Projet-5
Études de marché

(Entreprise d'agroalimentaire spécialisée dans le poulet)

Manu Sharma
Data Analyst (Openclassrooms)

Objectif du Projet

Identifier et regrouper les pays susceptibles d'exporter sur le marché international du poulet.

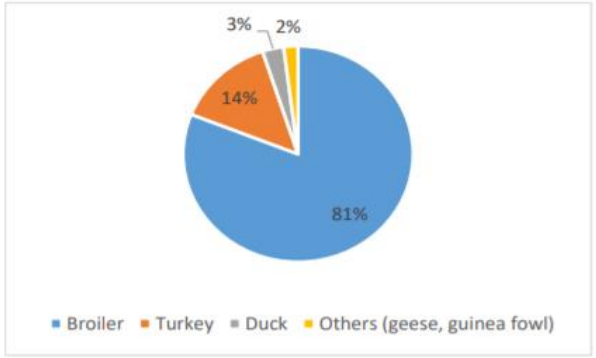
Mission du Projet

Pour réaliser ce projet, il existe différentes variables pour caractériser le régime alimentaire des pays.

- *Disponibilité alimentaire en calories par habitant*
- *Disponibilité alimentaire en protéines par habitant*
- *Rapport entre les protéines animales et les protéines végétal*
- *Population entre 2017 et 2018*
- *PIB par habitant*
- *Poulet export valeur*
- *Poulet import valeur*

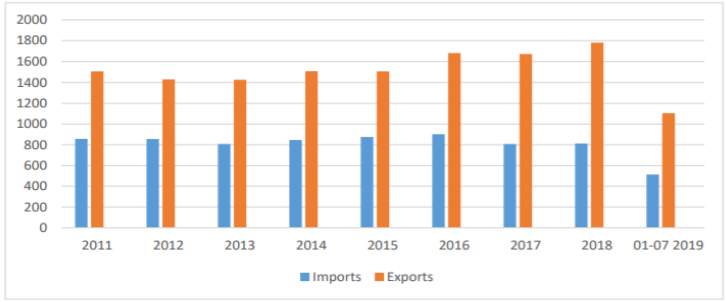
Tendance actuelle du marché du Poulet

Figure 1 – Poultry meat production in the EU

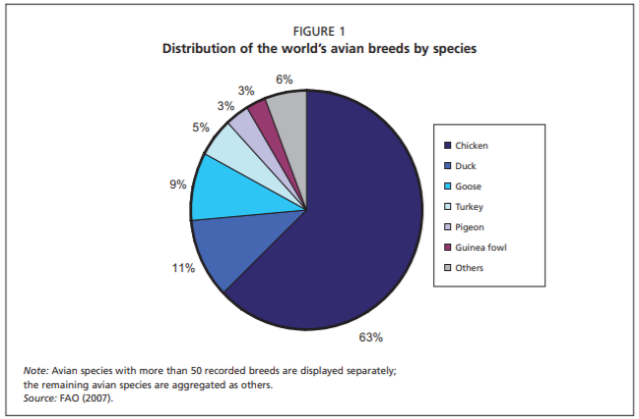


Data source: European Commission.

Figure 4 – EU poultry meat trade balance in 1 000 tonnes CWE



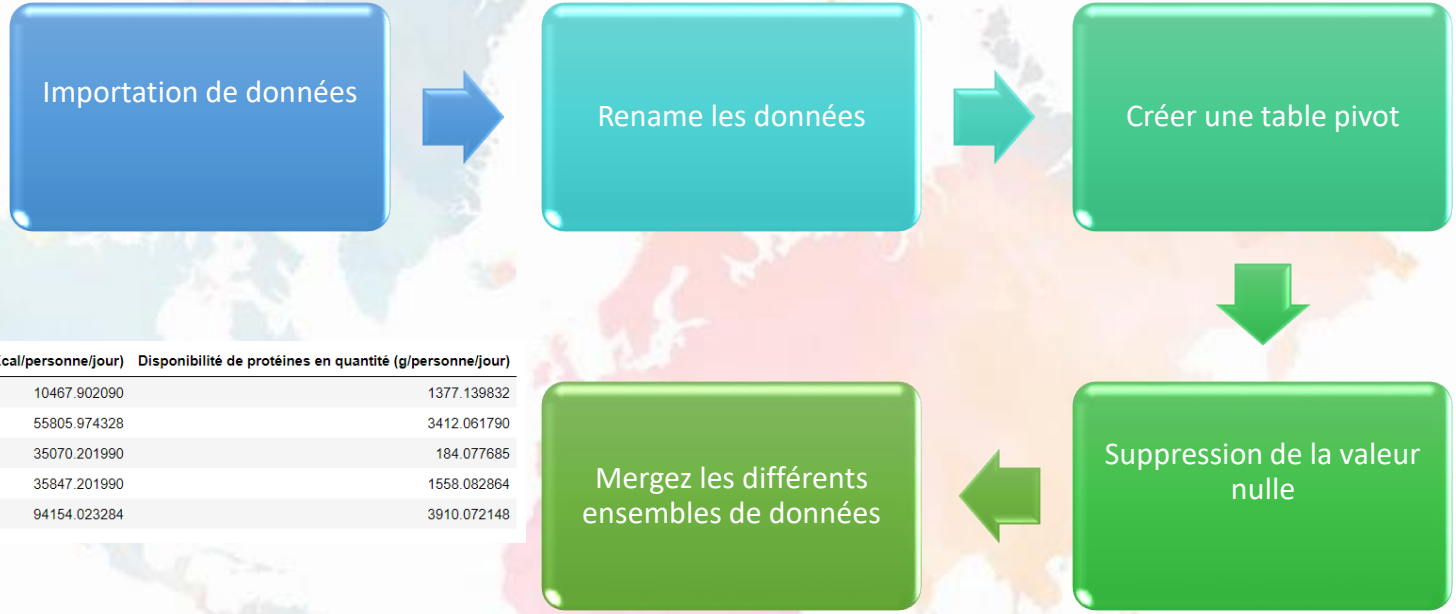
Data source: European Commission.



Dans ce graphique, nous pouvons voir la répartition des races aviaires du monde par espèce.

- *Nous avons analysé avec le diagramme que le poulet a un pourcentage élevé de distribution dans le monde.*
- *Les autres races suivantes ont une faible contribution*

Exportation des données, traitement et nettoyage des données



	Zone	Disponibilité alimentaire (Kcal/personne/jour)	Disponibilité de protéines en quantité (g/personne/jour)
0	Afghanistan	10467.902090	1377.139832
1	Afrique du Sud	55805.974328	3412.061790
2	Albanie	35070.201990	184.077685
3	Algérie	35847.201990	1558.082864
4	Allemagne	94154.023284	3910.072148

Zone	Disponibilité alimentaire (Kcal/personne/jour)	Disponibilité de protéines en quantité (g/personne/jour)	Code zone	ratio_evol_pop	ratio_protein_anim
Afghanistan	10467.902090	1377.139832	2	34.1	70.3
Afrique du Sud	55805.974328	3412.061790	202	16.1	46.5
Albanie	35070.201990	184.077685	3	-4.0	60.3
Algérie	35847.201990	1558.082864	4	21.6	59.1
Allemagne	94154.023284	3910.072148	79	2.5	51.8

