Projet n°5 :
« Produisez une étude de marché »

Jérémy Vangansberg - OpenClassrooms

Sommaire

Introduction : import des données et enjeux

Import des données

- Les données proviennent de la **FAO**
- Année étudiée **: 2017**
- Sections utilisées :

  « Nouveaux bilans alimentaire »

  « Production »

  « Population »

  « Statistiques-macro »

Enjeux & contraintes

Enjeux :

- Lancer une nouvelle filière de poulet **bio haut de gamme**
- Exporter ces produits à l'international

Contraintes :

- La production est faite en France
- Etablir 5 groupes, puis un découpage plus fin pour cibler quelques pays

Où s’implanter ?

Caractéristiques des pays recherchés :

Caractéristiques du pays	Variables étudiées
Pouvoir d’achat élevé de la population	PIB par habitant
Disponibilités alimentaires importantes	Disponibilité alimentaire en kcal et protéines
Faible production de volaille	Ratio entre la production de volailles et le nombre d’habitants
Forte habitude de consommation de viandes	% de protéines issues de la viande

La variable taux d’évolution démographique sera également étudiée. Elle ne constitue pas un critère de recherche directe mais elle est utile pour effectuer le clustering

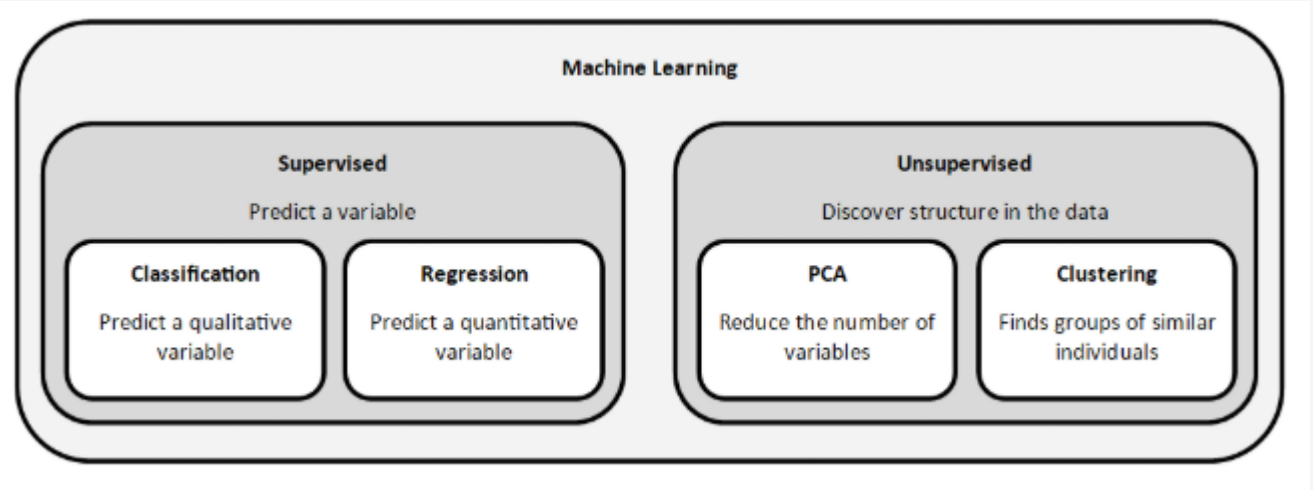
Méthodologie : détails des algorithmes utilisés

Méthodologie

Outils utilisés :

- Un **classement ascendant hiérarchique** et un **dendrogramme** pour le visualiser. C'est un outil de clustering
- **Analyse en composantes principales** (ACP). Elle permet de réduire les dimensions.

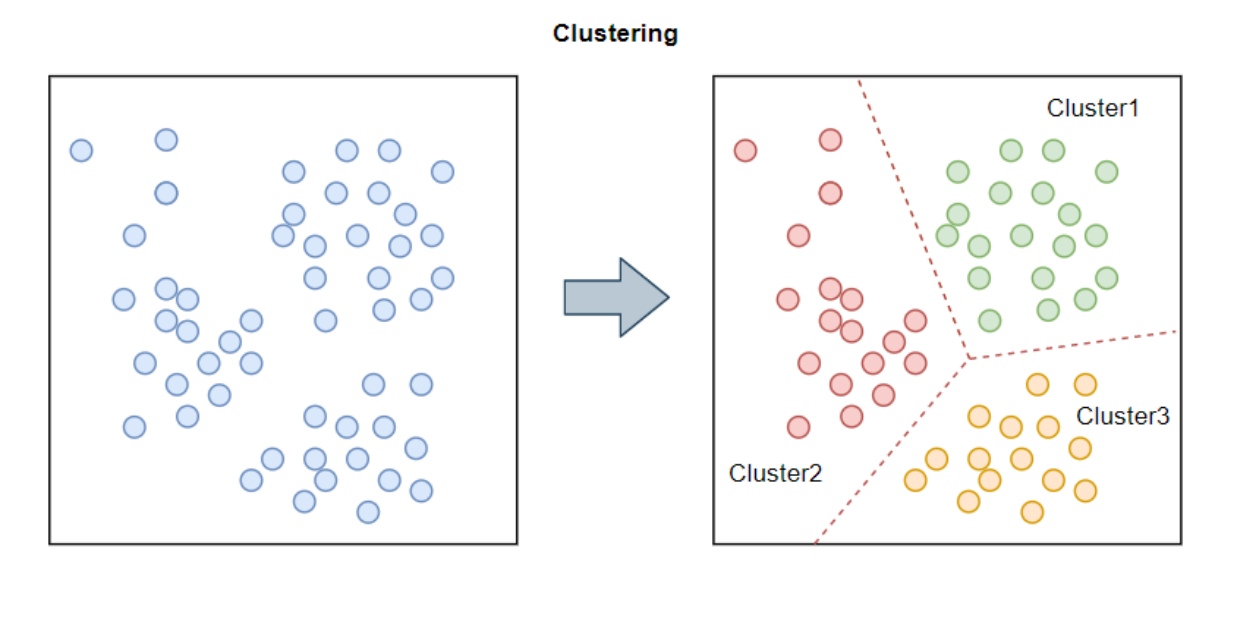
Le machine learning et ses composantes :



Source : https://openclassrooms.com/fr/courses/5869986-perform-an-exploratory-data-analysis/6435201-discover-the-principles-of-multivariate-exploratory-data-analysis

Clustering

Fonctionnement du clustering :



Source : https://www.ecloudvalley.com/mlintroduction/

Caractéristiques des données

- Individus : **167 pays**
- **6 variables** :

  Pourcentage de protéines animales consommées

  Disponibilité alimentaire en kcal par jour et par personne

  Disponibilité en protéines en gramme par personne et par jour

  Taux d'évolution démographique entre **2012 et 2017**

  PIB par habitant

  Le ratio de production de volailles par habitant

Aperçu

country	percent_prot_ani	food_avl_kcal_p_d	prot_avl_p_d	pop_growth_rate	PIB_hab	ratio_prod
Arménie	0.457692	3078.0	97.38	0.020996	3933.682101	3.599576
Afghanistan	0.195045	2000.0	54.09	0.164779	605.557362	0.758153
Albanie	0.554914	3399.0	119.55	-0.010270	4445.132198	4.476853
Algérie	0.276690	3349.0	92.92	0.107140	4051.244377	6.668746
Angola	0.304565	2270.0	54.11	0.187544	4100.291004	0.901204

Préparation des données

- Standardisation des 6 variables utilisées
- Les données sont exprimées dans des **unités différentes**, afin de les utiliser dans les algorithmes de l'ACP et du classement ascendant hiérarchique, il faut les **centrer et les réduire** :

$$X_{cr} = \frac{X - \bar{X}}{s_X}$$

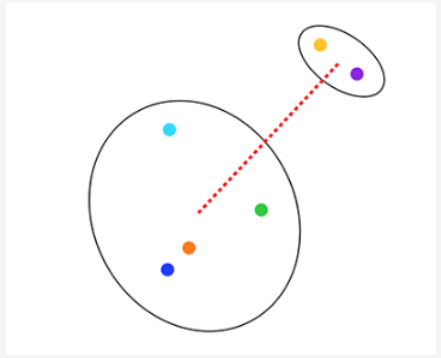
country	percent_prot_ani	food_avl_kcal_p_d	prot_avl_p_d	pop_growth_rate	PIB_hab	ratio_prod
Arménie	0.457692	3078.0	97.38	0.020996	3933.682101	3.599576
Afghanistan	0.195045	2000.0	54.09	0.164779	605.557362	0.758153
Albanie	0.554914	3399.0	119.55	-0.010270	4445.132198	4.476853
Algérie	0.276690	3349.0	92.92	0.107140	4051.244377	6.668746
Angola	0.304565	2270.0	54.11	0.187544	4100.291004	0.901204

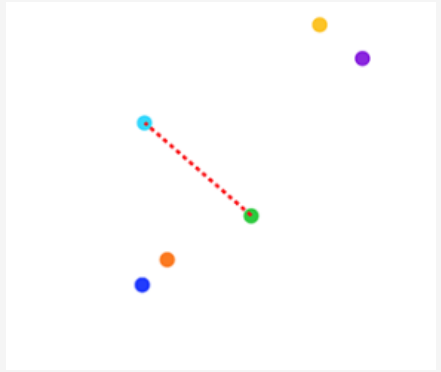
	percent_prot_ani	food_avl_kcal_p_d	prot_avl_p_d	pop_growth_rate	PIB_hab	ratio_prod
0	0.219210	-1.453557	0.838412	-0.933572	-0.756038	1.531385
1	0.479071	-1.889459	1.184358	1.074500	-1.296228	-0.944684
2	0.830685	-1.346889	1.945880	0.606338	-1.345883	0.029374
3	-0.748934	1.476802	-1.232921	0.584555	1.829198	-0.214936
4	-0.511060	-0.690449	-0.483493	-0.504724	-0.502080	0.052559
5	-0.732973	-0.908644	-0.678736	-0.543222	-0.899800	-0.893320

