Exportation de nos volailles à l'international

Démarche et premiers pays selectionnés

Contexte & Objectif

Démarche

Souhait de se développer à l'international

Sélectionner des pays à travers une analyse, puis réaliser une étude de marché

Importer et réaliser un retraitement des données

Analyse exploratoire des données, modifications et nettoyage éventuels

Importer des données supplémentaires pour affiner l'analyse

Choix des indicateurs pertinents

Appliquer une méthode de regroupement des lignes

Appliquer une méthode de réduction des dimensions

Clustering sur la réduction de données

Matrice des corrélations

- Outil permettant de montrer la force et la direction entre une ou plusieurs variables.

- Le coefficient de corrélation varie de -1 à +1 :

  -1 pour une corrélation négative parfaite,

  +1 pour une corrélation positive parfaite

  pas de corrélation entre les variables.

- Dans notre analyse, elle nous aide au choix des indicateurs à conserver pour effectuer un clustering plus cohérent.

Le clustering

diviser des données en sous ensemble homogènes et partageant des caractéristiques communes

- **Classification Ascendante Hiérarchique (clustering agglomératif) :** chaque individus est un cluster puis on les agglomère deux à deux pour ne former qu'un seul cluster. On utilise le plus souvent la méthode de Ward (calcul des distances entre les individus). Cette méthode ne nécessite pas de déterminer un nombre de clusters au préalable.

- **K-Means :** Nécessite de déterminer un nombre de clusters au préalable puisqu'il regroupe les points en k clusters. L'élément central est le centroïde (point que l'on choisit comme étant le centre d'un cluster). C'est en fonction du centroïde que nous définissons l'appartenance d'un individu à un cluster.

  - **Comment définir le nombre de clusters optimal ?**

    Méthode du coude : on cherche une cassure dans la courbe liant la variance intraclasse au nombre de clusters

    Silhouette score : Différence entre la distance moyenne entre les points du même groupe et entre les points des autres groupes voisins. Ce coefficient est la moyenne du coef de tous les points. Il varie entre –1 (pire classification) et 1 (meilleure classification)

  - **Comment savoir si le nombre de clusters est adéquat ?**

    On contrôle l'homogénéité des clusters ainsi que le rapport entre la variance intra et inter groupes (Calinski-Harabasz score)

Analyse en Composantes Principales (ACP ou PCA)

- Deux objectifs principaux :

  variabilité entre les individus : différences et ressemblances

  liaison entre les variables : y-a-t-il des groupes de variables très corrélées qui peuvent être regroupées en variables synthétiques ?

- Ces nouvelles variables sont appelées **composantes**. Cela permet de résumer l'information en réduisant le nombre de variables.

- Le **cercle des corrélation** nous permet de visualiser l'importance de chaque variable pour chaque composante.

  La direction de la flèche indique l'axe expliqué par la variable et le sens indique si la corrélation est positive ou négative. Plus la flèche touche le cercle plus cette variable est bien représentée. Les variables qui sont proches sur le cercle sont corrélées entre elles.

- On représente ensuite nos données sur les mêmes axes afin de pouvoir les analyser.

Analyse des pays cibles pour l'exportation

- Plusieurs essais ont été nécessaires pour sélectionner au mieux les pays à cibler :

1er essai : indicateurs de départ, optimiser les données en rechargeant les données puis sélection des variables corrélées et pertinentes à l'analyse. Premier clustering (CAH)

2ème essai : ajout d'indicateurs supplémentaires puis sélection des variables corrélées. Deuxième CAH

3ème essai : Réduire les dimensions avec l'ACP puis effectuer à nouveau une CAH

4ème essai : Effectuer une autre méthode de clustering (KMeans) pour comparer aux résultats précédents

1er essai

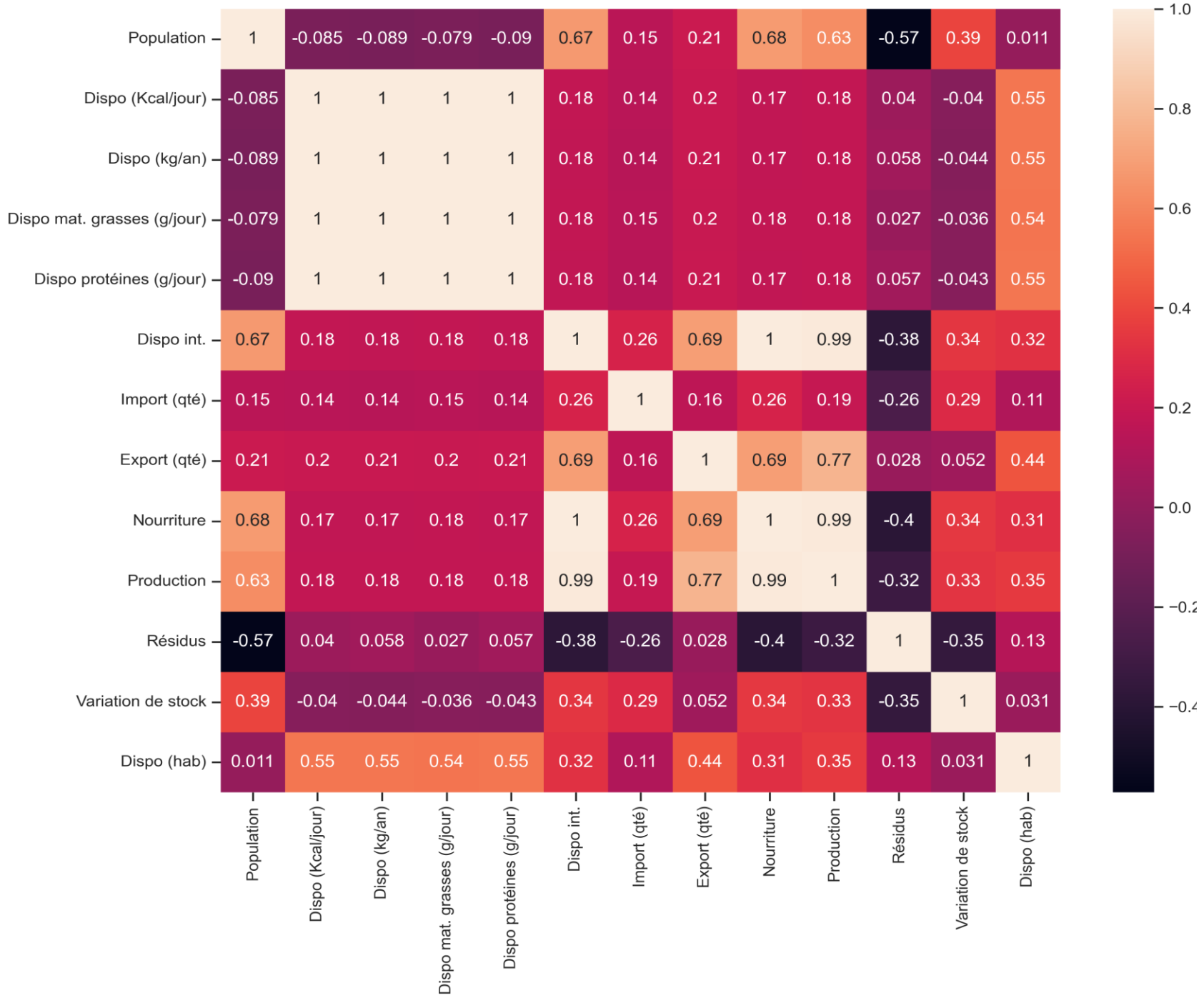
- Variables sélectionnées :

