Projet 2
Analysez des données de systèmes éducatifs

Guille Anaïs – Parcours Data Scientist

Mentor : Ahmed Tidiane Balde

Sommaire

I- Problématique

- ***Academy*** : start-up de la EdTech proposant des contenus de formations en ligne (niveau lycée et université)
- Projet d'expansion à l'international

→Réaliser une analyse pré-exploratoire des données sur l'éducation de la Banque Mondiale des Données

→Explorer les pays avec un fort potentiel de clients pour les services d'Academy et l'évolution de ce potentiel

II- Présentation du jeu de données

EdStatsCountry	• Liste des pays avec diverses informations (géographiques, économiques…) à leurs sujets • 241 lignes et 32 colonnes • Données manquantes mais aucun doublon
EdStatsCountry-Series	• Apporte des commentaires divers par pays et par indicateur • 613 lignes et 4 colonnes • Aucune données manquantes ou doublons dans les 3 premières colonnes
EdStatsData	• Contient les valeurs mesurées par pays, par indicateur et par année entre 1970 et 2100 • 886930 lignes et 70 colonnes • Données manquantes mais aucun doublon
EdStatsFootNote	• Apporte des commentaires divers par pays, par année et par indicateur • 643638 lignes et 5 colonnes • Aucune donnée manquante dans les 4 premières colonnes ou doublons
EdStatsSeries	• Contient une definition et une classification par topic ces indicateurs • 3665 lignes et 21 colonnes • Données manquantes mais aucun doublon

