La poule qui chante

se développer à l'international

une étude de marché avec Python

Ebru Mandrou

table des matières

data preparation

Choix des données : conformément à l'analyse PESTEL (pour les années entre 2016 -2018)

- **Political:**
  - Taux de stabilité politique
- **Economic:**
  - Taux de change
  - Coût de production des producteurs
  - Taux du PIB
  - Taux d'investissement direct étranger
- **Social:**
  - Comportement de consommation du Pays :
    - Disponibilité alimentaire (Kcal/personne/jour)
    - Disponibilité de protéines en quantité (g/personne/jour)
    - Consommation de protéines de base animale (calculée)
    - Production
    - Importation
  - Taux de croissance démographique
- **Legal:**
  - Taux d'importation de produits depuis la France

data preparation

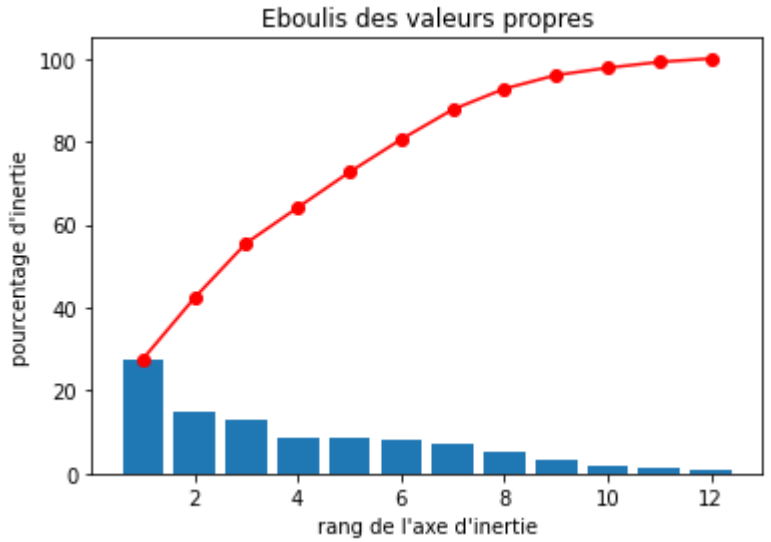
Nettoyage des données:

- Traitement des valeurs nulles
- Détection des valeurs aberrantes

principal component analysis (PCA)

- **Éboulis des valeurs propres**

  - Pour décider du nombre de caractéristiques que nous aimerions conserver en fonction du diagramme de variance cumulée.

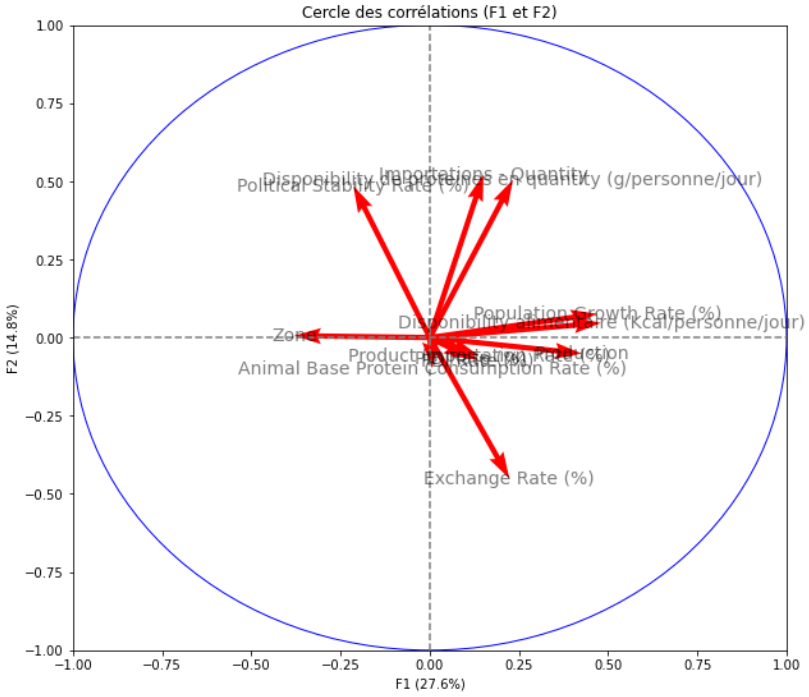


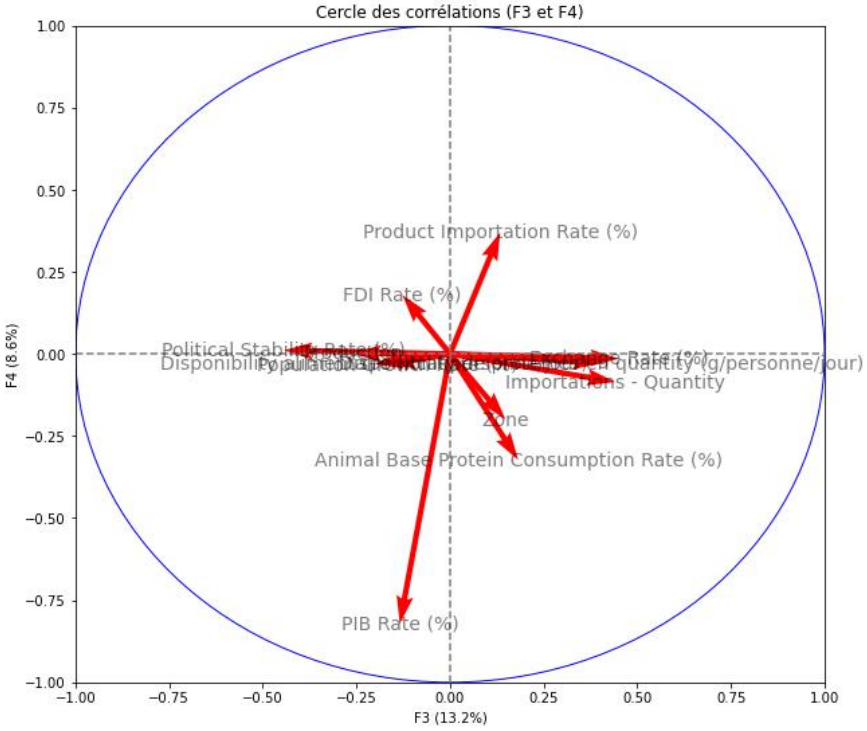
  - à conserver environ 80 % de la variance
    - 6 composants