Tableau final avec les variables nécessaires

Data Processing with Pandas


```
Zone                                                  0
Disponibilité alimentaire (Kcal/personne/jour)        0
Disponibilité de protéines en quantité (g/personne/jour)    0
ratio_evol_pop                                        0
ratio_protein_anim                                    0
Pib_par_habitant_2017                                 0
pib_par_habitant_2018                                 0
Importations Valeur 2017                              0
Importations Valeur 2018                              0
Exportations Valeur 2017 usd                          0
Exportations Valeur 2018 usd                          0
dtype: int64
```

```
data_full_final.shape
```

```
(96, 11)
```

	Zone	Disponibilité alimentaire (Kcal/personne/jour)	Disponibilité de protéines en quantité (g/personne/jour)	ratio_evol_pop	ratio_protein_anim	Pib_par_habitant_2017	pib_par_habitant_2018	Importations Valeur 2017	Importations Valeur 2018	Exportations Valeur 2017 usd	Exportations Valeur 2018 usd
0	Afrique du Sud	55805.974328	3412.061790	16.1	46.5	6.153459e+09	6.412963e+09	7394000.0	8671000.0	7992000.0	9671000.0
1	Allemagne	94154.023284	3910.072148	2.5	51.8	4.464274e+10	4.799347e+10	193759000.0	185487000.0	654322000.0	680352000.0
2	Antigua-et-Barbuda	47683.738109	276.093222	12.8	51.1	1.439025e+10	1.562905e+10	35000.0	84000.0	6000.0	2000.0
3	Arabie saoudite	43484.242388	1513.093222	30.2	51.8	2.090539e+10	2.331989e+10	17495000.0	19146000.0	12897000.0	7157000.0
4	Argentine	104493.657313	202.077685	10.7	27.7	1.451729e+10	1.160189e+10	14940000.0	13863000.0	1445000.0	29925500.0

Analyse des données

Clustering hiérarchique

Clustering hiérarchique - Partitionnement

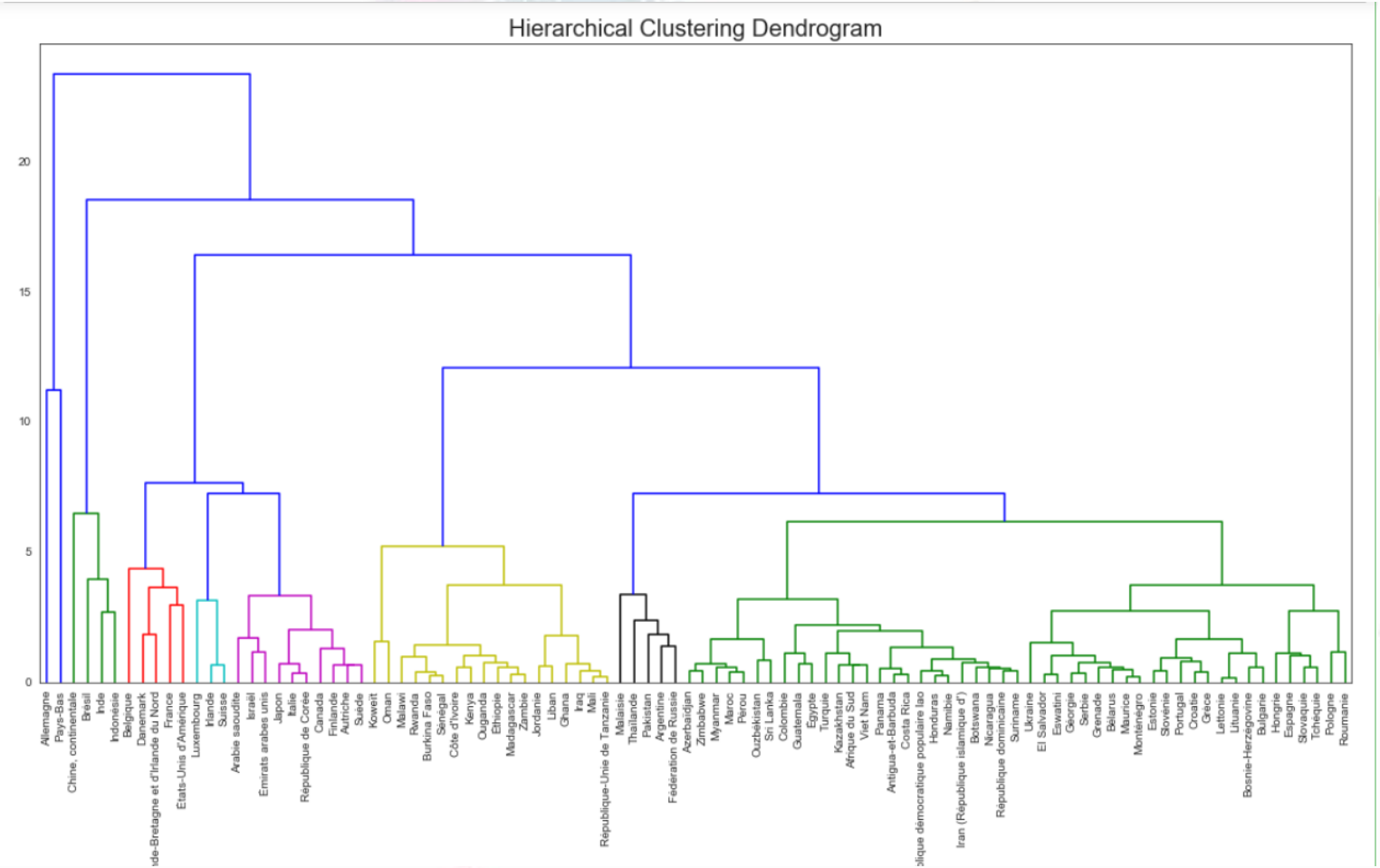
- *Pour la division en 5 groupes, j'ai utilisé la **méthode Ward**.*
- *Avec cette technique, un dénogramme a été réalisé pour identifier les groupes de pays les plus similaires.*
- *La taille de l'échantillon permet la réalisation du dendogramme bien que l'algorithme soit très complexe en temps et en espace.*

```
#Hierarchical clustering: creation of a link matrix using Ward's method
Y = linkage(X_scaled, method = 'ward', metric='euclidean')
```

