BECHEUR
YAZID

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

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 functions import *
from sklearn.metrics import silhouette_score
pd.options.mode.chained_assignment = None
import warnings
warnings.filterwarnings("ignore")
from kneed import KneeLocator
from yellowbrick.cluster import SilhouetteVisualizer
```

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)

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 3 datasets, « population », « disponibilité alimentaaire », « Pib »
7 variables finales seront utilisées pour cette analyse,

Zone	Disp_quanti_kg/per/an	Disp_alim_Kcal/per/jour	Disp_prot_g/per/jour	TAS (%)	TDI (%)	Croissance démographique (%)	Pib (%)
Afghanistan	1.53	5.0	0.54	49.122807	50.877193	0.247101	3.0
Afrique du Sud	35.69	143.0	14.11	78.706327	24.268178	0.092891	17.0
Albanie	16.36	85.0	6.26	27.659574	80.851064	-0.009564	9.0
Algérie	6.38	22.0	1.97	99.277978	0.722022	0.129027	6.0
Allemagne	19.47	71.0	7.96	87.061530	48.418631	0.020710	6.0

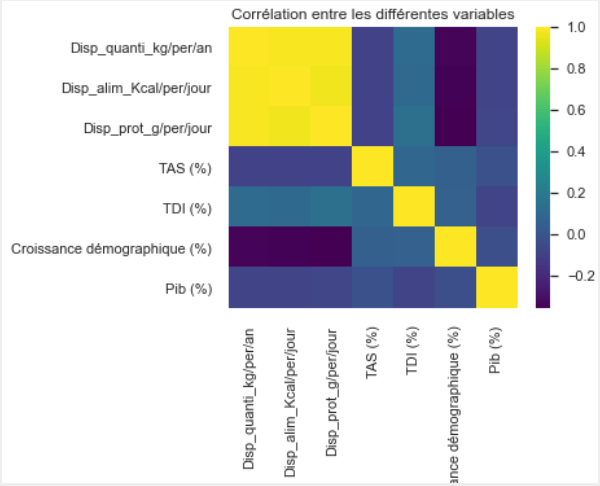
