ANALYSEZ DES DONNÉES DE SYSTÈMES ÉDUCATIFS

Projet 2 – OC/ Data Scientist

Audrey Terrien
08/12/2021

PLAN DE L'ÉTUDE

1. Contexte et questions stratégiques de la société
2. Préparation à l'exploitation du jeu de données
3. Analyse pré-exploratoire
4. Conclusion sur la pertinence du jeu de données final

1

CONTEXTE ET OBJECTIFS

CONTEXTE ET OBJECTIFS

QUI SOMMES-NOUS ?
Start-up dans le milieu de la EdTech

QUE VOULONS-NOUS ?
Vendre du contenu de formation en ligne

QUI EST-CE QUE NOUS VISONS ?
Jeunes gens entre 14-25 ans (niveau lycée et université)

CONTEXTE ET OBJECTIFS

OBJECTIF DE CETTE ÉTUDE

Analyser la viabilité de ce projet à travers une première analyse pré-exploratoire avec l'aide des données de la Banque Mondiale

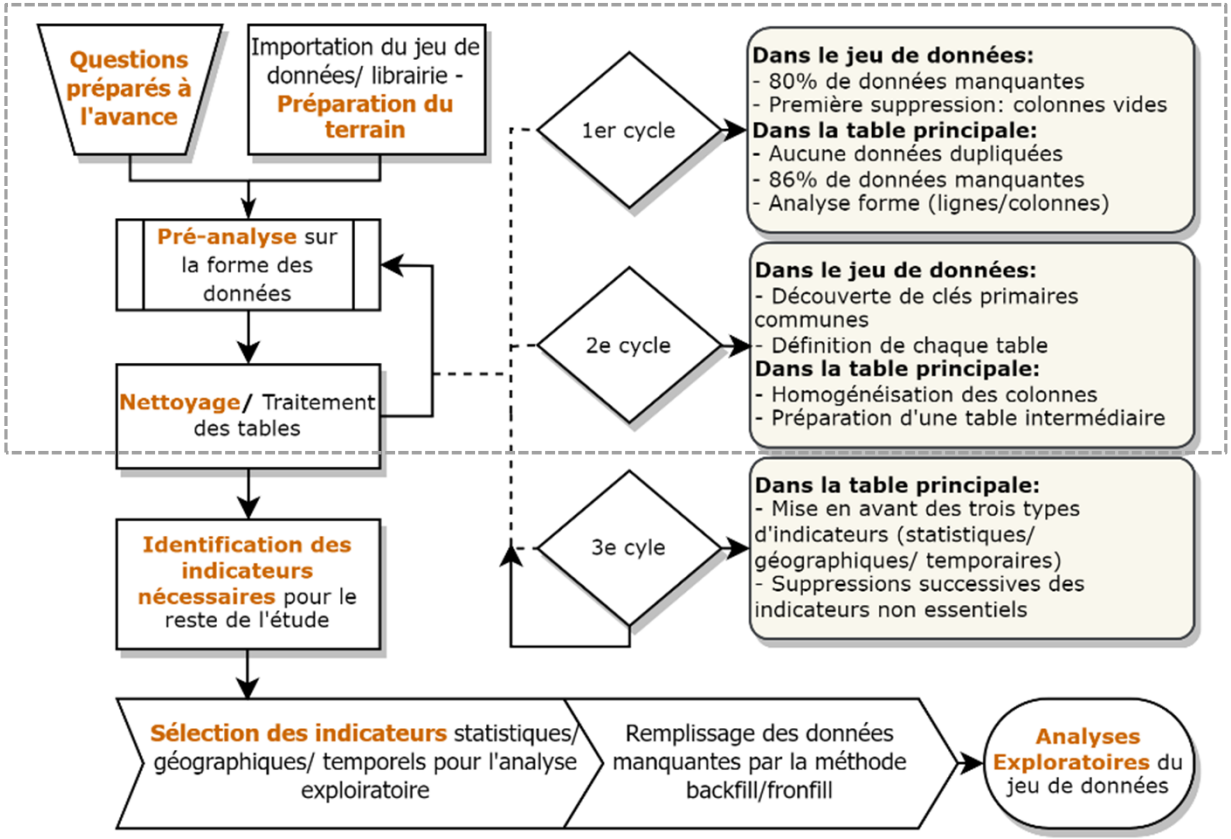
1. Quels sont les pays avec un fort potentiel de clients pour nos services ?
2. Pour chacun de ces pays, quelle sera l'évolution de ce potentiel de clients ?
3. Dans quels pays l'entreprise doit-elle opérer en priorité ?

BANQUE MONDIALE

2

PRÉPARATION À L'EXPLOITATION DU JEU DE DONNÉES

LE PROCESSUS POUR ARRIVER A L'ANALYSE EXPLORATOIRE

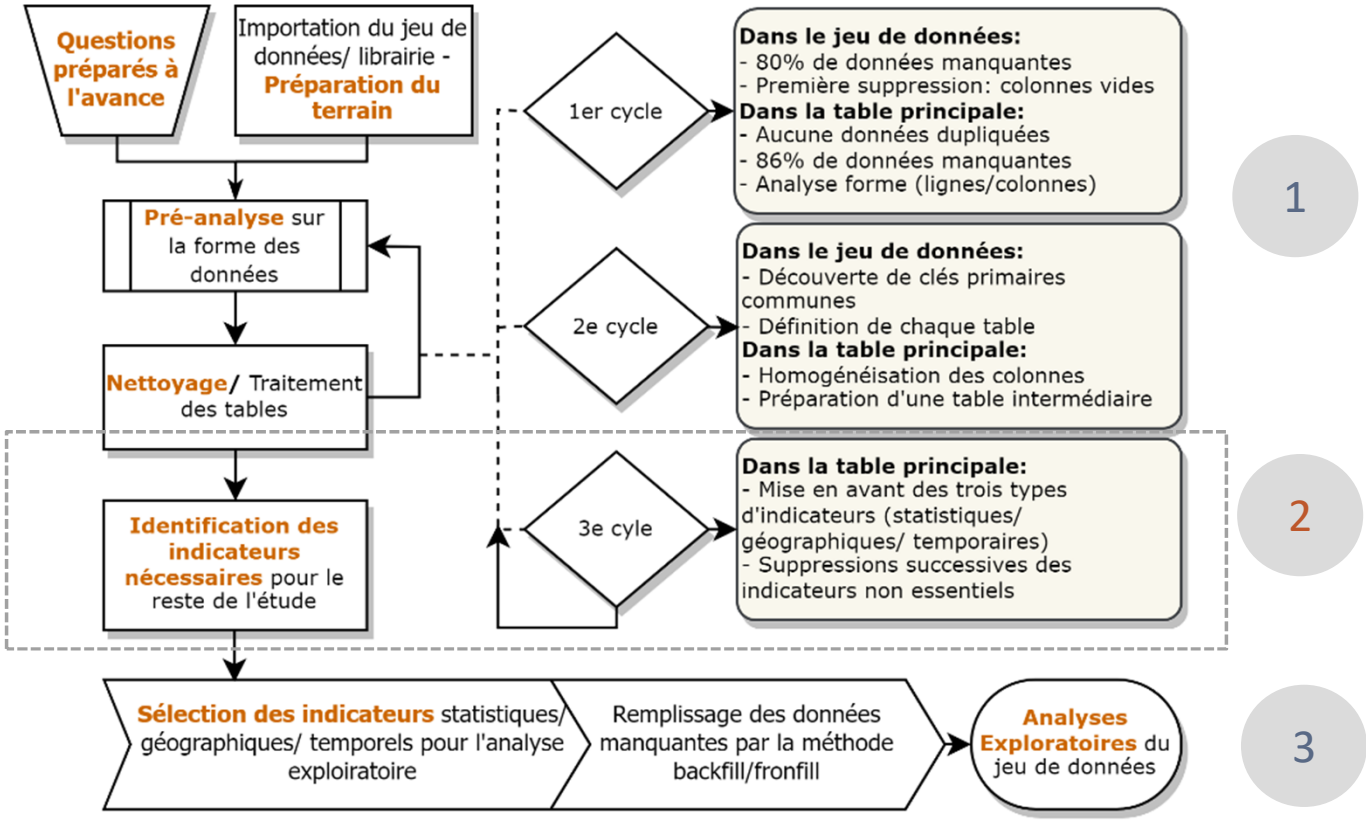


LES TABLES COMPOSANTS LE JEU DE DONNÉES

BANQUE MONDIALE

	Informations générales sur…
EdStatsCountry.csv	… l'économie de chaque zones géographiques et économiques dont les **pays** **32** colonnes et **241** lignes, aucun doublon**, qq valeur** manquante
EdStatsCountry-Series.csv	… la **provenance** des informations sur les zones géographiques utilisées dans **EdStatsCountrys.csv** **4** colonnes et **613** lignes, aucun doublon**, %** de Valeurs manquantes
EdStatsFootNote.csv	… **l'année** originale des données et les incertitudes sur les données **5** colonnes et **643,638** lignes, aucun doublon**,** aucune valeur manquante (sauf 1 colonne)
EdStatsSeries.csv	… les **indicateurs statistiques** utilisés du fichier **EdStatsData.csv** **21** colonnes et **3,665** lignes, aucun doublon**,** aucune valeur manquante (dont 1 colonne)
EdStatsData.csv	… les **indicateurs statistiques** en fonction **des zones géographiques** et les **années** étudiées (1970-2100) **70** colonnes et **886,930** lignes, aucun doublon**, plus 86%** de valeurs manquantes dont 6 colonnes quasi ou totalement vides

