Formation Openclassrooms
Projet 9 : Etude de marché avec R ou Python
AINOUNI Anwar

La poule qui chante

Sommaire

Préambule

Dans ce projet, je prends le rôle d'un data analyst qui travaille chez **"La poule qui chante"**, une entreprise française **d'agroalimentaire** qui souhaite se développer à l'international.

Mon objectif sera de proposer une première analyse **des groupements de pays** que l'on peut cibler **pour exporter les poulets**.

Pour la partie analyse, je vais tester la **classification ascendante hiérarchique**, avec un dendrogramme comme visualisation. Ensuite tester la méthode des **k-means** pour comparer les résultats des deux méthodes de clustering.

Pour finir je vais réaliser une **ACP** afin de visualiser les résultats , comprendre les groupes, les liens entre les variables, les liens entre les individus.

Sélection et préparation des données

Données utilisées :

- Dataset Population (2000-2018)
- Dataset Disponibilité alimentaire (année 2017)
- Score de la facilité de faire des affaires

Nouvelles variables ajoutées :

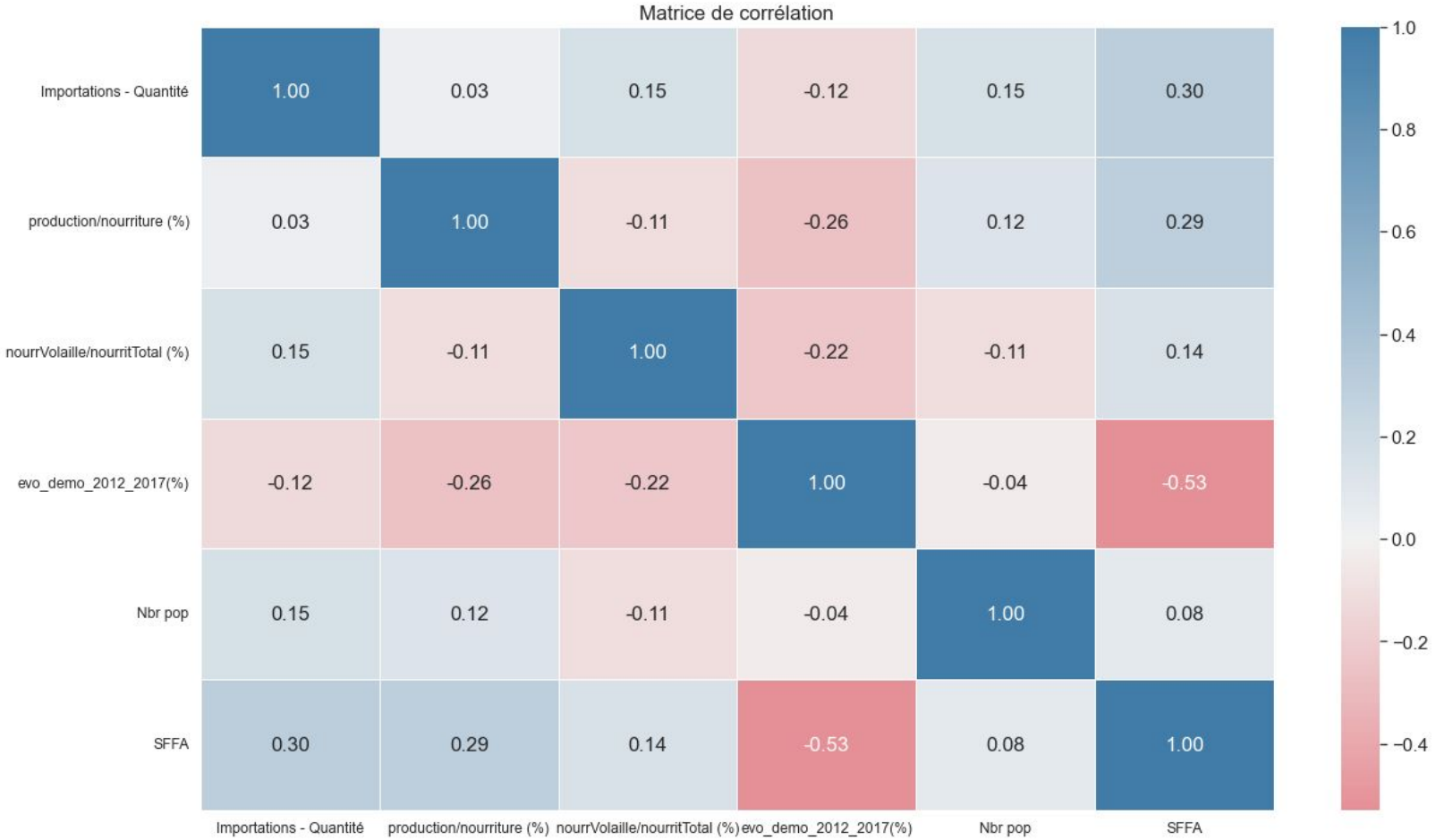
- Croissance démographique (%) sur la période 2012-2017
- Taux (nourriture en volaille / nourriture totale) x 100
- Taux (production volaille / nourriture en volaille) x 100

Sélection et préparation des données

Après jointure des 3 datasets, 6 variables finales seront utilisées pour cette analyse.