II. IMPORTATION DES DONNÉES (suite)

Après jointure des 3 datasets, « population », « disponibilité alimentaaire », « Pib »
7 variables finales seront utilisées pour cette analyse,

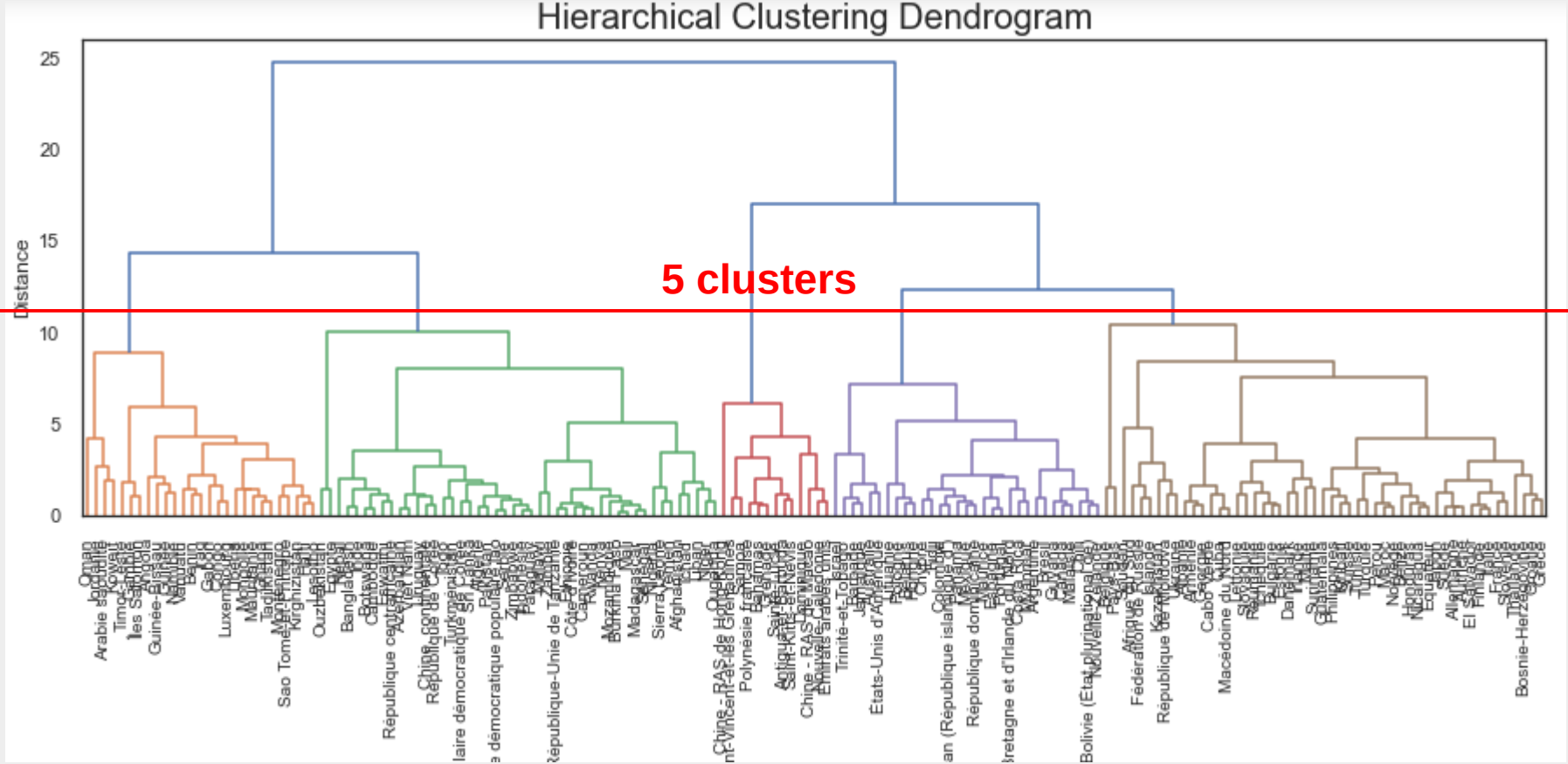
Zone	Disp_quanti_kg/per/an	Disp_alim_Kcal/per/jour	Disp_prot_g/per/jour	TAS (%)	TDI (%)	Croissance démographique (%)	Pib (%)
Afghanistan	1.53	5.0	0.54	49.122807	50.877193	0.247101	3.0
Afrique du Sud	35.69	143.0	14.11	78.706327	24.268178	0.092891	17.0
Albanie	16.36	85.0	6.26	27.659574	80.851064	-0.009564	9.0
Algérie	6.38	22.0	1.97	99.277978	0.722022	0.129027	6.0
Allemagne	19.47	71.0	7.96	87.061530	48.418631	0.020710	6.0

Corrélations entre les variables :

- Les disponibilités sont très corrélées entre elles
- La croissance démographique est négativement corrélé au disponibilités
- Le Pib est négativement corrélé avec toutes les variables



III. MÉTHODE DE CLASSIFICATION ASCENDANTE HIÉRARCHIQUE (CAH)

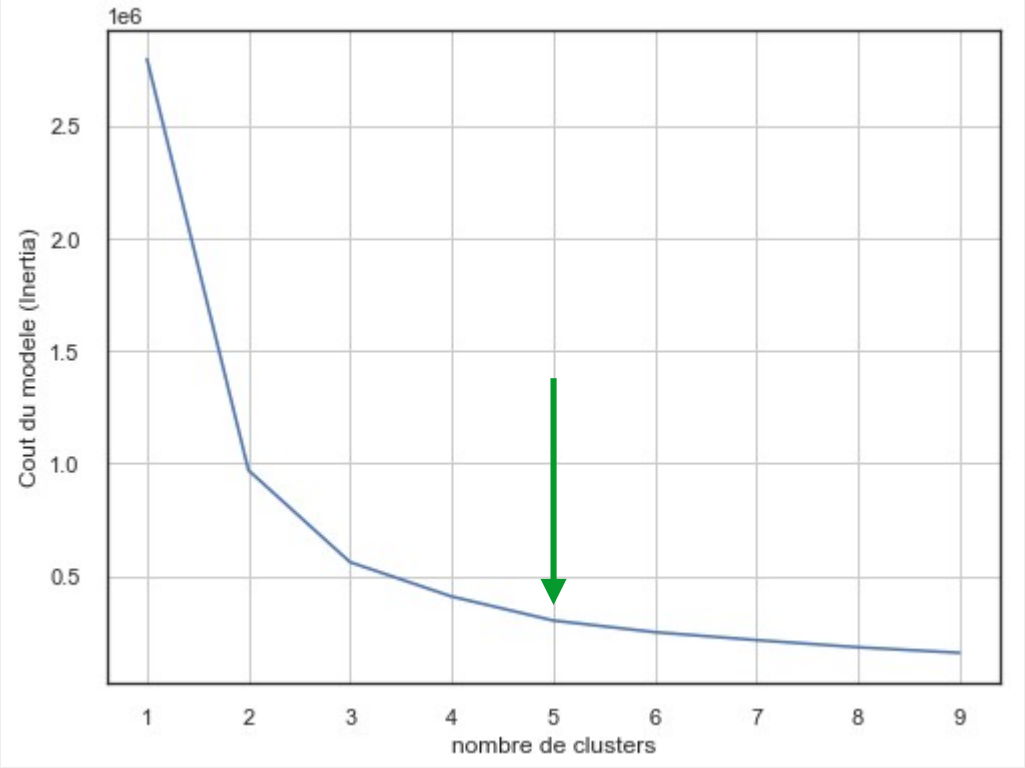


