Projet n°2 : “Analysez des données du système éducatif-Banque Mondiale”

SOUTENANCE DE PROJET

15 FÉVRIER 2022

academy

BANQUE MONDIALE

Programme

- I - Présentation du jeu de données et sa problématique
- II - Analyse des données
- III - Conclusion sur la pertinence du jeu de données

I- Présentation du jeu de données et sa problématique

Rappel de la problématique

- Academy est une start-up de la EdTech
- E-learnings : Contenus de formation de niveau lycée et université
- Objectif : l'expansion à l'international

academy

BANQUE MONDIALE

Objectif du projet :

Accompagner le projet d'expansion en réalisant une analyse pré exploratoire et déterminer si les données sur l'éducation de la Banque Mondiale répondent à l'objectif

Processus d'analyse pré exploratoire

1

Connaître les données

Quelles informations?
Quelles années?

2

Identifier les indicateurs exploitables

Quantités de données manquantes?

3

Comparer les pays

Quels indicateurs choisir?
Analyse des résultats obtenus
Quels sont les pays à cibler par Academy?

Présentation du jeu de données

EdStatsCountry.csv

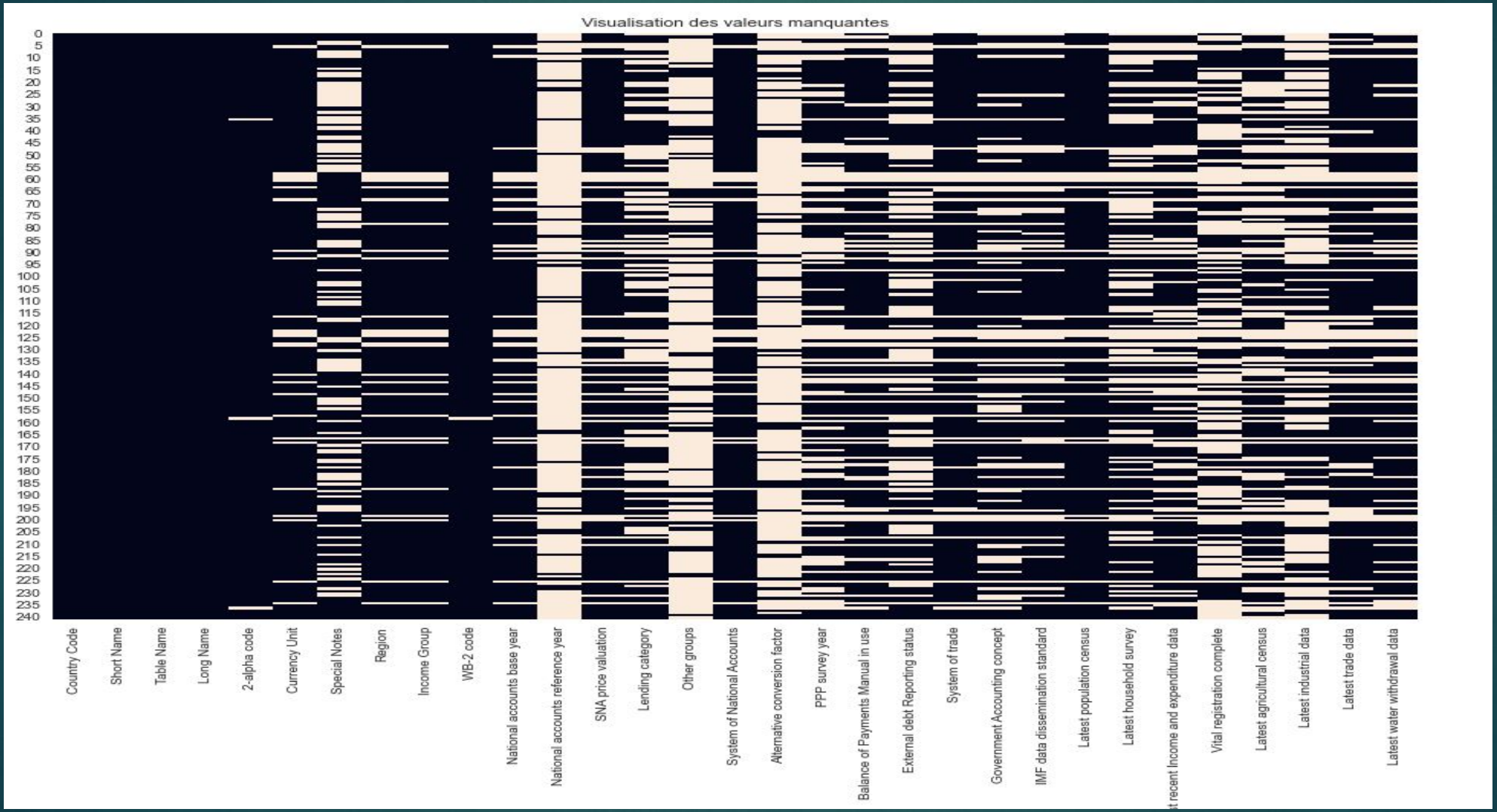
Ce jeu de données contient des informations géographiques sur les pays , les regroupements par 7 régions(Toutes les régions du monde sont représentées) et par groupe de revenus…, des données économiques globales et des dates de référence des dernières études + des statistiques.

