OpenClassrooms - Data Analyst
2021-2022

PROJET 9: ÉTUDE DE MARCHÉ AVEC PYTHON

Xiuting LIANG 28.07.2021

Le contexte du projet de data analyse

La poule qui chante

▶ L'entreprise française d’agroalimentaire, La poule qui chante ,
▶ Souhaite se développer à l'international.
▶ Notre mission d’analyse est de proposer une première analyse des groupements de pays que l’on peut cibler pour exporter nos poulets.

▶ On ultilise des données de la FAO(Food and Agriculture Organization) et des données open source avec les critères de l’analyse PEST.

▶ Pour la partie analyse, on ultilise les méthodes CAH(classification ascendante hiérarchique), k-means pour la classification et également réaliser une ACP afin de visualiser les résultats.

Dans Cette Présentation

Le contexte du projet

Préparation des données

- PEST Analyse
- Pivot

Normalizer des données

La classification ascendante hiérarchique (CAH)

- Dendrogramme

KMeans

- Elbow method
- Silhouette analyse

ACP

- Explained variance ratio
- Cercle de corrélation
- Visualisation pour CAH
- Visualisation pour Kmeans
- Visualisation avec Geopandas

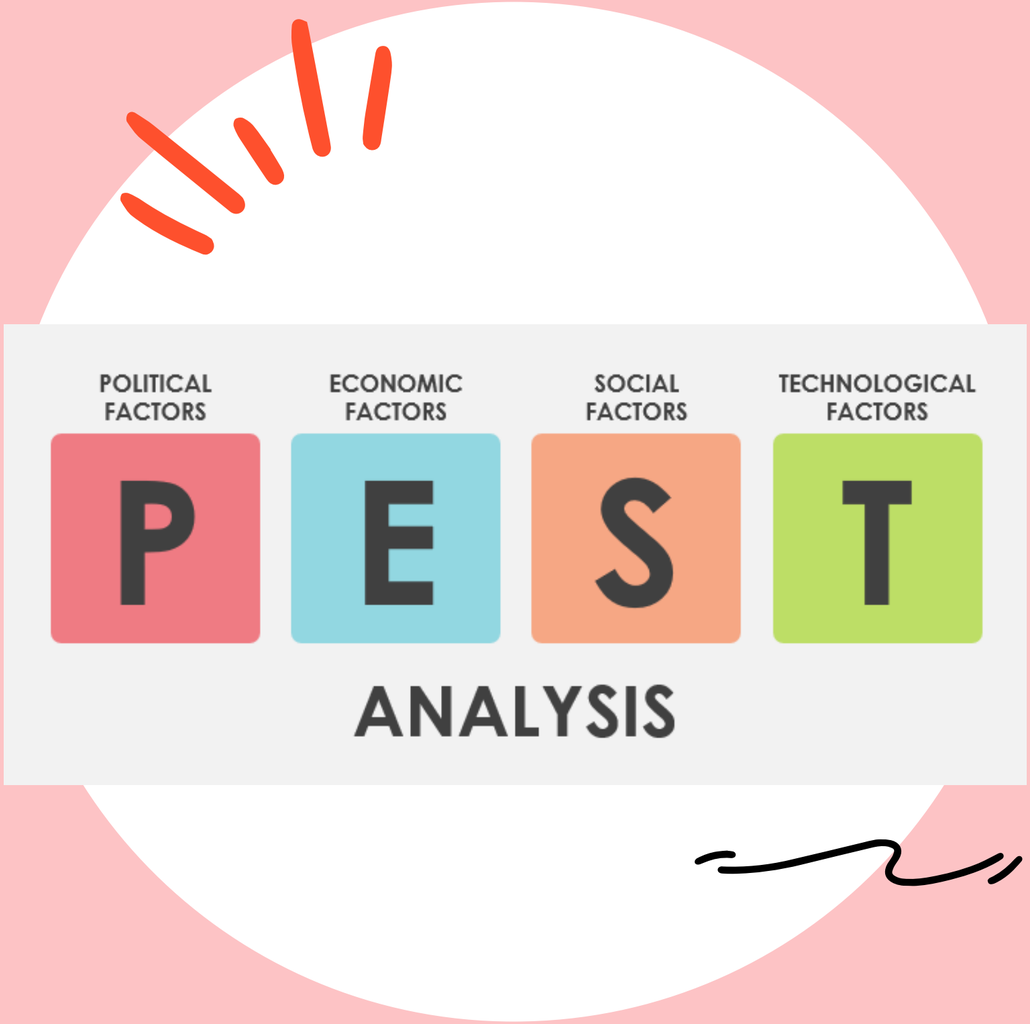
Partie supplémentaire: Recalculer les clusters après ACP avec CAH et kmeans

Heatmap avec clusters

Teste statistique

Recommandations

Préparation des données



Jeux de données

FAO(Food and Agriculture Organization) et des données **open source** avec les critères de l'analyse **PEST**

P	E	S	T
Political_Stability	PIB par habitant(USD), Prix poulet(1kg/USD)	Population(mille personnes)	Distance km

Autre: DisponibiliteAlimentaire_2017
Disponibilité alimentaire (Kcal/personne/jour),
Disponibilité de protéines en quantité (g/personne/jour),
Importations - Quantité(Milliers de tonnes),
Exportations - Quantité(Milliers de tonnes)

df_alimentaire

source: Données New Food Balances (FAO); données originals d'Openclassrooms.

--------->
- Suprimé les colonnes non pertinents
- Filtrer 'Viande de Volailles'
- **Traitement pour le pivot**

df_pop

source: Données New Food Balances (FAO); données originals d'Openclassrooms et transfer à version Anglais

--------->
- Suprimé les colonnes non pertinents
- Filtrer 2017
- **Suprimé les données doulons**

df_eco

source: FAO: https://www.fao.org/faostat/en/#data/MK

--------->
- Suprimé les colonnes non pertinents
- **Données FAO: la jointure de 'df_volailles', 'df_pop', 'df_eco'**
- **Traiter les valeurs manquants-fillna(0)**

df_prix

Source: 07/2022 https://www.numbeo.com/cost-of-living/country_price_rankings?itemId=19

--------->

df_poli

source: Données de P8, Openclassrooms

--------->
- La jointure des données open source
- **Traiter les valeurs manquants-fillna(means)**

--------->

df_distance

source: https://www.distancefromto.net/distance-from-france-country

Préparation de données

Avant traitement de pivot

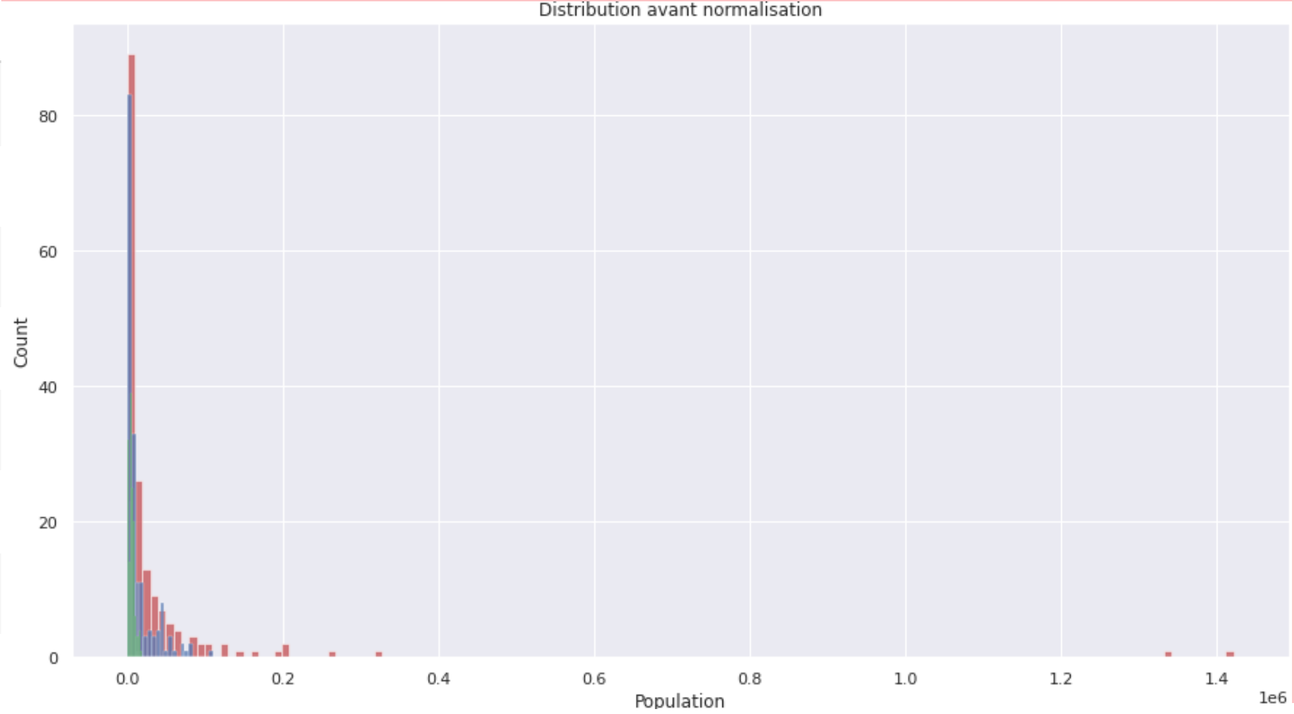
	Code zone	Zone	Valeur	Élément Unité
651	2	Afghanistan	28.0	Production(Milliers de tonnes)
652	2	Afghanistan	29.0	Importations - Quantité(Milliers de tonnes)
653	2	Afghanistan	0.0	Variation de stock(Milliers de tonnes)
654	2	Afghanistan	57.0	Disponibilité intérieure(Milliers de tonnes)
655	2	Afghanistan	2.0	Pertes(Milliers de tonnes)

Après traitement de pivot

	Élément Unité	Alimentation pour touristes(Milliers de tonnes)	Aliments pour animaux(Milliers de tonnes)	Autres utilisations (non alimentaire) (Milliers de tonnes)	Disponibilité alimentaire (Kcal/personne/jour) (Kcal/personne/jour)	Disponibilité alimentaire en quantité (kg/personne/an) (kg)	Disponibilité de matière grasse en quantité (g/personne/jour) (g/personne/jour)	Disponibilité de protéines en quantité (g/personne/jour) (g/personne/jour)	Disponibilité intérieure(Milliers de tonnes)	Exportations - Quantité(Milliers de tonnes)
Code zone	Zone									
1	**Arménie**	NaN	NaN	NaN	54.0	16.06	3.39	5.44	47.0	0.0
2	**Afghanistan**	NaN	NaN	NaN	5.0	1.53	0.33	0.54	57.0	NaN
3	**Albanie**	NaN	NaN	NaN	85.0	16.36	6.45	6.26	47.0	0.0
4	**Algérie**	0.0	NaN	NaN	22.0	6.38	1.50	1.97	277.0	0.0
7	**Angola**	0.0	NaN	NaN	35.0	10.56	2.22	3.60	319.0	0.0

Avant normalisation

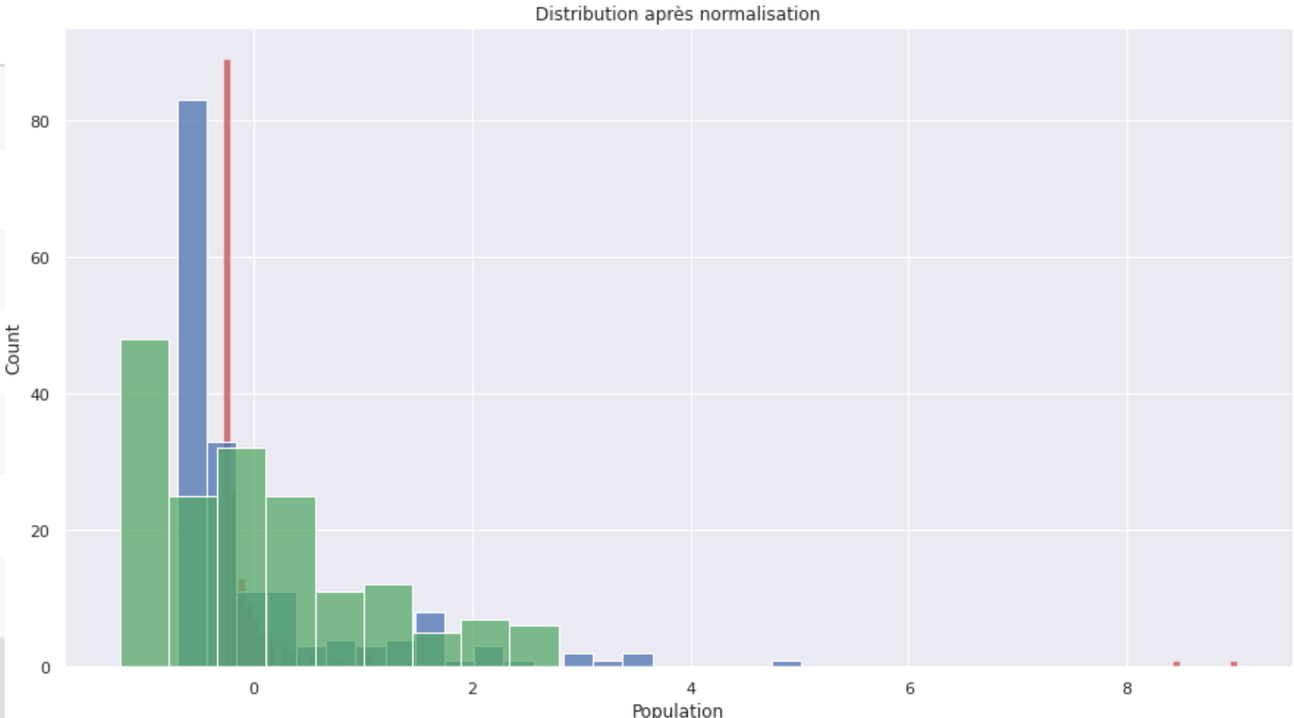
	Population
count	1.710000e+02
mean	4.295384e+04
std	1.535062e+05
min	5.204500e+01
25%	2.864792e+03
50%	9.729823e+03
75%	3.046071e+04
max	1.421022e+06



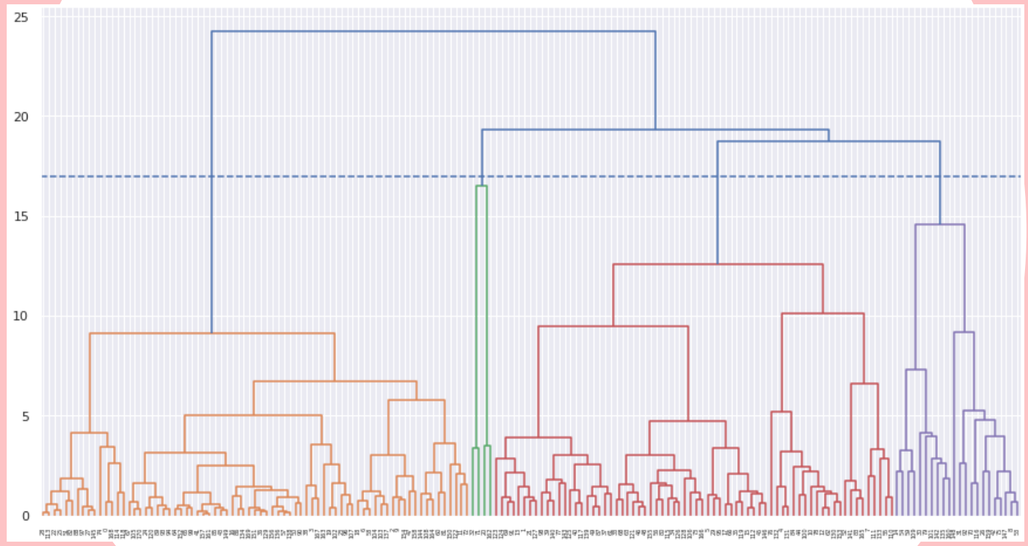
Après normalisation

le centrage et la réduction

	Population
count	1.710000e+02
mean	2.077610e-17
std	1.002937e+00
min	-2.803001e-01
25%	-2.619230e-01
50%	-2.170701e-01
75%	-8.162427e-02
max	9.003646e+00



