Produisez une étude de marché avec Python

Rémi Bardey - novembre 2025

Contexte & mission

Contexte

La Poule qui Chante : entreprise française de poulets BIO.

Activité uniquement en France, aucune présence export.

Le PDG veut explorer des opportunités d'exportation.

Aucun marché ciblé : toutes les zones sont ouvertes.

Besoin d'une analyse globale pour identifier les marchés les plus attractifs.

Mission

Identifier des groupes de pays pertinents pour l'export.

Collecter et sélectionner les données

Construire un dataset ≥ 8 variables, ≥ 100 pays, couvrant ≥ 60% de la population mondiale.

Réaliser une ACP, puis une CAH + K-means pour segmenter les pays.

Proposer des recommandations de pays et une méthodologie claire pour le COMEX

Sélection des données

Sources initiales

Population (2017–2022) – Évolution démographique sur 6 ans.

PIB (2017–2022) – Indicateur économique global sur la même période.

Stabilité politique (2017–2022) – Donnée géopolitique issue d'indicateurs internationaux.

Disponibilité alimentaire (2022) – Données FAO sur la production, importations, exportations et disponibilité.

Rôle de ces données

Constituent la base brute pour démarrer le projet.

Données volontairement larges, riches et hétérogènes pour permettre un travail complet :

Nettoyage, sélection, transformation, fusion dans un fichier unique.

Permettent de construire les variables clés nécessaires à l'analyse et au clustering.

Nettoyage & Préparation des données

Préparation & Standardisation

Suppression des colonnes inutiles pour réduire la complexité.

Vérification et gestion des valeurs manquantes (contrôle qualité).

Renommage des colonnes avec des intitulés explicites et uniformisés.

Vérification des types de données et correction si nécessaire.

Standardisation des unités pour rendre les valeurs comparables : milliers de tonnes ⟶ tonnes, population ⟶ millions, indicateurs économiques harmonisés.

Mise en cohérence des formats entre toutes les sources importées.

Restructuration & Enrichissement

Restructuration des données multi-annuelles (pivot) :

1 pays = 1 ligne, 1 année = 1 colonne.

Calcul d'indicateurs complémentaires : variations, tendances, ratios population/production, etc.

Harmonisation des noms de pays pour permettre la fusion.

Fusion progressive des fichiers nettoyés dans un dataset unique.

Vérifications finales de cohérence (doublons, valeurs aberrantes).

Obtention d'une base complète, homogène et prête pour ACP + clustering.

Fusion & finalisation du dataset

Fusions & Variables Initiales

Fusion des fichiers via la clé commune id_pays pour chaque source.

Harmonisation des noms de pays pour éviter les doublons ou les non-correspondances.

Contrôles de cohérence après fusion (doublons, types, valeurs manquantes).

Variables initiales conservées :

Disponibilité intérieure (tonnes), Exportations (tonnes), Importations (tonnes), Production (tonnes), Variation des stocks (tonnes), PIB 2022, Stabilité politique 2022

Variables Calculées & Dataset Final

Calculs ajoutés pour enrichir l'analyse :

% d'augmentation du PIB entre 2017 et 2022, Delta de stabilité politique (2017 → 2022), % d'augmentation de la population (2017 → 2022), Distance géographique entre la France et chaque pays

Vérification et nettoyage post-fusion pour éliminer les pays incomplets.

Création d'un dataset final propre, homogène et multivarié.

141 pays retenus après nettoyage, couvrant plus de 60% de la population mondiale → base robuste pour ACP & clustering. 12 variables.

Dataset final : extrait

pays	Dispo_int_tonne	Export_tonne	Import_tonne	Prod_tonne	Var_stock_tonne	pib_22	%_aug_pib_17_22	pol_sta_22	delta_polsta_17_22	pop_22	%_aug_pop_17_22	distance_km
Bulgarie	143000	44000	90000	113000	17000	32433.1	21.56	0.31	-0.01	6825864	-3.53	1837.144598
Burundi	8000	0	0	8000	0	829.4	-5.86	-1.19	0.81	13321097	15.77	6151.396386
Bélarus	336000	192000	16000	512000	0	26537.5	4.05	-0.8	-0.74	9173237	-3.4	1996.292762
Cambodge	31000	0	7000	26000	2000	6458.4	14.67	-0.1	-0.19	17201724	7.02	9934.941747
Cameroun	124000	1000	2000	123000	0	4843.7	0.95	-1.38	-0.29	27632771	14.52	4556.467295
Canada	1583000	151000	206000	1531000	2000	58321.1	3.02	0.78	-0.31	38821259	5.47	6069.215998
Cabo Verde	14000	0	13000	0	0	8850.1	7.47	0.93	0.17	519741	1.34	4141.058729
République centrafri	13000	0	7000	7000	0	1136.8	2.98	-2.21	-0.22	5098039	6.35	4706.168039
Sri Lanka	228000	1000	0	228000	0	13249.1	-9.08	-0.8	-0.72	22834965	3.38	8588.346624
Tchad	9000	0	3000	7000	0	2655.7	6.38	-1.47	-0.18	18455316	18.13	3800.05084
Chili	755000	188000	174000	769000	0	29569.5	6.01	0.13	-0.28	19553036	5.36	11193.97399
Colombie	1901000	0	81000	1820000	0	18458.7	8.67	-0.63	0.15	51737944	7.49	8476.929899
Comores	19000	0	14000	1000	-5000	3478.2	7.45	-0.23	-0.25	834188	10.24	7710.435771
Congo	165000	0	158000	7000	0	6205.1	-15.84	0.04	0.57	6035104	12.78	5380.290121
République démocra	101000	0	91000	10000	0	1385.5	10.56	-1.98	0.36	102396968	17.58	5585.249522
Costa Rica	169000	3000	20000	152000	0	24831.7	9.77	0.96	0.36	5081765	3.43	8924.002092
Croatie	79000	22000	34000	71000	4000	39861.8	23.21	0.67	-0.01	3907027	-4.22	1040.25475
Chypre	41000	1000	18000	27000	3000	51588.4	22.45	0.42	-0.11	1331370	6.15	2863.904463
Tchéquie	266000	34000	134000	172000	6000	48390.7	7.11	0.81	-0.18	10673213	1.25	1064.419415
Bénin	143000	0	133000	10000	0	3588.3	17.97	-0.34	-0.37	13759501	14.28	4070.834175

