ETUDE DE MARCHÉ AVEC PYTHON

La poule qui chante

SOMMAIRE

1 / Contexte

- L'entreprise d'agroalimentaire 'La poule qui chante' souhaite se développer a l'international.

- A partir des données de la FAO nous essayerons de trouver quels sont les principaux pays à cibler.

2/ Nettoyage et préparation des données

- Les données :
  - Données des disponibilité alimentaire de la FAO (Food and Agriculture Organization) pour l'année 2017
  - La population des pays 2017
  - Leur stabilité politique 2017

2/ Nettoyage et préparation des données

On nettoie et prépare les données au préalable :

- Traitement de potentiel doublons , valeurs nulles et outliers.
- Choix des variables
- Fusion des dataframes

Mon fichier pour les analyses comportera pour chaque pays des variables alimentaires pour les volailles :

- Importation
- Exportation
- La production
- Disponibilité globale
- Disponibilité par habitant
- Population
- Stabilité politique

3/ Analyse de données

A/ Analyse en composantes principales(ACP)

B/ Clustering hiérarchique agglomeratif

C/ Clustering K-means

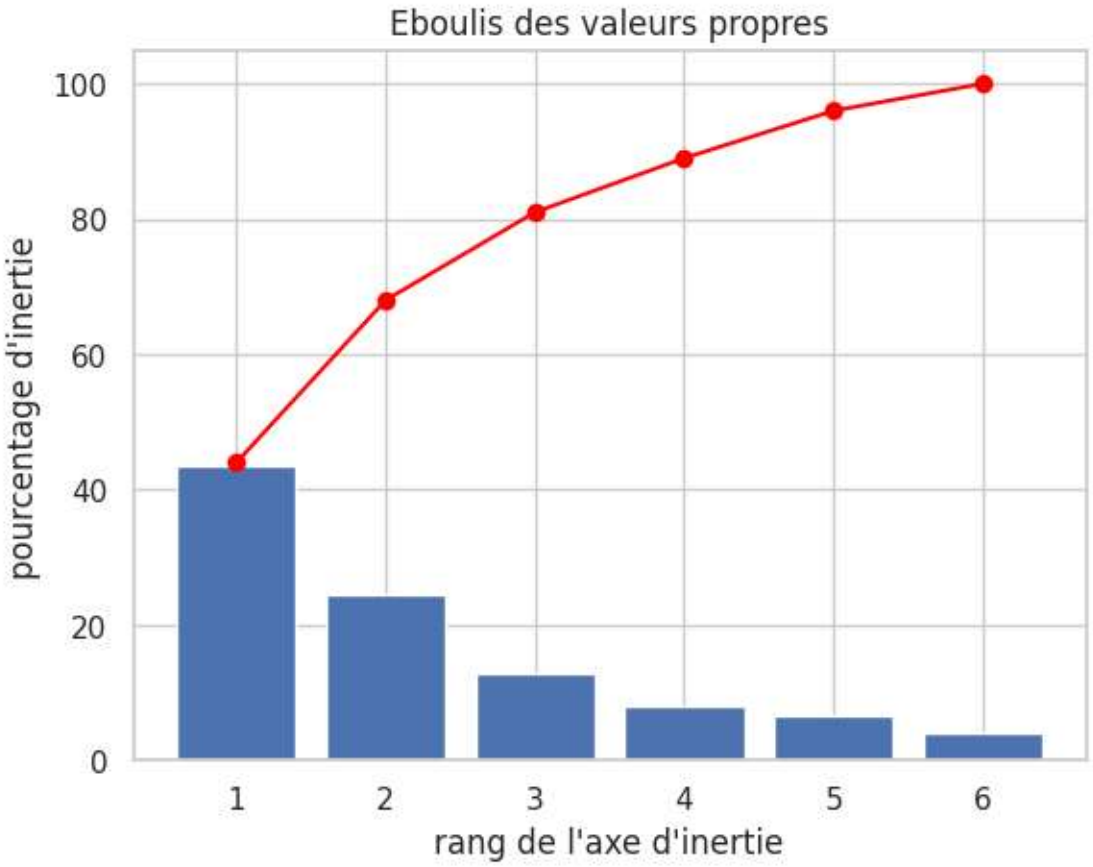
A/ Analyse en composantes principales (ACP)

- L'ACP est une technique d'analyse pour les données multivariées. En transformant les variables en nouvelles variables synthétiques appelées composantes.