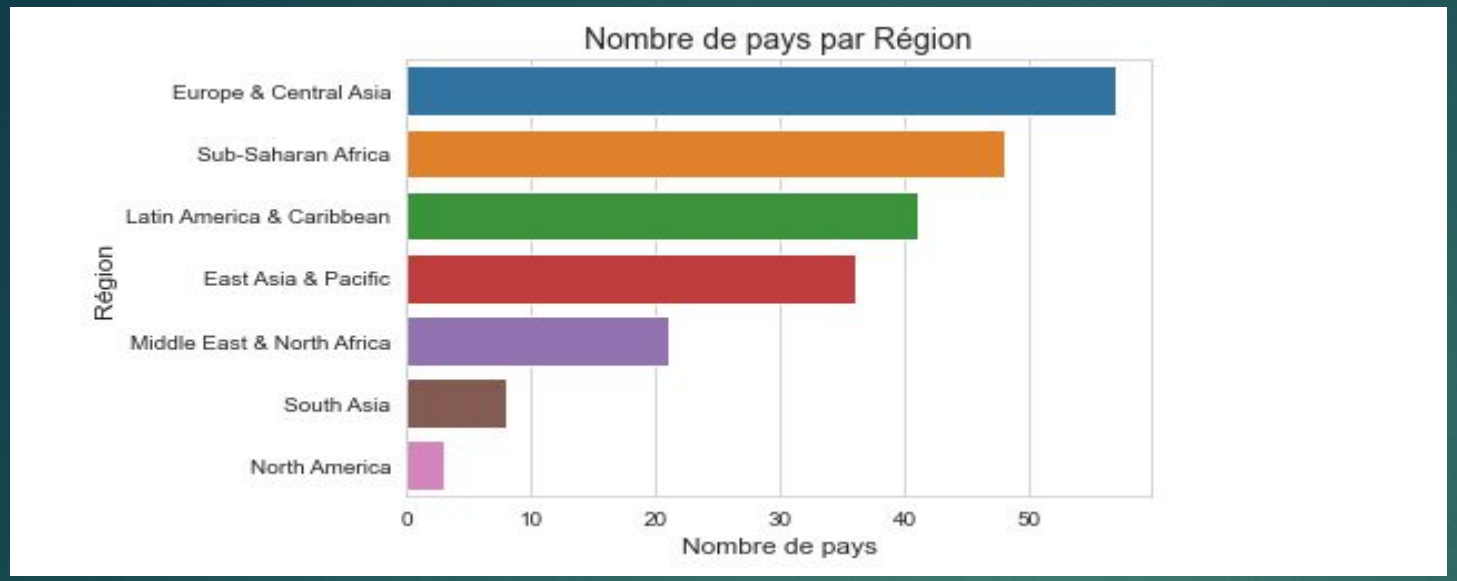
Taille : 241 lignes, 32 colonnes

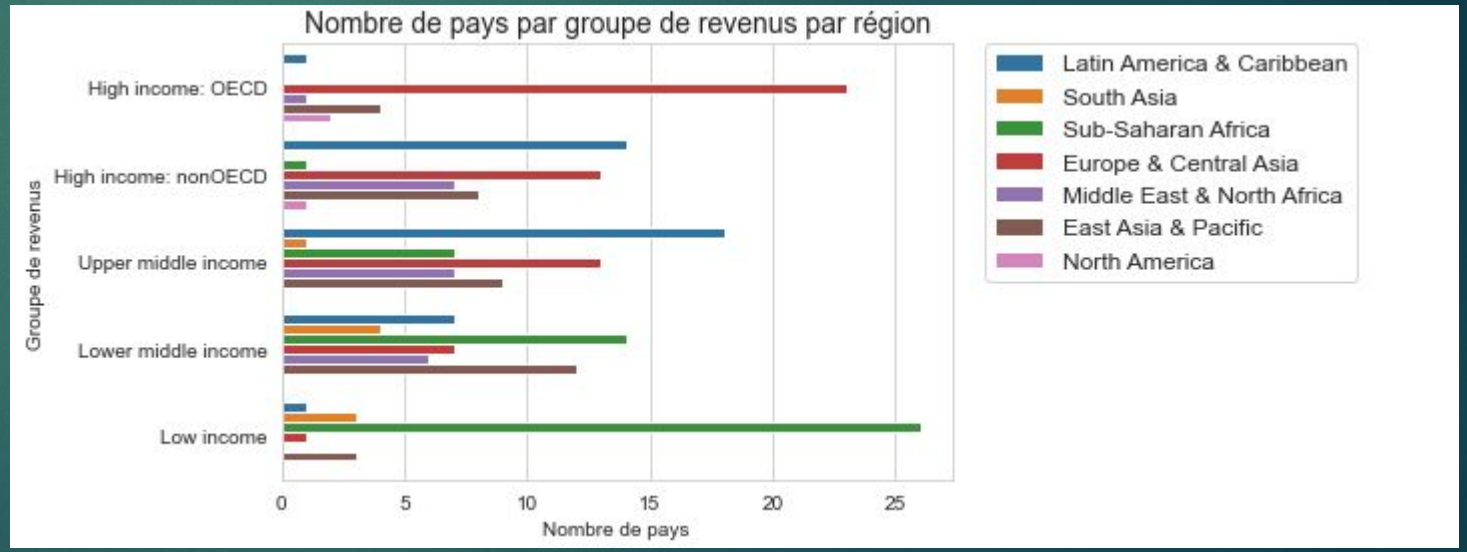
Toutes les régions du monde sont représentées.

28.28 % cellules manquantes . Aucun doublon

Visualisation des données manquantes




Nombre de pays par Région
Europe & Central Asia
Sub-Saharan Africa
Latin America & Caribbean
East Asia & Pacific
Middle East & North Africa
South Asia
North America
Région
0
10
20
30
40
50
Nombre de pays


Nombre de pays par groupe de revenus par région
High income: OECD
High income: nonOECD
Upper middle income
Lower middle income
Low income
Groupe de revenus
0
5
10
15
20
25
Nombre de pays
Latin America & Caribbean
South Asia
Sub-Saharan Africa
Europe & Central Asia
Middle East & North Africa
East Asia & Pacific
North America

Présentation du jeu de données

EdStatsCountry_Series.csv

Le jeu de données contient les références des sources de certains indicateurs par pays dans la colonne (Description), ces indicateurs sont présent dans le jeu de données EdStatsCountry.csv.

Taille : 613 lignes, 4 colonnes

30 pays qui manque par rapport à EdStatsCountry.csv
0.0% de cellules manquantes . Aucun doublon

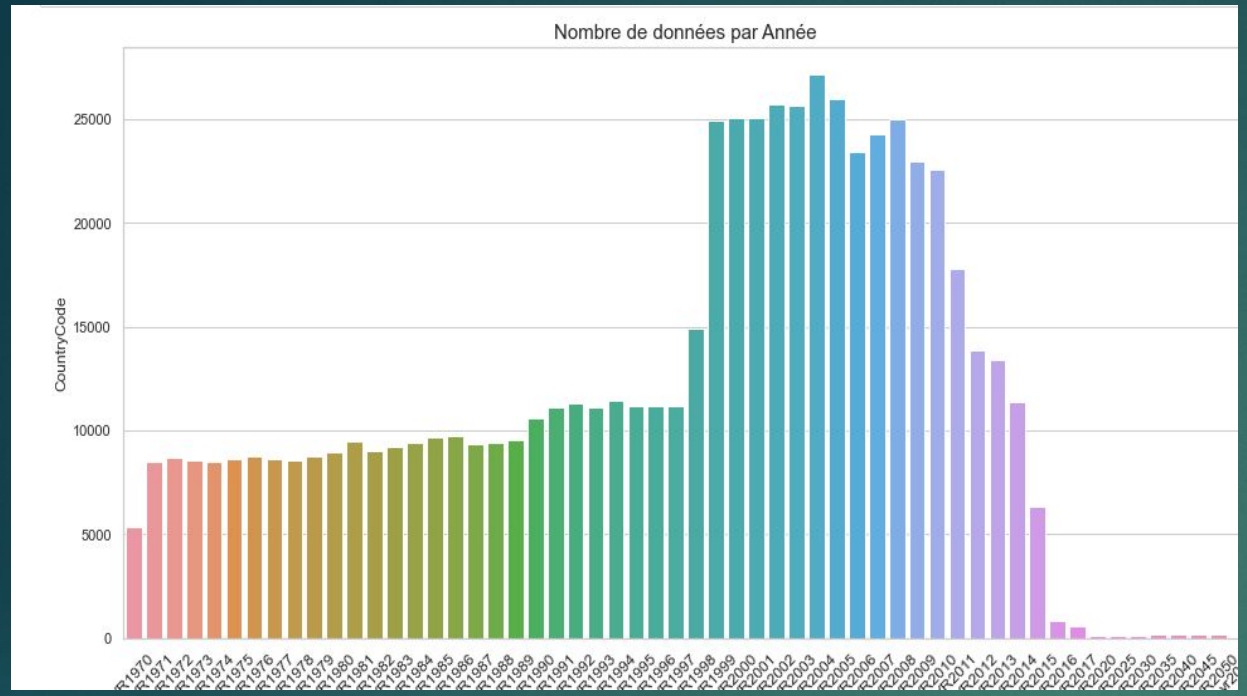
Présentation du jeu de données

EdStatsFootNote.csv

Le jeu de données contient des indicateurs et des séries par pays, les années de réalisation et des remarques sur les mises à jour.

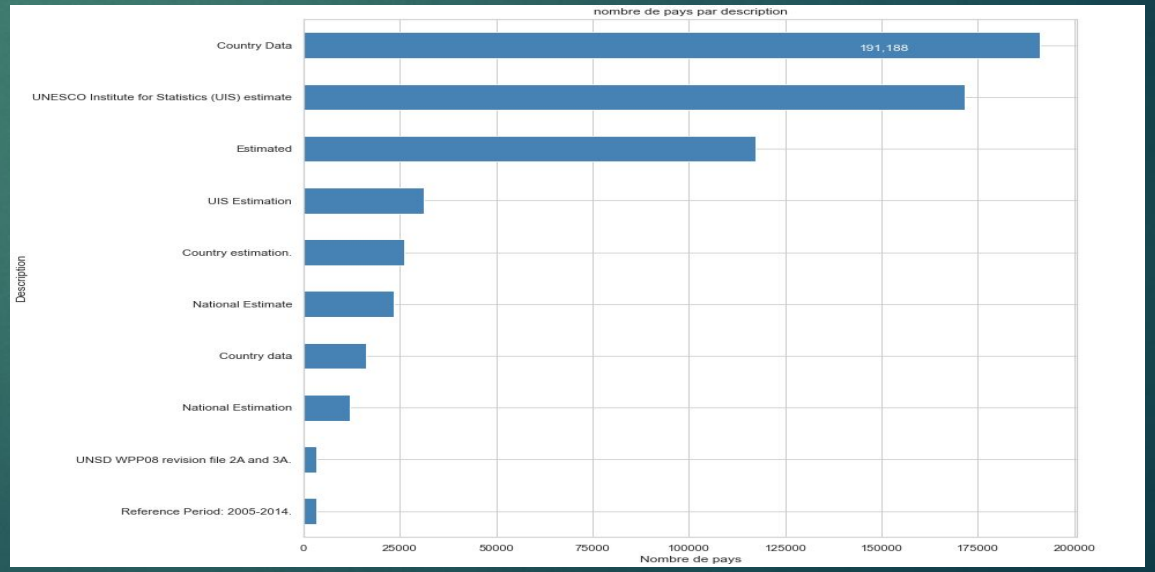
Taille : 643638 lignes, 5 colonnes
0.0% de cellules manquantes. Aucun doublon

