# Projet d'expansion à l’international d’Academy

Analyse des données de systèmes éducatifs
(données de la banque mondiale)

academy

# Contexte de ce travail

Academy souhaite étendre son activité à l'international. Pour cela, il est nécessaire de repérer les pays les plus propices à cette expansion.

Les données de la banque mondiale semblent être un bon point de départ pour mettre en place une stratégie.

Le but de ce travail est donc de mettre en forme ces données pour en extraire un classement de pays. Ce dernier permettra d'élaborer un début de réponse quant à l'implantation d'un service d'éducation en ligne pour des niveaux lycée et université.

# Description de la base de données de la banque mondiale

Ensemble de données sur le champ de l'éducation regroupé en 5 dataframes

| **Nom du jeu de données** | **Description du jeu de données** |
|---|---|
| EdStatsCountrySeries | Représente une collection de variables disponibles pour chaque pays, à partir desquelles les données sont extraites. Chaque ligne correspond à une variable décrite et contient le code pays, le nom de la variable et la description du mode de recueil. |
| EdStatsCountry | Contient un ensemble d'informations concernant les différents pays. Chaque ligne représente un pays avec plusieurs variables décrivant les données du pays, leur ancienneté, le système de récupération des données, etc. |
| EdStatsData | Chaque ligne représente la valeur d'un indicateur pour un pays et pour une année donnée. |
| EdStatsFootNote | Chaque ligne représente un code de série avec, en indication, le pays auquel il est rattaché, une description et une année. |
| EdStatsSeries | Chaque ligne représente un nom de série avec les informations qui lui sont rattachées pour la décrire. |

academy

**Code utilisé pour le traitement de données : Python**
**IDE: JupyterLab**

# Description de la base de données de la banque mondiale

Ensemble de données sur le champ de l'éducation regroupé en 5 dataframes

| **Nom du jeu de données** | **Description du jeu de données** |
|---|---|
| EdStatsCountrySeries | 613 lignes et 4 colonnes. |
| EdStatsCountry | 241 lignes et 32 colonnes. |
| EdStatsData | 886930 lignes et 70 colonnes. |
| EdStatsFootNote | 643638 lignes et 5 colonnes. |
| EdStatsSeries | 3665 lignes et 21 colonnes. |

academy

**Code utilisé pour le traitement de données : Python**
**IDE: JupyterLab**

# Vérifications réalisées

- Vérification de la présence de doublons : Aucun doublon détecté.
- Recherche des colonnes vides et suppression de ces colonnes :
  - EdStatsCountrySeries : “Unnamed: 3”
  - EdStatsCountry: “Unnamed: 31”
  - EdStatsData : “Unnamed: 69”
  - EdStatsFootNote : “Unnamed: 4”.
  - EdStatsSeries : “Other web links”, “Unnamed: 20”, “License Type”, “Notes from original source”, “Unit of measure”, “Related indicators”.
- Description de données qualitatives : l’intégralité des colonnes est décrite dans le code mis à disposition sur GitHub
- Suppression de données “pays” non valides dans le fichier EdStatsCountry

# Liste des données “pays” non valides

Dans la serie pays du fichier EdStatsCounrty un certain nombre d’items ne semblent pas valides et ont été supprimés. Par exemple :

- Des données correspondant à des régions (“Arab World”, "East Asia & Pacific", "Europe & Central Asia", ...)
- Des regroupements selon une typologie (“High income”, “Low income", ...)

L’ensemble de la liste des éléments supprimés est disponible dans le code fourni sur le GitHub en fin de document.

# Obtention d’une liste de variables utilisables

- À partir de la liste des pays, un nettoyage du dataframe EdStatsCountrySeries, grâce à un merge avec le dataframe EdStatsCountry nettoyé, a permis un nettoyage de l’ensemble des dataframes.

- Une sélection des topics intéressants pour une approche métier a été effectuée sur le dataframe EdStatsSeries

- La sélection a été faite pour retenir les sujets suivants : le niveau scolaire, les volumes de populations, les volumes financiers, les capacités de communication pour l'aspect distanciel et la place de l'enseignement privé.

