PROJET 2- ANALYSEZ DES DONNEES DE SYSTEMES EDUCATIFS

1-PROBLEMATIQUE

La start-up de la EdTech, nommée **academy**, propose des contenus de formation en ligne pour un public de niveau lycée et université.

Dans le cadre d'un projet d'expansion à l'international de l'entreprise,
une première mission d'analyse exploratoire est initiée, afin de déterminer si les données sur l'éducation de la banque mondiale permettent d'informer le projet d'expansion.

Ci-dessous les différentes questions à explorer :

- 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 ?

2-PRESENTATION DU JEU DU JEU DONNEES

1-Sources:

https://datacatalog.worldbank.org/dataset/education-statistics

http://datatopics.worldbank.org/education/

2-Fichiers: .

EdStatsData:

Est le fichier de base :

nous renseigne sur les codes pays, les codes series, les relevés des taux(réels ou estimés et prospectifs) sur une periode allant de 1970 à 2100.

nombfre de lignes: 886930
nombre de colonnes: 70
nombre de données dupliquées: 0

EdStatsCountry:

Contient des informations sur les pays, les regions et leur niveau de revenu.

nombfre de lignes: 241
nombre de colonnes: 32
nombre de données dupliquées:0

EdStatsCountry-Series:

contient des informations sur les code pays , les codes series et leur description.

nombfre de lignes: 613
nombre de colonnes: 4
nombre de données dupliquées: 0

EdStatsFootNote:

Contient des informations sur les pays, les series code ainsi la description des années de collecte de données.

nombfre de lignes: 643638
nombre de colonnes: 5
nombre de données dupliquées:0

EdStatsSerie:

Fournit des informations sur les codes series ainsi que la composition des populations.

nombfre de lignes: 3665
nombre de colonnes: 21
nombre de données dupliquées: 0

3-ANALYSE PRE-EXPLORATOIRE DU JEU DE DONNEES 1

- 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)

5- SELECTION DES VARIABLES

A-Nous selectionnons variables utiles et créons de nouveaux fichiers :

```
# 1-Nouveau fichier EdStatsDat
new_EdStatsData_df=EdStatsData_df[['Country Name', 'Country Code', 'Indicator Name',
'Indicator Code','1999', '2000', '2001', '2002', '2003', '2004', '2005',
'2006', '2007', '2008', '2009', '2010', '2011', '2012', '2013', '2014','2015',]]
# 2-Nouveau fichier EdStatsCountry
new_EdStatsCountry_df=EdStatsCountry_df[['Country Code', 'Short Name', 'Table Name',
'Long Name', '2-alpha code',
'Currency Unit', 'Special Notes', 'Region', 'Income Group', 'WB-2 code']]
# 3-Nouveau fichier EdStatsCountry
new_EdStatsFootNote_df=EdStatsFootNote_df[['CountryCode','SeriesCode', 'Year',
'DESCRIPTION']]
# 4-Nouveau fichier EdStatsCountry
new_EdStatsCountrySeries_df=EdStatsCountrySeries_df[['CountryCode', 'SeriesCode',
'DESCRIPTION']]
# 2-Nouveau fichier EdStatsCountry
```

```
new_EdStatsSeries_df=EdStatsSeries_df[['Series Code', 'Topic', 'Indicator Name',
'Short definition',
'Long definition', 'Unit of measure', 'Development relevance',
'Related indicators',]]
```

B-Ensuite nous procedons a la concatenation des fichiers

Nous retenons au final les fichiers suivants :

new_EdStatsData_df,

new_EdStatsSeries_df

que nous allons concatener par la suite

Les variables retenues pour l instant sont donc:

['Country Name', 'Country Code', 'Indicator Name_x', 'Indicator Code',

'1999', '2000', '2001', '2002', '2003', '2004', '2005', '2006', '2007',

'2008', '2009', '2010', '2011', '2012', '2013', '2014', '2015',

'Series Code', 'Topic',]

4- SELECTIONS DES INDICATEURS ET EVOLUTION

A-

Compte tenu du contexte les pays elligibles doivent:

*posseder des infrastructures internet en priorité pour avoir accès aux services assurés uniquement en ligne.

*avoir une population conséquente en nombre

*avoir une population elligible pour l'accès au lycée et université et ayant des moyens pour payer le service proposé.

Nous rechercherons donc des mots cles relatifs et d'autres critères au fur et à mesure de notre progression dans l'étude.

Ci-dessous les mots cles exploités en priorité :

-internet

-population

-secondary

-tertiary

-Income

et les indicateurs associés selectionnés:

IT.NET.USER.P2: Internet users (per 100 people) *Internautes (pour 100 personnes)*

SP.POP.1524.TO.UN: Population, ages 15-24, total

SP.POP.TOTL: Population, total

SP.POP.GROW: Population growth (annual %)

Croissance démographique (% annuel)

