Mission data international

Étude de marché via classification des pays

La poule qui chante

Sommaire

1. Contexte du projet
2. Données utilisées
3. Nettoyage des données et analyse exploratoire
4. Classification des pays
5. Recommandations

La poule qui chante

1. Contexte du projet

Objectif de l'entreprise : Étendre ses activités à l'international.

Mission : Nettoyage des données et analyse approfondie des groupements de pays pour l'exportation de viande de volaille.

Objectif du projet data : Etude de marché pour cibler un ou des groupes de pays.

Méthode : Analyse des données mondiales via notebook Python et classification de pays.

Résultat attendu : Sélection des pays à fort potentiel et premières recommandations pour le développement international de l'entreprise.

La poule qui chante

2. Données utilisées

Nous avons utilisé plusieurs données afin de classifier les pays et permettre de définir leur potentiel :

Leur disponibilité alimentaire en viande de volaille

Leur quantité d'exportation et d'importation en viande de volaille

Leur population

Leur stabilité politique

Leur PIB

Leur émission de CO2

Sources :

Organisation des Nations Unies pour l'alimentation et l'agriculture

LA BANQUE MONDIALE
BIRD • IDA | GROUPE DE LA BANQUE MONDIALE

La poule qui chante

3. Nettoyage des données et analyse exploratoire

Les valeurs manquantes

Nous avons pu récolter des données pour 172 pays identifiables comme acteurs sur le marché mondiale de volailles.
Toutefois, les jeux de données présentaient des valeurs manquantes (3.4% de données indisponibles) :

