Analyse des données démographiques

Awatef Edhib, Aida Haouas, Rida Asri

November 2024

1 Introduction

L'analyse des données démographiques est un domaine crucial qui permet de mieux comprendre les dynamiques de population à l'échelle mondiale. Ces informations sont essentielles pour anticiper les changements démographiques futurs. Cette étude explore les données démographiques en utilisant une variété d'approches analytiques afin de dégager des insights pertinents.

Dans cette étude, nous avons employé plusieurs techniques avancées pour analyser les données démographiques. Les méthodes utilisées incluent :

- **L'analyse temporelle**
- **L'analyse des patterns fréquents**
- **Clusters**
- **Les systèmes de recommandation**

2 Méthodologie

2.1 Exploration des données

Le jeu de données utilisé dans cette analyse provient de Gapminder. Ces données couvrent la période de 1950 à 2023 et concernent un large éventail de pays, offrant ainsi une vue d'ensemble globale des tendances démographiques et sanitaires. Le jeu de données comprend plusieurs variables clés qui mesurent divers aspects démographiques et sanitaires. Les principales variables incluent :

- **Population totale :** Mesurée en milliers, cette variable indique la taille de la population de chaque pays au début de chaque année.
- **Taux de natalité :** Exprimé en naissances pour 1 000 habitants, il reflète la fréquence des naissances dans un pays.
- **Taux de mortalité :** Mesuré en décès pour 1 000 habitants, il renseigne sur la fréquence des décès dans la population.
- **Espérance de vie :** Cette variable donne l'espérance de vie moyenne à la naissance, séparée pour les hommes et les femmes, et fournit une indication de la santé générale de la population.
- **Autres variables :** Le jeu de données inclut également des informations sur la fertilité, les naissances par tranche d'âge, et la mortalité infantile, entre autres.

Pour une bonne compréhension des données nous avons procédé de la mabière suivante:

- **Nettoyage des données**: Les valeurs manquantes ont été remplies par la moyenne des colonnes concernées pour garantir la continuité des données.

- **Matrice de corrélation**: Une matrice de corrélation a été générée pour examiner les relations entre les variables. La matrice a montré que l'espérance de vie, différenciée par sexe et âge, est fortement corrélée positivement, tandis que les taux de mortalité (infantile, avant 5 ans, etc.) forment un autre cluster cohérent. Une corrélation négative est observée entre l'espérance de vie et les taux de mortalité. De plus, le taux de fertilité total est négativement lié à l'espérance de vie et positivement aux taux de mortalité infantile.

- **Graphiques interactifs:**
  * Évolution de la population mondiale : Un graphique interactif figure1 montre une augmentation régulière et soutenue au fil des décennies, passant d'environ 2 milliards à plus de 8 milliards . Cette tendance reflète une croissance démographique mondiale marquée.

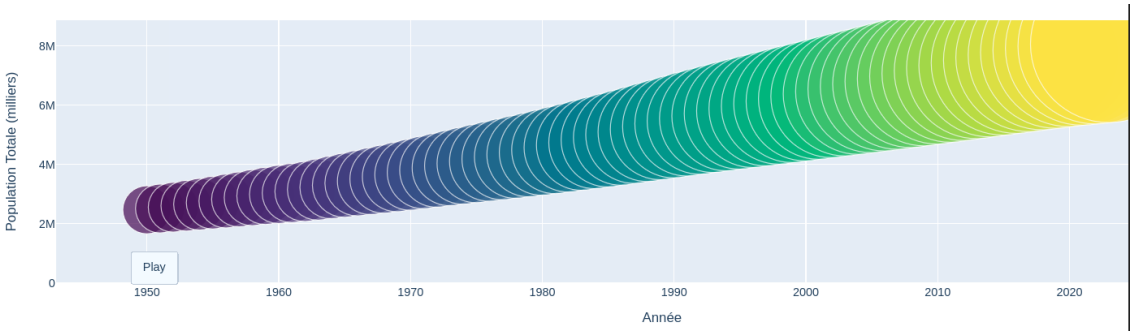


Figure 1: Évolution de la population mondiale au fil de temps

* Espérance de vie : une distribution unimodale de l'espérance de vie à la naissance, montrée par la figure 2, avec une concentration élevée autour de 70 ans. Cela indique que la majorité de la population a une espérance de vie autour de cette valeur. La distribution est asymétrique, légèrement étalée vers la gauche, ce qui suggère une minorité de personnes ayant une espérance de vie inférieure à la moyenne..

