Projet 2: Analysez des données de systèmes éducatifs

QUENTIN STEPNIEWSKI

OPENCLASSROOMS

03 JUIN 2020

Sommaire

1. Introduction

Academy – Startup de la Edtech proposant des services de formation en ligne

Problématique principale

- Expansion à l'international

Objectifs de l'étude

- Cibler les pays à fort potentiel pour une implantation
- Observer les potentielles évolutions de ces pays
- Déterminer dans quels pays opérer en priorité

2. Présentation de la base de données

Utilisation d'un dataset provenant de WorldBank.org :

"EdStats All Indicator Query" répertorie 4000 indicateurs internationaux décrivant l'accès à l'éducation

<u>Première visualisation de la base de données :</u>

```
data= pd.read_csv("EdStatsData.csv")
data.head()
```

	Country Name	Country Code	Indicator Name	Indicator Code	1970	1971	1972	1973	1974	1975	...	2060	2065	2070	2075	2080
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
1	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.F	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
2	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.GPI	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
3	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.M	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
4	Arab World	ARB	Adjusted net enrolment rate, primary, both sex...	SE.PRM.TENR	54.822121	54.894138	56.209438	57.267109	57.991138	59.36554	...	NaN	NaN	NaN	NaN	NaN

2. Présentation de la base de données

Utilisation d'un dataset provenant de WorldBank.org :

"EdStats All Indicator Query" répertorie 4000 indicateurs internationaux décrivant l'accès à l'éducation

<u>Caractéristiques de la base de données :</u>

Données représentées sur 3 axes:

- Pays
- Indicateurs
- Années

	Country Name	Country Code	Indicator Name	Indicator Code	1970	1971	1972	1973	1974	1975	...	2060	2065	2070	2075	2080
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
1	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.F	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
2	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.GPI	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
3	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.M	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
4	Arab World	ARB	Adjusted net enrolment rate, primary, both sex...	SE.PRM.TENR	54.822121	54.894138	56.209438	57.267109	57.991138	59.36554	...	NaN	NaN	NaN	NaN	NaN

2. Présentation de la base de données

Utilisation d'un dataset provenant de WorldBank.org :

"EdStats All Indicator Query" répertorie 4000 indicateurs internationaux décrivant l'accès à l'éducation

Caractéristiques de la base de données :

```
#Looking at the shape of Database
print("Caractéristiques de la base de données: \n",
      data.shape[0],"lignes \n",
      data.shape[1],"colonnes \n")
print(len(data["Country Name"].unique()), " pays différents")
print( len(data["Indicator Name"].unique()), " indicateurs différents")
```

```
Caractéristiques de la base de données:
 886930 lignes
 70 colonnes

242  pays différents
3665  indicateurs différents
```

	Country Name	Country Code	Indicator Name	Indicator Code	1970	1971	1972	1973	1974	1975	...	2060	2065	2070	2075	2080
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
1	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.F	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
2	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.GPI	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
3	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.M	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
4	Arab World	ARB	Adjusted net enrolment rate, primary, both sex...	SE.PRM.TENR	54.822121	54.894138	56.209438	57.267109	57.991138	59.36554	...	NaN	NaN	NaN	NaN	NaN

2. Présentation de la base de données

Utilisation d'un dataset provenant de WorldBank.org :

"EdStats All Indicator Query" répertorie 4000 indicateurs internationaux décrivant l'accès à l'éducation

Caractéristiques de la base de données :

On observe également un faible taux de remplissage :

```
Caractéristiques de la base de données:
 886930 lignes
 70 colonnes

242  pays différents
3665  indicateurs différents
```

```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 886930 entries, 0 to 886929
Data columns (total 69 columns):
Country Name      886930 non-null object
Country Code      886930 non-null object
Indicator Name    886930 non-null object
Indicator Code    886930 non-null object
1970              72288 non-null float64
1971              35537 non-null float64
1972              35619 non-null float64
1973              35545 non-null float64
1974              35730 non-null float64
1975              87306 non-null float64
1976              37483 non-null float64
1977              37574 non-null float64
1978              37576 non-null float64
1979              36809 non-null float64
1980              89122 non-null float64
1981              38777 non-null float64
1982              37511 non-null float64
1983              38460 non-null float64
1984              38606 non-null float64
1985              90296 non-null float64
1986              39372 non-null float64
1987              38641 non-null float64
1988              38552 non-null float64
1989              37540 non-null float64
1990              124405 non-null float64
1991              74437 non-null float64
1992              75543 non-null float64
1993              75793 non-null float64
1994              77462 non-null float64
1995              131361 non-null float64
1996              76807 non-null float64
1997              73453 non-null float64
```

3. Nettoyage de la base de données

Nettoyage de la base de données nécessaire selon trois axes :

3. Nettoyage de la base de données - Années

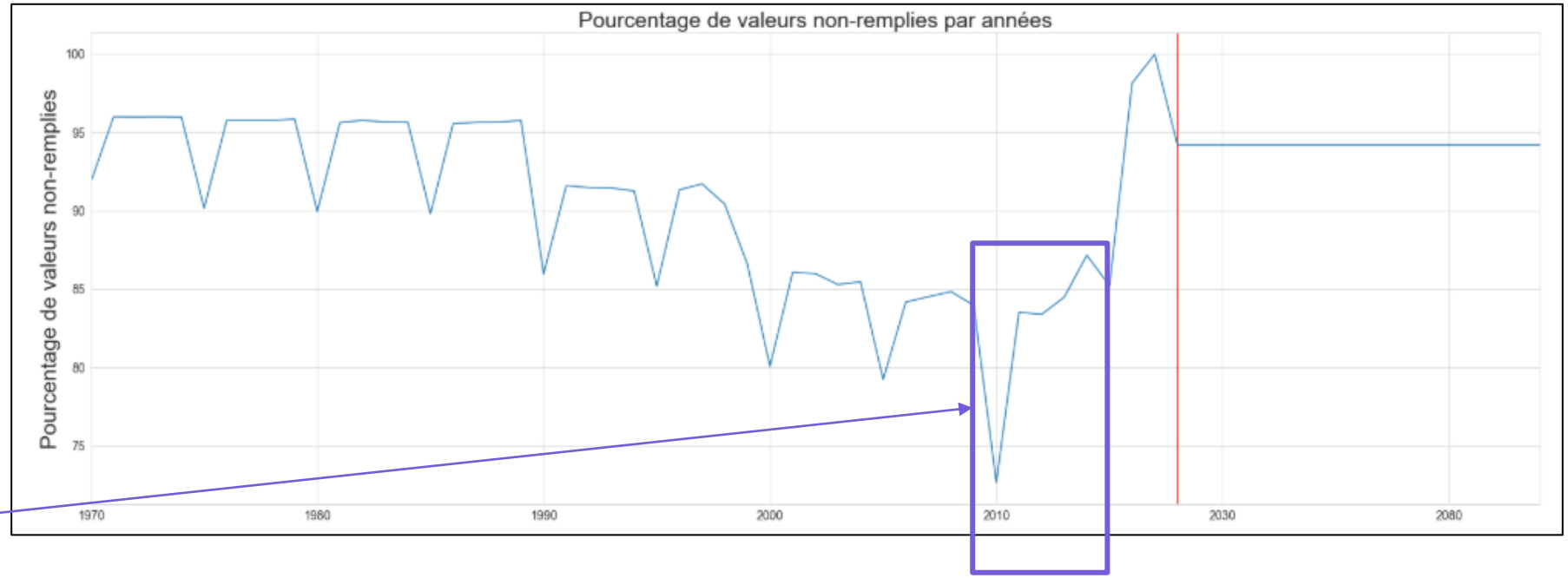
Etude du taux de remplissage par année :

```
#Display the amount of NANs in each column
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 886930 entries, 0 to 886929
Data columns (total 69 columns):
Country Name      886930 non-null object
Country Code      886930 non-null object
Indicator Name    886930 non-null object
Indicator Code    886930 non-null object
1970              72288 non-null float64
1971              35537 non-null float64
1972              35619 non-null float64
1973              35545 non-null float64
1974              35730 non-null float64
1975              87306 non-null float64
1976              37483 non-null float64
1977              37574 non-null float64
1978              37576 non-null float64
1979              36809 non-null float64
1980              89122 non-null float64
1981              38777 non-null float64
1982              37511 non-null float64
1983              38460 non-null float64
1984              38606 non-null float64
                  ...
2003              130363 non-null float64
2004              128814 non-null float64
2005              184108 non-null float64
2006              140312 non-null float64
2007              137272 non-null float64
2008              134387 non-null float64
2009              142108 non-null float64
2010              242442 non-null float64
2011              146012 non-null float64
2012              147264 non-null float64
2013              137509 non-null float64
2014              113789 non-null float64
2015              131058 non-null float64
2016              16460 non-null float64
2017              143 non-null float64
2020              51436 non-null float64
2025              51436 non-null float64
2030              51436 non-null float64
2035              51436 non-null float64
```

```
#Represent the amount of NaNs per year (showing year 2020 with a red vertical axis)
nan= data.isnull().sum(axis = 0)/lignes*100
nan.drop(nan.index[0:4],inplace=True)
```

```
g = nan.plot()
```



