Analyse des données de systèmes éducatifs pour l'entreprise "academy"

Openclassrooms
Formation Data Scientist
Projet 2

Serge DAVISTER
01/2023

academy

Academy est une Start-up de la **EdTech** qui propose des contenus de formation en ligne pour un public de niveau lycée et université.

Dans le cadre d'un projet d'expansion à l'international de l'entreprise, il y a lieu d'analyser les données de la banque mondiale sur l'éducation afin de déterminer :

- Quels seraient les pays avec un fort potentiel de clients pour nos services ?
- Pour chacun de ces pays, quelle serait l'évolution de ce potentiel de clients ?
- Dans quels pays l'entreprise devrait-elle opérer en priorité ?

academy

Pour la pré-analyse, :

Valider la qualité de ce jeu de données (comporte-t-il beaucoup de données manquantes, dupliquées ?)

Décrire les informations contenues dans le jeu de données (nombre de colonnes ? nombre de lignes ?)

Sélectionner les informations qui semblent pertinentes pour répondre à la problématique (quelles sont les colonnes contenant des informations qui peuvent être utiles pour répondre à la problématique de l'entreprise ?)

Déterminer des ordres de grandeurs des indicateurs statistiques classiques pour les différentes zones géographiques et pays du monde (moyenne/médiane/écart-type par pays et par continent ou bloc géographique)

academy

Pré-analyse des jeux de données

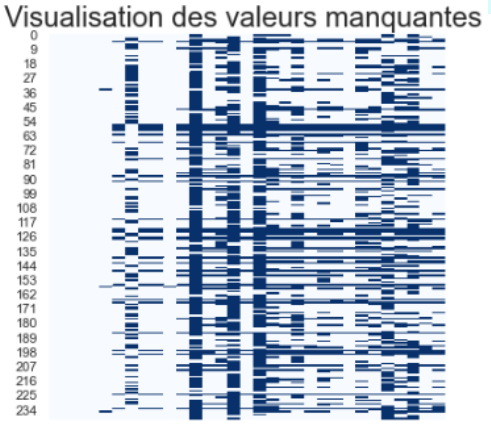
Fichier EdStatsCountry.csv

academy

Le jeu de donnees contient des informations sur :
les codes ISO-3166 alpha 2
les codes WB-2
les noms des **214 pays**
27 répartitions par région,par niveau de revenus,groupement économique,...
151 devises
5 catégories de niveau de revenu
des données économiques avec les dates des dernières études effectuées

la colonne 31 Unnamed qui ne contenait que des nan a été supprimée.

Le jeu de données restant contient :




```
nombre de lignes : 241                nombre de colonnes : 31
nombre de valeurs manquantes : 2113   % de valeurs manquantes : 28.28
taille totale du jeu de donnees : 7471
```

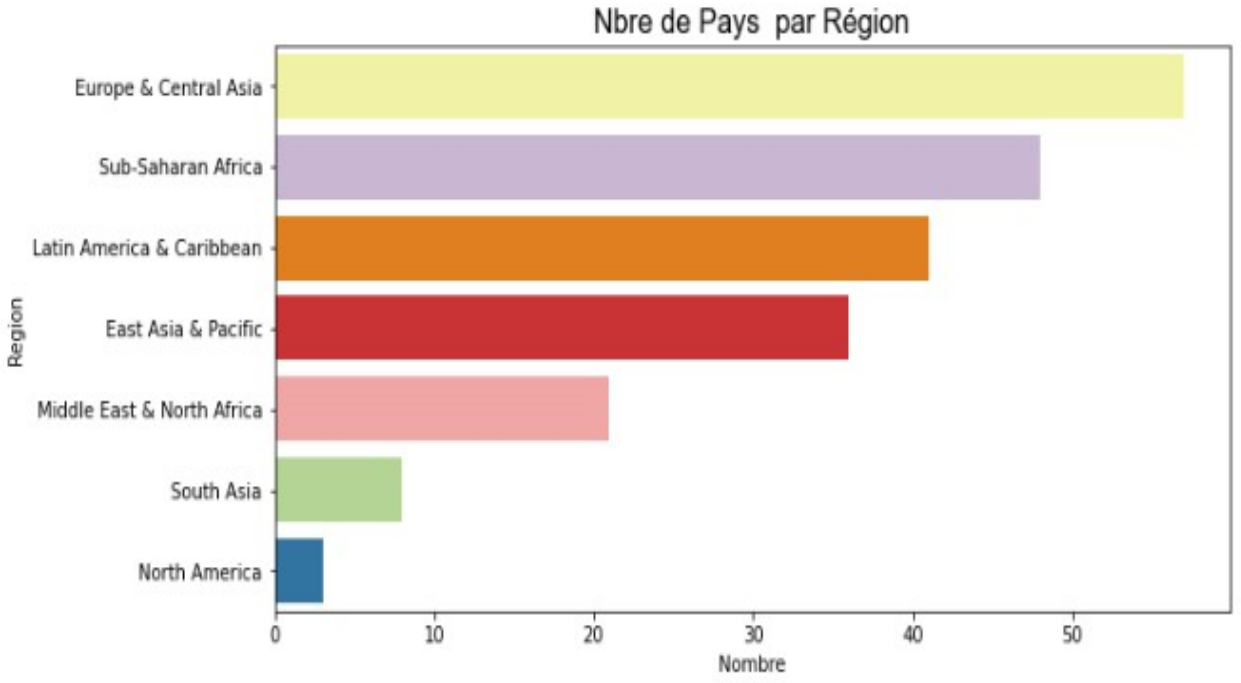
```
Variables
---------
```

	Nombre	%
object	28	90.32
float64	3	9.68

```
Nombre de doublons dans le dataset : 0
Nombre de doublons sur le Country Code : 0
```

