



Projet 2 OC :

Analysez des données de systèmes éducatifs




- **Start-up de la EdTech, nommée Academy**
- **Contenus de formation en ligne pour un public de niveau lycée et université**
- **Son but est de s'implanter à l'international**



Analysez des données de systèmes éducatifs

Mark d'Academy me demande par mail à partir des données de la Banque mondiale **“EdStats All Indicator Query”**
<https://datacatalog.worldbank.org/dataset/education-statistics>

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




Analyse pré-exploratoire

- Comporte-t-il beaucoup de données manquantes, dupliquées ?
- Nombre de colonnes ? nombre de lignes ?
- Sélectionner les informations qui semblent pertinentes pour répondre à la problématique
- 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)

Analyse pré-exploratoire

5 fichiers CSV



Nom/Nbr	Ligne	Colonne	Intérêt
EdStatsData	886930	70	indicateurs, pays, années, données numériques continues et qualitatives nominales, bcp de données pertinentes
EdStatsCountry	241	32	indicateurs, pays, notes, sources, peu de données pertinentes
EdStatsCountry-Series	613	4	pays, notes, sources, peu de données peu pertinentes
EdStatsFootNote	643638	5	pays, notes, sources, bcp de données, peu pertinentes
EdStatsSeries	3665	21	indicateurs, notes, sources, peu de données, peu pertinentes

Données dupliquées sur la colonne « Indicateurs » et informations

Data columns (total 70 columns):

#	Column	Non-Null Count	Dtype
0	Country Name	886930 non-null	object
1	Country Code	886930 non-null	object
2	Indicator Name	886930 non-null	object
3	Indicator Code	886930 non-null	object
4	1970	72288 non-null	float64
5	1971	35537 non-null	float64
6	1972	35619 non-null	float64
7	1973	35545 non-null	float64
8	1974	35730 non-null	float64
9	1975	87306 non-null	float64
10	1976	37483 non-null	float64
11	1977	37574 non-null	float64
12	1978	37576 non-null	float64
13	1979	36809 non-null	float64
14	1980	89122 non-null	float64
15	1981	38777 non-null	float64
16	1982	37511 non-null	float64
17	1983	38460 non-null	float64
18	1984	38606 non-null	float64
19	1985	90296 non-null	float64
20	1986	39372 non-null	float64

```
Entrée [68]: df1=pd.DataFrame(data1, columns=['Country Name'])
              df2=df1.groupby(by="Country Name")
              len(df2)
```

Out[68]: 242

```
Entrée [42]: len(data1['Country Name'].unique())
```

Out[42]: 242

```
Entrée [43]: len(data1['Country Name'])
```

Out[43]: 886930

```
Entrée [69]: df1=pd.DataFrame(data1, columns=['Indicator Name'])
              df2=df1.groupby(by="Indicator Name")
              len(df2)
```