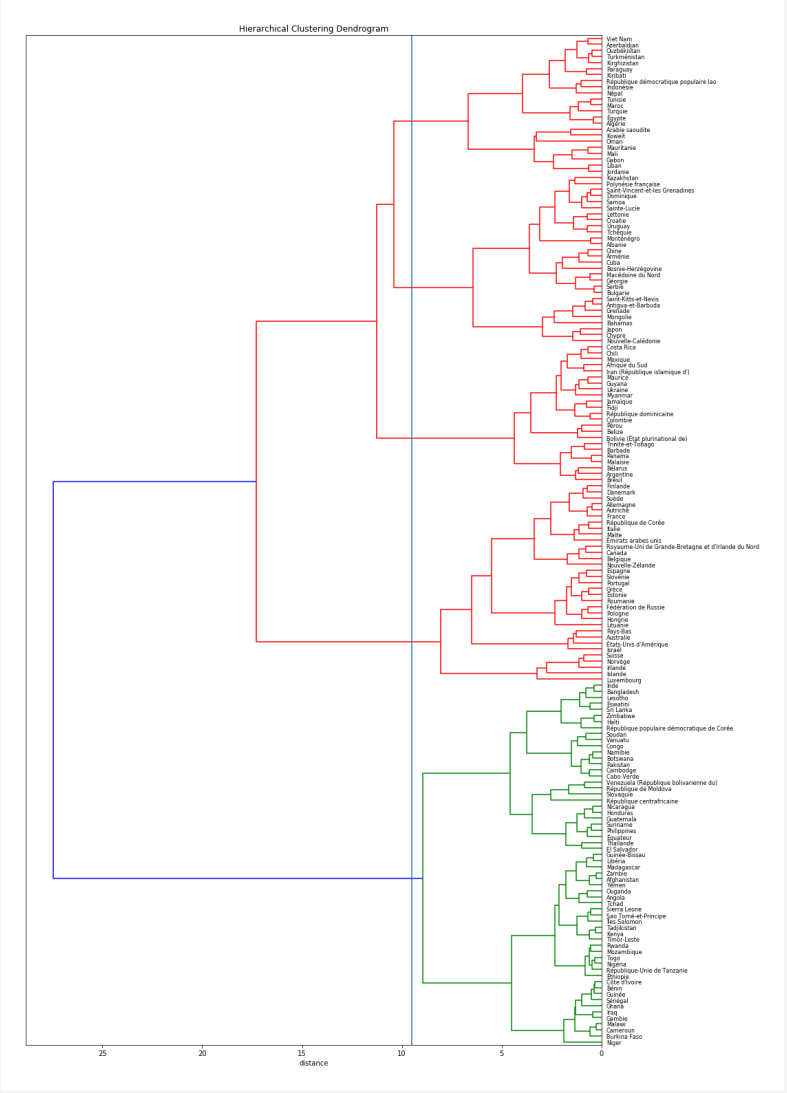
Classement Ascendant Hiérarchique : Dendrogramme

Caractéristiques :

- *Méthode de Ward* : elle permet de réduire l'inertie intra-classe



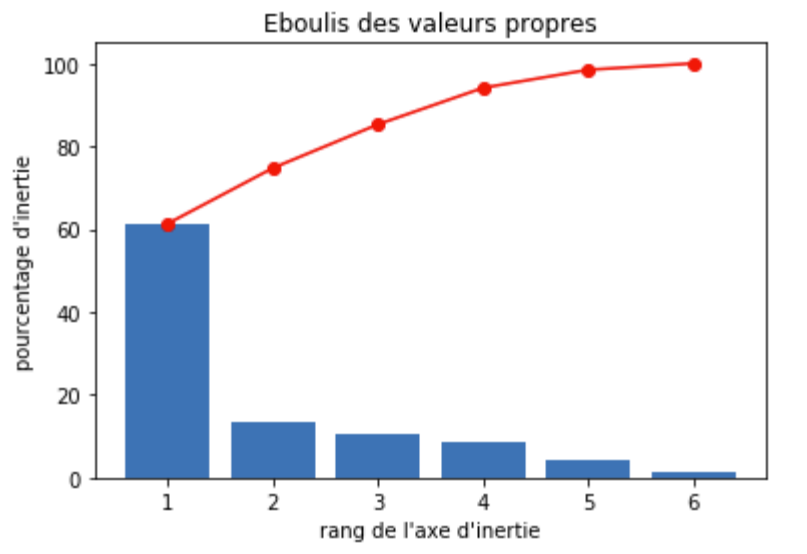




Parenthèses :

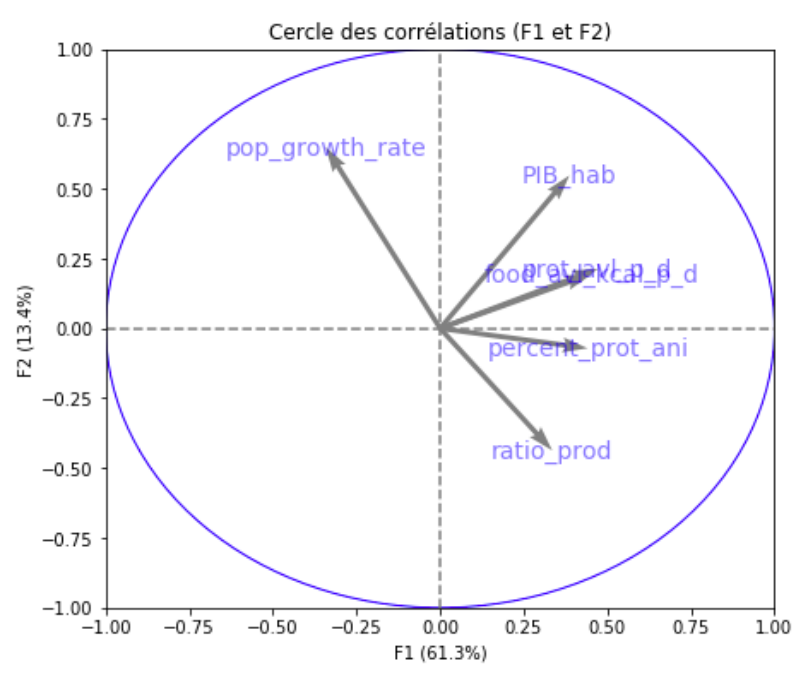
- *J'ai effectué un découpage en 5 clusters pour respecter les consignes du projet. En effet j'ai choisi d'étudié la dernière année disponible (2017) ce qui m'a amené à travailler sur la section « Nouveaux Bilans Alimentaire » et non pas la section « Bilans Alimentaire (Ancienne méthodologie et population) » sur lequel le projet est conçu. Ce qui m'a conduit à avoir un résultat légèrement différent de celui attendu dans l'énoncé du projet.*

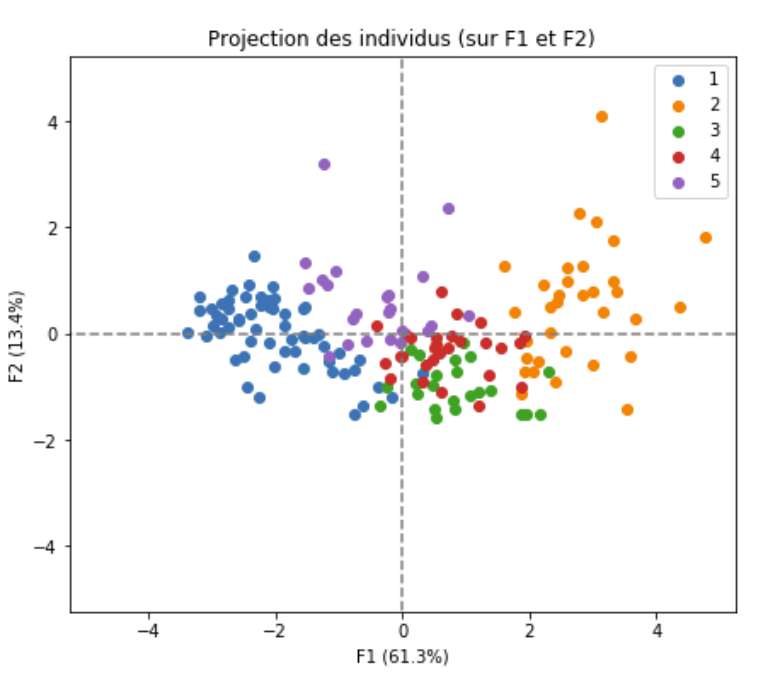
ACP – Eboulis des valeurs propres



- Le premier plan factoriel explique plus de 70% des données
- Les autres plan factoriel apporte très peu d'informations. En effet les axes inerties apportent moins de variabilité qu'une variable initiale
- *En utilisant la formule (100/p)% chaque variable initiale apportait 16,67% de l'information*

ACP – Cercle des corrélations et projection des individus





Combinaison linéaire des deux principaux axes d'inertie :

