Projet Python sur l'éducation mondiale
Analyse pré-exploratoire

Cyril REGAN

Base de données : http://datatopics.worldbank.org/education/

10 janvier 2020

Table of Contents

Problématique

Academy, start up de EdTech, propose des contenus de formation **en ligne** de **niveau lycée** et **université** et souhaite s'étendre à l'international.

Projet : Réaliser une analyse préparatoire pour répondre aux questions :

- 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 ?

Présentation du jeu de données

Les données présentent dans le fichier "EdStatsData.csv" sont compilées dans un tableau de 886930 lignes $\times$ 70 colonnes.

- Les lignes représentent les 3665 indices pour 242 pays
- Les colonnes sont les années [1970 - 2100] où sont calculées les indices.

Table of Contents

Plusieurs choix de traitement des données :

- **Supprimer les colonnes ou les lignes contenant trop de valeurs manquantes**
- Imputer les données manquantes par :
  - la moyenne de l'indicateur des autres pays
  - les plus proches voisins (géographiquement parlant)

Sélection de la plage d'années

Pas de données dupliquées dans le jeu de données.
La plage d'années [2000 - 2015] semble être optimale pour l'étude.

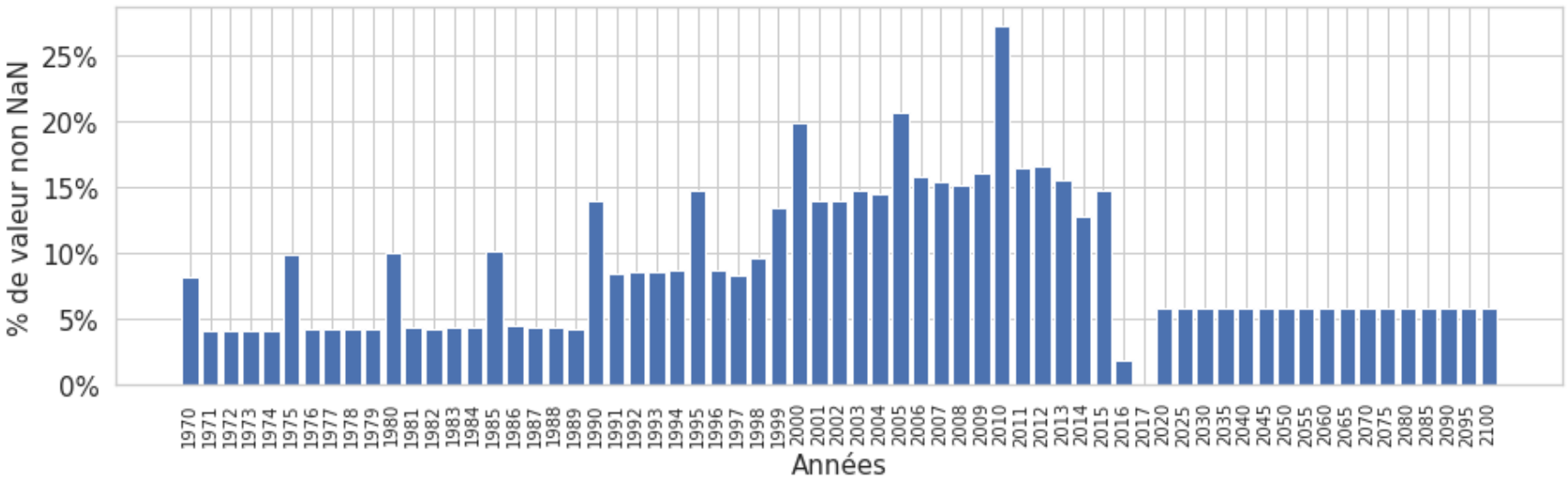


Figure – % d'indices non manquants par année

Sélection des couples pays/indicateur qui contiennent assez de données exploitables

Sur la plage [2000 - 2015], plusieurs type de selection ont été envisagés :

1. On selectionne les indicateurs dont la moyenne des données exploitables sur tous les pays respectent une certaine proportion : $\%_{NaN}$ acceptable

$$\frac{1}{242}\sum_{pays} \%_{NaN}(\text{indicateur}) \leq \%_{NaN} \text{ acceptable} \qquad (1)$$

