Formation de Data Science - Openclassrooms

Formation Ouverte et à Distance – FOAD par Pôle Emploi

Solutions 100% à distance

Projet 2 : Analysez des données de systèmes éducatifs

Étudiant : Maria Daniela Barrios

Mentor : Dan Slama

06 decembre, 2021

Contexte du problème

- ➔ On travaille pour une start-up EdTech appelée Academy, qui propose des contenus de formation en ligne pour des publics de niveau lycée et universitaire

- ➔ Mission : faire l'analyse exploratoire pour déterminer si les données sur l'éducation disponibles dans The World Bank peuvent informer l'expansion du projet

- ➔ Les trois questions principales :

1. Quels pays ont des clients potentiels pour nos services ?

2. Pour chacun de ces pays, comment évolueront les clients potentiels ?

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

Nous avons extrait les données du lien suivant :
https://datacatalog.worldbank.org/search/dataset/0038480

Stratégie pour réaliser la mission

➔ Phase pré-exploratoire : Analyse générale et découverte des fichiers
- Décrire les informations contenues dans l'ensemble de données: nombre de lignes et de colonnes

➔ Analyse de la qualité des données des indicateurs sélectionnés
- Exploration des valeurs dupliquées et des valeurs manquantes

➔ Sélection d'indicateurs possibles
- Proposition et sélection d'indicateurs pertinents

➔ Révision des indicateurs sélectionnés pour les années concernées
- Sélection des années qui seront utiles pour rapporter les quantités statistiques

➔ Analyse des données disponibles pour les indicateurs sélectionnés et les années concernées
- Détermination des ordres de grandeur des indicateurs statistiques classiques pour les différentes zones géographiques et pays du monde (moyenne/médiane/écart-type)

Phase pré-exploratoire : Analyse générale et découverte des fichiers

Cinq fichiers sur le jeu de données :

EdStatsCountry-Series.csv:

- **613** lignes et **4** colonnes

- Informations relatives à la méthode utilisée pour collecter les données et le series code pour chaque indicateur

EdStatsCountry.csv:

- **241** lignes et **32** colonnes

- Informations sur les pays, telles que le nom du pays, le code du pays (en conventions Alpha2 et Alpha3), l'unité monétaire, etc

EdStatsFootNote.csv:

- **643638** lignes et **5** colonnes

- Contient des informations supplémentaires sur la source de données.

EdStatsSeries.csv:

- **3665** lignes et **21** colonnes

- Contient des informations sur tous les **3665** indicateurs, les sujets, les définitions longues et courtes, leurs codes correspondants et les sources de données

EdStatsData.csv:

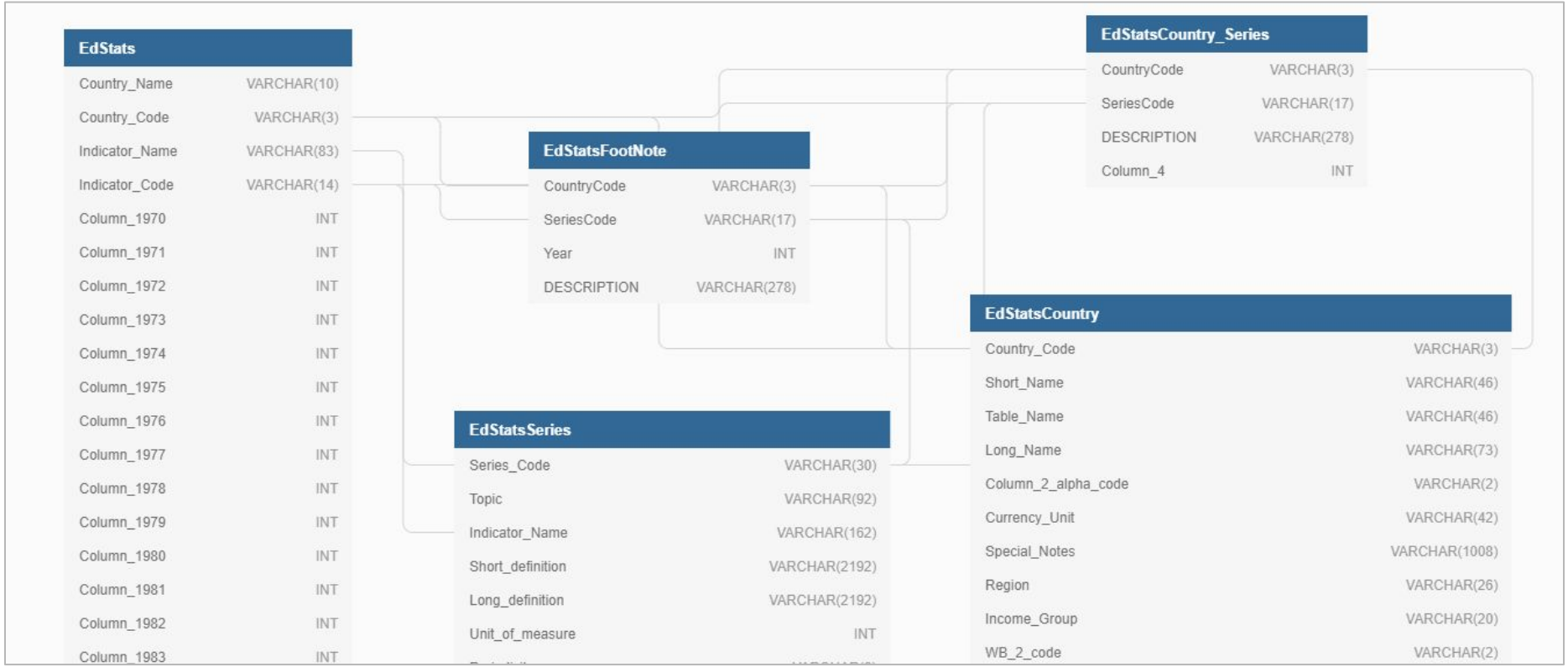
- **886930** lignes et **70** colonnes

- Contient tous les indicateurs pour chaque pays, y compris les codes pour ces pays et indicateurs. Les colonnes sont également désignées en fonction des années

Le tableau contient **3665** indicateurs pour **242** pays pendant **65** ans, de 1970 à 2017, et de 2020 à 2100

Phase pré-exploratoire : Analyse générale et découverte des fichiers

Les tables sont connectées entre elles comme suit : https://dbdiagram.io/home



Analyse de la qualité des données des indicateurs sélectionnés

Exploration des valeurs dupliquées

Vérifier s'il y a deux fois un indicateur pour le même pays dans le sous-ensemble de colonnes "Code de l'indicateur" et "Code du pays" → Si cette condition est vraie, elle sera affichée comme True dans un tableau

```
[31]  ed_stats_data[ed_stats_data.duplicated(subset=['Indicator Code','Country Code'])]
      ✓ 0.4s                                                                    Python
```

Country Name	Country Code	Indicator Name	Indicator Code	1970	1971	1972	1973	1974	1975	...	2060	2065	2070	2075	2080	2085	2090	2095	2100	Unnamed: 69

0 rows × 70 columns

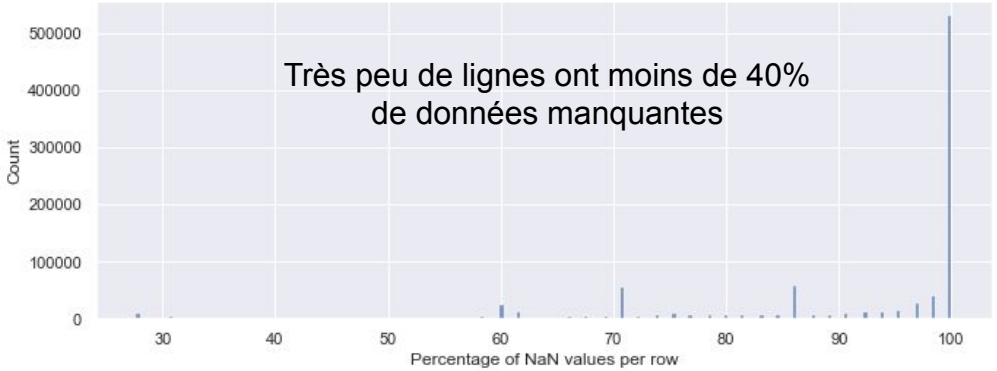
Exploration des valeurs manquantes

- Par ligne
- Par année (colonne)