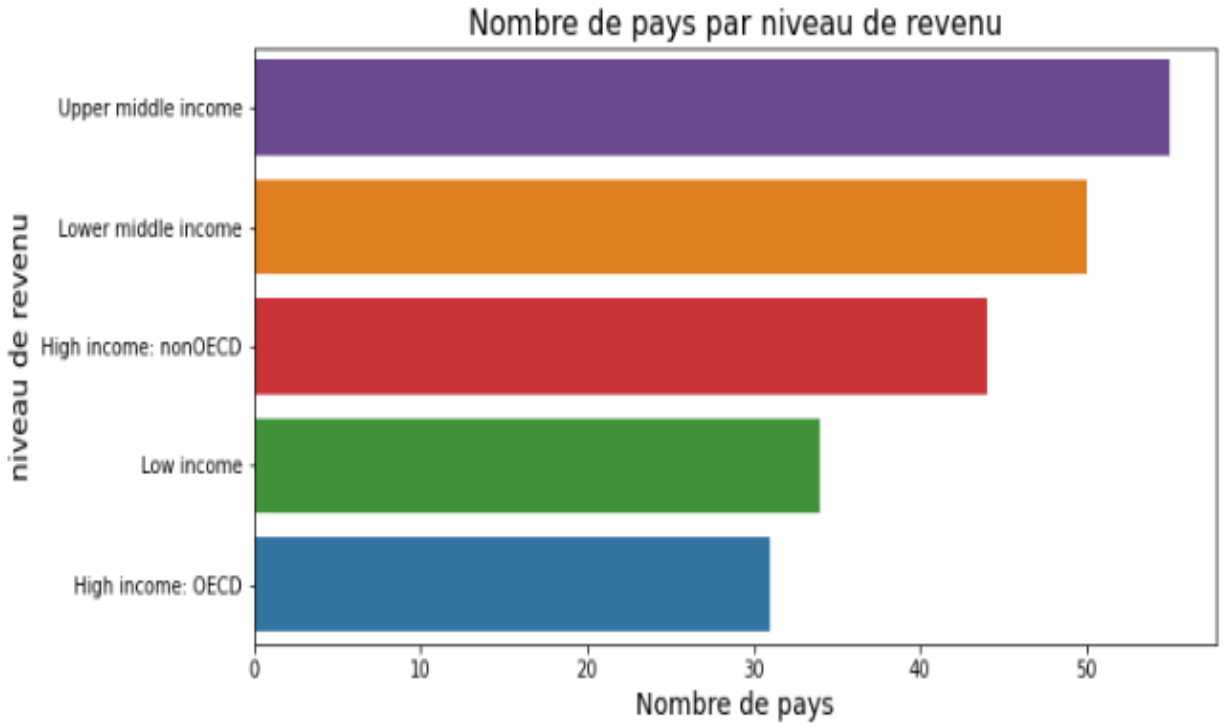
Répartition par région

academy



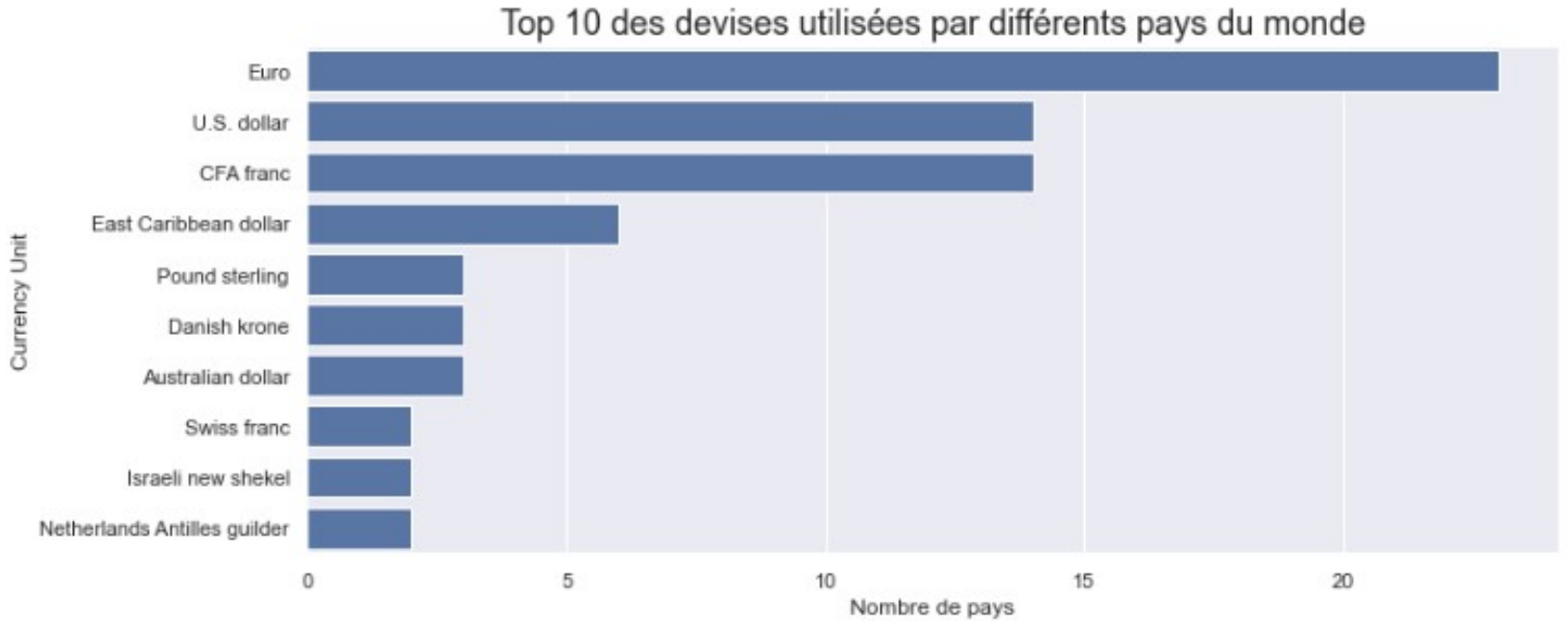
répartition par Income Group

academy



academy

Répartition par devise



Fichier EdStatsCountry-Series.csv

academy

Le jeu de donnees contient des informations sur :

```
211 countryCode correspondant à 211 pays
21 SeriesCode
97 DESCRIPTION qui indiquent la source des donnees
```

la colonne Unnamed :3 qui ne contenait que des nan a été supprimée.

Le jeu de données restant contient :

```
nombre de lignes : 613                          nombre de colonnes : 3
nombre de valeurs manquantes : 0          % de valeurs manquantes : 0.0
taille totale du jeu de donnees : 1839
```

