Khalid OURO-ADOYI
(DATA SCIENTIST)

academy

EdTech e-learning

Country Expansion

DATASETS

- EdStatsSeries
- EdStatsFootNote
- EdStatsData
- EdStatsCountry
- EdStatsCountry-Series

WORLD BANK GROUP

Importation Bases

```
import pandas as pd

# Import des fichiers CSV

series = pd.read_csv("/content/drive/MyDrive/Projet2/EdStatsSeries.csv")


footnote = pd.read_csv("/content/drive/MyDrive/Projet2/EdStatsFootNote.csv")

data = pd.read_csv("/content/drive/MyDrive/Projet2/EdStatsData.csv")

country = pd.read_csv("/content/drive/MyDrive/Projet2/EdStatsCountry.csv")

country_series =  pd.read_csv("/content/drive/MyDrive/Projet2/EdStatsCountry-Series.csv")
```

Informations Basiques

Dimensions

```
series.shape
```

```
(3665, 21)
```

5 Enrégistrements

```
footnote.head()
```

	CountryCode	SeriesCode	Year	DESCRIPTION	Unnamed: 4
0	ABW	SE.PRE.ENRL.FE	YR2001	Country estimation.	NaN
1	ABW	SE.TER.TCHR.FE	YR2005	Country estimation.	NaN
2	ABW	SE.PRE.TCHR.FE	YR2000	Country estimation.	NaN
3	ABW	SE.SEC.ENRL.GC	YR2004	Country estimation.	NaN
4	ABW	SE.PRE.TCHR	YR2006	Country estimation.	NaN

Valeurs manquantes

```
# Afficher le nombre de NA pour chaque caractéristique
footnote.isnull().sum()
```