	Zone	Importations - Quantité	production/nourriture (%)	nourrVolaille/nourritTotal (%)	evo_demo_2012_2017(%)	Nbr pop	SFFA
0	Afghanistan	29.0	50.909091	0.424121	14.146796	36296.113	44.2
1	Afrique du Sud	514.0	81.916462	6.414297	7.326988	57009.756	66.7
2	Albanie	38.0	27.659574	1.294409	-1.037630	2884.169	67.0
3	Algérie	2.0	104.166667	0.798645	9.677150	41389.189	48.5
4	Allemagne	842.0	94.095712	2.063138	2.039455	82658.409	79.3

Sélection et préparation des données

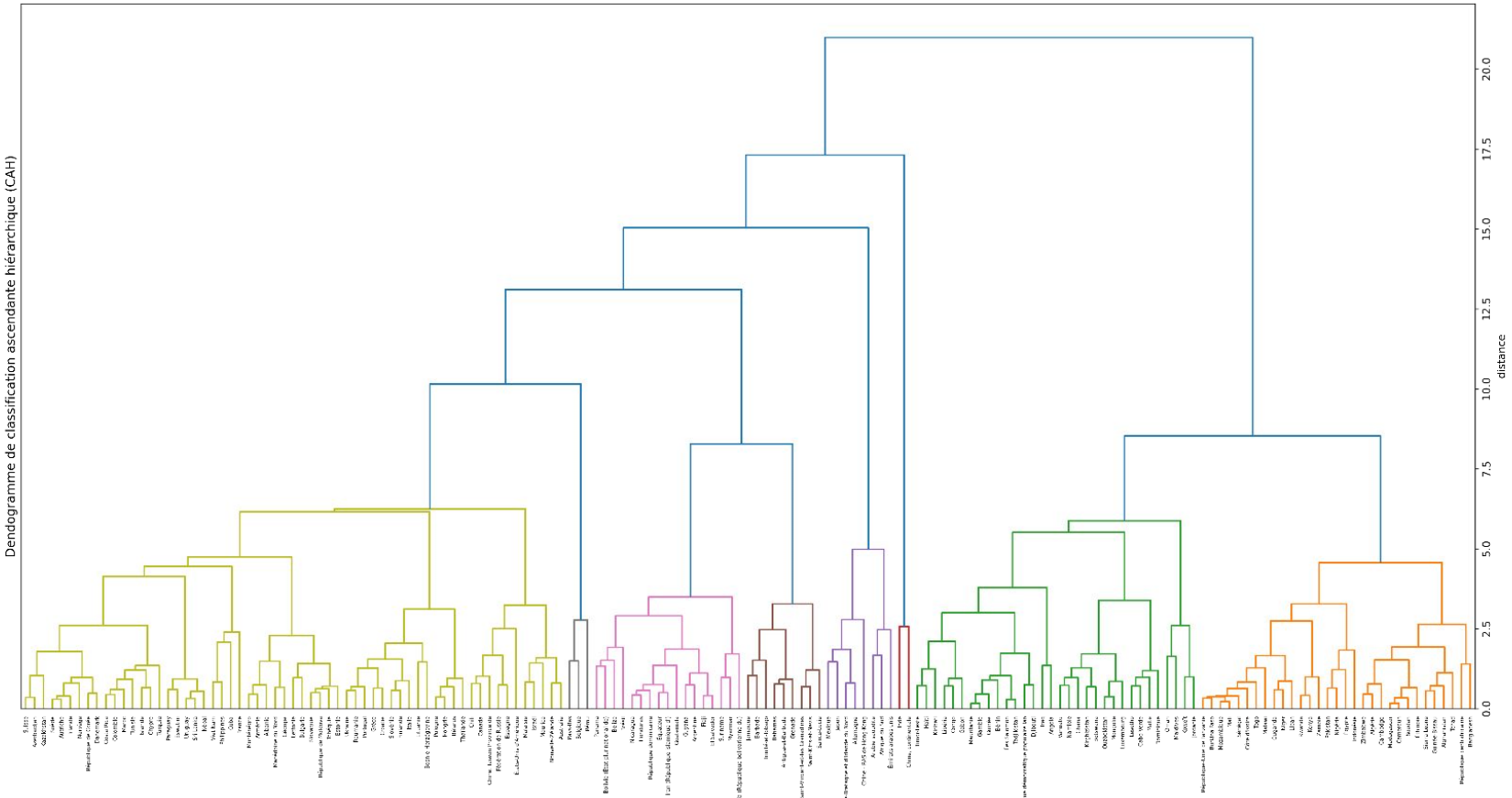


Clustering

Recherche du groupe qui se caractérise en ordre de priorité par :

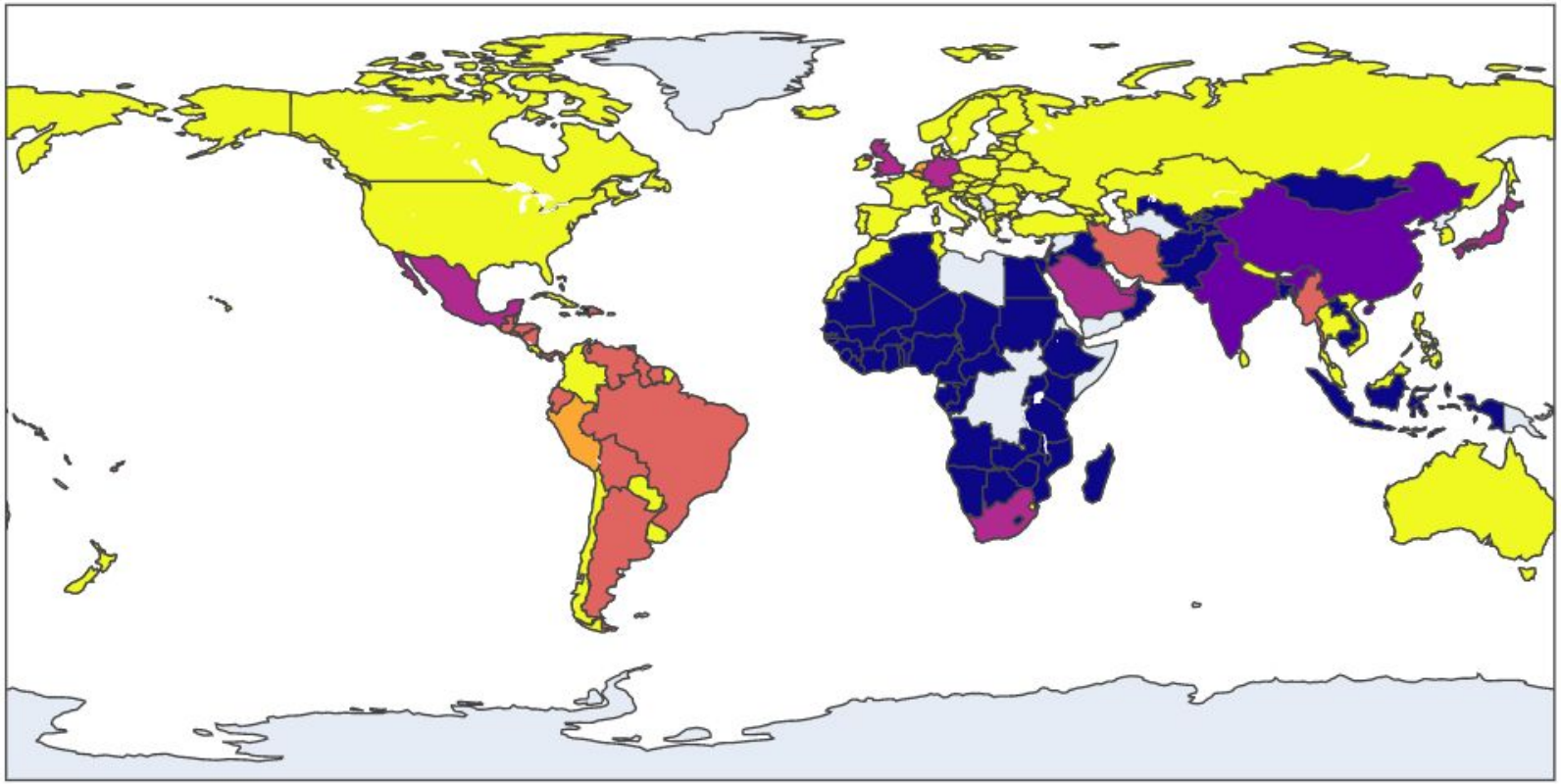
1. Quantité importante d'importation
2. Rapport production / nourriture bas
3. Score de facilité à faire des affaires élevées
4. Nombre de populations élevées
5. Taux de croissance démographique élevé
6. Rapport nourritureVollaile / nourritureTotale élevé

