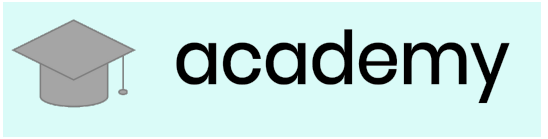
Projet 2 :
Analysez des données de systèmes éducatifs

Rosalba Juarez Mosqueda

01 Introduction

1.1 Description du projet



Academy est une start-up de la EdTech qui propose des contenus de formation en ligne pour un public de niveau lycée et université. L'entreprise a la vision de l'expansion et pour cela, il est nécessaire d'identifier :

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

La mission consiste à effectuer une analyse exploratoire pour déterminer si les données sur l'éducation de la banque mondiale permettent d'informer le projet d'expansion

1.2 Description du jeu de données

5 Dataframes:

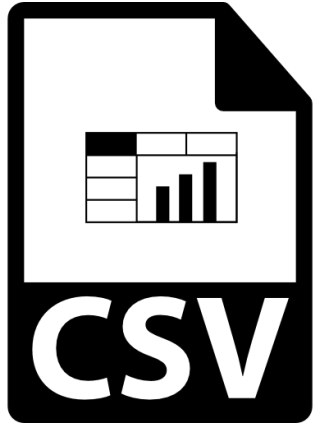
1) **Country (informations détaillées sur les pays) : 241 lignes, 32 colonnes.**

1) **Data (Indicateurs par année de mesure) : 886 930 lignes, 70 colonnes.**

1) **Series (Liste des indicateurs par pays) : 3665 lignes, 21 colonnes**

1) Country-Series (liste des indicateurs par pays) : 613 lignes, 4 colonnes

1) Footnotes (Détails années de mesure des indicateurs) : 643 638 lignes, 5 colonnes

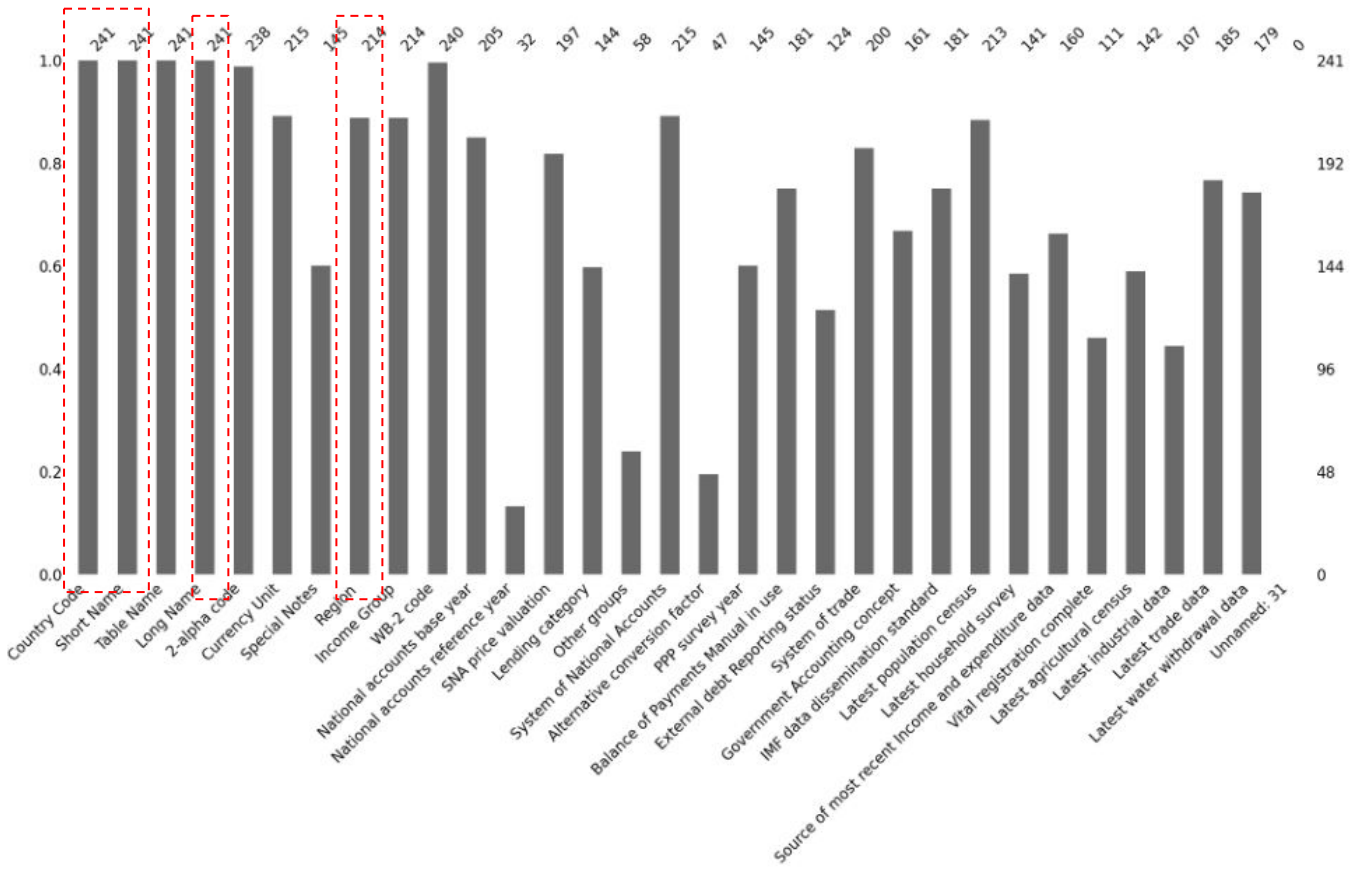


https://datacatalog.worldbank.org/search/dataset/0038480

02
Exploration de données

2.1.1 Description du jeu de données : Country

Country (informations détaillées sur les pays) : 241 lignes , 32 colonnes.



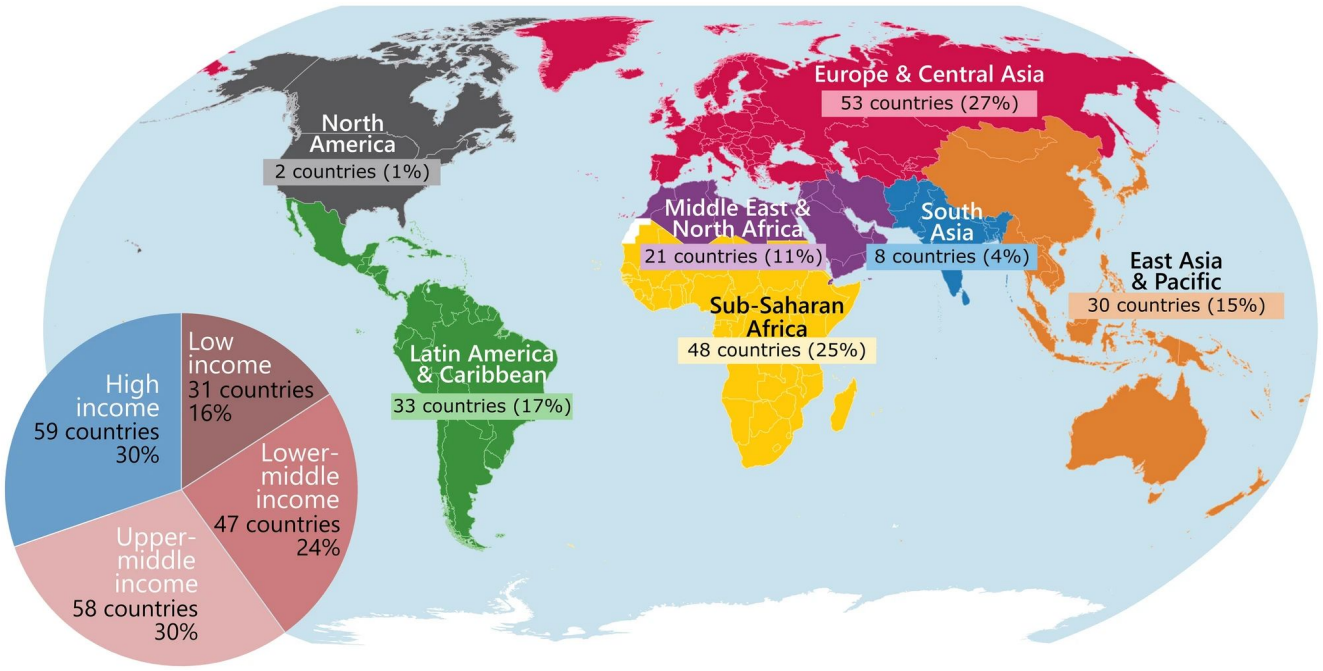
Liste des pays avec leur **nom** complet, leur **abréviation** et le nom de la **région** à laquelle ils appartiennent

Dans ce fichier il y a **214** pays, plus **27** groupements de pays (c.-à-d. qui ne sont pas vraiment un pays en soi, mais des “faux pays” regroupés par la banque mondiale)

2.1.1 Description du jeu de données : Country

"faux pays"

