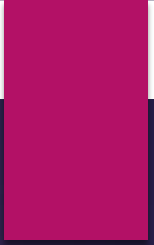


Projet n°2 : Analysez des données de systèmes éducatifs

ALEXANDRE JACQUELINE – DÉCEMBRE 2021

Programme

- ▶ Rappel de la problématique et présentation du jeu de données
- ▶ Analyse pré exploratoire
- ▶ Conclusions sur la pertinence du jeu de données



Rappel de la problématique et présentation du jeu de données

Rappel de la problématique

- ▶ Academy est un start-up de la EdTech
- ▶ E learnings: contenus de formation de niveau lycée et université
- ▶ Objectif d'expansion à l'international



Objectif du projet:

Informar 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

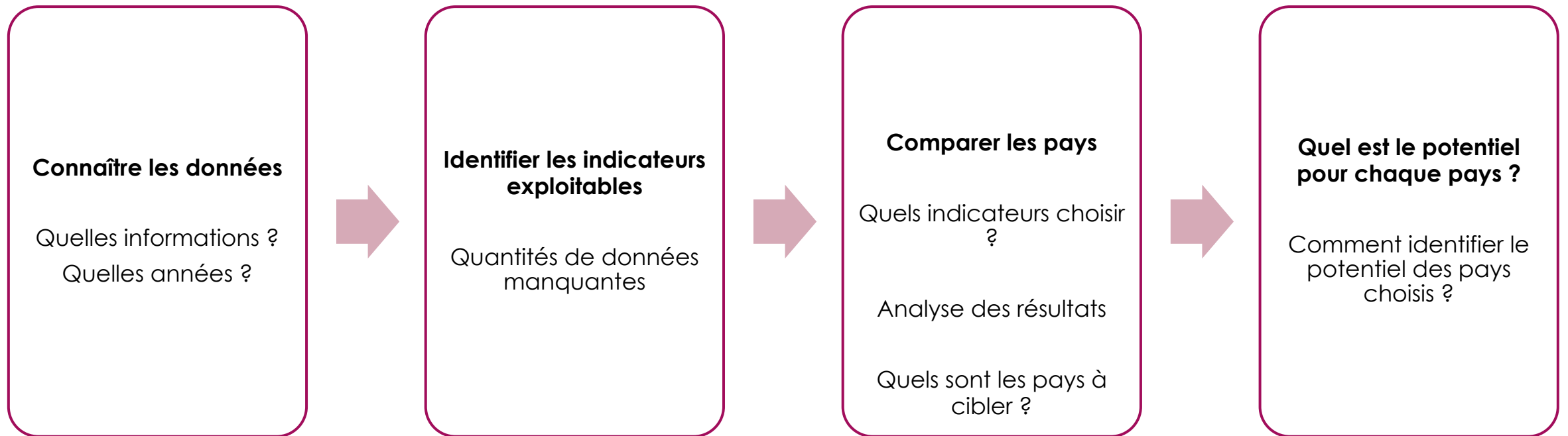


BANQUE MONDIALE

EdStatsCountry.csv	Informations globales sur l'économie de chaque pays du monde(et de zones géographiques) Taille: 241 lignes (1 pays / zone), 32 colonnes, quelques valeurs manquantes, aucun doublon
EdStatsCountry-Series.csv	Informations sur la source de données contenues dans EdStatsCountry Taille: 613 lignes, 4 colonnes, pas de valeur manquante(sauf Unnamed:3), 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, nombreuse valeurs manquantes , aucun doublons
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), aucun doublon
EdStatsSeries.csv	Informations sur les indicateurs socio économiques disponibles dans EdStatsDate Taille: 3665 lignes, 21 colonnes, 6 colonnes vides pour lesquelles il manque des valeurs, il manque plus de 80% des données dans 10 autres colonnes de la table, aucun doublon

Analyse Pré exploratoire

Processus d'analyse pré exploratoire



Outils utilisés pour l'analyse

Nom	Utilisation	Fonctions spécifiques
Jupyter Notebook	Structurer la démarche Exécuter code par étape Expliquer la démarche (markdown)	
Anaconda	Gestion des package Gestion de l'environnement virtuel	Condat: instal de package via le terminal
Python 3.7	Appel de librairies, Boucle for pour générer plusieurs graphe	Boucles, listes, dictionnaires, collections
Pandas	Manipulation de données Représentation des données	Manipulation de Df: création, copie, filtres, tris, description, concaténation, dépivotage
Matplotlib Seaborn	Génération de graphes	Barplot, Scatterplot, distplot, heatmap