LE PROCESSUS POUR ARRIVER A L'ANALYSE EXPLORATOIRE

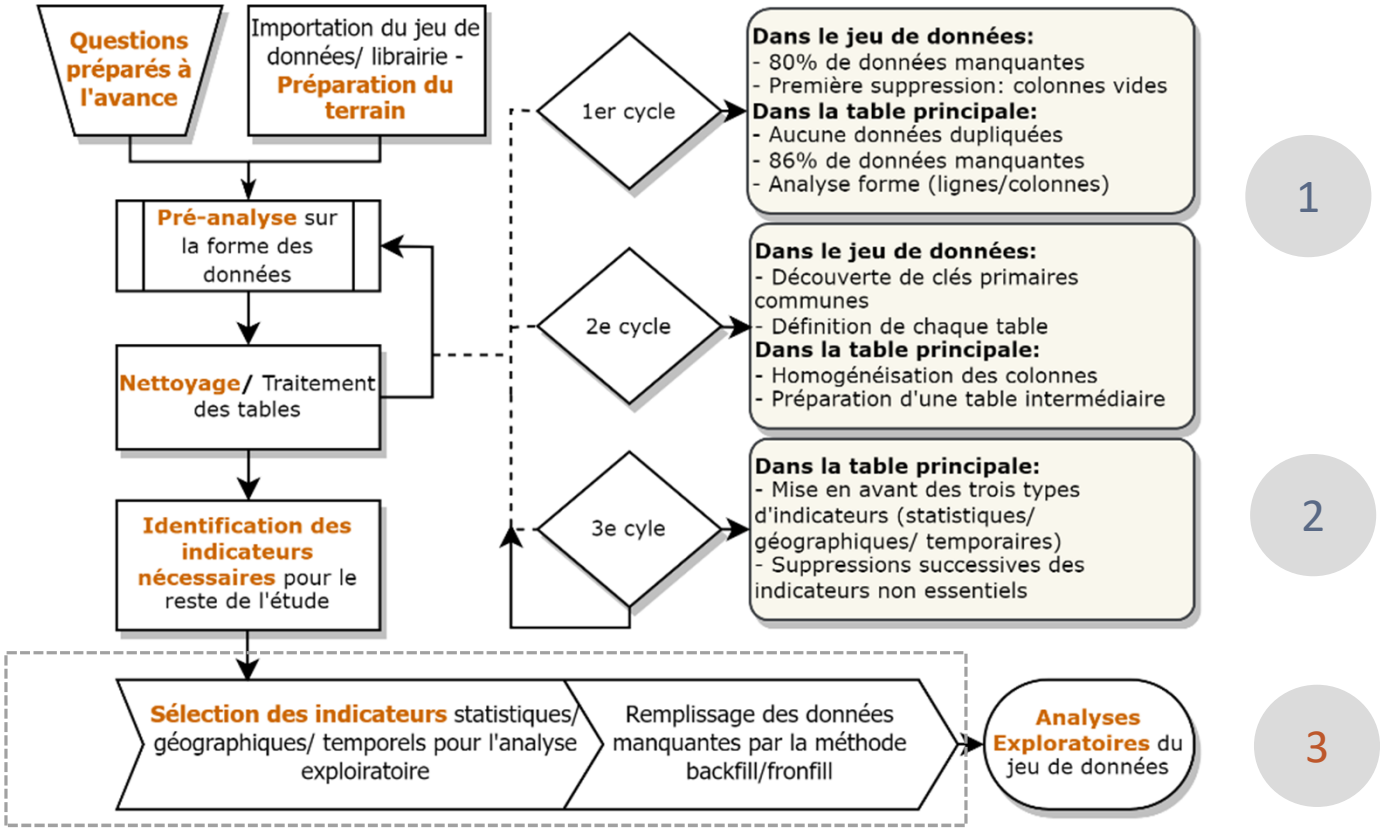


IDENTIFICATION DES INDICATEURS POTENTIELS

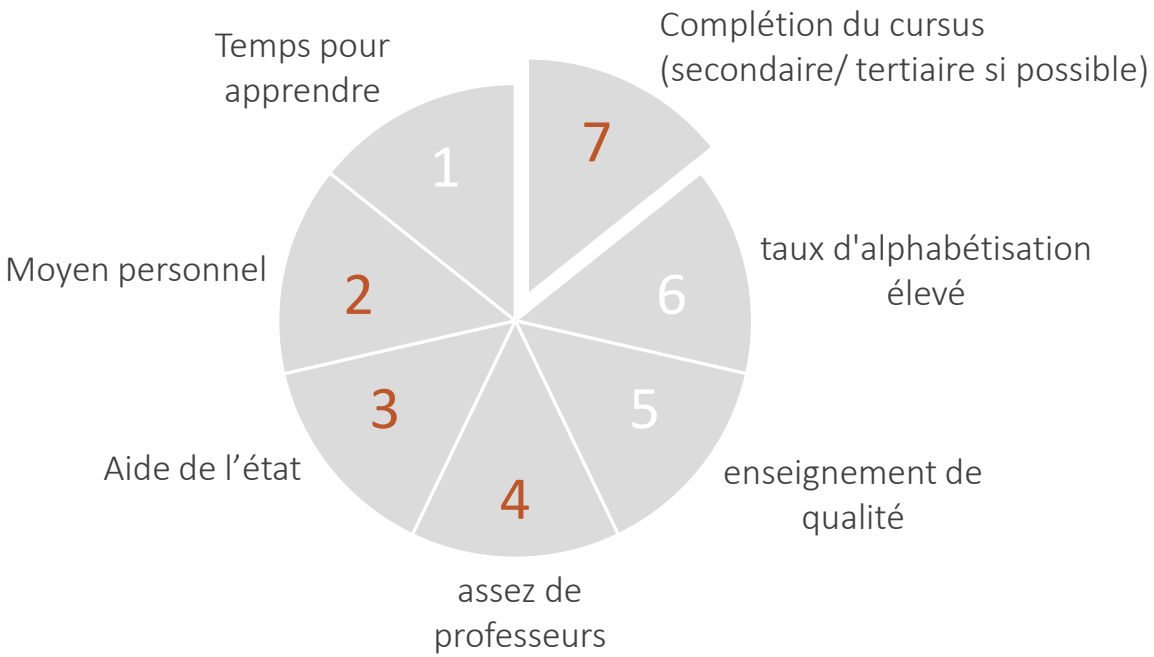
BANQUE MONDIALE

	Trois types d'informations à utiliser/confronter et qui sont les…
Indicateurs statistiques	3665 indicateurs statistiques sur l'éducation/le niveau social/l'accès à Internet/etc.
Zones géographiques (indicateurs géographiques)	241 zones géographiques/niveau social/îles – ne conserve que les pays, soit 113 pays
Années (indicateurs temporelles)	65 années entre 1970 et 2050 : - De 1970 à 2015: données récoltées - De 2015 à 2050: données prédites

LE PROCESSUS POUR ARRIVER A L'ANALYSE EXPLORATOIRE



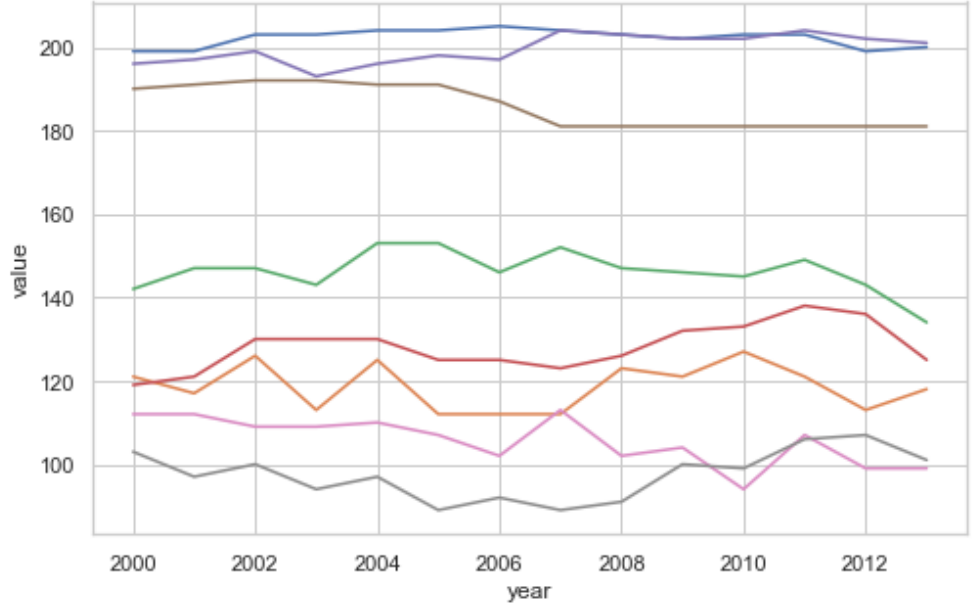
ORIENTATION CHOISIE POUR SELECTIONNER LES INDICATEURS STATISTIQUES



D'après un article de l'OECD iLibrary logo titré "Introduction: The indicators and their framework"[1]

SELECTION FINALE DE TOUS LES INDICATEURS

Nombre de valeurs existantes en fonction du temps et des indicateurs statistiques



8 INDICATEURS STATISTIQUES

- GDP per capita (current US$)
- Government expenditure on education as % of GDP (%)
- Gross enrolment ratio, secondary, both sexes (%)
- Gross enrolment ratio, tertiary, both sexes (%)
- Internet users (per 100 people)
- Population, ages 15-24, total
- Pupil-teacher ratio in secondary education (headcount basis)
- Pupil-teacher ratio in tertiary education (headcount basis)