	0
CountryCode	0
SeriesCode	0
Year	0
DESCRIPTION	0
Unnamed: 4	643638

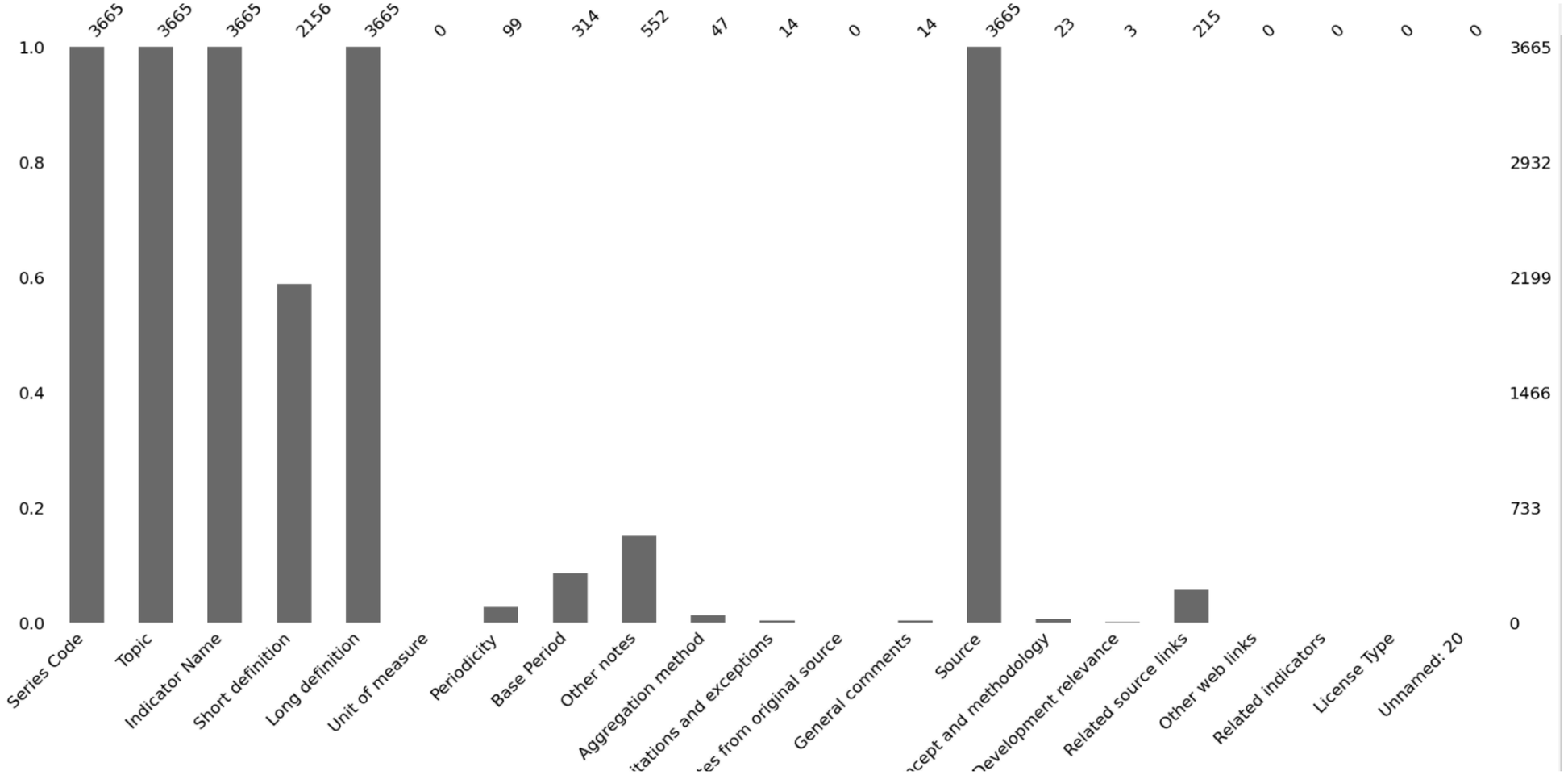
dtype: int64

Informations

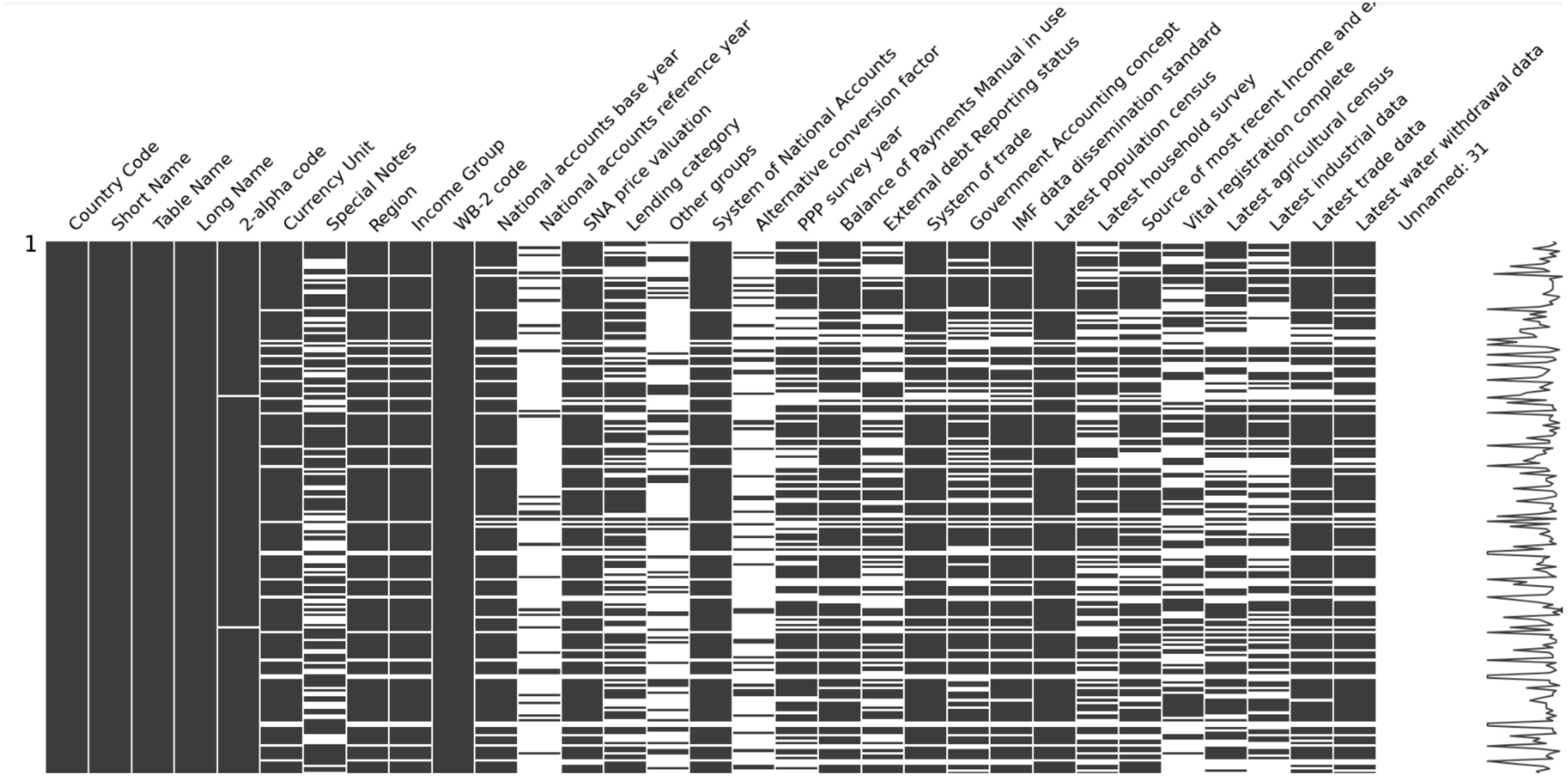
```
series.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3665 entries, 0 to 3664
Data columns (total 21 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Series Code                           3665 non-null   object
 1   Topic                                 3665 non-null   object
 2   Indicator Name                        3665 non-null   object
 3   Short definition                      2156 non-null   object
 4   Long definition                       3665 non-null   object
 5   Unit of measure                       0 non-null      float64
 6   Periodicity                           99 non-null     object
 7   Base Period                           314 non-null    object
 8   Other notes                           552 non-null    object
 9   Aggregation method                    47 non-null     object
 10  Limitations and exceptions            14 non-null     object
 11  Notes from original source            0 non-null      float64
 12  General comments                      14 non-null     object
 13  Source                                3665 non-null   object
 14  Statistical concept and methodology   23 non-null     object
 15  Development relevance                 3 non-null      object
 16  Related source links                  215 non-null    object
 17  Other web links                       0 non-null      float64
 18  Related indicators                    0 non-null      float64
 19  License Type                          0 non-null      float64
 20  Unnamed: 20                           0 non-null      float64
dtypes: float64(6), object(15)
memory usage: 601.4+ KB
```

