Projet 2 :

Analysez des données de systèmes éducatifs

Nom : TRABIS

Prénom : Mohamed

Intitulé de formation : Data Scientist

Encadré par : Mr. Christian NOUMSI

OPENCLASSROOMS

Table des matières

Introduction

Introduction

- **Contexte** :

  Vous êtes Data Scientist dans une **start-up de la EdTech**, nommée ***academy***, qui propose des contenus de formation en ligne pour un public de niveau lycée et université.

- **Mission** :

  Réaliser une analyse des données de la Banque mondiale pour répondre à ces différentes questions :

  → 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 de clients ?

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

Préparation des données

Préparation des données – Présentation générale

- La préparation des données permet d'obtenir les résultats suivants :

  - **Collecte de données :** Dans notre cas la collecte des données commence par le chargement de ces données, puis les affecter à des data frame.

  - **Découvrir et évaluer les données :** Lorsque les données ont été collectées, il est important de découvrir les différents datasets. Cette étape permet de mieux connaître les données et de déterminer le traitement à leur appliquer avant qu'elles deviennent exploitables dans un contexte particulier.

  - **Nettoyer et valider les données :** En général, le nettoyage des données est l'étape la plus longue du processus de préparation des données, mais cette opération est cruciale pour éliminer les données erronées et combler d'éventuelles lacunes. Lors du nettoyage, les tâches importantes sont :
    - ✓ Supprimer les données superflues et les valeurs aberrantes
    - ✓Ajouter les valeurs manquantes
    - ✓Adapter les données à une structure standard
    - ✓Masquer les données privées ou sensibles

Préparation des données – Présentation générale

- **Transformer les données :** cette étape consiste à mettre à jour les entrées de format ou de valeur de manière à obtenir un résultat clairement défini ou à rendre les données plus faciles à comprendre par un plus grand nombre.

- **Enrichir les données :** consiste à ajouter des données et à les relier à des données apparentées de manière à dégager des connaissances approfondies.

- **Stocker les données :** Lorsque la préparation des données est terminée, celles-ci peuvent être stockées dans un sous-échantillon.

Préparation des données - Évaluation et découverte des données

- Les fichiers CSV contenant les informations pour analyse :

1. dStatsCountry-Series.csv
2. EdStatsCountry.csv
3. EdStatsData.csv
4. EdStatsFootNote.csv
5. EdStatsSeries.csv

- La fichier CSV contenant les informations pour la pré-analyse est : *EdStatsData.csv*. Ce fichier contient :

→ 886930 Lignes et 70 colonnes

→ 242 Pays

→ 3665 Indicateurs

Préparation des données – Nettoyage et validation des données

- Les étapes effectuées pour nettoyer les données :

  - → Supprimer les lignes avec tous les champs 'NAN‘.
  - → Supprimer les colonnes avec tous les champs 'NAN‘.
  - → Supprimer les indicateurs qui contiennent les mots ‘**male**’, ‘**female**’ et ‘**Index parity**’.
  - → supprimer les lignes avec le champs 'Country Name' qui ne sont pas des pays (ex : ‘World’, ‘Arab World’, ‘South Asia ….’) en réalisant une jointure entre deux DataFrames.
  - → Conservez uniquement les lignes avec au moins 5 valeurs non 'NaN'.

- **Remarque :**

  Suite à ce nettoyage nous avons 148405 lignes au lieu de 886930.

Pré-analyse des données

Pré-analyse des données

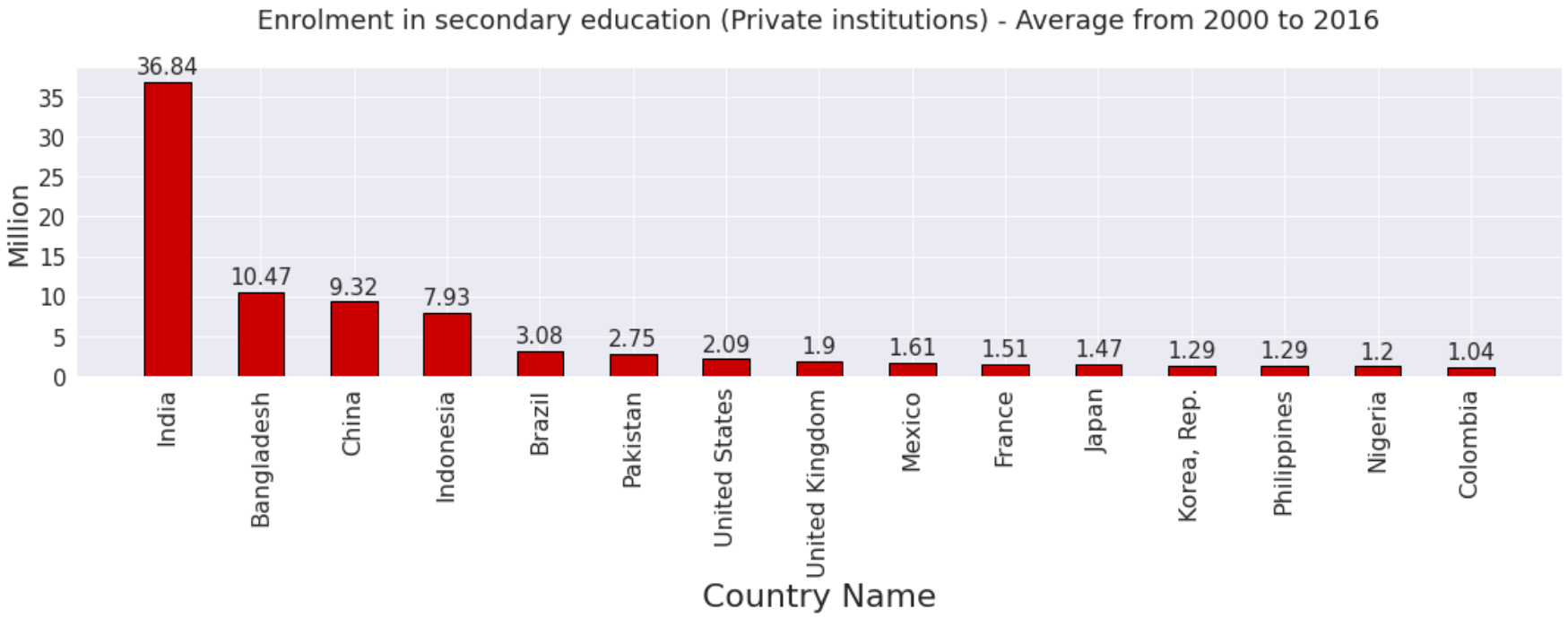
- Après avoir effectué le nettoyage et la validation des données nous allons nous baser sur :
  - Les indicateurs suivants qui vont nous permettre de répondre aux exigences et aux objectifs fixés par le centre de formation :

    ✓*Inscriptions dans les institutions privées*
    ✓Utilisateurs internet
    ✓Population totale
    ✓Croissance démographique annuelle
    ✓*Revenu national brut (RNB) par habitant en parité de pouvoir d'achat (PPA)*
    ✓*Rémunérations du personnel en % des dépenses totales*
    ✓*Salaires statutaires annuels des enseignants*
    ✓*Taux de réussite*
    ✓*Ratio nombre d'élèves par enseignant*

  - Le calcul de la moyenne de ces indicateurs de 2000 à 2016, car les données après 2016 ne sont pas enrichies.

