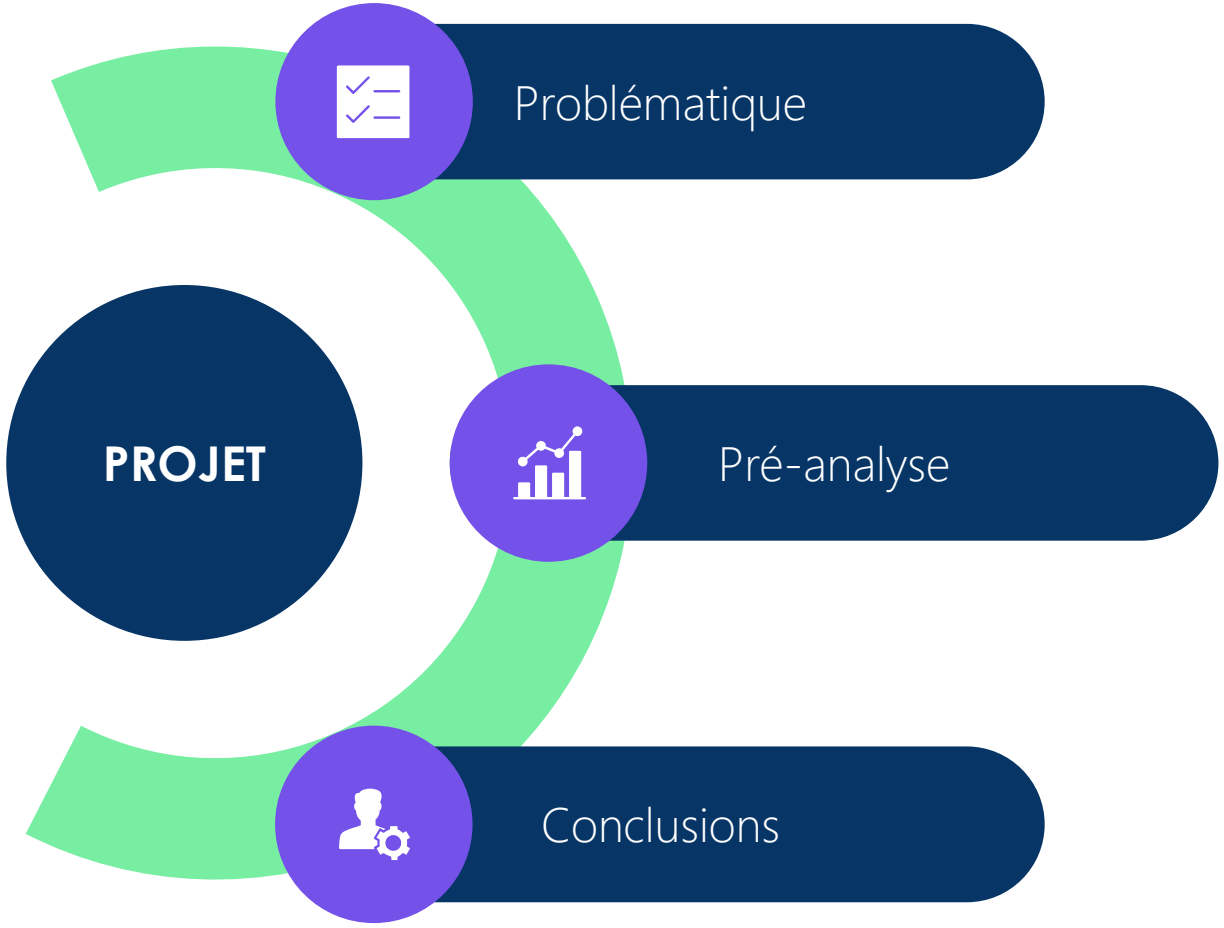
Projet N°2 :
Analysez des données de systèmes éducatifs

Agustin Bunader (autofinancé)
Soutenance de Projet
Octobre 2020

Programme



Problématique

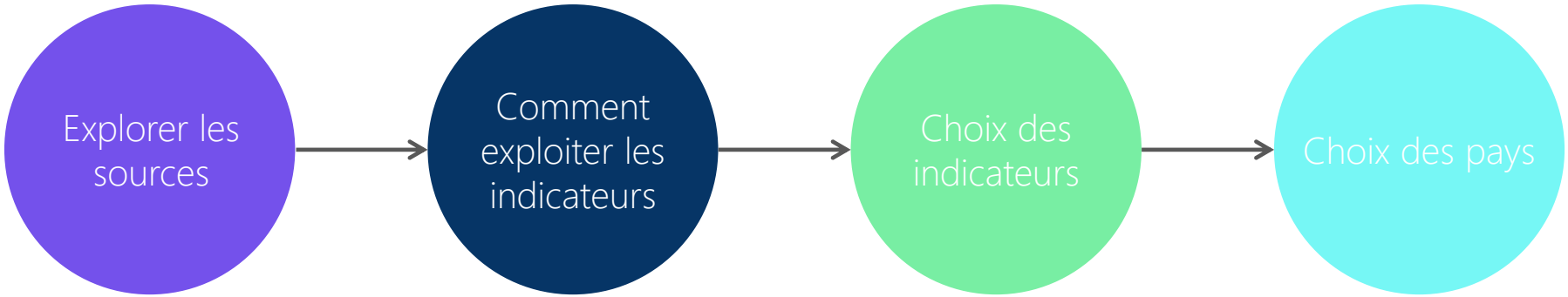
Academy est une start-up EdTech qui propose des contenus de formation en ligne pour un public de niveau lycée et universitaire.

Projet d'expansion à l'international de l'entreprise : une première mission d'analyse exploratoire pour déterminer si les données sur l'éducation de la *Banque Mondiale* permettent d'informer le projet d'expansion.

Pré-analyse exploratoire :

- Valider la qualité de ce jeu de données
- Décrire les informations contenues dans le jeu de données
- Sélectionner les informations qui semblent pertinentes pour répondre à la problématique
- Déterminer des ordres de grandeurs des indicateurs statistiques classiques pour les différentes zones géographiques et pays du monde

Pré-analyse - Etapes



- Outils utilisés
- Comment sont composés les fichiers sources ?
- Données manquantes ?

- Distribution par année
- Distribution par pays
- Taux de remplissage

- Quels indicateurs choisir ?
- Quels années choisir?

- Quels sont les pays à analyser ?
- Comment identifier leur potentiel ?

Pré-analyse – Outils utilisés

Nom	Description	Fonctions
Anaconda	Distribution libre et open source de Python appliqué aux sciences des données	Gestion des paquetages avec le système « conda »
Jupyter Notebook 6.1.1	Application web utilisée pour explorer et analyser des données en Python	Structurer la pré-analyse en exécutant le code par étape et en le commentant (avec des Markdowns)
Python 3.7	Langage de programmation interprété, multiparadigme et multiplateformes	Boucles (for), listes, dictionnaires et définitions (def)
Pandas 1.1.1	Structures de données et outils d'analyse de données performants et faciles à utiliser	Dataframe : création, filtrage, triage, copie, description, pivot table, comparaison
Matplotlib 3.3.1	Bibliothèque destinée à tracer et visualiser des données sous formes de graphiques	Améliorer la visualisation des graphiques générés avec Seaborn
Seaborn 0.11.0	Bibliothèque de visualisation des données basée sur Matplotlib	Barplot, scatterplot, lineplot et heatmap graphiques
Collections (module de Python)	Le module implémente des types de données de conteneurs spécialisés	Counter : sous-classe qui permet le dénombrement d'objets hachables.
Scipy 1.5.0	Ensemble des bibliothèques à usage scientifique	Linregress : méthode des moindres carrés
Numpy 1.19.1	Destinée à manipuler des matrices ou tableaux ainsi que des fonctions mathématiques	

