THE WORLD BANK

academy

Développement international

Analyse exploratoire

THE WORLD BANK

academy

Développement international

Analyse exploratoire

Valider jeu de données

Décrire les informations

Sélectionner les données

Indicateurs statistiques

Pertinence du jeu de données

Pays à fort potentiel clients

Evolution du potentiel

Pays cibles prioritaires

1 | Nettoyage & exploration

2 | Filtre et sélection des données

3 | Pertinence du jeu de données

4 | Scoring pays à fort potentiel clients

5 | Evolution du potentiel clients

6 | Indicateurs statistiques

1 | Nettoyage & exploration

2 | Filtre et sélection des données

3 | Pertinence du jeu de données

4 | Scoring pays à fort potentiel clients

5 | Evolution du potentiel clients

6 | Indicateurs statistiques

python™ 3.7.0 environnement

pandas 1.1.5 chargement des données et exploitation du dataframe

missingno 0.5.1 analyses générales et

pandas profiling 3.1.0 manquants

NumPy 1.21.5 modification jeu de données

matplotlib 3.5.1 visualisations graphiques et

seaborn 0.11.2 statistiques


construction
!!!

EdStatsFootnote.csv

footnote

5

643 638

Manquants 0%

CSV

commentaires

EdStatsCountry.csv

country

32

241

Manquants 30%

CSV

infos par pays (lien région)

EdStatsData.csv

data

71

886 930

Manquants 85%

CSV

données principales

EdStatsCountry-Series.csv

country_series

4

613

Manquants 0%

CSV

précisions particulières

EdStatsSeries.csv

series

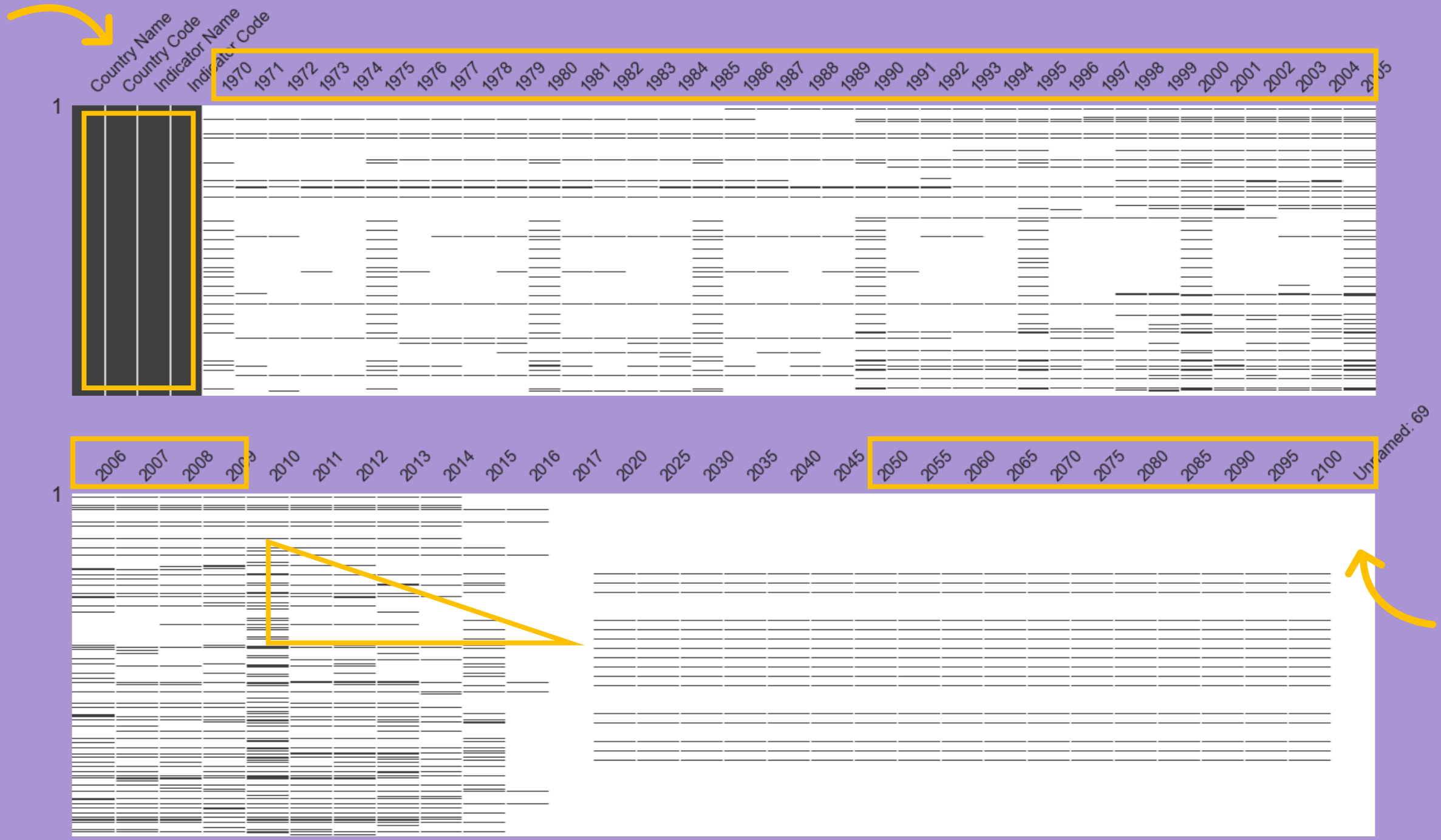
21

3665

Manquants 72%

CSV

définition des indicateurs


Country Name
Country Code
Indicator Name
Indicator Code
1
1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004
2006 2007 2008
2010 2011 2012 2013 2014 2015 2016 2017 2020 2025 2030 2035 2040 2045
2050 2055 2060 2065 2070 2075 2080 2085 2090 2095 2100
Unnamed: 69

242 Country Code = 217 pays +

	Country Name	Country Code	Indicator Name	Indicator Code
count	886930	886930	886930	886930
unique	242	242	3665	3665
top	High income	BTN	PIRLS: Fourth gr...	UIS.ROFST.H.2.Q3.F
freq	3665	3665	242	242

