ACP et Clustering

2023-06-23

POULET MONDIAL

“Poulet Mondial” est une entreprise française d’agroalimentaire. Elle souhaite se développer à l’international.

L’objectif de l’étude est de proposer une première analyse des groupements de pays que l’on peut cibler pour exporter les poulets.

La première partie de notre travail a été de récolter les données nécessaires pour cette étude, de les nettoyer, préparer et en faire une première analyse exploratoire.

La seconde partie de notre étude va consister à analyser de plus près les zones afin de définir un groupement de pays que l’on peut cibler pour exporter nos poulets.

Dans un premier temps, nous allons effectuer la standardisation des données pour rendre les variables comparables. Nous allons ensuite réaliser l’ACP (Analyse en Composantes Principales) pour réduire la dimension des données et visualiser les relations entre les pays.

Ensuite, nous allons utiliser plusieurs méthodes de clustering, notamment la Classification Ascendante Hiérarchique (CAH), la méthode des K-means et la méthode DBScan, pour identifier des groupes ou des clusters parmi les pays. Ces méthodes nous aident à regrouper les pays ayant des caractéristiques similaires, ce qui est essentiel pour notre objectif de sélection des pays cibles pour l’exportation de poulets.

Enfin, nous allons comparer les résultats de ces méthodes de clustering afin d’avoir une liste de pays cibles pour l’ expansion internationale.

1 Importation des Librairies

```
# Pour des opérations utilitaires
library(utils)

# Bibliothèques pour la manipulation de données et l'analyse statistique
library(dplyr)
library(tidyr)
library(Hmisc)

library(magrittr)

# Bibliothèques pour la visualisation des données
library(ggplot2)
library(plotly)
library(factoextra)
library(corrplot)
library(ggrepel)
library(gridExtra)

# Bibliothèques pour l'analyse de données multidimensionnelles et de clustering
library(FactoMineR)
library(cluster)
library(dbscan)

# Bibliothèque pour la personnalisation des tables
library(kableExtra)
```

2 Importation des données et dictionnaire des données

2.1 Importation des données

Nous avons importé les données préalablement préparée dans la première partie du projet “Préparation, Nettoyage et Analyse exploratoire des données”.

Nous avons identifié chaque pays avec son code codes pays à 3 lettres (alpha-3) comme définis par l’ISO 3166-1.

```
## character(0)
```

2.2 Dictionnaire des données

Variable	Descriptif	Unité
Population	Population totale de la zone.	Nombre de personnes
Stabilité politique	Indicateur de stabilité politique.	Mesure sans unité
PIB	Produit Intérieur Brut par habitant.	Dollars par habitant
Inflation	Taux d'inflation.	Pourcentage
Inflation décimale	Taux d'inflation décimal.	Décimal
Production	Production nationale de viande de volaille.	Milliers de tonnes
Importations	Quantité importée de viande de volaille.	Milliers de tonnes
Exportations	Quantité exportée de viande de volaille.	Milliers de tonnes
Variation de stock	Variation des stocks de viande de volaille.	Milliers de tonnes
TDI (Taux de Dépendance aux Importations %)	Pourcentage de dépendance aux importations de viande de volaille.	Pourcentage
TDI décimal	Taux de dépendance aux importations de viande de volaille décimal.	Décimal
TAS (Taux d'Auto-Suffisance %)	Pourcentage d'auto-suffisance en production de viande de volaille.	Pourcentage
TAS décimal	Taux d'autosuffisance en production de viande de volaille décimal.	Décimal
Disponibilité intérieure	Disponibilité intérieure.	Milliers de tonnes
Dispo_poulet_kg_par_hab	Disponibilité intérieure de poulet par habitant, calculée en divisant la disponibilité intérieure par la population.	Kilogrammes par habitant

3 Standardisation des données

Les données présentent des échelles différentes et ne peuvent être comparables dans leur état brut. Nous allons donc utiliser la technique de standardisation pour les rendre comparables sur une échelle commune. Ceci afin que toutes les variables aient le même poids dans la construction des plans de l'ACP.

Nous allons Sélectionner les variables numériques et les standardiser.

Pour standardiser les colonnes du dataframe merged_df , nous allons utiliser la fonction scale() qui standardise les valeurs en soustrayant la moyenne et en divisant par l'écart type.

4 Analyse en Composantes Principales (ACP)

L'analyse en composantes principales (ACP) est une technique de réduction de dimension qui vise à transformer des variables corrélées en un nouvel ensemble de variables non corrélées appelées composantes principales. Cette technique permet de visualiser et de résumer les relations complexes entre les variables dans un espace de dimension réduit.

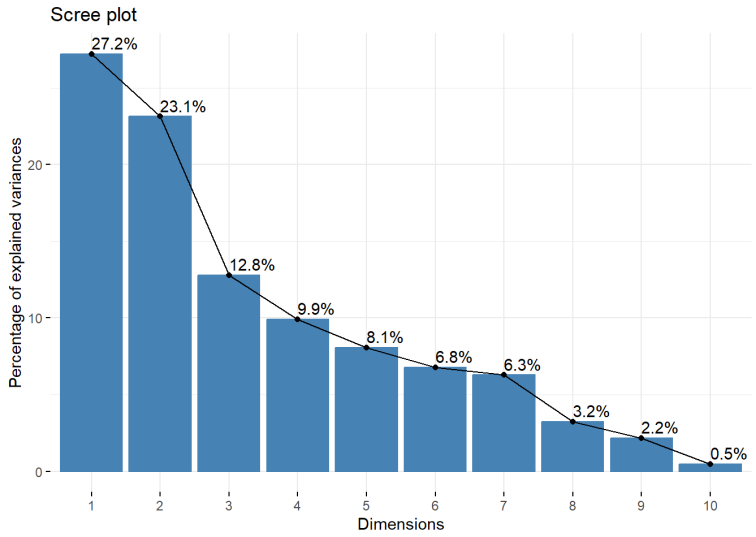
L'enjeu d'une ACP est de trouver le meilleur plan de projection ayant la plus grande inertie, c'est à dire limitant le plus la perte d'information originelle.

Nous pourrons ainsi visualiser les relations entre variables à l'aide d'un cercle de corrélation et la variabilité entre les pays.

4.1 ACP avant suppression des outliers

4.1.1 Répartition du Pourcentage d'Explication par les Plans Principaux

Pour sélectionner le nombre d'axes à conserver, on utilise le critère du coude, qui consiste à examiner le graphique de l'inertie expliquée en fonction du nombre d'axes. On retient les premiers axes jusqu'à observer un "décrochage" significatif dans la courbe de l'inertie expliquée.



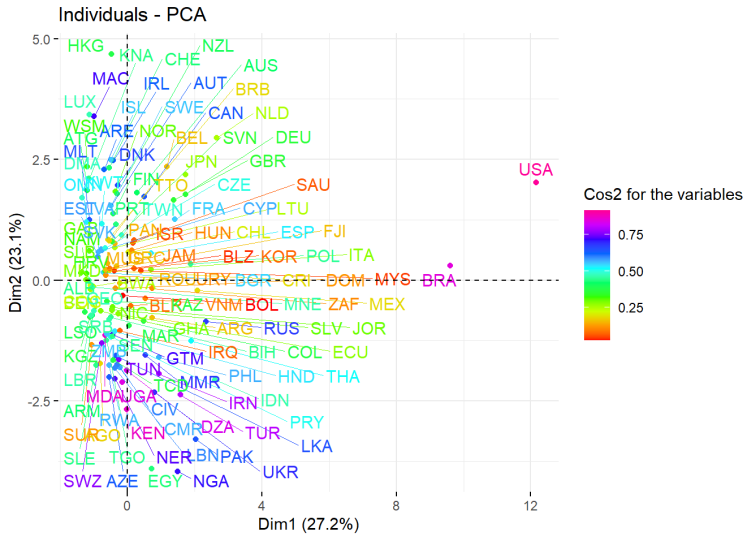
Dans notre cas, nous observons que les 3 premiers axes de l'ACP traduisent 63,3 % de l'inertie totale.

4.1.2 Graphiques de corrélation des variables et nuage des individus

Le cercle des corrélations est un outil pratique nous permettant de visualiser l'importance de chaque variable pour chaque axe de représentation.

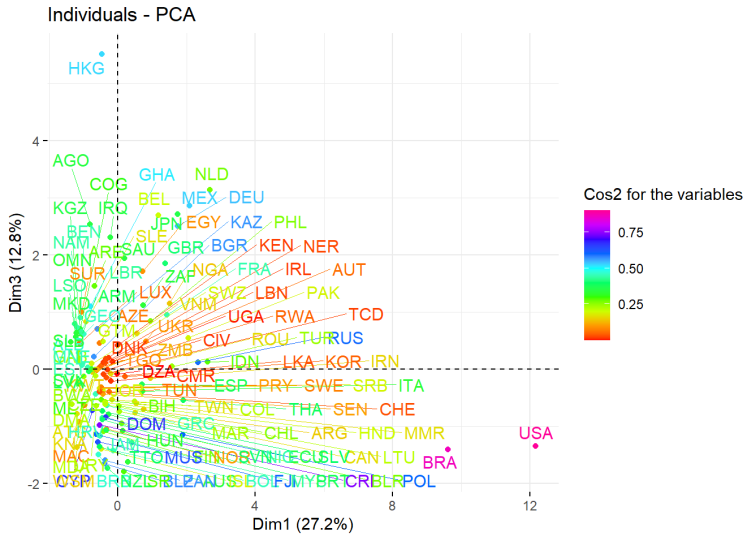
Les principes de lecture sont les suivants :

- plus une variable possède une qualité de représentation élevée dans l'ACP, plus sa flèche est longue;
- plus deux variables sont corrélées, plus leurs flèches pointent dans la même direction (dans le cercle de corrélation, le coefficient de corrélation est symbolisé par les angles géométriques entre les flèches);
- plus une variable est proche d'un axe principal de l'ACP, plus elle est liée à lui. Cette dernière règle permet généralement de donner un sens concret aux axes de l'ACP
- Plan (1,2)



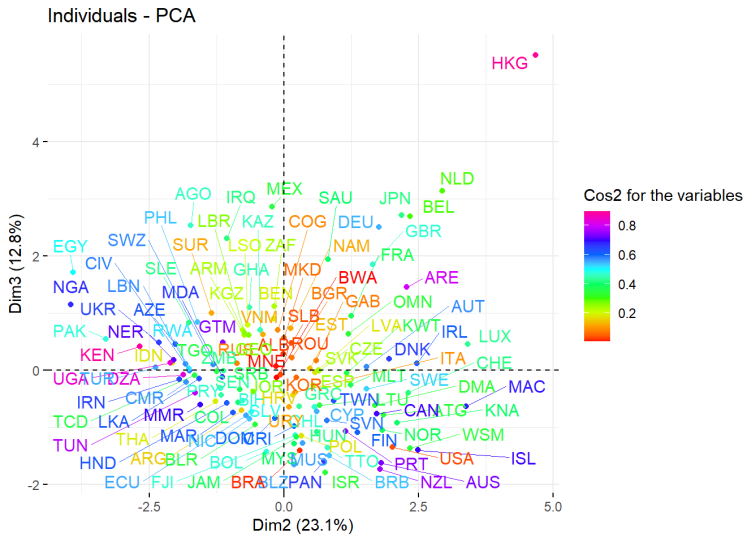
L'analyse du nuage des individus sur le plan (1,2) nous montre 2 valeurs extrêmes: USA “États-Unis d'Amérique” et BRA “Brésil”. Ceux sont de loin les pays qui produisent et exportent le plus.

