Produisez une étude de marché

La poule qui chante

- Contexte
  - Nettoyage et analyse préparatoire
    - Clustering CAH & K-means
      - Analyse en composantes principales
        - Analyse simplifiée en mode ‘Business’

Stéphane Srsa

Contexte

La poule qui chante

La poule qui chante est une entreprise française d'agroalimentaire.

Elle souhaite se développer à l'international. A l'heure actuelle, aucun pays particulier ni aucun continent n'ont été choisis.

La mission

Je suis Data Analyst chez « La poule qui chante »

Ma mission est de proposer une analyse préliminaire qui servira à une étude de marché approfondie.
Cette analyse permet de cibler un groupe de pays ayant les critères nécessaires pour exporter nos produits.

Pour cela, j'ai le choix des données à utiliser pour réaliser l'analyse.

Nettoyage et analyse préparatoire
Création des fonctions

La poule qui chante

Les fonctions

Affichage plus ‘esthétique’ au format HTML:

Text_message(message) et Titre _message(message)

Information sur le fichier:

infos-DF(DF) : format, valeurs manquantes/uniques

Informations sur les pays présents entre deux DF:

Pays_absents(DF1, DF2,L='Fr’): choix entre Fr et Eng

Création d’un tableau listant les indicateurs avec: valeurs = 0 et/ou des NaN à vérifier ou à supprimer

Verif_col(DF)

Graphique de corrélation & Projection sur plan factoriel:

correlation_graph(pca, x_y, features, palette="rocket", legend_fontsize=12, label_fontsize=10, Indicsize=8, arrow_alpha=0.8)

Plans_Factoriels(X_projected, x_y, pca=None, labels = None, clusters=None, alpha=1, figsize=[12,8], marker=".", palette='viridis’)

Nettoyage et analyse préparatoire
Création des fonctions

La poule qui chante

Les fonctions CAH & Kmeans

Dendrogramme et Score de Silhouette:

CAH(DF, scaler = preprocessing.StandardScaler())

Affichage du CAH à n clusters et liste des outliers:

CAH_groupes(DF, scaler = preprocessing.StandardScaler(), Nbc=5, Seuil_Outliers = 5)

CAH – Statistique (moyennes), Tendances des indicateurs, Imputation et liste des clusters:

CAH_Stats(DF, scaler = preprocessing.StandardScaler(), Nbc=5)

Méthode du coude et Scores de silhouette pour le choix du nombre de clusters

Kmeans(DF, scaler = preprocessing.StandardScaler())

Affichage des clusters/centroïdes (Kmeans) et liste des outliers:

Kmeans_Centroides(DF, scaler = preprocessing.StandardScaler(), Nbc=5, Seuil_Outliers=3)

Kmeans – Statistique (moyennes), Tendances des indicateurs, Imputation et liste des clusters:

Kmeans_Stats(DF, scaler = preprocessing.StandardScaler(), Nbc=5)

Nettoyage et analyse préparatoire
Choix des données

La poule qui chante

Datasets

Dataset contenant les données sur les disponibilités, l'import/export et la production.

Table_Volaille = pd.read_csv('Dispo_Alim2010-.csv', sep=';')

Source: FAOSTAT Mise à jour 2023-10-27

Dataset contenant la population totale par pays de 2000 à 2021

Pop_Evol = pd.read_csv('Evol_Demo2000.csv', sep=';')

Source: FAOSTAT Mise à jour 2022-11-10

Datasets contenant les données sur le PIB et sur la consommation de protéines animales

PIB_Hab = pd.read_csv('PIB_Hab.csv', sep=';')

Protein = pd.read_csv('Proteines.csv', sep=';')

Source: FAOSTAT Mise à jour 2023-08-23

Dataset contenant l'indice de stabilité politique de chaque pays

StabPol = pd.read_csv('Stabilité politique-TheWorldBank.csv', sep=';')

Source: THE WORLD BANK

Prix_Prod_Poulet = pd.read_csv('Prix_Prod_Poulet.csv', sep=';')

Dataset personnel permettant de vérifier la concordance 'Pays'

Pays_Etalon = pd.read_csv('PAYS_ETALON150.csv', sep=';')

Nettoyage et analyse préparatoire
Analyses spécifiques de certaines tables

La poule qui chante

Traitements spécifiques :

« Table_Volaille »

- Vérification des zones chinoises

- Utilisation de l'équation pour compléter des valeurs manquantes:
Exportations = Production + Importations + Variation de stock - disponibilité intérieure

« StabPol»

- Traitement des valeurs 'Indice Stabilité politique' notées Nan

Pour toutes les tables, je vérifie la concordance des pays avec le dataset « Pays_Etalon ». J'effectue un traitement si nécessaire.

Nettoyage et analyse préparatoire
Jointures - Ajout d'indicateurs

La poule qui chante

Création des indicateurs:

Taux d'importation :

Table_Volaille['%import']

= (Table_Volaille['Importations - Quantité / Milliers de tonnes'] / Table_Volaille['Disponibilité intérieure / Milliers de tonnes']) *100

Taux d'exportation :

Table_Volaille['%export']

= (Table_Volaille['Exportations - Quantité / Milliers de tonnes'] / Table_Volaille['Disponibilité intérieure / Milliers de tonnes']) *100

Taux de production :

Table_Volaille['%prod']

= (Table_Volaille['Production / Milliers de tonnes'] / Table_Volaille['Disponibilité intérieure / Milliers de tonnes']) *100

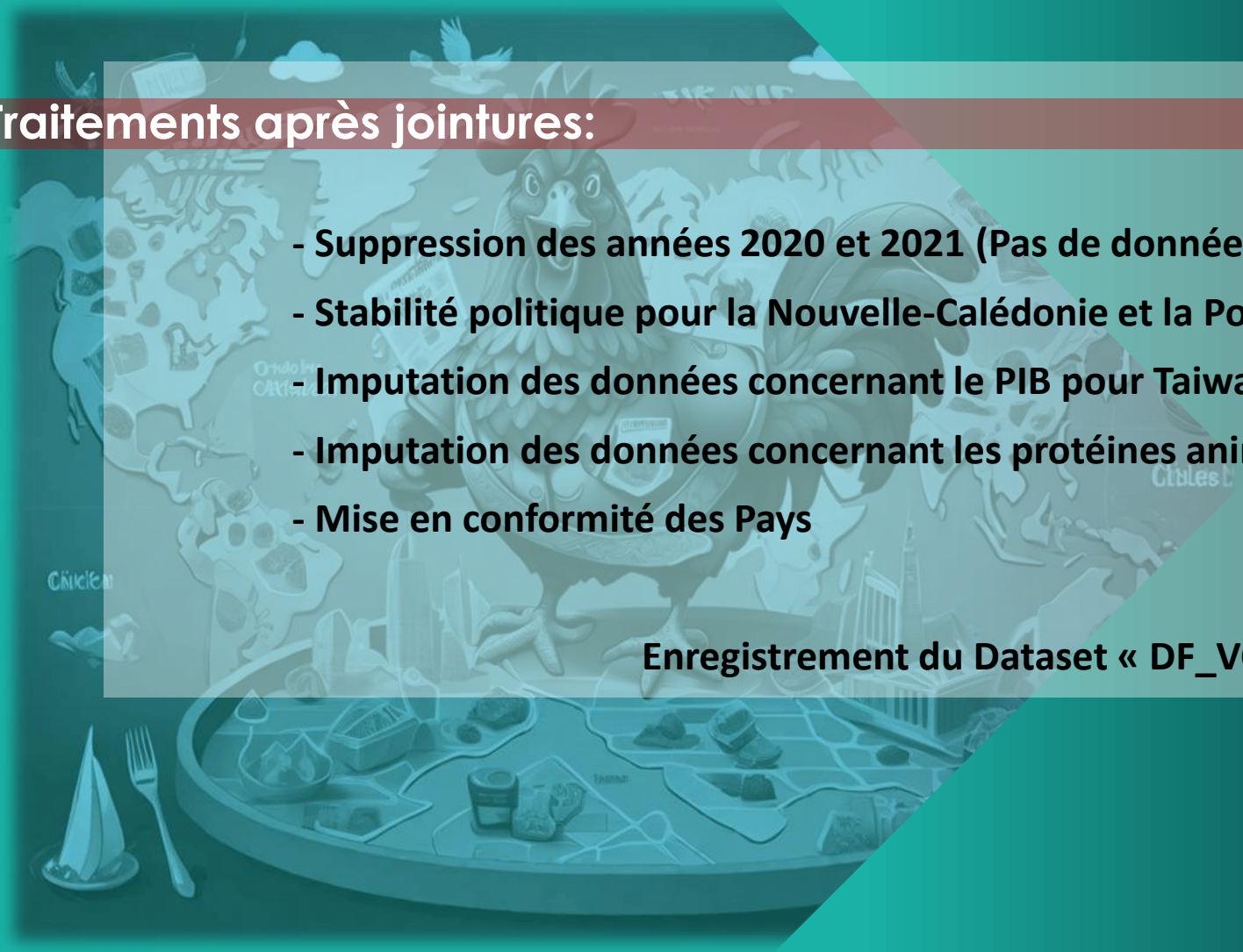
Taux de croissance sur 4 ans:

Debut_period = année - 4

TCCP['%Croiss_Pop'] = (TCCP[année] - TCCP[Debut_period]) / TCCP[Debut_period]) *100

Nettoyage et analyse préparatoire
Jointures – Traitements finaux

La poule qui chante



Traitements après jointures:

- Suppression des années 2020 et 2021 (Pas de données sur les protéines animales)
- Stabilité politique pour la Nouvelle-Calédonie et la Polynésie française
- Imputation des données concernant le PIB pour Taiwan
- Imputation des données concernant les protéines animales pour le Burundi:
- Mise en conformité des Pays

Enregistrement du Dataset « DF_VOLAILLE »

Clustering CAH & K-means
Filtrage des indicateurs

La poule qui chante

Choix de l'année pour l'étude de marché : de 2010 à 2019

J'ai décidé de prendre l'année la plus récente : 2019

Indicateurs à vérifier ou supprimer :