CAH - Dendrogramme - Méthode de Ward

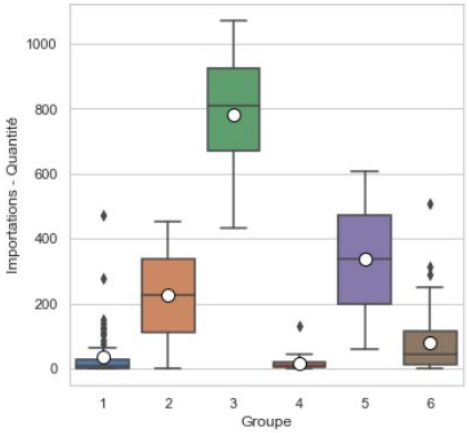


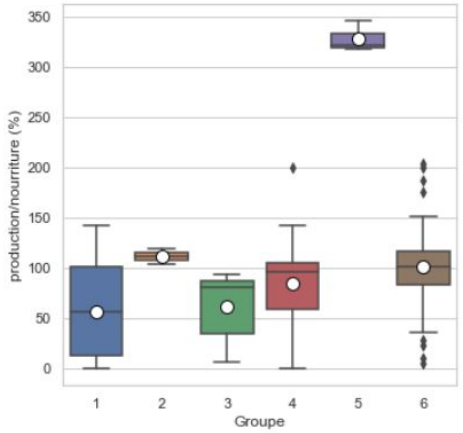
CAH

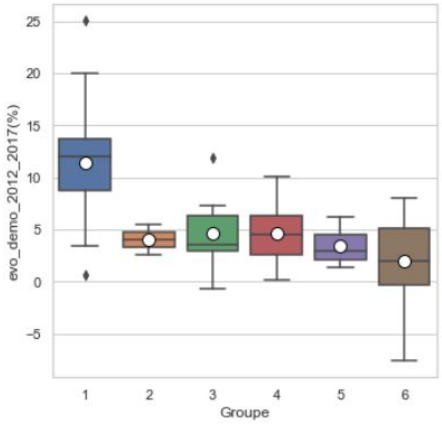
J’ai fait le choix de découper en 6 groupes:

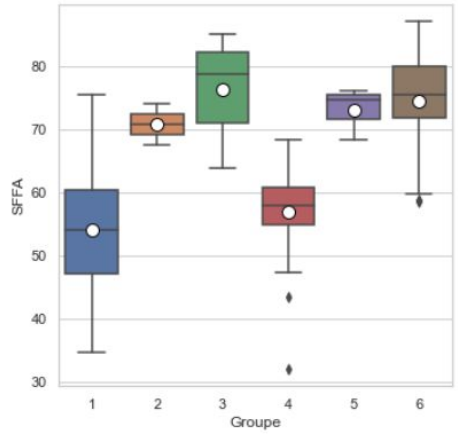


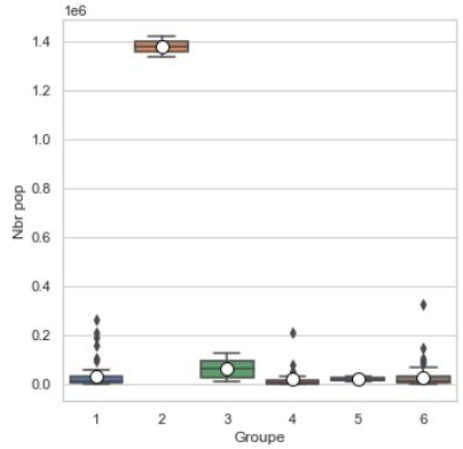
CAH - Distribution des variables par groupe

