ETUDE DE MARCHE

PASCAL BROCHART – Septembre 2024

CONTEXTE

- Entreprise d'agroalimentaire française spécialisé dans l'élevage et la vente de poulets
- Souhaite se développer à l'international
- Cibler les pays dont l'entreprise pourrait acquérir des parts de marché à l'exportation

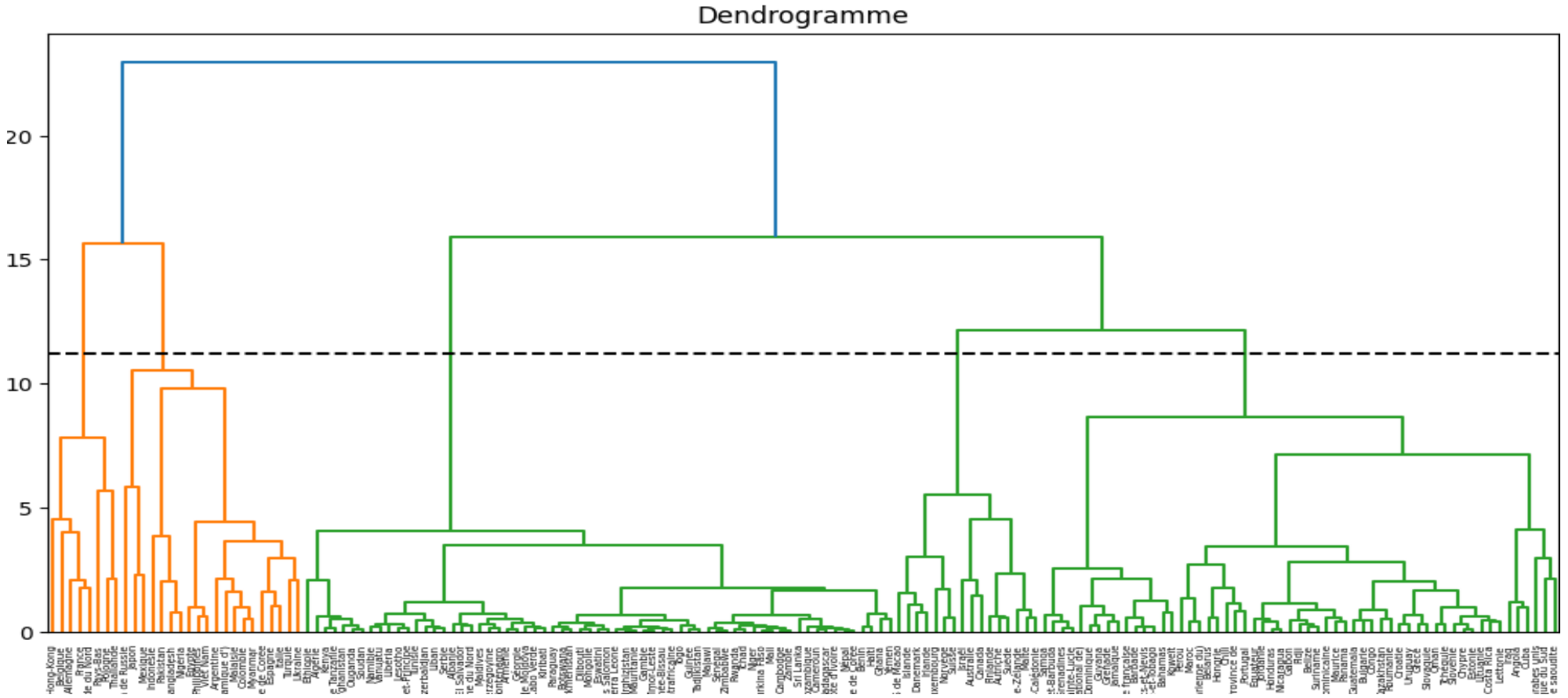
PRÉPARATION DES DONNÉES

- 3 fichiers de données (Population, Dispo alimentaire et FAOSTAT pour le PIB des pays)
- Aucune valeurs manquantes ou doublons à l'intérieur de chaque jeu de données
- Filtre sur la viande de volailles avec import/export, production et dispo alimentaire
- Suppression des colonnes inutiles
- Concaténation des 3 jeux de données
- Traitement de la population en millions d'habitants et calcul du PIB par habitant

NETTOYAGE DES DONNÉES

- Suppression des pays ou les données imports/export, production et dispo alimentaire sont manquantes après pivotement des données
- Ajustement de noms de pays et valeurs manquantes lorsque nécessaire
- Suppression des 4 pays les plus importants pour permettre une étude plus ciblée dans un périmètre moins concurrentiel et plus proche de la France
  - Etats-Unis d'Amérique, Brésil, Chine continentale, Inde

CLASSIFICATION ASCENDANTE HIÉRARCHIQUE



CLASSIFICATION ASCENDANTE HIÉRARCHIQUE

- Le dendrogramme ci-dessus permet de représenter un arbre de classification et de mettre en évidence une hiérarchisation des individus basée sur les données normalisées
- On place ici en pointillé une ligne permettant de définir combien de clusters seront utilisés

Cluster 0 : 20 pays
Argentine, Bangladesh, Colombie, Espagne, Fédération de Russie, Indonésie, Iran (République islamique d'), Italie, Japon, Malaisie, Mexique, Myanmar, Nigéria, Pakistan, Philippines, République de Corée, Turquie, Ukraine, Viet Nam, Égypte

