



La poule qui chante

22/04/2024

Objectif du projet et grandes étapes

Objectif

L'entreprise la Poule Qui Chante vend du poulet et cherche à s'exporter à l'international.

En tant que Data analyst nous devons identifier quelques pays potentiels dans lesquels exporter ces poulets .

Sommaire

Les Etapes

- 1ère étape : Trouver des données utiles et en libre accès
- 2ème étape : Préparer ces données pour les utiliser
- 3ème étape : Clusteriser les pays
- 4ème étape : Effectuer une Analyse en Composantes Principales (ACP)
- 5ème étape : Sélection des pays

Etape 1 : Trouver des données pertinentes

L'Organisation pour l'alimentation et l'agriculture (FAO) propose de nombreuses données en libre accès sur l'alimentation dans le monde.

J'ai choisis 4 bases de données utiles pour ce projet :

- La disponibilité alimentaire en viande de volailles en 2017 et 2021

Comprend la disponibilité en volailles, les importations, les exportations et la production, la variation de stock...

- Le PIB et le PIB/ habitant entre 2017 et 2022
- La population entre 2017 et 2022
- La stabilité politique des pays entre 2017 et 2023

Ces données nous permettent à la fois de faire un état des lieux en 2021 mais également de connaître l'évolution des tendances entre 2017 et 2021.

1) Simplification des fichiers

Dans la table des disponibilités alimentaires, on garde uniquement la dispo intérieure, la production et l'importation

On pivote chaque table pour que chacune ai ses propres variables et avec le pays en index

2) Nettoyage des tables

Lors du pivot, peu valeurs manquantes apparaissent dans la table de disponibilité alimentaire.

On fait le choix de supprimer ces lignes ce qui nous fait perdre 9 pays.

3) Création de nouveaux indicateurs

Pour chaque tables nous créons une version 2021 qui comprendra :

- Le pays
- La valeur en 2021
- L'évolution entre 2017 et 2021 de la valeur via un tx de croissance annuel moyen (Cagr)

Pour la table disponibilité alimentaire, on calcul en plus les ratios :

- Part de l'importation dans la disponibilité intérieure
- Part de la production dans la disponibilité intérieure

4) Assemblage des fichiers

On joint nos tables version 2021 avec les nouveaux indicateurs entre elles.

On effectue une jointure en prenant comme référence la disponibilité alimentaire pour garder toutes les infos de cette table.

On utilise comme clé primaire les pays.

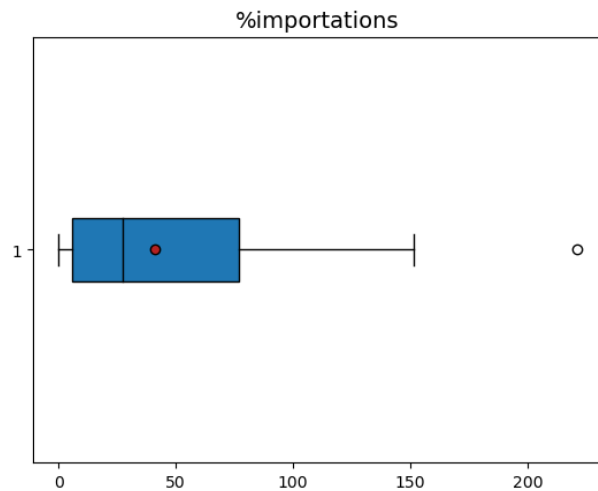
5) Nettoyage du fichier

Lors de la jointure, il manque certaines données pour 6 pays et notamment le cas de la Chine, Macao, Hong Kong...

On choisit de garder uniquement la Chine.

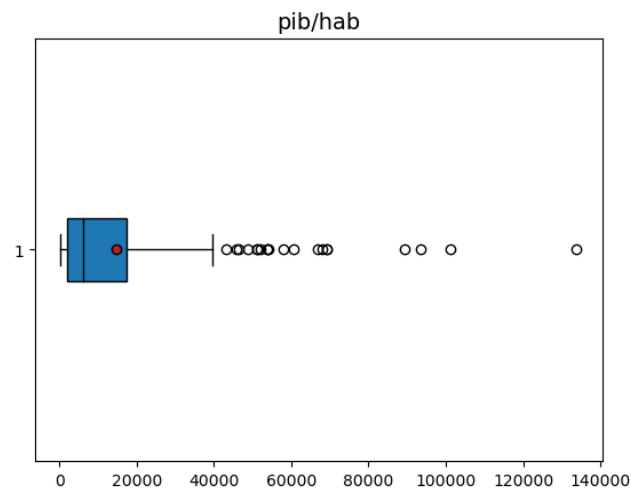
Pour la stabilité politique on choisit de remplacer la valeurs des outre-mers française par la stabilité politique de la France.

