1_preparation_nettoyage___analyse_exploratoire_072025

August 25, 2025

ANALYSE EXPLORATOIRE du DATASET "LAPOULE QUI CHANTE" POUR ETUDE DE MARCHE IMPLEMENTATION MONDIALE

Import de Librairies

[131]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
pd.options.display.max_columns=999
!pip install geopy
```

```
Requirement already satisfied: geopy in c:\users\famille\anaconda3\lib\site-
packages (2.4.1)
Requirement already satisfied: geographiclib<3,>=1.52 in
c:\users\famille\anaconda3\lib\site-packages (from geopy) (2.0)
```

Import des fichiers csv Lapoule

[144]:
```
df_population = pd.read_csv('Population_2000_2018.csv', sep=',')
df_stabilite_pol_PIB = pd.read_csv('Stabilite_pol_pib_2017_FAO.csv', sep=',')
df_dispo_alimentaire = pd.read_csv('DisponibiliteAlimentaire_2017.csv', sep=',')
```

[145]:
```
df_population.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4411 entries, 0 to 4410
Data columns (total 15 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Code Domaine            4411 non-null   object
 1   Domaine                 4411 non-null   object
 2   Code zone               4411 non-null   int64
 3   Zone                    4411 non-null   object
 4   Code Élément            4411 non-null   int64
 5   Élément                 4411 non-null   object
 6   Code Produit            4411 non-null   int64
 7   Produit                 4411 non-null   object
 8   Code année              4411 non-null   int64
 9   Année                   4411 non-null   int64
 10  Unité                   4411 non-null   object
```

```
 11  Valeur                  4411 non-null   float64
 12  Symbole                 4411 non-null   object
 13  Description du Symbole  4411 non-null   object
 14  Note                    258 non-null    object
dtypes: float64(1), int64(5), object(9)
memory usage: 517.0+ KB
```

[146]:
```
df_population.describe()
```

[146]:
```
          Code zone  Code Élément  Code Produit   Code année        Année  \
count  4411.000000        4411.0        4411.0  4411.000000  4411.000000
mean    132.202902         511.0        3010.0  2009.068238  2009.068238
std      75.854840           0.0           0.0     5.481539     5.481539
min       1.000000         511.0        3010.0  2000.000000  2000.000000
25%      68.000000         511.0        3010.0  2004.000000  2004.000000
50%     132.000000         511.0        3010.0  2009.000000  2009.000000
75%     195.000000         511.0        3010.0  2014.000000  2014.000000
max     299.000000         511.0        3010.0  2018.000000  2018.000000

              Valeur
count  4.411000e+03
mean   2.963004e+04
std    1.238029e+05
min    7.850000e-01
25%    3.921890e+02
50%    4.764741e+03
75%    1.821548e+04
max    1.427648e+06
```

[147]:
```
display(df_population.head())
```

```
  Code Domaine                      Domaine  Code zone         Zone  \
0           OA  Séries temporelles annuelles          2  Afghanistan
1           OA  Séries temporelles annuelles          2  Afghanistan
2           OA  Séries temporelles annuelles          2  Afghanistan
3           OA  Séries temporelles annuelles          2  Afghanistan
4           OA  Séries temporelles annuelles          2  Afghanistan

   Code Élément           Élément  Code Produit                Produit  \
0           511  Population totale          3010  Population-Estimations
1           511  Population totale          3010  Population-Estimations
2           511  Population totale          3010  Population-Estimations
3           511  Population totale          3010  Population-Estimations
4           511  Population totale          3010  Population-Estimations

   Code année  Année           Unité     Valeur Symbole  \
0        2000   2000  1000 personnes  20779.953       X
1        2001   2001  1000 personnes  21606.988       X
```

```
2        2002   2002  1000 personnes  22600.770       X
3        2003   2003  1000 personnes  23680.871       X
4        2004   2004  1000 personnes  24726.684       X

          Description du Symbole Note
0  Sources internationales sûres  NaN
1  Sources internationales sûres  NaN
2  Sources internationales sûres  NaN
3  Sources internationales sûres  NaN
4  Sources internationales sûres  NaN
```

[148]:
```
# Vérification de valeurs nulles
df_population.isna().sum()
```

[148]:
```
Code Domaine                    0
Domaine                         0
Code zone                       0
Zone                            0
Code Élément                    0
Élément                         0
Code Produit                    0
Produit                         0
Code année                      0
Année                           0
Unité                           0
Valeur                          0
Symbole                         0
Description du Symbole          0
Note                         4153
dtype: int64
```

[149]:
```
# Nettoyage pour garder que l'année 2017 (pour être en accord avec les autres df)
df_population = df_population.loc[df_population['Année'] == 2017, :]
```

[150]:
```
# Modification pour ne garder que les colonnes utiles pour la suite
df_population = df_population[['Zone', 'Valeur']]
```

[151]:
```
# Renommage de la colonne "valeur" pour "population"
df_population = df_population.rename(columns={'Valeur' : 'Population'})
```

[152]:
```
# Mis à l'échelle de la population
df_population['Population'] = df_population['Population'] * 1000

# Changement du dtype float en integer
df_population['Population'] = df_population['Population'].astype('int64')
```

[153]:
```
# Aperçu des modifications
display(df_population.head())
df_population.info()
```

```
             Zone  Population
17     Afghanistan    36296113
36  Afrique du Sud    57009756
55         Albanie     2884169
74         Algérie    41389189
93       Allemagne    82658409
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 236 entries, 17 to 4409
Data columns (total 2 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Zone        236 non-null    object
 1   Population  236 non-null    int64
dtypes: int64(1), object(1)
memory usage: 5.5+ KB
```

[154]:
```
df_population.describe()
```

[154]:

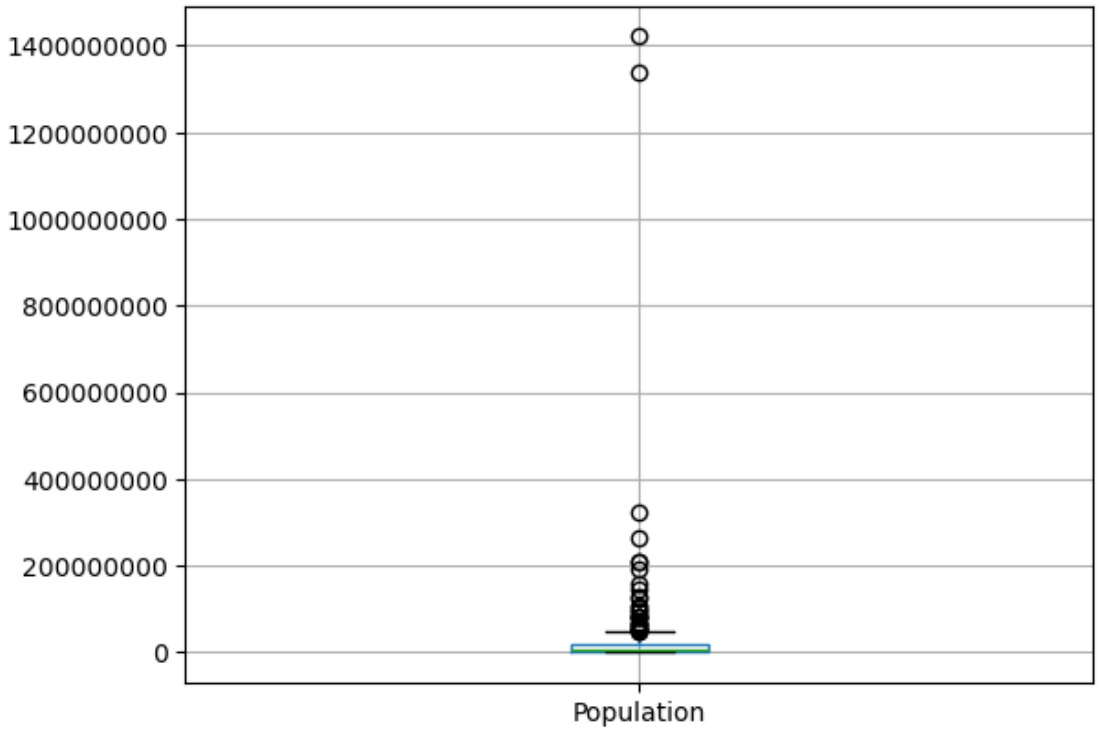
	Population
count	2.360000e+02
mean	3.198362e+07
std	1.318949e+08
min	7.930000e+02
25%	3.803032e+05
50%	5.203510e+06
75%	1.930842e+07
max	1.421022e+09

[155]:
```
# Vérification des valeurs maximum pour la population pour éviter les outliers
display(df_population.sort_values(by='Population', ascending=False))
```

	Zone	Population
823	Chine, continentale	1421021791
1952	Inde	1338676785
1230	États-Unis d'Amérique	325084756
1971	Indonésie	264650963
3029	Pakistan	207906209
…	…	…
2725	Montserrat	4984
1781	Îles Falkland (Malvinas)	3068
2896	Nioué	1609
4143	Tokélaou	1300
3672	Saint-Siège	793

