PROJET N°2 :

ANALYSEZ DES DONNÉES DE SYSTÈMES ÉDUCATIFS

Soutenance du P2: le 06/12/2021

Version notebook : **6.3.0**
Version Python : **3.8.8**
Version Pandas : **1.2.4**
Version Seaborn : **0.11.1**
Version Matplotlib: **3.3.4**

Plan

academy

- Problématique
- Mission
- Processus d’analyse exploratoire
- Présentation jeu données
- Sélection des indicateurs
- Méthode d’analyse et de traitement
- Comparaison par indicateur
- Représentation graphique d’indicateurs
- Sélection des pays / régions potentiels
- Conclusion

Problématique

Contexte:

academy

Academy une **start-up** de la **EdTech**, qui propose des formations en ligne pour un public de niveau lycée et université.

Objectif du projet: déterminer si les données sur l'éducation de la banque mondiale permettent d'informer le projet d'expansion et répondre à ces questions .

- Quels sont les pays avec un fort potentiel de clients?
- L'évolution de ce potentiel de clients?
- Dans quels pays l'entreprise doit elle opérer en priorité?

THE WORLD BANK
IBRD • IDA

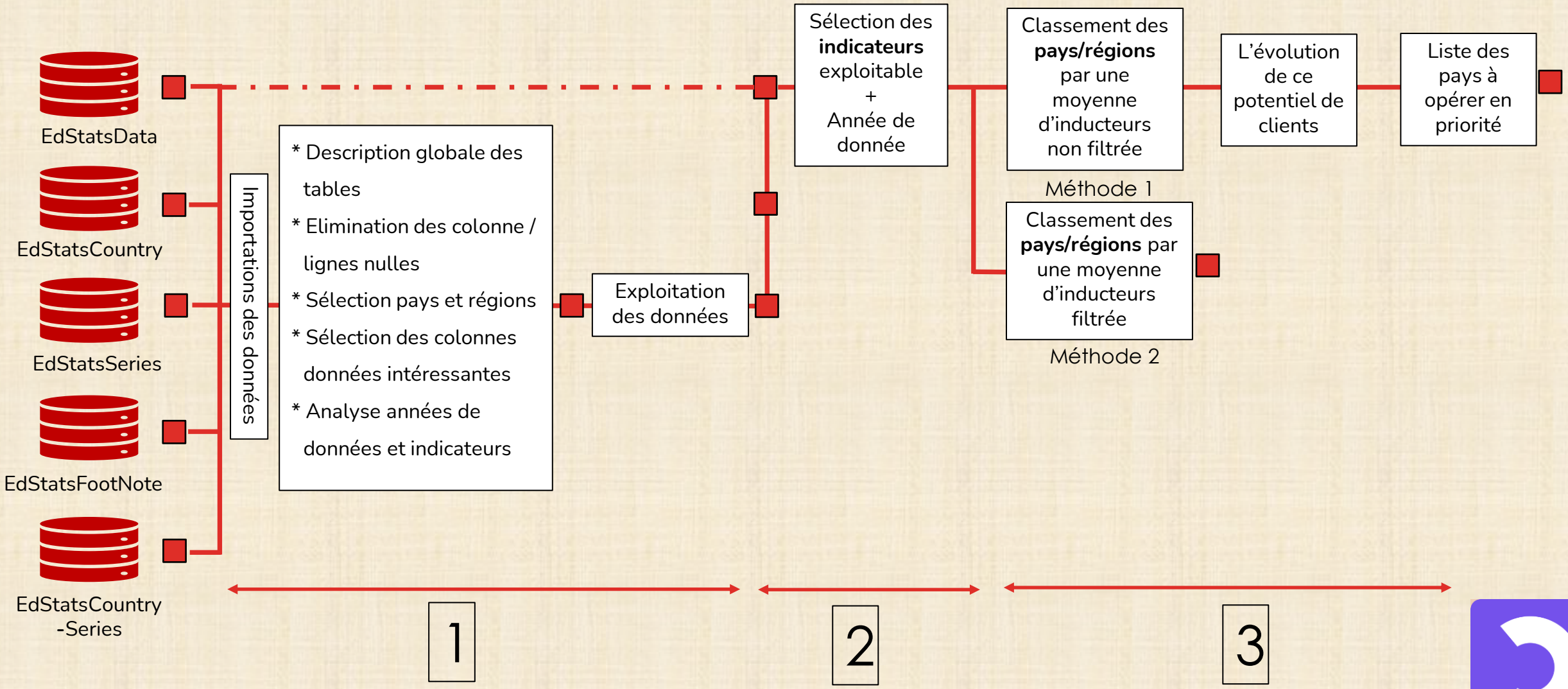
Mission

academy

Analyse exploratoire:

- Valider la qualité de ce jeu de données
- Décrire les informations contenues dans le jeu de données.
- 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*)

Processus d'analyse exploratoire



1

Présentation du jeu données

academy

EdStatsData

EdStatsData: le fichier le plus volumineux, il contient toutes les données et l'évolution de chaque inducteur.
Taille : 886930 lignes, 70 colonnes, nombre de cases: 62085100, nombre de valeurs nulles: 53455179, nombre de valeurs non nulles: 8629921, le pourcentage des valeurs nulles: 86.1%,le pourcentage des valeurs non nulles: 13.9 %
Données depuis 1970 pour 242 pays et régions

EdStatsCountry

EdStatsCountry: contient des infos globales sur l'économie des pays, pour certains pays il manque de données.
Taille : 241 lignes, 32 colonnes, nombre de cases: 7712, nombre de valeurs nulles: 2354, nombre de valeurs non nulles: 5358, le pourcentage des valeurs nulles: 30.5 %,le pourcentage des valeurs non nulles: 69.5 %
Données depuis 1970 pour 241 pays et régions, aucun doublon.

EdStatsSeries

EdStatsSeries: contient des informations sur les indicateurs socio économiques.
Taille : 3665 lignes, 21 colonnes, nombre de cases: 76965, nombre de valeurs nulles: 55203, nombre de valeurs non nulles: 21762, le pourcentage des valeurs nulles: 71.7 %, le pourcentage des valeurs non nulles: 28.3 %

EdStatsFootNote

EdStatsFootNote: contient des informations sur l'année d'origine des données indicateurs.
Taille : 643638 lignes, 5 colonnes, nombre de cases: 3218190, nombre de valeurs nulles: 643638, nombre de valeurs non nulles: 2574552, le pourcentage des valeurs nulles: 20 %, le pourcentage des valeurs non nulles: 80 %
Une colonne nulle « Unnamed 4 »

EdStatsCountry
-Series

EdStatsFootNote: contient des informations sur la source des données.
Taille : 613 lignes, 4 colonnes, nombre de cases: 2452, nombre de valeurs nulles: 613, nombre de valeurs non nulles: 1839, le pourcentage des valeurs nulles 25%, le pourcentage des valeurs non nulles: 75 %
Une colonne nulle « Unnamed 3 »

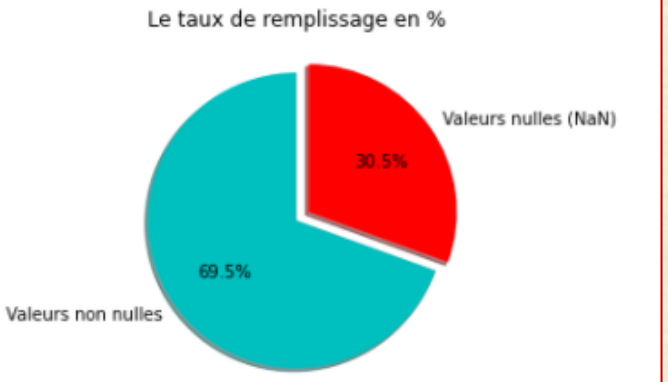
1

Présentation du jeu données

academy

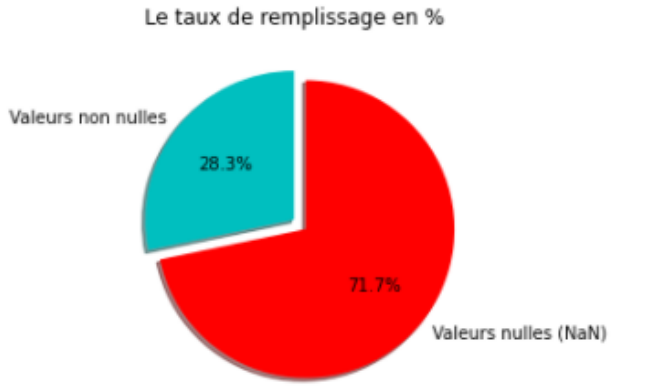
EdStatsCountry

```
calc_inf(country,True)

* Nombre de colonnes sans NaN -------: 4
* Nombre de colonnes NaN ------------: 1
* Nombre de colonnes mixtes----------: 27
* Nombre de ligne entièrement nulles : 0
* Nombre de ligne mixtes ---------- : 241
* Nombre de ligne sans NaN ----------: 0
* Nombre de lignes ------------------------: 241
* Nombre de colonnes ----------------------: 32
* Nombre de cases -------------------------: 7712
* Nombre de valeurs nulles ----------------: 2354
* Nombre de valeurs non nulles ------------: 5358
* le pourcentage des valeurs nulles ------: 30.5 %
* le pourcentage des valeurs non nulles --: 69.5 %
```



