Projet 9

Produisez une étude de marché avec R ou Python

Effectuer un clustering simple

Explorer des données pour synthétiser des variables

Livrables

✅ Notebook préparation des données

✅ Notebook clustering & visualisation

✅ Présentation du projet

Développement à l'international

Contexte & objectif

L'entreprise la poule qui chante :

- Entreprise française
- Spécialisé dans la commercialisation de poulet
- Souhaite se développer à l'international
    - En exportant uniquement

Objectif :

- Proposer une première analyse des groupements de pays que l'on peut cibler pour exporter nos poulets.

La poule qui chante

Sommaire

Importation des données

5 datasets (sources FAO & banque mondiale) :

- Disponibilité alimentaire 2017
  - Variables : Disponibilité en protéine de volaille, exportation et importation en quantité, production
- Population
  - Variables : Population 2017 et calcul variable taux évolution 2015-2016
- Sécurité alimentaire
  - Variables : Disponibilité en protéine animale, PIB/Hab, Indice de stabilité politique
- Valeur des importations de poulet
  - Variable : Importations en valeur
- Frais de douane
  - Variables : Taux de douane et Country code

	Code Domaine	Domaine	Code zone	Zone	Code Élément	Élément	Code Produit	Produit	Code année	Année	Unité	Valeur	Symbole	Description du Symbole
0	FBS	Nouveaux Bilans Alimentaire	2	Afghanistan	5511	Production	2511	Blé et produits	2017	2017	Milliers de tonnes	4281.0	S	Données standardisées
1	FBS	Nouveaux Bilans Alimentaire	2	Afghanistan	5611	Importations - Quantité	2511	Blé et produits	2017	2017	Milliers de tonnes	2302.0	S	Données standardisées
2	FBS	Nouveaux Bilans Alimentaire	2	Afghanistan	5072	Variation de stock	2511	Blé et produits	2017	2017	Milliers de tonnes	-119.0	S	Données standardisées

	Code Domaine	Domaine	Code zone	Zone	Code Élément	Élément	Code Produit	Produit	Code année	Année	Unité	Valeur	Symbole	Description du Symbole	Note
0	OA	Séries temporelles annuelles	2	Afghanistan	511	Population totale	3010	Population-Estimations	2000	2000	1000 personnes	20779.953	X	Sources internationales sûres	NaN
1	OA	Séries temporelles annuelles	2	Afghanistan	511	Population totale	3010	Population-Estimations	2001	2001	1000 personnes	21606.988	X	Sources internationales sûres	NaN
2	OA	Séries temporelles annuelles	2	Afghanistan	511	Population totale	3010	Population-Estimations	2002	2002	1000 personnes	22600.770	X	Sources internationales sûres	NaN

	Code Domaine	Domaine	Code zone (M49)	Zone	Code Élément	Élément	Code Produit	Produit	Code année	Année	Unité	Valeur	Symbole	Description du Symbole	Note
0	FS	Données de la sécurité alimentaire	4	Afghanistan	6123	Valeur	21014	Disponibilités protéines moyennes d'origine an...	20162018	2016-2018	g/personne/jour	10.7	E	Valeur estimée	NaN
1	FS	Données de la sécurité alimentaire	4	Afghanistan	6126	Valeur	22013	PIB par habitant, ($ PPA internationaux consta...	2017	2017	I$	2058.4	X	Ciffre de sources internationales	NaN
2	FS	Données de la sécurité alimentaire	4	Afghanistan	6125	Valeur	21032	Stabilité politique et absence de violence/ter...	2017	2017	indice	-2.8	X	Ciffre de sources internationales	NaN

	Code Domaine	Domaine	Code pays déclarant (M49)	Pays déclarants	Code pays partenaire (M49)	Pays partenaires	Code Élément	Élément	Code Produit (CPC)	Produit	Code année	Année	Unité	Valeur	Symbole	Descript Symb
0	TM	Matrices du commerce détaillées	4	Afghanistan	840	États-Unis d'Amérique	5622	Importations - Valeur	2151	Poulets	2018	2018	1000 US$	6	T	Chiffre r offi
1	TM	Matrices du commerce détaillées	4	Afghanistan	364	Iran (République islamique d')	5622	Importations - Valeur	2151	Poulets	2014	2014	1000 US$	3101	T	Chiffre r offi
2	TM	Matrices du commerce détaillées	4	Afghanistan	364	Iran (République islamique d')	5622	Importations - Valeur	2151	Poulets	2016	2016	1000 US$	3069	T	Chiffre r offi

	Pays	Country Code	Taux 2017
0	Djibouti	DJI	47.410000
1	Népal	NPL	35.677368
2	Palaos	PLW	29.880000
3	Libye	LBY	22.120000
4	Bangladesh	BGD	21.721500

Doublons, pivot table & merge

- J’ai fait des .pivot table sur mes données pour pouvoir les utiliser.
- J’ai travaillé les tables une par une, en traitant les doublons et valeur manquante à chaque importation de nouvelles données.
- J’ai utilisé comme clé primaire Pays pour chaque merge.

- Les données importantes à notre analyse étaient dans la fichier dispo alimentaire, je me suis basée sur cette table pour mes merges
- J’ai ajouté des variables en calculant les variables : valeurs import, et ratio dispo animale sur dispo volaille.

Pays	Disponibilité en protéines de volailles (g/pers/jour)	Exportation (Milliers de tonnes)	Importations (Milliers de tonnes)	Production (Milliers de tonnes)	Pop 2017	%Evol Pop 2016/2017	Disponibilité en protéines animales (g/pers/jour)	PIB/hab ($)	Indice de stabilité politique	Country Code	Taux douane	Valeur import ($/t)	%Protéines volailles/protéines animales

Variables calculés

Valeurs manquantes & doublons

- Pas de doublons dans les datasets de base.
- Lors des différents merge, des valeurs manquantes sont apparus.
  - Recherche de certaines valeurs sur internet (Banque mondiale)
  - Suppression de certains individus (selon importance pays)
  - Remplacement de certaines valeurs nan par 0.

160 pays retenus pour notre analyse (95,5% de la population mondiale 2017)

	Pays	Disponibilité en protéines de volailles (g/pers/jour)	Exportation (Milliers de tonnes)	Importations (Milliers de tonnes)	Production (Milliers de tonnes)	Pop 2017	%Evol Pop 2016/2017	Disponibilité en protéines animales (g/pers/jour)	PIB/hab ($)	Indice de stabilité politique
34	Chine, Taiwan Province de	11.01	8.0	161.0	652.0	23674.546	0.00	42.3	NaN	0.86
35	Chine, continentale	3.96	577.0	452.0	18236.0	1421021.791	0.00	40.0	NaN	NaN
41	Cuba	7.12	0.0	312.0	29.0	11339.254	0.00	33.0	NaN	0.65
107	Nouvelle-Calédonie	14.05	0.0	9.0	1.0	277.150	0.01	50.3	NaN	NaN
118	Polynésie française	16.20	1.0	15.0	1.0	276.102	0.01	62.3	NaN	NaN
127	République populaire démocratique de Corée	0.49	0.0	0.0	43.0	25429.825	0.00	10.0	NaN	-0.50
156	Venezuela (République bolivarienne du)	7.23	0.0	25.0	600.0	29402.484	-0.02	24.0	NaN	-1.27
158	Yémen	3.04	0.0	78.0	168.0	27834.819	0.02	9.7	NaN	-2.94

```
# modification de la valeur du PIB/habitant
data_pib_null.loc[data_pib_null.index[0], 'PIB/hab ($)'] = 25534
data_pib_null.loc[data_pib_null.index[1], 'PIB/hab ($)'] = 8817
#Moyenne continent pour la stabilité politique
data_pib_null.loc[data_pib_null.index[1], 'Indice de stabilité politique'] = -0.46
data_pib_null.loc[data_pib_null.index[2], 'PIB/hab ($)'] = 8543.4
data_pib_null.loc[data_pib_null.index[3], 'Indice de stabilité politique'] = 0.86
data_pib_null.loc[data_pib_null.index[3], 'PIB/hab ($)'] = 33875
data_pib_null.loc[data_pib_null.index[4], 'PIB/hab ($)'] = 19744
data_pib_null.loc[data_pib_null.index[4], 'Indice de stabilité politique'] = 0.86
data_pib_null.loc[data_pib_null.index[5], 'PIB/hab ($)'] = 31616
data_pib_null.loc[data_pib_null.index[6], 'PIB/hab ($)'] = 10742.7
data_pib_null.loc[data_pib_null.index[7], 'PIB/hab ($)'] = 893.7
```

	Pays	Disponibilité en protéines de volailles (g/pers/jour)	Exportation (Milliers de tonnes)	Importations (Milliers de tonnes)	Production (Milliers de tonnes)	Pop 2017	%Evol Pop 2016/2017	Disponibilité en protéines animales (g/pers/jour)	PIB/hab ($)	Indice de stabilité politique	Importations (1000$)
11	Australie	16.60	43.0	16.0	1269.0	24584.620	0.01	72.3	48398.5	0.90	NaN
60	Haïti	2.75	0.0	89.0	9.0	10982.366	0.01	10.3	3153.3	-0.67	NaN
66	Iraq	5.37	0.0	470.0	96.0	37552.781	0.03	14.3	10526.4	-2.31	NaN
82	Libéria	3.74	1.0	48.0	15.0	4702.226	0.03	10.7	1564.2	-0.33	NaN
112	Philippines	4.65	3.0	249.0	1272.0	105172.925	0.01	26.0	8120.9	-1.19	NaN
141	Timor-Leste	1.46	0.0	11.0	1.0	1243.258	0.02	14.6	3177.5	0.07	NaN
145	Turkménistan	1.53	0.0	9.0	20.0	5757.667	0.02	36.7	14205.0	-0.13	NaN
158	Îles Salomon	1.51	0.0	6.0	0.0	636.039	0.03	15.3	2663.5	0.20	NaN
162	Cuba	7.12	0.0	312.0	29.0	11339.254	0.00	33.0	8543.4	0.65	NaN

