

Projet 2 :

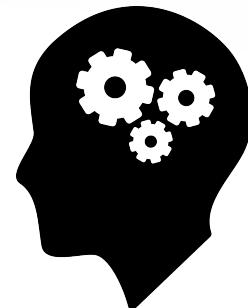
# Analysez des données de systèmes éducatifs.

Lecerf Defer Amandine



# Compétences évaluées

- Maîtriser les opérations fondamentales du langage Python pour la Data Science
- Mettre en place un environnement Python
- Manipuler des données avec des librairies Python spécialisées
- Effectuer une représentation graphique à l'aide d'une librairie Python adaptée
- Utiliser un notebook Jupyter pour faciliter la rédaction du code et la collaboration



# Plan

- I. Problématique
- II. Présentation du jeu de données
- III. Pré Analyse des données
- IV. Analyse
  - 1. Quels pays répondent à la problématique ?
  - 2. Confortation du choix de pays
- V. Conclusions





# Problématique



- Academy = start-up de la EdTech
  - Formation de niveau lycée et université en ligne
  - Projet de l'entreprise : Expansion à l'International
- **Mission :**  
Analyse exploratoire des données sur l'éducation de la banque mondiale
- **Objectif :**
    - Déterminer si ce jeu de données peut informer les décisions d'ouverture vers de nouveaux pays
    - Proposer de potentiels pays



GROUPE DE LA BANQUE MONDIALE

# Présentation du jeu de données



EdStatsCountry	<b>Informations globales sur l'économie de chaque pays</b> Taille : 241 lignes (1 par pays / zone) , 32 colonnes Quelques valeurs manquantes, Aucun doublon
EdStatsCountry-Series	<b>Sources des données contenues dans le fichier précédent</b> Taille : 613 lignes, 4 colonnes Pas de valeur manquante , Aucun doublon
EdStatsData	<b>Evolution de nombreux indicateurs pour chaque pays à partir de 1970</b> Taille : 886 930 lignes, 70 colonnes Nombreuses valeurs manquantes , Aucun doublon
EdStatsFootNote	<b>Informations sur l'année d'origine des données</b> Taille : 643 638 lignes, 4 colonnes Pas de valeur manquante, Aucun doublon
EdStatsSeries	<b>Informations sur les indicateurs socio-économiques étudiés</b> Taille : 3665 lignes, 21 colonnes (dont 6 vides) Beaucoup de données manquantes (surtout dans 10 colonnes), Aucun doublon

# Description plus détaillée

- **EdStatsData :**

- Année : 1970 à 2050
- Pays : 228
- Nombre indicateurs : **3665 indicateurs**
- Economie (Agriculture, Commerce, Marchés boursiers, aides internationales, PIB ...), Santé, Énergies, Science, Education , Démographie (inégalité population, marché du travail, ...)
- Données utilisées : noms des pays, noms des indicateurs, valeurs pour chaque année

- **EdStatsCountry :**

- Zone géographique : 23 dont certaines divisées selon le développement des pays
- Données utilisées : l'association pays-zone\_géographique