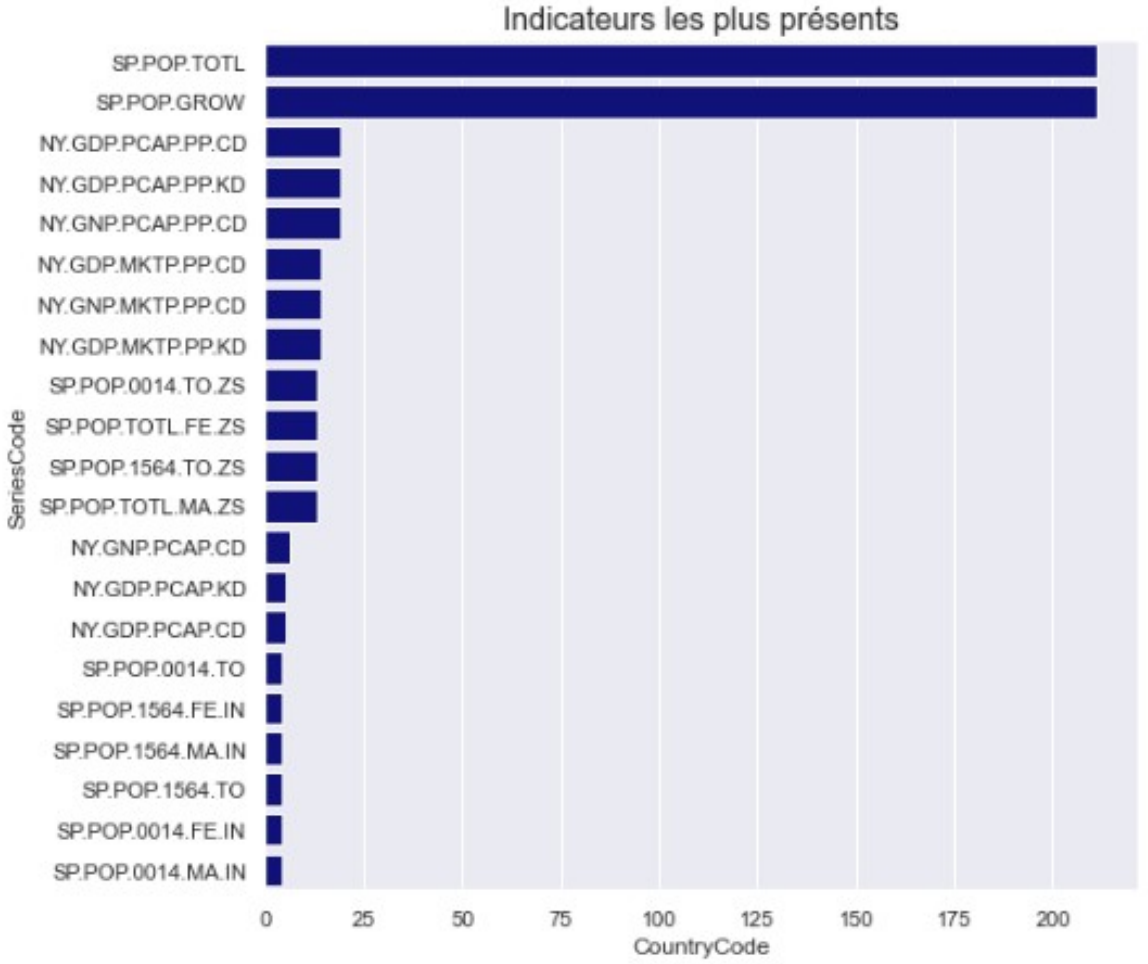
```
Variables
---------
```

	Nombre	%
object	3	100.0

```
Nombre de doublons dans le dataset : 0
```

Les indicateurs qui apparaissent le plus

academy



fichier EdStatsData.csv

academy

Le jeu de donnees contient des informations sur :

```
242 countryCode et CountryName dont
     211 pour des pays repris dans le dataset CountrySerie.
     31 concernent des regroupements écomnomiques,géographiques et des pays ou partie de pays.
3665 IndicatorCode et IndicatorName
65 colonnes contenant des données par année sur une période de 1970 à 2100
```

la colonne Unnamed :69 qui ne contenait que des nan a été supprimée.

Le jeu de données restant contient :

```
taille totale du jeu de donnees : 61198170
nombre de lignes : 886930                          nombre de colonnes : 69
nombre de valeurs manquantes : 52568249          % de valeurs manquantes : 85.9
```

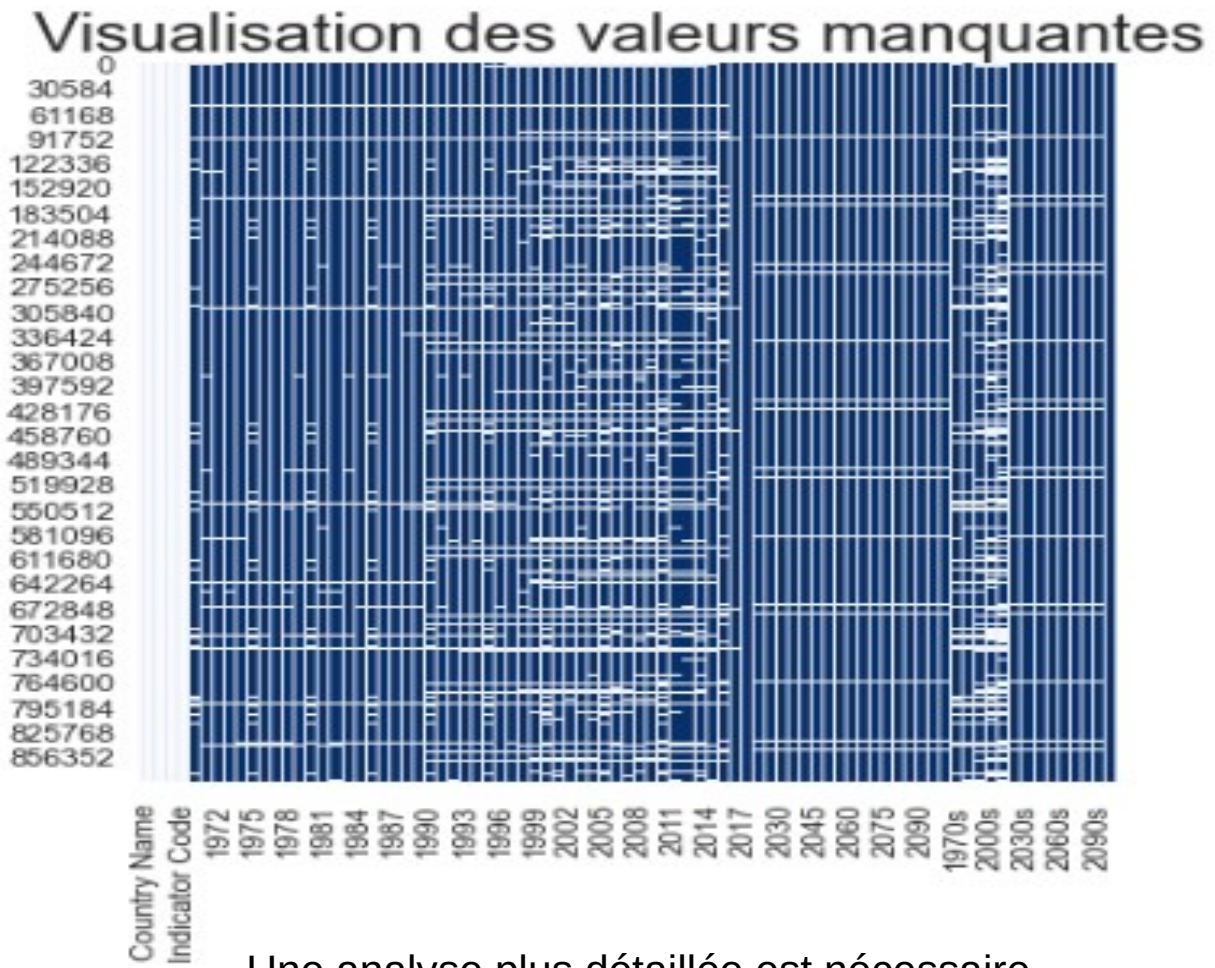
Il y a des années qui sont peu ou pas du tout documentées .Les années les plus riches en données sont les années entre 2000 et 2015

```
Variables
---------
```