Analyse en Composante Principale - ACP

Pourquoi faire une ACP ?

Nous avons un dataset riche et multidimensionnel (production, commerce, population, PIB, stabilité, distance…).

Certaines variables sont fortement corrélées entre elles → elles racontent « la même histoire ».

L'ACP permet de réduire la complexité en regroupant les informations dans quelques axes simples à interpréter.

Elle facilite la visualisation globale des pays et met en évidence les grandes tendances.

Intérêt dans notre projet

Identifier les grands profils de pays : producteurs, importateurs, stables politiquement, en forte croissance, etc.

Repérer les pays qui se ressemblent ou qui se distinguent fortement.

Préparer le terrain pour un clustering plus pertinent (CAH + K-means).

Garantir que les regroupements reposent sur des caractéristiques réellement significatives.

Comment ça fonctionne ?

L'ACP crée de nouveaux axes (appelés composantes) qui résument l'information :

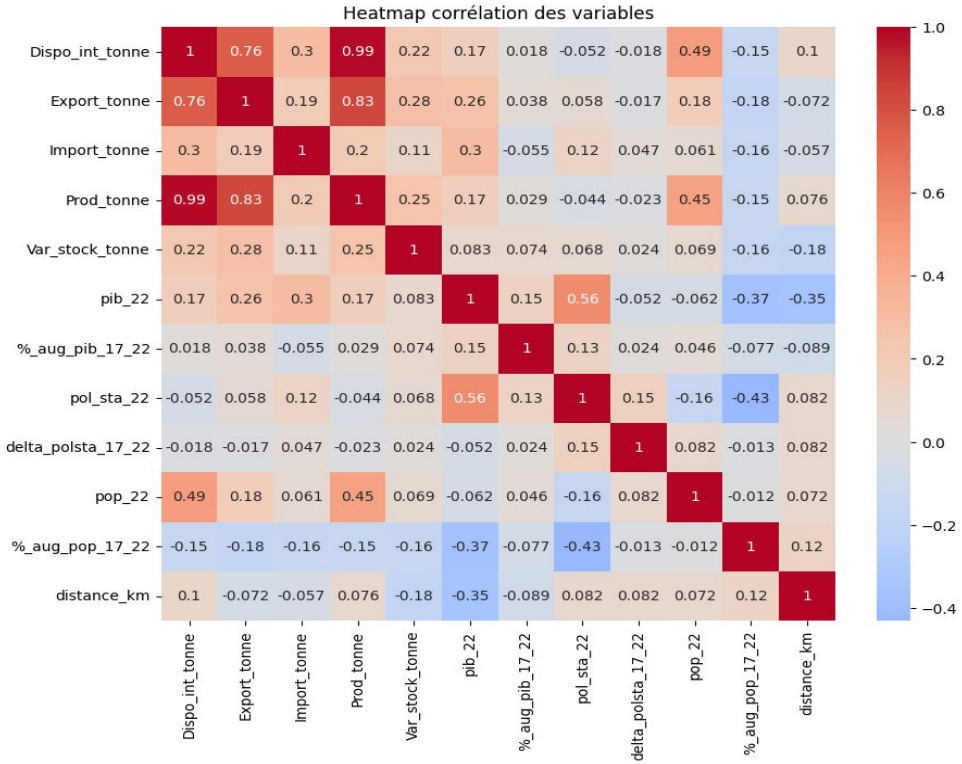
L'axe F1 = combinaison des variables qui expliquent le plus de différences entre les pays.

L'axe F2 = deuxième combinaison la plus informative, et ainsi de suite.

Chaque pays est ensuite projeté sur ces axes → cela permet une lecture visuelle rapide.

On obtient un cercle de corrélation (relations entre variables), une projection des pays

Heatmap corrélation des variables - ACP



Points clés de la heatmap de corrélation des Variables

Variables alimentaires (production, dispo interne, export) : très corrélées ⟶ mêmes dynamiques.

Importations : corrélation faible ⟶ comportement spécifique selon les pays.

PIB 2022 ↔ stabilité politique : lien modéré.

Population : corrélée à production/disponibilité ⟶ effet taille des pays.

Distance : très faible corrélation ⟶ variable indépendante.

Variables d'évolution (croissance PIB, population, politique) : peu corrélées ⟶ apportent une info complémentaire.

