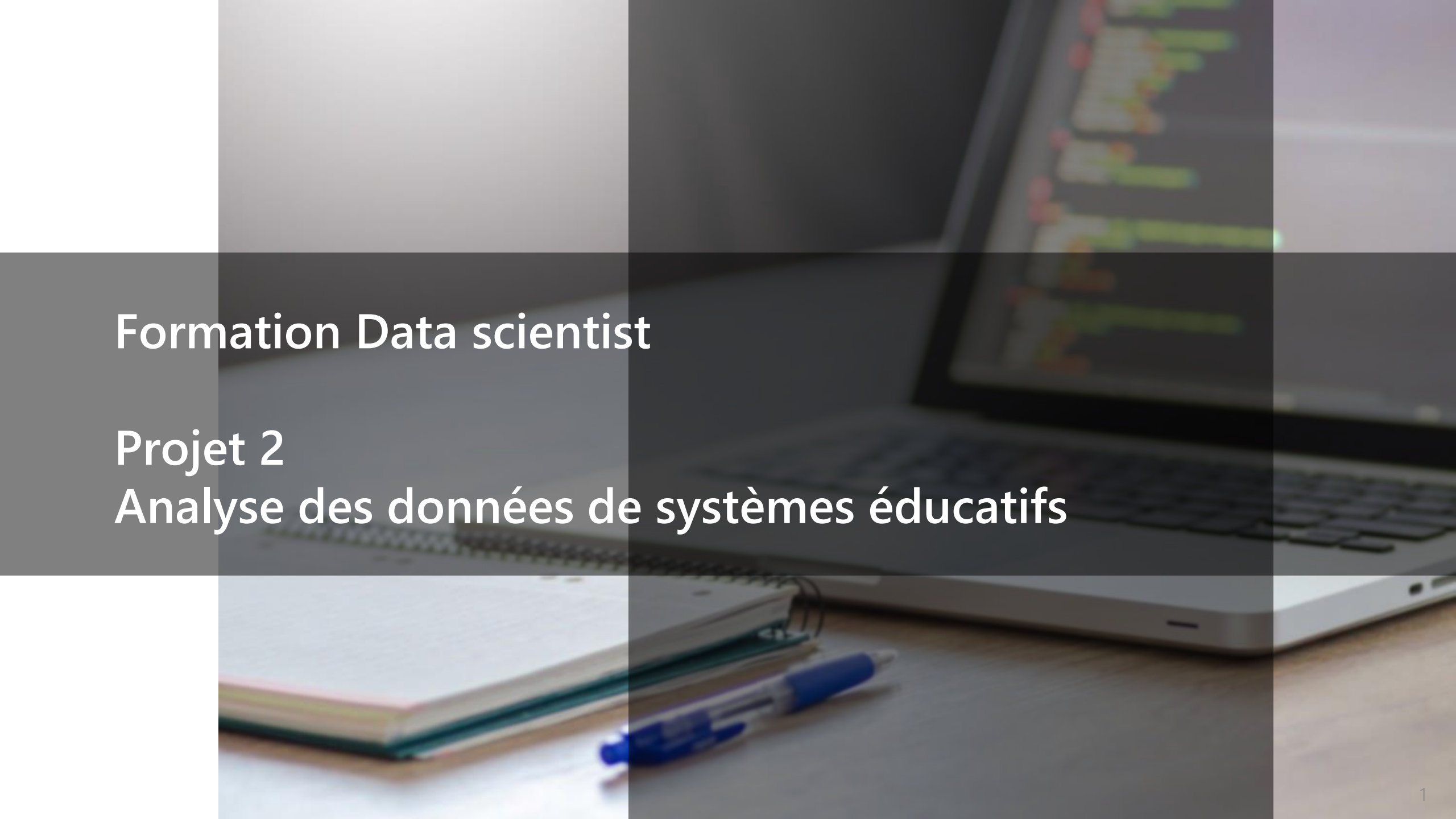


# Formation Data scientist

# Projet 2
# Analyse des données de systèmes éducatifs

## Sommaire

## Qualité du jeu de données

Description du jeu de données

Sélection des informations pertinentes pour la problématique

Ordres de grandeur statistique

## Country series

613 lignes, 4 colonnes

Toute la colonne ‘Unnamed: 3’ contient des valeurs nulles soit 25% du jeu de données.

Aucun doublon

## Country

241 lignes, 32 colonnes

2,354 données sont manquantes soit 30% des données totales.

Aucun doublon

## Data

886,930 lignes, 70 colonnes

53,455,179 données sont manquantes soit 86% des données totales.

Aucun doublon

## Foot note

643,638 lignes, 5 colonnes

Toute la colonne ‘Unnamed: 4’ contient des valeurs nulles soit 20% du jeu de données.

Aucun doublon

## Series

3,665 lignes, 21 colonnes

55,203 données sont manquantes soit 72% des données totales.

Aucun doublon

**83%** soit le nombre de données manquantes.

Semble traduire une mauvaise qualité des données.

Tendance à confirmer après sélection des données pertinentes.

Qualité du jeu de données

**Description du jeu de données**

Sélection des informations pertinentes pour la problématique

Ordres de grandeur statistique

## Country series

Donne des informations sur les sources des données, ce par pays et indicateur.

## Country

Donne des informations démographiques et économiques, par pays et par régions du monde

## Data

Expose des données concernant l'éducation de la population par pays et par région du monde.

3665 indicateurs sont référencés.

## Foot note

Donne des informations sur l'incertitude liées aux donnés.

## Series

Donne des informations sur les indicateurs utilisés (source et définition notamment).

Sur les 5 jeux de données présentés, seul 2 vont nous intéresser directement :

**Country et Data**.

Ils ont tous les deux pour clef les pays et régions du monde ce qui va faciliter leur éventuelle jonction.

Qualité du jeu de données

Description du jeu de données

**Sélection des informations pertinentes pour la problématique**

Ordres de grandeur statistique

# Le jeu de données Country

Le tri s'effectue au niveau des colonnes. Nous ne gardons que celles concernant la **région** à laquelle appartient le pays et son **groupe de revenu** divisé en 5 catégories :

- Revenu faible
- Revenu moyen-faible
- Revenu moyen-supérieur
- Revenu haut, hors pays de l'OCDE
- Revenu haut, pays de l'OCDE

| | Country Code | Table Name | Region | Income Group |
|---|---|---|---|---|
| 0 | ABW | Aruba | Latin America & Caribbean | High income: nonOECD |
| 1 | AFG | Afghanistan | South Asia | Low income |
| 2 | AGO | Angola | Sub-Saharan Africa | Upper middle income |
| 3 | ALB | Albania | Europe & Central Asia | Upper middle income |
| 4 | AND | Andorra | Europe & Central Asia | High income: nonOECD |

# Le jeu de données Data

L'utilisation de **3 mots clefs** nous permet de mettre en avant **12 critères** de comparaisons entre pays.

