Analysez des données de systèmes éducatifs

Formation Data scientist
Projet n°2
Bayram DONAT

Présentation du projet (1/2)

- **Contexte :**
  - Entreprise : Academy,
  - Activité : contenus de formation en ligne pour lycée et université
  - Enjeux : expansion à l'international
- **Demande :**
  - Analyse exploratoire de la banque de données mondiale
    - Quels pays avec un fort potentiel client?
    - Evolution de ce potentiel pour chacun de ces pays?
    - Priorité pour quels pays?

Présentation du projet (2/2)

- Description des informations contenues
- Analyse des données manquantes et dupliquées
- Sélection des critères importants pour la problématique
- Indicateurs statistiques par groupement et pays
- Réponses aux questions de l'entreprise

Description des informations contenues (1/2)

- **Données à analyser**
  - Un fichier zip contenant avec 5 fichiers CSV où les données sont séparées par des virgules
- **Méthode d'analyse**
  - Télechargement du fichier zip, décompression dans le répertoire « fichiers sources »
  - Importation de chaque fichier dans un dataframe pandas
  - Visualisation du nombre de lignes et de colonnes avec info() et des premières lignes avec head()
- **A retenir**
  - Aperçu du contenu des colonnes
  - Nombre des lignes non nulles par colonne

Comment ?

Description des informations contenues (2/2)

- **5 dataframes**
  - EdStatsCountry : généralités par pays : région géographique, dates des derniers indicateurs par catégorie économique…
  - EdStatsCountry_Series : origine des données statistiques des pays / fiabilités des données…
  - EdStatsData : indicateurs par pays et année de 1970 à 2100
  - EdStatsFootNote : dates et méthodes d'obtention des données
  - EdStats_Series : détails des indicateurs présents dans EdStatsData…
- **Pour l'analyse, on choisit EdStatsData**

Comment ?

Analyse des données manquantes et dupliquées (1/2)

- **Données dupliquées dans EdStatsData : 0 EdStatsData.duplicated().sum()**
- **Suppression des lignes**
  - pour les années bien remplies (>100 000 valeurs non nulles)
    - voir EdStatsData.info() => 1990, 1995, et de 1999 à 2015
    - Pour chaque indicateur, nb valeurs manquantes = 1 – nb de valeurs non nulles divisé par le nombre d'années (19)
    - Moyenne (nb valeurs manquantes) = 0.848136775168277
    - Suppression des indicateurs avec nombre valeurs nulles supérieur à la moyenne
- **Suppression des colonnes contenant plus de 73 % de données manquantes**
  - EdStatsData=EdStatsData[EdStatsData.columns[EdStatsData.isnull().mean()<0.73]]
  - Suppression de 40 colonnes sur 70 colonnes (57%)
  - Conservation des années de 1990 à 2015

Analyse des données manquantes et dupliquées (2/2)

- **Suppression des lignes renseignées sur moins de 2 années**
  - EdStatsData.dropna(thresh=6,inplace=True) : thresh=6 car les 4 premières colonnes non numériques
- **Suppression des lignes avec des distinctions homme/femmes**
  - EdStatsData.drop(EdStatsData[EdStatsData['Indicator Name'].str.contains('female|male|Male|GPI')].index, inplace=True)
- **Correction des pourcentages hors plage pour les indicateurs se terminant en ' (%)'**
  - Si Pourcentage<0 alors Pourcentage=0
  - Si Pourcentage>100 alors Pourcentage=100
- **Le dataframe EdStatsData contient avant nettoyage 886930 lignes et 70 colonnes. Après nettoyage, il passe à 67633 lignes (396 indicateurs) et 30 colonnes.**

Sélection des critères importants pour la problématique
Critères sélectionnés (1/2)

- ‘Internet users (per 100 people)',
- 'GDP per capita (constant 2005 US$)',
- 'Population growth (annual %)',
- ‘Population, total’
- 'Enrolment in upper secondary education, both sexes (number)',
- 'Pupil-teacher ratio in upper secondary education (headcount basis)',
- 'Teachers in upper secondary education, both sexes (number)',
- 'Enrolment in tertiary education, all programmes, both sexes (number)',
- 'Pupil-teacher ratio in tertiary education (headcount basis)',
- 'Teachers in tertiary education programmes, both sexes (number)'