Out[69]: 3665

```
Entrée [45]: len(data1['Indicator Name'].unique())
```

Out[45]: 3665

```
Entrée [78]: data1['Indicator Name'].nunique(dropna=True)
```

Out[78]: 3665

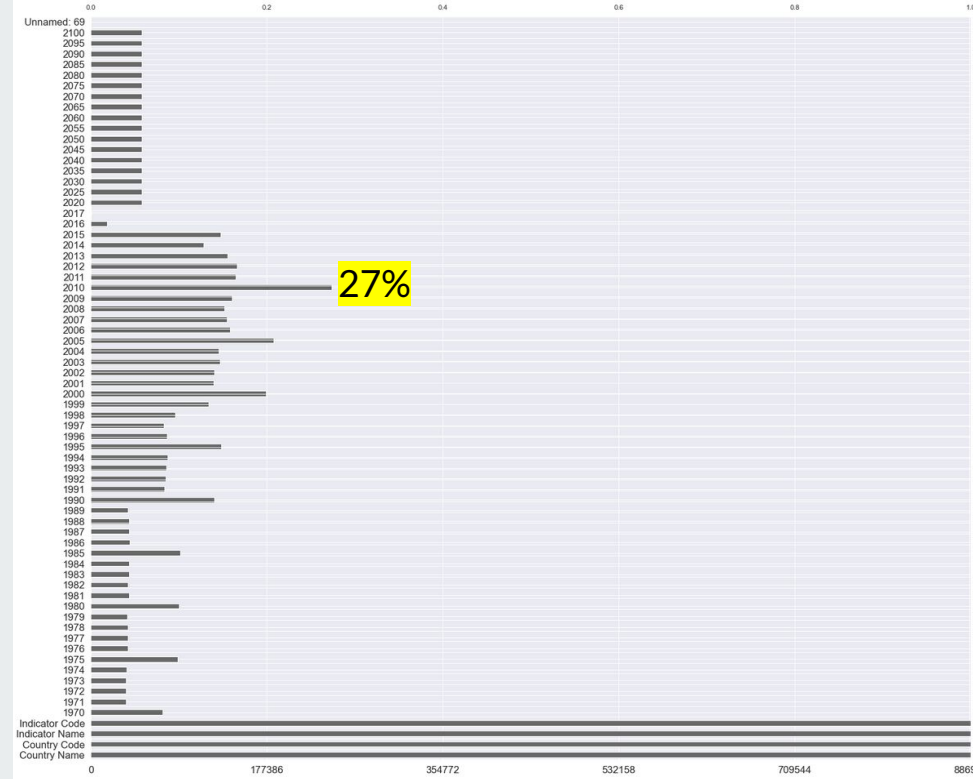
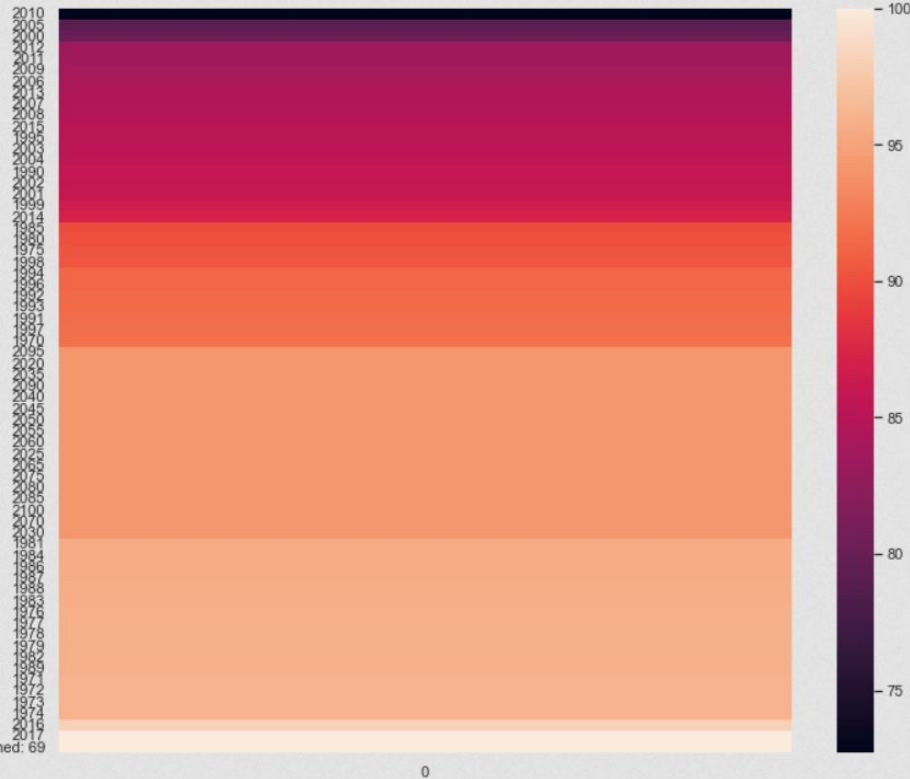
```
Entrée [87]: len(data1['Indicator Name'])
```

Out[87]: 886930

```
dups = data1.pivot_table(index = ['Country Name','Indicator Name'], aggfunc = 'size')
print (dups[dups>1])
```

ries([], dtype: int64)

Données manquantes sur les colonnes « Années »



Données manquantes sur la colonne « Indicateurs »



	2010
Indicator Name	
Population, total	93.5
Population growth (annual %)	93.5
Internet users (per 100 people)	93.8
GDP per capita (current US\$)	93.8
GDP per capita (constant 2005 US\$)	93.8
GDP at market prices (current US\$)	93.8
GDP at market prices (constant 2005 US\$)	93.8
Total outbound internationally mobile tertiary students studying abroad, all countries, both sexes (number)	93.9
Population of the official age for secondary education, both sexes (number)	94.0
Population of the official age for upper secondary education, both sexes (number)	94.0
Population of the official age for lower secondary education, both sexes (number)	94.0
Population of the official age for upper secondary education, female (number)	94.0
Population of the official age for upper secondary education, male (number)	94.0
Population, male (% of total)	94.0
Population, male	94.0
Population, female (% of total)	94.0
Population, ages 15-64, total	94.0
Population, ages 0-14, total	94.0
Population, ages 0-14, male	94.0
Population, female	94.0
Population, ages 0-14 (% of total)	94.0
Population, ages 15-64 (% of total)	94.0
Population, ages 0-14, female	94.0
Population, ages 15-64, male	94.0
Population, ages 15-64, female	94.0
Population of the official entrance age to secondary general education, both sexes (number)	94.1
Population of the official entrance age to secondary general education, female (number)	94.1
Population of the official entrance age to secondary general education, male (number)	94.1
Population of the official age for tertiary education, male (number)	94.1
Population of the official age for tertiary education, female (number)	94.1
Population of the official age for tertiary education, both sexes (number)	94.1
Population of the official age for secondary education, male (number)	94.1
Population of the official age for secondary education, female (number)	94.1

