


PROJET 2 : ANALYSEZ DES DONNÉES DE SYSTÈMES ÉDUCATIFS



Parcours Data Scientist - OpenClassrooms

Soutenance

SOMMAIRE

I – Rappel de la problématique et présentation du jeu de données

II – Analyse pré-exploratoire

- Nettoyage des données
- Étude des indicateurs
- Sélection des pays

III - Conclusions

I - RAPPEL DE LA PROBLÉMATIQUE ET PRÉSENTATION DU JEU DE DONNÉES



RAPPEL DE LA PROBLÉMATIQUE

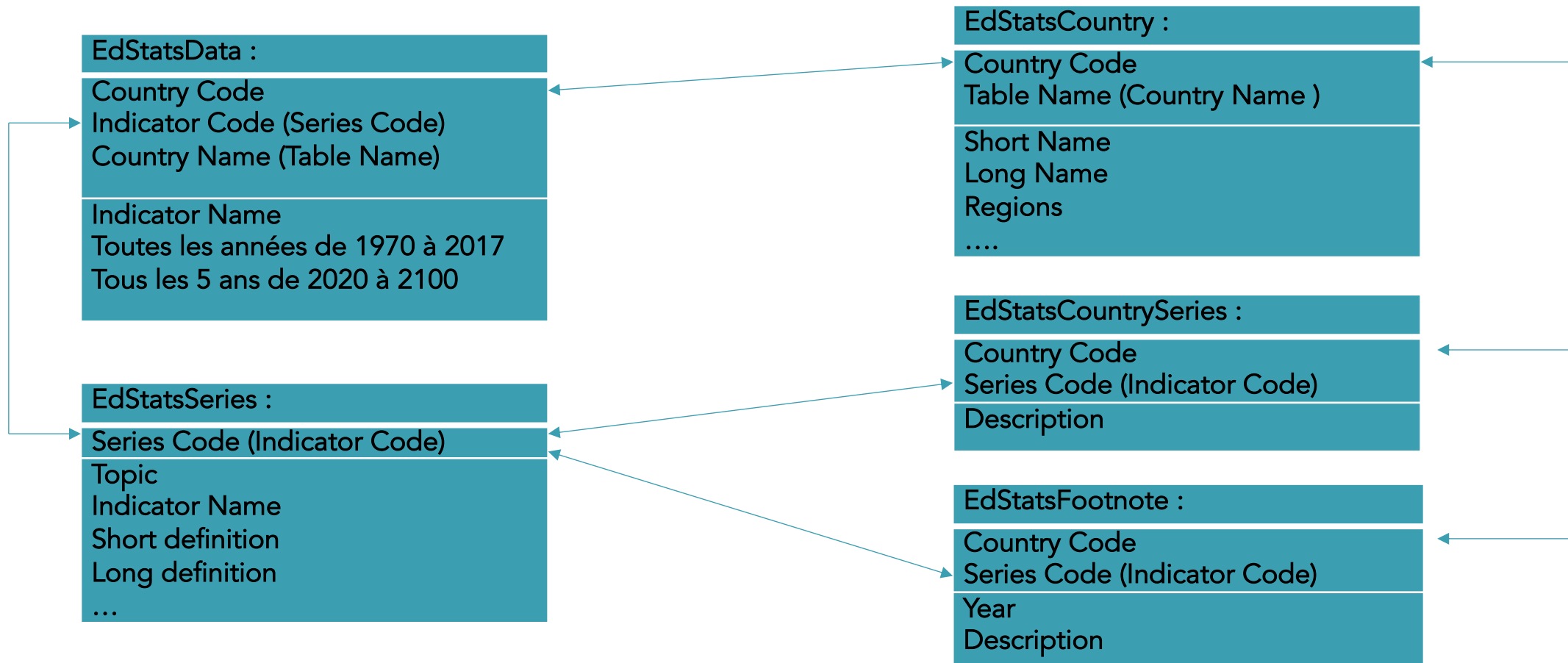
- start-up de la EdTech : *academy*
- Services : contenus de formation en ligne pour un public de niveau lycée et université
- Projet d'expansion à l'international

Mission : réaliser une première mission d'analyse exploratoire, pour déterminer si les données sur l'éducation de la banque mondiale permettent d'informer le projet d'expansion.



- « EdStats All Indicator Query » de la Banque mondiale
- Dataset : 5 tables différentes ayant des clés communes