  `Dispo (kg/an)`

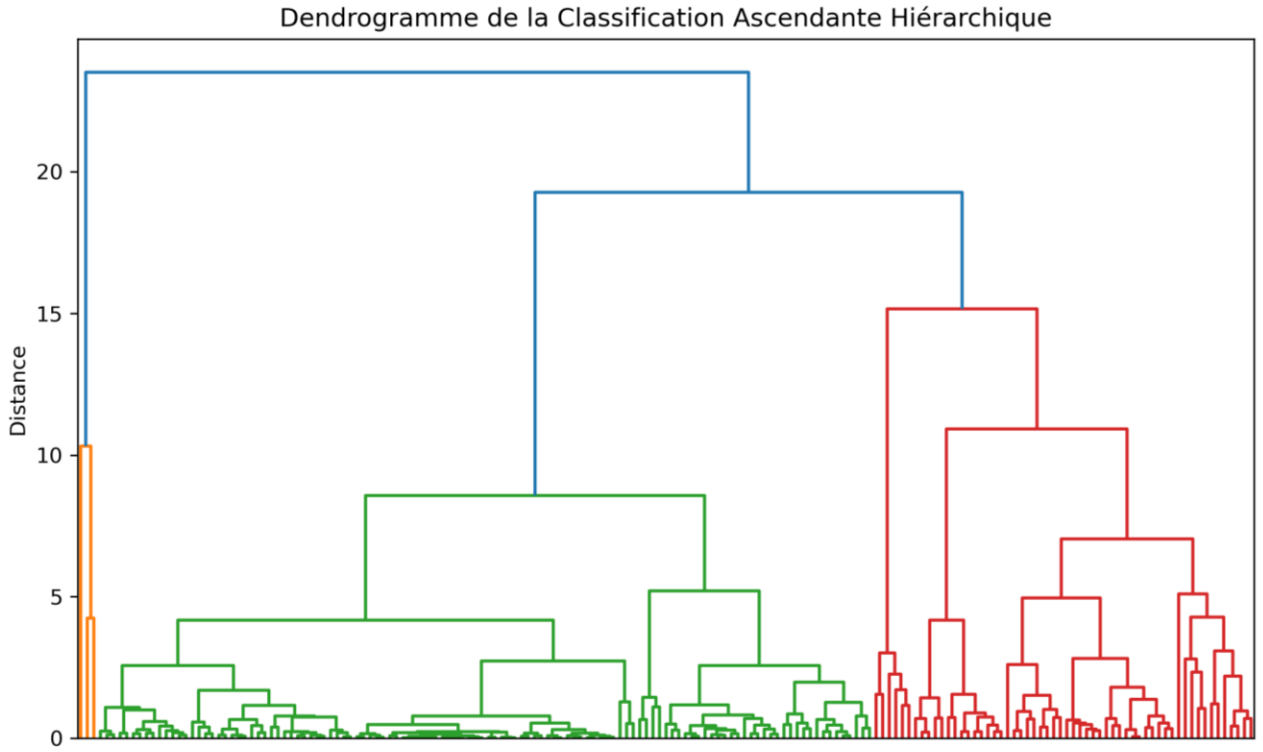
  `Dispo int.`

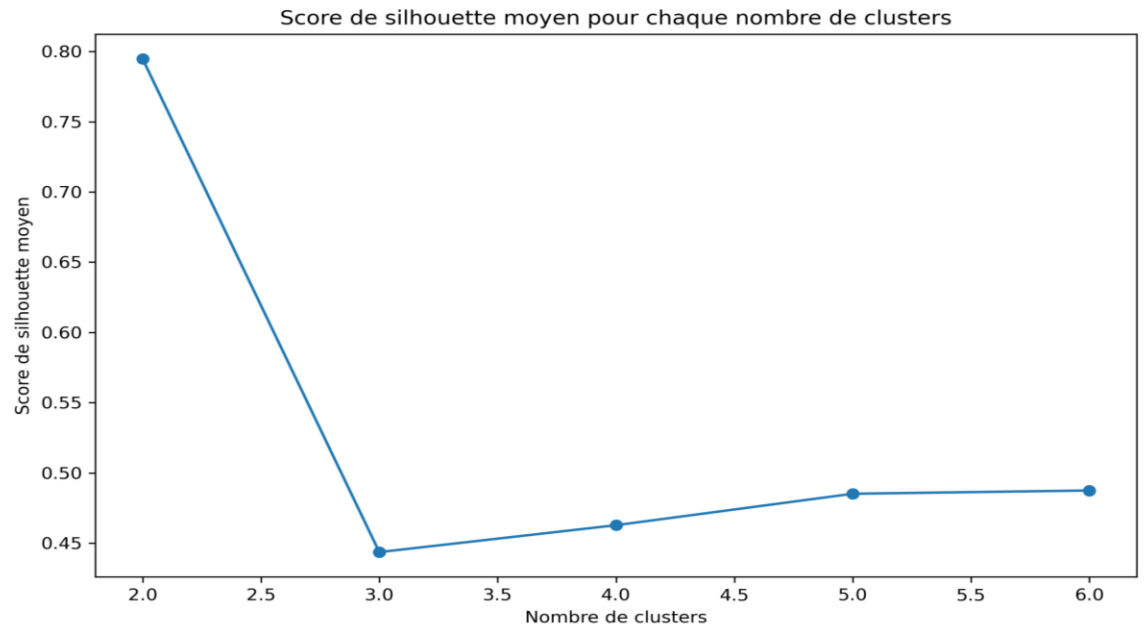
  `Export (qté)`

  `Import (qté)`



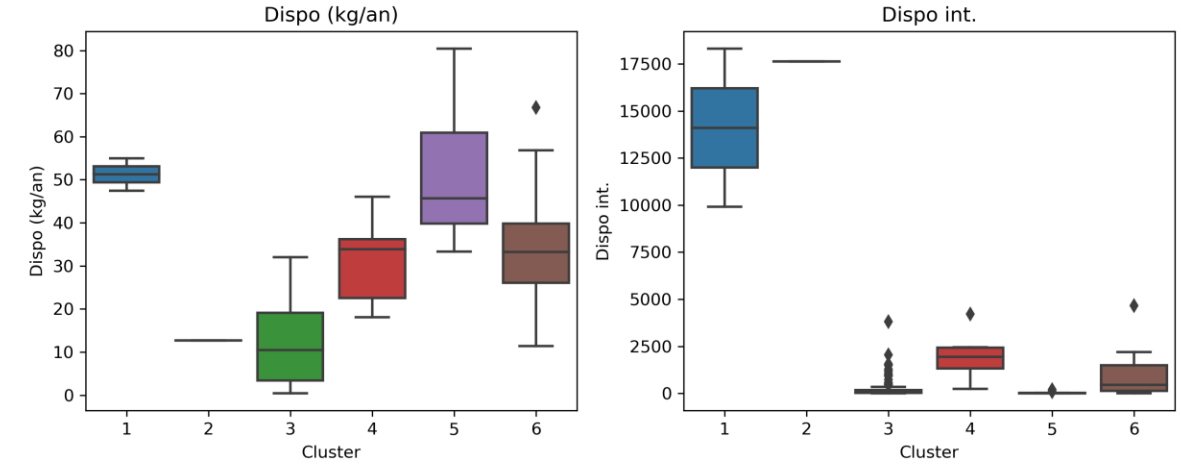
CAH

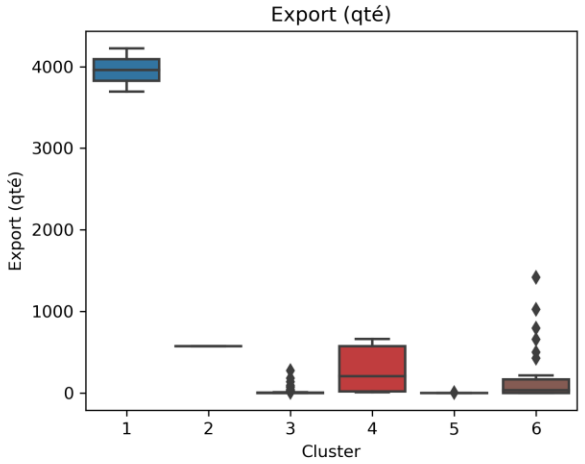


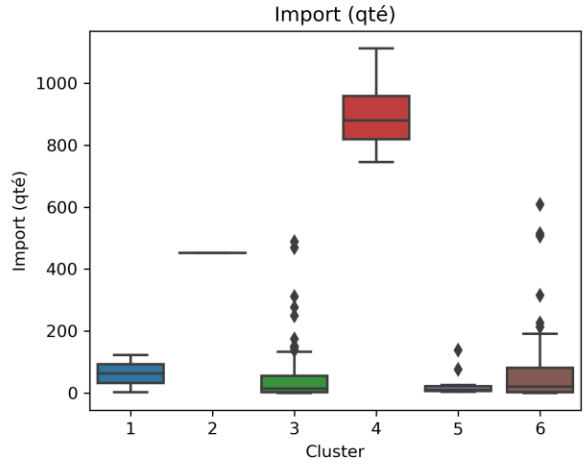


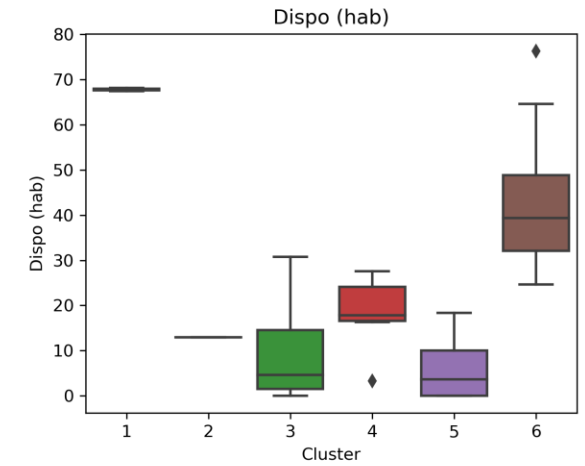

```
For n_clusters = 2, silhouette score is 0.795)
For n_clusters = 3, silhouette score is 0.446)
For n_clusters = 4, silhouette score is 0.449)
For n_clusters = 5, silhouette score is 0.472)
For n_clusters = 6, silhouette score is 0.501)
```

Résultats 1er essai









Homogeneity score: 0.828
Calinski Harabasz Score : 116.382

2ème essai

Variables sélectionnées :

Dispo (kg/an)

Dispo int.