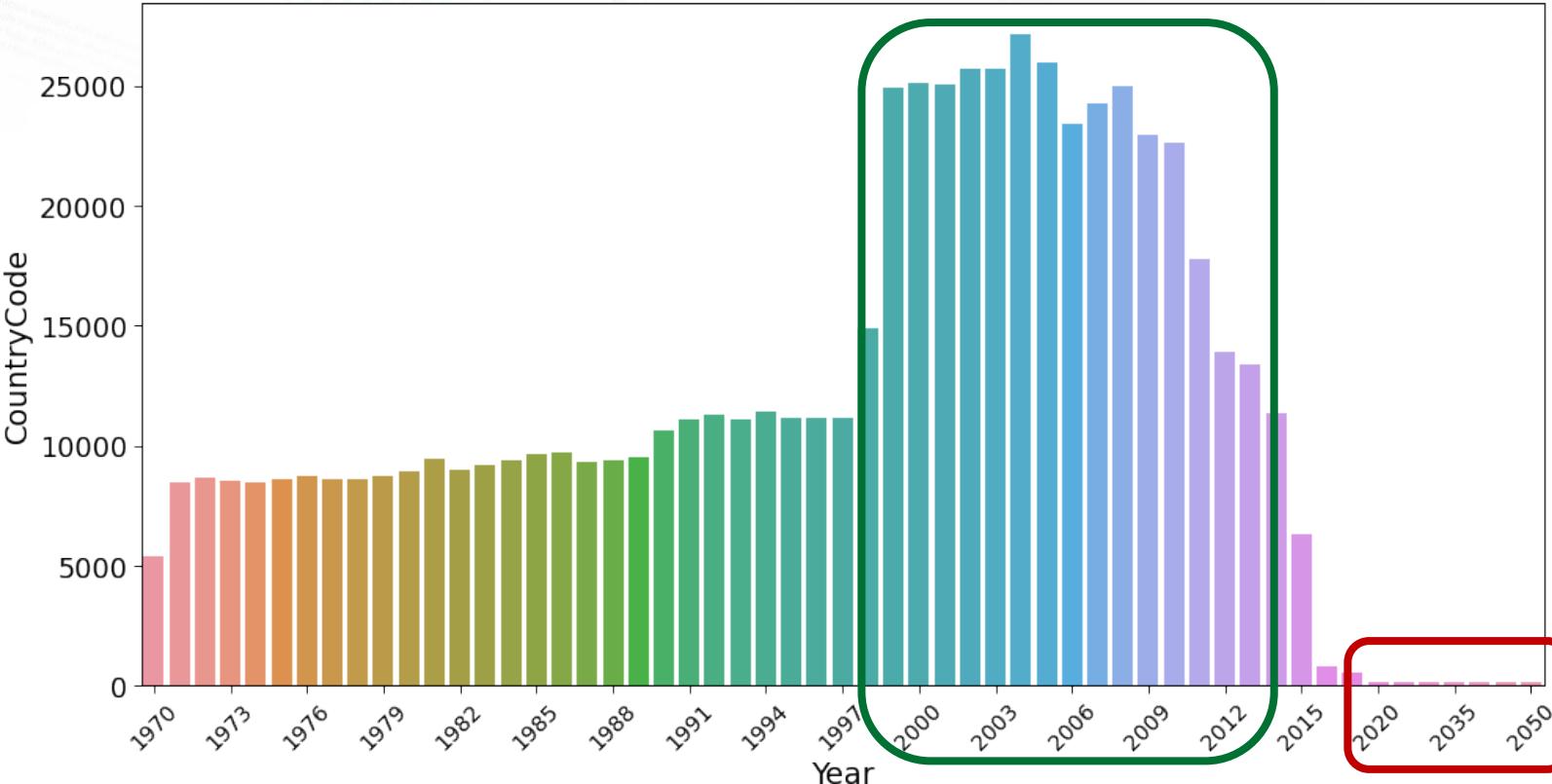


# Pré Analyse des données

# Choix de la période d'étude



Nombre de données par années présentes dans le dataset



Quantité de  
données pour  
chaque année :

**1998 à 2013**

**moyenne**

# Choix des pays

Nombre de données par pays pour la période allant de 1998 à 2013

- **64 premiers pays => moins de 6000 données :**
  - Les petits pays
  - Les nouveaux pays
  - Les territoires rattachés à un pays
- **Suppression de pays :**
  - Développement difficile de notre activité au vu de la situation du pays
  - Territoires de certains pays (st martin, french polynésia, ...)
  - Catégories qui représentent des regroupements de pays

=> 124 / 228 pays



# Choix des indicateurs

- De quels indicateurs avons-nous besoin pour cette étude ?

Démographique

Utilisation internet

Éducation

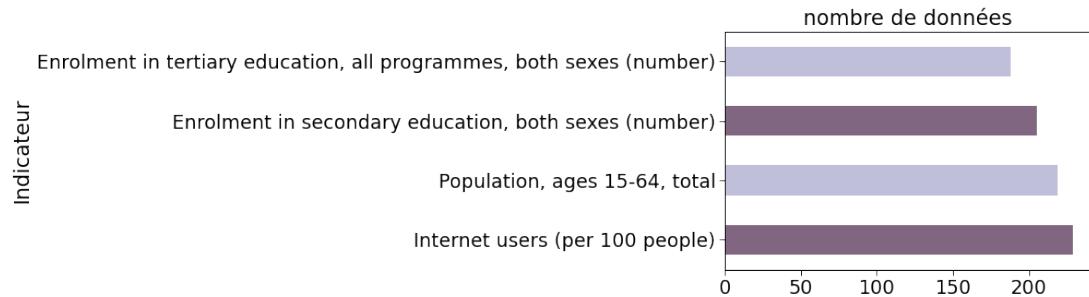
# Choix des indicateurs



4 indicateurs choisis = 200 données  
chacun = bons indicateurs

## Nombre de données pour chaque indicateur

Indicator Name	Indicator Code	Study_years
Internet users (per 100 people)	IT.NET.USER.P2	229
Population, ages 15-64, total	SP.POP.1564.TO	219
Enrolment in secondary education, both sexes (...)	SE.SEC.ENRL	205
Enrolment in tertiary education, all programme...	SE.TER.ENRL	188



