

Analysez des données de systèmes éducatifs

Data Scientist: Projet 2

Rim Bahroun



Analysez des données de systèmes éducatifs



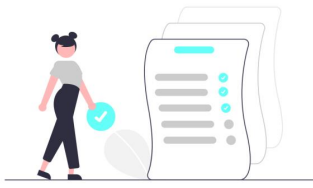
Academy : start-up de la EdTech

- des formations en ligne pour un public de niveau lycée et université.
- Objectif d'**expansion à l'international**.



Mission: analyse exploratoire des données de systèmes éducatifs de la banque mondiale pour définir une stratégie d'expansion.

Analysez des données de systèmes éducatifs



- **Partie 1: Inspection des données**

- EdStatsData.csv
- EdStatsCountry.csv
- EdStatsSeries.csv
- EdStatsCountry-Series.csv
- EdStatsFootNote.csv

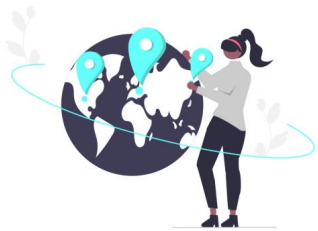


BANQUE MONDIALE

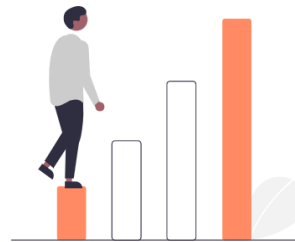
- **Partie 2: Exploration des données**

- Stratégie d'expansion : sélection des indicateurs
- Agrégation temporelle des indicateurs
- Calcul d'un score par pays
- Classement final des pays

- **Conclusion**



academy



OPENCLASSROOMS

1. Inspection des données



Données à disposition : 5 fichiers .csv

EdStatsData.csv

886930 lignes et 70 colonnes
Il donne les valeurs des indicateurs sur les années entre 1970 et 2100 pour chaque pays.

EdStatsCountry.csv

241 lignes et 32 colonnes
Il donne des informations sur chaque pays.

EdStatsSeries.csv

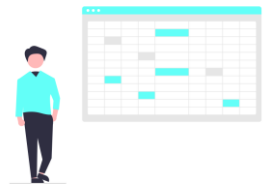
3665 lignes et 21 colonnes
Il donne des informations sur les indicateurs statistiques.

EdStatsCountry-Series.csv

613 lignes et 4 colonnes
Il donne la source des statistiques pour chaque pays et chaque indicateur.

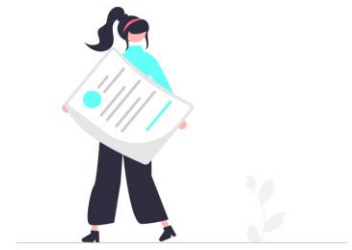
EdStatsFootNote.csv

643638 lignes et 5 colonnes
Il donne une information sur l'origine d'estimation des valeurs indiquées dans le EdStatsData.csv



OPENCLASSROOMS

1. Inspection des données: EdStatsData.csv



1. EdStatsData : Analyse de la forme des données

EdStatsData.csv

886930 lignes et 70 colonnes

Il donne les valeurs des indicateurs sur les années entre 1970 et 2100 pour chaque pays.

```
1 # Types de variables
2 data.dtypes.value_counts()
```

```
float64    66
object      4
dtype: int64
```



Les variables

1. Variables pays

- Country Name
- Country Code

2. Variables indicateurs

- Indicator Name
- Indicator Code

3. Variables données par année

- 66 années de 1970 à 2100



```
1 data = data_EdStatsData.copy()
2 data.head(2)
```

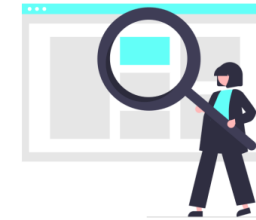
	Country Name	Country Code	Indicator Name	Indicator Code	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.F	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

1. Inspection des données : EdStatsData.csv



1.1 EdStatsData: Etude des variables pays

- Country Name : object (remplie à 100%)
- Country Code : object (remplie à 100%)



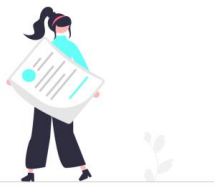
242 valeurs différentes : pour chaque 'Country Name' un unique 'Country Code'

- **25 groupes de pays** soit par région soit par income group
- **217 pays**

	Country Name	Country Code
0	Arab World	ARB
3665	East Asia & Pacific	EAS
7330	East Asia & Pacific (excluding high income)	EAP
10995	Euro area	EMU
14660	Europe & Central Asia	ECS
18325	Europe & Central Asia (excluding high income)	ECA
21990	European Union	EUU
25655	Heavily indebted poor countries (HIPC)	HPC
29320	High income	HIC
32985	Latin America & Caribbean	LCN
36650	Latin America & Caribbean (excluding high income)	LAC
40315	Least developed countries: UN classification	LDC
43980	Low & middle income	LMY
47645	Low income	LIC
51310	Lower middle income	LMC
54975	Middle East & North Africa	MEA
58640	Middle East & North Africa (excluding high inc...	MNA
62305	Middle income	MIC
65970	North America	NAC
69635	OECD members	OED
73300	South Asia	SAS
76965	Sub-Saharan Africa	SSF
80630	Sub-Saharan Africa (excluding high income)	SSA
84295	Upper middle income	UMC
87960	World	WLD



1. Inspection des données : EdStatsData.csv



1.2 EdStatsData : Etude des variables indicateurs

- Indicator Name : object (remplie à 100%)
- Indicator Code : object (remplie à 100%)

```
1 data = data_EdStatsData.copy()
2 data.head(2)
```

	Country Name	Country Code	Indicator Name	Indicator Code	1970	1971	1972	1973	1974
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	NaN	NaN	NaN	NaN	NaN
1	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.F	NaN	NaN	NaN	NaN	NaN



3665 valeurs différentes : pour chaque 'Indicator Name' un unique 'Indicator Code'

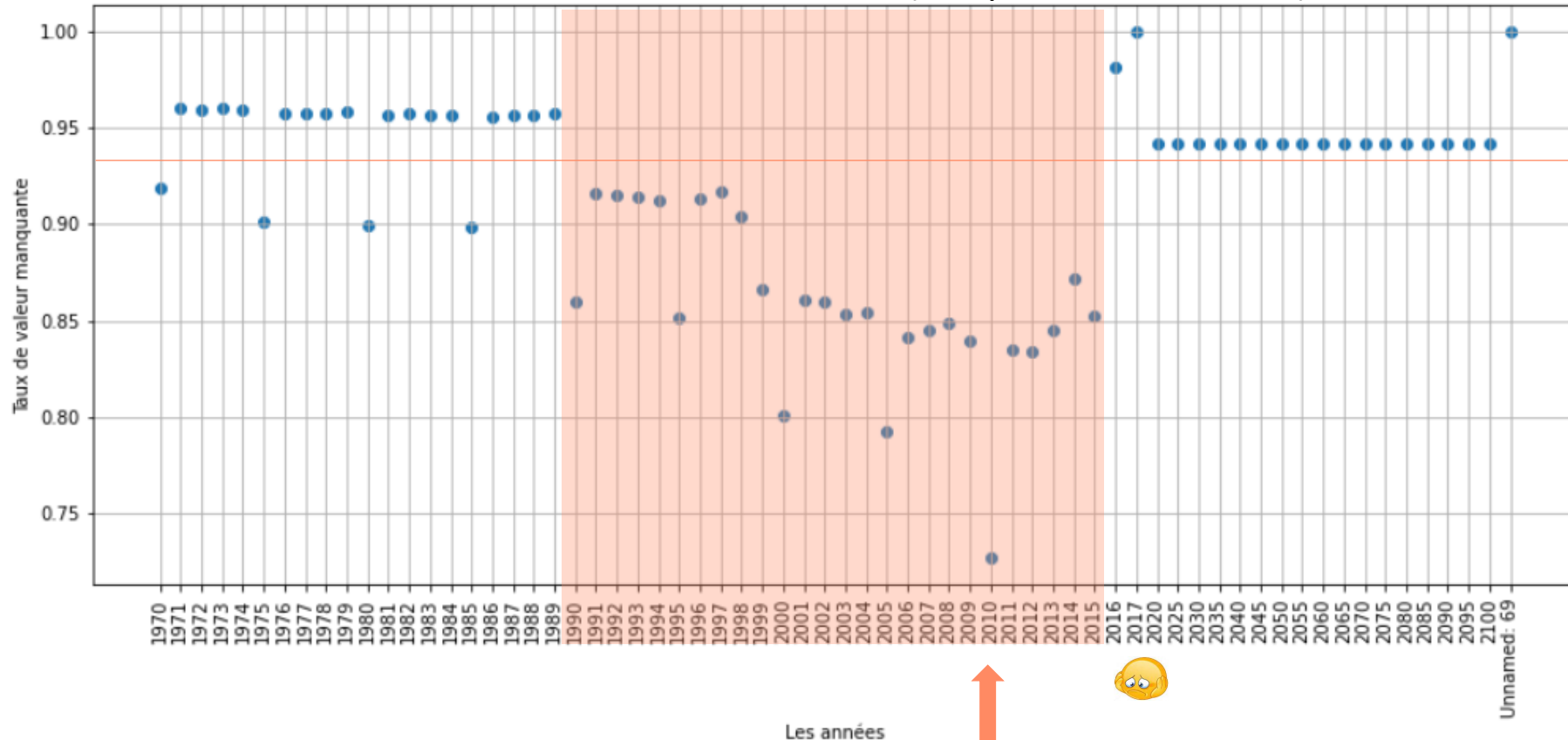


1. Inspection des données : EdStatsData.csv



1.3 EdStatsData : Etude et nettoyage sur les variables temps

66 années de 1970 à 2100 : float64 (remplie moins de 30%)



```
1 data = data_EdStatsData.copy()
2 data.head(2)
```

	Country Name	Country Code	Indicator Name	Indicator Code	1970	1971	1972	1973	1974
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	NaN	NaN	NaN	NaN	NaN
1	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.F	NaN	NaN	NaN	NaN	NaN



Années retenus
1990 à 2015

academy



→ L'année avec le plus de donnée est l'année **2010**



OPENCLASSROOMS

1. Inspection des données : EdStatsData.csv



1.4 EdStatsData : Etude et nettoyage sur les individus



3665 indicateurs x 242 pays = 886930 lignes

→ Pour chaque pays et pour chaque indicateur, on donne les valeurs des indicateurs en fonction de l'année.

```
1 data = data_EdStatsData.copy()
2 data.head(2)
```

	Country Name	Country Code	Indicator Name	Indicator Code	1970	1971	1972	1973	1974
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	NaN	NaN	NaN	NaN	NaN
1	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.F	NaN	NaN	NaN	NaN	NaN



→ **Eliminer les lignes vides**

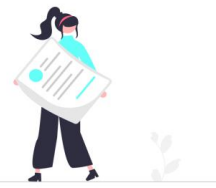
```
1 # Suppression des lignes vides sur les années
2 df = data.copy()
3 df = df.loc[:, Annee_list] # prendre que les colonnes années
4 print("Dimensions avant dropna:", df.shape)
5 df.dropna(axis=0, how='all', inplace = True) #eliminer les lignes avec 100% de valeurs manquantes
6 print("Dimensions après dropna:", df.shape)
7 df.head(2)
```

Dimensions avant dropna: (886930, 26)

