OPENCLASSROOMS

Produisez une étude de marché

Mentor : Claire Della Nova

Agenda

Construction de l’échantillon
- Valeurs imposées
- Valeurs complémentaires

Classification hiérarchique
- Préambule : analyses bivariées
- Dendrogramme
- Interprétation

Analyse en composantes principales
- Eboulis des valeurs propres, variance expliquée et variance cumulée
- Cercle des corrélations
- Représentation des individus
- Analyse et conclusions

Tests statistiques
- Test de Shapiro et représentations graphiques loi normale
- Tests de comparaison pour 'edb_2019’ sur clusters 1 et 5

Conclusion
- Pour aller plus loin

Construction de l’échantillon

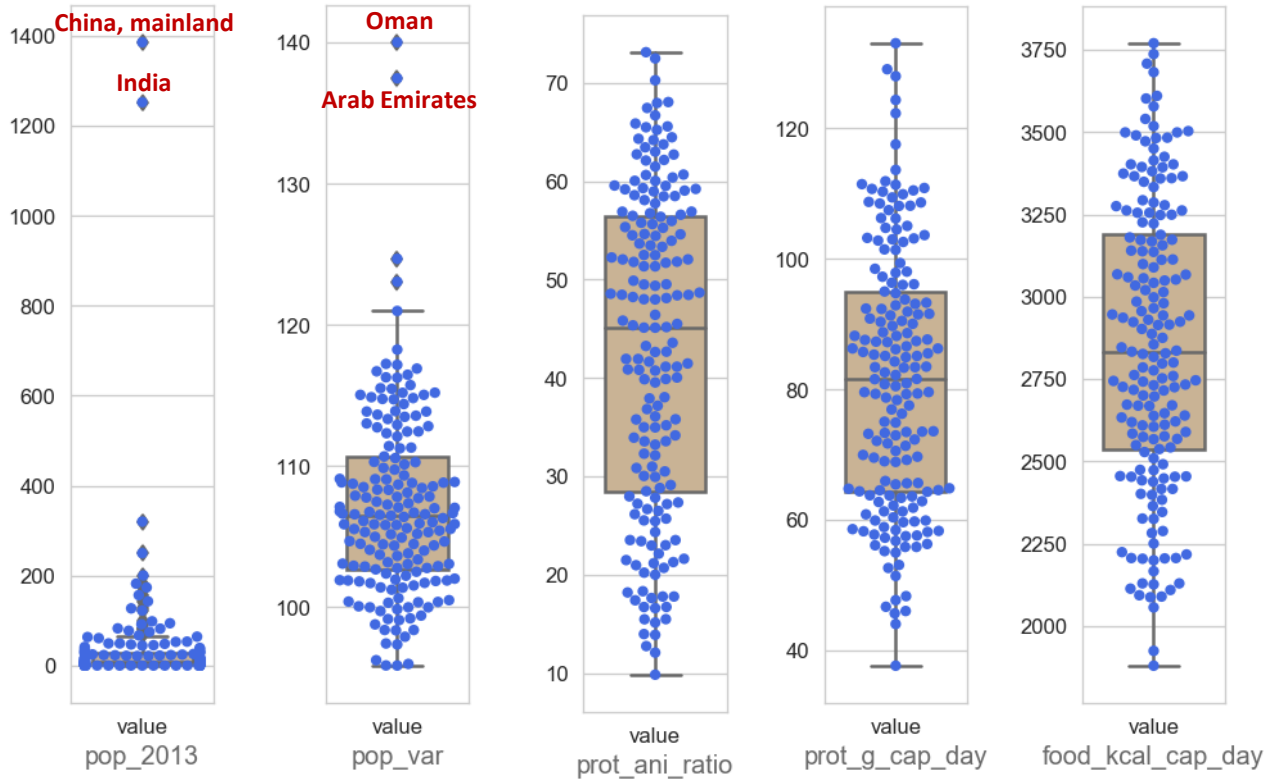
Nettoyage, préparation, investigation préléminaire

Construction de l'échantillon : valeurs imposées

Pour réaliser cette étude de marché, le commanditaire nous a fourni un set de données extrait des bilans alimentaires de la FAO (année 2013).

- **Pop_2013 :** population mondiale 2013
- **Pop_var :** évolution de la population 2008-2013
- **Prot_ani_ratio :** pourcentage de protéine animale dans la ration de protéines totale
- **Prot_g_cap_day :** ration individuelle quotidienne de protéines en grammes
- **Food_kcal_cap_day** : part destinée à la consommation humaine, ration individuelle quotidienne en kcal

Describe	pop 2013	pop var	prot ani ratio	prot (g.cap.day)	food (kcal.cap.day)
count	173.0	173.0	173.0	173.0	173.0
mean	40.2	107.2	42.8	81.4	2,850.3
std	146.0	6.7	16.2	20.0	438.1
min	0.1	95.9	9.8	37.7	1,879.0
25%	2.3	102.7	28.5	64.3	2,537.0
50%	9.4	106.5	45.1	81.6	2,833.0
75%	27.8	110.7	56.5	95.0	3,188.0
max	1,385.6	140.0	73.1	133.1	3,770.0

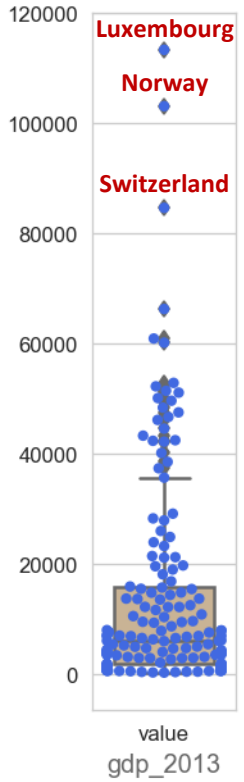


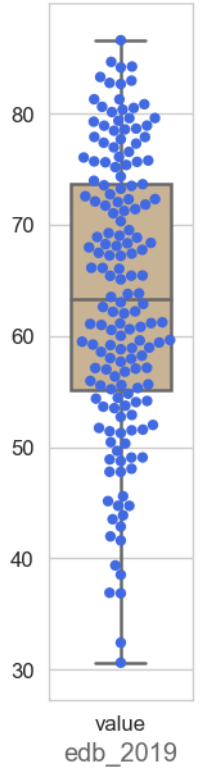
Construction de l'échantillon : valeurs complémentaires