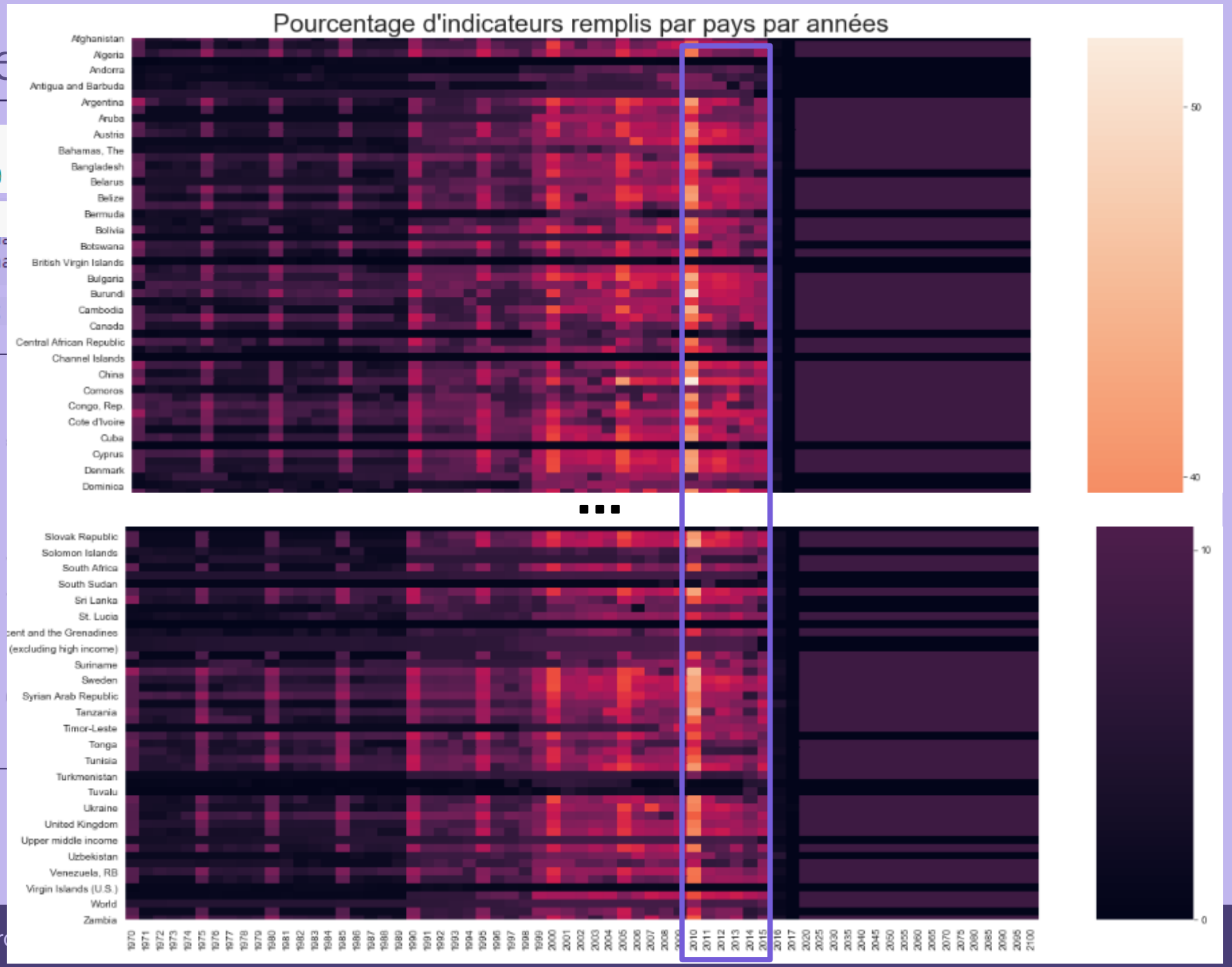
On cherche les années proches de 2020 ayant le meilleur taux de remplissage possible.

3. Nettoyage

```
# 2D view of NaNs with an heat map
data_nan = data.groupby(['Country Name']).count()/indicateurs*100
```

```
heat = sns.heatmap(data_nan)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 886930 entries, 0 to 886929
Data columns (total 69 columns):
Country Name      886930 non-null object
Country Code      886930 non-null object
Indicator Name    886930 non-null object
Indicator Code    886930 non-null object
1970              72288 non-null float64
1971              35537 non-null float64
1972              35619 non-null float64
1973              35545 non-null float64
1974              35730 non-null float64
1975              87306 non-null float64
1976              37483 non-null float64
1977              37574 non-null float64
1978              37576 non-null float64
1979              36809 non-null float64
1980              89122 non-null float64
1981              38777 non-null float64
1982              37511 non-null float64
1983              38460 non-null float64
1984              38606 non-null float64
                  ...
2003              130363 non-null float64
2004              128814 non-null float64
2005              184108 non-null float64
2006              140312 non-null float64
2007              137272 non-null float64
2008              134387 non-null float64
2009              142108 non-null float64
2010              242442 non-null float64
2011              146012 non-null float64
2012              147264 non-null float64
2013              137509 non-null float64
2014              113789 non-null float64
2015              131058 non-null float64
2016              16460 non-null float64
2017              143 non-null float64
2020              51436 non-null float64
2025              51436 non-null float64
2030              51436 non-null float64
2035              51436 non-null float64
```



3. Nettoyage de la base de données - Années

Résultat du nettoyage de la partie années :