Colonne	Zéro	NaN	
Disponibilité alimentaire (Kcal)	0	0	
Disponibilité alimentaire (Kcal/personne/jour)	0	0	
Disponibilité alimentaire en quantité (kg/personne/an)	0	0	
Disponibilité de matière grasse en quantité (g/personne/jour)	0	0	
Disponibilité de matière grasse en quantité (t)	0	0	
Disponibilité de protéines en quantité (g/personne/jour)	0	0	
Disponibilité de protéines en quantité (t)	0	0	
Disponibilité intérieure / Milliers de tonnes	0	0	
Exportations - Quantité / Milliers de tonnes	82	0	A Suprimer
Importations - Quantité / Milliers de tonnes	20	0	A Verifier
Nourriture / Milliers de tonnes	0	0	
Production / Milliers de tonnes	11	0	A Verifier
Résidus / Milliers de tonnes	175	0	A Suprimer
%import	20	0	A Verifier
%export	82	0	A Suprimer
%prod	11	0	A Verifier
Pop(Million)	0	0	
PIB / Croissance annuelle US$ par habitant %	0	0	
PIB / Croissance annuelle US$ %	0	0	
PIB / Valeur US $ par habitant USD	0	0	
PIB / Valeur US $ Millions d'USD	0	0	
Indice Stabilité politique	0	0	
Moy_Prot_Animale	0	0	
Moy_Prot	0	0	
%Prot_animale	0	0	
%Croiss_Pop	0	0	

Format de la Table : (182, 26)

Vérification des indicateurs :

Les pays avec %import = 0

Pays	%prod
Algérie	101.79
Bangladesh	100.0
Belize	100.0
Burkina Faso	100.0
Burundi	100.0
Comores	100.0
Inde	100.11
Indonésie	100.23
Israël	103.01
Kenya	100.0
Madagascar	100.0
Malawi	100.0
Népal	87.5
Ouganda	98.55
Pakistan	100.26
Rwanda	100.0
Soudan	100.0
Sri Lanka	100.54
Sénégal	99.15
Équateur	100.0

Les pays avec %prod = 0

Pays	%import
Antigua-et-Barbuda	100.0
Dominique	80.0
Lesotho	100.0
Luxembourg	108.33
Micronésie (États fédérés de	100.0
Mongolie	100.0
Nauru	100.0
Saint-Kitts-et-Nevis	100.0
Saint-Vincent-et-les Grenadin	100.0
Samoa	100.0
Îles Salomon	80.0

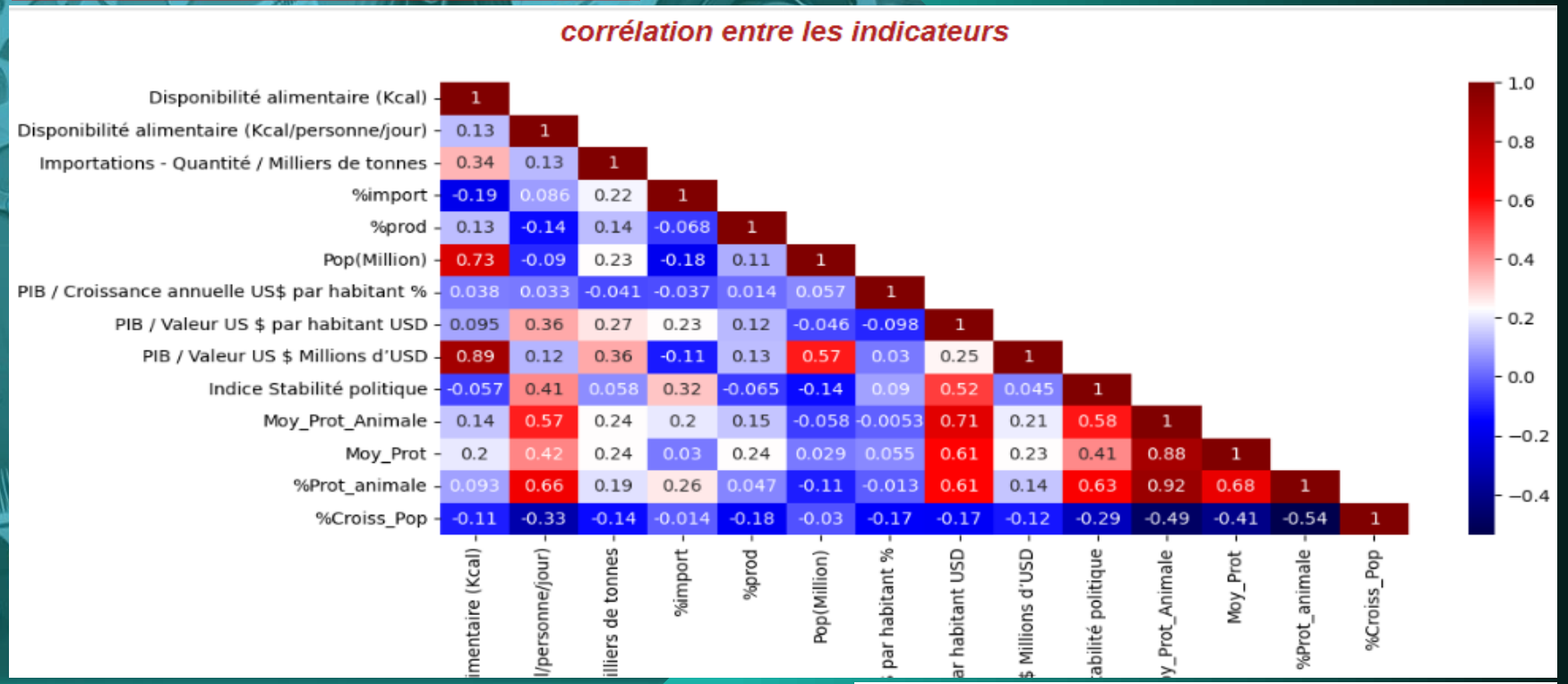
Clustering CAH & K-means
Matrice de corrélation

La poule qui chante

Suppression des indicateurs trop corrélés entre eux

Utilisation d'une boucle pour comparer la corrélation entre indicateurs

Matrice de corrélation :



Le dataset contient 14 indicateurs pour 182 Pays

Clustering CAH & K-means
Prétraitement des données

La poule qui chante

Prétraitement du dataset

Afin de trouver une répartition des clusters la plus homogène et de minimiser l'impact des outliers du fichier, nous testons plusieurs types de prétraitement de données:

- StandardScaler
- Transformation Logarithmique
- MinMaxScaler
- PowerTransformer
- Normalizer

Suite aux tests, nous conservons le prétraitement par Transformation Logarithmique

Clustering CAH & K-means

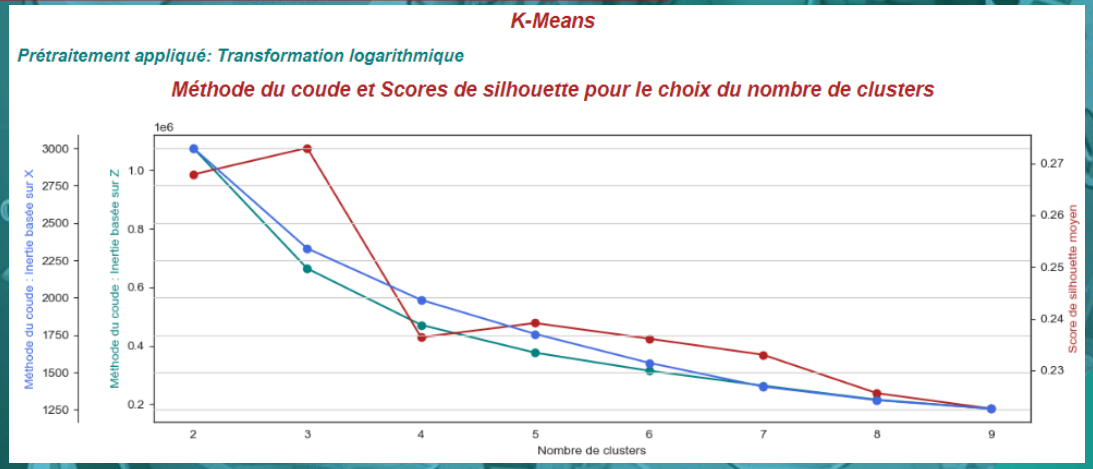
La poule qui chante

Mode projet : Utilisation d'un maximum de variables pour tester les méthodes CAH, K-means et ACP

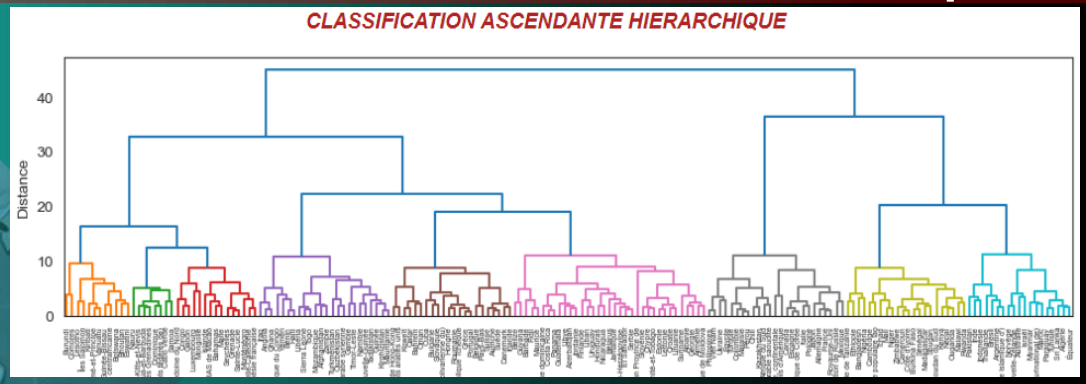
Choix du prétraitement à appliquer :

