

# **P2 : Analyser des données de systèmes éducatifs**

Présentation

Formation Data Scientist  
OpenClassrooms

Romain Vaillant

20 Juillet 2021 - 18 Mai 2022

# Sommaire

## **1. Rappel de la problématique**

## **2. Pré-analyse**

- a – Présentation du jeu de données
- b – Pré-filtrage
- c – Indicateurs sélectionnés

## **3. Traitement des données**

- a – Approche globale
- b – Filtrage
- c – Exemples

## **4. Score d'attractivité**

- a – Aperçu des pays retenus
- b – Méthode de calcul
- c – Pays potentiels Top 15

## **5. Conclusion**

# 1. Rappel de la problématique

- Data Scientist
- **Start-up de la EdTech** nommée **academy**
- Contenus de formation en ligne
- Public de niveau lycée et université

## **Mission d'analyse exploratoire :**

Déterminer si les données sur l'éducation de la banque mondiale permettent d'informer le projet d'expansion.



## 2. Pré-analyse

### a – Présentation du jeu de données

#### **EdStatsCountry.csv**

Données qualitatives

Régions

#### **EdStatsData.csv**

Données quantitatives

Fichier principal d'analyse

#### **EdStatsFootNote.csv**

Données qualitatives

Ignoré

#### **EdStatsCountry-Series.csv**

Données qualitatives

Ignoré

#### **EdStatsSeries.csv**

Données qualitatives

Ignoré

## 2. Pré-analyse

### a – Présentation du jeu de données

#### **EdStatsCountry.csv**

- 241 lignes (241 pays uniques, 7 régions uniques)
- 32 colonnes
- 32 colonnes qualitatives ou dates, ex : (*Income Group : Upper middle income, Latest agricultural census : 2010*)
- 30% de données manquantes

#### **EdStatsData.csv**

- 886 930 lignes (242 pays uniques, 3665 indicateurs uniques)
- 70 colonnes
- 4 colonnes qualitatives (*Country Name, Country Code, Indicator Name, Indicator Code*)
- 1970 - 2017 (+1), 2020 - 2100 (+5)
- 86% de données manquantes

## 2. Pré-analyse

### b – Pré-filtrage

Dans **EdStatsCountry.csv** :

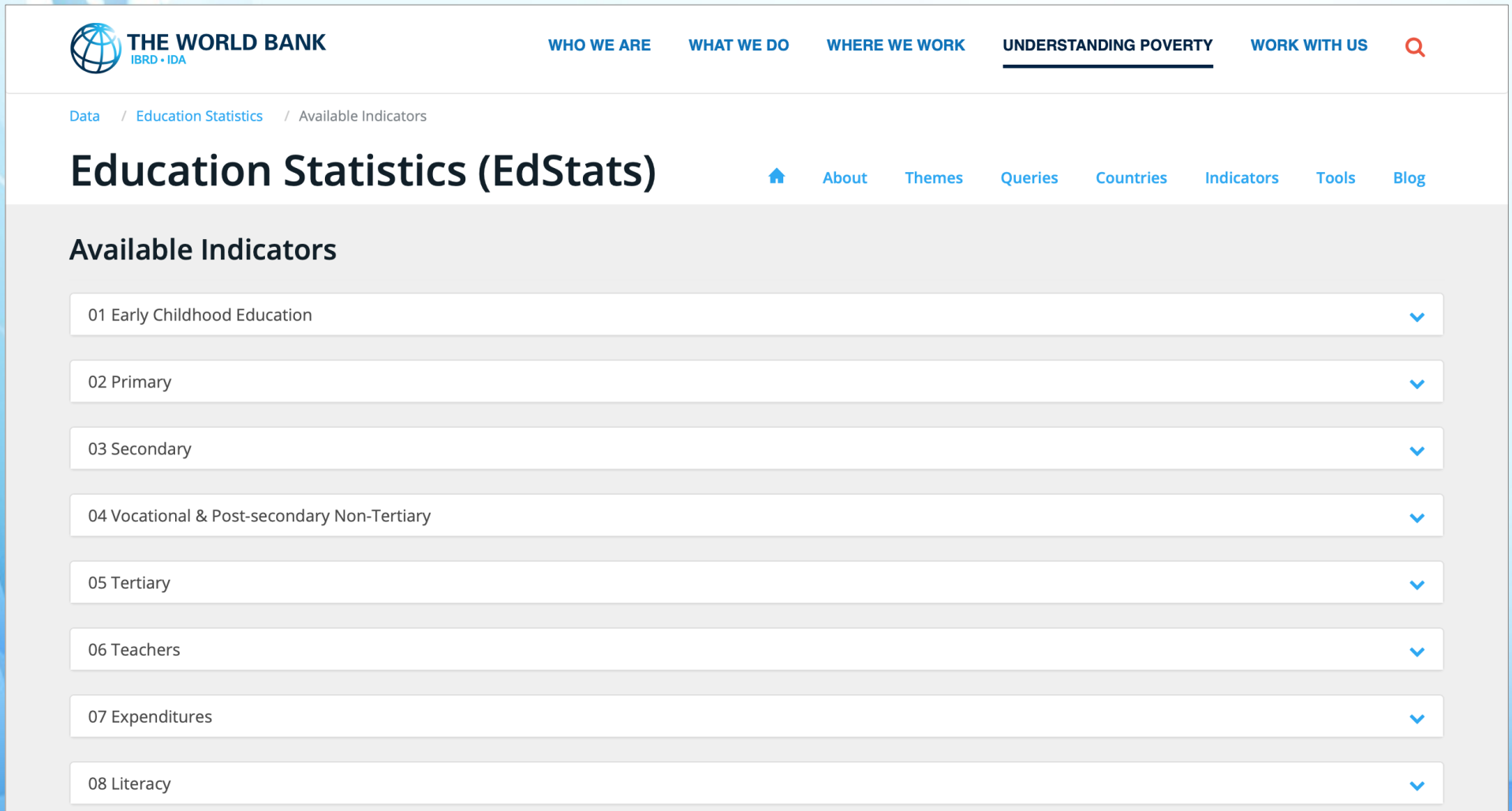
- 27 rangs/pays pour lesquels Region : NaN
- (Arab World, East Asia & Pacific (developing only), OECD members, World, ... )
- Exclure

