Analysez des données de systèmes éducatifs

Soutenance Projet 2

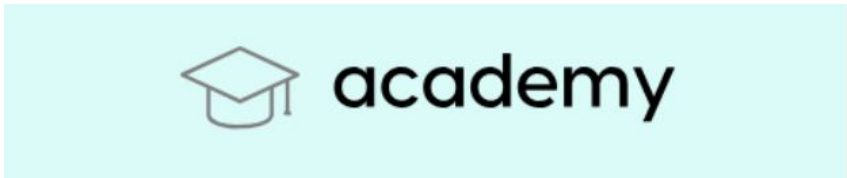
Analysez des données de systèmes éducatifs

1. Présentation du jeu de données
2. Analyse pré exploratoire
3. Conclusions sur le jeu de données

Presentation de la Mission

- Academy est une start-up de la EdTech
- En ligne: Contenus de formation de niveau lycée et université
- Objectif d’expansion à l’international

Analysis bivariée; media, median



Presentation de la Mission

Pour la pré-analyse:

- Valider la qualité de ce jeu de données (comporte-t-il beaucoup de données manquantes, dupliquées ?)
- Décrire les informations contenues dans le jeu de données (nombre de colonnes ? nombre de lignes ?)
- Sélectionner les informations qui semblent pertinentes pour répondre à la problématique (quelles sont les colonnes contenant des informations qui peuvent être utiles pour répondre à la problématique de l'entreprise ?)
- Déterminer des ordres de grandeurs des indicateurs statistiques classiques pour les différentes zones géographiques et pays du monde (moyenne/médiane/écart-type par pays et par continent ou bloc géographique)
- déterminer si ce jeu de données peut informer les décisions d'ouverture vers de nouveaux pays

Presentation de la Mission

Objectif du projet :

Déterminer si ce jeu de données peut informer les décisions d'ouverture vers de nouveaux pays

Rappel de la problématique

- Academy est une start-up de la EdTech
- Elearnings : Contenus de formation de niveau lycée et université
- Objectif d’expansion à l’international

Présentation du jeu de données 1 - 3

01. EdStatsCountry-Series

- Informations sur la source des données contenues dans EdStatsCountry Taille : 613 lignes, 4 colonnes
- Pas de valeur manquante (sauf Unnamed : 3” qui est une colonne uniquement composée de NaN)
- Aucun doublon

02. EdStatsCountry

- Informations globales sur l’économie de chaque pays du monde (et de zones géographiques)
- Taille : 241 lignes (1 par pays / zone) , 32 colonnes Quelques valeurs manquantes Aucun doublon

Présentation du jeu de données 2 - 3

03. EdStatsData

- Donne l’évolution de nombreux indicateurs pour tous les pays et certains groupes de pays Taille : 886 930 lignes, 70 colonnes données depuis 1970
- Nombreuses valeurs manquantes
- Aucun doublon

04. EdStatsFootNote

Contient des Informations sur l’année d’origine des données et les incertitudes sur les données) Taille : 643 638 lignes, 4 colonnes Pas de valeur manquante (sauf Unnamed : 4 qui est une colonne uniquement composée de NaN) Aucun doublon

Présentation du jeu de données 3 - 3

05. EdStatsSeries

- Informations sur les indicateurs socio économiques disponibles dans EdStatsData. Taille : 3665 lignes, 21 colonnes 6 colonnes vides pour lesquelles il manque toutes les valeurs.
- Il manque plus de 80 % des données dans 10 autres colonnes de la table
- Aucun doublon

II Analyse Pré Exploratoire

Processus d'analyse pré exploratoire

1. **Connaître les données**

   Quelles informations?

   Quelles années?

2. **Identifier les indicateurs exploitables**

   Quantités de données manquantes?

3. **Comparer les pays**

   Quels indicateurs choisir?

   Analyse des résultats obtenus

   Quels sont les pays à cibler par Academy?