Export (qté)

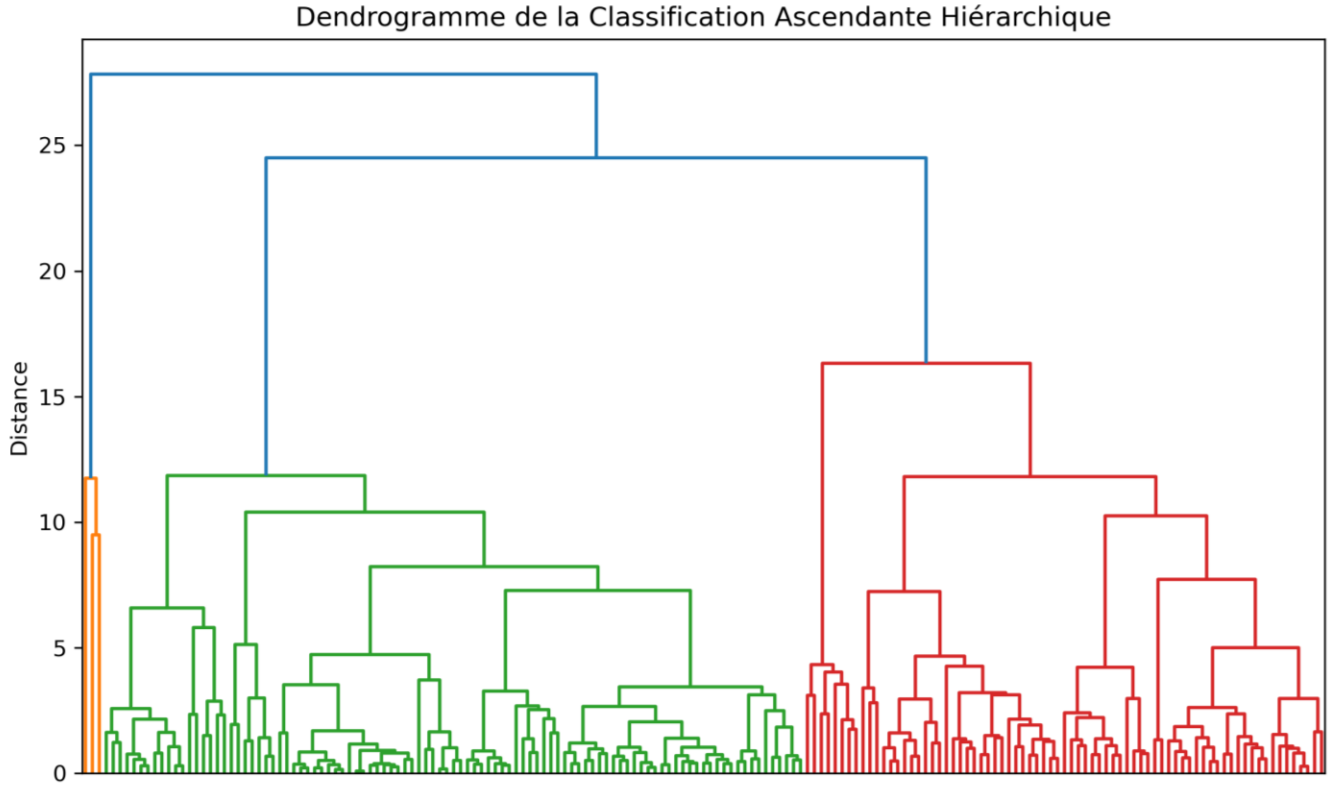
PIB ($)

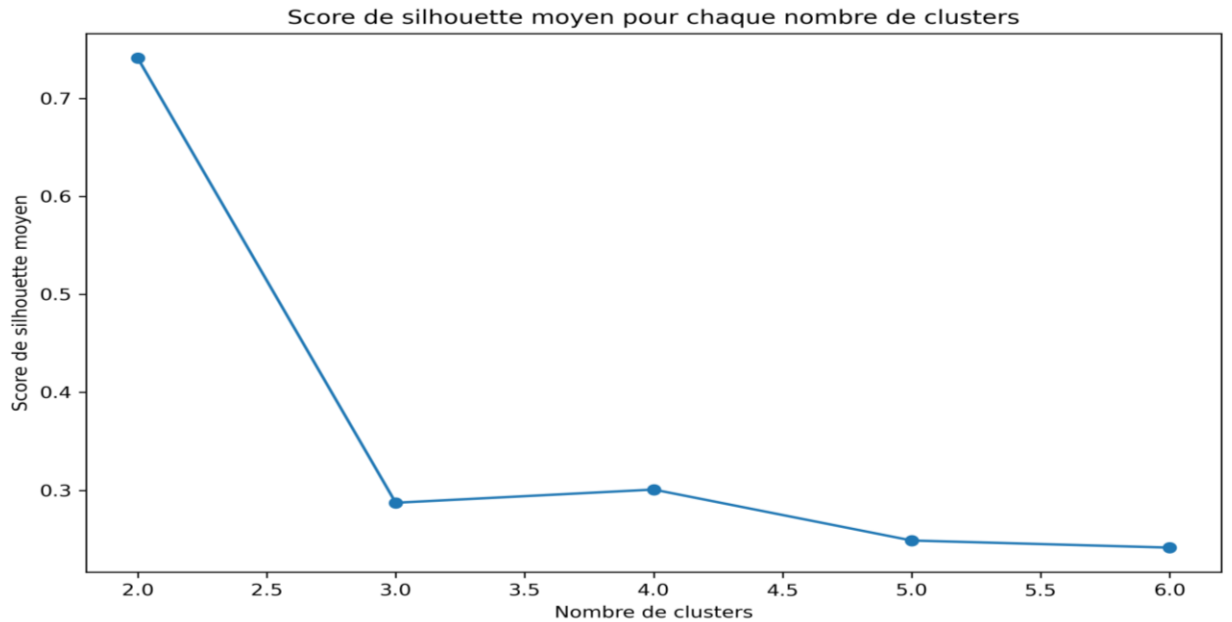
PIB PPA hab($)

Stabilité politique (indice)

PIB - Croissance (%)

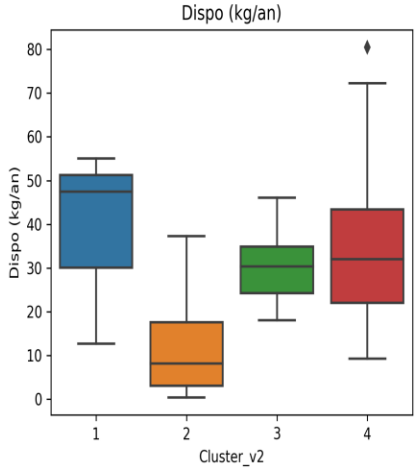
CAH

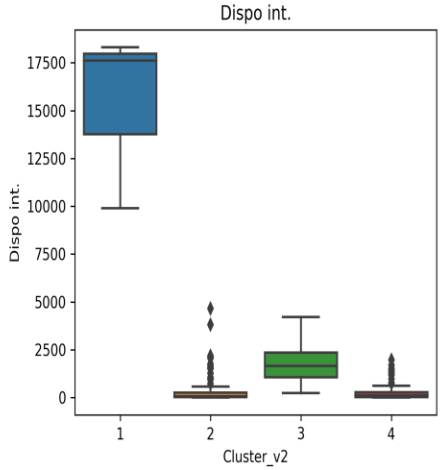


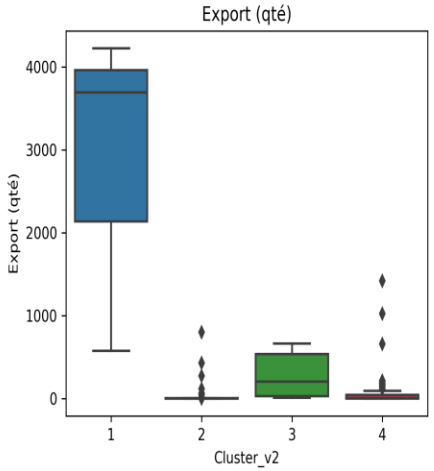


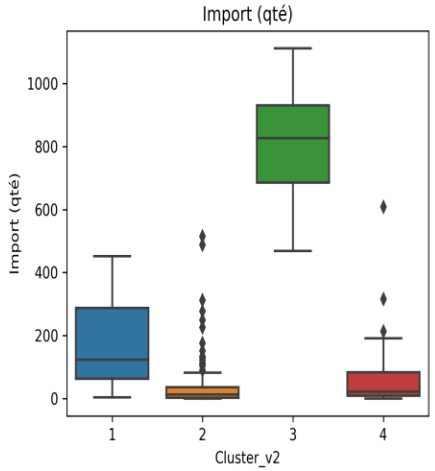

```
For n_clusters = 2, silhouette score is 0.741)
For n_clusters = 3, silhouette score is 0.295)
For n_clusters = 4, silhouette score is 0.308)
For n_clusters = 5, silhouette score is 0.28)
For n_clusters = 6, silhouette score is 0.261)
```

