RAPPORT PROJET

Obesity / CVD risk

DATA MINING

Massinissa MAOUCHE 12312071
Abdennour SLIMANI 12312007
Akram MEKBAL 12312227

1. INTRODUCTION (AKRAM MEKBAL)

1.1 Contexte et objectifs du projet

L'obésité est une problématique de santé publique mondiale, ayant des implications significatives pour les individus et les systèmes de santé. Ce projet s'inscrit dans une optique de mieux comprendre les comportements alimentaires et les caractéristiques démographiques associées à l'obésité à travers une analyse de données avancée. L'objectif principal est de :

- Identifier des segments d'individus ayant des comportements similaires (via clustering),
- Détecter des relations entre comportements et types d'obésité (règles d'association),
- Explorer des communautés relationnelles et des schémas spatiaux (analyse de graphes et spatialité).

1.2 Description des données

Les données consistent en l'estimation des niveaux d'obésité chez les personnes des pays du Mexique, du Pérou et de la Colombie, avec des âges entre 14 et 61 ans et diverses habitudes alimentaires et condition physique, les données ont été collectées à l'aide d'une plate-forme Web avec une enquête où des utilisateurs anonymes ont répondu à chaque question, puis les informations ont été traitées en obtenant 17 attributs et 2111 enregistrements. Ces variables incluent :

- Variables démographiques : âge, sexe, taille, poids.
- Variables comportementales : habitudes alimentaires (FCVC, CH2O), activité physique (FAF), temps sédentaire (TUE).
- Variables catégoriques : historique familial de surpoids, moyen de transport (MTRANS), types d'obésité (NObeyesdad).

1.3 Méthodologie utilisée

Le projet a été structuré en plusieurs étapes :

1. Prétraitement des données (encodage, normalisation),
2. Analyse exploratoire (corrélations et visualisations),
3. Réduction de dimensionnalité (PCA),

4. Clustering (K-Means, DBSCAN),
5. Extraction des règles d'association (Apriori),
6. Détection de communautés (graphe de similarité, Louvain),
7. Analyse spatiale (autocorrélation avec Moran's I).

1.3 Note sur la Collaboration de Groupe

Cette répartition ne représente pas une limitation stricte des contributions, car chaque membre a également apporté des idées et des suggestions sur les différentes sections, et repose sur un consensus établi concernant la stratégie et la méthodologie appliquées à chaque partie.

2. PRETRAITEMENT DES DONNEES (MASSINISSA MAOUCHE, AKRAM MEKBAL)

2.1 Nettoyage des données

Les données ont été vérifiées pour identifier des valeurs manquantes ou aberrantes. Les variables catégoriques ont été encodées en variables binaires (via one-hot encoding). Les variables numériques ont été normalisées pour garantir une échelle uniforme entre les caractéristiques.

2.2 Normalisation et sélection

Les variables quantitatives comme Age, Height, et Weight ont été standardisées à l'aide de StandardScaler. Cela a permis d'éliminer les biais dus à des échelles différentes.

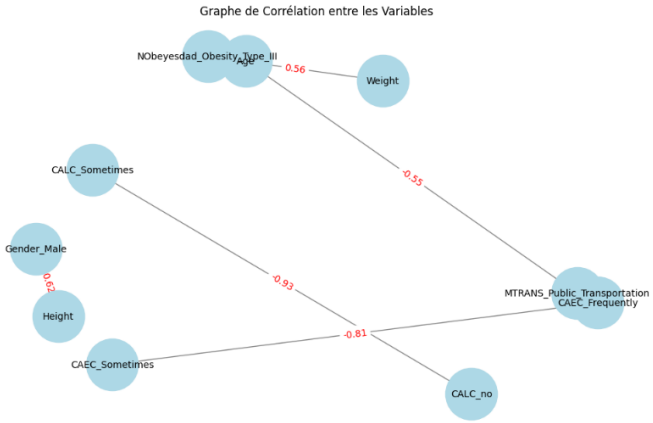
3. ANALYSE EXPLORATOIRE DES DONNEES (EDA) (MASSINISSA MAOUCHE, AKRAM MEKBAL)

- Corrélations entre les variables

Ces relations de corrélation significatives révèlent des schémas intéressants dans les données, notamment :

- Une opposition marquée entre les réponses "Sometimes" et "No" pour la consommation d'alcool (CALC), indiquant une dichotomie nette dans les comportements de consommation d'alcool.