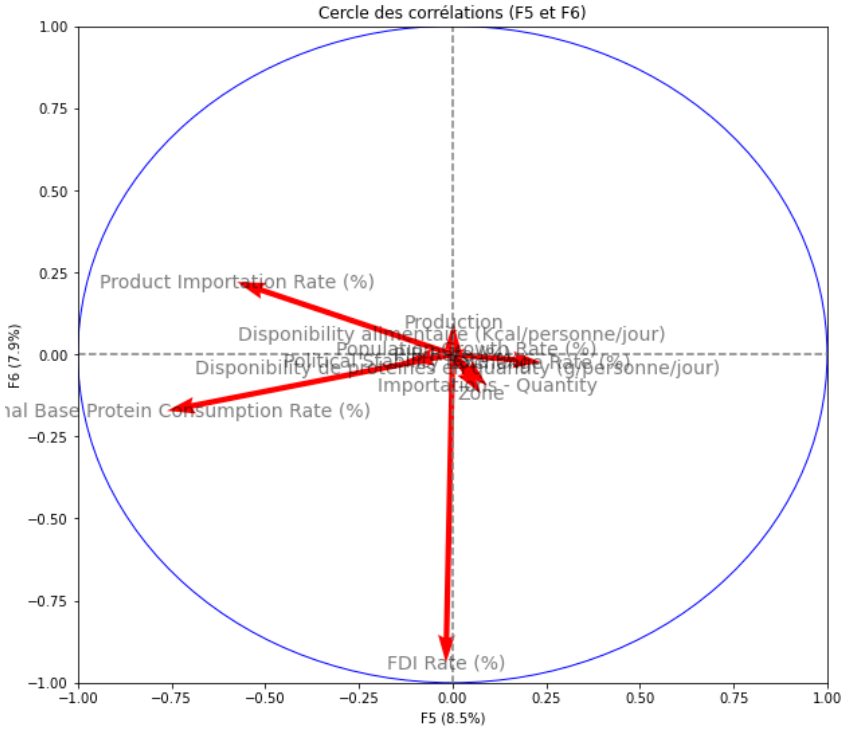
```
[0.27582531 0.14753026 0.13176628 0.08598178 0.08522806 0.07895971]
0.8052913991068827
```

principal component analysis (PCA)

Cercle des correlations







principal component analysis (PCA)

Correspondance des axes de synthèse:

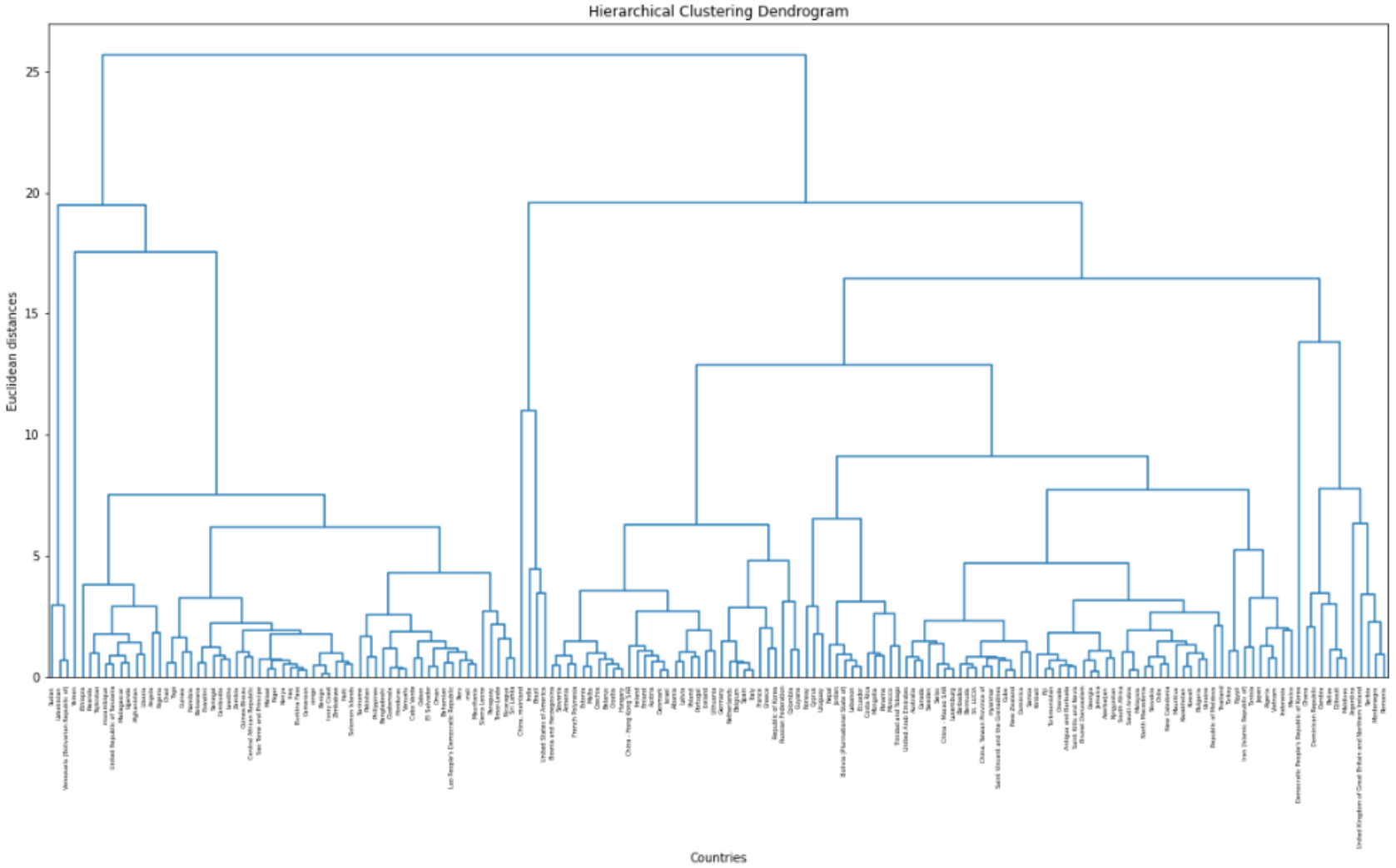
	Positive Correlation	Negative Correlation
F1: 27.6%	• Population Growth Rate • Disponibility_alimentaire (kcal/p/j) • Production	
F2: 14.8%	• Importation • Protein consumption (g/p/j) • Political Stability	• Exchange Rate
F3: 13.2%	• Exchange Rate • Protein consumption (g/p/j) • Importation	• Political Stability • Disponibility_alimentaire (kcal/p/j)
F4: 8.6%	• Product Importation from France	• PIB Rate
F5: 8.5%		• Product Importation Rate • Animal Base Consumption Rate
F6: 7.9%		• FDI Rate

principal component analysis (PCA)

