PRODUISEZ UNE ÉTUDE DE MARCHÉ AVEC R OU PYTHON

Par Claude Olukoya

Société ‘LA POULE QUI CHANTE’

Est une entreprise française d’agroalimentaire. Son activité principale est l’élevage et la vente de poulets sous le label “Poulet Agriculture Biologique”.

Son activité actuelle est franco-française mais Patrick, le PDG de l’entreprise souhaite évaluer la possibilité de se développer à l'international.

MA MISSION

En tant que le Data Analyst chez La Poule Qui Chante, j'ai pour mission de proposer une **analyse des groupements de pays** que l'on peut **cibler** pour exporter nos poulets:

1. Pars des données de la FAO *(Food and Agriculture Organization)*
2. Préparer et nettoyage des données
3. L'exploration des données (ACP, cercle de corrélation, classification hierarchique, K-Means

LES DONNÉES DE DÉPART

Je suis en autonomie sur ce projet donc c'est à moi de sélectionner les données basées sur (PESTEL) pour Politique, Economique, Socioculturel, Technologique, Ecologique et Légal:

Fichiers de départ

- Population.csv
- Disponibilité alimentaire.csv

Fichiers que j'ai cherchés du site (FAO)

- *Croissance annuelle.csv**
- *Stabilité politique.csv**

PARTIE I : PRÉPARATION & NETTOYAGE

Le DataFrame : ‘croissance_annuelle’

- Vérification des valeurs manquantes
- Exclusion des variables inutiles en ne gardant que (Zone, Valeur, Produit)
- Remodelage du DF en utilisant pivot table pour visualiser les valeurs de la colonne Elément

Le DataFrame : ‘stabilité_politique’

- Vérification des valeurs manquantes
- Exclusion des variables inutiles en ne gardant que (Zone, Valeur, Produit)
- Remodelage du DF en utilisant pivot table pour visualiser les valeurs de la colonne Elément

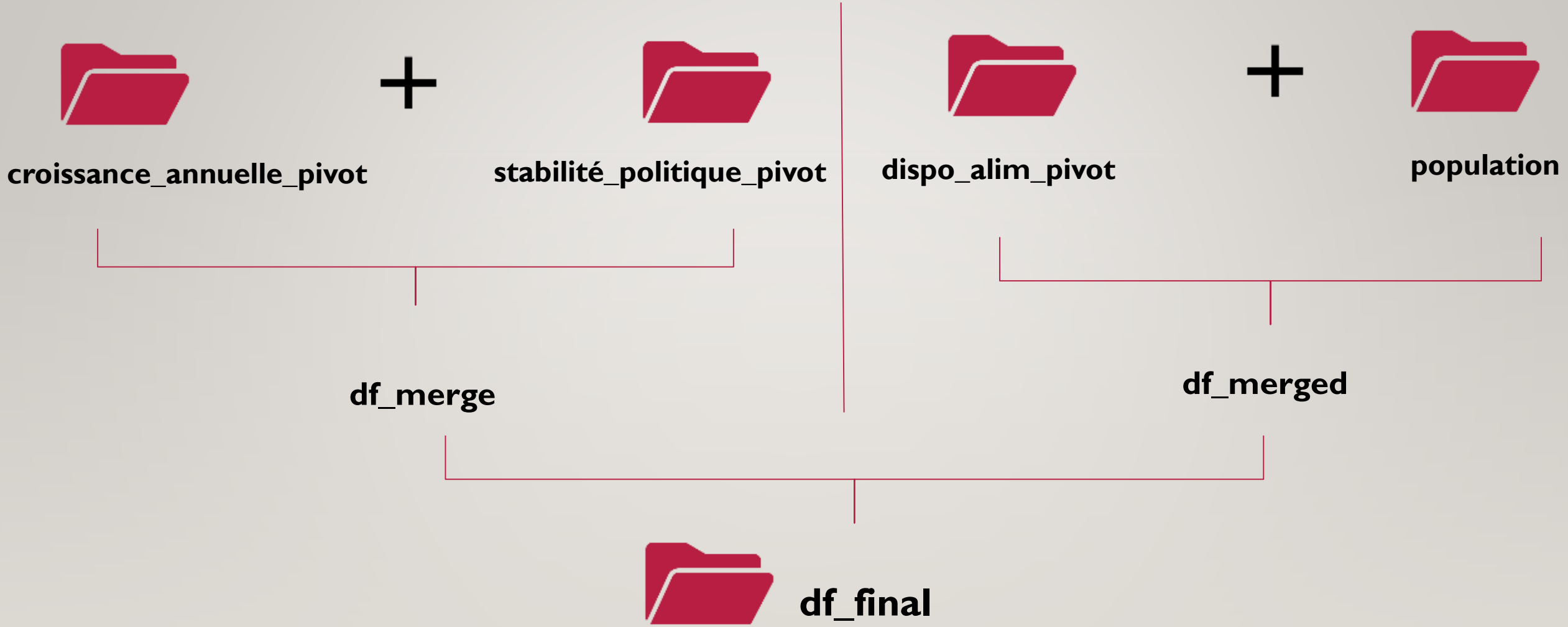
Le DataFrame : ‘population’

- Affichage des valeurs uniques dans chaque variable
- Recherche des valeurs manquantes dans chaque colonne et ligne
- Recherche des doublons sur axis 0 & 1
- Filtrage pour ne garder que les lignes où la population = 2017
- Exclusion des variables inutiles en ne gardant que (Zone, Valeur)
- Modification du nom la colonne ‘Valeur’ en ‘Population’
- Harmonisation des unités : multiplication de la colonne Population par 100

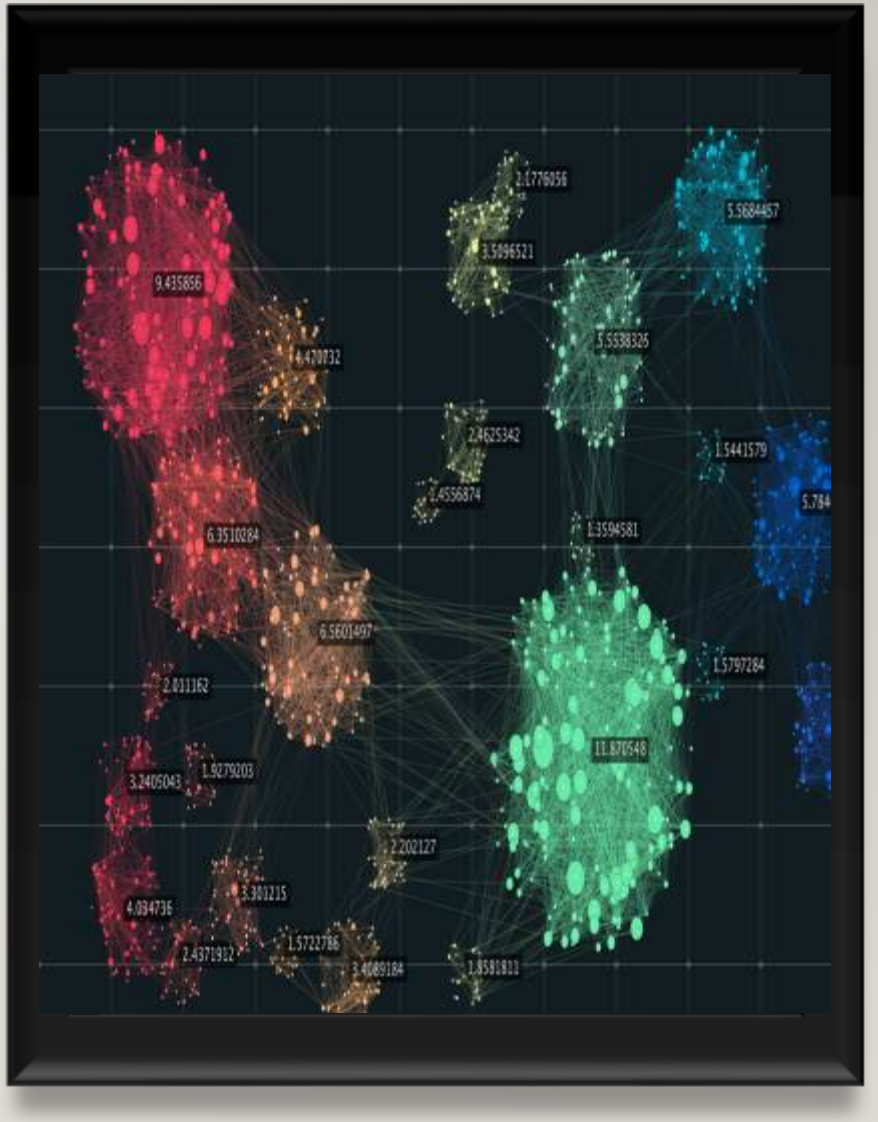
Le DataFrame : ‘dispo-alimentaire’