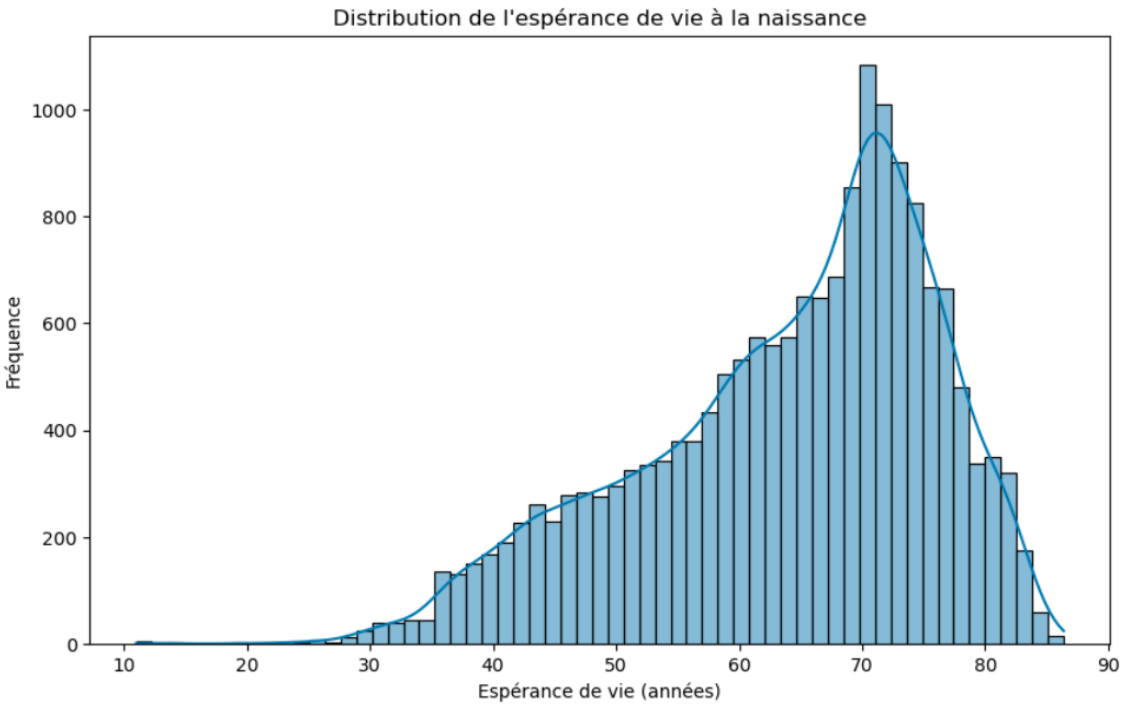


Figure 2: Distribution de l'espérance de vie à la naissance

* Taux de natalité et de mortalité : La figure 3 montre une nette différence entre les deux courbes : le taux de natalité est significativement plus élevé que le taux de mortalité sur toute la période, ce qui traduit une croissance démographique globale.

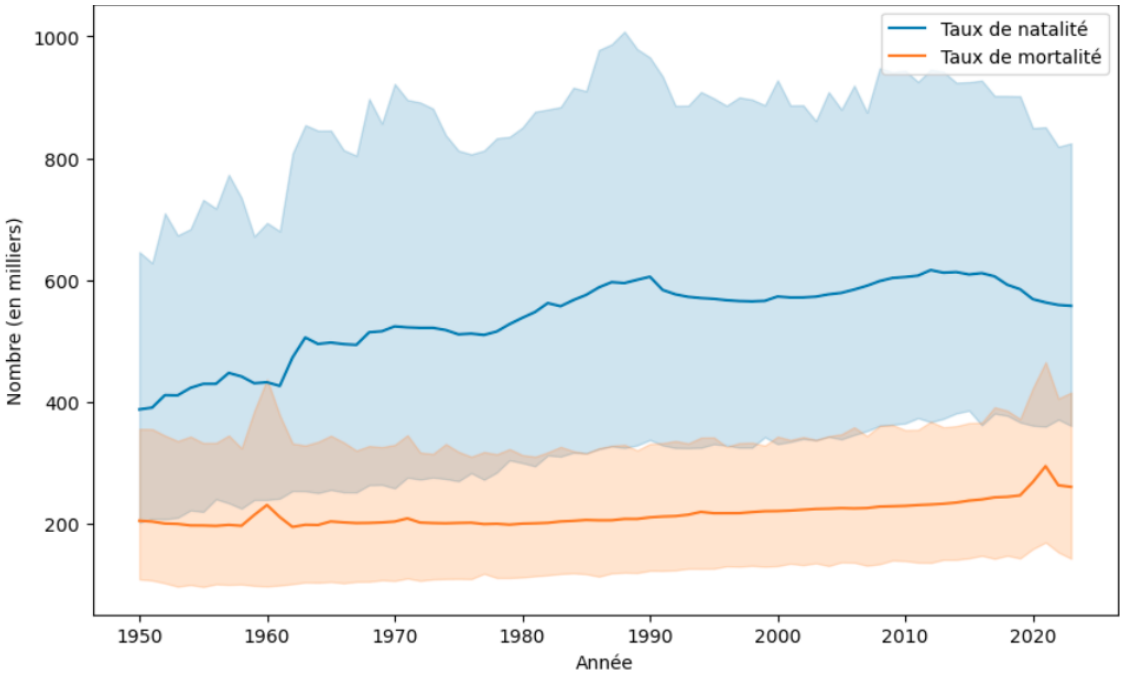


Figure 3: Évolution de la natalité et de mortalité au fil du temps

2.2 Analyse temporelle et Frequent Pattern

2.2.1 Analyse Tomporelle

Pour une analyse temporelle approfondie nous avons réalisé une décomposition des séries temporelles, analysé l'autocorrélation pour identifier des relations dans le temps et des clusters ont été formés avec DBSCAN.

- **Série temporelle**
  Les séries temporelles sont étudiées en commençant par le calcul de la moyenne de la population mondiale pour chaque année, ce qui donne lieu à une série temporelle globale. Ensuite, on procède à une décomposition additive de cette série, avec une période de 1 an réglée. Grâce à cette décomposition, on peut identifier les éléments clés : la tendance, la saisonnalité (bien que peu probable dans ce contexte annuel), et les résidus. Un graphique composite présente les résultats, permettant d'observer de manière claire l'évolution de la population mondiale au fil du temps. Afin d'approfondir l'analyse de la structure temporelle des données, on calcule et trace une fonction d'autocorrélation (ACF) pour la série de population mondiale. l'ACF est visualisée pour 20 lags (décalages temporels). Ce graphique permet d'identifier les corrélations significatives entre les valeurs de la série à différents intervalles de temps, mettant en évidence les motifs ou les cycles potentiels dans l'évolution de la population mondiale.
- **Clustering avec DTW et DBSCAN**
  Cette étude applique la méthode *Dynamic Time Warping* (DTW) et l'algorithme *DBSCAN* pour effectuer une analyse de clustering sur des séries temporelles de population. Les séries temporelles, regroupées par pays, sont converties en un format tridimensionnel (échantillons, séquences temporelles, caractéristiques).
  Afin de permettre une comparaison équitable entre les pays, les séries sont normalisées en ajustant la moyenne à 0 et l'écart-type à 1. Ensuite on calcule la matrice de distance DTW entre toutes les paires de séries temporelles normalisées.
  Pour le clustering, l'algorithme *DBSCAN* (*Density-Based Spatial Clustering of Applications with Noise*) est employé avec les paramètres suivants :
  - $\varepsilon = 0.5$, distance maximale entre deux points pour qu'ils soient considérés comme voisins,
  - `min_samples` $= 2$, densité minimale requise pour former un cluster.

  DBSCAN est choisi pour sa capacité à identifier des clusters de formes arbitraires et à détecter les points aberrants (*outliers*). Le cluster -1 représente les points considérés comme du bruit.
  Les résultats de l'analyse sont visualisés de deux manières :
  1. **Sous-graphiques par cluster** : Chaque cluster est représenté séparément.
  2. **Graphique interactif avec *Plotly*** : Toutes les séries temporelles sont affichées ensemble, colorées selon leur appartenance aux clusters. Cette visualisation interactive facilite l'exploration des résultats.