Dans **EdStatsData.csv** :

- Merge : `df_EdStatsCountry.csv['Region', 'Income Group'], df_EdStatsData.csv}`
- 886 930 → 784 310 lignes (-11%)

## 2. Pré-analyse

### c – Indicateurs sélectionnés



The screenshot displays the World Bank's Education Statistics (EdStats) website. The header includes the World Bank logo and navigation links: WHO WE ARE, WHAT WE DO, WHERE WE WORK, UNDERSTANDING POVERTY (underlined), and WORK WITH US. A search icon is also present. Below the header, a breadcrumb trail reads: Data / Education Statistics / Available Indicators. The main title is "Education Statistics (EdStats)". A secondary navigation bar contains links: Home, About, Themes, Queries, Countries, Indicators, Tools, and Blog. The "Available Indicators" section lists eight categories, each with a dropdown arrow:

- 01 Early Childhood Education
- 02 Primary
- 03 Secondary
- 04 Vocational & Post-secondary Non-Tertiary
- 05 Tertiary
- 06 Teachers
- 07 Expenditures
- 08 Literacy



## 2. Pré-analyse

### c – Indicateurs sélectionnés

- Internet users (per 100 people)
- Population, ages 15-64, total

**Effective internet users = Internet users (per 100 people) \* Population, ages 15-64, total**

- Enrolment in tertiary education, all programmes, both sexes (number)
- Enrolment in post-secondary non-tertiary education, both sexes (number)
- Enrolment in upper secondary education, both sexes (number)

**Total enrolment in education = Enrolment in tertiary + post-secondary + upper secondary**

- Government expenditure on education as % of GDP (%)
- GDP per capita, PPP (current international \$)
- **Population growth (annual %)**



# 3. Traitement des données

a – Approche d'analyse globale

b – Filtrage

c – Exemples

GDP per capita, PPP (current international \$)

Population growth (annual %)

