Projet N°2: Analyse des données de systèmes éducatifs

Problématique:

- academy, une start-up de la EdTech a un projet d'expansion à l'international.

academy

- Son activité : Elle propose des contenus de formation en ligne pour un public de niveau lycée et université.

- Notre objectif : Déterminer si les données de la banque mondiale permettent d'informer ce projet

BANQUE MONDIALE

Analyse pré-exploratoire

I) Présentation et description du jeu de données(types de variables, nombre de lignes et de colonnes)

II) Validation de la qualité du jeu de données (quantités de valeurs manquantes et de doublons)

III) Sélection des informations pertinentes pour la résolution de la problématique (choix des variables utiles).

IV) Analyse comparative des pays à travers les variables choisies

V) Conclusions

Présentation du jeu de données de la banque mondiale

5 tableaux csv situés dans la rubrique EdStats (Statistiques de l'éducation) :

EdStatsCountry.csv	▪ **Données économiques par pays ou par zone géographique (Niveau de revenu, monnaie, les types de prêts accordés aux pays etc.)** ▪ **Information sur les systèmes de prêts accordés aux pays** ▪ **Les systèmes de comptabilité, de collecte de données etc.** Taille : 241 lignes (pays ou zone géographique ou groupe de pays), 32 colonnes Aucun doublons mais il y a des valeurs manquantes
EdStatsCountry-Series.csv	▪ **Informations sur les sources de données de plusieurs indicateurs présentés par des codes. Exemple : SP.POP.TOTL (Population, total)** ▪ Sources officiellement reconnues (ONU, internationales , nationales) ➡ Données fiables

EdStatsData.csv	▪ Evolution par pays et zones géographiques depuis 1970 de plusieurs indicateurs sur l'éducation. Il s'agit principalement des proportions de personnes qu'on a aux différents niveau de scolarité. Taille : 886 930 lignes (étude de ces indicateurs pour chaque pays ou zone géographique) , 70 colonnes Aucun doublons mais il y a beaucoup de valeurs manquantes
EdStatsFootNote.csv	▪ Présentation de certaines données par : - leurs codes . Exemple : SE.PRE.ENRL.FE (Inscriptions à l'école, préscolaire c'est-à-dire maternelle, filles (% brut)) - leurs sources (nationale, UNESCO etc.) - leur année - leur incertitudes Taille : 643638 lignes et 5 colonnes. Il n'y a pas de doublons ni de valeurs manquantes hormis la colonne, Unnamed:4, qui contient uniquement des valeurs nulles.
EdStatsSeries.csv	Classement par thèmes de plusieurs indicateurs concernant les acquis scolaires, les niveaux scolaires, les inégalités d'accès à l'éducation entre les sexes etc. Taille : 3665 lignes et 21 colonnes Pas de doublons mais il y a beaucoup de valeurs nulles avec 6 colonnes vides.

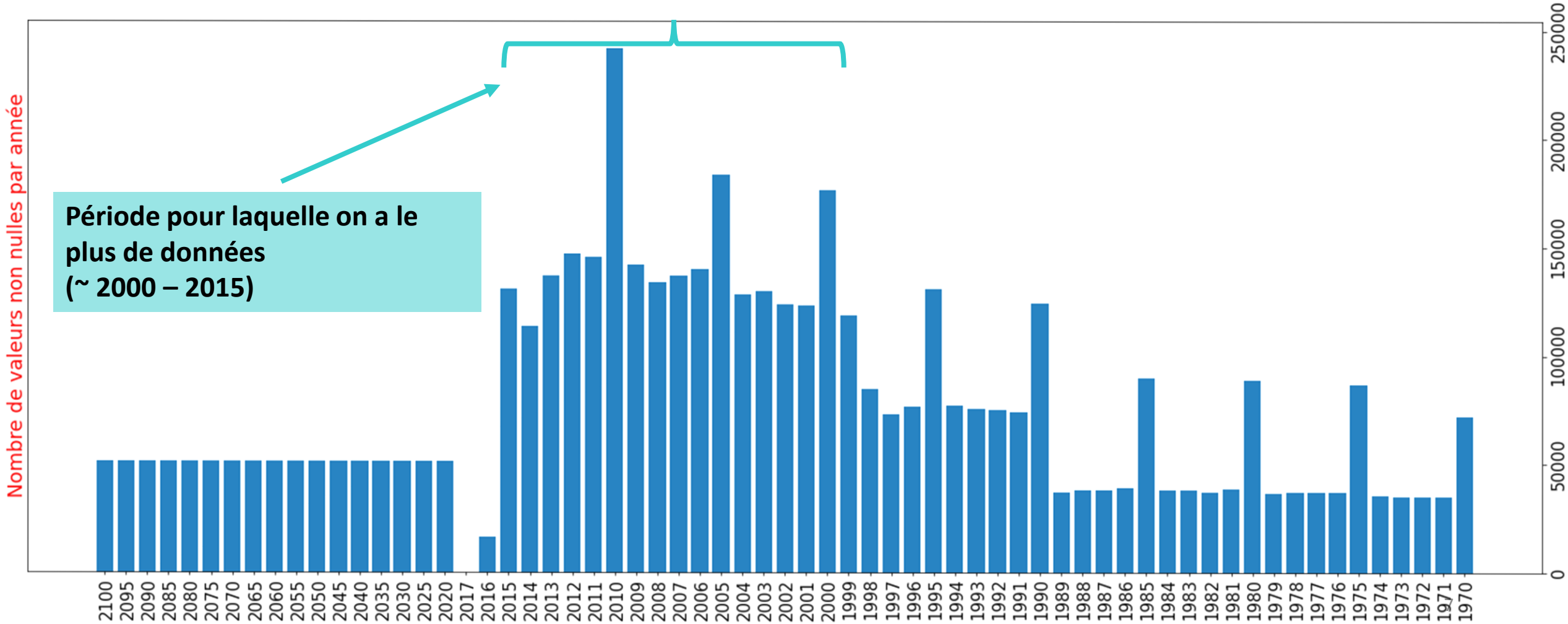
Notre problématique:

- Déterminer les pays à fort potentiel de clients pour les services que nous proposons (Cours en lignes de niveau lycée et supérieur).
- Données utiles : nombre d'étudiants au lycée et au niveau supérieur avec l'accessibilité au réseau internet.
- ➡ Choix du 3eme tableau (EdStatsData.csv) (seul tableaux à fournir ces données).
- ce tableau contient 3665 indicateurs qu'il faudra trier.

Validité des données du tableau EdStatsData.csv

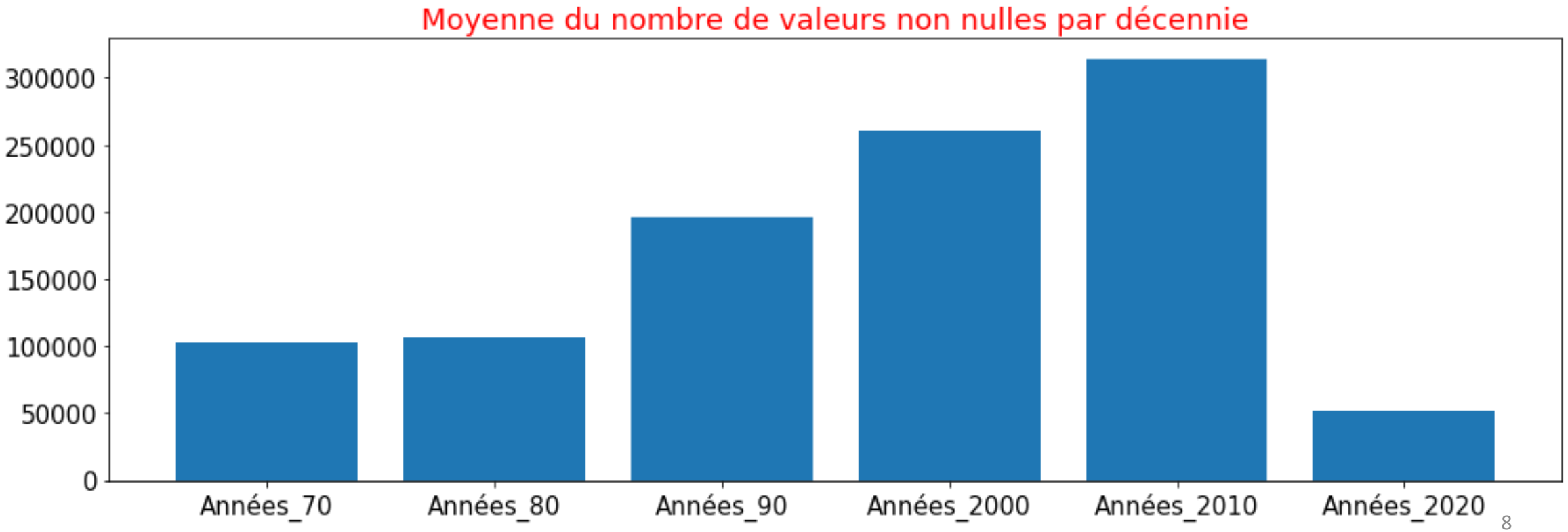
- Nombre de données non nulles :
  - 1ère étape : construction d'un histogramme représentant le nombre de valeurs non nulles par année.
  - 2ère étape : construction d'un histogramme représentant le nombre de valeurs non nulles par pays.

