Soutenance Projet 2

Analyse exploratoire sur The World Bank projet d'expansion à l'internationale

Michel Blazevic

Lundi 27 Décembre

academy

OPENCLASSROOMS - Formation Data Scientist

Ordre du jour

- Introduction – Problématique
- Etude pré-exploratoire des données
  - Présentation des datasets
  - Nettoyage
  - Sélectionbdes informations pertinentes
- Analyse
  - Observation sur les indicateurs
  - Classement par indicateur
  - Etude de l'évolution
- Conclusion

Introduction – Présentation de la problématique

Academy: Start-up EdTech
Formation en ligne niveau Lycée/Université

Volonté de s'étendre à l'international

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

Etude pré-exploratoire des données:

Présentation des datasets

Nettoyage

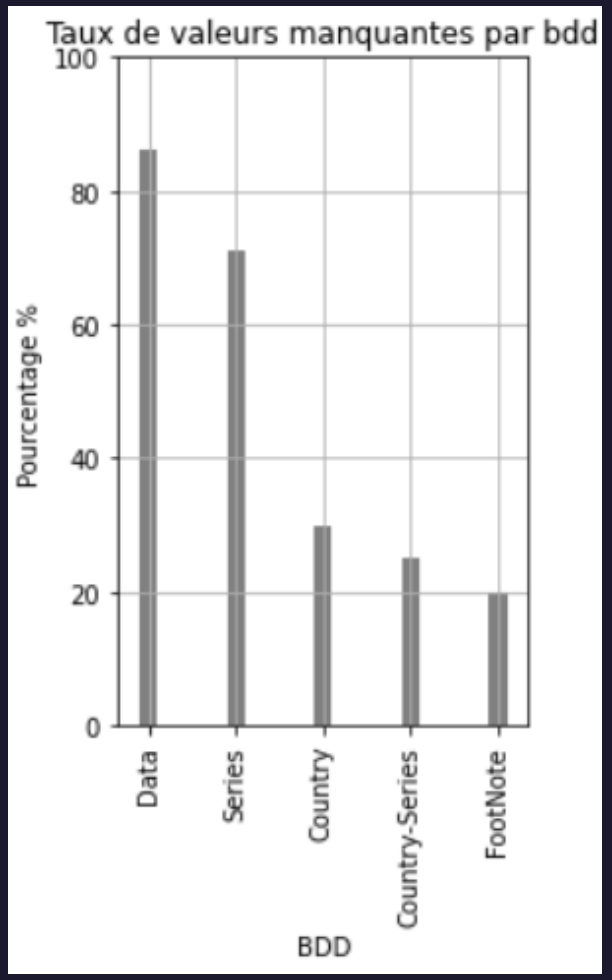
Sélection des informations pertinentes

Etude pré-exploratoire des données

Présentation des Datasets (World Bank)

- EdStatsCountry-Series: Quelques indicateurs par pays
- EdStatsFootNote: MétaData sur évolutions indicateurs par pays
- EdStatsCountry: Informations par rapport aux pays et aux régions par classe (ex: revenu...)
- **EdStatsData: Valeurs indicateurs par pays et par années (1970-2100)**
- **EdStatsSeries: Descriptions chaque indicateurs**

	Redondance	Valeurs manquantes	taux NA%	Nb ligne	Nb colonne
Country-Series	0	613	25	613	4
Country	0	2354	30	241	32
Data	0	53455179	86	886930	70
Series	0	55203	71	3665	21
FootNote	0	643638	20	643638	5



Présentation 2 datasets

Data & Series

Data

Date, année (1970-2100)

Indicateurs/Pays
& régions
(3665x240)

70 colonnes
879600 lignes
81% valeurs manquantes

Series

Informations indicateurs (définition etc…)

Indicateurs
3665

21 colonnes
3665 lignes
71% valeurs manquantes

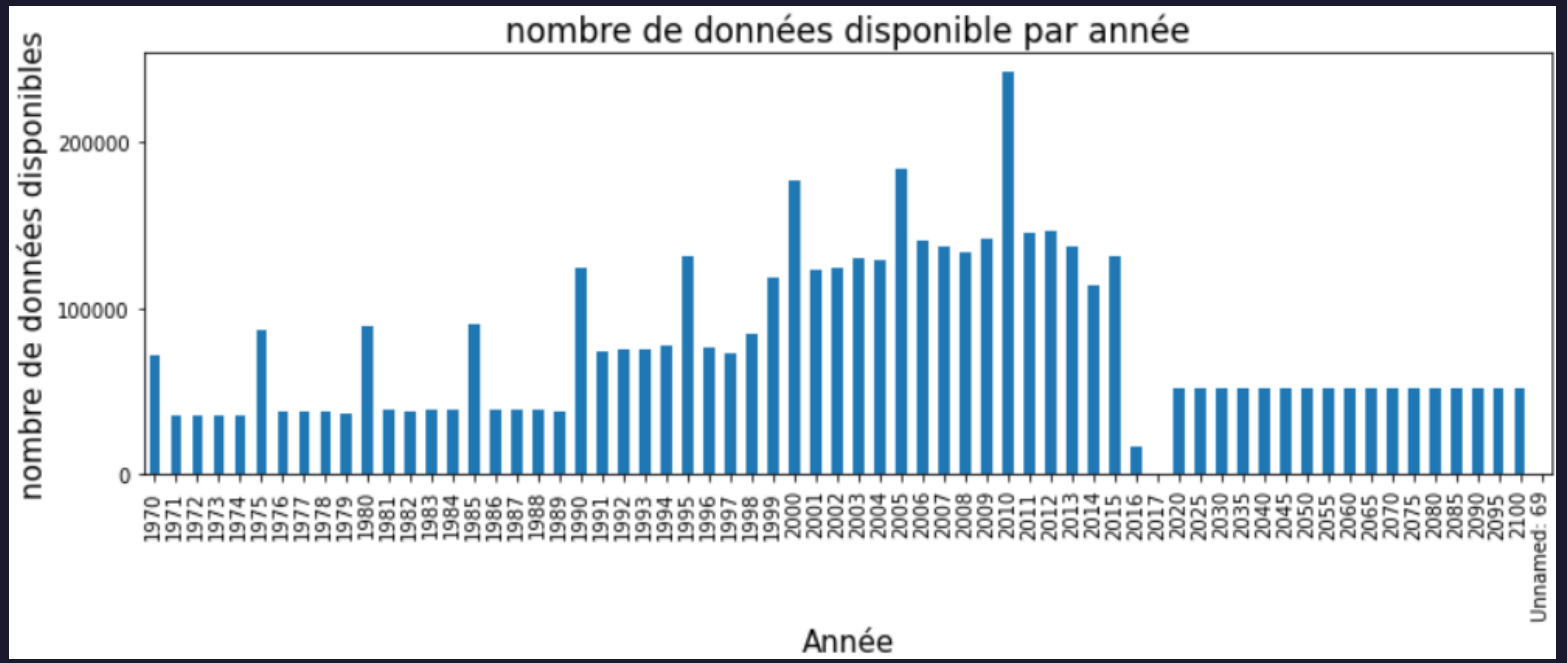
Etude des Datasets :

