Projet : clustering hiérachique

ALOUADI Alexandre, BLANC Solène, PETERSON Alexandre

February 2023

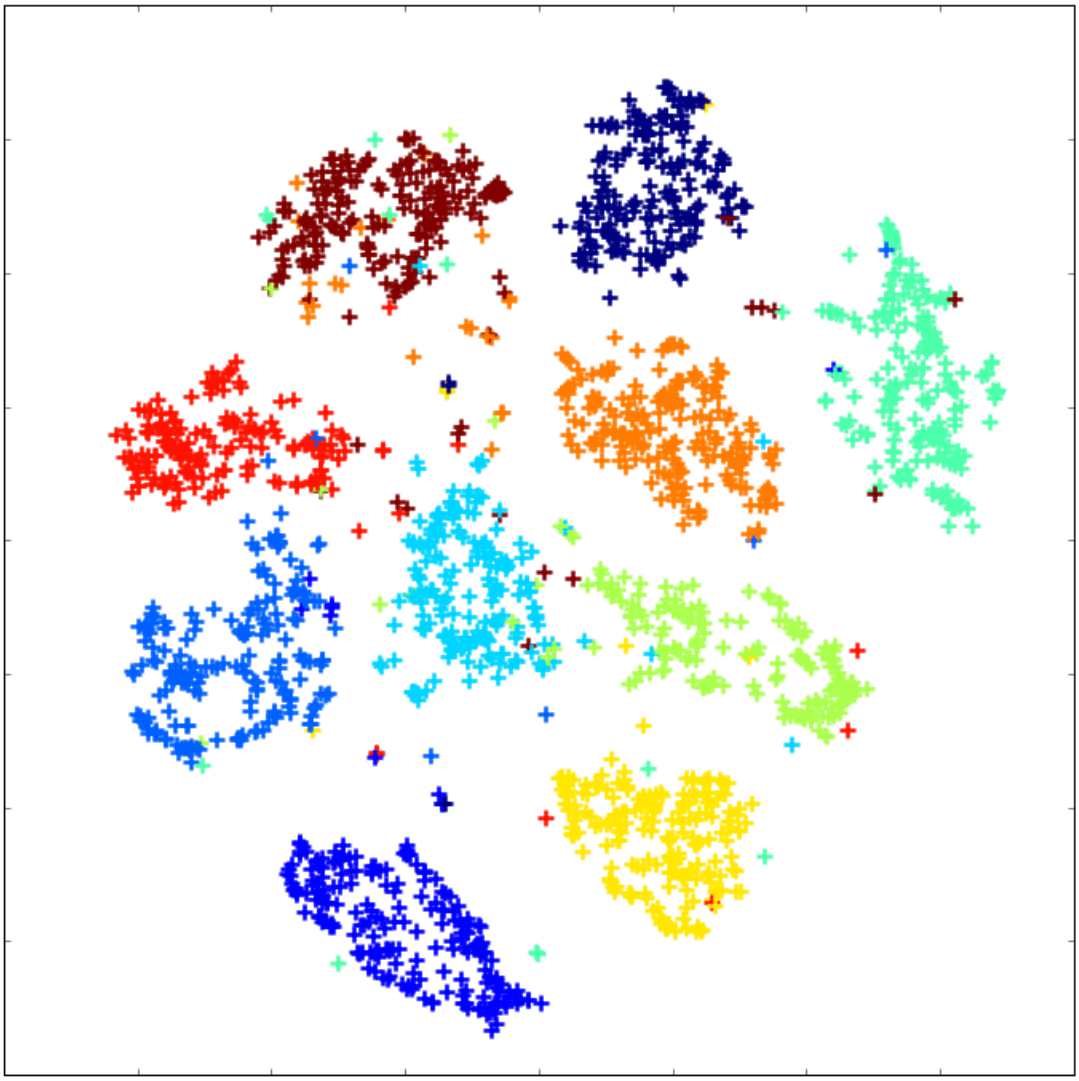


Figure 1: Illustration de classification de données

Table des matières

1 Introduction

Le clustering hiérarchique est un algorithme de classification non supervisée qui a de nombreuses applications pratiques dans différents domaines. De la segmentation de la clientèle en marketing à la découverte de motifs dans les données biomédicales, le clustering hiérarchique est un outil puissant pour organiser et explorer les données complexes. Cette méthode consiste à regrouper les éléments similaires en groupes appelés clusters, et à continuer à fusionner ces groupes jusqu'à ce qu'il n'y ait plus qu'un seul cluster ou un nombre prédéterminé de clusters. L'objectif de ce projet consiste à étudier en détail la méthode de clustering hiérarchique, à la mettre en œuvre numériquement et à l'appliquer sur un jeu de données.

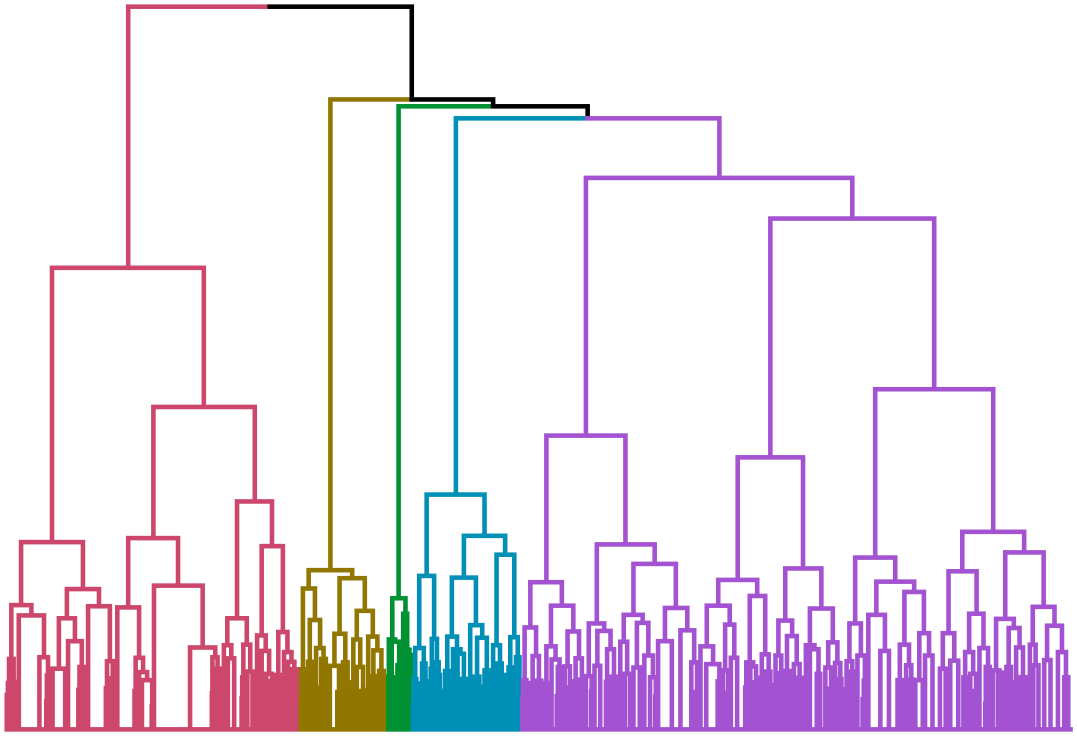


Figure 2: Illustration d'un dendrogramme

2 Étude Théorique

2.1 Présentation de la méthode de clustering hiérarchique

Le clustering hiérarchique est une méthode de classification itérative. À partir d'une base de données, on cherche à regrouper les individus qui se ressemblent pour créer des clusters. Les clusters sont ensuite organisés autour d'un arbre. Chaque groupe, ou nœud, est lié à deux groupes successeurs ou davantage. Tous les groupes sont donc imbriqués entre eux. En effet, les groupes d'un niveau, les enfants, rejoignent les groupes du niveau supérieur, les parents, toujours en fonction de leurs similitudes. Le processus se poursuit alors jusqu'à ce que tous les nœuds soient intégrés à l'arbre. Contrairement à la méthode du k-mean, le nombre total de clusters n'est pas prédéterminé avant le lancement de l'algorithme. Cette méthode permet de créer un dendrogramme, qui représente visuellement l'arbre décrit ci-dessus.

Il existe deux grands types de classification hiérarchique : la classification hiérarchique ascendante et la classification hiérarchique descendante. Dans la méthode de la classification hiérarchique ascendante, chaque point est considéré comme un cluster à part entière au départ. Ensuite, on cherche à regrouper le cluster le plus proche de chaque point (celui qui a le plus de similarités avec lui). On réitère ce processus jusqu'à ce que chaque point/individu de la base de données soit regroupé en un seul grand cluster. Dans la méthode de la classification hiérarchique descendante, on effectue le processus inverse. On commence par regarder le grand cluster qui regroupe tous les individus et on le subdivise progressivement jusqu'à ce qu'il y ait autant de clusters que de points.

2.2 Différentes variations de la méthode

Le clustering hiérarchique étant une famille de méthodes, il existe différentes variations de cette dernière. Dans la partie précédente, nous avons cité deux grandes méthodes : la méthode ascendante et la méthode descendante. Nous allons ici nous concentrer sur la méthode ascendante, qui est la plus couramment utilisée. Ensuite, nous expliquerons les avantages et inconvénients du clustering hiérarchique.

2.2.1 La méthode ascendante et descendante

La classification ascendante hiérarchique (CAH) est une méthode de classification itérative. Au début, chaque point est un cluster. Puis, on effectue le processus suivant :

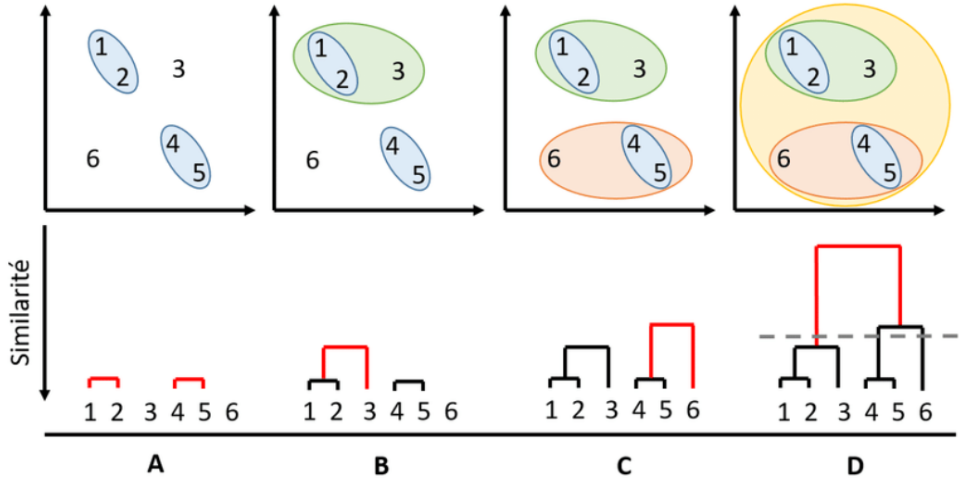
1. On calcule la dissimilarité entre les N objets.
2. Ensuite, on regroupe les deux objets dont le regroupement minimise un critère d'agrégation donné, créant ainsi une classe comprenant ces deux objets.
3. On calcule ensuite la dissimilarité entre cette classe et les N-2 autres objets en utilisant le critère d'agrégation. Puis, on regroupe les deux objets ou classes d'objets dont le regroupement minimise le critère d'agrégation.

On continue ce processus jusqu'à que tous les objets soient regroupés dans une seule classe. La manière de calculer la dissimilarité est expliqué dans la partie 2.6.
Mathématiquement parlant, notons H l'ensemble des classes, crée par l'algorithme. Nous avons les propriétés suivantes :

- $\Omega \in H$: si l'on veut obtenir une unique classe, tous les individus appartiennent à cette classe.
- $\forall \omega \in H, \omega \in H$, tous les individus forment un seul cluster en bas du dendrogramme.
- $\forall (h_1, h_2) \in H^2$, il y a 3 possibilités : $h_1 \cap h_2 = \emptyset$, $h_1 \subseteq h_2$ ou $h_2 \subseteq h_1$. En fait, si l'on considère deux clusters, soit ils n'ont pas d'individus en commun, soit ils sont inclus l'un dans l'autre.

