



DR. NASSERIN HAMIMED

Rapport Analyse des données

Système Éducatif du monde arabe

Le client



Academy est une start-up de la EdTech qui propose des contenus de **formation en ligne** de niveau lycée et université



Objectif du projet

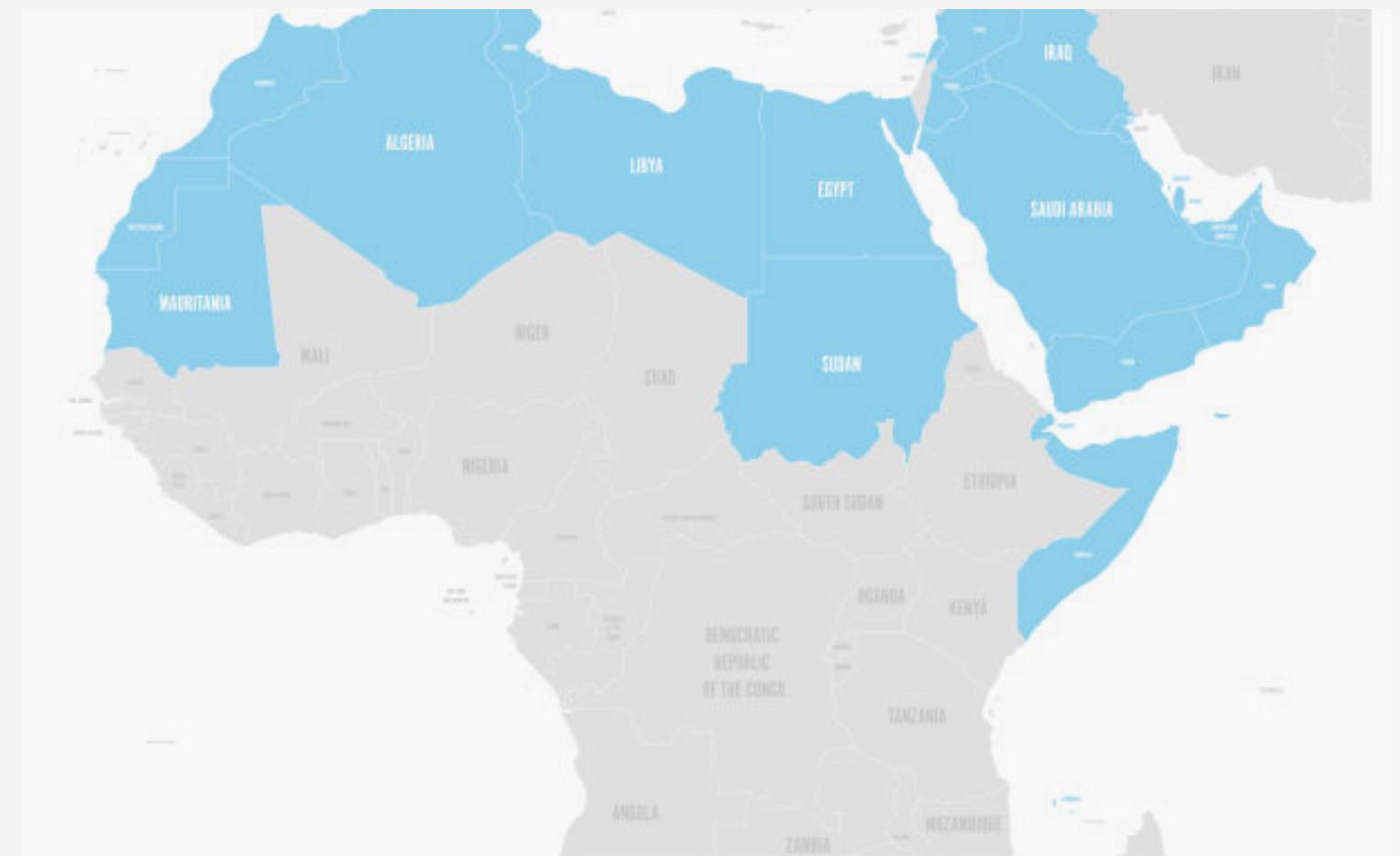
Expansion à l'international, dans les pays arabes.



Dans les prochains 5 minutes

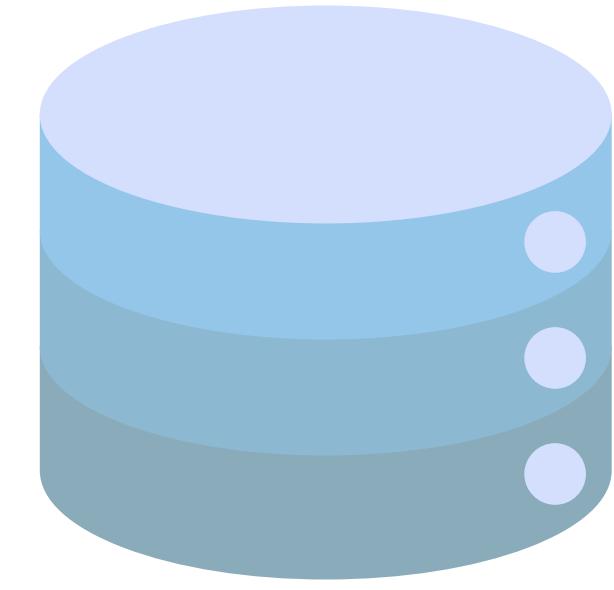
Notre objectif

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



Plan

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



Les données

Word Bank Group

Présentation du jeu de données



WORLD BANK GROUP

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

Valeurs manquantes : Beaucoup de valeurs manquantes
Aucun doublon

EdStatsCountry.csv

Informations globales sur l'économie de chaque pays du monde (et de zones géographiques)

Taille : 241 lignes, 32 colonnes
données depuis 1970

Valeurs manquantes : Quelques valeurs manquantes (sauf pour la variable Unamed 31 qui est vide)
Aucun doublon



Analyse pré exploratoire

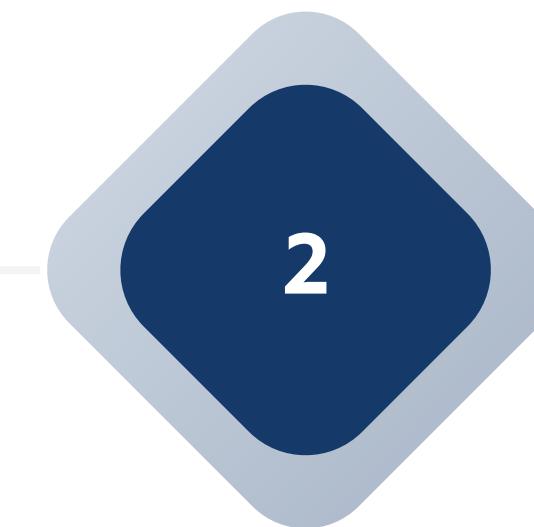
Qualité des données

Processus de l'analyse pré exploratoire



Connaître les données

- Quelles informations?
- Quelles années ?



Identifier les indicateurs exploitables

- Quantité de données manquantes



Comparer les pays

- Quels indicateurs choisir?
- Analyse des résultats obtenus?
- Quels sont les pays à cibler par la startup?



Quel est le potentiel de chaque pays

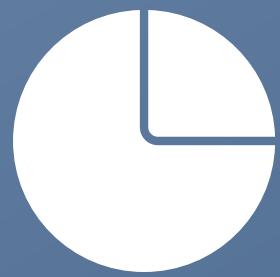
- Comment identifier le potentiel des pays choisis?

