OPENCLASSROOM

LA POULE QUI CHANTE

PROJET 9

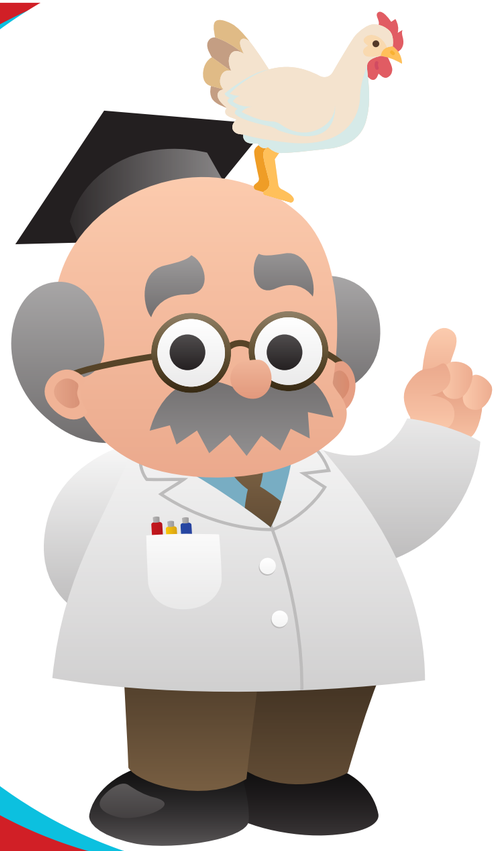
Etude de marché

Identification des pays propices à une insertion dans le marché du poulet

Réalisé par **YU TING HUANG**

LA POULE QUI CHANTE

Missions:
"La poule qui chante" est une entreprise francaise d'agroalimentaire fictive qui souhaite se développer à l'international.
Ils souhaitent avoir une première analyse des pays qui peut être intéressant pour l'export de poulet ainsi que approndir l'étude de marché.



PLAN

- **Pré-traitement des données**
- **Analyse exploratoire de données (AED)**
- **Exploration et partitionnement**
  - Clustering Ascendant Hiérarchique
  - K-means
- **Affinage des clusters**
  - Analyse en Composante Principale (ACP)

LA POULE QUI CHANTE

Données utilisées

Alimentaire

- **Dispo. alimentaire par habitant (calorie)**
- **Prop. de protéines d'origine de volaille (Protéine_volaille(%))**
  - Formule : protéine volaille / protéines totales (végétal + animal)

TDI & TAS (calculés)

- **Taux de dépendance aux importation**
  - Formule : importation / dispo. intérieur * 100
- **Taux d'auto-suffisance**
  - Formule : prod. / dispo. intérieur * 100

Population

- **Nombre de habitants - population en 2017**
- **Taux de croissance de la population sur 5 ans (Croissance_popu. (%))**
  - Formule : (popu. 2017 - popu. 2012) / 2012 popu. *100
- **Taux de croissance du pouvoir d'achat (PIB / HAB) sur 5 ans**
  - Formule : (PIB 2017 - PIB 2012) / 2012 population *100

Production de volaille

- **Quantité production par habitant**
- **Quantité de la production de volaille**
  -Formule : Prod. de volaille / popu. en 2017

Membre de l'EEE

(L'Espace économique européen (EEE))

- 0: non EEE
- 1: EEE

Indice de risque

- **CountryRisk**

Géographique

- **Distance à la France (pour chaque pays du monde)**

source FAO /WIKI/autre

Nettoyage

LA POULE QUI CHANTE

Filtre sur les pays:

- Suppression de la Corée du Nord
- Suppression des pays < 500 000 habitants
- **Traitement des doublons**
  - Nom avec l'orthographe différente
    - Différentes sources

```
df_merged5 = df_merged5[~(df_merged5['population_2017'] <= 500)]
df_merged5.head()
```

✓ 0.0s

	Zone	Croissance_popu(%)	population_2017	proteine_volaille(%)	Production
0	Afghanistan	16.477889	36296.113	0.998336	28.0
1	Afrique du	7.906280	57009.756	16.926583	1667.0

Traitement sur les données:

- **L'indice de risque** : **15** valeurs nulles
  - Imputation par la moyenne
  - Conversion en variable quantitative discrète
    - 0: Risque plus fort
    - 7: Risque plus faible
- **Dispo. intérieur** & **importation** : **2** valeurs nulles
  - imputation par la moyenne

```
df_country_risk.replace(['A1','A2','A3','A4','B','C','D','E'],range(8), inplace= True)
df_country_risk.head(3)
```

```
df_country_risk['CountryRisk'].replace(range(7,-1,-1),range(0,8), inplace=True)
```

✓ 0.0s

```
# Il y a 167 lignes dans 'c
df_country_risk.head()
```

✓ 0.0s

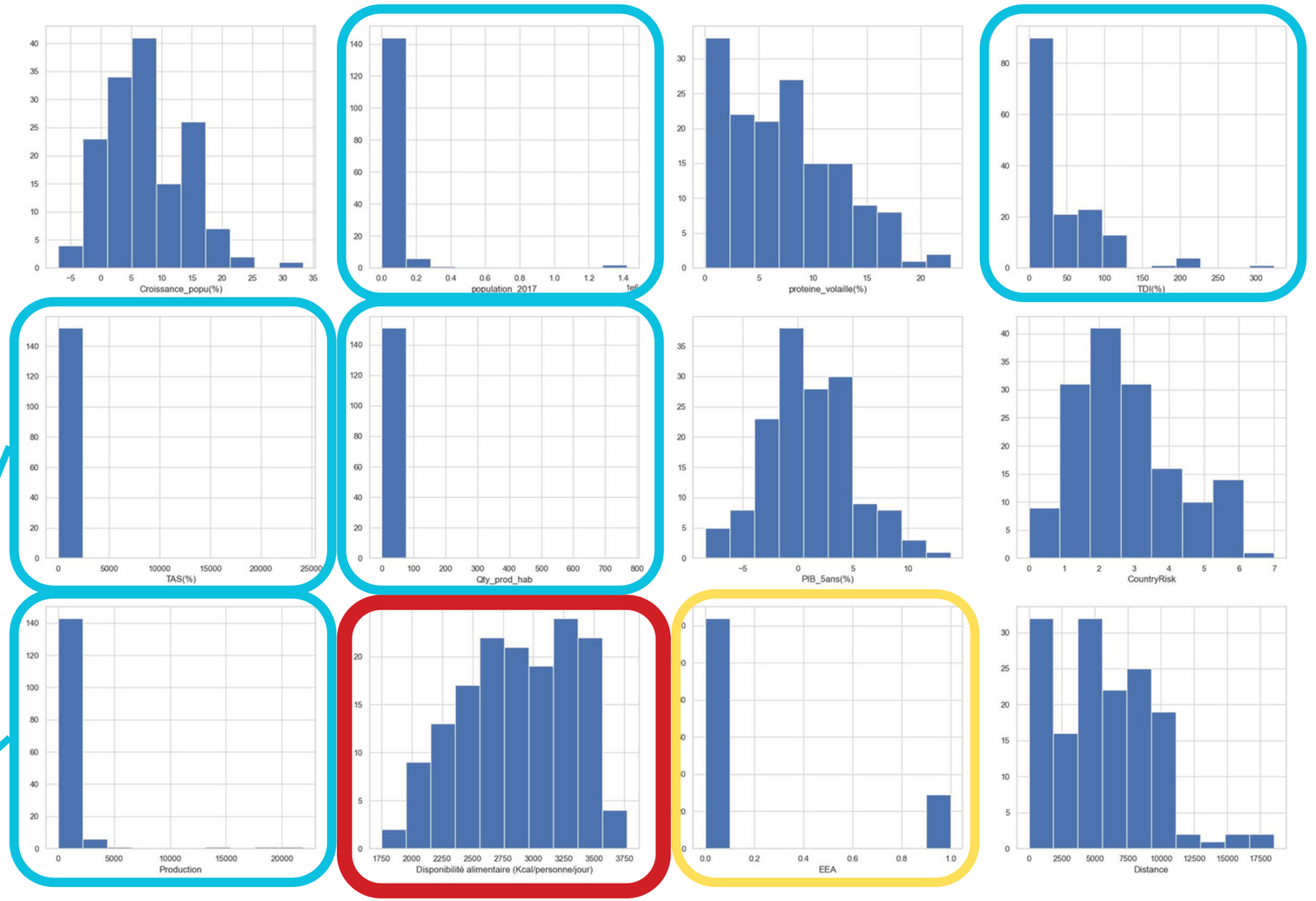
	Zone	CountryRisk
0	Afrique du Sud	2
1	Algerie	2
2	Angola	2
3	Bénin	3
4	Botswana	4

LA POULE QUI CHANTE

Analyse exploratoire de données (AED)

80%

Positive skew



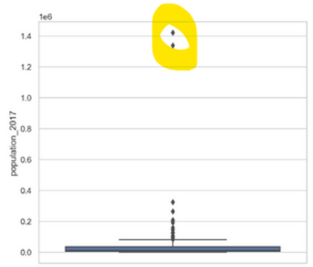
- **TAS (%)**
- **TDI (%)**
- **Population**
- **Production**
- **Quantité prod par hab**

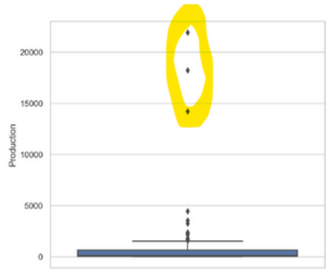
Valeurs extrèmes

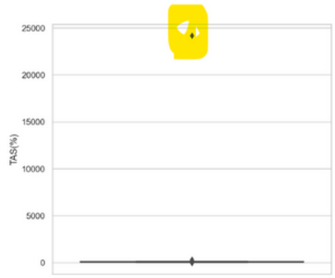
negative skew

LA POULE QUI CHANTE

Population 2017 & Production & TAS (%)







- **Suppression** :
  - États-Unis
  - Chine
  - Brésil
  - Inde

- **Suppression** : Djibouti

```
# Afficher les valeurs extremes de TAS(%)
df_finale.sort_values(by = ['TAS(%)'], ascending = False).head()
```
✓ 0.0s

	Zone	Croissance_popu(%)	population_2017	proteine_volaille(%)	Production	TDI(%)	TAS(%)
39	Djibouti	8.750001	944.099	1.384916	725.190476	100.000000	24173.015873
13	Belgique	3.016502	11419.748	4.509127	463.000000	222.368421	304.605263
106	Pays-Bas	1.366777	17021.347	8.120272	1100.000000	163.440860	295.698925

```
# Afficher les valeurs extremes de production
df_finale.sort_values(by = ['Production'], ascending = False).head()
```
✓ 0.1s