Visualisation de Valeurs Manquantes



Visualisation de Valeurs Manquantes



Valeurs Manquantes & Doublons

Doublons

```
#Suppression de doublons
footnote = footnote.drop_duplicates()
```

Valeurs manquantes

```
# Supression des lignes portant NA
data_final = data_filtered.dropna()
data_final.shape
```

Retention de la base EdStatData

```
8] data = data[['Country Name','Country Code','Indicator Name', 'Indicator Code', '2015']]
   data.head()
```

	Country Name	Country Code	Indicator Name	Indicator Code	2015
91625	Afghanistan	AFG	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	NaN
91626	Afghanistan	AFG	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.F	NaN
91627	Afghanistan	AFG	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.GPI	NaN
91628	Afghanistan	AFG	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.M	NaN
91629	Afghanistan	AFG	Adjusted net enrolment rate, primary, both sex...	SE.PRM.TENR	NaN

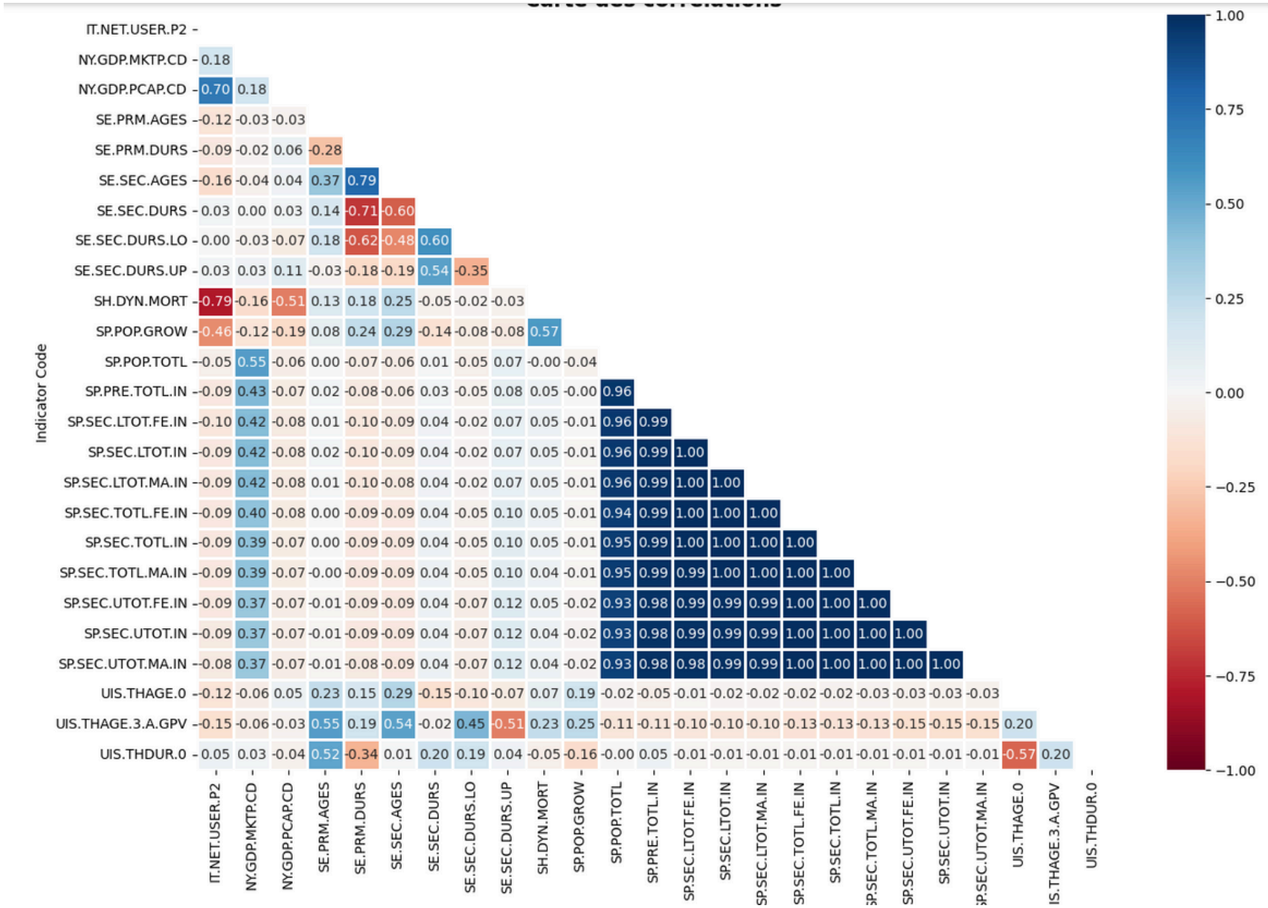
```
9] #Suppression de NA
   data_cleaned = data.dropna()
```

