# Analyse de données de systèmes éducatif

Projet d’expansion à l’international d’Academy

# Objectif : Identifier les pays pertinents pour une expansion à l'international

**Enjeux : Analyse pré-exploratoire de données issues de la Banque Mondiale**

Plan :

# 1. Description et analyse de la validité des jeux de données

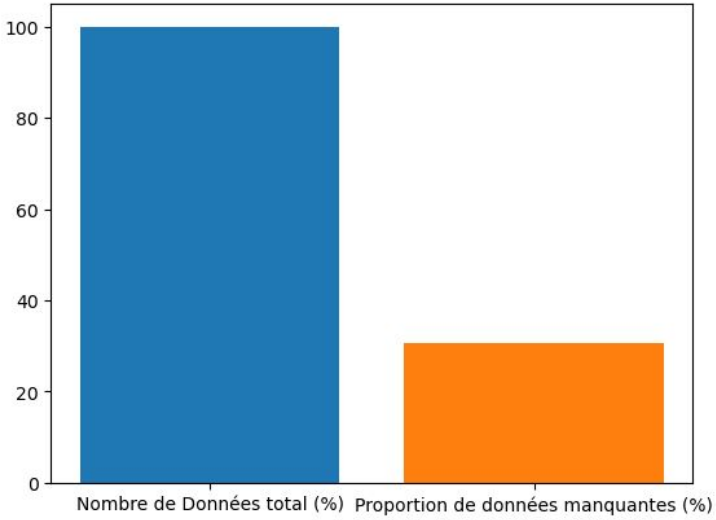
## 1.1. Informations de chaque jeu de données

Nous avions **5 jeux de données** à notre disposition :

- ‘Edstatscountry’, data frame de 241 lignes et 32 colonnes, comprenant 7712 éléments dont 30.52% de données manquantes.

  Ce data frame nous donne des informations sur les pays à disposition pour l’analyse, avec notamment la région et/ou le continent d’appartenance, la monnaie utilisée les dernières dates de recensement…

  On l’utilisera pour inclure la colonne indiquant la région de chaque pays à notre data frame principal.



# 1. Description et analyse de la validité des jeux de données

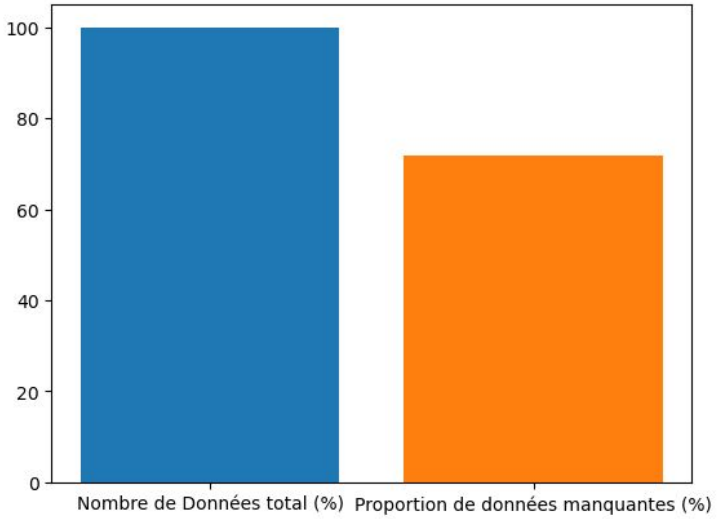
## 1.1. Informations de chaque jeu de données

Nous avions **5 jeux de données** à notre disposition :

- ‘EdStatsSeries’, data frame de 3665 lignes et 21 colonnes, comprenant 76965 éléments dont 71.72% de données manquantes.

  Ce data frame nous donne des informations sur les indicateurs à disposition : leur nom, leur définition, leur thème ainsi que leur source, entre autres.

  Il nous a été utile pour faire une sélection d’indicateur en fonction de leur thème



# 1. Description et analyse de la validité des jeux de données

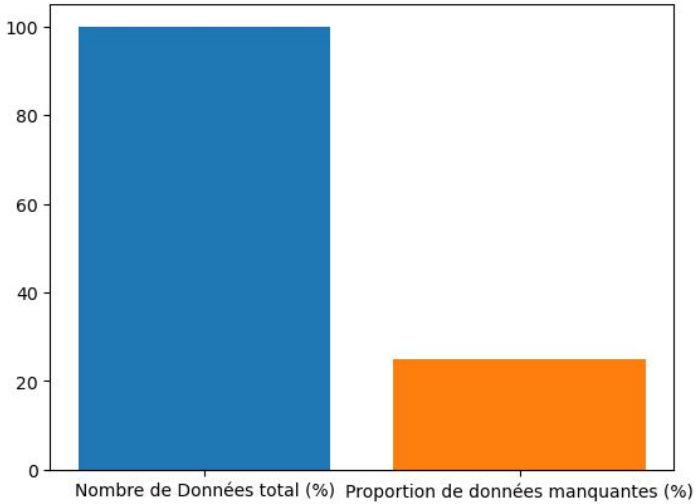
## 1.1. Informations de chaque jeu de données

Nous avions **5 jeux de données** à notre disposition :

- ‘EdStatsCountry-Series’, data frame de 613 lignes et 4 colonnes, comprenant 2452 éléments dont 25% de données manquantes.

  Ce data frame nous donne des informations sur les sources des indicateurs en fonction des pays et régions..

  Il ne nous a pas été utile pour l’analyse



# 1. Description et analyse de la validité des jeux de données

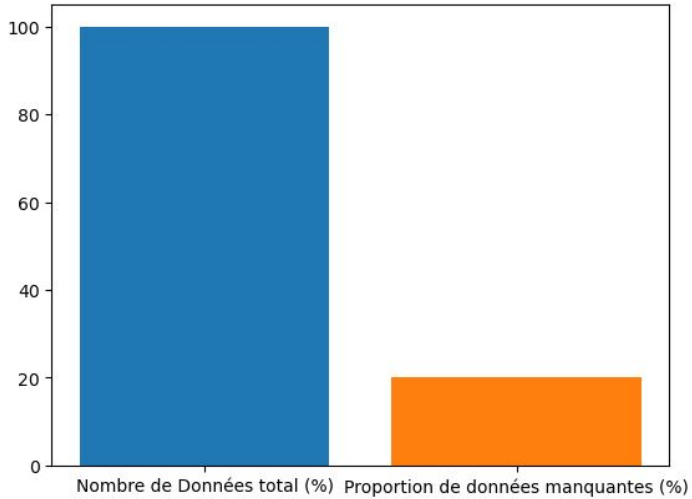
## 1.1. Informations de chaque jeu de données

Nous avions **5 jeux de données** à notre disposition :

- ‘EdStatsFootNote’, data frame de 643638 lignes et 5 colonnes, comprenant 3218190 éléments dont 20% de données manquantes.