'Arab World', 'East Asia & Pacific',
'East Asia & Pacific (excluding high income)', 'Euro area',
'Europe & Central Asia',
'Europe & Central Asia (excluding high income)', 'European Union',
'Heavily indebted poor countries (HIPC)', 'High income',
'Latin America & Caribbean',
'Latin America & Caribbean (excluding high income)',
'Least developed countries: UN classification',
'Low & middle income', 'Low income', 'Lower middle income',
'Middle East & North Africa',
'Middle East & North Africa (excluding high income)',
'Middle income', 'North America', 'OECD members', 'South Asia',
'Sub-Saharan Africa', 'Sub-Saharan Africa (excluding high income)',
'Upper middle income', 'World', 'British Virgin Islands',
'Gibraltar', 'Nauru'

25 valeurs sans Région associée :

- 22 agrégats géographiques & économiques +
- 2 territoires : 'British Virgin Islands', 'Gibraltar' → pays +
- 1 état insulaire : 'Nauru' → pays

Projection population cible

```
proj_indicators = data.loc[data["2050"] >= 0 ]["Indicator Name"].unique()
len(proj_indicators)
```

```
308
```

```
for i in proj_indicators:
    print(i)
```

```
Wittgenstein Projection: Mean years of schooling. Age 0-19. Female
Wittgenstein Projection: Mean years of schooling. Age 0-19. Male
```

```
maskproj = (data["Indicator Name"] == \
'Wittgenstein Projection: Population in thousands by highest level of educational attainment. Upper Secondary. Total')\
& (data["2050"] >= 0)
```

```
data.loc[maskproj].isna().mean()
```

```
Country Type      0.000000
Country Name      0.000000
Country Code      0.000000
Indicator Name    0.000000
Indicator Code    0.000000
2010              0.000000
2045              0.000000
2050              0.000000
Region            0.005988
dtype: float64
```

308

Indicateurs avec projection en 2050 (horizon de business plan)

Indicateur retenu pour prévision 2050 :

'Wittgenstein Projection: Population in thousands by highest level of educational attainment. Upper Secondary. Total'

```
series["Topic"].unique()
```

```
array(['Attainment', 'Education Equality',
       'Infrastructure: Communications', 'Learning Outcomes',
       'Economic Policy & Debt: National accounts: US$ at current pric
       'Economic Policy & Debt: National accounts: US$ at constant 201
       'Economic Policy & Debt: Purchasing power parity',
       'Economic Policy & Debt: National accounts: Atlas GNI & GNI per
       'Teachers', 'Education Management Information Systems (SABER)',
       'Early Child Development (SABER)',
```

```
mask = series["Topic"]=='Attainment' #boolean indexing
series.loc[mask,"Indicator Name"].unique() #localisation [lignes : pour les avoir toutes, colonnes]
```

```
array(['Barro-Lee: Percentage of female population age 15-19 with no education',
       'Barro-Lee: Percentage of population age 15-19 with no education',
       'Barro-Lee: Percentage of female population age 15+ with no education',
       'Barro-Lee: Percentage of population age 15+ with no education',
       'Barro-Lee: Percentage of female population age 20-24 with no education',
       'Barro-Lee: Percentage of population age 20-24 with no education',
```

27 indicateurs présélectionnés sur les 3665 présents :

'Population, total', 'Population growth (annual %)', 'Population, ages 14-19, total', \
'Population, ages 15-24, total', \
'Population of the official age for upper secondary education, both sexes (number)', \
'Population of the official age for tertiary education, both sexes (number)', \
'Population of the official age for post-secondary non-tertiary education, both sexes (number)', \
'Personal computers (per 100 people)', \
'Internet users (per 100 people)', 'GDP, PPP (constant 2011 international $)' #vérifier le second indicateur, \
'GDP per capita, PPP (constant 2011 international $)', 'GNI, PPP (current international $)', \
'GNI per capita, PPP (current international $)', \
'Official entrance age to upper secondary education (years)', \
'Barro-Lee: Population in thousands, age 15+, total', \
'Barro-Lee: Percentage of population age 15+ with secondary schooling. Completed Secondary', \
'Barro-Lee: Percentage of population age 15+ with tertiary schooling. Completed Tertiary', \
'Graduates from secondary education, both sexes (number)', \
'Graduates from tertiary education, both sexes (number)', \
'Enrolment in tertiary education per 100,000 inhabitants, both sexes', \
'Enrolment in upper secondary education, both sexes (number)', \
'Enrolment in upper secondary general, both sexes (number)', \
'Net enrolment rate, secondary, both sexes (%)', \
'Percentage of enrolment in upper secondary education in private institutions (%)', \
'School life expectancy, secondary, both sexes (years)', \
'School life expectancy, tertiary, both sexes (years)', \
'Projection: Population in thousands by highest level of educational attainment. Post Secondary. Total', \
'Wittgenstein Projection: Population in thousands by highest level of educational attainment. Upper Total'

1 | Nettoyage & exploration

2 | Filtre et sélection des données

3 | Pertinence du jeu de données

4 | Scoring pays à fort potentiel clients

5 | Evolution du potentiel clients

6 | Indicateurs statistiques

Filtre et découpage

‘data’ = jeu de données réduit à la sélection d’indicateurs

```
data = data[data["Indicator Name"].isin(indicator_select)]
```

Recopie de la dernière valeur non nulle à partir de 2010
jusqu'à la dernière année (méthode fillna() + ffill)
Affectation de cette valeur à une nouvelle colonne 'last_value'

```
data["last_value"] = data.loc[:,"2010":"2050"].fillna(method='ffill', axis=1).iloc[:,-1]
```

suppression des colonnes des années
‘last_value’ = donnée la plus récente depuis 2010.

