PRODUISEZ UNE ÉTUDE DE MARCHÉ AVEC PYTHON

Mame Diarra **DABO**

09/09/2023

La poule qui chante

PROBLÉMATIQUE

Data Analyst

La poule qui chante

L'objectif est de «proposer une première analyse des groupements de pays que l'on peut cibler pour exporter nos poulets.»

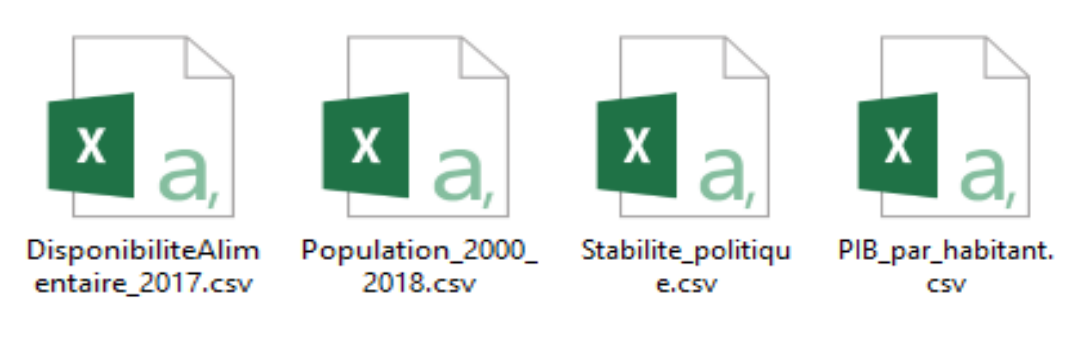
Mission:

- Tester la classification ascendante hiérarchique, avec un dendrogramme comme visualisation ;
- Utiliser la méthode des k-means, afin d'affiner l'analyse ;
- Comparer les résultats des deux méthodes de clustering ;
- Réaliser une ACP afin de visualiser les résultats de mon analyse, comprendre les groupes, les liens entre les variables, les liens entre les individus...

PLAN

- **Importer les données**
- **Nettoyages des données**
- **Répondre aux demandes: Analyses**
- **Recommandation et Conclusion**
- **Pistes d’amélioration**

LE JEU DE DONNÉES



- *Fichier Disponibilité Alimentaire: (176600, 14)*
- *Fichier Population : (4411, 15)*
- *Fichier Stabilité Politique : (196, 15)*
- *Fichier PIB : 186, 15)*

NETTOYAGES DES DONNÉES

Afficher chaque jeu de données

- **Nombre de lignes et colonnes :**

  => les types de données
  => Vérifier les doublons
  => Vérifier les valeurs manquantes
  => Sélectionner les colonnes pertinentes

