

# Formation Data Scientist

Projet 2

« Analyse des données de systèmes éducatifs »

Soutenance de Projet  
Samir HINOJOSA

24 avril 2021



OPENCLASSROOMS

# Plan de la soutenance

1. Mission
2. Présentation du jeu de données
3. Analyse pré-exploratoire du jeu de données
4. Conclusions sur la pertinence du jeu de données



1.

# Mission



# Mission

Une Start-up de la EdTech qui propose des contenus de formation **en ligne** pour un public de niveau **lycée et université**.



## Le projet d'expansion à l'international

**Analyse exploratoire pour déterminer:**

- Quels sont les pays avec un fort potentiel de clients pour nos services ?
- Pour chacun de ces pays, quelle sera l'évolution de ce potentiel de clients ?
- Dans quels pays l'entreprise doit-elle opérer en priorité ?

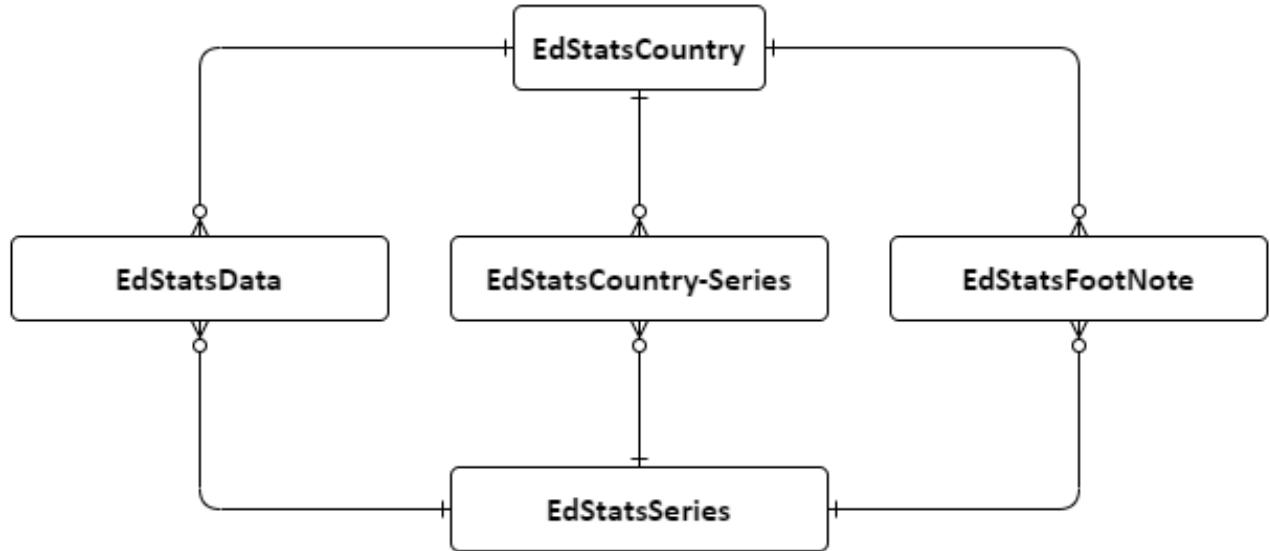


2.