### 3. Traitement des données

#### a – Approche d'analyse globale

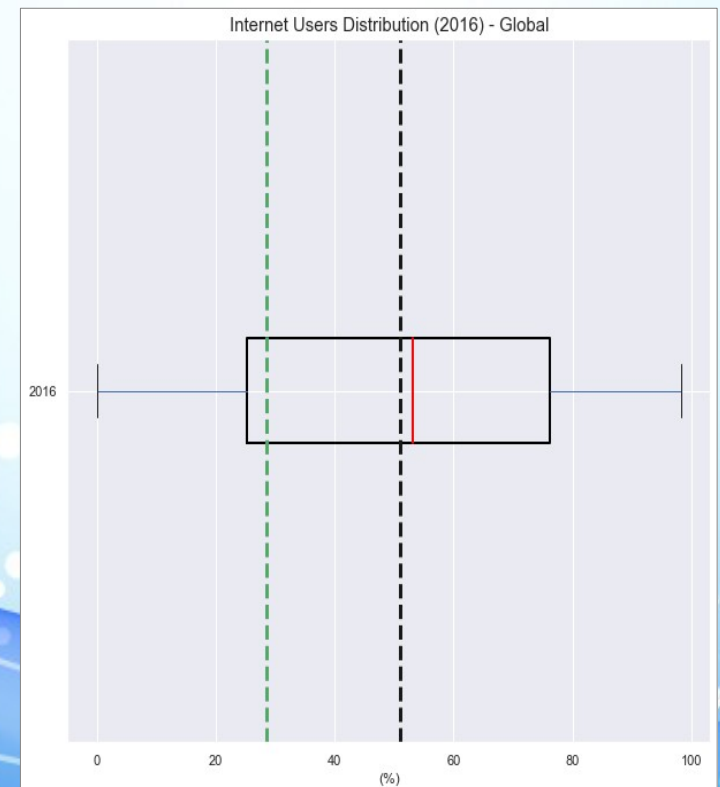
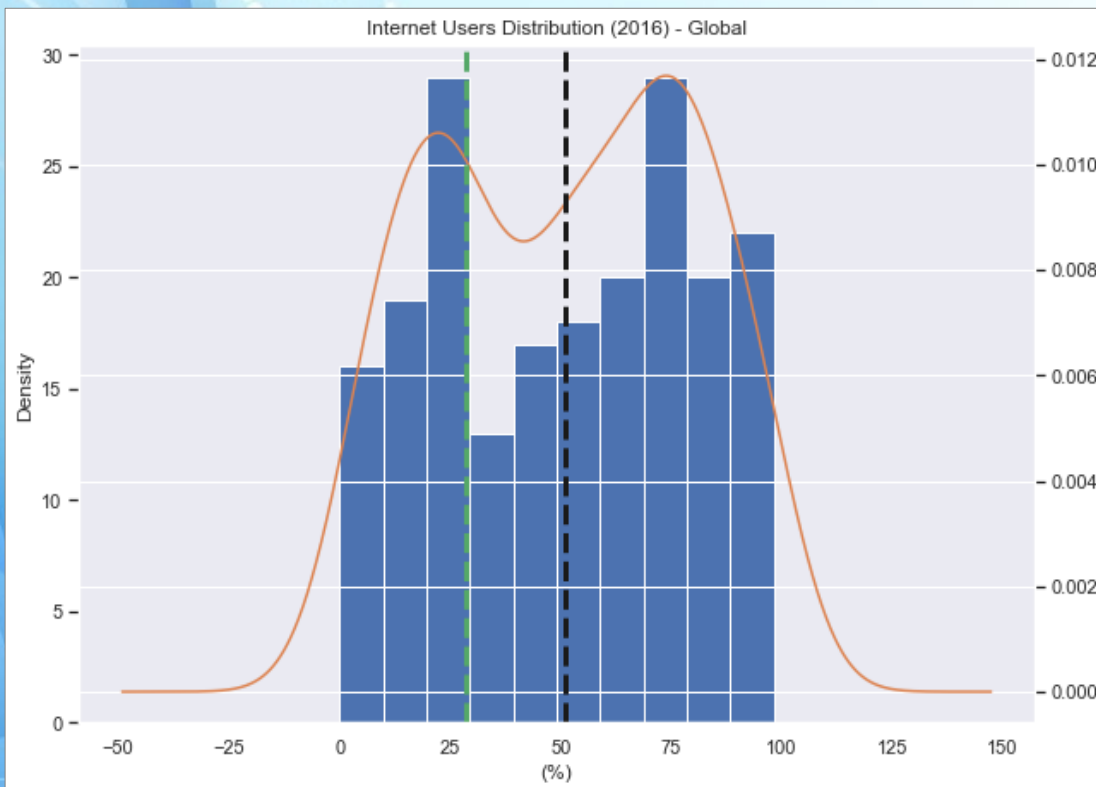
Indicateurs	Restriction temporelle	Clean 1	Clean 2	Filtrage
Internet users (%)	2006-2016	Elimination des rangs peuplés uniquement de valeurs NaN	Imputation par la dernière valeur non-NaN	Oui
Population, ages 15-64 (number)	2010-2015	Elimination des rangs peuplés uniquement de valeurs NaN	Imputation par la dernière valeur non-NaN	Non
Enrolment in tertiary education (total)	2010-2014	Elimination des rangs peuplés uniquement de valeurs NaN	Imputation par la moyenne	Non
Enrolment in post secondary non-tertiary education (total)	2010-2014	Elimination des rangs peuplés uniquement de valeurs NaN	Imputation par la moyenne	Non
Enrolment in upper secondary education (total)	2012-2014	Elimination des rangs peuplés uniquement de valeurs NaN	Imputation par la moyenne	Non
Government expenditure on education (%) of GDP	2005-2015	Elimination des rangs peuplés uniquement de valeurs NaN	Imputation par la moyenne	Non
GDP per capita PPP (\$)	2012-2016	Elimination des rangs peuplés uniquement de valeurs NaN	Imputation par la dernière valeur non-NaN	Oui
Population growth (%)	2012-2016	Elimination des rangs peuplés uniquement de valeurs NaN	Imputation par la moyenne	Non

# 3. Traitement des données

## b – Filtrage

### Variable « Internet users (per 100 people) »

- Selection des pays avec un taux d'utilisateurs supérieur à un certain seuil en 2016
- Seuil (22) = Moyenne (51) – Ecart-type (29)
- 203 → 160 pays



