Projet 2 : Analysez des données de systèmes éducatifs

15/05/2024

Soukaina GUAOUA ELJADDI

Parcours Data Scientist
OpenClassrooms

Plan:

- Problématique
- Description des jeux de données
- Sélection des indicateurs pertinents
- Nettoyage et Préparation du jeu de données
- Analyse statistique et graphique du jeu de données
- Conclusion

Problématique

Contexte : Projet d'expansion à l'international de formation en ligne de niveau lycée et supérieur d'une **start-up de la EdTech**, nommée ***academy***.

Objectif : Explorer les **pays à fort potentiel de clients**, leur **évolution** dans le temps. Proposer une liste de **pays prioritaires**.

Mission : **Analyse exploratoire** des données en se basant sur les données de la banque mondiale

Informations pertinentes? De qualité? Cohérentes?
Indicateurs statistiques pour départager les pays?
Conclusion : pays à fort potentiel ? Priorisation? Evolution?

Étape 1 : Menez une analyse générale des données

Description des jeux de données

EdStatsData.csv :

- Evolution des indicateurs par pays et par région sur plusieurs années avec projection (1970 à 2100) .
- 886930 lignes et 70 colonnes (85,8%NaN, Φ valeurs dupliquées)

EdStatsCountry.csv :

- Informations géographiques sur les pays par régions, données économiques globales et dates de référence
- 241 lignes et 32 colonnes (28,28% NaN, Φ valeurs dupliquées)

EdStatsCountry-Series.csv :

- Les références des sources des indicateurs par pays
- 613 lignes et 4 colonnes (Φ NaN, Φ valeurs dupliquées)

Description des jeux de données

EdStatsFootNote.csv:

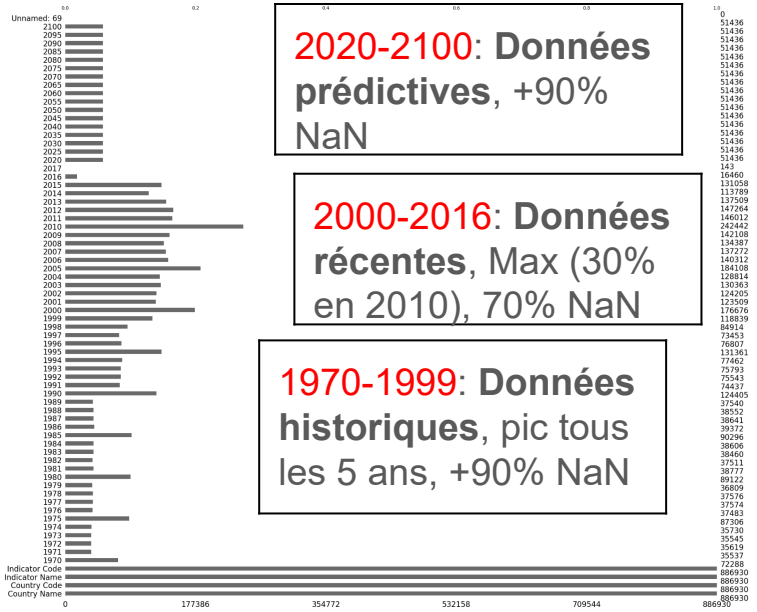
- Informations sur l'année de référence/incertitude des indicateurs par pays
- 643638 lignes et 5 colonnes (Φ NaN, Φ valeurs dupliquées)

EdStatsSeries.csv:

- Informations socio-éduco-économiques des indicateurs par thématique et leurs sources
- 3665 lignes et 21 colonnes (70.31%, Φ valeurs dupliquées)

Étape 2 : Sélectionnez les données pertinentes

Le jeu de données EdStatsData.csv



- 3665 indicateurs
- Plage temporelle 1970-2100
- 241 pays/régions/groupes
- **Extraction des données :**
  ⇨ Indicateurs pertinents
  ⇨ Plage temporelle significatif : 2010 à 2015
  ⇨ Variables utiles : Country Name, Indicator Code

Indicateurs pertinents

Sources : site banque mondiale, le fichier EdStatsSeries.csv

Mots clés : education, GDP, upper secondary, tertiary, Internet, computers, Population ages 15-24, Adult literacy

Économiques	Éducatifs	Numériques	Démographiques
SE.XPD.TOTL.GD.ZS : Dépenses publiques en éducation en pourcentage du PIB (%)	SE.SEC.ENRR.UP : Taux de scolarisation au niveau secondaire (Lycée) (%) SE.TER.ENRR : Taux de scolarisation au niveau universitaire (%)	IT.NET.USER.P2 : Utilisateurs Internet (pour 100 personnes) IT.CMP.PCMP.P2 : Ordinateurs personnels (pour 100 personnes)	SP.POP.1524.TO.UN : Population âgée de 15 à 24 ans, total SE.ADT.1524.LT.ZS : Taux d'alphabétisation des jeunes, population 15-24 ans, deux sexes (%)

Étape 3 : Créez une dataframe dans le but de mener une analyse

Nettoyage et Préparation du jeu de données

Etape 1	Etape 2	Etape 3	Etape 4	Etape 5
EdStatsData.csv : 886930 lignes, 70 colonnes Suppression colonne Unnamed: 69 ⇨ 886930 lignes 69 × colonnes	Filtration en fonction des 7 indicateurs choisis ⇨ 1694 lignes × 69 colonnes	Séparation des pays : 1519 lignes × 69 colonnes Filtration en fonction du plage temporaire (2010 à 2015) ⇨ 1519 lignes × 10 colonnes	Gestion des valeurs manquantes, en les remplaçant par la moyenne des autres valeurs indiquées ⇨ 1519 lignes × 10 colonnes	Suppression des lignes contenant que des valeurs manquantes ⇨ 1005 lignes × 10 colonnes

Nettoyage et Préparation du jeu de données

Création d'un pivot table pour l'année **2015** ⇨ 210 Lignes × 6 Colonnes

Indicator Code	IT.NET.USER.P2	SE.ADT.1524.LT.ZS	SE.SEC.ENRR.UP	SE.TER.ENRR	SE.XPD.TOTL.GD.ZS	SP.POP.1524.TO.UN
Country Name						
Afghanistan	8.260000	46.990050	42.613129	6.209390	3.31754	7252785.0
Albania	63.252933	99.011295	89.460274	58.109951	3.53944	556269.0
Algeria	38.200000	NaN	61.089111	36.922279	NaN	6467818.0
Andorra	96.910000	100.000000	NaN	NaN	3.25368	NaN
Angola	12.400000	77.431130	20.749240	9.308020	3.47623	4259352.0
...	...	...	...	...	...	...
Virgin Islands (U.S.)	54.839137	NaN	NaN	NaN	NaN	NaN
West Bank and Gaza	57.424192	99.340610	66.222519	44.283218	1.32232	1053004.0
Yemen, Rep.	24.085409	NaN	35.656956	10.274040	NaN	5995687.0
Zambia	21.000000	88.714560	NaN	3.993995	NaN	3068044.0
Zimbabwe	22.742818	90.679410	36.600655	8.433270	6.81781	3333716.0