- Objectifs :

- Connaitre les liaisons entre les variables
- Etudier la variabilité entre les individus

A/ Analyse en composantes principales (ACP)



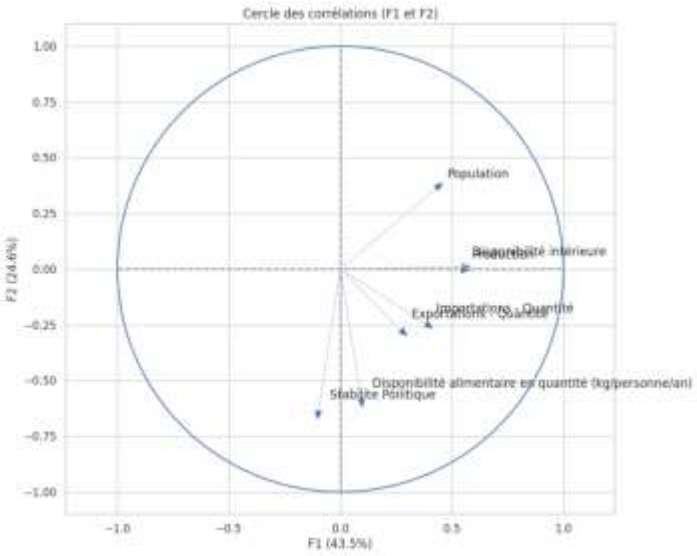
Le choix du nombre de composantes avec l'éboulis de la variance expliquée par les composantes principales.

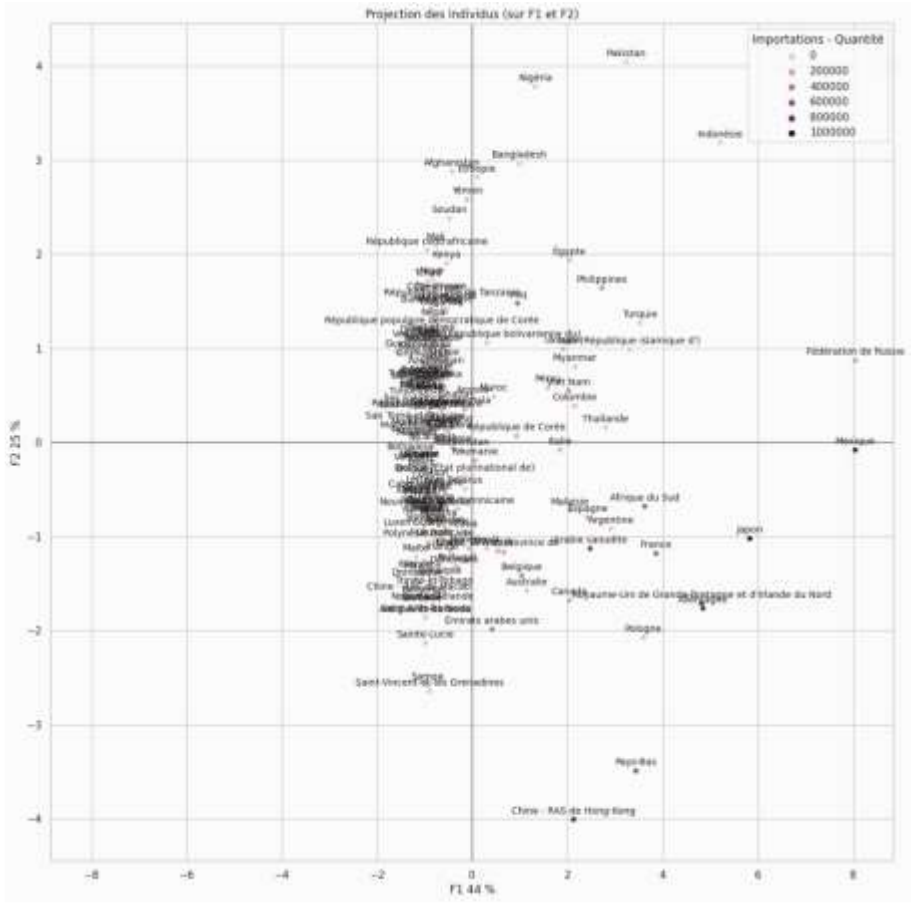
Avec la méthode du coude on recherche un cassure sur la courbe.