- **Gestion des valeurs**

  => Renommer des colonnes
  => Convertir au bon format de données

NETTOYAGES DES DONNÉES: Jointure

Un fichier en fonction de la disponibilité de poulet/zone ：172 zones
14 colonnes

- **Résultat 1**

Un fichier en fonction de la stabilité/zone：172 zones
15 colonnes

- **Résultat 2**

NETTOYAGES DES DONNÉES: Après la jointure

- **Résultat 3** → Un fichier en fonction du PIB/zone : 172 zones
  16 colonnes

Gestion des valeurs

Features engineering

- **manquantes :**

  => la valeur médiane par variables numériques

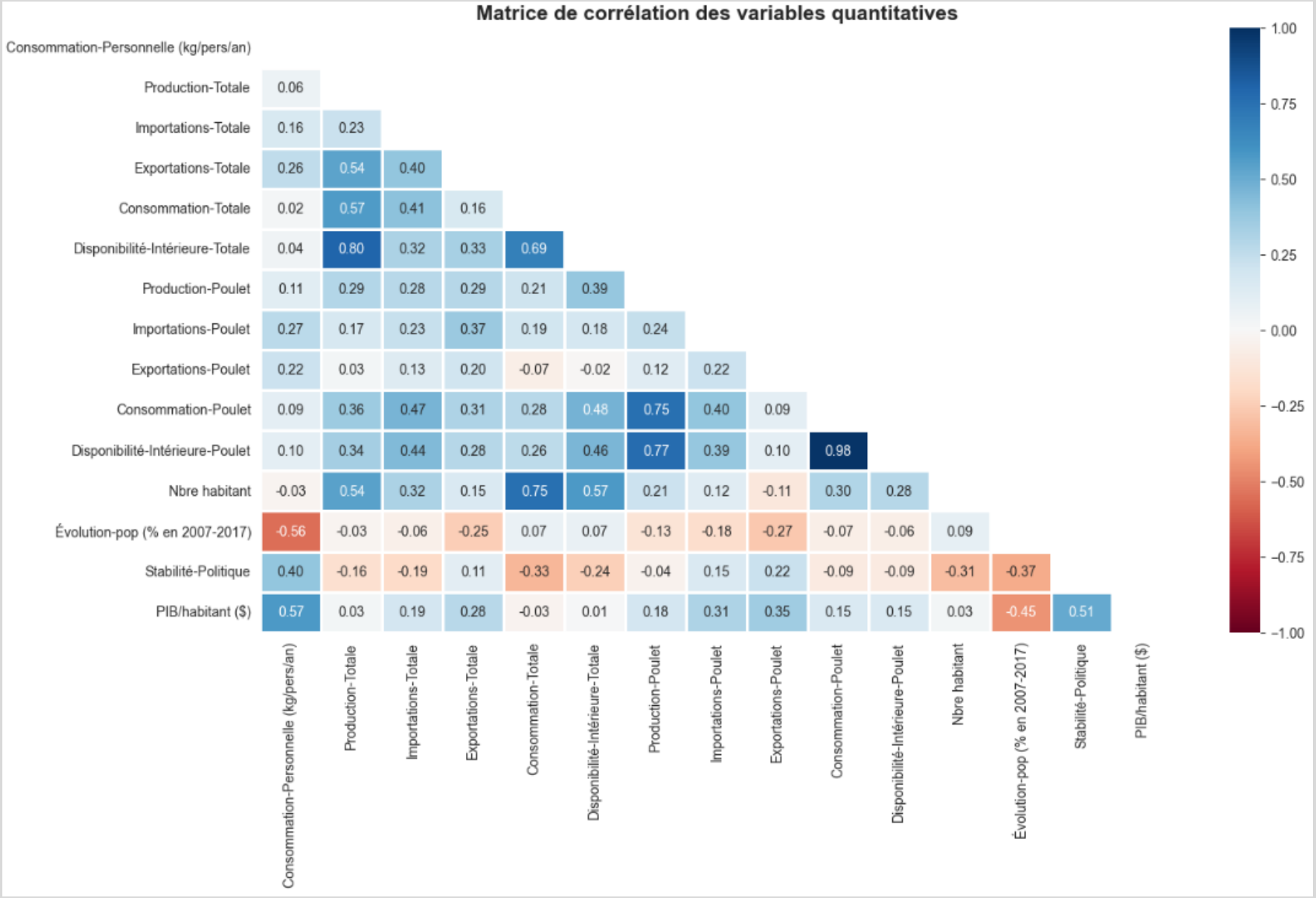
- **aberrantes :**

  => Traiter les valeurs aberrantes

Fichier final : 172 lignes
16 colonnes

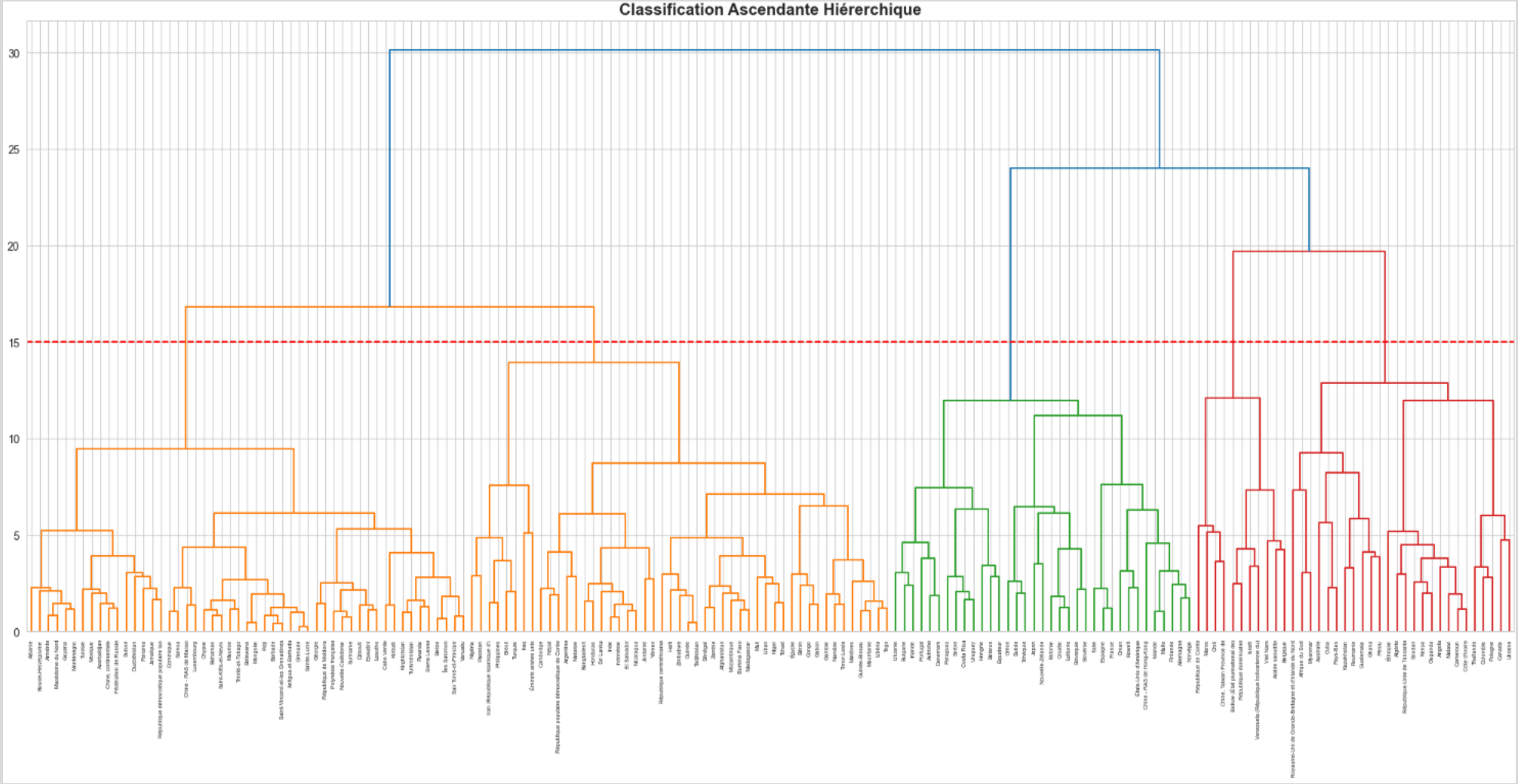
ANALYSE: Les corrélations linéaires

On remarque que la disponibilité intérieure de poulet est très fortement corrélé à la consommation de poulet, idem pour la disponibilité intérieure totale à la production totale. Cela est dû au fait de la forte consommation de poulet.



ANALYSE: Classification Hiérarchique Ascendante

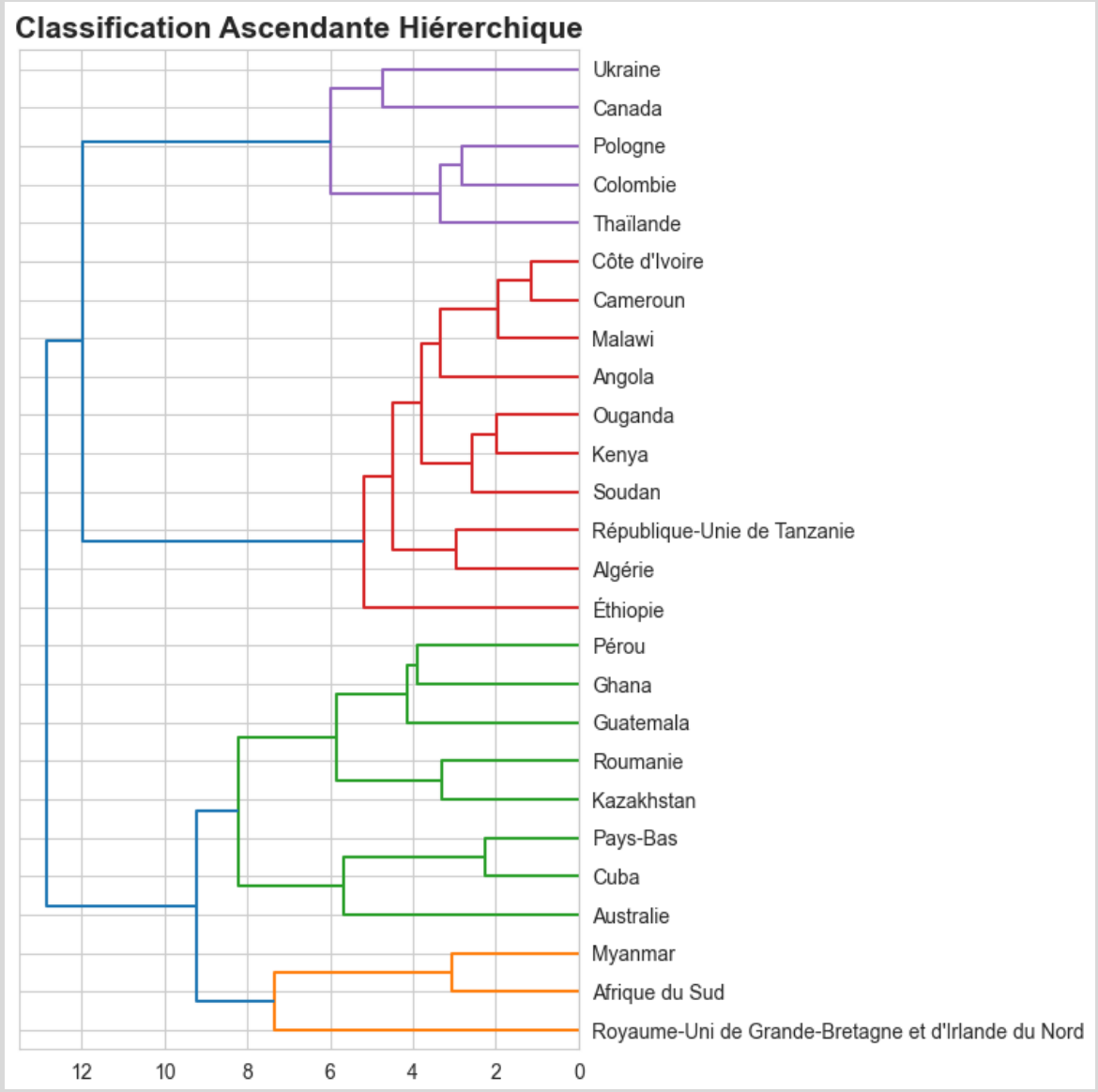
Visualisation du Dendrogramme



Le Dendrogramme est donc le type de diagramme en arborescence que l'on utilise pour présenter le clustering hiérarchique, à savoir les relations entre des ensembles de données similaires.

