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

Soutenance de Projet
en 2023

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**

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

Présentation du jeu de données

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

BANQUE MONDIALE

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 **Version : 1.11.0**	Gestion de package Gestion d'environnement virtuel	Conda : installation de package via le terminal !pip : installation ...
Jupyter Notebook **Version : 6.4.12**	Structurer la démarche Executer code par étape Expliquer la démarche (markdown)	
Python **Version : 3.9.13**	Appel aux librairies, Boucles for pour générer plusieurs graphes	Boucles, Listes, dictionnaires, collections (compteur de mot)
Pandas **Version : 1.4.4**	Manipulation de données Représentation des données	Manipulation de Dataframe : création, copie, filtres, tris, description, concaténation, dépivotage
Matplotlib **Version : 3.5.2** **Seaborn** **Version : 0.11.2**	Génération de graphes	Barplot, Scatterplot, lineplot, distplot, heatmap
Numpy **Version : 1.21.5**	les calculs scientifiques et mathématiques	array, shape, mean, std, sum

1 - Connaître les données

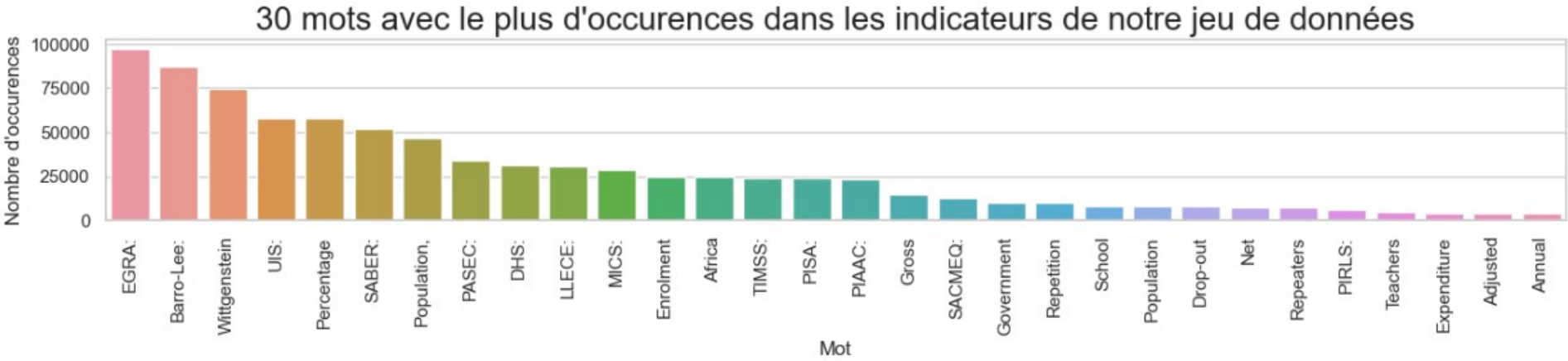
1 - Connaître les données - Préambule

Historique et prédictions de 1970 à 2050

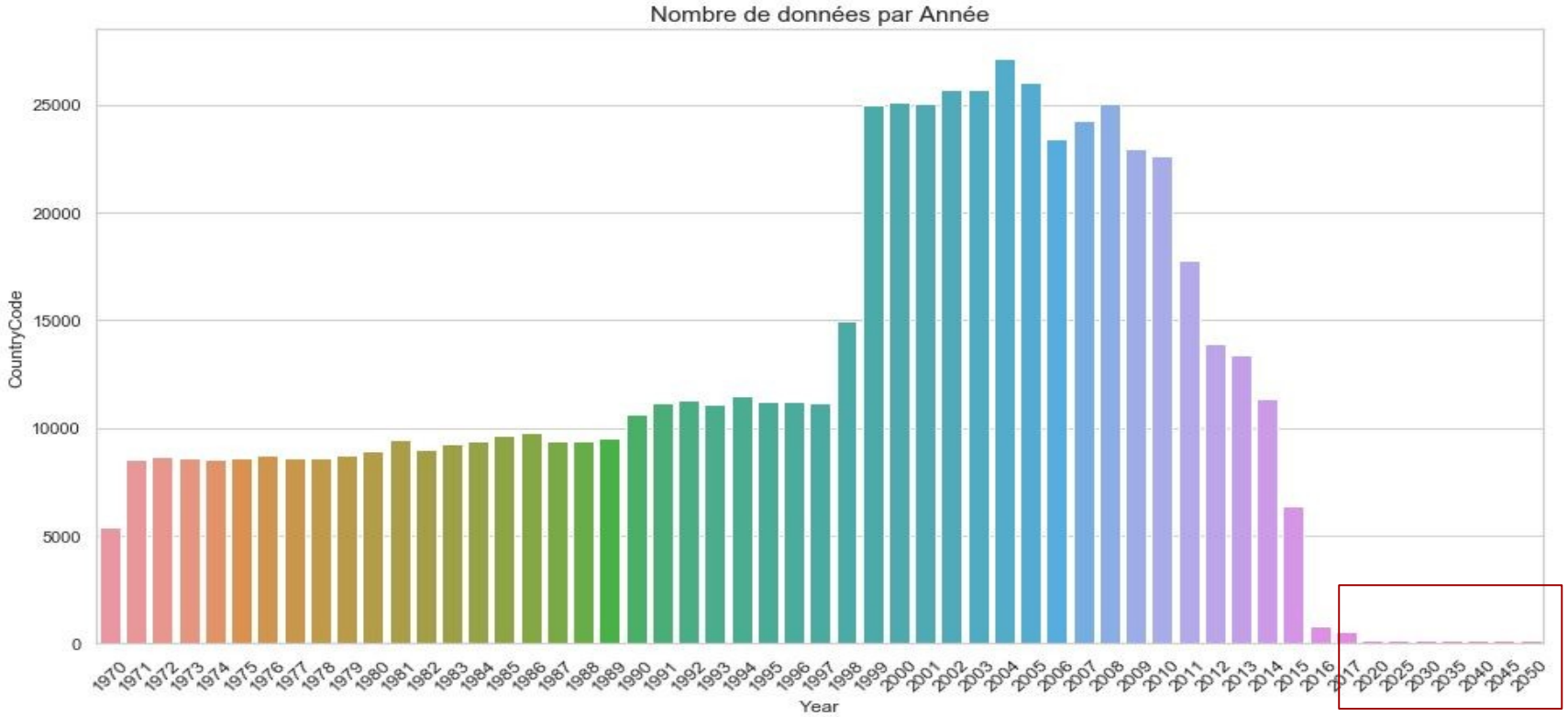
241 zones géographiques (dont pays)