- Groupe 1 : 27 pays - Groupe 2 : 46 pays - Groupe 3 : 13 pays - Groupe 4 : 31 pays - Groupe 5 : 51 pays

IV. MÉTHODE K-MEANS

Recherche et vérification du nombre de clusters

MÉTHODE DU COUDE

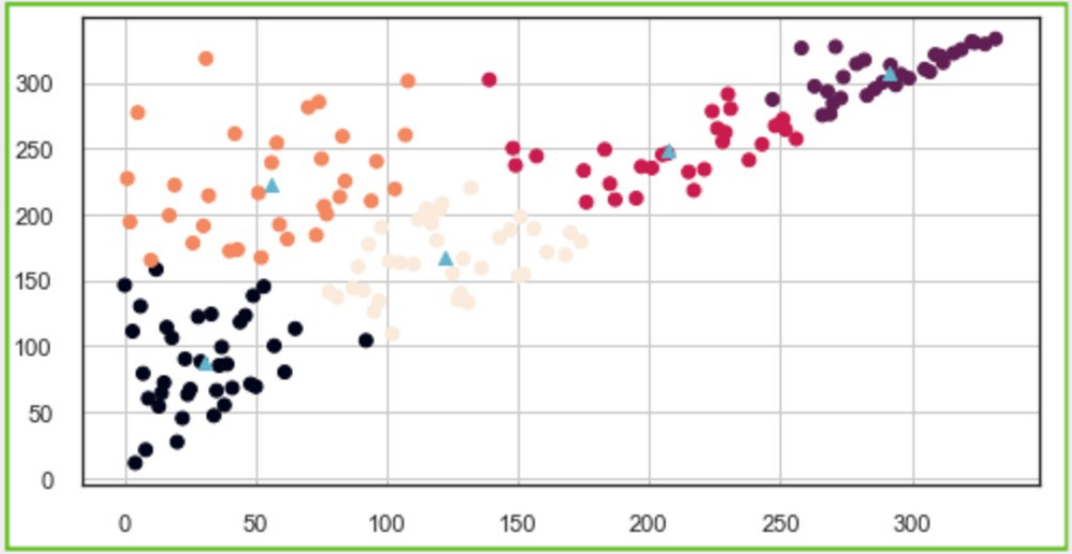


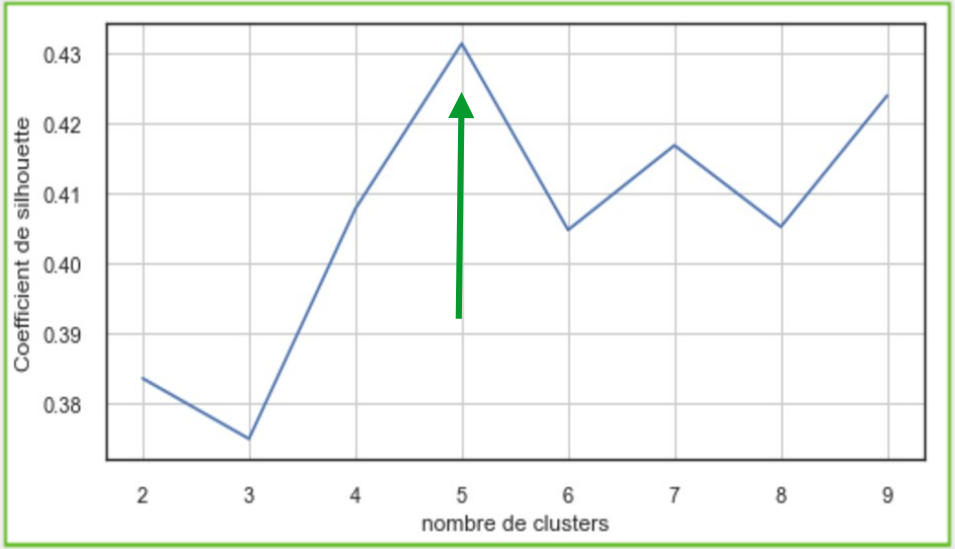
En se basant sur la notion d'inertie qui baisse avec l'augmentation du nombre de clusters jusqu'au point de stagnation qui nous indique le nombre idéal de clusters :

Sur ce graphique le point de stagnation est 5 clusters

IV. MÉTHODE K-MEANS (SUITE)

Après avoir implémenté le K-means et fixé le nombre de cluster à 5 nous obtenons les nuage de points ci dessous, avec affichage des cluster et leur centroïdes

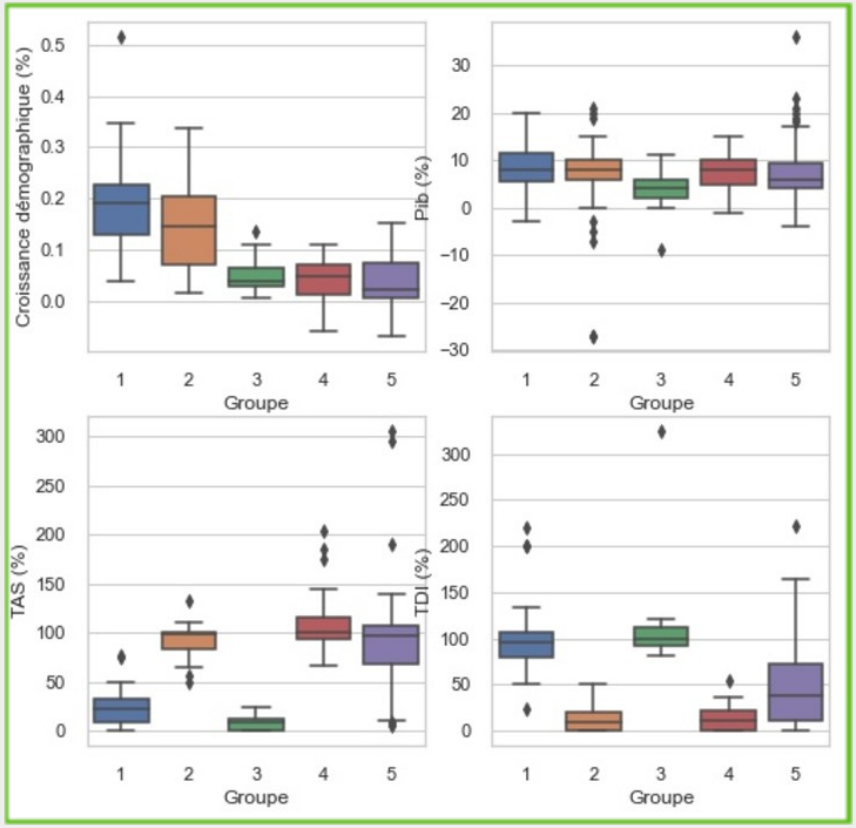


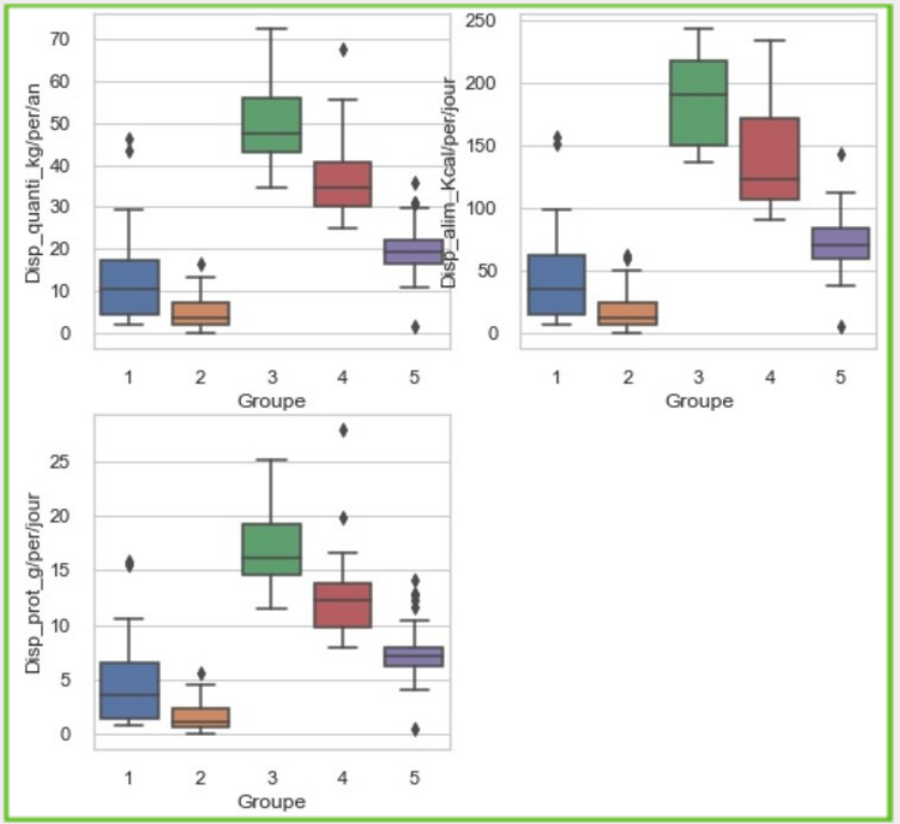


L'affichage de l'évolution du coefficient de silhouette en fonction du nombre de clusters :
- **Le nombre de 5 clusters donne bien le coefficient de silhouette le plus élevé : 0.43**

L'affichage du nuage de points avec les 5 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**

V. ANALYSE DES GROUPES (CLUSTERS)