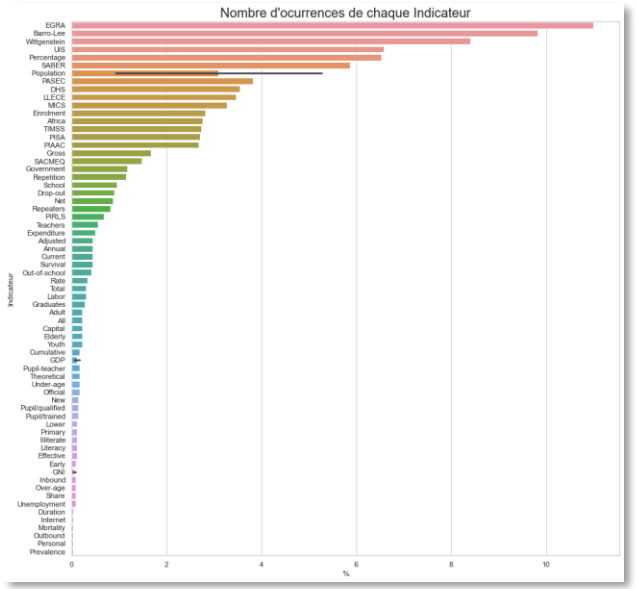
Pré-analyse – Jeu de données

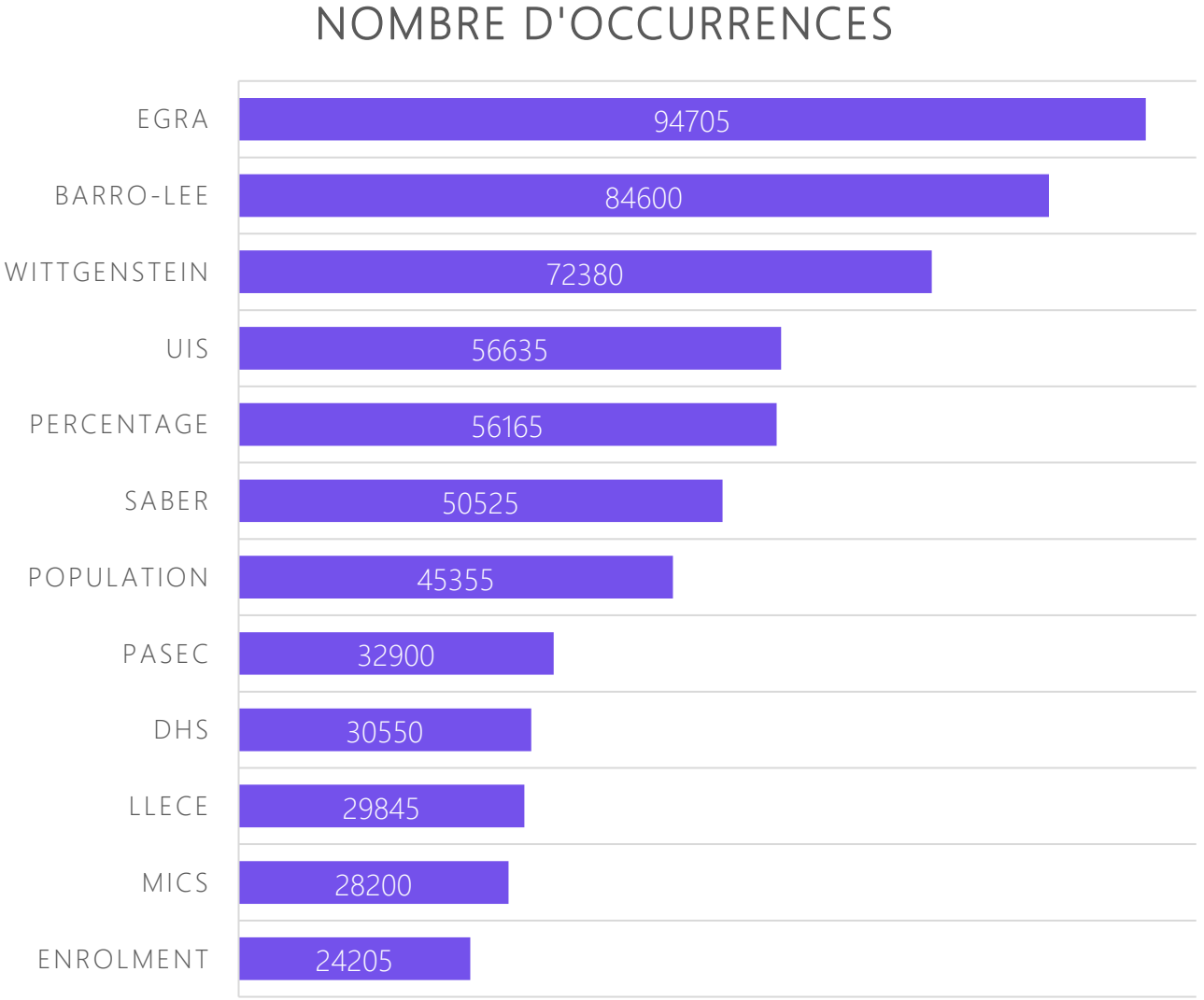
EdStatsCountry.csv	Indicateurs sur l'économie de chaque pays et zone géographique. Taille : 241 lignes (1 par pays / zone) et 32 colonnes (principalement des indicateurs) Certaines valeurs sont manquantes. Aucun doublon.
EdStatsCountry-Series.csv	Informations sur la source des données contenues dans EdStatsCountry.csv Taille : 613 lignes et 4 colonnes Seulement NaN sur la colonne « Unnamed: 3 » Aucun doublon.
EdStatsData.csv	Données de nombreux indicateurs pour tous les pays / zones depuis 1970 Taille : 886930 lignes et 70 colonnes (principalement des années) Quantité importante de données manquantes. Colonne « Unnamed: 69 » composée de NaN. Aucun doublon.
EdStatsFootNote.csv	Petite description, année par année, sur chaque indicateur. Taille : 643638 et 5 colonnes Seulement NaN sur la colonne « Unnamed: 4 » Aucun doublon.
EdStatsSeries.csv	Détails descriptifs des indicateurs présents dans EdStatsData.csv Taille : 3665 lignes (1 par indicateur) et 21 colonnes Quantité importante de données manquantes. 6 colonnes composées de NaN. Aucun doublon.

Pré-analyse – Exploiter les indicateurs

Résumé :

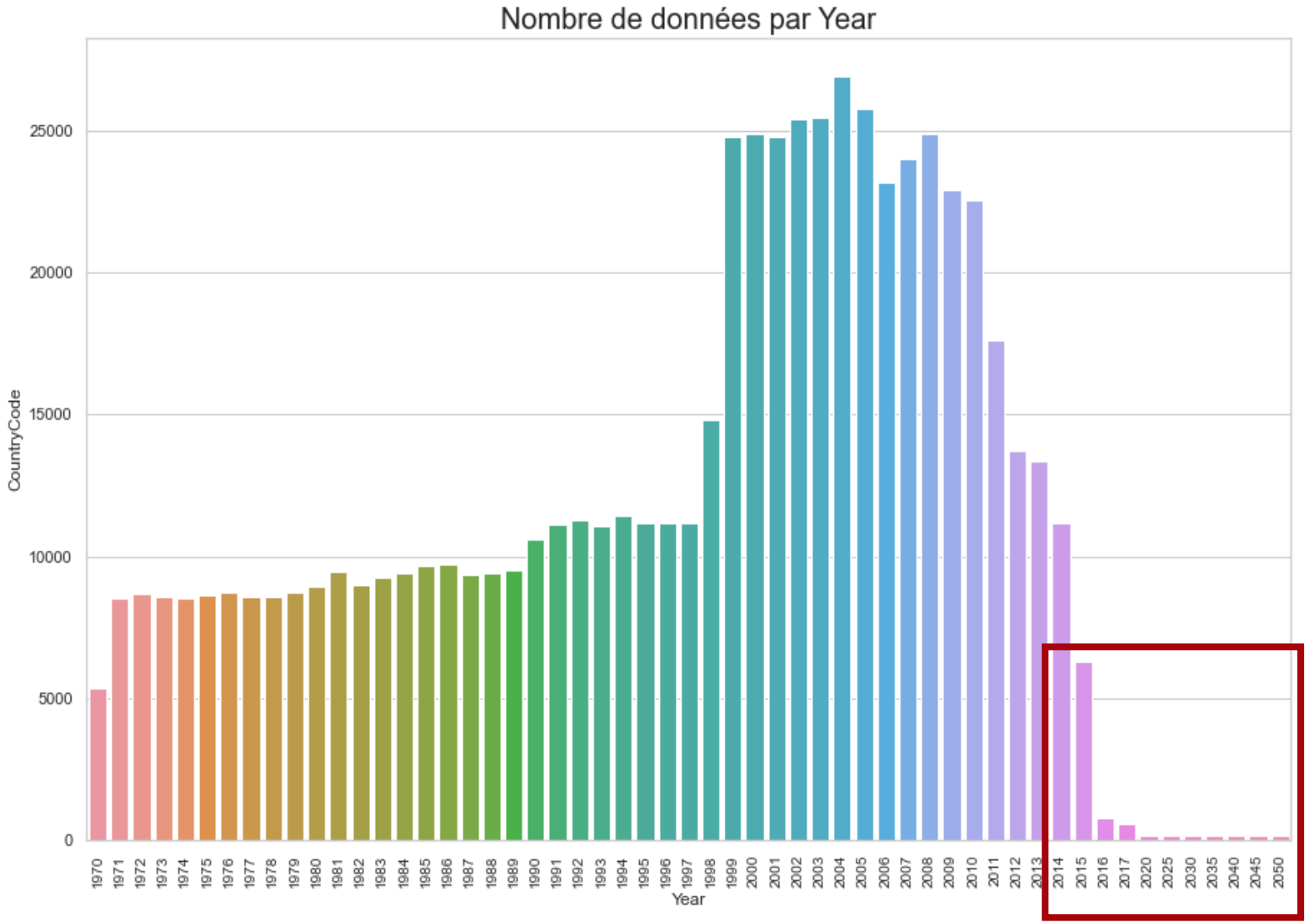
- 3665 indicateurs uniques
- 241 pays / zones géographiques
- Données historiques et prédictions (1970 à 2100)
- Indicateurs principalement relatifs à l'éducation





