**Etude de marché avec Python**

**4** - Réalisation d’une classification ascendante hiérarchique (CAH)

La première méthode que nous allons utiliser est la **classification ascendante hiérarchique** qui consiste à regrouper, de manière itérative, les individus (ici les pays), à partir de la « distance » qui les séparent.

Les distances entre individus sont regroupées dans une matrice calculée à partir des données initiales.

J’ai donc calculé la matrice des distances, aussi appelée matrice de liaison, puis j’ai tracé le résultat des différentes itérations sous forme d'un dendrogramme.

**5** - Réalisation d’un Coefficient de Silhouette

Pour choisir un nombre de *clusters* de manière plus objective, nous allons nous appuyer sur le **coefficient de Silhouette** qui rend compte de la qualité d'un regroupement. Il mesure à quel point les individus d'un même groupe sont similaires entre eux par rapport aux objets des autres groupes.

La méthode du score de Silhouette nous indique le nombre optimal de *clusters* est 3. Néanmoins, cela est trop peu par rapport au nombre de pays à grouper. Je choisis donc la deuxième meilleure valeur qui est de 13.

**6** - On visualise assez bien ce qui caractérise chacun des groupes. Les *clusters* 1 à 4 ont une majorité d'indicateurs ayant des valeurs assez élevées.

Nous n'allons retenir que ces derniers pour afficher un peu plus de détail grâce à une fonction que j’ai créer qui nous permettra de reproduire le graphique avec les résultats des autres classifications.

**7** - Les groupes 1 et 2 ont une forte consommation de poulet et une disponibilité importante.

Cependant le groupe 1 a un indice de stabilité, la plupart du temps, négatif et le groupe 2 a des revenus assez faibles comparé aux groupes 3 et 4.

De plus le second groupe est uniquement composé d'îles ce qui ne facilite pas la logistique.

Le groupe 3 a une dépendance importante aux importations avec des revenus assez haut.

Le groupe 4 a la population la plus forte avec les meilleurs revenus.

En conclusion de cette première méthode de *clustering*, on retient les pays des groupes 3 et 4

**8** - Réalisation d’une classification K-means

La seconde méthode que nous allons utiliser est la **classification k-means**. Elle va regrouper les individus en cherchant à minimiser l'inertie intraclasse de chaque groupe.

Pour cela, l'algorithme utilise les centres de gravité (centroïdes) des groupes. Au départ, les centroïdes sont placés aléatoirement puis, ils sont recaculés à chaque itération jusqu'à ce que les résultats de l'algorithme convergent.

Pour choisir le nombre optimal de *clusters* nous allons nous appuyer sur la **méthode du coude** dans laquelle nous allons calculer l'inertie intraclasse en fonction du nombre de groupes.

On cherchera alors à déterminer à partir de quand l'augmentation du nombre de *clusters* n'a plus un impact significatif sur la diminution de l'inertie.

Le graphique montre une cassure à 12 *clusters*. Pour confirmer ce résultat, calculons une nouvelle fois les scores de Silhouette.

**9** - Le résultat est identique. On choisit donc de travailler avec **12** clusters.

**10** - Enfin, on affiche les centroïdes des indicateurs, qui correspondent à leurs moyennes en réalisant une graphique heatmap, donc un graphique de couleurs

Cette fois, nous allons retenir les *clusters* 4, 5, 8 et 10 qui présentent une majorité d'indicateurs favorables.Nous allons les visualiser plus en détail dans un autre graphique.

**11** - Les groupes 4 et 8 ont une plus forte consommation de poulet et une plus grande disponibilité. Cependant le groupe 4 a un indice de stabilité souvent négatif et les pays ont des revenus plus faibles que ceux des groupes 5 et 10. De plus, le groupe 8 est uniquement composé d'îles ce qui ne facilite pas la logistique.

Le groupe 10 a une dépendance importante aux importations avec des revenus assez haut. Le groupe 5 a la population la plus forte avec les meilleurs revenus.