- Plan (1,3)



Sur ce plan, on retrouve nos 2 outliers précédent, mais aussi la HKG “Hong-Kong”. C'est un pays qui dépend fortement de l'importation, il importe beaucoup.

- Plan (2,3)



Sur ce 3ème plan on retrouve l'outlier “Hong-Kong”.

Zone	Production	Importations	Exportations	TDI	TAS	dispo_poulet_kg_par_hab
Brésil	14201	3	4223	0.03	142.28	48.02
Chine - RAS de Hong-Kong	24	907	663	338.43	8.96	35.04
États-Unis d'Amérique	21914	123	3692	0.67	119.45	56.68

Concernant le Brésil, c'est un pays qui importe peu, TDI est très faible et un TAS > 100%. Concernant les Etats-Unis, il fait parti des pays qui importe le plus bien qu'il ne dépende pas de l'importation TDI=0.67% et il produit beaucoup et est autosuffisant.

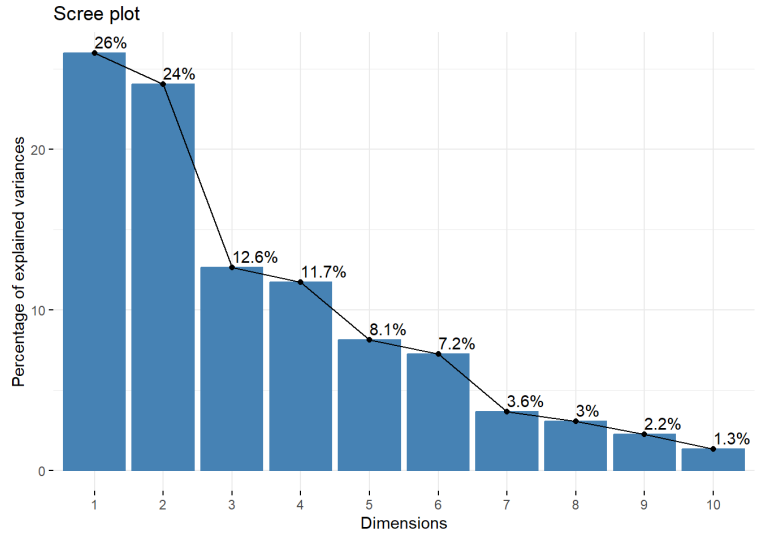
Nous décidons de les supprimer pour la suite de l'étude en raison de leurs valeurs extrêmes tout en gardant en tête que les Etats-Unis pourrait être un pays intéressant.

La “Chine - RAS de Hong-Kong”. C'est un pays qui dépend fortement de l'importation, il importe beaucoup. Ce pays est un sérieux candidat.

Après la suppression de ces outliers, nous relançons une nouvelle ACP avec les nouvelles données.

4.2 ACP après suppression des outliers

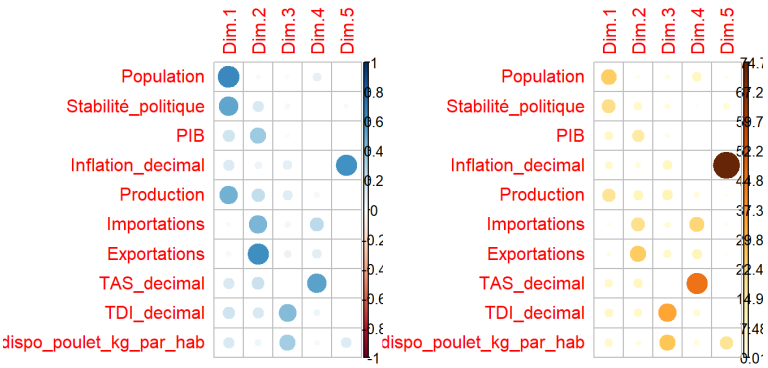
4.2.1 Répartition du Pourcentage d'Explication par les Plans Principaux



On voit ici un décrochage au niveau des 3 premières composantes qui représentent 62.6 % de l'inertie totale .

4.2.2 Analyse des variables et des individus

Visualisation de la qualité de la représentation des variables et de la contribution des variables aux axes



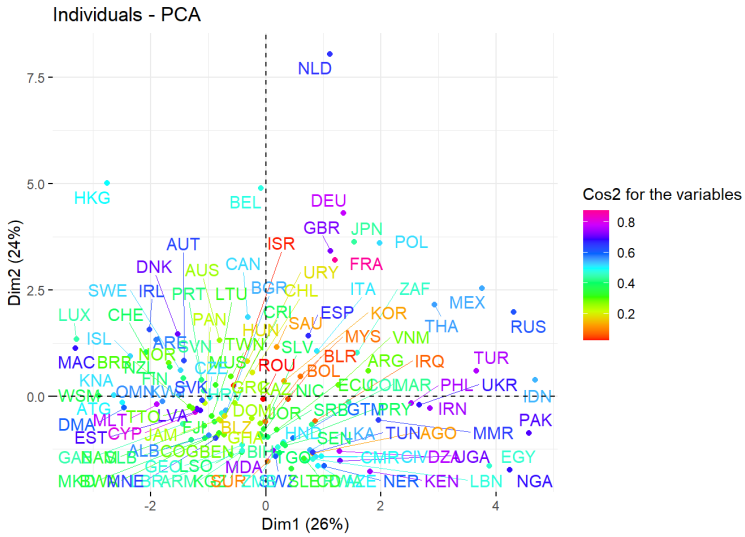
Cela nous indique que:

- Les variables "Production","Population" et "stabilité politique sont bien représentées sur la dimension 1 et contribuent le plus.
- Les variables "Exportation" et Importation" sont bien représentées sur la dimension 2 et contribuent le plus. Le "PIB" y est un peu moins bien représenté.
- Les variables "dispo_par_hab" et "TDI" sont bien représentées sur la dimension 3 et contribuent le plus.

Les variables "TAS" et "Inflation" sont males représentées dans les 3 premières dimensions

Graphiques de corrélation des variables et nuage des individus

- Plan (1,2)

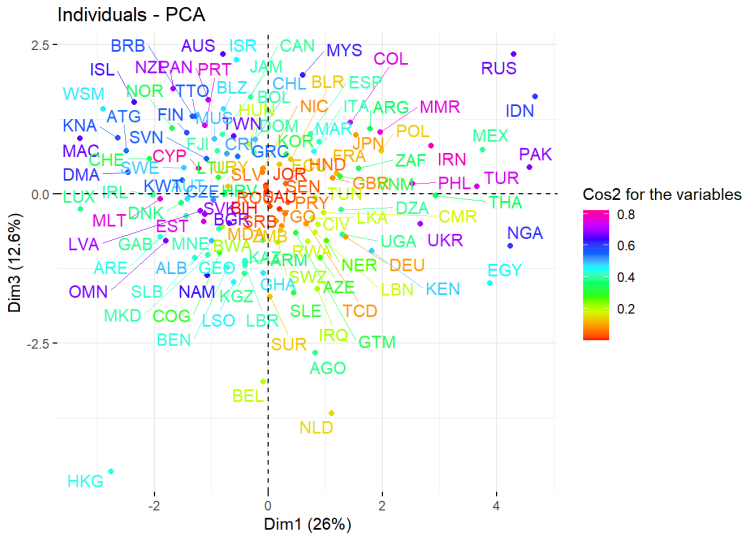


Sur le plan (1,2) , on note NLD(Pays-Bas), BEL (Belgique),DEU(Allemagne),GBR (Grande-Bretagne) qui sont des pays qui importent et exportent beaucoup.

RUS (Russie), IDN (Indonésie) et MEX (Mexique) pays qui produisent beaucoup et ont une forte population.

La Macédoine (MAC), l'Islande (ISL) et le Luxembourg (LUX) parmis les pays les plus stables politiquement.

- Plan (1,3)



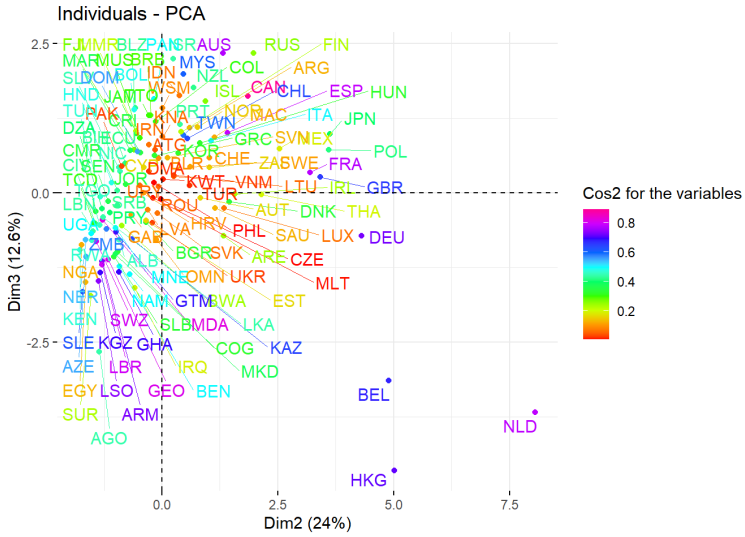
Sur le plan (1,3) , on note HKG (Hong-Kong) en tant que pays qui dépend beaucoup de l'importation.

On voit des pays comme la RUS (Russie),IDN (Indonésie) ou PAK(PAkistan) qui ont une forte population et produisent beaucoup.

La Macédoine (MAC)et le Luxembourg (LUX) parmis les pays les plus stables politiquement.

Des pays avec une importante 'dispo par habitant' comme la Nouvelle-Zélande NZL, Australie AUS, Israel ISR

- Plan (2,3)



Sur le plan (2,3) , on note 3 pays qui se démarquent nettement: BEL(Belgique), NLD(Pays Bas) et HKG(Hong-Kong). Ils font partie des pays qui à la fois importent beaucoup, exportent beaucoup et ont une forte dépendance à l'importation.