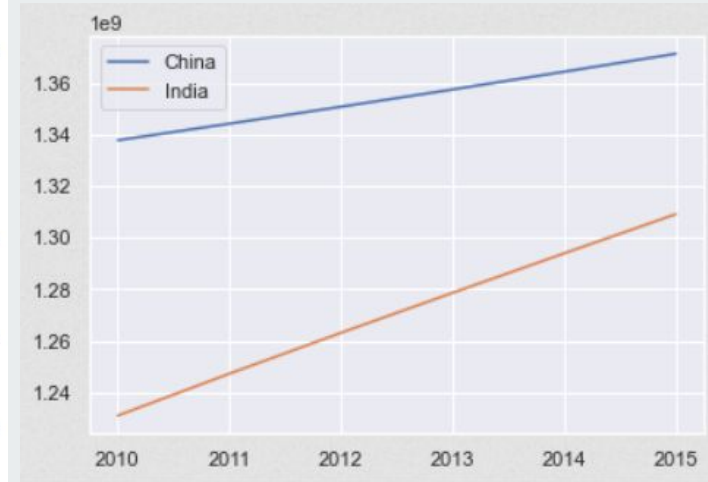
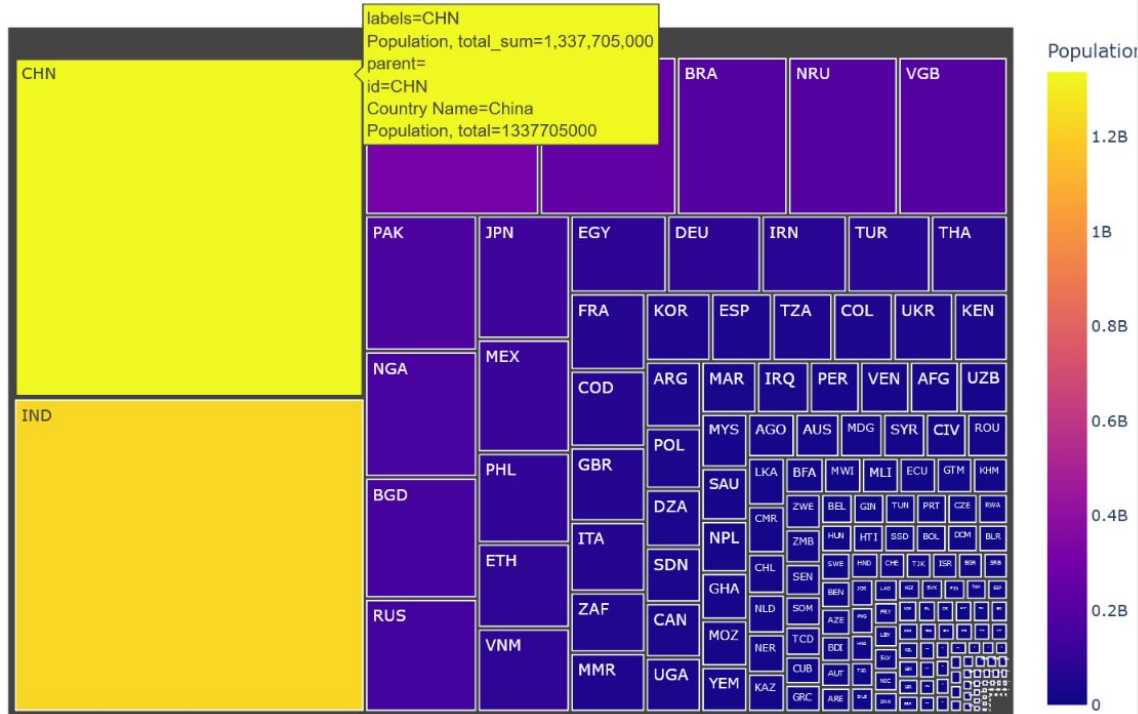
Statistiques sur les indicateurs sélectionnés