Indicateurs

Liste des indicateurs

```
Index(['SE.PRM.AGES', 'SE.PRM.DURS', 'SE.SEC.AGES', 'SE.SEC.DURS.UP',
       'SE.SEC.DURS', 'UIS.THDUR.0', 'SE.SEC.DURS.LO', 'SP.POP.TOTL',
       'SP.POP.GROW', 'UIS.THAGE.0', 'UIS.THAGE.3.A.GPV', 'IT.NET.USER.P2',
       'SP.SEC.UTOT.IN', 'SP.SEC.LTOT.IN', 'SP.SEC.TOTL.IN',
       'SP.SEC.LTOT.FE.IN', 'SP.SEC.UTOT.FE.IN', 'NY.GDP.MKTP.CD',
       'SP.SEC.LTOT.MA.IN', 'SP.SEC.TOTL.FE.IN', 'SP.SEC.TOTL.MA.IN',
       'NY.GDP.PCAP.CD', 'SP.SEC.UTOT.MA.IN', 'SP.PRE.TOTL.IN', 'SH.DYN.MORT'],
      dtype='object', name='Indicator Code')
```

Carte des corrélations



Indicateurs retenus

Statistique descriptive

```
 Analyse de IT.NET.USER.P2
Statistiques descriptives pour IT.NET.USER.P2:
count    195.000000
mean      48.102884
std       28.679733
min        1.083733
25%       21.563148
50%       48.884644
75%       72.867350
max       98.323610
Name: IT.NET.USER.P2, dtype: float64
```

NY.GDP.MKTP.CD PIB aux prix du marché (USD courants)

NY.GDP.PCAP.CD PIB par habitant (USD courants)

IT.NET.USER.P2 Utilisateurs d'Internet (pour 100 personnes)

UIS.THAGE.0 Âge officiel d'entrée dans l'enseignement préprimaire

SE.PRM.AGES Âge officiel d'entrée dans l'enseignement

UIS.THAGE.3.A.GPV Âge officiel d'entrée dans l'enseignement

SP.POP.GROW Croissance de la population (% annuel)

SP.POP.TOTL Population totale

SE.SEC.DURS.LO Durée théorique de l'enseignement secondaire inférieur

UIS.THDUR.0 Durée théorique de l'enseignement préprimaire

SE.PRM.DURS Durée théorique de l'enseignement primaire (années)

SE.SEC.DURS.UP Durée théorique de l'enseignement secondaire supérieur

Indicateurs retenus

Statistique descriptive

```
 Analyse de IT.NET.USER.P2
Statistiques descriptives pour IT.NET.USER.P2:
count    195.000000
mean      48.102884
std       28.679733
min        1.083733
25%       21.563148
50%       48.884644
75%       72.867350
max       98.323610
Name: IT.NET.USER.P2, dtype: float64
```

NY.GDP.MKTP.CD PIB aux prix du marché (USD courants)

NY.GDP.PCAP.CD PIB par habitant (USD courants)

IT.NET.USER.P2 Utilisateurs d'Internet (pour 100 personnes)

UIS.THAGE.0 Âge officiel d'entrée dans l'enseignement préprimaire

SE.PRM.AGES Âge officiel d'entrée dans l'enseignement

UIS.THAGE.3.A.GPV Âge officiel d'entrée dans l'enseignement

SP.POP.GROW Croissance de la population (% annuel)