- Grâce au lien fait via les codes de series du dataframe EdStatsCountrySeries, le dataframe EdStatsData a été réduit aux topics pertinents, limitant le nombre de variables disponibles.

# Sélection des années pour l'analyse et mise en forme du dataframe EdStatsData

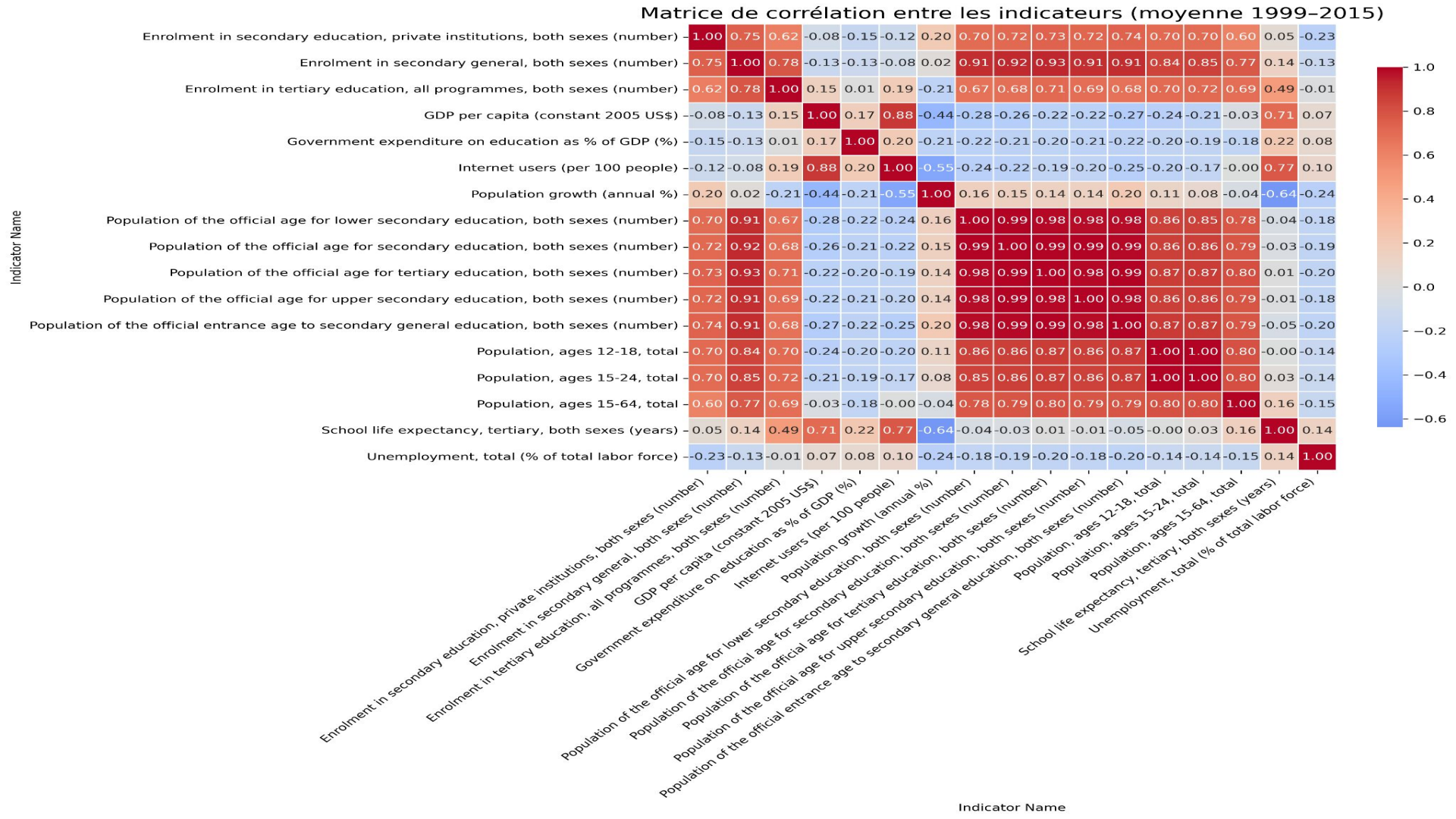
- Le dataframe contient de nombreuses années, mais toutes ne disposent pas d'un volume de données suffisant ; un certain nombre ont donc été supprimées.
- Les années 1999 à 2015, avec le plus grand nombre de données, ont été conservées.
- Le dataframe EdStatsData a été mis en forme selon la structure suivante : une ligne par pays et une colonne par indicateur. Les valeurs correspondent à la moyenne des données sur l'ensemble des années sélectionnées.