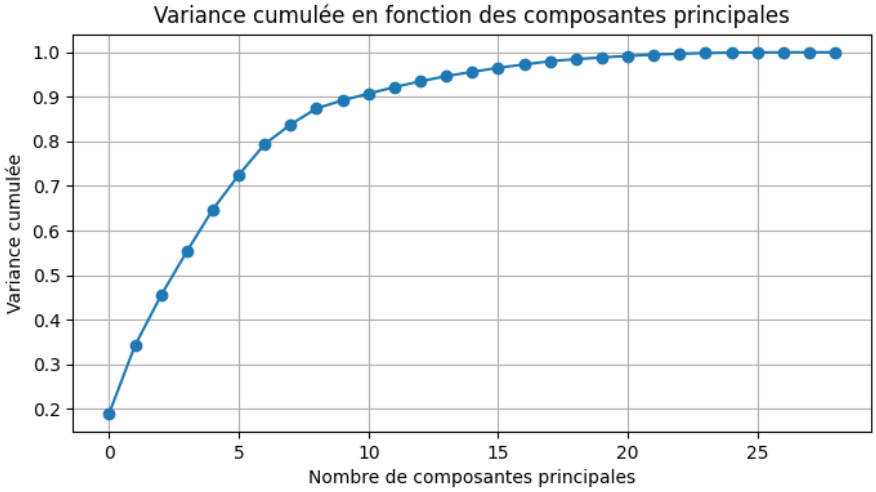
- Une opposition similaire entre les fréquences de consommation fréquente et occasionnelle de collations (CAEC), soulignant des préférences alimentaires distinctes.
- Une association positive entre la taille (Height) et le genre masculin (Gender_Male), confirmant une tendance biologique bien établie.
- Une corrélation positive entre le poids (Weight) et l'obésité sévère (NObeyesdad_Obesity_Type_III), reflétant l'influence du poids élevé dans cette catégorie d'obésité.
- Une corrélation négative entre l'âge (Age) et l'utilisation des transports publics (MTRANS_Public_Transportation), suggérant que les individus plus jeunes utilisent davantage les transports publics.



4. REDUCTION DE LA DIMENSIONNALITE (AKRAM MEKBAL, ABDENNOUR SLIMANI)

Ce résultat montre une structure complexe dans les données, nécessitant plusieurs dimensions pour conserver l'essentiel de l'information, mais pour simplifier l'analyse, on a décidé d'utiliser un nombre réduit de composantes. Dans notre cas, la réduction de dimension avec PCA a permis de :

- Faciliter l'application des algorithmes de clustering
- Fournir une visualisation claire des groupes détectés