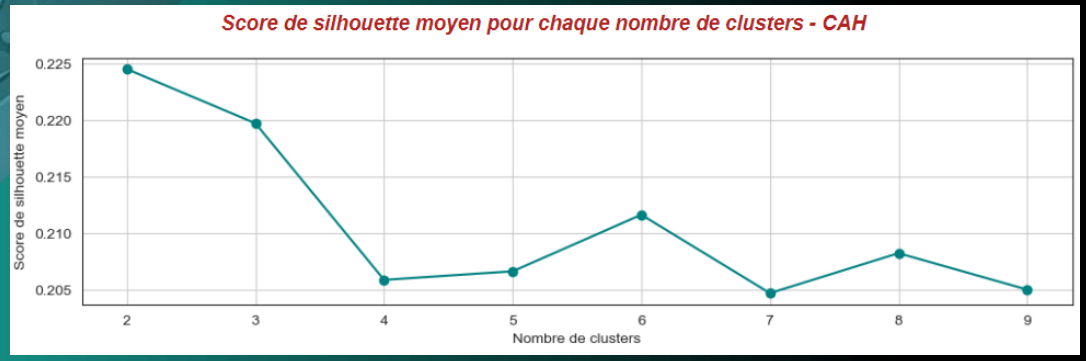
La transformation logarithmique atténue l'effet des outliers

K-means :



Classification ascendante hiérarchique :



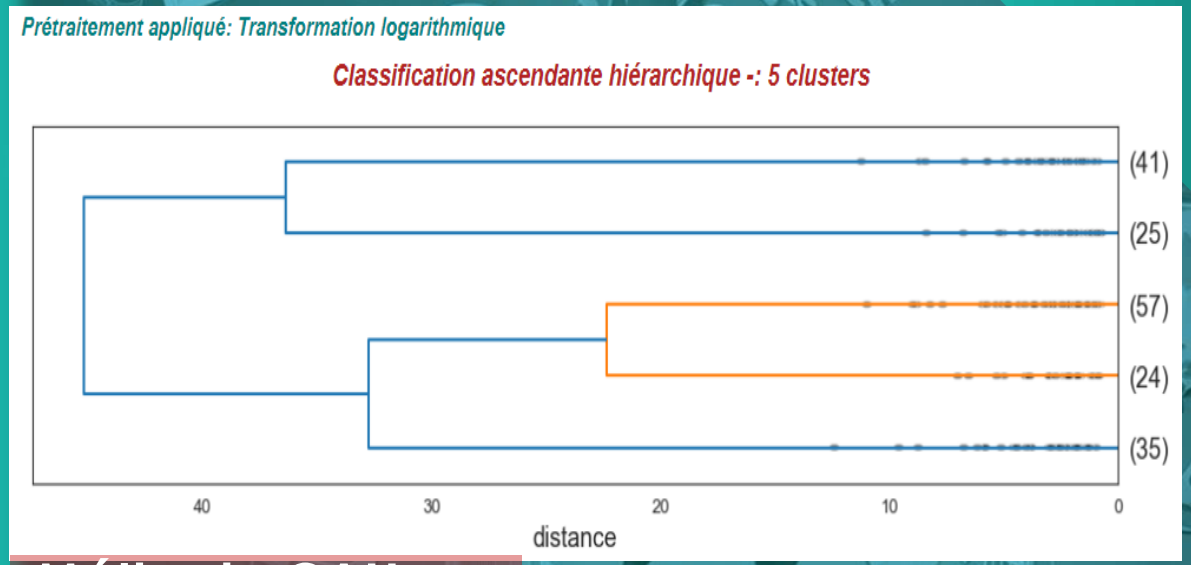


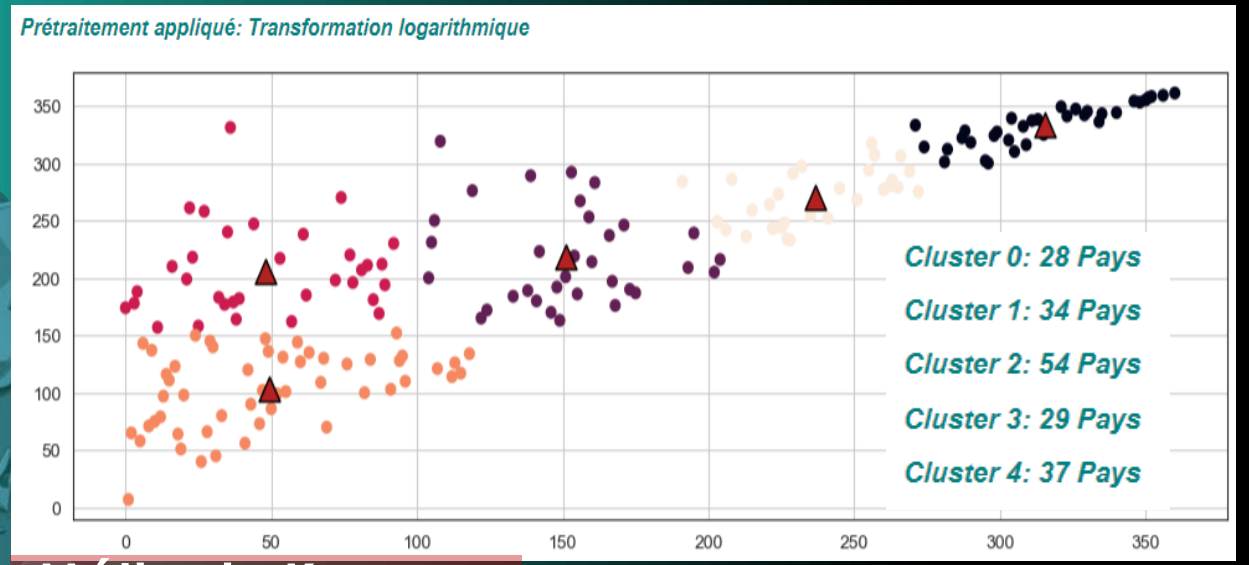
J'opte pour un nombre de clusters = 5 permettant d'obtenir des groupes plus distincts et mieux identifiés.

Clustering CAH & K-means Clusters

La poule qui chante

Choix de la méthode à appliquer :





Méthode CAH :

La CAH considère chaque élément comme un cluster individuel. Il combine progressivement les clusters similaires pour former des clusters plus larges. Le processus se poursuit jusqu'à ce que tous les points fassent partie d'un seul cluster global

Méthode K-mean :

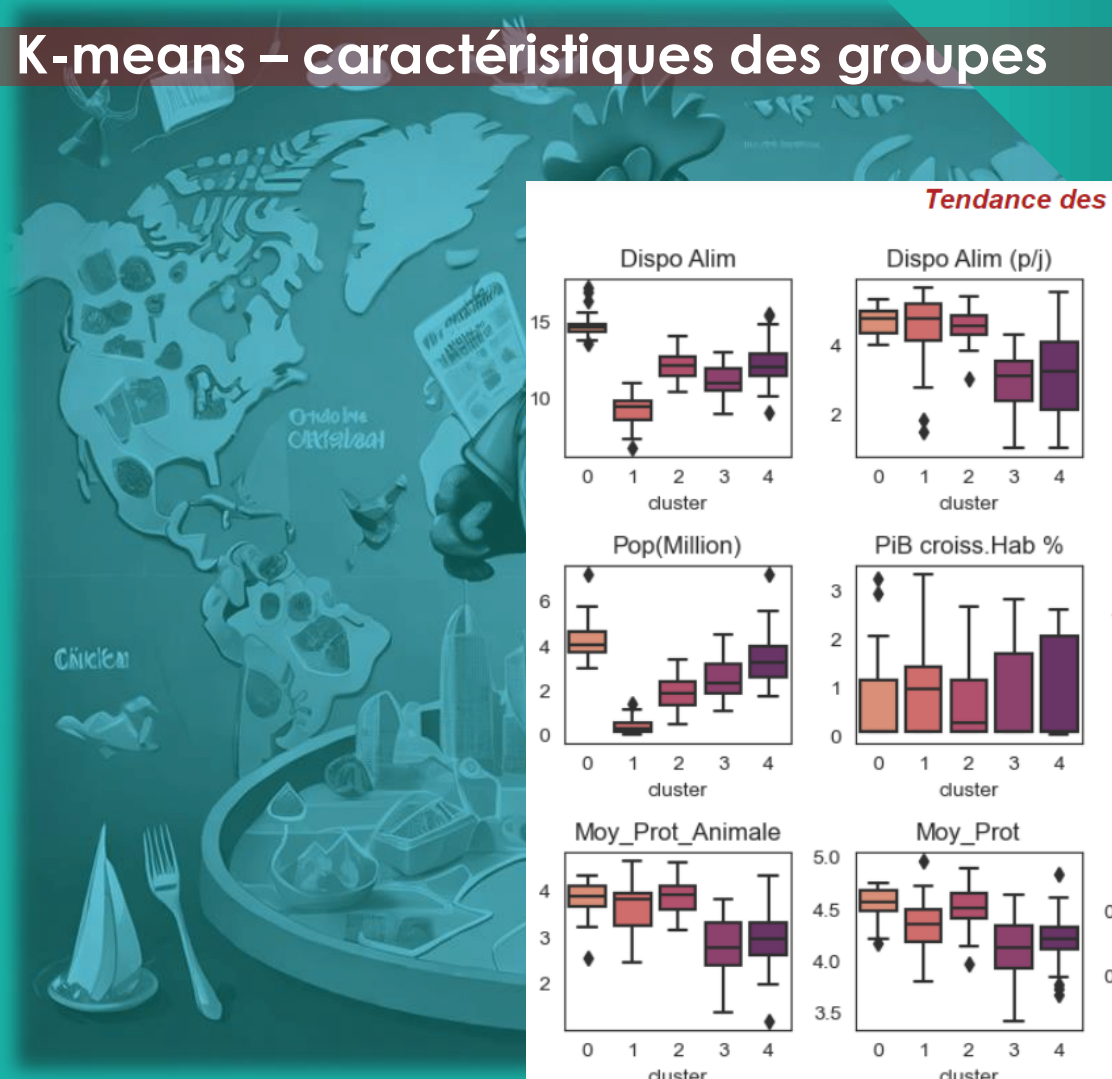
Le K-means commence par placer aléatoirement des centres de clusters, attribue les points de données aux clusters les plus proches, ajuste les centres des clusters et répète ce processus jusqu'à ce que les centres convergent vers des positions où les changements d'attribution des points sont minimes ou jusqu'à ce qu'un critère d'arrêt soit atteint