SP.POP.TOTL Population totale

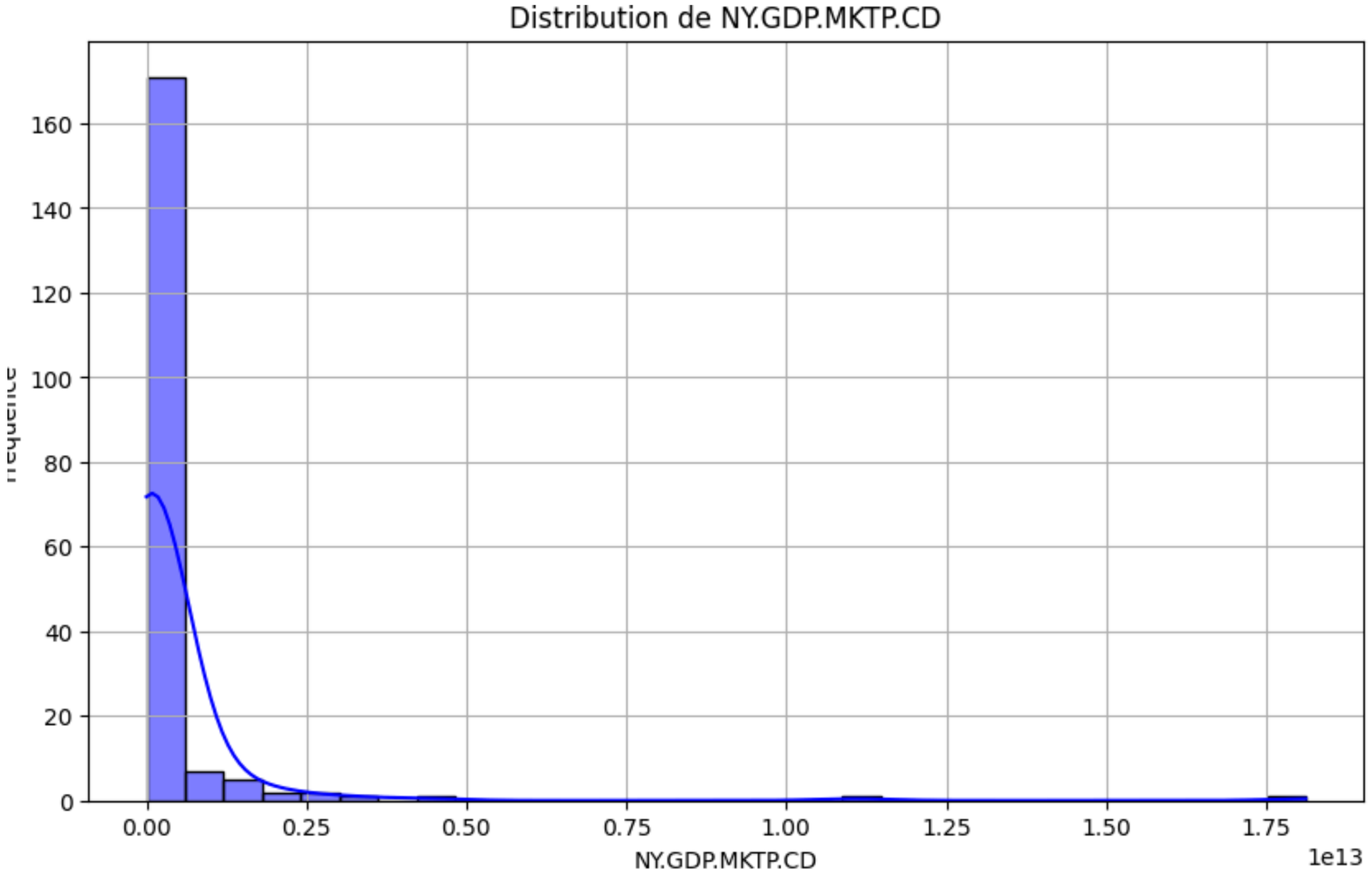
SE.SEC.DURS.LO Durée théorique de l'enseignement secondaire inférieur

UIS.THDUR.0 Durée théorique de l'enseignement préprimaire

SE.PRM.DURS Durée théorique de l'enseignement primaire (années)

SE.SEC.DURS.UP Durée théorique de l'enseignement secondaire supérieur

Histogramme de distribution “NY.GDP.MKTP.CD”



Le graphique montre une distribution très asymétrique du PIB (NY.GDP.MKTP.CD) :

- Concentration des faibles PIB : La majorité des pays ont un PIB faible.
- Longue traîne vers la droite : Quelques pays ont un PIB extrêmement élevé (ex. États-Unis, Chine).
- Inégalités économiques mondiales : Cela illustre de fortes disparités entre les économies, avec un petit nombre de pays dominant la production économique mondiale.

Normalisation des valeurs

Normalisation

```
# Normalisation min-max par indicateur pour ramener les valeurs entre 0 et 1
New_data["Normalized"] = New_data.groupby("Indicator Code")["2015"].transform(
    lambda x: (x - x.min()) / (x.max() - x.min())
)

New_data.head()
```

```
<ipython-input-167-bcb6bffeeadd>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-v
  New_data["Normalized"] = New_data.groupby("Indicator Code")["2015"].transform(
```

	Country Name	Country Code	Indicator Name	Indicator Code	2015	Normalized
92868	Afghanistan	AFG	GDP at market prices (current US$)	NY.GDP.MKTP.CD	1.921556e+10	0.001059
92870	Afghanistan	AFG	GDP per capita (current US$)	NY.GDP.PCAP.CD	5.695779e+02	0.001602
93000	Afghanistan	AFG	Internet users (per 100 people)	IT.NET.USER.P2	8.260000e+00	0.073800
93308	Afghanistan	AFG	Official entrance age to pre-primary education...	UIS.THAGE.0	3.000000e+00	0.000000
93309	Afghanistan	AFG	Official entrance age to primary education (ye...	SE.PRM.AGES	7.000000e+00	1.000000

```
New_data.shape
```

```
(2366, 6)
```

Aggrégation des scores

```
# Regrouper par 'Country Name' et calculer la somme de 'Normalized' pour tous les indicateurs
class_countries = (
    New_data.groupby(['Country Name'])['Normalized']
    .sum()
    .reset_index()
)

# Renommer les colonnes pour plus de clarté
class_countries.columns = ['Country Name', 'Total Normalized']

# Classer par ordre décroissant
class_countries = class_countries.sort_values(by='Total Normalized', ascending=False)
class_countries
```