Connaître les données

Connaître les données - Préambule



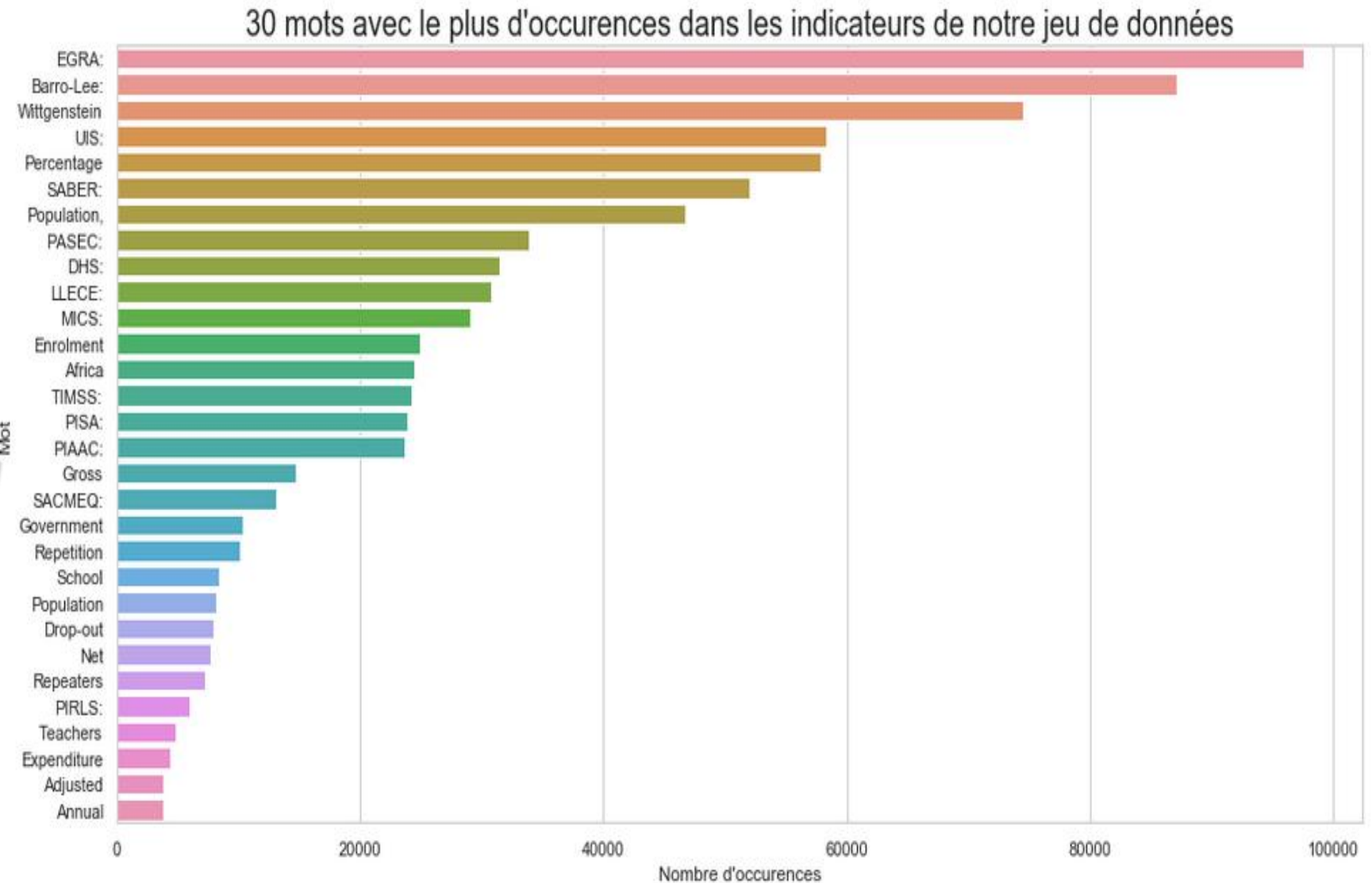
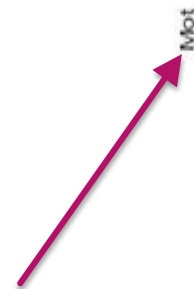
BANQUE MONDIALE

Historique et
Prédictions de
1970 à 2050

3665
Indicateurs
uniques

241 zones
Géographiques
(dont pays)