```
[236 rows x 2 columns]
```

[156]:
```
# Vérification graphique pour confirmer nos potentiels outliers
df_population.boxplot(column='Population')
plt.ticklabel_format(style='plain', axis='y')
plt.show()
```



Le cas de la Chine continentale et l'Inde sont confirmés et ont des valeurs extrêmes et peuvent être considérés comme outliers, cela sera traité par la suite.

[158]:
```
# Vérification
#df_population.describe()
df_population.describe().applymap(lambda x: f'{x:,.0f}')
```

```
C:\Users\FAMILLE\AppData\Local\Temp\ipykernel_3548\2889624685.py:3:
FutureWarning: DataFrame.applymap has been deprecated. Use DataFrame.map
instead.
  df_population.describe().applymap(lambda x: f'{x:,.0f}')
```

[158]:

	Population
count	236
mean	31,983,619
std	131,894,885
min	793

```
25%          380,303
50%        5,203,510
75%       19,308,418
max    1,421,021,791
```

ANALYSE EXPLORATOIRE FICHIER FAO Stabilité_Politique et PIB

[160]:
```
display(df_stabilite_pol_PIB.head())
```

```
  Code Domaine                              Domaine  Code zone (M49)  \
0           FS  Données de la sécurité alimentaire                4
1           FS  Données de la sécurité alimentaire                4
2           FS  Données de la sécurité alimentaire              710
3           FS  Données de la sécurité alimentaire              710
4           FS  Données de la sécurité alimentaire                8

             Zone  Code Élément Élément  Code Produit  \
0     Afghanistan          6126  Valeur         22013
1     Afghanistan          6125  Valeur         21032
2  Afrique du Sud          6126  Valeur         22013
3  Afrique du Sud          6125  Valeur         21032
4         Albanie          6126  Valeur         22013

                                             Produit  Code année  Année  \
0  PIB par habitant, ($ PPA internationaux consta…        2017   2017
1  Stabilité politique et absence de violence/ter…        2017   2017
2  PIB par habitant, ($ PPA internationaux consta…        2017   2017
3  Stabilité politique et absence de violence/ter…        2017   2017
4  PIB par habitant, ($ PPA internationaux consta…        2017   2017

      Unité    Valeur Symbole              Description du Symbole  Note
0  Int$/cap   2956.80       X  Chiffre de sources internationales   NaN
1       NaN     -2.79       X  Chiffre de sources internationales   NaN
2  Int$/cap  14823.60       X  Chiffre de sources internationales   NaN
3       NaN     -0.28       X  Chiffre de sources internationales   NaN
4  Int$/cap  14229.30       X  Chiffre de sources internationales   NaN
```

[161]:
```
# Vérification de valeurs nulles
df_stabilite_pol_PIB.isna().sum()
```

[161]:
```
Code Domaine          0
Domaine               0
Code zone (M49)       0
Zone                  0
Code Élément          0
Élément               0
Code Produit          0
Produit               0
```

```
Code année                  0
Année                       0
Unité                     196
Valeur                      0
Symbole                     0
Description du Symbole      0
Note                      386
dtype: int64
```

[162]:
```
# Modification pour ne garder que les colonnes utiles pour la suite
df_stabilite_pol_PIB = df_stabilite_pol_PIB[['Zone', 'Produit', 'Valeur']]
df_stabilite_pol_PIB.head()
```

[162]:

	Zone	Produit	Valeur
0	Afghanistan	PIB par habitant, ($ PPA internationaux consta…	2956.80
1	Afghanistan	Stabilité politique et absence de violence/ter…	-2.79
2	Afrique du Sud	PIB par habitant, ($ PPA internationaux consta…	14823.60
3	Afrique du Sud	Stabilité politique et absence de violence/ter…	-0.28
4	Albanie	PIB par habitant, ($ PPA internationaux consta…	14229.30

[163]:
```
df_stabilite_pol_PIB_pivot = pd.pivot(
    df_stabilite_pol_PIB,
    index=['Zone'],
    columns=['Produit'],
    values=['Valeur'])
df_stabilite_pol_PIB_pivot.head()
```

[163]:

	Valeur
Produit	PIB par habitant, ($ PPA internationaux constants de 2017)
Zone	
Afghanistan	2956.8
Afrique du Sud	14823.6
Albanie	14229.3
Algérie	13805.4
Allemagne	61563.6

	Valeur
Produit	Stabilité politique et absence de violence/terrorisme (indice)
Zone	
Afghanistan	-2.79
Afrique du Sud	-0.28
Albanie	0.37
Algérie	-0.92
Allemagne	0.57

[164]:
```
df_stabilite_pol_PIB_pivot.columns = df_stabilite_pol_PIB_pivot.columns.
↪droplevel()
```

[165]:
```
# Renommage des colonnes pour plus de lisibilité
df_stabilite_pol_PIB_pivot = df_stabilite_pol_PIB_pivot.rename(
    columns={'PIB par habitant, ($ PPA internationaux constants de 2017)': 'PIB␣
 ↪par habitant',
             'Stabilité politique et absence de violence/terrorisme (indice)':␣
 ↪'Stabilité politique'})

df_stabilite_pol_PIB_pivot.head()
```

[165]:

Produit	PIB par habitant	Stabilité politique
Zone		
Afghanistan	2956.8	-2.79
Afrique du Sud	14823.6	-0.28
Albanie	14229.3	0.37
Algérie	13805.4	-0.92
Allemagne	61563.6	0.57

[166]:
```
# Vérification des valeurs nulles, elles seront traitées par la suite
df_stabilite_pol_PIB_pivot[df_stabilite_pol_PIB_pivot.isna().any(axis=1)]
```

[166]:

Produit	PIB par habitant	Stabilité politique
Zone		
Chine	16773.5	NaN
Chine, continentale	16461.1	NaN
Cuba	NaN	0.64
République populaire démocratique de Corée	NaN	-0.50
Samoa américaines	NaN	1.18
Soudan du Sud	NaN	-2.45
Turkménistan	NaN	-0.13
Venezuela (République bolivarienne du)	NaN	-1.25
Yémen	NaN	-2.93
Érythrée	NaN	-0.73

[167]:
```
df_stabilite_pol_PIB_pivot.describe()
```

[167]:
```
Produit  PIB par habitant  Stabilité politique
count          190.000000          196.000000
mean         24374.291053           -0.068265
std          25524.973336            0.992503
min            908.700000           -2.930000
25%           5837.550000           -0.655000
50%          15282.600000            0.025000
75%          33736.925000            0.762500
max         135547.000000            1.930000
```

[168]:
```
# Vérification du pays avec le plus gros PIB par habitant
display(df_stabilite_pol_PIB_pivot.sort_values(by='PIB par habitant',␣
 ↪ascending=False))
```

```
Produit                                PIB par habitant  Stabilité politique
Zone
Chine - RAS de Macao                           135547.0                 1.41
Luxembourg                                     133845.0                 1.31
Singapour                                      113907.0                 1.60
Qatar                                          108024.0                 0.65
Bermudes                                        95310.9                 0.98
…                                                   …                    …
Soudan du Sud                                       NaN                -2.45
Turkménistan                                        NaN                -0.13
Venezuela (République bolivarienne du)              NaN                -1.25
Yémen                                               NaN                -2.93
Érythrée                                            NaN                -0.73

[198 rows x 2 columns]
```

ANALYSE EXPLORATOIRE Fichier Disponibilité Alimentaire

[170]:
```
display(df_dispo_alimentaire.head())
```

```
  Code Domaine                     Domaine  Code zone         Zone  \
0          FBS  Nouveaux Bilans Alimentaire          2  Afghanistan
1          FBS  Nouveaux Bilans Alimentaire          2  Afghanistan
2          FBS  Nouveaux Bilans Alimentaire          2  Afghanistan
3          FBS  Nouveaux Bilans Alimentaire          2  Afghanistan
4          FBS  Nouveaux Bilans Alimentaire          2  Afghanistan

   Code Élément                  Élément  Code Produit          Produit  \
0          5511               Production          2511  Blé et produits
1          5611   Importations - Quantité          2511  Blé et produits
2          5072        Variation de stock          2511  Blé et produits
3          5911   Exportations - Quantité          2511  Blé et produits
4          5301  Disponibilité intérieure          2511  Blé et produits

   Code année  Année              Unité  Valeur Symbole  \
```

```
0        2017   2017  Milliers de tonnes  4281.0       S
1        2017   2017  Milliers de tonnes  2302.0       S
2        2017   2017  Milliers de tonnes  -119.0       S
3        2017   2017  Milliers de tonnes     0.0       S
4        2017   2017  Milliers de tonnes  6701.0       S