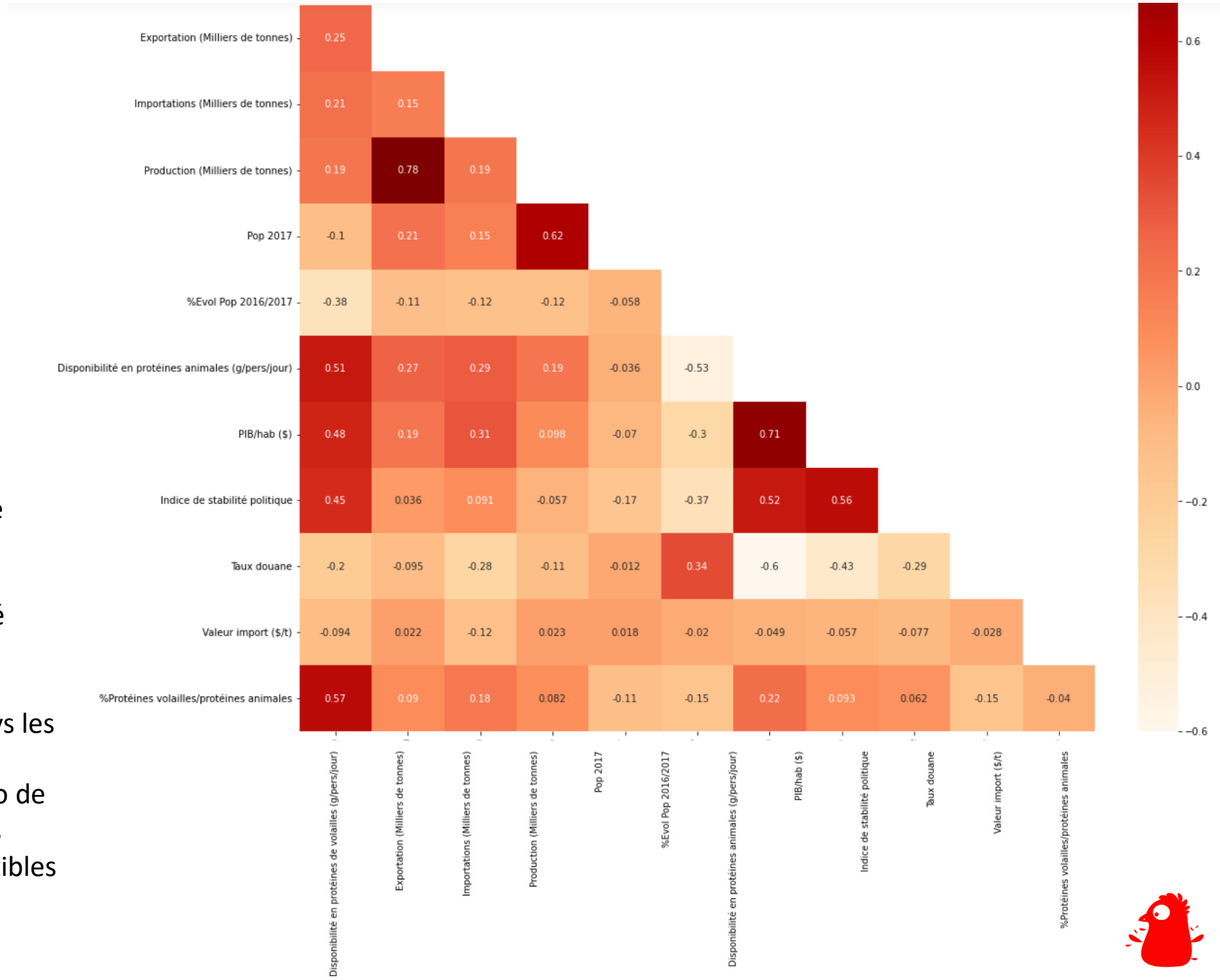
Élément	Pays	Disponibilité en protéines de volailles (g/pers/jour)	Exportation (Milliers de tonnes)	Importations (Milliers de tonnes)	Production (Milliers de tonnes)
43	Djibouti	0.92	NaN	3.0	NaN
92	Maldives	4.70	NaN	12.0	NaN
113	Ouzbékistan	0.63	NaN	NaN	NaN
122	Pérou	6.71	NaN	60.0	1465.0
130	République démocratique populaire lao	3.59	NaN	NaN	NaN

Matrice de corrélation

Fortes corrélations entre :

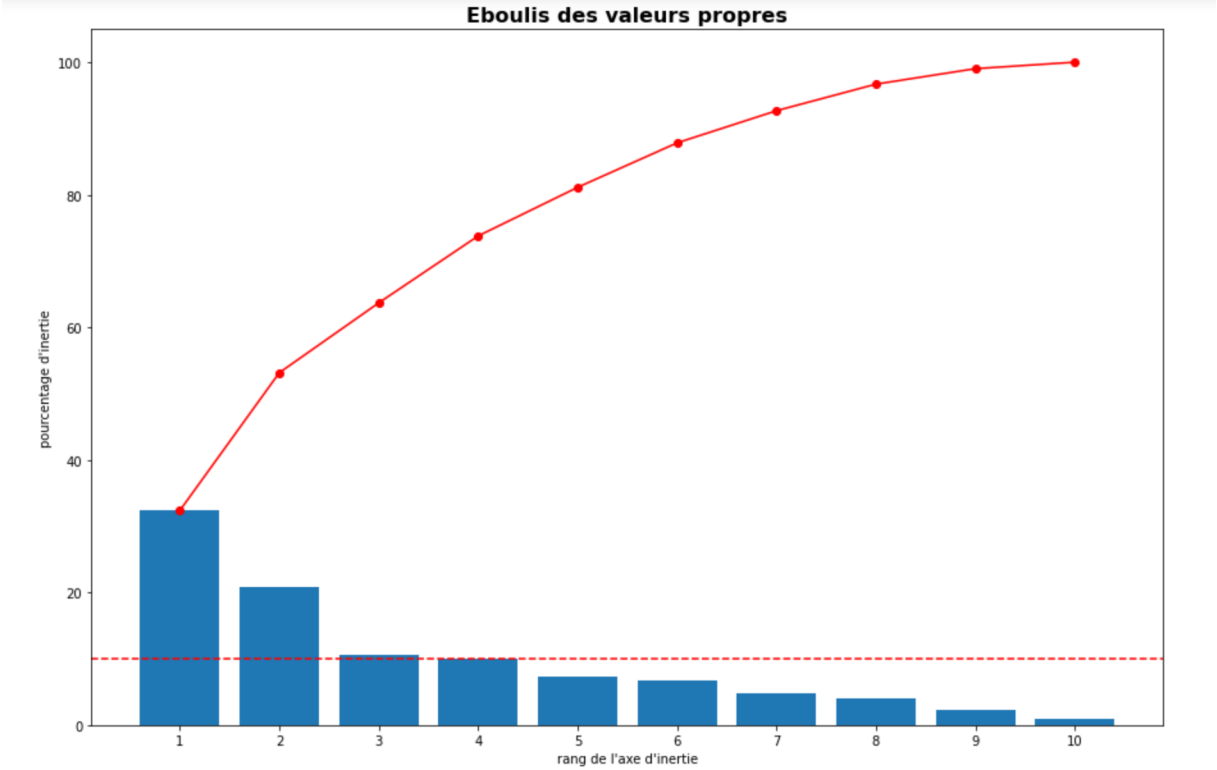
- la production et l'exportation
- la population et la production
- la dispo. en protéines et le ratio de protéines animales
- la dispo. en protéines et le PIB
- la dispo. en protéines et la stabilité politique

Nous sommes intéressés par les pays les plus susceptibles de consommer du poulet, donc ceux ayant un fort ratio de protéines animales. Les corrélations montrent que ces pays sont susceptibles d'avoir un fort PIB.



