OPENCLASSROOMS

P2 : Analyse des données de systèmes éducatifs chez l'entreprise Academy

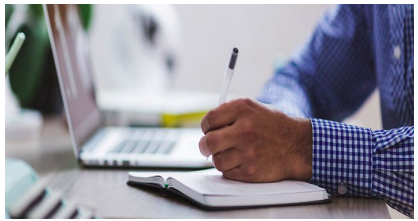
Juan David Briceno Guerrero
Août 2021

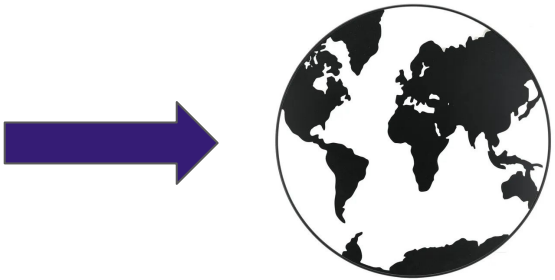
Contenu

1. Problématique : Besoin de l’entreprise

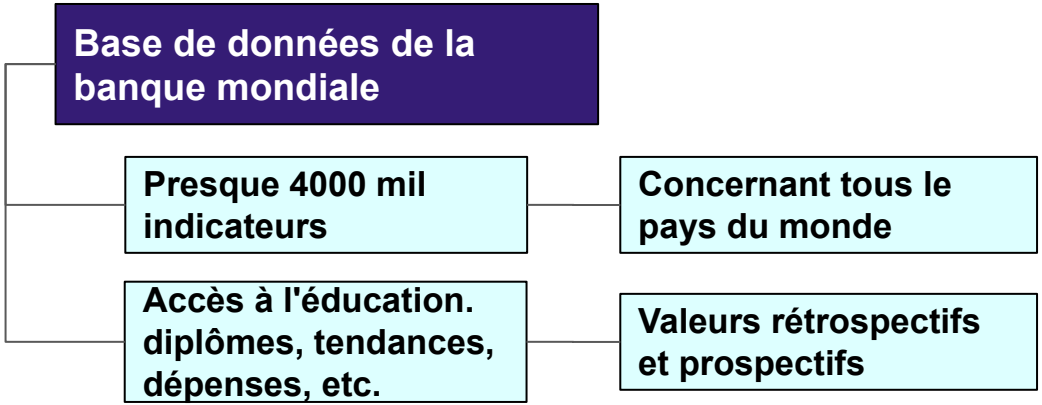
Academy

- Entreprise des formations en ligne visant à faire une expansion à l'international.





1. Problématique : Objectif de l'étude



Est-ce que les données permettent d'informer le projet d'expansion?

Quels sont les Pays les plus attractifs pour attirer des clients à les services d'Academy?

2. Jeux des données : Caractéristiques

edStatsCountry	**Information reliée aux pays** (nom, devise,SNA, statut de la dette extérieure, etc.)	Dimensions : 241 lignes, et 32 colonnes. Type: CSV Taille : 139,5 ko
edStatsCountry-Series	**Information associée aux indicateurs** et leur description (signification).	Dimensions : 613 lignes, et 4 colonnes. Type: CSV Taille : 49,0 ko
edStatsData	I**nformation relative aux indicateurs, leurs valeurs** par pays et par des differents années.	Dimensions : 886930 lignes, et 70 colonnes. Type: CSV Taille : 326,4 Mo

2. Jeux des données : Analyse préliminaire

	Colonnes pertinentes	Remarques	Importance
edStatsCountry	**External debt Reporting status, Latest population census, Currency units, SNA prices valuation.**	Valeurs manquants pour plusieurs de colonnes.	Incidence faible.
edStatsCountry-Series	**CountryCode, SeriesCode**	Relevant pour distinguer pays, continents, et groupes...	Incidence faible.
edStatsData	**Country Name, Indicator Name, séries temporelles appelées par année (1970, 1971,...)**	Indicateurs vidés pour quelques pays.	**Très importante (Il faut remplir, nettoyer, et filtrer ces données).**

