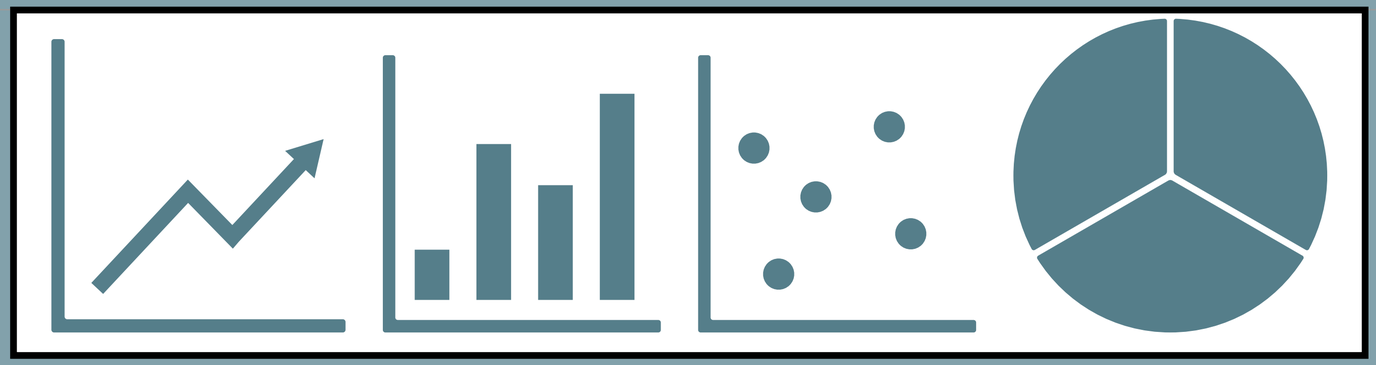
Data Scientist

Projet N°2: Analysez des données de systèmes éducatifs



academy

Sommaire

Présentation de l’entreprise

academy

C’est une start-up de la EdTech,
qui propose des contenus de formation en ligne pour un public de niveau lycée et université.

Notre manager, nous a convié à une réunion pour nous présenter le projet d’expansion à l’international de l’entreprise.
Il nous confie 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.

LA BANQUE MONDIALE
BIRD • IDA

source: https://datacatalog.worldbank.org/dataset/education-statistics

Problématique de l’entreprise

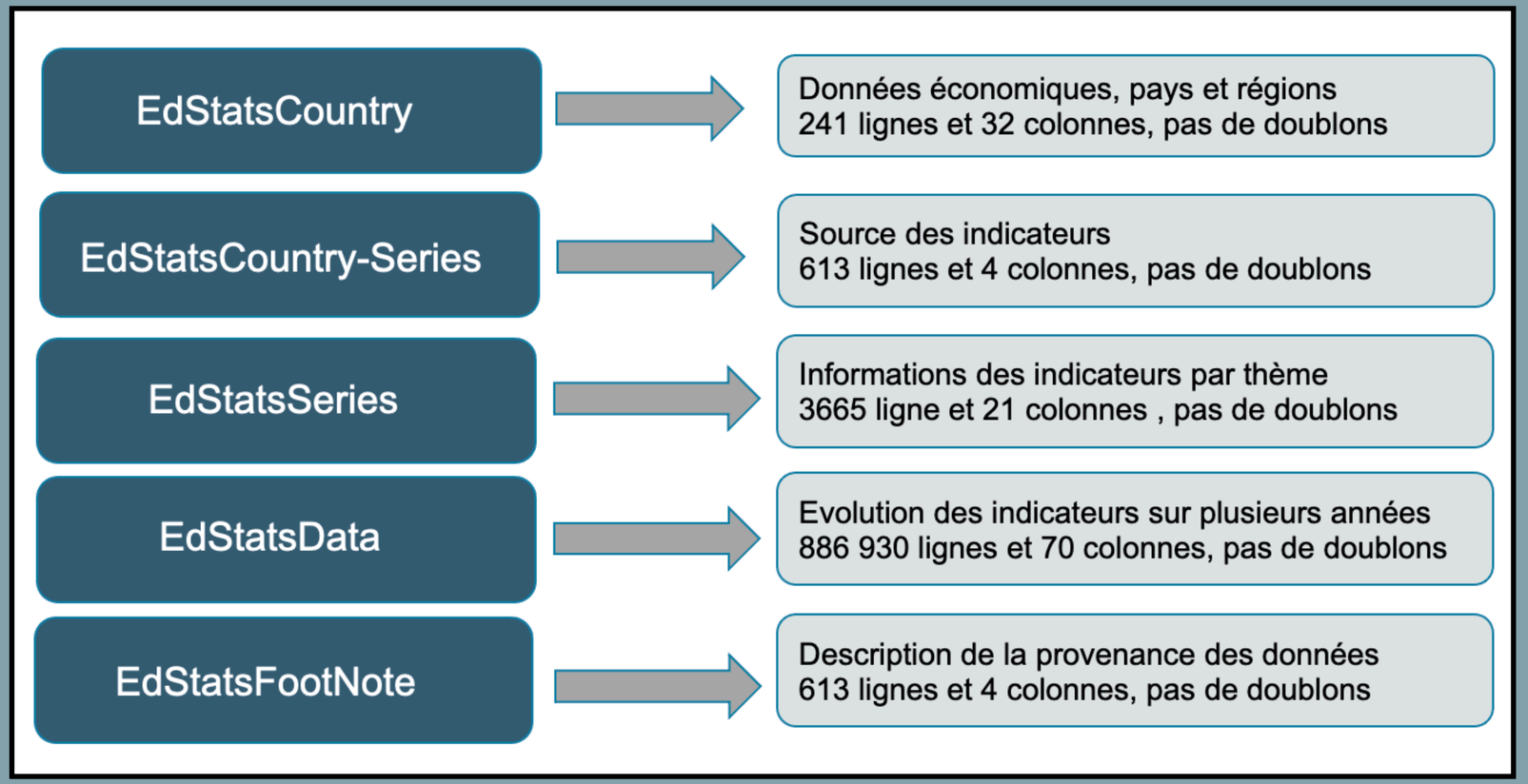
- Quels sont les pays avec un fort potentiel de clients pour nos services ?
- Pour chacun de ces pays, quelle sera l’évolution de ce potentiel de clients ?
- Dans quels pays l'entreprise doit-elle opérer en priorité ?

Quelques pré-Analyse pour pouvoir avancer aux niveaux des problématique

- Valider la qualité de ce jeu de données
- Décrire les informations contenues dans le jeu de données
- Sélectionner les informations qui semblent pertinentes
- Déterminer des ordres de grandeurs des indicateurs statistiques classiques pour les différentes zones géographiques et pays du monde

Le travail va nous permettre de déterminer si ce jeu de données peut informer les décisions d'ouverture vers de nouveaux pays.