Critère indispensable

Critères généralistes

Critères lycéens

Critères universitaires

Les classements se feront sur les dernières valeurs connues

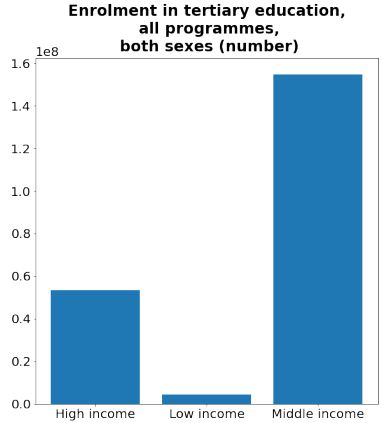
Sélection des critères importants pour la problématique
Groupement des pays (2/2)

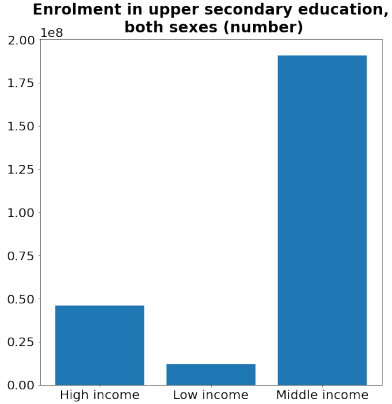
- **Niveau de développement**
  - ‘Low income’
  - ‘Middle income’
  - ‘High income’

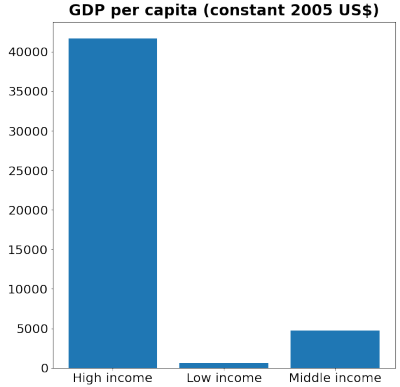
Comment ?

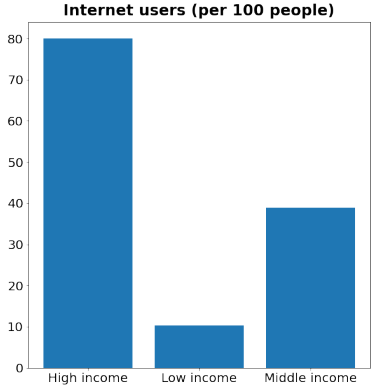
- **Zones géographiques**
  - 'East Asia & Pacific',
  - 'Europe & Central Asia',
  - 'Latin America & Caribbean',
  - 'Middle East & North Africa',
  - 'North America',
  - 'Sub-Saharan Africa'

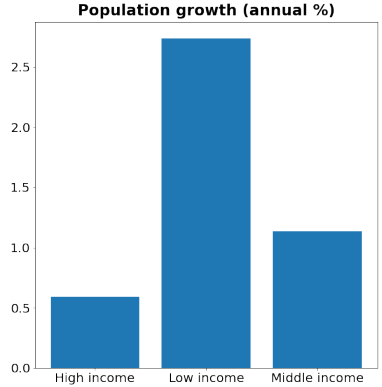
Indicateurs statistiques (1/3)
Niveau de développement

