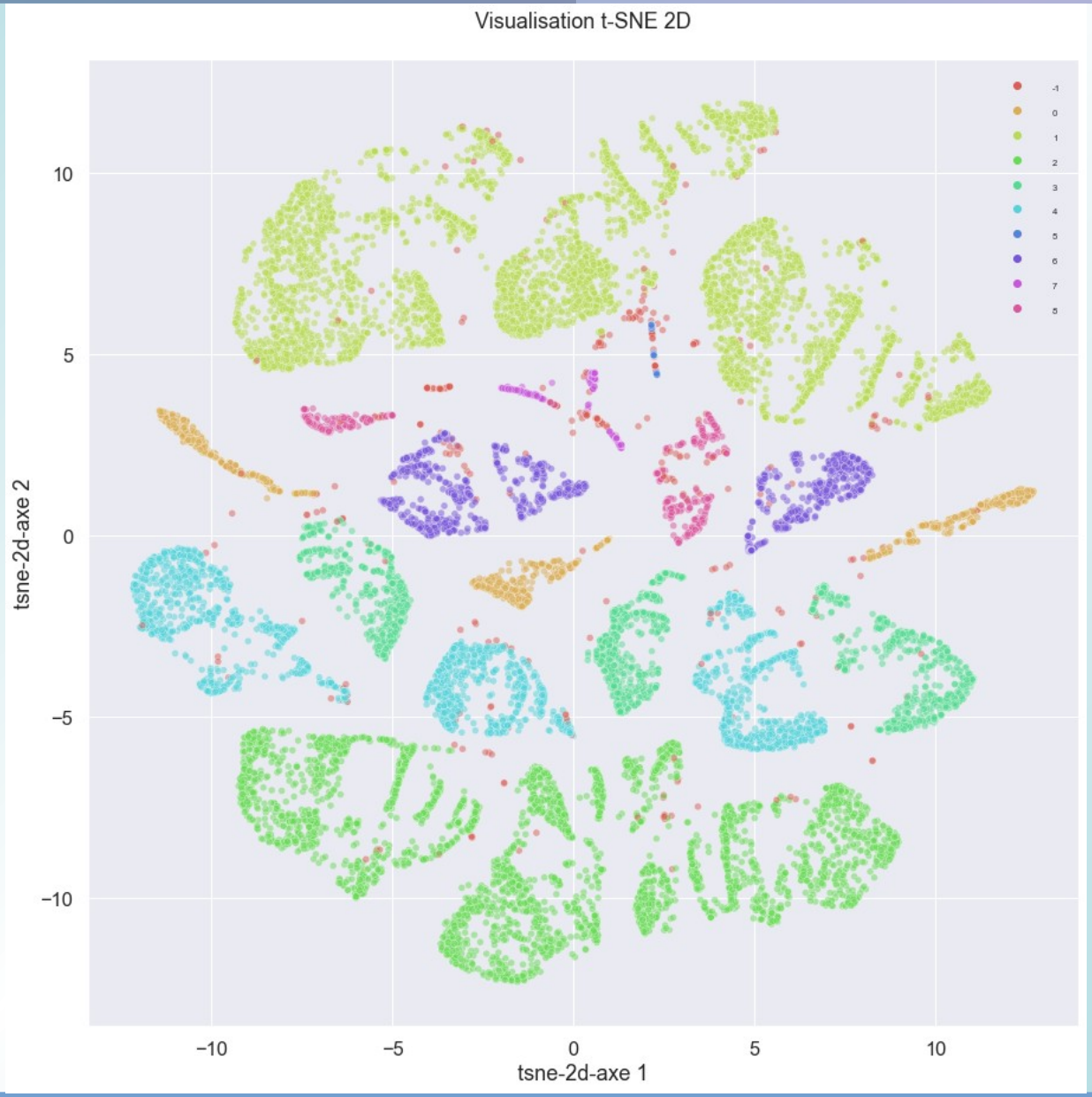
  - De manière empirique, calcul du score de silhouette optimal en calculant plusieurs modèle DBSCAN pour déterminer K et eps.

  => K = 9, eps = 0.54 pour un score de silhouette moyen = 0.413

Choix du nombre de clusters = 9 avec 594 points de bruit.

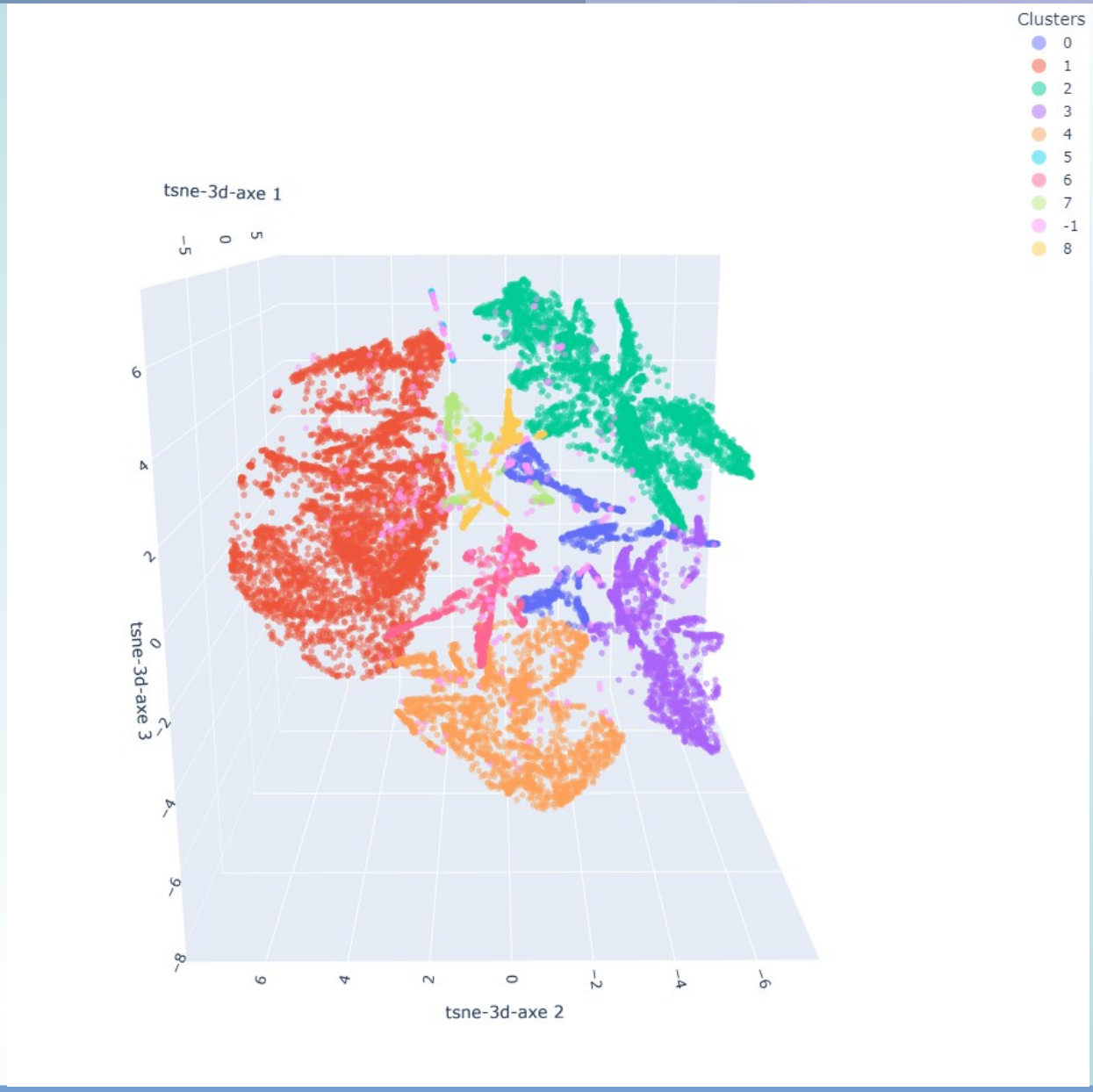
Hypothèse 1 - Visualisation T-SNE 2D

Modélisation de clustering DBSCAN



Hypothèse 1 – Visualisation T-SNE 3D

Méthode de clustering DBSCAN



Piste de modélisation – Hypothèse 1 - Principes

Méthode de clustering hiérarchique (agglomerative clustering)

- Modélisation la plus simple avec 22 variables en entrée.

- Standardisation des données.

- Détermination des paramètres :
  - Détermination du nombre de clusters K avec le dendrogram

  => K = 5, 6 ou 7

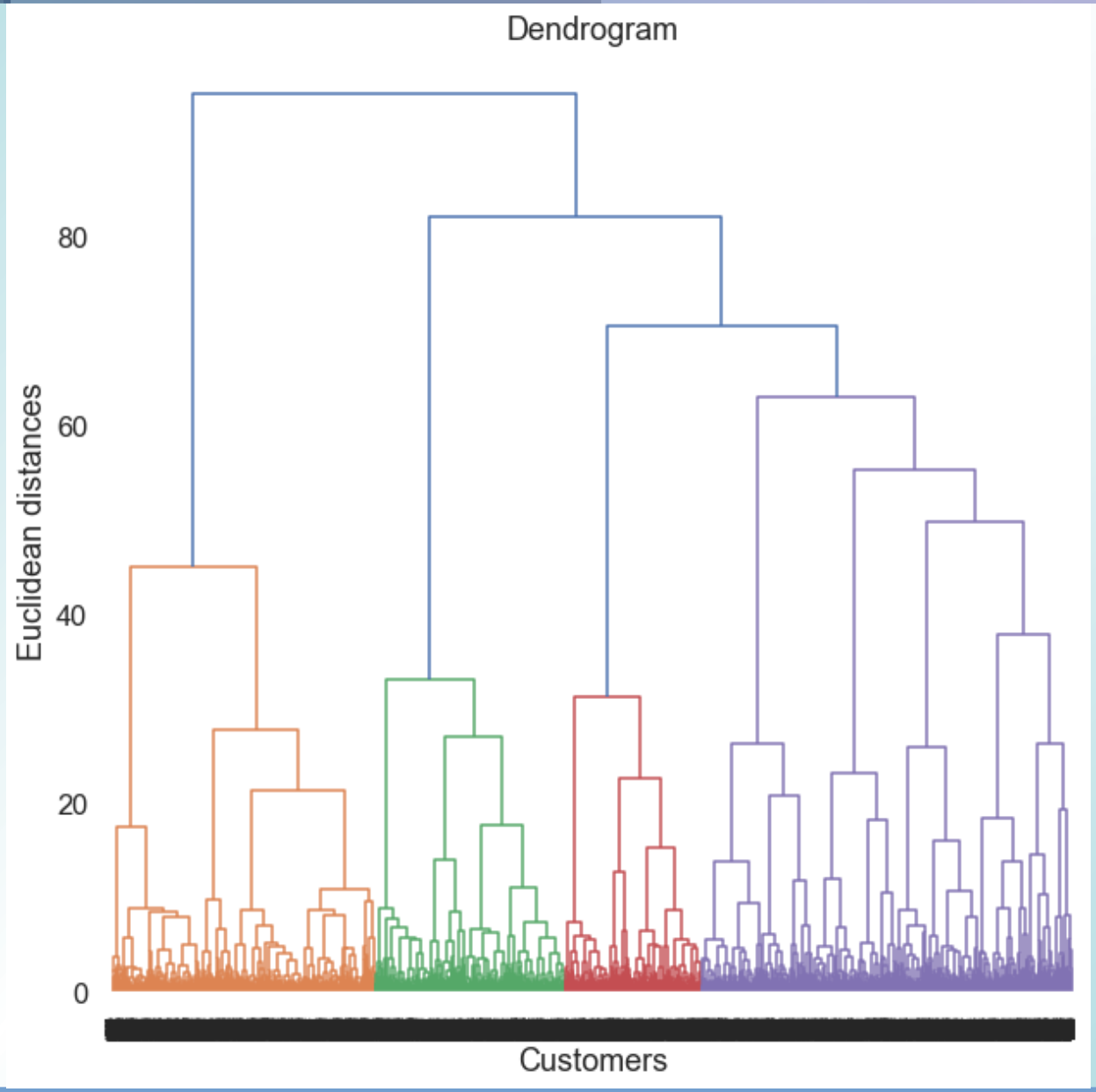
  - Calcul du score moyen de silhouette avec K = 5, 6 et 7

  => score de silhouette moyen optimal (0.41) pour K = 7

Choix du nombre de clusters = 7.