- Affichage des valeurs uniques dans chaque variable
- Recherche des valeurs manquantes dans chaque colonne et ligne
- Recherche des doublons sur axis 0 & 1
- Exclusion des variables inutiles en ne gardant que (Zone, Valeur, Elément, Produit)
- Mise à jour du DF pour garder les lignes où la variable ‘Produit’ = Volailles
- Remodelage du DF en utilisant pivot table pour visualiser les valeurs de la colonne Elément

4. JONCTION :


croissance_annuelle_pivot
+
stabilité_politique_pivot
dispo_alim_pivot
+
population
df_merge
df_merged
df_final

PARTIE II : ACP, CLUSTERING, VISUALISATIONS



Le but de cette partie est itératif pour visualiser le comportement de L'ACP, CAH, K-means clusterings :

Itération 1: Analyse 7 variables originales avec les outliers

Itération 2 : Analyse 12 variables : (3 variables dérivées avec les outliers)

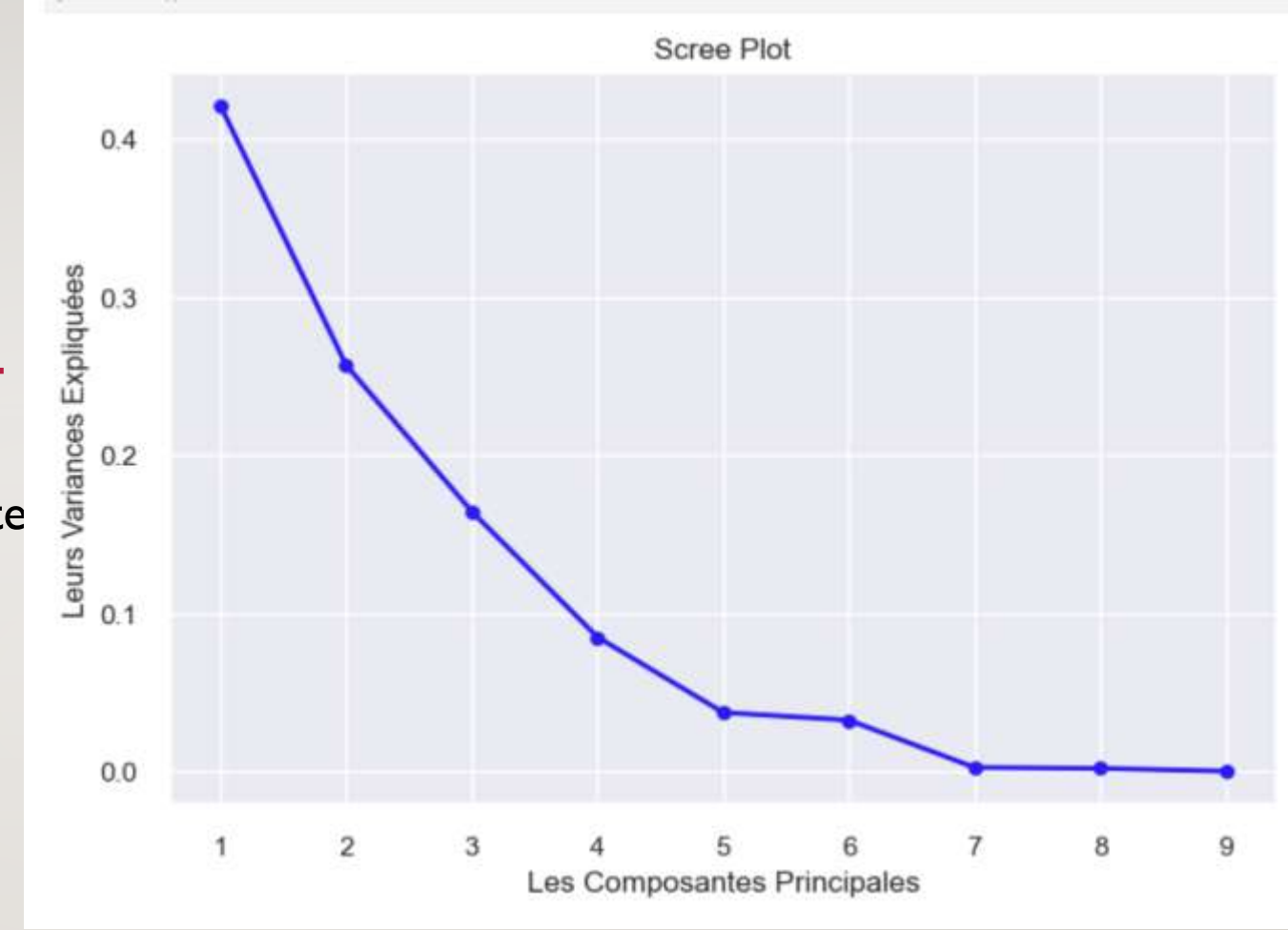
Itération 3: Analyse 9 variables originales sans les outliers

Puisqu'on a un maximum de 25 slides pour cette présentation, je ne vais souligner que l'itération 3.