Description des données



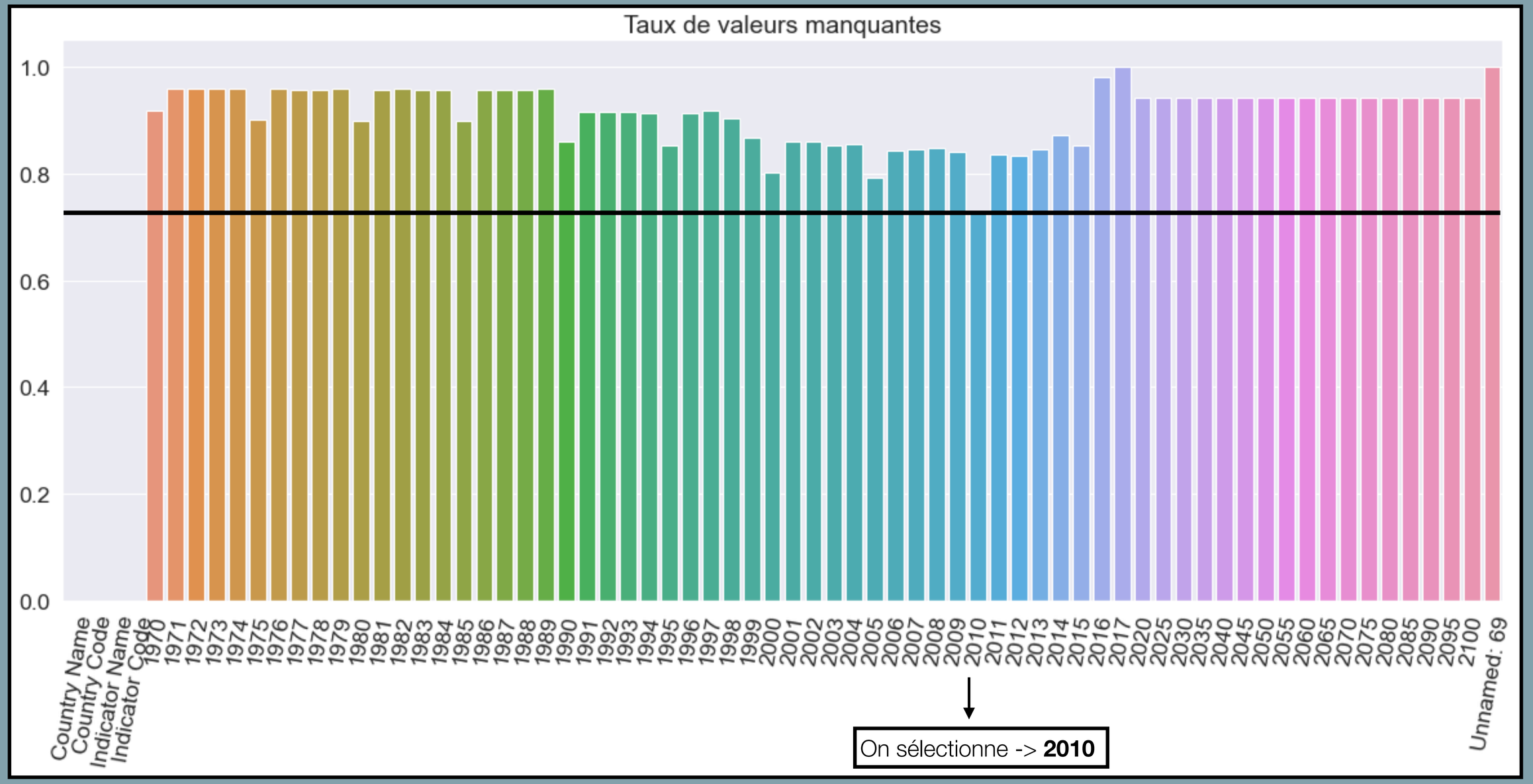
Les colonnes qui sont pertinent et qui peuvent être utile pour répondre à la problématique de l'entreprise sont:

- Long Name
- Income Group
- Region
- Country name
- Date
- Indicator Name

Le fichier N°1: **EdStatsCountry** contient des informations globales sur l'économie de chaque pays du monde et le classement de chaque pays par région. Des informations supplémentaires sont également pertinentes pour les groupes de pays par région ou par niveau de revenu.

Le fichier N°3: **EdStatsData** donne l'évolution de plusieurs indaicateurs sur les années.

Sélections d'indicateurs



Sélections d'indicateurs

Pour rappel le but de l'entreprise est de proposé des contenus de formation en ligne pour un public de niveau lycée et université.(Secondaire & Tertiaire)

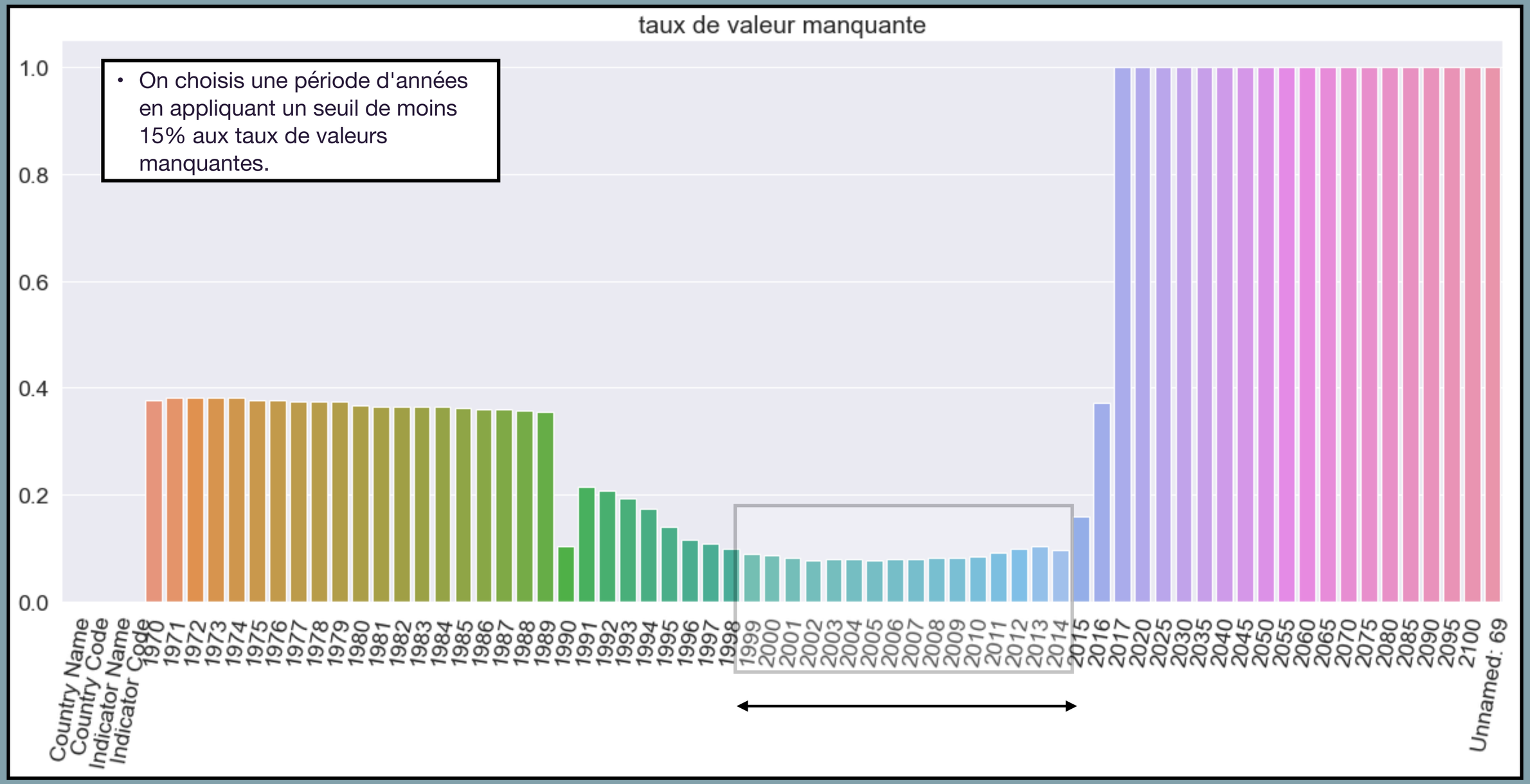
Donc pour cela on va sélectionner les indicateurs par mots clés qui contiennent les infos sur les situations:

- Économique (revenus par habitant, niveau de vie…).
- La population totale.
- La population d'une tranche d'âge donnée en général ici c'est (15-24 ans) (lycée et université).
- Accès à internet (posséder un ordinateur, électricité…).
- Le nombre des gens qui sont déjà inscrits ou éligibles d'être inscrits dans l'éducation secondaire & tertiaire. (nombre scolarisation lycée & université).