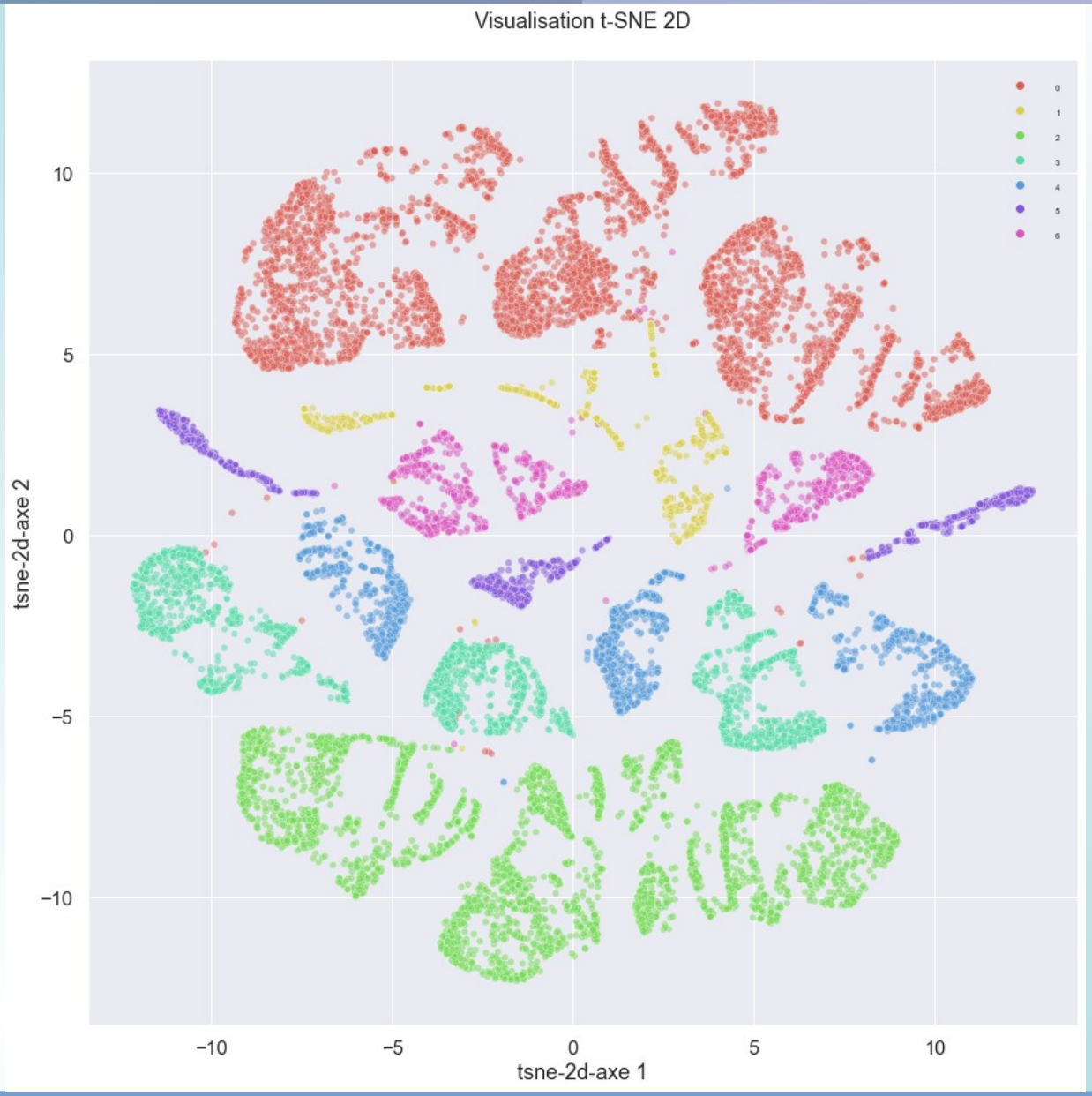
Piste de modélisation – Hypothèse 1 - Dendrogram

Méthode de clustering hiérarchique (agglomerative clustering)



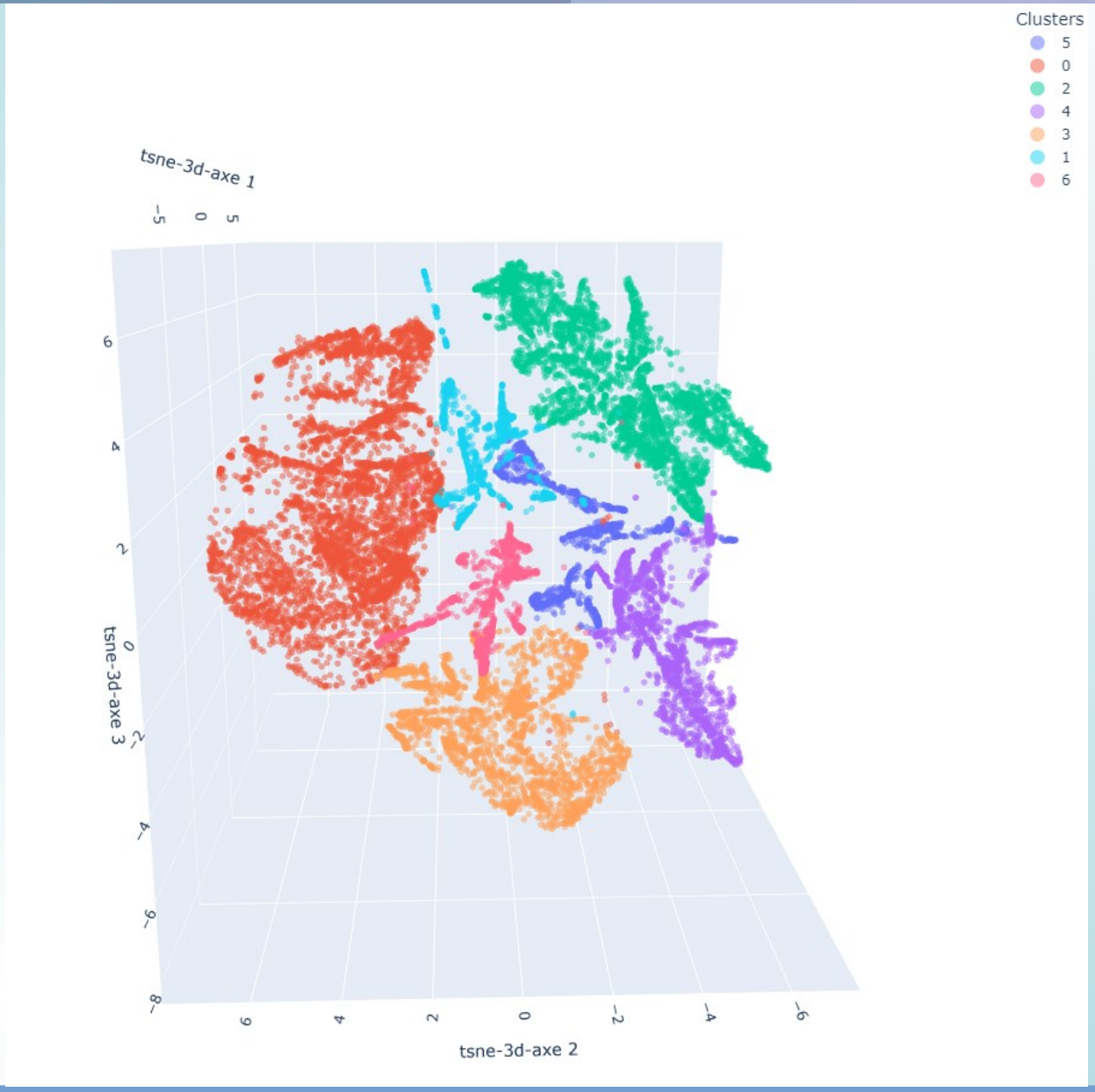
Hypothèse 1 - Visualisation T-SNE 2D

Modélisation de clustering hiérarchique (agglomerative clustering)



Hypothèse 1 – Visualisation T-SNE 3D

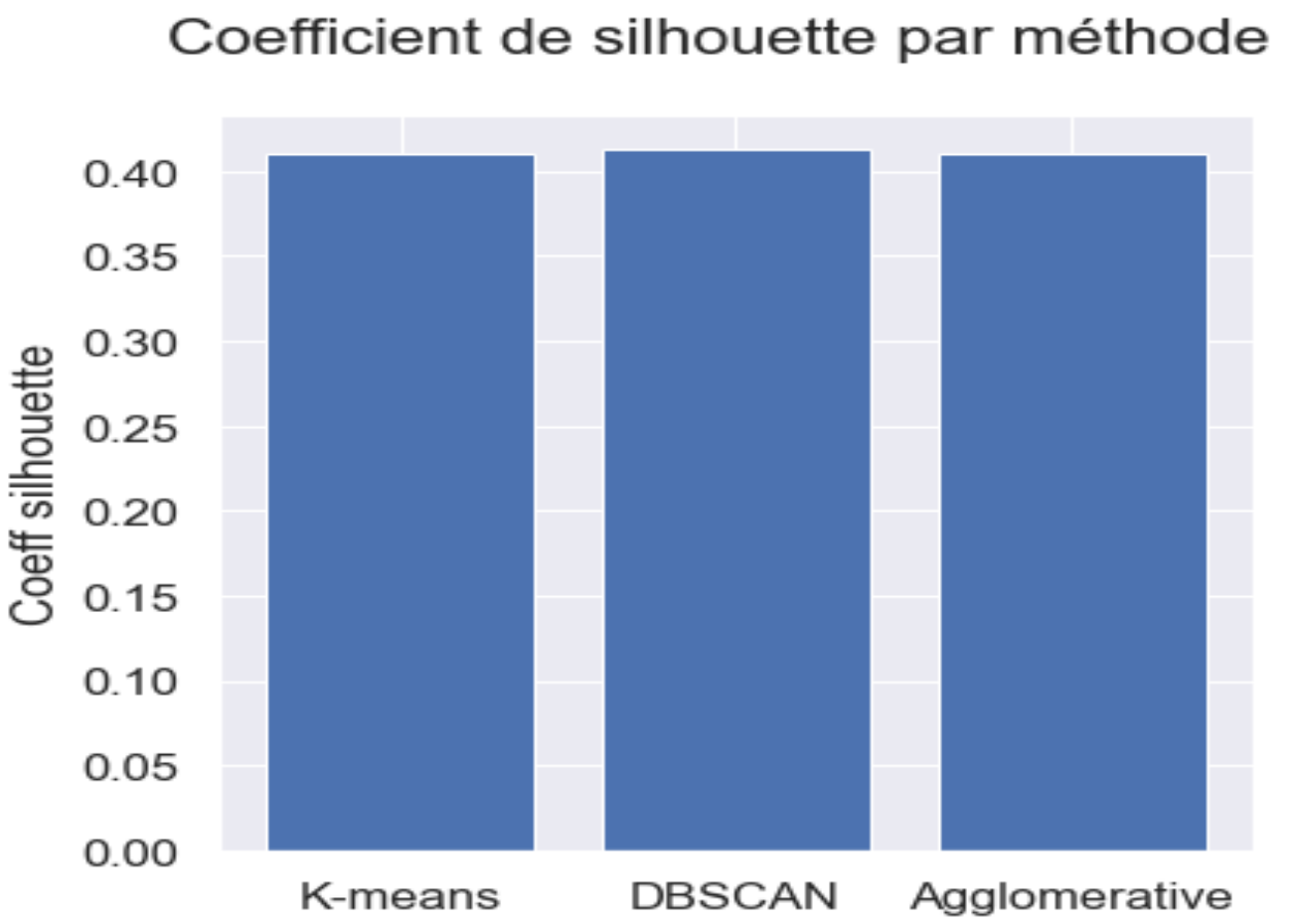
Méthode de clustering hiérarchique (agglomerative clsutering)



