La poule qui chante

Présenté par: ARTEMOV Pavel

La poule qui chante

Ordre du jour

I) Le contexte

Présentation du contexte et des objectifs de l'étude

II) La démarche utilisée

Description de toutes les étapes qui ont permis le clustering des pays en groupes homogènes.

III) Résultats et recommandations

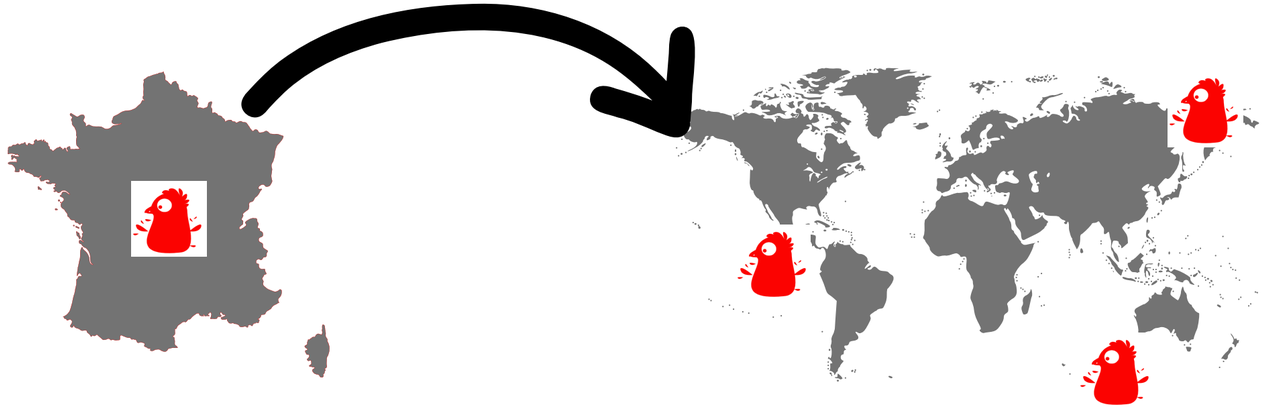
Présentation des groupes et de leurs caractéristiques.

La poule qui chante

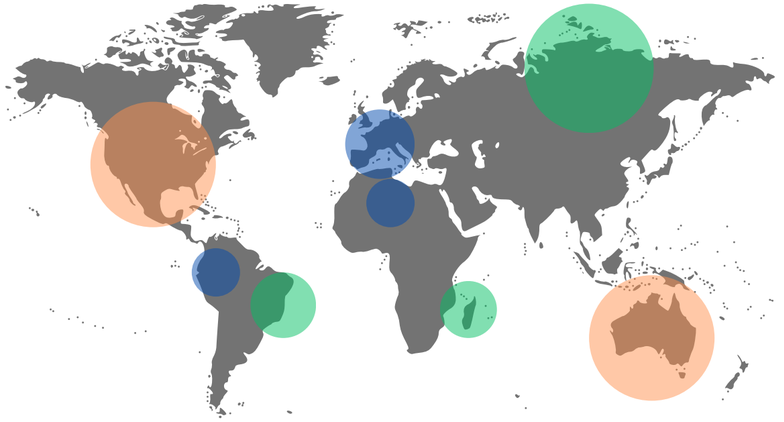
I) Le contexte

L'objectif de l'étude:

Le lancement à l'international de nos poulets



Export de nos poulets...