NY.GDP.PCAP.CD: gross domestic product gdp

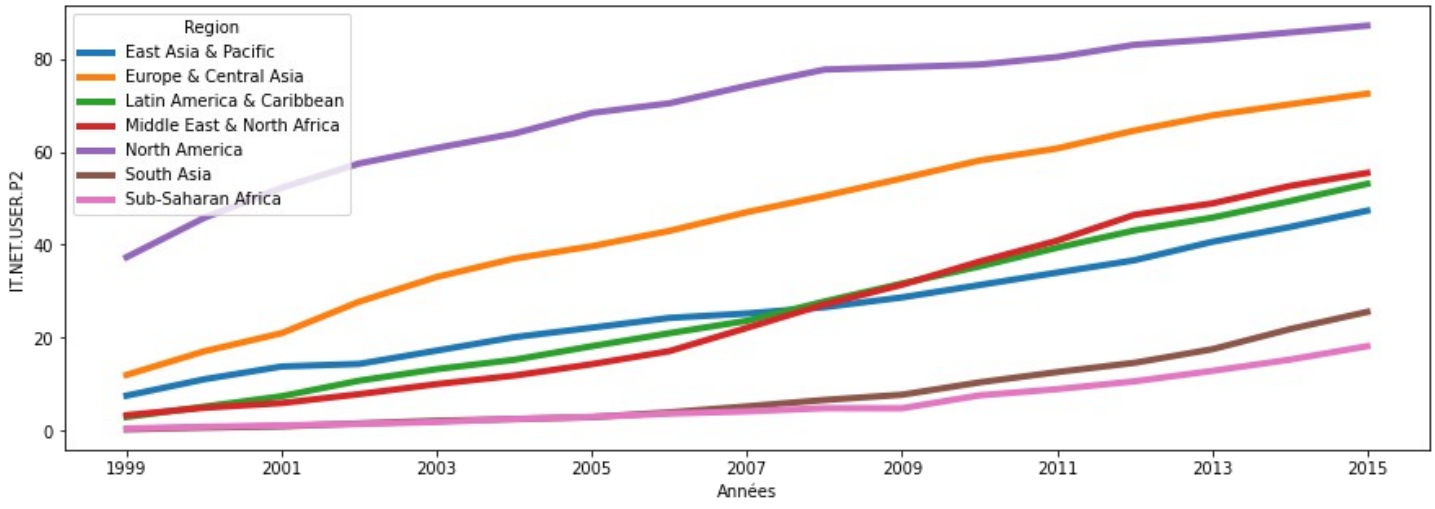
produit intérieur brut PIB

SE.SEC.ENRR Gross enrolment ratio, secondary, both sexes

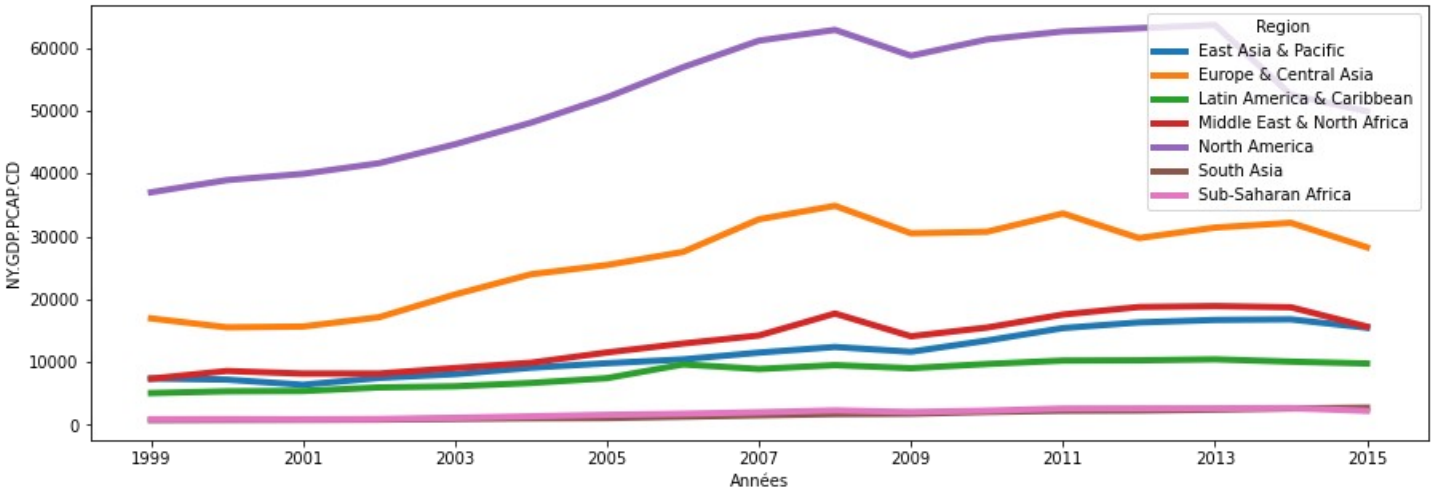
(%)*'Taux brut de scolarisation, secondaire, les deux sexes'*

SE.TER.ENRR Gross enrolment ratio, tertiary, both sexes (%) Taux brut de scolarisation, niveau supérieur, les deux sexes