210 rows × 6 columns

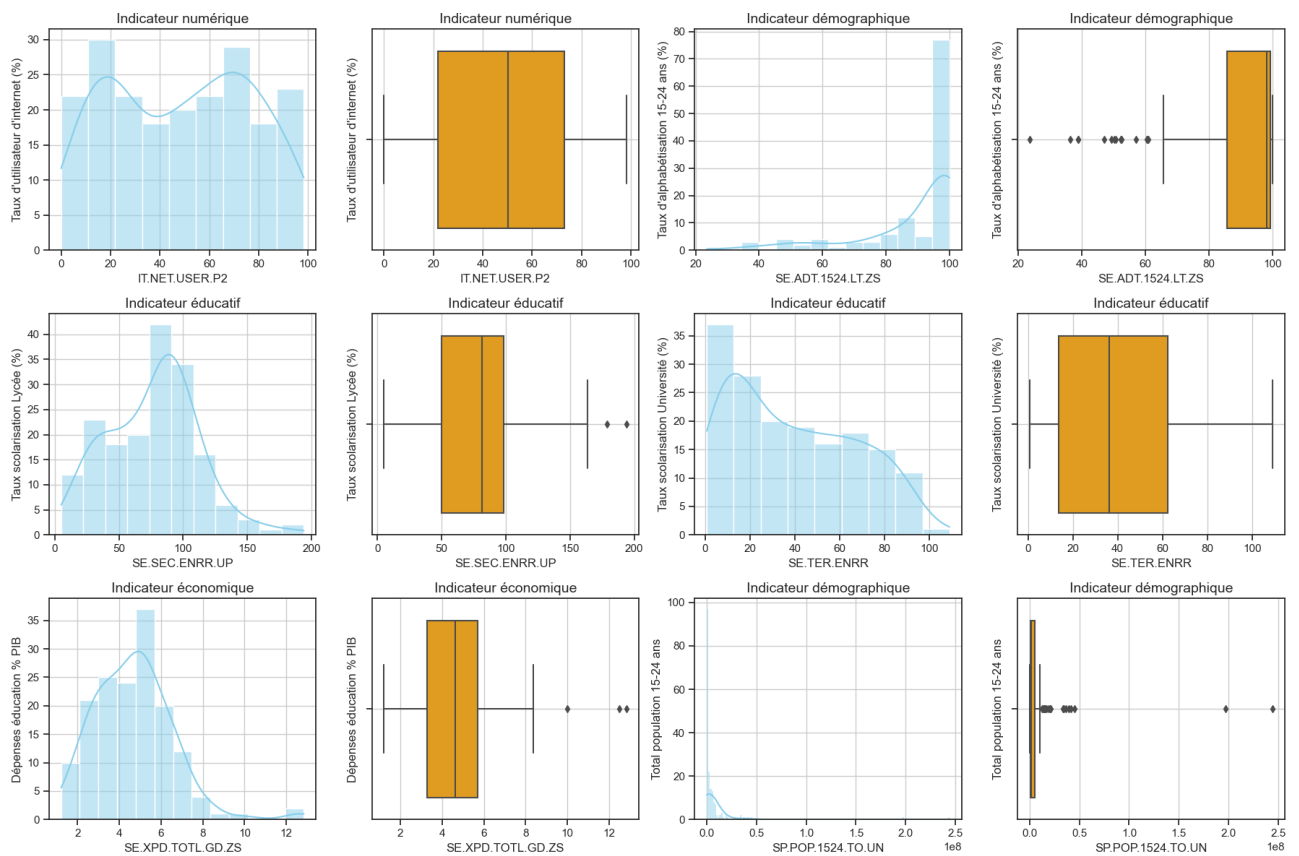
Etape 4 : Analysez le jeu de données

Analyse statistique :

Statistiques descriptives pour chaque indicateur

Indicator Code	IT.NET.USER.P2	SE.ADT.1524.LT.ZS	SE.SEC.ENRR.UP	SE.TER.ENRR	SE.XPD.TOTL.GD.ZS	SP.POP.1524.TO.UN
count	204.000000	121.000000	177.000000	165.000000	157.000000	1.810000e+02
mean	48.630114	89.158543	76.735901	39.361462	4.685562	6.654995e+06
std	28.558842	17.022032	35.421684	27.986252	1.891469	2.404044e+07
min	0.000000	23.523780	5.020860	0.745150	1.214010	1.445500e+04
25%	21.716942	85.716150	49.690262	13.602257	3.262120	4.897540e+05
50%	50.219659	98.287058	81.212189	36.228539	4.621675	1.331040e+06
75%	72.949675	99.300000	98.463112	62.300770	5.690636	4.945440e+06
max	98.323610	100.000000	194.101990	108.971404	12.837270	2.441202e+08

Analyse graphique :



Conclusions :

- Le taux de scolarisation dans l'enseignement supérieur est plus inégal entre les pays que le taux de scolarisation au niveau lycée.

- Pour l'indicateur numérique avec une forme bimodale, on distingue 2 groupes de pays répartis autour de 20% et 65%.

- Présence d'outliers pour les indicateurs démographique, économique et éducatif au niveau lycée.

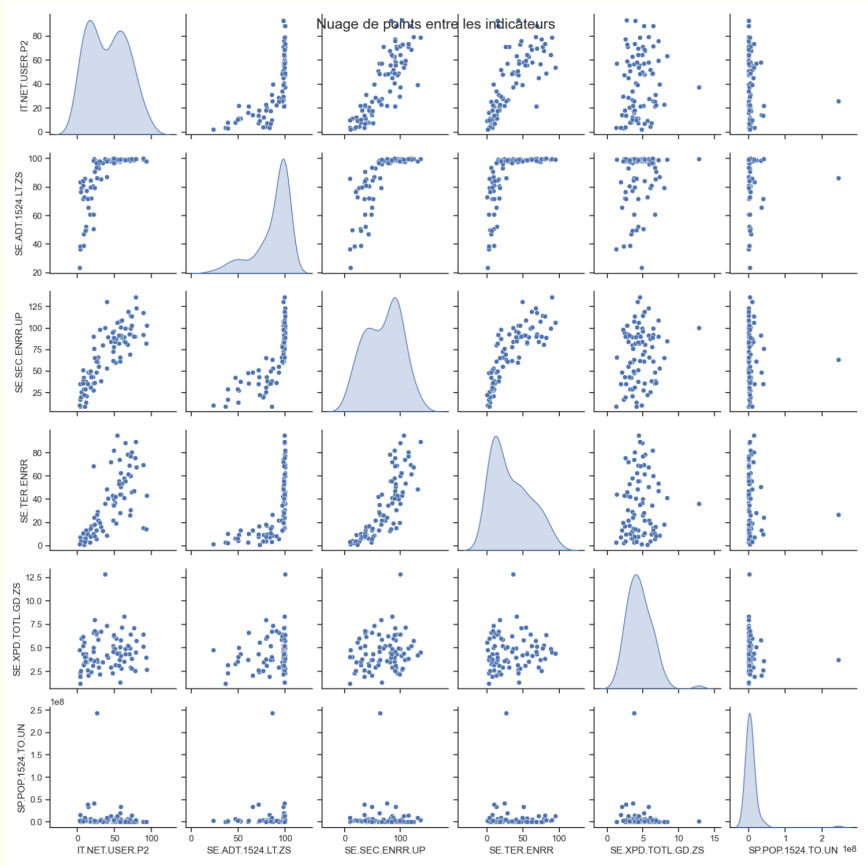
Analyse graphique :