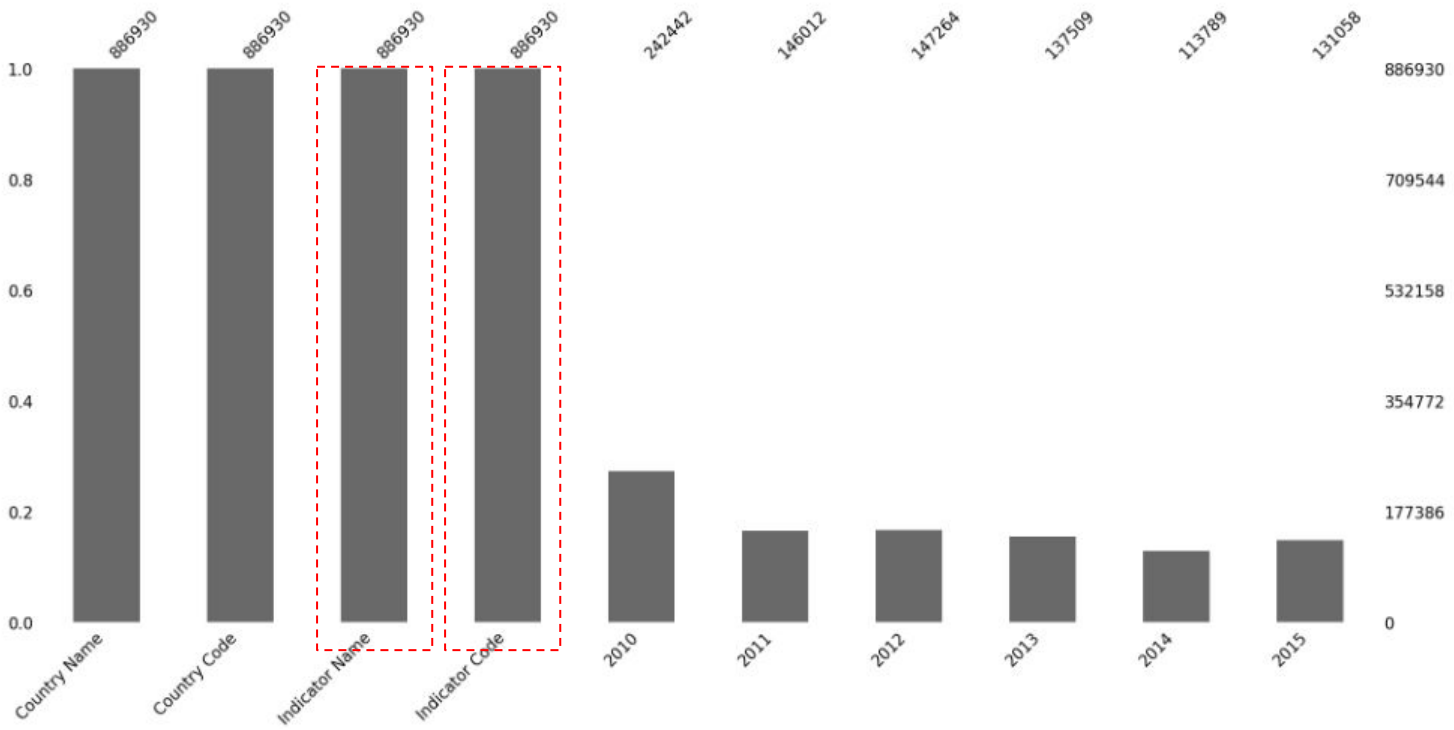
```
                                  Arab World
        East Asia & Pacific (developing only)
          East Asia & Pacific (all income levels)
          Europe & Central Asia (developing only)
        Europe & Central Asia (all income levels)
                                    Euro area
                                European Union
                                    Gibraltar
                                  High income
            Heavily indebted poor countries (HIPC)
      Latin America & Caribbean (developing only)
    Latin America & Caribbean (all income levels)
    Least developed countries: UN classification
                                   Low income
                             Lower middle income
                             Low & middle income
Middle East & North Africa (all income levels)
                                 Middle income
    Middle East & North Africa (developing only)
                                 North America
                                         Nauru
                                  OECD members
                                    South Asia
          Sub-Saharan Africa (developing only)
        Sub-Saharan Africa (all income levels)
                           Upper middle income
                                         World
```



Répartition des pays partenaires par catégorie de revenu et région de la Banque mondiale

2.1.2 Description du jeu de données : Data

Data (Indicateurs par année de mesure) : 886 930 lignes, 70 colonnes



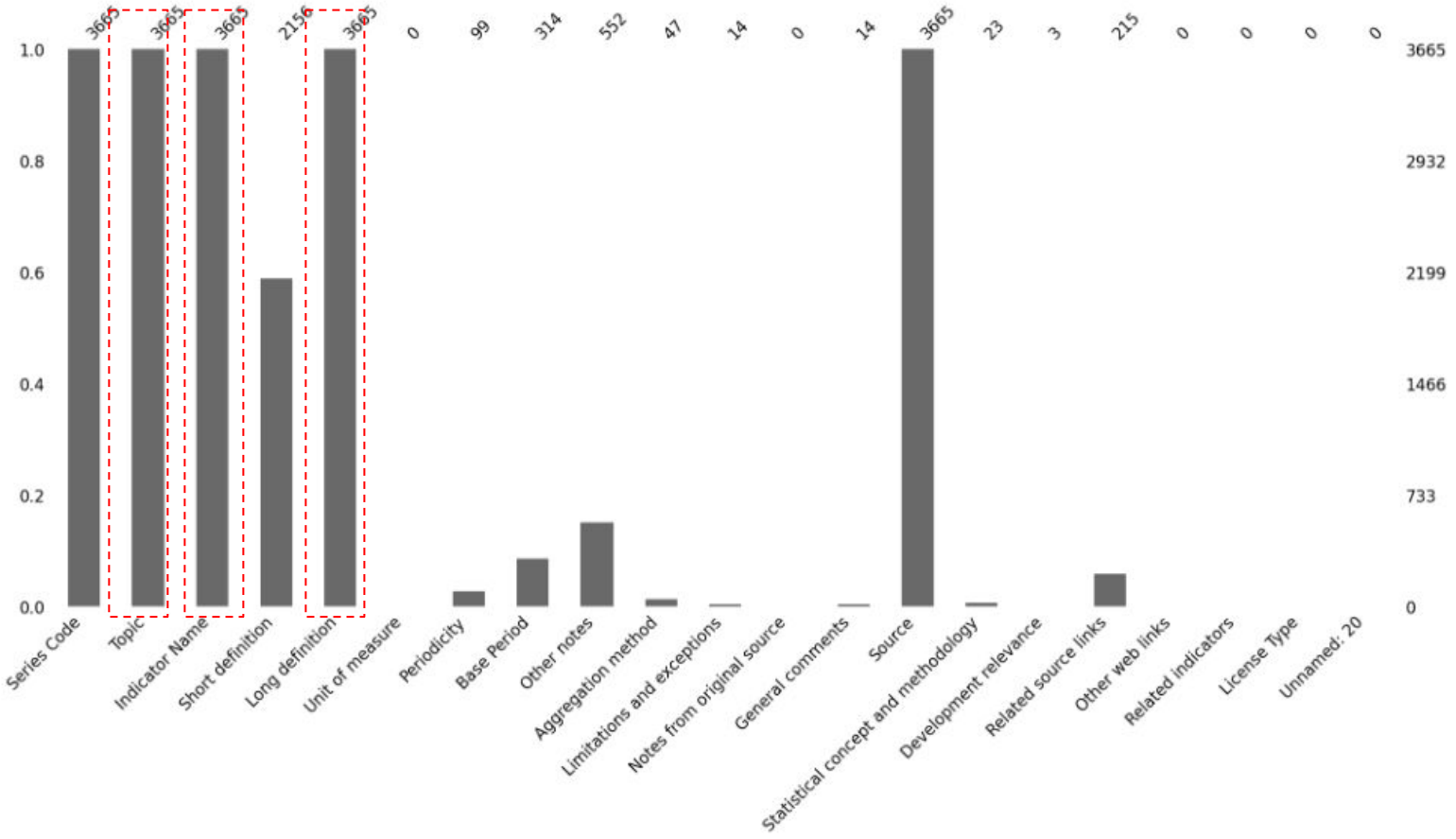
Ce jeu de données contient toutes les informations concernant les **indicateurs,** le nom et les codes des pays, ainsi que l'évolution de tous les indicateurs au fil des années (1970-2100).

Étant donné que **2010 - 2015** sont les années où il y a plus de données, (et où l'accès à Internet a prospéré) nous utiliserons ces années pour évaluer l'évolution des indicateurs d'intérêt.

Afin d'avoir les informations concernant uniquement les vrais pays, dans cette dataframe nous allons également éliminer tous les faux pays identifiés ci-dessus et listés dans la liste "faux pays"