Analyse de la qualité des données des indicateurs sélectionnés

Exploration des valeurs manquantes

- Par ligne



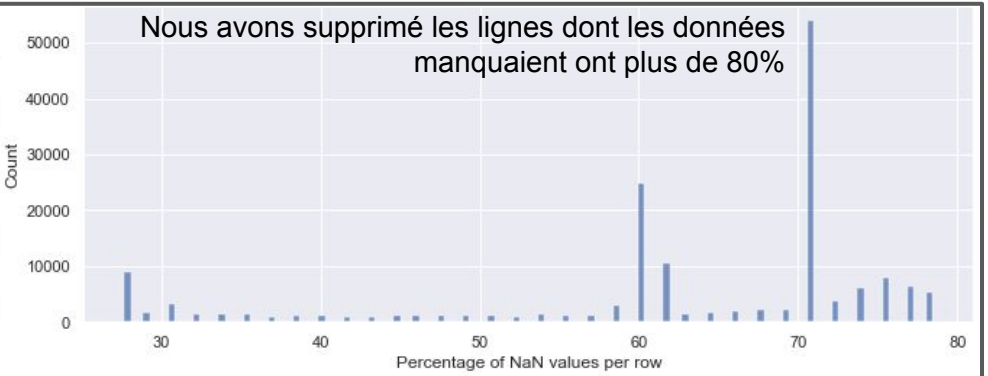
	Percentage of NaN values per row
count	886930.000000
mean	91.184456
std	15.590399
min	27.692308
25%	86.153846
50%	100.000000
75%	100.000000
max	100.000000

	Percentage of NaN values per row
count	719394.000000
mean	97.793326
std	4.633222
min	81.538462
25%	98.461538
50%	100.000000
75%	100.000000
max	100.000000

Supprimées

	Percentage of NaN values per row
count	162741.000000
mean	62.299586
std	14.095323
min	27.692308
25%	60.000000
50%	70.769231
75%	70.769231
max	78.461538

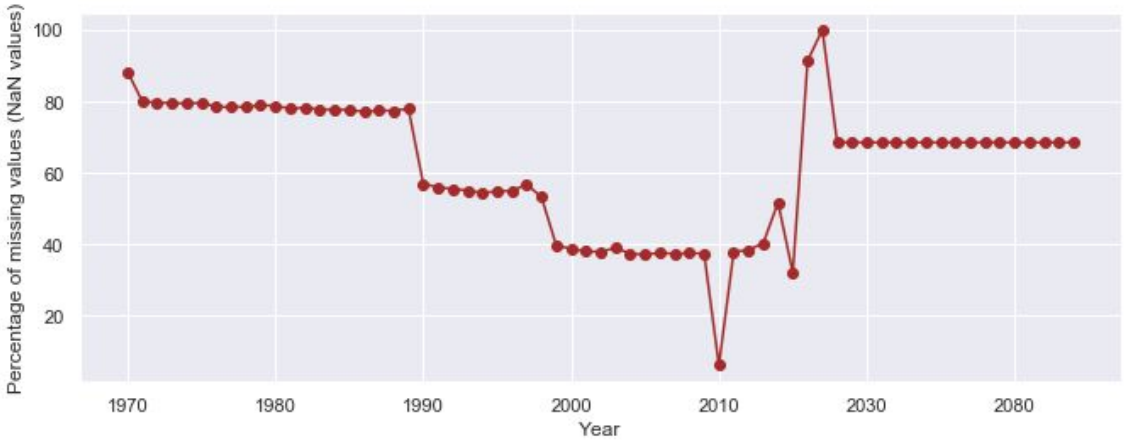
Gardées



Analyse de la qualité des données des indicateurs sélectionnés

Exploration des valeurs manquantes

- Par année



Nous avons déjà vu que tous les pays ne disposent pas de données pour toutes les années, nous devrons prendre une tranche d'années qui dépendra des données disponibles pour chaque indicateur

Sélection d'indicateurs possibles

Lien entre les questions et les indicateurs potentiels → Mots-clés associés aux questions de la réunion :

1. Quels pays ont des clients potentiels pour nos services ?

➔ Démographie ou **population totale par groupes d'âge**, **15-24** ans par exemple. Ces groupes pourraient faire partie de la clientèle cible.

➔ **Internet** et **ordinateur**, car le service de l'entreprise est une plateforme d'éducation en ligne, et les clients potentiels auront besoin d'ordinateurs et d'Internet.

➔ **Lycée** et **université**, parce que le contenu de la plateforme est destiné aux lycées et aux universités.

Sélection d'indicateurs possibles

Lien entre les questions et les indicateurs potentiels → Mots-clés associés aux questions de la réunion :

2. Pour chacun de ces pays, comment évolueront les clients potentiels ?

Question un peu ouverte → Faire des réflexions à partir des indicateurs qu'on extrait à partir des données en utilisant un modèle prédictif ou autre

Sélection d'indicateurs possibles

Lien entre les questions et les indicateurs potentiels → Mots-clés associés aux questions de la réunion :

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

➔ **Internet** et **ordinateur**, car le service de l'entreprise est une plateforme d'éducation en ligne, et les clients potentiels auront besoin d'ordinateurs et d'Internet.

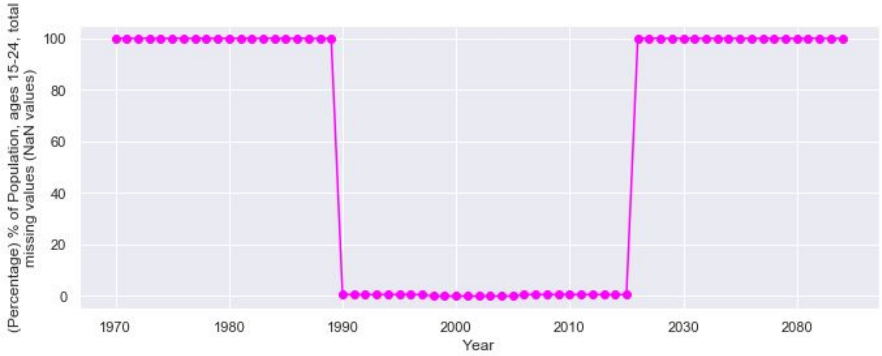
➔ **Lycée** et **université**, parce que le contenu de la plateforme est destiné aux lycées et aux universités.

➔ **PIB (Produit intérieur brut)** : exprime la valeur monétaire de la production de biens et de services pour la demande finale d'un pays ou d'une région pendant une période donnée.

Révision des indicateurs sélectionnés pour les années concernées

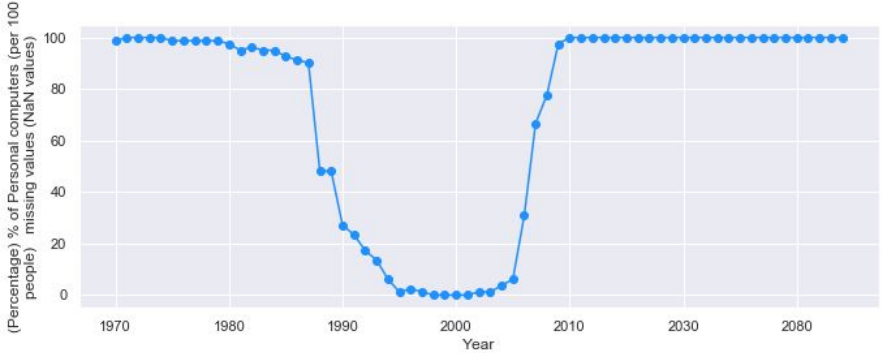
Exploration des valeurs manquantes par indicateur

- **Indicateur: Population, ages 15-24, total**



Les données de 1990 à 2015 contiennent très peu pourcentage de valeurs manquantes, notamment moins de 40%

- **Indicateur: Personal computers (per 100 people)**

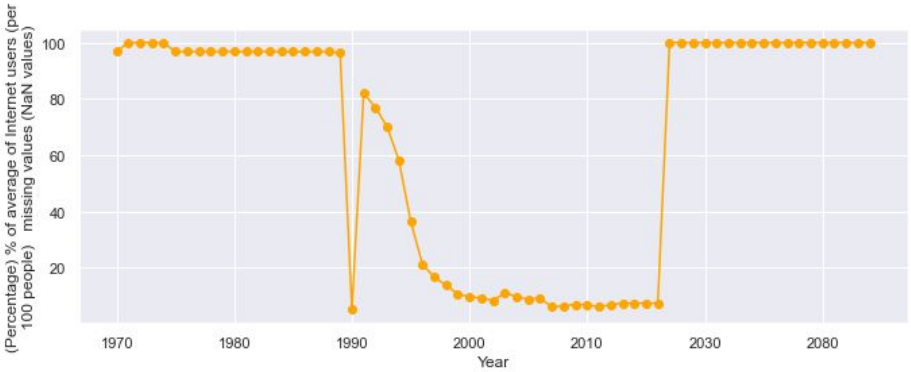


Les années entre 1990 et 2006 ont moins de 40%

Révision des indicateurs sélectionnés pour les années concernées

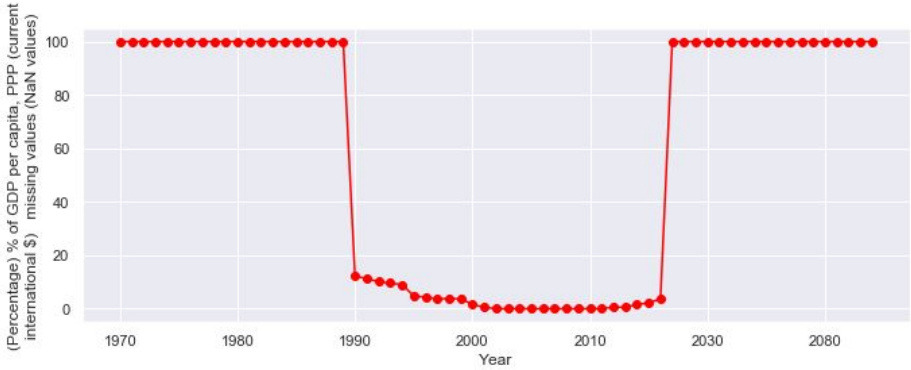
Exploration des valeurs manquantes par indicateur

