

# Présentation du projet 9 : Produisez une étude de marché avec Python



Préparé par :  
**MAGHLAZI Hanane**

# Projet 9 : Produisez une étude de marché avec Python

- **Introduction**
- **Choix des données**
- **Nettoyage du jeu de données**
- **Classification ascendante hiérarchique:**
  - **Dendrogramme**
- **Matrice de corrélation**
- **ACP**
- **Effectuer un clustering avec la méthode des k-means**
- **Correspondance CAH - K-Means**
- **Choix du cluster et heatmap**
- **Pays sur carte**
- **Conclusion**
- **Discussion**

## Introduction :

- **La poule qui chante, est une entreprise française d'agroalimentaire. Elle souhaite se développer à l'international.**
- **Tous les pays sont envisageables , et afin de les déterminer, une première analyse sera faite pour des groupements de pays que l'on peut cibler pour exporter les poulets.**



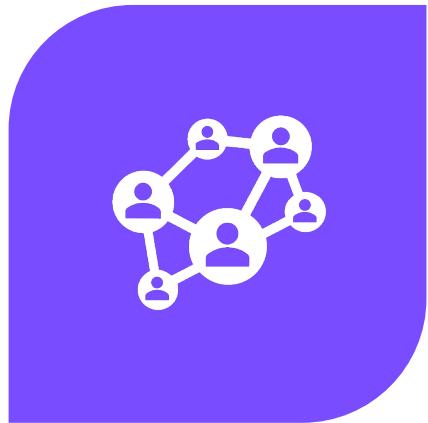
**La poule qui chante**

# Choix des données



# Choix des données

Afin de bien mener mon analyse et **cibler** les **pays** pour exporter les poulets, j'ai choisi 8 **variables** liées à deux indicateurs : disponibilités alimentaires et économiques.



## INDICATEURS LIÉS À LA DISPONIBILITÉS ALIMENTAIRES :

DISPONIBILITÉS ALIMENTAIRES (KCAL/PERSONNE/JOUR), DISPONIBILITÉS EN PROTÉINES (G/PERSONNE/JOUR), IMPORTATIONS (TONNES), EXPORTATIONS (TONNES) ET PRODUCTION (PAR PERSONNE).



## INDICATEURS ÉCONOMIQUES :

PIB, STABILITÉ POLITIQUE ET LA POPULATION CROISSANCE.

# Choix des données

	Pays	Disponibilitealimentairekcal	Diponibiliteproteines	Exportations	Importations	Production	population_croissance	PIB	stabilite_politique
0	Afghanistan	5.0	0.54	0.0	29.0	28.0	74.668889	1.862303e+04	-2.80
1	Afrique du Sud	143.0	14.11	63.0	514.0	1667.0	26.779324	3.490067e+05	-0.28
2	Albanie	85.0	6.26	0.0	38.0	13.0	-7.831734	1.301973e+04	0.38
3	Algérie	22.0	1.97	0.0	2.0	275.0	33.331859	1.700972e+05	-0.92
4	Allemagne	71.0	7.96	646.0	842.0	1514.0	1.544857	3.682602e+06	0.59
...	...	...	...	...	...	...	...	...	...
162	Égypte	50.0	4.51	1.0	110.0	1118.0	40.113909	1.951353e+05	-1.42
163	Émirats arabes unis	147.0	14.80	94.0	433.0	48.0	202.712678	3.856055e+05	0.62
164	Équateur	83.0	6.15	0.0	0.0	340.0	32.364941	1.042959e+05	-0.07
166	Éthiopie	0.0	0.04	0.0	1.0	14.0	60.664762	7.679452e+04	-1.68
167	îles Salomon	18.0	1.51	0.0	6.0	0.0	54.131488	1.215745e+03	0.20

164 rows x 9 columns

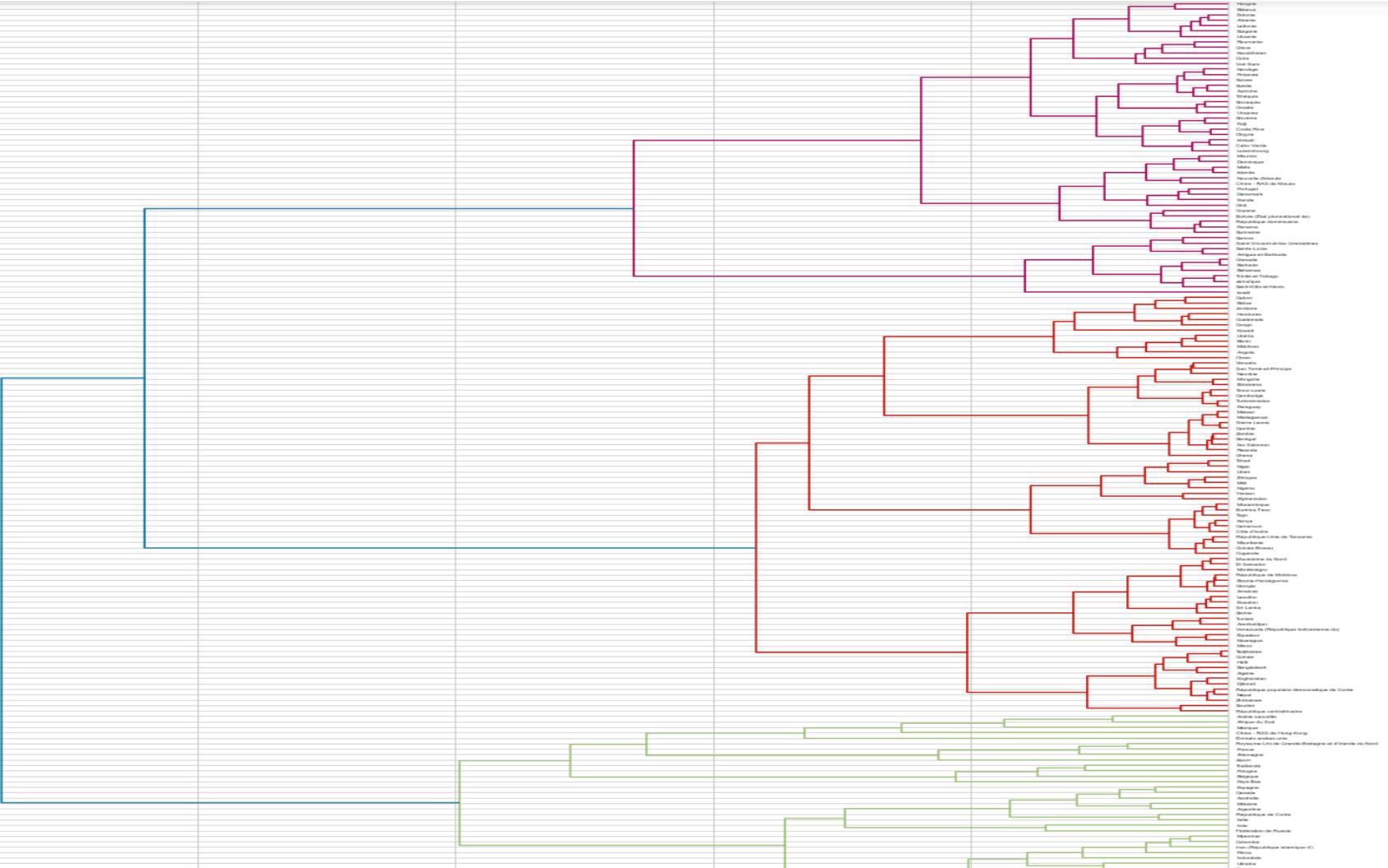
# Nettoyage du jeu de données

- Le nettoyage a été fait en remplaçant les NAN par 0: df\_final = df\_final.fillna(0)
- Deux lignes ont été supprimées car ne contenaient aucune donnée.
- Le remplacement des valeurs <0.1 par 0.
- Une nouvelle colonne ajoutée: population croissance qui est calculée.

# Classification ascendante hiérarchique:

- La Classification Ascendante Hiérarchique (CAH) est une **méthode d'analyse multivariée** automatique.
- Le principe de la CAH est d'effectuer un **regroupement** progressif des individus selon leur degré de ressemblance jusqu'à l'obtention d'une unique classe les regroupant tous .
- La CAH va ensuite rassembler les individus de manière itérative afin de produire un dendrogramme .

