PROJET 9

PRODUIRE UNE ÉTUDE DE MARCHE AVEC PYTHON

MISSION

EFFECTUER UNE PREMIÈRE ANALYSE D'UN GROUPEMENT DE PAYS CIBLES POUR UNE EXPORTATION DE POULET¶

PLAN DE TRAVAIL

I. **IMPORTATION DES LIBRAIRIES**

II. **IMPORTATION DES DONNÉES**

- Préparation et nettoyage des données
- Jointures des datasets
- Les Outliers

III. **ANALYSE DES COMPOSANTES PRINCIPALES (ACP)**

IV. **MÉTHODE DE CLASSIFICATION ASCENDANTE HIÉRARCHIQUE (CAH)**

V. **MÉTHODE K-MEANS**

VI. **ANALYSE DES GROUPES**

VII. **EXPLORATION DU CLUSTER SÉLECTIONNÉ**

VIII. **CONCLUSION**

OBJECTIF DE LA MISSION

- Dans le cadre de son développement à l'international l'entreprise française d'agroalimentaire « La poule qui chante » a besoin :
  - D'une première analyse sur un groupement de pays cibles pour exporter du poulet
  - L'étude du marché sera approfondie à l'issue de cette première analyse
- Les données de la FAO (Food and Agriculture Organization) seront utilisées dans cette étude
- Le langage utilisé est python

L'objectif final de cette étude est de mettre en évidence un groupe de pays homogène et répondant aux mêmes caractéristiques en terme de besoins d'importation de poulet