Variables	F1	F2
percent_prot_ani	0.441053	-0.071258
food_avl_kcal_p_d	0.451516	0.194122
prot_avl_p_d	0.472158	0.213361
pop_growth_rate	-0.342048	0.649058
PIB_hab	0.386436	0.547298
ratio_prod	0.335176	-0.436926

Noms des axes F1 et F2

Axe F1 : quantité de ressources alimentaires

Axe F2 : Taux d'évolution démographique

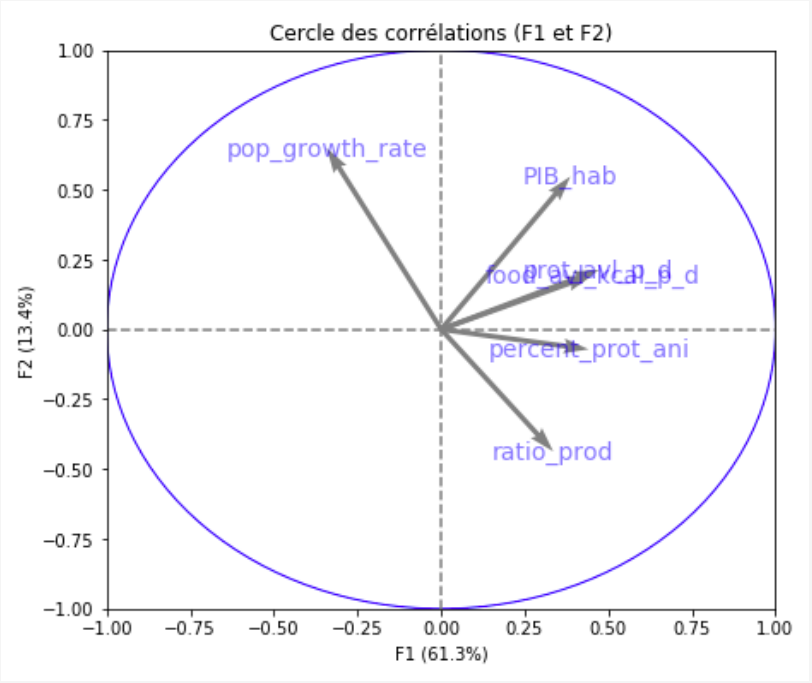
Lien entre le PIB et le taux d'évolution démographique ?

Pays qui ont un PIB habitant élevé, sont des petits pays. Le % d'évolution démographique des pays avec une population faible est sensible.

Exemple extrême de la variation d'une unité :

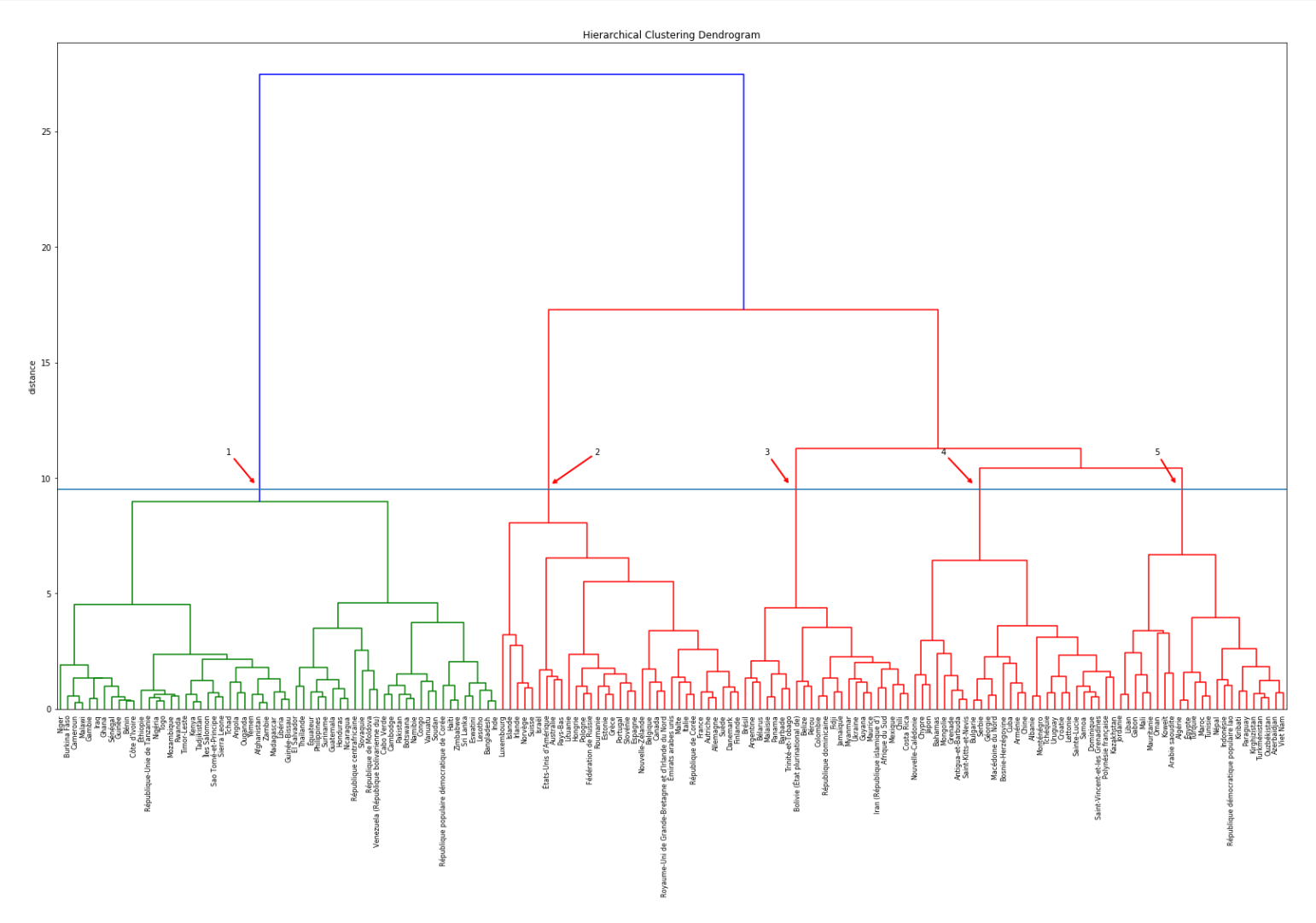
1 à 2, évolution de 100%

100 à 101, évolution de 1%



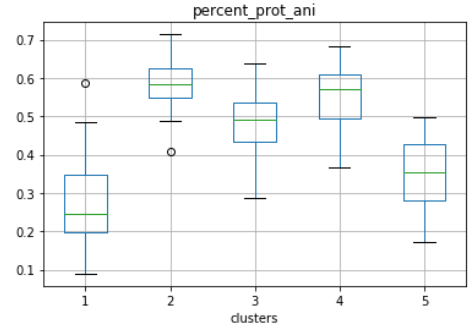
Définition et analyse des clusters

Dendrogramme

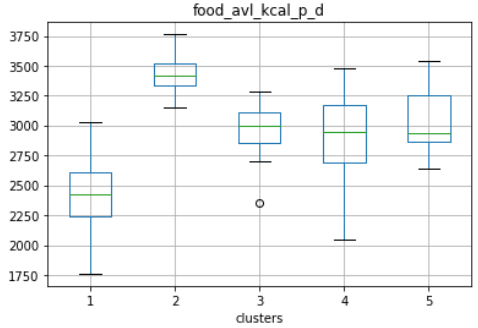


Vue d’ensemble des clusters : distribution sur chaque variable

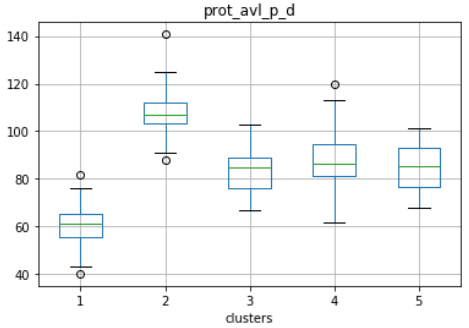
% de protéines animales par groupe



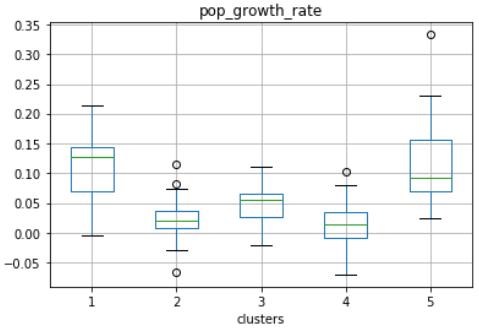
kcal de nourriture par jour et par habitant



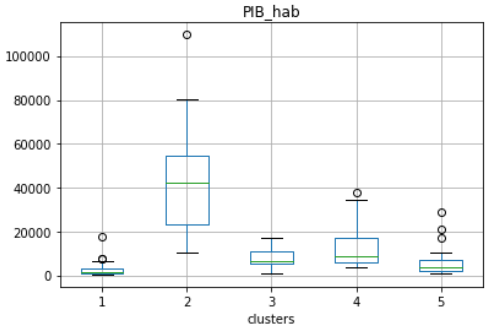
Protéines disponible en gramme par jour et par habitant