UPPER SECONDARY

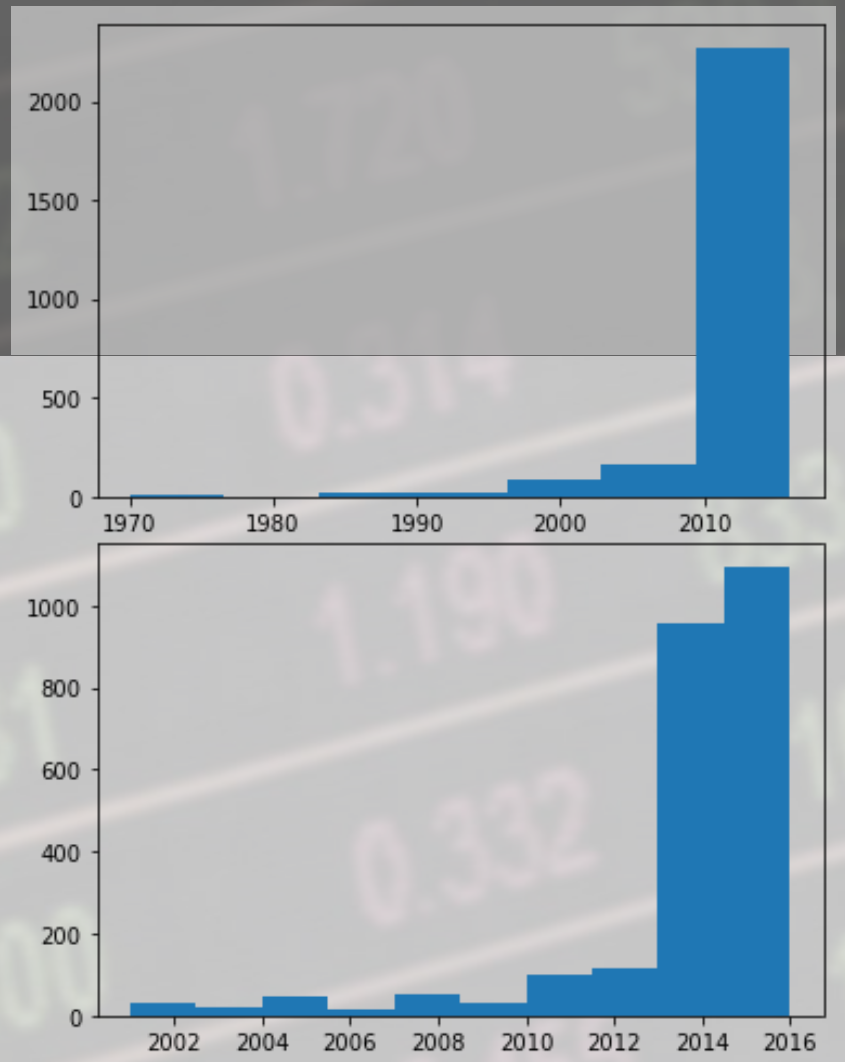
TERTIARY

INTERNET

Nous obtenons un jeu de données de 2662 lignes et 70 colonnes.

```
['Adjusted net enrolment rate, upper secondary, both sexes (%)',
 'Teachers in upper secondary education, both sexes (number)',
 'Enrolment in upper secondary education, both sexes (number)',
 'Rate of out-of-school youth of upper secondary school age, both sexes (%)',
 'Population of the official age for upper secondary education, both sexes (number)',
 'Gross enrolment ratio, upper secondary, both sexes (%)',
 'Enrolment in tertiary education, all programmes, both sexes (number)',
 'Gross enrolment ratio, tertiary, both sexes (%)',
 'Gross enrolment ratio, primary to tertiary, both sexes (%)',
 'Population of the official age for tertiary education, both sexes (number)',
 'Teachers in tertiary education programmes, both sexes (number)',
 'Internet users (per 100 people)']
```

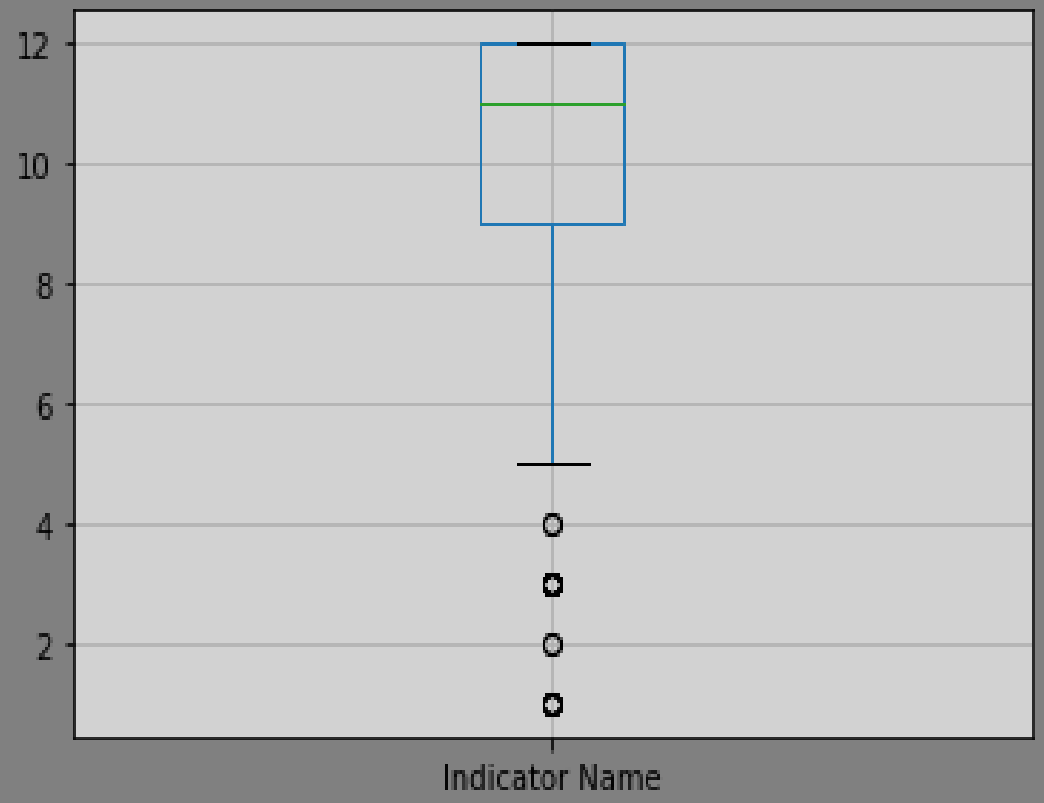
# Le jeu de données Data




Nous remettons en **une dimension** et gardons pour chaque paire pays-indicateur la **valeur la plus récente**.