Lien de corrélation entre :

- SE.SEC.ENRR.UP et SE.TER.ENRR

- IT.NET.USER.P2 et SE.SEC.ENRR.UP

- IT.NET.USER.P2 et SE.TER.ENRR



SE.XPD.TOTL.GD.ZS (Dépenses publiques en éducation en pourcentage du PIB))

SE.SEC.ENRR.UP (Taux de scolarisation au niveau secondaire (Lycée) (%)

SE.TER.ENRR (Taux de scolarisation au niveau universitaire)

IT.NET.USER.P2 (Utilisateurs Internet (pour 100 personnes)

SP.POP.1524.TO.UN (Population âgée de 15 à 24 ans, total)

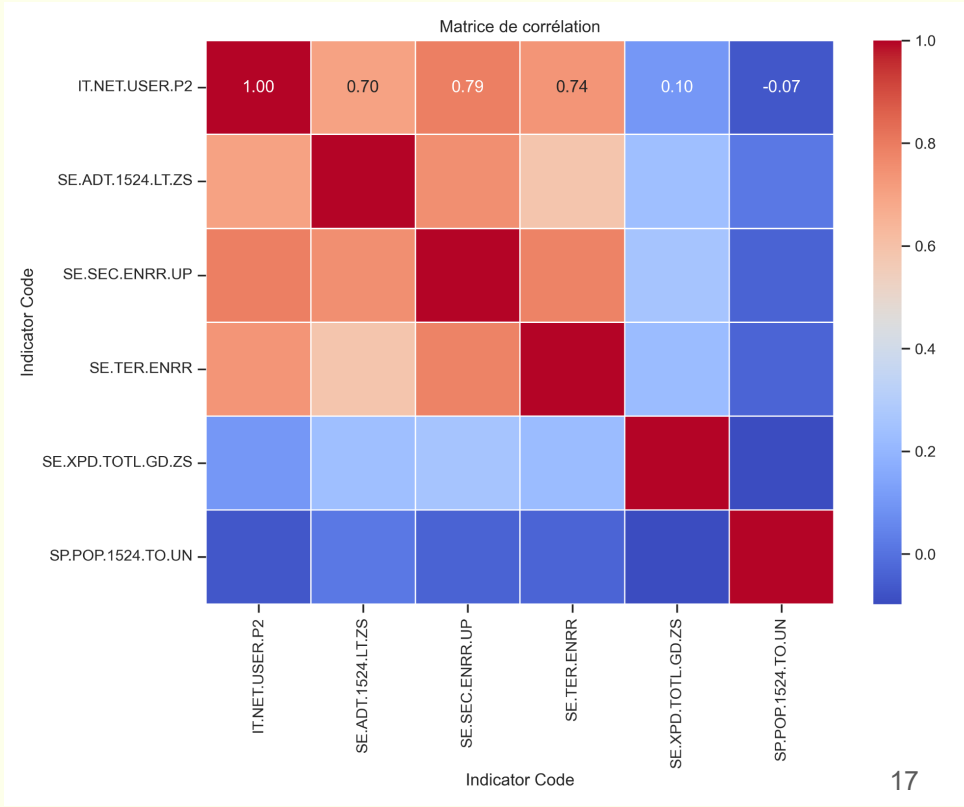
SE.ADT.1524.LT.ZS (Taux d'alphabétisation des jeunes, population 15-24 ans, deux sexes (%))