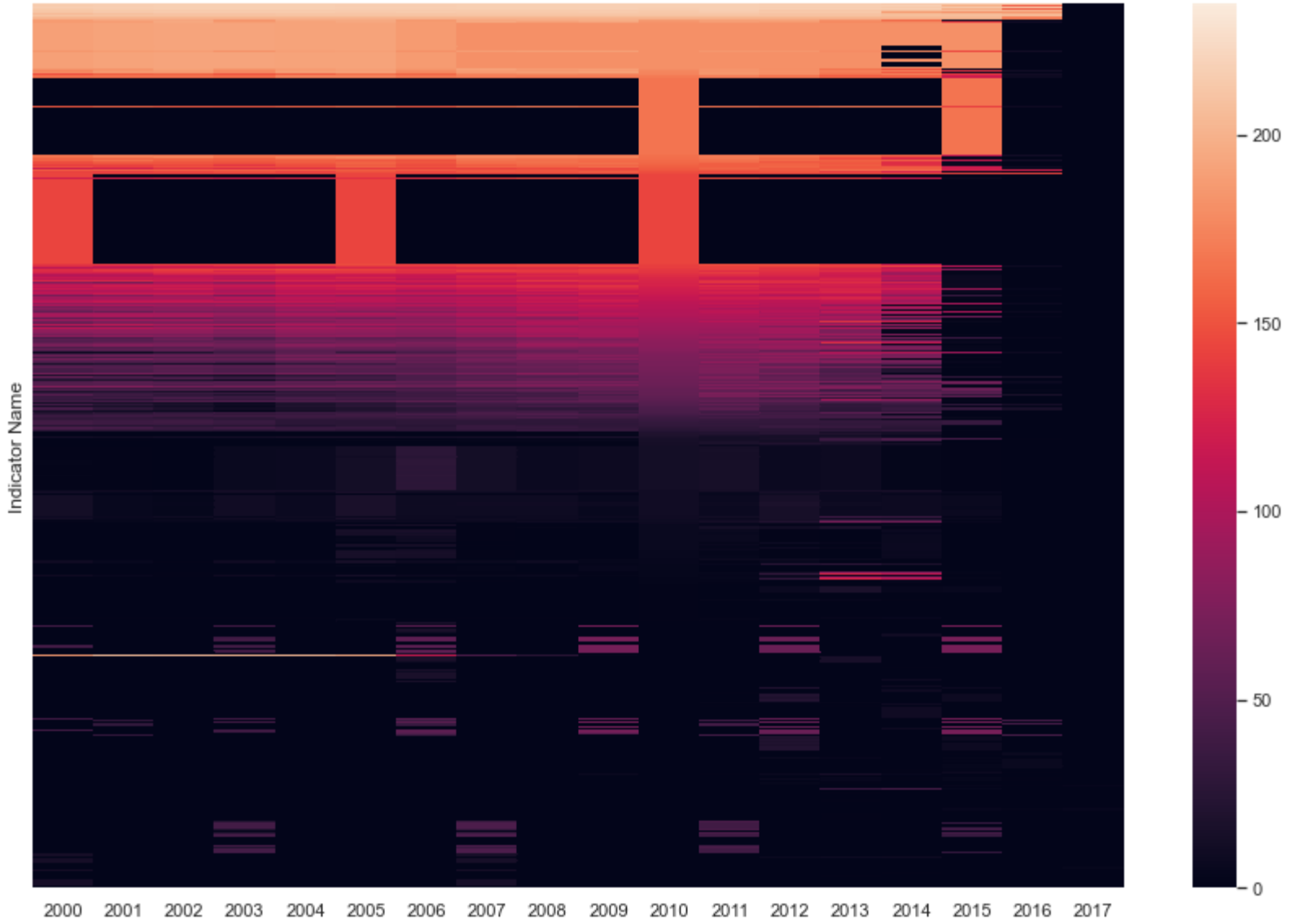
Pré-analyse – Exploiter les indicateurs

- A partir de 2015 la quantité des données disponibles par année diminue fortement
- Rendre le traitement des données plus gérable par décennie



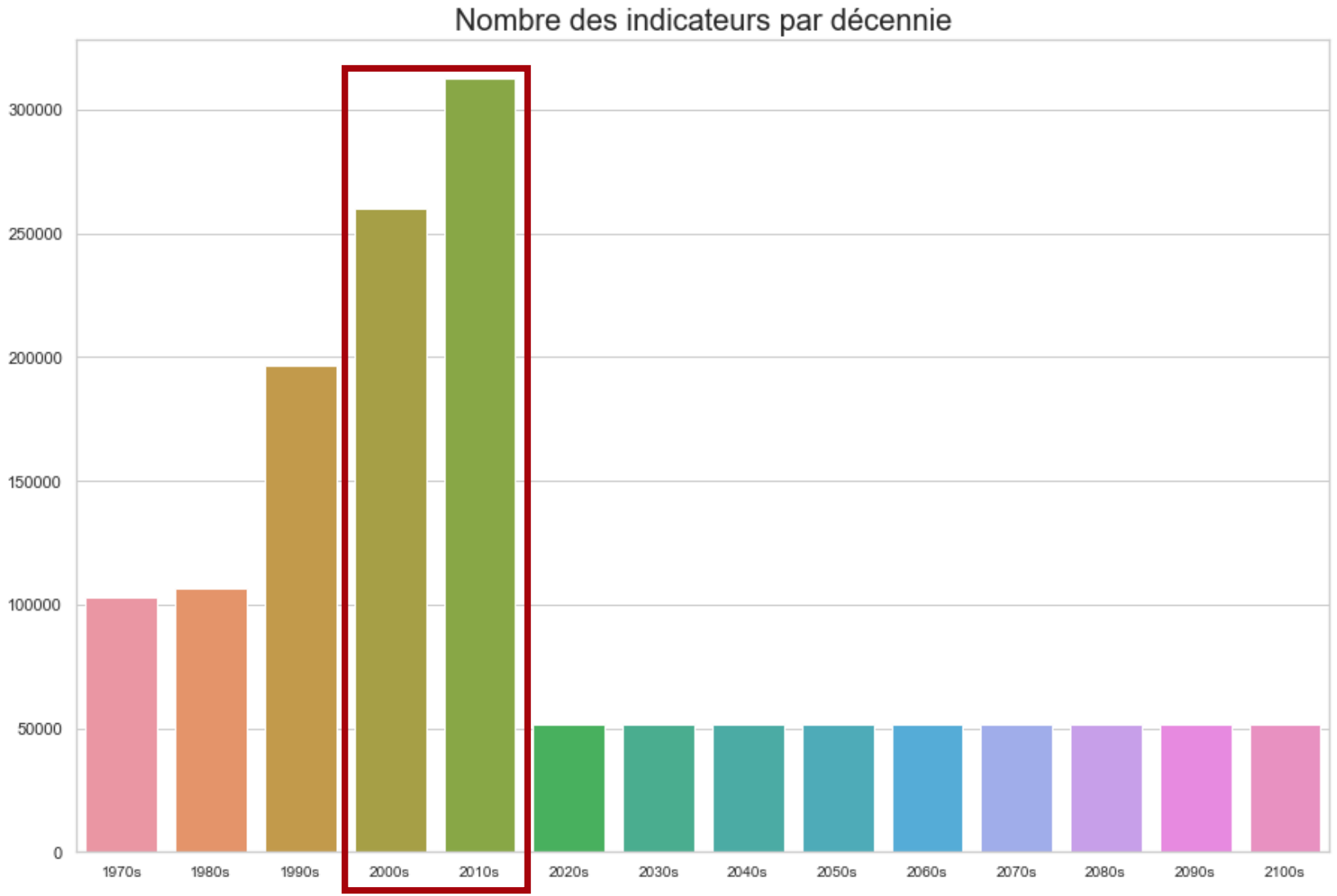
Pré-analyse – Exploiter les indicateurs

- A partir de 2015 la quantité des données disponibles par année diminue fortement
- Rendre le traitement des données plus gérable par décennie



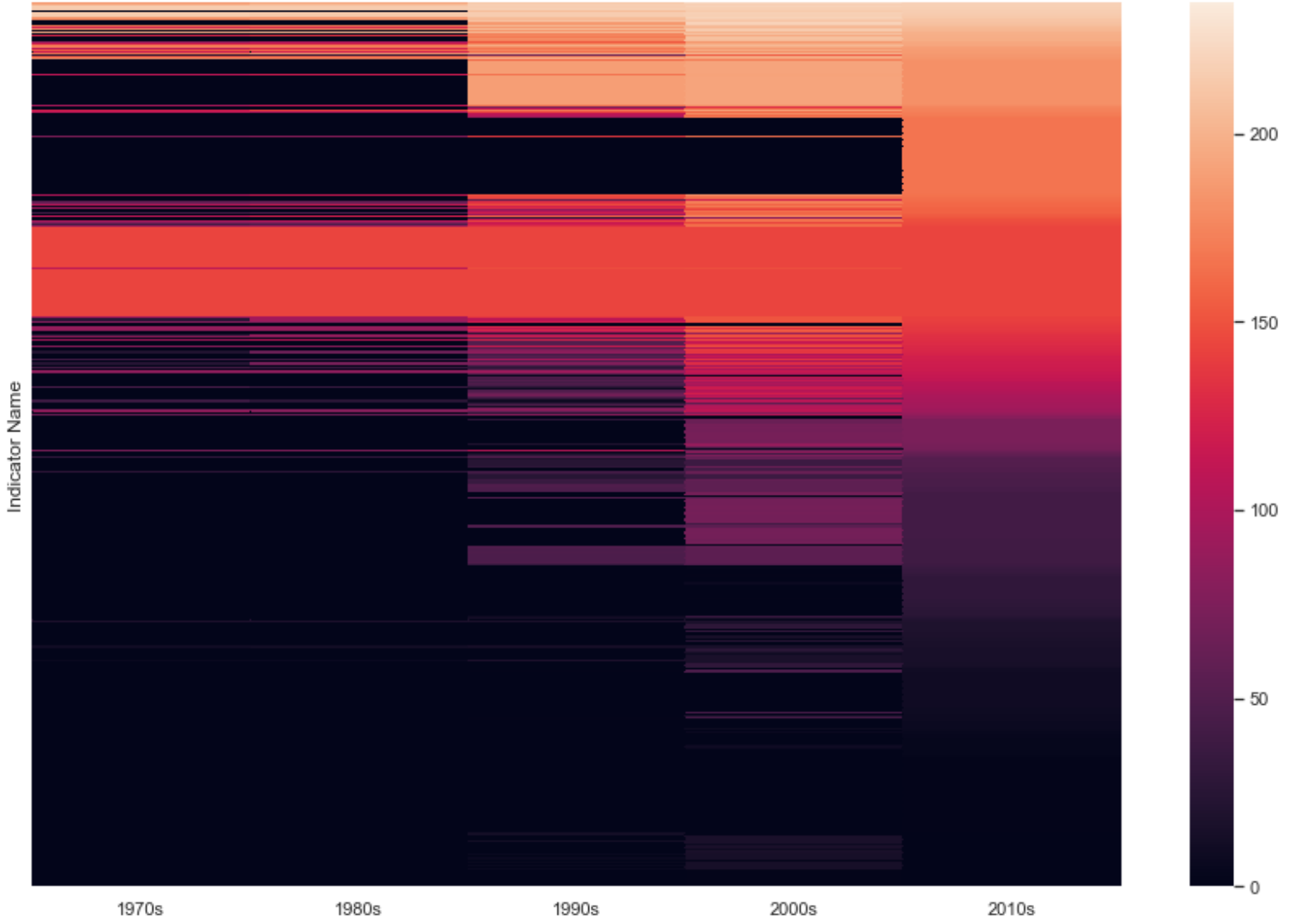
Pré-analyse – Exploiter les indicateurs

- Les décennies de 2000 à 2010 sont celles qui contiennent le plus grand nombre de données