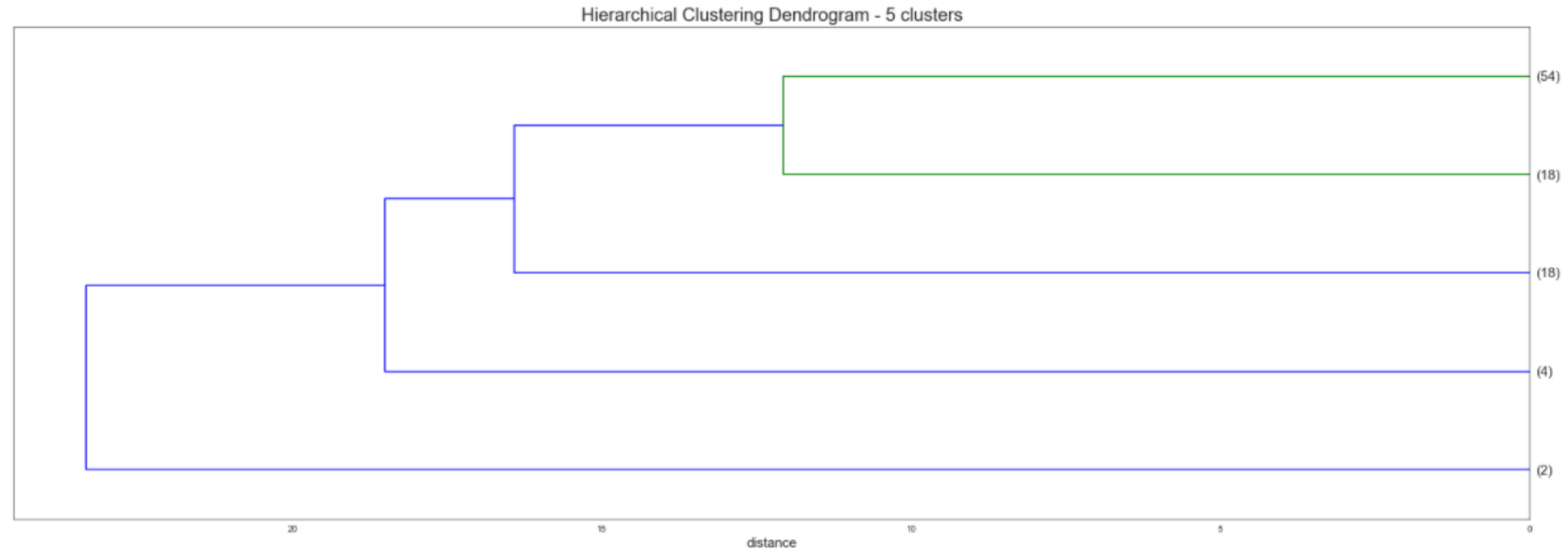
```
#Display of a first global dendrogram
fig =plt.figure(figsize=(20,10))
sns.set_style('white')
plt.title('Hierarchical Clustering Dendrogram',fontsize=20)
plt.xlabel('Distance')

dendrogram(Y, labels=df_cluster.index, leaf_font_size=10, color_threshold=7, orientation='top')
plt.show()
```

La méthode de Ward est en fait une méthode qui essaie de minimiser la variance au sein de chaque cluster. Dans K-means, lorsque nous essayions de minimiser le wcss pour tracer notre graphique de la méthode elbow, ici c'est presque la même chose, la seule différence est qu'au lieu de minimiser le wcss, nous minimisons les variantes à l'intérieur des clusters. C'est-à-dire la variance au sein de chaque cluster.

Hierarchical Clustering Dendrogram



Découpage du dendrogram en 5 groupes



- *Voici le dendogram, quand la division du groupe en 5 groups:*
- *1 cluster avec 54 pays*
- *2 cluster avec 18 pays*
- *3 cluster avec 18 pays*
- *4 cluster avec 4 pays*
- *5 cluster avec 2 pays*

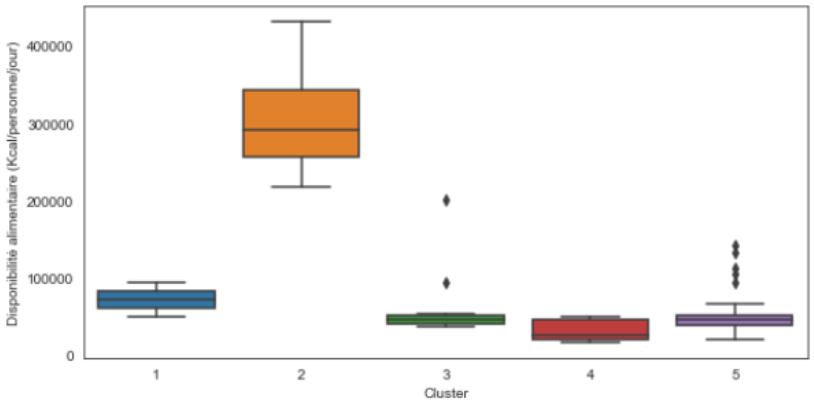
Description de 5 groupes de CAH

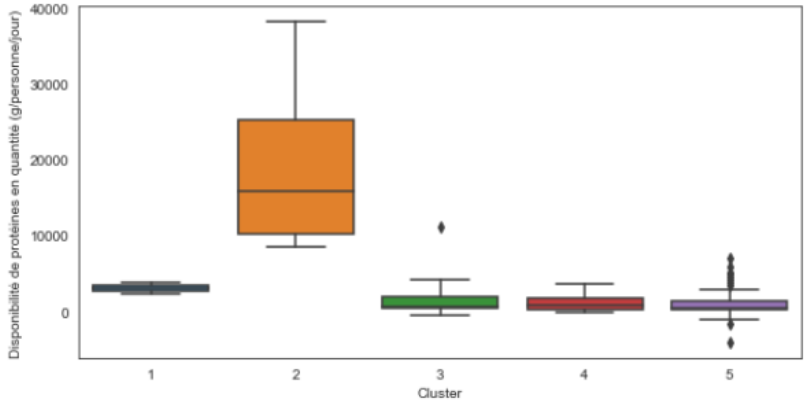
Cluster	Disponibilité alimentaire (Kcal/personne/jour)	Disponibilité de protéines en quantité (g/personne/jour)	ratio_evol_pop	ratio_protein_anim	F
1	71962.181791	3079.469200	2.750000	51.700000	
2	307369.047761	19620.857506	10.250000	11.200000	
3	55630.332482	1691.514668	10.983333	52.527778	
4	31059.047286	1071.086177	37.905556	61.844444	
5	50516.817000	1001.846848	7.109259	48.796296	

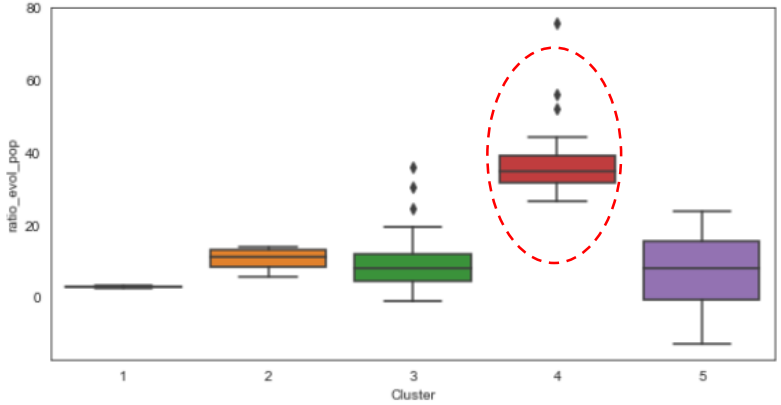
- *Avec l'aide du clustering hiérarchique, nous pouvons analyser que les clusters 3 et 4 ont un ratio élevé de protéines animales par rapport aux autres clusters.*

