Projet 2 : Analysez des données de systèmes éducatifs

Matthieu Gschwend

academy

Sommaire

1- Présentation de la problématique

Notre entreprise ***academy*** propose des contenus de formation en ligne pour un public de niveau lycée et université.
A l'aide des données de la ***Banque mondiale***, l'objectif final pour l'entreprise serait de pouvoir répondre aux questions 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 ?

Dans un premier temps notre rôle en tant que Data-Scientist est de d'effectuer une ***analyse exploratoire*** afin d'évaluer si oui ou non ce jeu de données peut informer les décisions d'ouverture vers de nouveaux pays"

Pour ce faire nous allons étudier le contenu des jeux de données, voir s'il est possible d'avoir des indicateurs à la fois représentatifs et pertinents pour répondre à cette problématique.

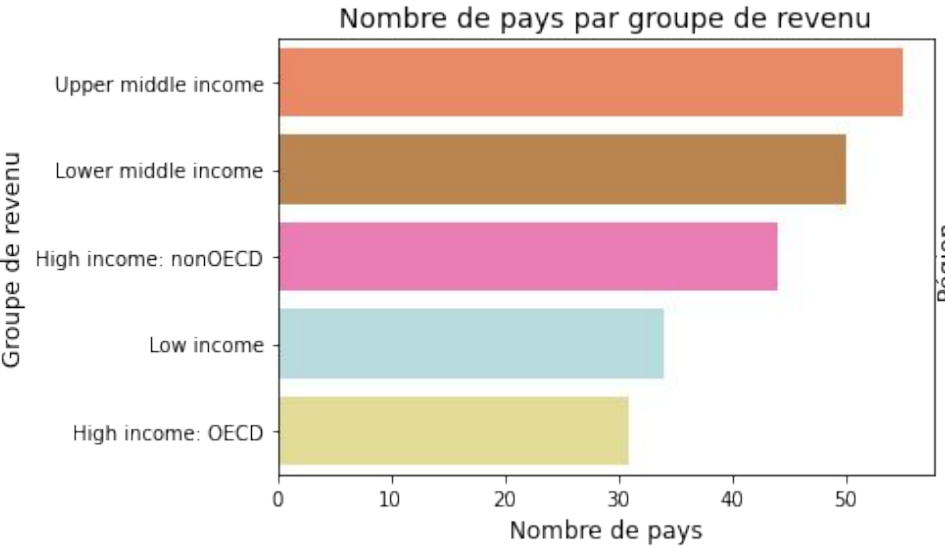
academy

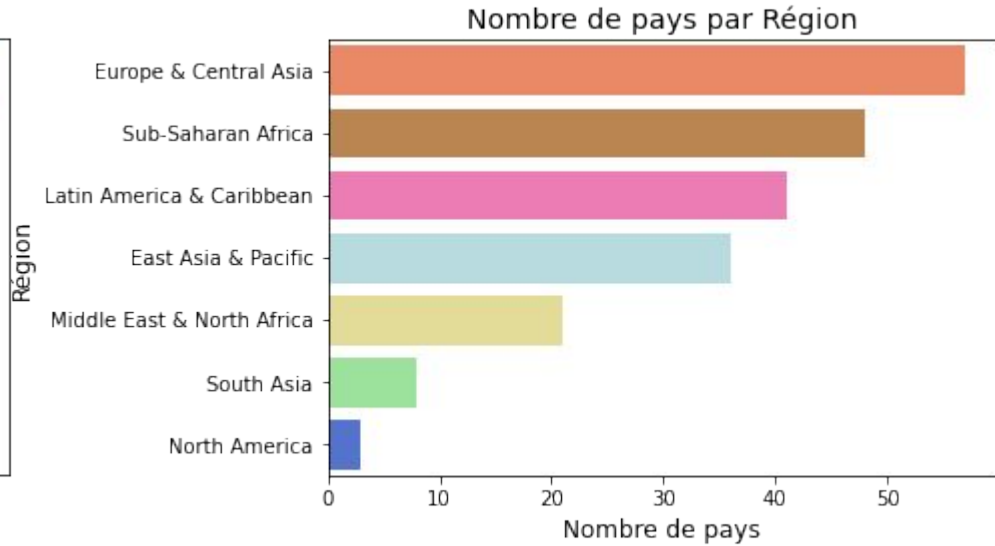
2- Analyse des jeux de données : Aperçu

EdStatsCountry	Informations générales sur les pays avec des données qualitatives en majorité	• 241 Pays dans 7 régions du monde. • Les dates des dernières mises à jour • Pas de ligne dupliquée • Il manque 30.5% des valeurs • 241 lignes et 32 colonnes
EdStatsContry_series	Provenance des indicateurs selon les pays	• Une colonne vide, sinon complet • 613 lignes et 4 colonnes • 100% des codes des pays sont présents dans EdStatsContry.
EdStatsFootNote	Dates d'actualisation des indicateurs. Description de la provenance (1970-2050)	• 1558 indicateurs • 239 Pays • 643638 lignes et 5 colonnes • Une colonne vide, sinon complet
EdStatsSeries	Thème des différents indicateurs, descriptions longues et sources	• 37 thèmes disponibles • 3665 indicateurs • Il manque 71% des valeurs • 3665 lignes et 21 colonnes
EdStatsData	Valeurs quantitatives des indicateurs par pays entre (1970 - 2100)	• 242 Pays • 3665 indicateurs • 886930 lignes et 70 colonnes • Il manque 86% des valeurs