3 pays qui manque par rapport à EdStatsCountry_Series



Nombre de données par Année

Nombre de pays par description



Présentation du jeu de données

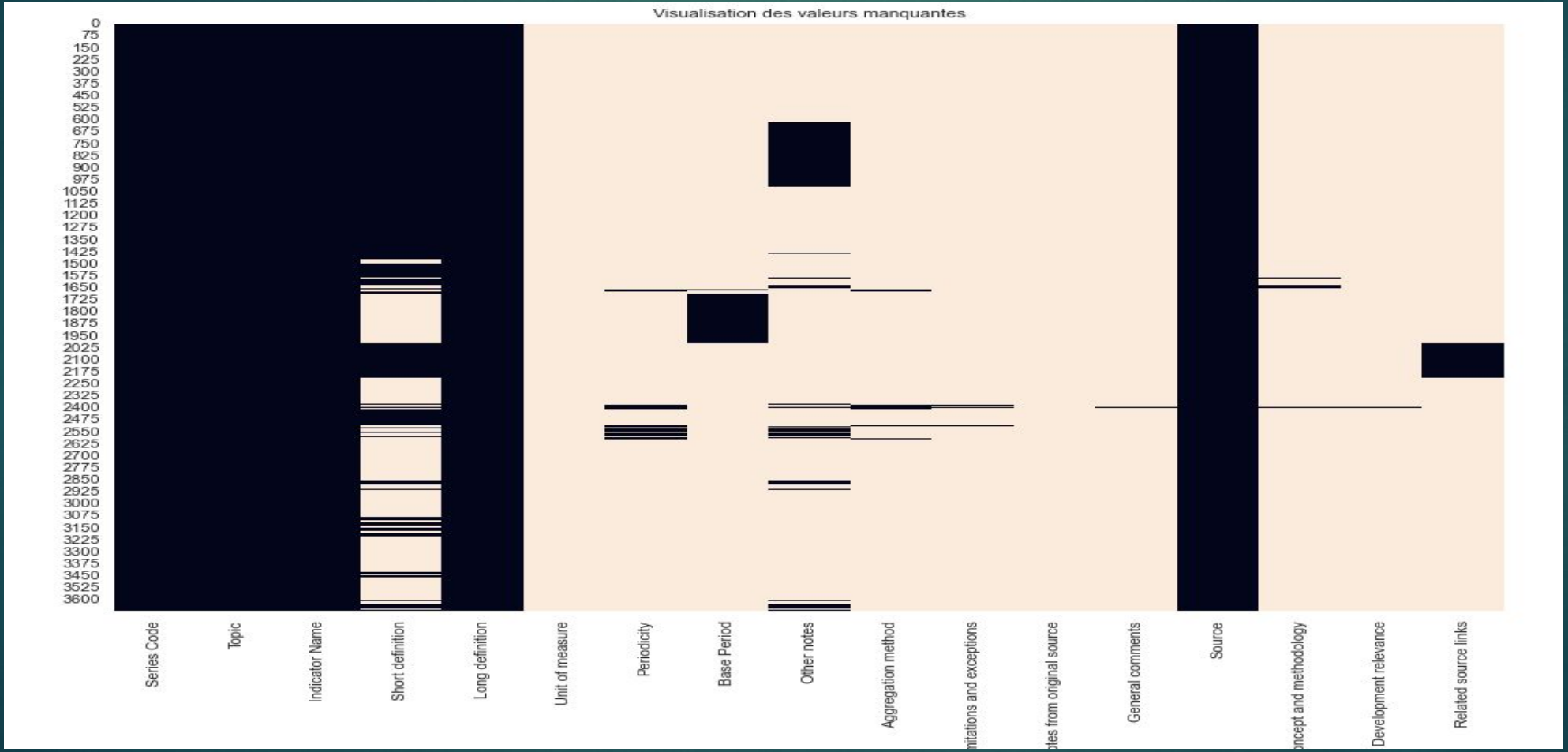
EdStatsSeries.csv

Le jeu de données permet de connaître les thèmes des indicateurs, les descriptions longues et les sources. Il donne des informations sur les indicateurs socio-éduco-économique classés en 37 thèmes.

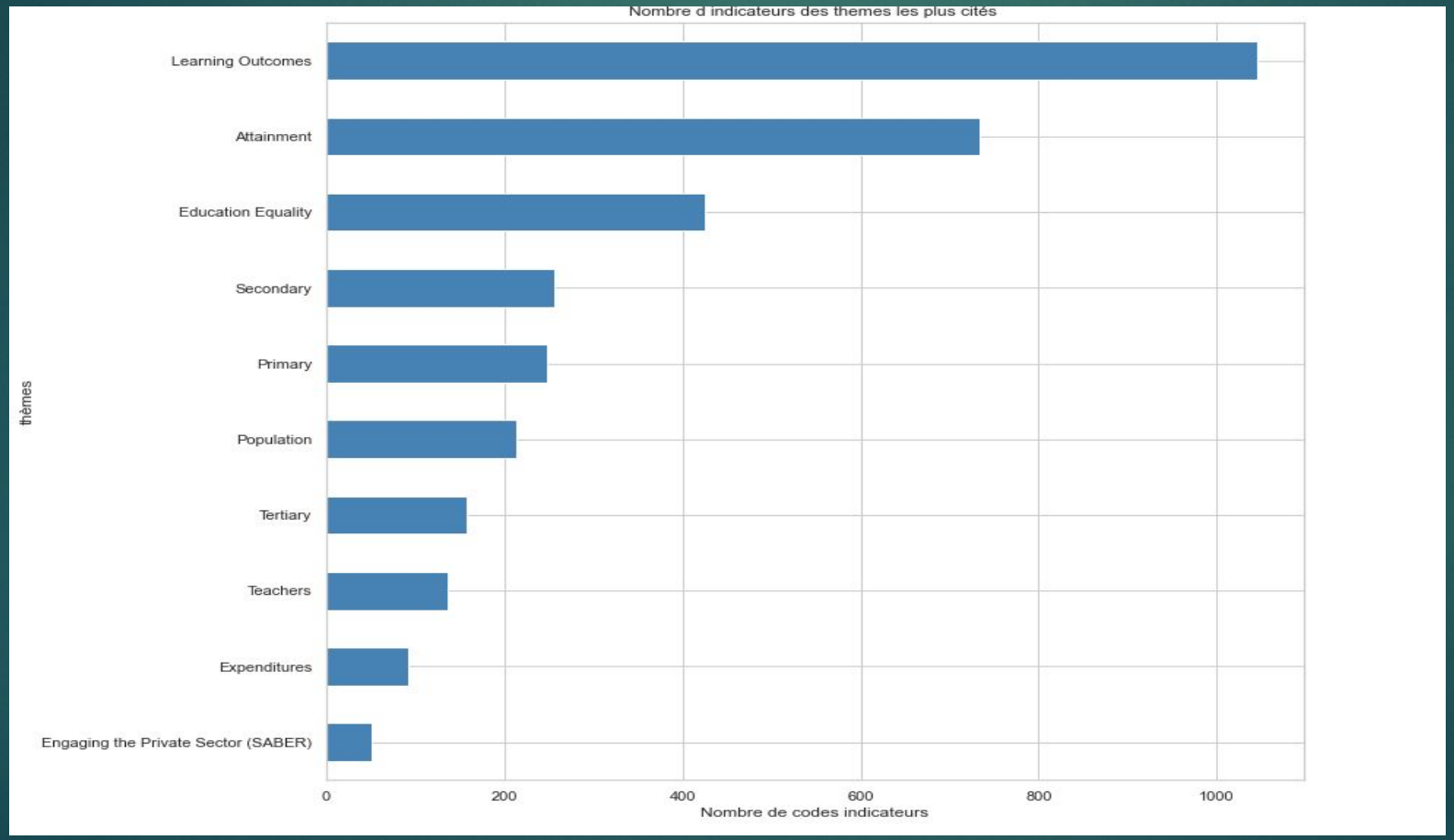
Taille : 3665 lignes, 17 colonnes
65.07% de cellules manquantes

Suppression des variables qui contiennent plus de 90%

Visualisation des valeurs manquantes



Nombre d'indicateur par thèmes



Présentation du jeu de données

EdStatsData.csv

Donne l'évolution de nombreux indicateurs pour tous les pays et certains groupes de pays

