

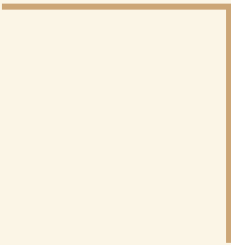
“Analysez des données de systèmes éducatifs”

Projet n°2

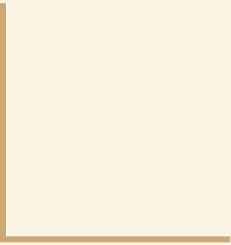
Léa ZADIKIAN
Parcours Data Scientist
7 Octobre 2022

Analysez des données de systèmes éducatifs

- I. Présentation de la problématique et du jeu de données
- II. Filtrage des données
- III. Analyse et pertinence du jeu de données



I. Présentation de la problématique et du jeu de données



Le projet d'expansion à l'international *d'academy*



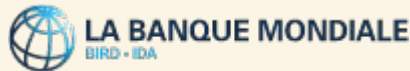
- **academy** : start-up membre de la EdTech.
- Propose des contenus de formation en ligne de niveau lycée et université.

⇒ **Son projet** : expansion des ses services à l'international.

- Quels sont les pays avec un fort potentiel de clients ?
- Pour chacun de ces pays, quelle sera l'évolution de ce potentiel de clients ?
- Dans quels pays opérer en priorité ?



Notre mission :



Les données de la **Banque Mondiale** permettent-elles d'informer le projet d'expansion ?

Jeu de données de la Banque Mondiale

EdStatsCountry.csv

- Informations économiques sur les pays / régions du monde : devises, niveau de revenu, date de divers recensements...
- 241 lignes (1 par pays / régions) et 32 colonnes correspondant aux informations sur ces pays.
- Pas de doublon.

EdStatsCountry-Series.csv

- Information sur la source de données des indicateurs.
- 613 lignes (combinaison de 21 indicateurs et 211 pays / régions) et 4 colonnes (pays/indicateur/ source).
- Pas de valeurs manquantes, exceptée la colonne *Unnamed: 3* entièrement vide.

EdStatsFootNote.csv

- Informations sur l'année des données et sur l'obtention des données, par pays/régions du monde et indicateurs.
- 643.638 lignes (combinaison de 1.558 indicateurs et 239 des régions / pays) et 5 colonnes.
- Pas de valeur manquante (exceptée la colonne *Unnamed: 4* entièrement vide).

EdStatsSeries.csv

- Description des 3.665 indicateurs socio-économiques disponibles sur la Banque Mondiale.
- 3.665 lignes (1 par indicateur) et 21 colonnes de description de l'indicateur.
- Nombreuses valeurs manquantes (hormis 5 colonnes entièrement renseignées : le code, le thème, le nom, la définition longue et la source de l'indicateur).

EdStatsData.csv

- Évolution dans le temps des indicateurs pour les pays / régions du monde.
- 886.930 lignes (3.665 indicateurs pour 242 pays/régions) et 70 colonnes (années de 1970 à 2100).
- Pas de doublon.
- Pas de valeur manquante pour les 4 premières variables (Country Name et Code, indicator Name et code).
Nombreuses valeurs manquantes pour les années 1970 à 2100. 2017 et *Unnamed:69* sont entièrement vides.

II. Filtrage des données

Environnement technique

- Notebook Jupyter 6.4.8
- Python 3.9.12
- Librairies : pandas, missingno

Filtrage des données en 3 temps

1. Choix des indicateurs

Sur les 3.665 indicateurs disponibles, quels sont les indicateurs pertinents pour quantifier le potentiel d'un pays en vue d'un développement commercial ?

