Projet 2 : Analysez des données de systèmes éducatifs

Raphaël GIRAUDOT

Sommaire :

1. Contexte et objectifs de la mission

- Contexte :
  - Data Scientist pour la Start-Up **“Academy”**
  - Formation en ligne niveau Lycée & Université
  - Projet d’expansion à l’international
- Objectifs :
  - Analyse exploratoire
  - Les données de la banque mondiale à disposition sont-elle utiles ?
  - Quels sont les pays à fort potentiel ?
  - Quelle évolution du potentiel pour ces pays ?
  - Dans quels pays la Start-Up doit elle opérer en priorité ?

2. Préparation et outils de travail

- Traitements en Python 3 via Notebook Jupyter

- Bibliothèques Python :
  - Matplotlib
  - Pandas
  - Numpy
  - Missingno (non inclue dans Anaconda 3)
  - Seaborn

Importation des bibliothèques dans le notebook :

```
%matplotlib inline
import pandas as pd
import matplotlib.pyplot as plt
import missingno as msno
import numpy as np
import seaborn as sns
```

- Installation de la bibliothèque “missingno” pour la visualisation des données manquantes :
  pip install missingno (dans la console python)

- Importation des fichiers .csv à disposition :

```
dataCountry = pd.read_csv("Data/EdStatsCountry.csv")
dataCS = pd.read_csv("Data/EdStatsCountry-Series.csv")
data = pd.read_csv("Data/EdStatsData.csv")
dataFN = pd.read_csv("Data/EdStatsFootNote.csv")
dataSeries = pd.read_csv("Data/EdStatsSeries.csv")
```

Possibilité de voir la version des bibliothèques :

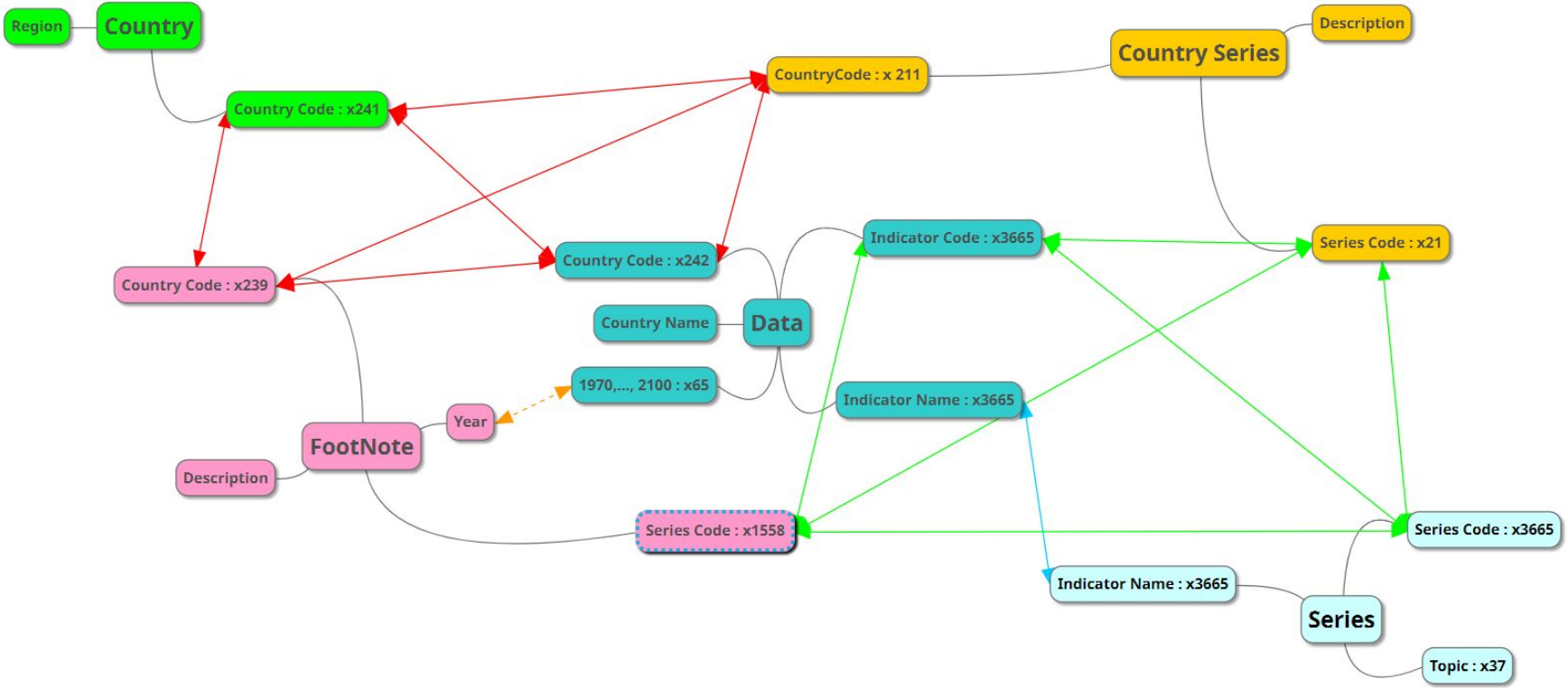
```
Entrée [2]: sns.__version__
Out[2]: '0.11.0'
```