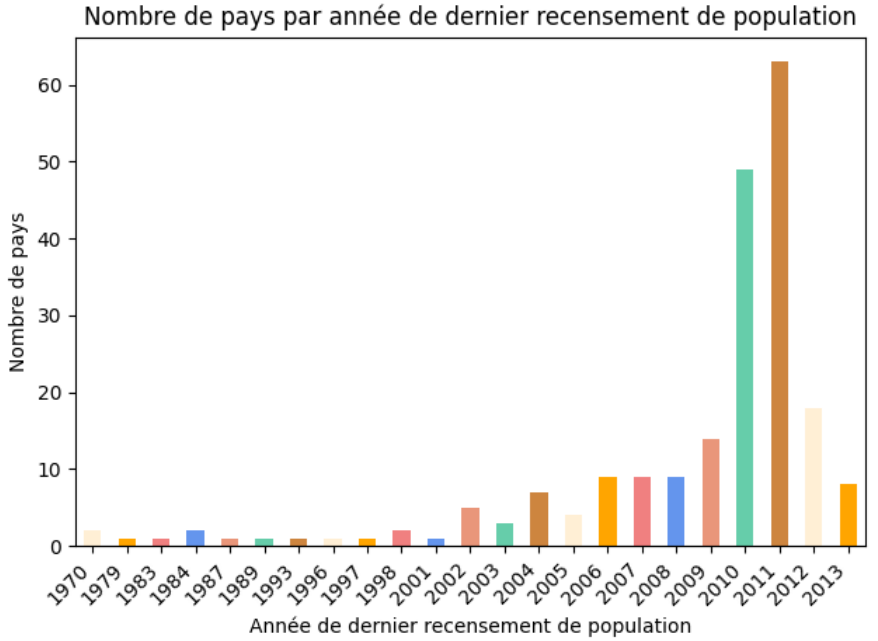
II- Présentation du jeu de données

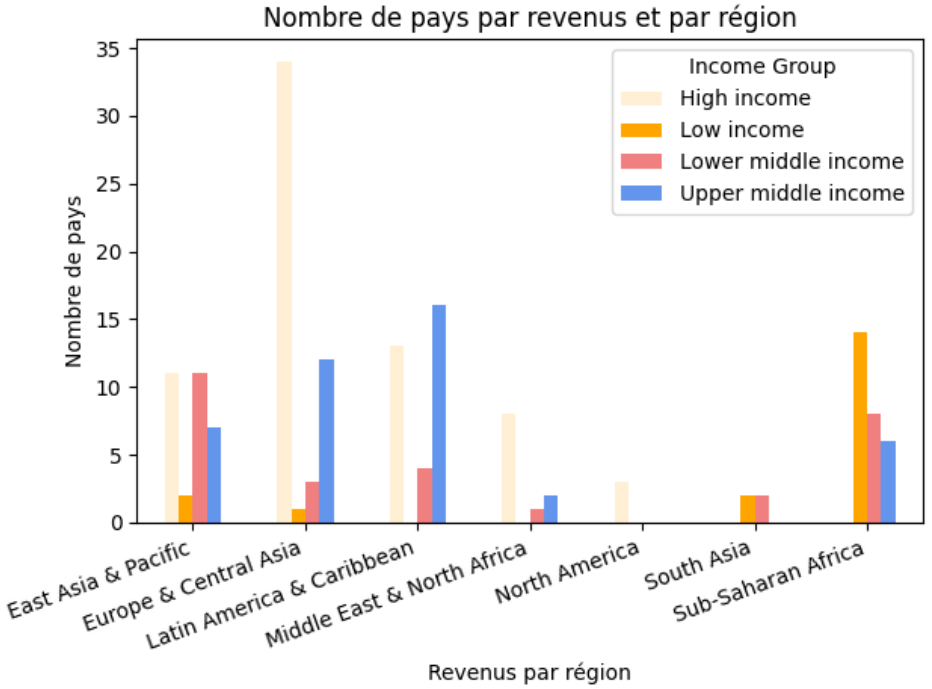
EdStatsCountry	• Liste des pays avec diverses informations (géographiques, économiques…) à leurs sujets • 241 lignes et 32 colonnes • Données manquantes mais aucun doublon
EdStatsCountry-Series	• Apporte des commentaires divers par pays et par indicateur • 613 lignes et 4 colonnes • Aucune données manquantes ou doublons dans les 3 premières colonnes
EdStatsData	• Contient les valeurs mesurées par pays, par indicateur et par année entre 1970 et 2100 • 886930 lignes er 70 colonnes • Données manquantes mais aucun doublon
EdStatsFootNote	• Contient les valeurs mesurées par pays, par indicateur et par année entre 1970 et 2100 • 643638 lignes et 5 colonnes • Aucune donnée manquante dans les 4 premières colonnes ou doublons
EdStatsSeries	• Contient une definition et une classification par topic ces indicateurs • 3665 lignes er 21 colonnes • Données manquantes mais aucun doublon

2.1- EdStatsCountry

- Suppression des colonnes inutiles à notre analyse
- Suppression des groupes de pays
- Suppression des pays avec une année de recensement de population antérieure à 2008
- Suppression des pays avec un revenu faible ou moyen faible.