165 INDICATEURS GÉOGRAPHIQUES

- PAYS : répartis en région et selon leur niveau économique

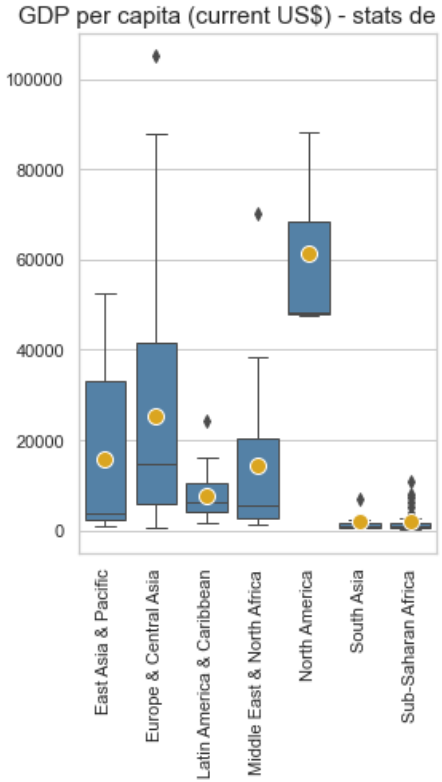
14 INDICATEURS TEMPORELS

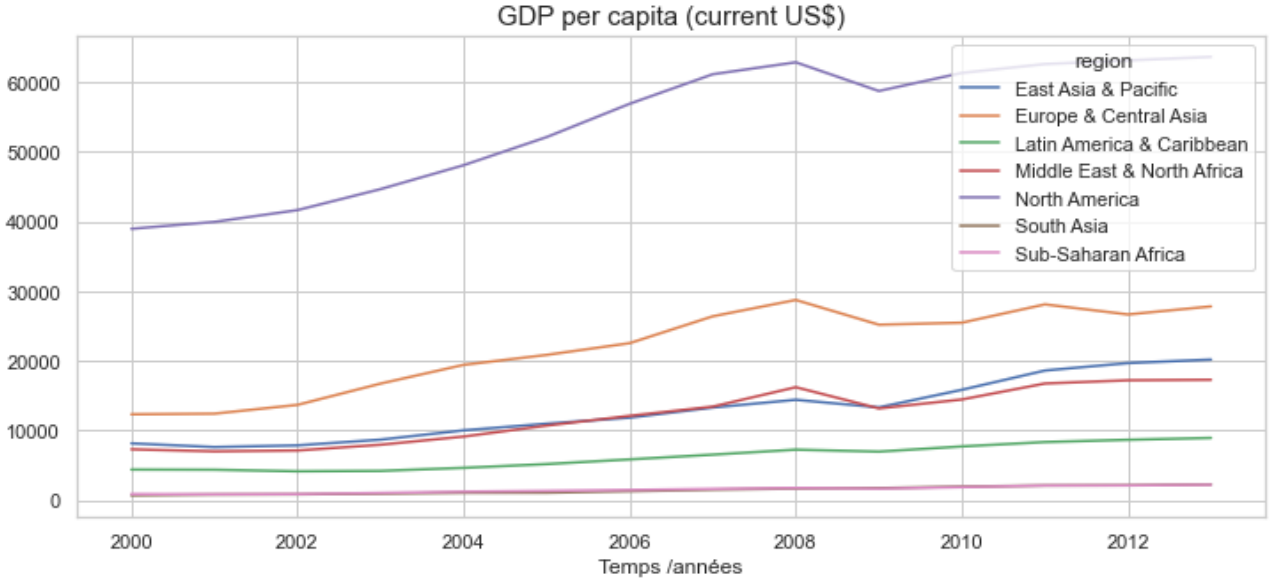
- ANNEE : De 2000 à 2013

3

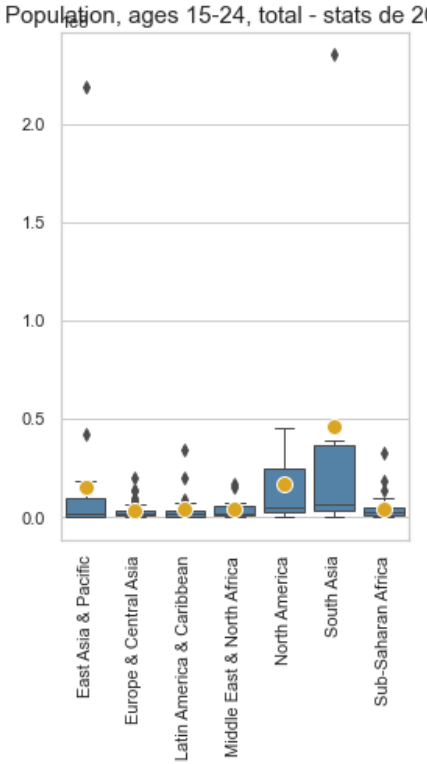
ANALYSE PRÉ-EXPLORATOIRE

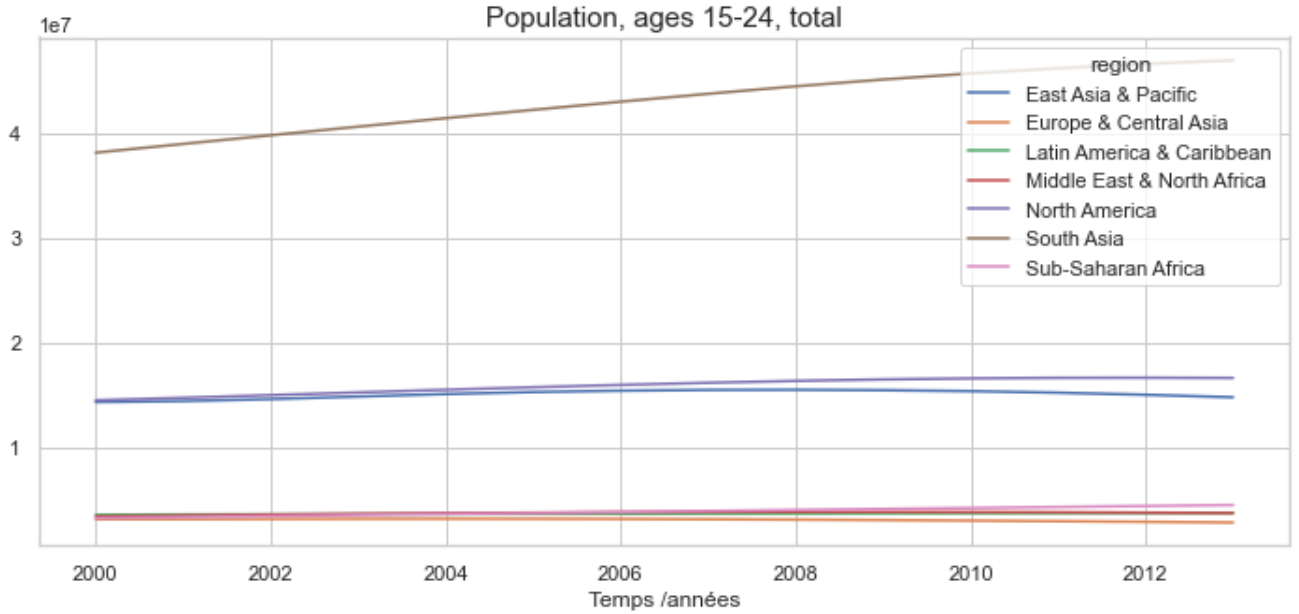
PIB par habitant dans le monde



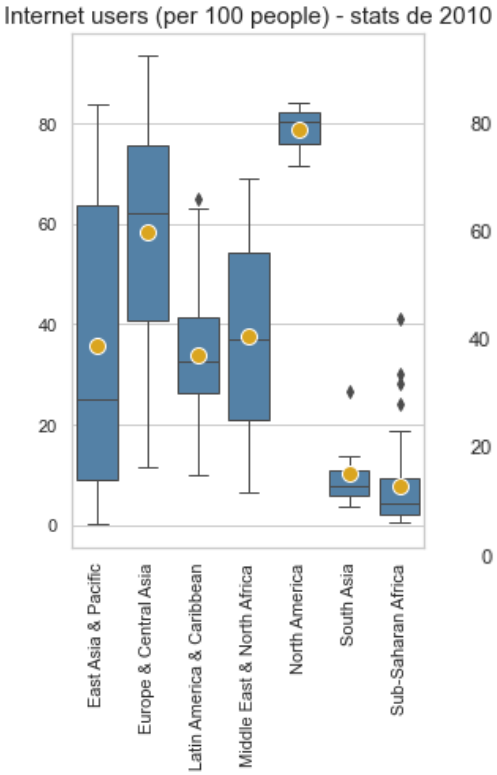


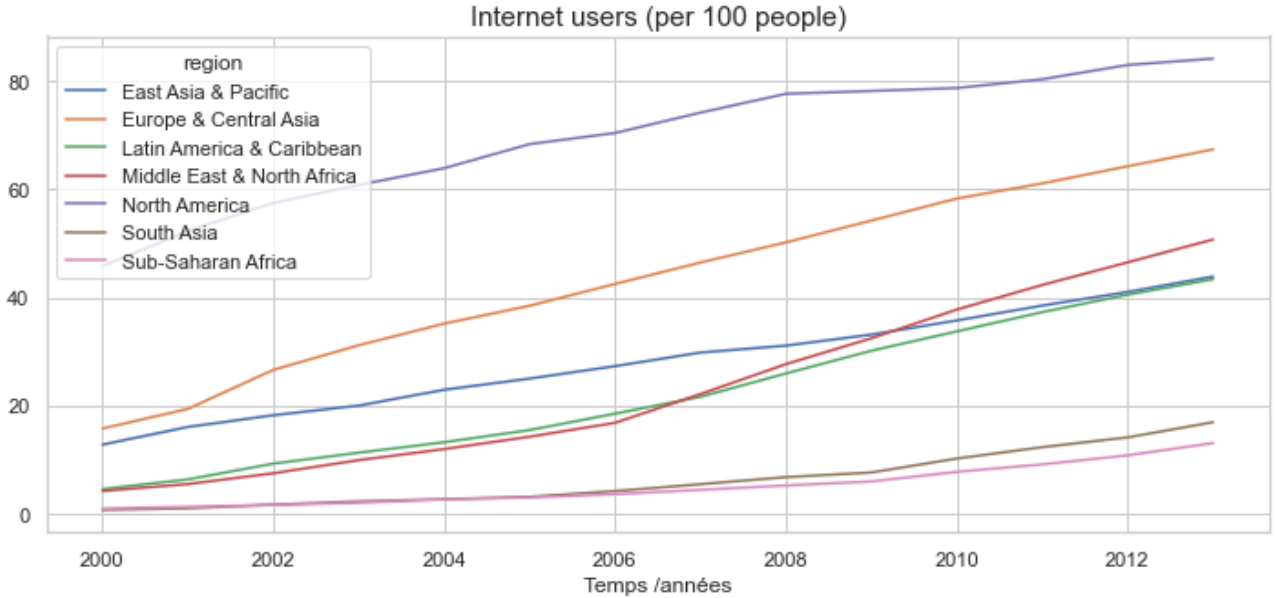
Population, entre 15-24 ans, dans le monde