Création d'un final dataset avec les scores des composants

Zone	F1	F2	F3	F4	F5	F6
Afghanistan	-2.968891	0.262997	0.338730	-0.170927	-0.486712	-0.478188
South Africa	0.550650	-0.791201	0.867292	-0.008651	-0.507840	-0.053089
Albania	2.743652	-1.322731	-0.173795	0.288962	-0.232491	0.215809
Algeria	0.089149	0.652312	-0.138205	-0.211687	0.097001	-0.367632
Germany	3.024856	1.397656	0.844625	-0.203860	0.044475	-0.022076
...	...	...	...	...	...	...
United Arab Emirates	1.211660	-0.078635	-0.162797	-0.201452	-0.142955	-0.135466
Ecuador	-1.258662	-0.082509	-0.371558	-1.092172	1.270242	-0.640393
United States of America	4.130186	4.153897	2.340002	-0.194177	0.048398	-0.329269
Ethiopia	-3.262282	0.816696	-0.302317	-1.461935	1.934545	-1.508563
Solomon Islands	-2.027609	-0.496434	0.510167	0.123591	0.251773	-0.525118

ascending hierarchical clustering (CAH)



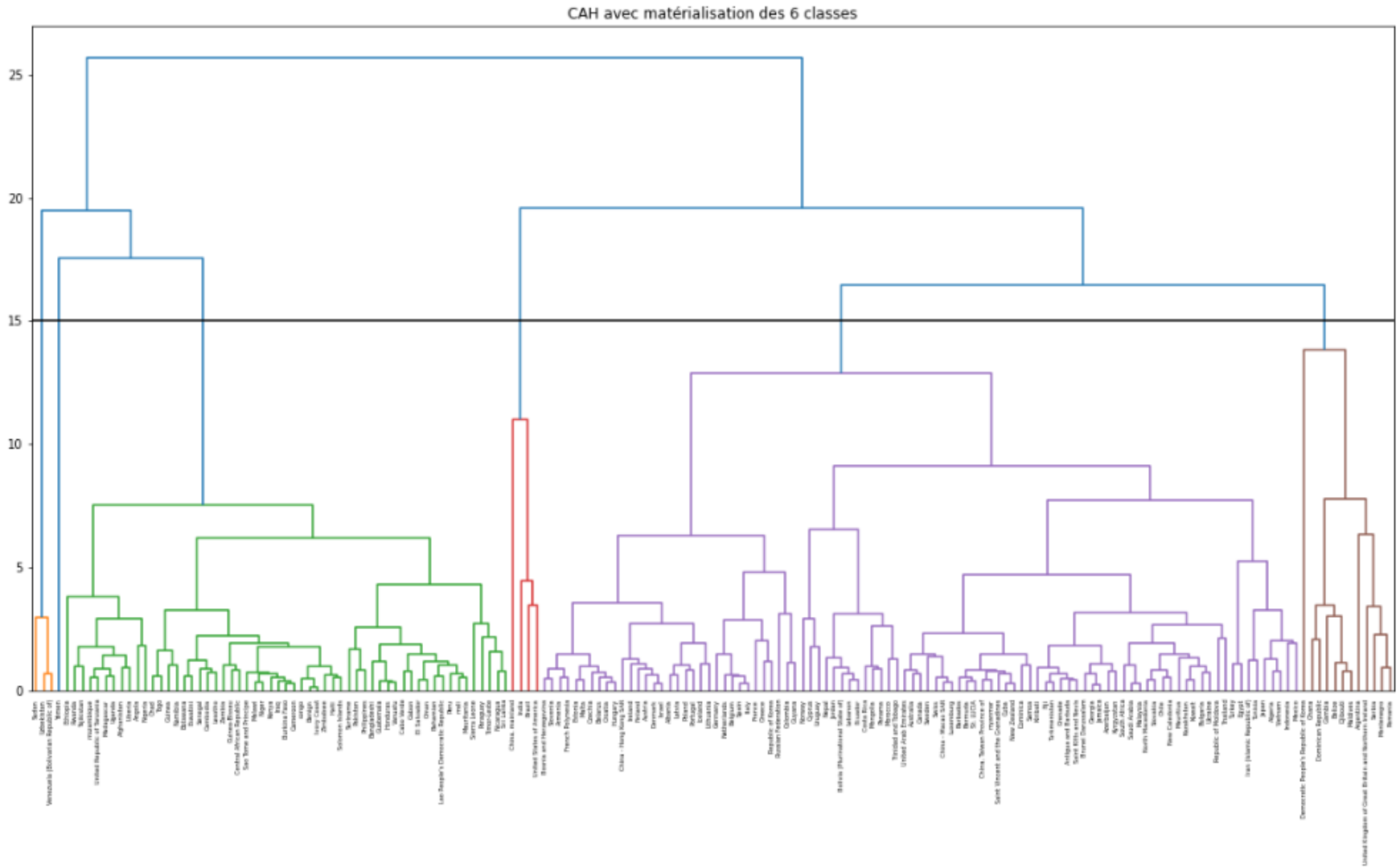
ascending hierarchical clustering (CAH)