  Description du Symbole
0  Données standardisées
1  Données standardisées
2  Données standardisées
3  Données standardisées
4  Données standardisées
```

[171]:
```
# Vérification de valeurs nulles
df_dispo_alimentaire.isna().sum()
```

[171]:
```
Code Domaine              0
Domaine                   0
Code zone                 0
Zone                      0
Code Élément              0
Élément                   0
Code Produit              0
Produit                   0
Code année                0
Année                     0
Unité                     0
Valeur                    0
Symbole                   0
Description du Symbole    0
dtype: int64
```

[172]:
```
# Modification pour ne garder que les colonnes utiles pour la suite
df_dispo_alimentaire = df_dispo_alimentaire[['Zone', 'Élément', 'Produit',␣
 ↪'Valeur']]
```

[173]:
```
# Aperçu des modifications
display(df_dispo_alimentaire.head(10))
```

```
         Zone                   Élément          Produit  Valeur
0  Afghanistan                Production  Blé et produits  4281.0
1  Afghanistan    Importations - Quantité  Blé et produits  2302.0
2  Afghanistan         Variation de stock  Blé et produits  -119.0
3  Afghanistan    Exportations - Quantité  Blé et produits     0.0
4  Afghanistan  Disponibilité intérieure  Blé et produits  6701.0
5  Afghanistan      Aliments pour animaux  Blé et produits    76.0
6  Afghanistan                  Semences  Blé et produits   344.0
7  Afghanistan                    Pertes  Blé et produits   642.0
```

```
8  Afghanistan                  Résidus  Blé et produits     0.0
9  Afghanistan                Nourriture  Blé et produits  5640.0
```

[174]:
```
# Aperçu des valeurs trouvées dans la colonne "Élément" pour ne garder que␣
 ↪celles utiles
df_dispo_alimentaire['Élément'].unique()
```

[174]:
```
array(['Production', 'Importations - Quantité', 'Variation de stock',
       'Exportations - Quantité', 'Disponibilité intérieure',
       'Aliments pour animaux', 'Semences', 'Pertes', 'Résidus',
       'Nourriture',
       'Disponibilité alimentaire en quantité (kg/personne/an)',
       'Disponibilité alimentaire (Kcal/personne/jour)',
       'Disponibilité de protéines en quantité (g/personne/jour)',
       'Disponibilité de matière grasse en quantité (g/personne/jour)',
       'Traitement', 'Autres utilisations (non alimentaire)',
       'Alimentation pour touristes'], dtype=object)
```

[175]:
```
# Création d'une liste pour ne garder que celles utiles
liste_elements = (
    'Production', 'Importations - Quantité', 'Variation de stock',
    'Exportations - Quantité', 'Disponibilité intérieure', 'Nourriture',
    'Disponibilité alimentaire en quantité (kg/personne/an)',
    'Disponibilité alimentaire (Kcal/personne/jour)',
    'Disponibilité de protéines en quantité (g/personne/jour)',)
```

[176]:
```
# Conservation des éléments cités dans la liste pour le df
df_dispo_alimentaire = df_dispo_alimentaire.loc[df_dispo_alimentaire['Élément'].
 ↪isin(liste_elements), :]

# Aperçu des modifications
display(df_dispo_alimentaire.head(10))
```

```
          Zone                                          Élément  \
0   Afghanistan                                       Production
1   Afghanistan                     Importations - Quantité
2   Afghanistan                          Variation de stock
3   Afghanistan                     Exportations - Quantité
4   Afghanistan                    Disponibilité intérieure
9   Afghanistan                                     Nourriture
10  Afghanistan  Disponibilité alimentaire en quantité (kg/pers…
11  Afghanistan     Disponibilité alimentaire (Kcal/personne/jour)
12  Afghanistan  Disponibilité de protéines en quantité (g/pers…
14  Afghanistan                                       Production

           Produit   Valeur
0   Blé et produits  4281.00
1   Blé et produits  2302.00
```

```
2   Blé et produits  -119.00
3   Blé et produits     0.00
4   Blé et produits  6701.00
9   Blé et produits  5640.00
10  Blé et produits   155.39
11  Blé et produits  1331.00
12  Blé et produits    35.88
14  Riz et produits   338.00
```

[177]:
```
# Aperçu des valeurs trouvées dans la colonne "Produit" pour ne garder que␣
 ↪celles utiles
df_dispo_alimentaire['Produit'].unique()
```

[177]:
```
array(['Blé et produits', 'Riz et produits', 'Orge et produits',
       'Maïs et produits', 'Seigle et produits', 'Avoine',
       'Millet et produits', 'Sorgho et produits', 'Céréales, Autres',
       'Pommes de Terre et produits', 'Ignames', 'Racines nda',
       'Sucre, canne', 'Sucre, betterave', 'Sucre Eq Brut',
       'Edulcorants Autres', 'Miel', 'Haricots', 'Pois',
       'Légumineuses Autres et produits', 'Noix et produits', 'Soja',
       'Arachides Decortiquees', 'Graines de tournesol',
       'Graines Colza/Moutarde', 'Graines de coton', 'Coco (Incl Coprah)',
       'Sésame', 'Olives', 'Plantes Oleiferes, Autre', 'Huile de Soja',
       "Huile d'Arachide", 'Huile de Tournesol',
       'Huile de Colza&Moutarde', 'Huile Graines de Coton',
       'Huile de Palmistes', 'Huile de Palme', 'Huile de Coco',
       'Huile de Sésame', "Huile d'Olive", 'Huile de Son de Riz',
       'Huile de Germe de Maïs', 'Huil Plantes Oleif Autr',
       'Tomates et produits', 'Oignons', 'Légumes, Autres',
       'Oranges, Mandarines', 'Citrons & Limes et produits',
       'Pamplemousse et produits', 'Agrumes, Autres', 'Bananes',
       'Pommes et produits', 'Ananas et produits', 'Dattes', 'Raisin',
       'Fruits, Autres', 'Café et produits', 'Feve de Cacao et produits',
       'Thé', 'Poivre', 'Piments', 'Girofles', 'Épices, Autres', 'Vin',
       'Bière', 'Boissons Fermentés', 'Boissons Alcooliques',
       'Alcool, non Comestible', 'Viande de Bovins',
       "Viande d'Ovins/Caprins", 'Viande de Suides',
       'Viande de Volailles', 'Viande, Autre', 'Abats Comestible',
       'Beurre, Ghee', 'Crème', 'Graisses Animales Crue', 'Oeufs',
       'Lait - Excl Beurre', 'Poissons Eau Douce',
       'Aliments pour enfants', 'Miscellanees', 'Manioc et produits',
       'Patates douces', 'Palmistes', 'Bananes plantains',
       'Huiles de Poissons', 'Huiles de Foie de Poisso', 'Perciform',
       'Poissons Pelagiques', 'Poissons Marins, Autres', 'Crustacés',
       'Cephalopodes', 'Mollusques, Autres', 'Animaux Aquatiques Autre',
       'Plantes Aquatiques', 'Sucre non centrifugé',
       'Viande de Anim Aquatiq'], dtype=object)
```

[178]:
```
# Ce qui nous intéresse, c'est "Viande de Volailles" pour cette étude de marché.
df_dispo_alimentaire = df_dispo_alimentaire.loc[df_dispo_alimentaire['Produit']␣
 ↪== 'Viande de Volailles', :]
```

[179]:
```
# Aperçu des modifications
display(df_dispo_alimentaire.head())
```

	Zone	Élément	Produit	Valeur
651	Afghanistan	Production	Viande de Volailles	28.0
652	Afghanistan	Importations - Quantité	Viande de Volailles	29.0
653	Afghanistan	Variation de stock	Viande de Volailles	0.0
654	Afghanistan	Disponibilité intérieure	Viande de Volailles	57.0
657	Afghanistan	Nourriture	Viande de Volailles	55.0

[180]:
```
df_dispo_alimentaire_pivot = pd.pivot(df_dispo_alimentaire,
                                      index=['Zone'],
                                      columns=['Élément'],
                                      values=['Valeur'])

display(df_dispo_alimentaire_pivot.head())
```

	Valeur	\
Élément	Disponibilité alimentaire (Kcal/personne/jour)	
Zone		
Afghanistan	5.0	
Afrique du Sud	143.0	
Albanie	85.0	
Algérie	22.0	
Allemagne	71.0	

	\
Élément	Disponibilité alimentaire en quantité (kg/personne/an)
Zone	
Afghanistan	1.53
Afrique du Sud	35.69
Albanie	16.36
Algérie	6.38
Allemagne	19.47

	\
Élément	Disponibilité de protéines en quantité (g/personne/jour)
Zone	
Afghanistan	0.54
Afrique du Sud	14.11
Albanie	6.26
Algérie	1.97
Allemagne	7.96

```
                                                                    \
Élément        Disponibilité intérieure Exportations - Quantité
Zone
Afghanistan                         57.0                     NaN
Afrique du Sud                    2118.0                    63.0
Albanie                             47.0                     0.0
Algérie                            277.0                     0.0
Allemagne                         1739.0                   646.0