La classification ascendante hiérarchique (CAH) et le dendrogramme

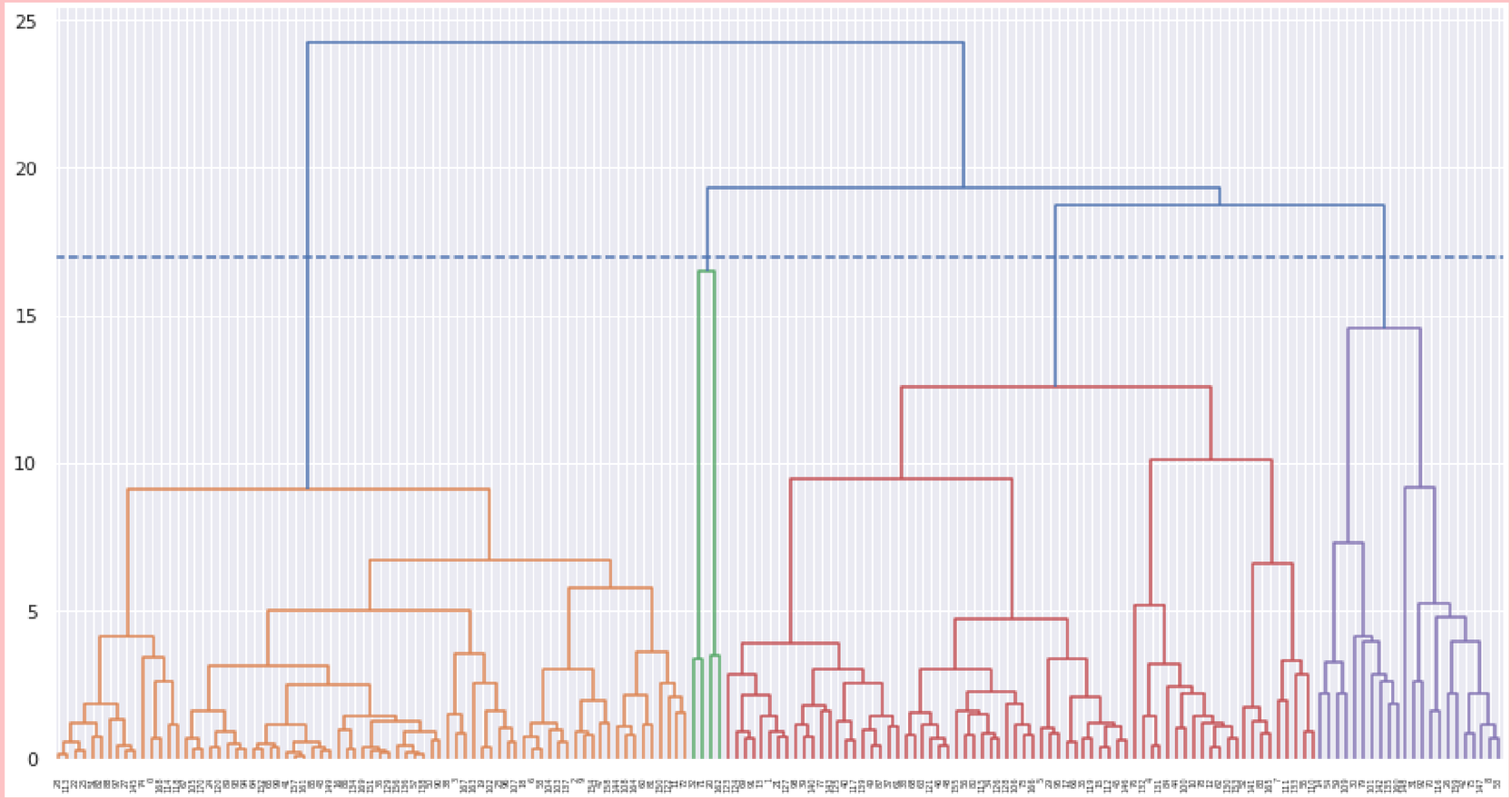


L'approche ascendante-- clustering agglomératif :

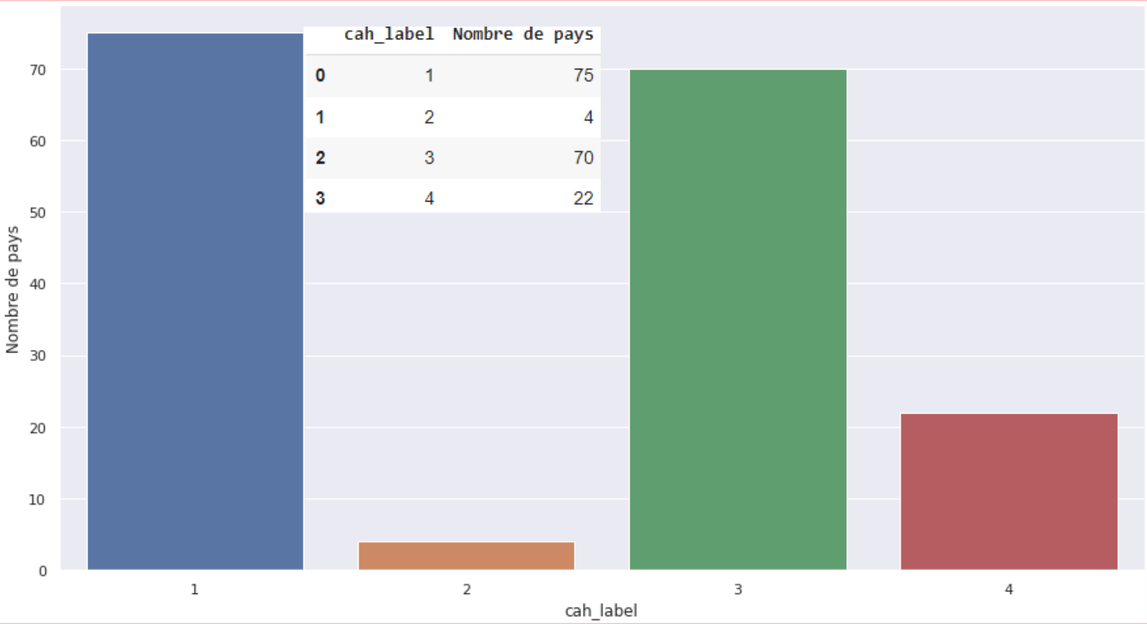
- Chaque point est un cluster. Chercher les deux clusters les plus proches.
- Agglomèrer les 2 clusters en un seul cluster.
- Répète les étapes ci-dessus jusqu'à ce que tous les points soient regroupés en un seul grand cluster.

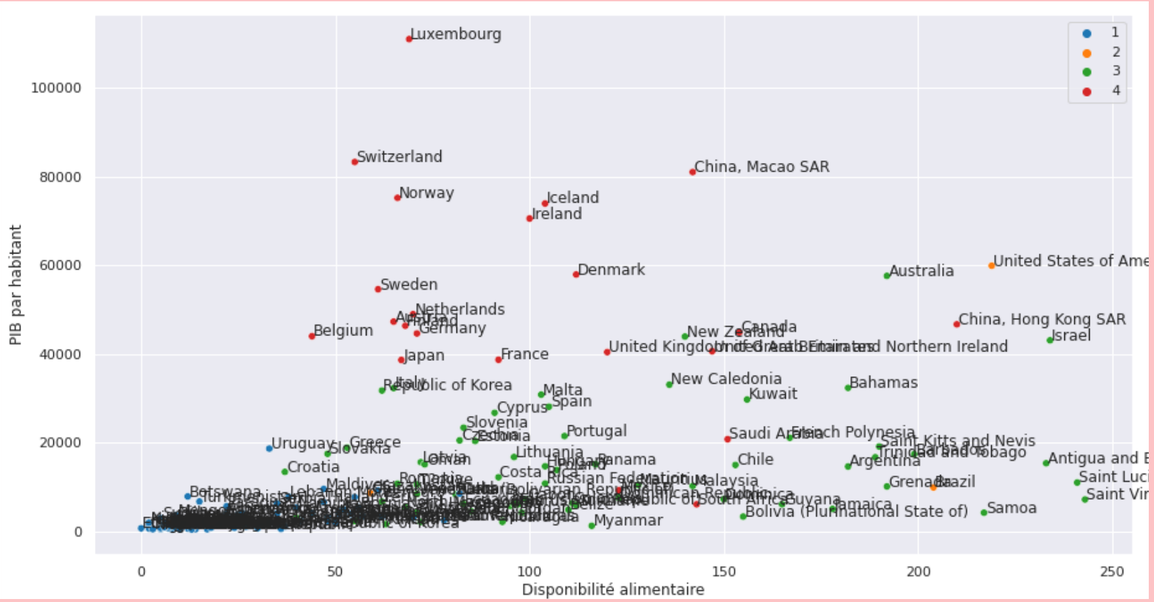
Dendrogrammes

Choisir 4 comme le nombre de clusters



Interprétation des 4 groupes par CAH





Groupe1: Pays sous-développé

Les pays avec la moindre de population, PIB est en bas, stabilité politique est unstable. Il y a le moindre de disponibilité alimentaire et le moindre de disponibilité de protéines.

Groupe2: Grands pays

Les 4 grands pays avec beaucoup de population, ils sont très différents des autres.

Groupe3: Les pays moyennes

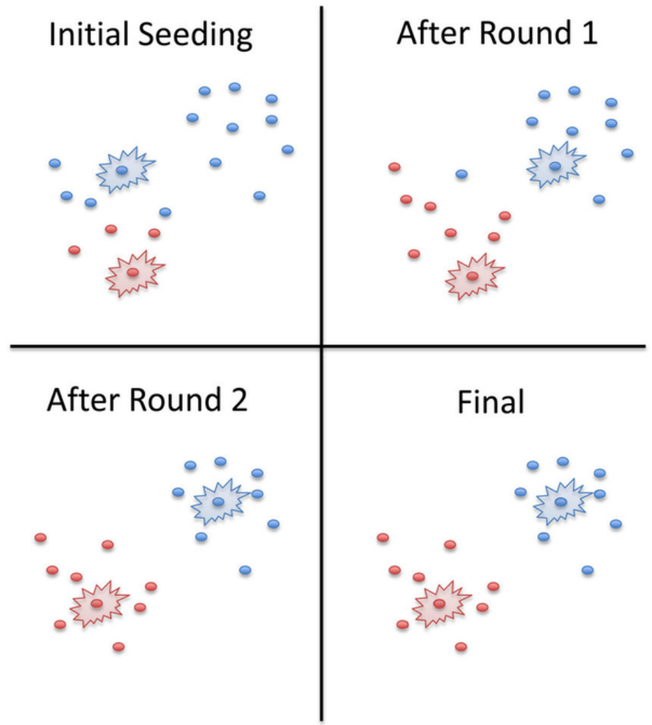
Ils sont les pays avec situation 'moyenne' pour tous les variable. Ils sont moyenne pour la disponibilité alimentaire, PIB par habitant, exportation, importation, stabilité politique, etc.

Groupe4: Les pays/région riches

Ils sont les pays riches, bien développés avec PIB par habitant les plus hautes, les prix de poulet les plus haute. Ils sont les pays/regions les plus proches avec stabilité politique stable.

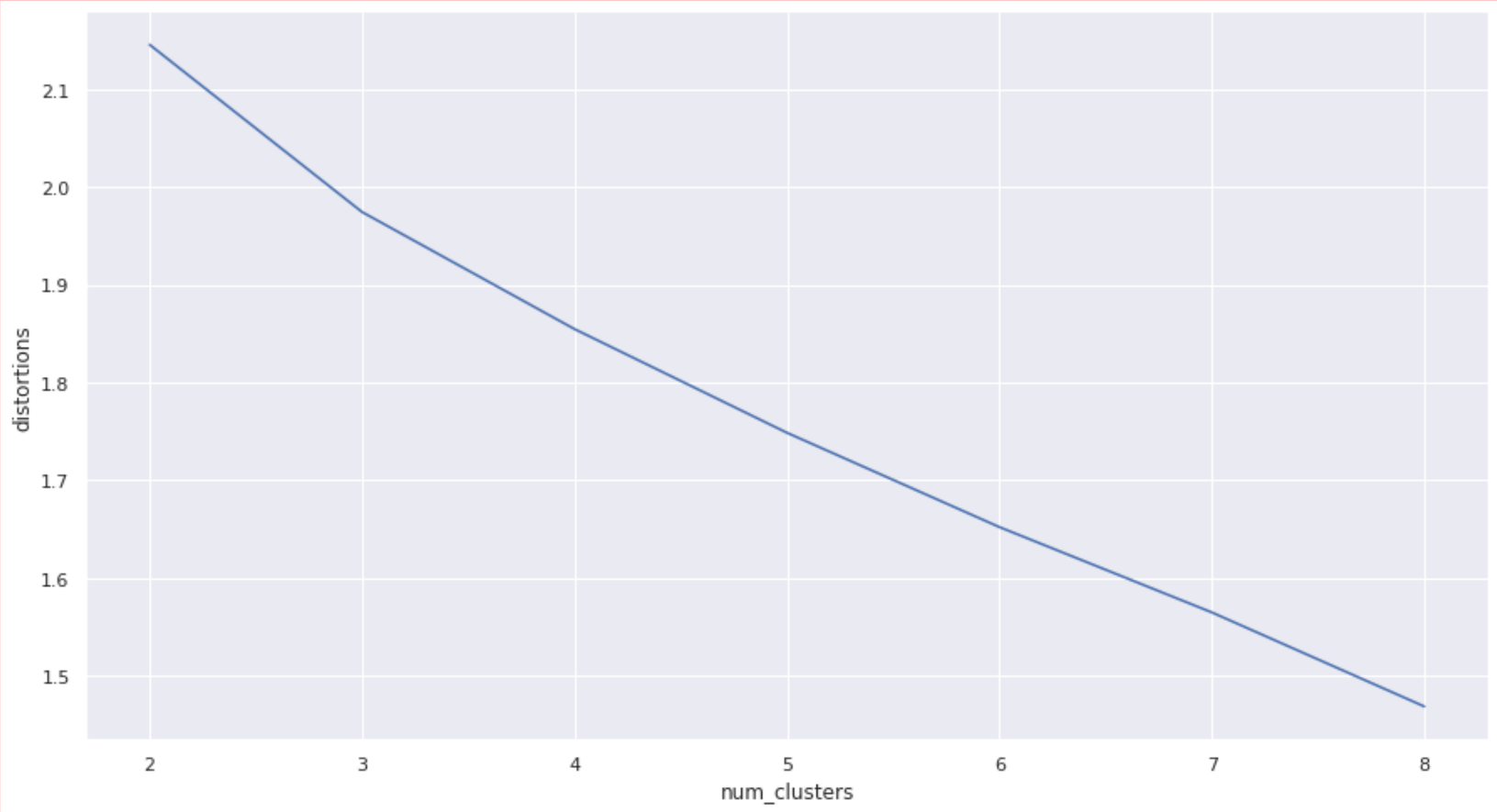
Kmeans

- Placer les centroïdes aléatoirement dans l'espace
- Prendre chaque point du nuage et lui associee le cluster du centroïde dont il est le plus proche. On obtient donc K groupes.
- Recalculer les centroïdes de chaque groupe quand les centroïdes bougent jusqu'à ce qu'ils ne bougent pas.



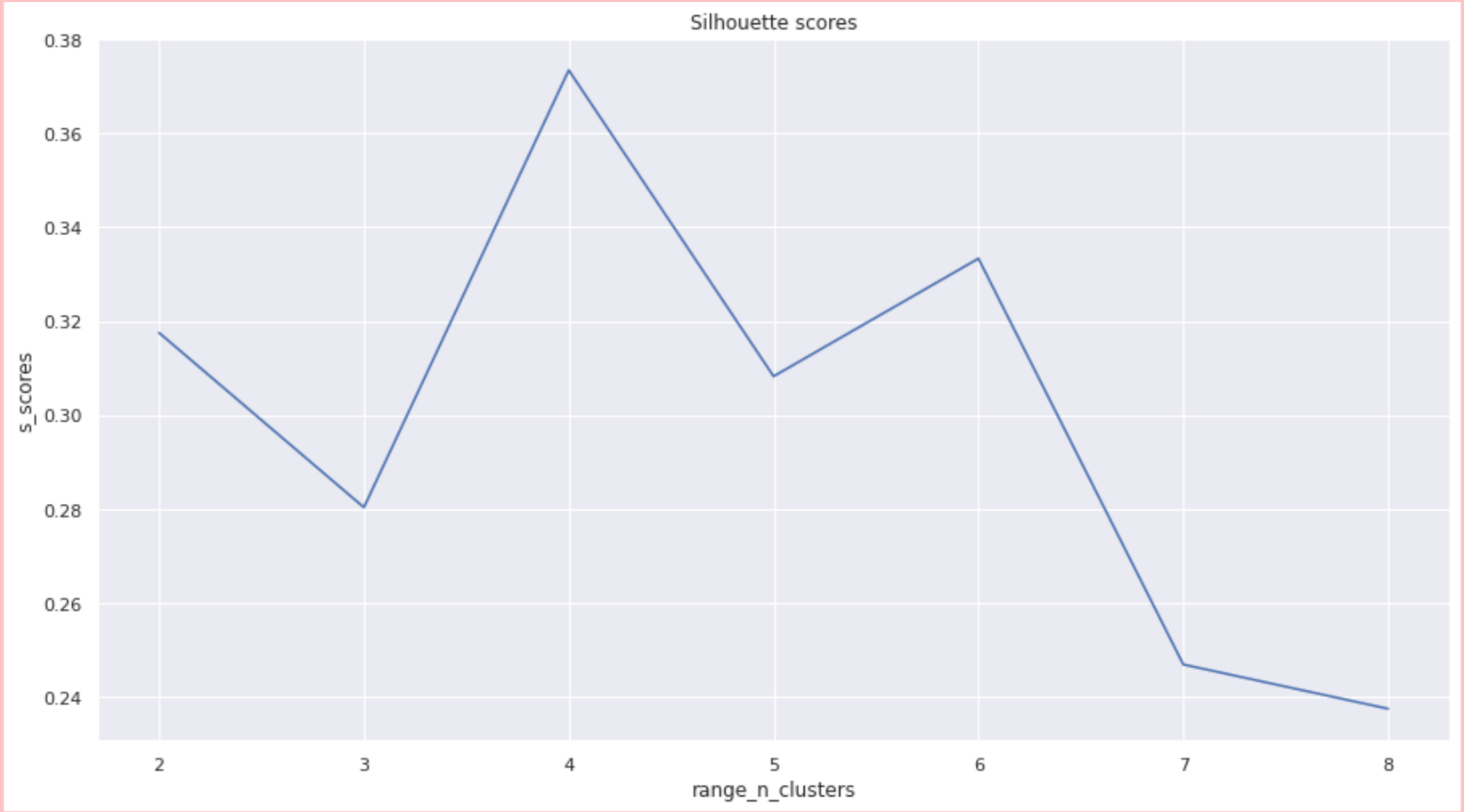
La méthode Elbow

On ne peut pas juger combien de groupes qu'on doit choisir par cette méthode.