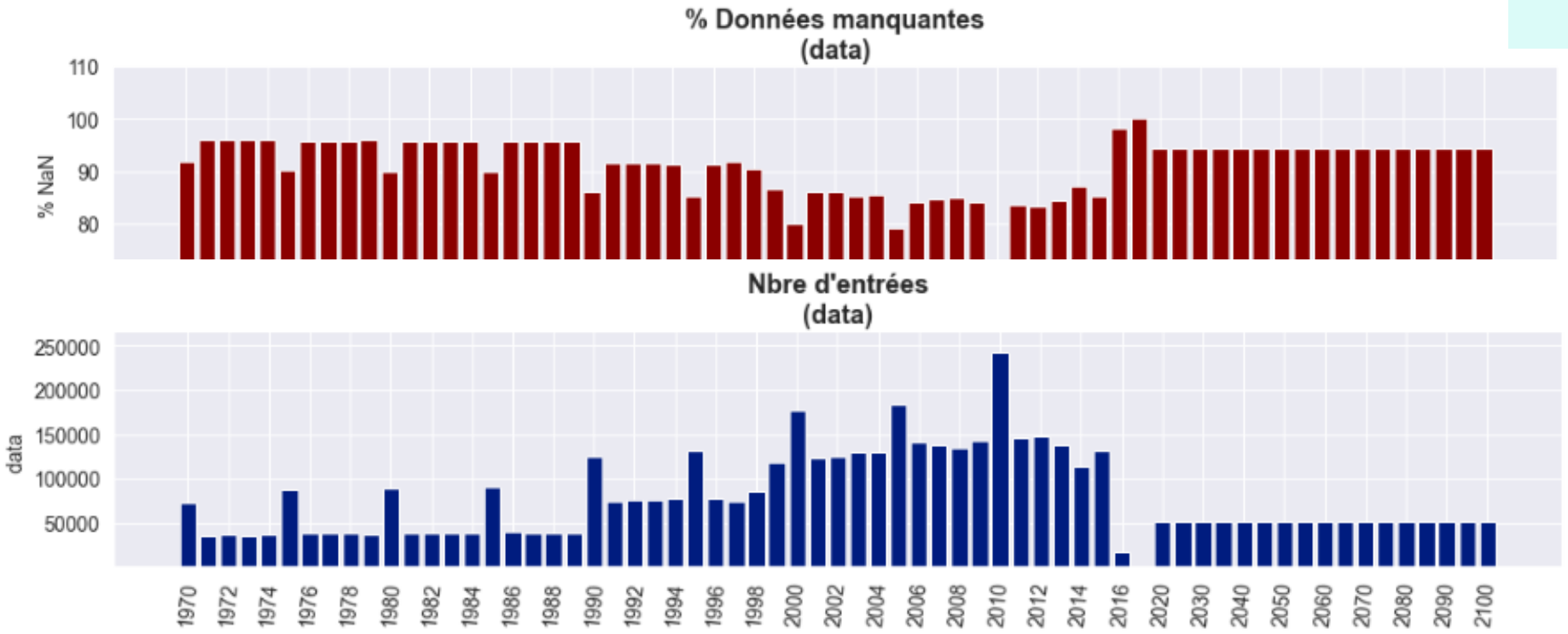
```
Nombre de doublons dans le dataset : 0
```

	Nombre	%
float64	65	94.2
object	4	5.8

academy



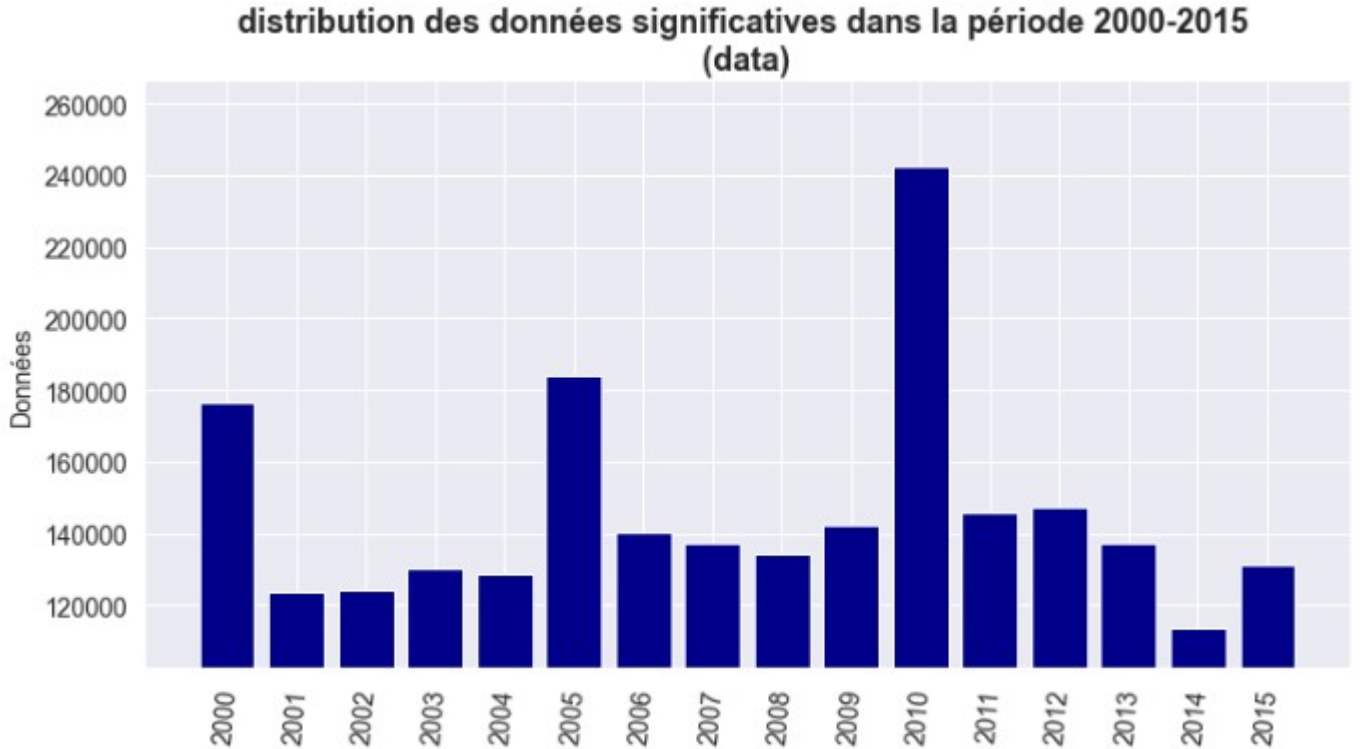
Une analyse plus détaillée est nécessaire