	Population, total	Population growth (annual %)	Internet users (per 100 people)	GDP per capita (current US\$)	Population of the official age for secondary education, both sexes (number)	School life expectancy, primary, both sexes (years)	Gross outbound enrolment ratio, all regions, both sexes (%)	Population, ages 0-14 (% of total)	Population of the official age for tertiary education, both sexes (number)	Mortality rate, under-5 (per 1,000)	Population, ages 13-19, total
count	2.160000e+02	216.000000	216.000000	216.000000	2.160000e+02	216.000000	216.000000	216.000000	2.160000e+02	216.000000	2.160000e+02
mean	3.366866e+07	1.426698	34.401353	15386.023056	5.970456e+06	5.989904	3.703767	28.926304	4.995565e+06	38.767147	4.701802e+06
std	1.288818e+08	1.558615	26.411632	22449.416483	1.565971e+07	1.041000	6.793553	10.273195	1.312045e+07	37.858081	1.558149e+07
min	1.053100e+04	-3.333512	0.250000	231.194326	1.310000e+03	2.461840	0.076720	11.935041	1.041000e+03	2.600000	1.006000e+04
25%	8.249985e+05	0.478990	10.000000	1619.189500	2.296595e+05	5.654540	0.534450	20.109418	1.591042e+05	10.375000	4.015512e+05
50%	6.166883e+06	1.264250	33.000000	6298.405565	9.037990e+05	5.973710	2.120460	28.808183	8.238175e+05	24.850000	1.534742e+06
75%	2.105204e+07	2.254904	53.075000	16621.446637	4.235514e+06	6.505418	3.379972	37.007866	3.312874e+06	54.075000	4.693412e+06
max	1.337705e+09	11.220686	93.390000	144246.368775	1.701458e+08	8.773300	55.501732	49.963344	1.296427e+08	208.000000	1.714163e+08

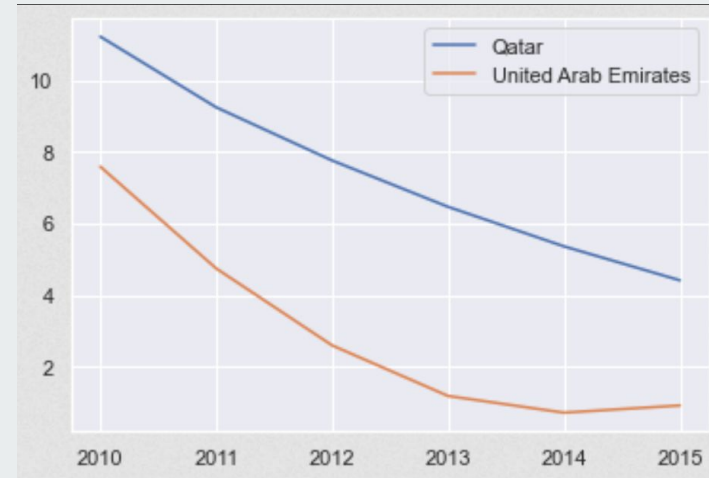
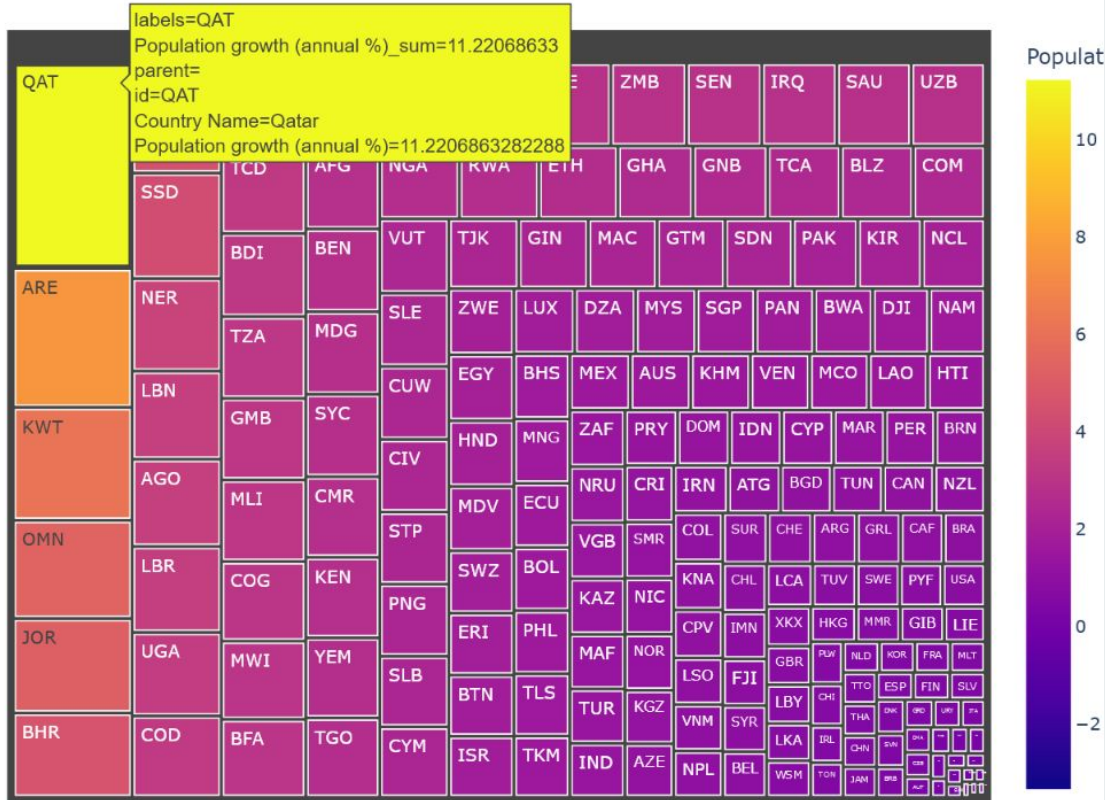
Heatmap et tendance sur 6 ans