	Indicator Name	Indicator Code	2010
0	Population growth (annual %)	SP.POP.GROW	240
1	Population, total	SP.POP.TOTL	240
2	GDP per capita (current US$)	NY.GDP.PCAP.CD	228
3	Internet users (per 100 people)	IT.NET.USER.P2	227
4	Population of the official age for secondary education, both sexes (number)	SP.SEC.TOTL.IN	219
5	Population of the official age for tertiary education, both sexes (number)	SP.TER.TOTL.IN	217
6	Population, ages 15-24, total	SP.POP.1524.TO.UN	181

Après avoir sélectionner les indicateurs on fais une nouvelles liste qu'avec les indicateurs choisis, pour pouvoir continuer la suite de notre analyse.

Les données



- On choisis une période d'années en appliquant un seuil de moins 15% aux taux de valeurs manquantes.

Les données

- Séparation listes, pour pouvoir avancée nous allons devoir faire deux listes séparée pour les pays et les zones géographiques/économiques.

Après avoir séparer les liste en 2, pays/régions

214 pays

```
array(['Afghanistan', 'Albania', 'Algeria', 'American Samoa', 'Andorra',
       'Angola', 'Antigua and Barbuda', 'Argentina', 'Armenia', 'Aruba',
       'Australia', 'Austria', 'Azerbaijan', 'Bahrain', 'Bangladesh',
       'Barbados', 'Belarus', 'Belgium', 'Belize', 'Benin', 'Bermuda',
       'Bhutan', 'Bolivia', 'Bosnia and Herzegovina', 'Botswana',
       'Brazil', 'Bulgaria', 'Burkina Faso', 'Burundi', 'Cabo Verde',
       'Cambodia', 'Cameroon', 'Canada', 'Cayman Islands',
       'Central African Republic', 'Chad', 'Channel Islands', 'Chile',
       'China', 'Colombia', 'Comoros', 'Costa Rica', 'Croatia', 'Cuba',
       'Cyprus', 'Czech Republic', 'Denmark', 'Djibouti', 'Dominica',
       'Dominican Republic', 'Ecuador', 'El Salvador',
       'Equatorial Guinea', 'Eritrea', 'Estonia', 'Ethiopia', 'Fiji',
       'Finland', 'France', 'French Polynesia', 'Gabon', 'Georgia',
       'Germany', 'Ghana', 'Greece', 'Greenland', 'Grenada', 'Guam',
       'Guatemala', 'Guinea', 'Guinea-Bissau', 'Guyana', 'Haiti',
       'Honduras', 'Hong Kong SAR, China', 'Hungary', 'Iceland', 'India',
       'Indonesia', 'Iraq', 'Ireland', 'Isle of Man', 'Israel', 'Italy',
       'Jamaica', 'Japan', 'Jordan', 'Kazakhstan', 'Kenya', 'Kiribati',
       'Kosovo', 'Kuwait', 'Kyrgyz Republic', 'Lao PDR', 'Latvia',
       'Lebanon', 'Lesotho', 'Liberia', 'Libya', 'Liechtenstein',
       'Lithuania', 'Luxembourg', 'Macao SAR, China', 'Madagascar',
       'Malawi', 'Malaysia', 'Maldives', 'Mali', 'Malta',
       'Marshall Islands', 'Mauritania', 'Mauritius', 'Mexico', 'Moldova',
       'Monaco', 'Mongolia', 'Montenegro', 'Morocco', 'Mozambique',
       'Myanmar', 'Namibia', 'Nepal', 'Netherlands', 'New Caledonia',
       'New Zealand', 'Nicaragua', 'Niger', 'Nigeria',
       'Northern Mariana Islands', 'Norway', 'Oman', 'Pakistan', 'Palau',
       'Panama', 'Papua New Guinea', 'Paraguay', 'Peru', 'Philippines',
       'Poland', 'Portugal', 'Puerto Rico', 'Qatar', 'Romania', 'Rwanda',
       'Samoa', 'San Marino', 'Saudi Arabia', 'Senegal', 'Serbia',
       'Seychelles', 'Sierra Leone', 'Singapore',
       'Sint Maarten (Dutch part)', 'Slovak Republic', 'Slovenia',
       'Solomon Islands', 'Somalia', 'South Africa', 'South Sudan',
       'Spain', 'Sri Lanka', 'St. Kitts and Nevis', 'St. Lucia',
       'St. Martin (French part)', 'St. Vincent and the Grenadines',
       'Sudan', 'Suriname', 'Swaziland', 'Sweden', 'Switzerland',
       'Syrian Arab Republic', 'Tajikistan', 'Tanzania', 'Thailand',
       'Timor-Leste', 'Togo', 'Tonga', 'Trinidad and Tobago', 'Tunisia',
       'Turkey', 'Turkmenistan', 'Turks and Caicos Islands', 'Tuvalu',
       'Uganda', 'Ukraine', 'United Arab Emirates', 'United Kingdom',
       'United States', 'Uruguay', 'Uzbekistan', 'Vanuatu', 'Vietnam',
       'West Bank and Gaza', 'Zambia', 'Zimbabwe'], dtype=object)
```