Pré-analyse des données – Enseignement Secondaire

- Les inscriptions de l'enseignement secondaire aux écoles privées en ordre décroissant (Moyenne de 2000 à 2016) :



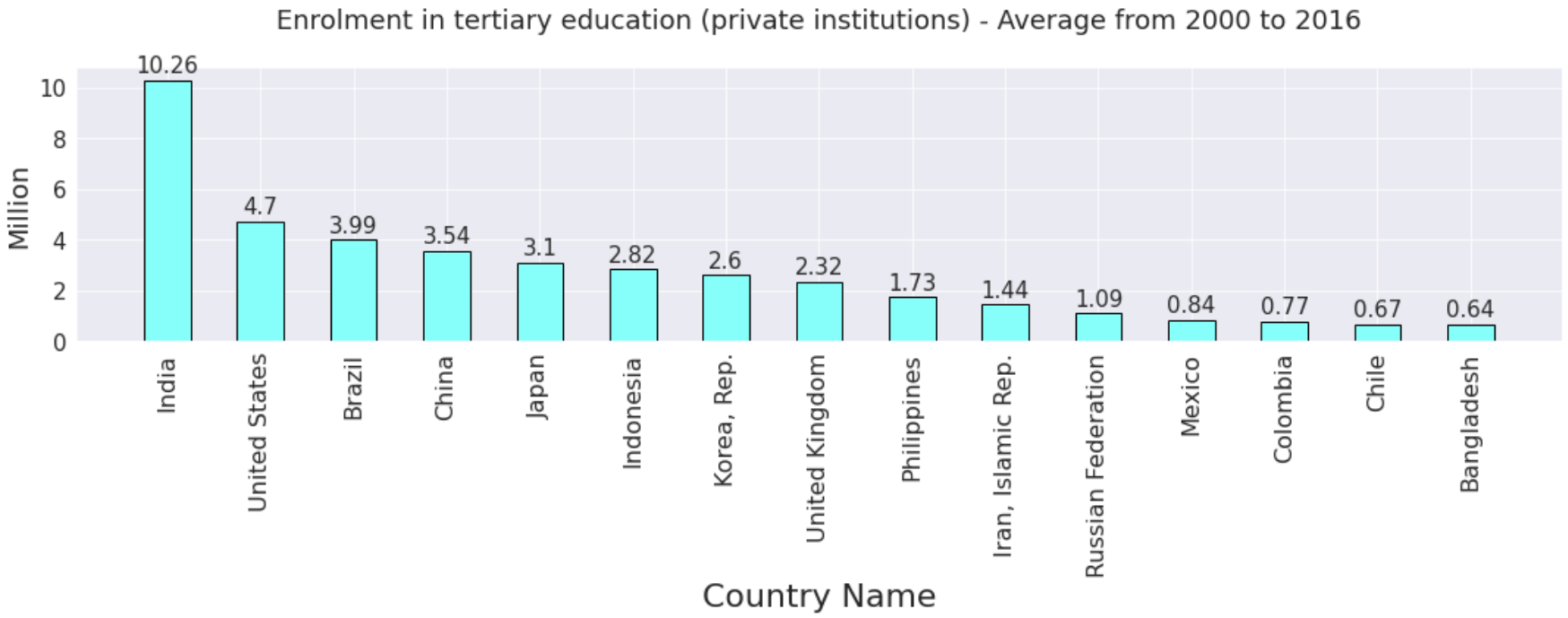
- **Remarque :**

  L'inde à plus de 36 millions d'inscriptions en moyenne dans des institutions privées.

  La Chine a environ 4 fois moins d'inscriptions par rapport à l'Inde.

Pré-analyse des données – Enseignement Supérieur

- Les inscriptions de l'enseignement supérieur aux écoles privées en ordre décroissant (Moyenne de 2000 à 2016) :



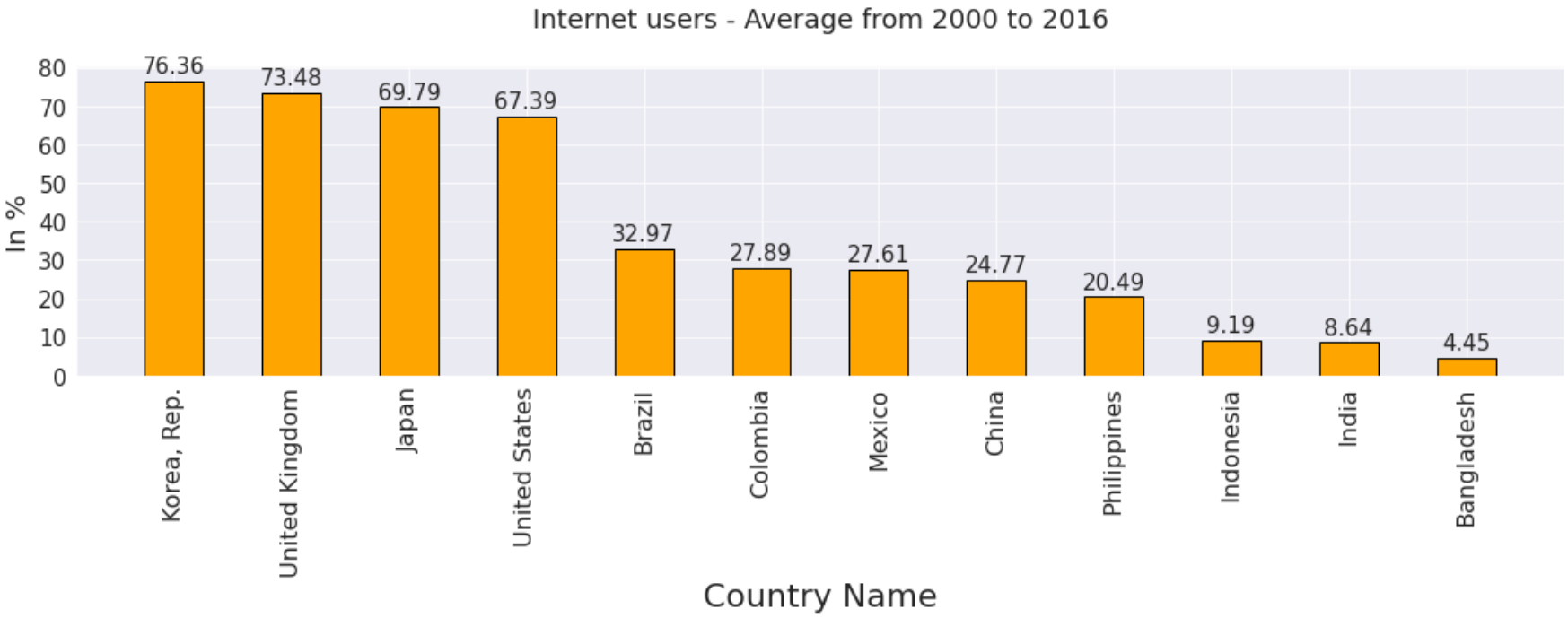
- **Remarque :**

  L'inde à plus de 10 millions d'inscriptions en moyenne dans des institutions privées supérieur.

