ACADEMY

Analysez des données de systèmes éducatifs
Projet d'expansion à l'international

SOMMAIRE

CONTEXTE ET PROBLÉMATIQUE

📚 Academy

- Academy est une start-up de la EdTech qui propose des formations en ligne pour les lycéens et les enseignements supérieurs.
- Projet d'expansion à l'international
- Identification du potentiel de nouveaux marchés : dans le cadre de son expansion, elle souhaite analyser les données de systèmes éducatifs de différents pays pour mieux cibler leurs futures implantations.

• Pour lever les interrogations …

- A partir des données de la Banque mondiale, réaliser une pré-analyse exploratoire permettant de répondre aux interrogations suivantes :

➢ **Quels sont les pays avec un fort potentiel de clients pour nos services ?**

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

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

THE WORLD BANK

MISSION ET OBJECTIFS

Mission:

Notre mission consiste à réaliser une analyse préliminaire des données éducatives de la Banque mondiale pour évaluer leur potentiel à éclairer la stratégie d'expansion internationale d'Academy, une start-up EdTech.

Les objectifs spécifiques sont :

- **Validation de la qualité des données : Vérifier la présence de données manquantes ou dupliquées.**
- **Description du jeu de données : Dénombrer les colonnes et les lignes et comprendre la nature des informations contenues.**
- **Sélection des données pertinentes : Identifier les colonnes utiles pour la problématique de l'entreprise.**
- **Analyse statistique : Évaluer les tendances et les indicateurs clés par pays et par région, en utilisant des mesures statistiques comme la moyenne, la médiane, et l'écart-type.**

Ces étapes guideront l'entreprise dans sa décision d'entrer dans de nouveaux marchés en se basant sur les données éducatives mondiales.

DESCRIPTION DES JEUX DE DONNÉES

Source du jeu de données:

- Les données utilisées pour cette analyse sont extraites de la base "EdStats All Indicator Query" de la Banque mondiale, qui comprend un ensemble de 4000 indicateurs internationaux sur l'accès à l'éducation, les niveaux de diplômes atteints, et des données sur les enseignants et les dépenses éducatives.

- 5 fichiers de données au format CSV Essentiel des données : EdStatsData.csv

Indicateurs	Régions	Thèmes	Années	Pays
3665 indicateurs internationaux	7 régions	37 sujets liés à l'éducation	1970-2100	242 pays et zones

DESCRIPTION DES JEUX DE DONNÉES

EdStatsData.csv	EdStatsSeries	EdStatsCountry	EdStatsFootNote	EdStatsCountry-Series
Valeurs des indicateurs pour tous les pays sur plusieurs années avec projection ➢ 886 930 lignes ➢ 70 variables Colonnes importantes : ➢ Pays ➢ Région ➢ Indicator Code ➢ Années de 1970 à 2100	Informations socio-éduco-économiques des indicateurs par thème ➢ 3 665 lignes ➢ 21 variables Colonnes importantes : ➢ Série Code ➢ Topic ➢ Indicator Name ➢ Long Définition	Informations globales sur l'économie de chaque pays/zone du monde ➢ 241 lignes ➢ 32 variables Colonnes importantes : ➢ Country Code ➢ Région	Informations sur l'année d'origine/incertitude des indicateurs par pays ➢ 643 638 lignes ➢ 5 variables	Informations sur la source des indicateurs par pays ➢ 613 lignes ➢ 4 variables

EXPLORATION PRÉLIMINAIRE DES DONNÉES

1. STRATEGIE

1. Pré-sélection des données

Sélection préliminaire de données qui pourraient s'avérer cruciales pour l'analyse.

2. Analyse de la qualité des données

Vérification de l'intégrité des données identifiées, en examinant le taux de remplissage des données présélectionnées.

3. Sélection affinée des données

Choix définitif des indicateurs et détermination des pays cibles pour l'analyse détaillée.