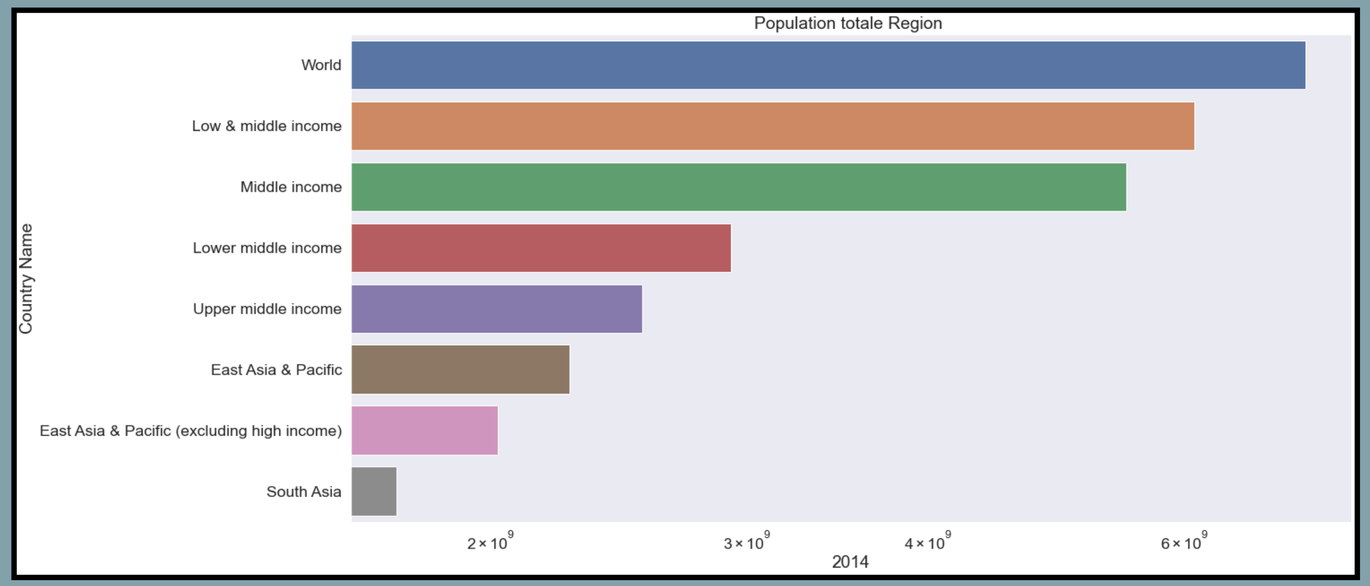
22 Régions

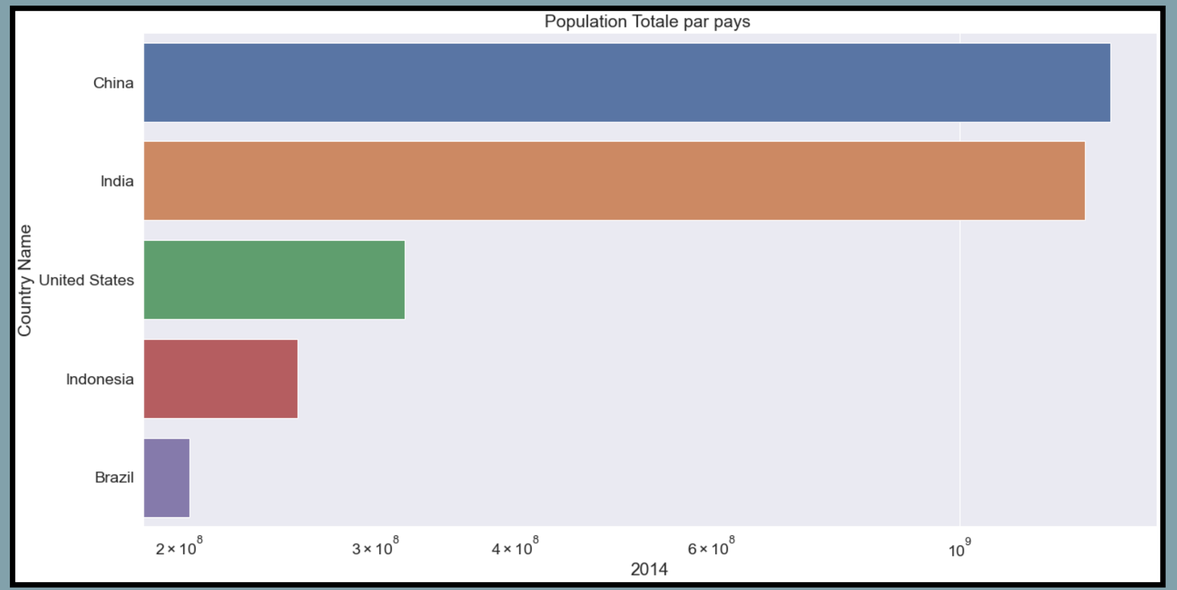
```
array(['East Asia & Pacific',
       'East Asia & Pacific (excluding high income)', 'Euro area',
       'Europe & Central Asia',
       'Europe & Central Asia (excluding high income)',
       'Heavily indebted poor countries (HIPC)', 'High income',
       'Latin America & Caribbean',
       'Latin America & Caribbean (excluding high income)',
       'Least developed countries: UN classification',
       'Low & middle income', 'Low income', 'Lower middle income',
       'Middle East & North Africa',
       'Middle East & North Africa (excluding high income)',
       'Middle income', 'North America', 'South Asia',
       'Sub-Saharan Africa', 'Sub-Saharan Africa (excluding high income)',
       'Upper middle income', 'World'], dtype=object)
```