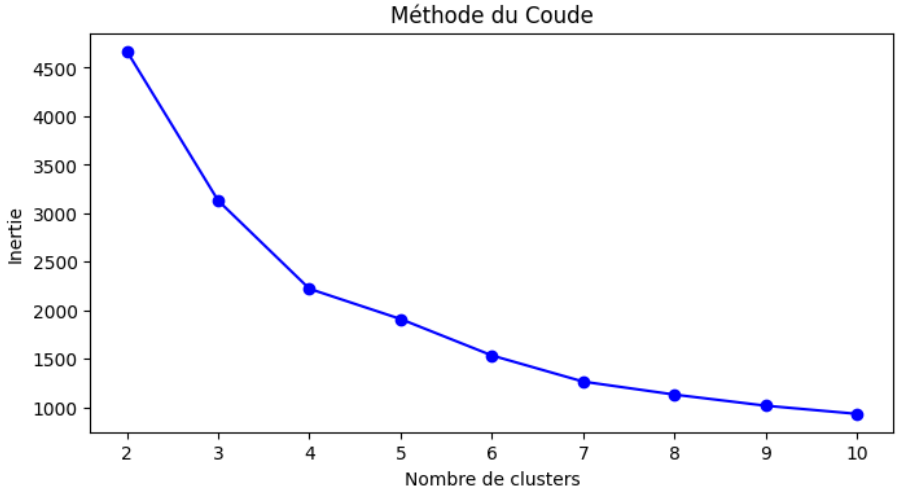


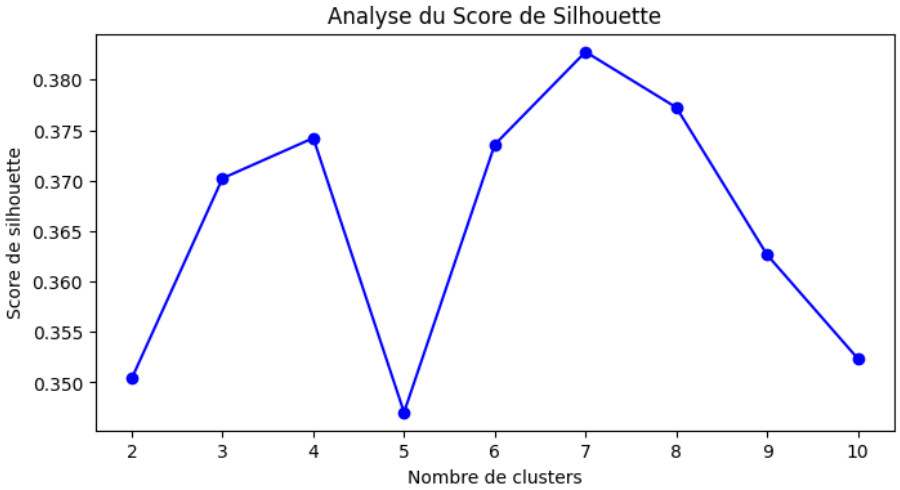
5. SEGMENTATION DES DONNEES

5.1 Clustering avec K-Means (Massinissa MAOUCHE, Abdennour SLIMANI)

Méthode du Coude

La méthode du coude indique un point d'inflexion entre 7 et 6 clusters, suggérant un nombre optimal de groupes. Le modèle K-Means a été configuré avec **7 clusters**, correspondant au score de silhouette maximal.