Les boxplot des différentes variables nous permettent de caractériser chaque groupe

V. ANALYSE DES GROUPES (CLUSTERS)

Disp_quanti_kg/per/an	13.545185
Disp_alim_Kcal/per/jour	45.703704
Disp_prot_g/per/jour	4.752222
TAS (%)	23.906126
TDI (%)	101.692533
Croissance démographique (%)	0.194808
Pib (%)	8.518519
Groupe	1.000000

GROUPE 1

- Un taux de dépendance à l'importation des plus élevé
- Un taux d'auto-suffisance des plus faibles
- Une croissance démographique des plus élevée
- Une disponibilité des plus faible
- Un PIB des plus élevé

Disp_quanti_kg/per/an	5.093696
Disp_alim_Kcal/per/jour	18.369565
Disp_prot_g/per/jour	1.710217
TAS (%)	91.170130
TDI (%)	12.330206
Croissance démographique (%)	0.145253
Pib (%)	6.500000
Groupe	2.000000

GROUPE 2

- Un taux de dépendance à l'importation des plus faible
- Un taux d'auto-suffisance des plus élevé
- Une croissance démographique élevée
- Une disponibilité des plus faible
- Un PIB élevé

Disp_quanti_kg/per/an	49.635385
Disp_alim_Kcal/per/jour	188.461538
Disp_prot_g/per/jour	17.216154
TAS (%)	7.418215