- 6 Années sélectionnées [2010,2011,2012,2013,2014,2015]

```
#Re-index table with keeping columns from 2010 to 2015
data=data.reindex(columns=['Country Name','Country Code','Indicator Name','Indicator Code',
                           '2010','2011','2012','2013','2014','2015'])
data.head()
```

	Country Name	Country Code	Indicator Name	Indicator Code	2010	2011	2012	2013	2014	2015
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	NaN	NaN	NaN	NaN	NaN	NaN
1	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.F	NaN	NaN	NaN	NaN	NaN	NaN
2	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.GPI	NaN	NaN	NaN	NaN	NaN	NaN
3	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.M	NaN	NaN	NaN	NaN	NaN	NaN
4	Arab World	ARB	Adjusted net enrolment rate, primary, both sex...	SE.PRM.TENR	85.211998	85.24514	86.101669	85.51194	85.320152	NaN

3. Nettoyage de la base de données - Indicateurs

Etude préliminaire des indicateurs :

- Visualisation des indicateurs pour en dégager un éventuel pattern non-désiré

- On observe beaucoup d'indicateurs genrés

```
#First overview of the different indicators available
test=data["Indicator Name"].unique()
for i in range(len(test)):
    print(test[i])
```

```
Adjusted net enrolment rate, lower secondary, both sexes (%)
Adjusted net enrolment rate, lower secondary, female (%)
Adjusted net enrolment rate, lower secondary, gender parity index (GPI)
Adjusted net enrolment rate, lower secondary, male (%)
Adjusted net enrolment rate, primary, both sexes (%)
Adjusted net enrolment rate, primary, female (%)
Adjusted net enrolment rate, primary, gender parity index (GPI)
Adjusted net enrolment rate, primary, male (%)
Adjusted net enrolment rate, upper secondary, both sexes (%)
Adjusted net enrolment rate, upper secondary, female (%)
Adjusted net enrolment rate, upper secondary, gender parity index (GPI)
Adjusted net enrolment rate, upper secondary, male (%)
Adjusted net intake rate to Grade 1 of primary education, both sexes (%)
Adjusted net intake rate to Grade 1 of primary education, female (%)
Adjusted net intake rate to Grade 1 of primary education, gender parity index (GPI)
Adjusted net intake rate to Grade 1 of primary education, male (%)
Adult illiterate population, 15+ years, % female
Adult illiterate population, 15+ years, both sexes (number)
Adult illiterate population, 15+ years, female (number)
```

➢ Ces indicateurs peuvent être supprimés

```
gender = ['male','female']
data= data[~data['Indicator Name'].str.contains('|'.join(gender ), na=False)]
print('Il reste: \n',len(data["Indicator Name"].unique()), 'indicateurs')
```

```
Il reste:
 2458 indicateurs
```

➢ On a donc déjà écarté 1/3 des indicateurs

3. Nettoyage de la base de données - Indicateurs

Etude du remplissage des indicateurs :

- Pour chaque indicateur, on va observer le pourcentage de pays à avoir renseigné l'indicateur pour chaque année. On obtient le tableau suivant :

Indicator Name	2010	2011	2012	2013	2014	2015	Mean
Adjusted net enrolment rate, lower secondary, both sexes (%)	0.490741	0.495370	0.472222	0.458333	0.393519	0.013889	0.387346
Adjusted net enrolment rate, lower secondary, gender parity index (GPI)	0.476852	0.490741	0.467593	0.453704	0.388889	0.013889	0.381944
Adjusted net enrolment rate, primary, both sexes (%)	0.606481	0.611111	0.629630	0.587963	0.601852	0.518519	0.592593
Adjusted net enrolment rate, primary, gender parity index (GPI)	0.509259	0.532407	0.537037	0.486111	0.439815	0.032407	0.422840
Adjusted net enrolment rate, upper secondary, both sexes (%)	0.263889	0.263889	0.245370	0.462963	0.416667	0.013889	0.277778

- Colonne 'Mean': Pourcentage moyen de réponses des pays pour ces 6 années par indicateur.

3. Nettoyage de la base de données - Indicateurs

Etude du remplissage des indicateurs :

- On va trier les indicateurs en fonction de leur remplissage moyen sur les 6 ans.
  - On gardera uniquement les indicateurs ayant un remplissage moyen supérieur à 60% (afin d'avoir une base de données solide sur laquelle appuyer l'étude).

```
#Drop indicator with average filling level inferior to a given percentage (here 60%)
study_indicators = study_indicators.where(study_indicators['Mean']>0.60)
study_indicators.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 2458 entries, Adjusted net enrolment rate, lower secondary, both sexes (%) to You
s, gender parity index (GPI)
Data columns (total 7 columns):
2010    129 non-null float64
2011    129 non-null float64
2012    129 non-null float64
2013    129 non-null float64
2014    129 non-null float64
2015    129 non-null float64
Mean    129 non-null float64
dtypes: float64(7)
memory usage: 153.6+ KB
```

- Il reste une liste de 129 indicateurs qu'on va pouvoir étudier pour choisir les plus pertinents.

3. Nettoyage de la base de données - Indicateurs

Choix des indicateurs finaux :

- Après étude, 4 indicateurs ont été sélectionnés :

- **Nombre d'habitants** – Pour sélectionner des pays avec une taille de marché conséquente
- **PIB par habitant** – Les formations demandant un investissement financier conséquent
- **Pourcentage d'utilisateurs d'internet** – Prérequis obligatoire pour la formation en ligne
- **Pourcentage d'inscription au cycle scolaire secondaire (Collège – Lycée)** – Une base minimum de connaissances sera nécessaire pour suivre ces formations de façon efficace

3. Nettoyage de la base de données - Pays

Etude de la pertinence des pays :

- Y a-t-il des pays non-pertinents ?
- Quels sont les pays présentant pas ou trop peu de données pour les indicateurs ciblés ?

1) <u>Pertinence des pays :</u>

Après observation des 242 pays, on observe 3 catégories différentes pour ceux-ci :

- Certains sont effectivement des pays (ex: 'Algeria', 'Austria', ...)
- Certains sont des zones ou des régions du monde ('Arab World', 'European Union')
- Certains sont des groupes classés selon des critères économiques ('High Income', 'OECD members', ...)

Ici, on basera notre étude uniquement sur les pays.

```
data_country = data[data["Country Name"].isin(list_country)]
nbCountry = len(data_country["Country Name"].unique())

data = data_country
print(nbCountry," pays restants")
```

```
216  pays restants
```

3. Nettoyage de la base de données - Pays

Etude de la pertinence des pays :

- Y a-t-il des pays non-pertinents ?
- Quels sont les pays présentant pas ou trop peu de données pour les indicateurs ciblés ?

2) <u>Remplissage des indicateurs par pays :</u>

On va compter le nombre de cellules vides par indicateur par pays.