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 172 entries, 0 to 171
Data columns (total 10 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Zone                      172 non-null    object
 1   Disponibilité intérieure  170 non-null    float64
 2   Exportations - Quantité   135 non-null    float64
 3   Importations - Quantité   170 non-null    float64
 4   Production                168 non-null    float64
 5   Variation de stock        169 non-null    float64
 6   Population                172 non-null    float64
 7   Stabilité politique       169 non-null    float64
 8   PIB                       168 non-null    float64
 9   Emission de CO2           168 non-null    float64
dtypes: float64(9), object(1)
memory usage: 13.6+ KB
```

170
135
170
168
169

169
168
168

pays pour lesquels l'émission de CO2 est disponible

172 pays étudiés

Solutions apportées pour suppléer à ces données manquantes :

1) Calculer les **exportations** manquantes (en fonction des importations, de la production, de la variation de stock et de la disponibilité intérieure).

2) **Estimer** les autres valeurs manquantes en utilisant les données des pays qui ont le même profil.

3) Calculer le **taux d'autosuffisance** afin de voir les pays qui serraient autosuffisants sur la viande de volaille.

4) Supprimer les **données superflues** pour la suite (Production et Variation de stock).

La poule qui chante

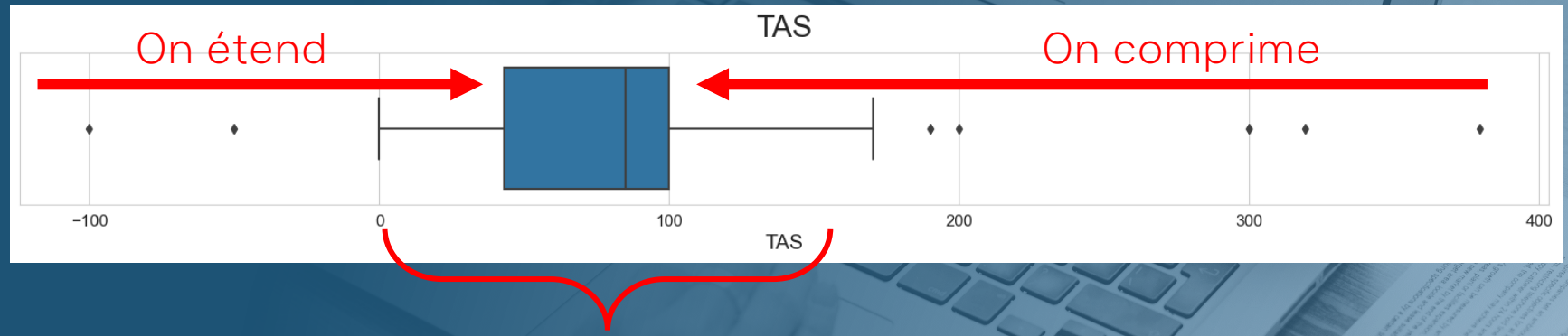
3. Nettoyage des données et analyse exploratoire

Les valeurs extrêmes

Afin d'éviter toute perte d'information, il a été choisi de garder l'ensemble des 172 pays – y compris ceux qui présentaient des valeurs extrêmes.

Nous avons donc choisi de modifier les valeurs de telle sorte que les valeurs trop importantes soient comprimées et que les valeurs trop faibles soient étendues.

Prenons l'exemple de la répartition des valeurs pour le taux d'autosuffisance :



La plage de distribution des valeurs va donc se retrouver réduite, cela va nous permettre de rendre notre modèle plus robuste face à des valeurs extrêmes.

La qualité des représentations graphiques sera également améliorée.

La poule qui chante

3. Nettoyage des données et analyse exploratoire

Standardisation des données

Afin d'obtenir des données comparables, il est nécessaire de rendre les variables indépendantes de leur unité (Kg, $, etc.) ou de leur échelle d'origine (milliers, millions, etc.) :

Sinon, certaines variables prendront plus d'importance que d'autres. Alors que chacune doit avoir la même importance pour l'analyse.

Pays	Disponibilité alimentaire	Population	PIB
USA	Kg	Nombre d'habitants	Millions US$
Chine	Kg	Nombre d'habitants	Millions US$

Standardisation

Pays	Disponibilité alimentaire	Population	PIB