Classification ascendante hiérarchique

Analyse des composantes principales



	1	2	3	4
Cumul de variance expliquée	32.37449	53.153281	63.691483	73.788135

	5	6	7	8
Cumul de variance expliquée	81.118105	87.826391	92.679413	96.687259

	9	10
Cumul de variance expliquée	99.024338	100.0

ACP

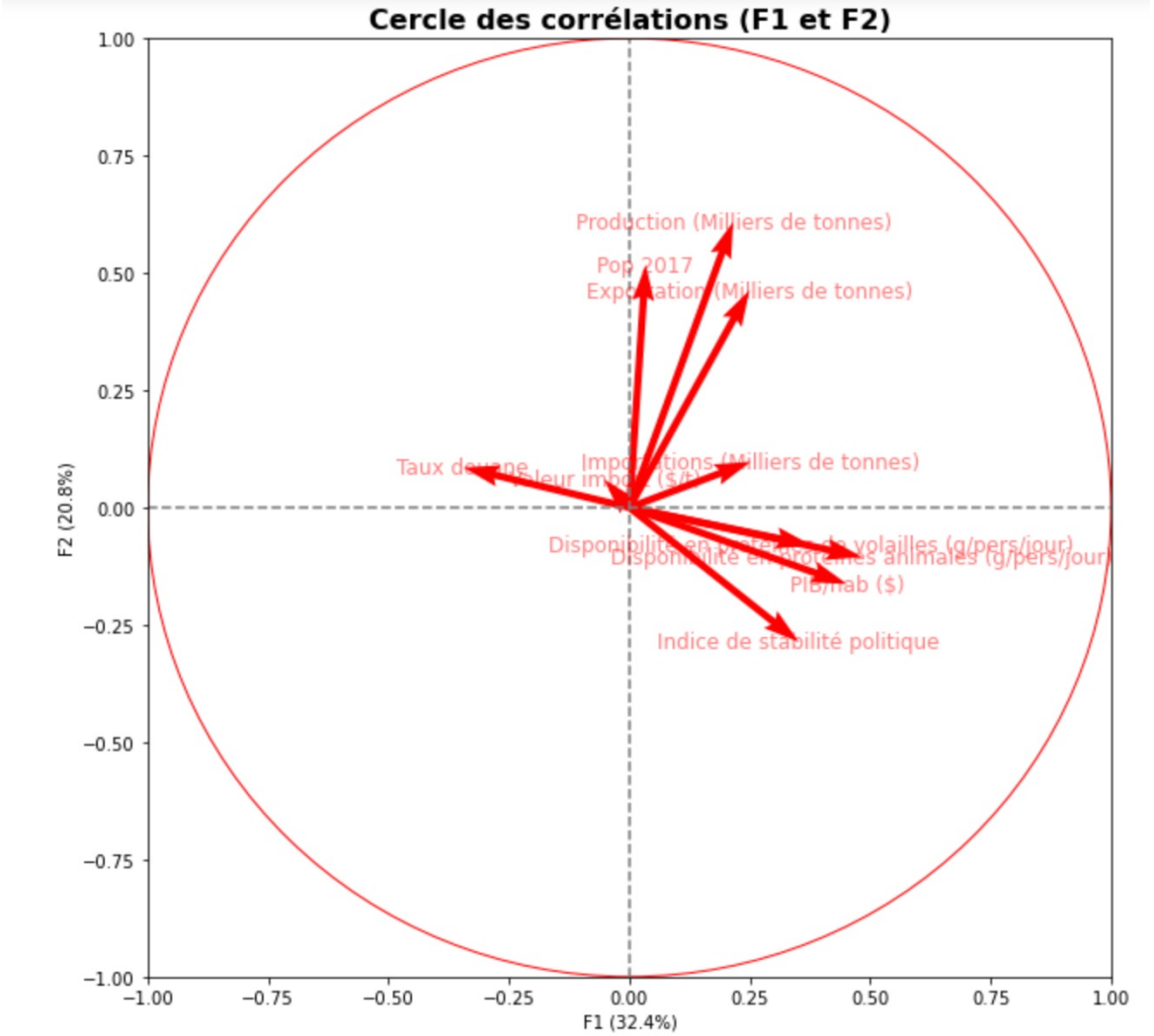
- Permet la réduction des variables
- Elle est utilisé pour visualiser et modéliser
- Sur des variables normalisés (et centré ici)

Variance expliquée

- En bleu la variance de chaque nouvelle composante, et en rouge la variance cumulée.
- 87% de la variance est comprise dans les 6 premières composantes.

Nous choisissons 6 composantes pour l'ACP.

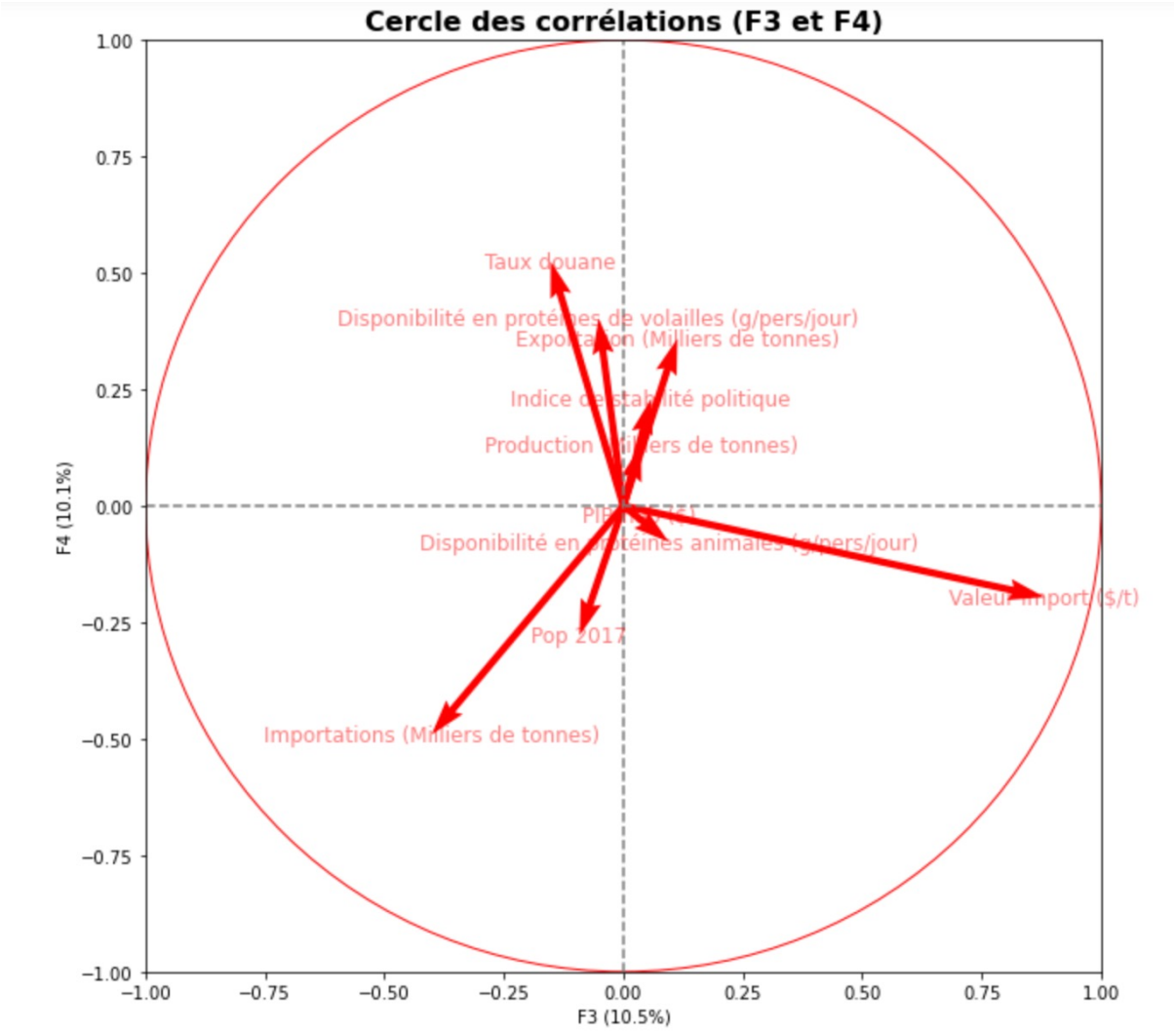
Représentation des cercles de corrélations F1 & F2



- La variable avec une contribution négative à la 1e CP correspondent à celles dont le taux doit être le plus faible possible : les frais de douanes. Nous cherchons effectivement à nous implanter dans des pays aux frais de douane faibles pour minimiser les couts. Les autres variables ont un effet positif, 4 avec un effet plus marqué : le PIB/habitant, l'indice de stabilité et les apports en protéine.
- La 2e CP est impacté fortement par la production, la population et les exportations elle caractérise donc l'auto-suffisance, minimisée par une stabilité politique basse et un PIB bas.

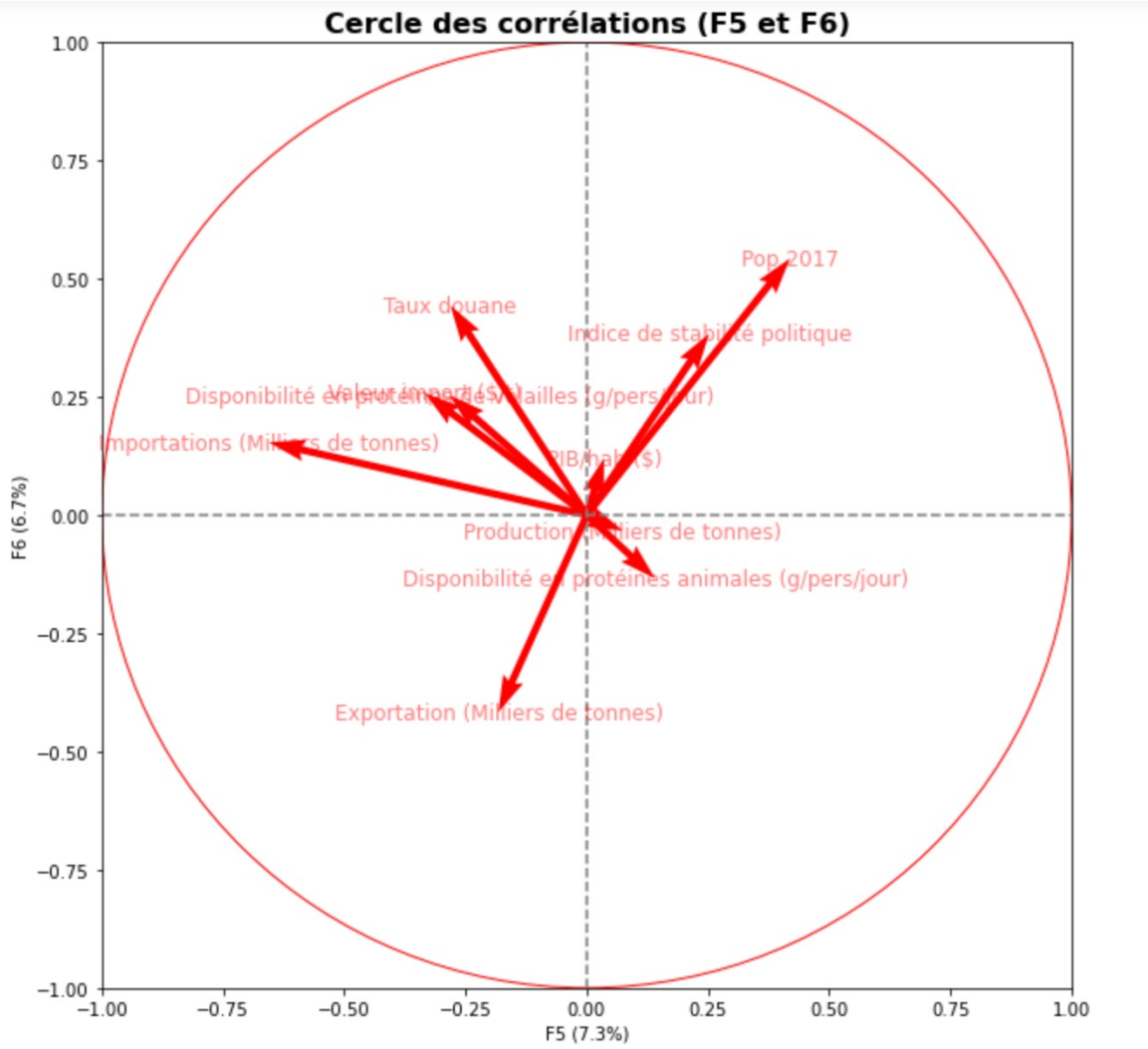
Comme vu sur la matrice de corrélation, il y a une corrélation entre PIB et consommation de viande. Les fortes valeurs de la 1e CP caractérisent les pays stables et riches où la consommation de viande est importante.