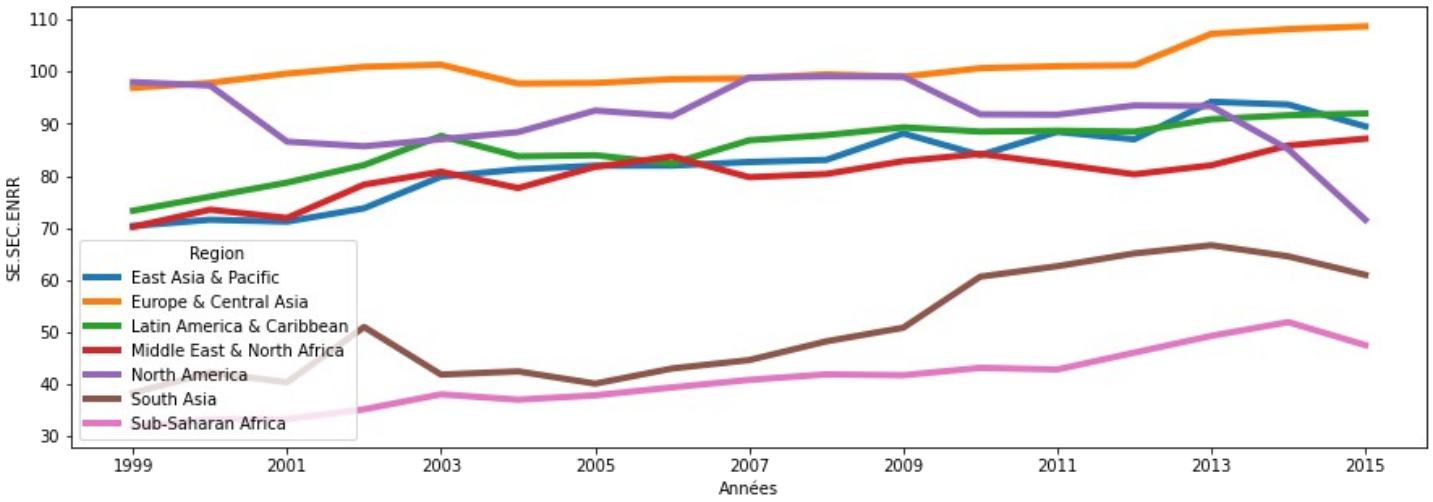
B-Evolution des indicateurs 1999-2015



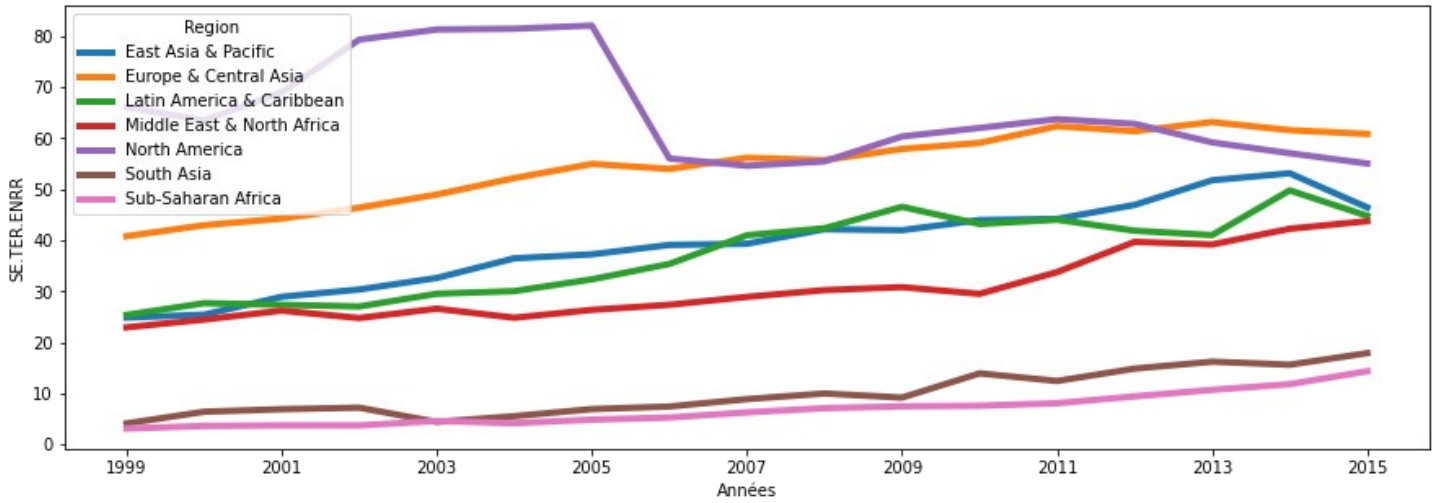
Sur ce graphique les régions 'Nord America' et 'Europe & Central Asia' sont nettement en tçete suivi par 3 autres regions. Ces 2 régions sont dignes d interêt en priorite