Nombre de valeurs non nulles par année



Discrétisation = agréger les valeurs d'une variable quantitative continue en classes

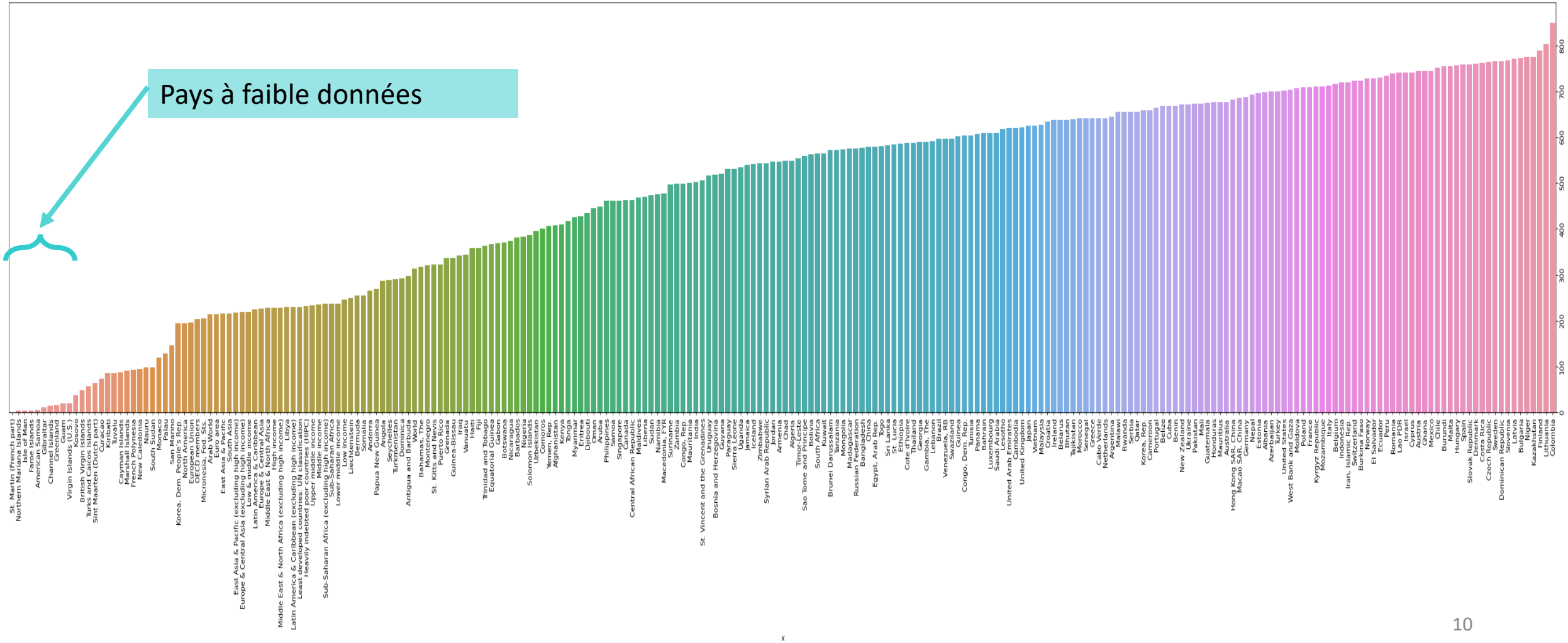
<u>Application</u> : On raisonne par décennie et on va passer d'une valeur par année à la moyenne des valeurs des années d'une même décennie



Limitation de l'analyse des données à La décennie 2010 (2010→ 2017)

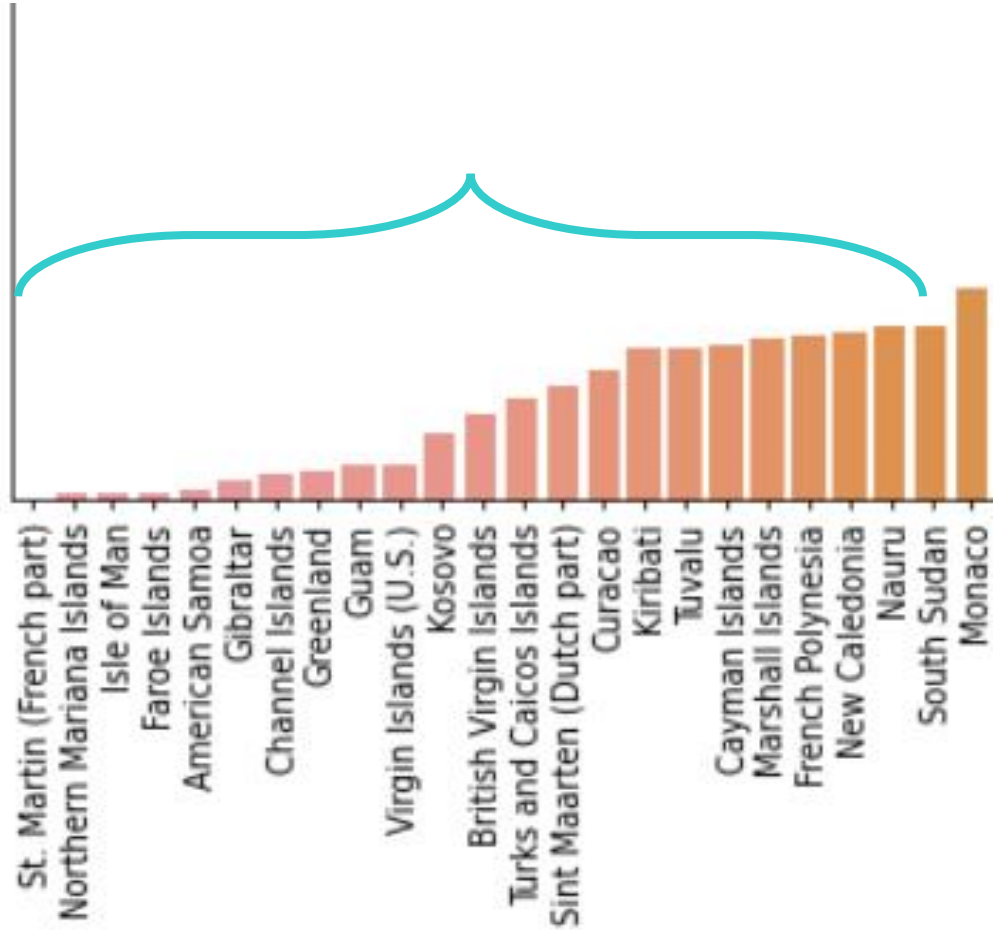
- La décennie 2010 (2010→ 2017) est celle qui contient le plus de valeurs non nulles
- Limitation de l'étude à cette période pour déterminer les pays à fort potentiel de clients.
- Suppression des autres années:
  - les années 70 étant trop anciennes pour apporter des informations pertinentes.
  - les autres décennies présentent moins de données.
- Néanmoins ces décennies vont nous intéresser pour représenter l'évolution dans le temps des indicateurs pertinents et pour pouvoir se projeter dans l'avenir.