3665 indicateurs uniques

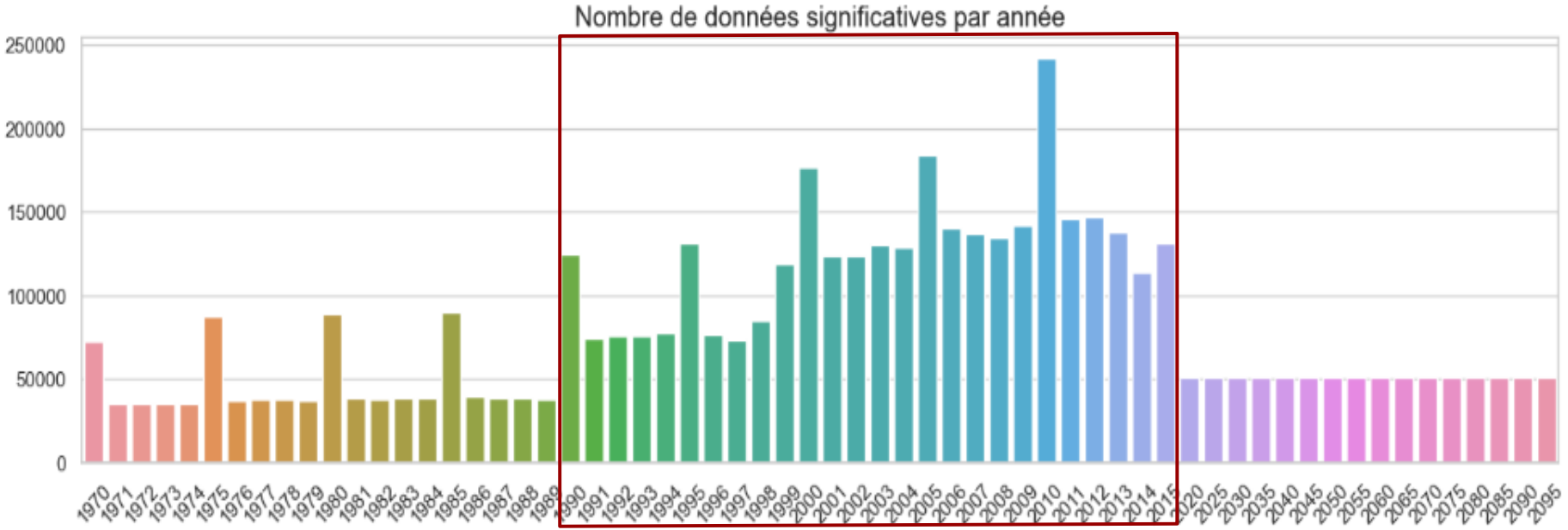
Indicateurs relatif à l'éducation :



1 - Connaître les données - Quantité de données par année fichier EdStatsFootNote.csv

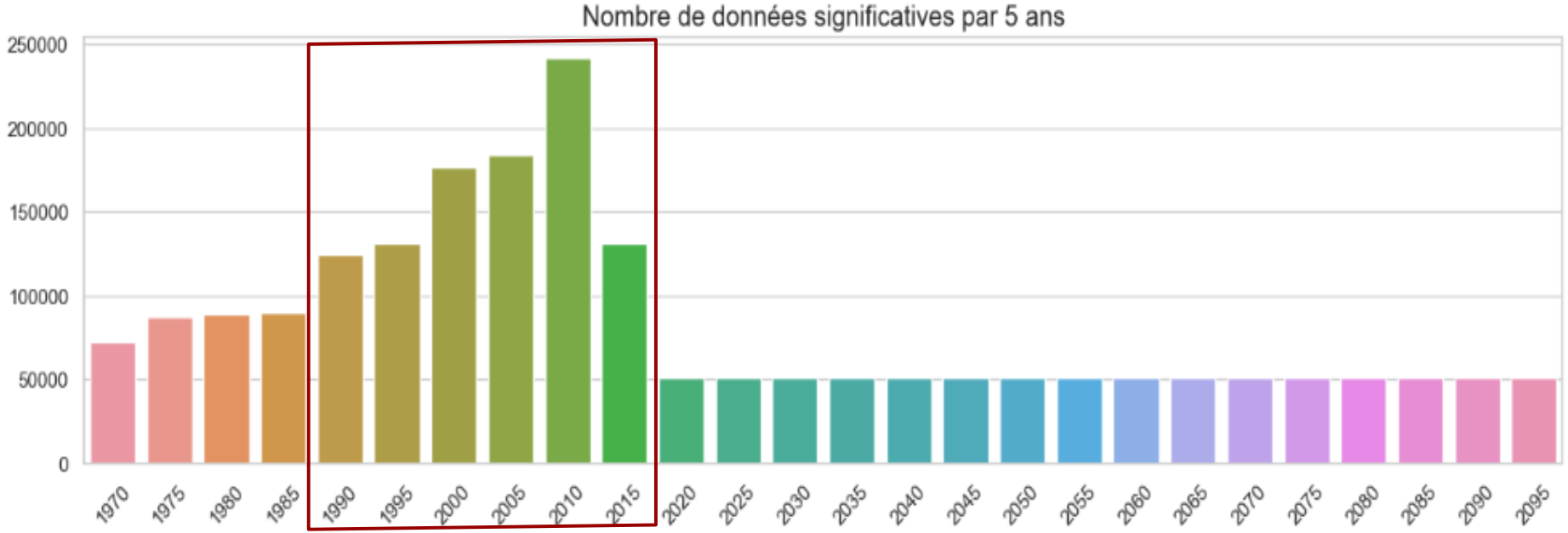


1 - Connaître les données - Nombre de données par année fichier EdStatsData.csv

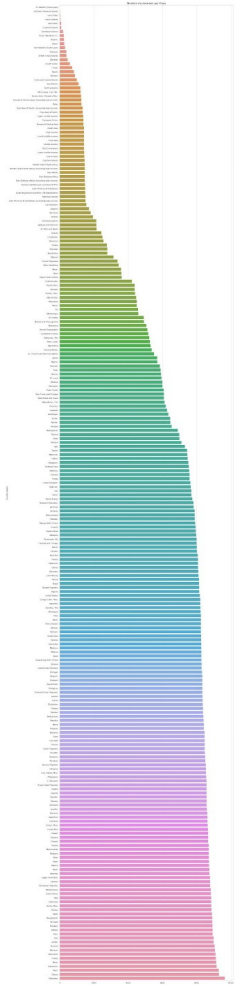


1 - Connaître les données - Nombre de données tous les 5 ans fichier EdStatsData.csv