2.1.3 Description du jeu de données : Series

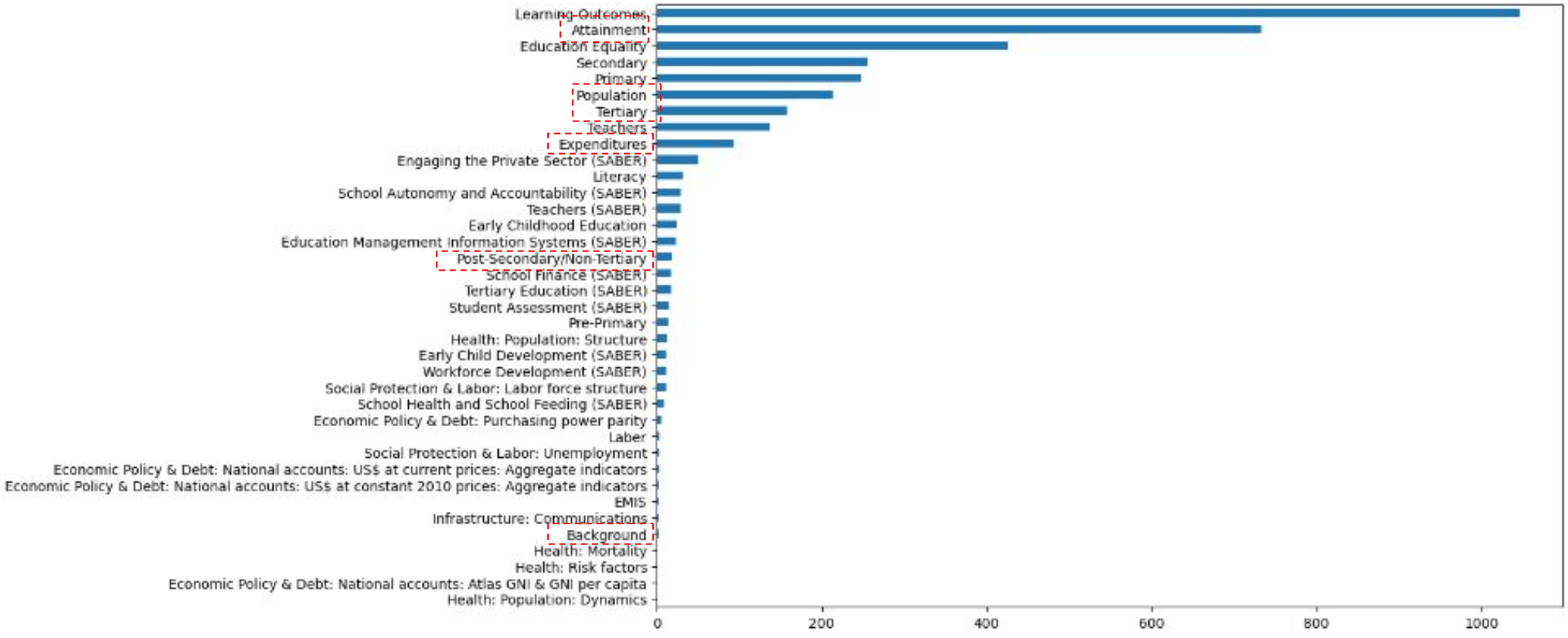
Series (Liste des indicateurs par pays) : 3665 lignes, 21 colonnes



Ce jeu de données donne des **informations plus détaillées sur les indicateurs** ; leur nom, leur définition et le sujet (Topic) auquel ils appartiennent

2.1.3 Description du jeu de données : Series

37 topics (catégories)



2.2 Sélection des indicateurs d'intérêt

Au total, il y a 3665 indicateurs, parmi eux on retrouve ceux qui peuvent être d'un grand intérêt pour notre analyse

07 Expenditures

- Government expenditure on education as % of GDP
- Government expenditure on secondary education as % of GDP (%)
- Government expenditure on tertiary education as % of GDP (%)
- Government expenditure on upper secondary education as a percentage of GDP (%)

05 Tertiary

- Enrolment in tertiary education, all programmes, both sexes (number)

04 Vocational & Post-secondary Non-Tertiary

- Gross enrolment ratio, post-secondary non-tertiary, both sexes (%)

12 Population

- Population, total
- School age population, secondary education, both sexes (number)
- School age population, tertiary education, both sexes (number)
- School age population, upper secondary education, both sexes (number) --> School age population, lower secondary education, both sexes (number) in Mexico was reported at 6710162 Persons in 2020.

14 Background

- GDP per capita (constant 2005 US$)
- Internet users (per 100 people)
- Personal computers (per 100 people)

2.2 Sélection des indicateurs d'intérêt

Accès à l'apprentissage en ligne

Internet users (per 100 people)

Indicateur économique (produit intérieur brut (PIB)) Mesure du niveau de vie (pouvoir d'achat)

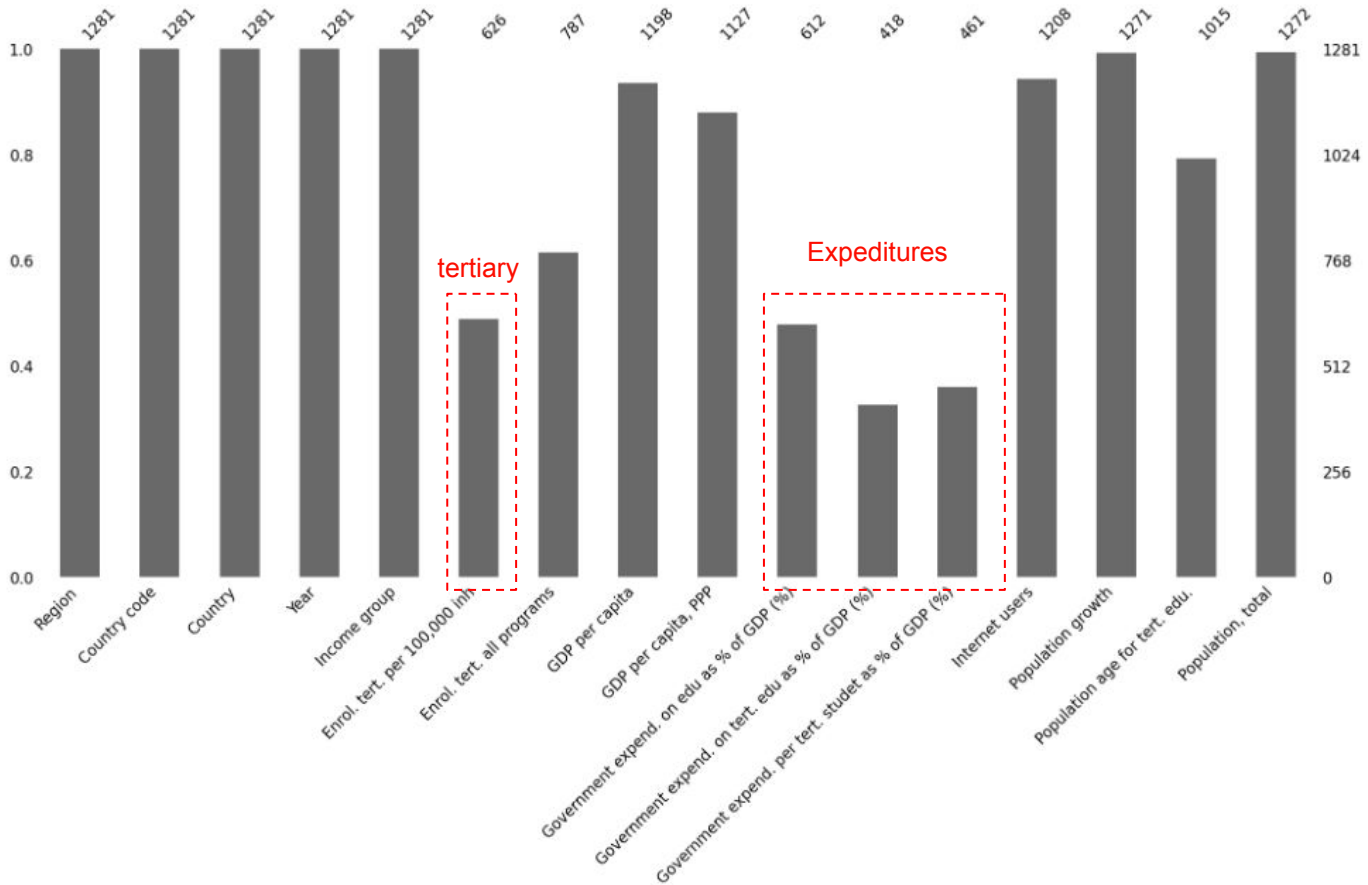
GDP per capita (current US$)
GDP per capita, PPP (current international $)
Government expenditure on education as % of GDP (%)
Government expenditure on tertiary education as % of GDP (%)
Government expenditure per tertiary student as % of GDP per capita (%)

Informations démographiques

Population, total
Population growth (annual %)
Population of the official age for tertiary education, both sexes (number)
Enrolment in tertiary education, all programmes, both sexes (number)
Enrolment in tertiary education per 100,000 inhabitants, both sexes

