Père Dodet

Etude du développement à l'international

Problématique

Quels sont les pays les plus susceptibles d'accueillir facilement nos produits ?

SOMMAIRE

Traitement des données

- Importation des données de la FAO
- Concaténation et sélection des variables utilisées pour l'étude

Premier clustering

- Dendrogramme et présentation des clusters
- Choix du cluster candidat
- Première ACP pour valider notre choix

Deuxième clustering

- Découpage du cluster candidat
- Sélection du cluster final
- Tests de comparaison pour démontrer la différence entre les classes
- Conclusion

Traitement des données

```
ani = pd.read_csv("Fichiers CSV/Produits Animaux 2013.csv")
veg = pd.read_csv("Fichiers CSV/Produits Vegetaux 2013.csv")
pop = pd.read_csv("Fichiers CSV/Population 2012-2013.csv")
```

Suppression de l'individu « Chine » car présent à la fois en tant que pays, et en tant qu'agrégation de ses différentes régions

	Country	Population 2012	Population 2013	Evo population %	Animal protein supply quantity (g/capita/day)	Vegetal protein supply quantity (g/capita/day)	Total protein supply (g/capita/day)	Ratio Animal protein / Total protein	Food supply (kcal/capita/day)
0	Afghanistan	29825000	30552000	2.437552	12.21	46.05	58.26	0.209578	2087.0
1	Albania	3162000	3173000	0.347881	59.41	51.96	111.37	0.533447	3188.0
2	Algeria	38482000	39208000	1.886596	24.98	66.94	91.92	0.271758	3293.0
3	Angola	20821000	21472000	3.126651	18.40	38.87	57.27	0.321285	2474.0
4	Antigua and Barbuda	89000	90000	1.123596	56.83	26.66	83.49	0.680680	2416.0

	Country	Evo population %	Ratio Animal protein / Total protein	Total protein supply (g/capita/day)	Food supply (kcal/capita/day)
0	Afghanistan	2.437552	0.209578	58.26	2087.0
1	Albania	0.347881	0.533447	111.37	3188.0
2	Algeria	1.886596	0.271758	91.92	3293.0
3	Angola	3.126651	0.321285	57.27	2474.0
4	Antigua and Barbuda	1.123596	0.680680	83.49	2416.0

Détails de certains calculs

- $Evo\ population\ \% = \frac{(Population\ 2013 - Population\ 2012)}{Population\ 2012}$

	Country	Population 2012	Population 2013	Evo population %
0	Afghanistan	29825000	30552000	2.437552
1	Albania	3162000	3173000	0.347881
2	Algeria	38482000	39208000	1.886596
3	Angola	20821000	21472000	3.126651
4	Antigua and Barbuda	89000	90000	1.123596

- $Ratio\ Animal\ protein/Total\ protein = \frac{Animal\ protein\ supply\ quantity}{Animal\ protein\ supply\ quantity + Vegetal\ protein\ supply\ quantity}$

Animal protein supply quantity (g/capita/day)	Vegetal protein supply quantity (g/capita/day)	Ratio Animal protein / Total protein
12.21	46.05	0.209578
59.41	51.96	0.533447
24.98	66.94	0.271758
18.40	38.87	0.321285
56.83	26.66	0.680680

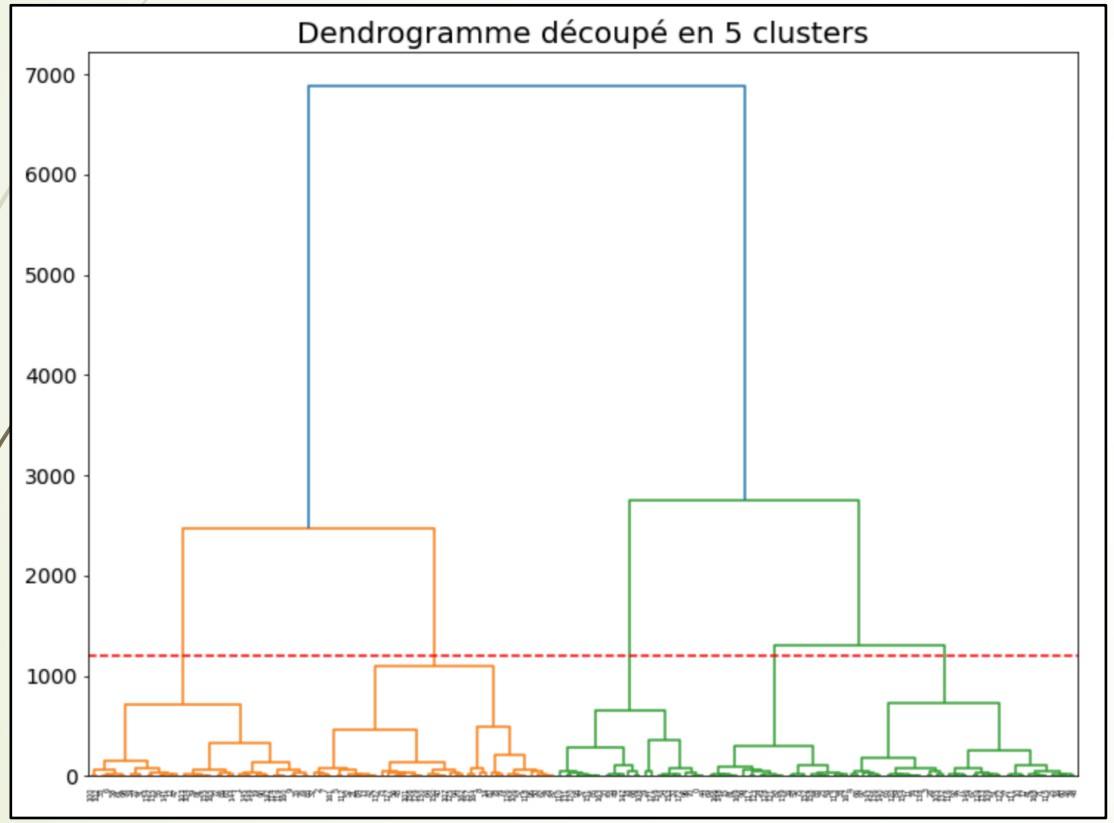
- « Food supply » est obtenue par la commande suivante :

```
kcal_by_country = data.groupby('country')['food_supply_(kcal/capita/day)'].sum()
```