# Dendrogramme



- Le dendrogramme suggère un découpage en 3 groupes

# Dendrogramme

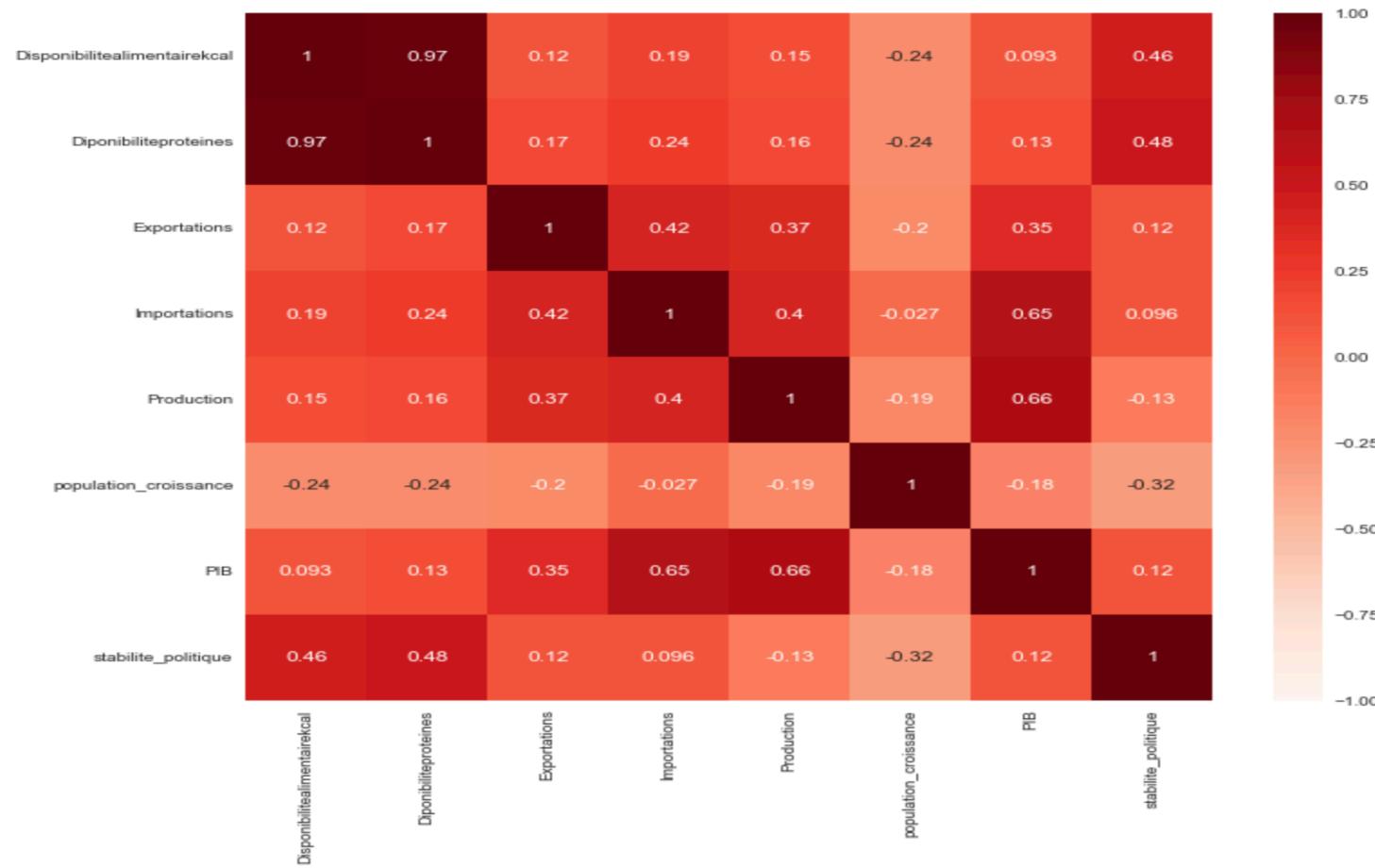
L'analyse des trois groupes du Dendrogramme montre :

Des pays qui exportent et importent beaucoup avec une production élevée, une bonne disponibilité en Kcal et disponibilité protéines et que économiquement ce sont des pays qui ont une croissance moyenne , un PIB élevé et stable politiquement.

Des pays qui n'exportent pas beaucoup voire pas du tout mais qui importent moyennement avec une production moyenne, une disponibilité en Kcal et protéines moyennes et que économiquement ce sont des pays en bonne croissance, plus instable politiquement que le groupe 1 avec un PIB plus bas que le groupe 1.

Des pays qui exportent et importent , mais moins que le groupe 1, plus d'import que d'export, une production moyenne un peu plus élevé que le groupe 2 et économiquement ce dont des pays stables politiquement, avec un PIB assez élevé et une croissance faible.

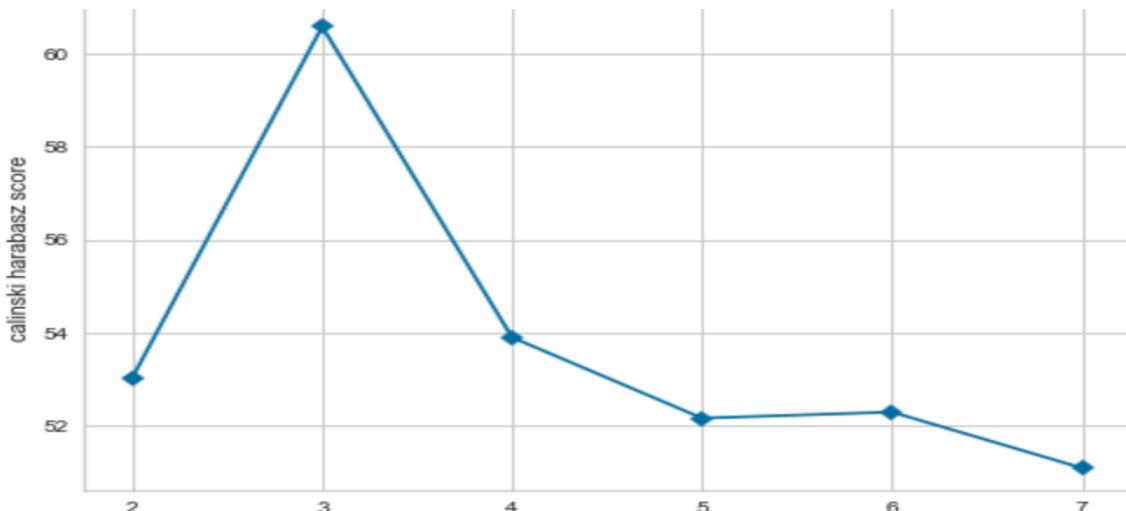
# Matrice de corrélation



- Une relation linéaire positive existe entre les variables Dispo alimentaires , dispo protéines , exportations, importations et production.
- Une relation linéaire négative existe pour les variables population croissance et stabilité politique avec des coefficients de corrélation de Pearson négatifs.
- Ce qui indique que, plus la population croissance et la stabilité politique augmentent, plus les variables Dispo alimentaires , dispo protéines , exportations, importations et production diminuent.

# ACP

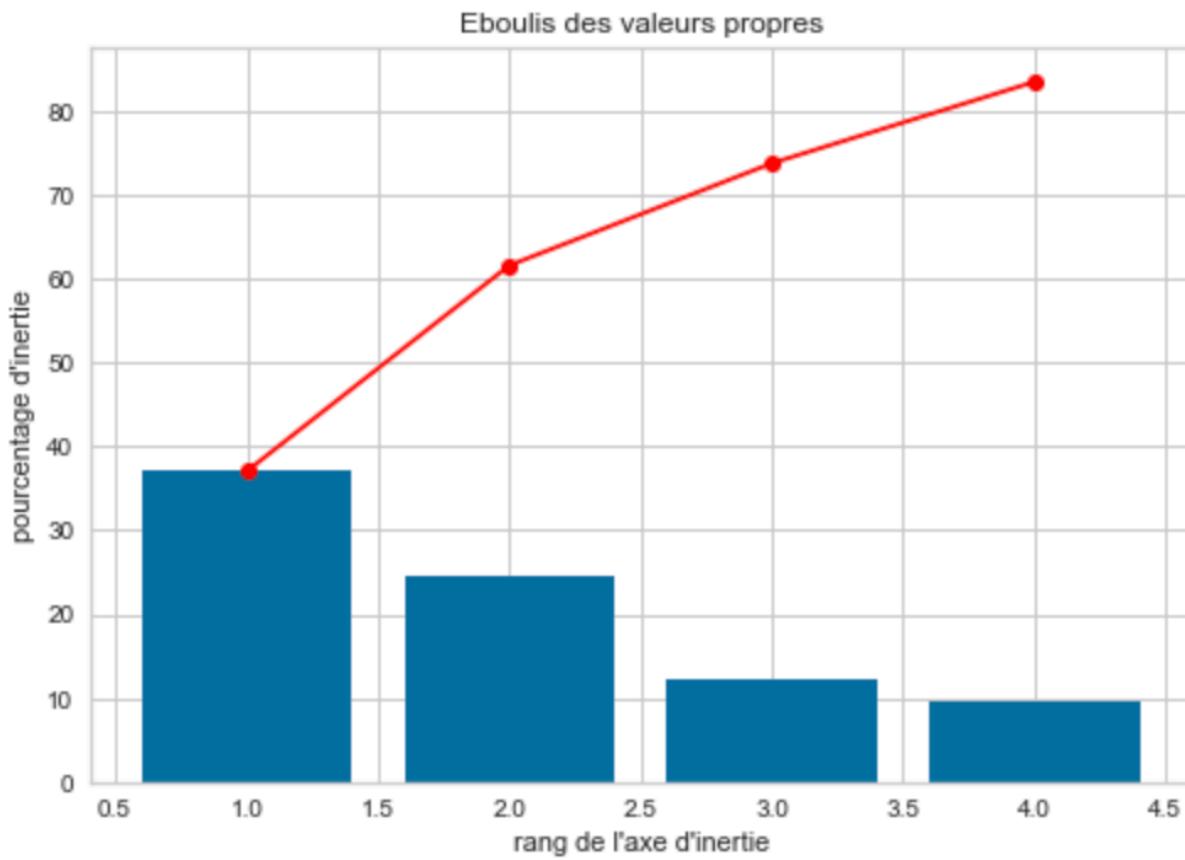
- l'Analyse en Composantes Principales permet d'étudier la variabilité entre les individus, c'est-à-dire quelles sont les différences et les ressemblances entre individus et les liaisons entre les variables .
- On projette les données sur 4 dimensions selon la méthode du coude :



```
047]: 1 #pourcentage de variance expliquée par chacune des composantes.  
2 print(pca.explained_variance_ratio_ )  
3 print(pca.explained_variance_ratio_.sum())
```

