Analyse des données de systèmes éducatifs

academy

OC Data Scientist
Projet 2

Clément Dubois
06/09/2022

Contexte générale

Contexte :

L'entreprise souhaite s'expandre à l'international.

- Quels sont les pays à fort potentiel pour l'installation de l'entreprise ?

Données :

Jeu de données sur l'éducation mis à disposition par The World Bank

Mission :

Analyse exploratoire afin de répondre à deux questions :

- Le jeu de données est il pertinent et suffisant pour déterminer les pays cibles ?
- Si oui, dans quels pays l'entreprise doit-elle opérer en priorité ?

Sommaire

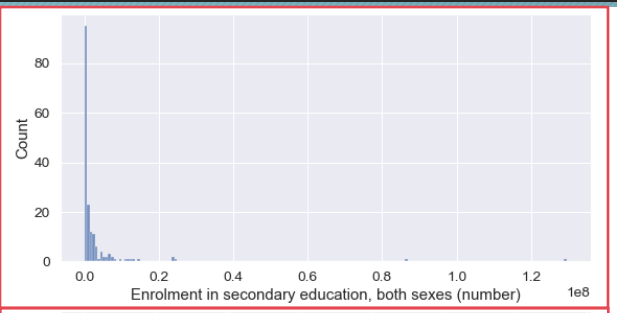
Présentation des données et des indicateurs sélectionnés

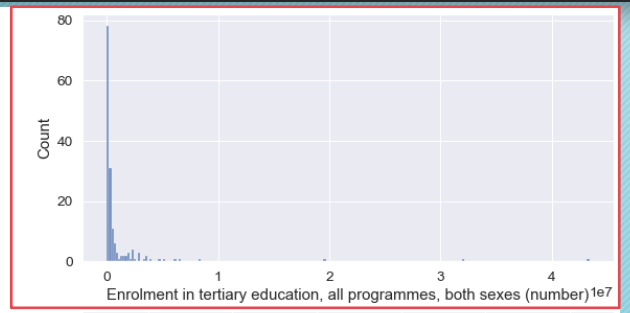
- **3665 Indicateurs**
- **217 pays**
- **1970 > 2016 (Projections de Wittgenstein misent à part)**

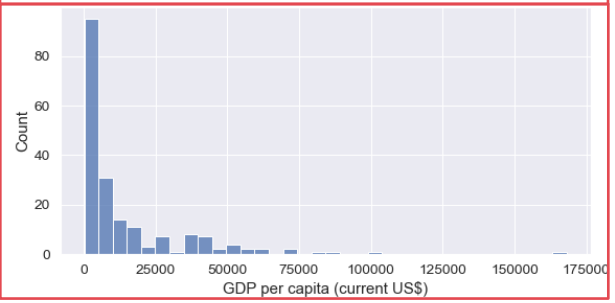
Educatifs			Economiques		Démographiques		Spécifique
Enrolment in secondary education, both sexes (number)	Enrolment in tertiary education, all programmes, both sexes (number)	Government expenditure on education as % of GDP (%)	GDP per capita (current US$)	Unemployment, total (% of total labor force)	Population growth (annual %)	Population, total	Internet users (per 100 people)
Nombre d'inscrits en Lycée	Nombre d'inscrits en études supérieures	Dépenses du gouvernement dans l'éducation (% du PIB)	PIB par habitant (en $ US)	Pourcentage de personnes sans emploi	Taux de croissance	Population totale	Nombre d'utilisateurs d'internet (%)

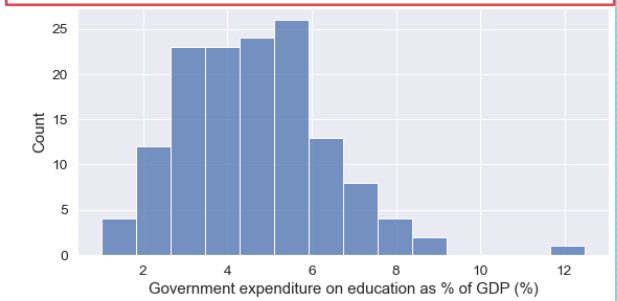
> Prise en compte de la dernière valeur connue entre 2012 et 2016.

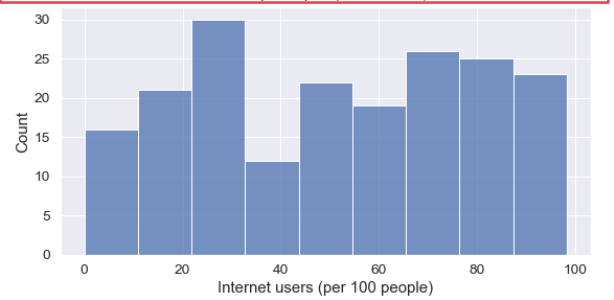
Analyse univariée : Histogramme et transformation des Indicateurs