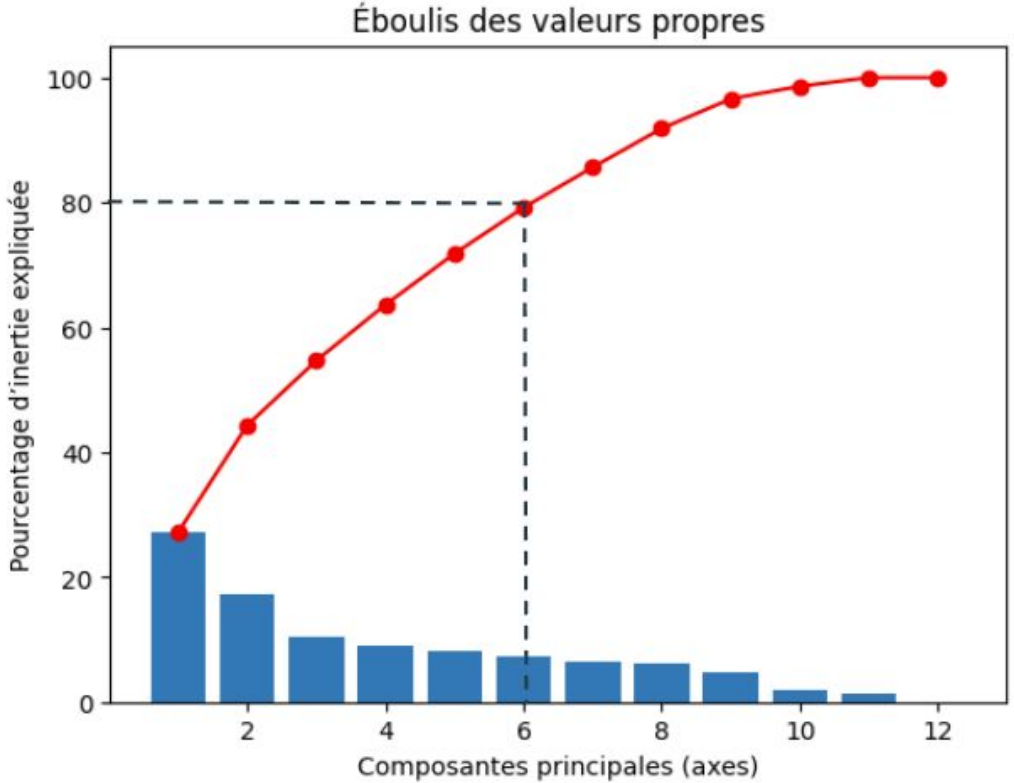
Éboulis des valeurs propres - ACP

Les 6 premières composantes expliquent ≈ 80 % de l'information → dataset riche et structuré.

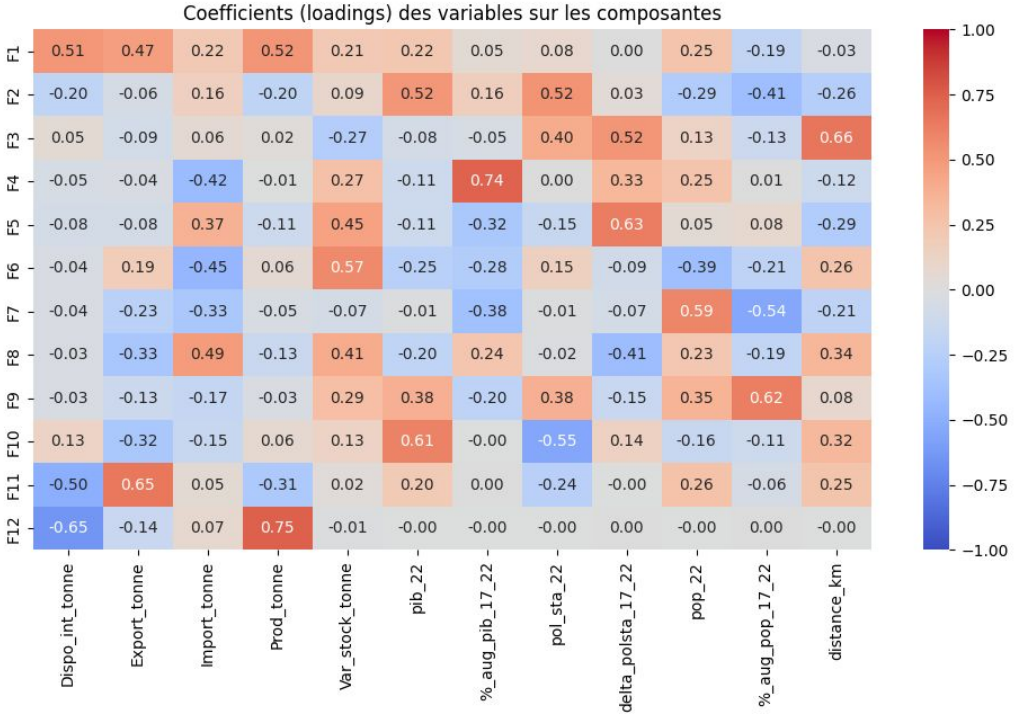
F1–F2 concentrent l'essentiel des tendances → base suffisante pour visualiser les relations entre variables et différencier les pays.

Les axes suivants (F3+) apportent une information plus fine mais moins déterminante.

L'analyse se focalise donc sur F1–F2 (et F3 - F4 si nécessaire) pour interpréter la structure des données avant de détailler les contributions des variables.