Taille : 886 930 lignes, 70 colonnes

Données depuis 1970 , Nombreuses valeurs manquantes
Aucun doublon

Préambule

3 665
indicateurs uniques

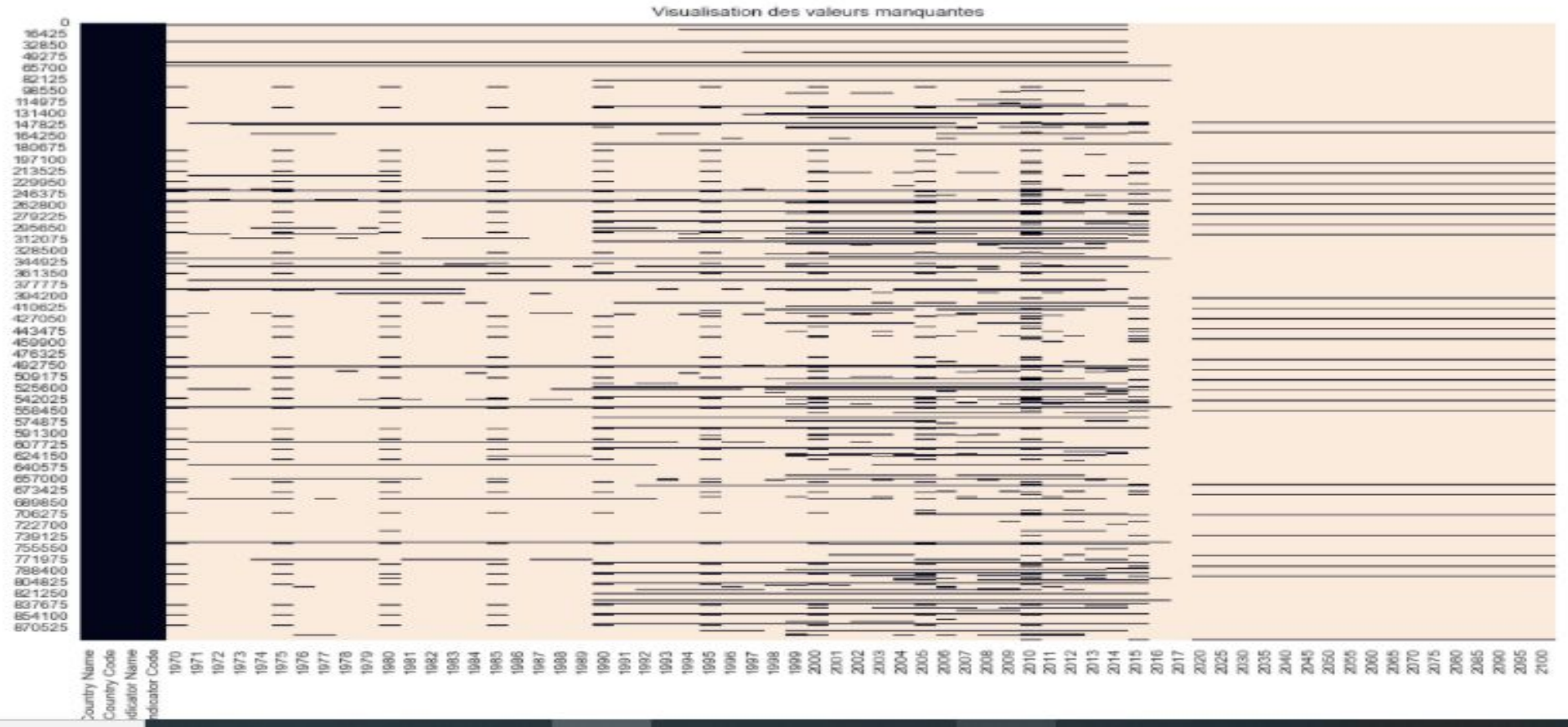
Historique et prédictions
de 1970 à 2100