Outils utilisés pour l'analyse

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

Description des données

Les données du monde arabe



2 Régions



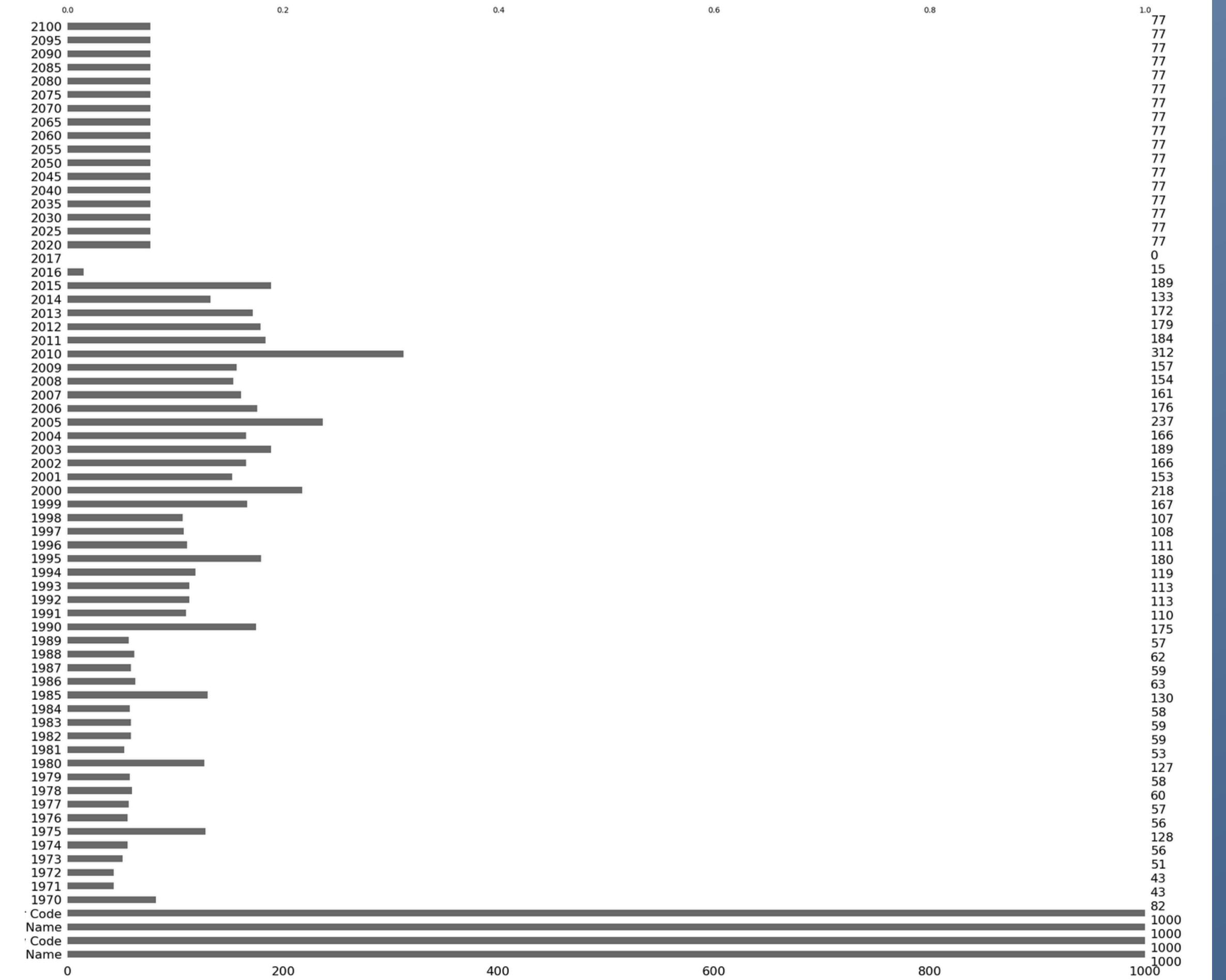
Afrique du
nord et moyen
orient



16 pays



3665 indicateurs
uniques

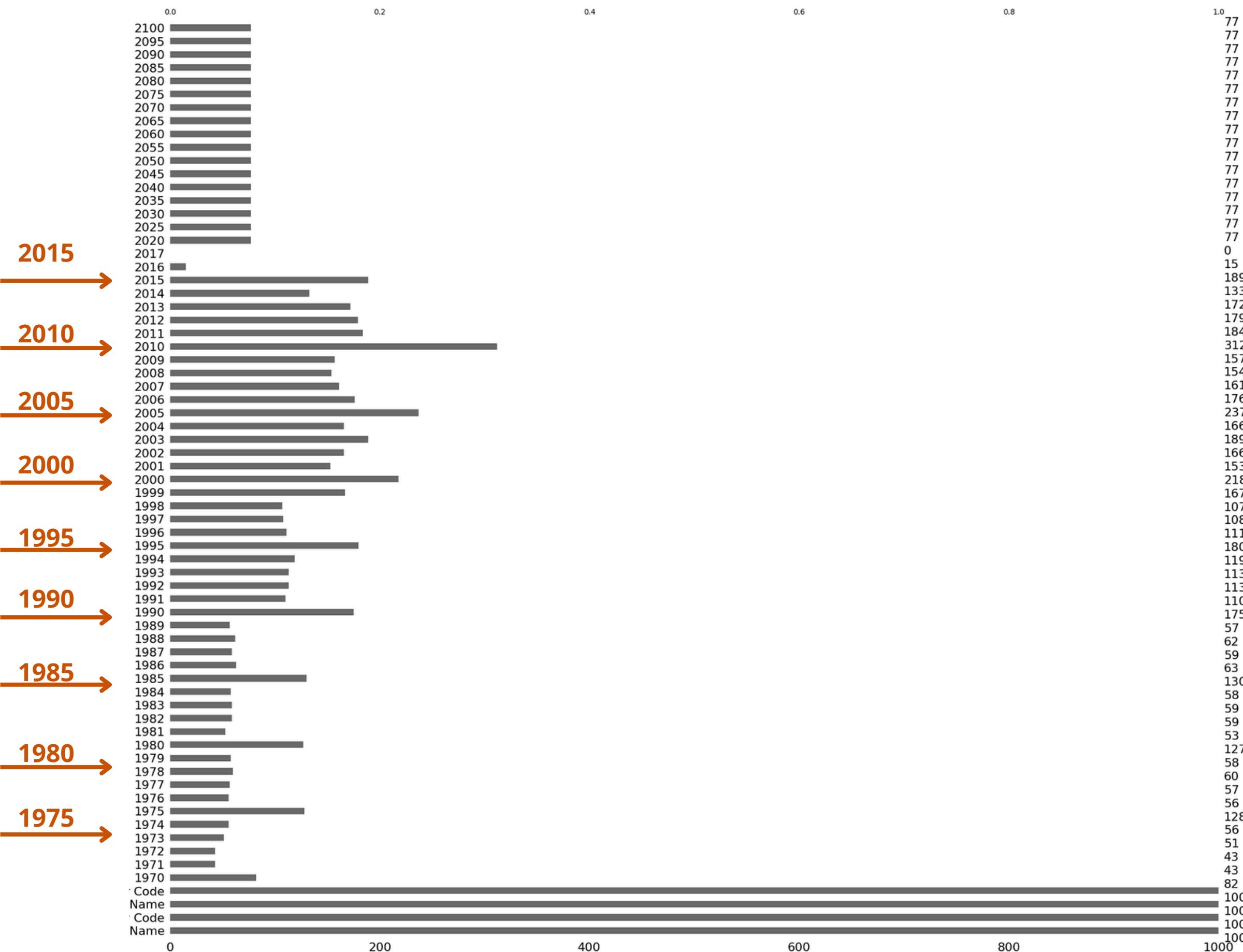


Taux de remplissage des colonnes

Les variables ayant un bon taux de remplissage sont :

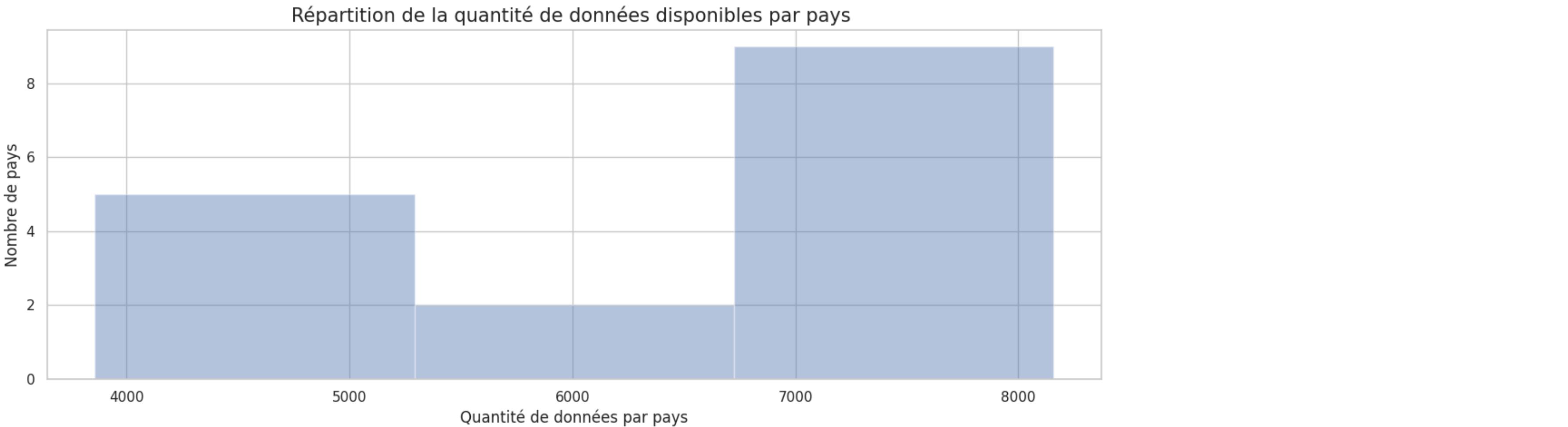
- Country Name
- Country Code
- Indicator Name
- Indicator Code
- les colonnes 1970 - 2015

Choix des années



Nous avons des pics dans le taux de remplissage des données avec une périodicité de **5 ans**