6) Analyse exploratoire univariée

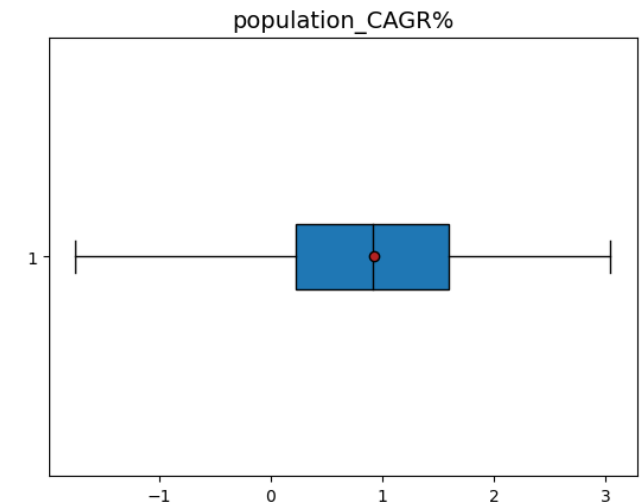


Les pays à taux d'importation élevé par rapport à leur disponibilité intérieure ne sont pas aberrants car ils sont également pour la plupart fort exportateur ou peu producteur

La moitié des pays ont un tx d'importation **>25%**



Les pays les plus riches ont un pib/hab très élevés par rapport aux autres pays

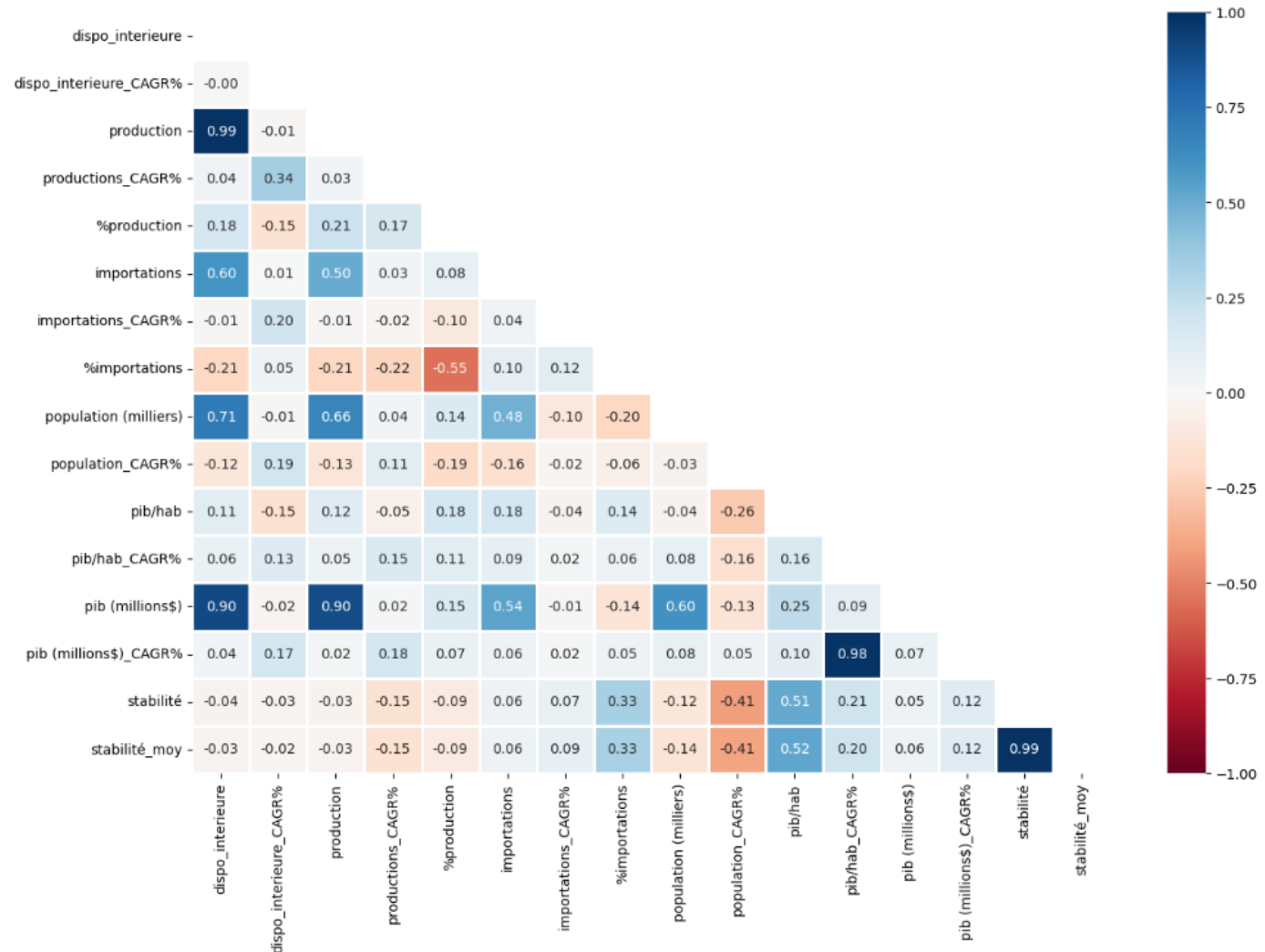


La population des pays évolue entre 2017 et 2021 de -2% à +3%

La moyenne de croissance au niveau mondiale est de 1% par an

7) Analyse exploratoire bivariée

Matrice de corrélations entre les variables



Plus le chiffre tend vers 1 (bleu foncé) plus la corrélation positive entre les 2 variables est forte. Inversement plus le chiffre tend vers -1 (rouge foncé) la corrélation est négative

Exemples :

- Production et disponibilité intérieure sont très liés.
- Pas de relation entre pib/hab et disponibilité intérieure
- La stabilité moyenne est très proche de la stabilité en 2021

Fichier finale à utiliser

Echantillons du fichier

	pays	dispo_interieure	dispo_interieure_CAGR%	production	productions_CAGR%	%production	importations	importations_CAGR%	%importations
139	Sainte-Lucie	11.0	0.00	1.0	-12.94	9.09	10.0	0.00	90.91