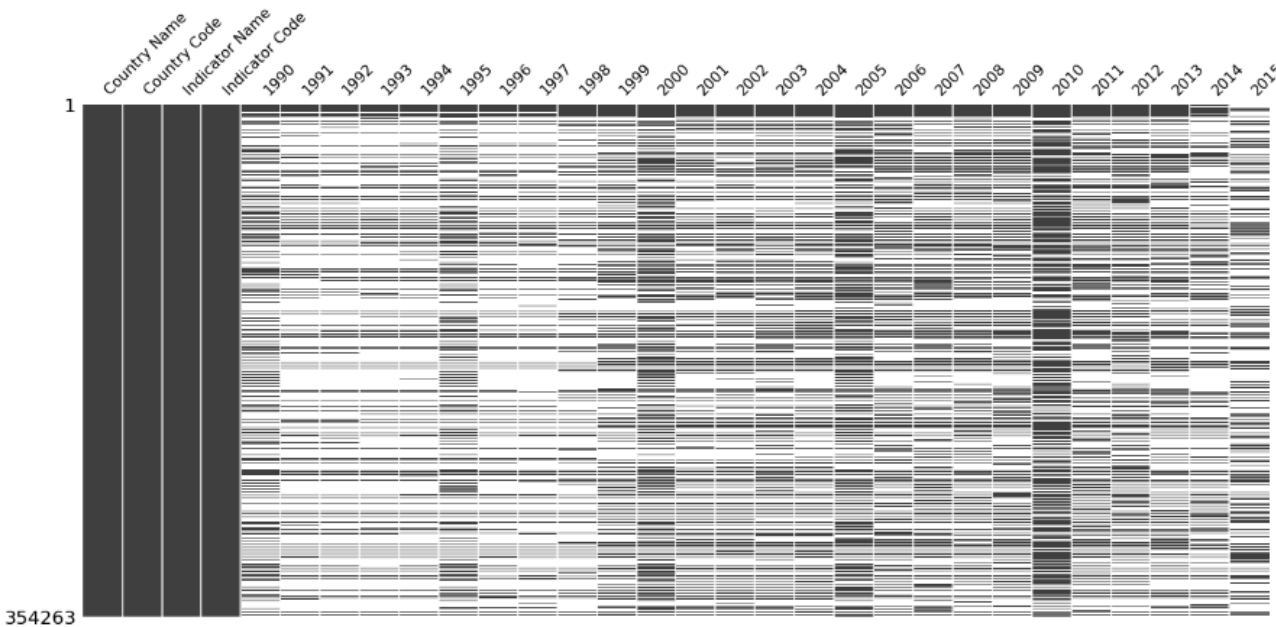
Dimensions après dropna: (354263, 26)



1. Inspection des données : EdStatsData.csv



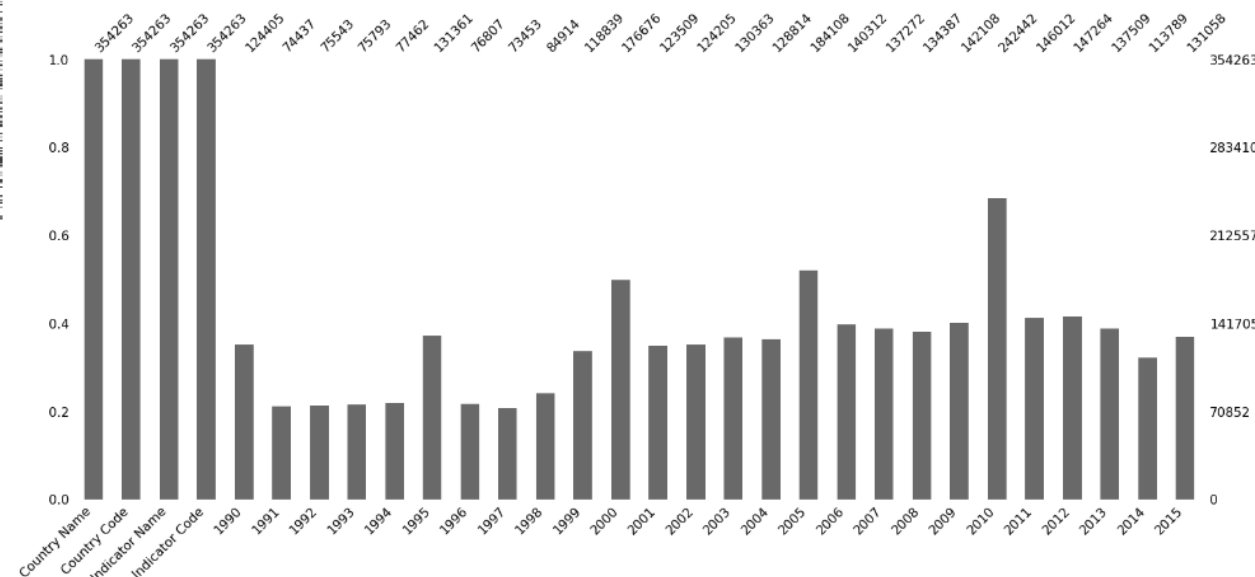
1.5 EdStatsData : graphiques après un premier nettoyage sur les variables



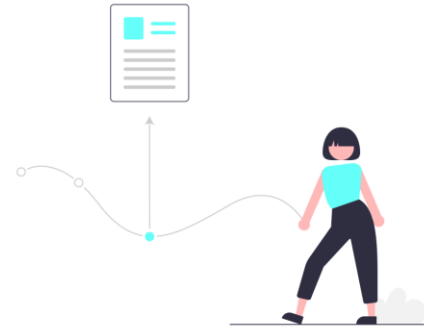
Données par variables

```
1 data = data_EdStatsData.copy()
2 data.head(2)
```

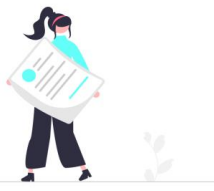
	Country Name	Country Code	Indicator Name	Indicator Code	1970	1971	1972	1973	1974
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	NaN	NaN	NaN	NaN	NaN
1	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.F	NaN	NaN	NaN	NaN	NaN



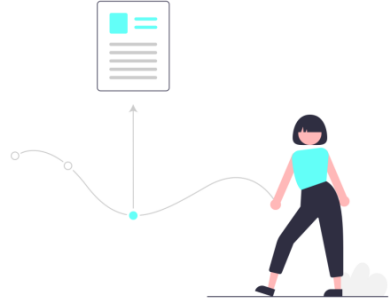
Quantité de données par variables



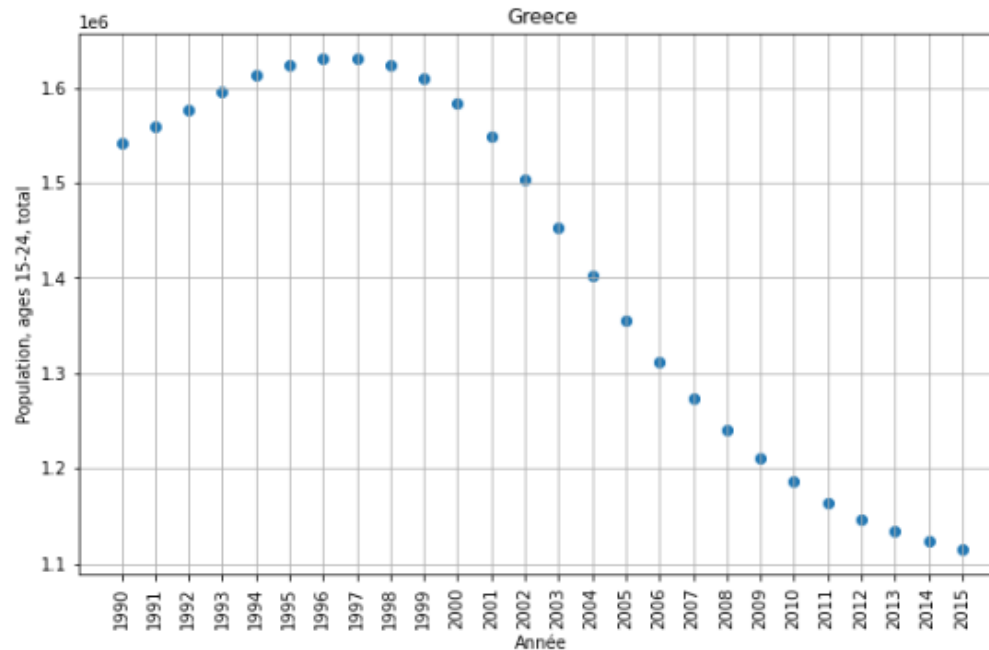
1. Inspection des données : EdStatsData.csv



1.5 EdStatsData : graphiques: exemple de données



```
data = StatsData.copy()
# pour 1 pays donné, 1 indicateur en fct de l'année
plt.figure(figsize=(10,6))
Country = Country_list[100]
Indicator = 'Population, ages 15-24, total'
df = data.loc[(data['Country Name'] == Country)&(data['Indicator Name'] == Indicator) , Annee_list]
plt.scatter(df.columns, df.values)
plt.title(Country)
plt.xlabel('Année')
plt.ylabel(Indicator)
plt.xticks(rotation='vertical')
plt.grid()
```

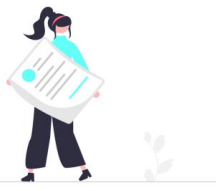


```
1 data = data_EdStatsData.copy()
2 data.head(2)
```

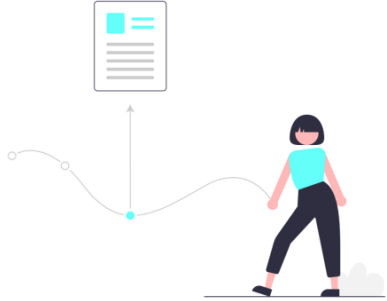
	Country Name	Country Code	Indicator Name	Indicator Code	1970	1971	1972	1973	1974
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	NaN	NaN	NaN	NaN	NaN
1	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.F	NaN	NaN	NaN	NaN	NaN



1. Inspection des données : EdStatsData.csv



1.5 EdStatsData : graphiques: exemple de données



```
data = StatsData.copy()
data = data.iloc[25:,:]

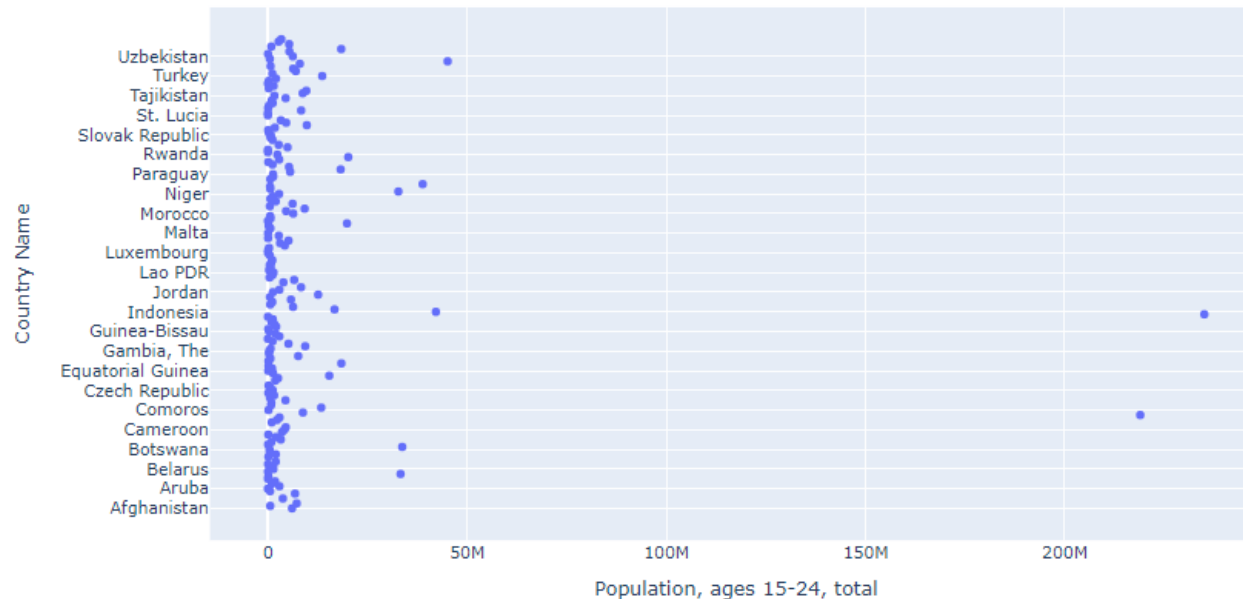
Indicator = 'Population, ages 15-24, total'#Indicator_list[0]
Annee = Annee_list[20]
df = data.loc[data['Indicator Name'] == Indicator ,:]

fig = px.scatter(df, x=df[Annee], y='Country Name' )

fig.update_layout(
    title=Annee,
    xaxis_title=Indicator,
    yaxis_title='Country Name')

fig.show()
```

2010



```
1 data = data_EdStatsData.copy()
2 data.head(2)
```

	Country Name	Country Code	Indicator Name	Indicator Code	1970	1971	1972	1973	1974
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	NaN	NaN	NaN	NaN	NaN
1	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.F	NaN	NaN	NaN	NaN	NaN



1. Inspection des données



Données à disposition : 5 fichiers .csv

EdStatsData.csv

886930 lignes et 70 colonnes

Il donne les valeurs des indicateurs sur les années entre 1970 et 2100 pour chaque pays.

EdStatsCountry.csv

241 lignes et 32 colonnes

Il donne des informations sur chaque pays.