Premier clustering

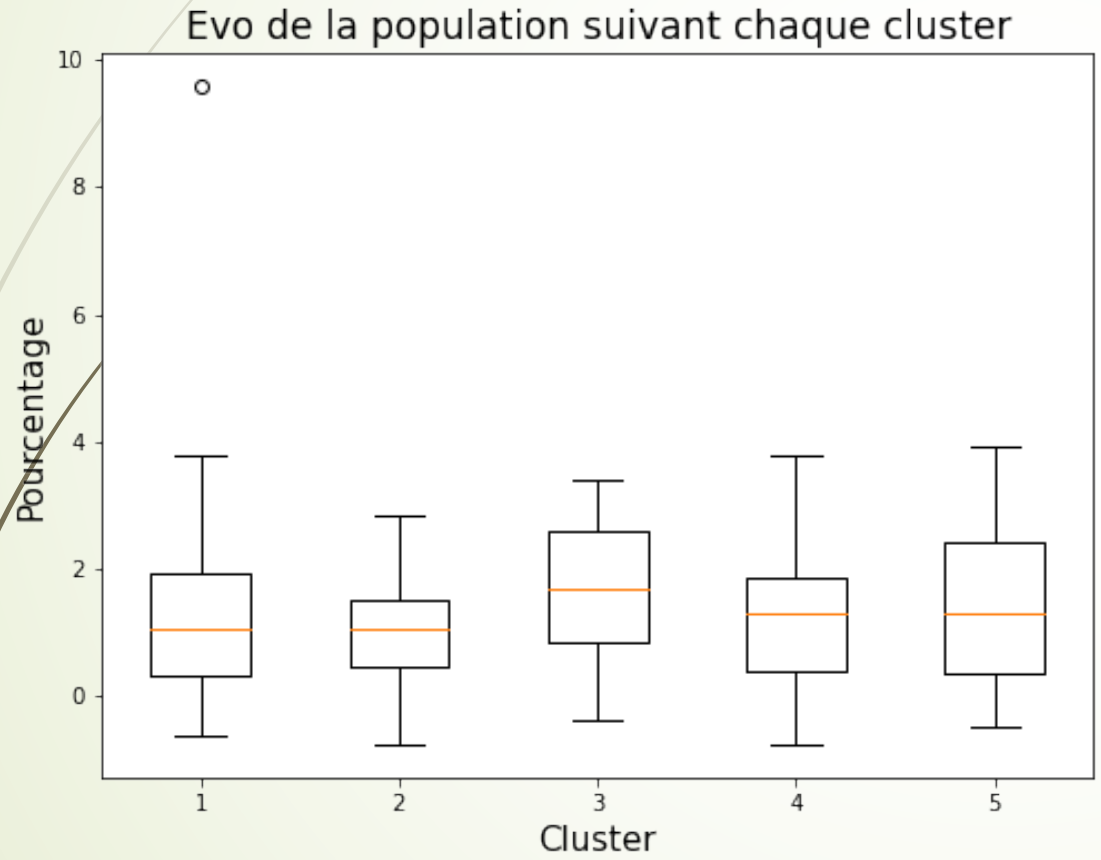
1er clustering

Construction du dendrogramme via la méthode « Ward », permettant de minimiser la variance intraclasse. Nous le découpons pour former 5 clusters.