...de manière intelligente

La poule qui chante

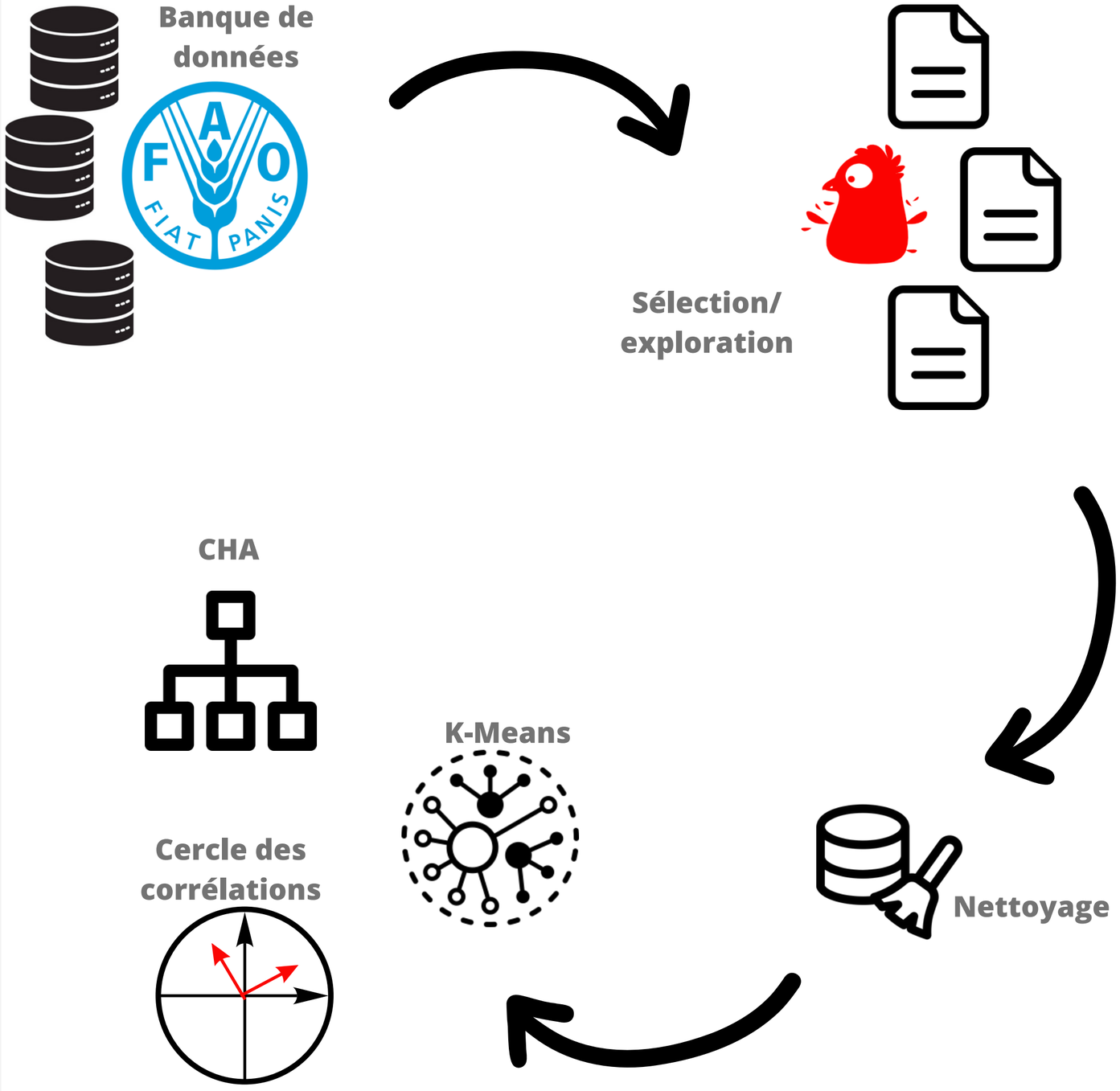
II) La démarche

A - Données:

- Sélection de la donnée (FAO)
- Traitement (nettoyage, vérification d'importation, format...)
- Choix des variables finales pour l'étude

B - Analyse exploratoire des données:

- Clustering (CHA, K-Means): groupement des individus
- Analyse en composante principale: groupement de variables en variables synthétiques
- Projection des clusters sur les axes synthétiques



La poule qui chante

A - Préparation des données

La poule qui chante

5 Data Frames:

- Population
- Disponibilité alimentaire
- Stabilité politique
- PIB
- Volaille

Extraction de 12 variables

- 11 variables quantitatives
- 1 variable qualitative (nom du pays)

Vérification et mise en conformité

- Valeurs manquantes
- Type de donnée
- Changement de nom ("Valeur")
- Création de nouvelles colonnes
- Merge et export

2.4 Vérification des valeurs dans les colonnes

```
df_final.shape
```

```
(166, 12)
```

```
# Comptage du nombre de valeurs manquantes par colonne
missing_values_count = df_final.isnull().sum()
missing_values_count
```

```
Zone                                              0
Population totale(1000 pers)                      0
%Femme                                            0
%Population urbaine                               0
PIB(Ma en US $)                                   0
PIB(US $ par habitant)                            0
Dispo interieure (KT)                             0
Disponibilité alimentaire (Kcal/personne/jour)    0
Production de volaille(milliers de T)             0
Importations de volaille(milliers de T)           0
Consommation de volaille(milliers de T)           0
Stabilité politique                               0
dtype: int64
```

```
df_final.dtypes
```

```
Zone                                              object
Population totale(1000 pers)                     float64
%Femme                                           float64
%Population urbaine                              float64
PIB(Ma en US $)                                  float64
PIB(US $ par habitant)                           float64
Dispo interieure (KT)                            float64
Disponibilité alimentaire (Kcal/personne/jour)   float64
Production de volaille(milliers de T)              int64
Importations de volaille(milliers de T)            int64
Consommation de volaille(milliers de T)            int64
Stabilité politique                              float64
dtype: object
```

```
#Export en format CSV

df_final.to_csv('df_final.csv', index=False)
```

La poule qui chante

B - Analyse exploratoire des données

La poule qui chante

a) Clustering

Clustering

- le but du clustering est de regrouper les pays avec des caractéristiques similaires ensemble tout en les différentiant des autres groupes.