Nettoyage (1)

Sélection des années

⇒ EdStatsData:

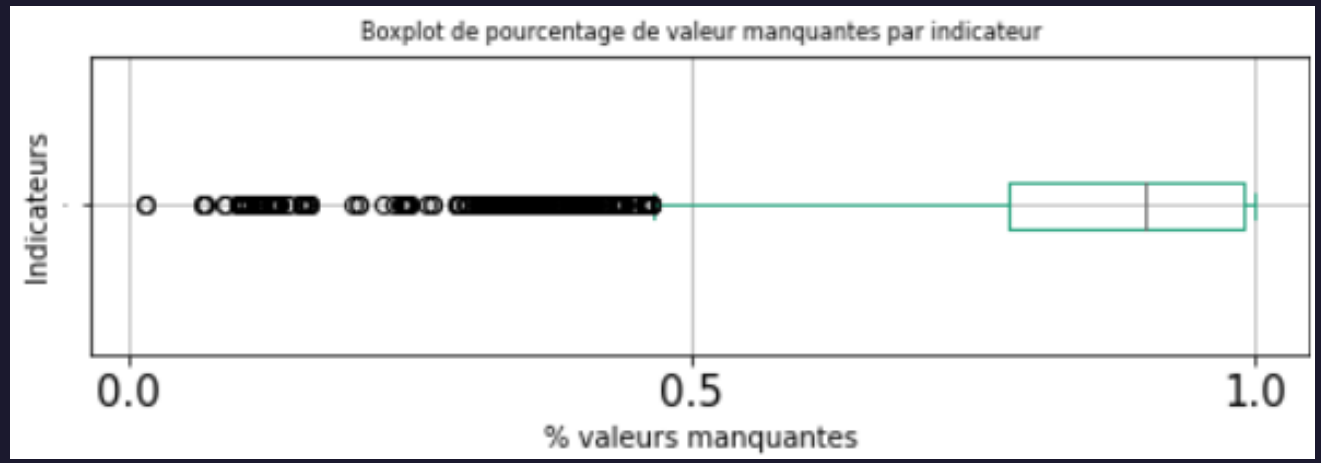
- Par rapport aux valeurs manquantes
- <2005: Pas cohérents pour marché
- >2016: faible nombre et prévision



Choix de garder ***2005*** à ***2015***

Etude des Datasets :
Nettoyage (2)

count	3665.000000
mean	0.830236
std	0.231603
min	0.015026
25%	0.780240
50%	0.902329
75%	0.990233
max	1.000000
dtype:	float64

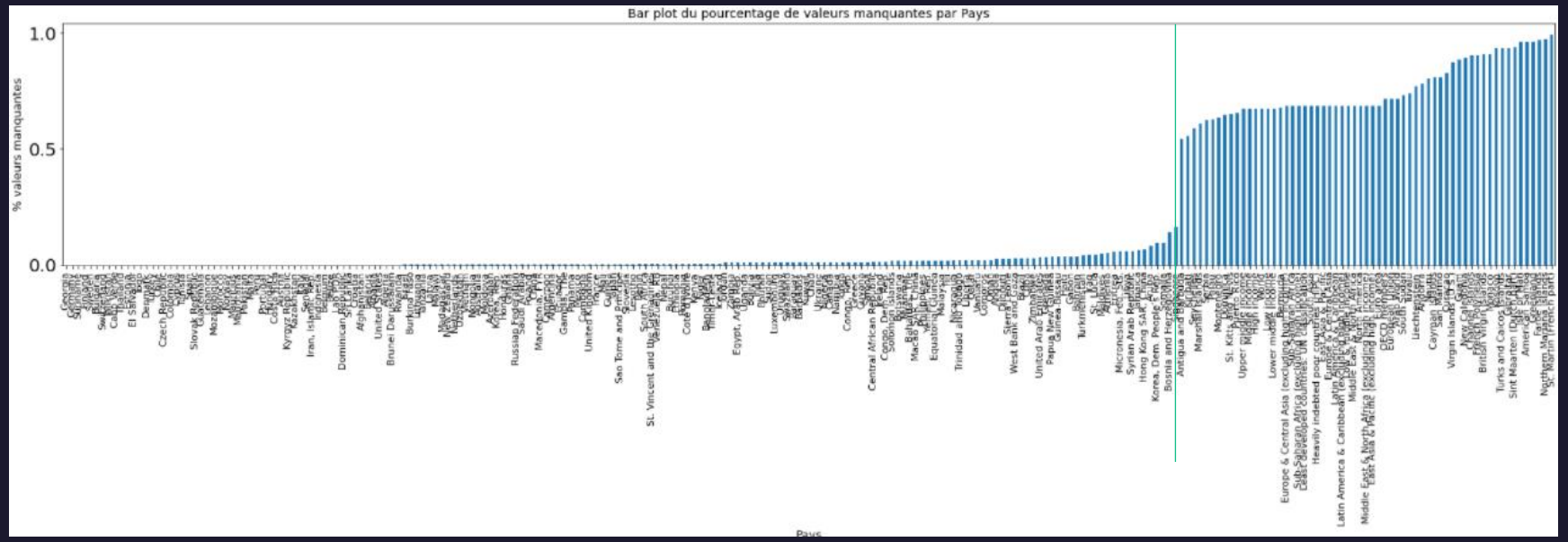


⇒ EdStatsData indicateurs par rapport aux valeurs manquantes:

- Beaucoup de valeurs aberrantes
- Beaucoup d'indicateurs peu renseignés
- Choix de garder %NA > ***min + std***

206 indicateurs

Etude des Datasets :
Nettoyage (3)



⇒ EdStatsData:

- Par rapport aux valeurs manquantes
- Beaucoup d'indicateurs peu renseignés
- Choix de garder %NA > 75% quartile

181 Pays

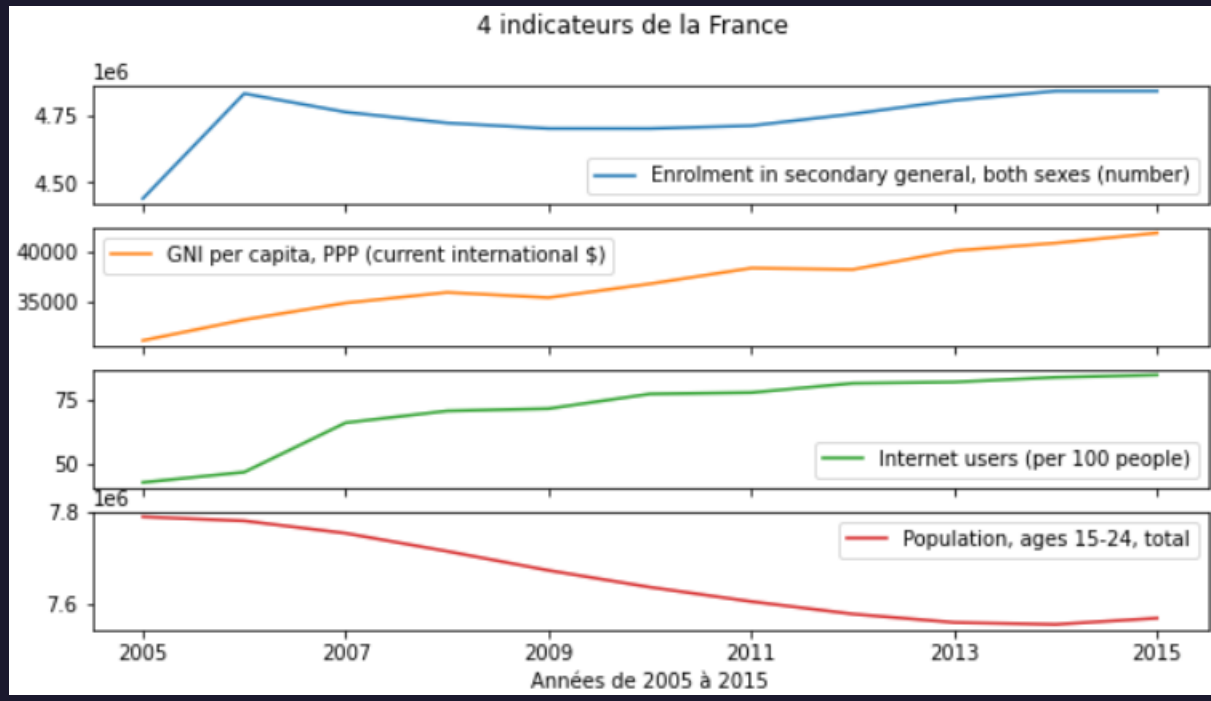
Sélection des indicateurs pertinents

Recherche par mots clefs dans les définitions longues de EdStatsSeries

(Démographie, niveau de vie et les potentiels clients)

```
Lst_keywords=['Demography','demography','15','20','23',
              'GPD','Economy','economy','Internet',
              'internet','school','School','Education','education']
```

- Population, ages 15-24, total
- Internet users (per 100 people)
- GNI per capita, PPP (current international $)
- Enrolment in secondary general, both sexes (number)



Analyse:

Observation sur les indicateurs

Classement par indicateurs

Etude de l’évolution

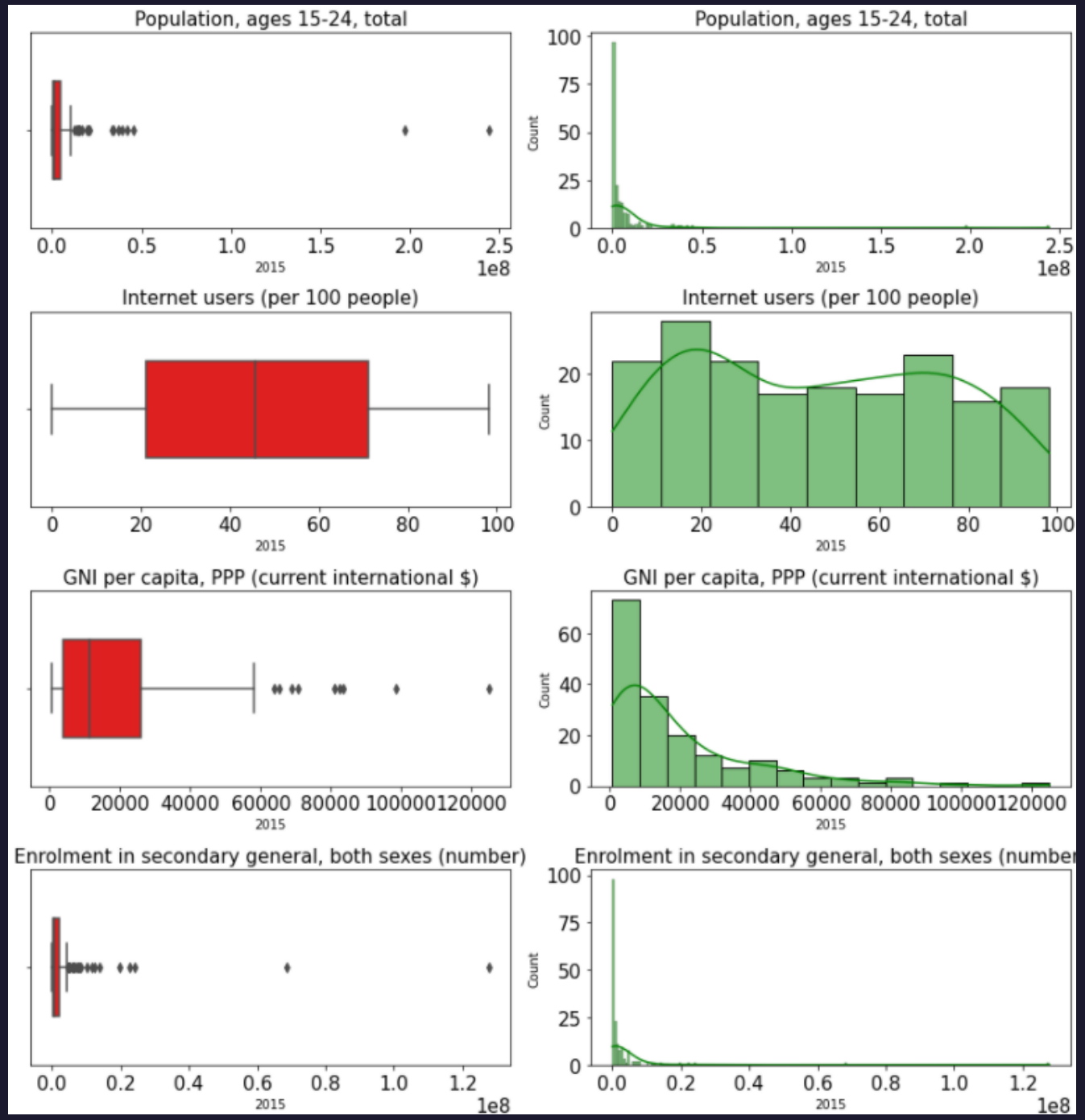
Analyse:

Observation sur les indicateurs

⇒ Population 15-24 ans ~ inscription lycée

⇒ Utilisateurs internet: disparate

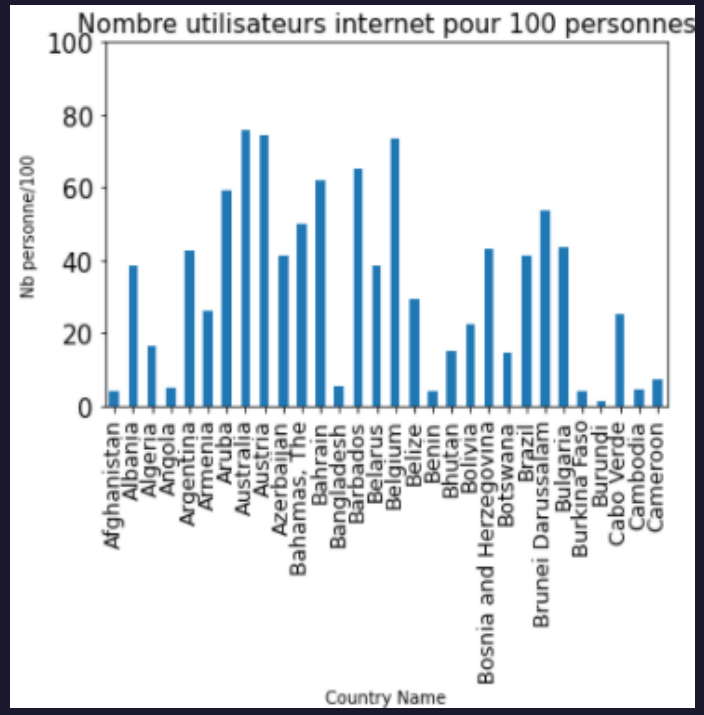
⇒ Nombreux pays avec niveau de vie « bas »
mais avec beaucoup d'outliers



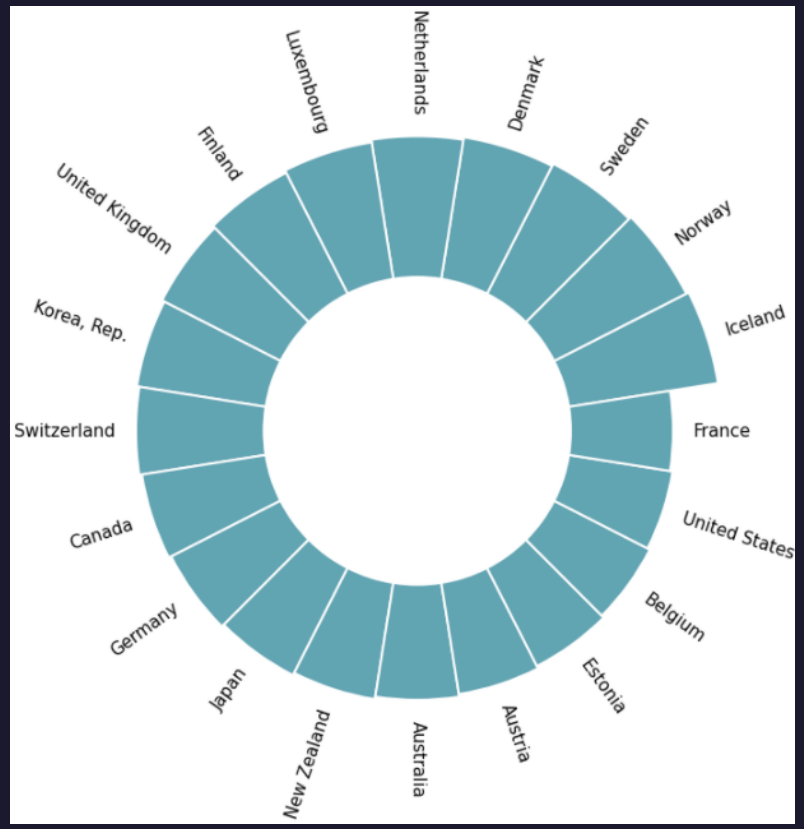
Analyse

Classement par indicateurs (1)