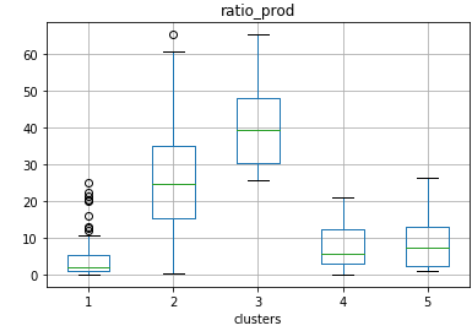
Taux de croissance de la population entre 2012 et 2017



PIB par habitant



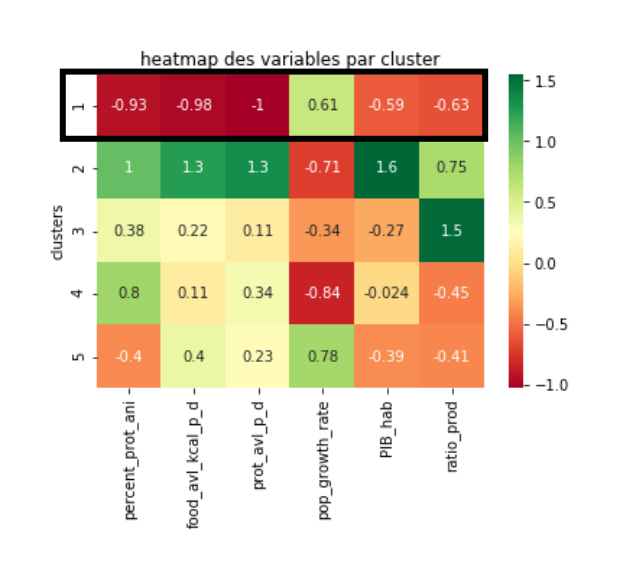
Ratio de production de volailles par habitant



Caractérisation : cluster 1

Liste de quelques pays :

country	percent_prot_ani	food_avl_kcal_p_d	prot_avl_p_d	pop_growth_rate	PIB_hab	ratio_prod	clusters
Afghanistan	0.195045	2000.0	54.09	0.164779	605.557362	0.758153	1
Angola	0.304565	2270.0	54.11	0.187544	4100.291004	0.901204	1
Bangladesh	0.198375	2596.0	60.29	0.057479	1491.673410	1.217569	1
Botswana	0.405799	2340.0	65.18	0.081157	7595.147598	2.086092	1
Îles Salomon	0.295565	2411.0	53.66	0.143825	1961.155429	0.462236	1
Cameroun	0.162131	2653.0	69.45	0.143392	1455.460927	3.112870	1
Cabo Verde	0.370637	2515.0	69.34	0.063857	3239.065521	1.838146	1



- Pourcentage de protéines animales : faible
- Quantité de nourriture disponible kcal p/j/h : faible
- Quantité de protéines disponible g p/j/h : faible
- Taux d'évolution démographique : élevé
- PIB par habitant : faible
- Ratio de production : faible

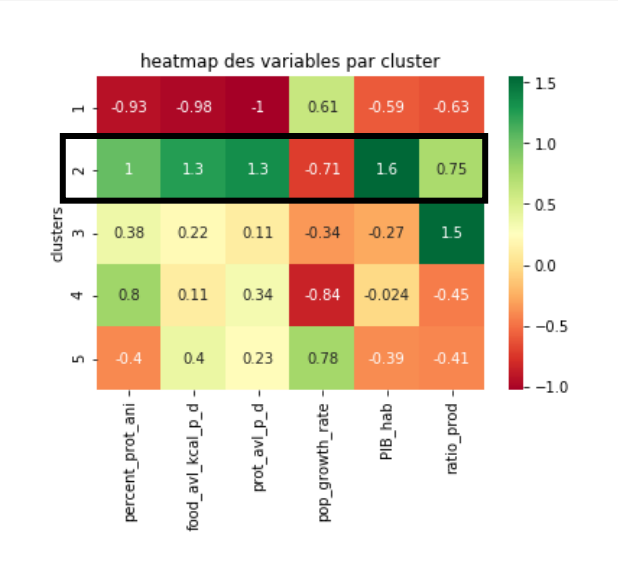
Résumé du groupe :

Ce sont des pays pauvres, avec peu de nourriture et une population qui croit rapidement

Caractérisation : cluster 2

Liste de quelques pays :

country	percent_prot_ani	food_avl_kcal_p_d	prot_avl_p_d	pop_growth_rate	PIB_hab	ratio_prod	clusters
Australie	0.659574	3311.0	108.10	0.073379	57917.093685	50.011145	2
Autriche	0.600018	3693.0	108.13	0.037364	47887.174380	14.421931	2
Canada	0.503656	3492.0	101.20	0.051832	45057.287226	33.682342	2
Danemark	0.647824	3384.0	113.04	0.021631	57454.289221	26.466285	2
Estonie	0.572000	3245.0	107.50	-0.002848	20508.891852	15.461691	2
Finlande	0.624384	3337.0	117.78	0.017840	46180.408610	21.918321	2
France	0.621643	3558.0	112.09	0.020110	38566.566085	17.916410	2



- Pourcentage de protéines animales : élevé
- Quantité de nourriture disponible kcal p/j/h : très élevé
- Quantité de protéines disponible g p/j/h : très élevé
- Taux d'évolution démographique : faible
- PIB par habitant : très élevé
- Ratio de production : élevé

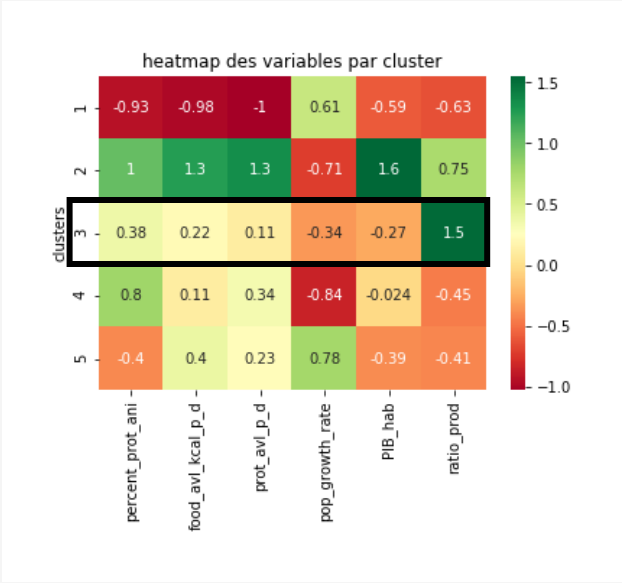
Résumé du groupe :

Ce sont des pays riches, avec beaucoup de nourriture et une population qui croit lentement

Caractérisation : cluster 3

Liste de quelques pays :

country	percent_prot_ani	food_avl_kcal_p_d	prot_avl_p_d	pop_growth_rate	PIB_hab	ratio_prod	clusters
Argentine	0.639630	3238.0	102.70	0.052256	14517.291248	48.159712	3
Barbade	0.564370	2896.0	89.25	0.008921	17423.251128	55.217446	3
Bolivie (État plurinational de)	0.453762	2353.0	68.45	0.078551	3393.959523	45.037035	3
Brésil	0.581114	3249.0	90.86	0.042885	9812.310980	65.472269	3
Belize	0.378833	2699.0	67.18	0.111743	4901.752333	52.282120	3
Myanmar	0.478522	2700.0	93.35	0.038294	1245.826398	28.099084	3
Chili	0.513712	3011.0	88.61	0.061498	15383.587714	33.898869	3



- Pourcentage de protéines animales : moyen
- Quantité de nourriture disponible kcal p/j/h : moyen
- Quantité de protéines disponible g p/j/h : moyen
- Taux d'évolution démographique : faible
- PIB par habitant : faible
- Ratio de production : très élevé

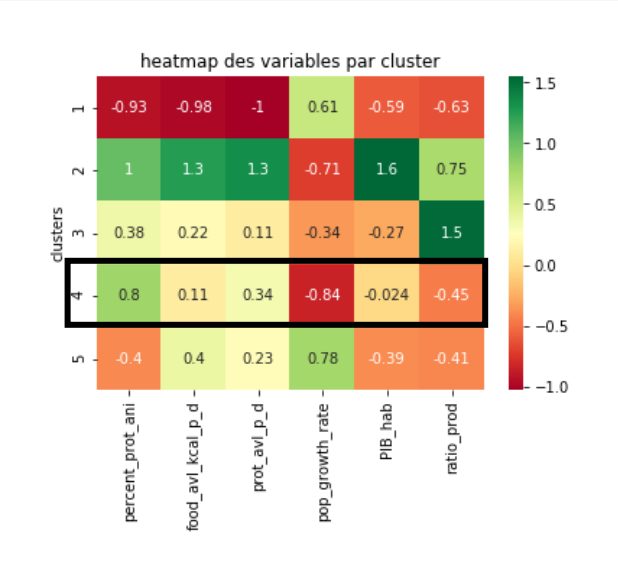
Résumé du groupe :

Ce sont des pays pauvres, avec des quantités de nourriture dans la moyenne, une population qui croit lentement et qui produisent beaucoup de volailles par rapport à leur population

Caractérisation : cluster 4

Liste de quelques pays :

country	percent_prot_ani	food_avl_kcal_p_d	prot_avl_p_d	pop_growth_rate	PIB_hab	ratio_prod	clusters
Arménie	0.457692	3078.0	97.38	0.020996	3933.682101	3.599576	4
Albanie	0.554914	3399.0	119.55	-0.010270	4445.132198	4.476853	4
Antigua-et-Barbuda	0.663720	2430.0	81.45	0.055492	14390.246028	1.006015	4
Bahamas	0.683423	2043.0	61.47	0.049977	30732.419232	18.412333	4
Bulgarie	0.499760	2828.0	83.34	-0.031573	8321.113046	12.253388	4
Cuba	0.368385	3410.0	88.44	0.007298	8433.092699	2.579535	4
Chypre	0.568635	2616.0	80.28	0.039322	26473.007605	21.276993	4