ANALYSE EN COMPOSANTES PRINCIPALES (ACP)

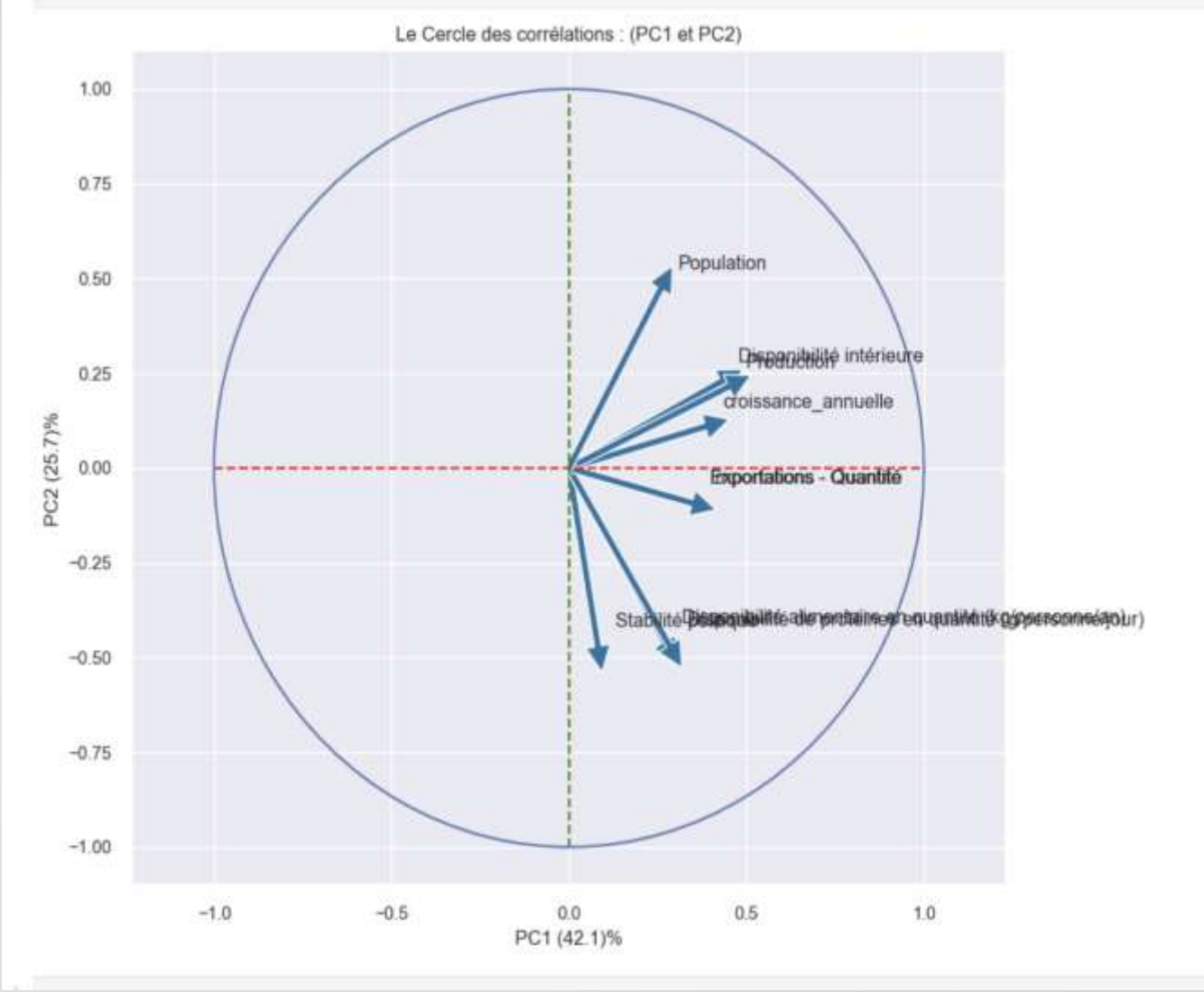
- Ici avec un Scree-plot, on visualise la proportion de variances que détiennent les 9 premières composante

- On observe que les 3 premières composantes capturent presque 85% de la variance totale



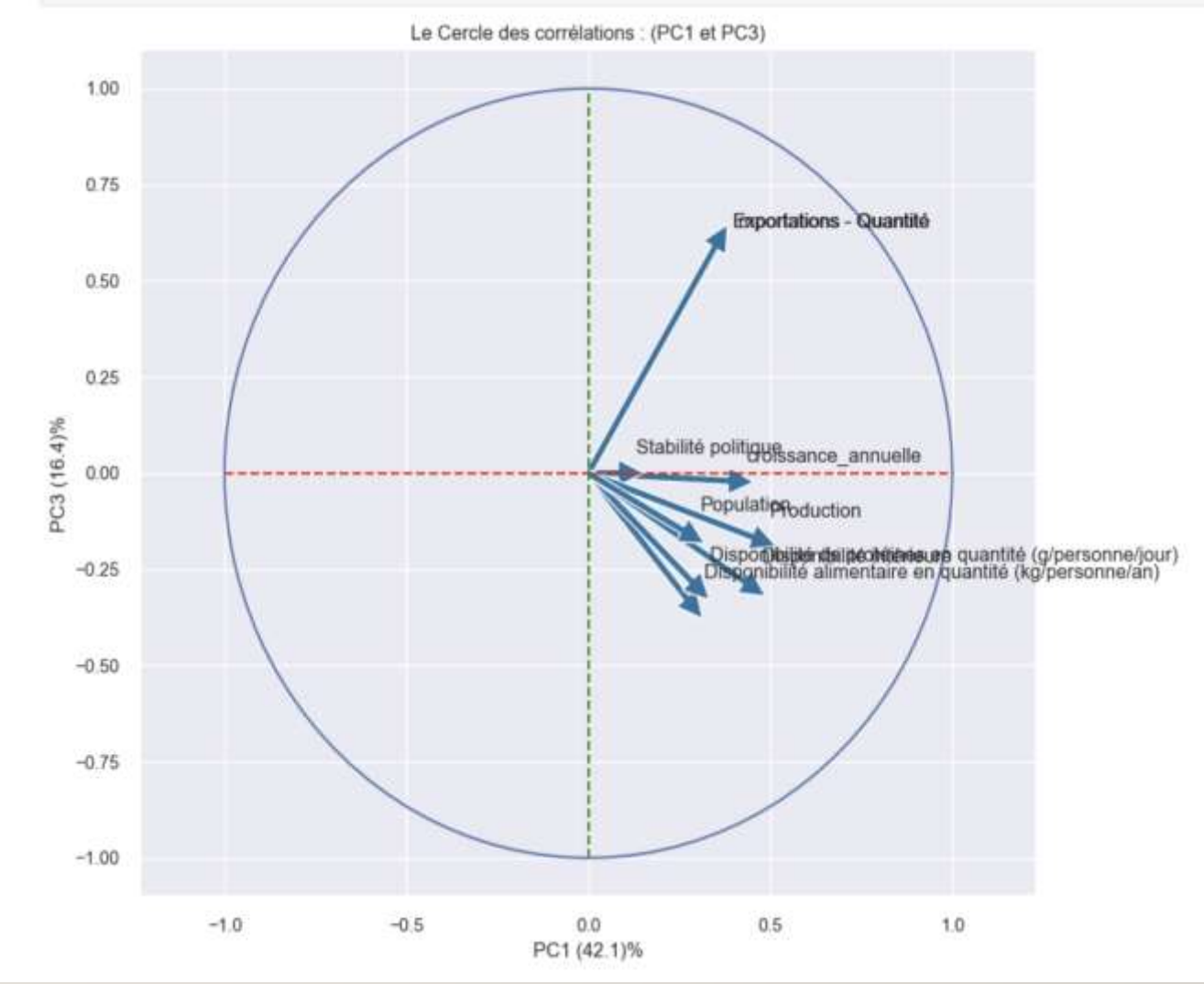
CERCLES DE CORRÉLATION PC1 & PC2

- La variable *Population* est un peu positivement corrélée avec PC2 (coefficient Pearson) à 0.55 mais très peu corrélée positivement avec PC1 (coefficient Pearson) à 0.3
- Les variables *Disponibilité intérieure, Production, croissance annuelle* quant à elles sont positivement corrélées avec une combination de PC1 & PC2
- La variable *Stabilité politique* est moyennement corrélé avec PC2 et un peu positivement corrélée avec PC1