# Choix judicieux ?



3 4



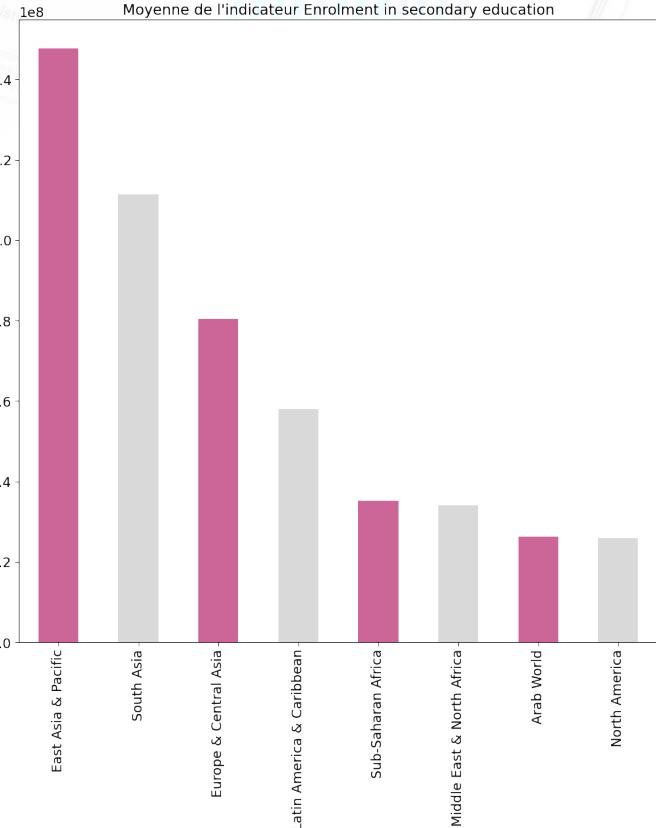
# Analyse des données



## Dans quels pays et régions s'implanter ?

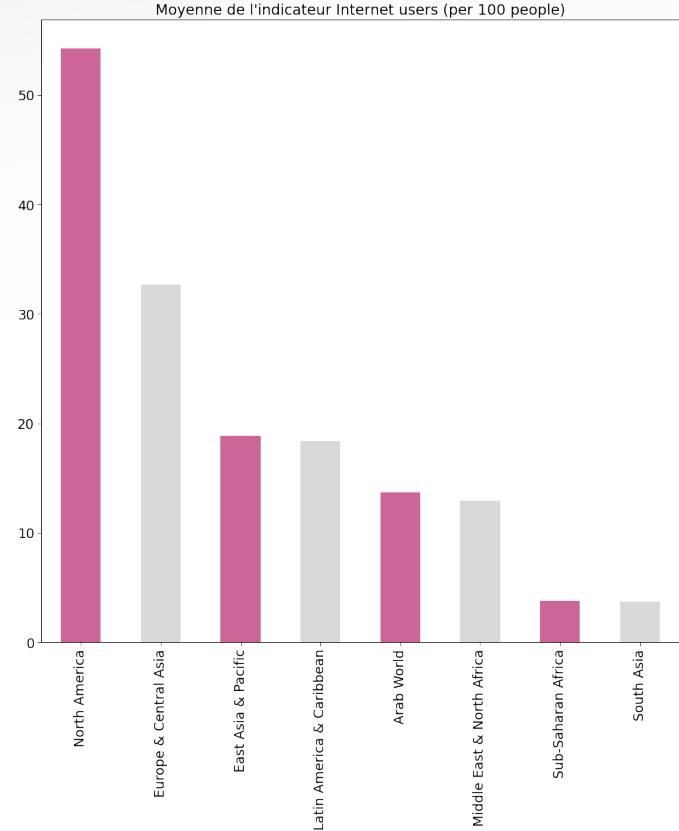
Comparaison des moyennes des indicateurs choisis  
sur la période définie (1998 – 2013)