Nombre de valeurs non nulles par pays pour la décennie 2010



Suppression des pays avec très peu de données

- Suppression de pays à faible données (- 15%) et à faible population (« petits pays », les iles etc.):
- Exemples: Virgin Islands (U.S.), St. Martin (French part), sint Maarten (Dutch part) etc.



Recherche d'indicateurs pertinents

- Exploration en détails des 3665 indicateurs en s'aidant d'une recherche par mots-clés (internet, secondary, tertiary …)

- Objectif : faire l'état du nombre de lycéens et d'étudiants par pays avec le pourcentage de personnes ayant accès à internet.

INDICATEURS CHOISIS:

- Population, total

- Enrolment in upper secondary education, both sexes (number)

- Internet users (per 100 people)

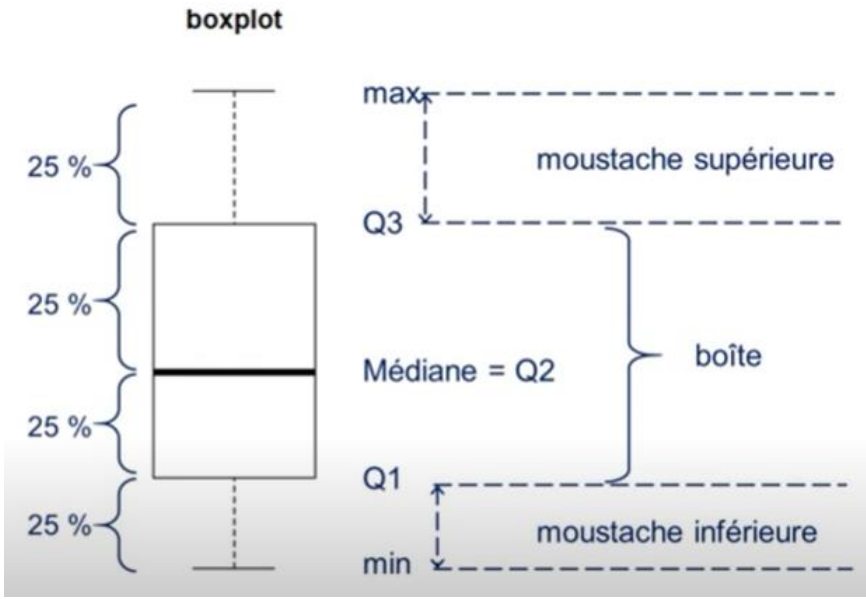
- Enrolment in post-secondary non-tertiary education, both sexes (number)

- Enrolment in tertiary education, all programmes, both sexes (number)