4. Elaboration du score et visualisation

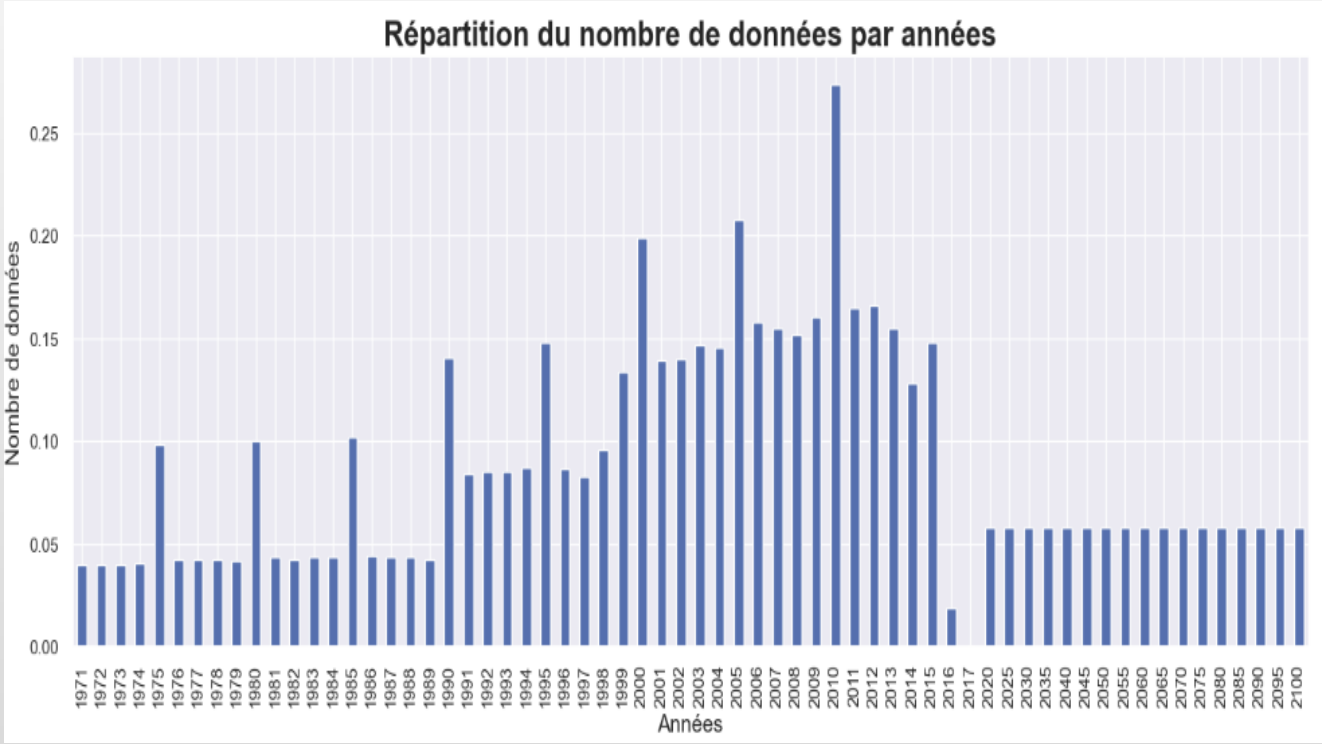
Pondération/score d’attractivité pour évaluer le potentiel de marché et présentation des résultats.

EXPLORATION PRÉLIMINAIRE DES DONNÉES

2. Pré-sélection des données

Dans cette étape, nous ciblons et extrayons les informations cruciales à partir des ensembles de données éducatives :

- **Importation des fichiers clés : Utilisation de EdStatsData.csv, EdStatsSeries.csv, et EdStatsCountry.csv pour notre analyse.**
- **Extraction des variables utiles :** Identification des variables essentielles pour l'analyse, en tenant compte des objectifs de la start-up EdTech, qui sont de s'étendre à l'international et de servir les étudiants du secondaire et de l'université.
- **Filtrage des indicateurs :** Sélection des mesures les plus significatives pour l'enseignement en ligne.
- **Plages temporelles :** Le Dataframe 'data' a été divisé en données historiques couvrant de 1970 à 2016 et en projections futures de 2020 à 2100.