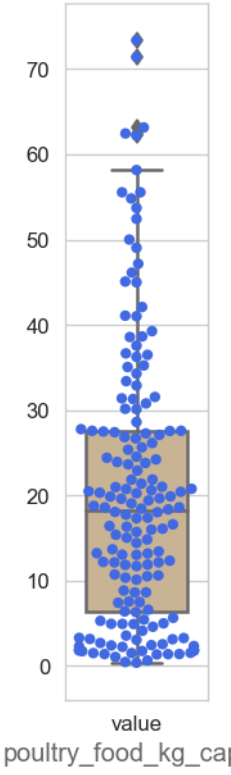
Les variables imposées n'indiquent que des tendances alimentaires, il manque un minimum d'information complémentaire pour réaliser une étude de marché.

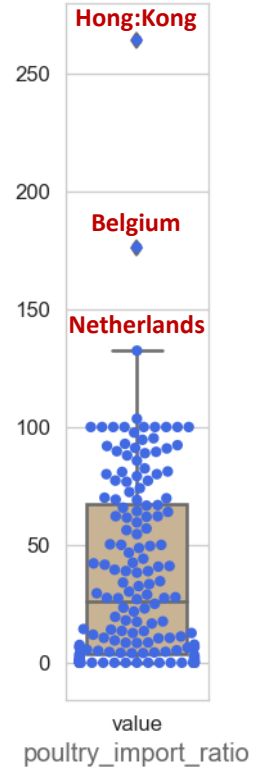
- **Poultry_food_kg_cap_y :** consommation annuelle de volaille par habitant et par an
- **Poultry_import_ratio :** Import Quantity/ Domestic supply quantity * 100
- **GDP_2013 :** PIB 2013 par pays et par habitant
- **Edb2019 :** Ease of doing Business de la Worldbank, indicateur global/ranking de risque pays.

Describe	gdp 2013	edb 2019	poultry food (kg.cap.y)	poultry import ratio
count	164.0	164.0	164.0	164.0
mean	**14,104.7**	**63.7**	**16.4**	**38.1**
std	19,645.9	12.4	16.4	41.1
min	314.7	30.6	0.4	-
25%	1,832.7	55.1	6.4	3.5
50%	**6,093,5**	**63.3**	**18.2**	**26.0**
75%	15,801.9	73.7	27.6	67.4
max	**113,341.2**	86.6	73.4	**264.1**