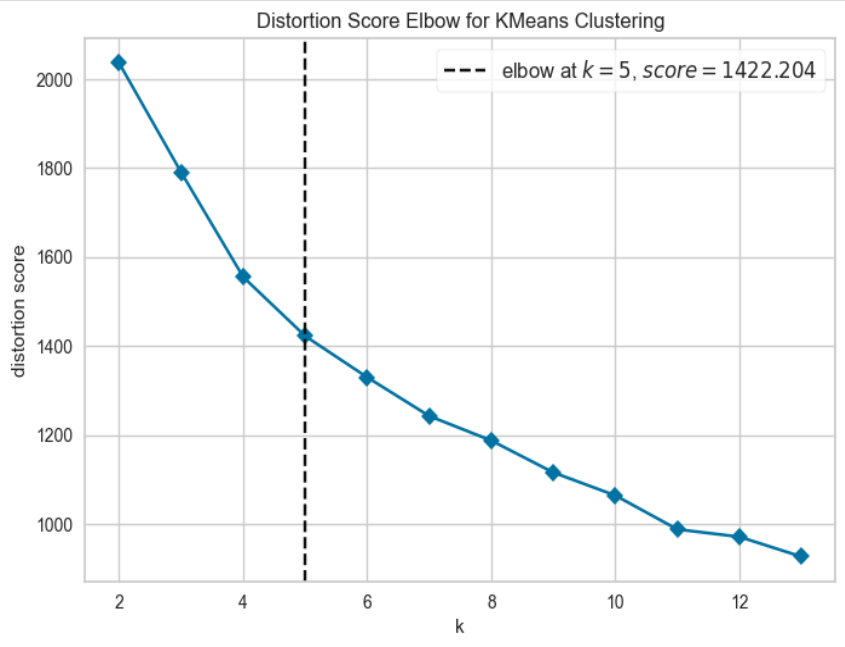
ANALYSE: Classification Hiérarchique Ascendante

Représentation d'une des zones des différentes clusters (cluster 1)



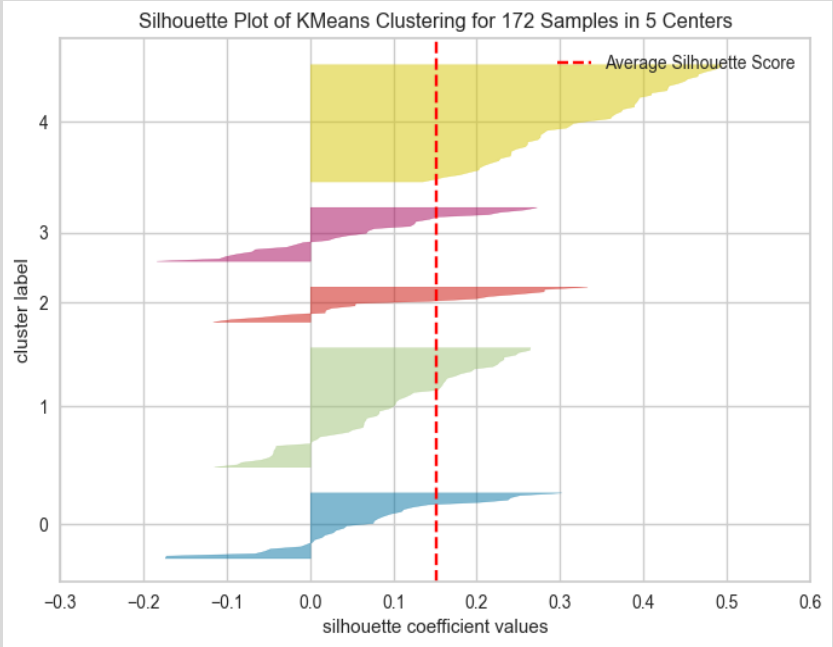
ANALYSE: K-means Clusters

Distorsion score Elbow



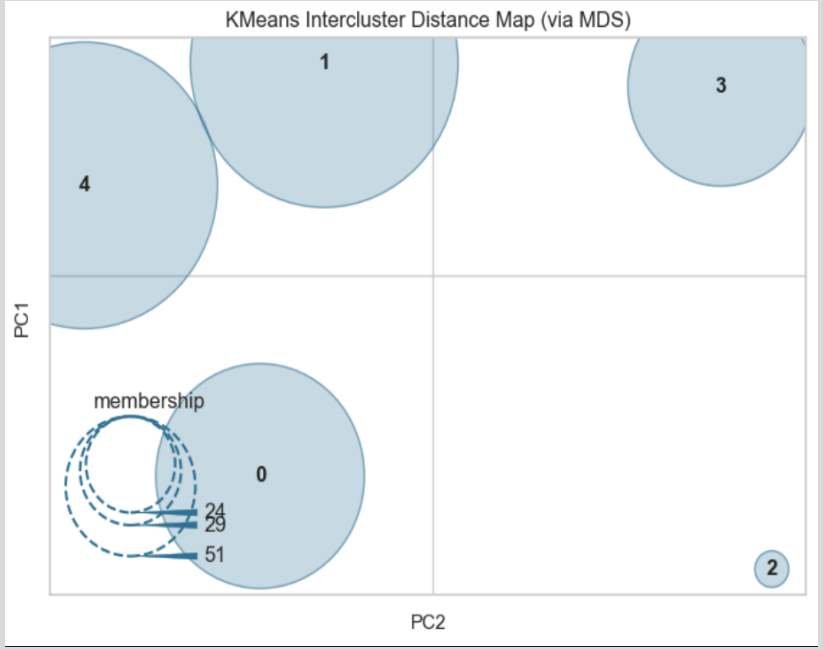
Grâce à la méthode du coude basée sur le score de distorsion *(somme moyenne des carrés des distances aux centres)*, une segmentation en K=5 clusters serait la meilleure option,

Silhouette



Silhouette : *rapport moyen entre la distance intra-cluster et la distance du cluster le plus proche*,