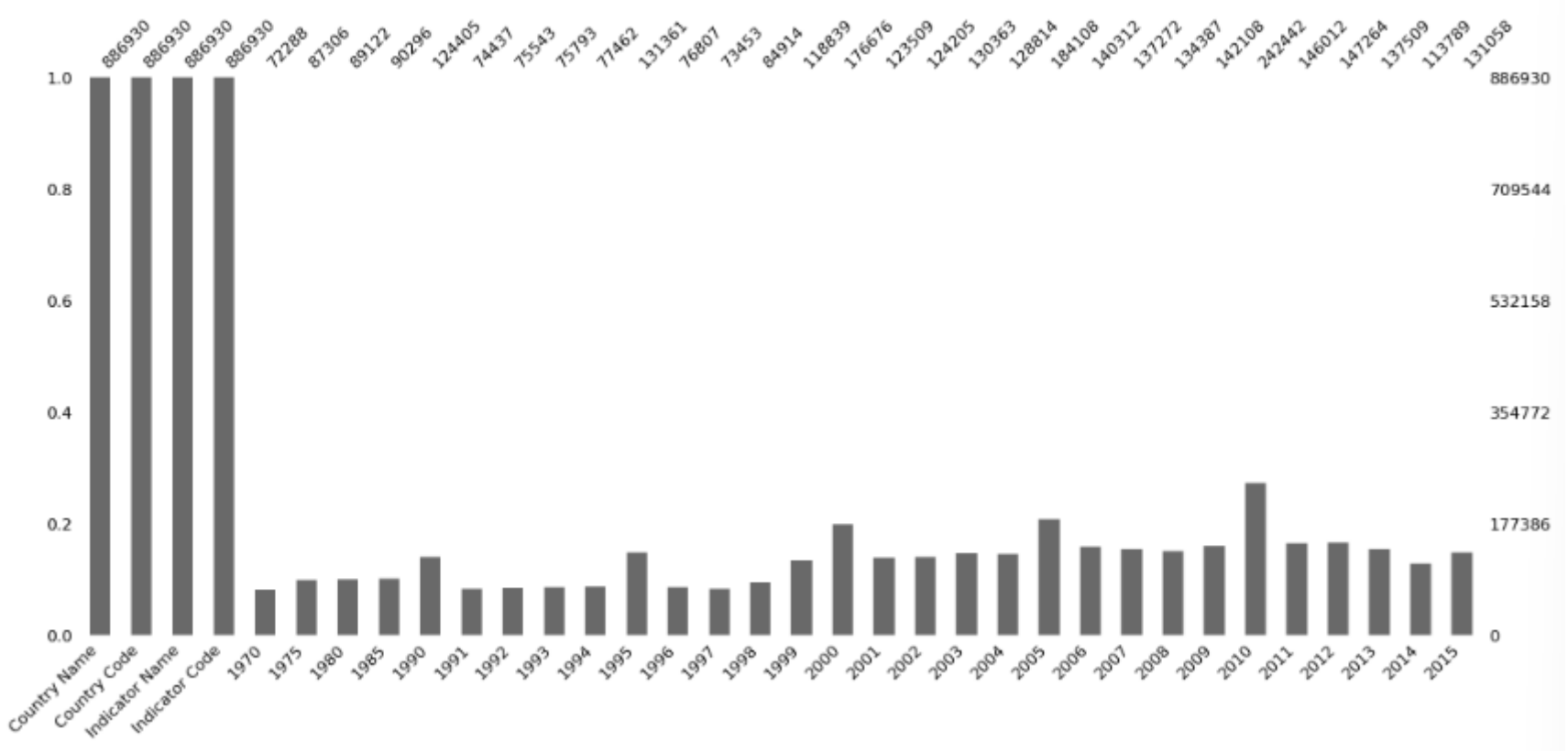
241 pays → 113 pays





2.2- EdStatsData

- Suppression des colonnes avec valeurs manquantes > 94% :