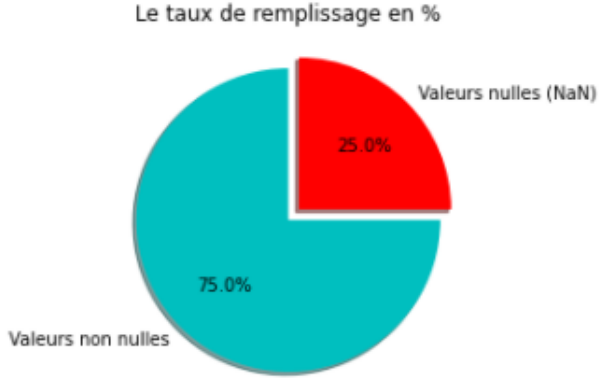
EdStatsSeries

```
calc_inf(series,True)

* Nombre de colonnes sans NaN -------: 5
* Nombre de colonnes NaN ------------: 6
* Nombre de colonnes mixtes----------: 10
* Nombre de ligne entièrement nulles : 0
* Nombre de ligne mixtes ---------- : 3665
* Nombre de ligne sans NaN ----------: 0
* Nombre de lignes ------------------------: 3665
* Nombre de colonnes ----------------------: 21
* Nombre de cases -------------------------: 76965
* Nombre de valeurs nulles ----------------: 55203
* Nombre de valeurs non nulles ------------: 21762
* le pourcentage des valeurs nulles ------: 71.7 %
* le pourcentage des valeurs non nulles --: 28.3 %
```



EdStatsCountry -Series

```
calc_inf(country_series, True)

* Nombre de colonnes sans NaN -------: 3
* Nombre de colonnes NaN ------------: 1
* Nombre de colonnes mixtes----------: 0
* Nombre de ligne entièrement nulles : 0
* Nombre de ligne mixtes ---------- : 613
* Nombre de ligne sans NaN ----------: 0
* Nombre de lignes ------------------------: 613
* Nombre de colonnes ----------------------: 4
* Nombre de cases -------------------------: 2452
* Nombre de valeurs nulles ----------------: 613
* Nombre de valeurs non nulles ------------: 1839
* le pourcentage des valeurs nulles ------: 25.0 %
* le pourcentage des valeurs non nulles --: 75.0 %
```



Présentation du jeu données

academy

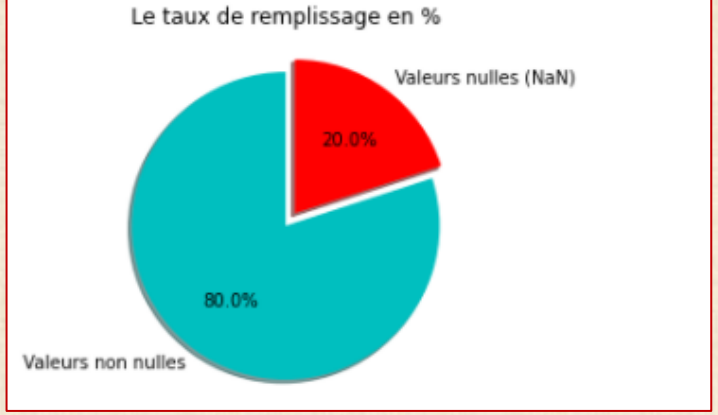
EdStatsData

```
calc_inf(EdStatsData, False)

* Nombre de colonnes sans NaN -------: 4
* Nombre de colonnes NaN ------------: 1
* Nombre de colonnes mixtes----------: 65
* Nombre de lignes ------------------------: 886930
* Nombre de colonnes ----------------------: 70
* Nombre de cases -------------------------: 62085100
* Nombre de valeurs nulles ----------------: 53455179
* Nombre de valeurs non nulles ------------: 8629921
* le pourcentage des valeurs nulles ------: 86.1 %
* le pourcentage des valeurs non nulles --: 13.9 %
```

L'argument de la fonction calc_inf() == False, pour désactiver le calcul des lignes, Le temps d'exécution est long
➔ le nombre de lignes énorme

EdStatsFootNote

```
calc_inf(FootNote, False)

* Nombre de colonnes sans NaN -------: 4
* Nombre de colonnes NaN ------------: 1
* Nombre de colonnes mixtes----------: 0
* Nombre de lignes ------------------------: 643638
* Nombre de colonnes ----------------------: 5
* Nombre de cases -------------------------: 3218190
* Nombre de valeurs nulles ----------------: 643638
* Nombre de valeurs non nulles ------------: 2574552
* le pourcentage des valeurs nulles ------: 20.0 %
* le pourcentage des valeurs non nulles --: 80.0 %
```



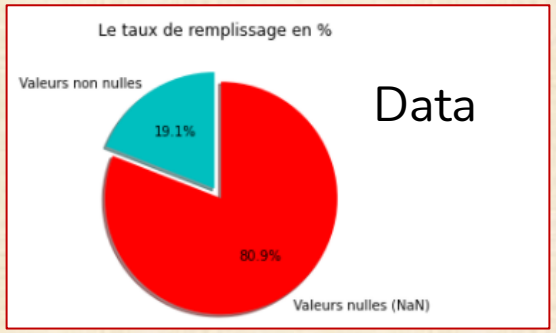
calc_inf(): donnes toutes les informations sur le fichier de données

1

Présentation du jeu données

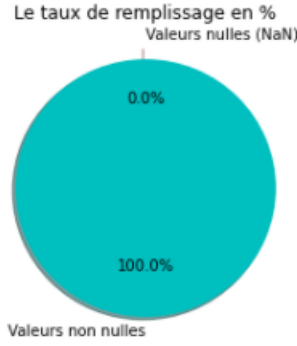
academy

- Le taux de remplissage égale 100% après avoir éliminer les données inutiles
- Concernant la table Data on est à 19.1% de remplissage c'est suffisant pour les indicateurs sélectionnés

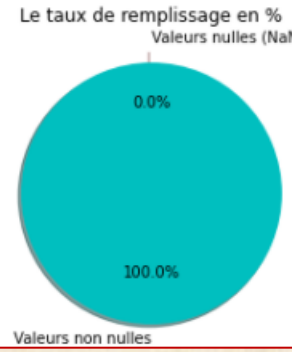



```
calc_inf(FootNote_flt, False)

* Nombre de colonnes sans NaN -------: 4
* Nombre de colonnes NaN ------------: 0
* Nombre de colonnes mixtes----------: 0
* Nombre de lignes ------------------------: 643638
* Nombre de colonnes ----------------------: 4
* Nombre de cases -------------------------: 2574552
* Nombre de valeurs nulles ----------------: 0
* Nombre de valeurs non nulles ------------: 2574552
* le pourcentage des valeurs nulles ------: 0.0 %
* le pourcentage des valeurs non nulles --: 100.0 %
```