Normalisation:

- Centrage (soustraire la moyenne)
- Réduction (diviser par écart-type)
- "coordonnées" de chaque pays dans n(=nombre de variables) dimensions

Méthodes de clustering utilisés

- Classification hiérarchique ascendante (CAH)
- K-Means

Data Frame initial

	Zone	Population totale(1000 pers)	%Femme	%Population urbaine	PIB(Ma en US $)	PIB(US $ par habitant)	Dispo interieure (KT)
0	Afghanistan	35643.418	49.419284	25.170066	1.889635e+04	520.616409	15139.0
1	Afrique du Sud	56641.209	51.510094	65.938130	3.490067e+05	6121.876572	66840.0
2	Albanie	2879.355	49.878532	60.431312	1.301973e+04	4514.204908	4879.0
3	Algérie	41136.546	49.044548	72.370223	1.700970e+05	4109.696001	45365.0
4	Allemagne	82624.374	50.679425	76.783788	3.690849e+06	44651.829102	174960.0

Array normalisé

```
# normaliser les données
scaler = StandardScaler()
df_final_norm = scaler.fit_transform(df_final.iloc[:, 1:])
```

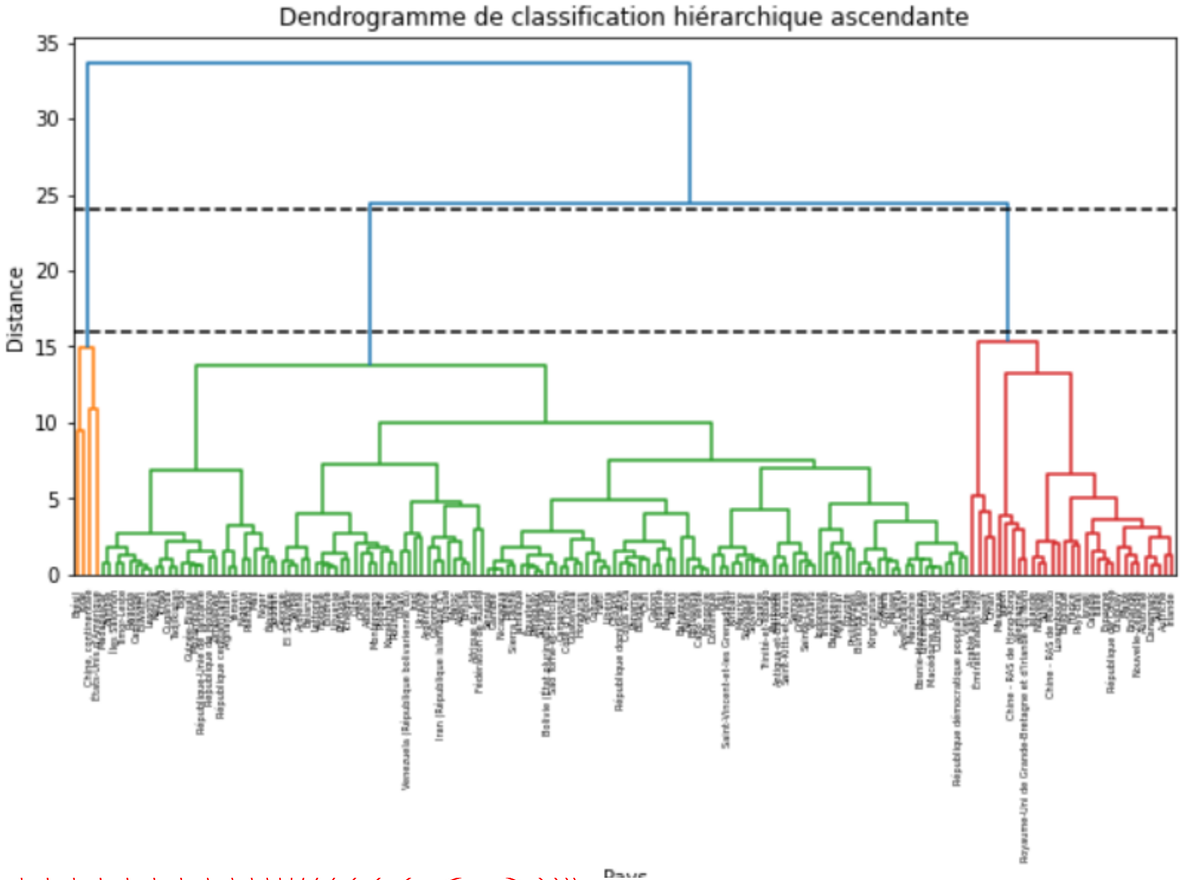
```
df_final_norm
```

```
array([[-0.05596219, -0.32522203, -1.51934607, ..., -0.31963147,
        -0.27926115, -3.03390623],
       [ 0.07902904,  0.48770123,  0.3363566 , ...,  2.27421287,
         0.62930597, -0.23763493],
       [-0.2665968 , -0.14666292,  0.08569429, ..., -0.36776466,
        -0.29302731,  0.49472184],
       ...,
       [-0.09202169, -0.33931704, -1.12304009, ..., -0.03618044,
        -0.19533193, -3.18925463],
       [-0.17390143,  0.17724062, -0.73189483, ..., -0.41054973,
        -0.28103743,  0.2395066 ],
       [-0.19027536,  1.04692333, -1.02071487, ..., -0.44263852,
        -0.27437638, -0.71477646]])
```