  L'inde compte 47 millions d'inscriptions dans des institutions privées (secondaire et supérieur)

Pré-analyse des données

- Les utilisateurs connectés par pays :

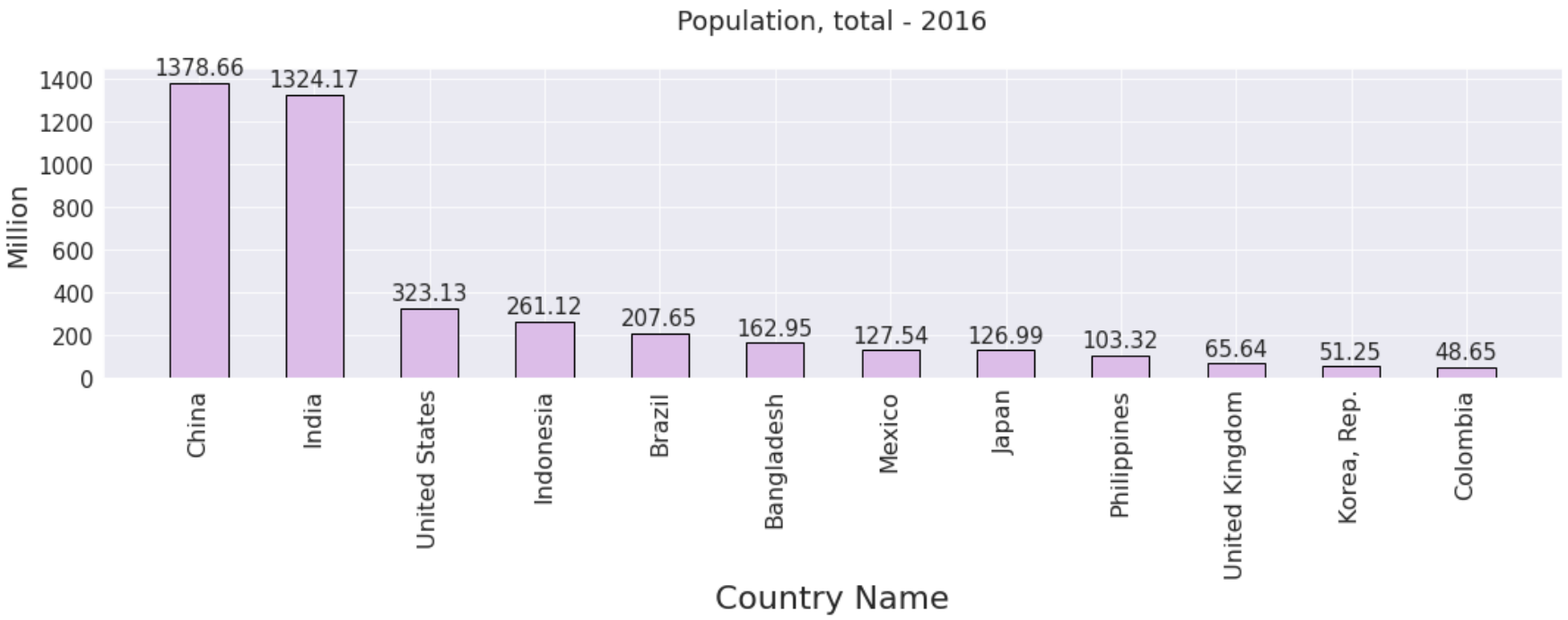


- **Remarque :**

  La Corée du Sud est le pays qui comporte le pourcentage le plus élevé des utilisateurs connectés, suivi de la Grande-Bretagne.

Pré-analyse des données

- Le nombre de la population totale :



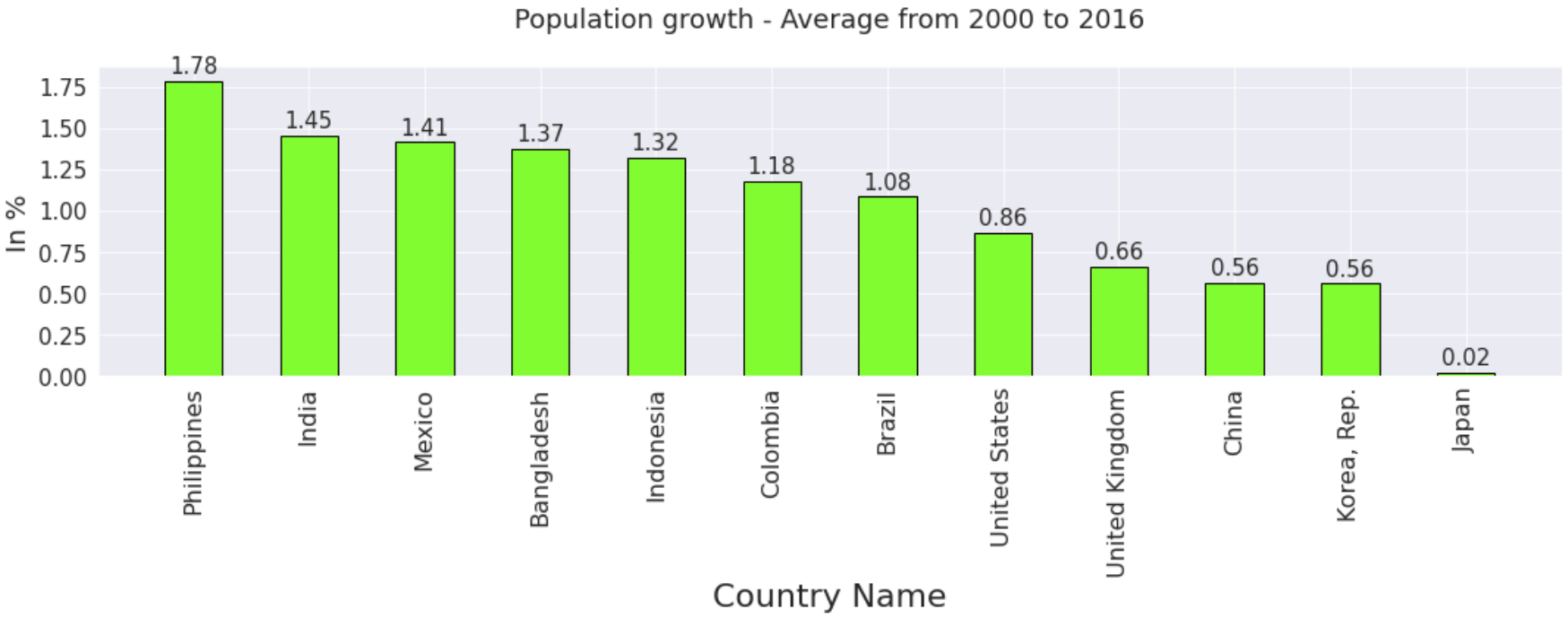
- **Remarque :**

  La Chine et l'Inde ont une population totale qui dépasse le 1,3 milliard, les Etats-Unis arrive en 3ème position avec plus 300 million (¼ de la population chinoise et indienne).

  Les 3 premiers pays représentent des clients avec un fort potentiel vu la population totale.