2. Jeux des données : Analyse préliminaire

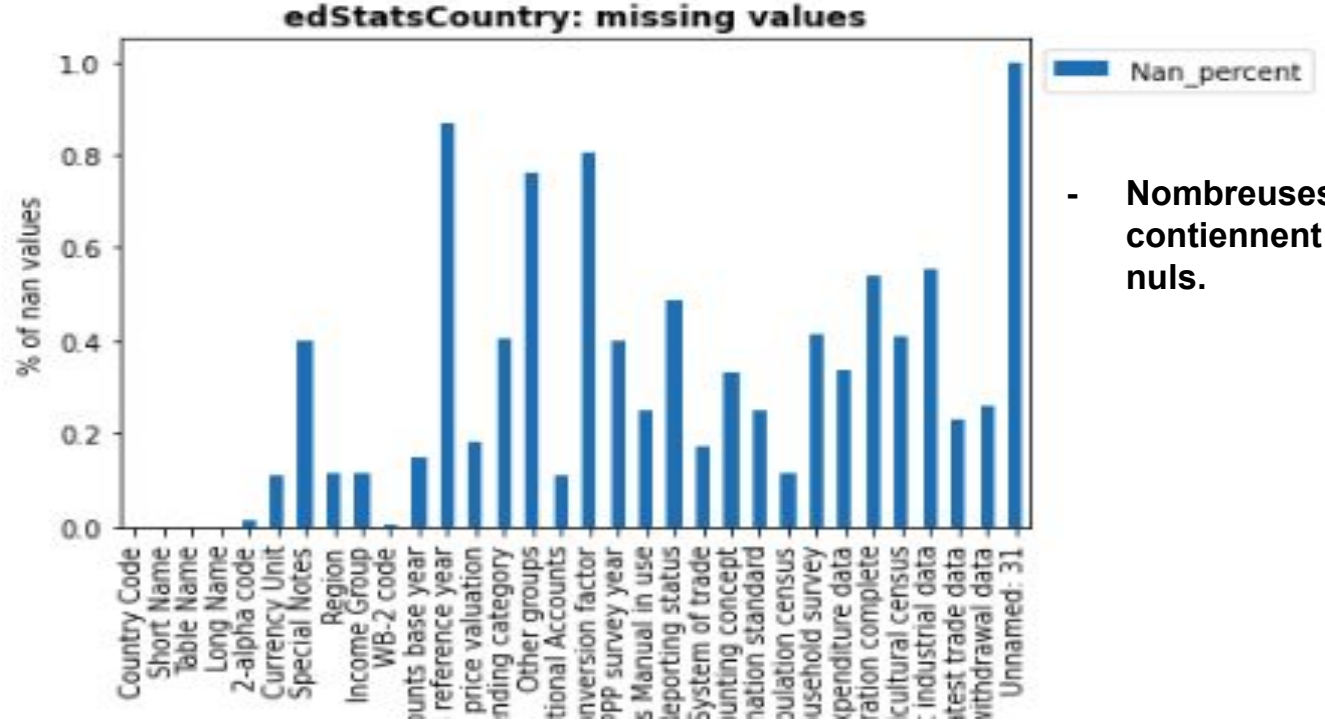
edStatsData → **Clé pour répondre à la problématique**

- Base des données en forme de série temporelle (indicateur-pays).
- Permet la quantification et la sélection des variables d'intérêt (ceci contient chiffres).