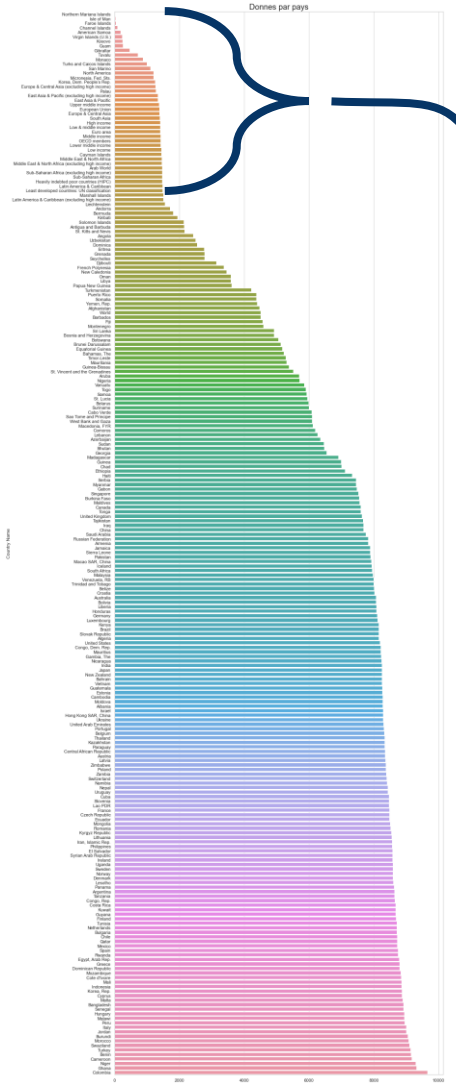
Pré-analyse – Exploiter les indicateurs

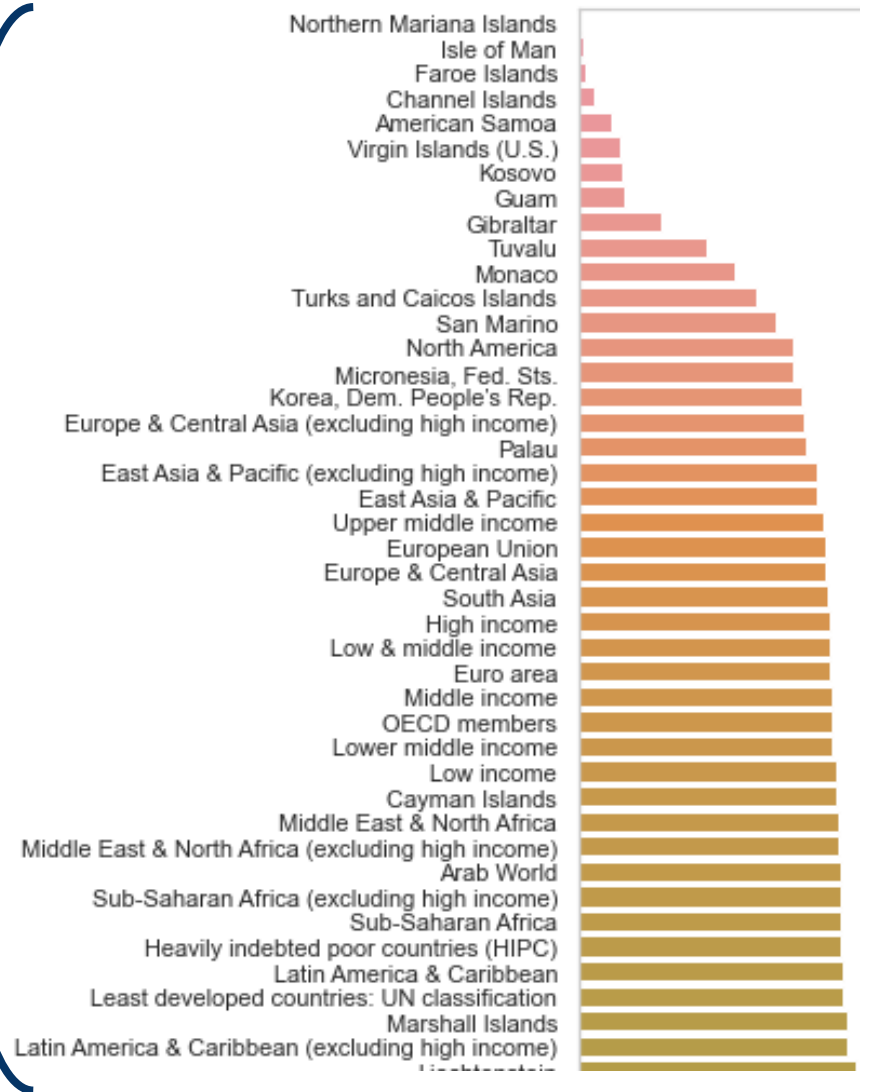
- Les décennies de 2000 à 2010 sont celles qui contiennent le plus grand nombre de données



Pré-analyse – Exploiter les indicateurs

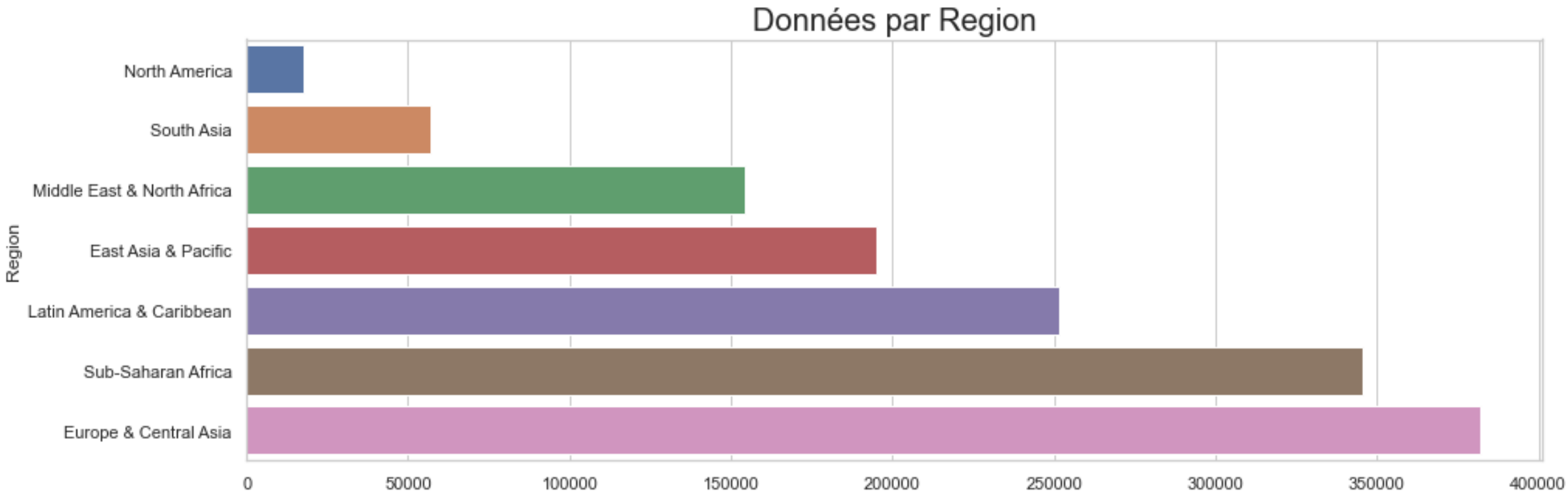
- 235 pays/régions sont choisis une fois consolidés les éléments dans les colonnes « Country Name », « Country Code » et « Countrycode » contenues dans les fichiers source
- 235 dont 35 éléments sont des régions composées de pays ou de territoires spéciaux
- 235 dont 200 sont des pays ou territoires spéciaux
- La répartition des données par pays n'est pas uniforme
- Les micro-états, les régions autonomes, les zones des pays et « nouveaux pays » sont ceux qui ont le moins de données