Pré-analyse des données

- Graphique de la croissance démographique :



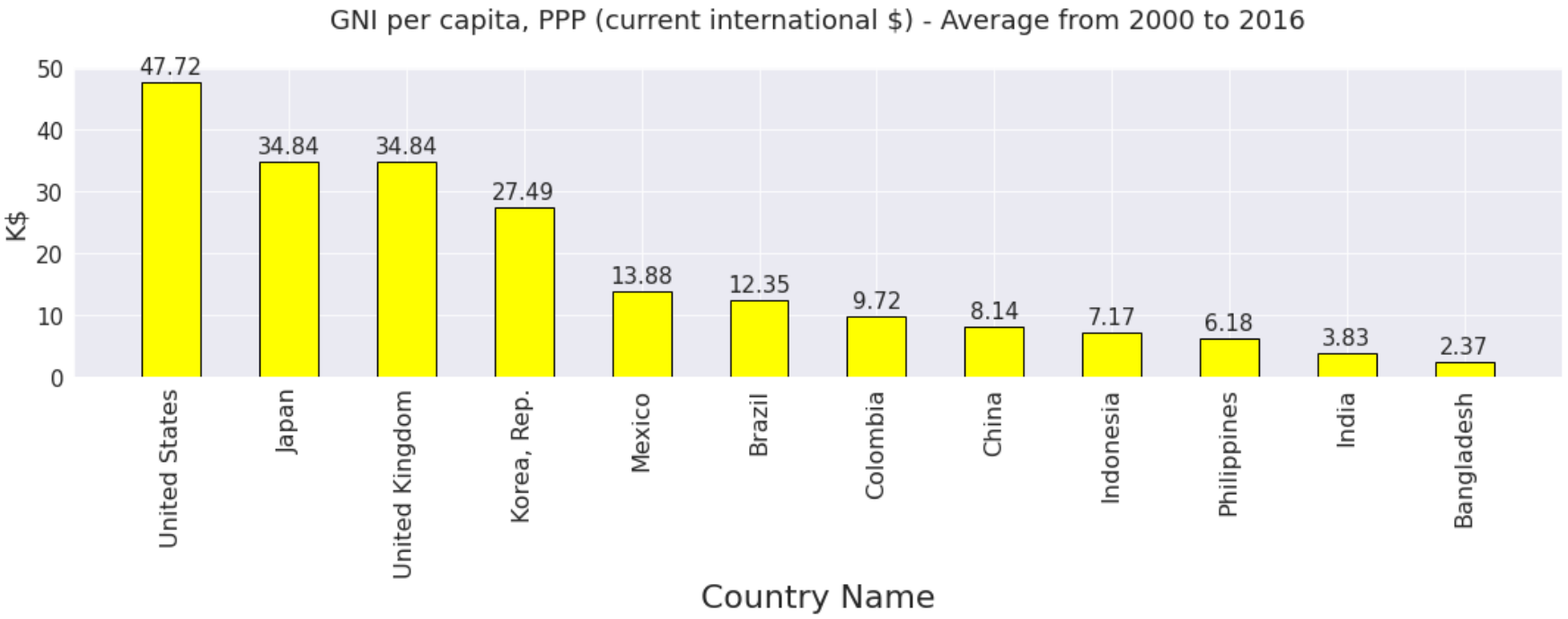
- **<u>Remarque :</u>**

  L'inde est 2ème du classement , avec une croissance démographique de plus de 1,45% en moyenne par an. Classé première en nombre d'inscriptions en secondaire et supérieur.

  L'évolution de ce potentiel client est en hausse grâce à sa forte croissance démographique.

Pré-analyse des données

- Le revenu national brut (RNB) par habitant en parité de pouvoir d'achat (PPA) :

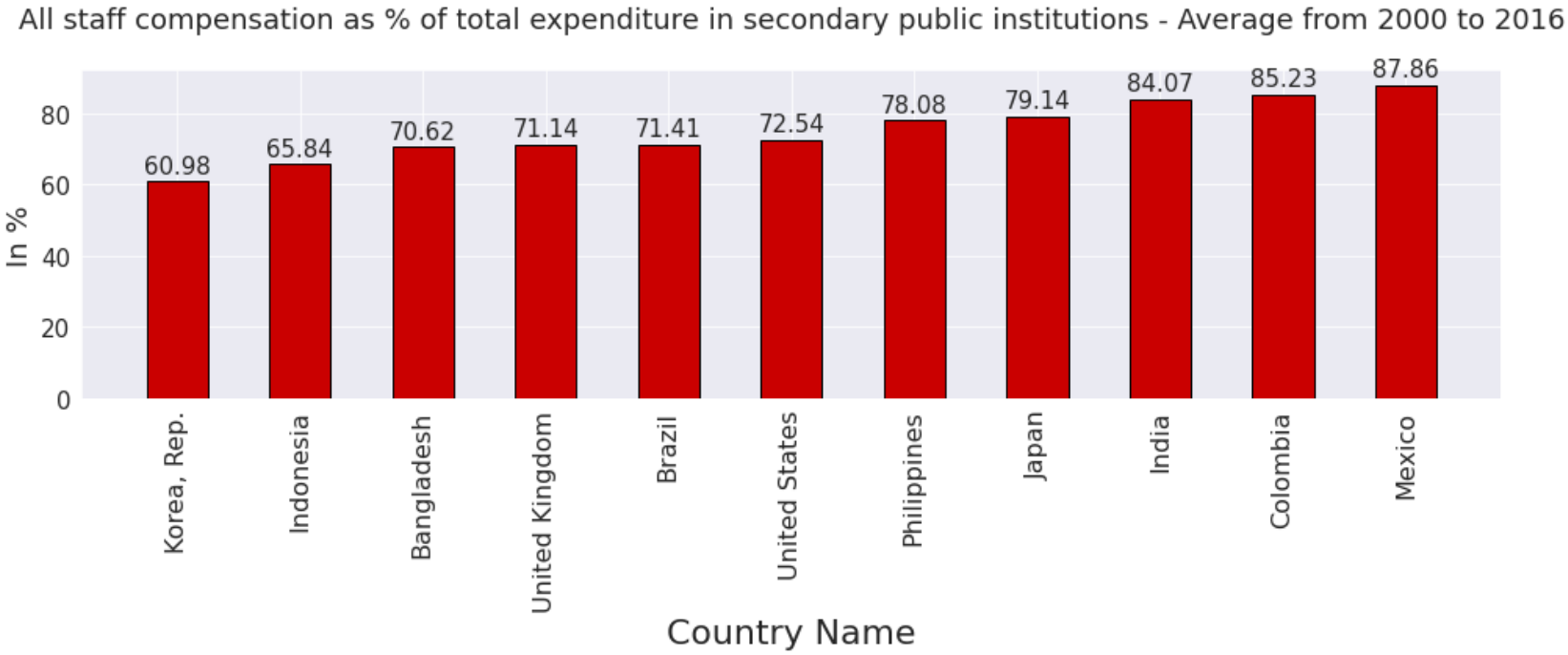


- **Remarque :**

  Les Etats-Unis a le RNB par habitant le plus élevé dans le monde, ce qui représente environ 12 fois le RNB de l'Inde et environ 6 fois le RNB de la Chine.