2. On selectionne les pays dont la moyenne des données exploitable sur les indicateurs restant après la selection 1, qui respectent une certaine proportion : $\%_{NaN}$ acceptable

$$\frac{1}{\text{nbr indic restants}} \sum_{\text{indic restants}} \%_{NaN}(\text{Pays}) \leq \%_{NaN} \text{ acceptable} \tag{2}$$

- Avantage des méthodes 1 et 2 enchainées : l'ensemble des indicateurs retenus est le même pour tous les pays retenus.
- Inconvénient : Elles n'assurent pas un respect du critère de proportion de valeur exploitable **pour chaque ligne.**

Une dernière méthode est d'assurer pour chaque ligne un nombre minimum de valeur exploitable :

3. On selectionne les couples indicateur/pays qui contiennent assez de données exploitables : $\%_{NaN}$ acceptable

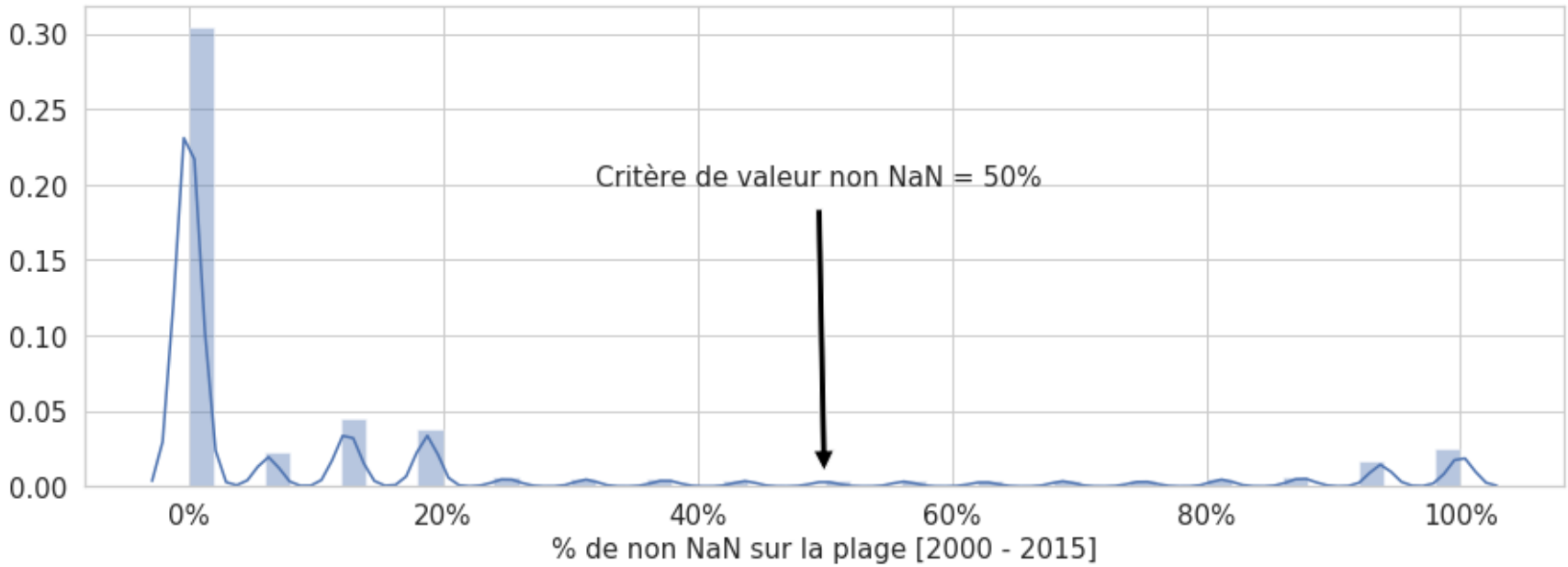


Figure – distribution des pays / % de non NaN sur [2000 - 2015]

On peut choisir un $\%_{NaN}$ acceptable de 50 % sans obtenir un ensemble de couple pays/indicateur vide.

A l'issue de la selection de la méthode 3, il reste 1175 indicateurs et 242 pays.