- Pourcentage de protéines animales : élevé
- Quantité de nourriture disponible kcal p/j/h : moyen
- Quantité de protéines disponible g p/j/h : moyen
- Taux d'évolution démographique : faible
- PIB par habitant : moyen
- Ratio de production : faible

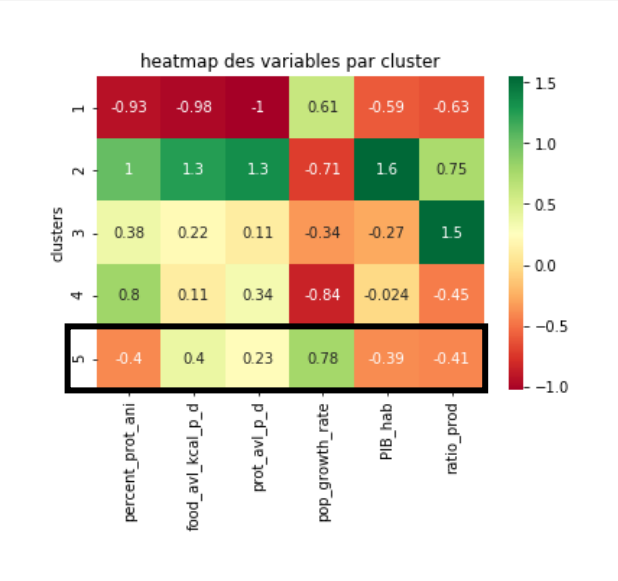
Résumé du groupe :

Ce sont des pays qui ont des quantités de nourriture dans la moyenne. Une part importante de leur protéines est de source animale. Ils ont un pouvoir d'achat modéré.

Caractérisation : cluster 5

Liste de quelques pays :

country	percent_prot_ani	food_avl_kcal_p_d	prot_avl_p_d	pop_growth_rate	PIB_hab	ratio_prod	clusters
Algérie	0.276690	3349.0	92.92	0.107140	4051.244377	6.668746	5
Azerbaïdjan	0.326042	3103.0	92.35	0.062671	4158.357869	10.600163	5
Égypte	0.249974	3321.0	96.33	0.115946	2000.297246	10.821101	5
Gabon	0.486758	2643.0	79.67	0.180113	7372.081131	1.888782	5
Kiribati	0.499256	3057.0	73.91	0.073216	1594.344097	7.507139	5
Indonésie	0.323281	2892.0	68.64	0.065201	3846.426581	8.532895	5
Jordanie	0.349154	2714.0	69.11	0.209628	4195.802289	19.669639	5



- Pourcentage de protéines animales : faible
- Quantité de nourriture disponible kcal p/j/h : moyen
- Quantité de protéines disponible g p/j/h : moyen
- Taux d'évolution démographique : élevé
- PIB par habitant : faible
- Ratio de production : faible

Résumé du groupe :

Ce sont des pays qui disposent de ressources alimentaires modérées mais ils mangent peu de viande. Leurs habitants sont pauvres et la population croit rapidement

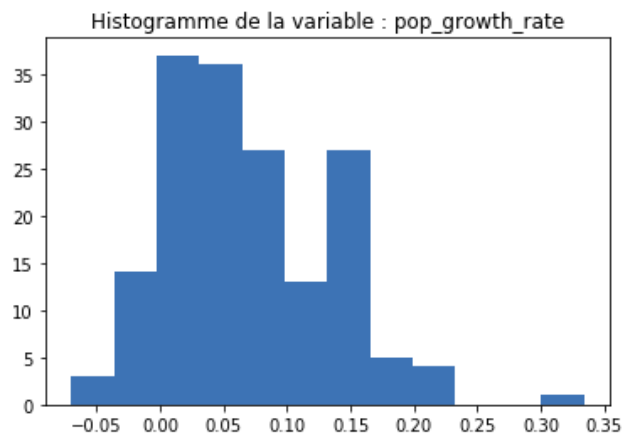
Tests statistiques

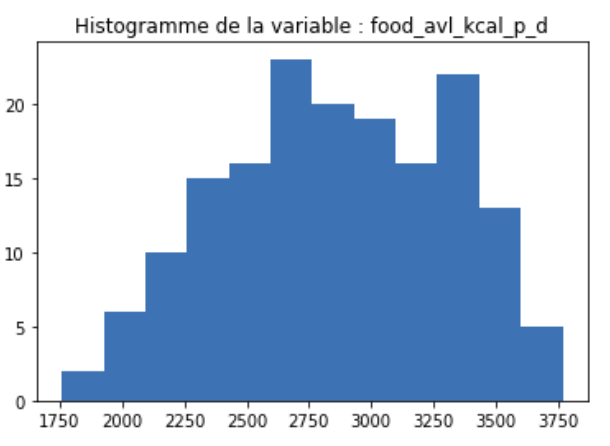
Test d’adéquation des variables à une loi normale

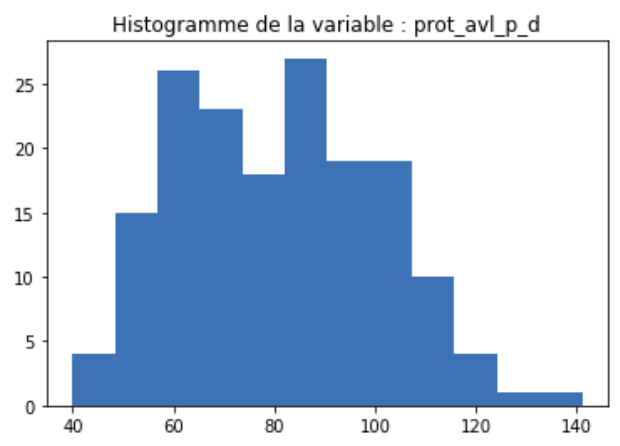
Pour un échantillon dont la loi de probabilité continue :

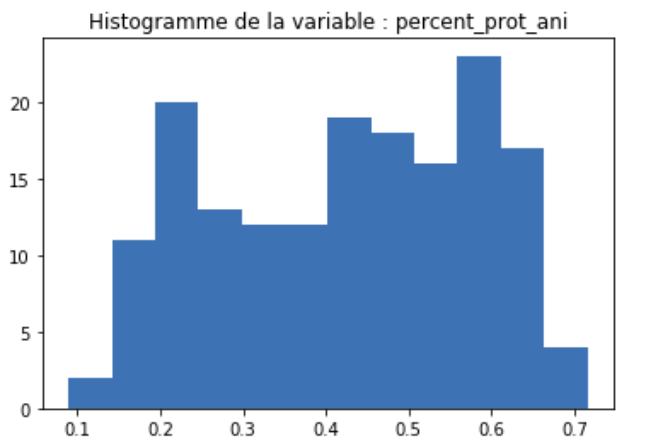
- Lorsque l’on souhaite tester l’adéquation d’un échantillon à une loi normale, le test le plus adapté est celui de Shapiro-Wilk
- Il existe également le test de Kolmogorov-Smirnov qui permet de tester n’importe quelle loi. Cependant il est moins puissant que le test de Shapiro-Wilk
- Seuil alpha = 0,05