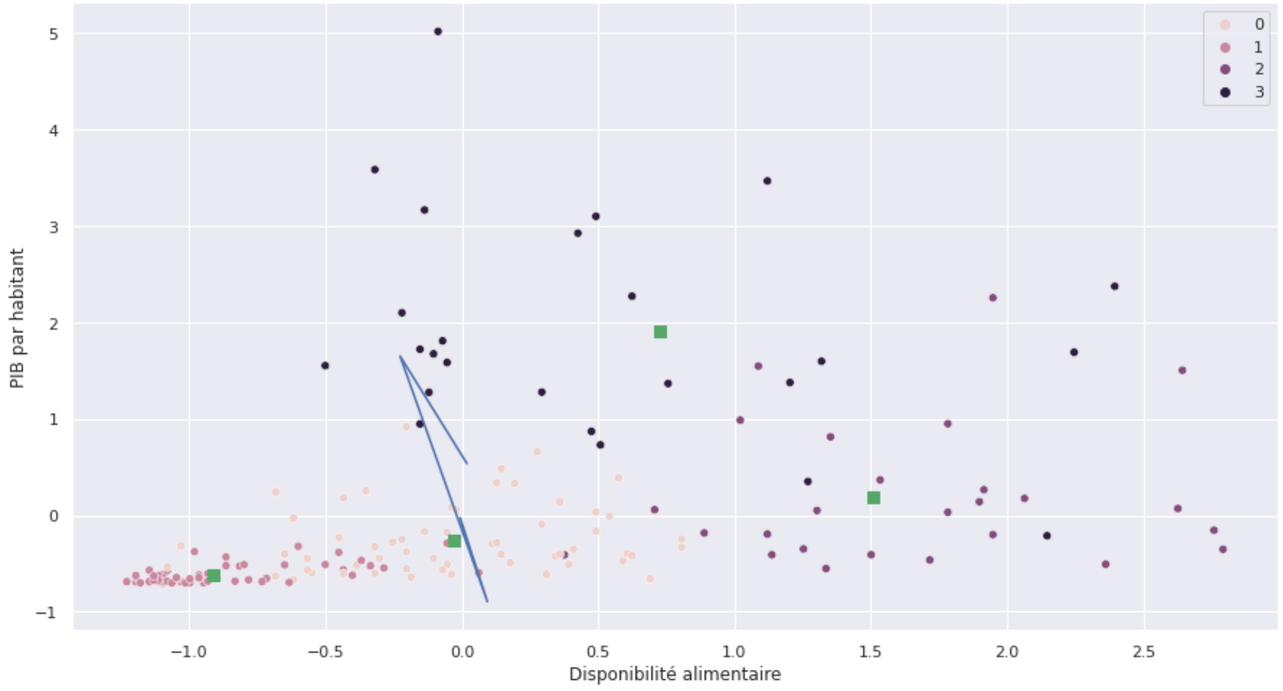


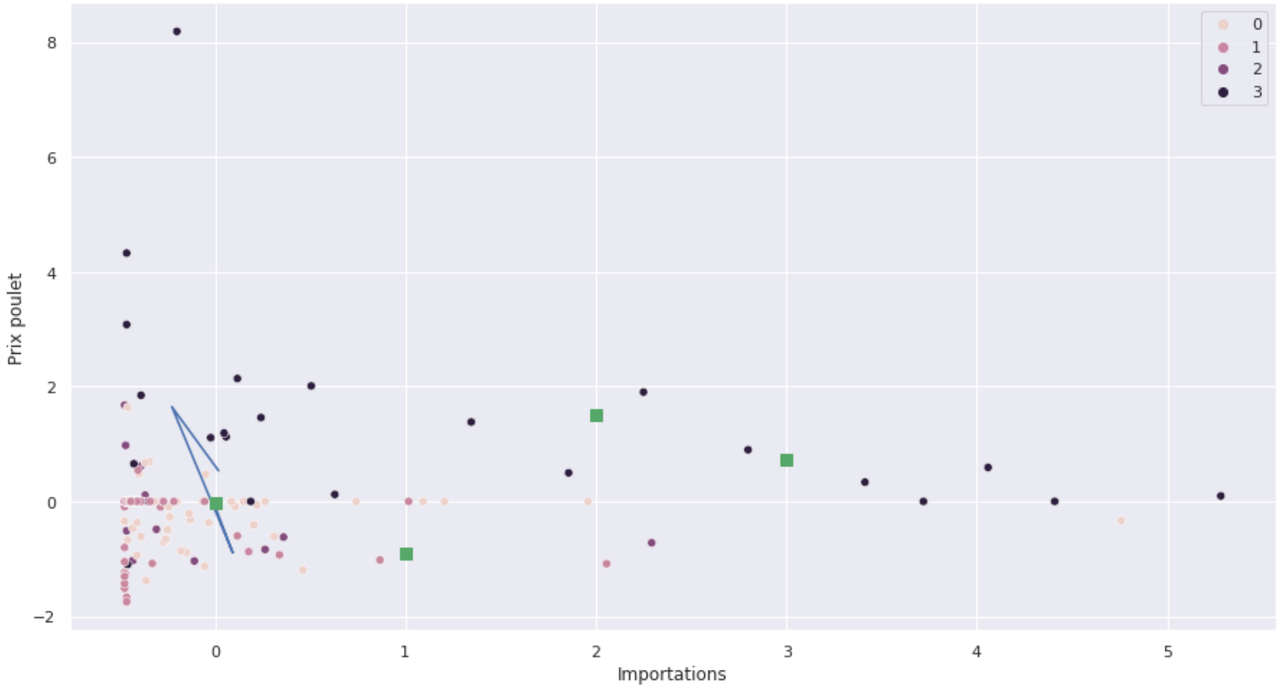
Silhouette analyse

Choisir 4 comme le nombre de clusters

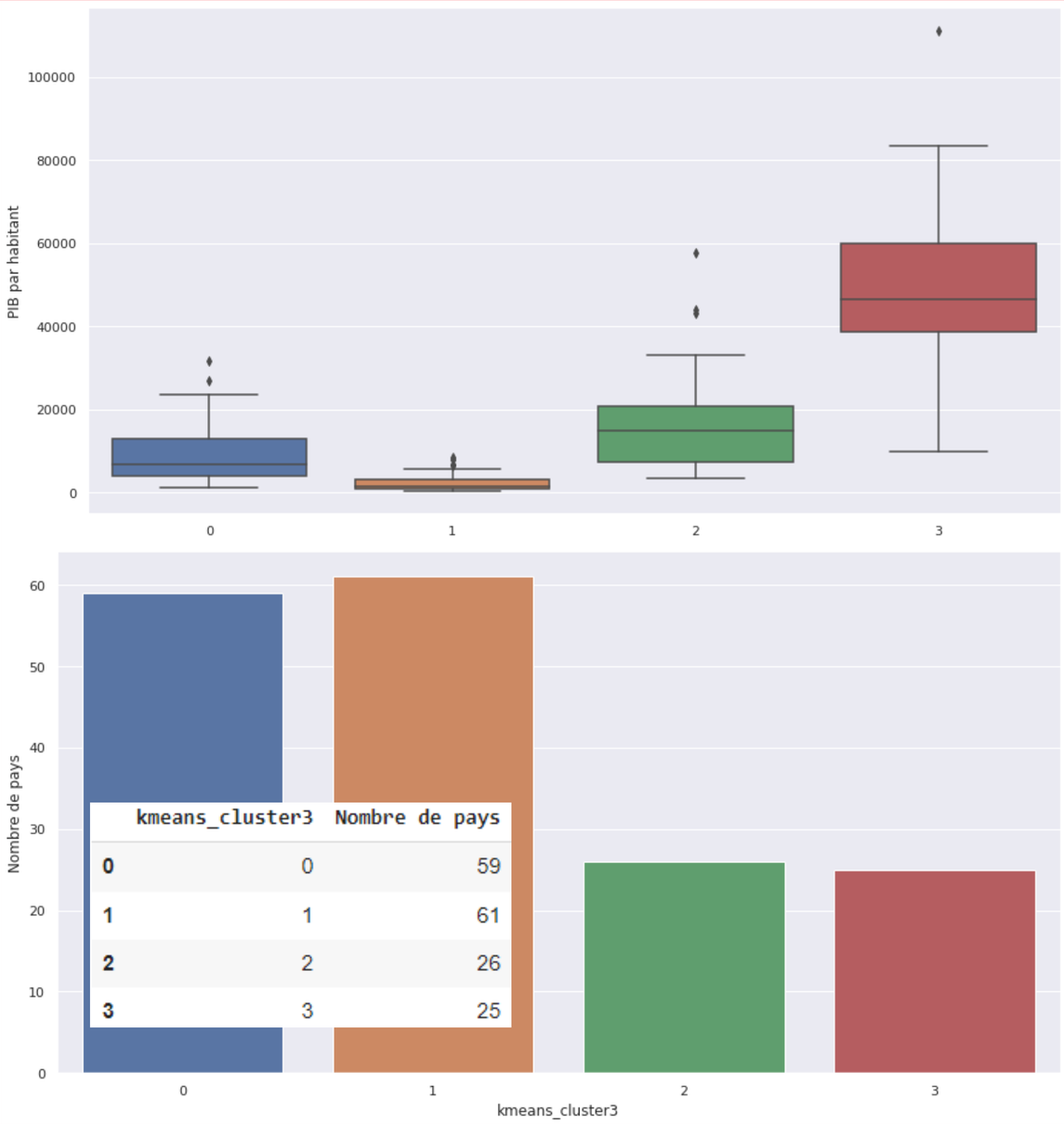


Présenter les centroîds





Interprétation de les 4 groupes par kmeans



	kmeans_cluster3	Nombre de pays
0	0	59
1	1	61
2	2	26
3	3	25

Groupe Pays sous-développé
Pour cette partie de pays; le PIB par habitant est en bas, stabilité politique est unstable. Il y a le moindre de disponibilité alimentaire et le moindre de disponibilité de protéines. Le prix de poulet est en bas mais la population est en haute.

Groupe Puissance agricole
Les pays avec disponibilité alimentaire et disponibilité de protéines en haute, ils exportent beaucoup de poulet aux autres pays. Ils ne sont pas notre cibles parce qu'il y a longs distances aussi.

Groupe Les pays moyennes
Ils sont les pays avec les moindre de population (pas grands pays). Ils sont moyenne pour la disponibilité alimentaire, PIB par habitant, exportation, importation, stabilité politique, etc.

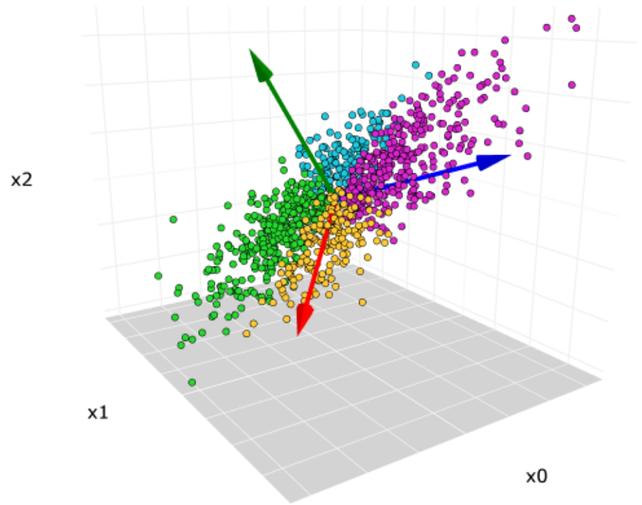
Groupe Les pays/région riches
Ils sont les pays riches, bien développés avec PIB par habitant les plus hautes, les prix de poulet les plus haute. Ils sont les pays/regions les plus proches avec stabilité politique stable. La grande partie des pays sont en Euroupe. Il y a aussi les grandes avec beaucoup des populations et les grands marchés.

L'ordres des groups peut changer à cause de la récalculation.

ACP

Analyse en composantes principales:

- Transformer des variables liées entre elles (dites « corrélées » en statistique) en nouvelles variables décorrélées les unes des autres.
- Elle permet au statisticien de résumer l'information en réduisant le nombre de variables.



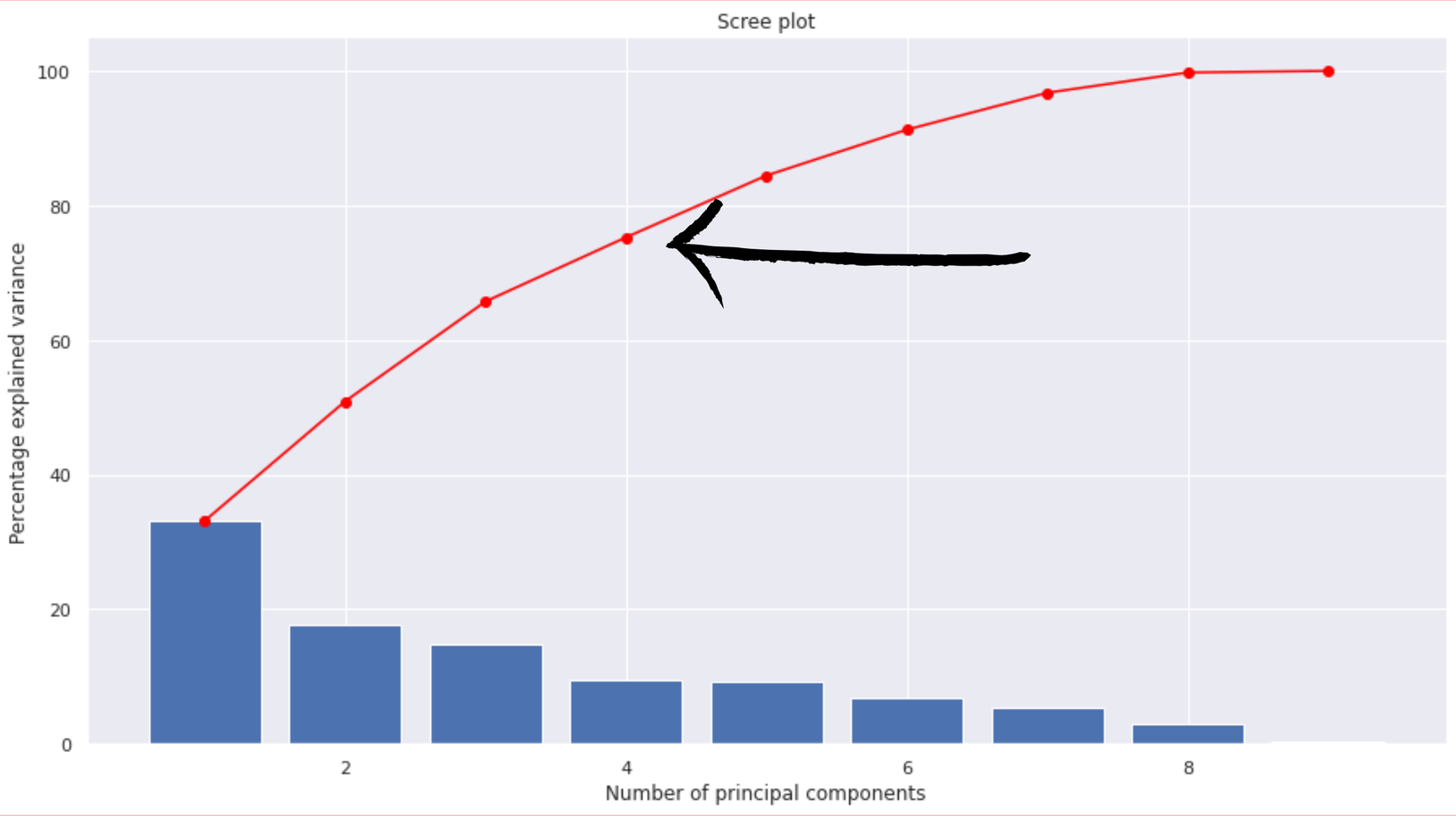
Explained variance ratio

```
pca.explained_variance_ratio_.round(2)
```

```
array([0.33, 0.18, 0.15, 0.1 , 0.09, 0.07, 0.05, 0.03, 0.  ])
```

```
[53] pca.explained_variance_ratio_.cumsum()
```