	Zone	Croissance_popu(%)	population_2017	proteine_volaille(%)	Production	TDI(%)	TAS(%)
150	États-Unis d'Amérique	3.515710	325084.756	17.542470	21914.0	0.673382	119.971532
30	Chine	2.659675	1421021.791	3.903785	18236.0	2.488850	100.412973
17	Brésil	4.288544	207833.823	17.255420	14201.0	0.030054	142.266079
47	Fédération de Russie	1.066844	145530.082	10.248356	4444.0	4.960492	97.541703
60	Inde	5.759020	1338676.785	1.141553	3545.0	0.000000	96.831467

- TDI ↘
- TAS ↗
- Grand producteur
- Indice de risque

```
# Afficher les valeurs extremes de population_2017
df_finale.sort_values(by = ['population_2017'], ascending = False).head()
```
✓ 0.2s Python

Zone	Croissance_popu(%)	population_2017	proteine_volaille(%)	Production	TDI(%)	TAS(%)	Disponibilité alimentaire (Kcal/personne/jour)	Qty_prod_hab	PIB_5ans(%)	Countryl
Chine	2.659675	1421021.791	3.903785	18236.0	2.488850	100.412973	3197.0	12.833019	7.983015	
Inde	5.759020	1338676.785	1.141553	3545.0	0.000000	96.831467	2515.0	2.648137	4.694857	

DataFrame :
148pays x 14 variables

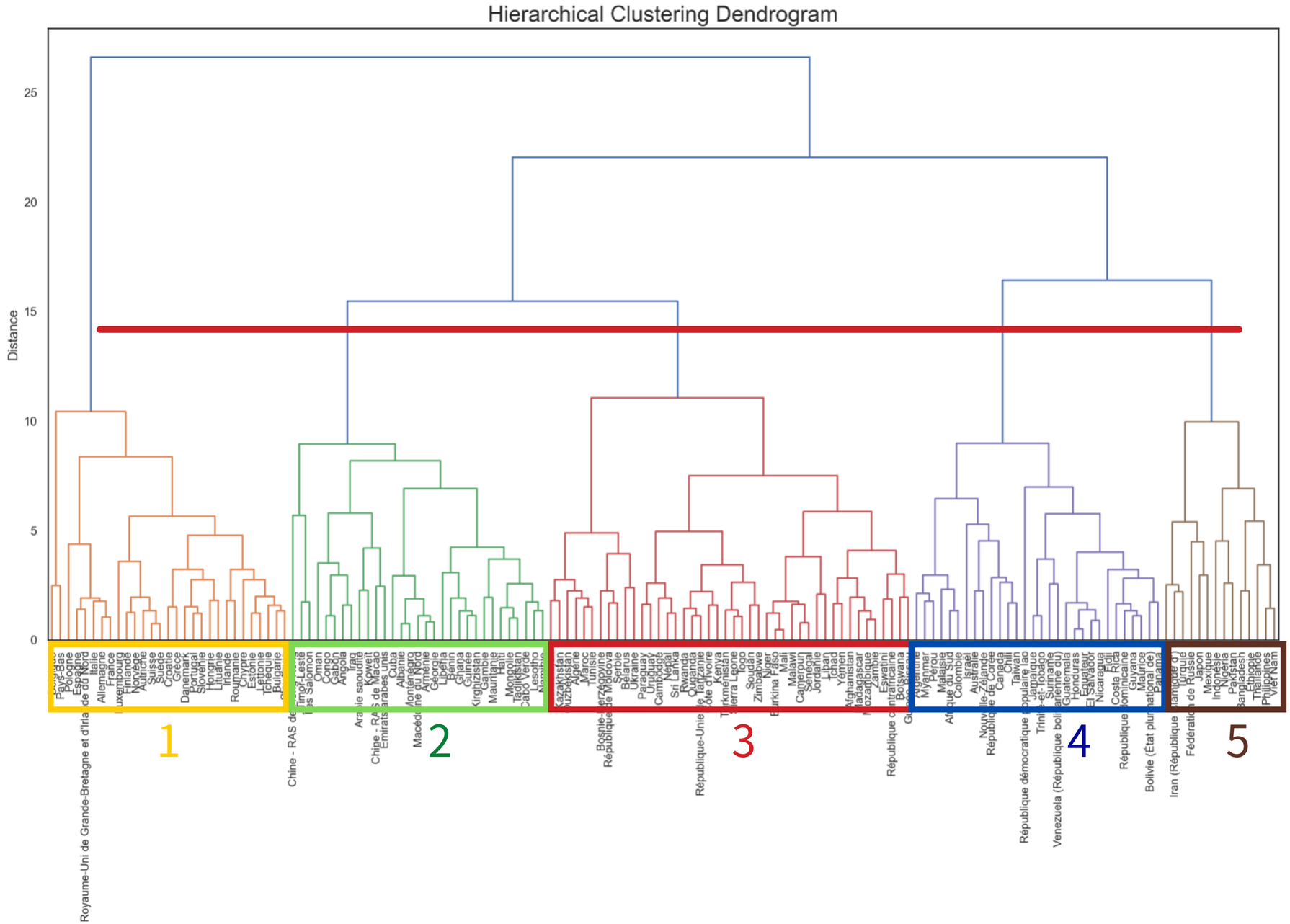
LA POULE QUI CHANTE

Corrélations entre les différentes variables

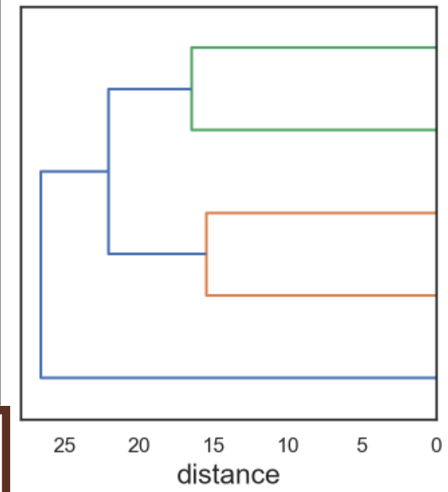
Pays à cibler :

- **TDI(%) élevé + TAS(%) faible**
- **TDI(%) élevé + TAS(%) élevé**
- **CountryRisk élevée**
- **Appartenance à l'EEE.**
- **Distance plus courte (aspect écologique)**
- **Taux de croissance démographique faible**
- **Disponibilité alimentaire élevée**

Clustering Ascendant Hiérarchique

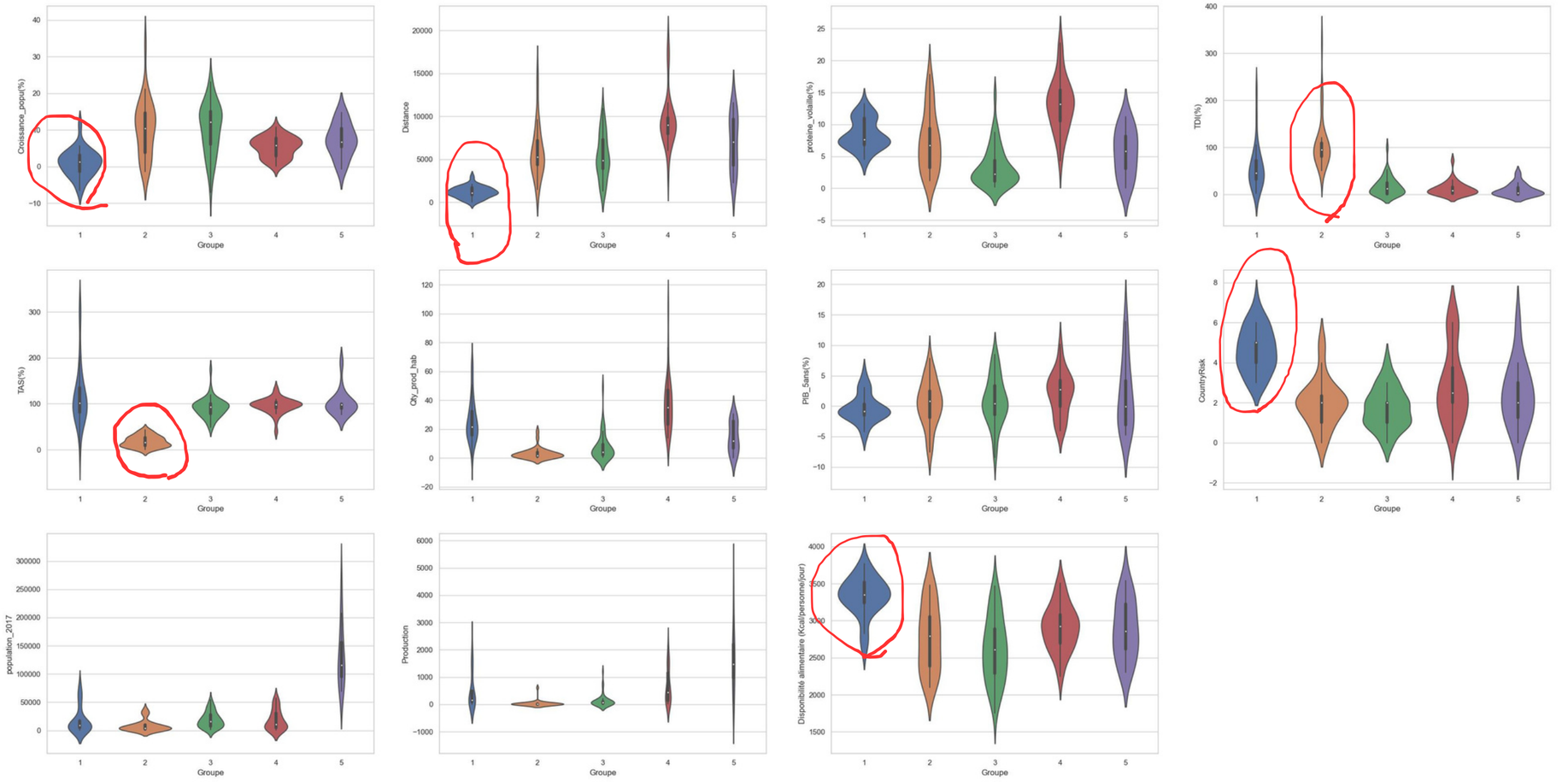


- Calcule des distance euclidienne
- Méthod de Ward



LA POULE QUI CHANTE

Quelles sont les groupes interessant ?