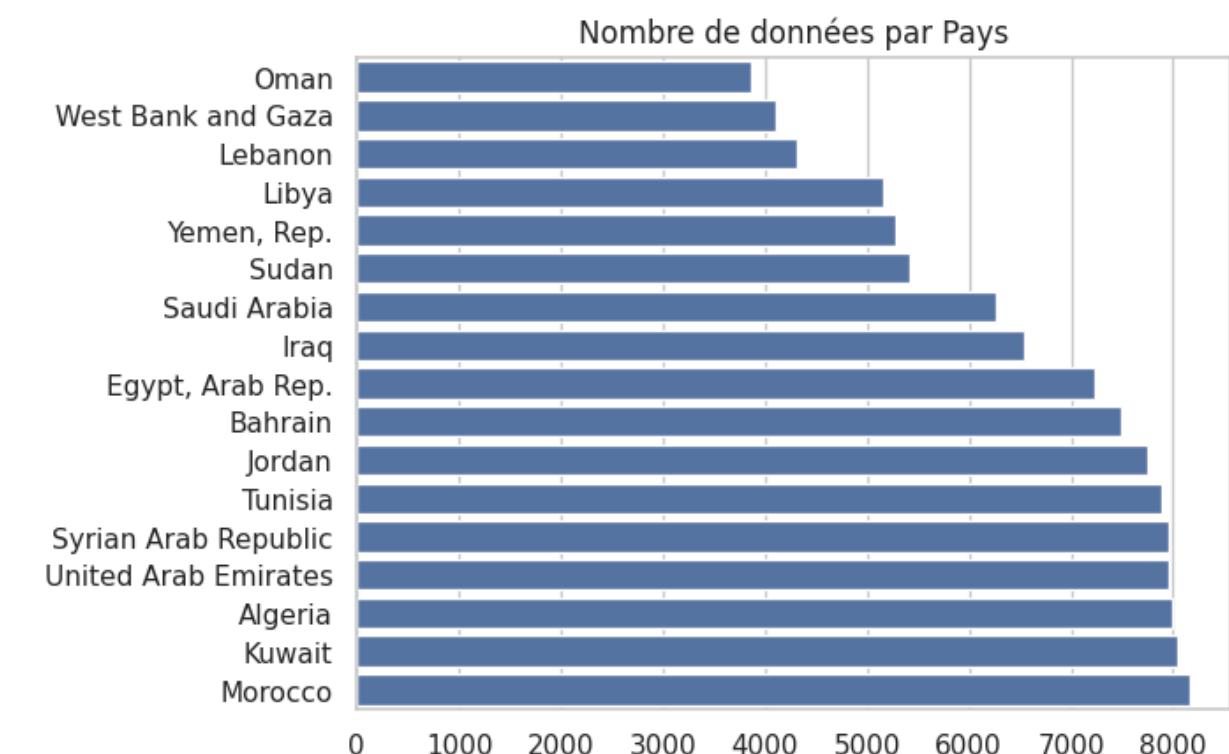
Inégalité du nombre de données par pays



Constat : Inégalité de répartition des données par pays.

Moins d'informations pour ~50 % des pays, il s'agit :

- Les petits pays
- Les pays en situation instable (Guerres, conflits politiques)



Quelles informations conserver?

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

- **EdstatsCountry** : l'association pays-régions

```
data_region = pd.merge(country[['Country Code','Region']],data, on = 'Country Code')  
data_region.head(3)
```

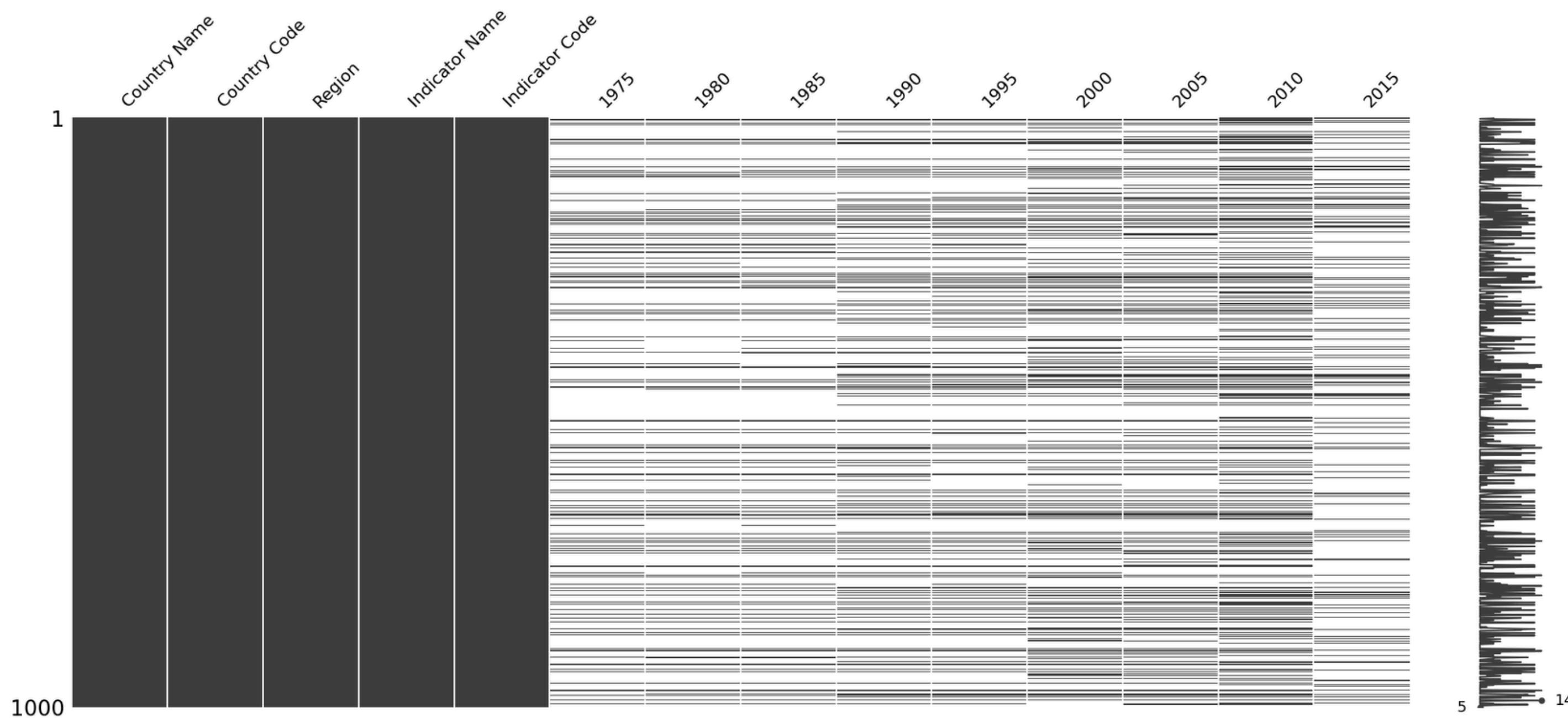
- **EdstatsData** : les noms des pays, indicateurs, régions et les valeurs pour les années 1975 à 2015

```
data_filtered= data_region.loc[:, ['Country Name','Country Code','Region','Indicator Name','Indicator Code',  
'1975','1980','1985','1990','1995','2000','2005','2010','2015']]  
data_filtered.head(3)
```

Autres données : non nécessaires à ce stade.

Identifier les indicateurs exploitables

Identifier les NaN graphiquement (Blanc = donnée manquante)

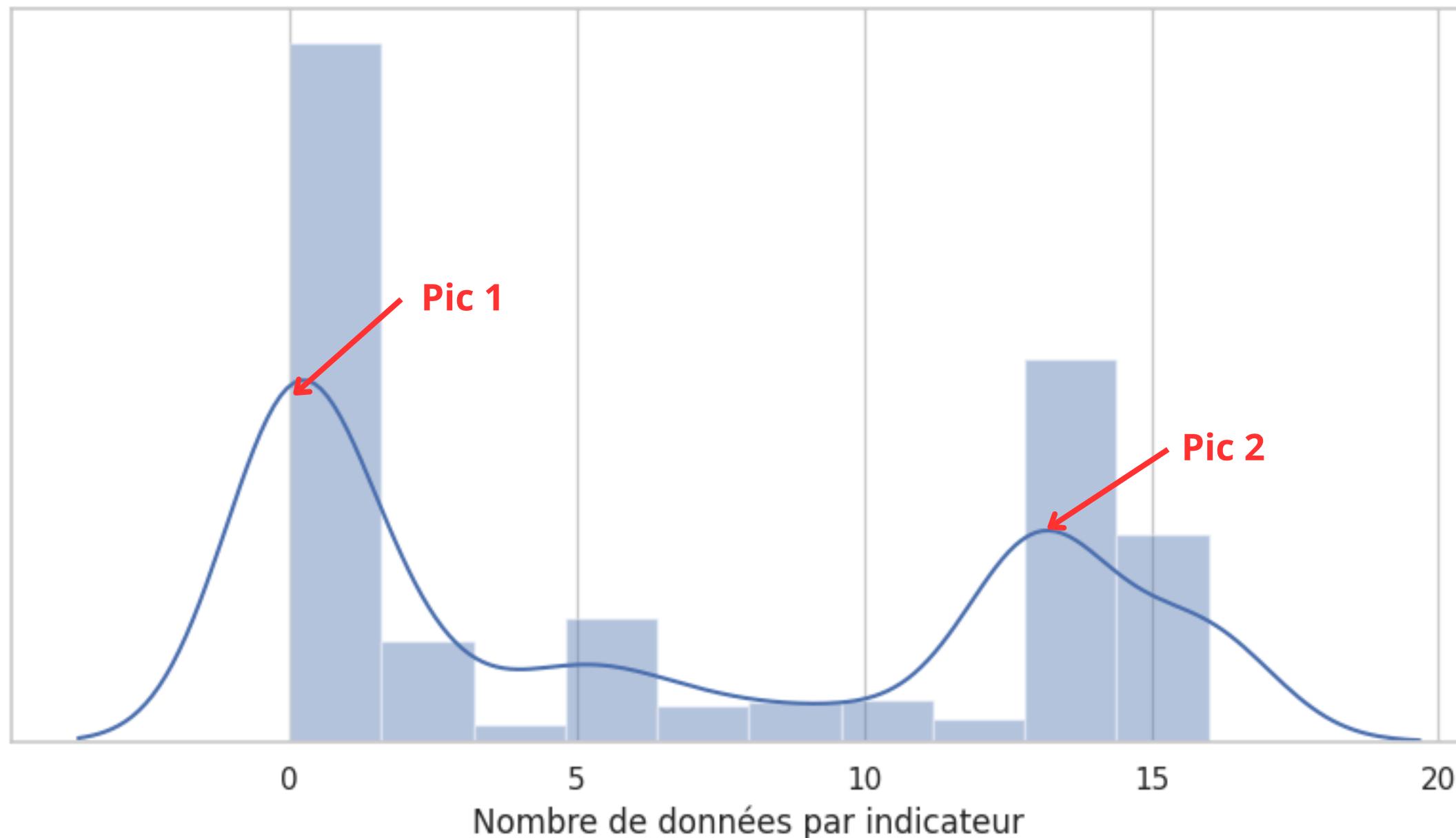


- Identification des indicateurs peu informatifs
- Identification des indicateurs les plus informatifs