52 568 249
Valeurs
Manquantes

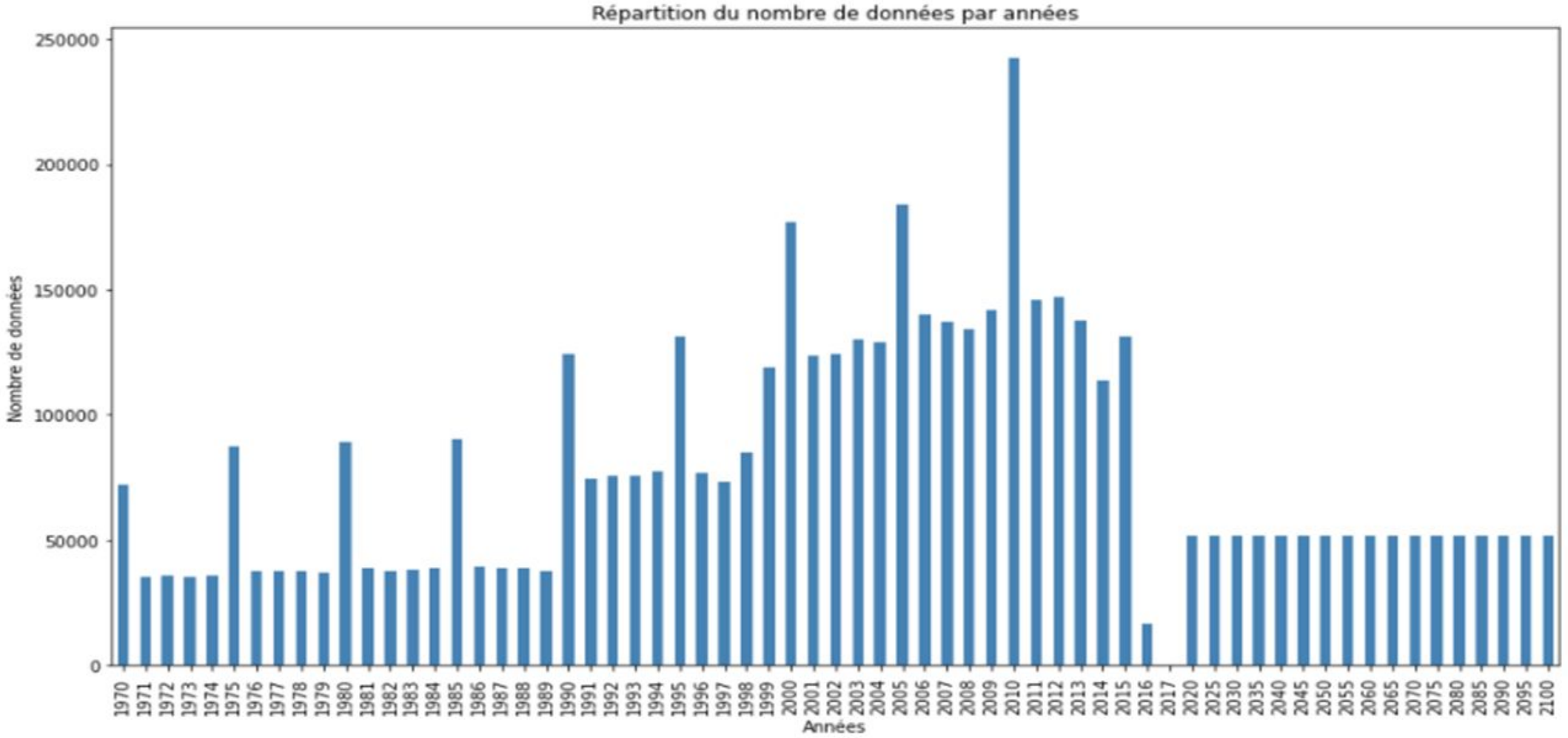
242 zones
Géographiques et pays

Visualisation des données manquantes

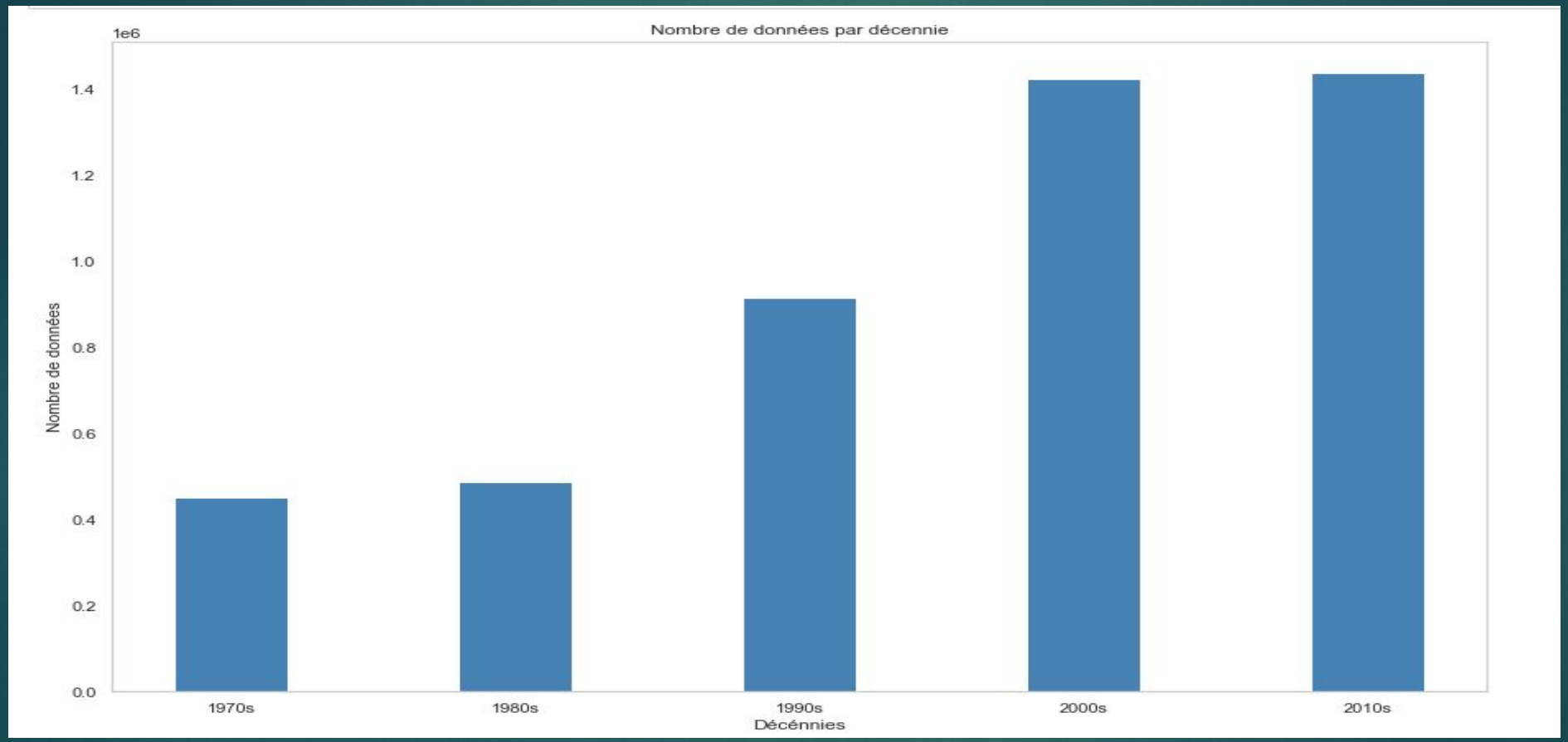

```
Out[339]: <AxesSubplot:title={'center':'Visualisation des valeurs manquantes'}>
```



Quantité de données par année



Nombre de données par décennie



II- **Analyser les données**

2.1 **Sélection des indicateurs**

Brainstorming

Notre Start Up cible

la population des 15-24 ans,

le nombre d étudiants scolarisés niveau lycée et université.

le niveau de vie des habitants comparable par pays

les moyens de communication accès à l'électricité, accès au réseau internet, possession d un ordinateur ou tablette personnel.

En parcourant le site de la banque mondiale:

Les différents groupes d'indicateurs à étudier sont :

SE : Social Education

SP : Social Population

IT : Infrastructure

NY : National Accounts, produits intérieurs et nationaux

Les différents mots clés à rechercher :

15 : pour la cible de la population des 15-19 ans

20 : pour la cible de la population des 20-24 ans

SEC : pour les regroupements par lycéens

TER : pour les regroupements par étudiants de l enseignement supérieur

IT : pour l'accès aux infrastructures techniques

Bilan sur les indicateurs pertinents

Démographique :

SP.POP.1524.TO.UN

Economique :

NY.GNP.PCAP.PP.CD

Educatif :

SE.SEC.ENRR

SE.TER.ENRR

Numérique :

IT.NET.USER.P2

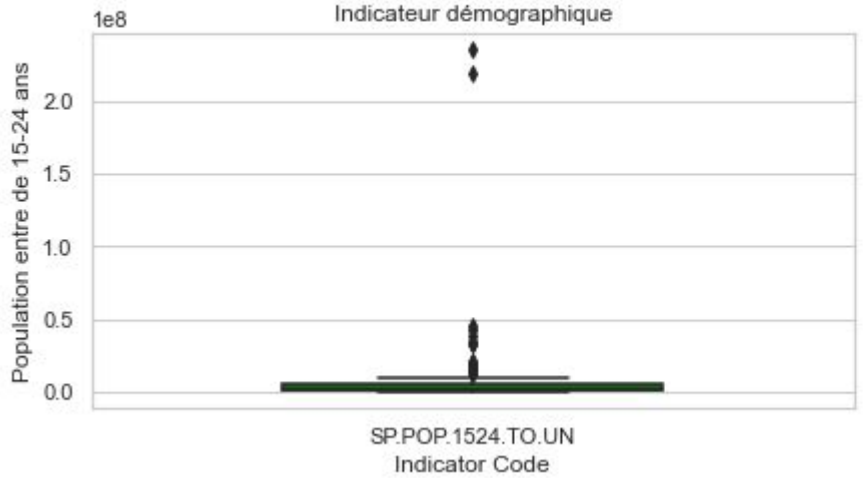
Application des filtres pour trouver des indicateurs pertinents

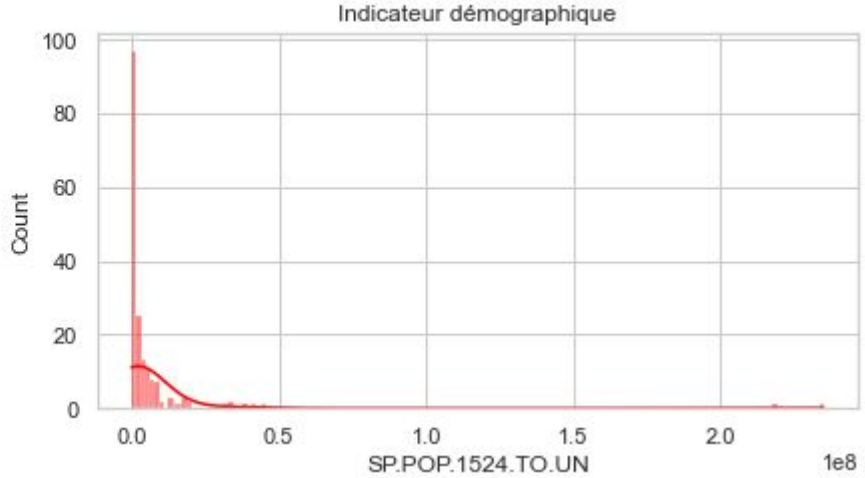
Code utilisé pour filtrer les indicateurs

```
#Application du filtre pour trouver un indicateur démographique
pd.set_option('max_colwidth', None)
mot_cle=[row for row in df_Ed_Ttl['Indicator Code'] if ('1524') in row]
df_Ed_Ttl[df_Ed_Ttl['Indicator Code'].isin(mot_cle)][['Indicator Code','Indicator Name']]
```

Indicateur démographique

```
Text(0.5, 0, 'SP.POP.1524.TO.UN')
```



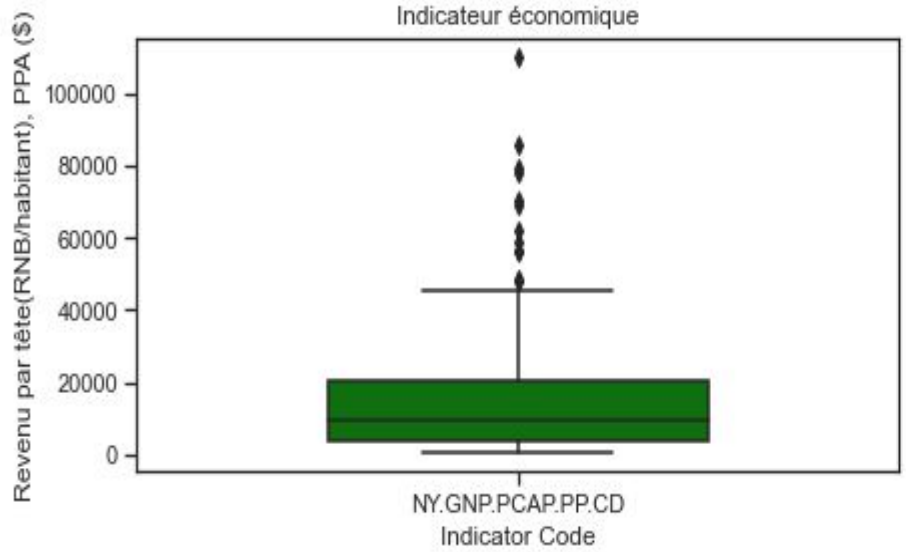


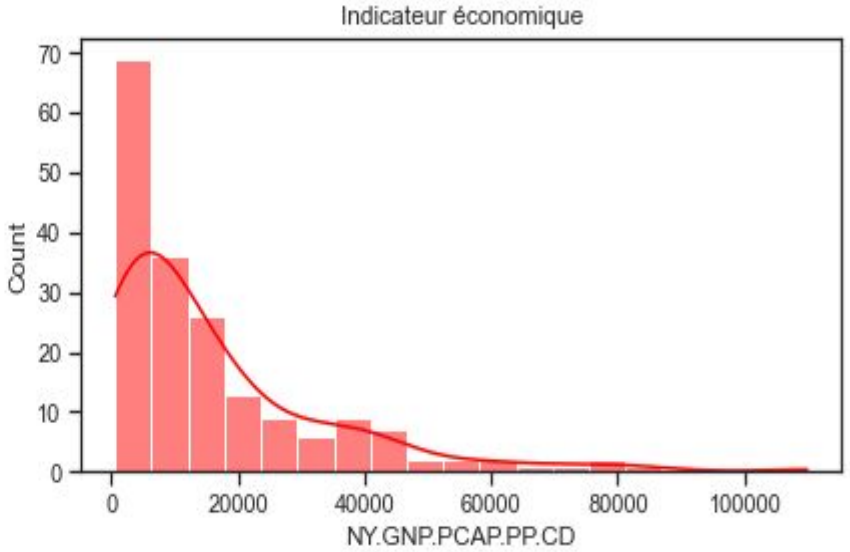
Remarques importantes:

- Très peu réparti
- Présence d'outliers
- Courbe asymétrique positive

Indicateur économique

```
Text(0.5, 0, 'NY.GNP.PCAP.PP.CD')
```



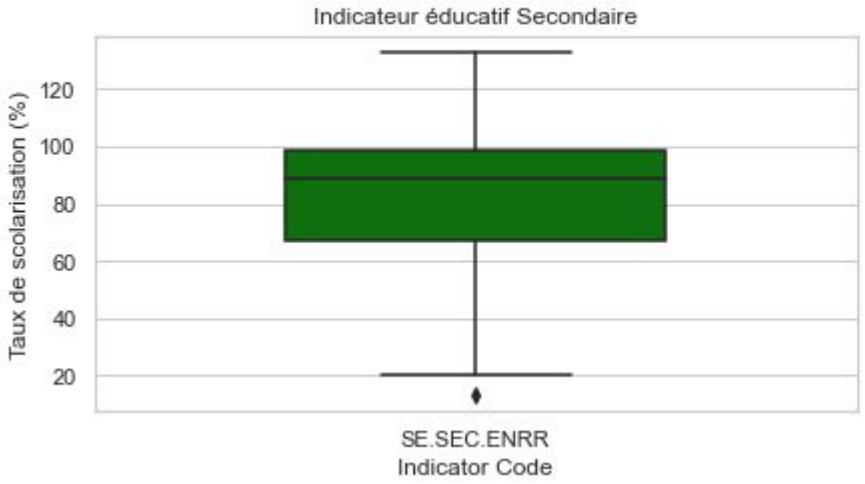


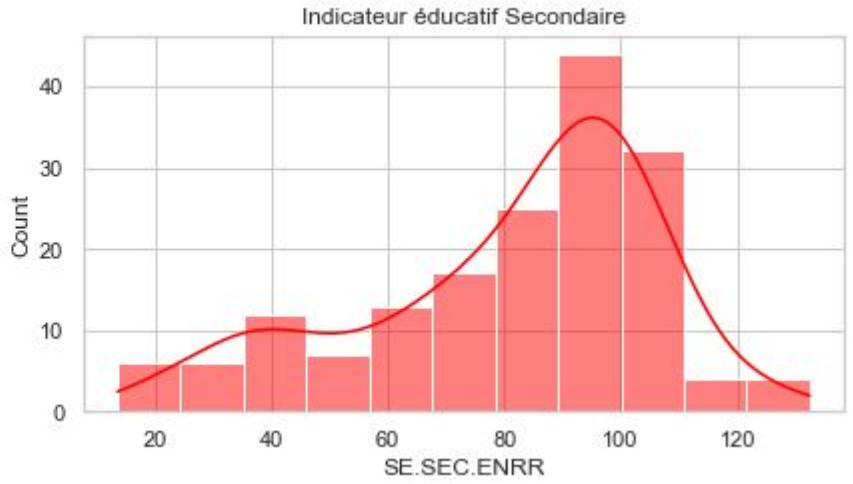
Remarques importantes::

- Très peu réparti
- Présence d'outliers
- Courbe asymétrique positive

Indicateur éducatif-Secondaire


```
Text(0.5, 0, 'SE.SEC.ENRR')
```



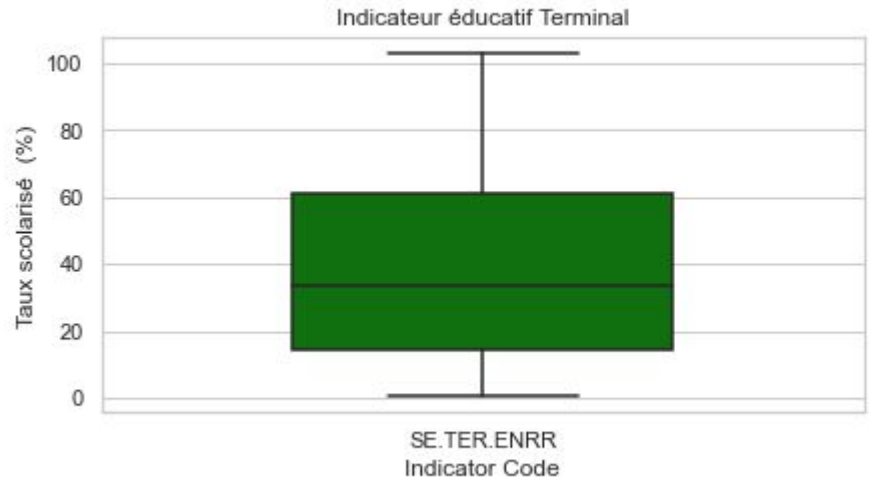


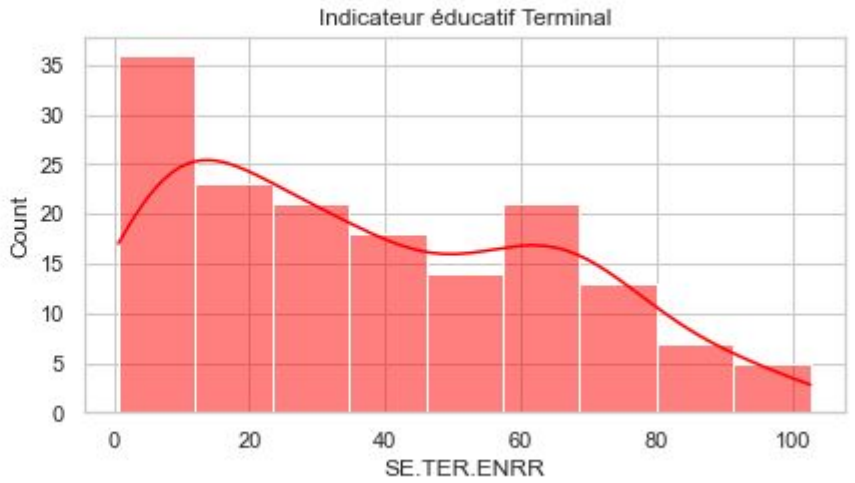
Remarques importantes:

- Bonne distribution
- Présence d'outlier
- Courbe asymétrique légèrement négative