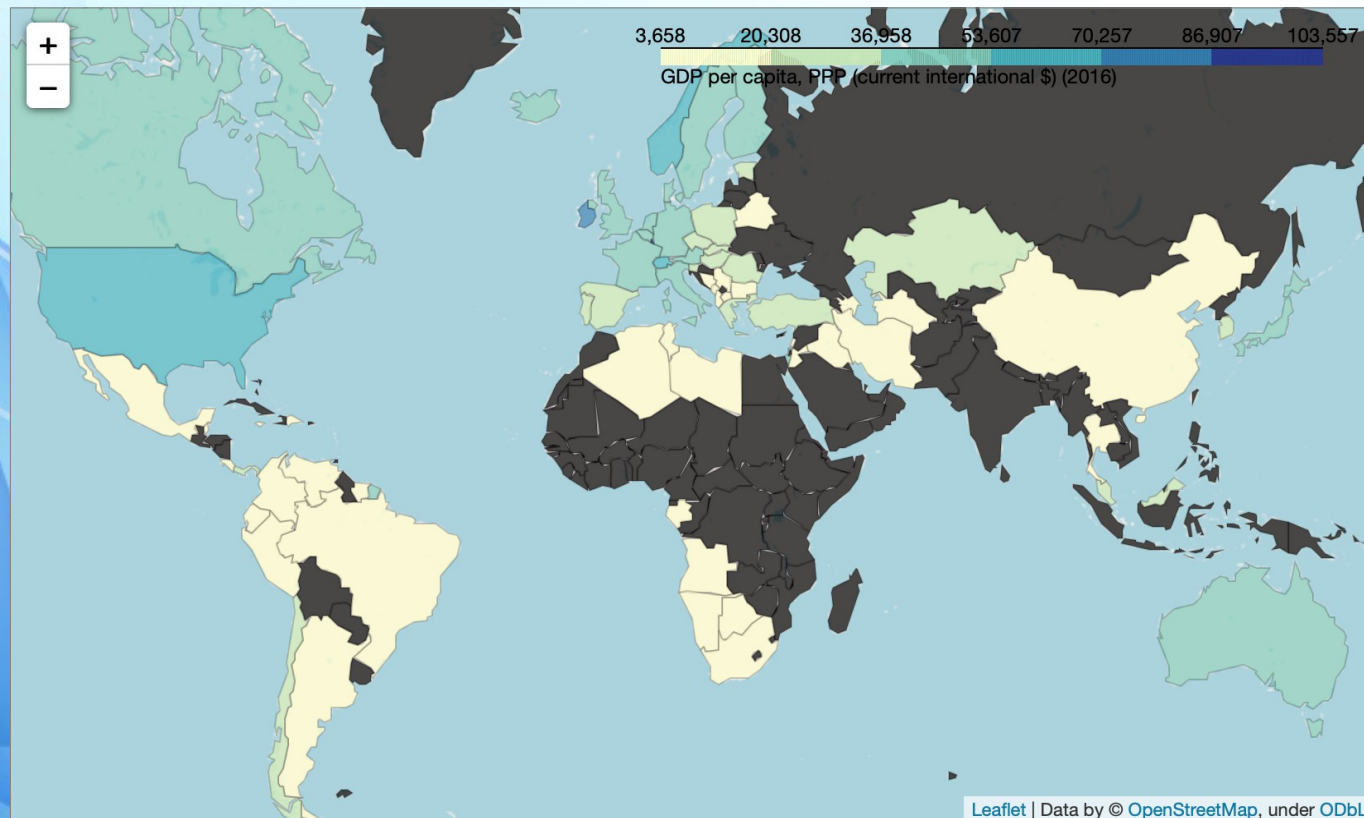


### 3. Traitement des données

#### b – Filtrage

**Variable « GDP per capita, PPP (current international \$) »**

- Income Group : 'High income: nonOECD', 'Low income', 'Upper middle income', 'Lower middle income', 'High income: OECD'
- 192 → 86 pays



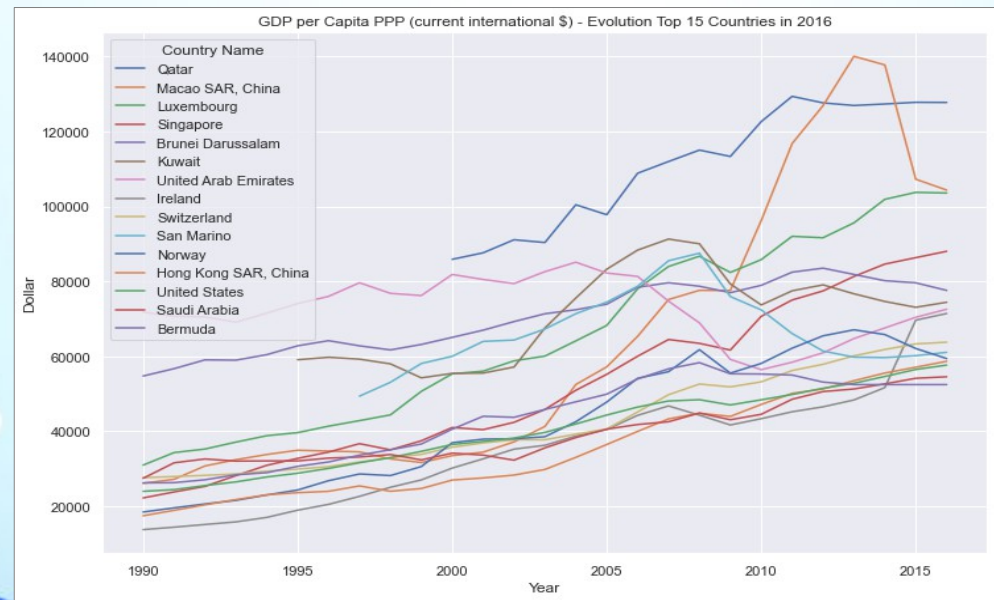
### 3. Traitement des données

#### c – Exemples

« GDP per capita, PPP (current international \$) »



*Valeurs manquantes, vue générale msno.bar()*



*Evolution temporelle des pays Top 15 (2016)*

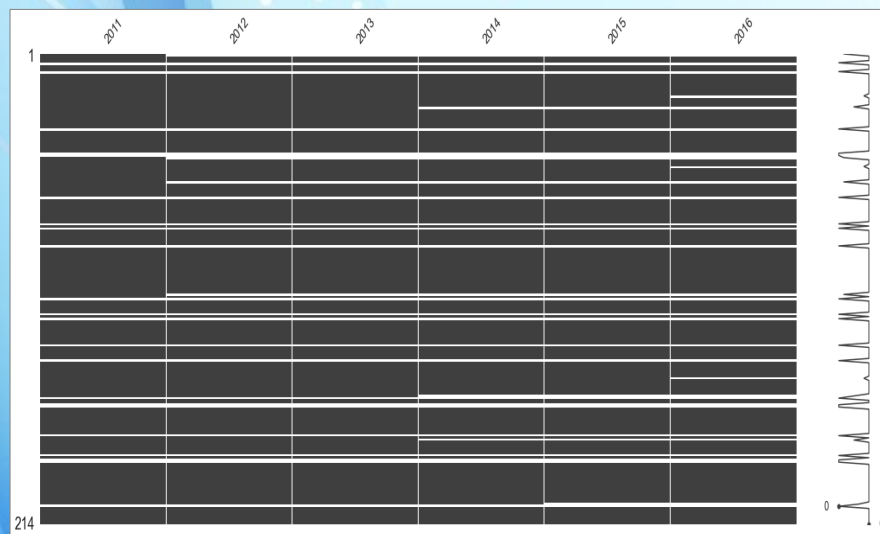
→ L'imputation par la dernière valeur non-NaN est la plus adaptée