Dernière étapes de selection : filtrer par chaines de caractère

Les indicateurs contenant ces chaines de caractères seront supprimées :

"primary" "lower secondary" "male" ""gender" ""at market prices" "PPP" "GDP" "GNI" "Mortality rate" "Official entrance age" "Under-age" "Education programmes" "in private institutions" "Percentage of students in upper secondary education enrolled in" "vocational" "general" "Graduates" "private" "public" "per 100,000" "mobility" "mobility" "post-secondary" "in secondary " "Net enrolment rate" "basic" "intermediate" "of the official age" "chool life expectancy" "Theoretical duration" "Total outbound international" "Gross" "Teacher" "Prevalence" "entrance age" "of compulsory" "Duration" "All staff compensation" "expenditure" "Health" "Welfare" "Agriculture" "Engineering" "Humanities" "unspecified fields" "childhood" "Science" "Services" "Social" "Pupil-teacher" "age 25+" "illiterate" "25+" "Out-of-school" "ISCED" "mobile" "25-64" "65+" "youth" "literacy"

Les 10 indicateurs selectionnés sont :

- "Enrolment in tertiary education, all programmes, both sexes (number) "
- "Enrolment in upper secondary education, both sexes (number)"
- "Internet users (per 100 people)"
- "Labor force, total"
- "Population growth (annual %)"
- "Population, total"
- "**Employment**, total (% of total labor force)" : recalculé avec l'indicateur de chômage
- "Labor force with advanced education (% of total)"
- "Population, ages 15-24, total"
- "Personal computers (per 100 people)"

Les pays sont partagés en 2 groupes :

- Les ensembles de pays (regions du monde).
  Les indicateurs ***"Population, ages 15-24, total"*** *et* ***"Personal computers (per 100 people)"*** *ne sont pas analysés par manque de données*
- Les pays uniques

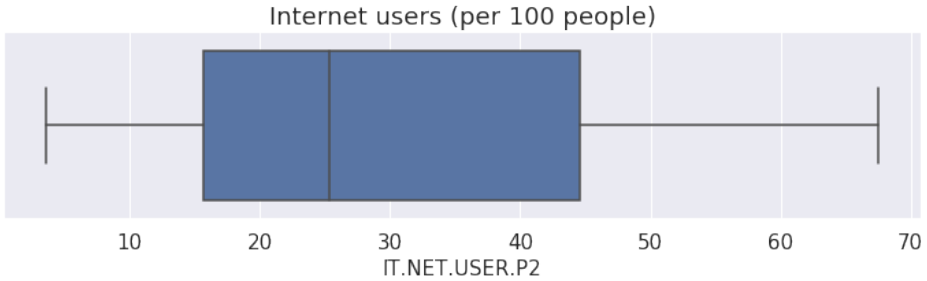
Table of Contents

Valeurs moyennées sur l'ensemble [2000 2015]

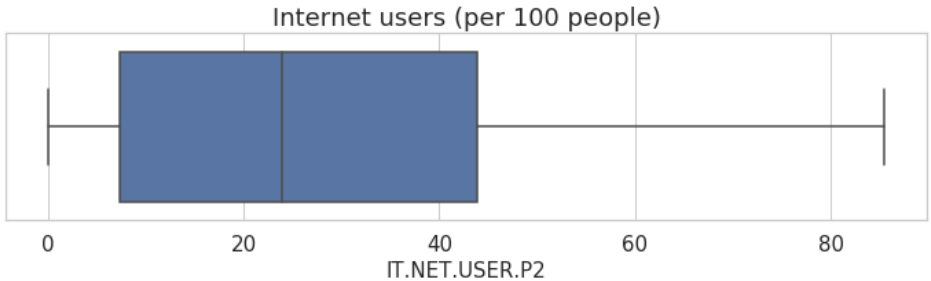
L'indicateur retenu le plus important est celui des utilisateurs d'internet. Les box-plots ou boite à moustaches permettent de représenter :

- une boite de quartiles
- les centiles 5 et 95 par des barres
- les éventuelles valeurs à l'extérieur des centiles extrêmes

Pour les ensembles de pays :



Pour les pays uniques :



Scatter plot des ensembles de pays moyennés sur [2000 - 2015]