	Country	Cluster
0	Afghanistan	3
1	Albania	1
2	Algeria	2
3	Angola	5
4	Antigua and Barbuda	5
...	...	...
169	Venezuela (Bolivarian Republic of)	5
170	Viet Nam	4
171	Yemen	3
172	Zambia	3
173	Zimbabwe	3

Détection d'un outlier via un boxplot

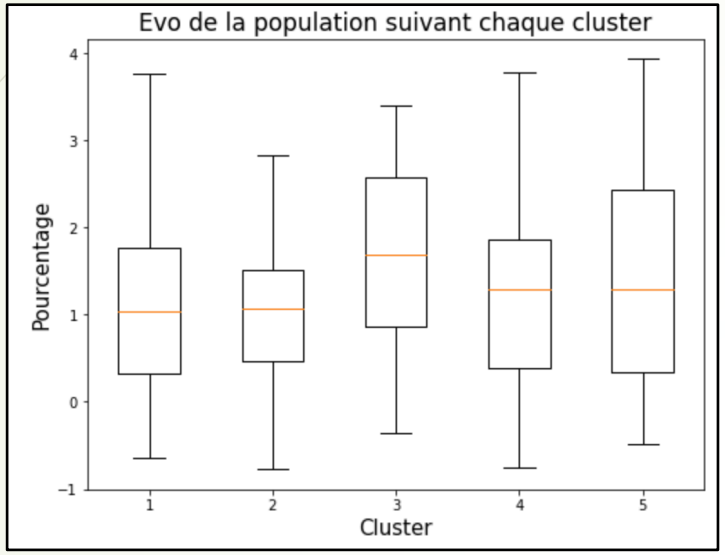


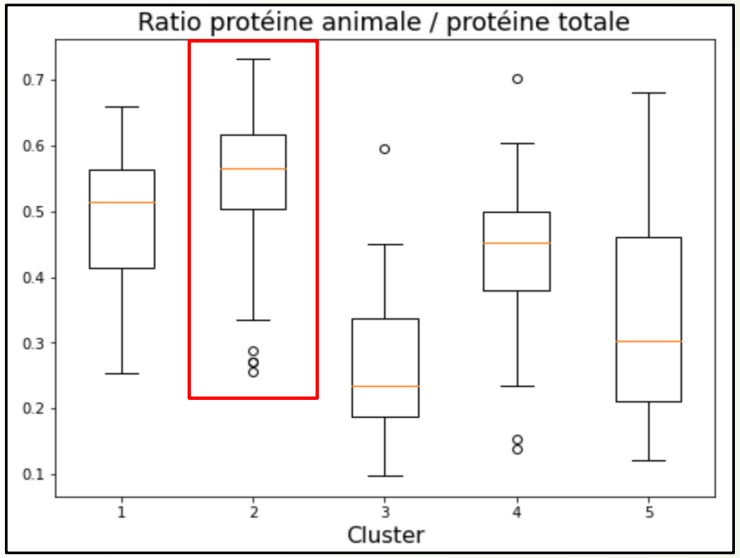
La population omanaise (Oman) a évolué très fortement entre 2012 et 2013 :

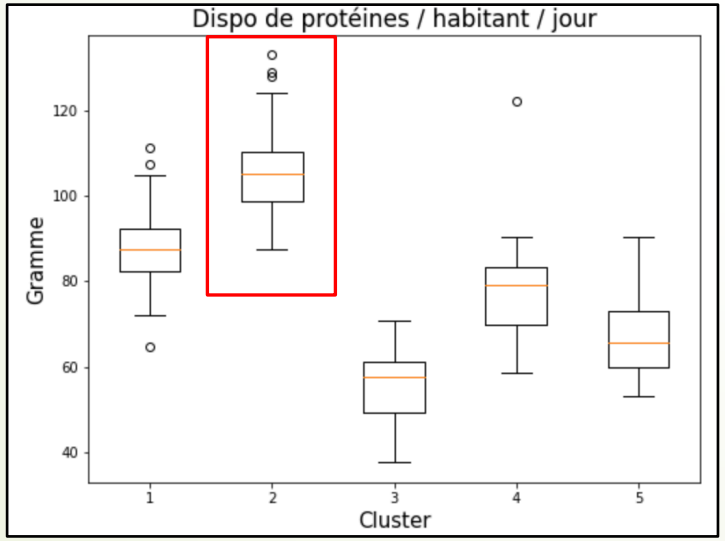
- → *Augmentation drastique de la production de pétrole,*
- → *Donc forte immigration.*