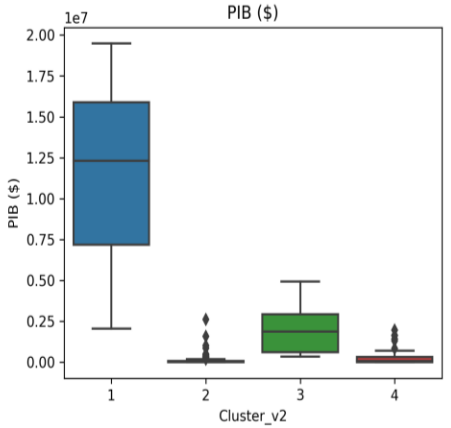
Résultats 2ème essai

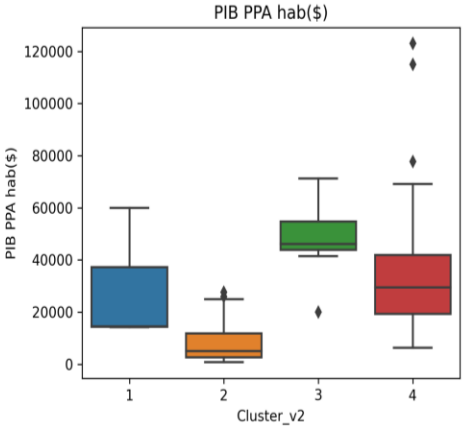


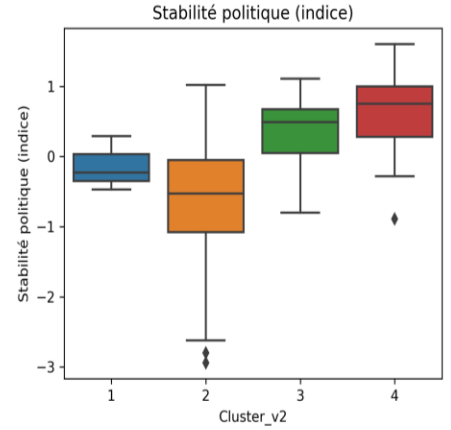


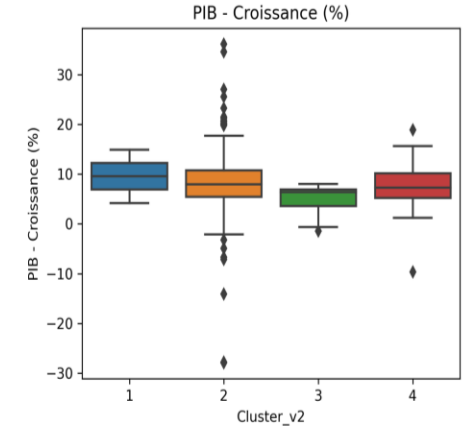