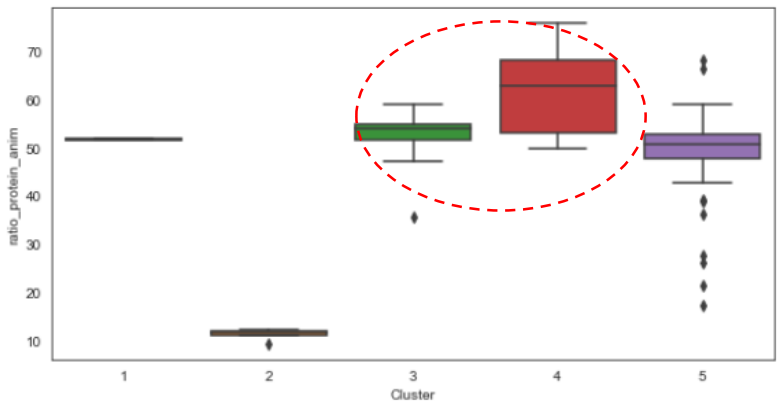
- *Donc, nous pouvons cibler ces clusters et analyser les résultats ultérieurs.*

Caractéristiques de 5 groupes avec box plot



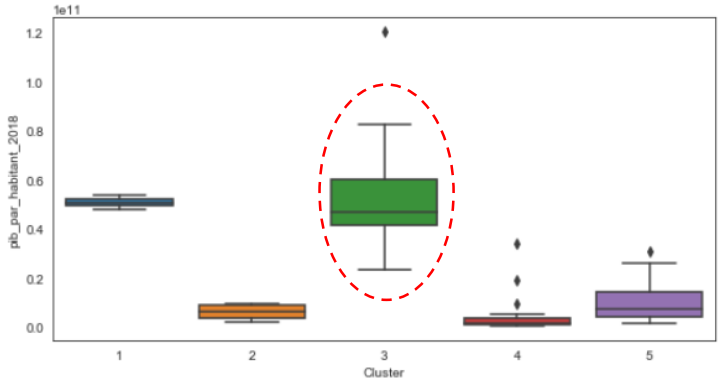


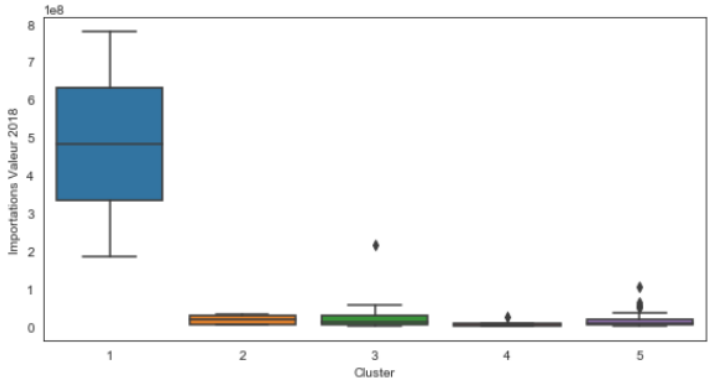


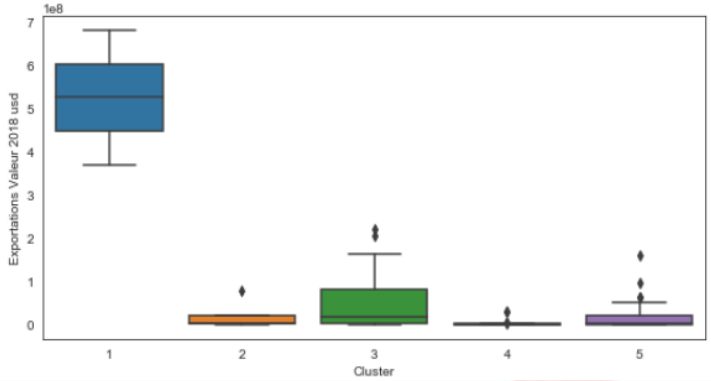


Nous pouvons analyser les caractéristiques des clusters à l'aide d'un box plot.
Ici, les ***clusters 3 et 4 s****emblent plus intéressants en termes de* ***population et de protéines de ratio****.*

Caractéristiques de 5 groupes avec box plot







*Ici, les **clusters 3** semblent plus intéressants en termes de **PIB_PAR_HABITANT_2018.***

Extraction des pays sur la base de 3&4 clusters

Extract countries on the basis of groupe (4,3)

```
df_HC_subset = df_groupes_HC.query('[3,4] in Cluster')
df_HC_subset.shape
```

```
(36, 12)
```

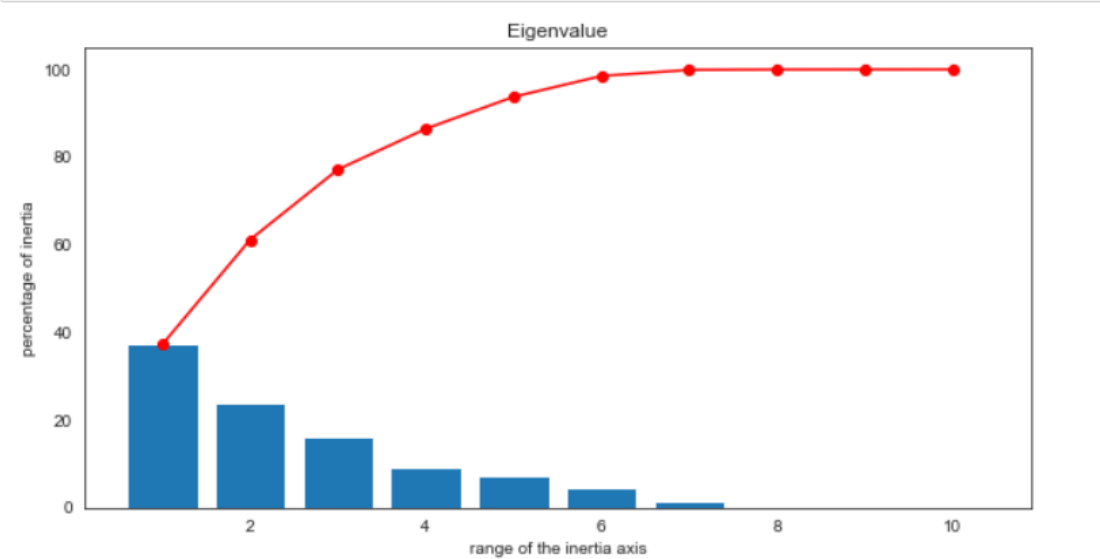
```
df_HC_subset.head()
```

	Zone	Disponibilité alimentaire (Kcal/personne/jour)	Disponibilité de protéines en quantité (g/personne/jour)	ratio_evol_pop	ratio_protein_anim	Pib_par_habitant_2017	pib_par_habitant_2018	Importations Valeur 2017	Importations Valeur 2018
3	Arabie saoudite	43484.242388	1513.093222	30.2	51.8	2.090539e+10	2.331989e+10	17495000.0	19146000.0
5	Autriche	40398.884975	240.088043	6.6	53.6	4.788717e+10	5.204726e+10	32714000.0	36944000.0
7	Belgique	44544.250945	2041.066611	6.5	56.1	4.411893e+10	4.722576e+10	207538000.0	213538000.0
12	Burkina Faso	16335.625473	-159.911957	34.5	72.8	7.382748e+08	8.201668e+08	233000.0	384000.0
14	Canada	50578.429652	3010.512506	11.2	54.7	4.505729e+10	4.634337e+10	52627000.0	57158000.0