Au vu de la distribution des données dans le temps, nous gardons les données des seules **10 dernières années** afin d'éviter de baser notre analyse sur des **données obsolètes**.

# Le jeu de données Data




On se retrouve finalement avec un jeu de données de 1876 lignes et 6 colonnes.

Concernant **236 pays**.

Avec **75%** de ces pays ayant au moins **9** des **12** critères retenus.

Les données concernant notre problématique apparaissent maintenant relativement qualitatives.

# La jonction des données

```
Country Name          0
Country Code          0
Indicator Name        0
Indicator Code        0
year                  0
Value                 0
Table Name            7
Region              290
Income Group        290
dtype: int64
```

Jonction gauche avec comme table principale la table Data.

- 7 lignes ayant la colonne « table name » vide (données manquantes pour les îles vierges britanniques)
- 283 lignes ayant les colonnes « Income group » et « Region » vides correspondant aux ensembles géographiques regroupés sous la colonne « Country name ».

Suppression des données manquantes et des pays pour lesquels nous avons moins de 6 indicateurs.

Le jeu de données fait **1862 lignes, 6 colonnes** et concerne **173 pays** avec 0 données manquantes.

| | country_name | indicator_name | year | Value | Region | income_group |
|---|---|---|---|---|---|---|
| 0 | Montenegro | Enrolment in tertiary education, all programme... | 2010 | 23786.000000 | Europe & Central Asia | Upper middle income |
| 1 | Montenegro | Gross enrolment ratio, tertiary, both sexes (%) | 2010 | 55.344589 | Europe & Central Asia | Upper middle income |
| 2 | Montenegro | Gross enrolment ratio, primary to tertiary, bo... | 2010 | 88.662643 | Europe & Central Asia | Upper middle income |
| 3 | Mauritius | Teachers in tertiary education programmes, bot... | 2010 | 1100.000000 | Sub-Saharan Africa | Upper middle income |
| 4 | Nigeria | Teachers in upper secondary education, both se... | 2010 | 112840.000000 | Sub-Saharan Africa | Lower middle income |

Qualité du jeu de données

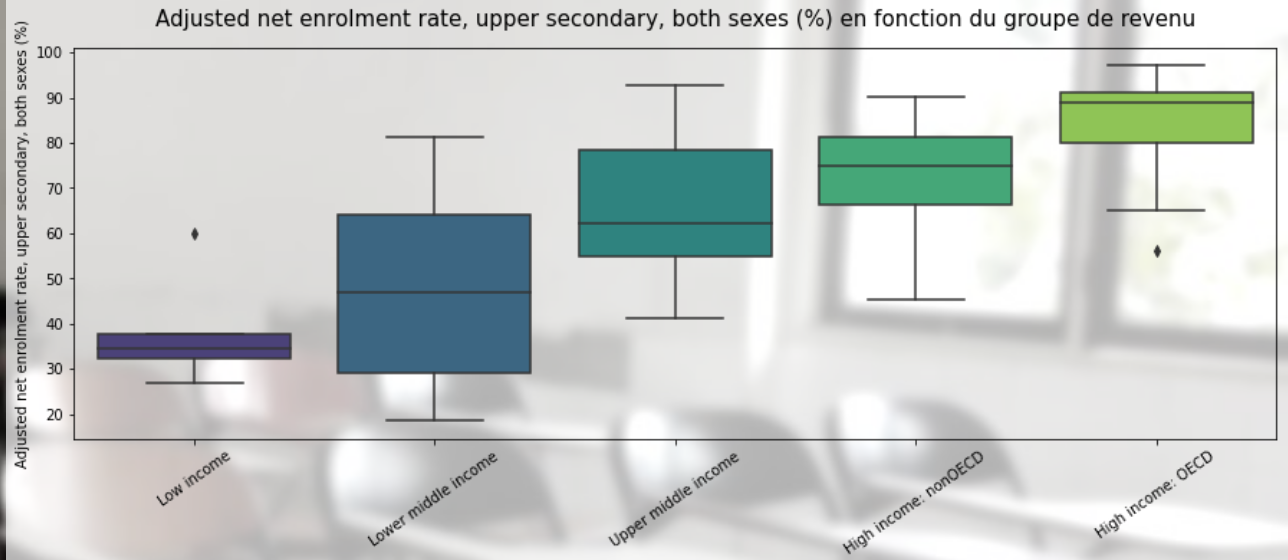
Description du jeu de données

Sélection des informations pertinentes pour la problématique

**Ordres de grandeur statistique**

# Corrélation entre le groupe de revenu et le taux d'inscription au lycée

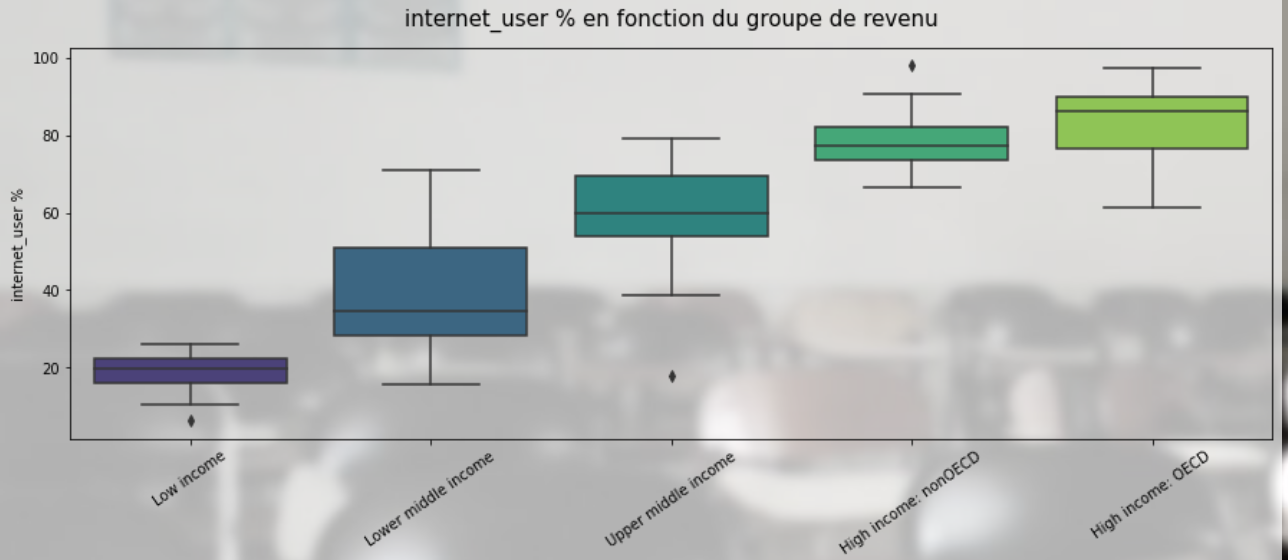
| income_group | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| High income: OECD | 25.0 | 83.546513 | 9.967533 | 56.299019 | 79.888199 | 85.990601 | 90.916893 | 97.071823 |
| High income: nonOECD | 22.0 | 72.827195 | 13.604522 | 43.202511 | 66.048872 | 73.007351 | 83.074898 | 90.123840 |
| Low income | 20.0 | 18.988762 | 14.931107 | 3.523080 | 6.807980 | 17.204289 | 28.177495 | 59.921841 |
| Lower middle income | 35.0 | 42.529119 | 22.808712 | 8.744770 | 25.007629 | 36.029129 | 61.765755 | 83.718147 |
| Upper middle income | 36.0 | 60.155881 | 20.757144 | 3.415430 | 50.048023 | 57.811029 | 78.014101 | 92.717438 |