Représentation des cercles de corrélations F3 & F4



- La 3e CP est influencée positivement par la valeur de l'import et négativement par les importations. (Cette CP caractérise les aspects financiers).
- La 4e CP est influencée positivement par les taux de douane (variables que nous voulons les plus faibles possibles) , la disponibilité en protéines de volailles, les exportations et négativement par les importations (que nous voulons plus fortes) et la population. Cette composante devra donc être la plus petite possible.

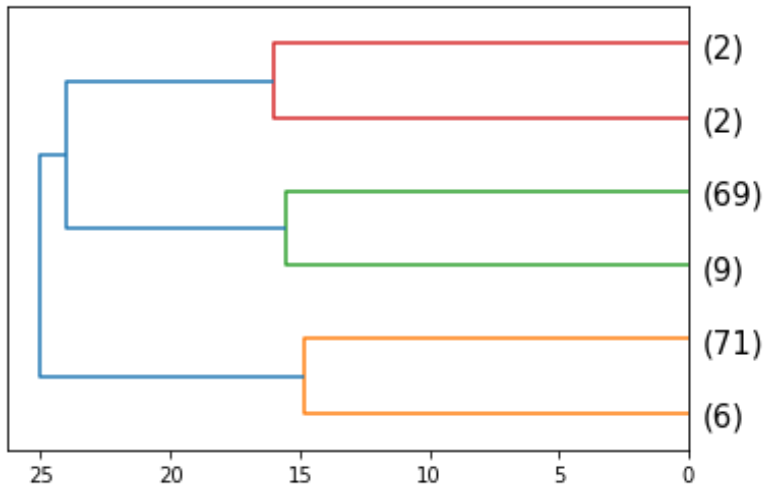
Représentation des cercles de corrélations F5 & F6

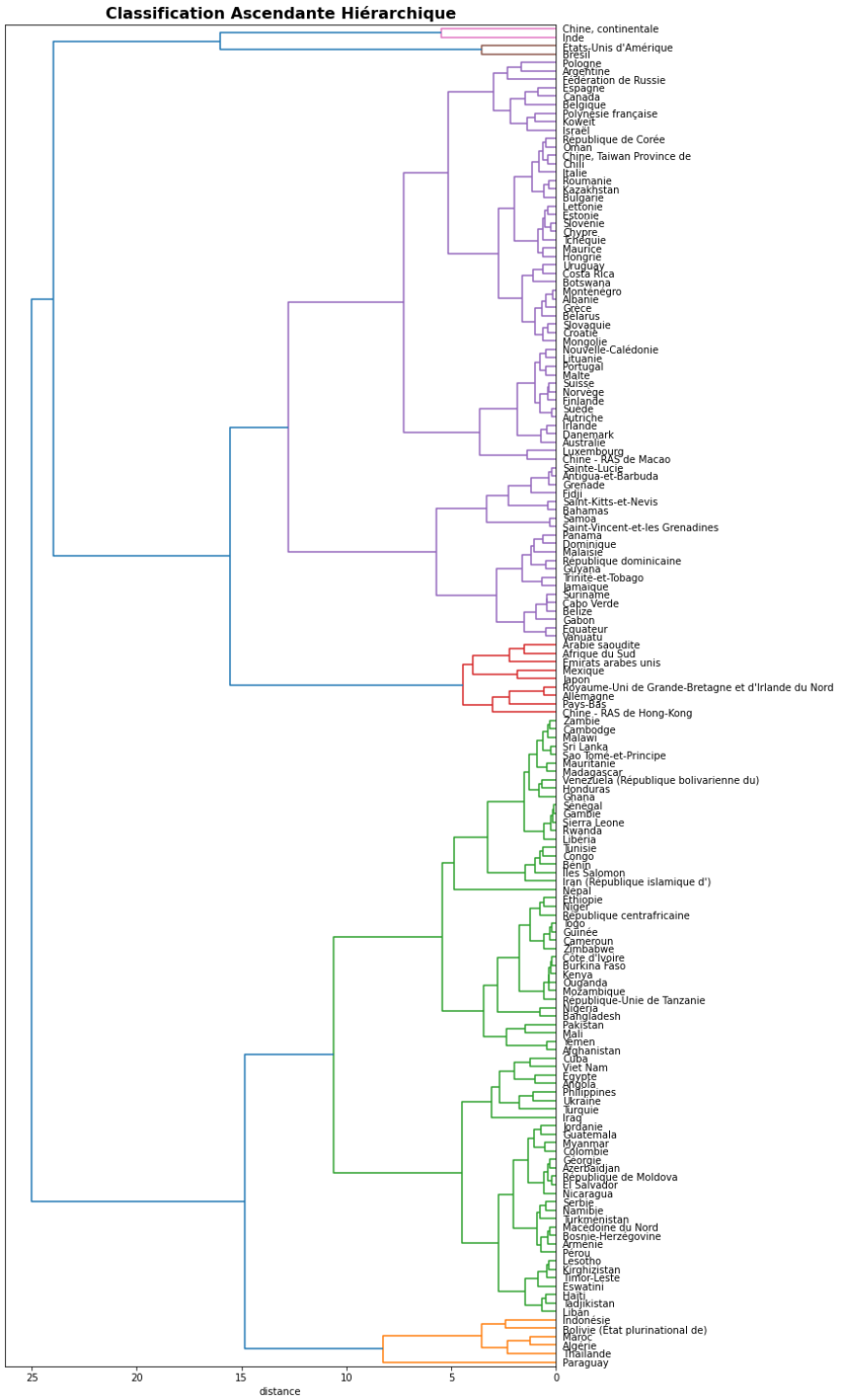


- La 5e CP est influencée positivement par la population et la stabilité politique et négativement par les importations, et la valeur de l'import.
- La 6e CP est influencée positivement par la population, les taux de douane, et la stabilité politique et négativement par les exportations.