Les pays présentant moins de 2 réponses sur un indicateur seront écartés.

```
data["Nans"]= 10-data.count(axis=1)
data.head()
```

	Country Name	Country Code	Indicator Name	Indicator Code	2010	2011	2012	2013	2014	2015	Nans
92870	Afghanistan	AFG	GDP per capita (current US$)	NY.GDP.PCAP.CD	5.533003e+02	6.035370e+02	6.690091e+02	6.317450e+02	6.120697e+02	5.695779e+02	0
92960	Afghanistan	AFG	Gross enrolment ratio, secondary, both sexes (%)	SE.SEC.ENRR	5.324683e+01	5.461618e+01	5.667734e+01	5.668866e+01	5.565616e+01	5.564441e+01	0
93000	Afghanistan	AFG	Internet users (per 100 people)	IT.NET.USER.P2	4.000000e+00	5.000000e+00	5.454545e+00	5.900000e+00	7.000000e+00	8.260000e+00	0
94158	Afghanistan	AFG	Population, total	SP.POP.TOTL	2.880317e+07	2.970860e+07	3.069696e+07	3.173169e+07	3.275802e+07	3.373649e+07	0
96535	Albania	ALB	GDP per capita (current US$)	NY.GDP.PCAP.CD	4.094359e+03	4.437178e+03	4.247614e+03	4.413082e+03	4.578667e+03	3.934895e+03	0

On va donc ici retirer une **60aine de pays** et on pourra commencer à exploiter la base de données !

4. Exploitation et Analyse

La base de données présente encore une certain nombre de cellules non-remplies.

On va donc chercher à remplir les données manquantes pour avoir une analyse finale uniforme.

On utilisera deux méthodes différentes :

- Un remplissage manuel de la base de données en fonction des cellules voisines
- Un remplissage basé sur une interpolation polynomiale.

4. Exploitation et Analyse

Méthode de remplissage manuel:

Mise en place d’un script simple de parcours de la base de données :

Dès qu’une cellule non-remplie est détectée, on remplira la cellule avec les données de l’année remplie la plus proche :

4. Exploitation et Analyse

125945	Aruba	ABW	Gross enrolment ratio, secondary, both sexes (%)	SE.SEC.ENRR	9.582354e+01	9.701512e+01	1.111745e+02	NaN	NaN	NaN

Méth

Mise en place d'un script simple de parcours de la base de données :

Dès qu'une cellule n données de l'année remplie la plus proch

```
for ind in range(0,len(data_manual.index)):
    for j in range(4,10):
        if np.isnan(data_manual.iat[ind,j]):

            if (j == 4):
                for k in range (5,10):
                    if ~(np.isnan(data_manual.iat[ind,k])):
                        data_manual.iat[ind,j] = data_manual.iat[ind,k]
                        break

            else:
                data_manual.iat[ind,j] = data_manual.iat[ind,j-1]

data_manual.head(40)
```

125945	Aruba	ABW	Gross enrolment ratio, secondary, both sexes (%)	SE.SEC.ENRR	9.582354e+01	9.701512e+01	1.111745e+02	1.111745e+02	1.111745e+02	1.111745e+02

4. Exploitation et Analyse

Méthode de remplissage manuel :

Mise en place d'un script simple de parcours de la base de données :

Dès qu'une cellule non-remplie est détectée, on remplira la cellule avec sa cellule voisine la plus proche :

Retour sur la méthode :

- Peu de « risques » de dégradation du niveau des indicateurs par pays
- Pas de prise en compte de l'évolution des différents indicateurs

4. Exploitation et Analyse

Méthode de remplissage par interpolation polynomiale :

Deux fonctions testées :

- Interpolation polynomiale de degré 3
- Interpolation de Lagrange

Retour sur les méthodes :

- Interpolation polynomiale de degré 3 : peu précise (pourcentage pouvant monter très haut – 231% en 2015 pour Aruba)

	125945	Aruba	ABW	Gross enrolment ratio, secondary, both sexes (%)	SE.SEC.ENRR	9.582354e+01	9.701512e+01	1.111745e+02	1.383114e+02	1.784355e+02	2.315563e+02

- Interpolation de Lagrange : plus précise mais présente des effets de bords importants (-53% en 2015 pour le Royaume-Uni)

844285	United Kingdom	GBR	Gross enrolment ratio, secondary, both sexes (%)	SE.SEC.ENRR	1.019089e+02	9.284833e+01	9.139314e+01	1.244260e+02	1.278113e+02	-5.371875e+01

4. Exploitation et Analyse

Mise en place de l'analyse :

Mettre en place un système de note par indicateur

La première étape consiste à étudier les échelles pour chaque indicateur afin de pouvoir noter les pays de façon pertinente :

```
for ind in range(0,len(relevant_indicators)):
    print(relevant_indicators[ind],": \n",
          data_manual[data_manual['Indicator Name'] == relevant_indicators[ind]].describe()," \n---------- \n")
```

```
Internet users (per 100 people) :
              2010        2011        2012        2013        2014        2015  \
count  159.000000  159.000000  159.000000  159.000000  159.000000  159.000000
mean    35.908779   38.623681   41.620346   44.295822   47.271121   50.380565
std     27.779035   28.095951   28.557673   28.791384   28.542199   28.136822
min      0.250000    0.700000    0.800000    0.900000    0.990000    1.083733
25%     10.350000   12.290000   14.515000   15.750000   21.000000   24.750000
50%     31.800000   37.438613   41.000000   44.030000   47.400000   51.919116
75%     56.425000   61.474999   64.900000   68.211450   71.055000   73.254350
max     93.390000   94.819687   96.209800   96.546800   98.160000   98.323610
```

On obtient un tableau similaire pour **chaque indicateur**.

4. Exploitation et Analyse

Mise en place de l'analyse :

Mettre en place un système de note par indicateur

Grâce à ce tableau, on peut observer 2 problèmes principaux pour notre notation :

- Les valeurs pour les inscriptions dans le secondaire sont parfois supérieures à 100%
- Les échelles pour le PIB et le nombre d'habitants sont très grandes
  (Population Min 2015 ~10000 / Max 2015 ~1,3 milliard)

Il faudra prendre en compte ces paramètres lors de la notation :

- Mise en place d'un seuil maximal de 100% (toute valeur supérieure à 100% obtiendra la note maximale)
- Transformation de l'échelle pour le PIB et le nombre d'habitants.

4. Exploitation et Analyse

Transformation de l'échelle pour le PIB et le nombre d'habitants :

Etude des quantiles pour dégager d'éventuels outliers :

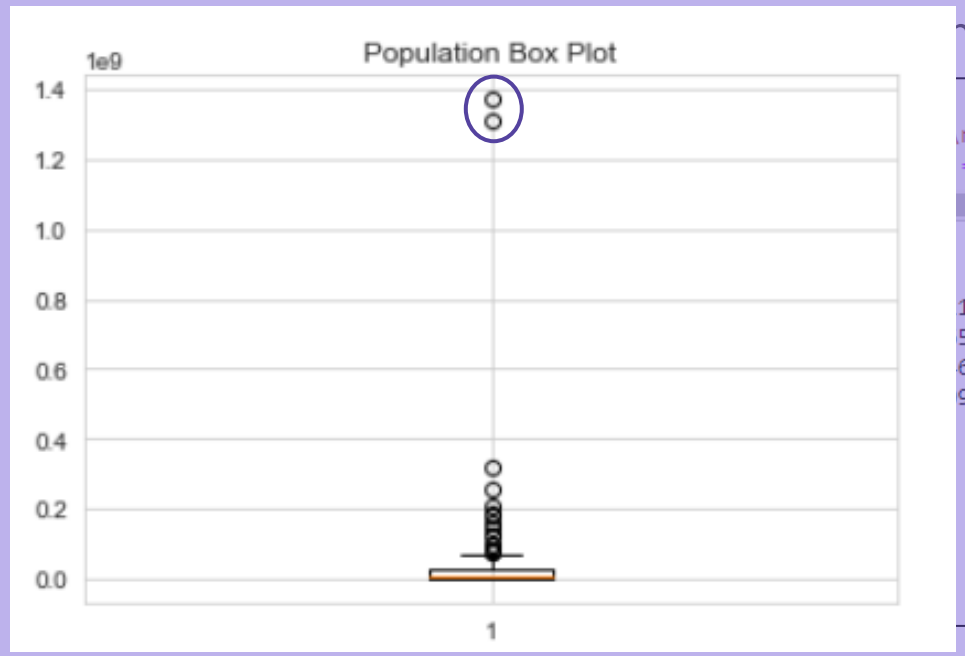
```
for ind in range(0,len(relevant_indicators)):
    print('Quantiles:',relevant_indicators[ind],": \n",
          data_manual[data_manual['Indicator Name'] == relevant_indicators[ind]].quantile([0.05,0.25,0.75,0.98],axis=0)," \n---
```