EdStatsSeries.csv

3665 lignes et 21 colonnes

Il donne des informations sur les indicateurs statistiques.

EdStatsCountry-Series.csv

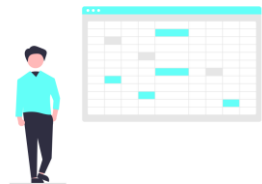
613 lignes et 4 colonnes

Il donne la source des statistiques pour chaque pays et chaque indicateur.

EdStatsFootNote.csv

643638 lignes et 5 colonnes

Il donne une information sur l'origine d'estimation des valeurs indiquées dans le EdStatsData.csv



OPENCLASSROOMS

1. Inspection des données : EdStatCountry



2. EdStatCountry : Analyse de la forme des données

EdStatsCountry.csv

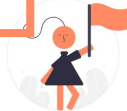
241 lignes et 32 colonnes

Il donne des informations sur chaque pays.

Une classification des pays par région et par groupe de revenu.

```
1 Country = data_EdStatsCountry.copy()
2 Country.head(2)
```

	Country Code	Short Name	Table Name	Long Name	2-alpha code	Currency Unit	Special Notes	Region	Income Group	WB-2 code	National accounts base year	National accounts reference year	SNA price valuation	Lending category	Other groups	System of National Accounts	Alternative conversion factor	PPP survey year	Balance of Payments Manual in use	External debt Reporting status	System of trade	Government Accounting concept	IMF data dissemination standard	Latest population census	Latest household survey	Source of most recent Income and expenditure data	
0	ABW	Aruba	Aruba	Aruba	AW	Aruban florin	SNA data for 2000-2011 are updated from official data	Latin America & Caribbean	High income: nonOECD	AW	2000	NaN	Value added at basic prices (VAB)	NaN	NaN	Country uses the 1993 System of National Accounts	NaN	NaN	IMF Balance of Payments Manual, 6th edition.	NaN	Special trade system	NaN	NaN	2010	NaN	NaN	
1	AFG	Afghanistan	Afghanistan	Islamic State of Afghanistan	AF	Afghan afghani	Fiscal year end: March 20; reporting period for...	South Asia	Low income	AF	2002/03	NaN	Value added at basic prices (VAB)	IDA	HIPC	Country uses the 1993 System of National Accounts	NaN	NaN	NaN	Actual	General trade system	Consolidated central government	General Data Dissemination System (GDSDS)	1979	Multiple Indicator Cluster Survey (MICS), 2010/11	Integrated household survey (IHS), 2008	...



academy



OPENCLASSROOMS

1. Inspection des données : EdStatCountry



2. EdStatCountry : Analyse de la forme des données

EdStatsCountry.csv

241 lignes et 32 colonnes

Il donne des informations sur chaque pays.

Une classification des pays par région et par groupe de revenu.



EdStatsCountry: Dictionnaire Région : liste des pays

```
1 # Liste des Régions
2 df = Country.loc[:,['Country Code', 'Region']]
3 Region_liste = df['Region'].dropna().unique().tolist()
4 print("Il y a ", len(Region_liste), " Régions: ")
5 Region_liste
```

Il y a 7 Régions:

```
: ['Latin America & Caribbean',
  'South Asia',
  'Sub-Saharan Africa',
  'Europe & Central Asia',
  'Middle East & North Africa',
  'East Asia & Pacific',
  'North America']
```



EdStatsCountry: Dictionnaire Income Groupe : liste des pays

```
1 # Liste des Income Groups
2 df = Country.loc[:,['Country Code', 'Income Group']]
3 IncomeG_liste = df['Income Group'].dropna().unique().tolist()
4 print("Il y a ", len(IncomeG_liste), " Income groups: ")
5 IncomeG_liste
```

Il y a 5 Income groups:

```
['High income: nonOECD',
 'Low income',
 'Upper middle income',
 'Lower middle income',
 'High income: OECD']
```



1. Inspection des données



Données à disposition : 5 fichiers .csv

EdStatsData.csv

886930 lignes et 70 colonnes

Il donne les valeurs des indicateurs sur les années entre 1970 et 2100 pour chaque pays.

EdStatsCountry.csv

241 lignes et 32 colonnes

Il donne des informations sur chaque pays.

EdStatsSeries.csv

3665 lignes et 21 colonnes

Il donne des informations sur les indicateurs statistiques.

EdStatsCountry-Series.csv

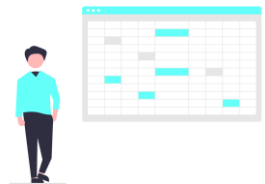
613 lignes et 4 colonnes

Il donne la source des statistiques pour chaque pays et chaque indicateur.

EdStatsFootNote.csv

643638 lignes et 5 colonnes

Il donne une information sur l'origine d'estimation des valeurs indiquées dans le EdStatsData.csv



OPENCLASSROOMS

1. Inspection des données : EdStatsSeries.csv

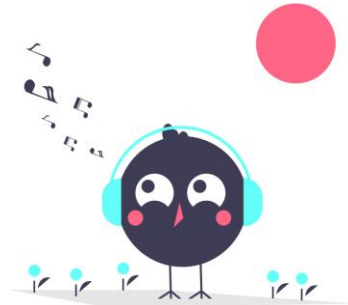
3. EdStatsSeries : Analyse de la forme des données

EdStatsSeries.csv

3665 lignes et 21 colonnes

Il donne des informations sur les indicateurs statistiques.

Une classification des indicateurs par thèmes.



```
1 Series = data_EdStatsSeries.copy()
2 Series.head(2)
```

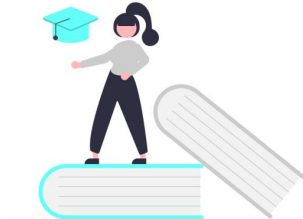
	Series Code	Topic	Indicator Name	Short definition	Long definition	Unit of measure	Periodicity	Base Period	Other notes	Aggregation method	Limitations and exceptions	Notes from original source	General comments	Source	Statistical concept and methodology	Development relevance	Related source links	Other web links	Related indicators	License Type	Unnamed: 20
0	BAR.NOED.1519.FE.ZS	Attainment	Barro-Lee: Percentage of female population age 15-19 with...	Percentage of female population age 15-19 with...	Percentage of female population age 15-19 with...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Robert J. Barro and Jong-Wha Lee: http://www.b...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	BAR.NOED.1519.ZS	Attainment	Barro-Lee: Percentage of population age 15-19 with no edu...	Percentage of population age 15-19 with no edu...	Percentage of population age 15-19 with no edu...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Robert J. Barro and Jong-Wha Lee: http://www.b...	NaN	NaN	NaN	NaN	NaN	NaN	NaN



1. Inspection des données : EdStatsSeries.csv

3. EdStatsSeries : Analyse de la forme des données

EdStatsSeries.csv

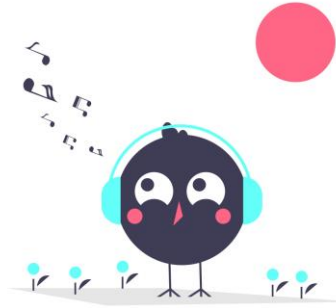


academy

```
1 # Liste des Topic
2 Topic_list = Series['Topic'].dropna().unique().tolist()
3 print("Il y a ", len(Topic_list), " Topic. ")
4 Topic_list
```

Il y a 37 Topic.

```
['Attainment',
'Education Equality',
'Infrastructure: Communications',
'Learning Outcomes',
'Economic Policy & Debt: National accounts: US$ at current prices: Aggregate indicators',
'Economic Policy & Debt: National accounts: US$ at constant 2010 prices: Aggregate indicators',
'Economic Policy & Debt: Purchasing power parity',
'Economic Policy & Debt: National accounts: Atlas GNI & GNI per capita',
'Teachers',
'Education Management Information Systems (SABER)',
'Early Child Development (SABER)',
'Engaging the Private Sector (SABER)',
'School Health and School Feeding (SABER)',
'School Autonomy and Accountability (SABER)',
'School Finance (SABER)',
'Student Assessment (SABER)',
'Teachers (SABER)',
'Tertiary Education (SABER)',
'Workforce Development (SABER)',
'Literacy',
'Background',
'Primary',
'Secondary',
'Tertiary',
'Early Childhood Education',
'Pre-Primary',
'Expenditures',
'Health: Risk factors',
'Health: Mortality',
'Social Protection & Labor: Labor force structure',
'Labor',
'Social Protection & Labor: Unemployment',
'Health: Population: Structure',
'Population',
'Health: Population: Dynamics',
'EMIS',
'Post-Secondary/Non-Tertiary']
```



OPENCLASSROOMS

1. Inspection des données



Données à disposition : 5 fichiers .csv

EdStatsData.csv

886930 lignes et 70 colonnes

Il donne les valeurs des indicateurs sur les années entre 1970 et 2100 pour chaque pays.

EdStatsCountry.csv

241 lignes et 32 colonnes

Il donne des informations sur chaque pays.

EdStatsSeries.csv

3665 lignes et 21 colonnes

Il donne des informations sur les indicateurs statistiques.

EdStatsCountry-Series.csv

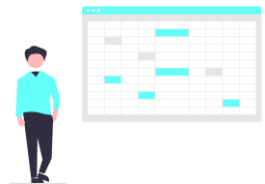
613 lignes et 4 colonnes

Il donne la source des statistiques pour chaque pays et chaque indicateur.

EdStatsFootNote.csv

643638 lignes et 5 colonnes

Il donne une information sur l'origine d'estimation des valeurs indiquées dans le EdStatsData.csv



OPENCLASSROOMS

1. Inspection des données : EdStatsCountry-Series.csv

4. EdStatsCountry-Series : Analyse de la forme des données

EdStatsCountry-Series.csv

613 lignes et 4 colonnes

Il donne la source des statistiques pour chaque pays et chaque indicateur.



```
1 Country_series = data_EdStatsCountrySeries.copy()
2 Country_series.head(3)
```

	CountryCode	SeriesCode	DESCRIPTION	Unnamed: 3
0	ABW	SP.POP.TOTL	Data sources : United Nations World Population...	NaN
1	ABW	SP.POP.GROW	Data sources: United Nations World Population ...	NaN
2	AFG	SP.POP.GROW	Data sources: United Nations World Population ...	NaN



```
1 Country_series['CountryCode'].unique().shape[0]
```

211

```
1 Country_series['SeriesCode'].unique().shape[0]
```

21

```
1 Country_series['DESCRIPTION'].unique().shape[0]
```

97



Pas très utile pour le projet



1. Inspection des données



Données à disposition : 5 fichiers .csv

EdStatsData.csv

886930 lignes et 70 colonnes

Il donne les valeurs des indicateurs sur les années entre 1970 et 2100 pour chaque pays.

EdStatsCountry.csv

241 lignes et 32 colonnes

Il donne des informations sur chaque pays.

EdStatsSeries.csv

3665 lignes et 21 colonnes

Il donne des informations sur les indicateurs statistiques.

EdStatsCountry-Series.csv

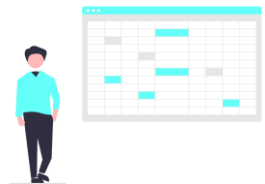
613 lignes et 4 colonnes

Il donne la source des statistiques pour chaque pays et chaque indicateur.

EdStatsFootNote.csv

643638 lignes et 5 colonnes

Il donne une information sur l'origine d'estimation des valeurs indiquées dans le EdStatsData.csv



OPENCLASSROOMS

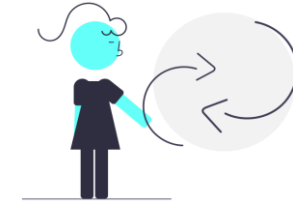
1. Inspection des données : EdStatsFootNote.csv

5. EdStatsFootNote : Analyse de la forme des données

EdStatsFootNote.csv

643638 lignes et 5 colonnes

Il donne une information sur l'origine d'estimation des valeurs indiquées dans le EdStatsData.csv



```
1 FootNote = data_EdStatsFootNote.copy()
2 FootNote.head()
```

	CountryCode	SeriesCode	Year	DESCRIPTION	Unnamed: 4
0	ABW	SE.PRE.ENRL.FE	YR2001	Country estimation.	NaN
1	ABW	SE.TER.TCHR.FE	YR2005	Country estimation.	NaN
2	ABW	SE.PRE.TCHR.FE	YR2000	Country estimation.	NaN
3	ABW	SE.SEC.ENRL.GC	YR2004	Country estimation.	NaN
4	ABW	SE.PRE.TCHR	YR2006	Country estimation.	NaN



Pas très utile pour le projet

 academy



OPENCLASSROOMS

1. Inspection des données



BANQUE MONDIALE

Conclusion



EdStatsData.csv 886930 lignes et 70 colonnes → **354263 lignes, 26 colonnes**
Pays : **217 pays, 25 groupes de pays**
Indicateurs : **3665**
Données par années : 1970-2100 → **1990 - 2015 (2010 +)**.