*En conclusion, ce premier partitionnement **montre que 36 pays sont susceptibles** de devenir une cible pertinente pour l'entreprise. La demande sera présente dans ces pays, notamment en termes de besoins **en protéines animales ratio**.*

ACP
Analyse en composantes principales

Eboulis Valeurs Propres




```
print(pca.explained_variance_ratio_)
```

```
[3.72516394e-01 2.38998864e-01 1.60205247e-01 9.27950341e-02
 7.29185831e-02 4.75043760e-02 1.38933917e-02 9.08073364e-04
 1.62904225e-04 9.71323246e-05]
```

```
print(pca.explained_variance_ratio_.cumsum())
```

```
[0.37251639 0.61151526 0.77172051 0.86451554 0.93743412 0.9849385
 0.99883189 0.99973996 0.99990287 1.        ]
```

*A l'aide du **scree plot**, nous pouvons analyser que les 3 premières composantes principales représentent **77% de la variance des données.***

Circle de Corrélation

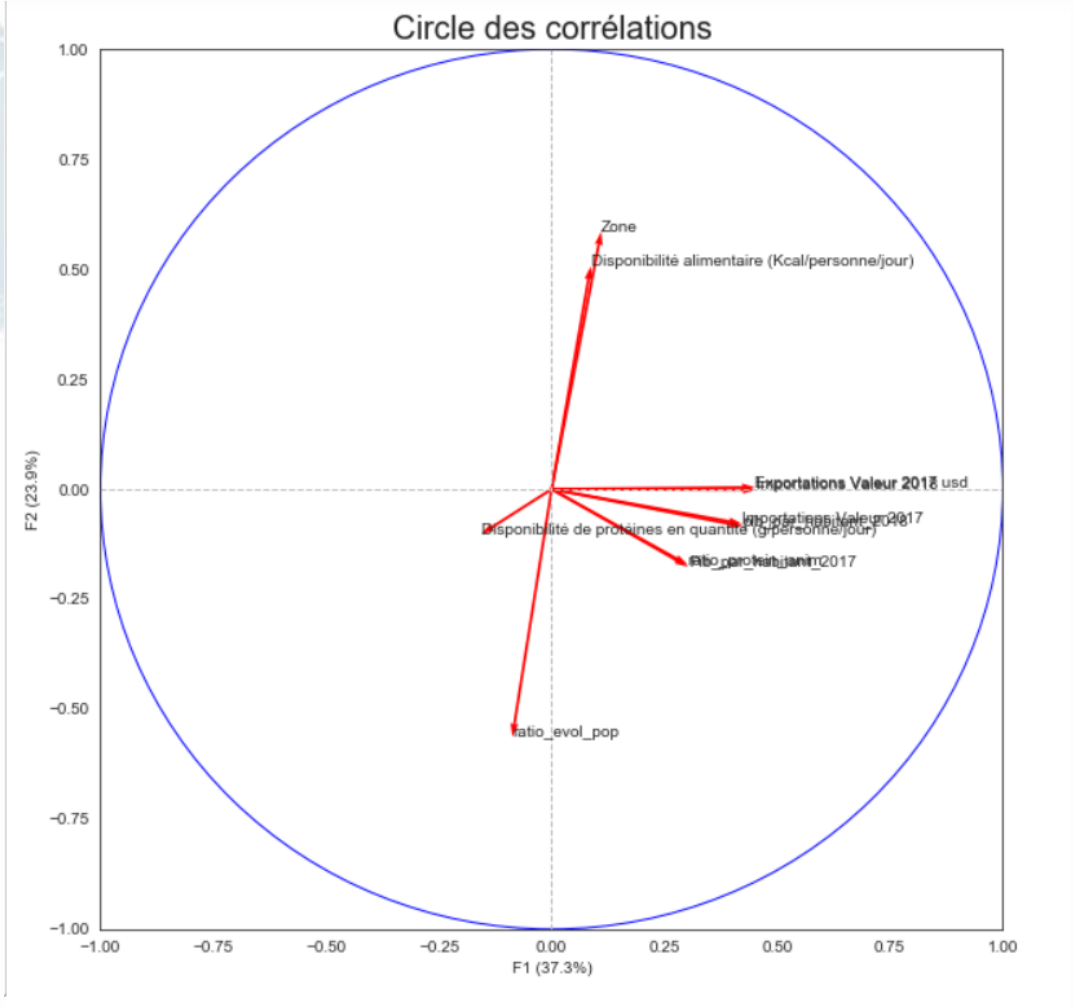
- *F1 montre 37 % de la variance et caractérise les pays pour lesquels les données alimentaires sont déjà importantes*
- *F2 montre 23% de variance et caractérise les pays pour lesquels l'évolution de la population est importante*

