OPENCLASSROOMS

Projet 2 : Analyse des données de systèmes éducatifs.

Ilham NOUMIR

Contexte:

une start-up de la EdTech, nommée academy;

Expansion à l’international

Mission:

- 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 ?

Objectifs de l’étude:

- 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

Plan du travail:

1.Vérification de la qualité des données :

- Description globale des données;
- Données manquantes et données dupliquées sur l'ensemble de la dataframe;
- Données manquantes sur les colones des dataframes;
- Elimination Des colonnes non remplies;
- Analyse Statistique globale des Dataframes avec la fonction describe()

2. Première visualisation des données: (Compréhension de l'objet des dataframes)

- Nombre de pays par région;
- Répartition des pays par unité monétaire;
- Nombre d'indicateurs par sujet (Topic);
- Nuage des mots;

3.Filtrage et regroupement des données :

- Filtrage de l'intervalle de l'étude : Quels sont les années qu'on peut utiliser pour notre étude;
- Filtrage des pays d'étude;
- Filtrage en fonction des indicateurs;
- Sélection minimale des indicateurs a analyser

4. Visualisation des indicateurs choisis:

5. Analyse bivariée

6. Classement des pays par potentiel d'implémentation via un scoring

7. Evaluation du potentiel des pays choisis;

Méthodologie du travail:

1.
Compréhension des données.

2.
Visualisation

3.
Filtrage des données

4.
Choix des indicateurs

5.
Scoring des pays

1. Compréhension et vérification de la qualité des données

5 jeu de données :

```
--------------------
country
La dimension de la dataframe country : (241, 32)
--------------------
indicateurs
La dimension de la dataframe indicateurs : (3665, 21)
--------------------
data
La dimension de la dataframe data : (886930, 70)
--------------------
source
La dimension de la dataframe source : (613, 4)
--------------------
description
La dimension de la dataframe description : (643638, 5)
```

Durée de 1970 à 2050

241 pays

3665 Indicateurs

1. Compréhension et vérification de la qualité des données

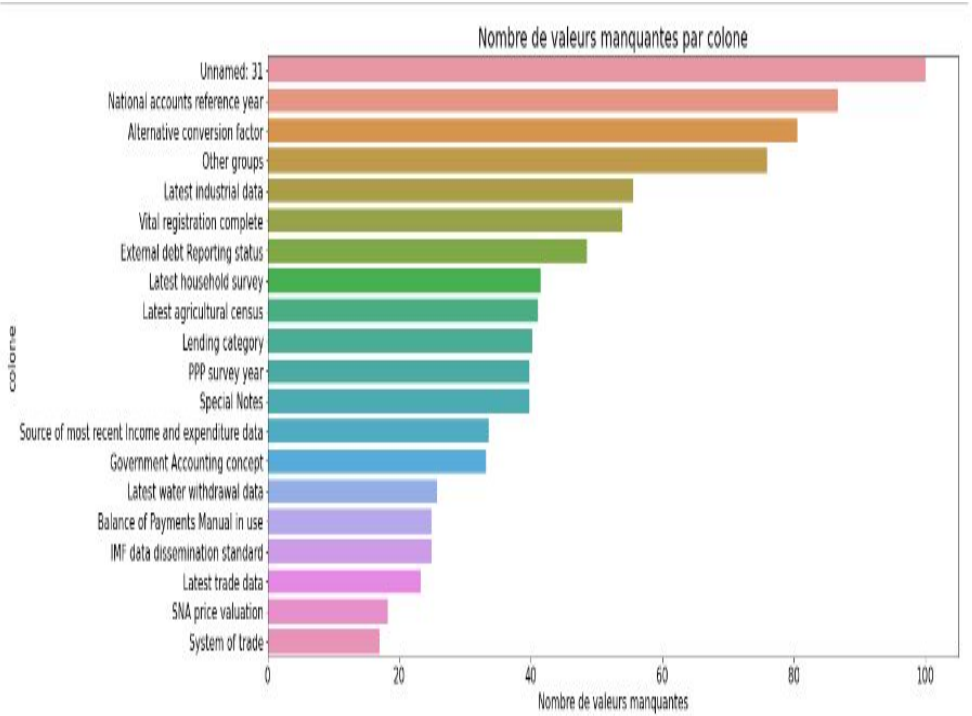
Valeurs manquantes et valeurs dupliquées

```
--------------------
country
Le percentage des NAN dans la totalité de la Dataframe country : 30.52
La somme des valeurs dupliquées dans la Dataframe country : 0
--------------------
indicateurs
Le percentage des NAN dans la totalité de la Dataframe indicateurs : 71.72
La somme des valeurs dupliquées dans la Dataframe indicateurs : 0
--------------------
data
Le percentage des NAN dans la totalité de la Dataframe data : 86.1
La somme des valeurs dupliquées dans la Dataframe data : 0
--------------------
source
Le percentage des NAN dans la totalité de la Dataframe source : 25.0
La somme des valeurs dupliquées dans la Dataframe source : 0
--------------------
description
Le percentage des NAN dans la totalité de la Dataframe description : 20.0
La somme des valeurs dupliquées dans la Dataframe description : 0
```

1. Compréhension et vérification de la qualité des données

Les valeurs manquantes pour chaque colonne de la Dataframe: Country

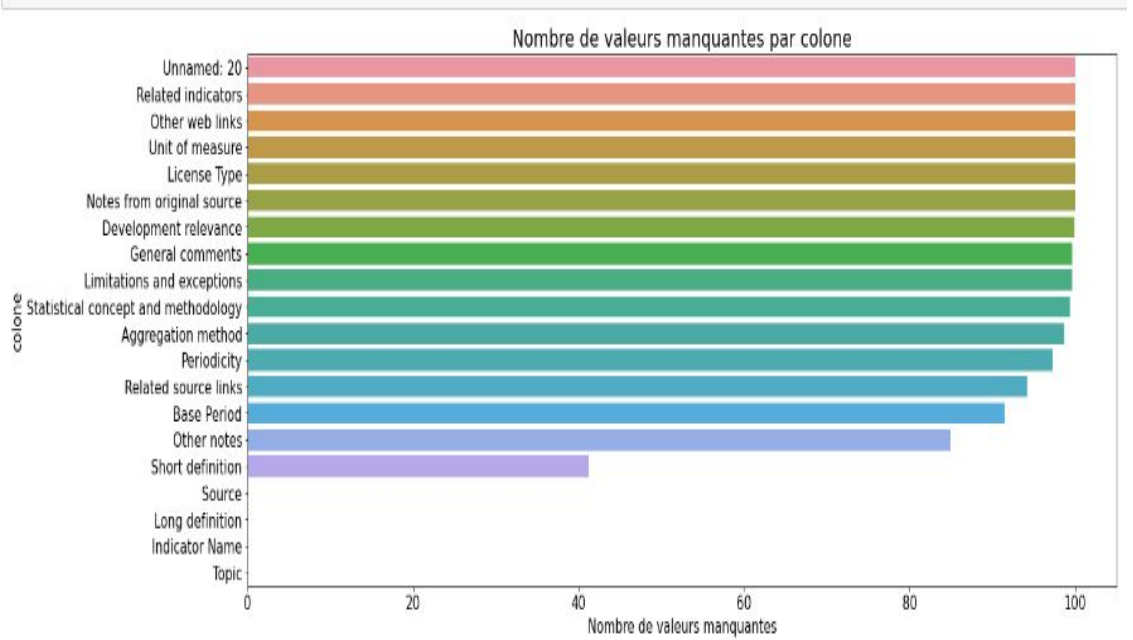
	% of Total Values
Unnamed: 31	100.000000
National accounts reference year	86.721992
Alternative conversion factor	80.497925
Other groups	75.933610
Latest industrial data	55.601660
Vital registration complete	53.941909
External debt Reporting status	48.547718
Latest household survey	41.493776
Latest agricultural census	41.078838
Lending category	40.248963
PPP survey year	39.834025
Special Notes	39.834025
Source of most recent Income and expenditure data	33.609959
Government Accounting concept	33.195021
Latest water withdrawal data	25.726141
Balance of Payments Manual in use	24.896266
IMF data dissemination standard	24.896266
Latest trade data	23.236515
SNA price valuation	18.257261



1. Compréhension et vérification de la qualité des données

Les valeurs manquantes pour chaque colonne de la Dataframe: Indicateurs

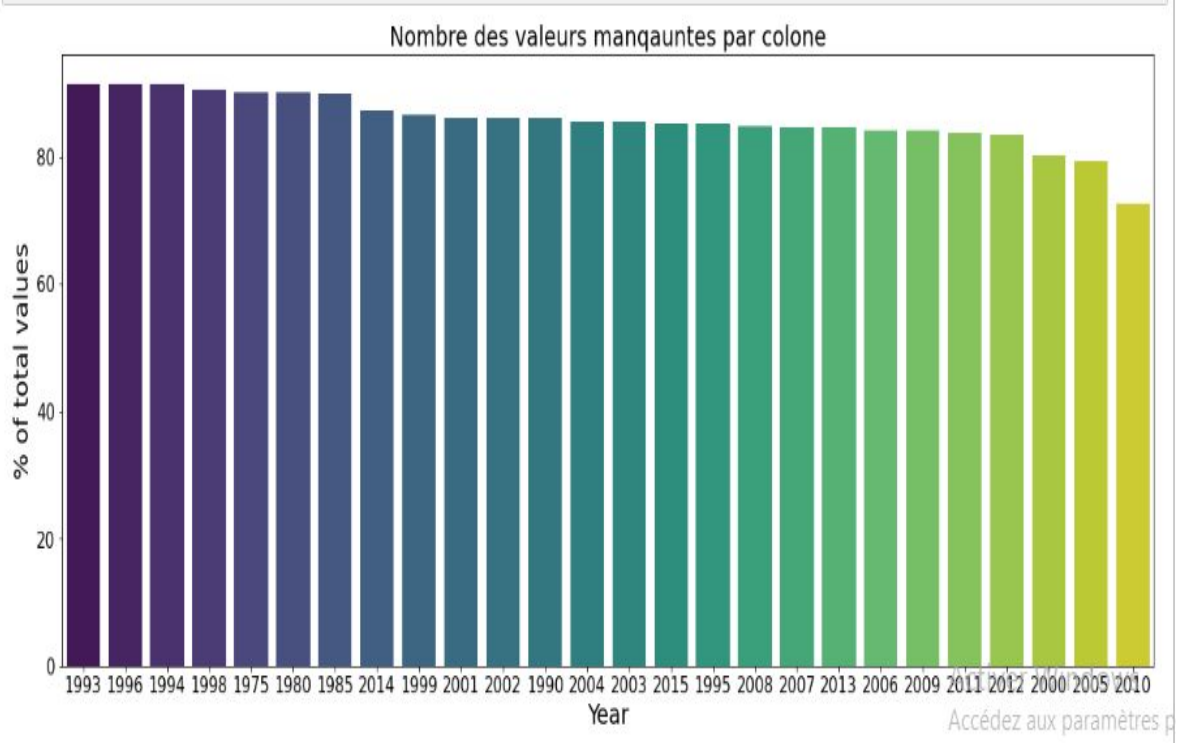
	% of Total Values
Unnamed: 20	100.000000
Related indicators	100.000000
Other web links	100.000000
Unit of measure	100.000000
License Type	100.000000
Notes from original source	100.000000
Development relevance	99.918145
General comments	99.618008
Limitations and exceptions	99.618008
Statistical concept and methodology	99.372442
Aggregation method	98.717599
Periodicity	97.298772
Related source links	94.133697
Base Period	91.432469
Other notes	84.938608
Short definition	41.173261
Source	0.000000
Long definition	0.000000
Indicator Name	0.000000
Topic	0.000000