	Country Type	Country Name	Indicator Name	Region	last_value
377	Aggregate	Arab World	Barro-Lee: Perce...	NaN	NaN
379	Aggregate	Arab World	Barro-Lee: Perce...	NaN	NaN
480	Aggregate	Arab World	Barro-Lee: Popul...	NaN	NaN

Transposition des données

transformation de ‘data’ en ‘data_scoring’ en passant par un pivot table

```
data_scoring = data.loc[data["Country Type"] == "Country", :].pivot_table(index=["Country Name","Region"], columns="Indicator N
```

```
data_scoring.columns = data_scoring.columns.droplevel(0) #remove last_value
data_scoring.columns.name = None                         #remove Indicator Name
data_scoring = data_scoring.reset_index()                #index to columns
```

```
data_scoring.set_index("Country Name")
```

Country Name	Region	Barro-Lee: Percentage of population age 15+ with secondary schooling. Completed Secondary	Barro-Lee: Percentage of population age 15+ with tertiary schooling. Completed Tertiary	Barro-Lee: Population in thousands, age 15+, total	Enrolment in tertiary education per 100,000 inhabitants, both sexes	Enrolment in upper secondary education, both sexes (number)	Enrolment in upper secondary general, both sexes (number)	GNI per capita, PPP (current international $)	GNI, PPP (current international $)	Graduates from tertiary education, both sexes (number)	...	Population growth (annual %)	Popu of the offici age f post- secor non- tertia educa both sexes (num
Afghanistan	South Asia	8.65	3.65	19299.0	831.156250	968769.0	943750.0	1900.0	6.588225e+10	24315.0	...	2.689163	1302
Albania	Europe & Central	42.90	0.93	2431.0	6015.172852	151937.0	125256.0	11670.0	3.357241e+10	33529.0	...	-0.159880	112

Minoration de l'évolution de la population cible

1. nb bacheliers < nb lycéens + étudiants
2. taux utilisateurs internet = figé au taux actuel

Population cible actuelle = (pop 15+ * taux de bacheliers de 15+) * taux utilisateurs internet actuel :

```
data_scoring["Target Pop, current, in thousands"] = (data_scoring['Barro-Lee: Population in thousands, age 15+, total'] *\
data_scoring['Barro-Lee: Percentage of population age 15+ with secondary schooling. Completed Secondary']/ 100)*\
data_scoring['Internet users (per 100 people)'] / 100
```

Population cible 2050 = (pop totale de bacheliers) * taux utilisateurs internet actuel :

```
data_scoring['Target Pop, 2050, in thousands'] =\
data_scoring['Wittgenstein Projection: Population in thousands by highest level of educational attainment. Upper Secondary. Total
* data_scoring['Internet users (per 100 people)'] / 100
```

Evolution = ((Population cible 2050 / Population cible actuelle) -1) * 100 :

```
data_scoring['Target Pop evol %'] = round((data_scoring['Target Pop, 2050, in thousands'] / \
                                          data_scoring['Target Pop, current, in thousands']-1)*100,1)
```

Vérification du taux de manquants

```
print(f"Ce jeu de données contient {round(data_scoring.isna().mean().mean()*100,2)}% de données manquantes")
```

```
Ce jeu de données contient 17.99% de données manquantes
```

```
print(f"Cela représente {data_scoring.isna().sum().sum()} données manquantes")
```

```
Cela représente 1078 données manquantes
```

Nous allons identifier les pays pour lesquels les données sont les moins fournies et les supprimer.
Pour cela nous créons une colonne 'is_null' contenant la moyenne par ligne des cellules NaN :

```
data_scoring["is_null"] = data_scoring.isnull().mean(axis=1)
```

manquants = 18%

Suppression des 117 pays au Taux de manquants > 0.05

```
# taux_limite_nan = float(input("Quel est le taux maximum acceptable pour les manquants par pays? (entre 0.00 et 1.00)")) #0.05
taux_limite_nan = 0.05
mask = data_scoring["is_null"] > taux_limite_nan
```

```
print(f"""
Il existe {mask[mask].shape[0]} pays dont le taux de NaN est supérieur à {taux_limite_nan*100}%.""")
```

```
Il existe 117 pays dont le taux de NaN est supérieur à 5.0%.
```

On supprime les pays dont le taux est supérieur à la limite fixée (on conserve ceux hors de 'mask') :

```
data_scoring = data_scoring[~mask]
```

```
print(f"{len(data_scoring.index)} pays ont un taux de manquant compris dans les limites définies pour l'analyse")
```

```
97 pays ont un taux de manquant compris dans les limites définies pour l'analyse
```

```
print(f"Après suppression des pays non retenus, ce jeu de données contient {data_scoring.isna().mean().mean()*100:.2f}% de donnée
```

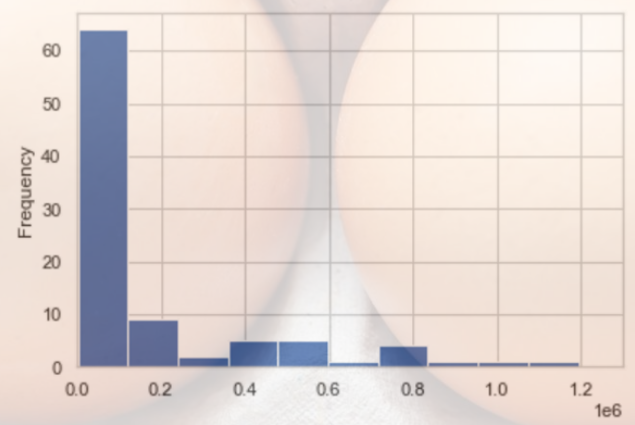
```
Après suppression des pays non retenus, ce jeu de données contient 1.17% de données manquantes
```

après suppression
manquants data_scoring = 1,17%

Imputation des données manquantes sur 5 indicateurs par la médiane :

```
data_scoring = data_scoring.fillna({"Graduates from tertiary education, both sexes (number)": \
                                    data_scoring["Graduates from tertiary education, both sexes (number)"].median()})
```

```
ax = data_scoring['Graduates from tertiary education, both sexes (number)'].plot.hist(bins=100, xlim=[0,1300000])
```



1 | Nettoyage & exploration

2 | Filtre et sélection des données

3 | Pertinence du jeu de données

4 | Scoring pays à fort potentiel clients

5 | Evolution du potentiel clients

6 | Indicateurs statistiques

Un jeu de données pertinent mais...

