



Analysez des données de systèmes éducatifs





SOMMAIRE

1- Objectif

2- Mission

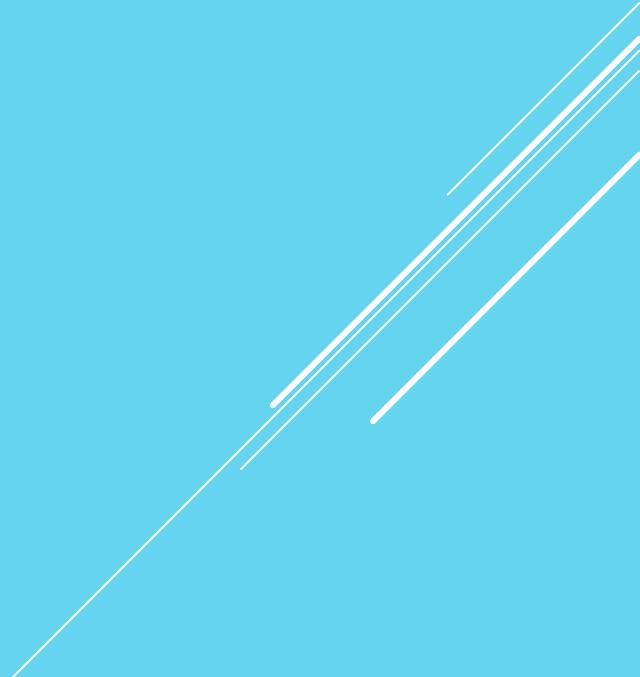
3- Le jeu de données

4- Analyse pré-exploratoire des données

5- Analyse des statistiques globales

6- Analyse des besoins spécifique de l'entreprise

7- Point de vue personnel et limite





1- Objectif

Information relative à l'entreprise: La start-up de la EdTech, nommé Academy est spécialisé dans la formation en ligne pour un public de niveau lycée et université.

Objectif: Projet d'expansion à l'international de l'entreprise

2- Mission

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. Comme par exemple, explorer les pays avec un fort potentiel de clients pour les services de academy, et comment ce potentiel pourrait évoluer.

Interrogations pertinentes:

- Quels sont les pays prioritaires pour les opérations de la start-up de la EdTech, nommée academy ?
- Quels pays présentent un fort potentiel de clientèle pour nos services ?
- Quelle sera l'évolution du potentiel client pour chacun de ces pays ?



3- Le jeu de données

Jeux de données extraits du site de la banque mondiale:

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

Nous avons récupéré cinq (5) ensembles de données, répartis de la manière suivante :

- **Un fichier Csv EdStatsFootNote:** Présente des informations de révision des méthodes de calcul et prise en compte des incertitudes de certains indicateurs, variant en fonction de l'année et du pays.
- **Un fichier Csv EdStatsCountry:** Présente des informations concernant les différents pays/zones présents dans le fichier.
- **Un fichier Csv EdStatsSeries:** Présente des informations relatives aux divers indicateurs présents dans le fichier.
- **Un fichier Csv EdStatsCountry-Series:** Présente des informations détaillant l'origine de certains indicateurs spécifiques pour certains pays.
- **Un fichier Csv EdStatsData:** Origine des données, Présente des informations couvrant divers pays et indicateurs pour la période allant de 1970 à 2100. »



4- Analyse pré-exploratoire des données

➤ Valider la qualité de ce jeu de données et Décrire les informations contenues dans le jeu de données.

| Nom_Fichier | Nbre_Lignes | Nbre_Colonne | Nbre_Duplicatas | Nbre_NaN | Nbre_NonNaN | Pourcentage_NaN | Nbre_Pays/Zones | Nbre_Indicateurs |
|------------------|-------------|--------------|-----------------|----------|-------------|-----------------|-----------------|------------------|
| df_Data | 357405 | 69 | 0 | 18149124 | 6511821 | 74 | 242 | 3665 |
| df_Country | 241 | 32 | 0 | 2354 | 5358 | 31 | 241 | 0 |
| df_CountrySeries | 613 | 3 | 0 | 0 | 1839 | 0 | 211 | 21 |
| df_StatsFootNote | 643638 | 4 | 0 | 0 | 2574552 | 0 | 239 | 1558 |
| df_StatsSeries | 3665 | 21 | 0 | 55203 | 21762 | 72 | 0 | 3665 |

Observation:

- ❖ Zéro (0) doublon ou duplicitas dans les jeux de données
- ❖ 3665 indicateurs
- ❖ 242 pays observer
- ❖ Plusieurs données manquantes (NaN) dans les jeux de données

- 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 ?

Pour le fichier Country:

- Prendre en compte la colonne de noms des pays.
- Conserver la colonne de groupement des pays par zone revenu élevé ou zone de richesse [Income Group].

| Short Name | |
|----------------------|---|
| Income Group | |
| High income: OECD | [Australia, Austria, Belgium, Canada, Switzerland, Chile, Czech Republic, Germany, Denmark, Spain, Estonia, Finland, France, United Kingdom, Greece, ...] |
| High income: nonOECD | [Aruba, Andorra, United Arab Emirates, Antigua and Barbuda, Bahrain, The Bahamas, Bermuda, Barbados, Brunei, Channel Islands, Curaçao, Cayman Islands, ...] |
| Low income | [Afghanistan, Burundi, Benin, Burkina Faso, Bangladesh, Central African Republic, Dem. Rep. Congo, Comoros, Eritrea, Ethiopia, Guinea, The Gambia, ...] |
| Lower middle income | [Armenia, Bolivia, Bhutan, Côte d'Ivoire, Cameroon, Congo, Cabo Verde, Djibouti, Egypt, Micronesia, Georgia, Ghana, Guatemala, Guyana, Honduras, ...] |
| Upper middle income | [Angola, Albania, Argentina, American Samoa, Azerbaijan, Bulgaria, Bosnia and Herzegovina, Belarus, Belize, Brazil, Botswana, China, Colombia, Costa Rica, ...] |

| Short Name | |
|----------------------------|---|
| Region | |
| East Asia & Pacific | [American Samoa, Australia, Brunei, China, Fiji, Micronesia, Guam, Hong Kong SAR, China, Indonesia, Japan, Cambodia, Kiribati, Korea, Lao PDR, Macau, Mongolia, Myanmar, Nauru, New Zealand, Palau, Papua New Guinea, Philippines, Samoa, Singapore, Sri Lanka, Thailand, Timor-Leste, Tuvalu, ...] |
| Europe & Central Asia | [Albania, Andorra, Armenia, Austria, Azerbaijan, Belgium, Bulgaria, Bosnia and Herzegovina, Belarus, Switzerland, Channel Islands, Cyprus, Czech Republic, Estonia, Georgia, Greece, Hungary, Italy, Malta, Montenegro, North Macedonia, Poland, Portugal, Romania, Russia, San Marino, Serbia, Slovakia, Slovenia, Turkey, Ukraine, ...] |
| Latin America & Caribbean | [Aruba, Argentina, Antigua and Barbuda, The Bahamas, Belize, Bolivia, Brazil, Barbados, Chile, Colombia, Costa Rica, Cuba, Curaçao, Cayman Islands, Dominican Republic, Ecuador, El Salvador, French Guiana, Grenada, Honduras, Jamaica, Martinique, Mexico, Nicaragua, Panama, Saint Lucia, Uruguay, Venezuela, ...] |
| Middle East & North Africa | [United Arab Emirates, Bahrain, Djibouti, Algeria, Egypt, Iran, Iraq, Israel, Jordan, Kuwait, Lebanon, Libya, Morocco, Malta, Oman, West Bank and Gaza Strip, ...] |
| North America | [Bermuda, Canada, United States] |
| South Asia | [Afghanistan, Bangladesh, Bhutan, India, Sri Lanka, Maldives, Nepal, Pakistan] |
| Sub-Saharan Africa | [Angola, Burundi, Benin, Burkina Faso, Botswana, Central African Republic, Côte d'Ivoire, Cameroon, Dem. Rep. Congo, Congo, Cabo Verde, Eritrea, Ethiopia, ...] |



Pour le fichier Stats Series (Indicateurs):

- La colonne "Topic" classe les 3665 indicateurs en 37 catégories distinctes.
- La colonne "Longue Définition" offre une compréhension approfondie des indicateurs.
- La Périodicité est notable car elle inclut seulement 99 indicateurs avec une mesure annuelle, dont 54 se rapportent à la population totale.

```
['GDP at market prices (constant 2005 US$)',  
 'GDP at market prices (current US$)',  
 'GDP per capita (constant 2005 US$)',  
 'GDP per capita (current US$)',  
 'GDP per capita, PPP (constant 2011 international $)',  
 'GDP per capita, PPP (current international $)',  
 'GDP, PPP (constant 2011 international $)',  
 'GDP, PPP (current international $)',  
 'GNI (current US$)',  
 'GNI per capita, Atlas method (current US$)',  
 'GNI per capita, PPP (current international $)',  
 'GNI, PPP (current international $)',  
 'Internet users (per 100 people)',  
 'Labor force, total',  
 'Mortality rate, under-5 (per 1,000)',  
 'Personal computers (per 100 people)',  
 'Population growth (annual %)',  
 'Population, age 0, total']
```

```
'Population, age 0, total',  
 'Population, age 1, total',  
 'Population, age 10, total',  
 'Population, age 11, total',  
 'Population, age 12, total',  
 'Population, age 13, total',  
 'Population, age 14, total',  
 'Population, age 15, total',  
 'Population, age 16, total',  
 'Population, age 17, total',  
 'Population, age 18, total',  
 'Population, age 19, total',  
 'Population, age 2, total',  
 'Population, age 20, total',  
 'Population, age 21, total',  
 'Population, age 22, total',  
 'Population, age 23, total',  
 'Population, age 24, total',  
 'Population, age 25, total',  
 'Population, age 3, total',  
 'Population, age 4, total',  
 'Population, age 5, total',  
 'Population, age 6, total',  
 'Population, age 7, total',  
 'Population, age 8, total',  
 'Population, age 9, total'
```

Analyse:

- Suite à un calcul, il apparaît que les indicateurs annuels sont renseignés à hauteur de 47%, tandis que les autres ne le sont qu'à 7%.
- Ces 54 indicateurs sont présents dans plus de 180 pays de la DataFrame df_Data.
- Les 45 indicateurs restants sont identiques, mais différenciés par le genre (une information utile pour des initiatives commerciales à venir).