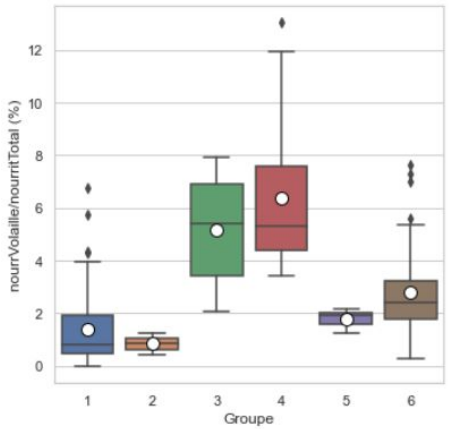




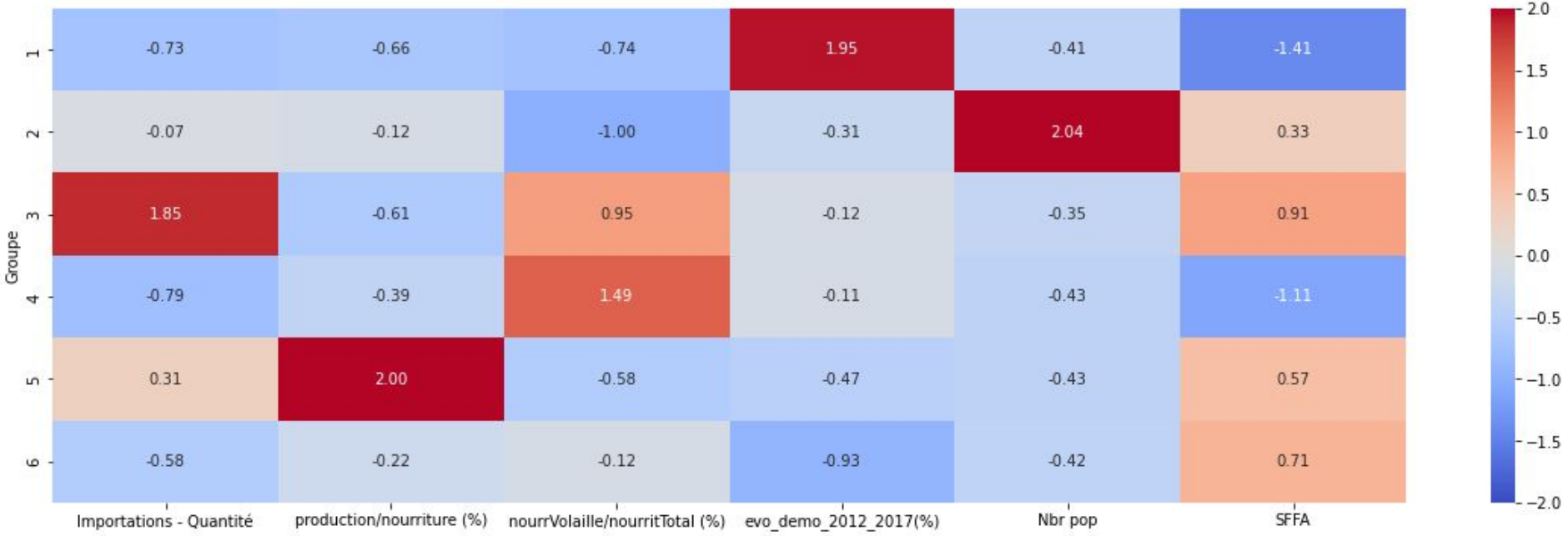








CAH

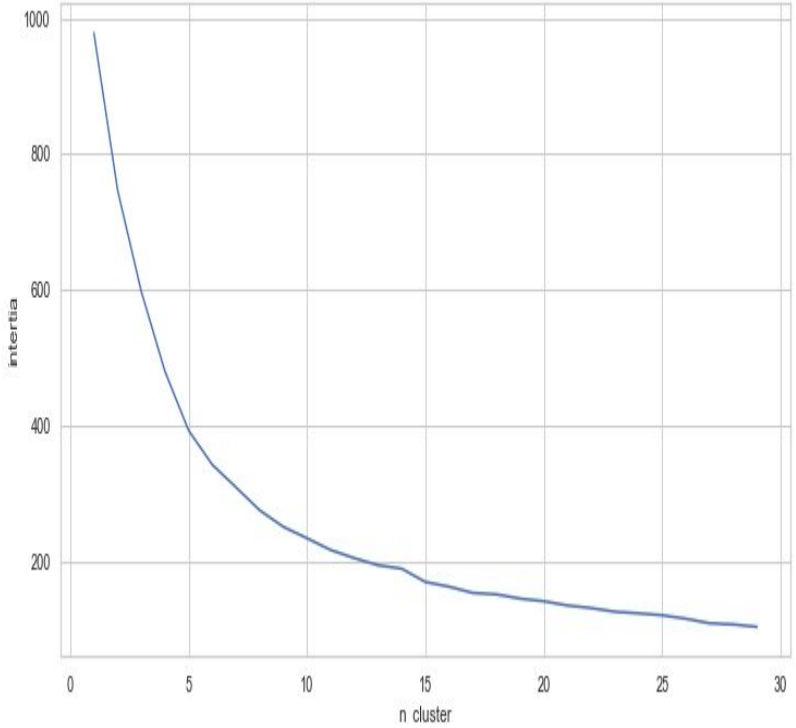


Le groupe qui répond le mieux à nos critères est le groupe 3 :

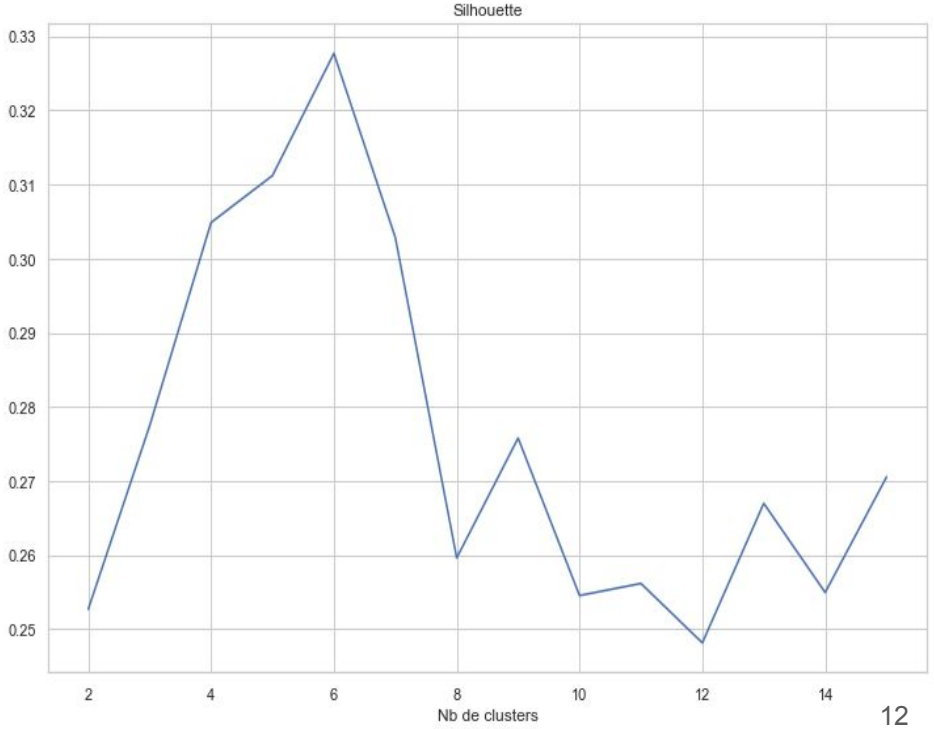
Afrique du Sud', 'Allemagne', 'Arabie saoudite', 'Chine - RAS de Hong-Kong', 'Japon', 'Mexique', 'Royaume-Uni de Grande-Bretagne et d'Irlande du Nord', 'Émirats arabes unis'.