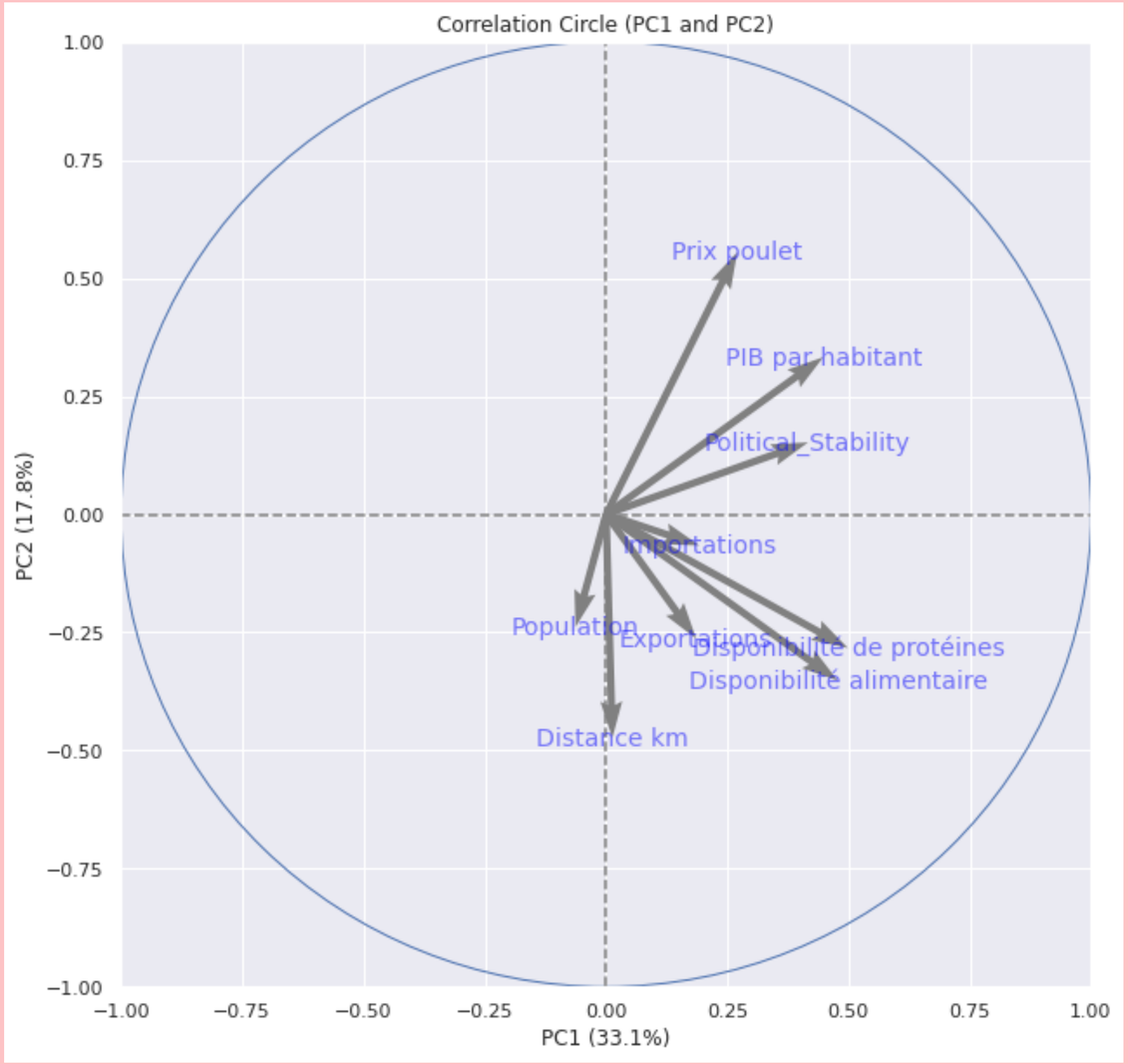
```
array([0.33140626, 0.5089476 , 0.65737281, 0.75265492, 0.84449873,
       0.91267085, 0.96751197, 0.99766941, 1.        ])
```

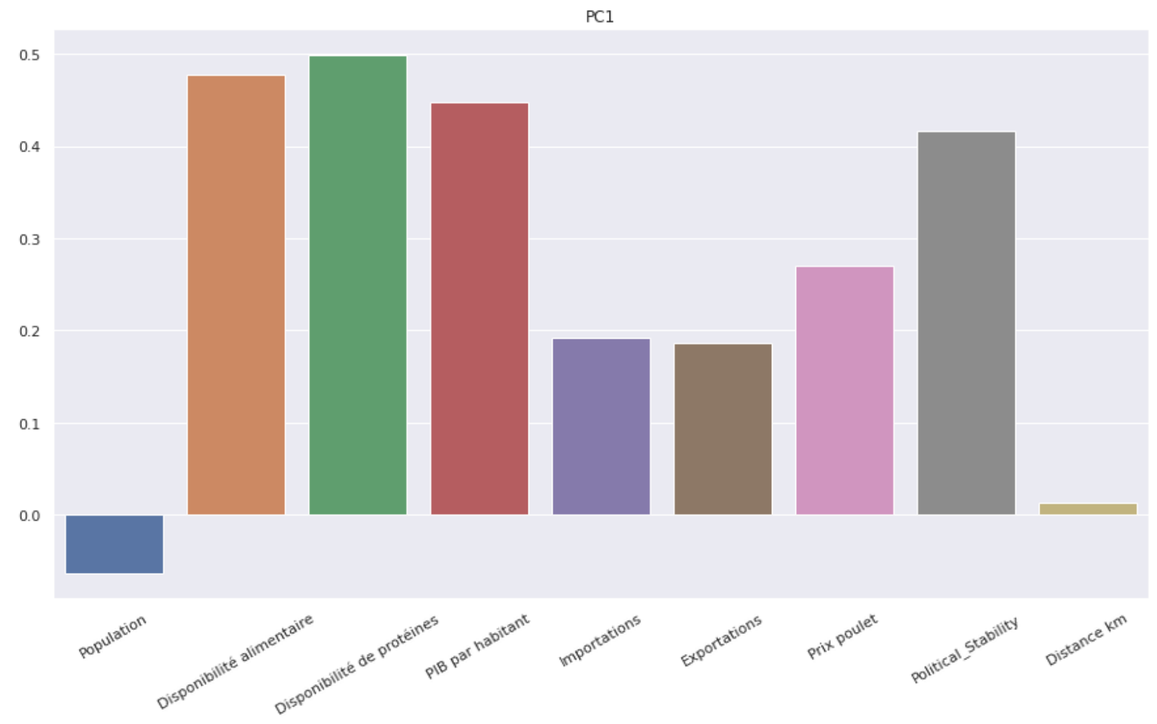


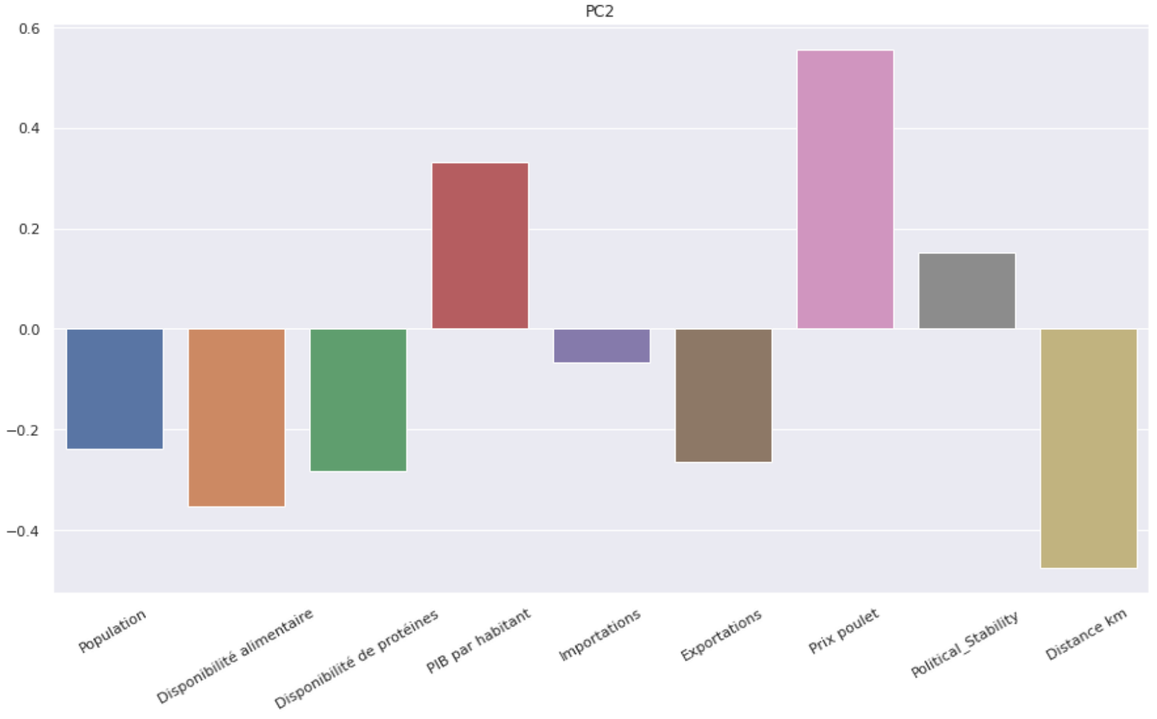
Choisir 4 comme le nombre de composantes principaux parce qu'il presente presque 80% de information (basé sur la loi de Pareto)

PC1, PC2

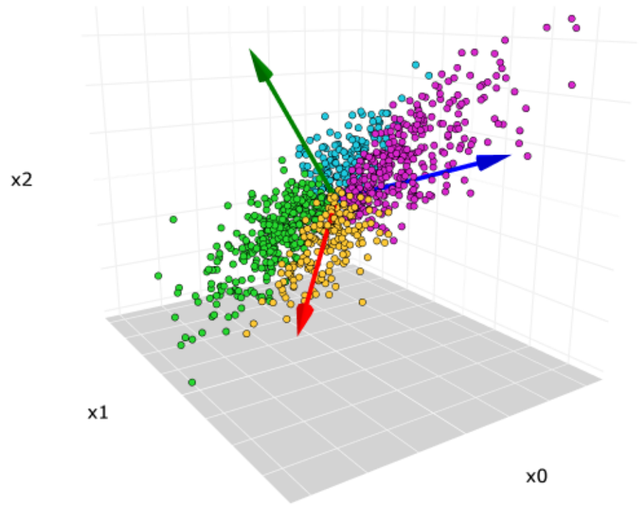
- Comprendre les liens entre les variables
- Choisir les PC1 et PC2, qui sont les 2 PC les plus importants et pertinentes