LIENS ENTRE LES DIFFÉRENTES TABLES

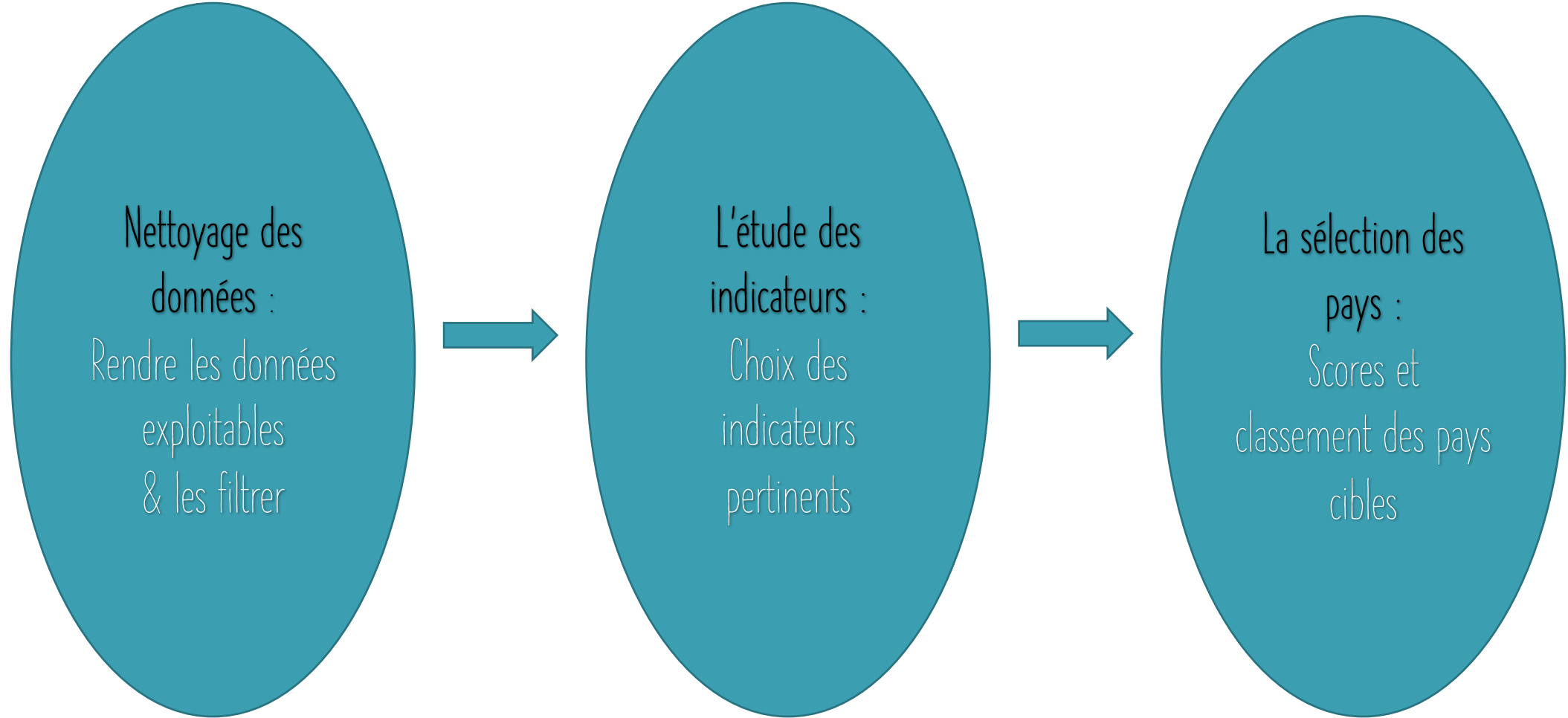


PRÉSENTATION DU JEU DE DONNÉES

EdStatsData	L'évolution de nombreux indicateurs pour tous les pays et certains groupes de pays ou subdivisions 886 930 observations et 70 variables. 3 665 indicateurs renseignés pour 242 pays
EdStatsCountry	Informations économiques concernant les différents pays. 241 observations et 32 variables.
EdStatsCountry-Series	Répertoire les sources des données concernant les pays représentés. 4 variables et 613 observations.
EdStatsSeries	Les indicateurs de la table <i>data</i> , ainsi que des informations les concernant (telles que leur définition, leur méthode d'agrégation ou leur source). 3665 observations et 21 variables.
EdStatsFootNote	Informations supplémentaires concernant les données comme leur année et leur description. 643 638 observations et 5 variables.

II - ANALYSE PRÉ- EXPLORATOIRE





NETTOYAGE DES DONNÉES

Data

Nombre total de cellules manquantes : 52568249
Nombre de cellules manquantes en % : 85.90%

Series

Nombre total de cellules manquantes : 51538
Nombre de cellules manquantes en % : 70.31%

Country

Nombre total de cellules manquantes : 2113
Nombre de cellules manquantes en % : 28.28%

Country_series

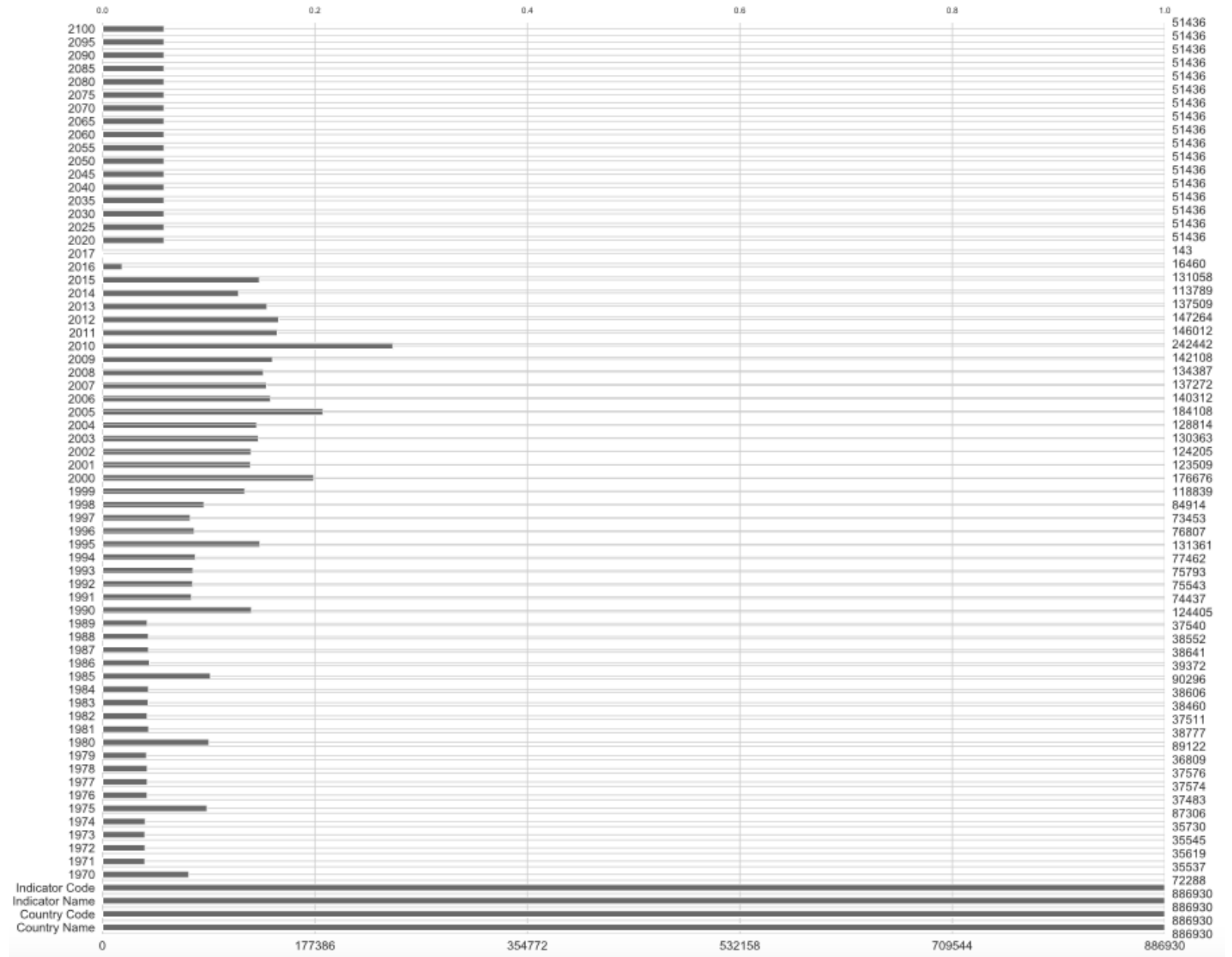
Nombre total de cellules manquantes : 0
Nombre de cellules manquantes en % : 0.00%
Nombre de cellules manquantes pour chaque variable :
CountryCode 0
SeriesCode 0
DESCRIPTION 0
dtype: int64

Footnote

Nombre total de cellules manquantes : 0
Nombre de cellules manquantes en % : 0.00%
Nombre de cellules manquantes pour chaque variable :
CountryCode 0
SeriesCode 0
Year 0
DESCRIPTION 0
dtype: int64

NETTOYAGE DES DONNÉES : SÉLECTION DES ANNÉES

- Critère d'accès à la technologie
- Internet apparait à partir des années 1990 -> essor à partir des années 2000.
- Années entre 2010 et 2017



NETTOYAGE DES DONNÉES : PREMIER FILTRAGE DES PAYS

	Country Name	last year
830823	Tuvalu	1.109700e+04
640243	Palau	2.150300e+04
761188	St. Martin (French part)	3.126400e+04
691553	San Marino	3.320300e+04
365368	Gibraltar	3.440800e+04
...
64838	Middle income	5.592833e+09
46513	Low & middle income	6.252106e+09
90493	World	7.442136e+09
193113	British Virgin Islands	NaN
596263	Nauru	NaN

242 rows x 2 columns

- pays < 1 million d'habitants -> 56 pays.
- pays < 1000 données renseignées -> 40 pays.
- il reste 146 pays potentiels.