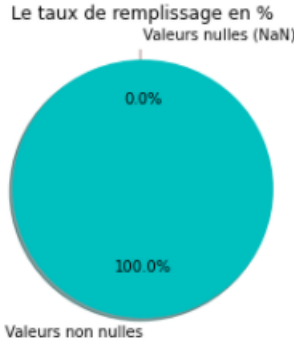

```
calc_inf(data_series,True)

* Nombre de colonnes sans NaN -------: 4
* Nombre de colonnes NaN ------------: 0
* Nombre de colonnes mixtes----------: 0
* Nombre de ligne entièrement nulles : 0
* Nombre de ligne mixtes ------------ : 0
* Nombre de ligne sans NaN ----------: 3665
* Nombre de lignes ------------------------: 3665
* Nombre de colonnes ----------------------: 4
* Nombre de cases -------------------------: 14660
* Nombre de valeurs nulles ----------------: 0
* Nombre de valeurs non nulles ------------: 14660
* le pourcentage des valeurs nulles ------: 0.0 %
* le pourcentage des valeurs non nulles --: 100.0 %
```




```
calc_inf(country_series, True)

* Nombre de colonnes sans NaN -------: 3
* Nombre de colonnes NaN ------------: 0
* Nombre de colonnes mixtes----------: 0
* Nombre de ligne entièrement nulles : 0
* Nombre de ligne mixtes ------------ : 0
* Nombre de ligne sans NaN ----------: 613
* Nombre de lignes ------------------------: 613
* Nombre de colonnes ----------------------: 3
* Nombre de cases -------------------------: 1839
* Nombre de valeurs nulles ----------------: 0
* Nombre de valeurs non nulles ------------: 1839
* le pourcentage des valeurs nulles ------: 0.0 %
* le pourcentage des valeurs non nulles --: 100.0 %
```




```
calc_inf(country,True)

* Nombre de colonnes sans NaN -------: 4
* Nombre de colonnes NaN ------------: 0
* Nombre de colonnes mixtes----------: 0
* Nombre de ligne entièrement nulles : 0
* Nombre de ligne mixtes ------------ : 0
* Nombre de ligne sans NaN ----------: 214
* Nombre de lignes ------------------------: 214
* Nombre de colonnes ----------------------: 4
* Nombre de cases -------------------------: 856
* Nombre de valeurs nulles ----------------: 0
* Nombre de valeurs non nulles ------------: 856
* le pourcentage des valeurs nulles ------: 0.0 %
* le pourcentage des valeurs non nulles --: 100.0 %
```



Sélection des indicateurs

academy

Les indicateurs utiles pour analyser et comparer:

- Le choix de l'année d'indicateurs est sélectionnée en fonction du taux de remplissage ainsi que les données récentes
- Le traitement des indicateurs est fait sur deux méthodes expliquées sur les slides suivants
- La sélection des indicateurs est faite en fonction du besoin du projet

Indicator Code	Description	Année
SP.POP.TOTL	Population, total	2016
IT.NET.USER.P2	Internet users (per 100 people)	2016
UIS.E.3	Enrolment in upper secondary education, both sexes (number)	2011
SE.TER.ENR	Enrolment in tertiary education, all programmes, both sexes (number)	2011
UIS.E.4	Enrolment in post-secondary non-tertiary education, both sexes (number)	2011
BAR.TER.CMPT.25UP.ZS	Barro-Lee: Percentage of population age 25+ with tertiary schooling. Completed Tertiary	2010
NY.GDP.PCAP.CD	GDP per capita (current US$)	2016

Méthode d'analyse et de traitement

academy

Filtrage par inducteur: exemple du premier indicateur population totale **SP.POP.TOTL**

1. Créer une nouvelle table pop_2016 par filtration avec l'indicateur SP.POP.TOTAL et l'année 2016 sélectionné comme illustré en dessous,

```
code_pop_total = ["SP.POP.TOTL"]
pop_tot = countries["Indicator Code"].isin(code_pop_total)
pop_total = countries[pop_tot]
```

```
# population totale par pays et région pour l'années 2016
pop_2016 = pop_total[["Country Name","Country Code","2016", "Region"]].copy()
```

2. Suppression des NaN
3. Refaire le même processus avec les autres indicateurs
4. Merger tous les indicateurs
5. Normalisation des indicateurs
6. Calcul de la moyenne géométrique
7. Classement des pays / régions par ordre

   + présentation graphique(bar, scatter,....)

```
lst_fnl_new
```

	Country Name	Country Code	Region	nb_users_int	nb_bar_ter	2016_gdp	nb_students
0	Afghanistan	AFG	South Asia	3.672058e+06	1.909547e+06	561.778746	917878.0
1	Albania	ALB	Europe & Central Asia	1.908680e+06	2.847340e+04	4124.982390	285011.0

```
for col in ["nb_users_int","nb_bar_ter","2016_gdp","nb_students"]:
    lst_fnl_new[col] = lst_fnl_new[col]/lst_fnl_new[col].sum()
```

```
# Maintenant on calcule la moyenne géométrique,
lst_fnl_new['moyenne'] = stats.gmean(lst_fnl_new.iloc[:, 3:6], axis=1)
lst_fnl_new['moyenne'] = lst_fnl_new['moyenne']/ lst_fnl_new['moyenne'].max()
lst_fnl_new= lst_fnl_new.sort_values(by="moyenne", ascending = False)
lst_fnl_new
```

	Country Name	Country Code	Region	nb_users_int	nb_bar_ter	2016_gdp	nb_students	moyenne
196	United States	USA	North America	0.078926	0.178061	0.038381	0.086860	1.000000
40	China	CHN	East Asia & Pacific	0.235176	0.059668	0.005409	0.197751	0.520140

Méthode 1

Méthode d'analyse et de traitement

academy

Filtrage par inducteur: exemple du premier indicateur population totale **SP.POP.TOTL**

1. Via la table pop_2016 crée dans la méthode 1, on calcule les ordres de grandeur statistiques
2. Fixer le seuil pour inducteur SP.POP.TOTL, puis prendre la population au-dessus de la médiane.
3. Suppression des NaN
4. Refaire le même processus pour le reste d'indicateurs
5. Merger tous les indicateurs
6. Normalisation des indicateurs
7. Calcul de la moyenne géométrique
8. Classement des pays / régions par ordre

+ présentation graphique(bar, scatter,....)

```
pop_2016.describe()
```

	2016
count	2.060000e+02
mean	3.585180e+07
std	1.378209e+08
min	1.109700e+04
25%	9.992810e+05
50%	6.741830e+06
75%	2.470270e+07
max	1.378665e+09

```
# le seuil choisi représente la median de la population totale
seuil_pop = float(pop_2016.median())
# comparaison au seuil median de la population tot 2016.
pop_2016_flt = pop_2016[pop_2016["2016"] > seuil_pop]
# mettre en ordre la pop_tot 2016 par pays.
pop_2016_flt = pop_2016_flt.sort_values(by=['2016'])
```

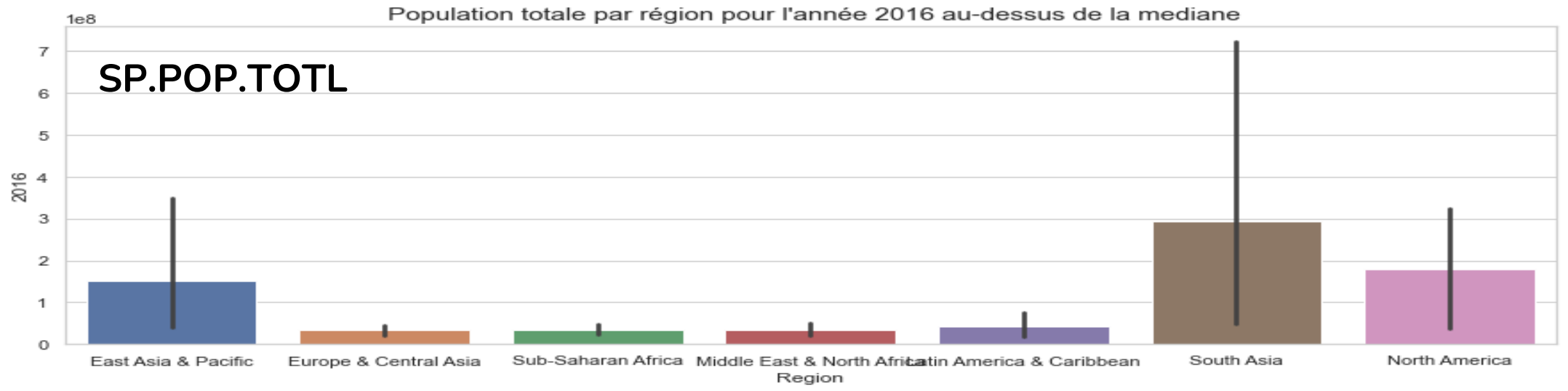
```
for col in ["nb_users_int","nb_bar_ter","2016_gdp", "nb_students" ]:
    lst_fnl_flt[col] = lst_fnl_flt[col]/lst_fnl_flt[col].sum()

lst_fnl_flt['moyenne'] = stats.gmean(lst_fnl_flt.iloc[:, 3:6], axis=1)
lst_fnl_flt['moyenne'] = lst_fnl_flt['moyenne']/ lst_fnl_flt['moyenne'].max()
lst_fnl_flt= lst_fnl_flt.sort_values(by="moyenne", ascending = False)
lst_fnl_flt
```