La hauteur à laquelle deux objets sont joints

h = 15

Selon le dendrogramme

6 clusters



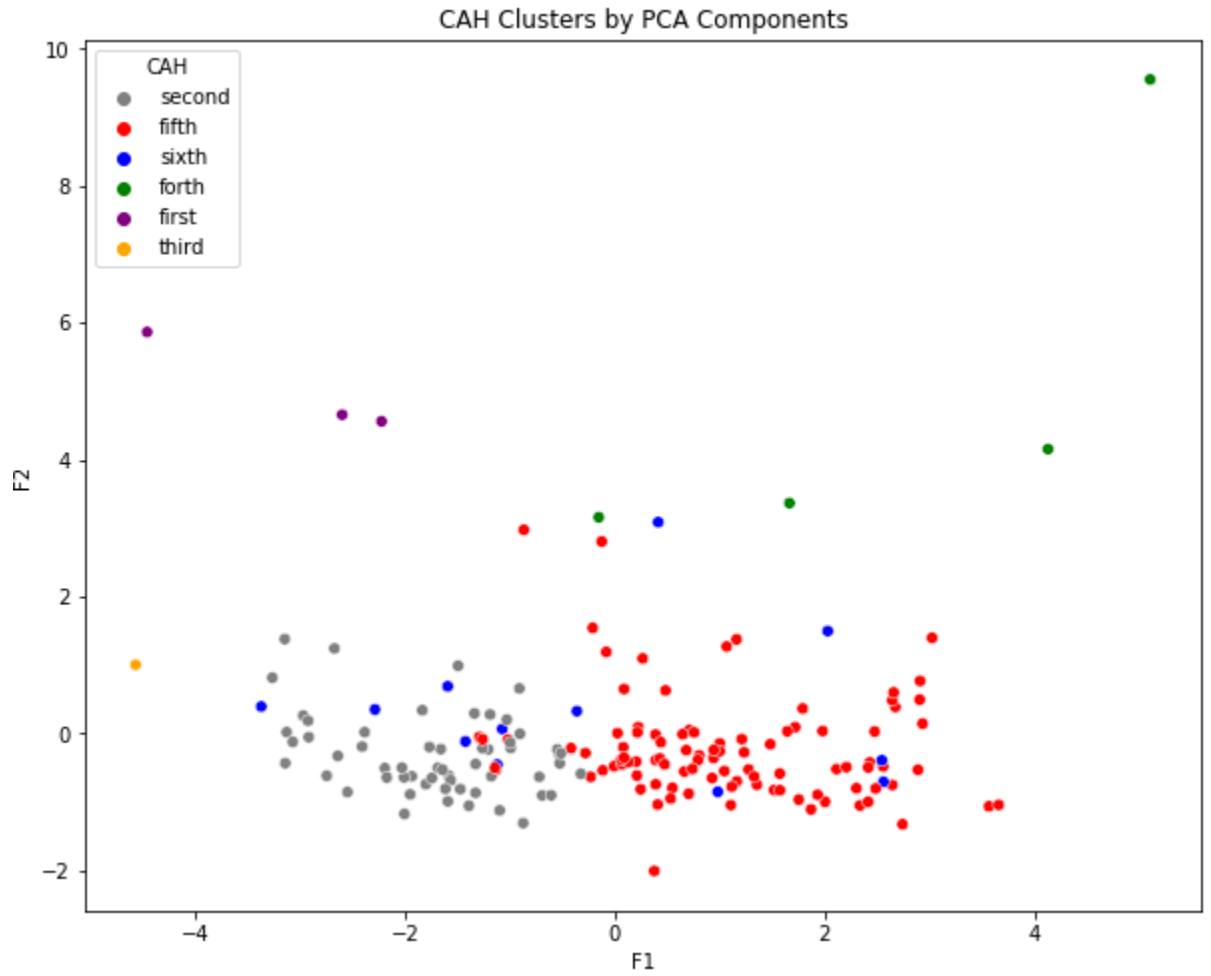
analyse des l'algorithme CAH

Création d'un nouvel dataset avec les composants PCA et leurs CAH_clusters correspondantes

Zone	F1	F2	F3	F4	F5	F6	CAH
Afghanistan	-2.968891	0.262997	0.338730	-0.170927	-0.486712	-0.478188	second
South Africa	0.550650	-0.791201	0.867292	-0.008651	-0.507840	-0.053089	fifth
Albania	2.743652	-1.322731	-0.173795	0.288962	-0.232491	0.215809	fifth
Algeria	0.089149	0.652312	-0.138205	-0.211687	0.097001	-0.367632	fifth
Germany	3.024856	1.397656	0.844625	-0.203860	0.044475	-0.022076	fifth
...	...	...	...	...	...	...	...
United Arab Emirates	1.211660	-0.078635	-0.162797	-0.201452	-0.142955	-0.135466	fifth
Ecuador	-1.258662	-0.082509	-0.371558	-1.092172	1.270242	-0.640393	fifth
United States of America	4.130186	4.153897	2.340002	-0.194177	0.048398	-0.329269	forth
Ethiopia	-3.262282	0.816696	-0.302317	-1.461935	1.934545	-1.508563	second
Solomon Islands	-2.027609	-0.496434	0.510167	0.123591	0.251773	-0.525118	second

analyse des l'algorithme CAH

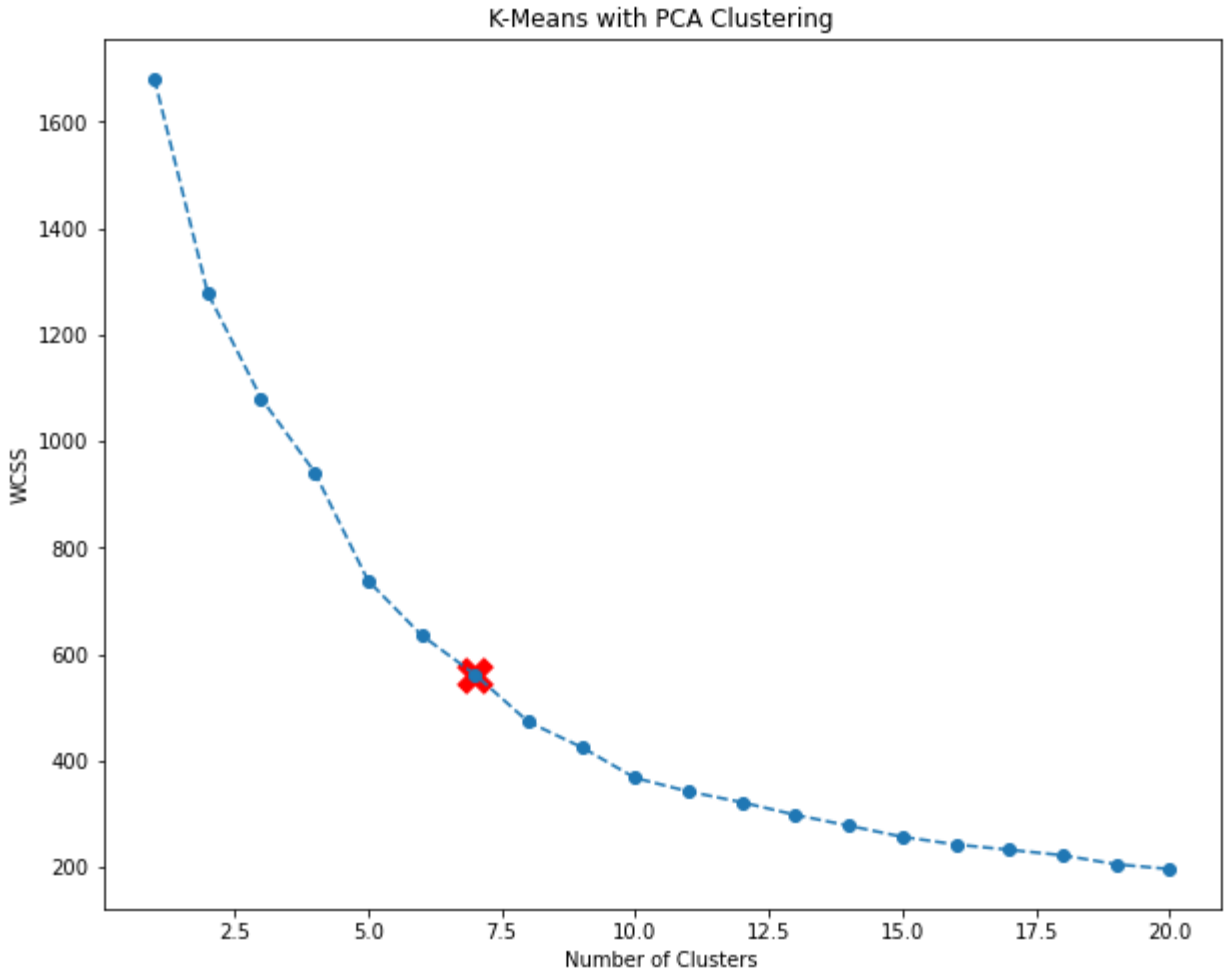
- Visualiser les CAH clusters sur un plan 2D (axes F1 & F2)



k-means

Méthode du coude:
pour déterminer le nombre optimal de clusters (k)