Analyse des loadings - ACP



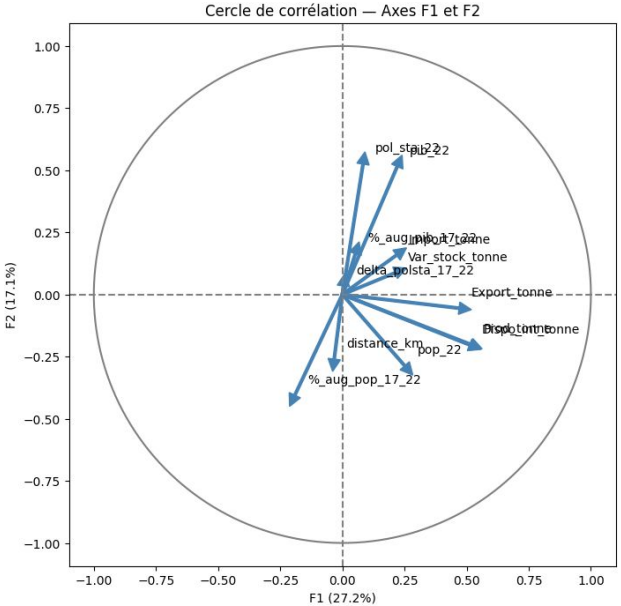
F1 – Axe alimentaire
Porté par la production, la disponibilité interne et les exportations. Oppose pays fortement productifs à pays dépendants des importations.

F2 – Axe économique & politique
Construit par le PIB, la stabilité politique et la variation du PIB. Oppose pays stables/économiquement solides aux pays plus fragiles.

F3 – Axe démographique
Influencé par la population totale et la croissance démographique. Met en évidence les marchés en expansion vs stagnants.

F4 – Axe “dynamique récente”
Porté surtout par la variation des stocks, et par les évolutions du PIB et de la stabilité politique.
Il reflète des changements récents plutôt que des niveaux absolus.

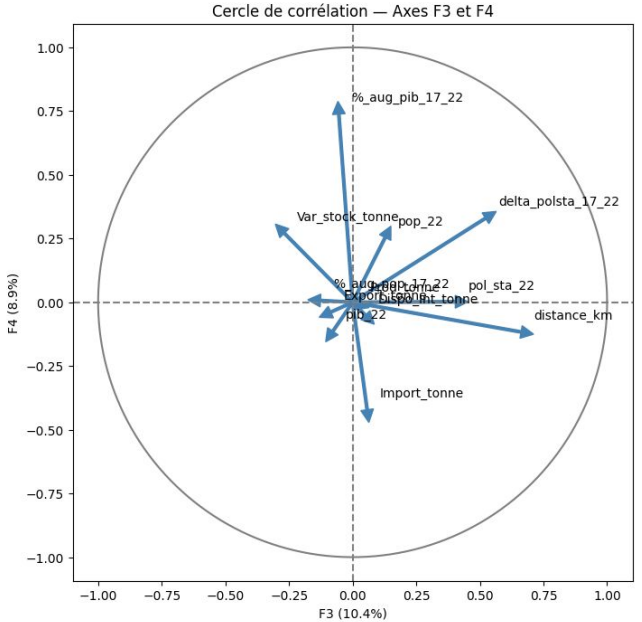
Cercles de corrélation - ACP



F1–F2 : axes principaux (≈ 45 % de l'info)

F1 oppose les pays producteurs/exportateurs aux pays plus dépendants.

F2 sépare les pays selon leur niveau économique et politique (PIB, stabilité). Ces deux axes révèlent les grandes dynamiques du marché mondial.



F3–F4 : axes secondaires (≈ 19 % de l'info)

F3 met en avant la démographie : taille de population, croissance.

F4 met en avant surtout les évolutions, Variation des stocks + évolutions du PIB et de la stabilité politique.

Projection des pays sur F1 F2 - ACP

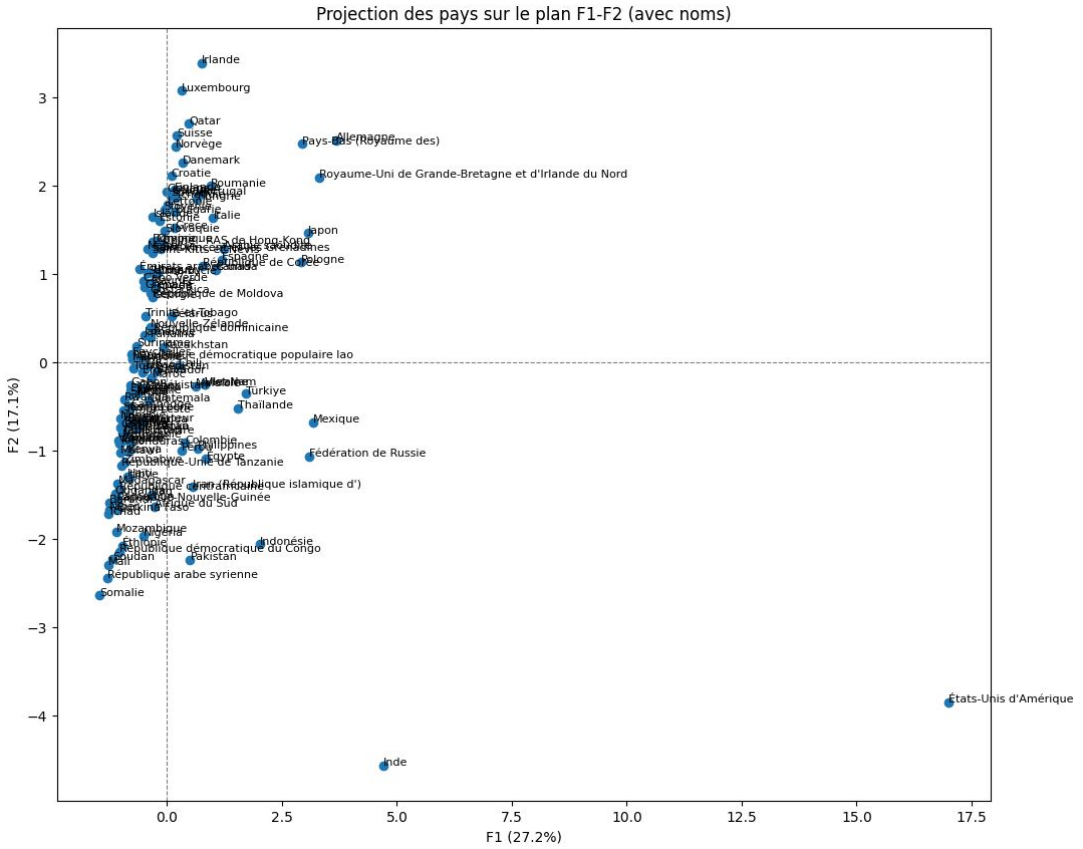
La projection permet de positionner les pays selon les deux axes majeurs de l'ACP.