academy

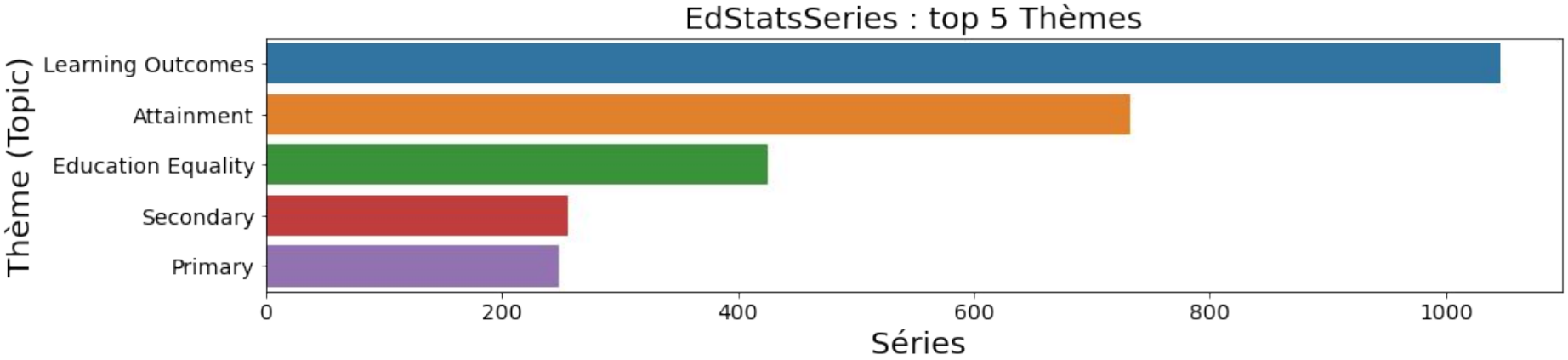
1.1 Visualisation EdstatsCountry

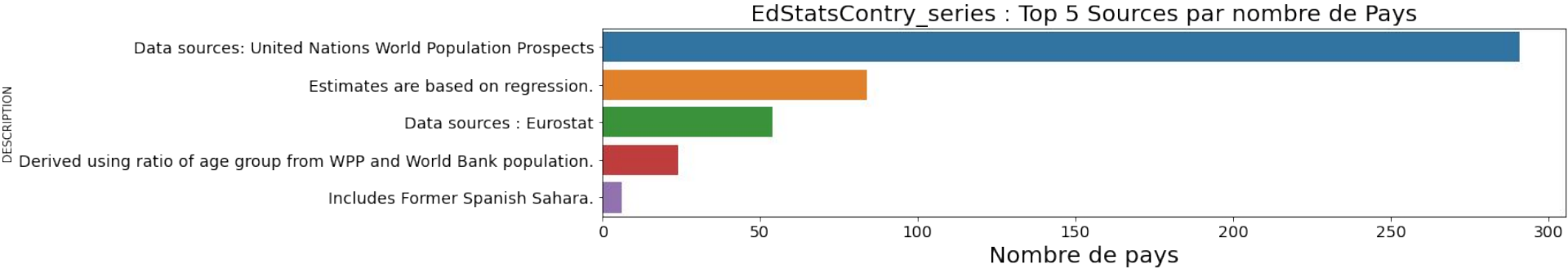




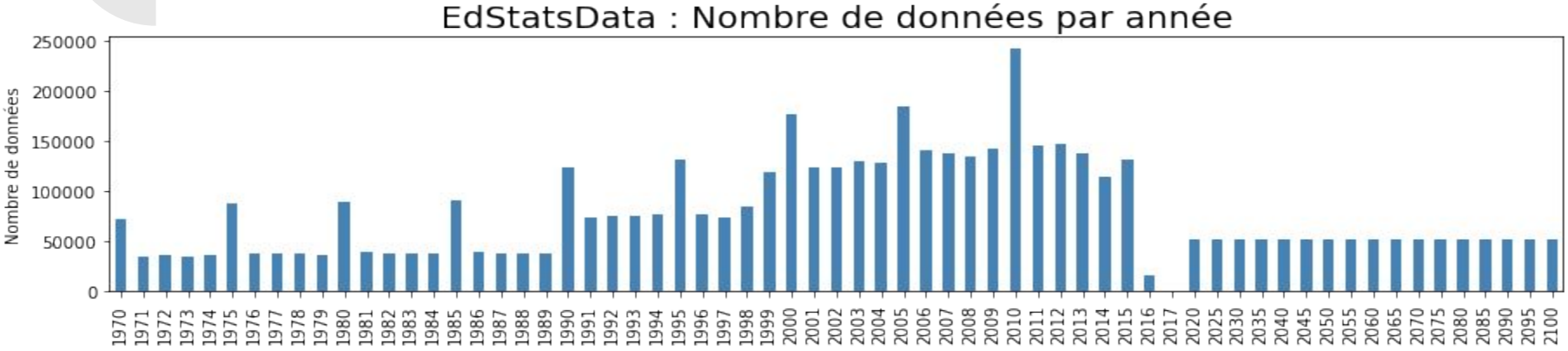
academy

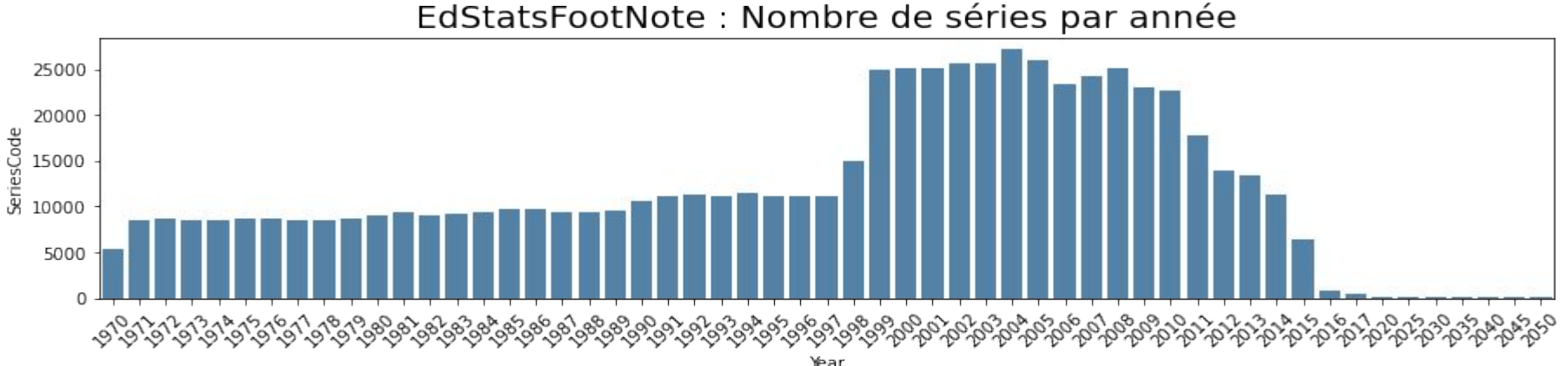
1.2 Visualisation : provenance et thème





1.3 Visualisation : Nombre de données par an

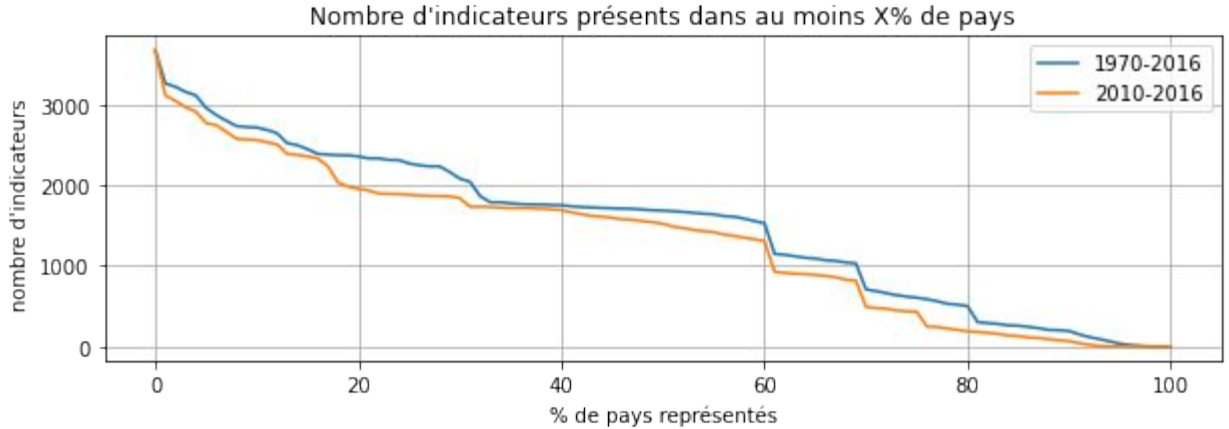




2 Etude et sélection des indicateurs

Démarche :

- Le fichier EdStatsData contient les données quantitatives utiles pour notre étude.
- Ce fichier contient énormément d'indicateurs mais :
  - Ils ne sont pas tous représentatifs (peu de pays sont renseignés).