2.2.2 Frequent Pattern: Règles d'Association (Apriori)

L'analyse des règles d'association est utilisée pour identifier des relations significatives entre le taux de natalité et l'espérance de vie dans différents pays. Cette approche permet de découvrir des motifs fréquents et des associations potentiellement intéressantes dans les données.

Deux étapes principales de préparation des données sont effectuées :

- **Catégorisation des variables** :
  - Le taux de natalité est catégorisé en trois niveaux :
    - *Faible* : inférieur à 10.
    - *Moyenne* : entre 10 et 20.

        - *Élevé* : supérieur à 20.
    * L'espérance de vie est également catégorisée en trois niveaux :
        - *Faible* : inférieur à 60 ans.
        - *Moyenne* : entre 60 et 80 ans.
        - *Élevé* : supérieur à 80 ans.
- **Création des transactions** :
    * Pour chaque pays, une transaction est créée contenant les catégories de taux de natalité et d'espérance de vie pour toutes les années disponibles.
    * Ces transactions sont ensuite encodées en format binaire.

L'algorithme *Apriori* est appliqué sur les données transactionnelles avec les paramètres suivants :

- **Support minimum** : 0.05 (5% des transactions).

Les règles d'association sont extraites à partir des *itemsets* fréquents identifiés par *Apriori*. Les critères suivants sont appliqués :

- **Métrique** : *Lift*, pour mesurer l'indépendance des règles.
- **Seuil minimum de *lift*** : 1, pour garantir une corrélation positive.

2.3 Clusters

2.3.1 choix des données

Pour un clustering qui permet d'identifier des groupes de régions ayant des profils de santé similaires, tout en capturant les différences importantes entre ces groupes afin de fournir une base solide pour une analyse approfondie des disparités de santé entre les régions et pourra guider des interventions ciblées en matière de santé publique, nous choisissons de sélectionner les indicateurs de santé suivant la population à différents stades de la vie.

- **Espérance de vie à la naissance**: Cet indicateur fournit une mesure globale de la santé et de la longévité d'une population.
- **Taux de mortalité infantile**: Il reflète la qualité des soins de santé maternelle et infantile, ainsi que les conditions socio-économiques générales.
- **Taux de fécondité total**: Cet indicateur donne un aperçu des tendances démographiques et peut être lié à divers facteurs socio-économiques et de santé.
- **Mortalité avant 60 ans**: Cet indicateur met en lumière les décès prématurés et peut refléter l'efficacité des systèmes de santé et les conditions de vie.
- **Mortalité entre 15 et 50 ans**: Il se concentre sur la santé des adultes en âge de travailler, ce qui peut avoir des implications importantes sur la productivité économique et le bien-être social.
- **Espérance de vie à 65 ans**: Cet indicateur est crucial pour évaluer la santé et la qualité de vie des personnes âgées.

Ces indicateurs se complètent mutuellement, offrant une vue d'ensemble de la santé de la population tout en évitant les redondances, ils sont largement utilisées et standardisées, facilitant les comparaisons entre les régions et aussi pour guider les politiques de santé publique et évaluer leur efficacité.

2.3.2 Approche utilisée

Dans notre Implementation, nous avons employé deux méthodes de clustering distinctes : **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise) et **K-means**.
Chaque méthode a été appliquée sur les données brutes et sur les données réduites par Analyse en Composantes Principales (PCA). Cette approche nous permet de comparer les résultats et d'obtenir une compréhension plus approfondie de la structure des données.

- **DBSCAN**: elle a été choisi pour sa capacité à identifier des clusters de forme arbitraire et à détecter les outliers. Cette méthode est particulièrement utile puisque le nombre de clusters n'est pas connu a priori. Pour améliorer potentiellement les résultats du clustering et réduire le bruit dans les données.
- **K-means**: elle a été utilisé pour sa simplicité et son efficacité computationnelle.
  Cependant, l'un des défis majeurs de K-means est la détermination du nombre optimal de clusters. Pour surmonter ce défi, nous avons employé deux techniques complémentaires : le score de silhouette et l'inertie indiqué dans la figure suivante 4:

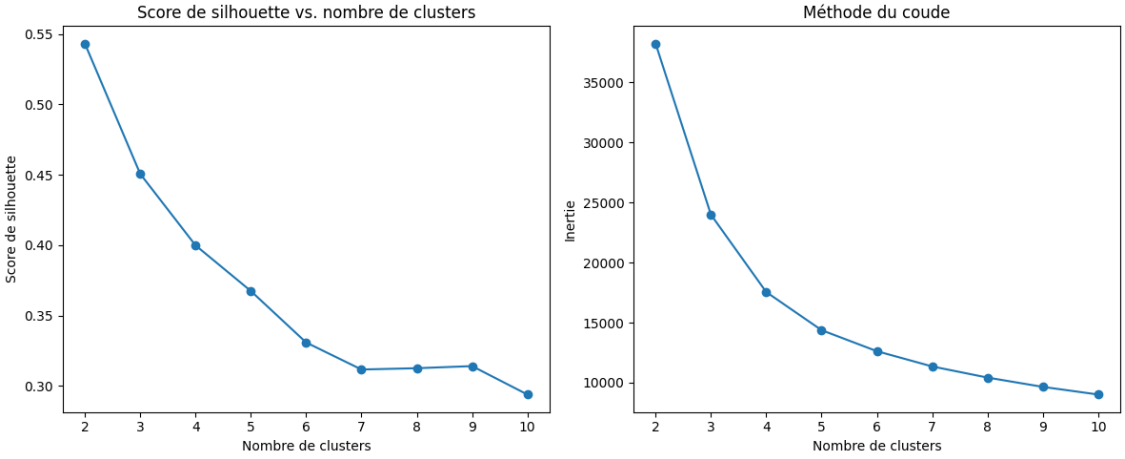


Figure 4: Choix du nombre optimal de clusters

- **Le score de silhouette** mesure la qualité et la cohérence des clusters. Il évalue à quel point un objet est similaire à son propre cluster par rapport aux autres clusters, en effet nous avons cherché à maximiser ce score pour déterminer le nombre optimal de clusters.
- **L'inertie**: elle quantifie la compacité globale des clusters, nous avons utilisé la méthode du "coude" sur le graphique de l'inertie pour identifier le point où l'ajout de clusters supplémentaires n'apporte plus d'amélioration significative.