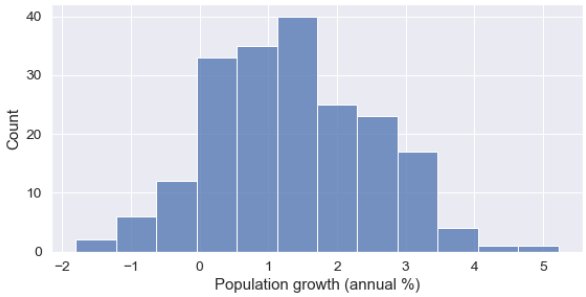


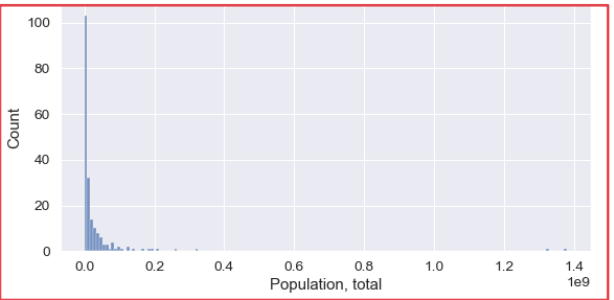


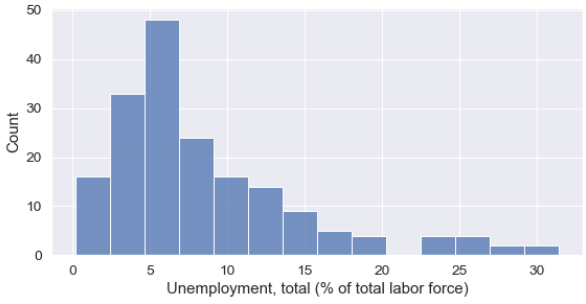








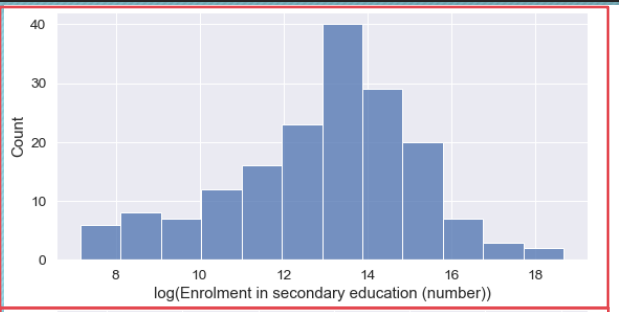


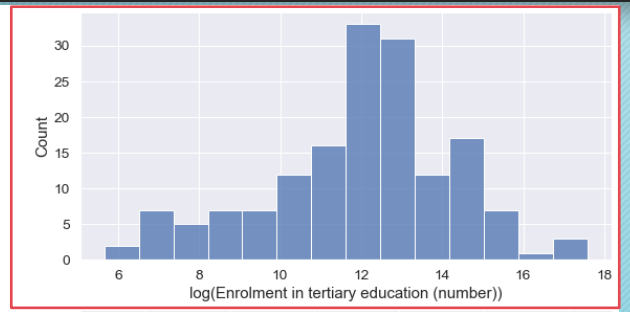


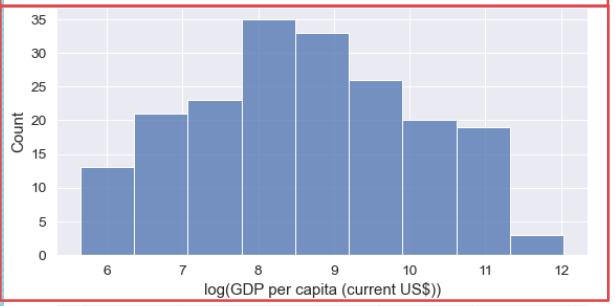
‘Enrolment in sec. Education’
‘Enrolment in tert. Education’
‘GDP per capita’
‘Population’

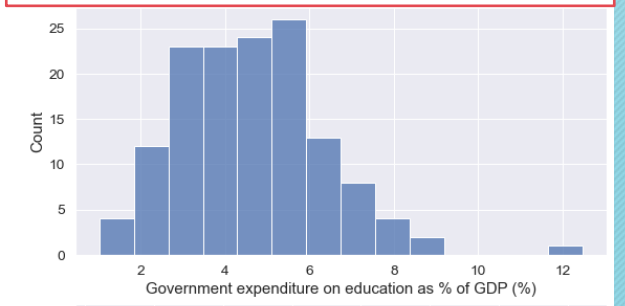
} **Distributions ne suivant pas une Loi Normal**

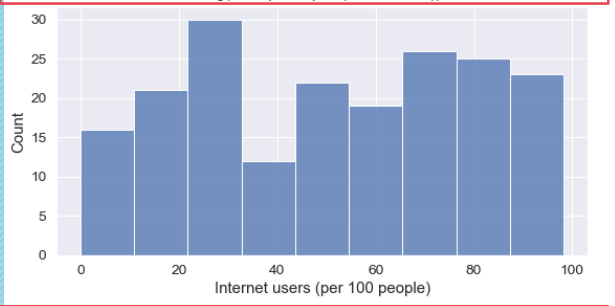
Analyse univariée : Histogramme et transformation des Indicateurs