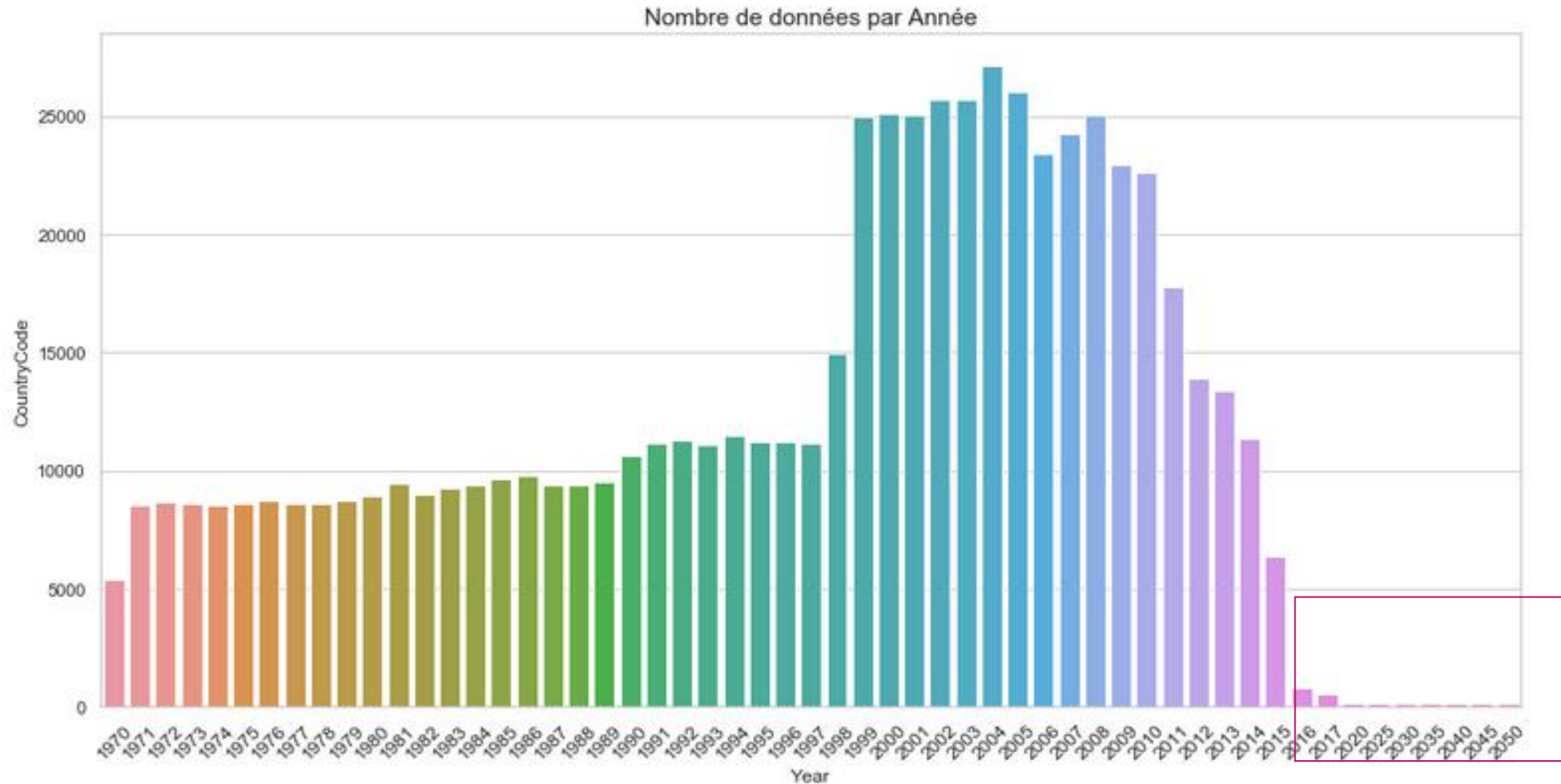
Indicateurs
Relatifs à
l'éducation



Connaître les données – Quantité de données par années



BANQUE MONDIALE



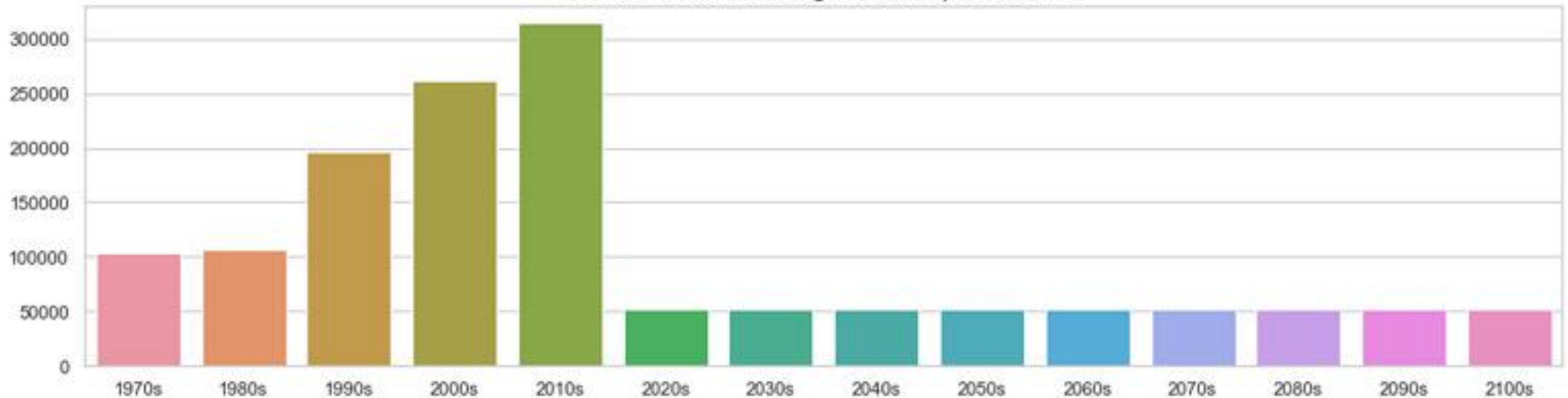
Connaître les données – Nombre de données par décennie