197 pays selon l'ONU

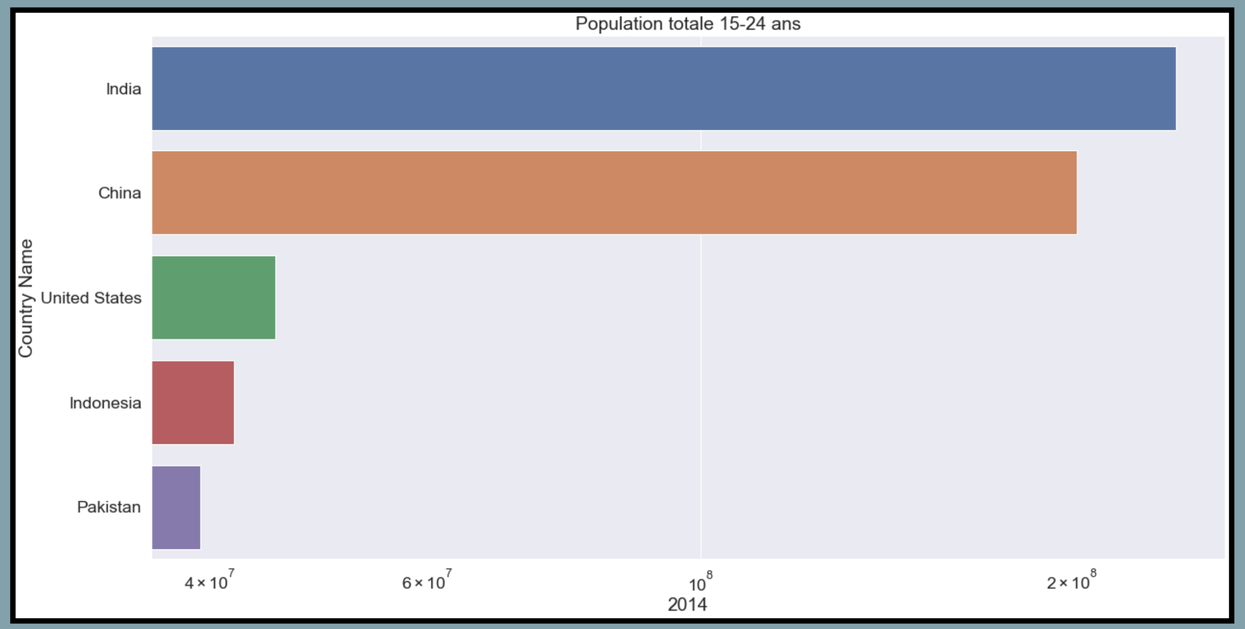
Représentation

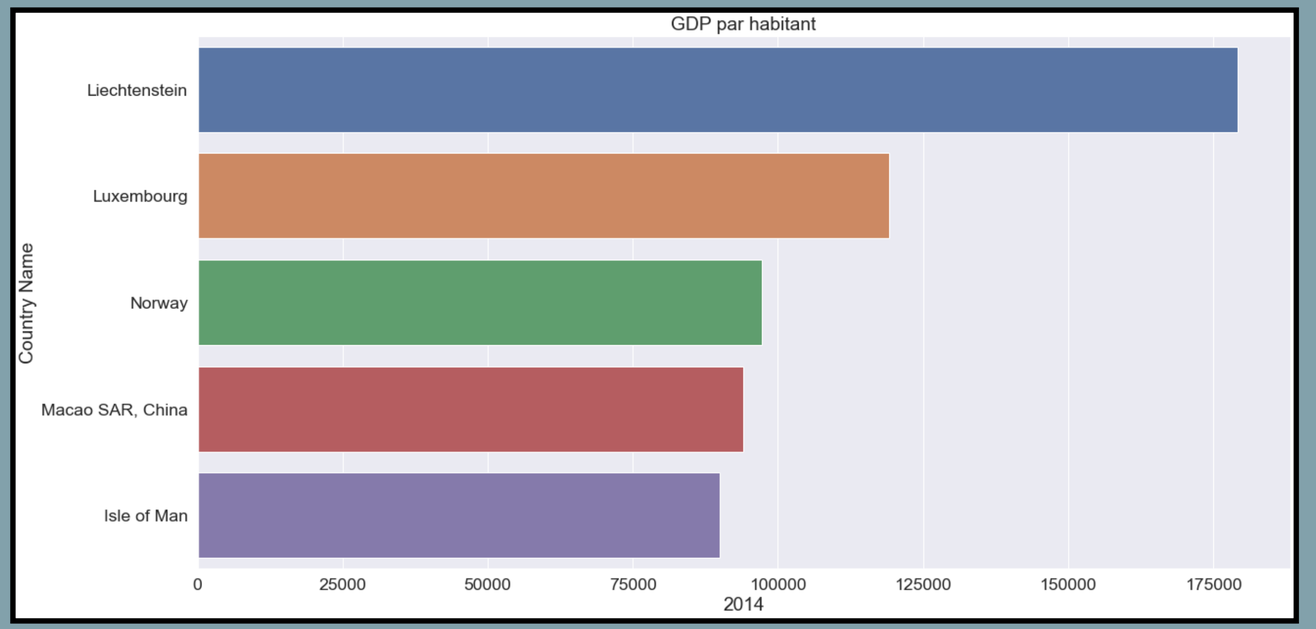
Étude de l'année 2014



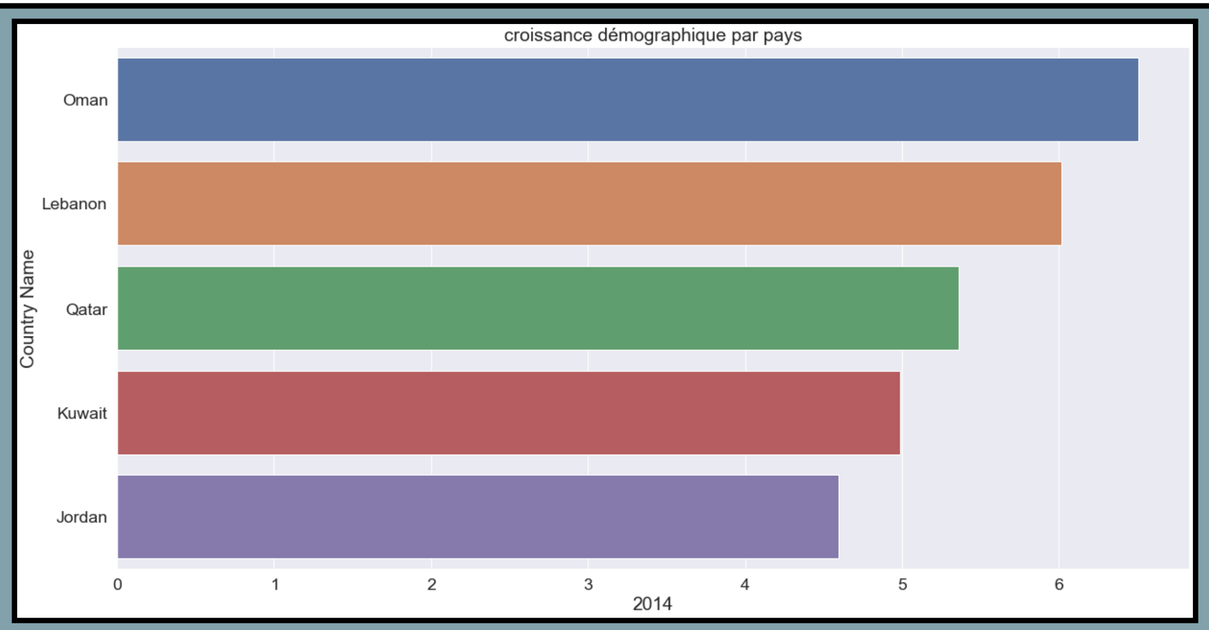


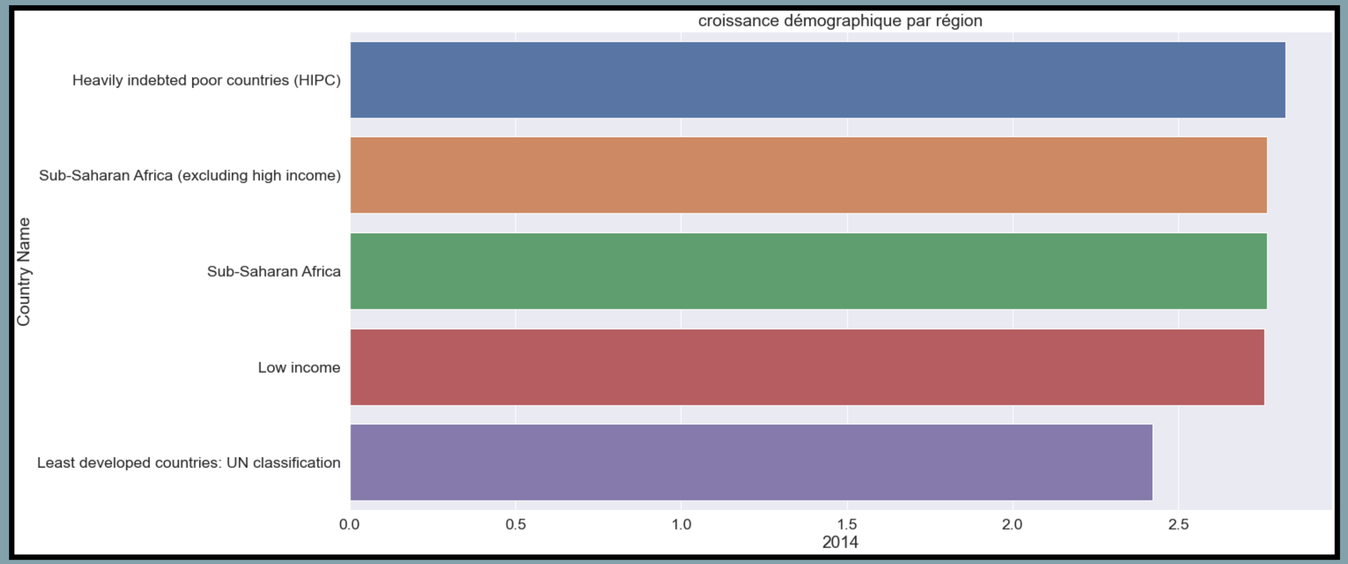
Représentation

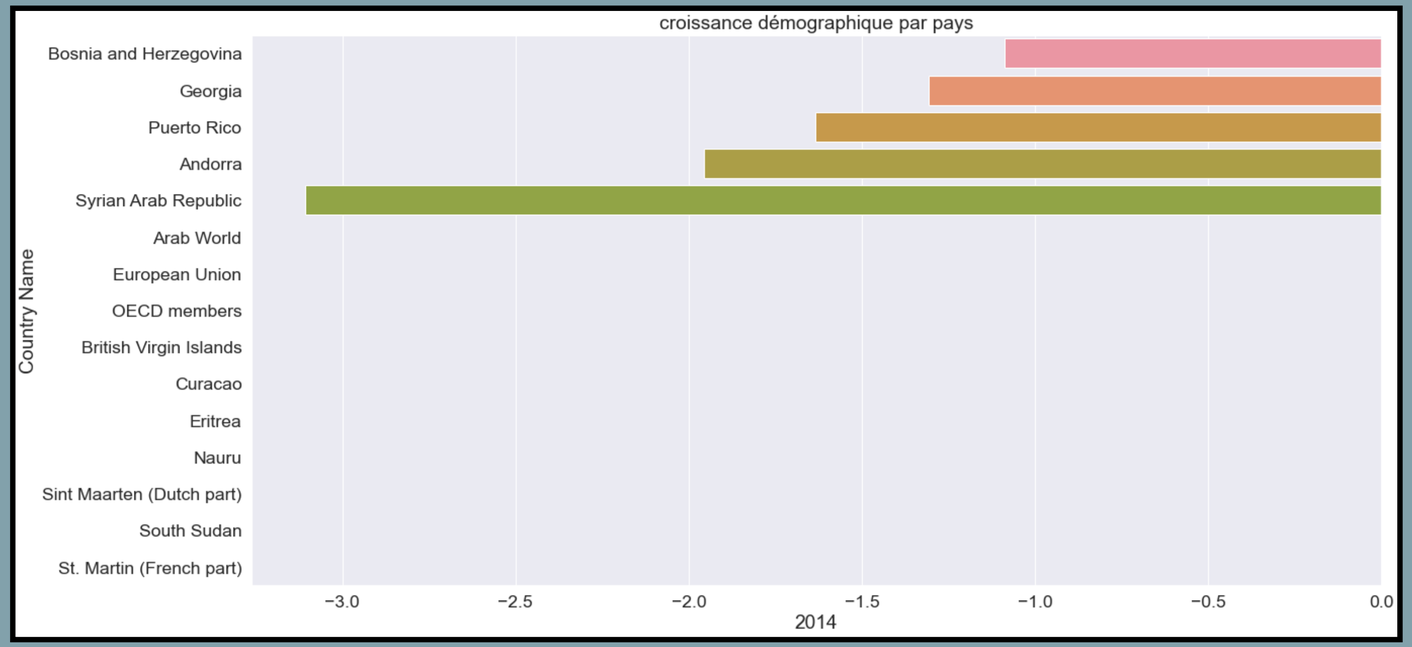




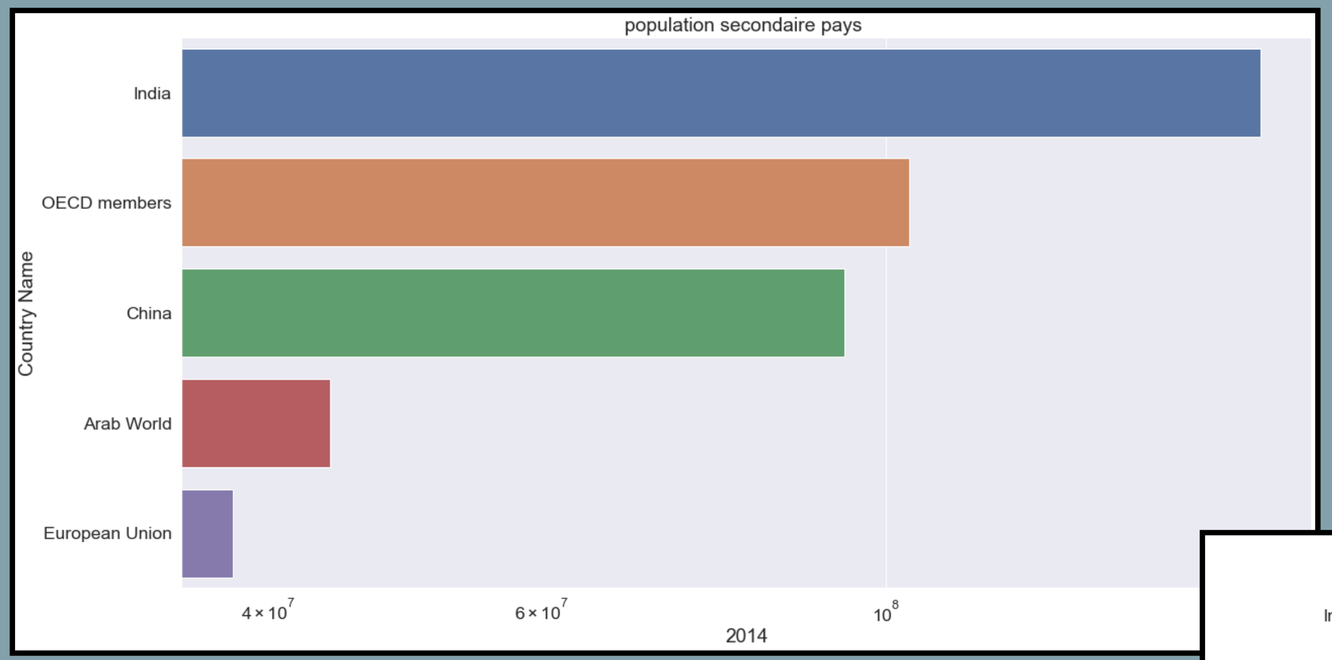
Représentation

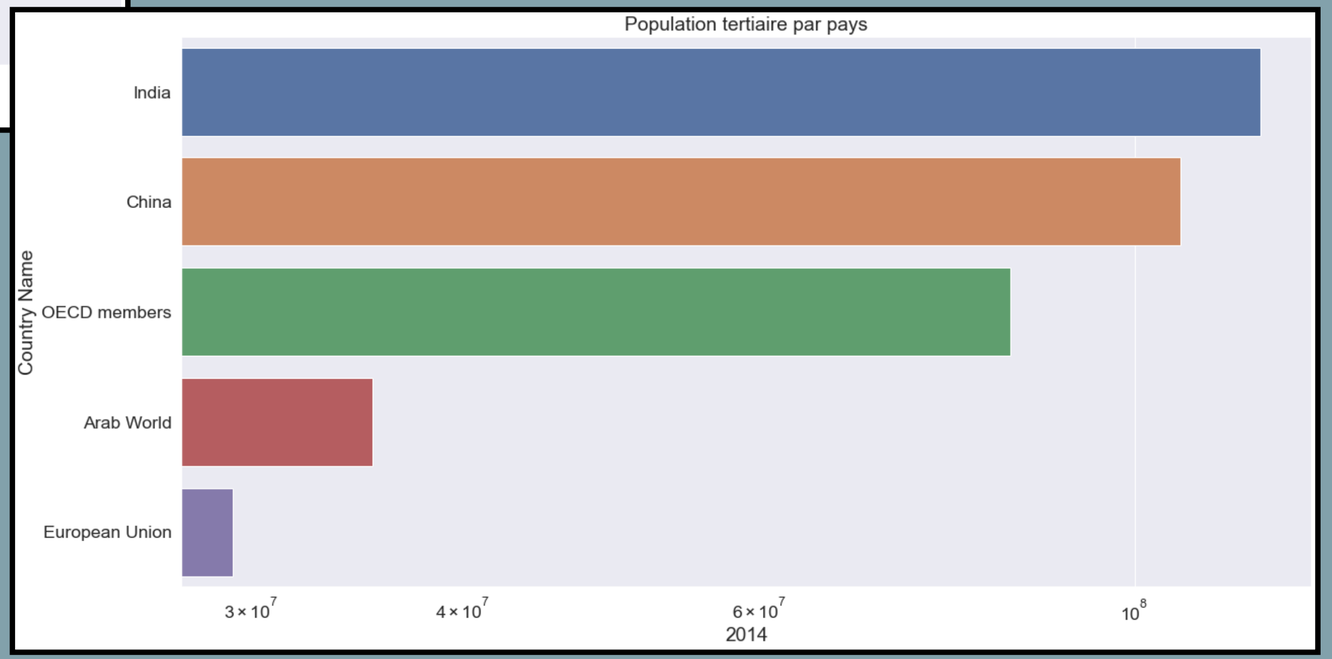




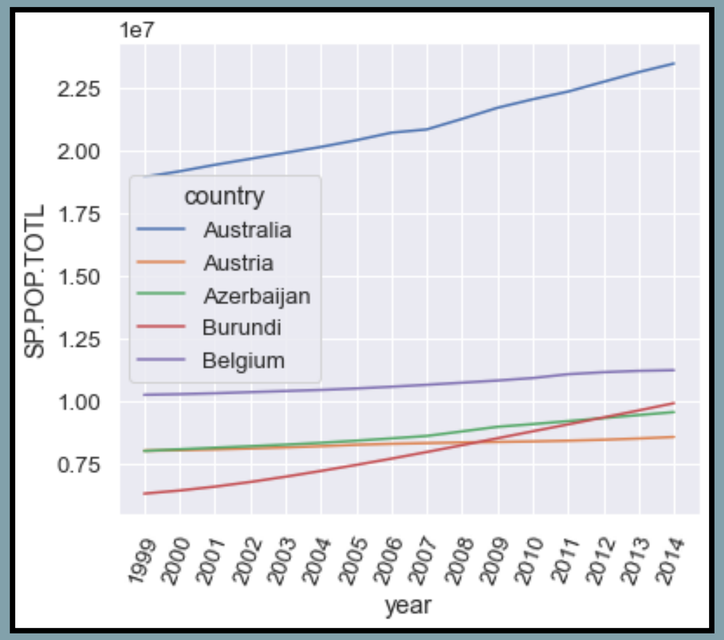


Représentation

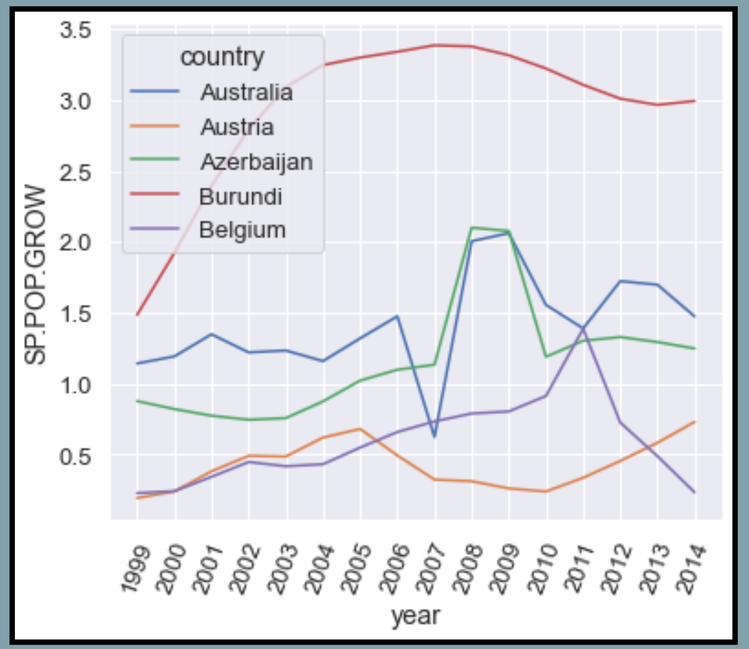


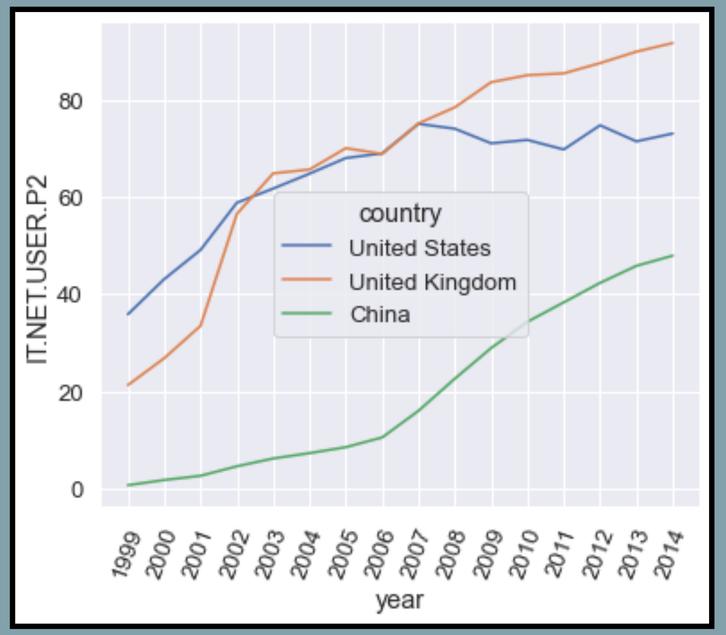


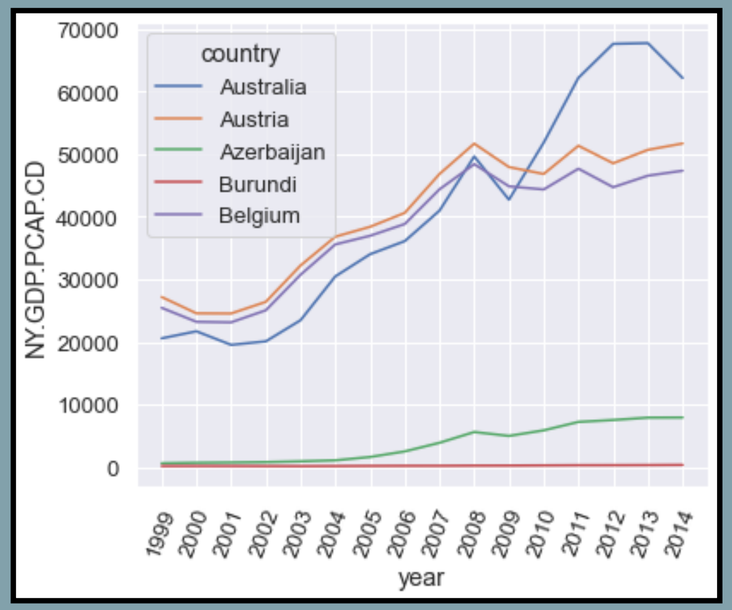
Représentation

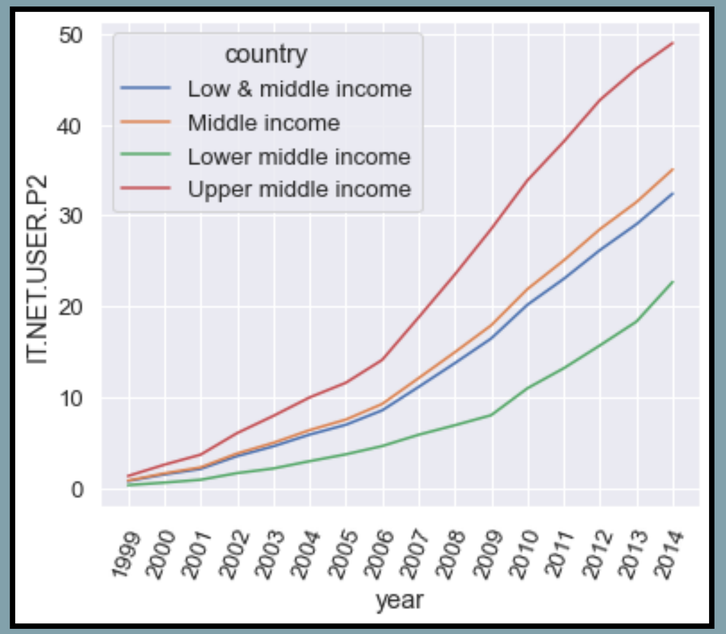


Quelques Évolution temporel
1999-2014

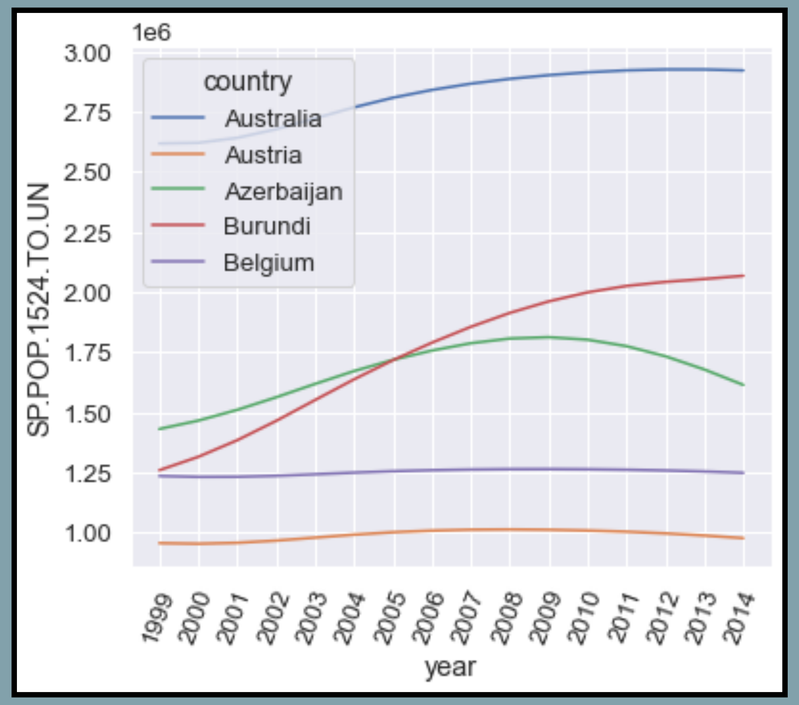


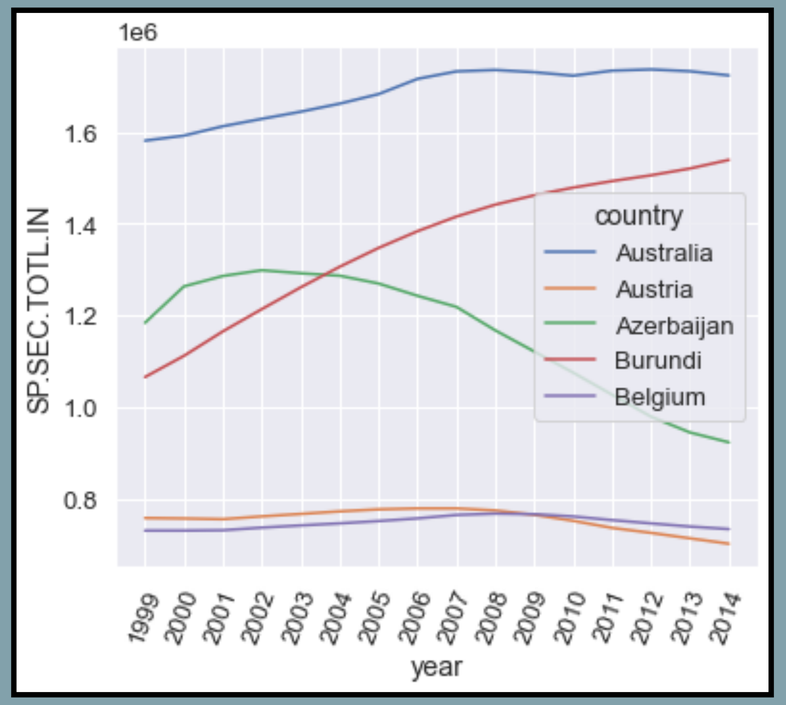


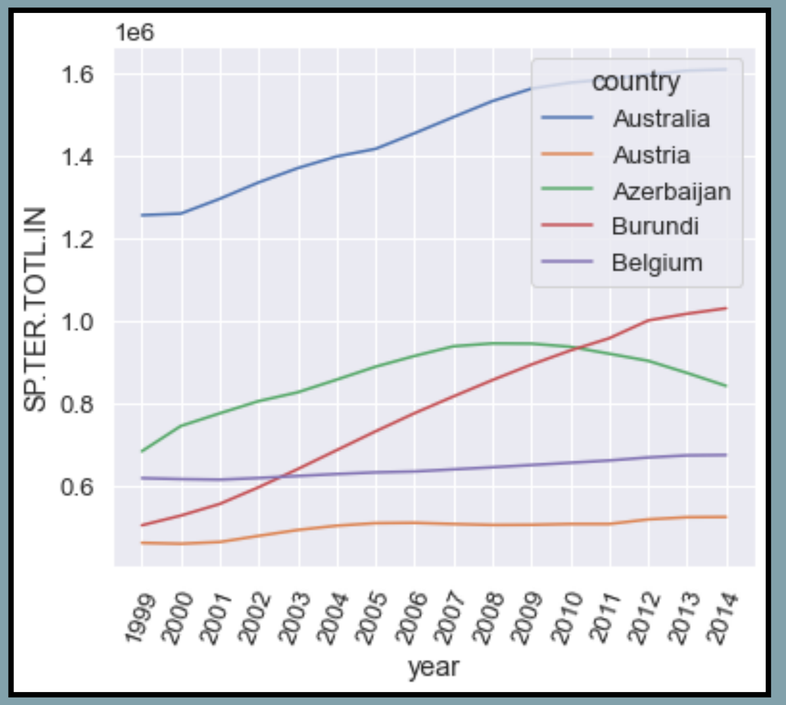




Représentation







Scoring