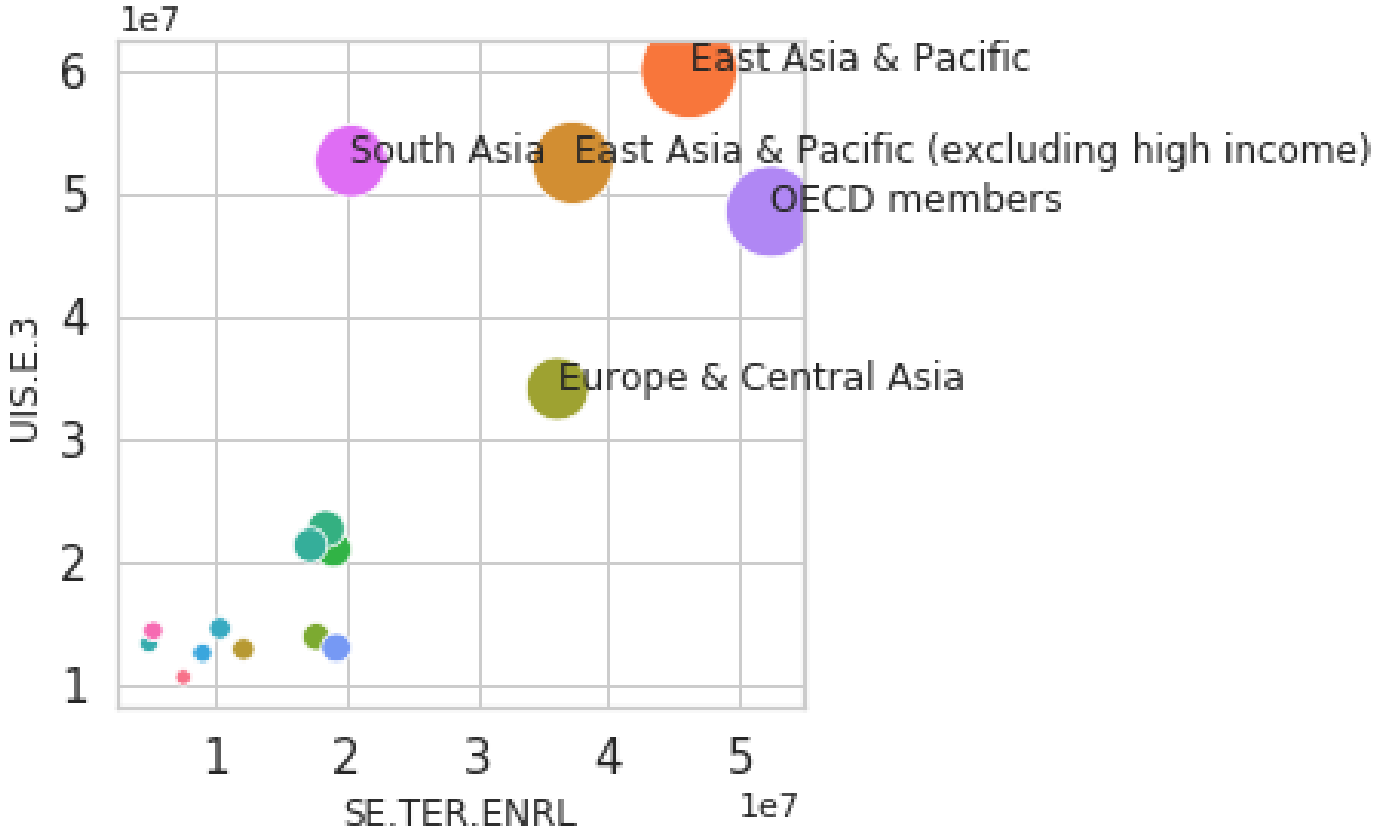


Figure – 'Enrolment in tertiary education, all programmes, both sexes (number)' ('SE.TER.ENRL') avec 'Enrolment in upper secondary education, both sexes (number)' ('UIS.E.3')

Scatter plot des pays uniques moyennés sur [2000 - 2015]