03
Préparation des données

3.1 Data Frame avec la liste des indicateurs d'intérêt

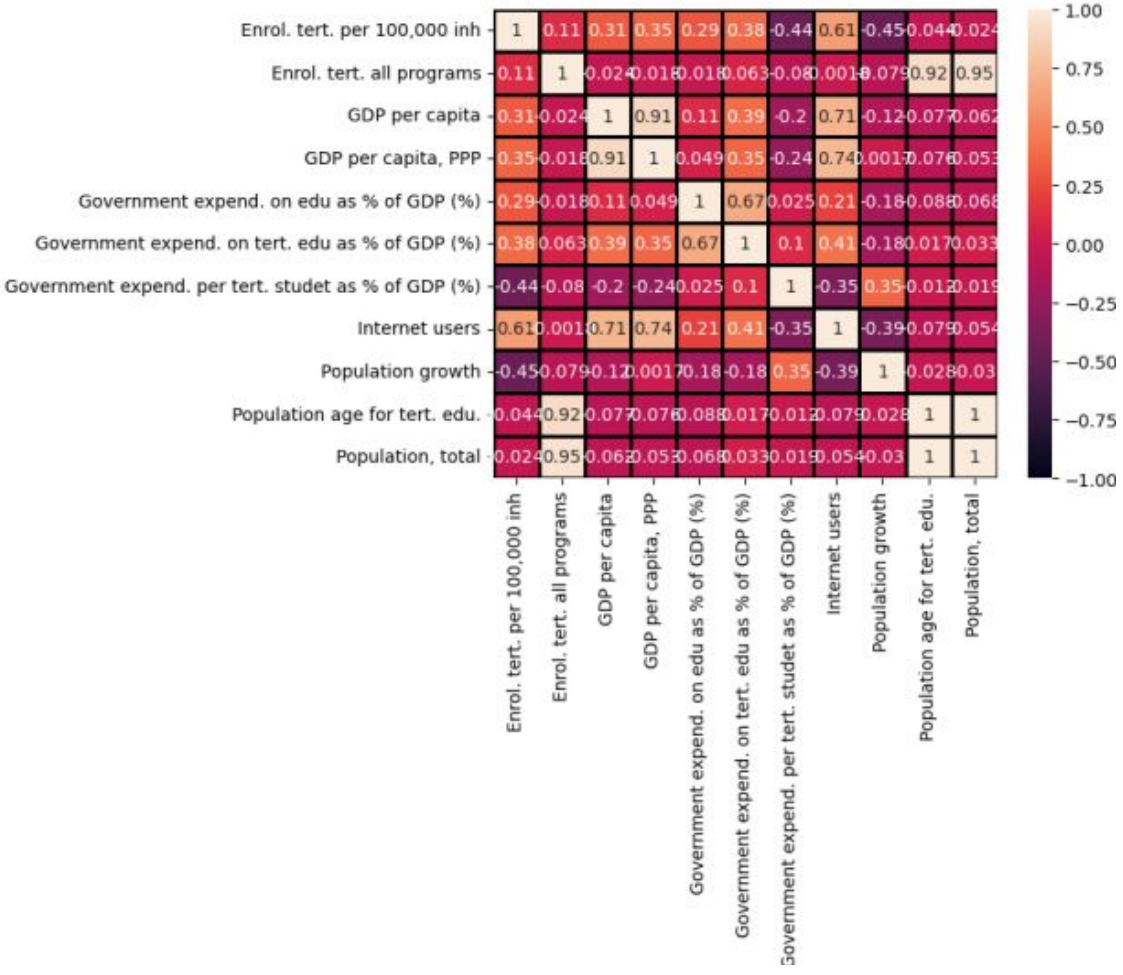


valeurs manquantes valeurs manquantes lors de l'utilisation de cette sélection d'indicateurs

Region	0.000000
Country code	0.000000
Country	0.000000
Year	0.000000
Income group	0.000000
Population, total	0.702576
Population growth	0.780640
Internet users	5.698673
GDP per capita	6.479313
GDP per capita, PPP	12.021858
Population age for tert. edu.	20.765027
Enrol. tert. all programs	38.563622
Enrol. tert. per 100,000 inh.	51.131928
Government expend. on edu as % of GDP (%)	52.224824
Government expend. per tert. studet as % of GDP (%)	64.012490
Government expend. on tert. edu as % of GDP (%)	67.369243

Les indicateurs de la catégorie "Expeditures" et "Teriary" n'atteignent pas 50% en termes d'informations disponibles.

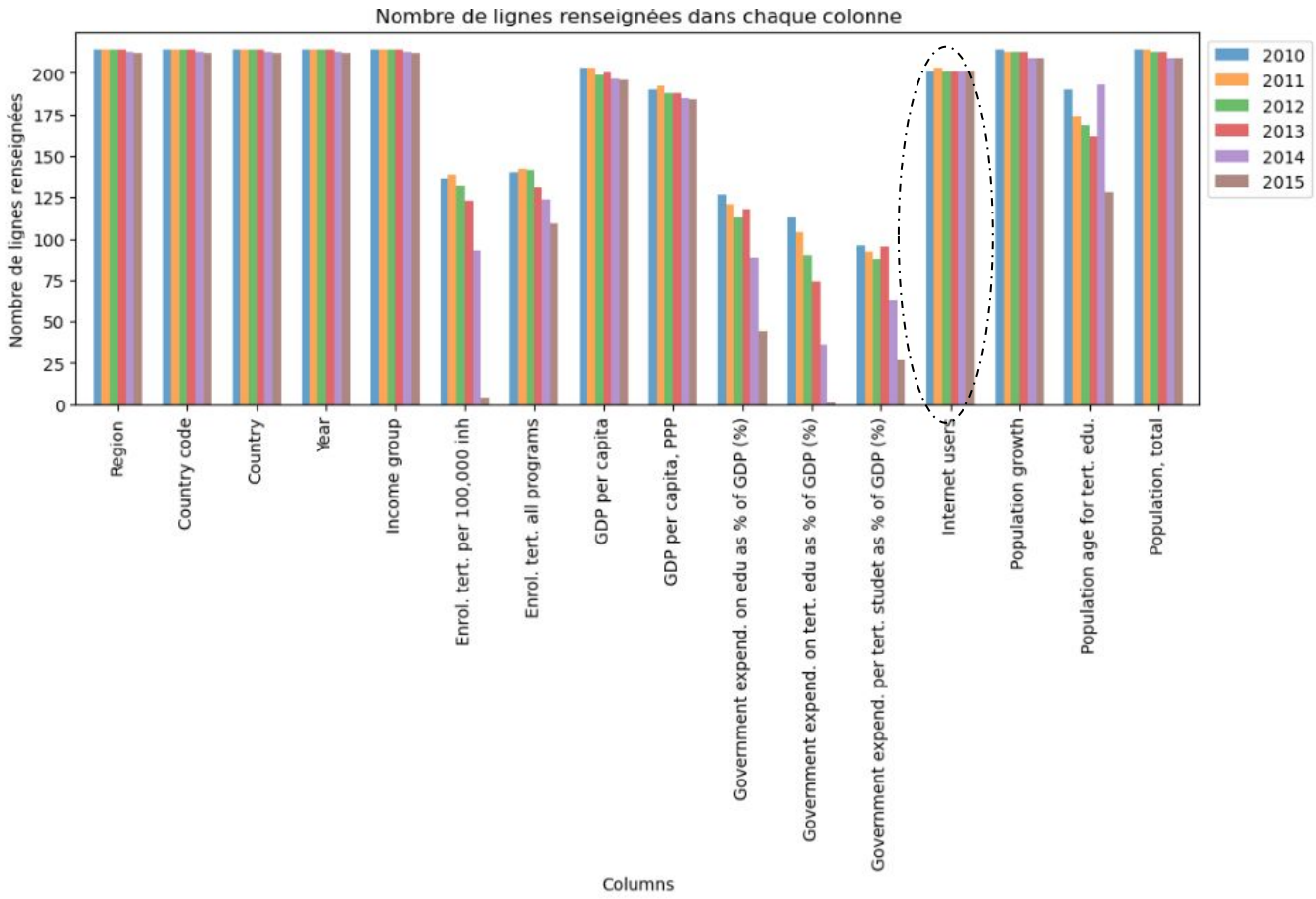
3.2 Corrélation entre les indicateurs sélectionnés



De cette analyse simple, nous pouvons rapidement remarquer qu'il existe une corrélation entre:

- "PIB par habitant" et le "PIB par habitant, PPP"(0.91)
- “Population d'âge officiel pour l'enseignement supérieur” et “Inscriptions dans l'enseignement supérieur” (0.92)
- "Population d'âge officiel pour l'enseignement supérieur " et "Population, total" (~1.00)

3.3 Valeurs manquantes par indicateur et par année

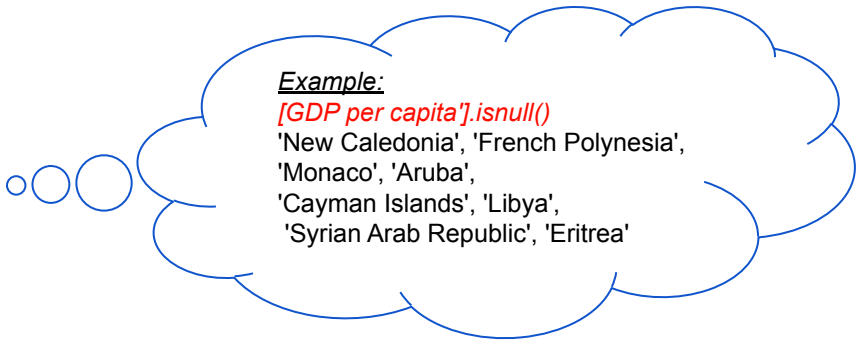


Étant donné que mon indicateur discriminant (le plus important) est l'accès à Internet et que cette colonne est presque également complète toutes les années, j'analyse le degré de remplissage dans le reste des colonnes pour décider quelles années sont les plus appropriées à prendre en compte.
Dans ce cas je comparerai l'évolution de mes indicateurs en analysant les années 2010 et 2013

3.4 Stratégies pour éliminer les valeurs manquantes

Pour travailler avec un jeu de données complet, les stratégies que j'utilise pour éliminer les valeurs manquantes sont :

1) dropna (pandas)

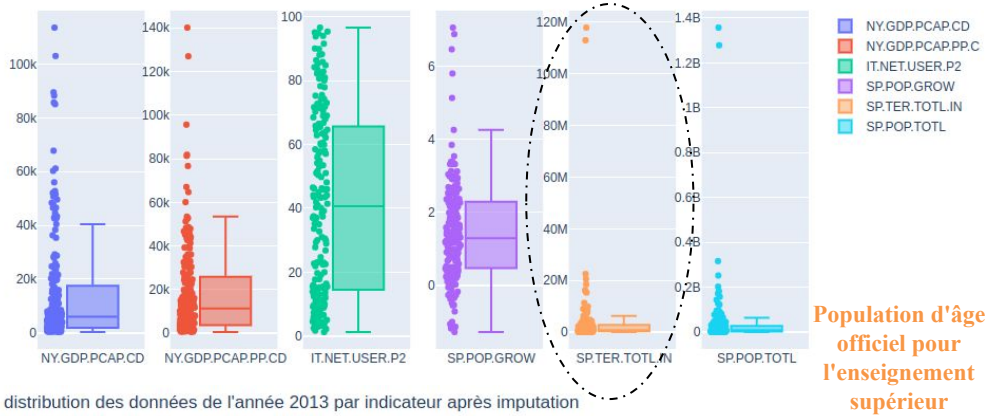


2) Nearest neighbors imputation: sklearn.impute.KNNImputer avec valeur moyenne de n_neighbors=3