Pour la méthode dite descendante, on divise le cluster en deux à chaque étape au lieu de le regrouper.

Illustrons nos propos par un exemple :



La partie haute de l'image correspond à la représentation des objets numérotés de 1 à 6 en fonction des critères choisis pour la classification, la partie basse décrit la création d'un dendrogramme. On constate que les individus 1 et 2 puis 4 et 5 sont les plus proches, d'où la formation de 2 clusters. On réitère le processus en regroupant $\{1, 2\}$ avec $\{3\}$; puis en regroupant $\{4, 5\}$ avec $\{6\}$. Enfin, on a les six individus regroupés dans une seule et même classe. La ligne grise représentant la coupe optimale, on constate que les 6 individus sont finalement regroupés en 3 clusters: $\{1, 2, 3\}$, $\{4, 5\}$ et $\{6\}$.

2.2.2 Avantages et inconvénients

La classification hiérarchique ascendante présente plusieurs avantages. Tout d'abord, cette méthode permet de choisir un critère de d'agrégation adapté au sujet étudié et à la nature des données. Ensuite, contrairement à la méthode du k-means, elle ne nécessite pas de déterminer un nombre de classes au préalable. En effet, en jouant sur la profondeur de l'arbre, on peut explorer différentes possibilités et choisir le nombre de classes qui convient le mieux à la base de données et au critère de dissimilarité choisi. Le nombre de classes peut donc être choisi en regardant le dendrogramme

Le principal problème de cette méthode est son implémentation. En effet, à chaque itération, c'est-à-dire à chaque fois que l'on divise un cluster en deux pour l'approche descendante, ou que l'on regroupe deux clusters pour l'approche ascendante, l'algorithme doit recalculer les distances de toutes les paires de points possibles entre les deux clusters en question. Cette méthode nécessite donc beaucoup de temps et d'espace mémoire, elle est donc mieux adaptée aux petits échantillons.

2.3 Types de distances possibles entre les individus

L'algorithme de classification hiérarchique consiste à regrouper des individus similaires ou à séparer ceux qui sont différents, suivant si la méthode est ascendante ou descendante. Pour fusionner ou séparer les classes, on utilise des critères de similitudes entre individus. Ces critères reposent donc sur la notion de distance entre individus.

Soit d'une distance, une application de E×E dans $\mathbb{R}^+$ telle que $\forall\, a, b, c \in E$:

$$\begin{aligned} d(a,b) = 0 &\iff a = b && \text{(séparation)} \\ d(a,b) &= d(b,a) && \text{(symétrie)} \\ d(a,c) &\leq d(a,b) + d(b,c) && \text{(inégalité triangulaire)} \end{aligned}$$

Si la distance est nulle, les individus sont identiques. Si elle est faible, ils sont proches et on voudra donc les regrouper. Si elle est grande, ils sont éloignés et on voudra les séparer. La notion de distance étant nécessaire, le choix de celle-ci est donc important. De nombreuses distances peuvent être considérées selon le type de données utilisées, les objectifs et le critère de similitude utilisé.

Voyons quelques distances, $\forall X, Y \in \mathbb{R}^n$:

- La distance de Minkowski : $\sqrt[p]{\sum_{i=1}^{n} |X_i - Y_i|^p}$

  Cette distance dépend de p $\in \mathbb{N}^*$, elle est efficace pour des classes compactes ou isolés, mais est sensible aux valeurs aberrantes (outliers). Elle est aussi sensible aux échelles des dimensions, cela peut poser un problème si la différence entre les échelles est trop importante.
  Plusieurs cas particuliers de cette distance sont très utilisés :

  - la distance euclidienne : $\sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2}$

    C'est la distance la plus couramment utilisée en pratique. La distance euclidienne est une mesure de la distance géométrique entre deux points dans un espace à n dimensions. Pour les données de clustering, la distance euclidienne mesure la distance entre deux individus en considérant leurs coordonnées dans un espace à n dimensions. Les coordonnées sont les différentes variables dans lesquelles chaque individu est mesuré. La distance euclidienne est facile à calculer et donne une mesure intuitive de la distance. Cependant, elle peut être sensible aux différences d'échelle entre les variables.
  - la distance de Manhattan : $\sum_{i=1}^{n} |X_i - Y_i|$

    La distance de Manhattan est une mesure de la distance entre deux individus qui est basée sur la somme des différences absolues entre les valeurs de chaque variable. Cette distance est utile lorsque les données ont des échelles différentes et sont mesurées dans des unités différentes. Elle peut également être utilisée pour des données qui ont des relations linéaires, mais qui sont fortement affectées par des valeurs aberrantes. Cependant, cette mesure ne prend pas en compte les corrélations entre les variables et peut-être moins utile pour des données hautement corrélées.
  - la distance de Tchebychev, p $\to +\infty$: $\max_i |X_i - Y_i|$

    La distance de Tchebychev est une mesure de la distance entre deux individus qui est basée sur la plus grande différence absolue entre les valeurs de chaque variable. Cette distance est utile pour des données avec des valeurs aberrantes ou des données qui ont des échelles différentes entre les variables. Cependant, elle peut également être sensible aux valeurs aberrantes et ne prend pas en compte les corrélations entre les variables.

- Distance euclidienne au carré : $\sum_{i=1}^{n} (X_i - Y_i)^2$

  La distance euclidienne au carré est une mesure alternative qui consiste à prendre la distance euclidienne et à la mettre au carré. Elle peut être préférable à la distance euclidienne normale lorsque l'on souhaite donner plus de poids aux distances plus éloignées. Cela peut aider à réduire l'influence des valeurs aberrantes et à obtenir des clusters plus compacts. Cependant, elle peut également être sensible aux différences d'échelle et peut être plus difficile à interpréter que la distance euclidienne normale.

- La distance de Pearson : $1 - \frac{\sum_{i=1}^{n} (X_i - \mathbb{E}(X))(Y_i - \mathbb{E}(Y))}{\sqrt{\sum_{i=1}^{n} (X_i - \mathbb{E}(X))^2} \sqrt{\sum_{i=1}^{n} (Y_i - \mathbb{E}(Y))^2}}$

  La distance de Pearson est une mesure de la similarité entre deux individus basée sur leurs corrélations. Elle mesure la différence entre deux vecteurs normalisés de données. Cette mesure est couramment utilisée dans le domaine de la biologie pour étudier les relations entre différentes espèces. La distance de Pearson est utile pour les données hautement corrélées, mais elle peut donner des résultats biaisés si les données ont des relations non linéaires.

- La distance de Canberra : $\sum_{i=1}^{n} \frac{|X_i - Y_i|}{X_i + Y_i}$

  La distance de Canberra est une mesure de la distance entre deux individus qui prend en compte la différence relative entre les valeurs dans chaque variable. Cette distance est utile lorsque les valeurs sont très dispersées et que les différences relatives sont plus importantes que les différences absolues. Elle peut également être utilisée pour des données qui ont des échelles différentes, car elle est moins sensible aux différences d'échelle que la distance euclidienne. Cependant, elle peut également être sensible aux données extrêmes et aux valeurs aberrantes.

- La distance cosinus : $1 - \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}}$

  La distance cosinus est une mesure de la similarité entre deux individus basée sur l'angle entre leurs vecteurs de données normalisés. Ces valeurs sont comprises entre 0 et 2. Lorsque le quotient $\frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}}$ vaut -1 (la distance est égal à 2), cela signifie que les vecteurs sont opposés. Lorsque le quotient vaut 0 et donc la distance 1, les vecteurs X et Y sont orthogonaux (indépendants). Enfin, lorsque $\frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}} = 1$, la distance est nulle et les vecteurs sont colinéaires de coefficient positif.

  Les valeurs intermédiaires évaluent le degré de similitude des vecteurs. Cette distance est utile pour les données avec de nombreuses variables et peut être robuste aux différences d'échelle entre les variables. Elle peut également être utilisée pour des données textuelles et de classification, mais elle ne prend pas en compte les différences dans la magnitude des valeurs des variables.

En résumé, chaque distance a ses propres spécificités, avantages et inconvénients. Il est important de choisir une distance qui convient le mieux aux données et a l'objectif du clustering. Les distances euclidiennes sont souvent utilisées, mais il peut être utile d'explorer d'autres distances, telles que la distance de Manhattan ou la distance cosinus, en fonction des caractéristiques de vos données.

2.4 Algorithme de clustering hiérarchique : explication

Algorithme 1 : Classification hiérarchique ascendante

Entrées : $x_1, ..., x_n$ (donnés), d (distance), S (critère de similitude)
$M \leftarrow \emptyset$;
pour $i \leftarrow 1$ **à** n **faire**
  $M \leftarrow M \cup \{\{x_i\}\}$;
fin
tant que $|M| > 1$ **faire**
  $C_1, C_2 \leftarrow \underset{\substack{A,B \in M \\ A \neq B}}{\operatorname{argmin}} \mathrm{S}(\mathrm{d})(A, B)$;
  $M \leftarrow ((M \backslash \{C_1\}) \backslash \{C_2\}) \cup \{C_1 \cup C_2\}$;
fin

Expliquons l'algorithme avec un exemple simple étape par étape :

- Entrées : seuls les données sont essentielles, la distance et le critère peuvent être directement implémentés dans l'algorithme.
  Les données sont des vecteurs qui correspondent à chaque individu et à chacune de ses "caractéristiques" que l'on veut comparer pour pouvoir les regrouper.
  Prenons 4 individus de dimension 1, la distance euclidienne et le minimum de cette distance comme critère.

x_1	x_2	x_3	x_4
4	3	15	6

- Initialisation : on crée une matrice M pour l'instant vide, que l'on va utiliser à chaque étape pour fusionner les classes.

$$M = \{\}$$

- Boucle : On insère chaque individu dans cette matrice, les individus sont séparés.

$$M = \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}\}$$

  On a 4 clusters.

- Boucle Tant que : On refait cette étape tant que l'on n'a pas obtenu un unique cluster, donc tant que M n'est pas de taille 1. Sachant que le min nous donne la plus petite dissimilitude entre les groupes selon la distance d et le critère S, argmin renvoie alors les deux classes les plus proches. On retire alors ces classes de M tout en rajoutant leur fusion.

  – Voici un tableau des distances :

	x_1	x_2	x_3	x_4
x_1	0	1	11	2
x_2	1	0	12	3
x_3	11	12	0	9
x_4	2	3	9	0

  Le minimum est 1 donc on va fusionner x_1 et x_2.

$$M = \{\{\{x_1\}, \{x_2\}\}, \{x_3\}, \{x_4\}\}$$

  On a 3 clusters.

  – Tableau des distances :

	x_1, x_2	x_3	x_4
x_1, x_2	0	11	2
x_3	11	0	9
x_4	2	9	0

  Le minimum est 2 donc on va fusionner (x_1, x_2) et x_4.

$$M = \{\{\{\{x_1\}, \{x_2\}\}, \{x_4\}\}, \{x_3\}\}$$

  On a 2 clusters.