Cette approche multi-méthodes augmente la fiabilité de nos résultats et nous permet d'avoir une vision plus complète de la structure des données, essentielle pour tirer des conclusions solides sur les groupements des régions en fonction de leurs indicateurs de santé.

2.4 Systèmes de Récommandation

2.4.1 choix des données

Lors de la conception de mon système de recommandation, le choix des données utilisées est crucial. Chaque champ sélectionné joue un rôle spécifique pour permettre au système de générer des recommandations cohérentes et pertinentes en se basant sur des similarités entre les entités (ici, les pays).

Les critères démographiques, sanitaires et sociaux choisis permettent de capturer une vue globale des caractéristiques principales des pays et de répondre aux besoins des utilisateurs, comme l'identification de pays ayant des contextes similaires.

- **Total Population**
  La taille de la population constitue un indicateur de base permettant de normaliser d'autres indicateurs, comme les taux de mortalité ou de natalité. Les pays ayant des populations similaires peuvent partager des défis communs en termes de gestion des ressources ou de politiques publiques.

- **Population Growth Rate**
  Ce champ met en lumière la dynamique démographique d'un pays, ce qui permet de rapprocher des pays en phase de croissance rapide ou, au contraire, de déclin.

- **Total Fertility Rate**
  Le taux de fécondité reflète les comportements reproductifs et les projections de croissance future. Les pays ayant des taux similaires peuvent avoir des défis communs liés à l'éducation ou à la planification familiale.

- **Life Expectancy at Birth, both sexes**
  Cet indicateur synthétise les progrès en matière de santé publique et de qualité de vie. Les pays avec une espérance de vie similaire partagent souvent des niveaux comparables de développement.

- **Infant Mortality Rate**
  Ce champ évalue la qualité des soins prénataux et postnataux ainsi que des politiques de santé maternelle et infantile.

- **Mean Age Childbearing**
  L'âge moyen des mères lors de l'accouchement reflète les comportements sociaux et culturels ainsi que l'accès à l'éducation ou à la planification familiale.

- **Mortality before Age 60, both sexes**
  Cet indicateur met en avant les décès prématurés souvent liés à des inégalités économiques ou à des maladies évitables.

- **Births by women aged 15 to 19**
  Le nombre de naissances chez les adolescentes est un indicateur de développement socio-économique et d'accès à l'éducation.

- **Infant Deaths, under age 1**
  Le nombre brut de décès d'enfants de moins d'un an donne une perspective complémentaire au taux de mortalité infantile.

- **Under-Five Deaths, under age 5**
  Le nombre total de décès d'enfants de moins de 5 ans fournit une vue plus large sur les enjeux de mortalité infantile et les conditions de vie des jeunes enfants.

- **Year**
  La prise en compte de l'année permet d'analyser les pays dans le contexte temporel pertinent, notamment pour des données évolutives.

– **Country**
Ce champ est utilisé pour identifier et différencier les entités analysées (les pays).

2.4.2 Approche utilisée

Pour développer un système de recommandation robuste, j'ai adopté une approche hybride en intégrant quatre méthodes complémentaires. Cette démarche permet d'exploiter les forces de chaque méthode pour obtenir des recommandations plus pertinentes et adaptées.

– **Recommandation basée sur les plus proches voisins (KNN)**
Cette méthode utilise les similarités entre les pays pour recommander des entités similaires. En comparant directement les caractéristiques des pays, elle identifie ceux qui partagent des profils proches en termes de démographie, de santé publique ou de développement social.

– **Factorisation matricielle (SVD)**
En décomposant les données en facteurs latents, cette méthode met en lumière les relations sous-jacentes entre les différents champs. Elle est particulièrement utile pour identifier des patterns moins évidents, comme des liens entre des pays ayant des indicateurs apparemment divergents mais partageant des tendances cachées.

– **Similarité cosinus (cosine_similarity)**
Cette méthode évalue les similarités entre les vecteurs représentant les caractéristiques des pays. Elle est adaptée pour mesurer des relations proportionnelles entre les champs, indépendamment de leur magnitude. Par exemple, elle peut identifier des pays ayant des taux de croissance démographique similaires, même si leurs populations totales diffèrent considérablement.

– **Regroupement par K-means**
Le clustering K-means regroupe les pays en clusters en fonction de la proximité de leurs caractéristiques. Chaque cluster représente un groupe de pays partageant des contextes similaires, facilitant l'identification de segments homogènes.

3 Résultats et Analyse

3.1 Analyse temporelle

Les résultats pour la décomposition de la série temporelle globale sont illustrés par la figure 5 :

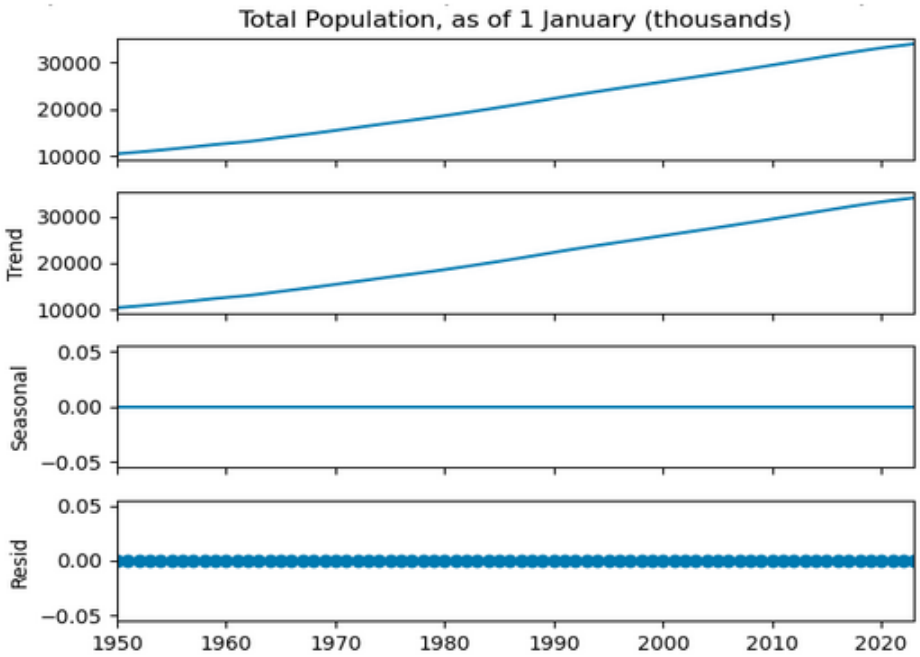


Figure 5: Décomposition de la série temporelle de la population mondiale.