Nombre de Pays : 242

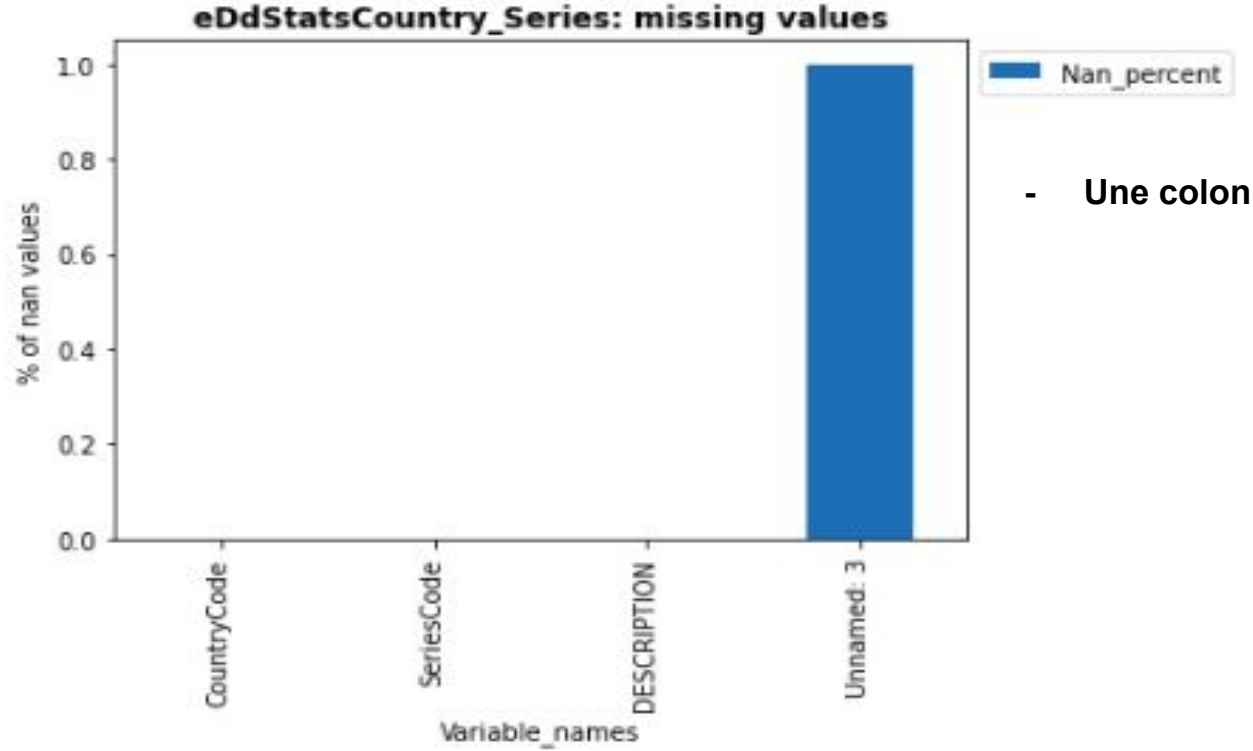
Nombre des Indicateurs : 3665

2. Jeux des données : Valeurs manquantes



- **Nombreuses colonnes contiennent valeurs nuls.**

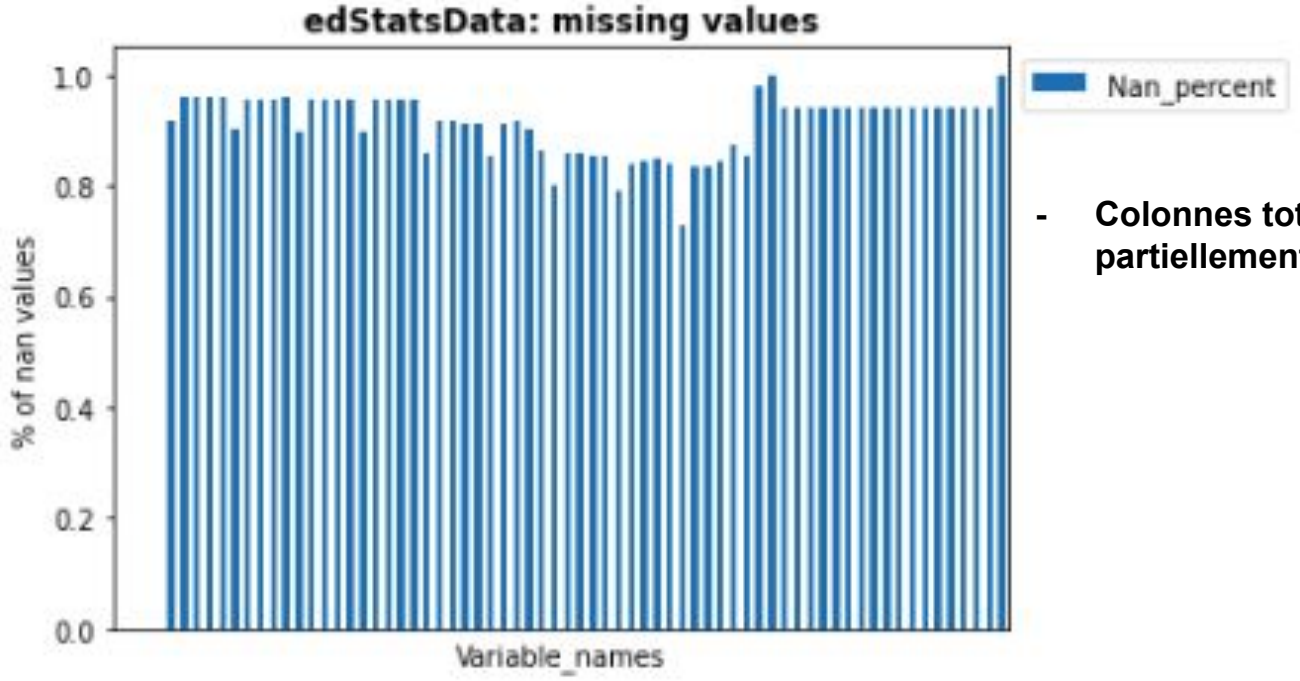
2. Jeux des données : Valeurs manquantes



-

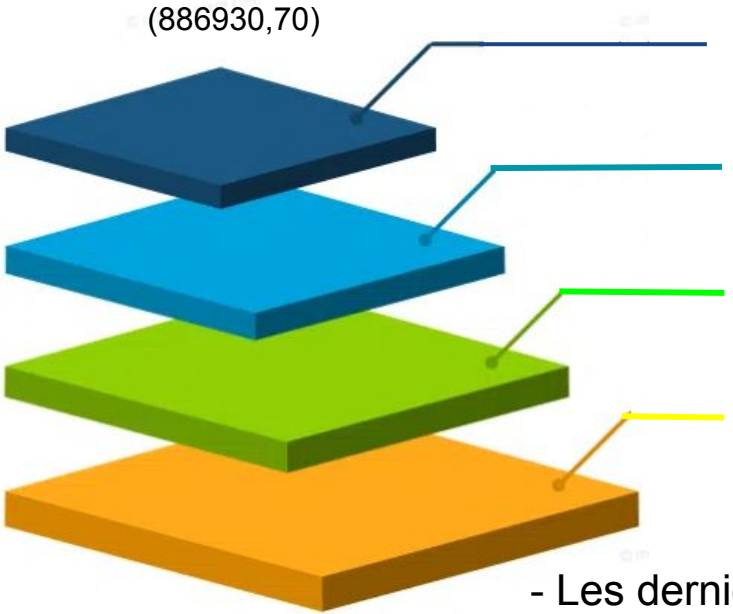
- **Une colonne est vidée.**

2. Jeux des données : Valeurs manquantes



- **Colonnes totalement et partiellement vidées.**

3. Filtrage des données : Méthodologie



1. Sélectionner de la base uniquement les pays.	(747660,70)
2. Eviter données qui ne sont pas entre (1990, et 2016).	(747660,33)
3. Effacement des indicateurs quasi vidés.	(149736,33)
4. Sélection des indicateurs les plus significatifs.	(13464,33)

- Les derniers indicateurs ont été sélectionnés à travers d'une fonction qui match avec des mots clés.

4. Sélection des indicateurs

Adjusted net enrolment rate in education.

Cumulative dropout rate.

Expenditure in education as % of total government expenditure.

GDP per capita

Expenditure in secondary education as % of the GDP.

Expenditure in tertiary education as % of the GDP.

Internet users (per 100 people).

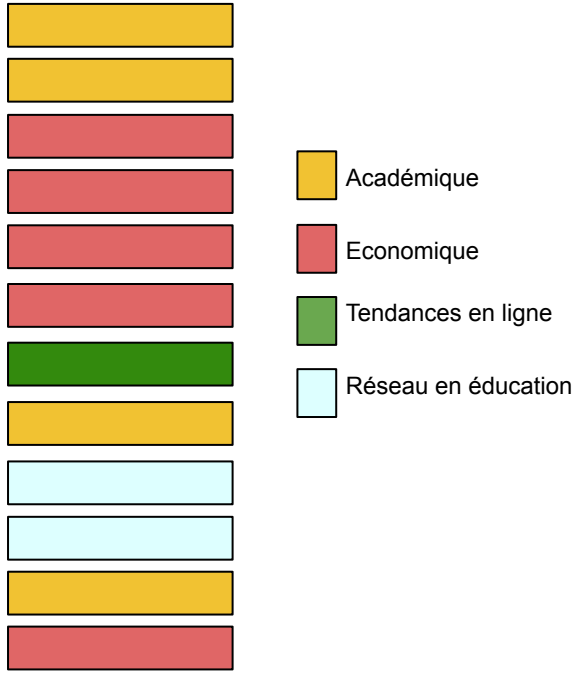
Out-of-school rate secondary education

Pupil teacher ratio secondary education

Pupil teacher ratio tertiary education

Total net enrolment rate secondary

Total invested money in education (secondary and tertiary).