Identifier les indicateurs exploitables

Les données manquantes

Répartition du nombre de données par indicateur pour la période 1975- 201



On constate deux distributions cela est dû au fait que :
--> Certains indicateurs ont un taux de remplissage à 100% tandis que d'autres sont renseignés uniquement pour certaines années (tous les 5 ans)

Sélection des indicateurs

Trier selon la quantité d'informations disponible

```
data_arab[['Indicator Name', 'All']].groupby('Indicator Name').count().sort_values(by='All', ascending=False)
```

Indicator Name	
Population growth (annual %)	240
Population, total	240
GDP per capita (current US\$)	233
GDP at market prices (current US\$)	233
Internet users (per 100 people)	233
Enrolment in primary education, both sexes (number)	232
Enrolment in secondary general, female (number)	232
Enrolment in secondary general, both sexes (number)	232
Enrolment in secondary education, both sexes (number)	231
Enrolment in primary education, female (number)	230

Sélection des indicateurs

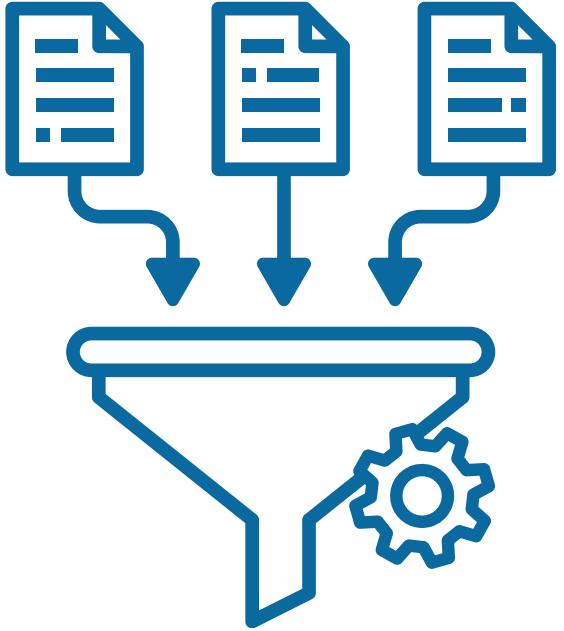
Après une phase d'observation des indicateurs : indicateurs retenus

Les indicateurs retenus pour cette étude sont :

- Population
- Enrolment in upper secondary education both sexes
- Enrolment in tertiary education, all programmes, both sexes
- Internet users (per 100 people)
- GDP per capita (current US\$)

Calcul de nouvelles variables

à partir des indicateurs choisis

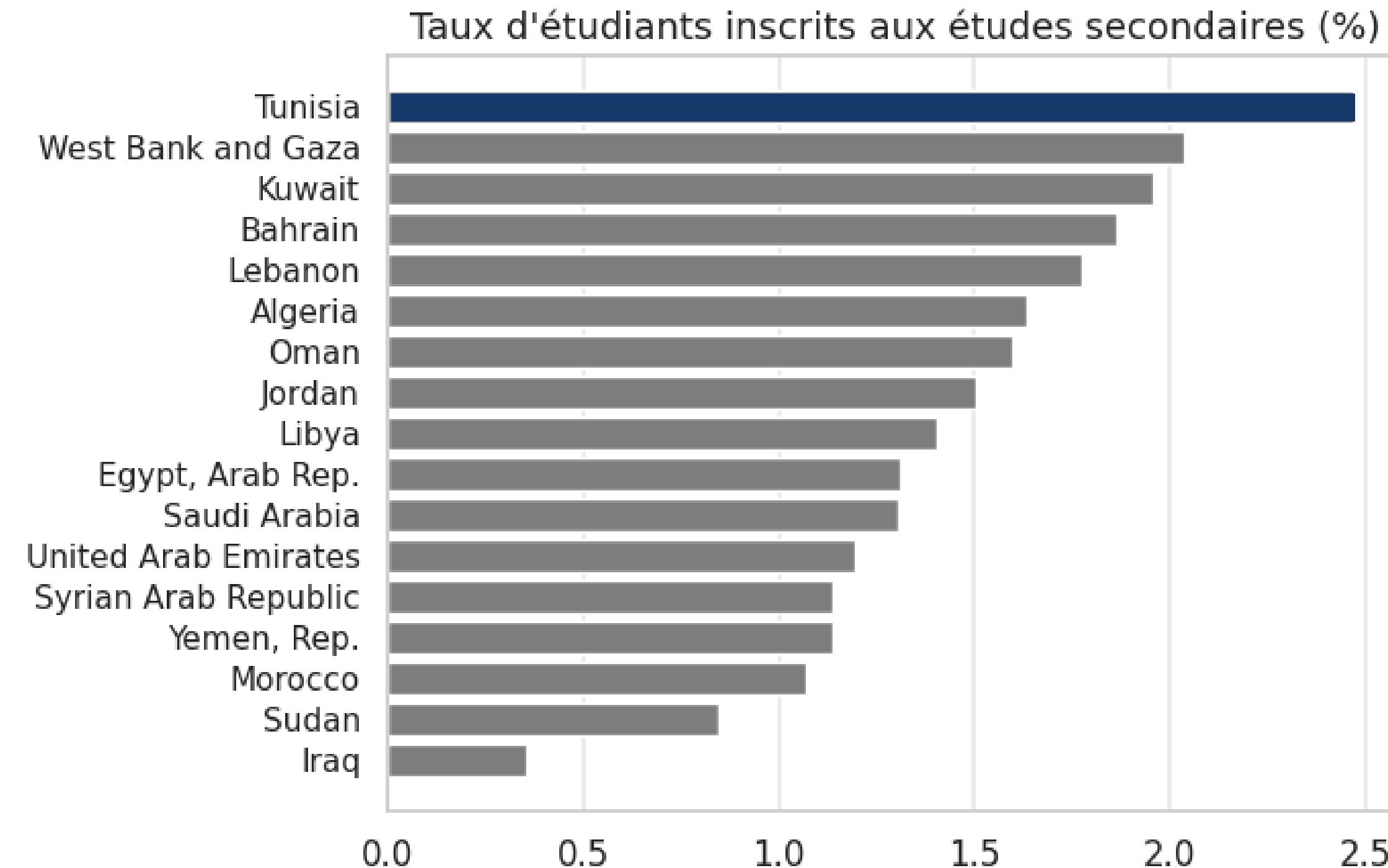




Inscription aux études secondaires et supérieures

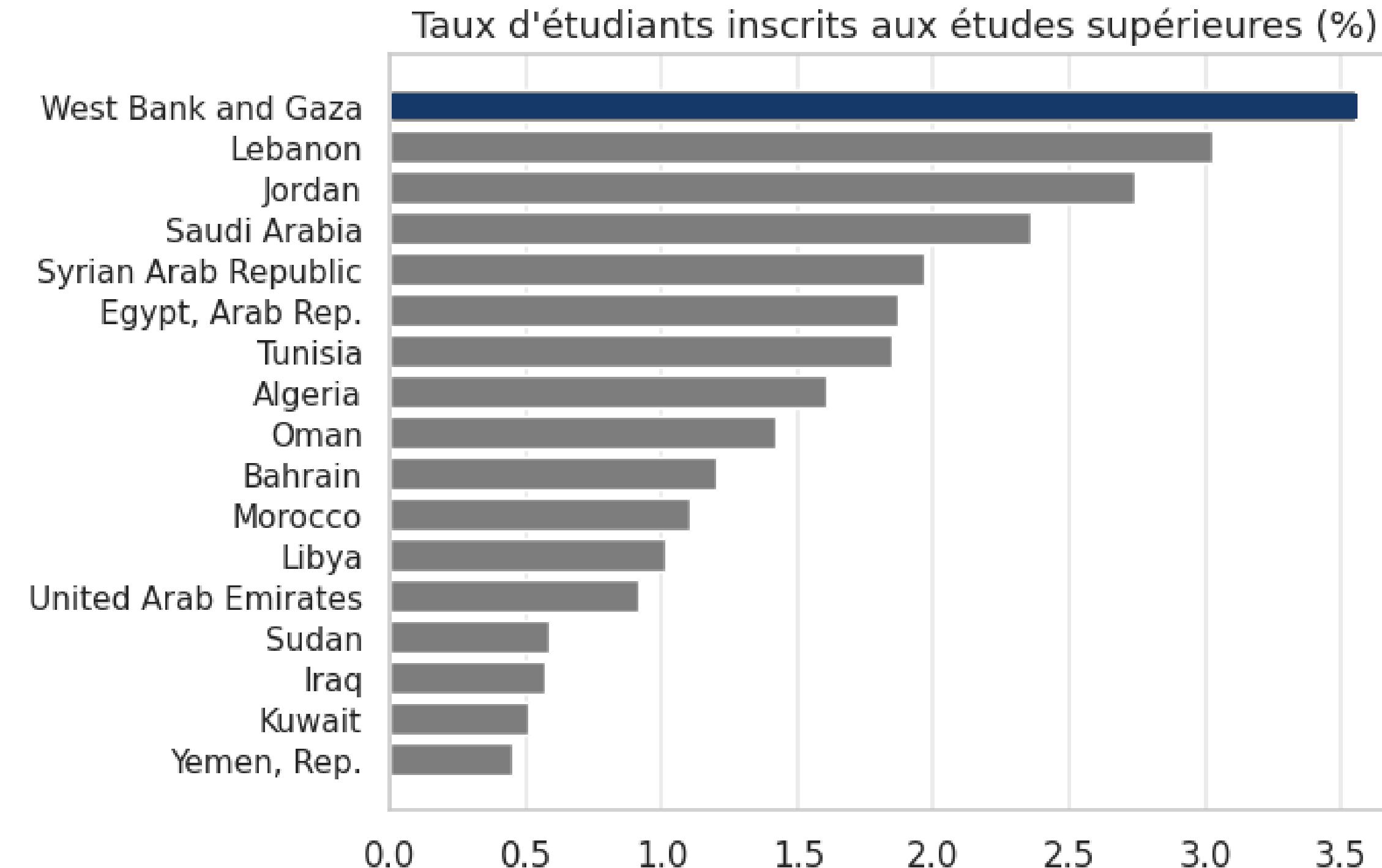
Taux d'étudiants inscrits aux études secondaires (%)