La figure 5 illustre la décomposition d'une série temporelle de la population mondiale en trois composantes : tendance, saisonnalité et résidus. Série temporelle observée : La courbe supérieure montre une augmentation constante de la population mondiale de 1950 à 2020, indiquant une croissance soutenue.

Tendance : La composante de tendance suit étroitement la courbe brute, révélant une croissance exponentielle modérée de la population mondiale.

Saisonnalité : La composante saisonnière est presque nulle, ce qui est attendu, car les données démographiques mondiales ne présentent pas de variations saisonnières significatives.

Résidus : Les résidus sont très faibles et proches de zéro, indiquant que le modèle de décomposition explique presque entièrement la série temporelle.

La série est principalement influencée par la tendance, avec peu d'irrégularités. Cela reflète la stabilité et la prévisibilité de la croissance démographique mondiale sur cette période.

L'analyse d'autocorrélation a révélé une forte autocorrélation positive pour les premiers lags, indiquant une dépendance temporelle significative à court terme. Comme montre la figure 6 l'autocorrélation diminue progressivement avec l'augmentation des lags, mais reste statistiquement significative sur plusieurs périodes. Cette structure suggère une forte persistance dans la série temporelle de la population mondiale, reflétant la nature continue et prévisible de la croissance démographique sur la période étudiée.

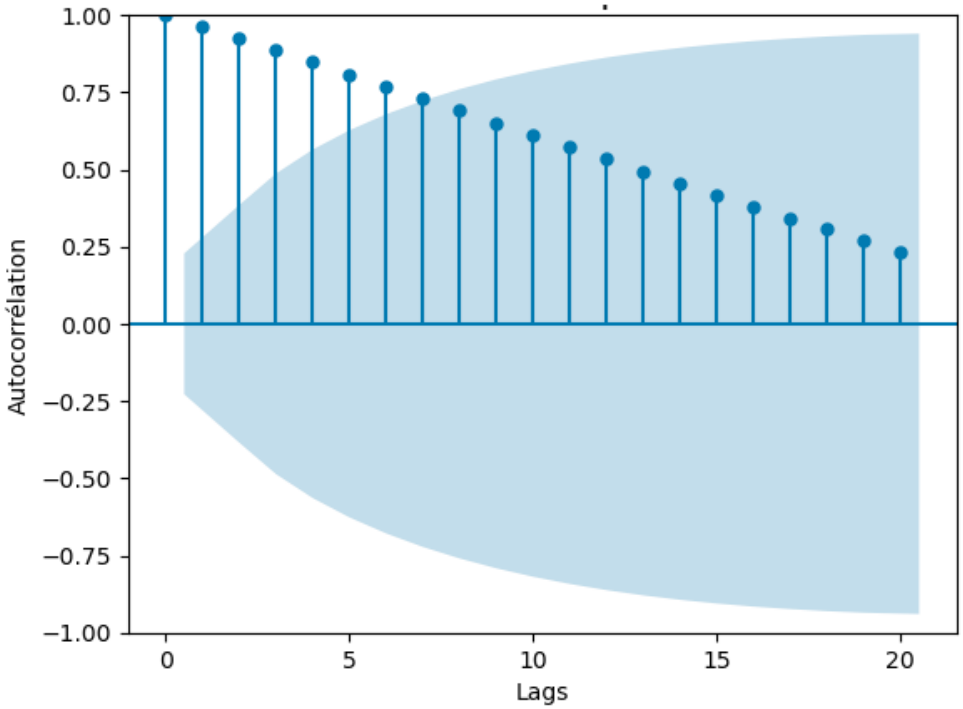


Figure 6: Autocorrélation de la population mondiale.

L'application du clustering DTW-DBSCAN a révélé plusieurs groupes distincts de trajectoires démographiques illustrés par la figure 7:

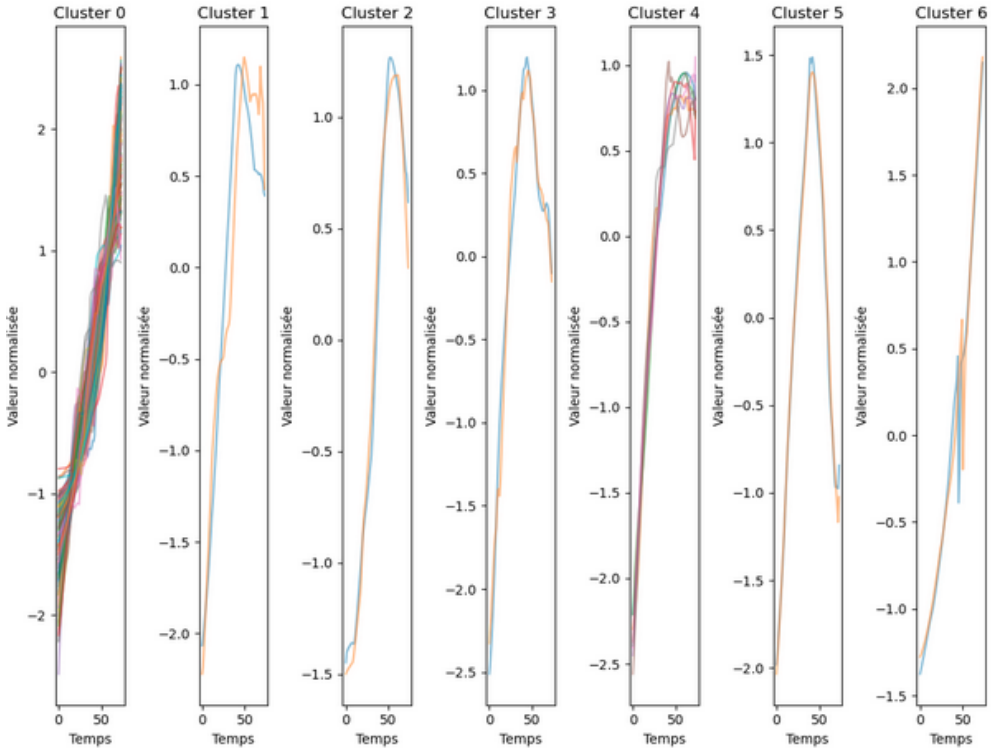


Figure 7: Séries temporelles: DBSCAN Clusters.

