# Projet n°2 : “Analysez des données de systèmes éducatifs”

Soutenance de Projet
30 octobre 2019

# Programme

I - Rappel de la problématique et présentation du jeu de données

II - Analyse pré exploratoire

III - Conclusions sur la pertinence du jeu de données

# I Rappel de la problématique et présentation du jeu de données

# Rappel de la problématique

- Academy est une **start-up de la EdTech**
- Elearnings : Contenus de formation de **niveau lycée et université**
- Objectif d'**expansion à l'international**

academy

**Objectif du projet :**

*Informer le projet d'expansion en réalisant une analyse pré exploratoire et déterminer si les données sur l' éducation de la Banque Mondiale conviennent*

BANQUE MONDIALE

# Présentation du jeu de données

BANQUE MONDIALE

| | |
|---|---|
| EdStatsCountry.csv | **Informations globales sur l'économie de chaque pays du monde (et de zones géographiques)**<br>Taille : 241 lignes (1 par pays / zone) , 32 colonnes<br>Quelques valeurs manquantes<br>Aucun doublon |
| EdStatsCountry-Series.csv | **Informations sur la source des données contenues dans EdStatsCountry**<br>Taille : 613 lignes, 4 colonnes<br>Pas de valeur manquante (sauf Unnamed : 3" qui est une colonne uniquement composée de NaN)<br>Aucun doublon |
| EdStatsData.csv | **Donne l'évolution de nombreux indicateurs pour tous les pays et certains groupes de pays**<br>Taille : 886 930 lignes, 70 colonnes<br>données depuis 1970<br>**Nombreuses valeurs manquantes**<br>Aucun doublon |
| EdStatsFootNote.csv | **Contient des Informations sur l'année d'origine des données et les incertitudes sur les données)**<br>Taille : 643 638 lignes, 4 colonnes<br>Pas de valeur manquante (sauf Unnamed : 4 qui est une colonne uniquement composée de NaN)<br>Aucun doublon |
| EdStatsSeries.csv | **Informations sur les indicateurs socio économiques disponibles dans EdStatsData**.<br>Taille : 3665 lignes, 21 colonnes<br>6 colonnes vides pour lesquelles il manque toutes les valeurs.<br>Il manque plus de 80 % des données dans 10 autres colonnes de la table<br>Aucun doublon |

# II Analyse Pré Exploratoire

# Processus d’analyse pré exploratoire

1

**Connaître les données**

Quelles informations?

Quelles années?

2

**Identifier les indicateurs exploitables**

Quantités de données manquantes?

3

**Comparer les pays**

Quels indicateurs choisir?

Analyse des résultats obtenus

Quels sont les pays à cibler par Academy?

4

**Quel est le potentiel pour chaque pays?**

Comment identifier le potentiel des pays choisis?

# Outils utilisés pour l'analyse

| Nom | Utilisation | Fonctions spécifiques |
|---|---|---|
| **Anaconda** | Gestion de package<br>Gestion d'environnement virtuel | Conda : installation de package via le terminal |
| **Jupyter Notebook 6.0.1** | Structurer la démarche<br>Executer code par étape<br>Expliquer la démarche (markdown) | |
| **Python 3.7** | Appel aux librairies,<br>Boucles for pour générer plusieurs graphes | Boucles, Listes, dictionnaires, collections (compteur de mot) |
| **Pandas 0.25.0** | Manipulation de données<br>Représentation des données | Manipulation de Dataframe : création, copie, filtres, tris, description, concaténation, dépivotage |
| **Matplotlib 3.1.0**<br>**Seaborn 0.9.0** | Génération de graphes | Barplot, Scatterplot, lineplot, distplot, heatmap |