On observe un coude pour 2 composantes, pour mon analyse ACP je choisirai 4 composante qui représentent 90% de la variances.

A/ Analyse en composantes principales (ACP)

On peut a partir de ces graphiques visualiser les différences et points communs entre les pays.





Projection des pays dans un espace a deux dimension qui représente 69% de la variance.

B/ Cluster hiérarchique agglomératif

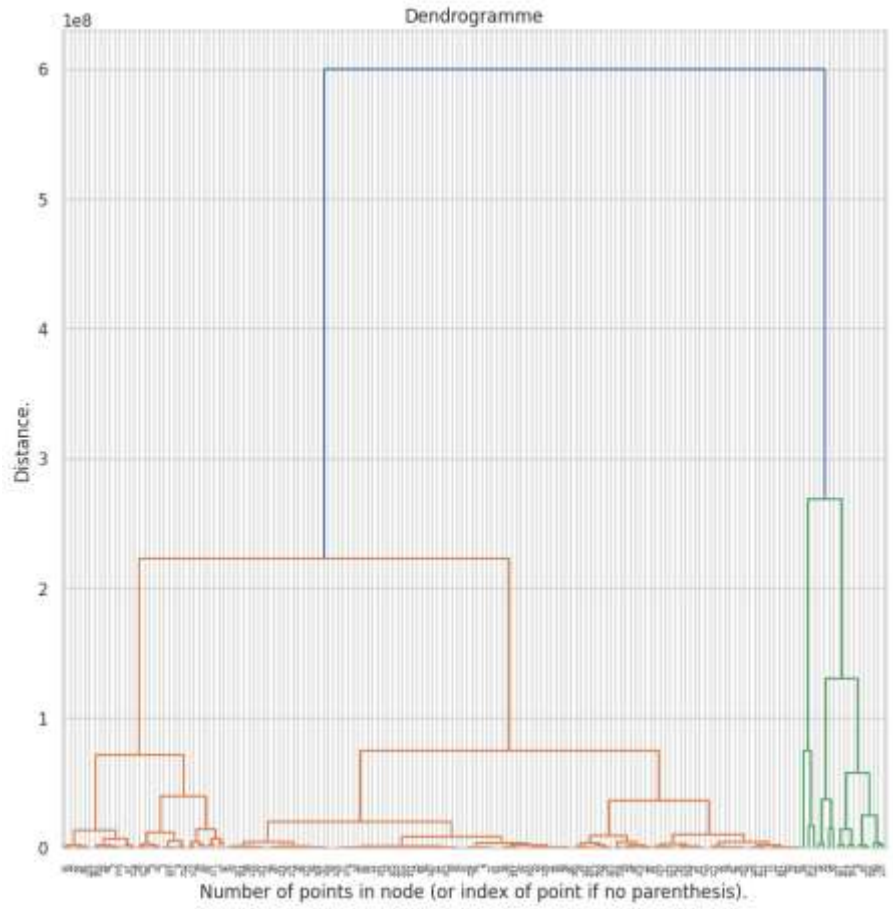
- Le clustering hiérarchique permet de regrouper les individus similaires en clusters selon la distance entre chaque individus.

- Le clustering agglomératif // divisif, part de chaque point et les regroupe un à un jusqu'à n'avoir qu'un cluster qui regroupe tous les points.

- Méthode de linkage utilisé → Ward qui minimise l'inertie intra-classe à chaque itération.