  - La date de l'actualisation est parfois ancienne donc non pertinent
- Réduction du jeu de données :
  - Garder uniquement les dernières valeurs disponibles entre 2010 et 2016
  - Garder uniquement les indicateurs présents dans plus de la moitié des pays



2.1 Construction du Data-Frame pour la sélection des indicateurs

A partir de **EdStatsData** nous effectuons plusieurs modifications :

- Seulement la dernière valeur de entre 2010 et 2016
- Pour chaque pays on ajoute sa région à partir de **EdStatsCountry**
  - Les “pays” qui ne sont pas associés à des régions seront supprimés car ce ne sont pas des pays mais des zones géographiques ou des groupes de revenu.
- Pour chaque indicateur nous associons sa définition à partir de **EdStatsSeries**.
  - 98.5 % des indicateurs présents dans EdStatsData le sont également dans EdStatsSeries. Donc nous aurons accès à l'énorme majorité des définitions
- Les indicateur n’étant pas présents dans plus de la moitié des pays sont supprimés

A la suite de cette opération notre Data-Frame contient : **214 Pays** et **529 indicateurs**

academy

2.2 A la recherche des indicateurs pertinents

Pour sélectionner les indicateurs pertinents nous utilisons une recherche par mot clé présent dans le nom des indicateurs.