Disponibility alimentaire

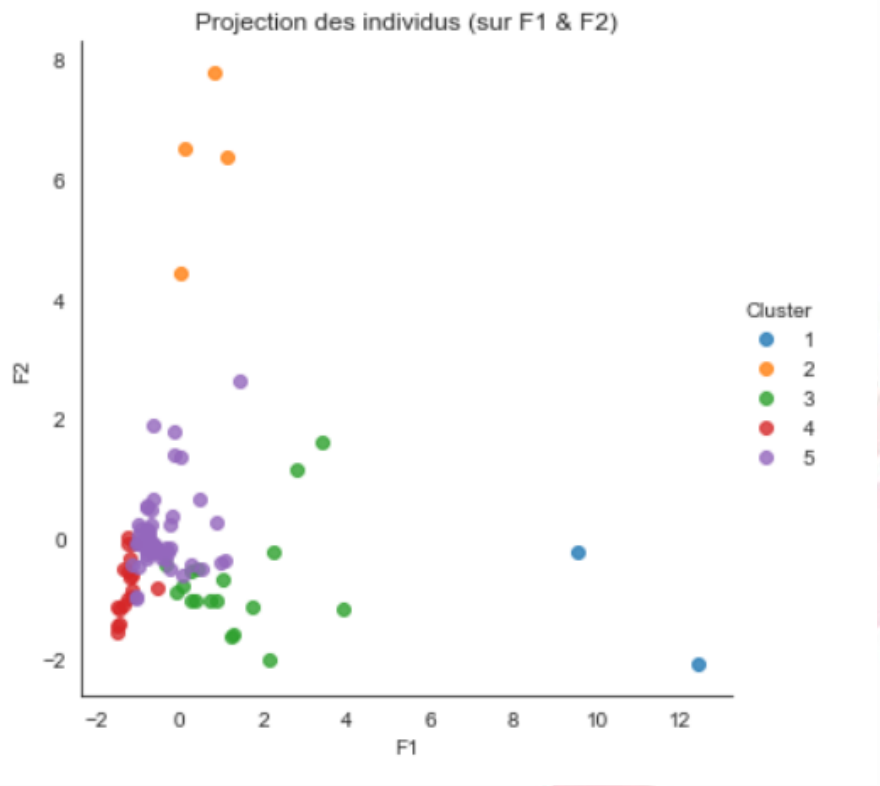
Proteines

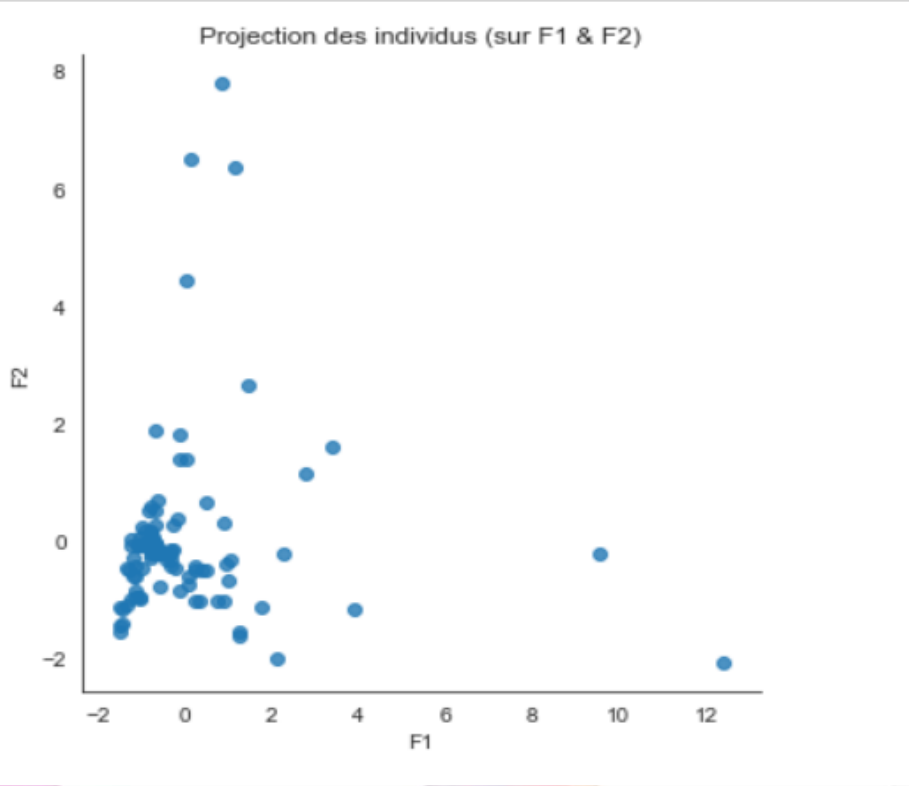
Ratio

Population



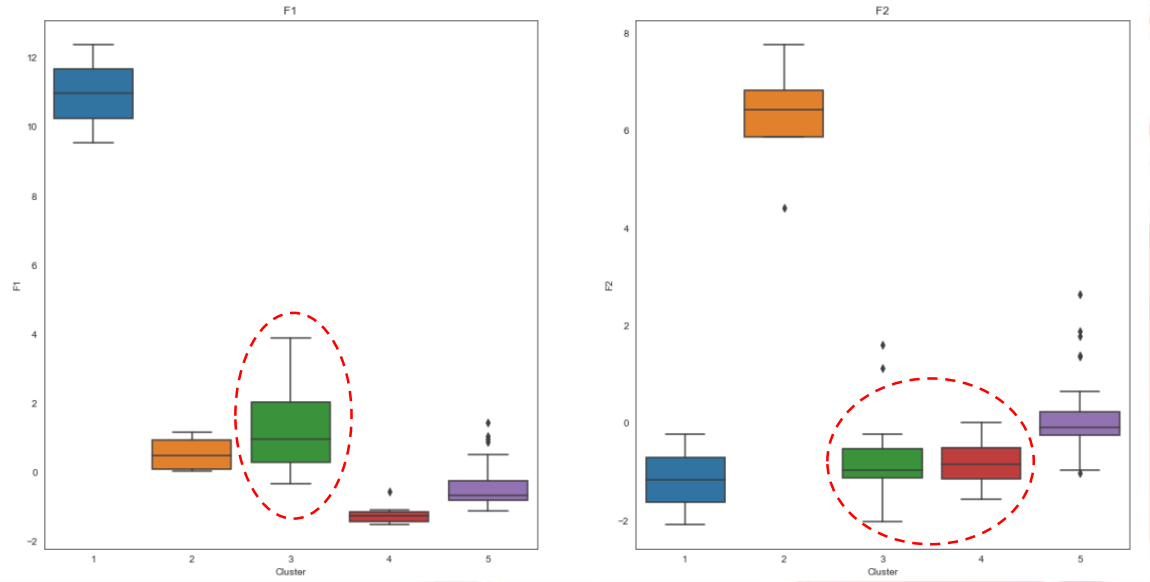
Projection des Individus et clusters





- *Le graphique montrant les composantes principales sur le **plan factoriel** avec le cluster également*
- *Le **cluster 3 semble présenter plus de variance dans les données***

Projection des composantes principales avec box plot



- *Le graphique montre les deux premières composantes principales dans les 5 clusters.*
- *Les Cluster **sélectionnées 3 & 4** ont des cibles attrayantes.*

	Cluster	Disponibilité alimentaire (Kcal/personne/jour)	Disponibilité de protéines en quantité (g/personne/jour)	ratio_evol_pop	ratio_protein_anim	Pib_par_habitant_2017	pib_par_habitant_2018	Importations Valeur 2017	Importations Valeur 2018
0	1	71962.181791	3079.469200	2.750000	51.700000	4.679521e+10	5.074926e+10	4.908070e+08	4.816465e+08
1	2	307369.047761	19620.857506	10.250000	11.200000	6.058593e+09	6.109294e+09	1.090125e+07	1.687400e+07
2	3	55630.332482	1691.514668	10.983333	52.527778	5.008495e+10	5.338347e+10	2.584611e+07	2.707967e+07
3	4	31059.047286	1071.086177	37.905556	61.844444	4.304459e+09	4.791116e+09	2.955778e+06	4.860416e+06
4	5	50516.817000	1001.846848	7.109259	48.796296	9.060775e+09	9.752331e+09	1.301169e+07	1.463718e+07