- tous les pays représentés
- données sur les dimensions clé
- données relativement récentes

- beaucoup de manquants
- indicateurs historiques ≠ prospectifs
- manque mesures de :
  - l'utilisation du web
  - la formation à distance
  - la taille et offre actuelles d'Academy

1 | Nettoyage & exploration

2 | Filtre et sélection des données

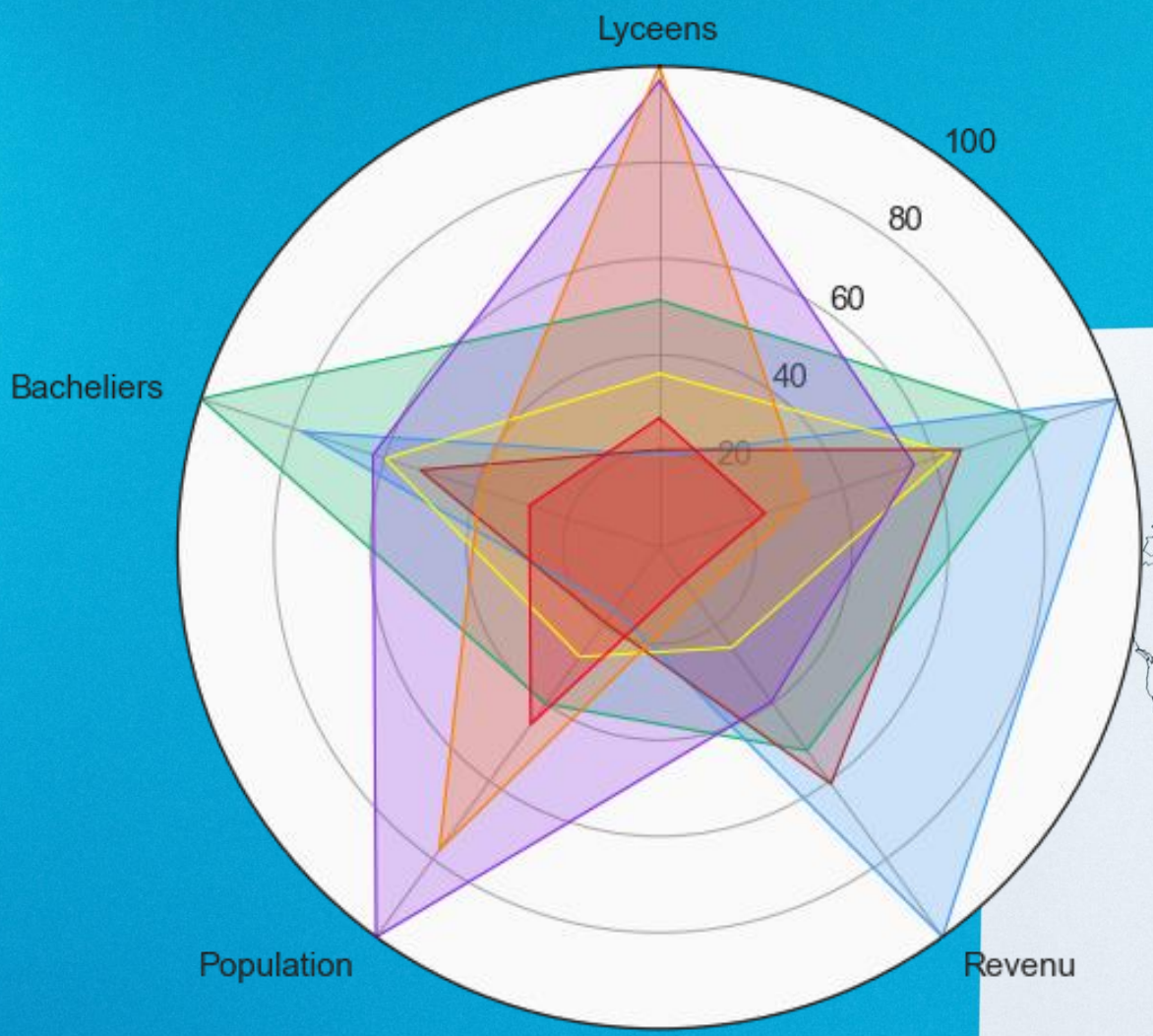
3 | Pertinence du jeu de données

4 | Scoring pays à fort potentiel clients

5 | Evolution du potentiel clients

6 | Indicateurs statistiques

Zones géographiques selon indicateurs clé



Le poids de chaque critère intervient fortement dans le classement final

Saisie des coefficients souhaités pour chaque indicateur

```
# On demande à l'utilisateur de saisir des coefficients
poids_pop, poids_edu, poids_web, poids_eco = \
input("Saisir le poids à donner aux indicateurs SCORE_POP/SCORE_EDU/SCORE_WEB/SCORE_ECO de 1 à 3 et séparés par un /" \
      +"\n"+"Par exemple : 1/3/3/3 :"+"\n").split("/",4)
```

```
Saisir le poids à donner aux indicateurs SCORE_POP/SCORE_EDU/SCORE_WEB/SCORE_ECO de 1 à 3 et séparés par un /
Par exemple : 1/3/3/3 :
1/3/3/3
```

Pop : 1 / Edu : 3 / Web : 3 / Eco : 3

Country Name	SCORE_SYNTH	SCORE_POP	SCORE_ECO	SCORE_EDU	SCORE_WEB
Qatar	0.684046	0.001622	1.000000	0.321658	0.957953
Switzerland	0.616699	0.005832	0.506477	0.641310	0.905934
Luxembourg	0.614008	0.000180	0.560946	0.493632	0.992056
United States	0.601799	0.234191	0.466570	0.696285	0.765077
Macao SAR, China	0.601388	0.000202	0.787996	0.393284	0.823280
Korea, Rep.	0.594312	0.036937	0.282243	0.745297	0.941188
Norway	0.592022	0.003554	0.497546	0.484703	0.989972
Hong Kong SAR, C...	0.578775	0.005088	0.482098	0.561939	0.883518
Sweden	0.575802	0.006942	0.396814	0.591903	0.928307
Denmark	0.574189	0.003916	0.405423	0.520782	0.986454
Japan	0.571502	0.091894	0.338563	0.602253	0.933558
United Kingdom	0.570796	0.047378	0.333012	0.590735	0.963114
Ireland	0.567248	0.003220	0.451846	0.609017	0.828889
Iceland	0.564389	0.000000	0.416606	0.464691	1.000000
Netherlands	0.555244	0.012105	0.402285	0.527855	0.916639