Indicateur éducatif-Terminal


```
Text(0.5, 0, 'SE.TER.ENRR')
```



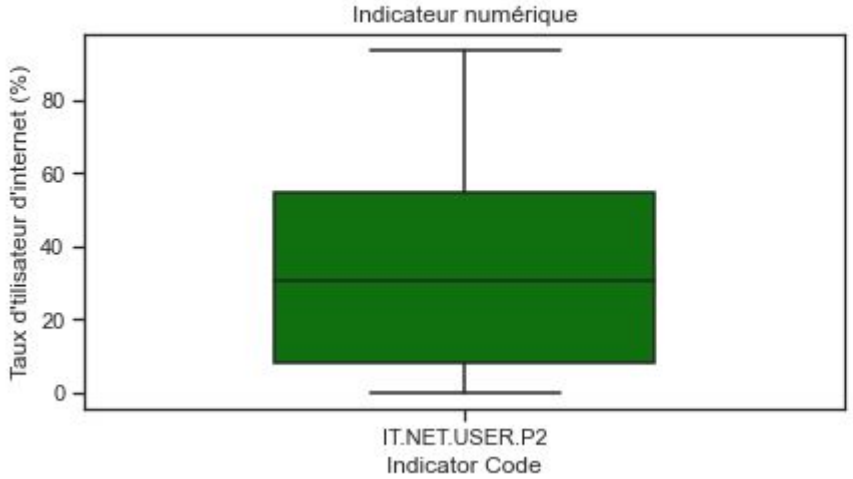


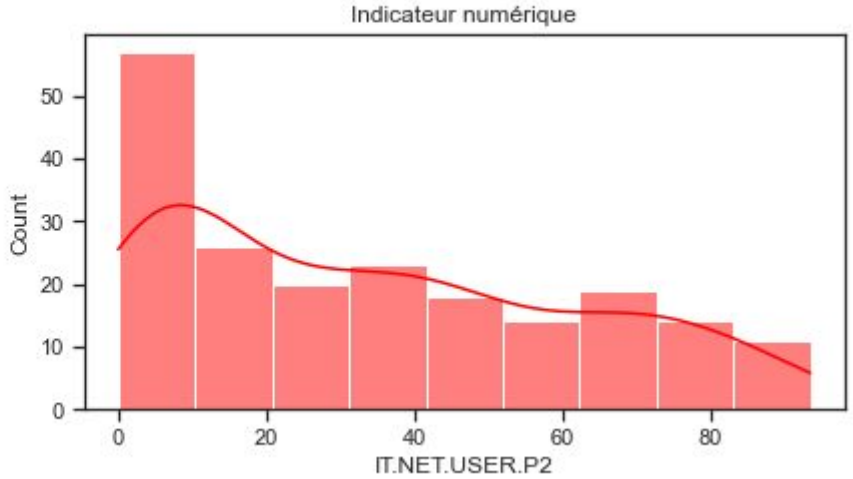
Remarques importantes:

- Une répartition non équilibrée
- Aucune présence d'outliers
- Courbe asymétrique positive bimodale

Indicateur numérique


```
Text(0.5, 0, 'IT.NET.USER.P2')
```





Remarques importantes:

- Une répartition non équilibrée
- Aucune présence d'outliers
- Courbe asymétrique positive multimodale

Statistiques descriptives des Indicateurs

	Descriptive	Stat_pop	Stat_eco	Stat_edu	Stat_eduS	Stat_tec
0	mean	6.685106e+06	1.636128e+04	82.048512	38.449436	34.259979
1	median	1.309299e+06	9.550000e+03	90.295853	36.018311	30.325000
2	var	5.992928e+14	3.373573e+08	657.207570	772.448783	748.295129
3	std	2.448046e+07	1.836729e+04	25.636060	27.792963	27.354984
4	skew	8.223737e+00	2.058465e+00	-0.860616	0.394172	0.482760
5	kurtosis	7.219921e+01	5.138663e+00	-0.011939	-0.993197	-1.004314

Pour l'indicateur démographique et économique : les courbes sont étalées à droite car le skewness est positif,elles sont moins aplaties que la distribution normale car le kurtosis empirique est positif et leur variance est élevée.
Pour l'indicateur éducatif niveau Secondaire: la courbe est multimodale, étalée à gauche car le skewness est négatif, elles sont plus aplatie que la distribution normale car le kurtosis empirique est négatif.
Pour l'indicateur éducatif Terminal : la courbe est bimodale et étalée à droite car le skewness est positif, elles sont plus aplatie que la distribution normale car le kurtosis empirique négatif.
Pour l'indicateur numérique : la courbe est multimodale et étalée à gauche car le skewness est négatif,elles sont plus aplatie que la distribution normale car le kurtosis empirique négatif.

Standardisation des données


```
# Normalisation de la data
z=pd.DataFrame(df_code_selection)
z_norm=stats.zscore(z)
z_norm
```

					2010
Indicator Code	Tec	Eco	Edu Sec	Edu Ter	Pop
Country Name					
Afghanistan	-1.079439	-0.786288	-0.230216	-0.921850	-0.013125
Albania	0.426232	-0.355230	0.620313	0.607037	-0.240966
Algeria	-0.767288	-0.191146	0.832198	0.102241	0.034902
Angola	-1.123507	-0.590254	-0.821962	-0.921850	-0.109354
Antigua and Barbuda	0.499679	0.119151	1.050722	-0.364678	-0.266255
...	...	...	...	...	...
Vietnam	-0.100753	-0.643324	-1.518094	-0.143081	0.499403
West Bank and Gaza	0.147132	-0.872392	0.553068	0.720755	-0.229707
Yemen, Rep.	-0.772796	-0.648739	-0.468995	-0.558909	-0.043516
Zambia	-0.859097	-0.709932	-1.518094	-0.921850	-0.153187
Zimbabwe	-0.991302	-0.805242	-1.518094	-0.719137	-0.125396

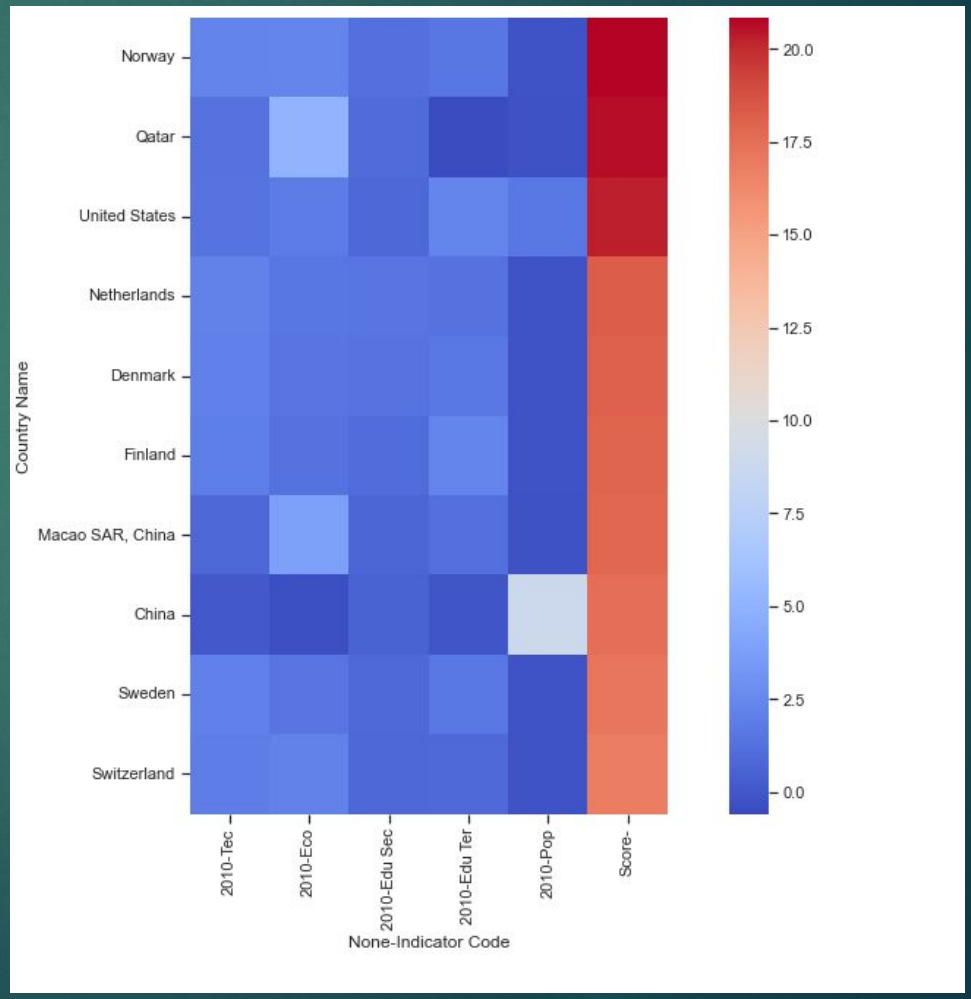
188 rows × 5 columns

II- Classement des pays potentiels

Score par pays


```
df_top10=z_norm.sort_values(by='Score',ascending=False).head(10)
df_top10
```