# 1 - Connaître les données

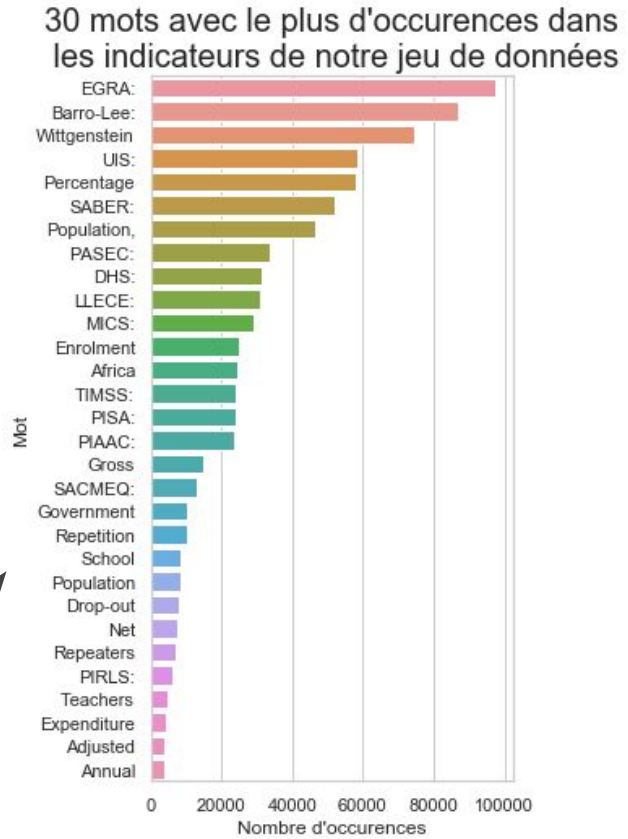
# 1 - Connaître les données - Préambule

Historique et prédictions de 1970 à 2050

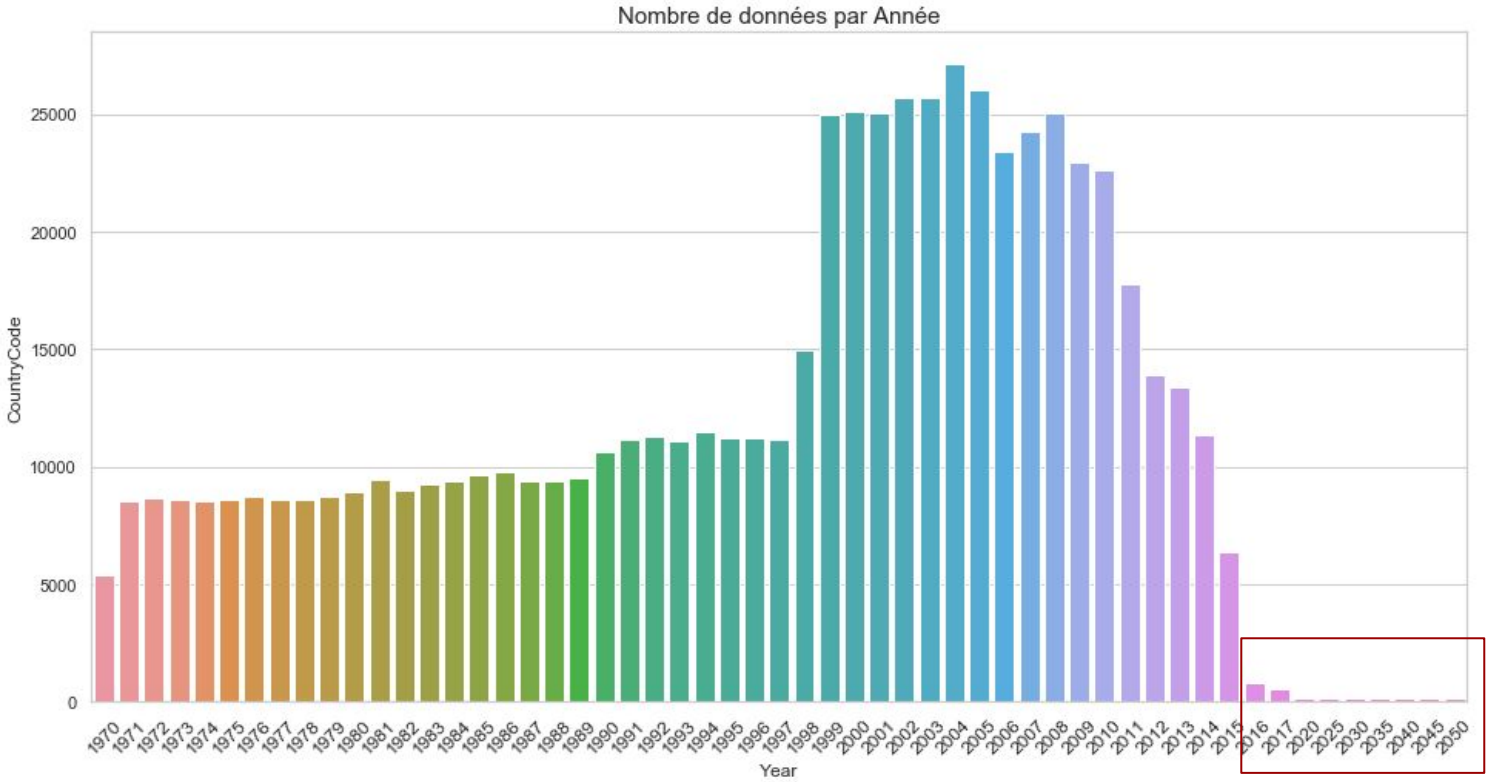
3665 indicateurs uniques

241 zones géographiques (dont pays)