I. IMPORTATION DES LIBRAIRIES

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import scipy.stats as st
from sklearn import preprocessing
from sklearn import decomposition
from sklearn import cluster, metrics
from sklearn.cluster import KMeans
from scipy.cluster.hierarchy import linkage, fcluster, dendrogram
from matplotlib.collections import LineCollection
from sklearn.metrics import silhouette_score
pd.options.mode.chained_assignment = None
import warnings
warnings.filterwarnings("ignore")
```

Librairies utilisées :

- pandas
- numpy
- Seaborn
- Matplotlib
- Scipy
- sklearn

II. IMPORTATION DES DONNÉES

Les données utilisées :

- Dataset Population (2000-2018)
- Dataset Dispinibilité alimentaire (année 2017)
- Dataset PIB (croissance anuelle par pays année 2017)

Dataset Stabilité politique (2021)

Nouvelles variables créées pour le besoin de l'analyse :

- Croissance démographique (%) sur la période 2012-2017
- Le taux de dépendance à l'importation (TDI %) = (Importation ÷ Disponibilité intérieure) x 100
- Taux d'auto-suffisance (TAS %) = (Production ÷ Disponibilité intérieure) x 100

Valeurs manquantes :

Les valeurs manquantes ont été remplacées par la moyenne de la variable concernée :

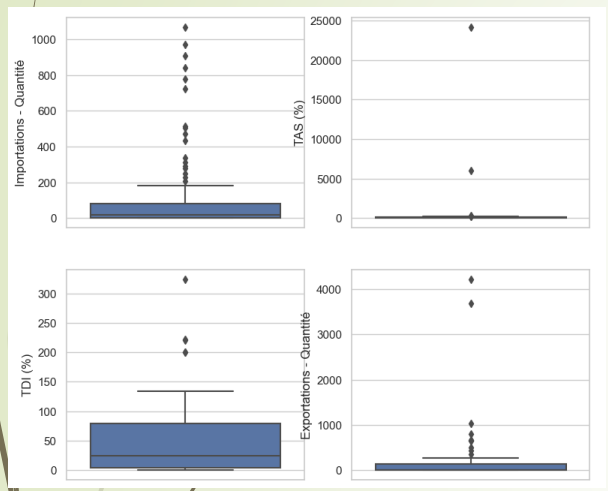
- PIB : 8 valeurs
- Disponibilité intérieur : 2 valeurs

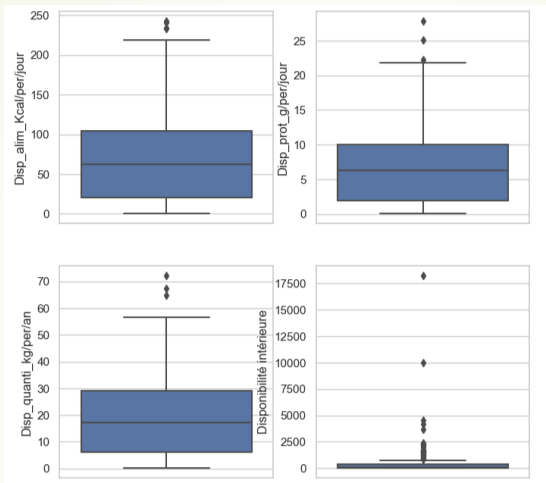
II. IMPORTATION DES DONNÉES (suite)

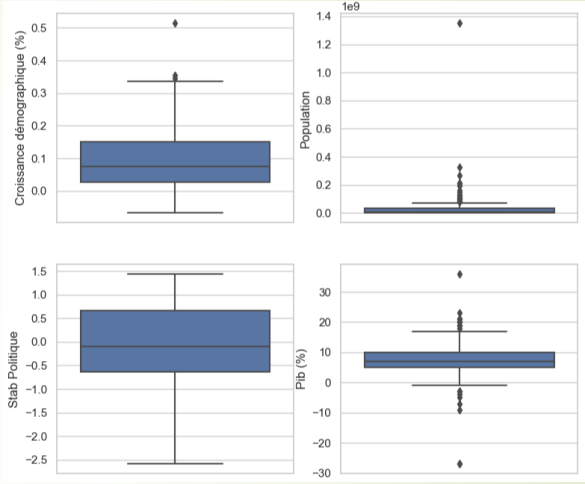
Après jointure des 4 datasets, « population », « disponibilité alimentaire », « Pib » et « stabilité politique », 8 variables finales seront utilisées pour cette analyse

	Zone	Croissance démographique (%)	Population	Stab Politique	Pib (%)	Disp_prot_g/per/jour	Disponibilité intérieure	TAS (%)	TDI (%)
0	Afghanistan	0.247101	37171921.0	-2.53	3.0	0.54	57.0	49.122807	50.877193
1	Afrique du Sud	0.092891	57792518.0	-0.71	17.0	14.11	2118.0	78.706327	24.268178
2	Albanie	-0.009564	2882740.0	0.11	9.0	6.26	47.0	27.659574	80.851064