1. Compréhension et vérification de la qualité des données

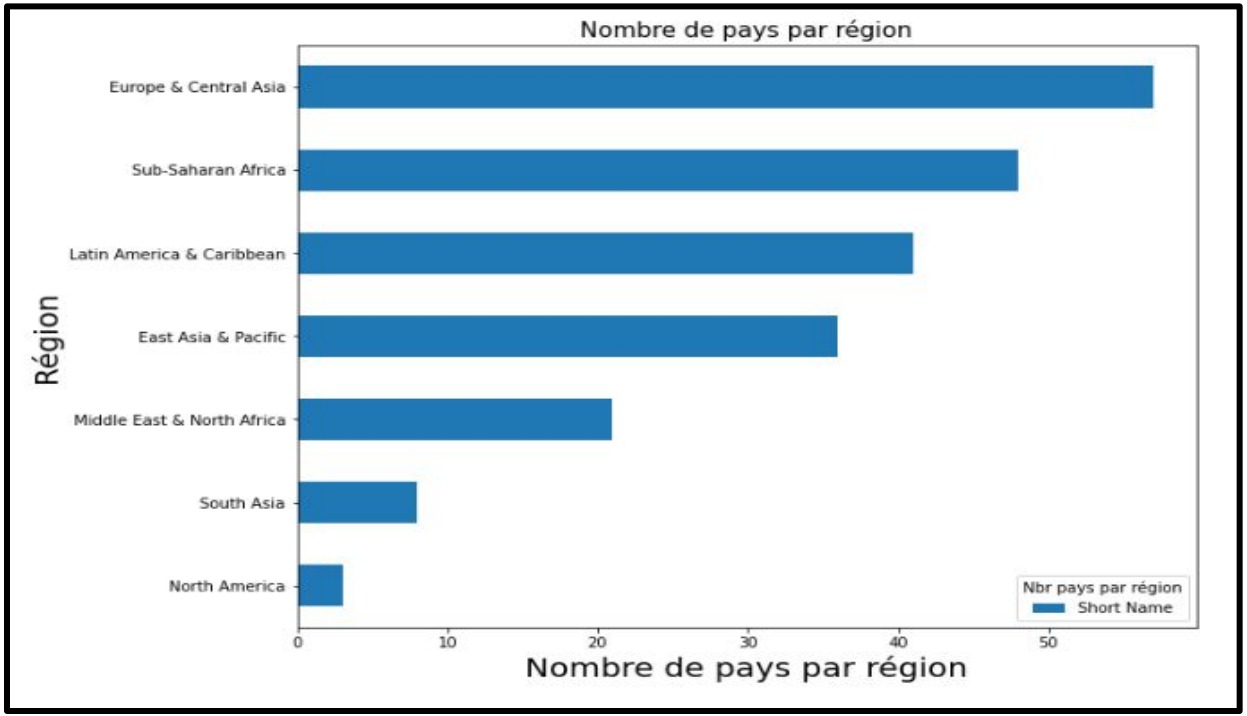
Les valeurs manquantes pour chaque colonne de la Dataframe: Data

	name_of_colone	% of Total Values
0	1993	91.454455
1	1996	91.340128
2	1994	91.266278
3	1998	90.426076
4	1975	90.156382
5	1980	89.951631
6	1985	89.819264
7	2014	87.170464
8	1999	86.601085
9	2001	86.074549
10	2002	85.996076
11	1990	85.973527
12	2004	85.476419
13	2003	85.301771
14	2015	85.223411
15	1995	85.189248
16	2008	84.848071
17	2007	84.522792
18	2013	84.496071
19	2006	84.180037
20	2009	83.977541
21	2011	83.537370
22	2012	83.396209
23	2000	80.080051



2. Visualisation des données

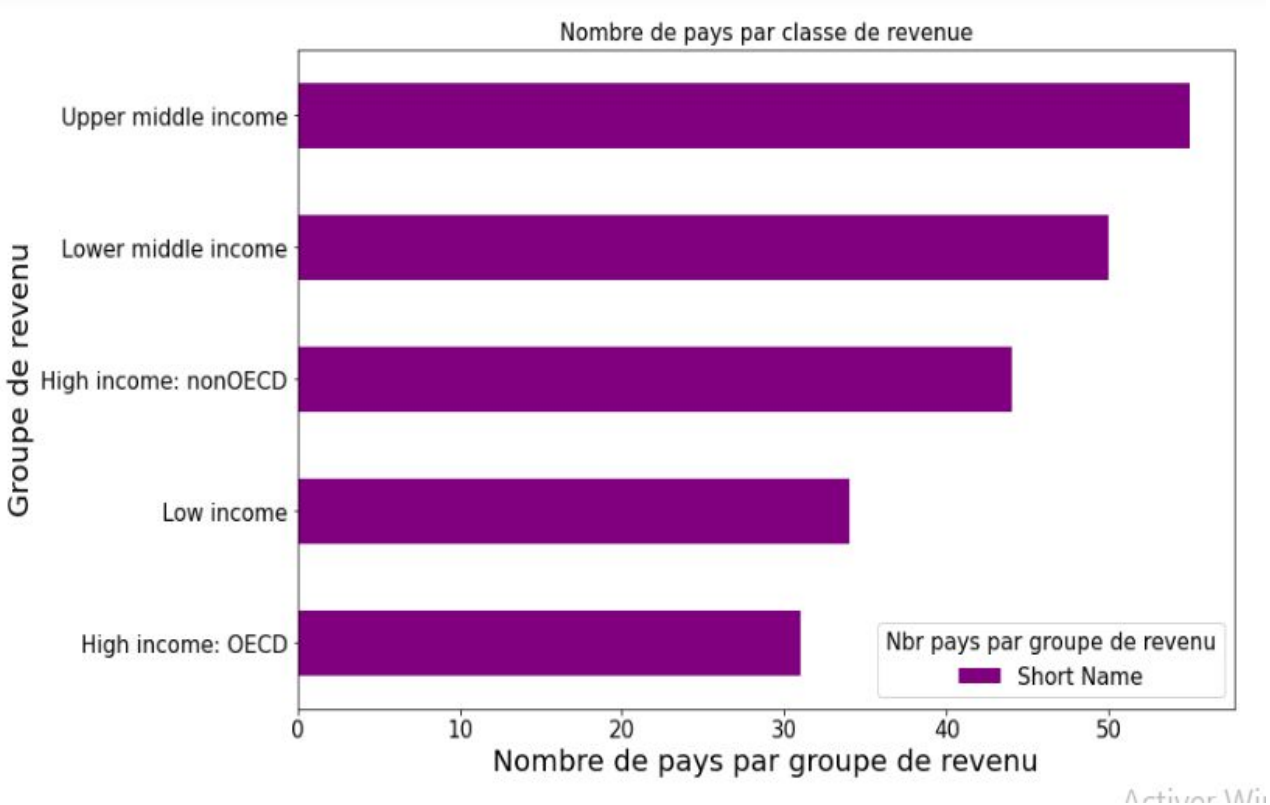
Les pays par région



On remarque que la région de l'Europe et le centre d'Asie contient plus de pays en matière de nombre