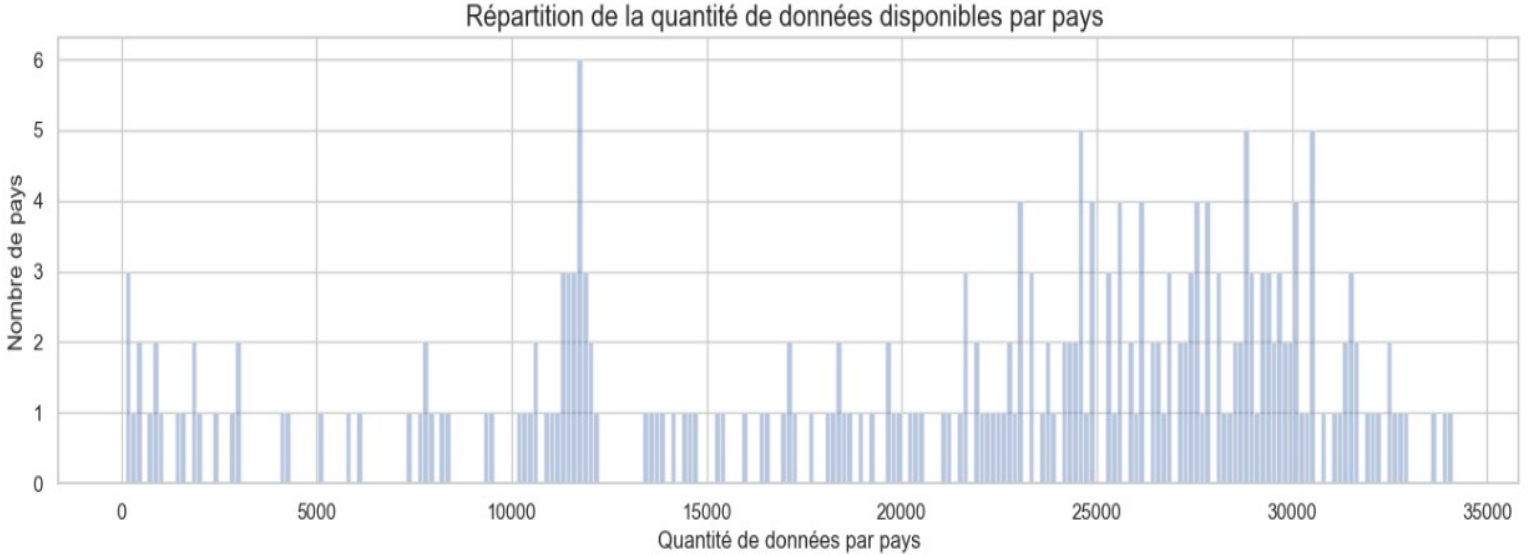
data_periode['1990a2015'] = data_c[[str(year) for year in [1990,1991,1992,1993,1994,1995,1996,1997,1998,1999,2000,2001,2002,2003,2004,2005,2006,2007,2008,2009,2010,2011,2012,2013,2014,2015,2016,2017]]].mean(1)



La période choisis est de 1990 à 2015



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_periode = data_periode.merge(right = country_c[['Country Code', 'Region']],on='Country Code', how='left')
```

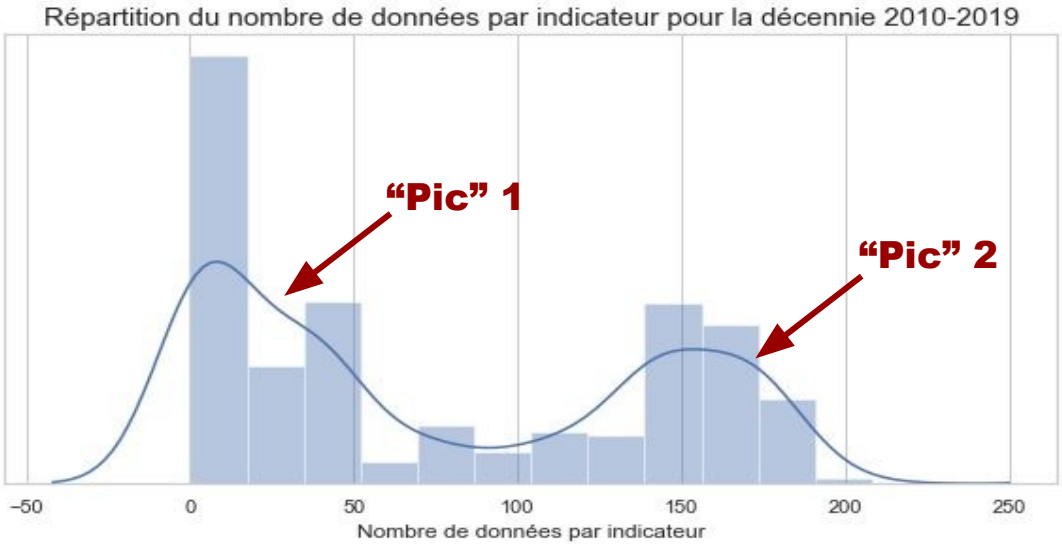
- EdstatsData : les noms de pays, d’indicateur, les valeurs pour la **période 1990 à 2015**