World Population Distribution



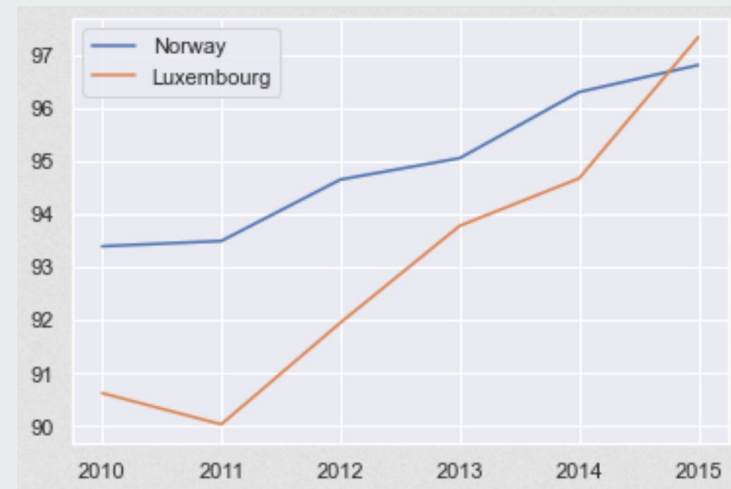
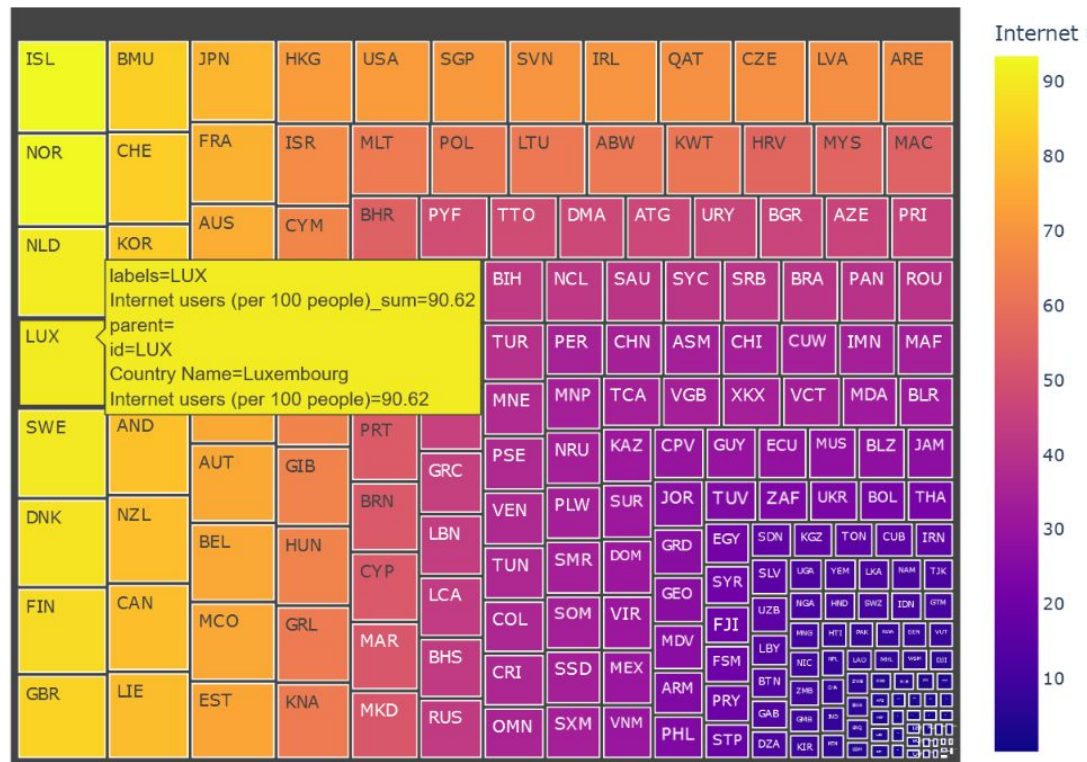
Heatmap et tendance sur 6 ans

Population growth (annual %)