Pour ce cas : 7 clusters
k = 7



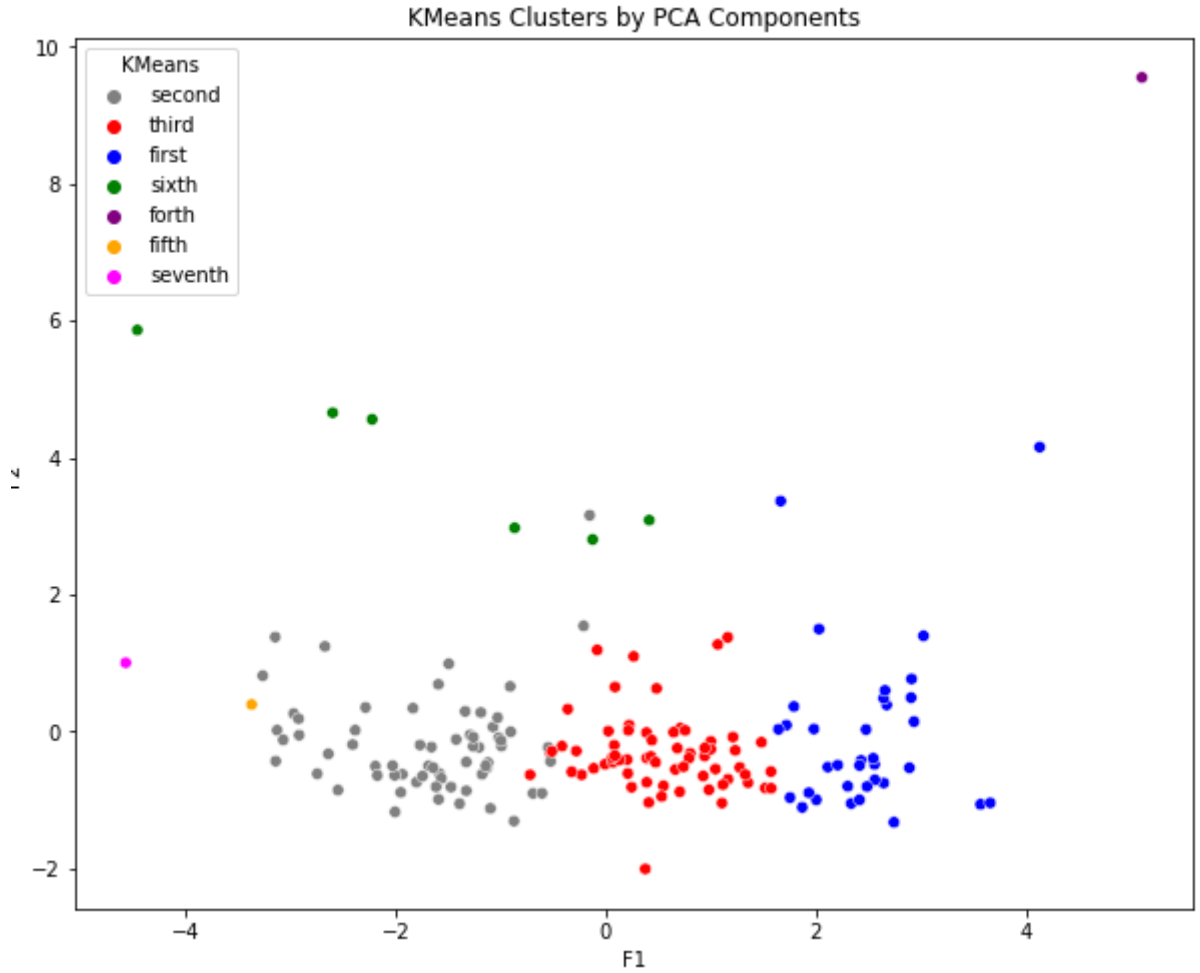
analyse des l'algorithme k-means

Création d'un nouvel dataset avec les composants PCA et leurs KMeans_clusters correspondantes

Zone	F1	F2	F3	F4	F5	F6	CAH	KMeans
Afghanistan	-2.968891	0.262997	0.338730	-0.170927	-0.486712	-0.478188	second	second
South Africa	0.550650	-0.791201	0.867292	-0.008651	-0.507840	-0.053089	fifth	third
Albania	2.743652	-1.322731	-0.173795	0.288962	-0.232491	0.215809	fifth	first
Algeria	0.089149	0.652312	-0.138205	-0.211687	0.097001	-0.367632	fifth	third
Germany	3.024856	1.397656	0.844625	-0.203860	0.044475	-0.022076	fifth	first
...	...	...	...	...	...	...	...	...
United Arab Emirates	1.211660	-0.078635	-0.162797	-0.201452	-0.142955	-0.135466	fifth	third
Ecuador	-1.258662	-0.082509	-0.371558	-1.092172	1.270242	-0.640393	fifth	second
United States of America	4.130186	4.153897	2.340002	-0.194177	0.048398	-0.329269	forth	first
Ethiopia	-3.262282	0.816696	-0.302317	-1.461935	1.934545	-1.508563	second	second
Solomon Islands	-2.027609	-0.496434	0.510167	0.123591	0.251773	-0.525118	second	second

analyse des l'algorithme k-means

- Visualiser les KMeans clusters sur un plan 2D (axes F1 & F2)



comparaison des clusters de CAH & KMeans

col_0	0	1	2	3	4	5	6
row_0							
1	0	0	0	0	0	3	0
2	0	53	4	0	0	0	0
3	0	0	0	0	0	0	1
4	2	1	0	1	0	0	0
5	30	6	59	0	0	2	0
6	3	5	2	0	1	1	0

col_0 représentent le cluster **KMeansles**
row_0 représentent le cluster **CAH**

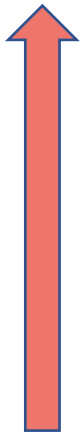
Correspondance CAH – K-Means :
le **groupe 1** du CAH & le **groupe 5** des K-Means
le **groupe 3** du CAH & le **groupe 6** des K-Means

des correspondances
mais n'est pas exactes

conclusion

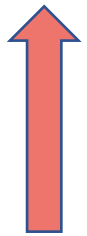
ACP: F1 & F2 ~42.4%

Premier Component: F1

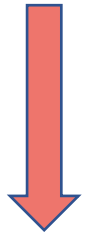


- Population Growth Rate
- Disponibility_alimentaire (kcal/p/j)
- Production

Seconde Component: F2



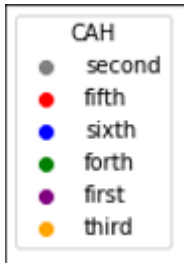
- Importation
- Protein consumption (g/p/j)
- Political Stability