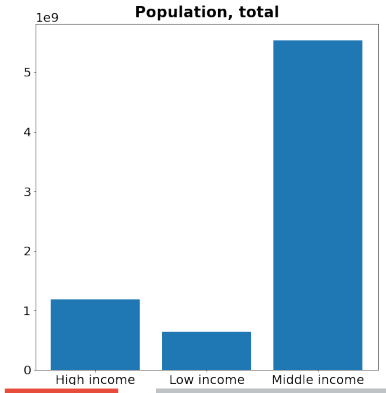


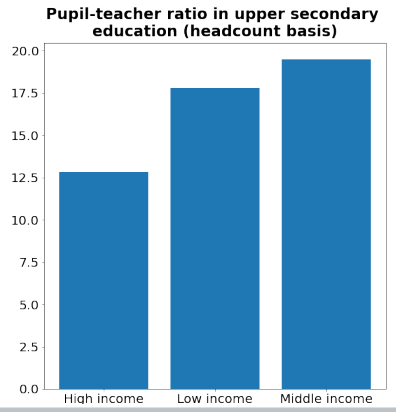


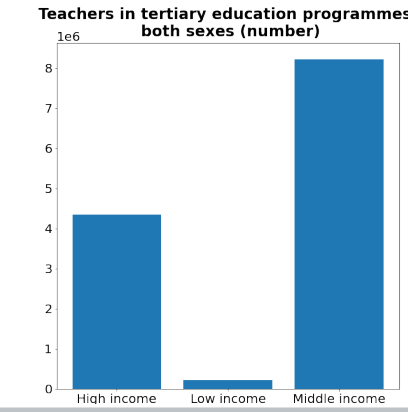


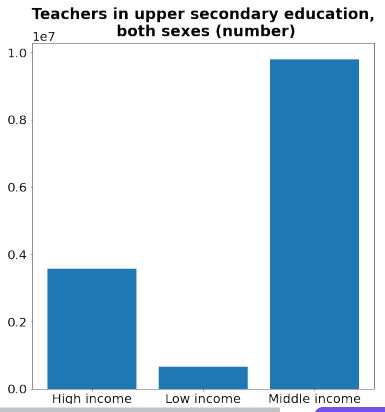




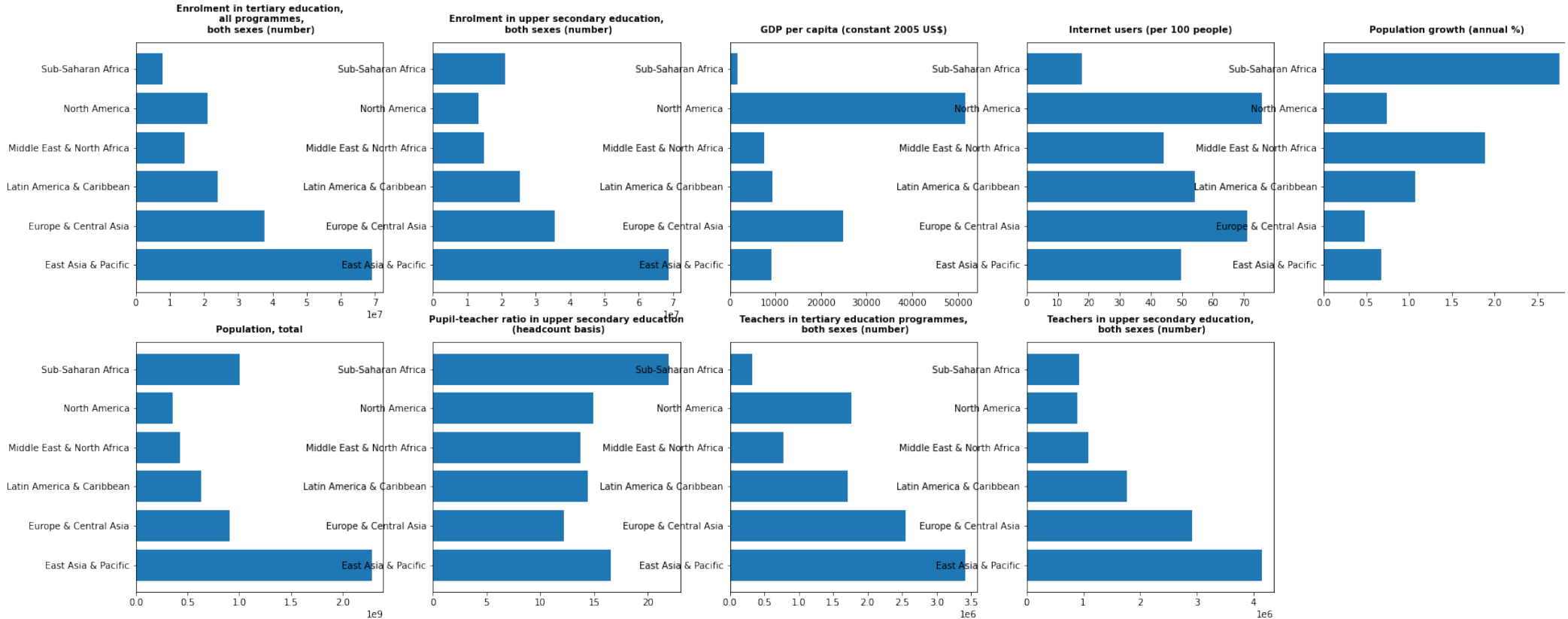








Indicateurs statistiques (2/3)
Zones géographiques



Indicateurs statistiques (3/3)
Pays

Critères communs | Critères lycéens | Critères universitaires

Indicator Name	count	mean	std	min	25%	50%	75%	max
Enrolment in tertiary education, all programmes, both sexes (number)	194.0	1.093541e+06	4.205167e+06	194.000000	19249.500000	1.713925e+05	4.793825e+05	4.336739e+07
Enrolment in upper secondary education, both sexes (number)	193.0	1.270340e+06	5.235376e+06	353.000000	33714.000000	1.791330e+05	7.306400e+05	5.522868e+07
GDP per capita (constant 2005 US$)	198.0	1.454980e+04	1.975842e+04	226.528058	1781.400727	5.971511e+03	1.697721e+04	1.076486e+05
Internet users (per 100 people)	205.0	4.852451e+01	2.852893e+01	0.000000	21.725834	5.013932e+01	7.290000e+01	9.832361e+01
Population growth (annual %)	215.0	1.319466e+00	1.230872e+00	-2.467847	0.496436	1.170823e+00	2.198431e+00	5.856170e+00
Population, total	215.0	3.409653e+07	1.338961e+08	11001.000000	839767.500000	6.234955e+06	2.297150e+07	1.371220e+09
Pupil-teacher ratio in tertiary education (headcount basis)	177.0	1.788112e+01	1.003847e+01	3.592590	10.874940	1.543690e+01	2.247714e+01	6.440642e+01
Pupil-teacher ratio in upper secondary education (headcount basis)	153.0	1.533053e+01	7.154532e+00	4.601230	10.207070	1.378374e+01	1.788640e+01	4.366559e+01
Teachers in tertiary education programmes, both sexes (number)	179.0	6.628849e+04	2.097578e+05	34.000000	1923.000000	1.033800e+04	3.271800e+04	1.606554e+06
Teachers in upper secondary education, both sexes (number)	145.0	8.560354e+04	2.798755e+05	54.000000	3665.000000	1.383200e+04	4.892663e+04	2.644952e+06