Pays avec cluster-3

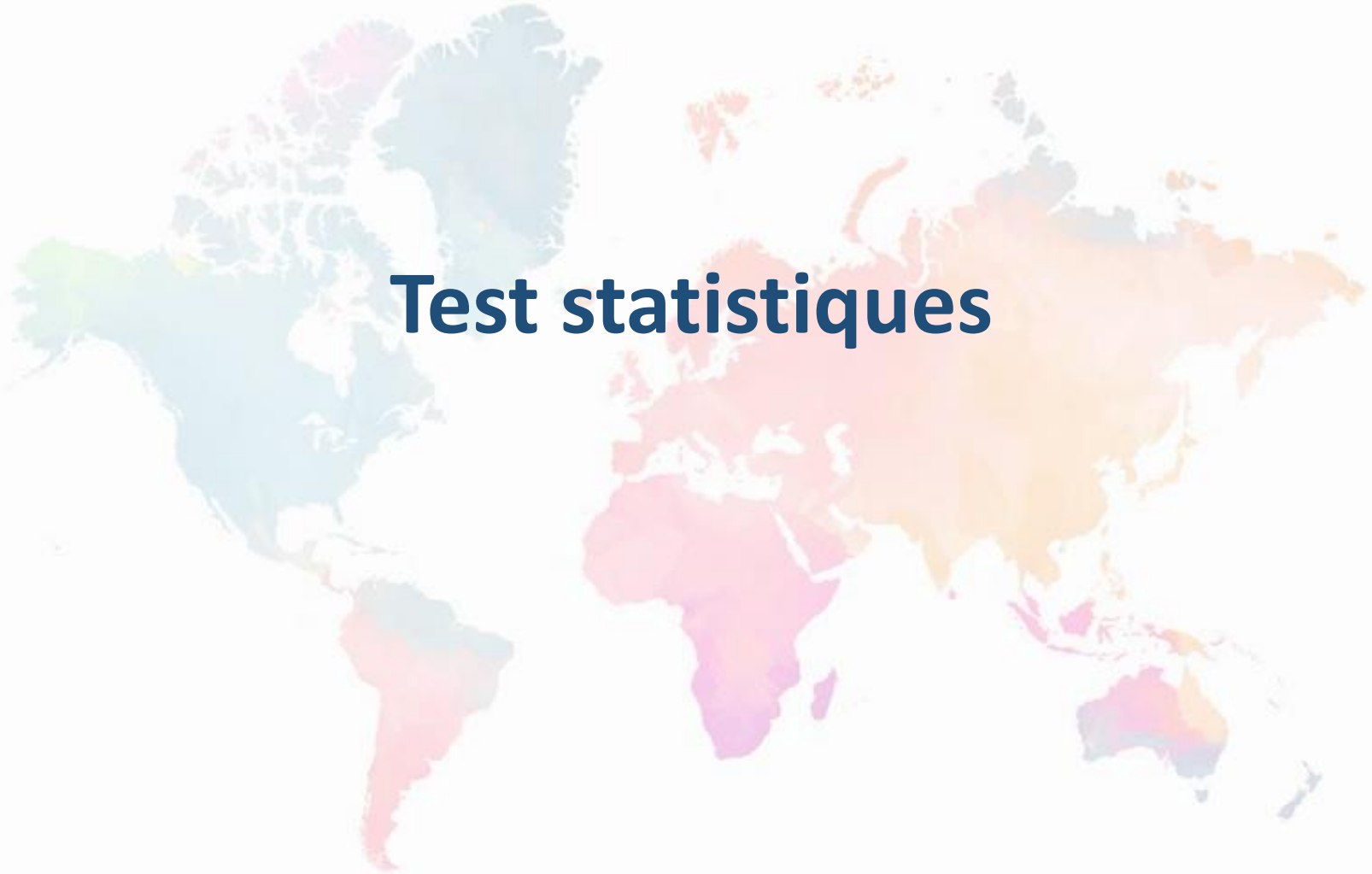
*Les pays du **cluster 3** sont susceptibles d'exporter sur le marché international du poulet en fonction de leur tendance attractive dans **Pib_par_habitant_2018 et présentent une variance dans les données.***

Zone	Disponibilité alimentaire (Kcal/personne/jour)	Disponibilité de protéines en quantité (g/personne/jour)	ratio_evol_pop	ratio_protein_anim	Pib_par_habitant_2017	pib_par_habitant_2018	Importations Valeur 2017	Importations Valeur 2018
Arabie saoudite	43484.242388	1513.093222	30.2	51.8	2.090539e+10	2.331989e+10	17495000.0	19146000.0
Autriche	40398.884975	240.088043	6.6	53.6	4.788717e+10	5.204726e+10	32714000.0	36944000.0
Belgique	44544.250945	2041.066611	6.5	56.1	4.411893e+10	4.722576e+10	207538000.0	213538000.0
Canada	50578.429652	3010.512506	11.2	54.7	4.505729e+10	4.634337e+10	52627000.0	57158000.0
Danemark	38144.340299	848.376611	4.6	54.9	5.745429e+10	6.180975e+10	9861000.0	9978000.0
Finlande	36133.429652	-109.911957	3.8	54.6	4.618041e+10	4.995534e+10	1212000.0	1626000.0
France	93271.478607	11235.062148	4.5	47.2	3.856657e+10	4.116865e+10	26061000.0	29312000.0
Irlande	43088.063682	773.262148	9.1	59.1	7.049292e+10	7.966162e+10	6194000.0	7219000.0
Israël	38481.063682	376.072506	19.4	55.4	4.245033e+10	4.384183e+10	8221000.0	4030000.0
Italie	52696.657313	496.072506	2.9	52.9	3.304935e+10	3.516357e+10	14882000.0	18280000.0
Japon	48969.884975	559.072506	-1.0	51.3	3.812191e+10	3.908724e+10	7082000.0	9035000.0
Luxembourg	46651.193433	260.077685	24.5	51.4	1.100032e+11	1.201380e+11	418000.0	434000.0
Royaume-Uni de Grande-Bretagne et d'Irlande du Nord	53222.519005	3684.056969	8.0	49.0	4.028657e+10	4.288939e+10	39531000.0	49859000.0
République de Corée	46526.697711	-471.927494	4.0	55.0	3.185231e+10	3.362666e+10	22636000.0	11095000.0
Suisse	37664.201990	464.093222	11.8	56.1	8.022063e+10	8.253017e+10	2515000.0	2204000.0
Suède	40565.746667	828.056969	8.0	54.5	5.454157e+10	5.570497e+10	5366000.0	5414000.0
Émirats arabes unis	45593.104080	561.077685	35.8	52.2	4.018034e+10	4.340764e+10	2297000.0	2141000.0
États-Unis d'Amérique	201331.795622	4139.062148	7.8	35.7	6.015982e+10	6.298139e+10	8580000.0	10021000.0

Pays avec cluster-4

*Les pays du **cluster 4** sont susceptibles d'exporter sur le marché international du poulet en fonction de leur tendance attractive dans **ratio_protien_animal et Ratio_population_evolution.***

	Zone	Disponibilité alimentaire (Kcal/personne/jour)	Disponibilité de protéines en quantité (g/personne/jour)	ratio_evol_pop	ratio_protein_anim	Pib_par_habitant_2017	pib_par_habitant_2018	Importations Valeur 2017	Importations Valeur 2018
12	Burkina Faso	16335.625473	-159.911957	34.5	72.8	7.382748e+08	8.201668e+08	233000.0	3.840000e
19	Côte d'Ivoire	20907.170149	1886.072506	27.9	65.6	1.570209e+09	1.727629e+09	3825000.0	3.873000e
28	Ghana	46555.331741	1411.082864	26.3	52.9	2.046043e+09	2.224259e+09	8247000.0	8.903000e
38	Iraq	42002.242388	1477.103580	35.4	54.8	4.980948e+09	5.395860e+09	2092000.0	2.069000e
43	Jordanie	43583.559403	267.088043	52.0	53.8	4.195802e+09	4.264119e+09	6345000.0	5.998000e
45	Kenya	28272.210547	2147.137327	29.2	60.2	1.584660e+09	1.725303e+09	732000.0	2.394000e
46	Koweït	47716.738109	392.098401	55.8	51.4	2.889724e+10	3.376074e+10	5836000.0	8.802000e
48	Liban	41211.697711	211.093222	44.0	55.5	8.778472e+09	9.257297e+09	5989000.0	6.028000e
51	Madagascar	19129.259502	211.088043	31.3	68.5	5.152856e+08	5.273866e+08	826000.0	1.109000e
53	Malawi	18654.170149	3548.082864	32.2	72.1	3.408700e+08	3.754904e+08	1111000.0	1.719000e
54	Mali	45554.331741	1621.108758	35.2	52.5	8.292383e+08	8.986854e+08	4237000.0	2.662000e
61	Oman	48849.282786	371.067327	75.6	49.9	1.709928e+10	1.907091e+10	3025000.0	3.188000e
62	Ouganda	20848.170149	-103.927494	40.4	65.7	6.347691e+08	6.798672e+08	3571000.0	6.054000e
72	Rwanda	15730.397811	7.082864	29.2	75.9	7.483501e+08	7.607298e+08	754000.0	1.053000e
76	République-Unie de Tanzanie	46839.470050	2512.072506	34.5	52.7	9.548774e+08	9.944472e+08	2076000.0	3.317000e
84	Sénégal	16491.625473	667.056969	32.1	74.0	1.321086e+09	1.461175e+09	716000.0	1.261000e
90	Zambie	19803.397811	897.062148	35.0	67.4	1.513276e+09	1.549354e+09	3498000.0	4.410000e
95	Éthiopie	20578.170149	1917.093222	31.7	67.5	7.315802e+08	7.466642e+08	91000.0	2.426348e