  – Tableau des distances :

	$(x_1, x_2), x_4$	x_3
$(x_1, x_2), x_4$	0	9
x_3	9	0

  Le minimum est 9 donc on va fusionner $((x_1, x_2), x_4)$ et x_3.

$$M = \{\{\{\{x_1\}, \{x_2\}\}, \{x_4\}\}, \{x_3\}\}$$

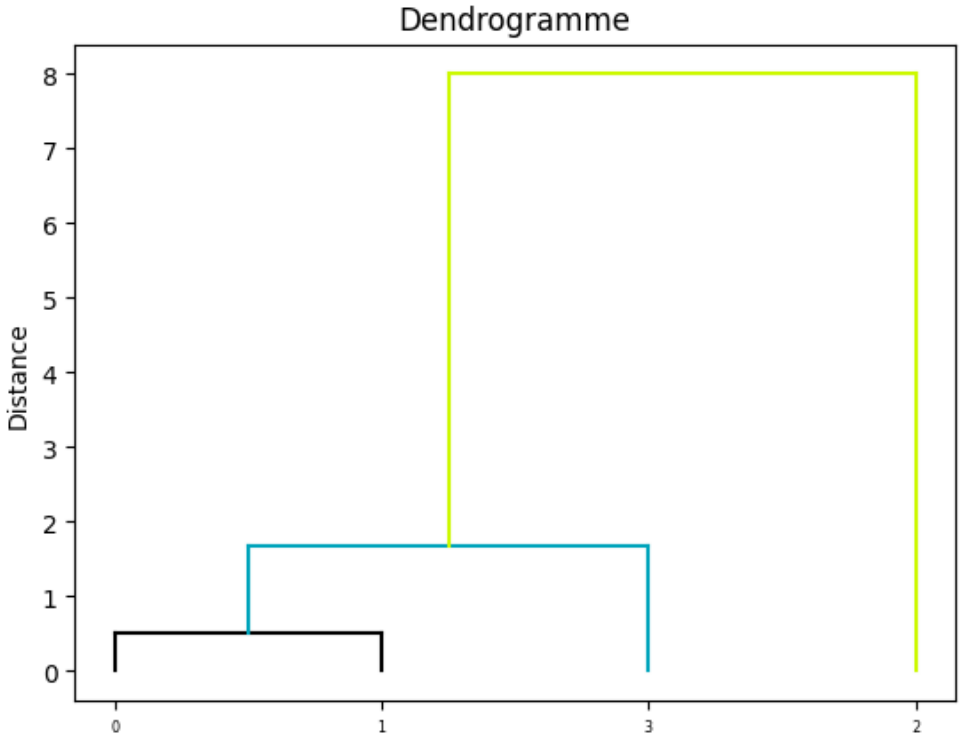
  On a 1 cluster, l'algorithme est donc terminé.

L'algorithme écrit tel quel ne renvoie rien, nous allons voir la sortie standard de la classification hiérarchique dans la prochaine partie : le dendrogramme.

2.5 Principe de calcul de dendrogramme

Le dendrogramme est l'arbre qui regroupe les clusters. Il est créé à partir de l'algorithme de clustering. L'axe des ordonnées représente la distance entre les clusters déterminée à partir du critère de similitudes. En haut de cet axe, il n'y un qu'un seul cluster, celui qui regroupe tous les individus, en bas, il y a autant de clusters que d'individus.

Si l'on reprend l'exemple précédent, on obtient le dendrogramme suivant, avec (x_1, x_2, x_3, x_4) représentés respectivement par $0, 1, 2, 3$:



Pour rajouter un dendrogramme à l'algorithme précédent, il faut rajouter la matrice de lien. Elle est de taille $(n-1) \times 4$ avec n le nombre d'individus. Chaque ligne correspond à une fusion de classe, et est éditée dès la fusion dans l'algorithme. Une ligne est de cette forme : [A, B, d, k] où A et B sont les deux clusters fusionnés, d est la distance entre ces deux clusters et k est le nombre total de singletons dans le cluster. On a alors juste besoin d'un programme qui, à partir de la matrice, trace les fusions de clusters dans l'ordre, à la hauteur de la distance entre les clusters.

Pour lire un dendrogramme, il suffit de le couper horizontalement de sorte à obtenir le nombre de clusters voulu. Nous verrons comment choisir le nombre de classes dans la partie sur les critères de coupe optimale 2.7.

2.6 Critères calculés à chaque étape pour fusionner les classes

Lors de la mise en œuvre de l'algorithme de clustering hiérarchique, il est nécessaire de définir un critère pour évaluer la similarité entre les clusters. En effet, les classes sont fusionnées itérativement pour former des groupes plus larges. À chaque étape, un critère est utilisé pour déterminer les deux classes qui seront fusionnées en une seule. Plusieurs critères de distance peuvent être utilisés, tels que la distance euclidienne, la distance minimale ou encore la distance de Ward. Une fois que la distance entre chaque paire de clusters a été calculée, il est nécessaire de choisir la paire de clusters à fusionner. Il existe plusieurs critères possibles pour déterminer la fusion des classes, les plus courants étant les suivants :

2.6.1 Critère de distance minimale

Le critère de la distance minimale est un critère de fusion de classes qui se base sur la distance minimale entre les éléments des deux classes à fusionner. Ce critère consiste à déterminer la distance minimale entre tous les points des deux classes et de prendre cette distance comme critère de fusion. Les deux groupes dont la distance minimale est la plus petite sont ceux qui seront fusionnés pour l'itération en cours. Formellement, la distance minimale entre deux clusters A et B est $d(A, B) = \min\{d(a, b) \mid a \in A, b \in B\}$

En reprenant l'exemple précédent, on voit d'après la matrice des distances que la plus petite distance vaut 1 et est entre le cluster x_1 et le cluster x_2. Donc d'après le critère de la distance minimale, ce sont ces deux clusters que nous fusionnerons avant de passer à la porchaine étape et de réiterer le processus

jusqu'à obtenir un seul et unique cluster.

Cependant, ce critère peut être sensible aux outliers. Les "outliers" (ou valeurs aberrantes) sont des points de données qui se trouvent à une distance anormalement éloignée des autres points de données. Ces points sont considérés comme des exceptions à la règle générale et peuvent être causés par des erreurs de mesure, des erreurs de saisie de données, des événements rares ou des phénomènes inhabituels.

Dans le cas du clustering hiérarchique, si un ou plusieurs outliers sont présents, ils peuvent être considérés comme des clusters à part entière et donc créer des clusters très étendus. Si ces points sont situés loin des autres points, ils peuvent également être considérés comme peu denses, car la densité est déterminée par la distance entre les points. Ainsi, l'ajout d'outliers peut modifier significativement la structure des clusters et rendre la méthode sensible aux valeurs aberrantes. Il est donc important de prendre en compte les limites de ce critère et de l'utiliser avec prudence, en particulier lorsque les données peuvent contenir des outliers.

2.6.2 Critère de distance maximale

Le critère de distance maximale, également appelé méthode de la liaison complète, considère la distance maximale entre deux observations appartenant à des clusters différents. L'idée est de fusionner deux clusters lorsque la distance maximale entre leurs observations est la plus petite possible. Les deux groupes dont la distance maximale est la plus petite sont ceux qui seront fusionnés pour l'itération en cours. Formellement, la distance maximale entre deux clusters A et B est $d(A, B) = \max\{d(a, b) \mid a \in A, b \in B\}$

Le critère de distance maximale est similaire à celui de la distance minimale, mais il est moins sensible aux outliers car il ne considère que la distance maximale entre les observations des deux clusters. Cela signifie que les clusters créés par cette méthode sont moins étendus et plus denses que ceux créés par la méthode de la distance minimale.

Lorsque l'on utilise le critère de la distance maximale pour fusionner des clusters, il peut arriver que certains points éloignés soient considérés comme appartenant à un même cluster simplement parce que leur distance maximale est inférieure à un certain seuil. Cela peut créer des clusters qui ne sont pas très cohérents en termes de similarité entre les observations. Par exemple, si nous avons deux groupes de données très éloignés, mais qu'ils sont reliés par une série de petites distances entre d'autres observations, le critère de distance maximale peut considérer qu'ils appartiennent à un même cluster, alors que cela ne devrait pas être le cas. Ce phénomène est appelé l'effet de chaîne ou d'encombrement, et peut affecter la qualité de la partition finale obtenue par clustering hiérarchique.

2.6.3 Critère de distance moyenne

Pour le critère de la distance moyenne, il s'agit de calculer la distance moyenne entre tous les points des deux classes à fusionner. Plus précisément, la distance entre deux classes A et B est calculée en prenant la moyenne de toutes les distances entre chaque point de la classe A et chaque point de la classe B. Lors de la fusion de deux classes, on choisit donc la paire de classes avec la plus petite distance moyenne. Formellement, la distance moyenne entre deux clusters A et B est $d(A, B) = \frac{1}{|A| \cdot |B|} \sum_{a \in A, b \in B} d(a, b)$

Ce critère est moins sensible aux outliers que la distance minimale, car il est basé sur la moyenne de toutes les distances, plutôt que sur la distance minimale entre deux points. Cependant, il peut également être affecté par des valeurs aberrantes qui ont des distances très élevées avec tous les autres points de la classe, car ces valeurs peuvent avoir un impact disproportionné sur la distance moyenne.

2.6.4 Critère de la distance de Ward

La méthode de Ward est un autre critère couramment utilisé pour le clustering hiérarchique agglomératif. Cette méthode vise à minimiser la variance totale à l'intérieur de chaque cluster en fusionnant des groupes qui minimisent l'augmentation de la variance totale. La distance entre deux groupes est définie comme la somme des carrés des différences entre les moyennes des observations de ces deux groupes, pondérée par le nombre d'observations dans chaque groupe. Plus précisément, si nous avons

deux clusters C_1 et C_2, avec n_1 et n_2 observations respectivement, et des moyennes m_1 et m_2, la distance de Ward entre ces deux clusters est donnée par :

$$d(C_1, C_2) = \frac{n_1 n_2}{n_1 + n_2} \cdot ||m_1 - m_2||^2 \tag{1}$$

Lorsque deux groupes sont fusionnés, la variance totale des observations dans le nouveau groupe est la somme des variances de chaque groupe précédent, plus la variance supplémentaire due à la fusion. La distance entre les groupes fusionnés est alors ajustée pour refléter cette variance supplémentaire.

En d'autres termes, la méthode de Ward favorise les fusions qui minimisent l'augmentation de la variance totale à l'intérieur de chaque groupe et donc la similarité entre les observations dans chaque groupe. Cette méthode est particulièrement utile lorsque les données ont des structures hiérarchiques avec des niveaux de variance différents entre les groupes.

Cependant, la méthode de Ward peut également souffrir de l'effet de chaîne, où les clusters peuvent être créés par des observations très éloignées les unes des autres, mais qui sont reliées par une série de petites distances entre les autres observations. Cela peut conduire à des clusters qui ne sont pas très cohérents du point de vue de la similarité entre les observations.

Il est important de noter que le choix du critère de fusion de classe peut avoir un impact significatif sur les résultats de clustering hiérarchique, et qu'il peut être nécessaire d'expérimenter avec différents critères pour trouver celui qui convient le mieux aux données.

2.7 Critères pour déterminer la coupe optimale en k clusters

Les critères pour déterminer la coupe optimale en k clusters sont une étape importante dans le processus de clustering hiérarchique, car ils permettent de déterminer le nombre optimal de clusters à utiliser dans l'analyse des données. Il existe plusieurs méthodes pour déterminer le nombre de clusters optimal, voici les deux principales approches :

2.7.1 La méthode du coude

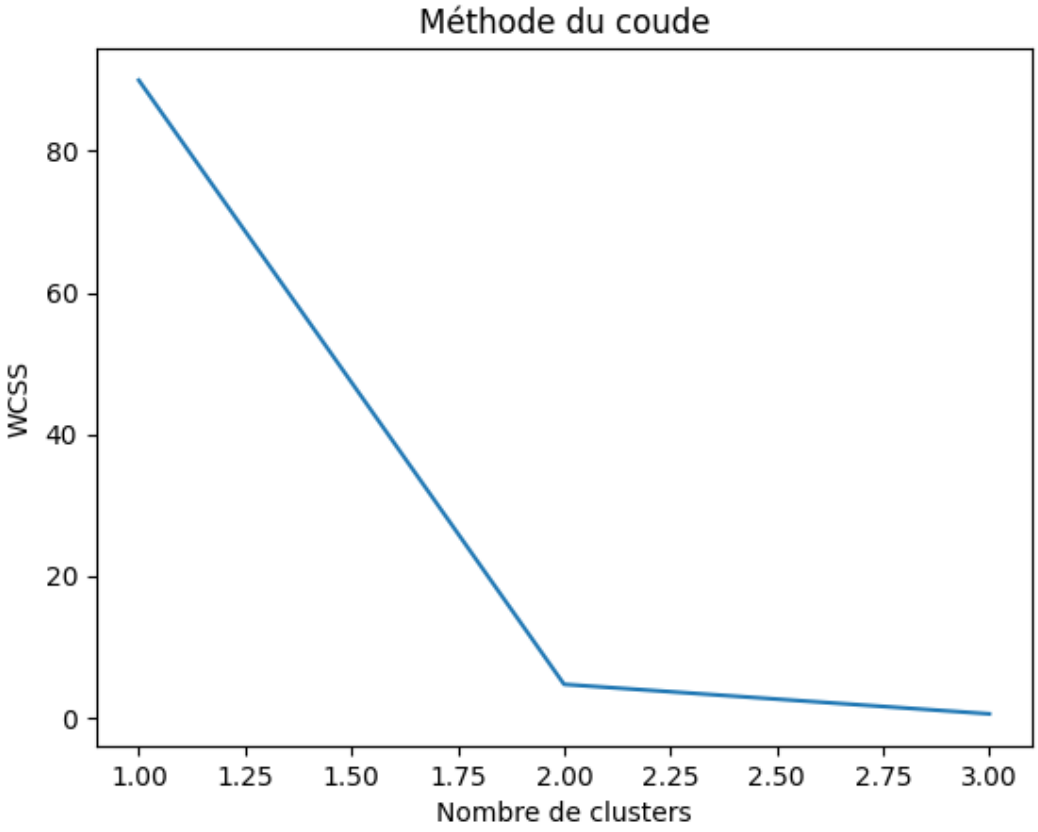
La méthode du coude est une technique courante pour déterminer le nombre optimal de clusters à utiliser lors de la segmentation de données. Elle est basée sur l'observation de la variation de la variance des différents clusters, aussi appelée somme des carrés des distances intra-classe, en fonction du nombre de clusters.