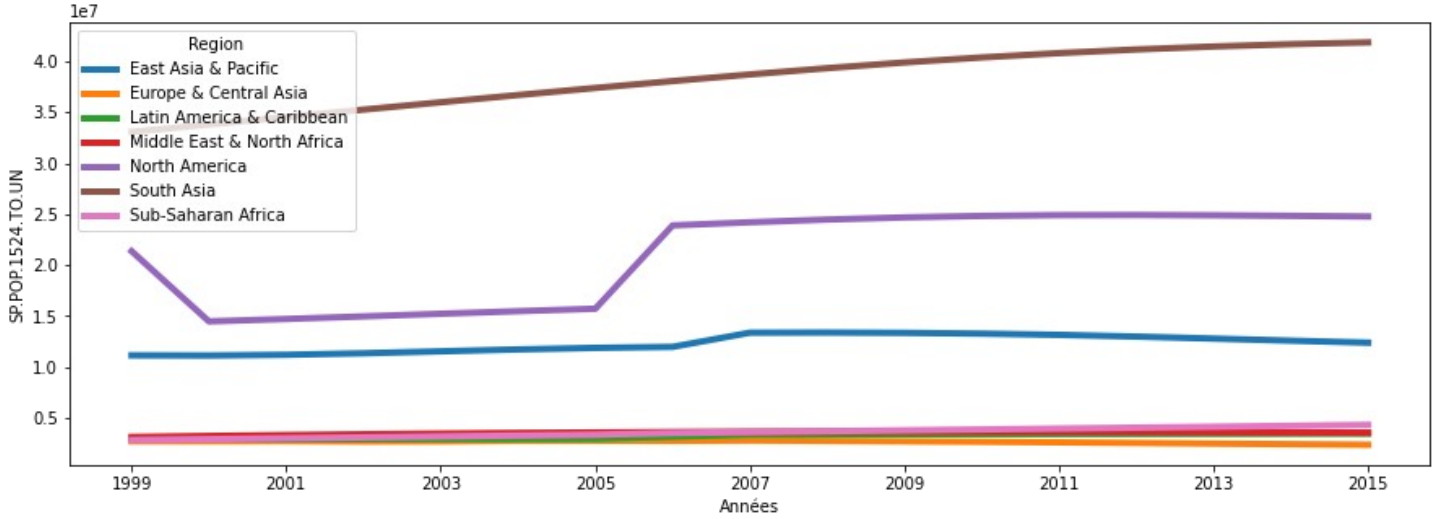
En ce qui concerne le PIB on trouve en tête les mêmes pays que précédemment. Ces pays possèdent un réel potentiel.



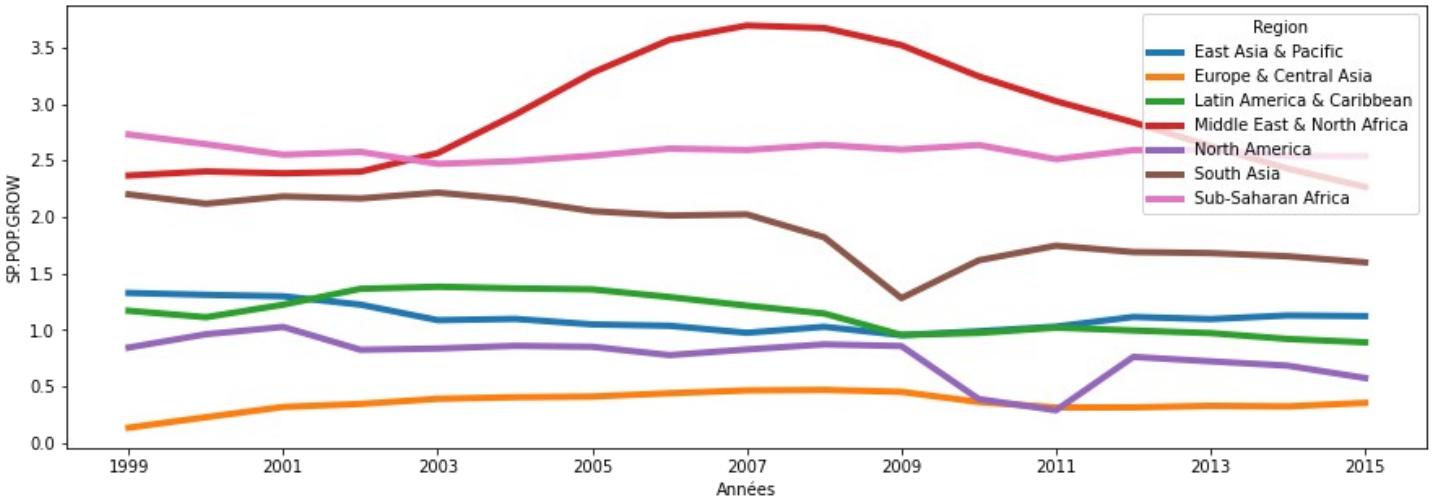
Pour le taux d inscription en secondaire la région 'Europe & Central Asia' se détache nettement, suivi de 3 régions : , 'Latin American & Caribbean','Nord America', 'Middle East & North Africa'.