indicateurs relatifs à l' éducation

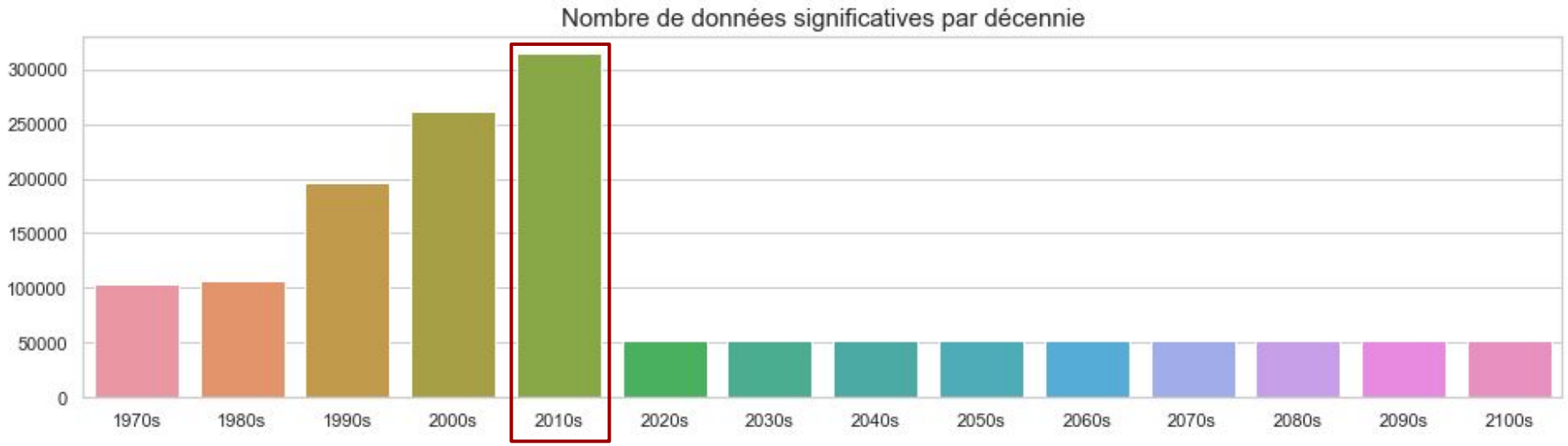


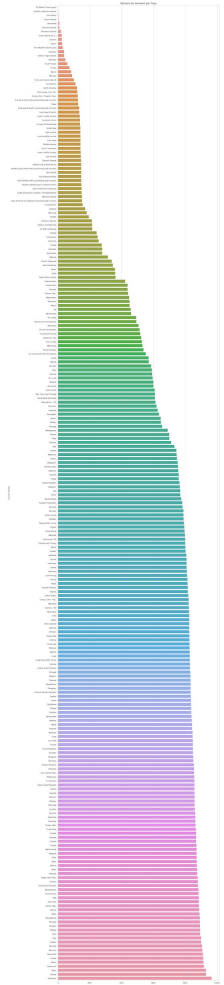
# 1 - Connaître les données - Quantité de données par année



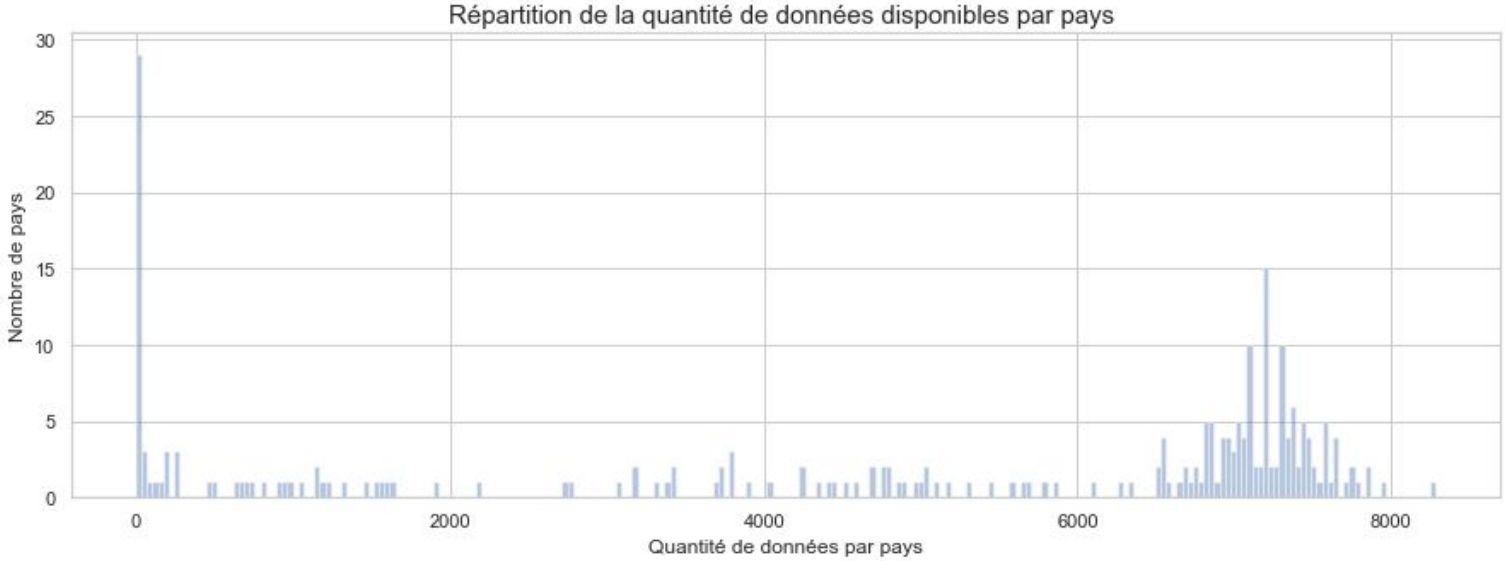
# 1 - Connaître les données - Nombre de données par décennie

```
data['1970s'] = data[[str(year) for year in range(1970,1980,1)]].mean(1)
```





# 1 - Connaître les données - Inégalité du nombre de données par pays




Constat : Inégalité de répartition des données par pays.
Moins d'information pour ~30 % des zones:

- les "petits pays";
- les nouveaux pays (Kosovo);
- les régions et groupes de pays (East Asia & Pacific, Upper Middle Income, etc.)

# 1 - Connaître les données - Quelles informations conserver?

Après analyse des colonnes de chaque partie du jeu de données:

- EdstatsCountry : l'association pays-régions

```
data = data.merge(right = country[['Country Code', 'Region']],
                  on='Country Code', how='left')
```