```
data_short = data_periode[['Country Name', 'Country Code', 'Indicator Name','Indicator Code', '1990a2015', 'Region']]
```

- Autres données : non nécessaires à ce stade.

2 - Identifier les indicateurs exploitables

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



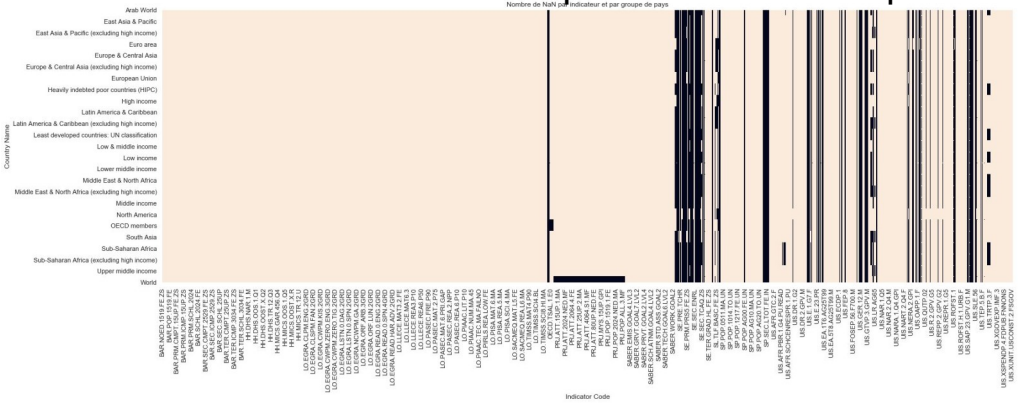
On voit qu'on a 2 pics intéressants :
il y a beaucoup d'indicateurs avec très peu de données (<50)
il y a beaucoup d'indicateurs avec plus de données (autour de 160) une correspondant aux pays (moyenne du nombre de données par indicateur autour de 150) , l'autre aux groupes de pays (moyenne du nombre de données inférieure à 50).

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

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

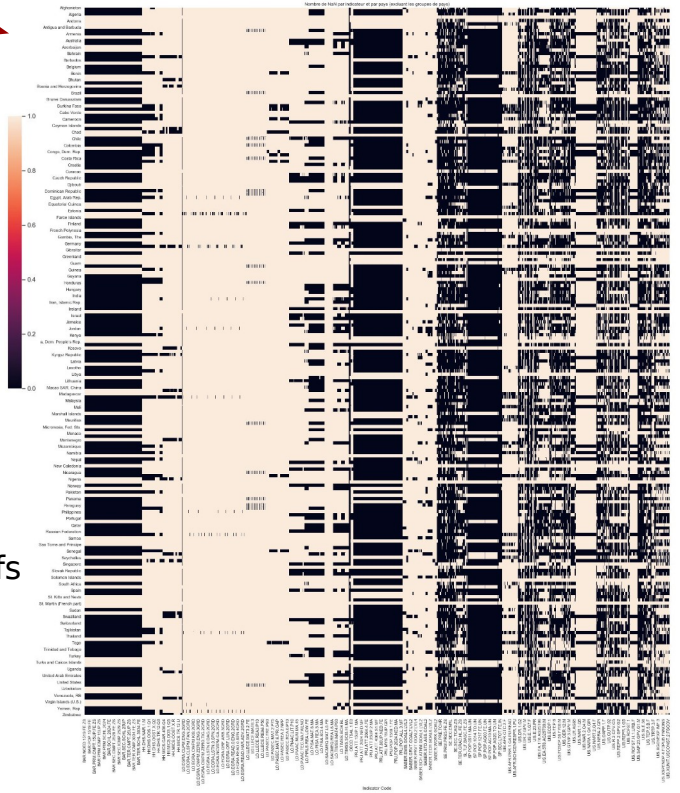
Par zone

Données par indicateur et par zone



Par pays

Données par indicateur et par pays

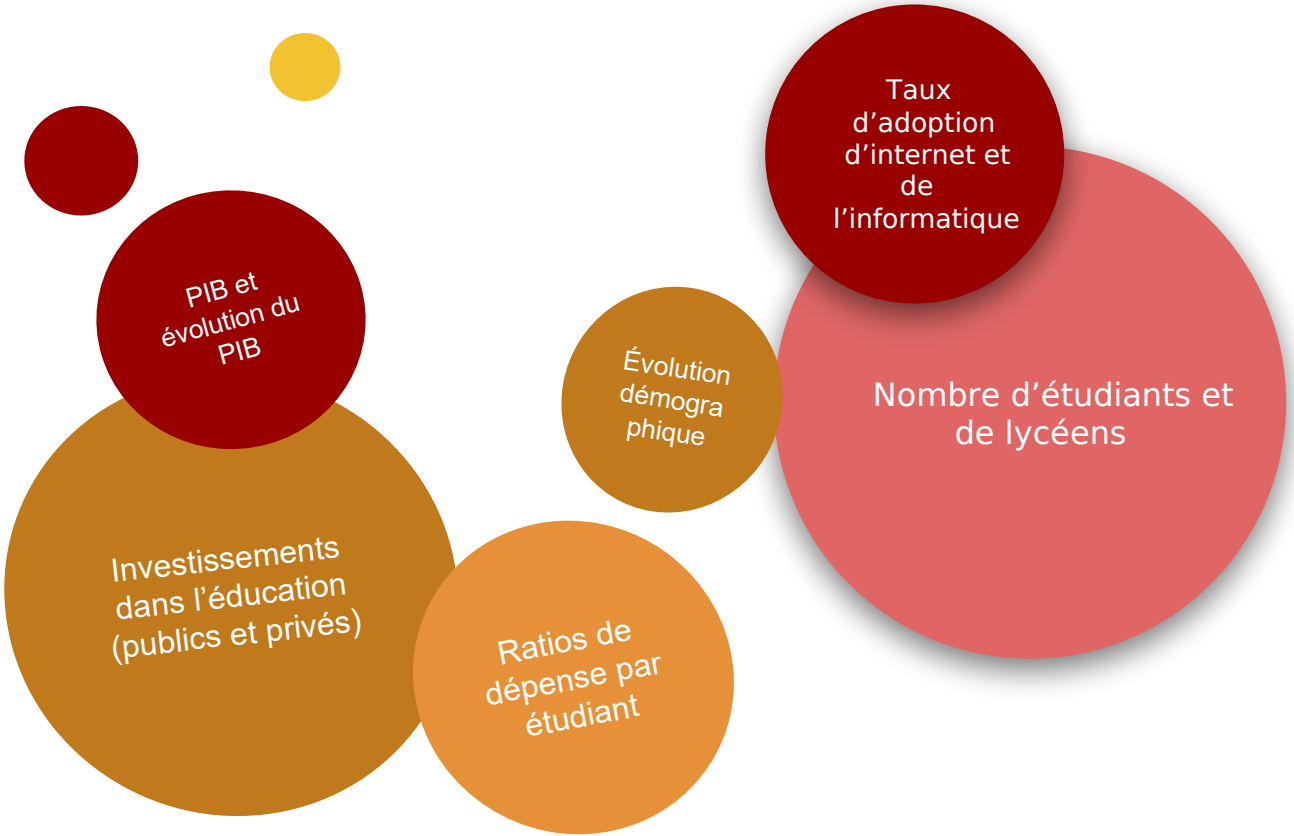


- Identification des préfixes d'indicateurs peu informatifs
- Identification des préfixes des indicateurs les plus informatifs

  => base pour sélectionner les indicateurs

3 - Comparer les pays

3 - Sélection des indicateurs -



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', '1990a2015']].
groupby(['Indicator Name', 'Indicator Code']).count().reset_index().sort_values(by='1990a2015',ascending=False)
```