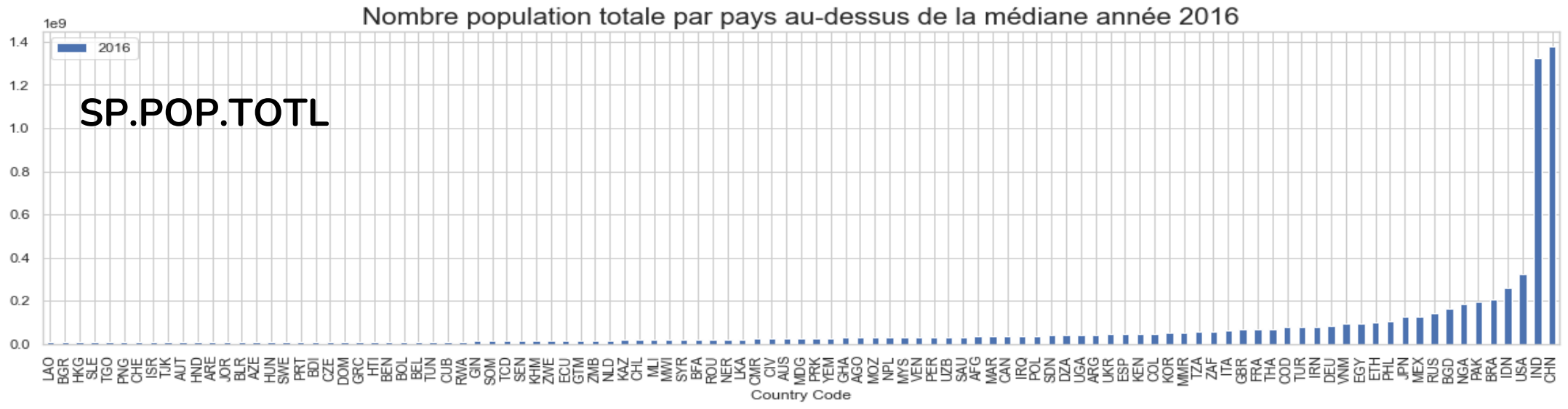
	Country Name	Country Code	Region	nb_users_int	nb_bar_ter	2016_gdp	nb_students	moyenne
30	United States	USA	North America	0.195643	0.312033	0.062228	0.257129	1.000000
16	Japan	JPN	East Asia & Pacific	0.092863	0.079034	0.041998	0.058357	0.432925
10	Germany	DEU	Europe & Central Asia	0.058903	0.041488	0.045420	0.020986	0.308002
23	Russian Federation	RUS	Europe & Central Asia	0.087661	0.116546	0.009445	0.090516	0.293958