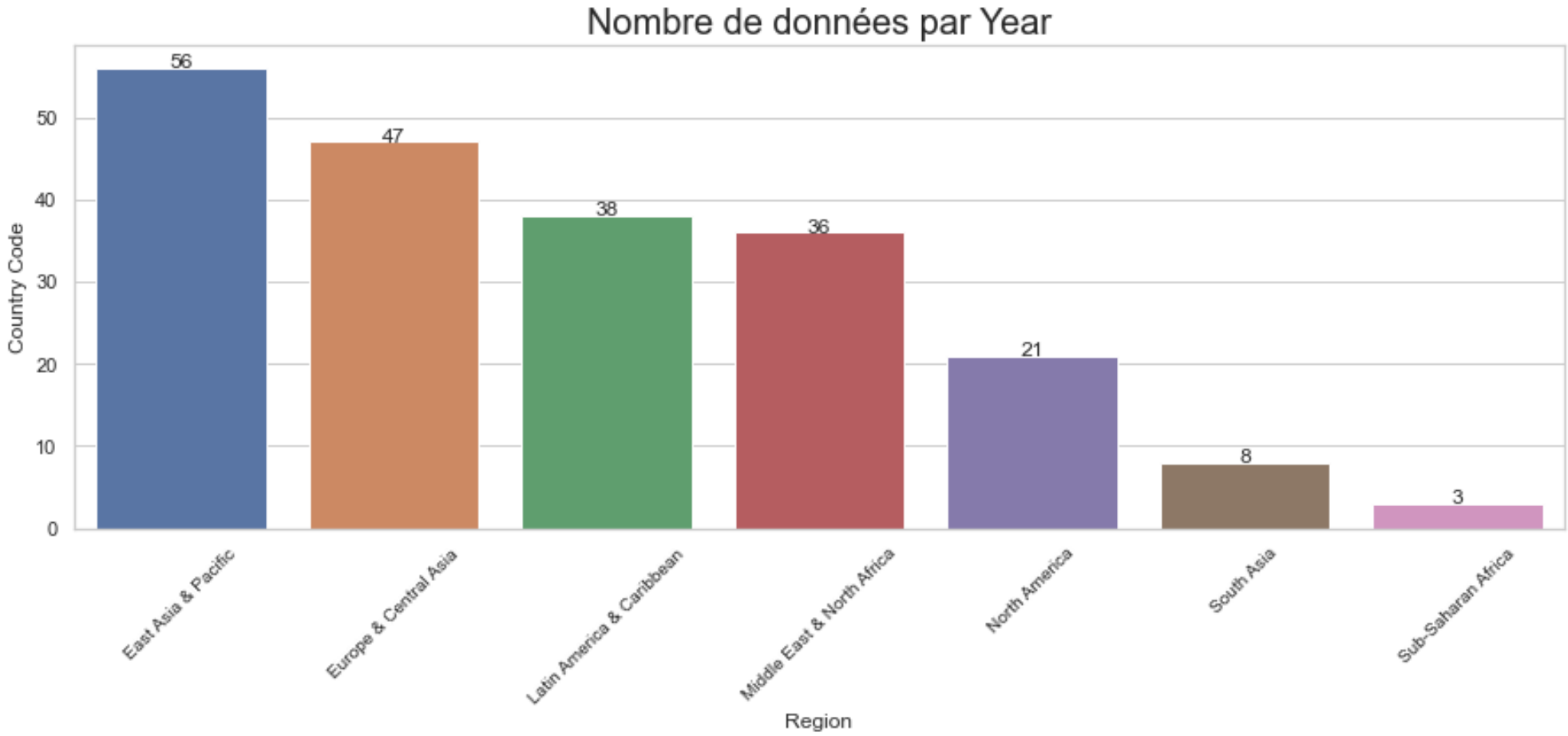
Pré-analyse – Exploiter les indicateurs

- La répartition des données par région est définie par le nombre de pays dans chacune



Pré-analyse – Exploiter les indicateurs

- La répartition des données par région par année montre une situation différente à celle montrée par les indicateurs par pays

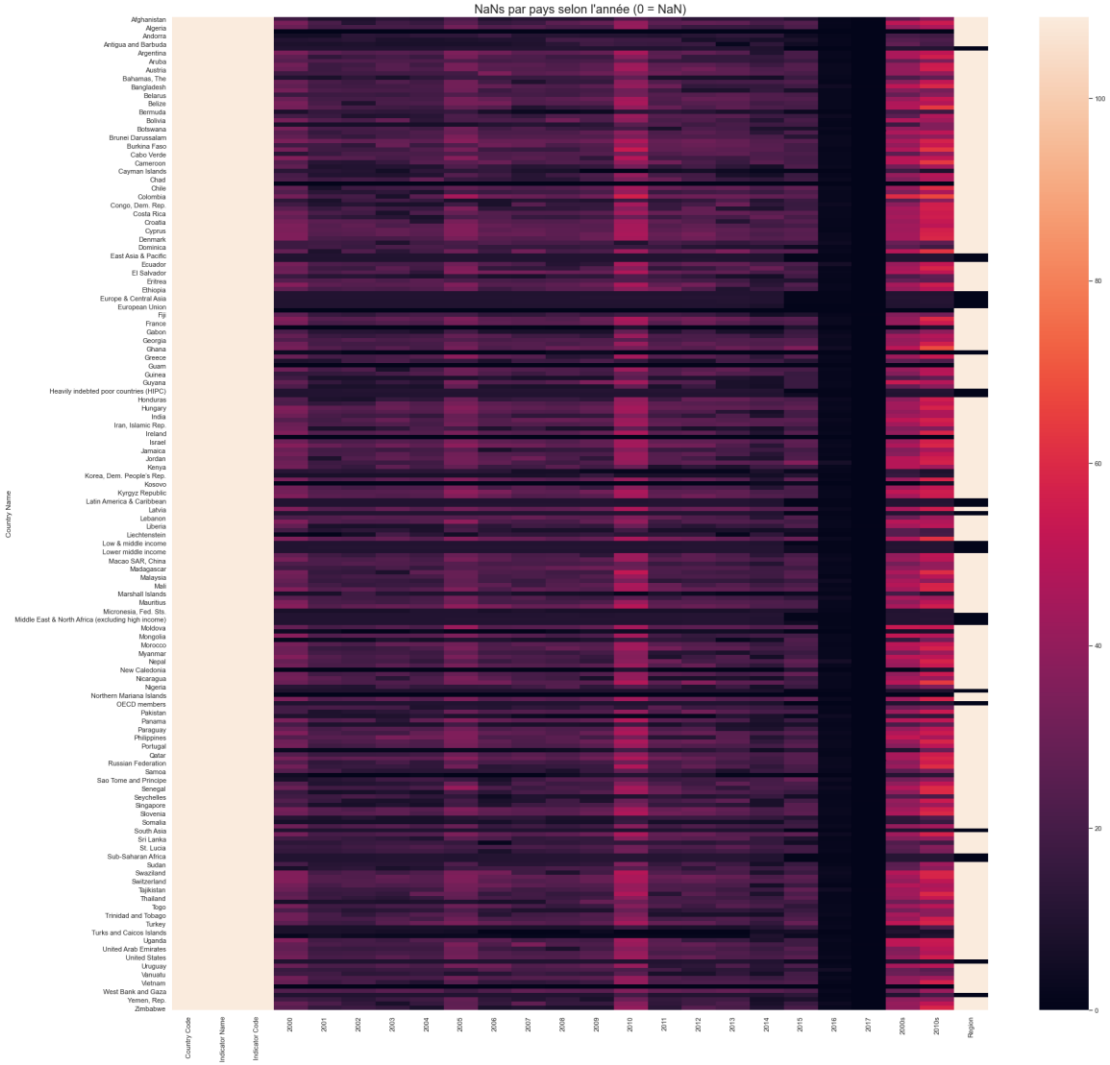


Pré-analyse – Exploiter les indicateurs

- Taux de remplissage des pays par année entre le 2000 et 2017

$$tr = \frac{F\ de\ chaque\ pays\ par\ année}{3665\ indicateus\ uniques} \cdot 100$$

- Tel que signalé précédemment, certains pays contiennent plus de données que d'autres
- La même situation s'applique aux régions