3	Algérie	0.129027	42228408.0	-0.88	6.0	1.97	277.0	99.277978	0.722022
4	Allemagne	0.020710	83124418.0	0.76	6.0	7.96	1739.0	87.061530	48.418631

Les Outliers







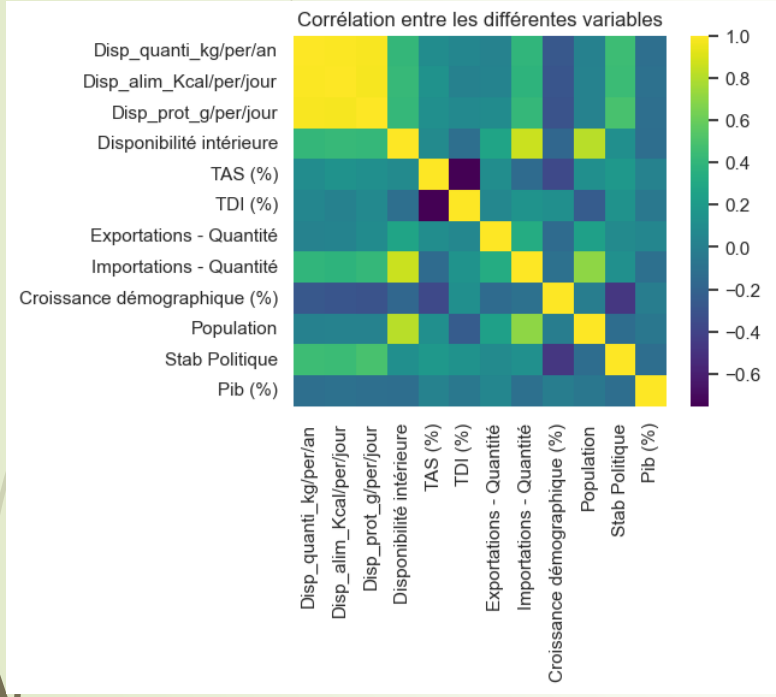
Conclusion : Des outliers présents dans toutes les variables excepté la stabilité politique

Les Outliers (suite)

Après analyse des valeurs extrêmes nous décidons de supprimer plusieurs pays pour différentes raisons :

- TDI élevé : Hong-Kong (sous emprise chinoise), Belgique (TAS trop élevé), Ile Salomon (Pop trop faible)
- TAS élevé : Belgique, Israël, Djibouti, Maldives, Pologne, Thaïlande.
- Dispo Quanti Kg élevé : Saint-Vincent et Grenadines, Samoa.
- Dispo Kcal élevé : Sainte-Lucie, Antigua-et-Barbuda.
- Croissance démographique élevé : Maldives (Population trop faible)
- Population très élevé (80M ou +) avec TAS à 100% et TDI à 0% : Nigéria, Bangladesh, Ethiopie.
- Dispo Intérieure trop élevé associé à un TAS élevé (95% ou +) et TDI faible (5% ou moins) : Etats-Unis, Brésil, Inde, Russie, Indonésie, Iran, Argentine, Turquie, Myanmar (Birmanie), Colombie, Malaisie, Pakistan, Pérou, Australie.
- Pays en guerre : Ukraine
- PIB fortement négatif avec un TAS élevé et TDI faible ou population trop faible : Ouzbékistan, Tunisie, Dominique.
- Pays inférieur à 2M d'habitants (potentiel commerciale trop faible) : Saint-Kitts-et-Nevis, Grenade, Kiribati, Sao Tomé-et-Principe, Barbade, Vanuatu, Islande, Belize, Bahamas, Malte, Cabo Verde, Suriname, Luxembourg, Monténégro, Chine - RAS de Macao, Guyana, Fidji, Djibouti, Eswatini, Chypre, Maurice, Timor-Leste, Estonie, Trinité-et-Tobago, Guinée-Bissau, Lettonie,
- Pays ou nous sommes déjà implanté : France

II. IMPORTATION DES DONNÉES (suite)

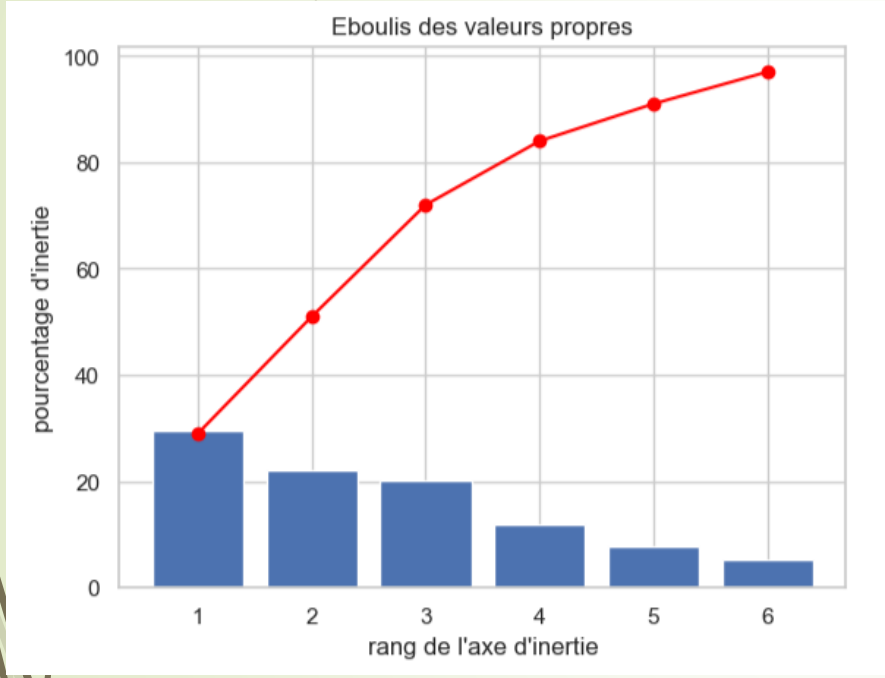


Corrélations entre les variables :

- Les disponibilités sont très corrélées entre elles
- Le TDI est négativement corrélé au taux d'auto-suffisance (TAS)