5. Traitement des données : Valeurs nuls

Les valeurs manquantes sont rempliées selon les critères suivants:

Cas	Solution
1. Il y a des zéros dans la série?	Remplacer par NaN, et remplir avec la moyenne des valeurs non nuls.
2. Il n'y pas de NaNs	Remplir les manquantes avec la moyenne typique.

- Le remplissage est fait avec des moyennes pour fuir de biaiser la fourchette des valeurs de l'indicateur (éviter valeurs zéros -> **réaliste**).

6. Définitions : Ensembles et variables

P : *Ensemble des pays.*

T: *Ensemble de périodes .*

I: *Ensemble des indicateurs.*

S_{ipt}: Données *normalizées par indicateur, pays, et périodes.*

μ_{ip}: *Moyenne de l'indicateur* $i \in I$, *dans le pays* $p \in P$.

a_i: *Importance de l'indicateur* $i \in I$ *en rapport* à *son type.*

σ_i: *Ecart type de l'indicateur* $i \in I$.

ω_i: *Ecart type* pondéré *de l'indicateur* $i \in I$.

A: *Matrice d'importances par indicateur.*

W: *Matrice d'ecarts types* pondérés *par indicateur.*

U: *Produit Hadamard entre les matrices d'importances, et d'ecarts types par indicateur.*

F_p: *Fonction d'*attractivité *par pays.*

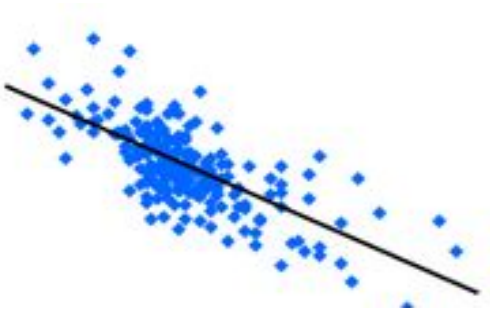
7. Fonction d'attractivité : Définition

$$U = A \circ W$$

$$F_p = MU^T$$

Attractivite

Variance



La fonction d'attractivité est un mélange qui donne de l'importance aux indicateurs tenant une grande variabilité. Les qualifications d'un pays sont après multipliées par les paramètres en question.

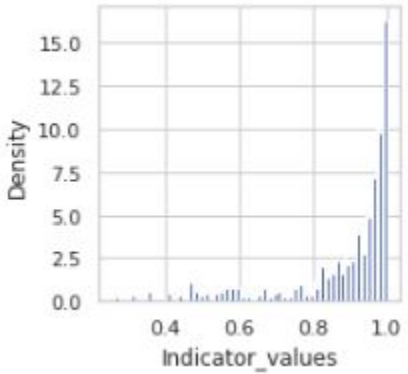
7. Fonction d'attractivité : Importance des indicateurs

Adjusted net enrolment rate in education.	6
Cumulative dropout rate.	-6
Expenditure in education as % of total government expenditure.	5
GDP per capita	5
Expenditure in secondary education as % of the GDP.	5
Expenditure in tertiary education as % of the GDP.	5
Internet users (per 100 people).	8
Out-of-school rate secondary education	-6
Pupil teacher ratio secondary education	7
Pupil teacher ratio tertiary education	7
Total net enrolment rate secondary	6
Total invested money in education (secondary and tertiary).	5

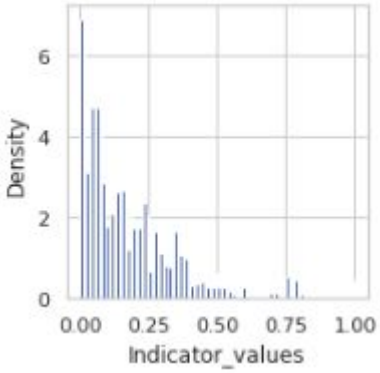
Académique	6
Economique	5
Tendances en lgine	8
Réseau en éducation	7

7. Fonction d'attractivité : Dispersion des indicateurs

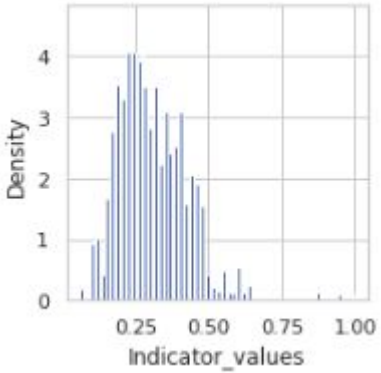
Adjusted net enrolment rate in education.



Cumulative dropout rate.



Expenditure in education as % of total government expenditure.