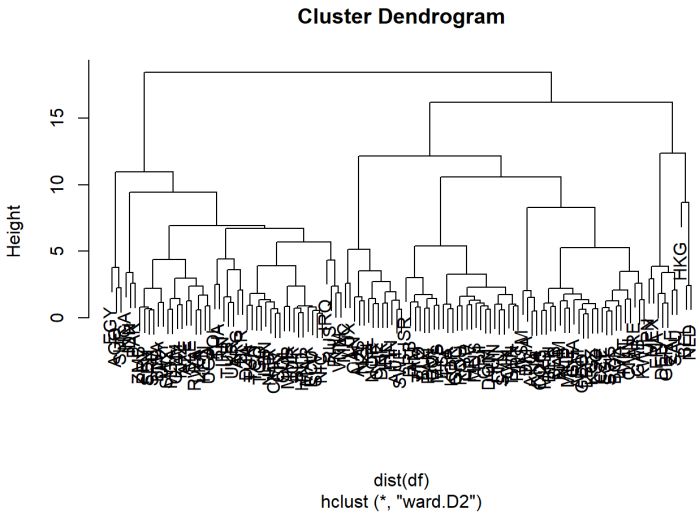
5 Clustering

5.1 Classification Ascendantes hiérarchiques (CAH)

La CAH est une méthode de regroupement des données qui commence par considérer chaque élément comme un cluster individuel. Ensuite, elle fusionne progressivement les clusters les plus similaires pour créer une hiérarchie de clusters, comme un arbre. Chaque fusion est représentée par une branche dans l'arbre.

L'objectif est de regrouper des pays similaires sur le critère de la distance séparant chaque point.

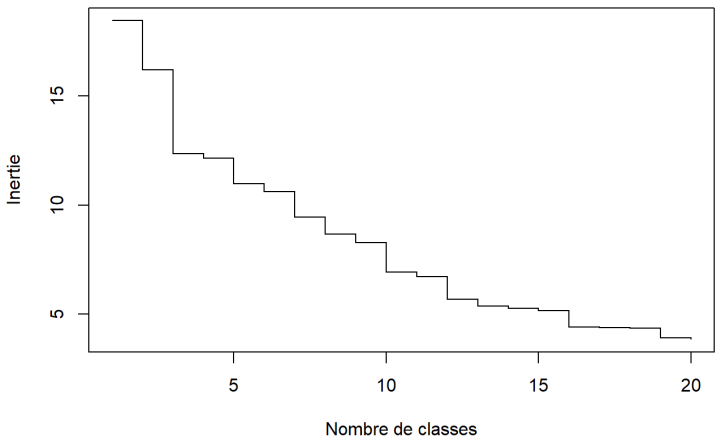
On utilise la méthode de Ward : cette méthode cherche à minimiser l'inertie intra-classe et à maximiser l'inertie inter-classe afin d'obtenir des classes les plus homogènes possibles.



En premier lieu, une analyse de la forme du dendrogramme pourra nous donner une indication sur le nombre de classes à retenir. 3 branches distinctes apparaissent sur l'arbre.

5.1.1 Détermination du nombre optimal de clusters

- Pour nous aider, nous pouvons représenter les sauts d'inertie du dendrogramme selon le nombre de classes retenues. Pour déterminer le nombre optimal de clusters, nous regardons les sauts d'inertie dans le dendrogramme. Chaque fusion dans le dendrogramme est associée à un saut d'inertie. Plus le saut est important, plus la fusion de clusters à cet endroit est significative en termes de variance. Cela signifie que le nombre optimal de clusters pourrait être associé à ce saut.



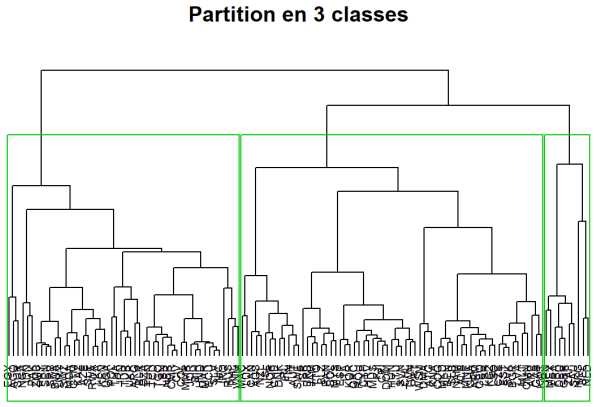
On voit un saut assez net à 3 classes.

- Utilisation de la fonction best.cutree

L'extension JLutils propose une fonction best.cutree qui calculera la meilleure partition en termes de perte relative d'inertie pour un nombre de clusters allant de 3 à 20 (par défaut), et renverra le nombre optimal de clusters.

```
## [1] 3
```

Un découpage en 3 clusters est confirmé.



Regardons le nombre de pays que comporte chaque cluster.

Var1	Freq
1	10
2	65
3	50

5.1.2 Indice de silhouette

L'objectif du clustering est d'obtenir des clusters de bonne qualité. Le clustering est de haute qualité si la distance dans les observations (intra-cluster) d'un cluster donné est minimale et la distance séparant les clusters eux-mêmes (inter-cluster) est maximale.

Le coefficient de silhouette est une mesure de la qualité d'une partition. Il évalue à quel point chaque observation est bien regroupée par rapport aux autres observations du même cluster, en prenant en compte également à quel point cette observation pourrait être mieux regroupée dans un autre cluster.

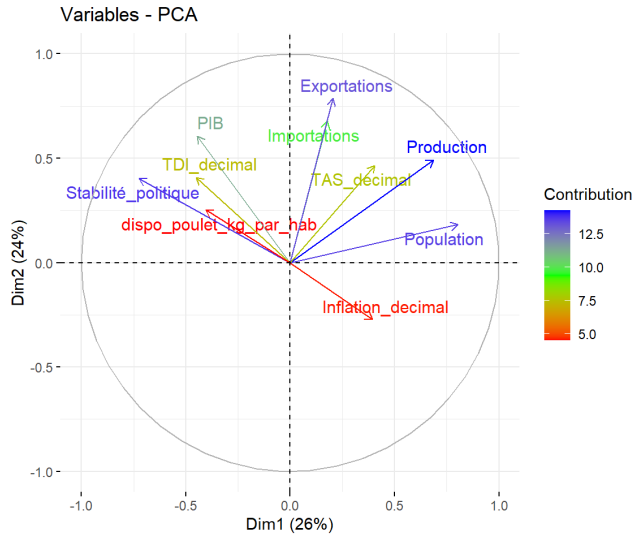
Un coefficient de silhouette proche de +1 indique une bonne séparation des clusters, tandis qu'un coefficient proche de -1 indique que l'observation serait mieux regroupée dans un autre cluster.

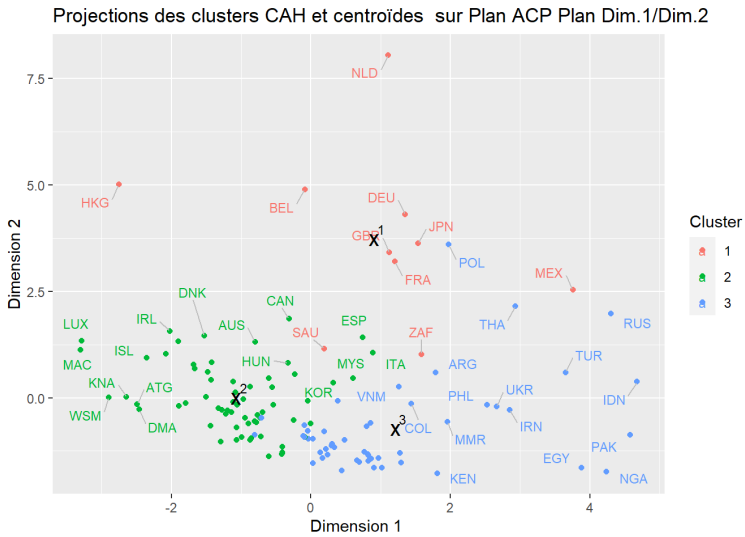
Cluster	Silhouette_Mean_CAH
1	0.07
2	0.24
3	0.18
Overall	0.20

5.1.3 Projection des clusters et centroÏdes sur les plans ACP

Nous allons calculer les coordonnées acp des centroïdes (moyenne) de chacun de ces groupes pour ensuite projeter nos clusters et leurs centroïdes sur les plans ACP.

- Plan (1,2)



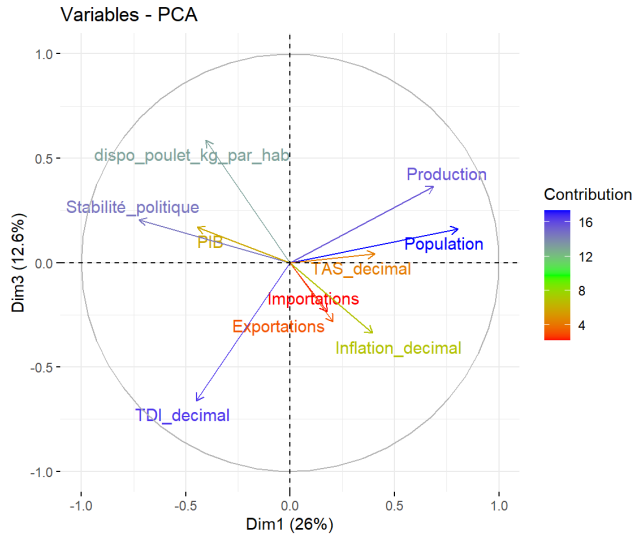


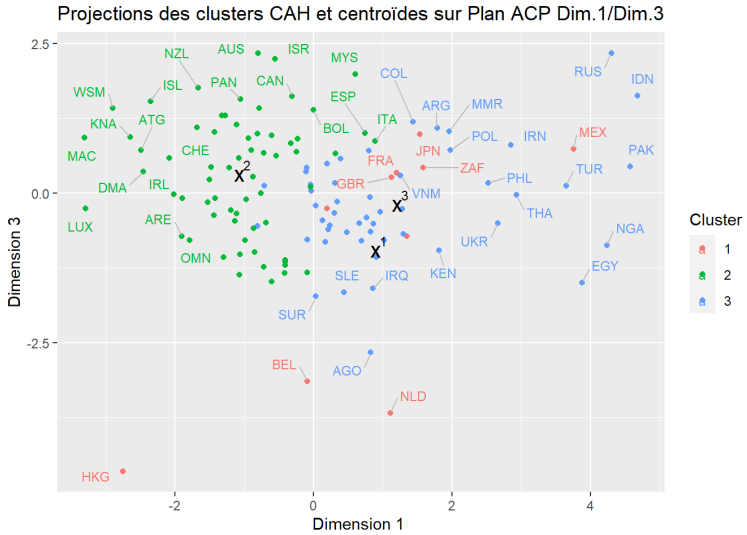
Le cluster 1 semble être composé des pays qui importent et exportent le plus.