R2 = 0,58

Le group de revenu du pays explique pour 58% le taux d'inscription au lycée.

# Corrélation entre le groupe de revenu et l'accès à internet



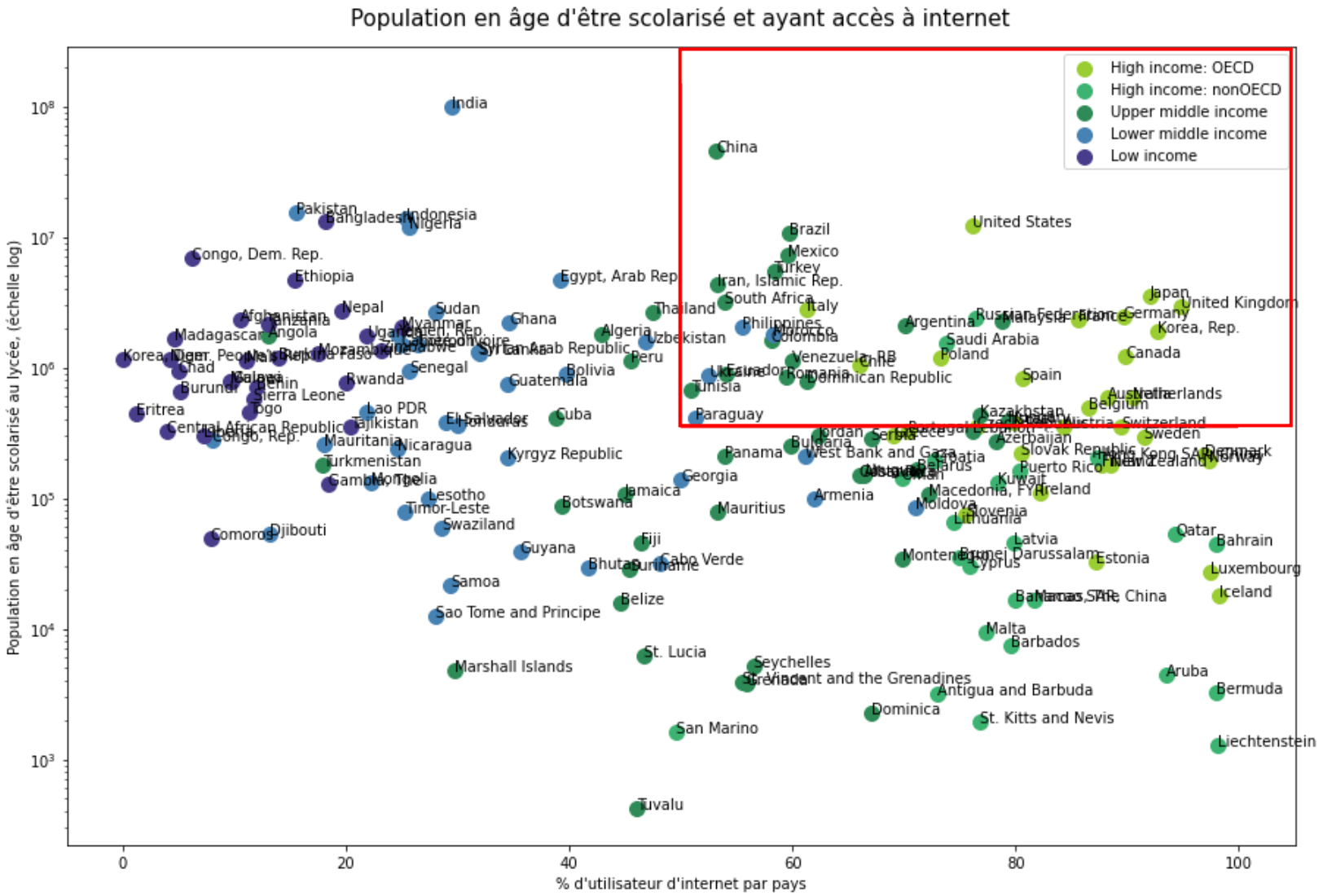

R2 = 0,82

Le groupe de revenu du pays explique pour 82% le taux d'accès à internet de la population du pays.

| income_group | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| High income: OECD | 31.0 | 84.506555 | 9.723361 | 61.324253 | 78.129996 | 87.237332 | 90.958894 | 98.240016 |
| High income: nonOECD | 24.0 | 79.585563 | 11.182103 | 49.600000 | 74.220210 | 77.828389 | 83.057325 | 98.093904 |
| Low income | 29.0 | 12.030246 | 6.900800 | 0.000000 | 6.209974 | 11.310000 | 18.246938 | 25.073304 |
| Lower middle income | 41.0 | 34.441505 | 14.273726 | 8.121949 | 25.366301 | 29.547163 | 41.772645 | 70.999999 |
| Upper middle income | 46.0 | 55.808429 | 14.514105 | 13.000000 | 46.562503 | 56.185558 | 66.279762 | 79.259401 |