	Country Name	Total Normalized
108	Liechtenstein	6.551846
178	Switzerland	6.285402
181	Tanzania	5.639223
177	Sweden	5.529741
195	United States	5.525025
...	...	...
27	British Virgin Islands	1.583333
3	American Samoa	1.387095
74	Guam	1.307687
66	French Polynesia	1.045267
89	Isle of Man	0.865336

Liste des pays prometteurs

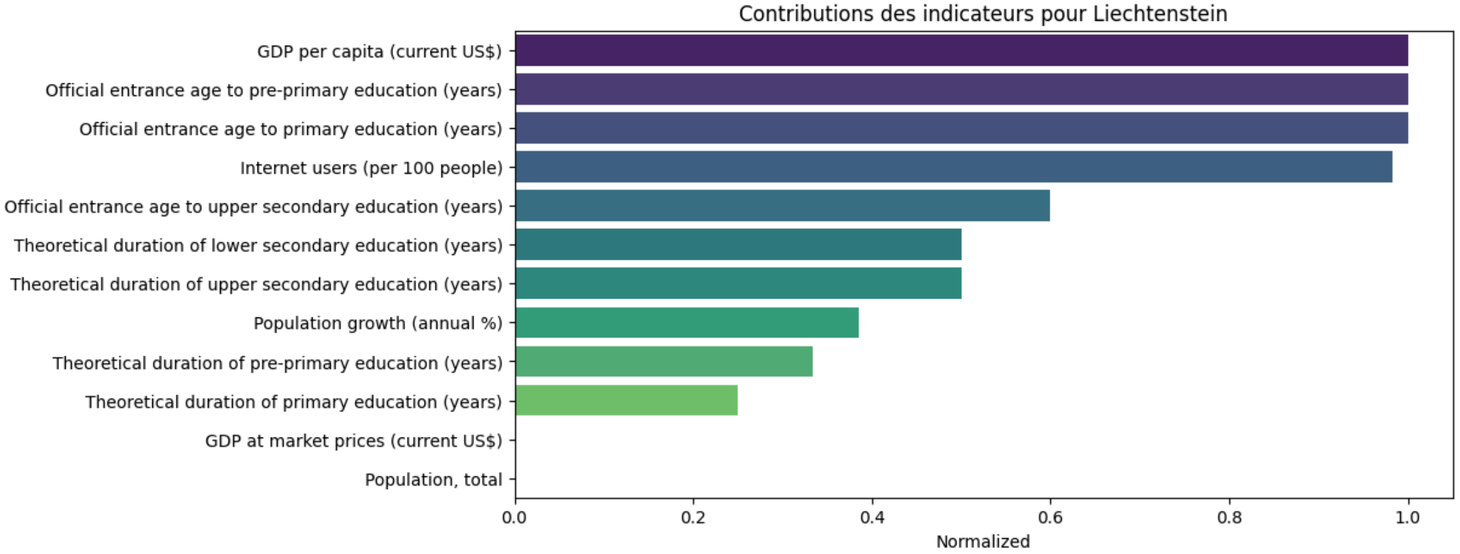
```
# Afficher les 10 premiers pays après le tri
top_10_countries = class_countries.head(10)
print(top_10_countries)
```

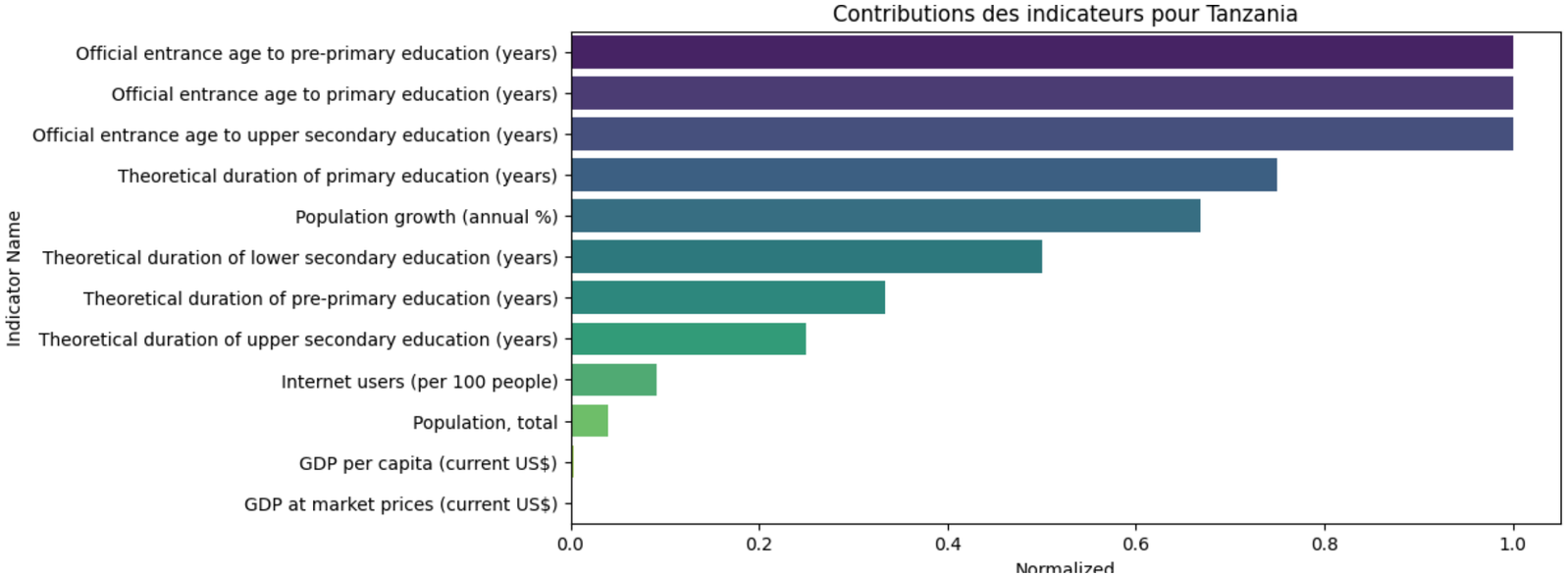
```
          Country Name  Total Normalized
108      Liechtenstein          6.551846
178        Switzerland          6.285402
181           Tanzania          5.639223
177             Sweden          5.529741
195      United States          5.525025
39               China          5.337209
64             Finland          5.331403
58   Equatorial Guinea          5.264083
136              Niger          5.253534
110         Luxembourg          5.242684
```