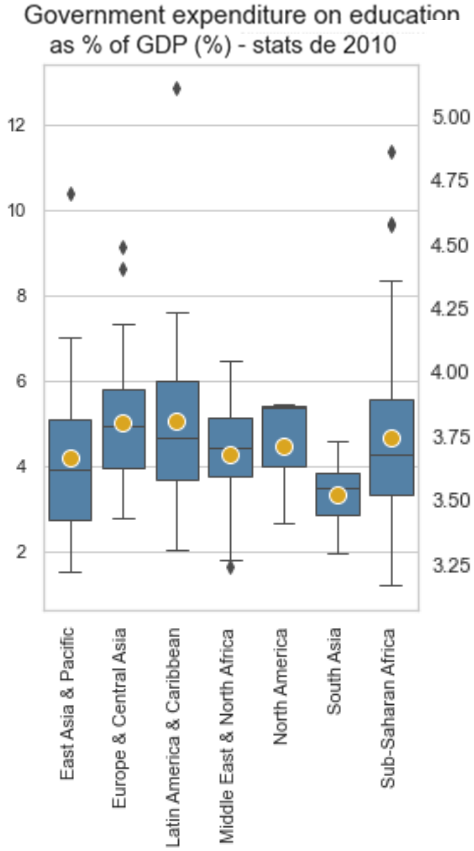


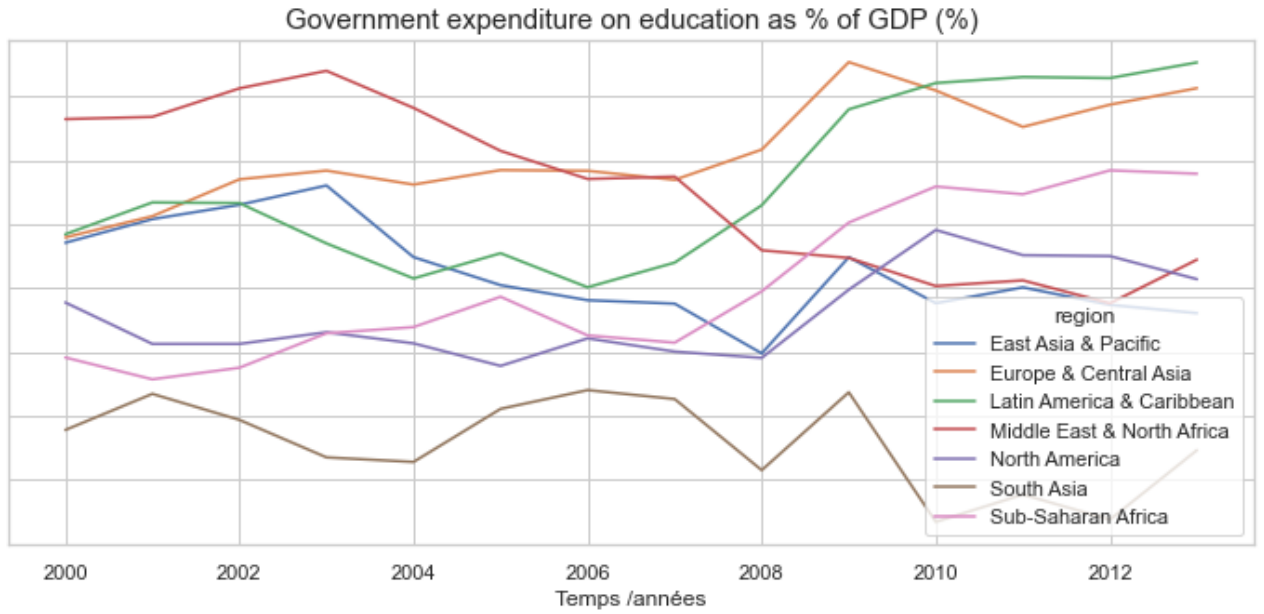
Utilisateurs d'internet (pour 100 p.) dans le monde



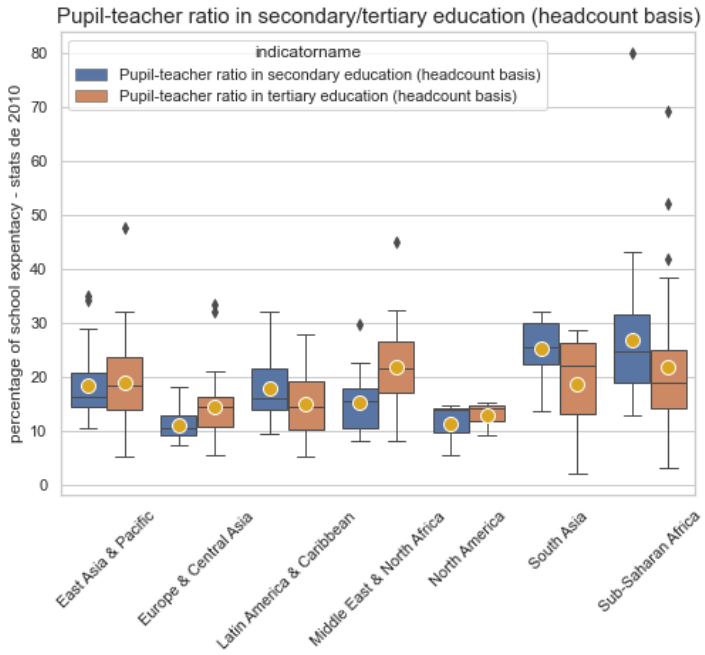


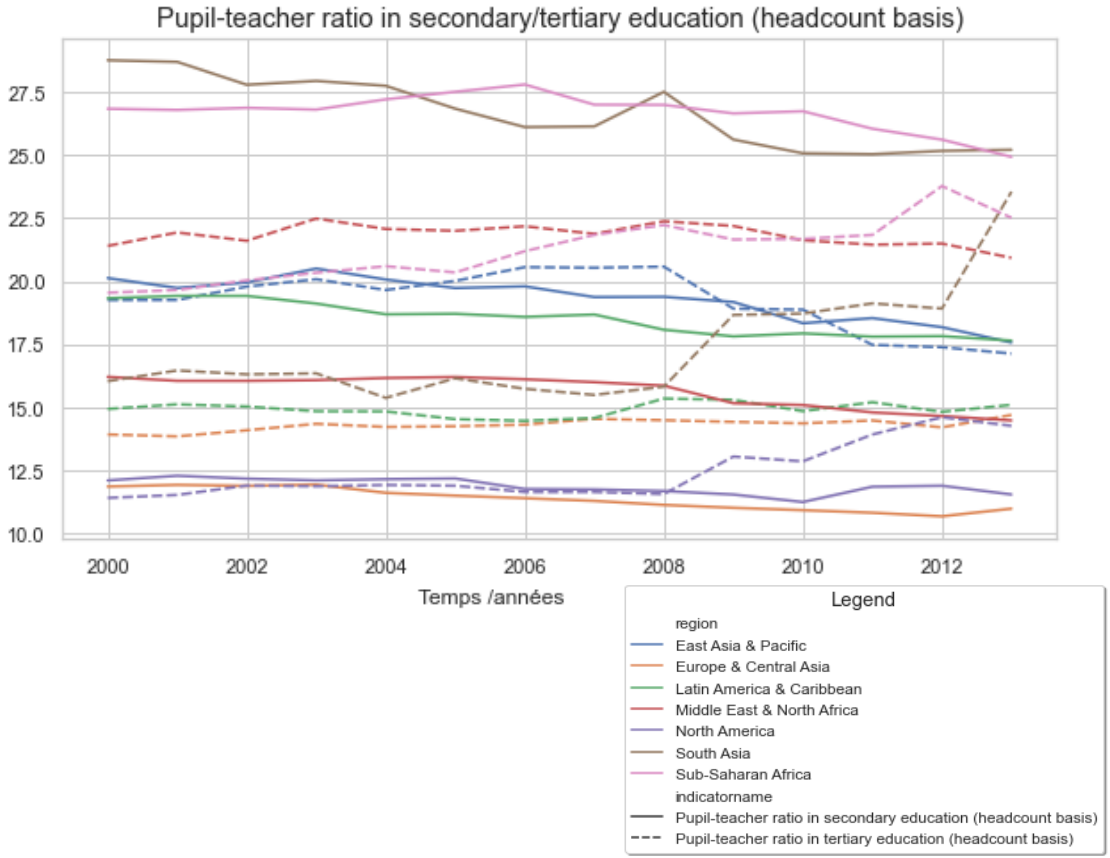
Pourcentage du PIB national dédié à l'education dans le monde