Le cluster 2 est composés de pays plus stables politiquement.

Le cluster 3 comportent des pays à forte population et production.

- Plan (1,3)



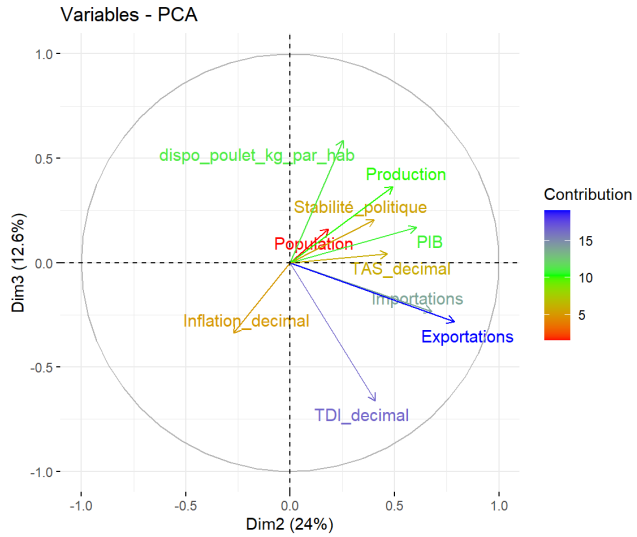


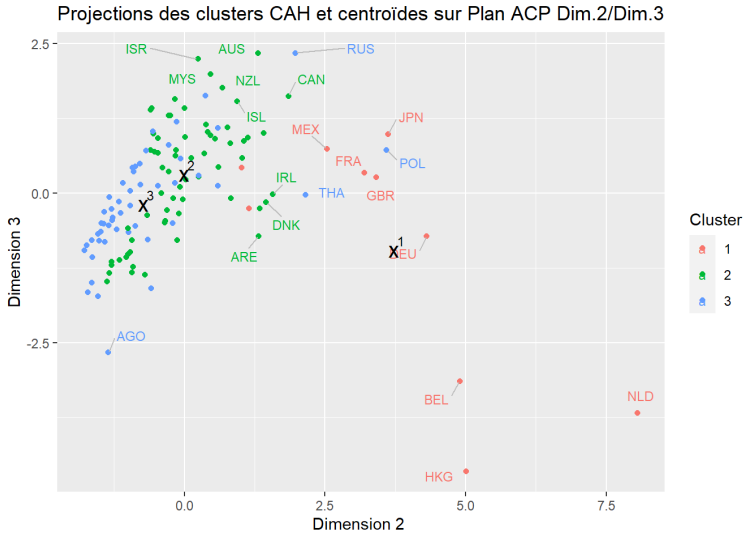
Le cluster 1 est celui qui dépend le plus des importations.

Le cluster 2 est composés de pays plus stables politiquement et plus de dispo par habitant.

Le cluster 3 comportent des pays à forte population et production.

- Plan (2,3)



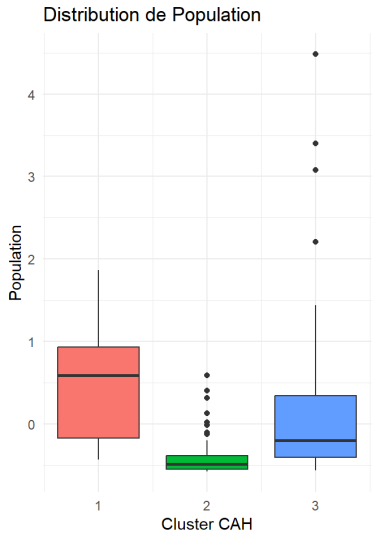


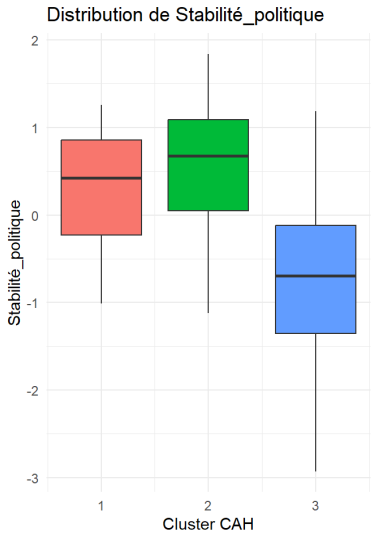
Le cluster 1 semble être composé des pays qui importent, exportent le plus et dépendent le plus de l'importation.

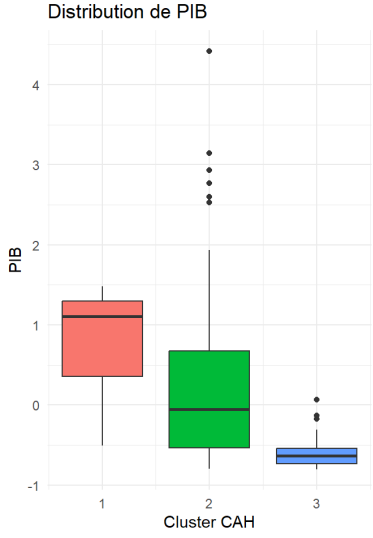
Le cluster 2 composés de pays qui a plus de dispo par habitant.

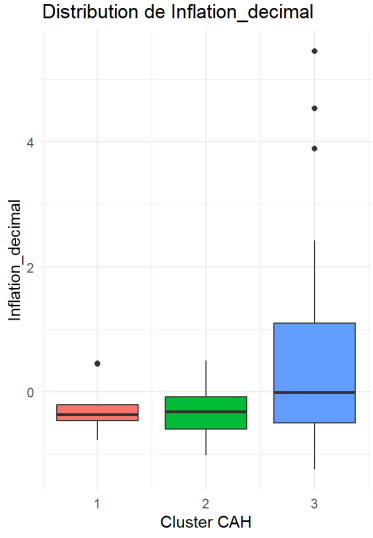
5.1.4 Caractérisation des clusters avec les variables initiales

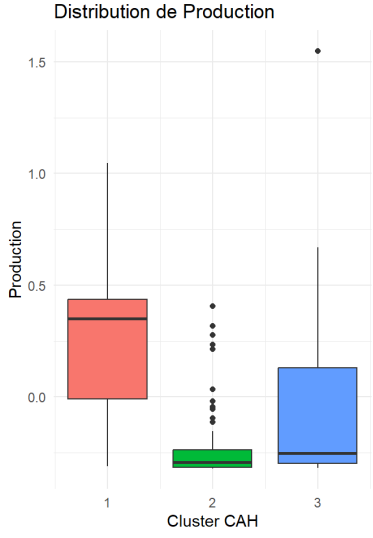
Nous allons caractériser chacun de ces clusters avec nos variables initiales.

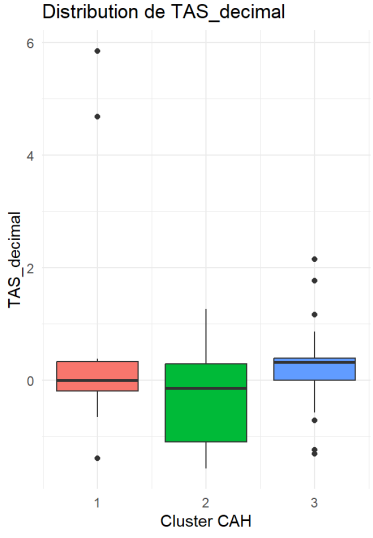

Distribution de Population
Population
Cluster CAH

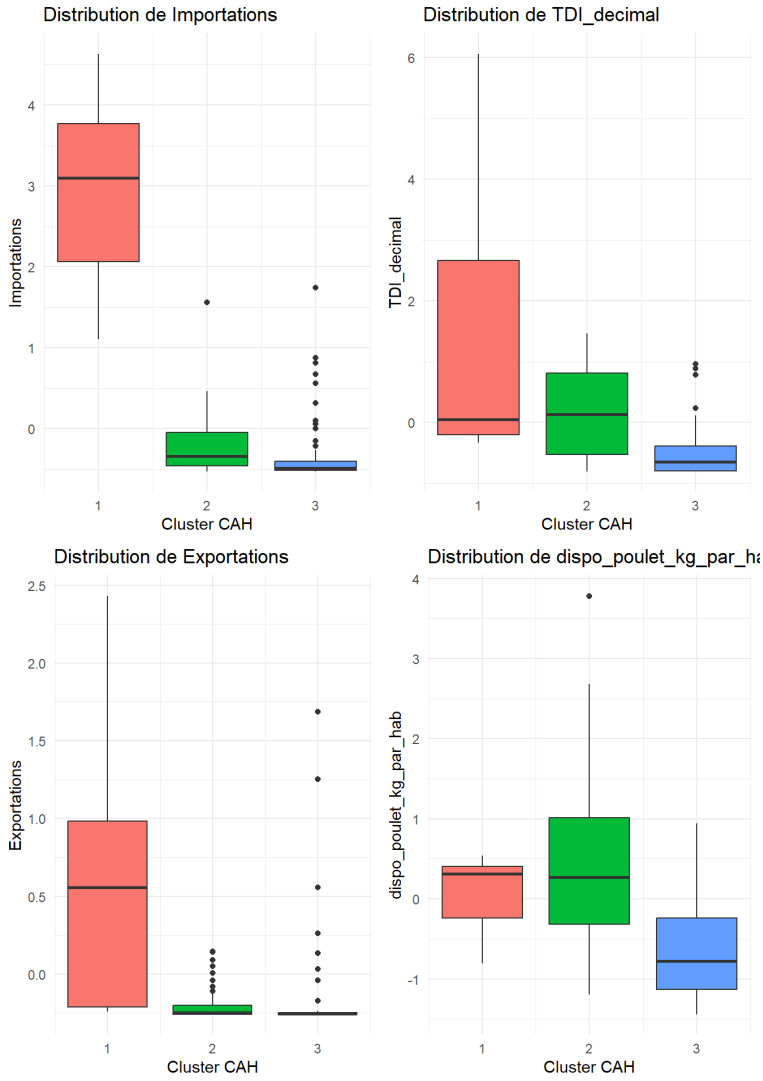

Distribution de Stabilité_politique
Stabilité_politique
Cluster CAH


Distribution de PIB
PIB
Cluster CAH


Distribution de Inflation_decimal
Inflation_decimal
Cluster CAH


Distribution de Production
Production
Cluster CAH


Distribution de TAS_decimal
TAS_decimal
Cluster CAH



```
## TableGrob (1 x 2) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
## TableGrob (1 x 2) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
## TableGrob (1 x 2) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
## TableGrob (1 x 2) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
## TableGrob (1 x 2) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
```

Cluster 1: Pays riches et stables politiquement qui importent beaucoup

- Population Fort
- Stabilité politique Forte
- PIB Fort
- Inflation Faible
- Production Fort
- TAS Moyen
- Importations Fort
- TDI Fort
- Exportations Fort
- Disponibilité de poulet Moyen

Cluster 2: Consommation de poulet forte mais Faibles importations

- Population Faible
- Stabilité politique Fort
- PIB Moyen
- Inflation Faible
- Production Faible
- TAS Faible
- Importations Faible
- TDI Moyen
- Exportations Faible
- Disponibilité de poulet Fort

Cluster 3: Pays en difficulté économique et politique

- Population Moyen
- Stabilité politique Faible
- PIB Faible
- Inflation Fort
- Production Moyen
- TAS Fort
- Importations Faible
- TDI Faible
- Exportations Faible
- Disponibilité de poulet Faible

Dans l'ensemble, le Cluster 1 semble présenter un environnement économique et politique favorable à l'exportation en raison de sa stabilité politique, de son PIB élevé, de ses faibles niveaux d'inflation et de ses fortes importations et exportations. Bien que la disponibilité de poulet soit classée comme "moyenne", elle peut toujours indiquer un marché potentiel.