Exemples d'ordres de grandeur (moyenne sur les années)



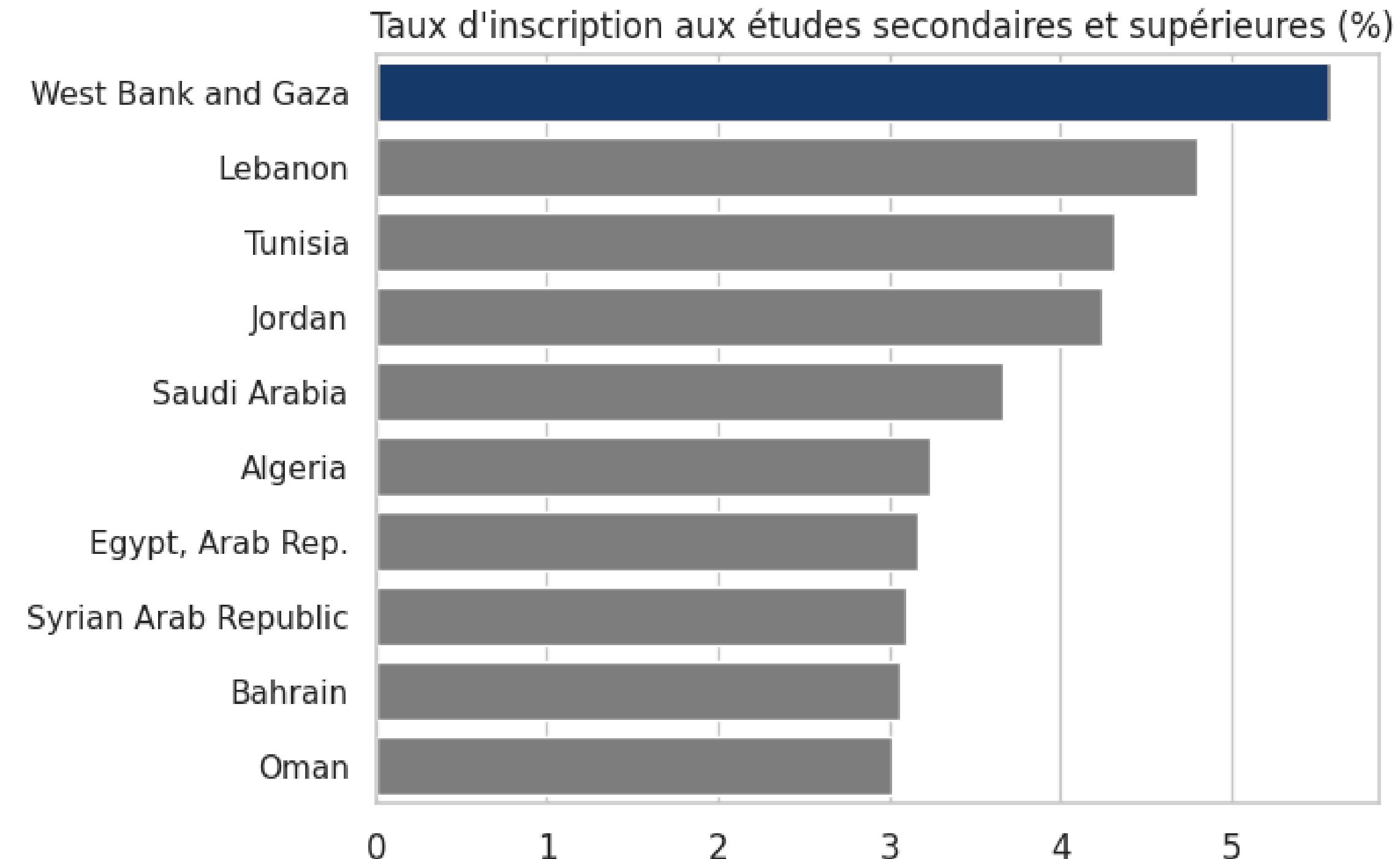
Taux d'étudiants inscrits aux études supérieure (%)

Exemples d'ordres de grandeur (moyenne sur les années)



Taux d'inscription aux études secondaires et supérieures (%)

Exemples d'ordres de grandeur (moyenne sur les années)