4. **Quel est le potentiel pour chaque pays?**

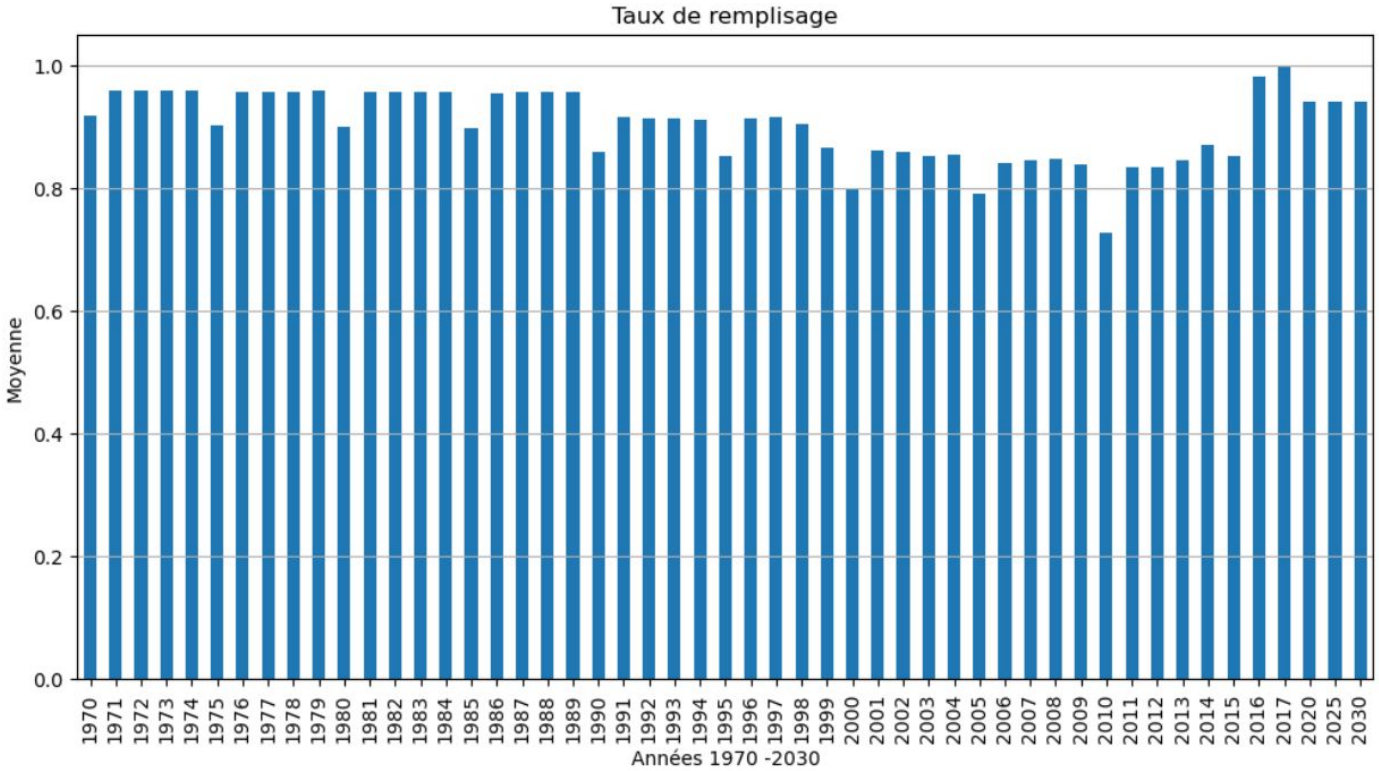
   Comment identifier le potentiel des pays choisis?

1 - Connaître les données

1 - Connaître les données - Préambule

- Historique et prédictions de 1970 à 2050
- 241 zones géographiques (la plupart des pays)
- 3665 indicateurs uniques

1 - Connaître les données - Quantité de données par année



2 - Identifier les indicateurs exploitables

2 - Identifier les indicateurs exploitables

1. EdStatsCountry_Series.sample(5)

	CountryCode	SeriesCode	DESCRIPTION	
349	MDV	SP.POP.GROW	Data sources: For 1960-1989, United Nations World Population Prospects. For ...	None
598	XKX	NY.GDP.MKTP.PP.KD	Estimates are based on regression.	None
74	BTN	SP.POP.GROW	Data sources: United Nations World Population Prospects	None
395	NCL	SP.POP.GROW	Data sources: Institute of Statistics and Economic Studies, 1983 Census, 198...	None
284	LBN	SP.POP.TOTL	Data sources : United Nations World Population Prospects	None

2 - Identifier les indicateurs exploitables

2. EdStatsCountry.sample(5)

Out[6]:

	Country Code	Short Name	Table Name	Long Name	2-alpha code	Currency Unit	Special Notes	Region	Income Group	WB-2 code	...	IMF data dissemination standard	Latest population census	Latest household survey	Sourc most re Income expend
6	ARE	United Arab Emirates	United Arab Emirates	United Arab Emirates	AE	U.A.E. dirham	April 2013 database update: Based on...	Middle East & North Africa	High income: nonOECD	AE	...	General Data Dissemination System (G...	2010		
35	CHI	Channel Islands	Channel Islands	Channel Islands		Pound sterling		Europe & Central Asia	High income: nonOECD	JG	...		Guernsey: 2009; Jersey: 2011.		
98	IND	India	India	Republic of India	IN	Indian rupee	Fiscal year end: March 31; reporting...	South Asia	Lower middle income	IN	...	Special Data Dissemination Standard ...	2011	Demographic and Health Survey (DHS),...	Integr house survey (I

