Analyser des données de systèmes éducatifs

PLAN

1/ Présentation de la problématique
2/ Présentation du jeu de données
3/ Analyse des données manquantes
4/ Nettoyage du jeu de données et des données manquantes
5/ Analyse et choix des indicateurs
6/ Scoring
7/ Analyse par région
8/ Projection

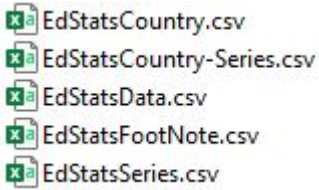
Start-up de la EdTech ➡ *academy*

Formations en ligne

Expansion internationale

Mission ➡ Explorer un jeu de données
- Fort potentiel ?
- Évolution ?
- Où opérer en premier ?

➡ 5 fichiers CSV



Fichier principal ➡ EdStatsData

✚ Autres fichiers renseignent sur :

Description des indicateurs
Informations sur les pays (alpha code, région etc.)
Sources des données

Dimension : 886 930 lignes et 70 colonnes

Colonnes :

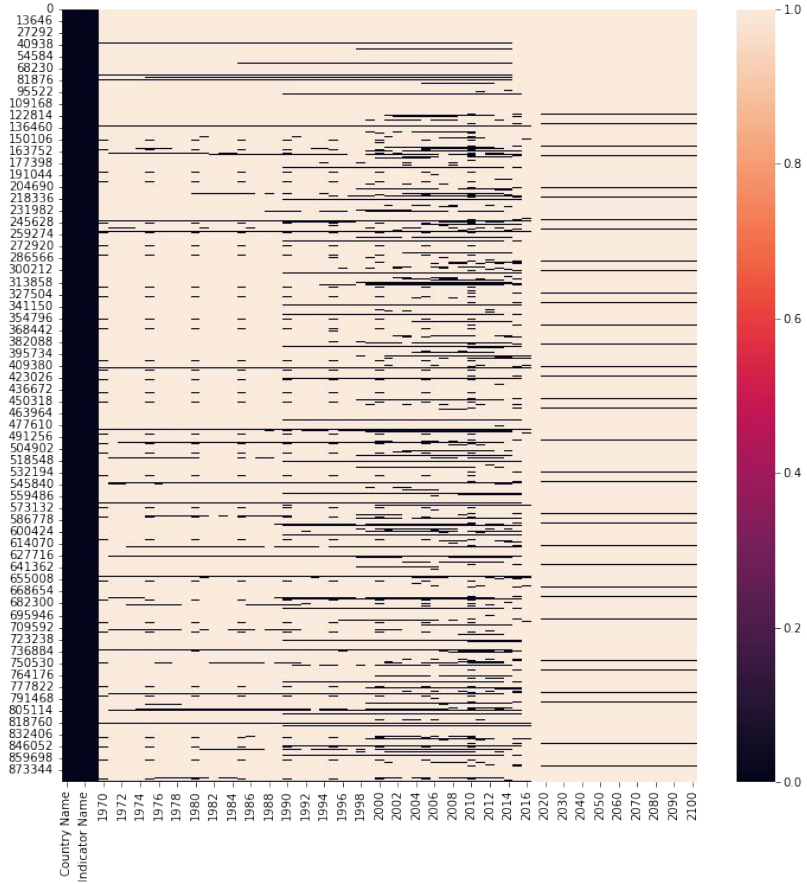
66 colonnes numériques (valeurs des indicateurs par année)
4 colonnes textuelles (noms et codes des pays et indicateurs)