F1 oppose pays producteurs/exportateurs et pays moins autonomes.

F2 distingue pays stables économiquement/politiquement des pays plus fragiles.

On observe : un groupe central de pays aux profils similaires, quelques pays atypiques très éloignés (ex : USA, Inde, Irlande).

Cette vue donne une première lecture des profils de pays avant le clustering.



Classification Ascendante Hiérarchique - CAH

CAH c'est quoi ?

Méthode de regroupement qui assemble les pays par similarité.

Chaque pays commence seul, puis la CAH les regroupe progressivement, du plus proche au plus éloigné.

Le résultat forme une arborescence qui montre comment les pays se rapprochent.

Pourquoi l'utiliser ?

Pour identifier des groupes naturels de pays sans fixer le nombre de clusters à l'avance.

Pour repérer les pays très proches et ceux totalement atypiques.

Pour déterminer visuellement le nombre optimal de clusters avant d'appliquer le K-means.

Comment l'interpréter ?

La CAH produit un dendrogramme : un arbre qui représente les étapes de regroupement.

Deux pays qui se rejoignent tôt sont similaires ; ceux qui se rejoignent tard sont très différents.

La hauteur à laquelle on "coupe" l'arbre permet de définir les clusters finaux.

Dendrogramme - CAH

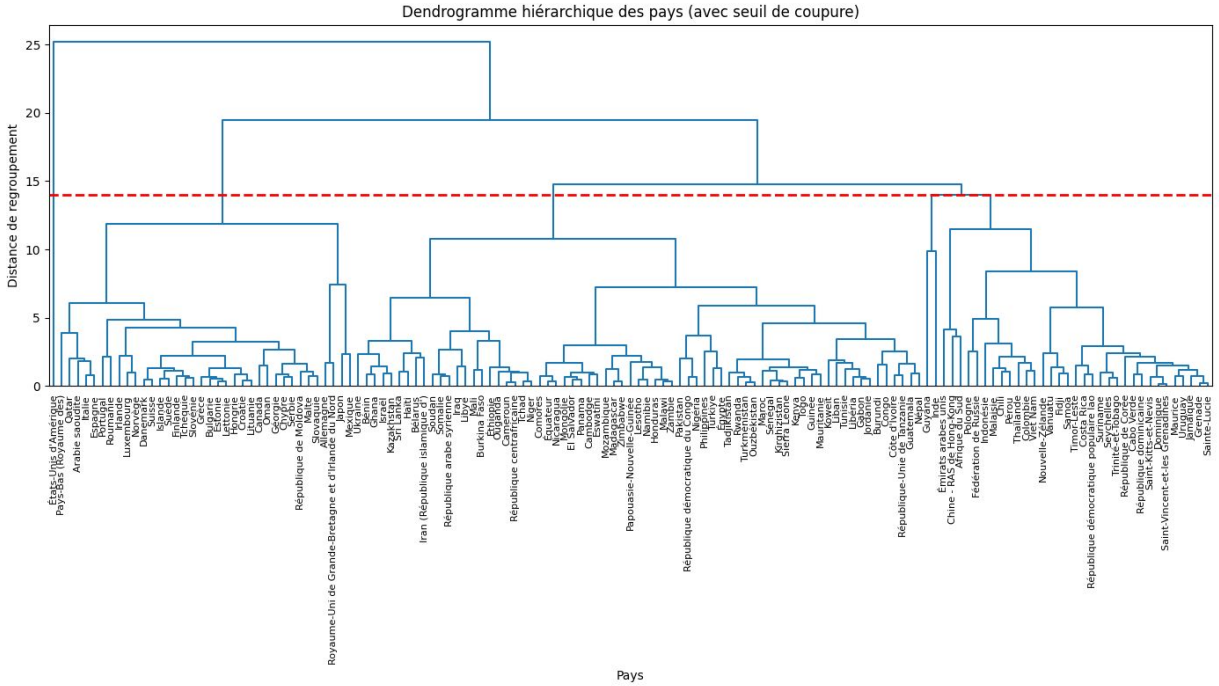
Le dendrogramme montre comment les pays se regroupent selon leur similarité.

Le seuil retenu fait apparaître 5 groupes cohérents, en ligne avec l'ACP.