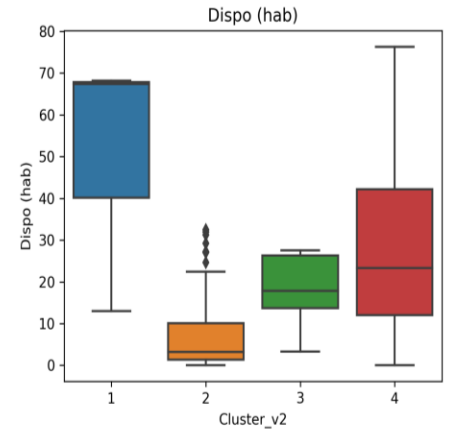






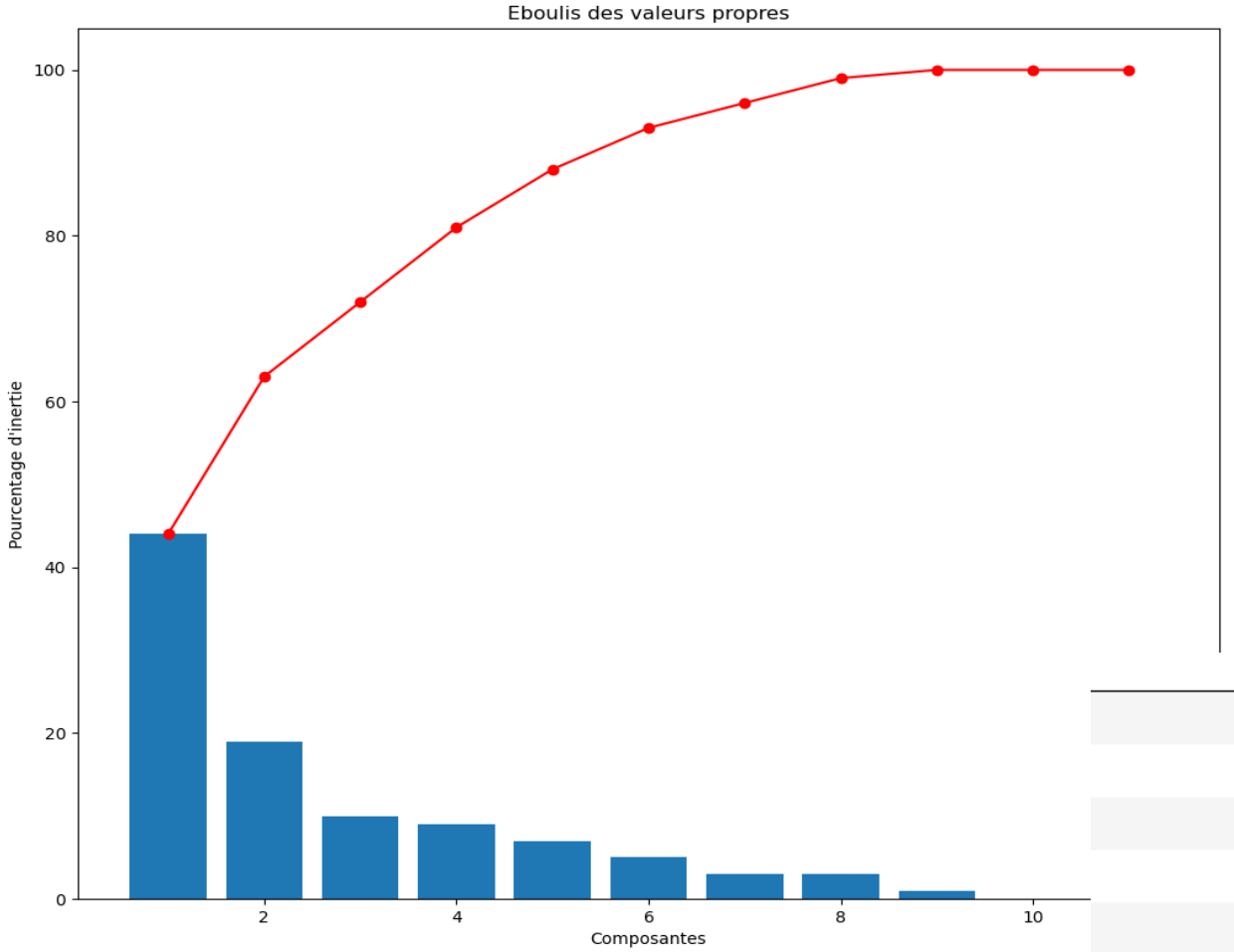




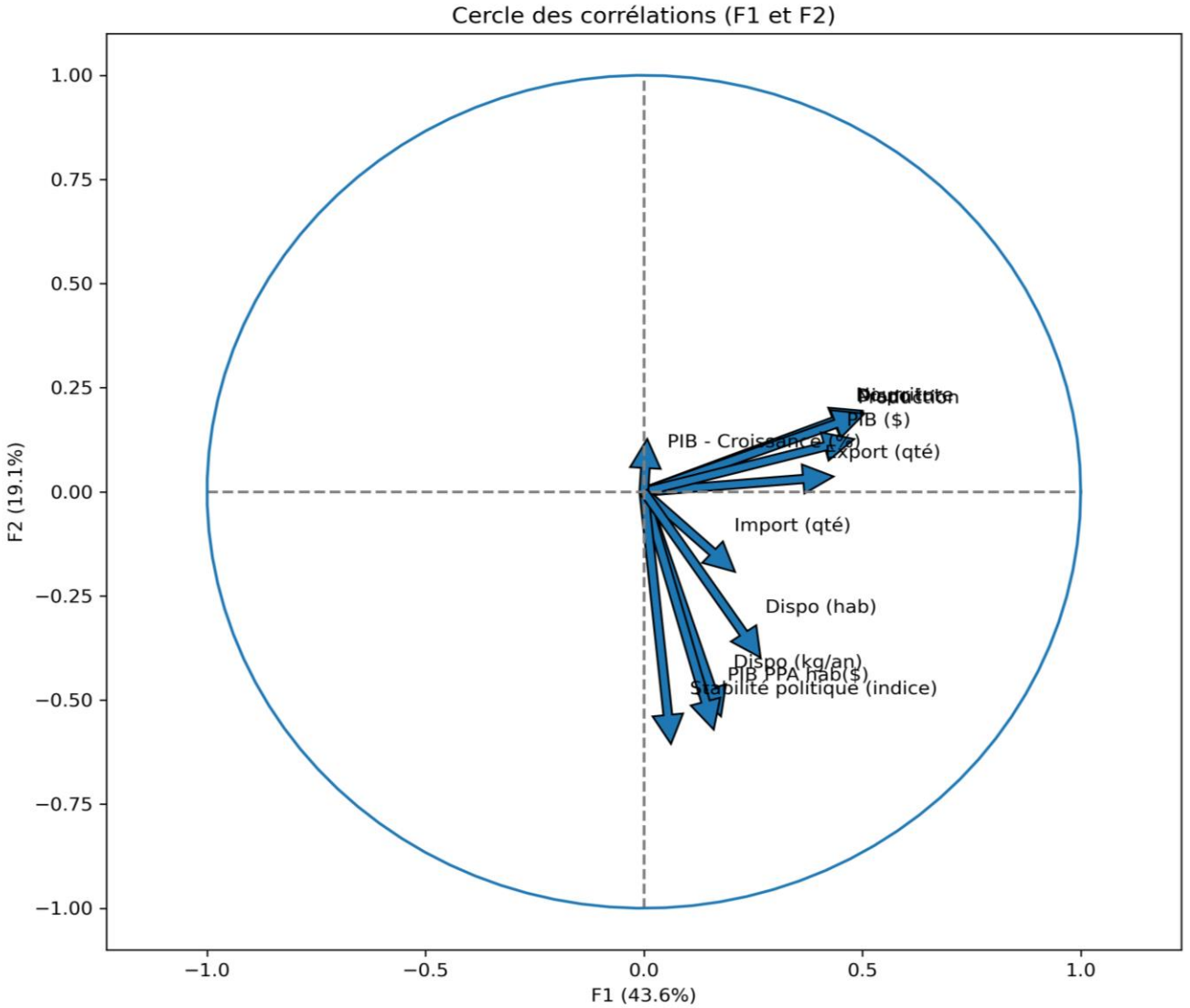


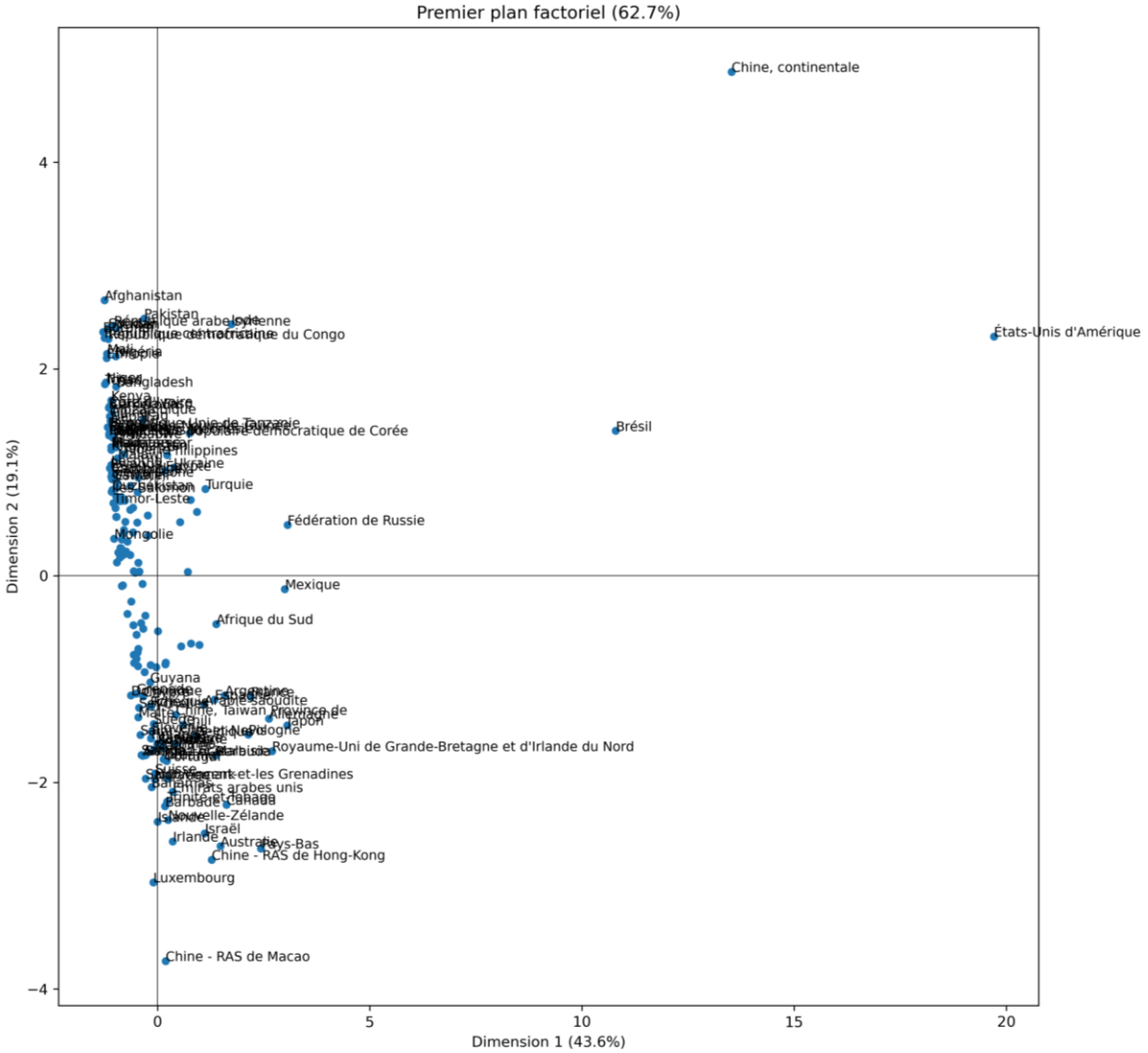
Homogeneity score: 0.849
Calinski Harabasz Score : 60.608