untry Code	Indicator Name	Indicator Code	1999	2000	2001	2002	2003	2004	...	2006	2007	2008	2009	2010	2011	2012	2013	2014	ref
ARB	GDP per capita (current US$)	NY.GDP.PCAP.CD	2332.1	2615.7	2515.1	2499.3	2754.3	3168.2	...	4414.1	5046.8	6255.8	5267.6	6033.0	7005.5	7571.4	7716.0	NaN	7716.0
ARB	Internet users (per 100 people)	IT.NET.USER.P2	0.6	1.1	1.6	2.7	3.6	7.0	...	11.7	14.2	18.6	23.0	26.7	29.7	34.0	36.9	NaN	36.9

Pour calculer d'abord le score on fais pour les indicateurs une valeur de référence qui nous permettra de remplacer les colonnes qui ne sont pas renseigné par la valeurs de la colonnes qui la précède.

Scoring

Indicator Code	Country Name	IT.NET.USER.P2	NY.GDP.PCAP.CD	SP.POP.1524.TO.UN	SP.POP.GROW	SP.POP.TOTL	SP.SEC.TOTL.IN	SP.TER.TOTL.IN
0	Afghanistan	7.0	612.1	7032072.0	3.2	32758020.0	4676453.0	3034517.0
1	Albania	60.1	4578.7	569427.0	-0.2	2889104.0	345644.0	277193.0