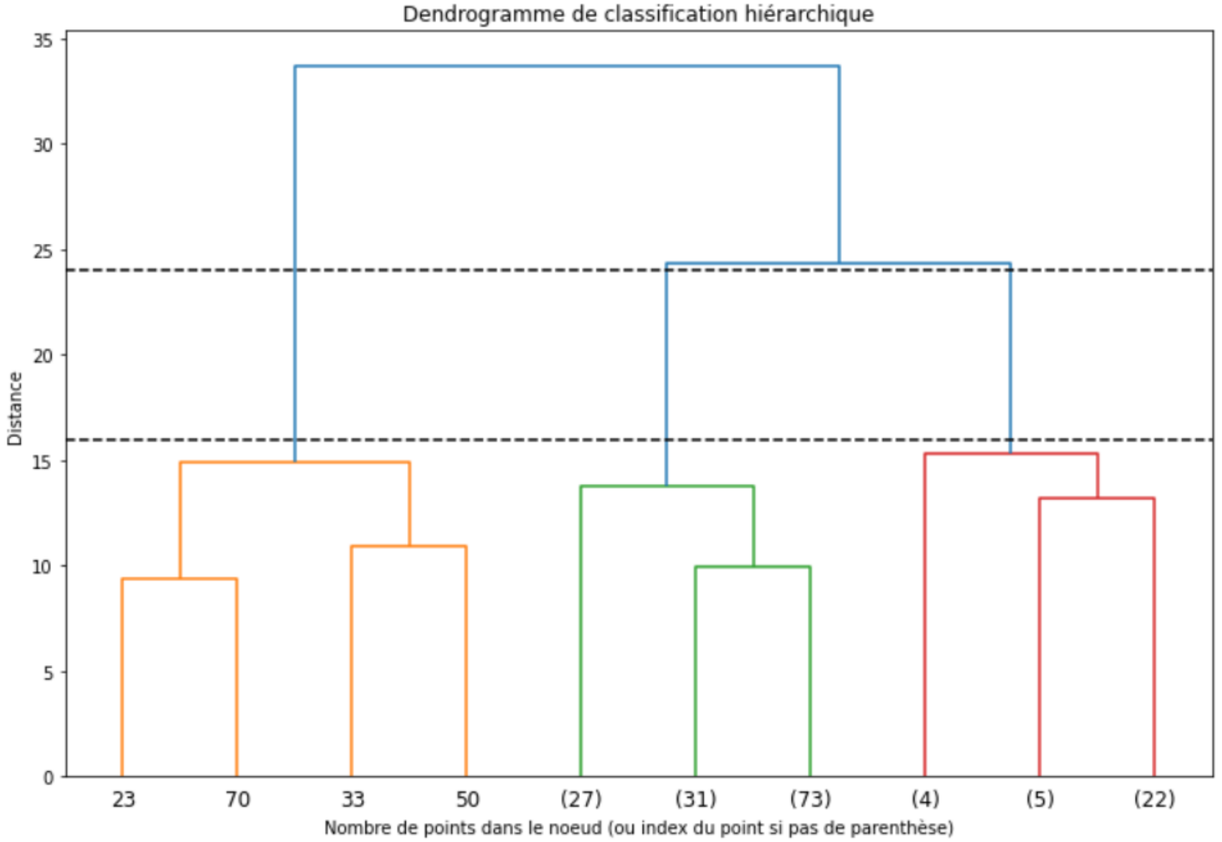
La poule qui chante

Classification ascendante hiérarchique

Détermination du nombre de clusters optimal



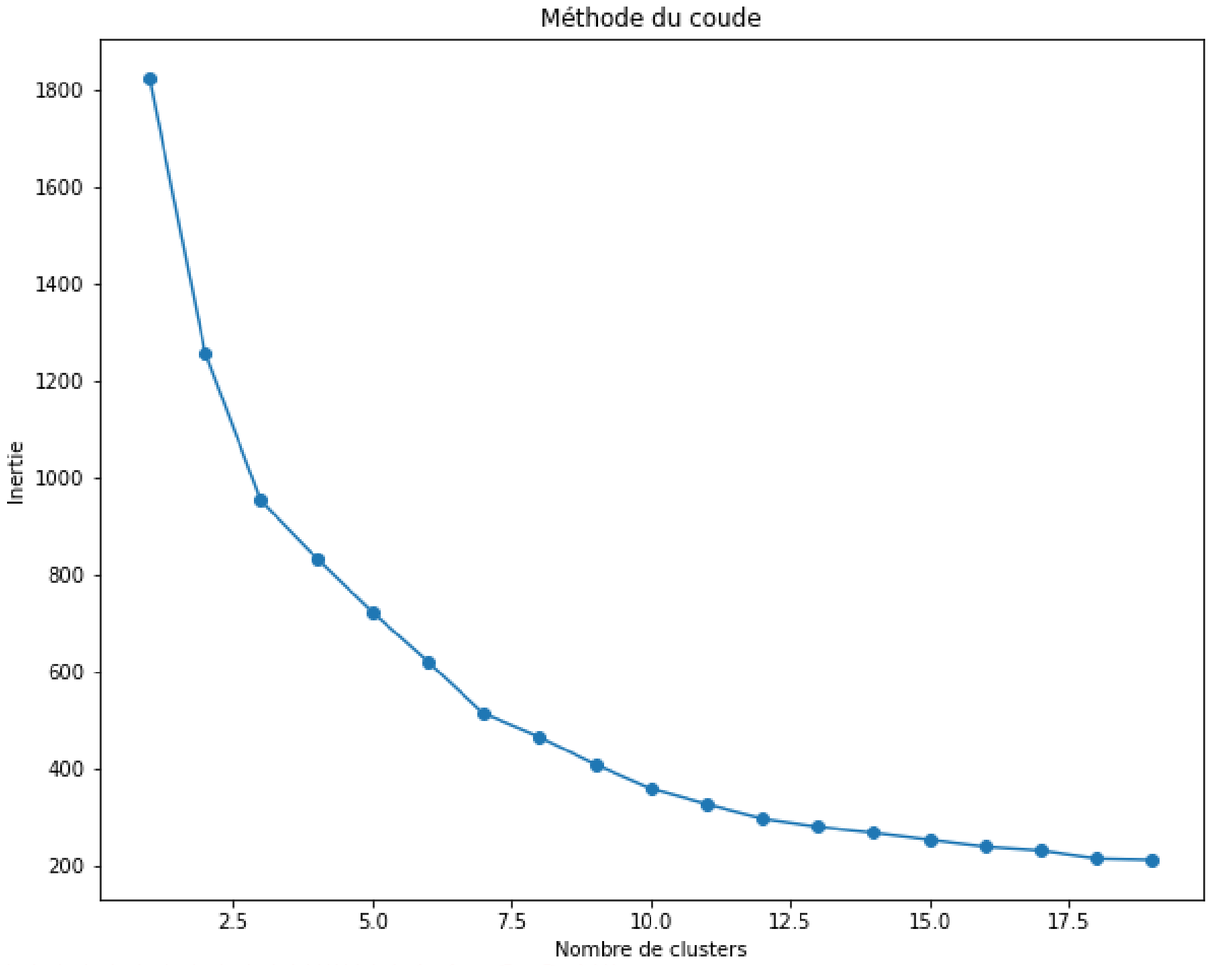
Visualiser les distances entre les clusters



La poule qui chante

K-Means

Détermination du nombre de clusters optimal



Attribution d'un cluster à chaque pays

Disponibilité alimentaire (Kcal/personne/jour)	Production de volaille(milliers de T)	Importations de volaille(milliers de T)	Consommation de volaille(milliers de T)	Stabilité politique	kmeans_group
1997.0	28	29	64	-2.80	1
2987.0	1665	514	2110	-0.28	2
3400.0	13	20	33	0.38	1
3345.0	284	2	286	-0.92	1
3559.0	1514	842	1492	0.59	2

Centroïdes

```
# créer les groupes
k = 3
kmeans = KMeans(n_clusters=k, random_state=40)
kmeans.fit(df_final_norm)

# obtenir les coordonnées des centroïdes
centroids = kmeans.cluster_centers_

# afficher les coordonnées des centroïdes pour chaque groupe
for i in range(k):
    print("Centroïde pour le groupe {}: {}".format(i, centroids[i]))
```

```
Centroïde pour le groupe 0: [ 5.0231259  -0.24682228  0.27681344  4.62077463  0.31023206  5.65149257
  0.69018078  5.48512074  0.29807785  5.37308322 -0.25150532]
Centroïde pour le groupe 1: [-0.1292726   0.0927865  -0.39124886 -0.21553385 -0.46066965 -0.16359183
 -0.41102311 -0.19030806 -0.30136624 -0.19651737 -0.28145597]
Centroïde pour le groupe 2: [-0.1099622  -0.22639813  1.02409344  0.15795216  1.20722934 -0.0750485
  1.03954553  0.01172429  0.78111147  0.03856174  0.77767786]
```