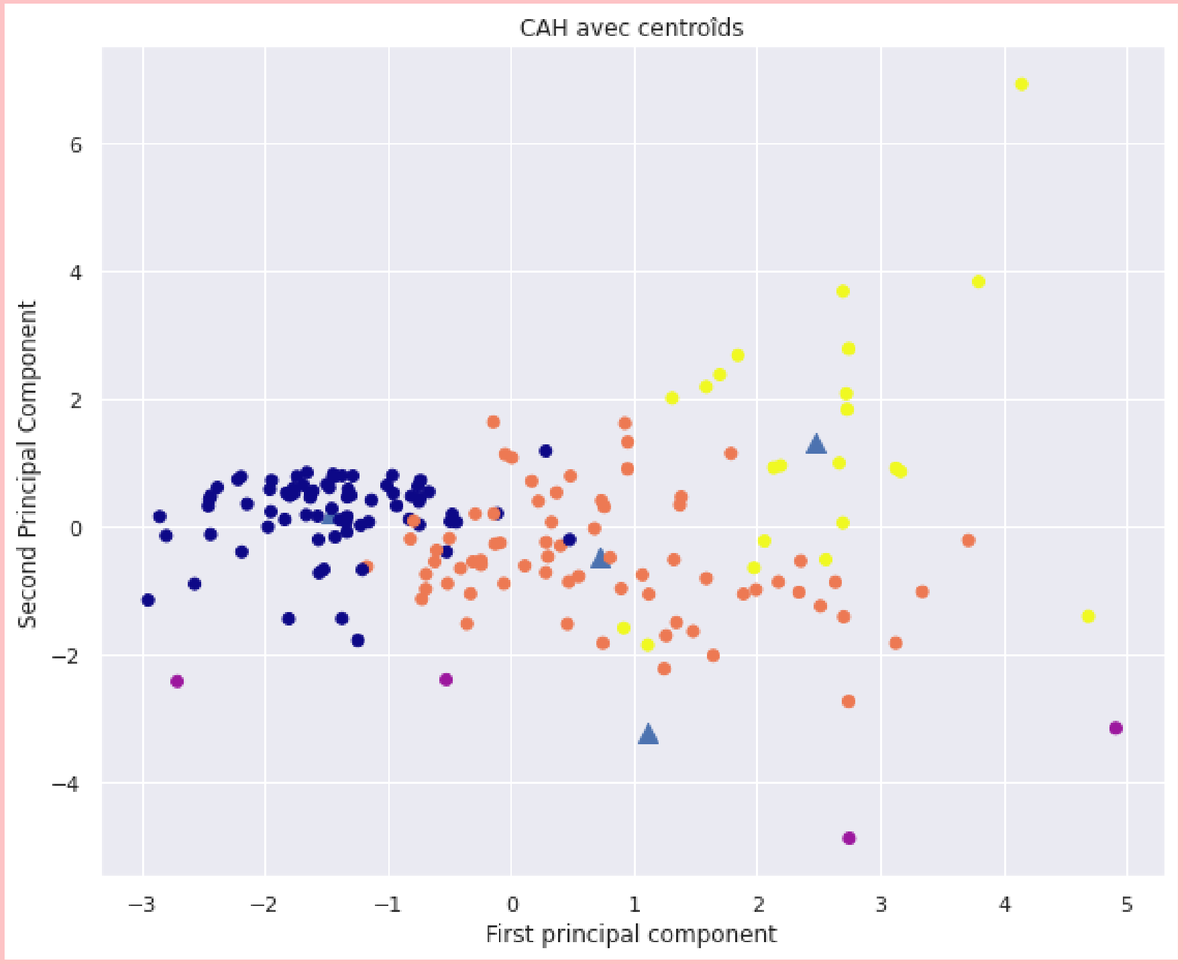


Résultats et visualisation avec ACP

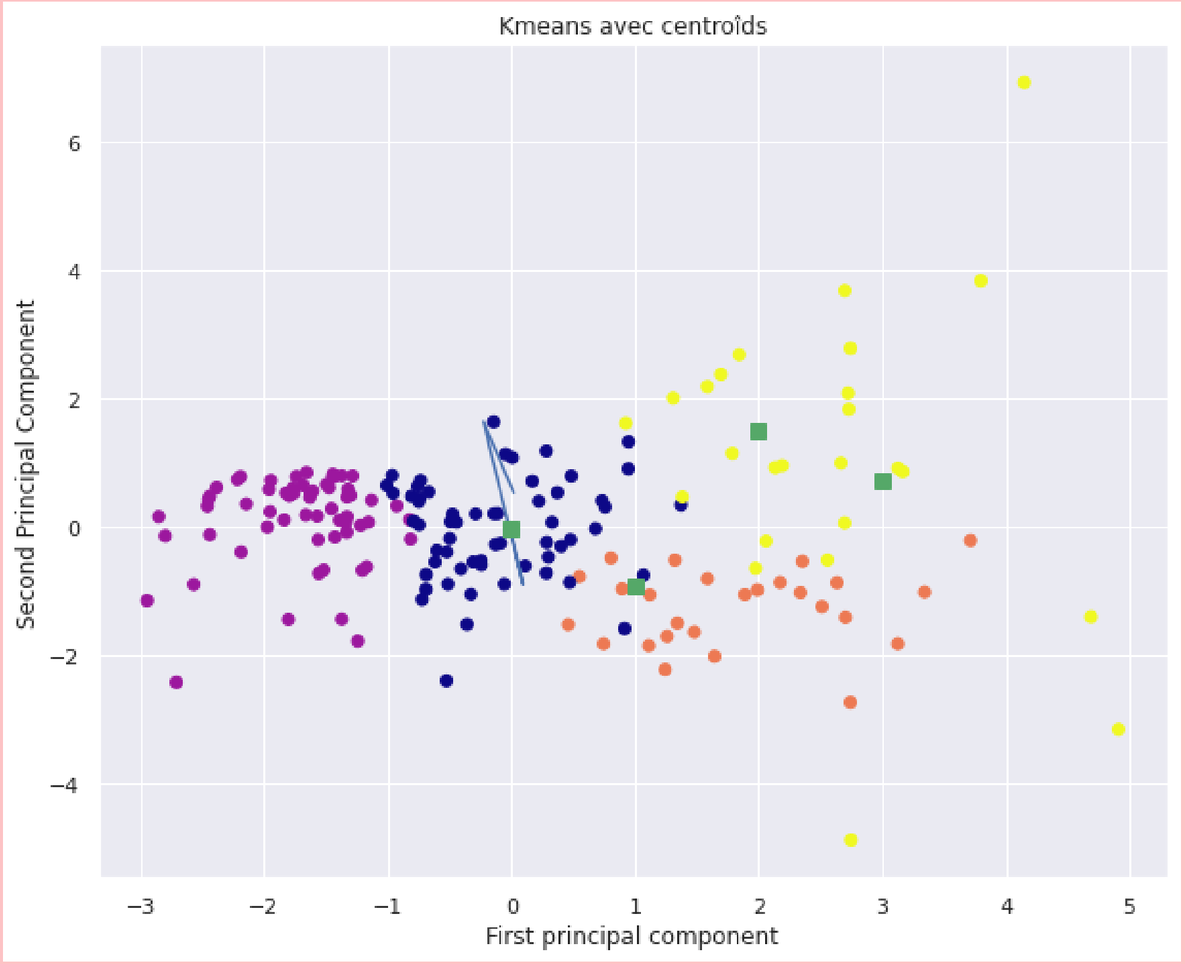


Comparer CAH, Kmeans

CAH

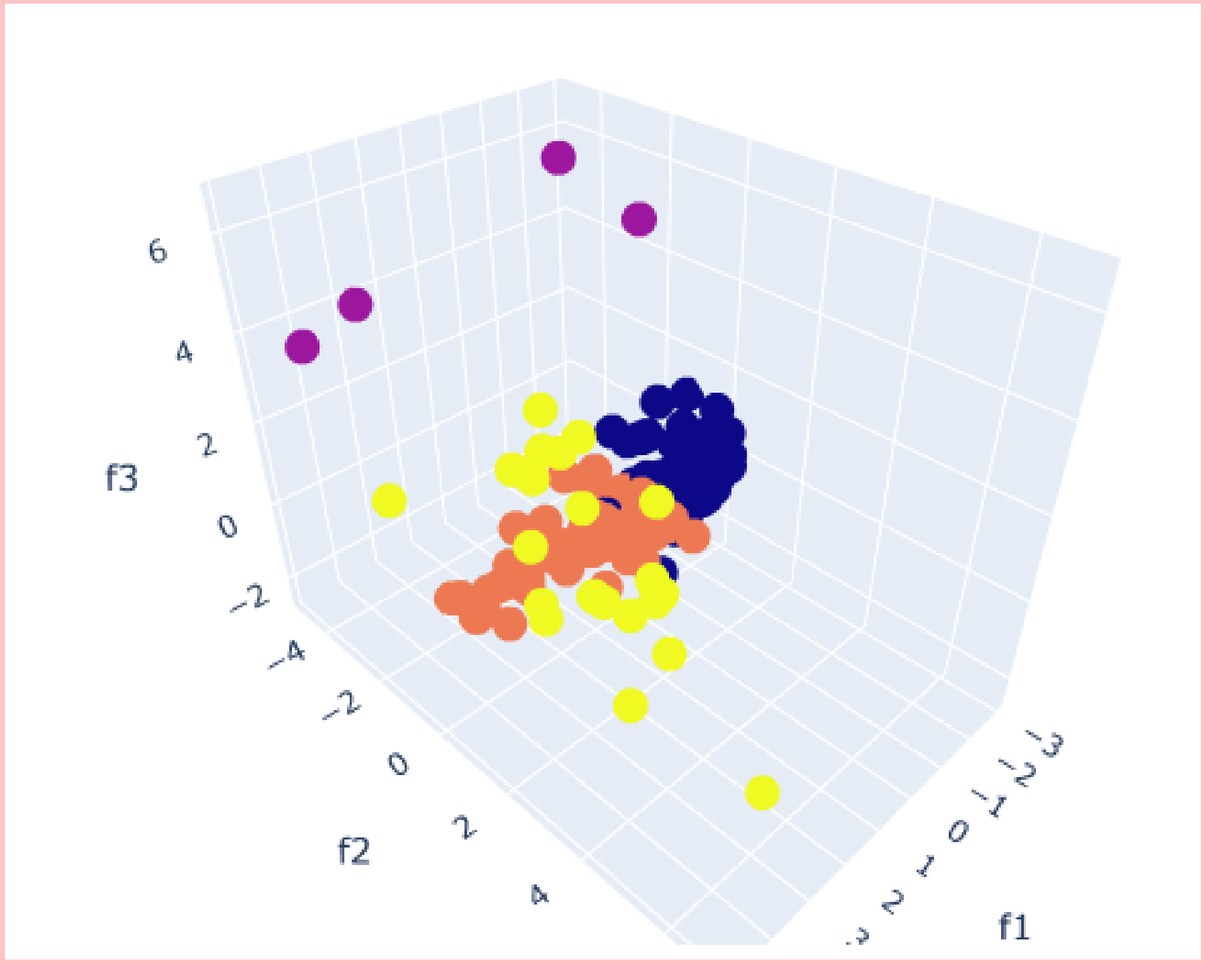


Kmeans

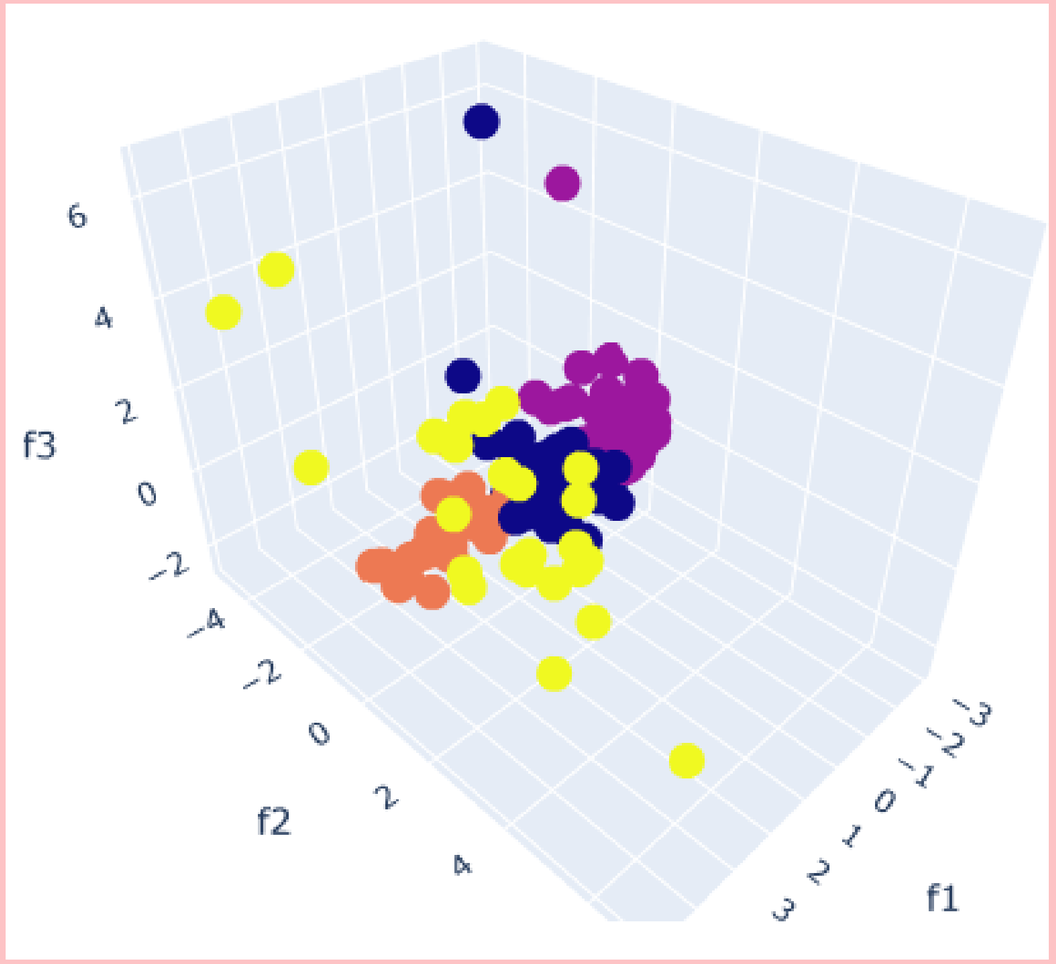


Comparer CAH, Kmeans 3D

CAH

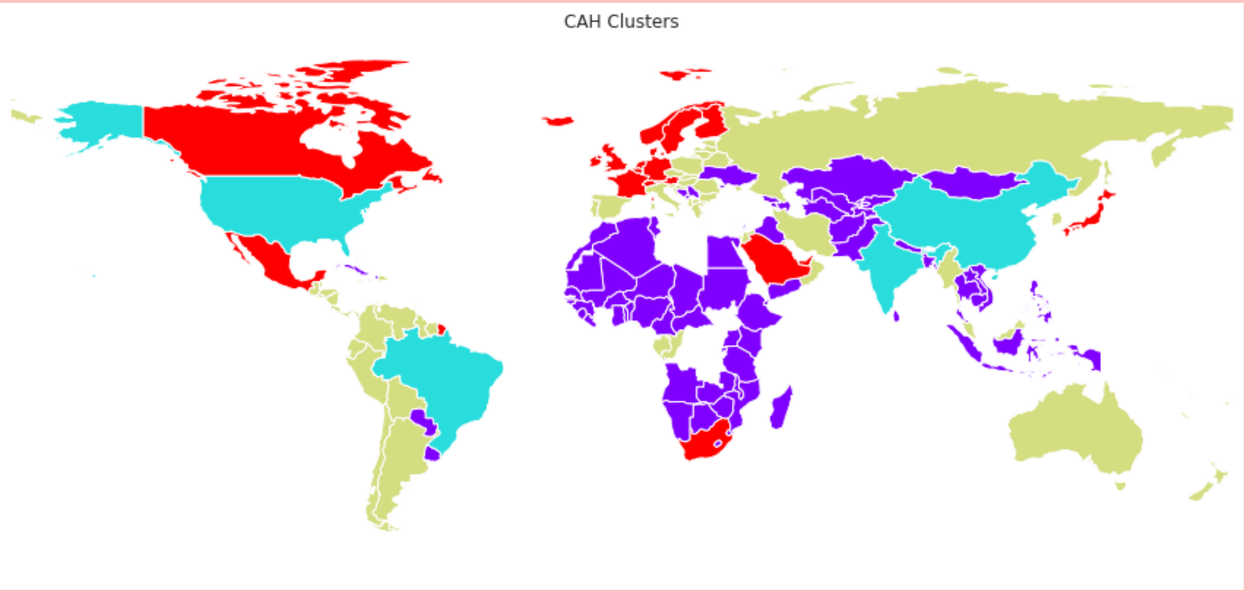


Kmeans

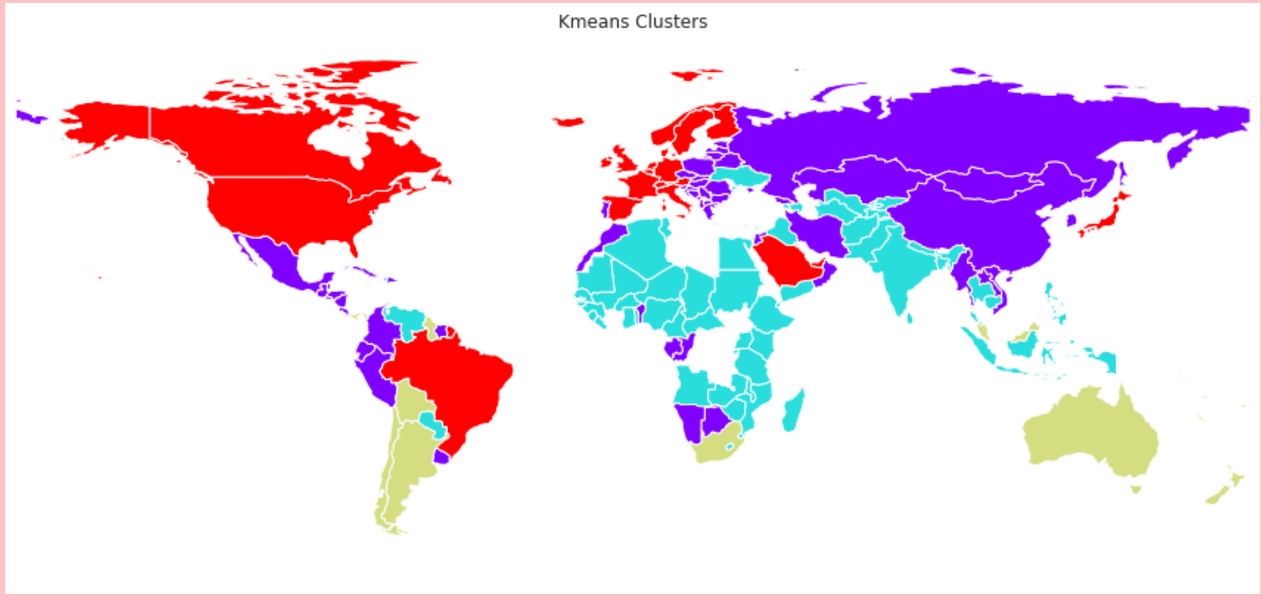


Comparer CAH, Kmeans

CAH

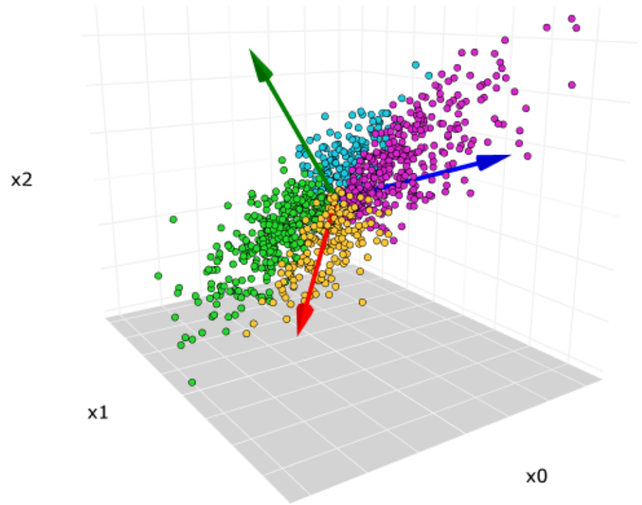


Kmeans



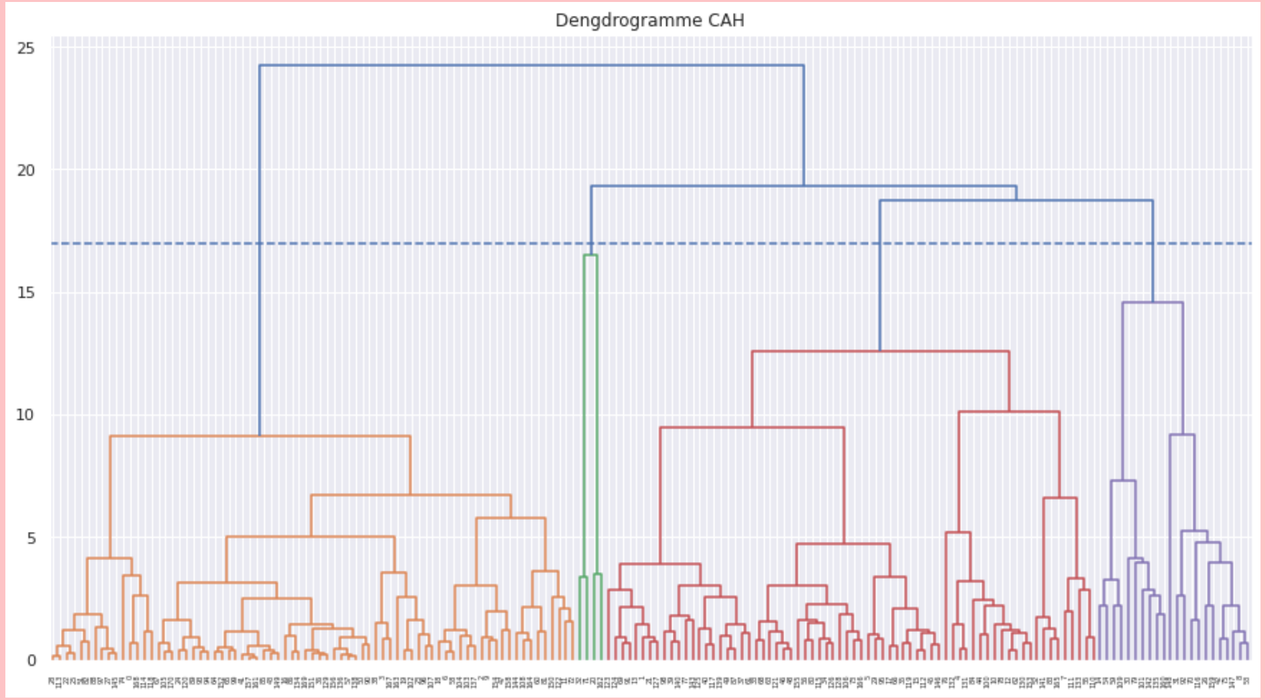
+ Partie supplémentaire:

Recalculer les clusters après ACP avec CAH et kmeans

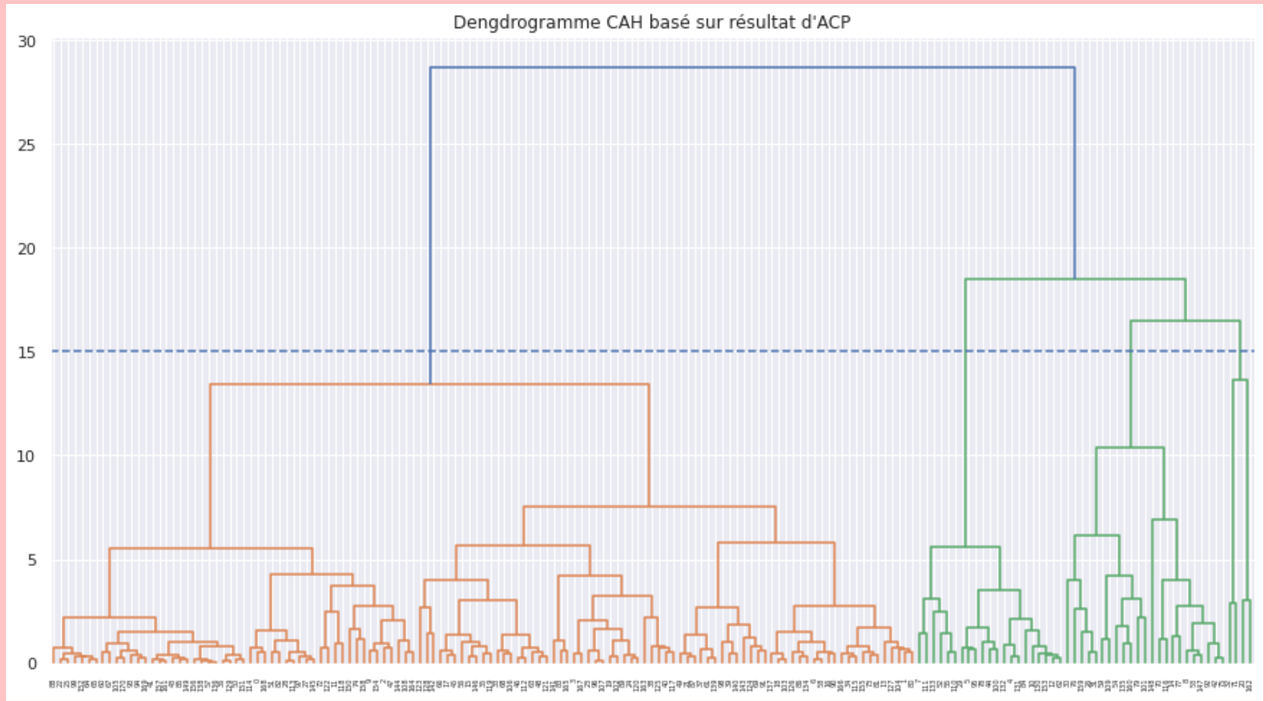


Partie supplémentaire: Recalculation après ACP-- CAH

Avant ACP

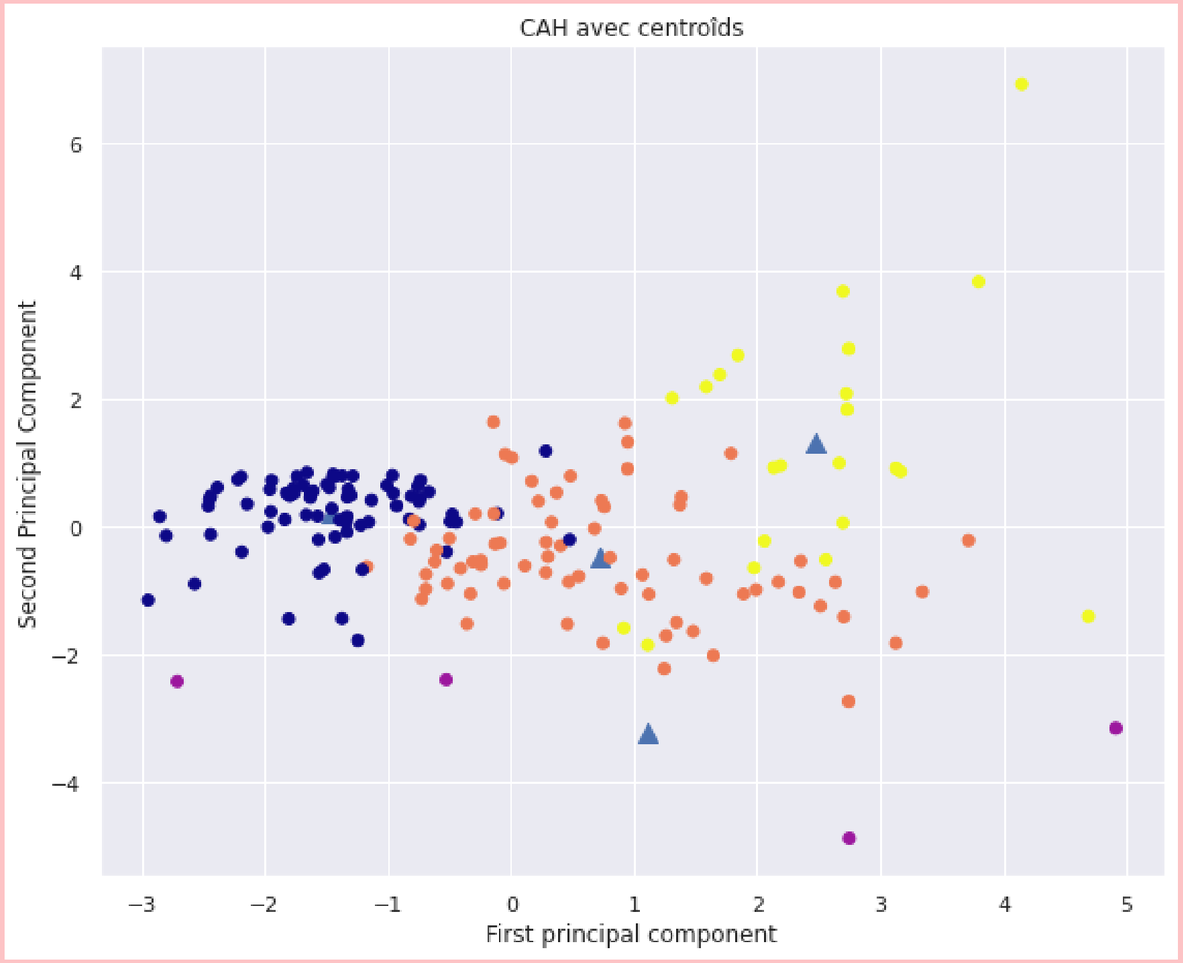


Après ACP

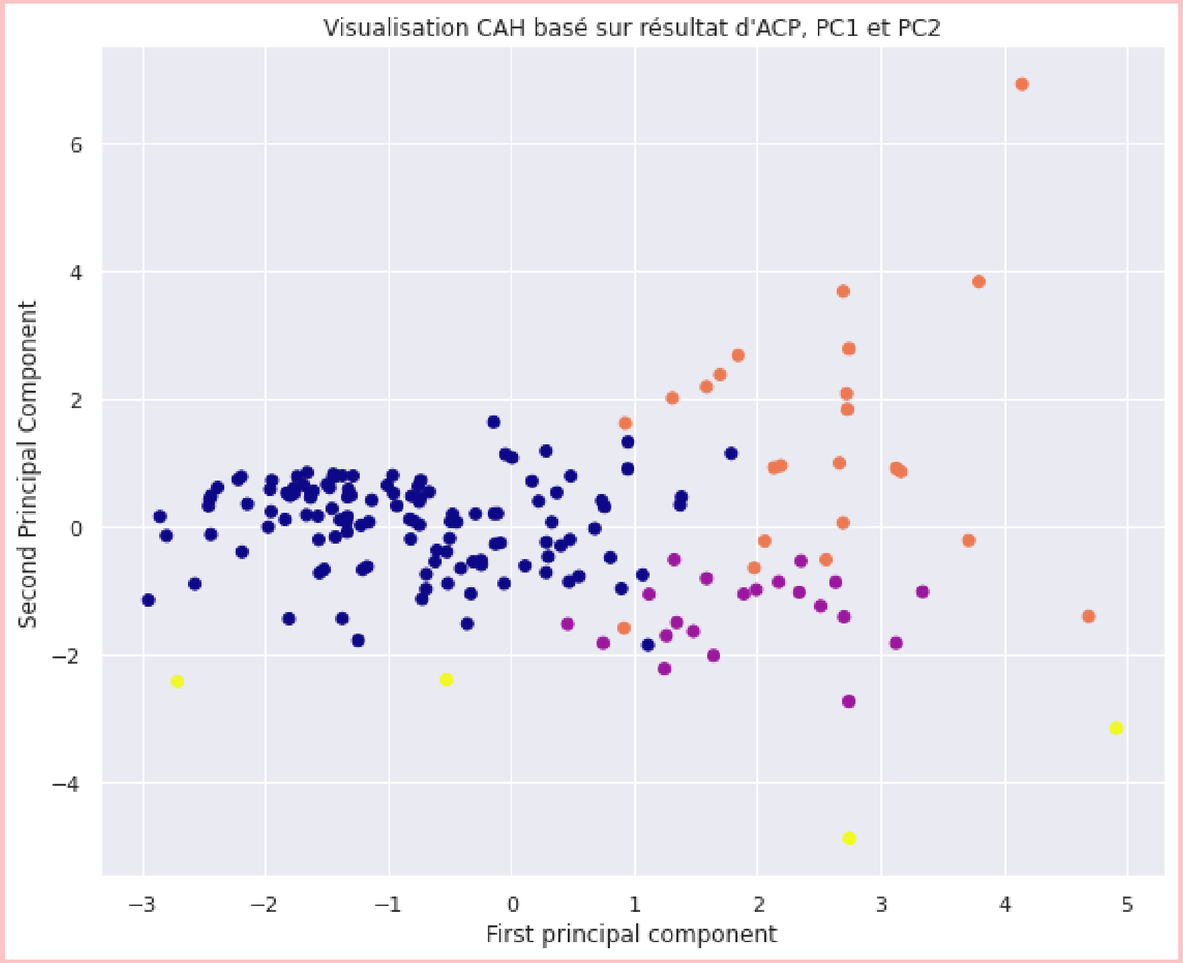


Partie supplémentaire: Recalculation après ACP-- CAH

Avant ACP

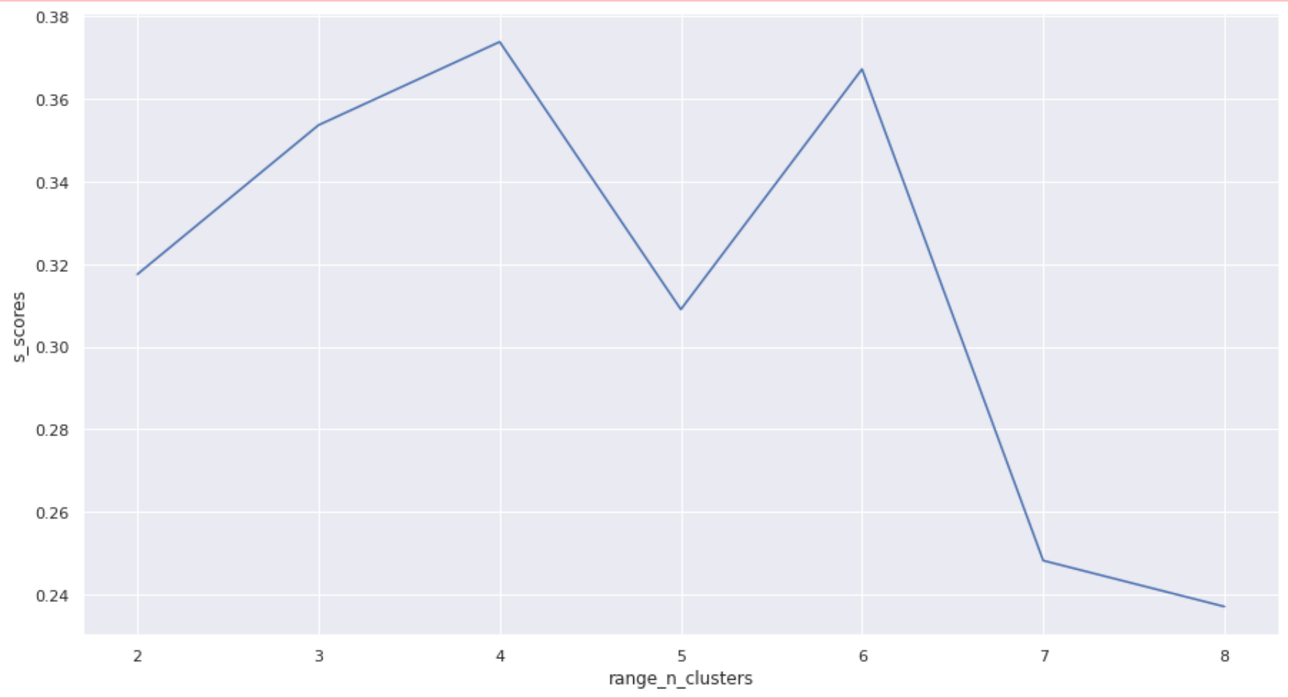


Après ACP

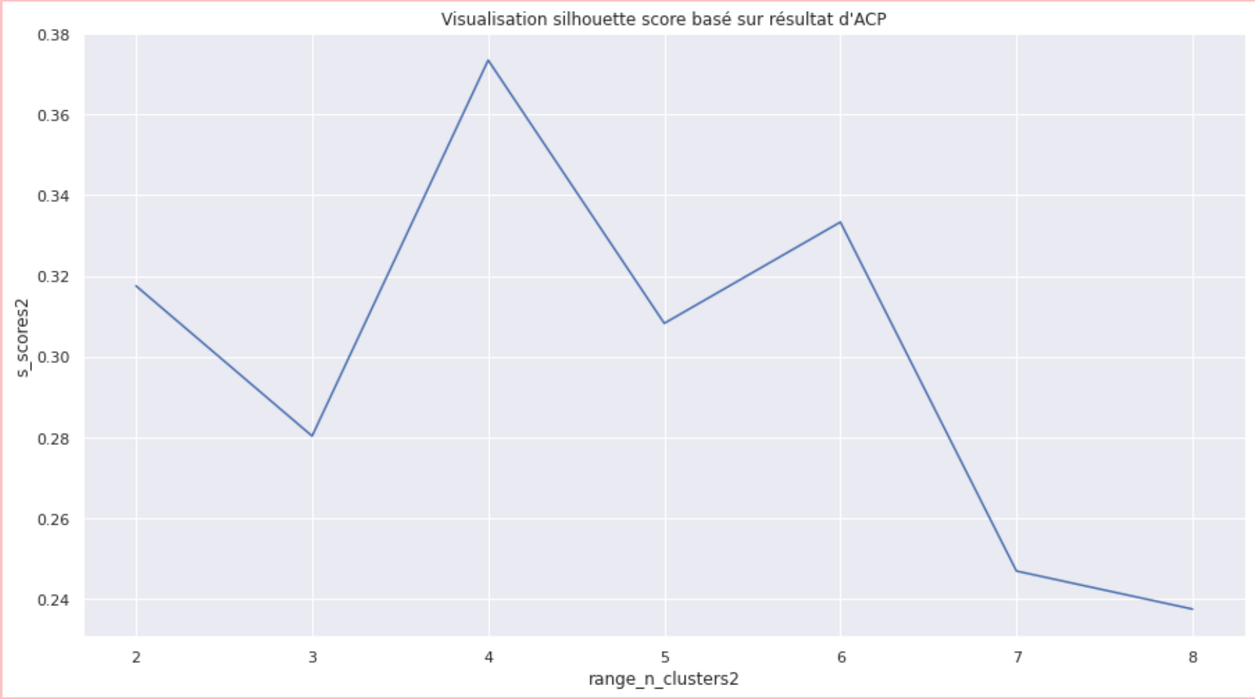


Partie supplémentaire: Recalculation après ACP-- Kmeans

Avant ACP

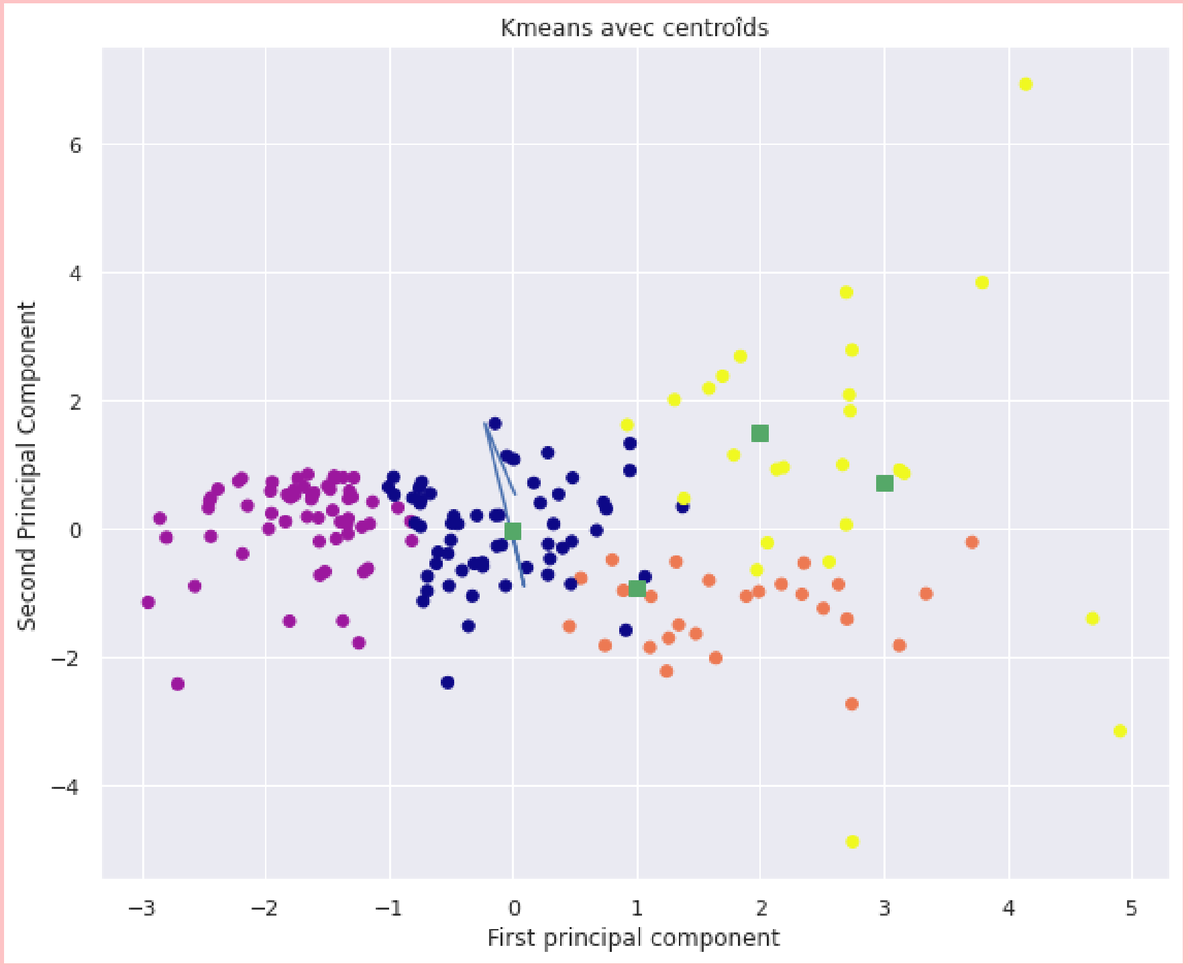


Après ACP

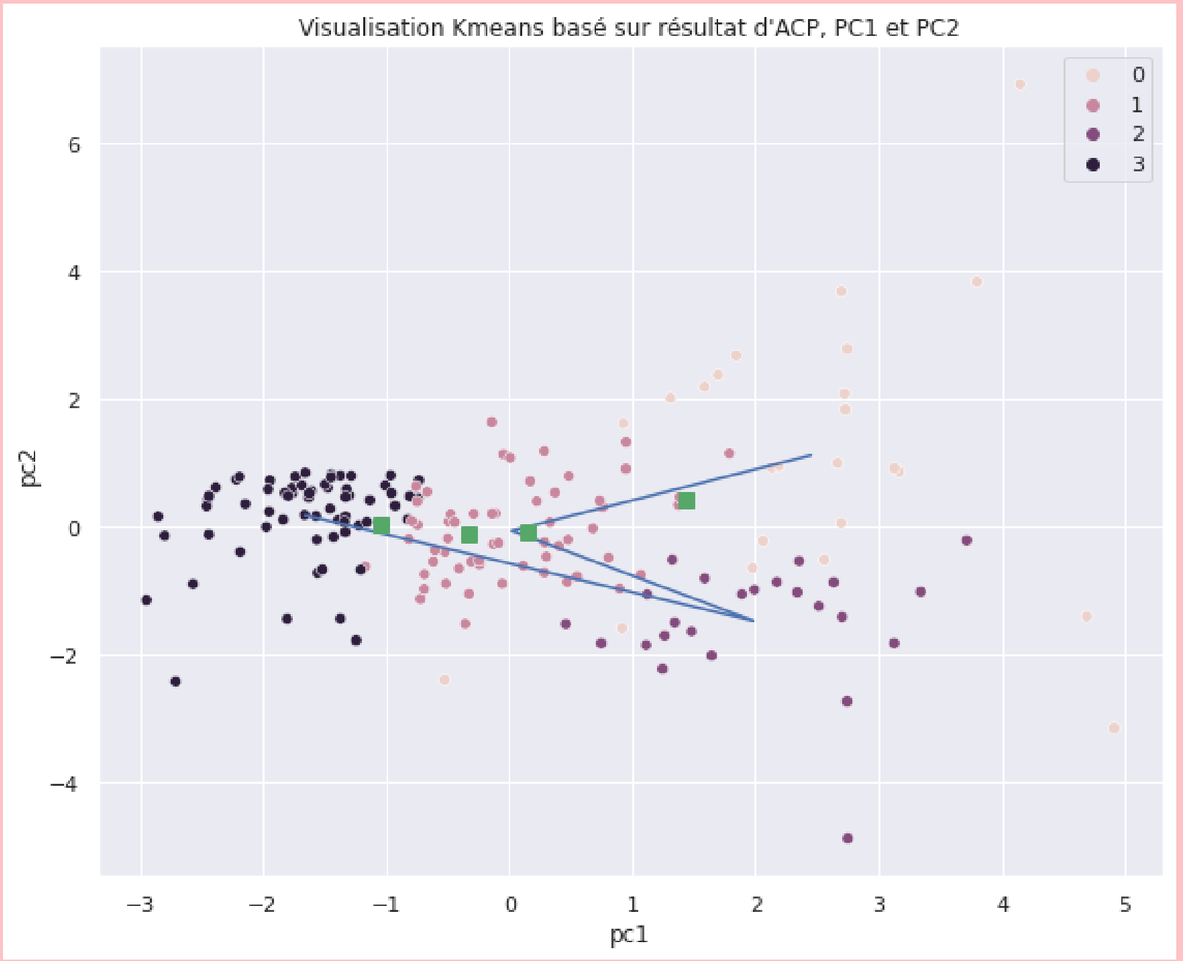


Partie supplémentaire: Recalculation après ACP-- Kmeans

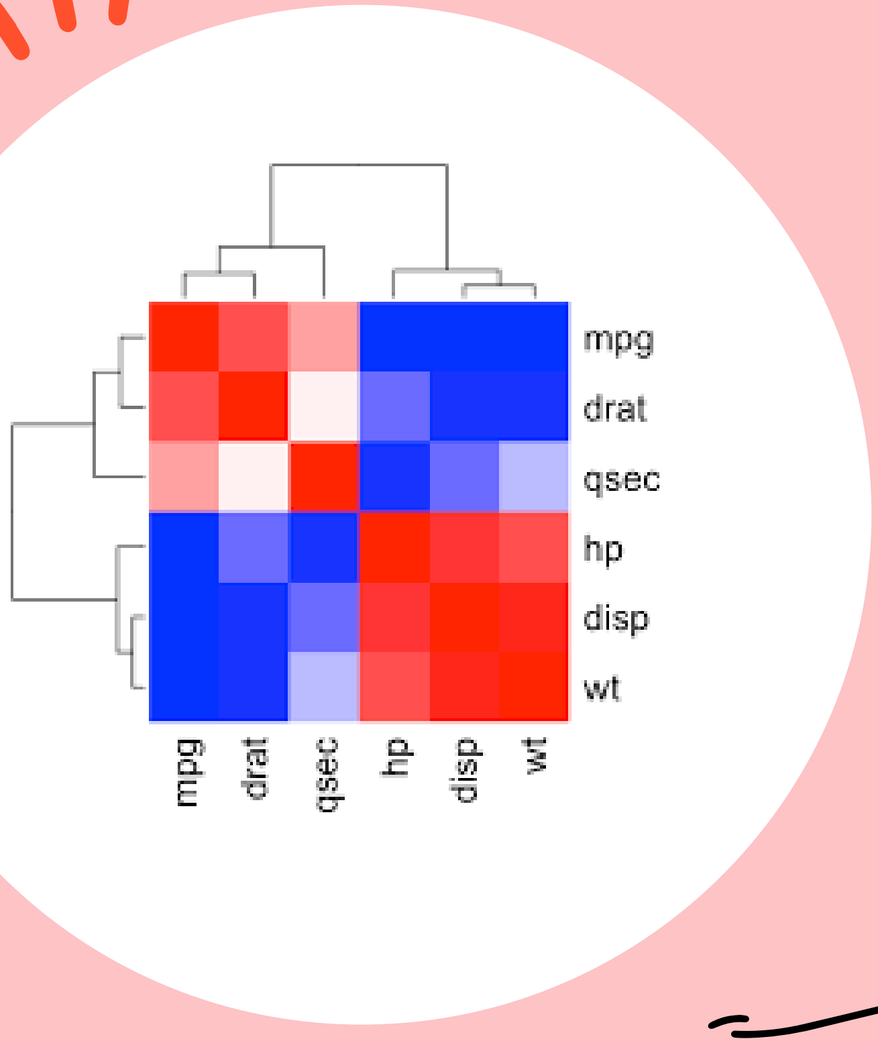
Avant ACP



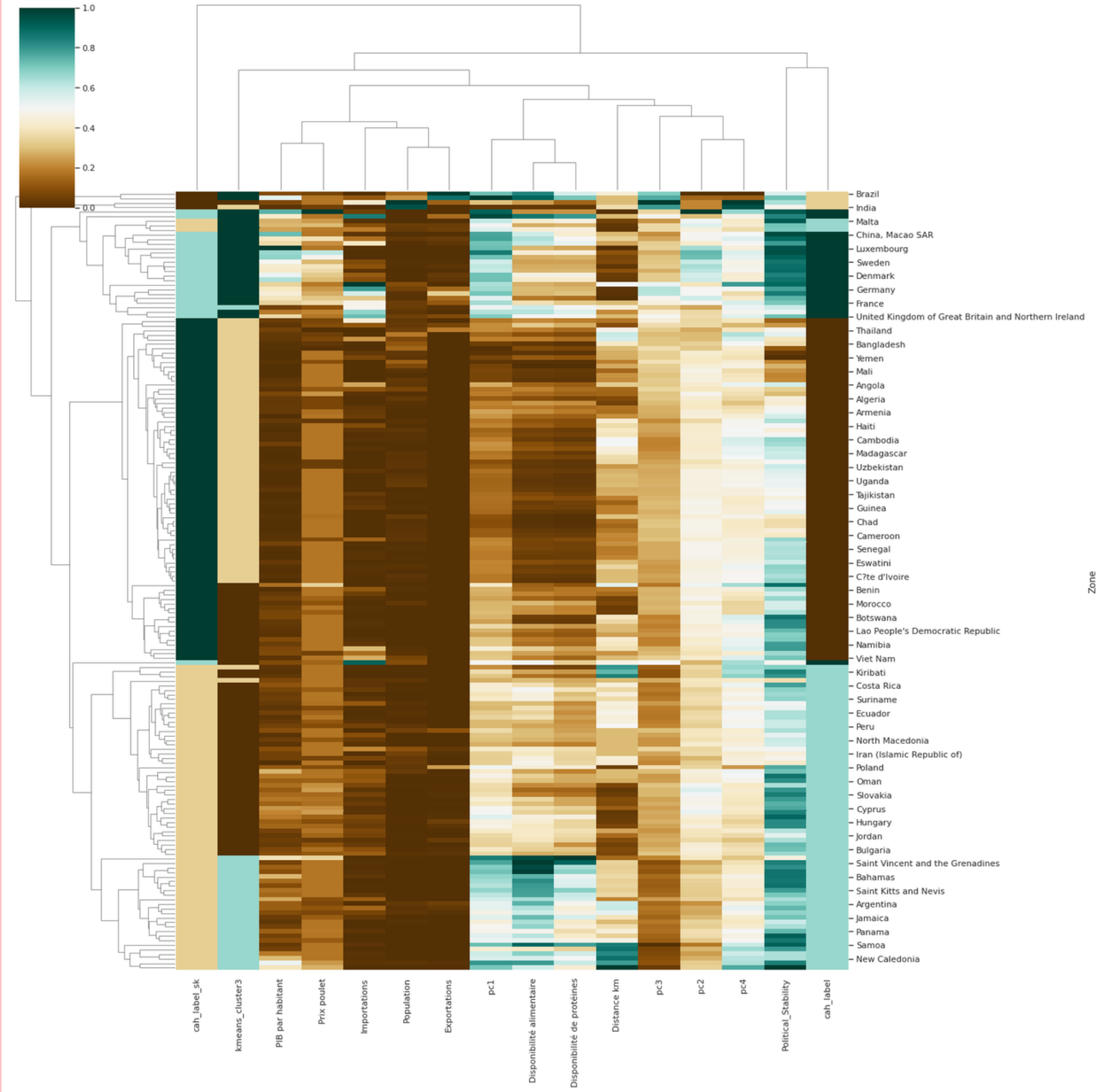
Après ACP



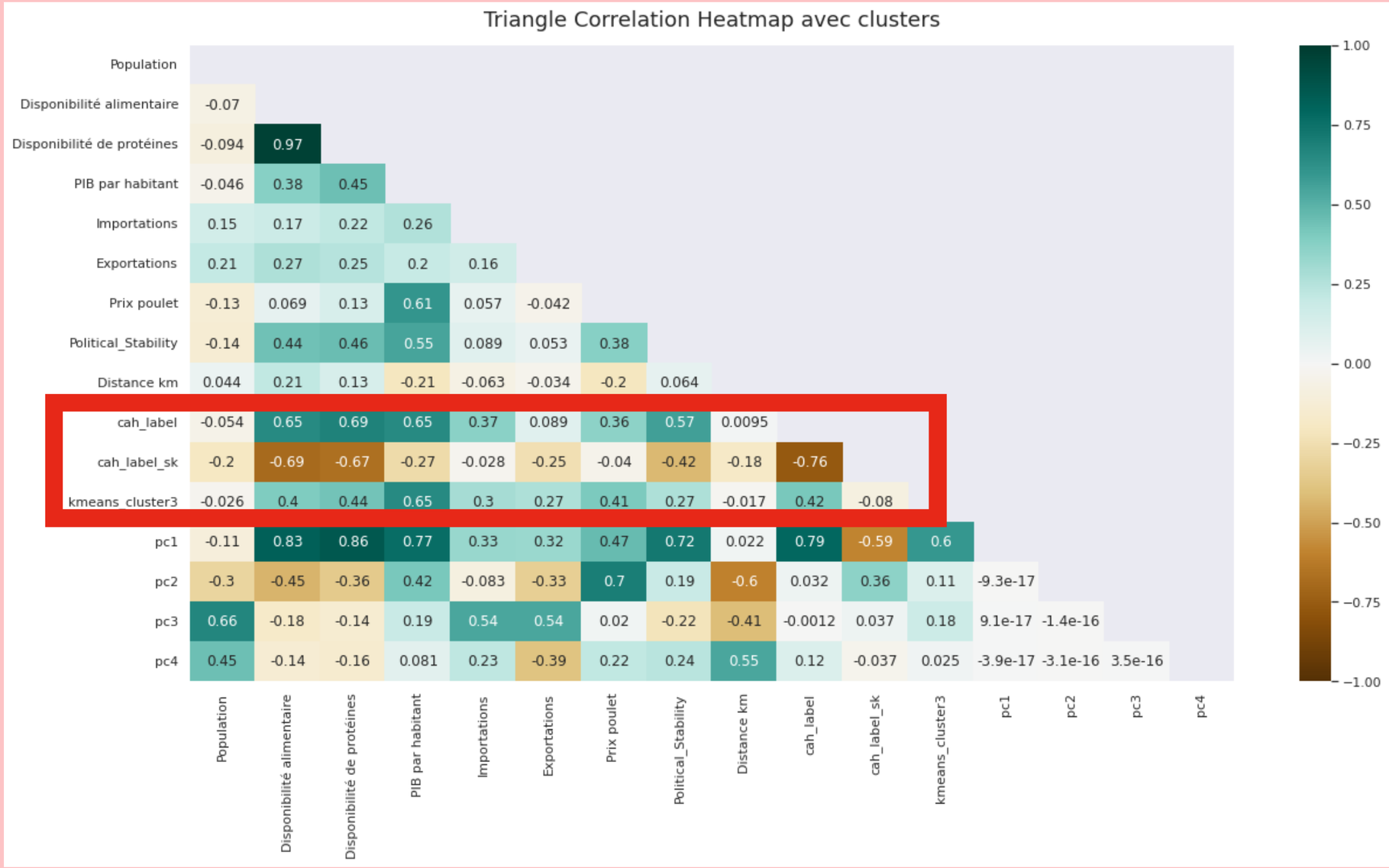
Heatmap



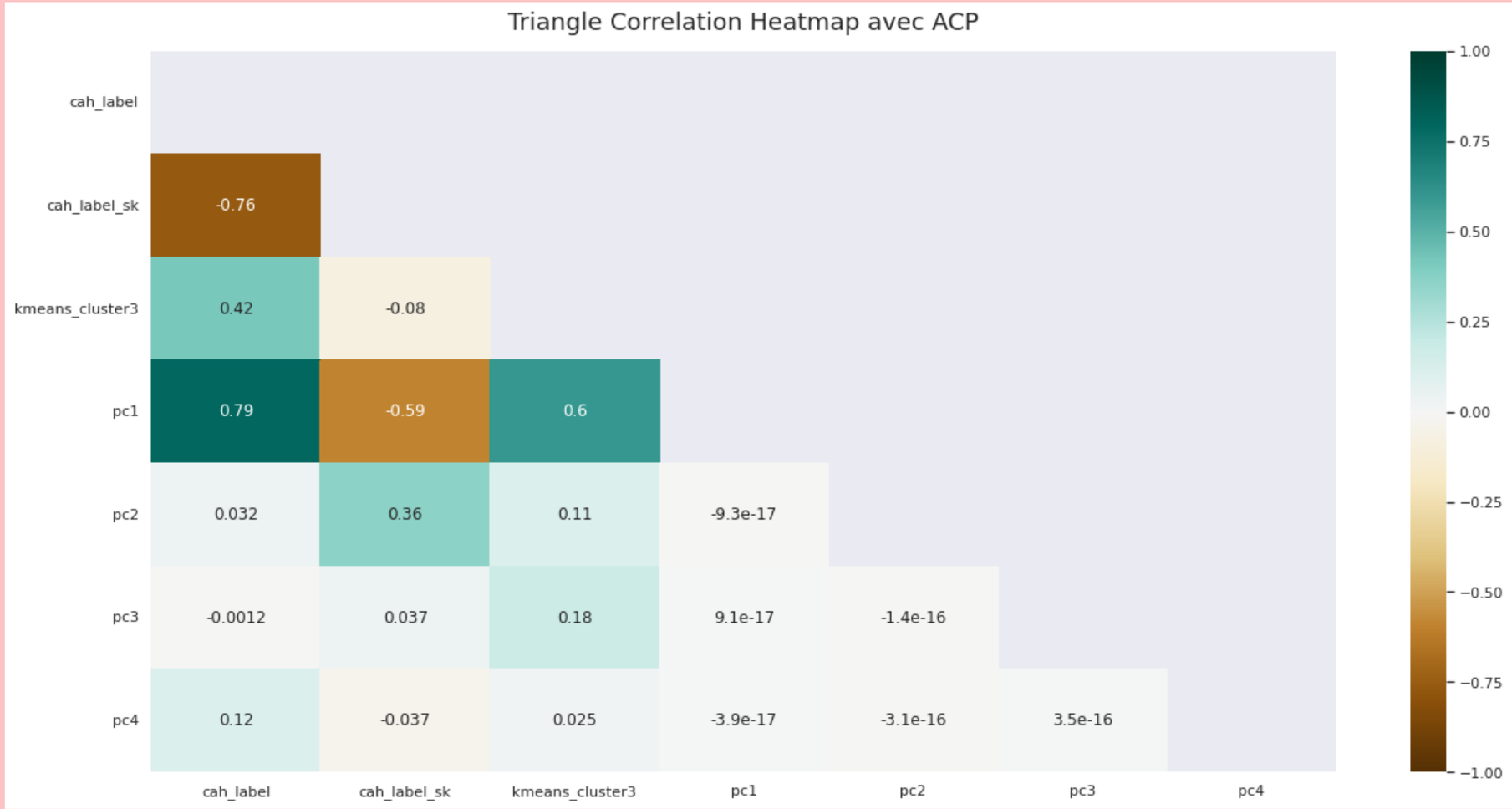
Heatmap avec clusters



Triangle Correlation Heatmap avec clusters



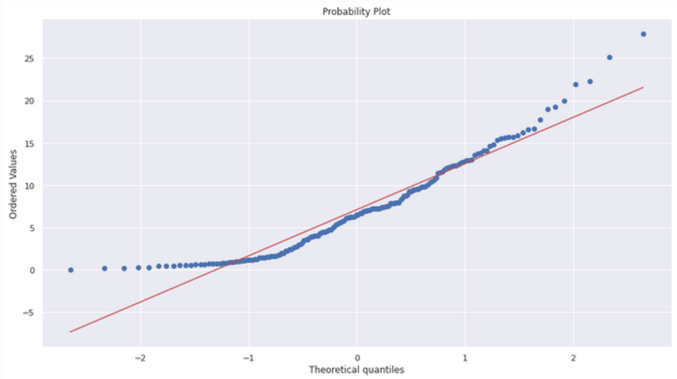
Triangle Correlation Heatmap avec clusters



Teste statistique

Teste loi normalité

- Q-Q plot
- Shapiro-Wilk Test



Shapiro-Wilk Test

```
# Teste loi normalité
stat_shapiro, p_shapiro = stats.shapiro(df['Disponibilité de protéines'])
print('Statistics=%.3f, p=%.3f' % (stat_shapiro, p_shapiro))
# Interpréter résultats
alpha = 0.05
if p_shapiro > alpha:
        print("L'hypothèse de normalité est donc tolérée (accepte H0)")
else:
        print("L'hypothèse de normalité est donc rejetée (rejete H0)")
```

```
Statistics=0.923, p=0.000