3.5 Répartition des données de chaque indicateur par année avant et après imputation

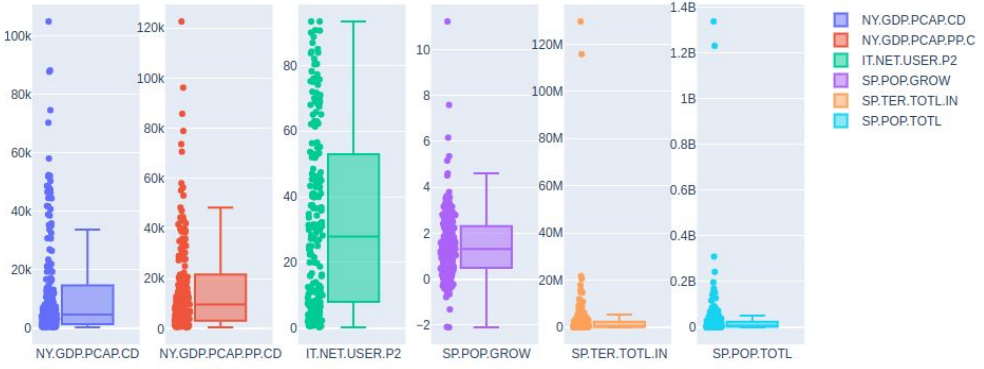
distribution des données de l'année 2010 par indicateur avant imputation

distribution des données de l'année 2013 par indicateur avant imputation

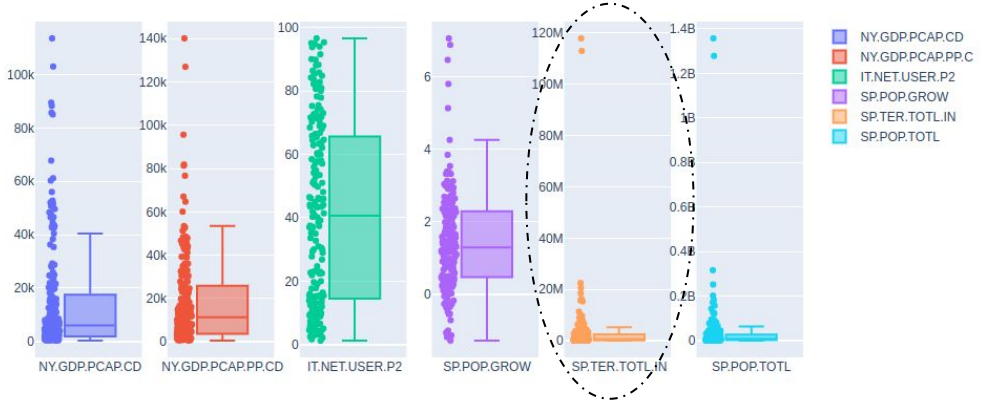


Population d'âge officiel pour l'enseignement supérieur

distribution des données de l'année 2010 par indicateur après imputation



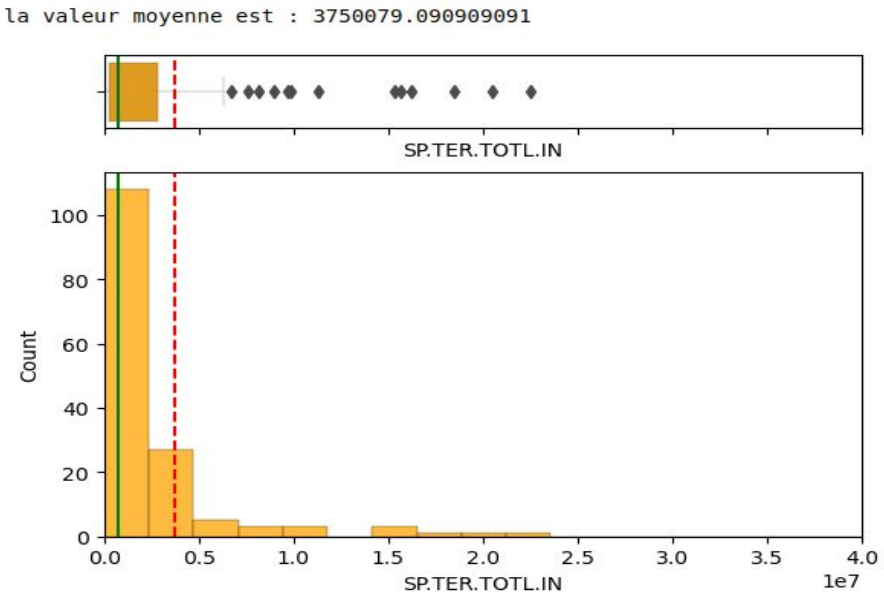
distribution des données de l'année 2013 par indicateur après imputation



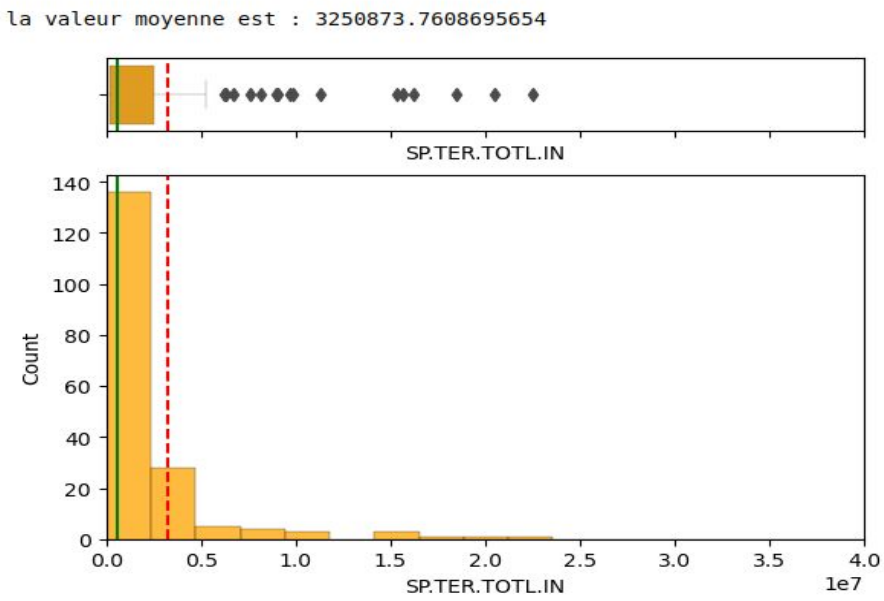
NY.GDP.PCAP.CD: GDP per capita (current US$) , **NY.GDP.PCAP.PP.CD : GDP per capita, PPP (current international $)** , **IT.NET.USER.P2 : Internet users (per 100 people)**, **SP.POP.GROW: Population growth (annual %)**, **SP.TER.TOTL.IN: Population of the official age for tertiary education, both sexes (number)** **SP.POP.TOTL: Population, total**

3.5 Distribution des données dans l'indicateur SP.TER.TOTL.IN dans année 2013

Avant imputation



Après imputation

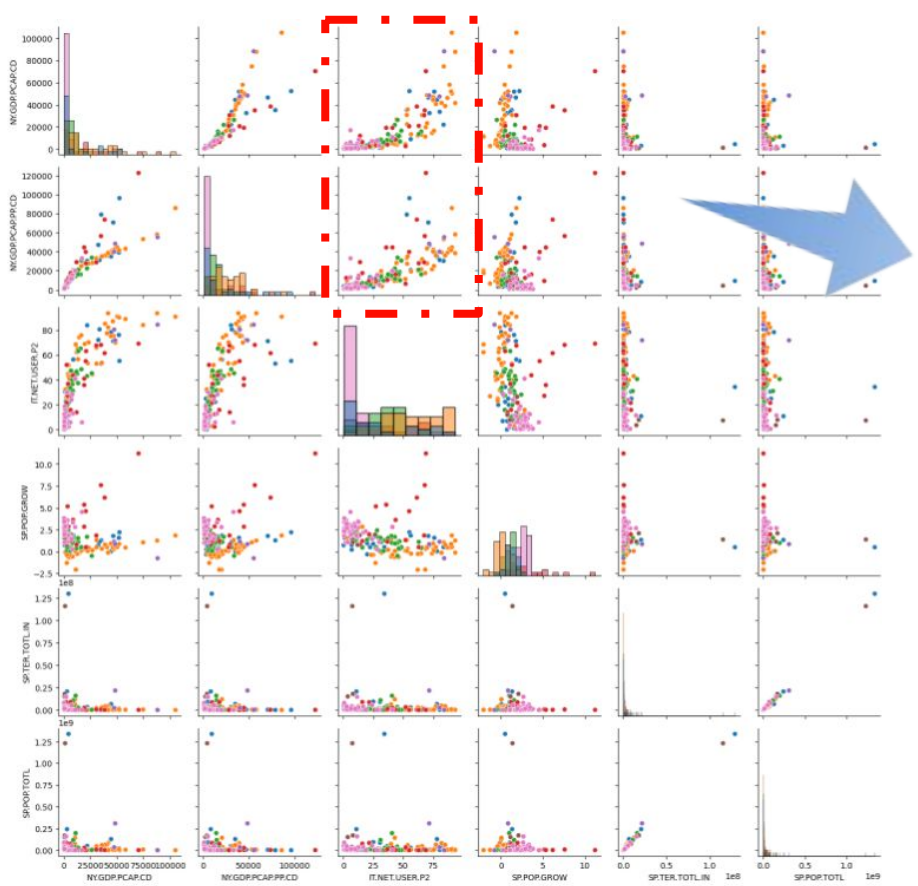


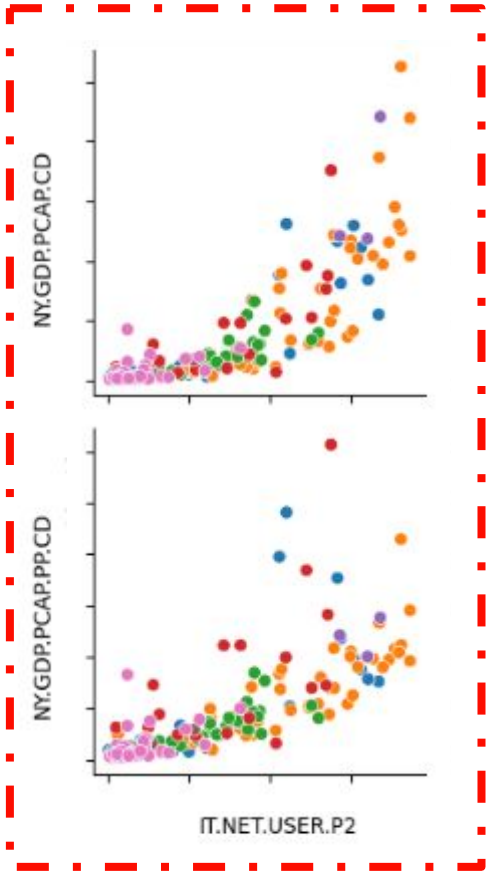
L'amputation des valeurs a fait baisser la moyenne de 15 %