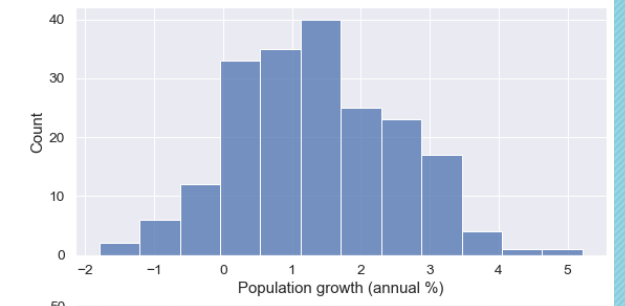


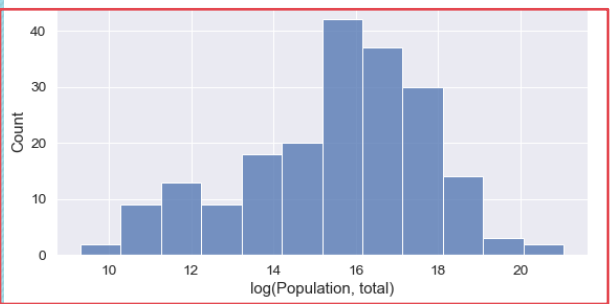


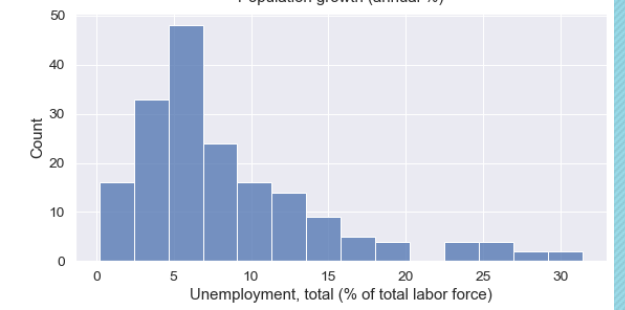










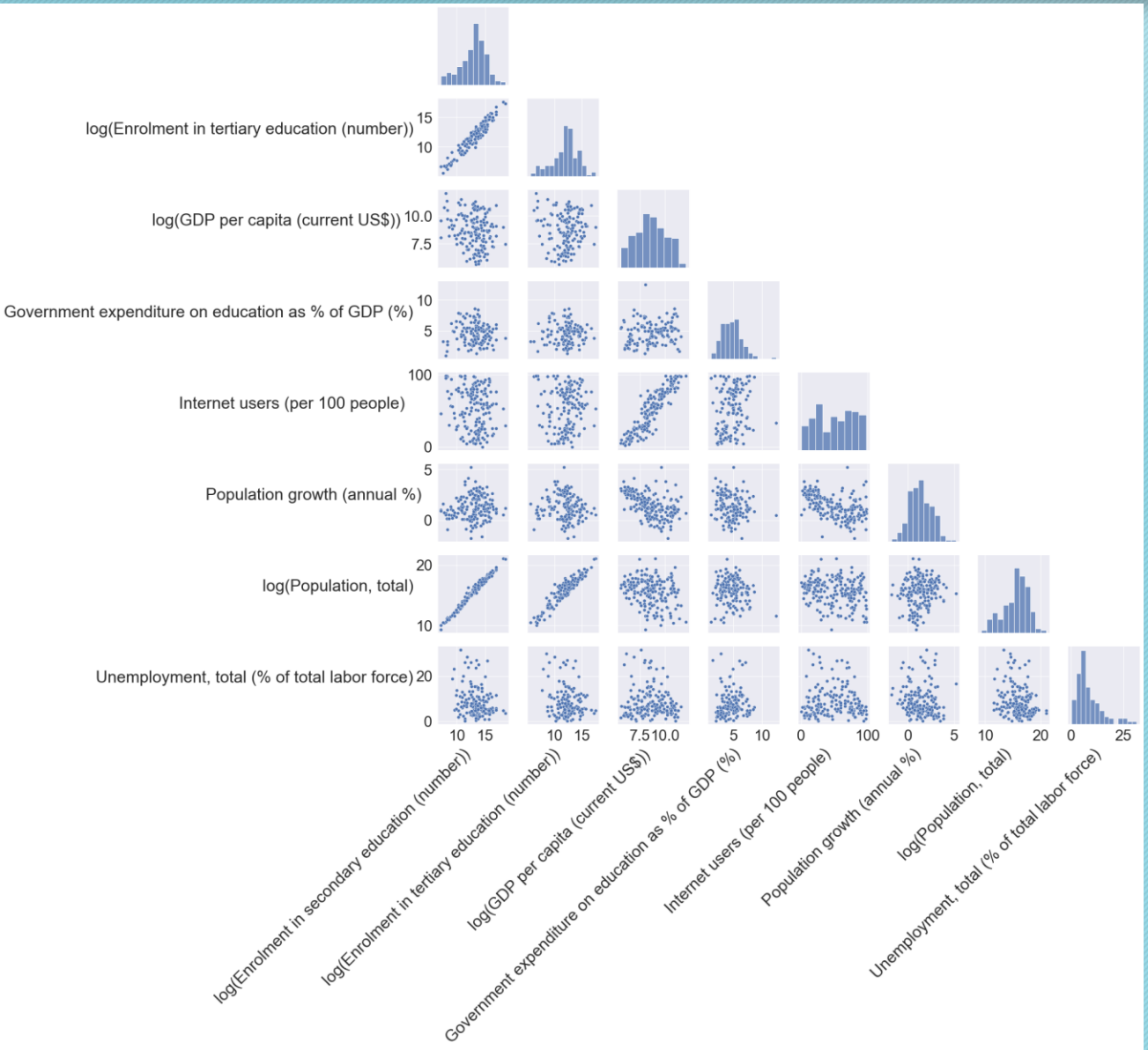


‘Enrolment in sec. Education’
‘Enrolment in tert. Education’
‘GDP per capita’
‘Population’