2 - Identifier les indicateurs exploitables

3. EdStatsData.sample(5)

Out[4]:

	Country Name	Country Code	Indicator Name	Indicator Code	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983
128476	Australia	AUS	Barro-Lee: Average years of secondary schooling, age 60-64, female	BAR.SEC.SCHL.6064.FE	2.38	NaN	NaN	NaN	NaN	2.99	NaN	NaN	NaN	NaN	3.7	NaN	NaN	NaN
431483	Ireland	IRL	SABER: (Engaging the Private Sector) Policy Goal 1 Lever 5: Curriculum Delivery	SABER.PRVT.GOAL1.LVL5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
			Wittgenstein Projection: Percentage															

2 - Identifier les indicateurs exploitables

4. EdStatsFootNote.sample(5)

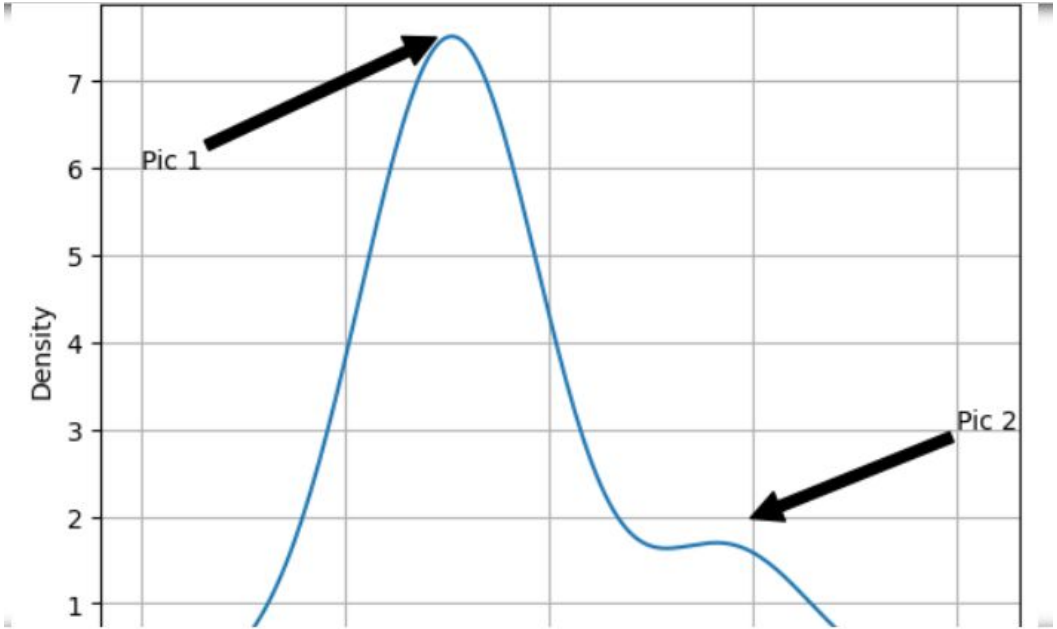
	CountryCode	SeriesCode	Year	DESCRIPTION	
547693	SSF	UIS.E.1.G2.F	YR1996	UNESCO Institute for Statistics (UIS) estimate	None
78093	BTN	SE.PRE.ENRL.FE	YR2005	Country estimation.	None
346251	LIE	SP.TER.TOTL.FE.IN	YR2006	National Estimation	None
154645	ECA	UIS.ROFST.1.F	YR1979	UNESCO Institute for Statistics (UIS) estimate	None
405482	MKD	SE.SEC.ENRL	YR1995	Country Data	None