Groupe 1

TDI (%) : 56.48

TAS (%) : 115.15

CountryRisk : 4.62

Distance à la France : 1192.11

Croissance démographique (%) : 0.97

Dispo. alimentaire : 3331.41

Pib (%) : -0.38

Groupe 2

TDI (%) : 109.25

TAS (%) : 17.97

CountryRisk : 2.0

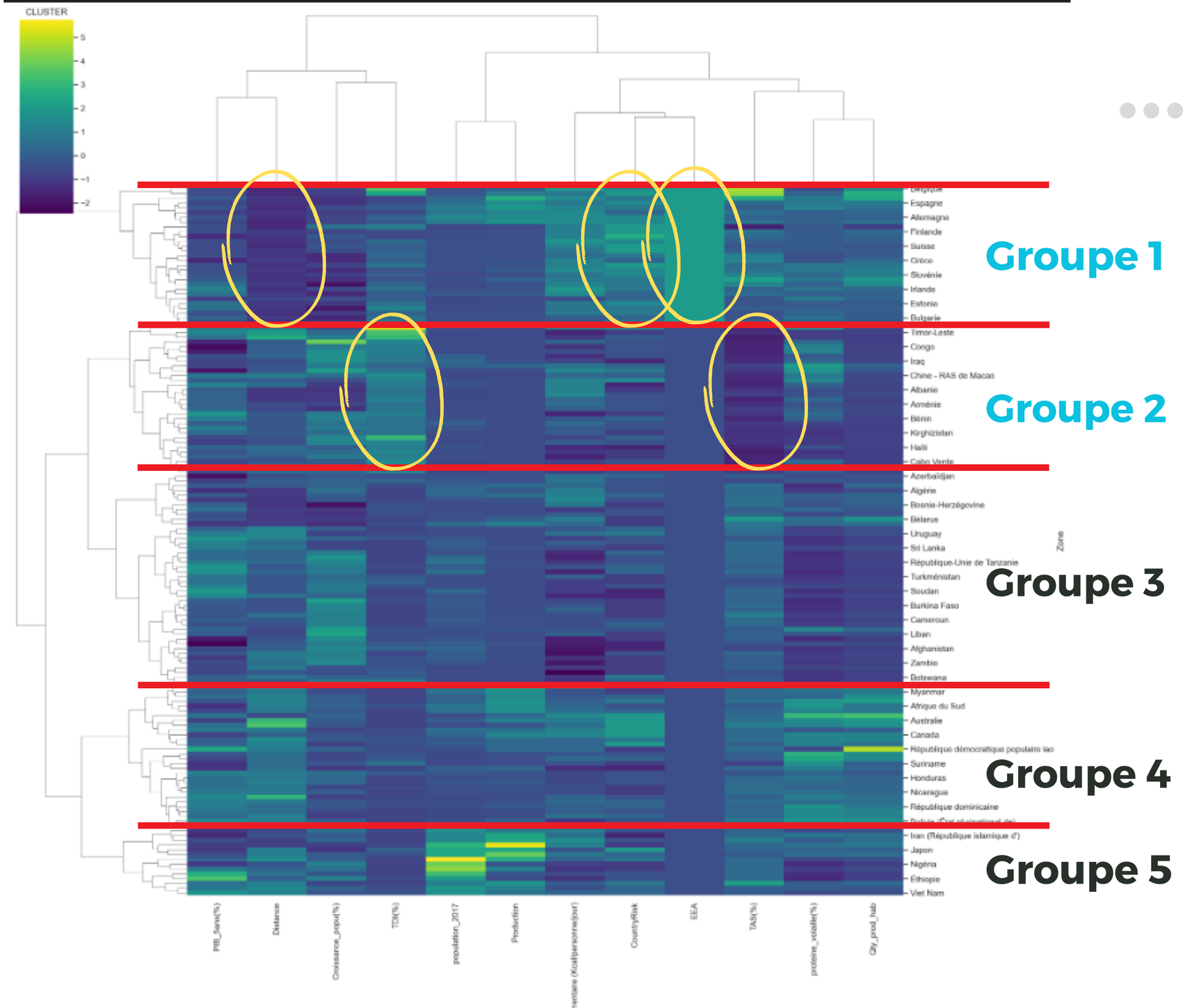
Distance à la France : 5876.74

Croissance démographique (%) : 10.46

Dispo. alimentaire : 2758.58

Pib (%) : 0.42

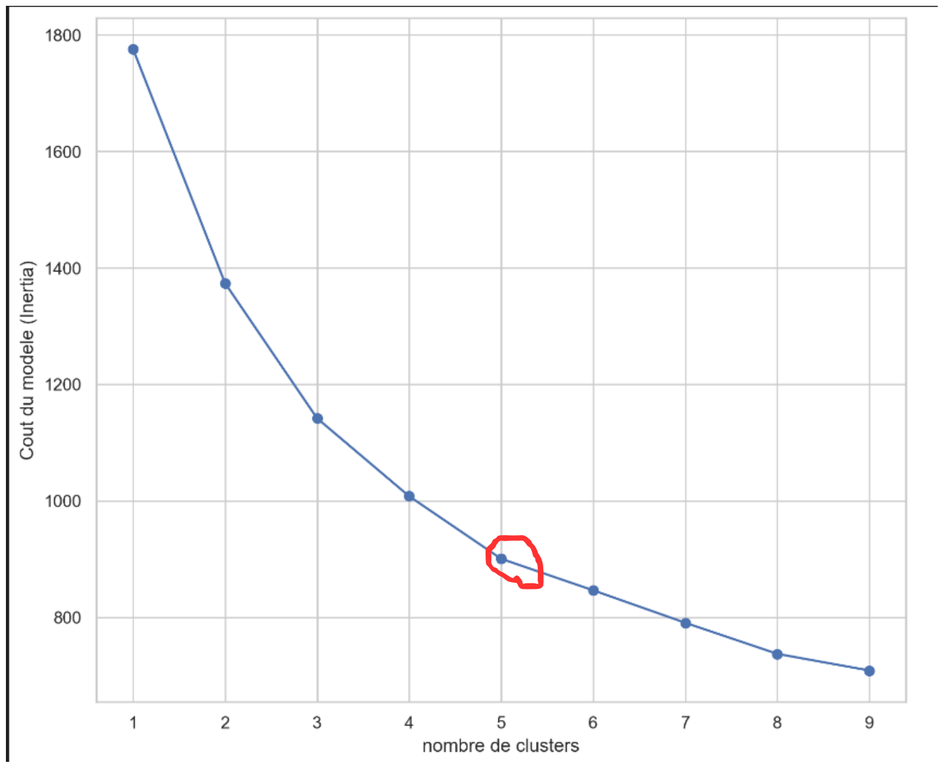
Heatmap Cluster



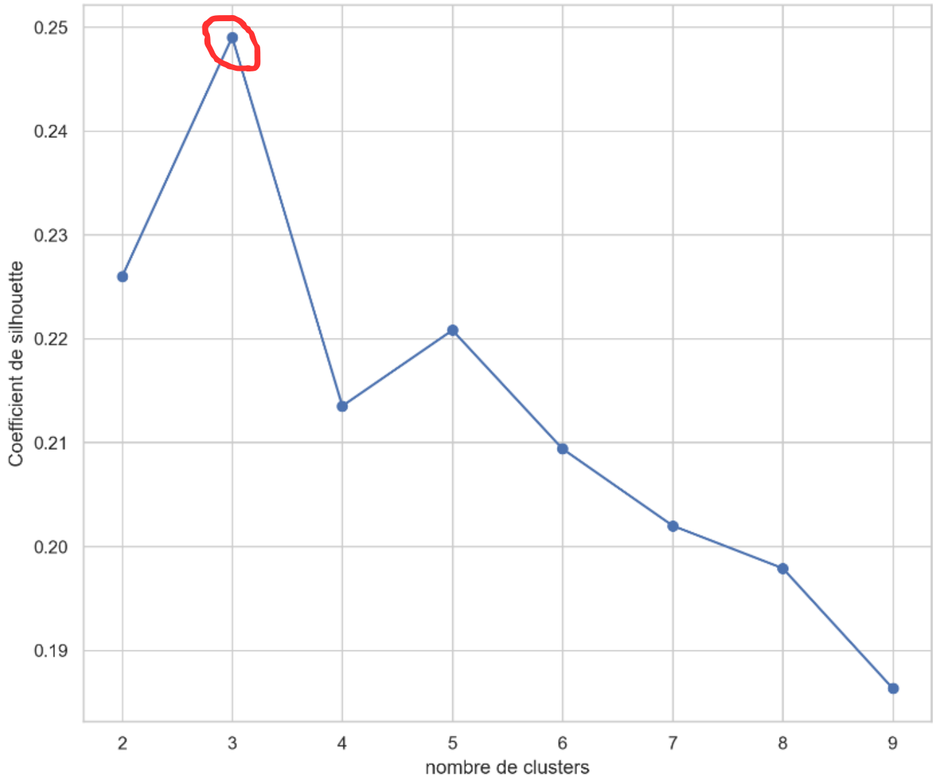
LA POULE QUI CHANTE

K-means (non déterministe)