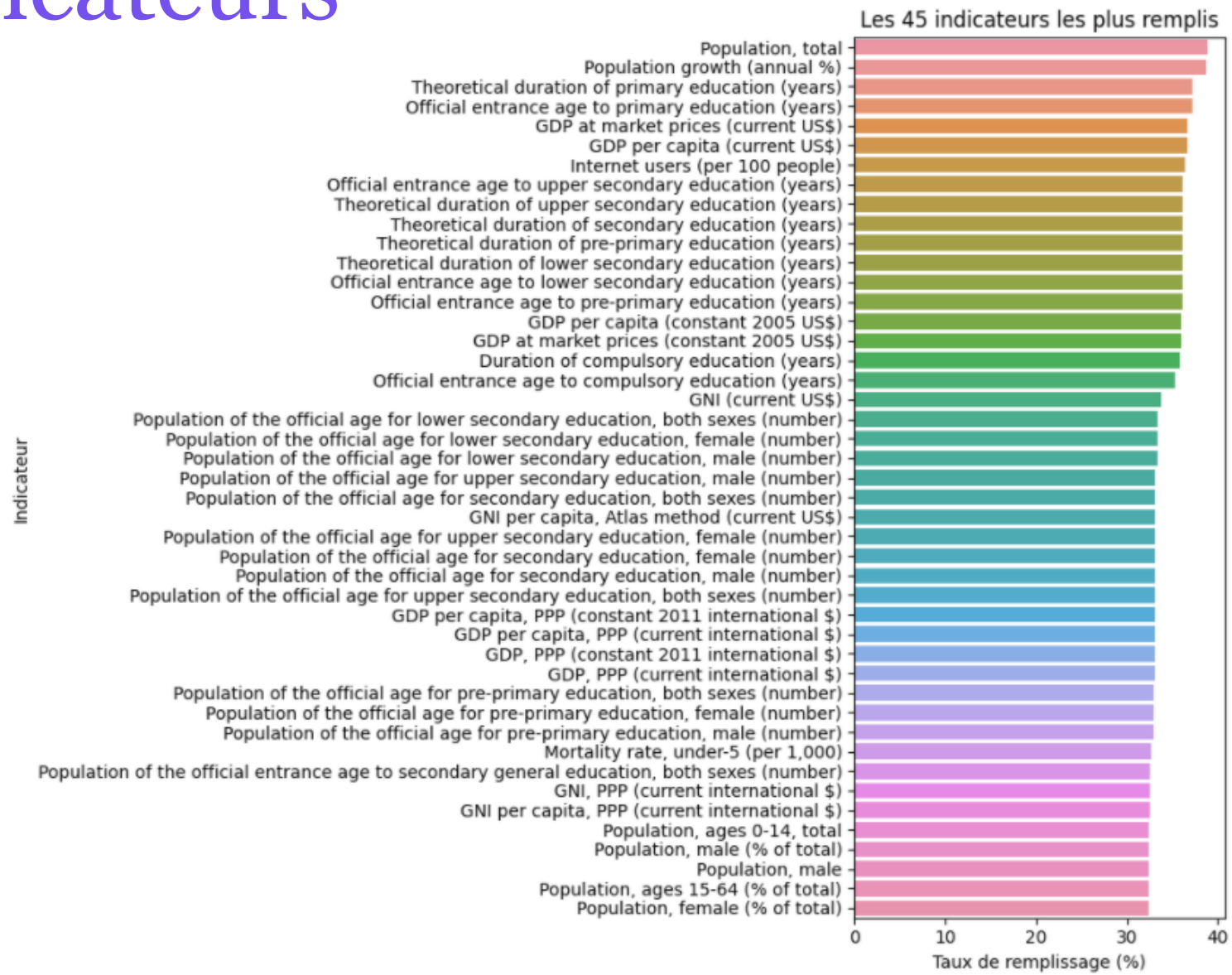
➢ Période temporelle conservée : 2010 à 2015

III- Choix des indicateurs

- Jointure interne entre ‘EdStatsCountry’ et ‘EdStatsData’ via ‘Country Code’