USA	Moyenne = 0	Moyenne = 0	Moyenne = 0
Chine	Moyenne = 0	Moyenne = 0	Moyenne = 0

La poule qui chante

3. Nettoyage des données et analyse exploratoire

Réduction des variables et interprétation

Nous avons ensuite réduit le nombre de variables : ici, avec **4 variables synthétiques,** on peut capter **86% de l'information.**

Après avoir étudié ces 4 variable synthétiques, on peut les interpréter de la manière suivante :

Variables synthétiques	Composante principale 1	Composante principale 2	Composante principale 3	Composante principale 4
Interprétation	Pays peuplés consommateurs de viande de volaille	Pays en développement et politiquement instables	Pays exportateur de viande de volaille sur le marché mondial	Pays qui polluent le plus

Néanmoins, le nombre de variables de base reste restreint.

Considérant l'objectif de permettre une décision éclairée pour le développement des activités de l'entreprise, la prudence et la volonté de limiter les biais d'analyse nous amènent à faire le choix de poursuivre l'étude avec les 8 variables de base.

La poule qui chante

4. Classification des pays

Quelles méthodes allons nous utiliser ?

Méthode n°1 : La classification ascendante hiérarchique

1. Réalisation d'un dendrogramme pour déterminer le nombre de groupes de pays

2. Etude des moyennes des groupes en fonction des variables de bases (interprétation graphique)

Méthode n°2 : La méthode du K-Means

1. Etude graphique des groupes sur le repère des variables synthétiques

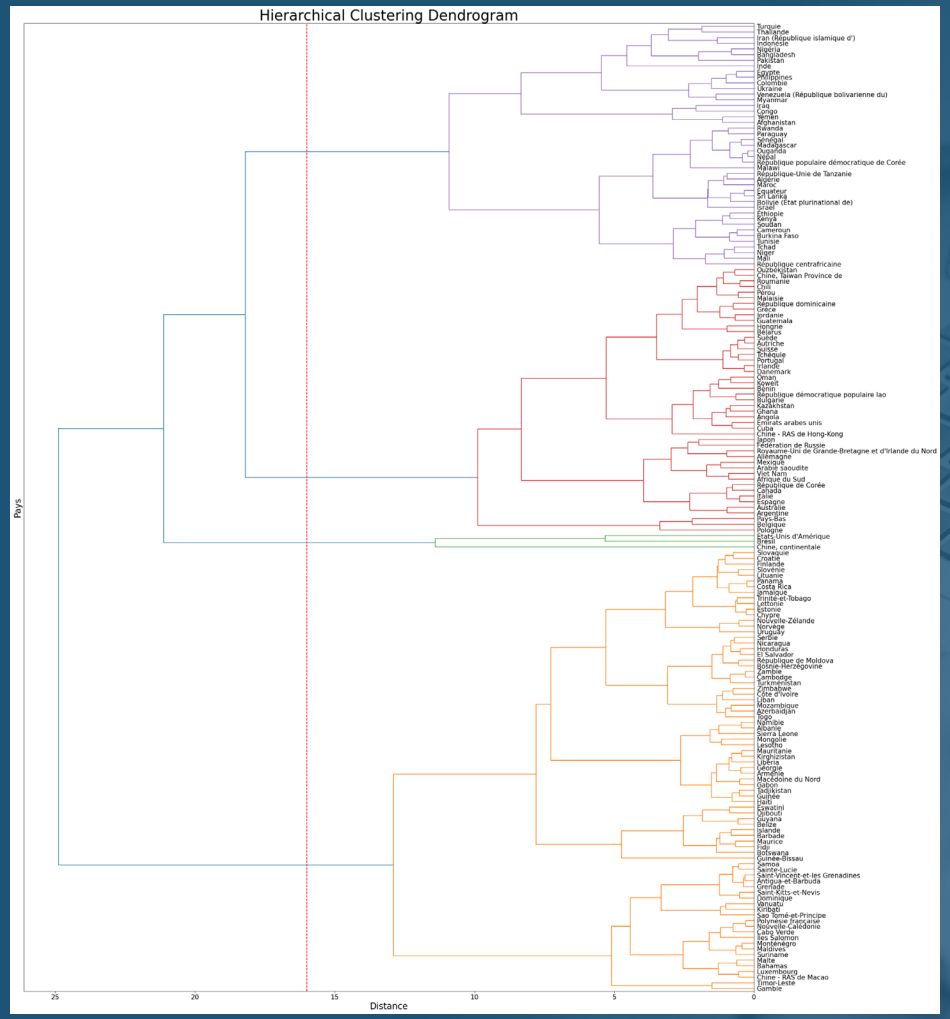
2. Etude des moyennes des groupes en fonction des variables de bases (interprétation graphique)

La poule qui chante

4. Classification des pays

Via la méthode n°1 : classification ascendante hiérarchique

└ *Dendrogramme*



Le **dendrogramme** nous permet de vérifier les pays qui présentent le plus de similitudes.

Il permet également de classifier les pays de sorte que les pays d'un même groupe soient suffisamment proches, et que chaque groupe soient suffisamment éloignés.

Il est alors possible de « couper » le dendrogramme (*cf. ligne rouge sur le graphique*) afin d'obtenir des groupes distincts.

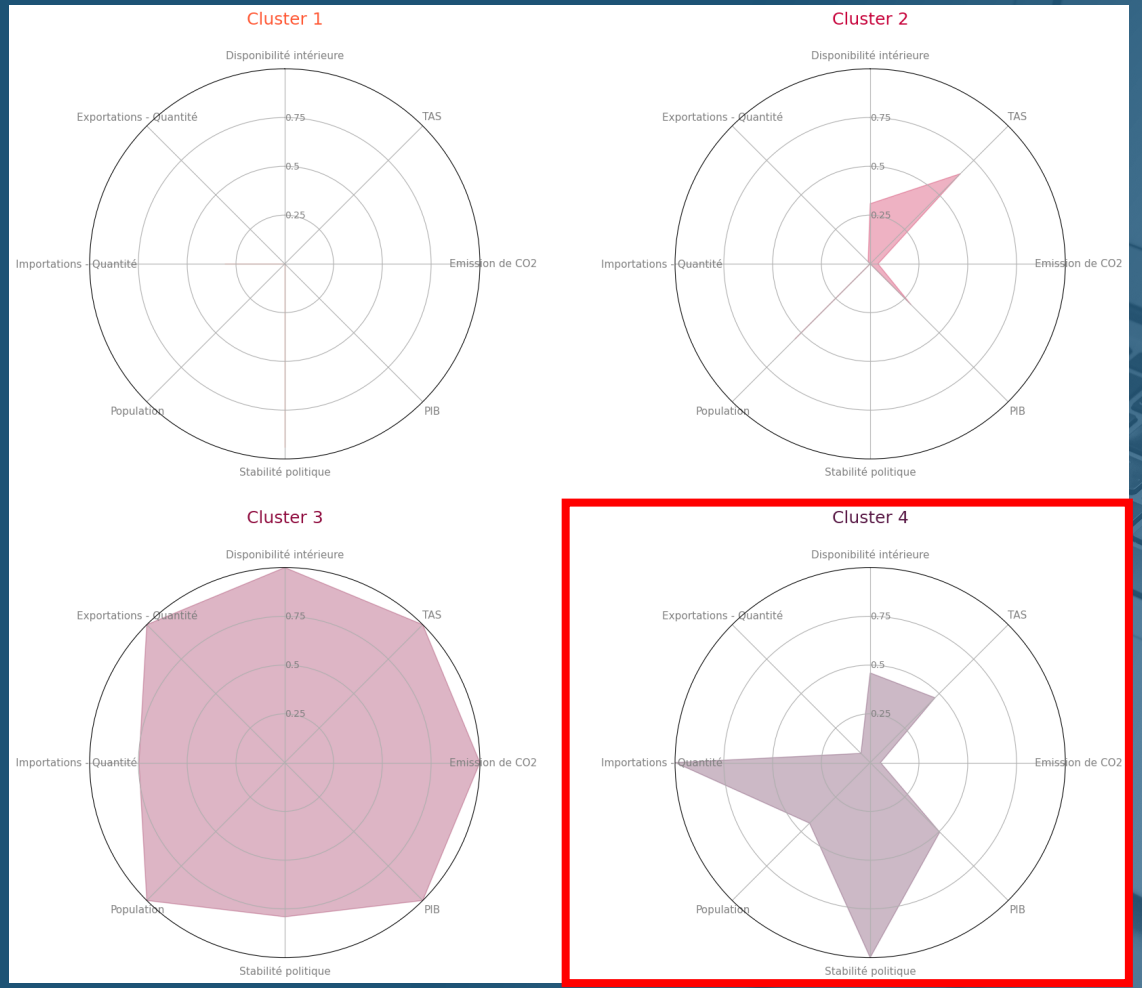
Nous étudierons donc 4 groupes de pays pour une meilleure homogénéité des groupes.

La poule qui chante

4. Classification des pays

Via la méthode n°1 : classification ascendante hiérarchique

└ Etude des 4 clusters avec graphiques en radar



Nous avons projeté les moyennes de ces groupes afin de pouvoir les interpréter en fonction des variables du jeux de données.

Nous retiendrons le **cluster 4** pour lequel les importations et la stabilité politique semblent les plus importants. Notons également la faible émission de CO_2 de ces pays.

Listes des pays candidats :