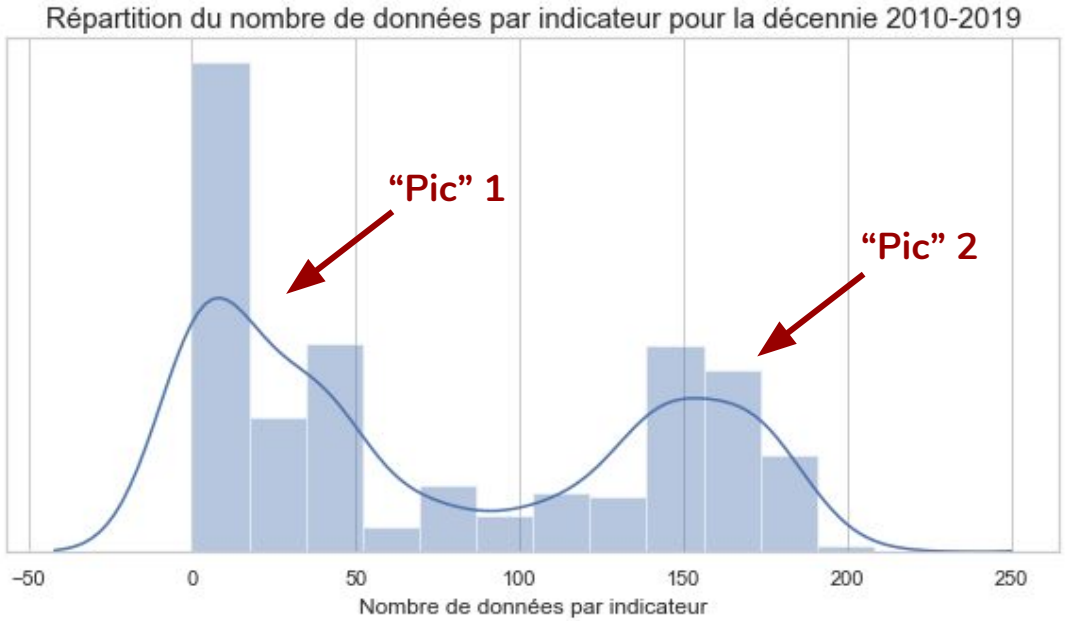
- EdstatsData : les noms de pays, d'indicateur, les valeurs pour la **décennie 2010**

```
data_short = data[['Country Name', 'Country Code', 'Indicator Name',
                   'Indicator Code', '2010s', 'Region']]
```

- Autres données : non nécessaires à ce stade.

# 2 - Identifier les indicateurs exploitables

# 2 - Identifier les indicateurs exploitables - Les données manquantes (NaN)



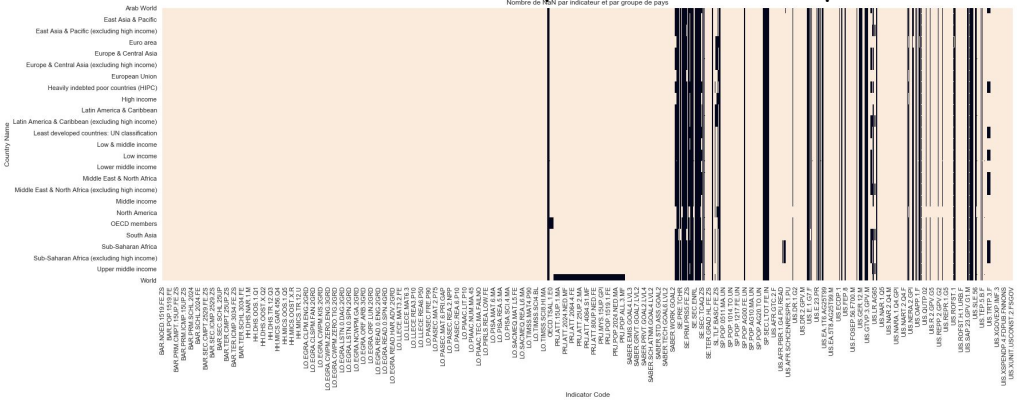

Intuition : Les indicateurs seraient-ils répartis selon la somme de 2 distributions : Zones géographiques (Pic 1) + Pays (Pic 2) ?