  Ce data frame nous donne des informations sur les sources  et années d’obtention des indicateurs en fonction des pays et régions..

  Il ne nous a pas été utile pour l’analyse



# 1. Description et analyse de la validité des jeux de données

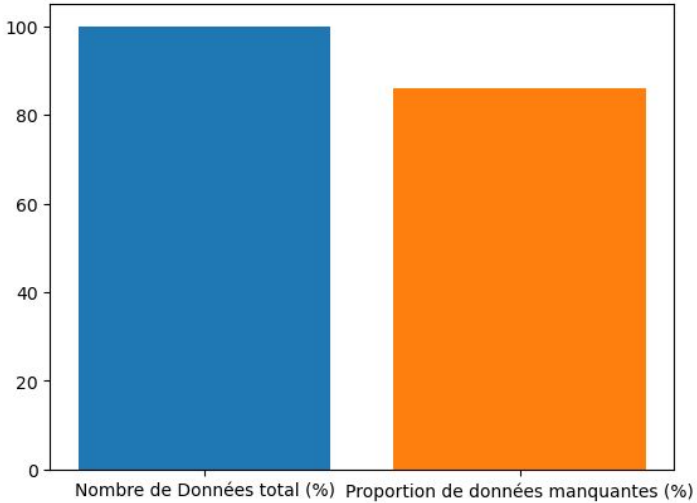
## 1.1. Informations de chaque jeu de données

Nous avions **5 jeux de données** à notre disposition :

- Et enfin '**EdStatsData**', data frame de 886930 lignes et 70 colonnes, comprenant 62 085 100 éléments dont 86.1% de données manquantes.

  C'est le **jeu de données principal** dans lequel se trouvent les informations qui nous ont été utiles pour l'analyse.

  Il comprend tous les indicateurs disponibles pour chaque pays et régions sur une période allant de 1970 à 2100.



# 1. Description et analyse de la validité des jeux de données

## 1.2. Sélection des variables pertinentes

C’est donc à partir du jeu de données Edstatsdata que nous avons réalisé notre analyse.

La problématique d’Academy nous a conduit à sélectionner des indicateurs répartis en 4 grands thèmes :

1) Démographie:
2) Scolarisation / Réussite scolaire
3) Technologie
4) Economie

Par ailleurs nous avons décidé, dans un premier temps, de restreindre notre analyse à la **période 2010-2025**.

# 1. Description et analyse de la validité des jeux de données

## 1.2. Sélection des variables pertinentes

Les indicateurs démographiques choisis :

- la population totale
- la population des 15-24 ans
- le taux de croissance de la population (en pourcentage annuel)

Les indicateurs technologiques choisis :

- Nombre d'ordinateurs personnel (pour 100 personnes)
- Nombre d'utilisateurs d'Internet (pour 100 personnes)
    - la Banque Mondiale définit cet indicateur comme suit : "les utilisateurs d'Internet sont des personnes ayant eu accès à Internet, depuis n'importe où, dans les 3 derniers mois. Cela peut être via un ordinateur, un smartphone etc.."

# 1. Description et analyse de la validité des jeux de données

## 1.2. Sélection des variables pertinentes

Les indicateurs de scolarisation choisis :

- le taux de scolarisation dans le secondaire (%)
  - Il s'agit du nombre total de personnes scolarisées dans le secondaire (collège et lycée) indépendamment de leur âge rapporté à la population ayant l'âge officiel de scolarisation dans le secondaire. C'est la raison pour laquelle l'indicateur peut dépasser les 100% : des personnes moins âgées et plus âgées peuvent aussi être scolarisées dans le secondaire, en raison d'une scolarisation précoce, de saut de classe, ou au contraire de redoublement ou difficultés scolaires.
- le taux de scolarisation dans le tertiaire (%)
  - Il s'agit du nombre total de personnes scolarisées dans le tertiaire (post bac et universités) indépendamment de leur âge rapporté à la population ayant l'âge officiel de scolarisation dans le tertiaire. C'est la raison pour laquelle l'indicateur peut dépasser les 100% : des personnes moins âgées et plus âgées peuvent aussi être scolarisées dans le secondaire, en raison d'une scolarisation précoce, de saut de classe, ou au contraire de redoublement ou difficultés scolaires.
- le taux de réussite scolaire dans le 'bas' secondaire (%)
  - Il s'agit du ratio du nombre de personnes ayant entre 3 et 5 ans de plus que l'âge auquel on atteint la dernière année dans le secondaire, ayant validé ce cycle, sur le total du nombre de personnes du même âge.

# 1. Description et analyse de la validité des jeux de données

## 1.2. Sélection des variables pertinentes

Les indicateurs économiques choisis :

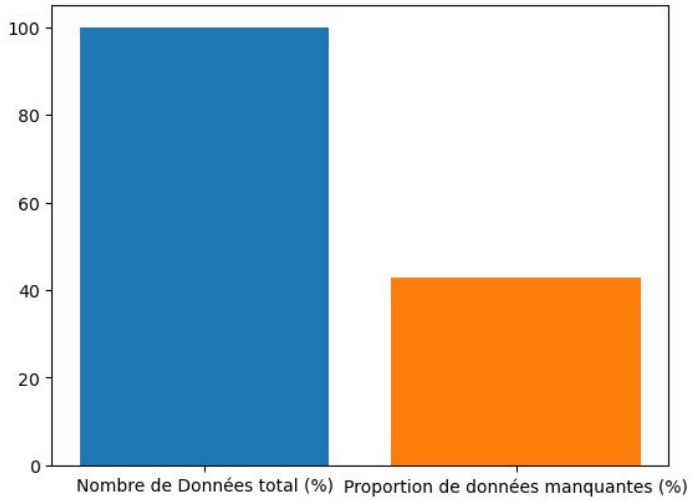
- le PIB par tête, en parité de pouvoir d'achat, et en dollars constant en base 2011.
  - Nous pensons qu'il s'agit là du meilleur indicateur économique pour effectuer des comparaisons entre pays. En effet, il est corrigé de l'inflation et la parité de pouvoir d'achat permet de "gommer" des différences de prix entre pays.
- la part des dépenses publiques en éducation, rapportée au PIB (%)
- la part des dépenses pour des projets de long terme dans les dépenses liées aux institutions publiques
  - Il s'agit des dépenses en biens et actifs physiques liés à l'éducation et ayant une durabilité moyenne supérieure à 1 an.
- la part des dépenses courantes dans les dépenses liées aux institutions publiques
  - Il s'agit des dépenses liées à l'éducation, devant servir et être consommées dans l'année.