```
Cluster 4
Nombre de pays dans le cluster : 47
['Afrique du Sud', 'Allemagne', 'Angola', 'Arabie saoudite', 'Argentine', 'Australie', 'Autriche', 'Belgique', 'Bulgarie', 'Bélarus', 'Bénin', 'Canada', 'Chili', 'Chine - RAS de Hong-Kong', 'Chine, Taiwan Province de', 'Cuba', 'Danemark', 'Espagne', 'Fédération de Russie', 'Ghana', 'Grèce', 'Guatemala', 'Hongrie', 'Irlande', 'Italie', 'Japon', 'Jordanie', 'Kazakhstan', 'Koweït', 'Malaisie', 'Mexique', 'Oman', 'Ouzbékistan', 'Pays-Bas', 'Pologne', 'Portugal', 'Pérou', 'Roumanie', "Royaume-Uni de Grande-Bretagne et d'Irlande du Nord", 'République de Corée', 'République dominicaine', 'République démocratique populaire lao', 'Suisse', 'Suède', 'Tchéquie', 'Viet Nam', 'Émirats arabes unis']
```

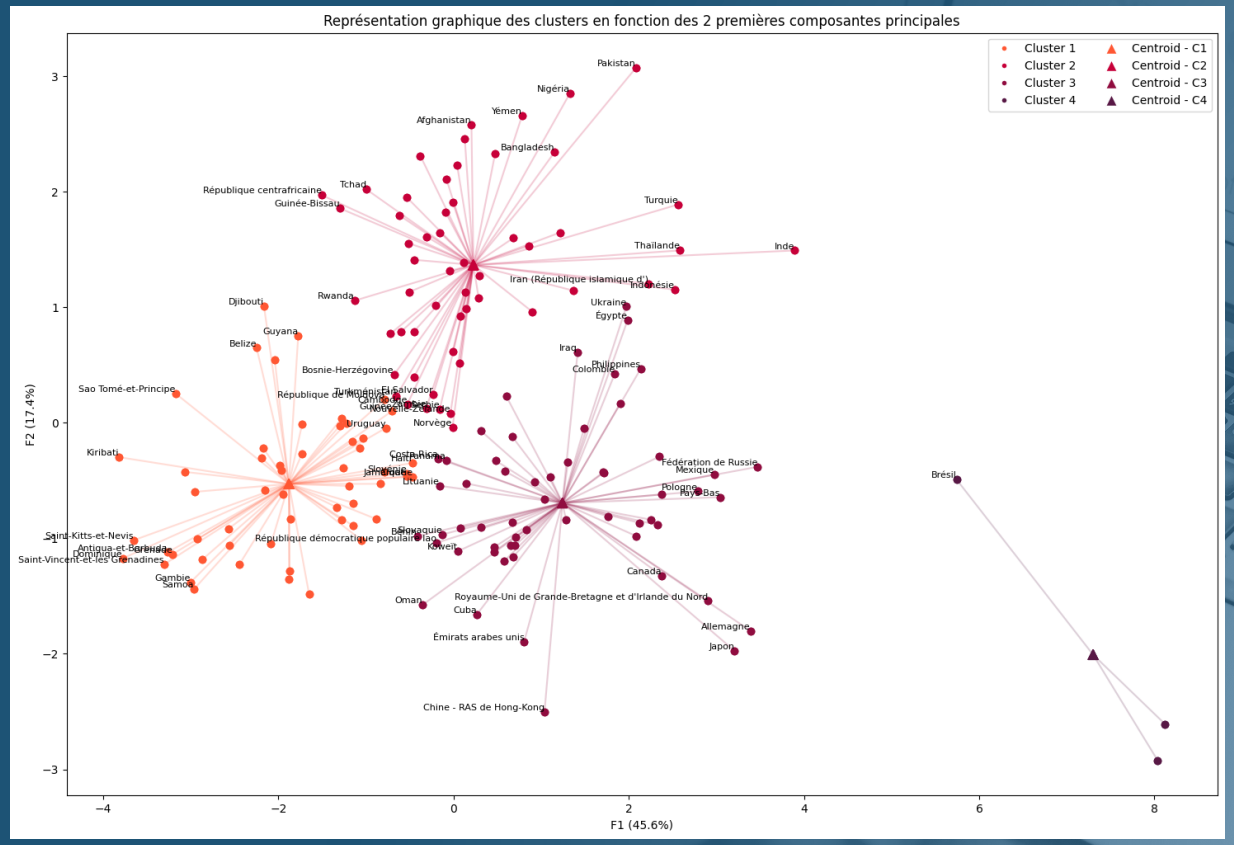
La poule qui chante

4. Classification des pays

Via la méthode n°2 : K-Means

L *Visualisation graphique des clusters*

Pour la méthode du K-Means, le nombre de groupe optimal est de 4.



Nous pouvons voir les différents pays de chaque groupe, chacun rattachés à leur centre (moyenne du groupe ou centroïde).

Le repère que nous avons choisi est le repère des composantes principales 1 et 2 que nous avons vu précédemment avec la réduction des variables.

Nous allons étudier les points centraux de ces 4 groupes afin de comprendre les tendances de chacun par rapport aux différentes variables.

La poule qui chante

4. Classification des pays

Via la méthode n°2 : K-Means

└ Etude des 4 clusters avec graphiques en radar

Ici, nous reprenons le même principe qu'avec la méthode précédente. On étudie les moyennes des différents groupes et on regarde les tendances de chacun des groupes par rapport aux variables de notre jeu de données.

Nous retiendrons le **cluster 3** pour lequel les importations et la stabilité politique semblent les plus importants. Notons également la faible émission de CO_2 de ces pays.

Liste des pays candidats :