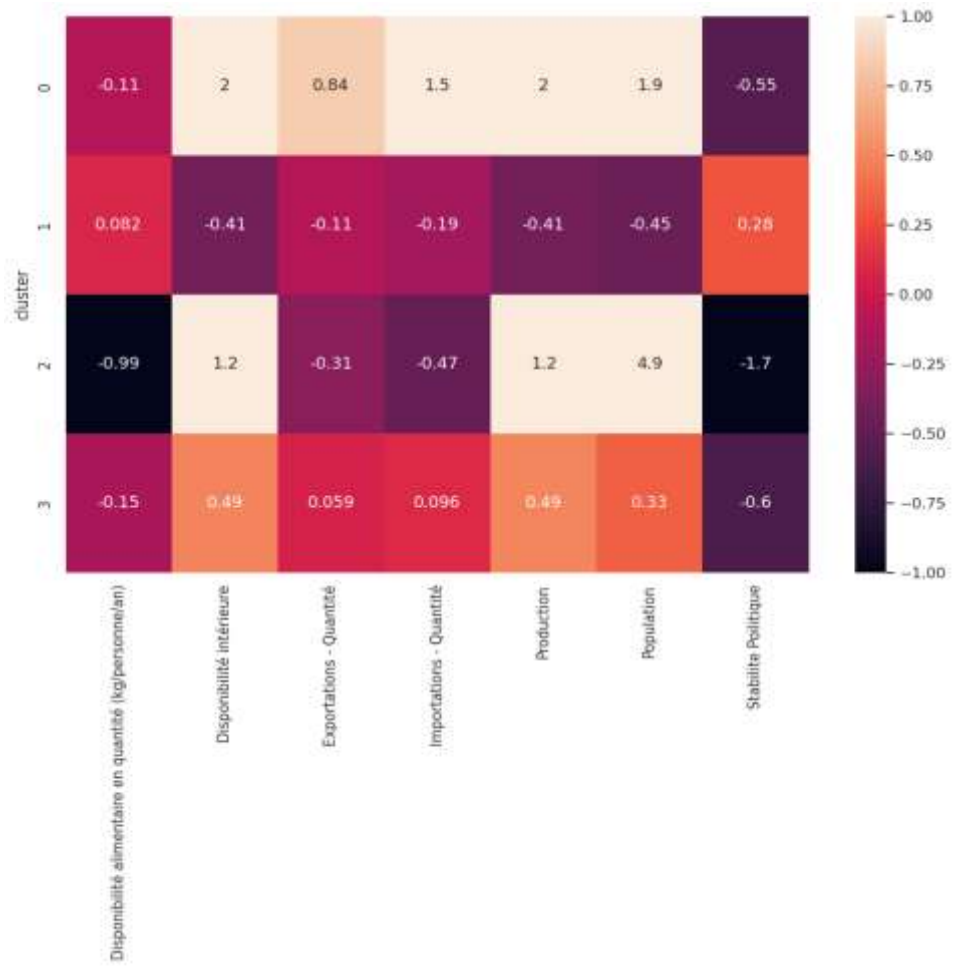
- L'avantage du clustering hiérarchique est qu'il n'y a pas besoin de sélectionner un nombre de cluster préalable.

Dendrogramme



B/ Cluster hiérarchique agglomératif

- A partir du dendogramme, je choisis de séparer mes données en 4 clusters.
- J'ajoute a chaque pays le cluster auquel il appartient
- Création d'une heat map pour chaque cluster en fonction des variables.
- Les clusters 0 et 3 sont intéressants pour l'expansion international de la société.



La poule qui chante

C/ Cluster hiérarchique agglomératif

Pays cibles

- Je ressors ensuite les pays des ces clusters.

Pour la *Poule qui chante* :

- On peut privilégier les pays plus proche de la France pour des raison logistique, donc : Italie, Allemagne, Espagne, Pologne, R-U.

Zone
Ghana
Viet Nam
République de Corée
Japon
Italie
Allemagne
France
Pologne
Espagne
Royaume-Uni de Grande-Bretagne et d'Irlande du Nord
Chine, Taiwan Province de
Canada
Malaisie
Argentine
Australie

C/ Clustering K-means

- K-means

- C'est une méthode de clustering qui consiste à diviser un ensemble de données en k clusters en minimisant la variance intra-cluster.
- On définit à l'avance le nombre de clusters souhaité.

- Fonctionnement

- L'algorithme k-means sélectionne aléatoirement k centres de cluster initiaux, puis attribue chaque observation à son centre de cluster le plus proche.
- Il calcul le centre de chaque cluster nouvellement créé et réassigne les observations aux centres de cluster les plus proches.
- Ainsi de suit jusqu'à ce que les clusters ne change plus. On obtient une partition des données.