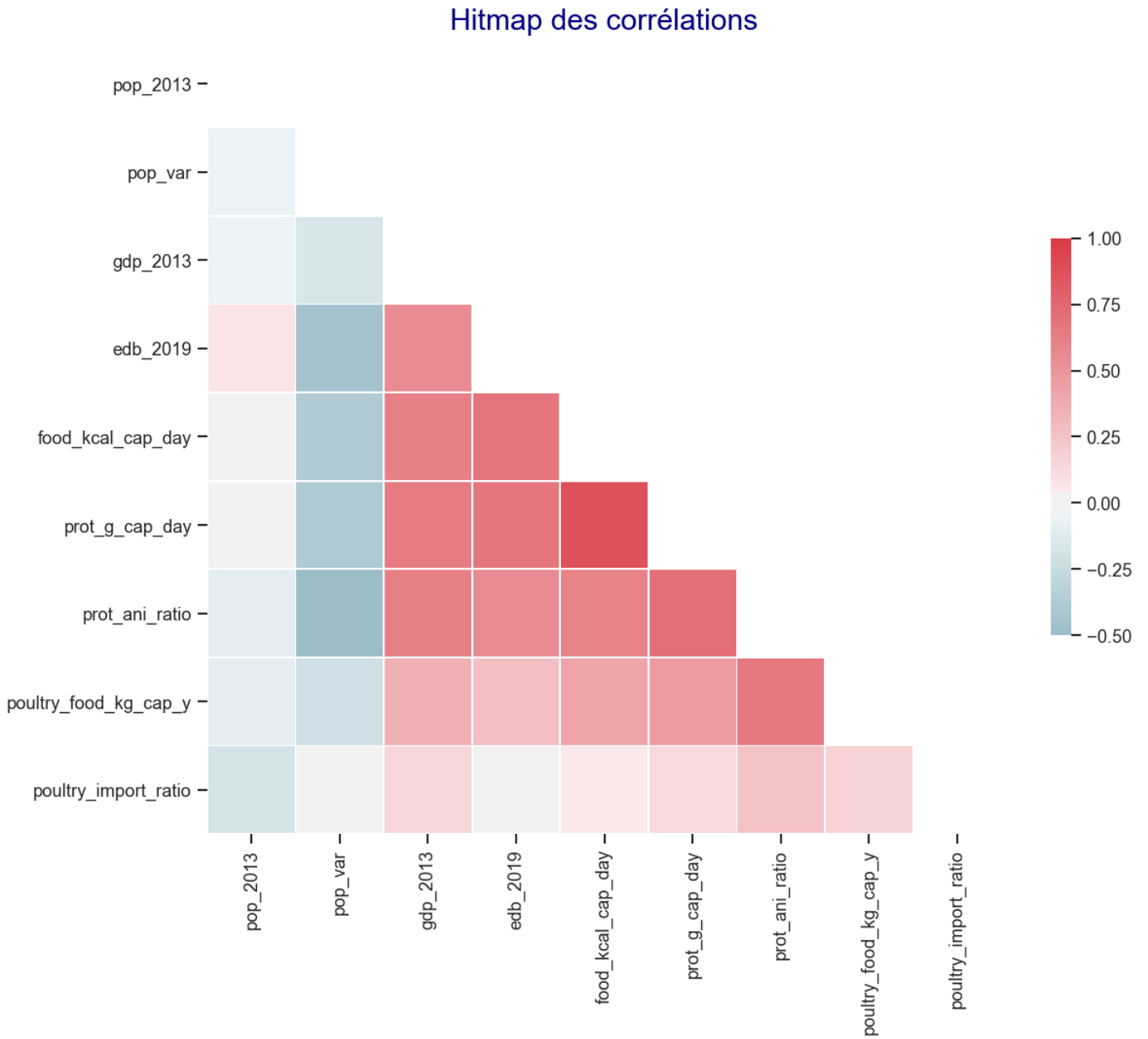


Classification hiérarchique

Analyses bivariées, dendrogramme, interprétation

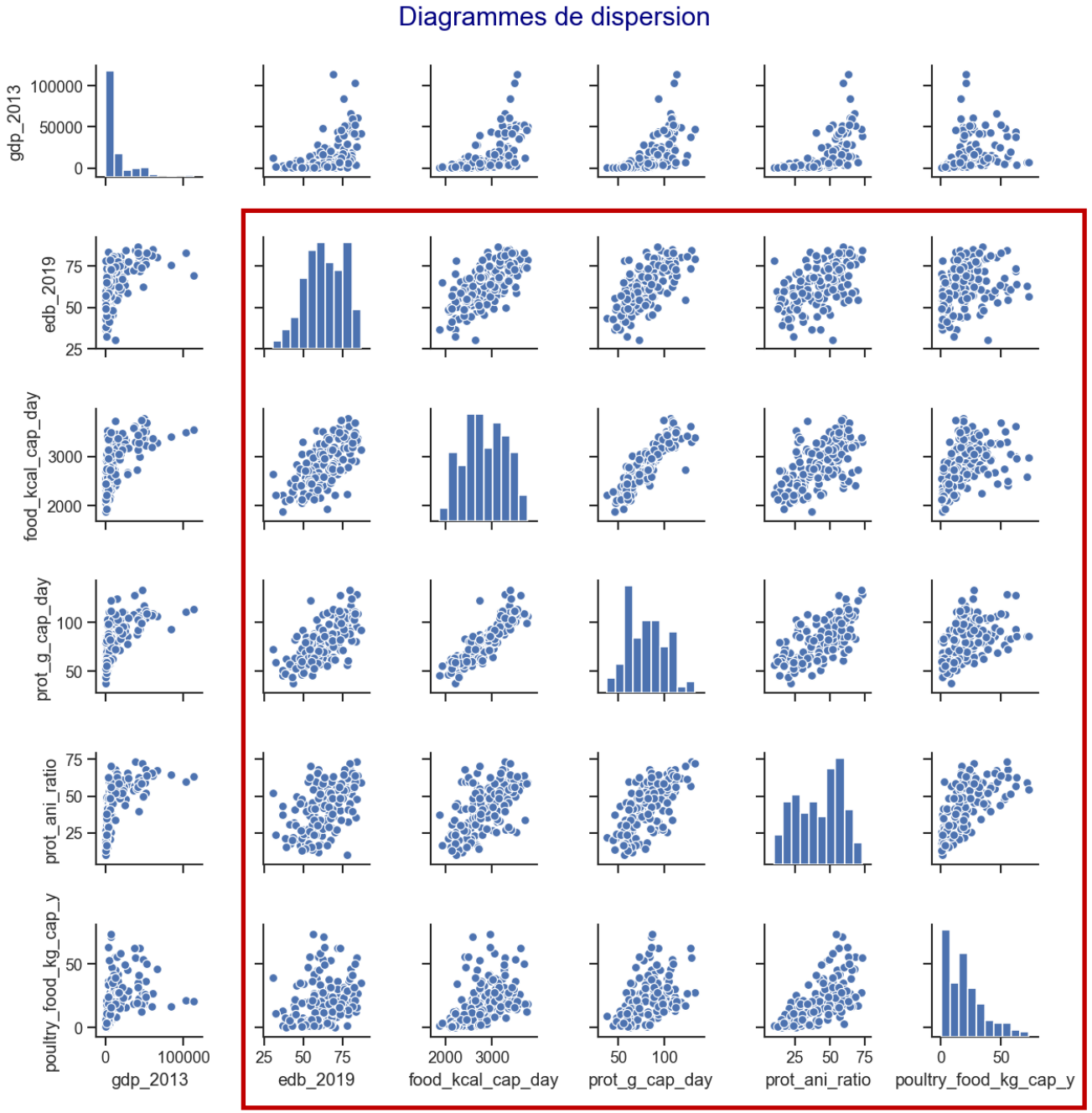
Préambule : analyses bivariées

- **pop_2013** et **poultry_import_ratio** sont proche de zéro, et n'ont donc aucune corrélation significative avec les autres variables
- **pop_var** a des valeurs plutôt négatives et est donc anti-corrélée avec les variables restantes
- Les variables restantes ont des corrélations positives entre elles à des degrés divers
- On notera par exemple **prot_g_cap_day** et **food_kcal_cap_day** sont fortement corrélées, ce qui nest a priori pas une surprise