2 - Identifier les indicateurs exploitables

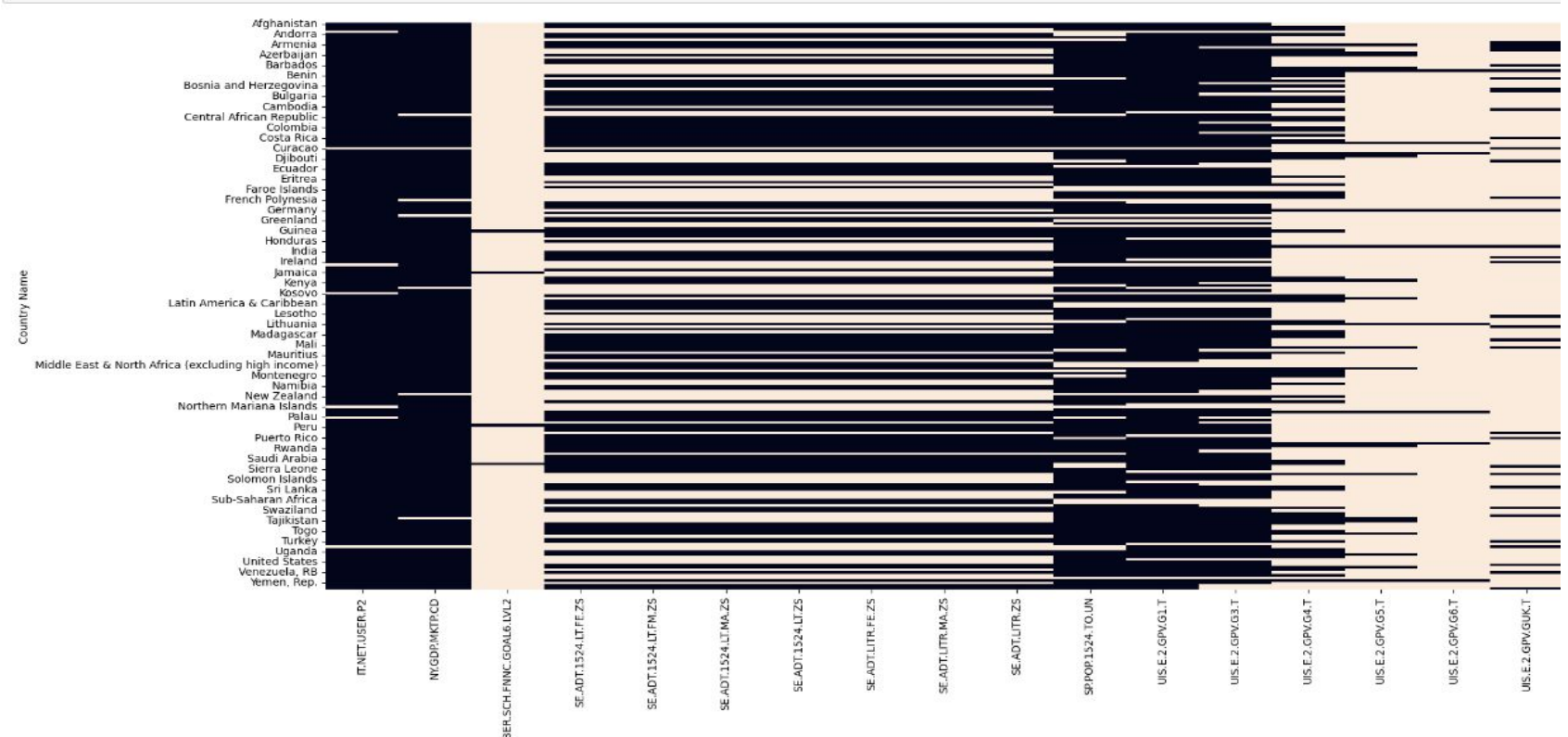
5. EdStatsSeries.sample(5)

	Series Code	Topic	Indicator Name	Short definition	Long definition	Unit of measure	Periodicity	Base Period	Other notes	Aggregation method	Limitations and exceptions	Notes from original source	Ge comm
3196	UIS.NART.2.Q5	Education Equality	UIS: Total net attendance rate, lower secondary, richest quintile, both sexe...	Total number of students of the official lower secondary school age group wh...	Total number of students of the official lower secondary school age group wh...	None						None	
2381	SE.XPD.TOTL.GD.ZS	Expenditures	Government expenditure on education as % of GDP (%)	Total general (local, regional and central) government expenditure on educat...	Total general (local, regional and central) government expenditure on educat...	None			All Levels			None	