*NB : 197 États reconnus par l’ONU, 22 “zones” dans notre dataset*

# 2 - Comparer les indicateurs - Identifier les NaN graphiquement (noir = donnée manquante)

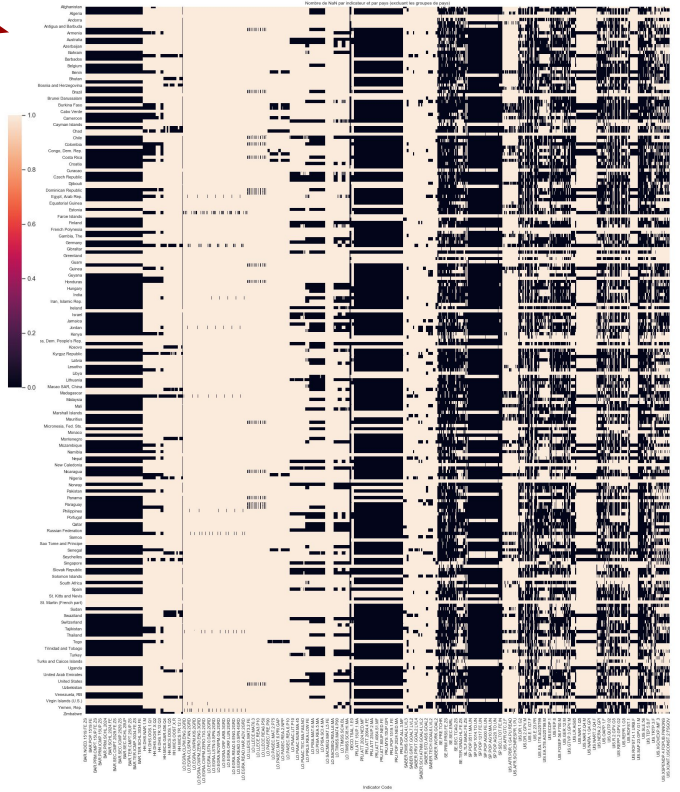
Par zone

Données par indicateur et par zone




Par pays

Données par indicateur et par pays

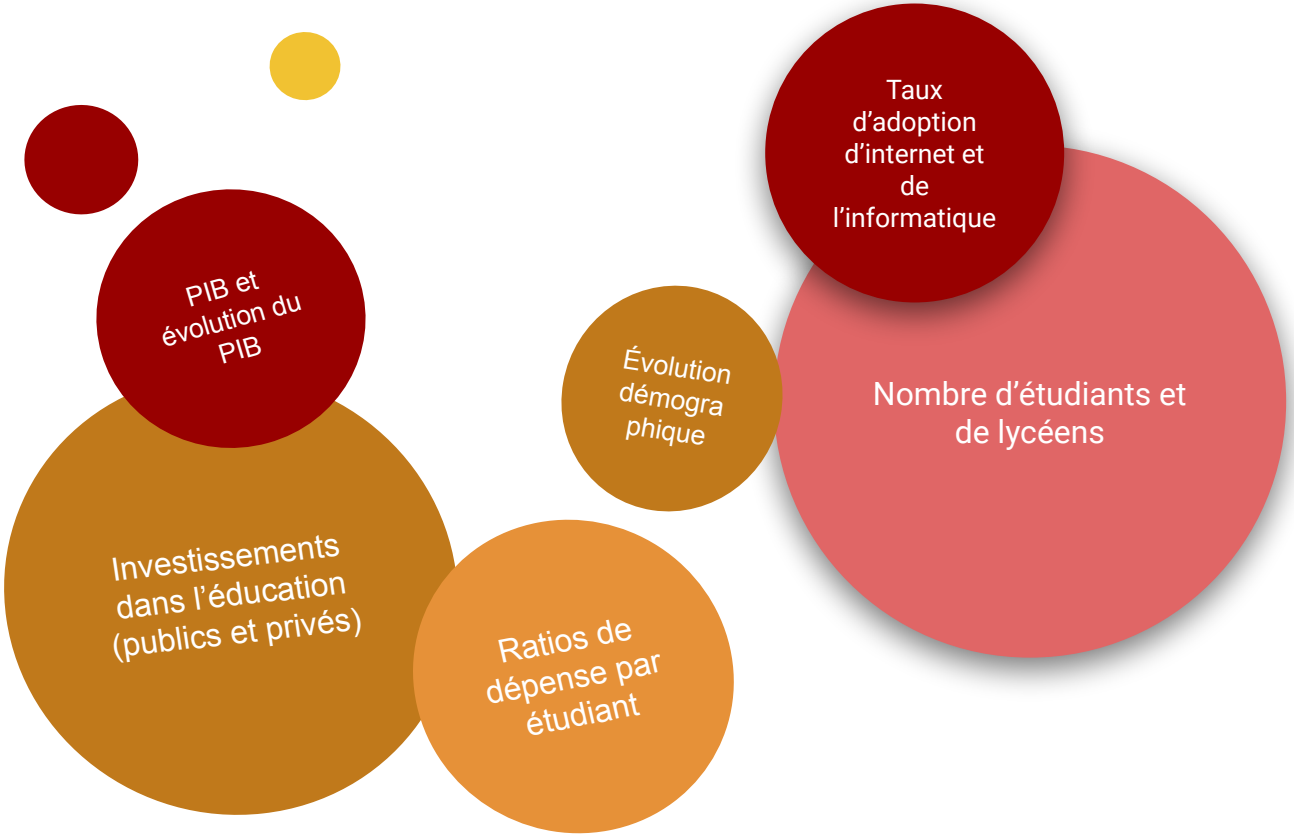



- Identification des préfixes d'indicateurs peu informatifs
- Identification des préfixes des indicateurs les plus informatifs

  => base pour sélectionner les indicateurs

# 3 - Comparer les pays

# 3 - Sélection des indicateurs - Brainstorming



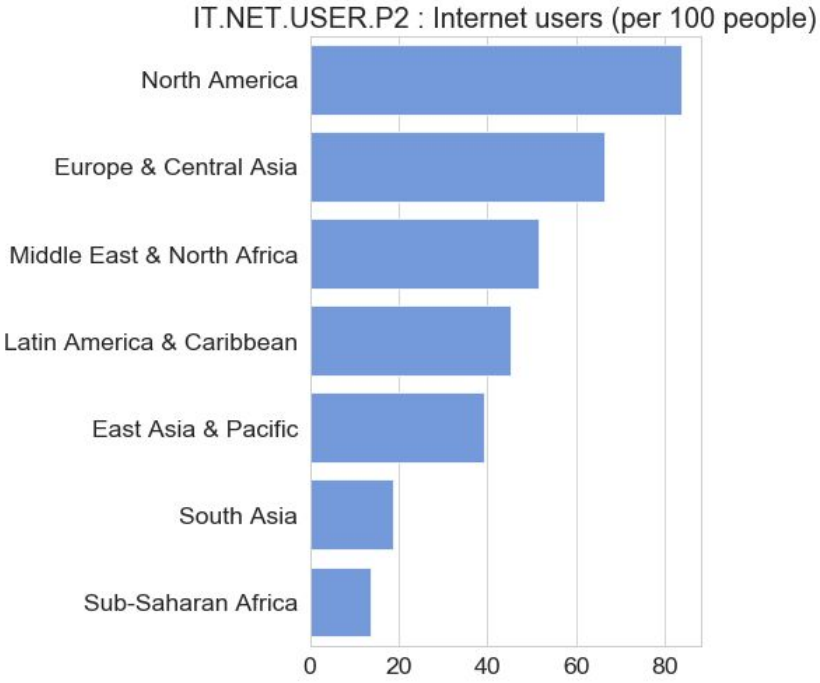
# 3 - Sélection des indicateurs - Indicateurs retenus