L'hypothèse de normalité est donc rejetée (rejete H0)
```

Teste homoscédasticité

- Teste levene pour disponibilité de protéines

Teste homoscédasticité - Teste levene pour disponibilité de protéines

Question: Les disponibilités de protéines sont-ils toujour pareils entre les classes differents ?

1. H0: Les disponibilités de protéines sont **pareils** entre les classes differents.
2. HA: Les disponibilités de protéines sont **differents** entre les classes differents.

```
[98] # grouper les samples
     group_km0 = df[df['kmeans_cluster3'] == 0]['Disponibilité de protéines']
     group_km1 = df[df['kmeans_cluster3'] == 1]['Disponibilité de protéines']
     group_km2 = df[df['kmeans_cluster3'] == 2]['Disponibilité de protéines']
     group_km3 = df[df['kmeans_cluster3'] == 3]['Disponibilité de protéines']

     # calculer le p-value
     F_group_km, p_value_group_km = stats.levene(group_km0, group_km1, group_km2, group_km3)
```

```
print('Statistics=%.3f, p=%.3f' % (F_group_km, p_value_group_km))
print("*P-value < 0.05, on rejette H0 et accept " +
            "HA: Les disponibilités de protéines sont differents entre les classes differents. " +
            "\nIl y a une corrélation entre les disponibilité de protéines et les clusters.")
```

```
Statistics=4.597, p=0.004
*P-value < 0.05, on rejette H0 et accept HA: Les disponibilités de protéines sont differents entre les classes differents.
Il y a une corrélation entre les disponibilité de protéines et les clusters.
```