Analysez les corrélations entre les différentes variables

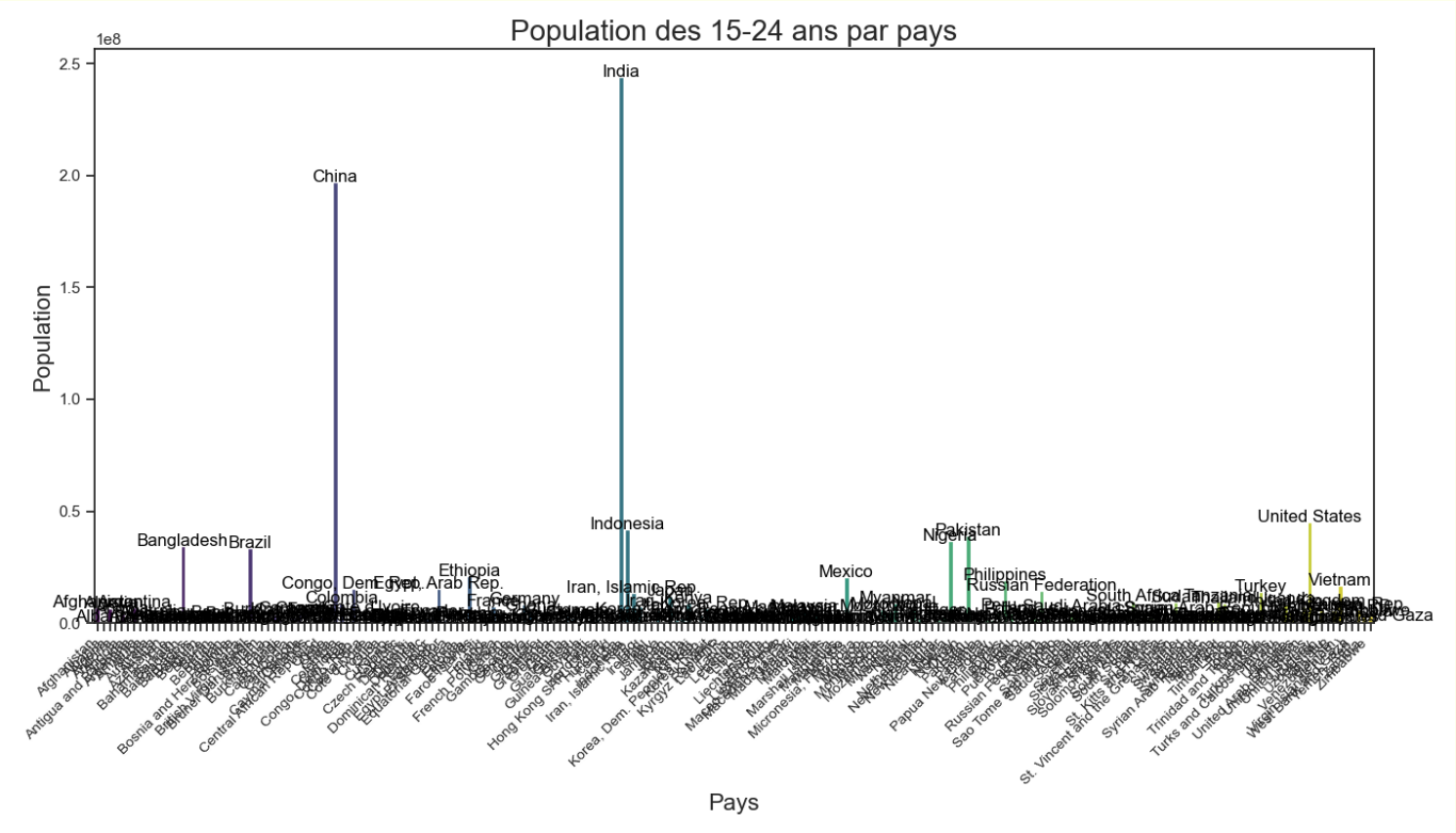
Indicator Code	IT.NET.USER.P2	SE.ADT.1524.LT.ZS	SE.SEC.ENRR.UP	SE.TER.ENRR	SE.XPD.TOTL.GD.ZS	SP.POP.1524.TO.UN
Indicator Code						
IT.NET.USER.P2	1.000000	0.703504	0.793854	0.736574	0.103787	-0.070287
SE.ADT.1524.LT.ZS	0.703504	1.000000	0.752351	0.584784	0.235403	0.015684
SE.SEC.ENRR.UP	0.793854	0.752351	1.000000	0.782352	0.254766	-0.040457
SE.TER.ENRR	0.736574	0.584784	0.782352	1.000000	0.220738	-0.037262
SE.XPD.TOTL.GD.ZS	0.103787	0.235403	0.254766	0.220738	1.000000	-0.097339
SP.POP.1524.TO.UN	-0.070287	0.015684	-0.040457	-0.037262	-0.097339	1.000000

On voit une **bonne corrélation** entre les indicateurs :
- "IT.NET.USER.P2" et "SE.SEC.ENRR.UP" avec une valeur de 0.79,
- "IT.NET.USER.P2" et "SE.TER.ENRR" avec une valeur de 0.73,
- "IT.NET.USER.P2" et "SE.ADT.1524.LT.ZS" avec une valeur de 0.70,

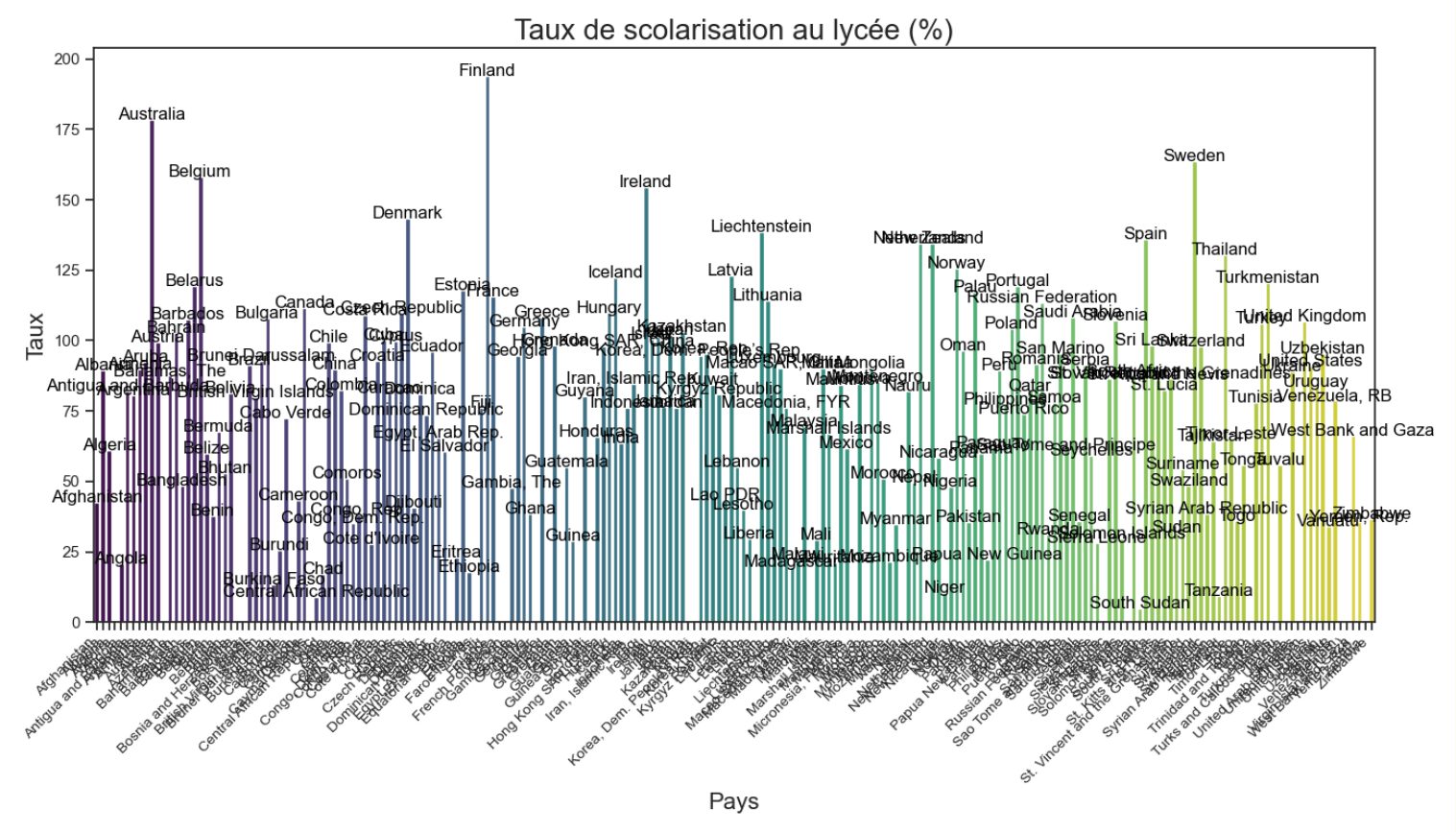
Et **pas de corrélation** entre les indicateurs :
- "IT.NET.USER.P2" et "SP.POP.1524.TO.UN" avec une valeur de -0.07,
- "IT.NET.USER.P2" et "SE.XPD.TOTL.GD.ZS" avec une valeur de 0.10,