EdStatsCountry.csv 241 lignes et 32 colonnes
Classification des pays **par région (7)** et **par groupe de revenu (5)**.

EdStatsSeries.csv 3665 lignes et 21 colonnes
Une classification des indicateurs **par thèmes (37)**.

EdStatsCountry-Series.csv 613 lignes et 4 colonnes
Il donne la source des statistiques pour chaque pays et chaque indicateur.

EdStatsFootNote.csv 643638 lignes et 5 colonnes
Il donne une information sur l'origine d'estimation des valeurs indiquées dans le EdStatsData.csv



OPENCLASSROOMS

Analysez des données de systèmes éducatifs

- **Partie 1: Inspection des données**

- EdStatsData.csv
- EdStatsCountry.csv
- EdStatsSeries.csv
- EdStatsCountry-Series.csv
- EdStatsFootNote.csv

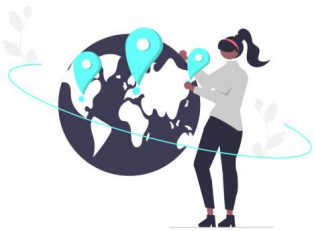


BANQUE MONDIALE

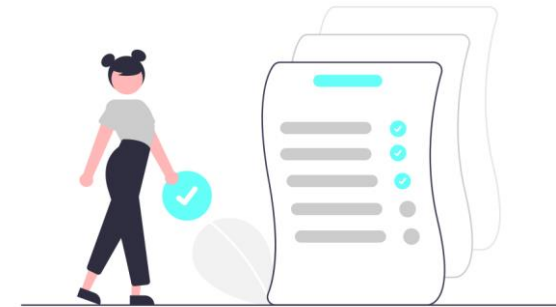
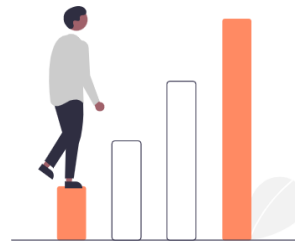
- **Partie 2: Exploration des données**

- Stratégie d'expansion : sélection des indicateurs
- Agrégation temporelle des indicateurs
- Calcul d'un score par pays
- Classement final des pays

- **Conclusion**

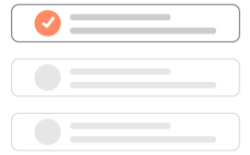


academy



OPENCLASSROOMS

2. Exploration des données



1. Stratégie d'expansion : sélection des indicateurs



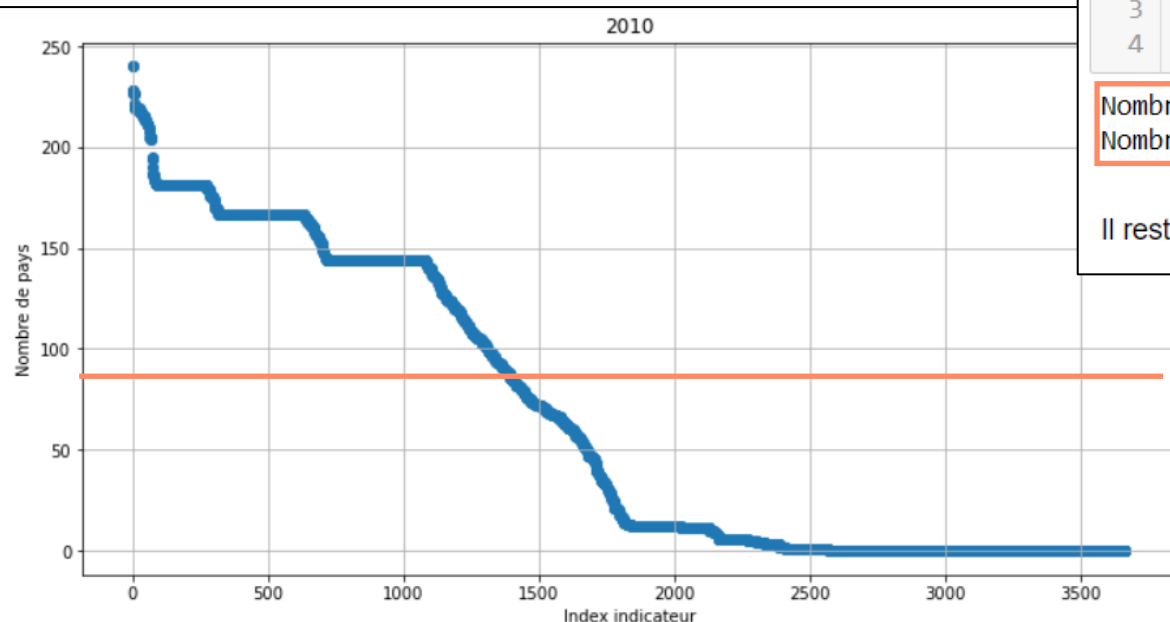
Suppression des indicateurs présents dans moins de **80 pays** (soit 1/3 des pays).

Pour la suite, je prendrai l'année **2010** qui a le moins de valeur manquante pour sélectionner mes indicateurs !

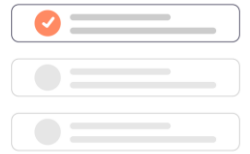
```
1 #Indicator_dp DataFrame avec les indicateurs présents dans plus de 80 pays en 2010
2 print("Nombre d'indicateur avant la suppression: ",Indicator_dp.shape[0])
3 Indicator_dp = Indicator_dp.loc[Indicator_dp['Nbre pays en 2010'] > 80,:]
4 print("Nombre d'indicateur après la suppression: ",Indicator_dp.shape[0])
```

Nombre d'indicateur avant la suppression: 3665
Nombre d'indicateur après la suppression: 1432

Il reste 1432 indicateurs à étudier!!



2. Exploration des données



1. Stratégie d'expansion : sélection des indicateurs

RAPPEL

ACADEMY propose des contenus de formation **en ligne** pour un public de niveau **lycée** et **université**.

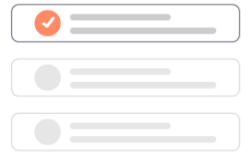
- **Problématique:** Dans quels pays l'entreprise doit-elle opérer en priorité ?
- **Stratégie:** Notre stratégie sera de classer les pays en prenant en compte 3 critères.



- **Démographique** : la population total, la population dans l'âge des lycéens et des universitaires
- **Infrastructure** de communication : L'accès à internet
- **Richesse** : Les moyens pour acheter la formation (Income groupe).



2. Exploration des données



1. Stratégie d'expansion : sélection des indicateurs

Pour faire la sélection, on se basera sur le classement des indicateurs retenus par thème '**Topic**'.



Repérage des indicateurs démographiques

```
# indicateurs démographiques retenus!  
Indicateur_demographique = Population + Lycee + University  
print("Nombre d'indicateurs démographiques retenus: ", len(Indicateur_demographique))  
display(Indicateur_demographique)
```

Nombre d'indicateurs démographiques retenus: 6

```
['Population, total',  
'Population, ages 15-24, total',  
'Population of the official age for secondary education, both sexes (number)',  
'Population of the official age for tertiary education, both sexes (number)',  
'Enrolment in secondary education, both sexes (number)',  
'Enrolment in tertiary education, all programmes, both sexes (number)']
```



Repérage des indicateurs d'infrastructure

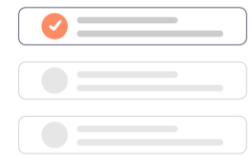
```
# liste des indicateurs infrastructure communications  
print("Nombre d'indicateurs d'infrastructure retenus: ",  
      len(Indicateur_infrastructure))  
display(Indicateur_infrastructure)
```

Nombre d'indicateurs d'infrastructure retenus: 1

```
['Internet users (per 100 people)']
```



2. Exploration des données



1. Stratégie d'expansion : sélection des indicateurs

On se basera pour le classement de richesse sur le **groupe de revenu** de chaque pays.



Repérage des indicateurs de richesse

```
df = data_EdStatsCountry.copy()
df = df.loc[:,['Long Name', 'Income Group']]
Country_groupIC_dp = df.groupby(['Income Group'])['Long Name'].apply(list).reset_index()
Country_groupIC_dp.rename(columns={'Long Name' : 'Country list'}, inplace=True)

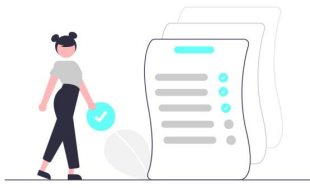
Country_groupIC_dp
```

	Income Group	Country list
0	High income: OECD	[Commonwealth of Australia, Republic of Austri...
1	High income: nonOECD	[Aruba, Principality of Andorra, United Arab E...
2	Low income	[Islamic State of Afghanistan, Republic of Bur...
3	Lower middle income	[Republic of Armenia, Plurinational State of B...
4	Upper middle income	[People's Republic of Angola, Republic of Alba...



Agrégation temporelle des indicateurs

Analysez des données de systèmes éducatifs



- **Partie 1: Inspection des données**

- EdStatsData.csv
- EdStatsCountry.csv
- EdStatsSeries.csv
- EdStatsCountry-Series.csv
- EdStatsFootNote.csv



BANQUE MONDIALE

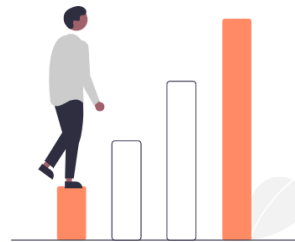
- **Partie 2: Exploration des données**

- Stratégie d'expansion : sélection des indicateurs
- Agrégation temporelle des indicateurs
- Calcul d'un score par pays
- Classement final des pays

- **Conclusion**



academy



OPENCLASSROOMS

2. Exploration des données



2. Agrégation temporelle des indicateurs



Agrégation des données entre 1990 et 2015 à la **dernière valeur renseignée** !

```
data_agg = data_etude.iloc[:,4]
#Agrégation des données entre 1990 et 2015 à la dernière valeur renseignée !
data_agg['Valeur'] = data_etude.agg(lambda x: x.dropna()[-1], axis = 1)
```

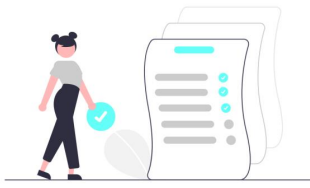
```
print(data_agg.shape)
data_agg.head(8)
```

(1572, 5)

	Country Name	Country Code	Indicator Name	Indicator Code	Valeur
56	Arab World	ARB	Enrolment in secondary education, both sexes (number)	SE.SEC.ENRL	3.097225e+07
63	Arab World	ARB	Enrolment in tertiary education, all programmes, both sexes (number)	SE.TER.ENRL	9.966484e+06
135	Arab World	ARB	Internet users (per 100 people)	IT.NET.USER.P2	3.686860e+01
257	Arab World	ARB	Population of the official age for secondary education, both sexes (number)	SP.SEC.TOTL.IN	4.378628e+07
260	Arab World	ARB	Population of the official age for tertiary education, both sexes (number)	SP.TER.TOTL.IN	3.537356e+07
287	Arab World	ARB	Population, total	SP.POP.TOTL	3.697615e+08
425	East Asia & Pacific	EAS	Enrolment in secondary education, both sexes (number)	SE.SEC.ENRL	1.542121e+08
432	East Asia & Pacific	EAS	Enrolment in tertiary education, all programmes, both sexes (number)	SE.TER.ENRL	6.909780e+07



Analysez des données de systèmes éducatifs



- **Partie 1: Inspection des données**

- EdStatsData.csv
- EdStatsCountry.csv
- EdStatsSeries.csv
- EdStatsCountry-Series.csv
- EdStatsFootNote.csv



BANQUE MONDIALE

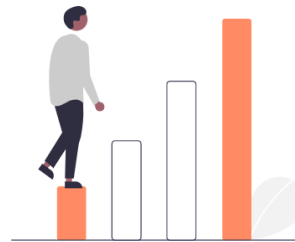
- **Partie 2: Exploration des données**

- Stratégie d'expansion : sélection des indicateurs
- Agrégation temporelle des indicateurs
- Calcul d'un score par pays
- Classement final des pays