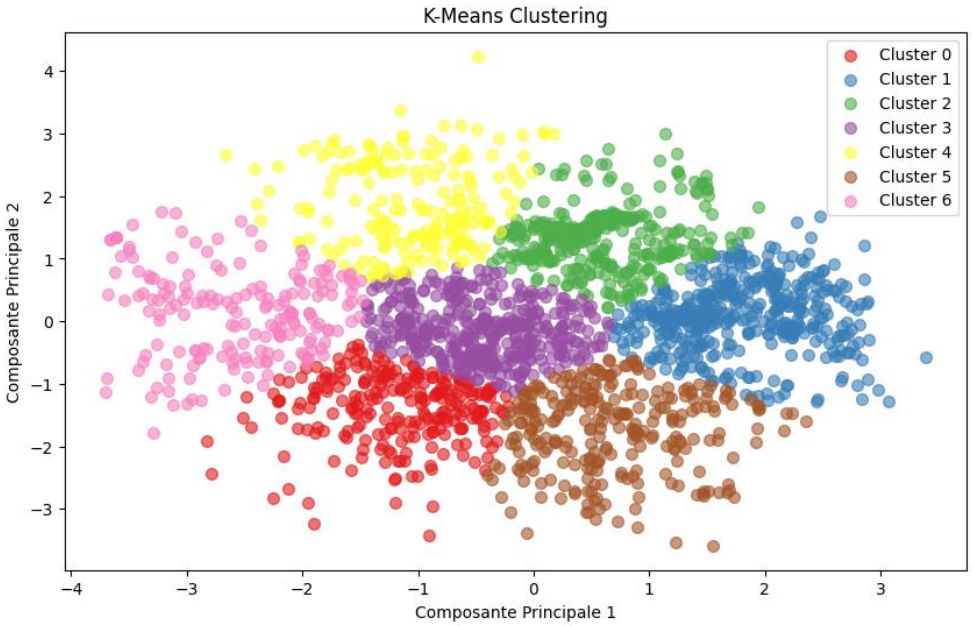




Résultats

- Cluster 0 : Principalement constitué de mâles, avec un historique familial d'obésité modéré (57.6 %). Majoritairement composé d'individus avec un poids normal (30.9 %) et un usage intensif des transports publics.
- Cluster 1 : Fortement dominé par des individus de sexe masculin (69.7 %) ayant des habitudes de marche limitées. Comprend un pourcentage élevé d'obésité de type II (40.5 %) et de type III (30.3 %).
- Cluster 2 : Faible activité physique et hydratation légèrement déficiente.
- Cluster 3 : Ce cluster présente les individus avec des comportements alimentaires diversifiés et une utilisation modérée des moyens de transport publics.
- Cluster 4 : Composé majoritairement de personnes en surpoids ou avec une obésité de type I ou II.

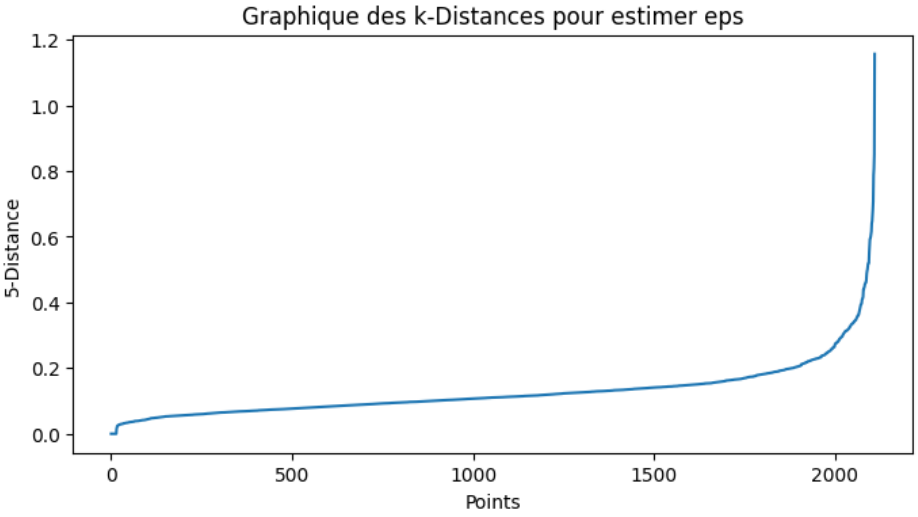
- Cluster 5 : Le sexe masculin est fortement représenté (86.4 %), avec une population ayant principalement un poids normal ou un surpoids léger, actifs physiquement avec une bonne hydratation.
- Cluster 6 : Constitué majoritairement de jeunes ayant une prévalence moindre d'obésité, mais souvent associés à des poids insuffisants.