Piste de modélisation – Hypothèse 1

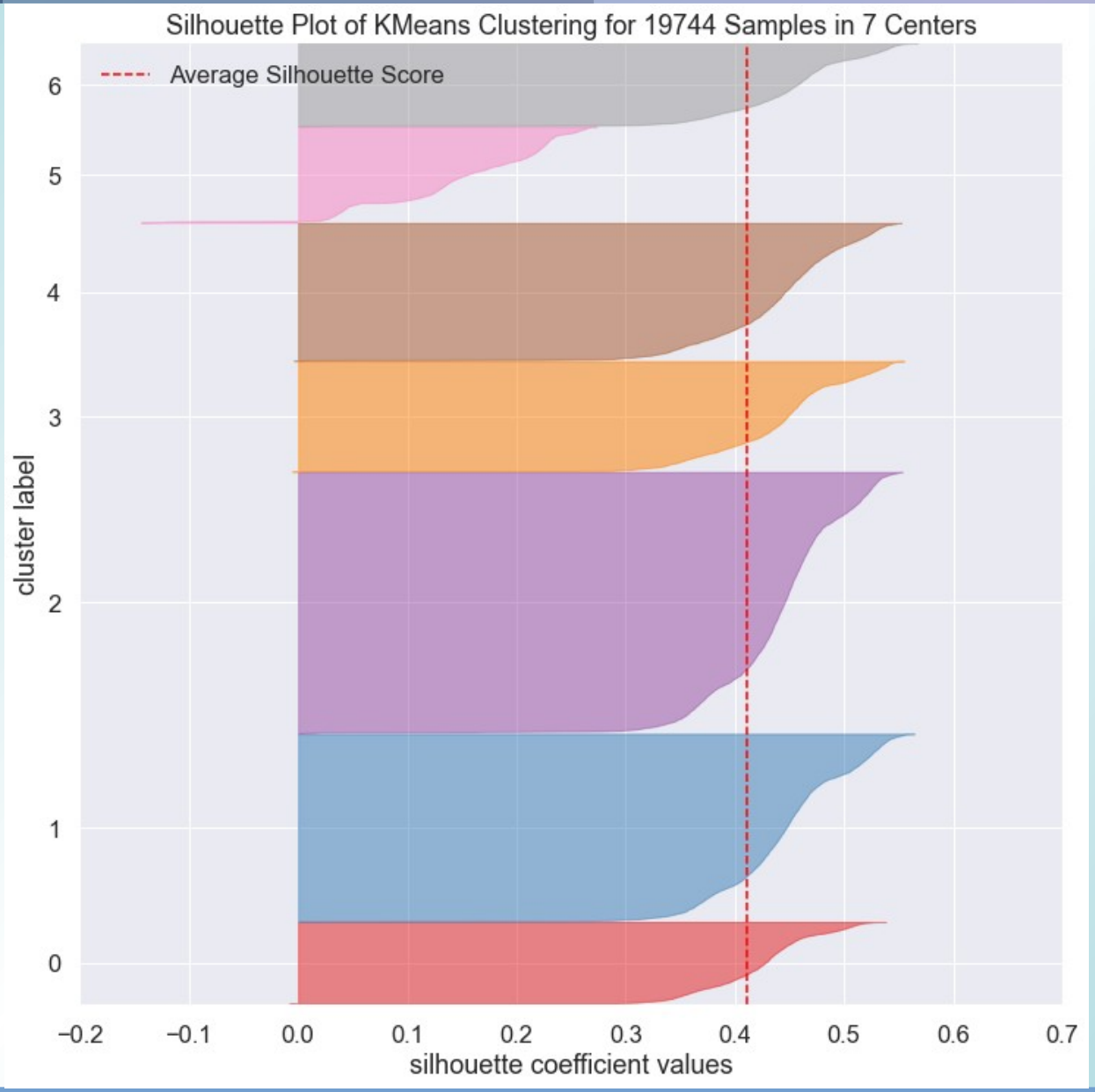
Qualité du clustering – Facteur de forme (coeff silhouette)

- Sur le coefficient de silhouette, **le modèle DBSCAN à 9 clusters** présente le meilleur coefficient moyen de silhouette, même si tous les modèles sont pratiquement à égalité.



Piste de modélisation – Hypothèse 1 - K-means

Qualité du clustering – Représentation graphique facteur forme



Piste de modélisation – Hypothèse 1

Qualité du clustering – Séparation des clusters

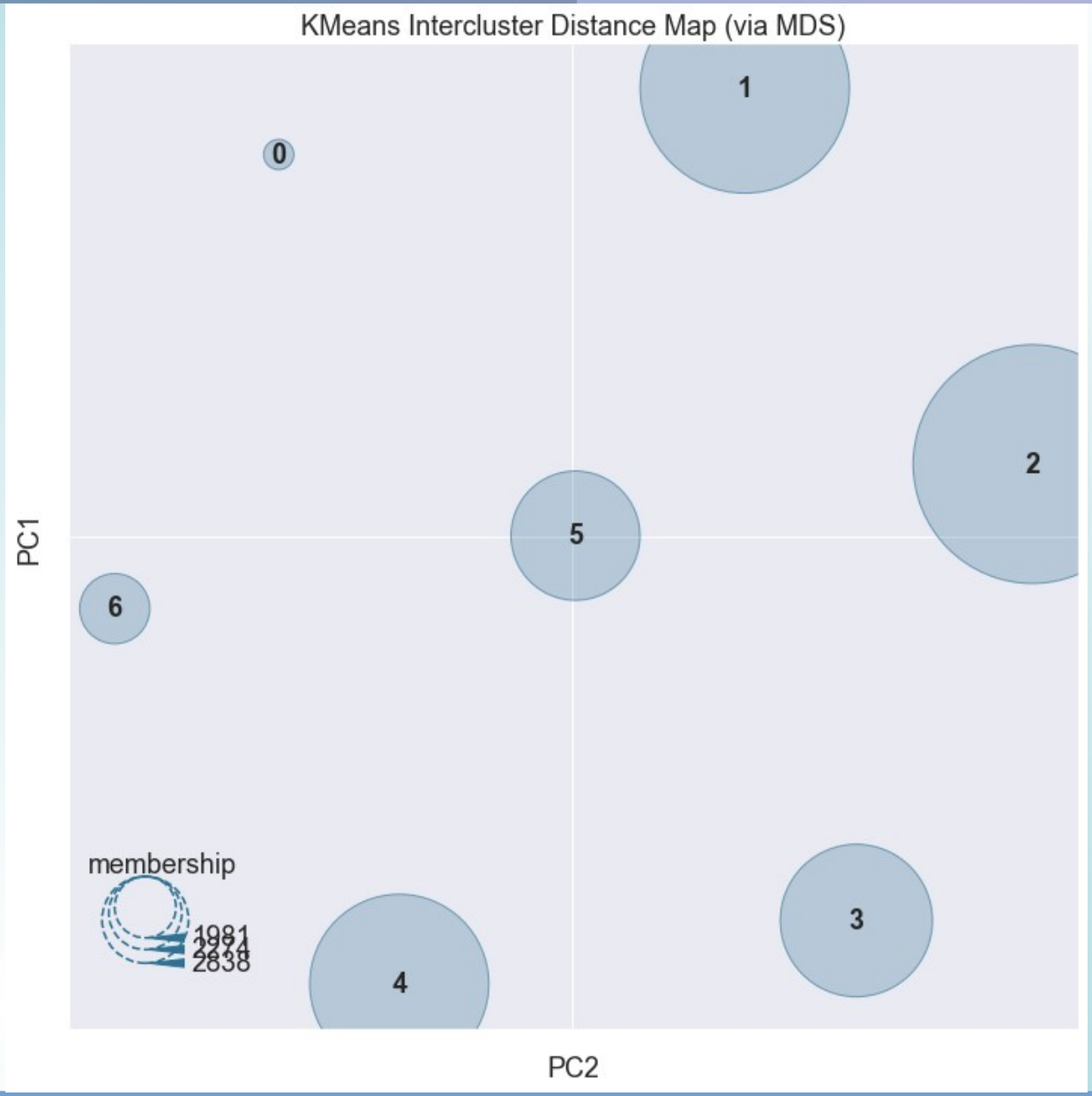
- Pour le modèles DBSCAN, la visualisation T-SNE 2D/3D montre des clusters peu marqués, avec du bruit.

- Les modèle K-means et clustering hiérarchique présentent pratiquement un clustering identique à la visualisation.

=> Même si le modèle hiérarchique est à égalité sur le critère score de silhouette, **le modèle k-means à 7 clusters** est considéré comme le meilleur choix sur ce critère, le clustering hiérarchique mobilisant beaucoup plus de ressources (CPU et mémoire).

Piste de modélisation – Hypothèse 1 - K-means

Qualité du clustering – Graphique distance inter-clusters



Piste de modélisation – Hypothèse 1

Qualité du clustering – Homogénéité des tailles de clusters

- Sur ce critère, **le modèle DBSCAN à 9 clusters** est écarté (petits clusters, difficulté d'exploiter et maintenir 9 clusters).

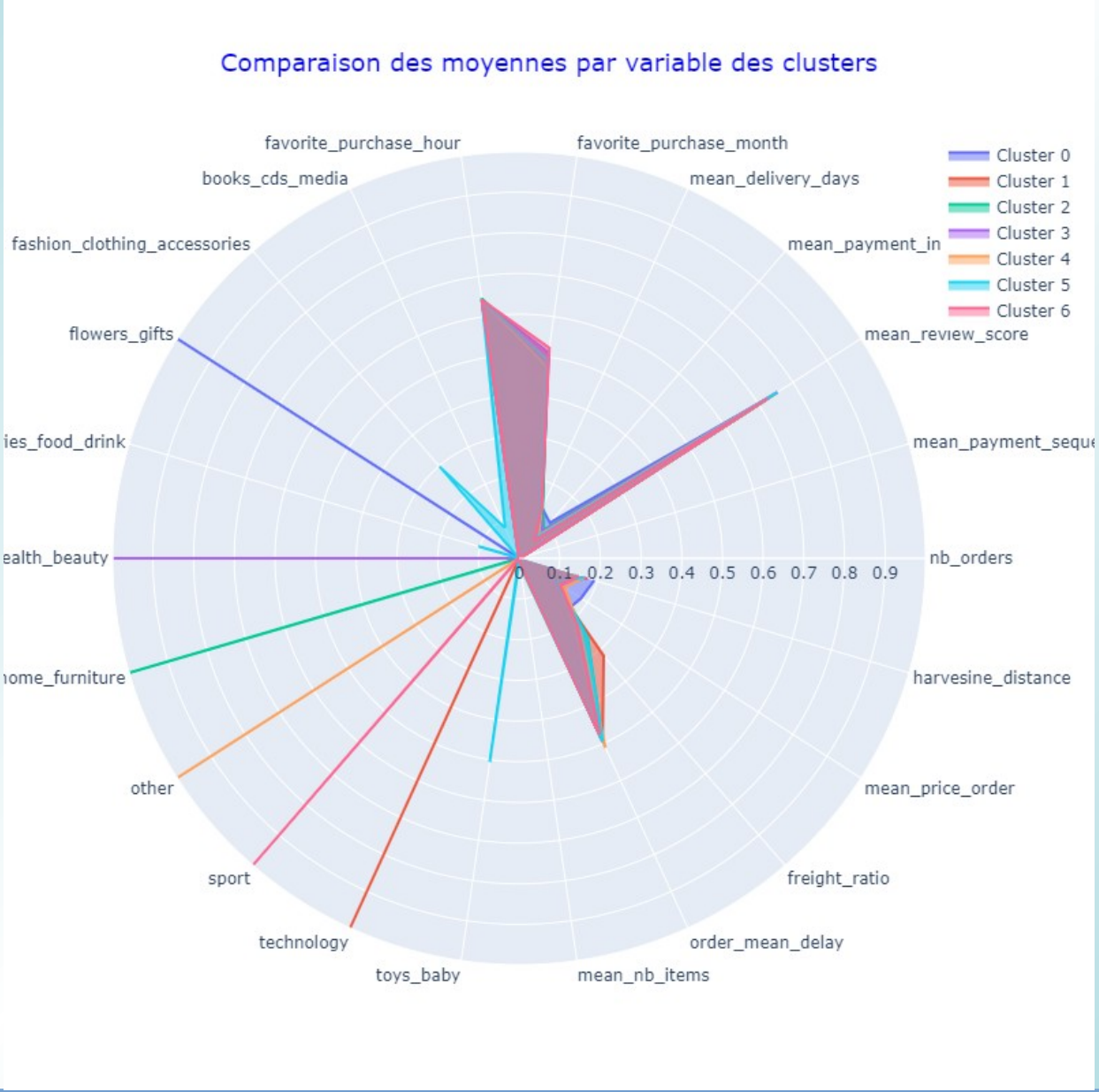
- Les modèles K-means et clustering hiérarchique présente un homogénéité dans les tailles des clusters.

Modèle	Taille cluster 0	Taille cluster 1	Taille cluster 2	Taille cluster 3	Taille cluster 4	Taille cluster 5	Taille cluster 6	Taille cluster 7	Taille cluster 8
K-means	1688	3869	5387	2274	2838	1981	1707		
DBSCAN	1679	5328	3834	2242	2790	168	1650	517	942
Agglomerative clustering	5443	1952	3868	2818	2273	1696	1694		

=> Choix final : le modèle K-means est retenu.