Pré-analyse des données – Enseignement Secondaire

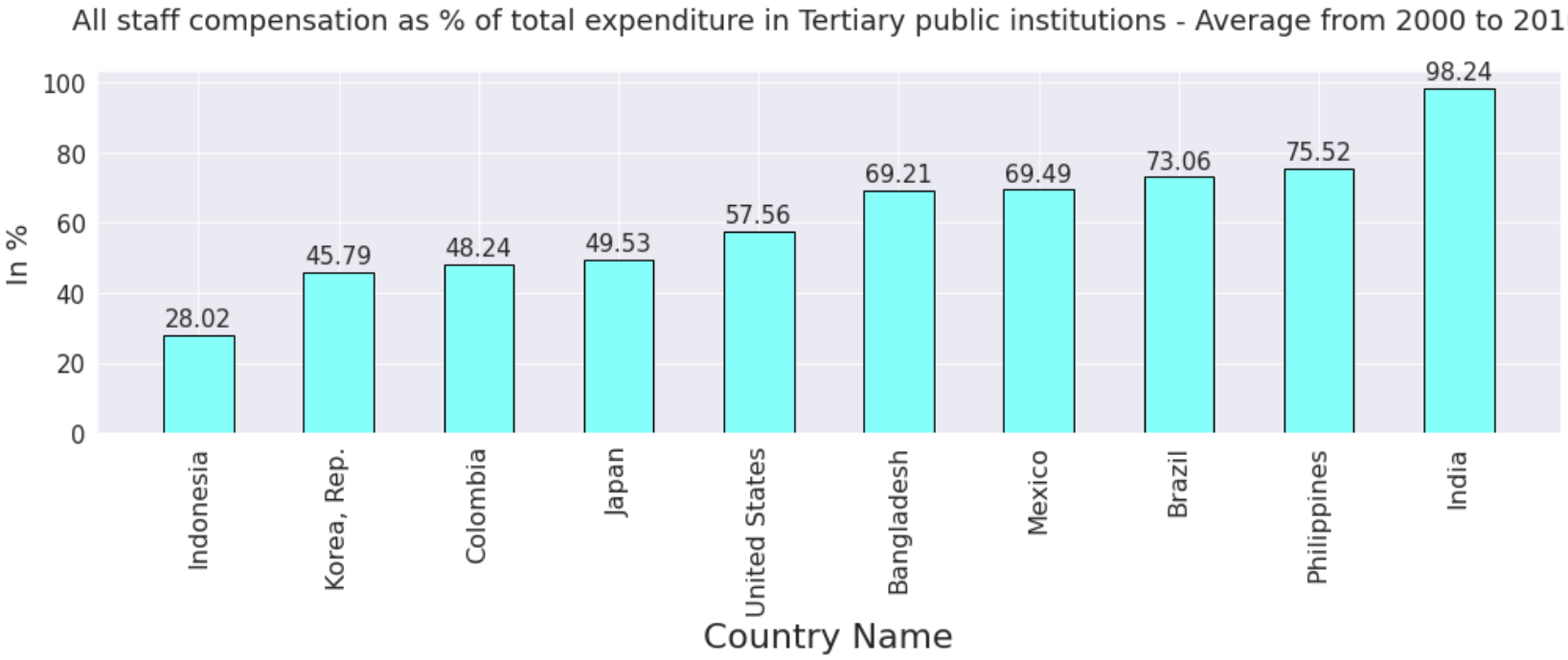
- La rémunération du personnel en % des dépenses totales dans les établissements publics de l'enseignement secondaires (%)



- **Remarque :** La rémunération du personnel dépasse les 80% des dépenses totale pour le Mexique, Colombie, et l'inde. Ce niveau élevé de dépenses est un atout important pour constituer, développer et entretenir un corps enseignant compétent et de qualité.

Pré-analyse des données – Enseignement Supérieur

- La rémunération du personnel en % des dépenses totales dans les établissements publics de l'enseignement supérieur (%)

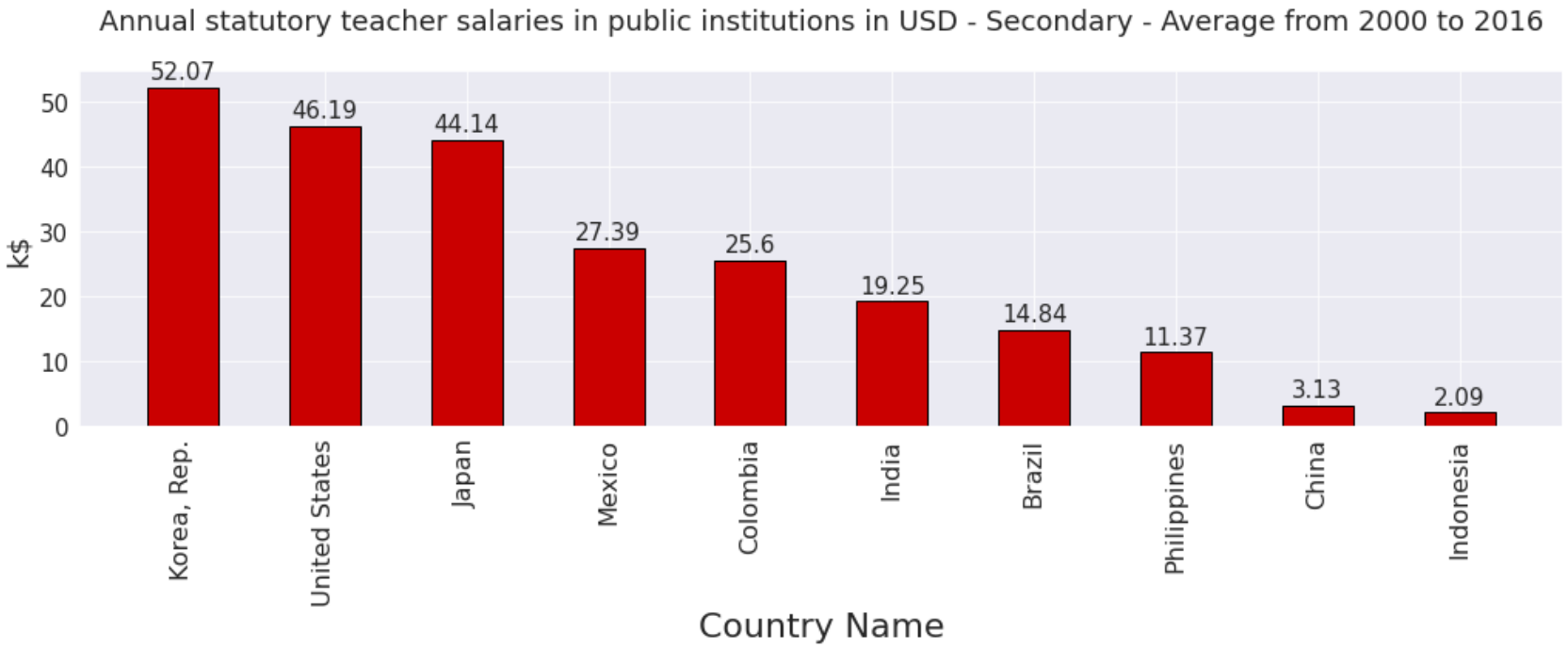


- **Remarque :**

  L'Inde dépasse les 98% des dépenses totale pour la rémunération du personnel. C'est un indicateur que l'Inde développe et entretien un corps enseignant compétent et de qualité. Or les Etats-Unis ne dépense que 57,56% de son budget.