SP.TER.TOTL.IN: Population d'âge officiel pour l'enseignement supérieur

04
Analyse

4.1 Analyse multivariée : Corrélations entre les indicateurs après imputation



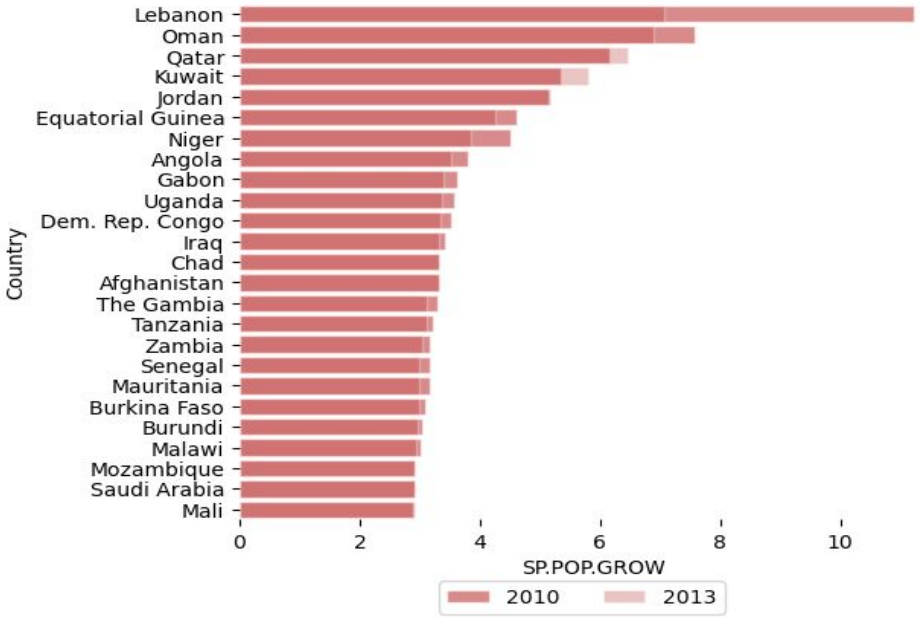


Il existe des relations étroites entre le produit intérieur brut (GDP) et l'accès à Internet

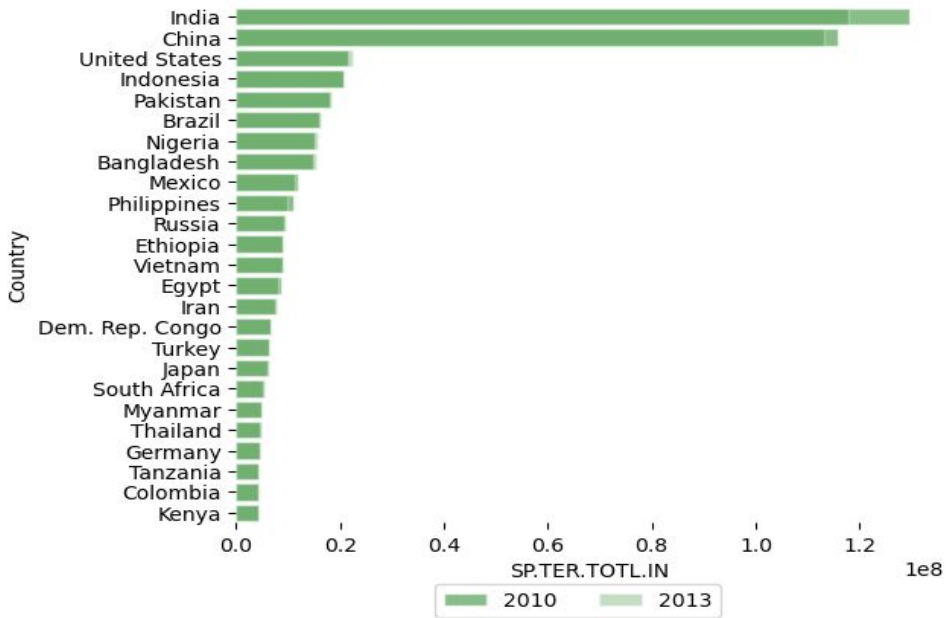
T.NET.USER.P2 : Internet users (per 100 people), **NY.GDP.PCAP.CD:** GDP per capita (current US$) , **NY.GDP.PCAP.PP.CD :** GDP per capita, PPP (current international $)

4.2 Analyse de l'évolution des indicateurs par pays: 2013 Vs 2010

Top 25 des pays avec la plus forte croissance démographique (en pourcentage)



Top 25 pays dont la population en âge d'étudier dans l'enseignement supérieur est la plus élevée

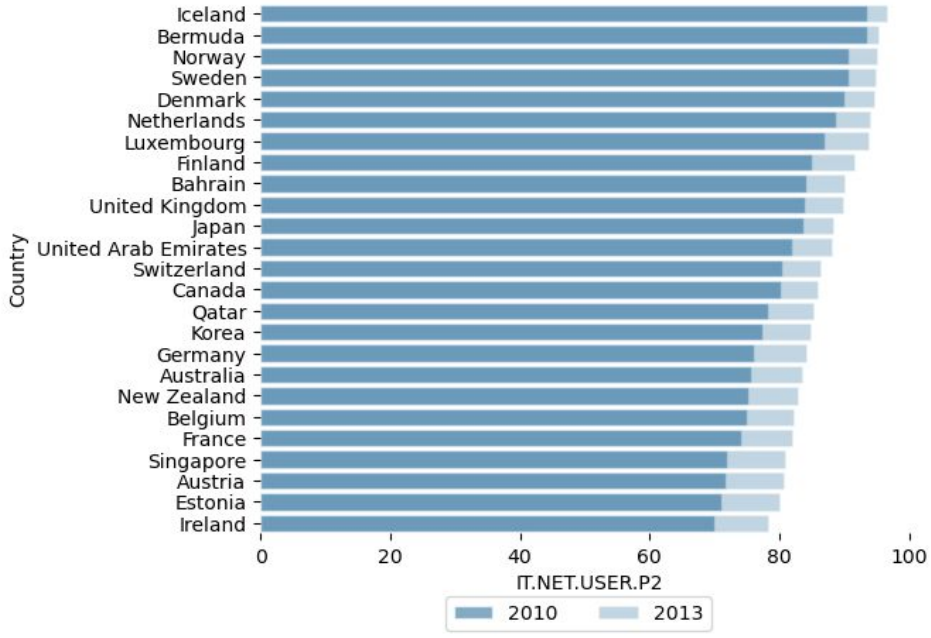


SP.POP.GROW: Population growth

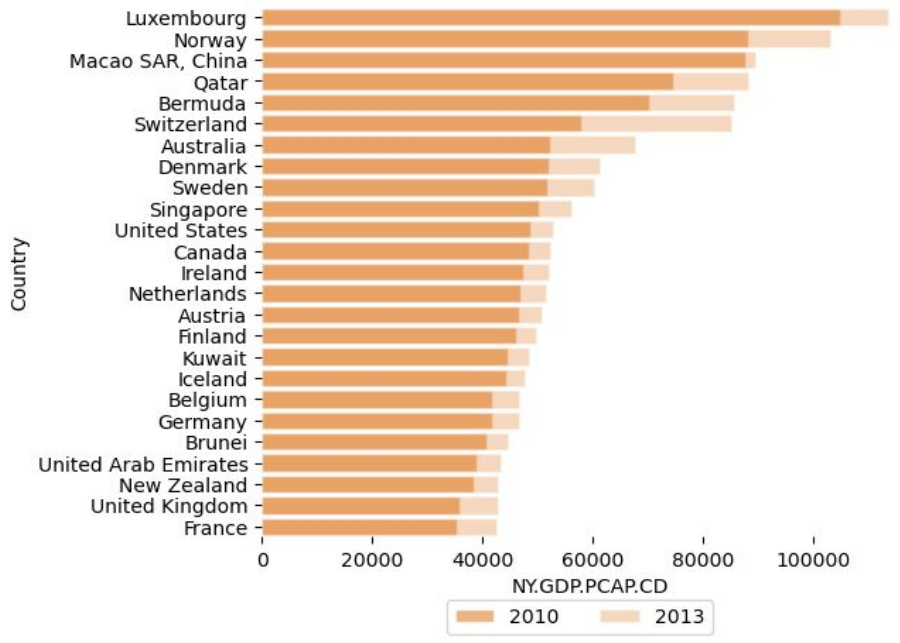
SP.TER.TOTL.IN: Population age for tert. edu