Hypothèse 1 - Modèle K-means à 7 clusters

Signification métier des clusters (1/2)



Hypothèse 1 - Modèle K-means à 7 clusters

Signification métier des clusters (2/2)

- **Cluster 0** : clients achetant des **fleurs et cadeaux**, **meilleurs clients en termes de dépenses d'achat,** privilégiant des frais de livraison réduits.
- **Cluster 1** : clients achetant des **produits technologiques**, client peu dépensiers, acceptant des frais de livraison élevés.
- **Cluster 2** : clients achetant des **fournitures pour la maison**, **clients peu actifs (fréquence et dépenses d'achat).**
- **Cluster 3** : clients achetant des produits de **santé/beauté**, **clients peu actifs (fréquence et dépenses d'achat)**.
- **Cluster 4** : clients achetant des **produits peu courants** (catégorie 'other'), **clients à surveiller, marchés de niche**.
- **Cluster 5** : clients achetant 3 catégories de produit de grande consommation. **Clients prometteurs ,clients peu actifs (fréquence et dépenses d'achat)**.
- **Cluster 6** : clients achetant des **articles de sports**, **clients peu actifs (fréquence et dépenses d'achat)**

Hypothèse 1 - Modèle K-means à 3 clusters

Evaluation de la stabilité temporelle du clustering (1/2)

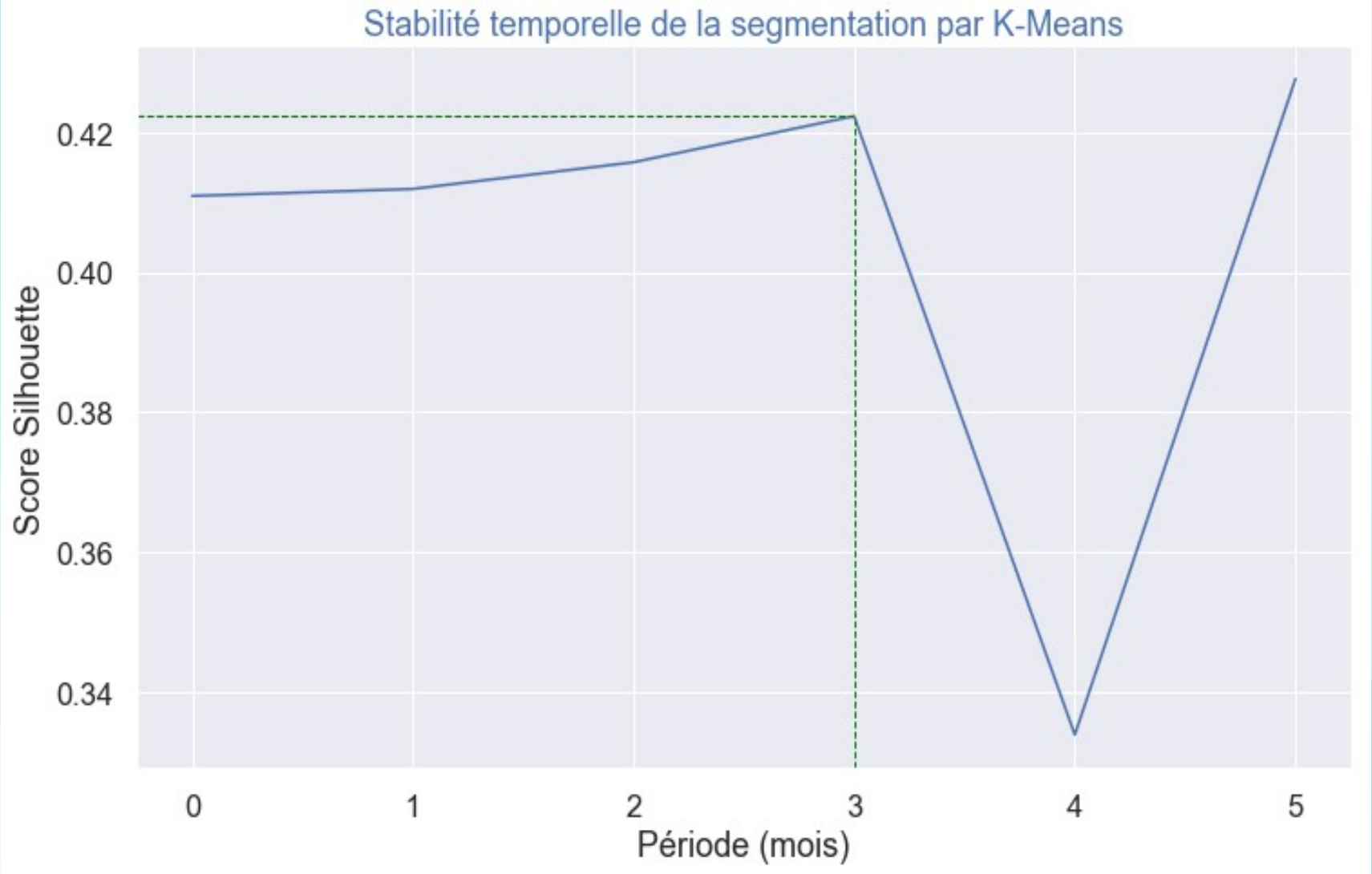
- **Méthode :**
  - Evaluation du score de silhouette sur 3 mois glissants par pas de 1 mois sur 5 mois.
  - Période initiale : 3 mois (du janvier à mars 2018).

- **Résultats** :
  - Mise à jour des clusters après le 4ème mois, **sans perte sur le score silhouette**.
  - Le pas 4 correspond à la période mai, juin, juillet 2018 => à partir du mois de mai, le score de silhouette se dégrade nettement (environ perte de 19%).

=> Le modèle k-means à 7 clusters est relativement limitée dans le temps.

Hypothèse 2 - Modèle K-means à 7 clusters

Evaluation de la stabilité temporelle du clustering (2/2)



Conclusion – Améliorations - Limites

- **Au final, la modélisation K-means avec 7 clusters montre :**
  - un score de silhouette moyen (0,41).
  - une stabilité du clustering limité dans le temps.
  - des profils client peu définis

- Compte tenu de la modélisation sur une période initiale de 3 mois, il est difficile de généraliser ces modèles en production.

- Il manque des informations dans le jeu de données (sexe, âge, profil socio-professionnel, …).

- Dans l'annexe, la piste de modélisation 2 (sans les catégories de produit) montre des résultats dégradés sur le score de si lhouette, la stabilité du clustering.

Annexes

Annexe 1 : PCA – cercle corrélations - 1er plan factoriel

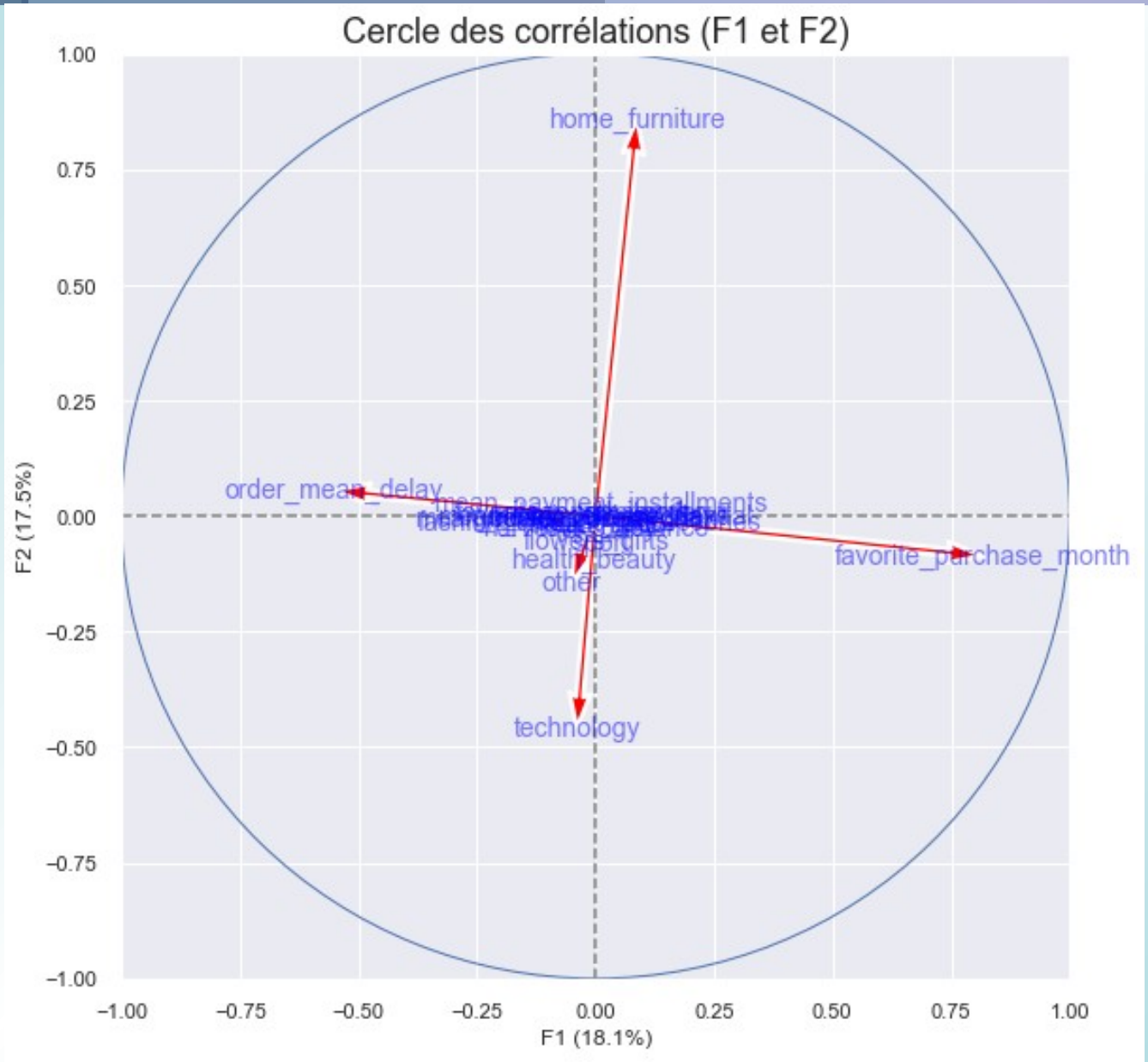
Annexe 2 : PCA – cercle corrélations - 2ème plan factoriel

Annexe 3 : PCA – cercle corrélations – 3ème plan factoriel

Annexe 4 : Piste modélisation – hyp 2 (slide 42 à 60)