3 665 indicateurs uniques

242 pays uniques

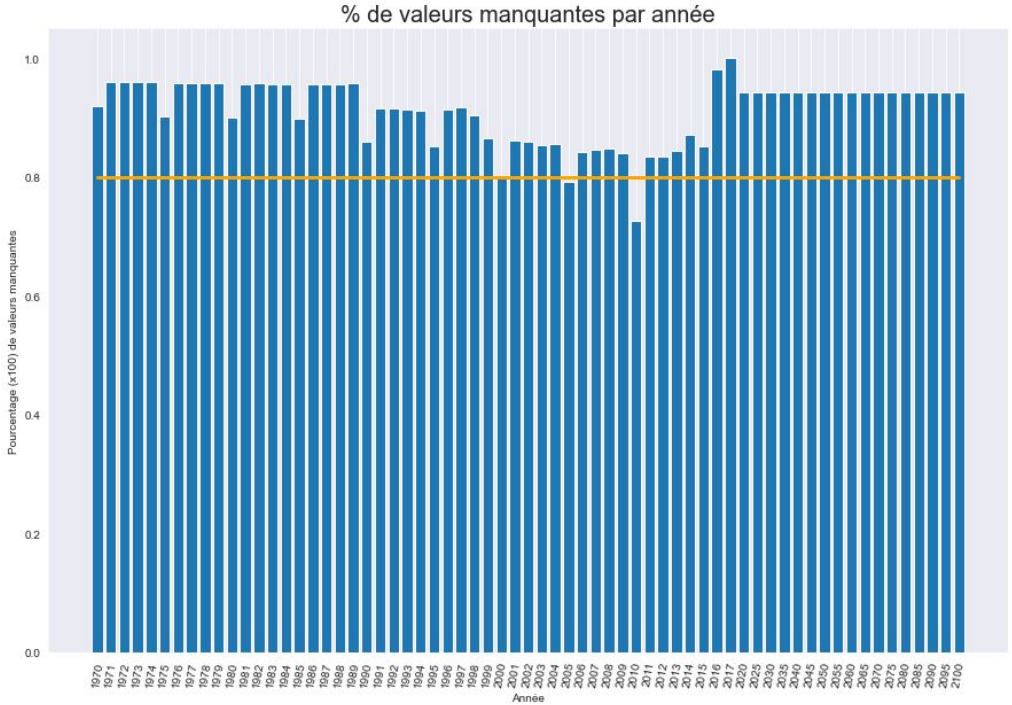
Pas de dupliqué

Analyse des données manquantes

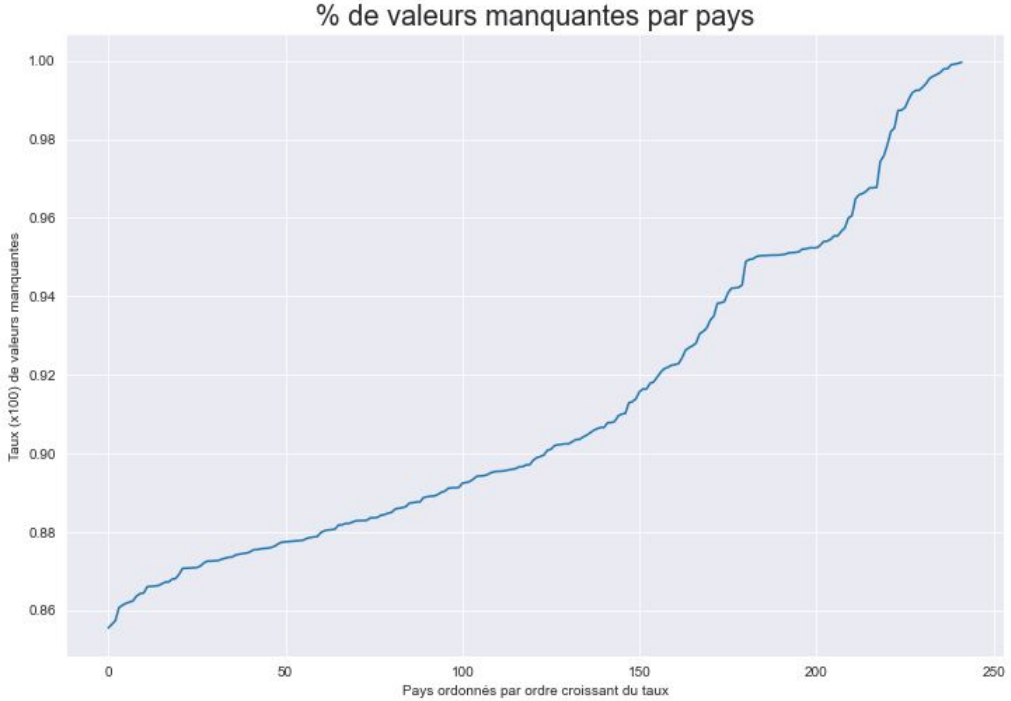


Analyse des données manquantes

Valeurs manquantes pour chaque année (colonnes numériques)