Cluster 1 : 57 pays
Afrique du Sud, Angola, Antigua-et-Barbuda, Arabie saoudite, Bahamas, Barbade, Belize, Bolivie (État plurinational de), Bulgarie, Bélarus, Chili, Chine, Taiwan Province de, Chypre, Congo, Costa Rica, Croatie, Cuba, Dominique, Estonie, Fidji, Gabon, Grenade, Grèce, Guatemala, Guyana, Honduras, Hongrie, Iraq, Jamaïque, Jordanie, Kazakhstan, Koweït, Lettonie, Lituanie, Maroc, Maurice, Nicaragua, Oman, Panama, Polynésie française, Portugal, Pérou, Roumanie, République dominicaine, Saint-Kitts-et-Nevis, Saint-Vincent-et-les Grenadines, Sainte-Lucie, Samoa, Slovaquie, Slovénie, Suriname, Tchéquie, Trinité-et-Tobago, Uruguay, Venezuela (République bolivarienne du), Émirats arabes unis, Équateur

Cluster 2 : 65 pays
Afghanistan, Albanie, Algérie, Arménie, Azerbaïdjan, Bosnie-Herzégovine, Botswana, Burkina Faso, Bénin, Cabo Verde, Cambodge, Cameroun, Côte d'Ivoire, Djibouti, El Salvador, Eswatini, Gambie, Ghana, Guinée, Guinée-Bissau, Géorgie, Haïti, Kenya, Kirghizistan, Kiribati, Lesotho, Liban, Libéria, Macédoine du Nord, Madagascar, Malawi, Maldives, Mali, Mauritanie, Mongolie, Monténégro, Mozambique, Namibie, Niger, Népal, Ouganda, Paraguay, Rwanda, République centrafricaine, République de Moldova, République populaire démocratique de Corée, République-Unie de Tanzanie, Sao Tomé-et-Principe, Serbie, Sierra Leone, Soudan, Sri Lanka, Sénégal, Tadjikistan, Tchad, Timor-Leste, Togo, Tunisie, Turkménistan, Vanuatu, Yémen, Zambie, Zimbabwe, Éthiopie, Îles Salomon

Cluster 3 : 8 pays
Allemagne, Belgique, Chine - RAS de Hong-Kong, France, Pays-Bas, Pologne, Royaume-Uni de Grande-Bretagne et d'Irlande du Nord, Thaïlande

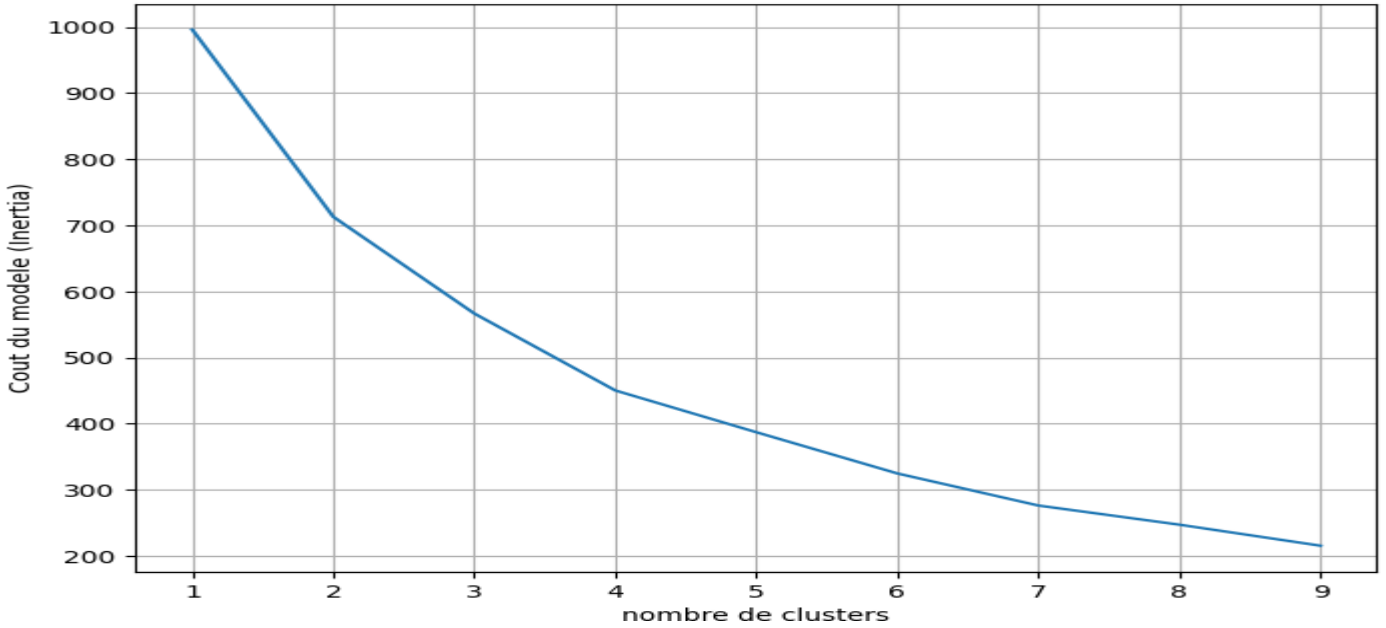
Cluster 4 : 16 pays
Australie, Autriche, Canada, Chine - RAS de Macao, Danemark, Finlande, Irlande, Islande, Israël, Luxembourg, Malte, Norvège, Nouvelle-Calédonie, Nouvelle-Zélande, Suisse, Suède