5.2 Clustering avec DBSCAN (Akram MEKBAL, Abdennour SLIMANI)

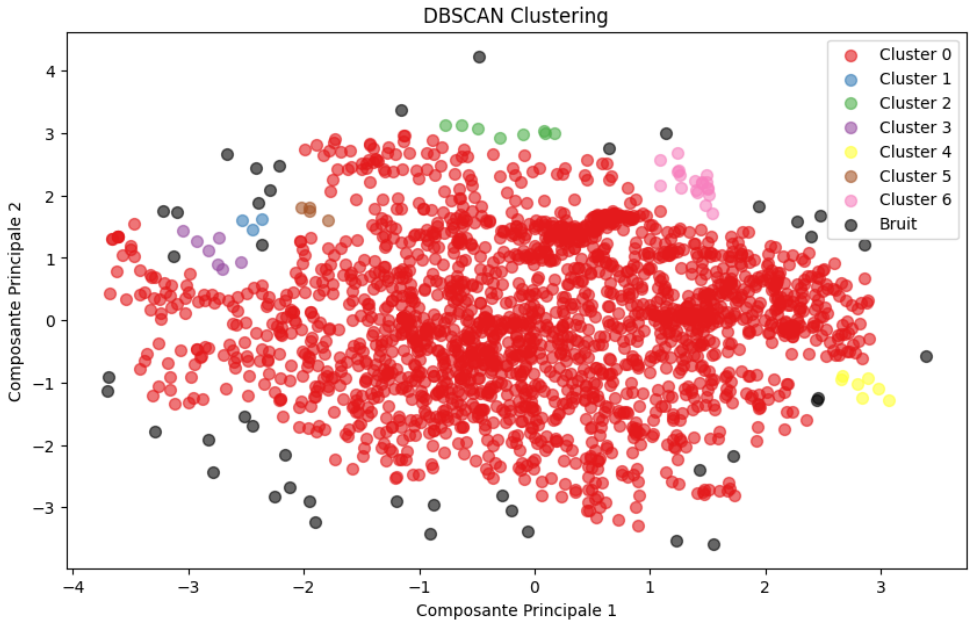
Graphique des k-Distances

Le choix de eps=0.25 a permis de bien discriminer les groupes dans un espace complexe, révélant des communautés distinctes sans compromettre les zones denses.



Résultats

- Les clusters identifiés par DBSCAN complètent l'analyse en capturant des sous-groupes potentiellement ignorés par K-Means.
- **Cluster 3** regroupe des individus avec des valeurs d'obésité modérées (Overweight_Level_II), **Cluster 1** : Âge moyen élevé, poids négatif, et faibles valeurs pour les habitudes alimentaires, **Cluster 4** : Très grande taille moyenne et poids élevé, probablement lié à un groupe spécifique, **Cluster 6** : Haute corrélation avec les individus masculins et lié à une grande taille.
- DBSCAN a également permis d'identifier des individus atypiques (bruit), ce qui est utile pour les analyses futures ou la gestion des exceptions.

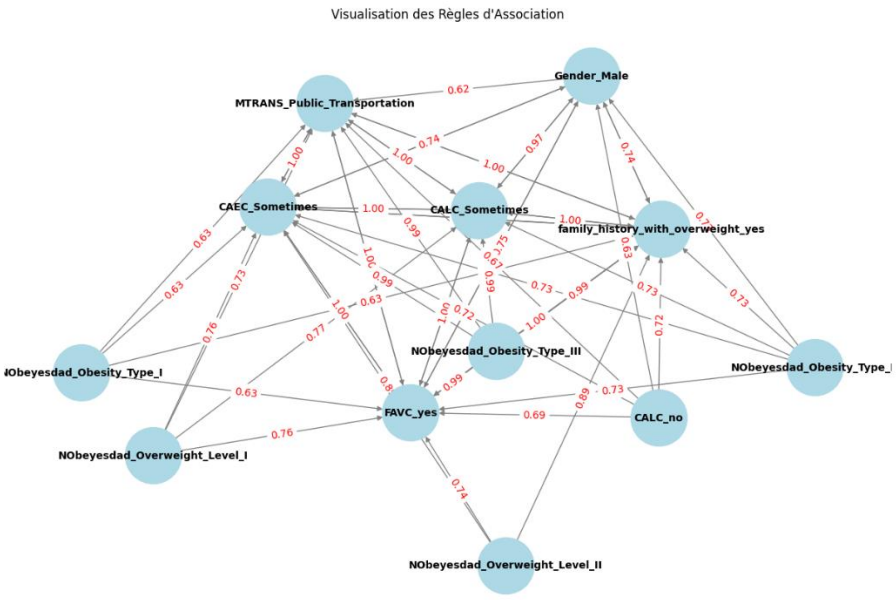


6. REGLES D'ASSOCIATION (MASSINISSA MAOUCHE, ABDENNOUR SLIMANI)

6.1 Extraction et visualisation des règles

Le graphe montre des relations significatives entre les variables binaires, notamment les habitudes alimentaires et le type d'obésité. Par exemple :

- La présence d'une consommation fréquente de collations (CAEC_Sometimes) est fortement associée à certains types d'obésité (Obesity_Type_I, Obesity_Type_III).
- Les hommes (Gender_Male) ayant un historique familial de surpoids (family_history_with_overweight_yes) tendent à être associés à des niveaux élevés d'obésité.