```
Cluster 3
Nombre de pays dans le cluster : 57
['Afrique du Sud', 'Allemagne', 'Angola', 'Arabie saoudite', 'Argentine', 'Australie', 'Autriche', 'Belgique', 'Bulgarie', 'Bél
arus', 'Bénin', 'Canada', 'Chili', 'Chine - RAS de Hong-Kong', 'Chine, Taiwan Province de', 'Colombie', 'Costa Rica', 'Cuba',
'Danemark', 'Espagne', 'Finlande', 'Fédération de Russie', 'Ghana', 'Grèce', 'Guatemala', 'Hongrie', 'Iraq', 'Irlande', 'Itali
e', 'Japon', 'Jordanie', 'Kazakhstan', 'Koweït', 'Lituanie', 'Malaisie', 'Mexique', 'Oman', 'Ouzbékistan', 'Panama', 'Pays-Ba
s', 'Philippines', 'Pologne', 'Portugal', 'Pérou', 'Roumanie', "Royaume-Uni de Grande-Bretagne et d'Irlande du Nord", 'Républiq
ue de Corée', 'République dominicaine', 'République démocratique populaire lao', 'Slovaquie', 'Suisse', 'Suède', 'Tchéquie', 'U
kraine', 'Viet Nam', 'Égypte', 'Émirats arabes unis']
```

La poule qui chante

4. Classification des pays

Analyse des groupes sélectionnés