### 3. Traitement des données

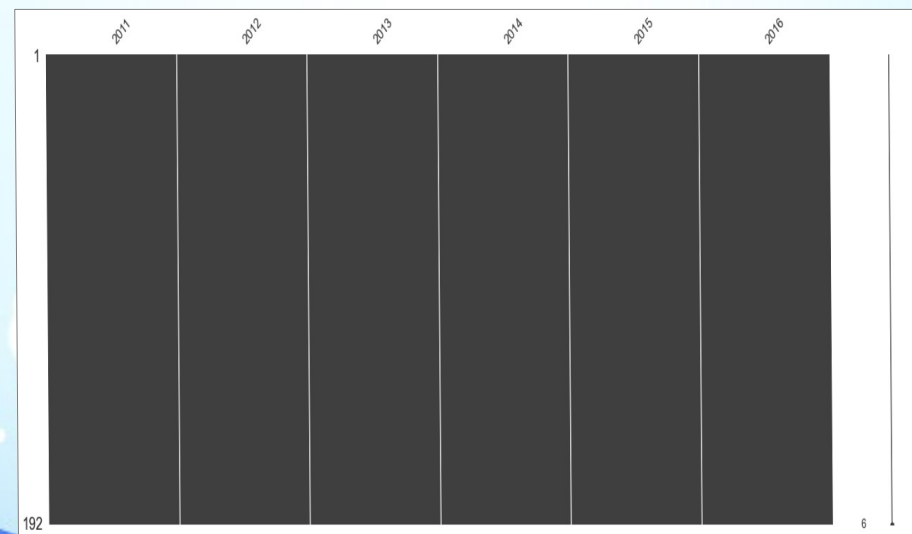
#### c – Exemples

##### « GDP per capita, PPP (current international \$) »

- Filtrage temporel 2011 – 2016
- Clean 1 – Elimination des rangs peuplés uniquement de valeurs NaN de 2006 à 2016
- Clean 2 - Imputation par la dernière valeur non-NaN pour chaque valeur NaN restante