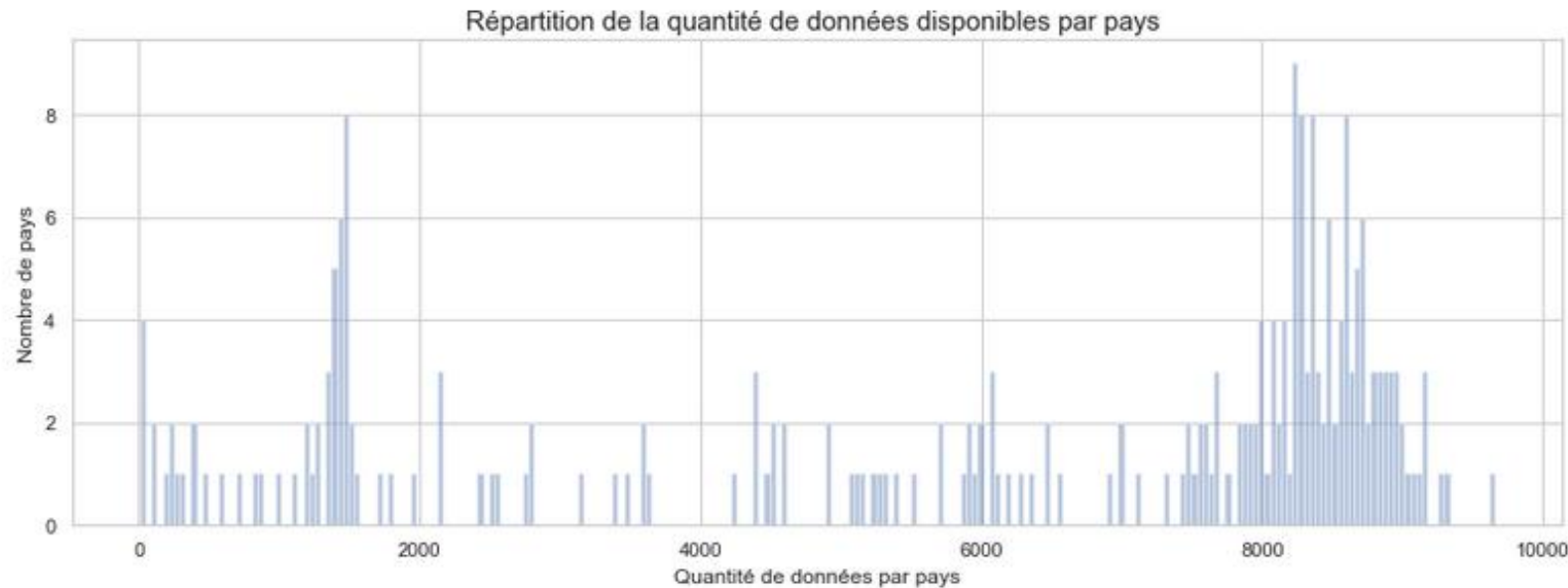
BANQUE MONDIALE

```
1 data['1970s'] = data[[str(year) for year in range(1970,1980,1)]].mean(1)
2
3 data['1980s'] = data[[str(year) for year in range(1980,1990,1)]].mean(1)
4 data['1990s'] = data[[str(year) for year in range(1990,2000,1)]].mean(1)
5 data['2000s'] = data[[str(year) for year in range(2000,2010,1)]].mean(1)
6 data['2010s'] = data[[str(year) for year in range(2010,2011,1)]]
```

Nombre de données significatives par décennie



Connaître les données – Nombre de données par pays



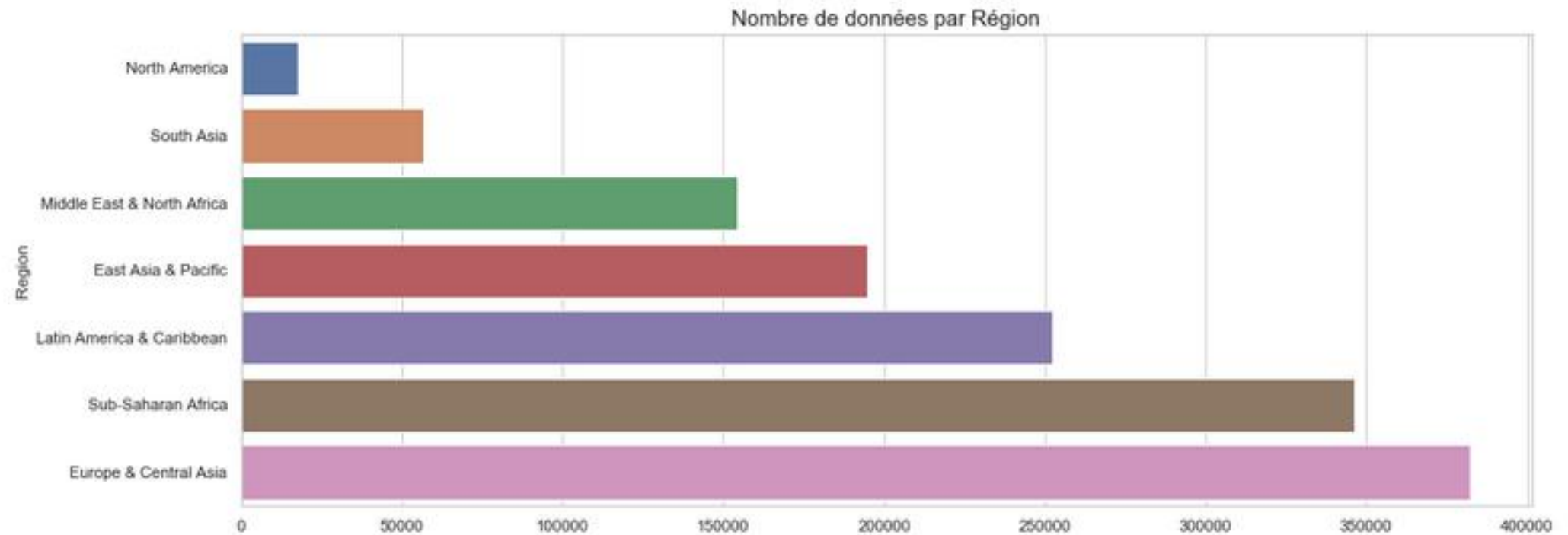
Constat: Inégalité de répartition des données par pays:
Moins de données pour environ 30% des zones:

- Les petits pays:
- Les nouveaux pays (Kosovo)
- Les régions et groupes de pays (East Asia & Pacific, Upper Middle Income, ect).

Connaître les données – Nombre de données par régions



BANQUE MONDIALE



Connaître les données – Quelles informations conserver ?



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

- EdStatsCountry: l'association pays-région

```
1 data = data.merge(right = country[['Country Code', 'Region']],  
2                   on='Country Code', how='left')  
3
```

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

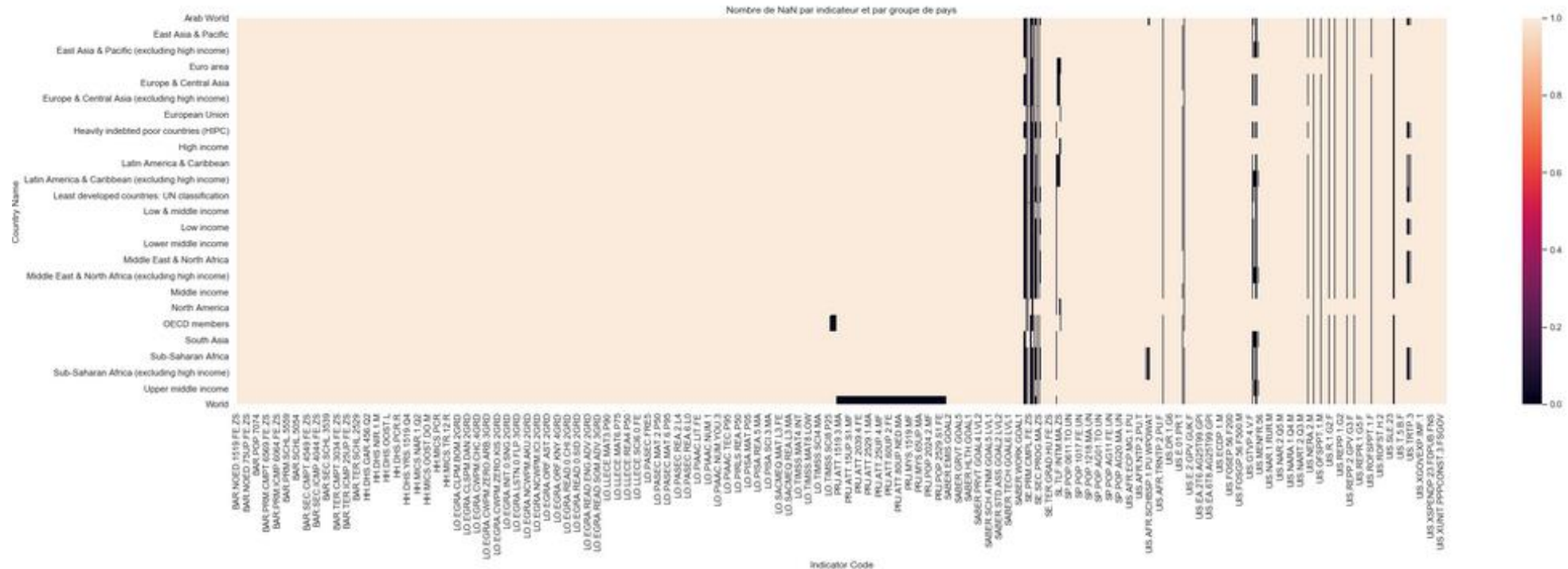
```
Entrée [90]: 1 data_short = data[['Country Name', 'Country Code', 'Indicator Name',  
2                               'Indicator Code', '2010s', 'Region']]
```

- Autres données: non nécessaires à ce stade



Identifier les indicateurs
exploitables

Sélection des indicateurs – Indicateurs retenus



Ce résultat est intéressant, on voit que pour de nombreux indicateurs, il n'y a aucun NaN pour les Groupe de pays! C'est le cas notamment de:

- Ceux avec le préfixe BAR
- Ceux avec le préfixe HH
- Ceux avec le préfixe LO
- Ceux avec le préfixe SP.POP
- Ceux avec le préfixe PRU (hors zone world)