TDI (%)	117.572815
Croissance démographique (%)	0.052654
Pib (%)	3.615385
Groupe	3.000000

GROUPE 3

- Un taux de dépendance à l'importation des plus élevé
- Un taux d'auto-suffisance très faible
- Une croissance démographique faible
- Une disponibilité très élevée
- Un PIB des plus faible

Disp_quanti_kg/per/an	37.122903
Disp_alim_Kcal/per/jour	139.870968
Disp_prot_g/per/jour	12.772258
TAS (%)	110.099701
TDI (%)	12.342959
Croissance démographique (%)	0.042824
Pib (%)	7.354839

GROUPE 4

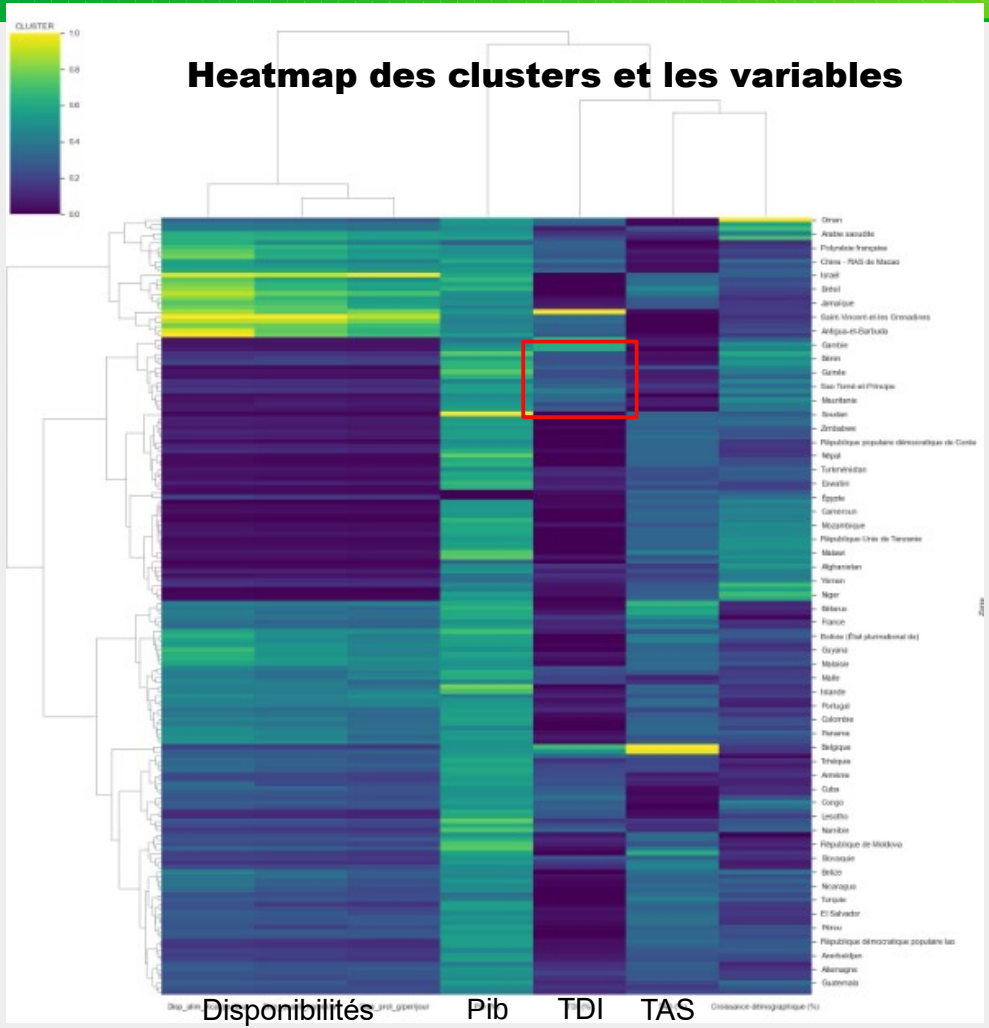
- Un taux de dépendance à l'importation des plus faible
- Un taux d'auto-suffisance des plus élevé
- Une croissance démographique faible
- Une diponibilité élevée
- Un PIB élevé

Disp_quanti_kg/per/an	19.822941
Disp_alim_Kcal/per/jour	72.666667
Disp_prot_g/per/jour	7.338627
TAS (%)	93.937923
TDI (%)	46.357254
Croissance démographique (%)	0.035712
Pib (%)	8.019608

GROUPE 5

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

V. ANALYSE DES GROUPES (CLUSTERS)



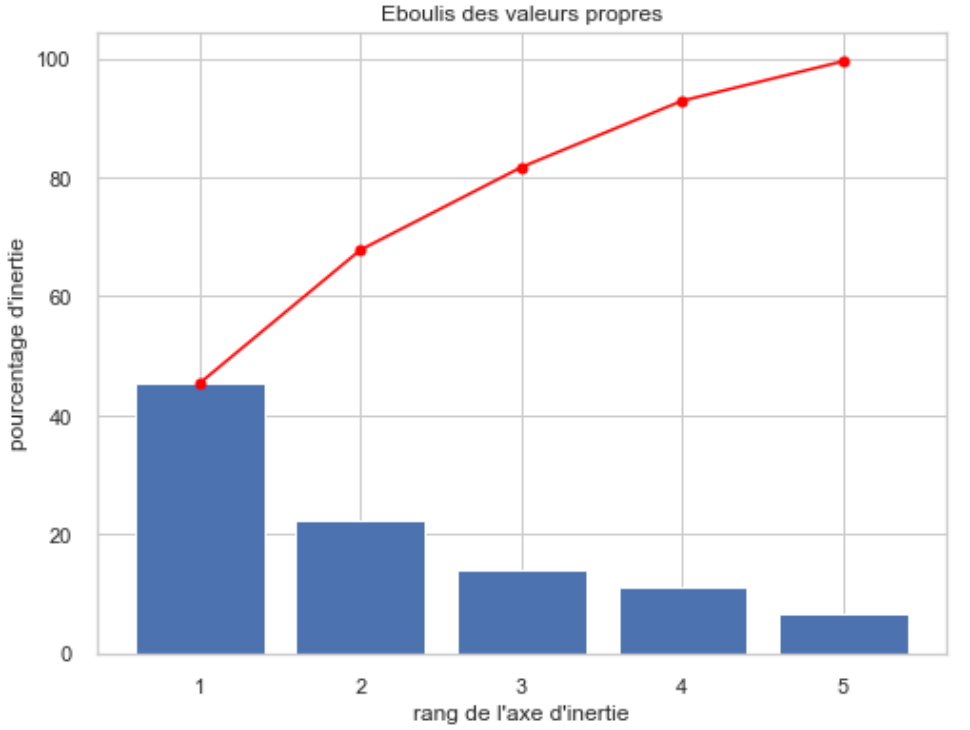