MÉTHODE K-MEANS

- La méthode de partitionnement K-Means, tout comme la CAH, est un algorithme non supervisé et vise à minimiser la somme des distances entre chaque individus et le centroïdes
- Le choix initial du nombre de centroïdes conditionne le résultat final
- Cet algorithme est itératif et permet de déplacer le centroïde jusqu'à la convergence, c'est-à-dire lorsque plus rien ne bouge entre deux itérations

MÉTHODE DU COUDE

- La méthode du coude permet de déterminer le nombre optimal de clusters à utiliser en affichant l'inertie interclasse obtenue
- On va s'intéresser particulièrement à une cassure dans la courbe qui nous permettra de déterminer où on va trop loin dans le nombre des clusters



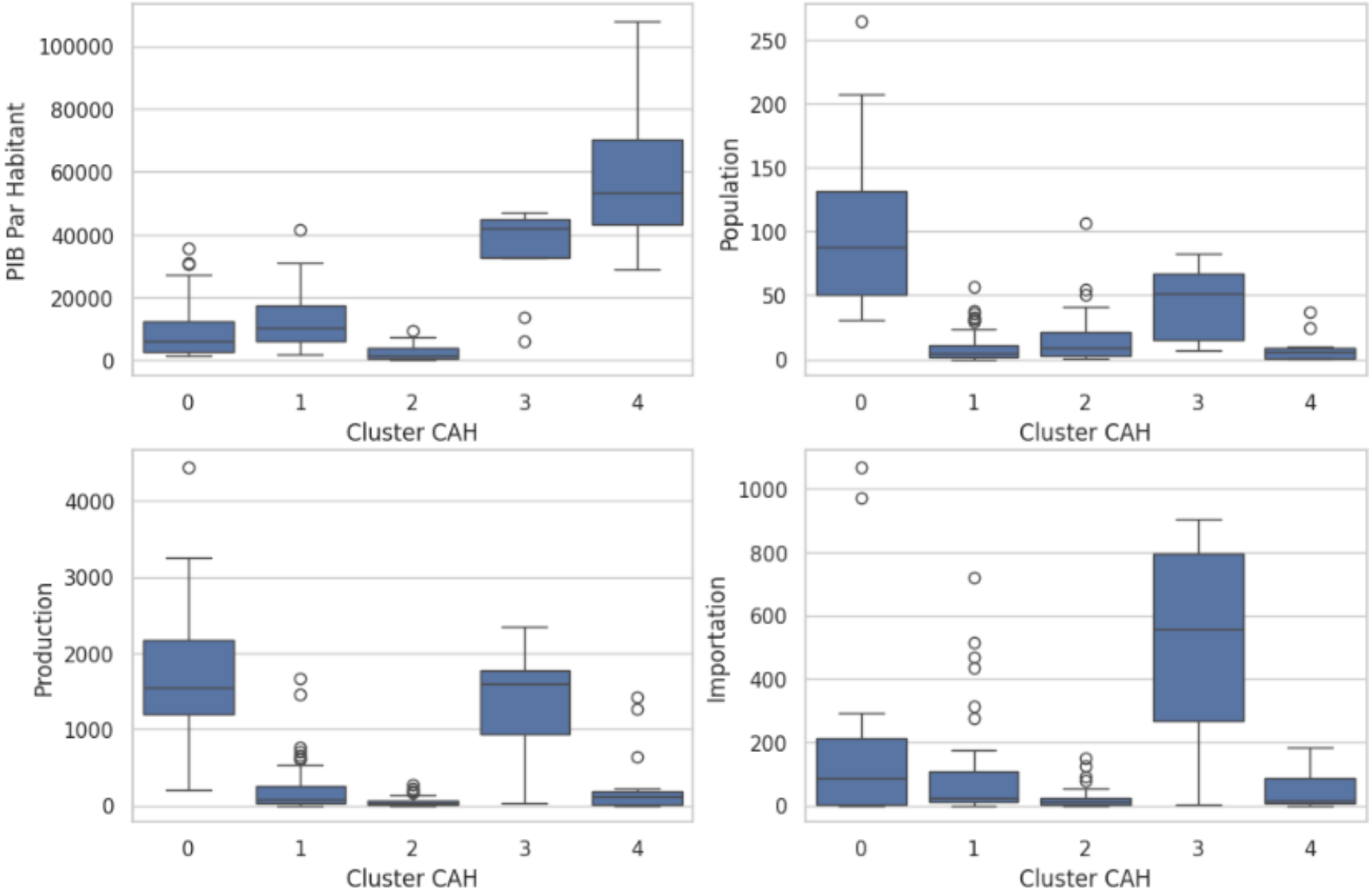
COMPARAISON DES CLUSTERS CAH ET K-MEANS

- Les graphiques ci-dessous illustrent les moyennes des différentes variables groupées par les clusters identifiés par la classification ascendante hiérarchique et par K-Means
- La répartition n'est pas identique mais on peut noter une forte similitude
- Les clusters en rouge sont strictement identiques entre les deux méthodes et ceux en vert sont très proches