Après une phase d'observation des indicateurs : indicateurs retenus

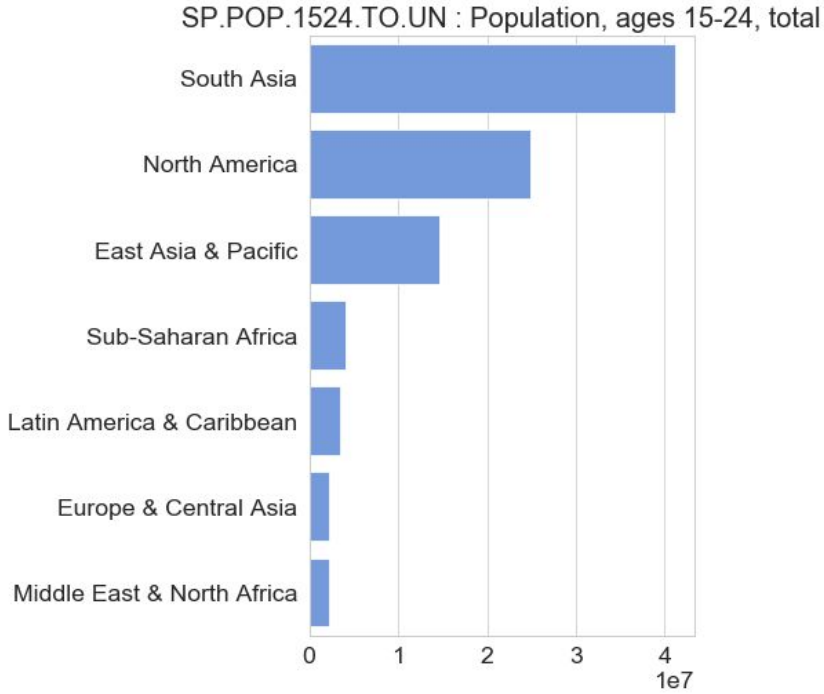
```
data_short[data_short['Indicator Code'].isin(indicateurs)][['Indicator Name', 'Indicator Code', '2010s']]
.groupby(['Indicator Name', 'Indicator Code']).count().reset_index().sort_values(by='2010s',ascending=False)
```

| | Indicator Name | Indicator Code | Nombre de valeurs |
|---|---|---|---|
| 6 | Population, total | SP.POP.TOTL | 240 |
| 3 | Internet users (per 100 people) | IT.NET.USER.P2 | 229 |
| 2 | Enrolment in upper secondary education, both sexes (number) | UIS.E.3 | 206 |
| 1 | Enrolment in tertiary education, all programmes, both sexes (number) | SE.TER.ENRL | 197 |
| 5 | Population, ages 15-24, total | SP.POP.1524.TO.UN | 181 |
| 0 | Enrolment in post-secondary non-tertiary education, both sexes (number) | UIS.E.4 | 137 |