3ème essai: ACP puis CAH



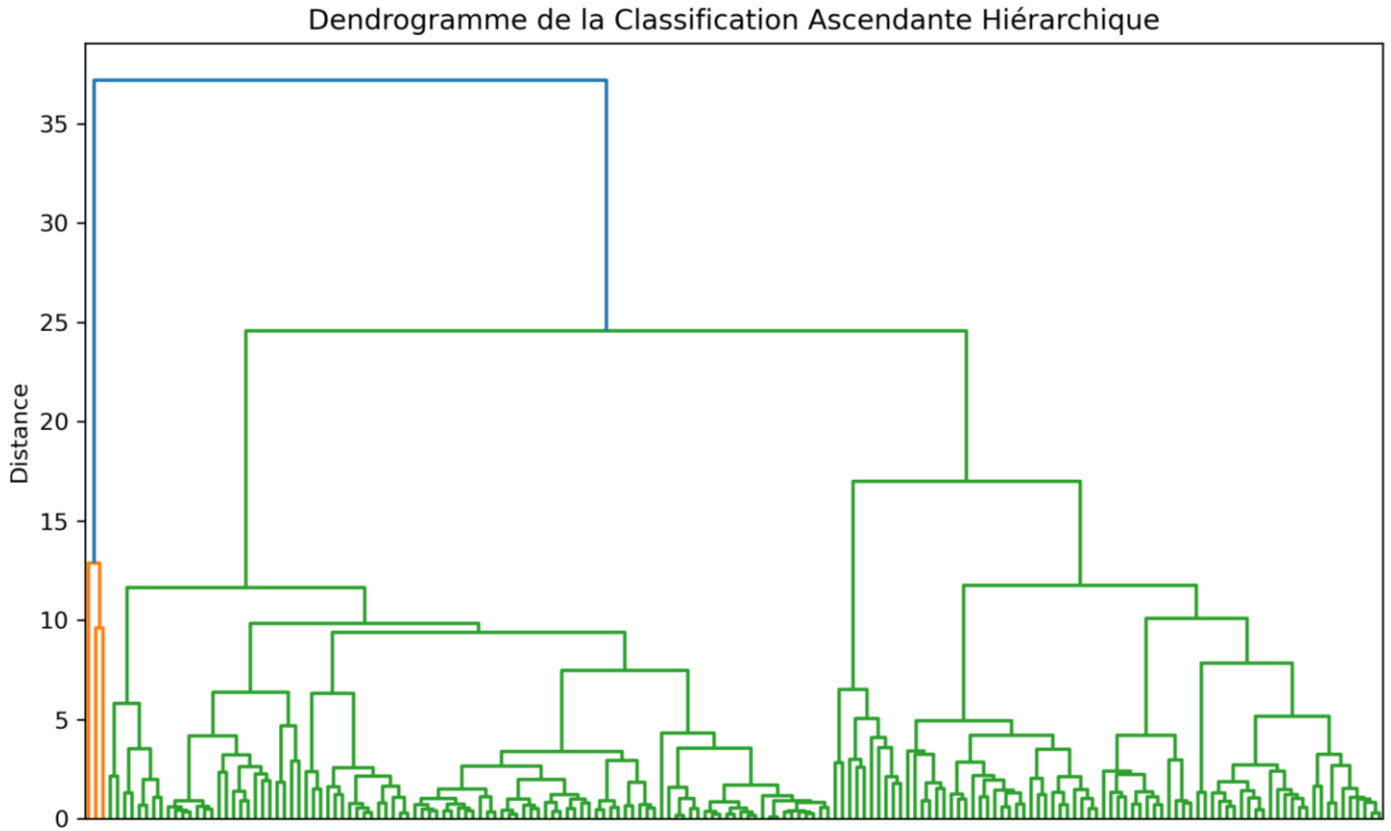
	F1	F2	F3	F4
Dispo (kg/an)	0.155	-0.473	-0.209	0.244
Dispo int.	0.436	0.166	0.021	0.008
Export (qté)	0.365	0.031	-0.166	0.070
Nourriture	0.435	0.169	0.022	0.008
Production	0.439	0.163	-0.046	0.050
Import (qté)	0.157	-0.145	0.586	-0.480
PIB ($)	0.414	0.109	0.158	-0.037
PIB PPA hab($)	0.141	-0.504	0.168	-0.282
Stabilité politique (indice)	0.055	-0.536	0.073	0.001
PIB - Croissance (%)	0.003	0.057	-0.610	-0.768
Dispo (hab)	0.228	-0.340	-0.390	0.179

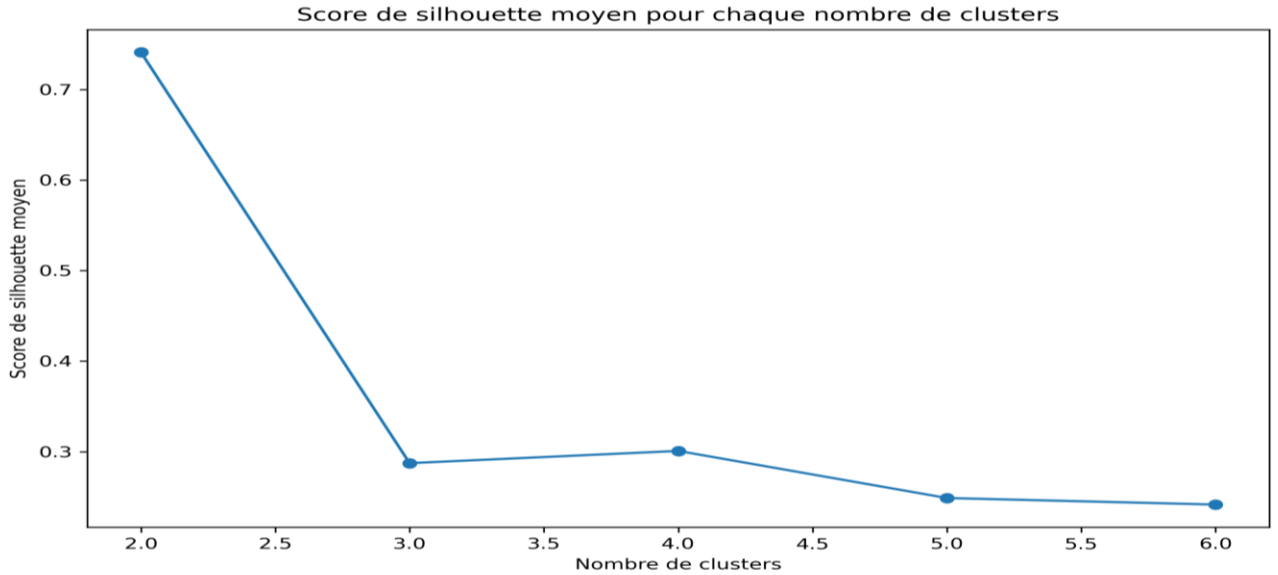




Les variables sont majoritairement bien représentées
La composante 1 est composé de Production, PIB, Nourriture , Export et Dispo intérieure. Elles sont positivement corrélées.
La stabilité politique, PIB PPA et Dispo (kg) sont négativement corrélées et composent le plus la composante 2

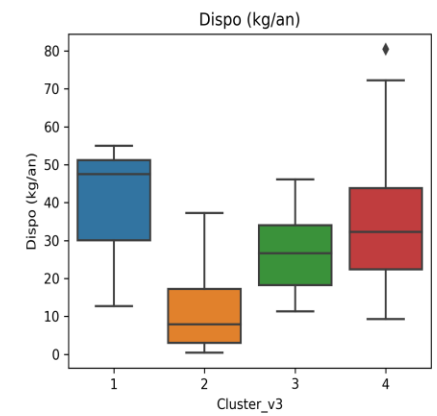
CAH

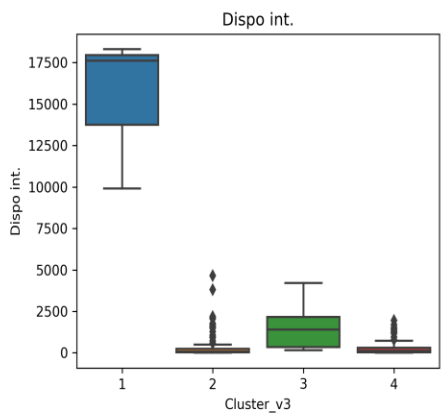


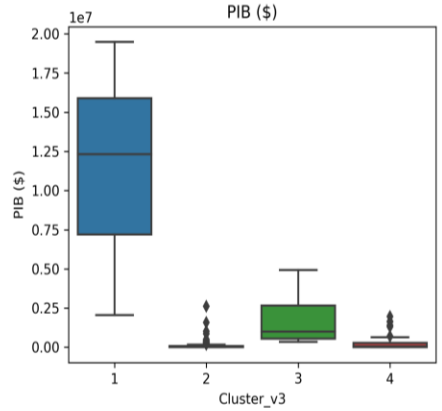


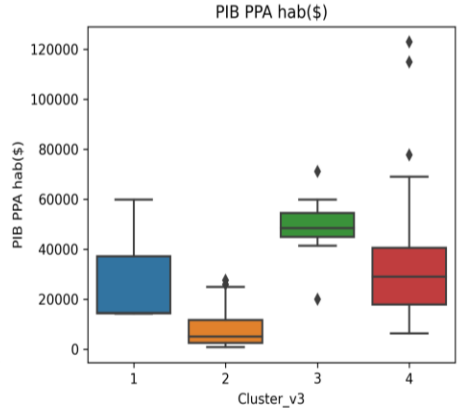

```
Score de silhouette - 3 clusters : 0.287
Score de silhouette - 4 clusters : 0.299
Score de silhouette - 5 clusters : 0.298
Score de silhouette - 6 clusters : 0.268
```