Ces relations fournissent des indications importantes sur les comportements et caractéristiques liés à l'obésité, pouvant être utiles pour des recommandations de santé publique. L'exécution du code permet d'avoir deux fichier csv : **association_rules.csv** et **frequent_itemsets.csv**.

6.2 Limites des Motifs Fréquents et Système de Recommandation

Dans une tentative d'utiliser les **motifs fréquents** identifiés pour construire un système de recommandation, nous avons observé que cette approche n'était pas pertinente dans ce contexte. Le système aurait simplement proposé des recommandations basées sur les comportements opposés à ceux déjà observés.

Exemple :

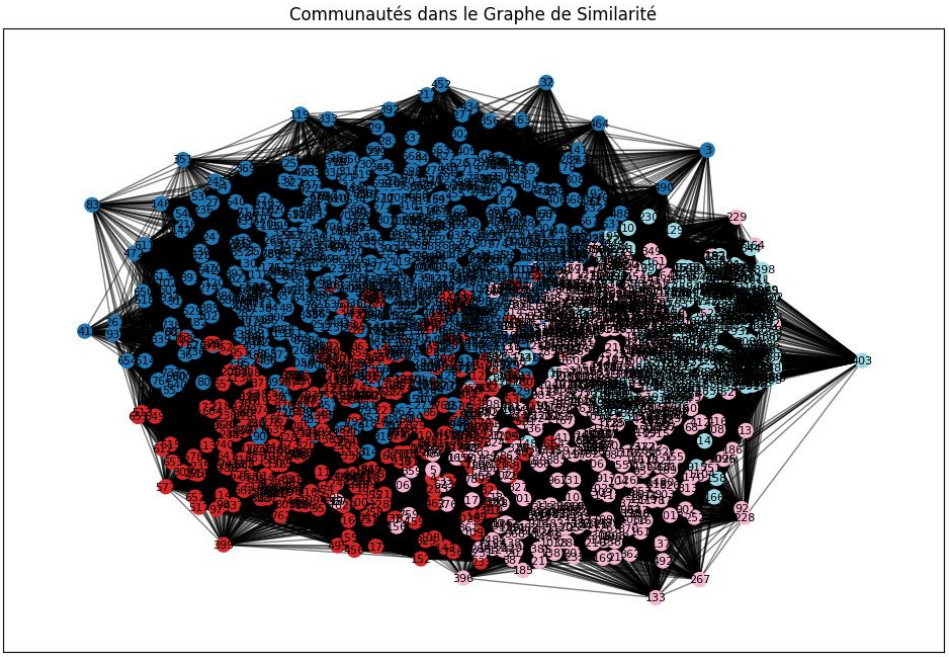
- Un individu ayant une consommation élevée de collations (CAEC_Frequently) serait conseillé de passer à une consommation occasionnelle ou nulle.

Intuitive, mais reste trop basique pour être réellement utile. Elle ne tient pas compte des facteurs plus complexes qui influencent les comportements, tels que les préférences ou les besoins médicaux. Un système de recommandation plus sophistiqué, intégrant des modèles prédictifs ou une approche hybride combinant les clusters identifiés, les règles d'association pour proposer des recommandations adaptées et efficaces serait plus adapté.

7. ANALYSE DE SIMILARITE ET COMMUNAUTES (ABDENNOUR SLIMANI, AKRAM MEKBAL)

7.1 Graphe de Similarité

Le graphe de similarité montre des communautés bien définies basées sur les relations cosinus entre les données. L'algorithme de Louvain a permis de détecter 4 communautés principales.



7.2 Analyse des communautés

Les moyennes des variables pour chaque communauté montrent des différences marquées :

- **Communauté 0** : Les individus de cette communauté semblent être dans un groupe relativement homogène, souvent associés à un faible niveau d'activité physique.
- **Communauté 1** : Des individus plus sédentaires ou à des zones urbaines.
- **Communauté 2** : Cela pourrait représenter un groupe plus âgé avec une faible mobilité et des habitudes alimentaires légèrement dégradées.
- **Communauté 3** : Cette communauté semble correspondre à un groupe avec des comportements à haut risque, nécessitant des interventions spécifiques.