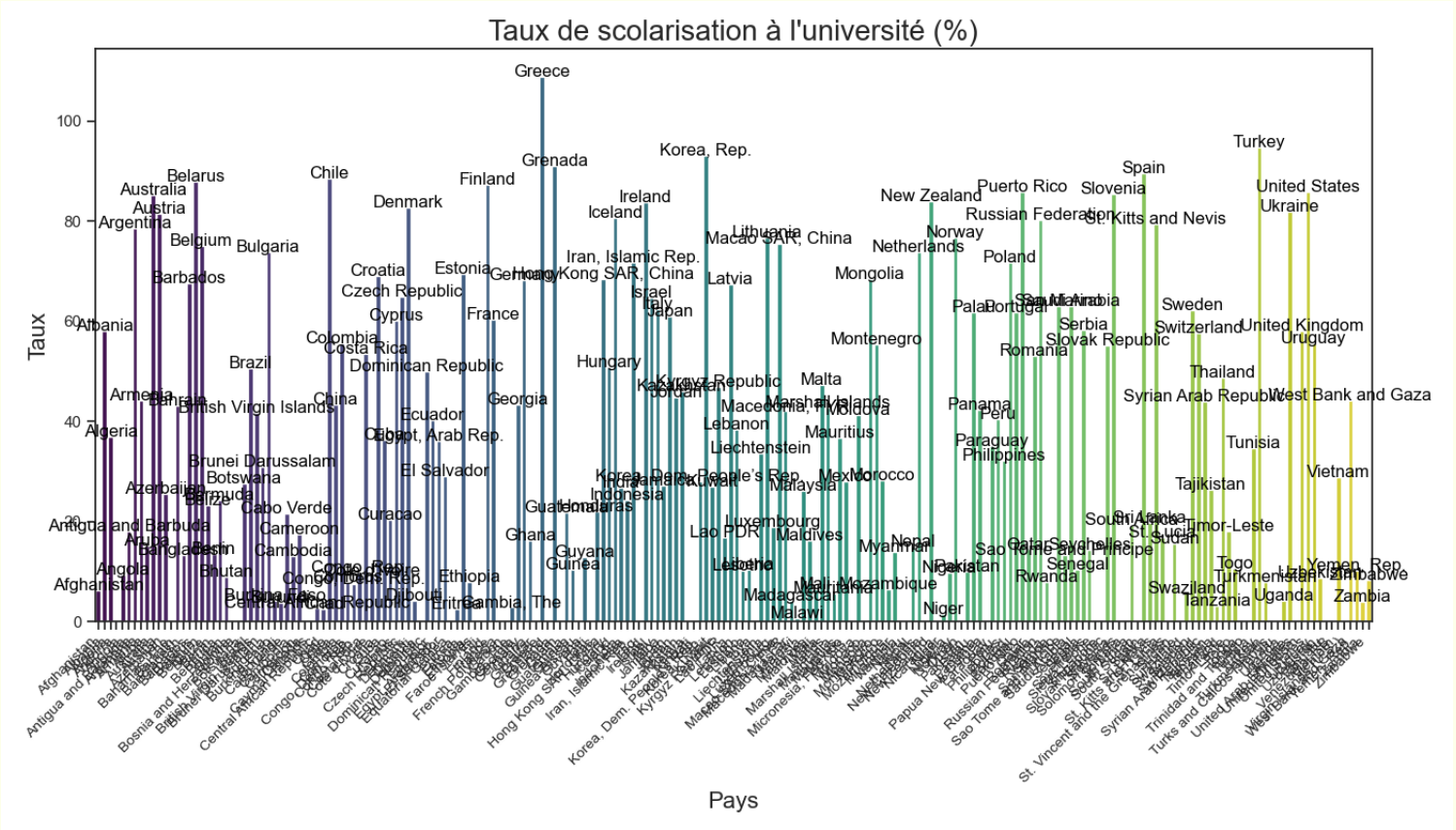
Analyse graphique :



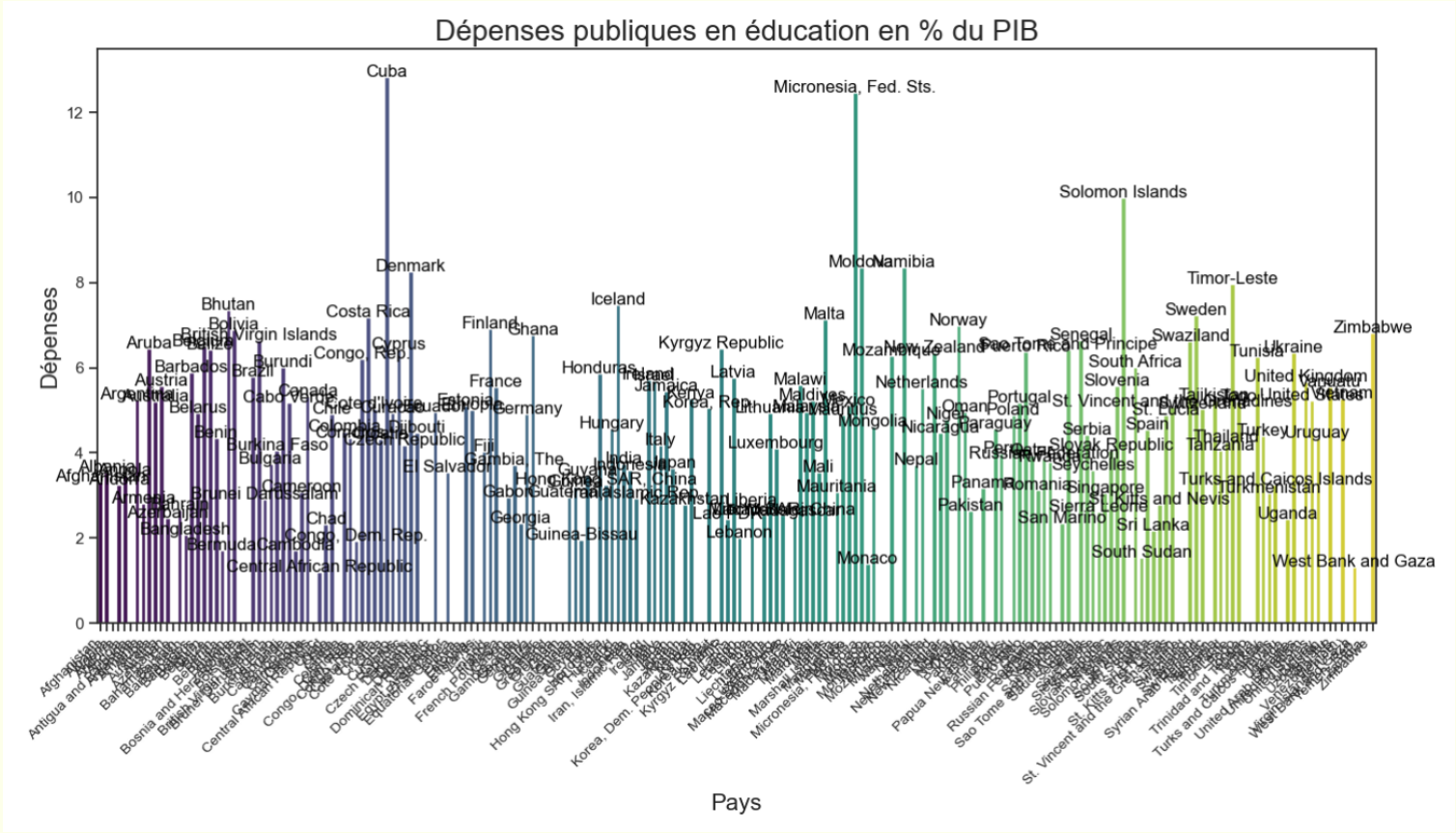
Analyse graphique :