La poule qui chante

b) ACP

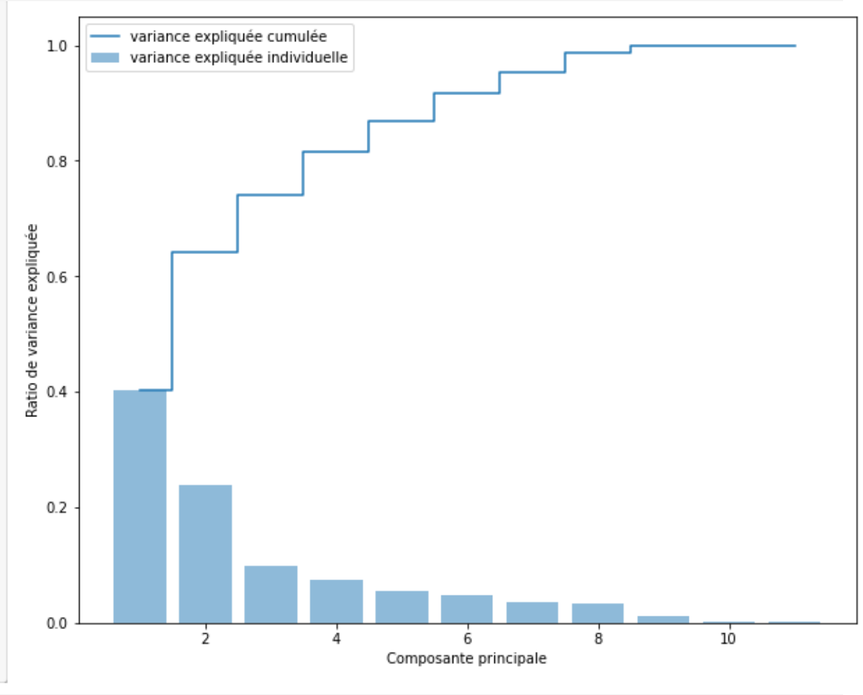
ACP (analyse en composante principale)

- réduction de dimensions
- projection des individus sur des axes synthétiques

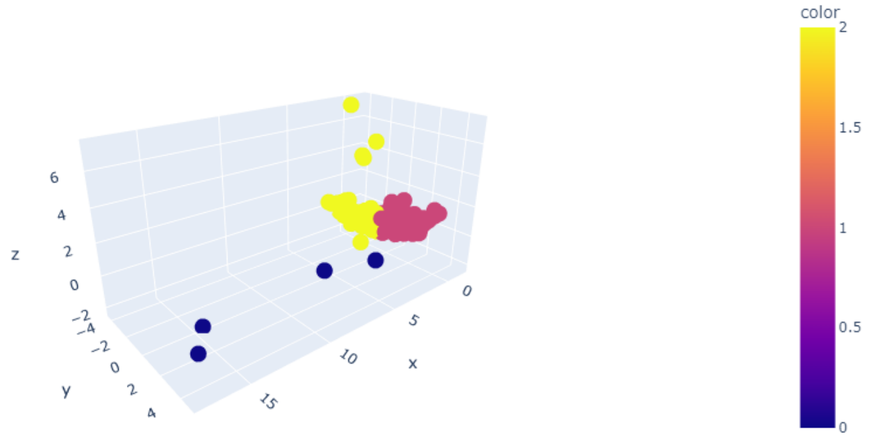
Outils de l'ACP:

- variance expliquée: aide à déterminer le nombre de composantes principales et comprendre le modèle
- Réduction (diviser par écart-type)
- "coordonnées" de chaque pays dans n(=nombre de variables) dimensions