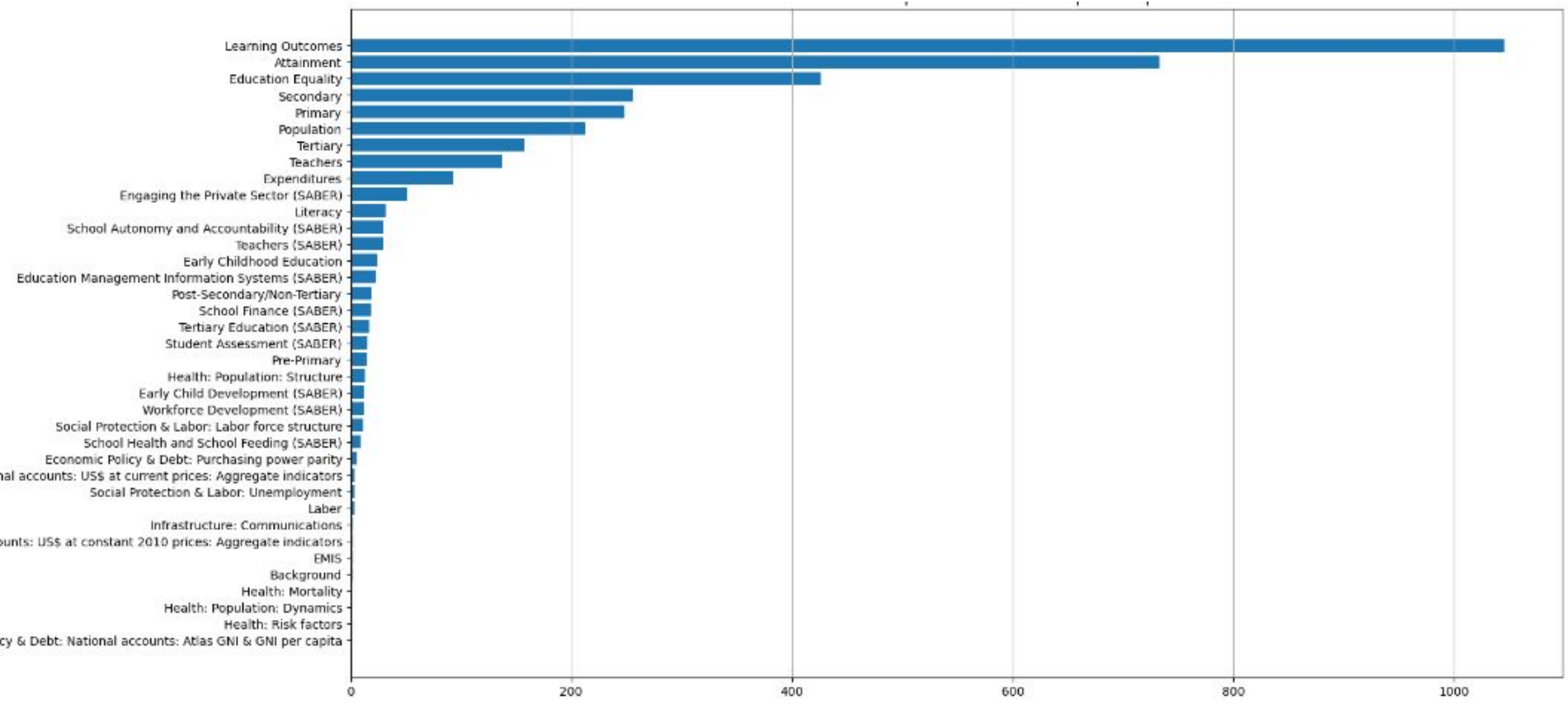
Taux de remplissage des indicateurs 2010 - 2015



2 - Comparer les indicateurs - Identifier les NaN graphiquement (saumon = donnée manquante)