144	Sierra Leone	61.0	16.03	23.0	6.23	37.70	38.0	22.10	62.30
136	République-Unie de Tanzanie	96.0	7.14	88.0	6.25	91.67	9.0	35.10	9.38
26	Bénin	133.0	1.41	16.0	2.71	12.03	117.0	0.70	87.97
141	Sao Tomé-et-Principe	5.0	20.11	1.0	0.00	20.00	4.0	14.87	80.00
123	Polynésie française	17.0	1.22	1.0	0.00	5.88	17.0	2.53	100.00
78	Jordanie	272.0	1.95	205.0	1.21	75.37	74.0	2.95	27.21
17	Belize	11.0	-10.35	19.0	-1.02	172.73	0.0	0.00	0.00
115	Ouzbékistan	221.0	26.96	124.0	16.41	56.11	97.0	64.72	43.89
161	Turkménistan	57.0	12.95	23.0	1.84	40.35	34.0	27.73	59.65

Les variables présentent dans le fichier

Data columns (total 17 columns):				
#	Column	Non-Null	Count	Dtype
0	pays	173	non-null	object
1	dispo_interieure	173	non-null	float64
2	dispo_interieure_CAGR%	173	non-null	float64
3	production	173	non-null	float64
4	productions_CAGR%	173	non-null	float64
5	%production	173	non-null	float64
6	importations	173	non-null	float64
7	importations_CAGR%	173	non-null	float64
8	%importations	173	non-null	float64
9	population (milliers)	173	non-null	float64
10	population_CAGR%	173	non-null	float64
11	pib/hab	173	non-null	float64
12	pib/hab_CAGR%	173	non-null	float64
13	pib (millions\$)	173	non-null	float64
14	pib (millions\$)_CAGR%	173	non-null	float64
15	stabilité	173	non-null	float64
16	stabilité_moy	173	non-null	float64