Méthode 2

Comparaison par l'indicateur SP.POP.TOTL

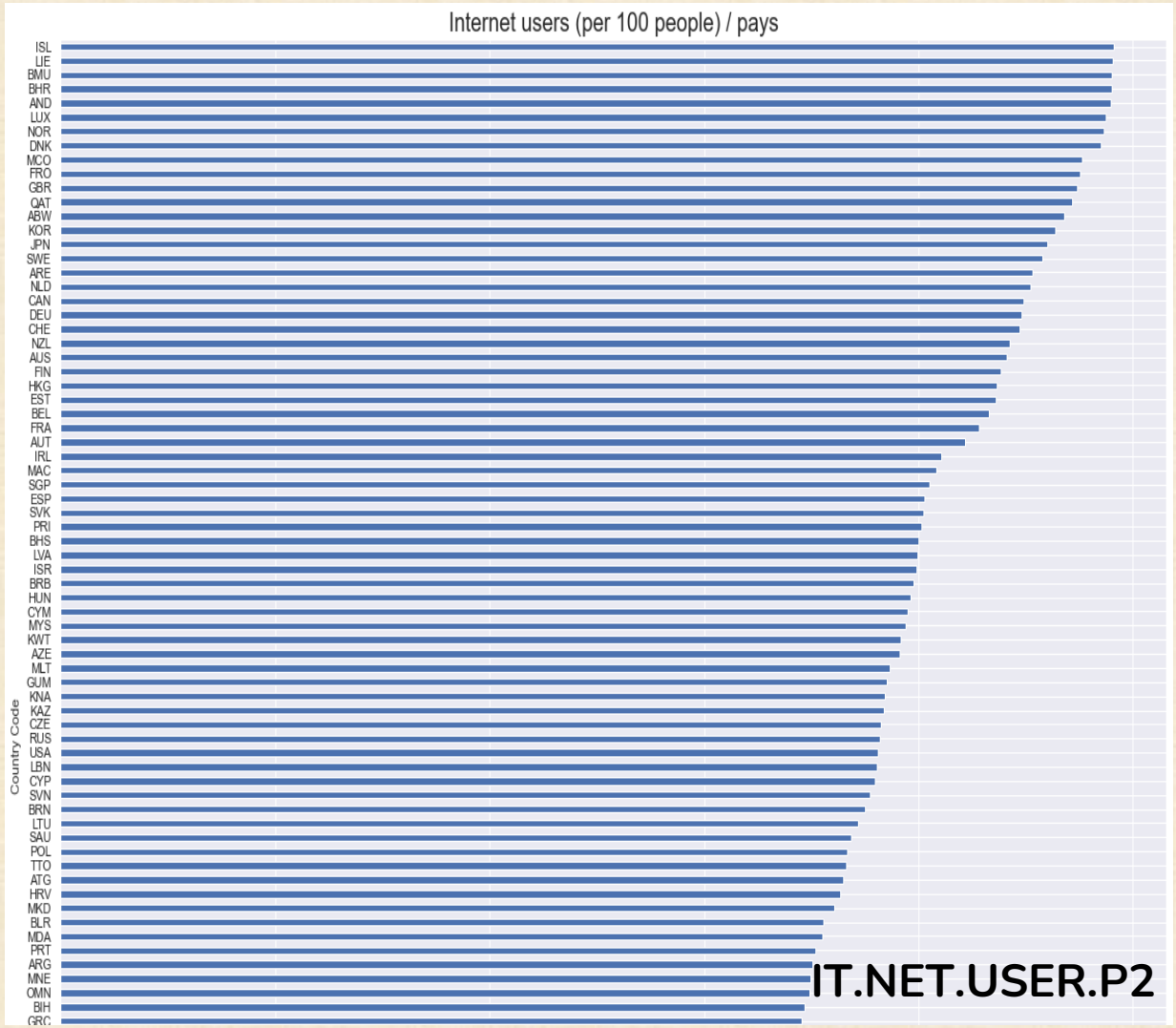
academy



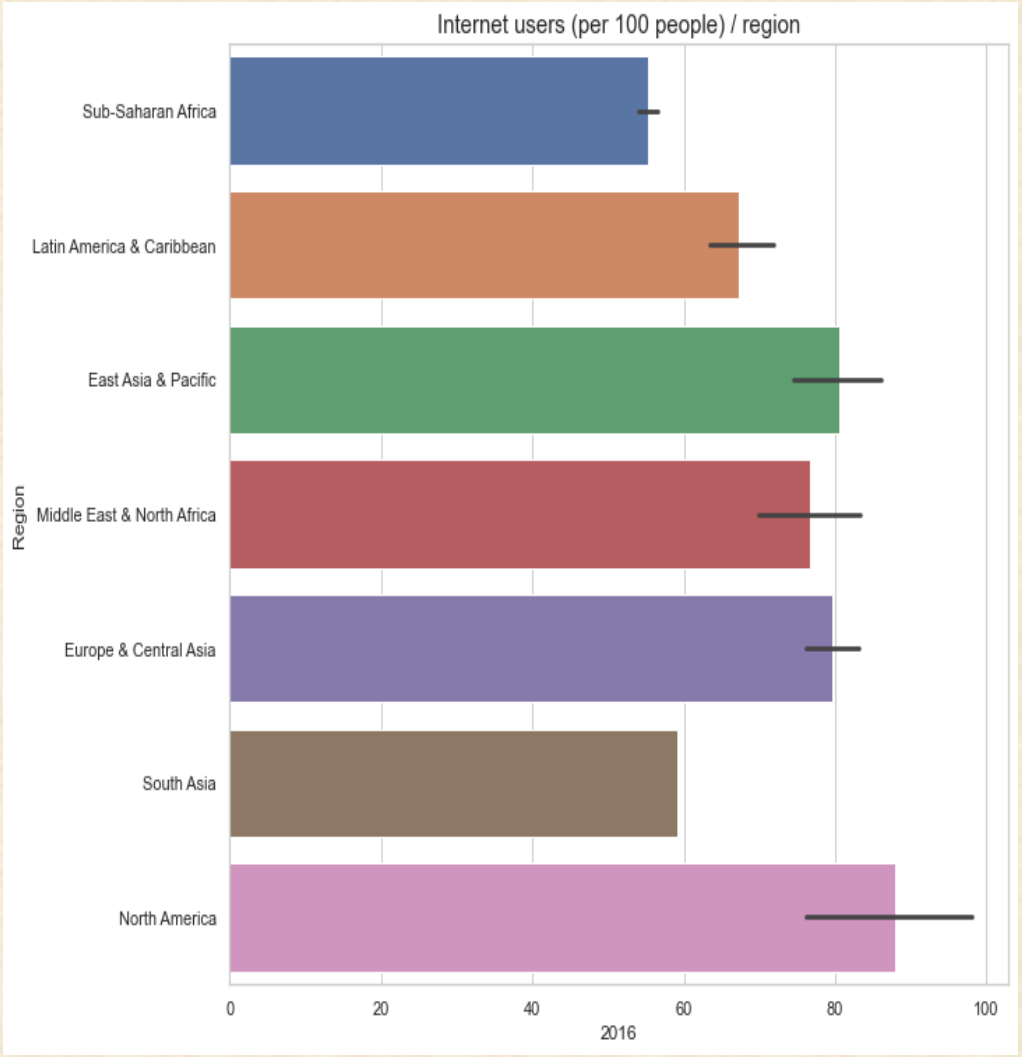


Comparaison par l'indicateur IT.NET.USER.P2

academy



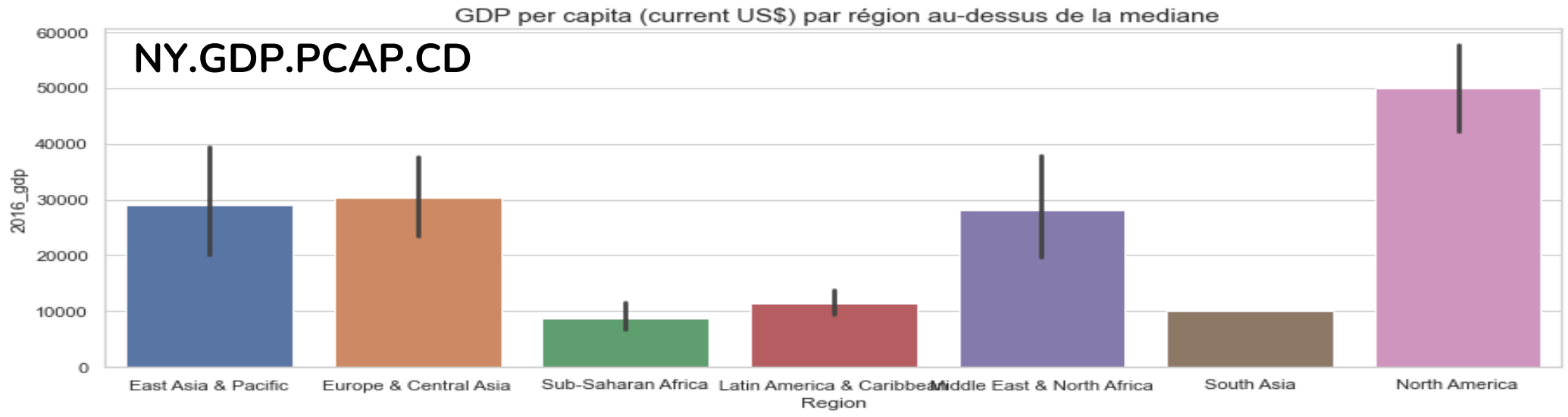
IT.NET.USER.P2

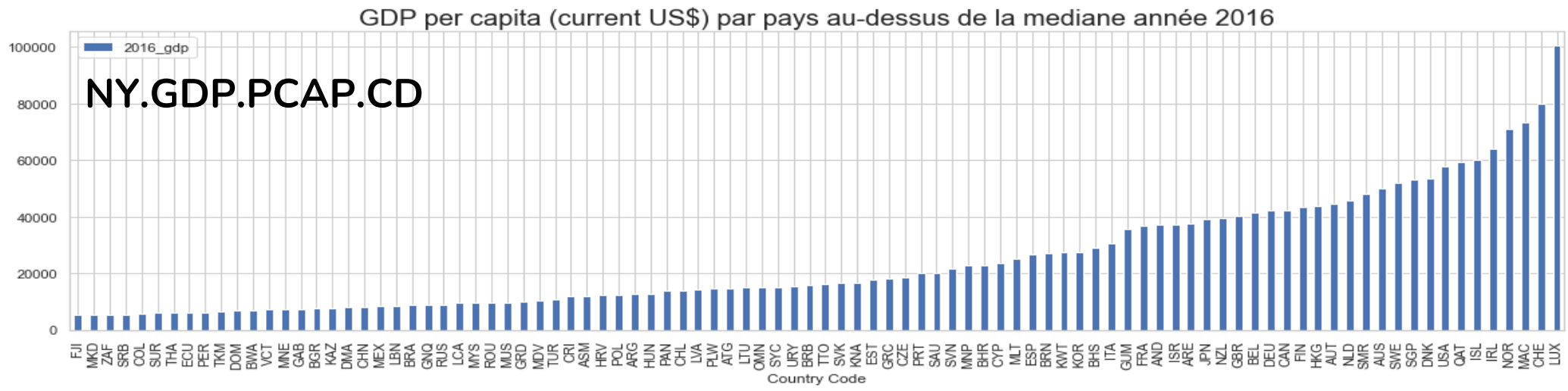


IT.NET.USER.P2 par pays/régions année 2016