# Etablissement d'une présélection

## Sélection des variables

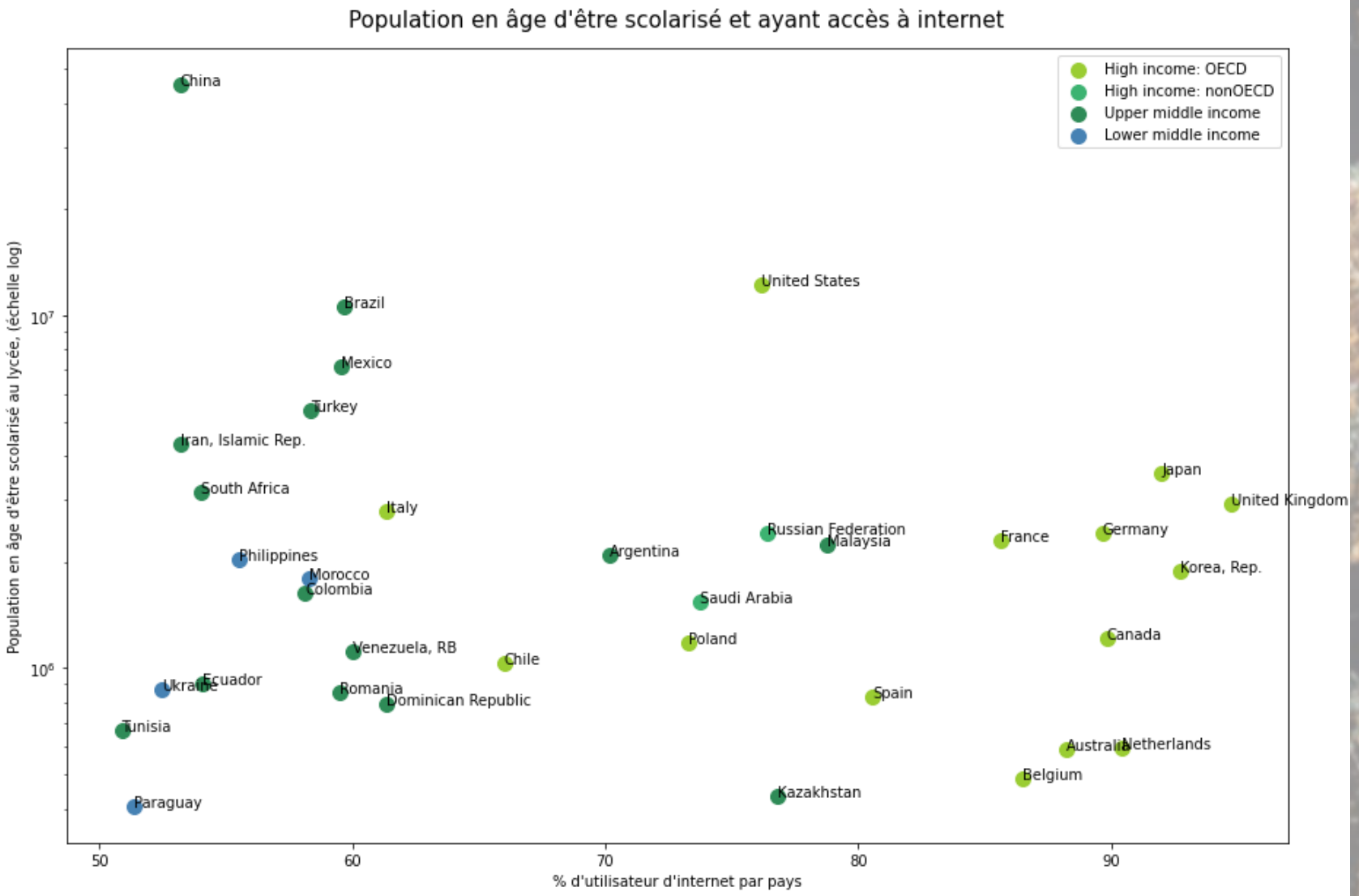



Nuage de points

- Abscisse : Taux d'utilisateur d'internet
- Ordonnée : population en âge d'être scolarisé au lycée (nombre et échelle logarithmique)