# 2. Visualisation, analyse et traitements des valeurs manquantes

## 2.1. Visualisation et analyse

Après sélection des variables, notre jeu de données comprend 2904 lignes et 14 colonnes.

Le taux de valeur manquante est de 42.94%.

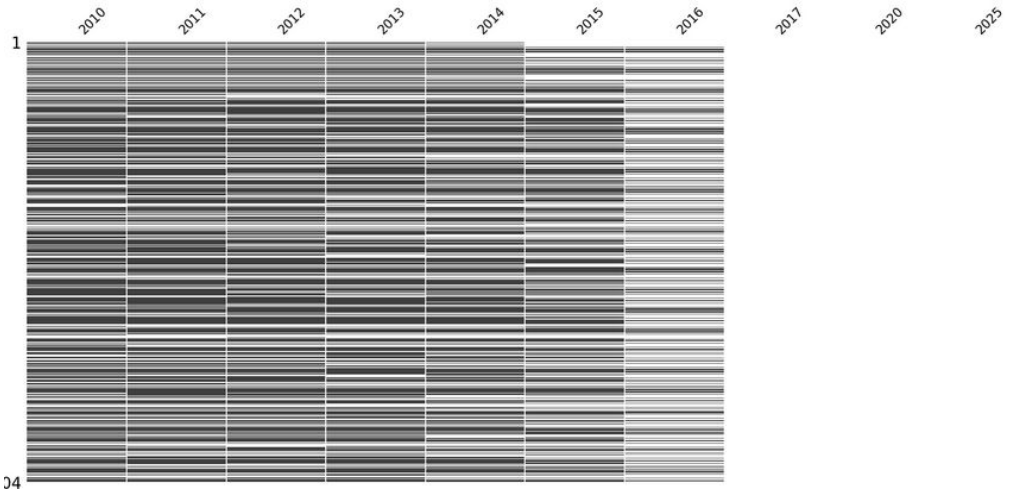


# 2. Visualisation, analyse et traitements des valeurs manquantes

## 2.1. Visualisation et analyse

La représentation graphique d'une matrice des données manquantes par année nous permet de voir qu'il n'y a **aucune donnée disponible au-delà de 2016**, tous indicateurs confondus.

Nous décidons donc de supprimer les colonnes de 2017 à 2020 et réalisons une analyse par variable ensuite.



# 2. Visualisation, analyse et traitements des valeurs manquantes

## 2.1. Visualisation et analyse

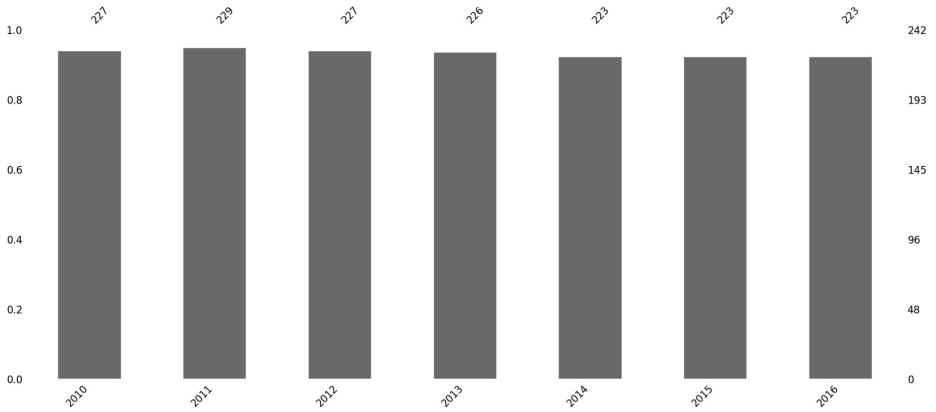
Cette analyse par variable nous a conduit à choisir l'année **2015 comme année d'étude**. Un nombre important de données étant manquante en 2016 pour certains indicateurs.

Par ailleurs elle nous conduit à **supprimer ces variables** :

- la part des dépenses pour des projets de long terme dans les dépenses liées aux institutions publiques
- la part des dépenses courantes dans les dépenses liées aux institutions publiques
- Nombre d'ordinateurs personnel (pour 100 personnes)

Elles sont en effet trop peu renseignées sur cette période. Pas du tout pour cette dernière.

Exemple d'analyse des données manquantes, par année, pour l'indicateur "Utilisateurs d'Internet (pour 100 personnes)" :



# 2. Visualisation, analyse et traitements des valeurs manquantes

## 2.2. Traitement des données manquantes

Etant donnée la diversité des pays que contient notre data frame, il nous semble plus pertinent de réaliser des les **imputations**, pour les données manquantes en 2015, **par la dernière donnée renseignée non nulle**, pour chaque pays.

Une imputation par la médiane ou la moyenne nous semble ici inappropriée.

Après l'application de ce traitement, nous avons regardé à nouveau quels étaient les pays ou groupes de pays pour lesquels la donnée était toujours manquante, pour chaque indicateur. C' était le cas lorsque la donnée n'était en fait jamais renseignée sur la période que nous avons sélectionnée.

# 2. Visualisation, analyse et traitements des valeurs manquantes

## 2.2. Traitement des données manquantes

Conséquences de ce traitement :

- **suppression des indicateurs suivants** :
  - taux de réussite scolaire dans le ‘bas’ secondaire (%) -> trop de valeurs manquantes et déjà 2 indicateurs ‘scolarisation’
  - part des dépenses publiques en éducation, rapportée au PIB (%)
- **suppression des pays** pour lesquels des données manquaient pour les indicateurs restants :
  - Exception : Nous choisissons de conserver le Canada, pour lequel il manquait le taux de scolarisation dans le tertiaire.

# 2. Visualisation, analyse et traitements des valeurs manquantes

## 2.2. Traitement des données manquantes

Etant donné la diversité des pays constituant notre data frame, nous nous attendions à une **grande dispersion des variables**, cela se confirmera dans la suite de l'analyse.

Il nous a semblé raisonnable de fixer un **seuil minimal de population totale** pour la suite de notre étude. Les pays trop peu peuplés offrant un potentiel de clients moindre.