- **Indicateur: Internet users (per 100 people)**



La tranche d'années qui contient moins de 40% de données manquantes se situe entre 1995 et 2016

- **Indicateur: GDP per capita, PPP (current international $)**

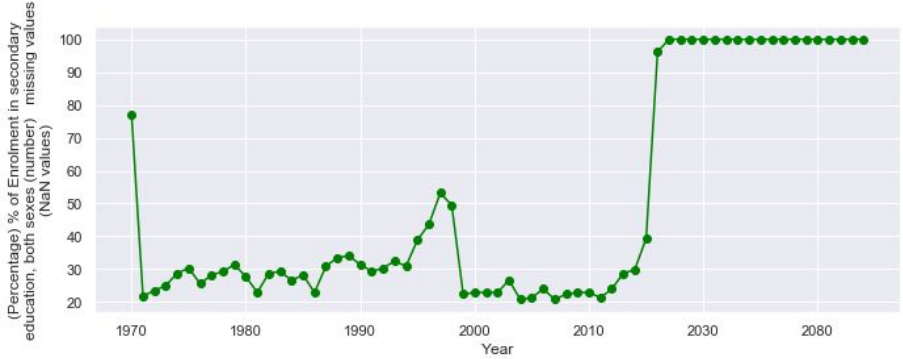


La tranche d'années qui contient moins de 40% de données manquantes se situe entre 1990 et 2016

Révision des indicateurs sélectionnés pour les années concernées

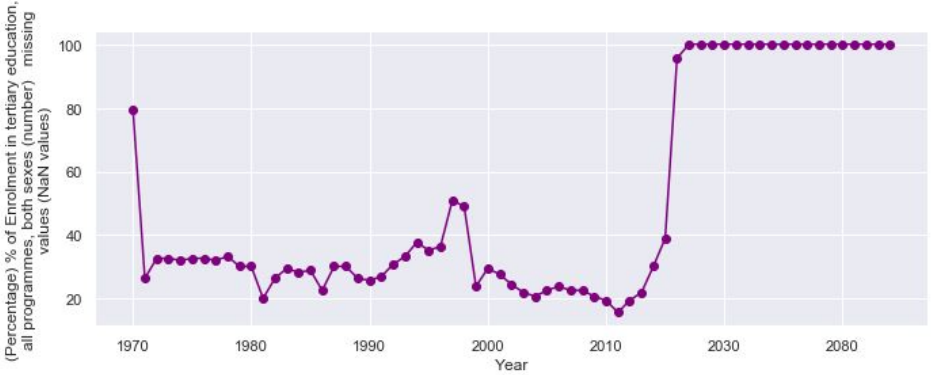
Exploration des valeurs manquantes par indicateur

- **Indicateur: Enrolment in secondary education, both sexes (number)**



La tranche d'années qui contient moins de 40% de données manquantes se situe entre 1999 et 2015

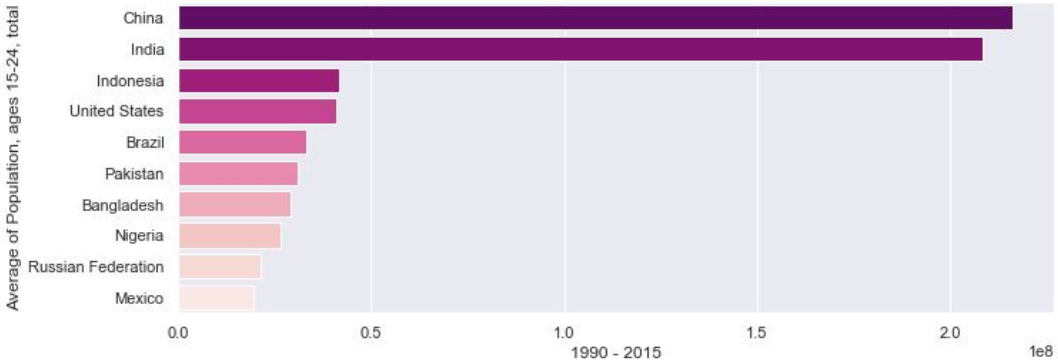
- **Enrolment in tertiary education, all programmes, both sexes (number)**



La tranche d'années qui contient moins de 40% de données manquantes se situe entre 1999 et 2015

Analyse des données disponibles pour les indicateurs sélectionnés et les années concernées

- **Indicateur: Population, ages 15-24, total**

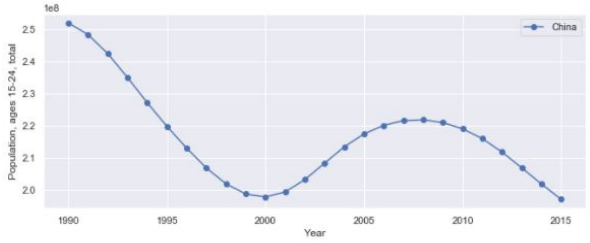


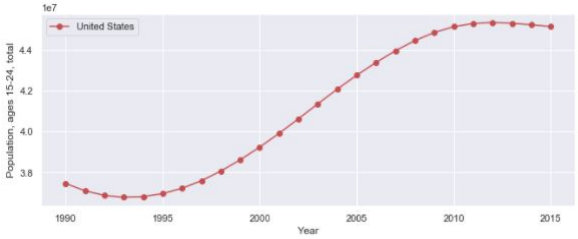
Country	Mean top 10 country	Standard deviation top 10 country
China	2.161629e+08	1.513938e+07
India	2.084619e+08	2.570123e+07
Indonesia	4.157602e+07	1.410595e+06
United States	4.105920e+07	3.435548e+06
Brazil	3.321308e+07	1.993811e+06
Pakistan	3.098327e+07	7.132889e+06
Bangladesh	2.927054e+07	3.888743e+06
Nigeria	2.673588e+07	6.040415e+06
Russian Federation	2.124383e+07	2.794309e+06
Mexico	1.959719e+07	5.843156e+05

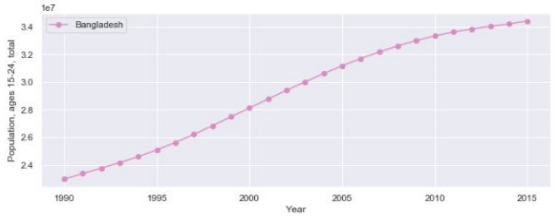
Pas de données disponibles pour les régions géographiques

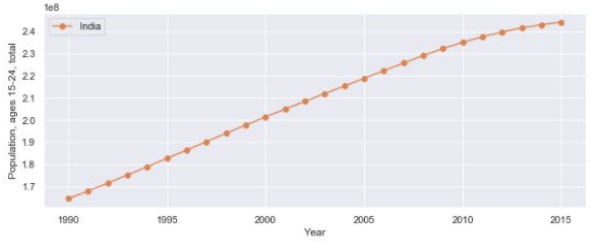
Analyse des données disponibles pour les indicateurs sélectionnés et les années concernées

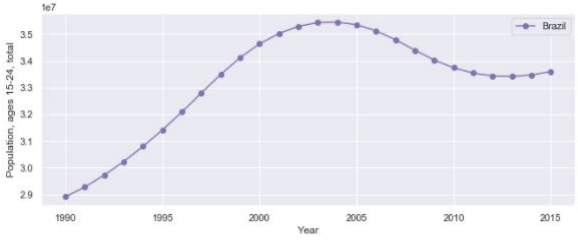
- **Indicateur: Population, ages 15-24, total**