2. Filtrage des années

3. Filtrage des pays

1. Choix de 9 indicateurs en lien avec la cible d'*academy*

	Code de l'indicateur	Nom de l'indicateur
Indicateurs démographiques	SP.POP.TOTL	Population Total
	SP.POP.1524.TO.UN	Population, ages 15-24, total
	BAR.POP.2529	Barro-Lee: Population in thousands, age 25-29, total
Niveau d'éducation	PRJ.POP.ALL.3.MF	Wittgenstein Projection: Population in thousands by highest level of educational attainment. Upper Secondary
	PRJ.POP.ALL.4.MF	Wittgenstein Projection: Population in thousands by highest level of educational attainment. Post Secondary
Moyens techniques	IT.NET.USER.P2	Internet users (per 100 people)
	IT.CMP.PCMP.P2	Personal computers (per 100 people).
Niveau de richesse	NY.GDP.MKTP.CD	GDP at market prices (current US dollar)
	NY.GDP.PCAP.KD	GDP per capita (constant 2010 US dollar)

Filtrage des données en 3 temps

1. Choix des indicateurs

Sur les 3.665 indicateurs disponibles, quels sont les indicateurs pertinents pour quantifier le potentiel d'un pays en vue d'un développement commercial ?

2. Filtrage des années

Restriction sur les années basée sur :

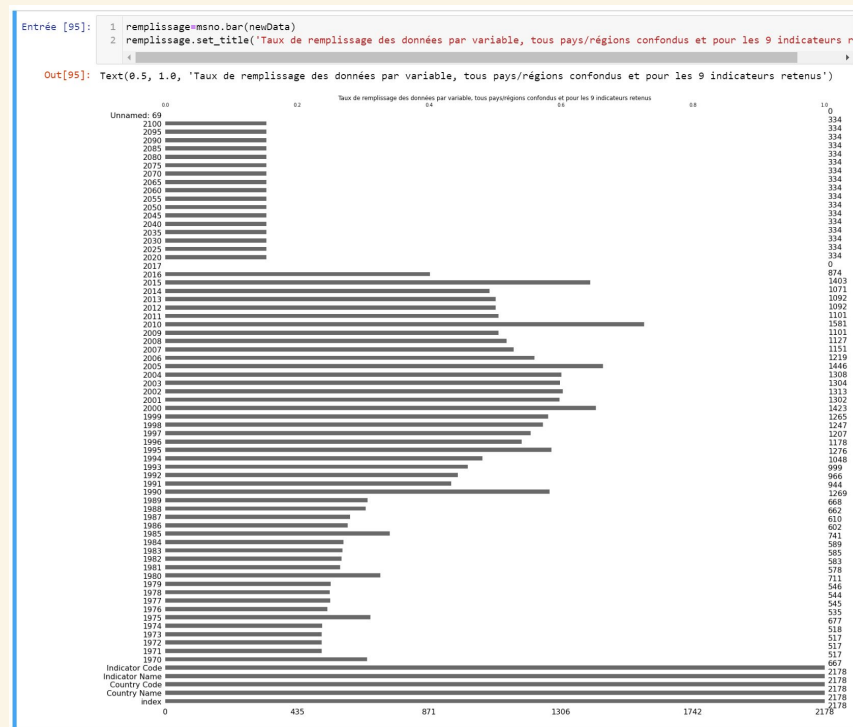
- Analyse des valeurs manquantes.
- Raison de pertinence par rapport à la problématique.

3. Filtrage des pays

2. Filtrage des années

Analyse du remplissage des données par année de 1970 à 2017, puis tous les 5 ans de 2020 à 2100 :

- Peu de valeurs avant 1990 (moins de 30%)
- Aucune valeur en 2017
- Peu de valeurs après 2020 (moins de 20%)



Taux de remplissage des données par année (librairie missingno)

2. Filtrage des années

Analyse des valeurs
manquantes

Critère de
pertinence (Accès
Internet & PC)

Restriction sur

les années 2000 à 2016 et 2020

```
Entrée [113]: 1 anneesRetenuesP = ['2000', '2001', '2002', '2003', '2004', '2005', '2006', '2007', '2008', '2009',  
2 '2010', '2011', '2012', '2013', '2014', '2015', '2016', '2020']  
3 colonnesRetenuesP = ['Country Name', 'Indicator Name', 'Indicator Code'] + anneesRetenuesP  
4 pastData = newData.filter(items=colonnesRetenuesP)  
5 pastData
```

Out[113]:

	Country Name	Indicator Name	Indicator Code	2000	2001	2002	2003	2004	2005	2006	...
0	Arab World	Barro-Lee: Population in thousands, age 25-29...	BAR.POP.2529	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
1	Arab World	GDP at market prices (current US\$)	NY.GDP.MKTP.CD	7.260126e+11	7.130072e+11	7.240099e+11	8.155385e+11	9.594050e+11	1.176100e+12	1.399823e+12	...
2	Arab World	GDP per capita (constant 2005 US\$)	NY.GDP.PCAP.KD	3.339713e+03	3.344289e+03	3.331486e+03	3.429226e+03	3.658931e+03	3.796022e+03	3.975785e+03	...
3	Arab World	Internet users (per 100 people)	IT.NET.USER.P2	1.139541e+00	1.561288e+00	2.693061e+00	3.645881e+00	7.006812e+00	8.393142e+00	1.165465e+01	...
4	Arab World	Personal computers (per 100 people)	IT.CMP.PCMP.P2	1.887733e+00	2.414015e+00	2.517600e+00	3.401317e+00	3.979838e+00	5.305007e+00	6.676681e+00	...
...
2173	Zimbabwe	Personal computers (per 100 people)	IT.CMP.PCMP.P2	1.559544e+00	1.588971e+00	4.746489e+00	4.892251e+00	5.514830e+00	6.687338e+00	6.798012e+00	...
2174	Zimbabwe	Population, ages 15-24, total	SP.POP.1524.TO.UN	2.921453e+06	3.003420e+06	3.080565e+06	3.151277e+06	3.213812e+06	3.266903e+06	3.309907e+06	...
2175	Zimbabwe	Population, total	SP.POP.TOTL	1.222225e+07	1.236616e+07	1.250052e+07	1.263390e+07	1.277751e+07	1.294003e+07	1.312427e+07	...
2176	Zimbabwe	Wittgenstein Projection: Population in thousan...	PRJ.POP.ALL.4.MF	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
2177	Zimbabwe	Wittgenstein Projection: Population in thousan...	PRJ.POP.ALL.3.MF	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...

2178 rows × 21 columns

Filtrage des données en 3 temps

1. Choix des indicateurs

Sur les 3.665 indicateurs disponibles, quels sont les indicateurs pertinents pour quantifier le potentiel d'un pays en vue d'un développement commercial ?

2. Filtrage des années

Restriction sur les années basée sur :

- Analyse des valeurs manquantes.
- Raison de pertinence par rapport à la problématique.

3. Filtrage des pays

En 2 étapes :

1. Elimination des régions du monde et des "petits pays".
2. Elimination des pays pour lesquels un indicateur (ou plus) ne comporte aucune donnée.

3. Filtrage des pays- 1^{ère} étape

1^{ère} étape :

- Elimination des régions du monde (ex : Arab World, East Asia & Pacific, Europe & Central Asia,...)
- Elimination des îles, dépendances... (ex : Saint-Martin, les Bahamas...)



Passage de 242 pays à 201 pays.

Entrée [128]:

```
1 # Liste des pays à garder en retirant Les 2 Listes précédentes des régions et des petits pays
2 paysNonRetenus=listeRegions+listePetitsPays
3 listePays=data.loc[~data['Country Name'].isin(paysNonRetenus),'Country Name']
4 # filtrage du dataframe sur La Liste de pays obtenue.
5 newData=pastData.loc[(pastData['Country Name'].isin(listePays)),:].reset_index()
6 newData
```

Out[128]:

	index	Country Name	Indicator Name	Indicator Code	2000	2001	2002	2003	2004
0	225	Afghanistan	Barro-Lee: Population in thousands, age 25-29,...	BAR.POP.2529	1.668000e+03	NaN	NaN	NaN	NaN
1	226	Afghanistan	GDP at market prices (current US\$)	NY.GDP.MKTP.CD	NaN	2.461666e+09	4.128821e+09	4.583644e+09	5.285466e+09

3. Filtre sur les pays – 2^{ème} étape

2^{ème} étape :

Élimination des pays pour lesquels au moins un des 9 indicateurs ne comporte aucune donnée.