3. Evaluation de la qualité des données

- En termes de qualité des données, un taux de remplissage de 87.9% est assez élevé, ce qui indique que la majorité des données sont présentes

```
taux_nan =data.isna().mean()
taux_nan_mean = taux_nan.mean()
print(f'Le taux de remplissage du dataframe data est de {1-taux_nan_mean:.2%}')
```

```
Le taux de remplissage du dataframe data est de 12.10%
```

EXPLORATION PRÉLIMINAIRE DES DONNÉES

4. Sélection des données pertinentes et choix d'indicateurs

Indicateur de connectivité à Internet: IT.NET.USER.P2

(Utilisateurs d'Internet (pour 100 personnes) en %)

Indicateur économique: NY.GDP.PCAP.CD

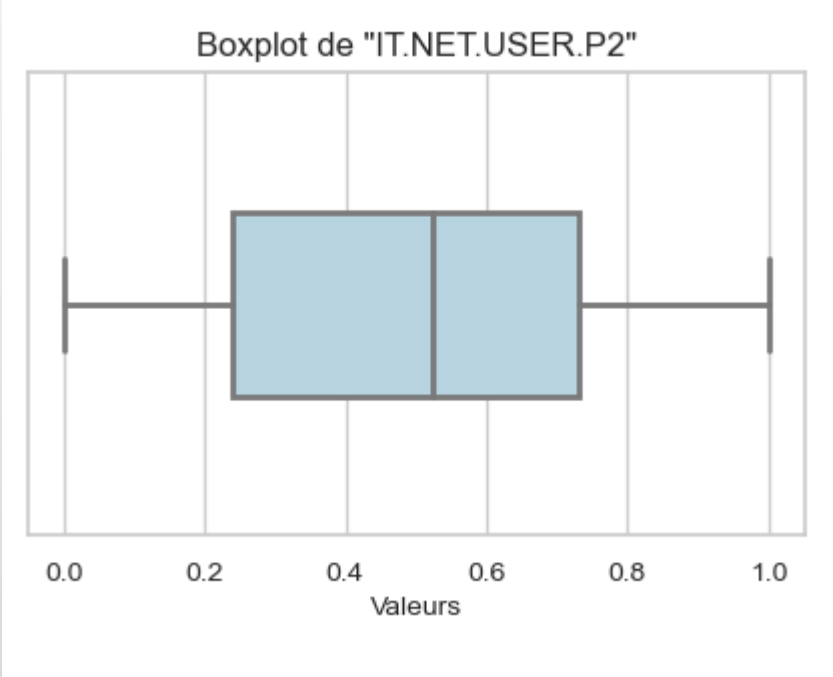
PIB par habitant (en dollars)

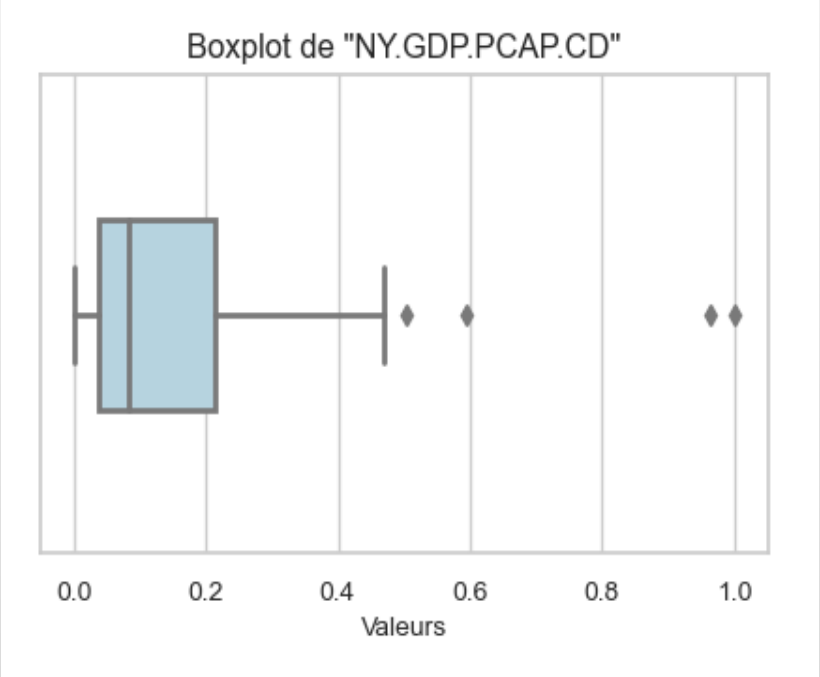
Indicateurs de Population Etudiante : UIS.E.3 et SE.TER.ENRL