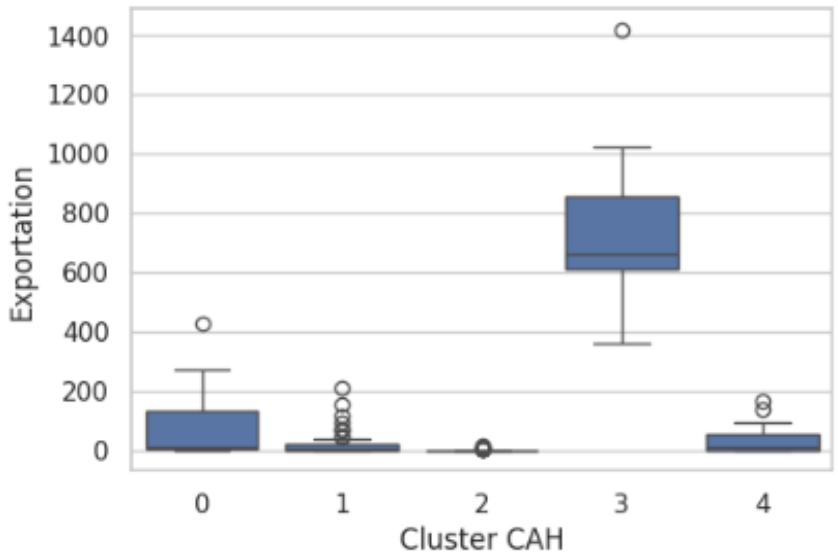
cluster_cah	Disponibilité alimentaire (Kcal/personne/jour) Viande de Volailles 2017	Exportations - Quantité Viande de Volailles 2017	Importations - Quantité Viande de Volailles 2017	Population totale millions Population-Estimations 2017	Production Viande de Volailles 2017	Valeur US $, aux prix du 2015 Produit Intérieur Brut Par Habitant 2017	cluster_kmeans
0	76.100000	77.150000	182.300000	102.958593	1680.200000	10807.087505	2.000000
1	115.912281	19.508772	88.736842	9.661815	215.105263	12266.496191	2.649123
2	24.569231	0.769231	17.738462	14.541804	44.830769	2462.517977	0.000000
3	95.125000	758.000000	504.625000	44.642348	1336.500000	35138.415877	3.000000
4	112.562500	36.312500	46.750000	7.812526	279.062500	57083.730541	1.562500

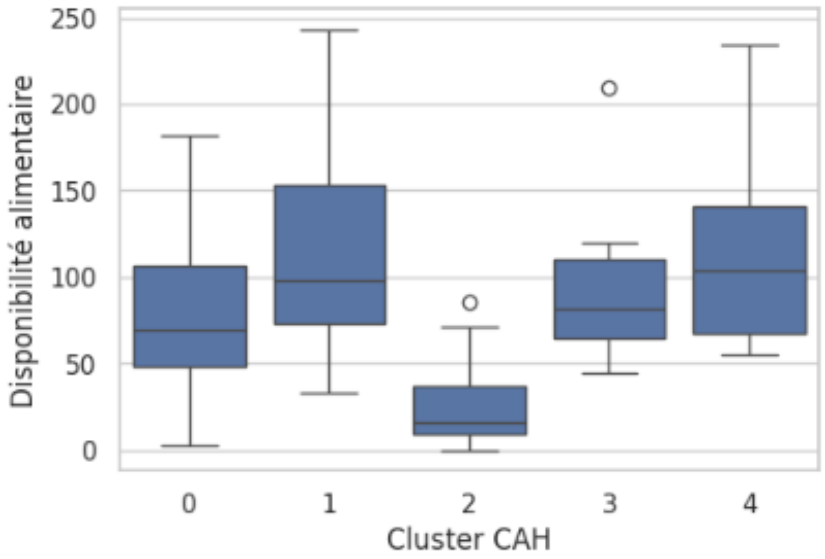
cluster_kmeans	Disponibilité alimentaire (Kcal/personne/jour) Viande de Volailles 2017	Exportations - Quantité Viande de Volailles 2017	Importations - Quantité Viande de Volailles 2017	Population totale millions Population-Estimations 2017	Production Viande de Volailles 2017	Valeur US $, aux prix du 2015 Produit Intérieur Brut Par Habitant 2017	cluster_cah
0	32.402439	3.634146	39.500000	14.668429	74.951220	3994.739072	1.792683
1	105.357143	48.000000	83.071429	8.966335	277.071429	60993.243532	3.785714
2	81.652174	70.304348	214.869565	94.814222	1624.000000	10845.756404	0.130435
3	95.125000	758.000000	504.625000	44.642348	1336.500000	35138.415877	3.000000
4	141.461538	17.923077	32.333333	4.385797	150.435897	14302.794346	1.230769