Comparaison entre pays pour chaque indicateur (moyenne 10 ans)



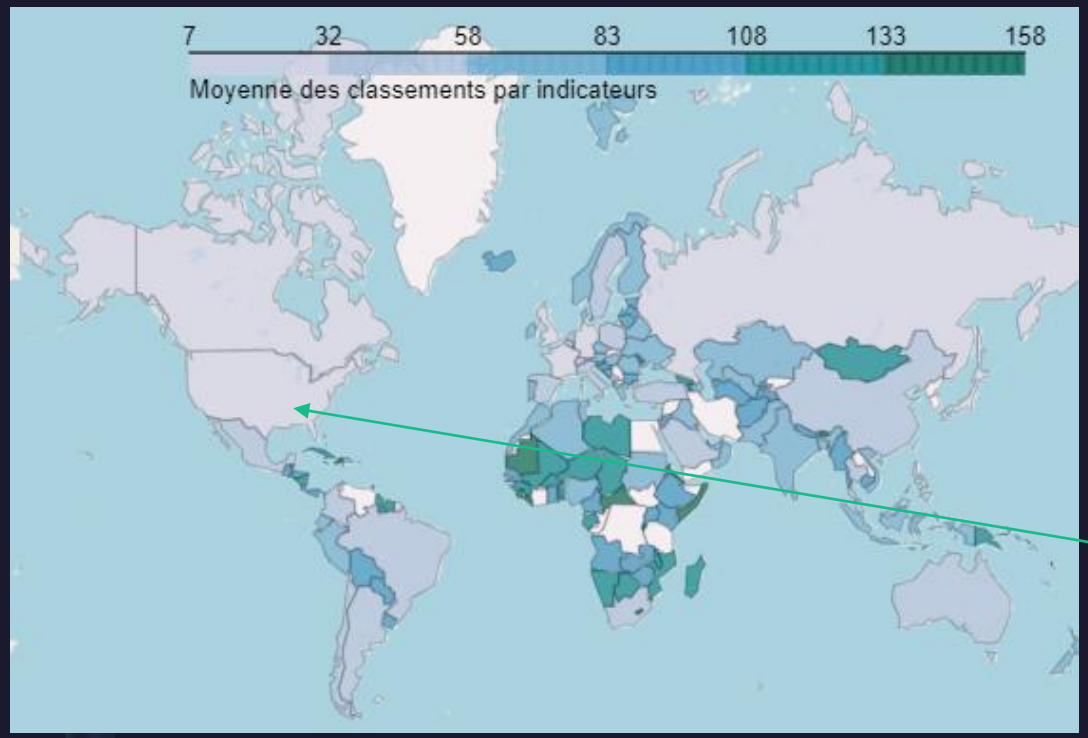
Classement attribué à chaque pays pour chaque indicateur