## Présentation du jeu de données



# Présentation du jeu de données



Il y a des « relations » entre les datasets à travers les colonnes

- Country code / CountryCode
- Series code / Indicator code

*Soyez prudent avec ces relations.*

# Présentation du jeu de données

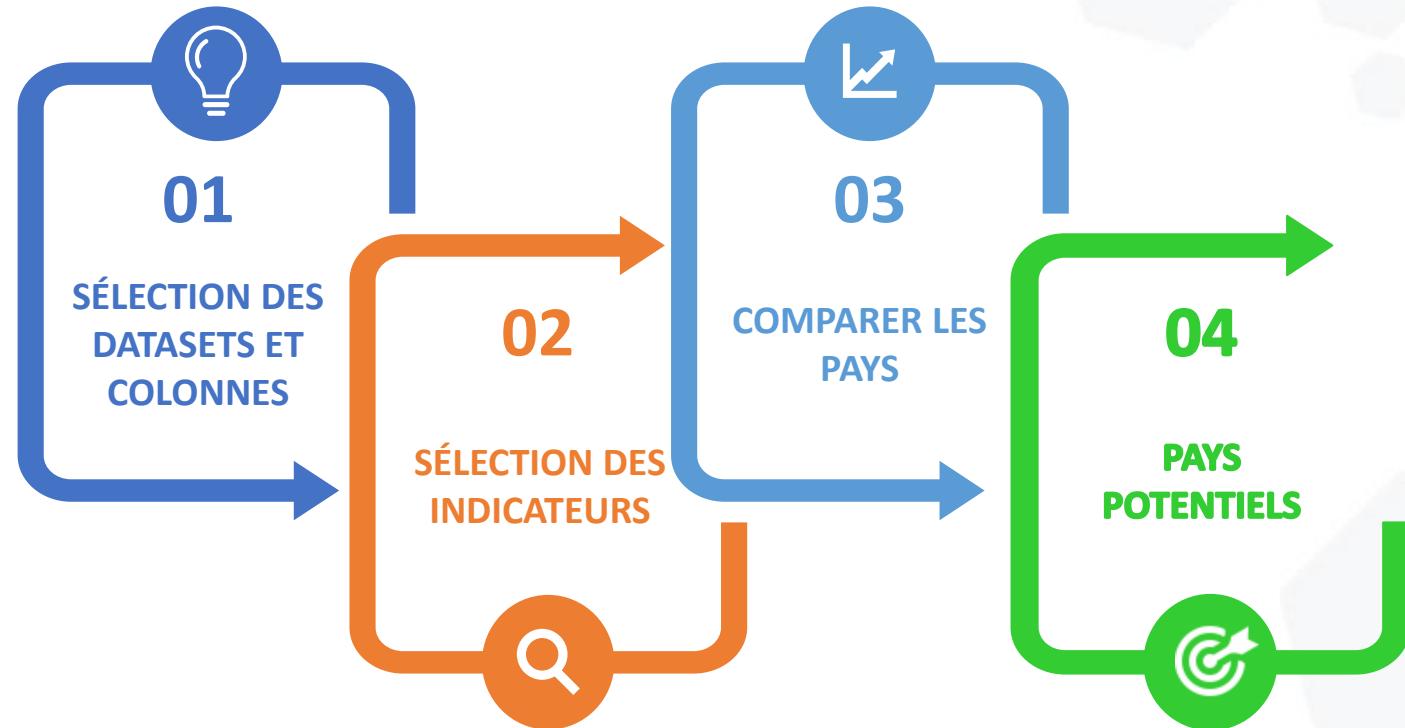
EdStatsCountry.csv	Informations économiques et géographiques générales sur les pays du monde.					
	Taille:	241x32	Percentage de NaN:	30.52 %	Doublon:	0
EdStatsSeries.csv	Informations générales sur les indicateurs.					
	Taille:	3665x21	Percentage de NaN:	71.72 %	Doublon:	0
EdStatsCountry-Series.csv	Informations relatives aux ensembles de données « Pays » et « Série ». (Contient les descriptions des indicateurs liés aux pays)					
	Taille:	613x4	Percentage de NaN:	25.0 %	Doublon:	0
EdStatsFootNote.csv	Contient l'année d'origine des données ainsi qu'une description des indicateurs.					
	Taille:	643638x5	Percentage de NaN:	20.0 %	Doublon:	0
EdStatsData.csv	Jeu de données principal qui contient en détail les informations sur les pays et les indicateurs par année.					
	Taille:	886930x70	Percentage de NaN:	86.1 %	Doublon:	0

3.