ÉTUDE DES INDICATEURS : CRITÈRES

Démographique :
Population totale ou par
tranche d'âge

Économique :
Revenus, PIB par habitant

Niveau d'éducation :
Nombre de lycéens et
d'étudiants

Accès à la technologie :
Accès à internet, niveau
d'utilisation d'internet

ÉTUDE DES INDICATEURS : SÉLECTION DES INDICATEURS

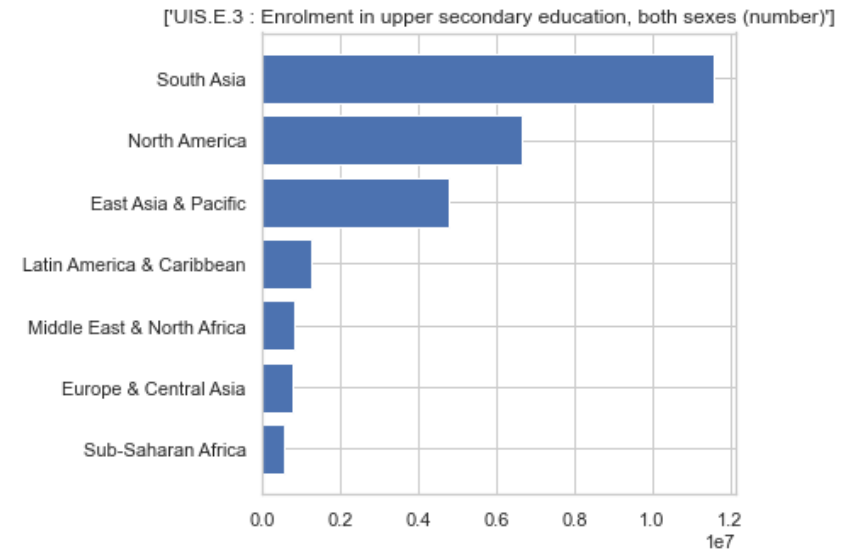
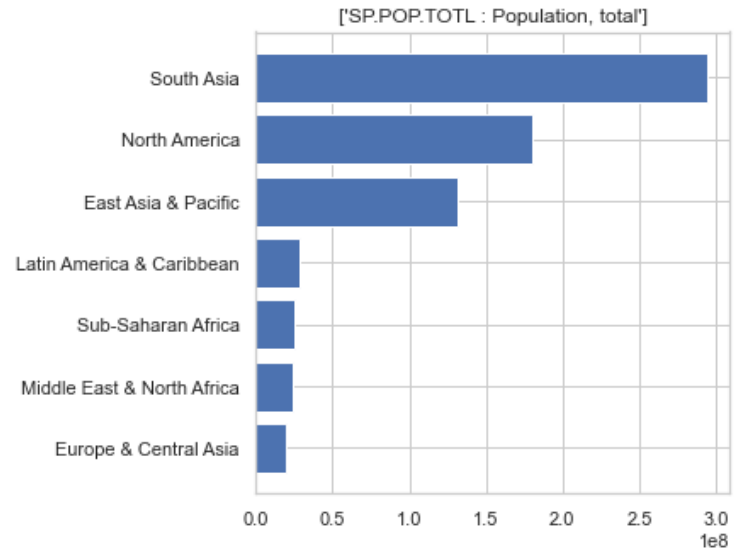
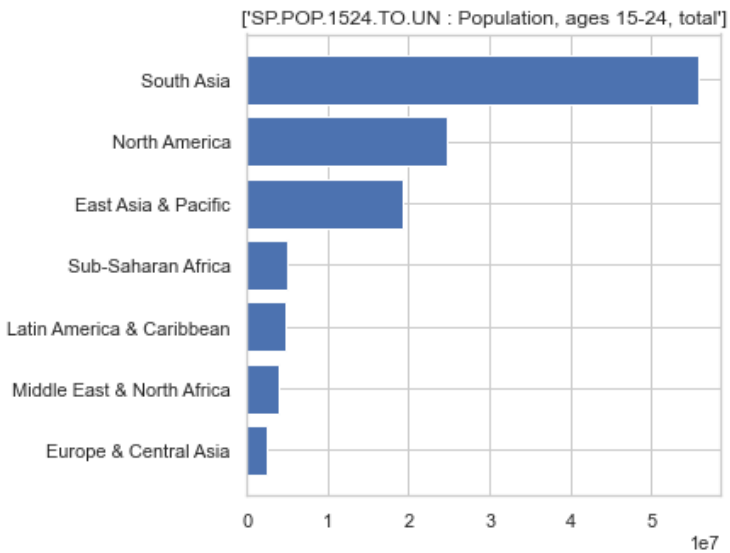
- démographique :
 - Population, âges 15-24, total : SP.POP.1524.TO.UN
 - Population, total : SP.POP.TOTL
- éducation :
 - Inscriptions dans l'enseignement secondaire supérieur, les deux sexes (nombre) : UIS.E.3
 - Inscriptions dans l'enseignement post-secondaire non supérieur, les deux sexes (nombre) : UIS.E.4
 - Inscriptions dans l'enseignement supérieur, tous programmes, les deux sexes (nombre) : SE.TER.ENRL
- économique :
 - PIB par habitant : NY.GDP.PCAP.CD
- accès à la technologie :
 - Utilisateurs d'internet (pour 100 personnes) : IT.NET.USER.P2

	Indicator Name	Indicator Code	last year
4	Internet users (per 100 people)	IT.NET.USER.P2	146
7	Population, total	SP.POP.TOTL	146
3	GDP per capita (current US\$)	NY.GDP.PCAP.CD	145
6	Population, ages 15-24, total	SP.POP.1524.TO.UN	145
1	Enrolment in tertiary education, all programme...	SE.TER.ENRL	135
2	Enrolment in upper secondary education, both s...	UIS.E.3	135
0	Enrolment in post-secondary non-tertiary educa...	UIS.E.4	87

ÉTUDE DES INDICATEURS : ORDRES DE GRANDEUR

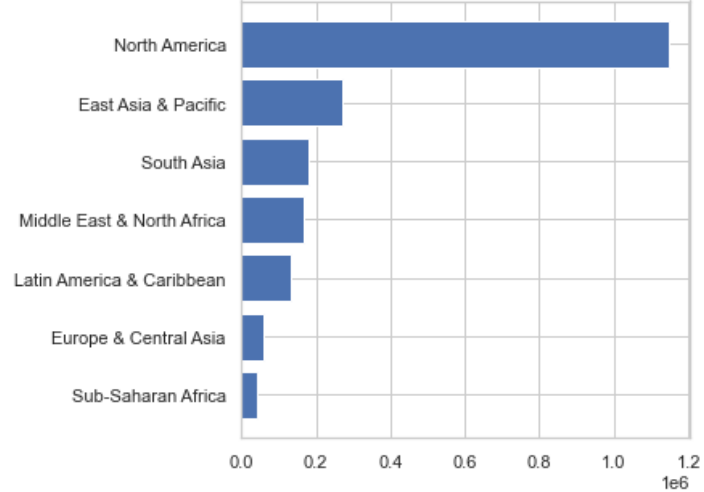
	count	mean	std	min	25%	50%	75%	max
Indicator Code								
IT.NET.USER.P2	146.0	5.070397e+01	2.797006e+01	4.000000e+00	2.527626e+01	5.361339e+01	7.579990e+01	9.799998e+01
NY.GDP.PCAP.CD	145.0	1.283372e+04	1.713675e+04	2.857274e+02	1.443625e+03	4.878576e+03	1.569241e+04	7.989052e+04
SE.TER.ENRL	135.0	1.524054e+06	4.980806e+06	5.001000e+03	1.256610e+05	2.682310e+05	8.600025e+05	4.336739e+07
SP.POP.1524.TO.UN	145.0	8.143981e+06	2.666002e+07	1.244930e+05	7.102940e+05	2.159790e+06	5.995687e+06	2.441202e+08
SP.POP.TOTL	146.0	4.971619e+07	1.617797e+08	1.170125e+06	5.152598e+06	1.137570e+07	3.697354e+07	1.378665e+09
UIS.E.3	135.0	1.782539e+06	6.194255e+06	2.672100e+04	1.472150e+05	3.631876e+05	1.182214e+06	5.522868e+07
UIS.E.4	87.0	1.216541e+05	2.849579e+05	1.870000e+02	4.427000e+03	1.438000e+04	6.648300e+04	1.564102e+06