DISTRIBUTION DES VARIABLES



DISTRIBUTION DES VARIABLES

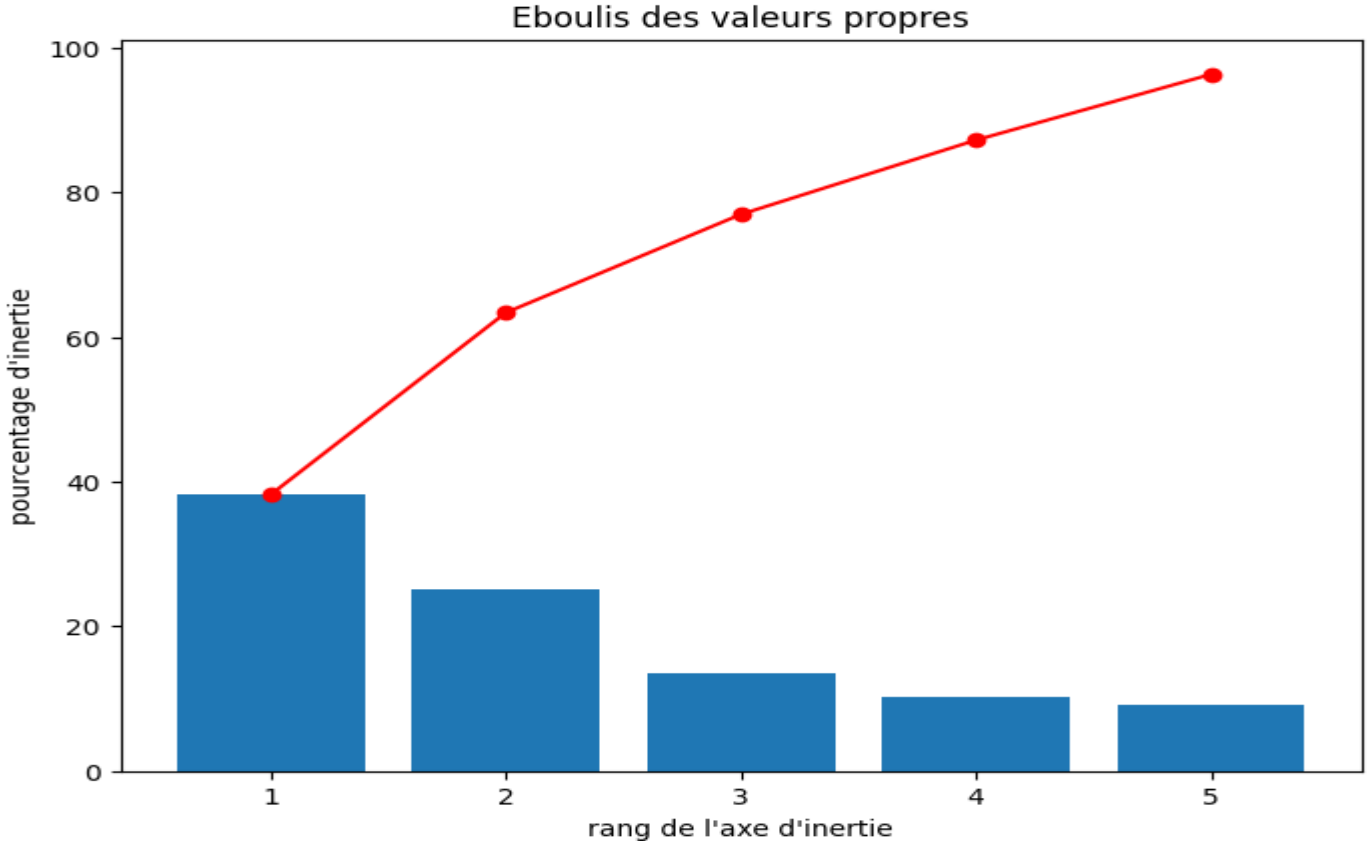




- Les graphiques ci-dessus représentent la répartition des variables par cluster
- On remarque dans la page précédente un PIB par habitant assez élevé pour le cluster 3 ainsi qu'une population moyenne
- En revanche le cluster 0 a une population bien plus importante mais un PIB par habitant faible

ANALYSE EN COMPOSANTES PRINCIPALES

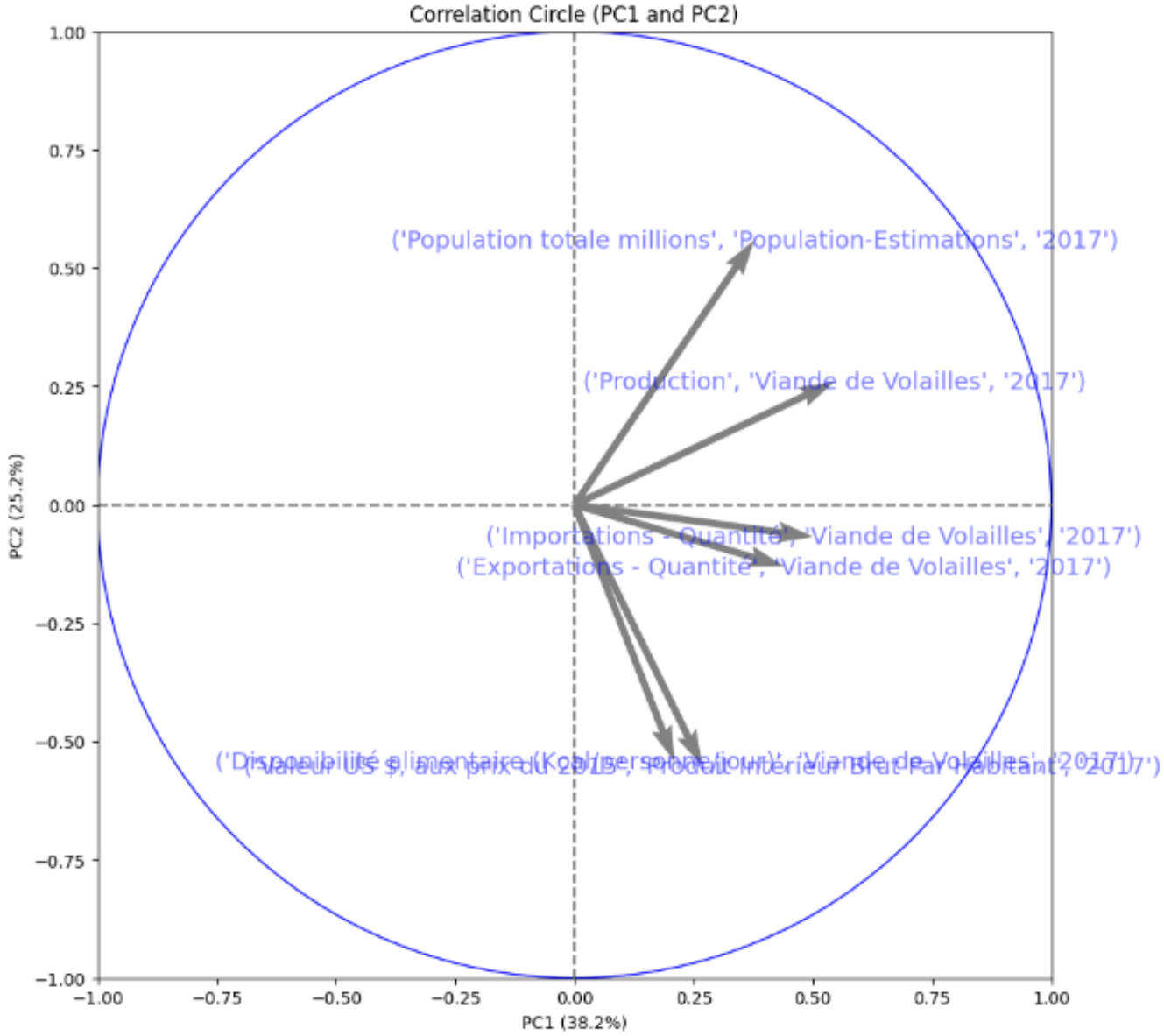
- L'ACP permet de réduire les dimensions d'un jeu de données dans un nouvel espace