- Le TDI est positivement (faiblement) corrélé aux disponibilités,
- Les pays dépendants à l'importation (TDI) sont ceux qui ont un taux d'auto-suffisance (TAS) le plus faible
- Les pays avec un TDI important ont des disponibilités relativement faibles

III. ANALYSE DES COMPOSANTES PRINCIPALES

Évaluation quantitative des informations apportées par chaque composante



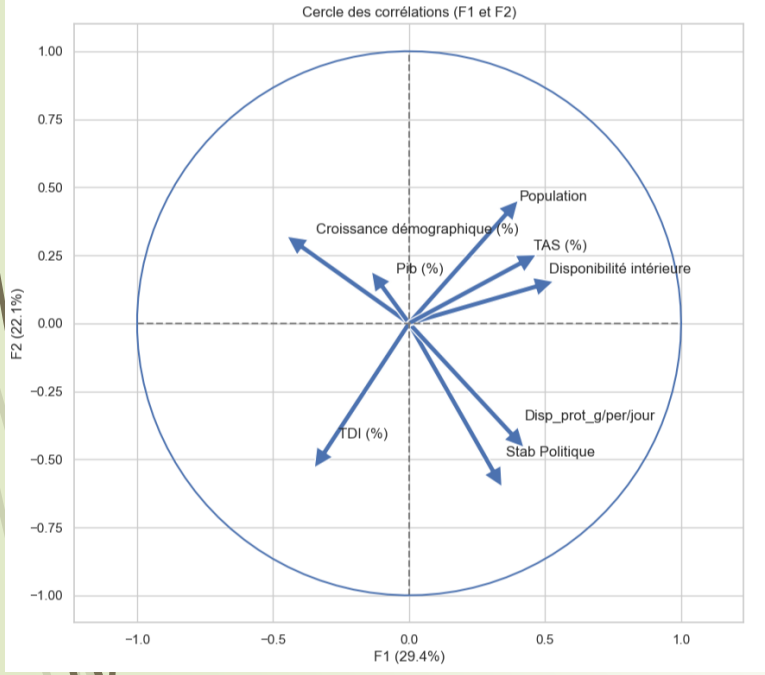
- Nous avons dans notre cas l'inertie totale répartie inéquitablement sur 6 axes :

- **Axe 1** : 29,4 % de l'inertie totale
- **Axe 2** : 22,0 % de l'inertie totale
- **Axe 3** : 20,1 % de l'inertie totale
- **Axe 4** : 12 % de l'inertie totale
- **Axe 5** : 7,9 % de l'inertie totale
- **Axe 6** : 5,2 % de l'inertie totale

A partir de 4 rangs nous avec un pourcentage d'inertie de 84%, nous nous concentrerons donc sur les 4 premières composantes.

III. ANALYSE DES COMPOSANTES PRINCIPALES

CERCLES DES CORRÉLATIONS



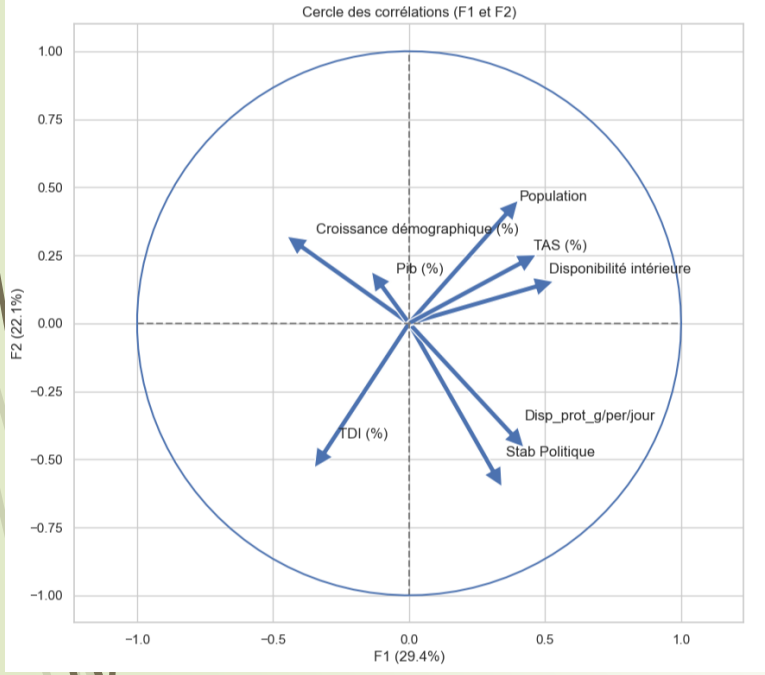
CORRÉLATIONS AVEC LES CP

F1 : La disponibilité intérieur a la plus grande contribution positive. Le TAS ainsi que la dispo en protéines ont également une contribution positive. Le croissance démographique a la plus forte contribution négative. La composante F1 peut être représenté par une notion de disponibilités. En projection, les points en bas de l'axe F1 auront des faibles dispos et un TAS faible également et en haut de F1 un croissance démographique élevé.

On peut dire que les pays qui ont une croissance démographique elevé ont en générale un TAS faible et de faible dispos

III. ANALYSE DES COMPOSANTES PRINCIPALES

CERCLES DES CORRÉLATIONS



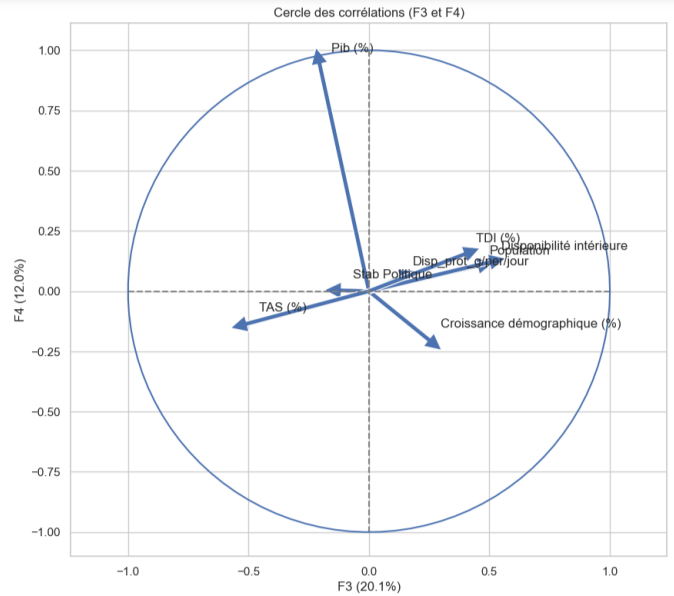
CORRÉLATIONS AVEC LES CP

F2 : La population a une forte contribution positive. La stabilité politique a une contribution négative importante. Le TDI est corrélé négativement à F2