Les informations obtenues des boxplot, des moyennes et la heatmap des groupes met en avant le groupe 1

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

VI. ANALYSE DES COMPOSANTES PRINCIPALES (ACP)

Évaluation quantitative des informations apportées par chaque composante



Nous avons dans notre cas l'inertie totale répartie inéquitablement sur 5 axes :

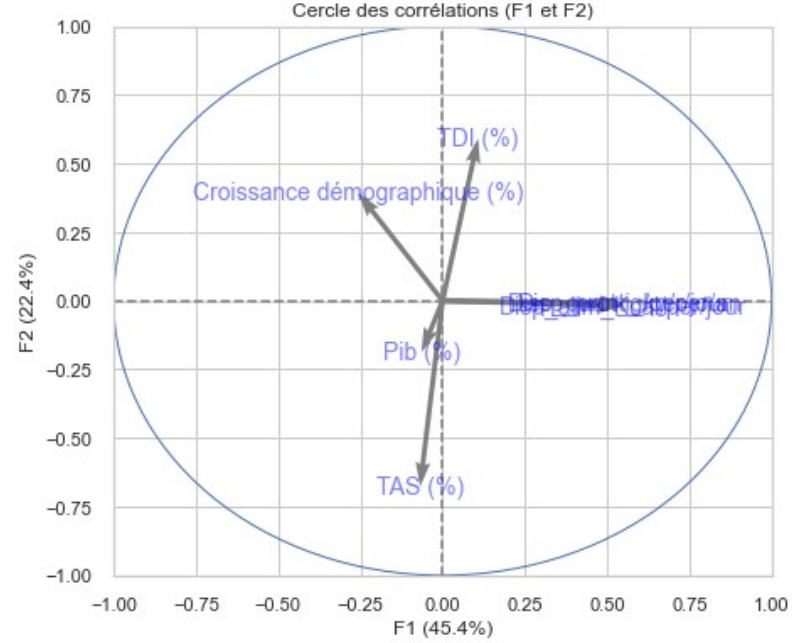
- **Axe 1** : 45,4 % de l'inertie totale
- **Axe 2** : 22,4 % de l'inertie totale
- **Axe 3** : 13,9 % de l'inertie totale
- **Axe 4** : 11,2 % de l'inertie totale
- **Axe 5** : 7,1 % de l'inertie totale

Nous appliquerons le critère de Kaiser (100/p) % pour le nombre de composantes à analyser

- Critère de Kaiser = (100/7) = 14,28 % : on étudiera les 2 première F1 et F2 = 67,8 % de l'information

VI. ANALYSE DES COMPOSANTES PRINCIPALES (ACP) (SUITE)

CERCLES DES CORRÉLATIONS



CORRÉLATIONS AVEC LES CP

F1 :

- **Variables corrélées positivement :**

Toutes les disponibilités sont corrélées (0.55), on peut dire que l'axe F1 représente les disponibilités