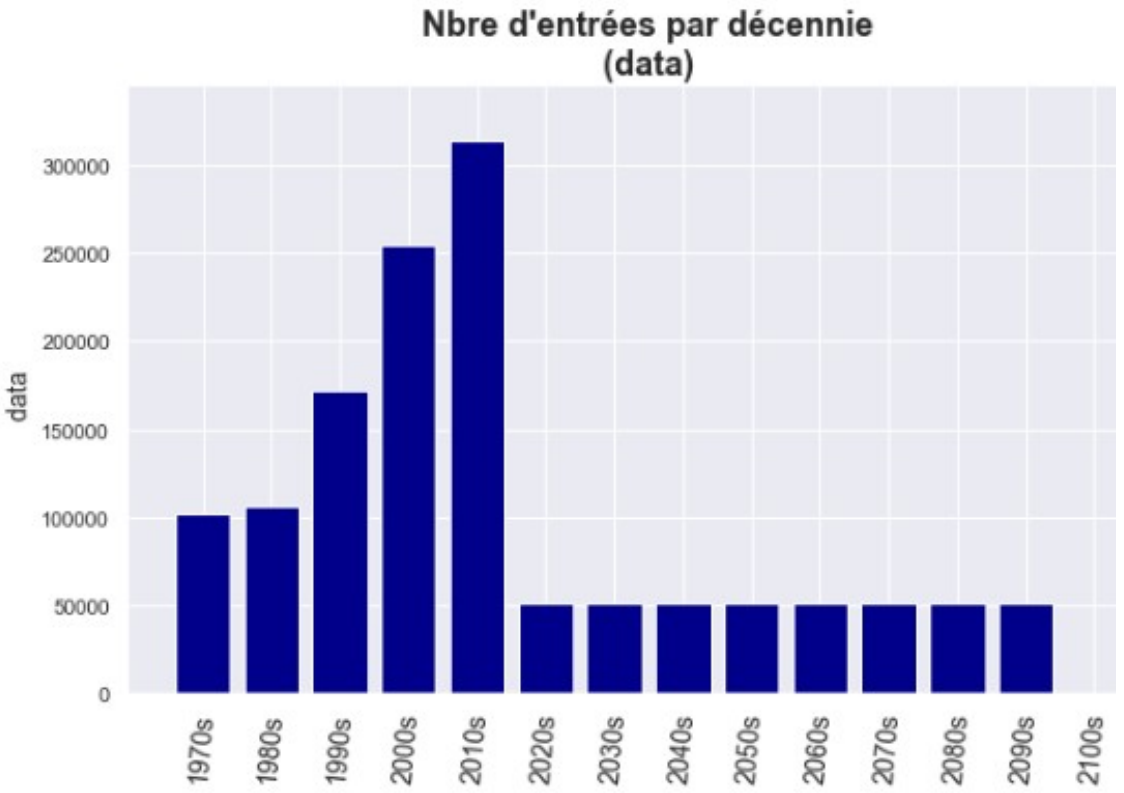


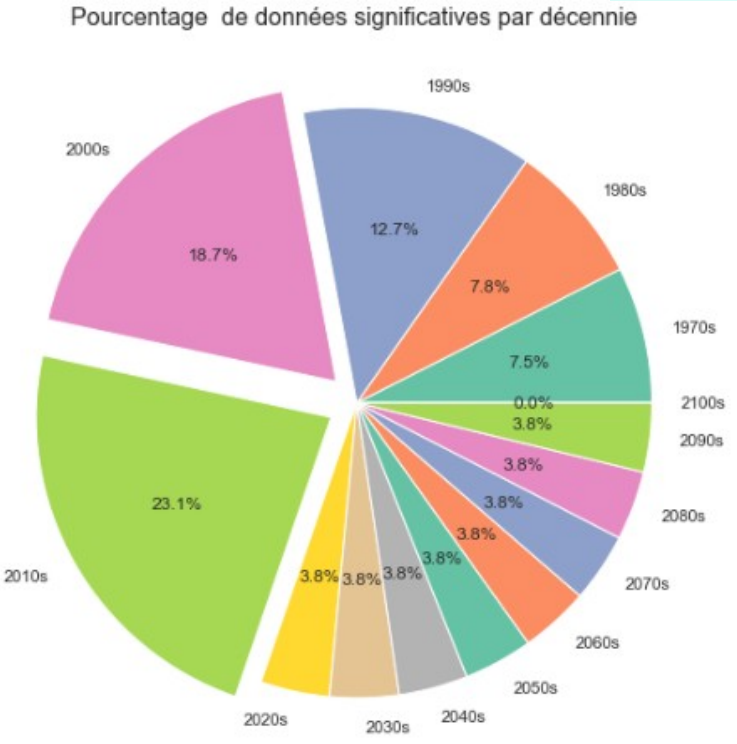
il y a des pics dans les quantités de données tout les 5 ans jusqu'en 2010
la quantité de données augmente à partie de 1990
l'augmentation la plus marquée commence en 2000 et atteint son maximum en 2010
2016 : très pêu de données (année de la collecte des data ?)
2017 à 2019 : pas de données
ensuite de 2020 à 2100 par intervalle de 5 ans , il y a peu de données
On peut regarder plus en détail la partie 2000 à 2020