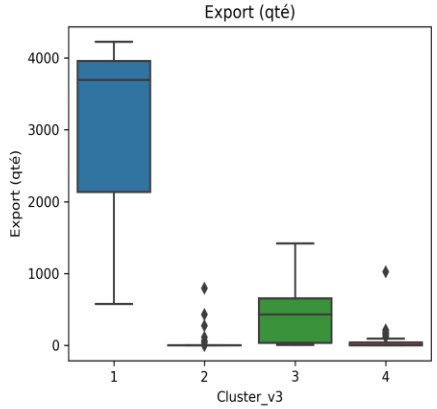
Résultats

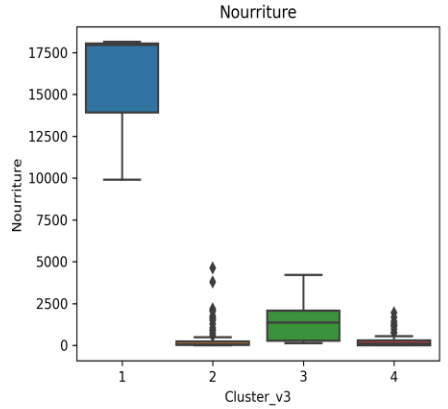


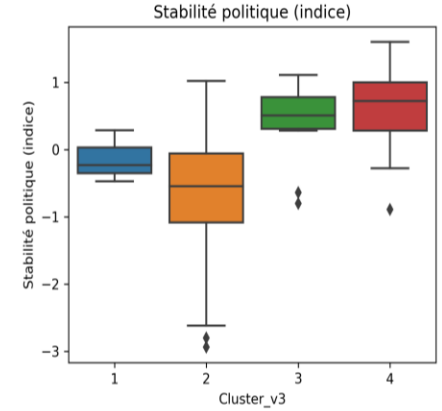


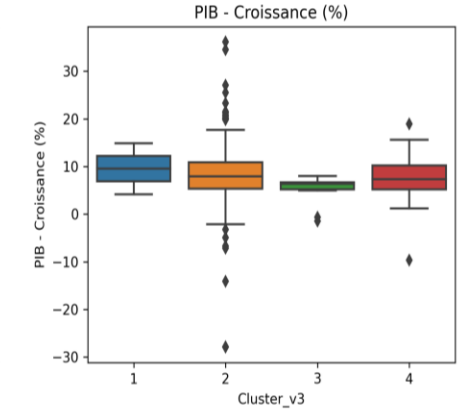


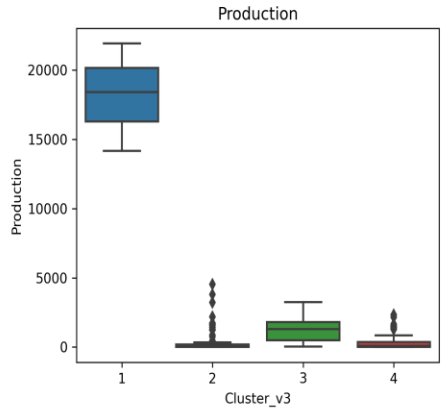


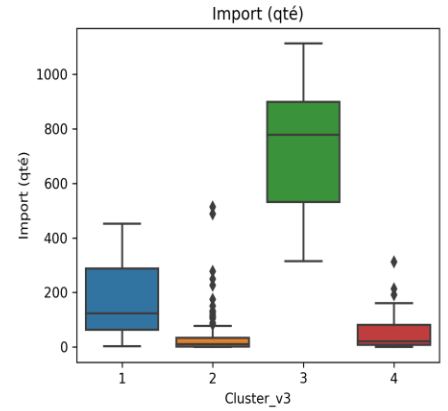


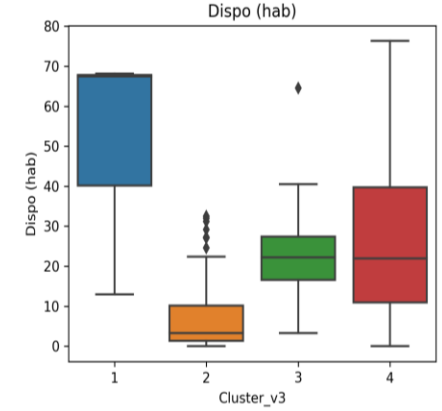






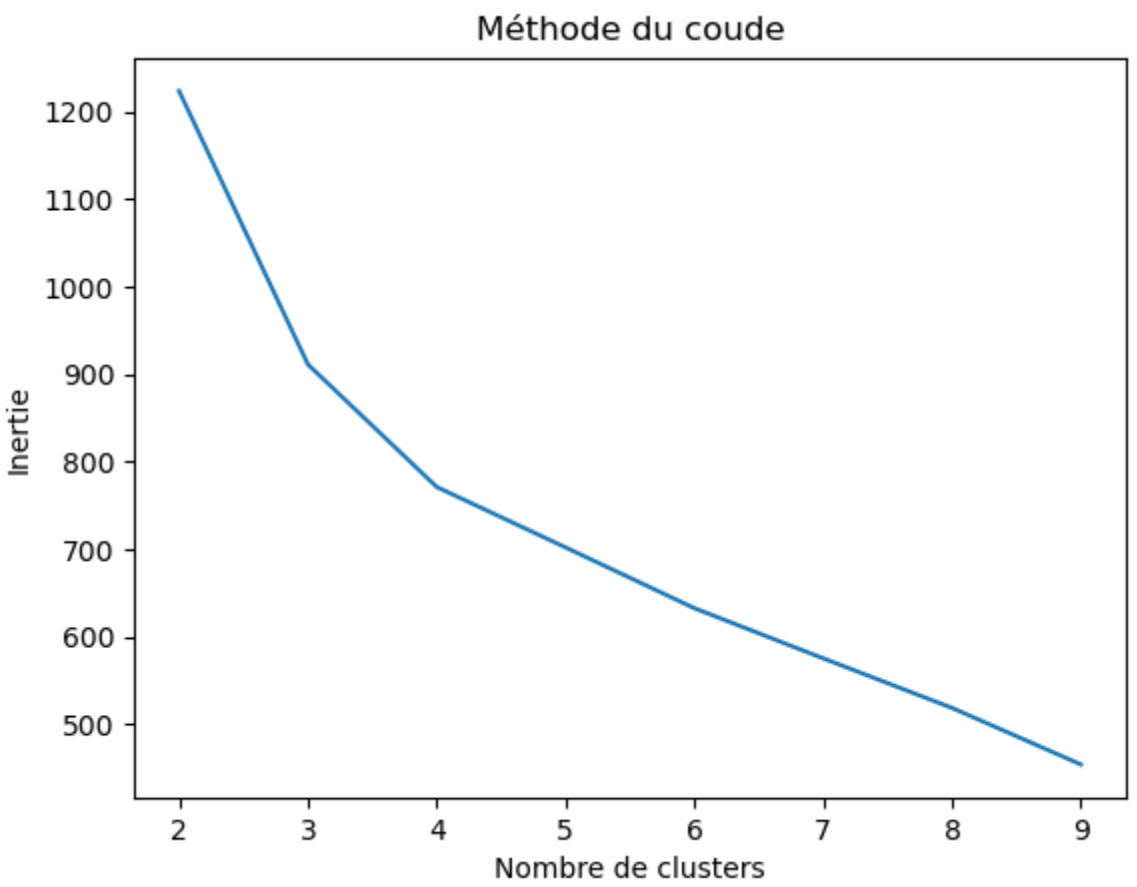


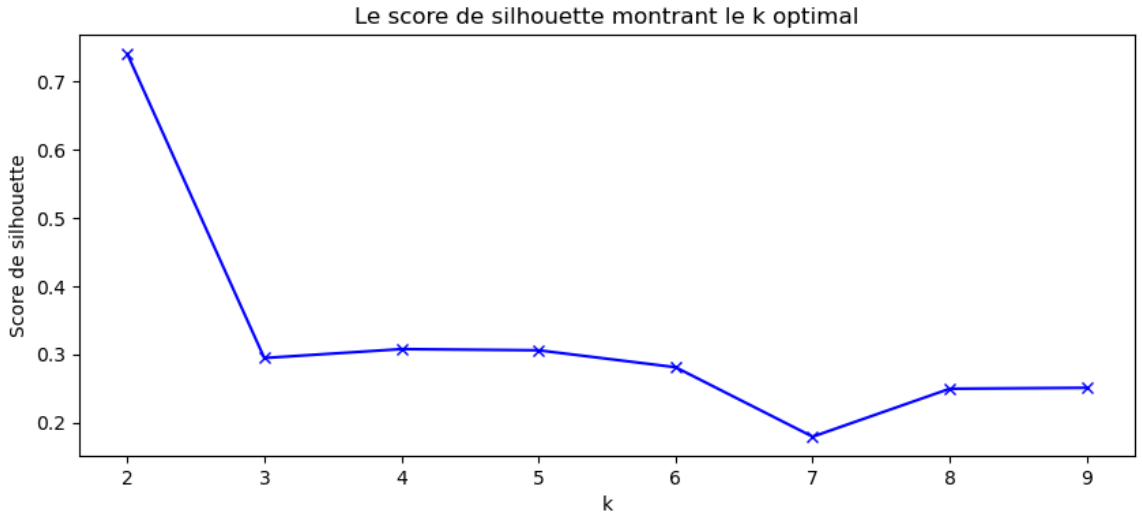




Homogeneity score: 0.861
Calinski Harabasz Score : 61.95

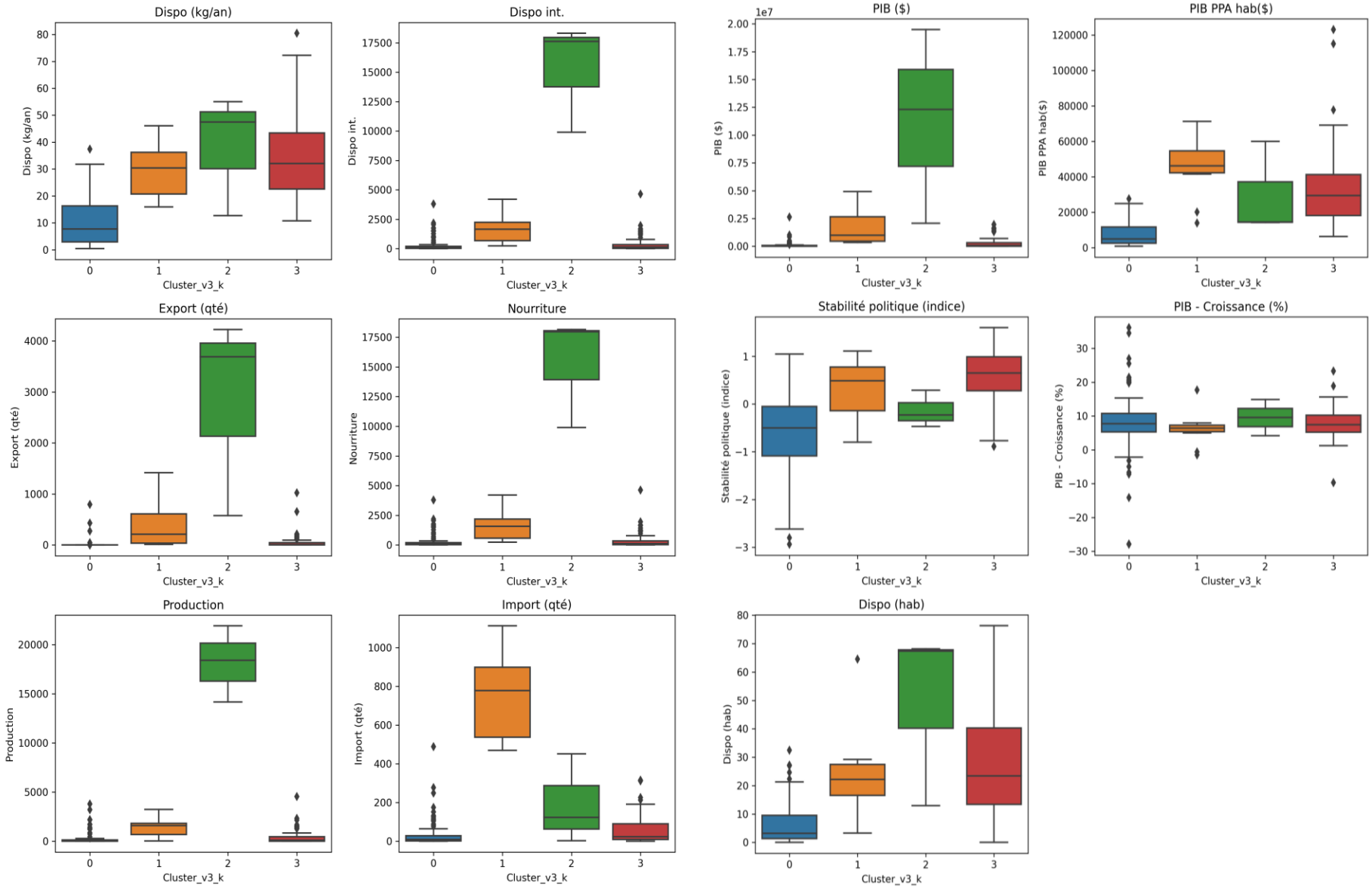
KMeans






```
For n_clusters = 2, silhouette score is 0.741)
For n_clusters = 3, silhouette score is 0.295)
For n_clusters = 4, silhouette score is 0.307)
For n_clusters = 5, silhouette score is 0.278)