} **Distributions ne suivant pas une Loi Normal**

- **Etape de transformation des données (log naturel)**

Analyse bivariée : Aperçu des relations entre les indicateurs

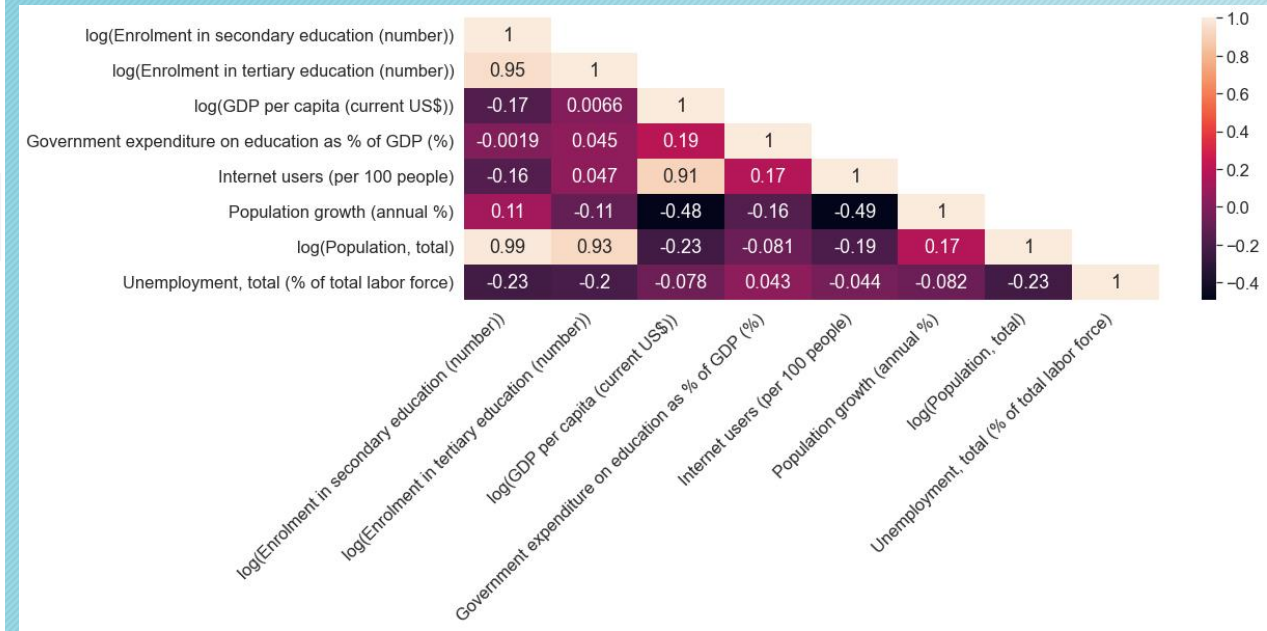


Certains indicateurs semblent très fortement corrélés

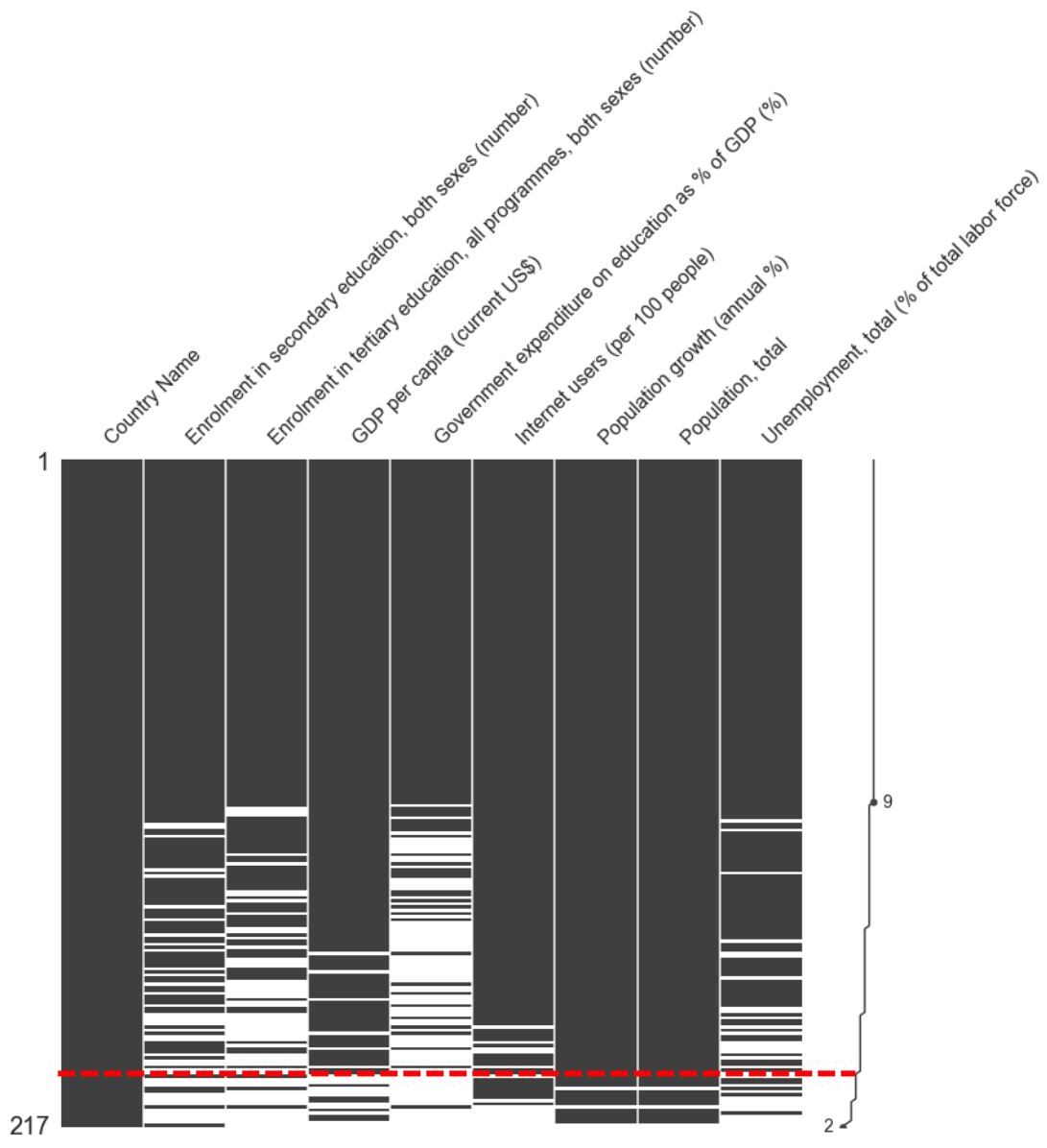
+ : Population, Nombre d'étudiants au Lycée et Nombre d'étudiants en Etude Sup.

+ : PIB par habitant et le % d'utilisateurs d'internet.

- : Taux de croissance avec le PIB et le % d'utilisateurs d'internet.



Gestion des valeurs manquantes



Nom de l'indicateur	Valeurs manquantes
Country Name	0
Enrolment in secondary education, both sexes (number)	39
Enrolment in tertiary education, all programmes, both sexes (number)	54
GDP per capita (current US$)	17
Government expenditure on education as % of GDP (%)	76
Internet users (per 100 people)	14
Population growth (annual %)	3
Population, total	3
Unemployment, total (% of total labor force)	31

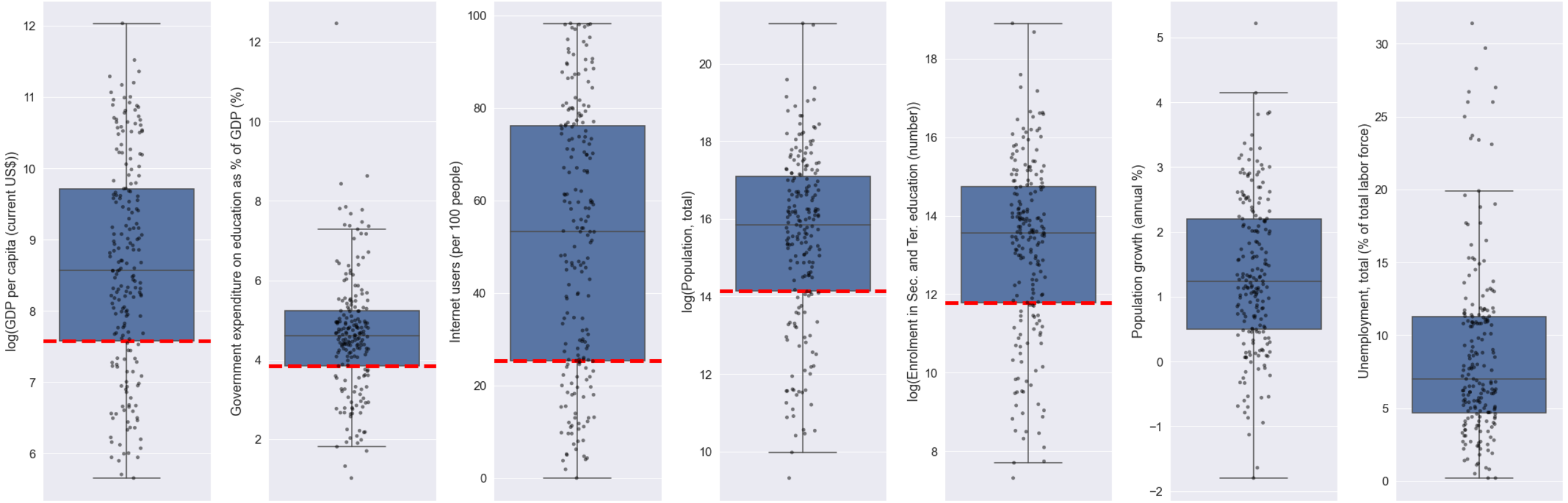
- 1er filtre : Plus de 3 valeurs manquantes (18 pays filtrés)

American Samoa, British Virgin Islands, Cayman Islands, Channel Islands, Eritrea, Faroe Islands, French Polynesia, Gibraltar, Greenland, Isle of Man, Kiribati, Kosovo, Libya, Nauru, New Caledonia, Northern Mariana Islands, Sint Maarten (Dutch part), St. Martin (French part)

- Prédictions des valeurs manquantes par un algorithme de régression linéaire itératif (IterativeImputer de scikit learn)

- Regroupement des indicateurs de recrutement des étudiants en secondaire et en tertiaire