Analyse graphique :



Analyse graphique :



Quels sont les pays avec un fort potentiel de clients pour 'Academy'?

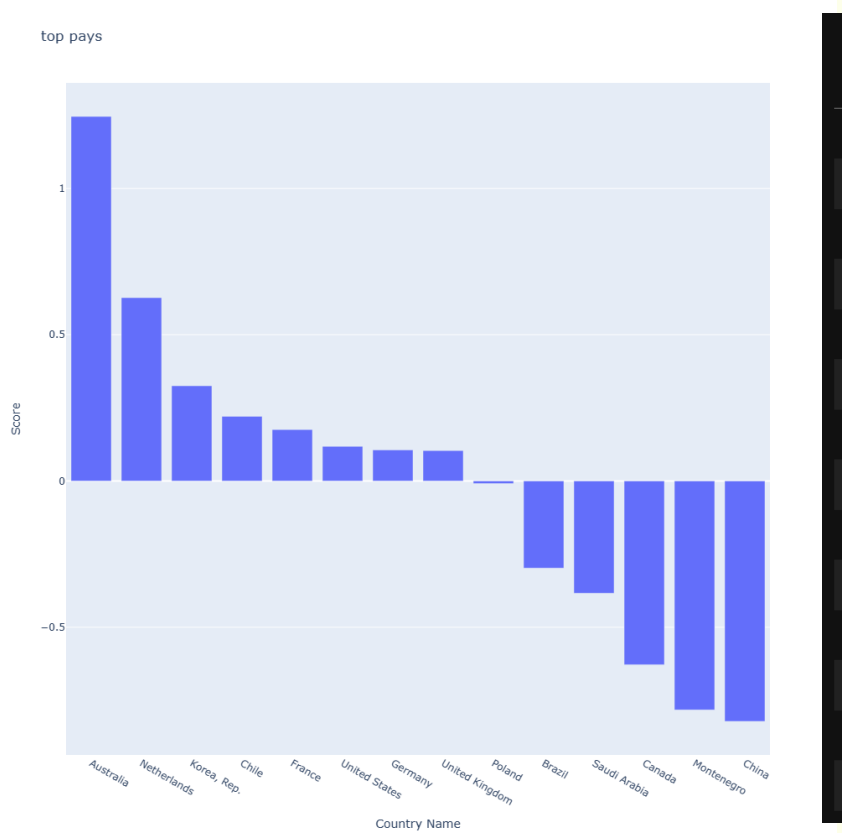
Quels sont les pays avec un fort potentiel de clients pour 'Academy'?

Résultats de filtrage par la valeur **médiane** de chaque indicateur et la suppression des pays ayant **plus de 2 valeurs manquantes** : **14 pays**

Indicator Code	Median
IT.NET.USER.P2	5.021966e+01
SE.ADT.1524.LT.ZS	9.828706e+01
SE.SEC.ENRR.UP	8.121219e+01
SE.TER.ENRR	3.622854e+01
SE.XPD.TOTL.GD.ZS	4.621675e+00
SP.POP.1524.TO.UN	1.331040e+06

Country Name	IT.NET.USER.P2	SE.ADT.1524.LT.ZS	SE.SEC.ENRR.UP	SE.TER.ENRR	SE.XPD.TOTL.GD.ZS	SP.POP.1524.TO.UN
Australia	84.560519	NaN	178.556668	85.332562	5.212932	2914620.0
Brazil	58.327952	98.440614	91.420326	50.604919	5.794417	33595574.0
Canada	88.470000	NaN	111.744696	NaN	5.321070	4373511.0
Chile	64.289000	99.228235	99.618759	88.577293	4.923240	2817084.0
China	50.300000	99.642290	89.651901	43.391769	NaN	197026759.0
France	84.694500	NaN	115.591806	60.335406	5.552842	7567872.0
Germany	87.589800	NaN	104.849243	68.265587	4.907592	8682394.0
Korea, Rep.	89.648631	NaN	95.281891	93.179138	5.052110	6456561.0
Montenegro	68.119581	99.211820	85.574402	55.344589	NaN	NaN
Netherlands	91.724138	NaN	134.458405	73.768501	5.534642	2005912.0
Poland	67.997000	NaN	104.092271	71.695662	4.903020	4404280.0
Saudi Arabia	69.616236	99.221770	108.314159	63.066219	NaN	5298036.0
United Kingdom	92.000300	NaN	106.872313	58.124581	5.684270	7731522.0
United States	74.554202	NaN	90.821086	85.795776	5.230070	45147517.0