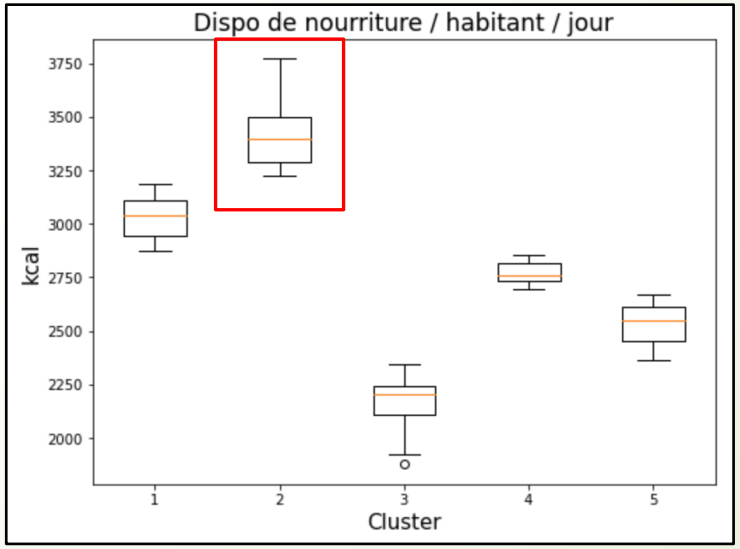
1er clustering

Description des clusters *(Oman a été retiré)*



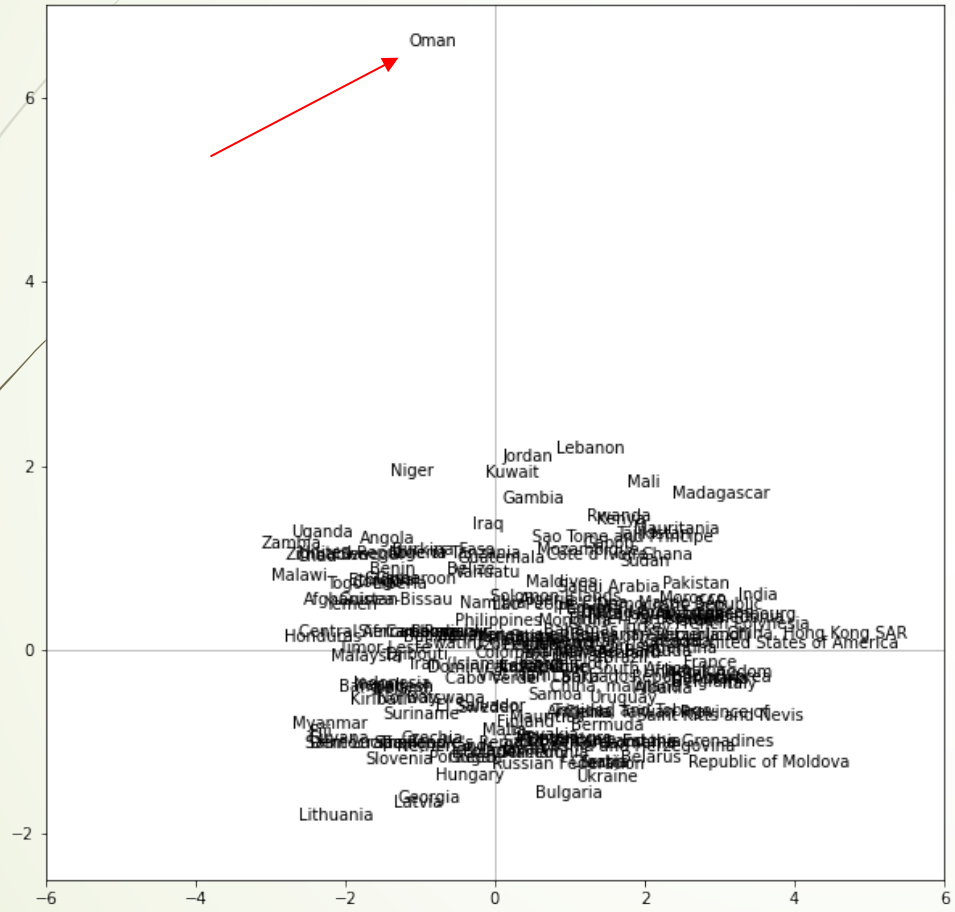






*Le **cluster 2** semble le plus intéressant dans toutes les variables liées aux habitudes alimentaires.*

Confirmation de l'outlier via l'ACP



On retire une nouvelle fois Oman pour plus de clarté dans notre représentation des individus dans le premier plan factoriel.