Analyse des données manquantes



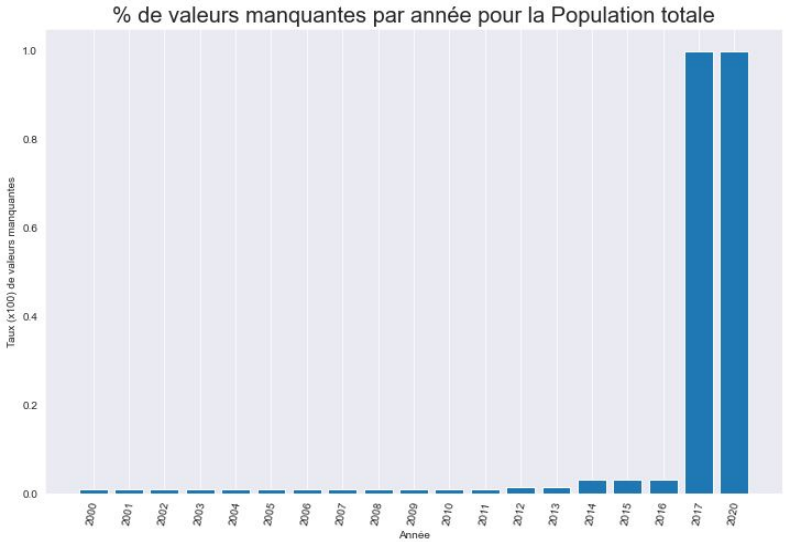
Valeurs manquantes pour chaque pays

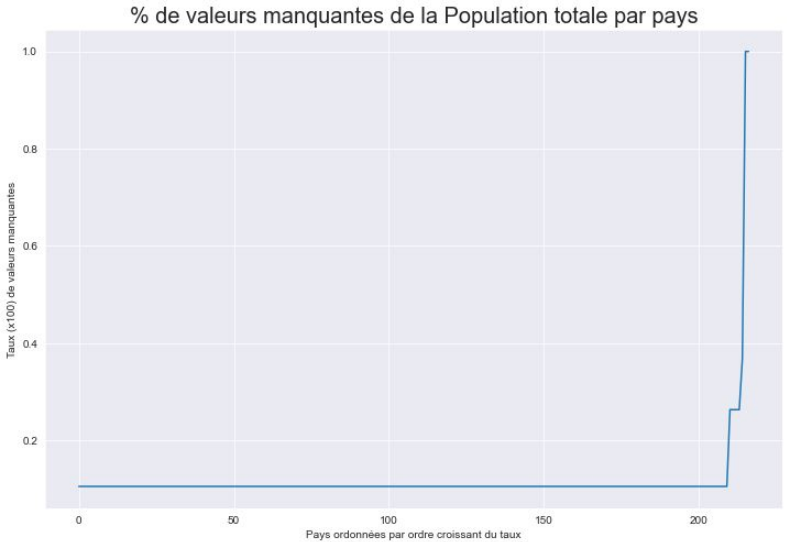
- Supprimer les années avant 2000 (début d'internet au début du 21ème siècle)
  (première formation en ligne de 1994, en France)*
  Et les années de projection (futures)

- Supprimer les pays qui n'en sont pas : régions géographiques (nb : 25)
  Arab World, East Asia & Pacific, East Asia & Pacific (excluding high income), Euro area, Europe & Central Asia, Europe & Central Asia (excluding high income), European Union, Heavily indebted poor countries (HIPC), High income, Latin America & Caribbean, Latin America & Caribbean (excluding high income), Least developed countries: UN classification, Low & middle income, Low income, Lower middle income, Middle East & North Africa, South Asia, Sub-Saharan Africa, Middle East & North Africa (excluding high income), Upper middle income, World, Middle income, North America, OECD members, Sub-Saharan Africa (excluding high income)

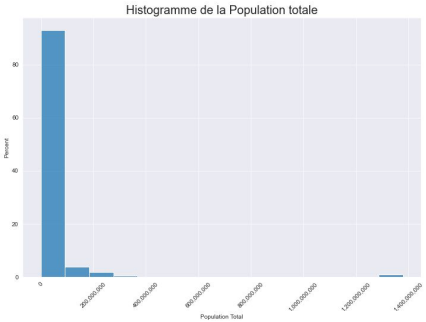
*Source : https://www.26academy.com/lhistoire-de-le-learning/

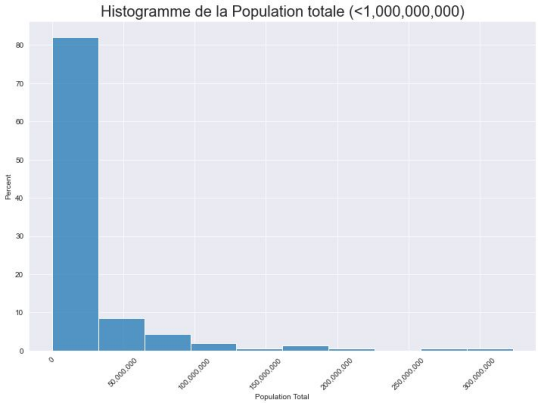
- Supprimer les “petits” pays : pays avec une faible population

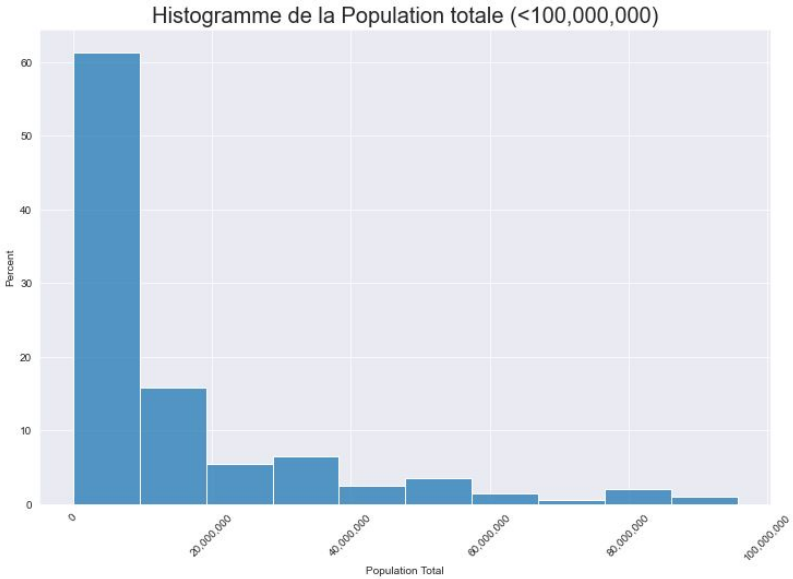




Analyse des données manquantes

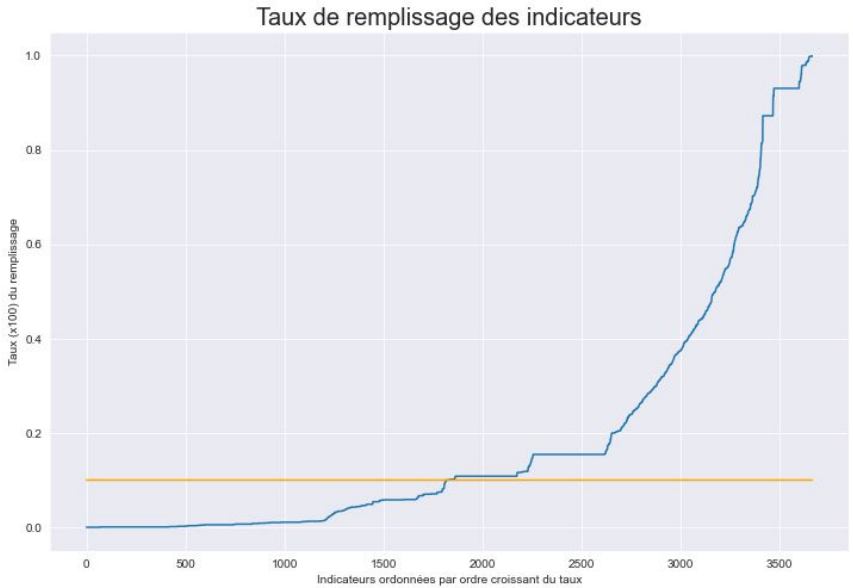






- Garder les pays avec au moins 10,000,000 d'habitants
  (88 pays restant)