- Exchange Rate

conclusion

CAH



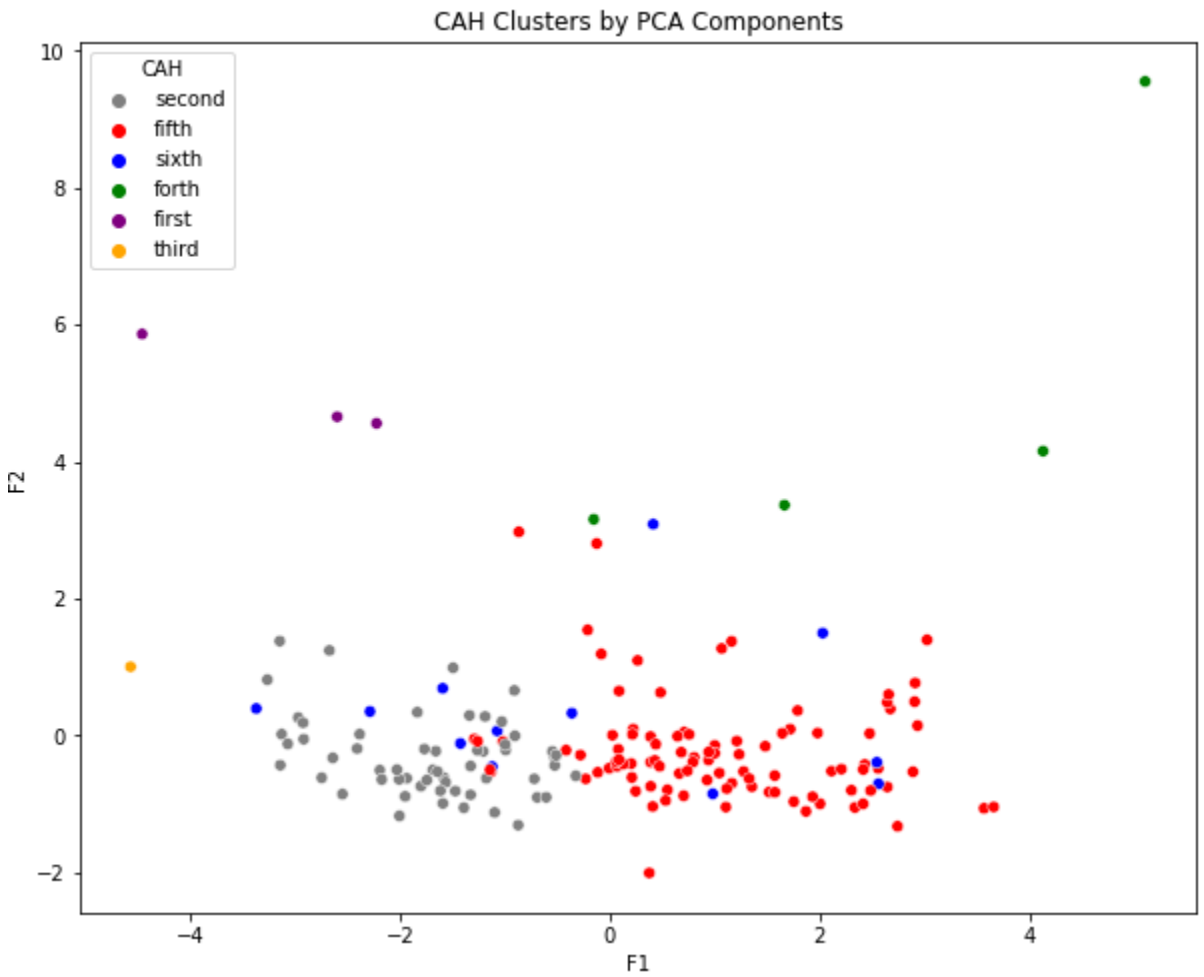
cluster	Zone
1	3
2	57
3	1
4	4
5	97
6	12

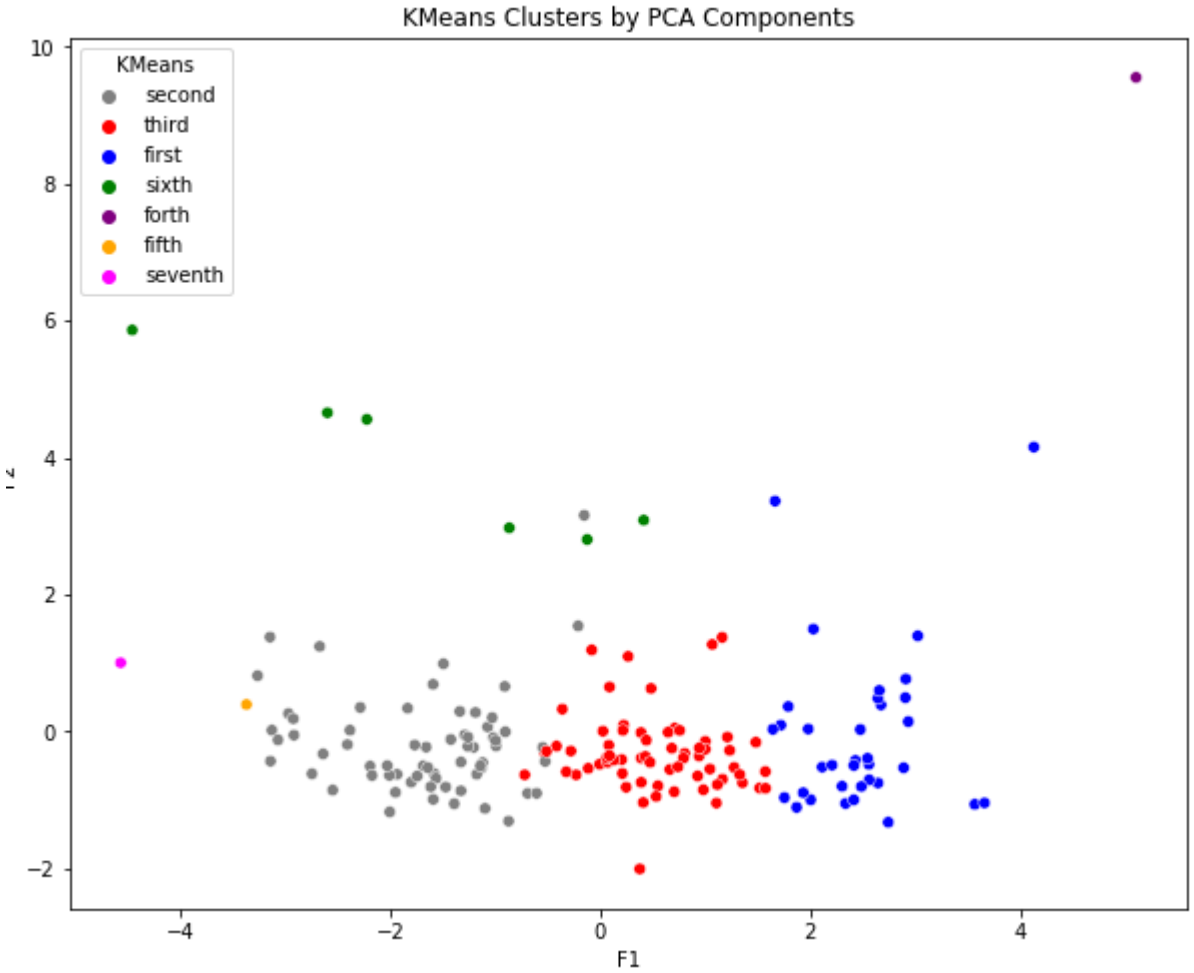
- Population Growth Rate
- Disponibility_alimentaire (kcal/p/j)
- Production