# Etablissement d'une présélection

## Représentation avec groupe de revenu

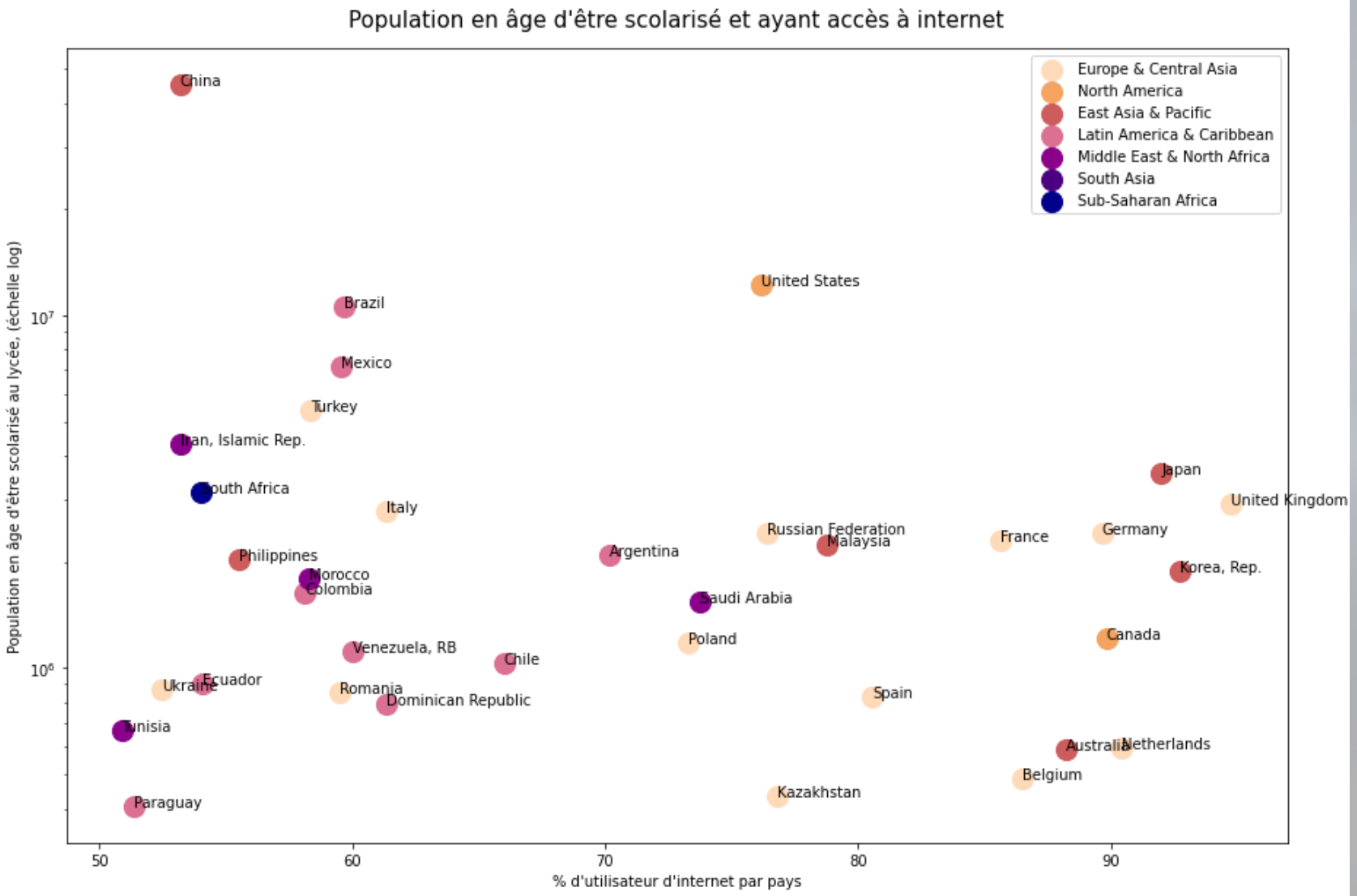



3 critères :

- Population en âge d'être en étude supérieur plus grand que la médiane (moins sensible aux outliers que sont l'Inde et la Chine)
- Population en âge d'être au lycée supérieur à la médiane
- Taux d'accès à internet supérieur à 50%

| income_group | |
|---|---|
| **All** | 35 |
| **Upper middle income** | 15 |
| **High income: OECD** | 14 |
| **Lower middle income** | 4 |
| **High income: nonOECD** | 2 |

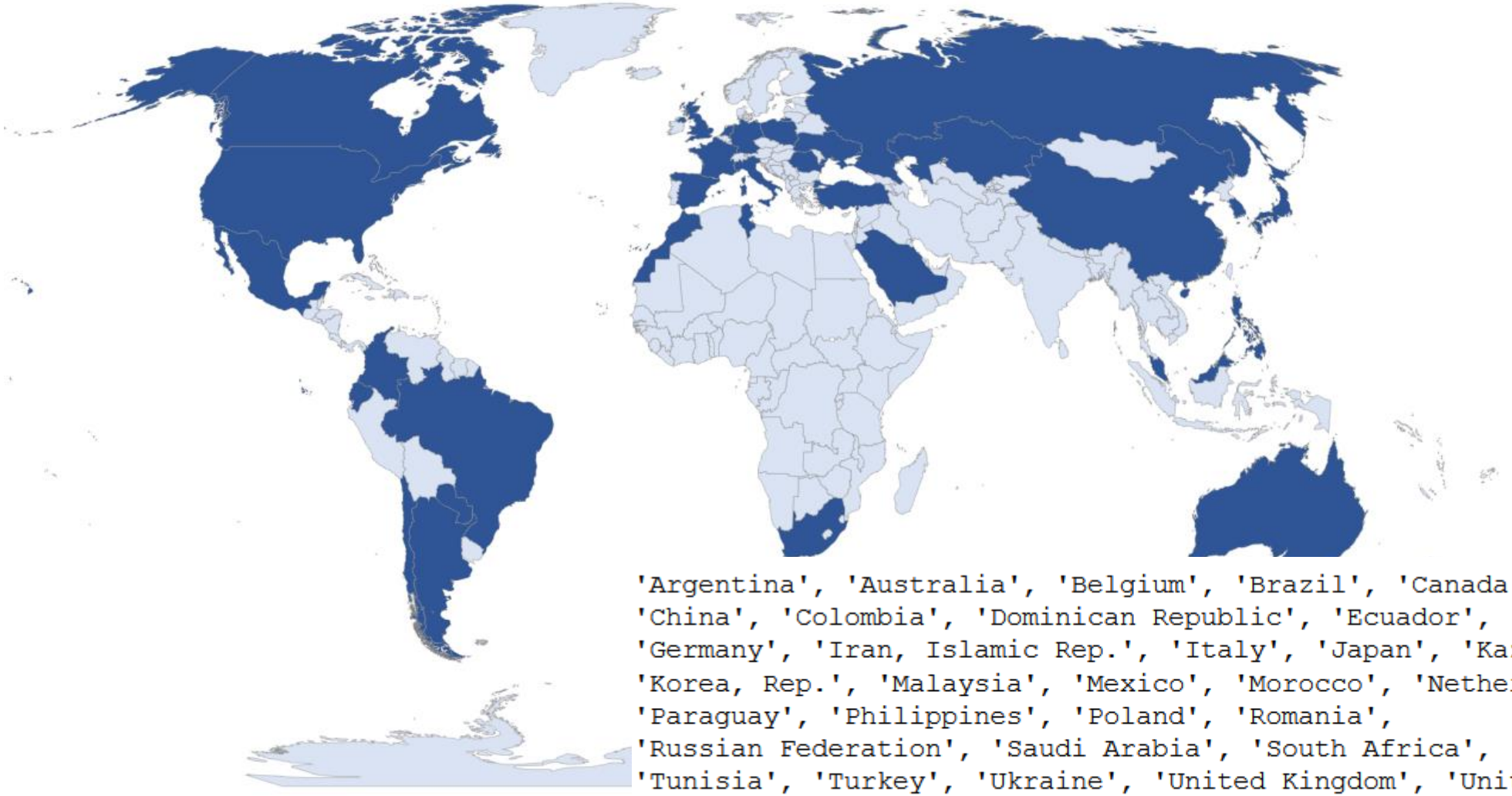
# Etablissement d'une présélection
## Représentation avec origine géographique




| Region | pays_liste | total_pays | % |
|---|---|---|---|
| **All** | 35 | 173 | 20.0 |
| **Europe & Central Asia** | 13 | 49 | 27.0 |
| **Latin America & Caribbean** | 9 | 34 | 26.0 |
| **East Asia & Pacific** | 6 | 22 | 27.0 |
| **Middle East & North Africa** | 4 | 18 | 22.0 |
| **North America** | 2 | 3 | 67.0 |
| **Sub-Saharan Africa** | 1 | 40 | 2.0 |