Concernant les tranches d'âges nous incluons des populations jeunes car il y a un décalage de parfois 11 ans (2010) avec 2021

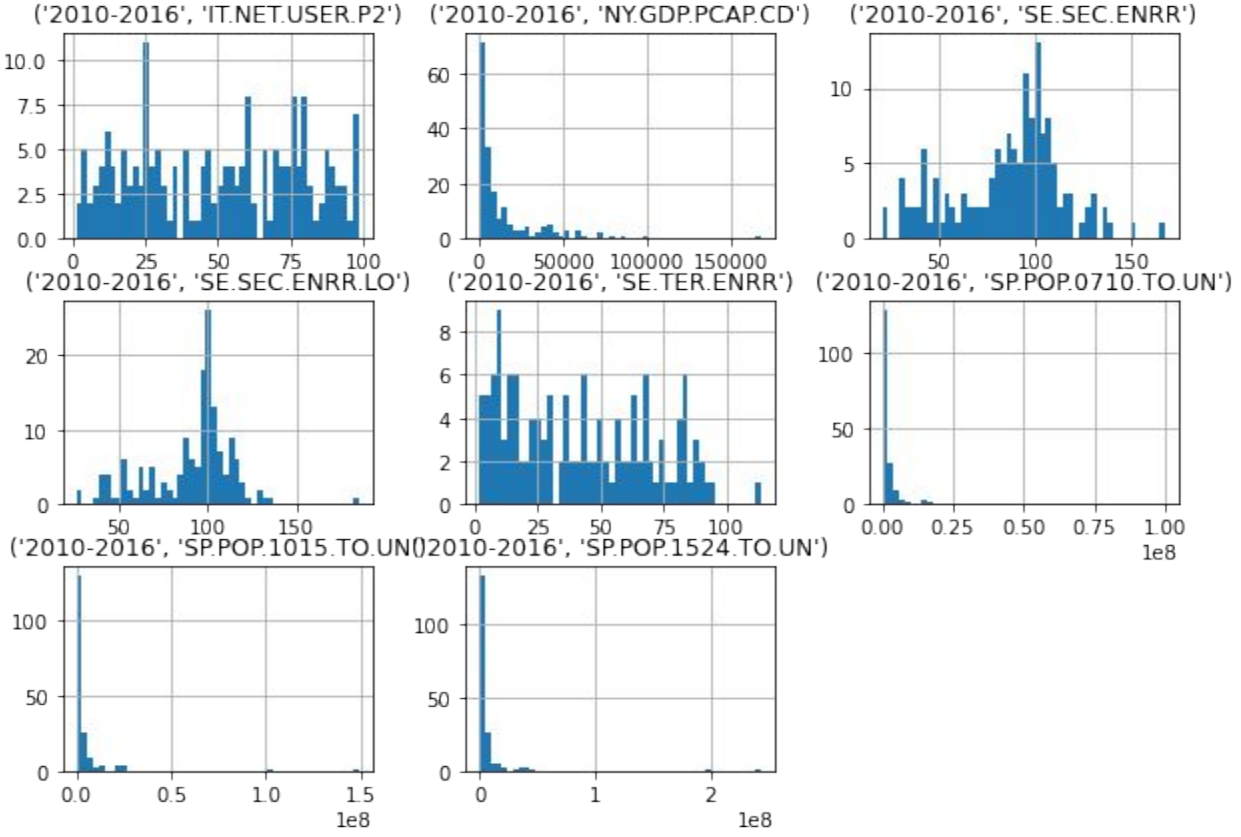
Catégorie		Mot clé	Indicateur retenu
1	Démographique	• 10 • 24	➢ SP.POP.1015.TO.UN (age 10-15) ➢ SP.POP.1524.TO.UN (âge 15-24) ➢ SP.POP.0710.TO.UN (âge 7-10) ➢ kfkfkf
2	Education	• SEC • TER	➢ SE.SEC.ENRR (% étudiant lycée) ➢ SE.SEC.ENRR.LO (% étudient collège) ➢ SE.TER.ENRR (% étudiants faculté)
3	Technologique	• IT	➢ IT.NET.USER.P2 (% utilisateur d'internet)
4	Economique	• NY	➢ NY.GDP.PCAP.CD (PIB par habitant)