ANALYSE EN COMPOSANTES PRINCIPALES

- Le graphique ci-dessus permet de représenter le pourcentage de données projetées sur les axes principaux et s'appelle « le diagramme d'éboulis de valeurs propre »
- Ainsi en choisissant cinq composantes nous obtenons presque 100% de la somme cumulée des inerties visualisable grâce à la courbe rouge
- Nous allons effectuer une ACP sur deux composantes ce qui représentent un peu plus de 60%
- Permettre la création de graphiques en 2D

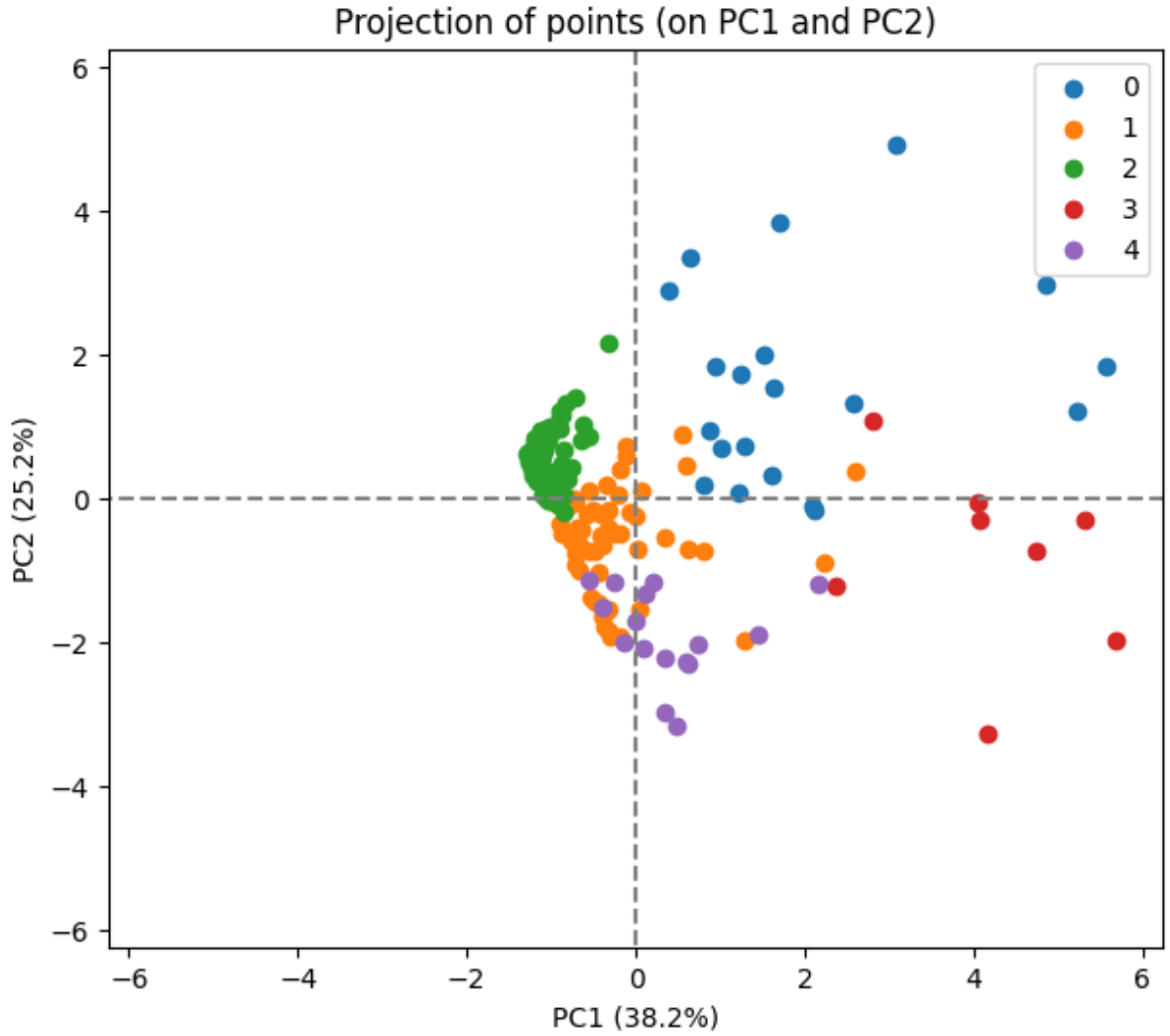
CERCLE DES CORRÉLATIONS



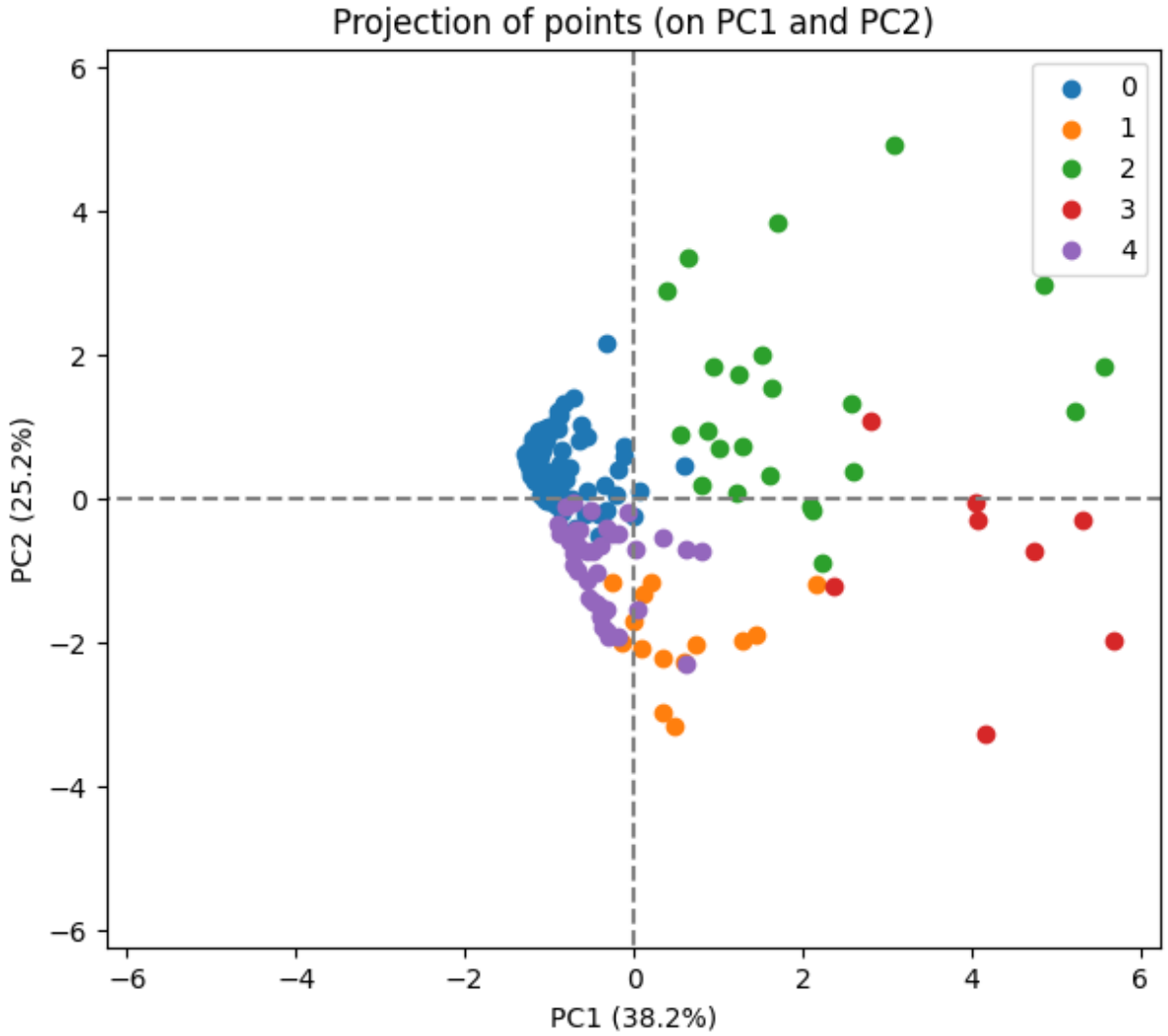
CERCLE DES CORRÉLATIONS

- Le cercle de corrélations ci-dessus permet d'identifier les corrélations des différentes variables avec les deux composantes de l'ACP représentées par PC1 et PC2
- On observe une forte corrélation de la production, importation et exportation avec PC1
- PC1 représente en quelque sorte des flux économiques
- Pour PC2 on observe une corrélation négative du PIB par habitant et de la disponibilité alimentaire et une corrélation positive de la population
- Ce qui signifie en d'autres termes que plus la population augmente et plus le PIB par habitant et la disponibilité alimentaire diminuent
- PC2 représente un indice démographique et économique