- Teste levene pour disponibilité alimentaire

```
print ("*P-value < 0.05, on rejette H0 et accept " +
            "HA: Les Disponibilité alimentaire sont differents entre les classes differents. " +
            "\nIl y a une corrélation entre les disponibilité alimentaire et les classes.")
```

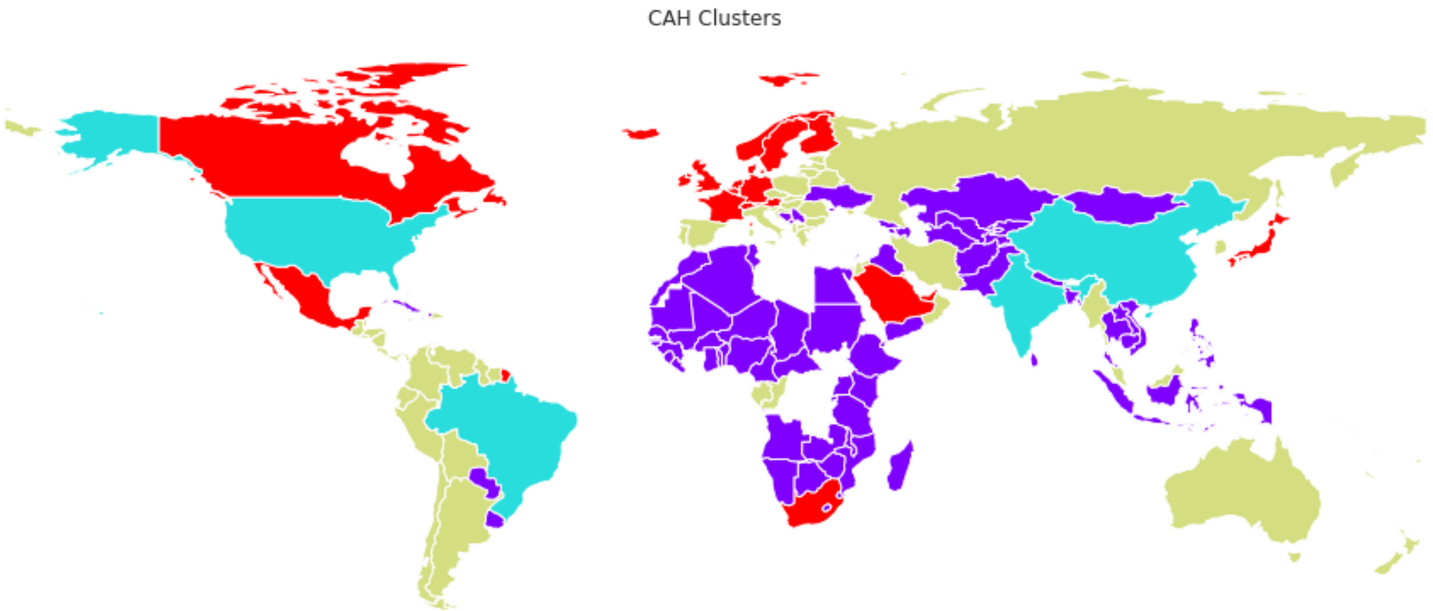
```
Statistics=17.001, p=0.000
*P-value < 0.05, on rejette H0 et accept HA: Les Disponibilité alimentaire sont differents entre les classes differents.
Il y a une corrélation entre les disponibilité alimentaire et les classes.
```

Recommandations

Recommandations sans tenir compte d'autres facteurs (22 pays)

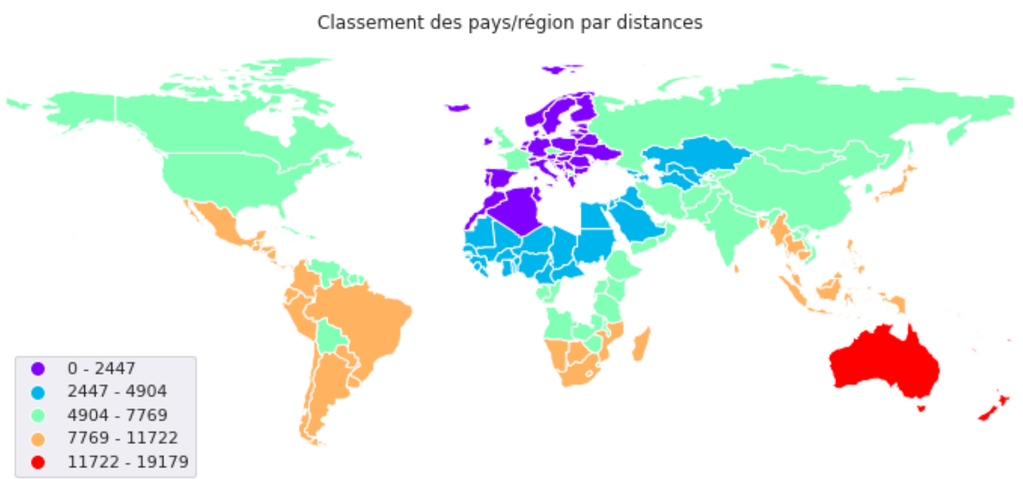
Groupe4- CAH-Rouge

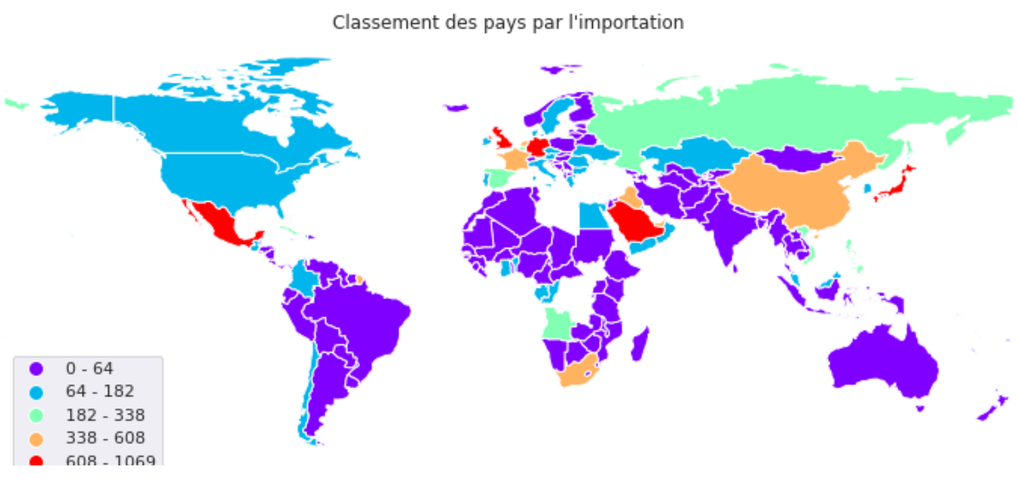
Les pays/région riches Ils sont les pays riches, bien développés avec PIB par habitant les plus hautes, les prix de poulet les plus haute. Ils sont les pays/regions les plus proches avec stabilité politique stable.



Recommandations prioritaires (7 pays):

1. **Pas de longue distance**
2. **Beaucoup d'importation**
3. **Bon prix -- "pays riche"**





Sélectionner les pays avec distances moins de 7769 kms pour choisir les 3 premiers classes

Recommandations prioritaires avec niveau de PIB par habitant

0	Canada
1	United Arab Emirates
2	Germany
3	Belgium
4	Netherlands
5	United Kingdom of Great Britain and Northern I...
6	Saudi Arabia

Sélectionner les pays avec importation plus 182 Milliers de tonnes pour choisir les 3 derniers classes

MERCI