La variance expliquée



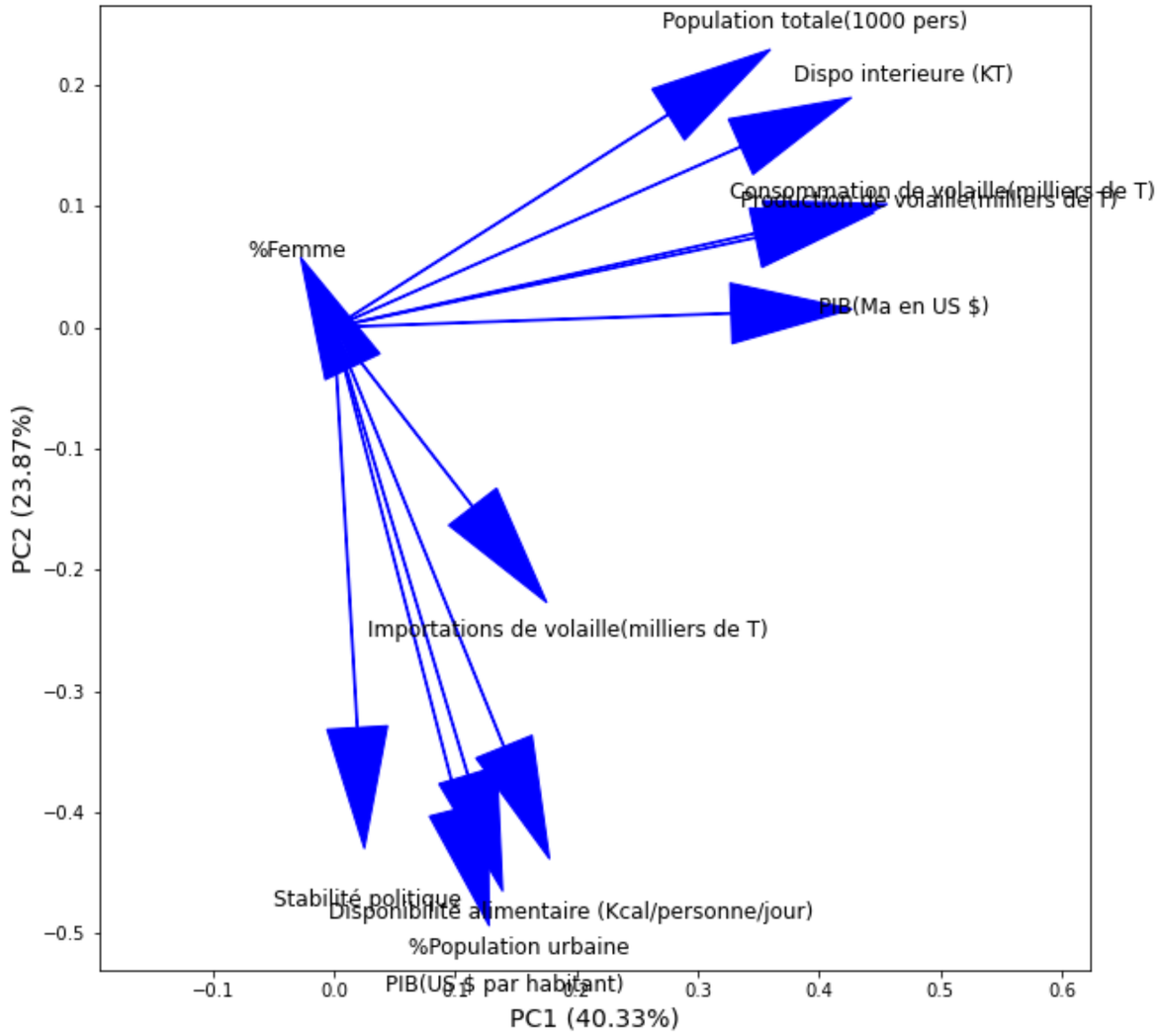
Visualisation des pays/cluster



La poule qui chante

Analyse en composante principale

Projection sur PC1 et PC2 (64% de variance)



PC1 (40% de la variance):

- Variables de volume (Population, PIB, Consommation...)
- Plus on se situe haut sur cet axe et plus le volume du pays est important