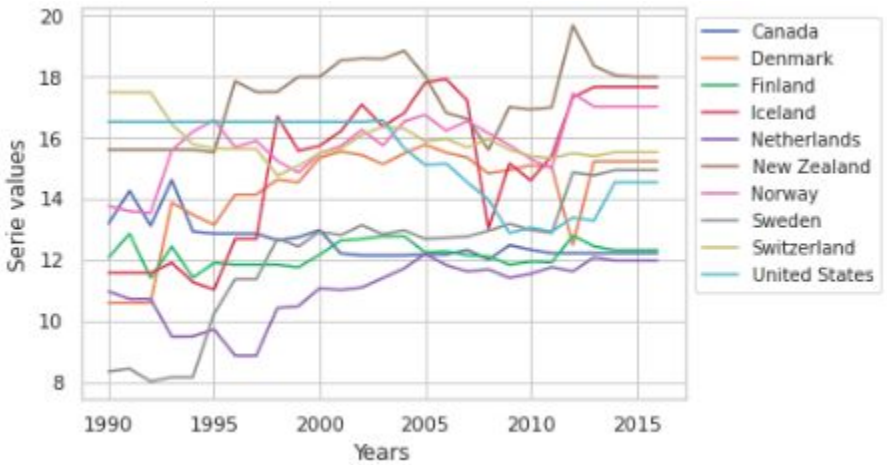


7. Fonction d'attractivité : Résultats

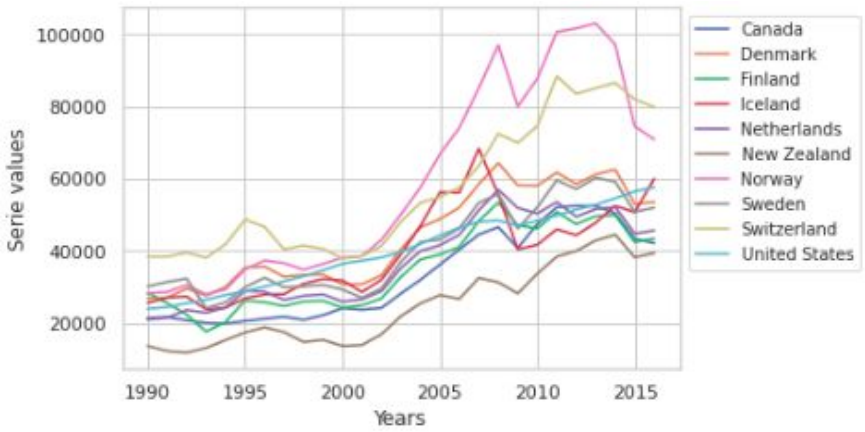
Nom pays	*Score d'attractivite*	*Continent*
Norvège	*0.65*	Europe
Danemark	*0.64*	Europe
Finlande	*0.62*	Europe
Suède	*0.62*	Europe
Islande	*0.62*	Europe
Suisse	*0.61*	Europe
Nouvelle-Zélande	*0.60*	Océanie
Canada	*0.59*	Amérique du Nord
États-Unis	*0.58*	Amérique du Nord
Pays-Bas	*0.57*	Europe

8. Evolution des indicateurs

Expenditure in education as % of total government expenditure.



GDP per capita



8. Conclusions

Est-ce que les données permettent d'informer le projet d'expansion?

- **Permettent de quantifier l'importance des indicateurs.**
- **Couvrent en globalité tous les aspects/domaines les plus relevants pour aborder un projet d'éducation en ligne.**
- **Trier les pays les plus intéressants au but de l'éducation virtuelle.**

Annexes : Normalization et calcul des variables

Les données sont normalisées par indicateur à travers de la valeur la plus grande entre les pays et leur périodes.

$$S_{ipt} = \frac{V_{ipt}}{\max\{V_{ipt}\}}; \forall i \in I, p \in P, t \in T$$

Pour chaque pays et indicateur, on calcule les moyennes des valeurs normalisées.

$$\mu_{ip} = \sum_{t} \frac{S_{ipt}}{|T|}; \forall i \in I, \forall p \in P \qquad \mu_{i} = \sum_{p} \sum_{t} \frac{S_{ipt}}{|T||P|}; \forall i \in I$$

Annexes : Normalization et calcul des variables

Pour chaque indicateur un score d'appréciation est donné, ainsi que le calcul d'écarts types.

$$a_i ; \forall i \in I$$

$$\sigma_i ; \forall i \in I$$

$$\sigma_i = \sqrt{\sum_p \sum_t \frac{(S_{ipt} - \mu_i)^2}{|T||P| - 1}} \; ; \; \forall i \in I$$

Des poids relatifs à la dispersion des données sont calculés comme une pondération.

$$\omega_i = \sigma_i / \sum_{j \in I} \sigma_j \; ; \; \forall i \in I$$

Annexes : Définition

Sachant les dernières calculs, les matrices suivantes sont définies.

$$M = \begin{pmatrix} \mu_{11} & \mu_{21} & \cdots & \mu_{|I|1} \\ \mu_{12} & \mu_{22} & \cdots & \cdots \\ \vdots & \cdots & \cdots & \cdots \\ \mu_{1|P|} & \cdots & \cdots & \mu_{|I||P|} \end{pmatrix}$$

$$A = (a_1 \quad a_2 \quad \cdots \quad a_{|I|})$$

$$W = (\omega_1 \quad \omega_2 \quad \cdots \quad \omega_{|I|})$$

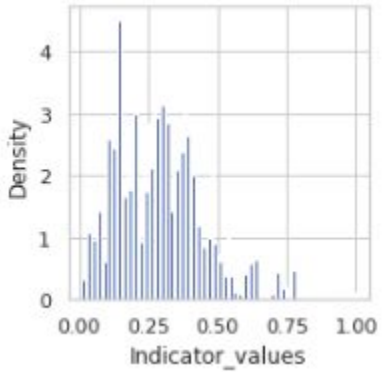
Les pays les plus attractifs sont lesquels qui ont les valeurs les plus grandes, selon la fonction déterminée par:

$$U = A \circ W$$

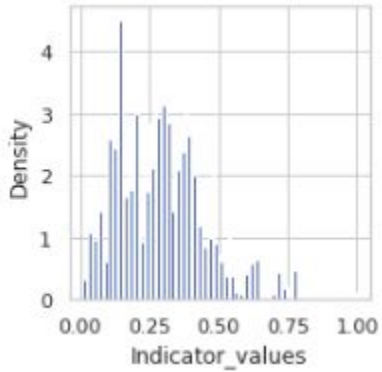
$$F_p = MU^T$$

Annexes : Dispersion des indicateurs

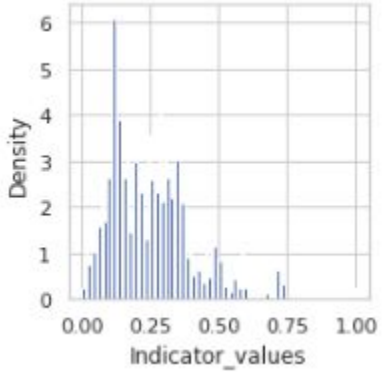
GDP per capita



Expenditure in secondary education as % of the GDP.

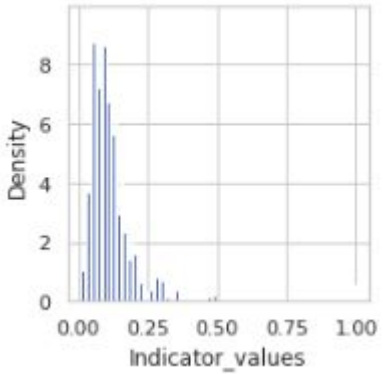


Expenditure in tertiary education as % of the GDP.

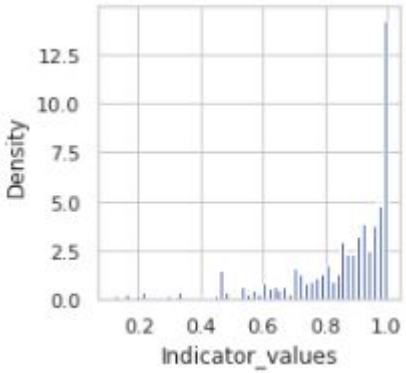


Annexes : Dispersion des indicateurs

Pupil teacher ratio tertiary education

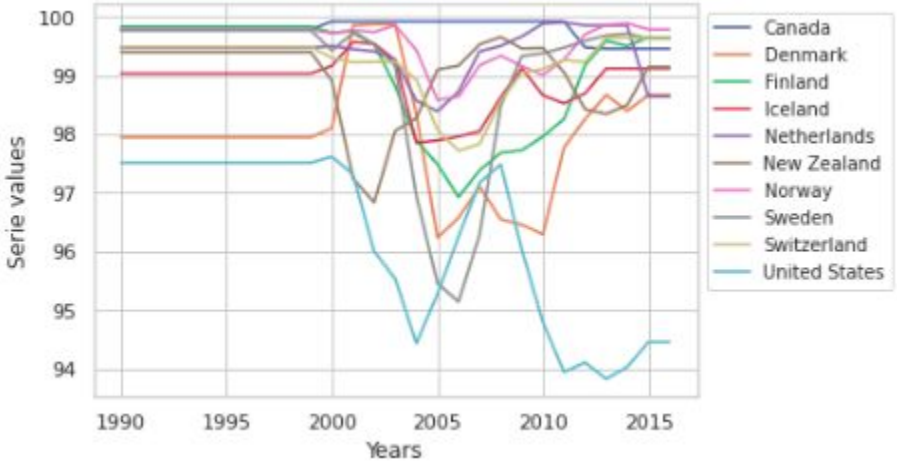


Total net enrolment rate secondary

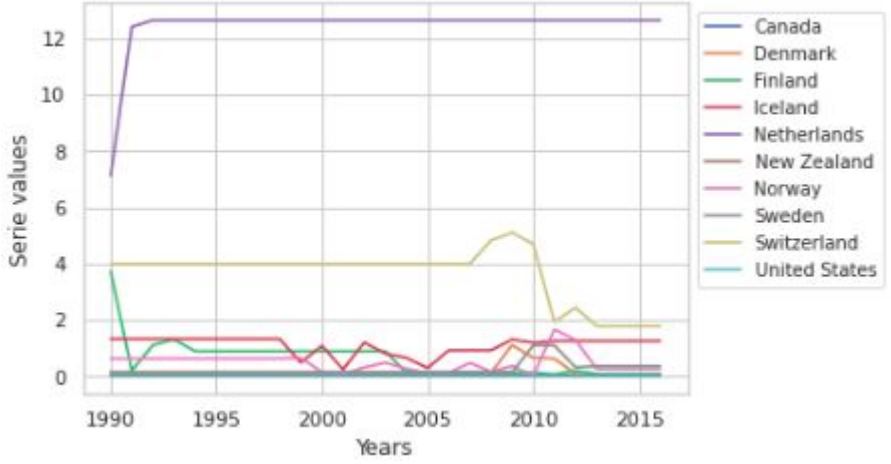


Annexes : Evolution des indicateurs

Adjusted net enrolment rate in education.