Nous aurions pu fixer ce seuil par une méthode statistique, ex : 'moyenne' - 2 x 'écart-type', mais les écarts étaient trop important pour cela, nous avons donc privilégié une approche "métier", subjective certes, en fixant ce seuil minimal à **2 millions d'habitants**.

L'application de ces différents traitement nous donne un **data frame d'analyse "final" comprenant 115 pays pour 7 indicateurs.**

| Indicator Name | PIB/tête, PPP ($ contant 2011) | Part des dépensens en éducation en % du PIB (%) | Tx scolarisation secondaire (%) | Tx scolarisation tertiaire (%) | Utilisateurs d'Internet (pour 100 personnes) | Tx achèvement secondaire (%) | Tx de croissance population | Population des 15-24 ans | Population totale |
|---|---|---|---|---|---|---|---|---|---|
| count | 180.00 | 144.00 | 164.00 | 152.00 | 184.00 | 163.00 | 186.00 | 171.00 | 186.00 |
| mean | 17808.31 | 4.69 | 83.63 | 39.64 | 46.97 | 75.79 | 1.41 | 6980662.03 | 39040157.64 |
| std | 19325.46 | 1.92 | 29.40 | 28.17 | 28.40 | 26.32 | 1.23 | 24695894.19 | 143358175.56 |
| min | 626.41 | 1.02 | 9.52 | 1.71 | 1.08 | 12.71 | -2.47 | 22956.00 | 11001.00 |
| 25% | 3532.16 | 3.30 | 60.99 | 13.43 | 21.24 | 52.82 | 0.53 | 502400.50 | 1999075.25 |
| 50% | 11302.37 | 4.59 | 90.22 | 35.99 | 47.77 | 84.37 | 1.25 | 1421630.00 | 8590910.00 |
| 75% | 25009.82 | 5.51 | 102.32 | 62.64 | 71.24 | 96.64 | 2.28 | 5315339.00 | 27972844.50 |
| max | 119749.43 | 12.84 | 166.81 | 113.87 | 98.20 | 133.78 | 5.86 | 244120201.00 | 1371220000.00 |

# 3. Analyse du jeu de données principal

## 3.1. Mise en forme finale du data frame

| | PIB/tête, PPP ($ contant 2011) | Tx scolarisation secondaire (%) | Tx scolarisation tertiaire (%) | Utilisateurs d'Internet (pour 100 personnes) | Tx de croissance population | Population des 15-24 ans | Population totale |
|---|---|---|---|---|---|---|---|
| count | 115.00 | 115.00 | 115.00 | 115.00 | 115.00 | 115.00 | 115.00 |
| mean | 18031.13 | 82.89 | 41.22 | 47.02 | 1.46 | 9751936.84 | 59543311.73 |
| std | 19006.20 | 31.91 | 29.56 | 28.77 | 1.20 | 29699910.66 | 179248776.54 |
| min | 626.41 | 17.38 | 1.71 | 1.08 | -0.94 | 129378.00 | 2063531.00 |
| 25% | 3445.34 | 55.67 | 11.24 | 21.06 | 0.55 | 1067861.00 | 6475798.50 |
| 50% | 11411.94 | 91.17 | 38.48 | 48.94 | 1.32 | 2817084.00 | 16939923.00 |
| 75% | 25953.90 | 103.29 | 64.86 | 72.11 | 2.42 | 6849630.00 | 44285897.00 |
| max | 119749.43 | 166.81 | 113.87 | 96.81 | 4.41 | 244120201.00 | 1371220000.00 |

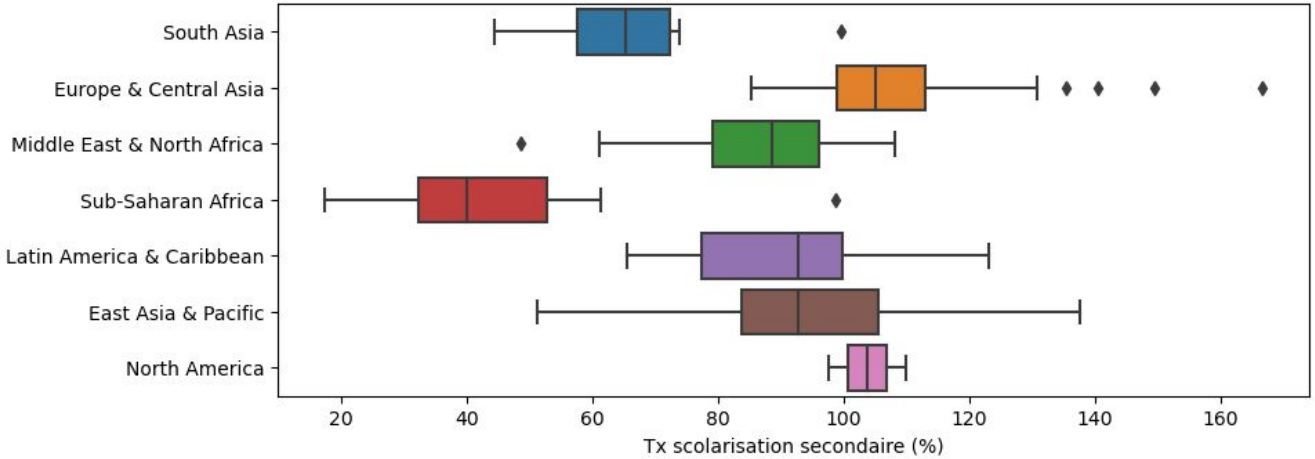
Nous obtenons donc le data frame suivant. Nous notons une grande dispersion pour chacune des variables, les populations des 15-24 ans et la population totale étant notamment impactées par la présence de la Chine et de l'Inde dans notre data frame.

Le contexte de notre étude nous conduit à considérer ces différents valeurs extrêmes comme atypiques, et non aberrantes. Nous les conservons donc.