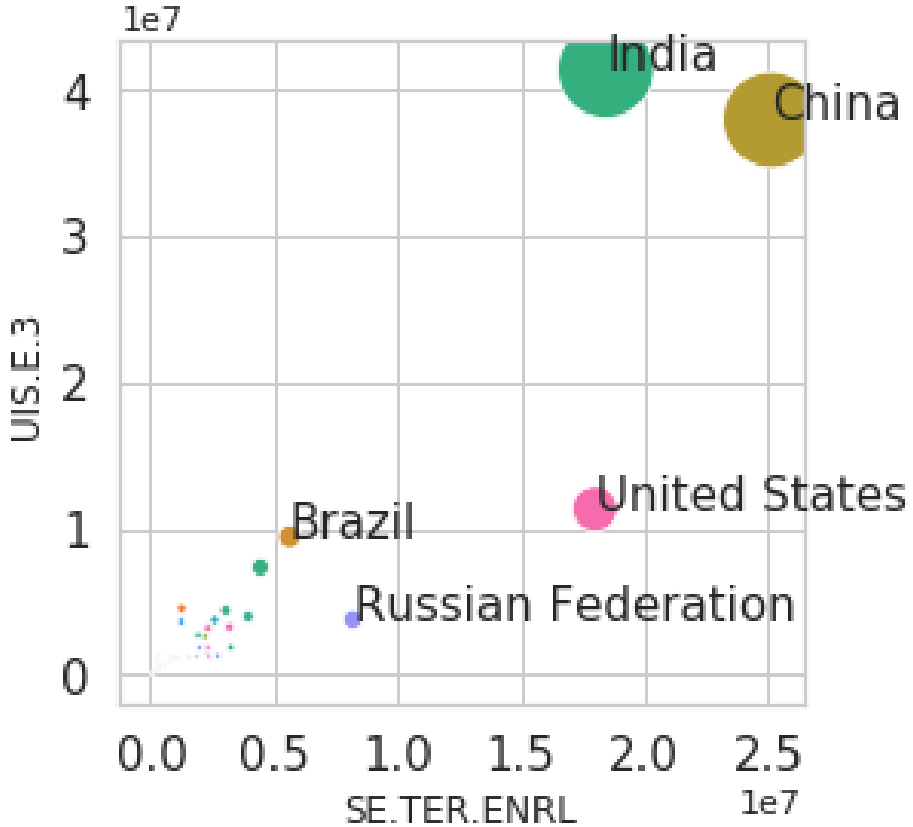


Figure – 'Enrolment in tertiary education, all programmes, both sexes (number)' : ('SE.TER.ENRL') avec 'Enrolment in upper secondary education, both sexes (number)' ('UIS.E.3')

Evolution des indicateurs

- Première étape :

  Imputation temporelle des valeurs sur [2000-2015] par interpolation de type spline de degré 1.

Figure – Exemple de splin de degré 1 sur 4 valeurs

- Seconde étape : Normalisation de tous les indicateurs par la valeur maximale.

- Troisième étape : Choix du poid de chaque indicateur pour créer un indicateur global.
    - Une pondération équipotentielle

Les régions ***Nord amérique***, ***les membres de l'OCDE*** et l'***Asie EST - Pacifique*** se détachent pour la pondération équipotentielle.