- **Conclusion**



academy



OPENCLASSROOMS

2. Exploration des données

3. Calcul d'un score par pays

Calcul d'un score démographiques



```
# DataFrame avec que les indicateurs démographiques
data_demog = data.loc[data['Indicator Name'].isin(Indicateur_demographique) , : ]
data_demog_piv = data_demog.pivot_table(index='Country Name', columns='Indicator Name', values = 'Valeur', aggfunc='sum')
data_demog_piv
```

Indicator Name	Enrolment in secondary education, both sexes (number)	Enrolment in tertiary education, all programmes, both sexes (number)	Population of the official age for secondary education, both sexes (number)	Population of the official age for tertiary education, both sexes (number)	Population, ages 15-24, total	Population, total
Country Name						
Afghanistan	2698816.0	262874.0	4850112.0	3199607.0	7252785.0	3.373649e+07
Albania	315079.0	160527.0	329011.0	276247.0	556269.0	2.880703e+06
Algeria	4572513.0	1289474.0	4056674.0	3492401.0	6467818.0	3.987153e+07
American Samoa	3643.0	1607.0	NaN	NaN	NaN	5.553700e+04
Andorra	4395.0	501.0				
...				
West Bank and Gaza	721414.0	221018.0				



```
data_demog_piv.describe()
```

Indicator Name	Enrolment in secondary education, both sexes (number)	Enrolment in tertiary education, all programmes, both sexes (number)	Population of the official age for secondary education, both sexes (number)	Population of the official age for tertiary education, both sexes (number)	Population, ages 15-24, total	Population, total
count	2.330000e+02	2.240000e+02	2.250000e+02	2.250000e+02	1.920000e+02	2.400000e+02
mean	1.659846e+07	6.201006e+06	2.292429e+07	1.810557e+07	6.274290e+06	2.070249e+08
std	6.424066e+07	2.325939e+07	8.707293e+07	6.876054e+07	2.338922e+07	8.025361e+08
min	1.028000e+03	1.940000e+02	1.263000e+03	8.680000e+02	2.825000e+03	1.100100e+04
25%	9.725400e+04	2.152300e+04	2.403030e+05	1.741000e+05	2.945968e+05	1.302206e+06
50%	6.434070e+05	2.176025e+05	9.289490e+05	6.970210e+05	1.158544e+06	8.590910e+06
75%	3.176320e+06	1.238699e+06	4.584447e+06	3.450626e+06	4.519916e+06	3.815176e+07
max	5.792067e+08	2.126700e+08	7.578950e+08	5.958385e+08	2.441202e+08	7.355220e+09

2. Exploration des données

3. Calcul d'un score par pays



Calcul d'un score
démographiques



Valeurs manquantes par indicateur

```
# Nombre de valeurs manquantes par variable démographique
data_demog_piv.shape[0] - data_demog_piv.describe().iloc[0,:]
```

Indicator Name	
Enrolment in secondary education, both sexes (number)	9.0
Enrolment in tertiary education, all programmes, both sexes (number)	18.0
Population of the official age for secondary education, both sexes (number)	17.0
Population of the official age for tertiary education, both sexes (number)	17.0
Population, ages 15-24, total	50.0
Population, total	2.0

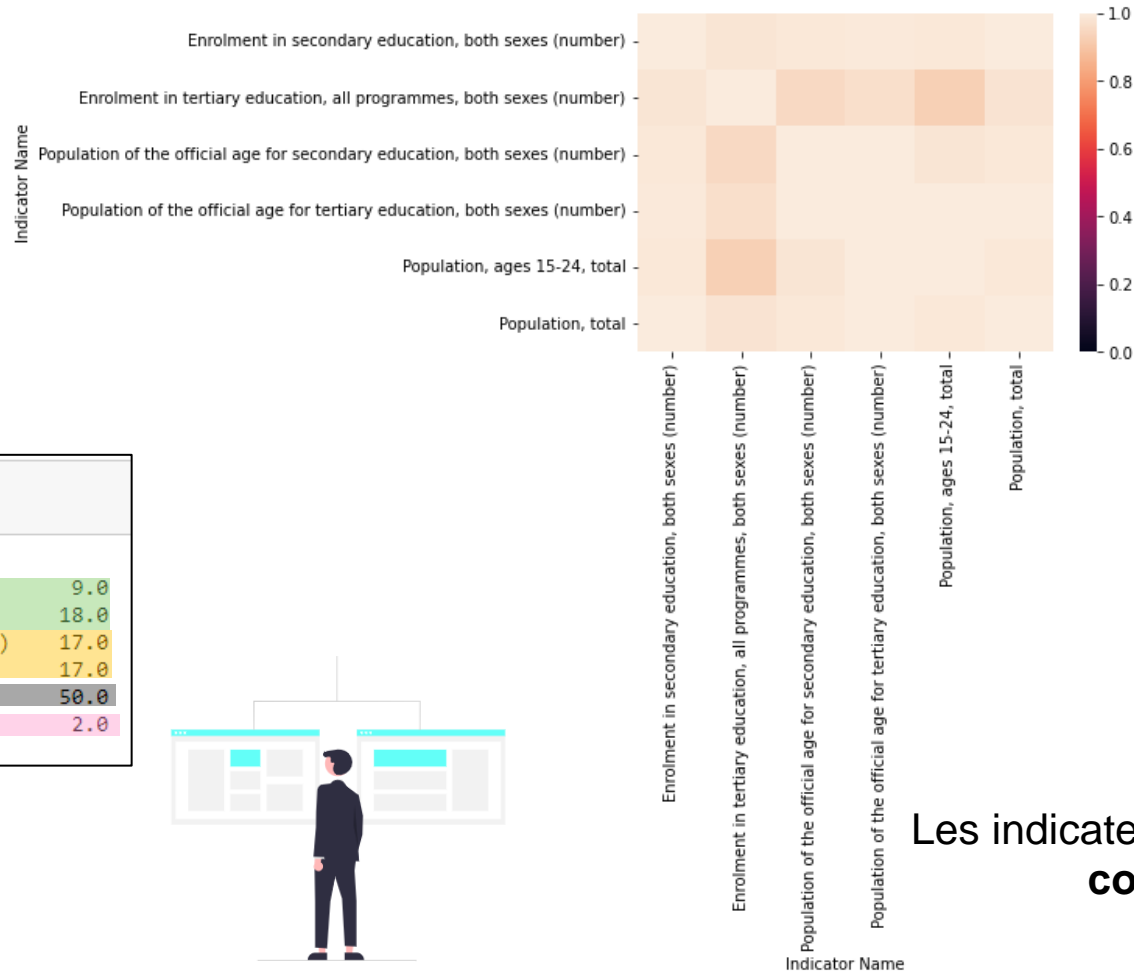
Name: count, dtype: float64



Corrélation entre les indicateurs

```
#Corrélation entre les indicateurs démographiques choisis
sns.heatmap(data_demog_piv.corr(),vmin=0, vmax=1)
```

<AxesSubplot:xlabel='Indicator Name', ylabel='Indicator Name'>



Les indicateurs démographiques sont
corrélés entre eux

OPENCLASSROOMS

2. Exploration des données



SCORE

3. Calcul d'un score par pays



Calcul d'un score
démographiques



'Population, ages 15-24, total': ne sera pas prise en compte dans le calcul du score vu qu'il y a environ 20% de valeurs manquantes.



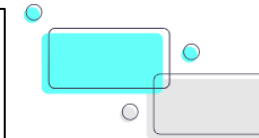
On additionne les deux variables :
'Enrolment in secondary education, both sexes (number)' et
'Enrolment in tertiary education, all programmes, both sexes (number)'
pour en avoir une seule :
'Nombre d'inscrit au lycée et à l'université'.



On additionne les deux variables :
'Population of the official age for secondary education, both sexes (number)' et
'Population of the official age for tertiary education, both sexes (number)'
pour en avoir une seule : **'Population âge de scolarisation au lycée et à l'université'**.

Indicator Name	Nombre d'inscrit au lycée et à l'université	Population âge de scolarisation au lycée et à l'université	Population, total
Country Name			
Afghanistan	2961690.0	8.049719e+06	3.373649e+07
Albania	475606.0	6.052580e+05	2.880703e+06
Algeria	5861987.0	7.549075e+06	3.987153e+07
American Samoa	5250.0	NaN	5.553700e+04
Andorra	4896.0	7.967000e+03	7.801400e+04

academy



OPENCLASSROOMS

2. Exploration des données



SCORE

3. Calcul d'un score par pays



Calcul d'un score démographiques



Mise à l'échelle: Scaling

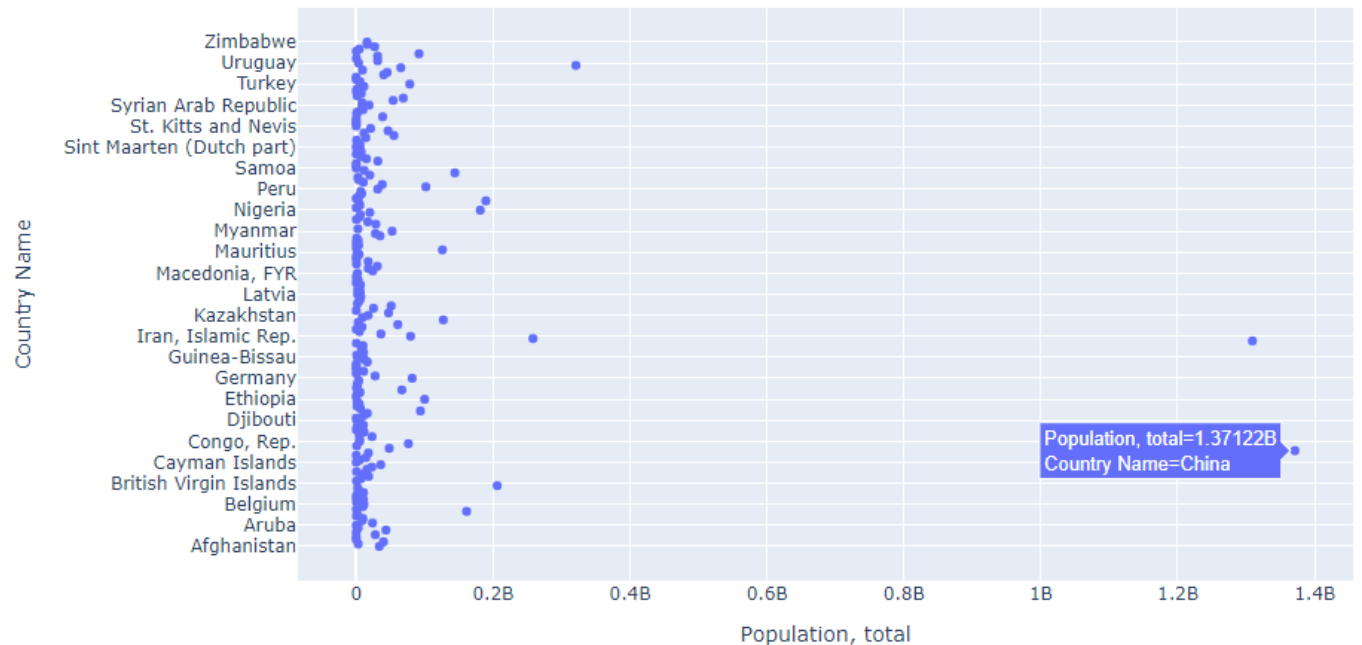
- * Scaling par groupe de pays
- * Scaling par pays



academy

Indicator Name	Nombre d'inscrit au lycée et à l'université	Population âge de scolarisation au lycée et à l'université	Population, total
Country Name			
Afghanistan	2961690.0	8.049719e+06	3.373649e+07
Albania	475606.0	6.052580e+05	2.880703e+06
Algeria	5861987.0	7.549075e+06	3.987153e+07
American Samoa	5250.0	NaN	5.553700e+04
Andorra	4896.0	7.967000e+03	7.801400e+04

```
data = data_demog_piv_red_pays.copy()
fig = px.scatter(data, x='Population, total', y=data.index)
fig.show()
```