	Indicator Name	Indicator Code	Nombre de valeurs
8	Population, total	SP.POP.TOTL	240
3	GDP per capita (current US$)	NY.GDP.PCAP.CD	234
5	Internet users (per 100 people)	IT.NET.USER.P2	233
2	Enrolment in upper secondary education, both s...	UIS.E.3	225
1	Enrolment in tertiary education, all programme...	SE.TER.ENRL	224
6	Personal computers (per 100 people)	IT.CMP.PCMP.P2	218
4	Government expenditure on education as % of GD...	SE.XPD.TOTL.GD.ZS	192
7	Population, ages 15-24, total	SP.POP.1524.TO.UN	192
0	Enrolment in post-secondary non-tertiary educa...	UIS.E.4	171

3 - Sélection des indicateurs - Indicateurs retenus

1 - Enrôlement dans l'enseignement tertiaire, tous programmes confondus (SE.TER.ENRL):
Cet indicateur mesure le nombre total d'étudiants inscrits dans les programmes d'enseignement tertiaire, y compris les universités, les écoles professionnelles et les établissements d'enseignement supérieur.

2 - Enrôlement dans l'enseignement secondaire supérieur, tous sexes confondus (UIS.E.3):
Cet indicateur mesure le nombre total d'étudiants inscrits dans l'enseignement secondaire supérieur, également appelé enseignement secondaire supérieur.

3 - PIB par habitant (en dollars courants des États-Unis) (NY.GDP.PCAP.CD): Cet indicateur mesure la richesse produite par une économie par habitant. Il est calculé en divisant le PIB d'un pays par sa population.

4 - Dépenses publiques pour l'éducation en pourcentage du PIB (SE.XPD.TOTL.GD.ZS): Cet indicateur mesure la part des dépenses publiques consacrée à l'éducation par rapport à la richesse produite par le pays.

5 - Utilisateurs d'internet (pour 100 personnes) (IT.NET.USER.P2): Cet indicateur mesure le pourcentage de la population ayant accès à Internet.

6 - Ordinateurs personnels (pour 100 personnes) (IT.CMP.PCMP.P2): Cet indicateur mesure le nombre d'ordinateurs personnels pour 100 personnes.

7 - Population âgée de 15 à 24 ans, totale (SP.POP.1524.TO.UN): Cet indicateur mesure le nombre total de personnes âgées de 15 à 24 ans dans une population.

8 - Population totale (SP.POP.TOTL): Cet indicateur mesure le nombre total de personnes vivant dans une région donnée.

3 - Sélection des indicateurs -
Définition de la moyenne, écart type, médiane

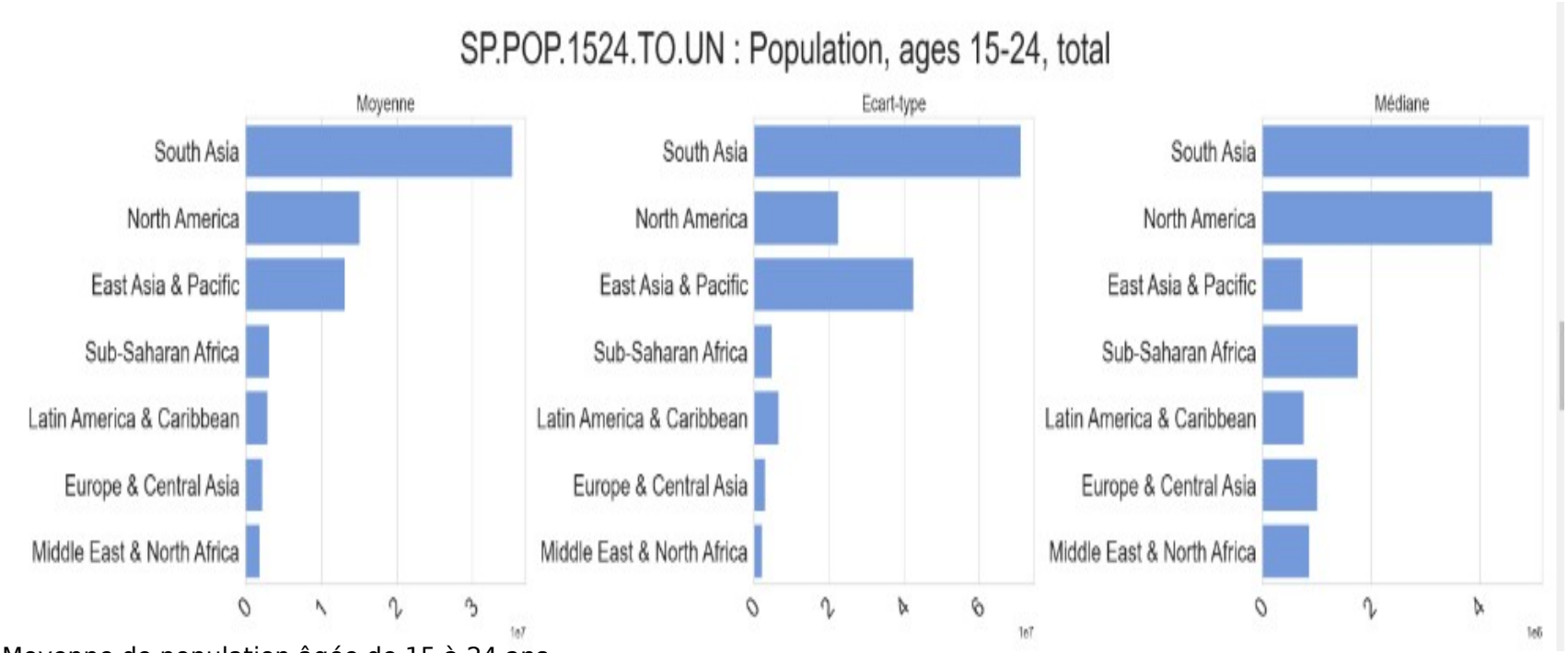
Moyenne : une mesure de la tendance centrale qui représente la somme de toutes les valeurs d'un ensemble de données, divisée par le nombre total de valeurs.