Choix des indicateurs du fichier Stats Series conserver pour l'entreprise:

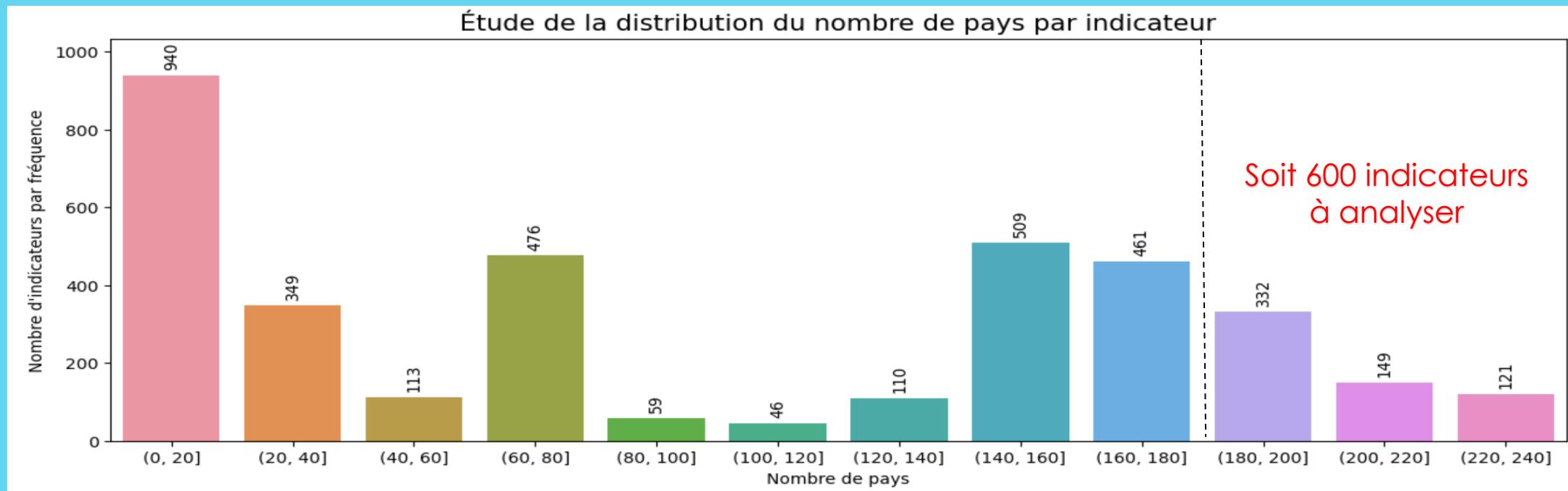
- Ratio d'ordinateurs et d'accès internet par 100 personnes : pertinent pour la commercialisation de cours en ligne.
- PIB par habitant : crucial car il influence les coûts associés aux cours en ligne.
- Main-d'œuvre totale et taux de chômage : utile pour ceux cherchant une mise à niveau professionnelle, surtout si l'aide gouvernementale est limitée.
- Population totale et croissance démographique : impact sur la taille nationale et le développement futur.
- Indicateurs basés sur l'âge : création d'indicateurs en fonction du niveau d'éducation pour évaluer la clientèle actuelle et potentielle.
- Autres indicateurs personnalisés : nombre d'utilisateurs d'internet et d'enfants hors système scolaire.

Pour le fichier EdStats Data (fichier principal):

- Conserver les colonnes de l'année 2000 à l'année 2020.
- Les données antérieures à l'année 2000 sont trop anciennes pour être significatives et fournir une perspective actuelle.
- Les données post-2020 sont des extrapolations, étant donné que nous sommes actuellement en 2024.



- Conserver les indicateurs englobant plus de 180 pays.



□ Diverses approches commerciales :

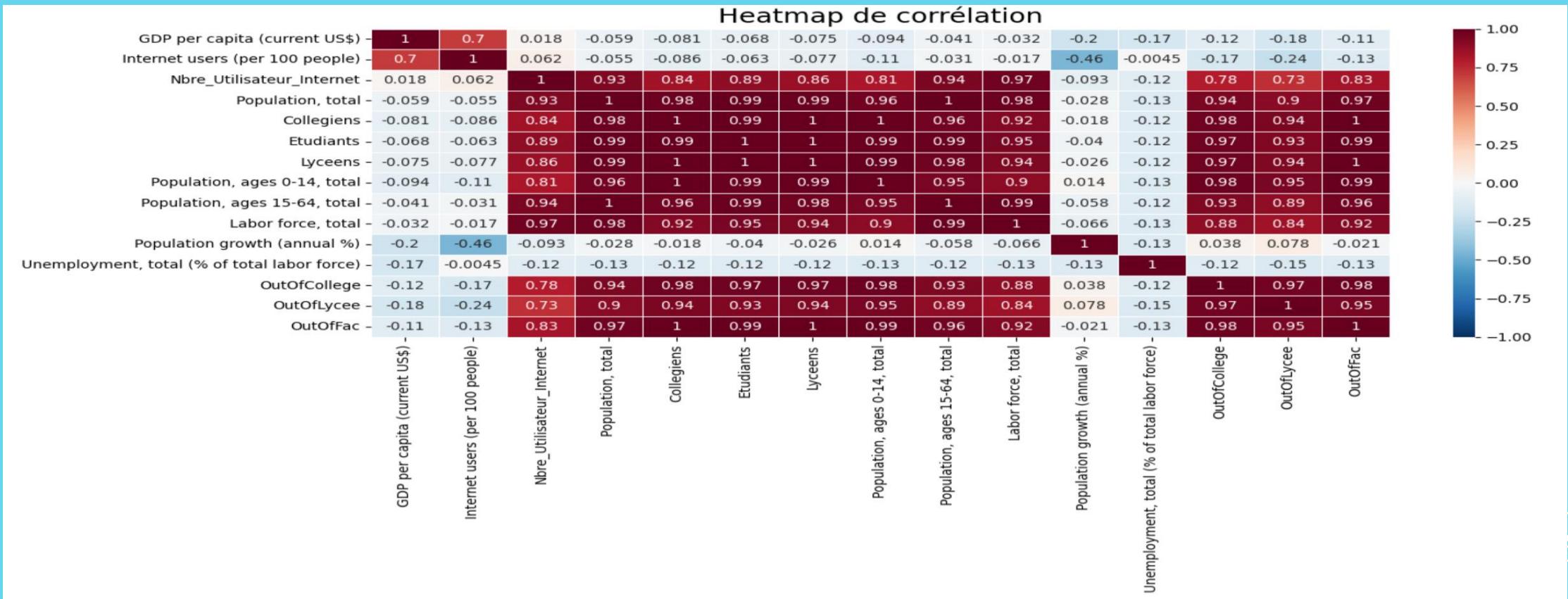
- Actuel nombre de clients disponibles pour ces cours.
- Potentiel nombre de clients futurs.
- Orientation des cours vers une mise à niveau (travailleurs en activité et chômeurs) ou spécifiquement pour les étudiants.
- Les deux autres fichiers présentent un intérêt, cependant, ils souffrent d'un manque d'informations crucial pour une compréhension approfondie. De plus, ils ne couvrent pas tous les indicateurs, années et pays de manière exhaustive.



5- Analyse des statistiques globales

| Region | East Asia & Pacific | Europe & Central Asia | Latin America & Caribbean | Middle East & North Africa | North America | South Asia | Sub-Saharan Africa | |
|--|---------------------|-----------------------|---------------------------|----------------------------|------------------|------------------|--------------------|--|
| Indicator Name | | | | | | | | |
| Collegiens | 4572561.300000 | 866627.300000 | 1411113.000000 | 1424183.700000 | 9363668.000000 | 16977249.100000 | 2003717.200000 | |
| Etudiants | 8927181.700000 | 1741995.300000 | 2370430.700000 | 2518463.100000 | 17899803.000000 | 28940538.200000 | 2812054.200000 | |
| GDP per capita (current US\$) | 15452.400000 | 28253.900000 | 10471.000000 | 15628.000000 | 61844.300000 | 2725.500000 | 2196.200000 | |
| Internet users (per 100 people) | 45.900000 | 72.500000 | 53.100000 | 55.400000 | 87.100000 | 25.600000 | 18.200000 | |
| Labor force, total | 42685180.400000 | 8891865.700000 | 9716732.100000 | 6854207.000000 | 90255019.500000 | 83628353.500000 | 8525866.000000 | |
| Lyceens | 3618886.600000 | 660198.900000 | 1045748.900000 | 1029320.000000 | 7113575.500000 | 12670593.000000 | 1369409.700000 | |
| Nbre_Utilisateur_Internet | 33486858.900000 | 12144127.700000 | 9256228.400000 | 9070581.300000 | 90340440.300000 | 50591095.400000 | 3845841.200000 | |
| OutOfCollege | 1573749.500000 | 44141.600000 | 170758.400000 | 444521.900000 | 2533880.000000 | 6617774.700000 | 1062460.200000 | |
| OutOfFac | 8990811.000000 | 927932.400000 | 1995482.600000 | 1713848.500000 | 13015004.000000 | 24305765.800000 | 2379717.200000 | |
| OutOfLycee | 1189521.300000 | 50602.000000 | 249192.700000 | 277232.700000 | 640612.500000 | 4613956.800000 | 865530.800000 | |
| Population growth (annual %) | 1.100000 | 0.400000 | 0.900000 | 2.300000 | 0.600000 | 1.600000 | 2.600000 | |
| Population, ages 0-14, total | 14567832.800000 | 3275381.300000 | 4612273.300000 | 6147492.300000 | 33690641.500000 | 64838693.900000 | 9160693.200000 | |
| Population, ages 15-64, total | 51216158.100000 | 12312542.300000 | 12039644.700000 | 13308844.900000 | 118297426.000000 | 141333867.500000 | 11470517.300000 | |
| Population, total | 62714691.200000 | 15919133.500000 | 15390836.800000 | 20426748.800000 | 118936821.000000 | 218023838.000000 | 21279529.000000 | |
| Unemployment, total (% of total labor force) | 6.100000 | 10.100000 | 8.700000 | 10.300000 | 6.100000 | 4.700000 | 9.500000 | |

- Nous constatons que les pays d'Amérique du Nord affichent le PIB par habitant le plus élevé, ainsi qu'un nombre élevé d'utilisateurs d'internet et une population active totale. Par conséquent, il est essentiel de prendre en considération cette région dans notre stratégie d'expansion à l'internationale de l'entreprise.

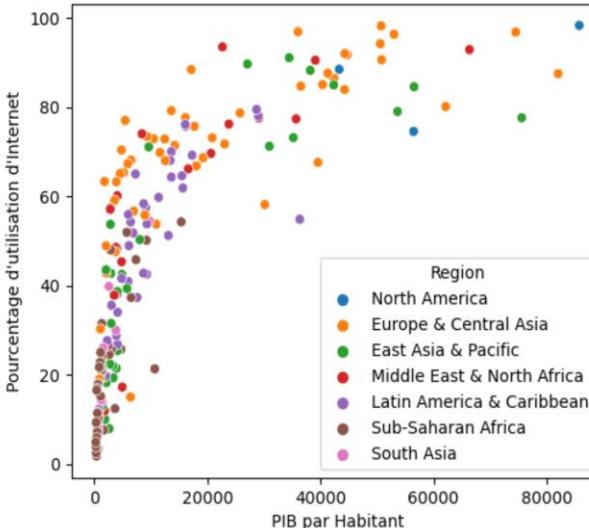


- Aucune corrélation n'est constatée entre les variables en pourcentage(%), soit la croissance démographique annuelle, le taux de chômage total et les autres indicateurs, bien que l'on note une inversion entre le pourcentage de la population utilisant Internet et le taux de croissance démographique (suggérant une possible causalité)
- Une corrélation entre le PIB par habitant et l'adoption d'Internet est observée.
- Des corrélations sont identifiées entre le nombre d'habitants, le nombre d'utilisateurs d'Internet, les effectifs scolaires (collégiens, lycéens, étudiants en études supérieures), la population dans les tranches d'âge 0-14 et 15-64, ainsi que la population active totale.