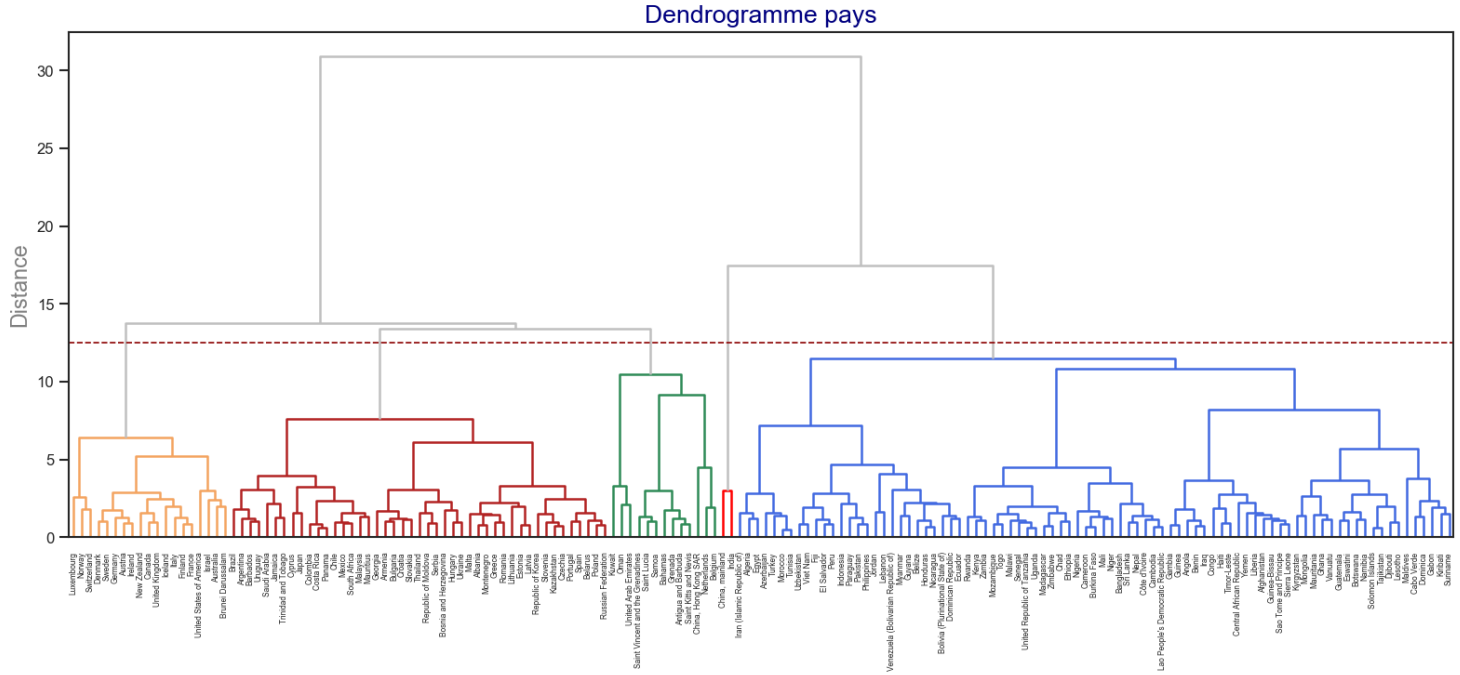
Préambule : analyses bivariées

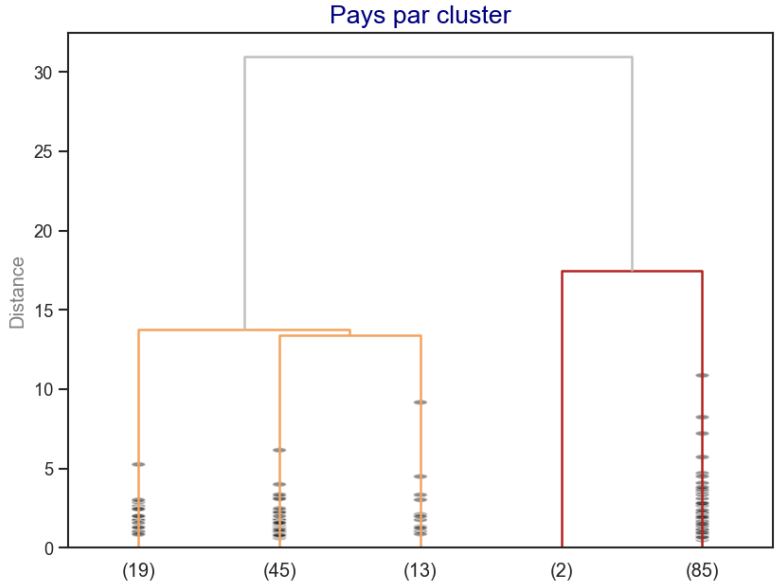
- **Pour une lecture plus commode du diagramme de dispersion, nous suprimons pop_2013, pop_var et poultry_import_ratio qui sont soit non corrélées soit anti-corrélées avec les autres variables**
- Mis à part pour **gdp_2013**, la representation en pairplots des diagrammes de dispersion vient confirmer une corrélation plus ou moins élevée entre ces variables laissant supposer des régressions linéaires plus ou moins fortes



Dendrogramme

- Au regard de ce graphique, la coupe en 5 clusters semble la plus pertinente, malgré le cluster 5 de 85 pays
- La coupe à 6 clusters divisant le cluster 5 en deux ne devrait guère apporter de précision vu leur proximité (solution testée)
- Le cluster 4 n'a que deux pays (China, India). Leur poids est dû à une population très au-dessus de la moyenne





- **Cluster 1 :** 19 pays
- **Cluster 2 :** 45 pays
- **Cluster 3 :** 13 pays
- **Cluster 4 :** 2 pays
- **Cluster 5 :** 85 pays

Interprétation

- **Clust1 (19 pays)** : GDP élevé (dans le 4e quartile), régime calorique et protéique élevé, forte consommation annuelle de volaille
- **Clust2 (45 pays)** : GDP dans la partie supérieure du 3e quartile, rations caloriques et protéiques dans la norme et consommation de volaille dans le 3e quartile
- **Clust3 (13 pays)** : Faible population, GDP dans le top 25%, proportion élevée de protéine animale et de consommation de volaille
- **Clust4 (2 pays)** : Très forte population (Chine, Inde), consommation de volaille très en-dessous de la moyenne et de la médiane, faible GDP, seulement 2 pays
- **Clust5 (85 pays)** : Faible GDP, indicateur Ease of Doing Business faible, faible consommation de volailles