L'idée est de tracer un graphique représentant la variation de la variance des différents clusters en fonction du nombre de clusters. En général, plus le nombre de clusters augmente, plus la variance des différents clusters diminue. Cependant, à un certain point, l'ajout de nouveaux clusters n'apporte pas beaucoup d'amélioration en termes de réduction de la variance des différents clusters. Le point où l'ajout d'un nouveau cluster ne donne plus une amélioration significative est appelé le point de coude.

Ce point de courbure représente le nombre de clusters optimal. Le nombre de clusters est déterminé en prenant le point où la courbe commence à s'aplanir, indiquant que la diminution de la somme des carrés des distances intra-cluster est moins importante pour chaque cluster ajouté. Ce point est donc généralement visuellement identifiable sur le graphique, où la variance des différents clusters diminue fortement jusqu'à un certain nombre de clusters, puis cette diminution ralentit et commence à former une courbe en coude.

Il est important de noter que le choix du nombre de clusters n'est pas toujours facile à déterminer avec cette méthode et que des interprétations subjectives sont souvent nécessaires. Il est donc commun d'utiliser d'autres méthodes de validation en plus de la méthode du coude pour confirmer la solution optimale.

Reprenons notre exemple précédent et déterminons le nombre optimal de clusters dans ce cas-là avec la méthode du coude.
Tout d'abord, nous trions les données par ordre croissant puis nous traçons un graphique représentant la variation de la variance des différents clusters en fonction du nombre de clusters. Nous obtenons alors :



Nous observons que la courbe des variances diminue rapidement lorsque le nombre de clusters augmente jusqu'à 2 passants de 90 à 5 WCSS (Within-Cluster Sum of Squares), puis diminue plus lentement à mesure que le nombre de clusters augmente, passant de 5 à environ 0. Le point où la diminution des variances ralentit et commence à former une courbe en coude est le point de coude. Dans cet exemple, le point de coude se situe à 2 clusters, car l'ajout d'un troisième cluster n'apporterait que très peu d'amélioration en termes de réduction des variances. Nous choisissons donc 2 comme nombre optimal de clusters pour segmenter ces données.

2.7.2 Méthode de la silhouette

La méthode de silhouette permet de mesurer à quel point chaque point d'un cluster est similaire aux points de son propre cluster par rapport aux points des autres clusters. Il s'agit de se demander si un point x est proche des points du cluster auquel il appartient ? Est-il loin des autres points ? Pour répondre à la première question, on calcule la distance moyenne de x à tous les autres points du cluster U_k auquel il appartient, qu'on note $a(x) = \frac{1}{|U_k|-1} \sum_{u \in U_k, u \neq x} d(u, x)$

Pour répondre à la deuxième, on calcule la plus petite valeur que pourrait prendre $a(x)$, si x était assigné à un autre cluster. On la note $b(x) = \min_{l \neq k} \frac{1}{|U_l|} \sum_{u \in U_l} d(u, x)$

Si x a été correctement assigné, alors $a(x) < b(x)$. Le coefficient de silhouette est donné par $s(x) = \frac{b(x)-a(x)}{\max(a(x),b(x))}$

Pour obtenir la silhouette moyenne pour l'ensemble des objets d'un cluster, on calcule la moyenne de la silhouette de chaque objet. Ensuite, pour obtenir la silhouette moyenne pour l'ensemble des clusters, on calcule la moyenne de la silhouette moyenne pour chaque cluster.

La méthode de la silhouette permet de comparer les résultats de différents algorithmes de clustering avec différents nombres de clusters et de choisir la solution la plus appropriée. Une silhouette moyenne élevée indique une bonne séparation des clusters, tandis qu'une silhouette moyenne faible ou négative indique une mauvaise séparation des clusters. Pour évaluer un clustering, on peut calculer son coefficient de silhouette moyen.

Les avantages de cette méthode sont qu'elle est facile à interpréter, ne nécessite pas de connaissance a priori du nombre de clusters et peut être appliquée à différents types de données. Cependant, elle peut être sensible à la taille des clusters, ne prend pas en compte la structure globale des données et ne fournit pas d'informations sur la pertinence des clusters obtenus. Elle fonctionne particulièrement bien lorsque les données ne sont pas trop grandes et que les clusters sont bien séparés. Elle peut également être utile pour évaluer la qualité des clusters obtenus à partir d'autres méthodes de clustering.

Cependant, elle peut être moins appropriée pour les données avec des clusters très proches ou avec des formes complexes, car cela peut rendre difficile la mesure de la similarité entre les clusters et donc la précision des résultats obtenus.

3 Implémentation numérique et Analyse de jeu de données réelles

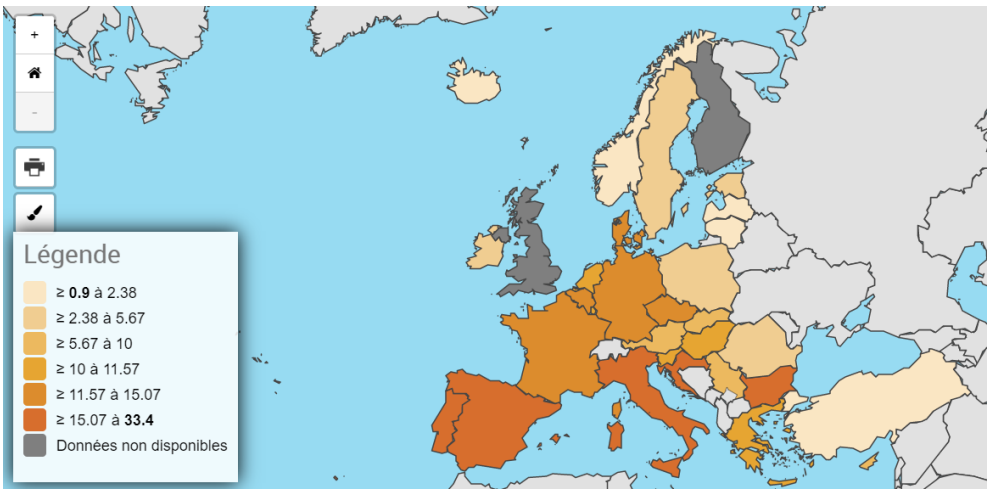
3.1 Description de la base de données

Dans cette partie, nous étudierons et implémenterons une base de données sur la consommation d'alcool dans l'Union Européenne, à l'exception de la Finlande, et incluant l'Islande, la Norvège et la Turquie, en 2019, provenant du site d'Eurostat. Cette base de données contient 30 lignes : la première ligne représente une moyenne des 27 pays de l'Union Européenne, les 29 lignes suivantes correspondent aux pays de l'UE, à l'exception de la Finlande, avec l'ajout de l'Islande, la Norvège et la Turquie. Nous avons créé deux tableaux distincts, l'un pour les hommes et l'autre pour les femmes, afin d'éviter tout problème de binarité. Nous avons choisi 20 critères basés sur l'éducation et la fréquence de consommation d'alcool comme variables explicatives. Les cinq fréquences de consommation sont : quotidien, hebdomadaire, mensuel, moins d'une fois par mois et jamais ou pas au cours des 12 derniers mois.

	Les 5 premières colonnes représentent la consommation d'alcool par les hommes, classée par leur fréquence de consommation					Les 5 colonnes suivantes représentent la consommation d'alcool par les hommes qui ont niveau d'éducation inférieur à l'enseignement p				
NC (Libellés)	Quotidiennement	Hebdomadaire	Mensuellement	Moins d'une fois par mois	Jamais ou pas au cours de	Quotidiennement	Hebdomadaire	Mensuellement	Moins d'une fois par mois	Jamais ou pas au cours de
GEO (Libellé										
Union europé	13,0	36,4	22,1	10,1	18,4	17,3	26,8	17,6	9,6	28,7
Belgique	13,5	47,2	16,2	6,8	16,5	12,6	34,1	16,4	8,9	28,0
Bulgarie	17,4	34,0	23,6	8,8	16,2	20,0	26,3	18,1	7,9	27,7
Tchéquie	12,9	44,2	22,3	10,7	9,9	6,7	22,5	21,8	18,7	30,2
Danemark	12,4	45,2	26,0	9,3	7,1	11,5	33,8	24,5	13,4	16,8
Allemagne (j	11,7	38,2	23,5	9,7	16,9	9,2	20,7	29,0	12,6	28,5
Estonie	2,4	30,7	30,7	19,0	17,1	3,3	24,3	24,9	19,0	28,5
Irlande	3,6	45,0	20,8	10,8	19,8	5,4	36,8	16,2	10,3	31,3
Grèce	10,0	35,5	28,1	8,2	18,2	14,3	24,5	22,7	9,0	29,5
Espagne	20,2	27,0	18,6	9,8	24,4	22,9	18,8	16,9	10,0	31,4
France	14,6	41,5	18,6	7,2	18,1	16,9	33,8	17,2	7,7	24,4
Croatie	18,2	26,5	21,0	9,2	25,1	25,2	19,3	16,2	6,9	32,5
Italie	18,8	37,4	16,5	5,8	21,5	22,8	31,0	10,7	5,6	29,9
Chypre	7,4	25,8	38,2	15,9	12,8	10,0	19,1	29,8	17,1	24,0
Lettonie	2,3	20,0	34,5	22,6	20,6	2,0	14,6	23,0	20,3	40,1
Lituanie	1,4	19,5	37,8	20,0	21,3	1,9	14,7	22,1	15,0	46,3
Luxembourg	11,5	50,1	18,6	7,2	12,6	13,9	34,3	17,4	12,4	22,1
Hongrie	11,5	28,8	23,6	17,7	18,6	10,9	16,5	19,1	21,5	32,1
Malte	11,4	34,9	19,1	11,9	22,7	14,6	28,7	15,7	12,8	28,1
Pays-Bas	10,0	55,0	9,8	8,0	17,1	9,4	41,8	9,2	8,7	31,0
Autriche	9,0	43,1	22,4	9,7	15,7	9,9	24,1	19,0	13,3	33,6
Pologne	3,0	27,0	34,4	18,6	17,0	3,1	17,7	20,0	19,5	39,7
Portugal	33,4	28,2	14,0	7,3	17,1	42,9	20,0	9,9	6,4	20,8
Roumanie	5,6	32,2	32,0	12,4	17,7	7,1	25,2	25,3	12,5	29,8
Slovénie	10,6	33,0	25,0	12,6	18,9	14,4	20,8	21,0	14,8	29,0
Slovaquie	7,5	30,6	25,1	17,1	19,7	6,3	13,9	21,0	12,8	46,1
Suède	2,5	40,5	27,2	12,8	17,1	3,0	27,5	27,0	15,3	27,3
Islande	1,5	27,2	37,0	16,0	18,3	1,2	17,6	34,7	19,7	26,8
Norvège	1,7	41,2	30,6	14,3	12,2	1,5	29,5	29,8	18,9	20,3
Turquie	0,9	6,0	7,9	8,5	76,7	1,2	4,7	5,9	7,0	81,3