# 3. Analyse du jeu de données principal

## 3.2. Analyse univariée

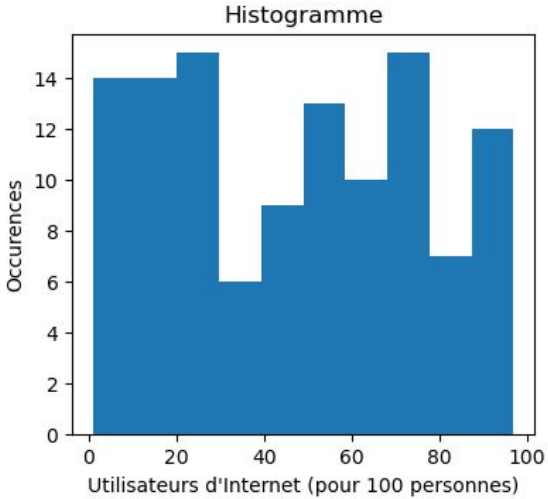
Nous avons par la suite étudié la distribution des différentes variables, une à une, en observant également les disparités régionales. Prenons l'exemple du taux de scolarisation secondaire :

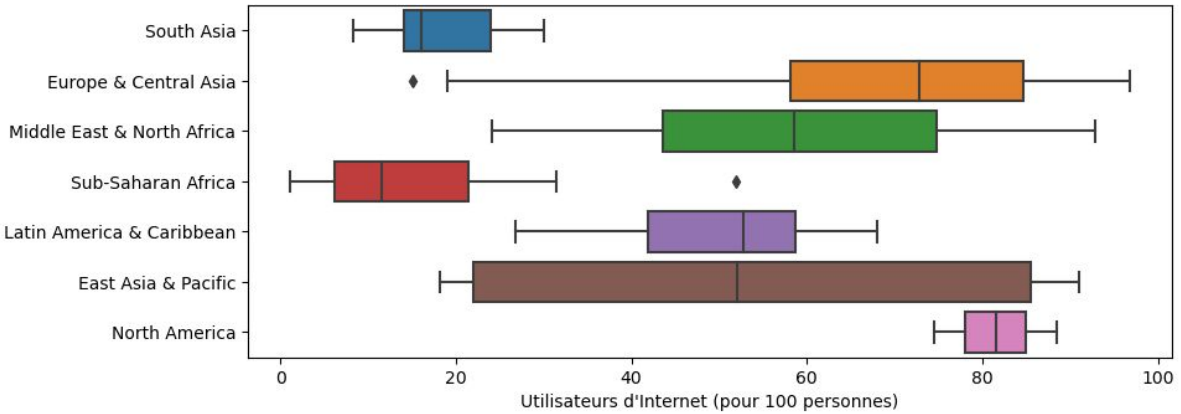


# 3. Analyse du jeu de données principal

## 3.2. Analyse univariée

La distribution des utilisateurs d'Internet est également intéressante à observer, elle est multimodale avec une dispersion élevée:

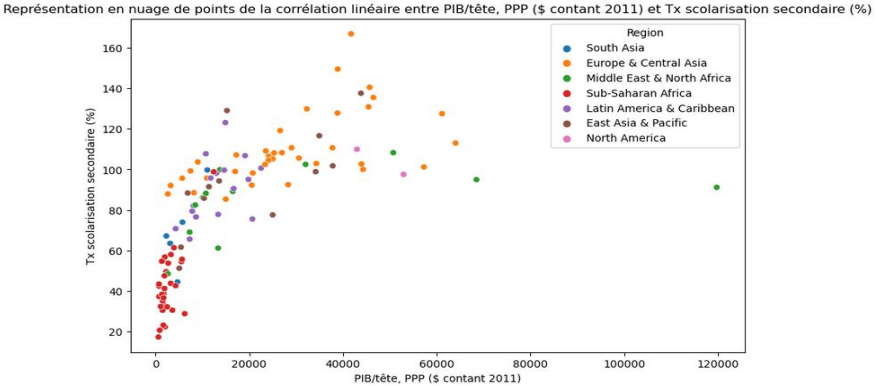


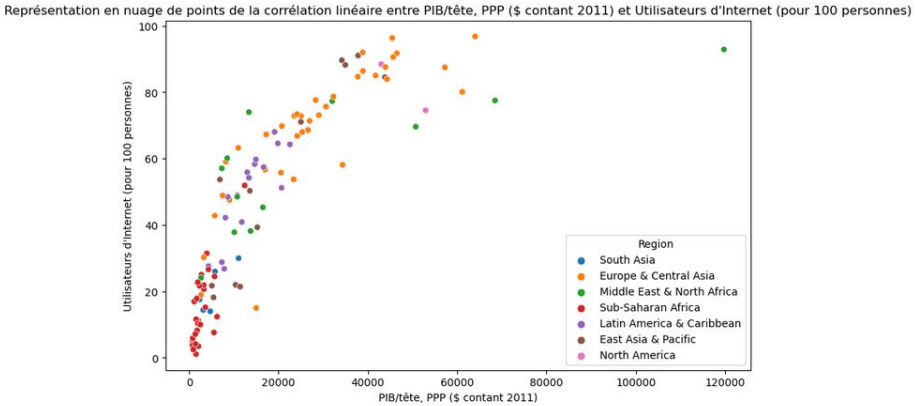



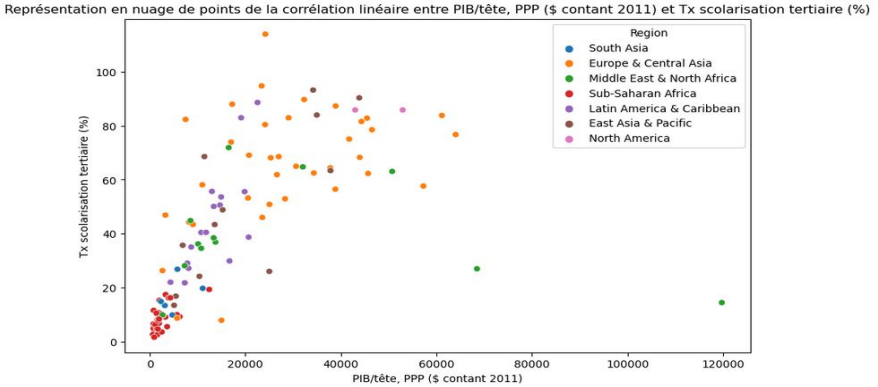
# 3. Analyse du jeu de données principal

## 3.3. Analyse bivariée

L'analyse bivariée nous a permis de mettre en évidence des **corrélations linéaires positives fortes** entre le PIB/tête, les indicateurs de scolarisation et notre indicateur technologique, voyons cela graphiquement :