- Suppression des indicateurs avec taux de remplissage des valeurs < 20%

3665 indicateurs → 536 indicateurs



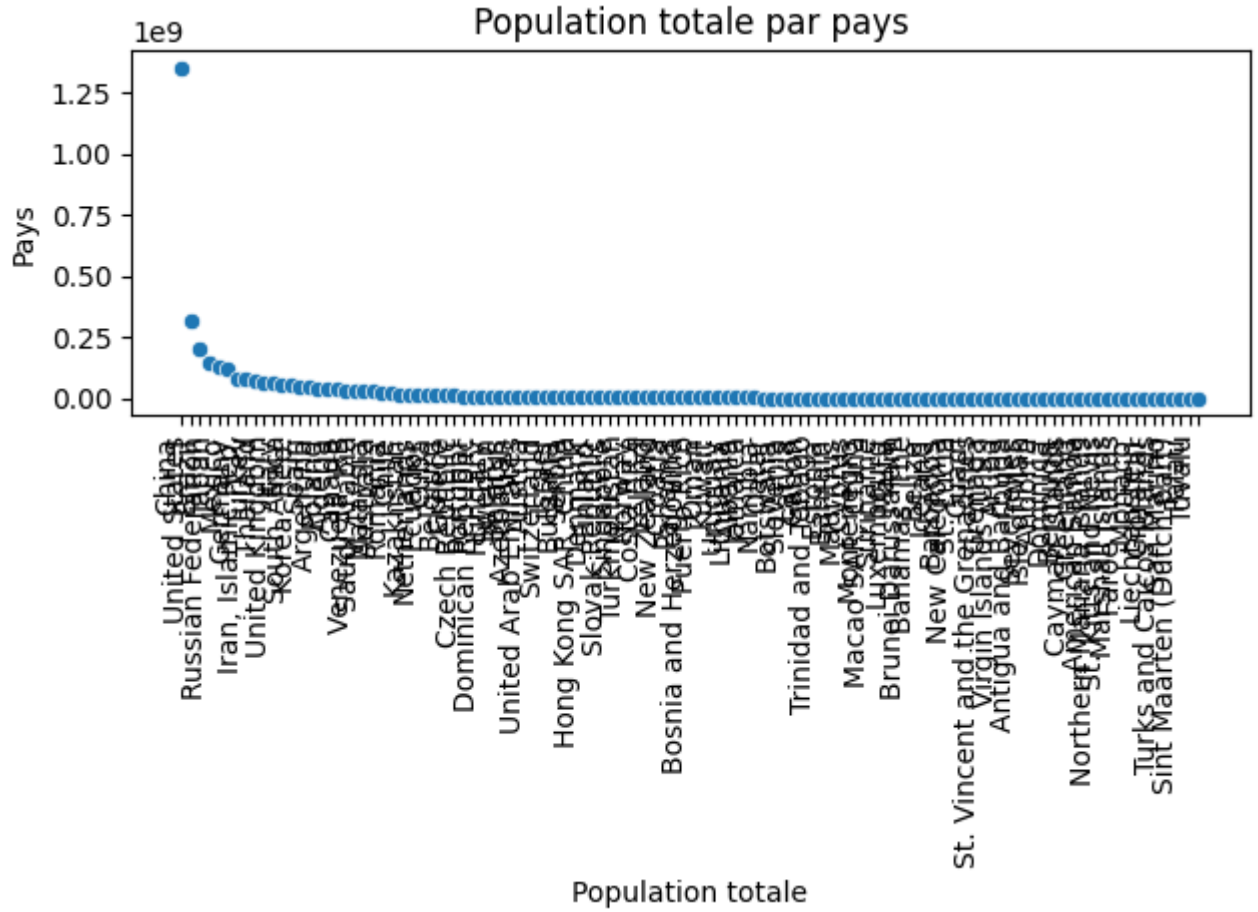
III- Choix des indicateurs

- **IT.NET.USER.P2** :Internet users (per 100 people)
- **SE.TER.ENRL** : Enrolment in tertiary education, all programmes, both sexes (number)
- **UIS.E.3** : Enrolment in upper secondary education, both sexes (number)
- **NY.GDP.PCAP.PP.CD** : GDP per capita, PPP (current international $)
- **SP.POP.TOTL**: Population, total
- **SP.POP.GROW** : Population growth (annual %)

3.1 Population totale

- Basé sur la moyenne entre 2010 et 2015
- Suppression des pays avec une population totale < 1 270 000 habitants (quartile 35%)