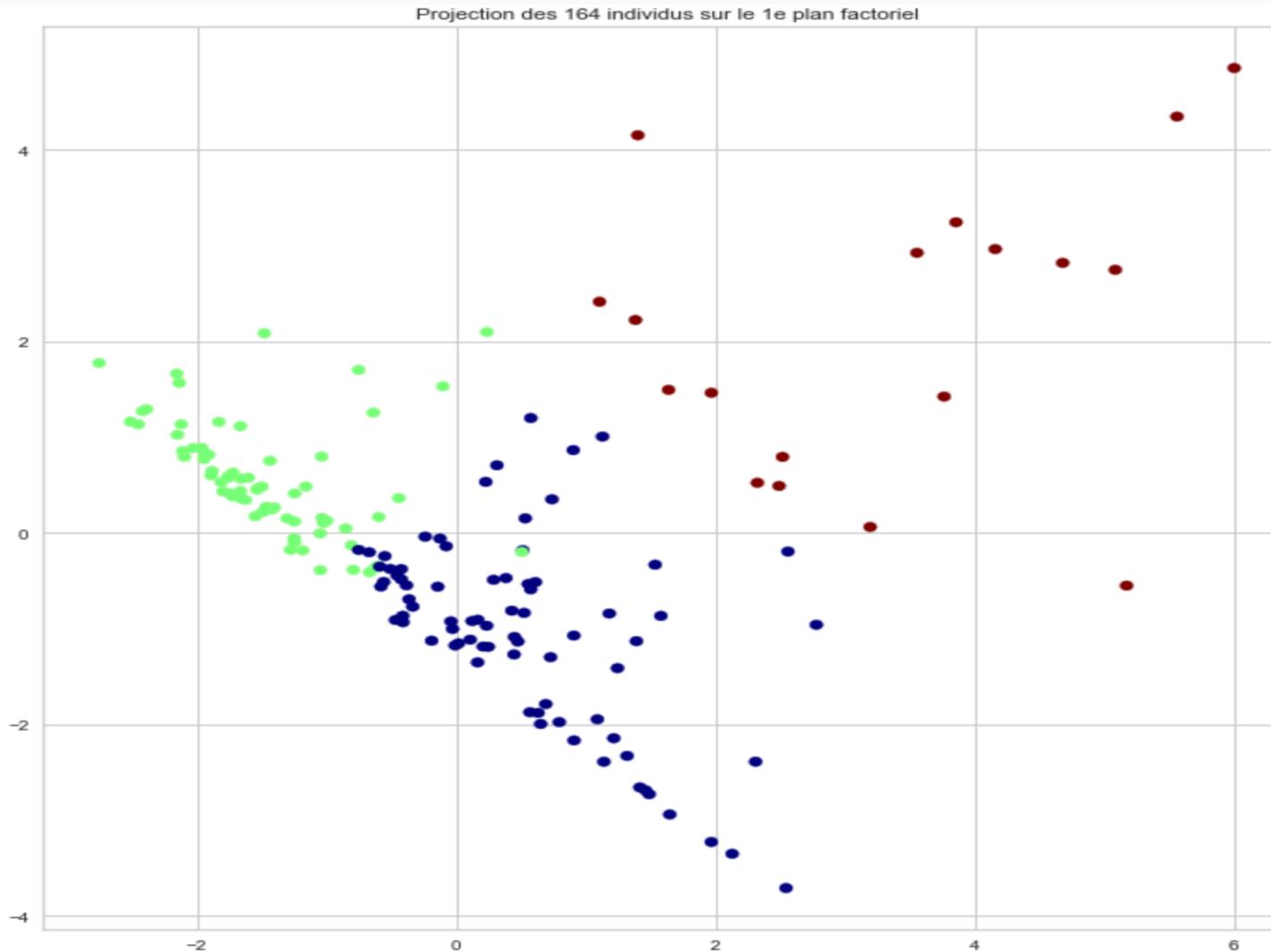
```
[0.37047608 0.24559709 0.12196883 0.09715781]  
0.8351998066998402
```

# ACP



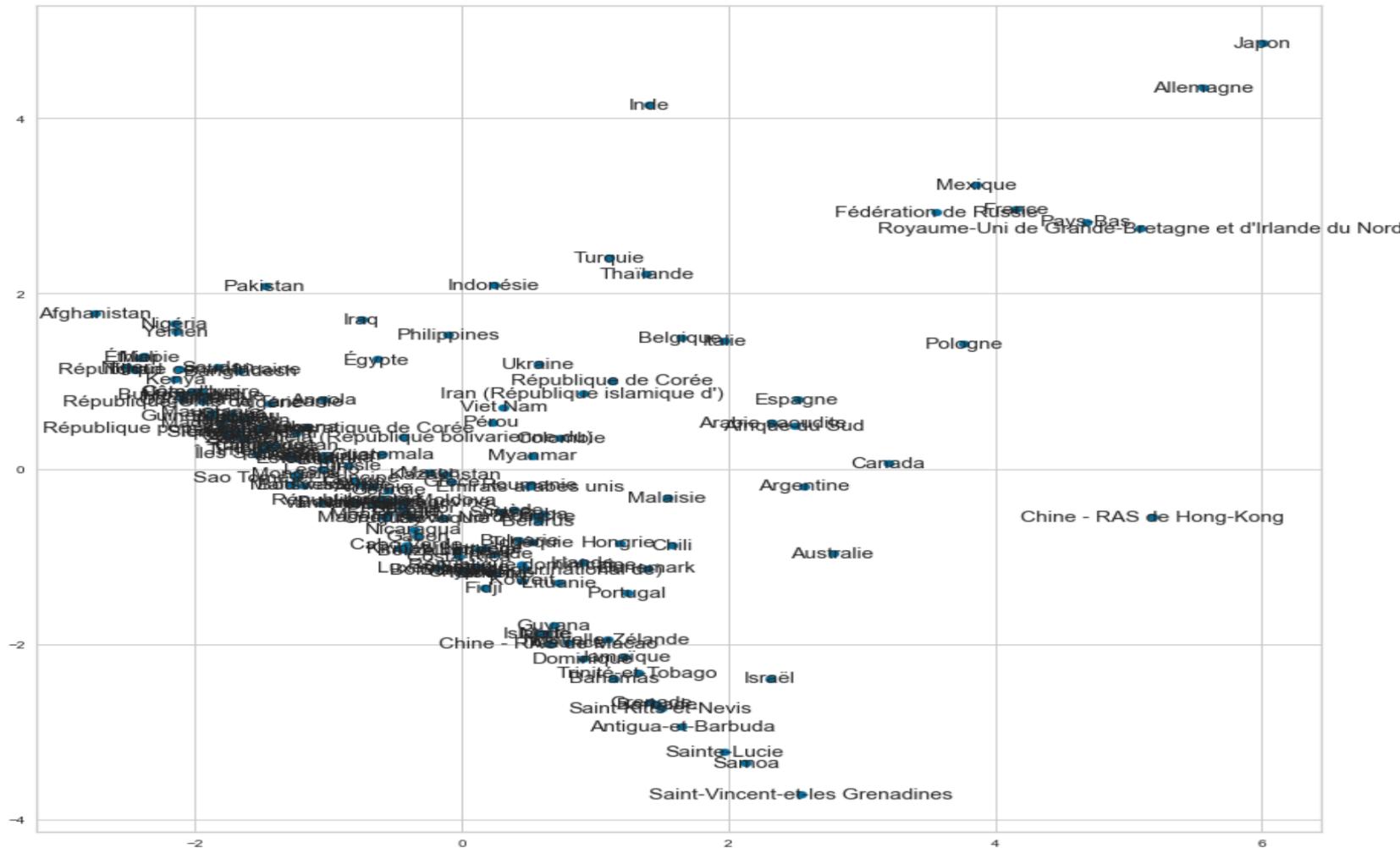
- L'inertie des axes factoriels indique d'une part si les variables sont structurées et suggère d'autre part le nombre judicieux de composantes principales à étudier.
- Les 2 premiers axes de l'ACP expriment 61,6% de l'inertie totale du jeu de données ; cela signifie que 61,6% de la variabilité totale du nuage des individus (ou des variables) est représentée dans ce plan. C'est un pourcentage assez important