5.1.5 Liste des pays du cluster ciblé avec la méthode CAH

Code_iso	Zone
BEL	Belgique
DEU	Allemagne
FRA	France
GBR	Royaume-Uni de Grande-Bretagne et d'Irlande du Nord
HKG	Chine - RAS de Hong-Kong
JPN	Japon
MEX	Mexique
NLD	Pays-Bas
SAU	Arabie saoudite
ZAF	Afrique du Sud

5.2 Kmeans

L'algorithme des k-means consiste à regrouper les individus dans k classes les plus homogènes possibles.

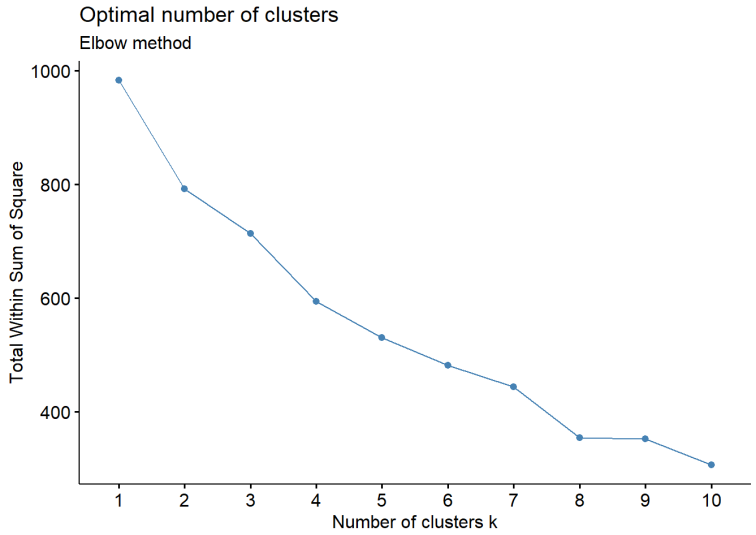
Au départ, elle place aléatoirement K centres (un pour chaque cluster). Ensuite, elle attribue chaque point au cluster dont le centre est le plus proche. Puis, elle déplace les centres pour minimiser la distance moyenne entre les points et leur centre. Elle répète ces étapes jusqu'à ce que les centres ne bougent plus beaucoup.

5.2.1 Détermination du nombre optimal de clusters

- Elbow method

On va générer le graphique qui montre la somme des carrés des distances intra-clusters (WCSS) pour différents nombres de clusters (k). En d'autres termes, le WCSS mesure à quel point les points de données à l'intérieur de chaque cluster sont proches les uns des autres. Plus le WCSS est faible, plus les points au sein d'un cluster sont similaires.

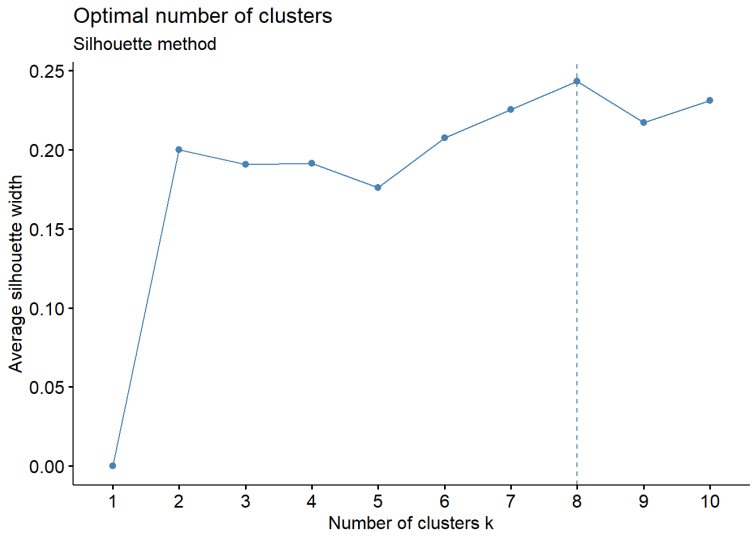
Le point d'inflexion sur le graphique du coude est utilisé pour déterminer le nombre optimal de clusters. Ce coude représente le point où la réduction du WCSS ralentit considérablement à mesure que vous augmentez le nombre de clusters.



La méthode du coude semble indiqué un découpage en 8 clusters.

- Méthode de silhouette

Nous allons confirmer avec une autre méthode de détermination: La méthode de silhouette. Elle évalue la qualité de la séparation des clusters en calculant les valeurs de silhouette pour différents nombres de clusters. La silhouette mesure à quel point chaque objet est similaire à son propre cluster par rapport aux autres clusters. Un score de silhouette élevé indique une meilleure séparation des clusters.



Les 2 méthodes nous indiquent un nombre optimal de 8 clusters. Voyons le nombre de pays par cluster que ce découpage nous donne.

8 clusters peut paraître beaucoup. Nous avons essayer la méthode avec un découpage en k=5 et k=6 clusters, cela a conduit à un regroupement qui inclut la Belgique, les Pays-Bas et Hong-Kong comme pays d'intérêt. Un découpage en 8 nous permet d'obtenir une liste de pays plus élaborée.

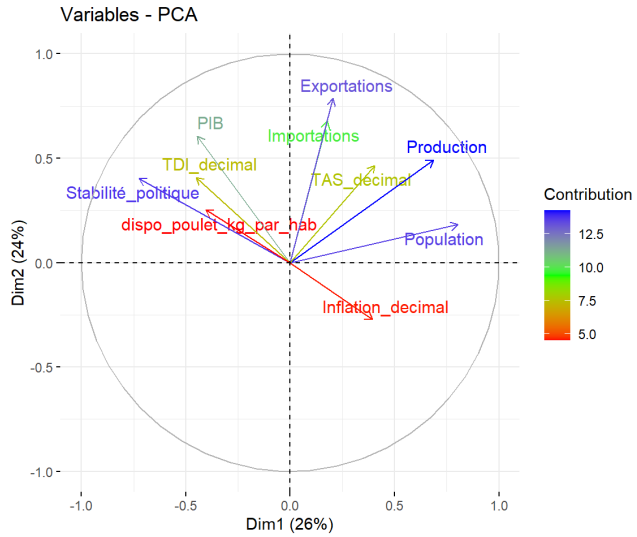
Var1	Freq
1	13
2	28
3	8
4	2
5	24
6	6
7	9
8	35

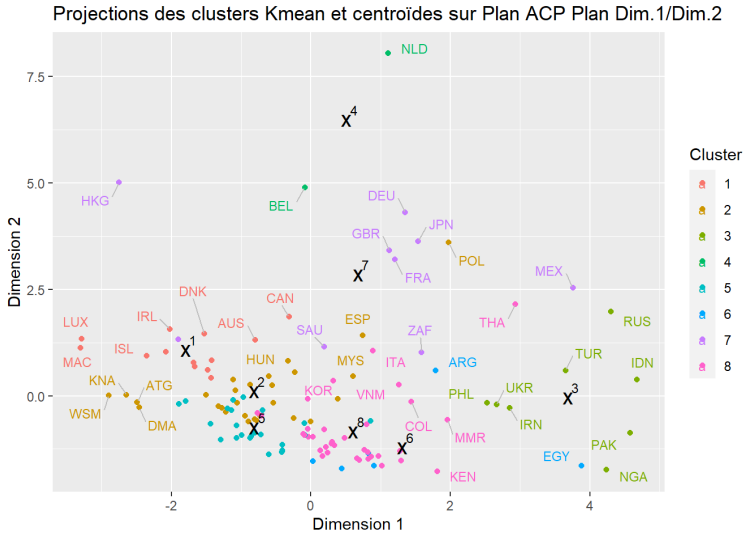
5.2.2 Indice de silhouette

Cluster	Silhouette_Mean_Kmeans
1	0.33
2	0.19
3	0.12
4	0.66
5	0.31
6	0.15
7	0.16
8	0.27
Overall	0.25

5.2.3 Projection des clusters K-means et centroïdes sur les plans ACP

- Plan (1,2)



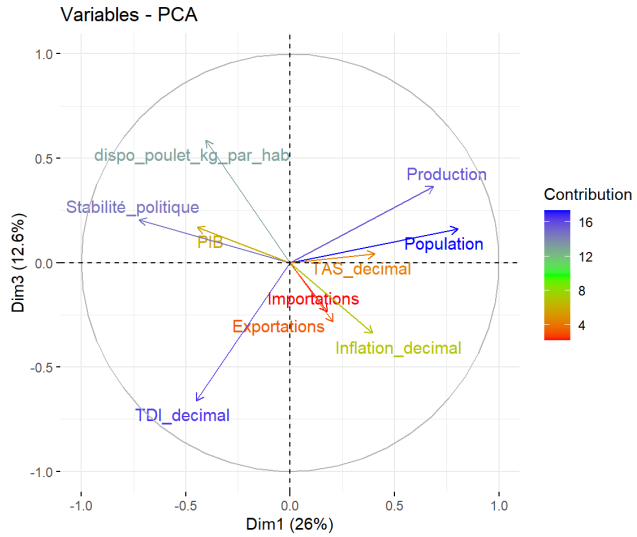


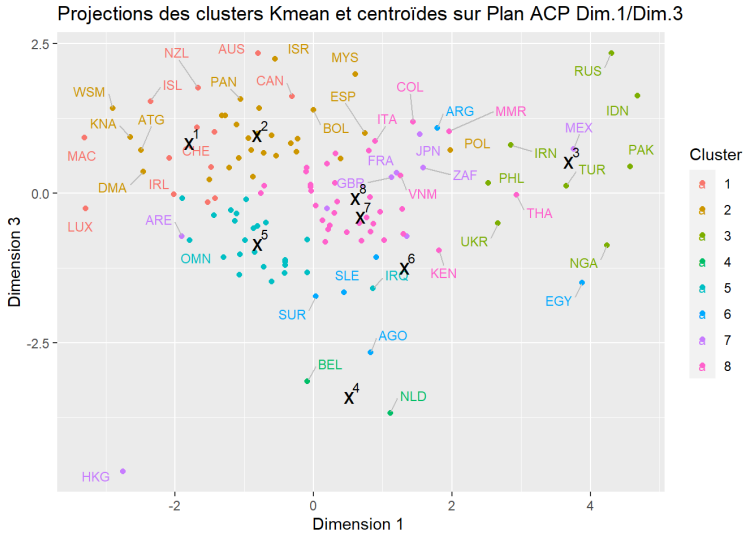
Le cluster 1 lui a une stabilité politique élevé.