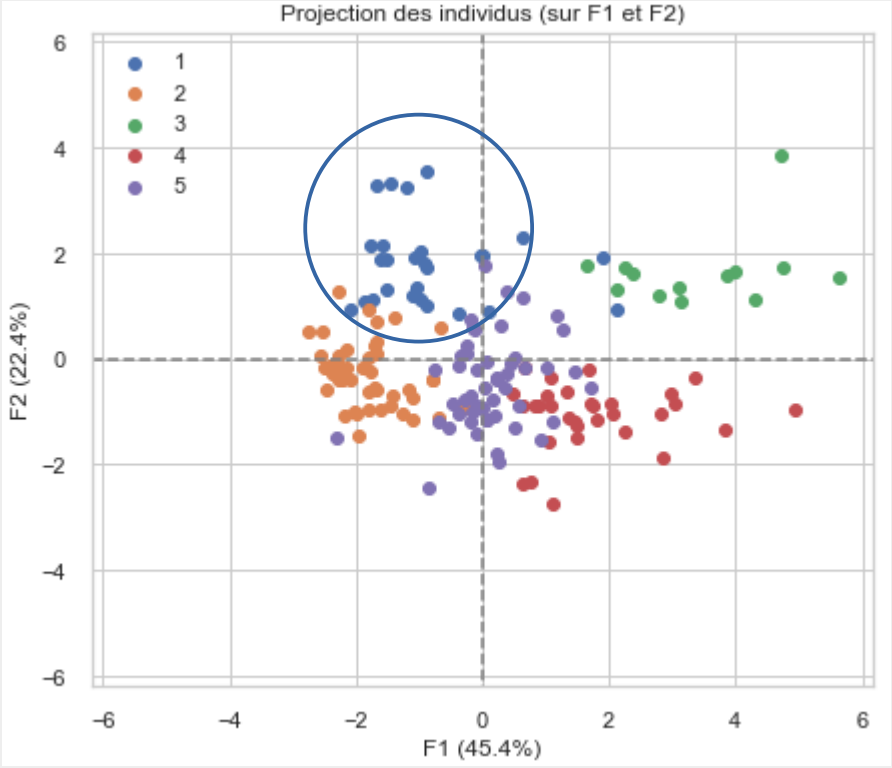
F2 :

- Variables corrélées positivement : Le TDI est fortement corrélé (0.59)
- variables corrélées négativement : Le TAS est fortement corrélé (- 0,67)

On peut dire que les pays avec un fort TDI ont une croissance démographique positive et un faible TAS

VI. ANALYSE DES COMPOSANTES PRINCIPALES (ACP) (SUITE)

PROJECTION DES GROUPE D'INDIVIDUS SUR F1 ET F2



Le groupe 1 présente bien les critères :

- TDI très élevé
- TAS faible
- Croissance démographique élevée
- Des disponibilités très faibles

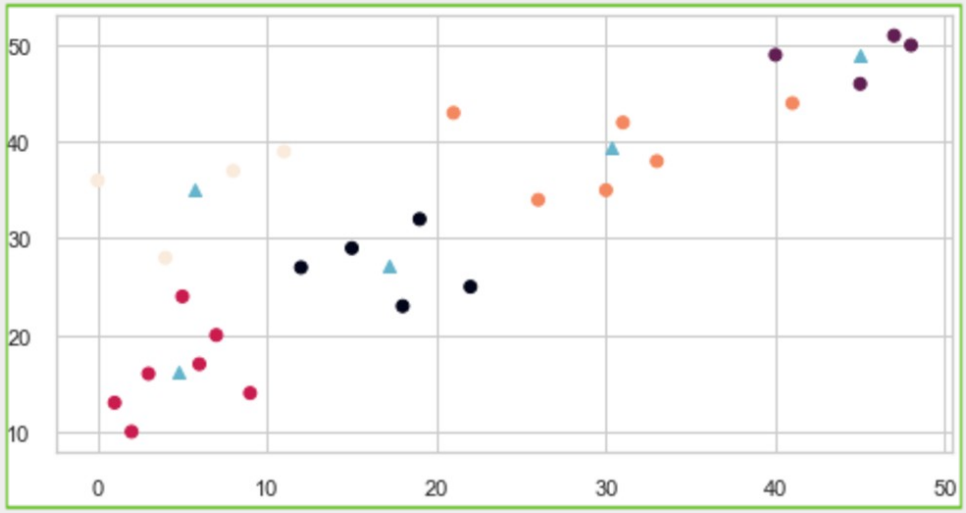
L'ACP confirme bien nos observations concernant le groupe 1

VII. EXPLORATION DU CLUSTER SÉLECTIONNÉ

Pour affiner d'avantage notre résultat, nous avons appliquer la même démarche sur le groupe 1

- 5 sous-groupes résultent de cette analyse

AFFICHAGES DES SOUS-CLUSTERS ET LEUR CENTROÏDES



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

- Le sous-groupes 4 présente :
  - TDI très élevé
  - TAS très faible
  - Disponibilités très faibles