2. Exploration des données



3. Calcul d'un score par pays



Calcul d'un score démographiques



Mise à l'échelle: Scaling

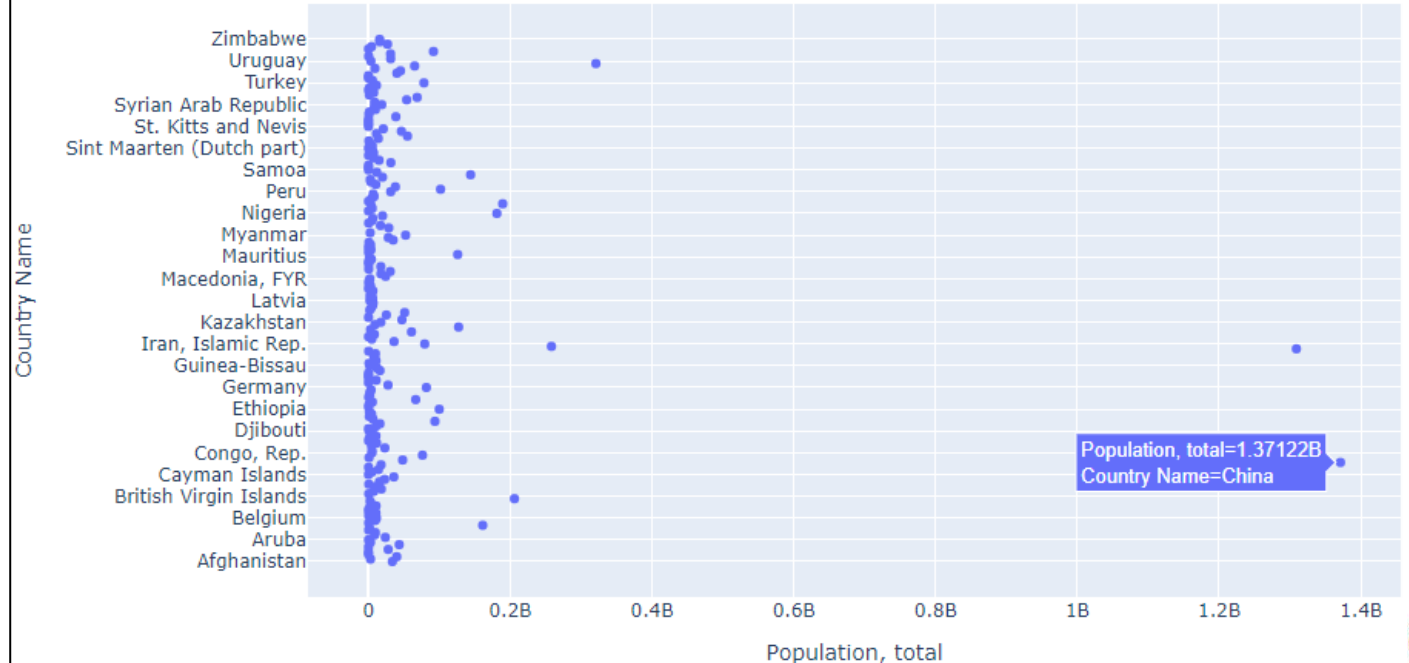
- * Scaling par groupe de pays
- * Scaling par pays



Indicator Name	Nombre d'inscrit au lycée et à l'université	Population âge de scolarisation au lycée et à l'université	Population, total
Country Name			
Afghanistan	2961690.0	8.049719e+06	3.373649e+07
Albania	475606.0	6.052580e+05	2.880703e+06
Algeria	5861987.0	7.549075e+06	3.987153e+07



```
data = data_demog_piv_red_pays.copy()
fig = px.scatter(data, x='Population, total', y=data.index)
fig.show()
```



La Chine et l'Inde ont une très grande population! Un Min-Max Scaling n'est pas très intéressant dans notre cas.

2. Exploration des données



3. Calcul d'un score par pays



Qcut scaling



Calcul d'un score démographique



Mise à l'échelle: Scaling

Scaling par pays

```
data_demog_piv_red_pays_scaled = data_demog_piv_red_pays.copy()
p = 20
for col in data_demog_piv_red_pays.columns:
    data_demog_piv_red_pays_scaled[col] = (pd.qcut(data_demog_piv_red_pays[col], p, labels=False) + 1)/p
```

data_demog_piv_red_pays_scaled

Indicator Name	Nombre d'inscrit au lycée et à l'université	Population âge de scolarisation au lycée et à l'université	Population, total
Country Name			
Afghanistan	0.80	0.85	0.85
Albania	0.45	0.40	0.40
Algeria	0.90	0.85	0.85
American Samoa	0.05	NaN	0.10



Score démographique: Moyenne des scores

```
data_demog_piv_red_pays_scaled['Score Démographique'] = data_demog_piv_red_pays_scaled.mean(axis = 1)
data_demog_piv_red_pays_scaled
```

Indicator Name	Nombre d'inscrit au lycée et à l'université	Population âge de scolarisation au lycée et à l'université	Population, total	Score Démographique
Country Name				
Afghanistan	0.80	0.85	0.85	0.833333
Albania	0.45	0.40	0.40	0.416667
Algeria	0.90	0.85	0.85	0.866667
American Samoa	0.05	NaN	0.10	0.075000
Andorra	0.05	0.05	0.10	0.066667

2. Exploration des données



3. Calcul d'un score par pays



Calcul d'un score démographique



```
df = data_demog_piv_red_pays_scaled.melt()

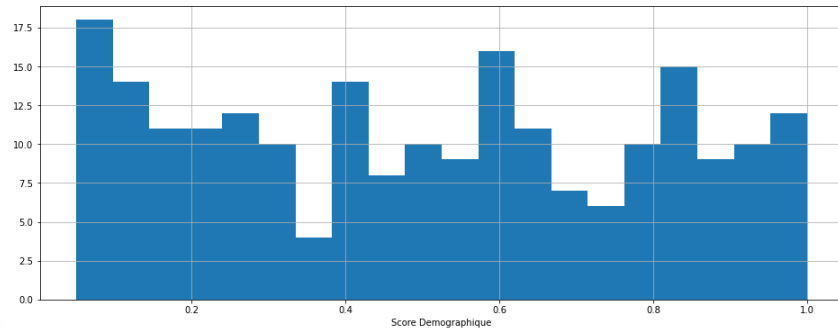
fig = px.box(df, x="Indicator Name", y="value", points="all")
fig.show()
```

SCORE

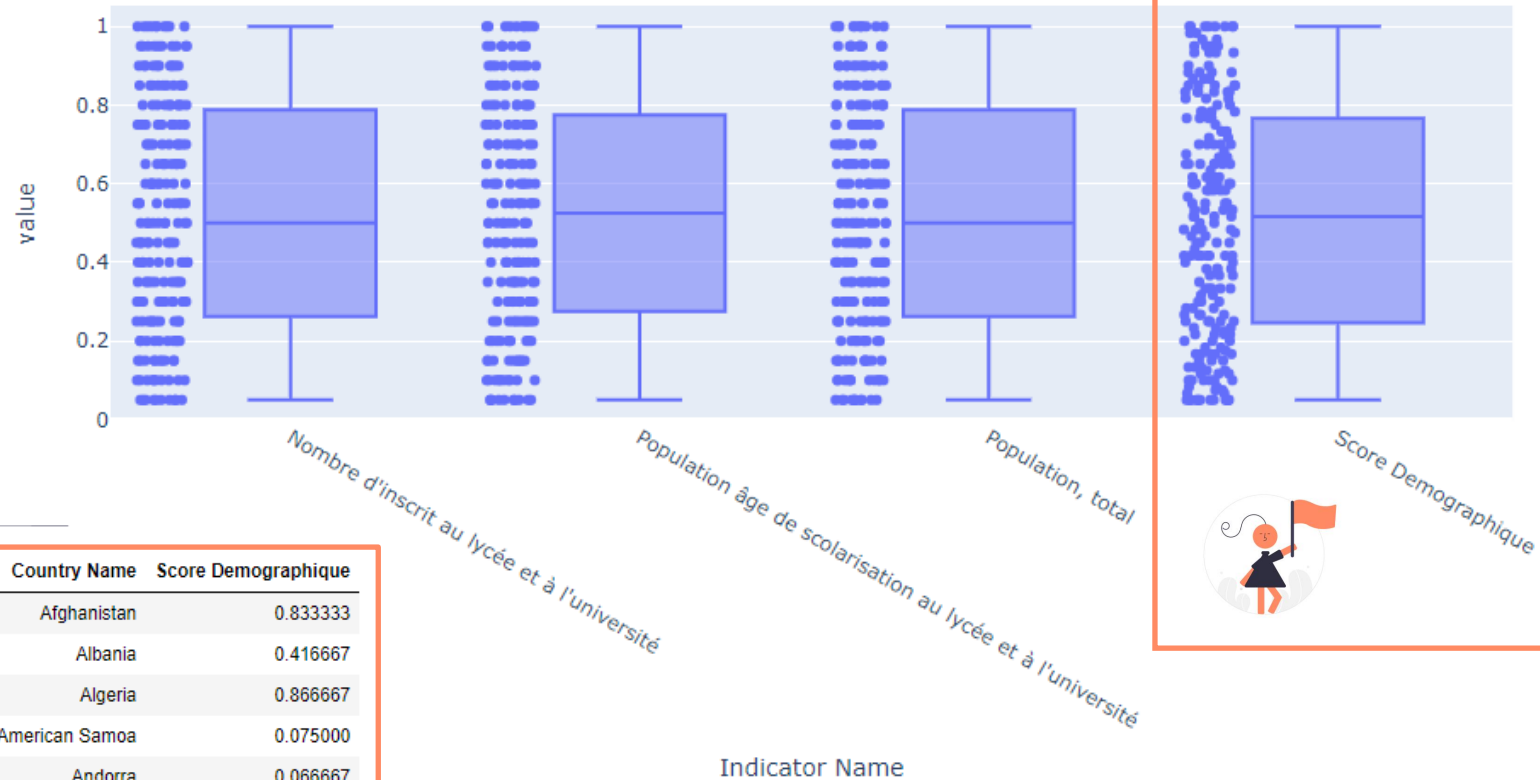
Mise à l'échelle: Scaling

Score démographique

```
plt.figure(figsize=(16,6))
plt.hist(data_scoring_demographique_pays['Score Demographique'], bins=20)
plt.xlabel('Score Demographique')
plt.grid()
```



academy



	Country Name	Score Demographique
0	Afghanistan	0.833333
1	Albania	0.416667
2	Algeria	0.866667
3	American Samoa	0.075000
4	Andorra	0.066667

2. Exploration des données



SCORE

3. Calcul d'un score par pays



Calcul d'un score
infrastructure



1 Indicateur: nombre d'utilisateurs d'internet pour 100 personnes

```
# DataFrame avec que les indicateurs infrastructures
data_inf = data.loc[data['Indicator Name'].isin(Indicateur_infrastructure) , : ]
data_inf_piv = data_inf.pivot_table(index='Country Name', columns='Indicator Name', values = 'Valeur', aggfunc='sum')
data_inf_piv
```

Indicator Name	Internet users (per 100 people)
Country Name	
Afghanistan	8.260000
Albania	63.252933
Algeria	38.200000
American Samoa	0.000000
Andorra	96.910000
...	...
West Bank and Gaza	57.424192
World	43.198456
Yemen, Rep.	24.085409
Zambia	21.000000
Zimbabwe	22.742818



academy

OPENCLASSROOMS

2. Exploration des données



SCORE

3. Calcul d'un score par pays



```
data_inf_piv_pays = data_inf_piv.loc[~data_inf_piv.index.isin(groupe_liste),:]
```

```
data_inf_piv_pays.shape
```