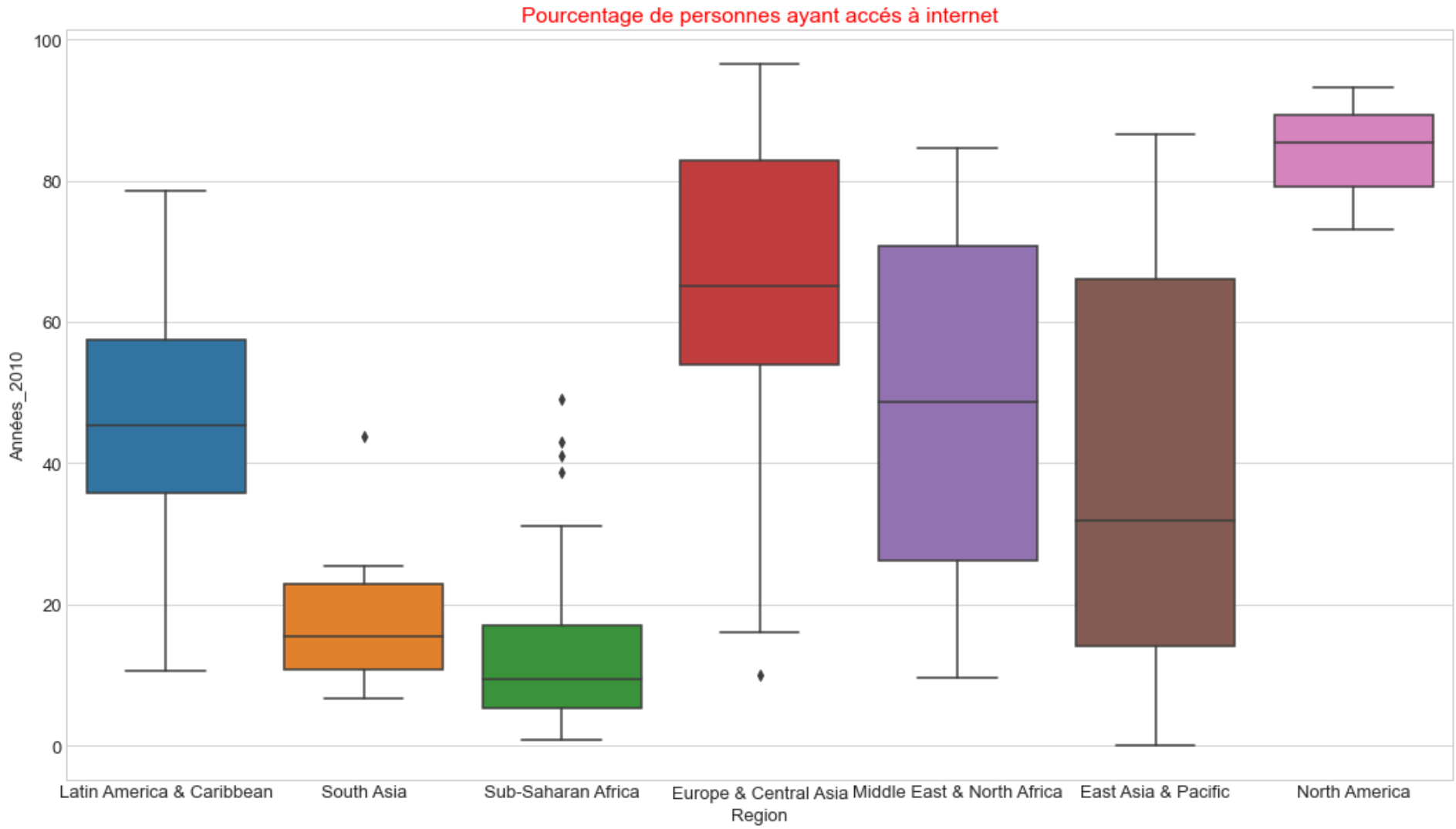
Ordres de grandeurs des indicateurs choisis pour les zones géographiques

- Etude comparative des zones géographiques spécifiées dans le jeu de données:
  - Latin America & Caribbean
  - South Asia
  - Sub-Saharan Africa
  - Europe & Central Asia
  - Middle East & North Africa
  - East Asia & Pacific
  - North America