# 3 - Sélection des indicateurs - Exemples d'ordres de grandeur (moyenne)

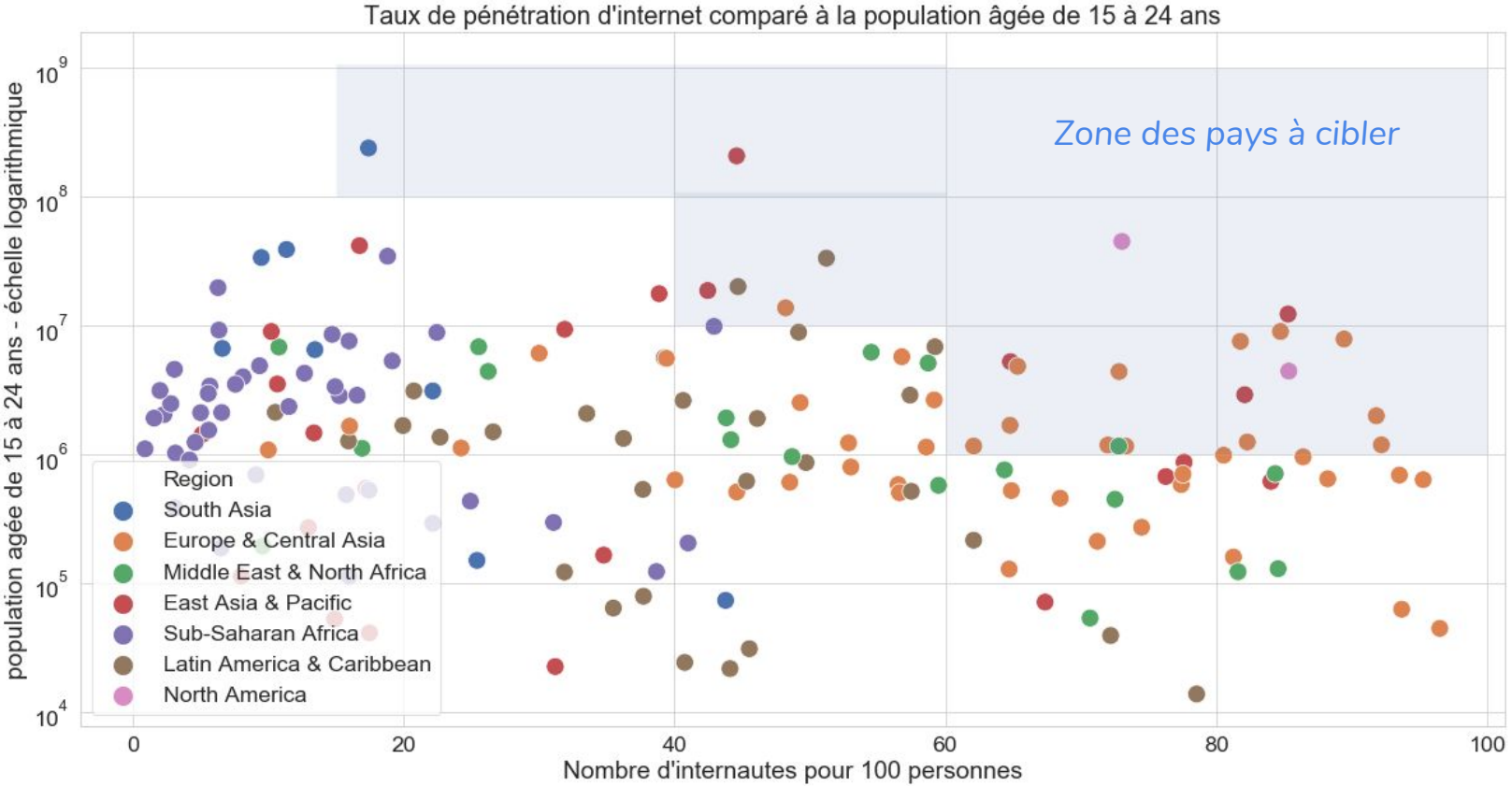



Taux de pénétration d'internet (%)




Moyenne de population âgée de 15 à 24 ans (multiple de 10 millions)

# 3 - Comparaison des pays - Intuition

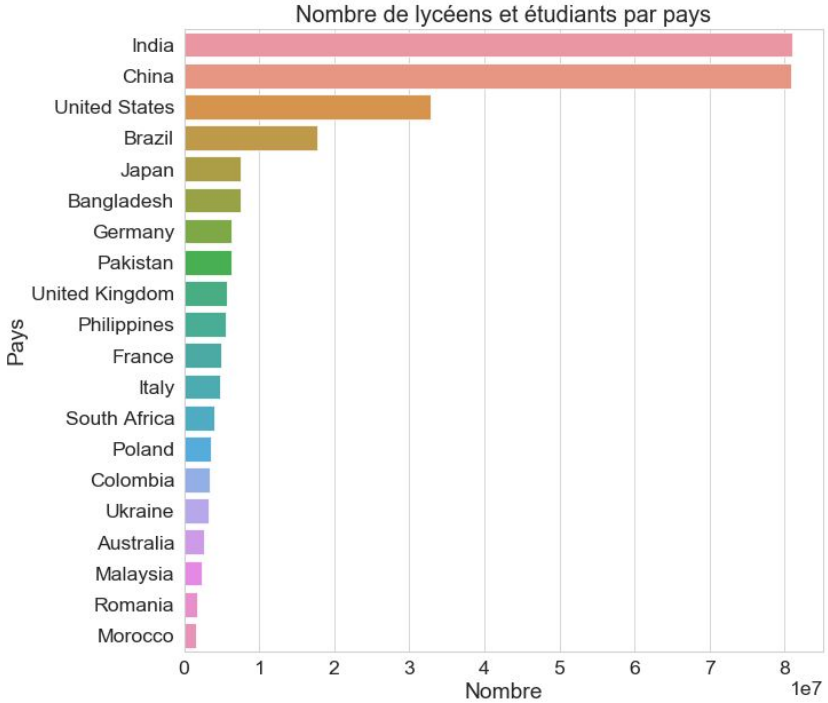



| Country Name |
|---|
| India |
| China |
| United States |
| Japan |
| Germany |
| United Kingdom |
| France |
| Malaysia |

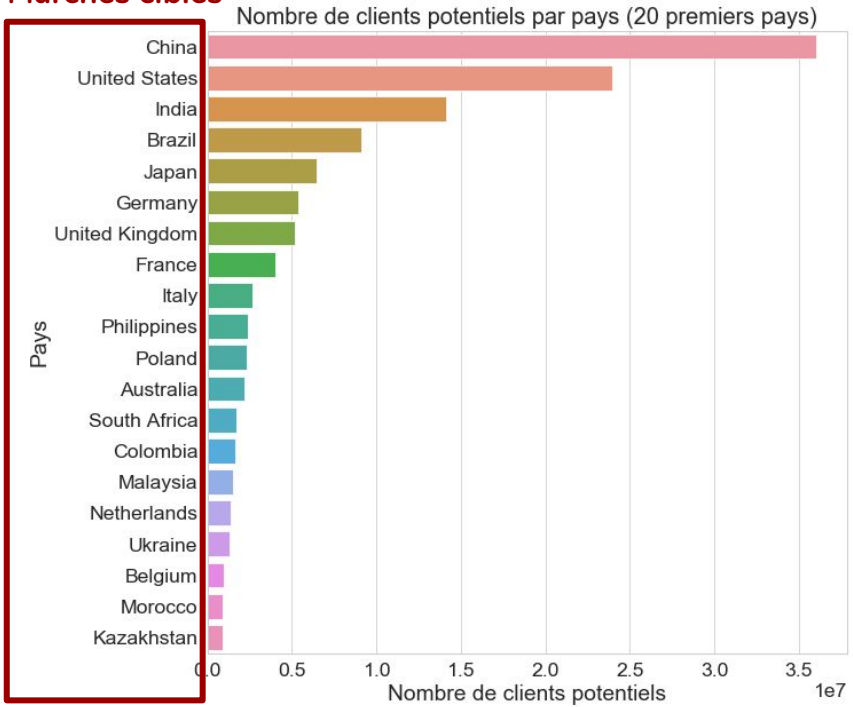
```
df_countries[((df_countries['IT.NET.USER.P2'] > 15) & (df_countries['SP.POP.1524.TO.UN'] > 100000000))  |  ((df_countr
```