La composante F2 peut être représenté par une notion d'évolution des populations. En projection, les points à droite de l'axe F2 auront des une population faible et à gauche de l'axe F2 une bonne stabilité politique et un TDI élevé.

Sur F1/F2 les points en bas à gauche seront les pays les plus adaptés à notre analyse (TAS faible, TDI élevé, bonne croissance, bonne stabilité, faibles disponibilités

III. ANALYSE DES COMPOSANTES PRINCIPALES

CERCLES DES CORRÉLATIONS



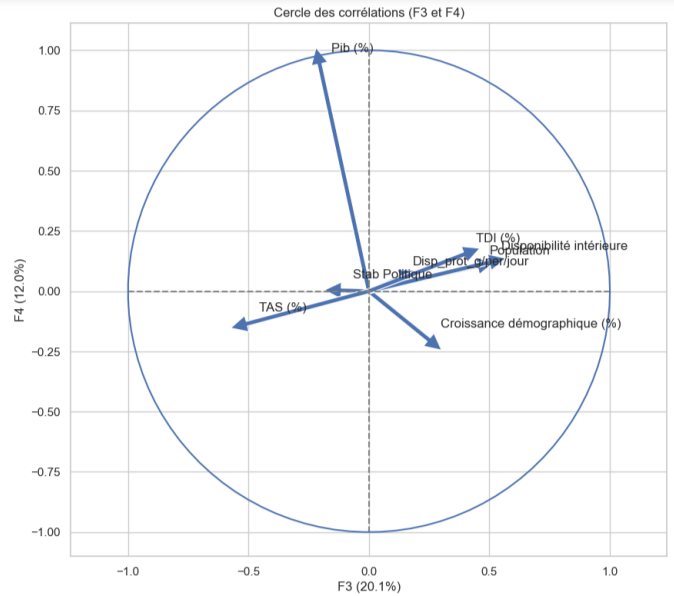
CORRÉLATIONS AVEC LES CP

F3 :

La disponibilité intérieure a une contribution positive élevée.
Le TAS a une contribution négative élevée.
Le TDI a une contribution positive élevée.
La composante F3 est aussi une notion de disponibilités.
En projection, les points en haut de l'axe F3 auront des faible dispo et une faible TDI en bas de l'axe F3 un TAS faible.

III. ANALYSE DES COMPOSANTES PRINCIPALES

CERCLES DES CORRÉLATIONS



CORRÉLATIONS AVEC LES CP

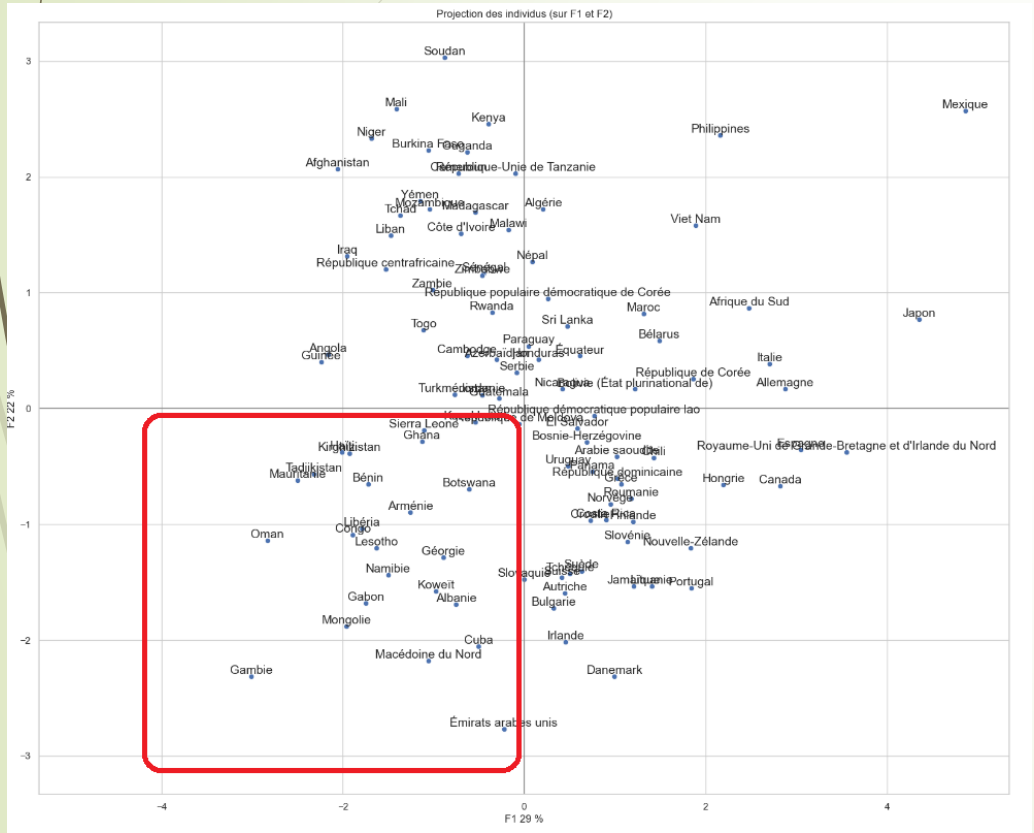
Le PIB a la plus grande contribution positive.
le croissance démographique a une contribution modérément négative.

F4 : La composante F4 peut être représenté par une notion de richesse du pays.
En projection, les points à gauche de l'axe F4 auront une bonne croissance démographique et à droite de l'axe F4 un PIB élevé.

Sur F3/F4 l'analyse est plus complexe car elle ne représente que 30% de l'inertie totale.

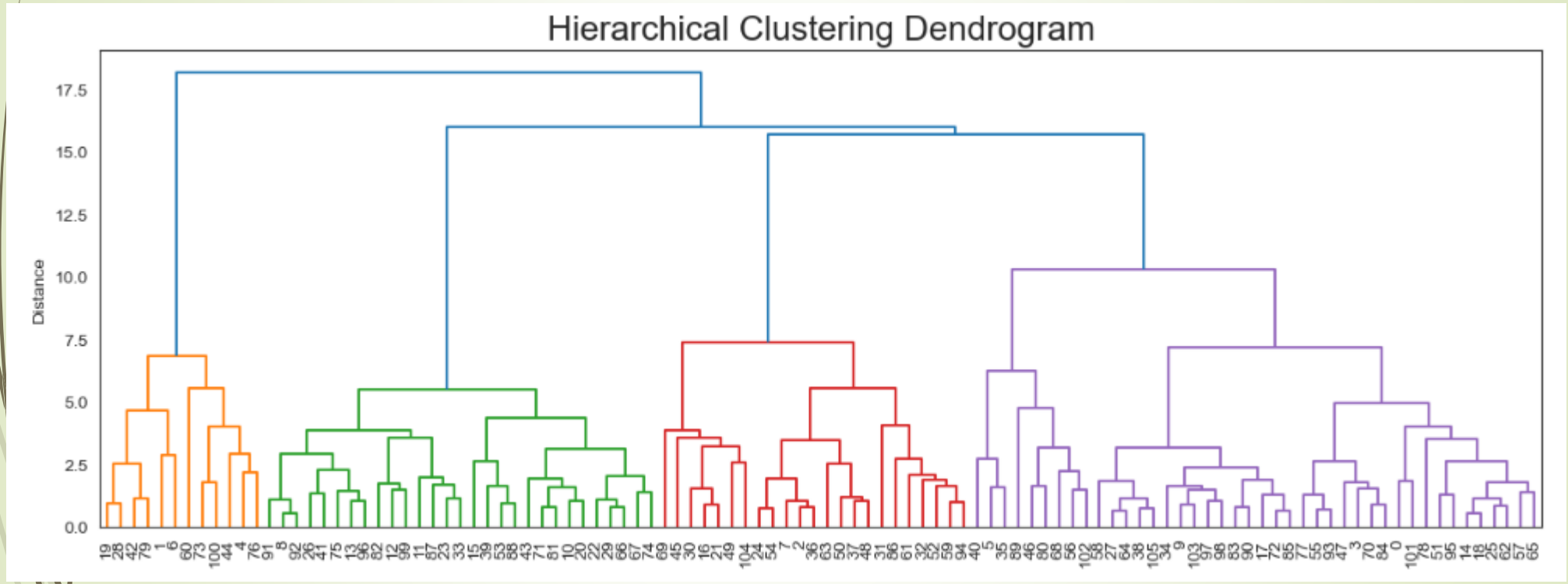
III. ANALYSE DES COMPOSANTES PRINCIPALES

Projection nuages de points des individus

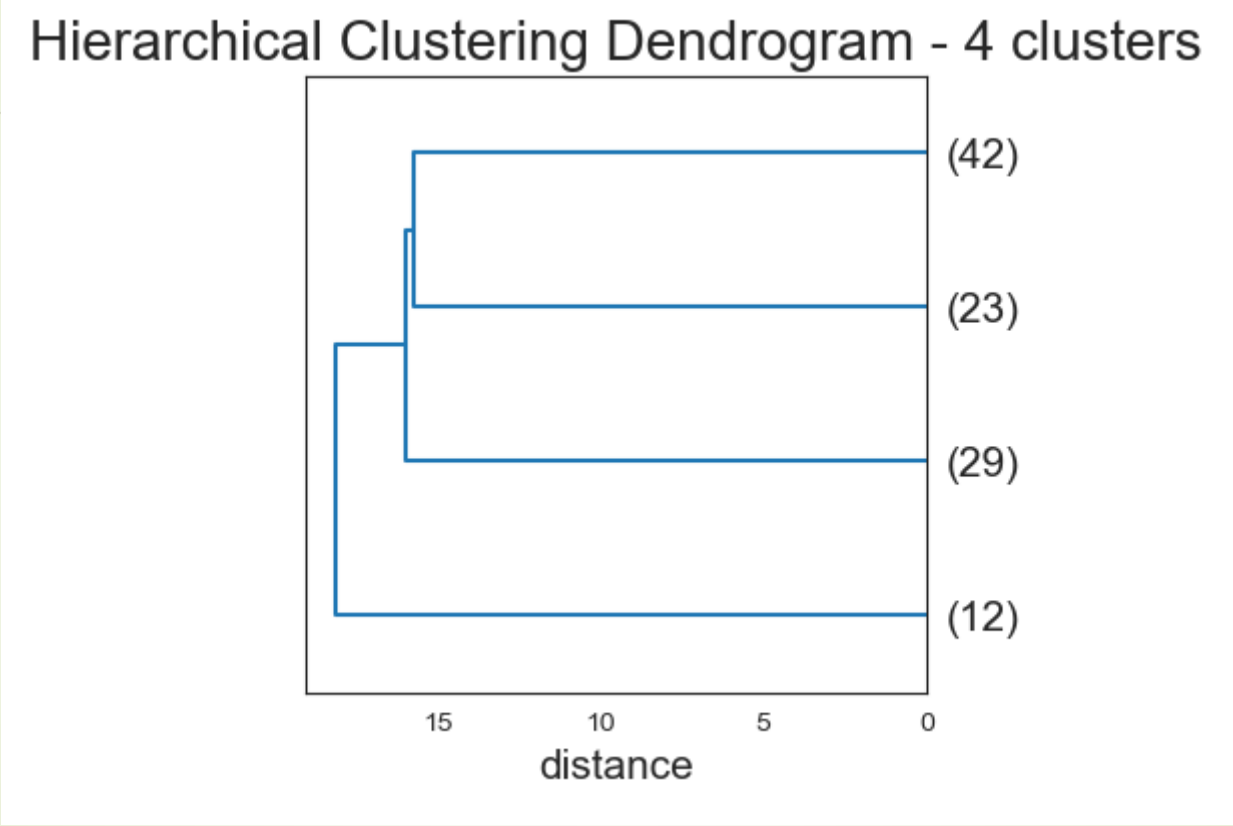


Par rapport au cercle des corrélations F1/F2 les pays en bas à gauche sont donc à priori des pays cibles intéressants pour notre analyse. Nous confirmerons cela avec la CAH.

IV. CLASSIFICATION ASCENDANTE HIÉRARCHIQUE



IV. CLASSIFICATION ASCENDANTE HIÉRARCHIQUE

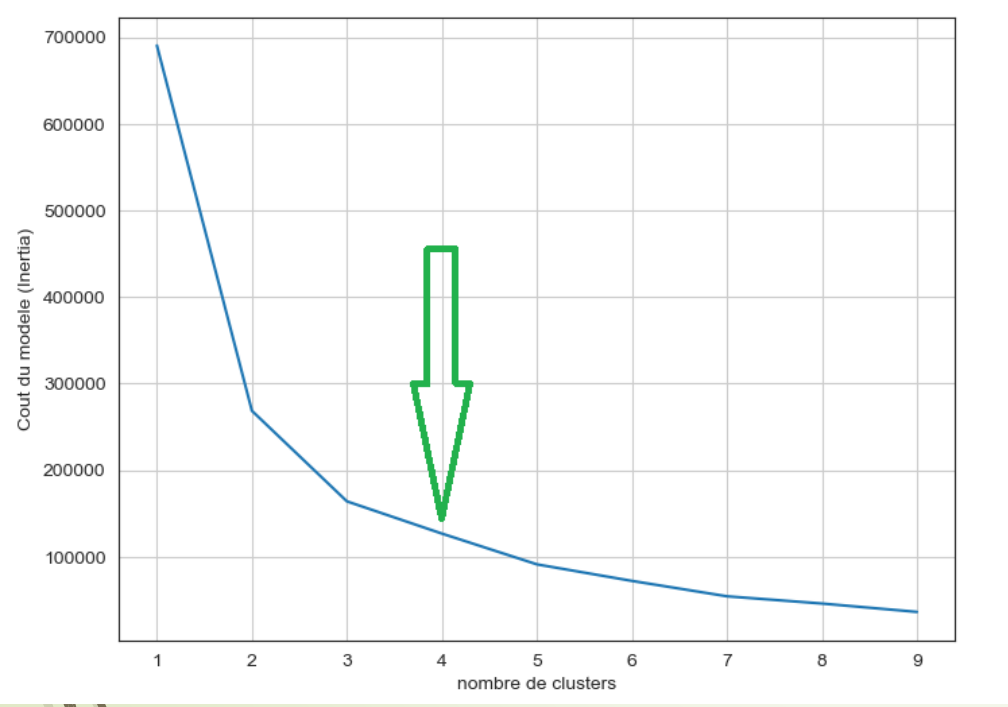


- Groupe 1 : 12 pays - Groupe 2 : 29 pays - Groupe 3 : 23 pays - Groupe 4 : 42 pays

IV. MÉTHODE K-MEANS