- Supprimer les années 2017 et 2020 : beaucoup de données manquantes
- Agréger (somme) de l'indicateur Barro-Lee sur la population pour affiner la tranche d'âge (15-29 ans)
- Chercher un indicateur externe au jeu de données :
  FDI (Foreing Direct Investment), ou l'investissement direct à l'étranger



Analyse et choix des indicateurs

- Conserver les indicateurs avec un taux de remplissage de 10% minimum

- Choisir des indicateurs pertinents

  Government expenditure on education as % of GDP (%),
  Capital expenditure as % of total expenditure in tertiary public institutions (%),
  Barro-Lee: Population in thousands, age 15-54, total,
  Unemployment, total (% of total labor force), GDP per capita (current US$),
  Internet users (per 100 people), FDI inward position (%GDP)

Indicator Name	Taux de remplissage
Barro-Lee: Population in thousands, age 15-29, total	0.154412
Capital expenditure as % of total expenditure in tertiary public institutions (%)	0.308155
Government expenditure on education as % of GDP (%)	0.570856
FDI inward position (%GDP)	0.628803
GDP per capita (current US$)	0.960561
Internet users (per 100 people)	0.977941
Unemployment, total (% of total labor force)	0.988636

Scoring

- Donnée la plus récente date pour chaque indicateur

- Procéder au scoring :
  Attribuer un rang à chaque pays sur chaque indicateur (plus petite valeur a le plus petit rang)
  Pondérer chaque indicateur (valeur rang)

Indicator Name	Weight
Unemployment, total (% of total labor force)	1
FDI inward position (%GDP)	2
Government expenditure on education as % of GDP (%)	3
Capital expenditure as % of total expenditure in tertiary public institutions (%)	4
GDP per capita (current US$)	5
Internet users (per 100 people)	6
Barro-Lee: Population in thousands, age 15-54, total	7