Statistiques descriptives du dataset

Describe	pop 2013	pop var	gdp 2013	edb 2019	food (kcal.cap.day)	prot (g.cap.day)	prot ani ratio	pltry food (kg.cap.y)	pltry imp ratio
count	164.0	164.0	164.0	164.0	164.0	164.0	164.0	164.0	164.0
mean	42.0	107.4	14,104.7	63.7	2,850.4	81.1	42.4	20.3	38.1
std	149.8	6.8	19,645.9	12.4	443.5	20.3	16.2	16.4	41.1
min	0.1	95.9	314.7	30.6	1,879.0	37.7	9.8	0.4	-
25%	2.9	102.9	1,832.7	55.1	2,523.2	63.8	27.9	6.4	3.5
50%	9.5	106.6	6,093.5	63.3	2,821.5	80.9	45.1	18.2	26.0
75%	29.1	111.3	15,801.9	73.7	3,223.0	96.0	56.1	27.6	67.4
max	1,385.6	140.0	113,341.2	86.6	3,770.0	133.1	73.1	73.4	264.1

Centroides des clusters

clust	pop 2013	pop var	gdp 2013	edb 2019	food (kcal.cap.day)	prot (g.cap.day)	prot ani ratio	pltry food (kg.cap.y)	pltry imp ratio
1	37.4	104.7	57,008.9	78.7	3,434.7	108.3	61.0	30.4	28.6
2	27.1	101.6	14,038.2	72.1	3,086.4	90.7	52.2	26.1	30.9
3	4.0	110.9	25,496.8	65.7	2,976.0	93.0	59.5	48.3	115.4
4	1,318.9	104.9	4,233.6	70.4	2,783.0	79.0	30.0	7.5	1.6
5	26.7	110.6	3,039.4	55.3	2,577.3	68.2	30.8	11.1	33.0

Conclusions

- **Les clusters 1 et 2 sont les plus intéressants**
- **Le cluster 3 est à observer**
- **Les clusters 4 et 5 peuvent être exclus**

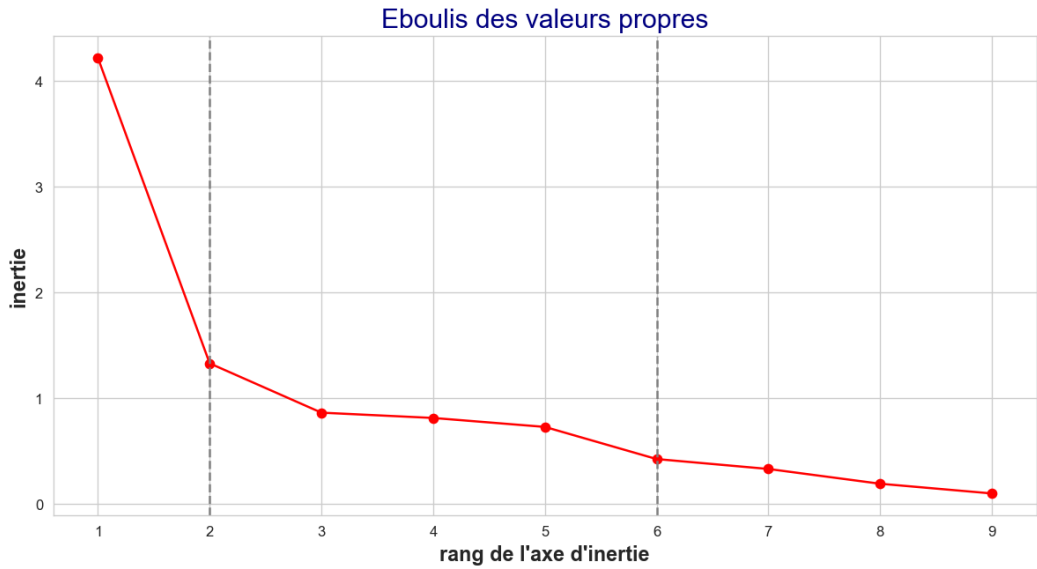
Analyse en composantes principales

Eboulis des valeurs propres, variances, cercle des corrélation, représentation des individus, analyse et conclusions

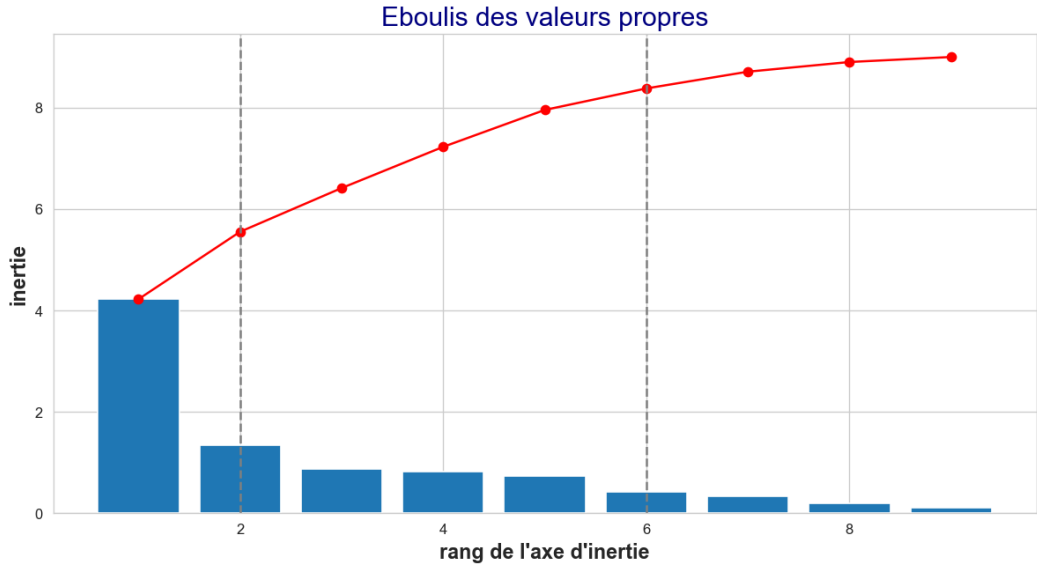
Eboulis des valeurs propres

- Le coude apparaît dès la deuxième composante, mais cela coïncide avec les demandes du commenditaire, nous n'étudierons donc que le premier plan factoriel avec les composantes 1 et 2
- Le premier plan factoriel représente 62 % de la variance cumulée