ÉTUDE DES INDICATEURS : ORDRES DE GRANDEUR

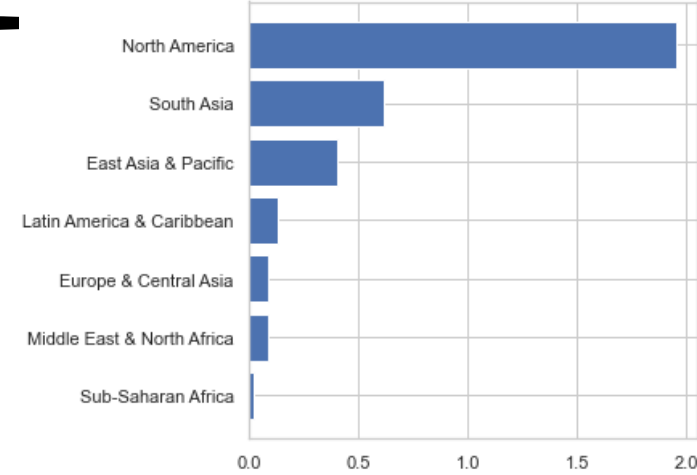


ÉTUDE DES INDICATEURS : ORDRES DE GRANDEUR

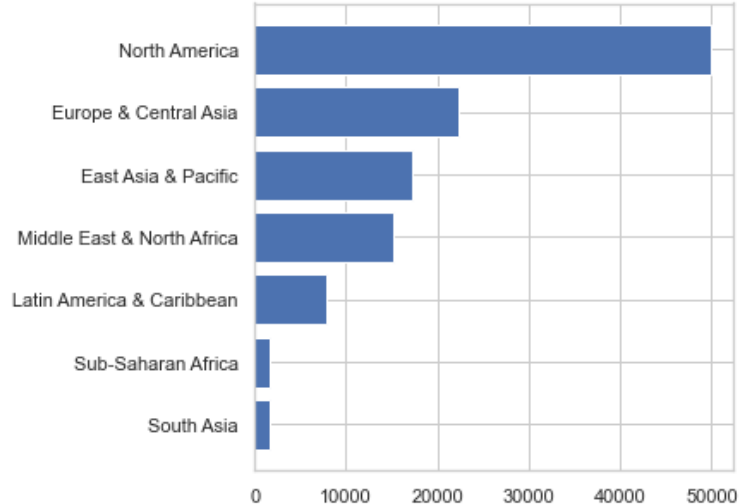
['UIS.E.4 : Enrolment in post-secondary non-tertiary education, both sexes (number)']



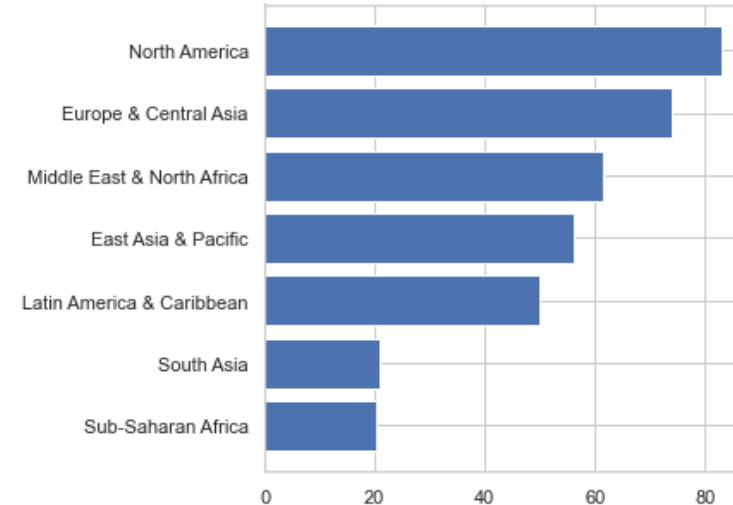
['SE.TER.ENRL : Enrolment in tertiary education, all programmes, both sexes (number)']



['NY.GDP.PCAP.CD : GDP per capita (current US\$)']



['IT.NET.USER.P2 : Internet users (per 100 people)']

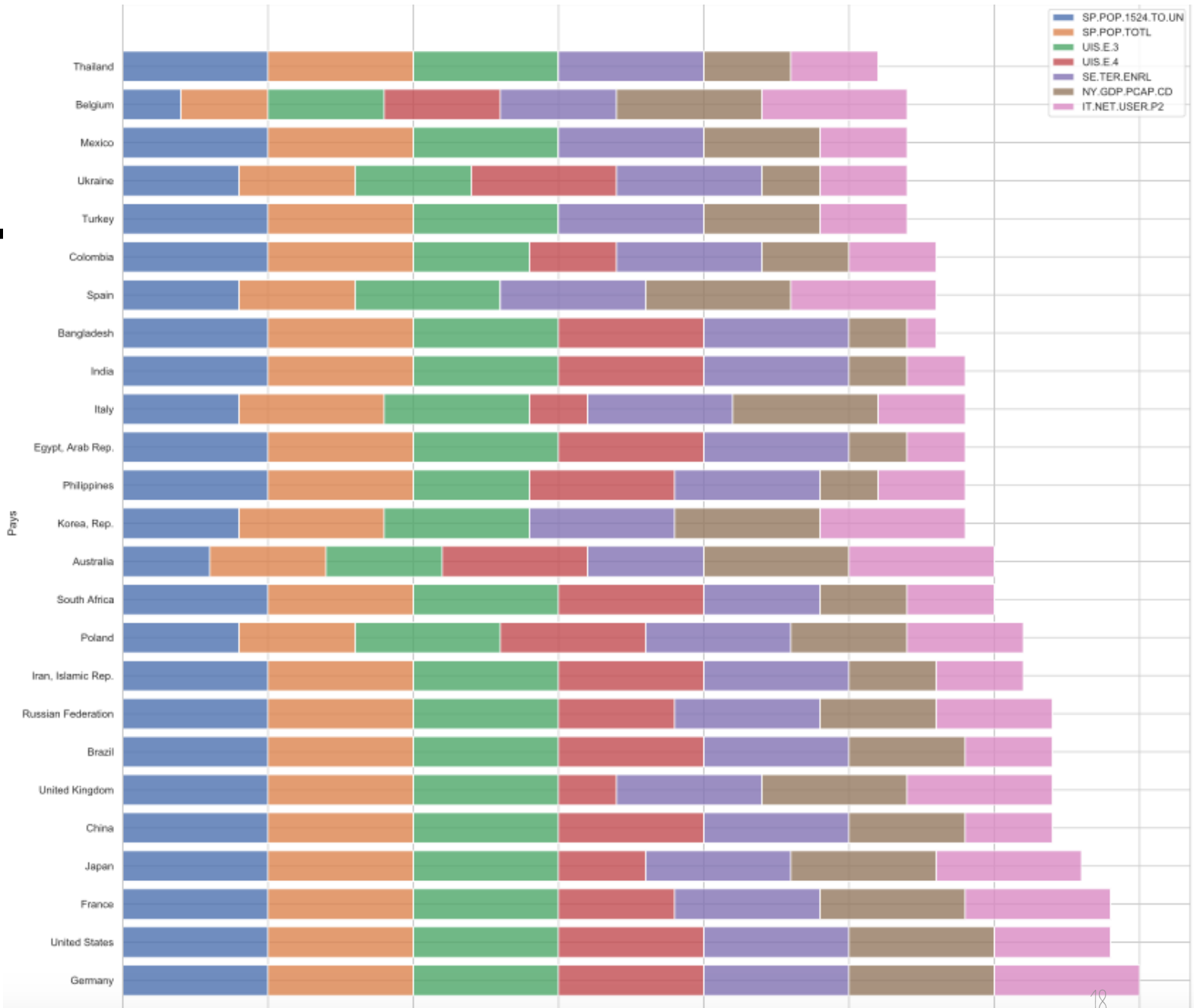


SCORING

- score de 1 à 5 pour chaque indicateur
- score total = somme des scores des 7 indicateurs (sur 35)
- *pd.qcut()* : fonction de discrétisation basée sur les quantiles