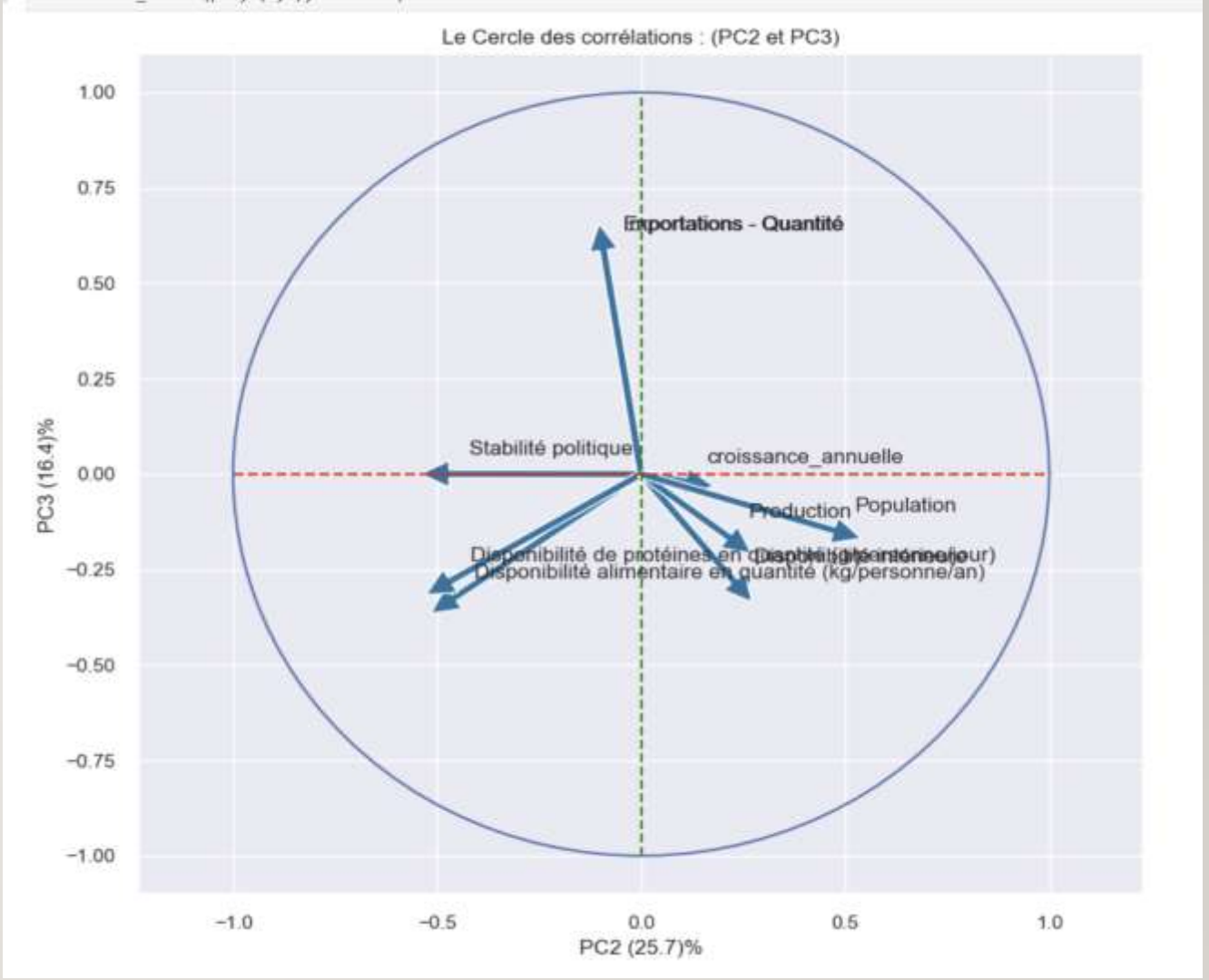
CERCLES DE CORRÉLATION PC1 & PC3

- La variable *Exportation Quantité* est moyennement positivement corrélée avec PC1 (coefficient Pearson) à environ 0.4 mais très corrélée positivement avec PC2 (coefficient Pearson) à 0.7