Describe	F1	F2	F3	F4	F5	F6	F7	F8	F9
count	164	164	164	164	164	164	164	164	164
mean	-	-	-	-	-	-	-	-	-
std	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
min	- 0.3	- 1.7	- 0.7	- 2.7	- 2.2	- 2.1	- 2.0	- 1.2	- 0.9
25%	- 0.3	- 0.7	- 0.6	- 0.7	- 0.7	- 0.9	- 0.9	- 0.9	- 0.8
50%	- 0.2	- 0.1	- 0.4	- 0.0	- 0.1	- 0.0	0.2	- 0.1	- 0.3
75%	- 0.1	0.6	0.1	0.8	0.8	0.7	0.9	0.4	0.7
max	9.0	4.8	5.1	1.9	2.1	2.6	1.9	3.3	5.5



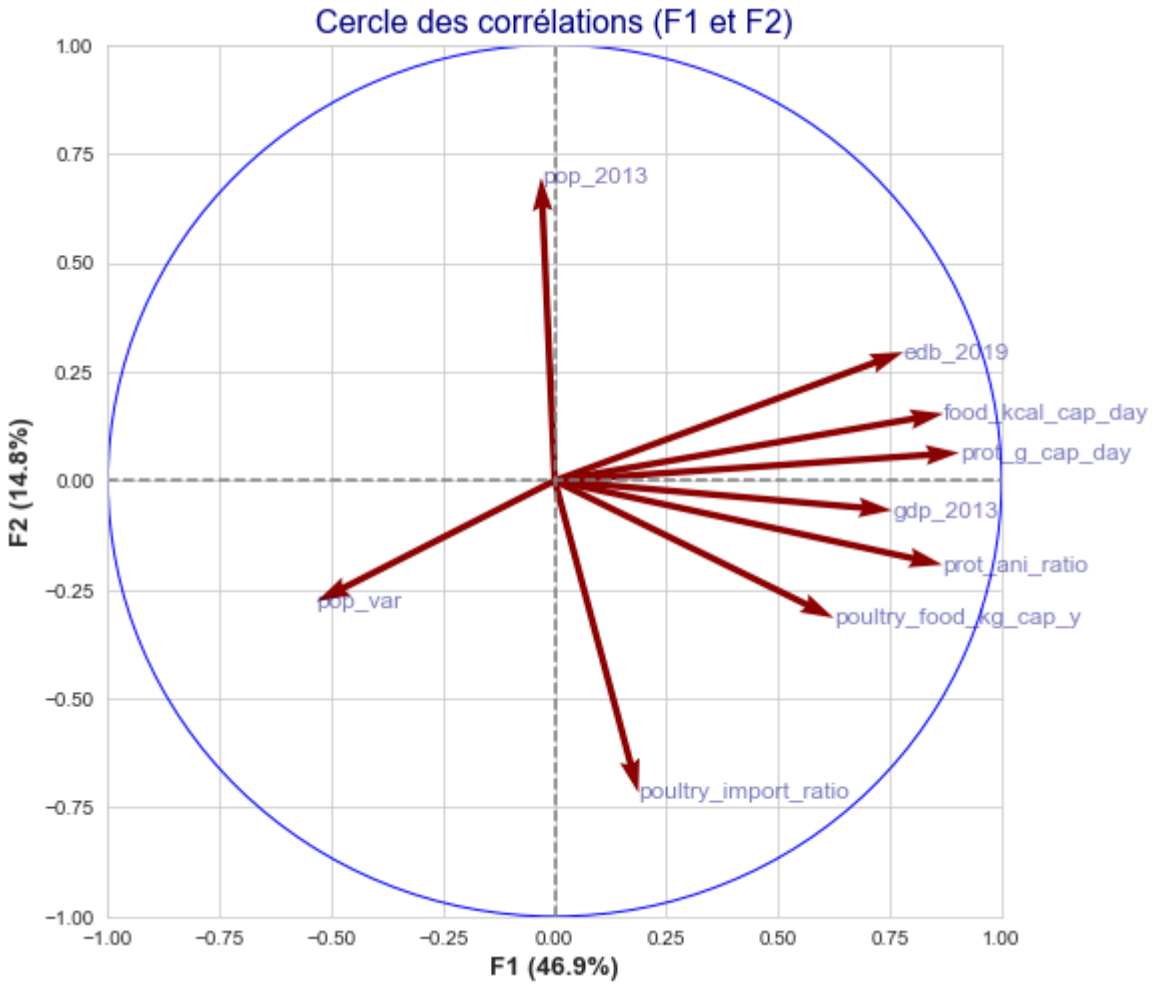
Variance expliquée : [47. 15. 10. 9. 8. 5. 4. 2. 1.]



Variance cumulée : [47. 62. 71. 80. 88. 93. 97. 99. 100.]