# Exemples d'ordre de grandeur (moyenne)

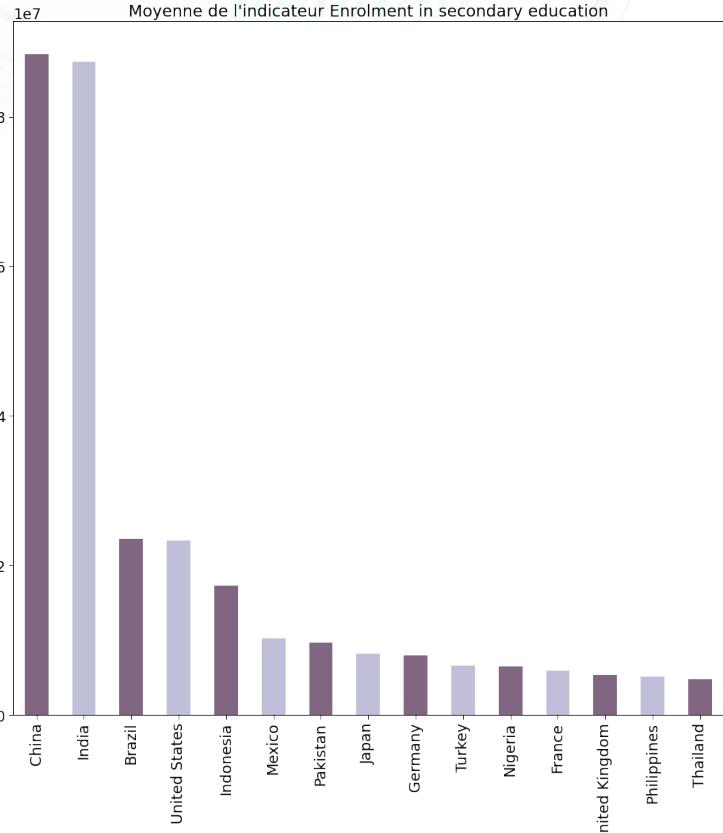


Quelles régions ?

- Asie de l'Est et Pacifique
- Asie du Sud
- Europe & Asie centrale
- Amérique du Nord
- Amérique latine & Caraïbes

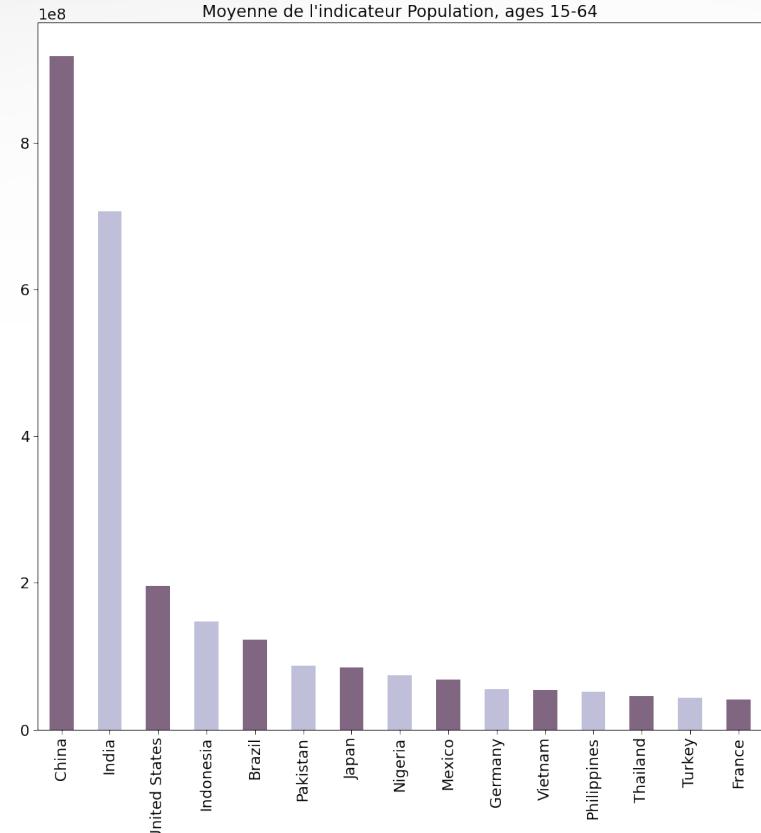


# Exemples d'ordre de grandeur (moyenne)



Quels pays ?