Certains pays fusionnent très tôt → profils proches (production, stabilité, population…).

D'autres fusionnent très tard → profils atypiques (USA, Inde, Irlande, Luxembourg).

Le dendrogramme confirme des familles de pays bien distinctes, justifiant l'étape suivante : K-means pour affiner les clusters.



Profils des clusters - CAH

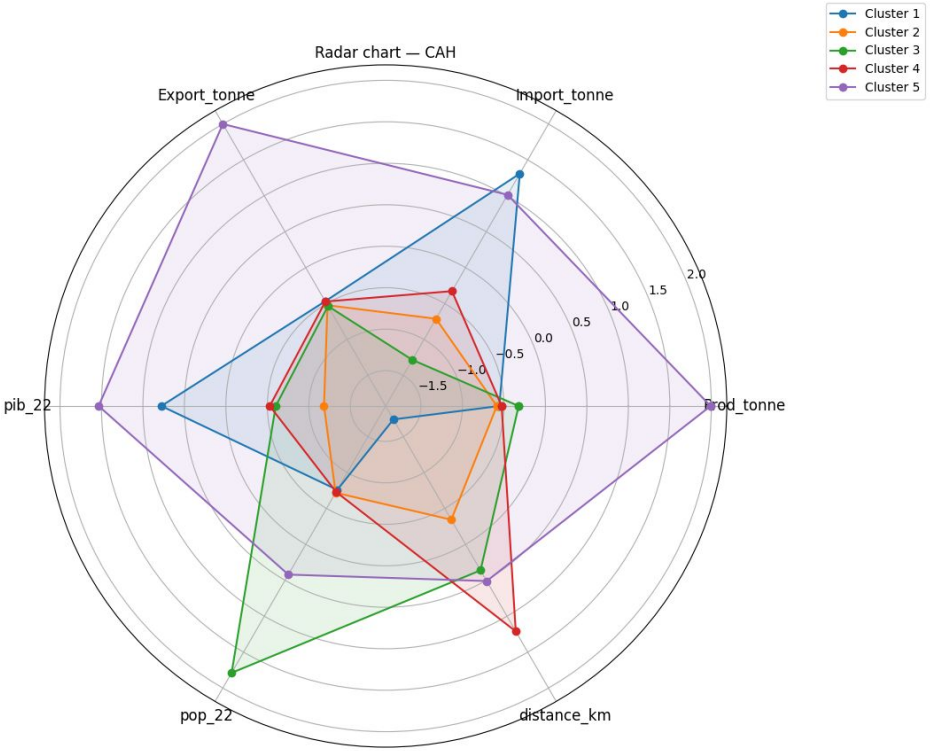
Cluster 1 — Importateurs solvables
Import élevé + PIB solide, production et population faibles. Pays dépendants des importations mais capables de payer.

Cluster 2 — Petits pays à faibles volumes
Production, export, PIB et population bas. Faible attractivité globale.

Cluster 3 — Grands pays peu productifs
Population très élevée, production/export faibles, distance élevée. Marchés immenses mais peu autonomes.

Cluster 4 — Pays moyens avec spécificités logistiques
Indicateurs globalement modérés, pas de dominance claire. Profils intermédiaires, hétérogènes.

Cluster 5 — Pays hors-norme (1 pays)
Production, export, PIB très élevés. Profil “superpuissance” type USA, incomparable aux autres groupes.



Algorithme K-MEANS

Définition

Méthode de segmentation qui regroupe automatiquement les pays en K groupes (clusters).

Chaque pays est affecté au cluster dont il est le plus proche, selon ses caractéristiques (production, PIB, population, etc.).

Le nombre de clusters K est choisi en amont, souvent après observation de la CAH ou de la méthode du coude.

Comment ça fonctionne ?

L'algorithme place K centres dans les données.

Chaque pays est attribué au centre le plus proche → formation d'un cluster.

Les centres sont réajustés jusqu'à ce que les regroupements deviennent cohérents et optimaux.

Résultat : une segmentation simple, lisible, et directement exploitable pour prioriser les pays cibles.

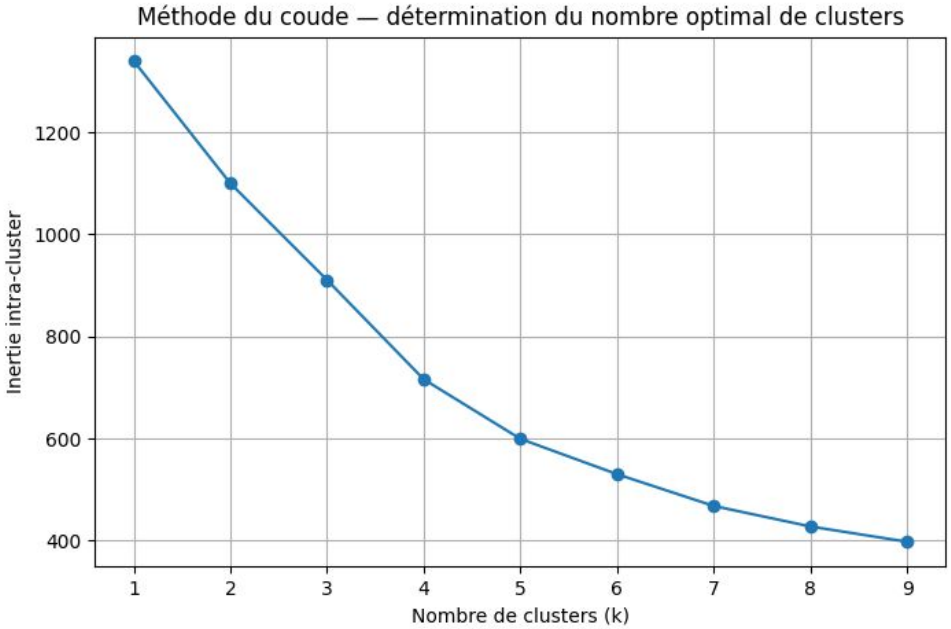
Pourquoi l'utiliser ?