  - PIB élevé

Ces pays peuvent être une cible pertinente pour l'exportation de viande de volaille

VIII. CONCLUSION

Le groupe de pays qui correspond au critères de sélection en terme de besoins en viande de volaille est le groupe 1.

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

Nous optons pour le sous-groupe 4

Zone	Disp_quanti_kg/per/an	Disp_alim_Kcal/per/jour	Disp_prot_g/per/jour	TAS (%)	TDI (%)	Croissance démographique (%)	Pib (%)
Ghana	7.24	16.0	2.26	28.436019	71.563981	0.162095	7.0
Haïti	8.91	31.0	2.75	9.183673	90.816327	0.086419	12.0
Kirghizistan	3.10	11.0	1.08	21.875000	78.125000	0.115462	13.0
Lesotho	7.93	27.0	2.72	11.764706	88.235294	0.037652	10.0
Libéria	10.67	36.0	3.74	30.000000	96.000000	0.198903	0.0
Mauritanie	5.14	11.0	1.59	22.727273	109.090909	0.218999	5.0
Mongolie	2.77	9.0	0.95	0.000000	111.111111	0.120583	2.0
Tadjikistan	4.45	18.0	1.45	5.000000	95.000000	0.161740	8.0

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 tout à fait à notre besoin.

Cette liste sera affinée avec les équipes métiers