Ces résultats mettent en évidence des segments bien définis parmi la population analysée. Les groupes varient en termes d'âge, d'habitudes alimentaires et d'activités physiques, ce qui offre une opportunité d'adopter des stratégies ciblées pour améliorer la santé et les comportements nutritionnels.

8. Analyse Spatiale (Abdennour SLIMANI, Massinissa MAOUCHE)

8.1 Moran's I

La faible valeur de Moran's I proche de 0, combinée à une p-valeur supérieure au seuil de signification (généralement 0.05), suggère que les attributs spatiaux ne présentent pas de regroupement géographique notable. Ces résultats indiquent une absence d'autocorrélation spatiale significative pour la variable analysée (Weight).

8.2 Moran's Scatterplot

Le Moran's Scatterplot visualise cette faible autocorrélation, avec des points dispersés de manière uniforme autour des quadrants, et une pente proche de zéro pour la régression linéaire. Cela confirme que la variable ne présente pas de structure spatiale évidente.

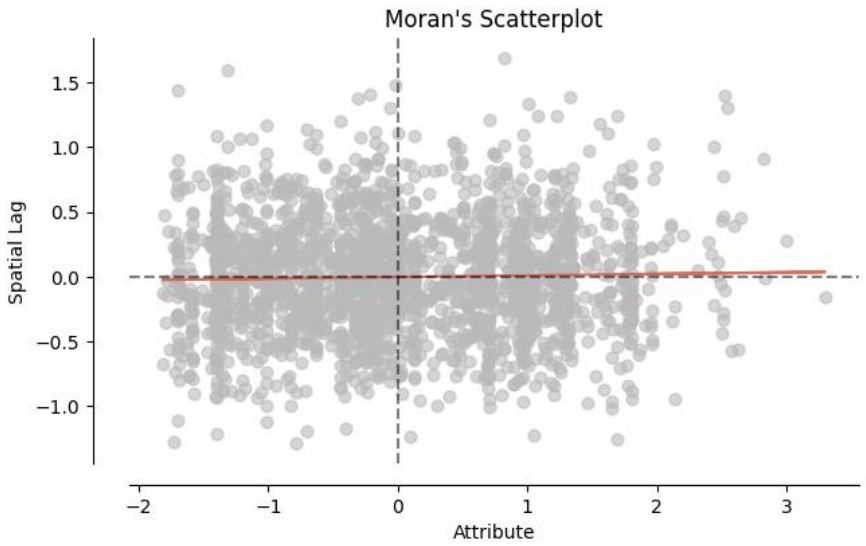


Figure 1: Moran's scatterplot

8.3 Limites

Les données géographiques utilisées dans cette analyse ont été générées artificiellement, ce qui limite la portée des conclusions. Toutefois, ces résultats pourraient révéler des schémas pertinents si des coordonnées réelles étaient utilisées. Cela ouvre des perspectives pour une analyse spatiale approfondie avec des données géographiques authentiques.

Note : la valeur de Moran's I change avec chaque exécution.

9. SYNTHESE ET CONCLUSIONS (MASSINISSA MAOUCHE)

9.1 Résumé des résultats principaux

1. Les clusters identifiés par K-Means et DBSCAN mettent en évidence des groupes distincts basés sur des variables comportementales et démographiques.
2. Les règles d'association révèlent des corrélations fortes entre les habitudes alimentaires et les types d'obésité.
3. L'analyse de similarité et des communautés met en lumière des segments sociaux ayant des comportements spécifiques.

9.2 Recommandations

- Intégrer des données géographiques réelles pour explorer davantage la dimension spatiale.
- Approfondir l'analyse avec des algorithmes non supervisés comme UMAP ou t-SNE.

9.3 Perspectives

Le projet pourrait être enrichi par :

- Une intégration de données socio-économiques.
- Une validation croisée des clusters détectés.
- L'automatisation des pipelines pour une application à des données en temps réel.