2.3 Description des indicateurs retenus

Les valeurs des indicateurs retenus ne sont pas homogènes.

SE.SEC.ERR et SE.SEC.ENRR.LO semblent avoir une distribution normale.

	Indicator Code	moyenne	Ecart type
2010-2016	IT.NET.USER.P2	5.013821e+01	2.844775e+01
	NY.GDP.PCAP.CD	1.434233e+04	2.202263e+04
	SE.SEC.ENRR	8.656679e+01	2.822941e+01
	SE.SEC.ENRR.LO	9.095767e+01	2.423526e+01
	SE.TER.ENRR	4.170360e+01	2.818212e+01
	SP.POP.0710.TO.UN	2.771086e+06	9.354284e+06
	SP.POP.1015.TO.UN	4.056589e+06	1.402581e+07
	SP.POP.1524.TO.UN	6.700539e+06	2.417073e+07



3 Construction des Scores

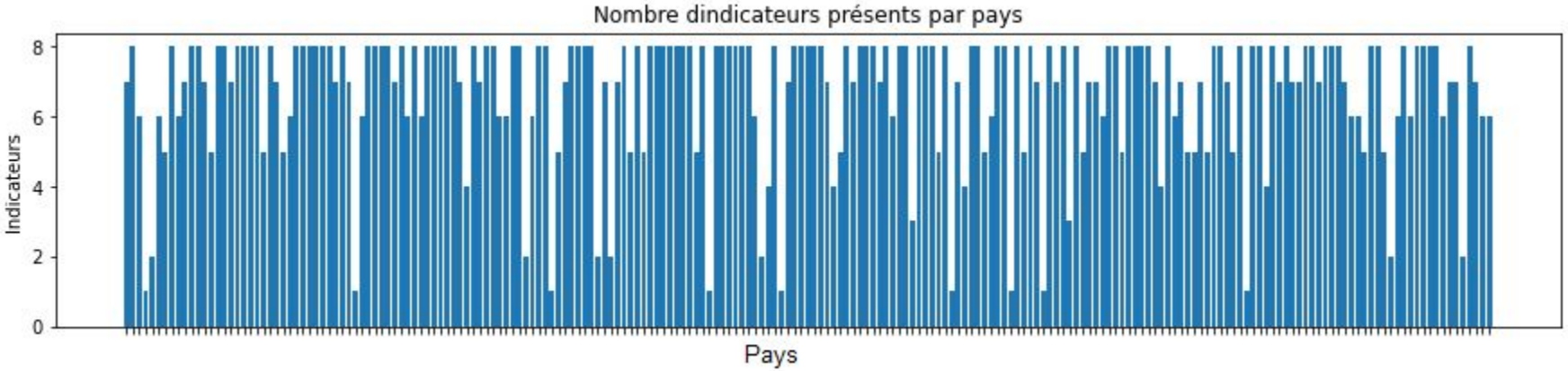
Établir des scores permet de savoir dans quels pays il est le plus judicieux d'investir pour notre entreprise. Cela implique une part de subjectivité : il faut assigner des coefficients "poids" à nos indicateurs.