                                                                  \
Élément        Importations - Quantité Nourriture Production
Zone
Afghanistan                       29.0       55.0       28.0
Afrique du Sud                   514.0     2035.0     1667.0
Albanie                           38.0       47.0       13.0
Algérie                            2.0      264.0      275.0
Allemagne                        842.0     1609.0     1514.0


Élément        Variation de stock
Zone
Afghanistan                   0.0
Afrique du Sud               -0.0
Albanie                       4.0
Algérie                       0.0
Allemagne                   -29.0
```

Il y a un “NaN” dans la colonne “Exportations - Quantité”, il faut regarder s’il n’y en a pas d’autres suite à ce pivot.

[182]:
```
df_dispo_alimentaire_pivot.columns = df_dispo_alimentaire_pivot.columns.
↪droplevel()
```

[183]:
```
df_dispo_alimentaire_pivot.isna().sum()
```

[183]:
```
Élément
Disponibilité alimentaire (Kcal/personne/jour)                0
Disponibilité alimentaire en quantité (kg/personne/an)        0
Disponibilité de protéines en quantité (g/personne/jour)      0
Disponibilité intérieure                                      2
Exportations - Quantité                                      37
Importations - Quantité                                       2
Nourriture                                                    2
Production                                                    4
Variation de stock                                            3
dtype: int64
```

Il y a 37 valeurs nulles compatibilisées pour la colonne “Exportations - Quantité”. Deux choix sont possibles :

Remplacer les valeurs nulles par 0
Appliquer une formule de calcul à l'aide d'informations déjà présentes dans ce df
Source du calcul

[185]:
```
# Tentative de formule à l'aide de ce calcul :
# Calcul 1 : (Production + Importations - Disponibilité alimentaire en quantité
 ↪+ Variation de Stock)
# Calcul 2 : Résultat du calcul 1 - Disponibilité intérieure

df_dispo_alimentaire_pivot['Exportations - Quantité'] =
 ↪df_dispo_alimentaire_pivot.apply(
    lambda x : ((x['Production'] + x['Importations - Quantité'] -
                 x['Disponibilité alimentaire en quantité (kg/personne/an)'] +
                 x['Variation de stock'])
                - x['Disponibilité intérieure'])
    if pd.isnull(x['Exportations - Quantité']) else x['Exportations -
 ↪Quantité'], axis=1)
```

[186]:
```
display(df_dispo_alimentaire_pivot.head())
df_dispo_alimentaire_pivot.isna().sum()
```

Élément	Disponibilité alimentaire (Kcal/personne/jour)
Zone	
Afghanistan	5.0
Afrique du Sud	143.0
Albanie	85.0
Algérie	22.0
Allemagne	71.0

Élément	Disponibilité alimentaire en quantité (kg/personne/an)
Zone	
Afghanistan	1.53
Afrique du Sud	35.69
Albanie	16.36
Algérie	6.38
Allemagne	19.47

Élément	Disponibilité de protéines en quantité (g/personne/jour)
Zone	
Afghanistan	0.54
Afrique du Sud	14.11
Albanie	6.26
Algérie	1.97
Allemagne	7.96

Élément	Disponibilité intérieure	Exportations - Quantité
Zone		
Afghanistan	57.0	-1.53

```
Afrique du Sud                    2118.0                    63.00
Albanie                             47.0                     0.00
Algérie                            277.0                     0.00
Allemagne                         1739.0                   646.00

Élément         Importations - Quantité  Nourriture  Production  \
Zone
Afghanistan                        29.0        55.0        28.0
Afrique du Sud                    514.0      2035.0      1667.0
Albanie                            38.0        47.0        13.0
Algérie                             2.0       264.0       275.0
Allemagne                         842.0      1609.0      1514.0

Élément         Variation de stock
Zone
Afghanistan                    0.0
Afrique du Sud                -0.0
Albanie                        4.0
Algérie                        0.0
Allemagne                    -29.0
```

[186]:
```
Élément
Disponibilité alimentaire (Kcal/personne/jour)             0
Disponibilité alimentaire en quantité (kg/personne/an)     0
Disponibilité de protéines en quantité (g/personne/jour)   0
Disponibilité intérieure                                   2
Exportations - Quantité                                    4
Importations - Quantité                                    2
Nourriture                                                 2
Production                                                 4
Variation de stock                                         3
dtype: int64
```

La formule a pu remplacer une grande partie des valeurs manquantes. Il reste encore des valeures nulles dans les exportations, c'est peut-être du au fait qu'il y a quelques valeures nulles dans les autres colonnes.

[188]:
```
# Vérification des valeurs nulles, elles seront traitées par la suite
df_dispo_alimentaire_pivot[df_dispo_alimentaire_pivot.isna().any(axis=1)]
```

[188]:
```
Élément                                 Disponibilité alimentaire
(Kcal/personne/jour)  \
Zone
Djibouti
9.0
Maldives
47.0
Ouzbékistan
```

```
6.0
Pérou
62.0
République démocratique populaire lao
41.0

Élément                                 Disponibilité alimentaire en quantité
(kg/personne/an)  \
Zone
Djibouti
2.68
Maldives
13.50
Ouzbékistan
1.96
Pérou
13.47
République démocratique populaire lao
10.91

Élément                                 Disponibilité de protéines en quantité
(g/personne/jour)  \
Zone
Djibouti
0.92
Maldives
4.70
Ouzbékistan
0.63
Pérou
6.71
République démocratique populaire lao
3.59

Élément                                 Disponibilité intérieure  \
Zone
Djibouti                                                     3.0
Maldives                                                    12.0
Ouzbékistan                                                  NaN
Pérou                                                     1523.0
République démocratique populaire lao                        NaN

Élément                                 Exportations - Quantité  \
Zone
Djibouti                                                     NaN
Maldives                                                     NaN
Ouzbékistan                                                  NaN
```

```
Pérou                                                    1.0
République démocratique populaire lao                    NaN

Élément                                 Importations - Quantité  Nourriture  \
Zone
Djibouti                                                    3.0         3.0
Maldives                                                   12.0         7.0
Ouzbékistan                                                 NaN         NaN
Pérou                                                      60.0       424.0
République démocratique populaire lao                       NaN         NaN

Élément                                 Production  Variation de stock
Zone
Djibouti                                       NaN                 0.0
Maldives                                       NaN                 0.0
Ouzbékistan                                    NaN                 NaN
Pérou                                       1465.0                 NaN
République démocratique populaire lao          NaN                 NaN
```

[189]:
```
df_dispo_alimentaire_pivot.describe()
```

[189]:
```
Élément  Disponibilité alimentaire (Kcal/personne/jour)  \
count                                        172.000000
mean                                          74.558140
std                                           60.571277
min                                            0.000000
25%                                           22.000000
50%                                           64.000000
75%                                          105.500000
max                                          243.000000

Élément  Disponibilité alimentaire en quantité (kg/personne/an)  \
count                                              172.000000
mean                                                20.213372
std                                                 15.860311
min                                                  0.130000
25%                                                  6.440000
50%                                                 18.090000
75%                                                 30.037500
max                                                 72.310000

Élément  Disponibilité de protéines en quantité (g/personne/jour)  \
count                                                  172.000000
mean                                                     7.121279
std                                                      5.603464
min                                                      0.040000
25%                                                      2.165000
```

```
50%                                                 6.505000
75%                                                10.365000
max                                                27.870000

Élément  Disponibilité intérieure  Exportations - Quantité  \
count                  170.000000                168.000000
mean                   687.594118                104.550595
std                   2187.184747                463.737451
min                      2.000000                -70.310000
25%                     30.500000                  0.000000
50%                    100.000000                  1.000000
75%                    368.250000                 16.602500
max                  18266.000000               4223.000000

Élément  Importations - Quantité    Nourriture    Production  \
count                 170.000000    170.000000    168.000000
mean                   89.529412    657.047059    725.190476
std                   186.669983   2136.545796   2501.457125
min                     0.000000      2.000000      0.000000
25%                     3.000000     28.500000     13.750000
50%                    16.000000     99.500000     70.000000
75%                    81.250000    365.250000    409.750000
max                  1069.000000  18100.000000  21914.000000

Élément  Variation de stock
count            169.000000
mean              13.668639
std               75.364884
min             -119.000000
25%                0.000000
50%                0.000000
75%                7.000000
max              859.000000
```

[190]:
```
# Vérification du pays avec une très grande quantité d'exportations
display(df_dispo_alimentaire_pivot.sort_values(by='Exportations - Quantité',␣
 ↪ascending=False))
```

```
Élément                                 Disponibilité alimentaire (Kcal/personne/
 ↪jour)  \
Zone
Brésil                                                                          ␣
 ↪204.0
États-Unis d'Amérique                                                           ␣
 ↪219.0
Pays-Bas                                                                        ␣
 ↪70.0
```