Les 5 colonnes suivantes représentent la consommation d'alcool par les hommes qui ont arrêté leurs études en deuxième cycle de l'ens					Les 5 colonnes suivantes représentent la consommation d'alcool par les hommes qui ont effectué des études supérieurs (niveaux 5-8), c				
Quotidiennement	Hebdomadaire	Mensuellement	Moins d'une fois par mois	Jamais ou pas au cours de	Quotidiennement	Hebdomadaire	Mensuellement	Moins d'une fois par mois	Jamais ou pas au cours de
10,7	37,8	24,8	11,1	15,6	11,6	45,9	23,1	8,9	10,5
11,3	47,5	17,2	8,3	15,7	16,3	56,3	15,3	3,4	8,7
18,1	38,1	23,3	7,8	12,6	12,7	31,1	30,5	12,5	13,2
14,9	44,5	22,6	9,7	8,3	8,7	55,1	21,7	10,0	4,4
11,4	46,2	27,2	9,8	5,4	14,2	49,6	25,5	6,9	3,7
11,9	37,4	22,9	10,4	17,5	12,8	49,8	21,5	6,9	9,1
2,4	31,3	31,1	19,4	15,8	2,0	33,2	33,5	18,6	12,6
2,2	44,1	23,7	12,0	18,1	3,5	52,9	21,2	9,9	12,5
8,7	40,0	30,0	8,0	13,3	7,8	40,0	30,8	7,5	13,9
17,2	31,8	20,4	9,8	20,9	18,0	37,3	20,2	9,4	15,2
10,5	45,9	22,2	7,8	13,7	13,0	52,1	19,1	6,0	9,8
17,6	27,2	19,6	9,8	25,9	14,0	30,8	30,0	9,2	16,1
15,1	42,0	20,7	5,8	16,4	16,3	45,5	23,5	6,4	8,2
7,5	25,1	39,8	15,8	11,8	5,5	31,1	41,7	15,2	6,5
2,6	19,2	37,1	23,8	17,4	1,8	26,3	37,4	21,3	13,1
1,7	19,7	40,2	19,2	19,2	0,8	21,0	39,8	23,1	15,3
11,6	49,0	18,8	8,2	12,5	10,2	57,2	19,3	4,8	8,6
12,4	27,5	23,5	18,1	18,5	9,7	40,1	26,7	14,0	9,5
9,6	38,6	22,9	11,2	17,7	6,4	45,0	23,1	10,4	15,1
9,7	58,5	9,6	9,0	13,2	11,0	64,1	11,1	5,9	7,9
9,4	43,5	22,9	9,4	14,9	8,1	51,0	23,1	8,8	9,1
3,7	27,4	36,7	18,3	14,0	1,0	31,6	37,1	19,3	11,1
22,4	35,5	18,3	9,3	14,5	17,1	45,5	22,0	7,2	8,2
5,6	36,0	33,1	11,6	13,6	3,3	27,8	38,8	15,8	14,4
13,2	29,9	20,7	11,7	24,4	8,8	37,1	27,3	12,3	14,5
8,1	33,6	25,3	16,4	16,6	6,0	28,7	26,3	21,1	17,9
2,5	45,3	28,1	13,0	11,1	2,4	48,9	27,9	10,6	10,2
1,8	29,4	38,4	14,1	16,3	1,6	34,2	37,6	14,7	11,9
2,1	41,2	34,0	14,4	8,2	1,7	53,0	27,5	9,3	8,4
0,8	6,2	8,3	8,2	76,5	0,6	8,4	11,9	12,5	66,7

Les cinq premières colonnes représentent la consommation d'alcool des hommes, classée par leur fréquence de consommation. Les 15 colonnes suivantes décrivent les différents niveaux d'éducation (basés sur la Classification Internationale Type de l'Éducation (CITE 2011)) pour chaque fréquence de consommation d'alcool. Les cinq colonnes suivantes représentent la consommation d'alcool des hommes ayant un niveau d'éducation inférieur à l'enseignement primaire, ceux ayant un enseignement primaire et ceux ayant suivi le premier cycle de l'enseignement secondaire (niveaux 0 à 2 de la CITE 2011, le niveau 2 correspondant au collège en France), classées selon leur degré de consommation d'alcool. Ensuite, les cinq colonnes suivantes représentent la consommation d'alcool des hommes ayant arrêté leurs études au deuxième cycle de l'enseignement secondaire (niveau 3, correspondant au lycée en France) et ceux ayant arrêté en enseignement post-secondaire non-supérieur (niveau 4, cela correspond au CAP, par exemple), classées selon leur périodicité de consommation d'alcool. Enfin, les cinq dernières colonnes correspondent à la consommation d'alcool des hommes ayant effectué des études supérieures (niveaux 5-8 de la CITE 2011, c'est-à-dire de la licence à la thèse), classées selon leur fréquence de consommation.

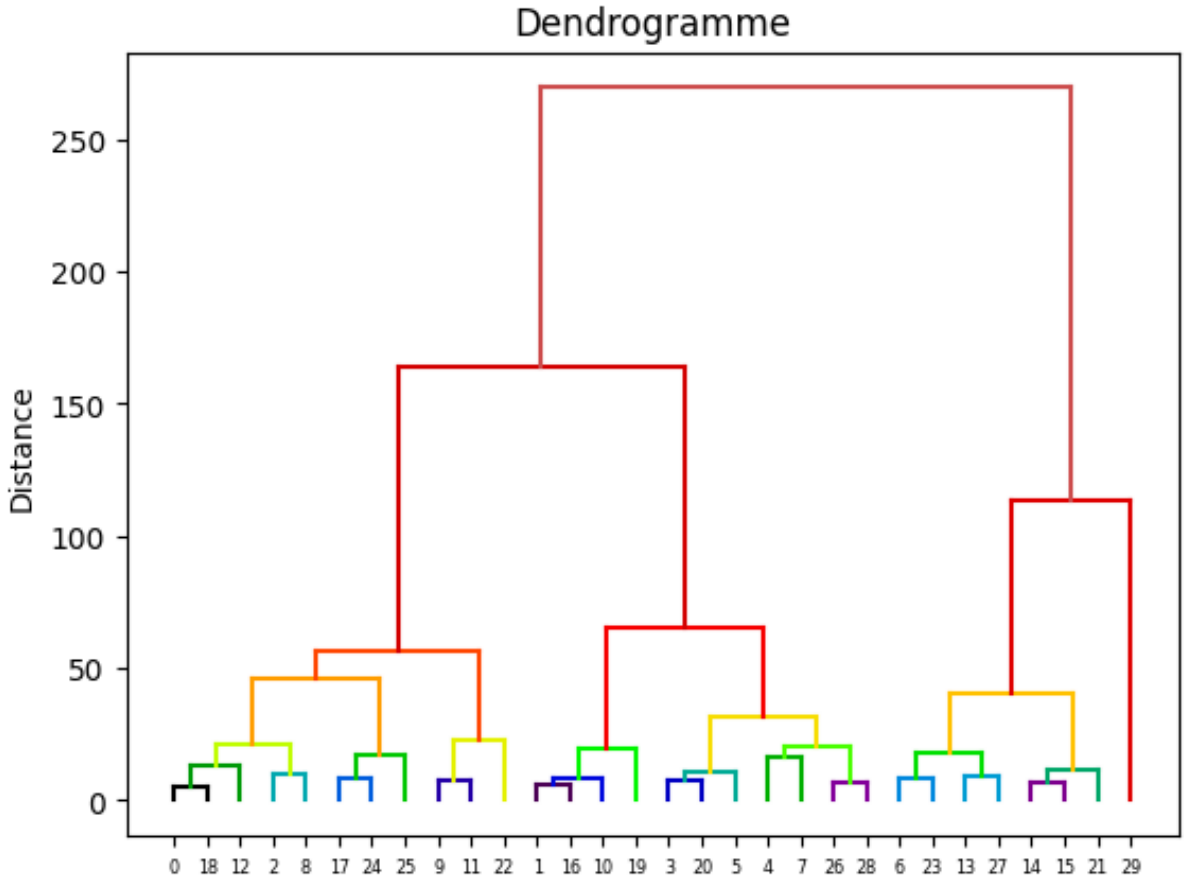


Carte représentant la consommation quotidienne d'alcool

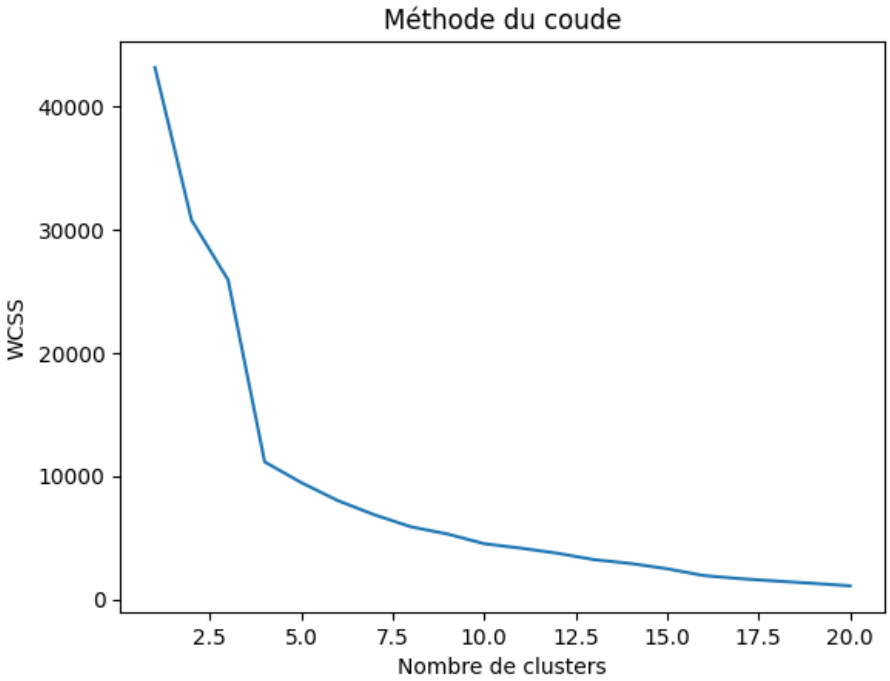
3.2 Implémentation de l'algorithme sur la base de données

3.2.1 Homme : critère de Ward et distance euclidienne

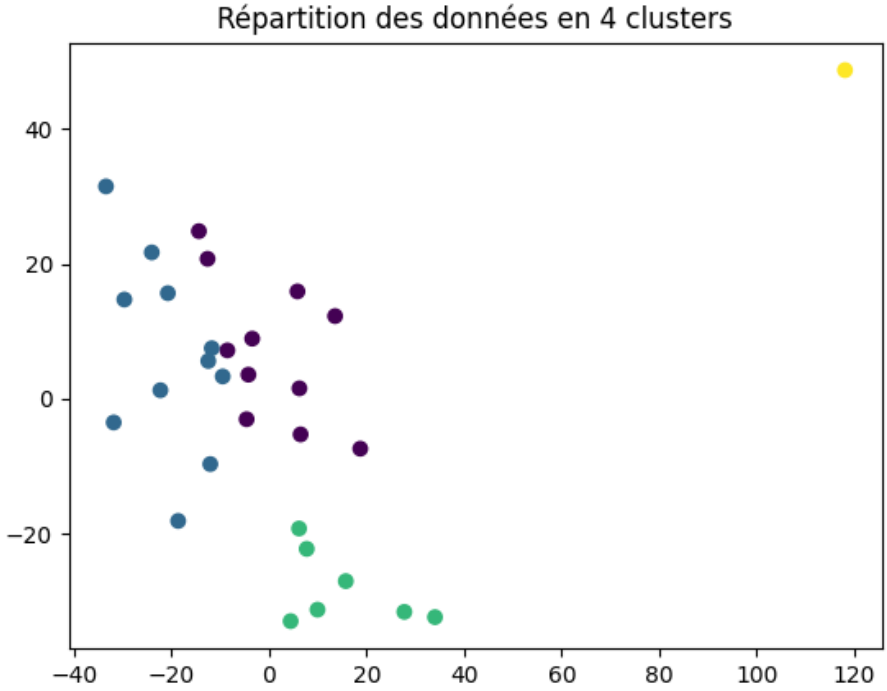
Ici, l'algorithme a été appliqué sur la base de données des hommes en utilisant le critère de Ward et la distance euclidienne. On obtient alors le dendrogramme suivant :