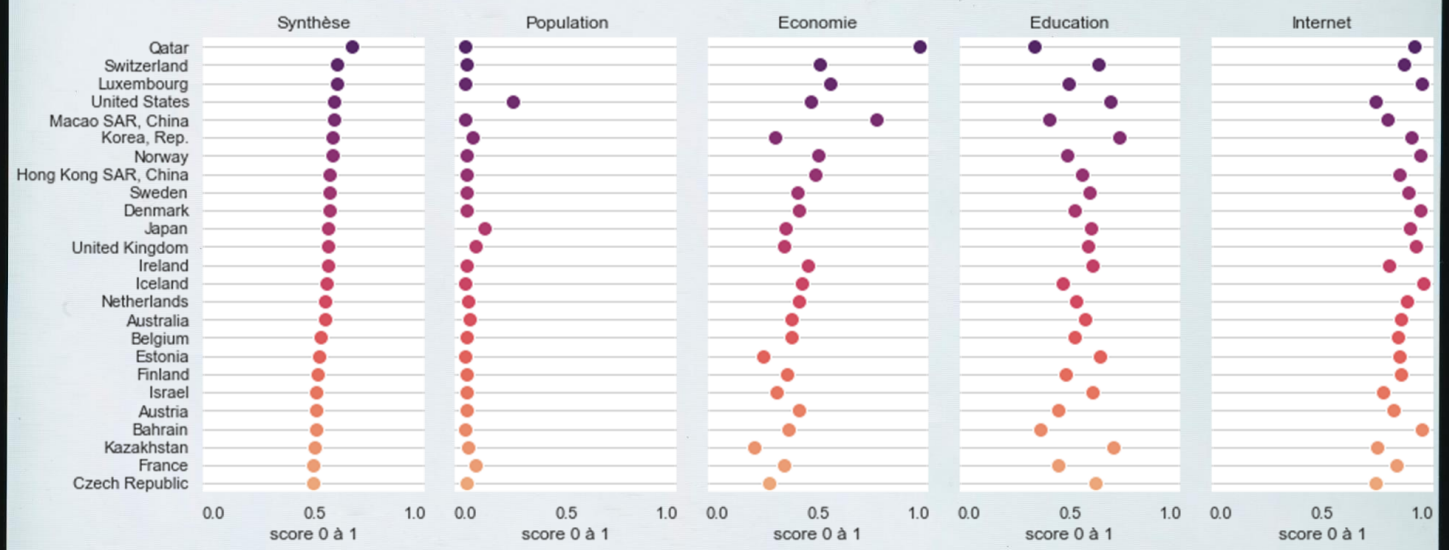
top5 ≠

hormis :
- Quatar
- United States

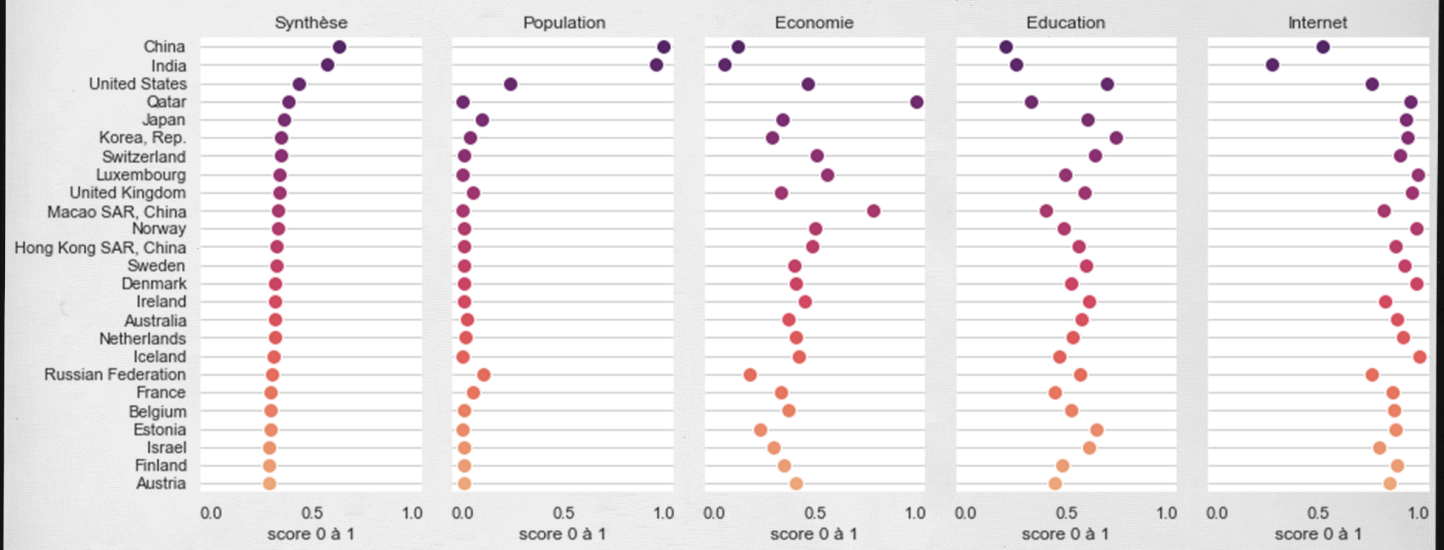
Pop : 3 / Edu : 1 / Web : 1 / Eco : 1

Country Name	SCORE_SYNTH	SCORE_POP	SCORE_ECO	SCORE_EDU	SCORE_WEB
China	0.638901	1.000000	0.118996	0.193982	0.520429
India	0.573805	0.960464	0.046585	0.246271	0.268581
United States	0.438418	0.234191	0.466570	0.696285	0.765077
Qatar	0.380746	0.001622	1.000000	0.321658	0.957953
Japan	0.358343	0.091894	0.338563	0.602253	0.933558
Korea, Rep.	0.346590	0.036937	0.282243	0.745297	0.941188
Switzerland	0.345203	0.005832	0.506477	0.641310	0.905934
Luxembourg	0.341196	0.000180	0.560946	0.493632	0.992056
United Kingdom	0.338166	0.047378	0.333012	0.590735	0.963114
Macao SAR, China	0.334194	0.000202	0.787996	0.393284	0.823280
Norway	0.330481	0.003554	0.497546	0.484703	0.989972
Hong Kong SAR, C...	0.323803	0.005088	0.482098	0.561939	0.883518
Sweden	0.322975	0.006942	0.396814	0.591903	0.928307
Denmark	0.320734	0.003916	0.405423	0.520782	0.986454
Ireland	0.316569	0.003220	0.451846	0.609017	0.828889