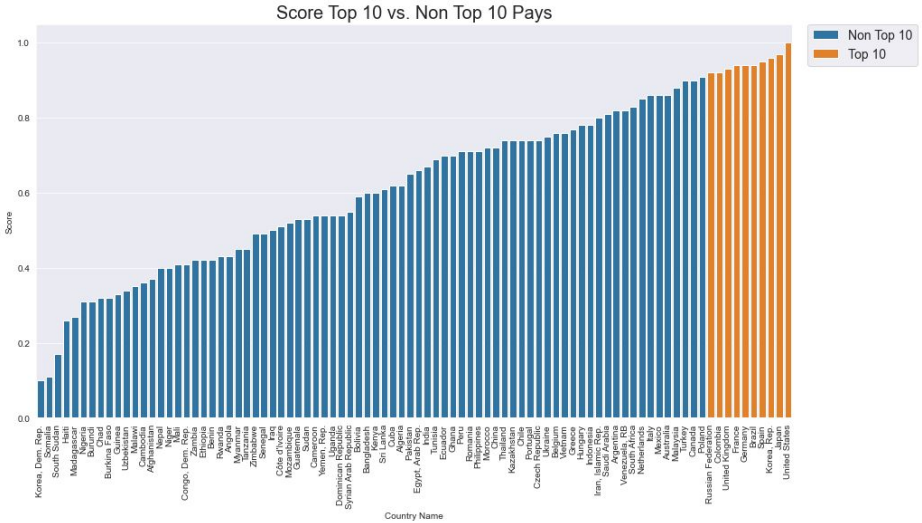
Score par pays

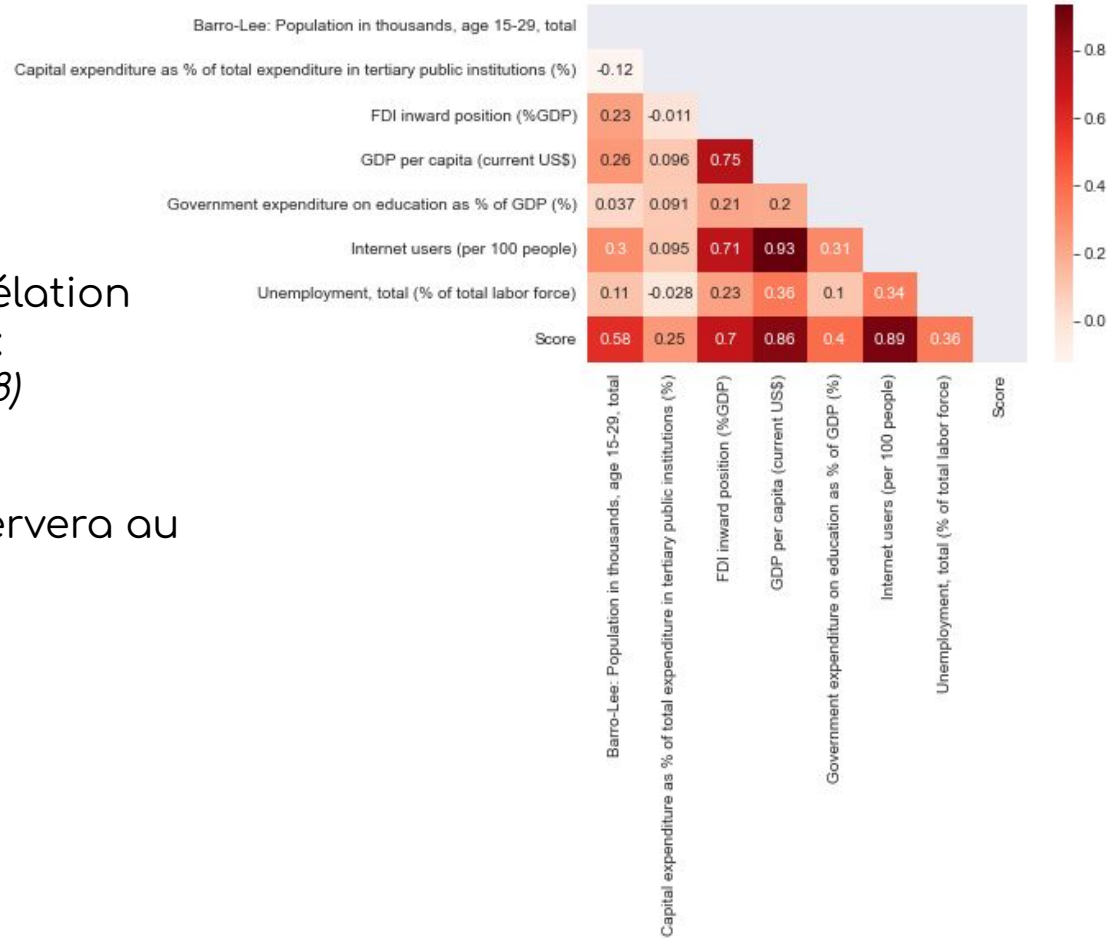
$$\frac{\sum_1^n (Weight_i * Rang_i)}{Nb\ d'indicateurs}$$

Score final = Score / max{Score}

Scoring

Country Name	Barro-Lee: Population in thousands, age 15-29, total	Capital expenditure as % of total expenditure in tertiary public institutions (%)	FDI inward position (%GDP)	GDP per capita (current US$)	Government expenditure on education as % of GDP (%)	Internet users (per 100 people)	Unemployment, total (% of total labor force)	Score
United States	86.0	37.0	72.0	87.0	59.0	73.0	50.0	1.00
Japan	78.0	63.0	60.0	83.0	26.0	85.0	23.0	0.97
Korea, Rep.	65.0	66.0	61.0	76.0	51.0	86.0	16.5	0.96
Spain	56.0	65.0	78.0	78.0	43.0	79.0	84.0	0.95
Germany	70.0	36.0	70.0	81.0	47.0	83.0	57.0	0.94
France	64.0	38.0	69.0	80.0	67.0	80.0	55.0	0.94
Brazil	83.0	40.0	71.0	65.0	70.0	63.0	70.0	0.94
United Kingdom	66.0	11.0	80.0	84.0	69.0	88.0	39.0	0.93
Russian Federation	81.0	44.0	73.0	69.0	32.0	74.0	54.0	0.92
Colombia	68.0	86.0	81.0	58.0	44.0	58.0	82.0	0.92