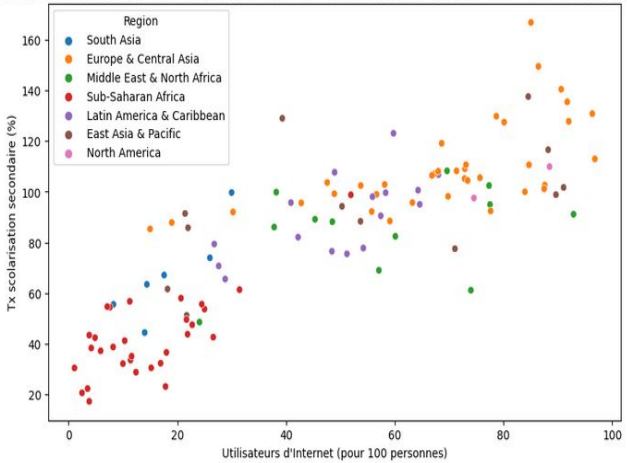



# 3. Analyse du jeu de données principal

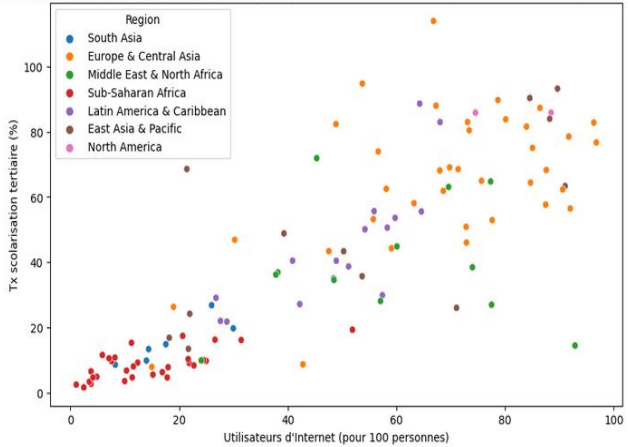
## 3.3. Analyse bivariée

On a pu également observer une **corrélation linéaire positive élevée** entre l'indicateur technologique et les indicateurs de scolarisation.

Représentation en nuage de points de la corrélation linéaire entre Utilisateurs d'Internet (pour 100 personnes) et Tx scolarisation secondaire (%)



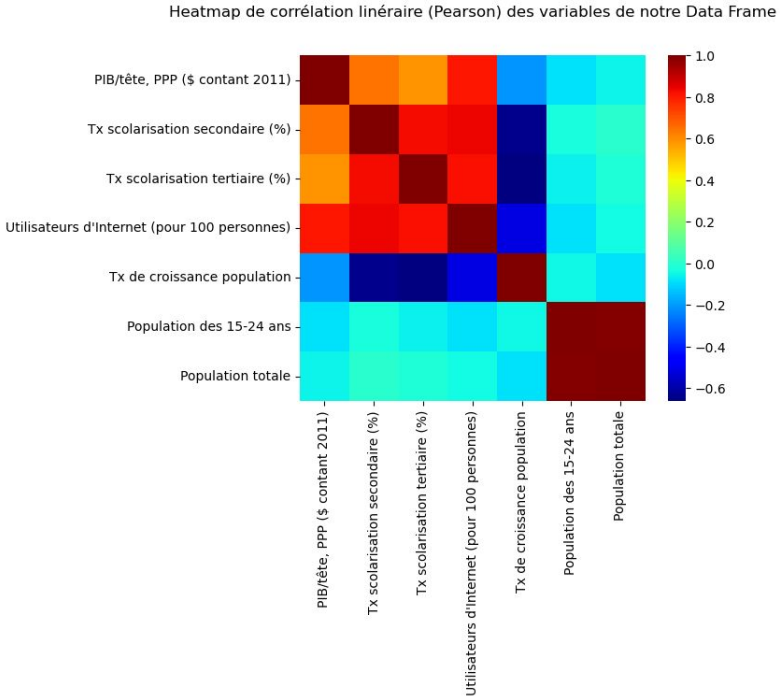

Représentation en nuage de points de la corrélation linéaire entre Utilisateurs d'Internet (pour 100 personnes) et Tx scolarisation tertiaire (%)



# 3. Analyse du jeu de données principal

## 3.3. Analyse bivariée

La heatmap des coefficients de corrélation permet d'illustrer l'ensemble :



# 4. Agrégation de variables et création des scores

## 4.1. Agrégation de variables

Nous souhaitons aboutir à un classement de pays en fonction d'un **score synthétique**, lui-même calculé à partir de **scores intermédiaires**. Pour la création de ces derniers, nous avons fait le choix de **transformer** certaines de nos **variables** :

- variable 'scolarisation' : moyenne des taux de scolarisation secondaire et tertiaire, en donnant un poids plus important au tertiaire, étant donnée le public cible d'Academy;
- Nous avons vu que les indicateurs de populations contenaient des valeurs extrêmes, notamment sur les maximum, en raison de la présence de la Chine et de l'Inde, pays tous deux très peuplés. pour corriger cet "effet taille", nous choisissons de créer une variable nommée 'population concernée' : ('population des 15-24 ans' / 'population totale') x 100

# 4. Agrégation de variables et création des scores

## 4.2. Création de scores intermédiaires

Nous créons **4 scores intermédiaires**, un pour chacun des thème de nos indicateurs. Nous les souhaitons **compris entre 0 et 1** et optons pour la formule suivante pour chaque pays :

- Score démographique = ‘population concernée’ / ‘valeur la plus élevée de ‘population concernée’’
- Score économique = ‘PIB/tête’ / ‘valeur la plus élevée du ‘PIB/tête’’
- Score technologique = ‘Nb Utilisateurs Internet’ / ‘valeur la plus élevée du ‘Nb Utilisateurs Internet’’
- Score ‘Scolarisation’ = ‘Scolarisation’ / ‘valeur la plus élevée de ‘scolarisation’’

# 4. Agrégation de variables et création des scores

## 4.3. Fonction de score final et classement

Le **score final** consiste en une **moyenne pondérée des scores intermédiaires**. Nous avons opté pour la création d'une fonction laissant libre choix aux décideurs d'Academy de donner plus ou moins de poids aux différents coefficient de pondération : scolaire, technologique, économique et enfin démographique.

Notre interprétation de la problématique de l'entreprise nous a conduit à **"privilégier" les indicateurs de scolarisation et technologique**, vis-à-vis des indicateurs économiques et démographiques.