- **Cluster 0 (majoritaire) :** Regroupe des pays présentant une croissance démographique stable et modérée.
- **Clusters 1-6 :** Identifient des motifs démographiques plus spécifiques :
  Cluster 1 : Pays avec un pic de population suivi d'un déclin rapide.
  Cluster 2 : Nations connaissant une croissance démographique accélérée.
  Clusters 3-6 : Divers modèles présentant des variations significatives (pics, creux, inflexions).

Cette méthode a permis de capturer efficacement les similitudes entre les trajectoires démographiques, même en présence de décalages temporels. Les résultats mettent en évidence la diversité des dynamiques de population à l'échelle mondiale, allant de modèles de croissance stable à des schémas plus complexes et variables. L'approche DTW-DBSCAN s'est révélée particulièrement utile pour identifier des groupes de pays partageant des caractéristiques démographiques communes, tout en isolant les cas atypiques.

3.2 Analyse Frequent Pattern

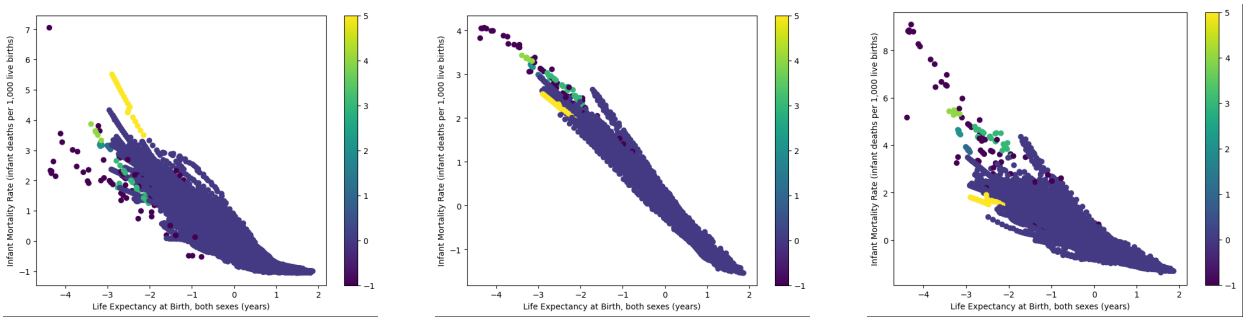
L'application de l'algorithme Apriori a révélé plusieurs associations entre les catégories de taux de natalité et d'espérance de vie :

- Un taux de natalité moyen est fortement associé à une espérance de vie faible (support : 99,58%, confiance : 99,58%).
- Un taux de natalité faible est systématiquement lié à une espérance de vie moyenne (support : 99,58%, confiance : 100%).
- Un taux de natalité élevé est invariablement associé à une espérance de vie moyenne (support : 90,72%, confiance : 100%).

Ces règles présentent toutes un lift de 1, indiquant une coexistence fréquente des catégories, mais sans dépendance significative. L'analyse suggère des tendances démographiques générales, où les pays à taux de natalité moyen tendent à avoir une espérance de vie plus faible, tandis que ceux à taux de natalité faible ou élevé sont associés à une espérance de vie moyenne. Cependant, l'absence de lift supérieur à 1 indique que ces associations, bien que fréquentes, ne sont pas nécessairement causales.

3.3 Analyse des Clusters

3.3.1 DBSCAN sans reduction



Les trois graphes montrent les clusters formés en fonction de différentes combinaisons de variables deux a deux, et identifiés sur la base des dimensions brutes normalisées.
Cela signifie que les points se situent dans un espace multidimensionnel, où les corrélations et les relations entre variables sont directement exploitées.
On constate que les données contenir du bruit élevé, ce qui peut conduire à des clusters qui ne sont pas bien séparés.

Le chevauchement entre clusters est plus difficile à évaluer, car les graphes ne montrent qu'une fraction de la variation entre les dimensions.

3.3.2 DBSCAN avec réduction par PCA

La réduction par PCA diminue les redondances dans les données et capture les principales variations, cela rend les clusters plus facilement identifiables visuellement indiqué dans la figure 8, en particulier dans des données à haute dimension.
On peut constaté que les clusters sont mieux séparés et plus distincts dans l'espace réduit.
Le bruit (-1) est visible et représenté en noir, ce qui permet de mieux comprendre la structure des données et la proportion des points non regroupés.

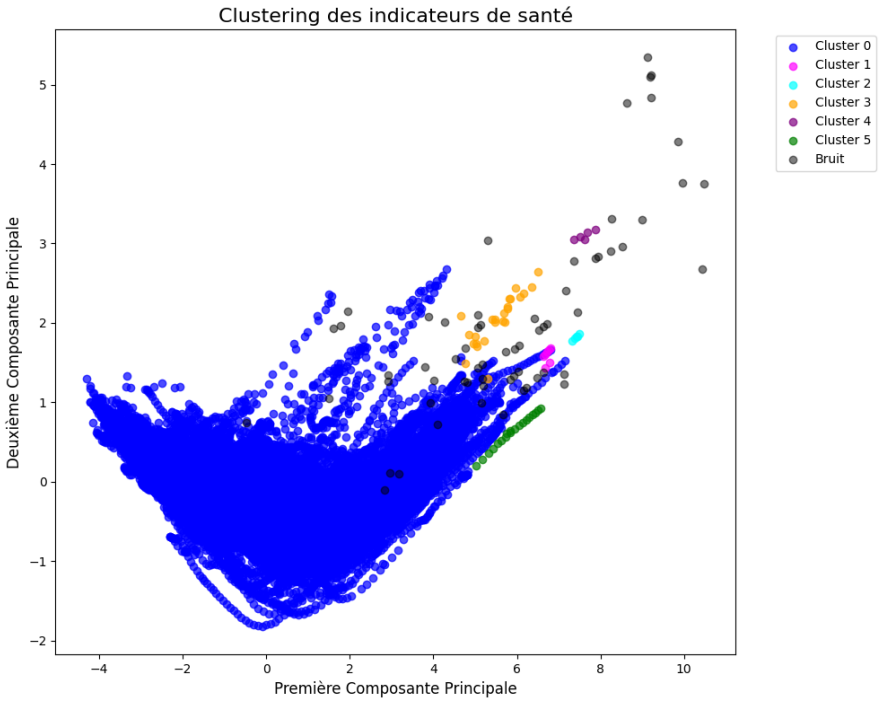
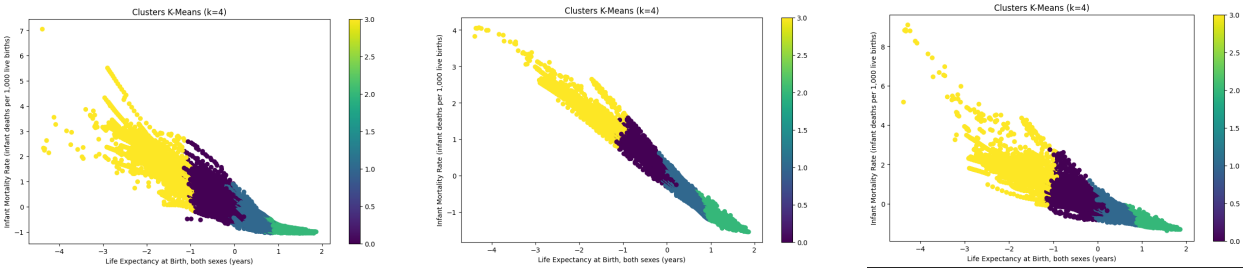