Écart type : une mesure de la dispersion ou de la variabilité d'un ensemble de données par rapport à sa moyenne.

Médiane : une mesure de la tendance centrale qui représente la valeur au centre d'un ensemble de données triées par ordre croissant ou décroissant.

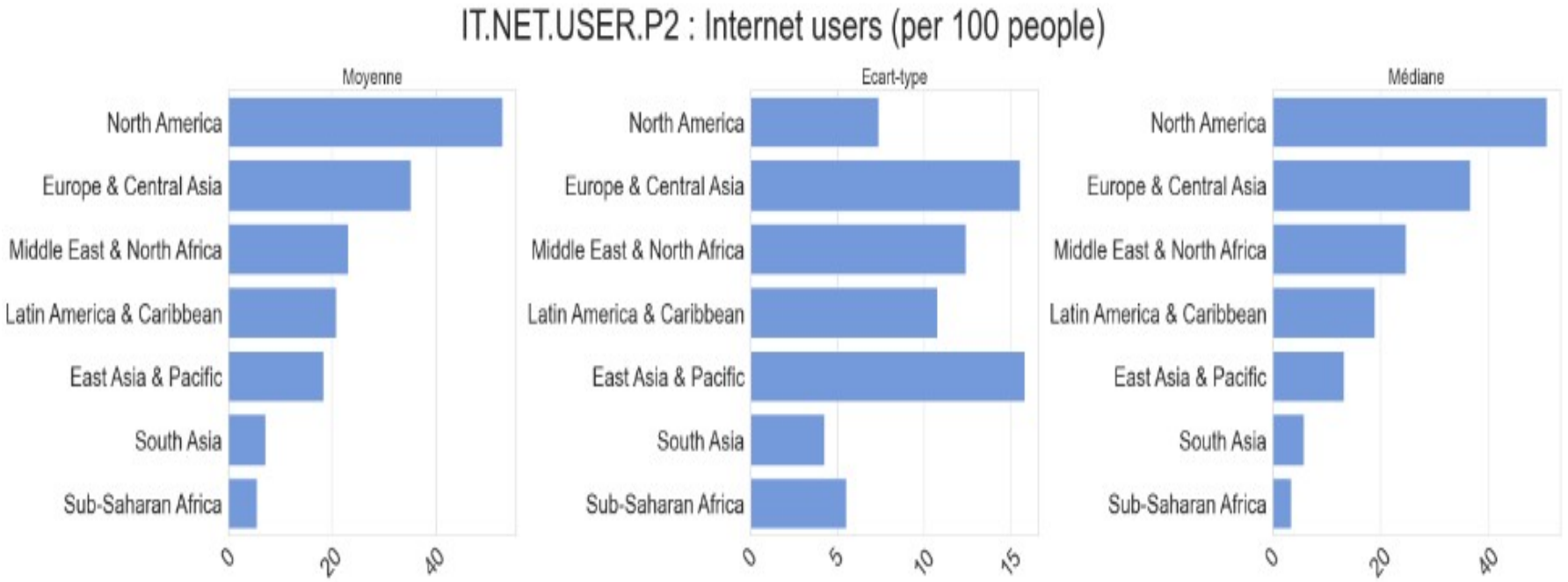
Ces concepts sont fondamentaux en data science pour décrire et analyser des ensembles de données, et sont utilisés pour prendre des décisions éclairées et tirer des insights significatifs (:"observations ou résultats pertinents" ou encore "perspectives significatives")