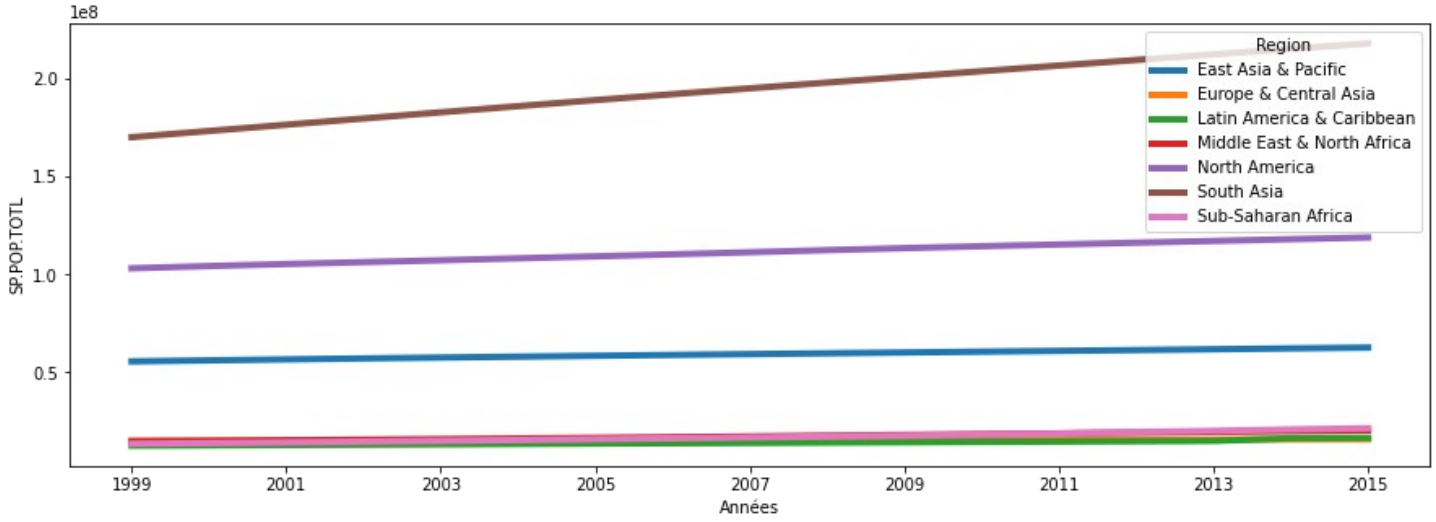
Taux d' inscription en tertiaire en tête : 'Europe & Central Asia' suivi de la région 'North America'



Evotion popûlation 15-24 : la région 'South Asia' domine suivi des reions 'Nord America' et 'East Asia & Pacific'



Nous avons successivement :' Sub Saharian Africa' ,'Middle East &North Africa' ,'South Asia'



Population totale : 'South Asia', 'North America', 'East Asia & Pacific' sont en tête.

Conclusion :

Si on privilégie les indicateurs

IT.NET.USER.P2

NY.GDP.PCAP.CD

SP.POP.1524.TO.UN

3 regions dans ce cas présentent beaucoup d'interêt :

'North America'

'Europe & Central Asia'

'Middle East & North Africa'.