```
Pologne                                                                 ␣
 ↪107.0
Thaïlande                                                               ␣
 ↪47.0
…                                                                       …
Saint-Vincent-et-les Grenadines                                         ␣
 ↪243.0
Djibouti                                                                ␣
 ↪ 9.0
Maldives                                                                ␣
 ↪47.0
Ouzbékistan                                                             ␣
 ↪ 6.0
République démocratique populaire lao                                   ␣
 ↪41.0

Élément                                 Disponibilité alimentaire en quantité (kg/
 ↪personne/an)  \
Zone
Brésil                                                                  ␣
 ↪  48.03
États-Unis d'Amérique                                                   ␣
 ↪  55.68
Pays-Bas                                                                ␣
 ↪  20.33
Pologne                                                                 ␣
 ↪  30.30
Thaïlande                                                               ␣
 ↪  12.95
…                                                                       ␣
 ↪  …
Saint-Vincent-et-les Grenadines                                         ␣
 ↪  72.31
Djibouti                                                                ␣
 ↪   2.68
Maldives                                                                ␣
 ↪  13.50
Ouzbékistan                                                             ␣
 ↪   1.96
République démocratique populaire lao                                   ␣
 ↪  10.91

Élément                                 Disponibilité de protéines en quantité (g/
 ↪personne/jour)  \
Zone
```

```
Brésil                                                                  ␣
  ↪ 15.68
États-Unis d'Amérique                                                   ␣
  ↪ 19.93
Pays-Bas                                                                ␣
  ↪  8.48
Pologne                                                                 ␣
  ↪ 12.14
Thaïlande                                                               ␣
  ↪  4.35
…                                                                       ␣
  ↪ …
Saint-Vincent-et-les Grenadines                                         ␣
  ↪ 25.10
Djibouti                                                                ␣
  ↪  0.92
Maldives                                                                ␣
  ↪  4.70
Ouzbékistan                                                             ␣
  ↪  0.63
République démocratique populaire lao                                   ␣
  ↪  3.59
```

Élément	Disponibilité intérieure
Zone	
Brésil	9982.0
États-Unis d'Amérique	18266.0
Pays-Bas	372.0
Pologne	1156.0
Thaïlande	881.0
…	…
Saint-Vincent-et-les Grenadines	8.0
Djibouti	3.0
Maldives	12.0
Ouzbékistan	NaN
République démocratique populaire lao	NaN

Élément	Exportations - Quantité
Zone	
Brésil	4223.00
États-Unis d'Amérique	3692.00
Pays-Bas	1418.00
Pologne	1025.00
Thaïlande	796.00
…	…
Saint-Vincent-et-les Grenadines	-70.31
Djibouti	NaN

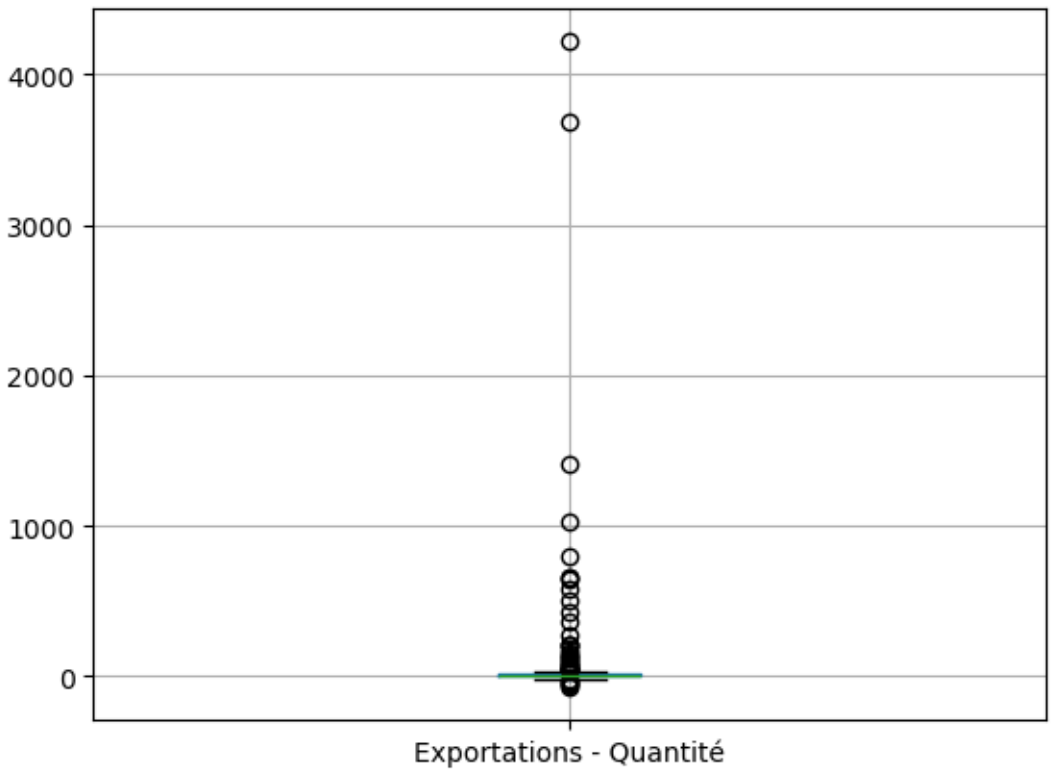
```
Maldives                                                NaN
Ouzbékistan                                             NaN
République démocratique populaire lao                   NaN

Élément                              Importations - Quantité  Nourriture  \
Zone
Brésil                                                   3.0      9982.0
États-Unis d'Amérique                                  123.0     18100.0
Pays-Bas                                               608.0       346.0
Pologne                                                 55.0      1150.0
Thaïlande                                                2.0       896.0
…                                                        …          …
Saint-Vincent-et-les Grenadines                          9.0         8.0
Djibouti                                                 3.0         3.0
Maldives                                                12.0         7.0
Ouzbékistan                                              NaN         NaN
République démocratique populaire lao                    NaN         NaN

Élément                              Production  Variation de stock
Zone
Brésil                                  14201.0                0.0
États-Unis d'Amérique                   21914.0               80.0
Pays-Bas                                 1100.0              -82.0
Pologne                                  2351.0              225.0
Thaïlande                                1676.0                1.0
…                                           …                  …
Saint-Vincent-et-les Grenadines             0.0                1.0
Djibouti                                    NaN                0.0
Maldives                                    NaN                0.0
Ouzbékistan                                 NaN                NaN
République démocratique populaire lao       NaN                NaN

[172 rows x 9 columns]
```

[191]:
```
# Vérification graphique pour confirmer nos potentiels outliers
df_dispo_alimentaire_pivot.boxplot(column='Exportations - Quantité')
plt.show()
```



Le Brésil et les Etats-Unis seront considérés à part également du à leurs valeurs extrêmes à différente(s) variable(s). Il y a donc actuellement, en pays qui seront traités à part :

```
Chine, continentale
Inde
Brésil
Les États-Unis
```

FUSION DES FICHIERS

MERGE “dispo_alimentaire et Stabilité Politique_PIB”

[195]:
```
# Première jointure entre les df "dispo alimentaire" et "PIB / stabilité␣
 ↪politique"
# Outer join pour garder tous les pays pour l'instant
df_dispo_alimentaire_stabilite_politique = pd.merge(df_dispo_alimentaire_pivot,␣
 ↪df_stabilite_pol_PIB_pivot,
                                                    on='Zone', how='outer')

display(df_dispo_alimentaire_stabilite_politique.head())
```

```
                Disponibilité alimentaire (Kcal/personne/jour)  \
Zone
Afghanistan                                              5.0
```

Zone	
Afrique du Sud	143.0
Albanie	85.0
Algérie	22.0
Allemagne	71.0

Zone	Disponibilité alimentaire en quantité (kg/personne/an)
Afghanistan	1.53
Afrique du Sud	35.69
Albanie	16.36
Algérie	6.38
Allemagne	19.47

Zone	Disponibilité de protéines en quantité (g/personne/jour)
Afghanistan	0.54
Afrique du Sud	14.11
Albanie	6.26
Algérie	1.97
Allemagne	7.96

Zone	Disponibilité intérieure	Exportations - Quantité
Afghanistan	57.0	-1.53
Afrique du Sud	2118.0	63.00
Albanie	47.0	0.00
Algérie	277.0	0.00
Allemagne	1739.0	646.00

Zone	Importations - Quantité	Nourriture	Production
Afghanistan	29.0	55.0	28.0
Afrique du Sud	514.0	2035.0	1667.0
Albanie	38.0	47.0	13.0
Algérie	2.0	264.0	275.0
Allemagne	842.0	1609.0	1514.0

Zone	Variation de stock	PIB par habitant	Stabilité politique
Afghanistan	0.0	2956.8	-2.79
Afrique du Sud	-0.0	14823.6	-0.28
Albanie	4.0	14229.3	0.37
Algérie	0.0	13805.4	-0.92
Allemagne	-29.0	61563.6	0.57

Deuxième jointure (résultat jointure précédente et population)

[197]:
```
# Deuxième jointure entre la population et la jointure ci-dessus
# Toujours en outer pour garder l'ensemble pour l'instant