academy

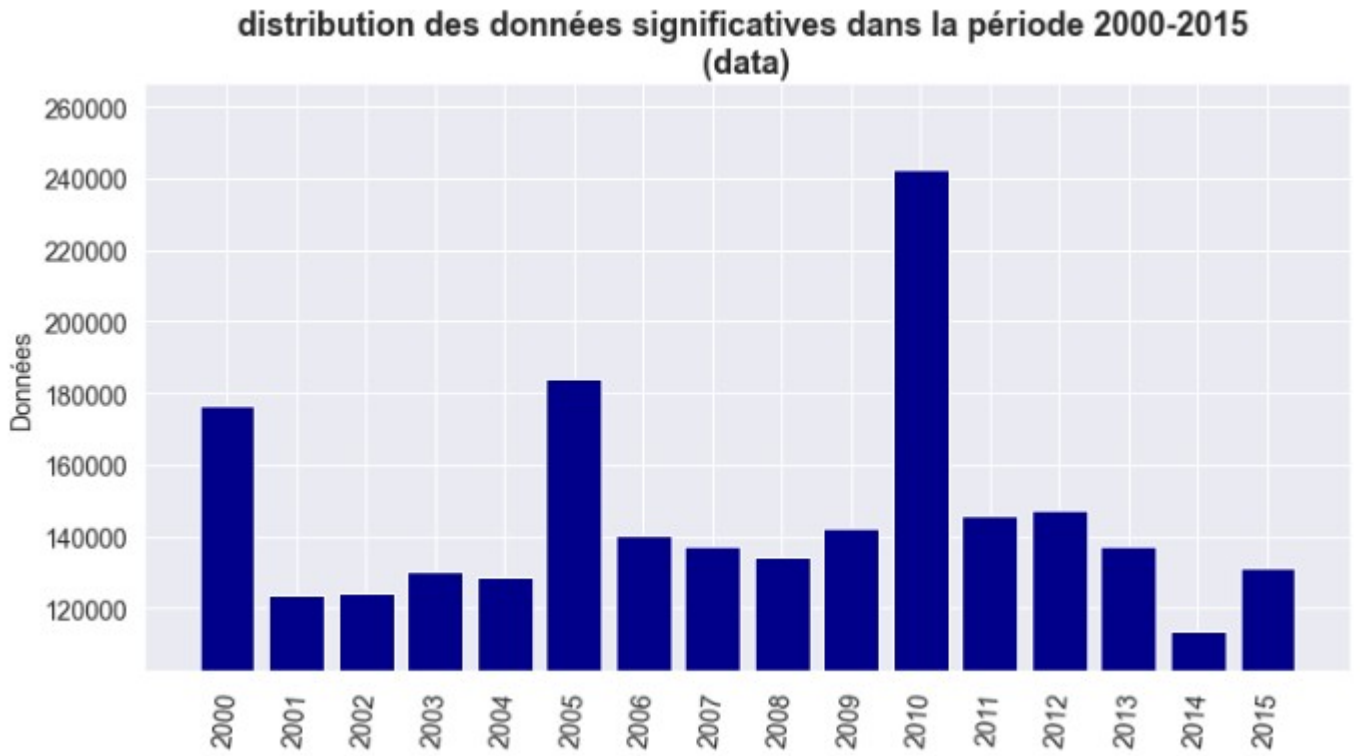
Les années les plus documentées







l'analyse des données par décennie confirme ce que nous avions constaté dans la répartition par année . Les décennies 2000s et 2010s contiennent le plus de données. On peut évaluer le pourcentage.



Valeurs statistiques sur la quantité de données pour la période 2000-2015

---

moyenne : 146239
écart type : 31419
médiane : 137390
valeur maximale : 242442
valeur minimale : 113789

Fichier EdStatsFootNote.csv

academy

Le jeu de donnees contient des enregistrements sur :

```
239 countryCode avec 210 pour des pays repris dans le dataset CountrySerie.
                     29 concernent des regroupements écomnomiques,géographiques ou autres .
1558 SeriesCode
1 colonne Year contenant une année sur une période de 1970 à 2050
1 colonne Description comportant 9102 DESCRIPTION
```

la colonne Unnamed :69 qui ne contenait que des nan a été supprimée.

Le jeu de données restant contient :

```
taille totale du jeu de donnees : 2574552
nombre de lignes : 643638                    nombre de colonnes : 4
nombre de valeurs manquantes : 0             % de valeurs manquantes : 0.0
```

```
Variables                    Nombre de doublons dans le dataset : 0
---------
```

	Nombre	%
object	4	100.0

Fichier EdStatsSeries.csv

academy

Le jeu de donnees contient des informations sur :

```
3665 indicateurs (Series code et Indicator Name ) répartis en
37 Topic (catégories portant sur la santé , l' éducation, l' économie ,les communications, la protection sociale...
Certain codes ont une définition (Short Definition et long definition)
Les autres colonnes traitent principalement de la collecte de données source,fréquence... sont peu intéressantes.
```

6 colonnes avec uniquement des nan ont été supprimées :
5 Unit of measure 11 Notes from original source 17 Other web links
18 Related indicators
19 License Type
20 Unnamed: 20
Le jeu de données restant contient :

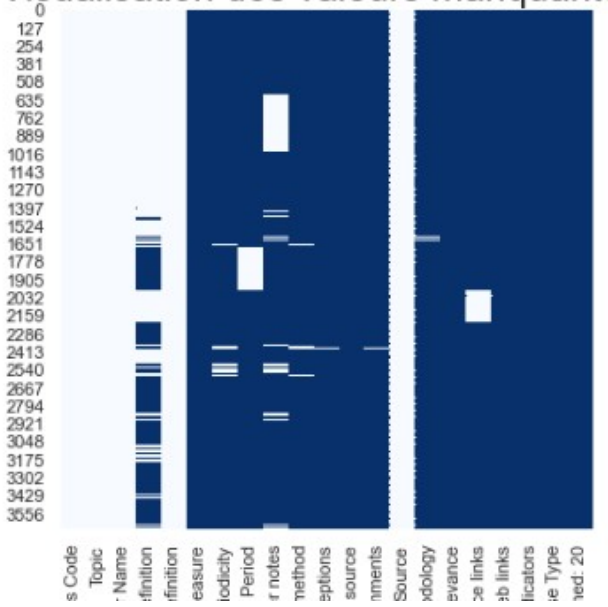
```
taille totale du jeu de donnees : 54975
nombre de lignes : 3665                          nombre de colonnes : 15
nombre de valeurs manquantes : 33213          % de valeurs manquantes : 60.41
```

```
Variables
---------
```