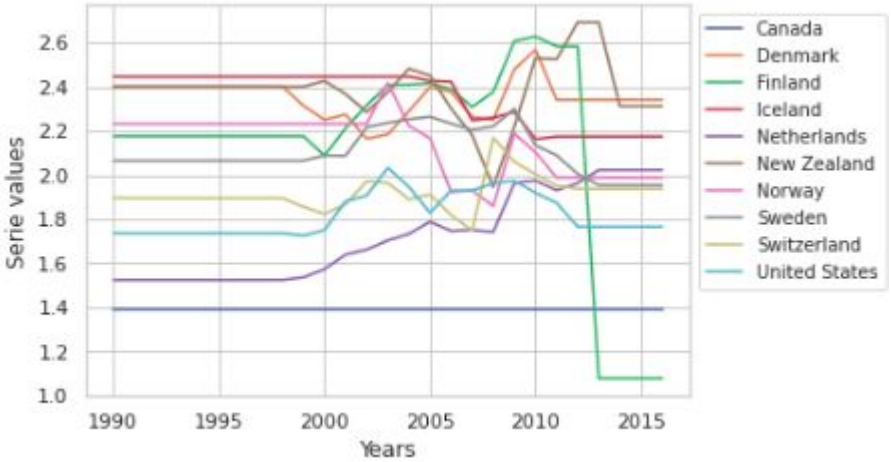


Cumulative dropout rate.

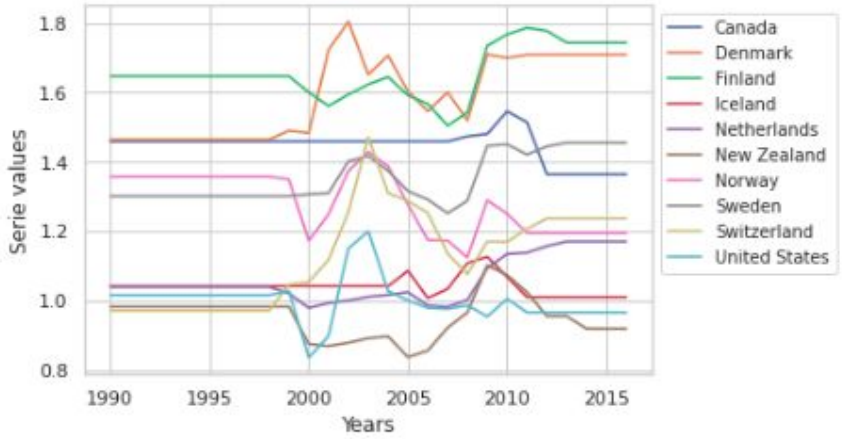


Annexes : Evolution des indicateurs

Expenditure in secondary education as % of the GDP.

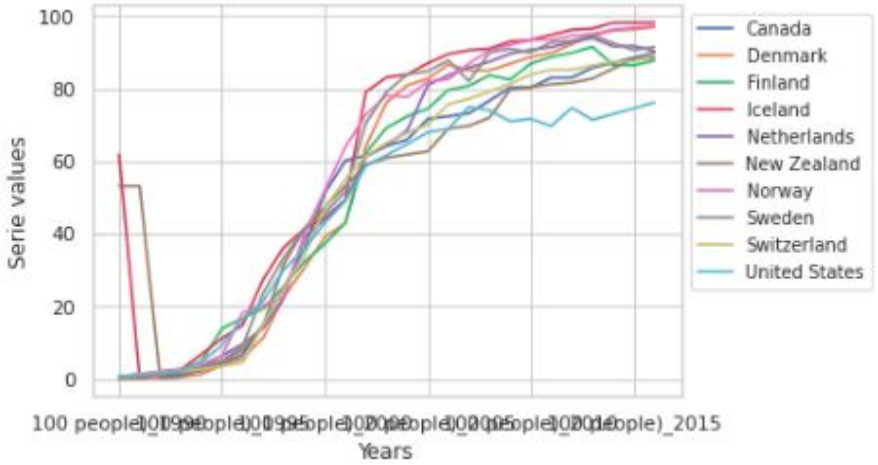


Expenditure in tertiary education as % of the GDP.

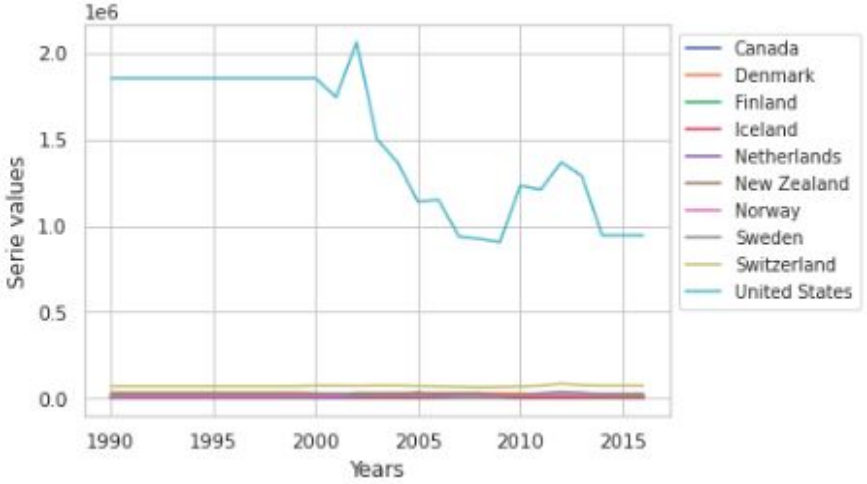


Annexes : Evolution des indicateurs

Internet users (per 100 people).