3 - Sélection des indicateurs -
Ordres de grandeur : Population, ages 15-24



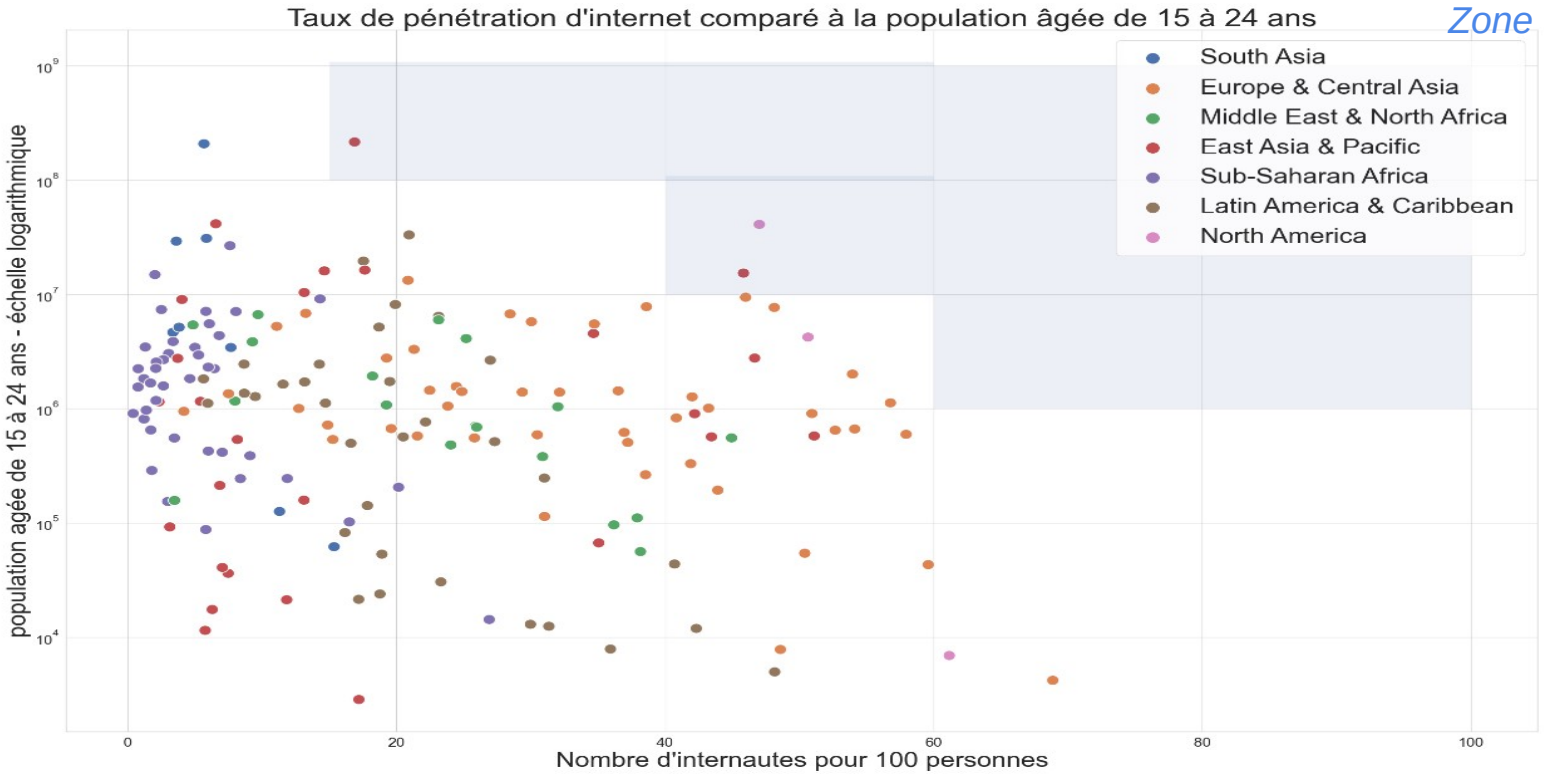
Moyenne de population âgée de 15 à 24 ans

3 - Sélection des indicateurs -
Ordres de grandeur : internet users (per 100 people)



Taux de pénétration d'internet (%)

3 - Comparaison des pays - Intuition



Zone des pays à cibler

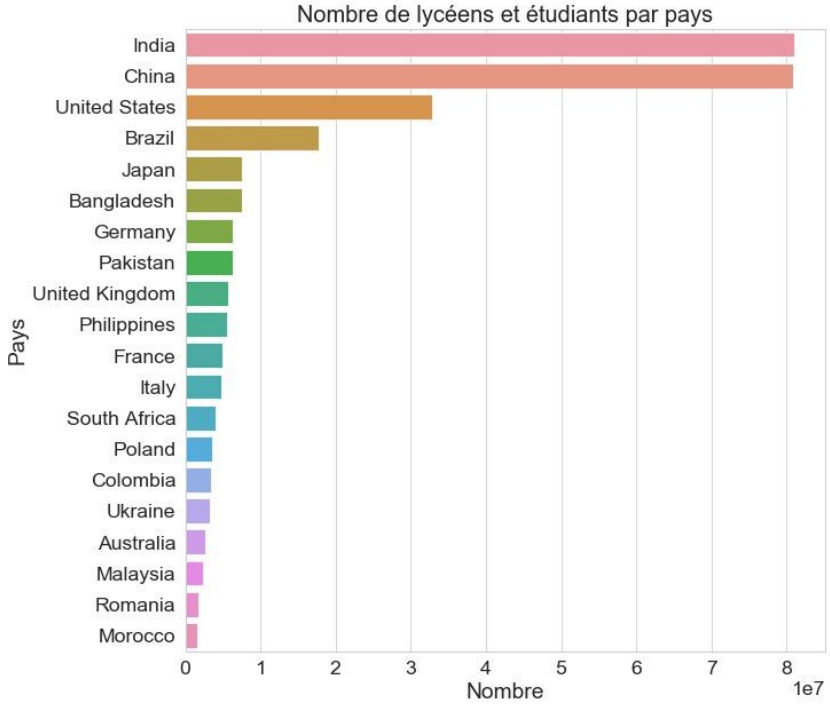
	Country Name
38	China
78	India
79	Indonesia
189	United States
25	Brazil
133	Pakistan
129	Nigeria
113	Mexico
193	Vietnam
139	Philippines
86	Japan
181	Turkey
175	Thailand