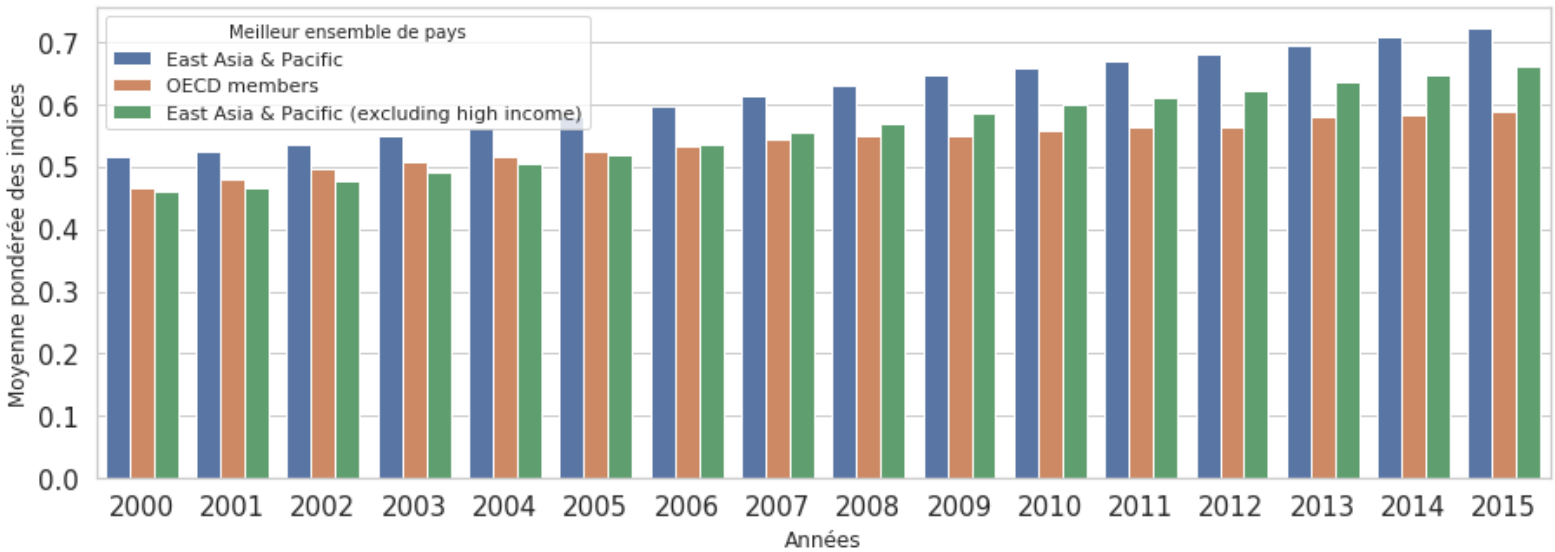


Figure – Trois meilleurs ensembles de pays pour l'indicateur global equipondéré

La ***Chine***, l'***Inde*** et les ***Etats Unis*** se détachent pour la pondération équipotentielle.