*Jeu de données initial*



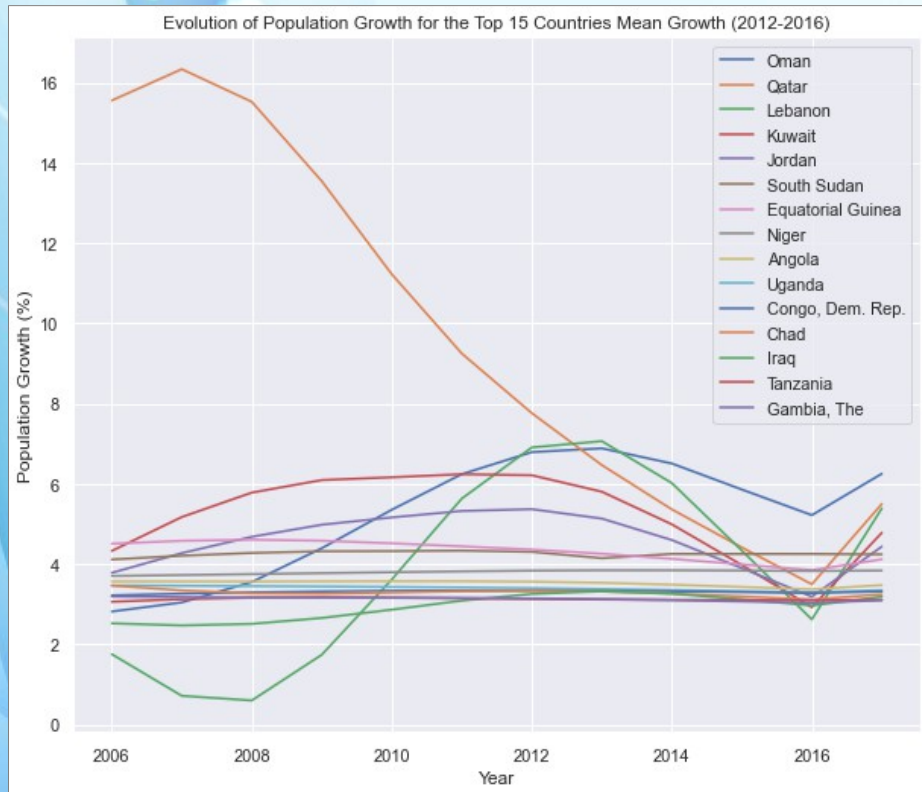
*Jeu de données final*



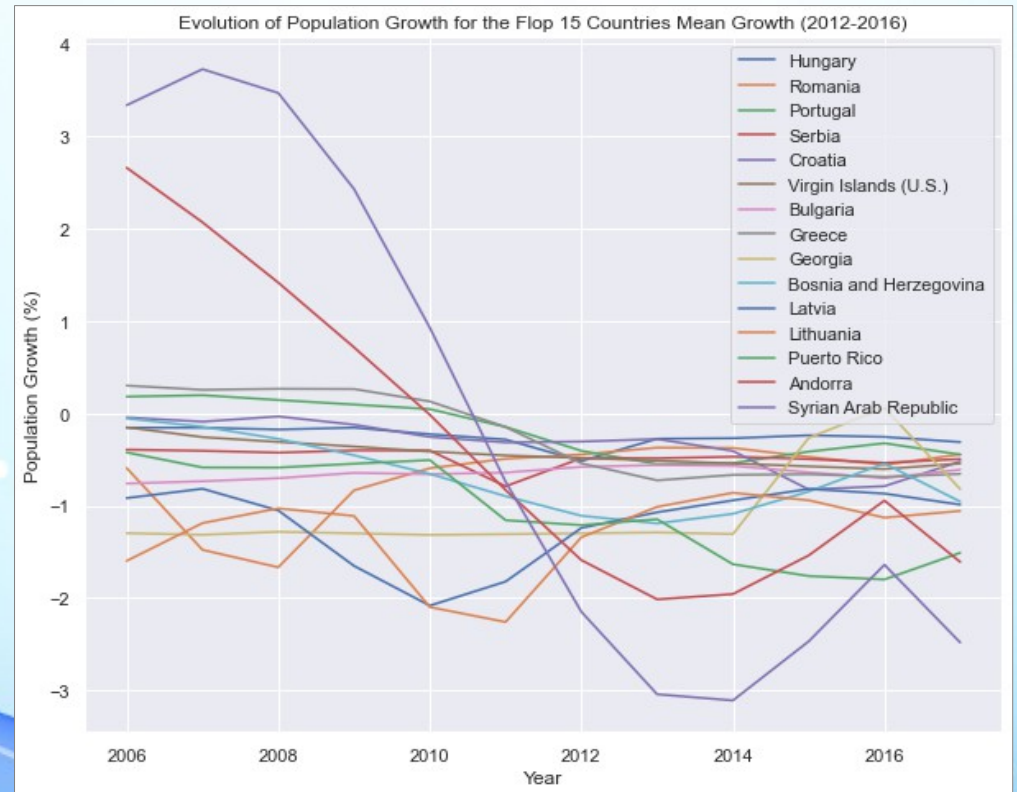
### 3. Traitement des données

#### c – Exemples

##### « Population growth (annual %) »



*Top 20 (Moyenne 2012-2016)*

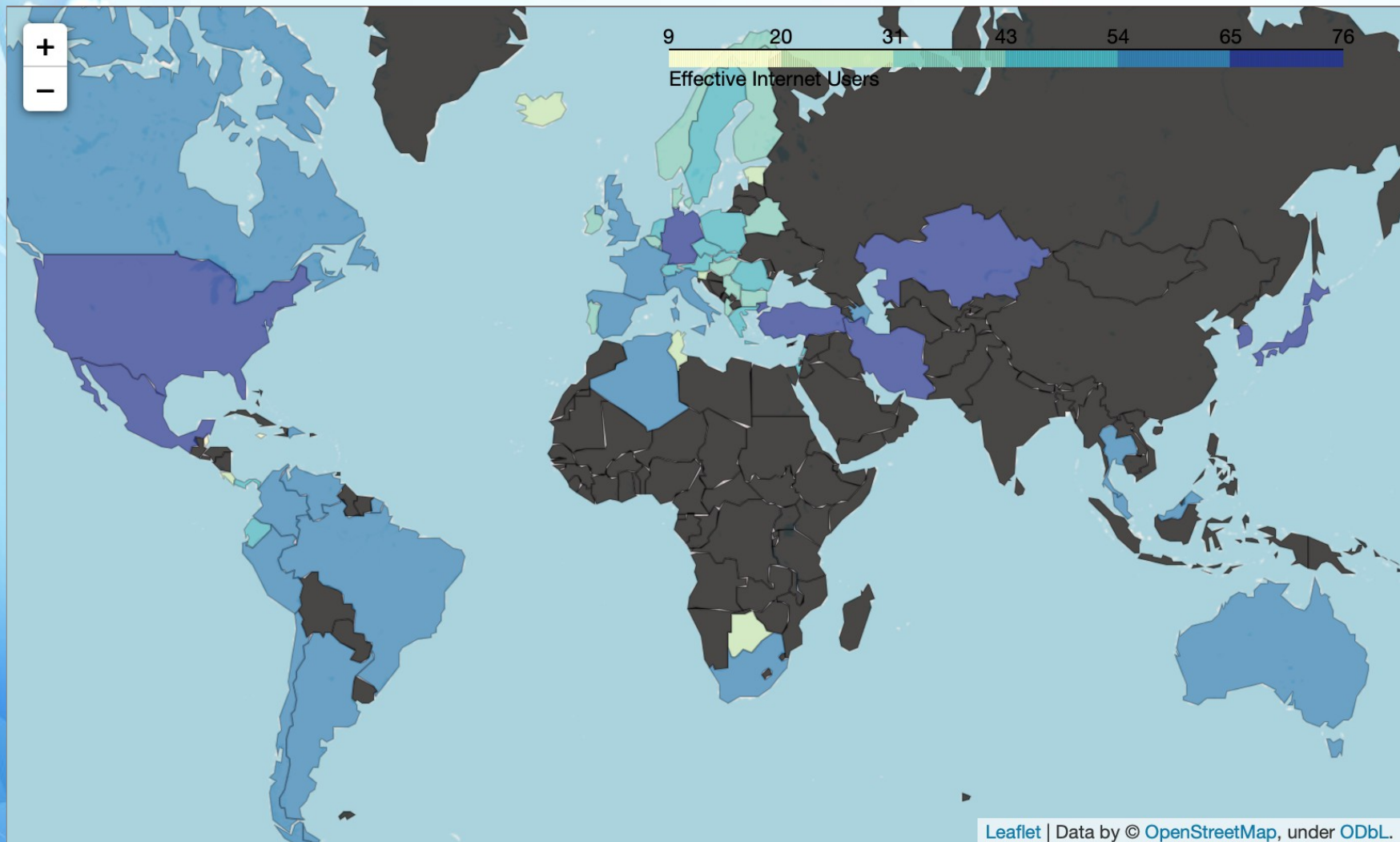


*Flop 20 (Moyenne 2012-2016)*

→ L'imputation par la moyenne est la plus adaptée

## 4. Score d'attractivité

a – **66 pays** retenus après filtrage



## 4. Score d'attractivité

### b – Méthode de calcul

Indicateur	Formule	Valeur	Pondération
<b>Effective internet users (nombre)</b>	$\frac{\text{Internet users (\%)}}{\text{Population 15-64 (nombre)}}$	Imputation par la dernière valeur non-NaN (→2016)	33%
<b>Total enrolment in education (nombre)</b>	$\begin{aligned} &\text{Tertiary (nombre)} \\ &+ \\ &\text{Post-secondary non-tertiary (nombre)} \\ &+ \\ &\text{Upper secondary (nombre)} \end{aligned}$	Moyenne (2010-2014)	17%
<b>Gouvernement expenditure on education (%)</b>	$100 - \text{valeur en centile* (Valeur complémentaire)}$	Moyenne (2010-2014)	33%
<b>GDP per capita (\$)</b>	-	Imputation par la dernière valeur non-NaN (→2016)	17%

1 - Remplacer chaque valeur (pays-indicateur) par son équivalent en centile pour l'indicateur correspondant (Normalisation /100)

2 - Normalisation du score d'attractivité (/150 → /100)