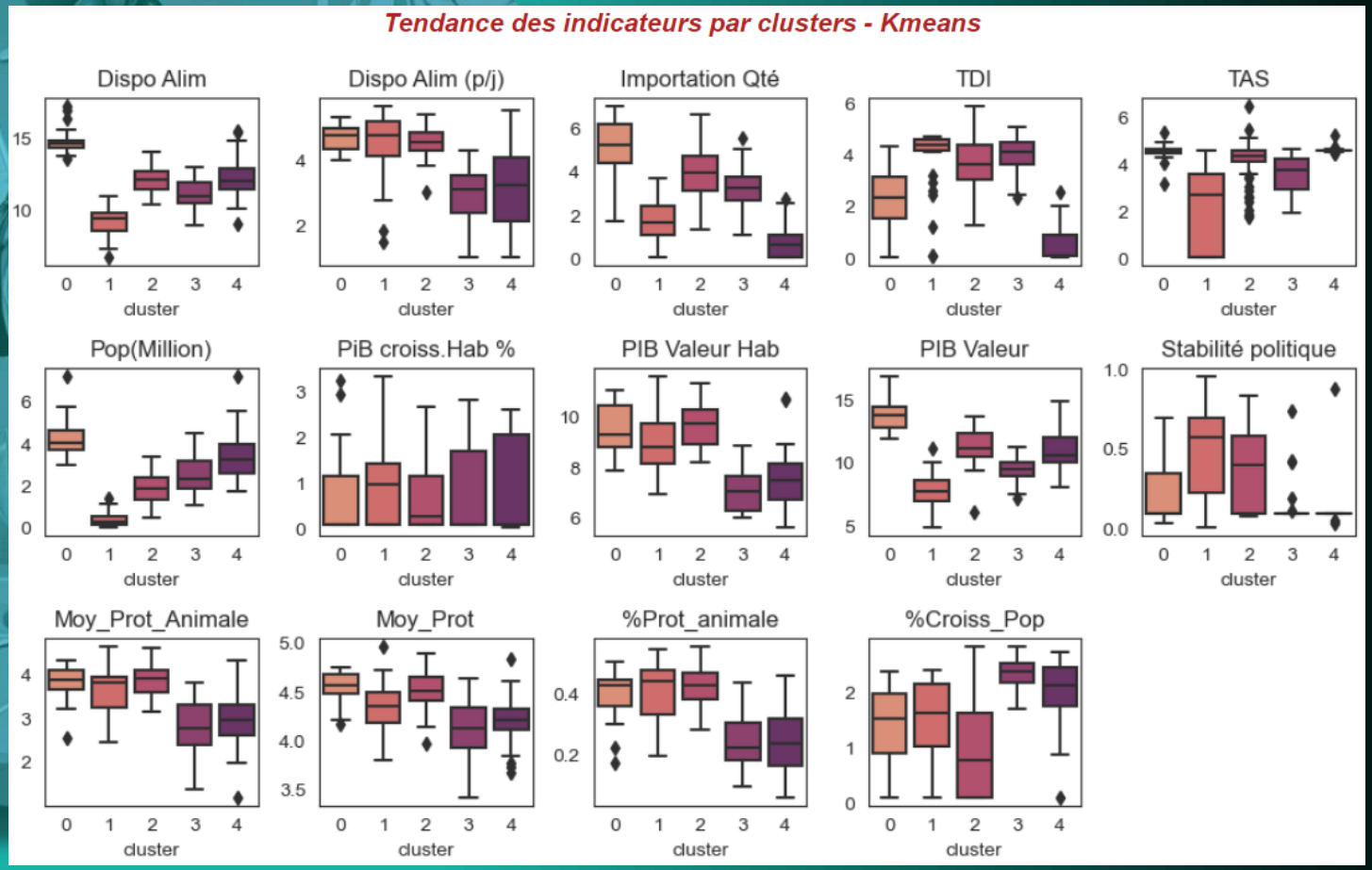
J'utilise le dataset contenant les clusters suivants la méthode K-means (clusters plus homogènes)

Clustering CAH & K-means

La poule qui chante

K-means – caractéristiques des groupes



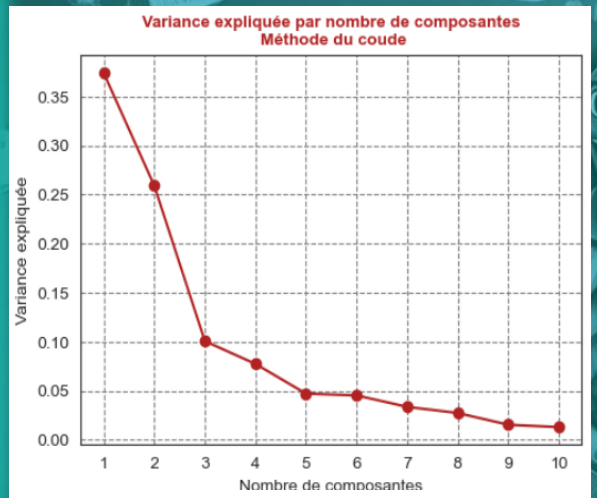


Analyse en composantes principales

La poule qui chante

Nombre de composantes :

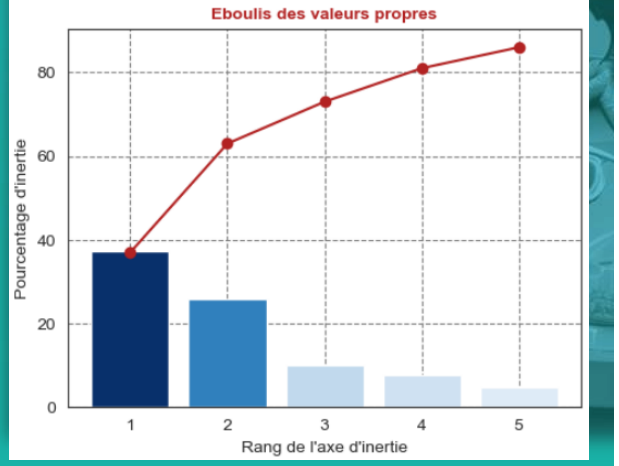
La méthode du coude nous indique que 5 composantes suffisent



Vérification moyennes à 0 et écarts type à 1

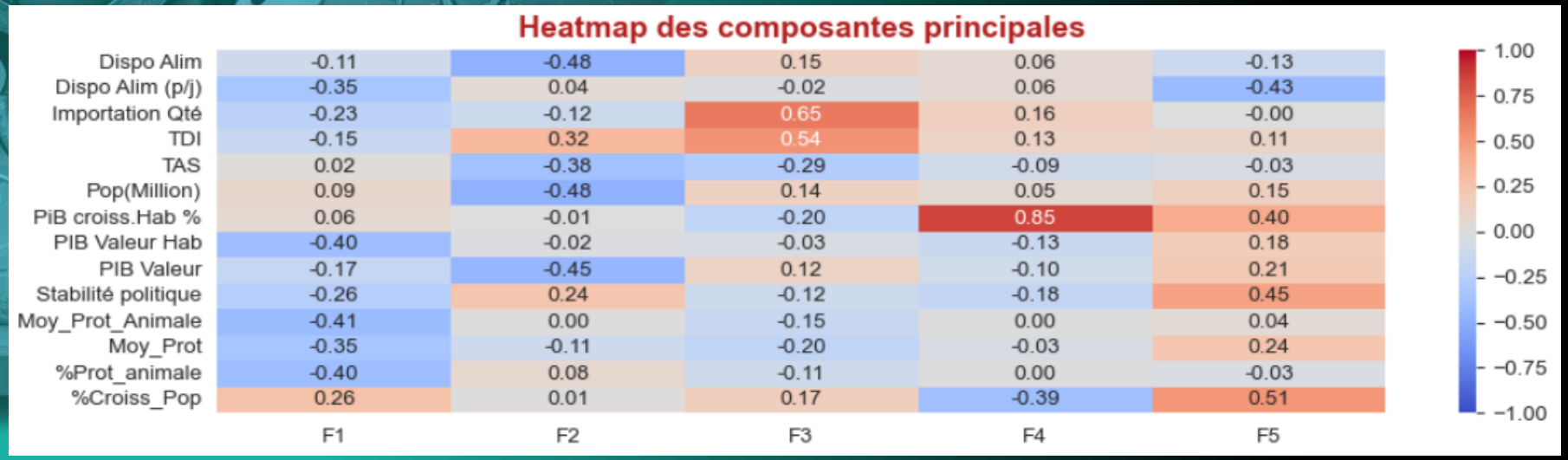
	0	1	2	3	4	5	6	7	8	9	10	11	12	13
mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.0	0.0	0.0	-0.0	-0.0	-0.0
std	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Cumul des valeurs propres : [37. 63. 73. 81. 86.]