Méthode du Coude



Silhouette Coefficient



- 3 clusters
- 5 clusters

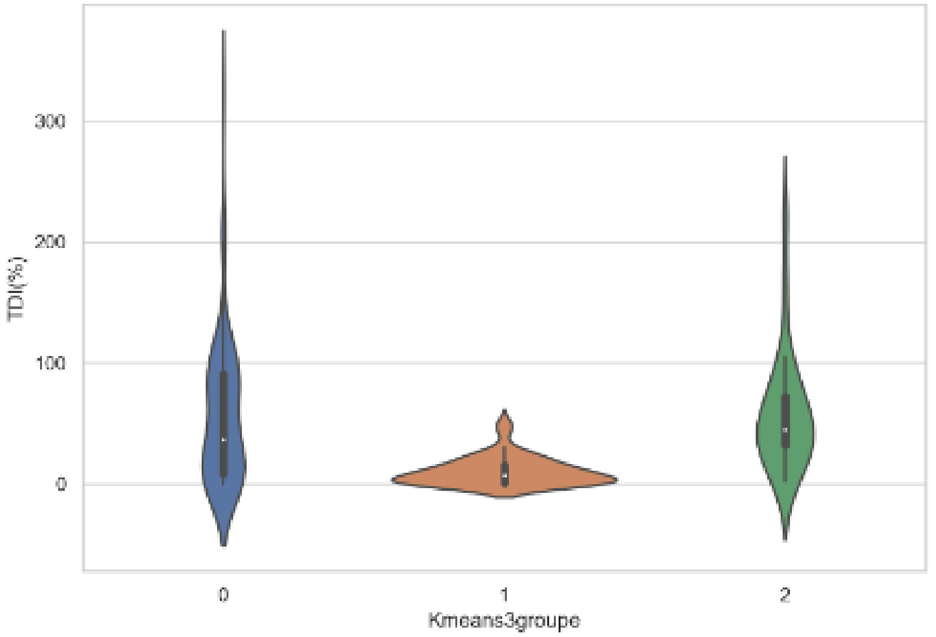
LA POULE QUI CHANTE

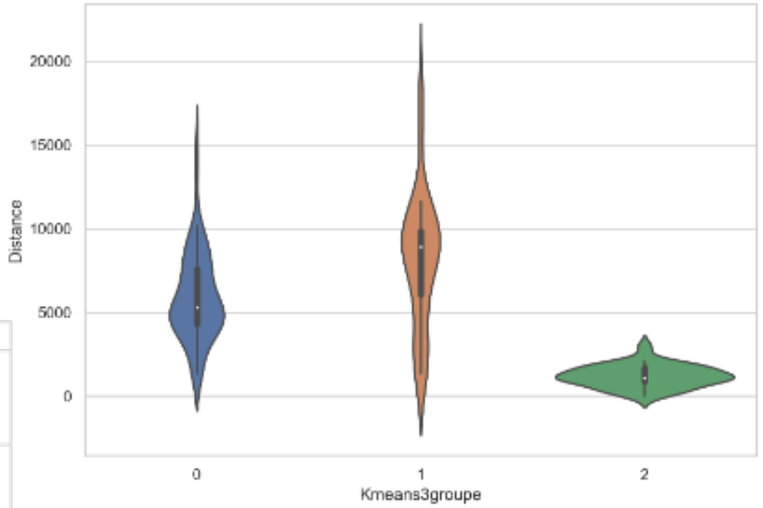
K-means - centroîde 3 Clusters

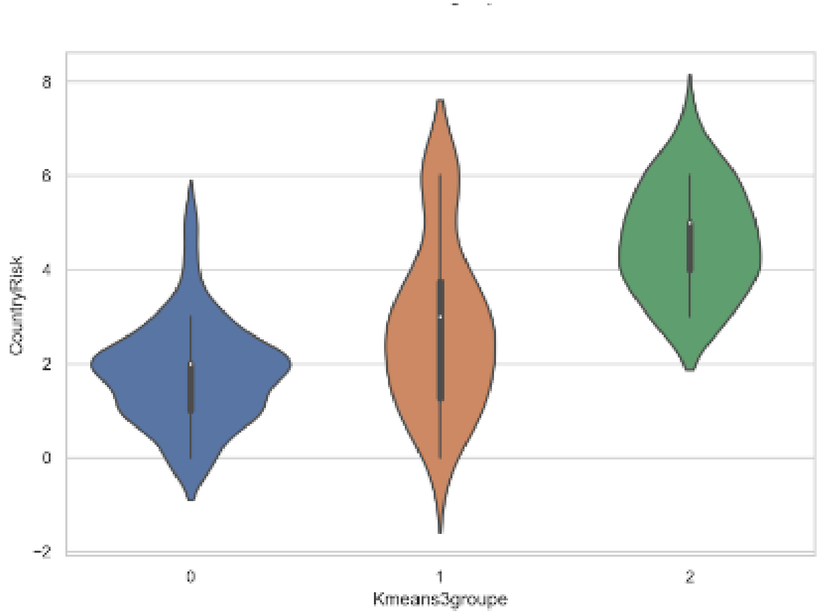
Groupe 1 : 73 pays- **Sierra Leone**

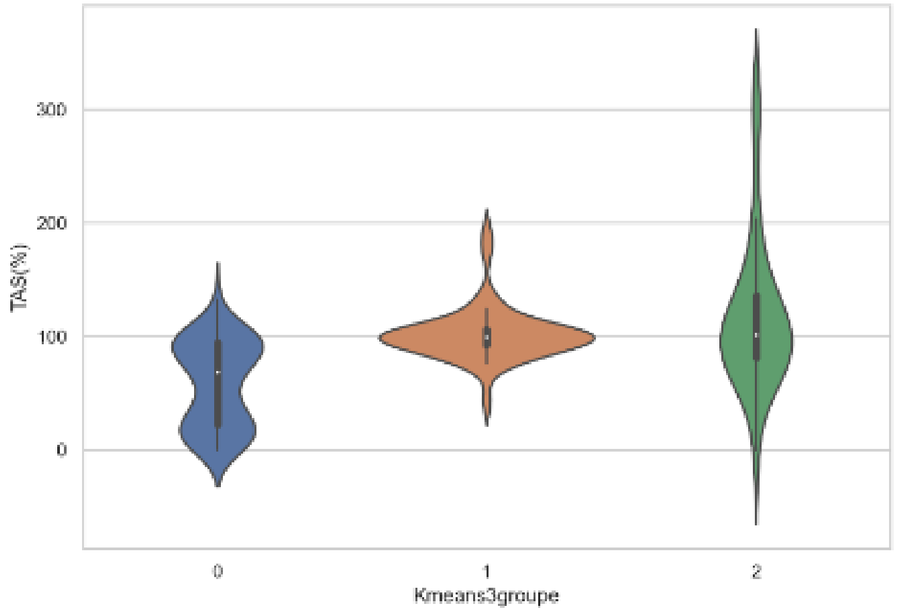
Groupe 2 : 46 pays-**Colombie**

Groupe 3 : 29 pays-**Slovénie**









K-means - centroîde 5 Clusters

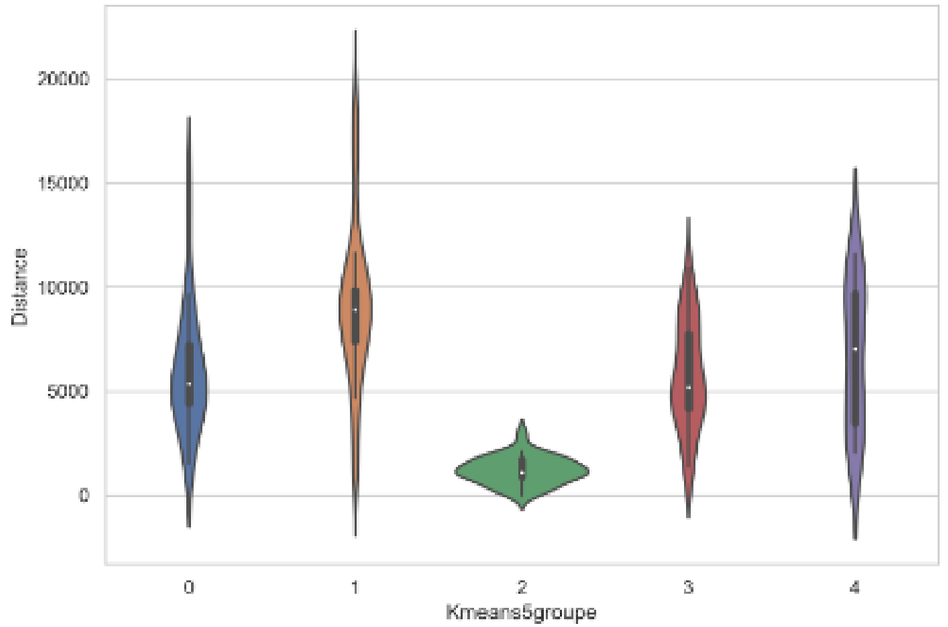
Groupe 1 : 32 pays -**Namibie**

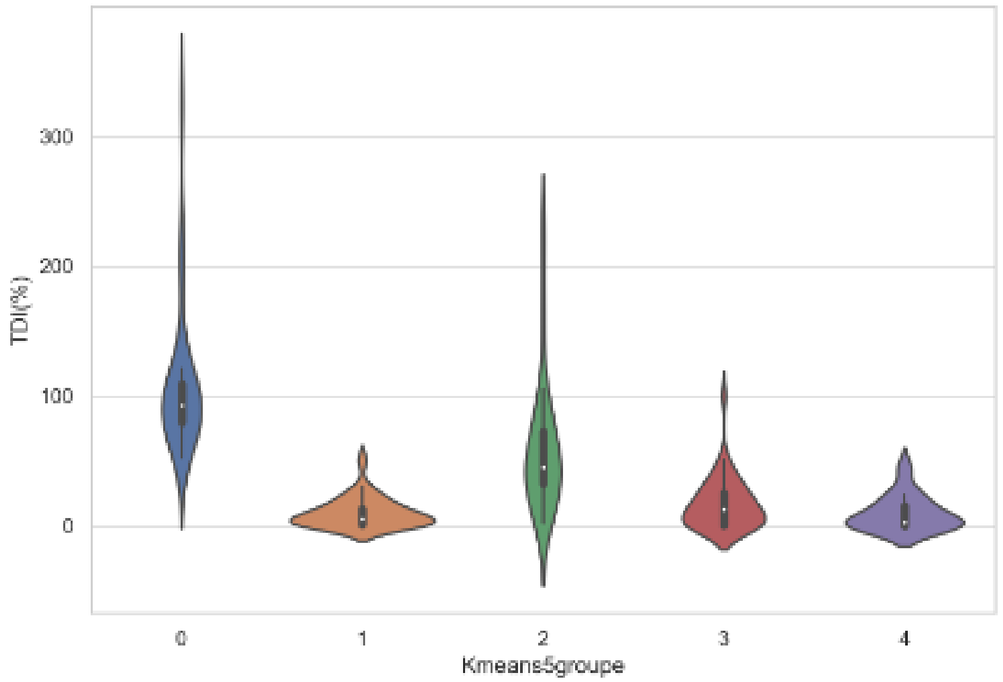
Groupe 2 : 30 pays -**Chili**

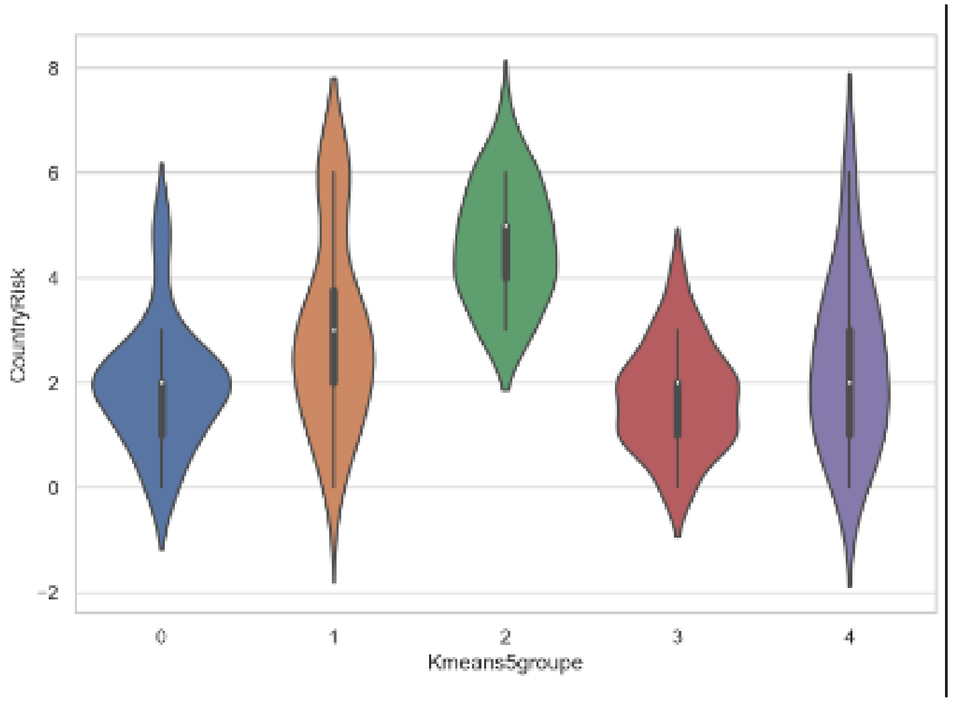
Groupe 3 : 29 pays-**Slovénie**

Groupe 4 :43 pays -**Cameroun**

Groupe 5 : 14 pays-**Philippines**

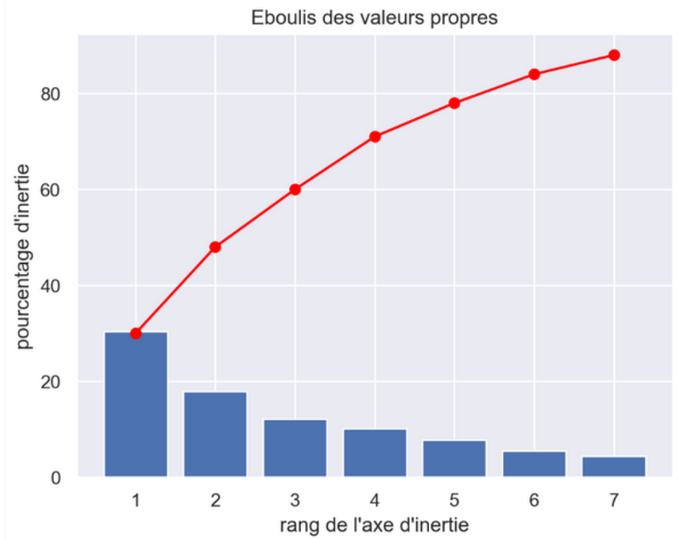






Analyse en composantes principales (ACP)