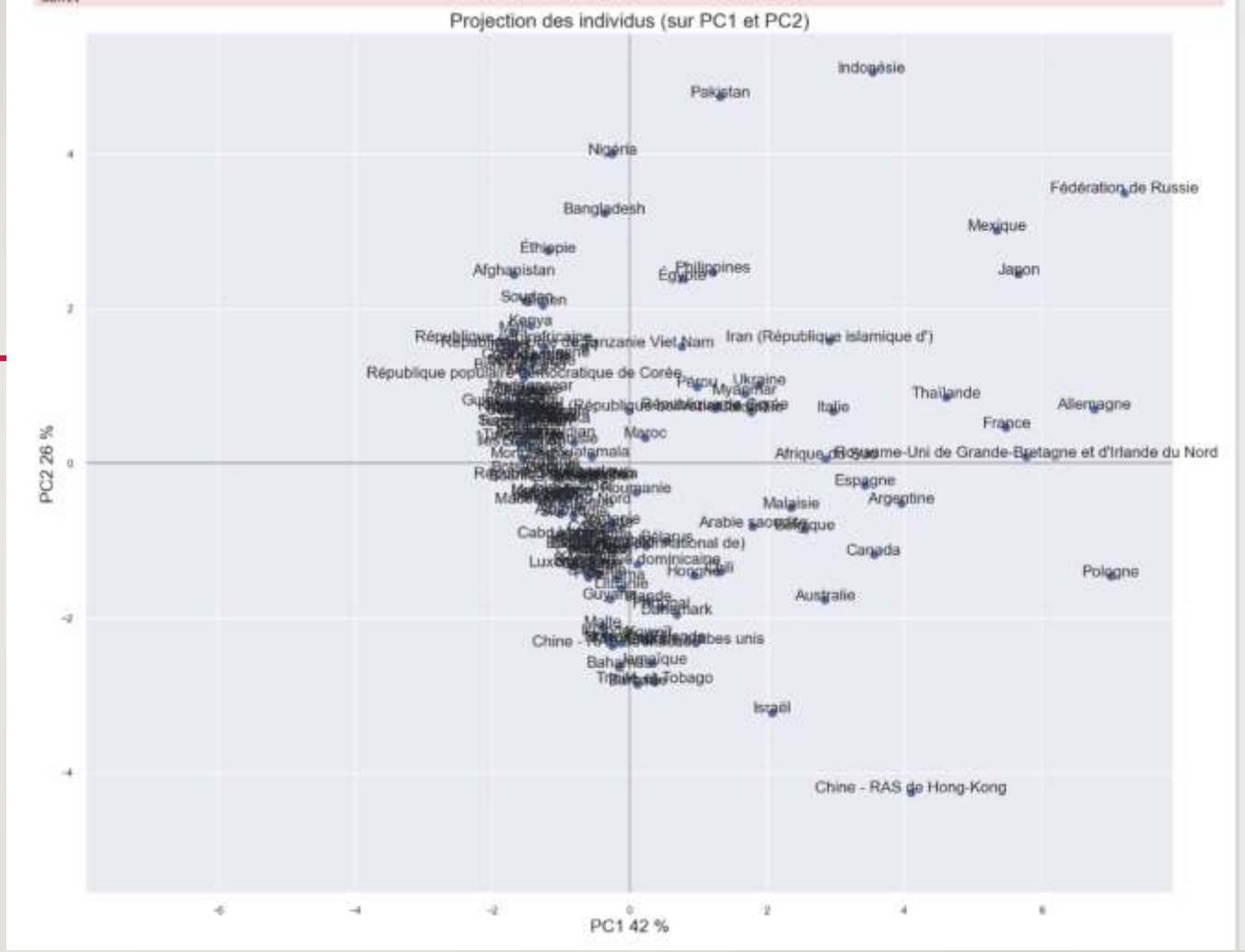
CERCLES DE CORRÉLATION PC2 & PC3

- La variable *Population* est un peu positivement corrélée avec PC1 (coefficient Pearson) à environ 0.2 mais très corrélée positivement avec PC2 (coefficient Pearson) à 0.7
- *Disponibilité de proteins* et *Disponibilité alimentaire* sont négativement corélées avec PC2 & PC3
- *Stabilité politique* est négativement corrélée avec PC2 et moyennement corrélée avec PC3

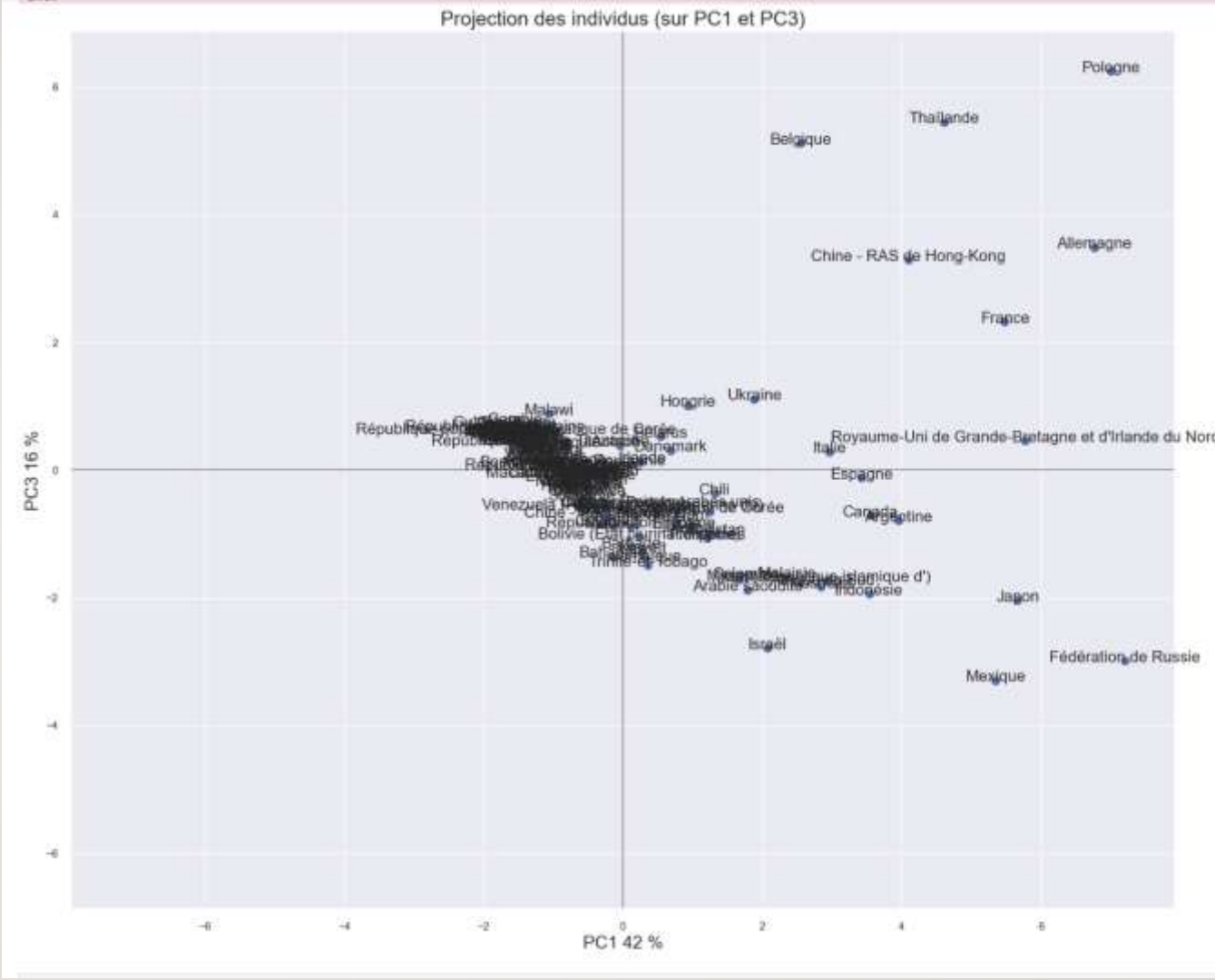


PROJECTION DES INDIVIDUS PC1 & PC2

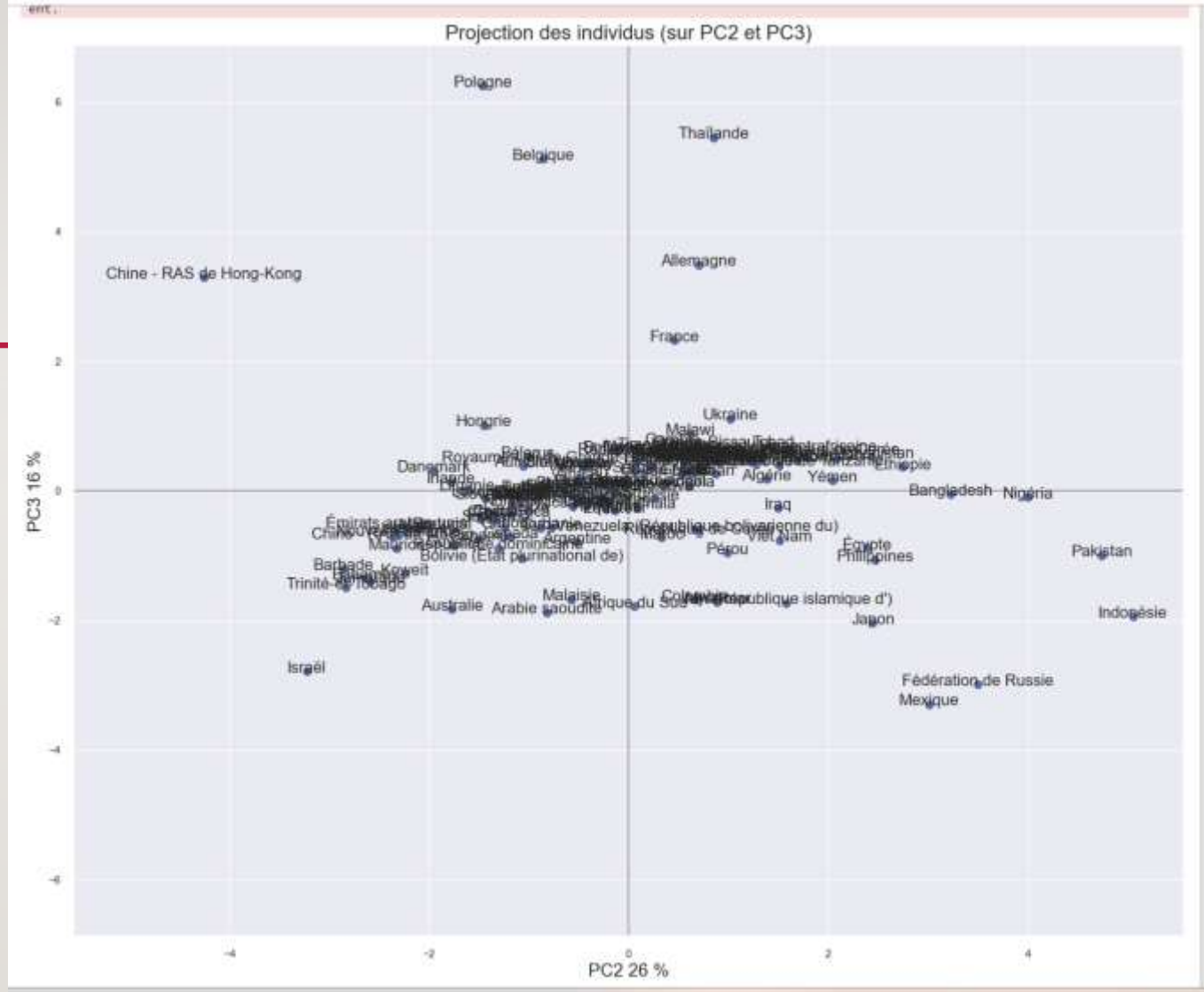
- On constate la convergence des pays plutôt vers le centre