```
Pourcentage de similitude : 82.46 %

Pays présents dans les deux listes : ['Portugal', 'Koweït', "Royaume-Uni de Grande-Bretagne et d'Irlande du Nord", 'Pérou', 'Ja
pon', 'Angola', 'Cuba', 'Canada', 'Suisse', 'Ghana', 'République dominicaine', 'République démocratique populaire lao', 'Espagn
e', 'Mexique', 'Kazakhstan', 'Roumanie', 'Ouzbékistan', 'Chine, Taiwan Province de', 'Oman', 'Jordanie', 'Bélarus', 'Tchéquie',
'Autriche', 'Afrique du Sud', 'Bénin', 'Pologne', 'Malaisie', 'Australie', 'Arabie saoudite', 'Guatemala', 'Chili', 'Italie',
'Pays-Bas', 'Danemark', 'Hongrie', 'Belgique', 'Bulgarie', 'Irlande', 'Fédération de Russie', 'Viet Nam', 'Chine - RAS de Hong-
Kong', 'Grèce', 'Suède', 'Émirats arabes unis', 'République de Corée', 'Argentine', 'Allemagne']

Pays seulement présents dans pays_cluster_CAH : []

Pays seulement présents dans pays_cluster_KMeans : ['Égypte', 'Panama', 'Colombie', 'Finlande', 'Ukraine', 'Lituanie', 'Philipp
ines', 'Costa Rica', 'Slovaquie', 'Iraq']
```

On peut voir ici qu'il y a un taux de similitude de **82,46%** entre les clusters sélectionnés par les deux méthodes.

La différence va être que le cluster du K-Means comporte plus de pays. Mais si on s'intéresse de plus prêt à la stabilité politique, on peut voir que les pays qui ont été rajoutés font baisser la stabilité politique du groupe.

Le choix du cluster de la méthode n°1 (par CAH) comportant moins de pays serait donc plus judicieuse.

4. Classification des pays

Comparaison et choix des groupes sélectionnés par les 2 méthodes

Groupe de pays de la méthode n°1

Nombre de pays : 47

Les +	Les -
Disponibilité intérieure	Exportations
Importations	Emission de CO2
Population	
Stabilité politique	
PIB	
Taux d'autosuffisance	

Groupe de pays de la méthode n°2

Nombre de pays : 57

Les +	Les -
Exportations	Disponibilité intérieure
Émission de CO2	Importations
	Population
	Stabilité politique
	PIB
	Taux d'autosuffisance

Ce qui nous importe : avoir le meilleur score sur le PIB, la stabilité politique et l'importation en quantité de viande de volaille.

Le groupe de la méthode n°1 présente de meilleur attributs sur ces points.

La poule qui chante

5. Recommandations

Analyse PESTEL

Politique :
Les données de **stabilité politique** aideront à identifier les pays où le potentiel de croissance est favorisé

Légal :
La **législation locale** et les **normes en vigueur** doivent être étudiées afin d'adapter la production, l'exportation et l'importation de la vente de viande de volaille

Economique :
Un **fort PIB** démontre une activité économique élevée, propice au développement d'entreprises

Ecologique :
Les données sur les **émissions de CO_2** peuvent être un départ pour sélectionner les pays importateurs. La question va notamment se poser sur le **choix du transport**

Social :
Les données liées à la **population** et la **disponibilité alimentaire** permettront d'envisager le comportement des consommateurs

Technologique :
L'entreprise devra peut-être **s'adapter sur le plan technologique.** Il n'est pas à négliger surtout si la demande devient importante

La poule qui chante

5. Recommandations

Pour l'étude de marché à venir

La liste des **47 pays** avec un fort potentiel reste subjective.

Le travail réalisé a permis de faire une **pré-sélection grâce à des critère précis**, mais il n'empêche pas de réaliser d'autres études par la suite.

Il est également important de bien comprendre que le groupement qui a été fait méritera certainement d'être subdivisé en d 'autres groupements suivant les objectifs de marché.

En effet, il conviendra d'étudier d'autre critères comme par exemple la **proximité géographique** afin de limiter les émissions de CO2 ou bien même les taxes à l'importation.

Analyse de données

Etude de marché

La poule qui chante