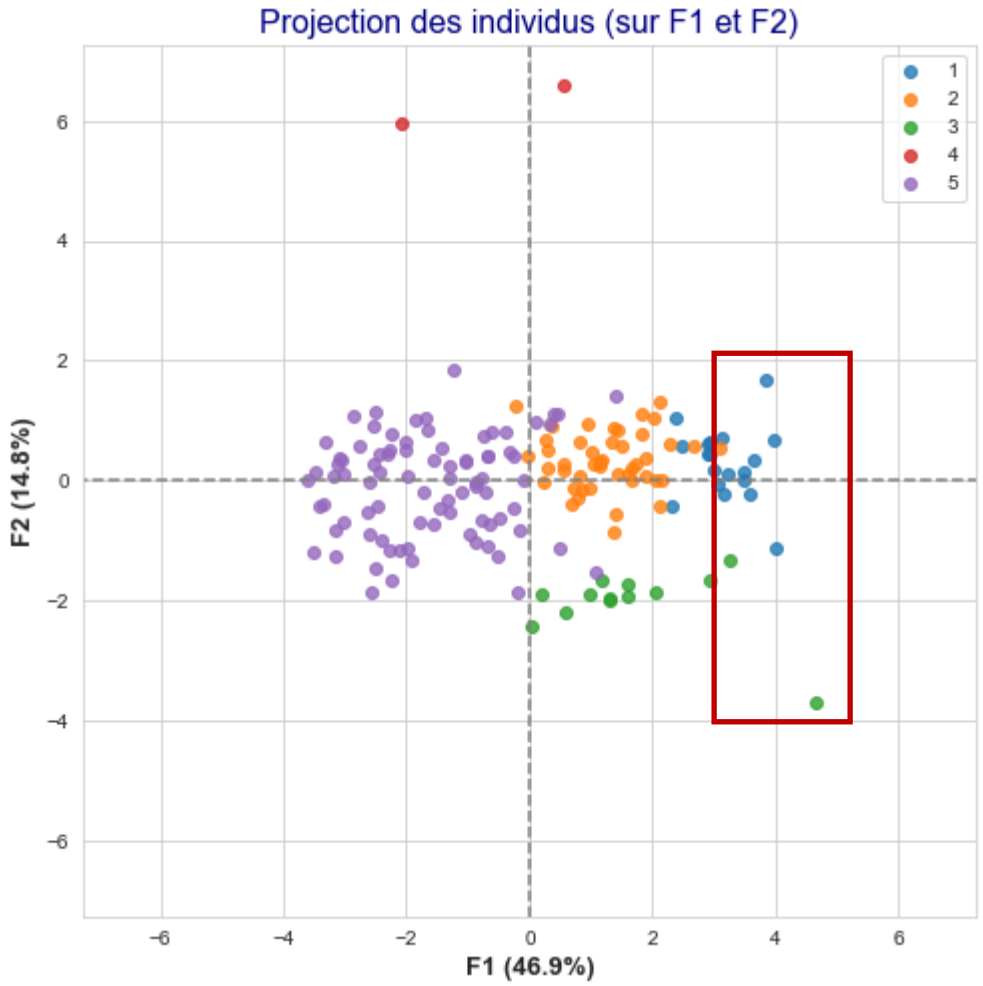
Cercle des corrélations

- Il apparaît clairement que les variables prot_g_cap_day, prot_ani_ratio , food_kcal_cap_day , edb_2019 , gdp_2013 et, dans une moindre mesure, poultry_food_ kg_cap_day sont très bien représentées sur l'axe de du premier plan
- Il faut également noter une anti-corrélation sur ce plan entre pop_var et edb_2019
- Sur le deuxième plan, nous avons une projection anti-corrélée entre pop_2013 et poultry_import_ratio
- Il faut privilégier les individus les mieux positivement représentés sur l'axe F1
- Nous devons donc chercher les individus représentés le plus à droite sur le premier plan factoriel F1



Projection des individus

- Selon le cercle des corrélations, nous cherchons donc les individus qui ont les valeurs les plus élevées sur l'axe F1
- Graphiquement, on constate que cela concerne le cluster 1, ainsi qu'un ou deux des éléments des clusters 2 et 3
- On notera sur l'axe F2 la forte valeur des 2 pays du cluster 2, en ligne avec la représentation de pop_2016 sur le cercle des corrélations



Analyse et conclusions

- Nous avons conclu que les pays à sélectionner sont ceux qui ont la meilleure représentation à droite de l'axe F1, nous sélectionons les 12 premiers
- Nous partons du principe que le comanditaire est européen, nous excluons Hong Kong, USA, Australia et Israel
- Luxembourg et Iceland ont moins d'un million d'habitants, cela semble insuffisant pour développer un marché
- Austria et Finland consomment moins de 20kg de volaille par an
- Poultry import ratio élevé pour Netherlands, le pays exporte 2 fois plus de volaille qu'il n'en consomme

country	F1
China, Hong Kong SAR	4.66
Luxembourg	4.02
Norway	3.97
United States of America	3.85
Iceland	3.63
Denmark	3.58
Israel	3.50
Australia	3.48
Netherlands	3.27
Austria	3.24
Ireland	3.15
Finland	3.13

Les trois pays retenus sont donc

Norway

Denmark

Ireland

Tests statistiques

Test de Shapiro et représentations graphiques loi normale, tests de comparaison pour 'edb_2019' sur clusters 1 et 5

Test de Shapiro et représentations graphiques loi normale

Pour les tests, nous avons retenu les clusters 1 et 5

Shapiro-Wilk est un test de normalité indiqué pour le traitement déchantillons réduits

Il retourne les valeurs

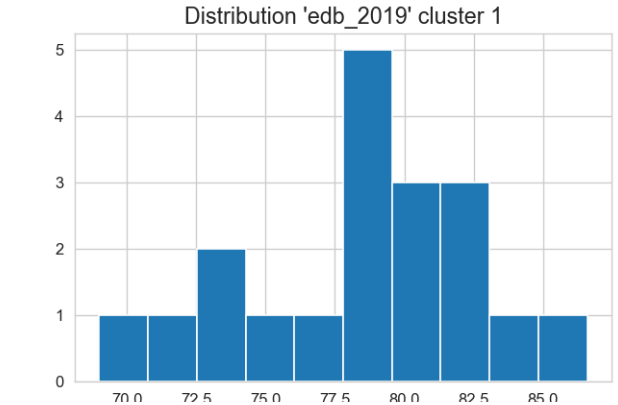
- **W = test statistique**
- **p-value = seuil hypothèse loi normale**