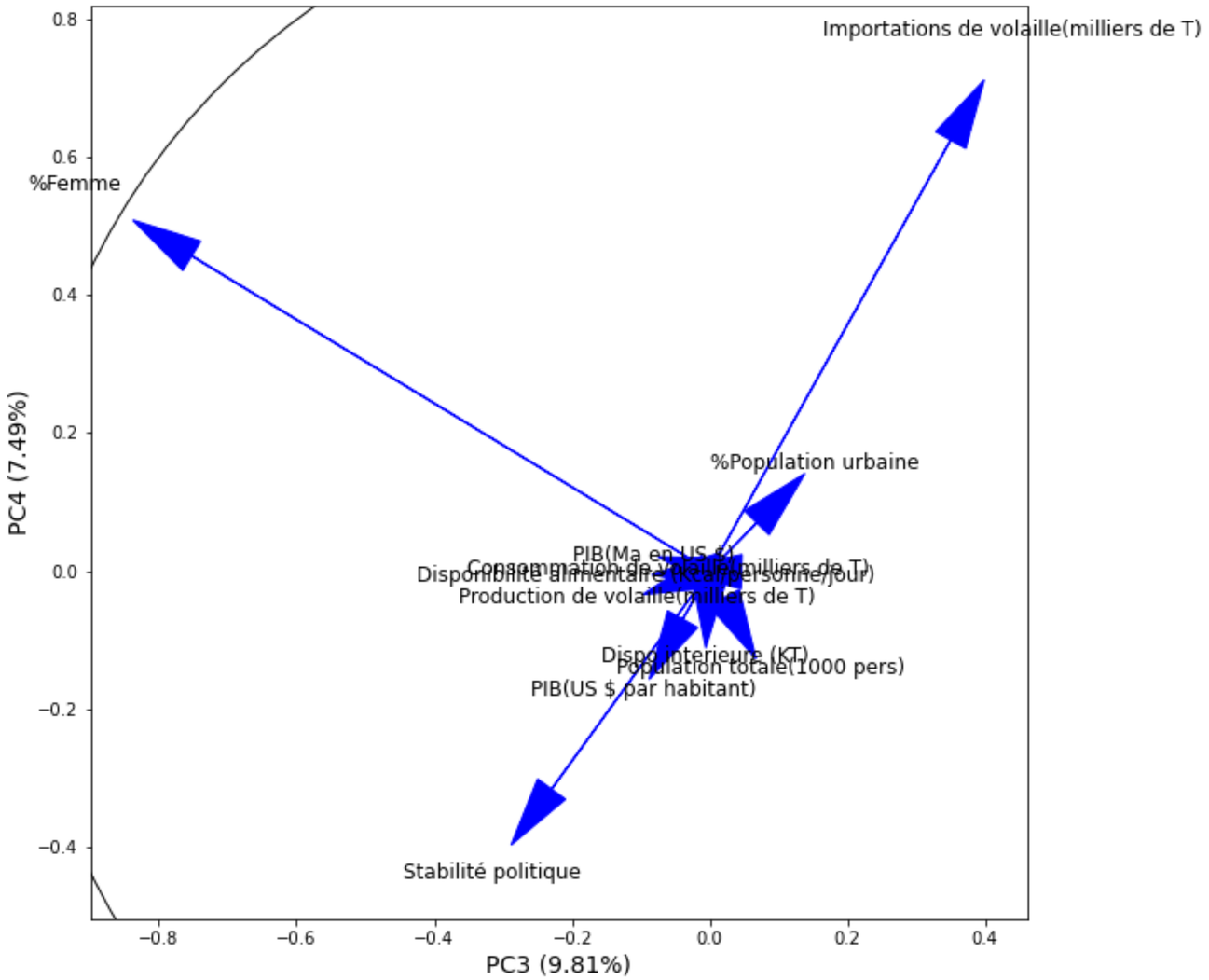
PC2 (24% de la variance):

- Richesse/habitant
- Variable est anti corrélée (plus est elle importante et plus la richesse/habitant du pays est faible)

La poule qui chante

Analyse en composante principale

Projection sur PC3 et PC4 (17% de variance)



PC3 (10% de la variance):

- % Femme dans la population
- Variable est anti corrélée (plus la proportion des femmes est importante et plus la valeur sur cet axe sera faible)

PC4 (7% de la variance):

- % Femme dans la population et importation de volaille en volume
- Plus la proportion des femmes et importation de volaille est importante et plus la valeur sur cet axe sera forte

La poule qui chante

III - Résultat et recommandations

La poule qui chante

A) Résultats

```
centroids_proj
```

	PC1	PC2	PC3	PC4
cluster_0	3.066465	0.990554	0.607673	-0.416582
cluster_1	-2.421155	1.978444	-1.289637	0.316029
cluster_2	-0.645310	-2.968998	0.681964	0.100554

	Caractéristiques	**Pays**
Cluster_0	PC1 est de loin le plus élevé. Il s'agit des pays avec le plus grand volume globale en valeur sur l'ensemble de nos variables. PC2: richesse par habitant est moyenne. PC3: la proportion des hommes est plus forte dans ces pays.	4 pays : Brésil, Chine, continentale, États-Unis d'Amérique, Inde
Cluster_1	PC1 est très faible. Il s'agit des pays avec peu de volume globale. PC2: la richesse/habitant est faible. PC3: La proportion des femmes est plus forte dans ces pays.	118 pays: Afghanistan, Iraq, Jamaïque, Ukraine, Vanuatu, Venezuela...
Cluster_2	PC1 est moyen. Il s'git des pays important mais d'une taille plus modeste. Les volumes (produits/consommés) sont moindre que ceux du cluster_0. PC2 est très faible, donc la richesse/habitant dans ces pays est la plus forte. PC3: La proportion des hommes est plus forte dans ces pays	44 pays: Afrique du Sud, Allemagne, France, Pologne...

La poule qui chante

B) Recommandations

	Volume	Richesse/habitant	Type de produit
Cluster_0	✓		• Qualité moyenne • Prix moyen
Cluster_1	✗	✗	• Qualité basse • Prix faible
Cluster_2		✓	• Haute qualité • Prix élevé

La poule qui chante

C) Limites du modèle

Importation de volaille

Variable très importante pour notre entreprise.

Analyse PESTEL

Besoin d'analyse de l'environnement de chaque pays.

	Zone	Importations de volaille(milliers de T)
23	Brésil	3
33	Chine, continentale	452
50	États-Unis d'Amérique	123
70	Inde	0

Merci de votre attention!

Présenté par: ARTEMOV Pavel