Passage de 201 à 128 pays.

```
Entrée [73]: 1 #Liste des pays pour lesquels il y a un indicateur ou plus pour lequel le fichier ne comporte aucune donnée
2 dfTemp=meltData.groupby(['Country Name','Indicator Code'],as_index=False).count()
3 listePaysNonRenseignes=dfTemp.loc[(dfTemp['Value']==0), 'Country Name'].unique()
4 listePaysNonRenseignes
```


```
Out[73]: array(['Afghanistan', 'American Samoa', 'Andorra', 'Angola', 'Aruba',
'Azerbaijan', 'Barbados', 'Belarus', 'Bermuda', 'Bhutan',
'Bosnia and Herzegovina', 'Botswana', 'Brunei Darussalam',
'Burkina Faso', 'Cabo Verde', 'Chad', 'Comoros', 'Curacao',
'Djibouti', 'Dominica', 'Equatorial Guinea', 'Eritrea', 'Ethiopia',
'Fiji', 'Georgia', 'Gibraltar', 'Greenland', 'Grenada', 'Guam',
'Guinea', 'Guinea-Bissau', 'Kazakhstan', 'Kiribati',
'Korea, Dem. People's Rep.', 'Kosovo', 'Lebanon', 'Liberia',
'Libya', 'Liechtenstein', 'Macedonia, FYR', 'Madagascar',
'Marshall Islands', 'Mauritania', 'Micronesia, Fed. Sts.',
'Monaco', 'Montenegro', 'Nauru', 'New Caledonia', 'Nigeria',
'Oman', 'Palau', 'Papua New Guinea', 'Puerto Rico', 'Samoa',
'San Marino', 'Sao Tome and Principe', 'Serbia', 'Seychelles',
'Sierra Leone', 'Solomon Islands', 'Somalia', 'South Sudan',
'Sri Lanka', 'St. Kitts and Nevis', 'Suriname',
'Syrian Arab Republic', 'Timor-Leste', 'Togo', 'Turkmenistan',
'Tuvalu', 'Uzbekistan', 'Vanuatu', 'Yemen, Rep.'], dtype=object)
```

```
Entrée [75]: 1 #On filtre à nouveau le jeu de données sur les pays en retirant les 73 pays pour lesquels un ou plusieurs indicateurs ne con
2 filteredData=meltData.loc[~meltData['Country Name'].isin(listePaysNonRenseignes)]
3 filteredData['Country Name'].nunique()
```

```
Out[75]: 128
```

Synthèse du filtrage des données

- 9 indicateurs
- Années 2000 à 2017 et 2020
- 128 pays

 **Pour chaque pays, et chaque indicateur, le jeu de données comporte au moins une valeur sur la période de temps sélectionnée.**



III. Analyse et pertinence du jeu de données



Score d'attractivité des pays

1) Moyenne de l'indicateur sur les années retenues

2) Normalisation des indicateurs :

- Population totale du pays en pourcentage du pays le plus peuplé.
- Population des 15-24 ans et des 25-29 ans en pourcentage de la population totale. idem pour le niveau d'étude.
- PIB et PIB par habitant en pourcentage du PIB le plus élevé des pays.

```
3  
4 #Création des colonnes avec L'indicateur normalisé (pour Les 7 indicateurs necessitant d'être normalisés)  
5  
6 #La population de chaque pays en % de la population du pays le plus peuplé  
7 normalizedData['POP TOT N']=pivotData['SP.POP.TOTL']/pivotData['SP.POP.TOTL'].max()*100  
8  
9 #Pourcentage des 15-24 ans sur la population totale  
10 normalizedData['POP 15 24 N']=pivotData['SP.POP.1524.TO.UN']/pivotData['SP.POP.TOTL']*100  
11  
12 #Pourcentage des 25-29 ans sur la population totale ( L'indicateur est exprimé en milliers)  
13 normalizedData['POP 25 29 N']=pivotData['BAR.POP.2529']*1000/pivotData['SP.POP.TOTL']*100  
14  
15 #Pourcentage de la population ayant terminé Le Lycée par rapport à la population totale ( L'indicateur est exprimé en millie  
16 normalizedData['UPPER SECONDARY N']=pivotData['PRJ.POP.ALL.3.MF']*1000/pivotData['SP.POP.TOTL']*100  
17  
18 #Pourcentage de la population ayant terminé Leur études supérieure par rapport à la population totale( L'indicateur est expr  
19 normalizedData['POST SECONDARY N']=pivotData['PRJ.POP.ALL.4.MF']*1000/pivotData['SP.POP.TOTL']*100  
20  
21 #PIB par habitant normalisé par Le max des PIB par habitant  
22 normalizedData['GDP PCAP N']=pivotData['NY.GDP.PCAP.KD']/pivotData['NY.GDP.PCAP.KD'].max()*100  
23  
24 #PIB normalisé par Le max des PIB le plus élevé des pays retenus  
25 normalizedData['GDP N']=pivotData['NY.GDP.MKTP.CD']/pivotData['NY.GDP.MKTP.CD'].max()*100  
26
```