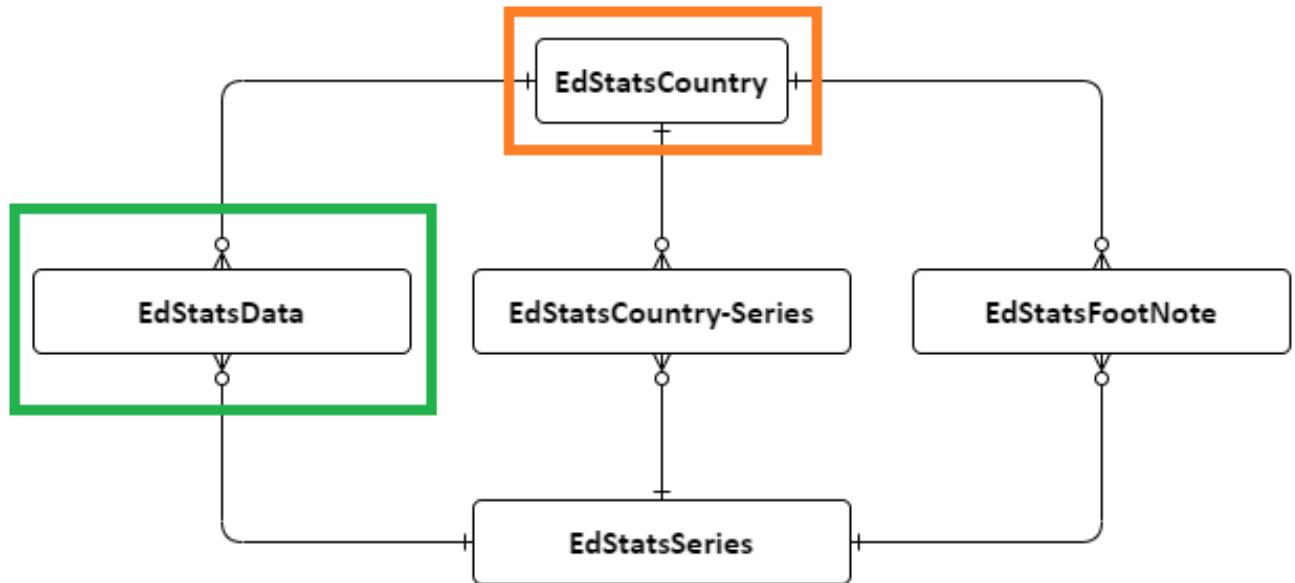
## Analyse pré-exploratoire du jeu de données



# Processus d'analyse



# 01. Sélection des jeux de données et les colonnes



Les jeux de données à travailler sont:

- **EdStatsData.csv** : Jeu de données principal
- **EdStatsCountry.csv** : Pour compléter les informations sur les pays dans le jeu de données "EdStatsData.csv".



Les colonnes par jeu de données à considérer sont:

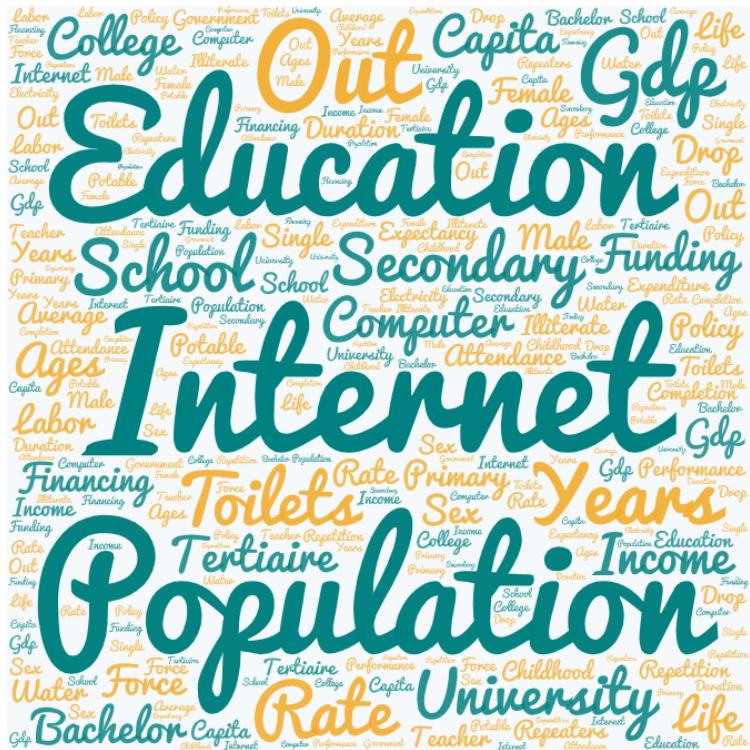
- |  |   |
|--|---|
| ▪ <b>EdStatsData.csv</b>   | ▪ <b>EdStatsCountry.csv</b>   |
| <ul style="list-style-type: none"><li>▪ Country Name</li><li>▪ Country Code</li><li>▪ Indicator Name</li><li>▪ Indicator Code</li><li>▪ Years until 2021</li></ul> | <ul style="list-style-type: none"><li>▪ Country Code</li><li>▪ Short Name</li><li>▪ 2-alpha code</li><li>▪ Region</li></ul> |