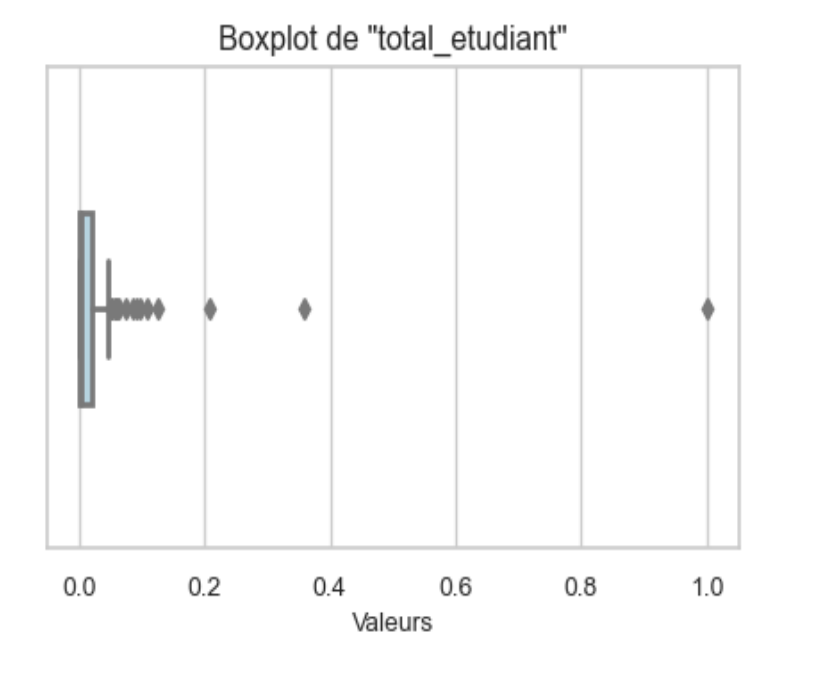
(Population étudiante dans le secondaire supérieur et dans le tertiaire)

EXPLORATION PRÉLIMINAIRE DES DONNÉES

5. Statistiques et indicateurs retenus







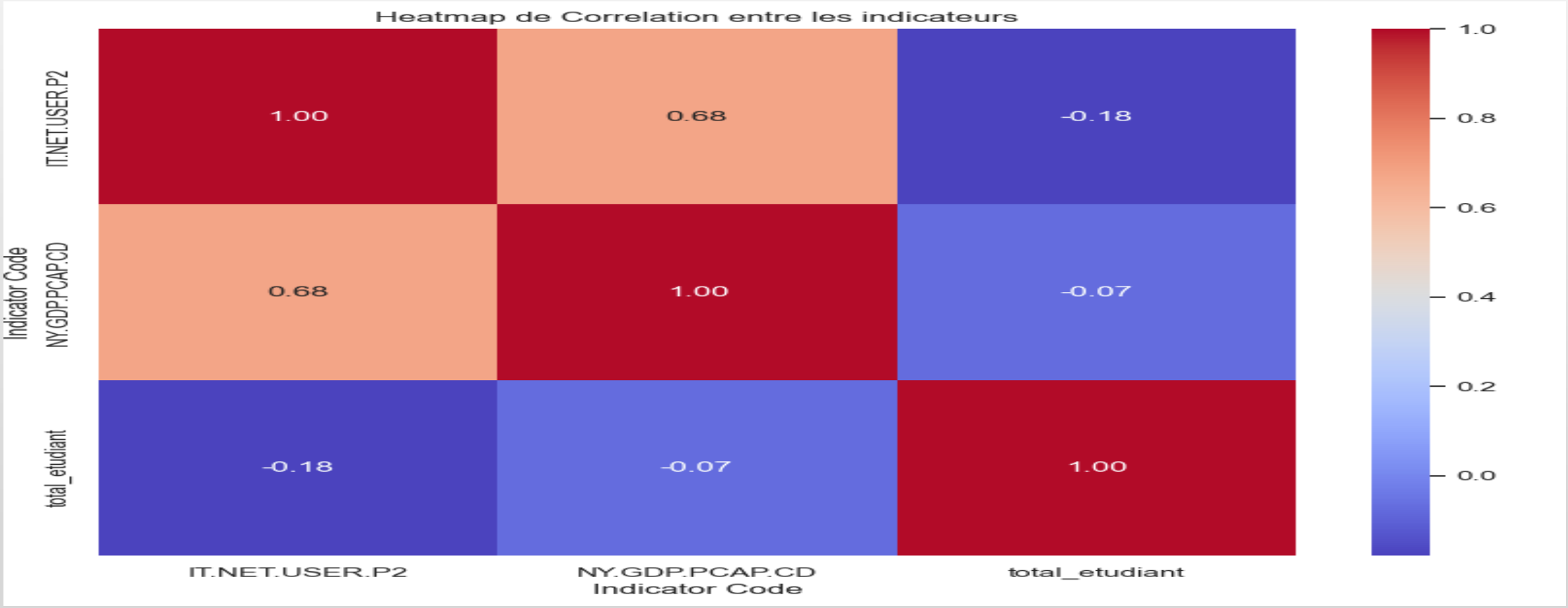
•Pour l'indicateur "IT.NET.USER.P2" : "Une médiane au-dessus de 50% pour l'utilisation d'Internet montre un terrain fertile pour l'expansion des services éducatifs en ligne dans de nombreux pays.