Filtre des pays les moins susceptibles d'accueillir l'entreprise

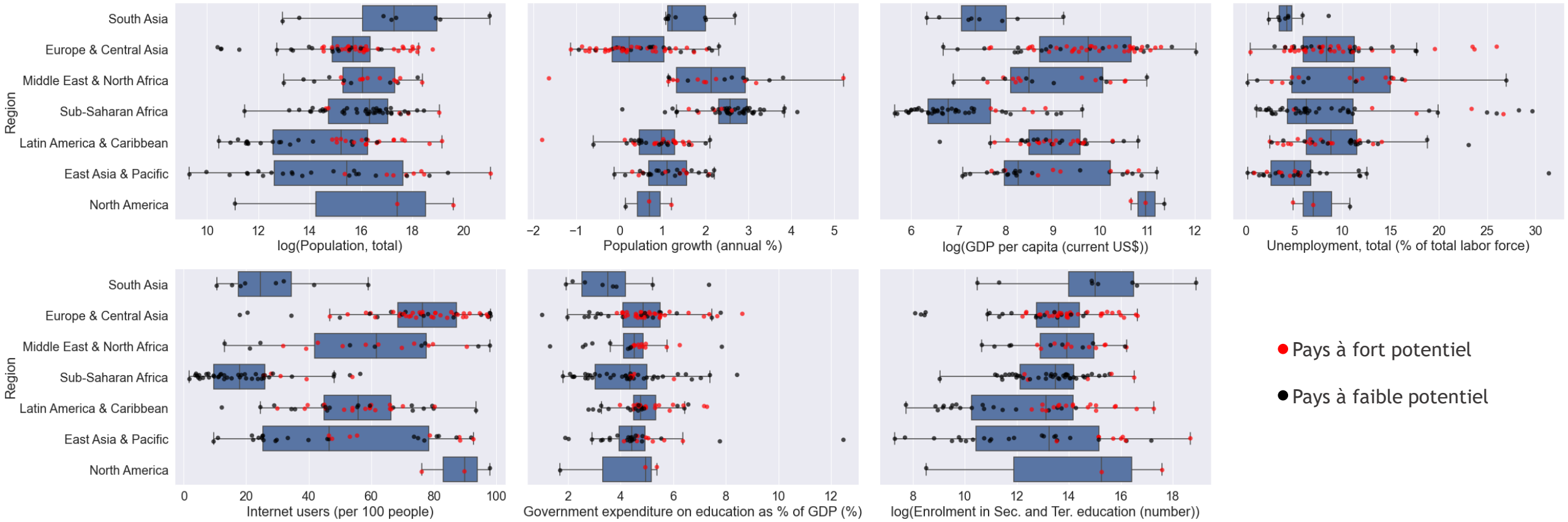


2ème filtre : Pays dont un indicateur principale au moins est inférieur au premier quartile

Après ces étapes : Liste de 77 pays à fort potentiel

- Dans quelles régions se situent ces différents pays ?

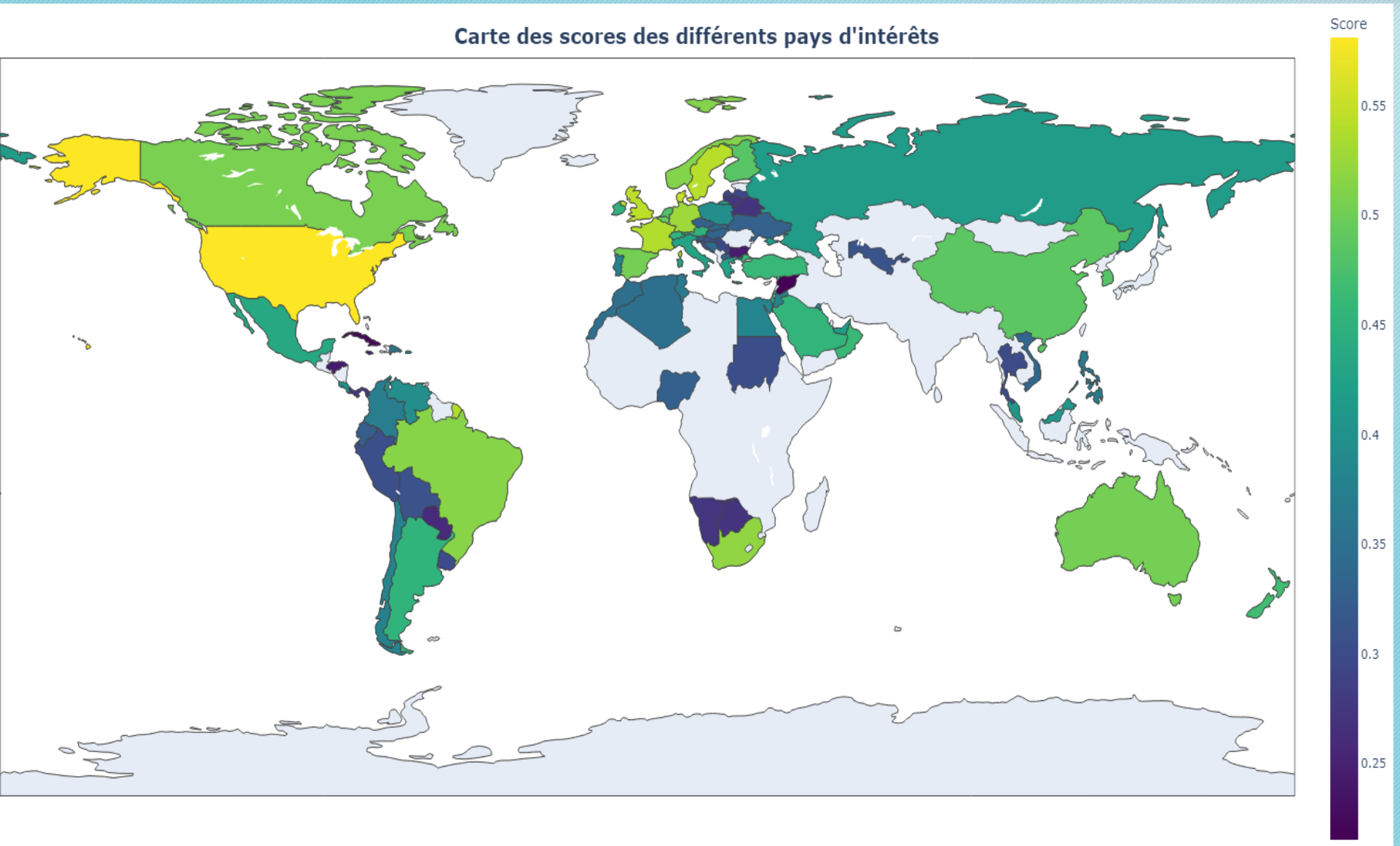
Comparaison des indicateurs entre les différentes régions



L' « Europe & l'Asie Centrale » semble être la région contenant le plus de pays à fort potentiel.