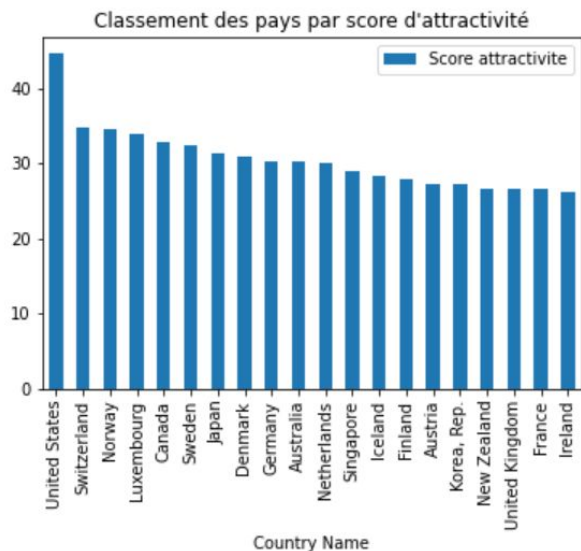
3) Calcul d'un score d'attractivité par pays :

- Moyenne des indicateurs normalisés.
- Moyenne pondéré des indicateurs normalisés.

Classement des pays par score d'attractivité

```
Entrée [151]: 1 # Représentation graphique du top 20 des pays
              2 candidateData.plot.bar(y='Score attractivite', title="Classement des pays par score d'attractivité")
```

```
Out[151]: <AxesSubplot:title={'center': 'Classement des pays par score d'attractivité'}, xlabel='Country Name'>
```



```
Entrée [90]: 1 # Calcul d'un score d'attractivité simple (non pondéré), égal à la moyenne des indicateurs normalisés
              2 normalizedData['score Simple']=normalizedData.mean(axis=1)
              3 candidateData=normalizedData.sort_values('Score Simple', ascending=False).head(20)
              4 candidateData
```

Out[90]:

	Indicator Code	POP TOT N	POP 15 24 N	POP 25 29 N	UPPER SECONDARY N	POST SECONDARY N	GDP PCAP N	GDP N	ITNETUSER.P2	ITCMPPCM.P2	Score Simple
Country Name											
United States	22.030247	14.299257	6.728848	43.468150	28.475998	47.178512	100.000000	67.385319	70.281675	44.527512	
Switzerland	0.582242	11.879196	5.768015	45.139501	19.588059	70.645988	3.574785	75.583391	80.694446	34.828403	
Norway	0.363636	12.407477	6.152416	36.274036	26.731556	85.118379	2.526460	85.045404	55.363449	34.442535	
Luxembourg	0.037553	11.502026	6.105131	29.864826	22.313919	100.000000	0.321148	74.480298	59.310195	33.772586	
Canada	2.522958	13.080803	5.583748	31.851887	47.347700	46.086758	9.424112	75.002310	63.038518	32.770977	
Sweden	0.701435	12.298037	6.037821	38.151345	30.000286	49.419552	3.068210	83.059642	69.478943	32.468464	
Japan	9.643628	10.661460	8.826514	38.302350	31.585480	43.266712	34.478957	69.788079	36.924583	31.275307	
Denmark	0.416091	11.662262	6.194228	38.556438	19.895482	56.789229	1.976333	81.444586	60.872642	30.867475	
Germany	6.187474	11.432721	6.016902	42.498532	25.565014	40.330710	21.745525	70.444952	48.376829	30.289295	
Australia	1.621449	13.187437	6.606782	39.599096	29.291319	48.964120	6.645482	72.753982	53.622604	30.254768	
Netherlands	1.246411	11.896408	6.136799	34.192405	22.090460	47.963725	5.080875	79.873268	60.913631	29.934022	
Singapore	0.361418	12.558882	5.834662	20.809141	51.024847	41.909478	1.367957	84.844272	60.929423	28.860009	
Iceland	0.023374	14.198601	7.006166	29.266371	30.017648	41.263714	0.101966	86.275284	45.860307	28.223715	
Finland	0.402366	12.286641	6.092445	30.870539	34.669457	43.876313	1.545131	76.654376	45.084037	27.942367	
Austria	0.628991	11.890467	6.268592	42.875583	21.831603	44.628127	2.428586	65.150804	50.036888	27.315516	
Korea, Rep.	3.711516	14.016990	8.011321	33.108807	30.466831	19.919562	7.023585	76.360066	51.682606	27.164723	
New Zealand	0.321400	13.947606	6.177049	40.732702	28.621138	32.720720	0.913627	72.001721	44.188646	26.625090	
United Kingdom	4.681247	12.551252	6.208063	10.348865	22.754108	38.150686	17.256740	73.480179	53.034963	26.496234	
France	4.849751	11.959272	6.148733	34.326882	21.145801	36.407382	16.323276	59.807078	44.224959	26.468105	
Ireland	0.328636	14.212090	7.336626	20.726621	37.700983	50.238011	1.518061	57.078669	46.117493	26.196688	