Cela s'est traduit par la formule suivante :

**Score final = (0.35 x Score Scolarisation) + (0.35 x Score Technologique) + (0.15 x Score Economique) + (0.15 x Score Démographique)**

# 4. Agrégation de variables et création des scores

## 4.3. Fonction de score final et classement

Nous aboutissons au **classement** suivant :

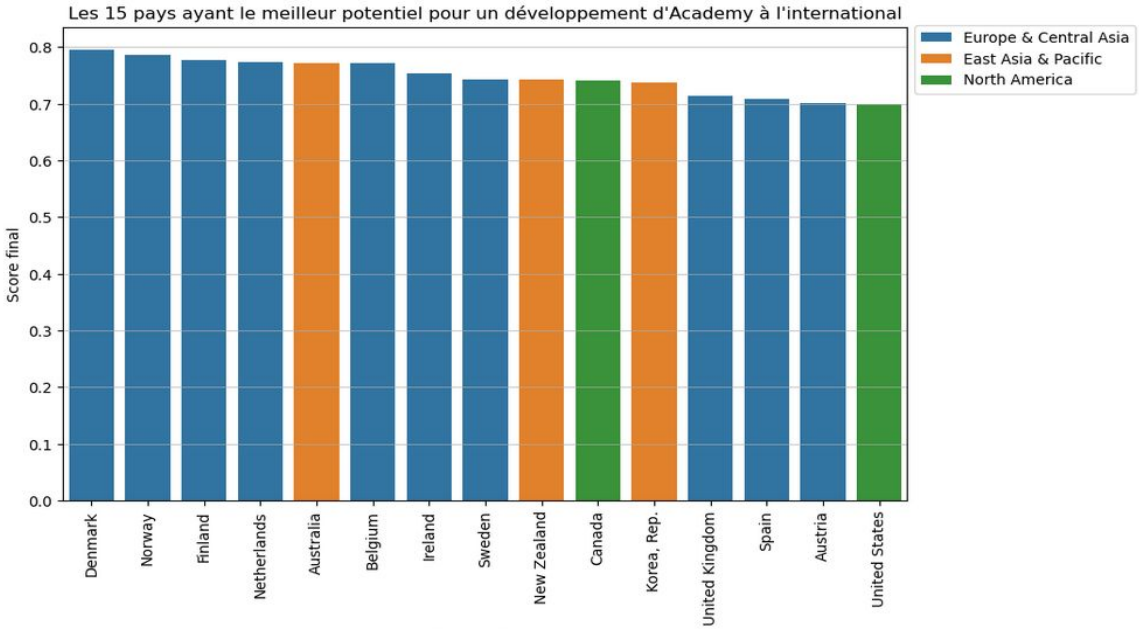
| | Country Name | Region | SCORE_POP | SCORE_SCO | SCORE_TECHNO | SCORE_ECO | SCORE_SYNT |
|---|---|---|---|---|---|---|---|
| **29** | Denmark | Europe & Central Asia | 0.48 | 0.91 | 1.00 | 0.38 | 0.80 |
| **78** | Norway | Europe & Central Asia | 0.48 | 0.81 | 1.00 | 0.53 | 0.79 |
| **36** | Finland | Europe & Central Asia | 0.45 | 1.00 | 0.89 | 0.32 | 0.78 |
| **74** | Netherlands | Europe & Central Asia | 0.45 | 0.90 | 0.95 | 0.39 | 0.77 |
| **6** | Australia | East Asia & Pacific | 0.47 | 0.97 | 0.87 | 0.37 | 0.77 |
| **10** | Belgium | Europe & Central Asia | 0.42 | 1.00 | 0.88 | 0.35 | 0.77 |
| **49** | Ireland | Europe & Central Asia | 0.48 | 0.90 | 0.83 | 0.51 | 0.75 |
| **98** | Sweden | Europe & Central Asia | 0.44 | 0.83 | 0.94 | 0.38 | 0.74 |
| **75** | New Zealand | East Asia & Pacific | 0.51 | 0.86 | 0.91 | 0.29 | 0.74 |
| **17** | Canada | North America | 0.47 | 0.85 | 0.91 | 0.36 | 0.74 |
| **56** | Korea, Rep. | East Asia & Pacific | 0.48 | 0.85 | 0.93 | 0.29 | 0.74 |
| **109** | United Kingdom | Europe & Central Asia | 0.45 | 0.76 | 0.95 | 0.32 | 0.71 |
| **95** | Spain | Europe & Central Asia | 0.35 | 0.94 | 0.81 | 0.27 | 0.71 |
| **7** | Austria | Europe & Central Asia | 0.43 | 0.79 | 0.87 | 0.37 | 0.70 |
| **110** | United States | North America | 0.54 | 0.81 | 0.77 | 0.44 | 0.70 |

# 4. Agrégation de variables et création des scores

## 4.3. Fonction de score final et classement

Graphiquement cela donne :

Le jeu de données de la Banque Mondiale nous a donc bien permis d'**identifier** et de **classer** différents **pays**, en sélectionnant les plus **pertinents pour un déploiement d'Academy à l'international.** L'analyse de ces données nous montre que les 5 pays à cibler en priorité seraient : le **Danemark, la Norvège, la Finlande, les Pays-Bas** et enfin **l'Australie.**



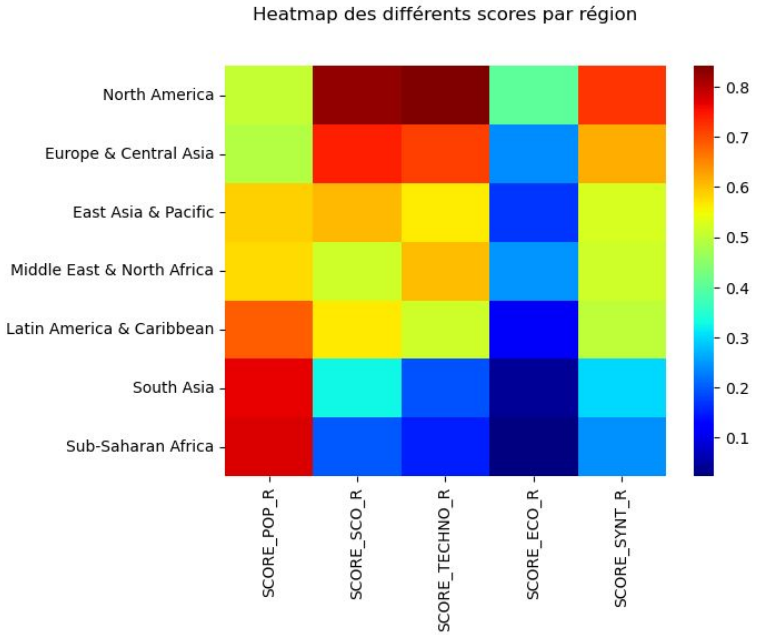
# 4. Agrégation de variables et création des scores

## 4.3. Fonction de score final et classement

Nous pouvons également illustrer la répartition des scores en fonction des régions/blocs continentaux :

**Ce graphique nous permet de voir que c'est la région Nord Américaine qui obtient le score global le plus élevé, suivie par l'Europe et l'Asie Centrale et enfin par l'Asie de l'Est et le Pacifique à égalité avec le Moyen-Orient et l'Afrique du Nord.**

Cela peut sembler étrange que ce soit l'Amérique du Nord qui atteigne le score le plus élevé par région, alors qu'aucun pays de ce bloc continental n'est dans le top 5 du classement pays, mais cela s'explique probablement par le fait que cette région ne comprend que deux pays : les USA et le Canada, malgré tout assez bien positionnés dans notre classement. Les autres régions comportent un plus grand nombre de pays, avec une plus grande diversité, ce qui impacte leur moyenne et tend à la faire baisser.