Nombre indicateurs pour les thèmes les plus fréquents



3 - Sélection des indicateurs - Indicateurs retenus

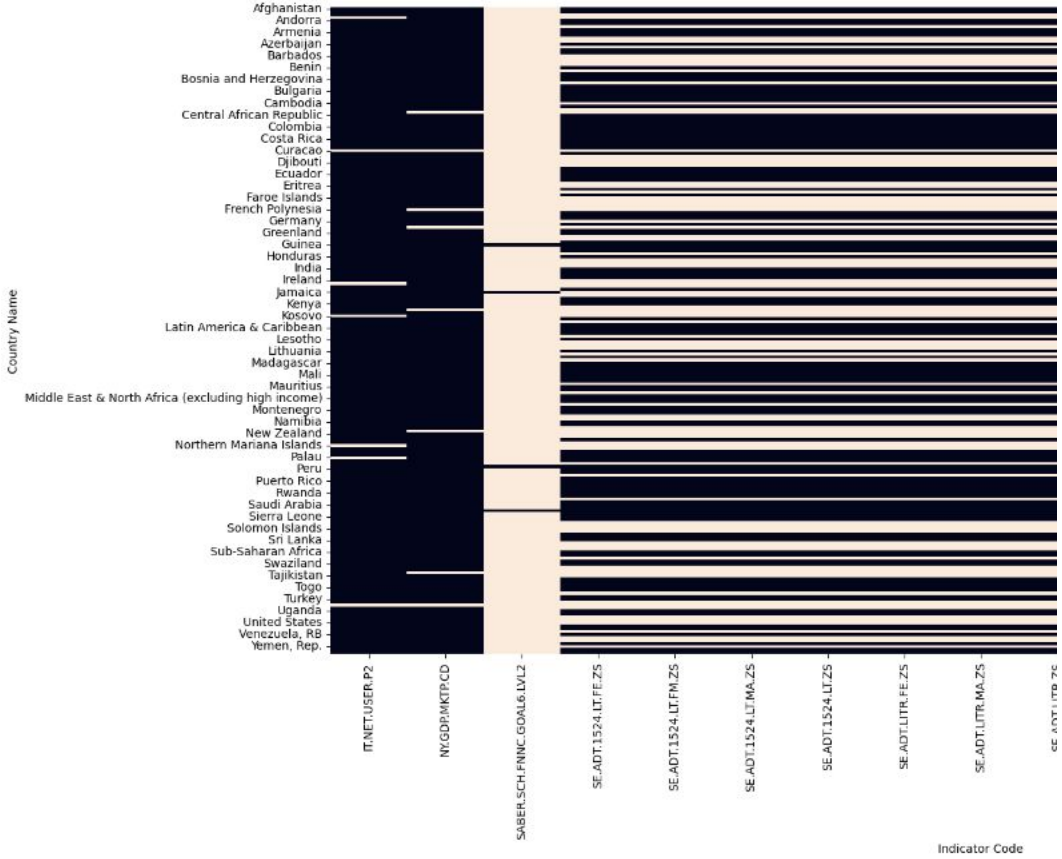
Après une phase d'observation des indicateurs : indicateurs retenus

```
# Definition de les indicateurs retenus
indicateurs_retenus = ["IT.NET.USER.P2","NY.GDP.MKTP.CD","SE.ADT.1524.LT.FM.ZS","
EdStatsSeries[["Series Code","Long definition"]].loc[EdStatsSeries["Series Code"]
```

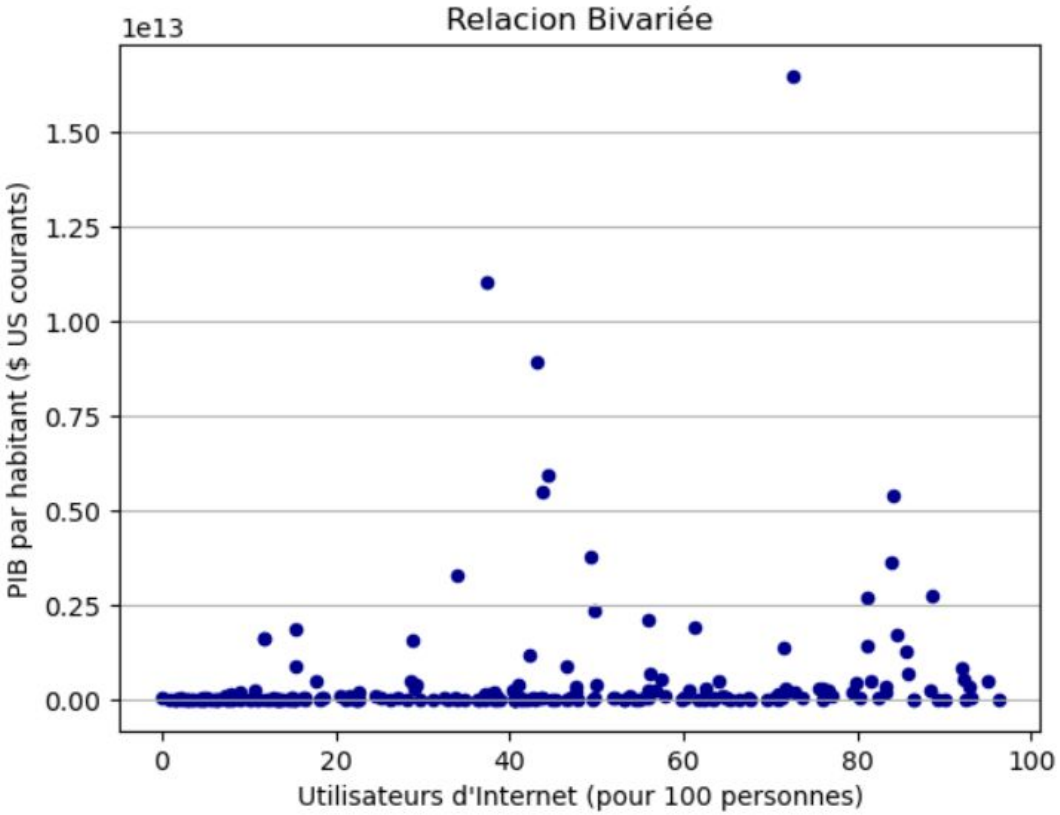
	Series Code	Long definition
611	IT.NET.USER.P2	Internet users are individuals who have used the Internet (from any location...
1658	NY.GDP.MKTP.CD	GDP at purchaser's prices is the sum of gross value added by all resident pr...
2210	SE.ADT.1524.LT.FM.ZS	Ratio of female youth literacy rate to male youth literacy rate. It is calcu...
2211	SE.ADT.1524.LT.MA.ZS	Number of males age 15 to 24 years who can both read and write with understa...
2212	SE.ADT.1524.LT.ZS	Number of people age 15 to 24 years who can both read and write with underst...
2806	UIS.E.2.GPV.G3.T	Total number of students enrolled in Grade 3 of lower secondary general educ...