Le cluster 3 représente des pays très peuplés qui produisent beaucoup.

Le cluster 4 est composé des pays qui ont un PIB élevé et qui exportent le plus , suivi du cluster 7.

- Plan (1,3)





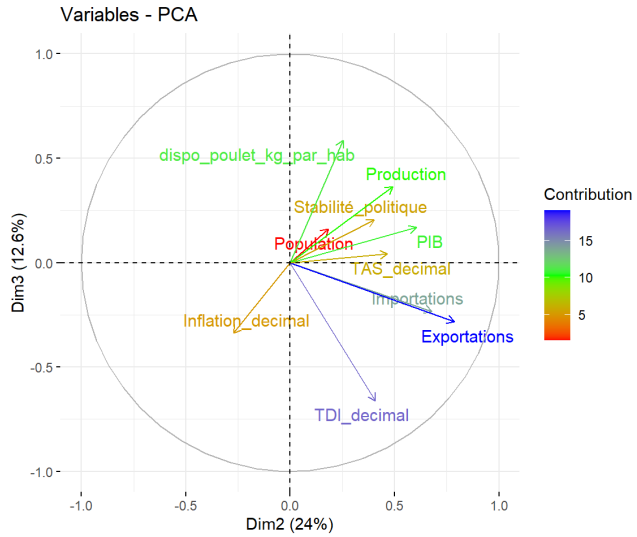
Le cluster 1 lui a une stabilité politique élevé.

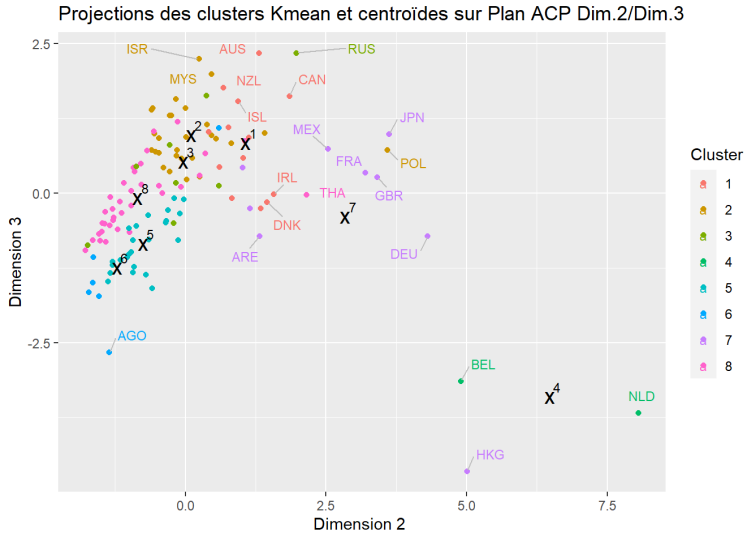
Le cluster 2 a une dispo en poulet élevé suivi du cluster 1.

Le cluster 3 a une population et production élevé.

Le cluster 4 dépend beaucoup de l'importation.

- Plan (2,3)



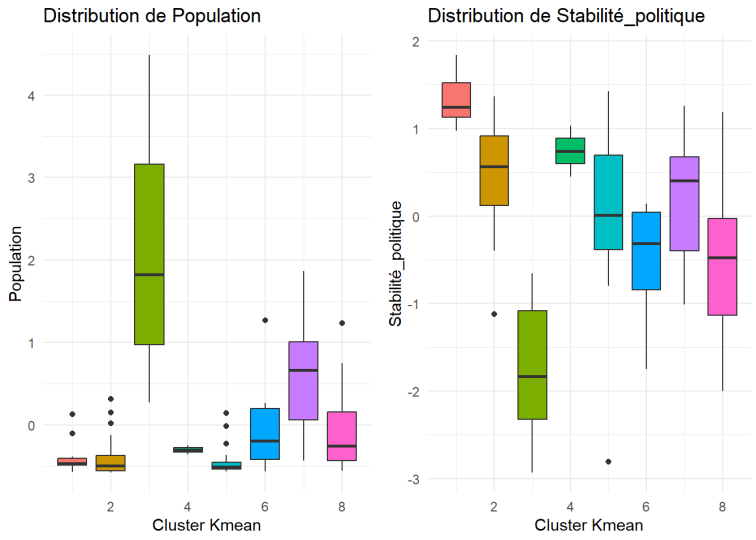


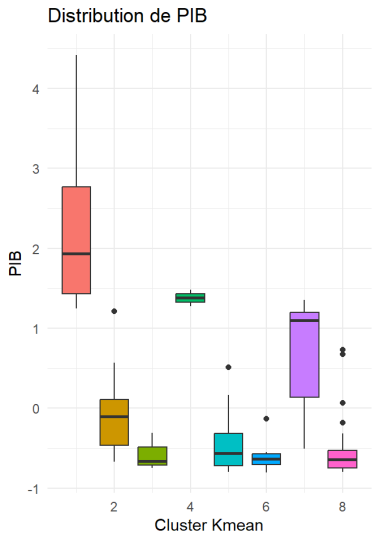
Le cluster 2 a une dispo en poulet élevé suivi du cluster 1.

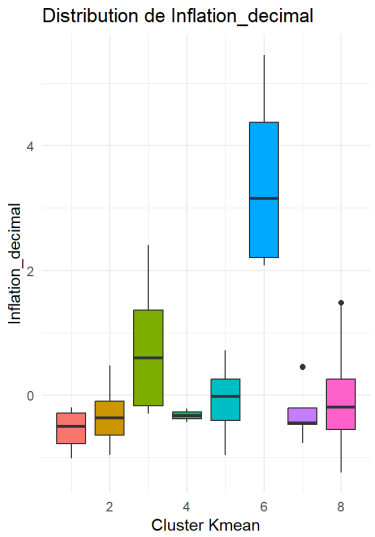
Le cluster 4 est composé de pays qui importent beaucoup, exportent beaucoup et dépendent beaucoup des importations.

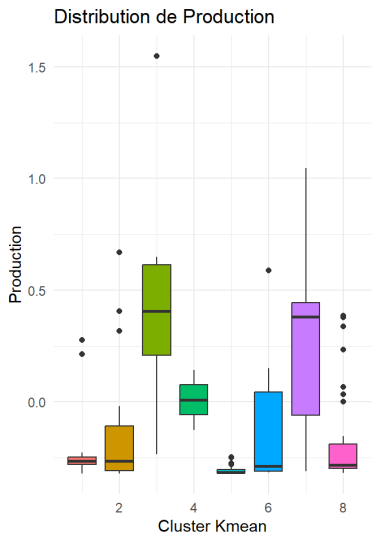
Le cluster 7 est composé de pays qui importent et exportent beaucoup.

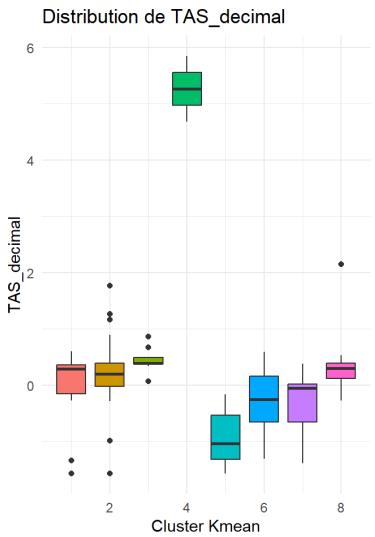
5.2.4 Caractérisation des clusters avec les variables initiales

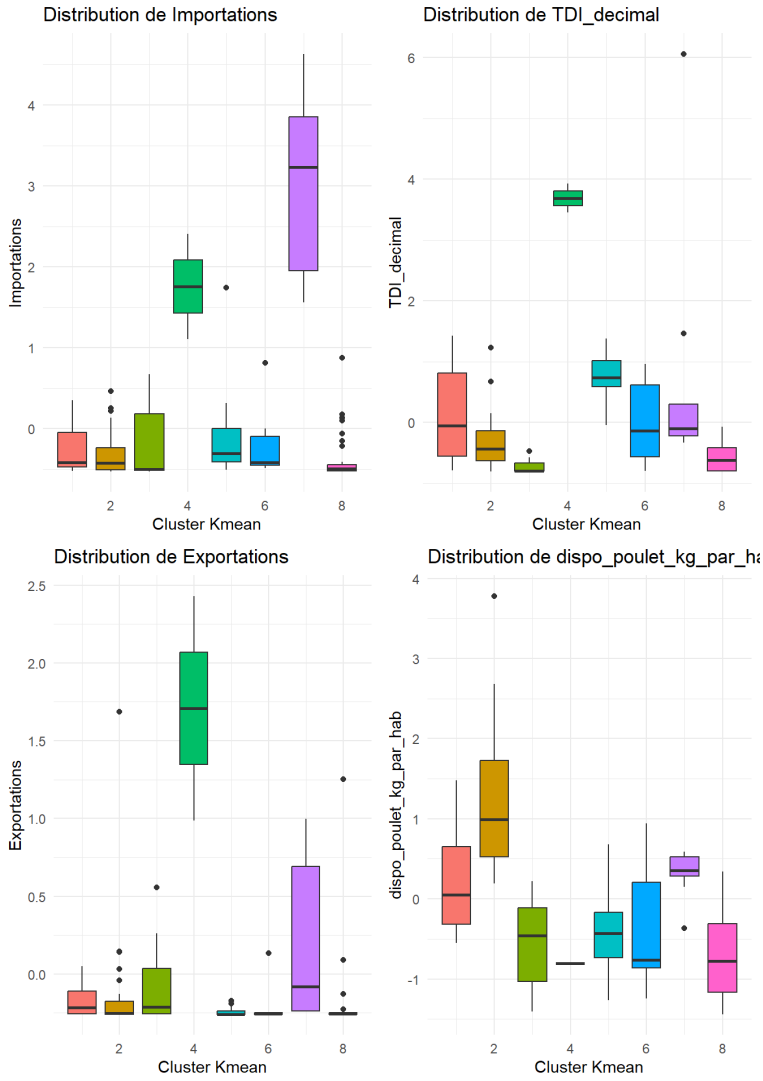












```
## TableGrob (1 x 2) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
## TableGrob (1 x 2) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
## TableGrob (1 x 2) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
## TableGrob (1 x 2) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
## TableGrob (1 x 2) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
```

```
## [[1]]
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
##
## [[2]]
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
##
## [[3]]
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
##
## [[4]]
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
##
## [[5]]
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
```