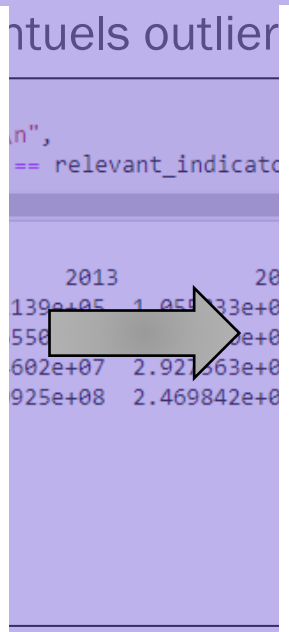
```
Quantiles: Population, total :
               2010          2011          2012          2013          2014  \
0.05  1.038902e+05  1.043246e+05  1.047136e+05  1.051139e+05  1.055833e+05
0.25  2.059661e+06  2.061938e+06  2.081982e+06  2.096550e+06  2.111640e+06
0.75  2.722441e+07  2.778258e+07  2.836814e+07  2.884602e+07  2.927563e+07
0.98  2.352077e+08  2.381842e+08  2.411517e+08  2.440925e+08  2.469842e+08

             2015          Mean
0.05     106161.7  8.996961e+04
0.25    2126976.5  1.791581e+06
0.75   29689718.5  2.453863e+07
0.98  249810112.2  2.079186e+08
----------
```

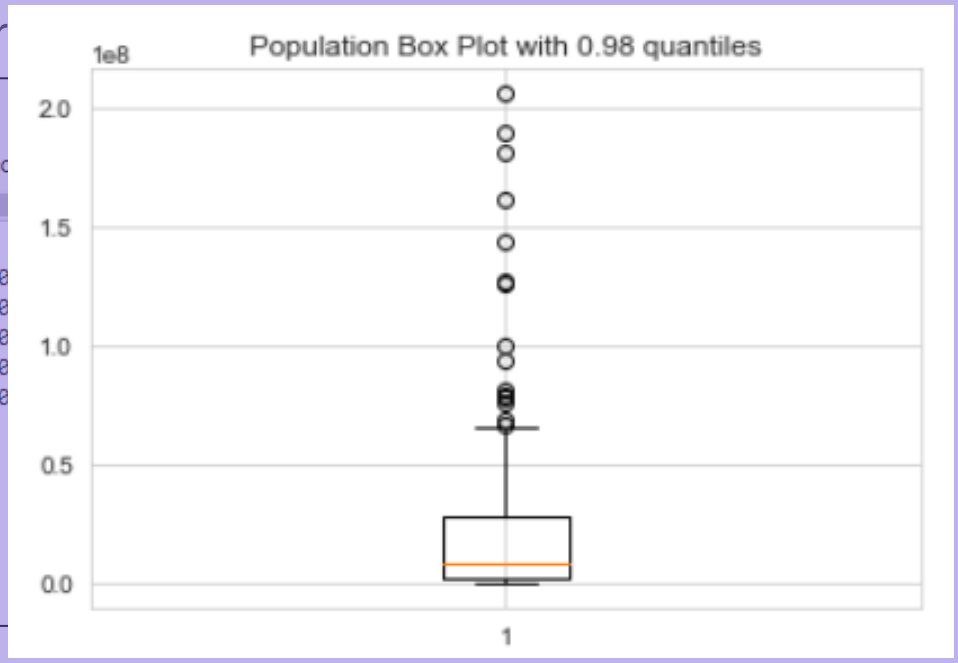
Ici on voit par exemple qu'on passe de 1,3 milliard en maximum de 2015 à 250 millions pour un quantile à 98% (98% des valeurs se trouvent en-dessous de 250 millions).

4. Exploitation et Analyse

Transformation de l'échelle pour le PIB et le nombre d'habitants :





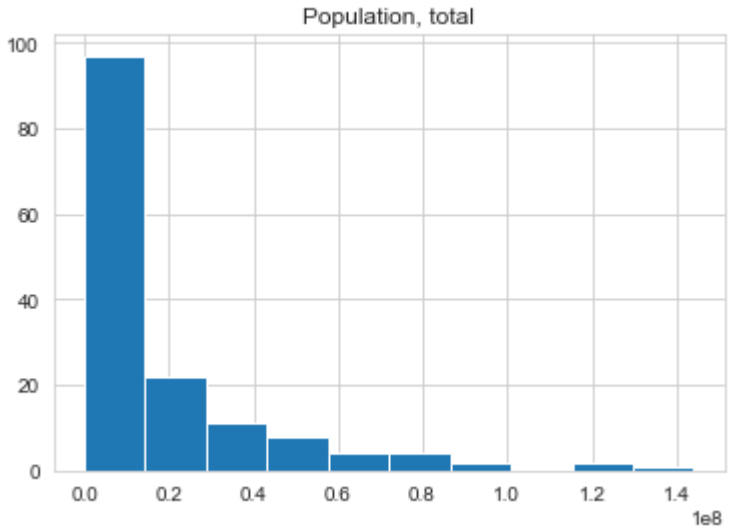


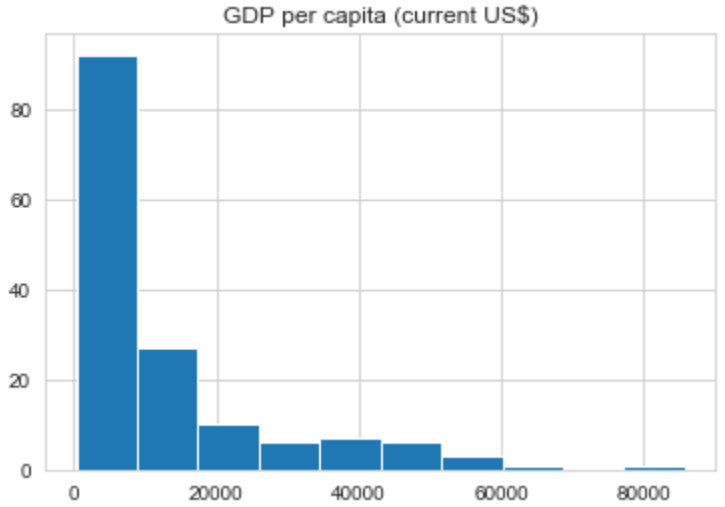
Ici on voit par exemple qu'on passe de 1,3 milliard en maximum de 2015 à 250 millions pour un quantile à 98% (98% des valeurs se trouvent en-dessous de 250 millions).