•Pour l'indicateur "NY.GDP.PCAP.CD" : Le PIB par habitant élevé dans certains pays, comme le montrent les valeurs aberrantes, indique un potentiel de marché pour des cours en ligne haut de gamme.

•Pour l'indicateur "total étudiant" : Une concentration basse d'étudiants dans la majorité des pays souligne l'opportunité d'étendre l'accès à l'éducation en ligne à une population plus large.

EXPLORATION PRÉLIMINAIRE DES DONNÉES

CORRELATIONS:

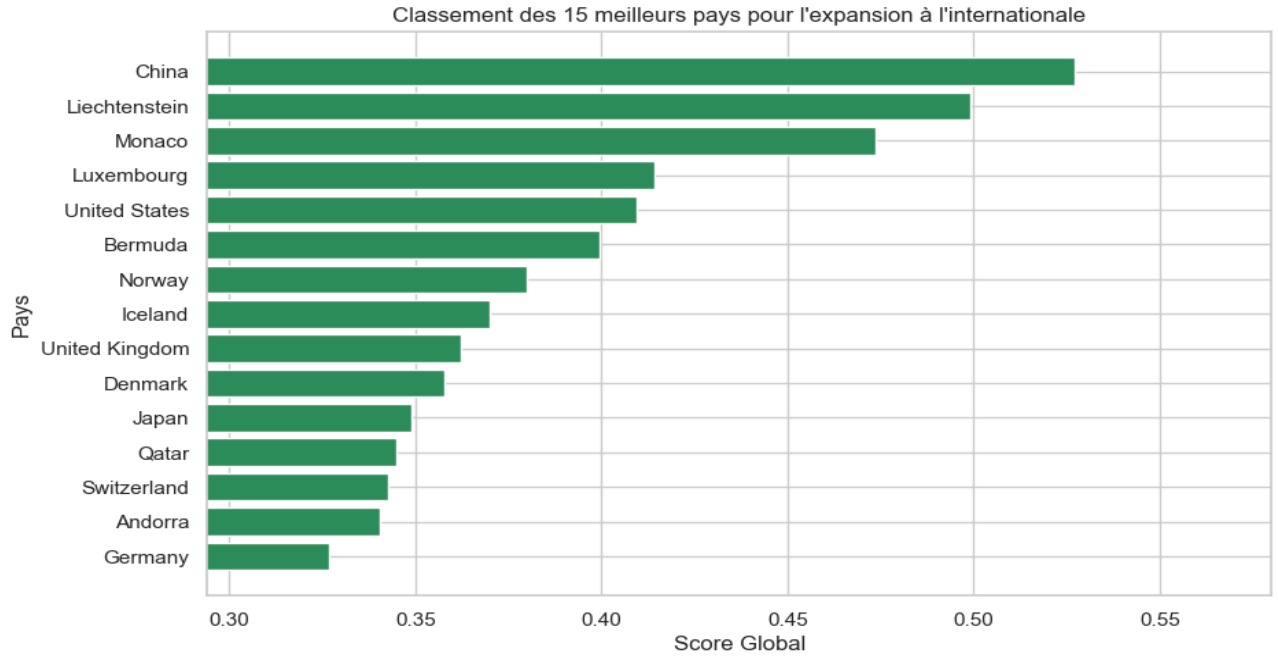


La heatmap de corrélation révèle une corrélation positive significative entre l'utilisation d'Internet (IT.NET.USER.P2) et le PIB par habitant (NY.GDP.PCAP.CD