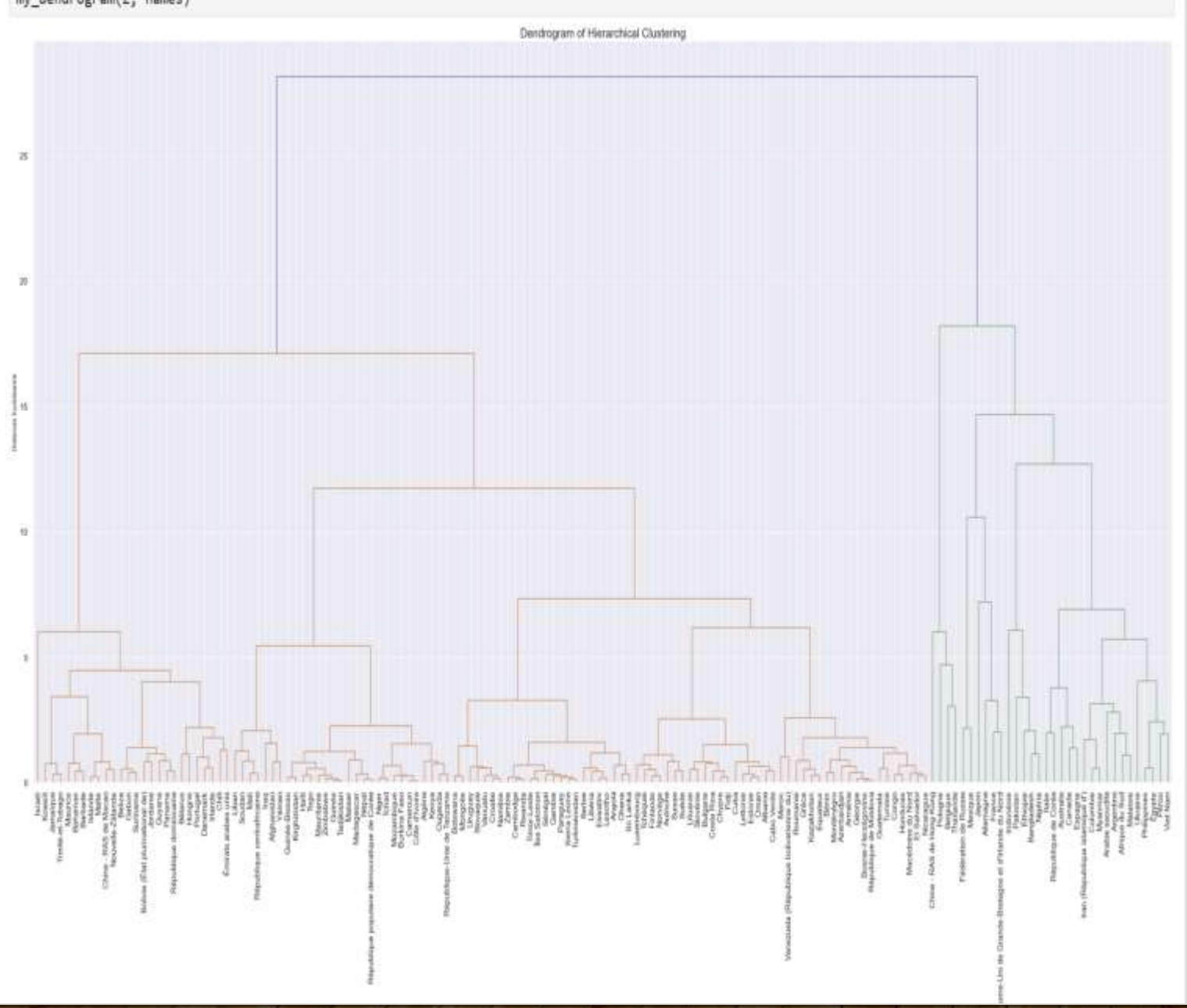


PROJECTION DES INDIVIDUS PC1 & PC3



PROJECTION DES INDIVIDUS PC2 & PC3





CLASSIFICATION HIERARCHIQUE

- Après avoir initialisé et normalisé les données, on utilise un dendrogramme pour regrouper les individus.
- On choisit 4 clusters pour l'harmoniser avec le nombre de clusters de K-Means

CLASSIFICATION ASCENDANTE HIÉRARCHIQUE – ANALYSE DES CLUSTERS

cluster	croissance_annuelle	Stabilité politique	Disponibilité alimentaire en quantité (kg/personne/an)	Disponibilité de protéines en quantité (g/personne/jour)	Disponibilité intérieure	Exportations - Quantité	Production	Importations - Quantité	Population
1	-0.26	0.68	1.31	1.27	-0.32	-0.11	-0.30	-0.11	-0.51
2	-0.32	-0.10	-0.51	-0.51	-0.43	-0.29	-0.45	-0.29	-0.33
3	0.27	0.42	0.65	0.86	0.26	4.90	1.02	4.90	0.13
4	1.26	-0.35	0.35	0.36	1.66	0.36	1.59	0.36	1.53

Cluster 1 : index population assez bas, index importation volailles négatif, index stabilité politique négatif, index croissance annuelle négatif