df = pd.merge(df_dispo_alimentaire_stabilite_politique, df_population,␣
 ↪on='Zone', how='outer')
df.info()
display(df.head())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 239 entries, 0 to 238
Data columns (total 13 columns):
 #   Column                                                Non-Null Count
Dtype
---  ------                                                --------------
-----
 0   Zone                                                  239 non-null
object
 1   Disponibilité alimentaire (Kcal/personne/jour)        172 non-null
float64
 2   Disponibilité alimentaire en quantité (kg/personne/an)  172 non-null
float64
 3   Disponibilité de protéines en quantité (g/personne/jour)  172 non-null
float64
 4   Disponibilité intérieure                              170 non-null
float64
 5   Exportations - Quantité                               168 non-null
float64
 6   Importations - Quantité                               170 non-null
float64
 7   Nourriture                                            170 non-null
float64
 8   Production                                            168 non-null
float64
 9   Variation de stock                                    169 non-null
float64
 10  PIB par habitant                                      190 non-null
float64
 11  Stabilité politique                                   196 non-null
float64
 12  Population                                            236 non-null
float64
dtypes: float64(12), object(1)
memory usage: 24.4+ KB
```

	Zone	Disponibilité alimentaire (Kcal/personne/jour)	\
0	Afghanistan	5.0	
1	Afrique du Sud	143.0	
2	Albanie	85.0	

```
3         Algérie                                                22.0
4       Allemagne                                                71.0

   Disponibilité alimentaire en quantité (kg/personne/an)  \
0                                               1.53
1                                              35.69
2                                              16.36
3                                               6.38
4                                              19.47

   Disponibilité de protéines en quantité (g/personne/jour)  \
0                                               0.54
1                                              14.11
2                                               6.26
3                                               1.97
4                                               7.96

   Disponibilité intérieure  Exportations - Quantité  Importations - Quantité  \
0                      57.0                    -1.53                      29.0
1                    2118.0                    63.00                     514.0
2                      47.0                     0.00                      38.0
3                     277.0                     0.00                       2.0
4                    1739.0                   646.00                     842.0

   Nourriture  Production  Variation de stock  PIB par habitant  \
0        55.0        28.0                 0.0            2956.8
1      2035.0      1667.0                -0.0           14823.6
2        47.0        13.0                 4.0           14229.3
3       264.0       275.0                 0.0           13805.4
4      1609.0      1514.0               -29.0           61563.6

   Stabilité politique  Population
0                -2.79  36296113.0
1                -0.28  57009756.0
2                 0.37   2884169.0
3                -0.92  41389189.0
4                 0.57  82658409.0
```

[198]:
```
# Modification pour ne garder que les colonnes utiles pour la suite
df = df[['Zone',
        'Disponibilité alimentaire en quantité (kg/personne/an)',␣
 ↪'Disponibilité de protéines en quantité (g/personne/jour)',
        'Disponibilité intérieure',
        'Exportations - Quantité', 'Importations - Quantité',
        'PIB par habitant', 'Stabilité politique', 'Population']]

# Aperçu des modifications
```

```
display(df.head())
df.info()
```

```
            Zone  Disponibilité alimentaire en quantité (kg/personne/an)  \
0     Afghanistan                                               1.53
1  Afrique du Sud                                              35.69
2         Albanie                                              16.36
3         Algérie                                               6.38
4       Allemagne                                              19.47

   Disponibilité de protéines en quantité (g/personne/jour)  \
0                                               0.54
1                                              14.11
2                                               6.26
3                                               1.97
4                                               7.96

   Disponibilité intérieure  Exportations - Quantité  Importations - Quantité  \
0                      57.0                    -1.53                     29.0
1                    2118.0                    63.00                    514.0
2                      47.0                     0.00                     38.0
3                     277.0                     0.00                      2.0
4                    1739.0                   646.00                    842.0

   PIB par habitant  Stabilité politique  Population
0            2956.8                -2.79  36296113.0
1           14823.6                -0.28  57009756.0
2           14229.3                 0.37   2884169.0
3           13805.4                -0.92  41389189.0
4           61563.6                 0.57  82658409.0
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 239 entries, 0 to 238
Data columns (total 9 columns):
 #   Column                                                Non-Null Count
Dtype
---  ------                                                --------------
-----
 0   Zone                                                  239 non-null
object
 1   Disponibilité alimentaire en quantité (kg/personne/an)    172 non-null
float64
 2   Disponibilité de protéines en quantité (g/personne/jour)  172 non-null
float64
 3   Disponibilité intérieure                              170 non-null
float64
 4   Exportations - Quantité                               168 non-null
float64
```

```
 5   Importations - Quantité                                 170 non-null
float64
 6   PIB par habitant                                        190 non-null
float64
 7   Stabilité politique                                     196 non-null
float64
 8   Population                                              236 non-null
float64
dtypes: float64(8), object(1)
memory usage: 16.9+ KB
```

Traitement des valeurs nulles

[200]:
```
# Regroumement des pays avec des valeurs nulles dans un autre DF pour les retraiter par la suite
countries_with_nans = df[df.isnull().any(axis=1)]

# Aperçu
display(countries_with_nans.head())
countries_with_nans.shape
```

```
                          Zone  \
5                      Andorre
7                     Anguilla
9   Antilles néerlandaises (ex)
13                       Aruba
18                     Bahreïn

    Disponibilité alimentaire en quantité (kg/personne/an)  \
5                                                  NaN
7                                                  NaN
9                                                  NaN
13                                                 NaN
18                                                 NaN

    Disponibilité de protéines en quantité (g/personne/jour)  \
5                                                  NaN
7                                                  NaN
9                                                  NaN
13                                                 NaN
18                                                 NaN

    Disponibilité intérieure  Exportations - Quantité  \
5                        NaN                      NaN
7                        NaN                      NaN
9                        NaN                      NaN
13                       NaN                      NaN
18                       NaN                      NaN
```

```
    Importations - Quantité  PIB par habitant  Stabilité politique  Population
5                       NaN           63176.7                1.39     77001.0
7                       NaN               NaN                 NaN     14584.0
9                       NaN               NaN                 NaN    275186.0
13                      NaN               NaN                 NaN    105366.0
18                      NaN           53451.0               -0.96   1494076.0
```

[200]:
```
(82, 9)
```

[201]:
```
# Enregistrement du csv dans le dossier "Etude_de_Marche"
#countries_with_nans.to_csv('Etude_de_Marche/countries_with_nans.csv')
#countries_with_nans.to_csv(
 #   r"C:\Users\FAMILLE\Desktop\Dossier Cours␣
 ↪Openclassrooms\Projet_11_Etude_de_marché_Python\DAN-P9-data\Etude_de_Marche\countries_with_
 ↪csv",
  #  index=False
#)
import os

# Définir le chemin complet du dossier de destination
folder_path = r"C:\Users\FAMILLE\Desktop\Dossier Cours␣
 ↪Openclassrooms\Projet_11_Etude_de_marché_Python"

# Créer le dossier s'il n'existe pas (optionnel mais sûr)
os.makedirs(folder_path, exist_ok=True)

# Chemin complet du fichier CSV à créer
file_path = os.path.join(folder_path, "countries_with_nans.csv")

# Enregistrer le DataFrame dans ce chemin
countries_with_nans.to_csv(file_path, index=False)

print(f"Fichier enregistré avec succès à l'emplacement : {file_path}")
```

```
Fichier enregistré avec succès à l'emplacement :
C:\Users\FAMILLE\Desktop\Dossier Cours
Openclassrooms\Projet_11_Etude_de_marché_Python\countries_with_nans.csv
```

[202]:
```
# Suppression des valeurs NaN dans le df final
df.dropna(inplace=True)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 157 entries, 0 to 234
Data columns (total 9 columns):
 #   Column                                                  Non-Null Count
Dtype
---  ------                                                  --------------
```

```
-----
 0   Zone                                                157 non-null
object
 1   Disponibilité alimentaire en quantité (kg/personne/an)    157 non-null
float64
 2   Disponibilité de protéines en quantité (g/personne/jour)  157 non-null
float64
 3   Disponibilité intérieure                            157 non-null
float64
 4   Exportations - Quantité                             157 non-null
float64
 5   Importations - Quantité                             157 non-null
float64
 6   PIB par habitant                                    157 non-null
float64
 7   Stabilité politique                                 157 non-null
float64
 8   Population                                          157 non-null
float64
dtypes: float64(8), object(1)
memory usage: 12.3+ KB
```

[203]:
```
df.head(100)
```

[203]:

	Zone	Disponibilité alimentaire en quantité (kg/personne/an)	Disponibilité de protéines en quantité (g/personne/jour)
0	Afghanistan	1.53	0.54
1	Afrique du Sud	35.69	14.11
2	Albanie	16.36	6.26
3	Algérie	6.38	1.97
4	Allemagne	19.47	7.96
..	…	…	…
126	Monténégro	15.98	5.79
127	Mozambique	3.59	1.20
128	Myanmar	30.37	9.93
129	Namibie	11.53	
131	Nicaragua	21.59	