K-means - Recherche et vérification du nombre de clusters

Méthode du coude

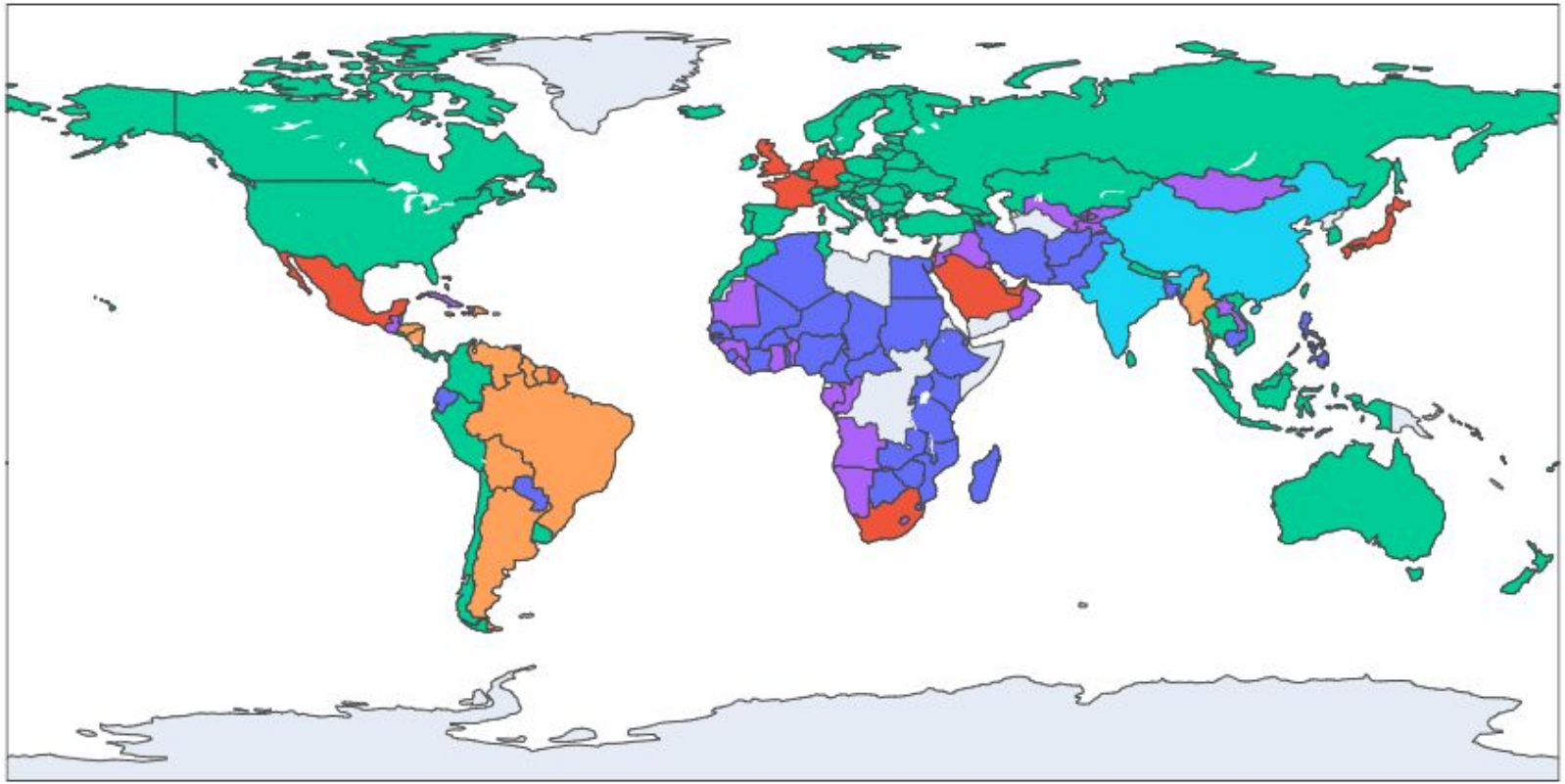


- Coefficient de silhouette

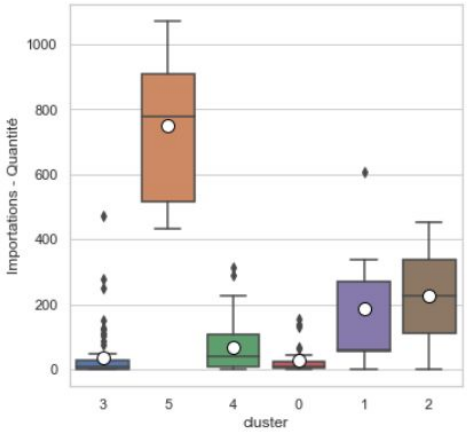


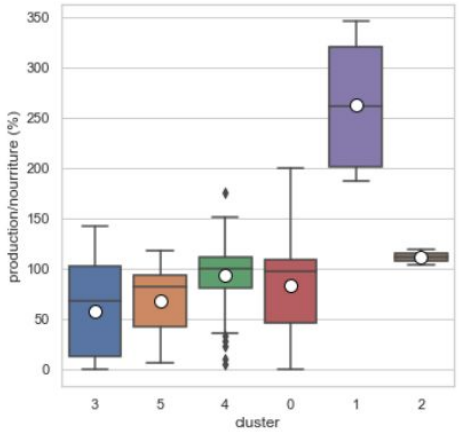
K-means

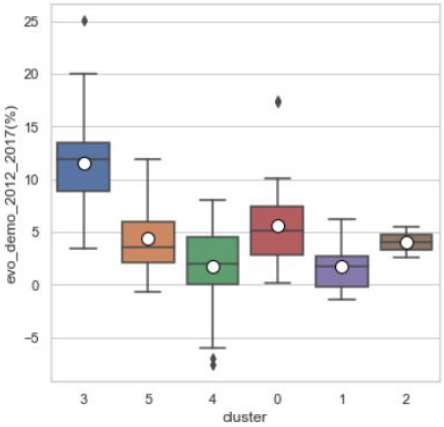
En couplant la méthode du coude et le coefficient de la silhouette, on décide de découper en 6 groupes :

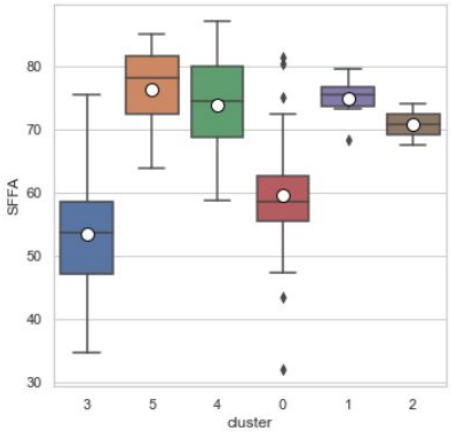