For n_clusters = 6, silhouette score is 0.279)
```

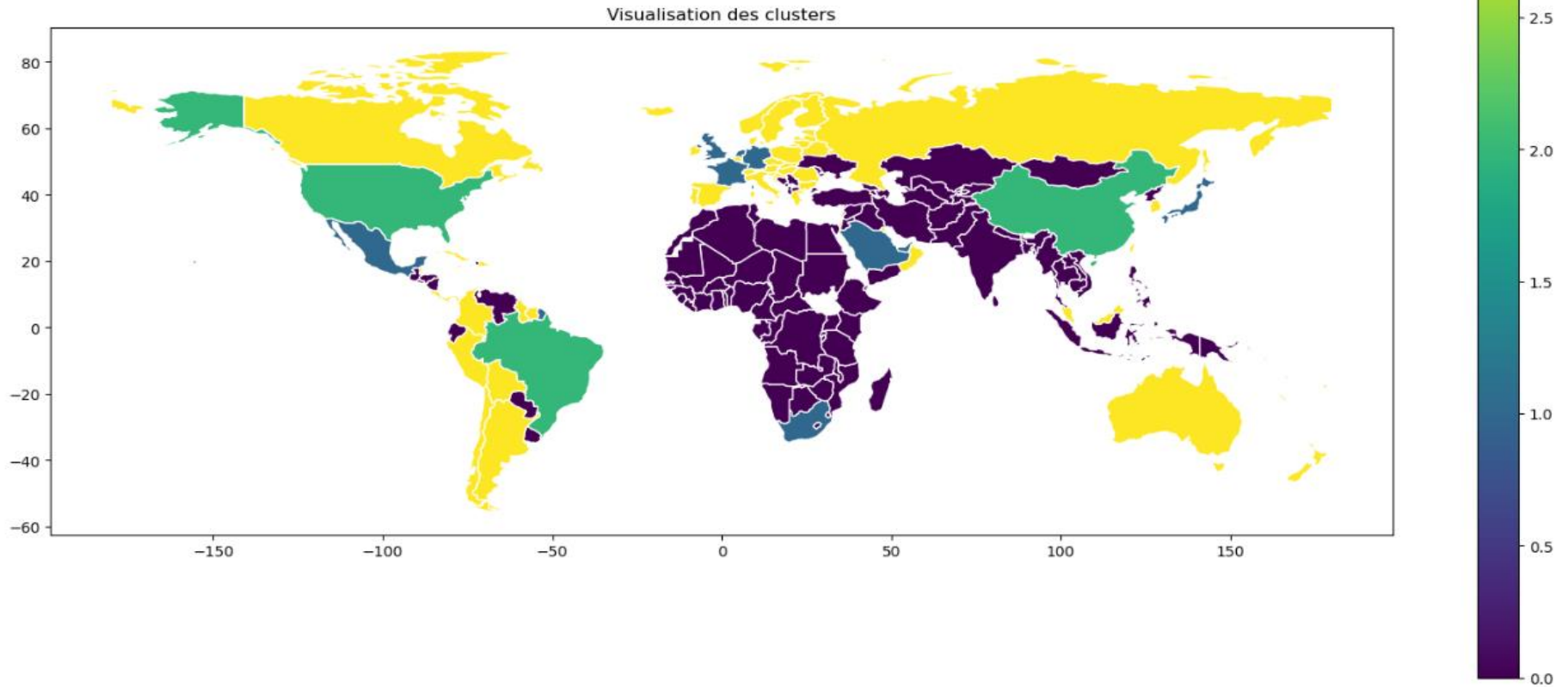
Résultat KMeans



Homogeneity score: 0.969
Calinski Harabasz Score : 61.95

Composition des clusters finaux

Pays cibles = cluster 3



Cluster_v3_k	Dispo (kg/an)	Dispo int.	Export (qté)	Nourriture	Production	Import (qté)	PIB ($)	PIB PPA hab($)	Stabilité politique (indice)	PIB - Croissance (%)	Dispo (hab)
0	10.29	282.91	16.61	281.04	278.34	33.60	105213.11	7459.43	-0.63	8.00	6.76
1	29.41	1667.50	375.30	1621.60	1395.20	749.60	1768933.24	45135.68	0.30	6.21	23.57
2	38.41	15276.00	2830.33	15333.00	18162.33	192.67	11283781.21	29545.40	-0.14	9.56	49.51
3	34.16	405.88	59.10	393.34	418.21	59.82	278890.01	33342.34	0.60	7.65	26.65