4. Exploitation et Analyse

Transformation de l'échelle pour le PIB et le nombre d'habitants :

Etude de la répartition des valeurs pour les deux indicateurs :





Ici on peut observer une répartition prenant une forme logarithmique, on pourra donc s'orienter vers une **échelle logarithmique** pour la notation de ces deux indicateurs.

4. Exploitation et Analyse

Finalisation du système de notation :

Mise en place d'une grille de notation :

On mettra de façon arbitraire une note plus élevée pour le nombre d'habitants afin de réduire les notes des pays ayant un très haut niveau de richesse et peu d'habitants.

```
notation = pd.DataFrame({'Note':[8,4,4,4]},index=relevant_indicators)
notation
```

	Note
Population, total	8
Internet users (per 100 people)	4
GDP per capita (current US$)	4
Gross enrolment ratio, secondary, both sexes (%)	4

On associe cette grille de notation à l'échelle logarithmique choisie précédemment pour la population et le PIB :

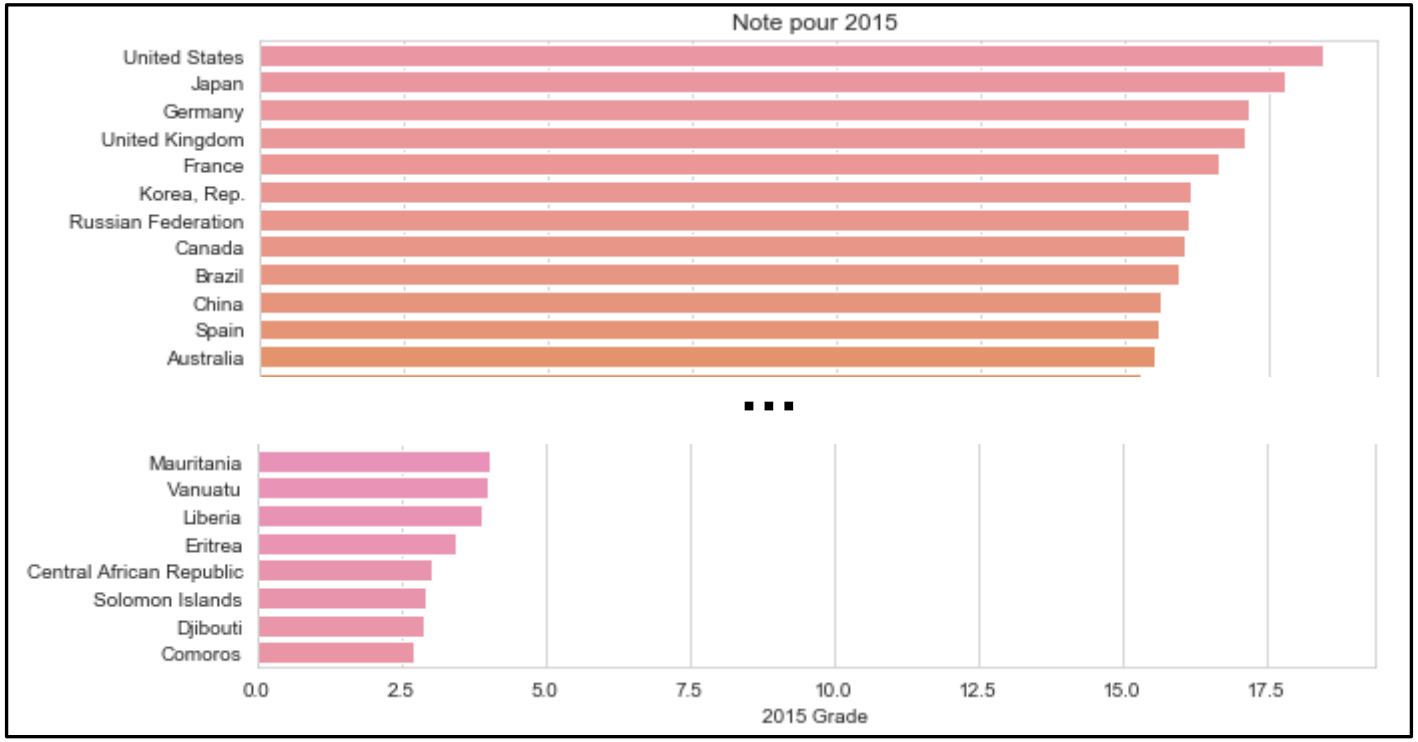
Par exemple pour la France (67 millions d'habitants) :

- 27% de la note maximale en échelle linéaire
- 76% de la note maximale en échelle logarithmique.

4. Exploitation et Analyse

Création d'un tableau de notation :

On va associer à chaque pays une note par indicateur par année et mettre en place un certain nombre de tableaux de bord via ceux-ci :