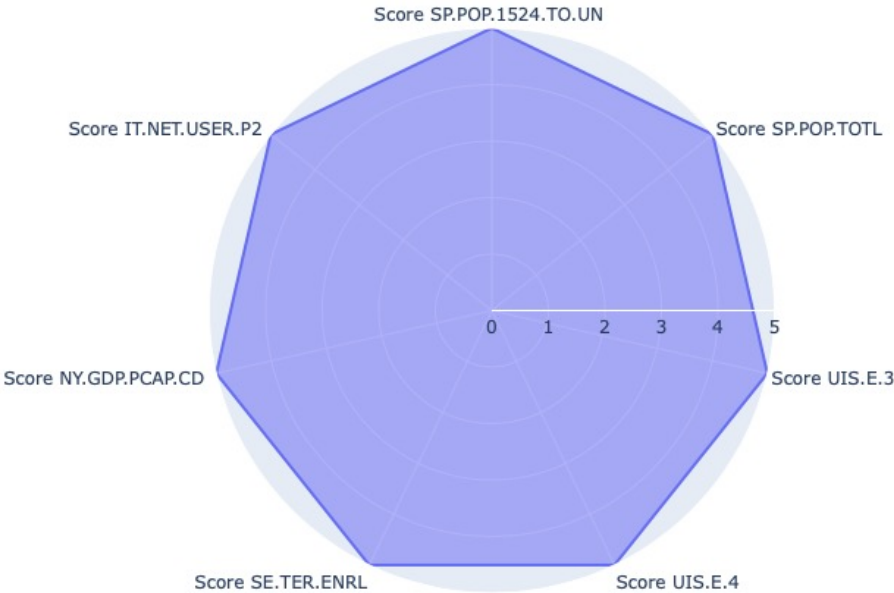
	Country Name	Score SP.POP.1524.TO.UN	Score SP.POP.TOTL	Score UIS.E.3	Score UIS.E.4	Score SE.TER.ENRL	Score NY.GDP.PCAP.CD	Score IT.NET.USER.P2	Score total
0	Afghanistan	4.0	4	4.0	5.0	3.0	1.0	1	22.0
1	Albania	1.0	1	2.0	1.0	2.0	3.0	4	14.0
2	Algeria	4.0	4	4.0	0.0	4.0	3.0	3	22.0
3	Argentina	4.0	4	5.0	0.0	5.0	4.0	4	26.0
4	Armenia	1.0	1	1.0	0.0	2.0	2.0	4	11.0
5	Australia	3.0	4	4.0	5.0	4.0	5.0	5	30.0
6	Austria	2.0	2	3.0	3.0	4.0	5.0	5	24.0
7	Azerbaijan	3.0	3	3.0	4.0	2.0	3.0	4	22.0
8	Bahrain	1.0	1	1.0	2.0	1.0	5.0	5	16.0
9	Bangladesh	5.0	5	5.0	5.0	5.0	2.0	1	28.0
10	Belarus	2.0	3	2.0	3.0	4.0	3.0	4	21.0

PAYS SÉLECTIONNÉS

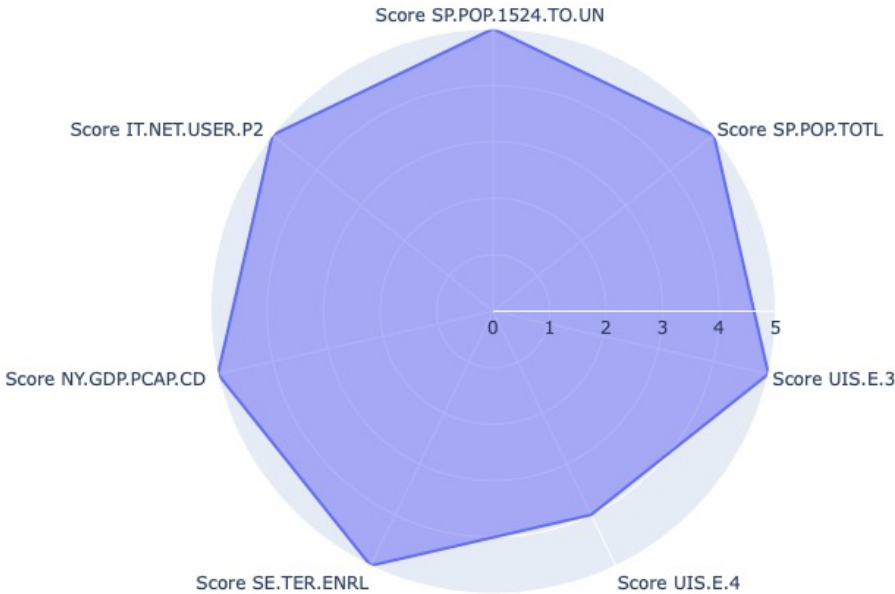


PAYS SÉLECTIONNÉS

Germany :

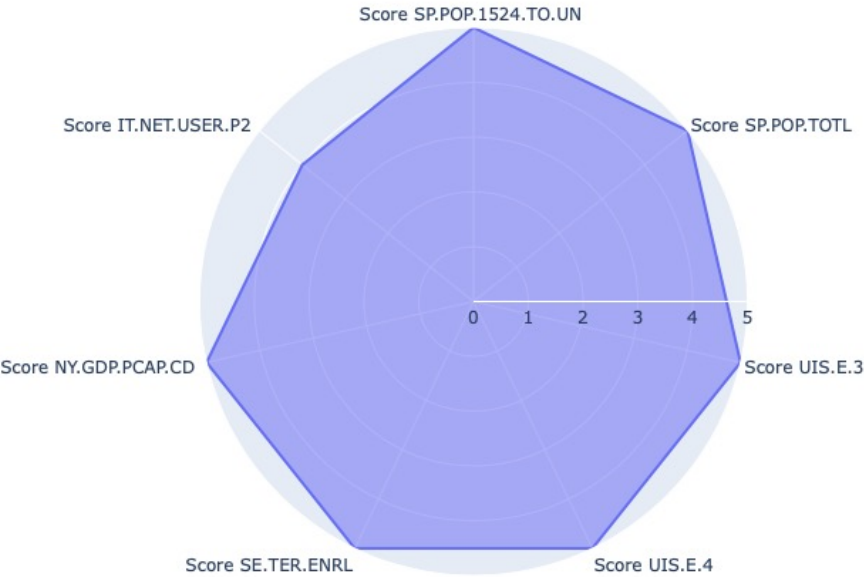


France :