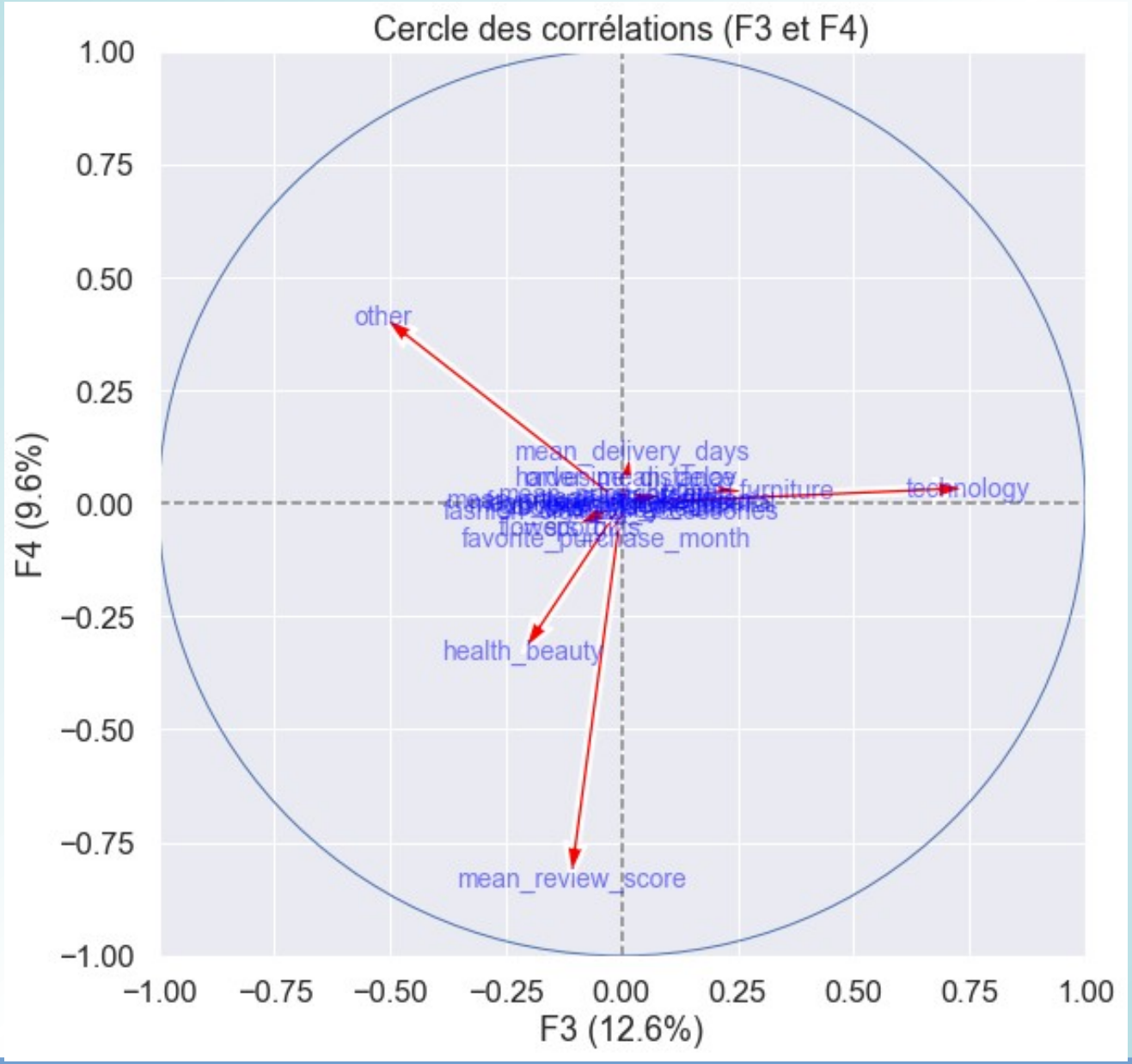
PCA – Cercle de corrélations

1er plan factoriel



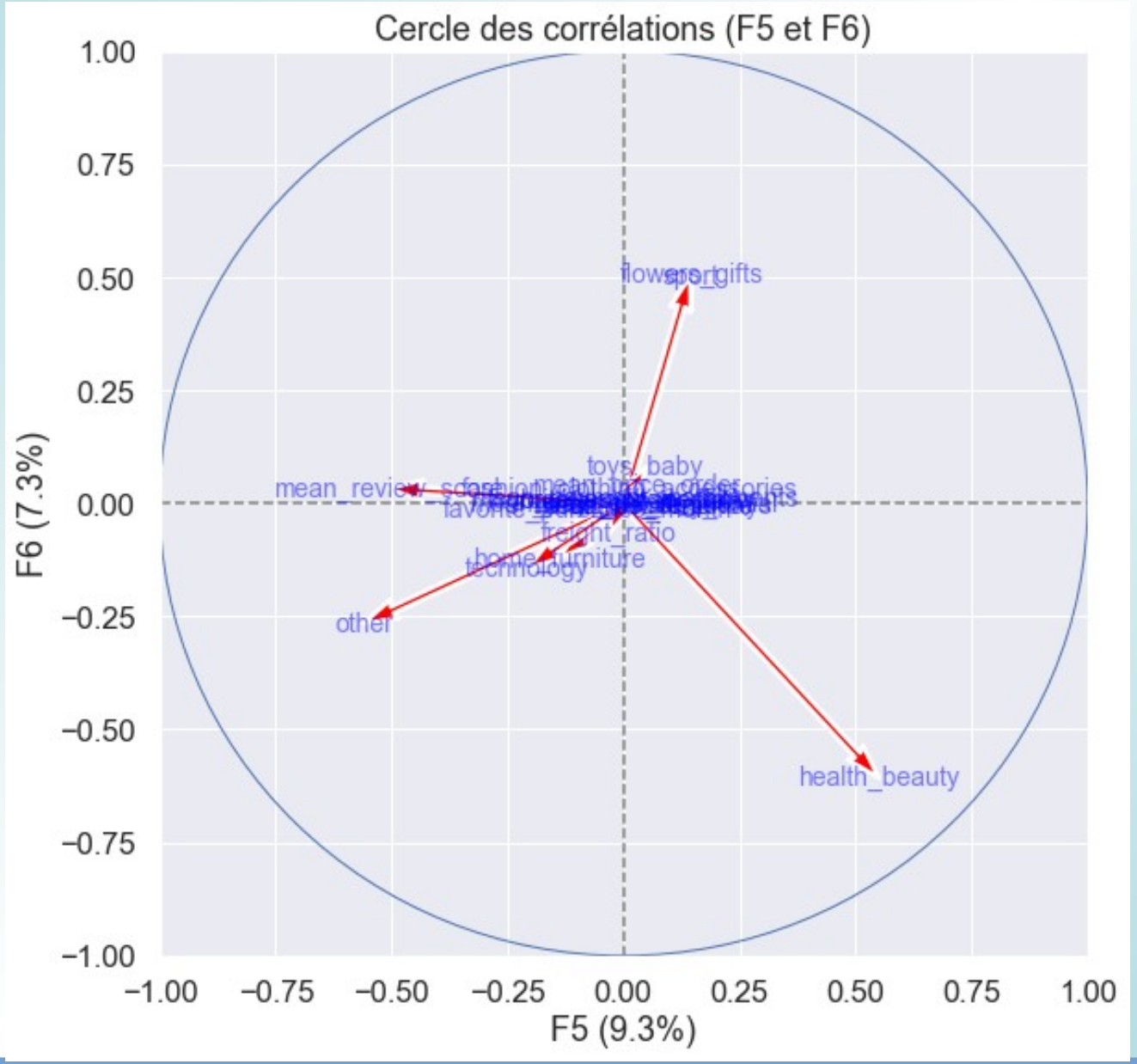
PCA – Cercle de corrélations

2ème plan factoriel



PCA – Cercle des corrélations

3ème plan factoriel



Piste de modélisation – Hypothèse 2 - Principes

Méthode de clustering K-means

- Modélisation la plus simple avec 12 variables en entrée.

- Standardisation des données.

- Détermination du nombre de clusters :
  - Par la méthode du coude en visuel **=> K entre 3 et 5**.
  - Par la méthode du coude basée sur le score de distorsion

  => K = 4

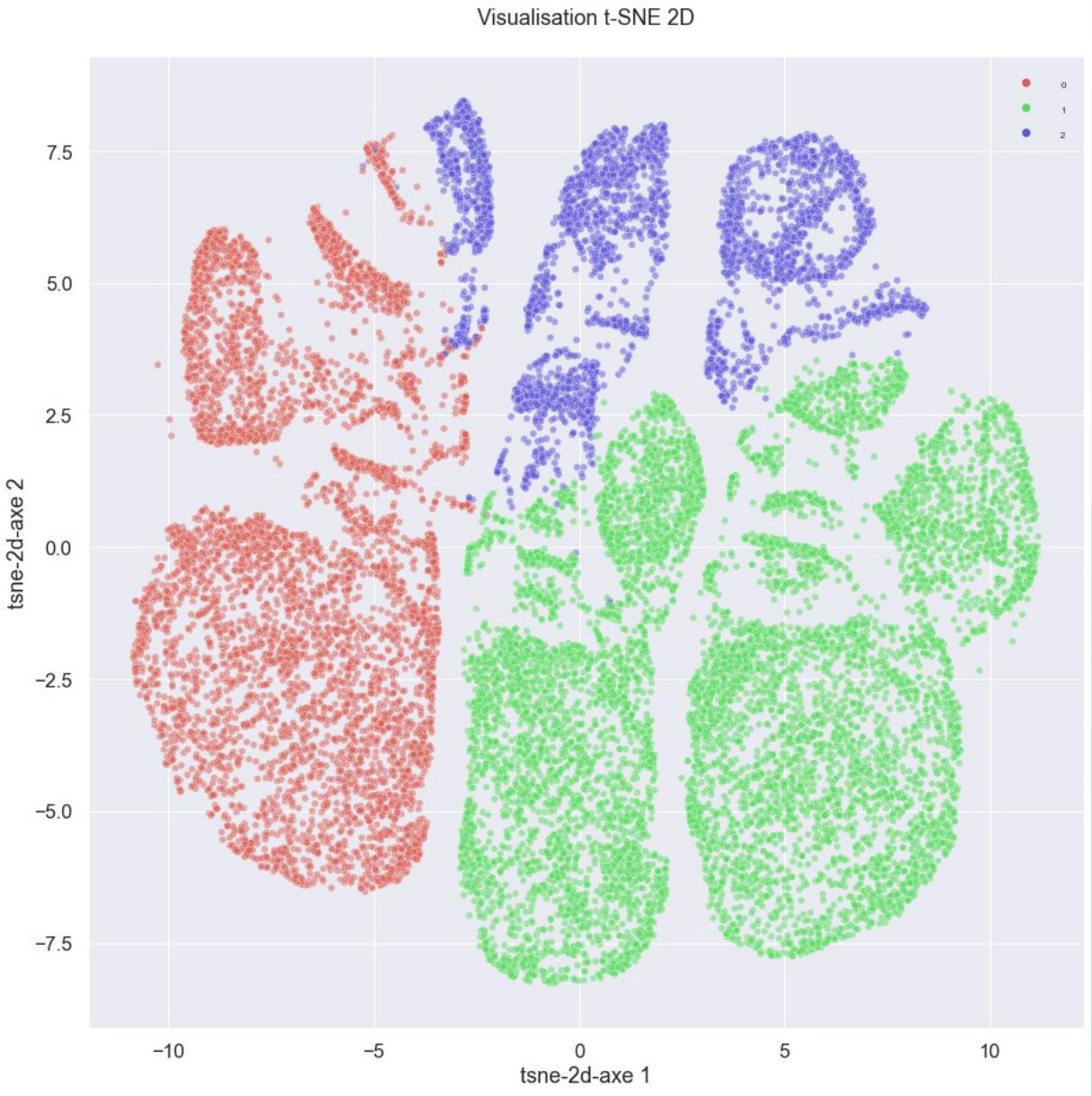
  - En trouvant le score de silhouette optimal

  => K = 3 pour un score de silhouette moyen = 0.3263

Choix du nombre de clusters = 3

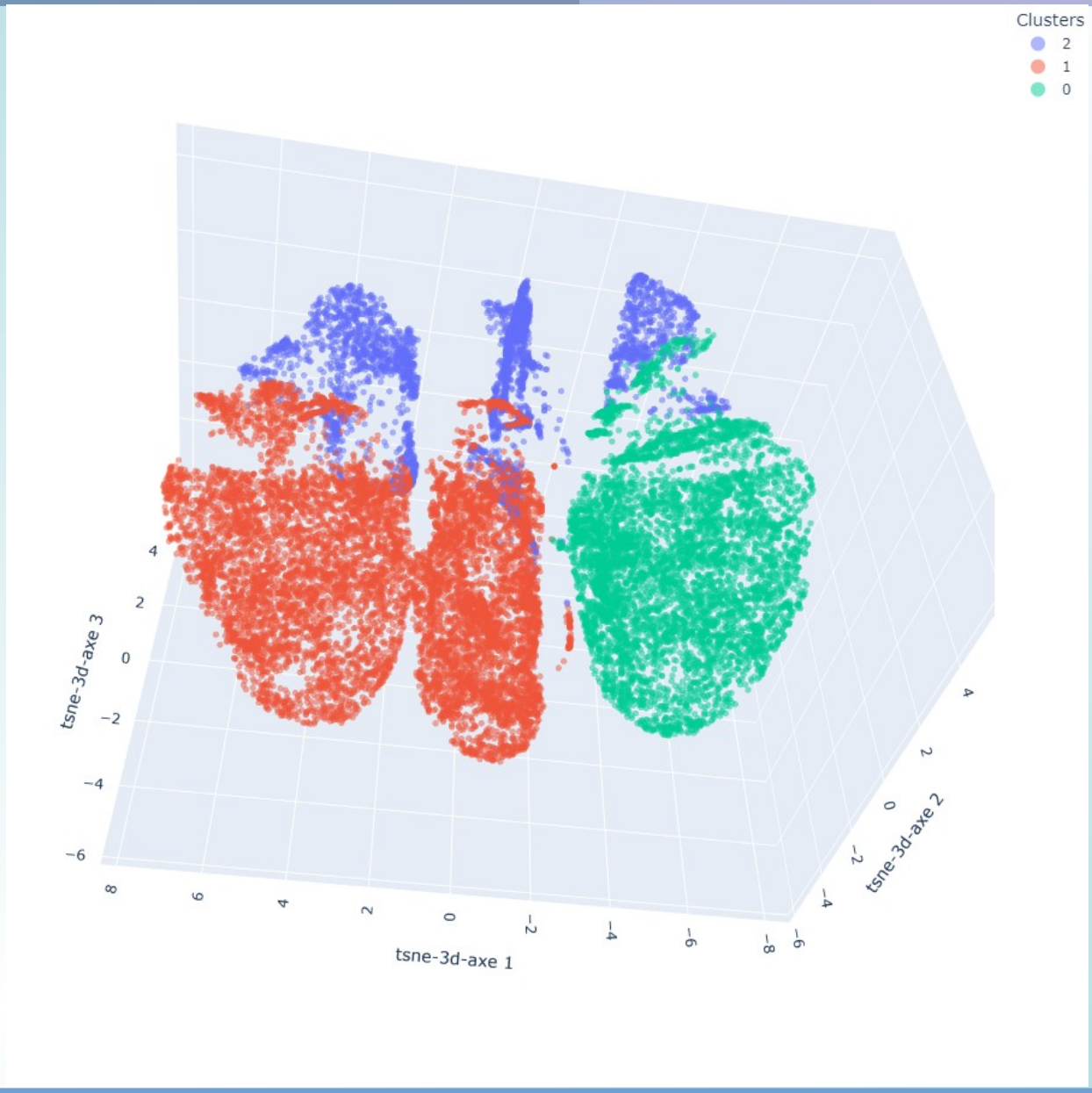
Hypothèse 2 - Visualisation T-SNE 2D

Modélisation de clustering K-means



Hypothèse 2 – Visualisation T-SNE 3D

Méthode de clustering K-means



Piste de modélisation – Hypothèse 2 - Principes

Méthode de clustering DBSCAN

- Modélisation la plus simple avec 12 variables en entrée.