Notre dataframe final couvre
99,17% de la population mondiale

Etape 3 : Clustering avec Kmeans

2) Algorithme K-means : définition nombre de cluster

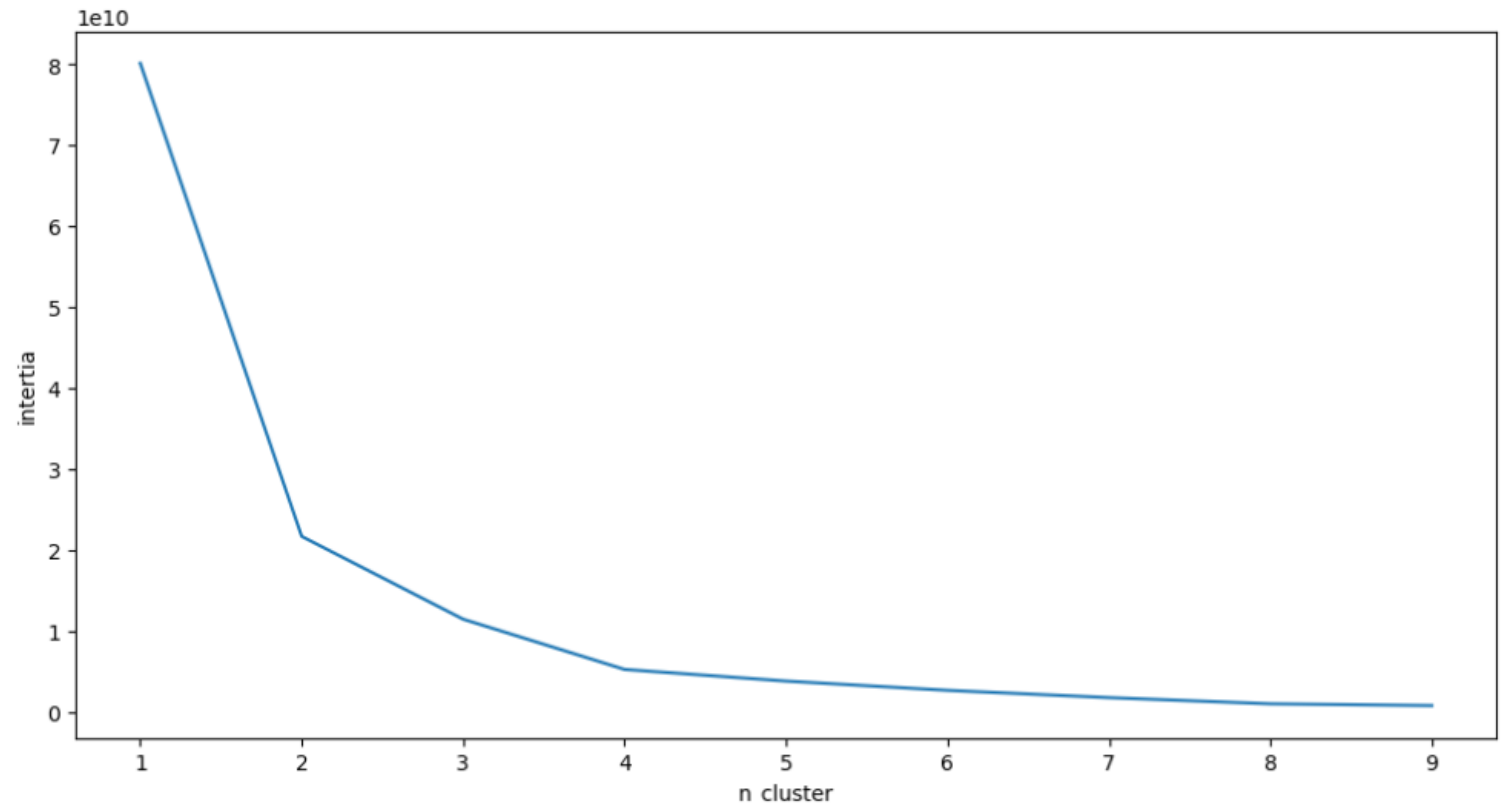
Objectif K-means :

L'algorithme du k-means est un algorithme non déterministe. On doit lui spécifier le nombre de cluster que l'on souhaite. Pour définir un nombre potentiel de cluster, on utilise la méthode « du coude »

Analyse elbow method :

Dans notre méthode du "coude", on cherche une "cassure" dans la courbe liant la variance intraclasse au nombre de clusters.