# Etablissement d'une présélection

## Représentation sur carte du monde



```
'Argentina', 'Australia', 'Belgium', 'Brazil', 'Canada', 'Chile',
'China', 'Colombia', 'Dominican Republic', 'Ecuador', 'France',
'Germany', 'Iran, Islamic Rep.', 'Italy', 'Japan', 'Kazakhstan',
'Korea, Rep.', 'Malaysia', 'Mexico', 'Morocco', 'Netherlands',
'Paraguay', 'Philippines', 'Poland', 'Romania',
'Russian Federation', 'Saudi Arabia', 'South Africa', 'Spain',
'Tunisia', 'Turkey', 'Ukraine', 'United Kingdom', 'United States',
'Venezuela, RB'], dtype=object)
```

# Approche alternative

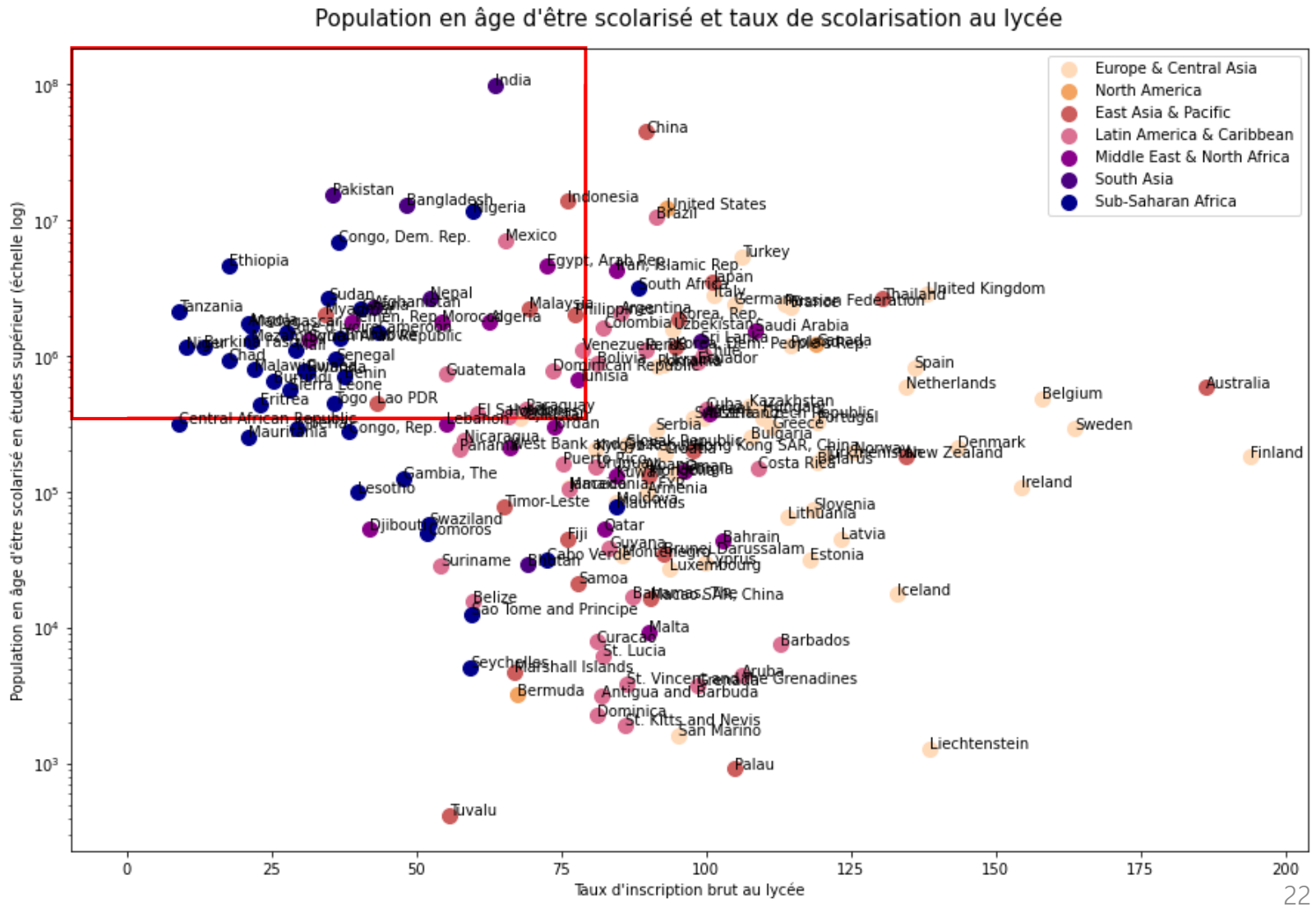
## Population en âge d'être scolarisée et taux de scolarisation

Nuage de points

- Abscisse : Taux d'inscription au lycée

- Ordonnée : population en âge d'être scolarisé au lycée (nombre et échelle logarithmique)