2. Visualisation des données

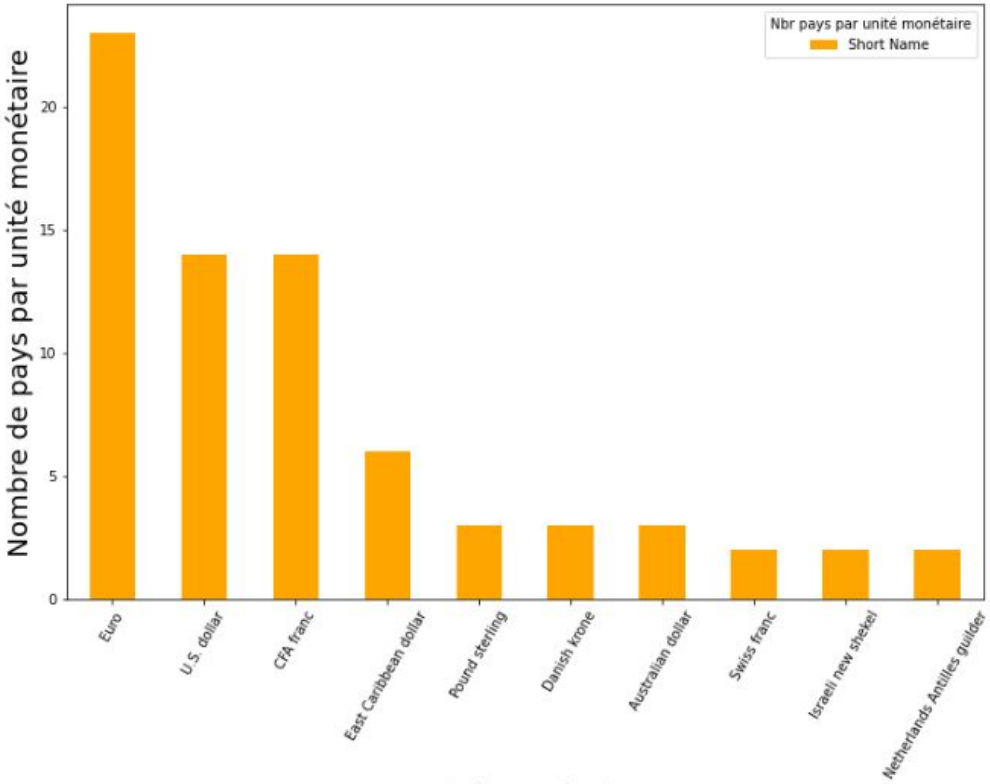
Les pays par classe de revenue



Beaucoup de pays ont des revenus moyens supérieurs

2. Visualisation des données

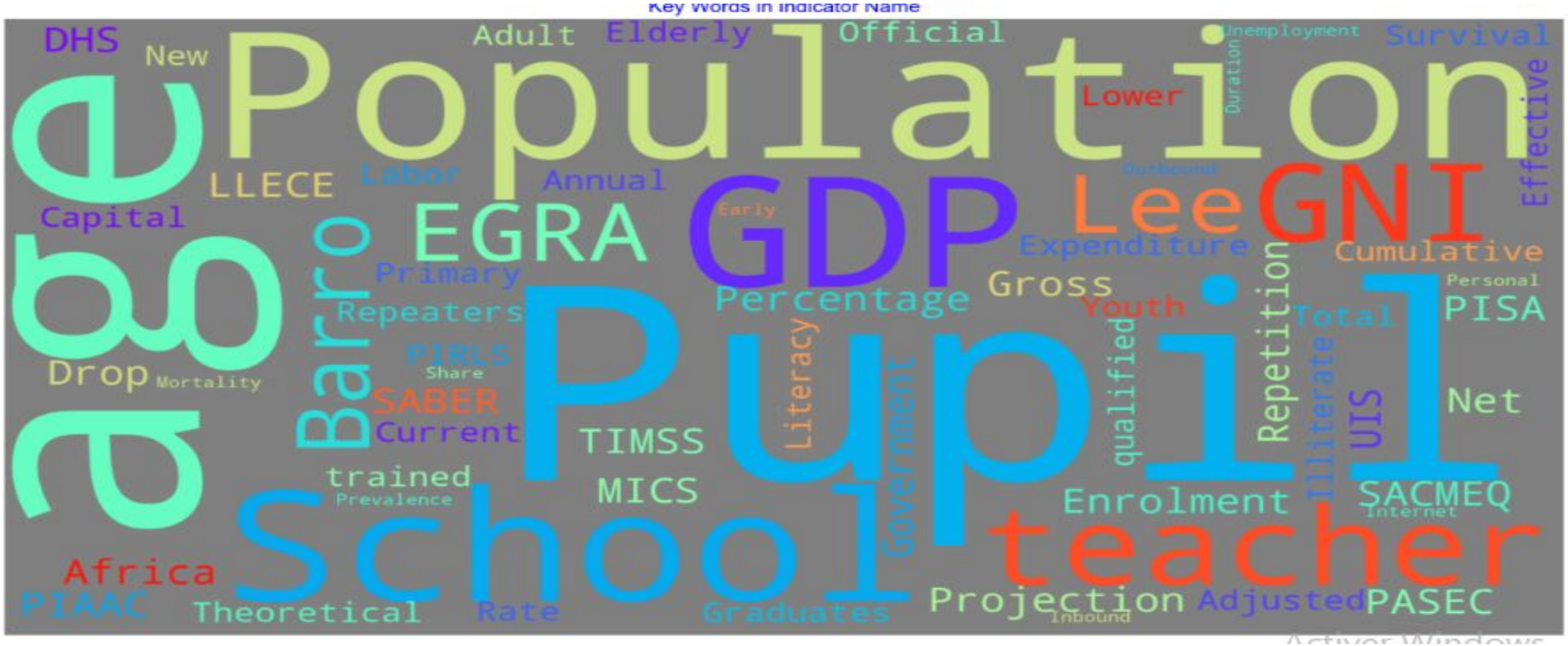
L’unité monétaire des pays



Nombreux sont les pays qui ont l’unité monétaire de l”Euro

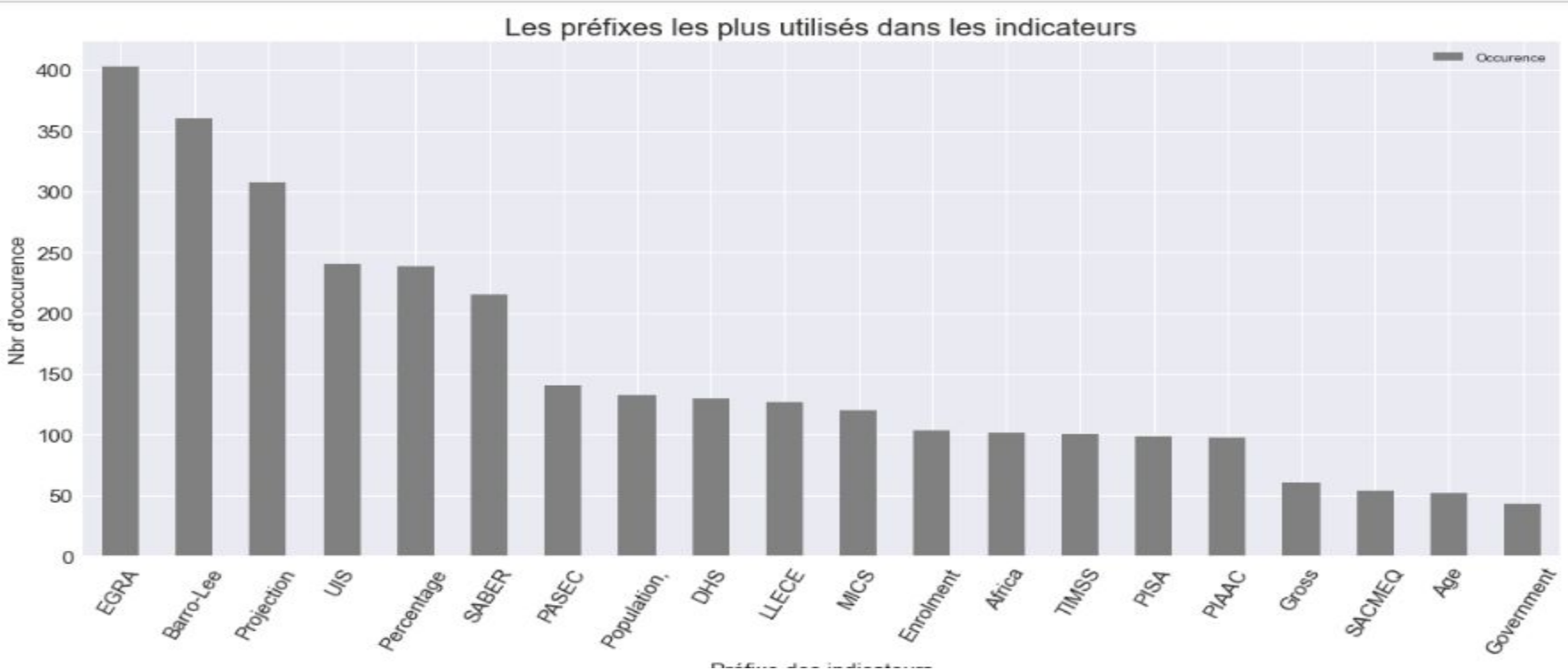
2. Visualisation des données

Nuage des points des mots les plus récurrents dans les indicateurs:



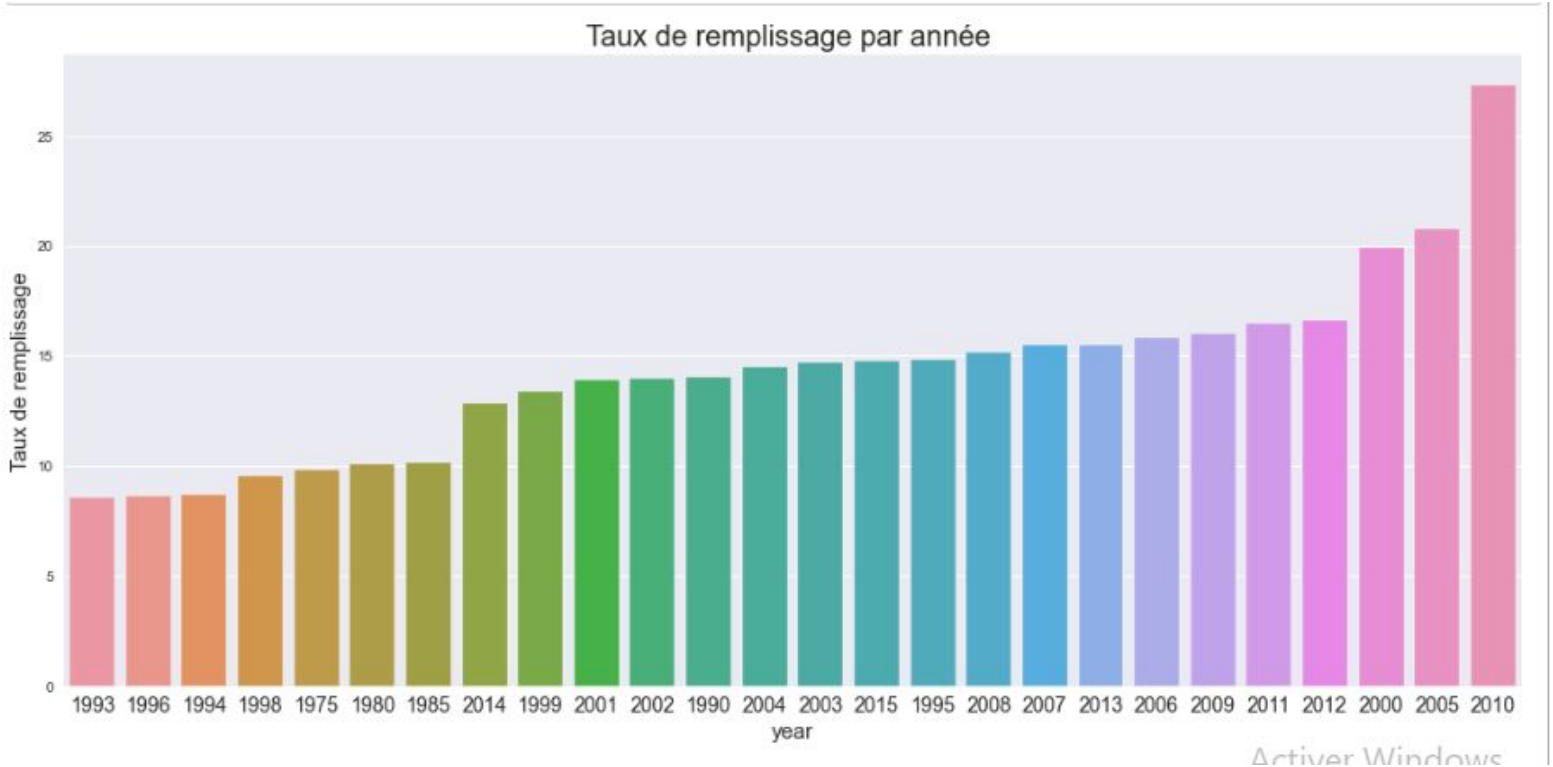
2. Visualisation des données

mots clés récurrents dans les indicateurs :



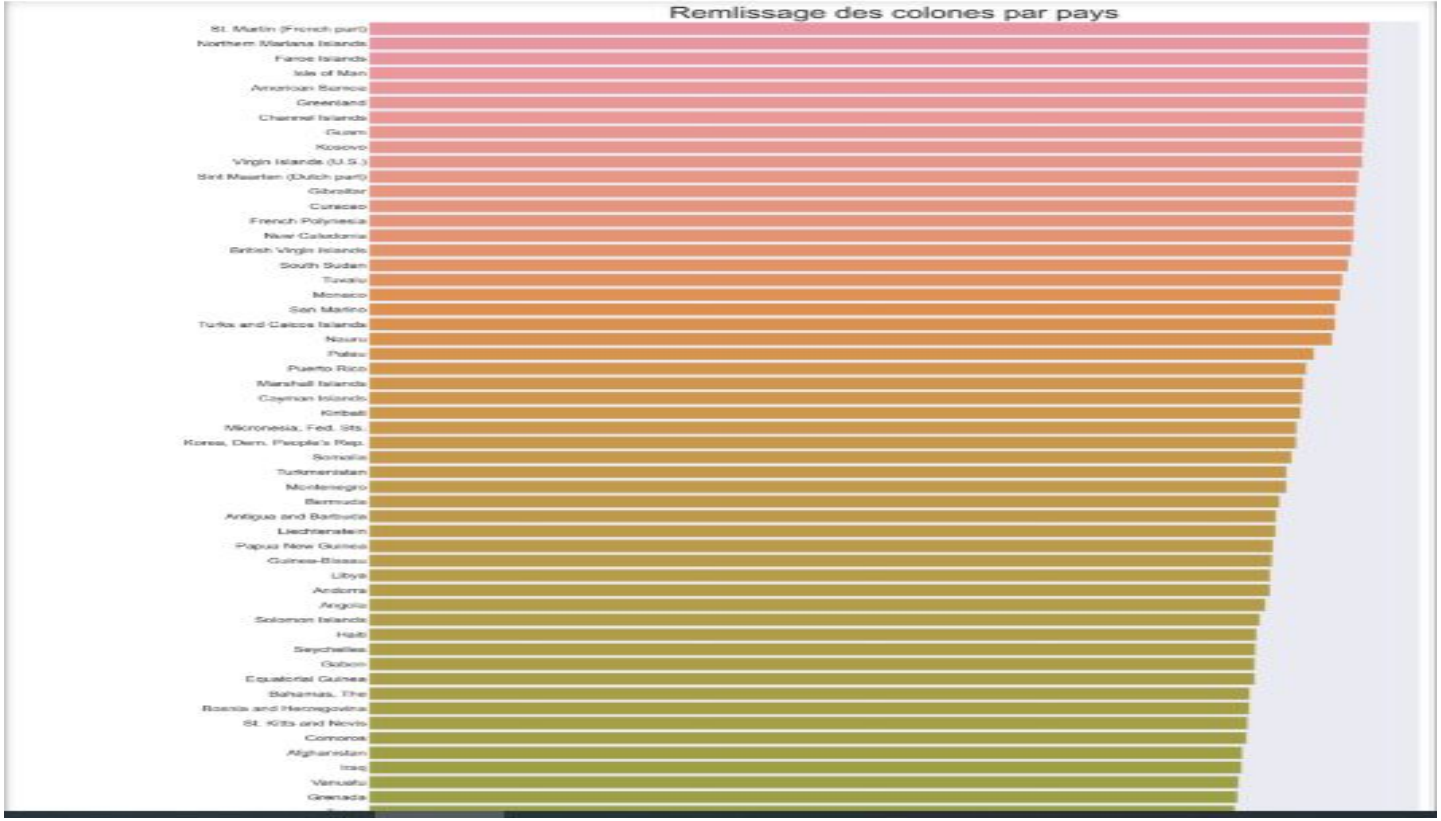
3. Filtrage des données

L’intervalle des années qu’on peut utiliser dans notre étude :



3. Filtrage des données

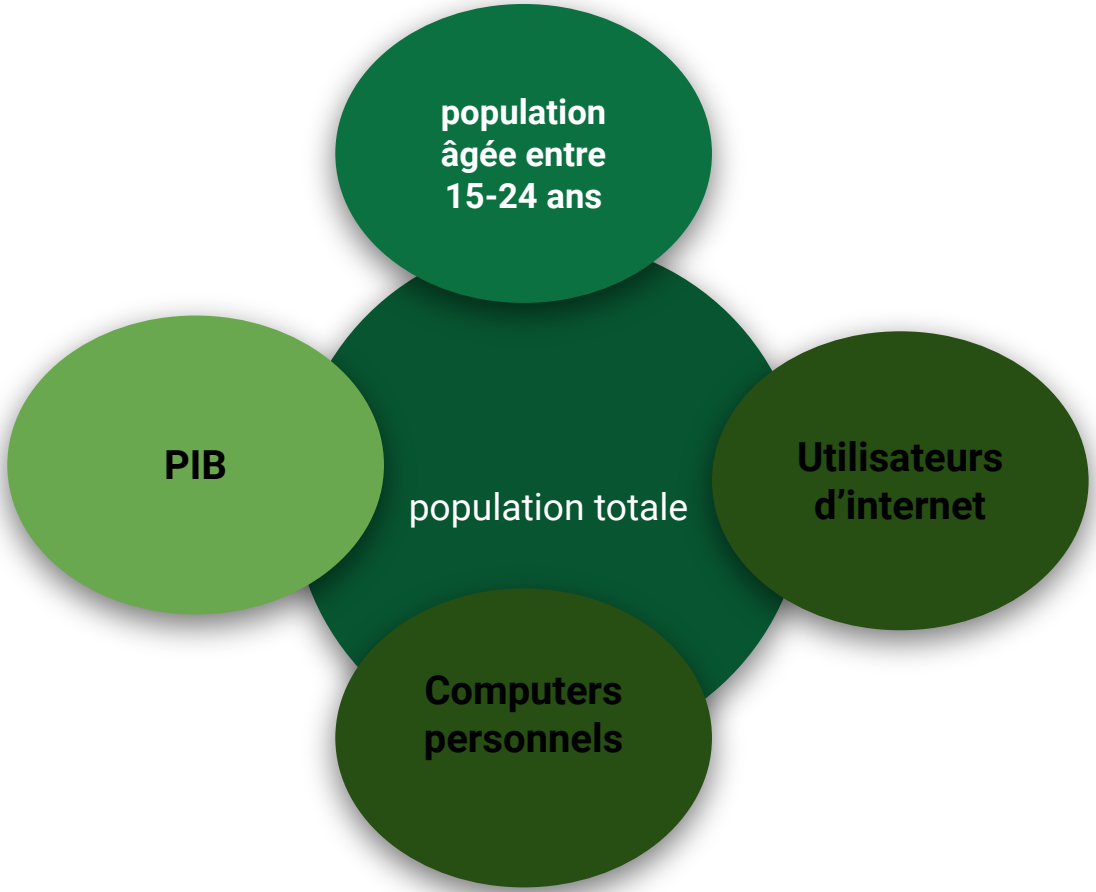
Filtrage par pays :