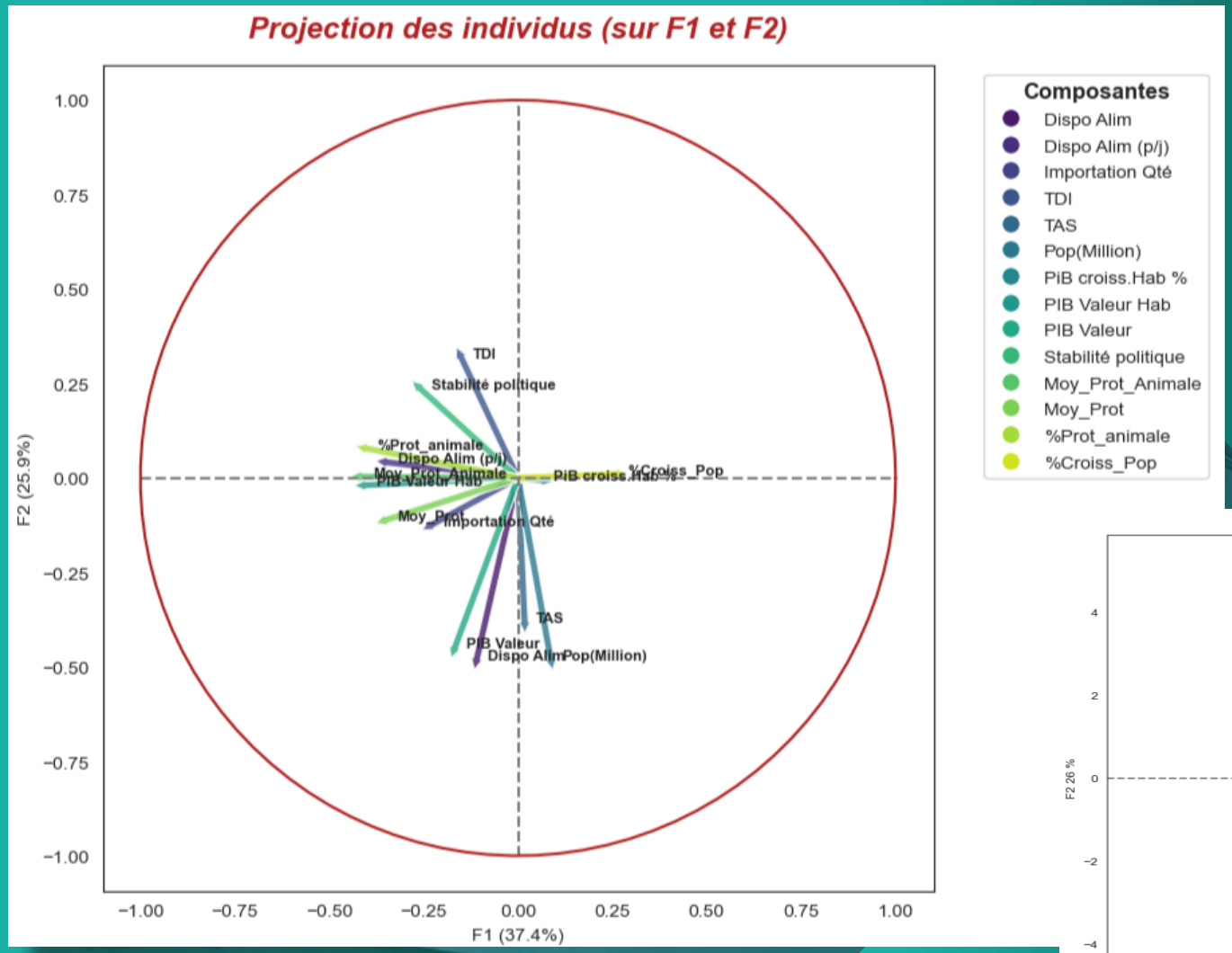
Méthode ACP :

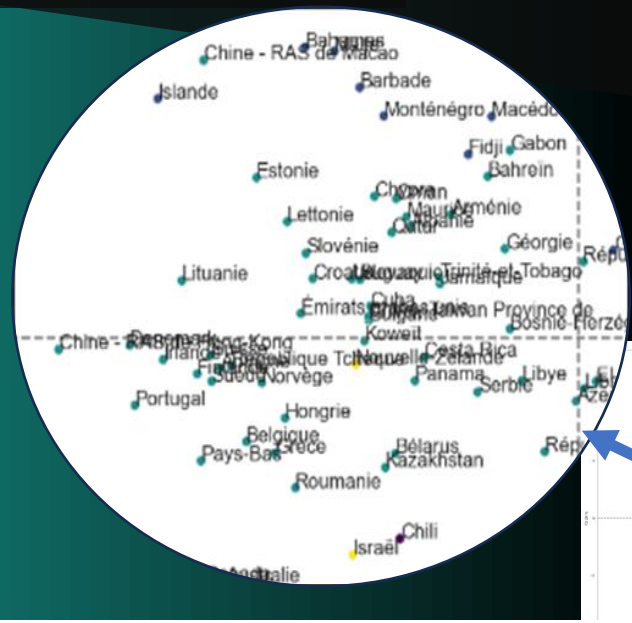
l'ACP simplifie de grandes quantités de données en quelques composantes principales tout en conservant l'essentiel de l'information. Cela permet de mieux comprendre les schémas, les similarités et les différences dans les données, facilitant ainsi leur interprétation et leur analyse.