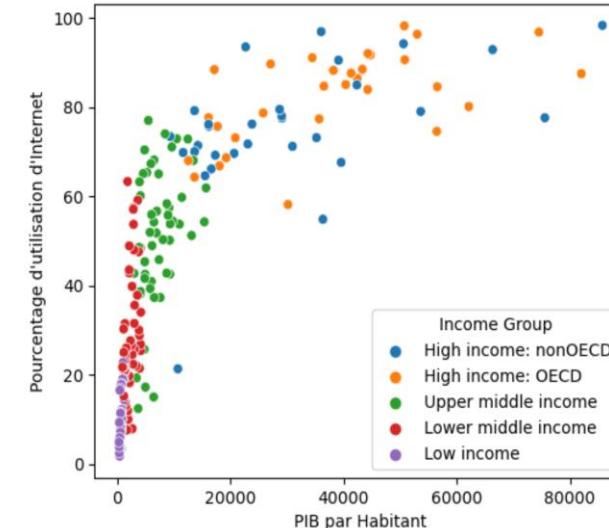


Etude entre les utilisateurs d'internet et du PIB, par région et par niveau de revenu (Income)

Étude de la corrélation entre les utilisateurs d'Internet (pour 100 habitants) et le PIB par habitant
Par Région



Par Groupe de Revenu



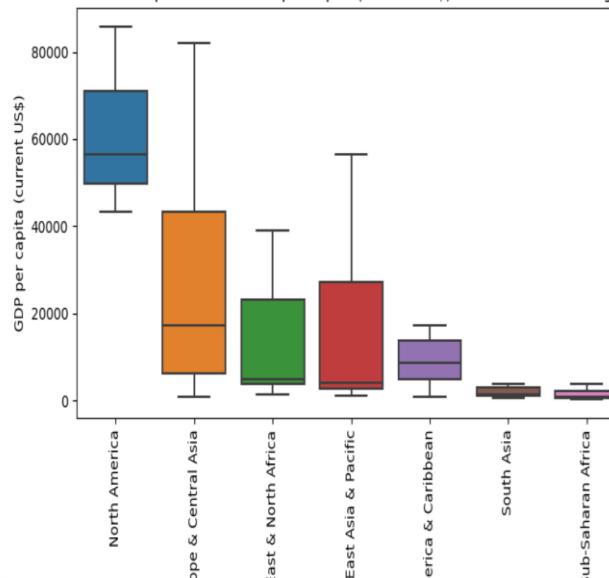
Observation:

- Les régions les plus prospères englobent à la fois les zones de l'OCDE et hors OCDE.
- Les zones géographiques et les niveaux de revenus qui affichent la plus grande utilisation d'Internet correspondent aux mêmes zones que celles présentant le PIB le plus élevé.
- Une corrélation significative de 0,7 est constatée entre ces deux indicateurs.

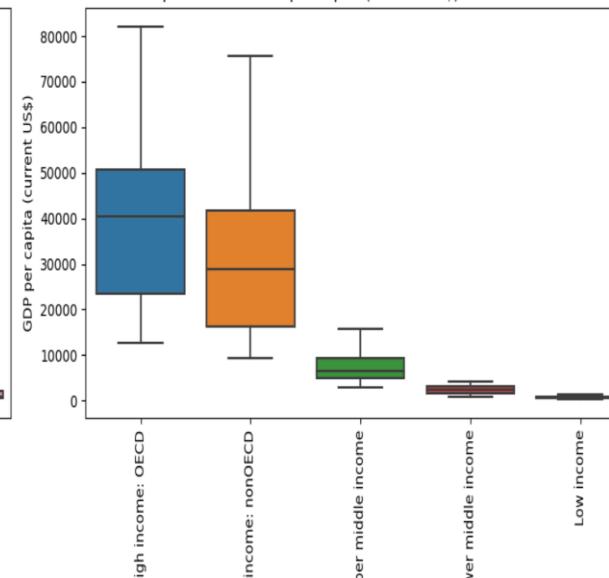
Conclusion:

- L'orientation privilégiée de l'entreprise pour son business plan devrait se porter vers les pays économiquement prospères.

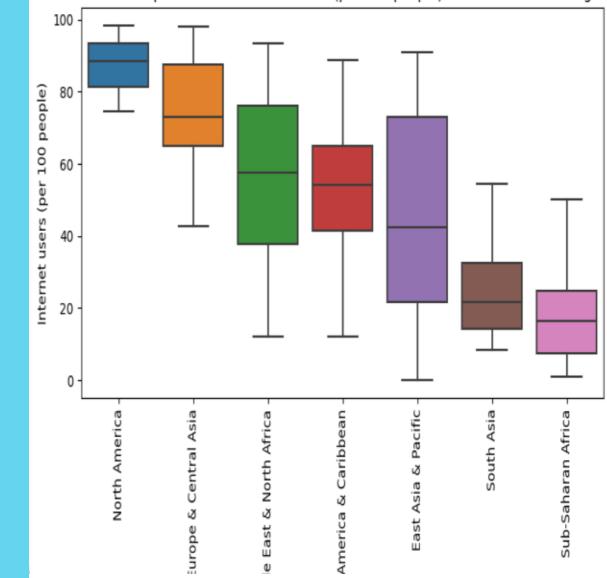
Étude de la répartition de "GDP per capita (current US\$)" en fonction : Des Régions.



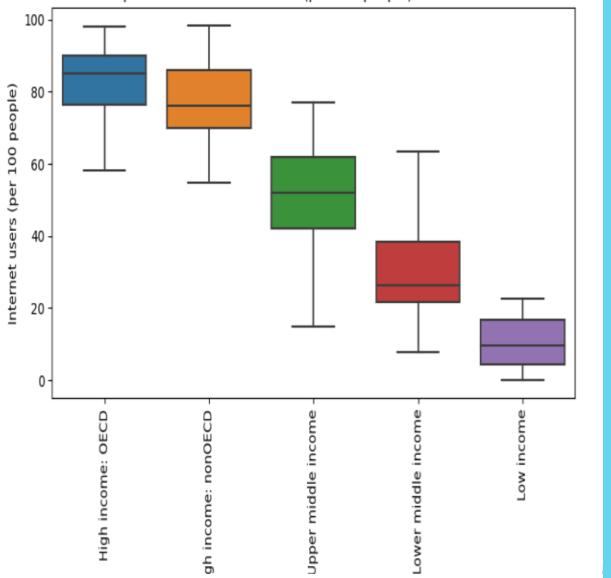
Étude de la répartition de "GDP per capita (current US\$)" en fonction : Des Incomes.



Étude de la répartition de "Internet users (per 100 people)" en fonction : Des Régions.

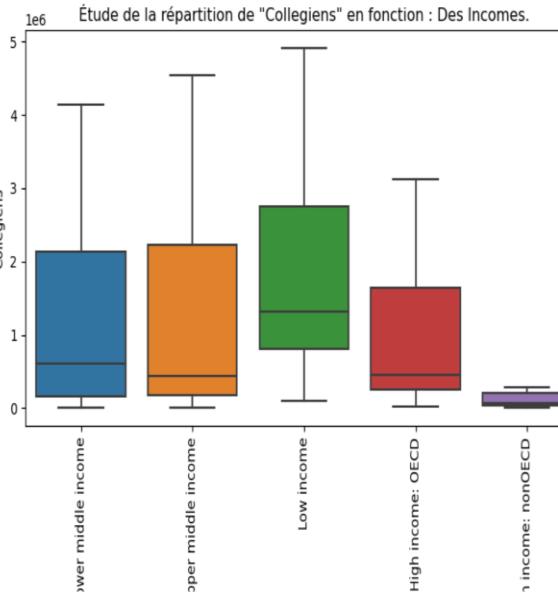
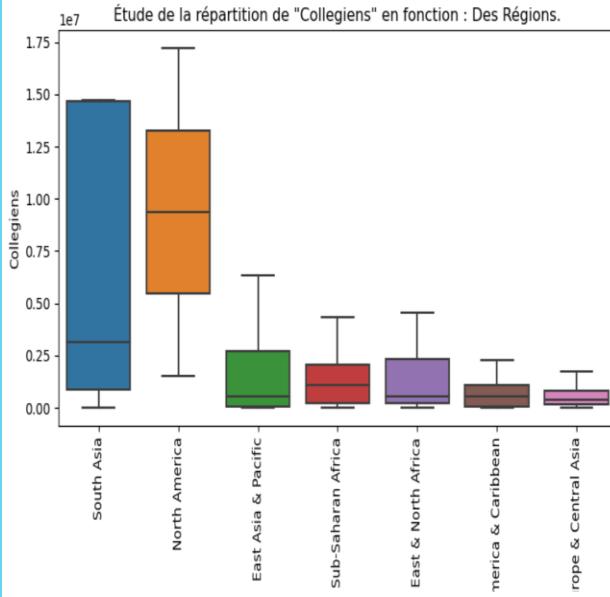


Étude de la répartition de "Internet users (per 100 people)" en fonction : Des Incomes.



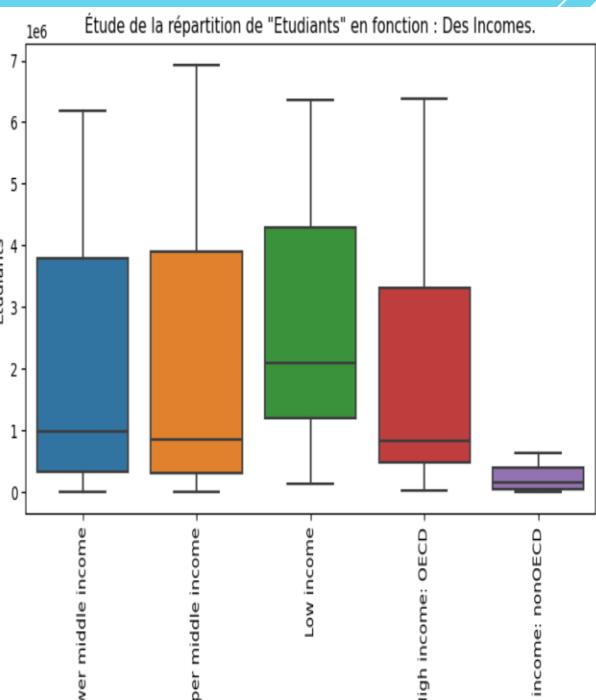
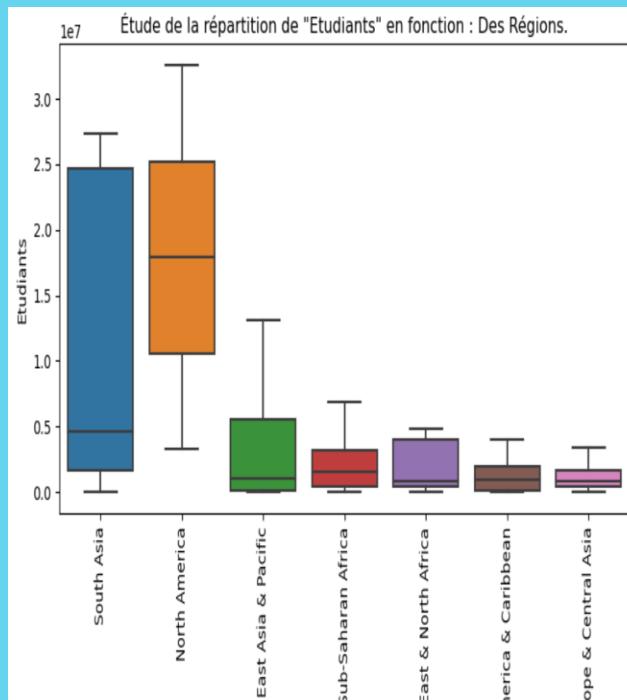
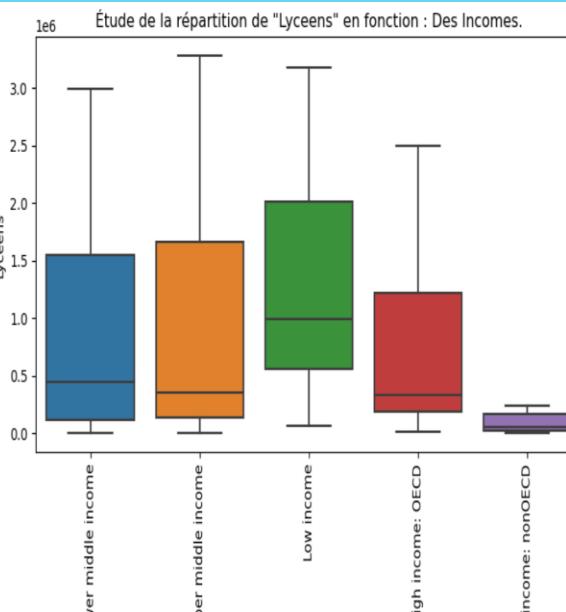
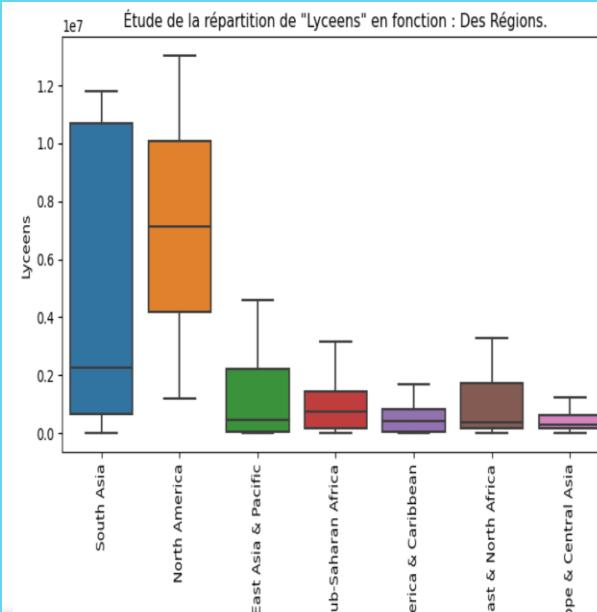