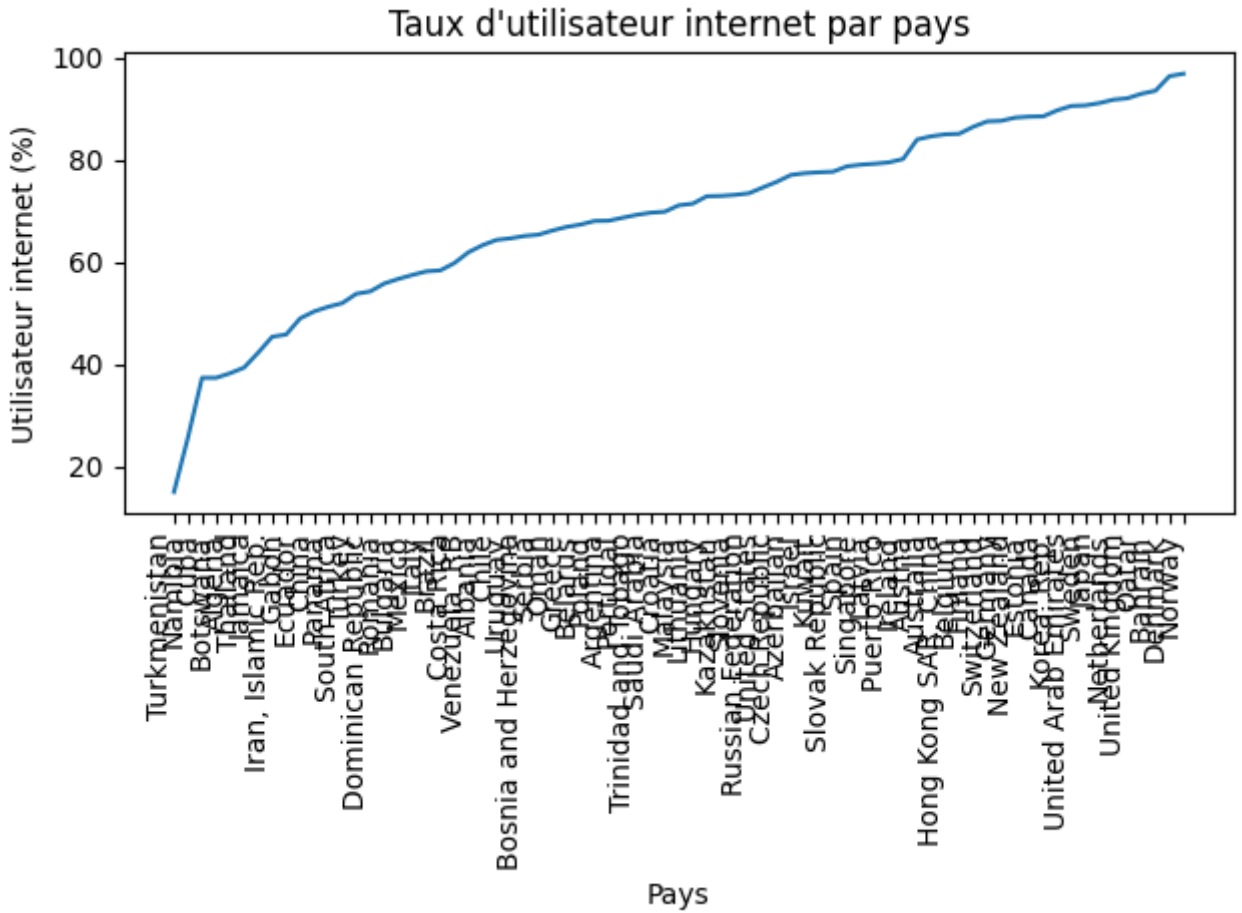
113 pays → 73 pays



3.2 Internet Users (per 100 people)

- Basé sur l'année la plus récente (2015)
- Suppression des pays avec utilisateurs internet <62% (Q2)