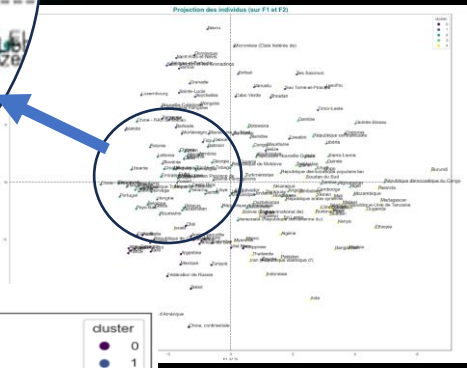


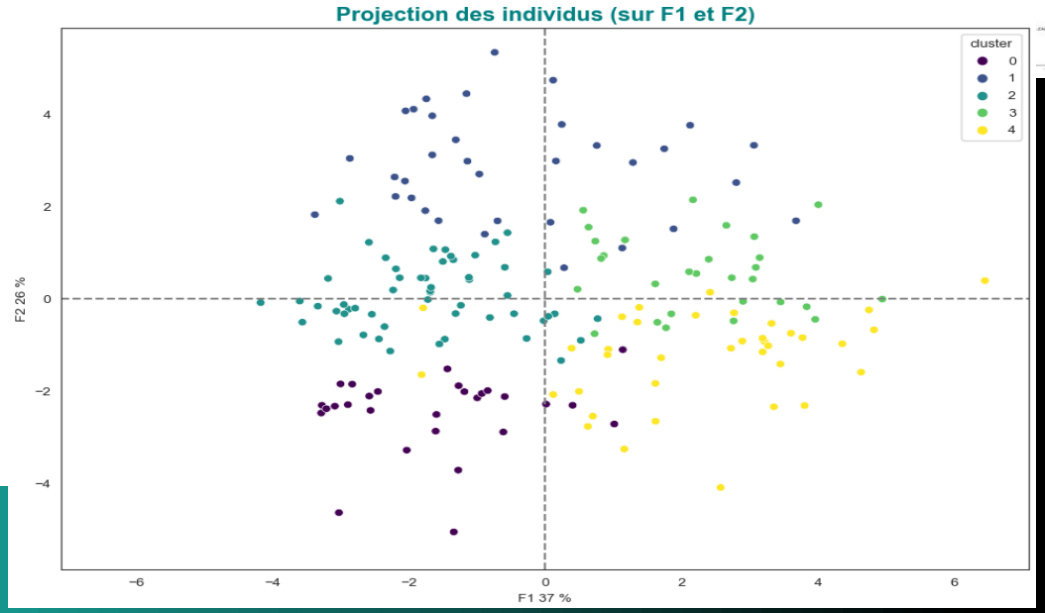
Analyse en composantes principales

La poule qui chante





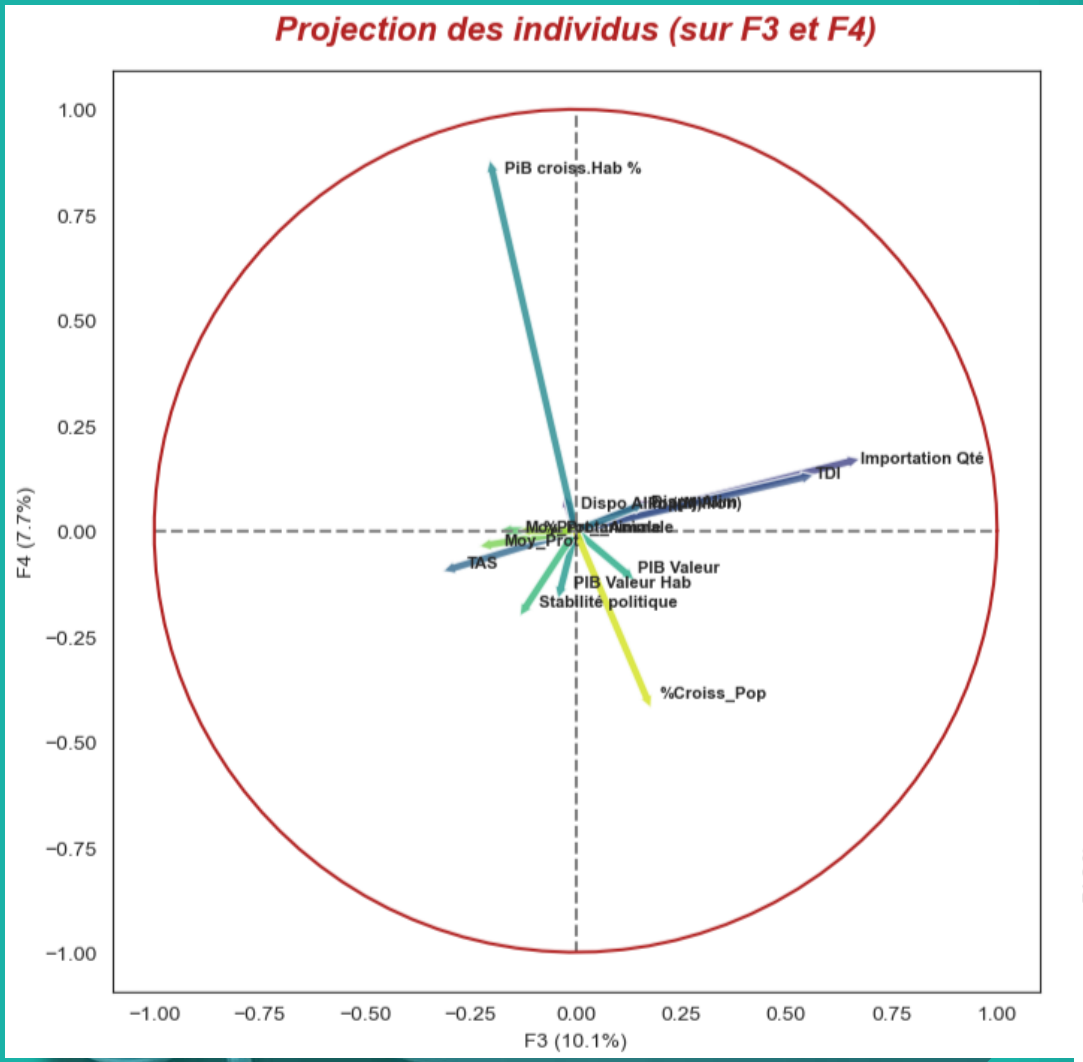


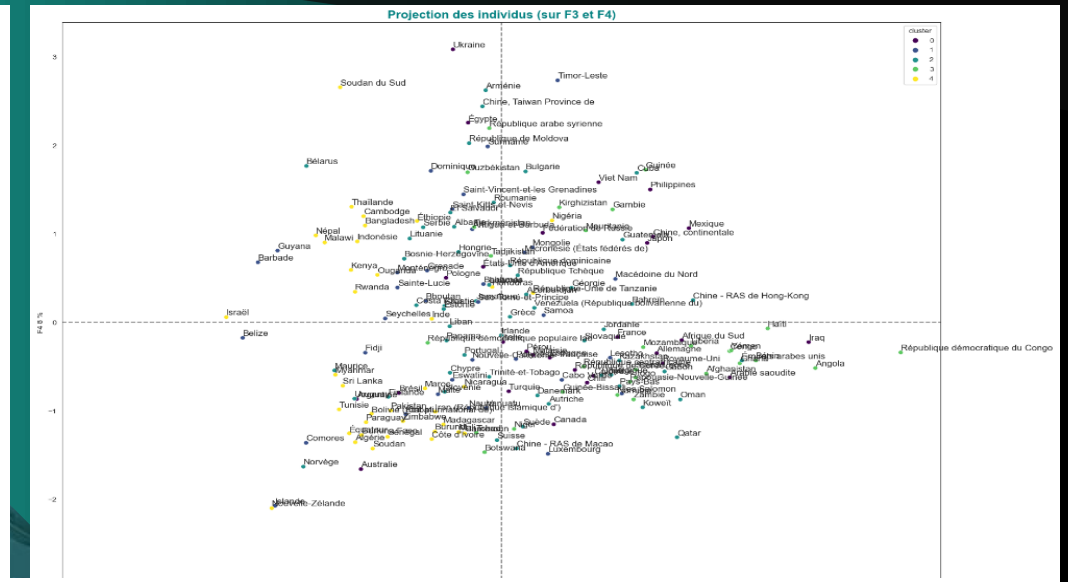


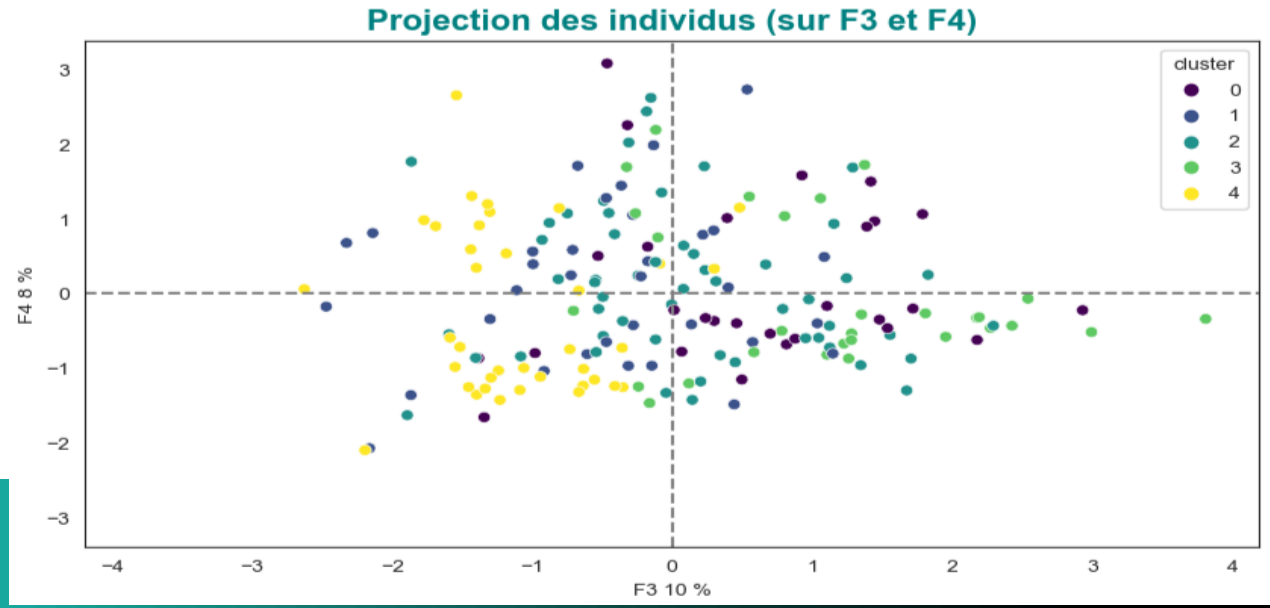
Le clusters 2 est le plus intéressant

Analyse en composantes principales

La poule qui chante







Analyse en composantes principales

La poule qui chante

calcul du score par pays :

['Score'] = (['Dispo Alim']* 0,1 + ['Dispo Alim (p/j)']*0.7
+ ['Importation Qté']*1.5 + ['TDI']*2 + ['TAS']*0.5
+ ['%Croiss_Pop']*1.3 + ['PiB croiss.Hab %'] *1.4
+ ['%Prot_animale']*2 + ['Stabilité_Politique']*0.4
+ ['Pop(Million']*0,1)

Critères de filtrage:

Pop(Million) > 4 Stabilité politique > -0,5 Score > 33

Recommandations :

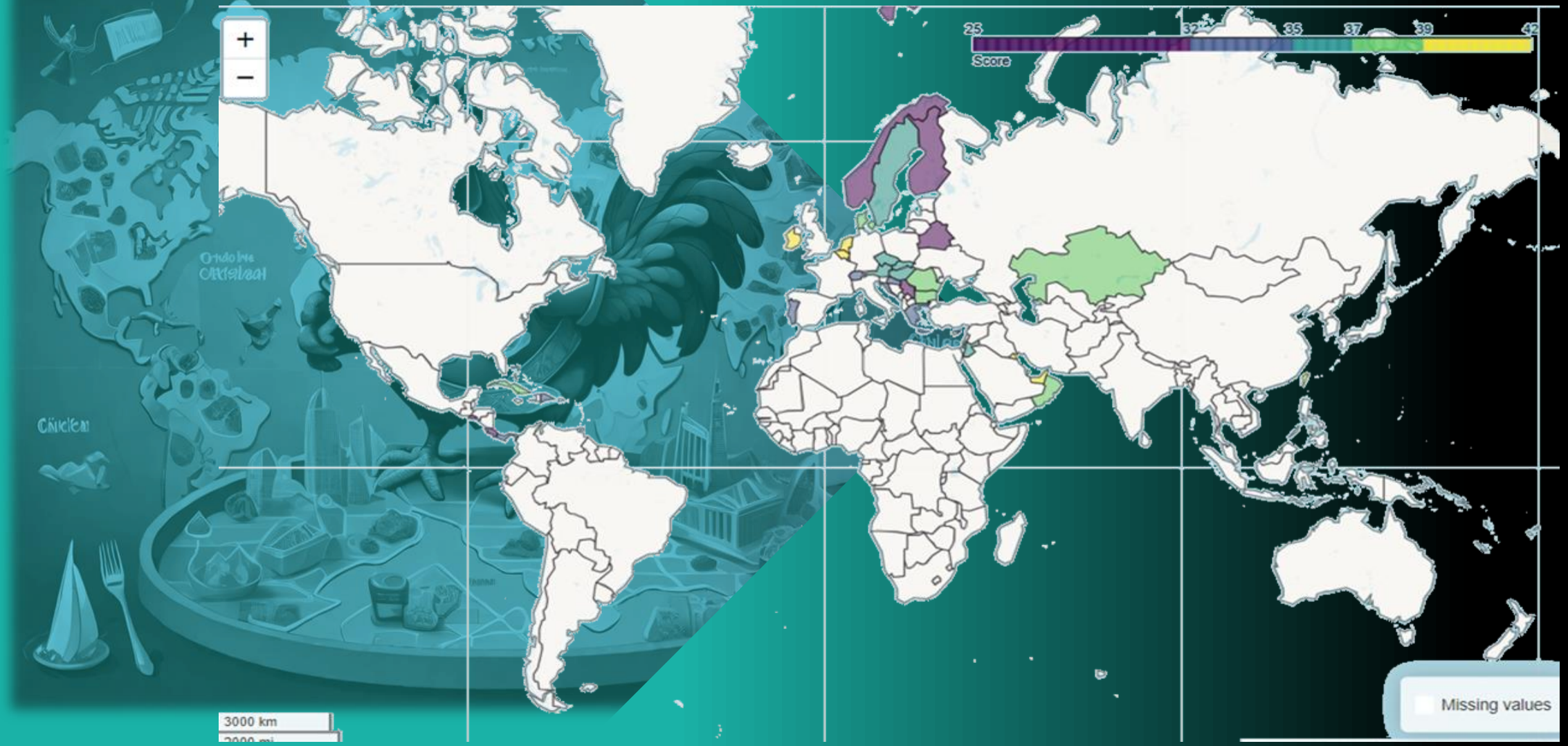
- Vérifier le niveau de coûts de production de volailles (Kg ou Tonne) par rapport à la France
- Vérifier auprès de la COFACE les informations économiques et sectorielles
- Les Pays ayant un score supérieur à 35 sont à privilégier
- Pour les Pays musulmans -- Conformité/Certification produits Halal
- D'un point de vue écologique (Transport), privilégier les Pays Européens + Ireland