Comparer les pays

Sélection des indicateurs – Brainstorming

PIB et
Évolution du PIB

Taux d'adoption
D'internet et de l'informatique

Ratios de
Dépense par
étudiant

Evolution
démographique

Invest dans
L'éducation
Publics et privé

Nb d'étudiants et
De lycéens

Evolution
démographique

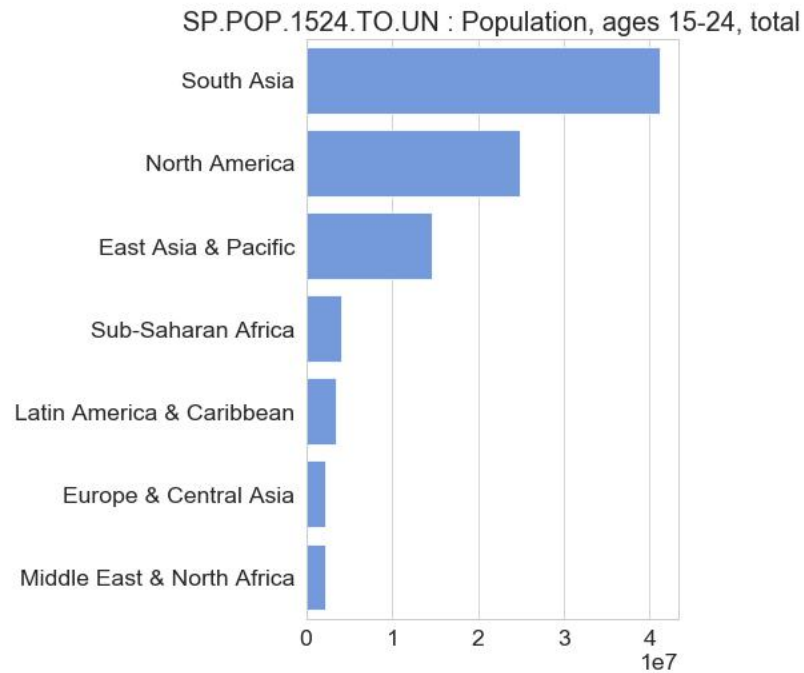
Sélection des indicateurs – Indicateurs retenus

```
Entrée [111]: 1 data_short[data_short['Indicator Code'].isin(indicateurs)]  
2 [['Indicator Name', 'Indicator Code', '2010s']].groupby(['Indicator Name', 'Indicator Code']).count()  
3 .reset_index().sort_values(by='2010s',ascending=False)
```

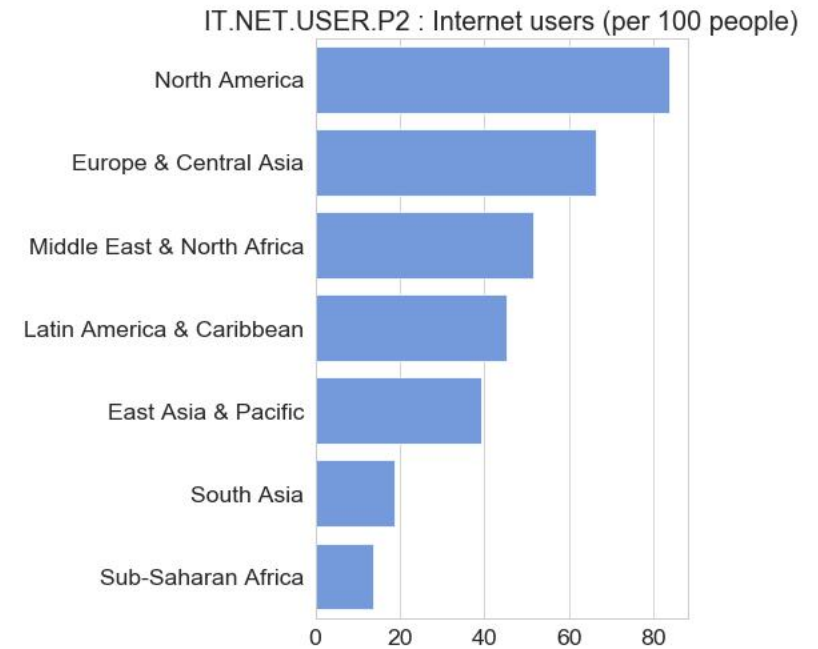
Out[111]:

	Indicator Name	Indicator Code	2010s
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 s...	UIS.E.3	206
1	Enrolment in tertiary education, all programme...	SE.TER.ENRL	197
5	Population, ages 15-24, total	SP.POP.1524.TO.UN	181
0	Enrolment in post-secondary non-tertiary educa...	UIS.E.4	137
4	Personal computers (per 100 people)	IT.CMP.PCMP.P2	0

Sélection des indicateurs – Exemple d'ordres de grandeur (moyenne)