Recherche et vérification du nombre de clusters

MÉTHODE DU COUDE

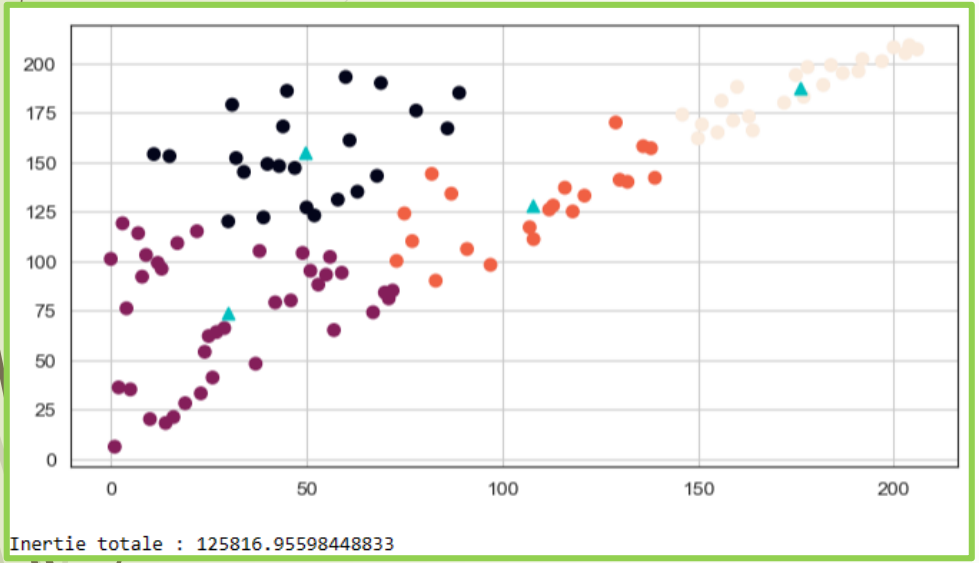


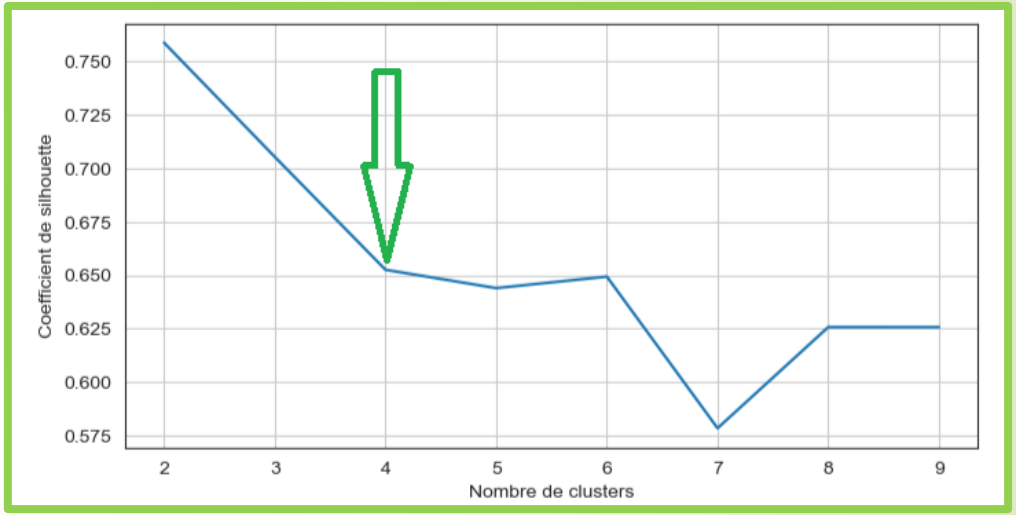
En se basant sur la notion d'inertie qui baisse avec l'augmentation du nombre de clusters jusqu'au point de stagnation qui nous indique le nombre idéal de clusters :

Sur ce graphique le point de stagnation se situe à 4 clusters

IV. MÉTHODE K-MEANS (Suite)

Après avoir implémenté le K-means et fixé le nombre de cluster à 4 nous obtenons les nuage de points ci dessous, avec affichage des cluster et leur centroïdes



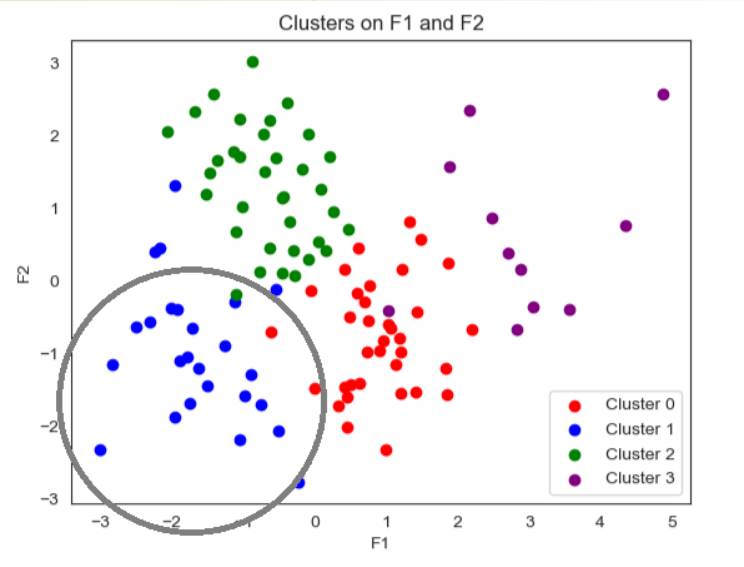


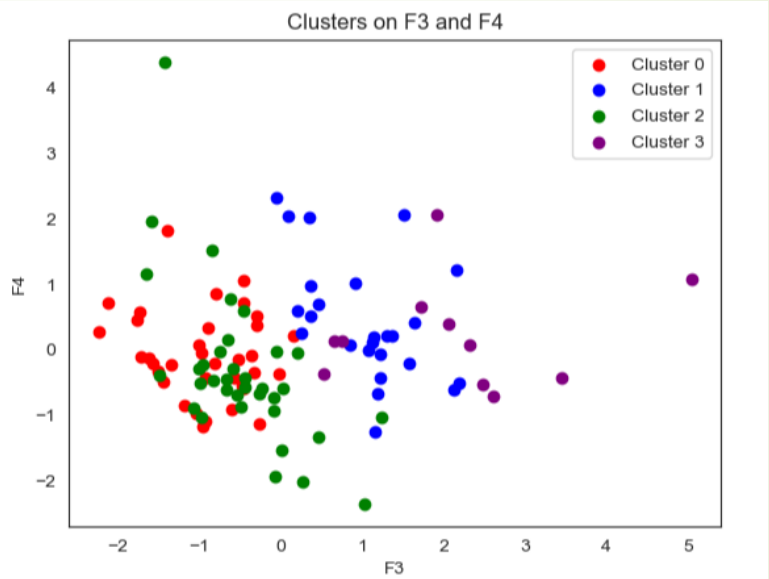
L'affichage de l'évolution du coefficient de silhouette en fonction du nombre de clusters :
- **Le nombre de 4 clusters donne bien le coefficient de silhouette le plus élevé une fois la courbe stabilisé : 0.65**

L'affichage du nuage de points avec les 4 clusters et leur centroïdes grâce à l'algorithme Kmeans :
- **Le nuage de points est étalé**
- **Le nombre de clusters est optimal, et centroïdes bien distants**

IV. MÉTHODE K-MEANS (Suite)

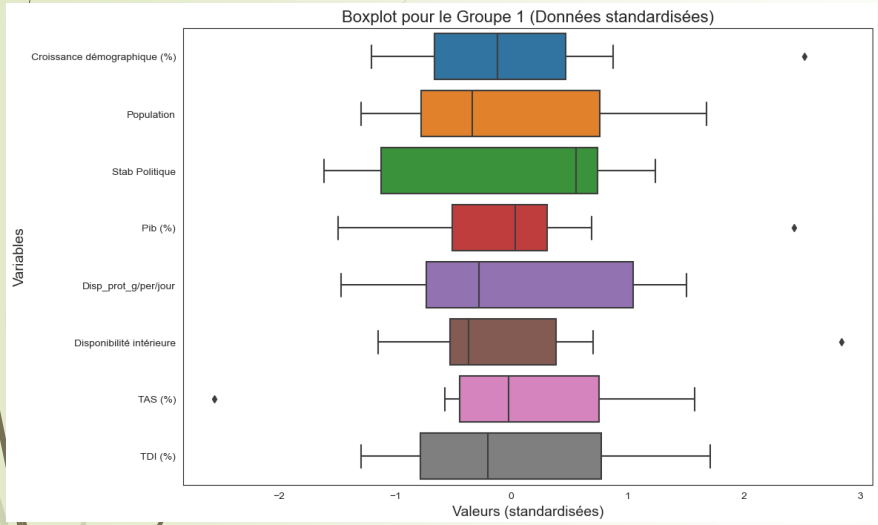
Projections des clusters sur les axes

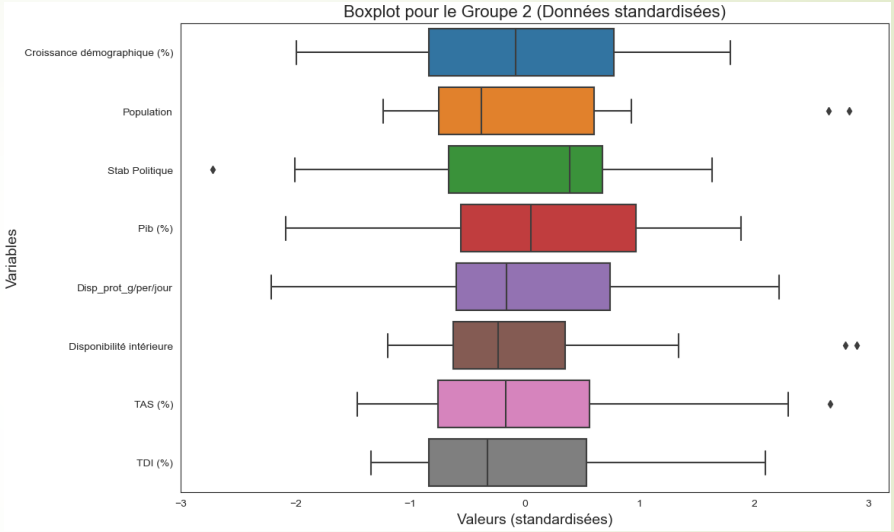




Sur l'axe F1/F2 nous avons bien un cluster en bas à gauche le cluster 1 qui se dégage.
Nous confirmerons par la suite en analysant chaque cluster

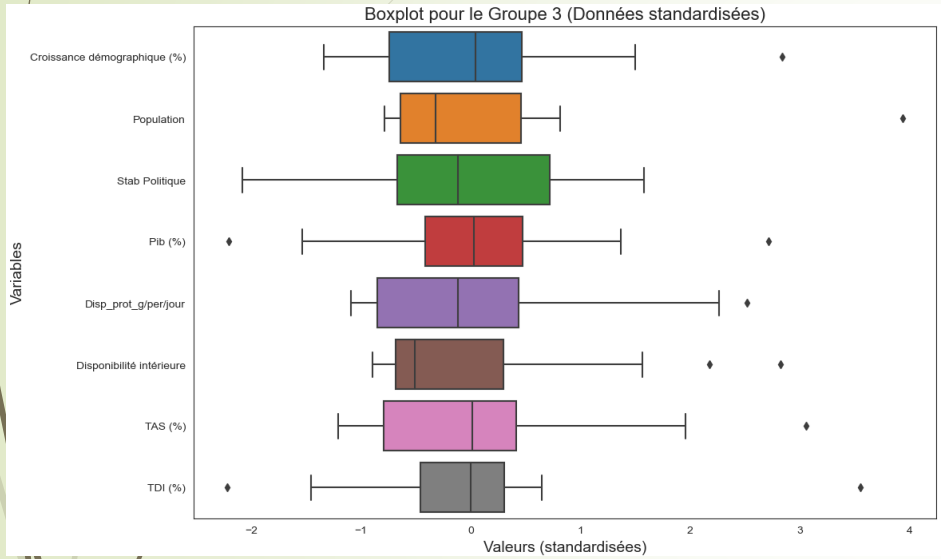
V. ANALYSE DES GROUPES (CLUSTERS)

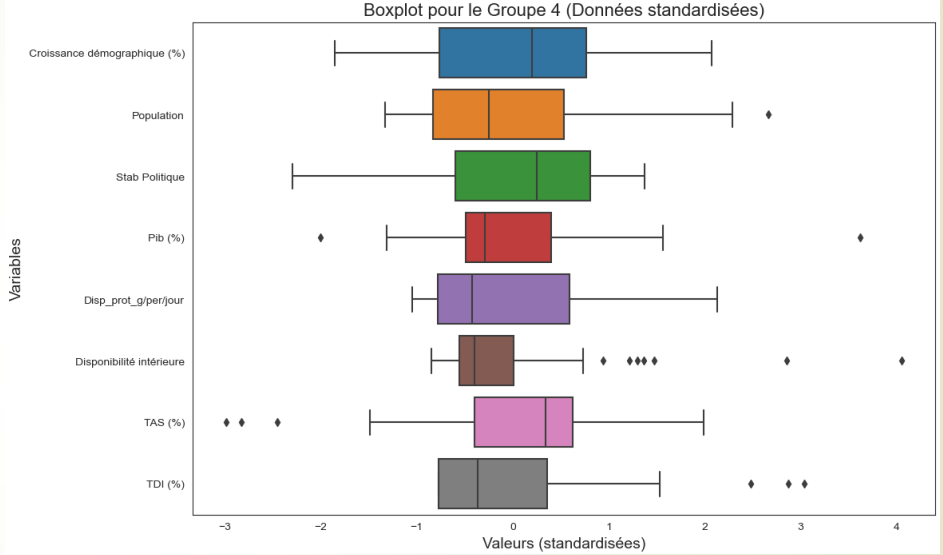




Les boxplot des différentes variables nous permettent de caractériser chaque groupe

V. ANALYSE DES GROUPES (CLUSTERS)





Les boxplot des différentes variables nous permettent de caractériser chaque groupe

V. ANALYSE DES GROUPES (CLUSTERS)

Croissance démographique (%)	0
Population	74409844
Stab Politique	0
Pib (%)	5
Disp_prot_g/per/jour	9
Disponibilité intérieure	1820
TAS (%)	85
TDI (%)	26
Groupe	1

GROUPE 1

- Un taux de dépendance à l'importation des plus faible