6- ANALYSE EXPLORATOIRE DU JEU DE DONNEES 2

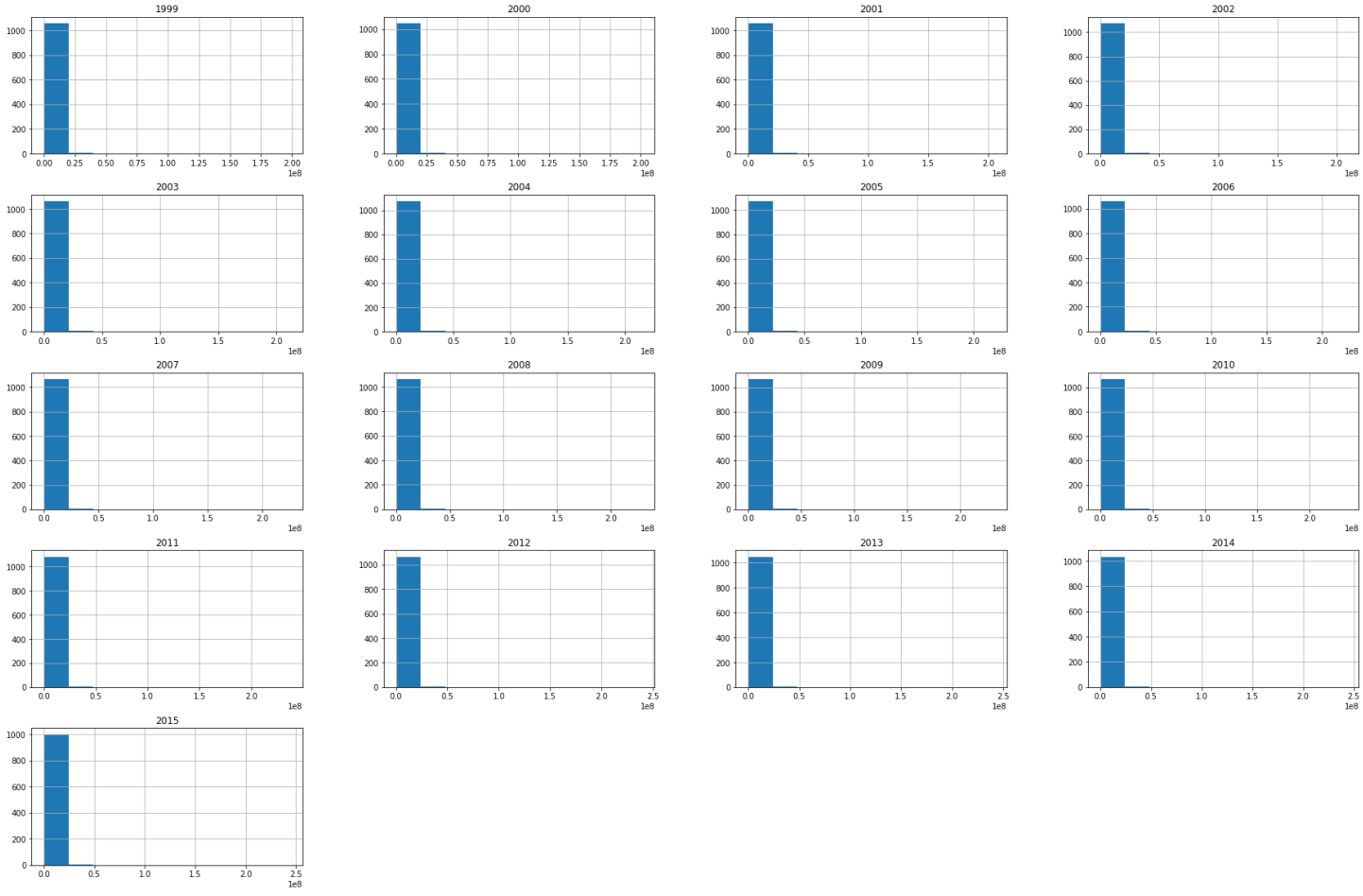
A-Variable Quantitative

 # Tracés correspondants a toutes les années: 1999-2015

```
# HISTOGRAMME SUR DONNEES NON NORMALISEES
```

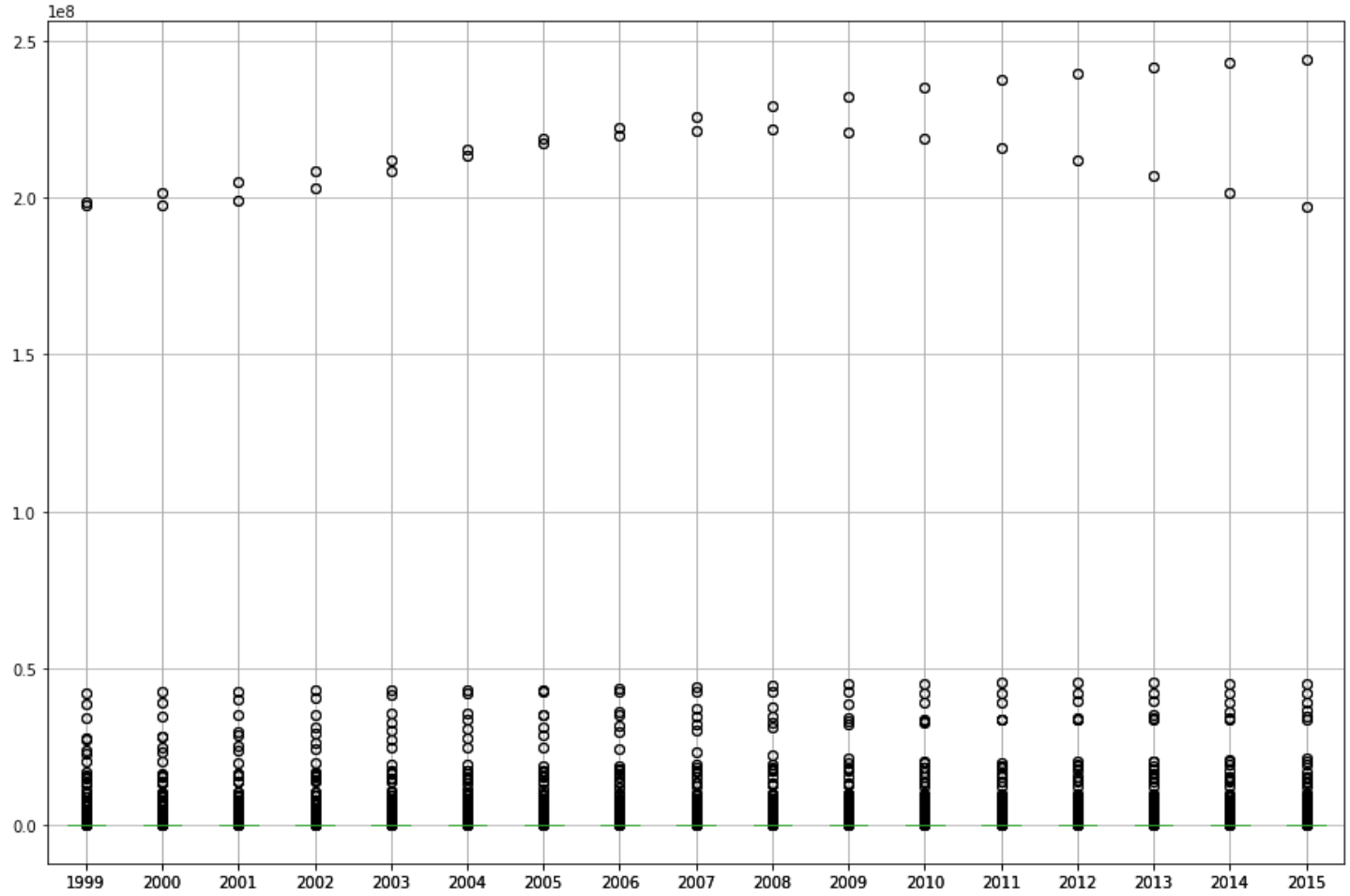
EdStats_final.hist(figsize=(30.5,20),bins=10)

plt.show()




```
#BOXPLOT SUR DONNEES NON NORMALISEES

EdStats_final.boxplot(figsize=(15,10))
EdStats_final.boxplot()
```