Cluster 1: Pays les plus riches et les plus stables politiquement, faibles importtaions

- Population faible
- Stabilité politique forte
- PIB fort
- Inflation faible
- Production faible
- TAS moyen
- Importations faibles
- TDI moyen
- Exportations faibles
- Disponibilité de poulet moyenne

Cluster 2: Pays les plus consommateurs mais les moins importateurs

- Population faible
- Stabilité politique forte
- PIB moyen
- Inflation faible
- Production faible
- TAS moyen
- Importations faibles
- TDI faible
- Exportations faibles
- Disponibilité de poulet élevée

Cluster 3: Pays très peuplés avec un PIB faible

- Population élevée
- Stabilité politique faible
- PIB faible
- Inflation moyenne
- Production élevée
- TAS moyen
- Importations faibles
- TDI faible
- Exportations faibles
- Disponibilité de poulet faible

Cluster 4: Pays qui importent beaucoup et exportent le plus, autosuffisant et dépendent le plus de l'importation

- Population faible
- Stabilité politique forte
- PIB fort
- Inflation faible
- Production moyenne
- TAS élevé
- Importations élevées
- TDI élevé
- Exportations élevées
- Disponibilité de poulet faible

Cluster 5: Pays qui produisent le moins et PIB faible

- Population faible
- Stabilité politique moyenne
- PIB faible
- Inflation faible
- Production faible
- TAS faible
- Importations faibles
- TDI moyen
- Exportations faibles
- Disponibilité de poulet faible

Cluster 6: Pays qui ont la plus forte inflation

- Population faible
- Stabilité politique moyenne
- PIB faible
- Inflation élevée
- Production faible
- TAS moyen
- Importations faibles
- TDI moyen
- Exportations faibles
- Disponibilité de poulet faible

Cluster 7: Pays qui importent le plus avec PIB et stabilité politique forts

- Population moyenne
- Stabilité politique forte
- PIB fort
- Inflation faible
- Production élevée
- TAS moyen
- Importations élevées
- TDI moyen
- Exportations moyennes
- Disponibilité de poulet moyenne

Cluster 8: Pays qui exportent le moins et PIB faible

- Population faible
- Stabilité politique moyenne
- PIB faible
- Inflation faible
- Production faible
- TAS moyen
- Importations faibles

- TDI faible
- Exportations faibles
- Disponibilité de poulet faible

Le Cluster 7 pourrait être intéressant à considérer, car il correspond aux pays qui importent le plus. Les caractéristiques de ce cluster comprennent une population moyenne, une stabilité politique forte, un PIB élevé, une inflation faible, une production élevée et des importations élevées. Cela pourrait indiquer une forte demande de produits importés.

Le cluster 4 pourrait être aussi être des pays intéressant. Pays qui importent aussi beaucoup et dépendent beaucoup de l'importation.

5.2.5 Liste des pays ciblés avec la méthode Kmeans

Le Cluster 7 est composé de:

Code_iso	Zone
ARE	Émirats arabes unis
DEU	Allemagne
FRA	France
GBR	Royaume-Uni de Grande-Bretagne et d'Irlande du Nord
HKG	Chine - RAS de Hong-Kong
JPN	Japon
MEX	Mexique
SAU	Arabie saoudite
ZAF	Afrique du Sud

Le cluster 4 est composé de:

Code_iso	Zone
BEL	Belgique
NLD	Pays-Bas

L'ensemble des pays ciblés par la méthode Kmean:

Code_iso	Zone
ARE	Émirats arabes unis
DEU	Allemagne
FRA	France
GBR	Royaume-Uni de Grande-Bretagne et d'Irlande du Nord
HKG	Chine - RAS de Hong-Kong
JPN	Japon
MEX	Mexique
SAU	Arabie saoudite
ZAF	Afrique du Sud
BEL	Belgique
NLD	Pays-Bas

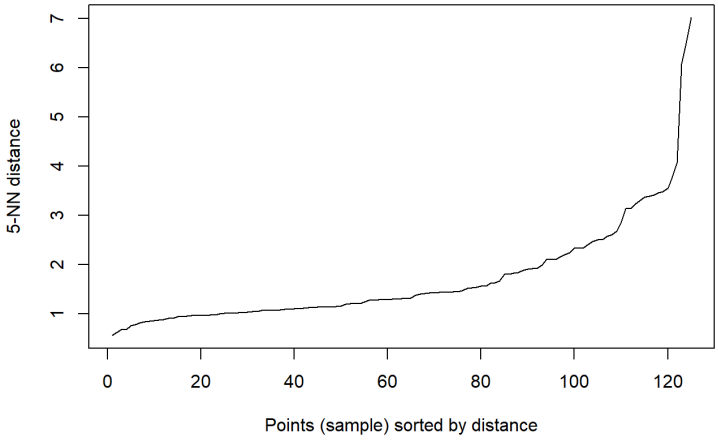
5.3 DBScan Density

Le DBSCAN est un algorithme de clustering basé sur la densité. Il regroupe les points de données qui sont proches les uns des autres dans un espace de grande densité, et marque les points isolés comme du bruit.

5.3.1 Détermination du nombre optimal de eps

Choisir les paramètres de l'algorithme DBSCAN : le nombre minimum de points (MinPts) requis pour former un cluster et le rayon (epsilon) autour d'un point à considérer pour la formation d'un cluster.

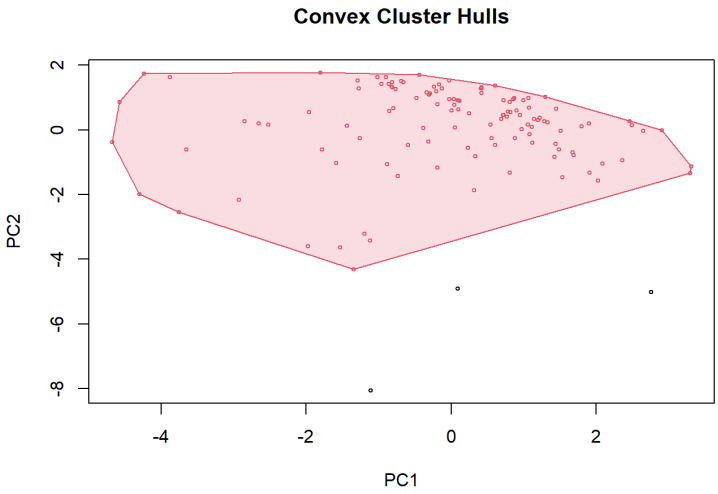
La courbe de k-distance peut être utilisée pour déterminer un bon choix de k pour l'algorithme DBSCAN. Le bon choix de k est généralement situé à un point où la courbe montre une forte augmentation des distances (un "coude"). Cela peut indiquer le nombre de voisins à partir duquel un point commence à être plus isolé des autres points et donc représente un bon candidat pour être un point de départ pour un nouveau cluster.



On lance le DBscan avec eps=4.

Var1	Freq
0	3
1	122

La méthode DBScan Density nous a isolé 3 pays.

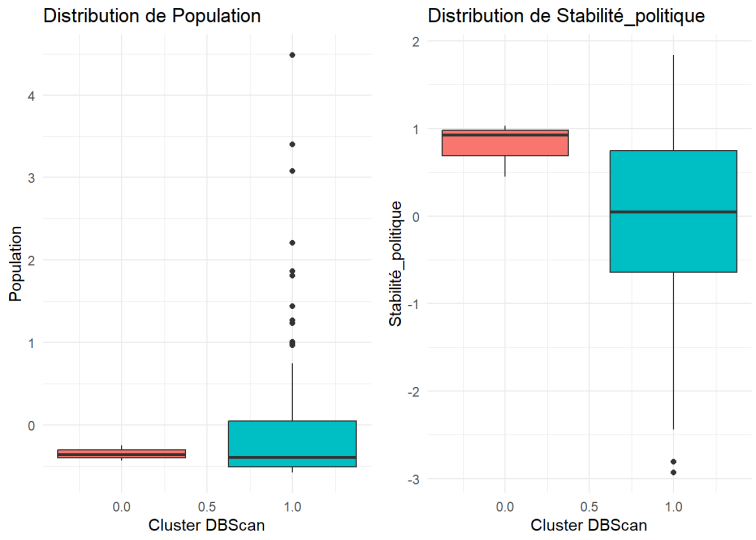


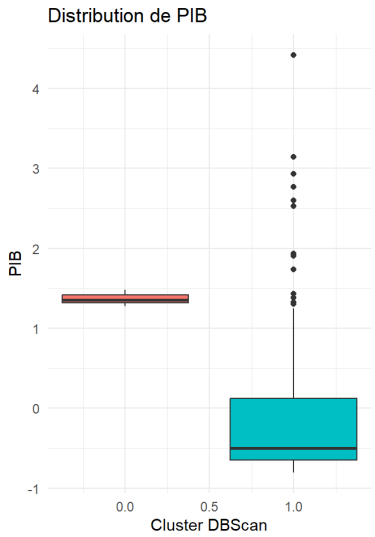
La methode a rassemblé un cluster et a identifié 3 individus comme étant du bruit.