- Un taux d'auto-suffisance des plus élevés
- Une disponibilité des plus élevés
- Un PIB des plus faible

Croissance démographique (%)	0
Population	7383551
Stab Politique	0
Pib (%)	7
Disp_prot_g/per/jour	8
Disponibilité intérieure	206
TAS (%)	107
TDI (%)	33
Groupe	2

GROUPE 2

- Un taux de dépendance à l'importation faible
- Un taux d'auto-suffisance des plus élevés
- Une disponibilité un peu faible
- Un PIB bon

Croissance démographique (%)	0
Population	6688903
Stab Politique	0
Pib (%)	6
Disp_prot_g/per/jour	5
Disponibilité intérieure	105
TAS (%)	21
TDI (%)	91
Groupe	3

GROUPE 3

- Un taux de dépendance à l'importation le plus élevé
- Un taux d'auto-suffisance le plus faible
- Une disponibilité la plus faible des groupes
- Un PIB bon

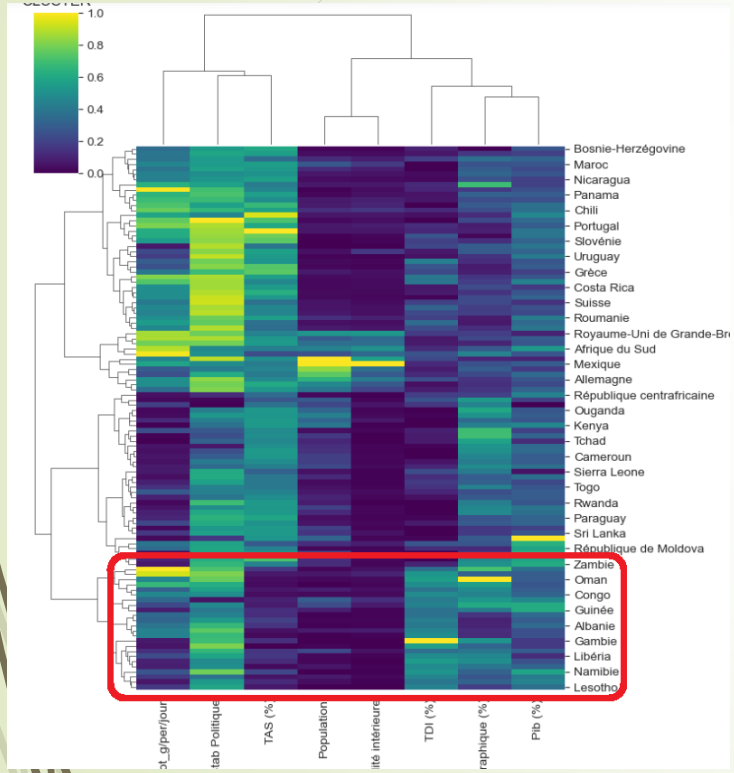
Croissance démographique (%)	0
Population	21428130
Stab Politique	0
Pib (%)	9
Disp_prot_g/per/jour	2
Disponibilité intérieure	134
TAS (%)	84
TDI (%)	17
Groupe	4

GROUPE 4

- Un taux de dépendance à l'importation des plus faible
- Un taux d'auto-suffisance des plus élevé
- Une diponibilité faible
- Un PIB des plus élevés

V. ANALYSE DES GROUPES (CLUSTERS)

Heatmap des clusters et les variables



Les informations obtenues des boxplot, des moyennes et la heatmap des groupes met en avant le groupe 3

Les caractéristique recherchées du groupe idéal en terme de besoins en viande de poulet :

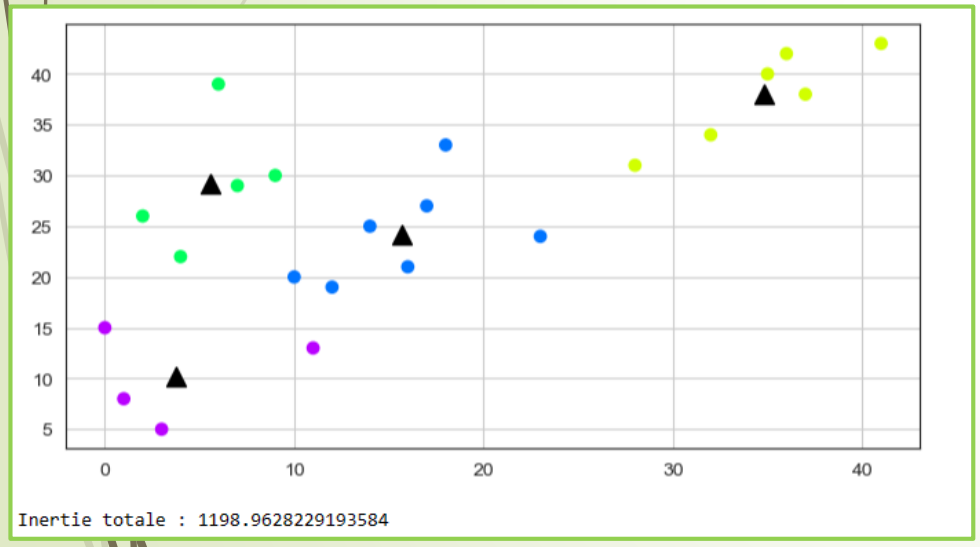
- Les disponibilités les plus faibles
- Une auto-suffisance des plus faibles
- Une dépendance à l'importation des plus élevée
- Une croissance démographique élevé
- Un PIB élevé

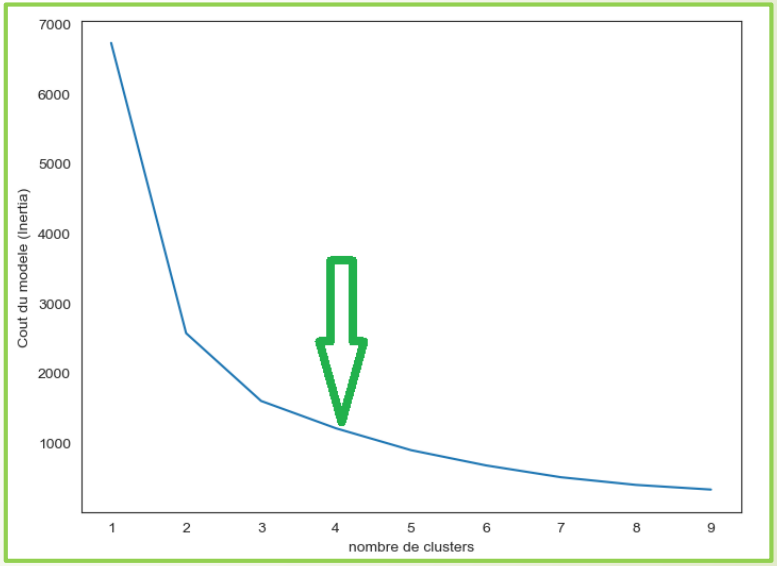
VII. EXPLORATION DU CLUSTER SÉLECTIONNÉ

Pour affiner d'avantage notre résultat, nous avons appliquer la même démarche sur le groupe 3

- 4 sous-groupes résultent de cette analyse

AFFICHAGES DES SOUS-CLUSTERS ET LEUR CENTROÏDES





VII. EXPLORATION DU CLUSTER SÉLECTIONNÉ