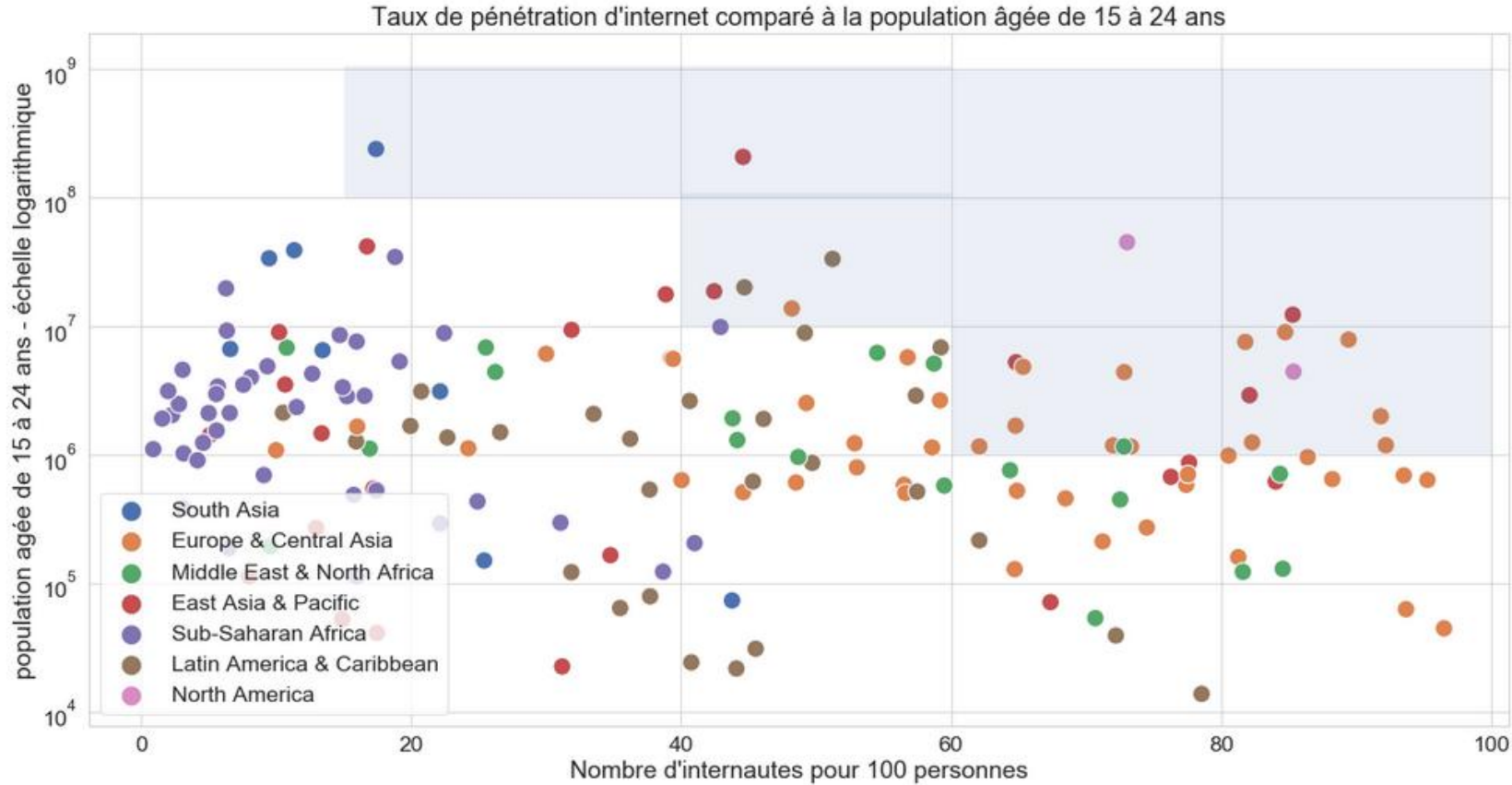


Moyenne de pop âgée de 15 à 24 ans
(Multiple de 10 millions)



Taux de pénétration d'internet (%)