Etude des collégiens, lycéens et étudiants dans le système éducatif scolaire



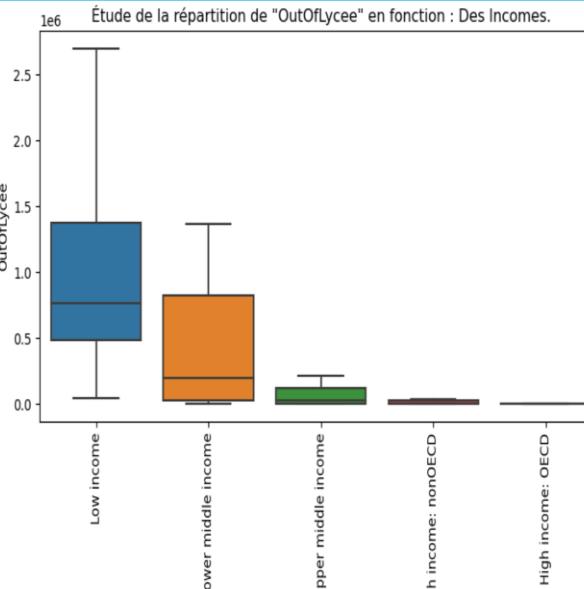
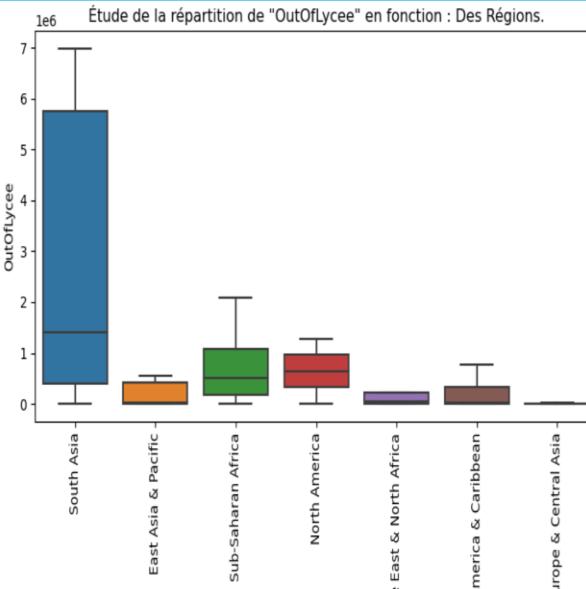
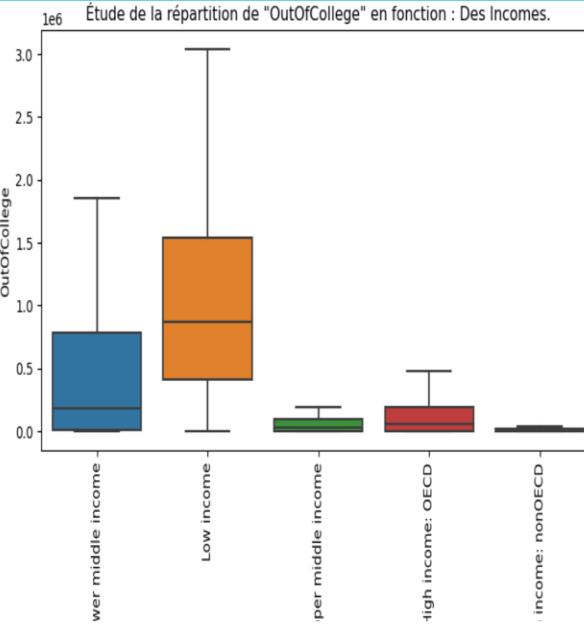
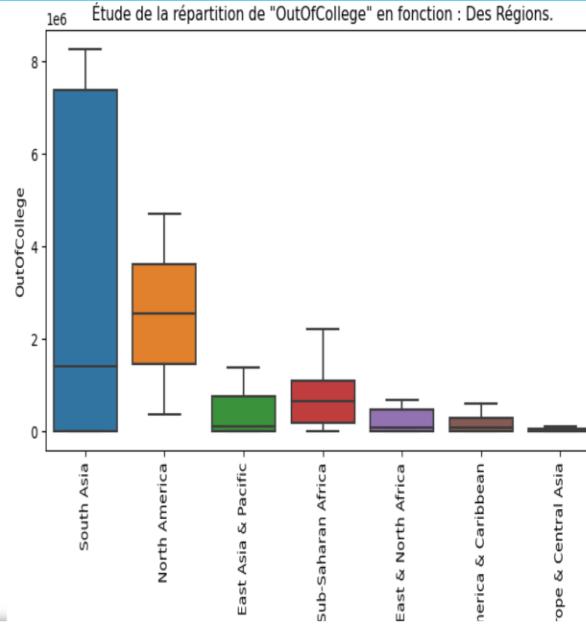
Observation:

- En Asie du Sud, en Amérique du Nord (en l'absence des données du Canada) et en Asie de l'Est et du Pacifique, on observe une forte fréquentation scolaire parmi les collégiens, lycéens et étudiants.
- Les régions les plus aisées affichent un nombre élevé de personnes scolarisées.



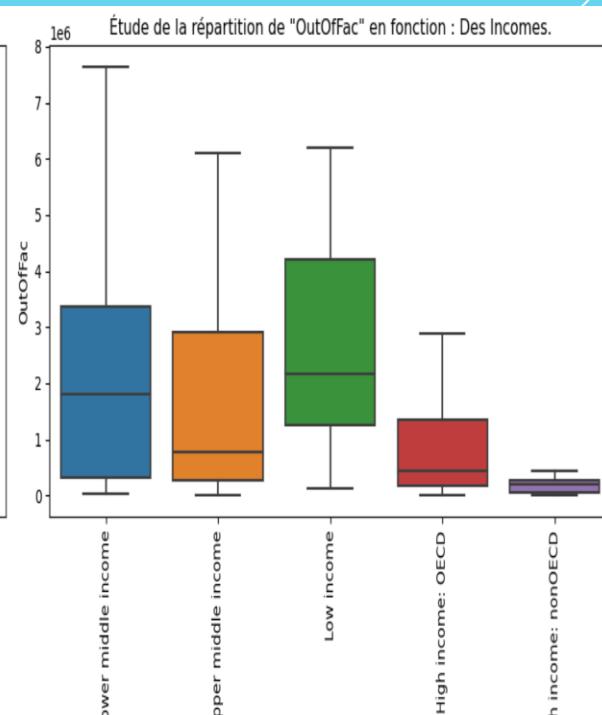
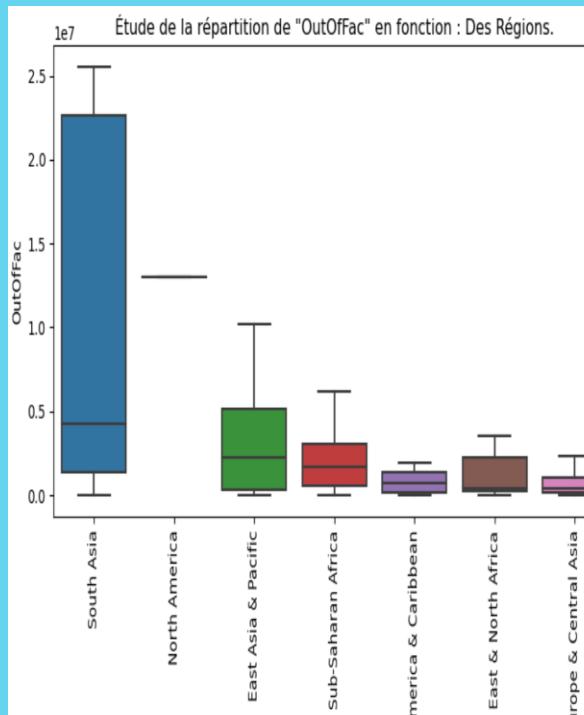


Etude des enfants, adolescents et jeunes ou adultes en dehors du système éducatif scolaire