Les pays ont été filtrés selon la liste des pays d'après ISO-3166-1<sup>1</sup>

1 - <https://datahub.io/core/country-list>

## 02. Sélection des Indicateurs



Au début, il y avait 3665 indicateurs uniques sur [EdStatsData.csv](#)



2 listes de mots-clés ont été définis

- **Mots-clés liés à l'objectif de la mission**
  - **Mots-clés PAS liés à l'objectif de la mission**



*Après avoir réalisé les filtres, 296 indicateurs ont été obtenus.*

# Liste d'indicateurs sélectionnés

Après une phase d'observation, les indicateurs retenus sont

Indicator Code	Indicator Name	Renamed Indicator
IT.NET.USER.P2	Internet users (per 100 people)	Internet users
SE.TER.ENRL	Enrolment in tertiary education, all programmes, both sexes (number)	Enrolment in tertiary education
UIS.E.3	Enrolment in upper secondary education, both sexes (number)	Enrolment in upper secondary education
NY.GDP.PCAP.PP.CD	GDP per capita, PPP (current international \$)	Gross domestic product per capita
SP.POP.1524.TO.UN	Population, ages 15-24, total	Ages 15-24 population
SP.POP.TOTL	Population, total	Total population

## 03. Comparaison des pays



*Quels sont les pays avec un fort potentiel de clients pour nos services ?*

 Sélectionner les colonnes minimales nécessaires pour travailler

 Sélectionner les données les plus récentes pour chaque pays par indicateurs

 Exclure les pays de moins de 10 millions d'habitants

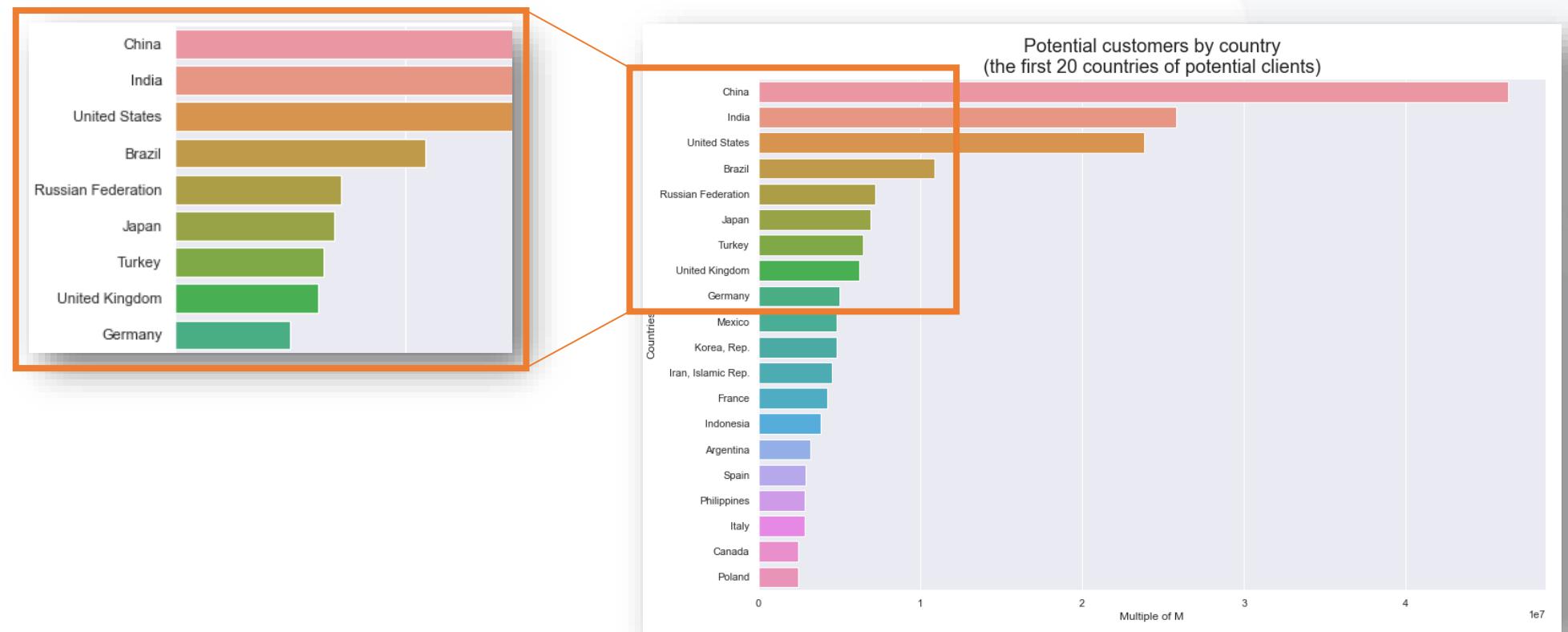
 *Après avoir réalisé les filtres, 86 pays ont été obtenus.*