KMeans

cluster	Zone
0	35
1	65
2	65
3	1
4	1
5	6
6	1

conclusion


CAH Clusters by PCA Components
CAH
second
fifth
sixth
forth
first
third
F1
F2


KMeans Clusters by PCA Components
KMeans
second
third
first
sixth
forth
fifth
seventh
F1

conclusion

De CAH: 101 pays

```
target_countries_CAH = G[G['CAH'].isin(['fourth', 'fifth']) == True]
target_countries_CAH
```

Zone	F1	F2	F3	F4	F5	F6	CAH	KMeans
South Africa	0.550650	-0.791201	0.867292	-0.008651	-0.507840	-0.053089	fifth	third
Albania	2.743652	-1.322731	-0.173795	0.288962	-0.232491	0.215809	fifth	first
Algeria	0.089149	0.652312	-0.138205	-0.211687	0.097001	-0.367632	fifth	third
Germany	3.024856	1.397656	0.844625	-0.203860	0.044475	-0.022076	fifth	first
Antigua and Barbuda	0.211471	-0.610732	-0.518515	0.033971	-0.212887	0.007841	fifth	third
...	...	...	...	...	...	...	...	...
Vietnam	0.484801	0.630625	0.401572	-0.029488	0.029361	-0.038214	fifth	third
Egypt	-0.865778	2.973963	-1.688547	-0.505330	-0.237367	-0.445795	fifth	sixth
United Arab Emirates	1.211660	-0.078635	-0.162797	-0.201452	-0.142955	-0.135466	fifth	third
Ecuador	-1.258662	-0.082509	-0.371558	-1.092172	1.270242	-0.640393	fifth	second
United States of America	4.130186	4.153897	2.340002	-0.194177	0.048398	-0.329269	fourth	first

101 rows × 8 columns

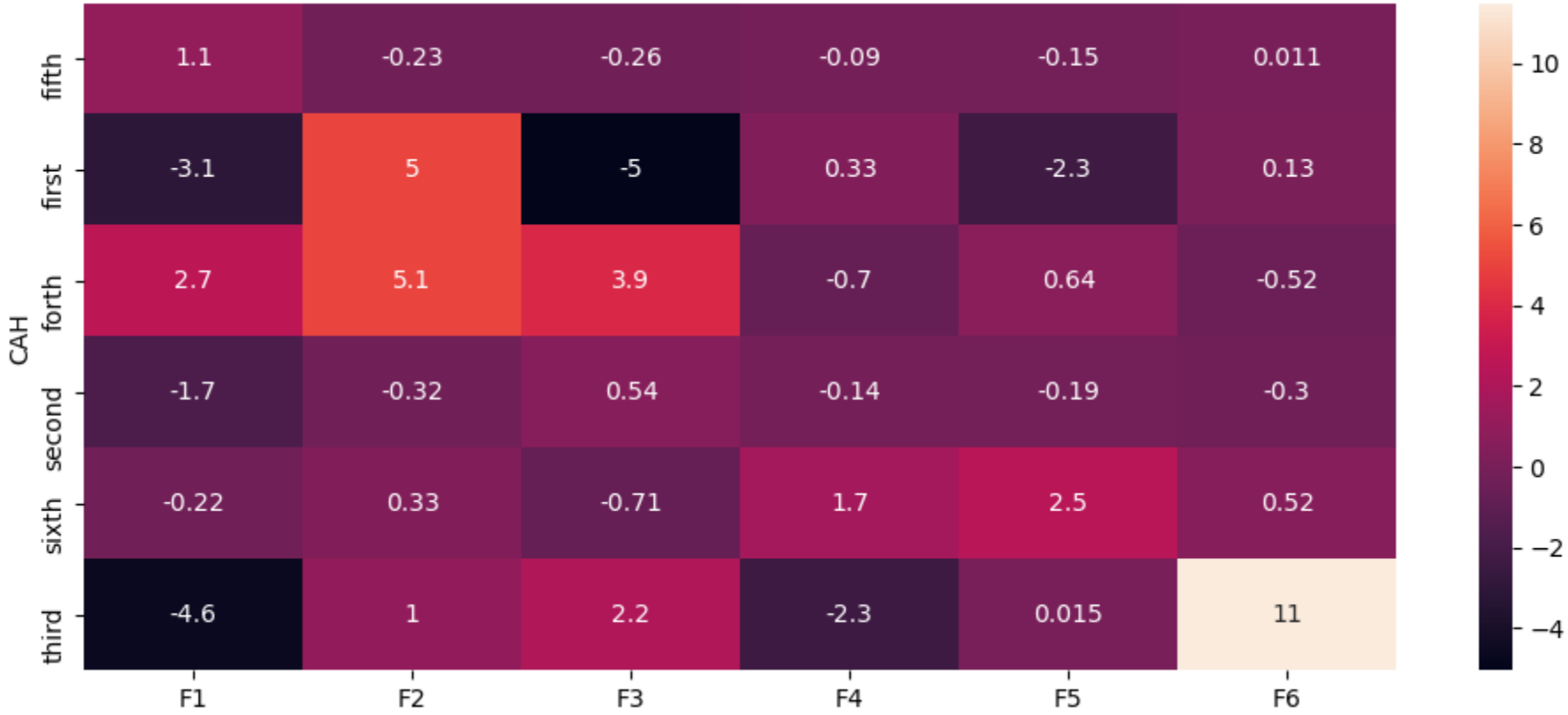
De Kmeans: 36 pays + 65 pays

```
target_countries_KMeans = G[G['KMeans'].isin(['first', 'fourth']) == True]
target_countries_KMeans
```