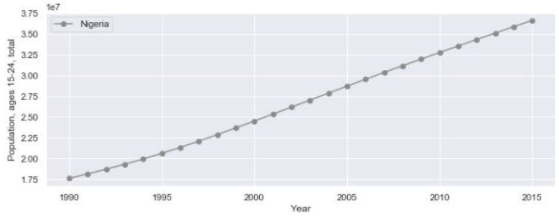


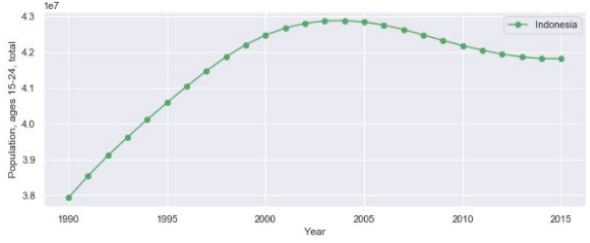


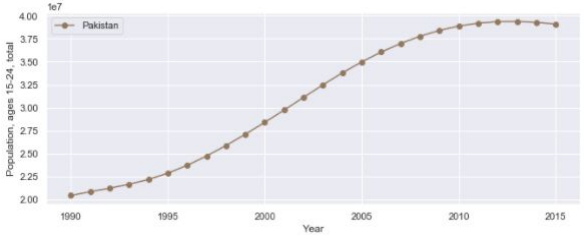


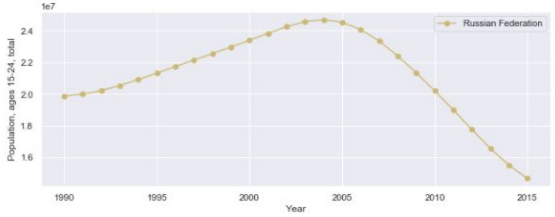


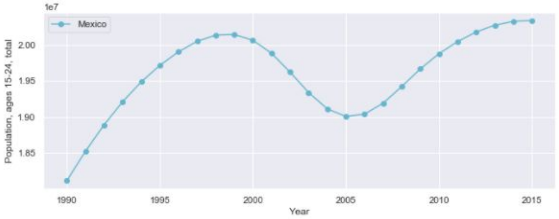






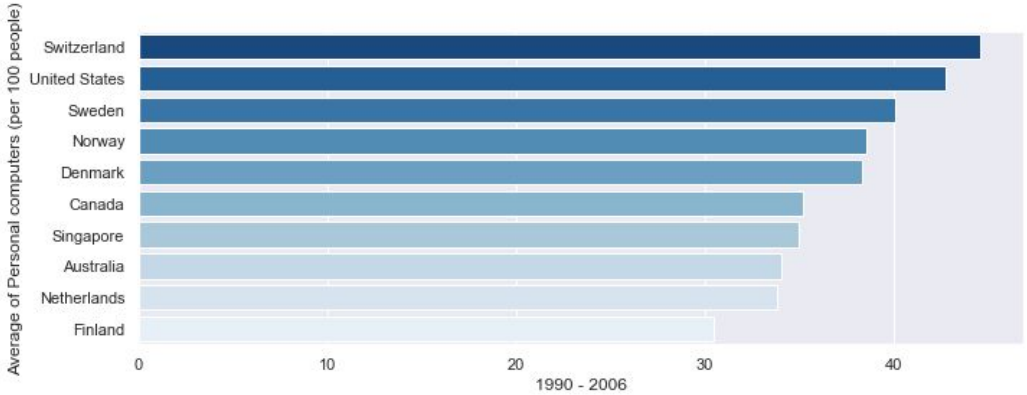




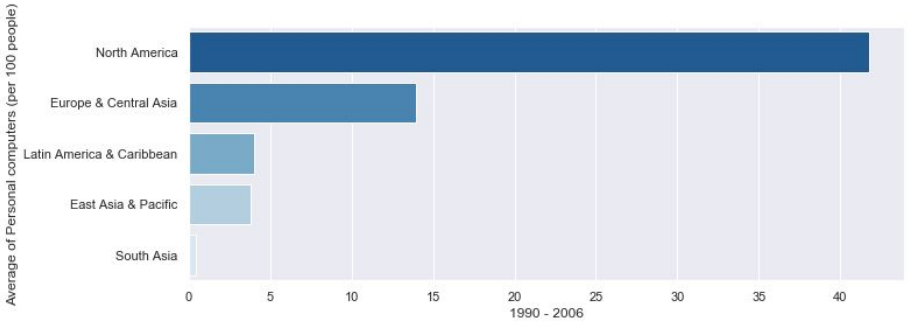


Analyse des données disponibles pour les indicateurs sélectionnés et les années concernées

- **Indicateur: Personal computers (per 100 people)**



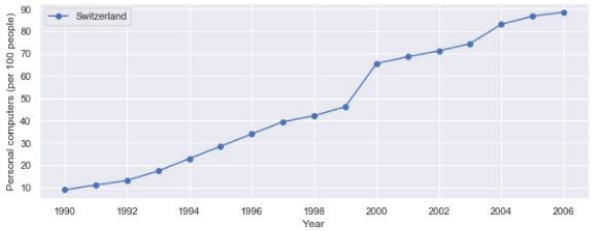
Country	Mean top 10 country	Standard deviation top 10 country
Switzerland	47.211879	28.165007
United States	45.246776	20.643806
Sweden	42.902756	26.118360
Norway	40.043380	16.339260
Denmark	40.331133	21.005607
Canada	38.654056	25.515943
Singapore	36.949846	22.370883
Australia	34.028258	15.178501
Netherlands	37.174808	25.101921
Finland	30.489937	14.212802

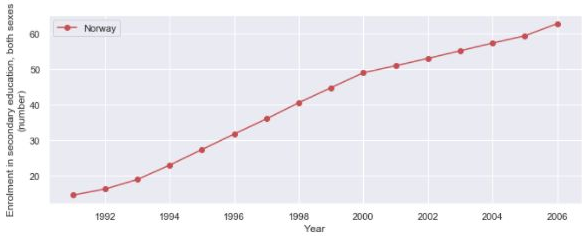


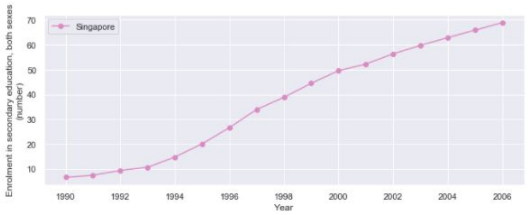
Country	Mean top 10 country	Standard deviation top 10 country
North America	44.444141	21.172083
Europe & Central Asia	15.169562	9.462203
Latin America & Caribbean	4.013894	3.197873
East Asia & Pacific	4.053468	2.670573
South Asia	0.541835	0.683510

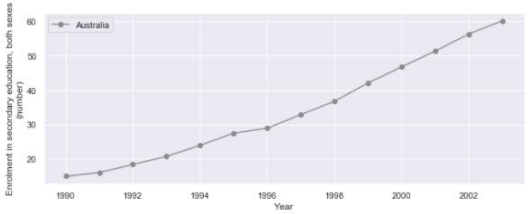
Analyse des données disponibles pour les indicateurs sélectionnés et les années concernées

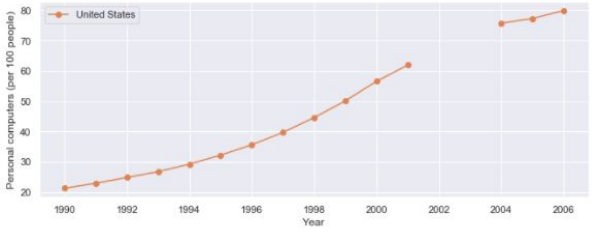
- **Indicateur: Personal computers (per 100 people)**