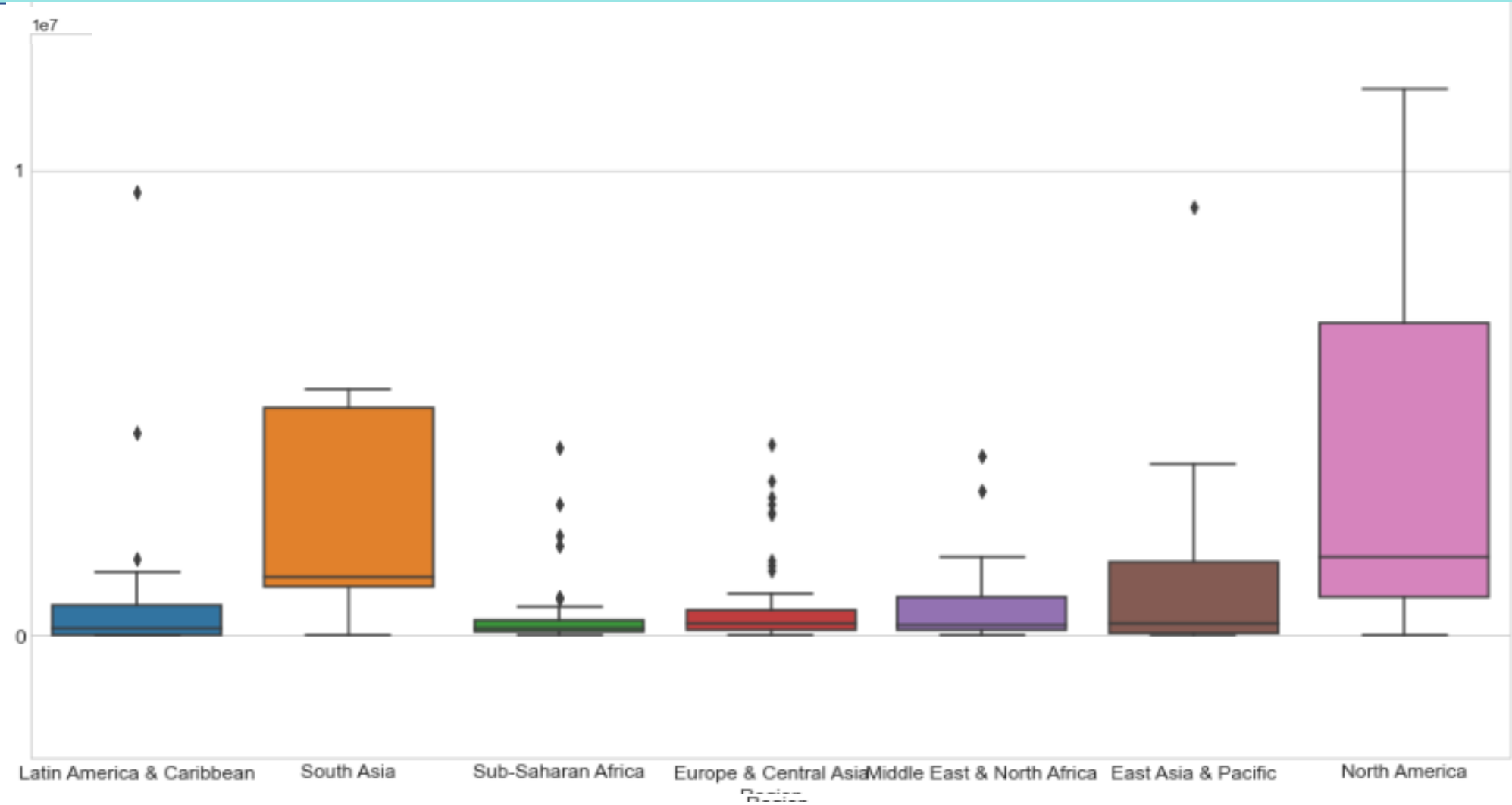
- la boîte à moustaches ou boxplot : Une représentation graphique intéressante pour représenter la dispersion de ces variables



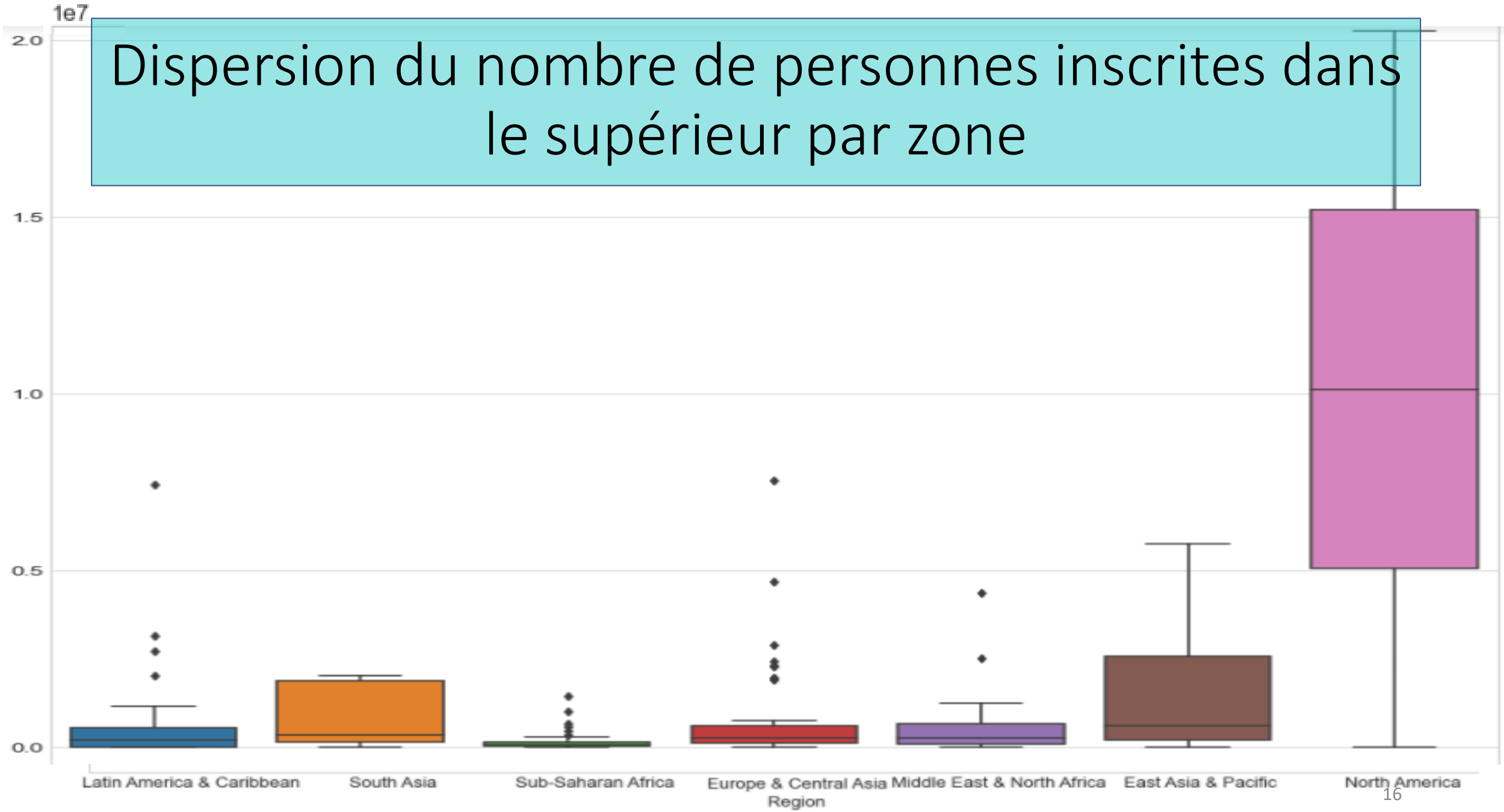
Dispersion du pourcentage de personnes ayant accès à internet par zone