```
129                                          4.25
131                                          6.91

     Disponibilité intérieure  Exportations - Quantité  \
0                        57.0                    -1.53
1                      2118.0                    63.00
2                        47.0                     0.00
3                       277.0                     0.00
4                      1739.0                   646.00
..                        …                       …
126                      10.0                     0.00
127                     116.0                    -4.59
128                    1666.0                     0.00
129                      28.0                    12.00
131                     138.0                     0.00

     Importations - Quantité  PIB par habitant  Stabilité politique  \
0                       29.0            2956.8                -2.79
1                      514.0           14823.6                -0.28
2                       38.0           14229.3                 0.37
3                        2.0           13805.4                -0.92
4                      842.0           61563.6                 0.57
..                        …                 …                   …
126                      8.0           23294.5                -0.06
127                     24.0            1510.4                -0.92
128                      3.0            5833.7                -1.08
129                     29.0           12147.4                 0.62
131                      6.0            7144.4                -0.05

     Population
0    36296113.0
1    57009756.0
2     2884169.0
3    41389189.0
4    82658409.0
..          …
126    627563.0
127  28649018.0
128  53382523.0
129   2402633.0
131   6384846.0

[100 rows x 9 columns]
```

Ajout de 3 variables feature engineering

1. Ratio export/import de volaille

[206]:
```
#Ratio export/import de volaille
df["ratio_export_import"] = df["Exportations - Quantité"] / (df["Importations -␣
↪Quantité"] + 1e-6)
df.head()
```

[206]:

	Zone	Disponibilité alimentaire en quantité (kg/personne/an)	Disponibilité de protéines en quantité (g/personne/jour)	Disponibilité intérieure	Exportations - Quantité	Importations - Quantité	PIB par habitant	Stabilité politique	Population	ratio_export_import
0	Afghanistan	1.53	0.54	57.0	-1.53	29.0	2956.8	-2.79	36296113.0	-0.052759
1	Afrique du Sud	35.69	14.11	2118.0	63.00	514.0	14823.6	-0.28	57009756.0	0.122568
2	Albanie	16.36	6.26	47.0	0.00	38.0	14229.3	0.37	2884169.0	0.000000
3	Algérie	6.38	1.97	277.0	0.00	2.0	13805.4	-0.92	41389189.0	0.000000
4	Allemagne	19.47	7.96	1739.0	646.00	842.0	61563.6	0.57	82658409.0	0.767221

Mesure l'ouverture commerciale : si >1, le pays exporte plus qu'il n'importe.Le +1e-6 évite une division par zéro.

2. Consommation apparente par habitant (kg/hab/an) C'est une forme de consommation par habitant calculée à partir de la disponibilité intérieure :

[209]:
```
df["conso_par_habitant"] = df["Disponibilité intérieure"] / df["Population"] *␣
↪1000  # car dispo est en tonnes ?
```

`Si "Disponibilité intérieure" est en milliers de tonnes et Population en individus, on multipli`

Sinon, si c'est déjà en tonnes, ajuste le facteur en fonction de l'unité réelle.

3. Score agro-économique (composite)

Un indicateur de potentiel "éco-agro" combinant :

`richesse (PIB/hab)`

stabilité politique

disponibilité alimentaire

[213]:
```
df["score_agro_eco"] = (
    0.4 * df["PIB par habitant"] +
    0.3 * df["Stabilité politique"] +  # attention : ici négatif = instable
    0.3 * df["Disponibilité alimentaire en quantité (kg/personne/an)"]
)
```

Un score combiné basé sur :

PIB par habitant (niveau de richesse économique)

Stabilité politique (important pour investir)

Disponibilité alimentaire (abondance de nourriture)

Ce score aide à comparer les pays selon plusieurs dimensions à la fois. Un score élevé → pays stable, riche, avec bonne dispo alimentaire.

[232]:
```
df
```

[232]:

	Zone	Disponibilité alimentaire en quantité (kg/personne/an)
0	Indonésie	7.19
1	Pakistan	5.86
2	Nigéria	1.01
3	Bangladesh	1.50
4	Fédération de Russie	30.98
..	…	…
95	Norvège	19.05
96	Congo	21.53
97	Costa Rica	26.52
98	Irlande	25.82
99	Libéria	10.67

```
    Disponibilité de protéines en quantité (g/personne/jour)  \
0                                                  2.42
1                                                  1.97
2                                                  0.31
3                                                  0.47
4                                                 10.44
..                                                  …
95                                                 7.81
96                                                 7.45
97                                                 7.93
98                                                11.60
99                                                 3.74

    Disponibilité intérieure  Exportations - Quantité  \
0                     2323.0                      0.0
1                     1282.0                      4.0
2                      202.0                      0.0
3                      250.0                     -2.5
4                     4556.0                    115.0
..                       …                       …
95                     102.0                      0.0
96                     110.0                      0.0
97                     134.0                      3.0
98                     128.0                     93.0
99                      50.0                      0.0

    Importations - Quantité  PIB par habitant  Stabilité politique  \
0                       1.0           11899.3                -0.50
1                       2.0            5191.9                -2.40
2                       0.0            5849.1                -2.00
3                       0.0            6020.3                -1.25
4                     226.0           36011.8                -0.64
..                      …                 …                    …
95                      2.0           85144.1                 1.15
96                    104.0            7426.4                -0.53
97                     17.0           22525.4                 0.60
98                     99.0           85225.2                 0.99
99                     48.0            1677.5                -0.32

     Population  ratio_export_import  conso_par_habitant  score_agro_eco  \
0   264650963.0         0.000000e+00            0.008778        4761.727
1   207906209.0         1.999999e+00            0.006166        2077.798
2   190873244.0         0.000000e+00            0.001058        2339.343
3   159685424.0        -2.500000e+06            0.001566        2408.195
4   145530082.0         5.088496e-01            0.031306       14413.822
..          …                    …                   …               …
```

```
95    5296326.0         0.000000e+00            0.019259       34063.700
96    5110695.0         0.000000e+00            0.021523        2976.860
97    4949954.0         1.764706e-01            0.027071        9018.296
98    4753279.0         9.393939e-01            0.026929       34098.123
99    4702226.0         0.000000e+00            0.010633         674.105

    Distance_France_km
0             12061.04
1              6040.23
2              4383.89
3              7854.91
4              5480.82
..                 …
95             1427.78
96             6107.36
97             8895.84
98              852.03
99             4904.59

[100 rows x 13 columns]
```

[256]:
```
#ajout manuellement valeur manquante distance France-Hong-Kong
df.loc[df['Zone'] == 'Chine-RAS de Hong-kong', 'Distance_France_km'] = 9779
```

[258]:
```
df[df['Zone'] == 'Chine-RAS de Hong-kong'][['Zone', 'Distance_France_km']]
```

[258]:
```
Empty DataFrame
Columns: [Zone, Distance_France_km]
Index: []
```

[260]:
```
df[df['Zone'].str.contains("hong", case=False, na=False)]
```

[260]:
```
                       Zone  \
74                  Hongrie
85  Chine - RAS de Hong-Kong

    Disponibilité alimentaire en quantité (kg/personne/an)  \
74                                              25.27
85                                              53.51

    Disponibilité de protéines en quantité (g/personne/jour)  \
74                                               9.80
85                                              22.26

    Disponibilité intérieure  Exportations - Quantité  \
74                     266.0                    210.0
85                     280.0                    663.0
```

```
    Importations - Quantité  PIB par habitant  Stabilité politique  \
74                     58.0           33940.3                 0.80
85                    907.0           65662.7                 0.82

    Population  ratio_export_import  conso_par_habitant  score_agro_eco  \
74   9729823.0             3.620690            0.027339       13583.941
85   7306322.0             0.730981            0.038323       26281.379

    Distance_France_km
74             1290.29
85                NaN
```

[262]:
```
df.loc[df['Zone'] == 'Chine - RAS de Hong-Kong', 'Distance_France_km'] = 9779
```

[266]:
```
df[df['Zone'] == 'Chine - RAS de Hong-Kong'][['Zone', 'Distance_France_km']]
```

[266]:
```
                        Zone  Distance_France_km
85  Chine - RAS de Hong-Kong              9779.0
```

[286]:
```
df.to_csv("top_100_pays_avec_distance_corrigee.csv", index=False,␣
↪encoding='utf-8-sig')
```

[288]:
```
from geopy.distance import geodesic
from geopy.geocoders import Nominatim
import time