# *La chine, l'inde, les US et le Brésil sont les pays avec le plus fort potentiel*

*Quels sont les pays avec un fort potentiel de clients pour nos services ?*



(Enrolment in tertiary education + Enrolment in upper secondary education) \* Internet users/100



# 04. Pays potentiels (*Définition de Score*)



*Dans quels pays l'entreprise doit-elle opérer en priorité ?*

*Pour chacun de ces pays, quelle sera l'évolution de ce potentiel de clients ?*



## Hypothèses

- L'indicateur le plus important est « Enrolment in education » car c'est le marché cible
- Le deuxième indicateur le plus important est « Internet users » car l'Académie propose une formation en ligne
- Le prix de la formation n'est pas cher

Renamed Indicator	Weighing	Indicator Code	Indicator Name
Enrolment in education	50 %	SE.TER.ENRL UIS.E.3	Enrolment in tertiary education, all programmes, both sexes (number) Enrolment in upper secondary education
Internet users	35 %	IT.NET.USER.P2	Internet users (per 100 people)
Gross domestic product per capita	15 %	NY.GDP.PCAP.PP.C D	GDP per capita, PPP (current international \$)

# *La Chine, l'Inde et les US sont les pays avec le plus fort potentiel*

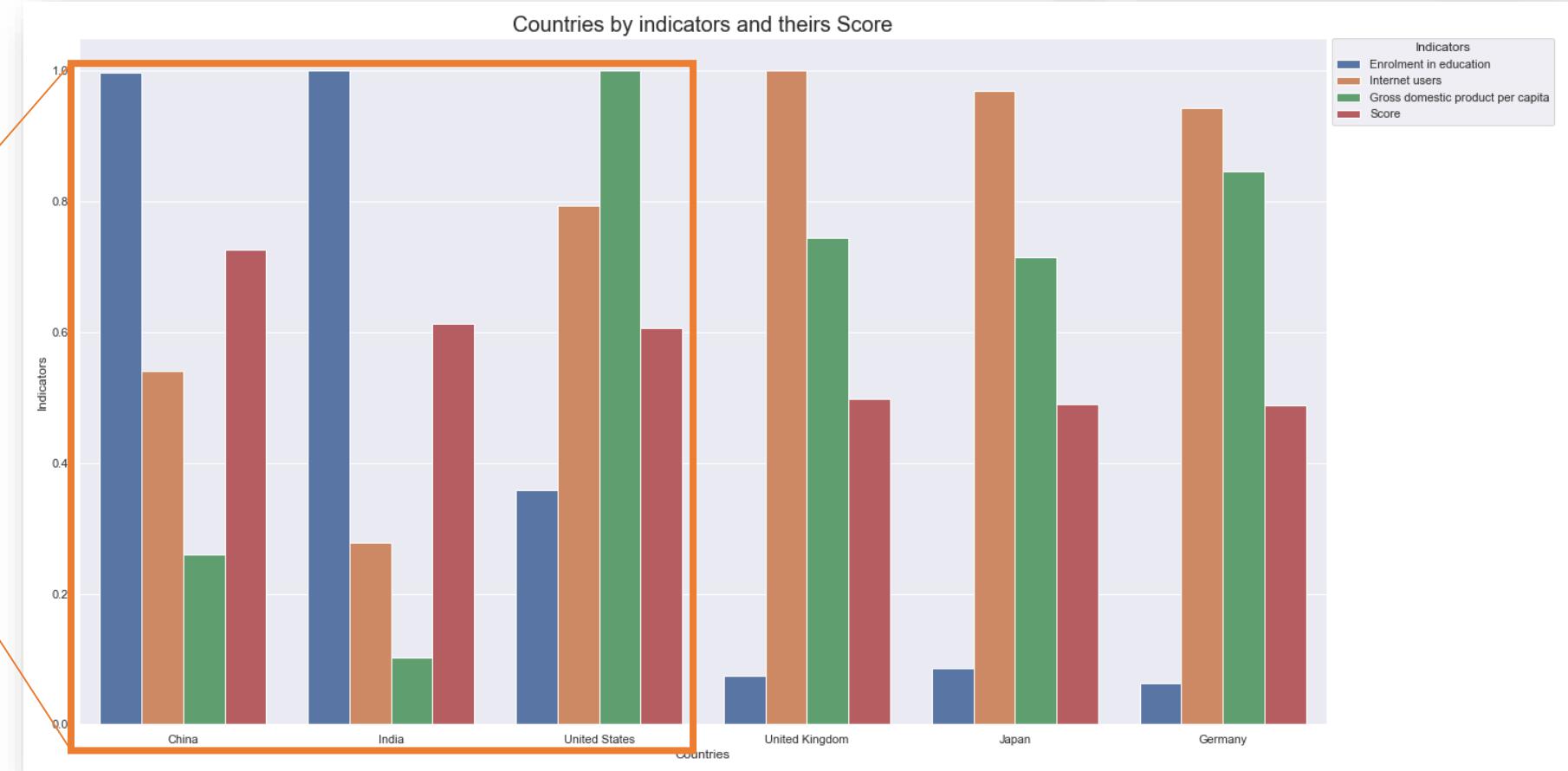
Dans quels pays l'entreprise doit-elle opérer en priorité ?