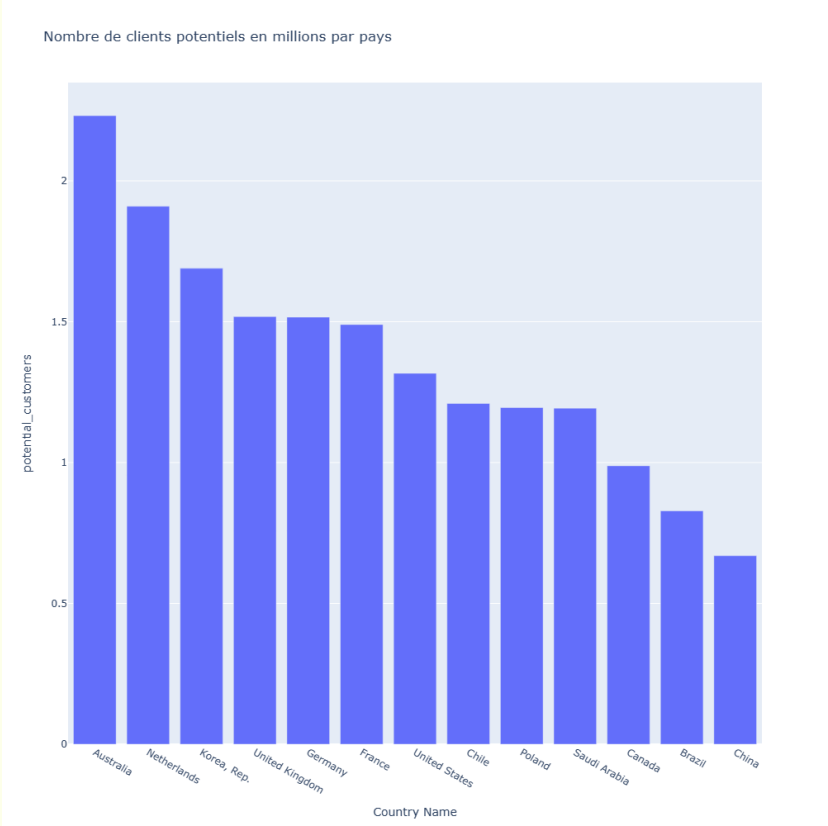
Quels sont les pays avec un fort potentiel de clients pour 'Academy'?



Country Name	Indicator Code Score
Australia	1.245777
Netherlands	0.626500
Korea, Rep.	0.325284
Chile	0.221025
France	0.175706
United States	0.118459
Germany	0.106072
United Kingdom	0.103845
Poland	-0.008919
Brazil	-0.297974
Saudi Arabia	-0.383616
Canada	-0.627963
Montenegro	-0.782594
China	-0.821602

Classification des pays en utilisant un **score** et poids d'indicateurs

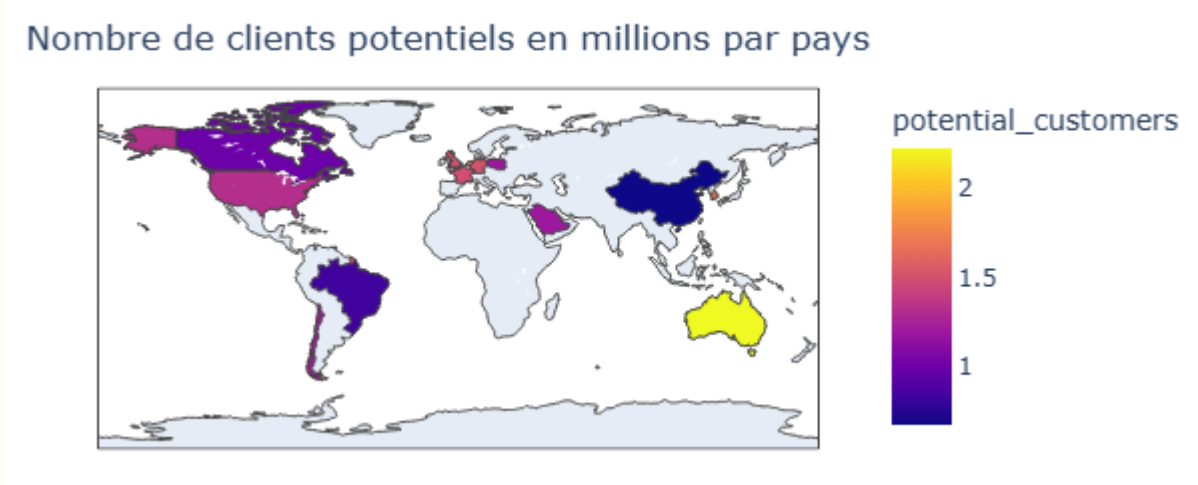
Quels sont les pays avec un fort potentiel de clients pour 'Academy'? (Taux de pénétration d'indicateurs)



Indicator Code	Country Name	potential_customers
0	Australia	2.231461
1	Netherlands	1.909943
2	Korea, Rep.	1.689527
3	United Kingdom	1.517976
4	Germany	1.516309
5	France	1.490007
6	United States	1.316753
7	Chile	1.209894
8	Poland	1.195305
9	Saudi Arabia	1.193086
10	Canada	0.988605
11	Brazil	0.828404
12	China	0.669210

Les pays avec un fort potentiel de clients pour 'Academy'?

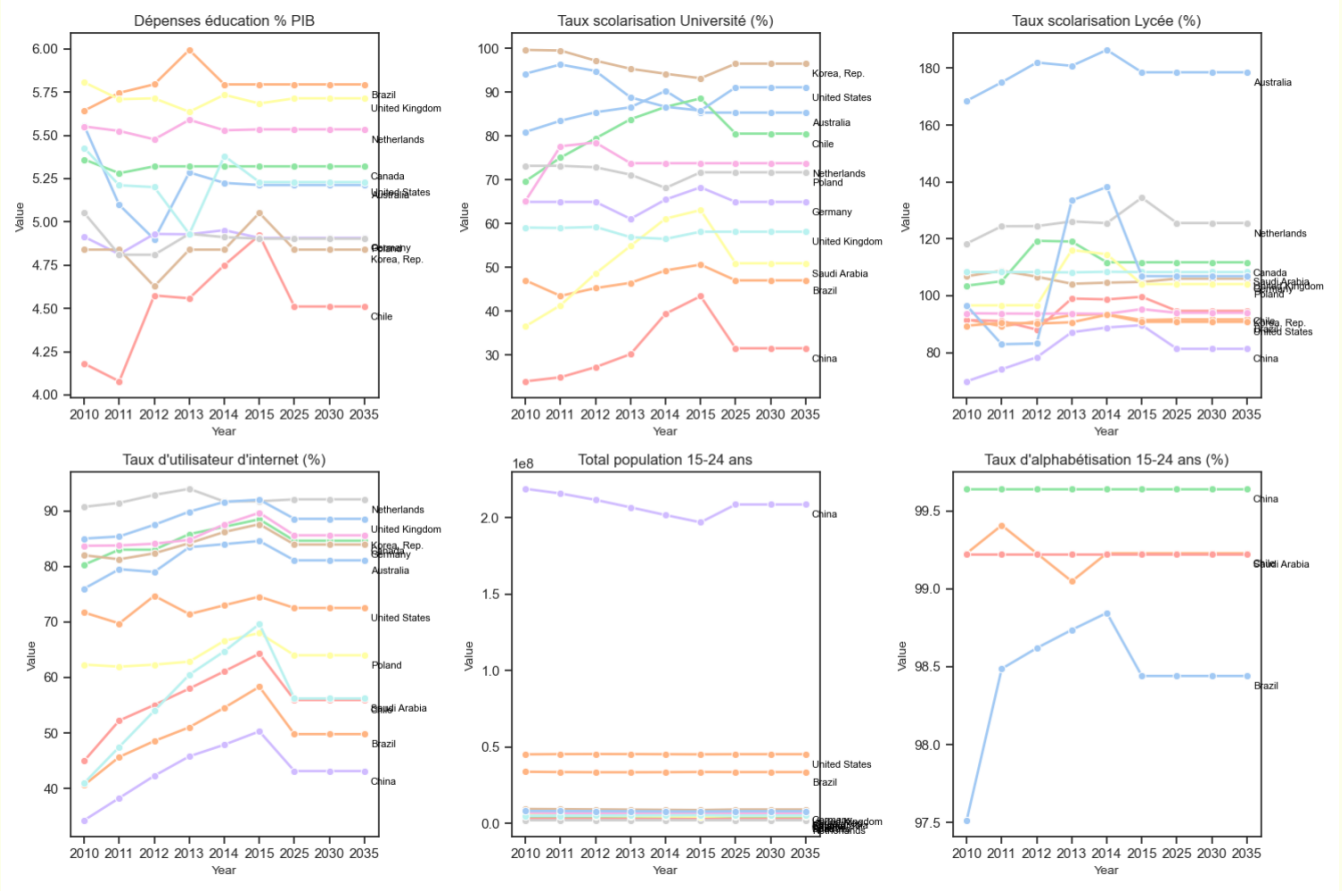
Priorisation des pays :