Indicator Code	IT.NET.USER.P2	NY.GDP.PCAP.CD	SP.POP.1524.TO.UN	SP.POP.GROW	SP.POP.TOTL	SP.SEC.TOTL.IN	SP.TER.TOTL.IN
count	229.0	230.0	191.0	240.0	240.0	221.0	220.0
mean	45.3	17367.2	6319050.0	1.3	204583062.5	23329098.2	18634293.0
std	28.2	25913.7	23588520.2	1.3	793202442.8	87785686.5	70045019.0
min	0.0	312.7	2825.0	-3.1	10908.0	1244.0	868.0
25%	17.8	2050.9	295571.5	0.5	1286556.2	250782.0	209243.2
50%	46.2	6570.6	1161310.0	1.2	8452160.0	986751.0	804793.5
75%	69.3	20331.2	4644766.0	2.2	37806368.5	4600755.0	3486137.2
max	98.2	179308.1	243002731.0	6.5	7268986175.7	757088256.0	602372672.0

Le top des pays potentiels pour Academy sans coefficient

Indicator Code	Country Name	scoring
0	India	65.0
1	China	62.1
2	United States	29.2
3	Qatar	27.7
4	Luxembourg	26.7
5	Macao SAR, China	22.9
6	Norway	21.6
7	Kuwait	20.9
8	Switzerland	20.4
9	Nigeria	18.8
10	Brazil	18.6

Conclusion

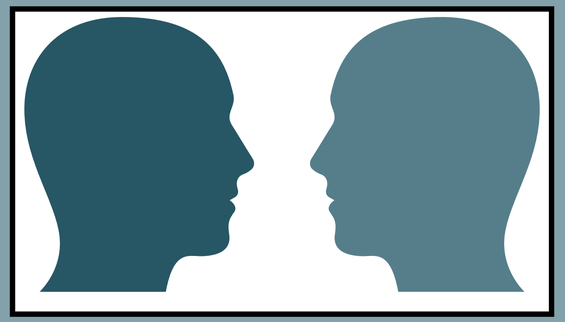