3. Récapitulatif des données à disposition

5 Fichiers :

- EdStatsCountry.csv : Informations générales liées aux Pays
  - Individus : *Pays*
- EdStatsCountry-Series.csv : Renseignements sur des indicateurs généraux liés aux Pays
  - Individus : Combinaison d'un *indicateur "général"* et d'un *pays*
- EdStatsData.csv : Valeurs des indicateurs pour chaques pays pour les années de 1970 à 2100
  - Individus : Combinaison d'un *indicateur* et d'un *pays*
- EdStatsFootNote.csv : Informations complémentaires pour les observations des indicateurs des pays en fonctions de l'année d'observation
  - Individus : Combinaison d'un *indicateur*, d'un *pays* et d'une *année*
- EdStatsSeries.csv : Informations complémentaires sur les indicateurs
  - Individus : *Indicateurs*

4. Relations entre les fichiers



5. Présélection d'indicateurs

Variables :

- Country Code
- Short Name
- Région
- Income Group

Indicateurs "Éducation" :

25 indicateurs sur 7 sujets :

- Indicateurs Historiques :
  1. Dépenses publiques (x3)
  2. Evaluation du niveau, PIAAC (x2)
  3. Taux de scolarisation (x3)
  4. % +15 ans au Lycée/Université (x4)
  5. % +25 ans par plus haut niveau d'éducation (x3)
  6. Professeurs (x8)
- Indicateurs de Projections :
  7. Plus haut-niveau d'éducation atteint (x2)

Indicateurs "Généraux" :

- % Utilisateurs d'Internet
- Population Totale

6. Tableau historique

Création du tableau :

```
#on ajoutes les colonnes "Code Indicateur + max year" et "Code Indicateur + last value" pour chaque indicateur sélectionnés
#à notre DataFrame de base
for codeIndic in tabIndic:
        colMy = codeIndic + ' max year'
        colLv = codeIndic + ' last value'
        df_temp = data[data['Indicator Code'] == codeIndic].reset_index(drop=True)
        df_temp[colMy] =  df_temp[years].apply(lambda x: x[x.notnull()].index[-1] if len(x[x.notnull()])>0 else np.NAN, axis=1)
        df_temp[colLv] = df_temp[years].apply(lambda x: x[x.notnull()].iloc[-1] if len(x[x.notnull()])>0 else np.NAN, axis=1)
        workCountryHist = workCountryHist.merge(df_temp[['Country Code', colMy, colLv]], on="Country Code", how='left')
```

Extrait du tableau :

	Country Code	Short Name	Region	Income Group	UIS.XGDP.3.FSGOV max year	UIS.XGDP.3.FSGOV last value	UIS.XGDP.4.FSGOV max year	UIS.XGDP.4.FSGOV last value	UIS.XGDP.56.FSGOV max year	UIS.XGDP.56.FSGOV last value
0	ABW	Aruba	Latin America & Caribbean	High income: nonOECD	NaN	NaN	NaN	NaN	2005	0.53760
1	AFG	Afghanistan	South Asia	Low income	NaN	NaN	NaN	NaN	NaN	NaN
2	AGO	Angola	Sub-Saharan Africa	Upper middle income	NaN	NaN	NaN	NaN	2006	0.19851
3	ALB	Albania	Europe & Central Asia	Upper middle income	NaN	NaN	NaN	NaN	2013	0.77585