Scoring top 25 | Pop : 1 / Edu : 3 / Web : 3 / Eco : 3



Scoring top 25 | Pop : 3 / Edu : 1 / Web : 1 / Eco : 1



1 | Nettoyage & exploration

2 | Filtre et sélection des données

3 | Pertinence du jeu de données

4 | Scoring pays à fort potentiel clients

5 | Evolution du potentiel clients

6 | Indicateurs statistiques

Evolution du potentiel client | Pop : 1 / Edu : 3 / Web : 3 / Eco : 3

Country Name	SCORE_SYNTH	SCORE_POP	SCORE_ECO	SCORE_EDU	SCORE_WEB	Region	Target Pop, current, in thousands	Target Pop, 2050, in thousands	Target Pop evol %
Qatar	0.684046	0.001622	1.000000	0.321658	0.957953	Middle East & No...	154.237451	988.161489	540.7
Switzerland	0.616699	0.005832	0.506477	0.641310	0.905934	Europe & Central...	2680.730474	3775.382567	40.8
Luxembourg	0.614008	0.000180	0.560946	0.493632	0.992056	Europe & Central...	108.630857	266.460661	145.3
United States	0.601799	0.234191	0.466570	0.696285	0.765077	North America	68827.170355	116778.564524	69.7
Macao SAR, China	0.601388	0.000202	0.787996	0.393284	0.823280	East Asia & Pacific	109.056741	188.799404	73.1
Korea, Rep.	0.594312	0.036937	0.282243	0.745297	0.941188	East Asia & Pacific	13436.289646	11809.918313	-12.1
Norway	0.592022	0.003554	0.497546	0.484703	0.989972	Europe & Central...	1513.159285	1927.156331	27.4
Hong Kong SAR, China	0.578775	0.005088	0.482098	0.561939	0.883518	East Asia & Pacific	2454.924588	2472.502861	0.7
Sweden	0.575802	0.006942	0.396814	0.591903	0.928307	Europe & Central...	3395.897550	3668.938827	8.0
Denmark	0.574189	0.003916	0.405423	0.520782	0.986454	Europe & Central...	1702.203532	2641.596406	55.2
Japan	0.571502	0.091894	0.338563	0.602253	0.933558	East Asia & Pacific	41772.241752	29606.234800	-29.1
United Kingdom	0.570796	0.047378	0.333012	0.590735	0.963114	Europe & Central...	22542.969709	12030.072448	-46.6
Ireland	0.567248	0.003220	0.451846	0.609017	0.828889	Europe & Central...	715.344264	1013.045580	41.6
Iceland	0.564389	0.000000	0.416606	0.464691	1.000000	Europe & Central...	60.182129	121.179060	101.4
Netherlands	0.555244	0.012105	0.402285	0.527855	0.916639	Europe & Central...	4762.737024	6281.310411	31.9

Evolution du potentiel client | Pop : 3 / Edu : 1 / Web : 1 / Eco : 1

Country Name	SCORE_SYNTH	SCORE_POP	SCORE_ECO	SCORE_EDU	SCORE_WEB	Region	Target Pop, current, in thousands	Target Pop, 2050, in thousands	Target Pop evol %
China	0.638901	1.000000	0.118996	0.193982	0.520429	East Asia & Pacific	132586.822466	159397.962360	20.2
India	0.573805	0.960464	0.046585	0.246271	0.268581	South Asia	61315.525625	166534.310424	171.6
United States	0.438418	0.234191	0.466570	0.696285	0.765077	North America	68827.170355	116778.564524	69.7
Qatar	0.380746	0.001622	1.000000	0.321658	0.957953	Middle East & No...	154.237451	988.161489	540.7
Japan	0.358343	0.091894	0.338563	0.602253	0.933558	East Asia & Pacific	41772.241752	29606.234800	-29.1
Korea, Rep.	0.346590	0.036937	0.282243	0.745297	0.941188	East Asia & Pacific	13436.289646	11809.918313	-12.1
Switzerland	0.345203	0.005832	0.506477	0.641310	0.905934	Europe & Central...	2680.730474	3775.382567	40.8
Luxembourg	0.341196	0.000180	0.560946	0.493632	0.992056	Europe & Central...	108.630857	266.460661	145.3
United Kingdom	0.338166	0.047378	0.333012	0.590735	0.963114	Europe & Central...	22542.969709	12030.072448	-46.6
Macao SAR, China	0.334194	0.000202	0.787996	0.393284	0.823280	East Asia & Pacific	109.056741	188.799404	73.1
Norway	0.330481	0.003554	0.497546	0.484703	0.989972	Europe & Central...	1513.159285	1927.156331	27.4
Hong Kong SAR, China	0.323803	0.005088	0.482098	0.561939	0.883518	East Asia & Pacific	2454.924588	2472.502861	0.7
Sweden	0.322975	0.006942	0.396814	0.591903	0.928307	Europe & Central...	3395.897550	3668.938827	8.0
Denmark	0.320734	0.003916	0.405423	0.520782	0.986454	Europe & Central...	1702.203532	2641.596406	55.2
Ireland	0.316569	0.003220	0.451846	0.609017	0.828889	Europe & Central...	715.344264	1013.045580	41.6

1 | Nettoyage & exploration

2 | Filtre et sélection des données

3 | Pertinence du jeu de données

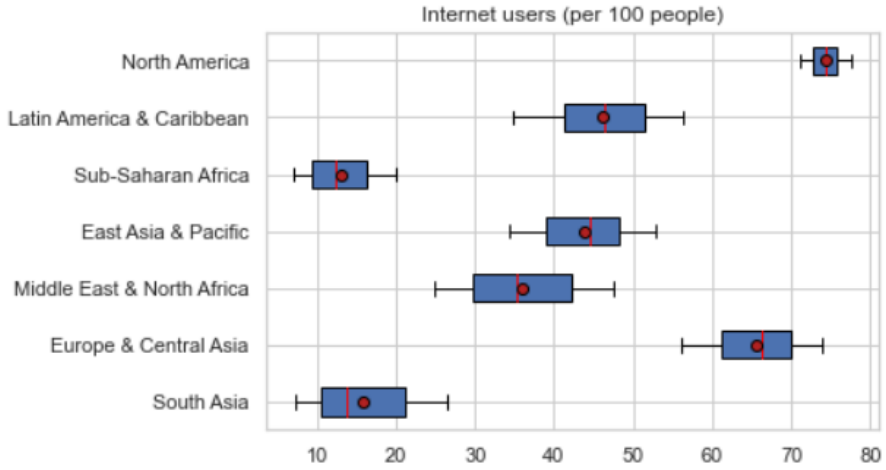
4 | Scoring pays à fort potentiel clients