# 3 - Comparaison des pays - Estimation du nombre de clients - 20 premiers pays

```
df_countries['customers'] = df_countries['UIS.E.3'] + df_countries['UIS.E.4'] + df_countries['SE.TER.ENRL']
df_countries['potential_customers'] = df_countries['customers'] * df_countries['IT.NET.USER.P2']/100
```



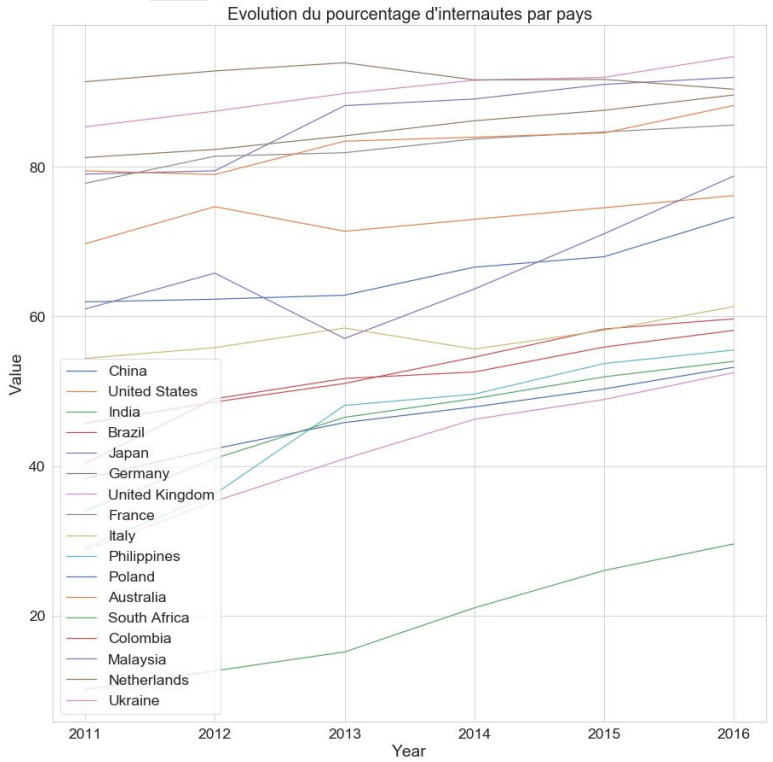

Marchés cibles

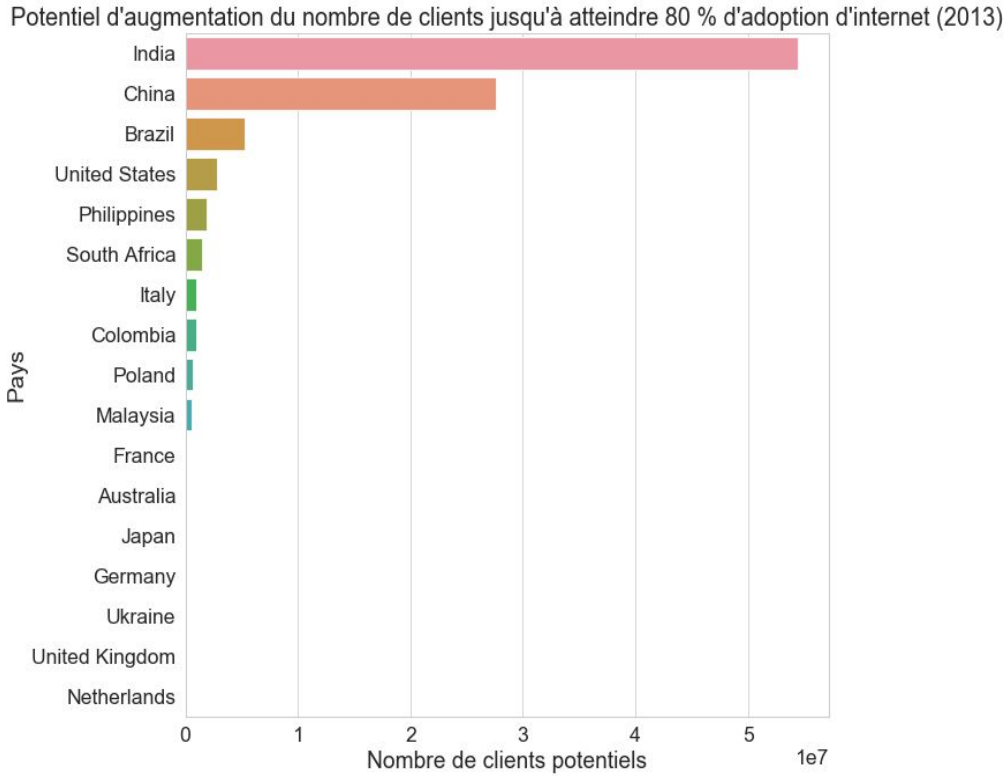


# 4 - Quel potentiel?

# 4 - Quel potentiel pour ces pays?

Exemple avec augmentation de pénétration d'internet



# III Conclusions

# Le jeu de données permet-il de répondre aux attentes de ACADEMY?

**Pertinence du jeu de données**

- tous les pays
- données relatives à l'éducation et utiles + données complémentaires
- sources

**Limites**

- Certains indicateurs inutilisables (beaucoup de données manquantes pour comparer)
- Manque certains indicateurs business : pénétration Moocs, dépense internet, proportion d'élève se formant en dehors de leur établissement, etc.
- Aucune information sur la société Academy pour guider l'étude (proximité géographique, concurrence, langue, etc.)

# Merci de votre attention