Comparaison par l'indicateur NY.GDP.PCAP.CD

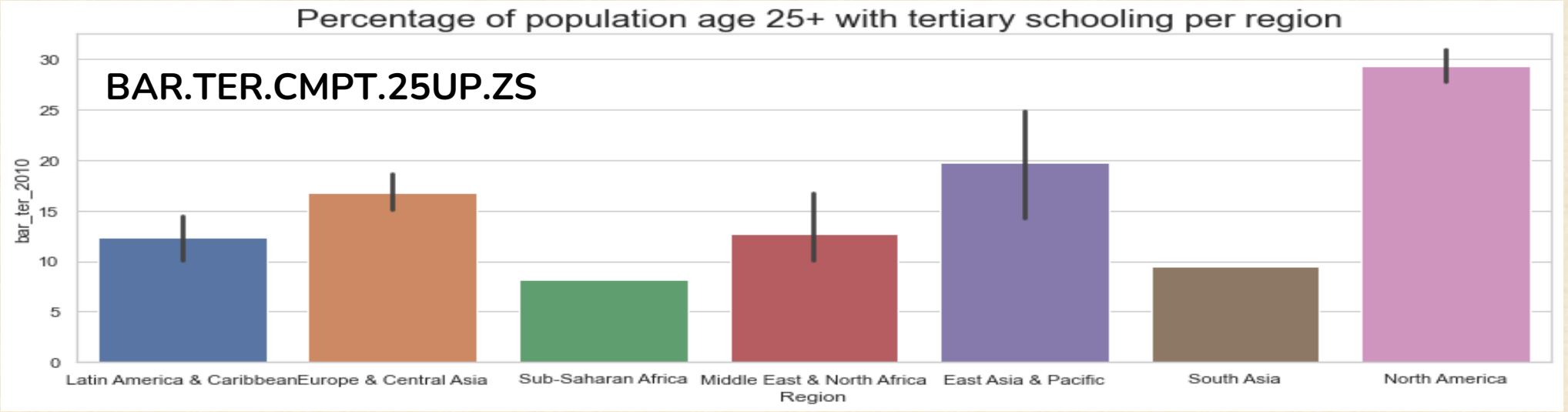
academy

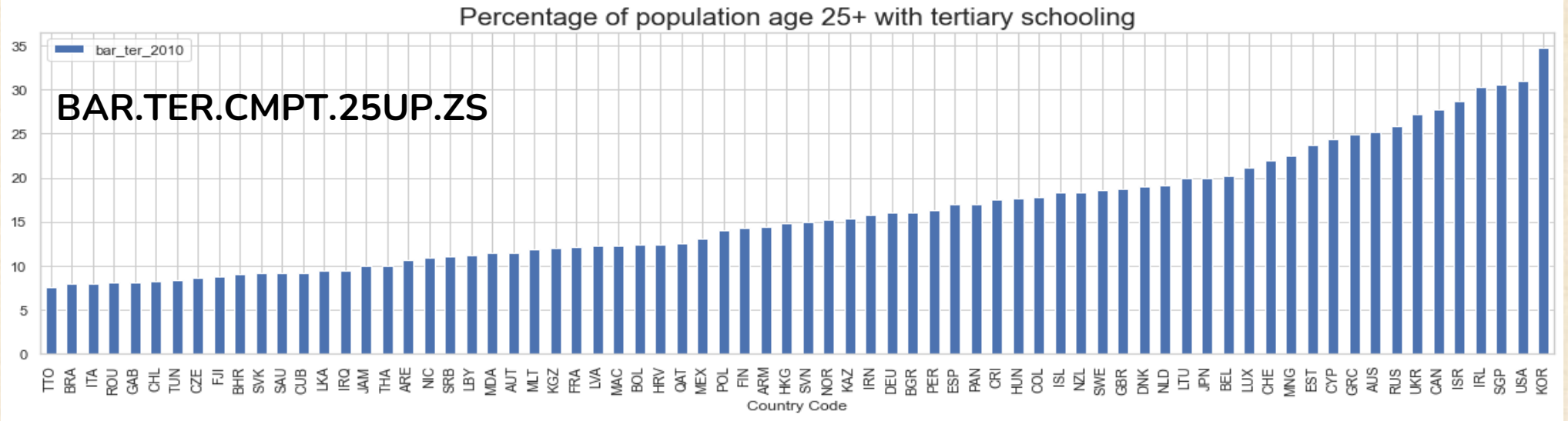




Comparaison par l'indicateur BAR.TER.CMPT.25UP.ZS

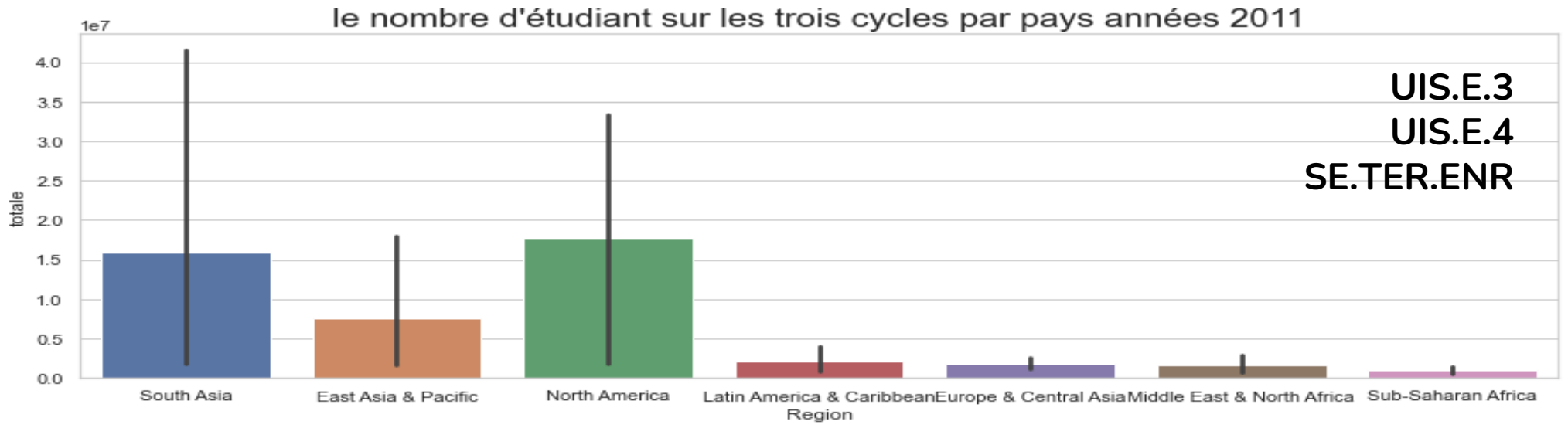
academy





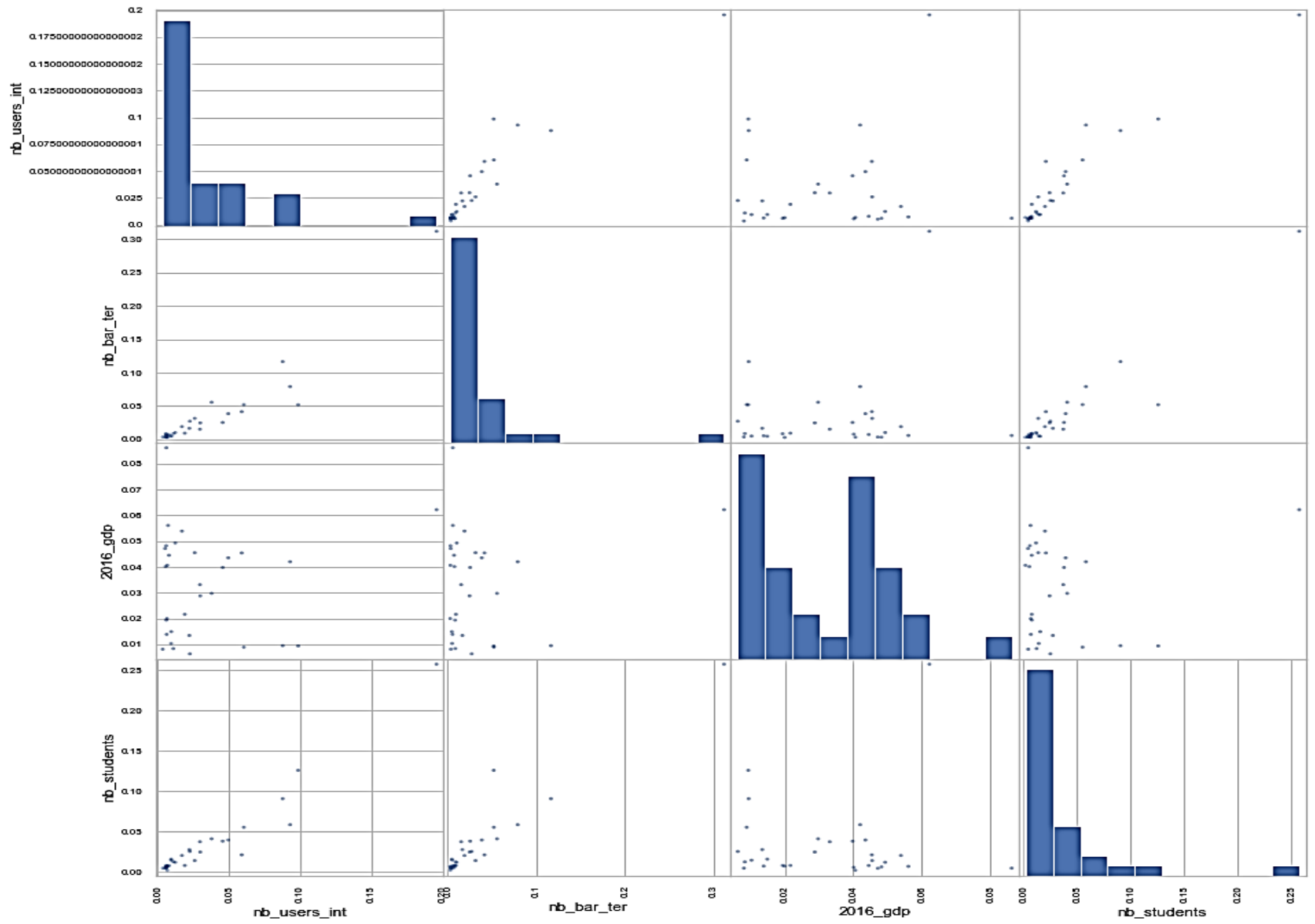
Comparaison par la somme des 3 indicateurs