Dendogramme

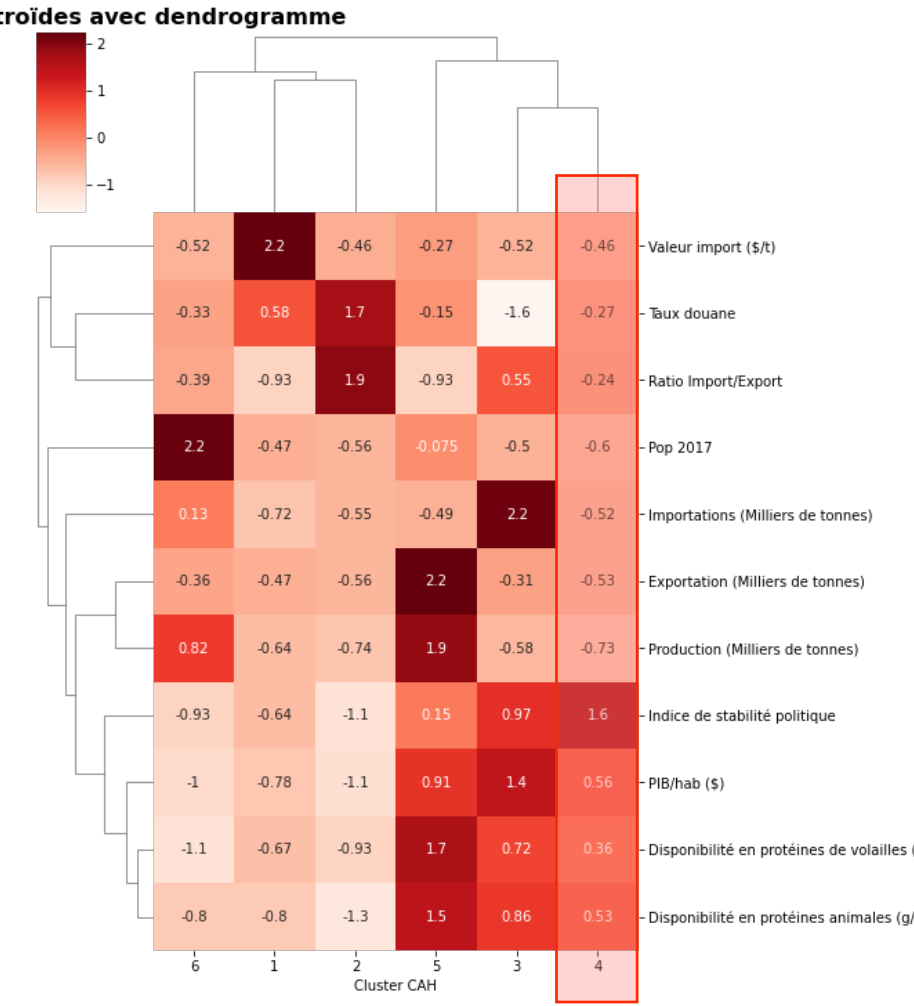
Distance entre les groupes





Heatmaps et clusters CAH

Cluster CAH	Disponibilité en protéines de volailles (g/pers/jour)	Exportation (Milliers de tonnes)	Importations (Milliers de tonnes)	Production (Milliers de tonnes)	Pop 2017	Disponibilité en protéines animales (g/pers/jour)	PIB/hab ($)	Indice de stabilité politique	Taux douane	Valeur import ($/t)	Ratio Import/Export
1	4.876667	134.166667	1.666667	932.000000	7.148186e+04	27.283333	11403.716667	-0.470000	6.385965	2294.691667	0.012422
2	3.439296	11.746479	47.521127	272.000000	3.032531e+04	20.185915	6753.536620	-0.705915	8.396797	60.084507	4.045564
3	12.605556	363.666667	760.666667	1360.777778	5.839908e+04	53.566667	44876.700000	0.304444	2.652222	10.430000	2.091659
4	10.588406	56.985507	56.202899	357.492754	1.122991e+04	48.421739	31850.100000	0.602174	4.914991	58.093188	0.986267
5	17.805000	3956.500000	63.000000	18057.500000	2.664593e+05	63.150000	37219.700000	-0.090000	5.125000	219.980000	0.015923
6	2.355000	290.000000	226.000000	10890.500000	1.379849e+06	27.350000	7499.950000	-0.610000	4.805000	9.400000	0.779310



Nombre de pays par clusters :

- Cluster CAH 1 : 6 pays
- Cluster CAH 2 : 71 pays
- Cluster CAH 3 : 12 pays
- Cluster CAH 4 : 67 pays
- Cluster CAH 5 : 2 pays
- Cluster CAH 6 : 2 pays

Test statistique CAH

Test de d'homoscédasticité – Test de Levene > H1 : au moins une des variances diffère des autres

Analyse de la variance – Test Anova > H0 : la distribution suit une loi normale

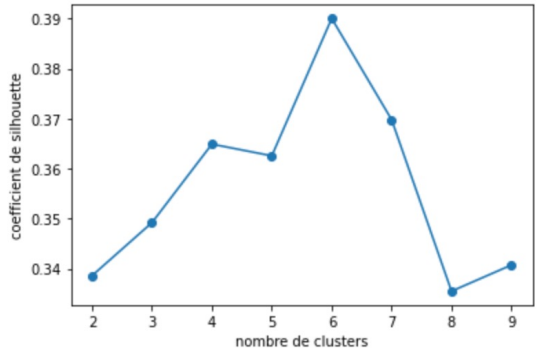
Test de normalité – Test Shapiro-Wilk > H1 : au moins une des moyennes diffère

Classification ascendante hiérarchique

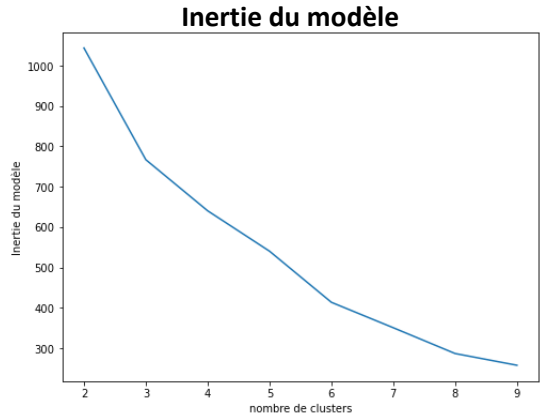
Methode K-means : choix du nombre de clusters et projection

	2	3	4	5	6	7	8	9
coefficient de silhouette	0.338622	0.349221	0.364898	0.362562	0.390011	0.369784	0.335513	0.340745

Evolution du coefficient de silhouette en fonction du nombre de clusters

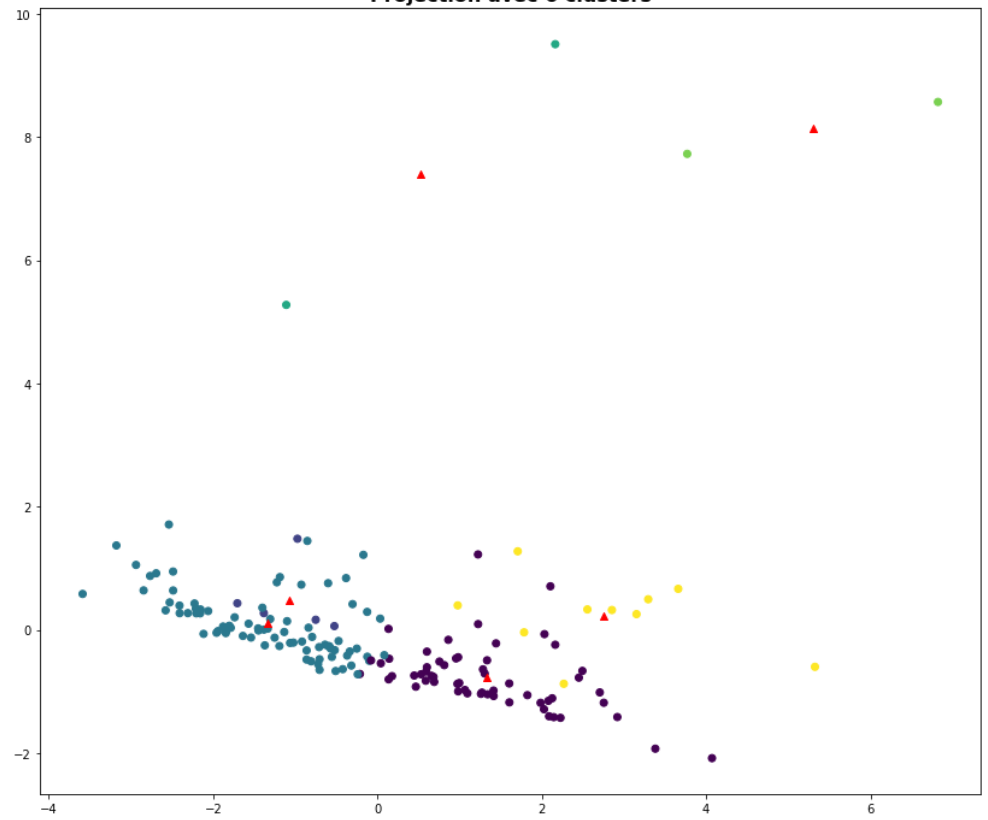


Inertie du modèle



Avec la méthode du coefficient de silhouette > recommande une analyse de 6 clusters

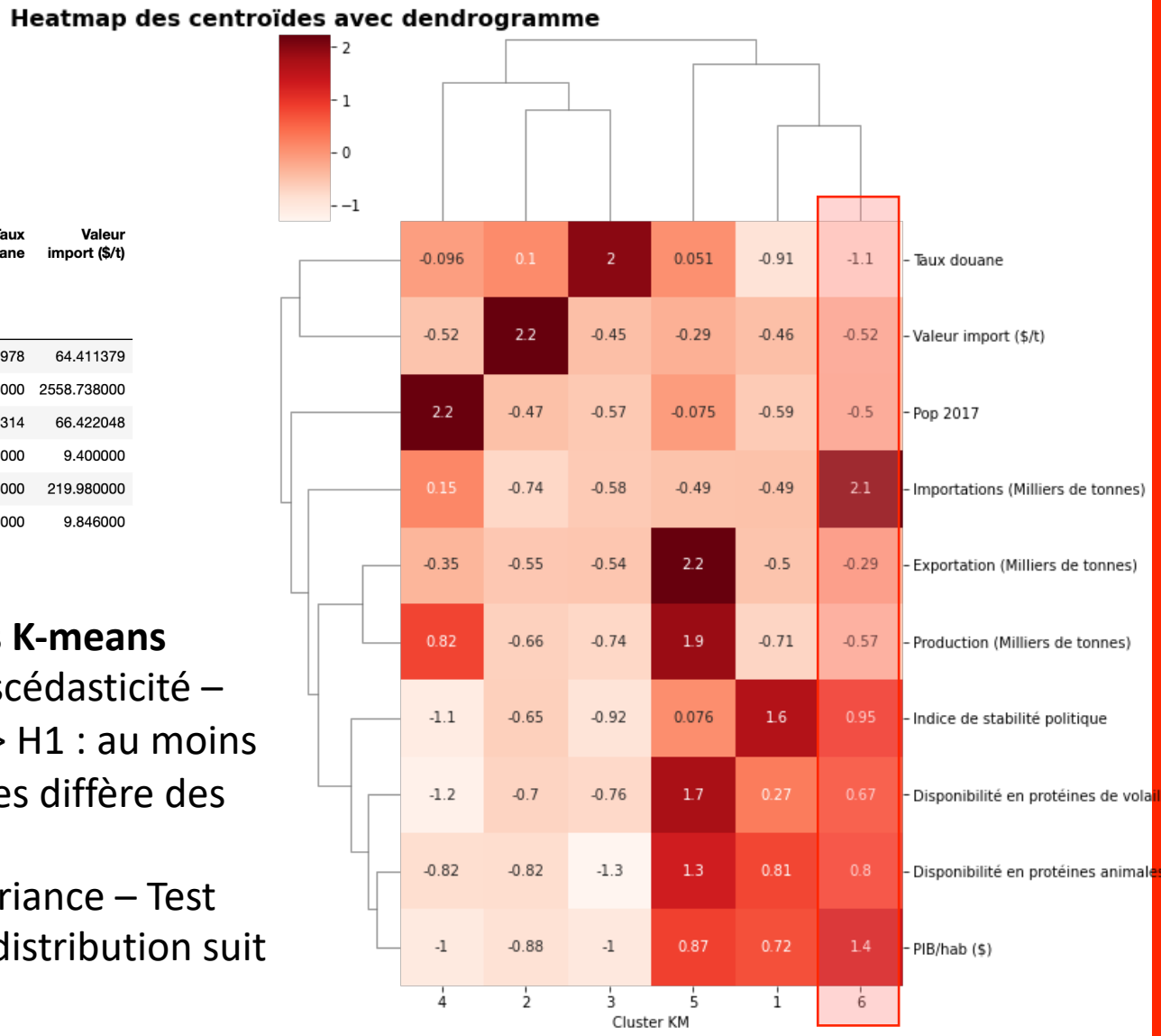
Projection avec 6 clusters



Les clusters sont biens distincts.