Pré-analyse des données – Enseignement Secondaire

- Salaire moyen statutaires annuel des enseignants :

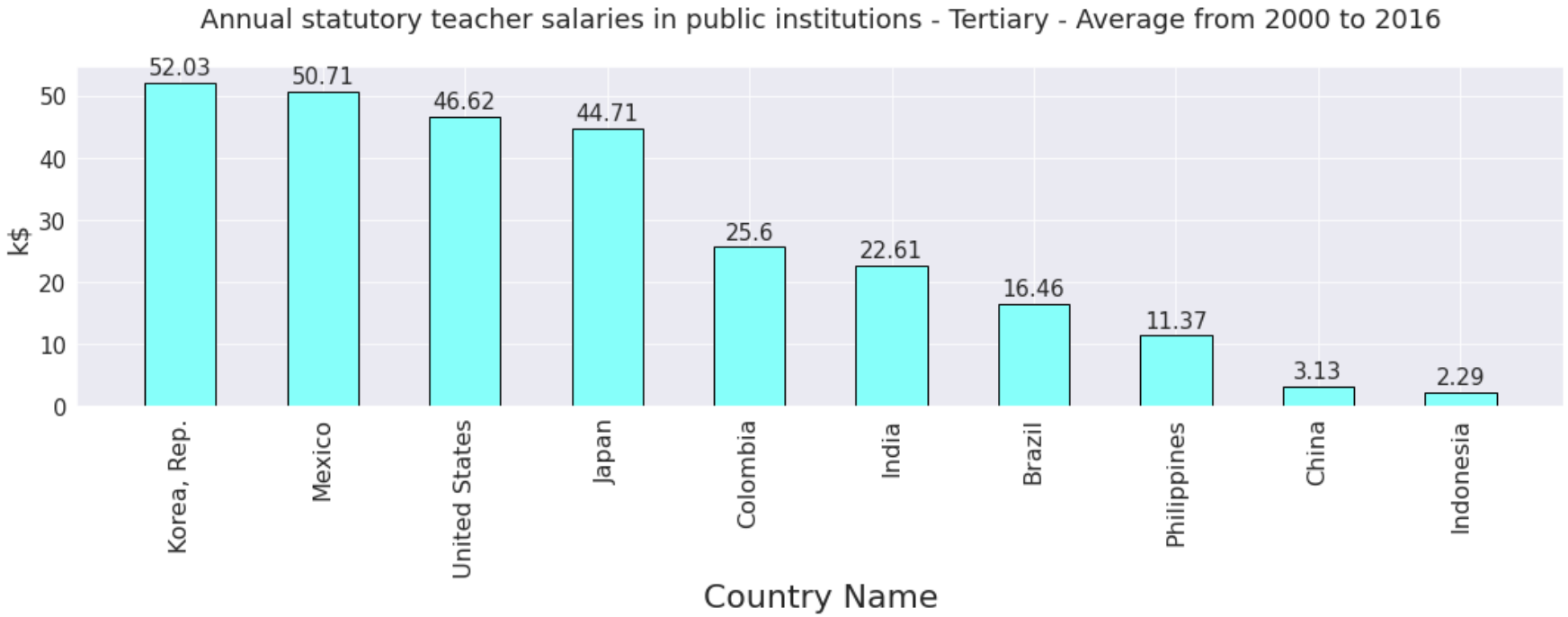


- **Remarque :**

  L'Indonésie et la Chine ont le salaire moyen d'enseignants du secondaire le plus bas (entre 2K$ et 3K$ par an). Par contre les enseignants aux Etats-Unis et la Corée-du-Sud ont le salaire le plus élevé.

Pré-analyse des données – Enseignement Supérieur

- Salaire moyen statutaires annuel des enseignants :



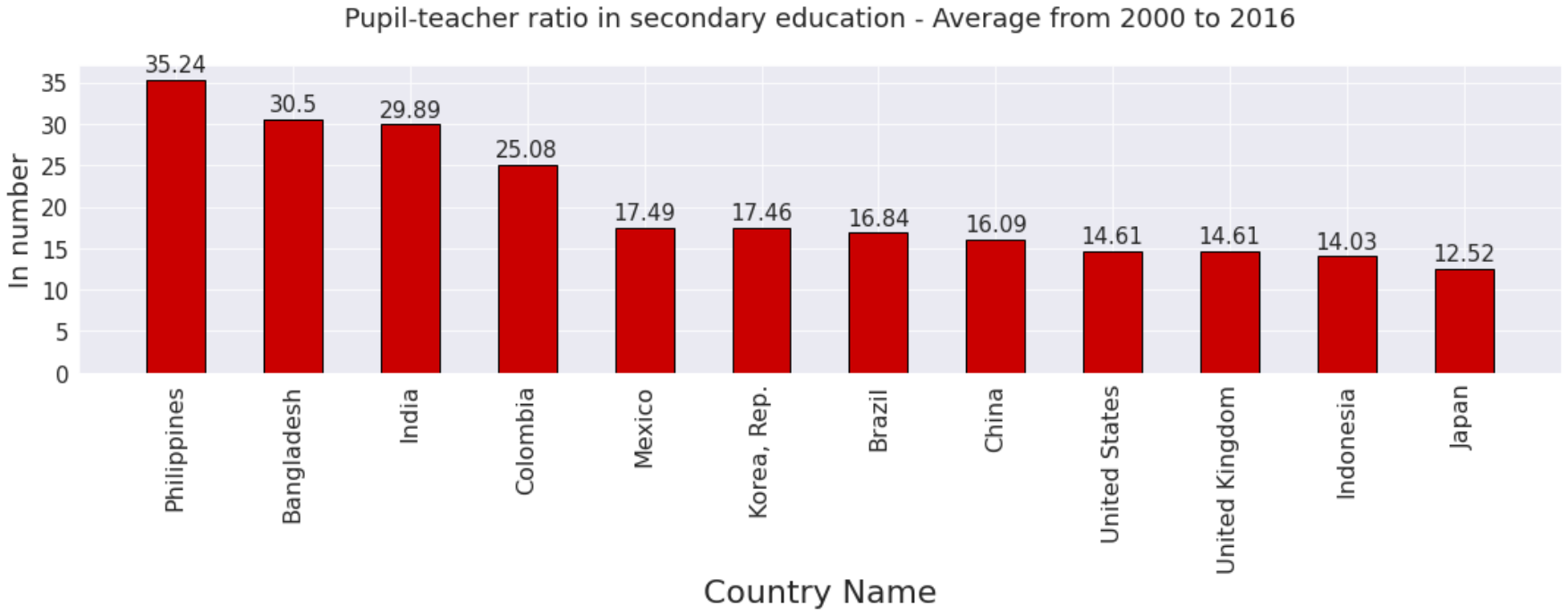
- **Remarque :**

  Les enseignants au Mexique et la Corée du Sud ont un salaire moyen de 50K$ par an.

  Comme pour le secondaire l'Indonésie et la Chine ont le salaire moyen d'enseignants le plus bas (entre 2K$ et 3K$ par an).

Pré-analyse des données – Enseignement Secondaire

- La ratio élèves-enseignant :