Dispersion du nombre de personnes inscrites dans le secondaire par zone



Dispersion du nombre de personnes inscrites dans le supérieur par zone



Bilan de l’étude des dispersions des indicateurs

- Disparité plus ou moins marquée entre les pays d’une même zone pour les indicateurs.
- Nécessité de faire une étude comparative par pays

Nombre de clients potentiels

- ***Nombre de clients potentiels*** = [Nombre de lycéens + nombre d'étudiants] x $\frac{[\text{ \% personnes ayant accès à internet}]}{100}$

- **INDICATEURS CHOISIS:**

  - **Enrolment in upper secondary education, both sexes (number)**

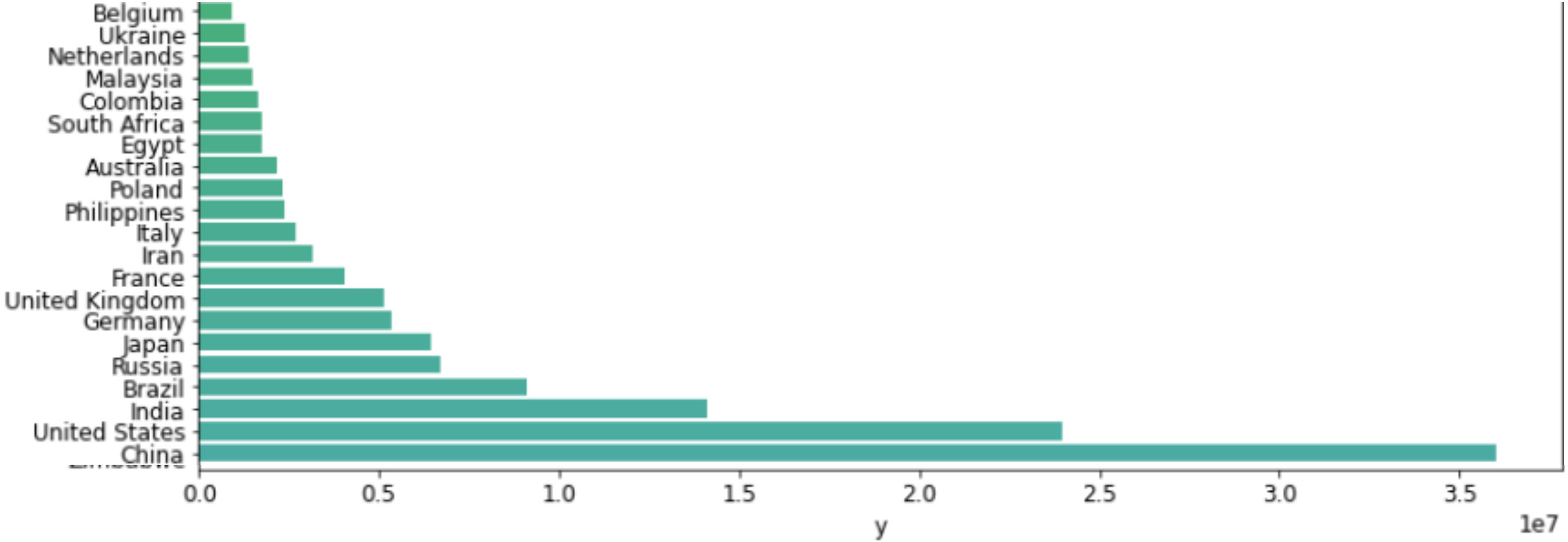
  - **Internet users (per 100 people)**

  - **Enrolment in post-secondary non-tertiary education, both sexes (number)**

    +

    - Enrolment in tertiary education, all programmes, both sexes (number)