La Chine

L'Inde

Les États-Unis

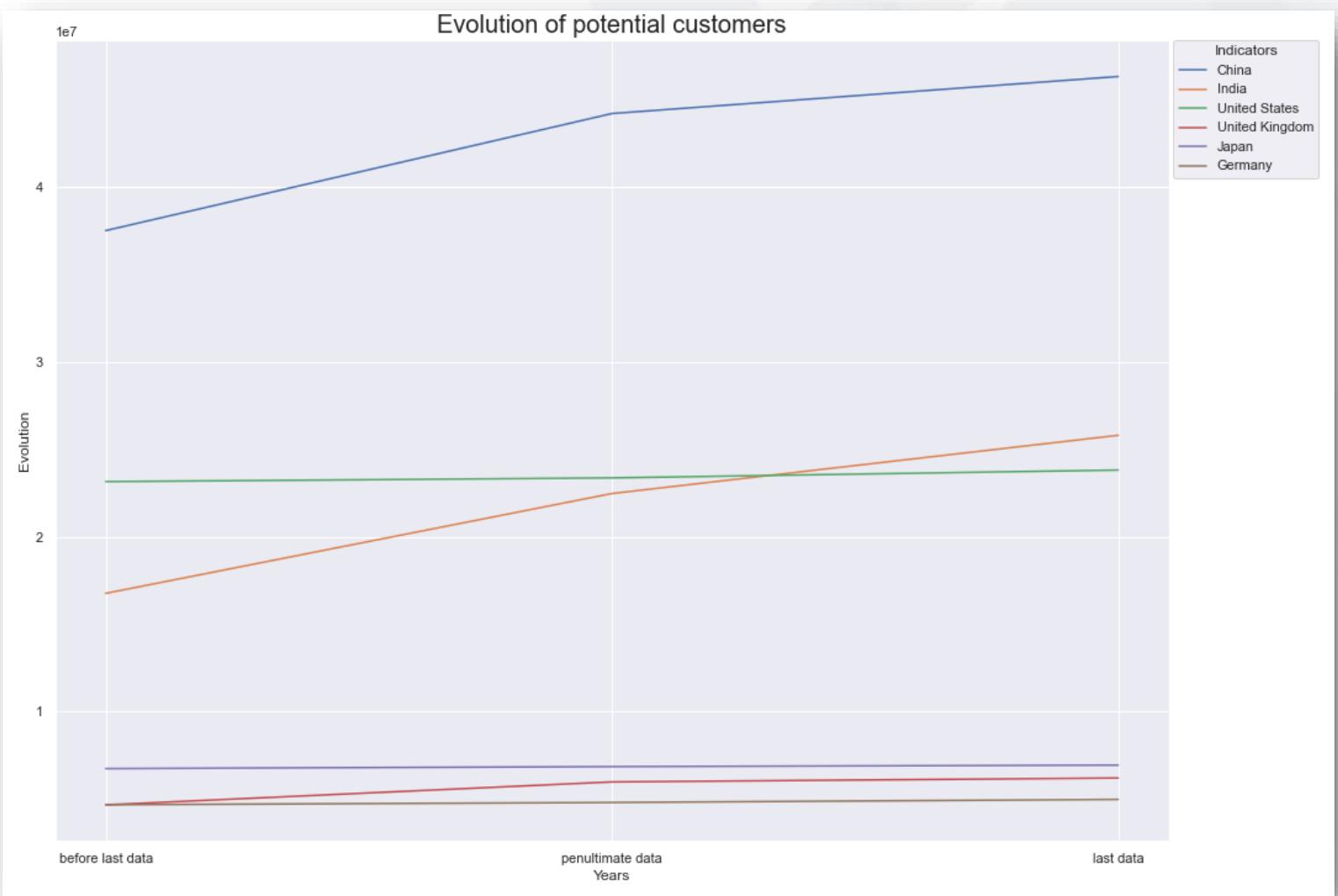


# La chine, l'inde et les US ont plus d'évolution de clients potentiels

*Pour chacun de ces pays, quelle sera l'évolution de ce potentiel de clients ?*



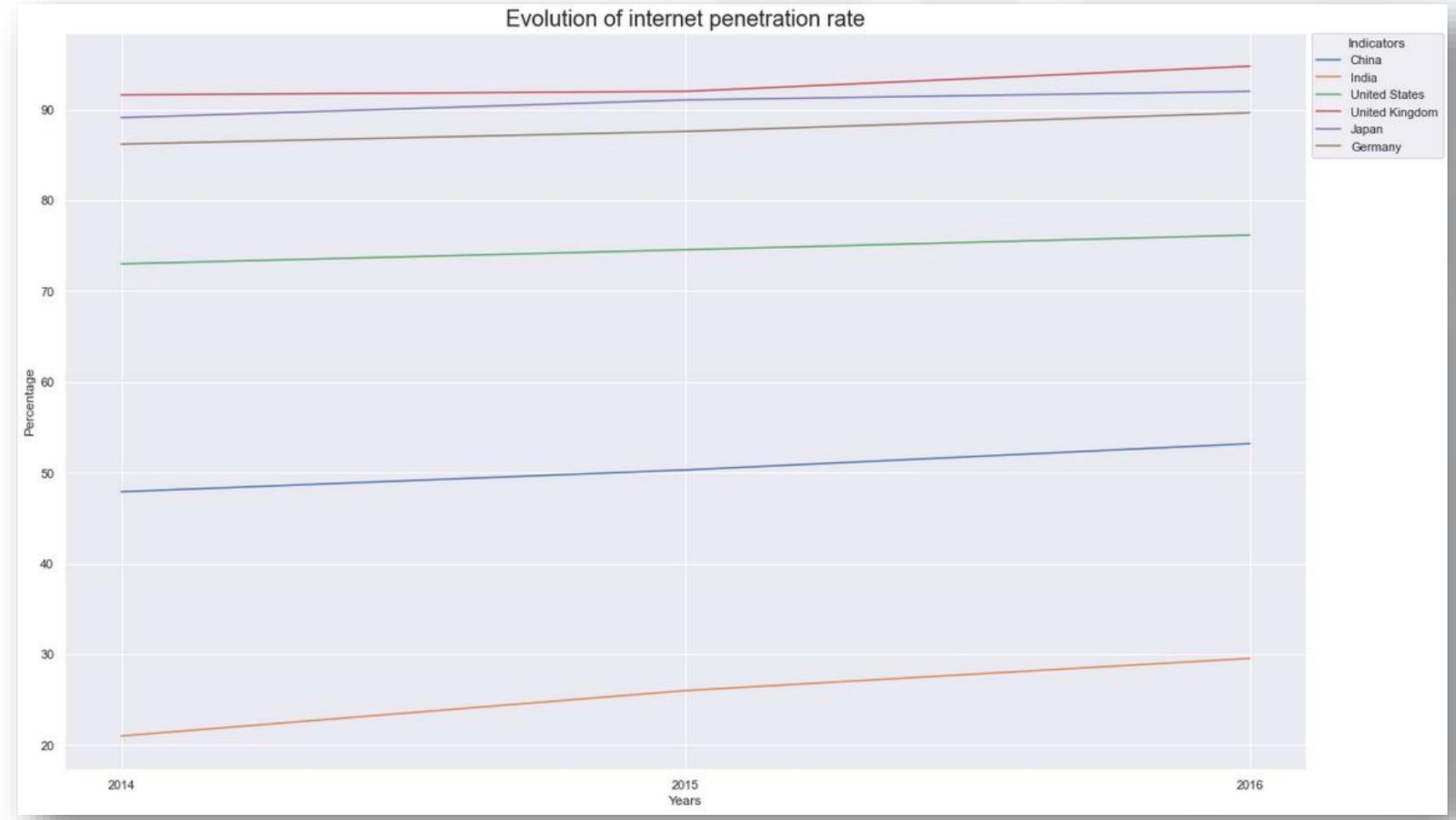
Sélection des trois dernières données disponibles pour chaque pays par les indicateurs “Enrolment in education” et “Internet users”