- Standardisation des données.

- Détermination des paramètres :
  - Pour min_samples, application de la règle 2*dimensions

  => min_samples = 24

  - Pour epsilon (eps), application de la méthode des plus proches voisins

  => eps environ 0,3

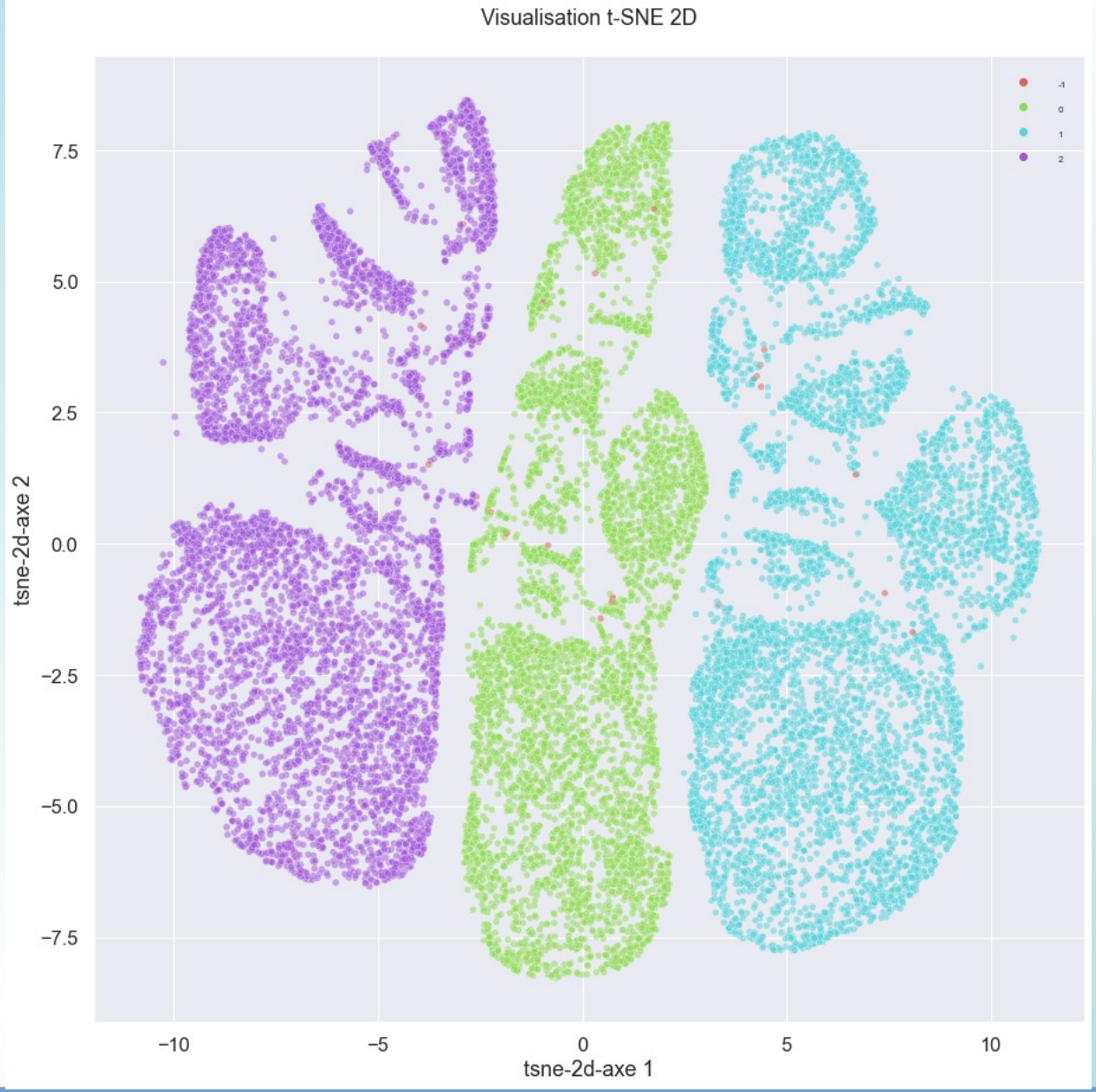
  - De manière empirique, calcul du score de silhouette optimal en calculant plusieurs modèle DBSCAN pour déterminer K et eps.

  => K = 3, eps = 0.5 pour un score de silhouette moyen = 0.252

Choix du nombre de clusters = 3 avec 31 points de bruit.

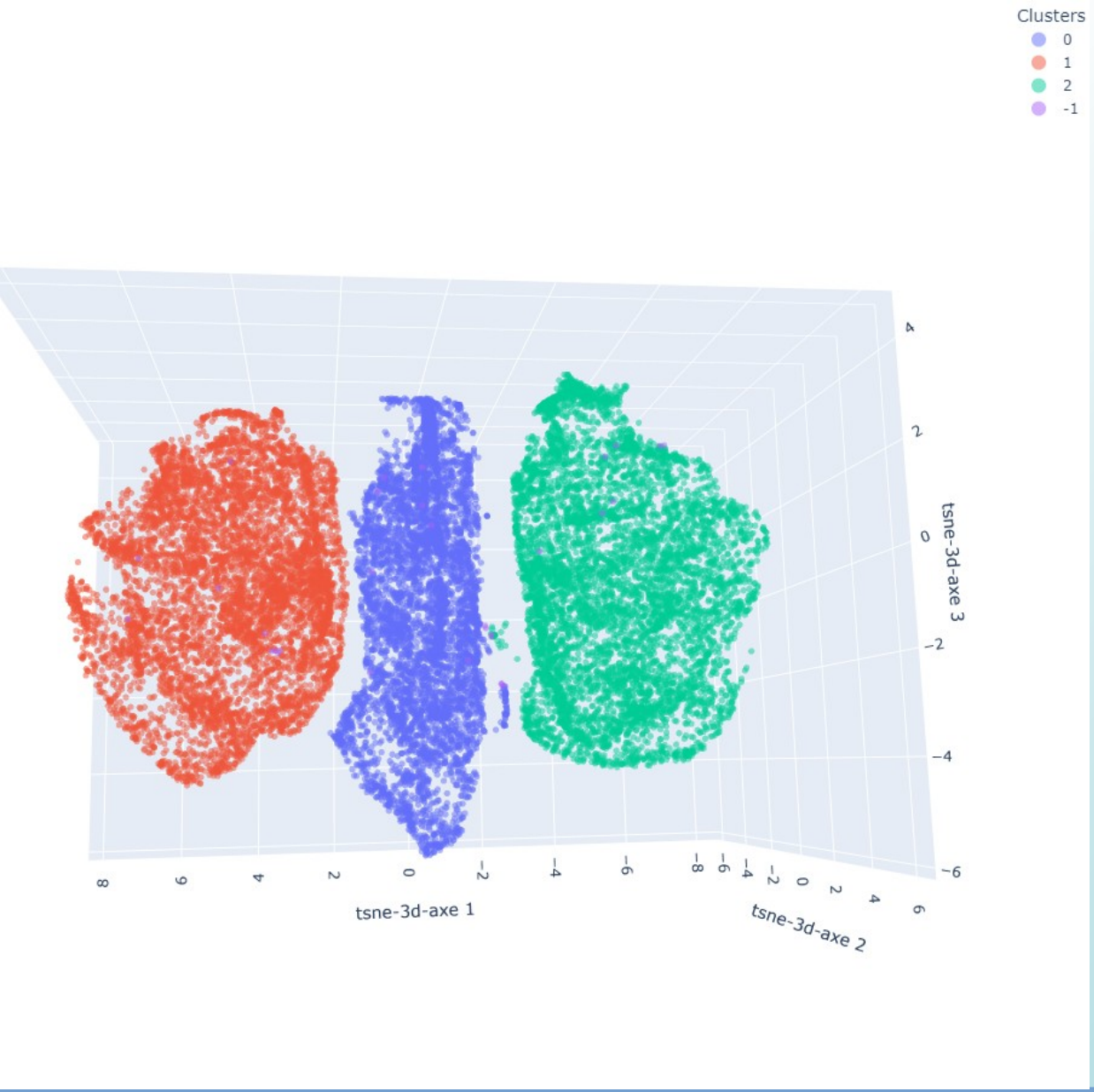
Hypothèse 2 - Visualisation T-SNE 2D

Modélisation de clustering DBSCAN



Hypothèse 2 – Visualisation T-SNE 3D

Méthode de clustering DBSCAN



Piste de modélisation – Hypothèse 2 - Principes

Méthode de clustering hiérarchique (agglomerative clustering)

- Modélisation la plus simple avec 12 variables en entrée.

- Standardisation des données.

- Détermination des paramètres :
  - Détermination du nombre de clusters K avec le dendrogram

=> K = 3 ou 5

  - Calcul du score moyen de silhouette avec K = 3 et K = 5

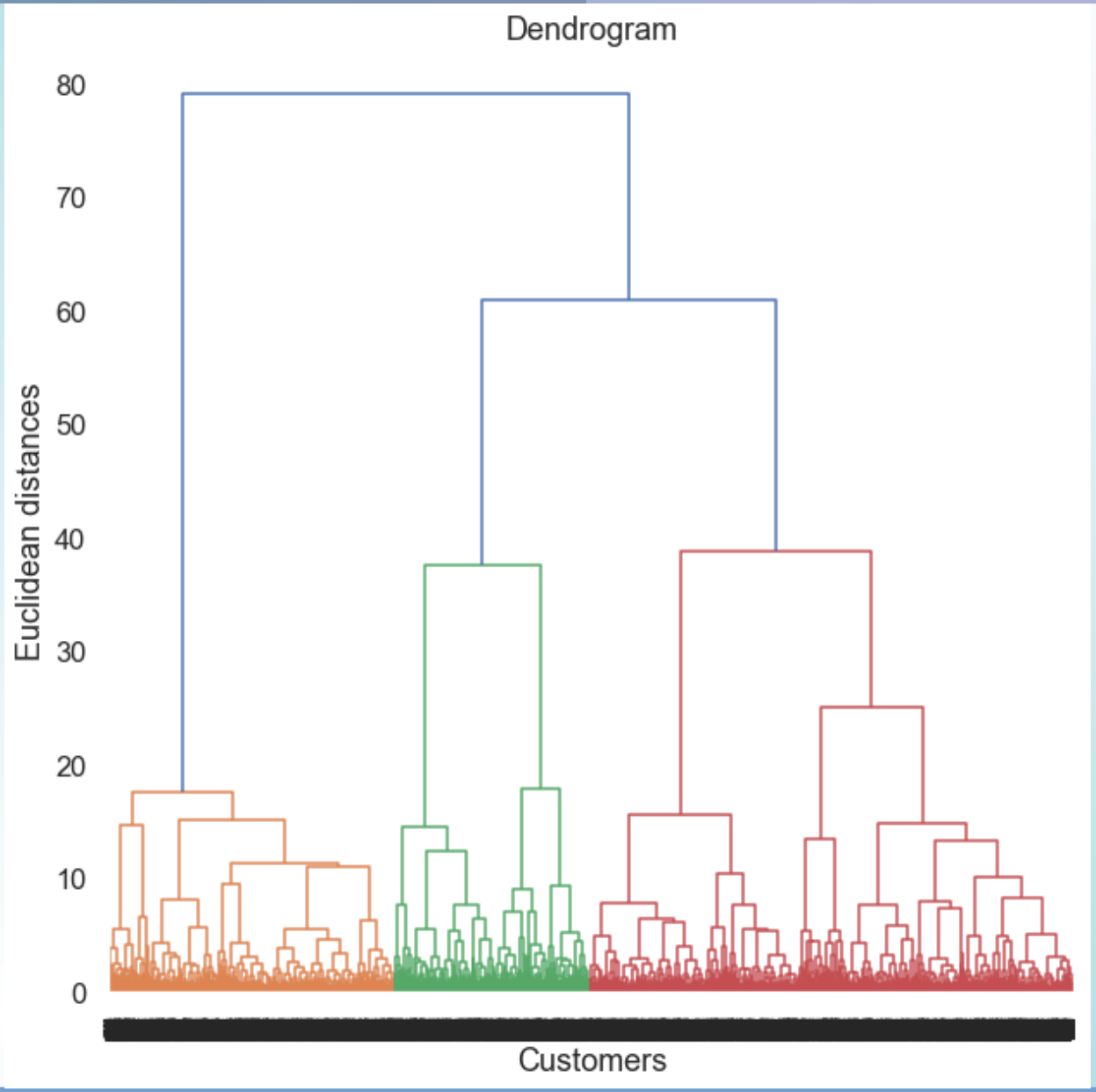
=> pour K = 3, score de silhouette moyen = 0.307

=> pour K = 5, score de silhouette moyen = 0.288

Choix du nombre de clusters = 3.