Création d’un heatmap pour trouver les valeurs manquantes.

NAN en saumon.



Analisis Univariée.



Analisis Bivariée

```
In [76]: table.corr()
```

Out[76]:

Indicator Code	IT.NET.USER.P2	NY.GDP.MKTP.CD	SE.ADT.1524.LT.FE.ZS	SE.ADT.1524.LT.FM.ZS	SE.ADT.1524.LT.MA.ZS
Indicator Code					
IT.NET.USER.P2	1.00000	0.15978	0.67479	0.37300	0.50402
NY.GDP.MKTP.CD	0.15978	1.00000	0.13438	0.06038	0.09233
SE.ADT.1524.LT.FE.ZS	0.67479	0.13438	1.00000	0.88942	0.96647
SE.ADT.1524.LT.FM.ZS	0.37300	0.06038	0.88942	1.00000	0.77648
SE.ADT.1524.LT.MA.ZS	0.50402	0.09233	0.96647	0.77648	1.00000
SE.ADT.1524.LT.ZS	0.50564	0.08938	0.99456	0.85534	0.98892
SE.ADT.LITR.FE.ZS	0.73172	0.12750	0.95786	0.80753	0.92737
SE.ADT.LITR.MA.ZS	0.70894	0.13817	0.94771	0.74528	0.96640
SE.ADT.LITR.ZS	0.72950	0.13360	0.96154	0.78808	0.95129
SP.POP.1524.TO.UN	-0.07149	0.45599	0.01462	-0.00286	-0.00500