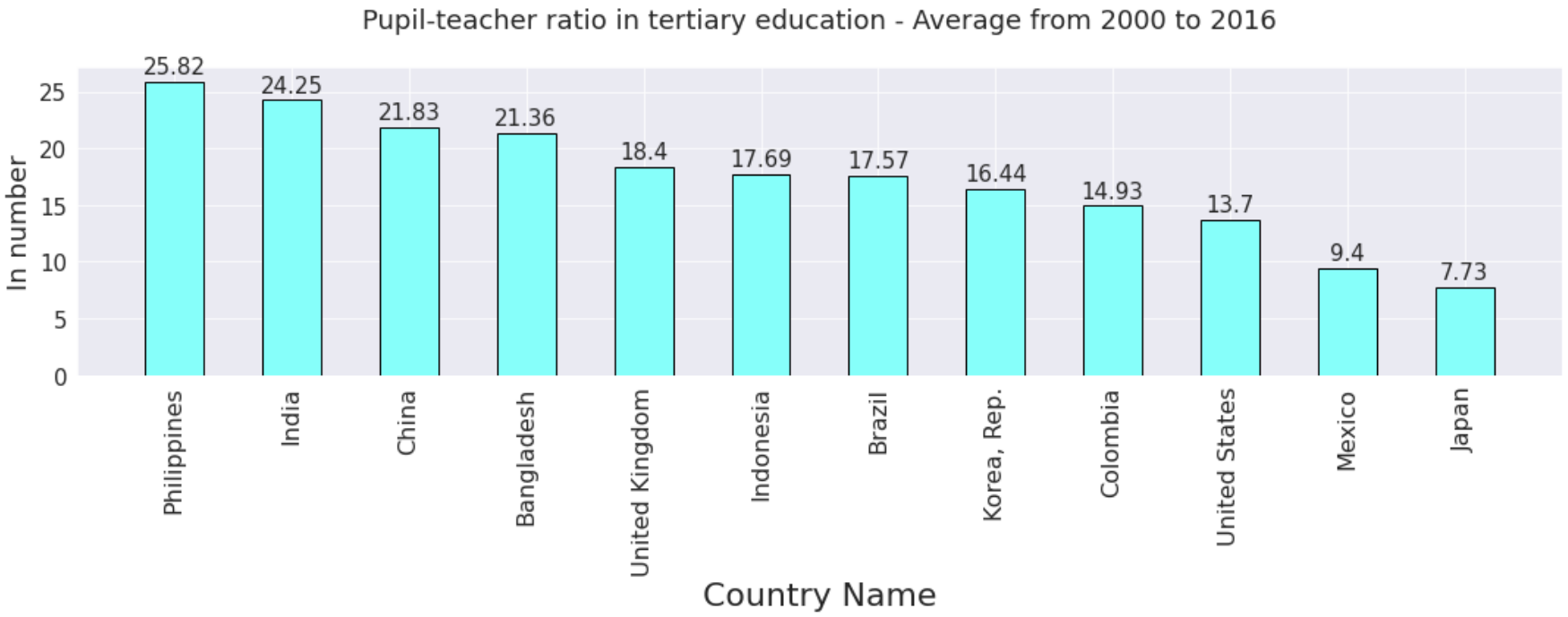


- **Remarque :** Les Philippines, le Bangladesh et l'Inde ont le ratio élèves-enseignant le plus élevé (au moins de 29 élèves par enseignant).

  Cet indicateur mesure la charge de travail des enseignants et les allocations de ressources humaines dans les établissements d'enseignement

Pré-analyse des données – Enseignement Supérieur

- La ratio élèves-enseignant :

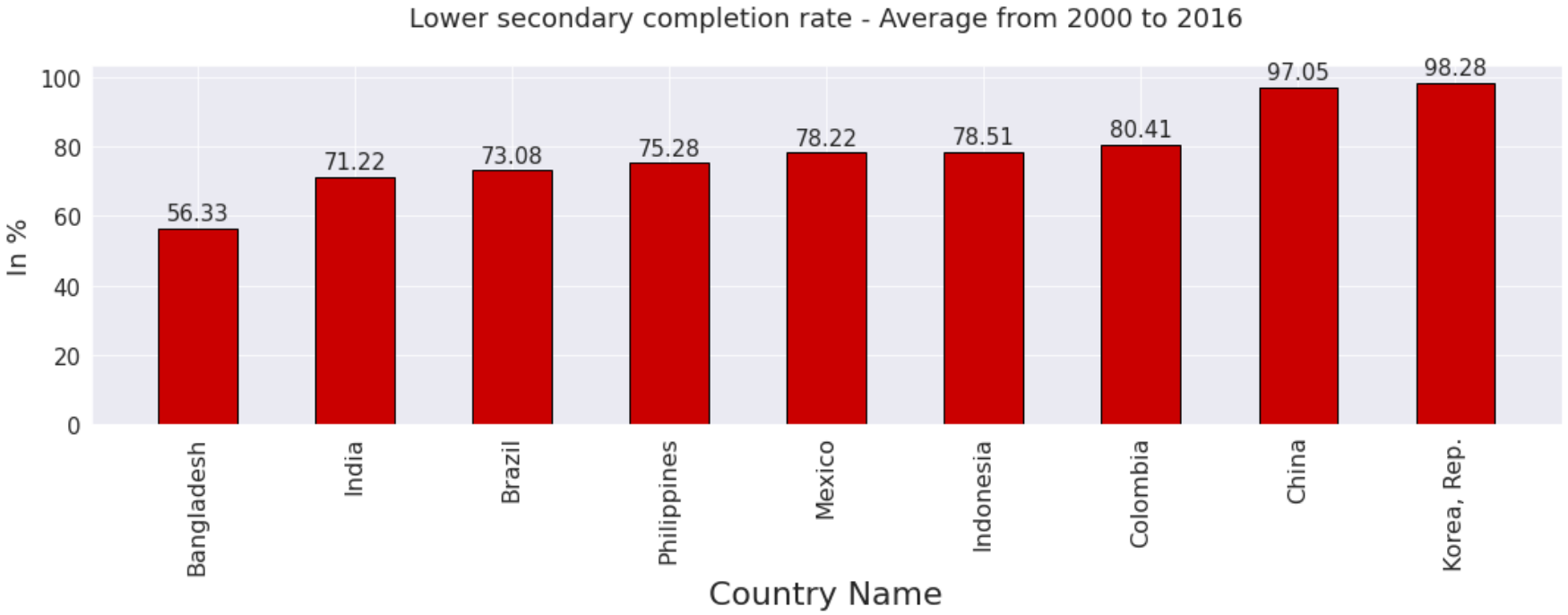


- **Remarque :** Les Philippines et l'Inde ont le ratio élèves-enseignant le plus élevé (au moins de 24 élèves par enseignant).

  La charge de travail des enseignants est plus élevée dans ces pays

Pré-analyse des données – Enseignement Secondaire

- Le taux de réussite du premier cycle du secondaire :