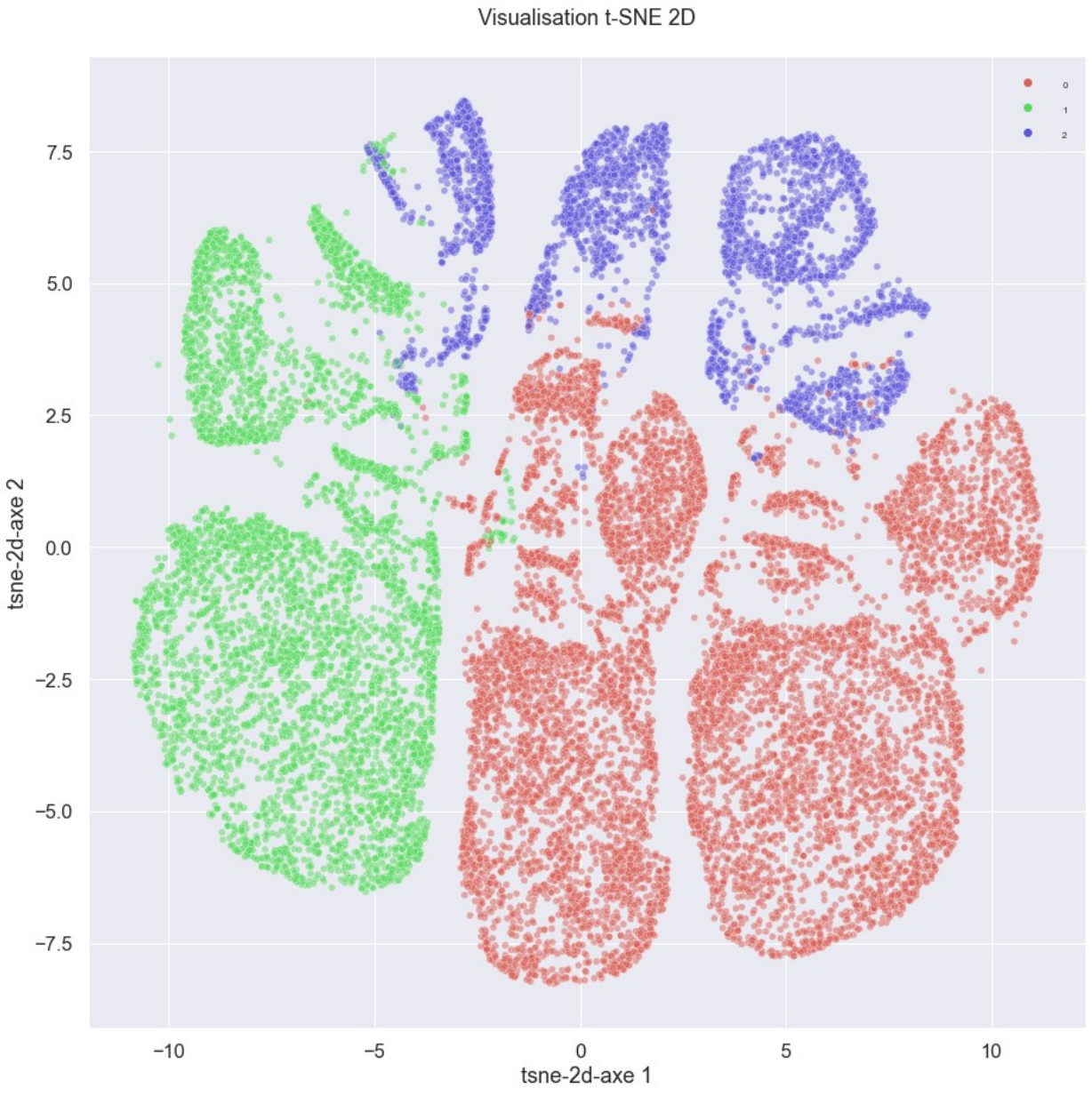
Piste de modélisation – Hypothèse 2 - Dendrogram

Méthode de clustering hiérarchique (agglomerative clustering)



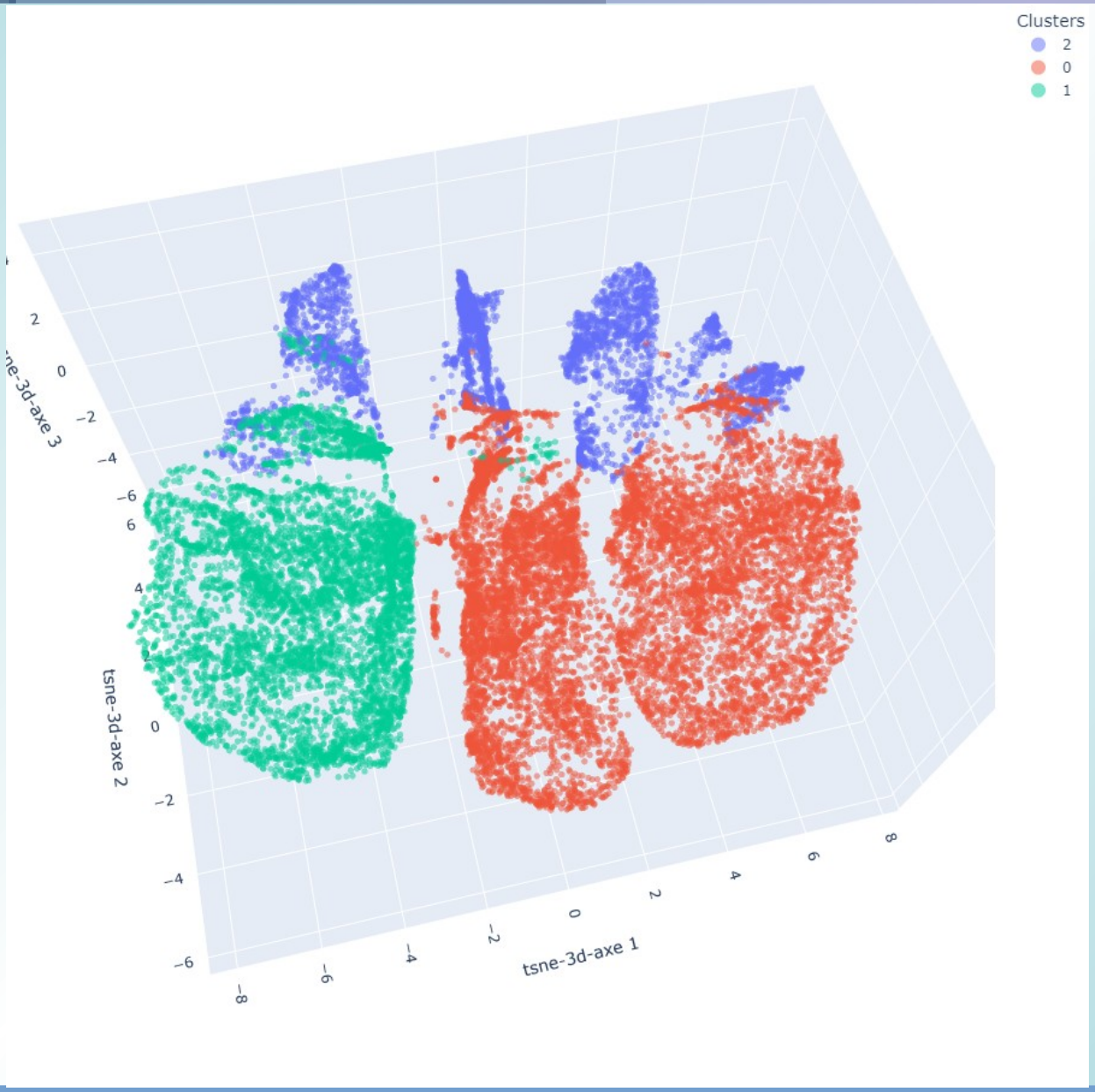
Hypothèse 2 - Visualisation T-SNE 2D

Modélisation de clustering hiérarchique (agglomerative clustering)



Hypothèse 2 – Visualisation T-SNE 3D

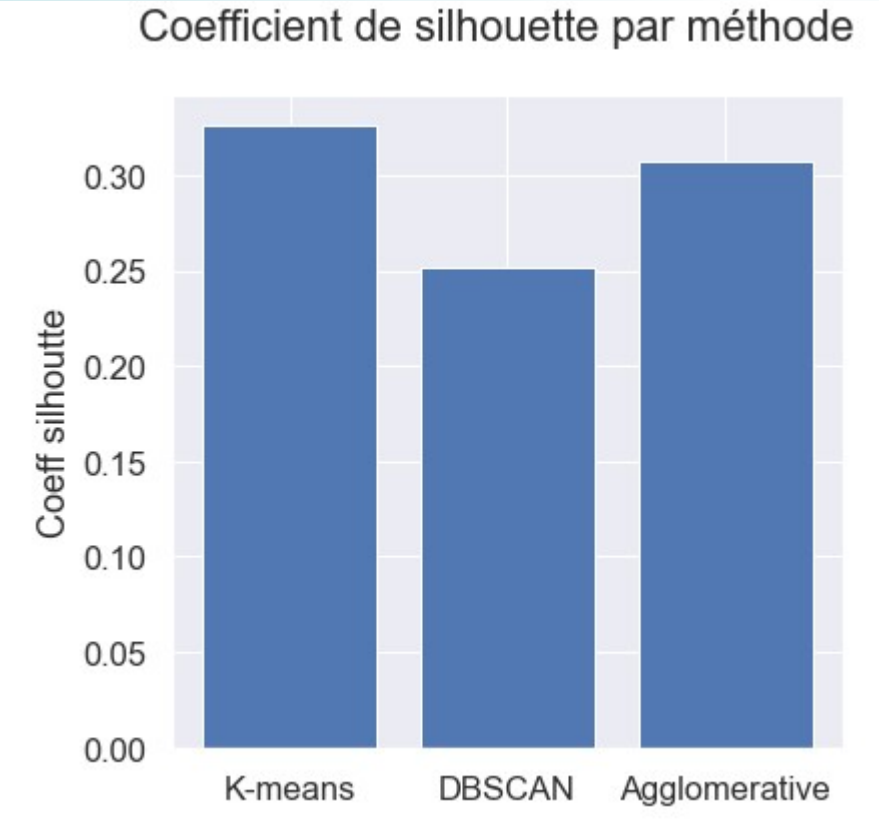
Méthode de clustering hiérarchique (agglomerative clsutering)