4.2 Analyse de l'évolution des indicateurs par pays: 2013 Vs 2010

Top 25 pays avec le plus d'accès à Internet



Top 25 pays vingt pays avec le plus grand GDP per capita



T.NET.USER.P2 : Internet users

NY.GDP.PCAP.CD: GDP per capita

4.3 Modèle pour identifier pays avec le plus grand potentiel

SCORE = ⅓ *

IT.NET.USER.P2

Accès à l'apprentissage en ligne

+

(NY.GDP.PCAP.CD + NY.GDP.PCAP. PP. CD) / 2

Indicateur économique
(produit intérieur brut (PIB))
Mesure du niveau de vie
(pouvoir d'achat)

+

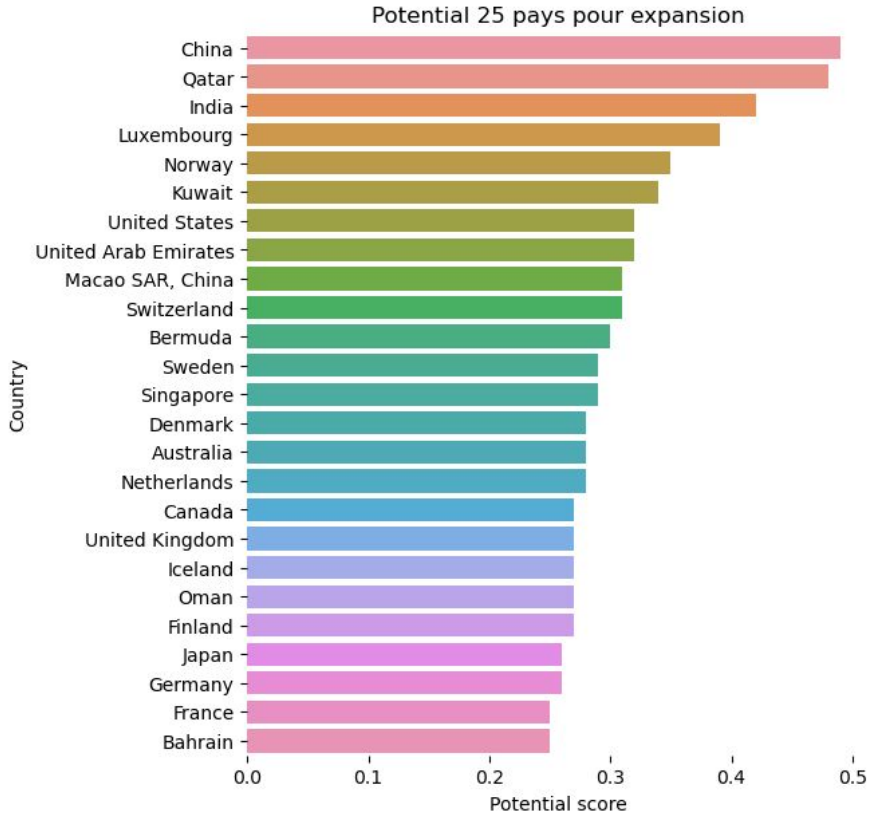
SP.POP. GROW + SP.TER.TOTL.IN + SP.POP.TOTL

Informations démographiques
(population cible)

IT.NET.USER.P2 : Internet users (per 100 people)
NY.GDP.PCAP.CD: GDP per capita (current US$
NY.GDP.PCAP.PP.CD : GDP per capita, PPP (current international $)
SP.POP.GROW: Population growth (annual %)
SP.TER.TOTL.IN: Population of the official age for tertiary education, both sexes (number)
SP.POP.TOTL: Population, total

4.5 Les 25 pays avec un fort potentiel pour acquérir les services d'Academy

Score = (IT.NET.USER.P2 + (NY.GDP.PCAP.CD + NY.GDP.PCAP. PP. CD)/2 + (SP.POP. GROW + SP.TER.TOTL.IN+ SP.POP.TOTL)) / 5



4.5 Le TOP 10 pays avec un fort potentiel pour acquérir les services d'Academy

05
Conclusions

5.1 Les 10 pays avec un fort potentiel de clients pour les services sont :

Chine :

- *population total*
- *Population ayant l'âge officiel d'accéder à l'enseignement supérieur*
- *Utilisateurs de l'internet*

Inde :

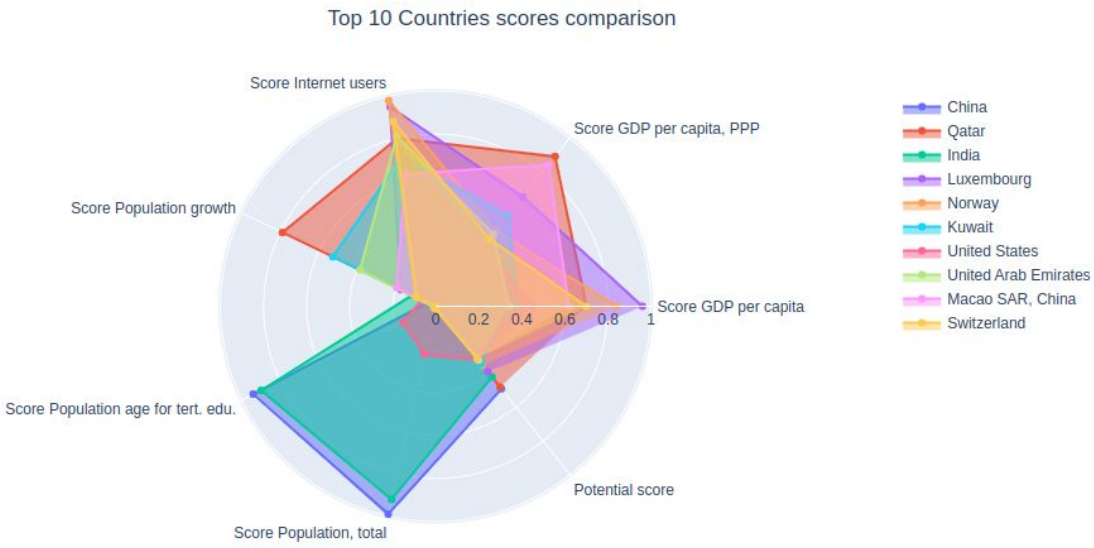
- *population total*
- *Population ayant l'âge officiel d'accéder à l'enseignement supérieur*

Qatar :

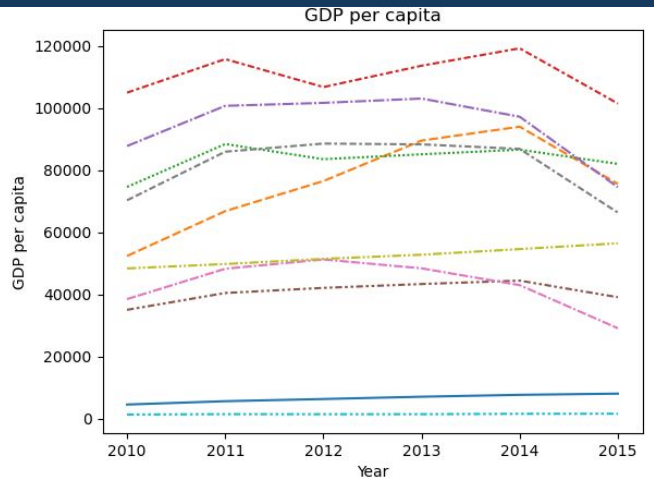
- *PIB par habitant (GDP per capita, PPP)*
- *Utilisateurs de l'internet*

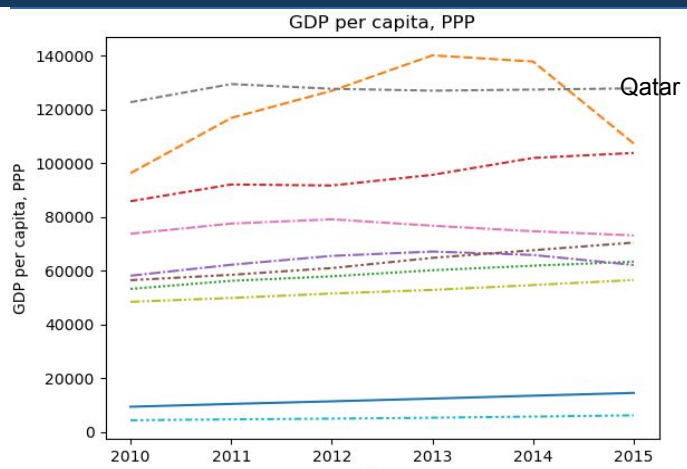
Luxembourg, Norvège, Kuwait, États-Unis, Émirats arabes unis, RAS de Macao Chine, Suisse

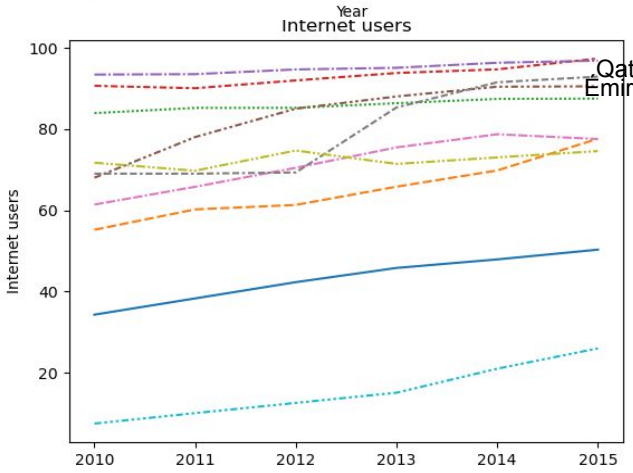
- *PIB par habitant (GDP)*
- *Utilisateurs de l'internet*