```
Groupe 1 Croissance démographique (%)          0
Population                               4782895
Stab Politique                                 0
Pib (%)                                        7
Disp_prot_g/per/jour                           4
Disponibilité intérieure                      53
TAS (%)                                       17
TDI (%)                                       98
Groupe                                         3
Sous_Groupes                                   1
dtype: int32
Groupe 2 Croissance démographique (%)          0
Population                               8807734
Stab Politique                                 0
Pib (%)                                        0
Disp_prot_g/per/jour                           5
Disponibilité intérieure                     152
TAS (%)                                       66
TDI (%)                                       34
Groupe                                         3
Sous_Groupes                                   2
dtype: int32
Groupe 3 Croissance démographique (%)          0
Population                               4483392
Stab Politique                                 0
Pib (%)                                        8
Disp_prot_g/per/jour                          11
Disponibilité intérieure                     151
TAS (%)                                       17
TDI (%)                                       91
Groupe                                         3
Sous_Groupes                                   3
dtype: int32
Groupe 4 Croissance démographique (%)          0
Population                              16912065
Stab Politique                                 0
Pib (%)                                        6
Disp_prot_g/per/jour                           8
Disponibilité intérieure                     321
TAS (%)                                       16
TDI (%)                                       89
Groupe                                         3
Sous_Groupes                                   4
dtype: int32
```

Comme pour l'analyse des groupes, certains sous-groupes présentent des caractéristiques plus favorables à notre objectif.

Nous décidons de garder les sous groupes 1,3 et 4 car le sous groupe 2 présente un TAS beaucoup plus élevé que les 3 autres groupes (66% contre 17% en moyenne) et un TDI bien plus faible (34% contre 93% en moyenne) et un PIB de 0 quand les 3 autres sous-groupes ont des PIB positifs.

VIII. CONCLUSION

Le groupe de pays qui correspond au critères de sélection en terme de besoins en viande de volaille est le Groupe 3.

De ce groupe nous avons sélectionné les pays qui correspondent le mieux au profil recherché.

Zone	Croissance démographique (%)	Population	Stab Politique	Pib (%)	Disp_prot_g/per/jour	Disponibilité intérieure	TAS (%)	TDI (%)	Groupe	Sous_Groupes
Albanie	-0.009564	2882740.0	0.11	9.0	6.26	47.0	27.659574	80.851064	3	1
Arménie	0.019728	2951745.0	-0.84	9.0	5.44	47.0	23.404255	74.468085	3	1
Bénin	0.210597	11485044.0	-0.30	7.0	4.98	161.0	11.180124	76.397516	3	1
Congo	0.192011	5244359.0	-0.61	8.0	7.45	110.0	6.363636	94.545455	3	1
Gabon	0.256554	2119275.0	-0.09	6.0	10.59	78.0	5.128205	97.435897	3	1
Gambie	0.234405	2280094.0	0.18	2.0	1.24	8.0	25.000000	200.000000	3	1
Géorgie	-0.012698	4002942.0	-0.42	7.0	5.19	61.0	36.065574	83.606557	3	1
Haïti	0.086419	11123178.0	-1.10	12.0	2.75	98.0	9.183673	90.816327	3	1
Kirghizistan	0.115462	6304030.0	-0.43	13.0	1.08	32.0	21.875000	78.125000	3	1
Lesotho	0.037652	2108328.0	-0.22	10.0	2.72	17.0	11.764706	88.235294	3	1
Libéria	0.198903	4818973.0	-0.24	0.0	3.74	50.0	30.000000	96.000000	3	1
Macédoine du Nord	0.003793	2082957.0	0.12	5.0	7.01	41.0	4.878049	97.560976	3	1
Mauritanie	0.218999	4403313.0	-0.67	5.0	1.59	22.0	22.727273	109.090909	3	1
Mongolie	0.120583	3170216.0	0.65	2.0	0.95	9.0	0.000000	111.111111	3	1
Namibie	0.115822	2448301.0	0.55	19.0	4.25	28.0	39.285714	103.571429	3	1
Tadjikistan	0.161740	9100835.0	-0.61	8.0	1.45	40.0	5.000000	95.000000	3	1
Koweït	0.345821	4137312.0	0.30	10.0	15.87	189.0	29.629630	72.486772	3	3
Oman	0.514953	4829473.0	0.51	7.0	7.38	114.0	6.140351	110.526316	3	3
Cuba	0.007383	11338134.0	0.43	5.0	7.12	342.0	8.479532	91.228070	3	4
Ghana	0.162095	29767102.0	0.07	7.0	2.26	211.0	28.436019	71.563981	3	4
Émirats arabes unis	0.110274	9630959.0	0.65	7.0	14.80	412.0	11.650485	105.097087	3	4

Pour tous ces pays le taux de dépendance à l'importation est élevé et inversement le taux d'auto-suffisance est faible.

Les pays ayants les plus faibles disponibilités alors qu'ils sont très dépendants de l'importation pourraient correspondre.

VIII. CONCLUSION

Dans ce groupe là, nous pouvons encore isoler les pays ayant un besoin encore plus important (TAS < 10% et TDI > 90%) que nous pourrons ensuite transposer avec les autres variables (Stabilité politique notamment) afin d'affiner notre choix.

Zone	Croissance démographique (%)	Population	Stab Politique	Pib (%)	Disp_prot_g/per/jour	Disponibilité intérieure	TAS (%)	TDI (%)	Groupe	Sous_Groupes
Mongolie	0.120583	3170216.0	0.65	2.0	0.95	9.0	0.000000	111.111111	3	1
Oman	0.514953	4829473.0	0.51	7.0	7.38	114.0	6.140351	110.526316	3	3
Cuba	0.007383	11338134.0	0.43	5.0	7.12	342.0	8.479532	91.228070	3	4
Macédoine du Nord	0.003793	2082957.0	0.12	5.0	7.01	41.0	4.878049	97.560976	3	1
Gabon	0.256554	2119275.0	-0.09	6.0	10.59	78.0	5.128205	97.435897	3	1
Congo	0.192011	5244359.0	-0.61	8.0	7.45	110.0	6.363636	94.545455	3	1
Tadjikistan	0.161740	9100835.0	-0.61	8.0	1.45	40.0	5.000000	95.000000	3	1
Haïti	0.086419	11123178.0	-1.10	12.0	2.75	98.0	9.183673	90.816327	3	1

Les choix idéaux si nous souhaitons travailler dans un pays avec une stabilité politique positive sont :

Macédoine du Nord / Mongolie / Oman / Cuba
L'equipe métier affinera ce choix pour la décision finale,