Figure 8: Clustering des pays par le profil de santé

3.3.3 K-means sans reduction

Le clustering est affiché directement sur deux dimensions normalisées parmi les variables initiales.



Bien que les clusters soient visibles, les deux dimensions choisies peuvent ne pas capturer toute la variabilité présente dans l'ensemble de données multivarié. Cette approche utilise toutes les variables d'origine pour calculer les distances entre les points, ce qui préserve l'intégrité des relations dans l'espace multidimensionnel, mais seules deux dimensions sont affichées parmi toutes les variables, ce qui peut induire une perte de compréhension visuelle du regroupement global.

3.3.4 K-means avec reduction

Le clustering est projeté sur les deux premières composantes principales après réduction de dimension via PCA, ces deux composantes expliquent la majorité de la variance présente dans les données ce qui permet une visualisation simplifiée et claire pour des nombreuses dimensions et une bonne séparation visuelle affiché dans la figure **??**

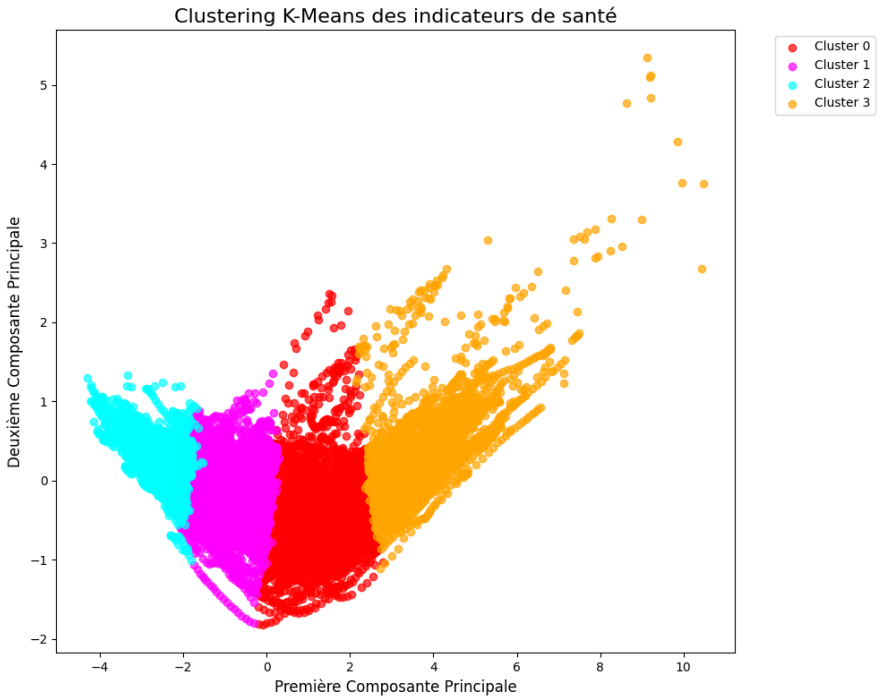


Figure 9: Clustering K-means des pays par le profil de santé

D'apres notre analyse comparative des méthodes de clustering, DBSCAN avec réduction par PCA semble être la meilleure option pour ce clustering pour les raisons suivants:

- Les clusters sont mieux définis et distincts, surtout après la réduction de dimension.
- La gestion des outliers est importante pour analyser des données bruitées, et DBSCAN le fait naturellement.
- Ne nécessite pas de spécifier un nombre de clusters, ce qui est avantageux dans notre situation où le nombre est inconnu ou difficile à estimer.

3.4 Analyse des Systèmes de Récommandation

3.4.1 Recommandation basée sur les plus proches voisins (KNN)

La similarité entre les pays dans K-Nearest Neighbors (KNN) est également mesurée en fonction de la distance entre leurs caractéristiques. Cependant, dans le cas du KNN, les pays les plus similaires à un pays donné sont identifiés par la proximité des points dans l'espace des caractéristiques. Plus précisément, un pays est classé en fonction des pays les plus proches (voisins) selon une mesure de distance (par exemple, la distance euclidienne). Les pays voisins les plus proches partagent des caractéristiques similaires, ce qui permet de prédire des comportements ou des catégories similaires pour le pays donné.

Exemple d'Application :

Pays recommandés	Distances	Pays recommandés	Distances
Mongolie	0.6739	Royaume-Uni	0.3709
Kiribati	0.7756	Allemagne	0.5734
Papouasie-Nouvelle-Guinée	0.8697	Belgique	0.5881
Myanmar	0.9109	Danemark	0.6292
Bolivie (État plurinational)	1.024	Grèce	0.6520

Table 1: Recommandations de pays similaires avec les distances associées pour les pays Maroc et France.

Le Maroc est recommandé avec des pays comme la Mongolie, Kiribati, Papouasie-Nouvelle-Guinée, Myanmar et la Bolivie, qui partagent des caractéristiques démographiques et sanitaires similaires. Les distances montrent des similarités globales, bien que des différences existent, notamment avec la Bolivie (distance de 1.02).