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

Création d'un score afin de déterminer les pays à fort potentiel



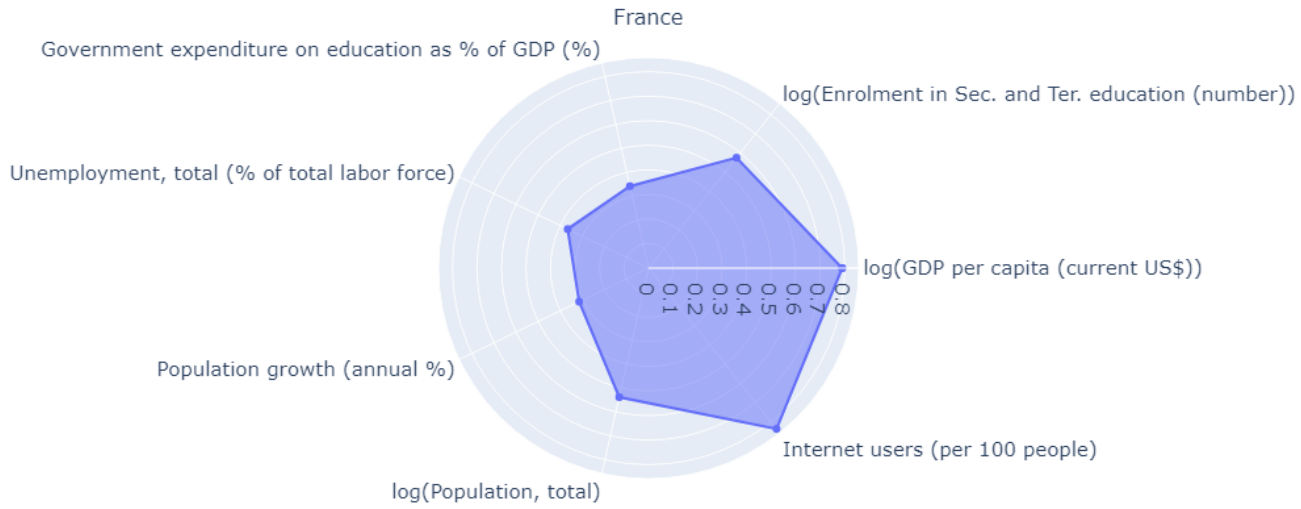
Moyenne arithmétique sur données normalisées (*min-max scaler*)

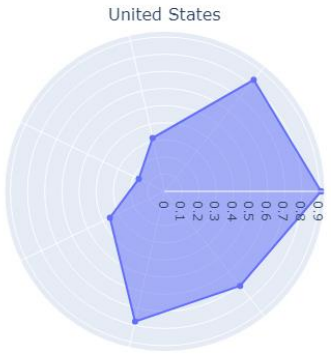
$$Score = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i - x_{\min i}}{x_{\max i} - x_{\min i}}$$

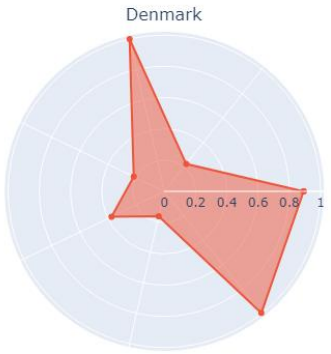
Liste des 6 pays à fort potentiel (cibles) :

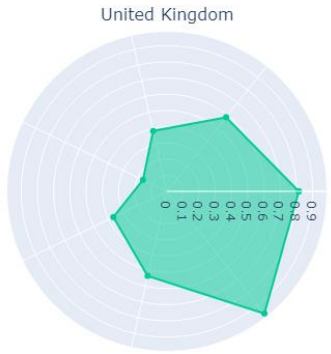
Country Name	Score
United States	0.5812
Denmark	0.5525
United Kingdom	0.5448
Sweden	0.5411
France	0.5376
Germany	0.5267

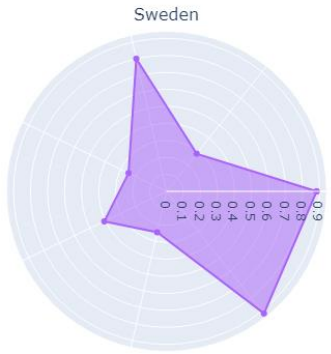
Comparaison des indicateurs des pays cibles