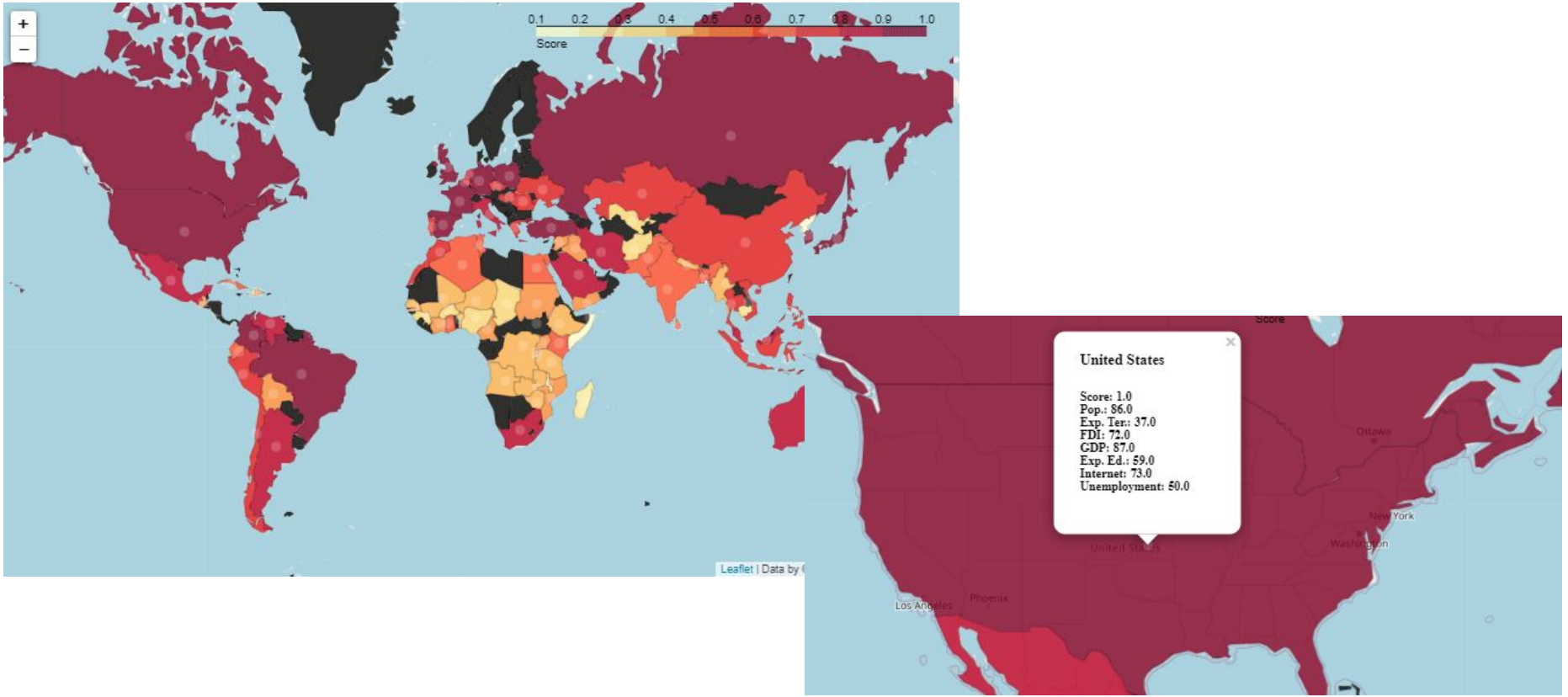




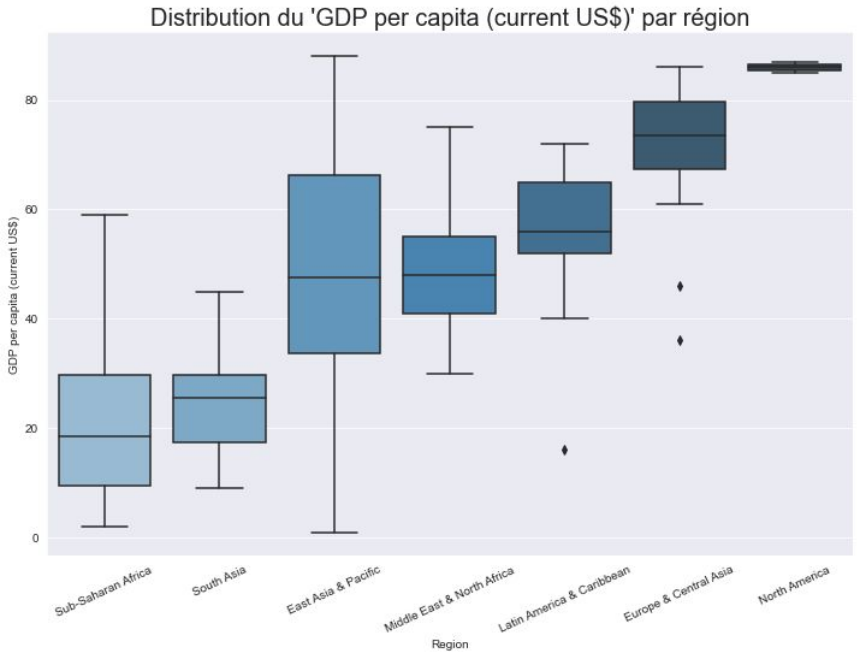
- **Score** a une forte corrélation avec deux indicateurs :
  - *GDP per capita (0.88)*
  - *Internet users (0.91)*

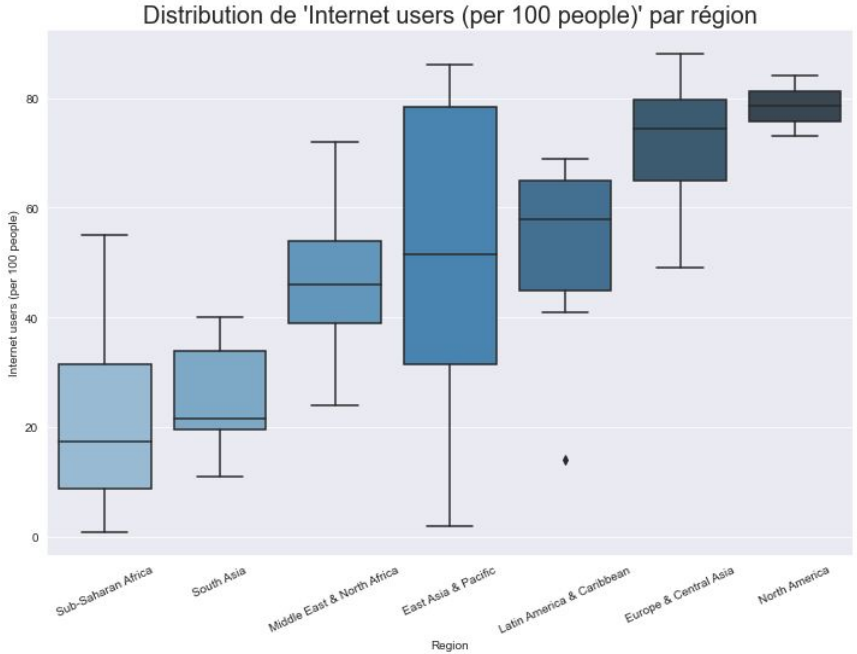
- Indicateurs qu'on observera au niveau des régions