# 1. Charger le fichier
df = pd.read_csv("top_100_pays.csv")

# 2. Coordonnées de Paris (France)
paris_coords = (48.8566, 2.3522)

# 3. Initialiser le géolocaliseur
geolocator = Nominatim(user_agent="geo_distance_app")

# 4. Fonction pour obtenir les coordonnées d'un pays
def get_country_coordinates(country_name):
    try:
        location = geolocator.geocode(country_name)
        if location:
            return (location.latitude, location.longitude)
    except:
        return None
    return None

# 5. Calculer les distances
```

```
country_distances = {}
for country in df['Zone'].unique():
    coords = get_country_coordinates(country)
    if coords:
        distance = geodesic(paris_coords, coords).kilometers
        country_distances[country] = round(distance, 2)
    else:
        country_distances[country] = None
    time.sleep(1)  # pause pour éviter surcharge de l'API

# 6. Ajouter la colonne de distance au DataFrame
df['Distance_France_km'] = df['Zone'].map(country_distances)

# 7. Sauvegarder le nouveau fichier
df.to_csv("top_100_pays_avec_distance.csv", index=False, encoding='utf-8-sig')

# 8. Afficher un aperçu
df[['Zone', 'Distance_France_km']].head()
```

[288]:

	Zone	Distance_France_km
0	Indonésie	12061.04
1	Pakistan	6040.23
2	Nigéria	4383.89
3	Bangladesh	7854.91
4	Fédération de Russie	5480.82

[289]:
```
# Liste des pays à exclure
pays_a_exclure = ["Chine, continentale", "Inde", "Brésil", "États-Unis␣
 ↪d'Amérique"]

# Suppression des lignes dont la valeur de 'Zone' est dans la liste
df_sans_outliers = df[~df['Zone'].isin(pays_a_exclure)]
```

[290]:
```
#vérification des pays outliers exclus
print(df_sans_outliers['Zone'].unique())
```

```
['Indonésie' 'Pakistan' 'Nigéria' 'Bangladesh' 'Fédération de Russie'
 'Japon' 'Mexique' 'Éthiopie' 'Philippines' 'Égypte' 'Viet Nam'
 'Allemagne' "Iran (République islamique d')" 'Thaïlande'
 "Royaume-Uni de Grande-Bretagne et d'Irlande du Nord" 'France' 'Italie'
 'Afrique du Sud' 'République-Unie de Tanzanie' 'Myanmar'
 'République de Corée' 'Kenya' 'Colombie' 'Espagne' 'Ukraine' 'Argentine'
 'Algérie' 'Ouganda' 'Soudan' 'Pologne' 'Iraq' 'Canada' 'Afghanistan'
 'Maroc' 'Arabie saoudite' 'Pérou' 'Malaisie' 'Angola' 'Ghana'
 'Mozambique' 'Népal' 'Madagascar' 'Australie' 'Cameroun' "Côte d'Ivoire"
 'Niger' 'Sri Lanka' 'Roumanie' 'Burkina Faso' 'Mali' 'Chili' 'Kazakhstan'
 'Malawi' 'Guatemala' 'Zambie' 'Équateur' 'Cambodge' 'Sénégal' 'Tchad'
 'Zimbabwe' 'Guinée' 'Rwanda' 'Tunisie' 'Belgique'
```

```
 'Bolivie (État plurinational de)' 'Bénin' 'Haïti' 'Tchéquie' 'Grèce'
 'République dominicaine' 'Portugal' 'Suède' 'Azerbaïdjan' 'Jordanie'
 'Hongrie' 'Émirats arabes unis' 'Bélarus' 'Honduras' 'Tadjikistan'
 'Serbie' 'Autriche' 'Suisse' 'Israël' 'Togo' 'Sierra Leone'
 'Chine - RAS de Hong-Kong' 'Bulgarie' 'Paraguay' 'Liban' 'El Salvador'
 'Nicaragua' 'Kirghizistan' 'Danemark' 'Finlande' 'Slovaquie' 'Norvège'
 'Congo' 'Costa Rica' 'Irlande' 'Libéria']
```

[291]:
```
top_100_pays = df_sans_outliers.sort_values(by='Population', ascending=False).
 ↪head(100)
```

[292]:
```
print(top_100_pays[['Zone', 'Population']])
```

```
                    Zone   Population
0              Indonésie  264650963.0
1               Pakistan  207906209.0
2                Nigéria  190873244.0
3             Bangladesh  159685424.0
4   Fédération de Russie  145530082.0
..                   …          …
95               Norvège    5296326.0
96                 Congo    5110695.0
97            Costa Rica    4949954.0
98               Irlande    4753279.0
99               Libéria    4702226.0

[100 rows x 2 columns]
```

[293]:
```
print(top_100_pays.columns)
```

```
Index(['Zone', 'Disponibilité alimentaire en quantité (kg/personne/an)',
       'Disponibilité de protéines en quantité (g/personne/jour)',
       'Disponibilité intérieure', 'Exportations - Quantité',
       'Importations - Quantité', 'PIB par habitant', 'Stabilité politique',
       'Population', 'ratio_export_import', 'conso_par_habitant',
       'score_agro_eco', 'Distance_France_km'],
      dtype='object')
```

[294]:
```
import os

folder_path = r"C:\Users\FAMILLE\Desktop\Dossier Cours␣
 ↪Openclassrooms\Projet_11_Etude_de_marché_Python"
os.makedirs(folder_path, exist_ok=True)

# Enregistrer countries_with_nans
countries_with_nans.to_csv(os.path.join(folder_path, "countries_with_nans.
 ↪csv"), index=False)
```

```
# Enregistrer top_100_pays
top_100_pays.to_csv(os.path.join(folder_path, "top_100_pays.csv"), index=False)

print("Les deux fichiers ont été enregistrés dans le dossier :", folder_path)
```

```
Les deux fichiers ont été enregistrés dans le dossier :
C:\Users\FAMILLE\Desktop\Dossier Cours
Openclassrooms\Projet_11_Etude_de_marché_Python
```

[295]:
```
# Vérifier si "Chine" et "Inde" sont présents dans la colonne 'Zone'
pays_a_verifier = ["Chine", "Inde","Brésil", "États-Unis d'Amérique"]
for pays in pays_a_verifier:
    if pays in top_100_pays['Zone'].values:
        print(f"{pays} est présent dans la colonne 'Zone'")
    else:
        print(f"{pays} n'est PAS présent dans la colonne 'Zone'")
```

```
Chine n'est PAS présent dans la colonne 'Zone'
Inde n'est PAS présent dans la colonne 'Zone'
Brésil n'est PAS présent dans la colonne 'Zone'
États-Unis d'Amérique n'est PAS présent dans la colonne 'Zone'
```

[296]:
```
top_100_pays.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 13 columns):
 #   Column                                                Non-Null Count
Dtype
---  ------                                                --------------
-----
 0   Zone                                                  100 non-null
object
 1   Disponibilité alimentaire en quantité (kg/personne/an)    100 non-null
float64
 2   Disponibilité de protéines en quantité (g/personne/jour)  100 non-null
float64
 3   Disponibilité intérieure                              100 non-null
float64
 4   Exportations - Quantité                               100 non-null
float64
 5   Importations - Quantité                               100 non-null
float64
 6   PIB par habitant                                      100 non-null
float64
 7   Stabilité politique                                   100 non-null
float64
 8   Population                                            100 non-null
float64
```

```
 9   ratio_export_import                                 100 non-null
float64
 10  conso_par_habitant                                  100 non-null
float64
 11  score_agro_eco                                      100 non-null
float64
 12  Distance_France_km                                  99 non-null
float64
dtypes: float64(12), object(1)
memory usage: 10.3+ KB
```

[307]:
```
# 1. Nettoyer les noms
df['Zone'] = df['Zone'].str.strip()

# 2. Modifier directement dans df (le bon DataFrame)
df.loc[df['Zone'].str.contains('Hong-Kong', case=False, na=False),␣
 ↪'Distance_France_km'] = 9779

# 3. Vérifier que c'est bien dans df
print(df.loc[df['Zone'].str.contains('Hong-Kong', case=False, na=False),␣
 ↪['Zone', 'Distance_France_km']])
```

```
                        Zone  Distance_France_km
85  Chine - RAS de Hong-Kong              9779.0
```

[309]:
```
df.to_csv("top_100_pays_avec_distance_corrigee.csv", index=False,␣
 ↪encoding='utf-8-sig')
```

[311]:
```
df[df['Distance_France_km'].isna()]
```

[311]:
```
Empty DataFrame
Columns: [Zone, Disponibilité alimentaire en quantité (kg/personne/an),
Disponibilité de protéines en quantité (g/personne/jour), Disponibilité
intérieure, Exportations - Quantité, Importations - Quantité, PIB par habitant,
Stabilité politique, Population, ratio_export_import, conso_par_habitant,
score_agro_eco, Distance_France_km]
Index: []
```

[ ]:
```

```