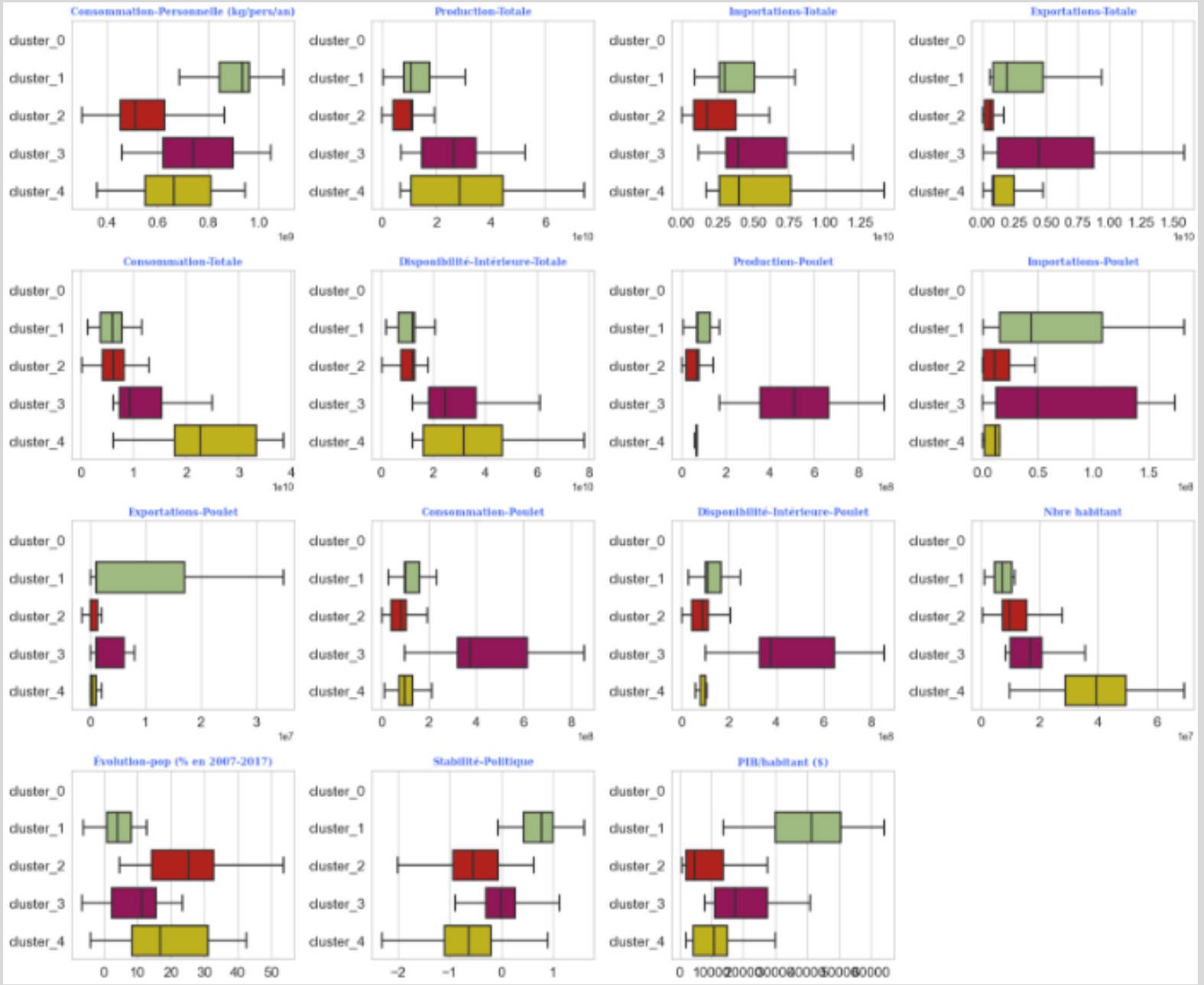
Distance inter cluster



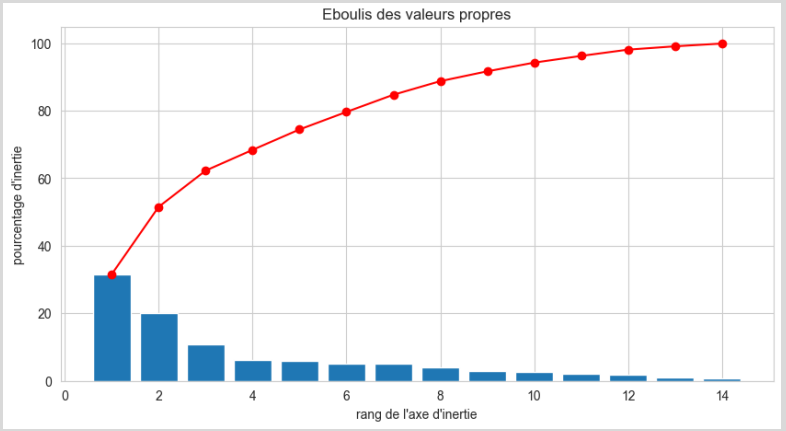
Cette projection en 2D, on remarque que deux clusters ne sont pas bien séparés sur les 2 premières composantes principales. Le clustering semble donc performant et il faut a présent **identifier les composantes métier de chaque cluster**.

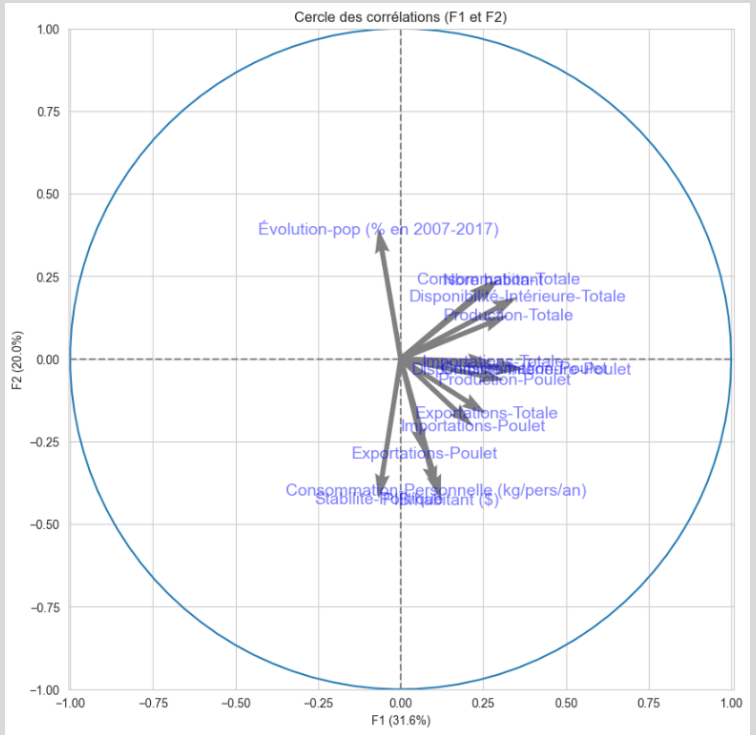
ANALYSE: K-means Clusters (vue globale)

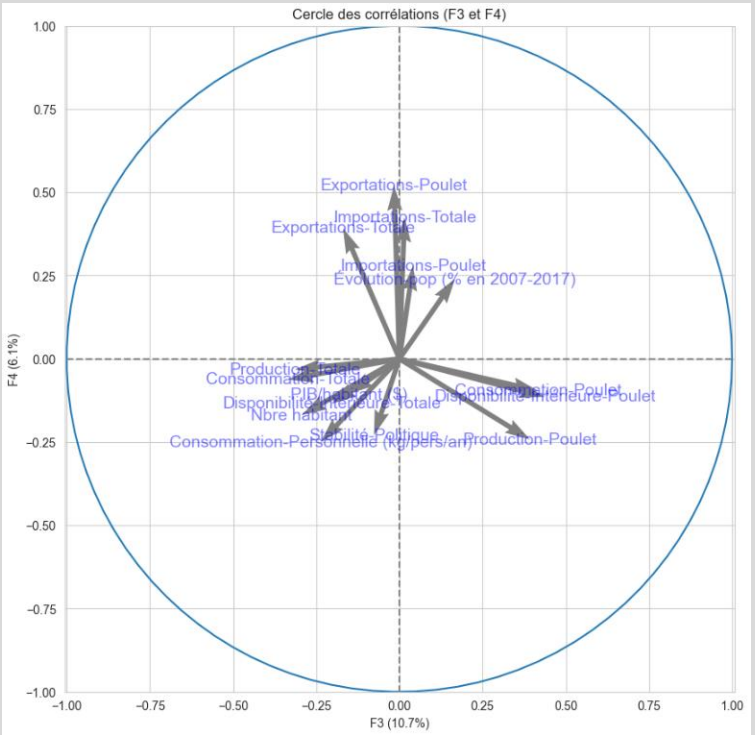
Boxplot de la contribution de chaque feature (variable) par clusters

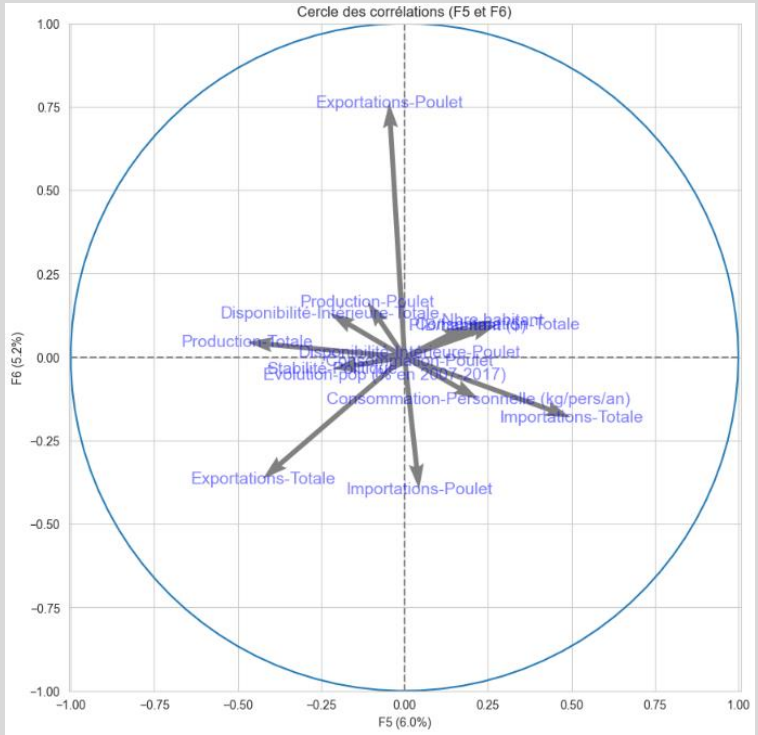


ANALYSE: Réduction de dimensions (PCA)