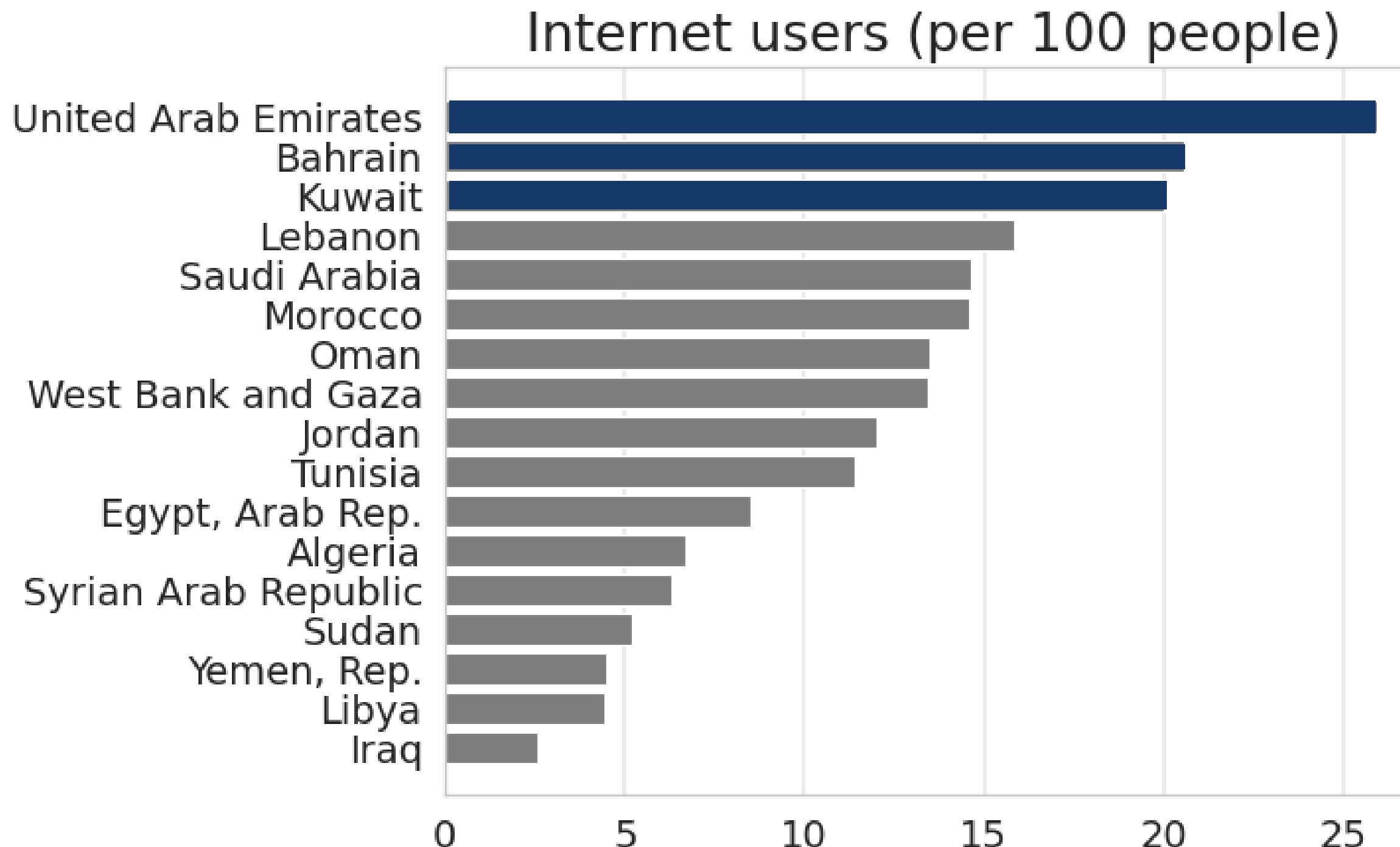




Pénétration d'internet

Internet

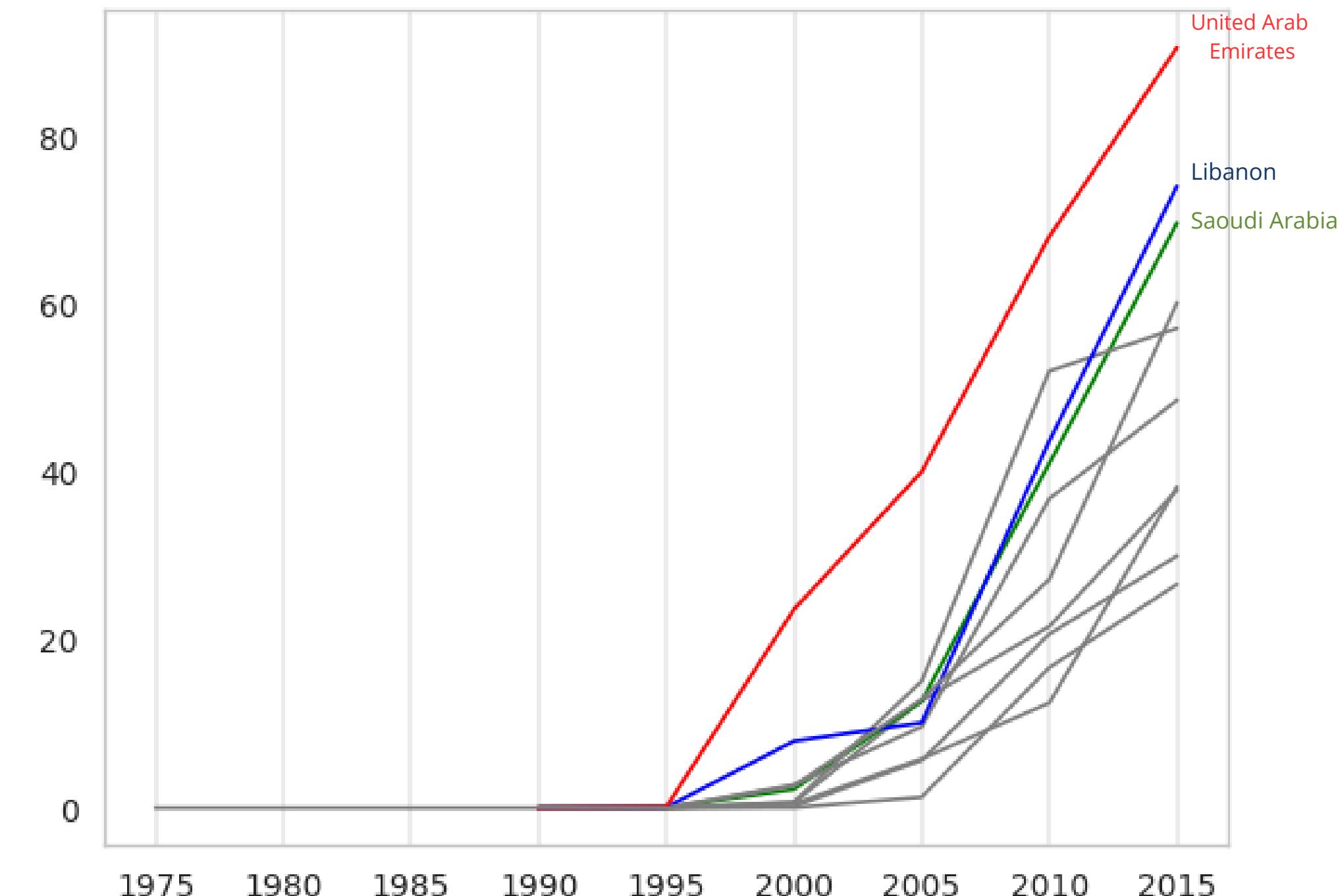
Exemples d'ordres de grandeur (moyenne sur les années)



Pénétration d'internet dans chaque pays

2me indicateur

Evolution de l'utilisation d'internet.

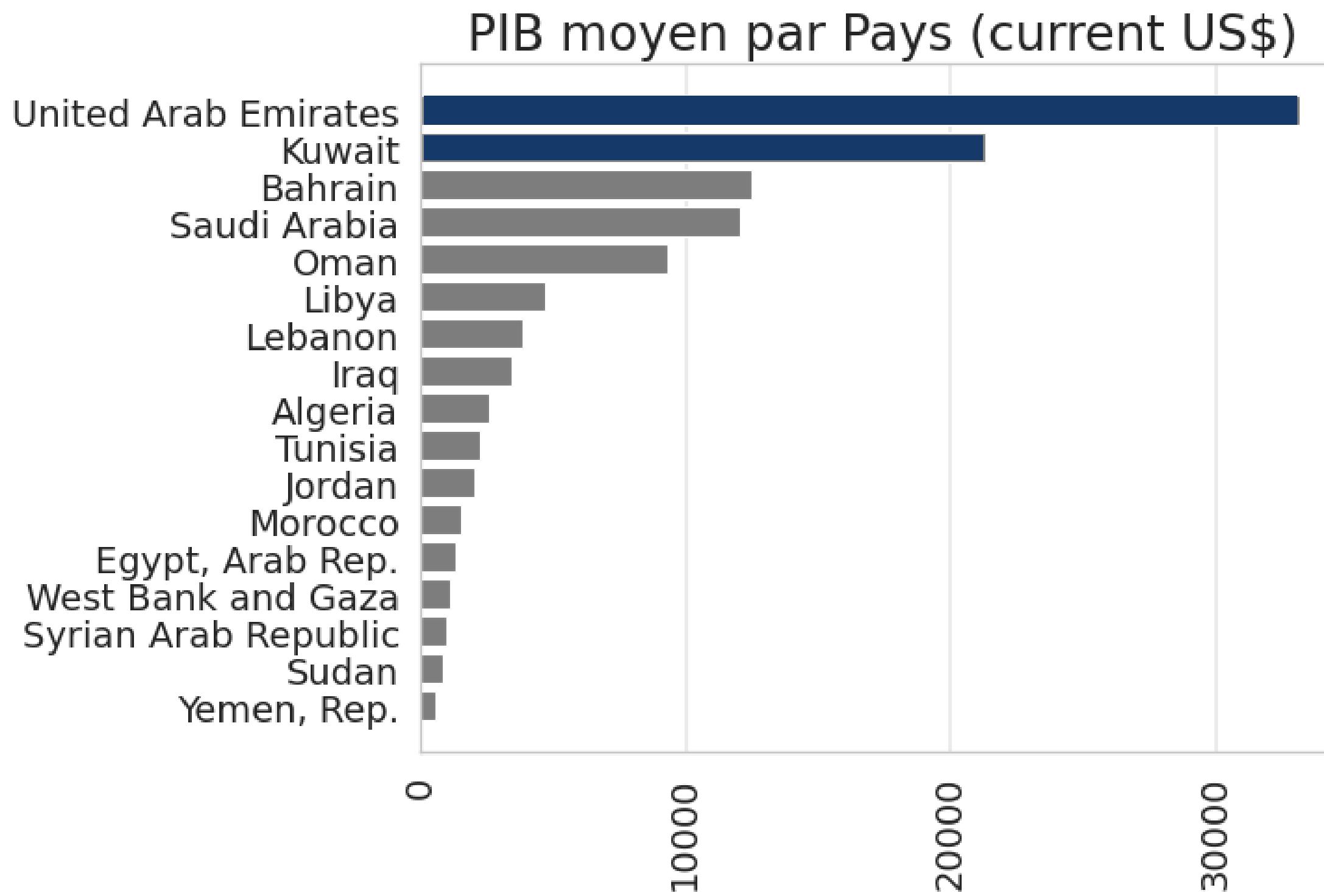




PIB

PIB

Exemples d'ordres de grandeur (moyenne sur les années)





Clients potentiels ?