Autre méthode, (moins optimale):

```python
#fonction qui renvoie le numéro de la colonne ayant les données les plus récentes :
def indVal(ligne):
    for i in range(67,3,-1):
        if not(pd.isna(ligne.iloc[0,i])):
            return i
#fonction qui renvoie un dataframe avec 3 colonnes le 'Country Code', 'la valeur de l'indicateur', 'l'année de l'indicateur'
#passé en paramètre

#paramètres : le dataframe 'historique', le dataframe des données, le code de l'indicateur que l'on veut enregistrer
def ajoutIndicHist(tabHist, tabData, codeIndic):
    years = tabData.columns
    colV = codeIndic + ' Valeur'
    colA = codeIndic + ' Année'
    tAjout=pd.DataFrame(columns=['Country Code',colV, colA])
    i=0
    for codePays in tabHist['Country Code']:
        lPays = tabData[(tabData["Indicator Code"]==codeIndic) & (tabData["Country Code"]==codePays)]
        if indVal(lPays) == None:
            tAjout.loc[i,colV] = np.NaN
            tAjout.loc[i,colA] = np.NaN
            tAjout.loc[i,'Country Code'] = codePays
        else:
            ans = years[indVal(lPays)]
            val = lPays.iloc[0,indVal(lPays)]
            tAjout.loc[i,colV] = val
            tAjout.loc[i,colA] = ans
            tAjout.loc[i,'Country Code'] = codePays
        i=i+1
    return tAjout
```

7. Qualité des données

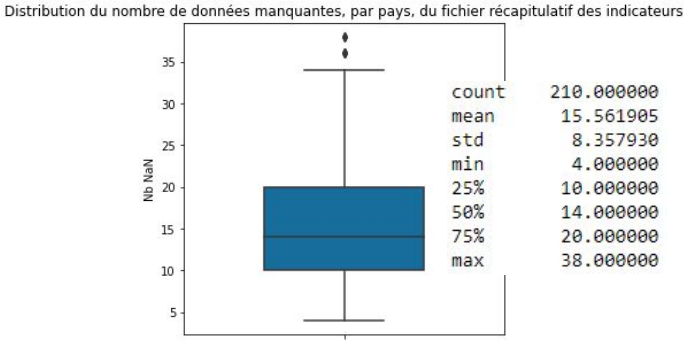
1. Retrait du tableau des indicateurs présentant 100% de données manquantes : (4x2) indicateurs

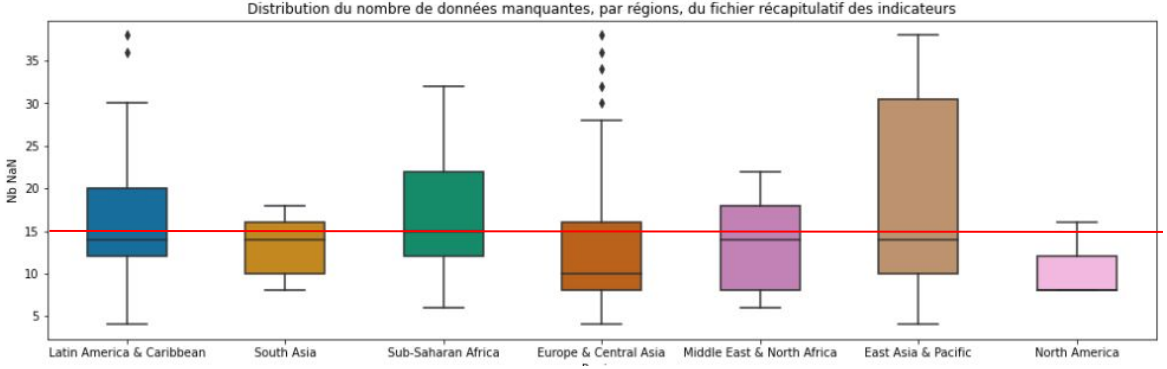
	Nb Null	% Null
UIS.TATTRR.3.T last value	214	100.000000
UIS.ASTAFF.6T8 last value	214	100.000000
UIS.XGDP.3.FSGOV max year	214	100.000000
UIS.XGDP.3.FSGOV last value	214	100.000000
UIS.T.5 last value	214	100.000000
UIS.ASTAFF.6T8 max year	214	100.000000
UIS.TATTRR.3.T max year	214	100.000000
UIS.T.5 max year	214	100.000000