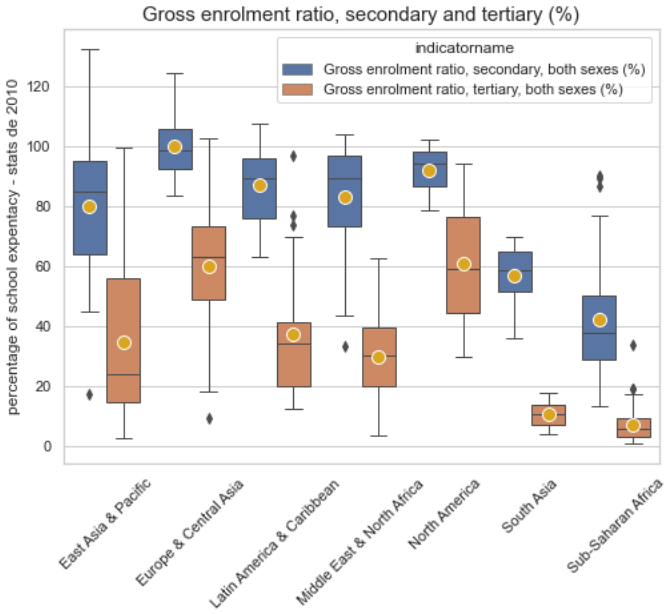


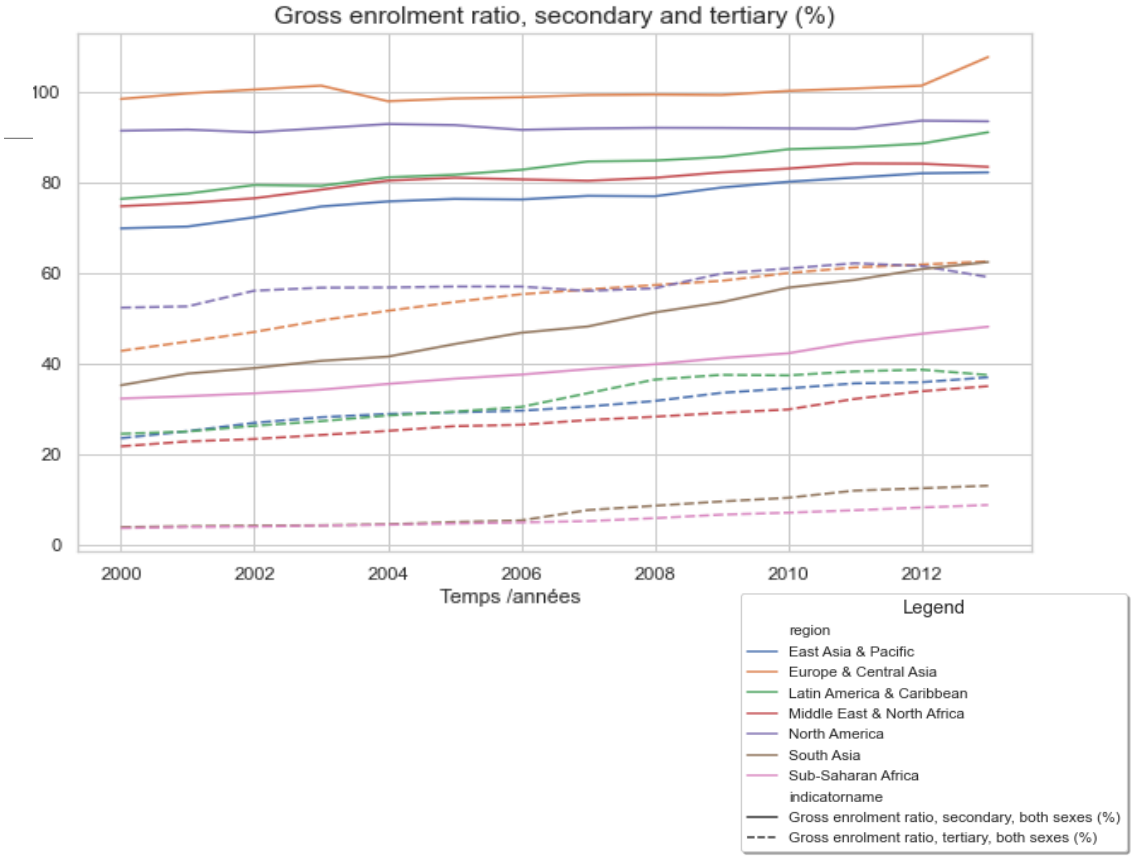
Ratio élèves/prof dans l'éducation secondaire/tertiaire dans le monde





Ratio d'élèves rentrant dans l'éducation secondaire/tertiaire dans le monde





CONCLUSION

Sur l'influence des régions du monde sur les indicateurs statistiques

Le niveau économique a un impact important sur plusieurs des indicateurs statistiques.

→ L'Amérique du Nord et l'Europe (de l'Ouest et Centrale) sont souvent les mieux classés

→ Exception: La taille de la population où le Sud de l'Asie est mieux classé (Inde et Indonésie)

Sur les indicateurs statistiques à retirer pour la suite de l'analyse sur les pays

→ Le PIB liée à l'éducation est forcément lié au PIB total

Un pays pauvre/ communisme peut se retrouver en tête alors qu'il n'est pas forcément le premier choix

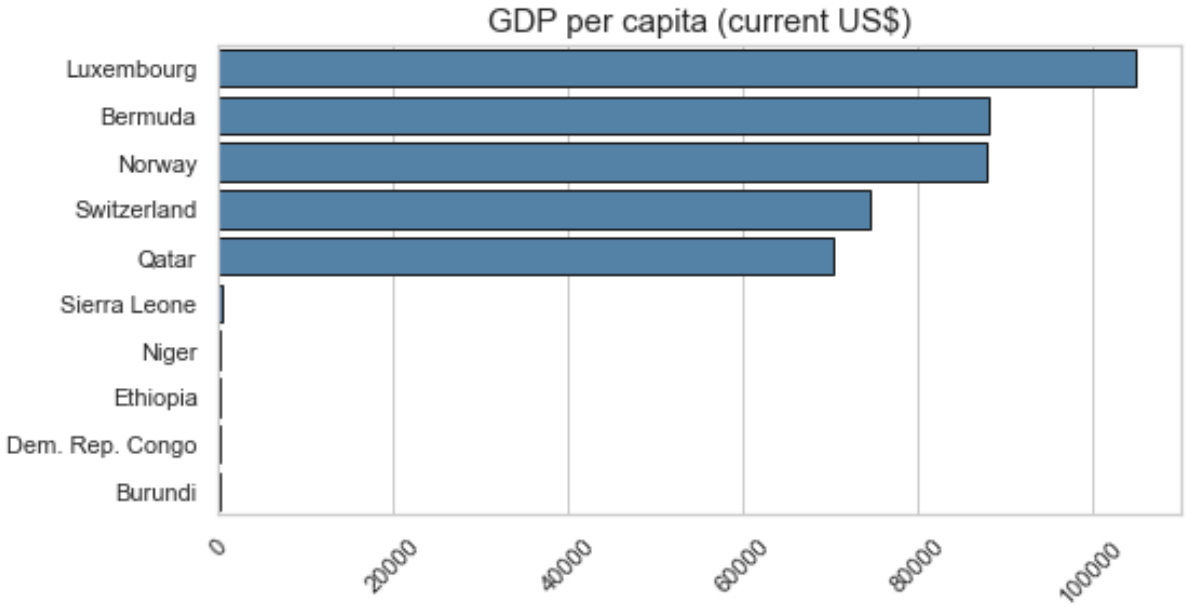
→ Le ratio élèves/professeur dans l'éducation secondaire/tertiaire

Le ratio élèves/professeur est très lié à l'éducation sur place – l'EdTech se situe dans un cadre à part

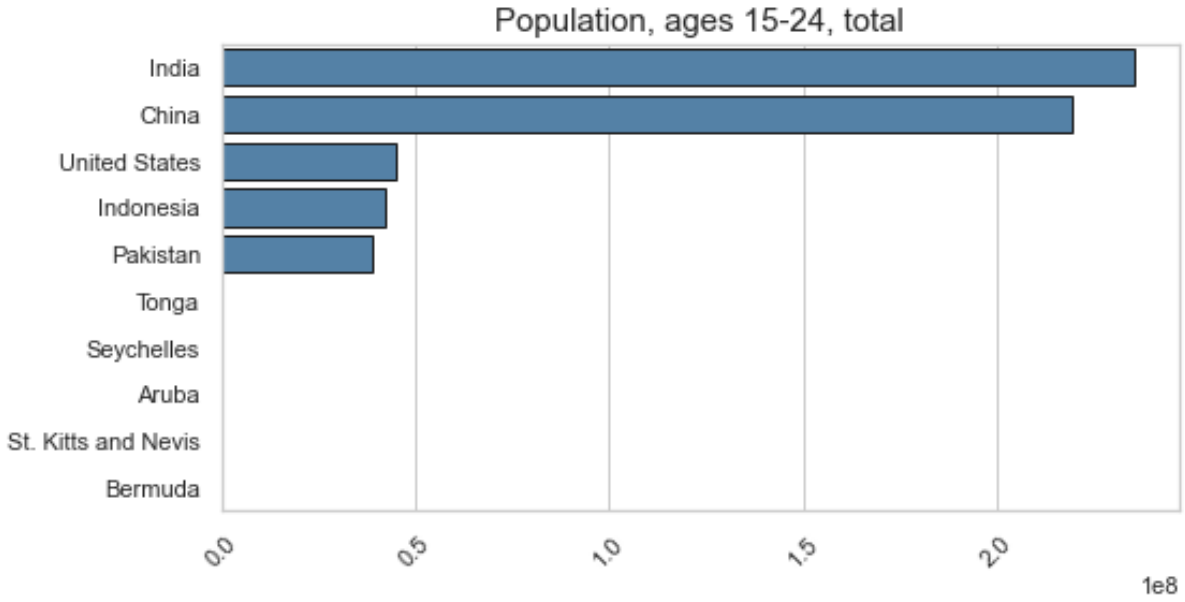
→ Le taux d'élèves rentrant dans l'éducation secondaire/tertiaire

Très lié au niveau économique d'un pays (sinon biais avec les redoublements/ expatriés)