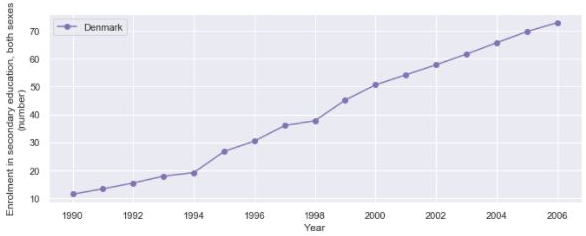


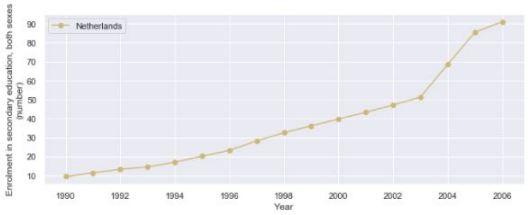


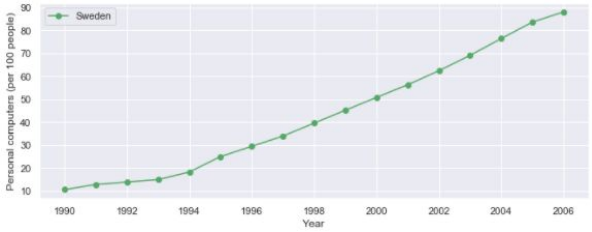


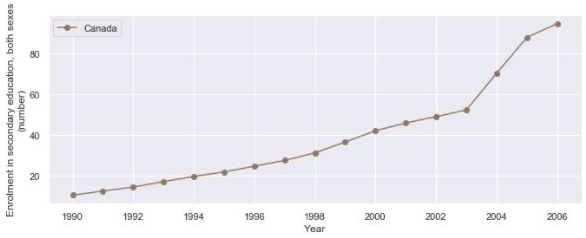


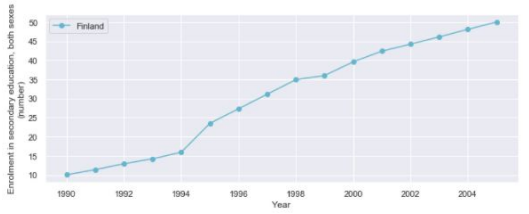






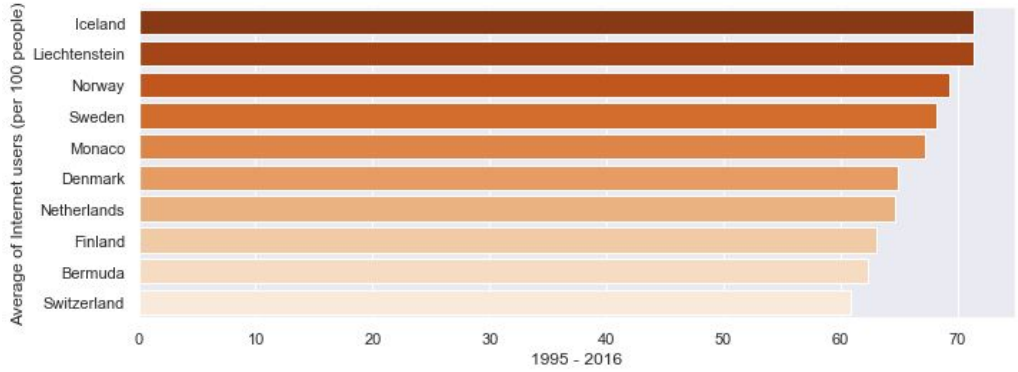




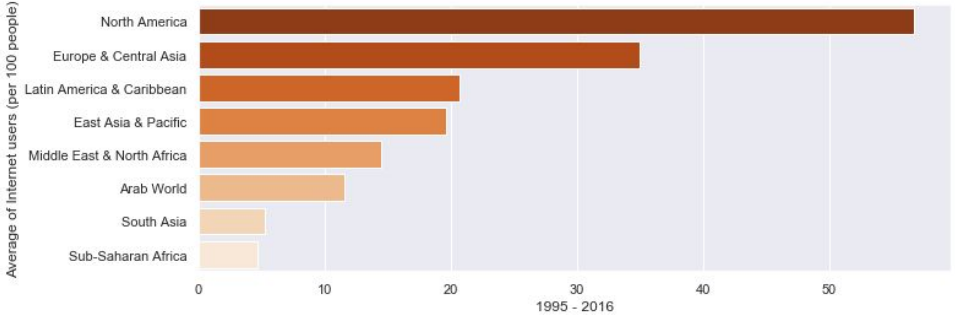


Analyse des données disponibles pour les indicateurs sélectionnés et les années concernées

- **Indicateur: Internet users (per 100 people)**



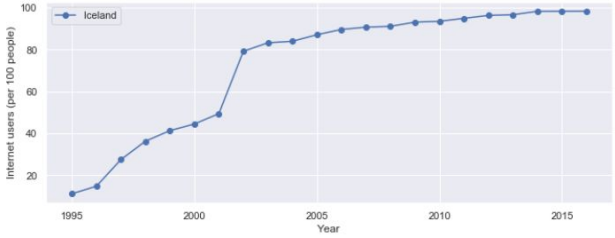
Country	Mean top 10 country	Standard deviation top 10 country
Iceland	72.624538	29.800863
Liechtenstein	72.925792	18.364638
Norway	70.610154	29.922482
Sweden	69.308354	29.596093
Monaco	68.911736	18.307843
Denmark	66.306427	32.681570
Netherlands	65.884234	30.446903
Finland	64.136867	27.572996
Bermuda	63.945523	28.150209
Switzerland	62.132696	28.236486

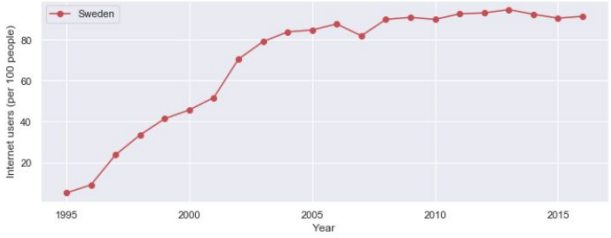


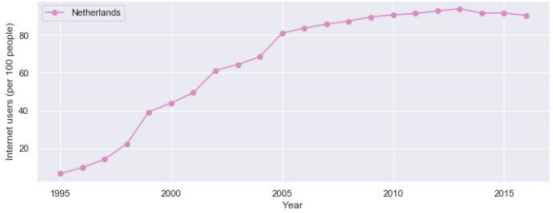
Country	Mean top 10 country	Standard deviation top 10 country
North America	57.681379	21.911036
Europe & Central Asia	36.725208	24.882466
Latin America & Caribbean	22.272095	19.209622
East Asia & Pacific	21.141932	17.887083
Middle East & North Africa	15.960277	15.820633
Arab World	11.592177	12.728905
South Asia	6.216173	7.873298
Sub-Saharan Africa	5.388914	6.222082

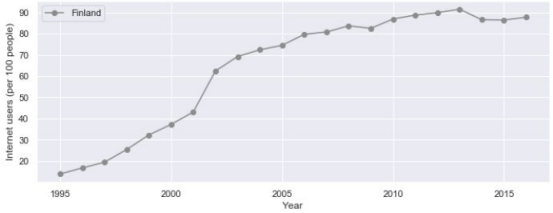
Analyse des données disponibles pour les indicateurs sélectionnés et les années concernées

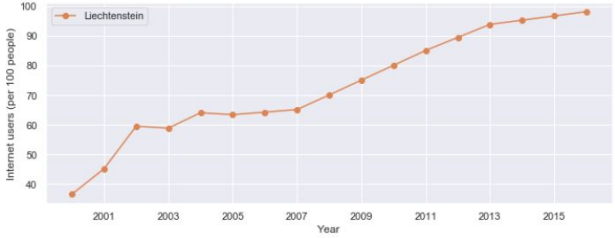
- **Indicateur: Internet users (per 100 people)**