2. Retrait du tableau des pays présentant 40 données manquantes :

	Country Code	Short Name	Region	Nb NaN
87	IMN	Isle of Man	Europe & Central Asia	40
209	XKX	Kosovo	Europe & Central Asia	40
118	MAF	St. Martin (French part)	Latin America & Caribbean	40
34	CHI	Channel Islands	Europe & Central Asia	40

3. Nombres de données manquantes, par pays, pour (20x2) indicateurs :





8. Affinage des indicateurs & Sélection des pays 1/2

Choix de quelques indicateurs parmis la 1ère sélection :

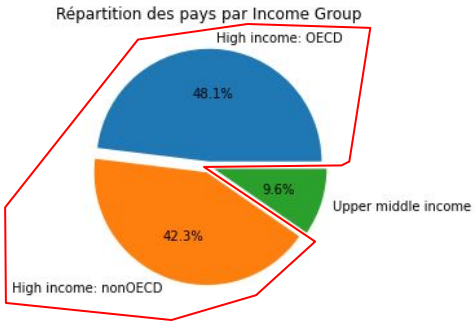
- % d'utilisateurs internet :
  - Formation en ligne
- Income Group :
  - Capacité à pouvoir acheter une formation
- %+15 ans au Lycée et + :
  - Proportion de la population "ciblable" : formation de niveau Lycée, au moins
- %+15 ans à l'Université :
  - Proportion de la population "ciblable" : formation de niveau Universitaire

8. Affinage des indicateurs & Sélection des pays 2/2

Variables	% d'utilisateurs Internet	→	Income Group	→	%+15 ans au Lycée et +	→	%+15 ans à l'Université

Valeurs

	% d'utilisateurs Internet	%+15 ans au Lycée et +	%+15 ans à l'Université
count	207.000000	36.000000	18.000000
mean	50.195395	56.704722	21.887222
std	28.959319	13.871902	6.908480
min	0.000000	31.090000	4.630000
25%	25.036652	45.262500	18.117500
50%	53.200000	58.550000	23.880000
75%	76.005180	64.497500	26.180000
max	98.240016	85.750000	30.980000



Années

	% d'utilisateurs Internet	%+15 ans au Lycée et +	%+15 ans à l'Université
count	207.000000	36.0	18.0
mean	2015.516908	2010.0	2010.0
std	3.242385	0.0	0.0
min	1990.000000	2010.0	2010.0
25%	2016.000000	2010.0	2010.0
50%	2016.000000	2010.0	2010.0
75%	2016.000000	2010.0	2010.0
max	2016.000000	2010.0	2010.0

Nb Pays sélectionnés	52	→	47	→	18	→	9

: Seuil de sélection (≥)

9. Pays sélectionnés

- 9 Pays (8 Europe + Australie)

Classement selon “SE.TER.TCHR last value” :