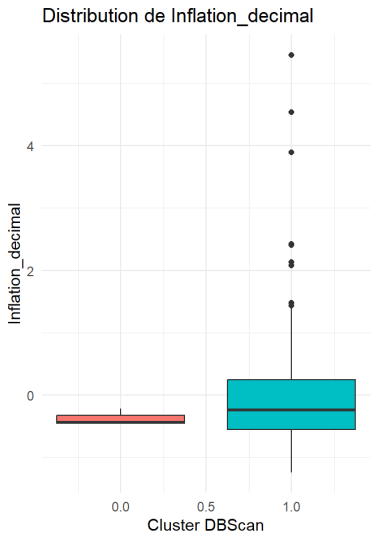
5.3.2 Indice de silhouette

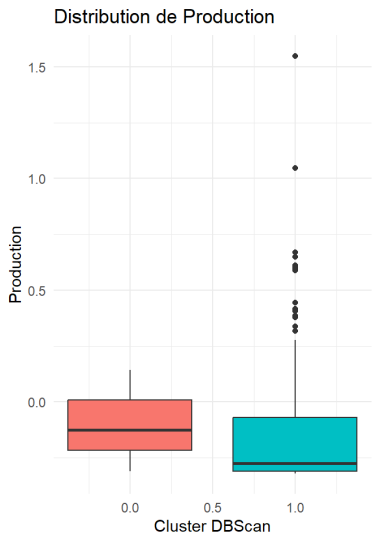
Cluster	Silhouette_Mean_dbscan
0	0.26
1	0.57
Overall	0.56

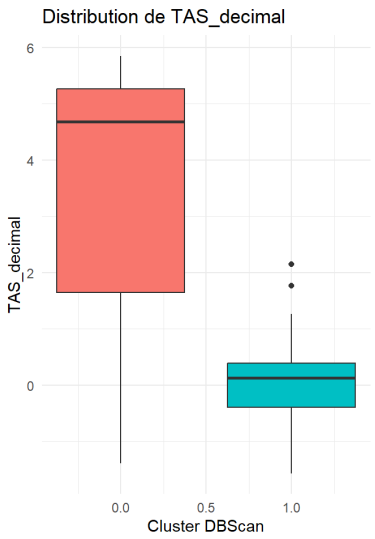
5.3.3 Caractérisation des clusters avec les variables initiales

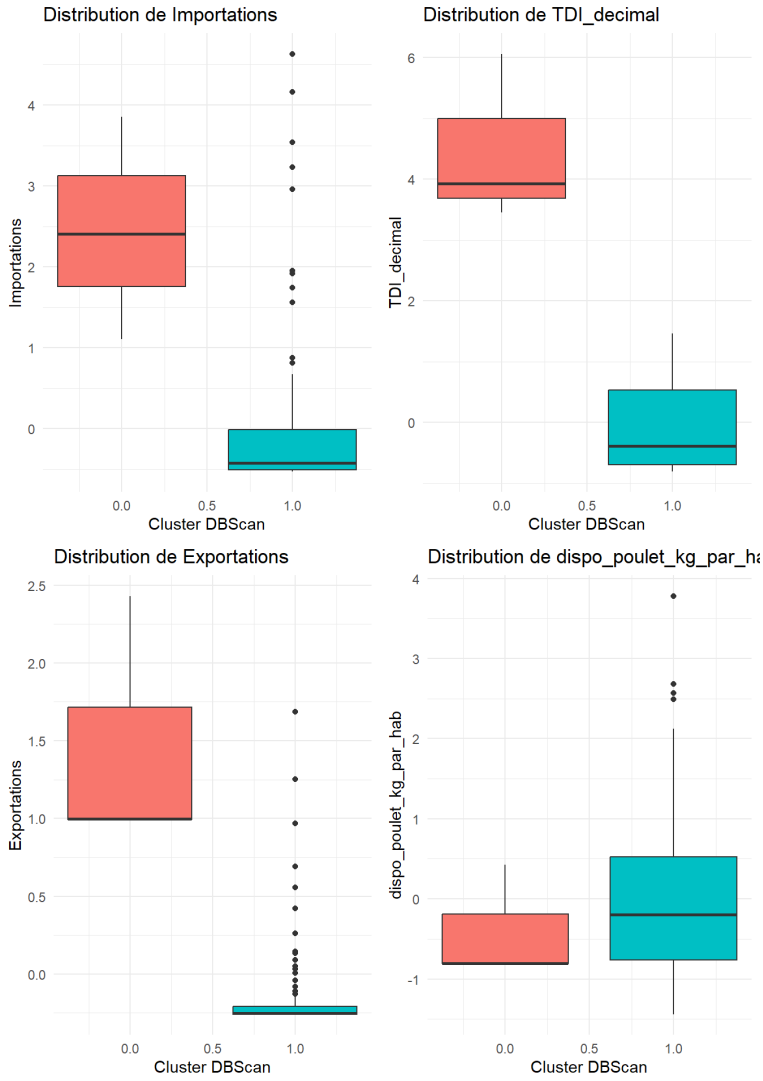












```
## TableGrob (1 x 2) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
## TableGrob (1 x 2) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
## TableGrob (1 x 2) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
## TableGrob (1 x 2) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
## TableGrob (1 x 2) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
```

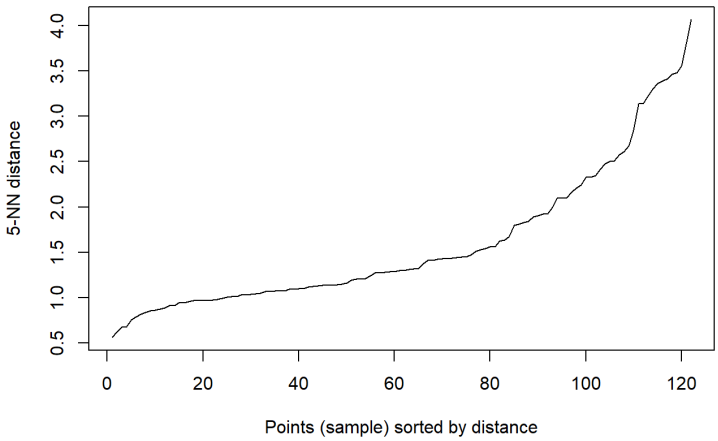
Nous constatons que ces trois pays se caractérisent par : une population et une inflation faibles, une production moyenne et des niveaux élevés de stabilité politique, de PIB, de TAS, d'importations, de TDI et d'exportations.

5.3.4 Liste des pays ciblés avec la méthode DBScan density

Code_iso	Zone
BEL	Belgique
HKG	Chine - RAS de Hong-Kong
NLD	Pays-Bas

On relance l'analyse après avoir supprimé ces 3 pays de notre liste afin d'essayer d'identifier d'autres clusters.

5.3.5 Relance nouveau DBScan sans les 3 premiers pays ciblés



On choisit eps=3.5

Var1	Freq
1	122

La méthode ne nous isole plus aucun pays.

5.4 Analyse Comparative des Résultats des 3 Méthodes de Clustering

5.4.1 Comparaison des Pays sélectionnés par Chaque Méthode de Clustering

Zone	CAH	Kmeans	DBScan
Belgique	BEL	BEL	BEL
Allemagne	DEU	DEU	
France	FRA	FRA	
Royaume-Uni de Grande-Bretagne et d'Irlande du Nord	GBR	GBR	
Chine - RAS de Hong-Kong	HKG	HKG	HKG
Japon	JPN	JPN	
Mexique	MEX	MEX	
Pays-Bas	NLD	NLD	NLD
Arabie saoudite	SAU	SAU	
Afrique du Sud	ZAF	ZAF	
Émirats arabes unis		ARE	

Les trois méthodes ont identifié des pays comme la Belgique, les Pays-Bas et Hong-Kong comme des cibles potentielles pour l'exportation de poulets.

5.4.2 Comparaison des Moyennes des Indices de Silhouette pour les Différentes Méthodes de Clustering

Cluster	Silhouette_Mean_CAH	Silhouette_Mean_Kmeans	Silhouette_Mean_DBScan
0	NA	NA	0.26
1	0.07	0.33	0.57
2	0.24	0.19	NA
3	0.18	0.12	NA
4	NA	0.66	NA
5	NA	0.31	NA
6	NA	0.15	NA
7	NA	0.16	NA
8	NA	0.27	NA
Overall	0.20	0.25	0.56

L'indice de silhouette n'est pas une mesure adaptée pour DBSCAN car il ne crée pas nécessairement des partitions cohérentes dans l'espace des données comme le font K-Means ou la classification ascendante hiérarchique (CAH). Au lieu de cela, DBSCAN identifie les régions de densité élevée et sépare les zones de faible densité, ce qui peut donner lieu à des groupes de formes et de tailles différentes, et certaines données peuvent être considérées comme du bruit.

On peut voir que le clustering DBScan a le silhouette score le plus élevé (0.61), suivi par le clustering KMeans (0.25) et le clustering CAH (0.20). Malgré le score de silhouette plus élevé pour DBScan, nous choisissons le clustering Kmeans qui a potentiellement mieux séparé les clusters par rapport aux autres méthodes.

6 Liste finale des pays à étudier

En complément des pays sélectionnés, d'autres pays pourraient également avoir de la pertinence pour votre marché comme La Chine continentale qui a été exclue en raison de l'absence de données politiques.

Les pays tels que l'Allemagne, Belgique, Pays-Bas et le Royaume-Uni offrent des avantages logistiques, tels que des liaisons de transport efficaces ou des frontières terrestres partagées, qui pourraient réduire les coûts de transport et les délais de livraison.

Vous trouverez ci-dessous un tableau répertoriant tous ces pays avec leurs données initiales ainsi que leur distance par rapport à Paris en km et les moyens de transport :

Zone	Population	Inflation	Stabilité_politique	PIB	Production	Importations	Exportations	TAS	TDI	dispo_poulet_kg_par_hab	Distance_Paris_cap
Afrique du Sud	57009756	6.90	-0.28	6723.93	1667	514	63	78.71	24.27	37.15	9
Allemagne	82658409	2.76	0.59	44670.22	1514	842	646	88.54	49.24	20.34	1
Arabie saoudite	33101179	-0.82	-0.64	20138.15	616	722	10	46.39	54.37	36.86	4
Belgique	11419748	1.27	0.43	44162.26	463	338	656	319.31	233.10	12.17	
Chine - RAS de Hong-Kong	7306322	1.13	0.83	45737.48	24	907	663	8.96	338.43	35.04	9
Chine, continentale	1421021791	-0.19	NA	8729.14	18236	452	576	100.68	2.50	12.71	8
Émirats arabes unis	9487203	1.21	0.62	42522.38	48	433	94	12.40	111.89	38.05	5
Japon	127502725	0.70	1.11	38928.95	2215	1069	10	67.65	32.65	32.41	9
Mexique	124777324	6.97	-0.80	9434.38	3249	972	9	77.14	23.08	33.71	9
Pays-Bas	17021347	2.66	0.92	48460.51	1100	608	1418	379.31	209.66	12.22	
Royaume-Uni de Grande-Bretagne et d'Irlande du Nord	66727461	2.26	0.39	40617.68	1814	779	359	81.20	34.87	33.48	

L'analyse a permis d'identifier plusieurs pays cibles potentiels pour l'exportation de poulets. La décision finale pourrait reposer sur d'autres paramètres propres à l'entreprise, tels que la compétitivité sur les marchés locaux et internationaux, les coûts liés à l'exportation…

L'entreprise "Poulet Mondial" peut maintenant utiliser ces informations pour prendre des décisions éclairées sur sa stratégie d'exportation.