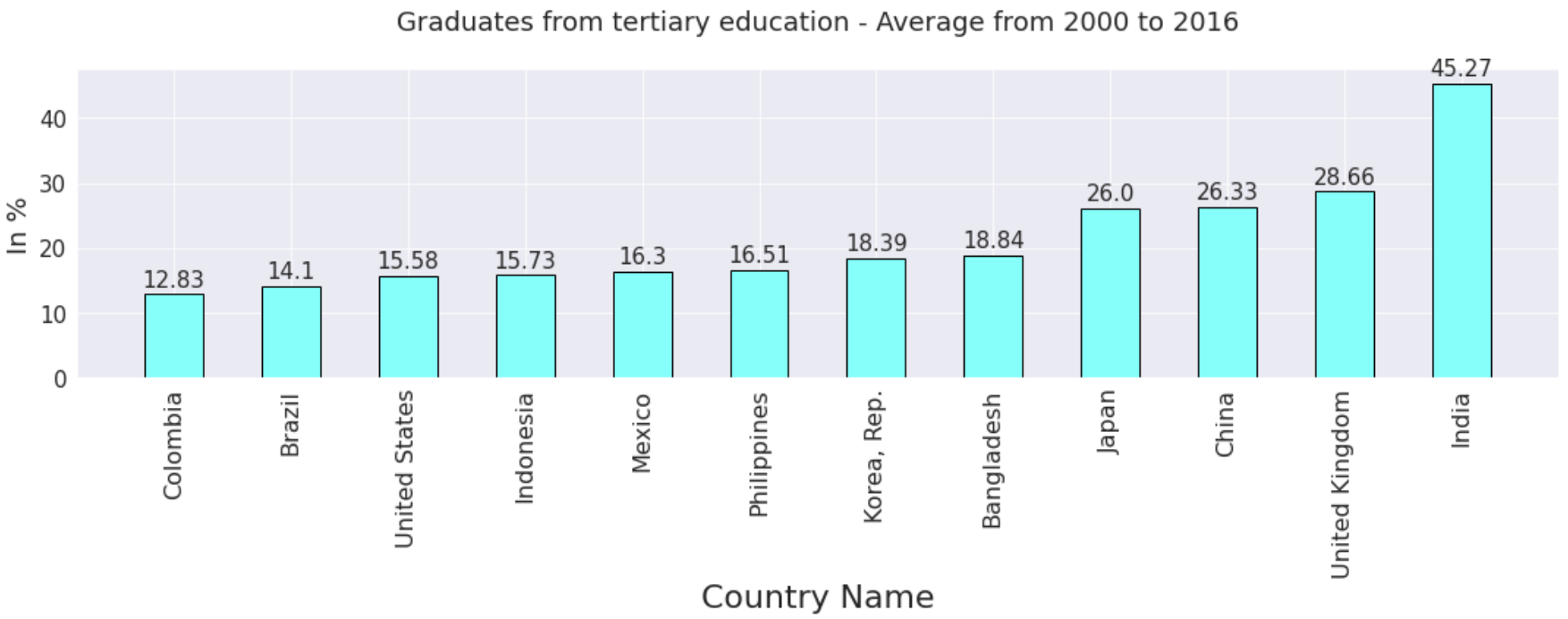


- **Remarque :**

  Des indicateurs, comme le taux de réussite sont souvent utilisés pour déterminer la qualité de l'enseignement.

Pré-analyse des données – Enseignement Supérieur

- Le taux de diplômés :



- **Remarque :**

  Le taux de diplômés indique la qualité de l'enseignement et des enseignants.

Conclusions

Conclusions

- Dans ces conclusions nous allons traiter les points ci-dessous pour chaque pays, on se basant sur les graphiques précédents:
  - Les pays avec un fort potentiel de clients pour nos services qui propose des contenus de formation en ligne pour un public de niveau lycée et université.
  - Les pays avec une évolutions croissante
  - Les pays dans lesquels l'entreprise doit opérer en priorité

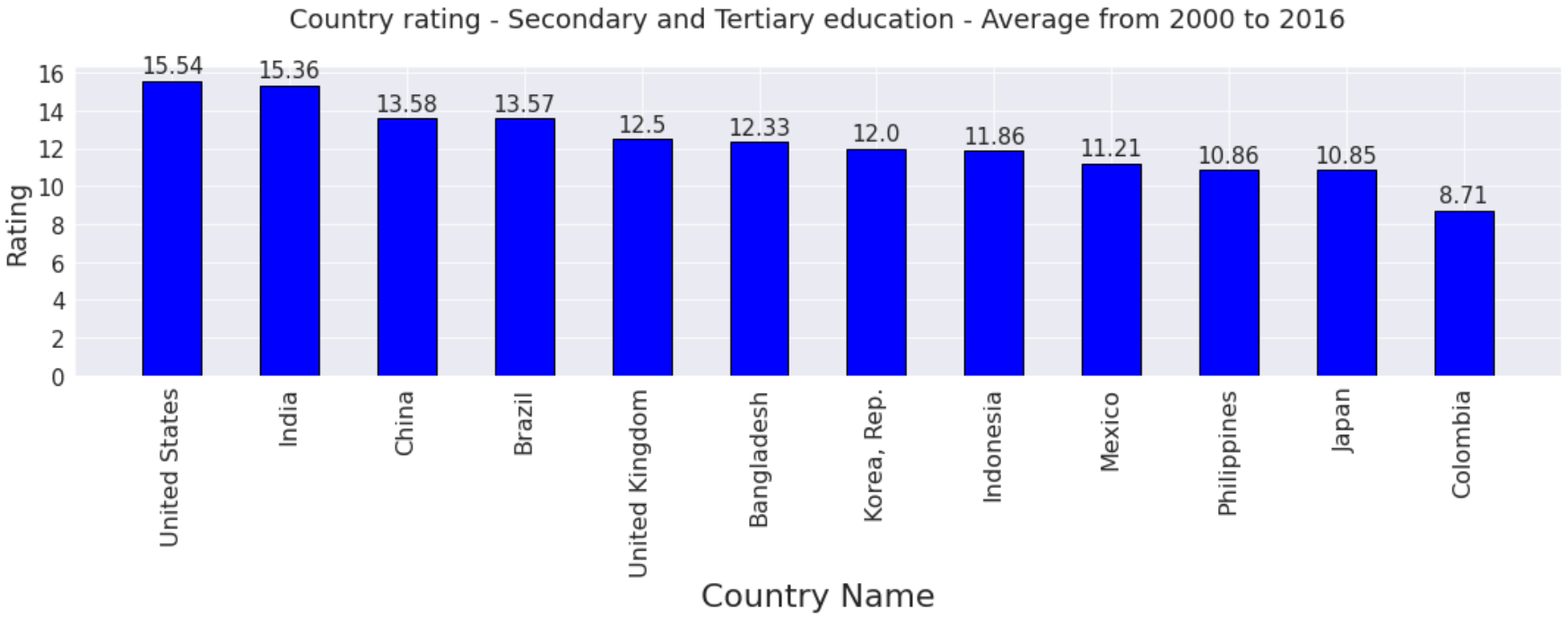
Conclusions

- Pour cibler les pays les plus attractifs pour le centre de formation, il faut noter les pays par rapport à l'importance des indicateurs (coefficient).
  - Ci-dessous les coefficients des indicateurs :

Indicateurs	Coefficient
Inscriptions dans les institutions privées	3
Utilisateurs d'internet	1
Population totale	2
Croissance démographique annuelle	2
Revenu national brut (RNB) par habitant en parité de pouvoir d'achat	2
Rémunérations du personnel en % des dépenses totales	1
Salaires statutaires annuels des enseignants	2
Taux de réussite et diplômés	1
Ratio nombre d'élèves par enseignant	2

Conclusions

- Suite à l'utilisation du système de notation par rapport au coefficient, ci-dessous le résultat de notation par pays :



Conclusions

- La conclusion après l'analyse des données et des graphiques, les pays avec un fort potentiel de clients pour la formation en ligne sont (par ordre de priorité) :

1. **Les Etats-Unis** : Ce pays compte plus de 303 millions d'habitants et plus de 6,5 millions d'inscriptions dans des établissements privés (secondaire et supérieur). Le revenu national brut par habitant est le plus élevé dans le monde (47,72 K$) représente un fort potentiel de clients pour les services du centre de formation.

2. **L'Inde :** Avec Une population qui dépasse le 1,3 milliard, et une croissance démographique de plus de 19 millions par an en moyenne, ce pays aura une évolution croissante malgré un revenu brut par habitant (3.86K$) 12 fois inférieur à celui des Etats-Unis.

3. **La Chine :** Est le pays le plus peuplé du monde avec 12,86 millions d'inscriptions dans les écoles privées (secondaire et supérieur), représente un fort potentiel avec un revenu brut par habitant environ 2 fois supérieur à celui de l'Inde.