Nombre de doublons dans le dataset : 0

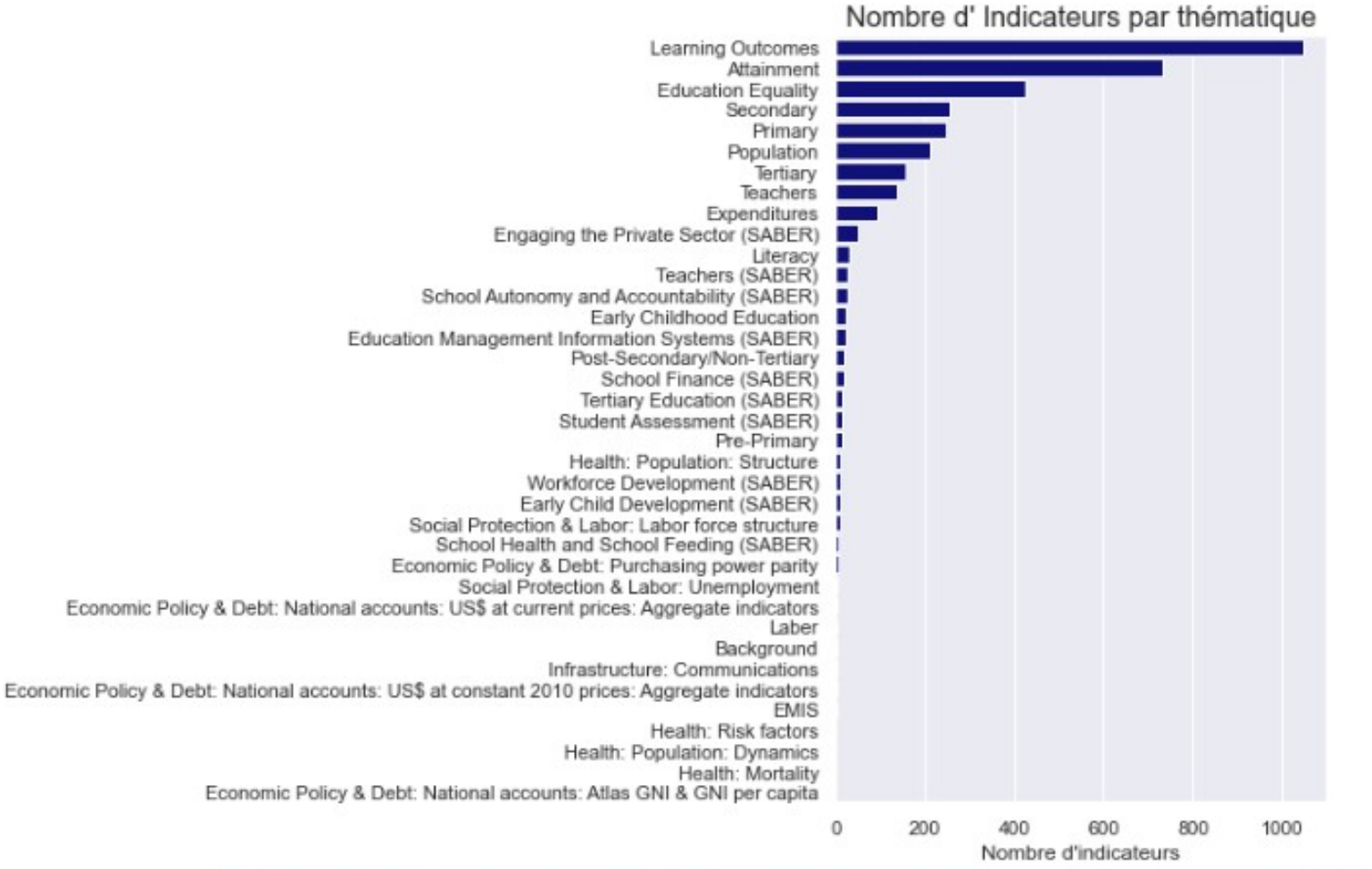
	Nombre	%
object	15	100.0

Visualisation des valeurs manquantes



Les 3665 indicateurs se répartissent dans 37 catégories

academy



Conclusions de la pré-analyse

academy

Les données à notre disposition contiennent suffisamment d'information pour réaliser une analyse afin de répondre au projet d'expansion à l'international de l'entreprise.

Nous allons utiliser le fichier **EdstatsData** pour exploiter les données.

Le fichier **EdstatsCountry** et le fichier **EdStatsSerie** seront utilisés pour des jointures internes avec **Edstatsdata** . Ils possèdent des colonnes qui peuvent être utilisées comme clé primaire lors des jointures.

academy

Dans le fichier EdStatsData.csv, la période la plus documentée de 2000 à 2015 sera analysée.

1.suppression des lignes pour lesquelles toutes les colonnes 'année' ont une valeur nan

2. Recherche des indicateurs les plus représentés dans le dataset
ajout d'une colonne Total
ajout du colonne % (taux de remplissage)

3.Seuil de remplissage fixé à 50 % . Il reste 450 indicateurs

Indicateurs restants : 450

statistiques descriptives des colonnes TOT et %

	count	mean	std	min	25%	50%	75%	max
TOT	3635.0	595.089959	860.858988	1.0	28.0	288.0	748.0	3582.0
%	3635.0	16.316094	23.703299	0.0	1.0	8.0	21.0	99.0

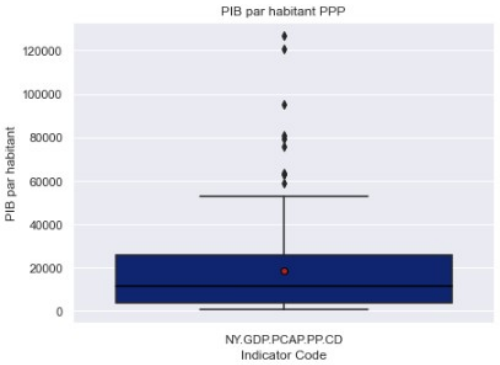
4. indicateurs les plus pertinents pour notre analyse

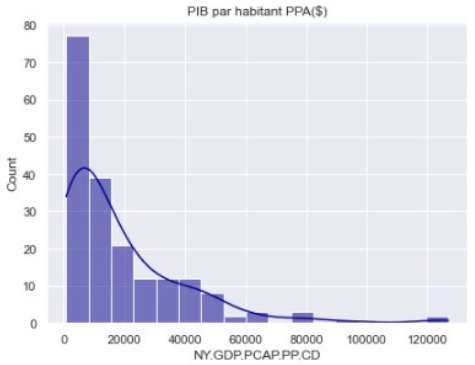
academy

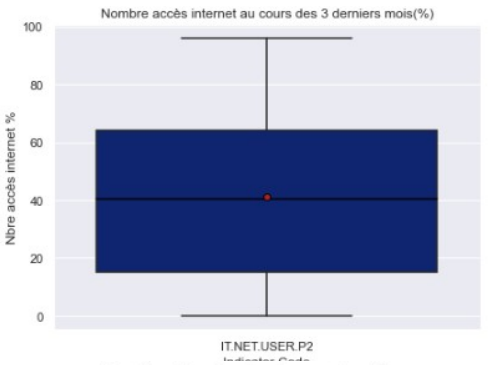
FACTEURS **ECONOMIQUES** : le pib du pays,le revenu moyen de la population.
FACTEURS **SOCIAUX** : la stabilité politique, le taux de scolarisation,la tranche de population entre 15 -25 ans
FACTEURS **ENERGETIQUES** : l'accès à l'électricité ,déploiement de l'internet, réseaux téléphonie fixe et mobile ...
Facteurs **STRUCTURELS** : accès aux infrastructures (développement réseau routier,ferroviaire et aérien, aux universités , bibliothèques , centres culturels....
En regardant sur le site de la banque mondiale de données, j'ai relevé plusieurs indicateurs et isolé des racines qui seront utilisées comme mot-clé pour les recherches.