Interprétation

- **Hypothèse nulle H0 : la variable suit une loi normale : seuil α = 5%**
- **Si p:value < α : l'hypothèse nulle est rejetée**
- **Si la p-value > α : on ne doit pas rejeter l'hypothèse nulle**

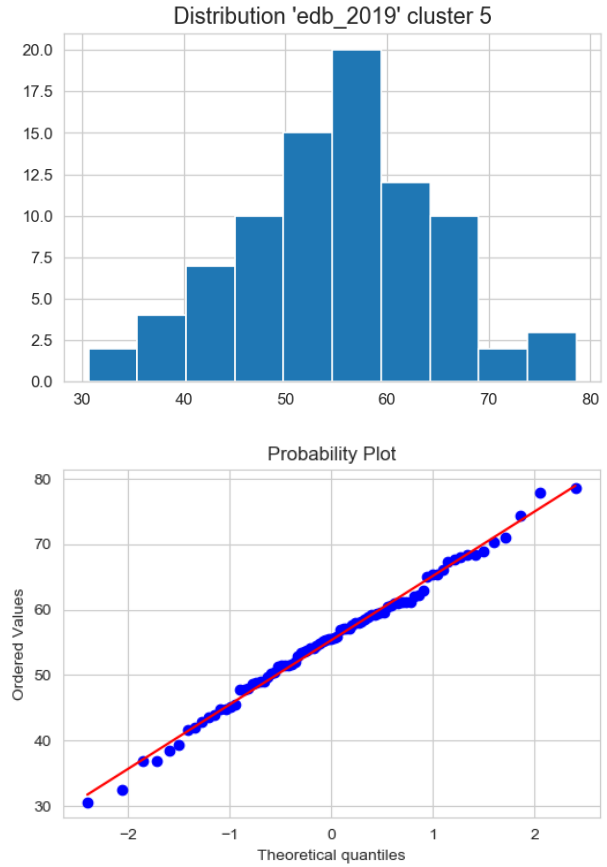
hypothèses nulles rejetées cluster 1

- **pop_2013** : 7.178522309914115e-07
- **gdp_2013** : 0.0007586915162391961
- **poultry_food_kg_cap_y** : 0.003183872438967228
- **poultry_import_ratio** : 0.02942577376961708



hypothèses nulles rejetées cluster 5

- **pop_2013** : 6.778163642341684e-14
- **gdp_2013** : 3.1339180139866585e-08
- **food_kcal_cap_day** : 0.036934543401002884
- **prot_g_cap_day** : 0.0017485303105786443
- **prot_ani_ratio** : 0.021738886833190918
- **poultry_food_kg_cap_y** : 2.7289121362628066e-07
- **poultry_import_ratio** : 2.499110030029783e-09



Tests de comparaison pour 'edb_2019' sur clusters 1 et 5

2.1 Test de Bartlett

- Comparation des variances avec le Test de Bartlett
- Hypothèse nulle (H0) = Egalité des variances de la variable 'edb_2019' sur les groupes 1 et 5

```
b = stats.bartlett(cl1['edb_2019'], cl5['edb_2019'])
print("p-value (test de Bartlett) pour la variable 'edb_2019' =", b[1])
```

```
p-value (test de Bartlett) pour la variable 'edb_2019' = 0.0006119503668720269
```

Hypothèse nulle rejetée

- Les conditions de validité du ttest ne sot pas remplies
- Le Student test sur l'égalité des moyennes n'est pas nécéssaire
- Toutefois, nous terminons la démonstration

2.2 t-test

```
t = stats.ttest_ind(cl1['edb_2019'], cl5['edb_2019'], equal_var=True)
print("p-value t-test de la variable 'edb_2019' sur les clusters 1 et 5 =", t[1])
```

```
p-value t-test de la variable 'edb_2019' sur les clusters 1 et 5 = 2.2683141666114416e-17
```

Hypothèse nulle rejetée

- Différence significative de la moyenne, conditions non valides

2.3 Test de Wilcoxon

Le test des rangs signés de Wilcoxon est une alternative **non-paramétrique** au test de Student pour des échantillons appariés. Il s'intéresse à un paramètre de position : la médiane, le but étant de tester s'il existe un changement sur la médiane.

Le test de Wilcoxon sur 2 samples est aussi appelé test Mann-Whitney

```
from scipy.stats import mannwhitneyu

# Comparation des samples
stat, p = mannwhitneyu(cl1['edb_2019'], cl5['edb_2019'])
print('Statistics=%.3f, p=%.3f' % (stat, p))

# interprétation
alpha = 0.05
if p > alpha:
    print("Même distribution (l'hypothèse H0 ne peut pas être rejetée)")
else:
    print("Distribution différente (rejet de H0)")
```

```
Statistics=19.000, p=0.000
Distribution différente (rejet de H0)
```

Hypothèse nulle rejetée

Les distributions sont bien différentes

Conclusion

Pour aller plus loin

Pour aller plus loin

Nous devons considérer cette étude comme préliminaire. En effet, les données proposées ont une autre destination, l'étude de la nutrition dans le monde. Voici quelques pistes pour un dataset plus complet et mieux adapté à une étude maketing

- Longitude et lattitude pour déterminer les distances dapprovisionnement
- Données du commanditaire : Sites de production, pays de destination (clients), type de produits (entiers, en découpe, frais ou congelé, produits transformés)
- Une version plus détaillée du risque pays (edb) infrastructures, corruption, sécurité financière, facilité dimplantation
- Des données plus récentes et plus précises des modes et habitudes de consommation (calories, protéines, quantité et type de produit volailler – entier, à la découpe, transformé, frais ou surgelé), voire au détail vs destiné à des industriels. soit une étude marketing ou des données produites par le syndicat de la fillière plutôt que des données de la FAO destinées à létude de la sécurité alimentaire
- On peut conserver le PIB mais une variable de parité de pouvoir d'achat serait intéressante

Des questions ?

lgourdon@gmail.com (+34 659 34 05 15)