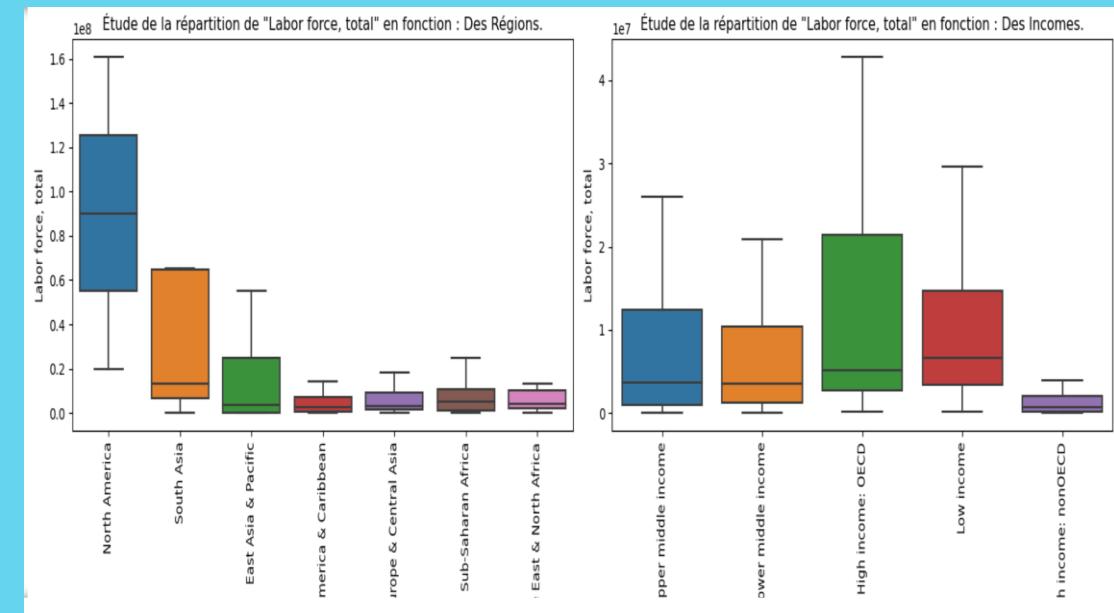
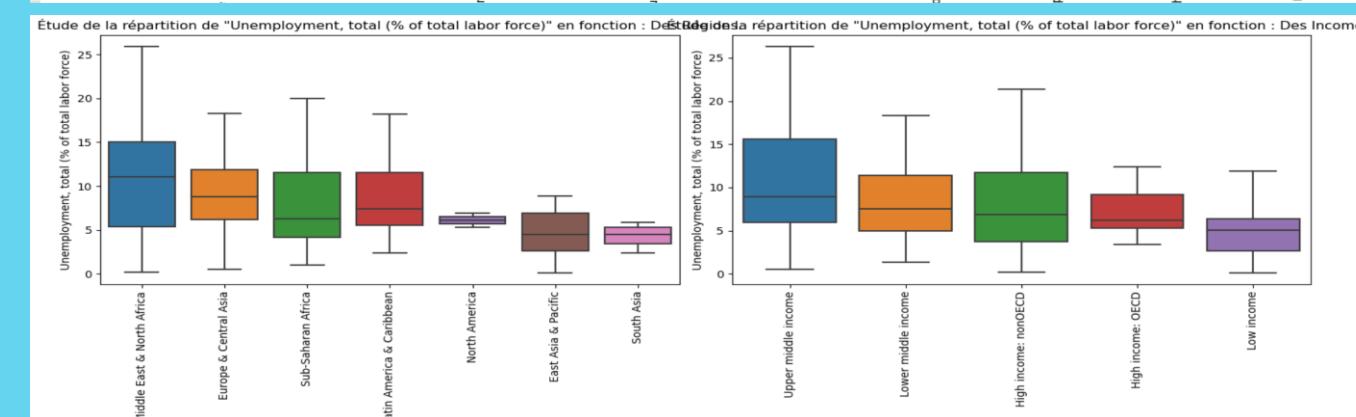
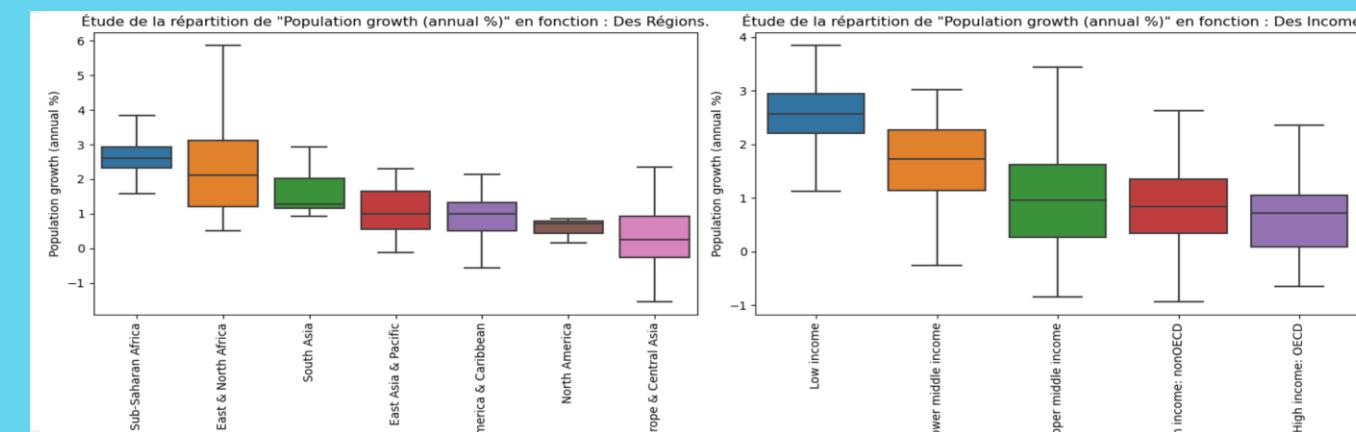
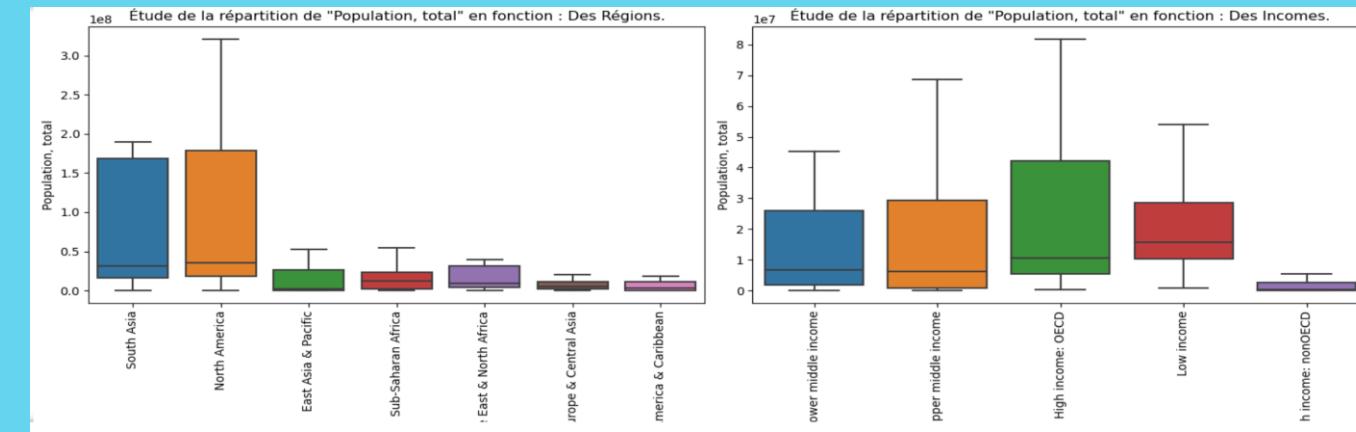
Observation:

- En Asie du Sud, en Amérique du Nord (à l'exception des données du Canada) et en Afrique, on observe une fréquentation scolaire moins importante chez les enfants.
- Les régions les plus défavorisées connaissent un taux d'absentéisme scolaire plus élevé.





Etude de la population totale, du taux de chômage, de la croissance démographique et de population active total



Observations:

- Les régions les plus densément peuplées se situent en Asie et en Amérique du Nord, ainsi que dans des zones moyennement riches.
- Les régions les plus affectées par le chômage incluent l'Afrique, l'Asie centrale, l'Europe et des zones moyennement riches.
- Les régions présentant la croissance démographique la plus importante sont l'Afrique et les pays des zones pauvres.
- L'Amérique du Nord et l'Asie du Sud se distinguent par la population active la plus importante parmi les différentes régions.



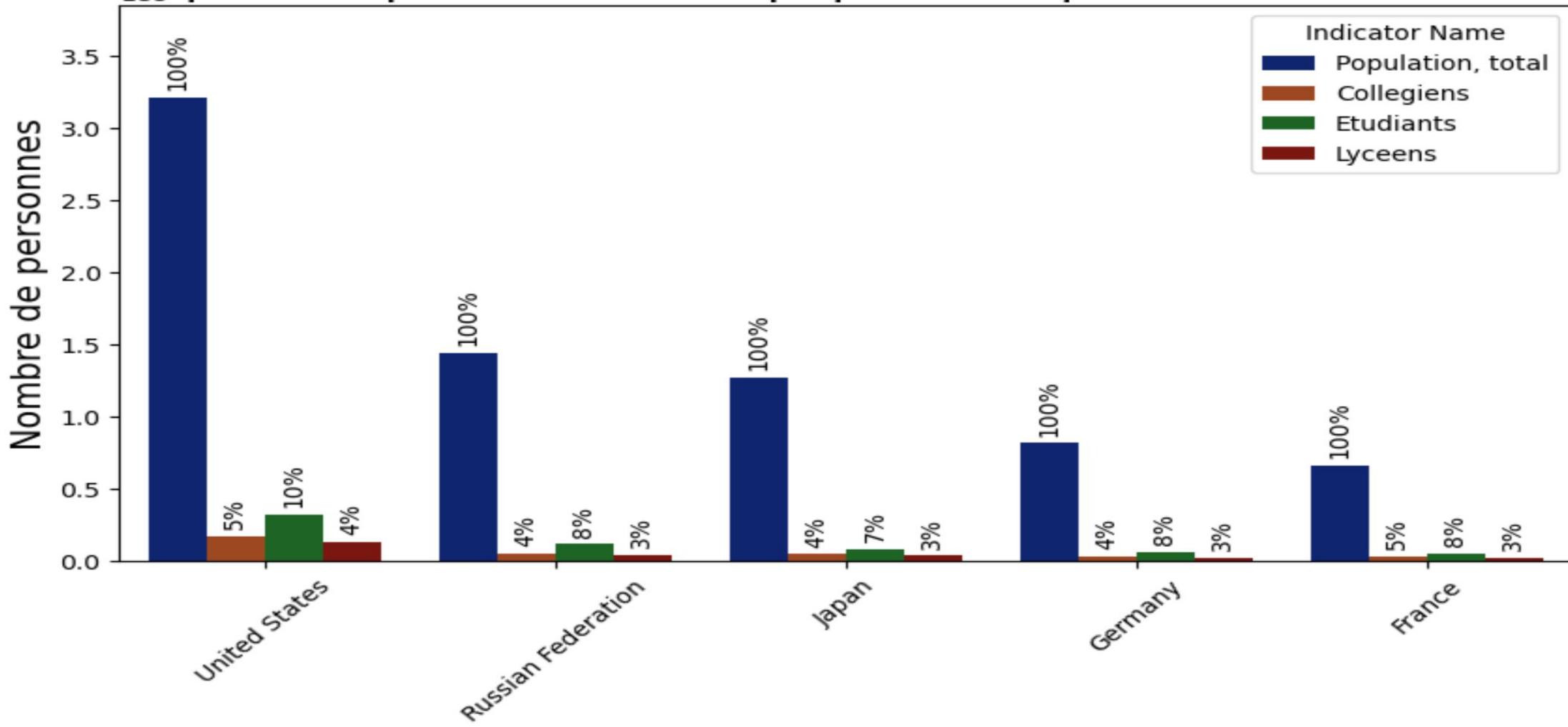
6- Analyse des besoins spécifique de l'entreprise

- L'analyse se concentre sur les pays les plus prospères pour lesquels des données sont disponibles.
- Étant donné la diversité linguistique de ces pays, il est recommandé de limiter la sélection aux pays ayant une population suffisamment importante pour justifier des investissements dans la création et la traduction d'un site web.
- Les données indiquent que parmi les 35 pays riches, la moitié compte plus de 3 millions d'habitants, avec une moyenne de 17 millions d'habitants.
- Par conséquent, les analyses ultérieures se basent sur les pays ayant une population supérieure à 3,5 millions d'habitants.

| | |
|-------|-------------|
| count | 75.0 |
| mean | 17318900.0 |
| std | 44617004.0 |
| min | 31264.0 |
| 25% | 162778.0 |
| 50% | 2904910.0 |
| 75% | 10078631.0 |
| max | 320896618.0 |

```
['Australia', 'Austria', 'Belgium', 'Chile', 'Croatia', 'Czech Republic', 'Denmark', 'Finland', 'France', 'Germany', 'Greece', 'Hong Kong SAR, China', 'Ireland', 'Israel', 'Italy', 'Japan', 'Korea, Rep.', 'Kuwait', 'Lithuania', 'Netherlands', 'New Zealand', 'Norway', 'Oman', 'Poland', 'Portugal', 'Russian Federation', 'Saudi Arabia', 'Slovak Republic', 'Spain', 'Sweden', 'Switzerland', 'United Arab Emirates', 'United Kingdom', 'United States', 'Uruguay']
```

Top 5 : Répartition de la population par zone éducative



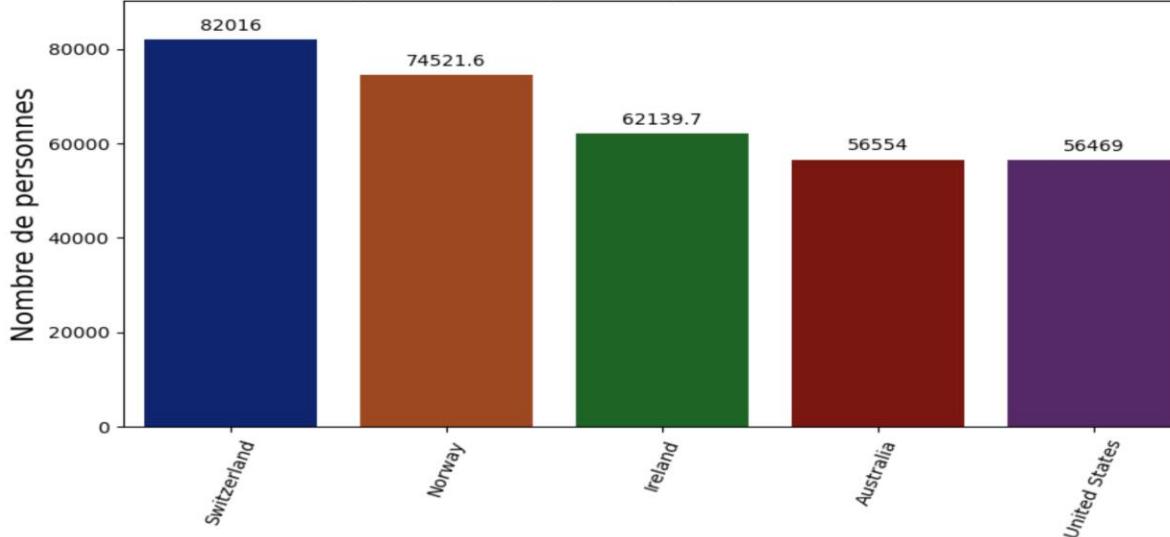
Observation:

- On observe une répartition équivalente de collégiens, lycéens et étudiants parmi les divers pays.