Heatmap & clusters KM

Cluster KM	Disponibilité en protéines de volailles (g/pers/jour)	Exportation (Milliers de tonnes)	Importations (Milliers de tonnes)	Production (Milliers de tonnes)	Pop 2017	Disponibilité en protéines animales (g/pers/jour)	PIB/hab ($)	Indice de stabilité politique	Taux douane	Valeur import ($/t)
1	10.139655	67.810345	65.258621	444.465517	1.365594e+04	54.837931	34925.646552	0.602414	3.042978	64.411379
2	4.982000	1.800000	1.600000	783.200000	7.193626e+04	27.400000	10199.860000	-0.414000	5.240000	2558.738000
3	4.711325	19.626506	41.795181	239.469880	2.656778e+04	19.522892	8058.957831	-0.533253	9.288314	66.422048
4	2.355000	290.000000	226.000000	10890.500000	1.379849e+06	27.350000	7499.950000	-0.610000	4.805000	9.400000
5	17.805000	3956.500000	63.000000	18057.500000	2.664593e+05	63.150000	37219.700000	-0.090000	5.125000	219.980000
6	12.240000	377.300000	735.200000	1399.700000	5.904342e+04	54.580000	44846.740000	0.302000	2.583000	9.846000



Nombre de pays par clusters :

- Cluster KM 1 : 58 Pays
- Cluster KM 2 : 5 Pays
- Cluster KM 3 : 83 Pays
- Cluster KM 4 : 2 Pays
- Cluster KM 5 : 2 Pays
- Cluster KM 6 : 10 Pays

Test statistiques K-means

Test de d'homoscédasticité – Test de Levene > H1 : au moins une des variances diffère des autres (10/12)

Analyse de la variance – Test Anova > H0 : la distribution suit une loi normale

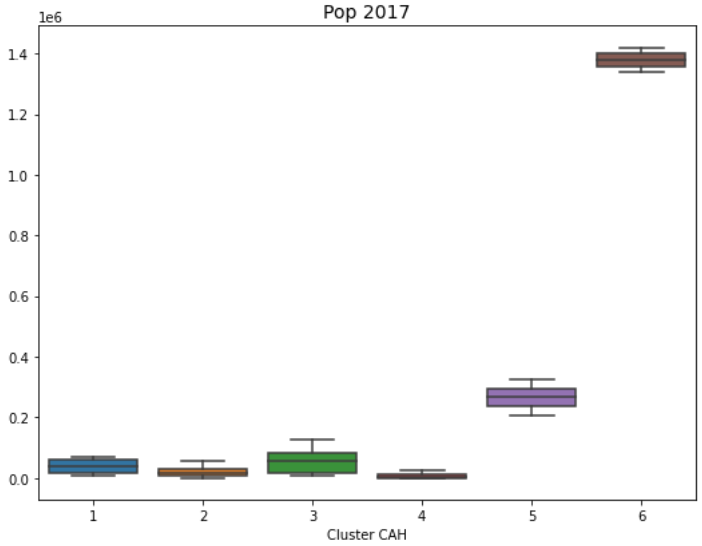
Test de normalité – Test Shapiro-Wilk > H1 : au moins une des moyennes diffère

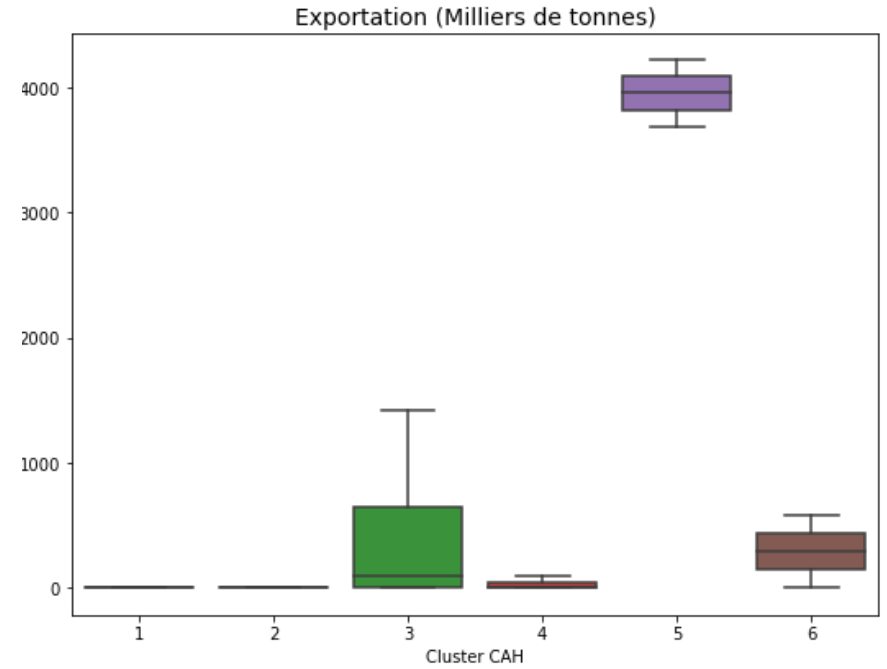
Classification ascendante hiérarchique

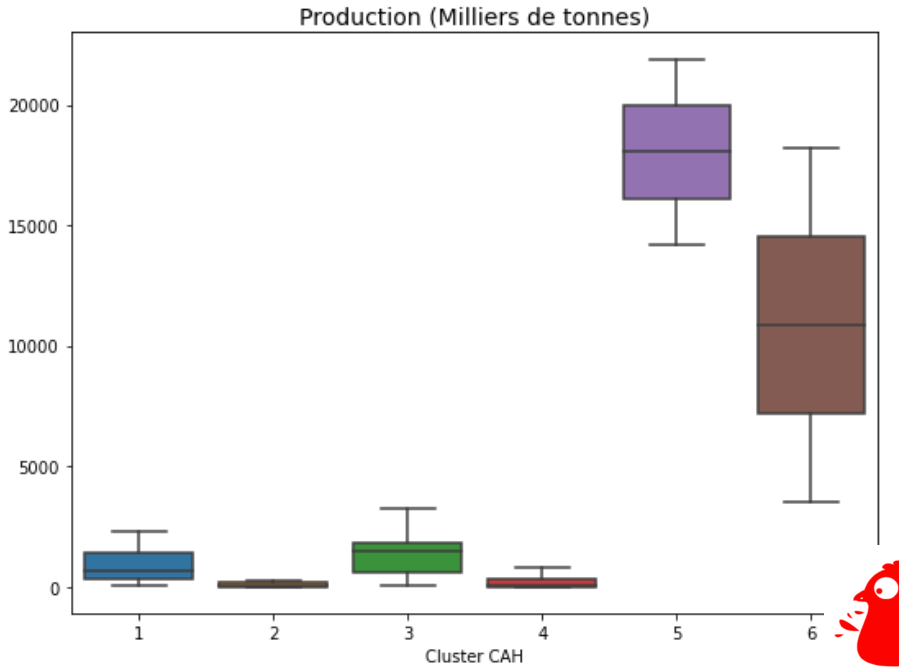
Analyse des clusters

Cluster CAH 5,6 & Cluster KM 4,5

- **CAH 5, KM4 :** Les EUA et le Brésil, pays producteurs de volaille (dépendance aux importations très faible) avec une forte exportation.
- **CAH6, KM5 :** L'Inde et la Chine (population supérieure à 1 Milliard). La consommation de volailles y est faible