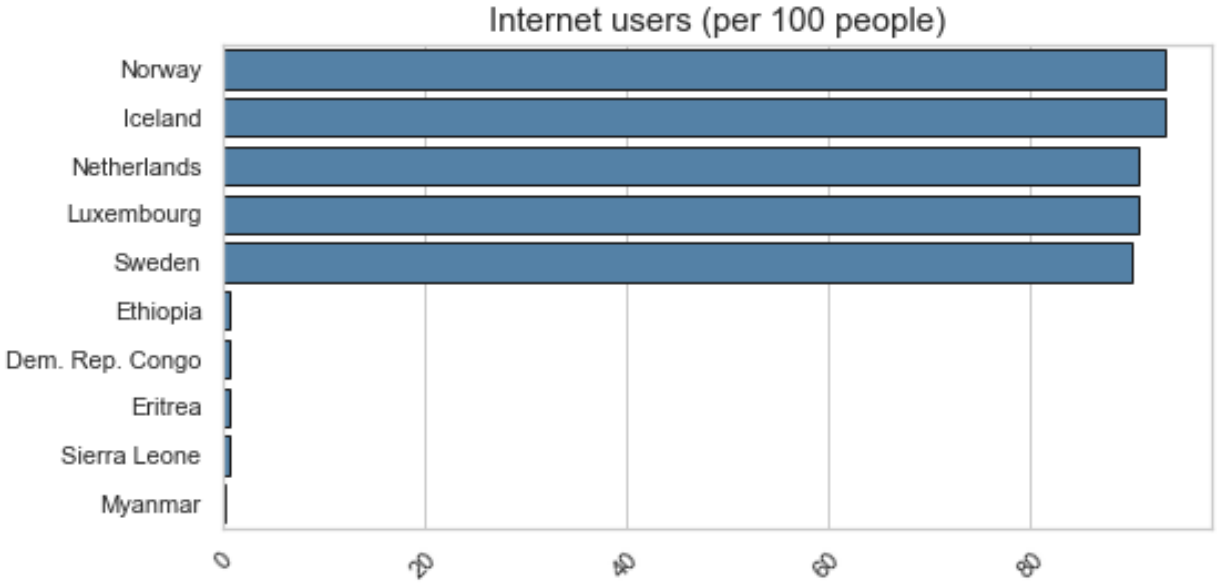
PIB par habitant par pays



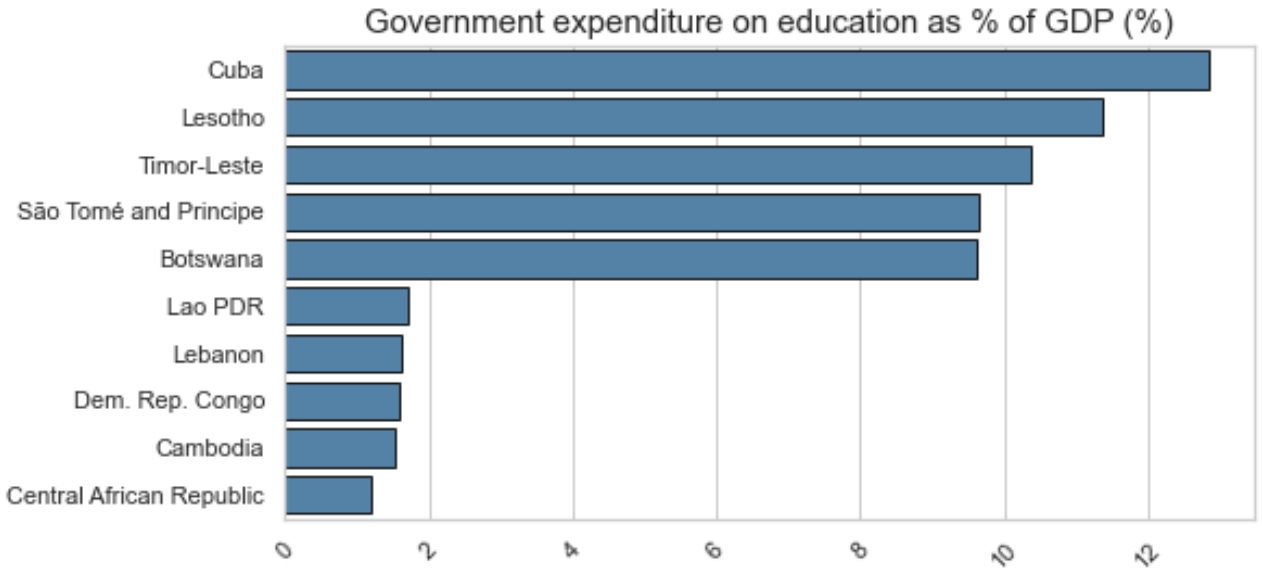
Population, entre 15-24 ans, par pays



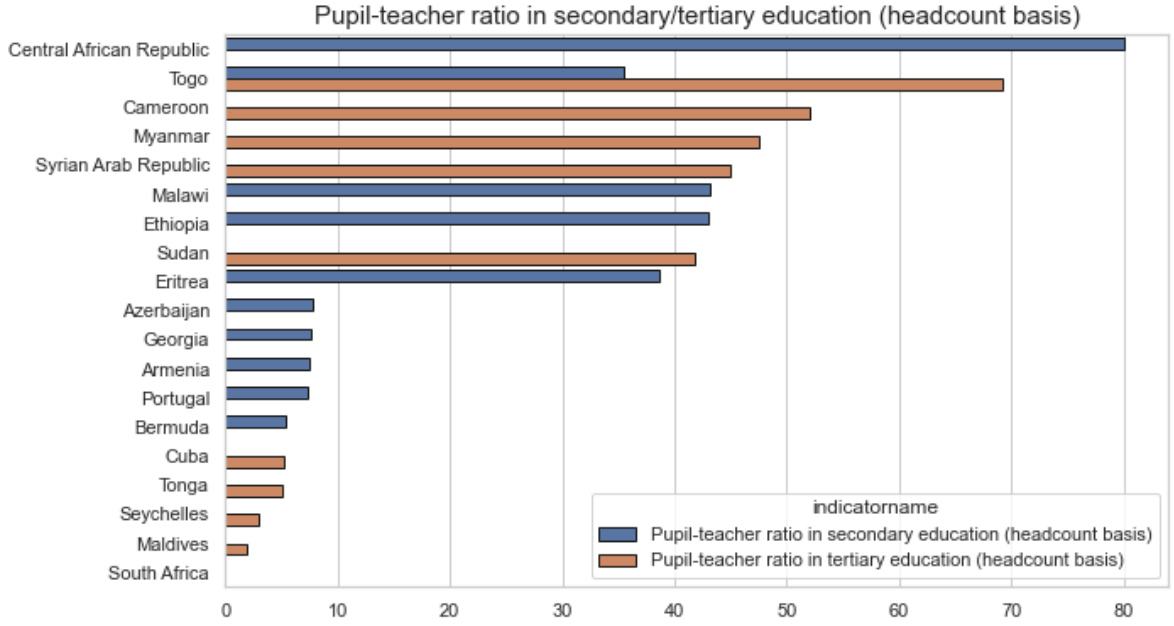
Utilisateurs d’internet par pays



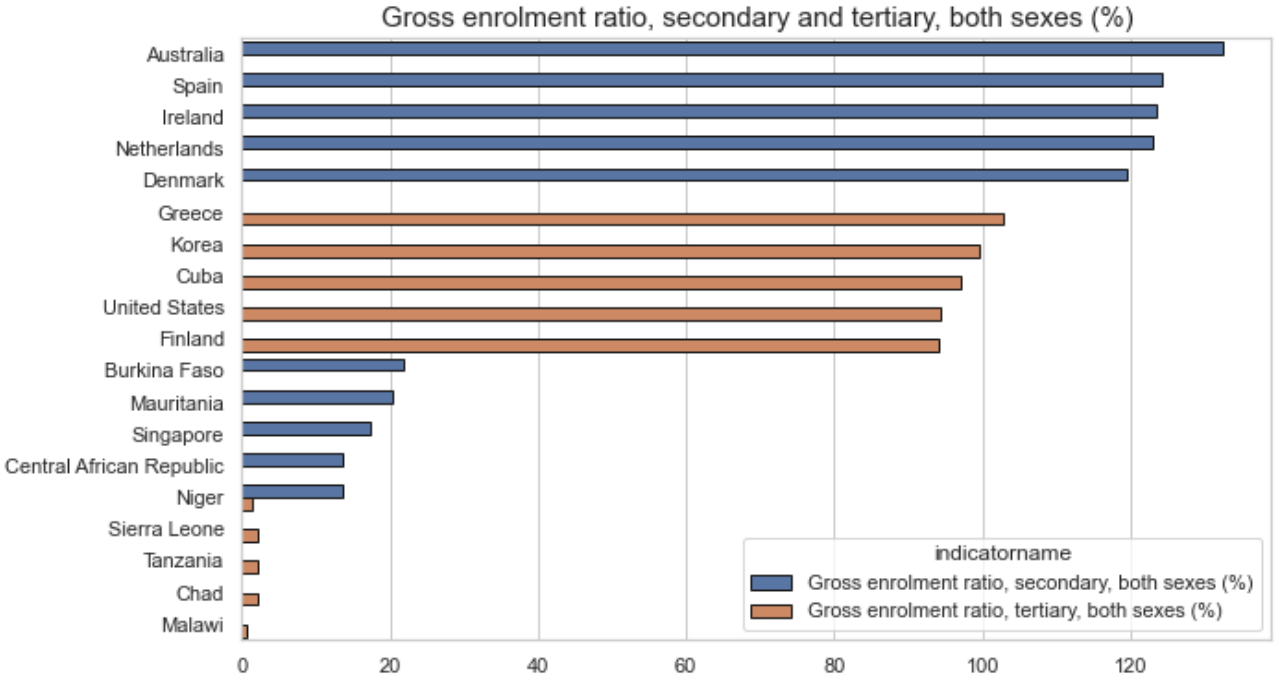
Pourcentage du PIB national dédié à l'éducation par pays