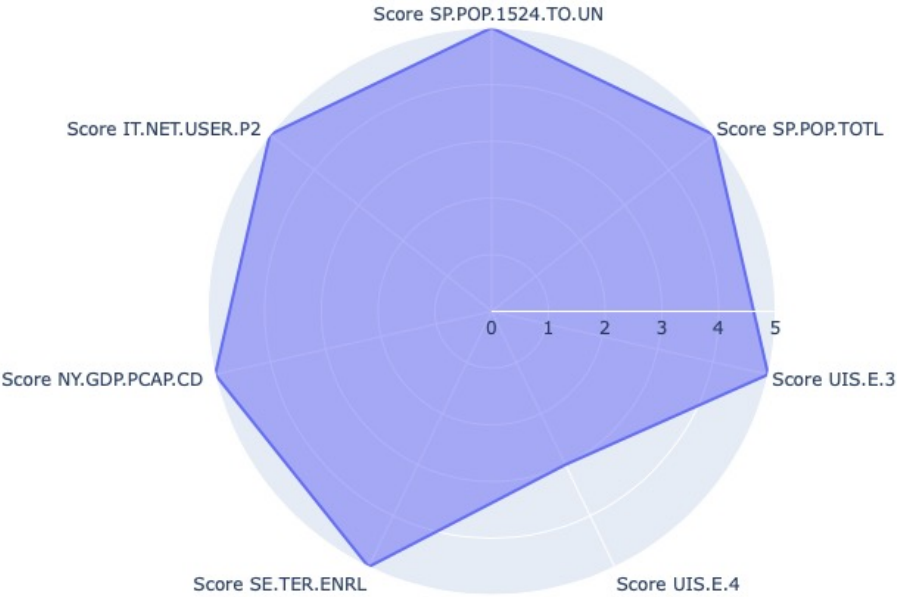


PAYS SÉLECTIONNÉS

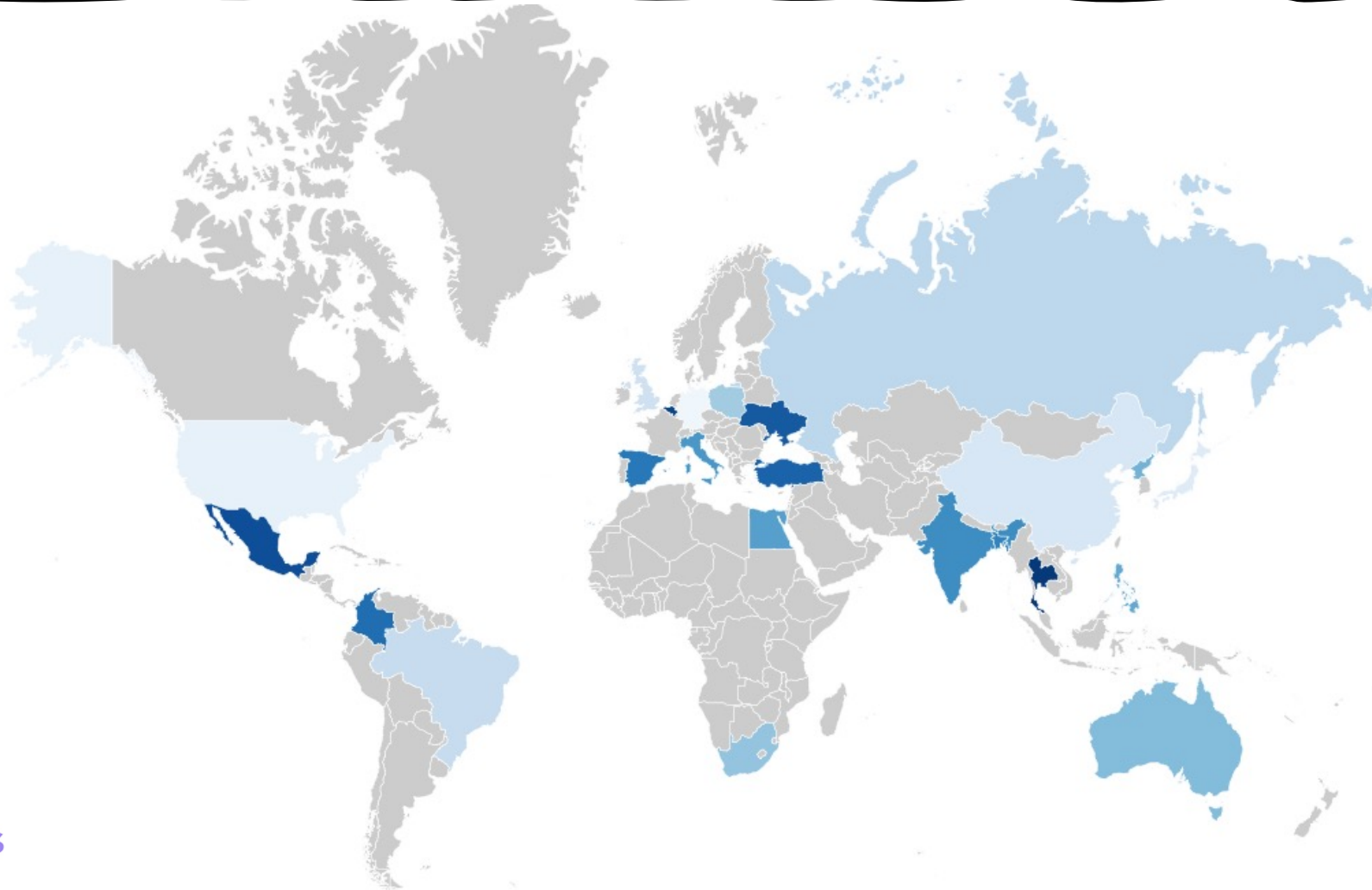
United States :