# Sélection des variables pour l'analyse

2 phases ont été réalisées pour la sélection des variables d'intérêts :

1) Une approche métier qui contient des variables de comme les niveaux de population, les niveaux scolaires visés, le niveau d'utilisation d'internet et les données économiques.

2) Une approche statistique avec des matrices de corrélation pour ne sélectionner qu'un nombre réduit de variable.


Matrice de corrélation entre les indicateurs (moyenne 1999–2015)
Indicator Name
Indicator Name
Enrolment in secondary education, private institutions, both sexes (number)
Enrolment in secondary general, both sexes (number)
Enrolment in tertiary education, all programmes, both sexes (number)
GDP per capita (constant 2005 US$)
Government expenditure on education as % of GDP (%)
Internet users (per 100 people)
Population growth (annual %)
Population of the official age for lower secondary education, both sexes (number)
Population of the official age for secondary education, both sexes (number)
Population of the official age for tertiary education, both sexes (number)
Population of the official age for upper secondary education, both sexes (number)
Population of the official entrance age to secondary general education, both sexes (number)
Population, ages 12-18, total
Population, ages 15-24, total
Population, ages 15-64, total
School life expectancy, tertiary, both sexes (years)
Unemployment, total (% of total labor force)
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

# Sélection des variables pour l'analyse

4 variables sont gardés et utilisé pour la construction d'un classement de pays d'intérêt :

1) "Enrolment in secondary general, both sexes (number)" : utilisé dans l'analyse principale.
2) "School life expectancy, tertiary, both sexes (years)" : utilisé dans l'analyse principale.
3) "Unemployment, total (% of total labor force)" : utilisé dans une analyse secondaire.
4) "Population growth (annual %)" : utilisé dans une analyse secondaire.

Les deux premières variables sont utilisées pour établir un classement des pays qui sont le mieux dotés pour une implantation allant du lycée jusqu'à l'université.

Les deux variables suivantes permettent un éclairage des conditions d'évolution à prendre en compte et de la stratégie possible concernant le chômage et l'apport de la formation à la problématique. Cette analyse a été appliquée pour les 20 premiers pays du classement

# Classement des pays selon les variables d'intérêts

| Country Name | Enrolment in secondary general, both sexes (number) | School life expectancy, tertiary, both sexes (years)_y | Classement moyen | Rang final | Note croissance | Note chômage |
|---|---|---|---|---|---|---|
| United States | 3.0 | 3.0 | 3.0 | **1** | Faible | Tension modérée |
| Russian Federation | 8.0 | 14.0 | 11.0 | **2** | Negatif | Tension modérée |
| Korea, Rep. | 26.0 | 1.0 | 13.5 | **3** | Faible | Tension faible |
| Ukraine | 25.0 | 12.0 | 18.5 | **4** | Negatif | Tension modérée |
| Argentina | 23.0 | 16.0 | 19.5 | **5** | Stable | Accès difficile à l'emploi |

# Classement des pays selon les variables d'intérêts

Selon l'analyse réalisée, les pays à privilégier sont les suivants :

United States, Russian Federation, Korea Rep., Ukraine, Argentina, Germany, Japan, United Kingdom, France, Spain, Australia, Italy, Canada, Poland, Turkey, Thailand, Belarus, Chile, Greece, Brazil

Bien entendu, les données étant datées de 2015, certains pays peuvent avoir connu une variation selon plusieurs variables non répertoriées ici (par exemple le contexte géopolitique).

# Ressources

Lien Github vers le code d’analyse

- https://github.com/Thomas-Auvin/Projet2

Lien vers les jeux de données

- https://datacatalog.worldbank.org/search/dataset/0038480