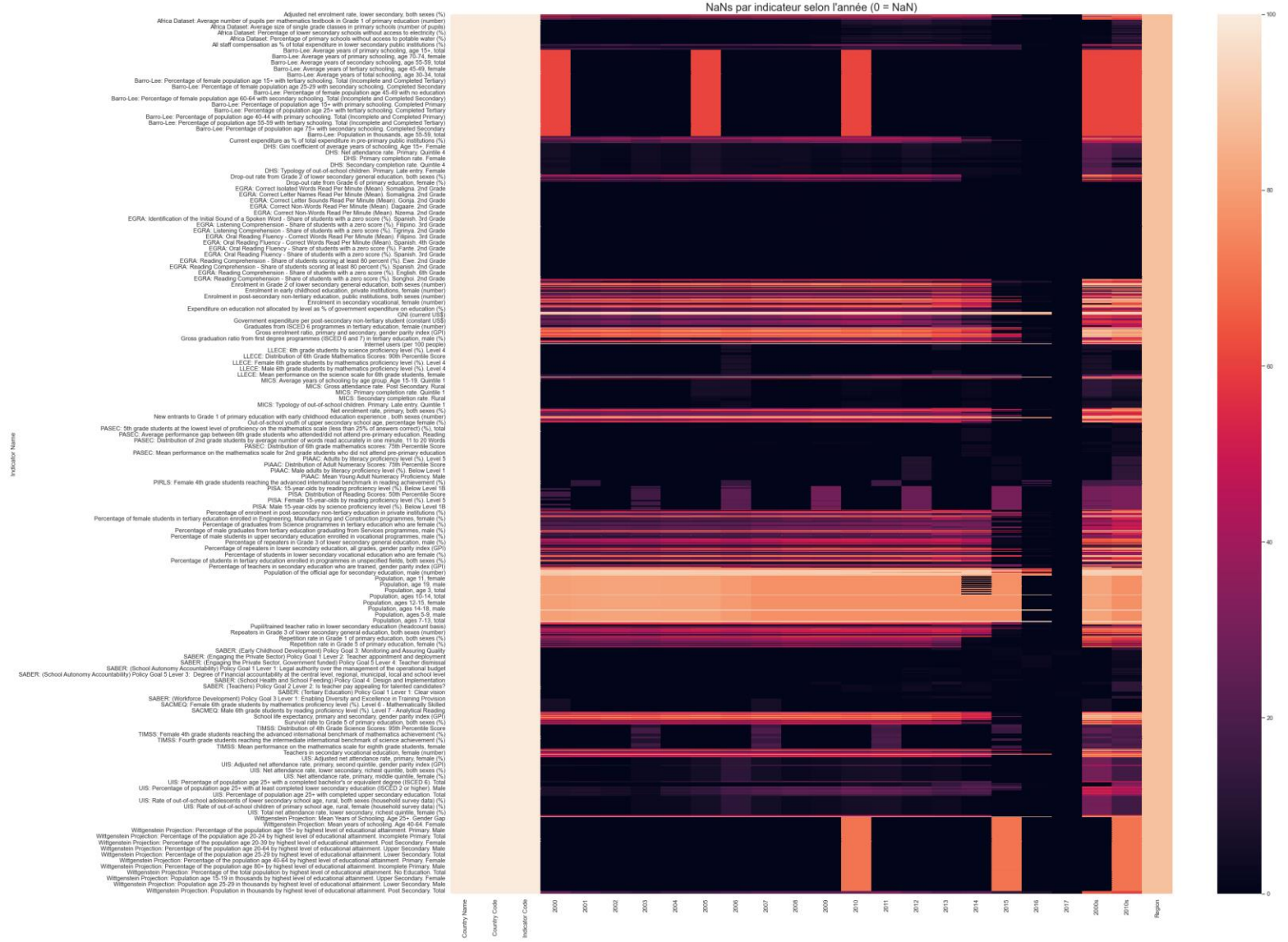


Pré-analyse – Exploiter les indicateurs

- Taux de remplissage des indicateurs par année entre le 2000 et 2017

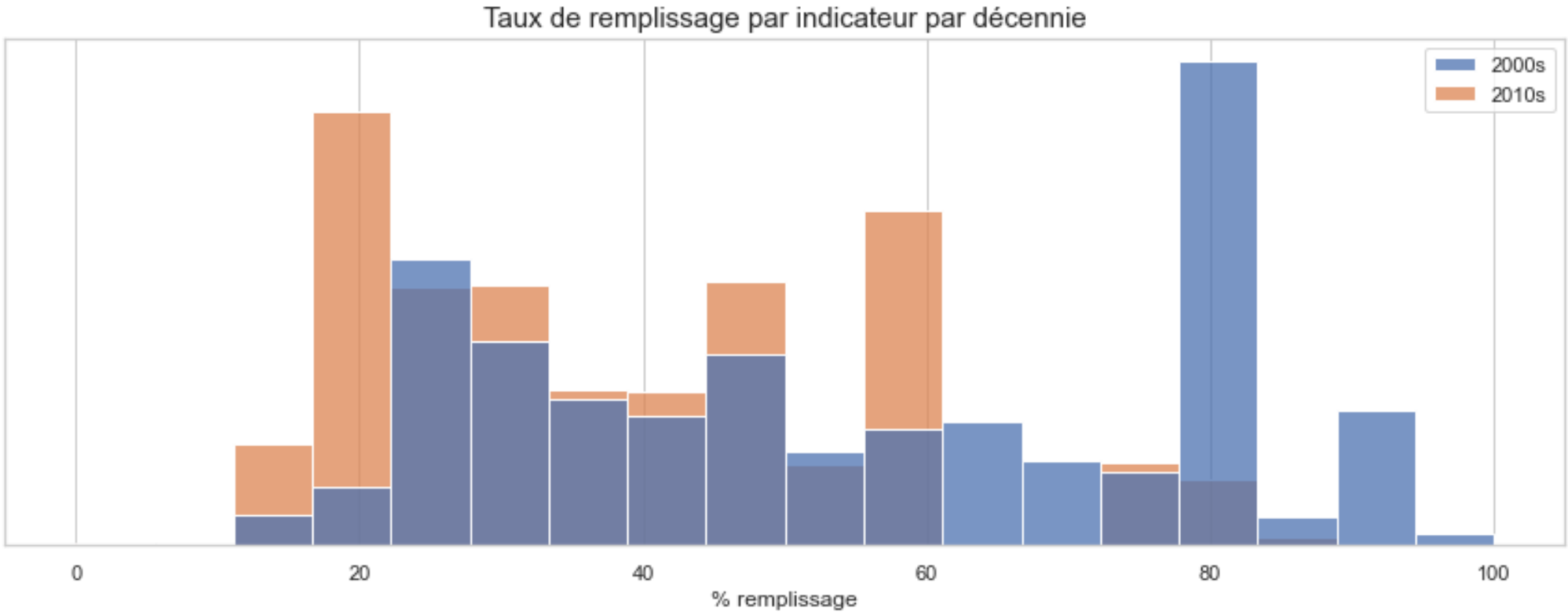
$$tr = \frac{F\ de\ chaque\ indicateur\ par\ année}{Q\ des\ pays\ ou\ régions} \cdot 100$$

- Les indicateurs à conserver doivent avoir un taux de remplissage supérieur à 20%
- 930 indicateurs à conserver (proportion conservée de 25%)



Pré-analyse – Choix des indicateurs

- Taux de remplissage des 930 indicateurs choisis par décennie 2000 et 2010



Pré-analyse – Choix des indicateurs

- Taux de remplissage des 930 indicateurs choisis par décennie 2000 et 2010
- Préfix des indicateurs les plus utilisés

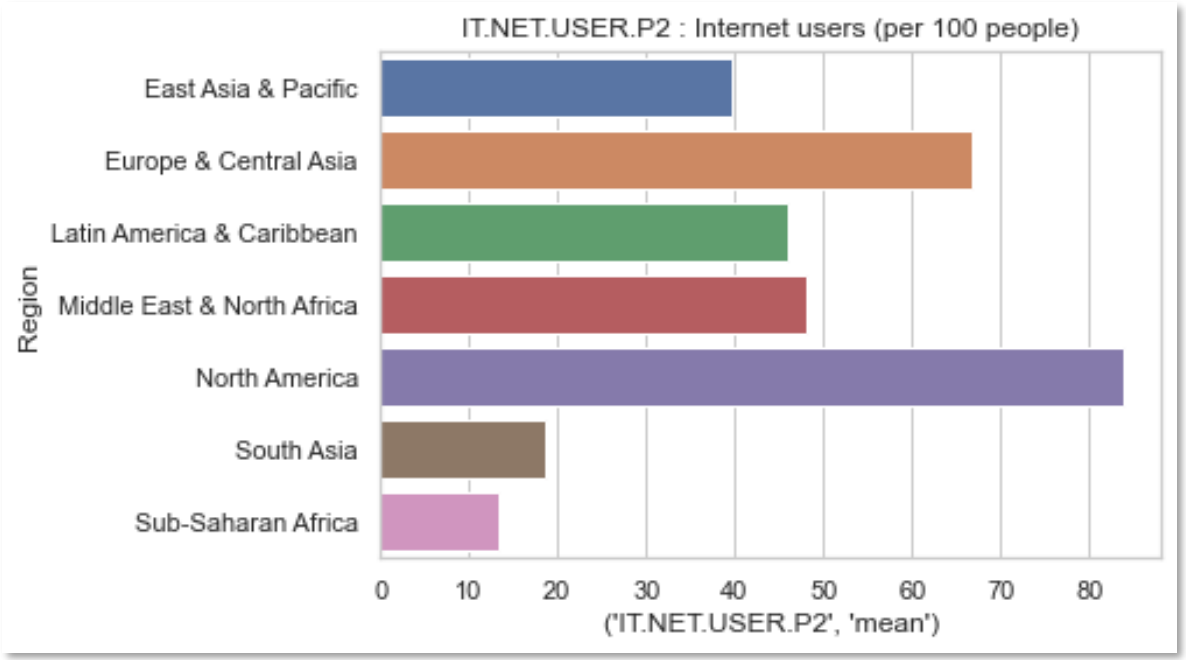
```
[('UIS', 122435),
 ('SP', 50525),
 ('SE', 38070),
 ('SL', 3290),
 ('NY', 2820),
 ('XGDP', 470),
 ('IT', 470),
 ('SH', 470)]
```

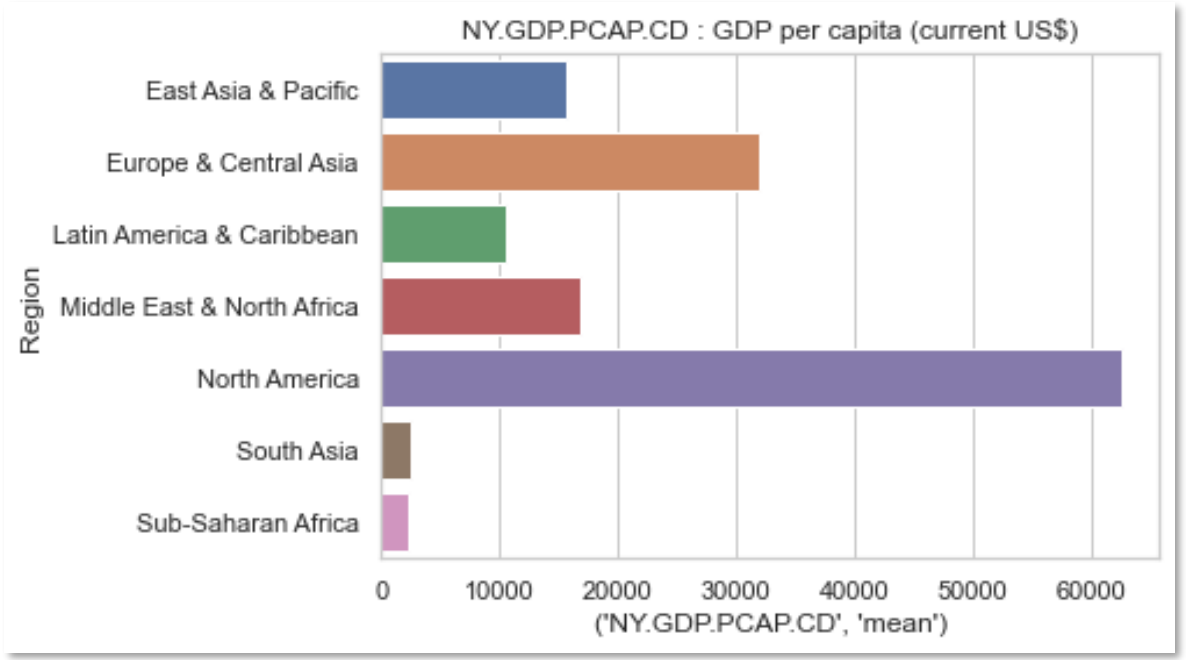
Indicateurs choisis	Description
UIS.E.5.B	Enrolment in tertiary education, ISCED 5 programs, both sexes (number)
UIS.E.4	Enrolment in post-secondary non-tertiary education, both sexes (number)
UIS.E.3	Enrolment in upper secondary education, both sexes (number)
SE.TER.ENRL	Enrolment in tertiary education, all programs, both sexes (number)
SP.POP.GROW	Population growth (annual %)
SP.POP.1524.TO.UN	Population, ages 15-24, total
SE.XPD.TOTL.GB.ZS	Expenditure on education as % of total government expenditure (%)
IT.NET.USER.P2	Internet users (per 100 people)
NY.GDP.PCAP.CD	GDP per capita (current USD)
NY.GDP.PCAP.PP.CD	GDP per capita, PPP (current international USD)