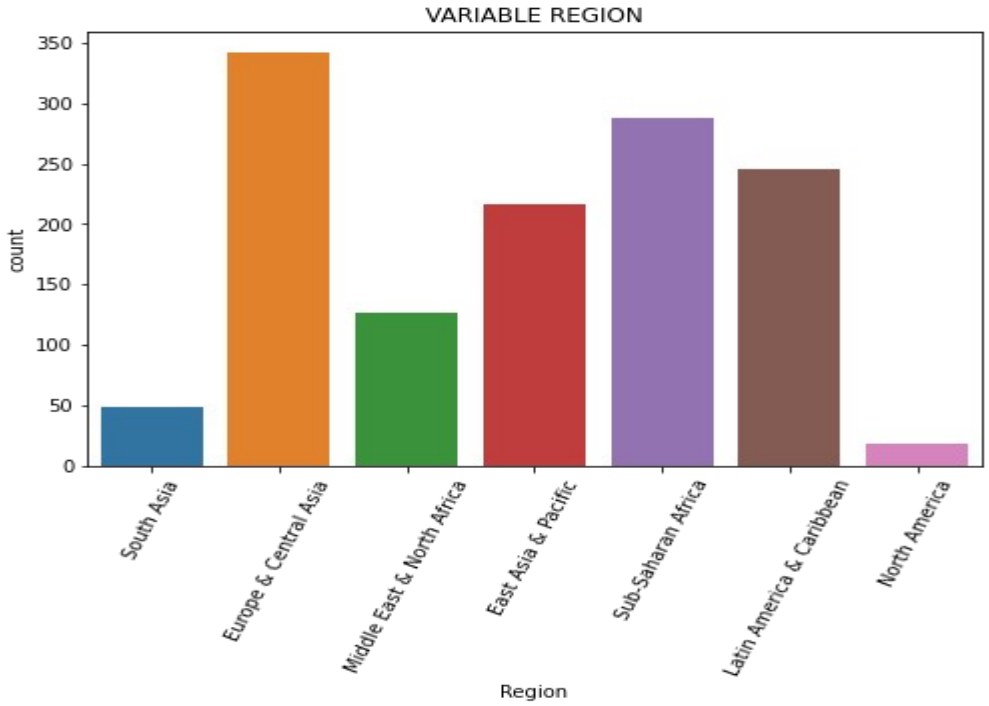


STATISTIQUE :CAS DE LA FRANCE

	Country Name	Indicator Code	MOY	MEDIAN	VAR	STD	SKEW	KURT
338555	France	IT.NET.USER.P2	5.53e+01	6.07e+01	5.70e+02	1.15e+02	4.18	18.93
338425	France	NY.GDP.PCAP.CD	3.56e+04	3.65e+04	5.38e+07	1.17e+07	4.28	19.74
338515	France	SE.SEC.ENRR	1.10e+02	1.11e+02	2.35e+00	2.42e+01	-2.95	3.57
338519	France	SE.TER.ENRR	5.65e+01	5.52e+01	8.35e+00	1.11e+01	-2.68	3.42
339662	France	SP.POP.1524.TO.UN	7.67e+06	7.68e+06	5.44e+09	1.18e+09	4.28	19.75
339487	France	SP.POP.GROW	5.98e-01	5.78e-01	1.06e-02	1.62e-01	-1.55	15.48