Ratio élèves/prof dans l'éducation secondaire/tertiaire par pays



Ratio d’élèves rentrant dans l’éducation secondaire/tertiaire par pays



CONCLUSION

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

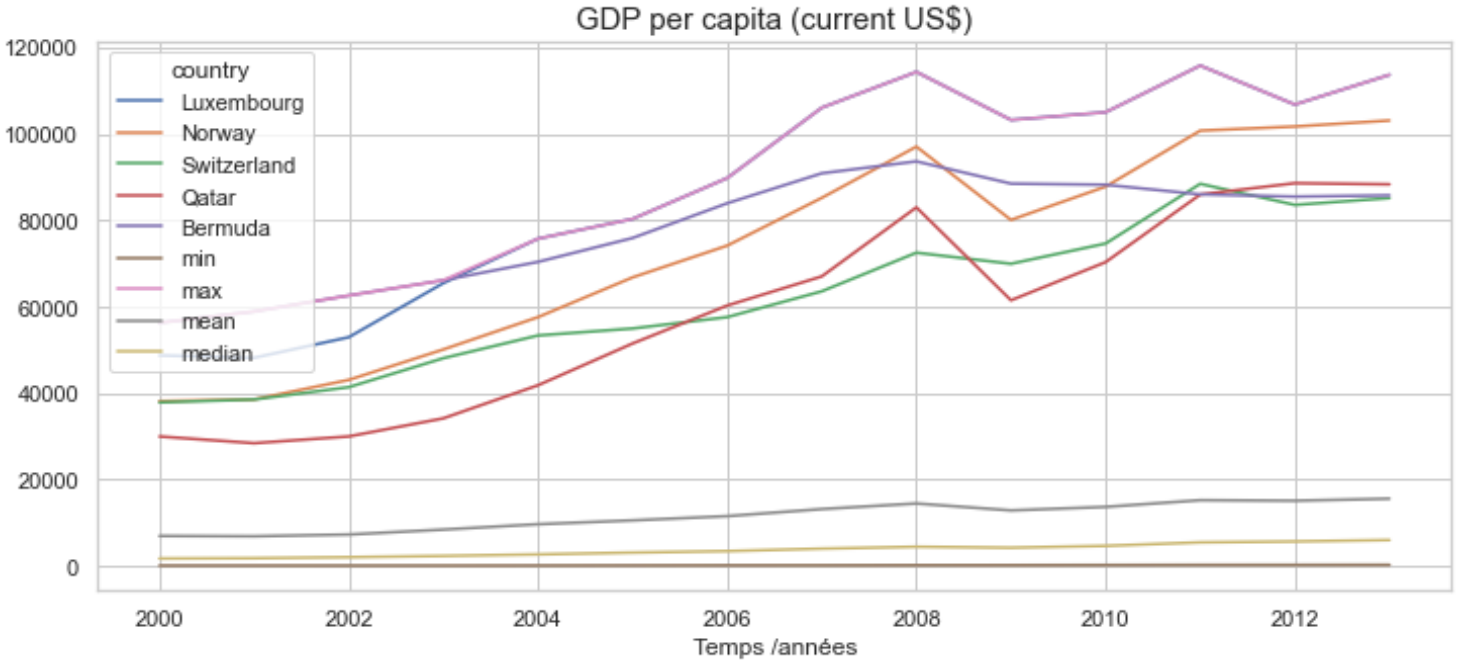
Deux facteurs principaux vont jouer :

- la taille de la population cible (15-24 ans)
- le PIB par habitant

Avec ces deux facteurs en tête les pays au fort potentiel de clients sont:

- L'Inde
- La Chine (dictature)
- Les Etats-Unis
- L'Indonésie
- Le Pakistan
- La Norvège
- Le Luxembourg
- La Suisse
- Le Danemark
- Le Qatar

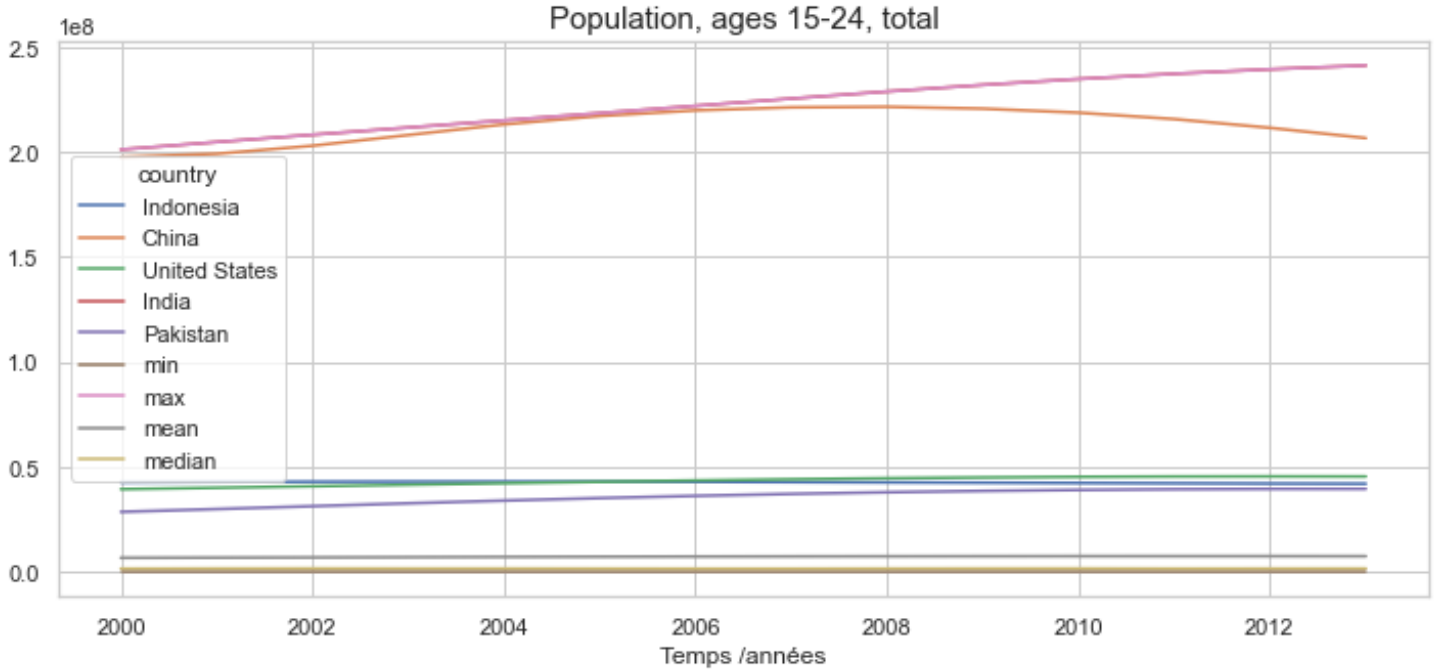
PIB par habitant par pays



La crise de 2008 a eu un impact négatif sur tous les pays sélectionnés mais à partir de 2011, les pays se sont relevés ou se sont stabilisés.

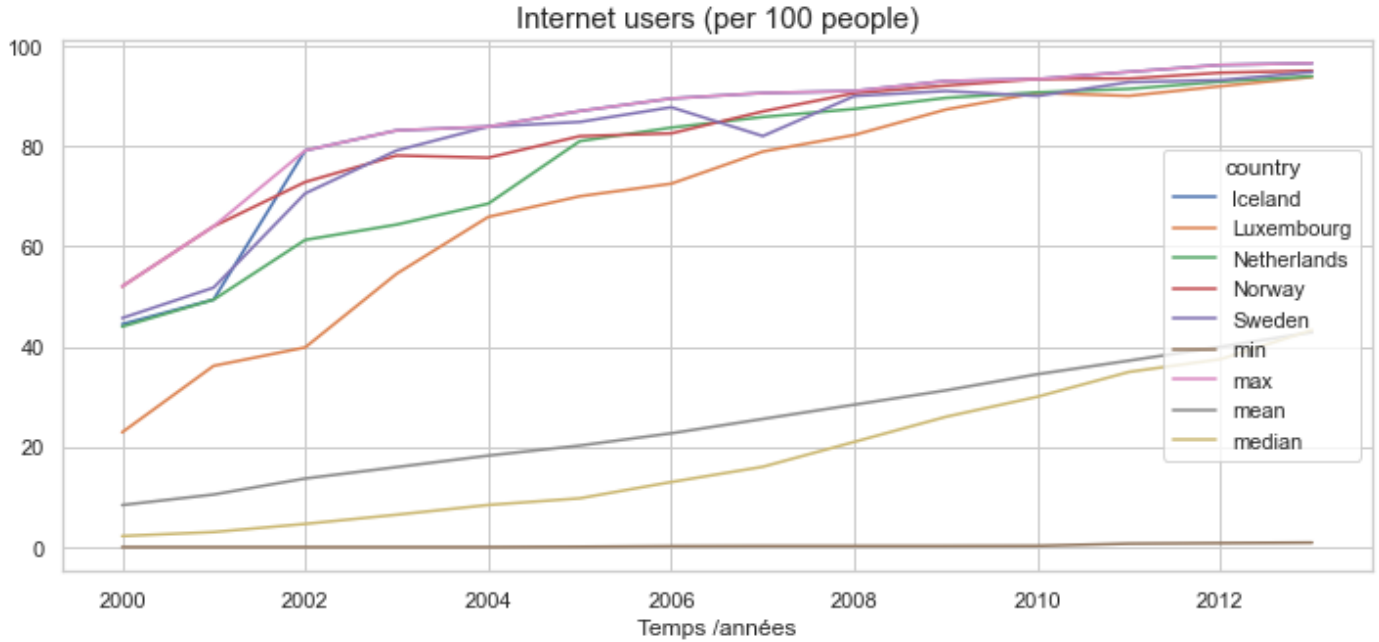
Petits pays surreprésentés – surtout ceux provenant d'Europe