12 variables quantitatives




```
scree = (pca.explained_variance_ratio_*100).round(2)
scree
```

✓ 0.0s

```
array([30.36, 17.87, 12.2 , 10.14,  7.8 ,  5.59,  4.43])
```

```
n_components = scree.size
scree_cum = scree.cumsum().round()
scree_cum
```

✓ 0.0s

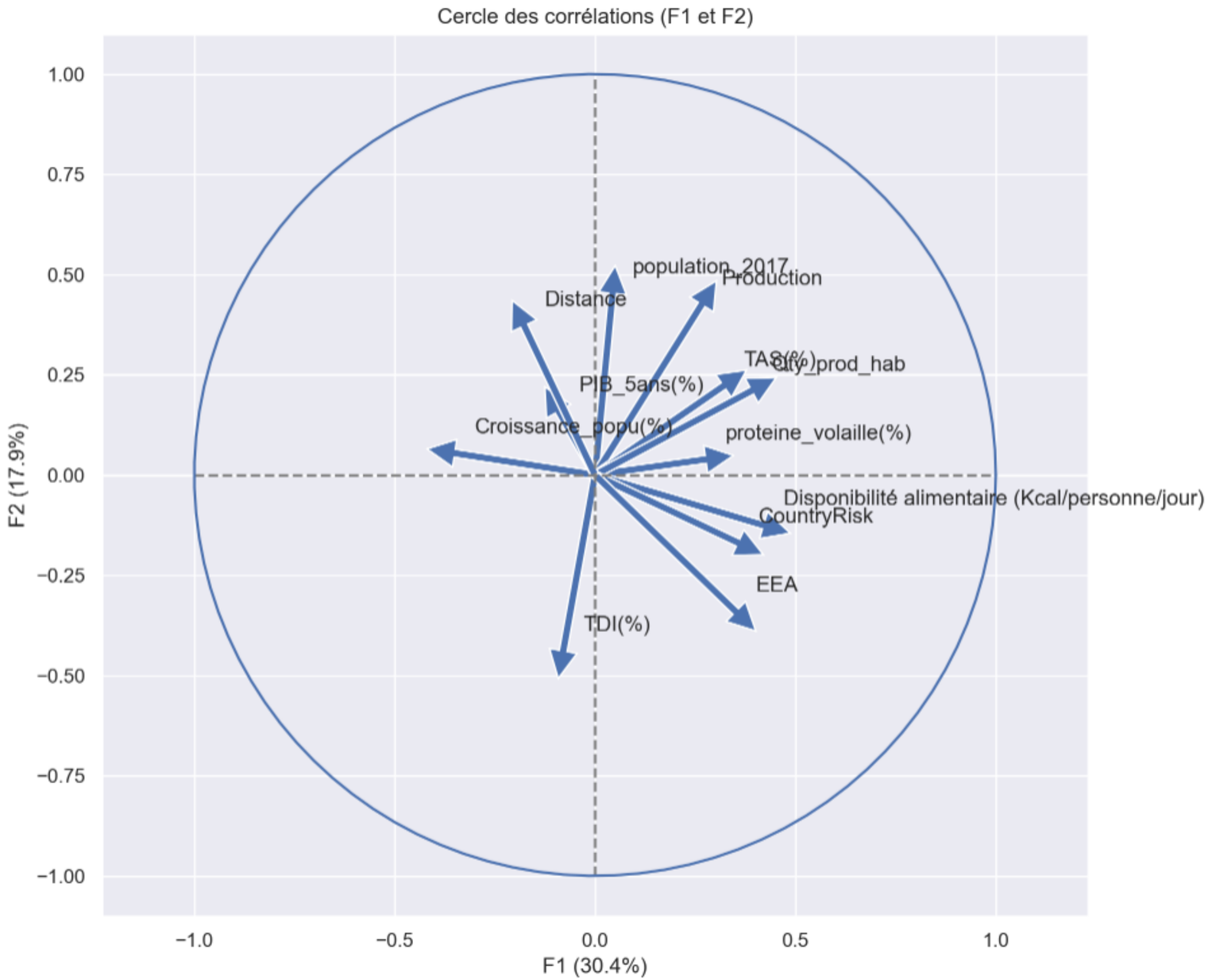
```
array([30., 48., 60., 71., 78., 84., 88.])
```

Corrélations des variables avec les composantes principales

	F1	F2	F3	F4	F5	F6	F7
Croissance_popu(%)	-0.35	0.05	-0.08	0.22	-0.15	-0.65	-0.16
population_2017	0.04	0.45	-0.42	0.18	0.43	-0.00	-0.05
proteine_volaille(%)	0.28	0.04	0.52	0.37	-0.12	0.11	-0.18
Production	0.27	0.43	-0.17	0.35	0.20	0.13	0.17
TDI(%)	-0.08	-0.44	0.16	0.22	0.48	0.02	0.62
TAS(%)	0.32	0.22	-0.17	-0.39	-0.26	-0.27	0.47
Disponibilité alimentaire (Kcal/personne/jour)	0.42	-0.13	-0.01	0.10	0.14	0.11	-0.35
Qty_prod_hab	0.39	0.21	0.32	-0.07	-0.24	-0.05	0.18
PIB_5ans(%)	-0.09	0.16	0.24	-0.65	0.45	0.12	-0.25
CountryRisk	0.36	-0.17	0.08	-0.00	0.33	-0.62	-0.23
EEA	0.35	-0.34	-0.24	-0.16	0.08	-0.11	0.05
Distance	-0.18	0.37	0.49	0.01	0.22	-0.25	0.19

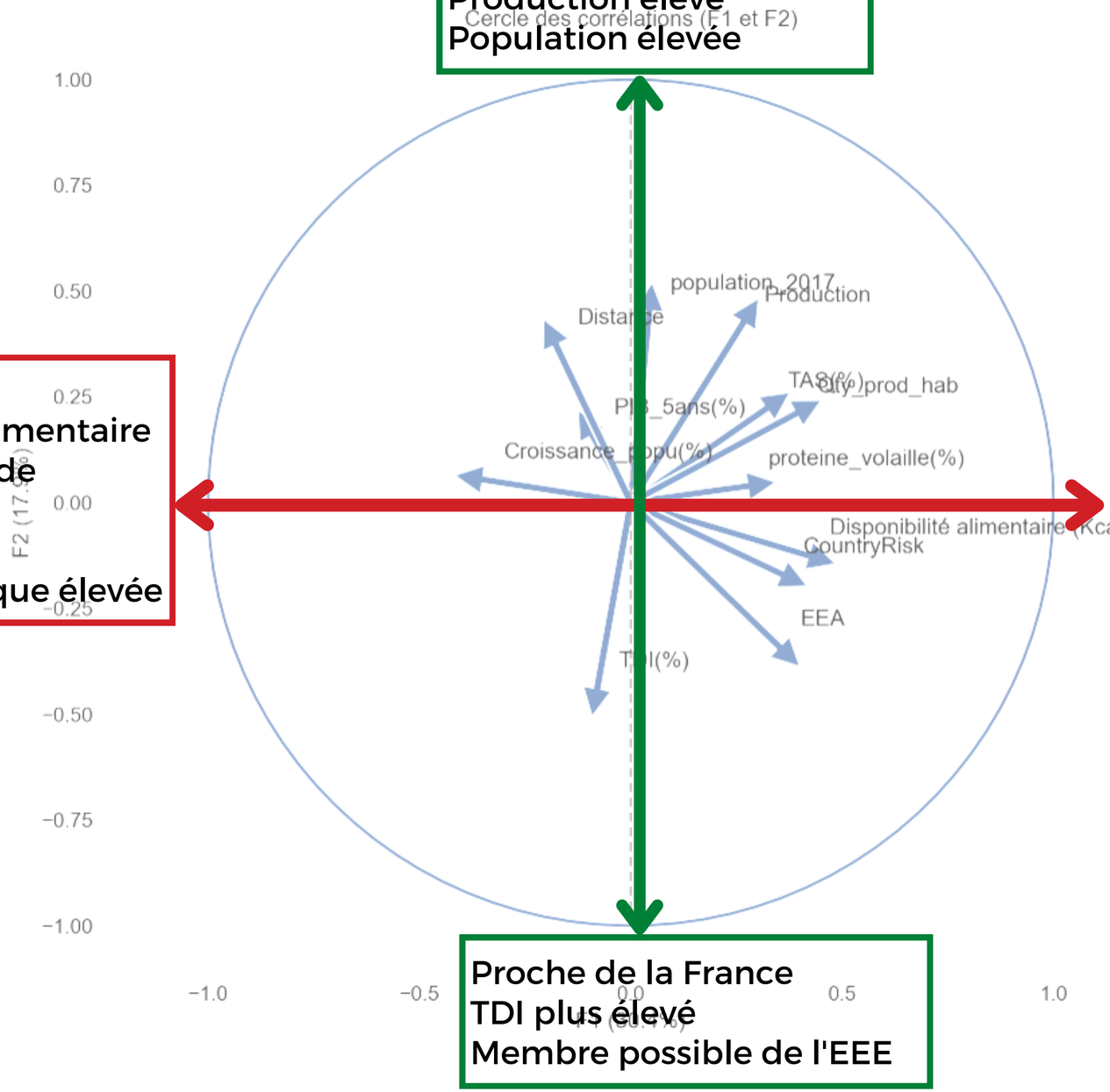
LA POULE QUI CHANTE

Cercles des corrélation



Cercles des corrélation

Eloigné de la France
TDI plus faible
Production élevé
Population élevée



Pauvre
Faible dispo.alimentaire
Mauvais score de CountryRisk
Croissance Démographique élevée