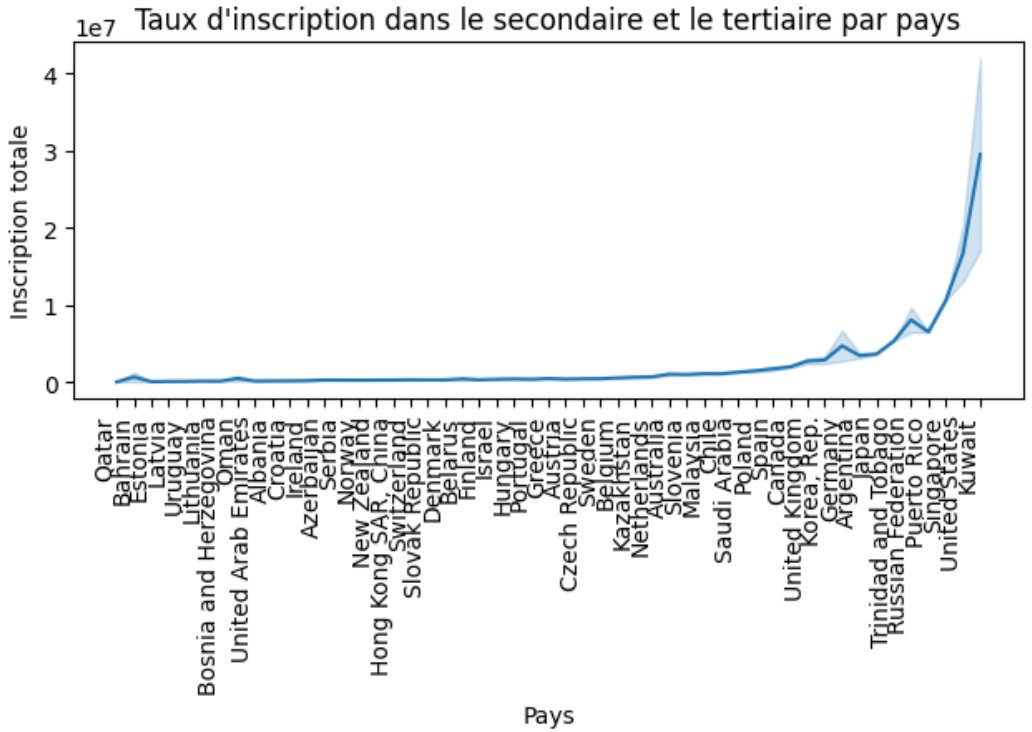
73 pays → 51 pays

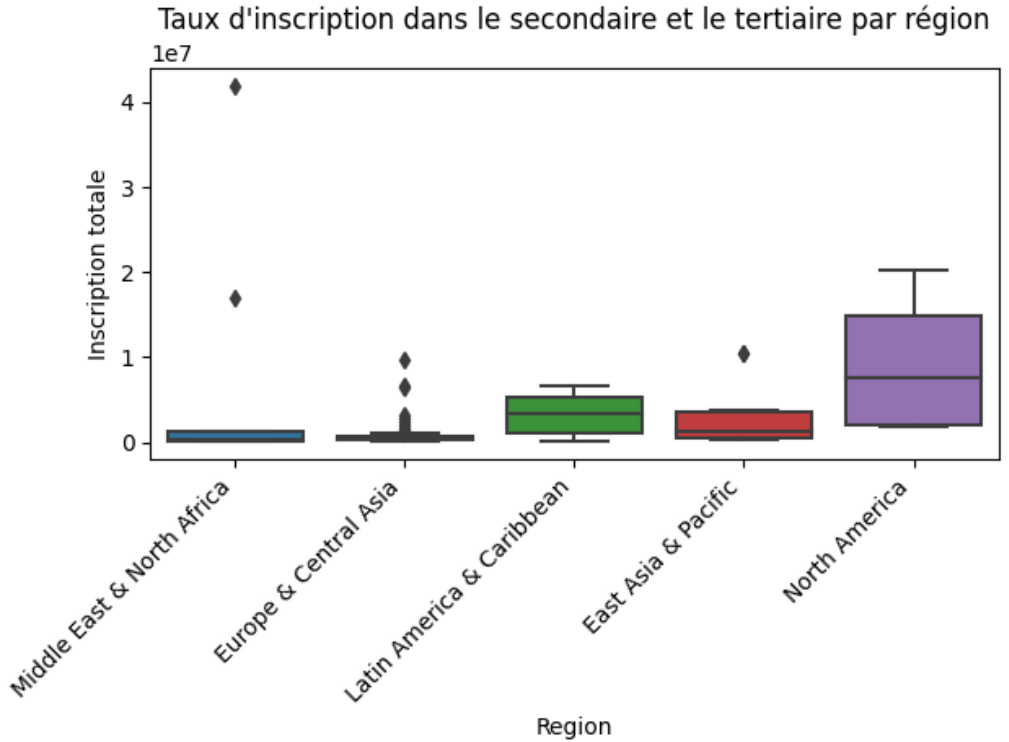


3.3 Taux d'inscription dans le secondaire et le tertiaire

- Basé sur la moyenne entre 2010 et 2015
- Somme des inscriptions dans le secondaire et le tertiaire par pays
- Suppression des pays avec taux d'inscription <878 000 (Q2)