Pré-analyse – Choix des indicateurs

Pré-analyse – Choix des pays

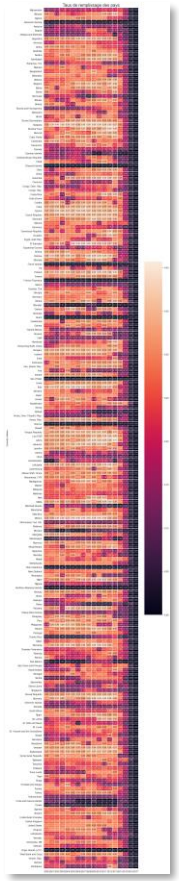
Exemples des ordres de grandeur (moyenne) par région pour la décennie de 2010s utilisés lors de l'analyse

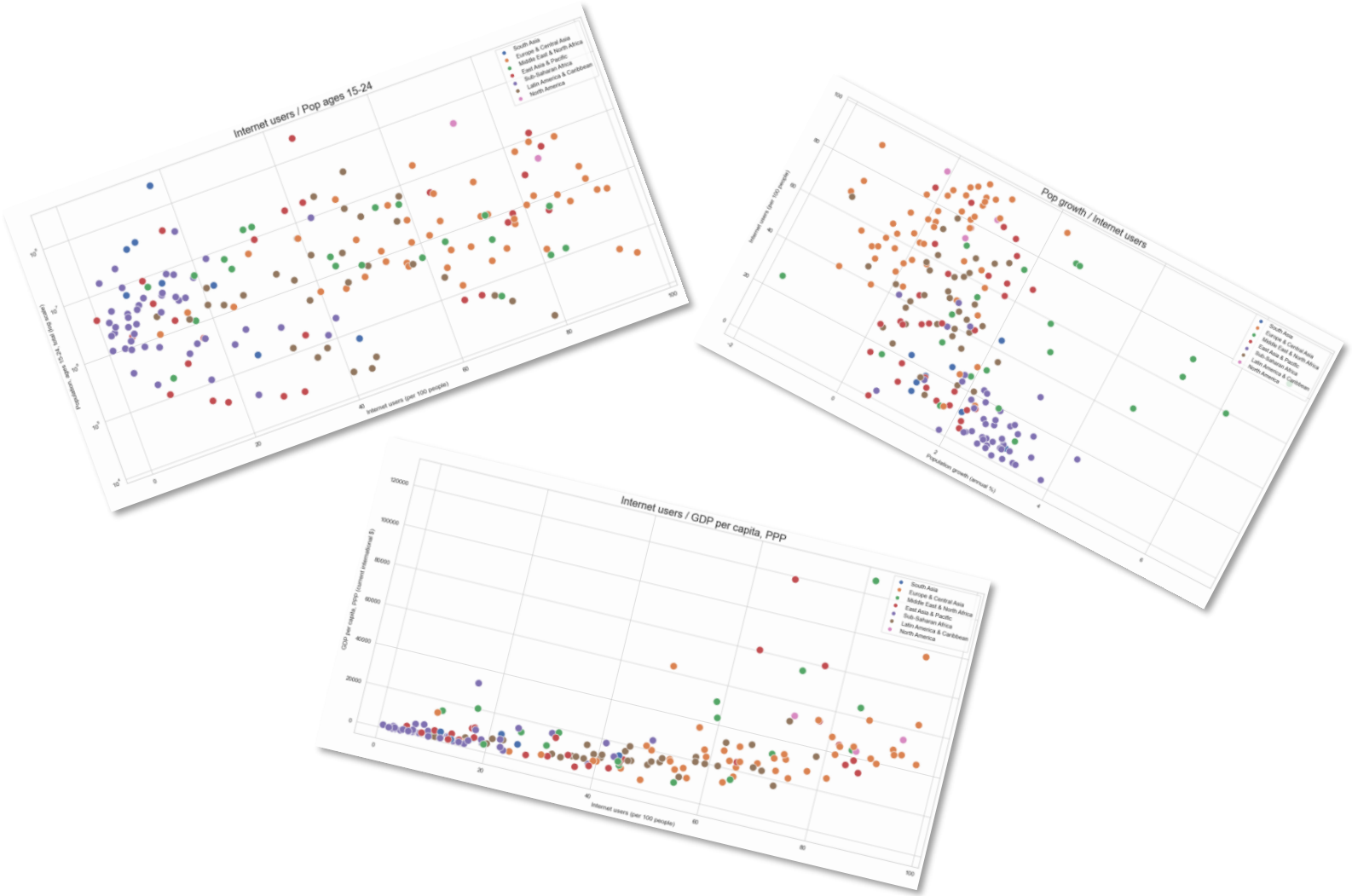




Pré-analyse – Choix des pays

- Analyse intuitive des données avec des scatterplots de la quantité des utilisateurs d'internet





Pré-analyse – Choix des pays

Les pays ont été choisis selon les paramètres suivants :

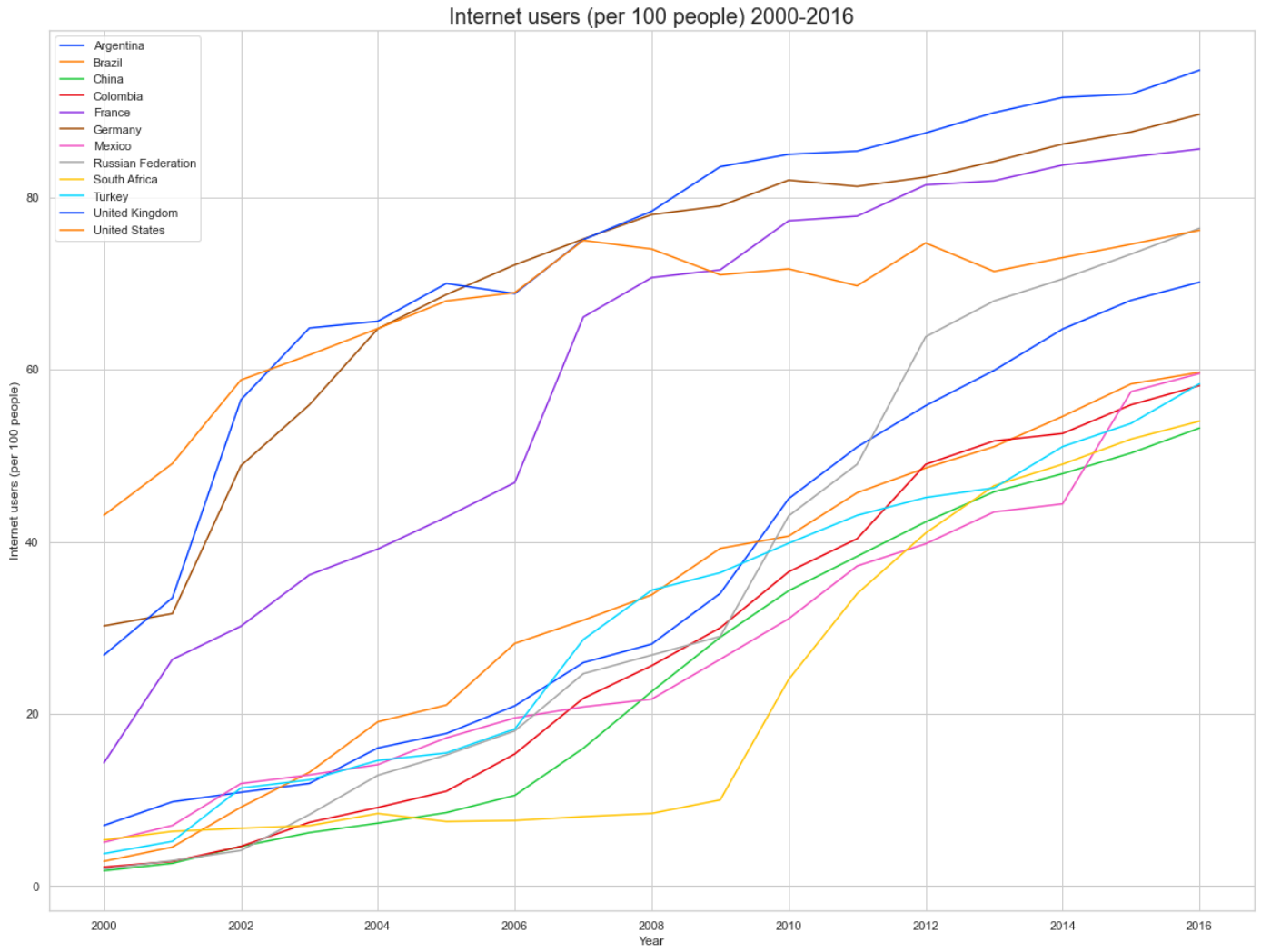
- Internautes (pour 100 personnes) : Supérieurs à la médiane mondiale
- Croissance démographique (% annuel) : Positive (supérieure à 0.1%)
- Population âgée de 15 à 24 ans : Supérieure à la moyenne mondiale
- PIB per capita à PPA : Supérieur à la médiane mondiale de 12229 USD

	Country Name	NY.GDP.PCAP.PP.CD
0	United States	53015.345
1	Germany	45000.383
2	United Kingdom	39396.005
3	France	39069.030
4	Russian Federation	23958.898
5	Turkey	21665.777
6	Argentina	19694.726
7	Mexico	16618.882
8	Brazil	15367.941
9	South Africa	12728.272
10	Colombia	12627.248
11	China	12416.258

Pré-analyse – Choix des pays

Il est possible de mesurer le potentiel d'un pays avec l'augmentation des utilisateurs d'internet.