Pour la France, les recommandations incluent des pays européens comme le Royaume-Uni, l'Allemagne, la Belgique, le Danemark et la Grèce. Les distances sont faibles, indiquant des profils proches, mais elles augmentent légèrement, suggérant des variations dans les politiques publiques ou les systèmes de santé.

3.4.2 Factorisation matricielle (SVD)

La SVD décompose la matrice des données en trois matrices : une pour les pays, une pour les caractéristiques, et une pour les valeurs singulières. Elle réduit la complexité des données tout en capturant les informations essentielles. Pour chaque pays, on obtient des vecteurs représentant ses caractéristiques latentes dans un espace réduit. Pour recommander des pays similaires, on calcule la distance entre ces vecteurs. Moins la distance est grande, plus les pays sont similaires. Cette approche permet d'identifier des relations cachées entre les pays, même si leurs caractéristiques sont initialement différentes.

Exemple d'Application :

Figure 10: Similarité des pays à France

Le Royaume-Uni et l'Italie sont proches de la France en raison de similarités démographiques et sanitaires, comme l'indiquent leurs faibles distances. Ces pays partagent des caractéristiques économiques et sociales communes.
La Thaïlande, l'Ukraine et la Turquie ont des distances plus élevées, suggérant des différences notables, notamment dans les systèmes de santé, le développement social et d'autres facteurs socio-économiques.

3.4.3 Similarité cosinus (cosine_similarity)

La similarité cosinus est une méthode utilisée pour mesurer la similitude entre deux vecteurs dans un espace multidimensionnel. Dans le cas des données démographiques et sanitaires des pays, chaque pays peut être représenté comme un vecteur où chaque dimension correspond à une caractéristique telle que la population, le taux de fertilité, l'espérance de vie, etc.

Pour calculer la similarité cosinus entre deux pays, on commence par considérer leurs vecteurs de caractéristiques normalisés, c'est-à-dire que les valeurs sont ajustées pour éliminer l'impact des différences d'échelles. Ensuite, la similarité cosinus est calculée en prenant le produit scalaire des deux vecteurs et en le divisant par le produit de leurs normes respectives. Ce calcul donne une valeur entre -1 et 1. Un score proche de 1 indique que les deux pays sont très similaires en termes de leurs caractéristiques démographiques et sanitaires, tandis qu'un score plus faible indique une différence plus marquée.

Exemple d'Application :

Country	Year	Similarity
United Kingdom	2019	0.997690
Italy	2009	0.997209
Italy	2006	0.997096
Italy	2010	0.996946
Germany	2017	0.996850

Table 2: Recommandations de pays similaires avec les distances associées pour le France.

Pour la France (année 2020), les pays les plus similaires sont le Royaume-Uni et l'Italie, avec des similitudes élevées proches de 1 (0.9976 et 0.9972). Cela reflète des caractéristiques démographiques et sanitaires semblables entre ces pays. D'autres pays comme l'Allemagne et l'Italie (2009 et 2006) sont également proches.

3.4.4 Regroupement par K-means

La similarité entre les pays dans K-means est mesurée par la distance euclidienne entre leurs caractéristiques. Les pays proches dans cet espace de caractéristiques sont regroupés dans le même cluster, indiquant qu'ils partagent des profils similaires (par exemple, des taux de croissance de la population ou de fécondité proches). Les pays dans un même cluster sont donc considérés comme similaires.

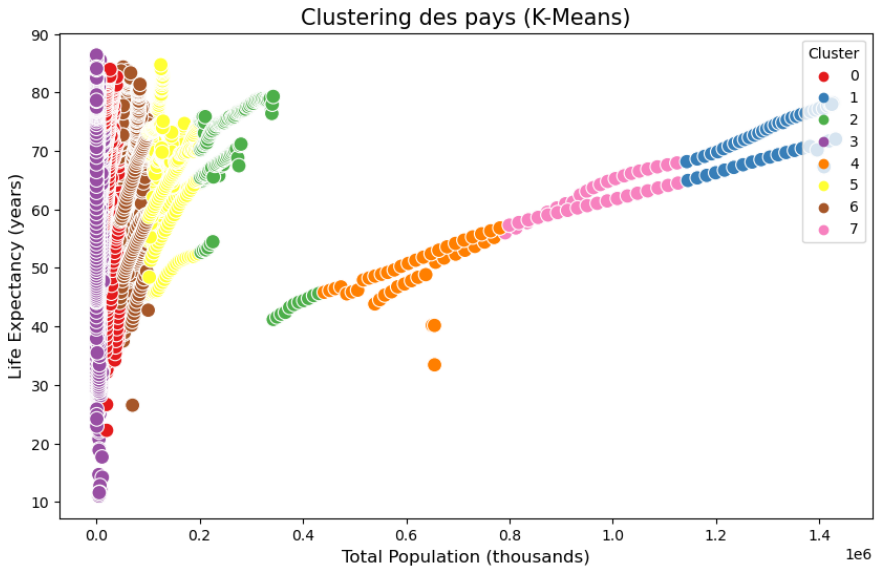


Figure 11: Regroupement des pays

Exemple d'Application :

Prenons l'exemple du Maroc :

Les pays similaires, tels que Burundi, Comoros, Djibouti, Eritrea, et Kenya, sont dans le même cluster. Cela signifie qu'ils partagent des caractéristiques démographiques et sanitaires proches, comme des taux de mortalité infantile ou une espérance de vie similaires.

4 Conclusion

Cette analyse a permis de mettre en lumière les dynamiques de population à l'échelle mondiale, en utilisant des techniques avancées de data mining. À travers l'analyse temporelle, nous avons observé une augmentation continue de la population mondiale, passant de 2 milliards en 1950 à plus de 8 milliards en 2023, illustrant ainsi une croissance démographique significative. L'application de méthodes telles que le clustering et l'analyse des patterns fréquents a révélé des relations importantes entre le taux de natalité et l'espérance de vie, tout en identifiant des groupes de pays partageant des caractéristiques démographiques similaires. Le système de recommandation que nous avons développé utilise des critères démographiques et sanitaires pour identifier des pays partageant des caractéristiques similaires, Ces résultats offrent des perspectives précieuses pour guider les politiques de santé publique et anticiper les changements démographiques futurs,

Répartition des taches:

- **Exploration des données:** Awatef Edhib, Aida Haouas, Rida Asri
- **Analyse temporelle et Frequent pattern:** Awatef Edhib
- **Clusters:** Aida Haouas
- **Systèmes de recommandation:** Rida Asri