1er clustering

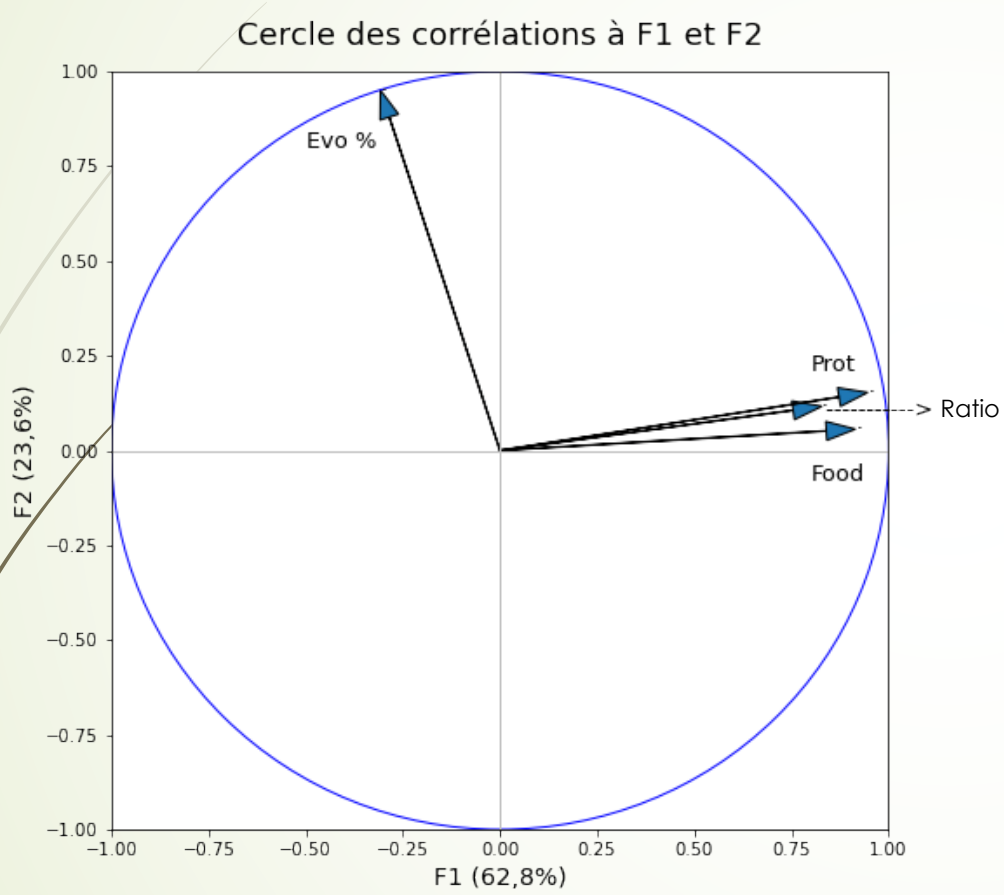
Ratio : *Ratio Animal protein / Total protein*