Country	Region
1. Australia	East Asia and Pacific
2. Netherlands	Europe and Central Asia
3. Korea, Rep	East Asia and Pacific
4. United Kingdom	Europe and Central Asia
5. Germany	Europe and Central Asia
6. United States	North America
7. Chile	Latin America and Caribbean
8. Poland	Europe and Central Asia
9. Saudi Arabia	Middle East & North Africa
10.Canada	North America
11. Brazil	Latin America and Caribbean
12. China	East Asia and Pacific

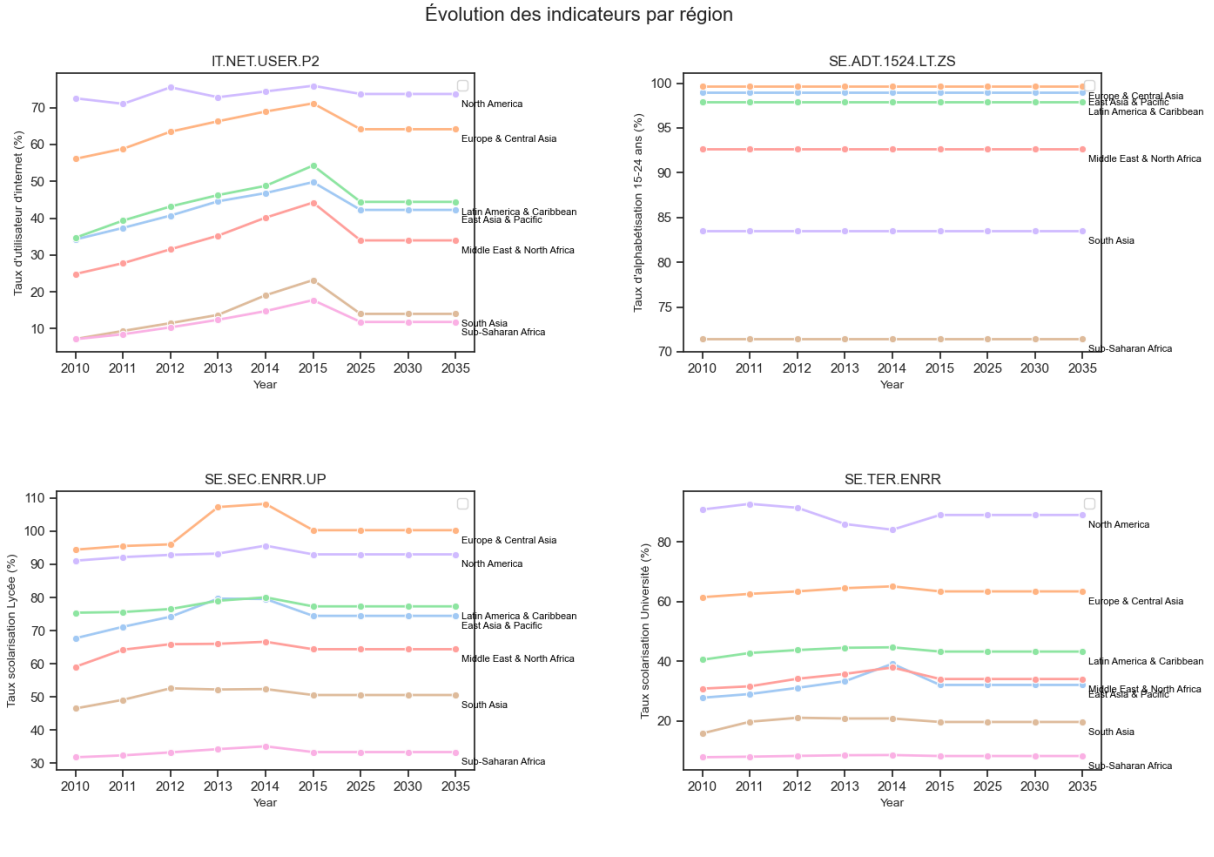
Analyse de l’évolution des indicateurs par pays

Quelle sera l’évolution de nos indicateurs par pays ?



Analyse de l’évolution des indicateurs par régions

Quelle sera l'évolution de nos indicateurs par région ?



COUNTRY	REGION
1. Australia	East Asia and Pacific
2. Netherlands	Europe and Central Asia
3. Korea, Rep	East Asia and Pacific
4. United Kingdom	Europe and Central Asia
5. Germany	Europe and Central Asia
6. United States	North America
7. Chile	Latin America and Caribbean
8. Poland	Europe and Central Asia
9. Saudi Arabia	Middle East & North Africa
10.Canada	North America
11. Brazil	Latin America and Caribbean
12. China	East Asia and Pacific

Conclusion :

Le jeu de données a une bonne capacité informatrice qui permet de répondre aux questions stratégiques de l'entreprise en matière d'expansion internationale.

Forces : jeu de données avec une large couverture géographique et fiable.
Limites : les lacunes dans les indicateurs disponibles pour certains pays.

A étudier :

- l'éventuelle concurrence surtout dans les pays développés
- la stratégie d'entreprise : langue, proximité géographique d'implantation
- la politique de l'entreprise envers les petits pays, qui ont une faible population, mais une bonne couverture d'internet et un bon investissement dans l'éducation