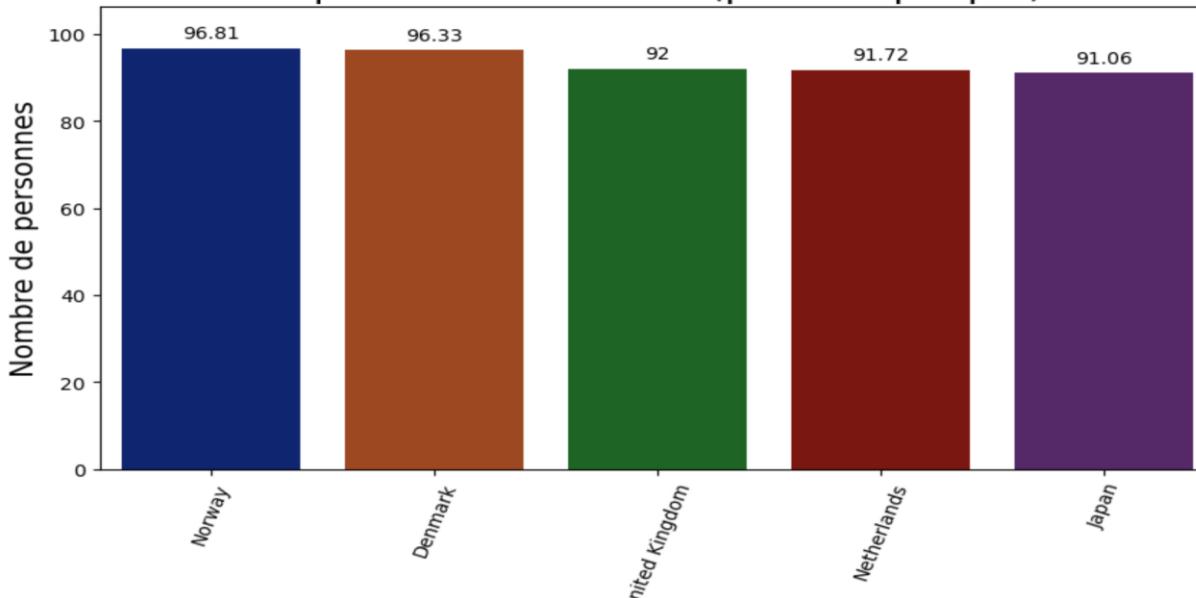


Etude du PIB et des utilisateurs d'internet

Top 5 : GDP per capita (current US\$)



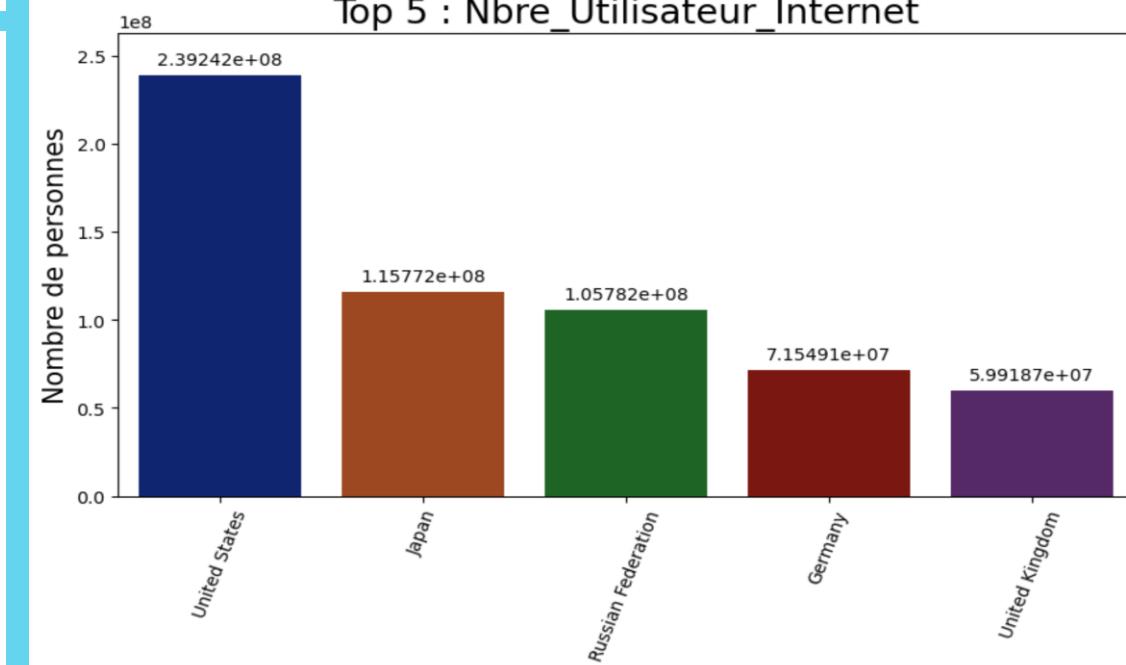
Top 5 : Internet users (per 100 people)



Observation:

- La Suisse se positionne en tant que pays le plus riche, suivi de près par la Norvège.
- La Norvège et le Danemark se distinguent en tant que principaux utilisateurs d'Internet.
- Cependant, en termes de nombre d'utilisateurs, les États-Unis et le Japon prennent la tête en raison de leur population plus importante.

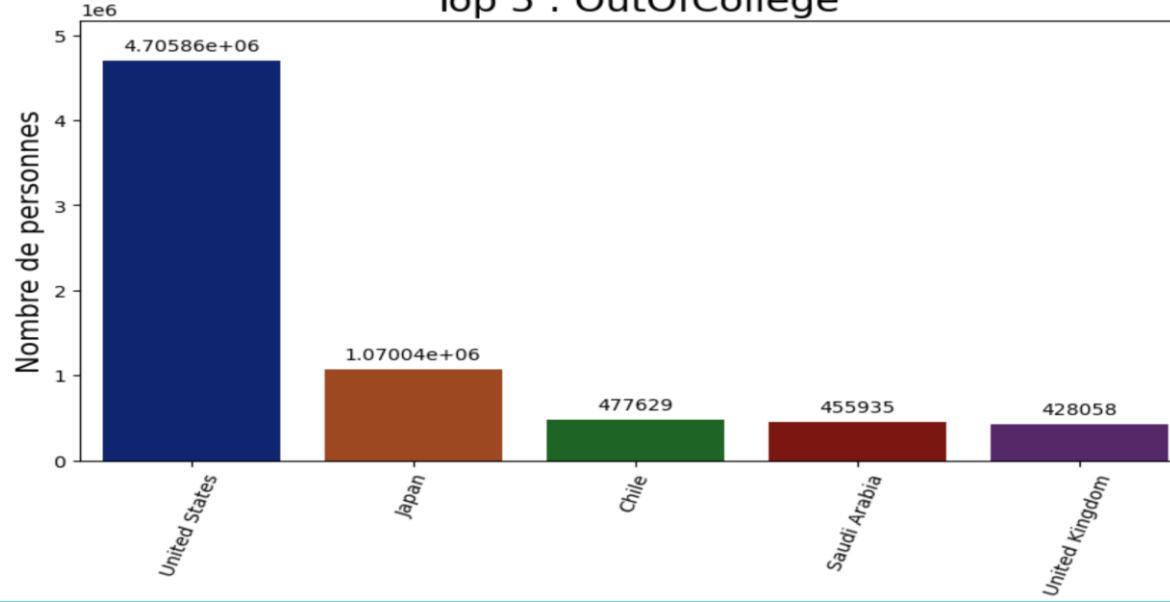
Top 5 : Nbre_Utilisateur_Internet



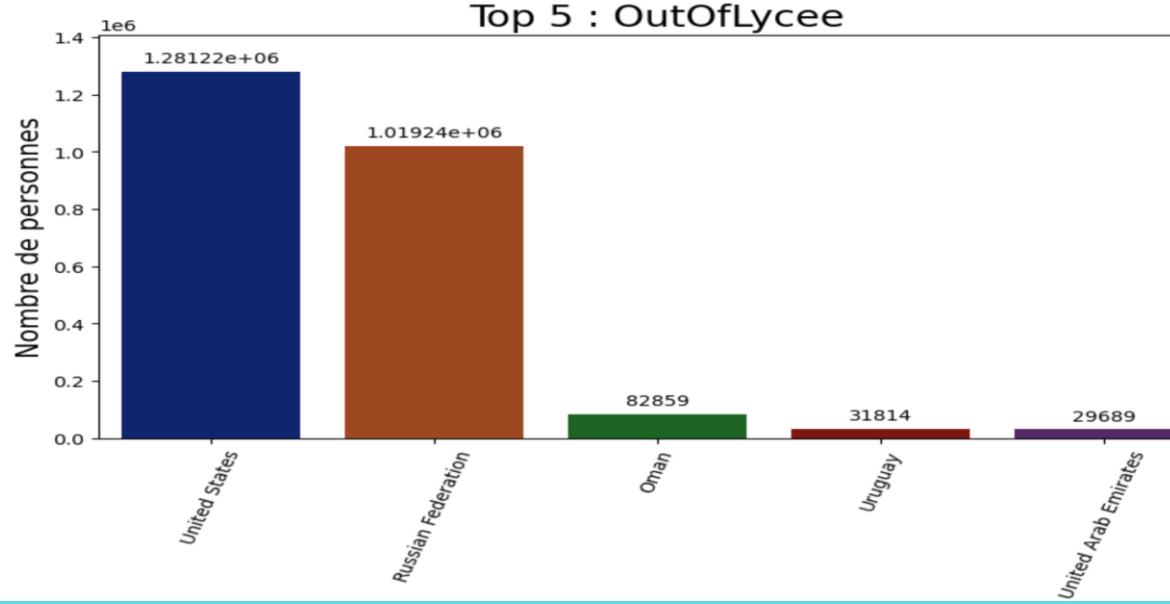


Etude des indicateurs OutOfEducation (hors éducation)

Top 5 : OutOfCollege



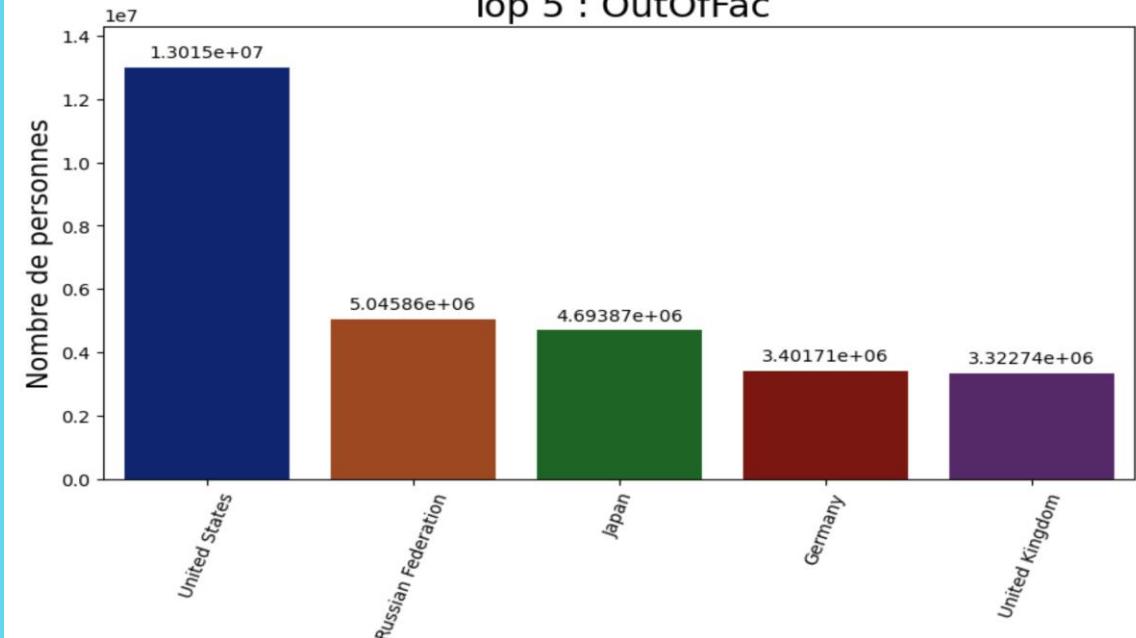
Top 5 : OutOfLycee



Observations:

- Les États-Unis comptent le plus grand nombre de collégiens, lycéens et étudiants qui ne fréquentent pas le système scolaire.
- Ensuite, la Russie et le Japon suivent dans cette catégorie.