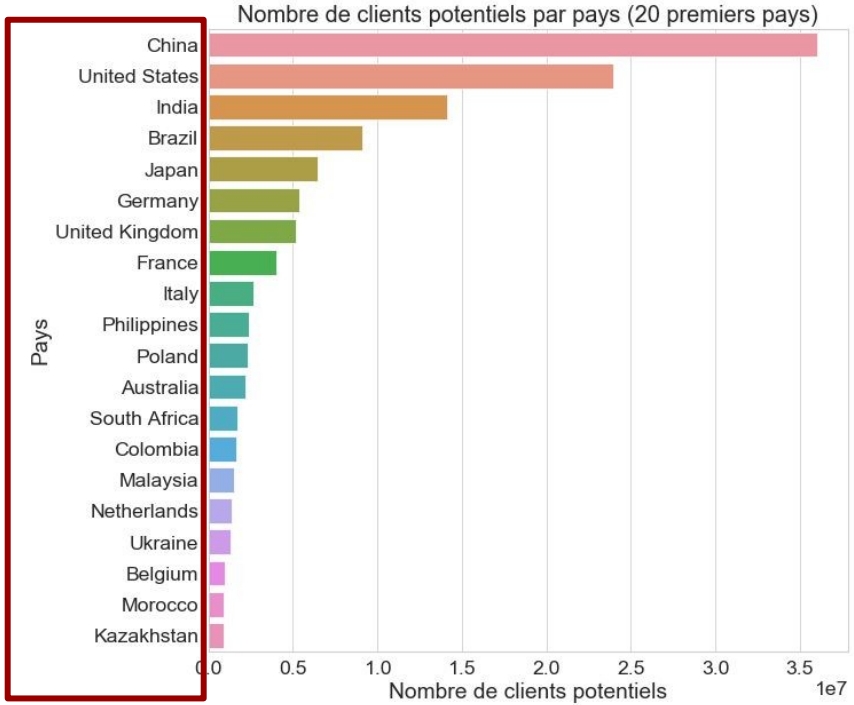
```
df_countries[((df_countries['IT.NET.USER.P2'] > 5) & (df_countries['SP.POP.1524.TO.UN'] > 10000000))|
        ((df_countries['IT.NET.USER.P2'] > 60) & (df_countries['SP.POP.1524.TO.UN'] > 1000000))].
        sort_values(by='SP.POP.1524.TO.UN',ascending = False)[['Country Name']]
```

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

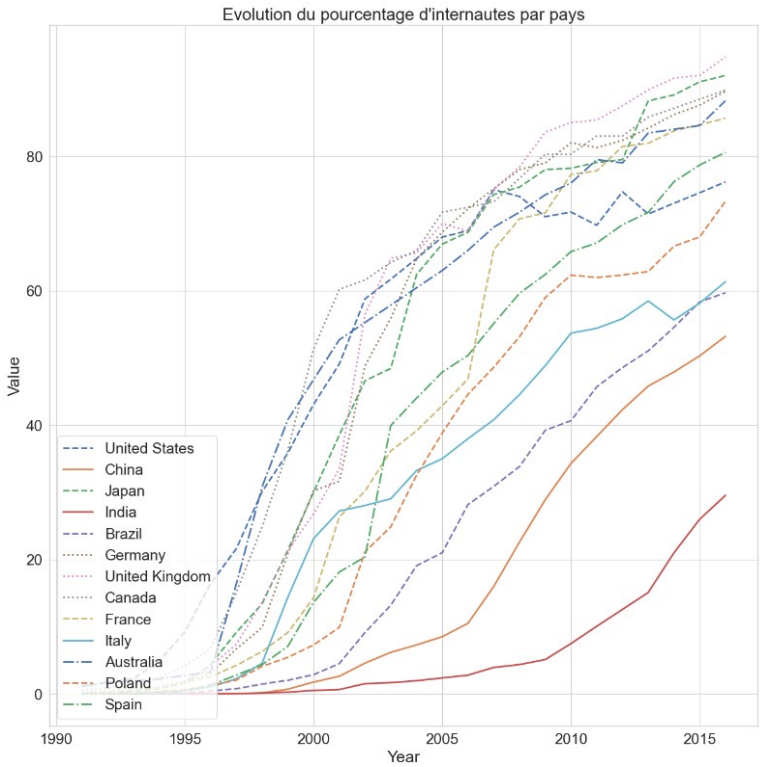


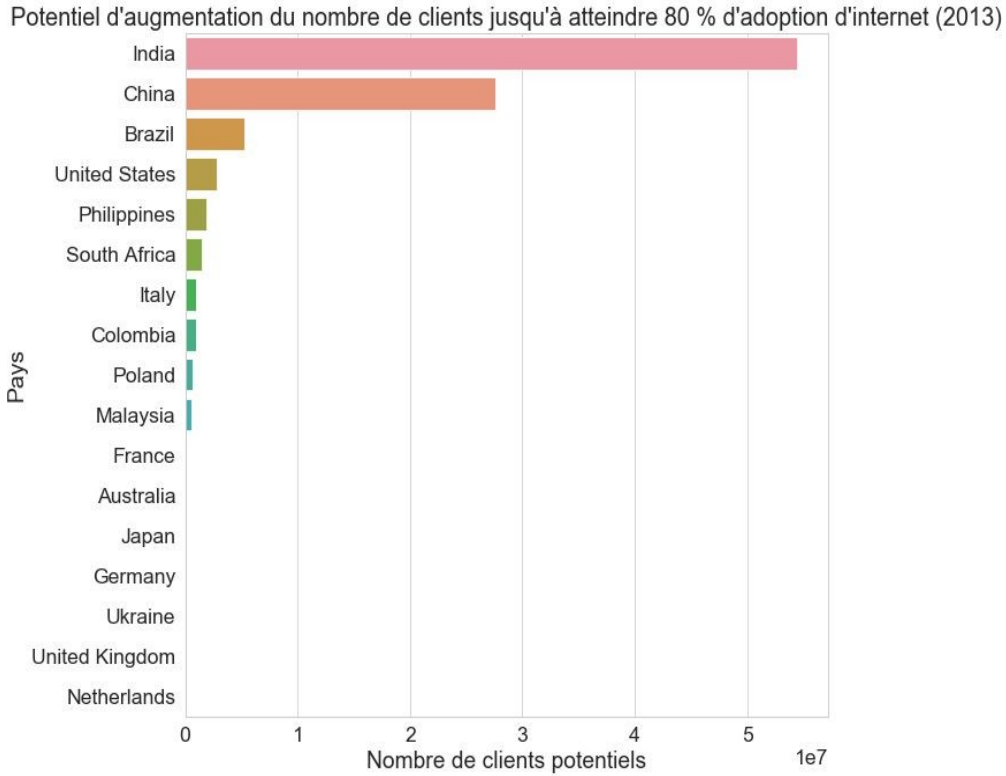


4 - Quel potentiel?

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?

Pays sélectionnés

- Chine, Inde, Etats Unis, Brésil.

Pertinence du jeu de données

- tous les pays sont sourcés
- données relatives à l'éducation et utiles + données complémentaires
- Sources de réussites à des tests, données démographiques , investissement dans l'éducation

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