SCORING

Création d'un Dataframe pour le scoring

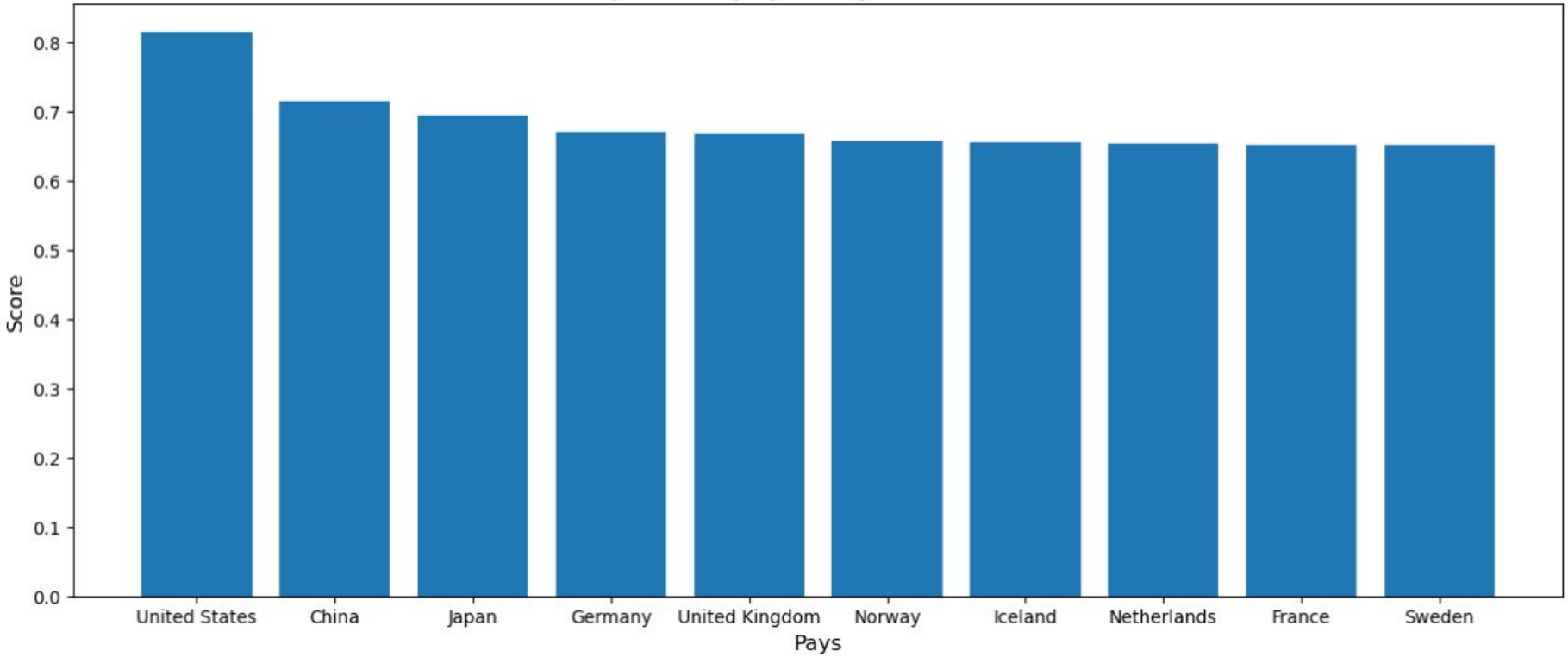
```
#Pivot du Dateframe sur "Country Name"
table = pd.pivot_table(EdStatsDataDfRed,values='moy', index=['Country Name'],
                       columns=['Indicator Code'])

#Indicator Code
table = table.reset_index()
```

4 - Quel potentiel pour ces pays?

```
#Graphic
table.sort_values("score", ascending=False).head(20)
```

Indicator Code	Country Name	IT.NET.USER.P2	NY.GDP.MKTP.CD	SE.ADT.1524.LT.FM.ZS	SE.ADT.1524.LT.MA.ZS	SE.ADT.1524.LT.ZS	UIS.E.2.GPV.G3.T	n
208	United States	72.51228	1.64738e+13	1.00000	97.93447	98.15793	4.35005e+06	
39	China	43.15002	8.89800e+12	0.99903	99.68962	99.64229	1.69228e+07	
97	Japan	84.19080	5.40805e+12	1.00000	97.93447	98.15793	1.21068e+06	
72	Germany	83.92830	3.62292e+12	1.00000	97.93447	98.15793	8.26946e+05	
207	United Kingdom	88.55240	2.72853e+12	1.00000	97.93447	98.15793	7.00667e+04	
149	Norway	94.94894	4.74616e+11	1.00000	97.93447	98.15793	6.41688e+04	
87	Iceland	96.22105	1.52650e+10	1.00000	97.93447	98.15793	4.42500e+03	

Top 10 les pays les plus attractifs



III Conclusions

Conclusion :

Après avoir effectué une analyse exploratoire des données de la Banque Mondiale sur l'éducation, nous pouvons tirer certaines conclusions importantes pour le projet d'expansion d'Academy . Nous avons identifié les dix pays les plus prometteurs, parmi lesquels les:

1. **Etats-Unis**
2. **China**
3. **Japan**
4. Germany
5. United Kingdom
6. Norway
7. Iceland
8. Netherlands
9. France
10. Sweden

Ces pays sont parmi les plus émergents du monde, ce qui en fait des choix attractifs pour l'expansion de l'entreprise.

En outre, en examinant les données par région, nous avons constaté que l'Amérique du Nord, l'Europe et l'Asie centrale sont les régions présentant les scores les plus élevés et les plus grands potentiels. Cela confirme les résultats précédents, car ces régions ont des économies solides et des niveaux socio-économiques élevés, ce qui se traduit par des niveaux d'éducation élevés pour leur population.

En résumé, ces résultats suggèrent que Academy devrait envisager d'ouvrir des succursales dans ces pays pour profiter de leur potentiel de croissance élevé. Il est également important de noter que ces résultats sont basés sur des données réelles et fiables, ce qui renforce leur pertinence pour le projet d'expansion d'Academy .

Remerciements