# Conclusion

Nous avons donc pu établir un classement des pays ayant le meilleur potentiel pour un déploiement d'Academy à l'international, en guise de conclusion et de piste pour une éventuelle suite de ce travail, nous avons esquissé une **analyse de projection**, à partir des 10 pays les mieux classés dans notre étude.

Nous avons en effet que le jeu de données initial portait sur une période allant jusqu'à 2100 et contenait donc des indicateurs de projection. Nous en avons sélectionné un et l'avons appliqué à nos 10 pays.

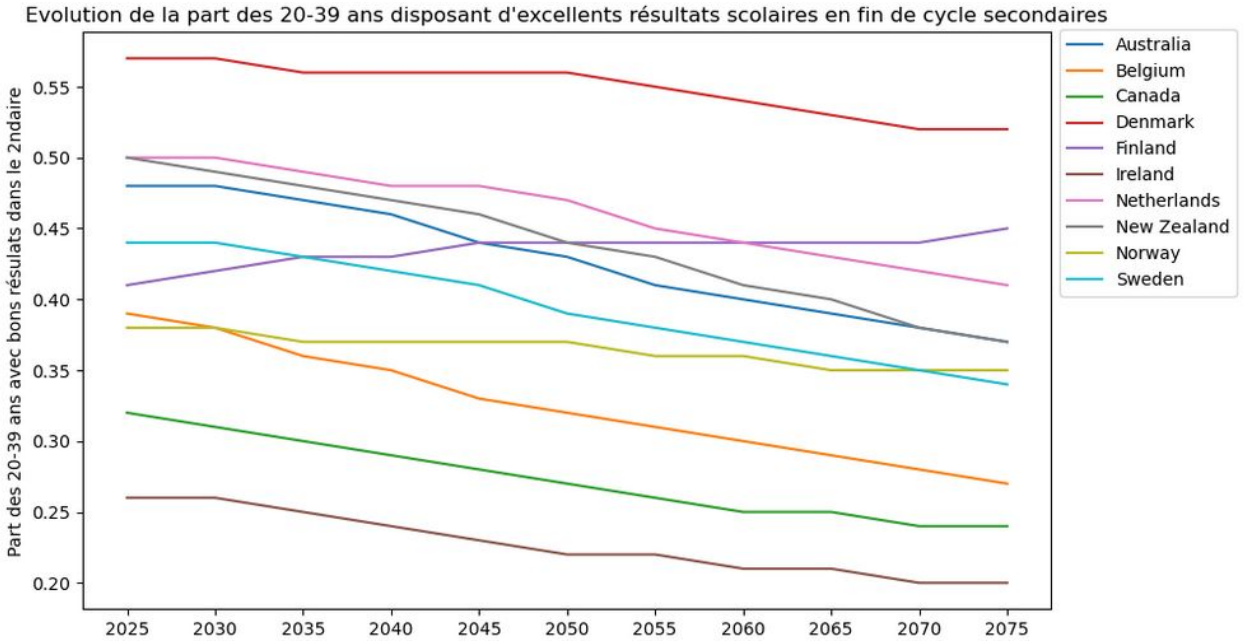
L'indicateur choisi est le suivant : Wittgenstein Projection: Percentage of the population age 20-39 by highest level of educational attainment. Upper Secondary. Total

Définition : Il s'agit de la part de la population, des 20-39 ans, qui a terminé un cursus secondaire (niveau lycée) ou un non terminé des études post-bac, avec les meilleurs résultats. Les projections sont réalisées à partir de données récoltées par des enquêtes et sondage réalisés autour de 2010 et par le "Medium Shared Socioeconomic Pathways (SSP2) projection model".

# Conclusion

Cela nous donne l’évolution suivante, à partir de 2025 :

On note ici que la **Finlande** est le seul pays à avoir un indicateur en progression sur la période considérée. Le **Danemark**, bien que l'évolution de l'indicateur soit en baisse, se positionne particulièrement haut. Cela sera peut-être à prendre en compte pour la sélection finale des pays les plus pertinents pour Academy.



# Conclusion

Comme indiqué plus haut, cette analyse nous a permis de sélectionner les pays les plus à même d'avoir un potentiel de clients intéressant pour Academy. On a pu affiner brièvement ce classement avec une analyse prospective qui, dans notre top 3 ferait peut-être passer la Finlande devant la Norvège.

Ce travail a néanmoins quelques **limites** que nous exposons ici :

- Le choix des indicateurs retenus est subjectif;
- Les dernières données utilisables, pour les indicateurs choisis, datent de 2015 et 2016. Cela qui commence à être assez ancien;
- Notre choix d'imputation de certaines données manquantes, en imputant la dernière valeur disponible, peut se discuter;
- Nous aurions pu chercher des indicateurs hors de ce jeu de données.

Des pistes d'**améliorations** peuvent s'envisager telles que :

- l'inclusion d'un indicateur portant sur la stabilité politique des pays;
- l'inclusion d'un indicateur économique différent, éclairant plus sur le niveau de vie, tel que l'IDH;
- La recherche de données plus récentes;
- Une méthode d'imputation des données manquantes différentes, telle qu'une estimation à partir des données historiques, par exemple.

# Annexes

Version de Python utilisée : 3.9.18

Librairies utilisées :

- pandas, version : 2.1.1
- numpy, version : 1.26.1
- matplolib, version : 3.8.0
- seaborn, version : 0.13.0
- missingno, version : 0.5.2

Source jeux de données : https://datacatalog.worldbank.org/dataset/education-statistics

Création d'un environnement virtuel via le terminal de commande Anaconda