# *Le Royaume-Uni, le Japon, l'Inde et l'Allemagne ont la meilleure évolution d'internet*



Sélection des trois dernières données disponibles pour chaque pays par l'indicateur "Internet users"



# *L'Academy doit opérer en priorité en Chine*



**La Chine a la plus forte croissance de clients potentiels au cours des trois dernières années**



**Le prix de la formation n'est pas cher donc elle peut être acquise par plus de clients**



## 4.

### Conclusions sur la pertinence du jeu de données



# La pertinence du jeu de données



Certains datasets n'ajoutent pas de valeur

- EdStatsCountry-Series.csv
- EdStatsFootNote.csv
- EdStatsSeries.csv



Il manque des informations sur Academy pour rendre l'analyse plus précise

- La langue des cours
- Le prix de la formation



Le jeu de données permet de répondre aux attentes d'Academy

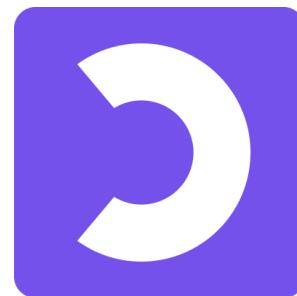
# Avez-vous des questions ?



**MERCI**

Soutenance de Projet  
Samir HINOJOSA

24 avril 2021



**OPENCLASSROOMS**