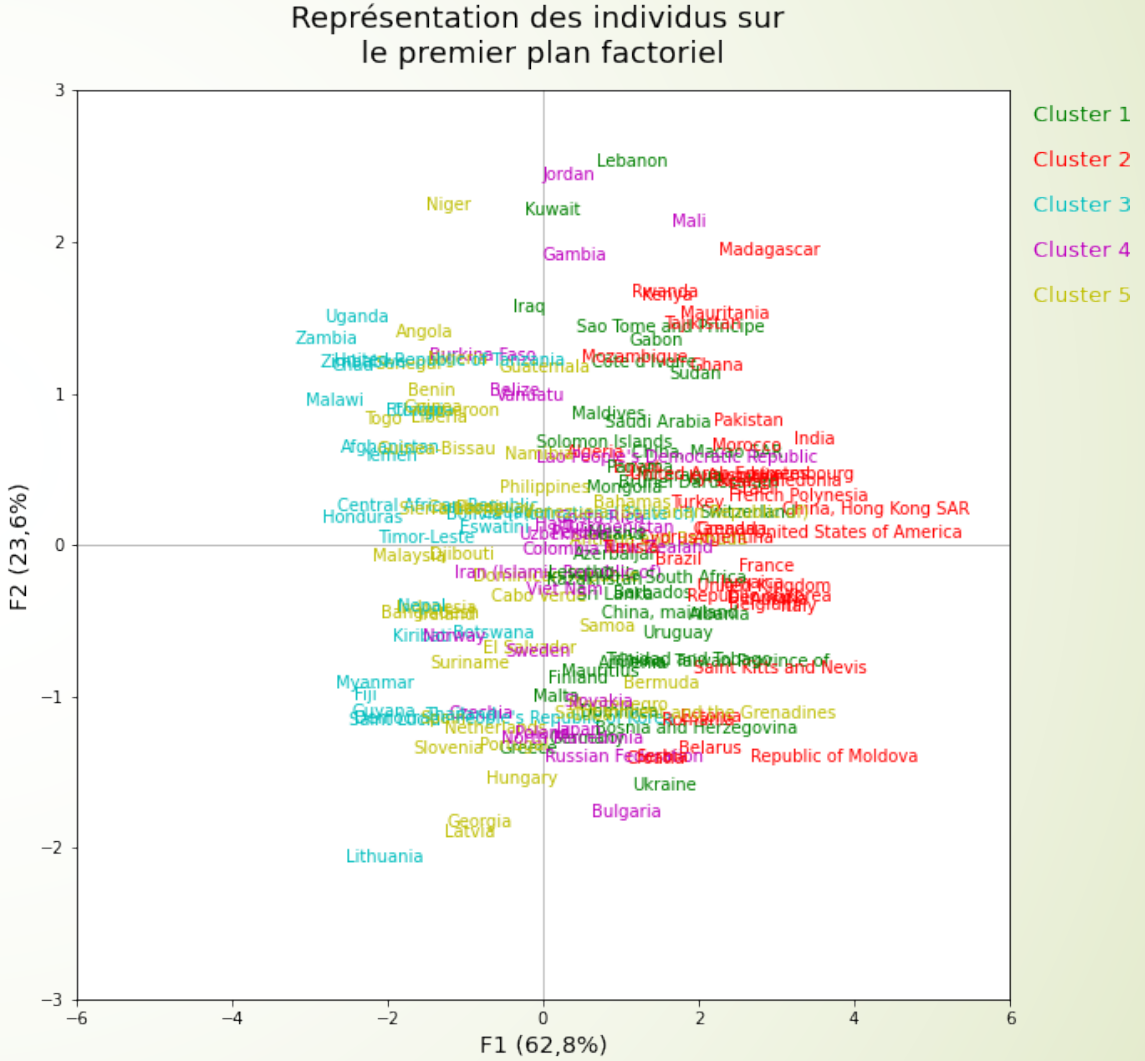
Prot : *Total protein supply (g)*

Food : *Food supply (kcal)*

Evo % : *Evo population %*

Description des clusters





- Les variables fortement corrélées à F1 sont celles décrivant les habitudes alimentaires.
- La variable fortement corrélée à F2 est l'évolution de la population.

2ème clustering

Découpage du cluster 2

- Le dataset initial est maintenant réduit aux seuls individus du cluster 2. Oman n'en faisant pas partie, l'avoir retiré n'impacte en rien la suite de notre étude.
- Nous procédons, avec la même méthode que précédemment, en un découpage en 4 nouveaux clusters.
- Nous vérifierons aussi que ces clusters diffèrent bien via des tests de comparaison dans le cas de variables gaussiennes.