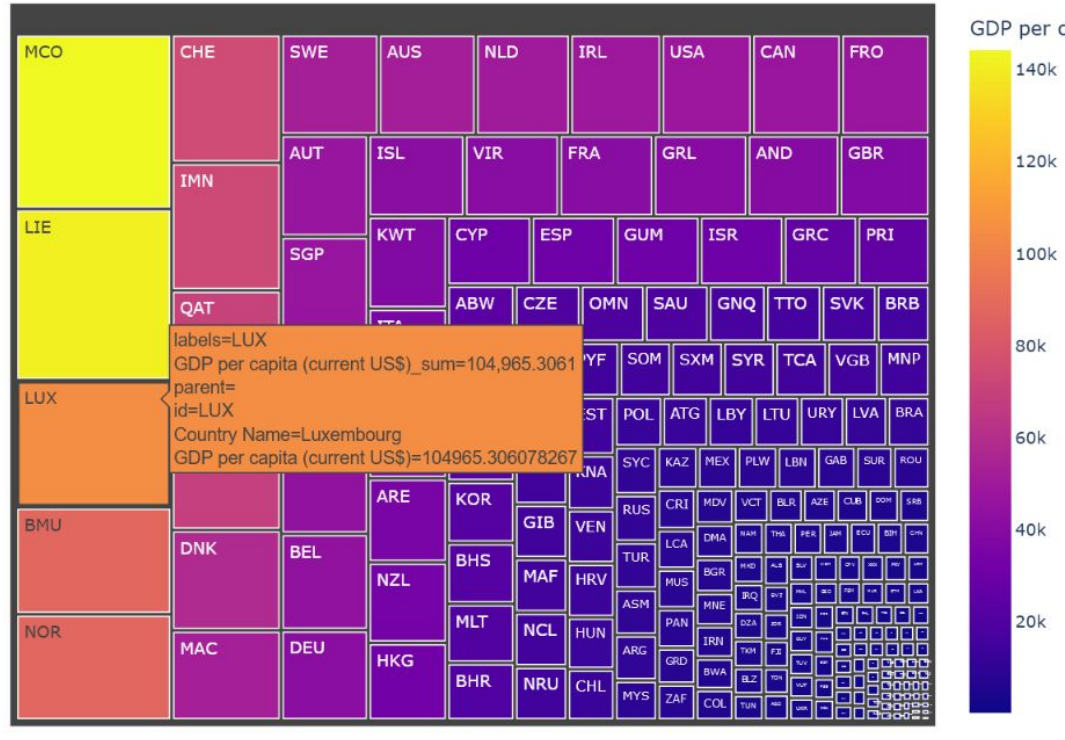
Heatmap et tendance sur 6 ans

Internet users (per 100 people)



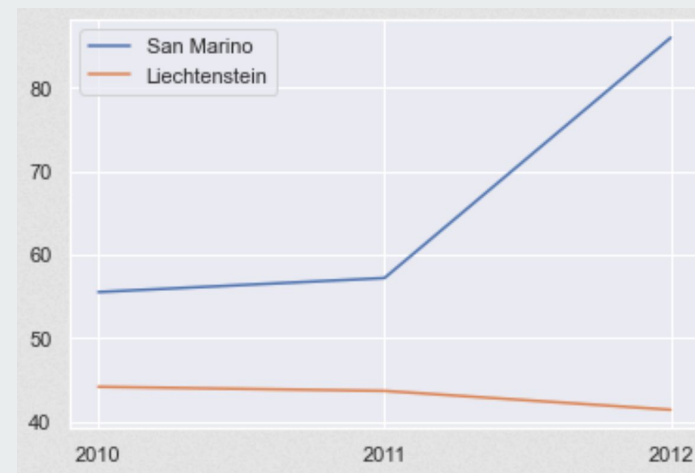
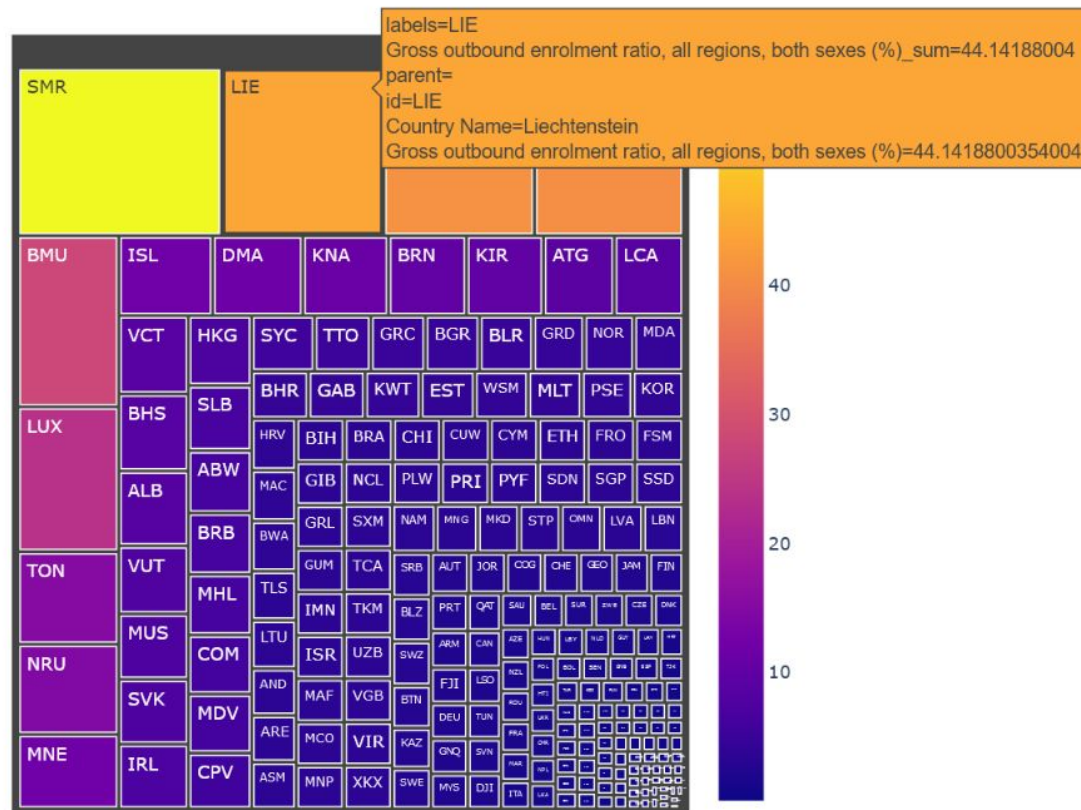
Heatmap et tendance sur 6 ans