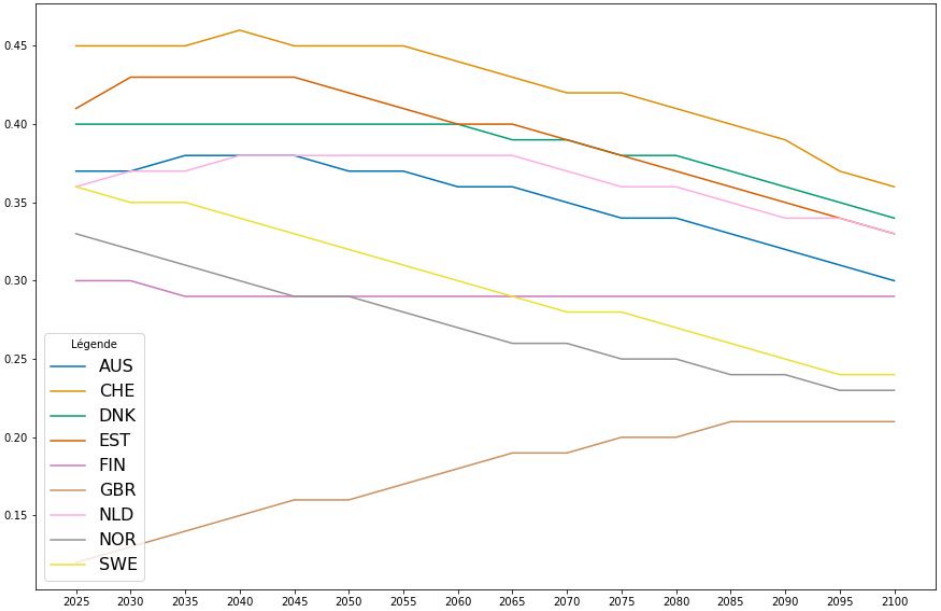
C’est à dire le nombre de professeurs dans l’enseignement tertiaire en proportion de la population totale (du - élevé au + élevé)

	Country Code	Short Name	Region
0	AUS	Australia	East Asia & Pacific
1	GBR	United Kingdom	Europe & Central Asia
2	FIN	Finland	Europe & Central Asia
3	SWE	Sweden	Europe & Central Asia
4	NLD	Netherlands	Europe & Central Asia
5	CHE	Switzerland	Europe & Central Asia
6	EST	Estonia	Europe & Central Asia
7	NOR	Norway	Europe & Central Asia
8	DNK	Denmark	Europe & Central Asia

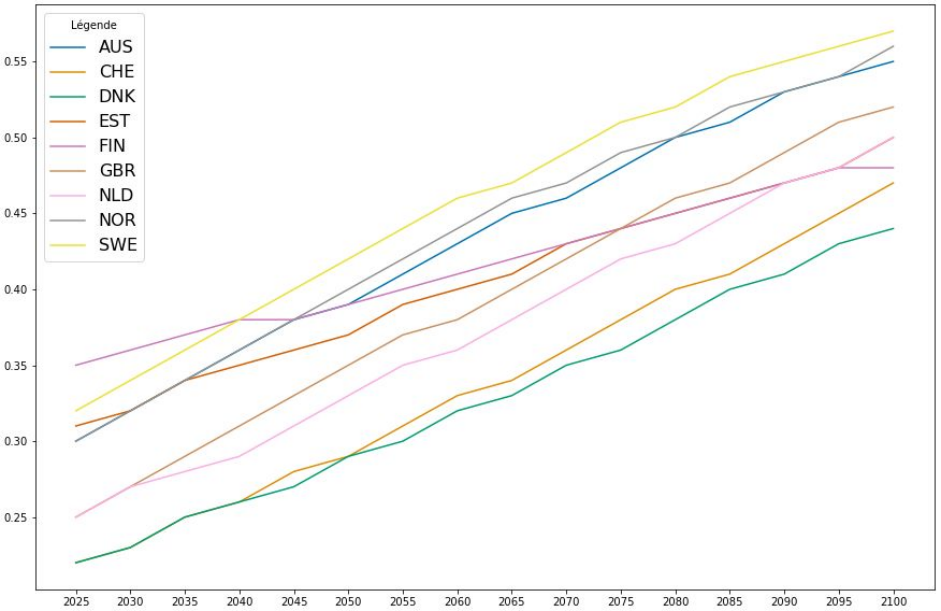
10. Projections

Wittgenstein Projection : Pourcentage de la population totale par niveau d'instruction le plus élevée

Upper Secondary (Lycée) : PRJ.ATT.ALL.3.MF



Post Secondary (Post-Bac) : PRJ.ATT.ALL.4.MF



11. Conclusion

Bon potentiel à long terme pour tous les pays :

Population qui va avoir tendance, avec le temps, à suivre de plus en plus des études supérieures (Post-Bac)

Priorisation:

1. Suisse : Utilisation du contenu FR → Utilisation du contenu existant
2. Pays Anglophone : Toucher plusieurs pays avec le même contenu
3. Zone Euro : Même zone économique et même monnaie