Ici nous allons utiliser la cassure au niveau de 4 clusters pour être le plus précis possible.



Etape 3 : Clustering avec Kmeans

3) Algorithme K-means : les clusters

Nos pays sont répartis en 4 cluster nommés a,b,c ou d.

cluster	
a	122
b	17
c	4
d	30

Après analyse, il y a très peu de différences entre le clustering kmeans et le clustering CAH

Exemples de pays pour chaque cluster

	%production	pib/hab	%importations	stabilité	import/pop	cluster
pays						
Vanuatu	14.29	3072.572221	85.71	0.79	1.880070	a
Malaisie	97.54	11100.943725	5.82	0.14	0.288915	a
Bénin	12.03	1360.911474	87.97	-0.30	0.900215	a
Équateur	100.00	5965.132871	0.23	-0.27	0.005619	a

	%production	pib/hab	%importations	stabilité	import/pop	cluster
pays						
Canada	98.27	52112.061496	12.89	0.94	0.526798	b
Japon	75.53	39649.927505	35.22	1.03	0.911626	b
Australie	103.59	66915.859201	0.23	0.85	0.011574	b
Nouvelle-Zélande	110.68	48823.507780	0.49	1.44	0.019494	b

	%production	pib/hab	%importations	stabilité	import/pop	cluster
pays						
Suisse	71.24	93525.366191	30.07	1.13	0.529258	c
Norvège	96.64	89241.713938	3.36	1.10	0.074033	c
Luxembourg	0.00	133745.401502	116.67	1.21	2.189823	c
Irlande	114.79	101108.989159	57.75	0.86	1.644431	c
pays						
Trinité-et-Tobago	71.21	16032.502768	27.27	0.15	1.179815	d
Nouvelle-Calédonie	10.00	34994.265834	90.00	0.37	3.127172	d
Italie	108.48	35578.851067	7.05	0.58	0.150235	d
Koweït	29.95	32150.178824	72.81	0.30	3.717547	d

Etape 4 : Analyse en composantes principales

1) l'ACP : nombre de composantes identifiées

Objectif ACP :