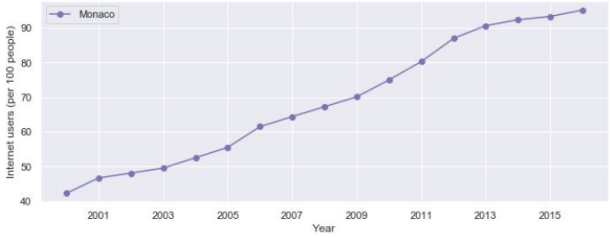


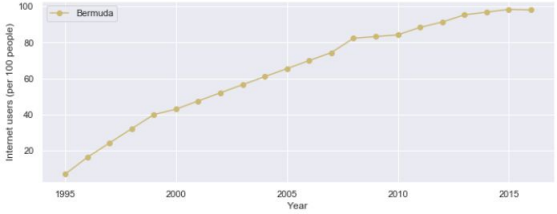


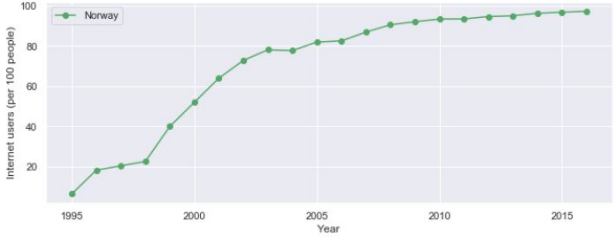


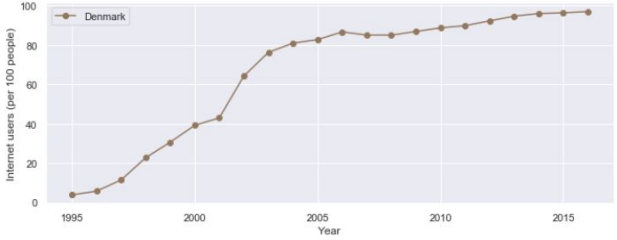


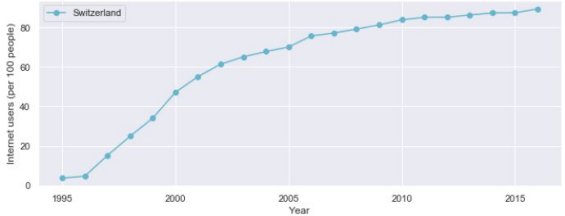






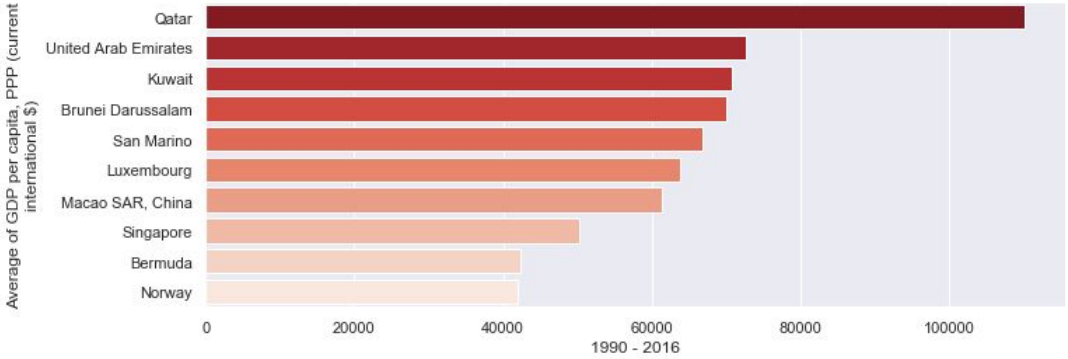




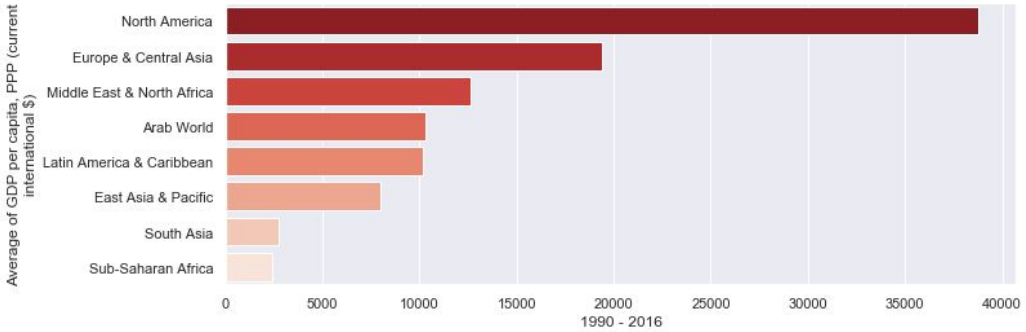


Analyse des données disponibles pour les indicateurs sélectionnés et les années concernées

- **Indicateur: GDP per capita, PPP (current international $)**



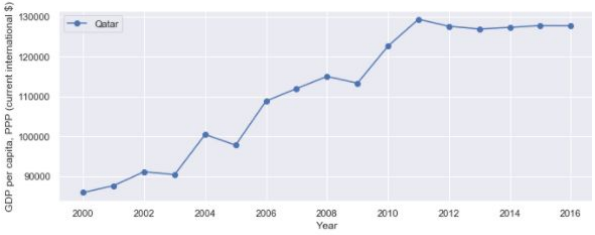
Country	Mean top 10 country	Standard deviation top 10 country
Qatar	111264.196449	16091.762916
United Arab Emirates	72691.586713	7901.955671
Kuwait	71019.249189	12016.330645
Brunei Darussalam	70403.738411	9088.472433
San Marino	66489.608408	10127.523139
Luxembourg	65286.344227	24402.422207
Macao SAR, China	62931.680647	37518.674075
Singapore	51669.312992	21107.583450
Bermuda	42339.929991	11329.487408
Norway	42601.836485	16944.783804

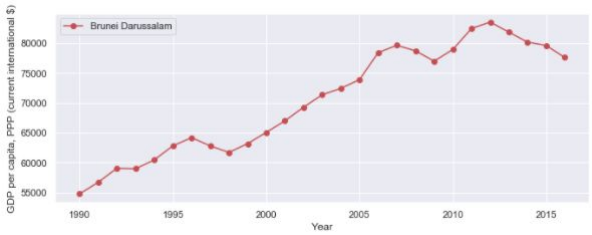


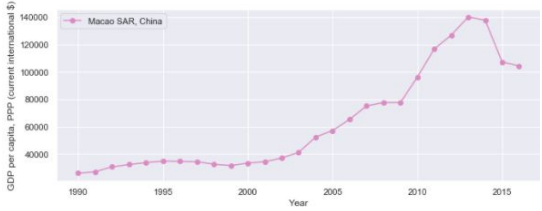
Country	Mean top 10 country	Standard deviation top 10 country
North America	39429.896960	10443.316993
Europe & Central Asia	19854.656893	6942.982733
Middle East & North Africa	12855.726548	3837.522738
Arab World	10300.657721	3176.896343
Latin America & Caribbean	10359.759086	3202.661893
East Asia & Pacific	8295.853561	4276.124416
South Asia	2882.244161	1492.423923
Sub-Saharan Africa	2447.902450	766.914051

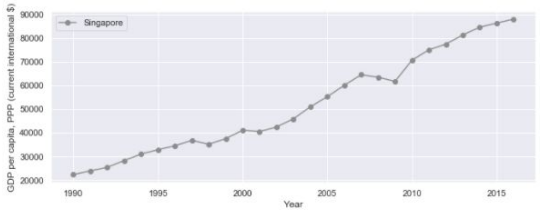
Analyse des données disponibles pour les indicateurs sélectionnés et les années concernées

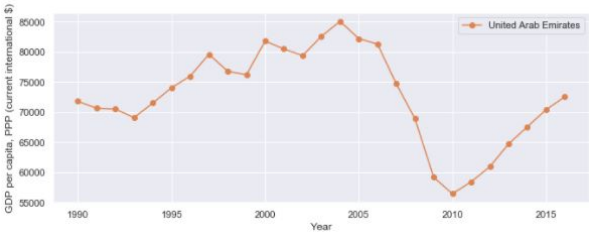
- **Indicateur: GDP per capita, PPP (current international $)**