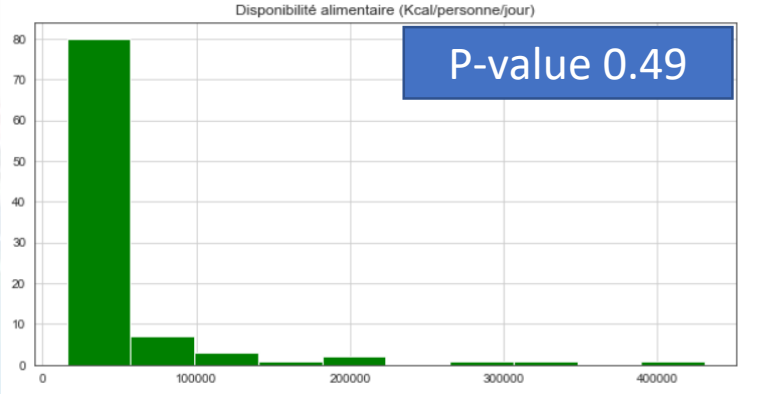
Test statistiques

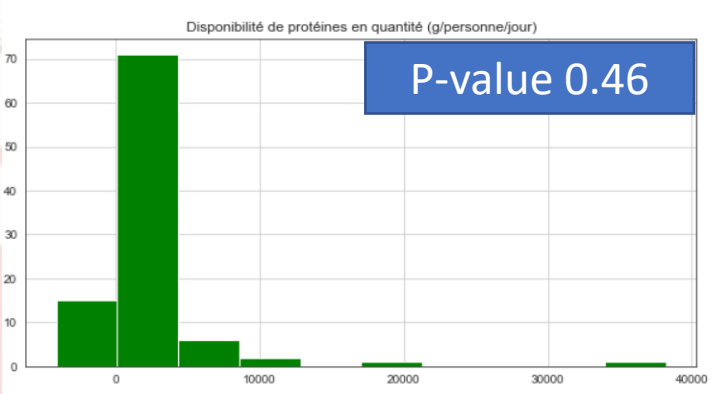
Comparisons Variables

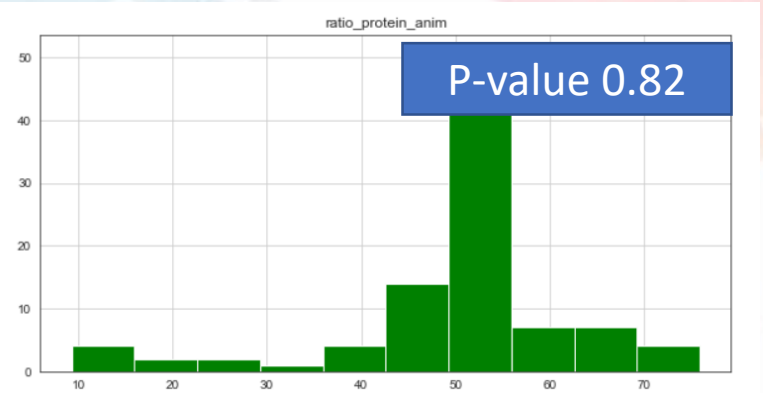
Shapiro Wilk

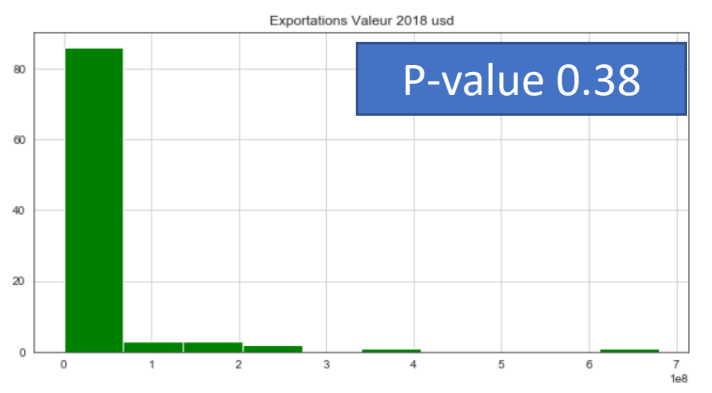
Dans le test d'adéquation, j'ai utilisé le test de distribution normale sur 4 variables :

- *Disponibilité alimentaire (Kcal/personne/jour)*
- *Disponibilité de protéines en quantité (g/personne/jour)*
- *Ratio_protein_animal*
- *Exportations Valeur 2018*









On ne peut pas rejeter l' hypothèse de normalité au niveau de test des variables

Inter-Cluster Comparisons Variables

Cluster 3&4

L'hypothèse nulle est rejetée pour les groupes 3 et 4 pour le ratio variable de protéines animales.

Plus,

H0 l'hypothèse d'égalité des moyennes est rejetée au niveau de test 5%.

```
cluster_test1 = data_clusters[data_clusters['Cluster'] == 3]['ratio_protein_anim']
cluster_test2 = data_clusters[data_clusters['Cluster'] == 4]['ratio_protein_anim']
```

```
#On teste tout d'abord l'égalité des variances à l'aide de la commande
from scipy.stats import bartlett
stat, p = bartlett(cluster_test1, cluster_test2)
print('Statistics=%.3f, p=%.3f' % (stat, p))

#Interprétation
alpha = 0.05
if p > alpha:
    print('On ne rejette donc pas H0, l'égalité des variances au niveau de test 5%')
else:
    print('H0 est rejetée au niveau de test 5%')
```

```
Statistics=5.231, p=0.022
H0 est rejetée au niveau de test 5%
```

```
#On teste ensuite l'égalité des moyennes à l'aide de la commande
from scipy.stats import ttest_ind
stat, p = ttest_ind(cluster_test1, cluster_test2, equal_var=True)
print('Statistics=%.3f, p=%.9f' % (stat, p))

#Interprétation
alpha = 0.05
if p > alpha:
    print('On ne rejette donc pas H0, l'égalité des moyennes de nos 2 clusters au niveau de test 5%')
else:
    print('H0 l\'hypothèse d'égalité des moyennes est rejetée au niveau de test 5%')
```

```
Statistics=-3.845, p=0.000504293
H0 l'hypothèse d'égalité des moyennes est rejetée au niveau de test 5%
```

Merci