Pour obtenir une segmentation claire, stable et facile à interpréter.

Le K-means affine les regroupements de la CAH en optimisant la cohésion interne de chaque cluster.

Idéal pour transformer l'analyse en groupes opérationnels utilisables en stratégie d'expansion.

Elbow method - K Means



L'inertie diminue fortement entre k = 1 et k = 4, puis la courbe commence à s'aplatir.

Après k = 4, le gain de qualité devient de plus en plus faible : chaque cluster supplémentaire n'apporte qu'une amélioration marginale.

Le point d'inflexion visible ("le coude") indique que k = 4 est le nombre optimal de clusters.

Cette valeur est cohérente avec la CAH, qui suggérait également 4 grands groupes naturels.

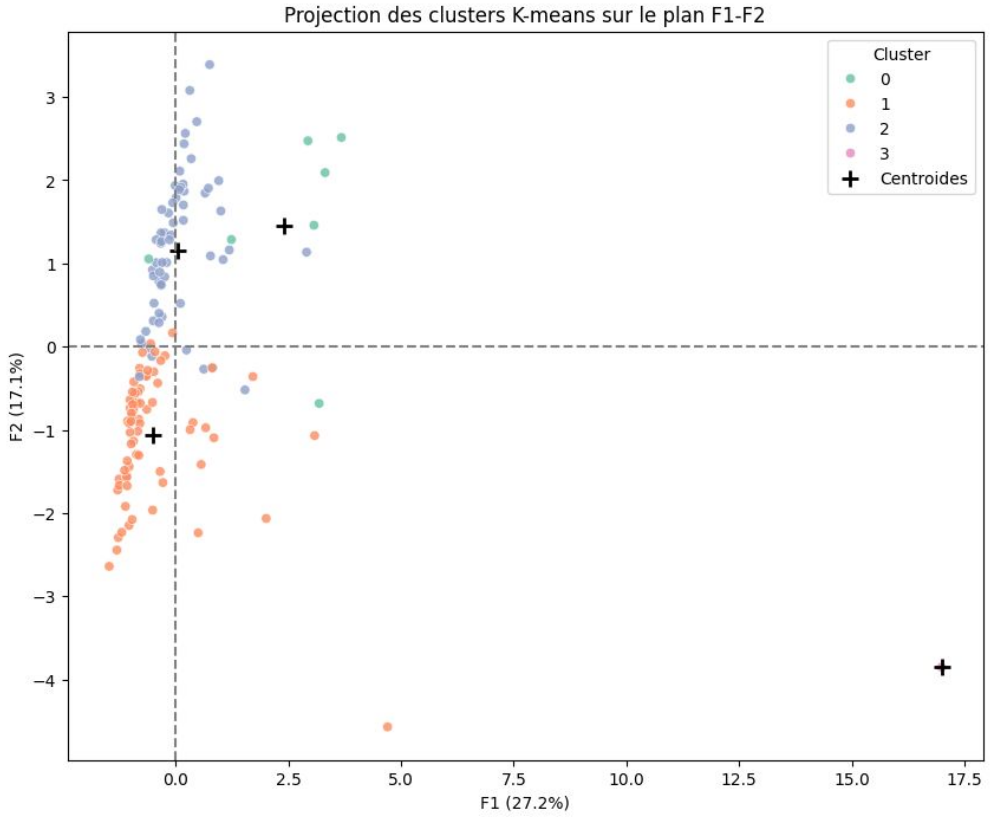
Projection des clusters sur F1/F2 - K MEANS

Cluster 0 — Pays producteurs solides
F1 élevé, F2 moyen ⟶ forts en production/export, économie correcte.