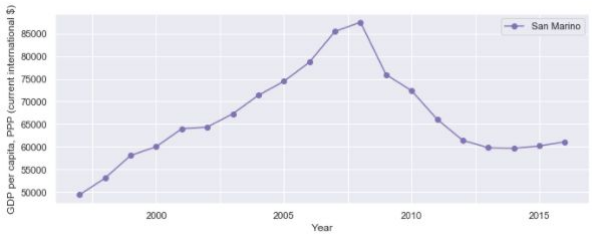


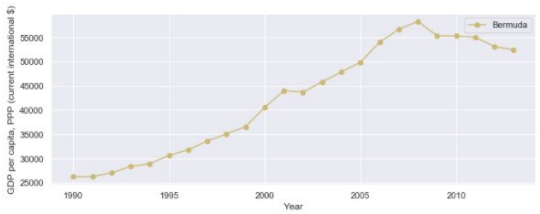


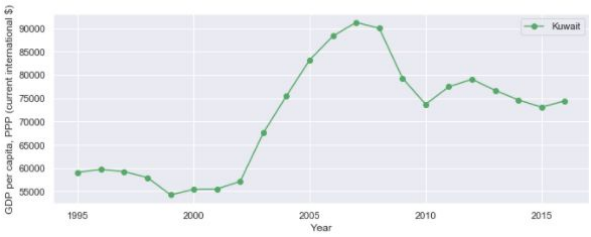


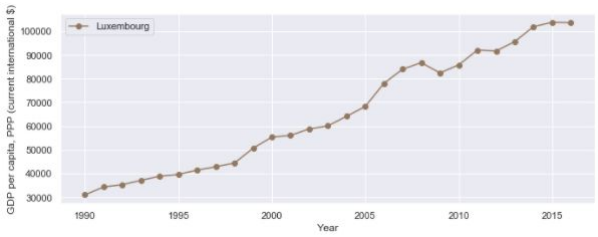


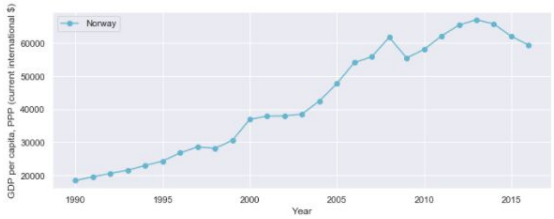






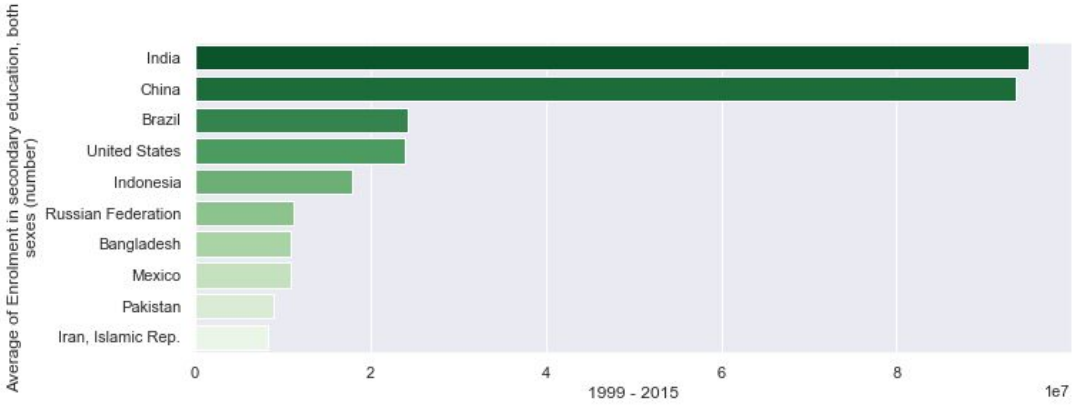




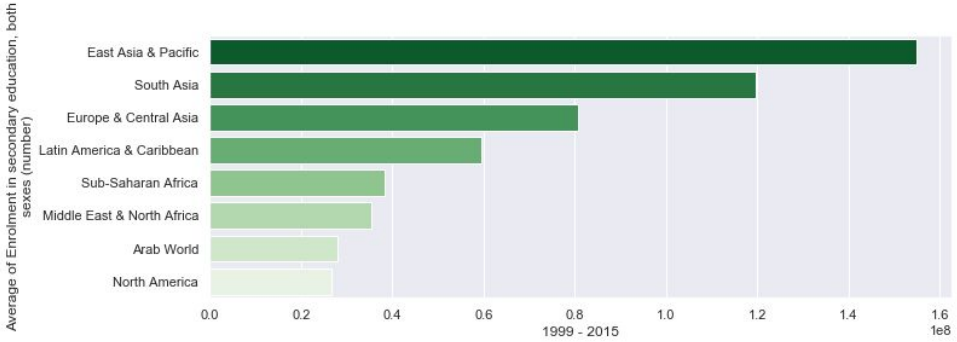


Analyse des données disponibles pour les indicateurs sélectionnés et les années concernées

- **Indicateur: Enrolment in secondary education, both sexes (number)**



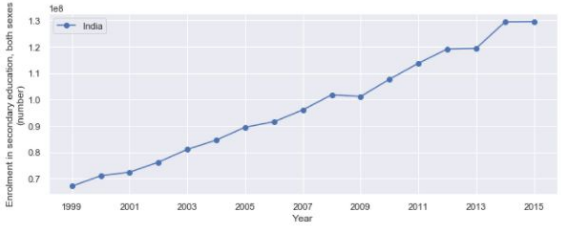
Country	Mean top 10 country	Standard deviation top 10 country
India	9.713101e+07	2.037191e+07
China	9.316490e+07	7.698240e+06
Brazil	2.421283e+07	1.025687e+06
United States	2.394684e+07	7.223936e+05
Indonesia	1.833100e+07	3.075603e+06
Russian Federation	1.110251e+07	2.313065e+06
Bangladesh	1.112852e+07	1.274770e+06
Mexico	1.087905e+07	1.244716e+06
Pakistan	9.192320e+06	1.894644e+06
Iran, Islamic Rep.	8.236126e+06	1.512713e+06

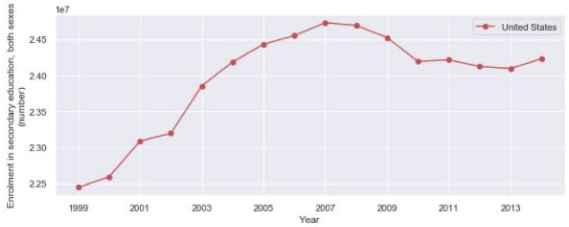


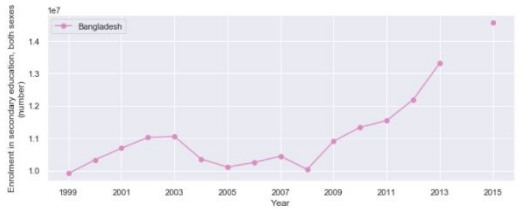
Country	Mean top 10 country	Standard deviation top 10 country
East Asia & Pacific	1.549664e+08	9.896861e+06
South Asia	1.196900e+08	2.260532e+07
Europe & Central Asia	8.068638e+07	4.516576e+06
Latin America & Caribbean	5.955585e+07	2.574362e+06
Sub-Saharan Africa	3.824453e+07	1.110534e+07
Middle East & North Africa	3.549594e+07	1.642820e+06
Arab World	2.803258e+07	2.945425e+06
North America	2.656114e+07	7.742312e+05

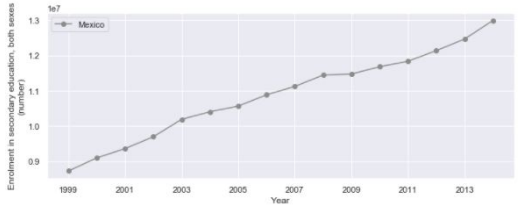
Analyse des données disponibles pour les indicateurs sélectionnés et les années concernées

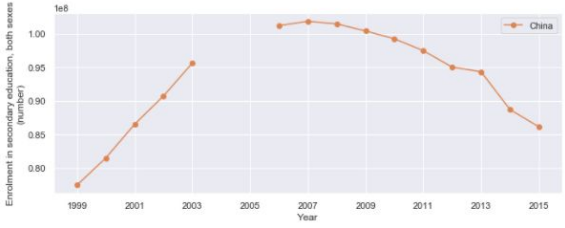
- **Indicateur: Enrolment in secondary education, both sexes (number)**