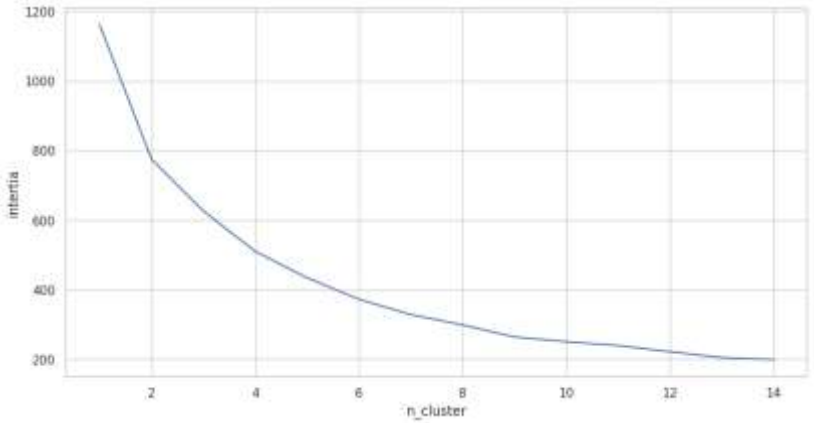
C/ Clustering K-means

- Choix du nombre de cluster

- Méthode du coude, on observe ou il y a une cassure sur la courbe d'inertie en fonction du nombre de cluster.

- Score de Davies-Bouldin de 2 à 10 clusters.

Inertie/nbre de cluster

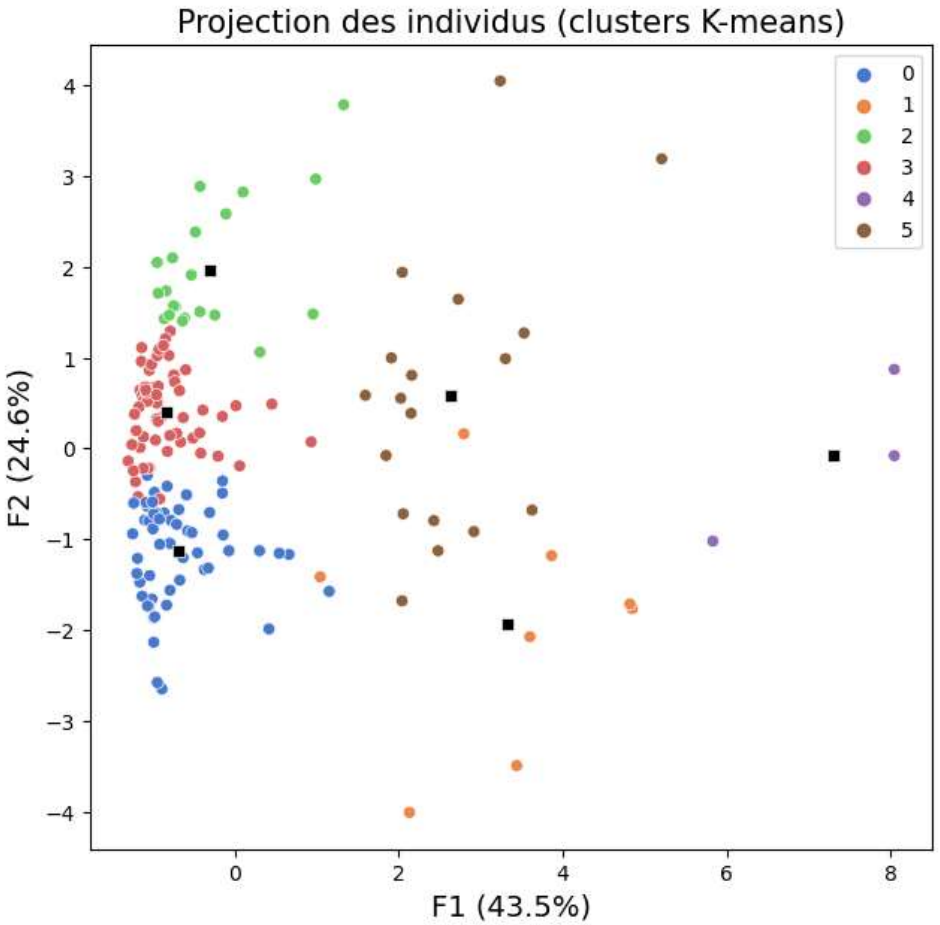


Score Bouldin/nbre de cluster

```
Davies-Bouldin score for 2 clusters: 1.181
Davies-Bouldin score for 3 clusters: 1.094
Davies-Bouldin score for 4 clusters: 1.156
Davies-Bouldin score for 5 clusters: 1.072
Davies-Bouldin score for 6 clusters: 1.004
Davies-Bouldin score for 7 clusters: 1.039
Davies-Bouldin score for 8 clusters: 1.011
Davies-Bouldin score for 9 clusters: 1.049
Davies-Bouldin score for 10 clusters: 1.030
```