Population, entre 15-24 ans, par pays



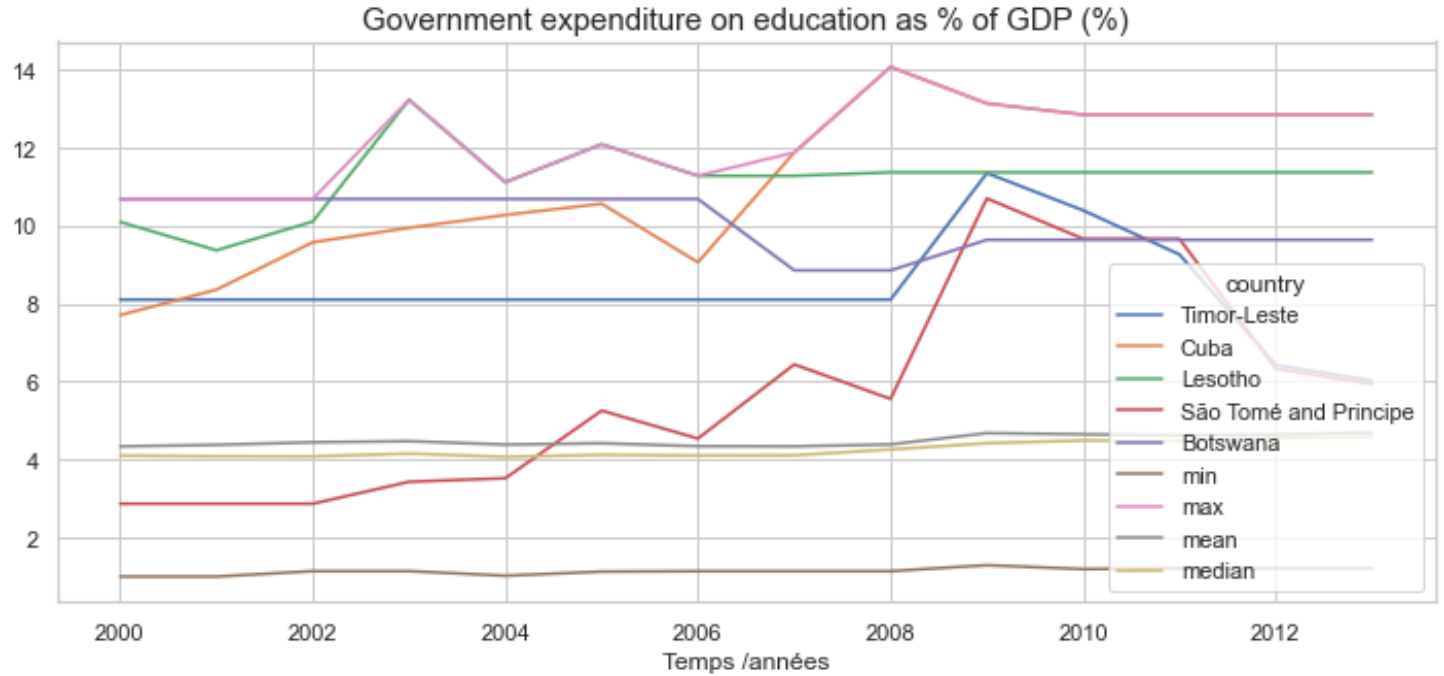
Inde et Chine sont les pays avec le plus de population entre 15-24 ans.
Les USA, le Pakistan et l'Indonésie sont aussi d'excellents candidats.

Utilisateurs d'internet par pays



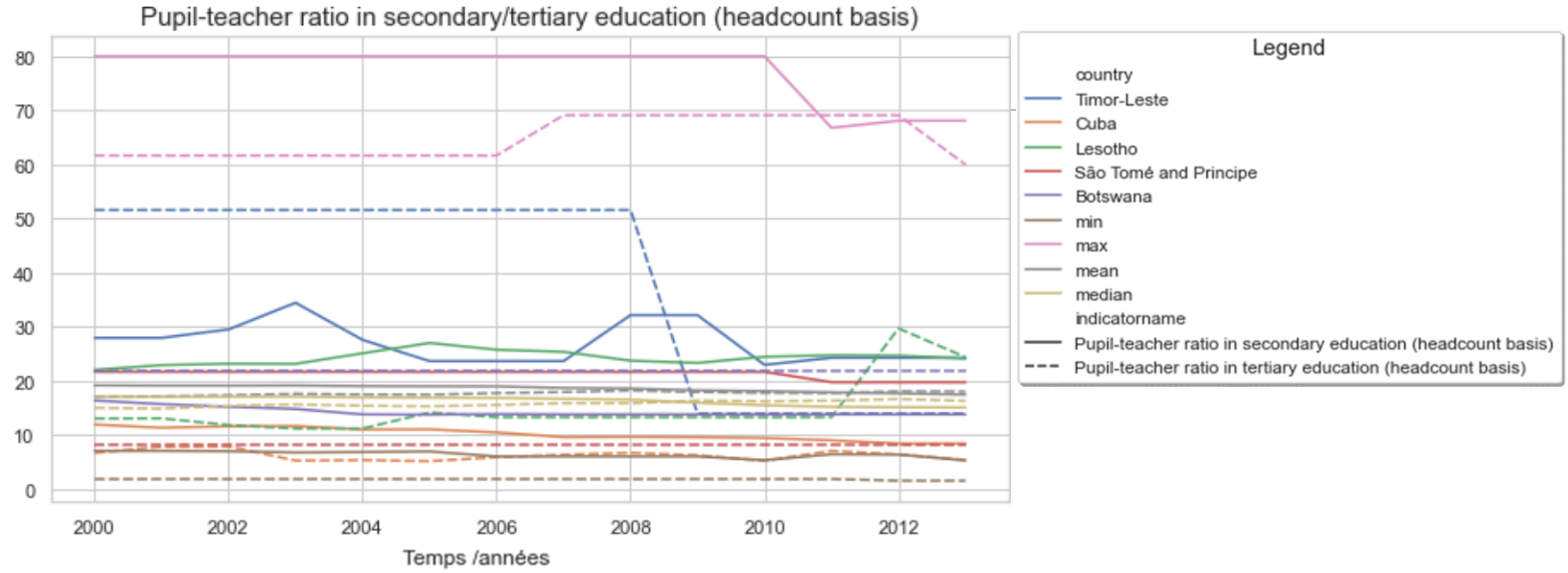
Petits pays surreprésentés – tous provenant d'Europe

Pourcentage du PIB national dédié à l'éducation pays



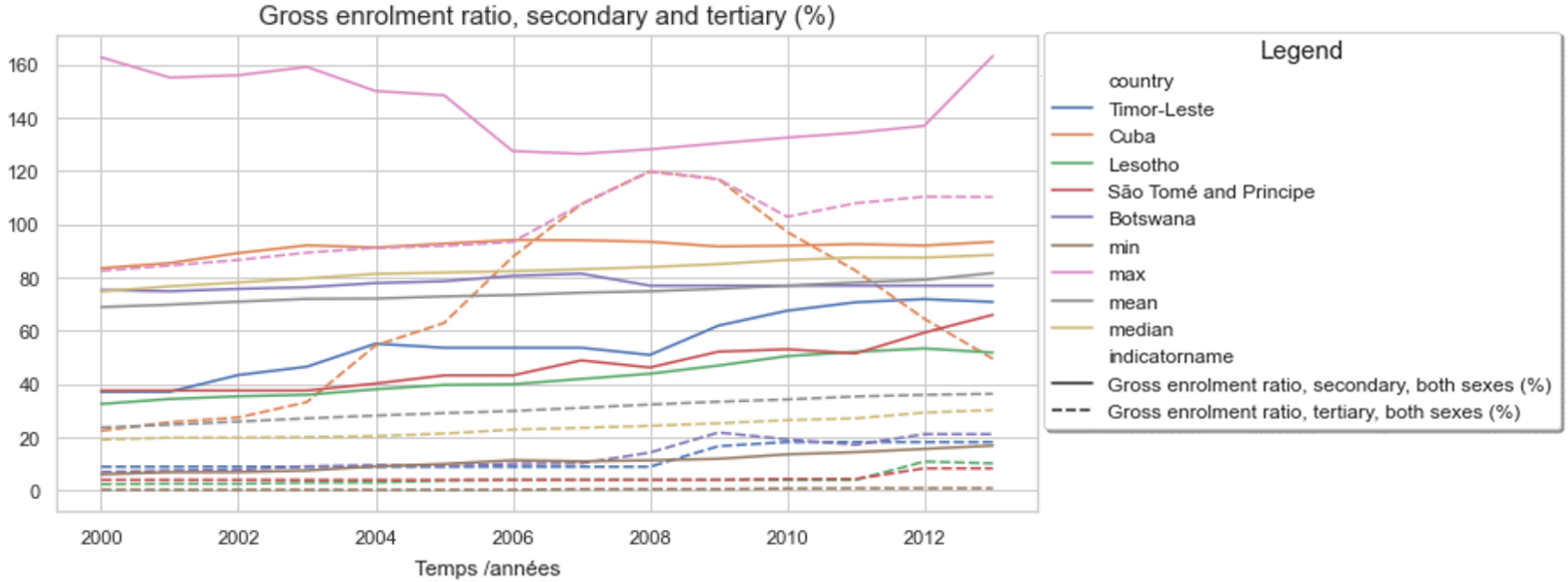
Petits pays et assez pauvre (ou avec une dictature en place) surreprésentés

Ratio élèves/prof dans l'éducation secondaire/tertiaire par pays



Petits pays et assez pauvre (ou avec une dictature en place) surreprésentés

Ratio d'élèves rentrant dans l'éducation secondaire/tertiaire par pays



Petits pays et assez pauvre (ou avec une dictature en place) surreprésentés

CONCLUSION

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

- L'Inde
  - Population en augmentation
- La Chine (dictature)
  - Population en diminution (mais toujours 2nd plus gros pool)
- Les Etats-Unis
  - Population stable
- L'Indonésie
  - Population stable
- Le Pakistan
  - Population stable

PIB par habitant en augmentation ou stagnation pour tous ces pays:

- Le Luxembourg
- La Norvège
- Le Danemark
- La Suisse
- Le Qatar

4

CONCLUSION

CONCLUSION

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

Deux axes:

- Soit les petits pays (en Europe) où le niveau de vie est très élevés:
  - Luxembourg
  - Norvège
  - Danemark

- Soit les pays avec le plus d'individus pouvant devenir des futurs clients :
  - L'Inde
  - L'Indonésie
  - Les Etats-Unis
  - La Chine (dictature – fermeture de plus en plus étroite sur les sociétés étrangères)

CONCLUSION

FORCE DE CE JEU DE DONNÉES…

- Quantité d'informations sur l'éducation
- Quantité d'informations sur les pays
- Qualité de l'information (Banque Mondiale)

… ET LES LIMITES DE CETTE ETUDE

- Pas assez d'analyses fines (ou bivariées entre deux indicateurs statistiques) et aucune analyse de corrélation ou d'ANOVA par manque de temps

… ET CES LIMITES

- Enormément de données manquantes (+ 80%)
- Manque d'informations de type sécurité/business/politique
- Manque d'information sur l'entreprise
- N'a pas les informations de ces dernières années (ou qu'en prédiction) – En 7/8 ans les tendances peuvent changer radicalement

MERCI POUR VOTRE ATTENTION

A vos questions !