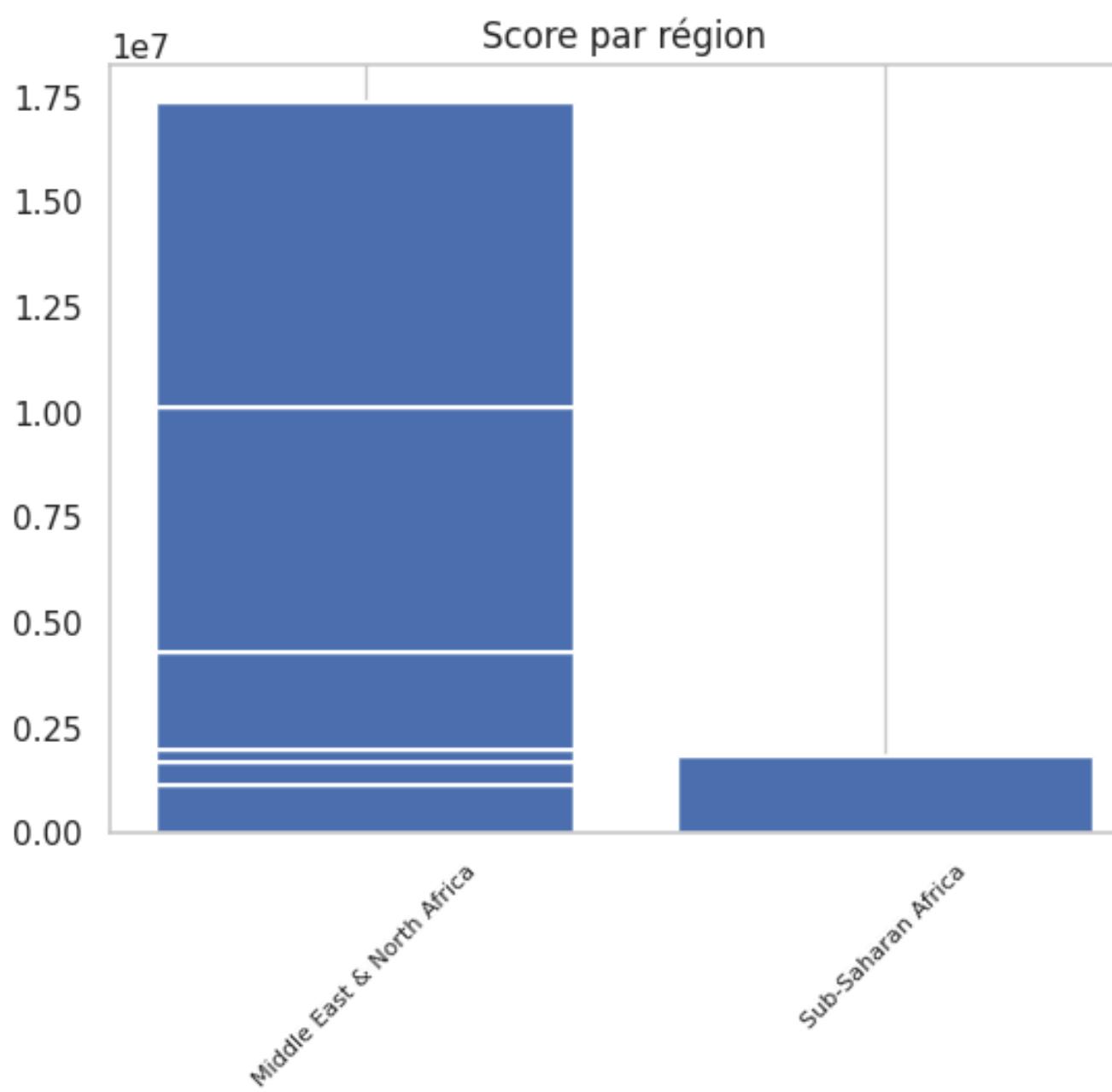
Méthode 1

Calculer à partir du taux de d'inscriptions et la pénétration d'internet

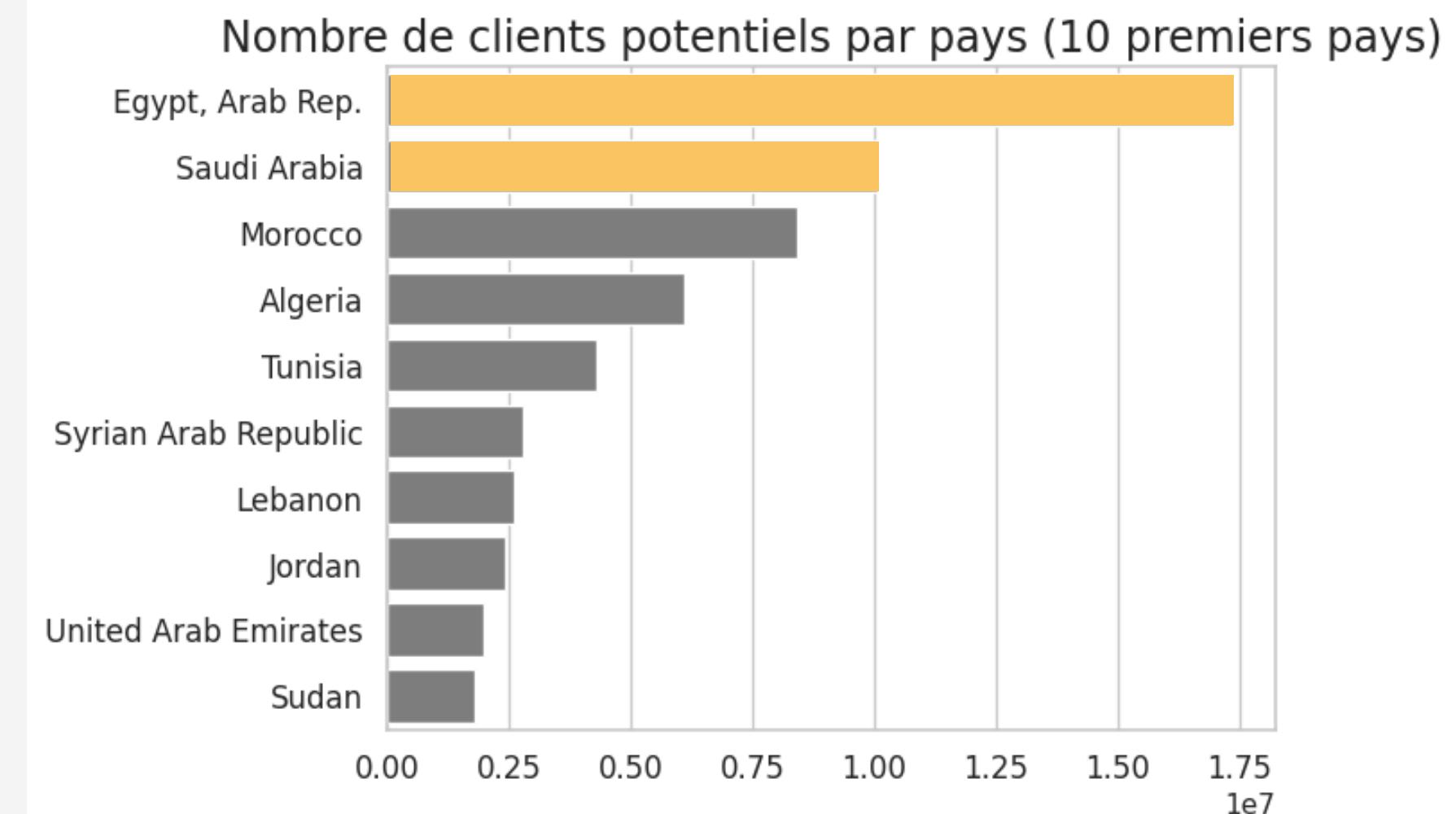
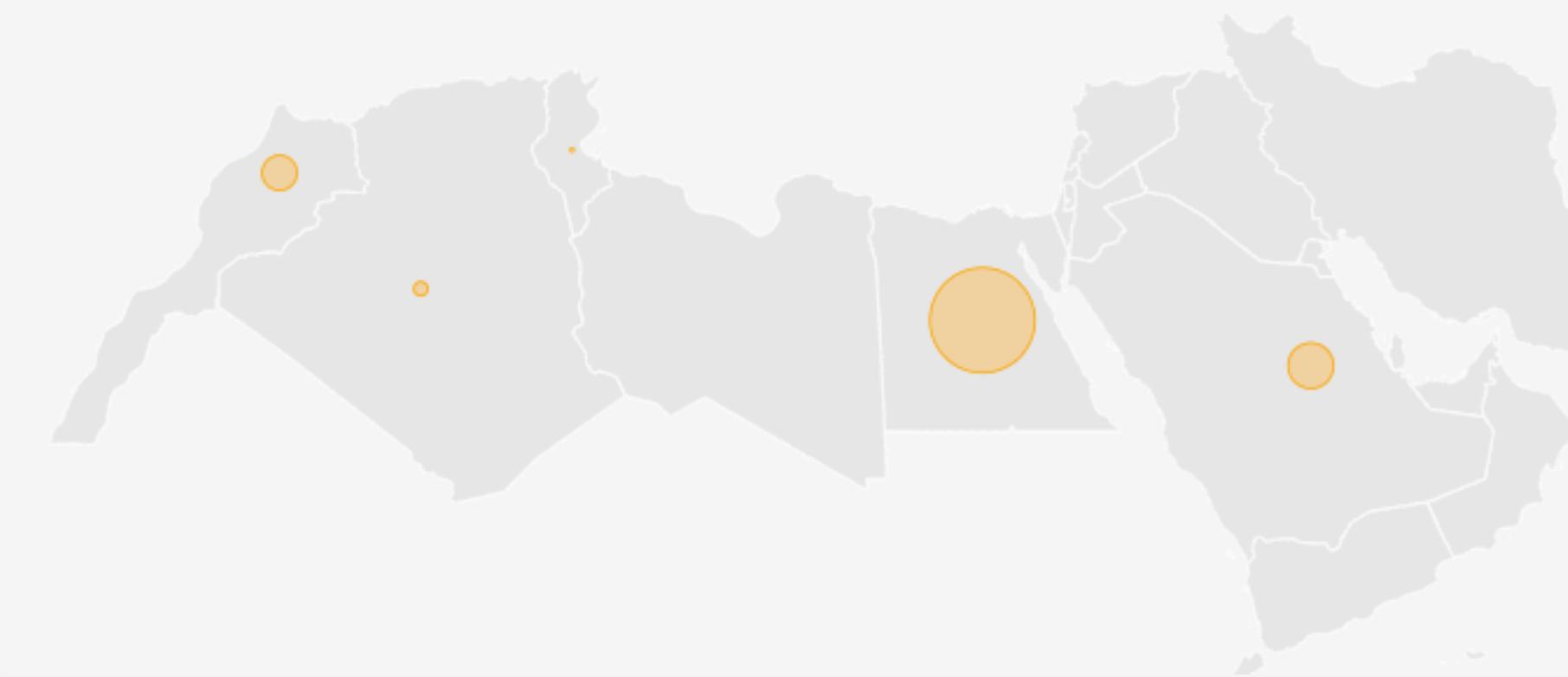
Les indicateurs utilisés pour ce calcul sont :