En utilisant la méthode du coude pour déterminer la coupe optimale du dendrogramme, on obtient que le nombre optimal de clusters est de 4 :



Nous avons donc 4 clusters :



Les 4 clusters sont les suivants :

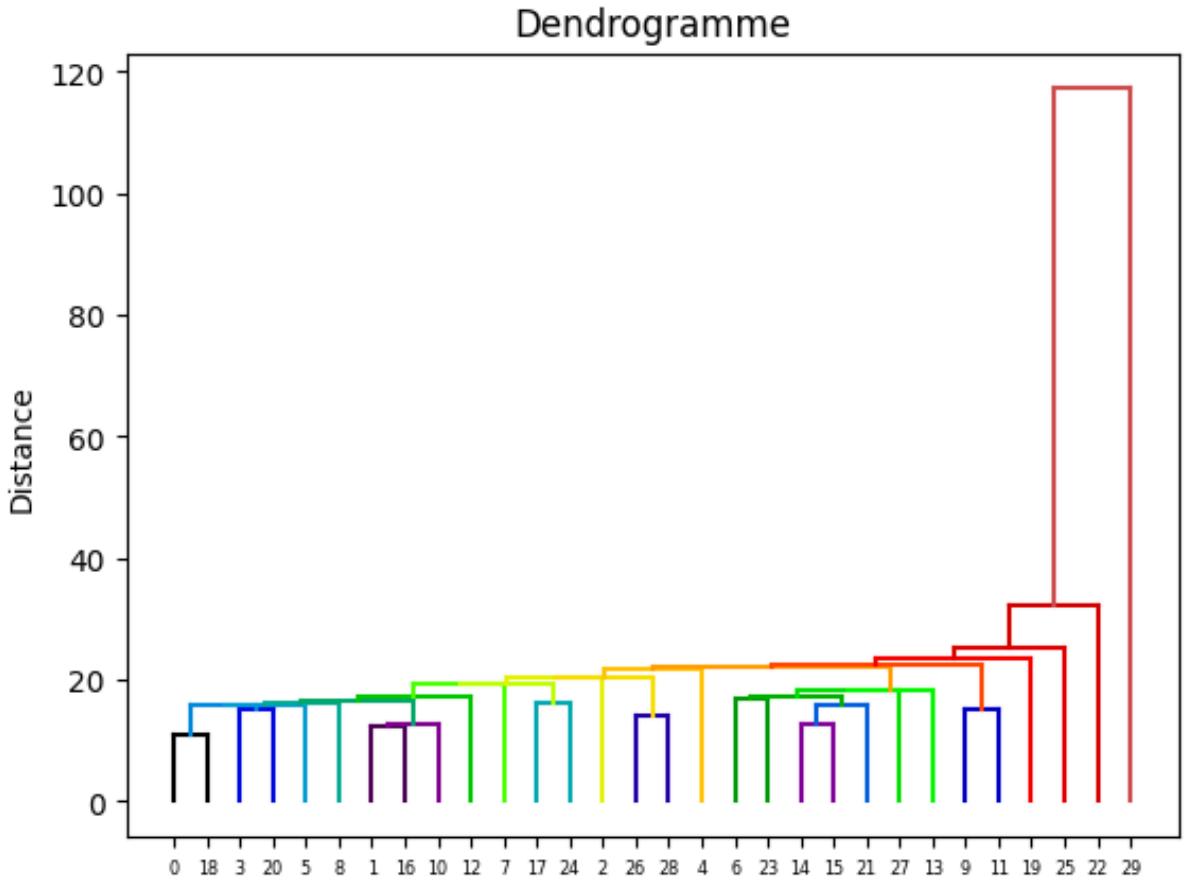
- le premier regroupe 11 individus : la moyenne sur l'UE, Malte, l'Italie, la Bulgarie, la Grèce, la Hongrie, la Slovénie, la Slovaquie, l'Espagne, la Croatie et le Portugal
- le second cluster est constitué de 11 pays : Belgique, Luxembourg, France, Pays-Bas, République Tchèque, Autriche, Allemagne, Danemark, Irlande, Suède, Norvège
- le troisième réunit 7 pays : Estonie, Roumanie, Chypre, Islande, Lettonie, Lituanie, Pologne
- le dernier est composé uniquement d'un seul pays : la Turquie.

Il est intéressant de noter la proximité géographique des pays au sein de chaque cluster, ainsi que la composition des clusters qui peut s'expliquer par des facteurs tels que l'histoire commune et les habitudes culturelles. En 2019, date de l'étude et de la base de données, l'étude Optimar indiquait qu'il y avait

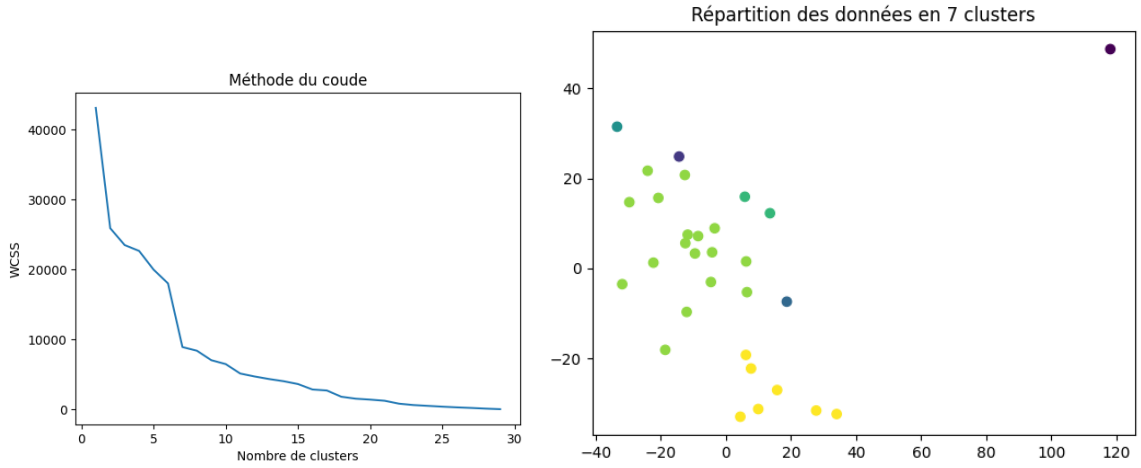
89,5% de musulmans en Turquie, ce qui peut expliquer les taux de consommation d'alcool relativement bas et la singularité de ce pays par rapport aux autres, aucun autre pays n'ayant un pourcentage aussi élevé de musulmans.

3.2.2 Homme : critère de distance minimale et distance euclidienne

Observons ce qu'il se passe en essayant avec d'autres paramètres : critère de distance minimale et distance euclidienne :



La coupe optimale du dendrogramme est de 7 clusters :



Les 7 clusters sont les suivants :

- le premier regroupe 17 individus : la moyenne sur l'UE, Malte, la République Tchèque, l'Autriche, l'Allemagne, la Grèce, la Belgique, le Luxembourg, la France, l'Italie, l'Irlande, la Hongrie, la Slovénie, la Bulgarie, la Suède, la Norvège et le Danemark

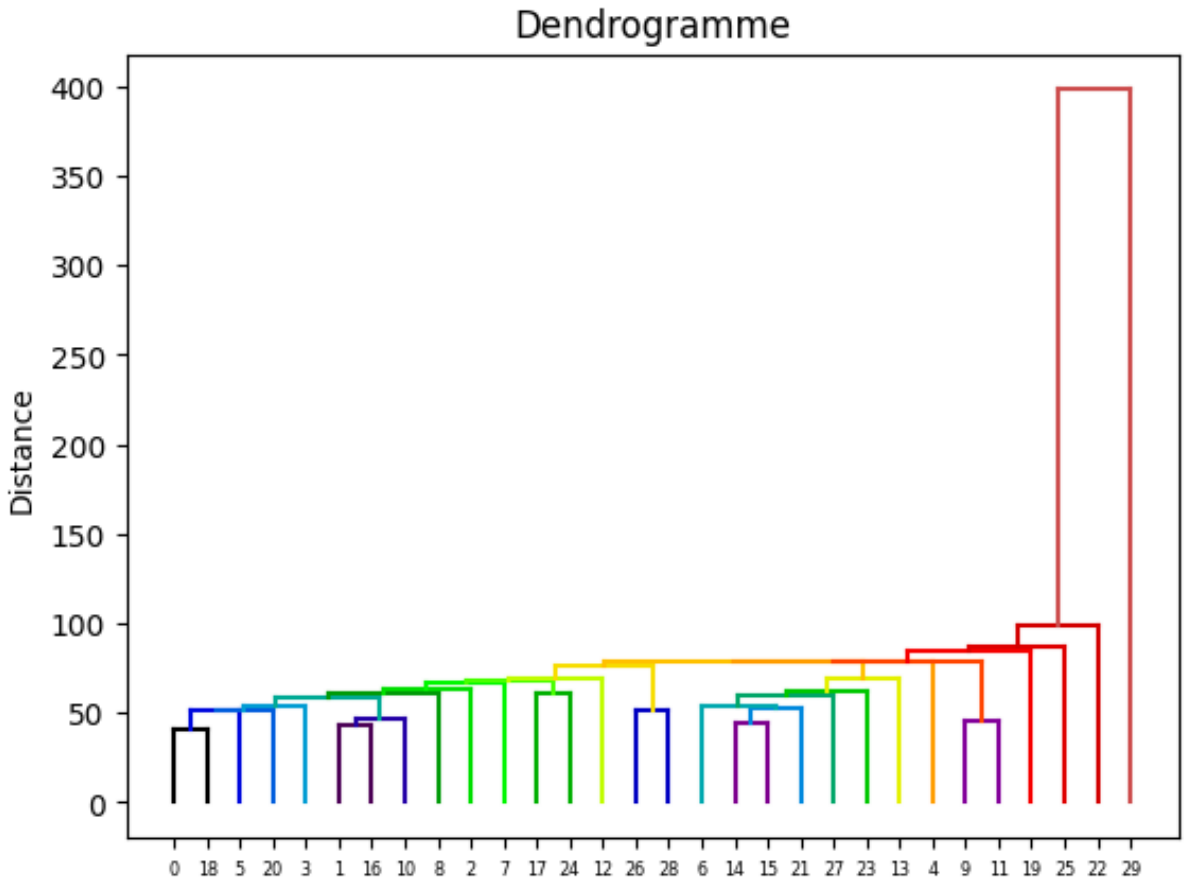
- le deuxième réunit 2 pays : l'Espagne et la Croatie
- le troisième cluster est constitué de 7 pays : Estonie, Roumanie, Lettonie, Lituanie, Pologne, Islande, Chypre
- les quatre derniers cluters sont est composé uniquement d'un seul pays chacun :
    1. les Pays-Bas
    2. la Slovaquie
    3. le Portugal
    4. la Turquie