PROJECTION DES INDIVIDUS AVEC LES CLUSTERS CAH



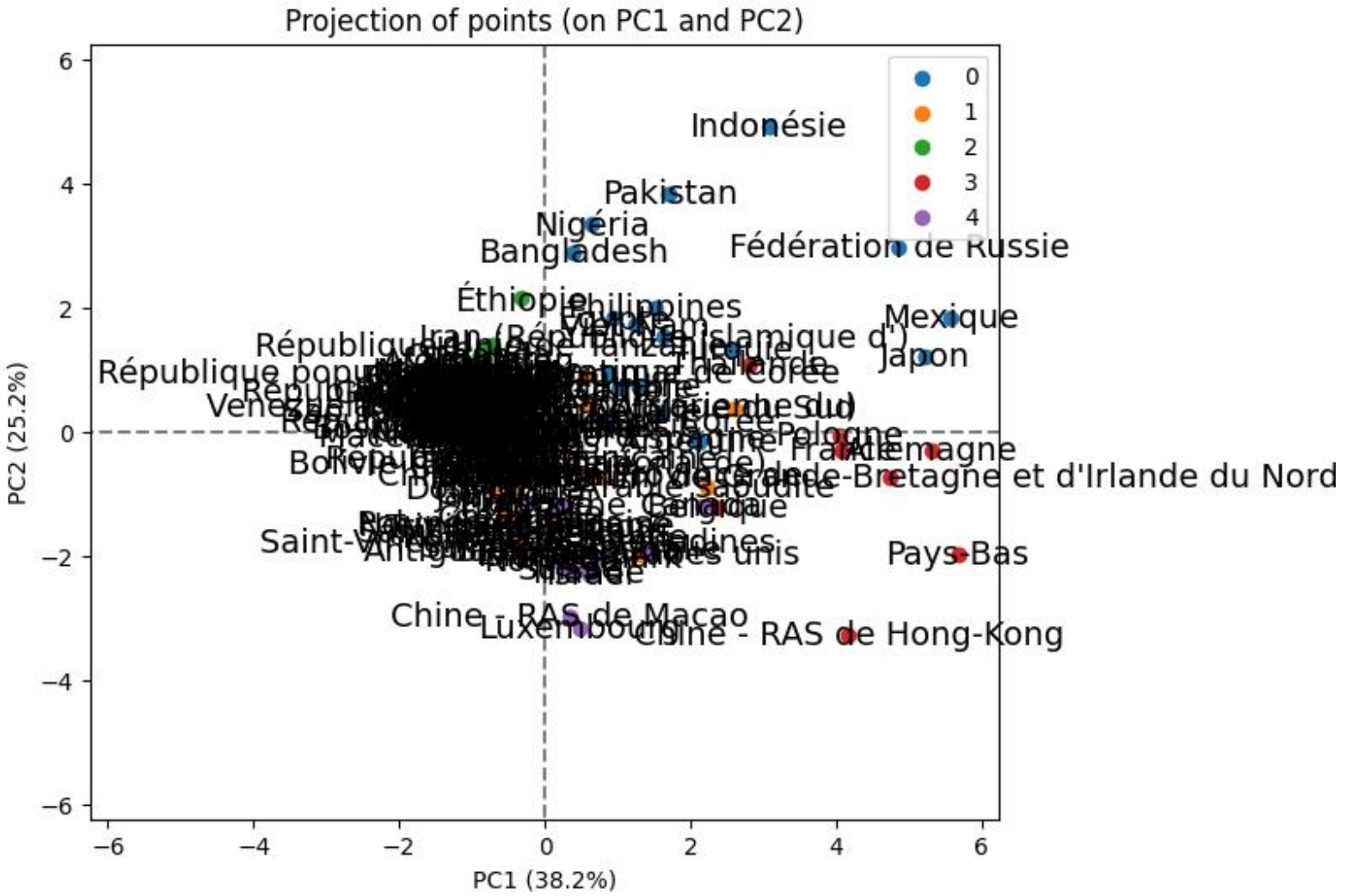
PROJECTION DES INDIVIDUS AVEC LES CLUSTERS K-MEANS



ANALYSE DES PROJECTIONS

- Les graphiques ci-dessus démontrent que la répartition des clusters avec les méthodes CAH et K-Means est relativement proche
- Nous voyons que les clusters les plus éloignés du centre sont ceux qui ont subi le moins de changements entre les deux méthodes, c'est-à-dire le cluster 3 et le cluster 0 pour CAH et 2 pour K-Means
- C'est assez logique car les individus près du centre sont très rapprochés et il devient difficile de garder la même répartition avec des algorithmes différents

PROJECTION DES INDIVIDUS AVEC LE NOM DE PAYS



ANALYSES ET LISTE DES PAYS RETENUS

- Le cluster en vert n'est pas très intéressant car la population est assez élevé avec un PIB par habitant assez faible
- Le cluster en rouge possède des avantages intéressants comme:
  - Un PIB par habitant assez élevé
  - Une population normale
  - La plupart de ces pays sont situés en zone euro et la logistique à mettre en place pour l'exportation est assez simple
  - Enfin le risque d'avoir une réglementation spécifique est quasi inexistant
- Les pays retenus sont:
  - Allemagne, Belgique, Pays-Bas, Royaume-Uni de Grande-Bretagne et d'Irlande du Nord