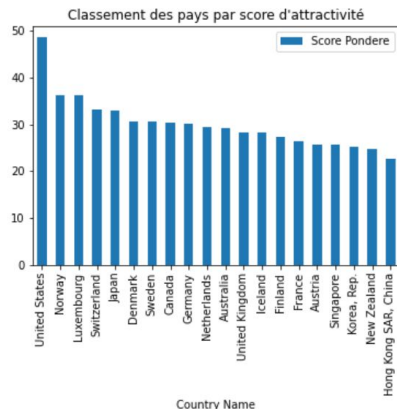
Score d'attractivité avec moyenne simple

Classement des pays par score d'attractivité

- Score d'attractivité avec moyenne pondérée.
- Poids plus importants pour :
 - Population totale
 - Population 15-24 ans
 - Accès à internet
 - PIB et PIB/habitant

```
Entrée [89]: 1 #Liste des poids des indicateurs
2 weight=[5,5,1,1,1,5,1,5,5]
3 #ajout d'une colonne contenant le score pondéré
4 normalizedData['Score Pondere']=(
5     weight[0]*candidateData['POP TOT N']
6     +weight[1]*candidateData['POP 15 24 N']
7     +weight[2]*candidateData['POP 25 29 N']
8     +weight[3]*candidateData['UPPER SECONDARY N']
9     +weight[4]*candidateData['POST SECONDARY N']
10    +weight[5]*candidateData['IT.NET.USER.P2']
11    +weight[6]*candidateData['IT.CMP.PCMP.P2']
12    +weight[7]*candidateData['GDP PCAP N']
13    +weight[8]*candidateData['GDP N'])/sum(weight)
14
15 weightCandidateData=normalizedData.sort_values('Score Pondere',ascending=False).head(20)
16 weightCandidateData.plot.bar(y='Score Pondere', title="Classement des pays par score d'attractivité")
```

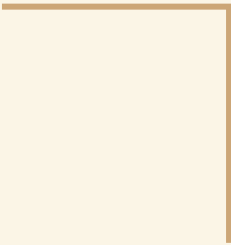
```
Out[89]: <AxesSubplot:title={'center':'Classement des pays par score d'attractivité'}, xlabel='Country Name'>
```



Score d'attractivité avec moyenne pondérée

Conclusion

- Jeu de données pertinent pour répondre à la problématique : les données sont renseignées pour les indicateurs, années et pays pertinents pour notre étude.
- Informations complémentaires intéressantes à étudier :
 - Accès à internet plus précis que sur les 3 derniers mois.
 - Concurrence dans les pays potentiels en termes de cours en ligne.
 - Dépenses pour la formation dans les pays potentiels.



Merci pour votre attention !
Des questions ?