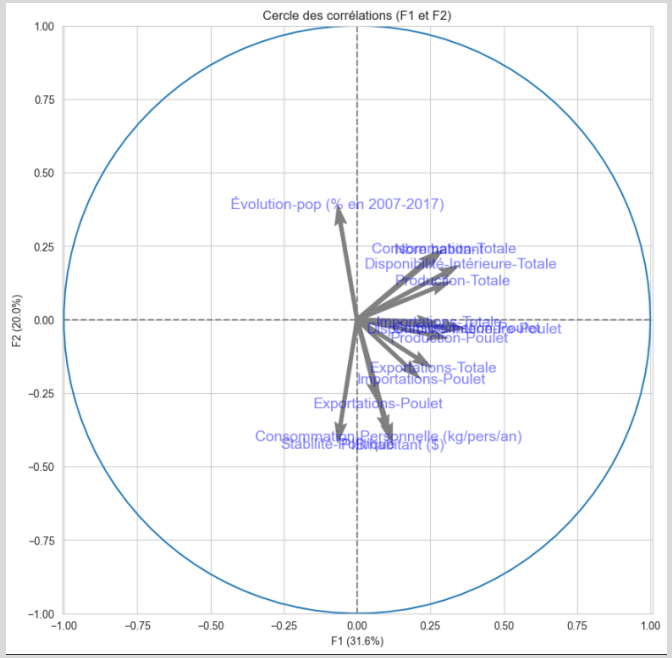


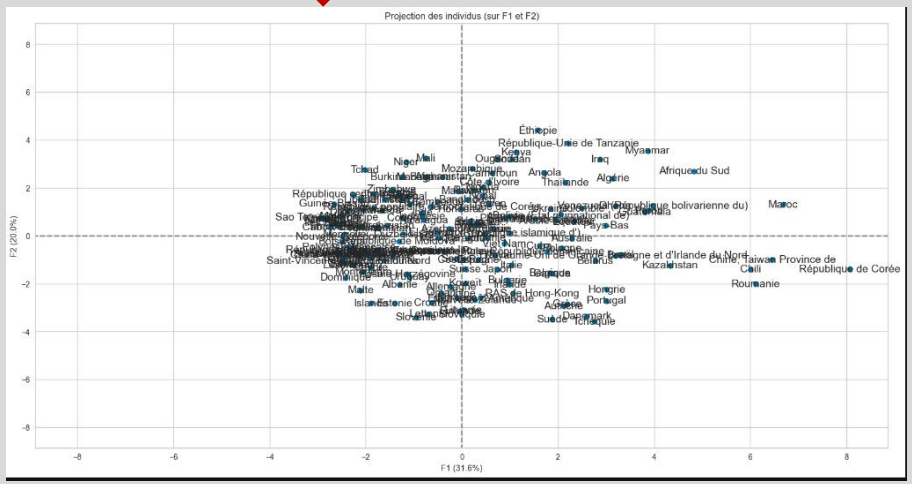


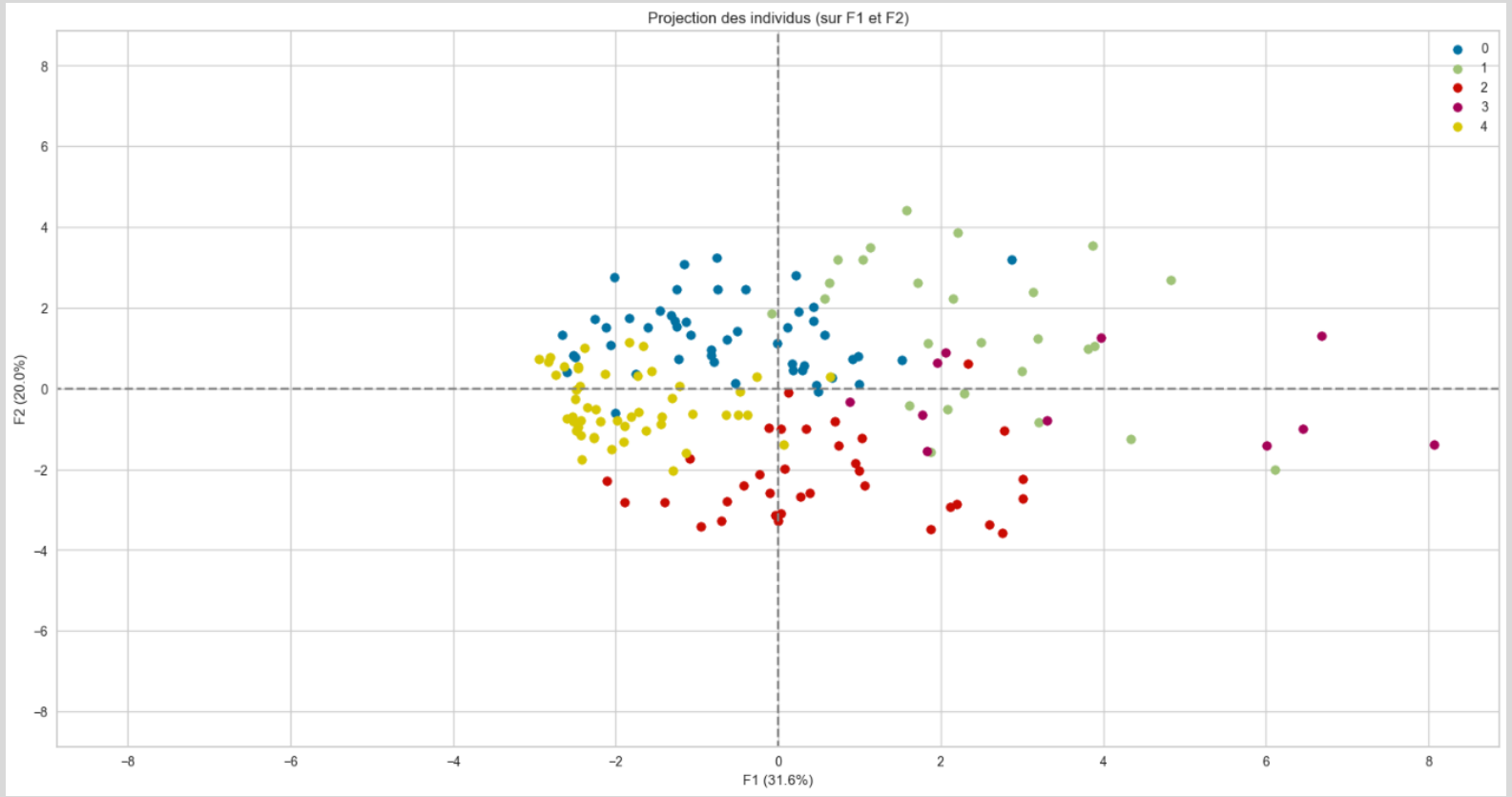




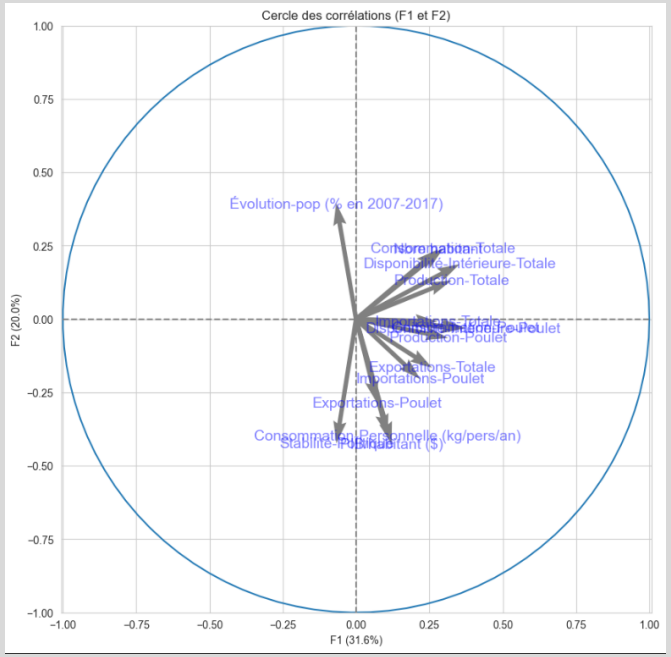
ANALYSE: Projection des individus avec la CAH