Riche
Forte Dispo. alimentaire
Qty.prod. hab
Production
Bon score de Country Risk
TAS élevé
Croissance démographique plus faible

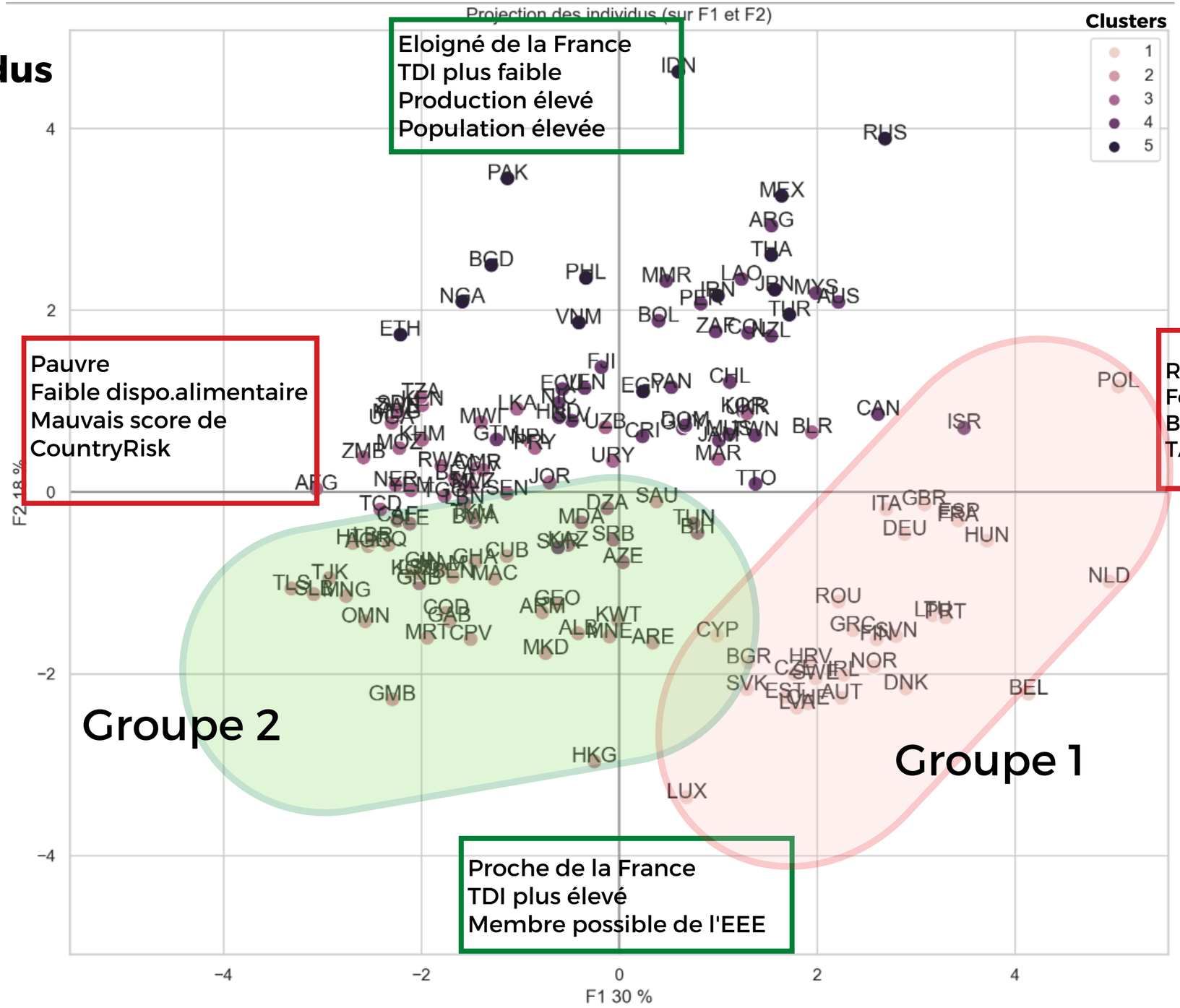
Proche de la France
TDI plus élevé
Membre possible de l'EEE

LA POULE QUI CHANTE

LA POULE QUI CHANTE

Projection des individus

CAH de 5 clusters



Eloigné de la France
TDI plus faible
Production élevé
Population élevée

Pauvre
Faible dispo.alimentaire
Mauvais score de CountryRisk

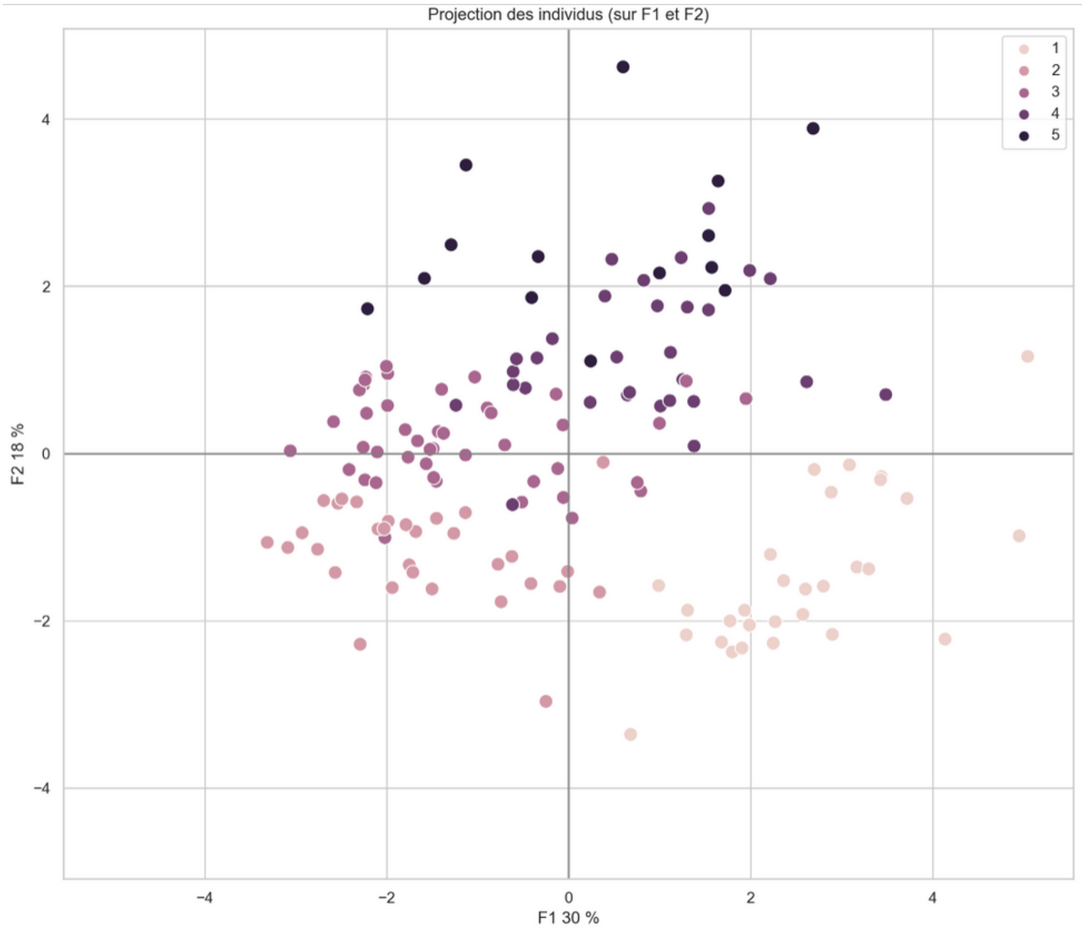
Riche
Forte Dispo. alimentaire
Bon score de Country Risk
TAS élevé