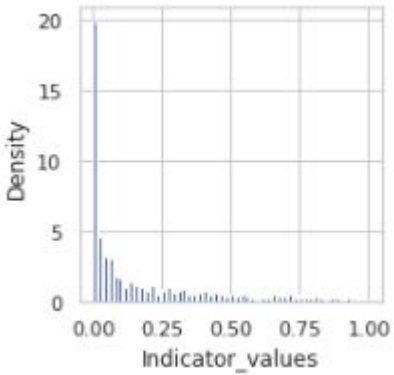


Out-of-school rate secondary education

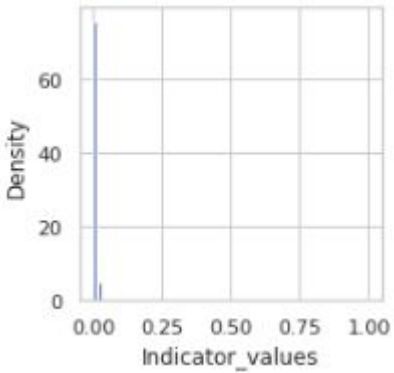


Annexes : Dispersion des indicateurs

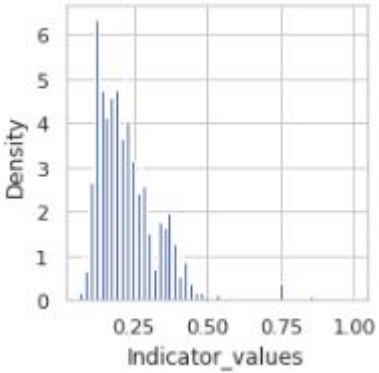
Internet users (per 100 people).



Out-of-school rate secondary education

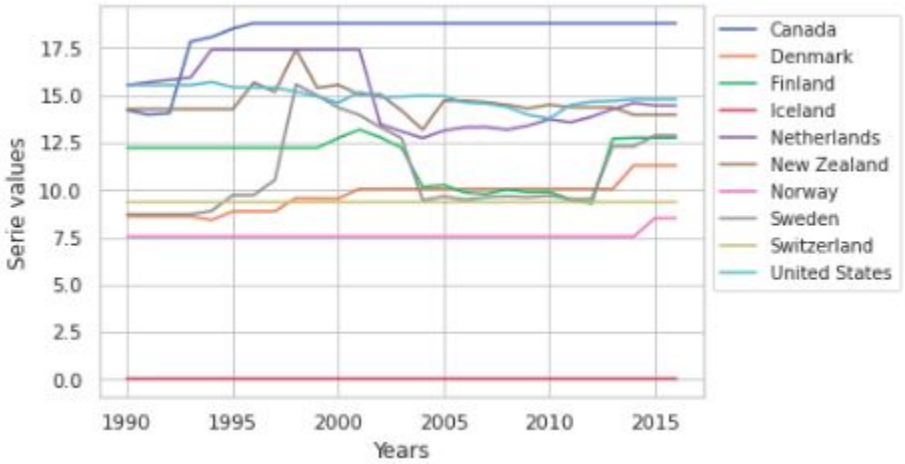


Pupil teacher ratio secondary education

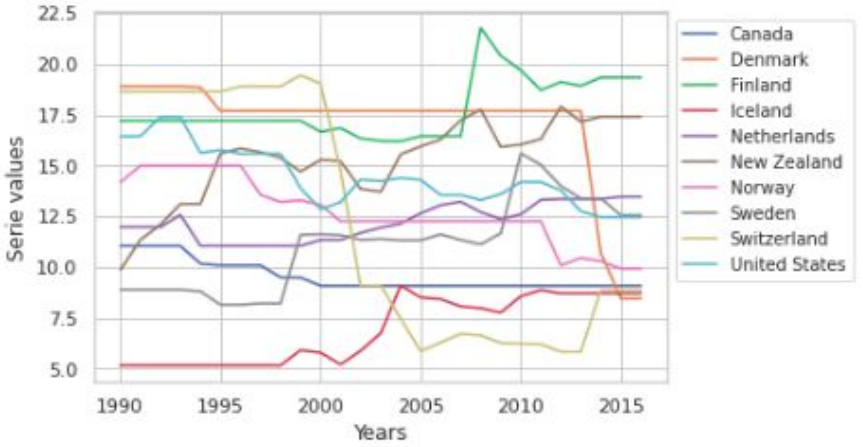


Annexes : Evolution des indicateurs

Pupil teacher ratio secondary education

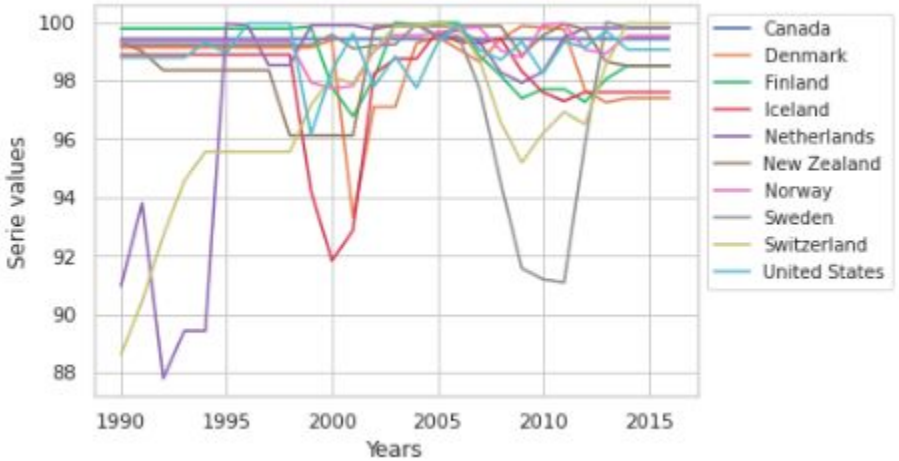


Pupil teacher ratio tertiary education



Annexes : Evolution des indicateurs

Total net enrolment rate secondary



Total invested money in education (secondary and tertiary).