Le top des pays potentiels pour Academy avec coefficient

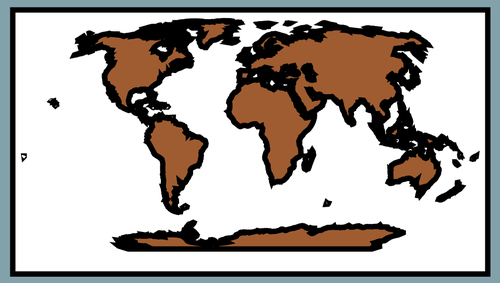
```
coefs = {"NY.GDP.PCAP.CD":3.,
         "SP.POP.1524.TO.UN":1.5,
         "SP.SEC.TOTL.IN":2.,
         "SP.TER.TOTL.IN":2.5,
         "INTPOPTOTALE":4.5,
         "SP.POP.GROW":1.5}
```

	Indicator Code	Country Name	scoring
0		China	24.2
1		India	23.8
2		Luxembourg	19.0
3		Qatar	18.5
4		Macao SAR, China	16.0
5		United States	15.4
6		Norway	15.3
7		Switzerland	14.3
8		Kuwait	13.3
9		Bermuda	13.0
10		Australia	12.1

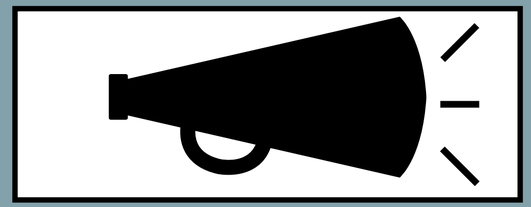
En opérant sur les pays en priorité, Academy peut choisir:

* United States
* Qatar
* Luxembourg
* Norvège
* Suisse





À noter qu'on peut penser que l'entreprise "Academy" peut cibler des clients dans des pays qui seront plus simples, comme par exemple les pays avec la même langue commune (Anglais, Espagnol, Arabe), ou géographiquement proche.



Merci