3. Filtrage des données

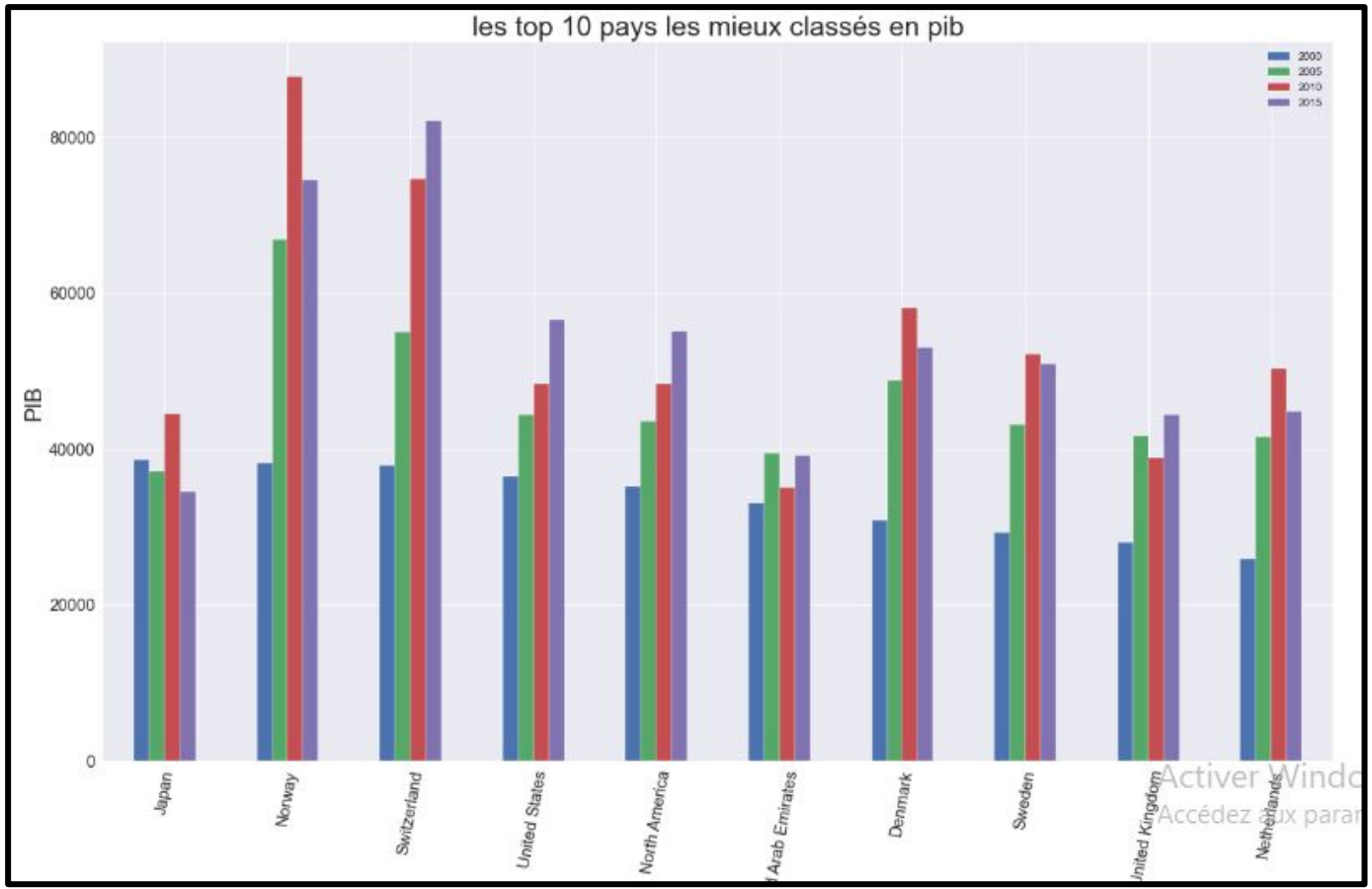
Indicator Name	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	NaN
Population, total	215	215	215	215	215	215	215	215	215	215	215	215	214	214	210	210	217
Population growth (annual %)	215	215	215	215	215	215	215	215	215	215	215	214	214	214	210	210	217
Official entrance age to primary education (years)	208	208	208	208	208	208	208	208	208	208	208	208	208	208	208	208	217
Theoretical duration of primary education (years)	208	208	208	208	208	208	208	208	208	208	208	208	208	208	208	208	217
Theoretical duration of pre-primary education (years)	205	205	205	205	205	205	205	205	205	205	205	205	205	205	205	205	217
Official entrance age to lower secondary education (years)	205	205	205	205	205	205	205	205	205	205	205	204	204	205	205	205	217
Theoretical duration of lower secondary education (years)	205	205	205	205	205	205	205	205	205	205	205	204	204	205	205	205	217
Theoretical duration of secondary education (years)	205	205	205	205	205	205	205	205	205	205	205	204	204	205	205	205	217
GDP at market prices (current US$)	199	199	203	203	204	204	205	204	203	202	203	203	199	200	197	196	217
GDP per capita (current US$)	199	199	203	203	204	204	205	204	203	202	203	203	199	200	197	196	217
Internet users (per 100 people)	196	197	199	193	196	198	197	204	203	202	202	204	202	201	201	201	217
Population, ages 15-64 (% of total)	194	194	194	194	194	194	194	194	194	194	194	194	193	193	193	193	217
Population, ages 0-14 (% of total)	194	194	194	194	194	194	194	194	194	194	194	194	193	193	191	191	217
Population, ages 0-14, total	194	194	194	194	194	194	194	194	194	194	194	194	193	193	191	191	217
Population, ages 15-64, total	194	194	194	194	194	194	194	194	194	194	194	194	193	193	191	191	217
Population of the official age for tertiary education, both sexes (number)	193	194	196	195	195	194	194	194	194	195	192	175	169	163	195	129	217
GDP at market prices (constant 2005 US$)	192	193	198	198	198	198	198	198	198	198	202	198	196	196	194	193	217
GDP per capita (constant 2005 US$)	192	193	198	198	198	198	198	198	198	198	202	198	196	196	194	193	217
Mortality rate, under-5 (per 1,000)	192	192	192	192	192	192	192	192	192	192	192	192	192	192	191	191	217
Population of the official age for pre-primary (years)	191	193	195	194	194	194	194	195	194	195	193	196	194	194	194	193	217
Personal computers (per 100 people)	162	182	182	179	179	171	99	48	27	3	0	0	0	0	0	0	217
Population, ages 12-18, total	190	191	192	192	191	191	187	181	181	181	181	181	181	181	181	181	217
Population, ages 13-16, total	190	191	192	192	191	191	187	181	181	181	181	181	181	181	181	181	217
Population, ages 13-17, total	190	191	192	192	191	191	187	181	181	181	181	181	181	181	181	181	217
Population, ages 13-18, total	190	191	192	192	191	191	187	181	181	181	181	181	181	181	181	181	217
Population, ages 13-19, total	190	191	192	192	191	191	187	181	181	181	181	181	181	181	181	181	217
Population, ages 14-18, total	190	191	192	192	191	191	187	181	181	181	181	181	181	181	181	181	217
Population, ages 14-19, total	190	191	192	192	191	191	187	181	181	181	181	181	181	181	181	181	217

3. Filtrage des données / Indicateurs choisis



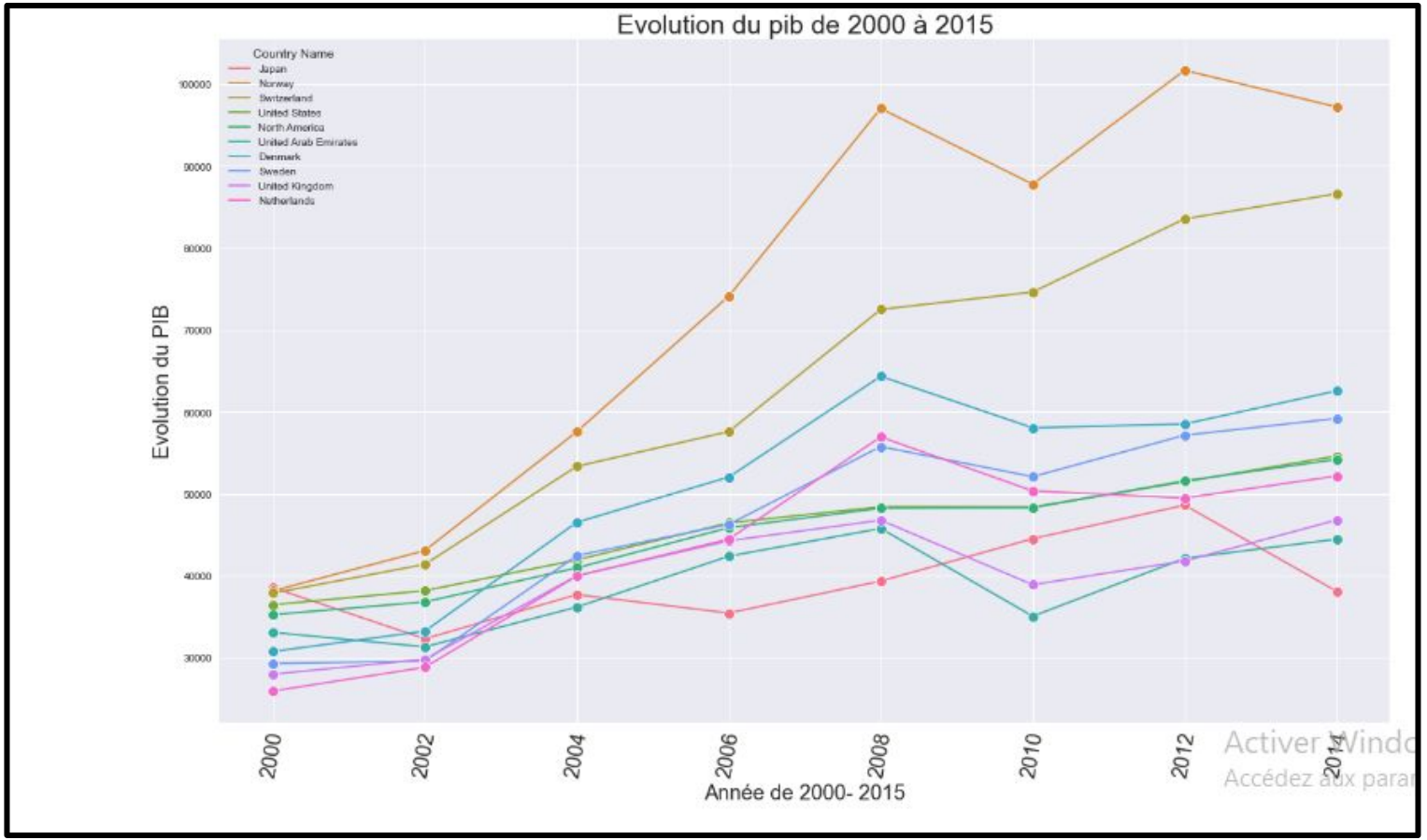
4. Visualisation des indicateurs choisis:

PIB



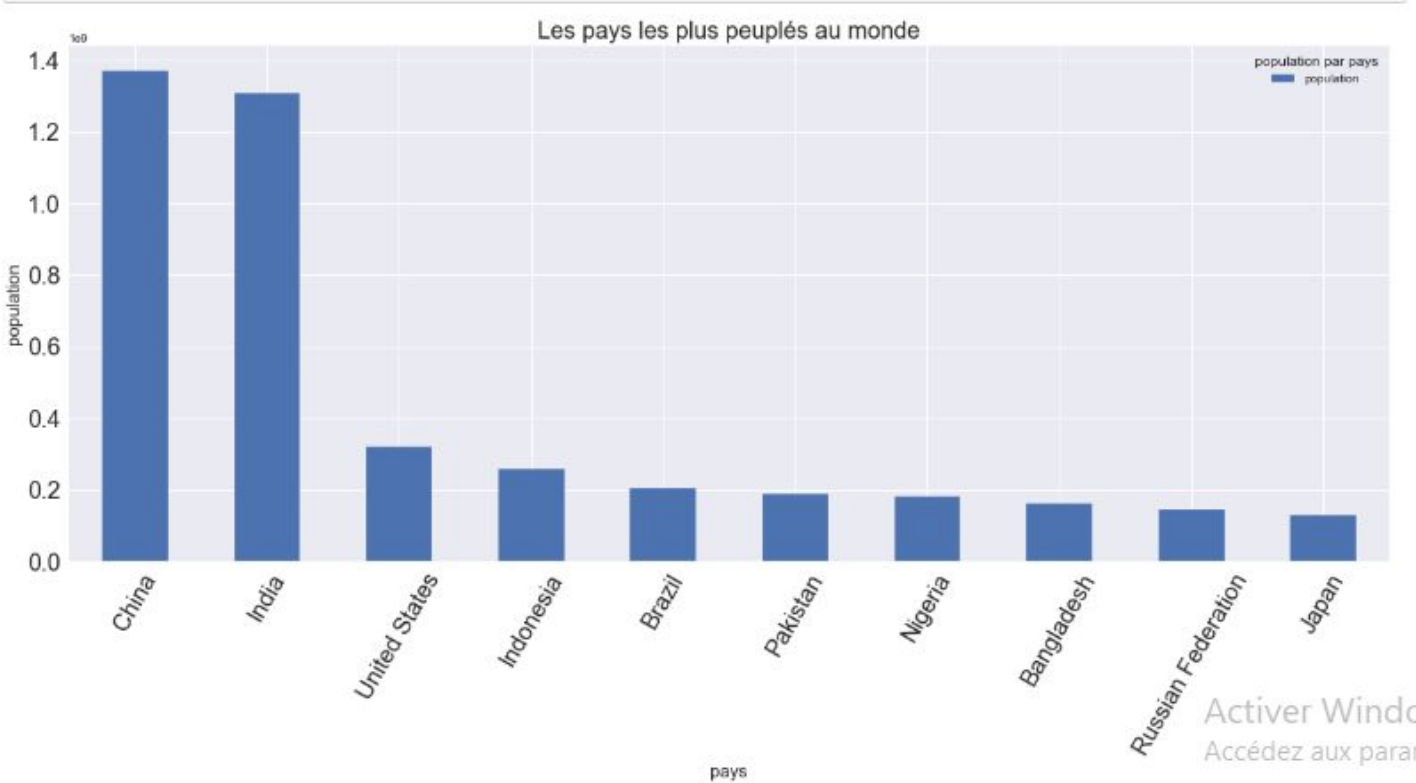
4. Visualisation des indicateurs choisis:

PIB



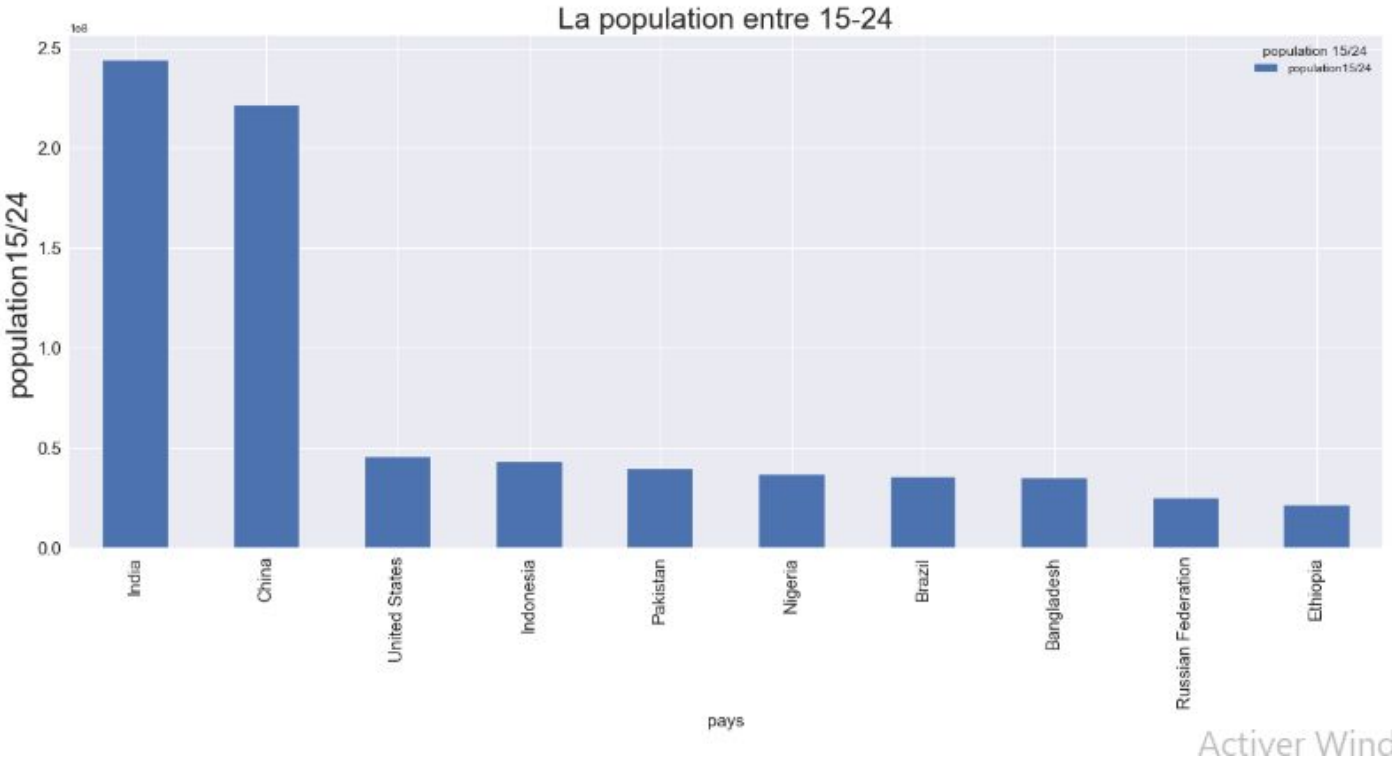
4. Visualisation des indicateurs choisis:

Population



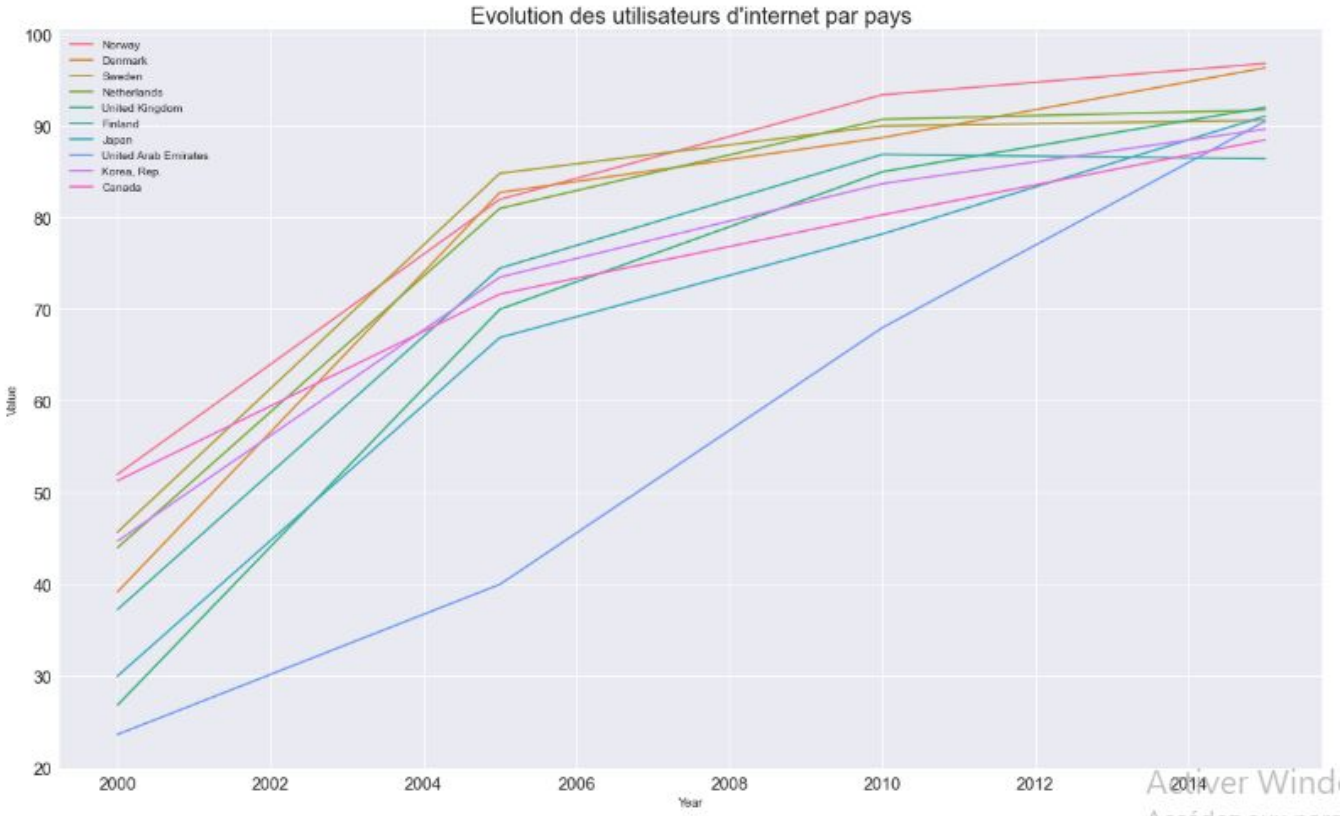
4. Visualisation des indicateurs choisis:

Population 15-24



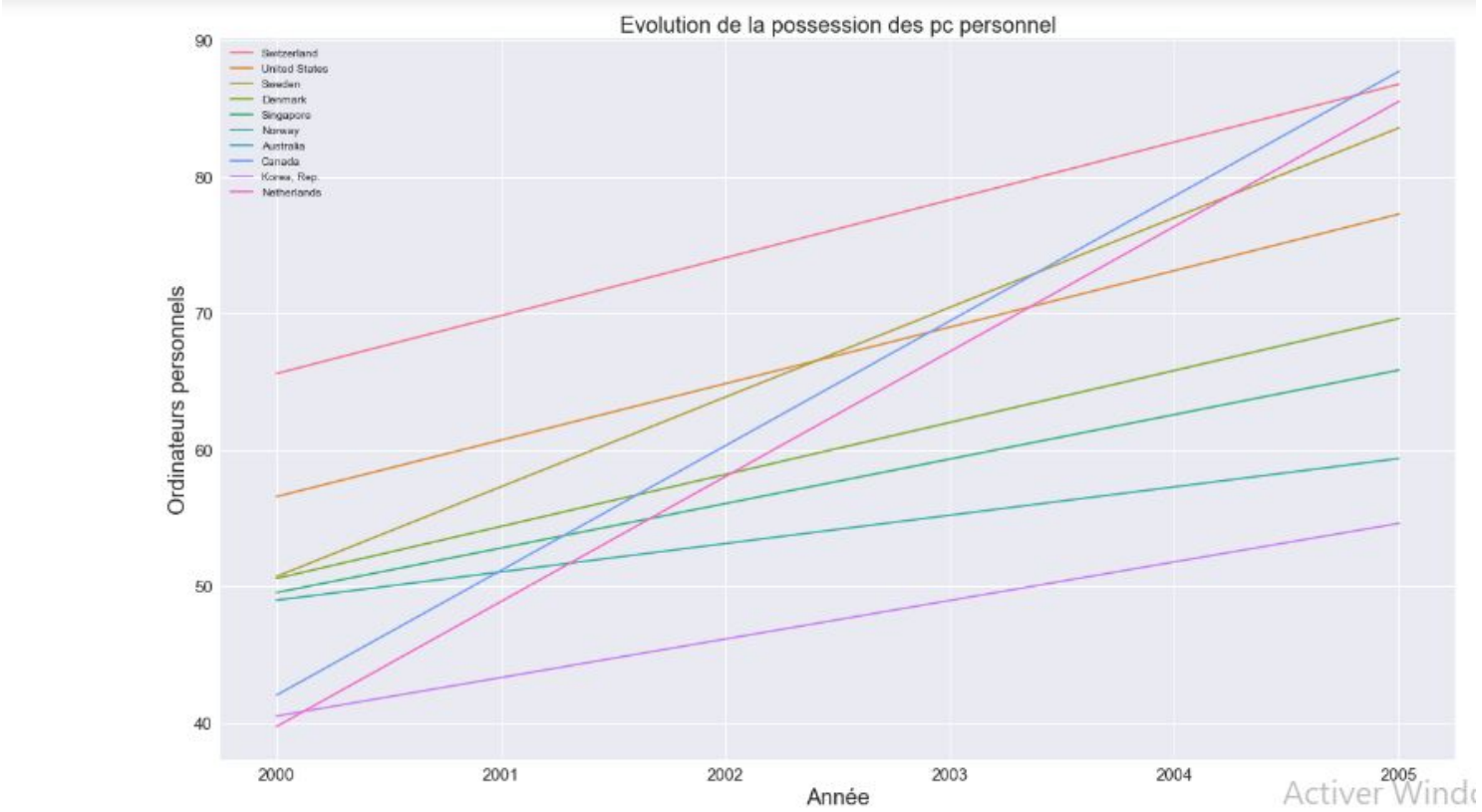
4. Visualisation des indicateurs choisis:

Internet Users



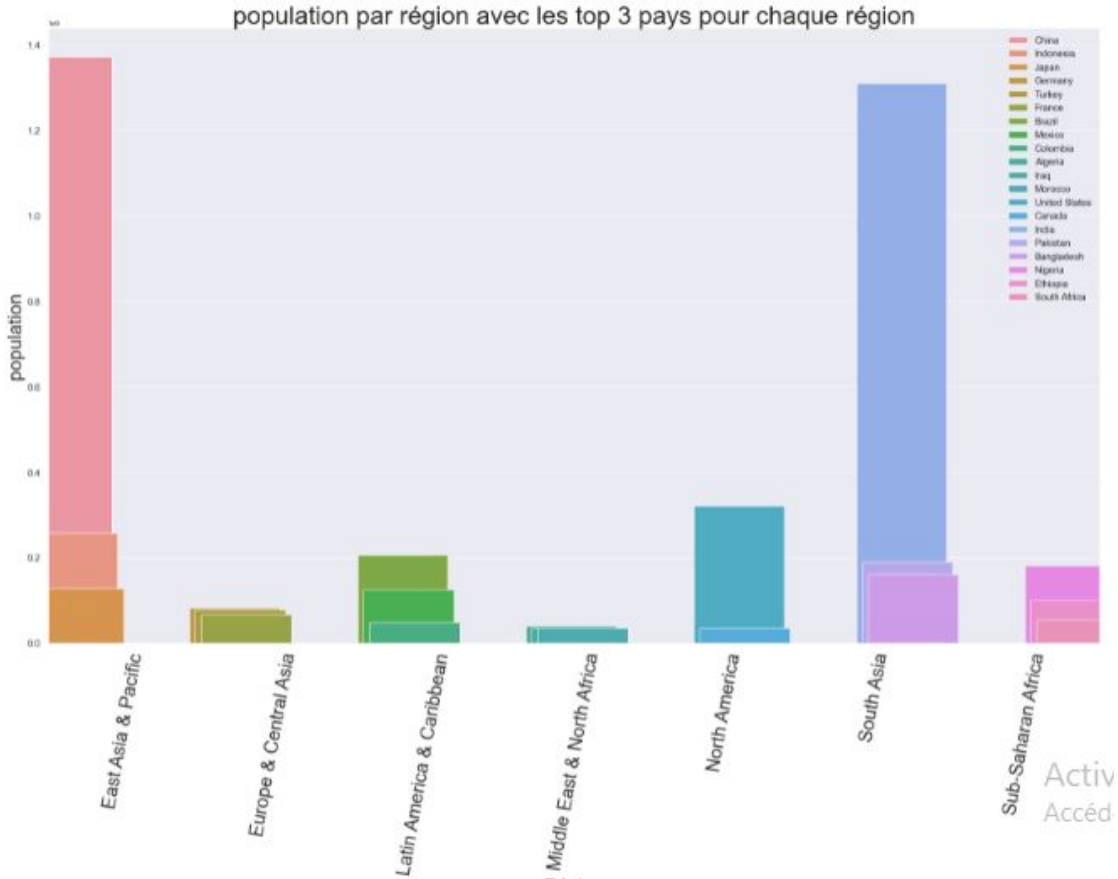
4. Visualisation des indicateurs choisis:

Ordinateurs personnels



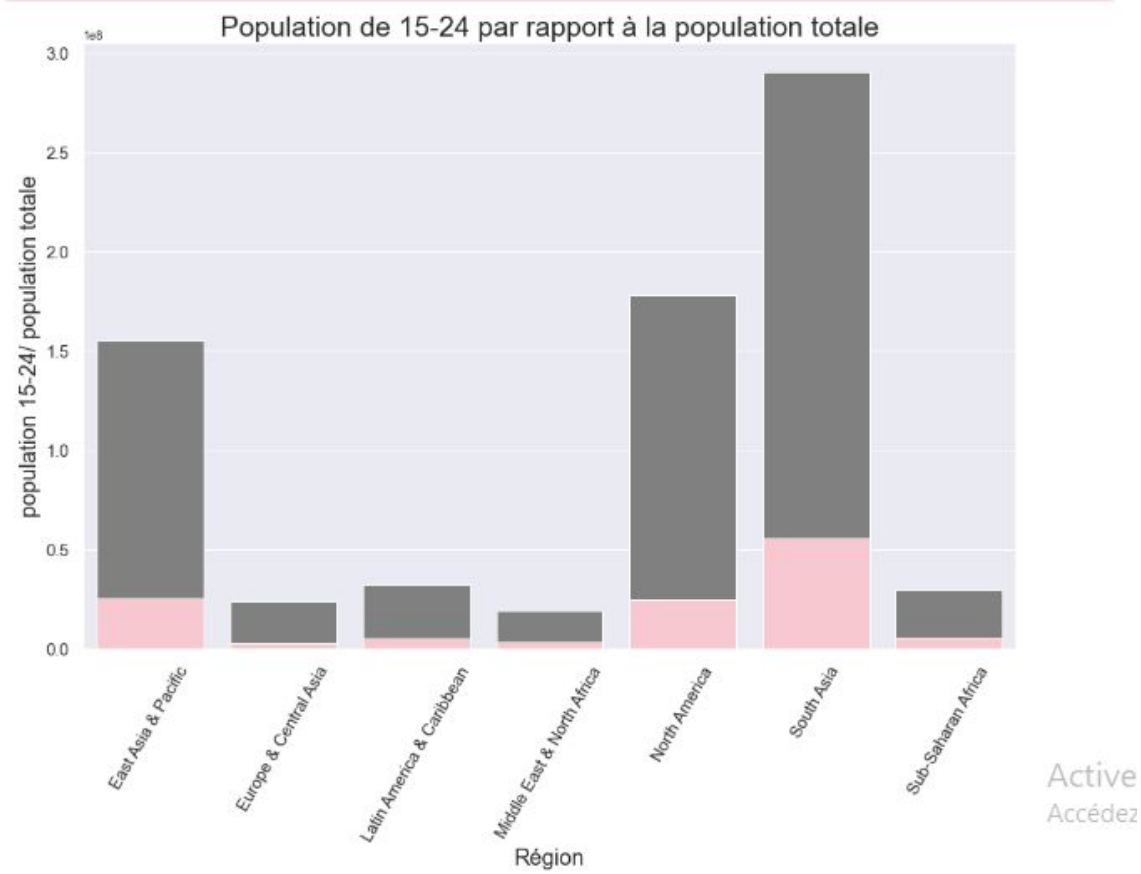
4. Visualisation des indicateurs choisis:

Population par région



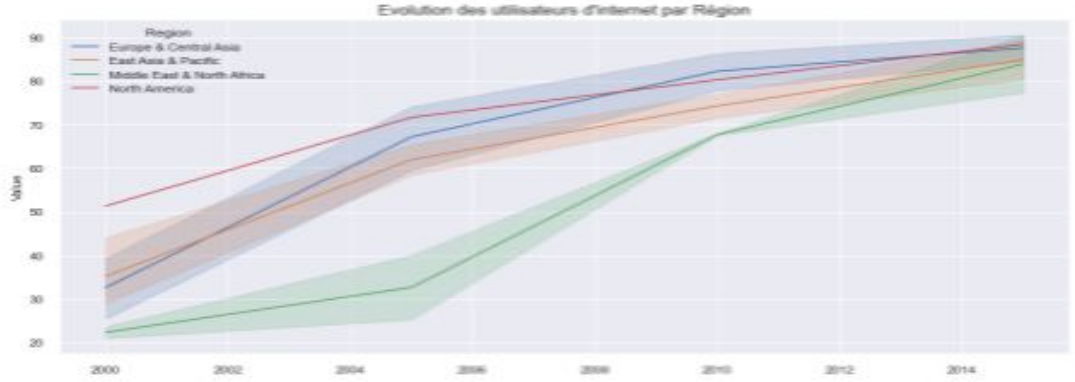
4. Visualisation des indicateurs choisis:

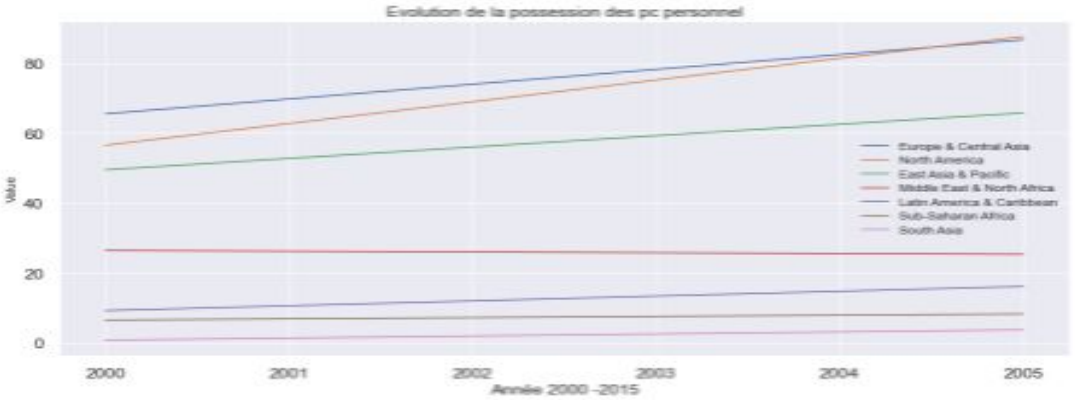
Population 15-24 par région



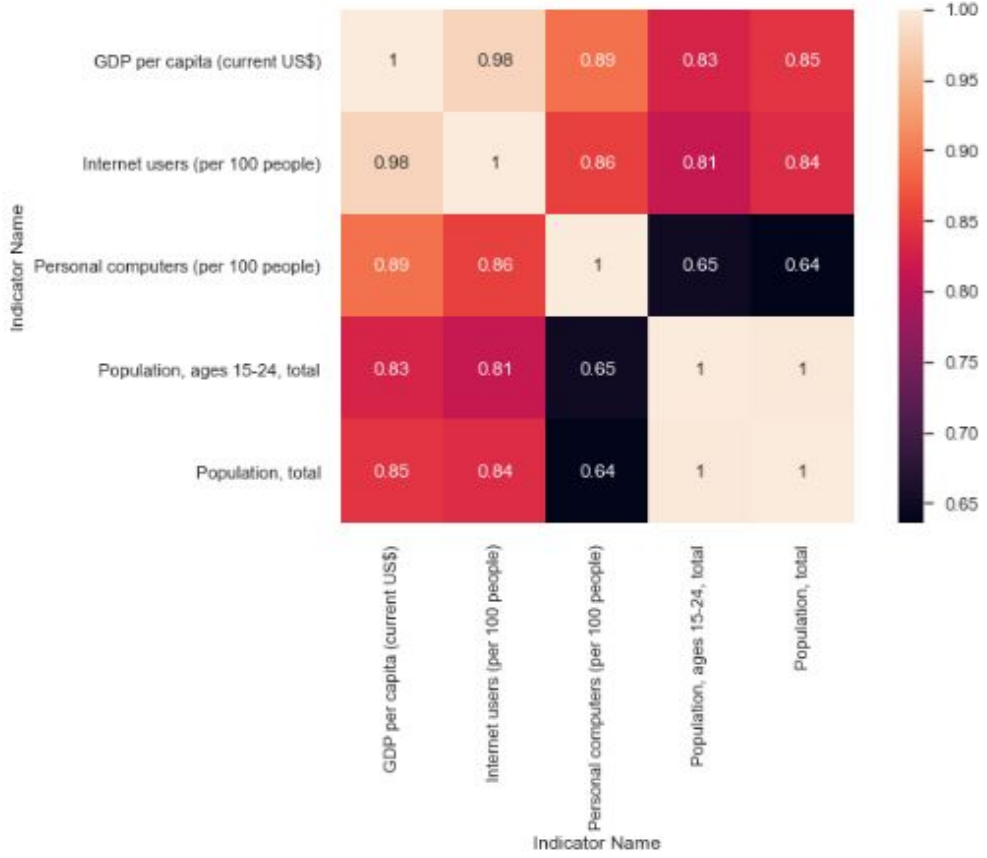
4. Visualisation des indicateurs choisis:

Utilisateurs d'Internet et ordinateurs personnels par région

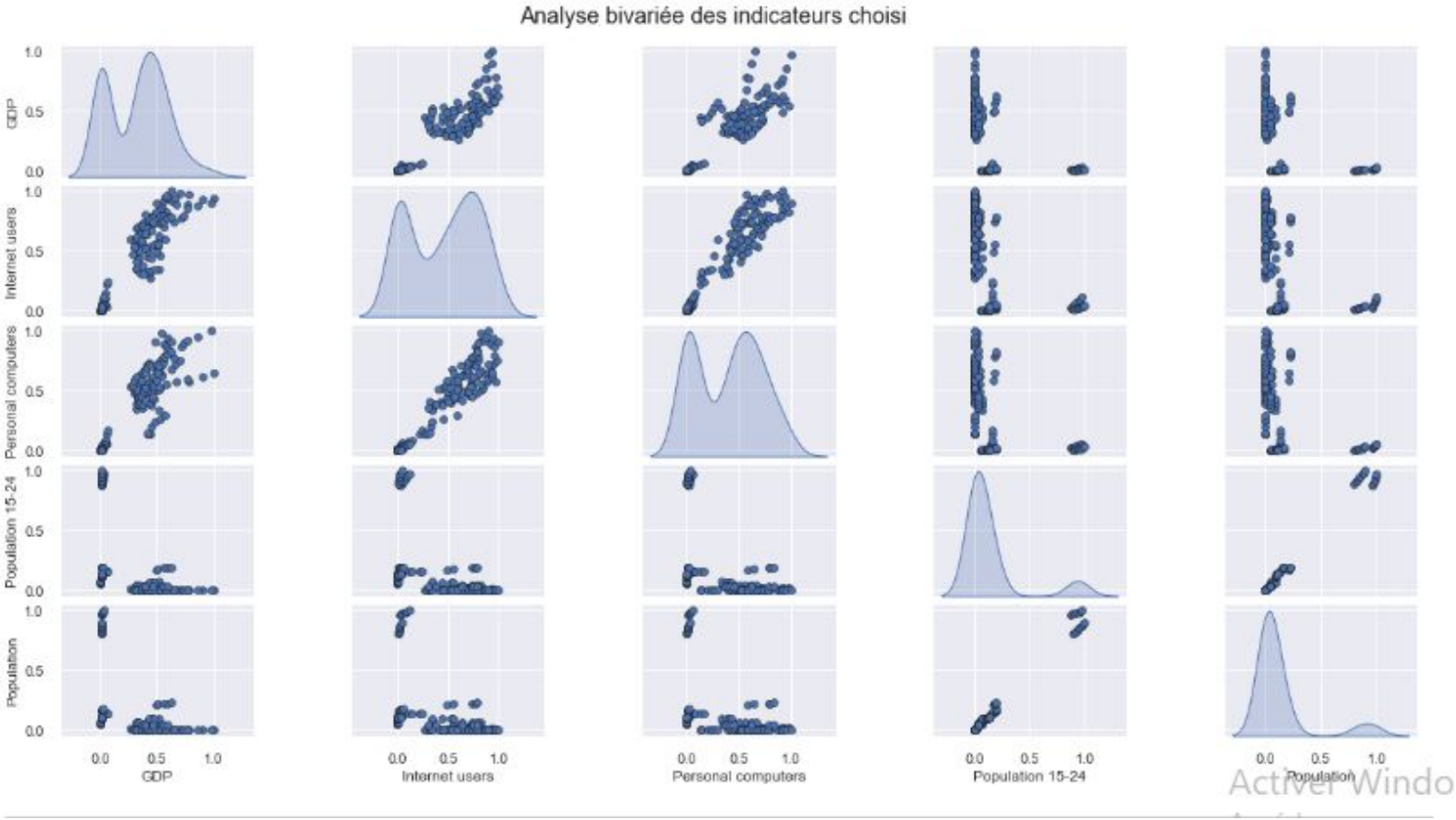




5. Analyse bivariée des indicateurs choisis :



5. Analyse bivariée des indicateurs choisis :



6 . Classement des pays par potentiel d'implémentation via un scoring :

PIB	Population	population 15-24	Utilisateurs d'Internet	Ordinateurs personnel
1.Japan 2. Norway 3.Switzeland 4.United States 5.United Arab Emirates 6.Denmark 7.Sweden 8.United Kingdom 9.Netherlands 10.Hong Kong SAR	1.China 2. India 3.United States 4Indonesia 5. Brazil 6. pakistan 7.Nigeria 8.Bangladesh 9.Russin Federation 10.Japan	1.India 2.China 3.United States 4Indonesia 5. Brazil 6. pakistan 7.Nigeria 8.Bangladesh 9.Russin Federation 10.Ethiopia	1.Norway 2. Denmark 3.Sweden 4. Netherlands 5.United Kingdom 6.Finland 7.Japan 8.United Arab Emirates 9.Canada 10.Germany	1.Switezerland 2.United States 3.Sweden 4.Denmark 5.Singapaore 6.Australia 8.Canada 9.Netherlands 10.Finld

6 . Classement des pays par potentiel d’implémentation via un scoring :

mais quels pays choisir parmi ces listes de pays ??

6 . Classement des pays par potentiel d'implémentation via un scoring :

Attribution des poids pour chaque pays parmi les top 10 relatifs à chaque indicateurs

	Australia	Bangladesh	Brazil	Canada	China	Denmark	Ethiopia	Finland	Germany	Hong Kong SAR, China	India	Indonesia	Japan	Netherlands
poids_pib	0.0	0.0	0.0	0.0	0.0	5.0	0.0	0.0	0.0	1.0	0.0	0.0	10.0	2.0
poids_population15_24	0.0	3.0	4.0	0.0	9.0	0.0	1.0	0.0	0.0	0.0	10.0	7.0	0.0	0.0
poids_population	0.0	3.0	6.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0	9.0	7.0	1.0	0.0
poids_Internet	0.0	0.0	0.0	2.0	0.0	9.0	0.0	5.0	1.0	0.0	0.0	0.0	4.0	7.0
poids_computer	4.0	0.0	0.0	3.0	0.0	7.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0

6 . Classement des pays par potentiel d'implémentation via un scoring :

Pour les 5 indicateurs réaliser des itérations et à chaque itération on attribue des poids allant de 1 à 5 qui se diffère d'une itérations à une autre.

```
poids_indicateur= []

for i in list(itertools.permutations([1, 2, 3,4,5])):
    array_i= np.array(i)
    poids_indicateur.append(array_i)
```

```
Entrée [112]: poids_indicateur

Out[112]: [array([1, 2, 3, 4, 5]),
           array([1, 2, 3, 5, 4]),
           array([1, 2, 4, 3, 5]),
           array([1, 2, 4, 5, 3]),
           array([1, 2, 5, 3, 4]),
           array([1, 2, 5, 4, 3]),
           array([1, 3, 2, 4, 5]),
           array([1, 3, 2, 5, 4]),
           array([1, 3, 4, 2, 5]),
           array([1, 3, 4, 5, 2]),
           array([1, 3, 5, 2, 4]),
           array([1, 3, 5, 4, 2]),
           array([1, 4, 2, 3, 5]),
           array([1, 4, 2, 5, 3]),
           array([1, 4, 3, 2, 5]),
           array([1, 4, 3, 5, 2]),
           array([1, 4, 5, 2, 3]),
           array([1, 4, 5, 3, 2]),
           array([1, 5, 2, 3, 4]),
           array([1, 5, 2, 4, 3]),
```

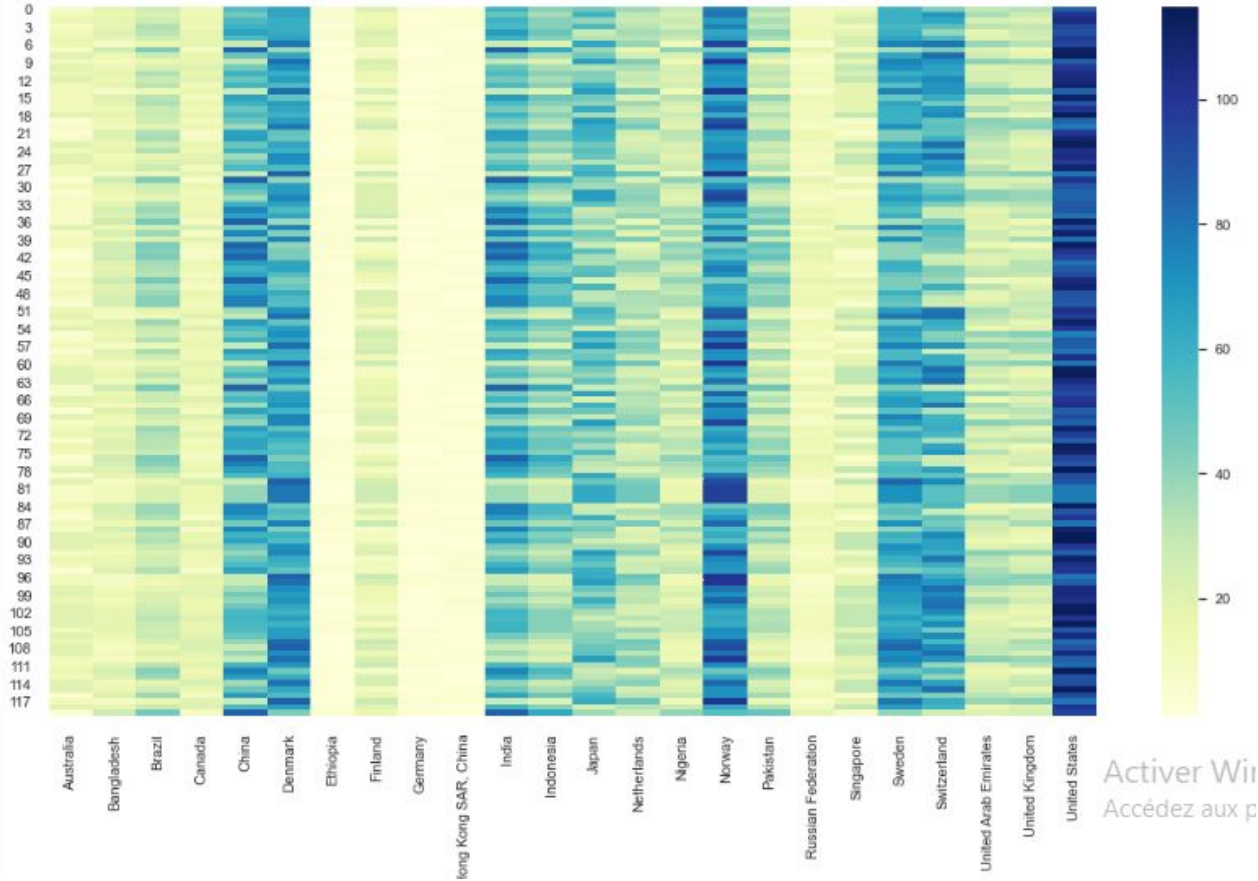
6 . Classement des pays par potentiel d'implémentation via un scoring :

Chaque ligne de la matrice représente le produit matriciel de l'itération i du vecteur poids indicateur multiplié par la matrice poids pays