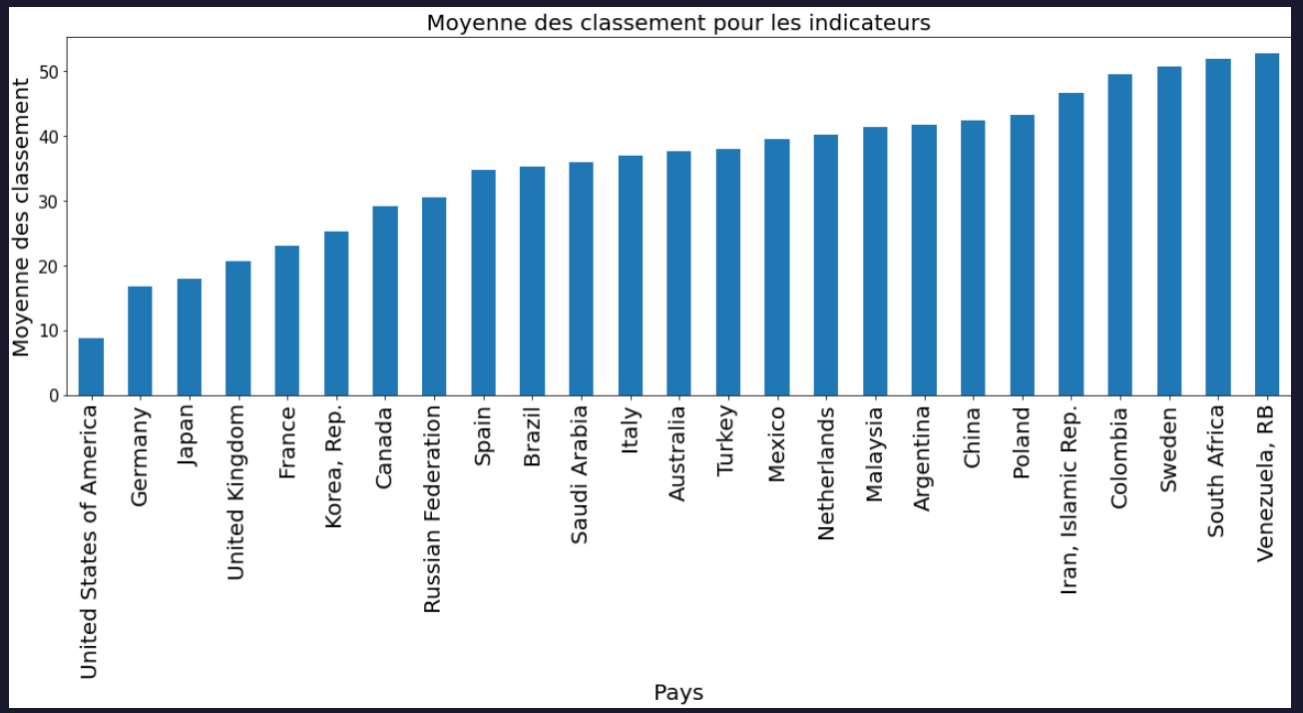


Moyenne des classements de chaque indicateur pour faire un classement général

Analyse

Classement par indicateur





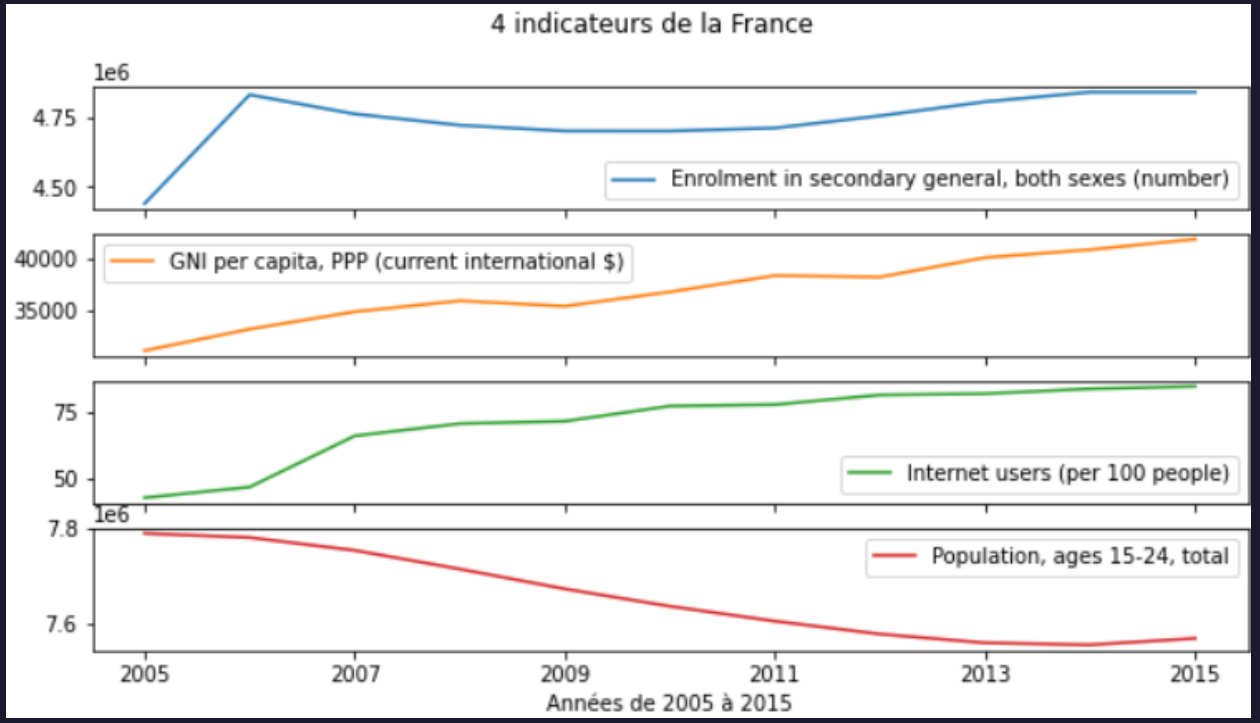
- ***Quels sont les pays avec un fort potentiel de clients pour nos services ?***

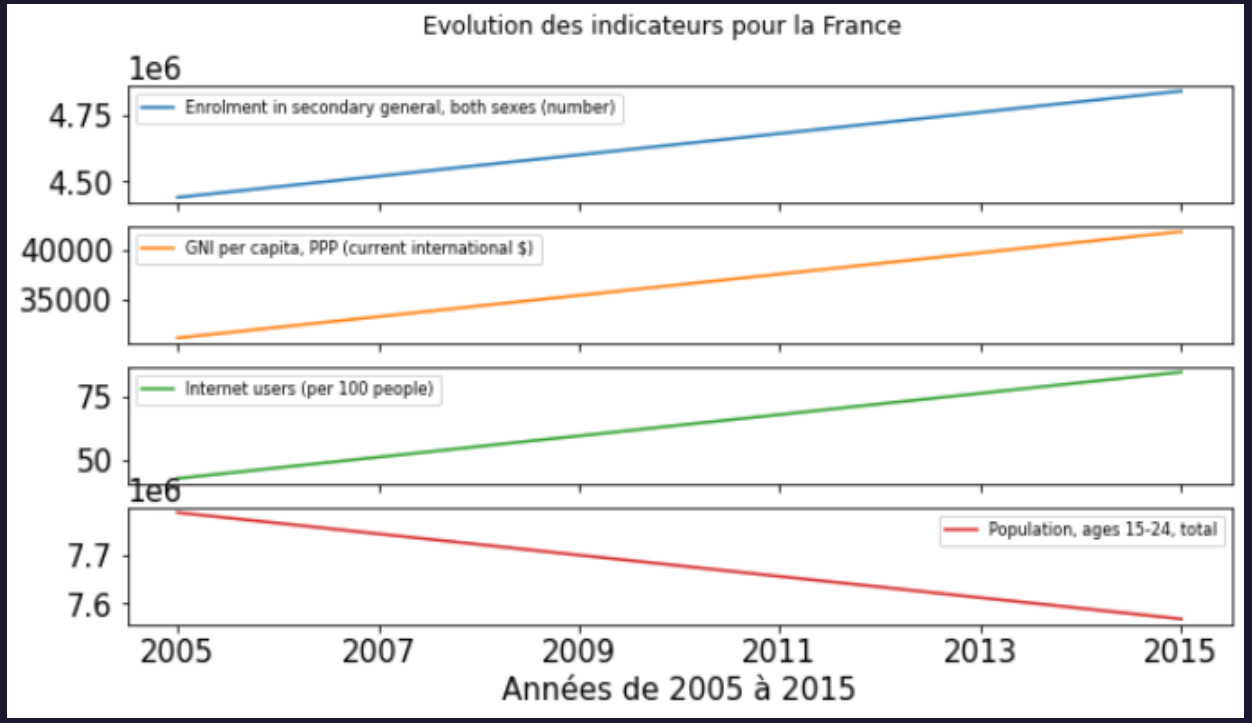
Moyenne des classements pour chaque indicateur:

=> USA + fort potentiel client

Analyse

Classement par évolution

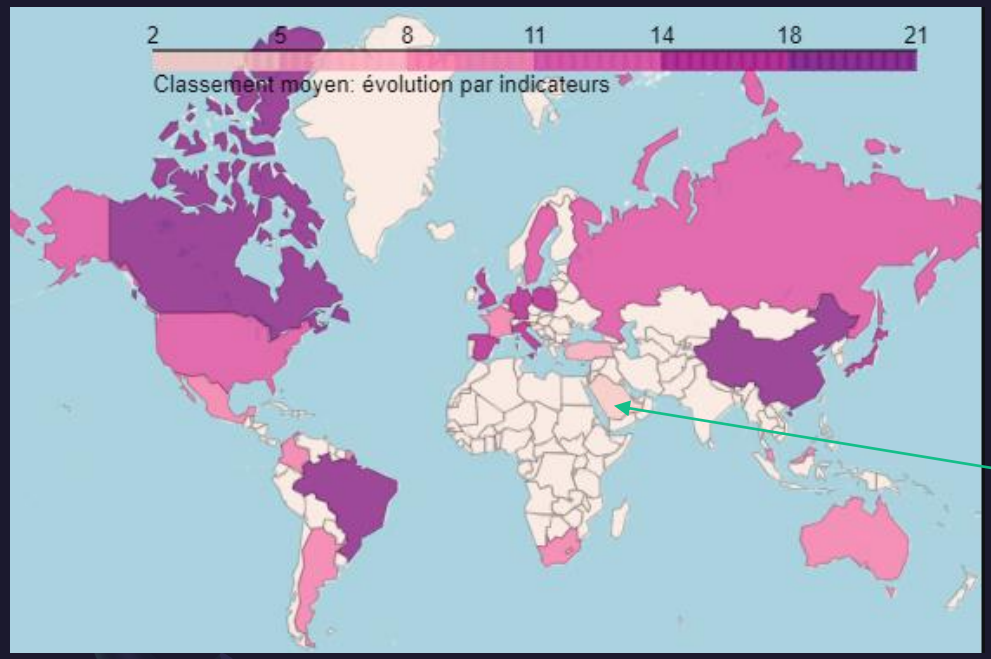


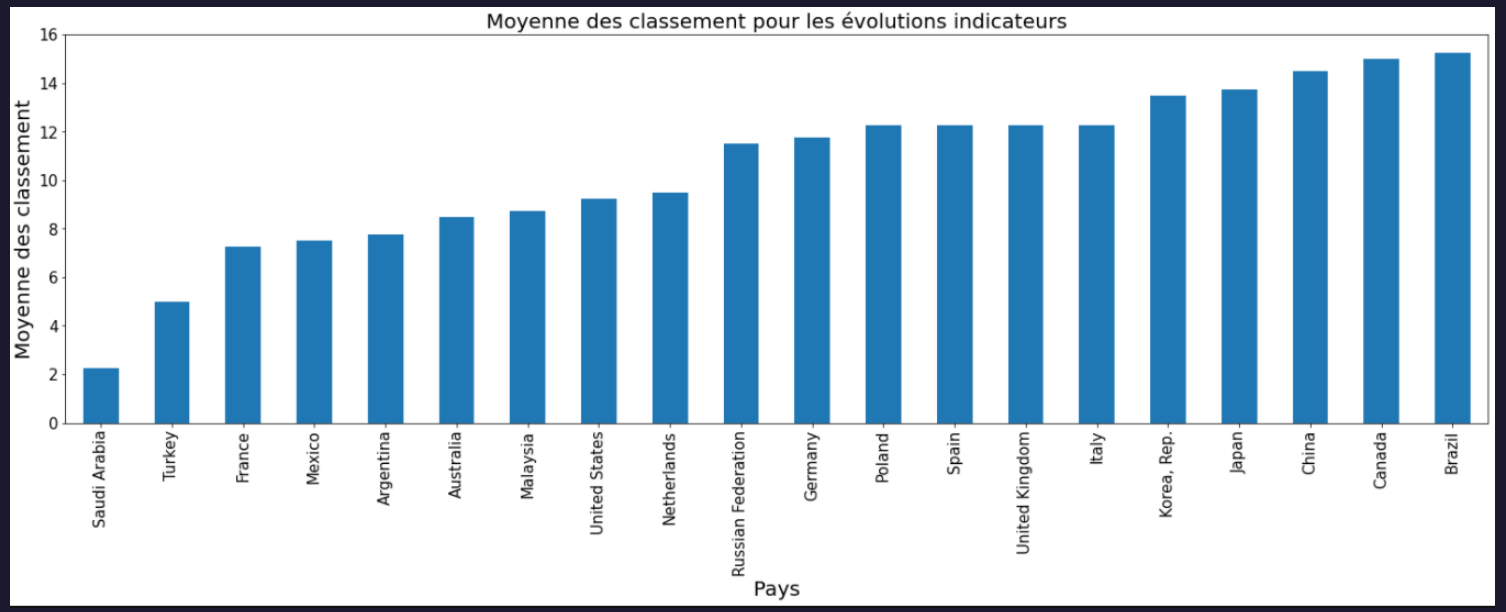


On associe l'évolution à la pente
(tendance de l'indicateur)