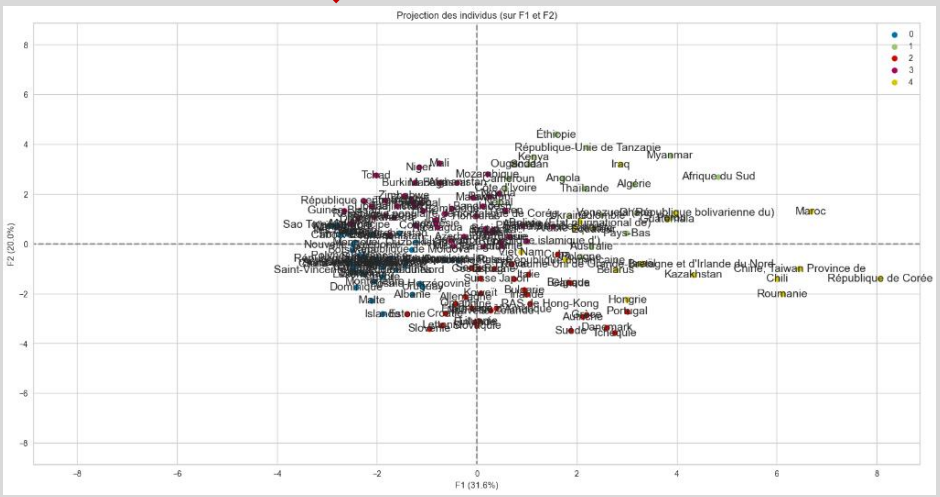


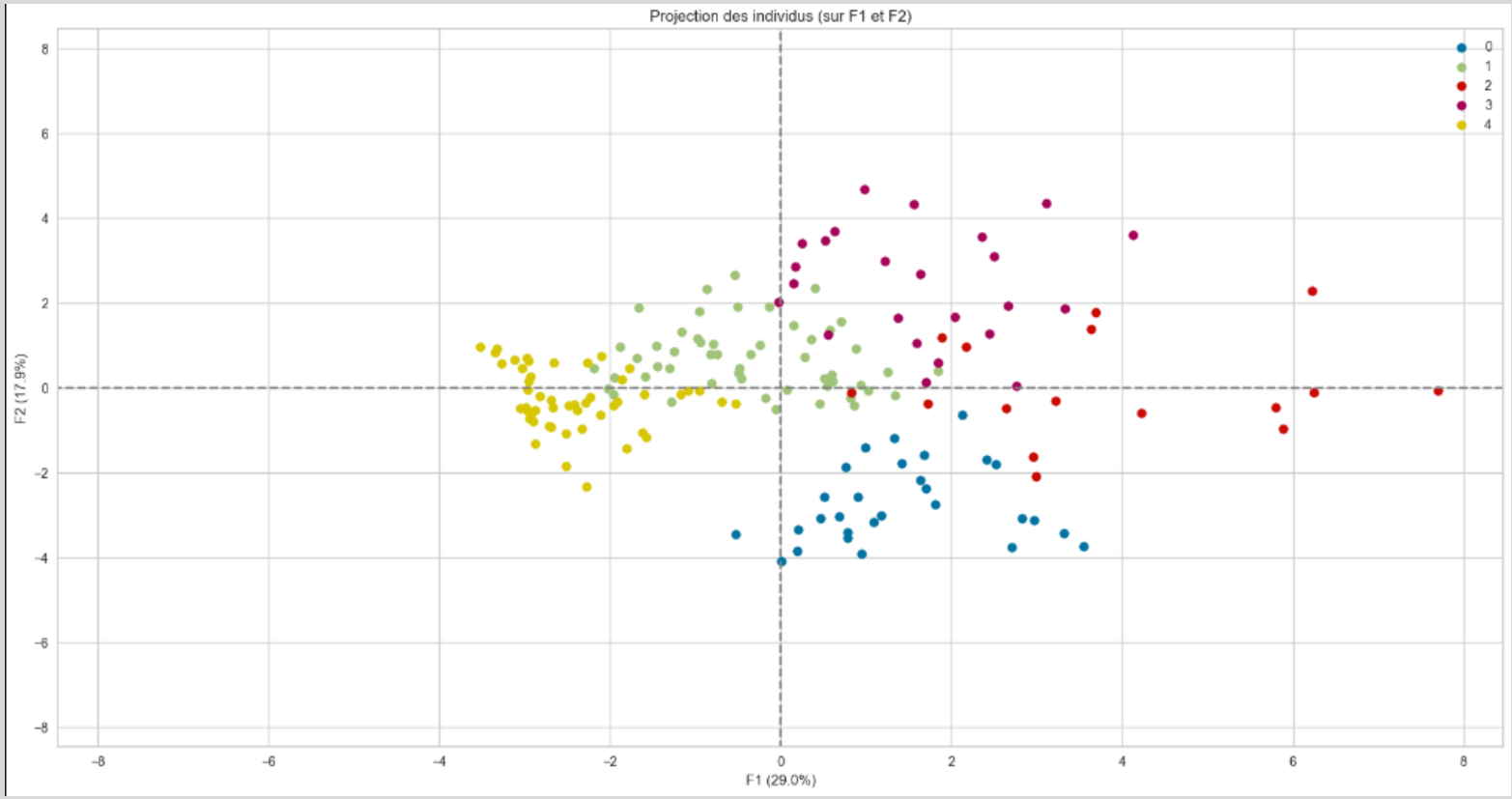




ANALYSE: Projection des individus avec le K-means

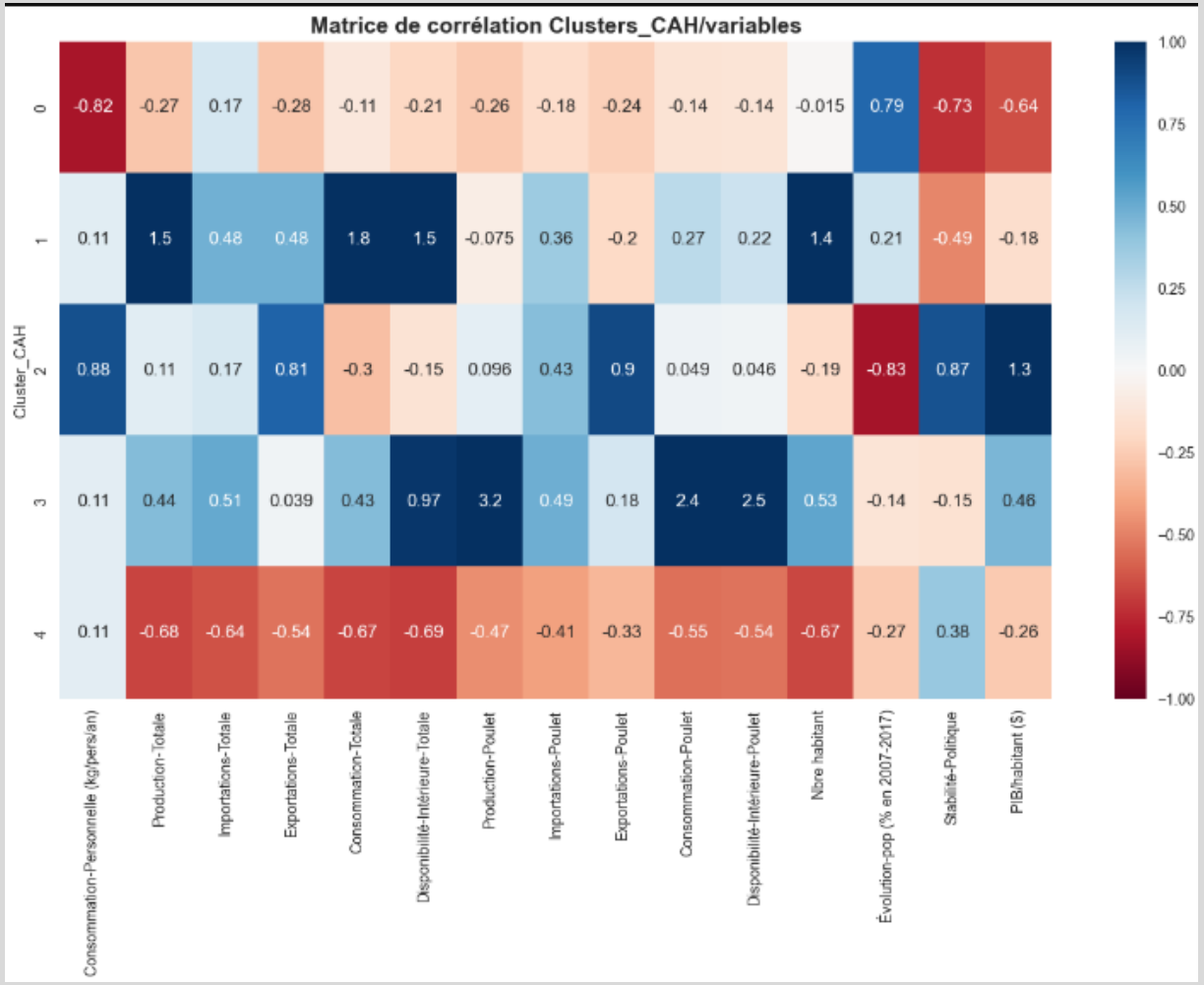




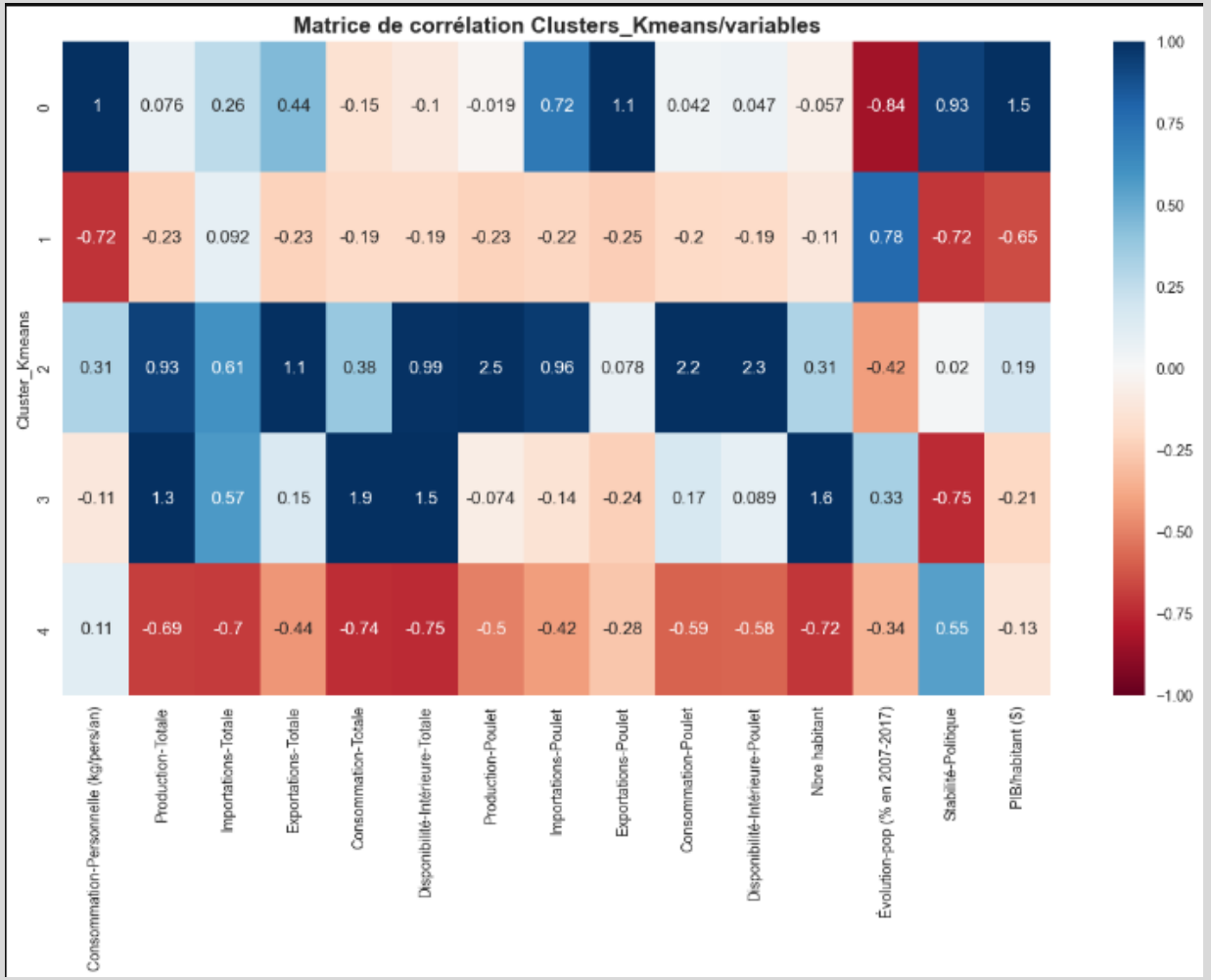


ANALYSE: Matrice des corrélations avec les clusters

Clusters CAH

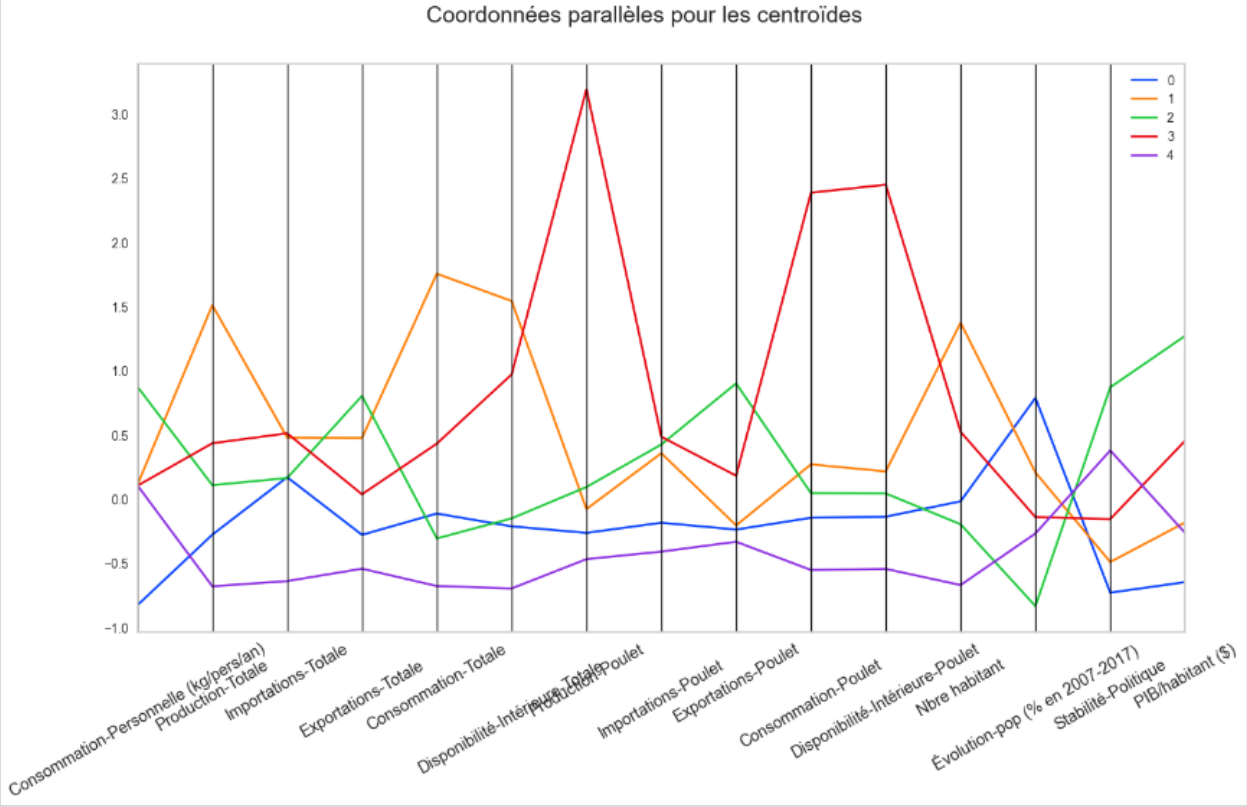


Clusters K-means

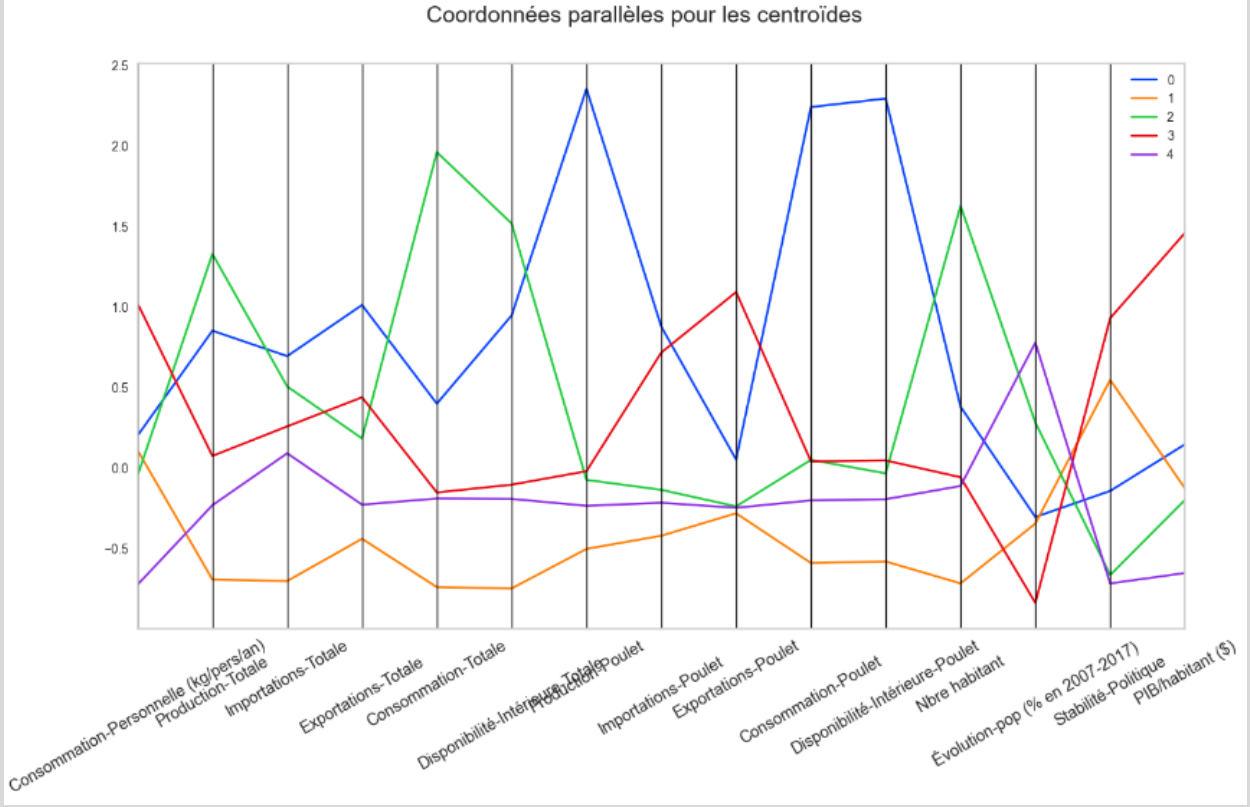


ANALYSE: Coordonnées parallèles pour les centroïdes"

Coordonnées avec la CAH



Coordonnées avec le K-means



ANALYSE: Comparaison des deux méthodes

Clusters avec la CAH

```
Nbre de clusters avec la CAH:
 4    51
0     49
2     35
1     26
3     11
```

Clusters avec le K-means

```
Nbre de clusters avec le K-means:
 1    52
4     51
0     29
3     24
2     16
```