Scoring

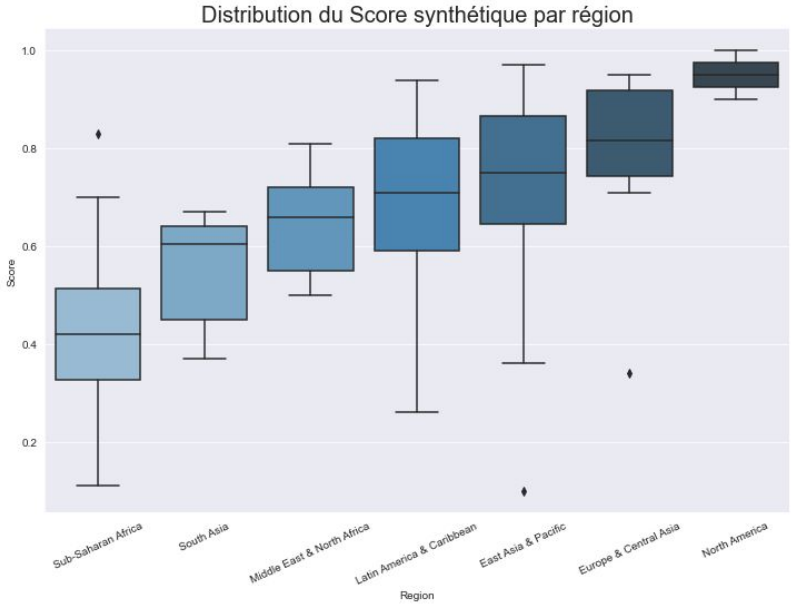

0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Score
Leaflet | Data by
United States
Score: 1.0
Pop.: 86.0
Exp. Ter.: 37.0
FDI: 72.0
GDP: 87.0
Exp. Ed.: 59.0
Internet: 73.0
Unemployment: 50.0
Ottawa
New York
Washington
Phoenix
Los Angeles

- S'intéresser à l'étendue des valeurs des indicateurs qui corrèlent le plus avec le score et le score





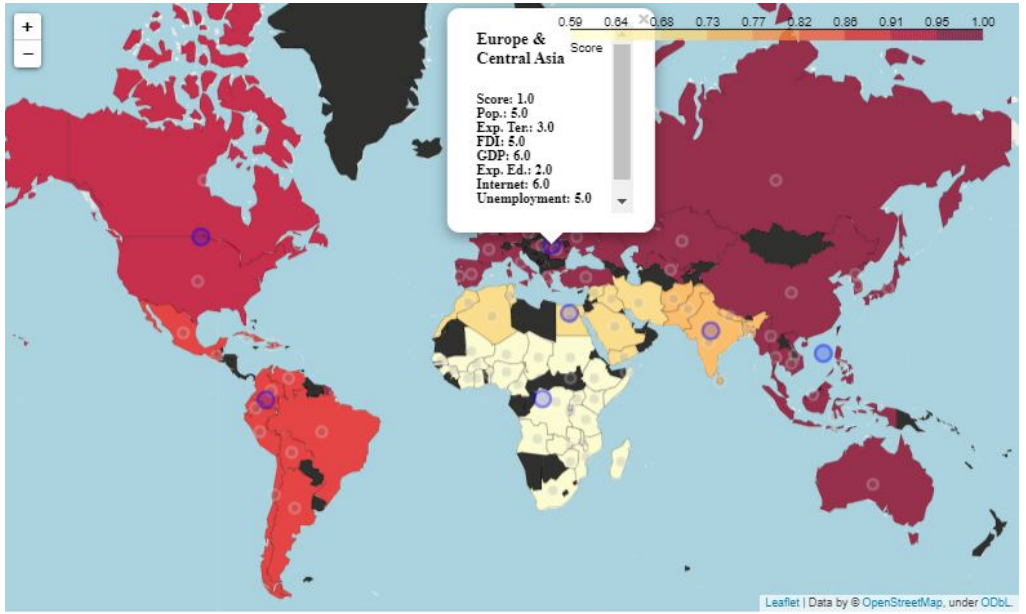
Analyse par région



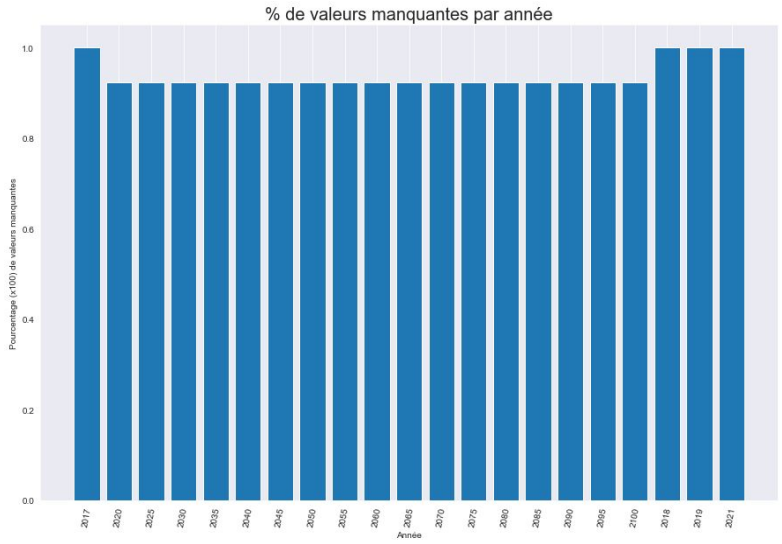
- On peut observer que les régions Latin America & Carribbean et Europe & Central Asia ont des dispersions plus petites sur les indicateurs ; Europe & Central Asia a une dispersion plus petite sur le Score