Etapes :

- ❏ Remplir les **valeurs manquantes**
- ❏ **Homogénéiser** les valeurs
- ❏ Choisir les **coefficients**

3.1 Valeurs manquantes

➢ Tous les pays présents ne disposent pas des 8 indicateurs retenus



Nous conservons uniquement les pays avec au moins 5 indicateurs.
Les valeurs manquantes sont ensuites remplacées par celles de la région associée.
(Pour calculer la valeur d'une région on utilise la médiane des Pays de la région)

3.2 Homogénéisation des valeurs et pondération

Homogénéisation
Les valeurs des colonnes ne sont pas homogènes : on ne compare pas des pourcentages avec un PIB par habitant.

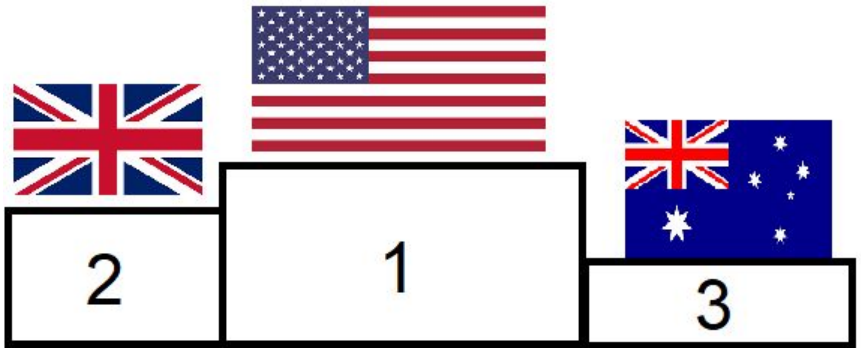
Nous décidons d'utiliser une technique qui "découpe" les valeurs par colonnes en 20 groupes. Associant une valeur de 5,10, 15 ... 95, 100 en fonction du groupe.

Plusieurs méthodes existent mais ici on ne prend pas en compte les valeurs extrêmes.

Catégorie	Indicateur	Coefficient
Technologique	IT.NET.USER.P2	5
Economique	NY.GDP.PCAP.CD	5
Education	SE.SEC.ENRR	2
	SE.SEC.ENRR.LO	1
	SE.TER.ENRR	2
Démographique	SP.POP.0710.TO.UN	2
	SP.POP.1015.TO.UN	2
	SP.POP.1524.TO.UN	1

academy

3.3 Résultats par Pays Région



Country Name	Score
United States	1849.663066
United Kingdom	1830.169256
Australia	1798.923393
Japan	1797.492867
Netherlands	1791.484436
Belgium	1790.811367
Germany	1771.567534
Spain	1765.700105
Korea, Rep.	1760.979822
France	1760.833130

Region	Score Final
North America	1769.050955
Europe & Central Asia	1406.400509
Middle East & North Africa	1216.600011
Latin America & Caribbean	1165.374409
East Asia & Pacific	1160.295315
South Asia	907.153016
Sub-Saharan Africa	692.868076

Conclusion et ouverture

Cette analyse conclut que les données disponibles sur la Banque mondiale sont pertinentes pour répondre à notre problématique sur plusieurs aspect :

- Presque tous les pays du monde y sont présents.
- Nous avons trouvé au moins 8 indicateur pertinents et représentatifs sur de années récentes.

Les scores sont à relativiser car ils ne prennent pas en considération l'émergence récente des technologies sur le continent Africain et Asiatique. Qui vont également subir croissance démographique exponentielle durant les années à venir.

Pour la suite du projet de l'entreprise il sera intéressant de développer des méthodes de prévisions sur les indicateurs : modèle ARIMA, Lissage exponentiel, VAR...