Comparaison des pays-

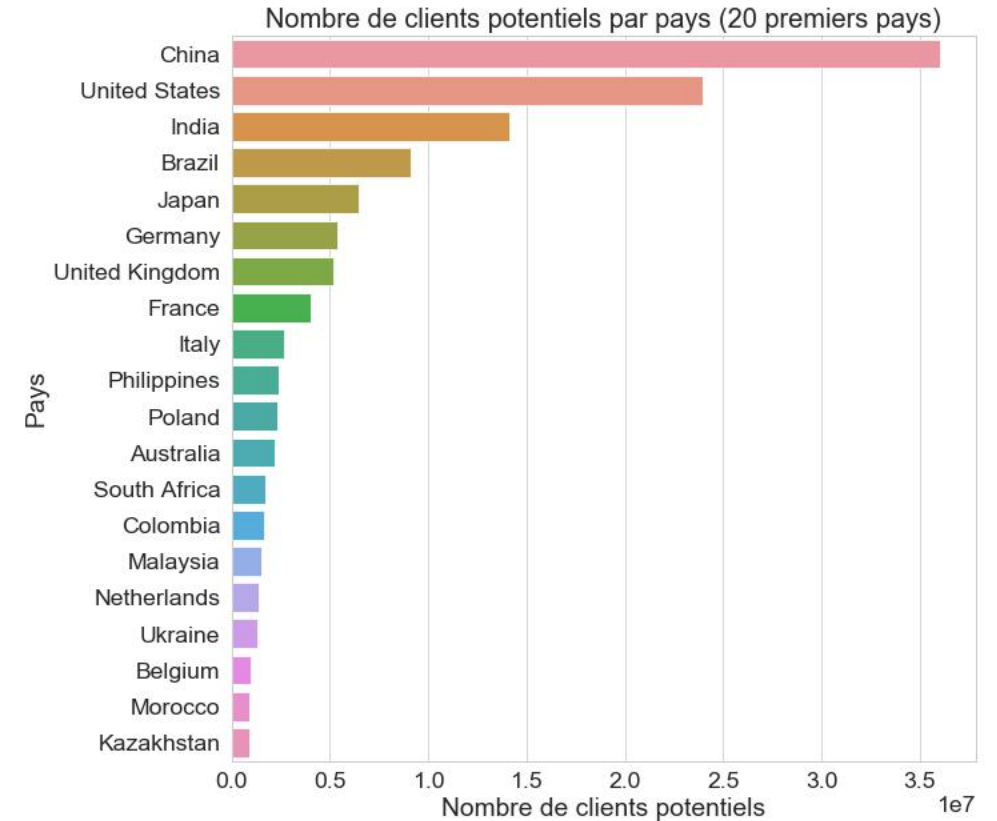
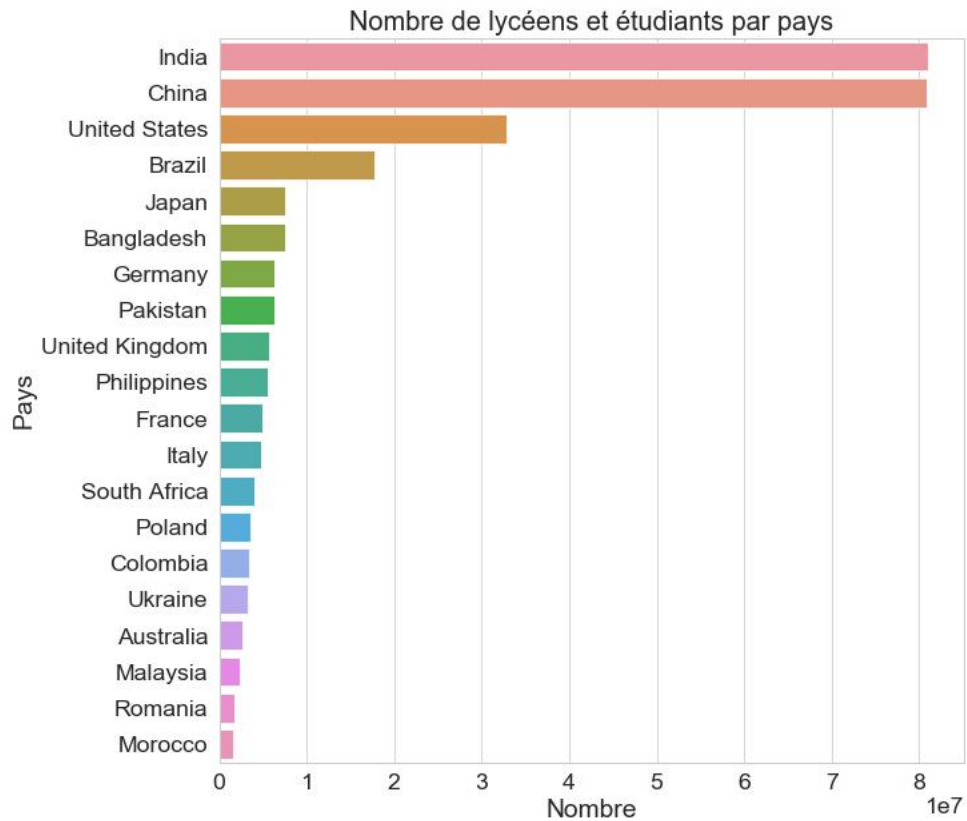


Country Name	
78	India
38	China
189	United States
86	Japan
62	Germany
188	United Kingdom
r pour faire défiler la sortie ;	
106	Malaysia
140	Poland
32	Canada
161	Spain
10	Australia
124	Netherlands
12	Azerbaijan
17	Belgium
170	Sweden
76	Hungary
141	Portugal
83	Israel
45	Czech Republic



Quel potentiel pour ces pays ?

Comparaison des pays- Estimation du nb de clients – 20 premiers pays



```
1 #on va multiplier ce nombre avec le taux de pénétration d'internet pour avoir une estimation du nombre de clien
2 df_countries['potential_customers'] = df_countries['customers'] * df_countries['IT.NET.USER.P2']/100
```


Conclusion

Le jeux 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)
- Besoin d'autres indicateurs business: pénétration Mooc, dépense internet, proportion d'élève se formant en dehors de leur établissements, ect.. Structure du marché, nouvelle tech..
- Besoin d'information sur la société Academy pour guider l'étude (géo, conçu, langue; ect..)
- Corrélation entre pays semi , tech et autre