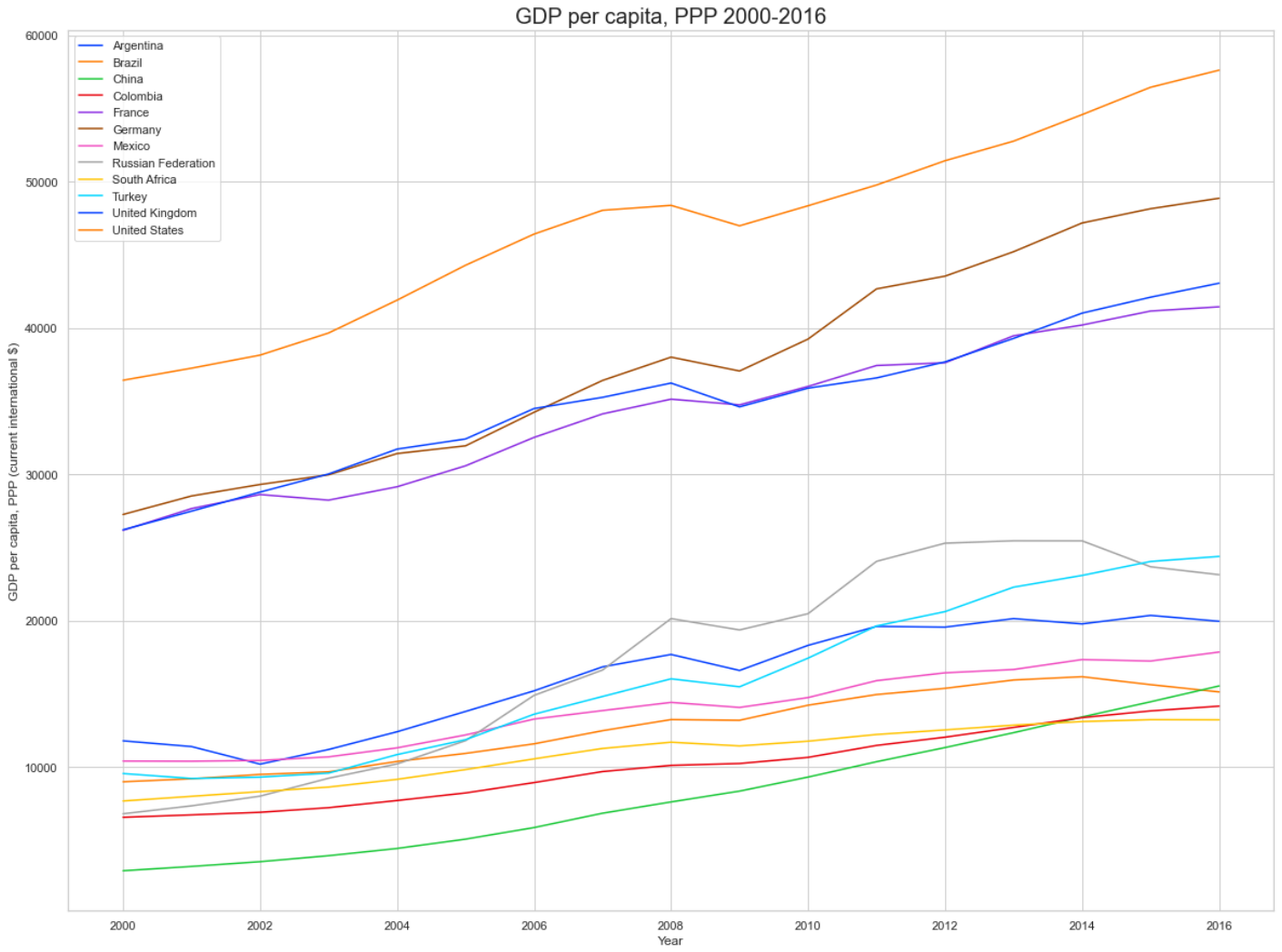
On peut supposer que dans tous les pays le nombre d'internautes va continuer à augmenter et par conséquent le nombre de clients potentiels également.



Pré-analyse – Choix des pays

Il est possible de mesurer le potentiel d'un pays en regardant si les citoyens de chaque pays ont les ressources pour payer une des options des services proposés.

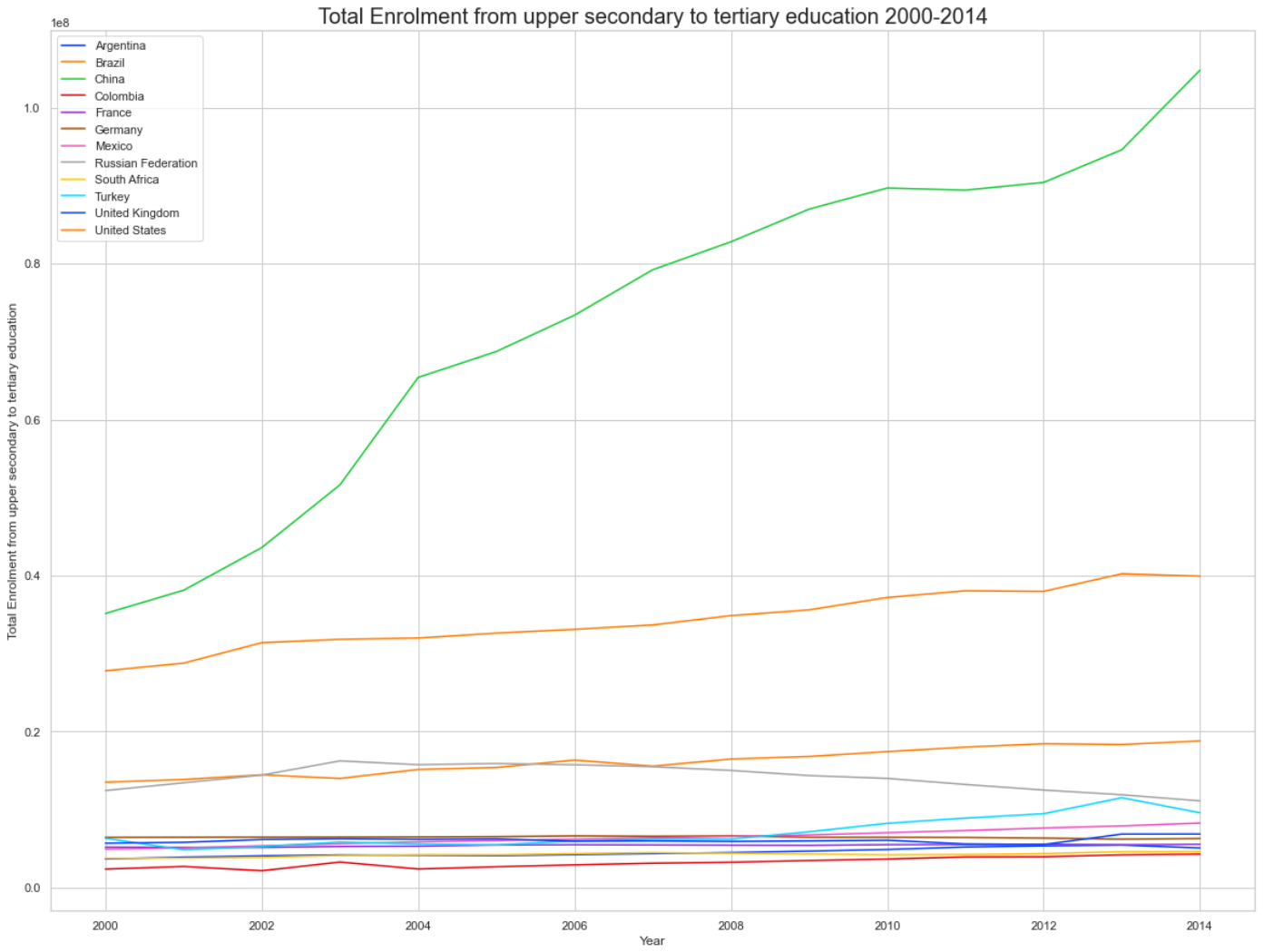
On peut supposer que l'augmentation du PIB per capita à PPA implique une augmentation des ressources disponibles pour la population qui peut potentiellement être investie dans l'éducation.



Pré-analyse – Choix des pays

Il est possible de mesurer le potentiel d'un pays avec la quantité total des inscriptions dans les cycles de détermination à l'enseignement supérieur.

On peut supposer que comme tous les pays choisis ont une croissance démographique positive, ils maintiendront une tendance à la hausse en ce qui concerne le nombre de clients potentiels.



Pré-analyse – Choix des pays

La choix final des pays est fait en priorisant :

- les trois principaux PIB per capita à PPA
- le pays avec la plus grande population potentielle de clients et la plus grande projection de clients potentiels.

1. Etats-Unis

2. Allemagne

3. Royaume-Uni

4. Chine

Conclusion

Sources :

- Tous les pays du monde sont compris
- Données relatives à l'éducation et des autres indicateurs complémentaires
- Détails de chaque indicateur

Limites de la source et de la pré-analyse :

- Des indicateurs avec une forte quantité des données manquantes (inutilisables)
- Il manque des indicateurs relatifs à l'éducation en ligne
- Absence de réalité géopolitique pour réaliser un analyse approfondie des coûts et des avantages de l'expansion dans chacun des pays