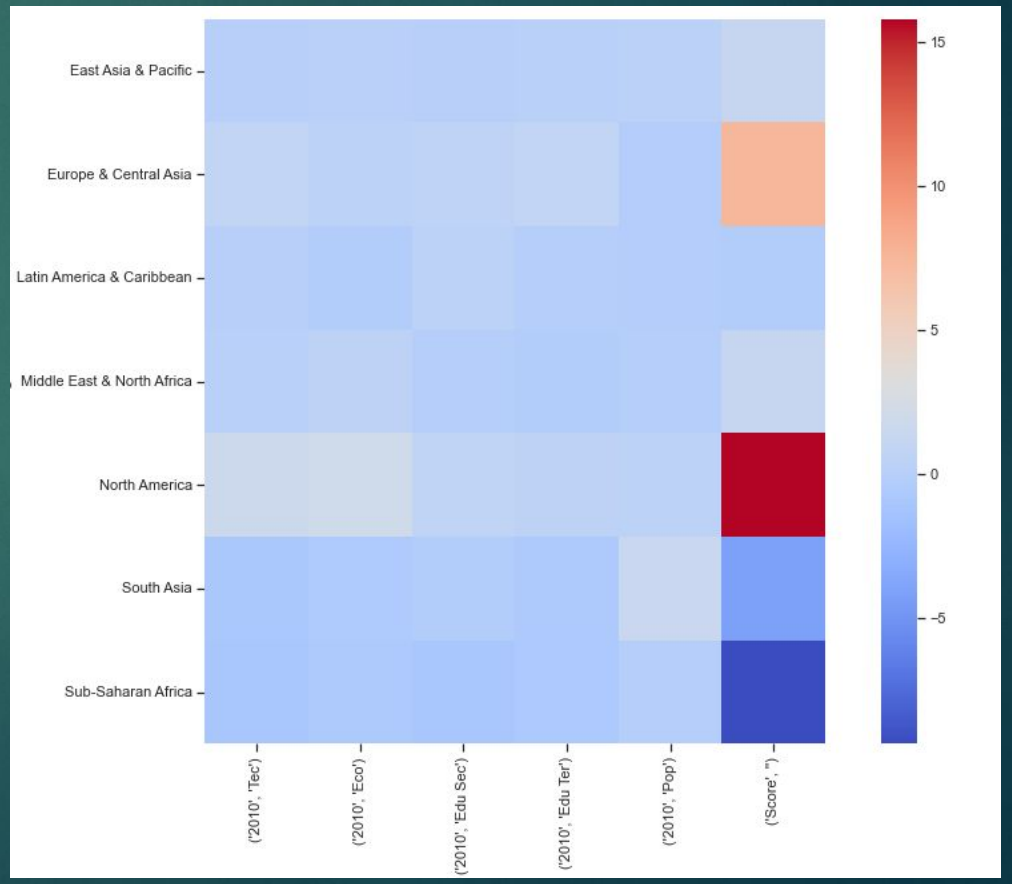
					2010	Score
Indicator Code	Tec	Eco	Edu Sec	Edu Ter	Pop	
Country Name						
Norway	2.203291	2.302061	1.220607	1.580041	-0.240243	20.840158
Qatar	1.307600	5.080655	0.928321	-0.583850	-0.261364	20.638582
United States	1.406387	1.774611	0.763008	2.312588	1.609788	20.320149
Netherlands	2.105239	1.561789	1.460087	1.314766	-0.183249	18.289529
Denmark	2.031791	1.498971	1.373924	1.605193	-0.238390	18.105533
Finland	1.964587	1.257990	1.076502	2.308851	-0.238980	17.925064
Macao SAR, China	0.800740	3.773400	0.750441	1.199066	-0.262949	17.896276
China	0.033288	-0.369310	0.535889	-0.099822	8.835013	17.567385
Sweden	2.078798	1.450234	0.856838	1.641732	-0.215147	17.232736
Switzerland	1.854783	2.167762	0.800502	0.890295	-0.226606	16.850801



Score par région


```
df_score_region.sort_values(by=('Score', ''),ascending=False).head(10)
```

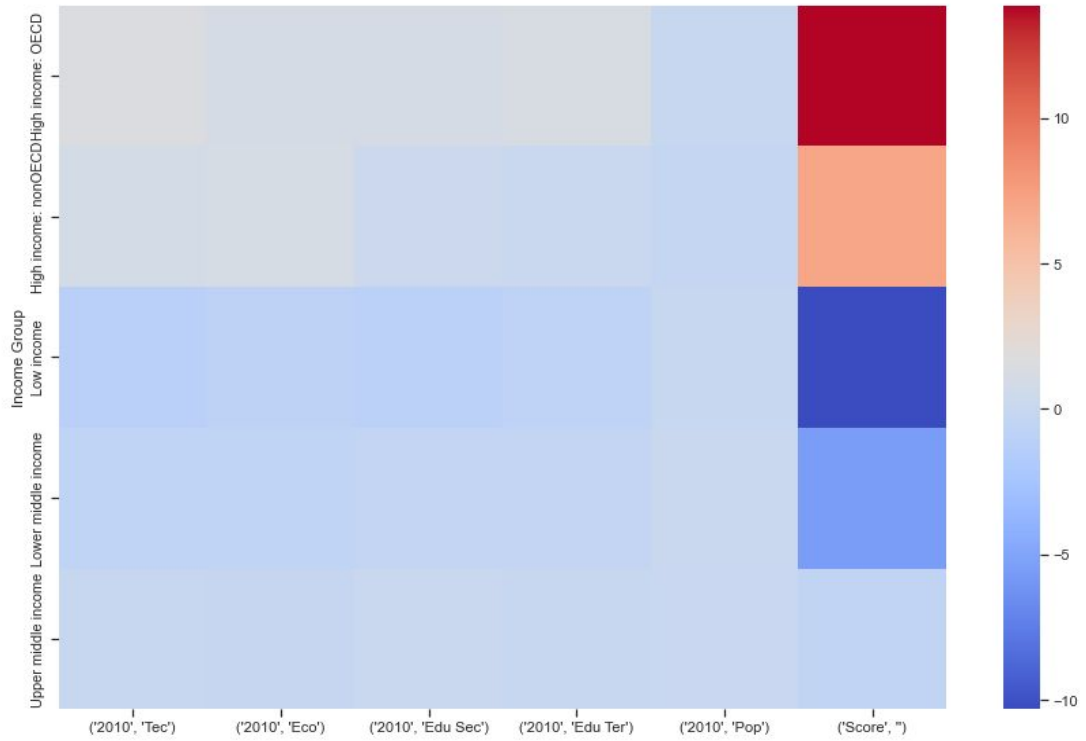
	(2010, Tec)	(2010, Eco)	(2010, Edu Sec)	(2010, Edu Ter)	(2010, Pop)	(Score,)	Region
630	2.203291	2.302061	1.220607	1.580041	-0.240243	20.840158	Europe & Central Asia
685	1.307600	5.080655	0.928321	-0.583850	-0.261364	20.638582	Middle East & North Africa
890	1.406387	1.774611	0.763008	2.312588	1.609788	20.320149	North America
605	2.105239	1.561789	1.460087	1.314766	-0.183249	18.289529	Europe & Central Asia
235	2.031791	1.498971	1.373924	1.605193	-0.238390	18.105533	Europe & Central Asia
295	1.964587	1.257990	1.076502	2.308851	-0.238980	17.925064	Europe & Central Asia
505	0.800740	3.773400	0.750441	1.199066	-0.262949	17.896276	East Asia & Pacific
180	0.033288	-0.369310	0.535889	-0.099822	8.835013	17.567385	East Asia & Pacific
805	2.078798	1.450234	0.856838	1.641732	-0.215147	17.232736	Europe & Central Asia
810	1.854783	2.167762	0.800502	0.890295	-0.226606	16.850801	Europe & Central Asia



Score par revenu


```
# Constitution du dataframe de visualisation du regroupement par Income Group
df_score_income_Group = df_score_region.groupby('Income Group').mean()
```

```
fig = plt.figure(1,figsize=(15, 10))
sns.heatmap(df_score_income_Group, square=True, cmap='coolwarm', fmt="d", )
plt.show()
```



Pays , Groupe, région à fort potentiel

Country Name	Region	Income Group
United States	North America	High income: OECD
Sweden	Europe & Central Asia	High income: OECD
Qatar	Middle East & North Africa	High income: nonOECD
Norway	Europe & Central Asia	High income: OECD
Macao SAR, China	East Asia & Pacific	High income: nonOECD
Finland	Europe & Central Asia	High income: OECD
China	East Asia & Pacific	Upper middle income

Conclusion

Les jeux de données d'une manière générale sont fiables pour mener un comparatif du fait qu'ils possèdent beaucoup de données importantes sur les pays et les régions, entre autres des facteurs économiques, démographiques et technologiques. Néanmoins ils restent difficiles à modeler.

Quels sont les pays avec un fort potentiel de clients pour nos services ?

United States, Switzerland, Sweden, Qatar, Norway, Netherlands, Finland, Denmark, China

Pour chacun de ces pays, quelle sera l'évolution de ce potentiel de clients ?

La trajectoire de ces pays peut être semblable au niveau de la globalité des indicateurs, qui influencent l'évolution de l'éducation mais avec un grand avantage pour les USA et la chine dû à la population (du point de vu business qui se caractérise sur le marché par le nombre des clients).

Dans quels pays l'entreprise doit-elle opérer en priorité ?

En se basant sur le score par revenu et par région , la priorité est donnée aux pays suivants:

United States,Sweden, Qatar, Norway, Finland, China

Merci de votre attention