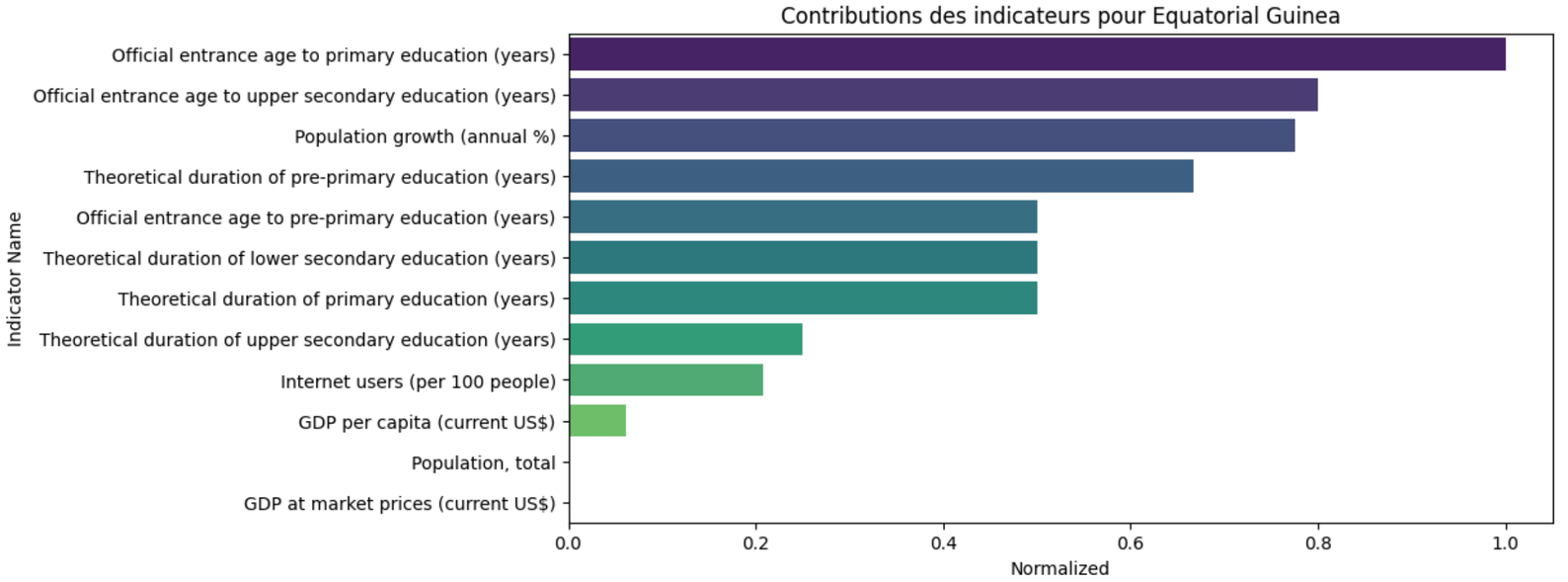
Pays surprenants

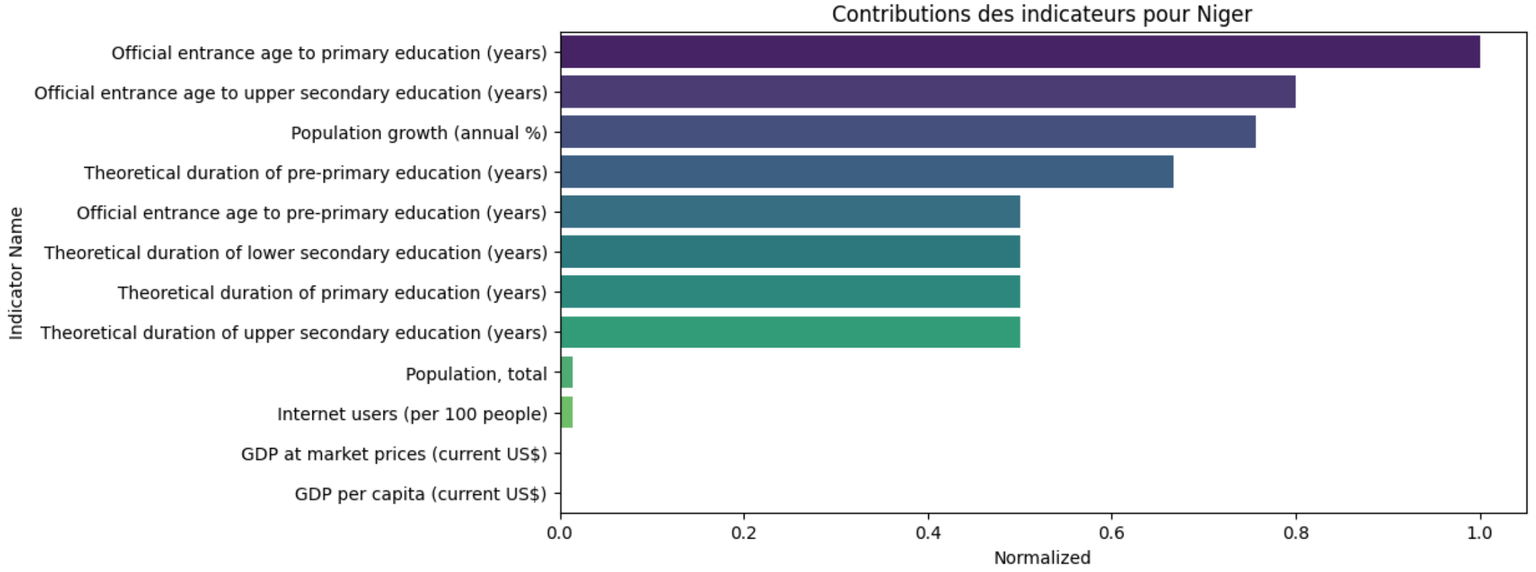
- Liechtenstein
- Tanzanie
- Guinée Equatoriale
- Niger

Distributions






Contributions des indicateurs pour Equatorial Guinea
Official entrance age to primary education (years)
Official entrance age to upper secondary education (years)
Population growth (annual %)
Theoretical duration of pre-primary education (years)
Official entrance age to pre-primary education (years)
Theoretical duration of lower secondary education (years)
Theoretical duration of primary education (years)
Theoretical duration of upper secondary education (years)
Internet users (per 100 people)
GDP per capita (current US$)
Population, total
GDP at market prices (current US$)
Indicator Name
0.0
0.2
0.4
0.6
0.8
1.0
Normalized


Contributions des indicateurs pour Niger
Official entrance age to primary education (years)
Official entrance age to upper secondary education (years)
Population growth (annual %)
Theoretical duration of pre-primary education (years)
Official entrance age to pre-primary education (years)
Theoretical duration of lower secondary education (years)
Theoretical duration of primary education (years)
Theoretical duration of upper secondary education (years)
Population, total
Internet users (per 100 people)
GDP at market prices (current US$)
GDP per capita (current US$)
Indicator Name
0.0
0.2
0.4
0.6
0.8
1.0
Normalized

Explications

Liechtenstein :

- Très petite population mais un PIB par habitant extrêmement élevé, ce qui tire son score global vers le haut.
- Excellents indicateurs en éducation et accès technologique (utilisation d'Internet très élevée).

Tanzanie :

- Croissance démographique rapide : Cela peut être vu comme un potentiel économique à long terme.
- Indicateurs éducatifs alignés sur des standards globaux en termes d'âge d'entrée et de durée des cycles scolaires.

Guinée équatoriale :

- PIB par habitant relativement élevé, grâce à l'exploitation pétrolière.
- Malgré cela, le développement humain et l'accès aux services restent faibles, ce qui rend sa présence ici surprenante.

Niger :

- Croissance démographique très élevée, ce qui booste son score.
- Bien que son PIB par habitant soit très faible, les indicateurs éducatifs (âge d'entrée et durées théoriques) respectent les standards internationaux.

En somme:

Ces pays ont des atouts spécifiques qui compensent leurs faiblesses :

- Le Liechtenstein se distingue par sa richesse.
- La Tanzanie et le Niger se démarquent par leur démographie et des indicateurs éducatifs stables.
- La Guinée équatoriale bénéficie de sa richesse pétrolière.

Conclusions

L'examen des indicateurs a montré que des pays telles que le Liechtenstein, le Luxembourg, les États-Unis et la Suisse, avec leur PBI élevé et un accès Internet considérable, constituent des marchés intéressant pour notre expansion.

Inversement, bien que des pays tels que la Tanzanie, le Niger et la Guinée équatoriale affichent un PIB modeste, ils bénéficient d'un potentiel considérable grâce à leur importante expansion démographique et des indicateurs d'éducation encourageants.

Ces informations facilitent notre ciblage de marchés : des nations prospères pour des contenus spécialisés et des pays en croissance pour des produits destinés à un auditoire plus vaste et en forte progression.