Les clusters ont des tailles différentes avec un faible écart dans les deux méthodes .

Recommandations et Conclusion

Pays avec la CAH

```
Cluster_CAH 2
['Allemagne', 'Autriche', 'Bulgarie', 'Bélarus', 'Chine - RAS de Hong-Kong', 'Costa Rica', 'Croatie', 'Danemark', 'Espagne', 'Estonie', 'Finlande', 'France', 'Grèce', 'Hongrie', 'Irland
e', 'Islande', 'Italie', 'Japon', 'Koweït', 'Lettonie', 'Lituanie', 'Malte', 'Norvège', 'Nouvelle-Zélande', 'Oman', 'Paraguay', 'Portugal', 'Serbie', 'Slovaquie', 'Slovénie', 'Suède', 'T
chéquie', 'Uruguay', 'Équateur', "États-Unis d'Amérique"]
```

Pays avec le K-means

```
Cluster_Kmeans 0
['Allemagne', 'Autriche', 'Belgique', 'Bulgarie', 'Canada', 'Chine - RAS de Hong-Kong', 'Croatie', 'Cuba', 'Danemark', 'Espagne', 'Estonie', 'Finlande', 'France', 'Grèce', 'Irlande', 'It
alie', 'Japon', 'Koweït', 'Lettonie', 'Lituanie', 'Norvège', 'Nouvelle-Zélande', 'Oman', 'Slovaquie', 'Slovénie', 'Suisse', 'Suède', 'Tchéquie', "États-Unis d'Amérique"]
```

- Ces pays-là sont recommandés, ce sont des pays avec une politique assez stable, un PIB par habitant relativement élevé. Cela signifie qu'ils sont plus développés, plus riche, et que donc il sera plus facile de travailler avec ces pays. Avec une production de poulet moyenne et un niveau d'importation important, ils permettent d'éviter une concurrence trop rude qui pourrait rendre difficile l'implantation.

AUTRES

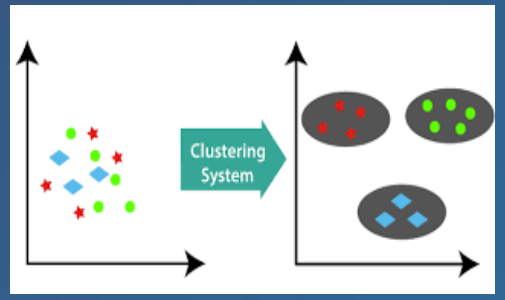
Piste à améliorer

- Tester plus de paramètres
- Plus de modèle
- Faire plus de visualisations

BILAN

- Données

- Démarche



- Réponses