- Chine
- Inde
- Etats-Unis
- Brésil
- Indonésie

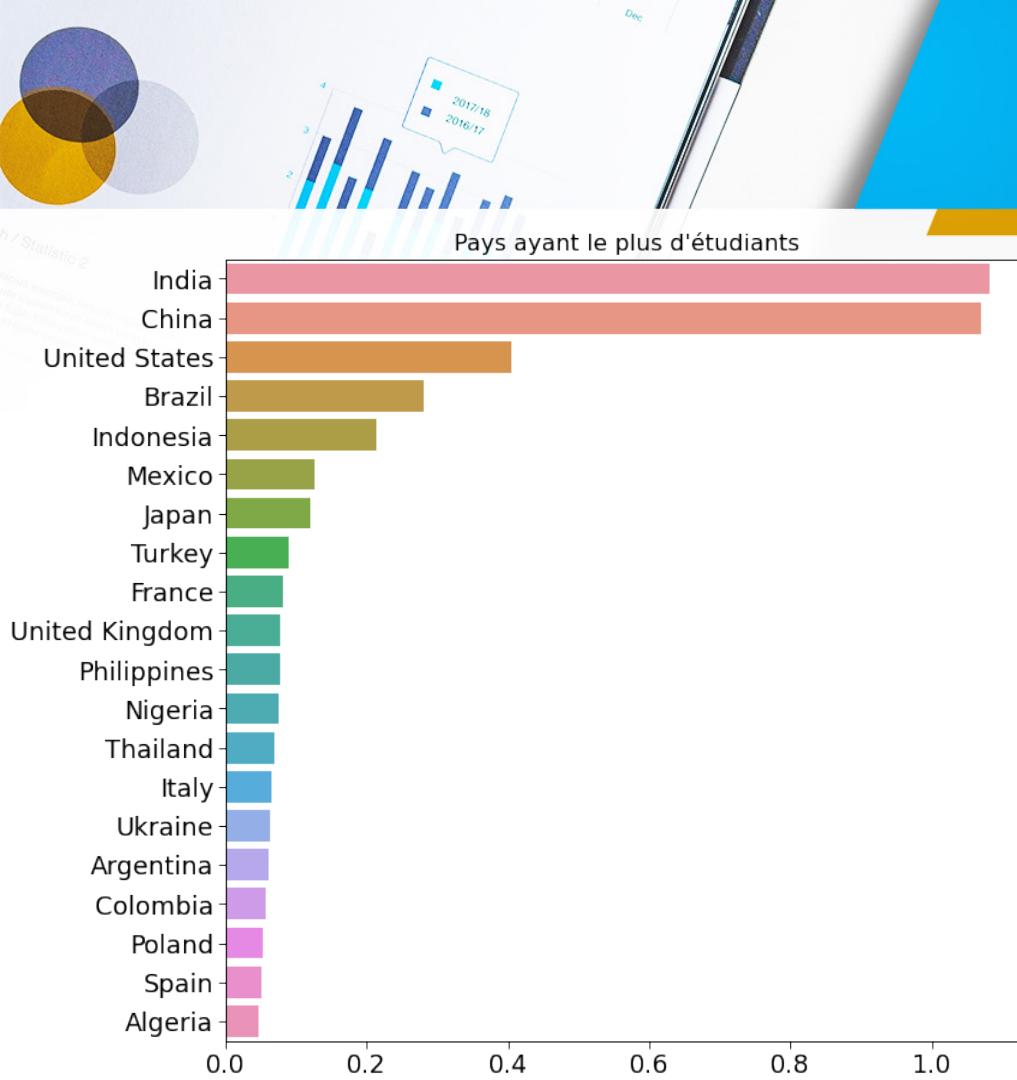




## Dans quels pays et régions s'implanter ?

Pays avec un fort potentiel de clients pour nos services

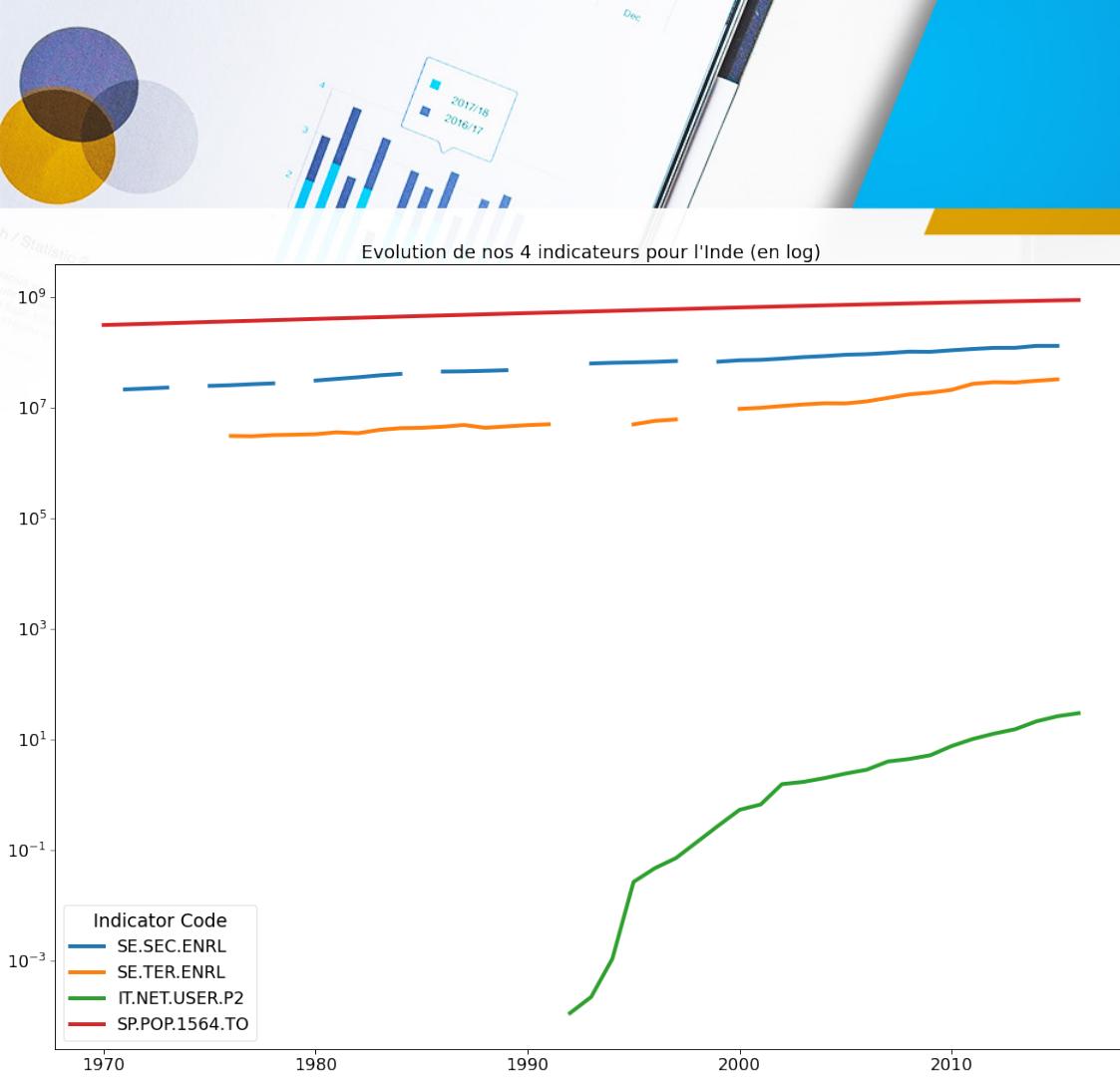
# Clients potentiels actuellement



Classement des pays ayant le plus d'étudiants (secondaire + tertiaire) :

- E – learning en particulier aux personnes en cours de scolarité
- Les pays :
  - Inde
  - Chine
  - Etats-Unis
  - Brésil
  - Indonésie.