Pays	iso_a3	Score	Pop(Million)	Stabilité politique
Pays-Bas	NLD	41.70	17.36	0.82
Chine - RAS de Hong-Kong	HKG	41.04	7.50	-0.23
Chine, Taiwan Province de	TWN	40.46	23.78	0.79
Émirats arabes unis	ARE	39.93	9.21	0.67
Belgique	BEL	39.72	11.51	0.46
Irlande	IRL	38.96	4.90	0.96
Koweït	KWT	38.79	4.44	0.18
Cuba	CUB	38.53	11.32	0.61
Kazakhstan	KAZ	37.72	18.75	-0.17
Danemark	DNK	37.65	5.80	0.97
Oman	OMN	37.49	4.60	0.59
Bulgarie	BGR	37.04	7.05	0.56
Roumanie	ROU	36.62	19.52	0.54
Jordanie	JOR	36.43	10.70	-0.27
Autriche	AUT	36.27	8.88	0.89
République Tchèque	CZE	35.78	10.54	0.94
Suède	SWE	35.25	10.27	1.01
Slovaquie	SVK	34.97	5.45	0.67
Hongrie	HUN	34.87	9.77	0.76
République dominicaine	DOM	34.59	10.88	-0.00
Portugal	PRT	34.07	10.29	1.05
Suisse	CHE	33.74	8.58	1.31
Panama	PAN	33.33	4.23	0.29
Grèce	GRC	33.22	10.57	0.16

Analyse en composantes principales

La poule qui chante

Visualisation géographique :



Analyse orientée ‘Business’

La poule qui chante

Mode ‘Réalité’ : Utilisation d’indicateurs pertinents

Démarche :

- Sélectionner les indicateurs les plus pertinents de la matrice de corrélations.

DF_F[['Importation Qté', 'TDI', 'TAS', '%Croiss_Pop', 'PiB croiss.Hab %' ,'Moy_Prot' , '%Prot_animale’]

L’indice de stabilité politique et la population en million sont non traités et serviront pour filtrer les pays retenus.

- Chercher une répartition homogène qui minimise l'effet des outliers parmi les différents prétraitements.

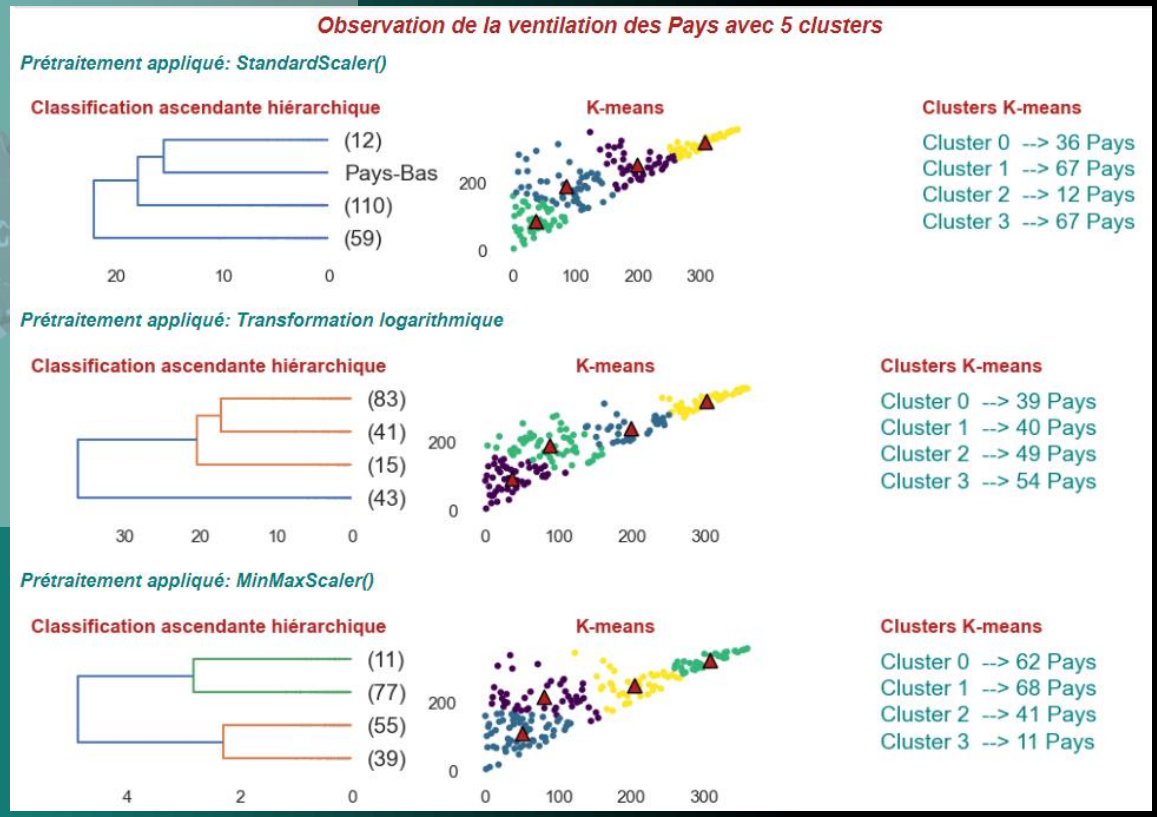
Voir ci-contre (3 prétraitements sur 5)

- Définir le prétraitement et la méthode les plus adaptés

Transformation logarithmique et K-means

- Analyse en composantes principales

- Résultats



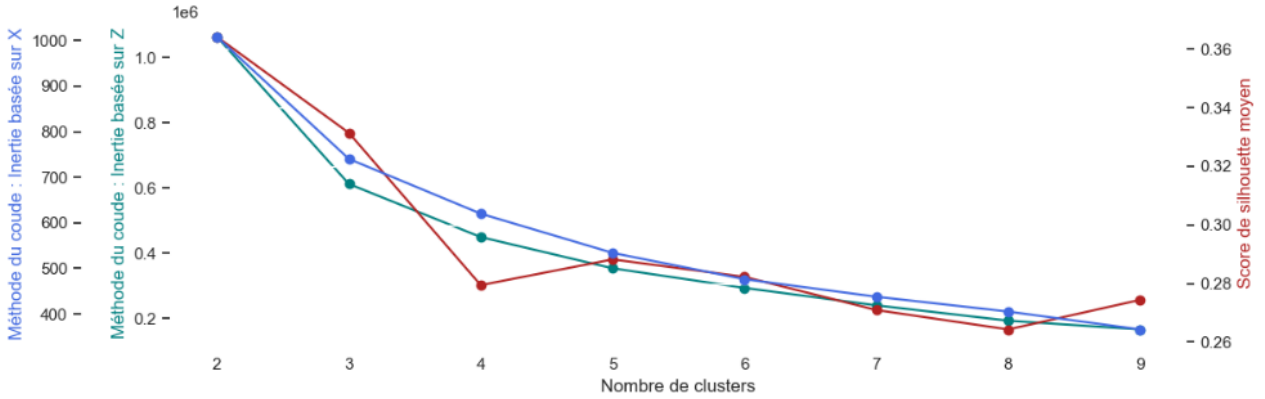
Analyse orientée 'Business'

La poule qui chante

K-Means

Prétraitement appliqué: Transformation logarithmique

Méthode du coude et Scores de silhouette pour le choix du nombre de clusters



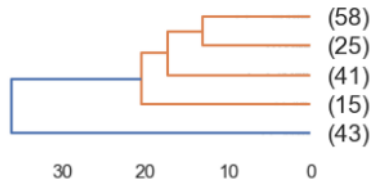
Nombre optimal de clusters conseillé: 2

Pour notre analyse, nous avons besoin d'au moins 4 clusters pour obtenir des groupes bien distincts

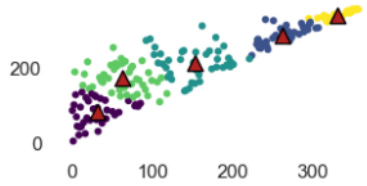
Je choisis 5 clusters (score plus élevé que 4)

Prétraitement appliqué: Transformation logarithmique

Classification ascendante hiérarchique



K-means



Clusters K-means

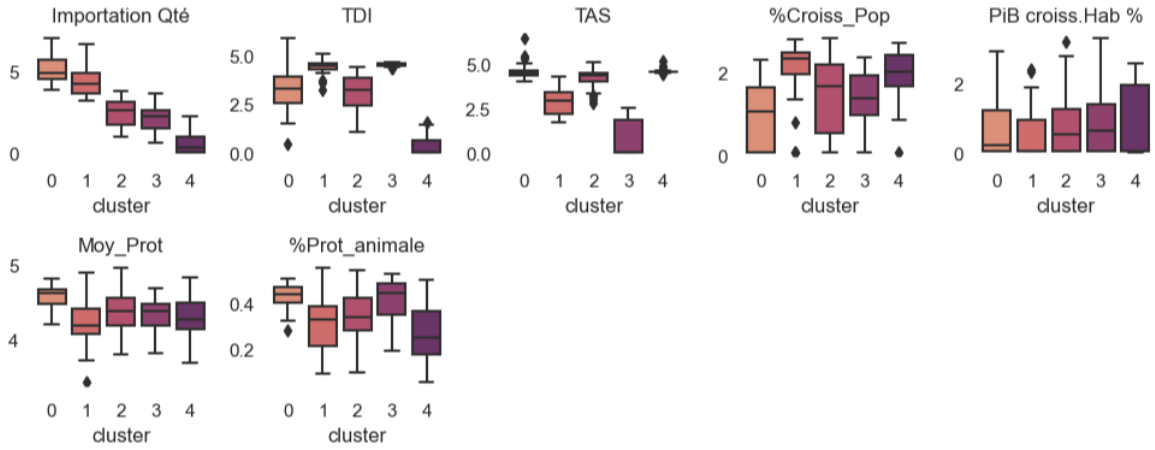
Cluster 0 --> 40 Pays
Cluster 1 --> 30 Pays
Cluster 2 --> 55 Pays
Cluster 3 --> 17 Pays
Cluster 4 --> 40 Pays