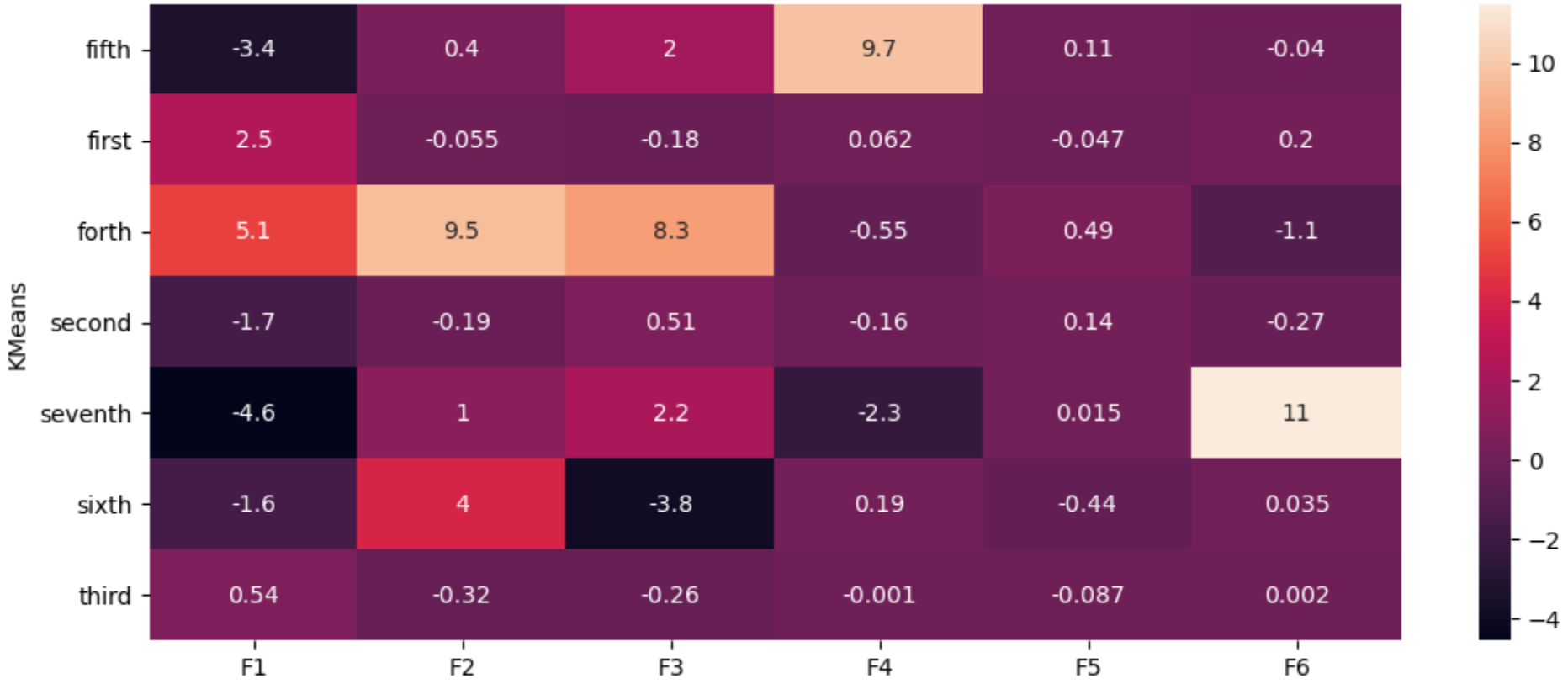
Zone	F1	F2	F3	F4	F5	F6	CAH	KMeans
Albania	2.743652	-1.322731	-0.173795	0.288962	-0.232491	0.215809	fifth	first
Germany	3.024856	1.397656	0.844625	-0.203860	0.044475	-0.022076	fifth	first
Australia	1.720653	0.095096	-0.690481	-0.237192	0.004462	-0.025771	fifth	first
Austria	2.431273	-0.420822	-0.705833	-0.544933	-0.090279	0.105955	fifth	first
Belgium	2.674457	0.389820	0.211183	0.501575	-0.505837	0.257224	fifth	first
Brazil	1.665599	3.365556	1.301401	-1.367437	1.572549	-1.031222	fourth	first
Canada	1.791187	0.366926	-0.116092	0.019470	-0.036341	0.011350	fifth	first
China - Hong Kong SAR	2.477960	0.030973	-1.220886	-0.037030	-0.008845	0.365356	fifth	first
China, mainland	5.103438	9.548113	8.323571	-0.544738	0.494889	-1.083665	fourth	fourth
Croatia	1.754803	-0.963065	-0.325415	0.097781	-0.155079	0.114587	fifth	first
Denmark	2.416838	-0.492872	-0.695416	-0.024824	-0.282109	0.203920	fifth	first
Spain	2.646060	0.485922	0.279113	0.160651	-0.030721	0.177500	fifth	first
Estonia	2.005146	-0.994299	-0.393120	0.056167	0.118558	0.206356	fifth	first
Finland	2.645711	-0.749247	-0.918612	-0.662766	-0.462729	0.406171	fifth	first
France	2.906310	0.496134	0.585978	-0.872284	-1.754234	-0.179569	fifth	first
Russian Federation	2.934691	0.146359	1.364682	1.858764	-1.466437	0.806903	fifth	first

Limitation: Les axes F1 et F2 ne représentent que 42,2 % des variables

heatmap


CAH
fifth
first
forth
second
sixth
third
1.1
-0.23
-0.26
-0.09
-0.15
0.011
-3.1
5
-5
0.33
-2.3
0.13
2.7
5.1
3.9
-0.7
0.64
-0.52
-1.7
-0.32
0.54
-0.14
-0.19
-0.3
-0.22
0.33
-0.71
1.7
2.5
0.52
-4.6
1
2.2
-2.3
0.015
11
F1
F2
F3
F4
F5
F6
10
8
6
4
2
0
−2
−4

heatmap


KMeans
fifth
first
forth
second
seventh
sixth
third
-3.4
0.4
2
9.7
0.11
-0.04
2.5
-0.055
-0.18
0.062
-0.047
0.2
5.1
9.5
8.3
-0.55
0.49
-1.1
-1.7
-0.19
0.51
-0.16
0.14
-0.27
-4.6
1
2.2
-2.3
0.015
11
-1.6
4
-3.8
0.19
-0.44
0.035
0.54
-0.32
-0.26
-0.001
-0.087
0.002
F1
F2
F3
F4
F5
F6
10
8
6
4
2
0
−2
−4