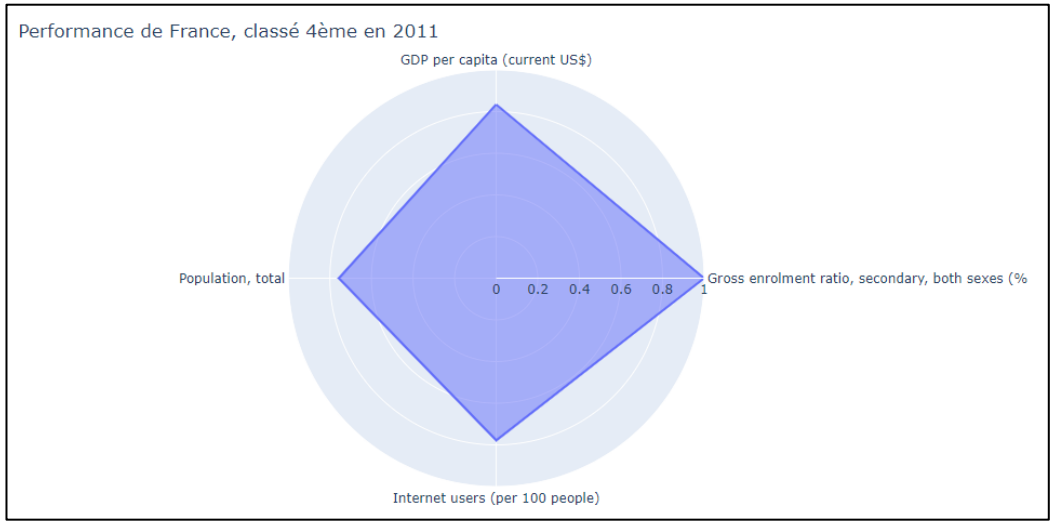


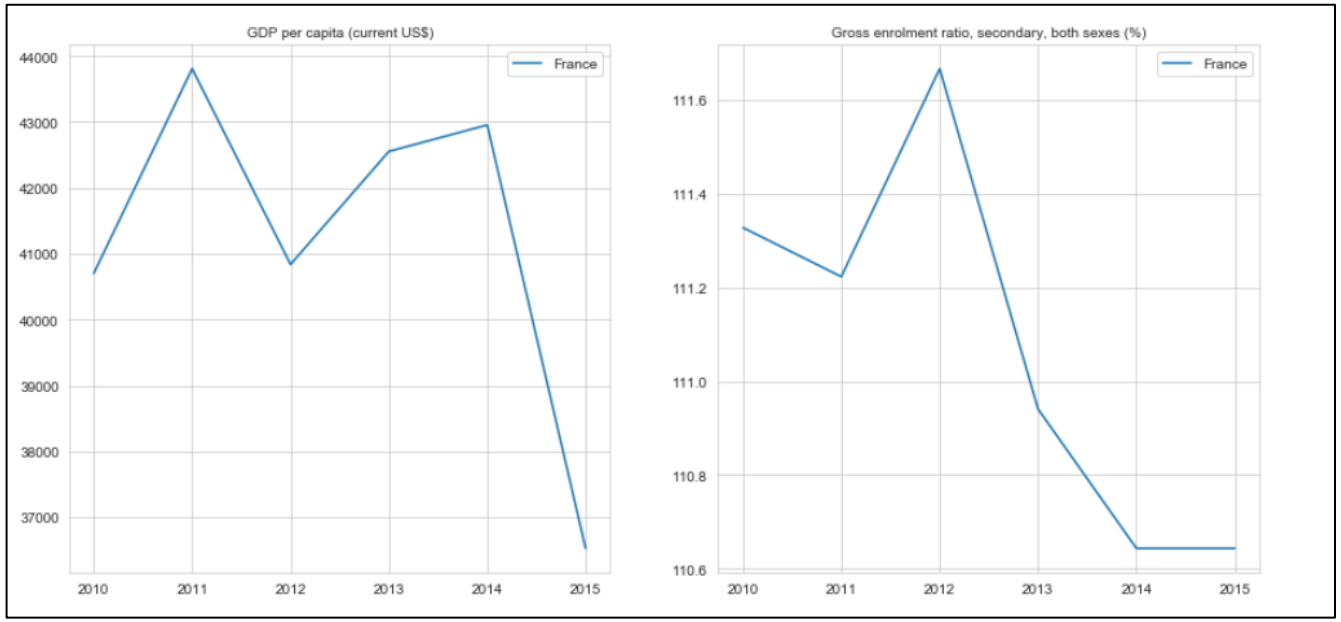
Country Name	2010 Grade	2011 Grade	2012 Grade	2013 Grade	2014 Grade	2015 Grade
Afghanistan	7.159523	7.299148	7.447201	7.513513	7.562340	7.654920
Albania	8.130209	8.454763	8.715789	8.972236	9.141534	9.103244
Algeria	10.905023	11.309891	11.492277	11.678741	11.986975	12.121745
Angola	6.943165	7.185823	7.432806	7.618741	7.703561	7.627295
Antigua and Barbuda	8.086407	8.268573	8.550924	8.743850	8.971902	9.107972
Argentina	13.245950	13.689904	13.913808	14.093455	14.249208	14.481527
Aruba	9.128439	9.493585	9.812981	10.008981	10.204181	10.399430
Australia	15.007267	15.327808	15.407134	15.611367	15.578795	15.537162
Austria	13.435162	13.637338	13.629576	13.745934	13.803489	13.827038
Bahrain	8.903032	9.996726	10.359102	10.694778	10.837027	10.953239
Bangladesh	9.434259	9.512717	9.656447	9.929796	10.309029	10.650600
Barbados	9.049347	9.140664	9.324711	9.337466	9.468338	9.499102
Belarus	10.118950	10.499199	10.843550	11.257305	11.489408	11.526968
Belgium	13.809184	14.157416	14.075741	14.176457	14.308012	14.177312

4. Exploitation et Analyse

Création d'un tableau de notation :

On va associer à chaque pays une note par indicateur par année et mettre en place un certain nombre de tableaux de bord via ceux-ci :





5. Conclusion et Perspectives

On peut finalement créer un classement des pays par année :

Ce tableau peut servir de base de travail pour orienter l'implantation à l'international.

Pour cela, on peut prendre plusieurs directions possibles :

- S'orienter sur les top countries directement (par exemple les 5 pays ayant le meilleur classement sur ces 6 années)
- S'orienter vers les pays de cette liste ayant une langue commune
- Suivre les pays présentant une forte évolution sur ces années.

	2010	2011	2012	2013	2014	2015
1	United States	United States	United States	United States	United States	United States
2	Japan	Japan	Japan	Japan	Japan	Japan
3	Germany	Germany	Germany	Germany	Germany	Germany
4	United Kingdom	France	France	United Kingdom	United Kingdom	United Kingdom
5	France	United Kingdom	United Kingdom	France	France	France
6	Canada	Canada	Russian Federation	Russian Federation	Russian Federation	Korea, Rep.
7	Korea, Rep.	Korea, Rep.	Canada	Canada	Canada	Russian Federation
8	Brazil	Brazil	Korea, Rep.	Brazil	Brazil	Canada
9	Italy	Australia	Brazil	Korea, Rep.	Korea, Rep.	Brazil
10	Spain	Italy	Australia	Australia	Spain	China
11	Netherlands	Spain	Spain	Italy	Australia	Spain
12	Australia	Netherlands	Italy	Spain	China	Australia
13	Russian Federation	Russian Federation	Netherlands	China	Italy	Italy
14	Sweden	China	China	Netherlands	Netherlands	Netherlands
15	China	Sweden	Sweden	Sweden	Sweden	Mexico
16	Norway	Switzerland	Switzerland	Turkey	Turkey	Turkey
17	Switzerland	Belgium	Norway	Switzerland	Switzerland	Argentina
18	Poland	Norway	Belgium	Mexico	Mexico	Saudi Arabia
19	Belgium	Poland	Mexico	Saudi Arabia	Saudi Arabia	Switzerland
20	Denmark	Mexico	Poland	Belgium	Belgium	Sweden

5. Conclusion et Perspectives

On peut finalement créer un classement des pays par année :

Ce tableau peut servir de base de travail pour orienter l'implantation à l'international.

Pour cela, on peut prendre plusieurs directions possibles :

- S'orienter sur les top countries directement (par exemple les 5 pays ayant le meilleur classement sur ces 6 années)
- S'orienter vers les pays de cette liste ayant une langue commune
- Suivre les pays présentant une forte évolution sur ces années.

	2010	2011	2012	2013	2014	2015
1	United States	United States	United States	United States	United States	United States
2	Japan	Japan	Japan	Japan	Japan	Japan
3	Germany	Germany	Germany	Germany	Germany	Germany
4	United Kingdom	France	France	United Kingdom	United Kingdom	United Kingdom
5	France	United Kingdom	United Kingdom	France	France	France
6	Canada	Canada	Russian Federation	Russian Federation	Russian Federation	Korea, Rep.
7	Korea, Rep.	Korea, Rep.	Canada	Canada	Canada	Russian Federation
8	Brazil	Brazil	Korea, Rep.	Brazil	Brazil	Canada
9	Italy	Australia	Brazil	Korea, Rep.	Korea, Rep.	Brazil
10	Spain	Italy	Australia	Australia	Spain	China
11	Netherlands	Spain	Spain	Italy	Australia	Spain
12	Australia	Netherlands	Italy	Spain	China	Australia
13	Russian Federation	Russian Federation	Netherlands	China	Italy	Italy
14	Sweden	China	China	Netherlands	Netherlands	Netherlands
15	China	Sweden	Sweden	Sweden	Sweden	Mexico
16	Norway	Switzerland	Switzerland	Turkey	Turkey	Turkey
17	Switzerland	Belgium	Norway	Switzerland	Switzerland	Argentina
18	Poland	Norway	Belgium	Mexico	Mexico	Saudi Arabia
19	Belgium	Poland	Mexico	Saudi Arabia	Saudi Arabia	Switzerland
20	Denmark	Mexico	Poland	Belgium	Belgium	Sweden

5. Conclusion et Perspectives

On peut finalement créer un classement des pays par année :

Ce tableau peut servir de base de travail pour orienter l'implantation à l'international.