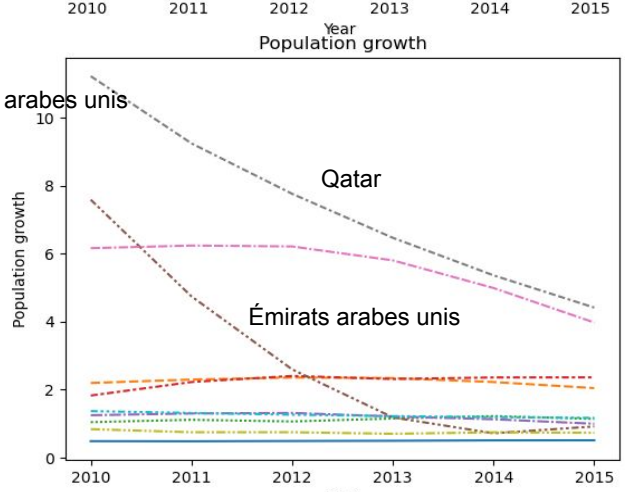


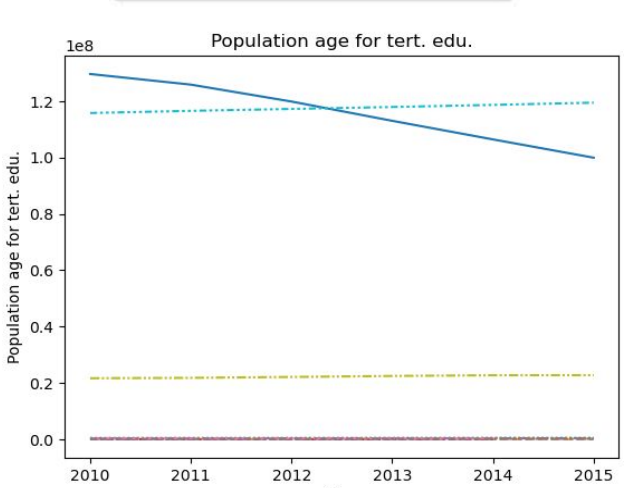
5.2 Pour chacun de ces pays, quelle sera l'évolution de ce potentiel de clients ?









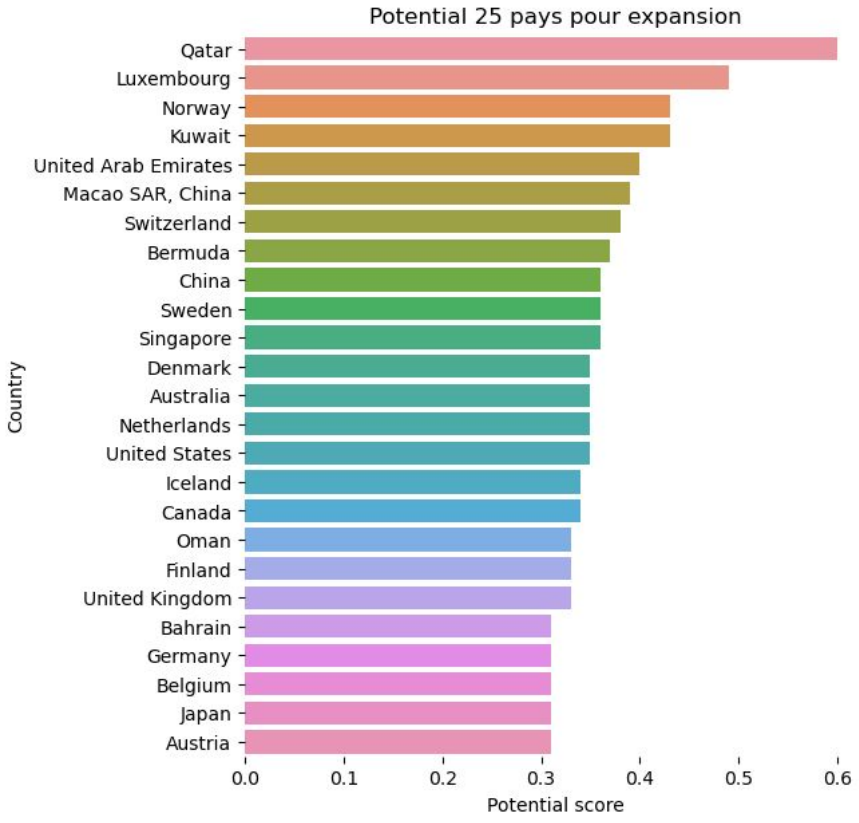


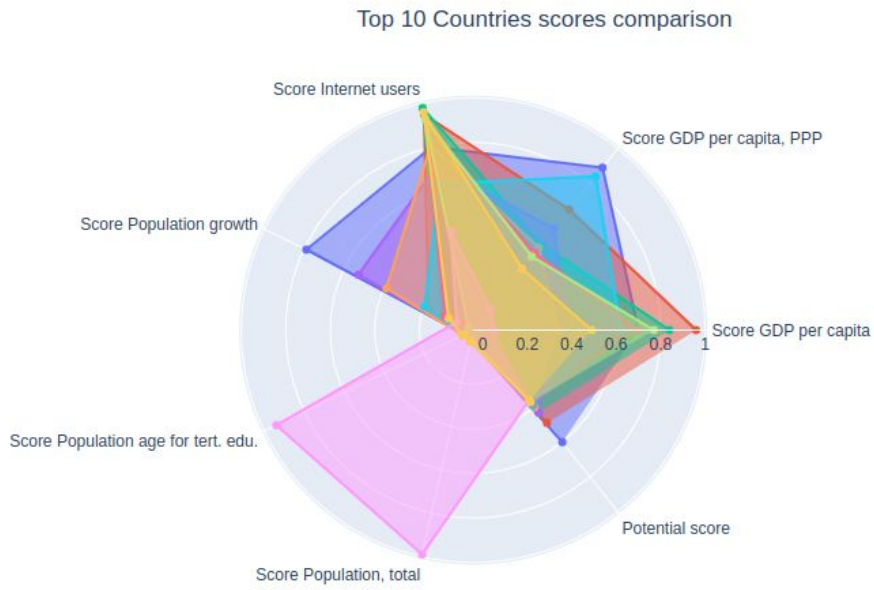
Merci !

06
Annexes

Model 2: Top 25 pays avec un fort potentiel pour acquérir les services d'Academy

Score = (IT.NET.USER.P2 + (NY.GDP.PCAP.CD + NY.GDP.PCAP. PP. CD)/2 + (SP.POP. GROW + SP.TER.TOTL.IN+)) / 4





5.1 Les 5 pays avec un fort potentiel de clients pour les services sont :

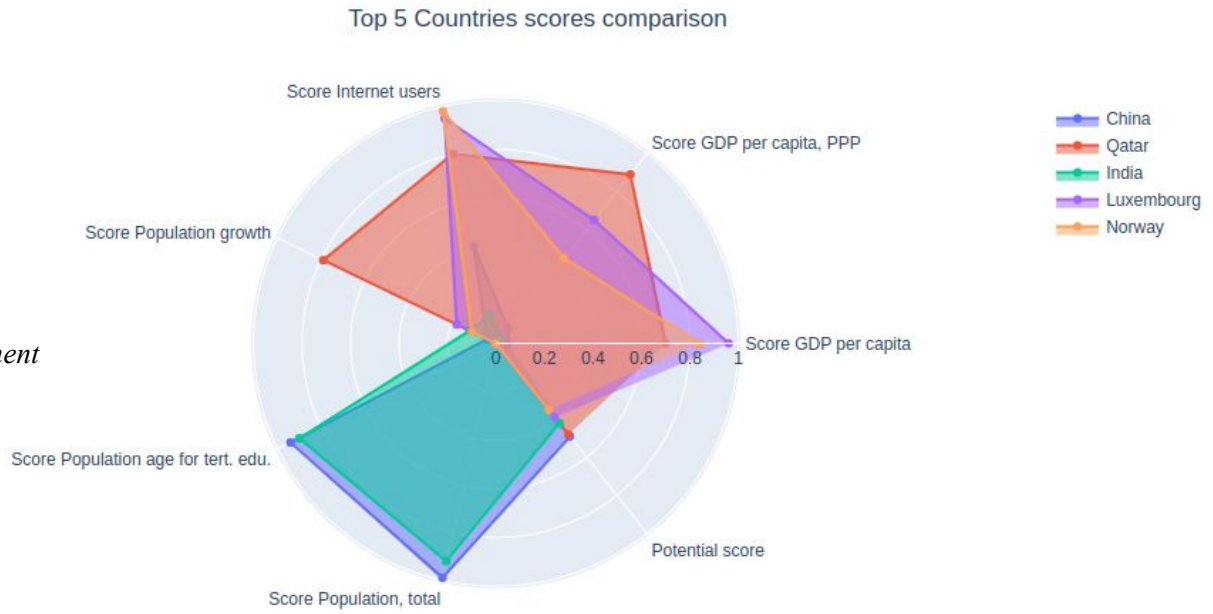
1. **China (Chine):**
   population total
   Population ayant l'âge officiel d'accéder à l'enseignement supérieur, hommes et femmes (nombre)
   Utilisateurs de l'internet

2. **Qatar :**
   PIB par habitant (GDP per capita, PPP)
   Utilisateurs de l'internet

3. **India (Inde) :**
   Population total
   Population ayant l'âge officiel d'accéder à l'enseignement supérieur, hommes et femmes (nombre)

4. **Luxembourg :**
   PIB par habitant (GDP)
   Utilisateurs de l'internet

5. **Norway (Norvège)**
   Utilisateurs de l'internet
   PIB par habitant



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