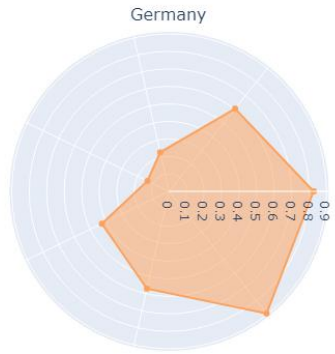






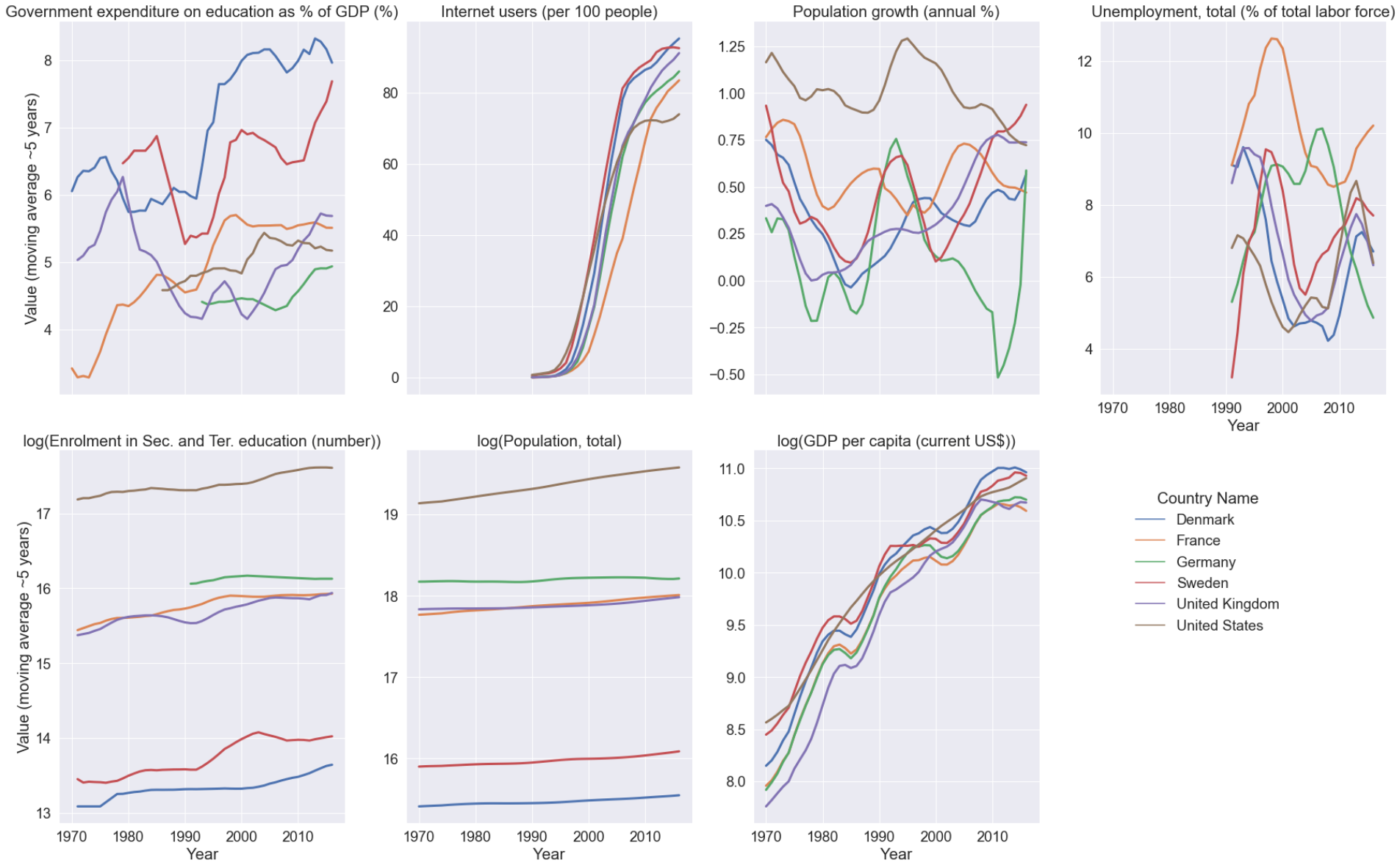






- La France, le Royaume-Uni et l'Allemagne partagent globalement les mêmes motifs (indicateurs similaires).
- Le Danemark et la Suède se démarquent par un fort investissement de l'état dans l'éducation.
- Les Etats-Unis se démarquent par leur forte population (totale et étudiante)

Evolution des indicateurs des pays cibles dans le temps



- **Dépenses publiques dans l'éducation :** ↗ Danemark et Suède
- **Utilisation d'internet :** ↗ expo
- **Taux de croissance :** stable en moyenne (voir légère diminution)
- **Taux de chômage :** ↘ récente sauf pour la France
- **Nombre d'étudiants (Lycée + Supérieur) :** ↗ globale
- **Population :** ↗ globale
- **PIB par habitant :** ↗ puis stabilisation globale

Conclusion

Le jeu de données :

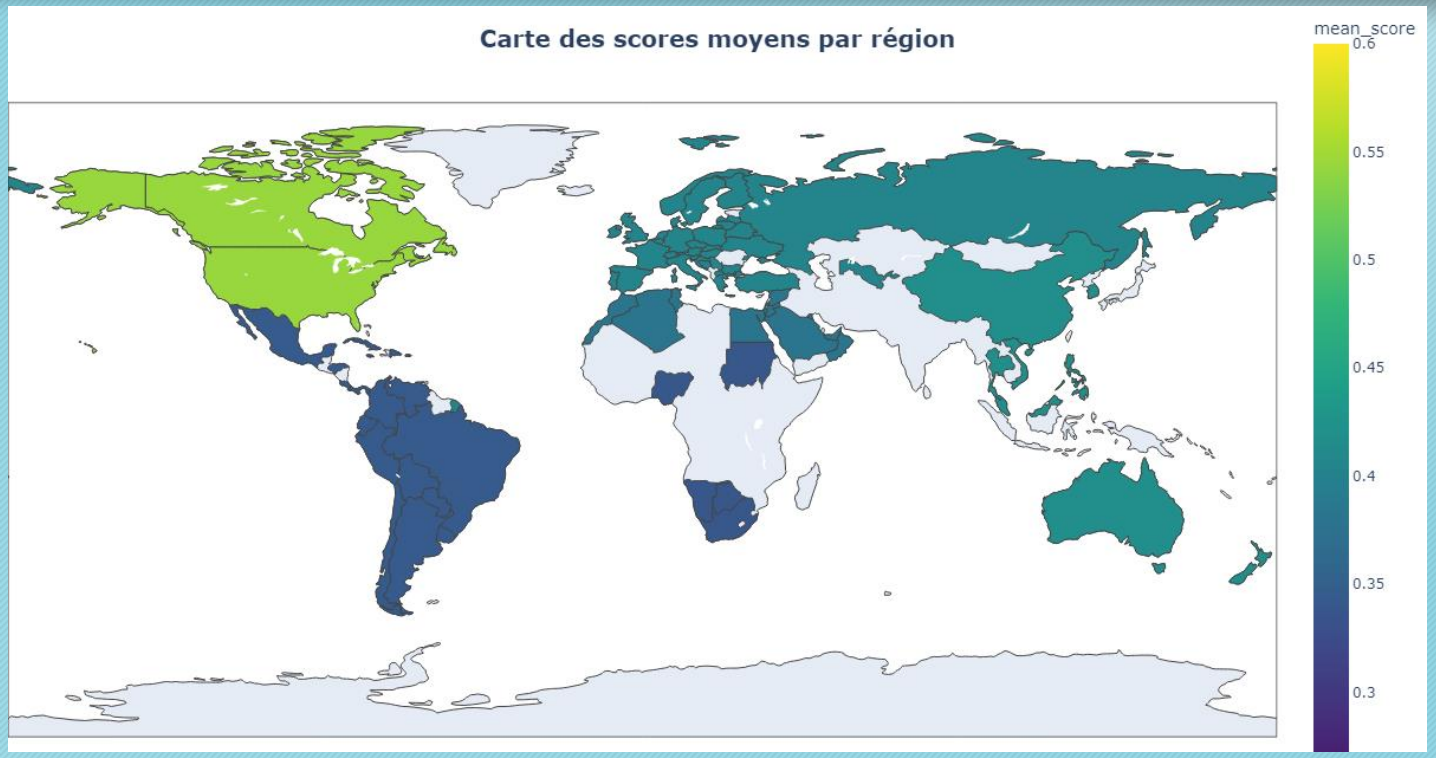
- Présente des indicateurs intéressants pour la problématique de l'entreprise.
- Apporte des infos disponibles pour un grand nombre de pays.
- Suffisant pour créer un score et déterminer une liste de pays à fort potentiel.