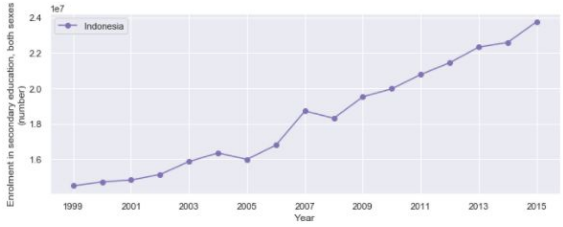


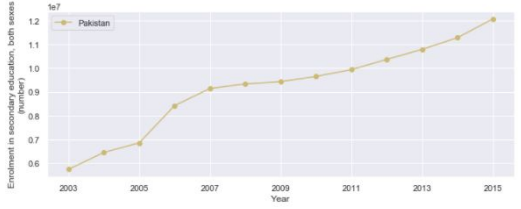


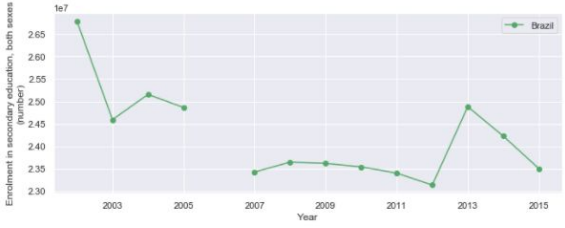


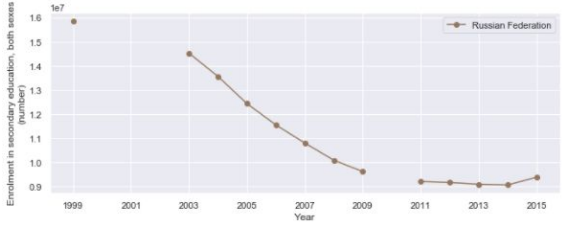


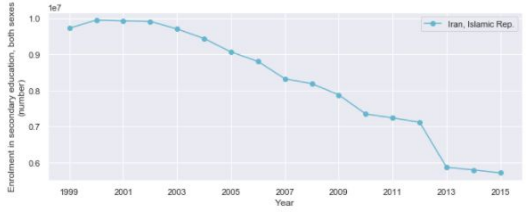






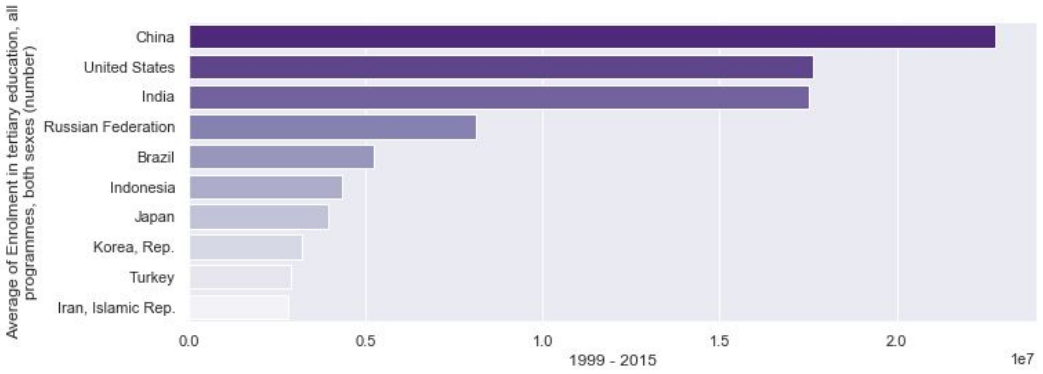




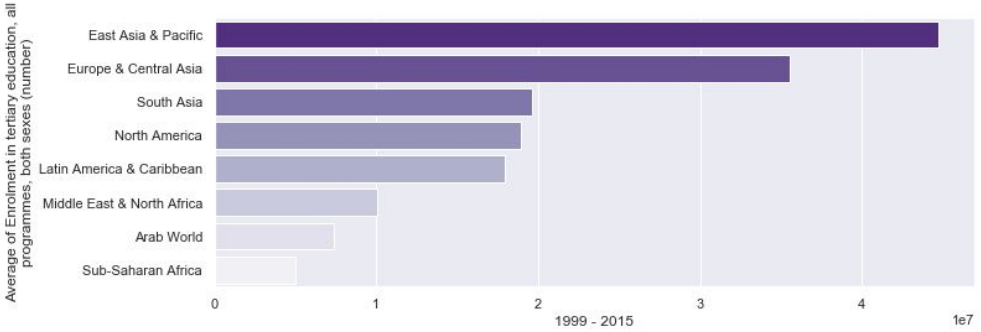


Analyse des données disponibles pour les indicateurs sélectionnés et les années concernées

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



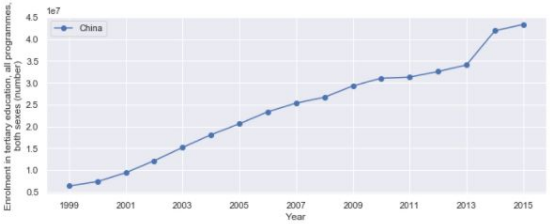
Country	Mean top 10 country	Standard deviation top 10 country
China	2.400992e+07	1.139670e+07
United States	1.773646e+07	2.522035e+06
India	1.841976e+07	8.163148e+06
Russian Federation	8.030212e+06	1.182779e+06
Brazil	5.422263e+06	1.929808e+06
Indonesia	4.374798e+06	1.215088e+06
Japan	3.948332e+06	7.557522e+04
Korea, Rep.	3.211176e+06	1.294091e+05
Turkey	3.075755e+06	1.463067e+06
Iran, Islamic Rep.	2.928386e+06	1.260035e+06

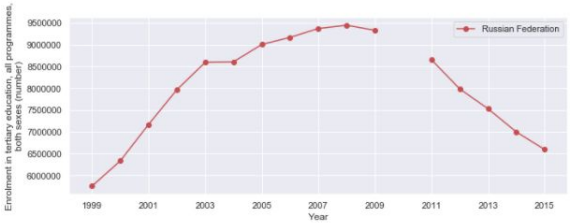


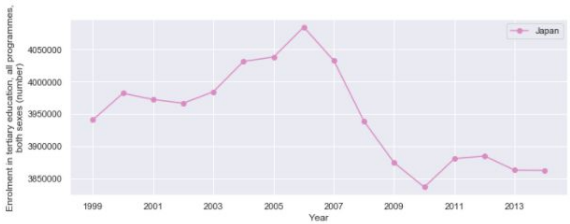
Country	Mean top 10 country	Standard deviation top 10 country
East Asia & Pacific	4.474047e+07	1.376660e+07
Europe & Central Asia	3.552121e+07	3.977190e+06
South Asia	1.965941e+07	8.796769e+06
North America	1.893470e+07	2.621734e+06
Latin America & Caribbean	1.792066e+07	4.630562e+06
Middle East & North Africa	1.006085e+07	2.719691e+06
Arab World	7.361873e+06	1.651123e+06
Sub-Saharan Africa	5.037613e+06	1.756842e+06

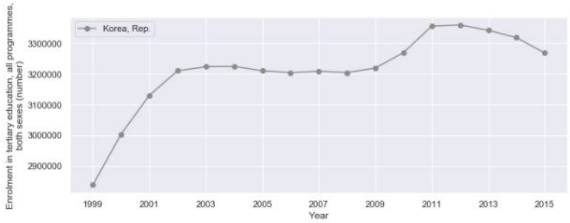
Analyse des données disponibles pour les indicateurs sélectionnés et les années concernées

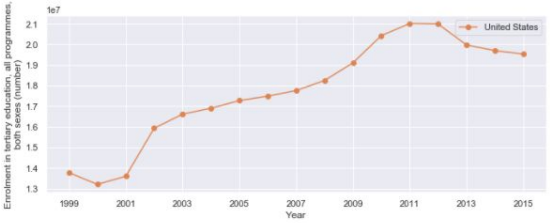
- **Indicateur: Enrolment in tertiary education, all programmes, both sexes (number)**

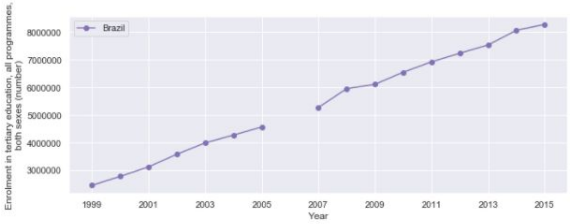


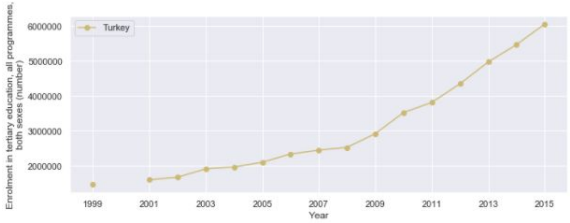


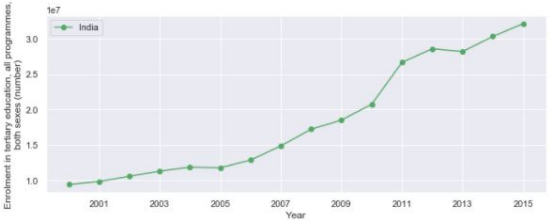


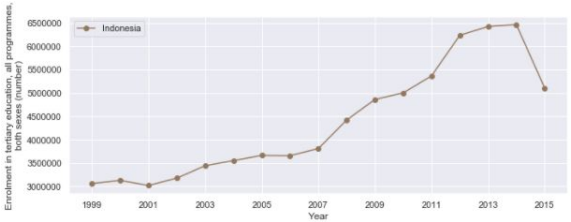


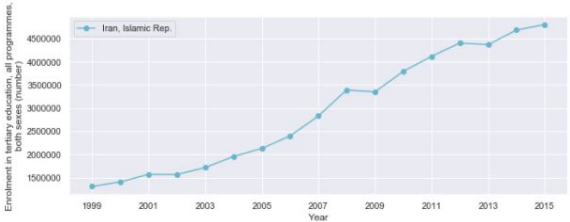












Conclusions

- ➔ Nous disposons d'un ensemble de données composé de 5 tableaux qui présentent différents indicateurs couvrant différents sujets pour chaque pays de 1970 à 2100

- ➔ Le jeu de données formé par 5 tableaux contient plus de 80% de valeurs manquantes

- ➔ Les tableaux contiennent des indicateurs qui pourraient fournir des informations pour les besoins de l'entreprise

- ➔ Les tables sont reliées entre elles par le “Country Name”, “Country Code”, “Indicator Name” and “Indicator Code”

- ➔ Nous avons sélectionné des indicateurs pertinents liés aux besoins de l'entreprise
  - Nous avons extrait des informations sur les quantités statistiques

Conclusions

➔ Top 10 des pays par indicateur

Ranking	1	2	3	4	5	6	7	8	9	10
15-24 year old	China	India	Indonesia	United States	Brazil	Pakistan	Bangladesh	Nigeria	Russian Federation	Mexico
PC users	Switzerland	United States	Sweden	Norway	Denmark	Canada	Singapore	Australia	Netherlands	Finland
Internet Users	Iceland	Liechtenstein	Norway	Sweden	Monaco	Denmark	Netherlands	Finland	Bermuda	Switzerland
GDP per capita	Qatar	United Arab Emirates	Kuwait	Brunei Darussalam	San Marino	Luxembourg	Macao SAR, China	Singapore	Bermuda	Norway
Secondary Education	India	China	Brazil	United States	Indonesia	Russian Federation	Bangladesh	Mexico	Pakistan	Iran, Islamic Rep.
Tertiary Education	China	United States	India	Russian Federation	Brazil	Indonesia	Japan	Korea, Rep.	Turkey	Iran, Islamic Rep.

Conclusions

1. **Quels pays ont des clients potentiels pour nos services ?**

2. **Pour chacun de ces pays, comment évolueront les clients potentiels ?**

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

➔ Afin de répondre aux questions de la mission à partir des indicateurs, il est recommandé de compléter l'étude avec des données plus récentes et plus complètes sur les pays signalés