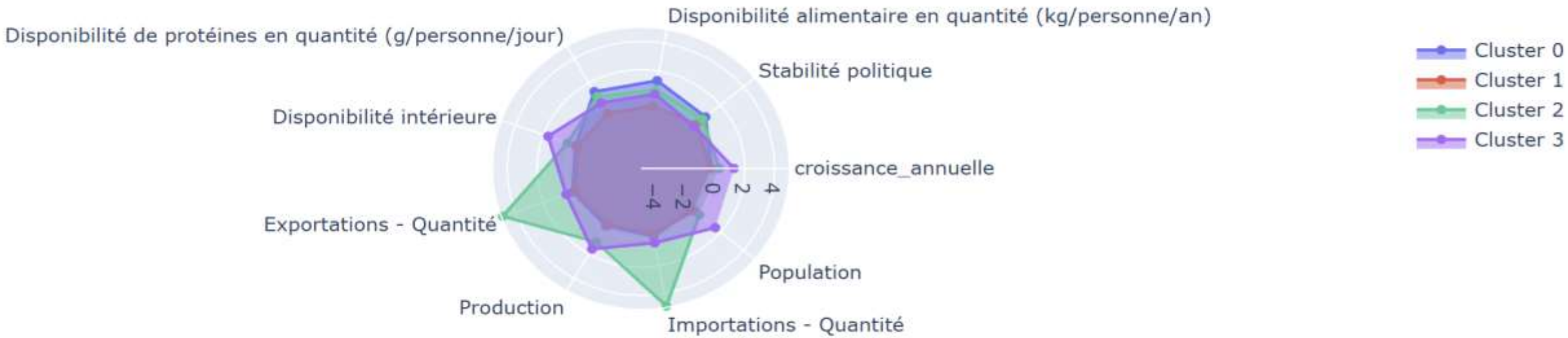
Cluster 2 : index population assez bas, index importation volailles négatif, index stabilité politique négatif, index croissance annuelle négatif

Cluster 3 : index population positif, index importation volailles max, index stabilité politique fort, index croissance annuelle positif

Cluster 4 : index population maximum, index importation volailles positif, index stabilité politique négatif, index croissance annuelle positif

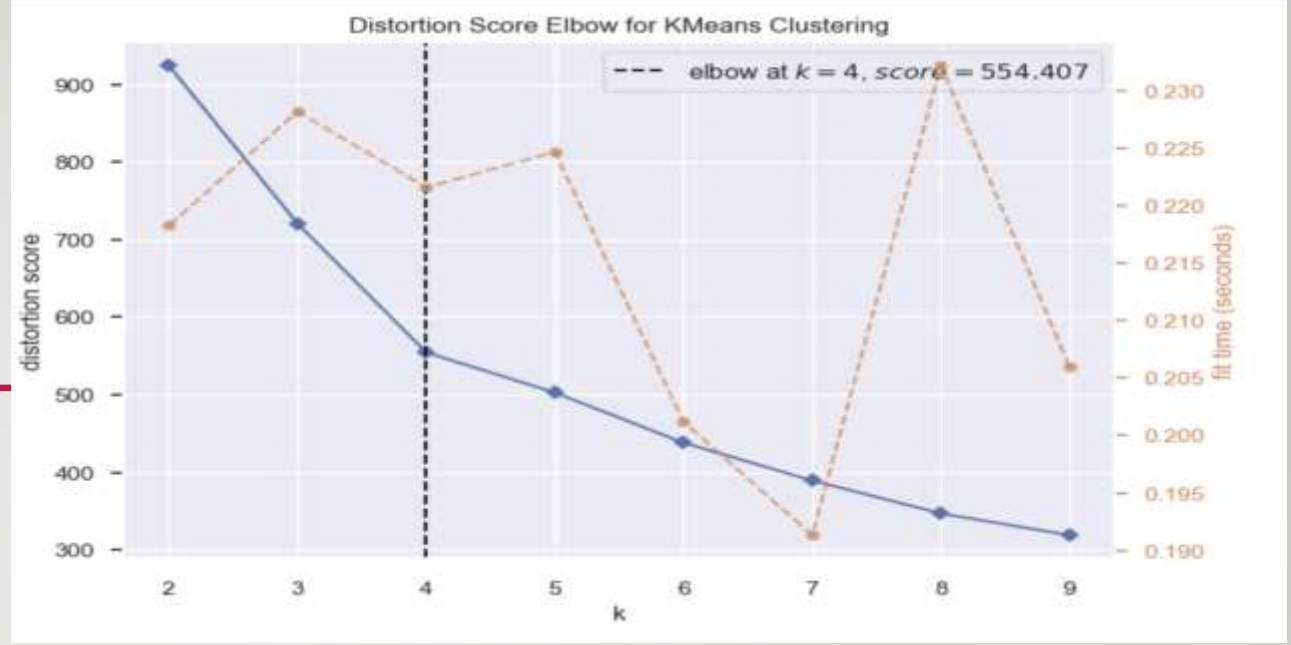
Le Cluster 3 semble le plus pertinent !

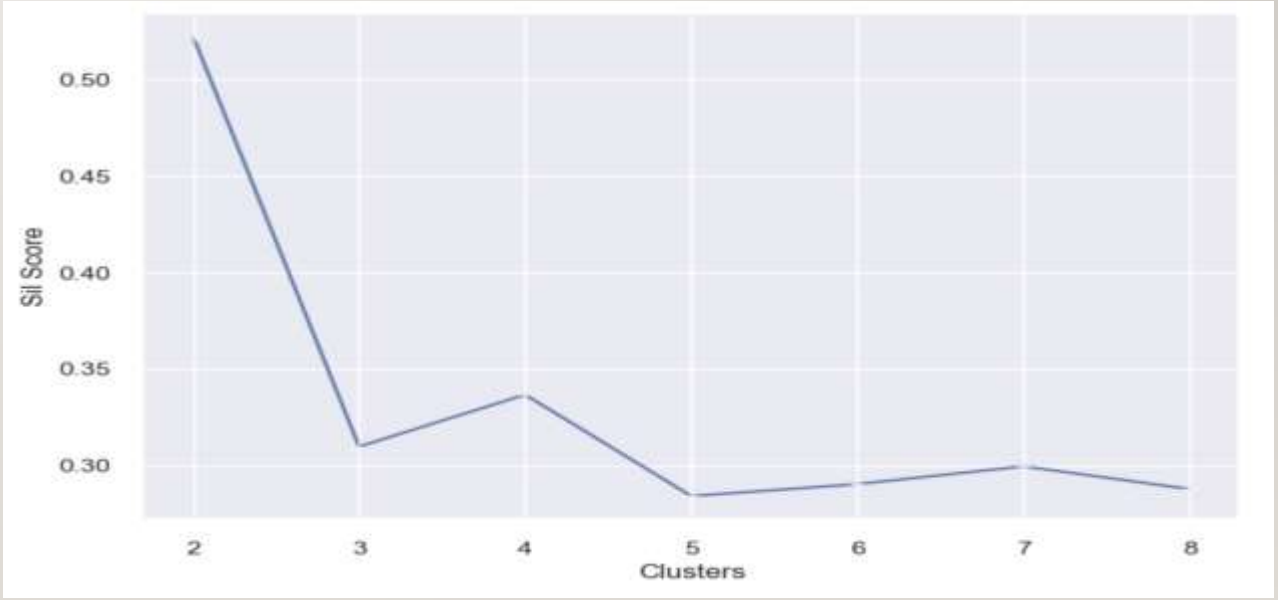
SCATTER-POLAR DES CLUSTERS (CAH)



K-MEANS

J'ai utilisé les méthodes Elbow (Coude) et Silhouette pour trouver le nombre optimal de k (clusters). Ici, les deux méthodes nous indiquent que k= 4