academy

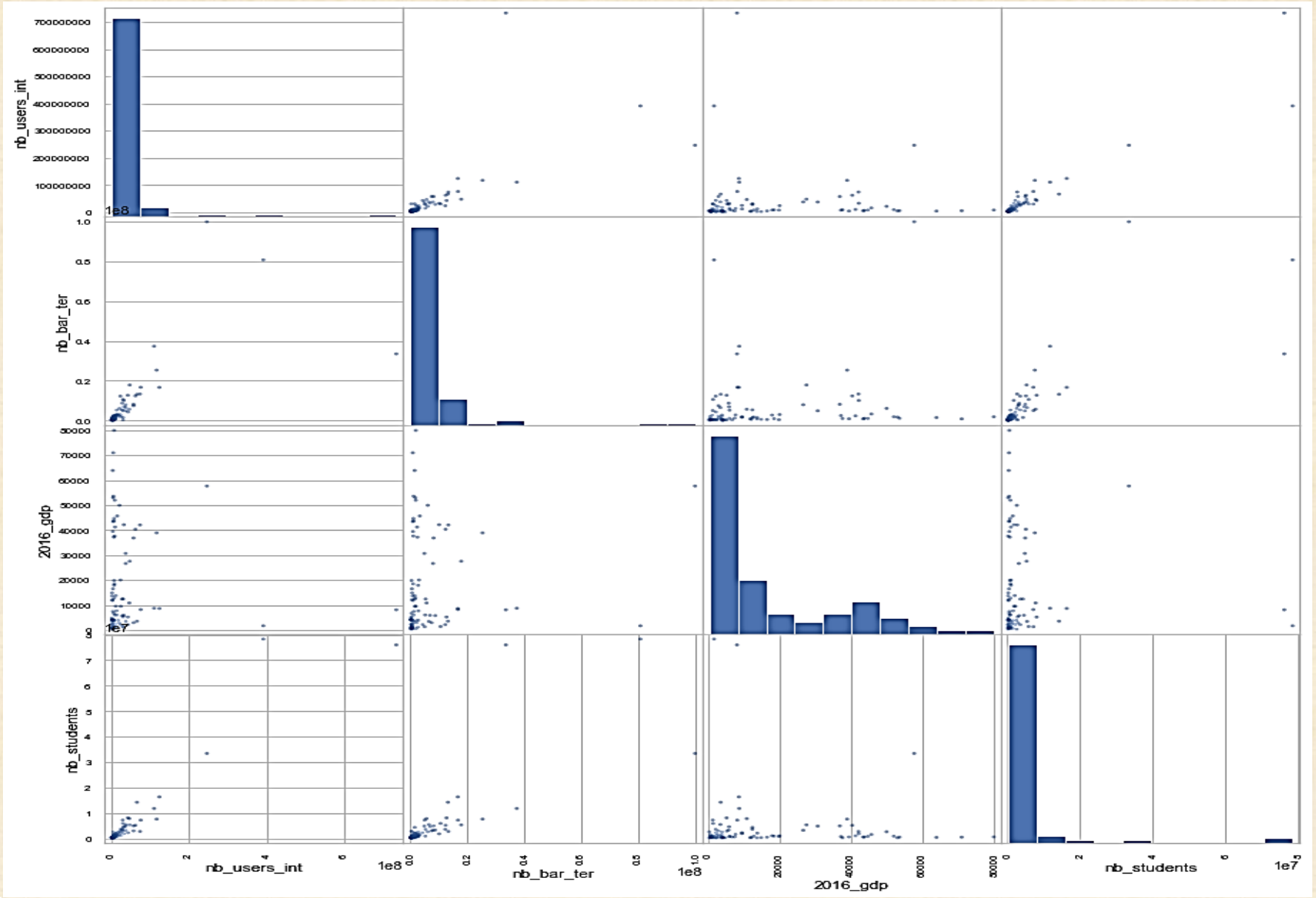


Représentation graphique des indicateurs filtrés / 31 pays

academy

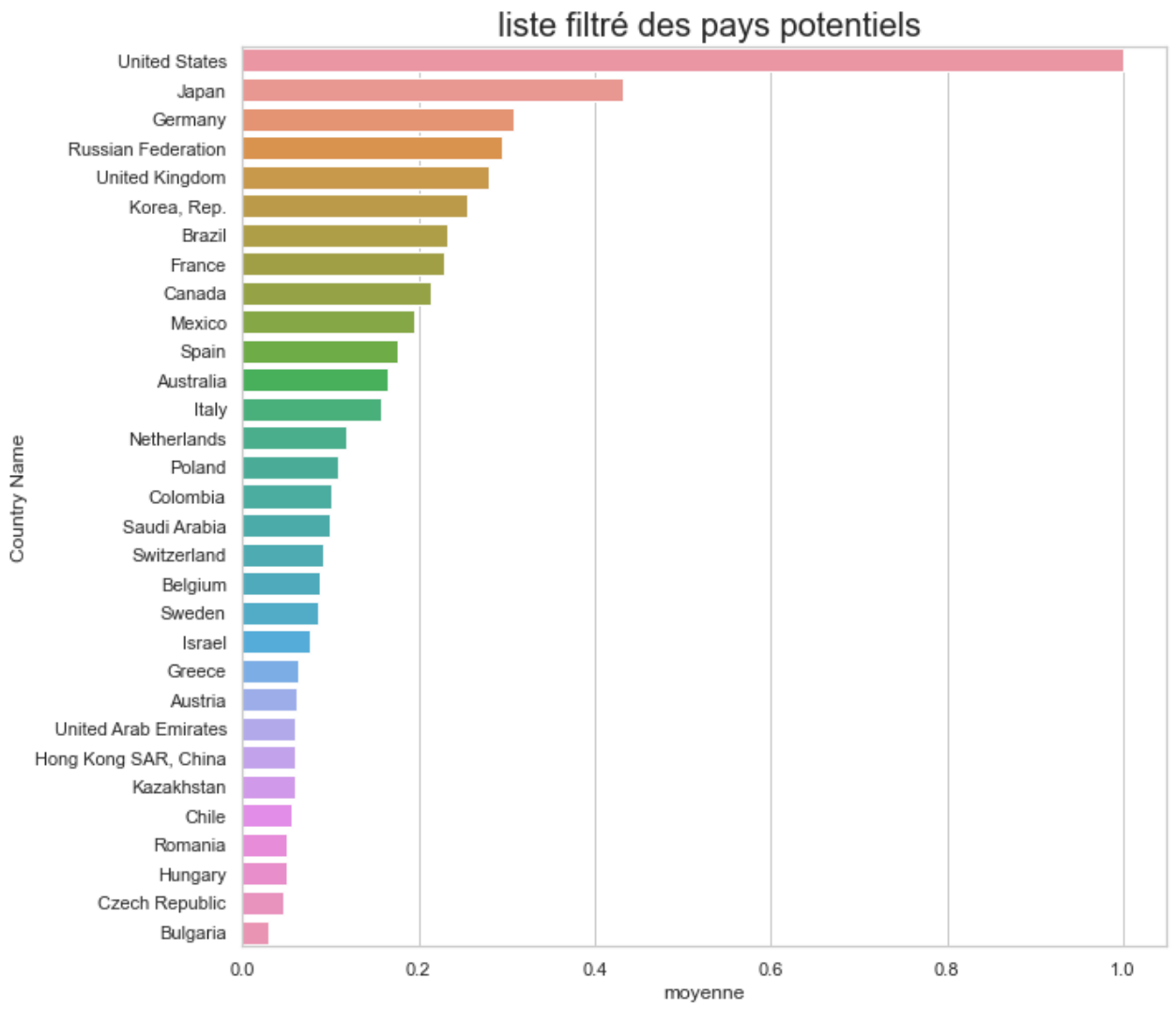


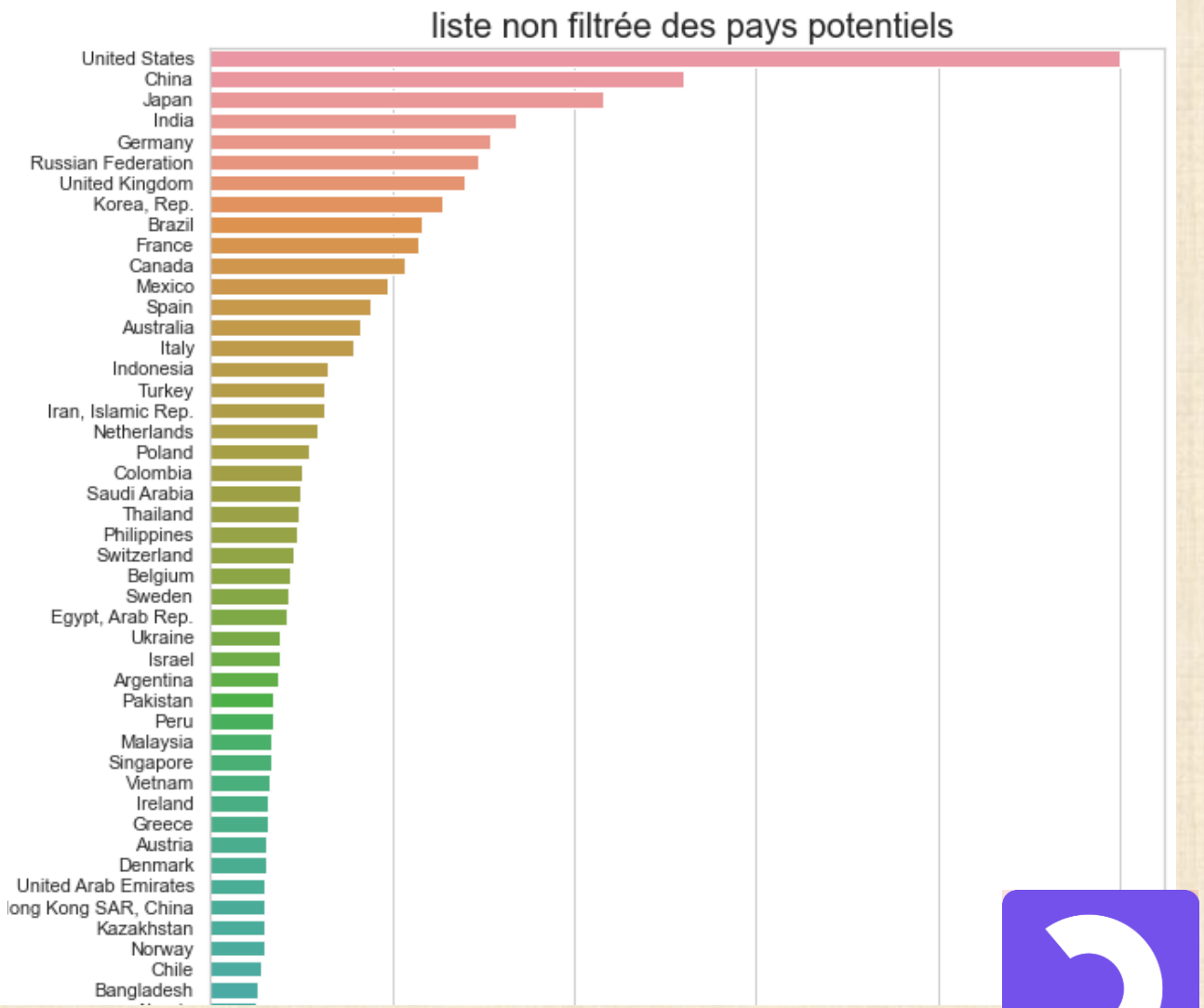
Représentation graphique des indicateurs non filtrés / 90 pays