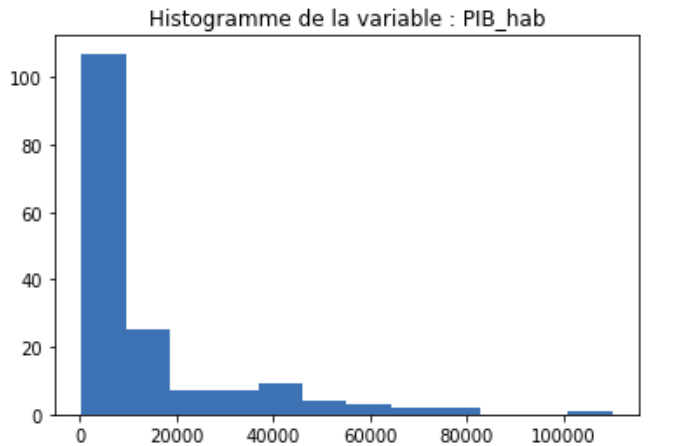
Aperçu des histogrammes par variable

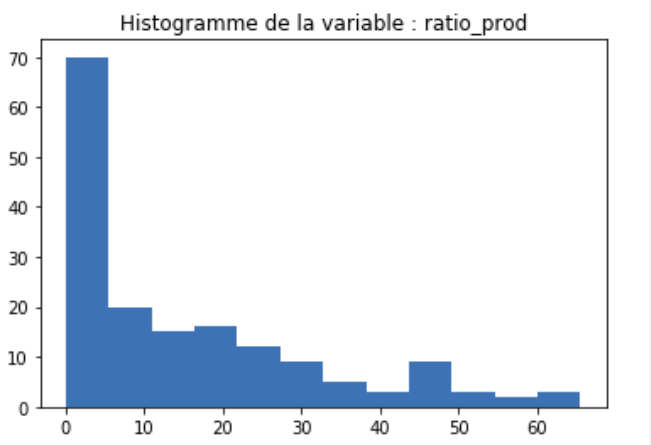




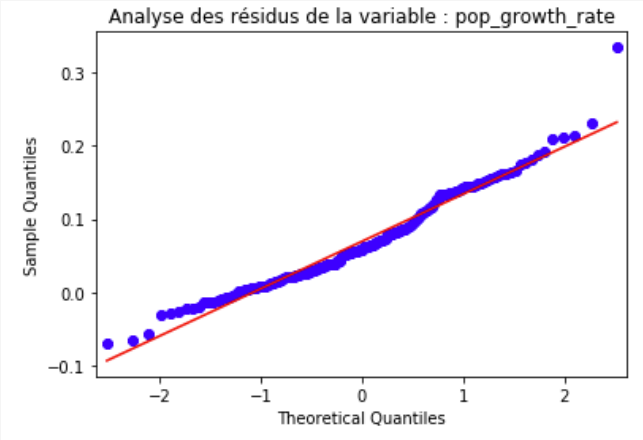


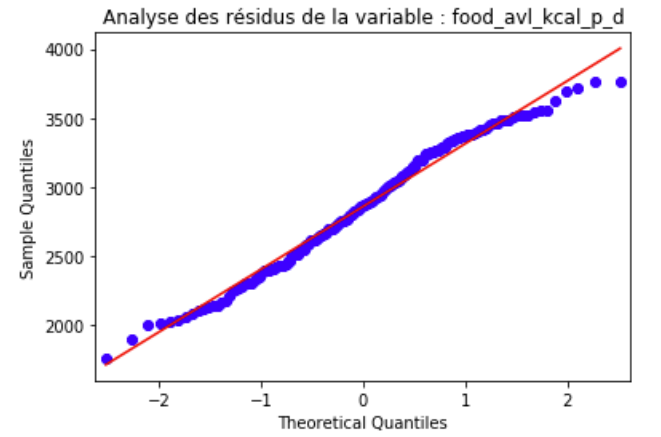


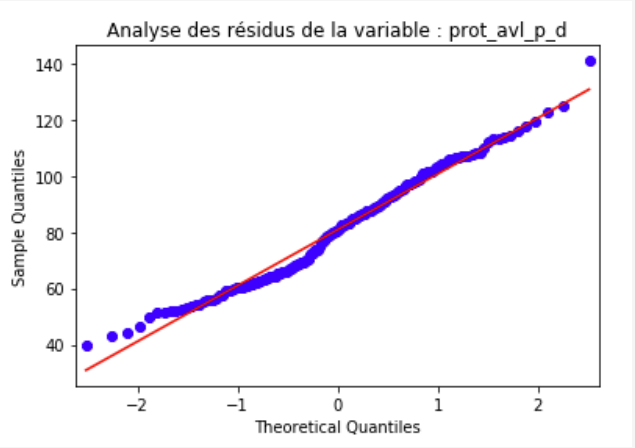


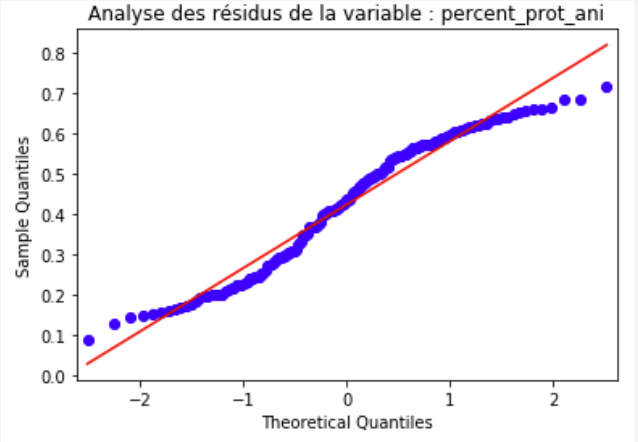


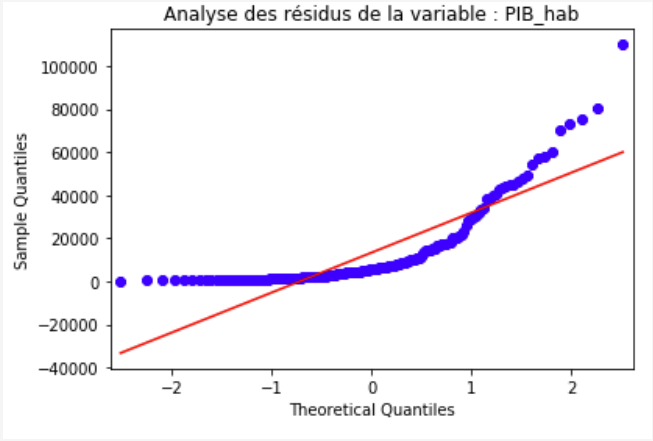
Aperçu des qqplot par variable

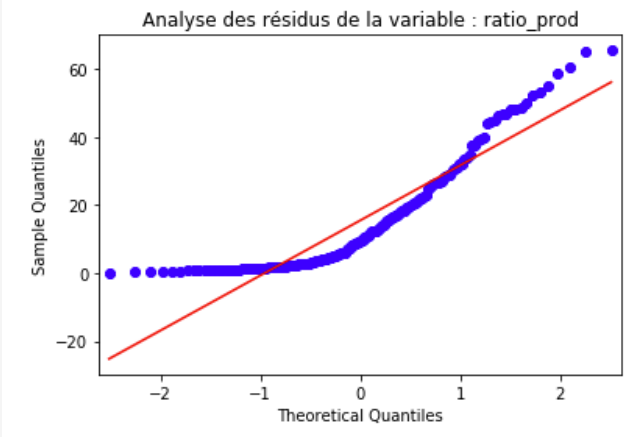




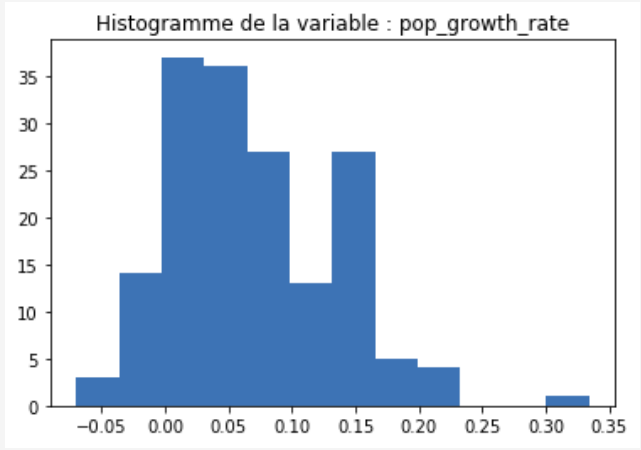


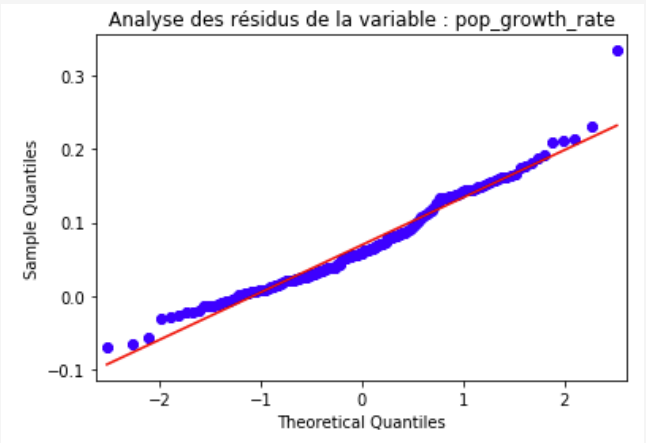




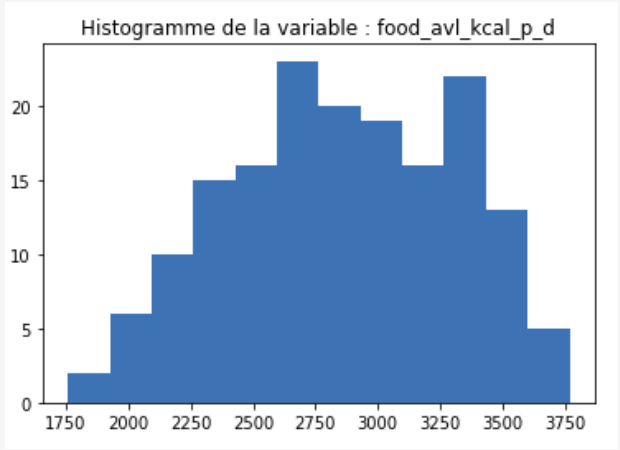


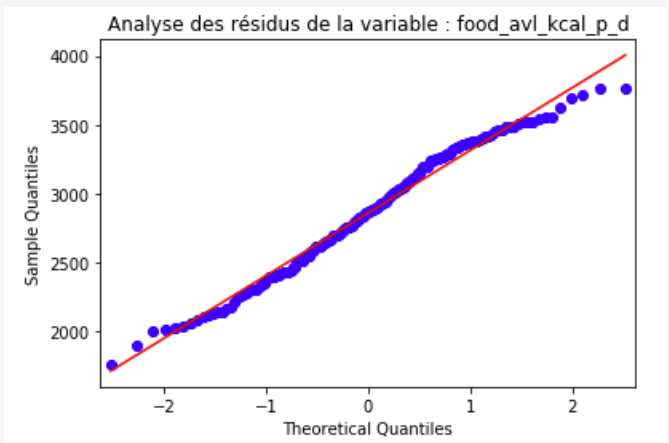
Aperçu des histogrammes par variable



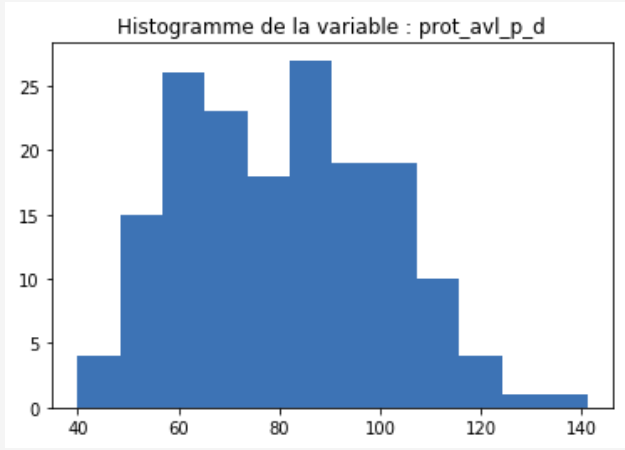


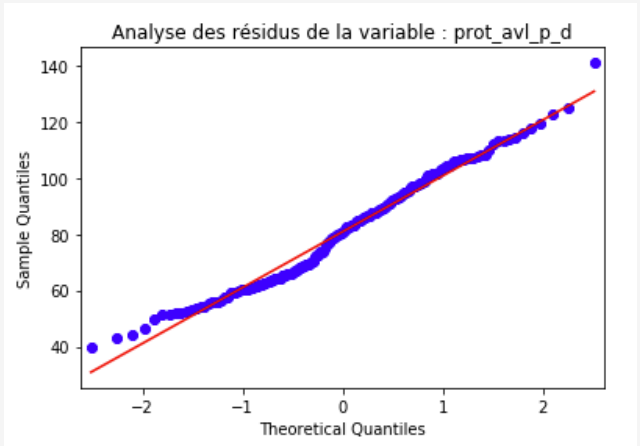
- **Test de Shapiro :**
- Valeur = 0.9676
- P-value =6,1 * 10^-4





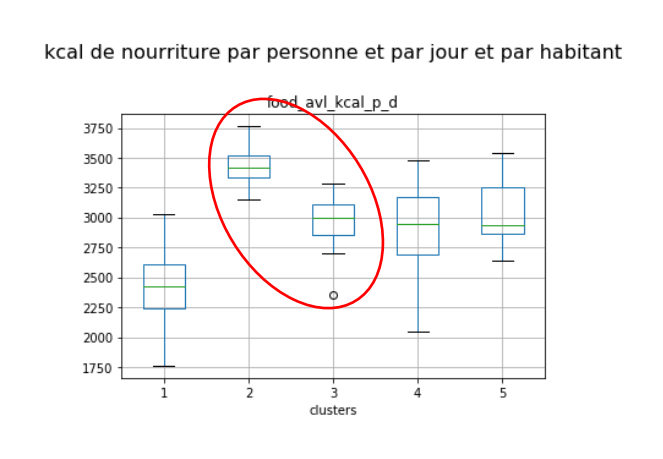
- **Test de Shapiro :**
- Valeur = 0.9810
- P-value =2,18 * 10^-2





- **Test de Shapiro :**
- Valeur = 0.9789
- P-value =1,21 * 10^-2

Test de comparaison



Comparaison des variances :

- $H_0 : \sigma_1^2 = \sigma_2^2$
- $H_1 : \sigma_1^2 \neq \sigma_2^2$

Test de Bartlett :

- Statistique = 2,543
- P-valeur = 0,11

Comparaison des moyennes :

- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 \neq \mu_2$

Test de Student :

- Statistique = 9,281
- P-valeur = 8,95 * 10^-13

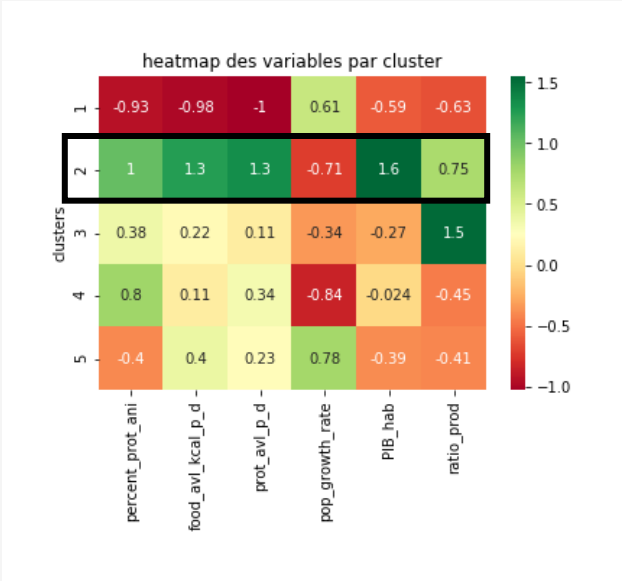
Je rejette donc l'hypothèse que les échantillons des clusters 2 et 3 sur la variable « *food_avl_kcal_p_d* » suivent la même distribution, car j'ai rejeté l'hypothèse d'égalité des moyennes

Sélection d’un cluster et recommandations

Rappel des caractéristiques du cluster 2

clusters	percent_prot_ani	food_avl_kcal_p_d	prot_avl_p_d	pop_growth_rate	PIB_hab	ratio_prod
1	0.277152	2412.750000	60.365333	0.108944	2407.636014	5.188151
2	0.585148	3433.151515	107.667576	0.023498	42241.754394	27.574467
3	0.482432	2960.130435	83.049565	0.047588	8495.746651	40.377523
4	0.548620	2910.714286	87.598214	0.015136	12963.225831	8.100017
5	0.360445	3042.217391	85.512609	0.119809	6240.039620	8.882178
mean	0.423273	2859.958084	80.866048	0.069377	13415.146487	15.455210

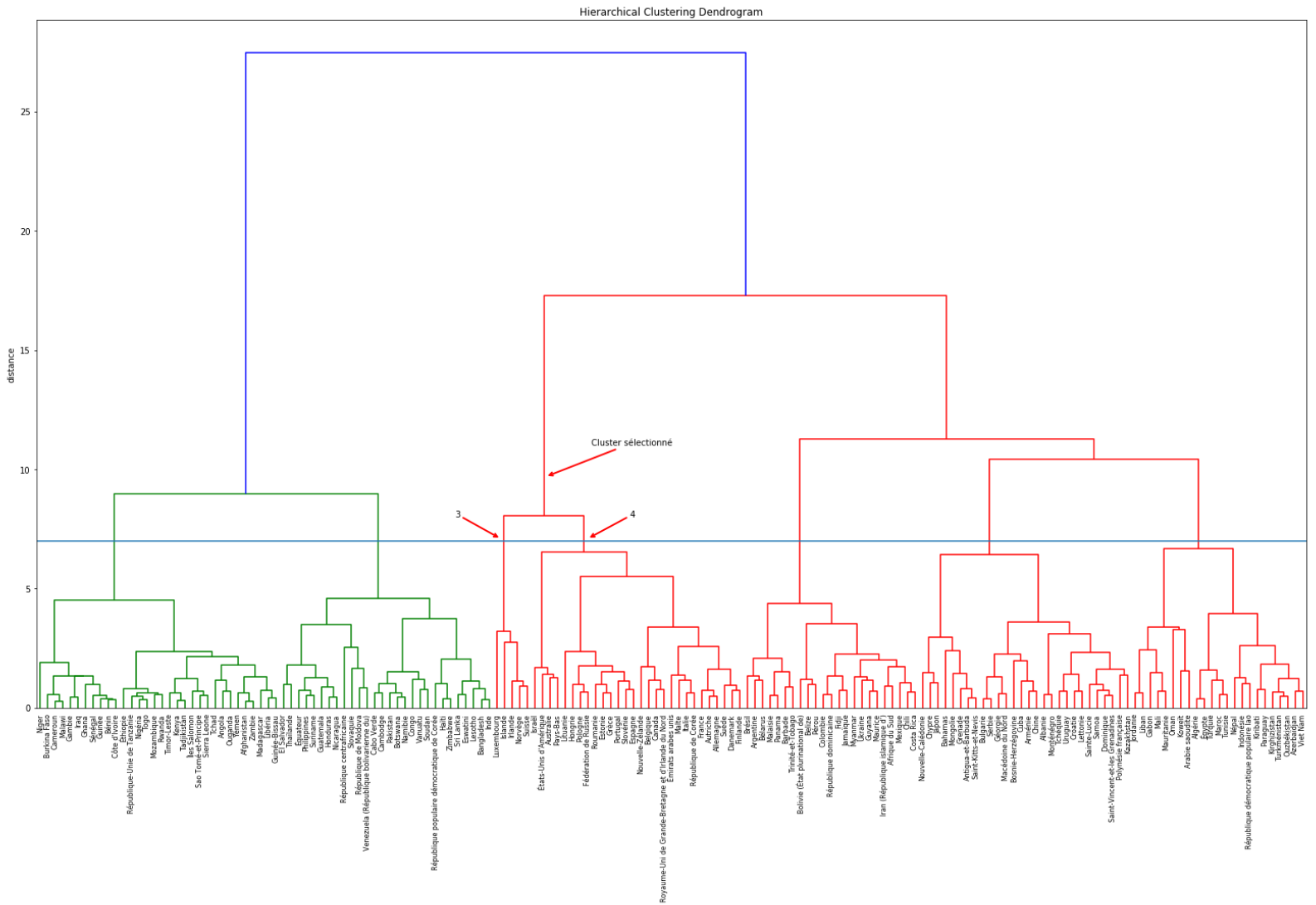
Caractéristiques du pays
Pouvoir d'achat élevé de la population
Disponibilités alimentaires importantes
Faible production de volaille
Forte habitude de consommation de viandes