Prétraitement appliqué: Transformation logarithmique

Statistiques (moyennes) des clusters - Kmeans

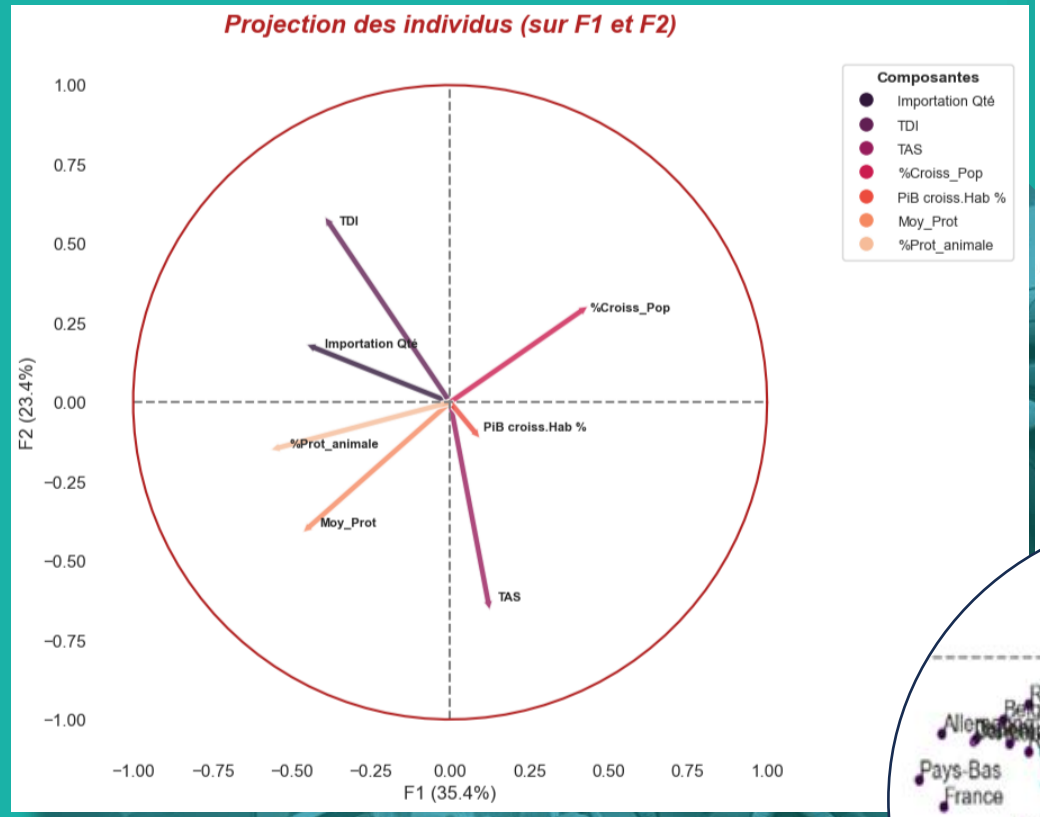
cluster	Importation Qté	TDI	TAS	%Croiss_Pop	PiB croiss.Hab %	Moy_Prot	%Prot_animale	cluster
0	5.212606	3.256072	4.603791	1.026070	0.739611	4.576360	0.434239	0.0
1	4.471941	4.451326	2.952480	2.060729	0.580497	4.240438	0.317105	1.0
2	2.497266	3.146295	4.238884	1.482076	0.857633	4.371715	0.343742	2.0
3	2.194206	4.574036	0.790539	1.382359	0.951318	4.322698	0.419249	3.0
4	0.622164	0.417106	4.647505	1.985784	0.821335	4.295447	0.279321	4.0

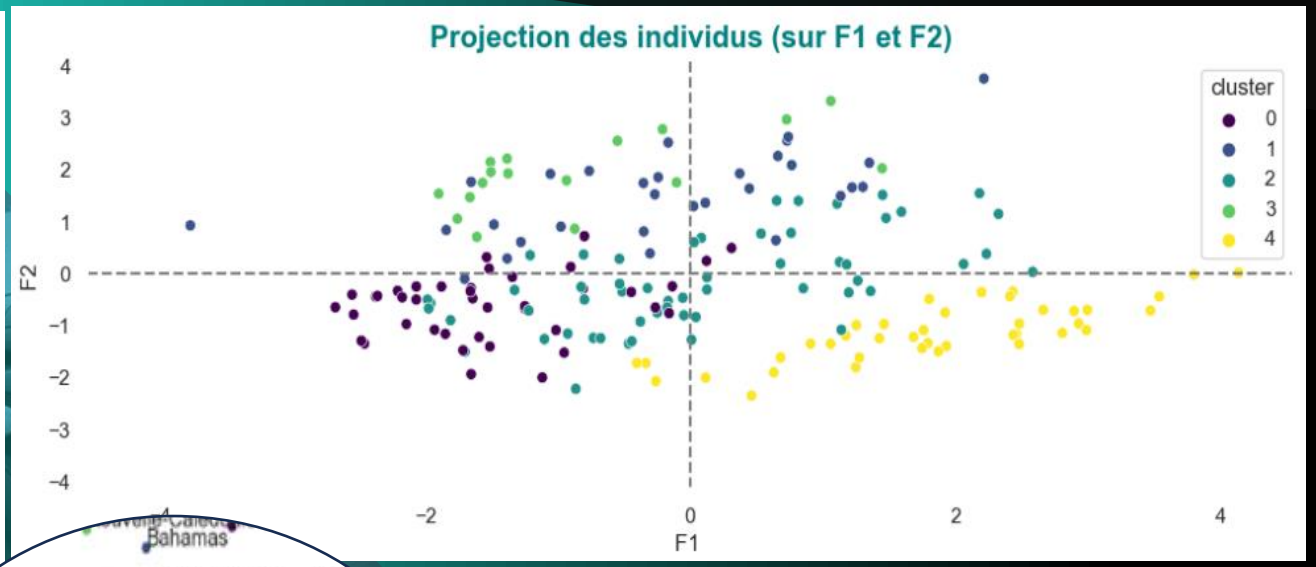
Tendance des indicateurs par clusters - Kmeans

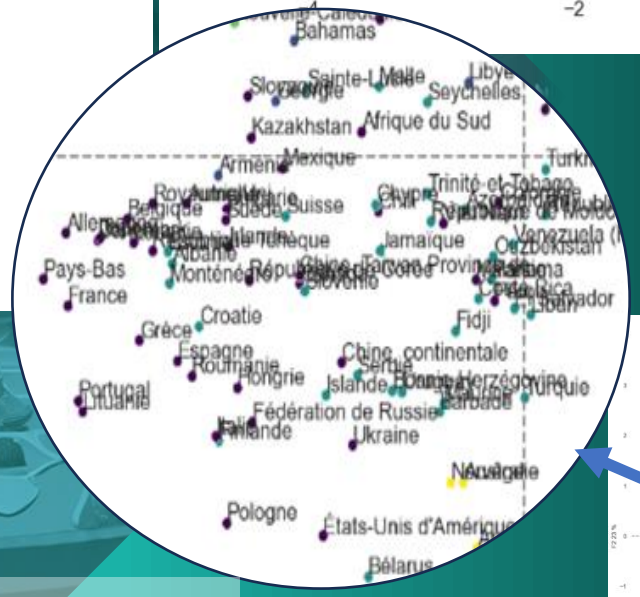


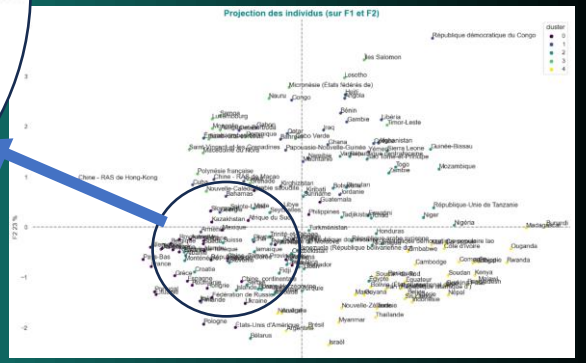
Analyse orientée ‘Business’

La poule qui chante









Le clusters 0 est le plus intéressant

Analyse orientée ‘Business’

La poule qui chante

Calcul du score par pays :

['Score'] = (['Importation Qté']*1.5 + ['TDI']*2 + ['TAS’] *1

+ ['%Croiss_Pop']*1.5 +

['PiB croiss.Hab %']*2 + ['%Prot_animale']*2)

Critères de filtrage:

Pop(Million) > 4 Stabilité politique > 0 Score > 20

Recommandations :

- Vérifier le niveau de coûts de production de volailles (Kg ou Tonne) par rapport à la France

- Vérifier auprès de la COFACE les informations économiques et sectorielles

- Les Pays ayant un score supérieur à 24 sont à privilégiér (Allemagne, Belgique, Irelande, Pays-Bas, UK - Japon, Taïwan et Viet Nam)

- D'un point de vue écologique (Transport), privilégier les Pays Européens + UK

Pays sélectionnés pour l'étude de marché

Pays	Score	Stabilité politique	Pop(Million)
Pays-Bas	30.70	0.82	17.36
Belgique	27.36	0.46	11.51
Chine, Taiwan Province de	26.14	0.79	23.78
Japon	26.00	1.02	125.79
Irlande	25.72	0.96	4.90
Allemagne	25.14	0.55	83.15
Viet Nam	25.02	0.04	95.78
Royaume-Uni	24.56	0.53	66.78
Bulgarie	24.28	0.56	7.05
Danemark	24.22	0.97	5.80
Autriche	23.21	0.89	8.88
France	23.04	0.27	64.40
Roumanie	22.77	0.54	19.52
République Tchèque	21.90	0.94	10.54
Suède	21.88	1.01	10.27
Slovaquie	21.85	0.67	5.45
Canada	21.41	0.99	37.52
Hongrie	21.39	0.76	9.77
République de Corée	20.59	0.55	51.80
Espagne	20.32	0.29	47.13

Analyse orientée ‘Business’

La poule qui chante

Visualisation géographique :