Sélection des pays / régions potentiels

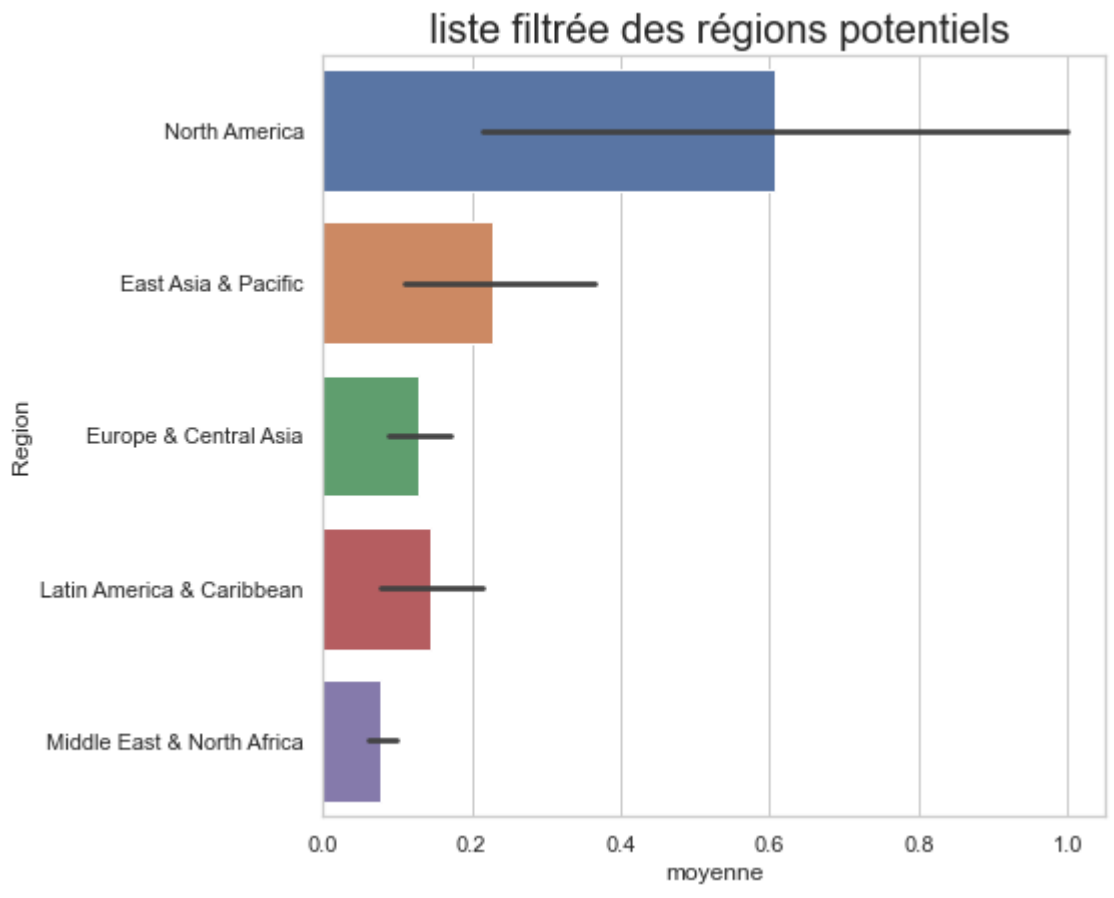
academy

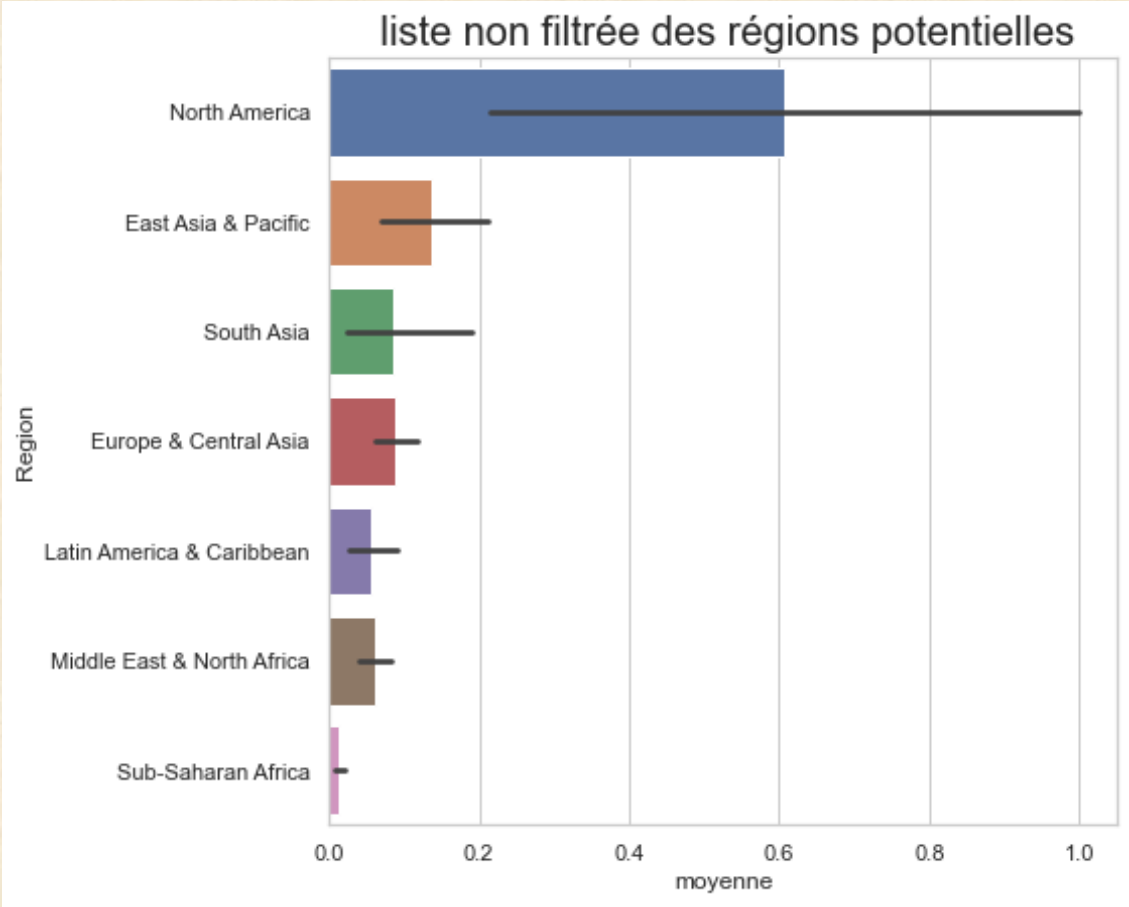




Sélection des pays / régions potentiels

academy





Conclusion

academy

La pertinence du jeu de données:

- Contient des informations utiles pour répondre à la problématique de l'entreprise
- Tous les pays du monde sont abordés
- La traçabilité des sources de données
- Contient des données détaillées sur l'éducation pour chaque pays / régions

Aspect négatif:

- La répartition de NaN distincte sur l'ensemble des indicateurs
- Le manque d'infos sur les entreprises de formations privées (concurrence)