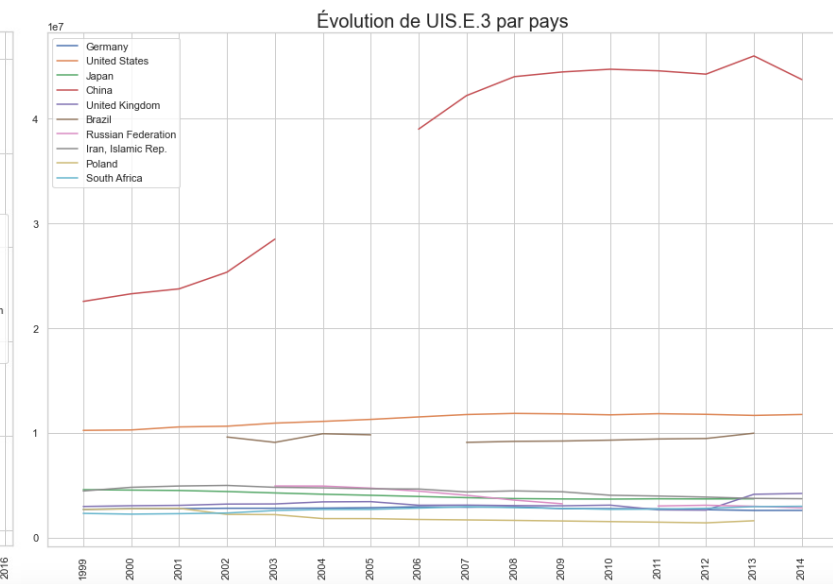
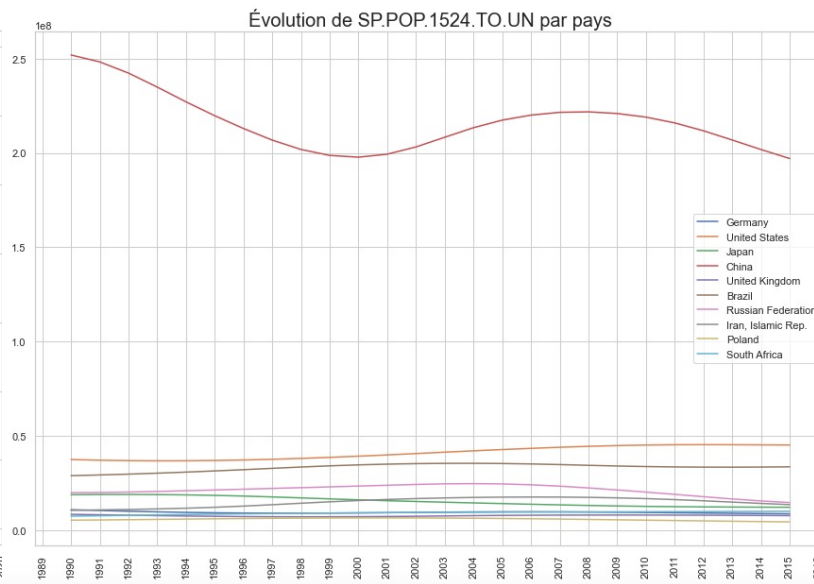
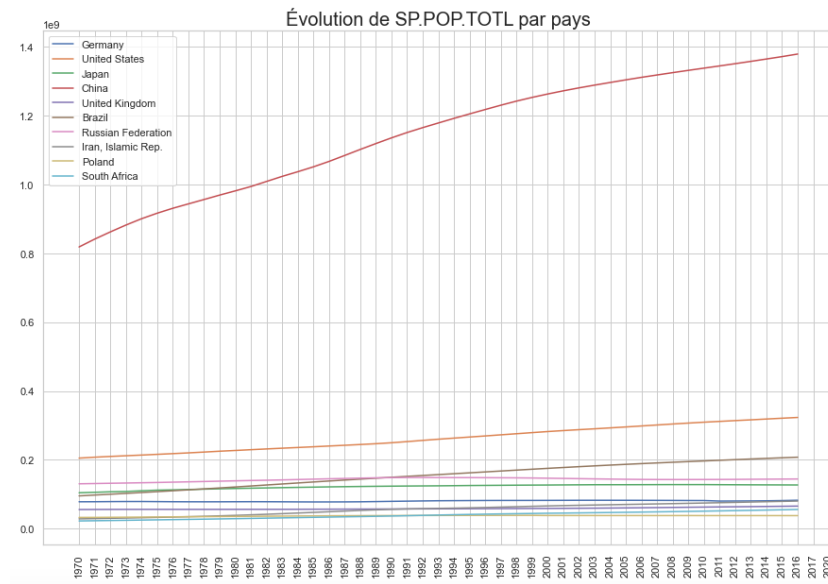
Japan :



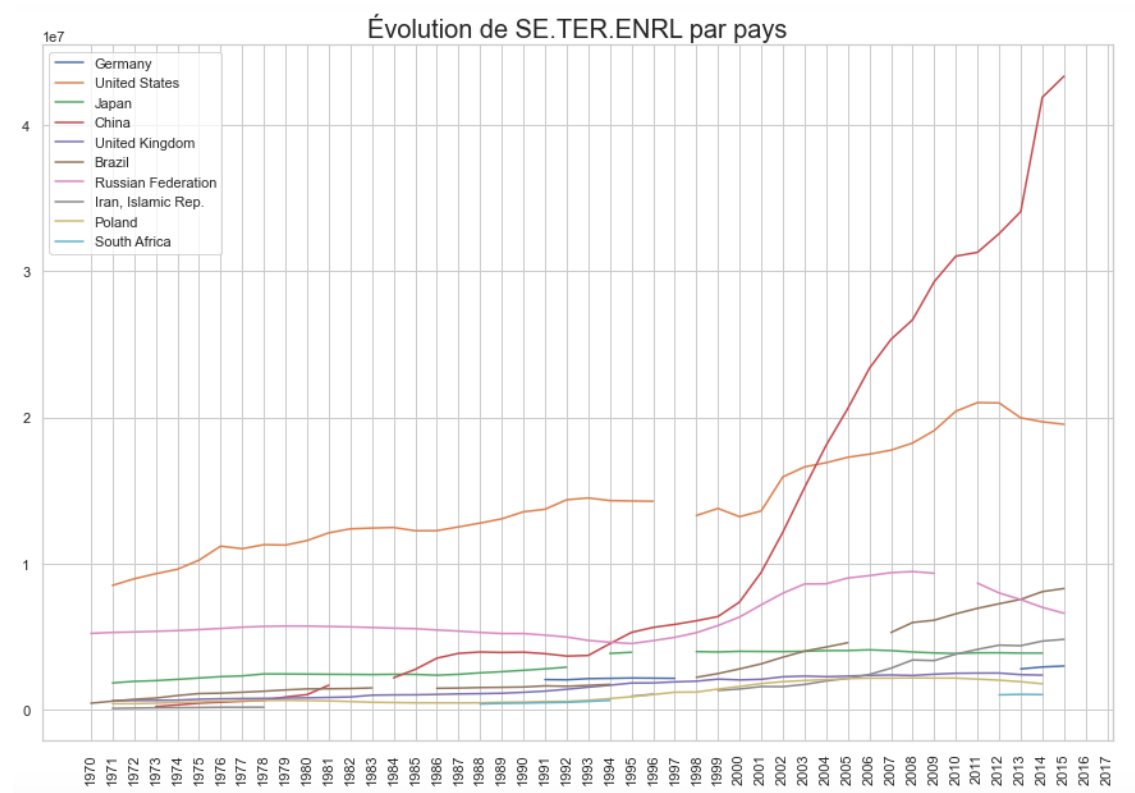
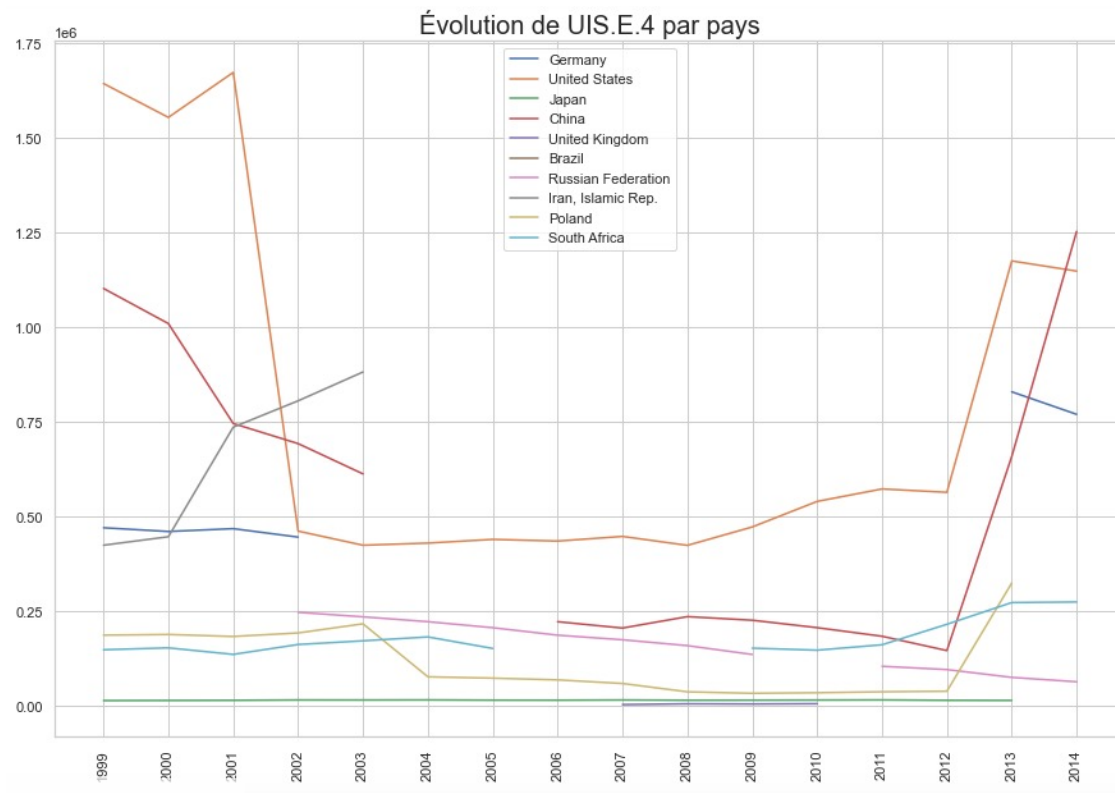
PAYS SÉLECTIONNÉS