Pour cela, on peut prendre plusieurs directions possibles :

- S'orienter sur les top countries directement (par exemple les 5 pays ayant le meilleur classement sur ces 6 années)
- S'orienter vers les pays de cette liste ayant une langue commune
- Suivre les pays présentant une forte évolution sur ces années.

	2010	2011	2012	2013	2014	2015
1	United States	United States	United States	United States	United States	United States
2	Japan	Japan	Japan	Japan	Japan	Japan
3	Germany	Germany	Germany	Germany	Germany	Germany
4	United Kingdom	France	France	United Kingdom	United Kingdom	United Kingdom
5	France	United Kingdom	United Kingdom	France	France	France
6	Canada	Canada	Russian Federation	Russian Federation	Russian Federation	Korea, Rep.
7	Korea, Rep.	Korea, Rep.	Canada	Canada	Canada	Russian Federation
8	Brazil	Brazil	Korea, Rep.	Brazil	Brazil	Canada
9	Italy	Australia	Brazil	Korea, Rep.	Korea, Rep.	Brazil
10	Spain	Italy	Australia	Australia	Spain	China
11	Netherlands	Spain	Spain	Italy	Australia	Spain
12	Australia	Netherlands	Italy	Spain	China	Australia
13	Russian Federation	Russian Federation	Netherlands	China	Italy	Italy
14	Sweden	China	China	Netherlands	Netherlands	Netherlands
15	China	Sweden	Sweden	Sweden	Sweden	Mexico
16	Norway	Switzerland	Switzerland	Turkey	Turkey	Turkey
17	Switzerland	Belgium	Norway	Switzerland	Switzerland	Argentina
18	Poland	Norway	Belgium	Mexico	Mexico	Saudi Arabia
19	Belgium	Poland	Mexico	Saudi Arabia	Saudi Arabia	Switzerland
20	Denmark	Mexico	Poland	Belgium	Belgium	Sweden

Slides additionnelles

Retour sur la méthode de notation

$$log_{maxi}\left(\frac{indicatorvalue}{divider}\right) * note$$

	Note	maxi	indType	divider
Population, total	8	249.810112	1	1000000
Internet users (per 100 people)	4	100.000000	2	0
GDP per capita (current US$)	4	92.860121	1	1000
Gross enrolment ratio, secondary, both sexes (%)	4	100.000000	2	0

Exemple pour la France :

$$log_{249,81}\left(\frac{67000000}{1000000}\right) * 8$$

$$log_{249,81}(67) * 8$$

$$0{,}76 * 8 = 6{,}08$$

```
if ind_type == 1:
    for j in range(11,17):
        if data_manual.iat[ind,j-7]>divider:

            # if the indicator value is not higher than our scale maximum scale
            # we give a grade according to the logarithmic scale (example with France)
            # Log_base249(67000000/1000000)*8 = 6.08

            if (data_manual.iat[ind,j-7]<(divider*maxi)):
                data_manual.iat[ind,j]= note * math.log(data_manual.iat[ind,j-7]/divider,maxi)
            else:
                data_manual.iat[ind,j]= note

        else:
            data_manual.iat[ind,j]=0
else:
    for j in range(11,17):
        if data_manual.iat[ind,j-7]<=100:
            data_manual.iat[ind,j]=data_manual.iat[ind,j-7]*note/100
        else:
            data_manual.iat[ind,j]=note
```

Slides additionnelles

Méthode de remplissage par interpolation polynomiale :

Deux fonctions testées :

- Interpolation polynomiale de degré 3
- Interpolation de Lagrange

```
def prediction(row):

    global data_interpolate

    x = [year for year in range(2010,2016) if not np.isnan(row.loc[str(year)])]
    y = [row.loc[str(year)] for year in range(2010,2016) if not np.isnan(row.loc[str(year)])]
#     f = interpolate.interp1d(x,y,bounds_error =False,fill_value ="extrapolate")
#     f = interpolate.lagrange(x,y)
    coefs = np.polyfit(x,y,3)

    for year in range(2010,2016):
        if np.isnan(row.loc[str(year)]):
#             print('Row:',row.name,'Nan trouvé :',row)
#             data_interpolate.loc[row.name,str(year)]=f(year)
            data_interpolate.loc[row.name,str(year)]=np.polyval(coefs,year)
```

```
data_interpolate.apply(prediction,axis=1)
```

Slides additionnelles

```
# Final tool of the study, here the goal is to have an overview of a country with update only the 2 first variables :
# my_country and year_studied


my_country = 'France'
#choose between following values : 2010,2011,2012,2013,2014,2015
year_studied =2011
studied = str(year_studied)+' Grade'

graph_table = data_manual.reindex(columns=['Country Name','Indicator Name','2010','2011','2012','2013','2014','2015'])
print('Représentation graphique des indicateurs pour le pays sélectionnée: ',my_country,"\n --------- \n")
fig, axs = plt.subplots(2,2, figsize=(15, 15), facecolor='w', edgecolor='k')
axs = axs.ravel()

# Using a for loop to create 1 graph per indicator in the 'relevant_indicator' list
for ind in range(0,len(relevant_indicators)):
    graph_table = data_manual.reindex(columns=['Country Name','Indicator Name','2010','2011','2012','2013','2014','2015'])
    graph_table = graph_table.loc[graph_table['Indicator Name']==relevant_indicators[ind]]
    graph_table = graph_table.loc[graph_table['Country Name']==my_country]
    graph_table = graph_table.drop(columns=['Indicator Name'])
    graph_table = graph_table.set_index('Country Name')
    graph_table = graph_table.transpose()

    sns.lineplot(data=graph_table,ax=axs[ind])
    axs[ind].set_title(relevant_indicators[ind],fontsize=10)

# This part is used to represent the chosen country performances with a radar plot

data_radar = data_manual.copy()
data_radar = data_radar.reindex(columns=['Country Name','Indicator Name',studied])
data_radar = data_radar.loc[data_radar['Country Name']==my_country]

# We want the graph to display a pourcentage of the grade for each indicator
# so that the radar graph can be obersved without knowing what is the maximum grade for each indicator
for i in range(len(relevant_indicators)):
    data_radar.iat[i,2] /= notation.at[data_radar.iat[i,1],'Note']

# This part is used to be able to display the rank of the country in the title
ranking = data_manual.reindex(columns=['Country Name','2010 Grade','2011 Grade','2012 Grade','2013 Grade','2014 Grade','2015 Gra
ranking = ranking.groupby(['Country Name']).sum()
ranking['rank']=ranking[studied].rank(ascending=False)
ranking = ranking.astype({"rank": int})

fig = px.line_polar(data_radar, r=studied, theta='Indicator Name', line_close=True)

fig.update_layout(title="Performance de {}, classé {}ème en {}".format(my_country,ranking.loc[my_country,'rank'],year_studied))
fig.update_traces(fill='toself')
fig.show()
```