En conclusion de cette deuxième méthode de *clustering*, on retient les pays des groupes 5 et 10 que nous allons sauvegarder dans une liste pour la comparer avec le résultat des autres méthodes de classification.

**12** - Réalisation d’une **ACP** avec réductions des dimensions et d’un clustering

Dans la partie précédente, nous n'avons pas pu représenter facilement nos résultats car le nombre de dimensions de notre jeu de données est un peu trop important (nous avons 8 indicateurs). En utilisant l'ACP, nous allons chercher à synthétiser les informations tout en réduisant le nombre de dimensions à analyser.

Les quatre premières valeurs propres correspondent à un pourcentage de variabilité de 76 %. La représentation sur **quatre dimensions** est donc de qualité suffisante, c'est ce nombre que nous choisissons de garder pour la suite.

**13** - Analyse d’un **cercle des corrélations**

On étudie ensuite les liaisons entre nos variables d'origine et les composantes principales en nous aidant du cercle des corrélations grâce à une fonction que j’ai créer qui va nous permettre de tracer ce cercle en fonction des composantes principales que l'on souhaite afficher.

On trace le cercle pour les deux premières dimensions

Les variables les plus corrélées à F1 sont :

* la disponibilité ;
* la part de consommation de volaille ;
* l'indice de stabilité politique ;
* le revenu par habitant.

Les variables les plus corrélées à F2 sont :

* le taux de croissance ;
* l'évolution démographique.

**14**- On trace ensuite le cercle pour les deux autres composantes

Les variables les plus corrélées à F3 sont :

* le taux de dépendance aux importations ;
* la taille de la population ;
* la part de consommation de volaille ;
* la disponibilité.

Les variables les plus corrélées à F4 sont :

* le taux de croissance ;
* la population ;
* le revenu par habitant.

**15** - On peut également afficher les corrélations sous forme de matrice

La matrice confirme les observations faites sur les cercles des corrélations.

**16** - **Projection des individus** en utilisant les données de l’ **ACP** (regroupements des pays)

Maintenant que nous avons créé de nouvelles variables, nous pouvons projeter les pays sur des plans composés par ces dernières.

J’ai créer une autre fonction qui va nous permettre de choisir les dimensions qui vont composer le plan que l'on souhaite représenter.

donc la partie projection des individus.

On va choisir de projecter nos pays sur un plan composé de F1 et de F3. En effet :

* plus un pays aura une valeur élevée sur F1, plus il aura une appétence pour le poulet tout en étant économiquement développé et fiable ;
* plus un pays aura faible valeur sur F3, plus aura besoin d'importer pour satisfaire ses besoins.

Cela devrait donc nous permettre d'identifier des pays intéressants pour notre objectif.

Sur la zone située à droite du plan, et avec une valeur sur F3 assez faible, on retouve des pays que nous pourrions cibler.

Pour confirmer cela, on peut faire la même projection mais en affichant les *clusters* créés grâce à la CAH et aux k-means.

**17** - On projette nos groupes sur les dimensions F1 et F3 pour les deux méthodes de classification

Clusters retenus avec la CAH : [3, 4]

**18** - Clusters retenus avec les k-means : [5, 10]

Dans les deux cas, les *clusters* que nous avions retenus sont situés dans la zone du plan factoriel que nous avions identifié comme propice.

**19** - *Pour rappel, notre entreprise souhaite démarrer son activité à l'internationale*.

Pour affiner notre recommandation, on retire dans un premier temps les pays qui sont loin de la France (logistique) et/ou les pays ayant une population trop importante (capacité à répondre à la demande).

Notre top 5 final porte sur les pays suivants :

1. Allemagne
2. Pays-Bas
3. Belgique
4. Suède
5. Autriche

Le Royaume-Uni a été retiré car suite au Brexit, les échanges commerciaux sont plus compliqués.

L'Italie a également été retirée car elle a l'un des plus faible taux de dépendance aux importations et le revenu par habitant le plus faible.