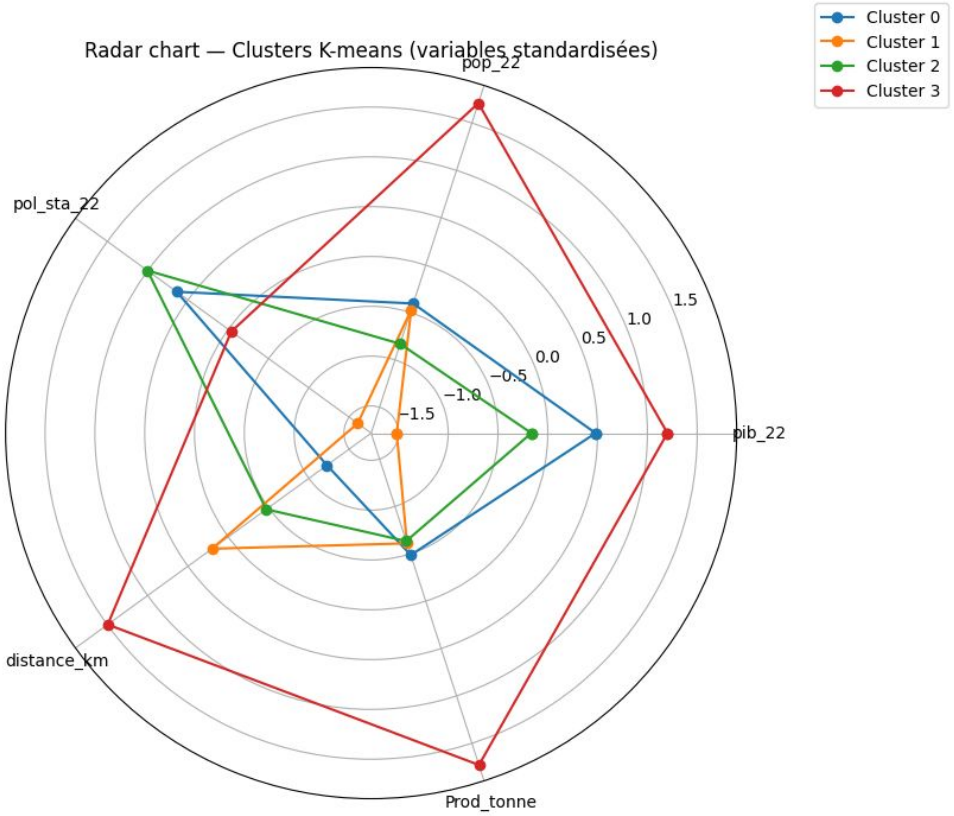
Cluster 1 — Petits pays fragiles
F1 bas, F2 bas ⟶ faibles volumes et faible stabilité/PIB.

Cluster 2 — Importateurs solvables
F1 bas, F2 élevé ⟶ dépendants des importations mais économiquement solides.

Cluster 3 — Pays hors-norme (1 pays)
F1 très élevé, F2 très bas ⟶ profil extrême type USA.



Radar des clusters - K Means



Cluster 0 — Pays stables intermédiaires
PIB / population moyens, stabilité élevée.
⟶ Intérêt : MODÉRÉ à FORT

Cluster 1 — Pays fragiles et peu solvables
PIB bas, faible production, instabilité.
⟶ Intérêt : FAIBLE

Cluster 2 — Pays stables à production moyenne
Stabilité forte, volumes intermédiaires.
⟶ Intérêt : MODÉRÉ

Cluster 3 — Très grands marchés producteurs
PIB, production, population très élevés.
⟶ Intérêt : FAIBLE (sauf niches premium)

Pays du Cluster 0

pays	Dispo_int_tonne	Export_tonne	Import_tonne	Prod_tonne	Var_stock_tonne	pib_22	%_aug_pib_17_22	pol_sta_22	delta_polsta_17_22	pop_22	%_aug_pop_17_22	distan
Japon	3510000.0	4000.0	1163000.0	2372000.0	20000.0	44972.3	1.05	1.03	-0.07	124997578	-1.63	9882.7
Mexique	4950000.0	8000.0	1158000.0	3800000.0	0.0	21392.1	-1.77	-0.69	0.12	128613117	4.22	9220.3
Royaume-Uni de Grande-Bretagne et d'Irlande du...	2507000.0	267000.0	947000.0	1952000.0	125000.0	52982.2	2.68	0.53	0.15	68179315	2.75	915.1
Pays-Bas (Royaume des)	408000.0	1365000.0	897000.0	876000.0	0.0	71324.0	8.62	0.73	-0.18	17904421	3.40	743.1
Allemagne	1462000.0	664000.0	789000.0	1507000.0	170000.0	62932.0	1.52	0.63	0.06	84086227	1.18	728.8
Émirats arabes unis	759000.0	133000.0	638000.0	53000.0	-201000.0	68867.8	-0.01	0.74	0.14	10242086	10.91	5209.3
Arabie saoudite	1659000.0	21000.0	551000.0	1130000.0	1000.0	67178.6	15.94	-0.36	0.29	32175352	4.53	4459.6

Tri des pays selon 6 critères clés :
import élevé, faible production, croissance population, PIB élevé, stabilité politique, proximité.

Résultats top 5 :
Japon, Mexique, Royaume-Uni, Pays-Bas, Allemagne.

Conclusion

ACP — Comprendre la structure mondiale

Nous avons d'abord réalisé une Analyse en Composantes Principales (ACP) pour :

réduire la complexité des données, identifier les axes qui différencient vraiment les pays,

visualiser les grandes tendances :
➜ pays producteurs/autosuffisants,
➜ pays importateurs/dépendants,
➜ pays stables vs fragiles,
➜ marchés en forte croissance.

Cette étape a posé les bases d'une segmentation cohérente.

CAH & K-means — Segmenter les pays

Nous avons ensuite appliqué deux méthodes de clustering : CAH pour révéler les groupes naturels issus des données, K-means pour consolider des clusters homogènes et stables.

Les deux méthodes convergent vers les mêmes familles de pays, confirmant 4 profils distincts : grands producteurs, importateurs solvables, pays intermédiaires, pays fragiles.

Le cluster le plus favorable réunit les pays : importateurs importants, à faible production locale, avec un PIB élevé, stables politiquement.

Pays recommandés pour l'ouverture internationale

Après un tri multicritères au sein du cluster le plus attractif, 7 pays émergent. Parmi eux, 3 pays européens se distinguent comme les meilleurs candidats pour une première expansion : **Royaume-Uni, Pays-Bas, Allemagne**

Forte dépendance aux importations
Marchés solvables
Forte stabilité
Compatible avec positionnement bio
Proximité logistique ⟶ démarrage plus simple et moins risqué