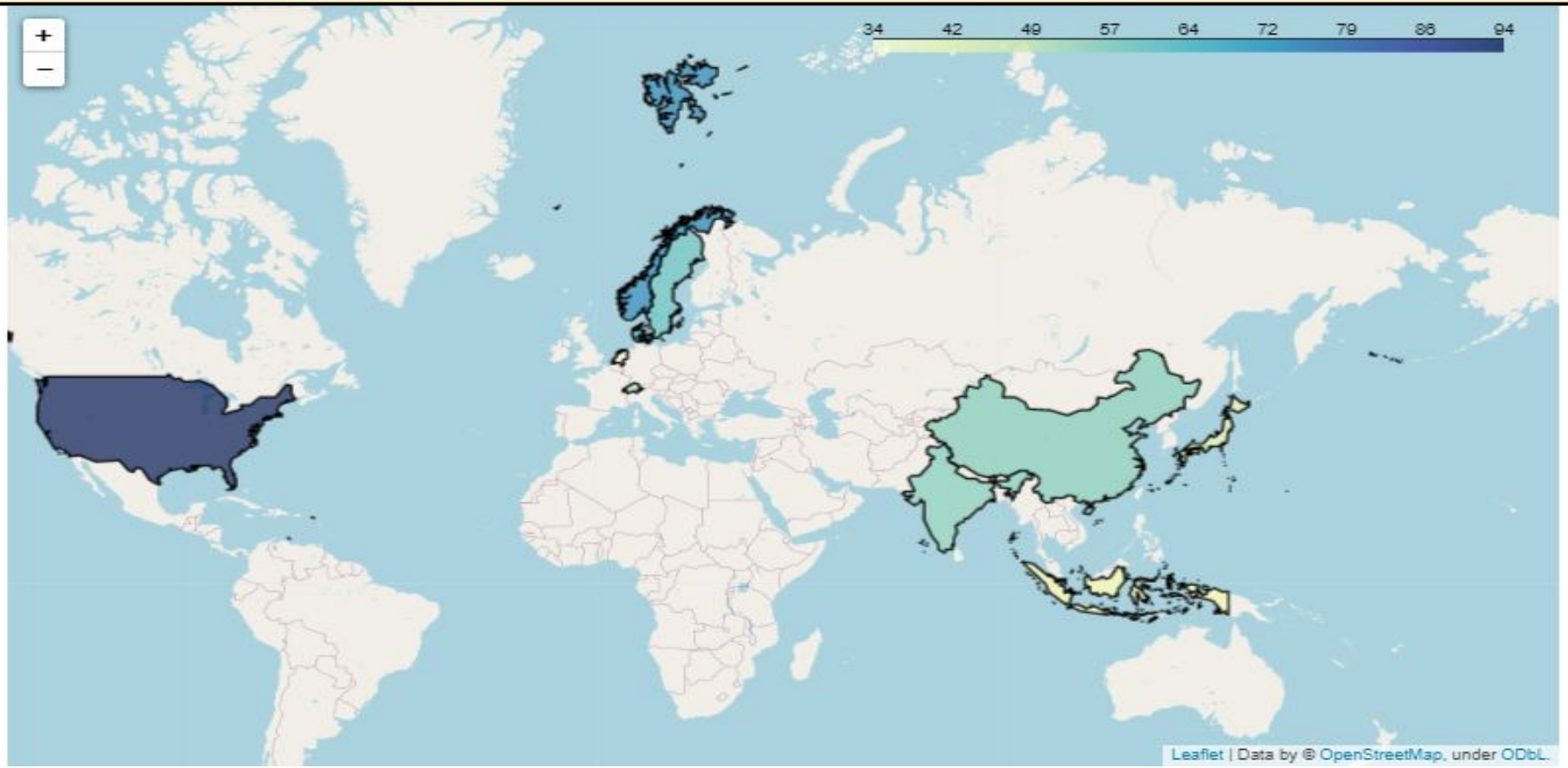
Australia	Bangladesh	Brazil	Canada	China	Denmark	Ethiopia	Finland	Germany	Hong Kong SAR, China	India	Indonesia	Japan	Netherlands	Nigeria	Norway	Pakistan	Russian Federation	S
20.0	15.0	26.0	23.0	48.0	76.0	2.0	25.0	4.0	1.0	47.0	35.0	29.0	40.0	22.0	74.0	27.0	10.0	
16.0	15.0	26.0	22.0	48.0	78.0	2.0	29.0	5.0	1.0	47.0	35.0	33.0	45.0	22.0	79.0	27.0	10.0	
20.0	18.0	32.0	21.0	58.0	67.0	2.0	20.0	3.0	1.0	56.0	42.0	26.0	33.0	26.0	64.0	32.0	12.0	
12.0	18.0	32.0	19.0	58.0	71.0	2.0	28.0	5.0	1.0	56.0	42.0	34.0	43.0	26.0	74.0	32.0	12.0	
16.0	21.0	38.0	18.0	68.0	60.0	2.0	19.0	3.0	1.0	65.0	49.0	27.0	31.0	30.0	59.0	37.0	14.0	
12.0	21.0	38.0	17.0	68.0	62.0	2.0	23.0	4.0	1.0	65.0	49.0	31.0	36.0	30.0	64.0	37.0	14.0	
20.0	15.0	24.0	23.0	47.0	76.0	3.0	25.0	4.0	1.0	48.0	35.0	28.0	40.0	23.0	74.0	28.0	10.0	
16.0	15.0	24.0	22.0	47.0	78.0	3.0	29.0	5.0	1.0	48.0	35.0	32.0	45.0	23.0	79.0	28.0	10.0	
20.0	21.0	36.0	19.0	67.0	58.0	3.0	15.0	2.0	1.0	66.0	49.0	22.0	26.0	31.0	54.0	38.0	14.0	

6 . Classement des pays par potentiel d'implémentation via un scoring :

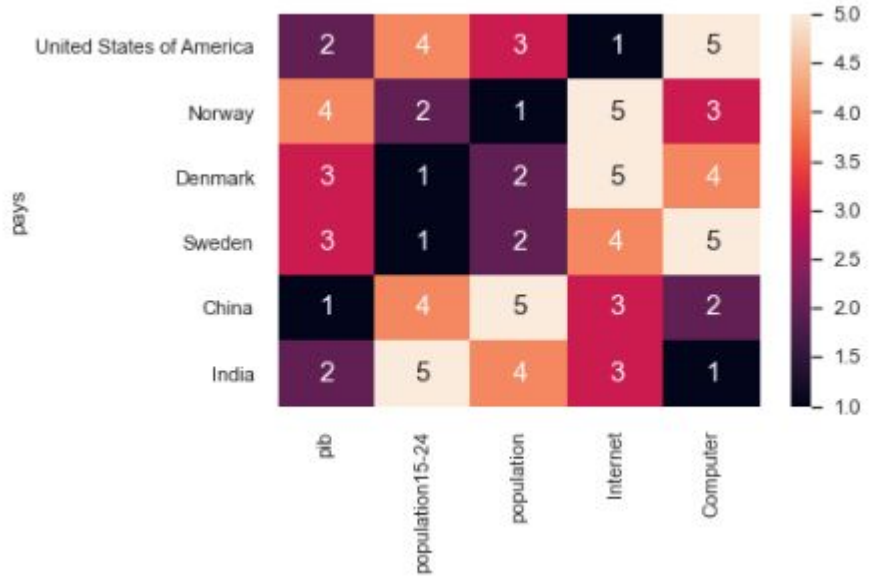
	pays	scores
0	United States	97.033333
1	Norway	73.233333
2	Denmark	63.725000
3	Sweden	60.766667
4	Switzerland	57.050000
5	India	55.025000
6	China	55.016667
7	Japan	46.575000
8	Indonesia	40.541667
9	Netherlands	32.708333



6 . Classement des pays par potentiel d’implémentation via un scoring :



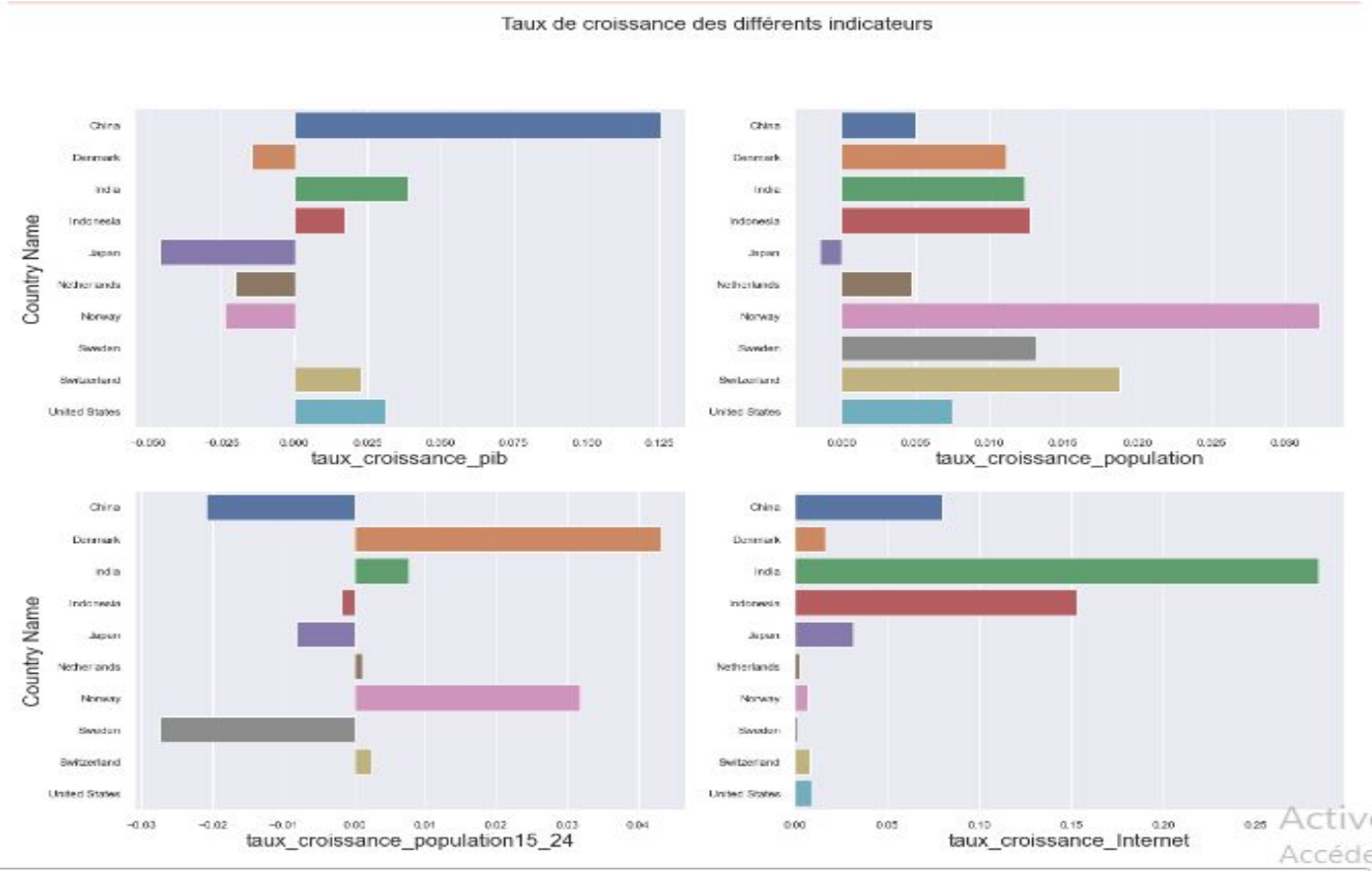
6 . Classement des pays par potentiel d'implémentation via un scoring :



7 . Evaluation du potentiel des pays choisis :

```
pib_change = dataN.groupby('Country Name')['GDP'].apply(lambda x: x.pct_change().mean()).reset_index(name='taux_croissance_pib'
pib_change['categorie_pib'] = np.where(pib_change['taux_croissance_pib'] < 0, 'decreasing', np.where(pib_change['taux_croissanc


population_change = dataN.groupby('Country Name')['Population'].apply(lambda x: x.pct_change().mean()).reset_index(name='taux_c
population_change['categorie_population'] = np.where(population_change['taux_croissance_population'] < 0, 'decreasing', np.wher

population15_24_change = dataN.groupby('Country Name')['Population 15-24'].apply(lambda x: x.pct_change().mean()).reset_index(n
population15_24_change['categorie_population15_24'] = np.where(population15_24_change['taux_croissance_population15_24'] < 0, '

Internet_users_change = dataN.groupby('Country Name')['Internet users'].apply(lambda x: x.pct_change().mean()).reset_index(name
Internet_users_change['categorie_Internet'] = np.where(Internet_users_change['taux_croissance_Internet'] < 0, 'decreasing', np.

Computers_change = dataN.groupby('Country Name')['Personal computers'].apply(lambda x: x.pct_change().mean()).reset_index(name=
Computers_change['categorie_computers'] = np.where(Computers_change['taux_croissance_Computers'] < 0, 'decreasing', np.where(Co


from functools import reduce
dfs= [pib_change, population_change, population15_24_change,Internet_users_change, Computers_change]
df_final = reduce(lambda left,right: pd.merge(left,right,on='Country Name'), dfs)
df_final= df_final.set_index('Country Name')
df_final
```

7 . Evaluation du potentiel des pays choisis :



7 . Evaluation du potentiel des pays choisis :

Country Name	categorie_pib	categorie_population	categorie_population15_24	categorie_Internet	Nbr_Increasing
China	increasing	increasing	decreasing	increasing	3
Denmark	decreasing	increasing	increasing	increasing	3
India	increasing	increasing	increasing	increasing	4
Indonesia	increasing	increasing	decreasing	increasing	3
Japan	decreasing	decreasing	decreasing	increasing	1
Netherlands	decreasing	increasing	increasing	increasing	3
Norway	decreasing	increasing	increasing	increasing	3
Sweden	decreasing	increasing	decreasing	increasing	2
Switzerland	increasing	increasing	increasing	increasing	4
United States	increasing	increasing	increasing	increasing	4

Conclusion :

- Qualité de ce jeu de données

Le jeu contient beaucoup de données non renseignées

- Description des informations

5 Jeu de données : Country , data , indicateurs , description et source

- Informations pertinentes à la problématique

Des indicateurs pertinents ont été sélectionnés

- Ordres de grandeurs de quelques indicateurs pertinents

Des informations intéressantes ont pu être mises en évidences