GDP per capita (current US\$)



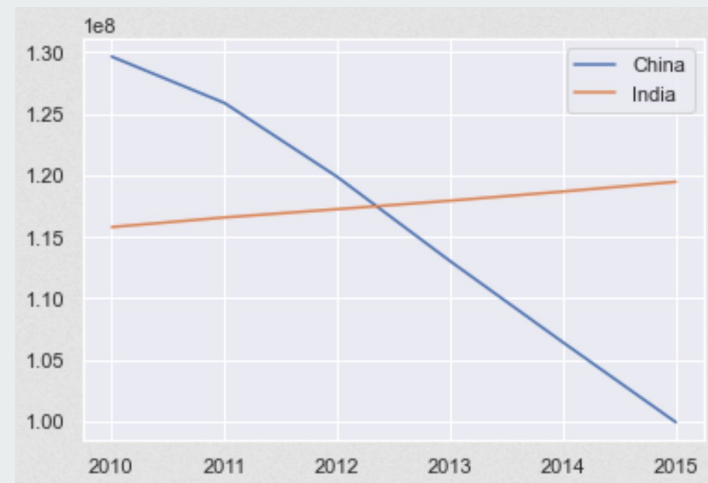
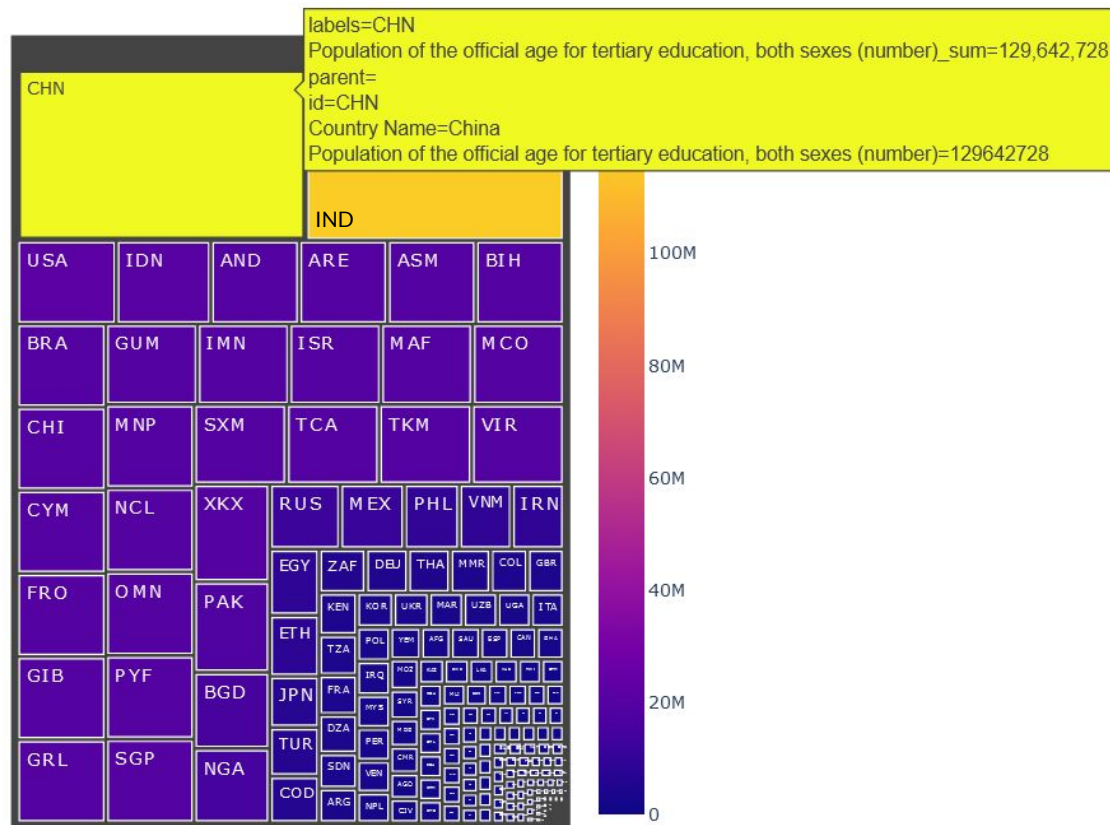
Heatmap et tendance sur 6 ans