# ACP



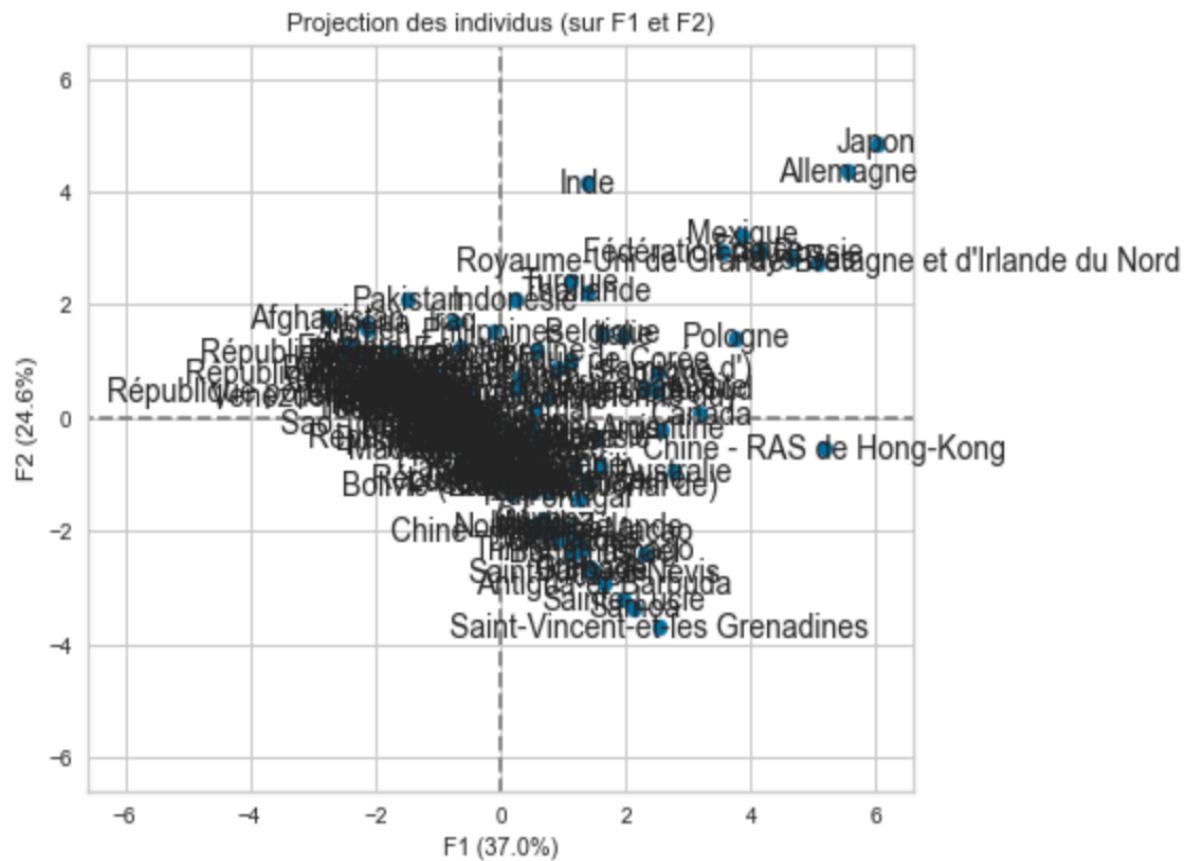
- Projection des individus sur le plan factoriel

**ACP**



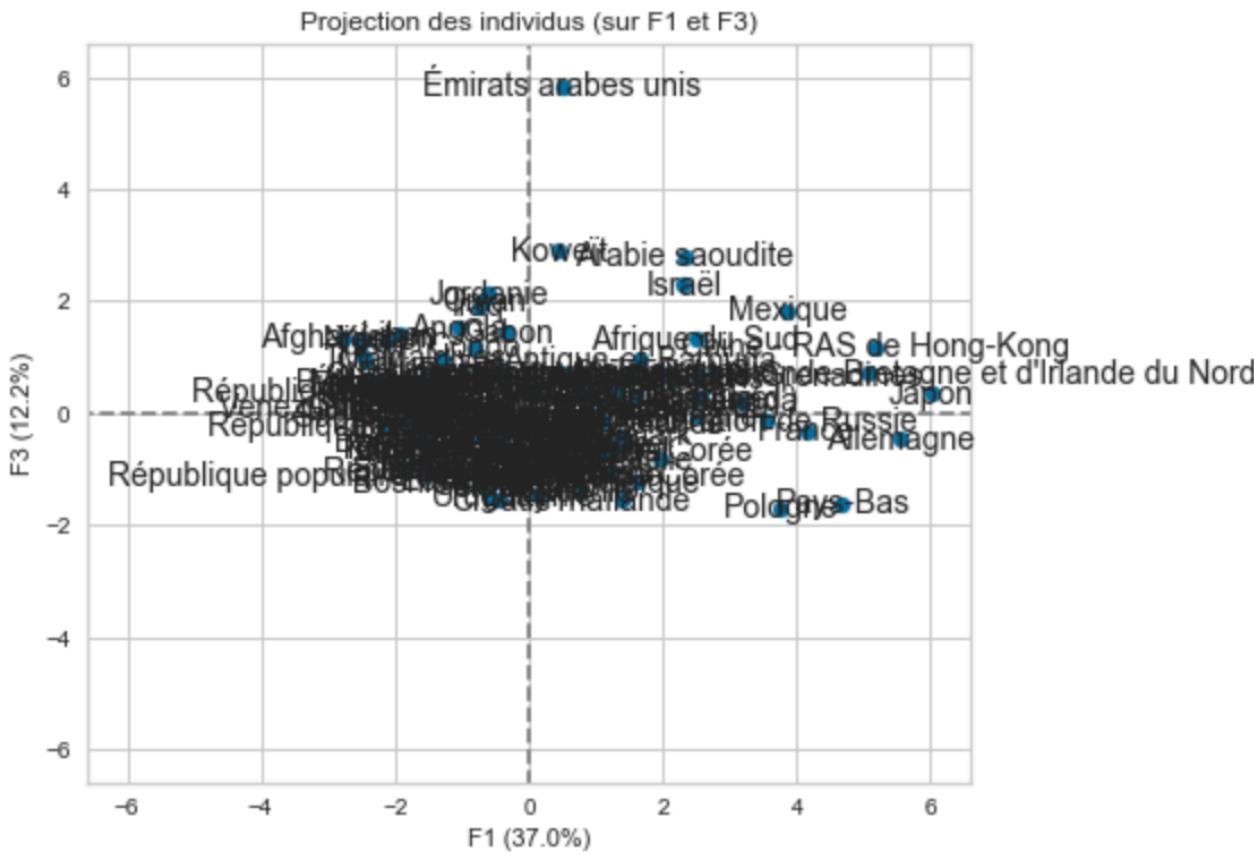
-Projection des individus  
avec les noms des pays