EXPLORATION PRÉLIMINAIRE DES DONNÉES

6. Pondération/Score d'attractivité

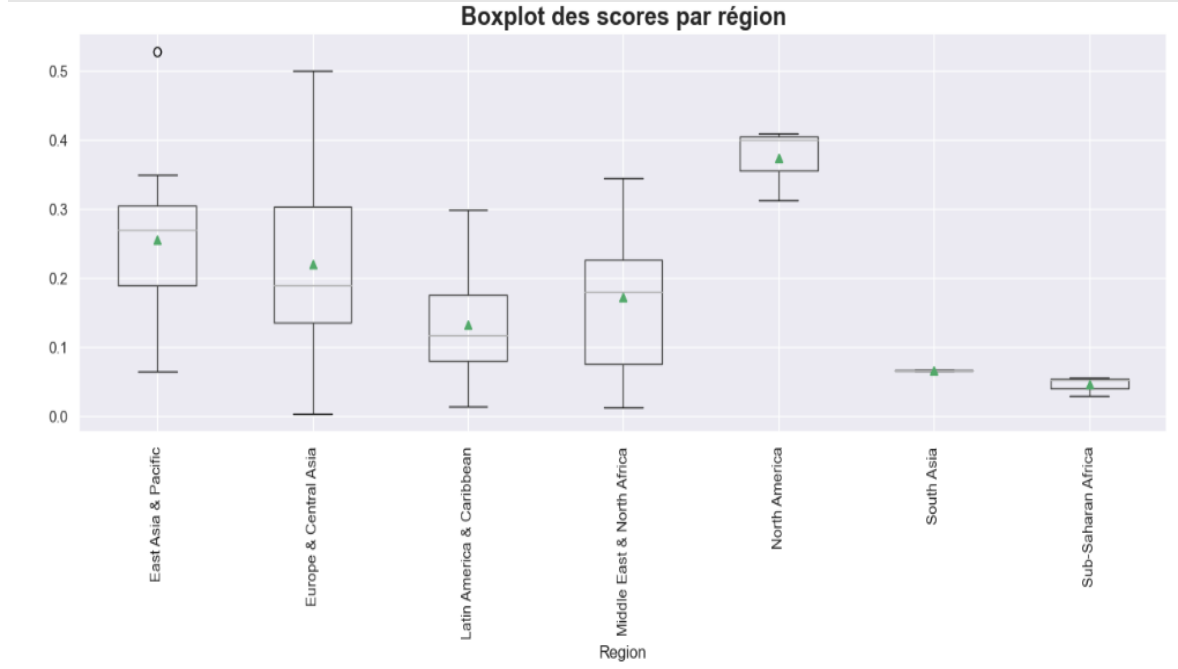
Indicator Code	Country Name	Region	IT.NET.USER.P2	NY.GDP.PCAP.CD	total_etudiant	weighted_score
22	China	East Asia & Pacific	0.066303	0.037432	1.000000	0.527377
55	Liechtenstein	Europe & Central Asia	0.996971	1.000000	0.000026	0.499104
66	Monaco	Europe & Central Asia	0.937149	0.963090	0.000012	0.473769
57	Luxembourg	Europe & Central Asia	0.984533	0.593536	0.000347	0.414241
101	United States	North America	0.542620	0.335274	0.359084	0.409383
14	Bermuda	North America	0.995024	0.504361	0.000033	0.399396
72	Norway	Europe & Central Asia	0.980476	0.415118	0.005947	0.380140
43	Iceland	Europe & Central Asia	1.000000	0.349343	0.000482	0.370109
100	United Kingdom	Europe & Central Asia	0.928185	0.231231	0.075195	0.362299
28	Denmark	Europe & Central Asia	0.973626	0.310682	0.007175	0.357812



Mise en place du scoring (attribution d'un score par pays)