C/ Clustering K-means

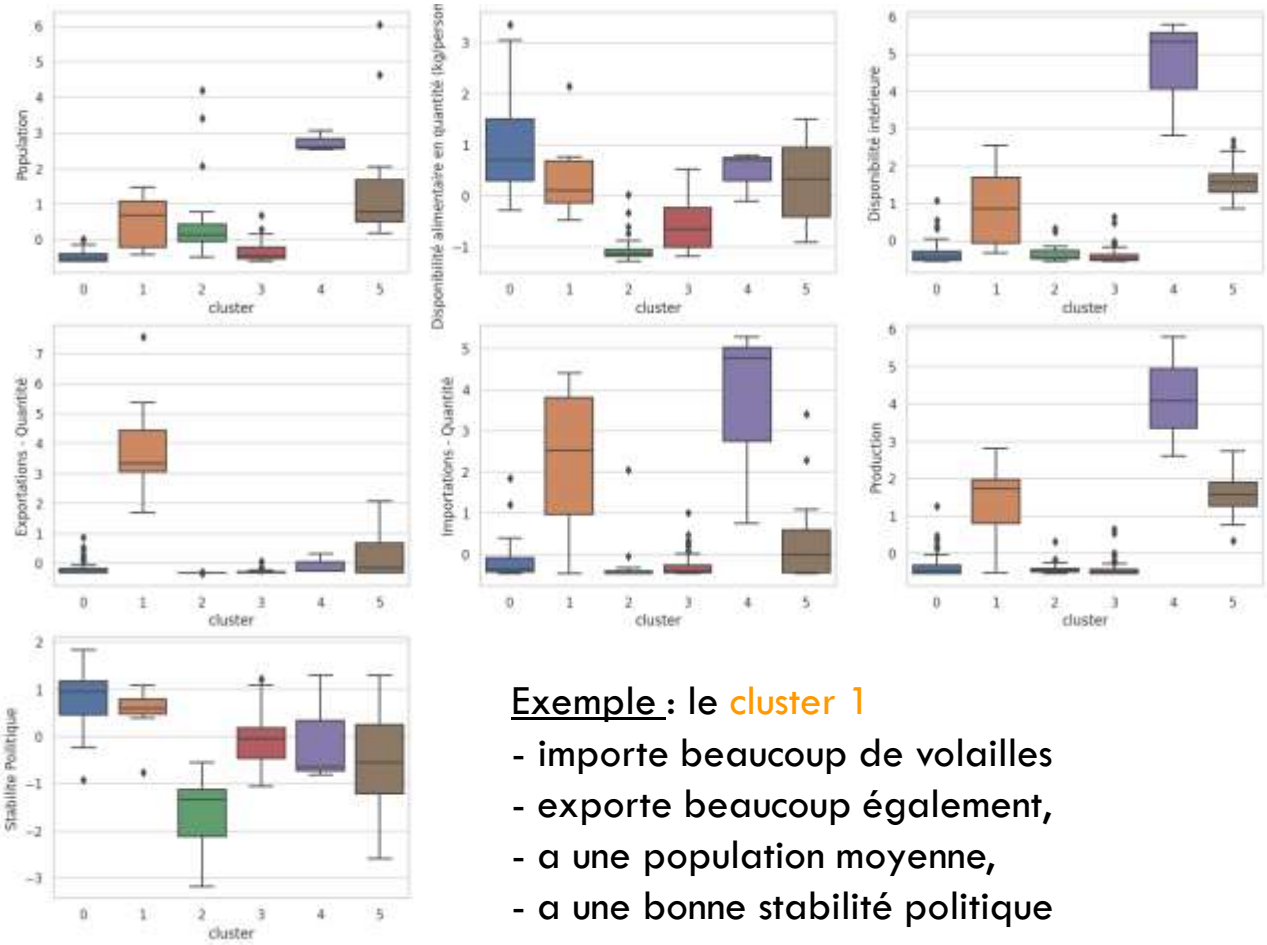
- On lance l'algorithme k-means.
- On projette ensuite les centroides sur la représentation de l'ACP, et on distingue les clusters.



C/ Clustering K-means
Analyse des clusters

• Analyser les clusters par variables avec des boxplot, nous permet d'avoir une représentation des clusters pour chaque variables.