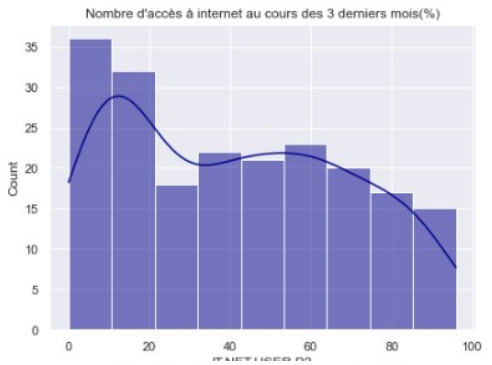
SE.SEC : enseignement dans le secondaire (lycée)
SE.TER : enseignement supérieur
1519 : population âgée de 15 à 19 ans
1524 : population âgée de 15 à 24 ans
2024 : population âgée de 20 à 24 ans
SP.POP: Population
IT.NET: Infrastructure technique
NY : National Accounts, produits intérieurs et nationaux
EG : Energie

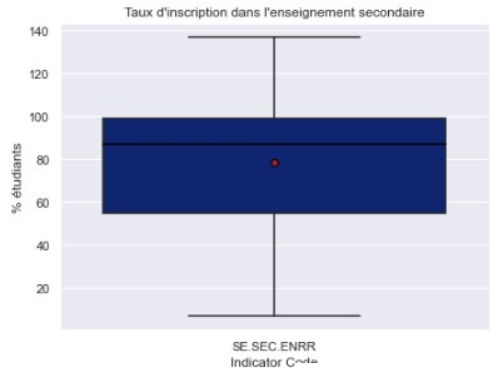
5.Analyse descriptive des données et score pour chaque indicateur

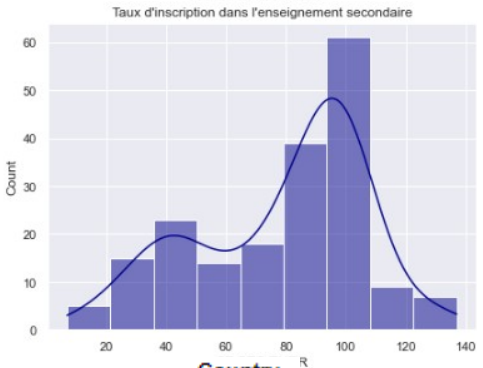


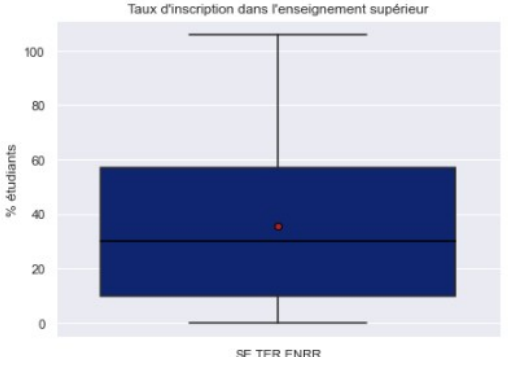


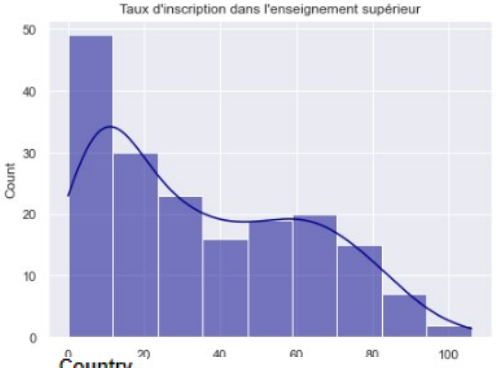












score_sec	Country Name
1.000000	Australia
0.994737	Belgium
0.989474	Netherlands
0.984211	Spain
0.978947	Finland

score_ter	Country Name
1.000000	Greece
0.994444	Korea, Rep.
0.988889	Finland
0.983333	United States
0.977778	Belarus

	score_ter	2000	2001	2002	200
count	181.00000	181.000000	181.000000	181.000000	181.00000
mean	0.50000	25.779006	26.756906	27.397790	28.39226
std	0.29108	20.870921	21.767879	22.740732	23.69120
min	0.00000	0.000000	0.000000	0.000000	0.00000
25%	0.25000	8.000000	8.000000	8.000000	8.00000
50%	0.50000	21.000000	20.000000	21.000000	21.00000
75%	0.75000	41.000000	43.000000	43.000000	44.00000
max	1.00000	82.000000	84.000000	86.000000	89.00000

score_pib	Country Name
1.000000	Qatar
0.994764	Macao SAR, China
0.989529	Luxembourg
0.984293	Brunei Darussalam
0.979058	Singapore

score_it	Country Name
1.000000	Iceland
0.995074	Norway
0.990148	Denmark
0.985222	Luxembourg
0.980296	Sweden

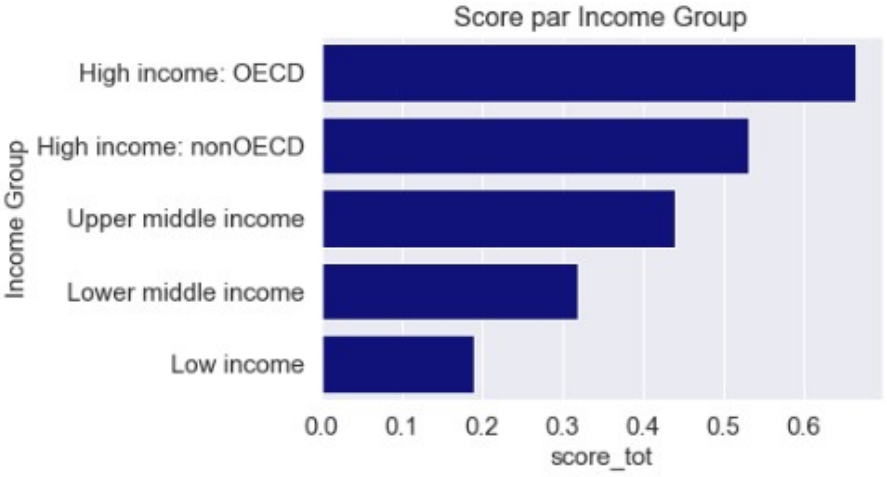
academy

academy

6. ajout d'un score total et affichage du Top 10

les 10 meilleurs candidats pour l'expansion à l'international sont :

	Country Name	score_1419	score_1524	score_sec	score_ter	score_it	score_pib	score_tot
0	Australia	0.659686	0.670157	1.000000	0.961111	0.891626	0.890052	0.734618
1	United States	0.989529	0.989529	0.631579	0.983333	0.832512	0.937173	0.729021
2	Germany	0.863874	0.890052	0.847368	0.816667	0.926108	0.895288	0.727066
3	Netherlands	0.586387	0.596859	0.989474	0.861111	0.975369	0.916230	0.722301
4	United Kingdom	0.853403	0.863874	0.905263	0.755556	0.955665	0.853403	0.721421
5	Spain	0.722513	0.732984	0.984211	0.972222	0.827586	0.811518	0.720548
6	France	0.858639	0.853403	0.926316	0.766667	0.886700	0.848168	0.713978
7	Korea, Rep.	0.827225	0.827225	0.700000	0.994444	0.945813	0.801047	0.711422
8	Japan	0.910995	0.910995	0.831579	0.761111	0.916256	0.842932	0.710479
9	Denmark	0.397906	0.382199	0.973684	0.933333	0.990148	0.905759	0.698829



Les régions North America et Europe & Central Asia ont les meilleurs scores.

De même les High income OECD et non OECD ont les meilleurs scores