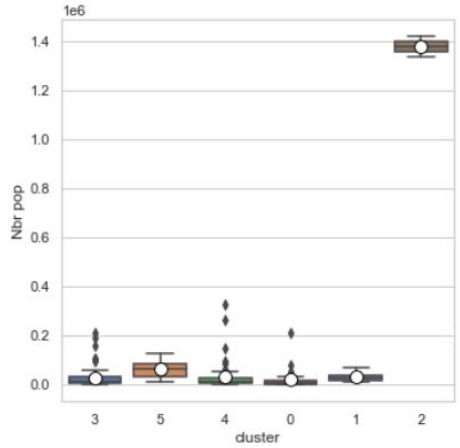
K-means - Distribution des variables par groupe

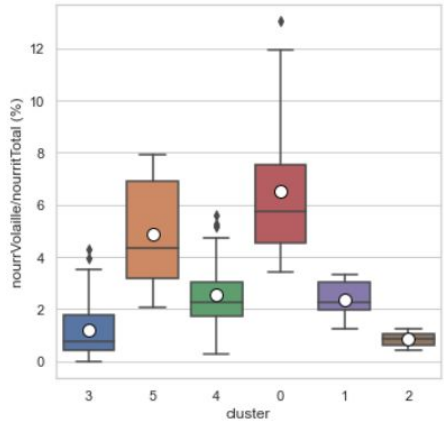




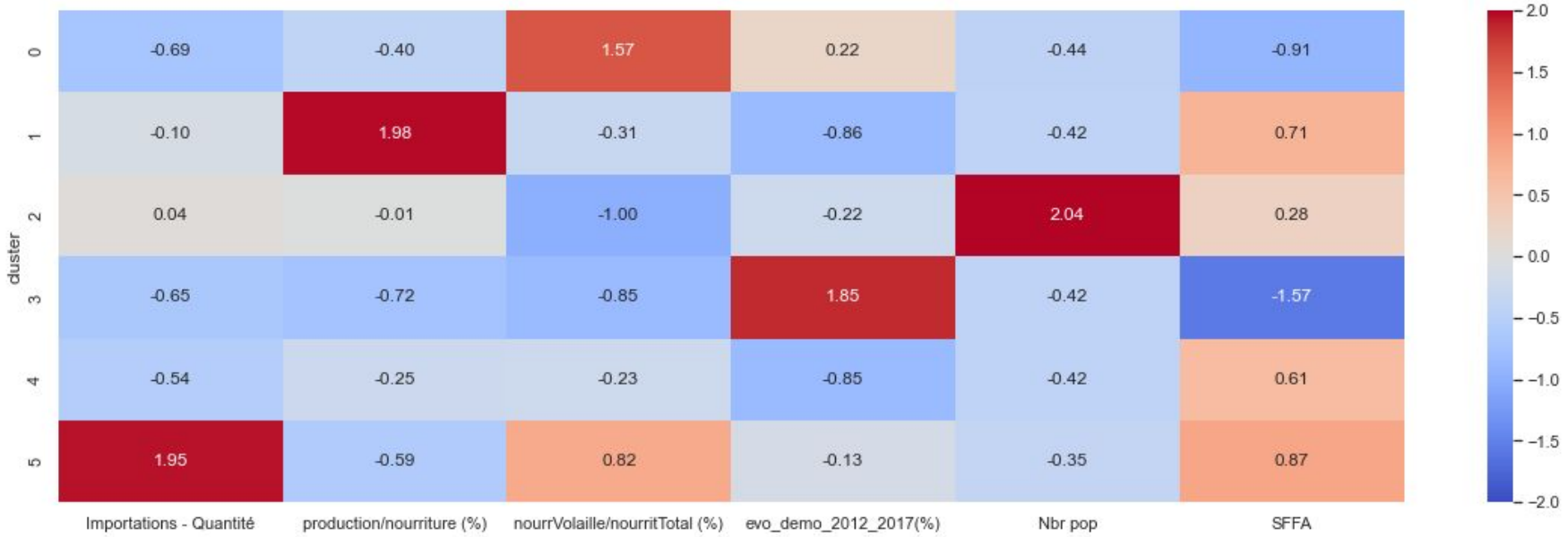








K-means



Le groupe qui répond le mieux à nos critères est le groupe 5 :

'Afrique du Sud', 'Allemagne', 'Arabie saoudite', 'Chine - RAS de Hong-Kong', 'France', 'Japon', 'Mexique', 'Royaume-Uni de Grande-Bretagne et d'Irlande du Nord', 'Émirats arabes unis'.