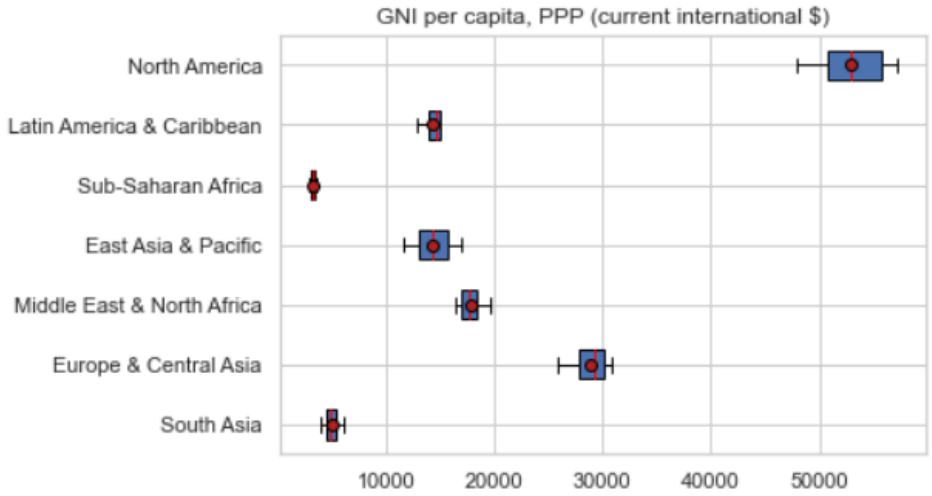
5 | Evolution du potentiel clients

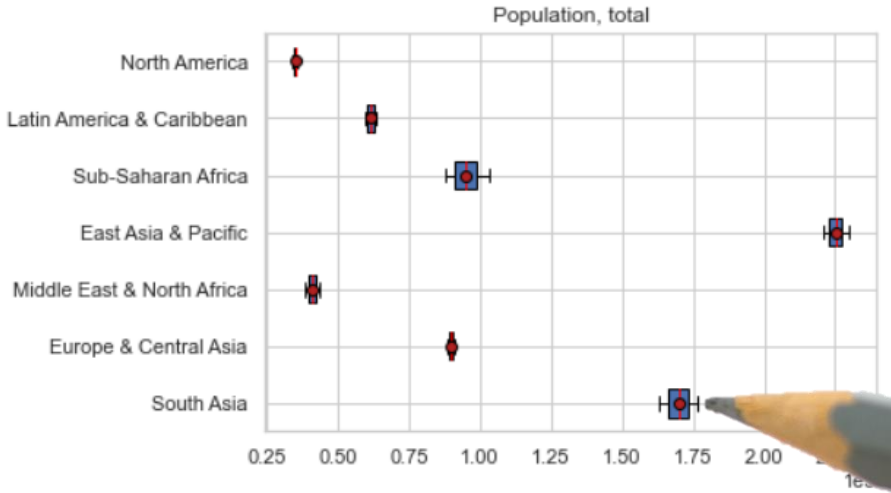
6 | Indicateurs statistiques

Statistiques par région

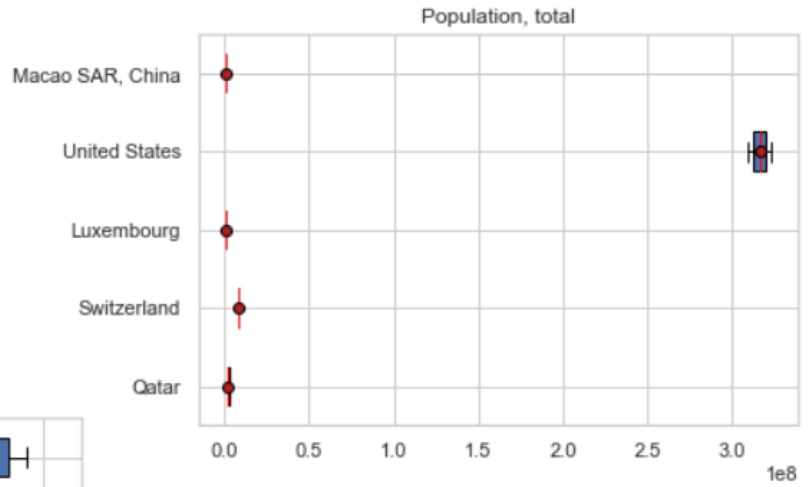
données d'éducation = non disponibles à l'échelle des régions