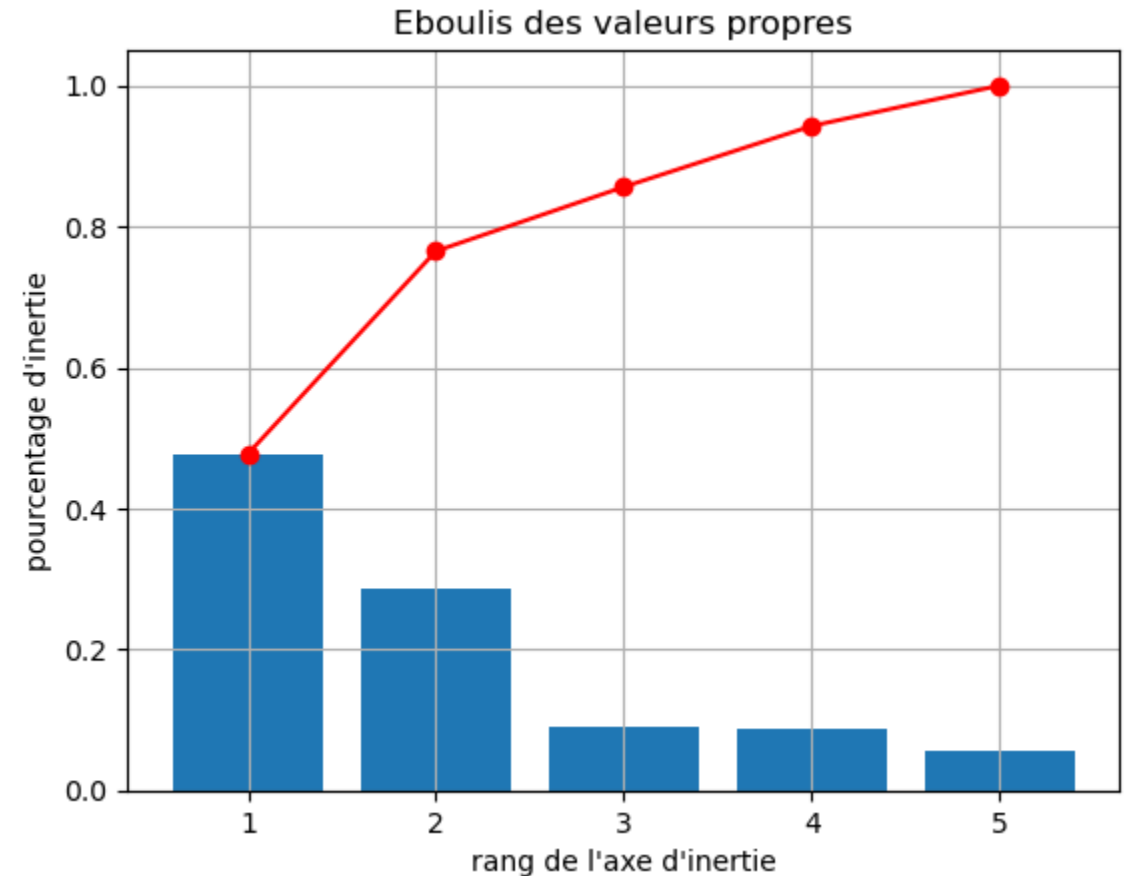
L'ACP permet de créer des dimensions synthétiques calculées à partir des dimensions initiales et ainsi de réduire le nombre de dimensions en perdant le moins d'information possible.

L'A.C.P. permet d'explorer les liaisons entre variables et les ressemblances entre individus.

Pour choisir le nombre de composantes pertinentes, on crée un éboulis de valeur propre

Analyse de l'éboulis :

Dans notre cas, notre éboulis de valeur propre nous montre que 2 variables suffisent à interpréter près de 80% de l'inertie totale