Analyse

Classement par évolution





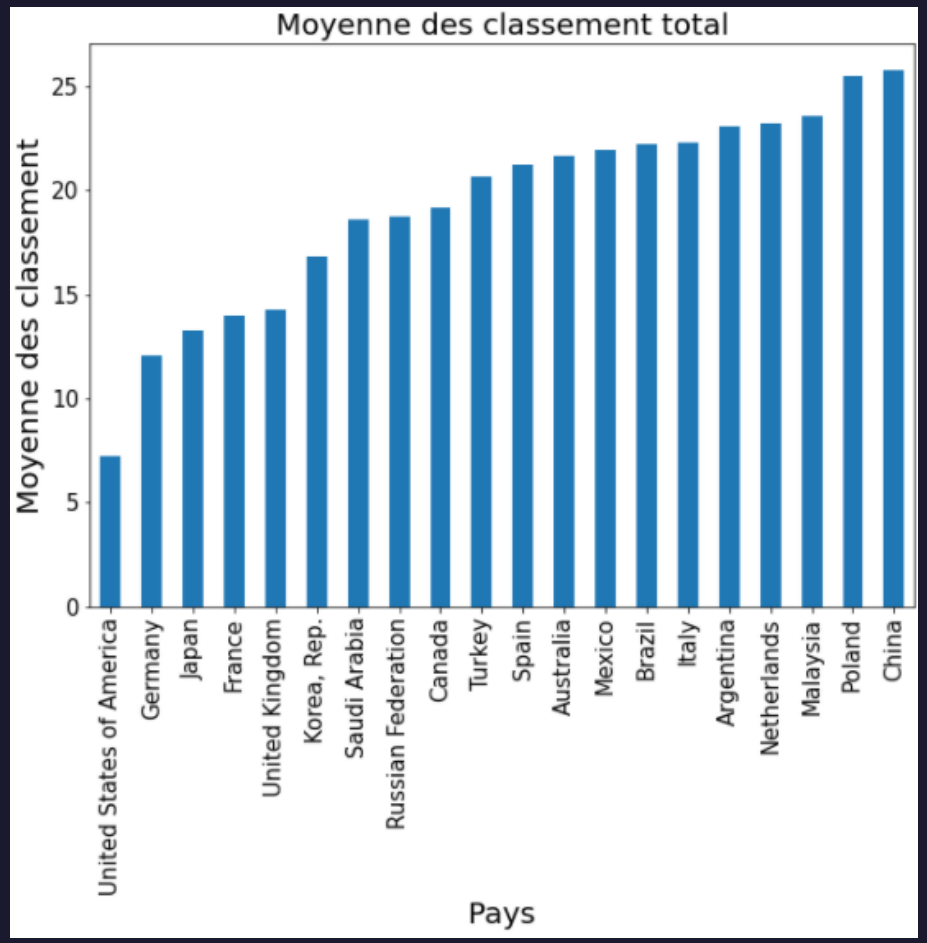
- ***Pour chacun de ces pays, quelle sera l'évolution de ce potentiel clients ?***

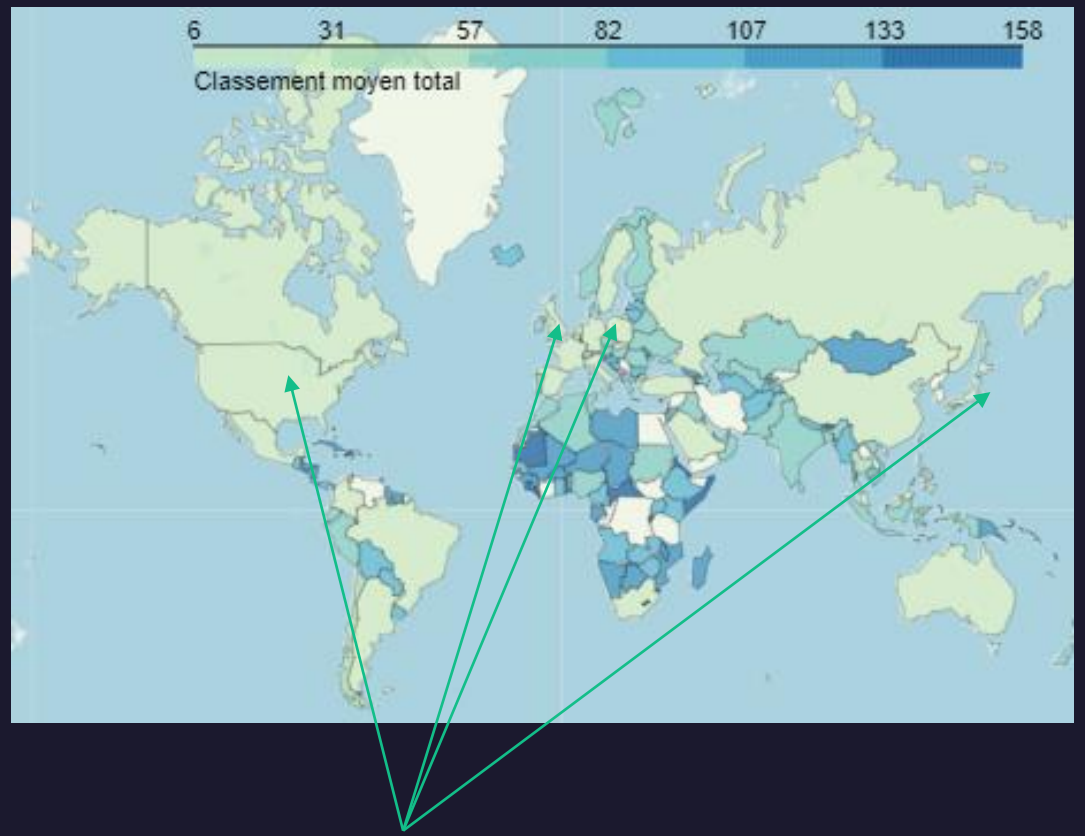
Moyenne des classements pour chaque évolution d'indicateur:

- Arabie Saoudite + fort potentiel client
- France présente

Conclusion:

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





Pays conseillés pour l'international:

⇒ USA

⇒ Allemagne

⇒ Japon

⇒ Royaume Uni

Merci!