Comparaison des deux méthodes

Les correspondances ci-dessous montrent une similitude entre les deux approches.

Les deux méthodes sont très utilisées en clustering, cependant il n'est pas nécessaire de spécifier un nombre de clusters initiaux pour lancer l'algorithme avec la classification hiérarchique contrairement à la méthode K-means.

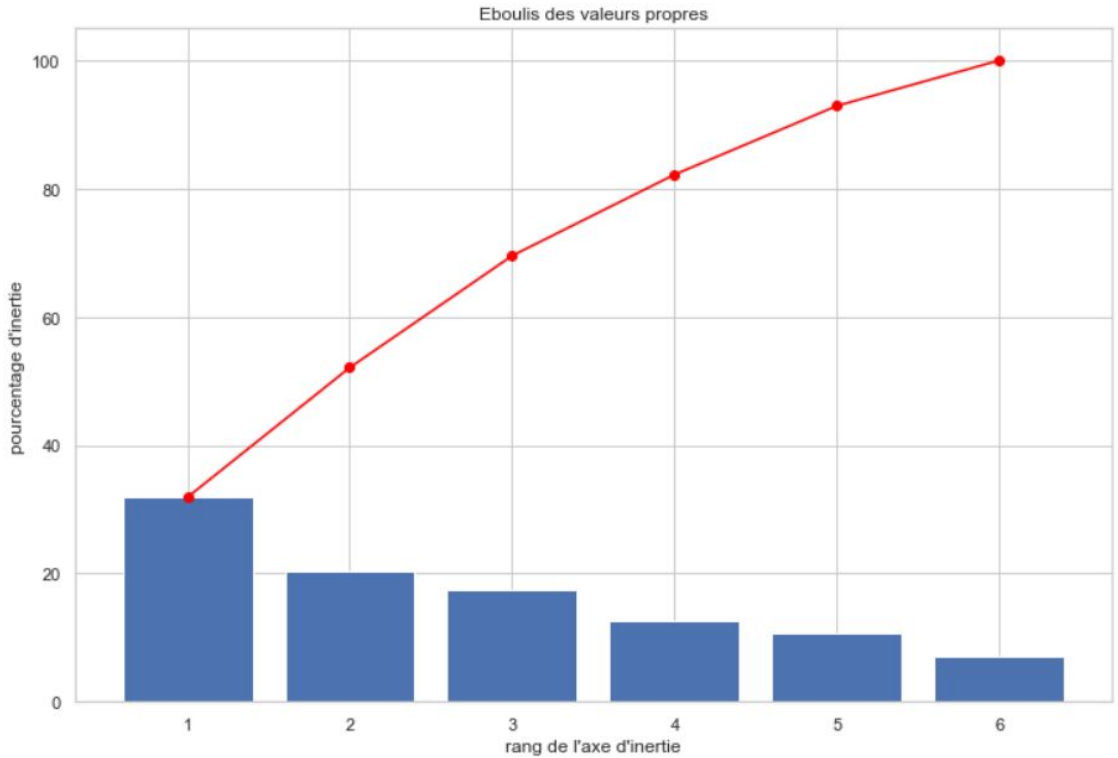
CAH

K-means

col_0	1	2	3	4	5	6
row_0						
0	3	0	0	24	0	4
1	0	0	0	0	3	3
2	0	2	0	0	0	0
3	58	0	0	0	0	2
4	2	0	0	2	0	51
5	0	0	8	0	0	1