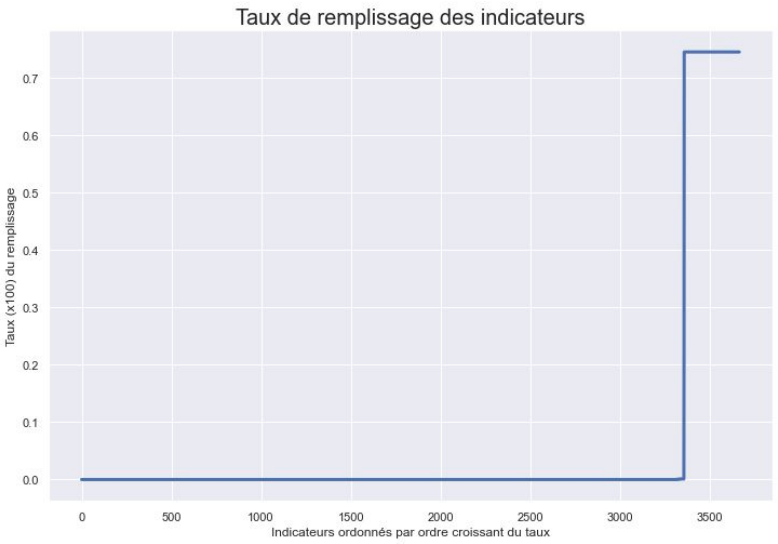
Analyse par région

Region	Barro-Lee: Population in thousands, age 15-29, total	Capital expenditure as % of total expenditure in tertiary public institutions (%)	FDI inward position (%GDP)	GDP per capita (current US$)	Government expenditure on education as % of GDP (%)	Internet users (per 100 people)	Unemployment, total (% of total labor force)	Score Region
Europe & Central Asia	5.0	3.0	5.0	6.0	2.0	6.0	5.0	1.00
East Asia & Pacific	7.0	5.0	2.0	4.0	4.0	4.0	1.0	0.97
North America	1.0	1.0	7.0	7.0	6.0	7.0	4.0	0.93
Latin America & Caribbean	4.0	2.0	4.0	5.0	5.0	5.0	6.0	0.90
South Asia	6.0	7.0	1.0	2.0	1.0	2.0	3.0	0.75
Middle East & North Africa	2.0	4.0	3.0	3.0	7.0	3.0	7.0	0.72
Sub-Saharan Africa	3.0	6.0	6.0	1.0	3.0	1.0	2.0	0.59



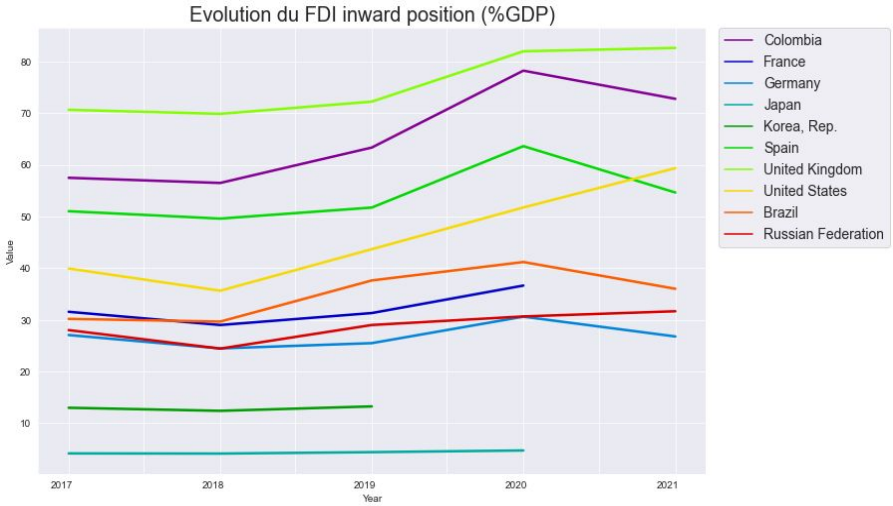
- Est-ce que notre Top 10 a des indicateurs évoluant positivement dans le futur ?





Projection

Indicator Name	Taux de remplissage
Unemployment, total (% of total labor force)	0.000000
Capital expenditure as % of total expenditure in tertiary public institutions (%)	0.000000
Barro-Lee: Population in thousands, age 15-29, total	0.000000
Internet users (per 100 people)	0.000000
Government expenditure on education as % of GDP (%)	0.000000
GDP per capita (current US$)	0.000000
FDI inward position (%GDP)	0.222898



- Difficile de tirer des conclusions avec un seul indicateur

Conclusion

- Deux approches de déploiement du projet sont envisageables :

  Au niveau des **pays** : on s'intéresse à notre Top 10

  Au niveau des **régions** : on croise notre Top 10 avec les régions les mieux classées

- La région **Europe & Central Asia** semble être la première cible (avec 5 de ses pays dans le Top 10), **l'Espagne** est le premier pays à cibler

- Utilisation d'une **map automatisée** pour analyser chaque région