51 pays → 25 pays





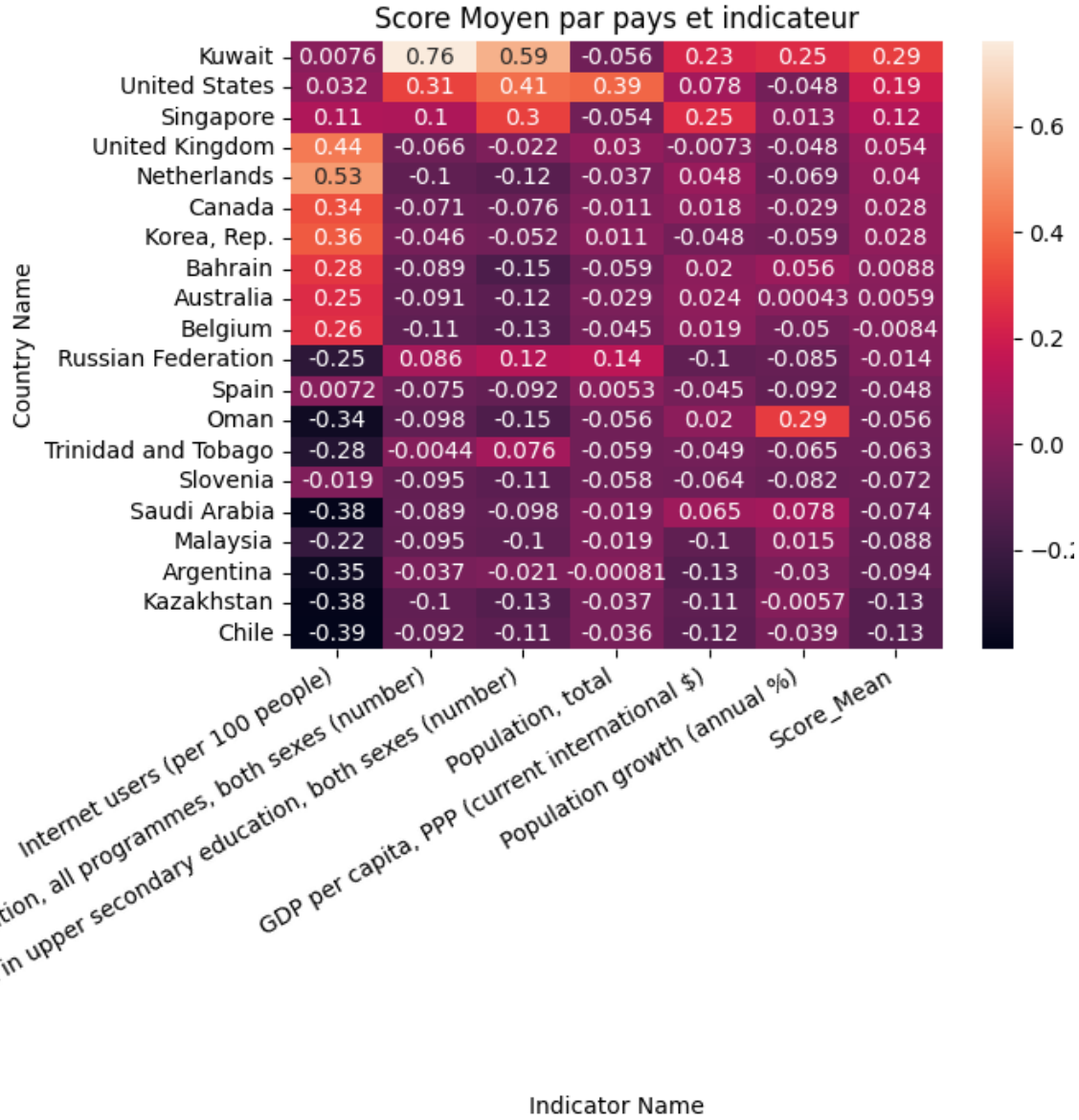
3.4 Croissance démographique

- Basé sur la croissance entre 2010 et 2015
- Suppression des pays en décroissance

25 pays → 20 pays

IV- Score d'attractivité

- Normalisation des données avec un z-score
- Pondération des indicateurs
- Score moyen par pays



Conclusion

- Le Koweit, les États-Unis, Singapoure, le Royaume-Uni et le Pays-Bas émergent comme les destinations les plus prometteuses.

- Richesse des informations fournies par le jeu de données cruciale pour notre analyse

- Cependant, le jeu de données présente quelques lacunes : *Problème de structuration de données, valeurs manquantes, manque d'indicateur sur les perspectives d'évolution d'Academy dans les pays.*

→ Pertinence du jeu de données offrant à Academy des perspectives d'expansions internationales intérressantes.