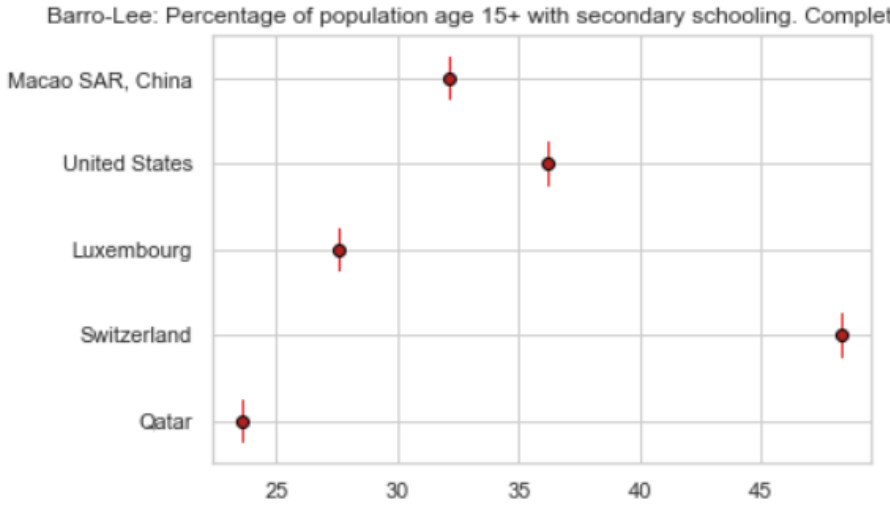


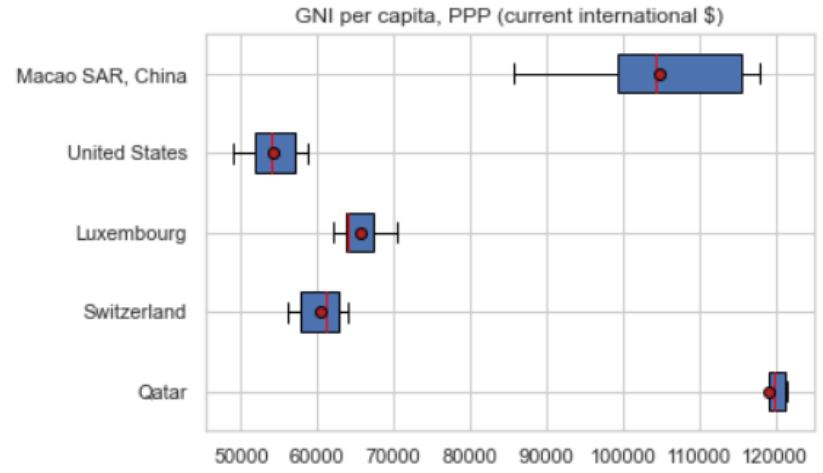


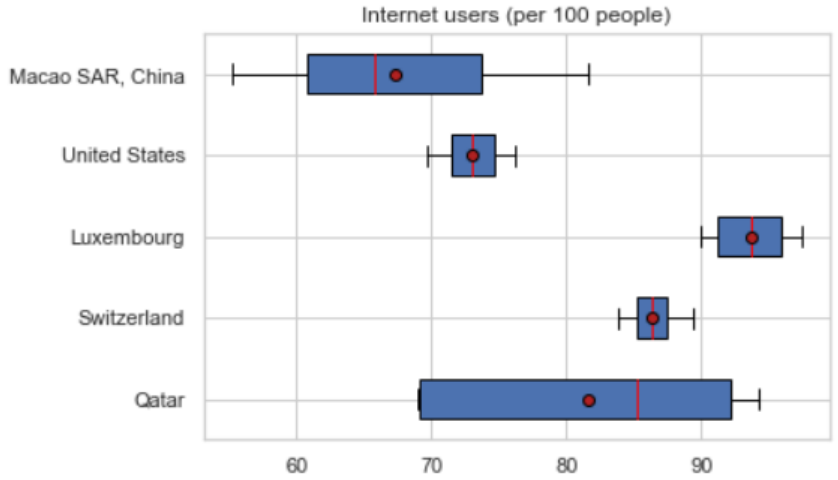


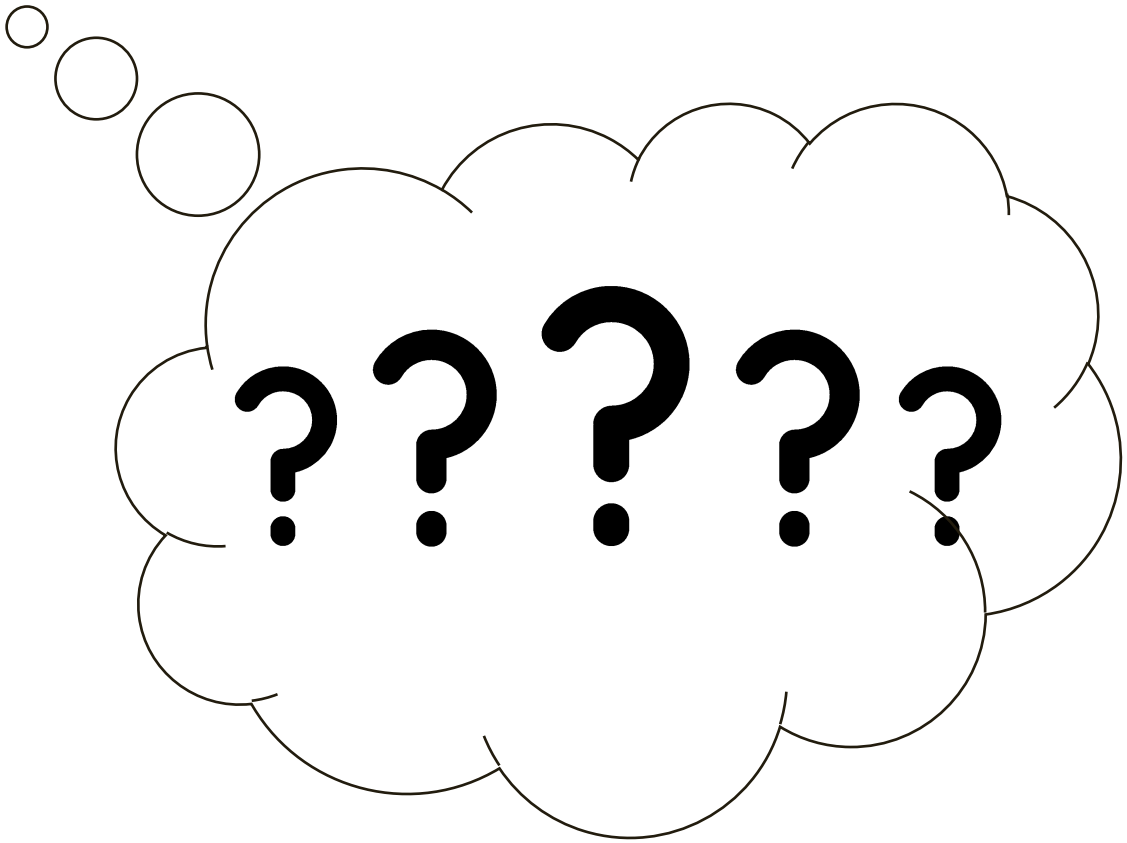
Statistiques par pays du top 5












academy
MERCI !