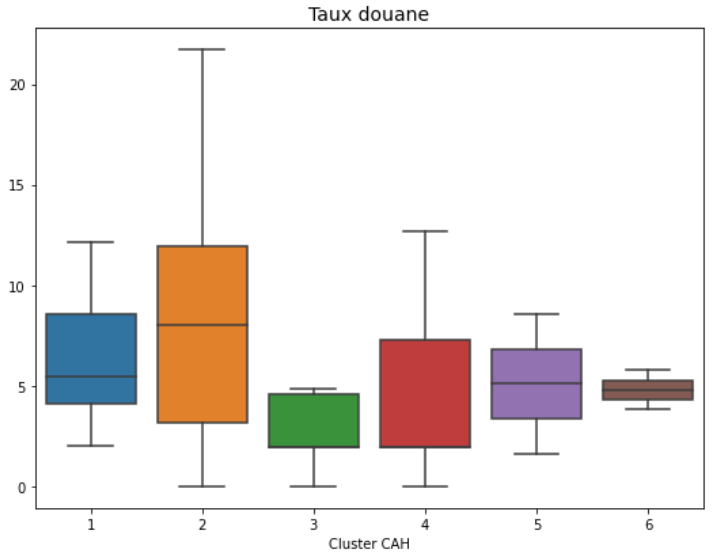


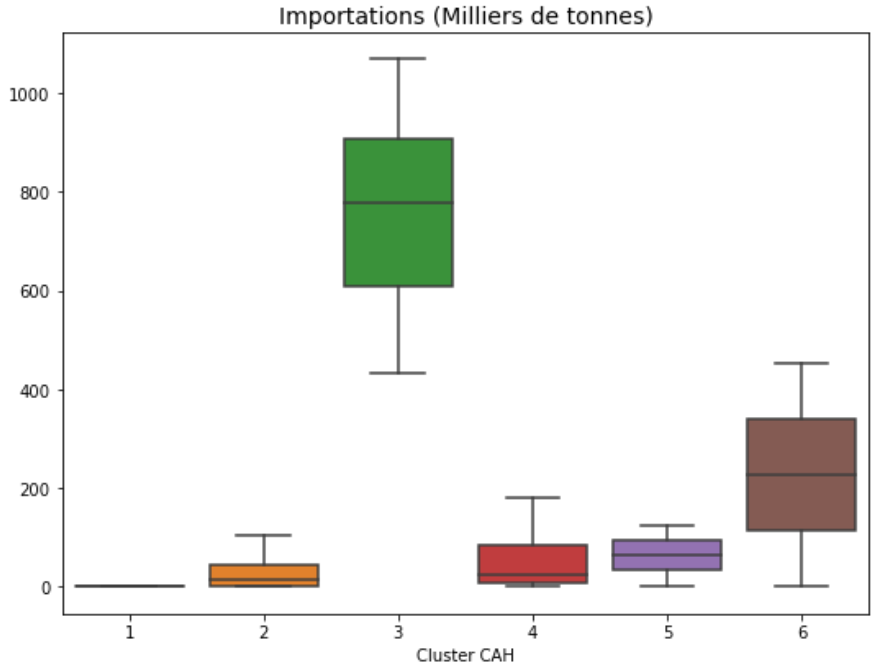


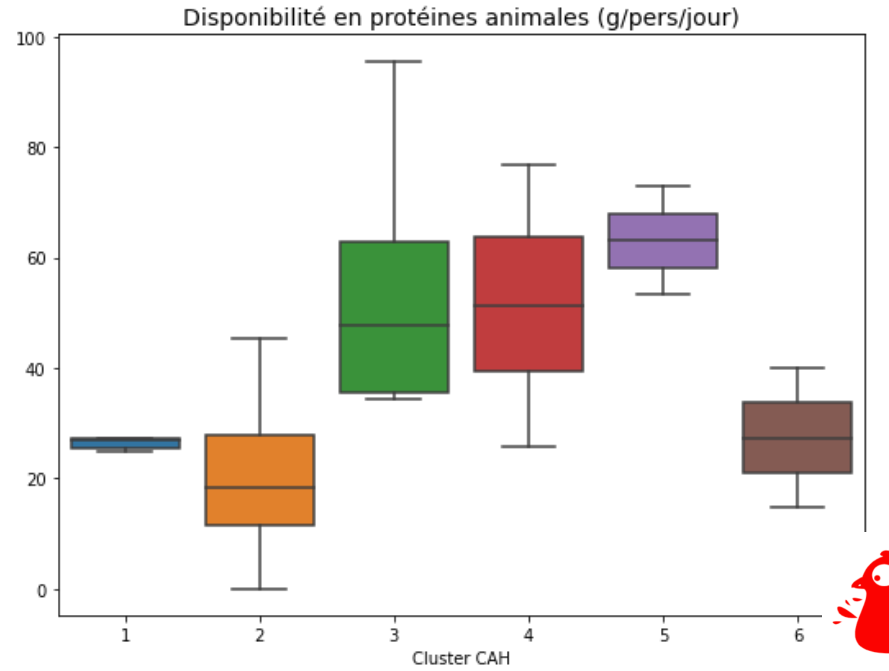
Analyse des clusters

Cluster CAH 3,4 & Cluster KM 6,1

- **CAH3, KM6** : Pays riches et stables politiquement avec une forte dépendance aux importations et à un prix élevé. Les taux de douanes sont faibles, ce sont les pays de ce cluster que nous devons cibler.
- **CAH4,KM1 :** Petits pays stables politiquement avec un PBI élevé. Ils importent autant qu'ils exportent et on une dépendance moyenne à l'importation. Valeur d'import supérieur au cluster 3 (cible).



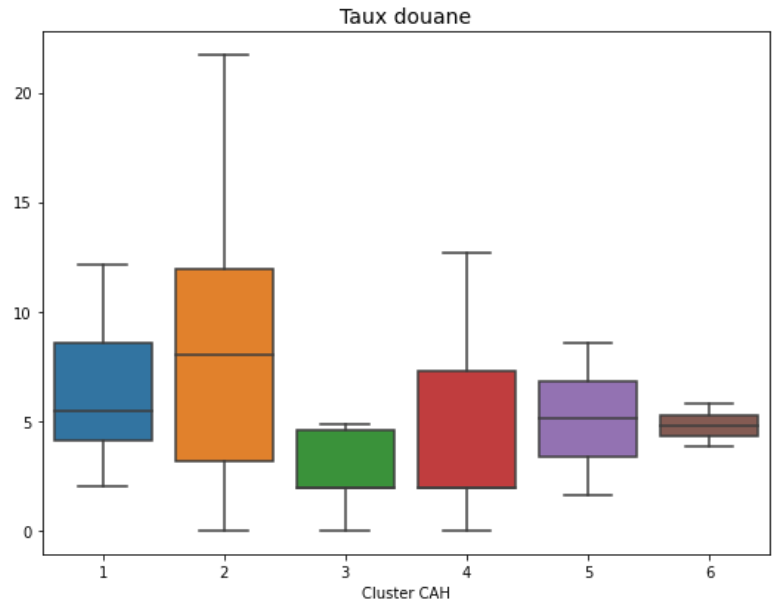


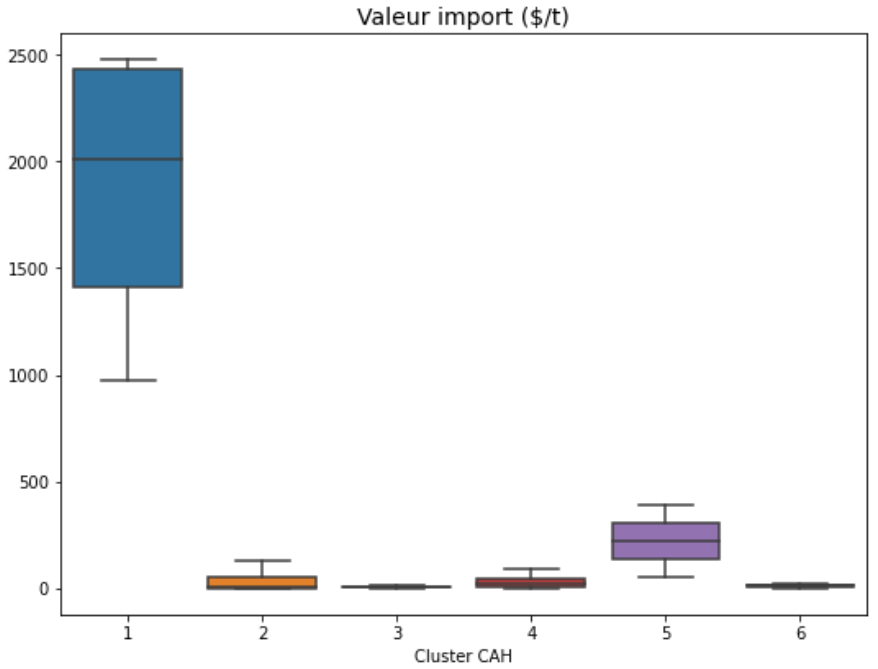


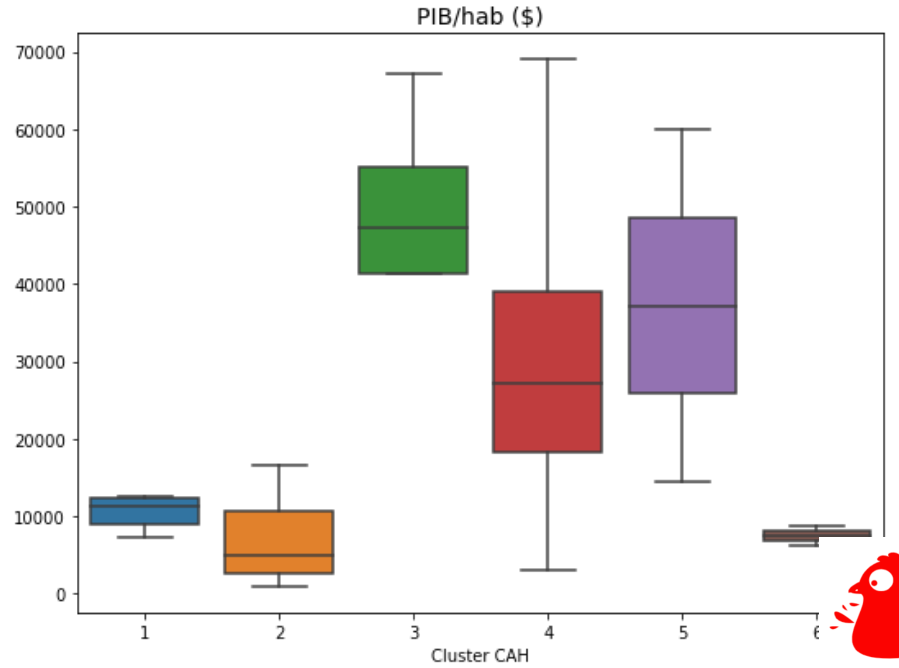
Analyse des clusters

Cluster CAH 2 & Cluster KM 3

Pays en voie de développement à la stabilité politique basse. Les frais de douane y sont élevés et le prix à l'import est le plus faible. La demande en viande de volailles pourrait y être forte (production faible, forte dépendance aux importations) mais les difficultés économiques écartent ces pays pour nos exportations.