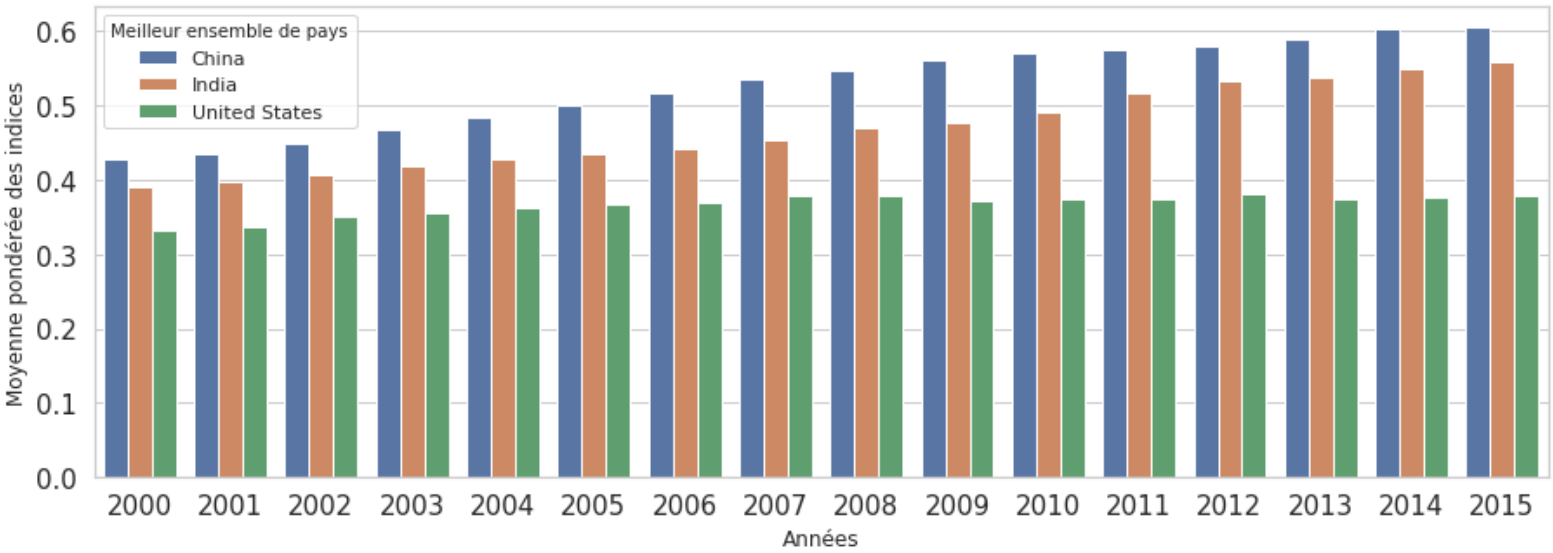


Figure – Trois meilleurs pays pour l'indicateur global equipondéré

- Des poids différents

	Indicator Name	Indicator Code	weight
0	Enrolment in tertiary education, all programme...	SE.TER.ENRL	0.172414
1	Enrolment in upper secondary education, both s...	UIS.E.3	0.172414
2	Internet users (per 100 people)	IT.NET.USER.P2	0.344828
3	Labor force, total	SL.TLF.TOTL.IN	0.034483
4	Population growth (annual %)	SP.POP.GROW	0.034483
5	Population, total	SP.POP.TOTL	0.034483
6	Employment, total (% of total labor force)	SL.EM.TOTL.ZS	0.034483
7	Labor force with advanced education (% of total)	SL.TLF.ADVN.ZS	0.172414

Figure – Pondération des indicateurs pour les ensembles des pays

Les régions ***Nord américaine***, les ***membres de l'OCDE*** et l'***Asie EST - Pacifique*** sont identifiées sur les années [2013 - 2015].