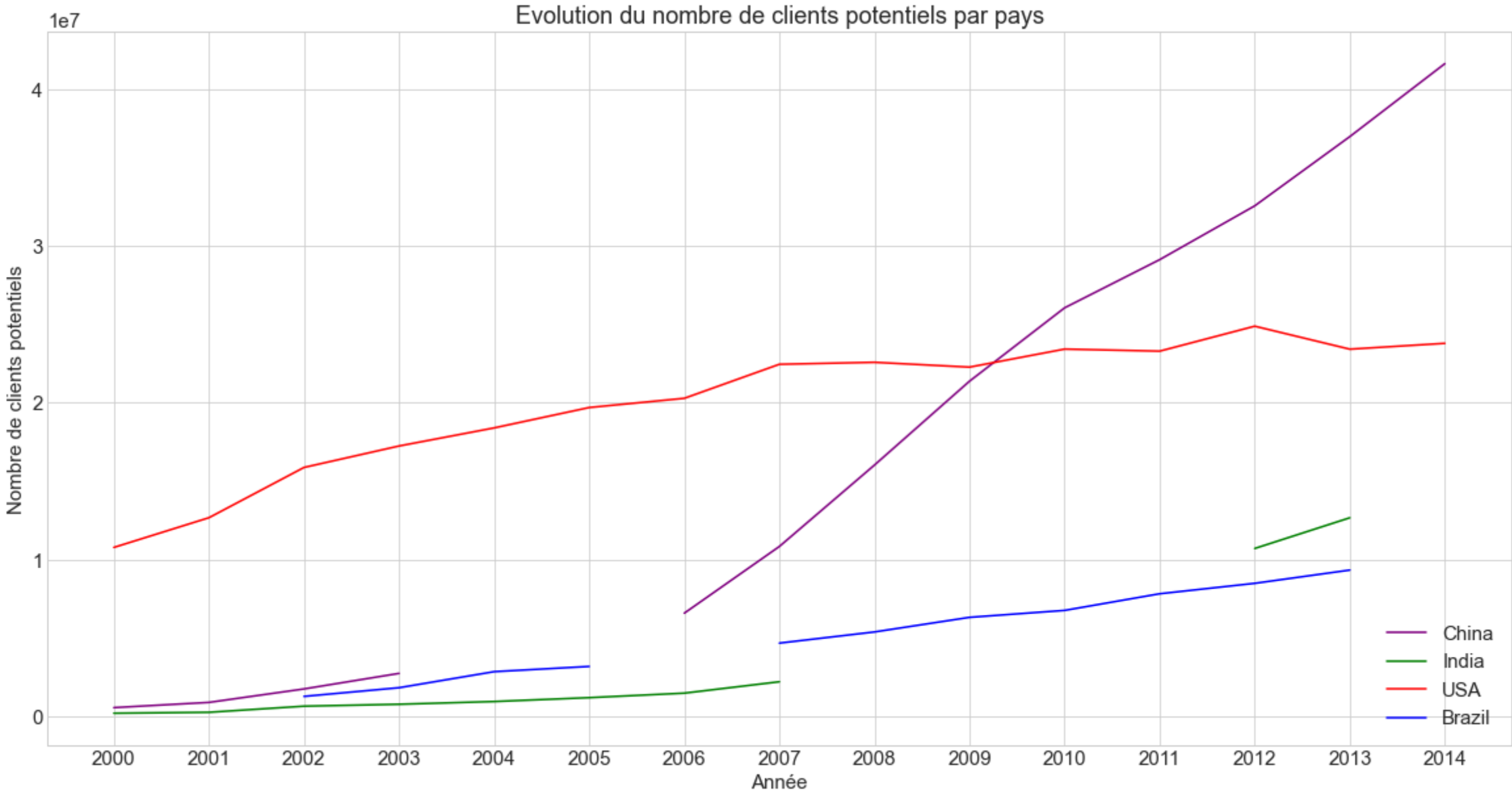
Classements suivant le nombre de clients potentiels (21 premiers pays)



Bilan

- Sur la décennie 2010, on a un classements des pays selon le potentiel de clients qu'il présentent. (China, United States, India, Brazil etc.).

- Il faut tout de même étudier l'évolution de ces pays par rapport aux indicateurs sur une grande période afin de voir la dynamique de chaque pays.


Evolution du nombre de clients potentiels par pays
1e7
Nombre de clients potentiels
Année
0
1
2
3
4
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
China
India
USA
Brazil

Conclusions

- Le jeu de données permet de répondre à notre problématique car les informations sur certains indicateurs permettent d'avoir un nombre de clients potentiels pour notre activité.

- Toutefois nous avons besoin d'autres informations pour conforter nos choix:
    - étude de la demande locale pour nos services : nombre d'élèves se formant en dehors de leur établissement, dépenses moyennes de ces élèves pour assurer leur réussite scolaire etc.
    - barrière de la langue, étude de la capacité de la start-up academy à traduire les supports en d'autres langues que l'anglais.
    - recrutement de formateurs avec les coûts que cela impliquera.
    - législation du pays en matière de formation et d'enseignement; procédure de délivrance des licences et autorisations obligatoires à l'exercice de l'activité.