**ACP**



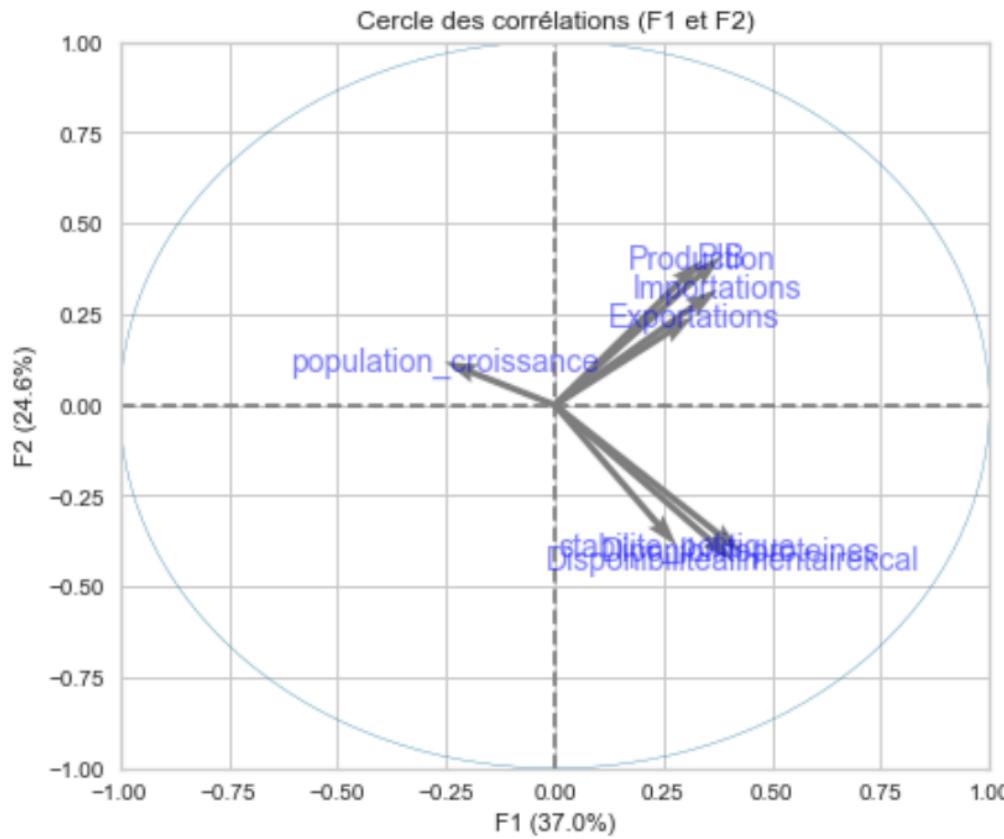
- La dimension 1 oppose des individus tels que Belgique, Allemagne, Mexique et Pologne (à droite du graphe, caractérisés par une coordonnée fortement positive sur l'axe) à des individus comme Afghanistan, Pakistan, Iraq, Philippines (à gauche du graphe, caractérisés par une coordonnée fortement négative sur l'axe).
  - La dimension 2 oppose des individus tels que Indonésie, Turquie et Thaïlande (en haut du graphe, caractérisés par une coordonnées fortement positive sur l'axe) à des individus comme saint-Vincent les grenadines, Samoa et Saint Lucie (en bas du graphe, caractérisés par une coordonnées fortement négative sur l'axe).

# ACP



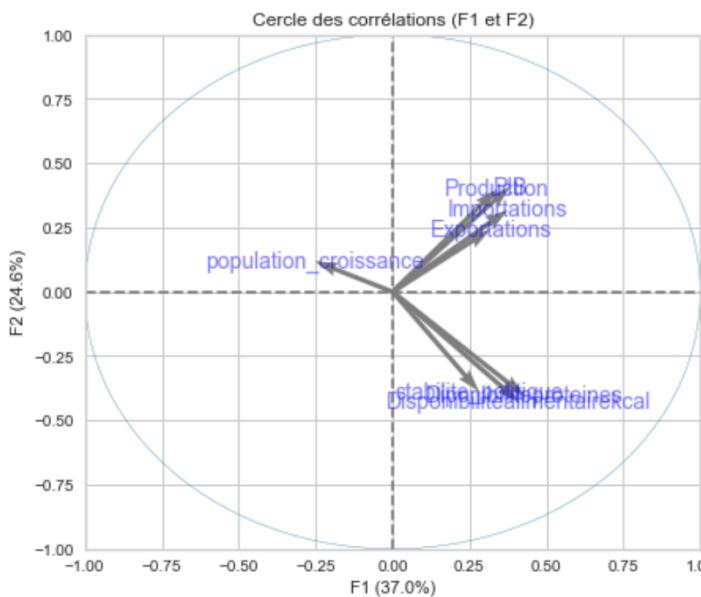
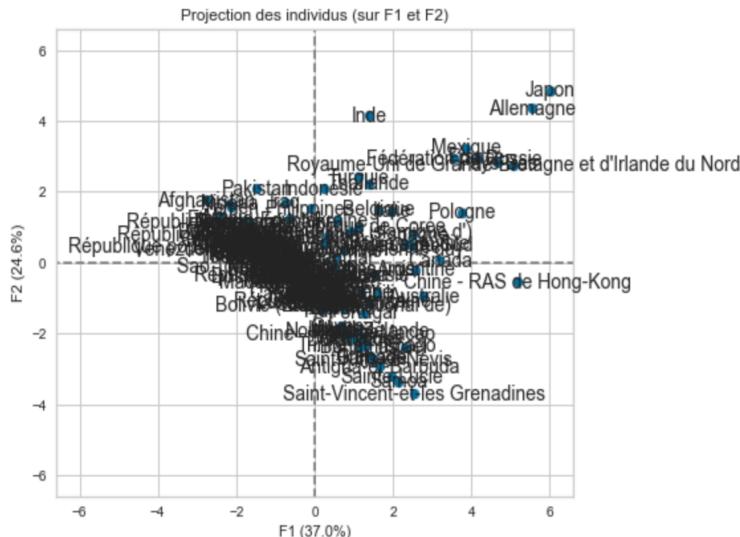
- La dimension 3 oppose des individus tels que Canada, Afrique du sud, Myanmar (à droite du graphe, caractérisés par une coordonnée fortement positive sur l'axe) à des individus comme Guatemala, Tunisie (à gauche du graphe, caractérisés par une coordonnée fortement négative sur l'axe).

# ACP



- Exportations, importations, production et PIB sont corrélés aux variables synthétiques F1, F2 (coeff corrélation entre 0,25 et 0,3)
- La variable population croissance est corrélée négativement à F1 c'est-à-dire que, quand population croissance croît, alors F1 décroît. Cela se traduit par un coefficient de corrélation proche de -1 (ici -0.25).
- Les variables les plus corrélées à F1 sont : dispo Kcal, dispo protéines, exports, imports , production ,stabilité et PIB : Elles sont corrélées positivement à F1

**ACP**



- Le groupe auquel les individus Belgique, Allemagne, Mexique et Pologne appartiennent (caractérisés par une coordonnée positive sur l'axe) partage :

- De fortes valeurs pour les variables Exportations, importation, production , Pib , dispo alimentaires kcal et protéines
  - De faibles valeurs pour la variables population croissance.

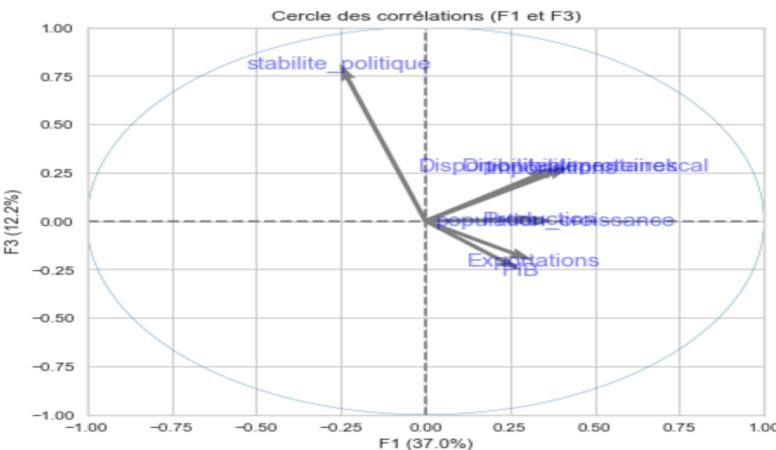
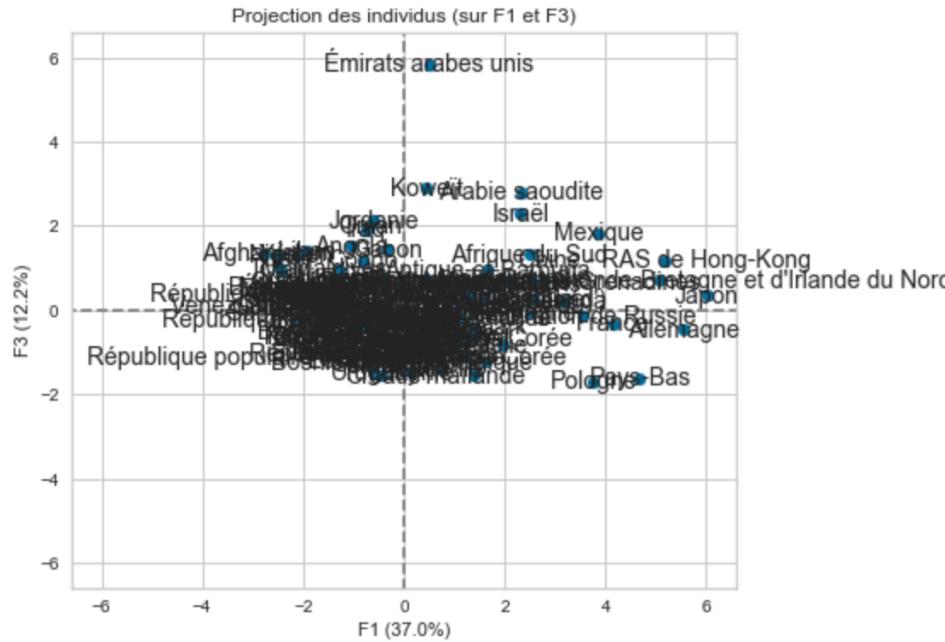
- Le groupe auquel les individus Afghanistan, Pérou, Iran, NOOL appartiennent (caractérisés par une coordonnées négative sur l'axe) partage :