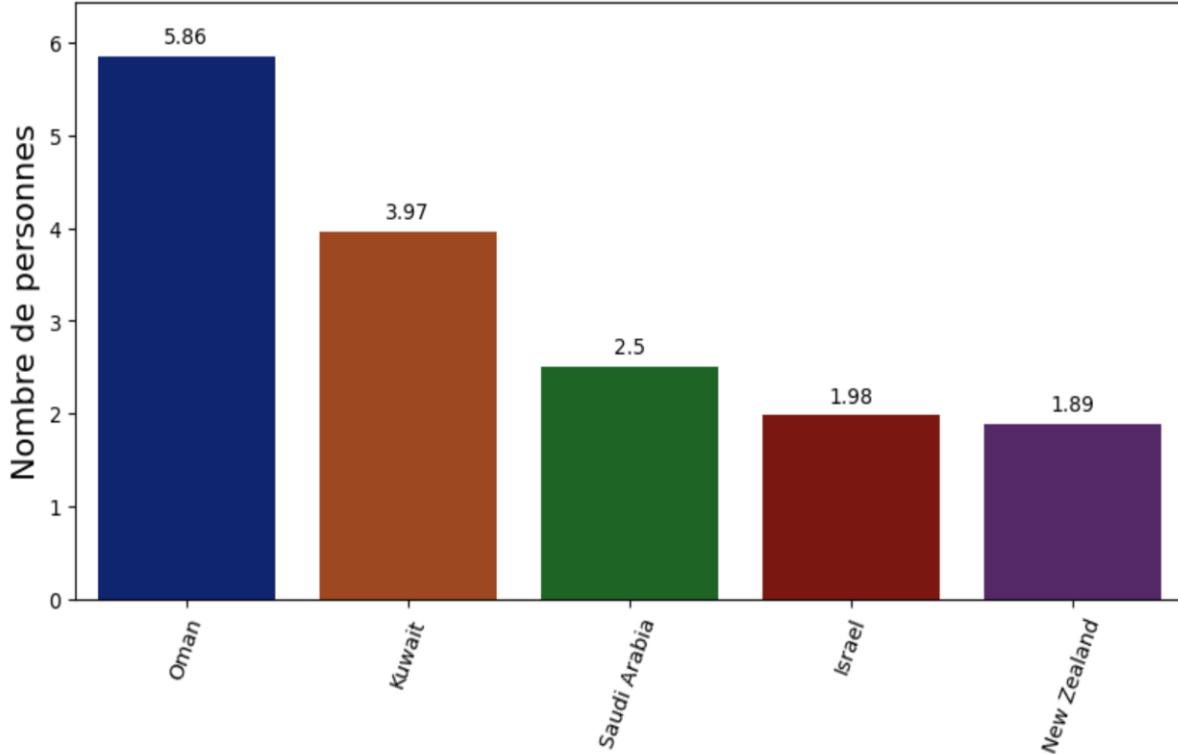
Top 5 : OutOfFac



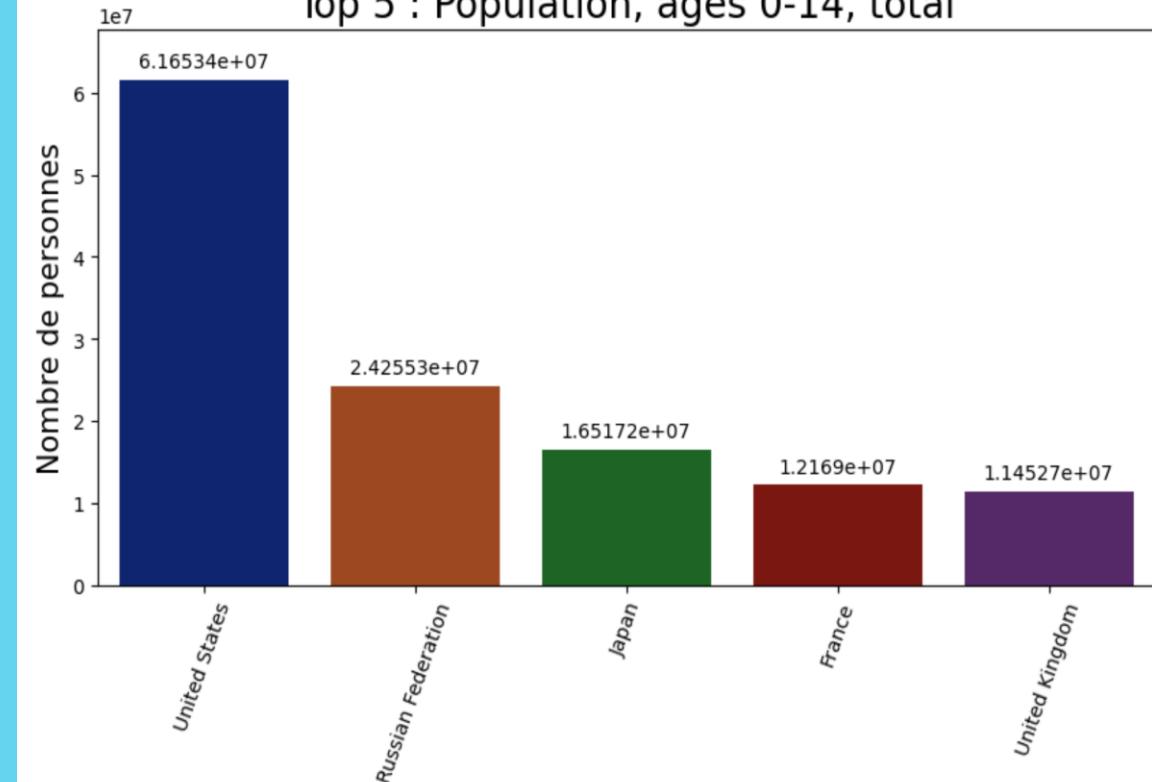


Etude des indicateurs des futurs clients

Top 5 : Population growth (annual %)



Top 5 : Population, ages 0-14, total



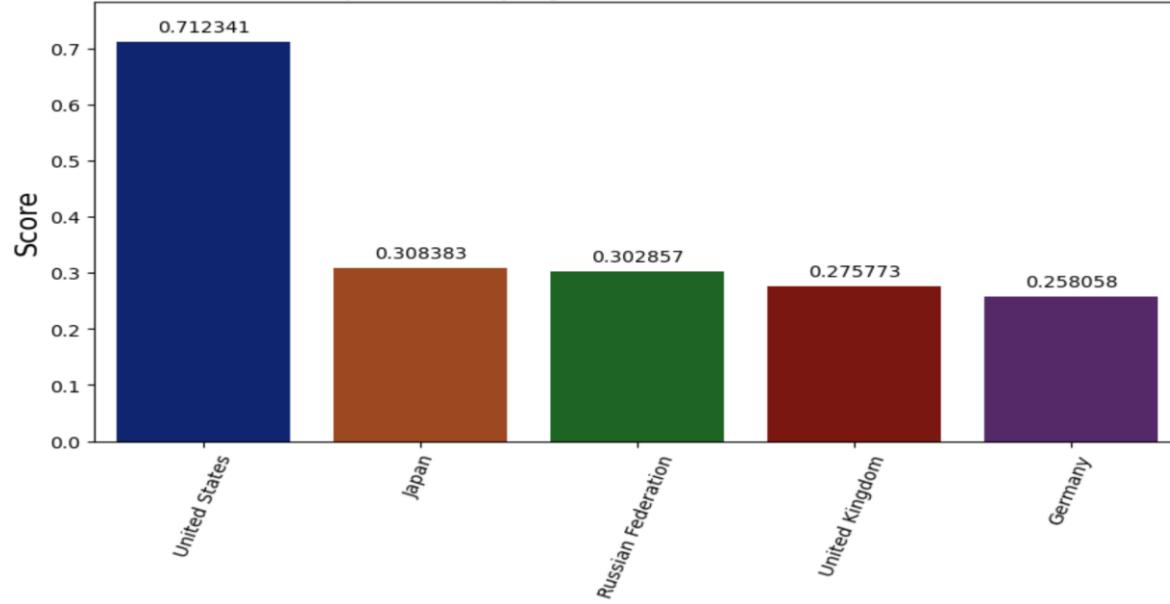
Observations:

- Oman et le Koweït se distinguent par la plus forte croissance démographique parmi les pays analysés.
- La majorité des pays les plus peuplés sont inclus dans l'étude de la population âgée de 0 à 14 ans.