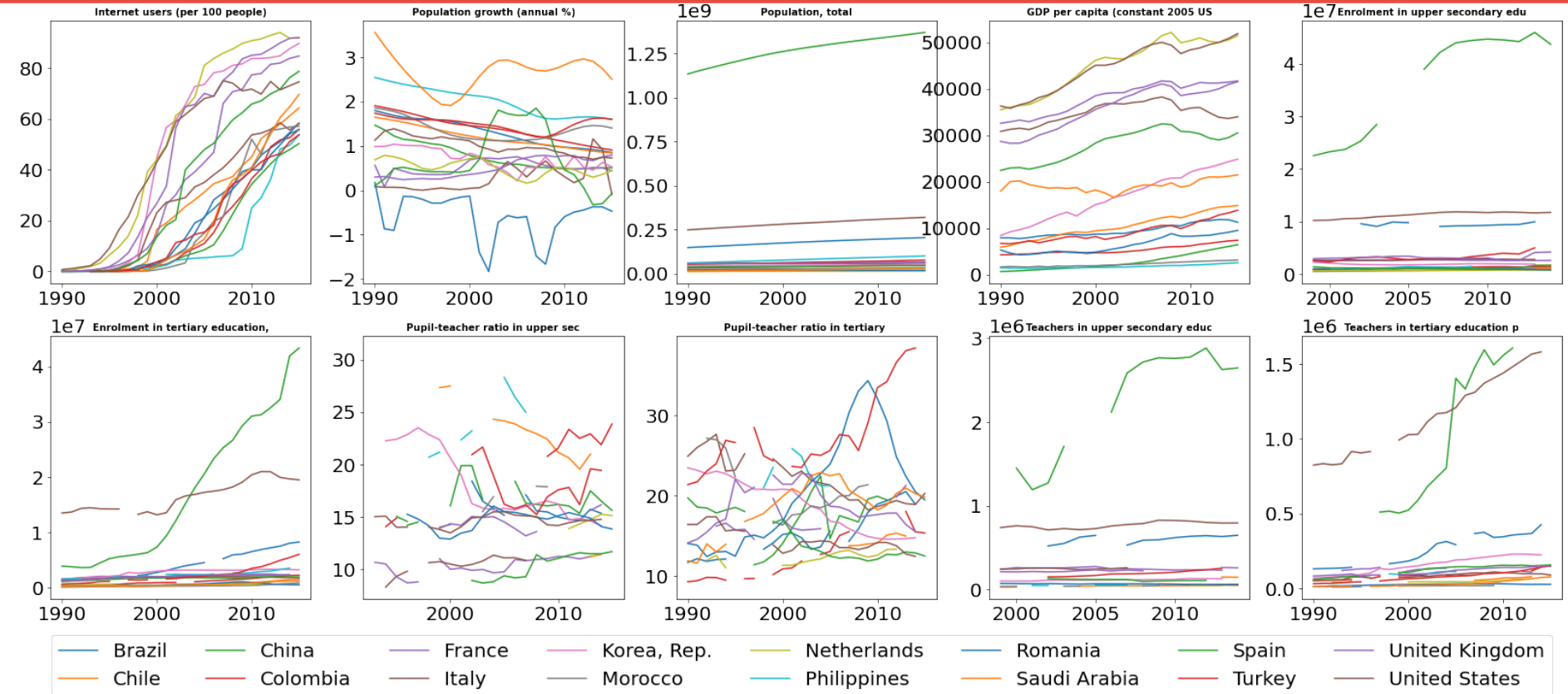
Réponses aux questions de l'entreprise (1/3)
Quels pays avec un fort potentiel client?

Pays_communs=pd.merge(Pays_Lycee,Pays_Universite)

- **20 pays pour le lycée**
  - Brazil, Chile, China, Colombia, France, Germany, Italy, Japan, Korea, Rep., Malaysia, Mexico, Morocco, Netherlands, Philippines, Romania, Saudi Arabia, Spain, Turkey, United Kingdom, United States
- **26 pays pour l'Université**
  - Argentina, Australia, Belgium, Brazil, Chile, China, Colombia, Dominican Republic, France, Greece, Italy, Kazakhstan, Korea, Rep., Malaysia, Morocco, Netherlands, Philippines, Poland, Russian Federation Romania, Saudi Arabia, South Africa, Spain, Turkey, United Kingdom, United States

- **16 pays pour le lycée et l'université**
  - Brazil, Chile, China, Colombia, France, Italy, Korea, Rep., Morocco, Netherlands, Philippines, Romania, Saudi Arabia, Spain, Turkey, United Kingdom, United States
  - 3 en Amérique du sud, 1 en Amérique du Nord, 2 au Moyen-Orient et Afrique du Nord, 3 en Asie de l'Est, 7 en Europe et Asie Centrale

Réponses aux questions de l'entreprise (2/3)
Evolution de ce potentiel pour chacun de ces pays?



Réponses aux questions de l'entreprise (3/3)
Priorité pour quels pays?

```
Pays_communs['lycee et universite']=Pays_communs['Enrolment in tertiary education, all programmes, both sexes (number)']+Pays_communs['Enrolment in upper secondary education, both sexes (number)']
Pays_communs.sort_values(by='lycee et universite', ascending=False)
```