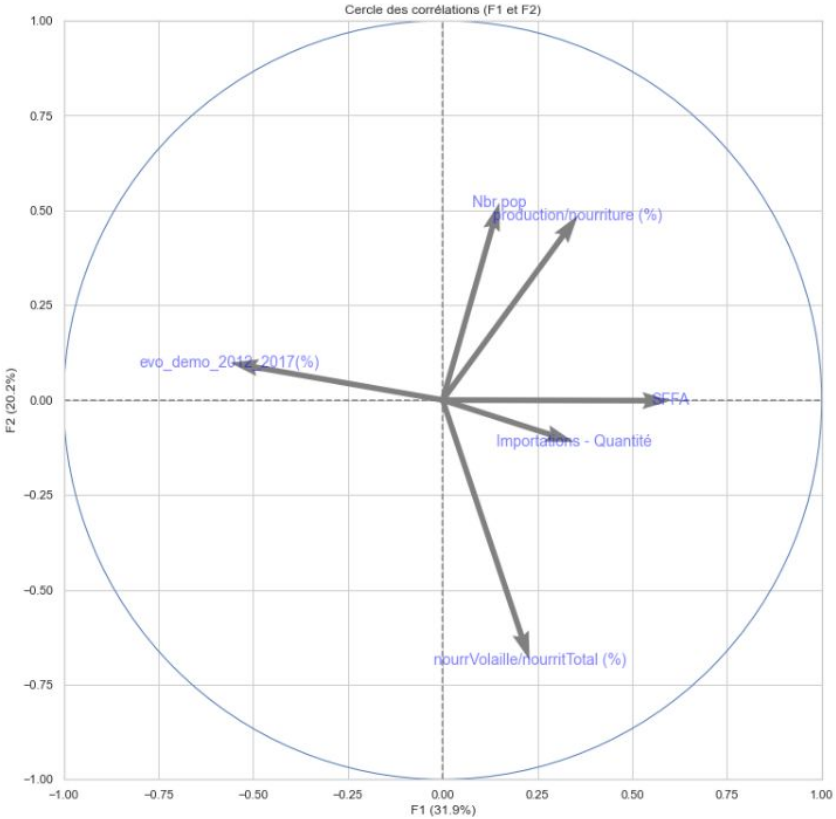
ACP

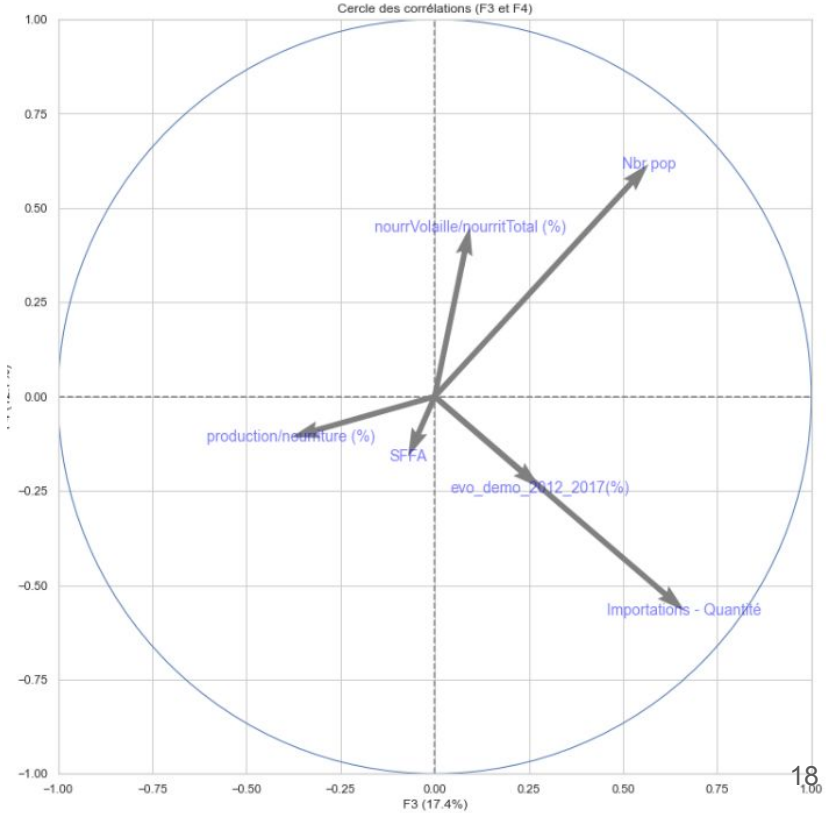
Diagramme de l’éboulis des valeurs propres:



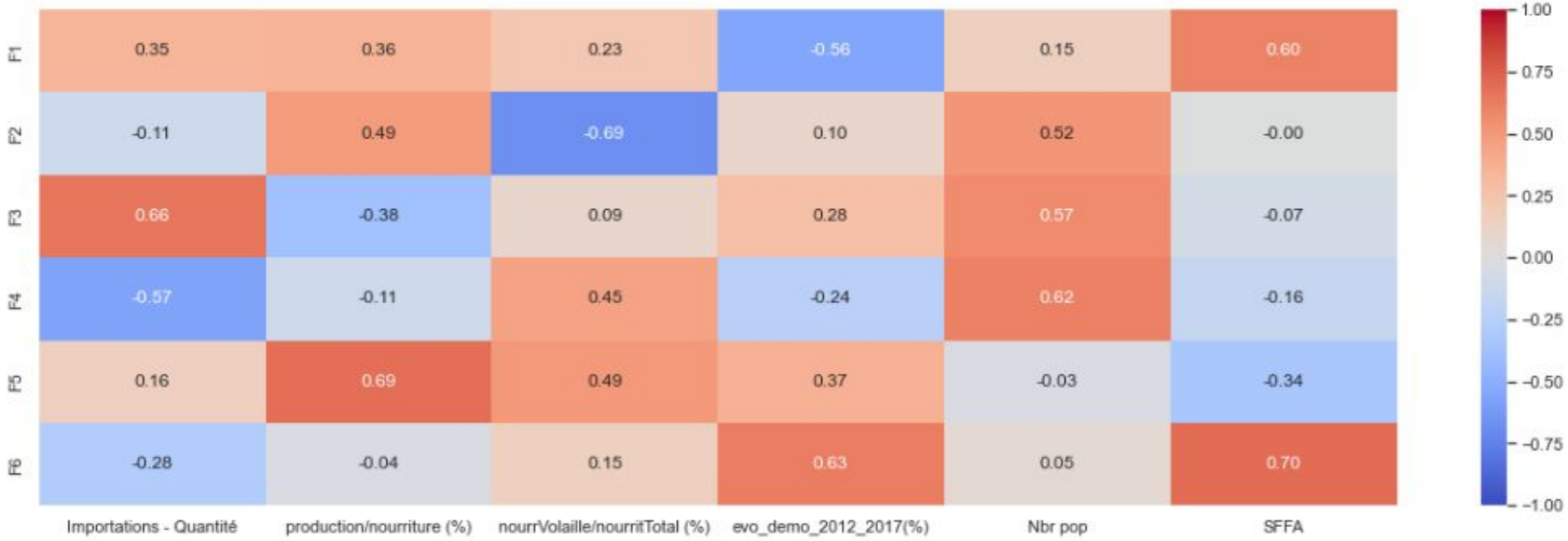
ACP

Cercle des corrélations:





ACP



F1 = représentatif des pays développés qui importent de grandes quantités de volailles

F2 = représentatif des pays dits émergents qui importent une moyenne quantité de volailles

F3 = représentatif des pays dits émergents qui importent de grandes quantités de volailles

F4 = représentatif des pays en voie de développement qui importent une faible quantité de volailles

Conclusion

On peut observer sur la projection des clusters avec la méthode CAH sur l'axe F1 et F2, un groupe vert, positif sur l'axe F1 et négatif sur l'axe F2

Ce cluster présente bien les critères suivants :

1. Quantité importante d'importation
2. Rapport “production / nourriture” bas
3. Score de facilité à faire des affaires élevées
4. Nombre de populations élevées
5. Taux de croissance démographique élevé
6. Rapport “nourritureVollaile / nourritureTotale” élevé

Dans ce groupe, on retrouve ces pays :

Afrique du Sud', 'Allemagne', 'Arabie saoudite', 'Chine - RAS de Hong-Kong', 'Japon', 'Mexique', 'Royaume-Uni de Grande-Bretagne et d'Irlande du Nord', 'Émirats arabes unis’.