Cible les pays ayant la **marge de progression** la plus importante



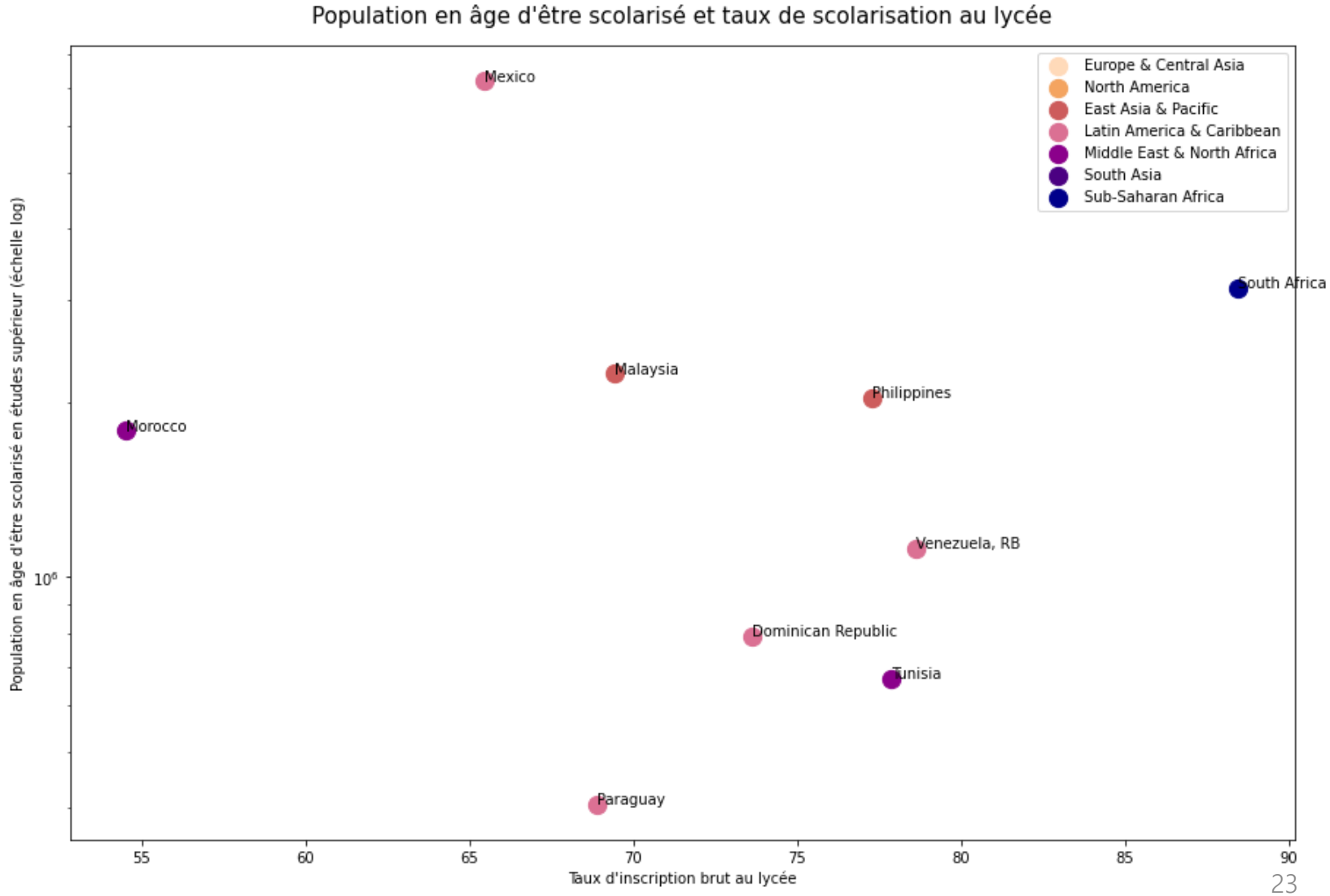
# Approche alternative

## Liste alternative avec représentation géographique

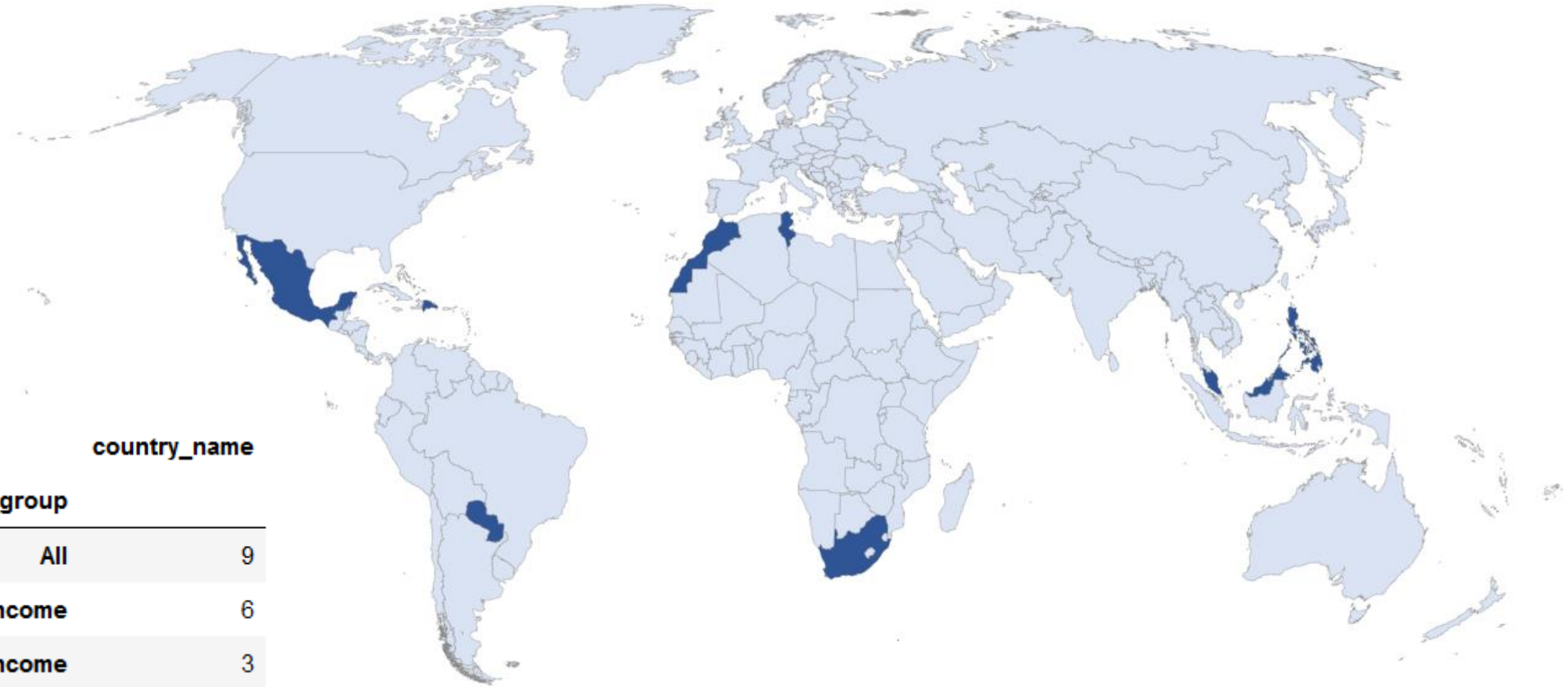
3 critères :

- Population en âge d'être en étude supérieur et au lycée plus grand que la médiane
- Taux d'inscription au lycée et en étude supérieur inférieur à la moyenne
- Taux d'accès à internet supérieur à 50%

| Region | pays_liste | total_pays | % |
|---|---|---|---|
| **All** | 9 | 173 | 5.0 |
| **Latin America & Caribbean** | 4 | 34 | 12.0 |
| **East Asia & Pacific** | 2 | 22 | 9.0 |
| **Middle East & North Africa** | 2 | 18 | 11.0 |
| **Sub-Saharan Africa** | 1 | 40 | 2.0 |



# Approche alternative
## Représentation sur carte du monde



| income_group | country_name |
|---|---|
| All | 9 |
| Upper middle income | 6 |
| Lower middle income | 3 |

```
['Dominican Republic', 'Malaysia', 'Mexico', 'Morocco', 'Paraguay',
 'Philippines', 'South Africa', 'Tunisia', 'Venezuela, RB'],
dtype=object)
```

## Conclusion

Les pays présent dans la deuxième liste le sont également dans la première. Ils apparaissent comme des cibles prioritaires.

Cette liste est susceptible d’être modifiée en fonction de nouveaux critère, notamment linguistiques, qui ici n’ont pas été pris en compte.