Gross outbound enrolment ratio, all regions, both sexes (%)



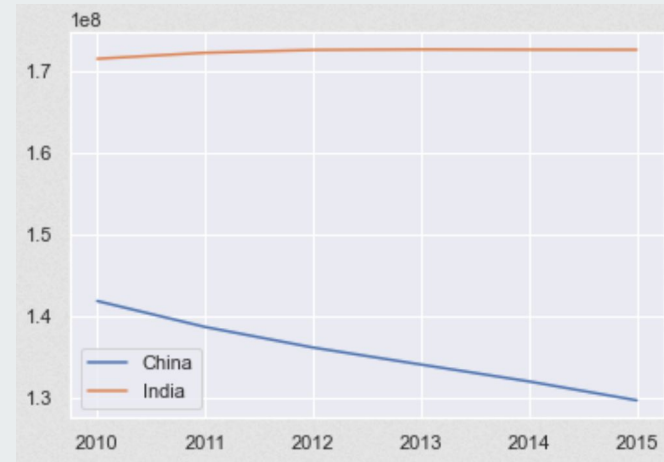
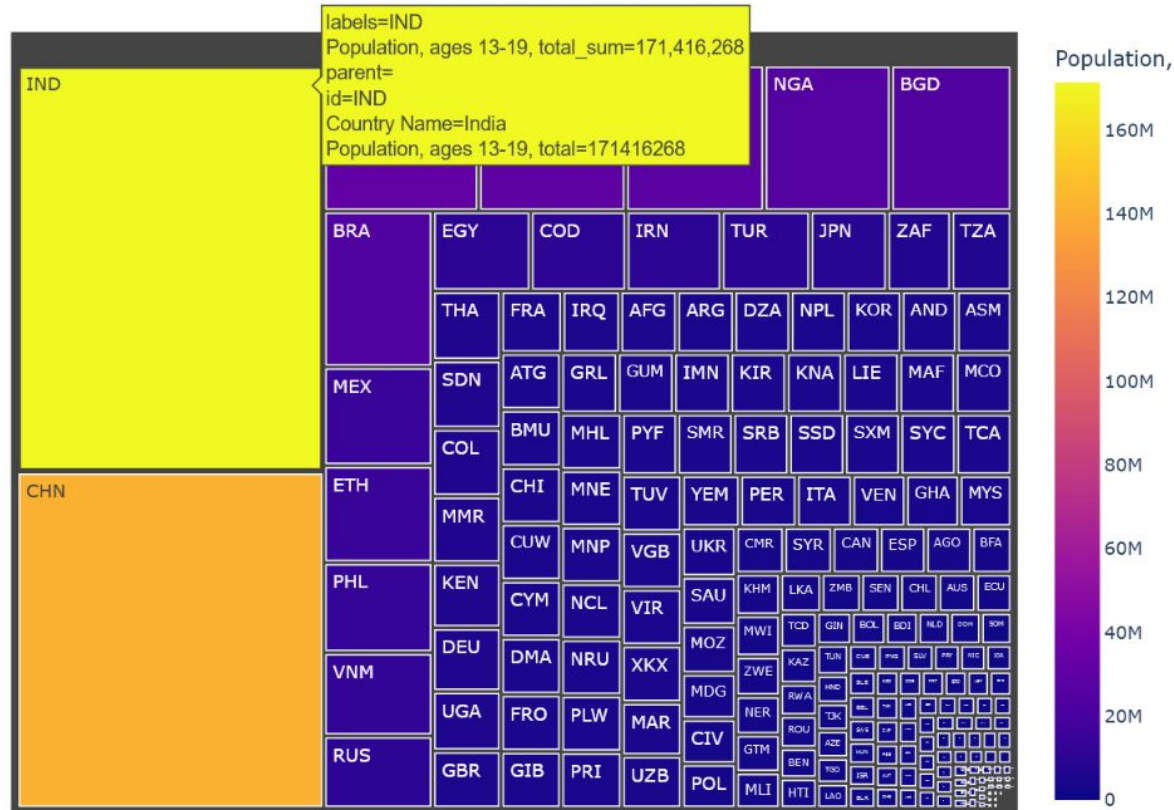
Heatmap et tendance sur 6 ans

Population of the official age for tertiary education, both sexes (number)



Heatmap et tendance sur 6 ans

Population, ages 13-19, total





Projet 2 OC :

Analysez des données de systèmes éducatifs

Questions / Réponses