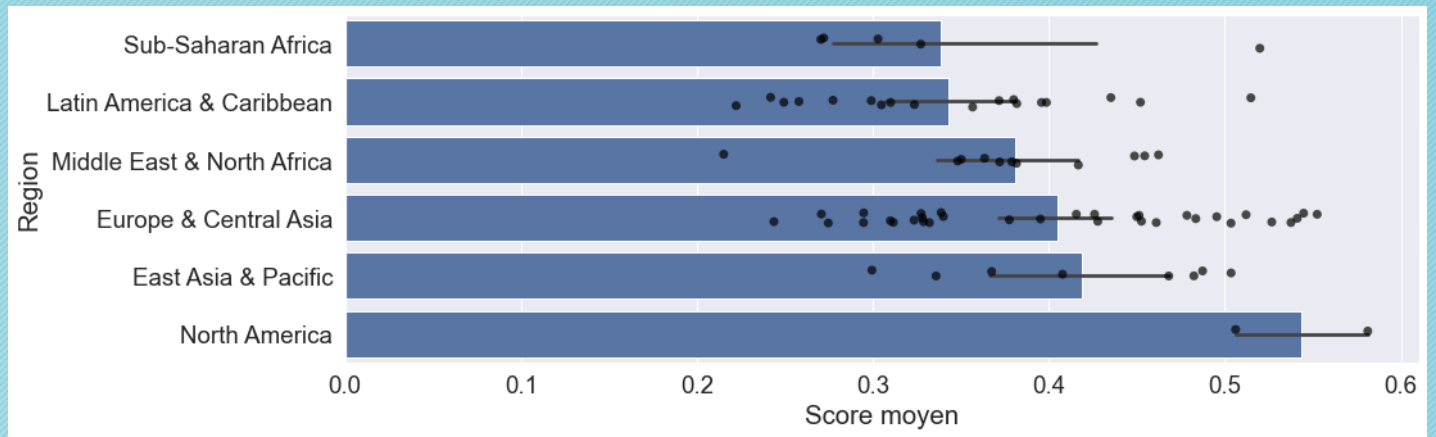
Etats Unis, Danemark, Royaume Uni

Axes d'amélioration :

- Données relativement anciennes (Données plus récentes sur The WorldBank)
- Pourraient être complétées par d'autres indicateurs (Stabilité politique, facilité à se développer dans le pays, langue des habitants…)
- La multiplication des indicateurs peut s'accompagner de nouvelles analyses (Analyse en composantes principales et clustering)

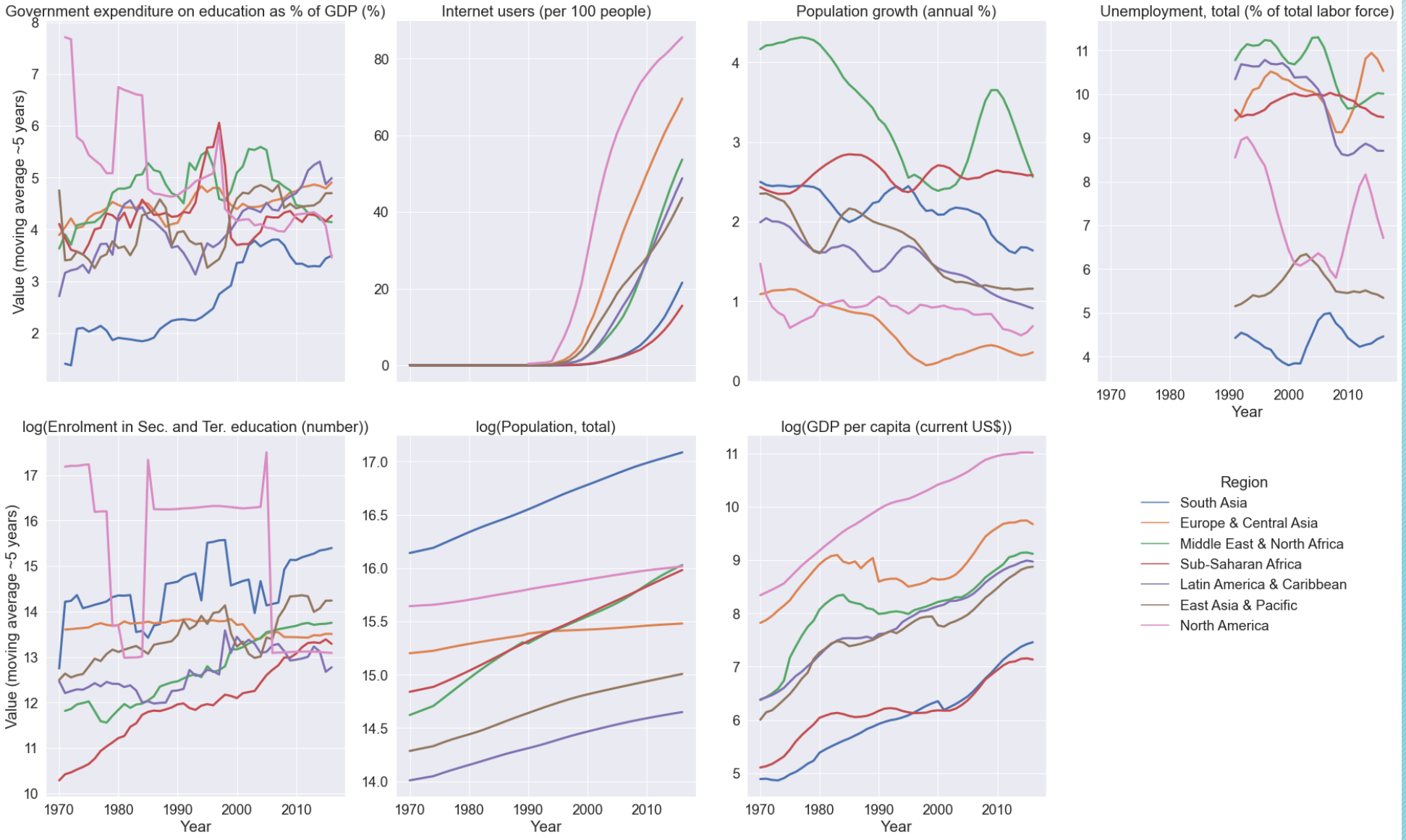
Annexe 1 : Présentation des scores par régions géographiques





- **L'Amérique du Nord dispose du score moyen le plus élevé (mais seulement deux pays)**
- **On observe une grande disparité à l'échelle régionale.**

Annexe 2 : Evolution des indicateurs des régions (moyenne par pays) dans le temps



- **Dépenses publiques dans l'éducation :** convergence vers 4%
- **Utilisation d'internet :** ↗ expo. Retard pour Afrique/Asie
- **Taux de croissance :** stable en moyenne (voir légère diminution)
- **Taux de chômage :** Stable
- **Nombre d'étudiants (Lycée + Supérieur) :** ↗ globale
- **Population :** ↗ globale
- **PIB par habitant :** ↗ puis stabilisation globale