ANALYSE PROSPECTIVE

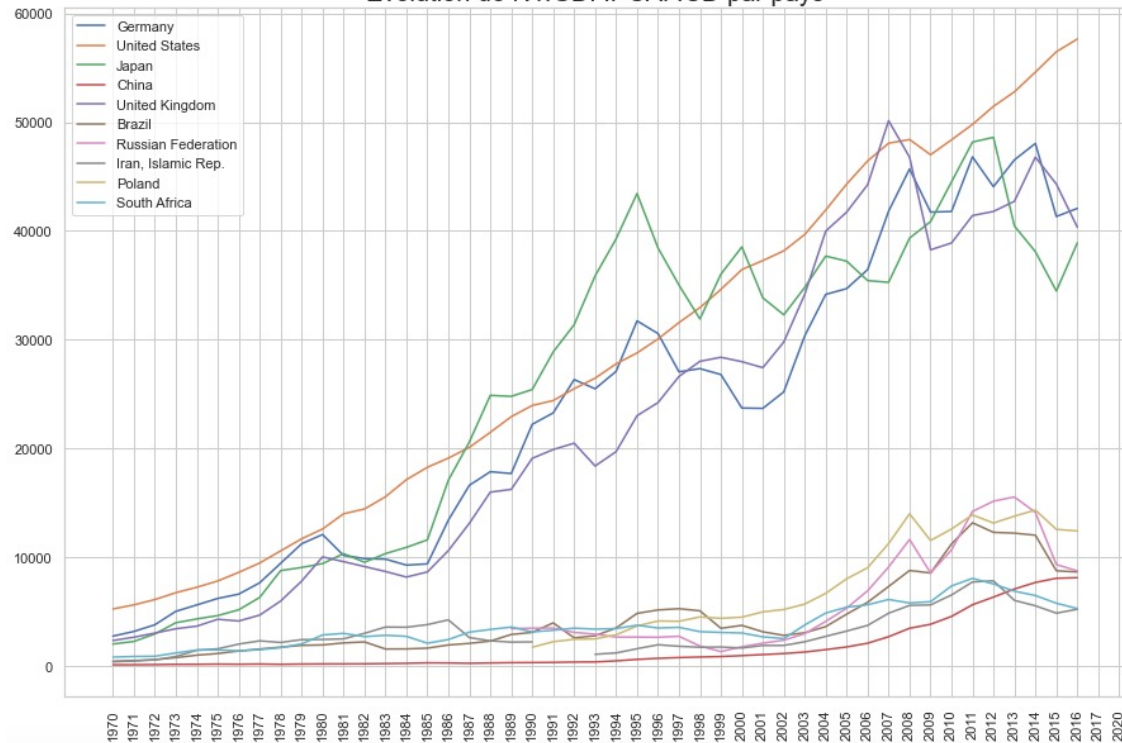


ANALYSE PROSPECTIVE

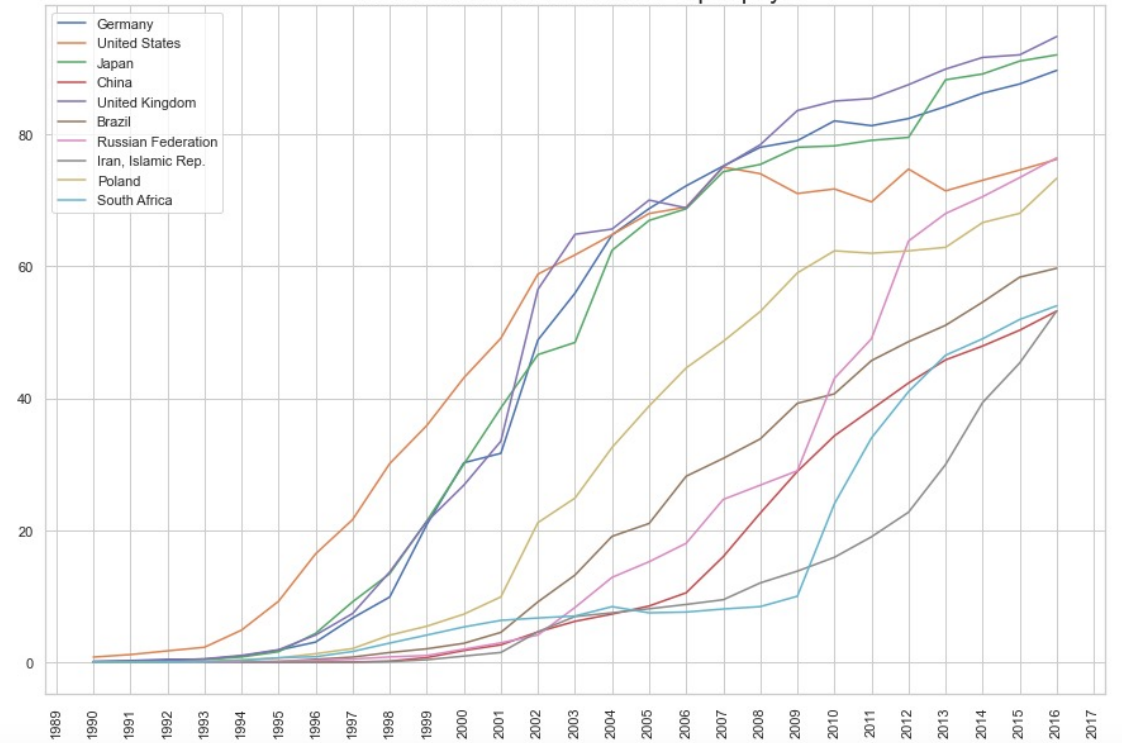


ANALYSE PROSPECTIVE

Évolution de NY.GDP.PCAP.CD par pays



Évolution de IT.NET.USER.P2 par pays



III - CONCLUSIONS



```
['Germany',  
 'United States',  
 'Japan',  
 'China',  
 'United Kingdom',  
 'Brazil',  
 'Russian Federation',  
 'Iran, Islamic Rep.',  
 'Poland',  
 'South Africa',  
 'Australia',  
 'Korea, Rep.',  
 'Philippines',  
 'Egypt, Arab Rep.',  
 'Italy',  
 'India',  
 'Bangladesh',  
 'Spain',  
 'Colombia',  
 'Turkey',  
 'Ukraine',  
 'Mexico',  
 'Belgium',  
 'Thailand']
```

- mission d'analyse exploratoire : déterminer si les données sur l'éducation de la banque mondiale permettent d'informer le projet d'expansion d'academy
- données portant sur le domaine éducatif et tous les pays sont représentés
- beaucoup de valeurs manquantes
- pas possible de réaliser d'analyse prospective

MERCI DE VOTRE ATTENTION