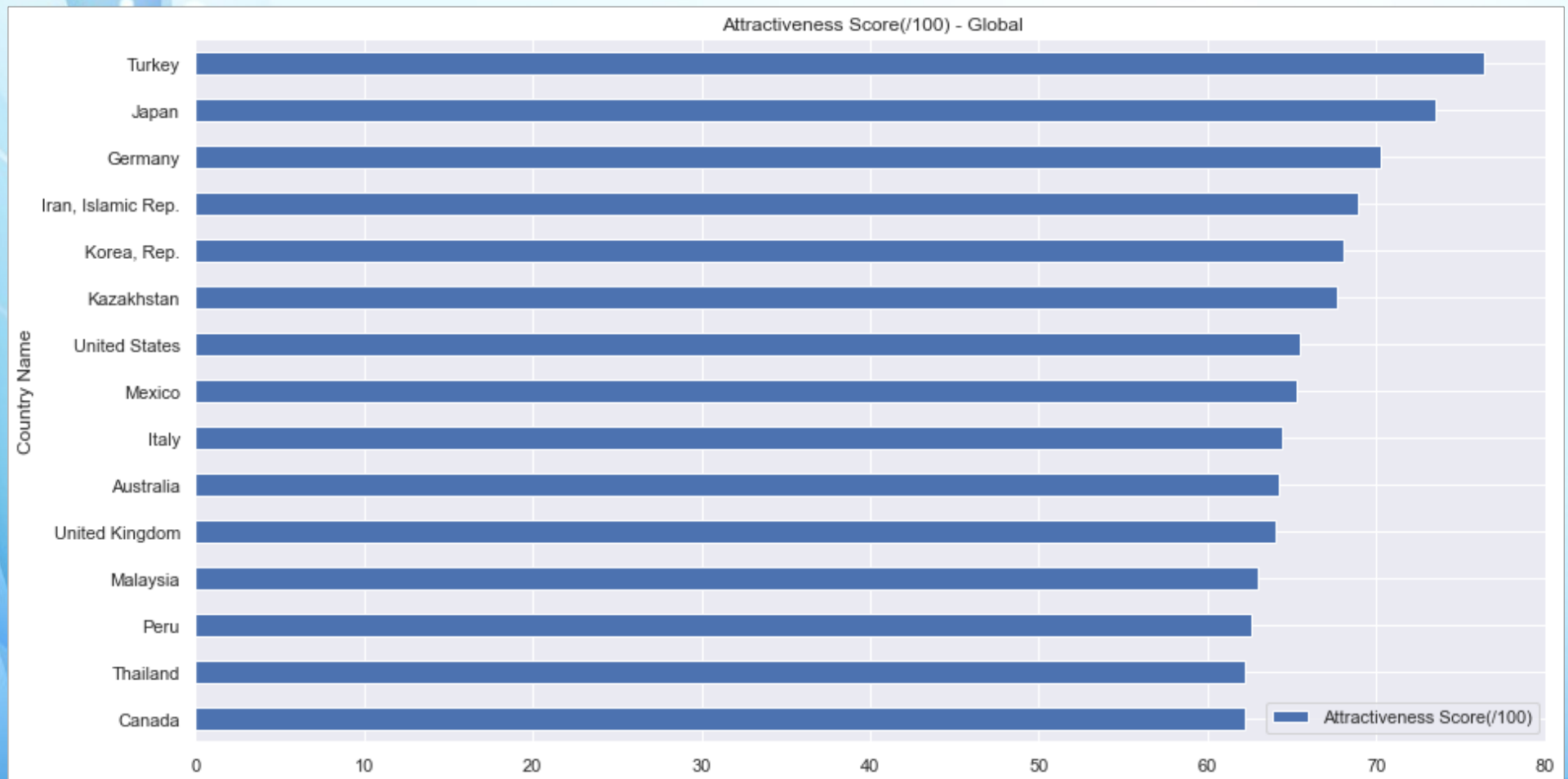
→ **66 pays finaux**

\*Après calcul du rang en centile



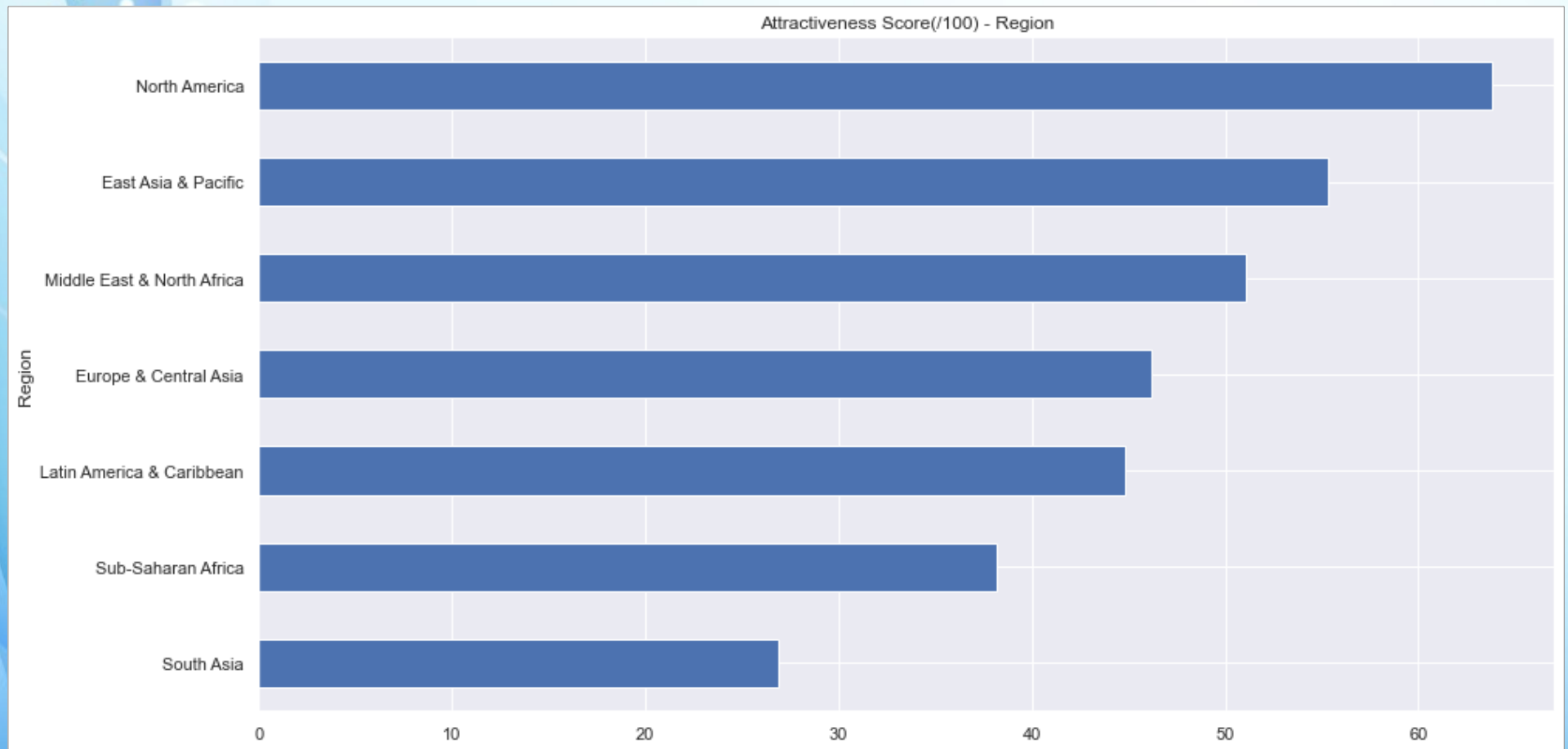
## 4. Score d'attractivité

c – Pays potentiels - Top 15



## 4. Score d'attractivité

c – Régions potentielles



## 4. Score d'attractivité

Discussion :

+

Principe de comparaison synthétique

Score normalisé évocateur

-

Sensible aux changements d'indicateurs

Pondération aléatoire

Assez discriminant ?



# 5. Conclusion

Nombreuses indicateurs non renseignées pourtant utiles !

## 06. Teachers

« Percentage of qualified teachers in upper secondary education, both sexes (%) »

« Proportion of teachers with the minimum required qualifications in upper secondary education, both sexes (%) »

## 03. Secondary

« Proportion of upper secondary schools with access to Internet for pedagogical purposes (%) »

## 11. Technology Skills

« Proportion of youth and adults who have found, downloaded, installed and configured software, both sexes (%) »

Données disponibles jusqu'en 2016 seulement

Certains pays importants éliminés au cours des merge (Chine, Thaïlande)

Peu d'informations sur l'entreprise **academy**

- Concurrents potentiels dans les pays proposés ?
- Présence de l'entreprise à l'étranger ?

Merci de votre attention

**Questions ?**