Etape 4 : Analyse en composantes principales

2) l'ACP : projection des composantes

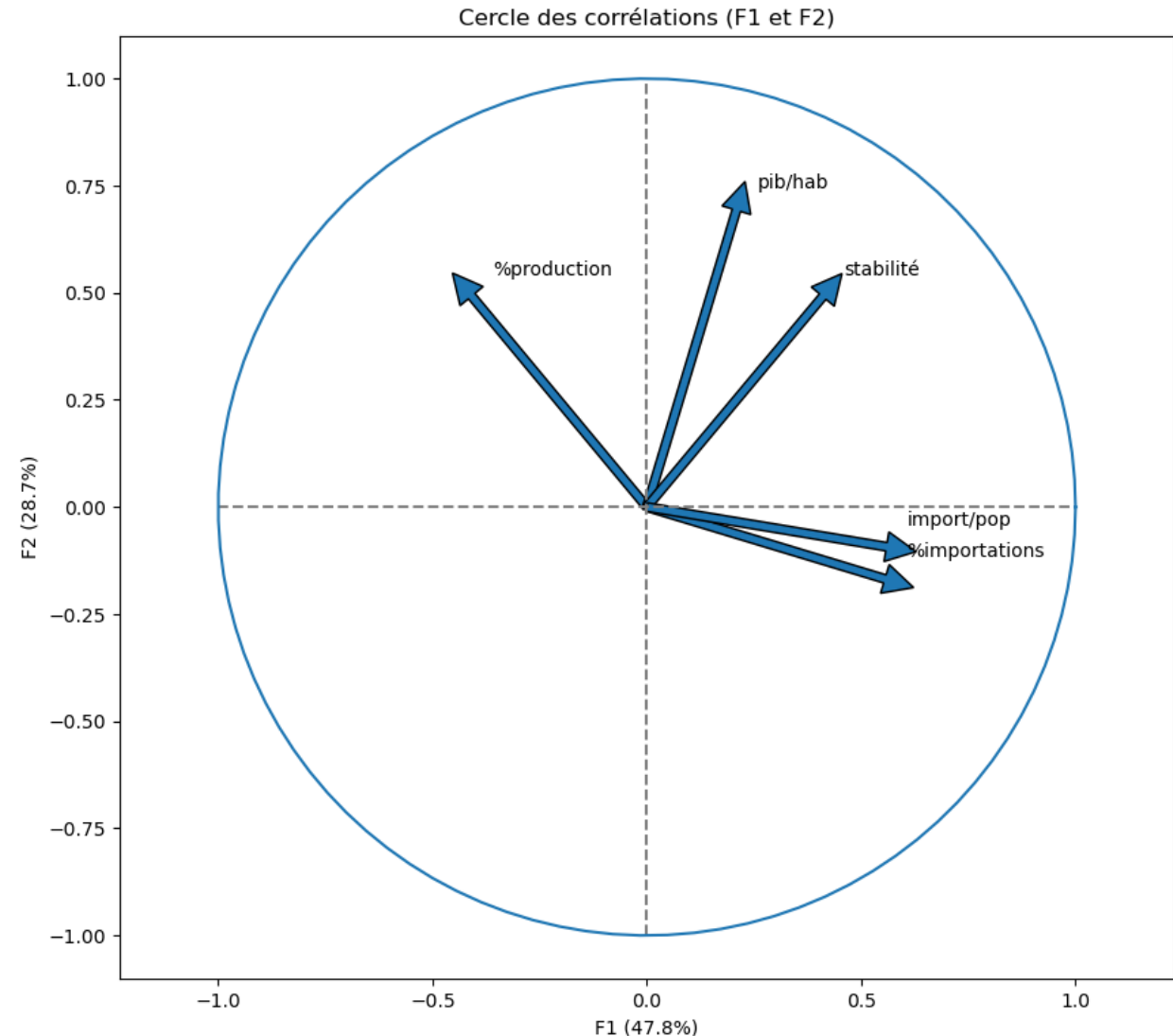
Cercle des corrélations

Le cercle des corrélations sert à projeter les variables sur un plan factoriel et permet d'étudier les liaisons entre les variables.

On cherche à savoir s'il y a des groupes de variables qui sont fortement corrélées entre elles (deux à deux).

Analyse du cercle

- L'axe F1 semble regrouper les variables liées à la forte importation.
- L'axe F2 semble lié fortement à la richesse du pays par habitants



Etape 4 : Analyse en composantes principales

3) Projections des clusters sur axes F1 et F2

On cherche à représenter nos individus sur les nouvelles dimensions F1 et F2.

Dans notre cas, nous choisissons de représenter, par la couleur, les clusters de nos individus

Les groupes d'individus qui nous intéressent pour notre analyse, sont les pays qui sont situés en haut à droite de la projection :

- Ils sont importateurs
- Ce sont les pays plutôt riches et stables

Les pays sont principalement dans les clusters rouge, orange et vert.

Ce sont ces groupes de pays qu'il faut étudier de plus près



Etape 5 : Les pays potentiels

Nous allons garder les pays des clusters précédents et affiner la sélection des pays avec le tx d'importations, l'évolution des importations et de la production.

A ce stade, nous avons 51 pays potentiels.

1) Tx d'importations

Critère : Dans l'échantillon des 3 clusters (51 pays) nous choisissons qui importe au moins 15% de leur disponibilité intérieure. Ce qui correspond au $\frac{3}{4}$ des pays

2) Evolution des importations







Critère : Dans les 38 pays restants, nous gardons les pays qui ont une évolution des importations positives entre 2017 et 2021

3) Evolution de la production

Critère : On réduit encore cette liste de moitié avec les pays qui ont une évolution de la production inférieure à la médiane

Conclusion

Nous obtenons une liste de 9 pays dont 6 sont en Europe (excl. France)

Pays	Importations <i>(milliers de t)</i>	Evol. Import. <i>(2017-2021 cagr %)</i>
 Lituanie	54	4,18
 Luxembourg	14	3,13
 Portugal	763	0,48
 Slovaquie	68	1,54
 Slovenie	23	2,83
 Pays-Bas	763	4,65

Après avoir ciblé ce groupe d'entreprises, il est nécessaire de faire une revue stratégique par pays pour identifier les les plus pertinents