Résumé du groupe :

Ce sont des pays riches, avec beaucoup de nourriture et une population qui croit lentement

Dendrogramme découpé en 7 clusters

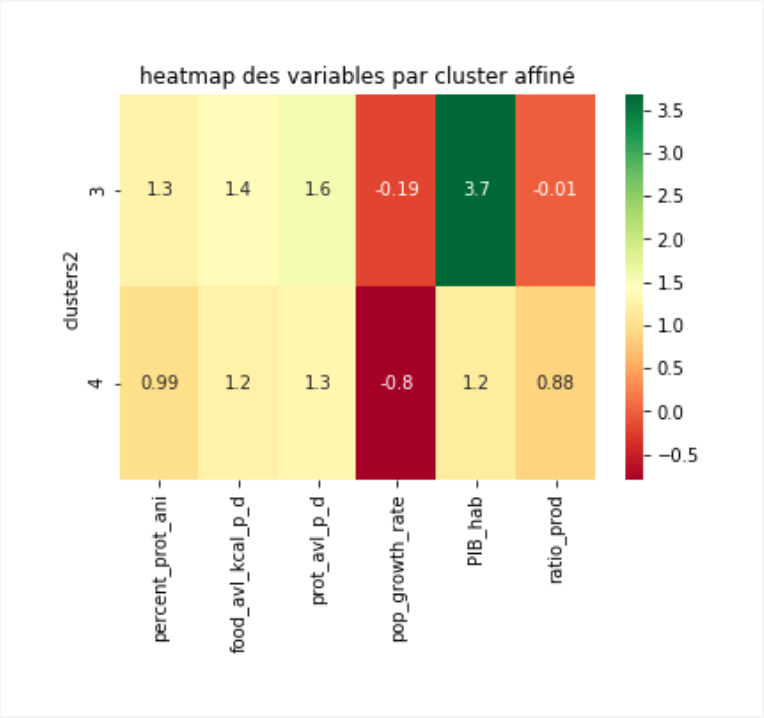


Pourquoi 7 clusters ?

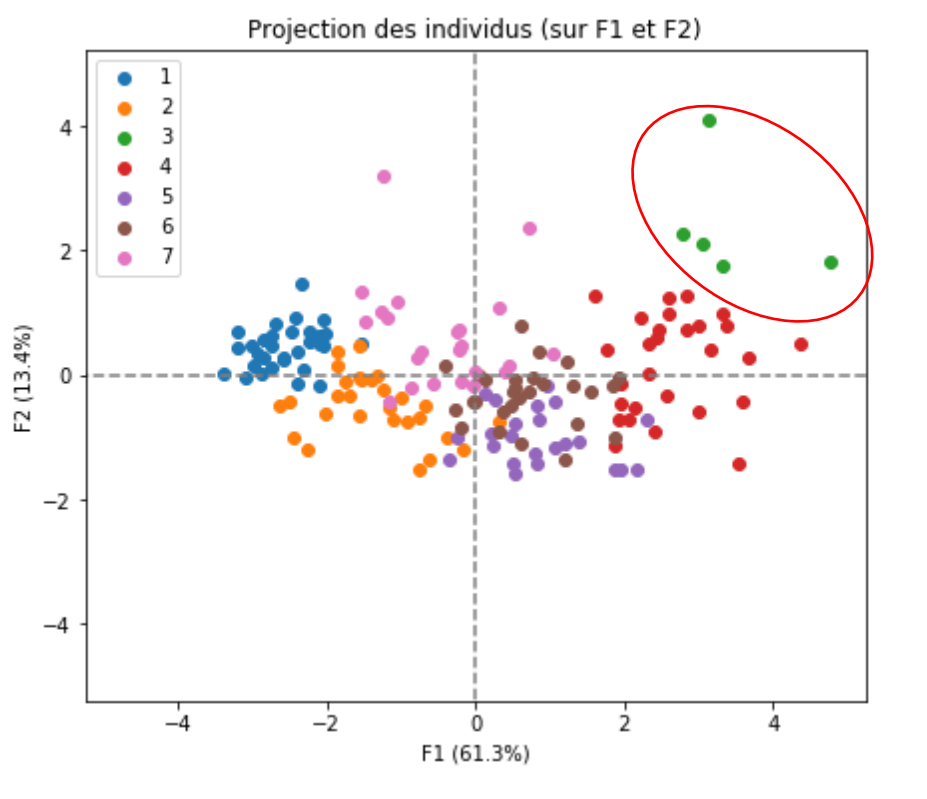
Résultat d'un processus itératif où je sais qu'en découpant ainsi mon dendrogramme, je pourrais isoler un groupe qui m'intéresse

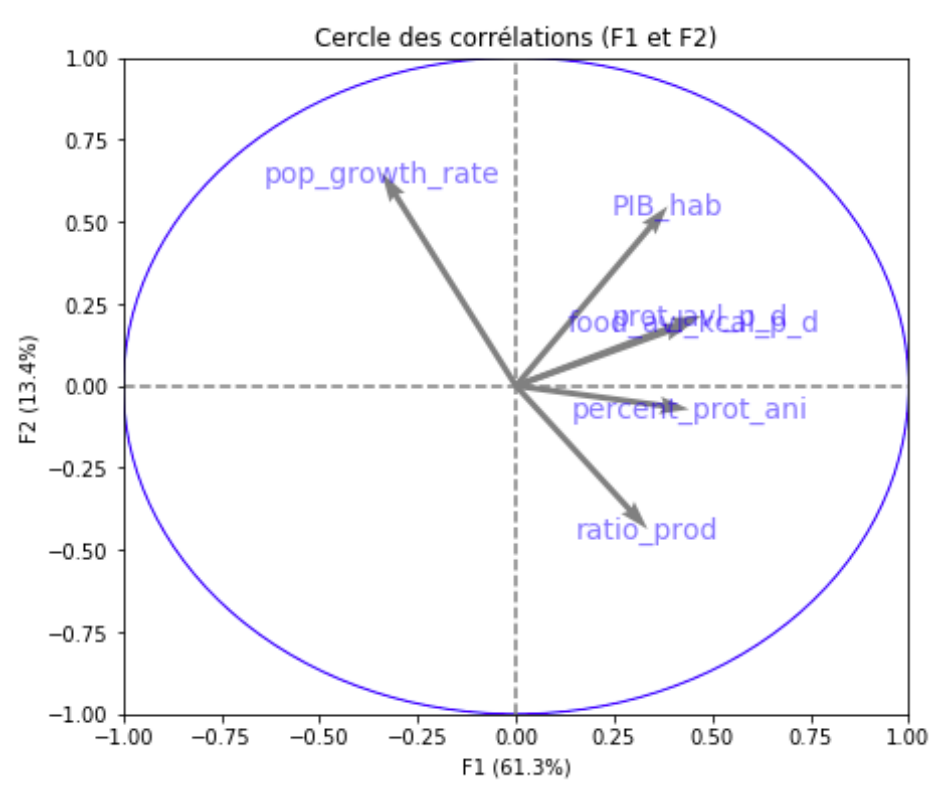
Focus sur les deux branches du secteur sélectionné

clusters2	percent_prot_ani	food_avl_kcal_p_d	prot_avl_p_d	pop_growth_rate	PIB_hab	ratio_prod
3	0.623852	3495.6	111.750000	0.057124	81783.457453	15.287540
4	0.578237	3422.0	106.938571	0.017493	35180.735991	29.768561



Représentation du cluster retenu sur le premier plan factoriel





Liste des pays ciblés

country	percent_prot_ani	food_avl_kcal_p_d	prot_avl_p_d	pop_growth_rate	PIB_hab	ratio_prod	clusters2	pop_1000_2017
Islande	0.715694	3627.0	141.01	0.026842	73108.083207	28.998813	3	334.393
Irlande	0.570458	3717.0	105.95	0.031483	70492.921726	19.155829	3	4753.279
Norvège	0.585603	3385.0	110.16	0.056369	75092.402428	17.171148	3	5296.326
Suisse	0.612053	3414.0	95.58	0.055918	80220.632483	10.655758	3	8455.804
Luxembourg	0.635455	3335.0	106.05	0.115008	110003.247419	0.456150	3	591.910

Total du nombre d'habitants : 19 431 712

Petits pays = bon marché-témoin

Avantages :

- En fonction des ventes obtenus l'entreprise pourra décider d'élargir le marché
- Permet de prévoir les volumes de ventes

Conclusion

5 pays cibles :

- Luxembourg
- Suisse
- Irlande
- Norvège
- Islande

Les pays sont relativement proches de la France :

- Deux sont limitrophes

Limites de l’analyse : le cas de l’Irlande

- L’Irlande se retrouve dans le classement des pays avec les PIB par habitants les plus élevé.
- Cependant cela résulte des conséquences d’un **dumping fiscal agressif**.
- L’Irlande attire des grandes sociétés grâce à des taux de taxe faible par rapport au reste de l’Europe.
- Ces grandes sociétés qui basent leur siège en Irlande font gonfler comptablement les indicateurs macro-économique de cette dernière mais on peut s’interroger quant à leur influence sur l’économie réelle.