# Exemple d'évolution de nos indicateurs



Evolution des 4 indicateurs retenus pour l'Inde :

- Pop 15 - 64 ans +
- Secondaire ++
- Tertiaire +
- Internet user ++



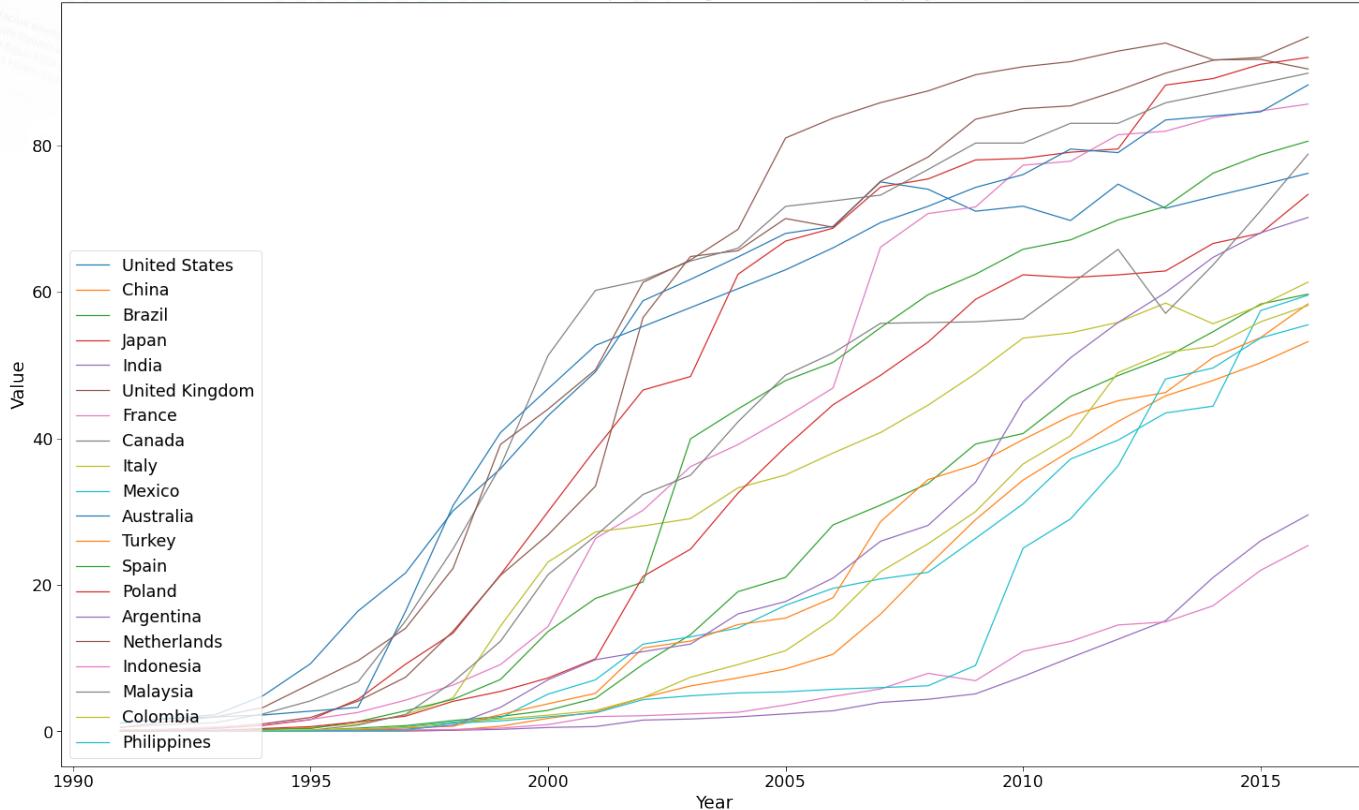
## Dans quels pays et régions s'implanter ?

Exemple de potentiel : Evolution du pourcentage d'utilisateur d'internet par pays

# Evolution d'utilisateurs internet



Evolution du pourcentage d'internautes par pays



Evolution du nombre d'utilisateurs d'internet par pays :

- Pas de projection après 2016
- Nombres d'étudiants ++  
= internet ++



# Conclusions



## Conclusion de l'étude:



### Région :

Asie de l'Est et Pacifique, Asie du Sud, Europe et Asie centrale,  
Amérique du Nord et Amerique Latine & Caraïbes

### Pays :

Chine, Inde, Etats-Unis et Brésil





## Conclusion sur le Dataset :

Le jeu de données permet-il de répondre au souhait d'extension ? :

### Pertinence du jeu de données :

- Tous les pays
- Nombreuses données relatives à l'éducation et l'accès à internet
- Toutes les données ont leur source = fiabilité

### Limits :

- Présence d'indicateurs peu ou pas exploitables = données manquantes
- D'autres indicateurs auraient été plus utiles = utilisation des e-learning, élèves se formant hors établissements scolaires, ...

# Fin de la présentation



Merci pour  
votre attention