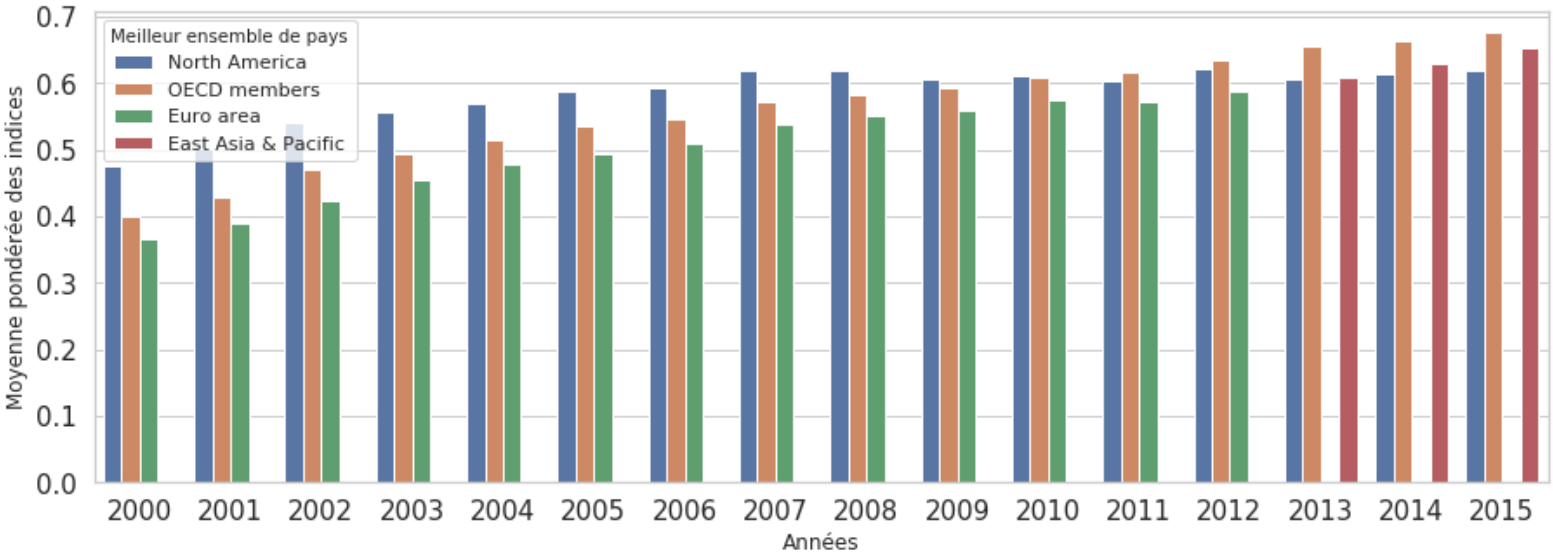


Figure – Trois meilleurs ensembles de pays pour l'indicateur global

	Indicator Name	Indicator Code	weight
0	Enrolment in upper secondary education, both s...	UIS.E.3	0.147059
1	Internet users (per 100 people)	IT.NET.USER.P2	0.294118
2	Labor force, total	SL.TLF.TOTL.IN	0.029412
3	Population growth (annual %)	SP.POP.GROW	0.029412
4	Population, ages 15-24, total	SP.POP.1524.TO.UN	0.073529
5	Population, total	SP.POP.TOTL	0.029412
6	Employment, total (% of total labor force)	SL.EM.TOTL.ZS	0.029412
7	Enrolment in tertiary education, all programme...	SE.TER.ENRL	0.147059
8	Personal computers (per 100 people)	IT.CMP.PCMP.P2	0.073529
9	Labor force with advanced education (% of total)	SL.TLF.ADVN.ZS	0.147059

Figure – Pondération des indicateurs pour les pays uniques

La ***Chine***, les ***Etats Unis*** et la ***Suisse*** sont identifiées sur les années [2009 - 2015].

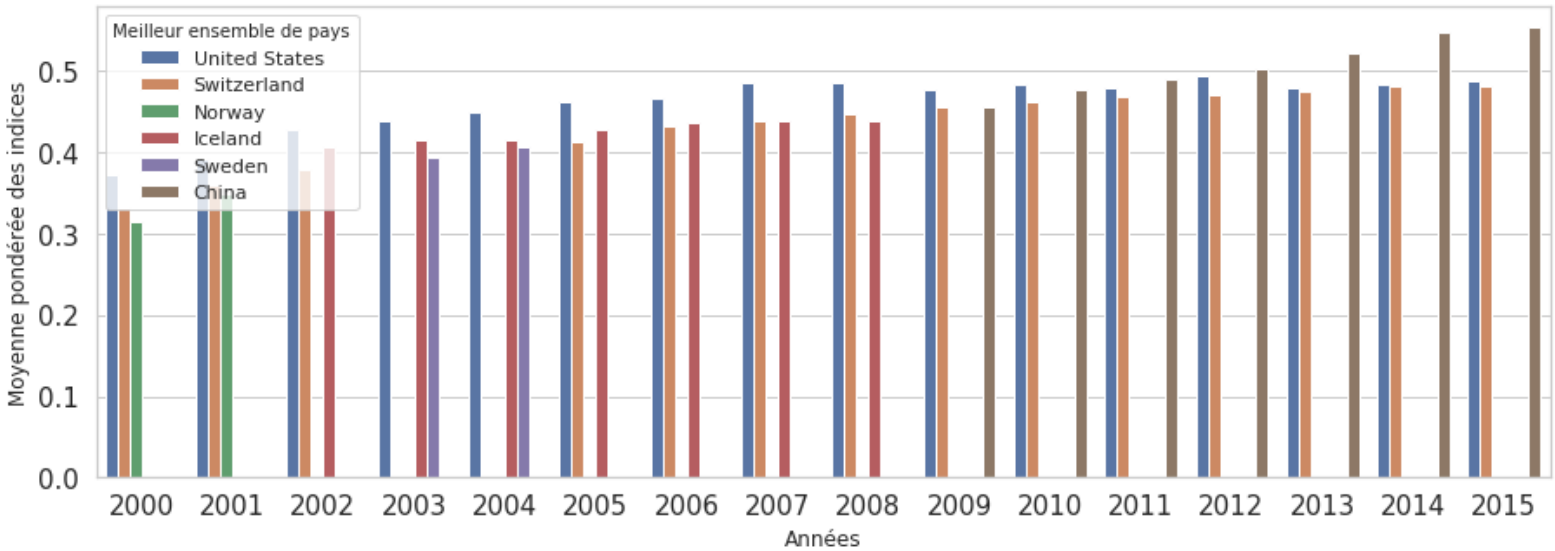


Figure – Trois meilleurs pays pour l'indicateur global

Table of Contents

Conclusion

- Jeu de donnée suffisament complet.
- Régions du monde identifiées : [Nord américaine, les membres de l'OCDE et d'Asie EST - Pacifique].
- Pays identifiés : [La Chine, les Etats Unis et la Suisse].

Perspectives

Faire la même analyse avec :

- un groupement de pays uniques en région pour étudier la distributions de chaque indicateurs dans ces régions,
- un jeu de donnée étendu aux années actuelles,
- d'autres indicateurs (PIB, dépenses dans l'éducation...).

Merci de votre attention