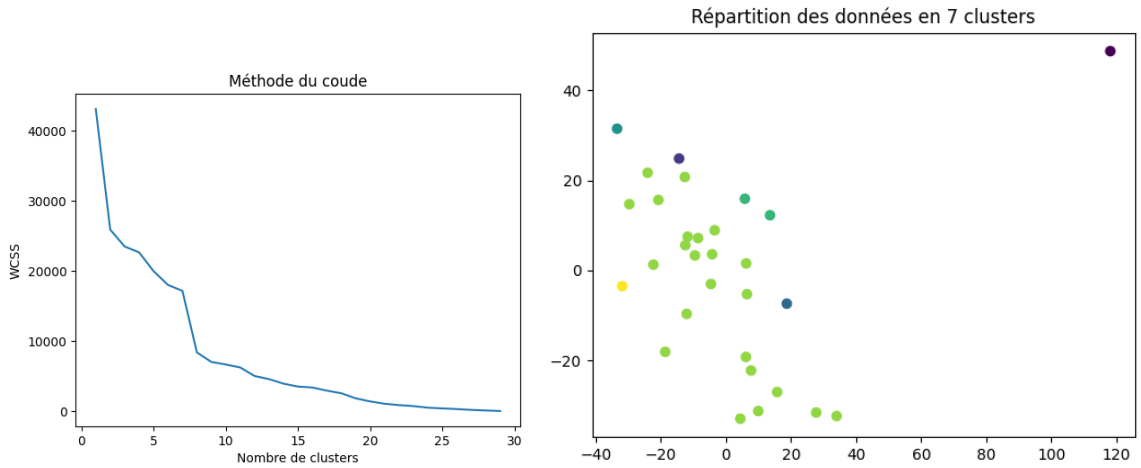
En modifiant le critère de distance entre les clusters de la distance de Ward à la distance minimale tout en conservant la distance euclidienne, le nombre de clusters est différent : 4 contre 7. Les deux premiers clusters de la méthode de Ward ont quasiment fusionné pour former le premier cluster de la distance minimale, et les cinq pays qui ne sont pas inclus dans les 17 individus du premier cluster forment soit un cluster seul, soit un cluster de deux. Une similitude importante : le troisième cluster reste le même, seuls les rapprochements à l'intérieur du cluster étant différents. On peut également noter que les distances entre les individus sur le dendrogramme sont beaucoup plus petites, ce qui implique que les individus sont plus proches les uns des autres.

Ces résultats sont cohérents avec la théorie sur le critère de distance minimale (section 2.6.1), qui montre sa sensibilité aux outliers. En effet, sur le graphe affichant les clusters, on constate que les clusters seuls sont tous (à l'exception d'un) assez éloignés des autres individus et peuvent être considérés comme des "outliers".

3.2.3 Homme : critère de distance minimale et distance de Manhattan

On peut se demander l'impact que peut avoir un changement de distance entre les individus, tout en conservant le même critère de distance pour les clusters.

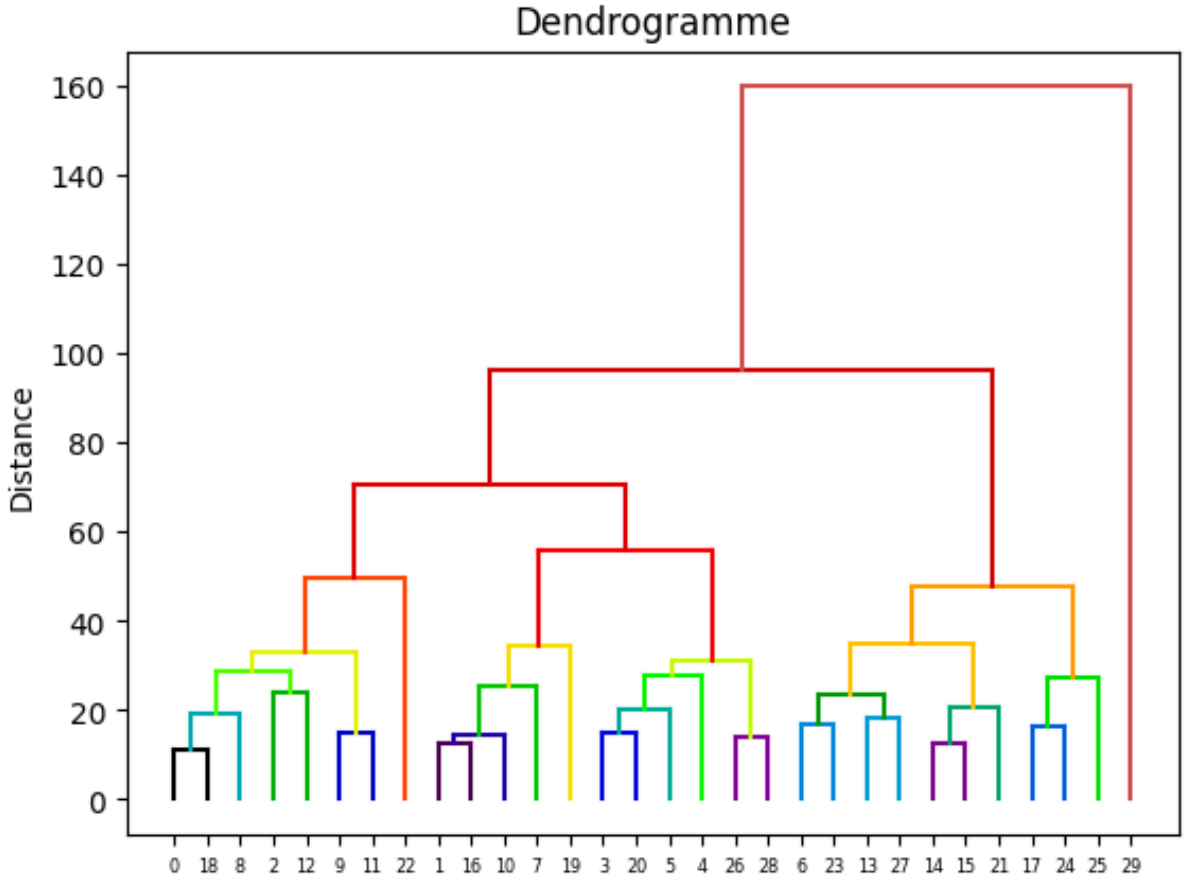


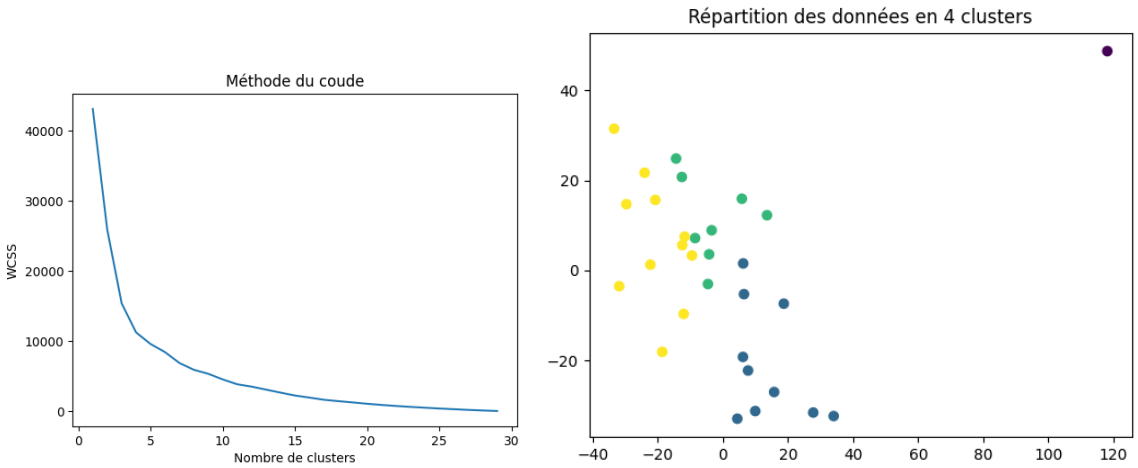


En conservant les 7 clusters, on constate que la distance minimale entre les individus a eu un impact sur la composition des clusters. Le deuxième cluster, composé de deux pays, ainsi que les quatre derniers clusters, composés chacun d'un seul pays, restent identiques. Cependant, le premier et le troisième cluster ont été regroupés pour former un cluster de 23 individus, tandis qu'un nouveau cluster composé uniquement du Danemark a été créé. Ainsi, nous avons maintenant un cluster important de 23 individus, suivi d'un cluster de 2 pays, puis de 5 clusters composés chacun d'un individu. Bien que la distance de Manhattan permette d'atténuer l'effet des valeurs aberrantes, il est possible que notre échantillon ne soit pas adapté au critère de distance minimale, en raison de la présence de trop d'outliers.

3.2.4 Homme : critère de distance maximale et distance euclidienne

Dans cette section, regardons à nouveau les effets d'un changement de critère de distance sur les clusters, tout en conservant la distance euclidienne entre les individus pour comparer avec les résultats trouvés dans la section 3.2.1 et la section 3.2.2.





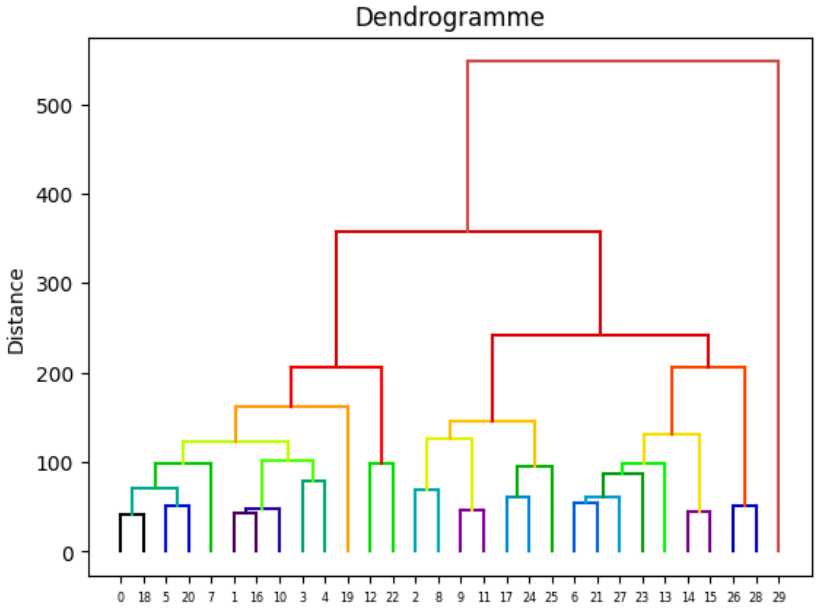
Comme pour la méthode de Ward, nous avons 4 clusters :

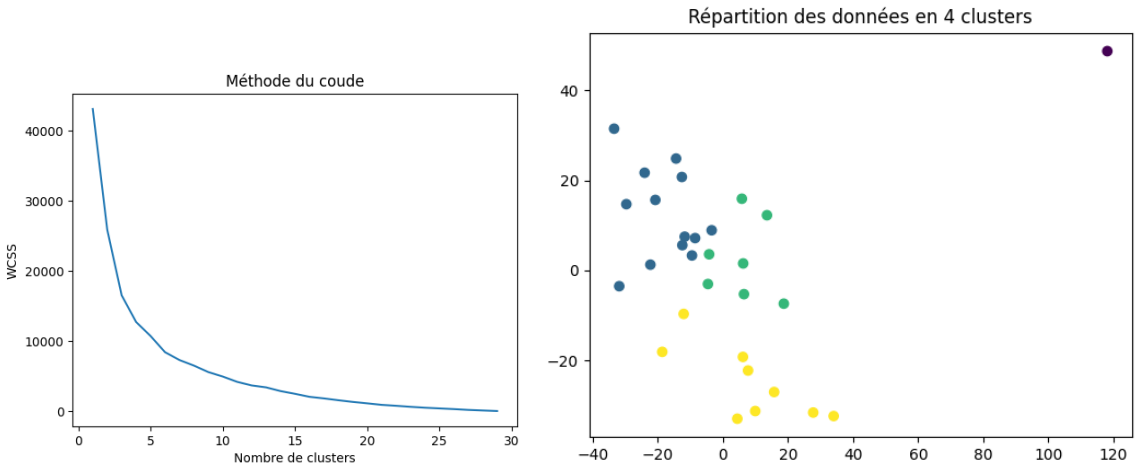
- le premier regroupe 8 individus : la moyenne sur l'UE, Malte, l'Italie, la Bulgarie, la Grèce, l'Espagne, la Croatie et le Portugal
- le second cluster est constitué de 11 pays : Belgique, Luxembourg, France, Pays-Bas, République Tchèque, Autriche, Allemagne, Danemark, Irlande, Suède, Norvège
- le troisième réunit 10 pays : Estonie, Roumanie, Chypre, Islande, Lettonie, Lituanie, Pologne, la Hongrie, la Slovénie et la Slovaquie
- le dernier est composé uniquement d'un seul pays : la Turquie

Mis à part trois pays (la Hongrie, la Slovénie et la Slovaquie), qui ont été déplacés du premier au troisième cluster, les clusters obtenus avec la distance maximale sont identiques à ceux obtenus avec la méthode de Ward. Les distances entre les individus sont également similaires. Cette similarité implique que l'on observe les mêmes différences et similitudes entre les critères maximal et minimal, et le critère de Ward et minimal, tel que présenté dans la section 3.2.2. En outre, comme indiqué dans la section 2.6.2, le critère de distance maximale est moins sensible aux valeurs aberrantes. Toutefois, il y a un petit effet de chaîne observé sur la Hongrie, la Slovénie et la Slovaquie, qui visuellement sont plus proches du premier cluster, mais sont classés dans le troisième cluster.