- Enrolment in upper secondary education both sexes (rate)
- Enrolment in tertiary education, all programmes, both sexes (rate)
- Internet users (per 100 people)

Clients Potentiels



Carte des clients potentiels



Taux de pénétration d'internet comparé aux nombre d'étudiants inscrits



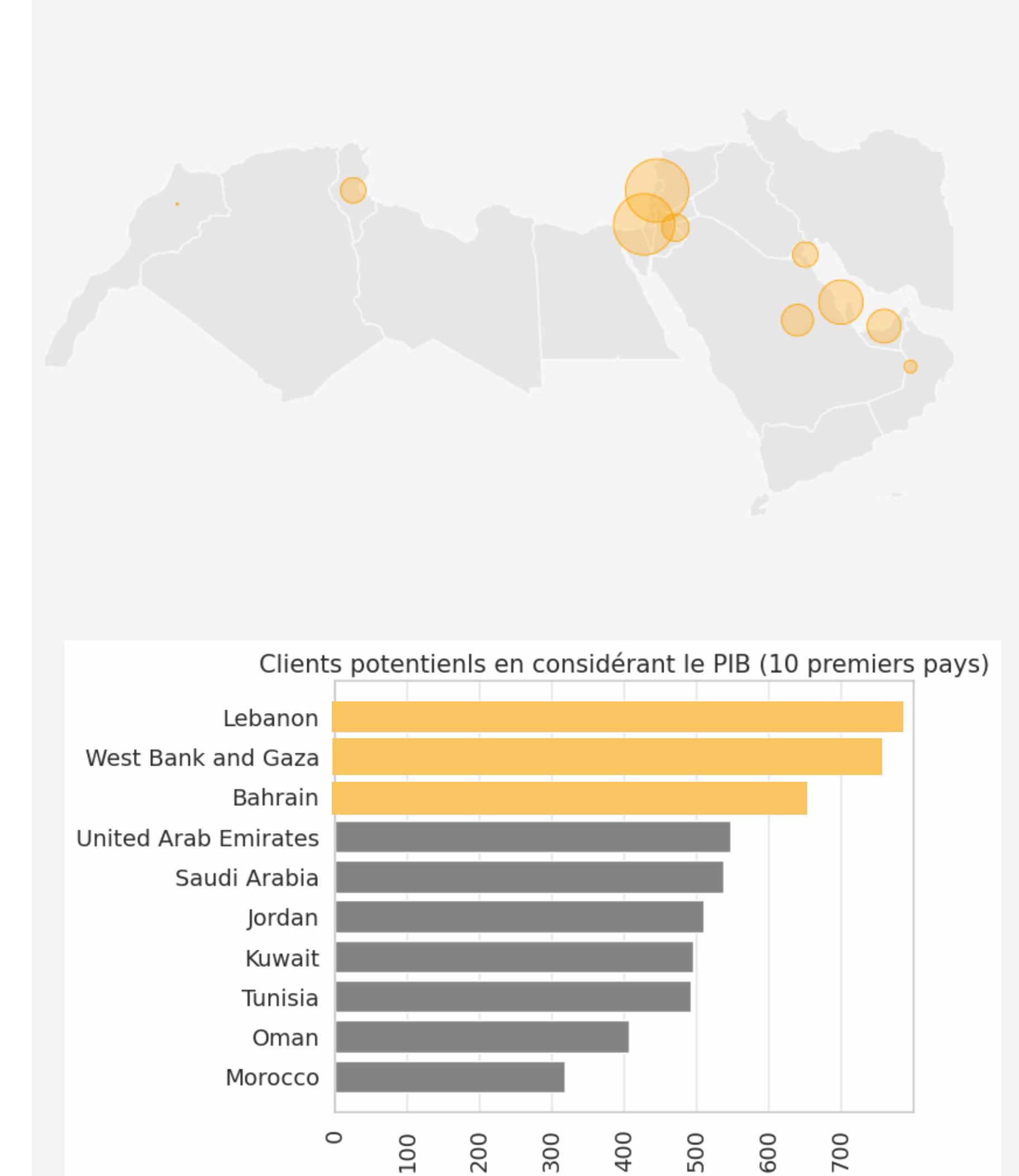
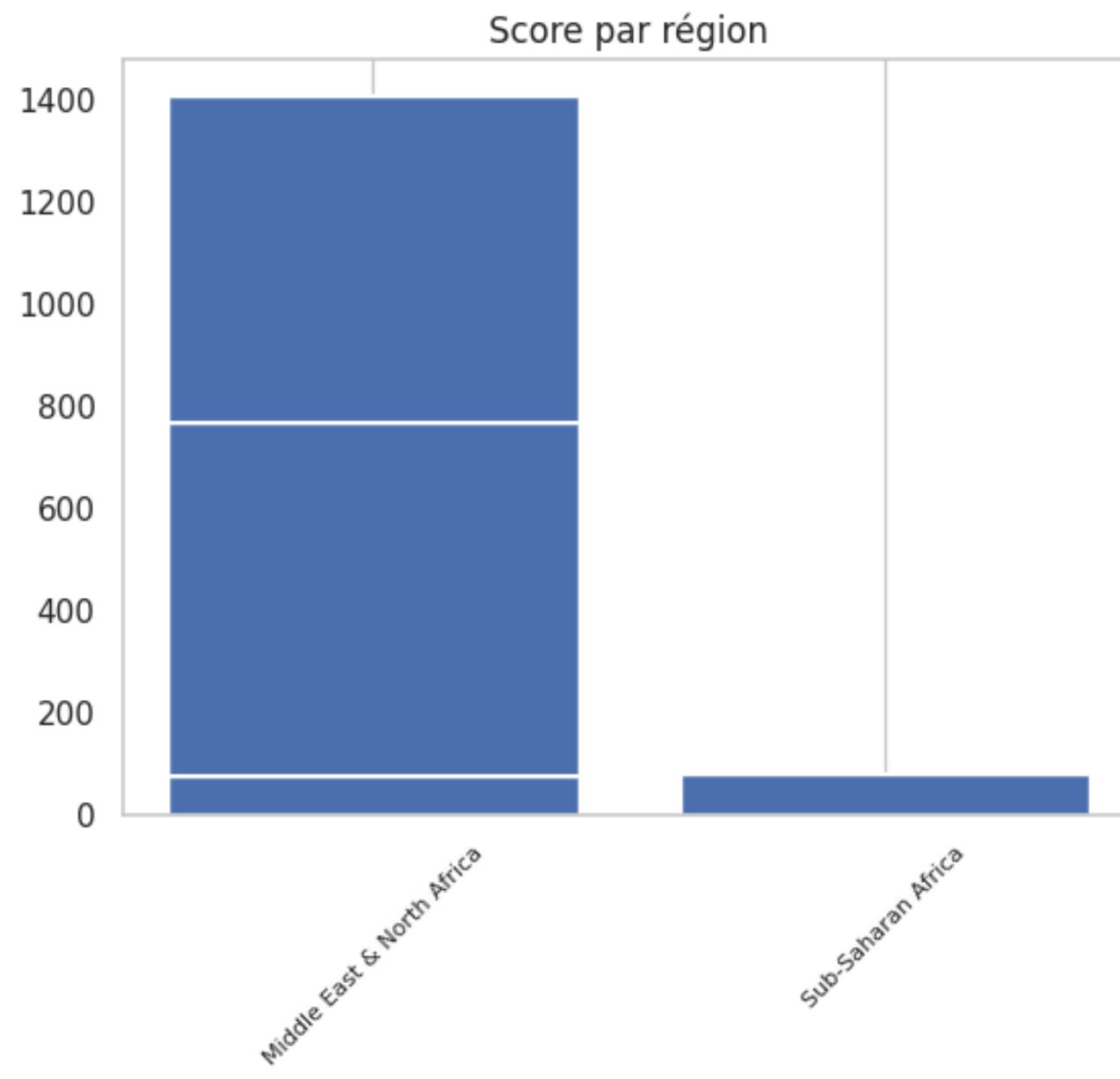
Méthode 2

Prise en compte du PIB

Les indicateurs utilisés pour ce calcul sont :

- Enrolment in upper secondary education both sexes (rate)
- Enrolment in tertiary education, all programmes, both sexes (rate)
- Internet users (per 100 people)
- GDP per capita (current US\$)

Clients Potentiels



Taux de pénétration d'internet comparé au PIB

Taux d'inscription des étudiants en prenant en compte le PIB



Middle East & North Africa
Sub-Saharan Africa

Conclusion

Je conseille de commencer par les pays du **Moyen-Orient**, dont le PIB est élevé ainsi que la pénétration d'internet, tels que les Émirats arabes unis, l'Arabie Saoudite, le Kouwait, la Jordanie et le Liban.

Puis s'installer en **Afrique** en commençant par le Maroc et la Tunisie.



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

Pertinence du jeu de données

- ◆ données relatives à l'éducation et utiles

- ◆ données complémentaires

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

Limites

- Il manque des données pour certains pays (comme le Soudan)
- Certains indicateurs inutilisables (beaucoup de données manquantes pour comparer)
- Aucune information sur la société Academy pour guider l'étude (proximité géographique, concurrence, langue, etc.)