• Ici il va être intéressant de choisir des clusters avec de forte importations, une bonne stabilité politique et potentiellement une population élevée pour plus de client potentiel.

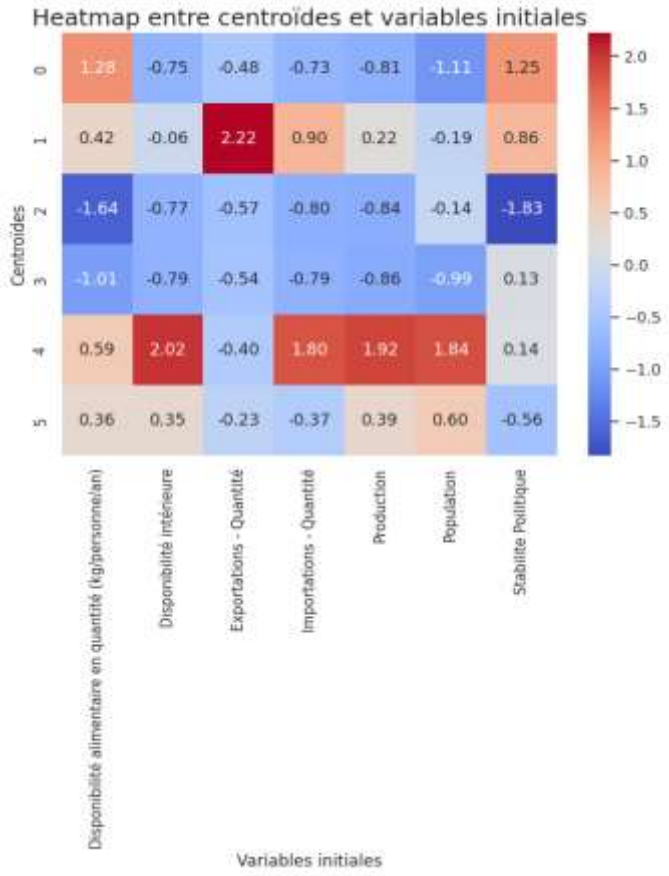


Exemple : le cluster 1
- importe beaucoup de volailles
- exporte beaucoup également,
- a une population moyenne,
- a une bonne stabilité politique

C/ Clustering K-means
Analyse des centroides

Analyse par les centroïdes, les représentants des clusters.

Avec une heat map on peut voir la représentation des variables pour chaque cluster.
On peut identifier lesquelles ont le plus d'influence sur la formation des clusters.



On retrouve ici le cluster 1 qui importe et exporte beaucoup avec une population moyenne et une bonne stabilité politique

C/ Clustering K-means
Pays cibles

- Je ressors les pays des clusters 1 et 4.

- On sélectionne les pays avec une stabilité politique ≥ 0.

- Comme pour le Clustering hiérarchique on peut privilégier les pays proche de la France :
La Belgique, l'Allemagne, la Pologne, le R-U .

Avec le k-means

Zone
Belgique
Japon
Allemagne
Pays-Bas
France
Pologne
Royaume-Uni de Grande-Bretagne et d'Irlande du Nord
Chine - RAS de Hong-Kong

Comparaison des résultats

On retrouve 5 pays communs au deux méthodes de clustering.

Il y a plus de pays dans la méthode CAH car j'ai choisis de partitionner les individus en 4 clustering, donc il y a plus de pays par cluster.

La ou le K-means avec 6 cluster sera plus affiné et moins de pays par cluster.

Avec le k-means

Zone
Belgique
Japon
Allemagne
Pays-Bas
France
Pologne
Royaume-Uni de Grande-Bretagne et d'Irlande du Nord
Chine - RAS de Hong-Kong

Avec le CHA

Zone
Ghana
Viet Nam
République de Corée
Japon
Italie
Allemagne
France
Pologne
Espagne
Royaume-Uni de Grande-Bretagne et d'Irlande du Nord
Chine, Taiwan Province de
Canada
Malaisie
Argentine
Australie

Conclusion

Pour l'expansion de La poule qui chante à l'international :

La société a plusieurs choix selon la méthode de clustering, avec des pays de différents continents notamment Europe et l'Asie.

On peut conseiller la société de choisir parmi les pays Européen qui peuvent faciliter les échanges par rapport aux autres pays.

La Pologne, l'Allemagne ou le Royaume-Uni sont de bons candidats.