- De fortes valeurs pour les variables population croissance.
  - De faibles valeurs pour les variables Exportations, importation, production , Pib , dispo alimentaires kcal et protéines

# ACP

- Le groupe auquel les individus Indonésie, Turquie et Thaïlande appartiennent (caractérisés par une coordonnée positive sur l'axe) partage :
  - De fortes valeurs pour la variable population croissance.
- Le groupe auquel les individus saint Vincent les grenadines, Samoa et Saint Lucie appartiennent (caractérisés par une coordonnées négative sur l'axe) partage :
  - De faibles valeurs pour la variable population croissance.

# ACP



- Le groupe auquel les individus Canada, Afrique du sud, Myanmar appartiennent (caractérisés par une coordonnée positive sur l'axe) partage :

- De fortes valeurs pour les variables Exportations, importation, production, Pib , dispo alimentaires kcal et protéines
- De faibles valeurs pour la stabilité politique

- Le groupe auquel les individus Guatemala, Tunisie appartiennent (caractérisés par une coordonnées négative sur l'axe) partage :

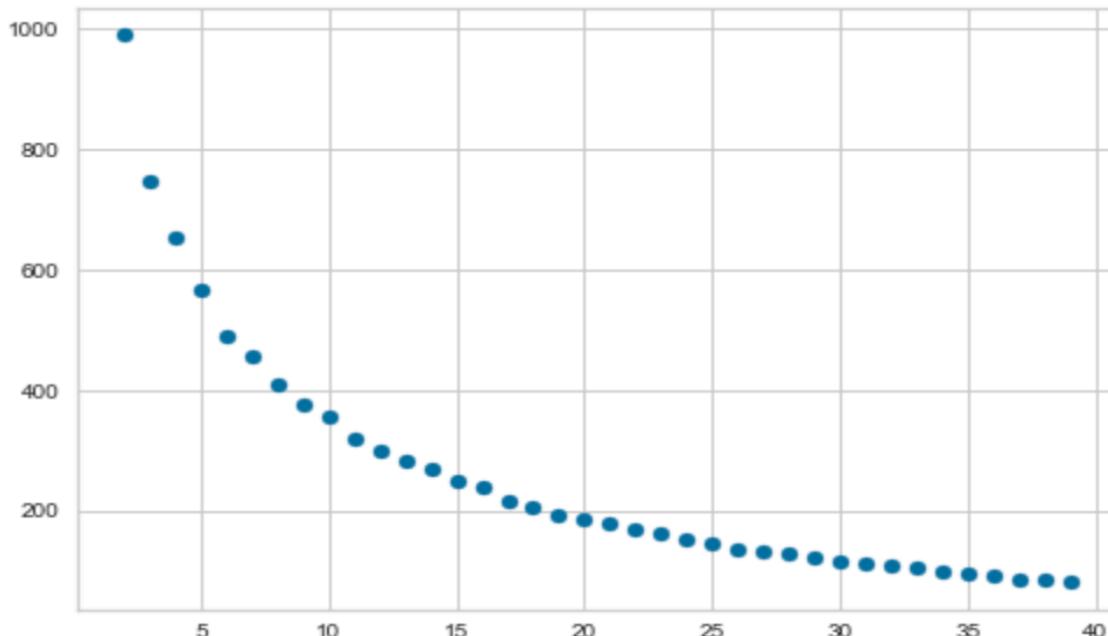
- De fortes valeurs pour les variables population croissance.
- De faibles valeurs pour les variables Exportations, importation, production , Pib , dispo alimentaires kcal et protéines

# Effectuer un clustering avec la méthode des k-means

- k-means est un algorithme non supervisé de clustering.
- Il permet de regrouper en K clusters distincts les observations du dataset. Ainsi les données similaires se retrouveront dans un même cluster.
- Pour effectuer un clustering avec la méthode des K-means, il faut tout d'abord déterminer combien de groupes on souhaite trouver : on appelle ce nombre K .
- L'objectif principal des K-means est de trouver des groupes en faisant en sorte de minimiser l'inertie intra classe.

# Effectuer un clustering avec la méthode des k-means

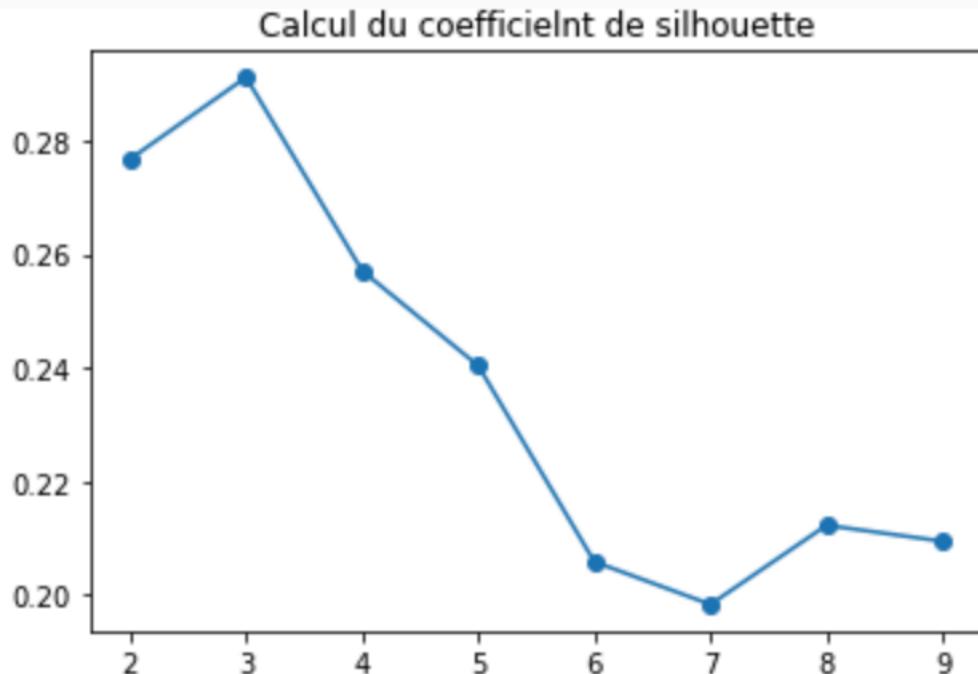
- Nous aurions pu nous servir des résultats de la CAH précédente pour choisir la valeur de  $k$ , mais nous allons plutôt observer la décroissance de l'inertie intra-classe pour déterminer la "meilleure" valeur de  $k$ . En effet, en traçant la courbe de l'inertie intra-classe en fonction de  $k$ , nous cherchons alors à identifier les étapes où l'on observe une rupture dans la décroissance de cette courbe.



- Ici, nous voyons que la partie la plus marquée du coude est entre  $k=5$  et  $k=8$ .

# Effectuer un clustering avec la méthode des k-means

- Une autre façon de quantifier à quel point un clustering répond à ces deux exigences (homogénéité et séparation) est de mesurer le coefficient de silhouette. Pour un point x donné, le coefficient de silhouette permet d'évaluer si ce point appartient au « bon » cluster .