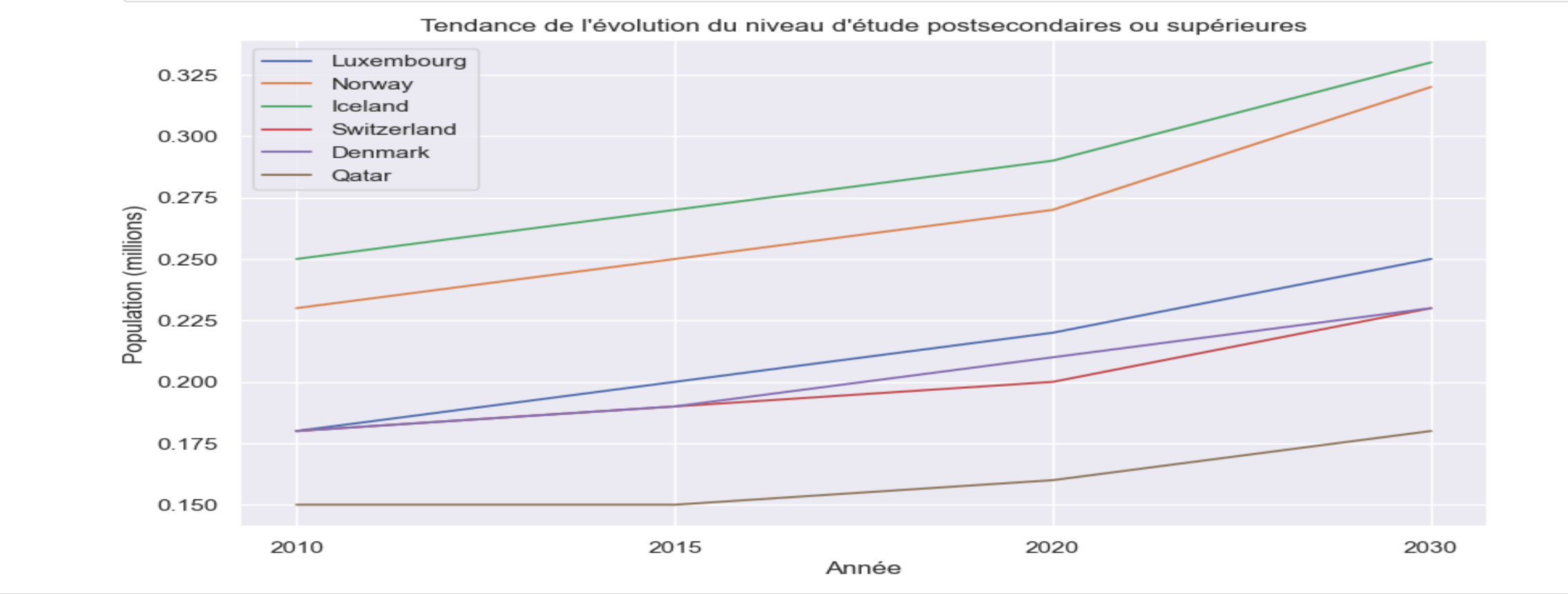
```
# Définir les poids pour chaque indicateur
poids = {
    'IT.NET.USER.P2': 0.3,
    'NY.GDP.PCAP.CD': 0.2,
    'total_etudiant': 0.5
}

# Calculer le score pondéré
dataPivot['weighted_score'] = (dataPivot[colonnes_a_normaliser] * pd.Series(poids)).sum(axis=1)
```



•Globalement, les résultats semblent logiques pour ces pays , en tenant compte de leurs caractéristiques économiques, démographiques et sociales.

SYNTHÈSE ET ÉVOLUTION FUTURE



Le graphique montre une tendance à la hausse pour la population étudiante dans l'enseignement postsecondaire ou supérieur jusqu'en 2030, avec des pays comme le Luxembourg, la Norvège et l'Islande présentant une croissance notable. Cela suggère une opportunité croissante pour les services d'éducation en ligne ciblant cette démographie en expansion.

CONCLUSION ET PERSPECTIVES

Propositions pour répondre aux enjeux professionnels et perspectives d'amélioration

Potentiel des marchés ciblés : Les analyses indiquent des opportunités dans les pays à haut revenu et forte connectivité Internet, tels que la Chine, l'Inde et les États-Unis.

Priorités d'expansion : Les pays avec les scores d'indicateurs les plus élevés devraient être la priorité pour une première phase d'expansion.

Tendance et croissance : L'Inde se démarque avec une croissance significative, suggérant un potentiel de marché en pleine expansion.

Stratégie de localisation : Considérer la langue et la proximité géographique pour une stratégie d'implantation efficace.

Données supplémentaires : Intégrer des données plus récentes et des indicateurs supplémentaires tels que les tendances de la concurrence et les données spécifiques au secteur éducatif pour affiner l'analyse.

Perspectives : Étudier les indicateurs de compétence numérique, les dépenses en éducation par habitant, et la démographie des apprenants pour une meilleure compréhension du marché.

MERCI

Des questions ?