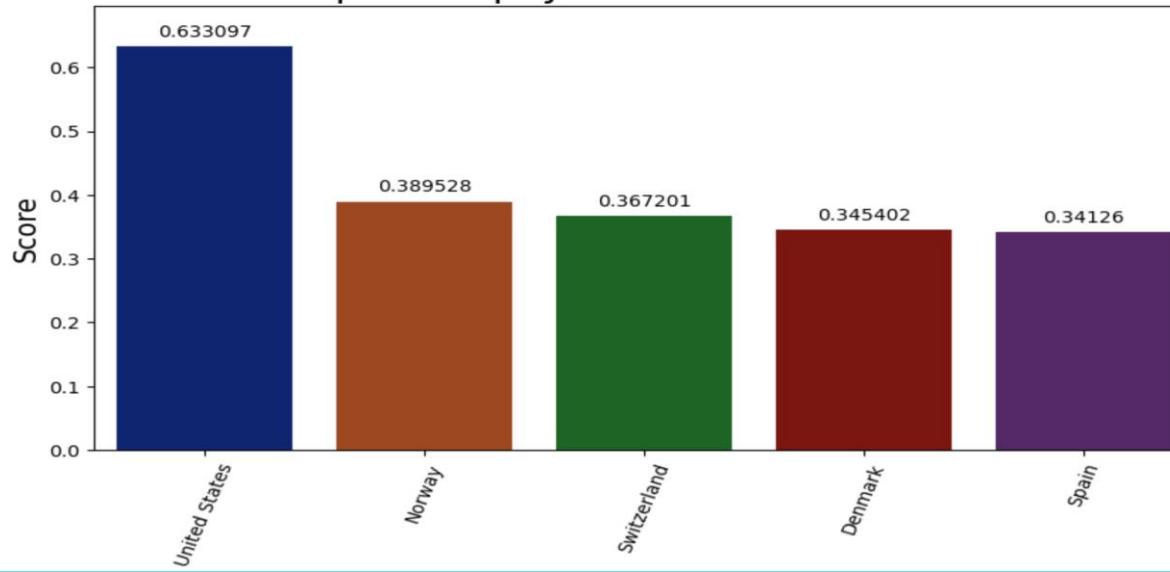


Etude des scores regroupant les indicateurs

Top 5 des pays où investir : Global

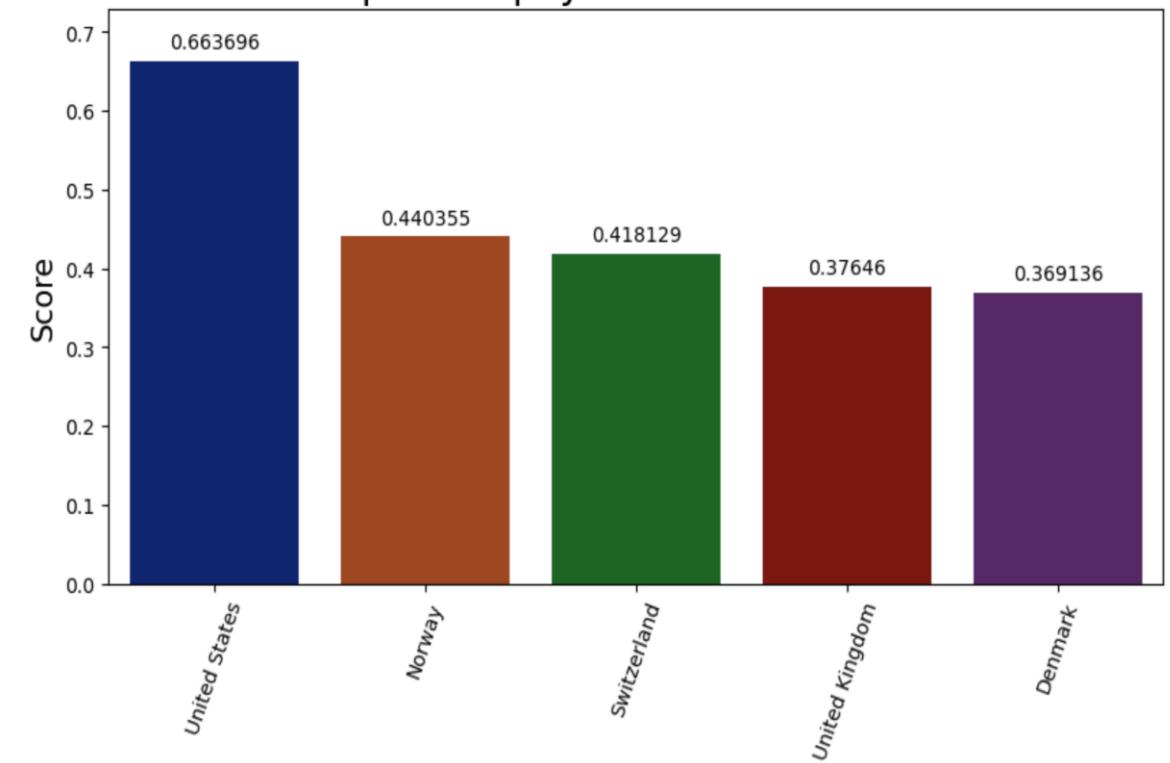


Top 5 des pays où investir : Present



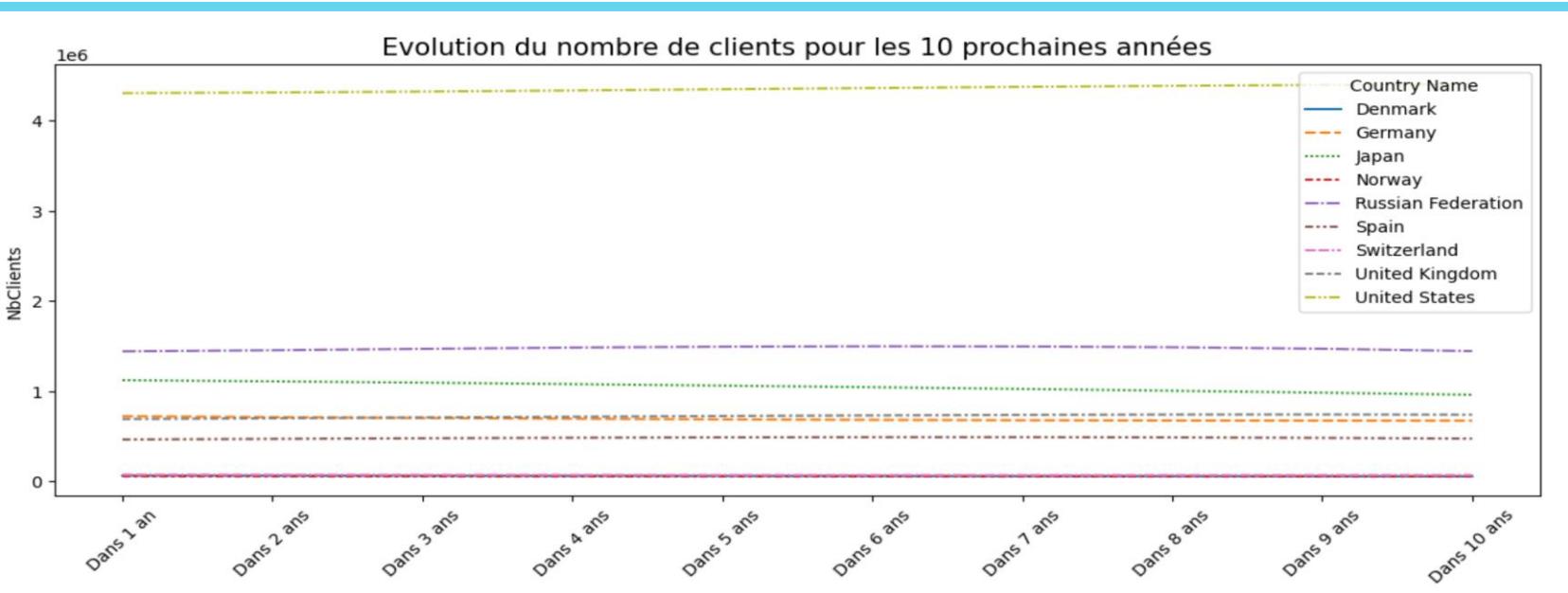
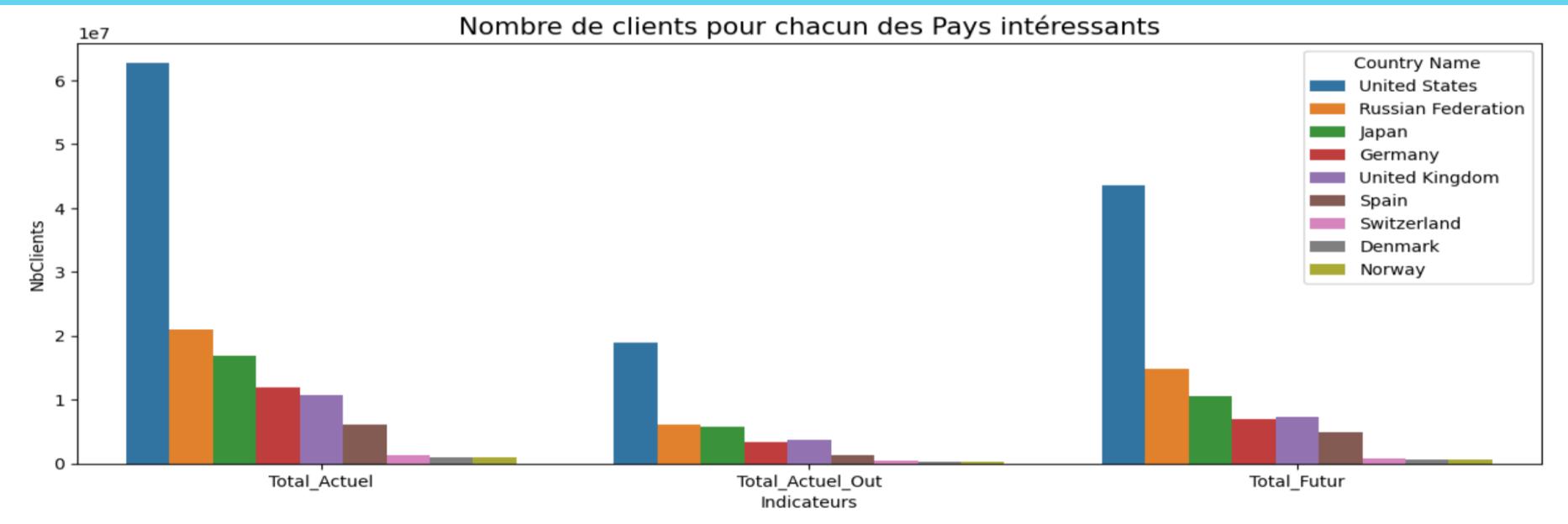
- En termes d'investissement, les pays à privilégier de manière générale sont les États-Unis, la Norvège, la Suisse, l'Angleterre, le Japon et le Danemark.

Top 5 des pays où investir : Futur





Les pays propices à l'investissement et à la clientèle future.



| Country Name | Indicateurs | Valeur |
|--------------------|------------------|------------|
| United States | Total_Actuel | 62763595.0 |
| United States | Total_Actuel_Out | 19002090.0 |
| United States | Total_Futur | 43546263.0 |
| Russian Federation | Total_Actuel | 20926472.0 |
| Russian Federation | Total_Actuel_Out | 6065103.0 |
| Russian Federation | Total_Futur | 14751594.0 |
| Japan | Total_Actuel | 16897623.0 |
| Japan | Total_Actuel_Out | 5763909.0 |
| Japan | Total_Futur | 10500115.0 |
| Germany | Total_Actuel | 11934828.0 |
| Germany | Total_Actuel_Out | 3401706.0 |
| Germany | Total_Futur | 6918085.0 |
| United Kingdom | Total_Actuel | 10658551.0 |
| United Kingdom | Total_Actuel_Out | 3750801.0 |
| United Kingdom | Total_Futur | 7249091.0 |
| Spain | Total_Actuel | 6099219.0 |
| Spain | Total_Actuel_Out | 1277354.0 |
| Spain | Total_Futur | 4829252.0 |
| Switzerland | Total_Actuel | 1302862.0 |
| Switzerland | Total_Actuel_Out | 473973.0 |
| Switzerland | Total_Futur | 729040.0 |
| Denmark | Total_Actuel | 979127.0 |
| Denmark | Total_Actuel_Out | 214007.0 |
| Denmark | Total_Futur | 617827.0 |
| Norway | Total_Actuel | 891979.0 |
| Norway | Total_Actuel_Out | 243604.0 |
| Norway | Total_Futur | 580120.0 |

Observation:

- Le nombre de clients potentiels est directement lié à la population totale des pays.



7- Point de vue personnel et limite

Point de vue personnel :

- Etant en 2024, la croissance exponentielle des utilisateurs d'internet entre 2000 et 2020 indique un potentiel fort pour les formations en ligne.
- Des régions comme l'Asie et l'Amérique du Nord, où cette croissance a été significative, pourraient être des marchés cibles stratégiques.
- Une analyse approfondie des pays avec une forte présence en ligne, comme la Norvège et le Danemark, pourrait révéler des opportunités stratégiques intéressantes.

Limites à prendre en compte :

- L'ajout de données récentes jusqu'à 2022 souligne la nécessité de rester attentif à une concurrence en constante évolution.
- Les nouveaux entrants et les changements rapides dans les préférences des consommateurs exigent une adaptabilité constante.
- La croissance des utilisateurs d'internet ne garantit pas automatiquement le succès des formations en ligne, nécessitant une analyse approfondie des facteurs tels que la diversité culturelle, les langues et les préférences d'apprentissage.
- La disponibilité des infrastructures technologiques et la situation économique des pays doivent être prises en compte pour assurer la qualité des services et la capacité financière du public cible.