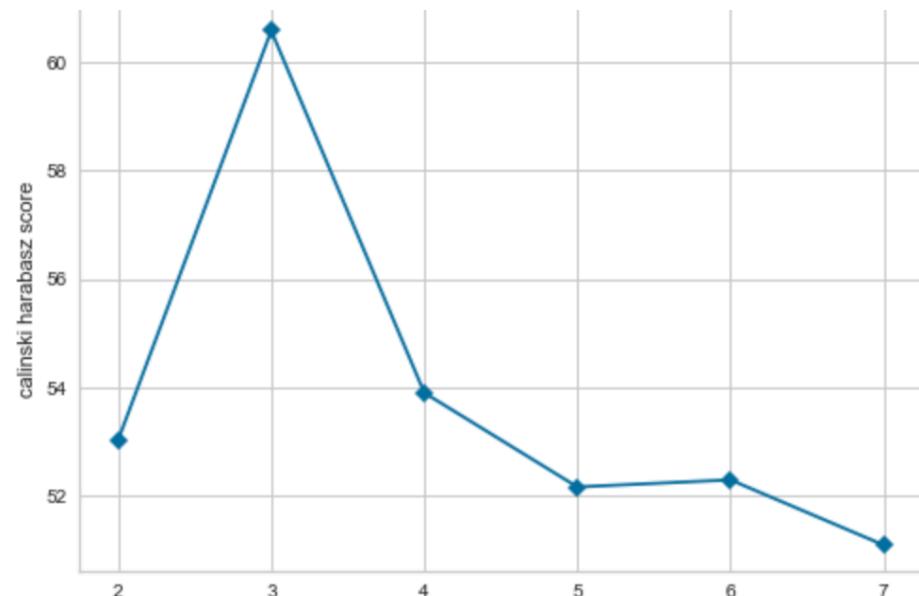


**K-MEANS, à la différence de la CAH, ne fournit pas d'outils d'aide à la détection du nombre de classes. Nous devons les programmer . On fait varier le nombre de groupes et on surveille l'évolution d'un indicateur de qualité qui est l'aptitude des individus à être plus proches de ses congénères du même groupe que des individus des autres groupes. Dans ce qui suit, on calcule la métrique « silhouette » pour différents nombres de groupes issus de la méthode des centres mobiles**

- Le coefficient de silhouette décroît avec le nombre de cluster, il remonte pour 2 et 7
- Le nombre de cluster indique 3 clusters , le nombre pour lequel le coefficient de silhouette est plus élevé

# Effectuer un clustering avec la méthode des k-means

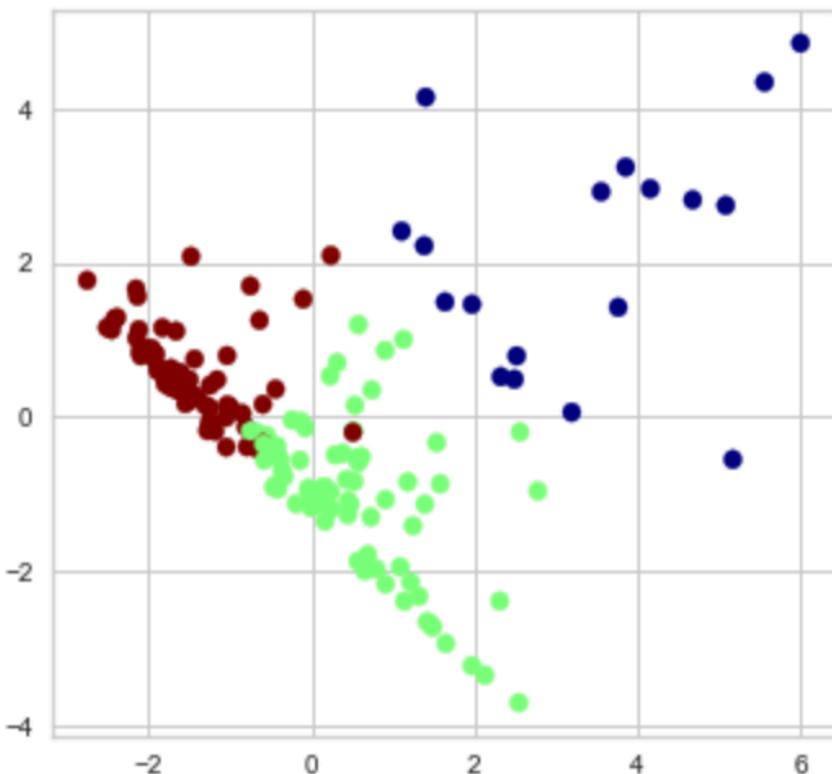
- Une autre méthode consiste à déterminer le nombre optimal de cluster dans lesquelles les données peuvent être regroupées est la méthode du coude qui est l'une des méthodes les plus populaires pour déterminer cette valeur optimale de k.



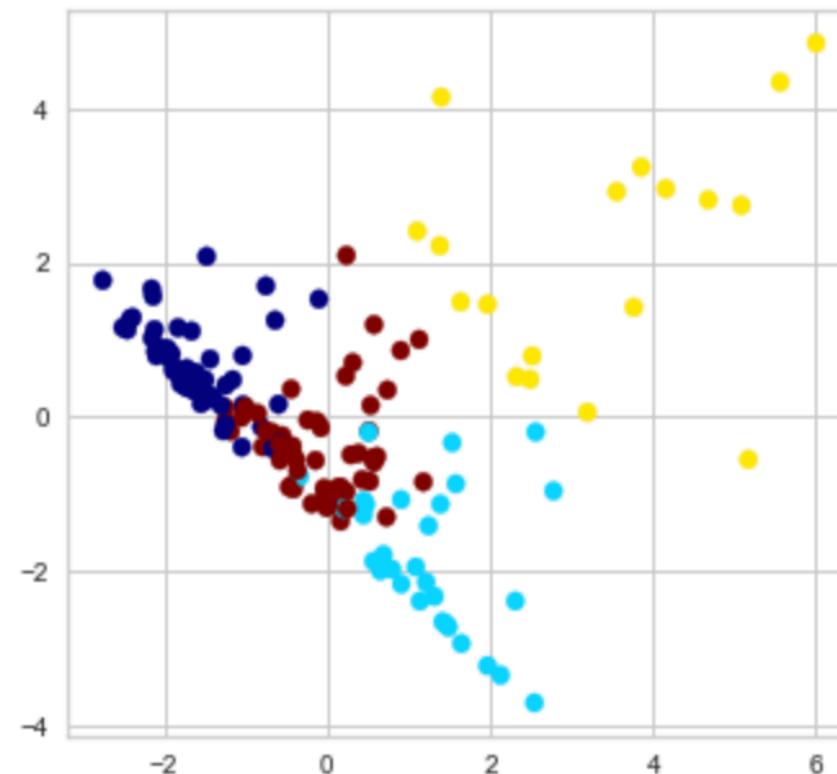
- Il faut sélectionner la valeur de k au «coude», c'est-à-dire le point après lequel l'inertie commence à diminuer de façon linéaire.

-Le point où la forme du coude est créée est 4 .

# Projection des données selon les clusters



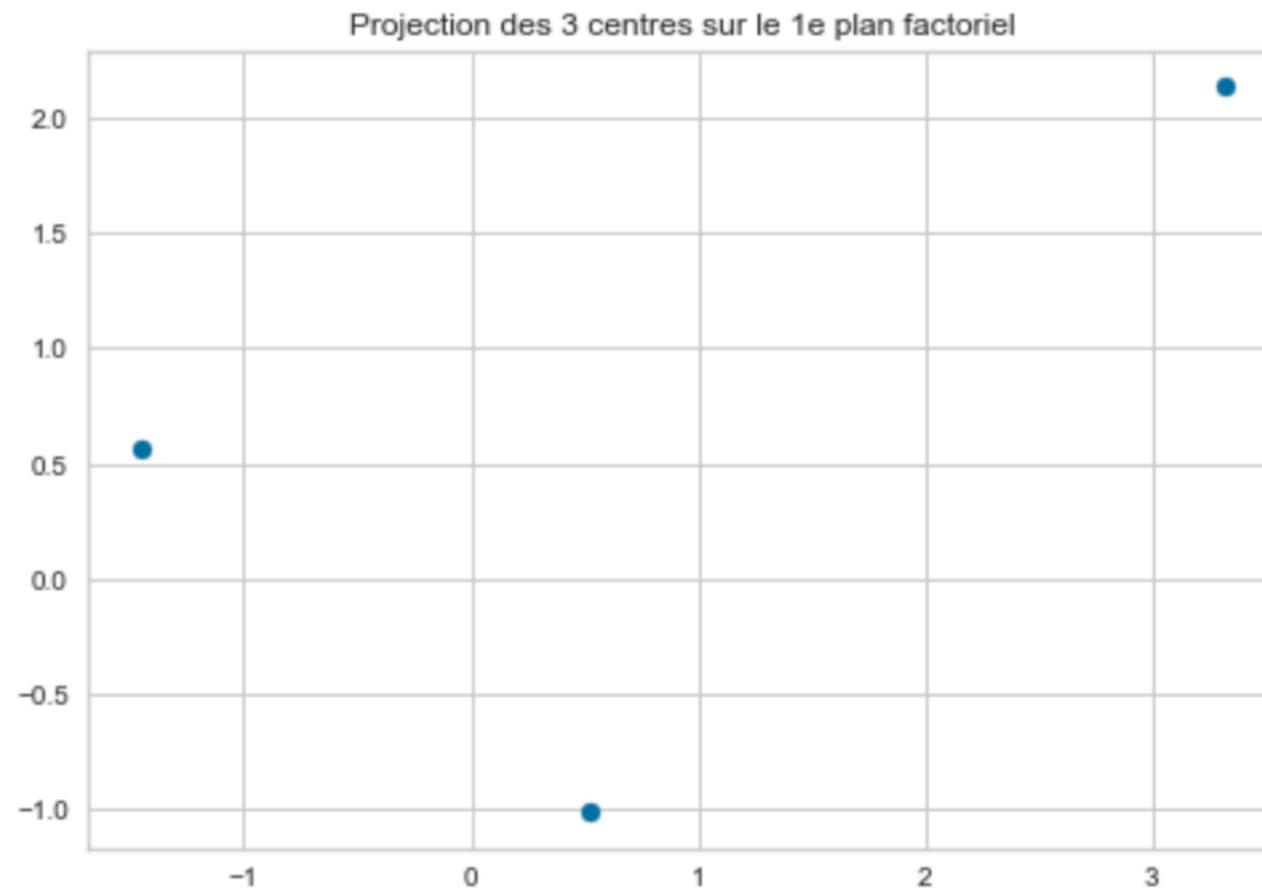
n\_clusters=3



n\_clusters=4

- La méthode de K-means qui est très sensible aux outliers (tel que les états unis et le Brésil) m'a permis de détecter des outliers et ceci en augmentant le nombre de clusters.