(208, 1) Sur 217 pays



Calcul d'un score
infrastructure



Mise à l'échelle: Scaling

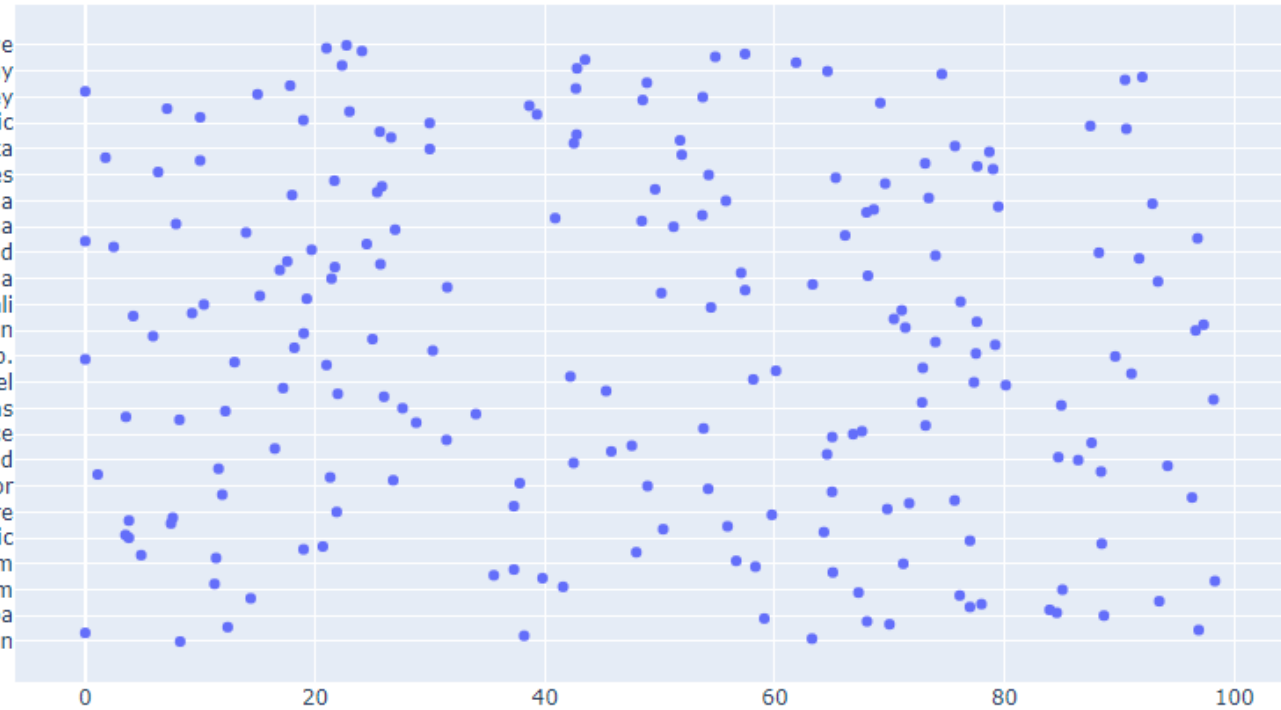
- * Scaling par groupe de pays
- * Scaling par pays



```
df = data_inf_piv_pays.copy()
fig = px.scatter(df, x='Internet users (per 100 people)', y=df.index)
fig.show()
```

Country Name

Zimbabwe
Uruguay
Turkey
Syrian Arab Republic
Sri Lanka
Seychelles
Romania
Panama
New Zealand
Mongolia
Mali
Liechtenstein
Korea, Rep.
Israel
Honduras
Greece
Finland
Ecuador
Cote d'Ivoire
Central African Republic
Brunei Darussalam
Belgium
Aruba
Afghanistan



Internet users (per 100 people)

2. Exploration des données



3. Calcul d'un score par pays

MinMax scaling 😊

SCORE



Calcul d'un score infrastructure



Mise à l'échelle: Scaling
Scaling par pays

```
# Min Max scaler
scaler = preprocessing.MinMaxScaler()
arr_scaled = scaler.fit_transform(data_inf_piv_pays)
data_inf_piv_scaled_pays = pd.DataFrame(arr_scaled, columns=data_inf_piv_pays.columns, index=data_inf_piv_pays.index)
data_inf_piv_scaled_pays
```

Country Name	Score_infrastructure
Afghanistan	0.084008
Albania	0.643314
Algeria	0.388513
American Samoa	0.000000
Andorra	0.985623
...	...
Virgin Islands (U.S.)	0.557741
West Bank and Gaza	0.584033
Yemen, Rep.	0.244961
Zambia	0.213580
Zimbabwe	0.231306



Score infrastructure

2. Exploration des données



3. Calcul d'un score par pays

MinMax scaling 😊

SCORE



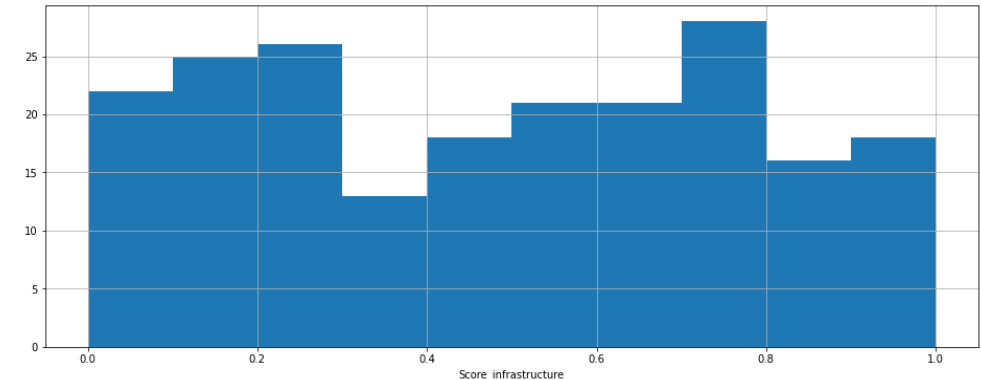
Calcul d'un score infrastructure



Mise à l'échelle: Scaling
Scaling par pays

```
# Min Max scaler
scaler = preprocessing.MinMaxScaler()
arr_scaled = scaler.fit_transform(data_inf_piv_pays)
data_inf_piv_scaled_pays = pd.DataFrame(arr_scaled, columns=data_inf_piv_pays.columns, index=data_inf_piv_pays.index)
data_inf_piv_scaled_pays
```

Country Name	Score_infrastructure
Afghanistan	0.084008
Albania	0.643314
Algeria	0.388513
American Samoa	0.000000
Andorra	0.985623
...	...
Virgin Islands (U.S.)	0.557741
West Bank and Gaza	0.584033
Yemen, Rep.	0.244961
Zambia	0.213580
Zimbabwe	0.231306



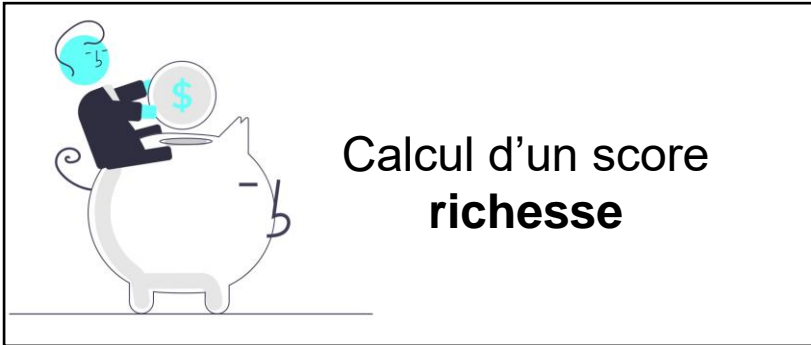
Score infrastructure

2. Exploration des données



SCORE

3. Calcul d'un score par pays



On utilisera les données de **EdStatsCountry** en gardant que les colonnes: nom de pays 'Table Name' et '**Income Groupe**'.



On attribuera à chaque pays un score comme suit:

- np.nan: ----- 0 ,
- 'Low income': ----- 0.2,
- 'Lower middle income': ----- 0.5,
- 'Upper middle income': ----- 0.8,
- 'High income: nonOECD': --- 1,
- 'High income: OECD': ----- 1.



academy




OPENCLASSROOMS

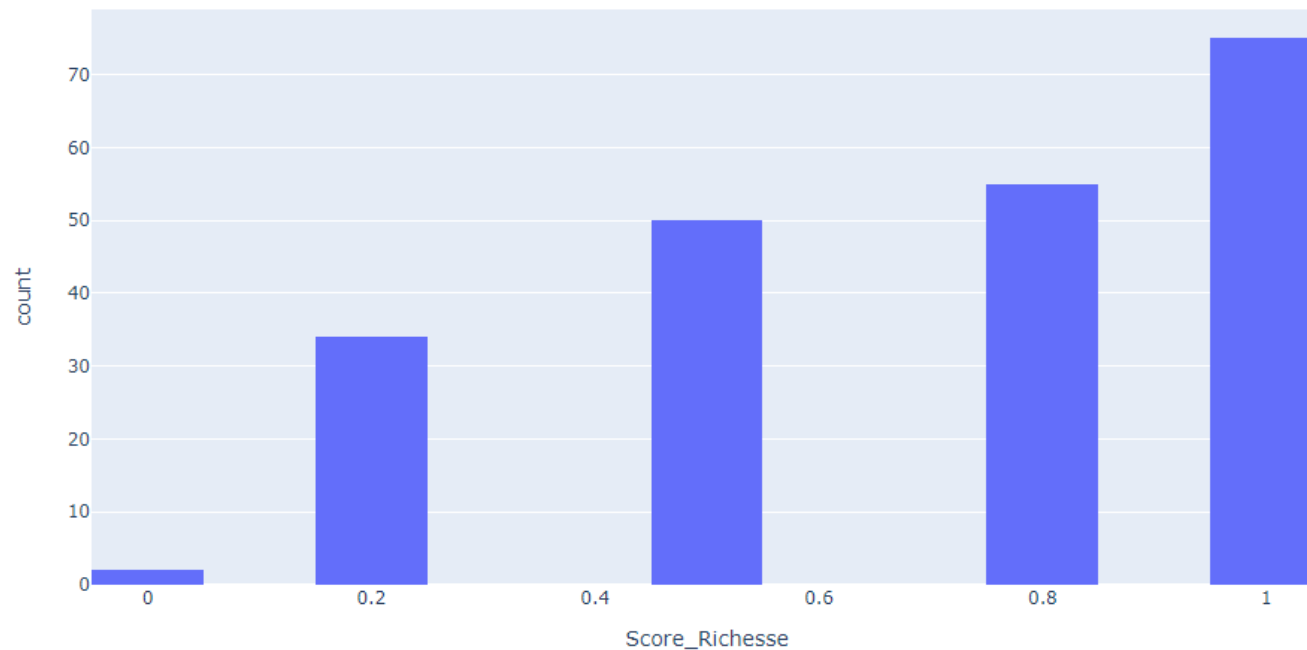
2. Exploration des données



3. Calcul d'un score par pays



Calcul d'un score
richesse



Country Name	Score_Richesse
Aruba	1.0
Afghanistan	0.2
Angola	0.8
Albania	0.8
Andorra	1.0
...	...
Kosovo	0.5
Yemen, Rep.	0.5
South Africa	0.8
Zambia	0.5
Zimbabwe	0.2

SCORE



OPENCLASSROOMS

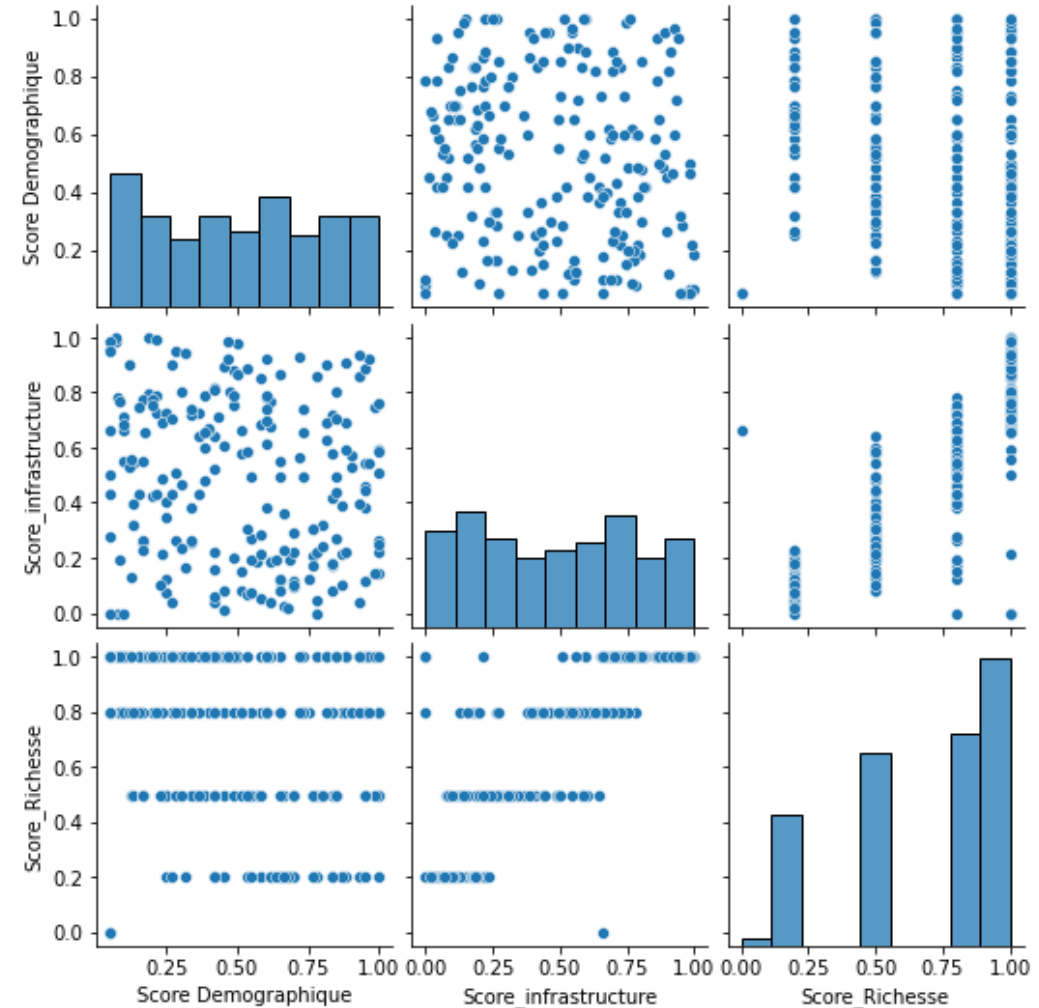
2. Exploration des données



3. Calcul d'un score par pays



	Country Name	Score Demographique	Score_infrastructure	Score_Richesse
0	Afghanistan	0.833333	0.084008	0.2
1	Albania	0.416667	0.643314	0.8
2	Algeria	0.866667	0.388513	0.8
3	American Samoa	0.075000	0.000000	0.8
4	Andorra	0.066667	0.985623	1.0
5	Angola	0.750000	0.126114	0.8
6	Antigua and Barbuda	0.100000	0.711935	1.0
7	Argentina	0.883333	0.692032	0.8
8	Armenia	0.383333	0.601101	0.5
9	Aruba	0.116667	0.901729	1.0
10	Australia	0.783333	0.860023	1.0