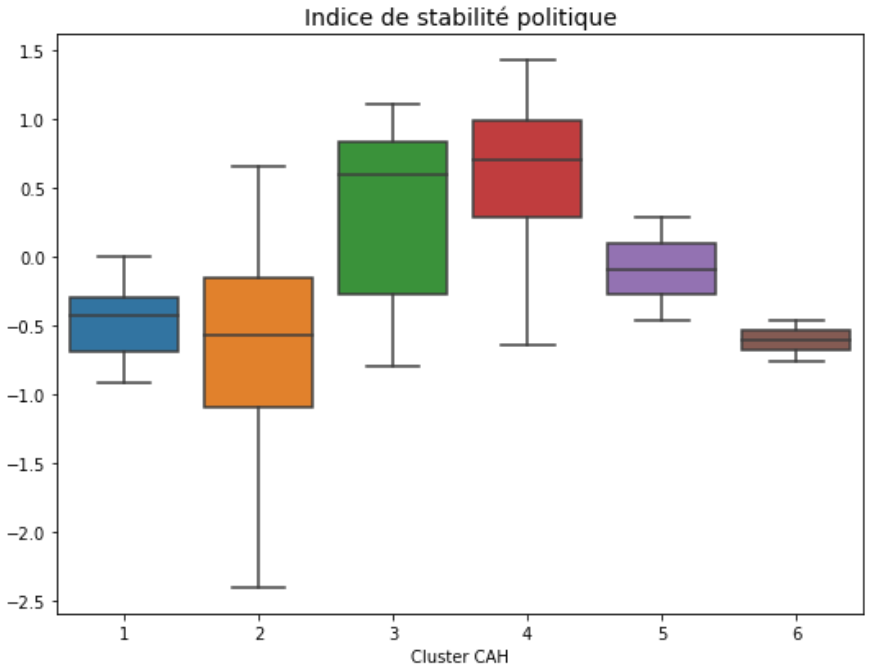


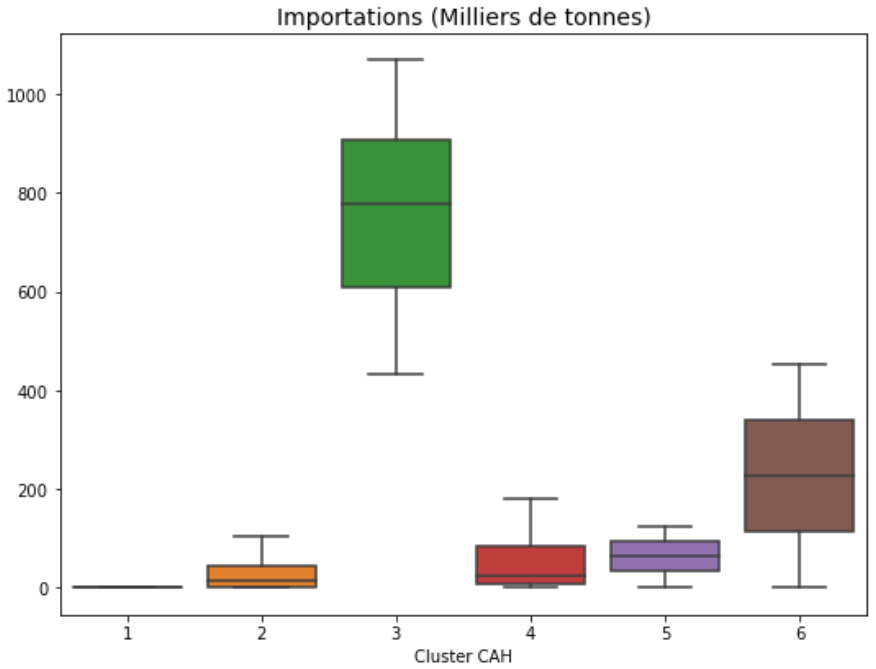


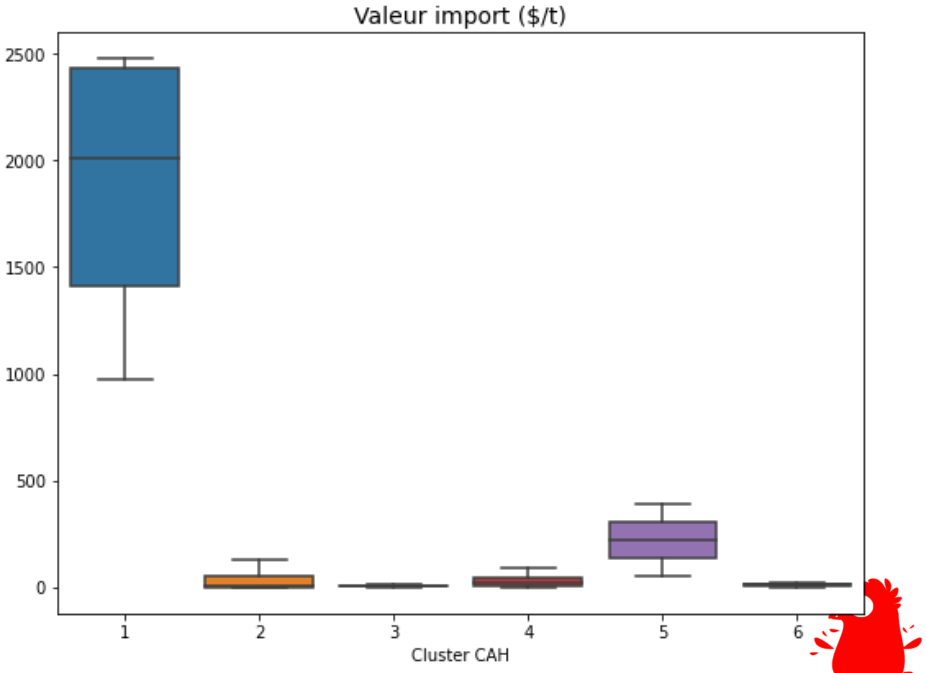
Analyse des clusters

Cluster CAH 1 & Cluster KM 2

Pays producteur de volaille qui importe très peu. Ou la valeur d'import est cependant très élevé mais une stabilité politique faible.







Analyse des clusters CAH et KM retenus (CAH3 & KM6)

Pays	Disponibilité en protéines de volailles (g/pers/jour)	Exportation (Milliers de tonnes)	Importations (Milliers de tonnes)	Production (Milliers de tonnes)	Pop 2017	Disponibilité en protéines animales (g/pers/jour)	PIB/hab ($)	Indice de stabilité politique	Taux douane	Valeur import ($/t)
Afrique du Sud	14.11	63.0	514.0	1667.0	57009.756	35.7	13860.3	-0.28	4.61	3.32
Allemagne	7.96	646.0	842.0	1514.0	82658.409	63.0	53071.5	0.59	1.96	15.85
Arabie saoudite	15.57	11.0	722.0	616.0	33101.179	34.6	47306.2	-0.64	4.85	1.92
Chine - RAS de Hong-Kong	22.26	663.0	907.0	24.0	7306.322	95.7	59842.2	0.83	0.00	1.73
France	8.95	500.0	506.0	1750.0	64842.509	63.7	44577.1	0.28	1.96	4.59
Japon	7.24	10.0	1069.0	2215.0	127502.725	47.7	41444.2	1.11	2.51	1.46
Mexique	9.26	8.0	972.0	3249.0	124777.324	44.0	19721.3	-0.80	1.24	4.88
Pays-Bas	8.48	1418.0	608.0	1100.0	17021.347	68.0	55088.6	0.92	1.96	55.48
Royaume-Uni de Grande-Bretagne et d'Irlande du Nord	13.77	359.0	779.0	1814.0	66727.461	59.0	46372.4	0.39	1.99	8.82
Émirats arabes unis	14.80	95.0	433.0	48.0	9487.203	34.4	67183.6	0.62	4.75	0.41
Belgique	4.57	655.0	338.0	463.0	11419.748	57.7	50442.3	0.43	1.96	116.48
Fédération de Russie	10.44	115.0	226.0	4444.0	145530.082	54.0	25926.4	-0.64	3.61	14.53

Conclusion & recommandations

1) Cibler les pays d'Europe dans un premier temps pour un développement plus rapide : Allemagne, Pays-Bas & Belgique.
   - Ces pays sont dans l'Union Européenne (facilité monétaire, logistique, etc…), et répondent à une position dominante en termes de dispo. en protéines animales, PIB/habitant, mais aussi sur la capacité d'importation.
2) Ensuite pour des raisons logistique et de cout de transport : Arabie-Saoudite, Emirats arabes unis peuvent être un groupe d'exportation.
3) Ensuite pour l'exportation dans ces pays il faudra faire une analyse supplémentaire pour voir si des pays autour pourraient être ciblés : Chine - RAS de Hong-Kong, Japon, Mexique , Royaume-Uni de Grande-Bretagne et d'Irlande du Nord et Fédération de Russie.
4) Les clusters qui sont le plus proche de ceux sélectionnés, si on veut augmenter la liste des pays seraient les clusters CAH4 ou KM1.