ANALYSE PAR CLUSTERS (K-Means)

	croissance_annuelle	Stabilité politique	Disponibilité alimentaire en quantité (kg/personne/an)	Disponibilité de protéines en quantité (g/personne/jour)	Disponibilité intérieure	Exportations - Quantité	Production	Importations - Quantité	Population
	mean	mean	mean	mean	mean	mean	mean	mean	mean
Cluster									
0	-0.35	-0.49	-0.74	-0.75	-0.43	-0.32	-0.44	-0.32	-0.13
1	1.18	-0.24	0.52	0.51	1.94	0.20	1.90	0.20	1.41
2	1.65	0.48	0.50	0.71	0.73	4.43	1.26	4.43	0.48
3	-0.21	0.77	0.81	0.81	-0.33	-0.16	-0.35	-0.16	-0.49

Cluster 0 : index croissance annuelle négatif, index population négatif, index importation volailles négatif, index stabilité politique négatif

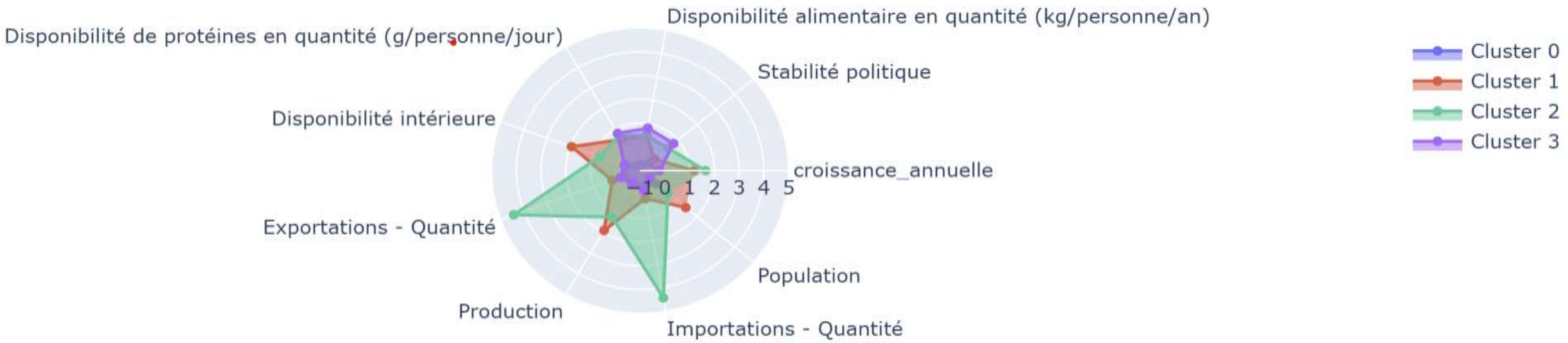
Cluster 1 : index population positif, index importation volailles positif, index stabilité politique positif, index croissance annuelle positif

Cluster 2 : index population positif, index importation volailles maximum, index stabilité politique positif, index croissance annuelle max

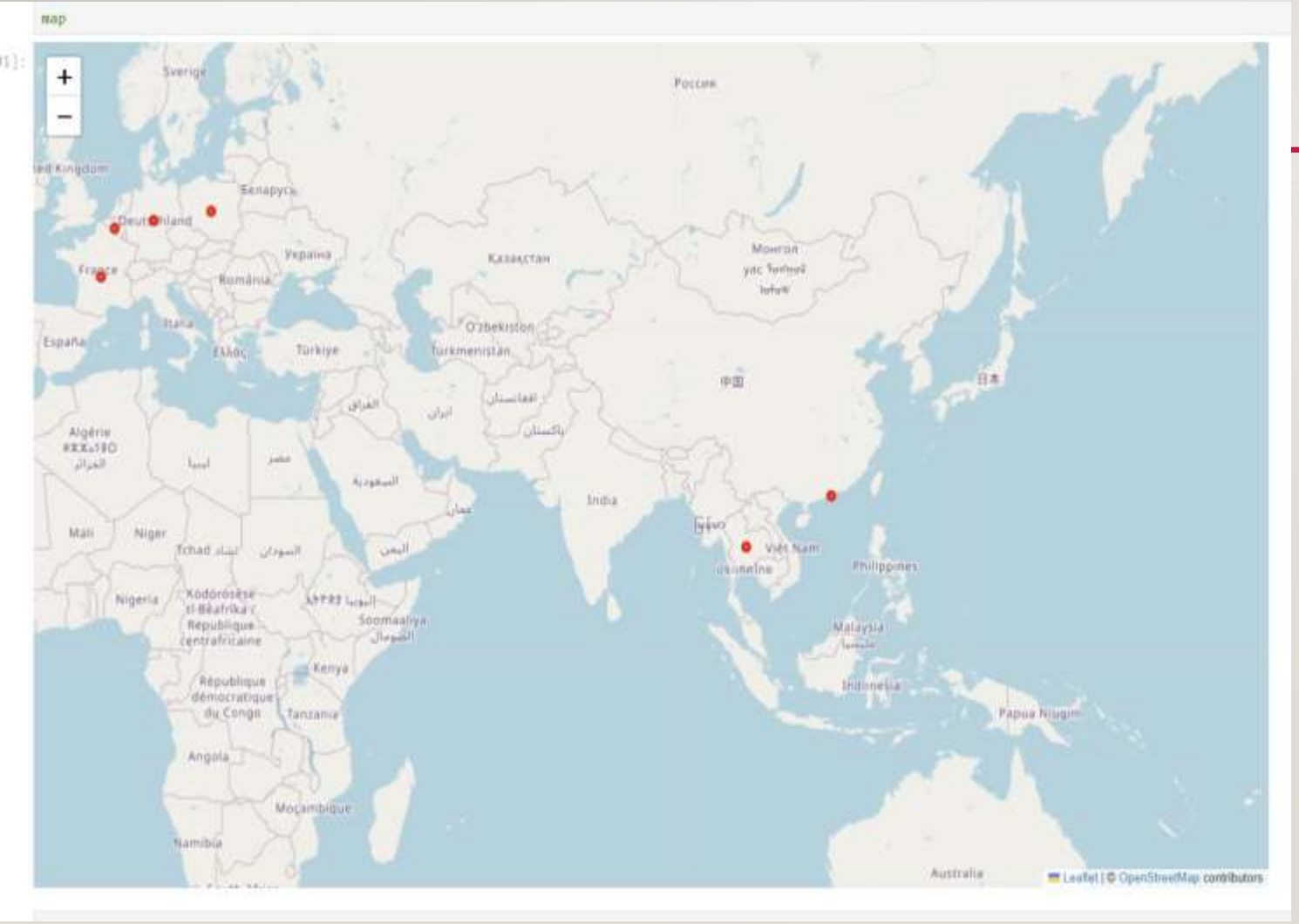
Cluster 3 : index population négatif, index importation volailles négatif, index stabilité politique positive, index croissance annuelle négatif

Le Cluster 2 semble le plus pertinent

SCATTER-POLAR DES CLUSTERS (K-Means)



CARTE DU MONDE AVEC LES PAYS CHOISIS



Les pays sélectionnés sont ***Allemagne, Belgique Chine - RAS de Hong-Kong, France, Pologne, Thaïlande.***