2ème clustering

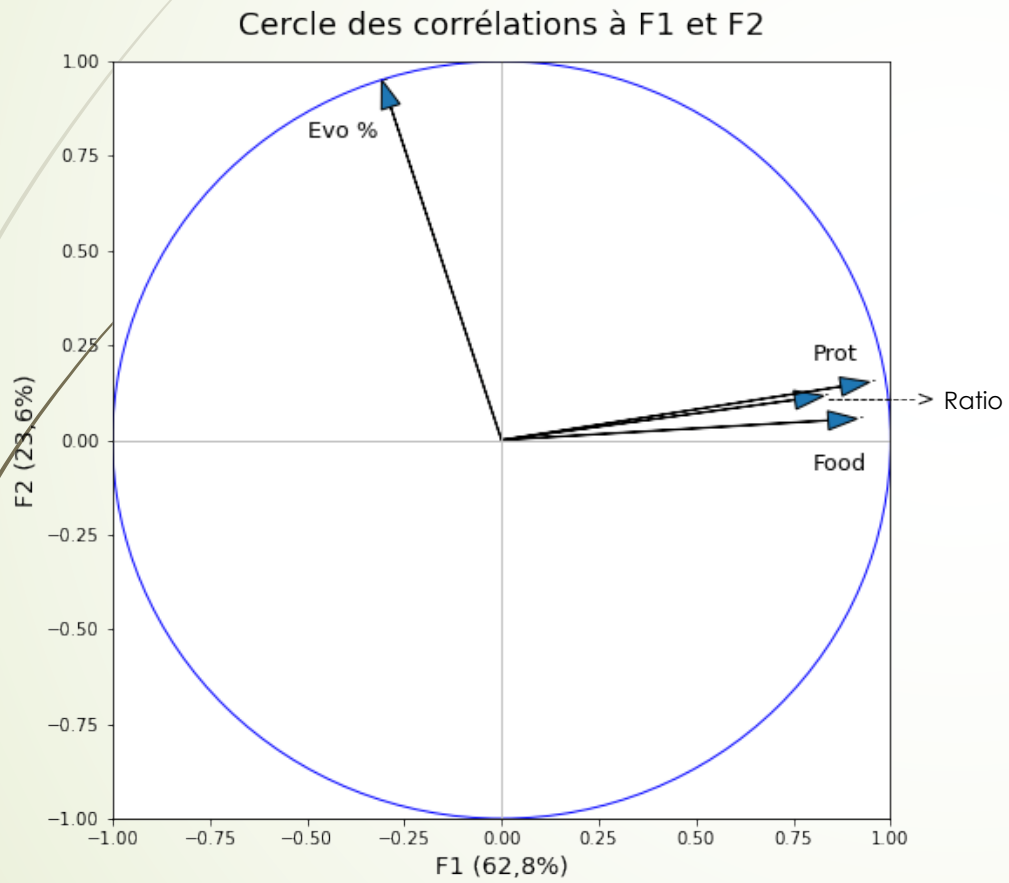
Description des clusters

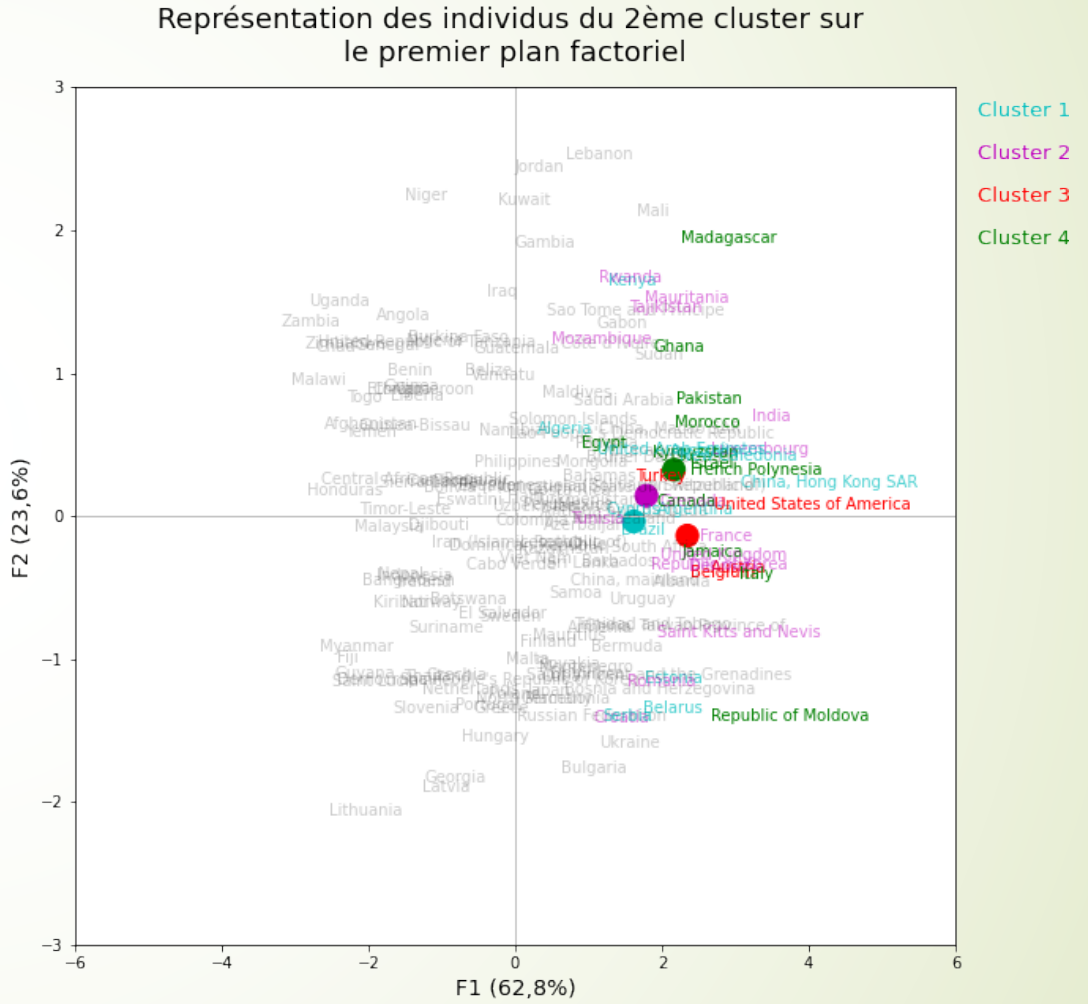
Ratio : *Ratio Animal protein / Total protein*

Prot : *Total protein supply (g)*

Food : *Food supply (kcal)*

Evo % : *Evo population %*





Le cluster 4 semble être le plus favorable à F1 et F2 simultanément.