Proche de la France
TDI plus élevé
Membre possible de l'EEE

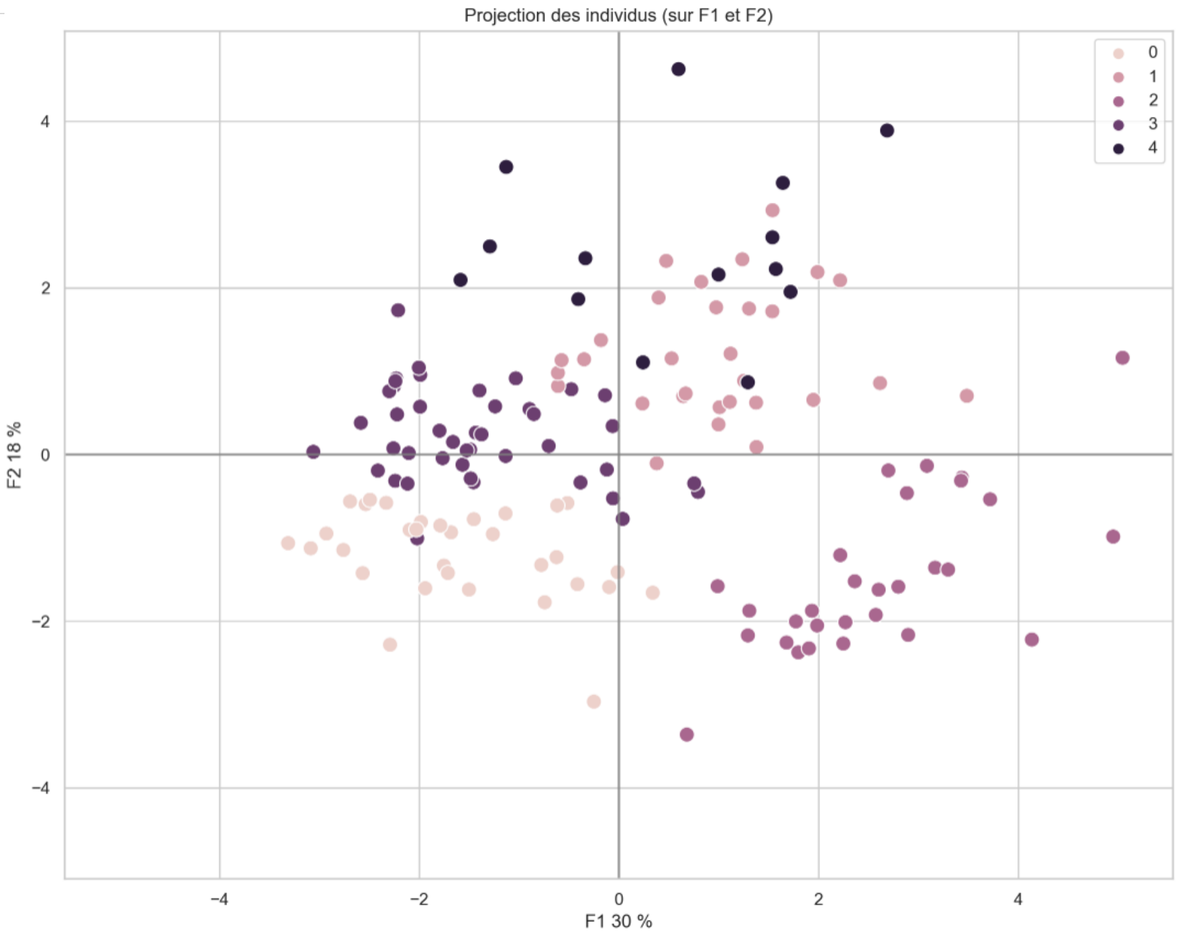
CAH v.s K-MEANS

LA POULE QUI CHANTE

Projection sur plan factorielle avec regroupement CAH de 5 clusters



Projection sur plan factorielle avec regroupement k-means de 5 clusters

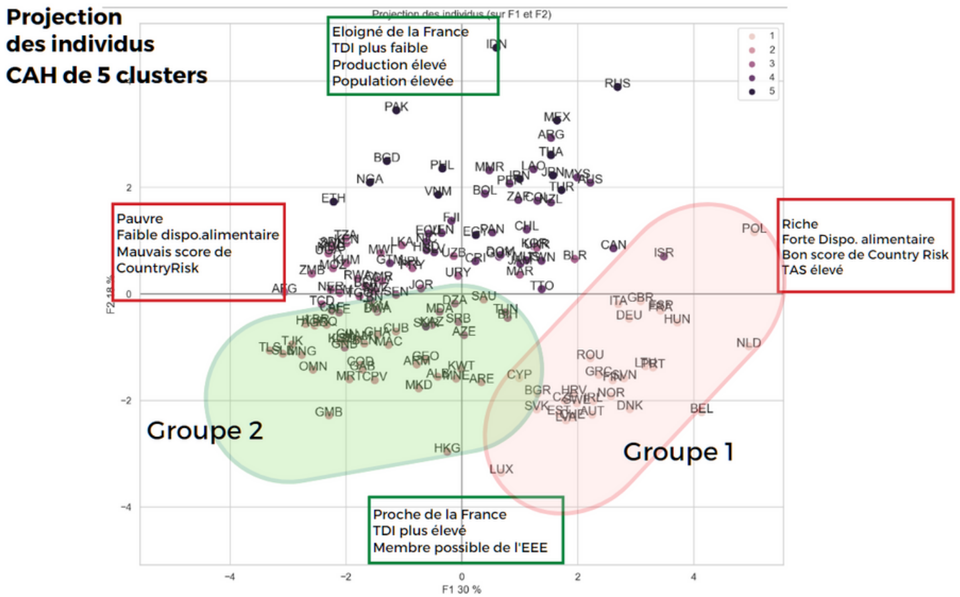


LA POULE QUI CHANTE

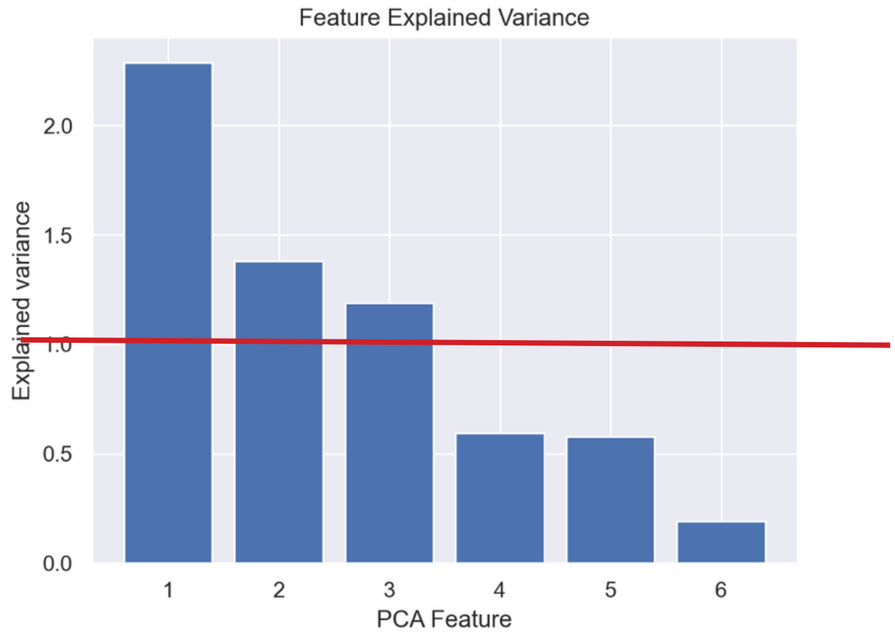
Exploration des clusters selectionnées

- Groupe 1
- Groupe 2

Pour l'affinage de la selection, on garde les clusters fait avec CAH.



- **6 variables pour analyse**
  - **TDI**
  - **TAS**
  - **Dispo. alimentaire**
  - **CountryRisk**
  - **Distance**
  - **Production**



Affinage de la selection sur le cluster 1

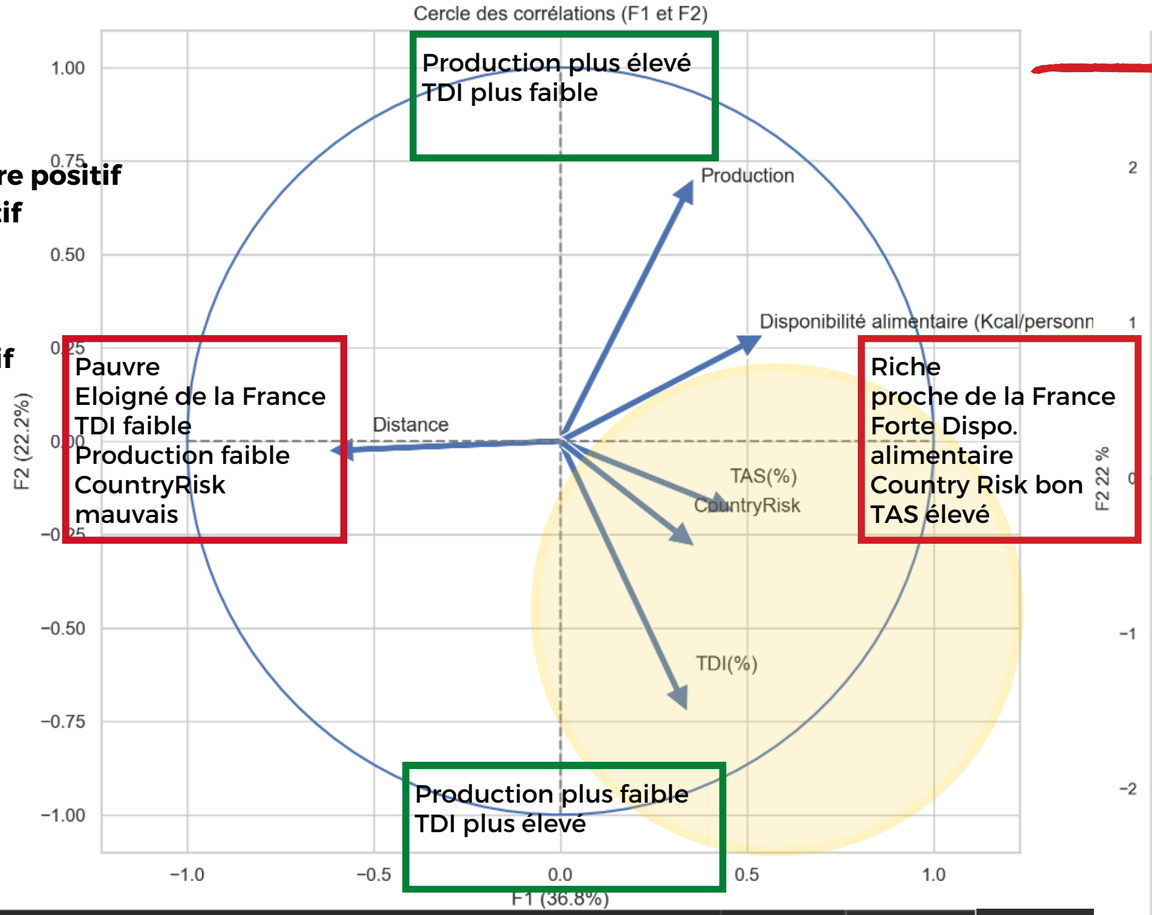
TDI ↗
TAS ↗
dispo. alimentaire ↗
proche de la France

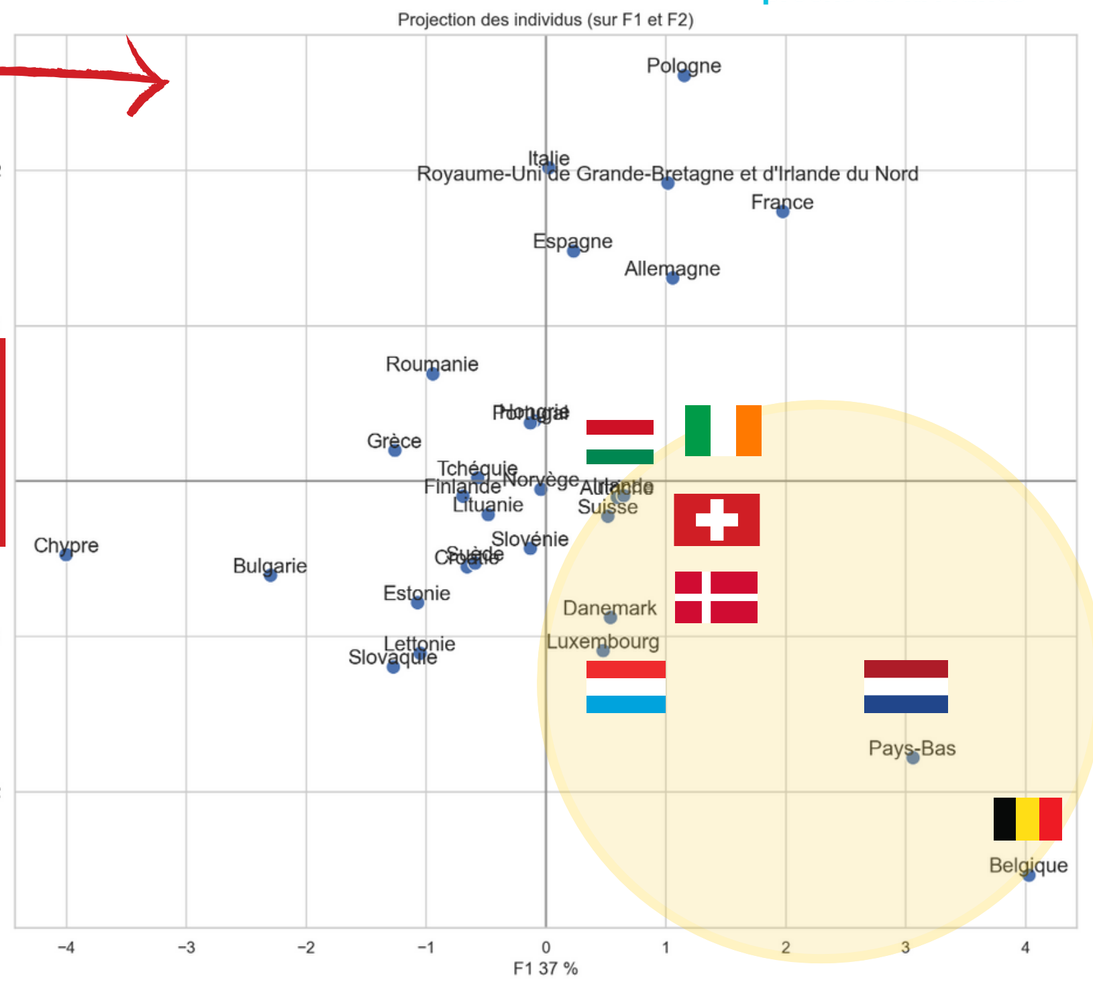
- **F1**

- **TDI positif**
- **TAS positif**
- **Dispo. alimentaire positif**
- **Production positif**
- **Distance négatif**