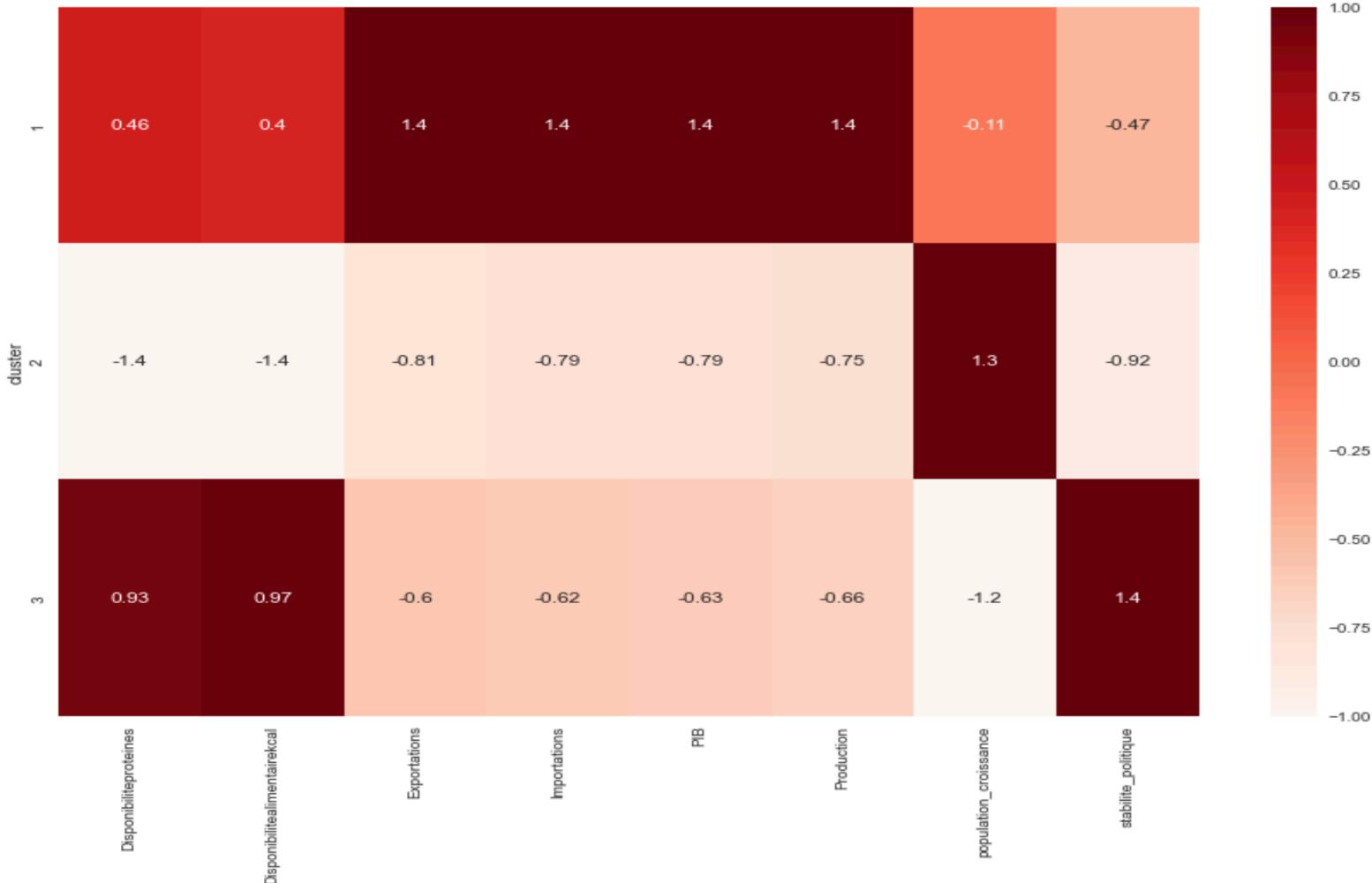
# Projection des données selon les clusters



# **Correspondance CAH – K-Means**

- Afin d'identifier les pays cibles , les correspondance CAH - K-Means indiquent que les groupes de la CAH coïncident avec les groupes des K-Means.
- Des groupes assez larges et assez bien caractérisés par les indicateurs économiques.
- Certains évidents outliers ont été exclus.

# Choix du cluster et heatmap



- **Cluster 1** : dispo Kcal et protéines assez élevés, exportations importations PIB et production fortes, faible croissance et stable politiquement.

- **Cluster 2** : faible en dispo Kcal et protéines, exportations importations PIB et production faibles, croissance élevée et moins stable politiquement.

- **Cluster 3** : dispo Kcal et protéinés très élevés, exportations importations PIB et production assez faibles, très faible croissance et très stable politiquement.

# Choix du cluster et heatmap

- Mon choix se porte sur le cluster 2 contenant des pays qui importent et exportent moins avec une faible disponibilité alimentaire mais pour autant une forte croissance et une bonne stabilité politique.
- Voici la liste des 20 pays les plus intéressés par l'exportation des poulets.

Pays	Disponibilitealimentairekcal	Diponibiliteproteines	Exportations	Importations	Production	population_croissance	PIB	stabilite_politique
Algérie	22.0	1.97	0.0	2.0	275.0	33.331859	170097.205312	-0.92
Angola	35.0	3.60	0.0	277.0	42.0	81.859749	122123.858628	-0.33
Koweït	156.0	15.87	4.0	137.0	56.0	98.330320	120687.635824	-0.05
Maroc	70.0	7.29	1.0	3.0	762.0	23.573146	109682.751915	-0.37
Équateur	83.0	6.15	0.0	0.0	340.0	32.364941	104295.862000	-0.07
Sri Lanka	26.0	2.64	2.0	0.0	192.0	12.517206	87428.117558	-0.07
Kenya	2.0	0.23	0.0	0.0	35.0	57.115088	78964.992630	-1.13
Éthiopie	0.0	0.04	0.0	1.0	14.0	60.664762	76794.516911	-1.68
Guatemala	71.0	6.11	7.0	129.0	235.0	45.183616	71611.975017	-0.65
Oman	73.0	7.38	16.0	126.0	7.0	105.731197	70598.026283	0.75
Ghana	16.0	2.26	0.0	151.0	60.0	51.053906	58994.855221	0.09
République-Unie de Tanzanie	6.0	0.63	0.0	2.0	105.0	63.169185	54723.795208	-0.56
Côte d'Ivoire	8.0	0.86	0.0	7.0	58.0	48.513905	51588.154312	-1.09
Serbie	35.0	3.50	7.0	12.0	85.0	0.000000	44179.075779	0.09
Jordanie	98.0	10.10	10.0	64.0	210.0	91.036728	41408.450704	-0.50
Azerbaïdjan	44.0	4.45	0.0	27.0	104.0	21.206868	40866.632048	-0.75
Tunisie	57.0	6.28	4.0	0.0	213.0	17.769168	39802.143071	-1.02
Paraguay	22.0	1.80	4.0	1.0	45.0	29.002474	39008.900667	0.00
Turkménistan	15.0	1.53	0.0	9.0	20.0	27.491086	37915.175476	-0.13
Cameroun	11.0	1.07	0.0	0.0	81.0	58.348331	35009.259798	-1.10

# Pays sur carte

Pays cibles



# Conclusion

- **Les clusters identifiés contiennent des pays assez larges mais bien différencier par le statut économique : le PIB, la croissance et la stabilité politique.**
- **Cette analyse permettra de proposer une première analyse des groupements de pays que l'on peut cibler pour exporter les poulets.**  
**L'étude de marché sera approfondie par la suite.**



**La poule qui chante**

# Discussion :



**FIN**

• **MERCI**