3.2.5 Homme : critère de distance maximale et distance de Manhattan

À présent, examinons l'effet de la modification de la distance entre les individus tout en maintenant le même critère de distance pour les clusters.





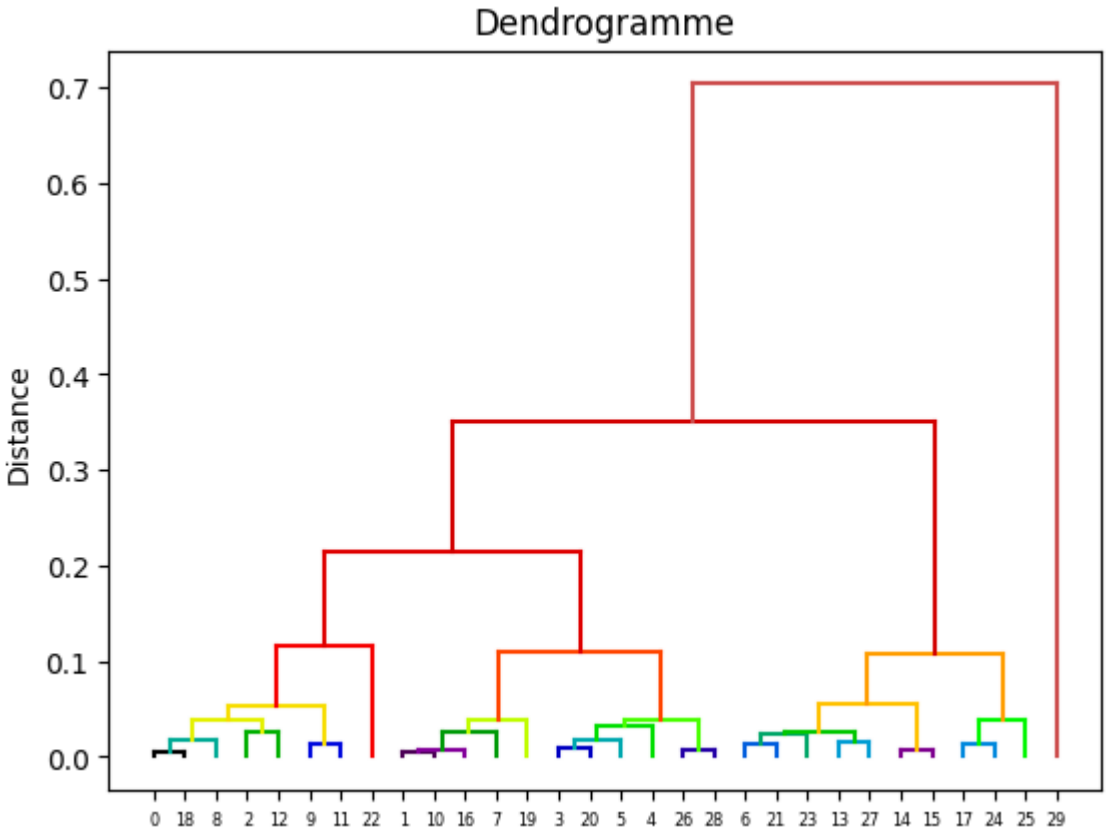
Comme avec la distance euclidienne, on obtient 4 clusters. Cependant, leur composition est différente :

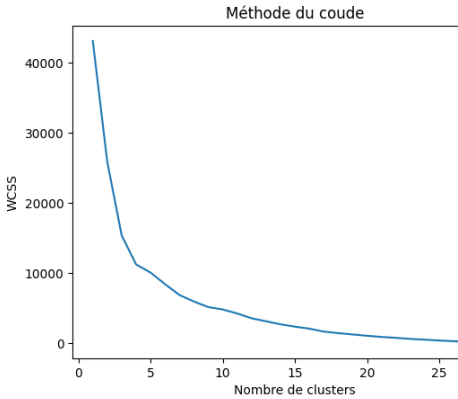
- le premier regroupe 13 individus : la moyenne sur l'UE, Malte, l'Allemagne, l'Autriche, l'Irlande, la Belgique, le Luxembourg, la France, la République Tchèque, le Danemark, les Pays-Bas, l'Italie et le Portugal
- le second cluster est constitué de 7 pays : la Bulgarie, la Grèce, l'Espagne, la Croatie, la Hongrie, la Slovénie et la Slovaquie
- le troisième réunit 9 pays : Estonie, Pologne, Roumanie, Chypre, Islande, Lettonie, Lituanie, Suède, Norvège
- le dernier est composé toujours uniquement de la Turquie

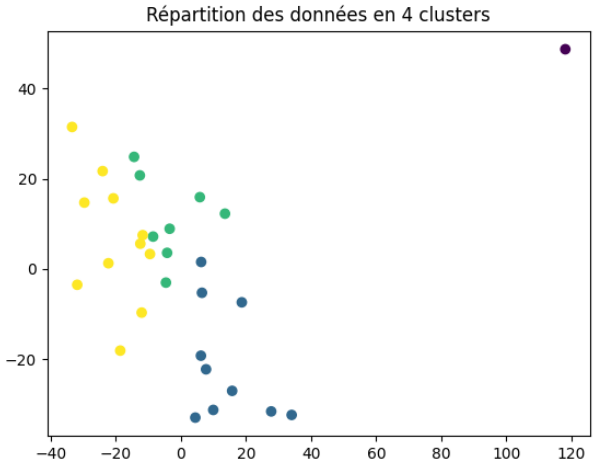
Bien que l'on puisse observer des similarités, les clusters sont assez différents en termes de taille et de composition, mais surtout, les individus sont très éloignés les uns des autres par rapport au calcul avec la distance euclidienne. De même que pour la comparaison en gardant la distance minimale et en changeant de la distance euclidienne à la distance de Manhattan (section 3.2.3), le changement de distance entre les individus a un impact important sur la méthode de clustering.
Il est à noter que lors de la comparaison avec la section 3.2.3, où la distance entre les individus est la même, mais le critère pour les clusters diffère, les résultats sont hétéroclites, avec un nombre de clusters différent, une composition de clusters très différente et des distances entre les individus encore plus exacerbées avec le critère de distance maximale.

3.2.6 Homme : critère de distance maximale et distance cosinus

Enfin, on étudie une nouvelle distance et comparer, à critère fixé, quel est l'impact de la distance cosinus, qui a une méthode de calcul différente des autres distances que l'on a utilisées, puisqu'elle regarde l'angle entre les vecteurs.







Pour toutes les distances, nous obtenons quatre clusters :

- le premier regroupe 8 individus : la moyenne sur l'UE, Malte, la Grèce, la Bulgarie, l'Italie, l'Espagne, la Croatie et le Portugal
- le second cluster est constitué de 11 pays : la Belgique, la France, le Luxembourg, l'Irlande, les Pays-Bas, la République Tchèque, l'Autriche, le Danemark, l'Allemagne, la Suède et la Norvège
- le troisième réunit 10 pays : Estonie, Pologne, Roumanie, Chypre, Islande, Lettonie, Lituanie, Hongrie, Slovénie, Slovaquie
- le dernier est composé uniquement de la Turquie

Ces clusters sont identiques à ceux obtenus avec la distance euclidienne. En utilisant la distance cosinus, on peut évaluer le degré de corrélation entre les variables. Ici, les distances sont comprises entre 0 et 0,7. La distance entre la Turquie et les autres pays est proche de 1, ce qui indique que les individus sont assez éloignés (leurs vecteurs sont proches de l'orthogonalité). Cependant, dans l'ensemble, les autres pays sont assez similaires (corrélation positive), car la distance est comprise entre 0 et 0,4.

3.2.7 Conclusion sur l'implémentation numérique

En conclusion, l'application de l'algorithme de clustering hiérarchique sur les données de consommation d'alcool en Europe a permis de mettre en évidence des regroupements géographiques entre les pays, reflétant leurs similitudes culturelles. Les choix de distance et de critères de regroupement ont un impact significatif sur la composition et le nombre de clusters obtenus. Néanmoins, certaines tendances récurrentes ont été observées, telles que la position de Malte comme pays le plus proche de la moyenne de l'UE et la formation d'un cluster unique pour la Turquie. Cette étude met en lumière l'importance de la compréhension des critères de clustering et de l'interprétation des résultats pour une utilisation pertinente et efficace de cette technique en analyse de données.

4 Conclusion

En conclusion, l'algorithme de clustering hiérarchique est une approche efficace pour regrouper des données en clusters hiérarchiques. Cette méthode commence par considérer chaque point de données comme un cluster individuel et fusionne progressivement les clusters les plus proches jusqu'à ce qu'un nombre prédéfini de clusters soit atteint. Les résultats obtenus montrent que cette approche peut être utilisée pour identifier des structures de données complexes et révéler des relations intéressantes entre les données. Cependant, comme pour tout algorithme de clustering, il est important de choisir judicieusement les paramètres tels que la mesure de distance et le critère de fusion pour obtenir des résultats optimaux.

En somme, l'algorithme de clustering hiérarchique utilisant la méthode ascendante est un outil puissant pour l'analyse des données, mais il existe d'autres approches de clustering différentes qui peuvent être utilisées. La méthode du k-means et le DBSCAN en sont des bien connues qui peuvent offrir des résultats différents et intéressants. Il est important de choisir la méthode la plus appropriée en fonction des objectifs de l'analyse.

5 Références

- Source des données : Eurostat : `https://ec.europa.eu/eurostat/databrowser/view/HLTH_EHIS_AL1E/default/table?lang=fr`
- Clustering — Python pour la data-science : `https://pythonds.linogaliana.fr/clustering/`
- OpenClassRooms : Partitionnez vos données avec un algorithme de clustering hiérarchique : `https://openclassrooms.com/fr/courses/4379436-explorez-vos-donnees-avec-des-algorithmes-non-su 4379561-partitionnez-vos-donnees-avec-un-algorithme-de-clustering-hierarchique`
- EpiMed Open Course - Clustering hiérarchique, méthode non supervisée du machine learning — Cycle avancé IA n9 : `https://www.youtube.com/watch?v=_CYaoWTc8HU&t=212s`
- Gabor J. Székely et Maria L. Rizzo, "*Hierarchical clustering via Joint Between-Within Distances : Extending Ward's Minimum Variance Method*", Journal of Classification, septembre 2005
- Leo Breiman, J. H. Friedman, R. A. Olshen et C. J. Stone, "*Classification and Regression Trees, Monterey*", CA, Wadsworth, 1984
- Ryan P. Adams, "*Hierarchical Clustering*" : `https://www.cs.princeton.edu/courses/archive/fall18/cos324/files/hierarchical-clustering.pdf`
- Ali Seyed Shirkhorshidi, Saeed Aghabozorgi et Teh Ying Wah, "*A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data*" : `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4686108/`
- F.-G. Carpentier, "*Classification Ascendante Hiérarchique*" : `http://www.normalesup.org/~carpenti/Cours/PSY-M1-Ana-mult-4.pdf`