Piste de modélisation – Hypothèse 2

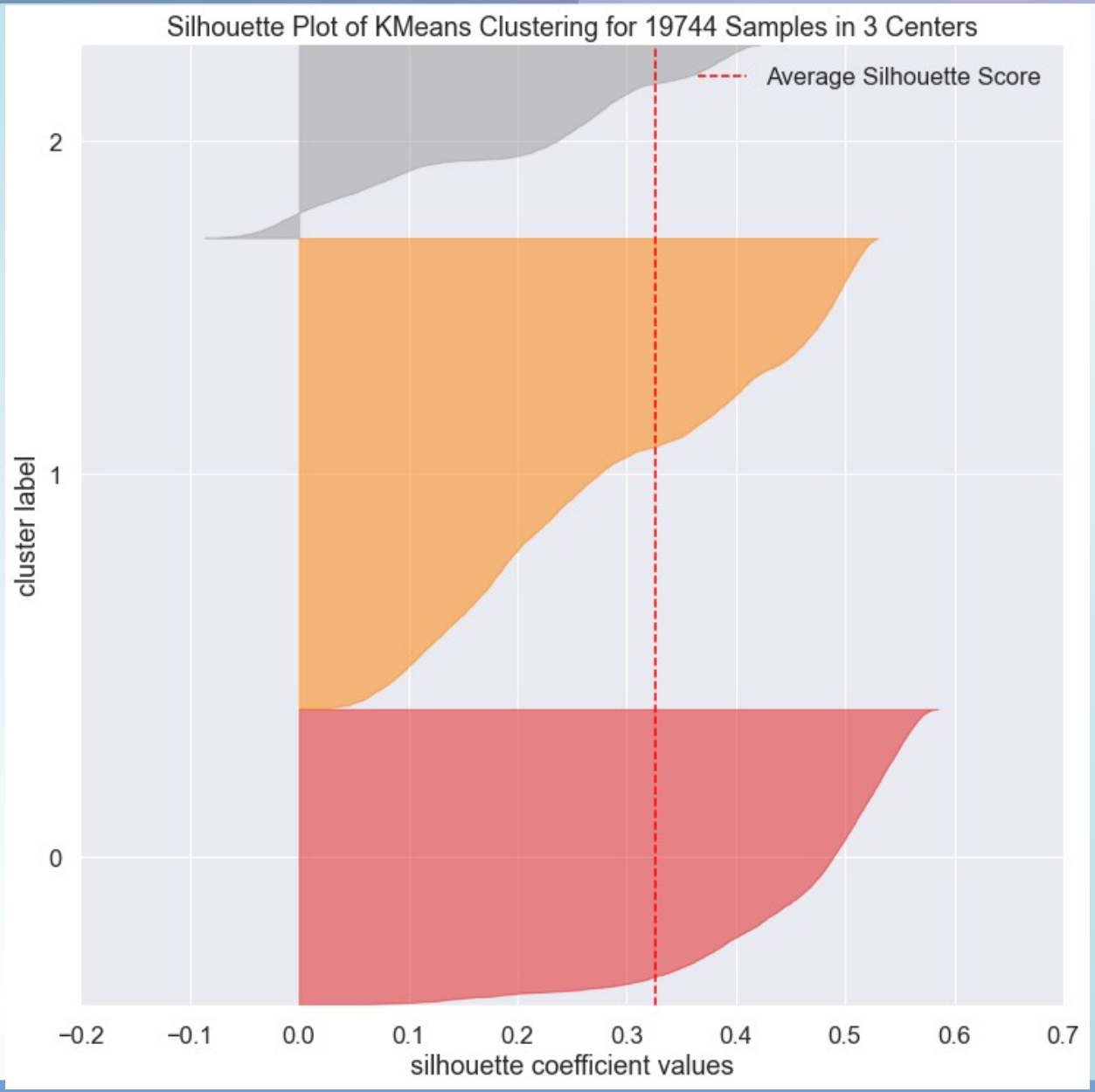
Qualité du clustering – Facteur de forme (coeff silhouette)

- Sur le coefficient de silhouette, **le modèle K-means à 3 clusters** présente le meilleur coefficient moyen de silhouette.



Piste de modélisation – Hypothèse 2 - K-means

Qualité du clustering – Représentation graphique facteur forme



Piste de modélisation – Hypothèse 2

Qualité du clustering – Séparation des clusters

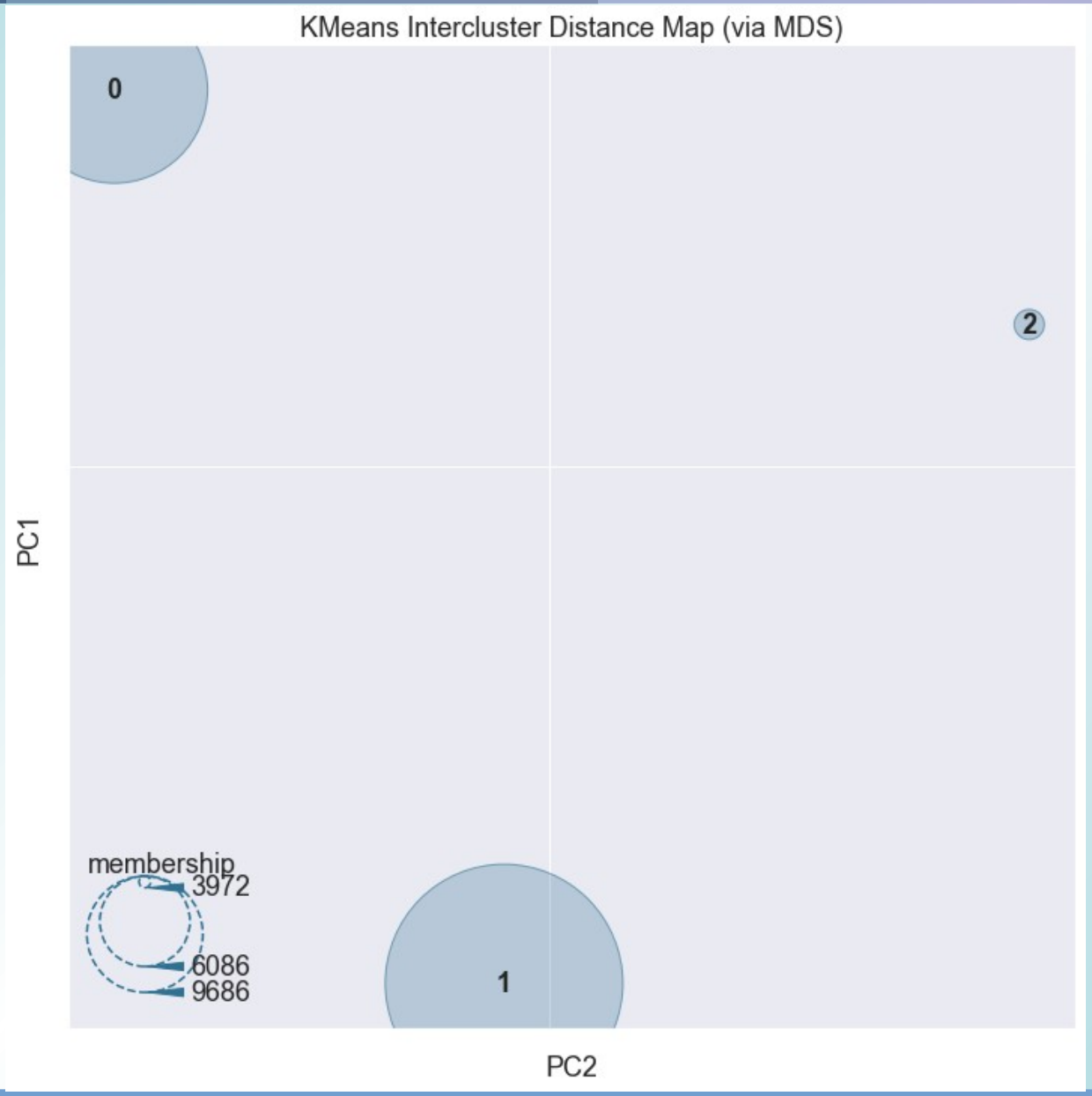
- Pour le modèle clustering hiérarchique, la visualisation T-SNE 2D/3D montre des clusters peu marqués.

- Pour le modèle DBSCAN, le clusters semblent plus distincts à la visualisation.

=> Compte tenu du score de silhouette (critère principal), **le modèle k-means à 3 clusters** est considéré comme le meilleur choix sur ce critère (voir graphe des distances inter-clusters.

Piste de modélisation – Hypothèse 2 - K-means

Qualité du clustering – Graphique distance inter-clusters



Piste de modélisation – Hypothèse 2

Qualité du clustering – Homogénéité des tailles de clusters

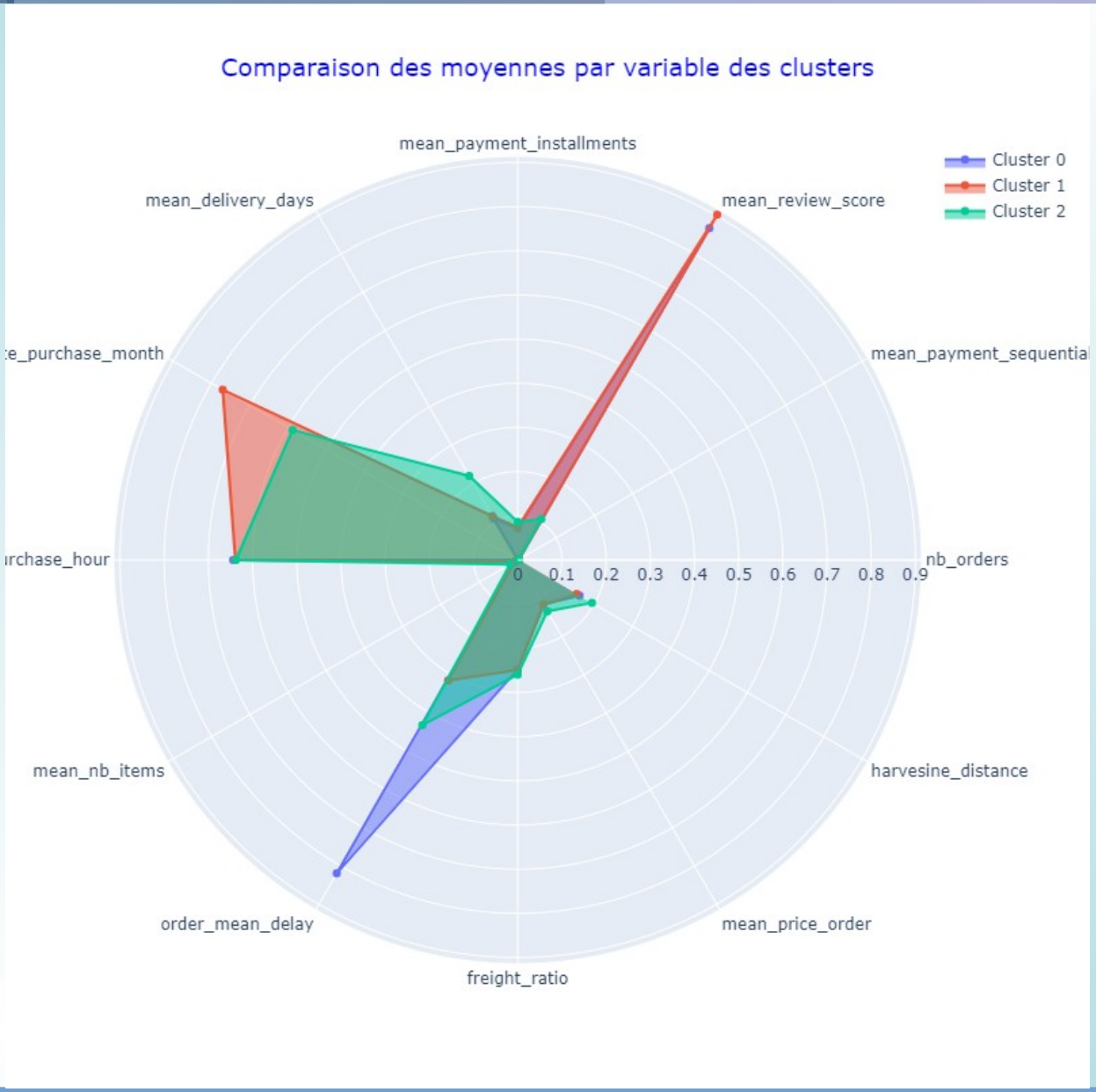
- Sur ce critère, **le modèle DBSCAN à 3 clusters** est le meilleur modèle (homogénéité presque parfaite des tailles de clusters).

Modèle	Taille cluster 0	Taille cluster 1	Taille cluster 2
K-means	6086	9686	3972
DBSCAN	6236	6672	6805
Agglomerative clustering	9899	5843	4002

=> Choix final pour la piste de modélisation 1 : le modèle K-means est retenu (meilleur modèle sur 2 critères sur 3).

Hypothèse 2 - Modèle K-means à 3 clusters

Signification métier des clusters (1/2)



Hypothèse 2 - Modèle K-means à 3 clusters
Signification métier des clusters (2/2)

- **Cluster 0** :
  - Clients satisfaits (score avis client)
  - Clients qui achètent peu souvent = **clients peu actifs**
  - Clients livrés en environ 13 jours

- **Cluster 1** :
  - Clients très satisfaits.
  - **Clients fidèles** (fréquence d'achat tous les 28 jours en moyenne sur la période).
  - Clients livrés en environ 13 jours.

- **Cluster 2** :
  - Client très insatisfaits.
  - **Clients prometteurs** (fréquence d'achat tous les 39 jours en moyenne sur la période).
  - Clients livrés en environ 25 jours.

=> Il n'y a pas de profils de clientèle très marqués avec ce modèle de clustering.

=> La piste de modélisation 2 va intégrer les catégories de produit pour essayer de faire un clustering plus précis.

Hypothèse 2 - Modèle K-means à 3 clusters

Evaluation de la stabilité temporelle du clustering (1/2)

- **Méthode :**
  - Evaluation du score de silhouette sur 5 mois glissants par pas de 1 mois.
  - Période initiale : 3 mois (du janvier à mars 2018).

- **Résultats** :
  - Mise à jour du clustering tous les 2 mois

  => perte de 6,2 % sur le score de silhouette à partir du pas 2 (période mars-avril-mai 2018)

  => Le modèle k-means à 3 clusters n'est pas stable dans le temps.

Hypothèse 1 - Modèle K-means à 3 clusters

Evaluation de la stabilité temporelle du clustering (2/2)