- **F2**

- **Production positif**
- **TDI négatif**
- **TAS négatif**





Zone	Croissance_popu(%)	population_2017	proteine_volaille(%)	Production	TDI(%)	TAS(%)	PIB_5ans(%)	CountryRisk	Distance
Autriche	3.736362	8819.901	6.928129	148.0	63.583815	85.549133	-0.995045	5.0	1036.711400
Belgique	3.016502	11419.748	4.509127	463.0	222.368421	304.605263	-0.994720	5.0	264.600808
Danemark	2.163128	5732.274	11.486008	173.0	79.640719	103.592814	-0.748765	6.0	1028.875241
France	2.011018	64842.509	7.985368	1750.0	32.167832	111.252384	-1.781165	5.0	0.000000
Irlande	3.148324	4753.279	10.963047	110.0	77.343750	85.937500	5.409817	4.0	780.840481
Luxembourg	11.500837	591.910	6.796484	0.0	100.000000	0.000000	-0.836909	6.0	287.972186
Pays-Bas	1.366777	17021.347	8.120272	1100.0	163.440860	295.698925	-1.337900	6.0	430.886817
Suisse	5.591800	8455.804	6.798659	91.0	38.345865	68.421053	-1.470107	6.0	490.305157

Les moyennes du cluster 1

TDI(%)	TAS(%)	Disponibilité alimentaire (Kcal/personne/jour)	Qty_prod_hab	PIB_5ans(%)	CountryRisk
56.483878	115.152440	3331.413793	25.838983	-0.384796	4.620690

Croissance_popu(%)	population_2017	proteine_volaille(%)	Production
0.972926	18068.880034	8.465894	513.551724

Affinage de la selection sur le cluster 2

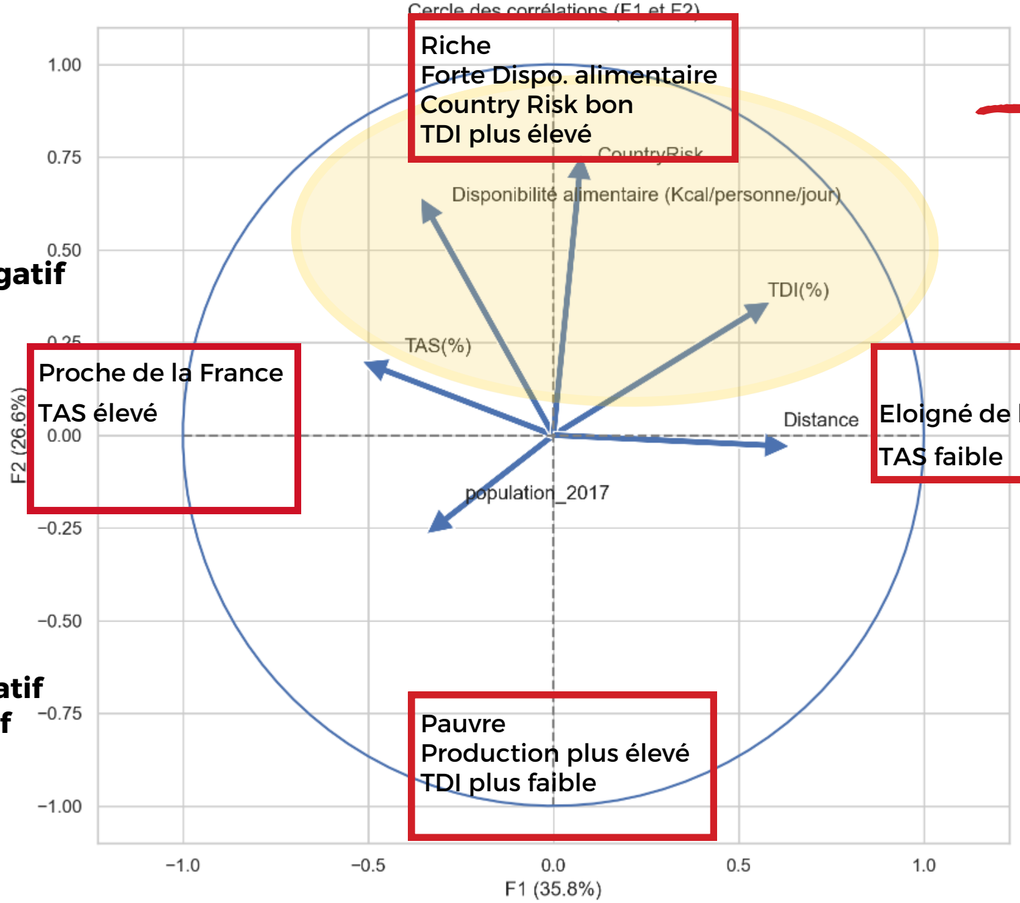
TDI ↗
TAS ↘
CountryRisk ↗

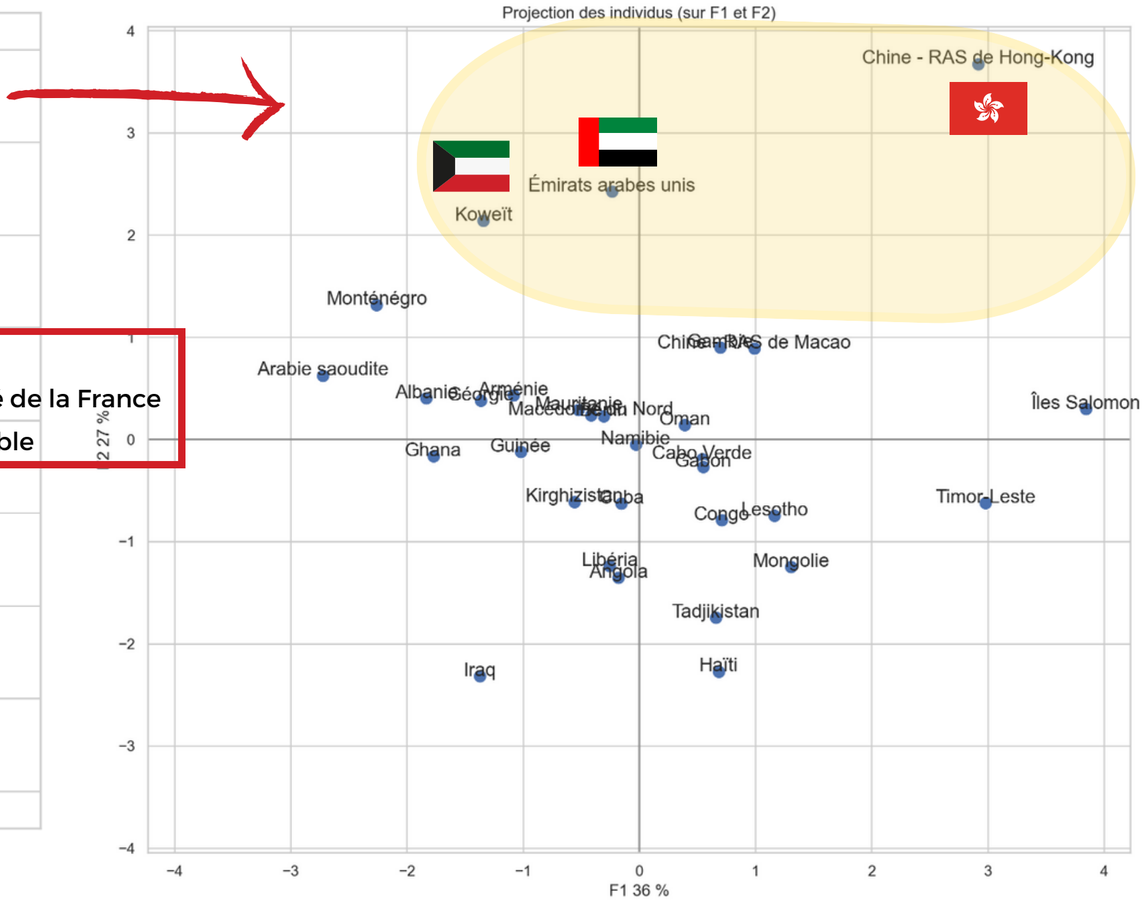
- **F1**

- TDI positif
- Distance positif
- TAS négatif
- Dispo. alimentaire négatif

- **F2**

- CountryRisk positif
- Dispo. alimentaire négatif
- Population 2017 négatif





Zone	Croissance_popu(%)	population_2017	proteine_volaille(%)	Production	TDI(%)	TAS(%)	Qty_prod_hab	PIB_5ans(%)	CountryRisk	Distance
RAS de Hong-Kong	3.682129	7306.322	16.144473	24.0	323.928571	8.571429	3.284826	4.697363	5.0	9645.855848
Koweït	21.119052	4056.099	15.919350	56.0	72.486772	29.629630	13.806369	-6.816247	4.0	4441.333447
Émirats arabes unis	3.780598	9487.203	13.798247	48.0	105.097087	11.650485	5.059447	0.798578	5.0	5251.803242
France	2.011018	64842.509	7.985368	1750.0	32.167832	111.252384	26.988468	-1.781165	5.0	

Les moyennes du cluster 2

Croissance_popu(%)	population_2017	proteine_volaille(%)	Production	TDI(%)	TAS(%)	Disponibilité alimentaire (Kcal/personne/jour)	Qty_prod_hab	PIB_5ans(%)	CountryRisk	Distance
10.465579	8300.350065	7.395255	36.451613	109.254367	17.976349	2758.580645	3.172838	0.428848	2.000620	5876.742948

LA POULE QUI CHANTE

Conclusion

- Sélection de 2 clusters
  - Groupe 1
    - Belgique
    - Pays-Bas
    - Luxembourg
    - Suisse
    - Danemark
  - Groupe 2
    - Hong-Kong
    - Émirats arabes unis
    - Koweït

MERCI

END OF SLIDE