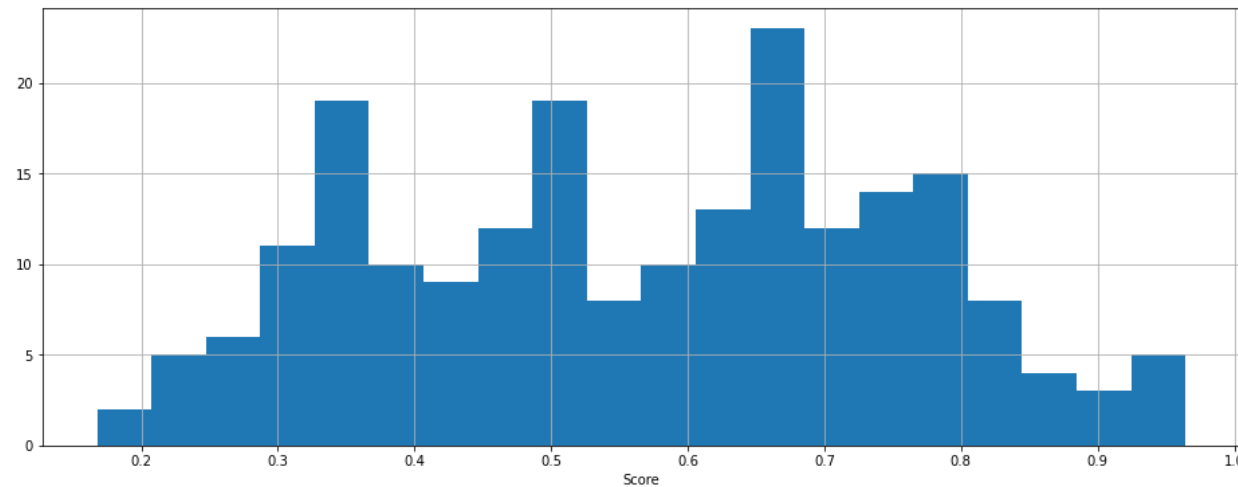


2. Exploration des données

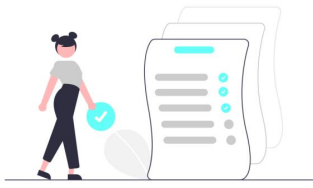
3. Calcul d'un score par pays



Score pays : moyenne arithmétique des 3 scores



Analysez des données de systèmes éducatifs



- **Partie 1: Inspection des données**

- EdStatsData.csv
- EdStatsCountry.csv
- EdStatsSeries.csv
- EdStatsCountry-Series.csv
- EdStatsFootNote.csv

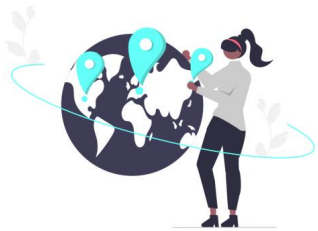


BANQUE MONDIALE

- **Partie 2: Exploration des données**

- Stratégie d'expansion : sélection des indicateurs
- Agrégation temporelle des indicateurs
- Calcul d'un score par pays
- Classement final des pays

- **Conclusion**



academy

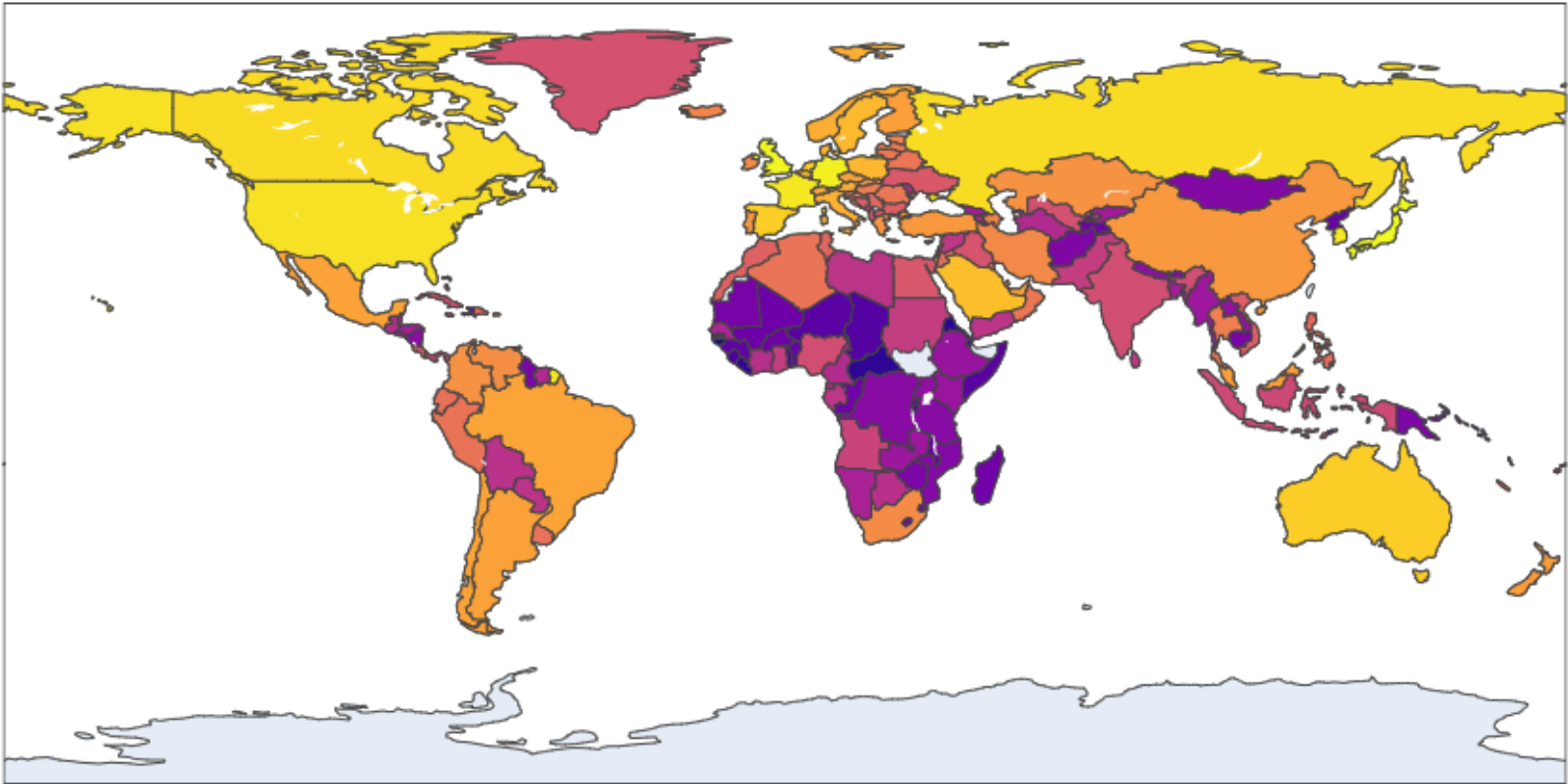


OPENCLASSROOMS

2. Exploration des données



4. Classement final des pays



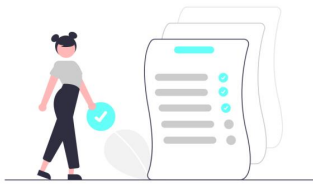
SCORE



SCORE

	Country Name	SCORE
0	Japan	0.964257
1	United Kingdom	0.956341
2	Germany	0.946944
3	Korea, Rep.	0.931701
4	France	0.931573
5	United States	0.919418
6	Russian Federation	0.909983
7	Canada	0.905484
8	Spain	0.883437
9	Netherlands	0.883182
10	Australia	0.881119
11	Saudi Arabia	0.852677
12	Sweden	0.840517
13	Belgium	0.838343
14	Poland	0.836077
15	Denmark	0.826576
16	Italy	0.824888
17	Norway	0.817092
18	Austria	0.812349
19	Switzerland	0.807680

Analysez des données de systèmes éducatifs



- **Partie 1: Inspection des données**

- EdStatsData.csv
- EdStatsCountry.csv
- EdStatsSeries.csv
- EdStatsCountry-Series.csv
- EdStatsFootNote.csv

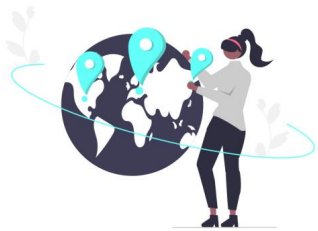


BANQUE MONDIALE

- **Partie 2: Exploration des données**

- Stratégie d'expansion : sélection des indicateurs
- Agrégation temporelle des indicateurs
- Calcul d'un score par pays
- Classement final des pays

- **Conclusion**



academy



OPENCLASSROOMS

Analysez des données de systèmes éducatifs



Conclusion



Academy : start-up de la EdTech

- des formations en ligne pour un public de niveau lycée et université.
- Objectif d'**expansion à l'international**.



Inspecter les données de systèmes éducatifs (5 fichiers csv)

-> 3 fichiers intéressants pour l'étude

-> beaucoup de données récentes sont manquantes

-> l'année exploitable la plus récente est **2015** (agrégation temporelle)



BANQUE MONDIALE



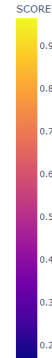
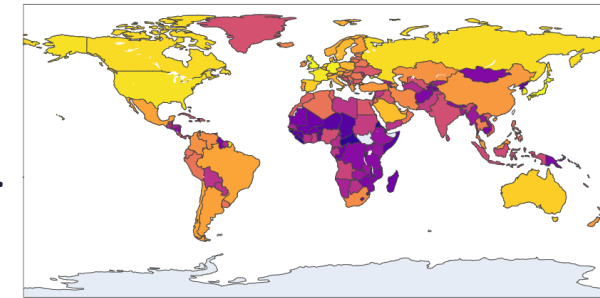
Calcul d'un score entre 0 et 1 pour chaque pays. Ce score est une moyenne arithmétique de 3 scores: un score **démographique**, un score **infrastructure** et un score **richesse**.



Un classement des pays du monde selon leur potentiel pour notre entreprise Academy.



academy



SCORE

	Country Name	SCORE
0	Japan	0.964257
1	United Kingdom	0.956341
2	Germany	0.946944
3	Korea, Rep.	0.931701
4	France	0.931573
5	United States	0.919418
6	Russian Federation	0.909983
7	Canada	0.905484
8	Spain	0.883437
9	Netherlands	0.883182
10	Australia	0.881119
11	Saudi Arabia	0.852677
12	Sweden	0.840517
13	Belgium	0.838343
14	Poland	0.836077
15	Denmark	0.826576
16	Italy	0.824888
17	Norway	0.817092
18	Austria	0.812349
19	Switzerland	0.807680

OPENCLASSROOMS