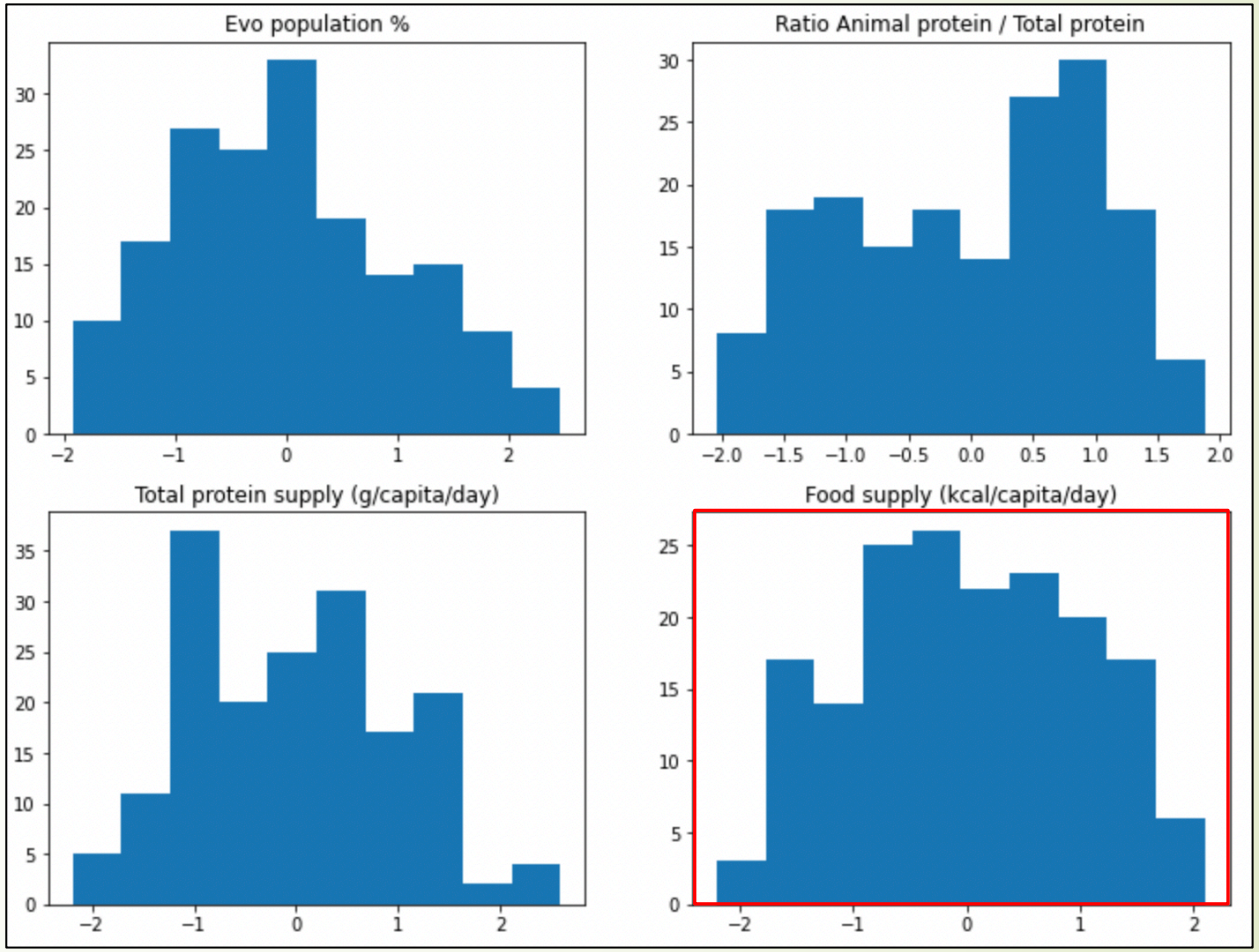
2ème clustering

Etude graphique de l'adéquation à une loi normale

Cherchons parmi nos 4 variables laquelle suit une loi normale.

Graphiquement, la variable « Food supply » semble suivre une loi normale.

Vérifions-le grâce à un test d'adéquation.



2ème clustering

Tests d'adéquation à une loi normale

Pour cela appliquons l'algorithme de Kolmogorov-Smirnov sur la variable « Food supply ».

- **L'hypothèse H0** que nous formulons : « Food supply » suit une loi normale (ou que sa fonction de répartition est celle d'une loi normale).
- Après centrage et réduction de la variable « Food supply » , via la commande kstest de Python, on obtient :

```
KstestResult(statistic=0.05689295619893475, pvalue=0.6093950240362977)
```

Nous obtenons une p-value de **0.6**, ce qui **ne nous permet pas** de rejeter l'hypothèse H0 pour tous les niveaux de tests classiques (5%, 10%, …)

« Food supply » suit donc une loi normale.

2ème clustering

Tests de comparaison

Testons maintenant si les clusters sont bien différents via un test de comparaison sur la variable « Food supply ».

Exemple entre les clusters 3 et 4 :

Test de la variance :

- **L'hypothèse H0** est : variance_cluster_3 = variance_cluster_4
- Cette statistique de test converge vers la loi de Fisher, ce qui donne sous Python une p-value de **0.76** > 0.05 (pour un niveau de test à 5%). On ne peut rejeter l'hypothèse H0.

Test de la moyenne :

- **L'hypothèse H0** est : moyenne_cluster_3 = moyenne_cluster_4
- Cette statistique de test converge vers la loi de Student, ce qui donne sous Python via la commande ttest_ind une p-value de **0,0000027** < 0.05. On rejette donc l'hypothèse H0.

Ces deux clusters sont bien différents.

2ème clustering

Conclusion

- Nous avons réparti le groupe des individus en 5 clusters. Le 2ème cluster semblait répondre le plus à nos exigences.
- Le nombre de pays étant encore trop important, nous avons découpé ce groupe en 4 nouveaux clusters.
- Le cluster 4 est le plus favorable à accueillir notre entreprise.
- Tous les clusters sont bien différents car ils ne valident pas les tests de comparaison effectués.

Les 12 pays à considérer sont :

Canada, Egypte, Polynésie française, Israël, Italie, Jamaïque, Kyrgyzstan, Moldavie, Madagascar, Ghana, Pakistan et Maroc.