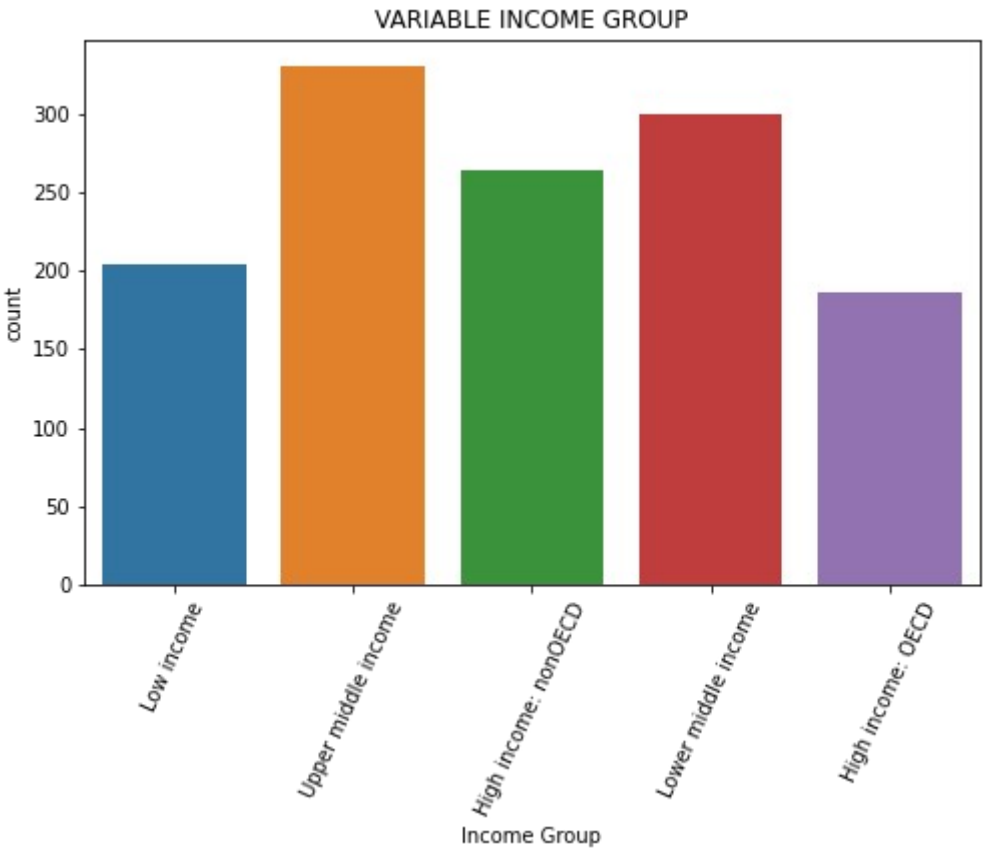
B-Variable Qualitative

rEGION

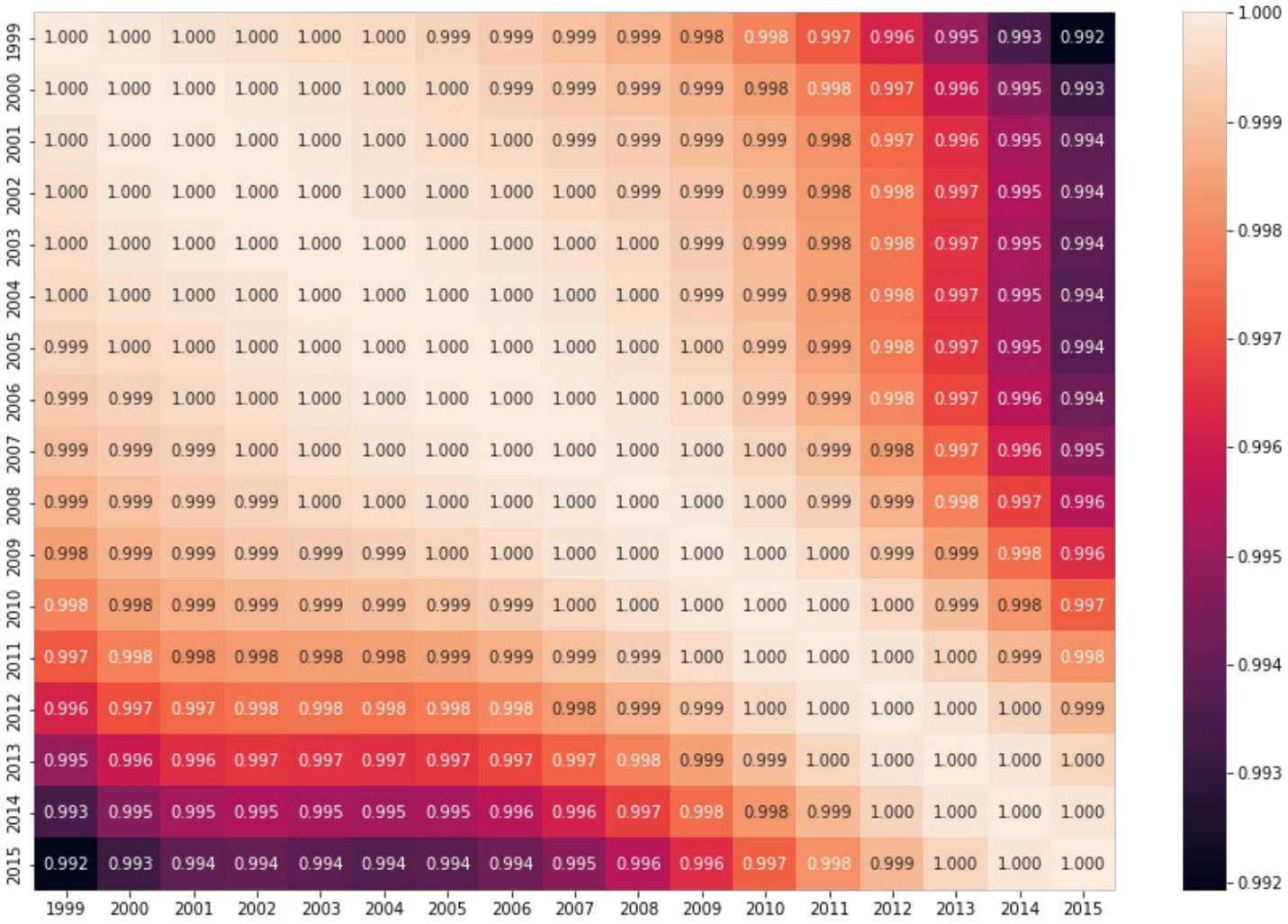


Domination des regions [Europe Asie Centrale] et [Afrique Sub Saharienne]

```
#Graphique Income Group
plt.figure(figsize=(8,5))
sns.countplot(x="Income Group", data=EdStats_final)
plt.title('VARIABLE INCOME GROUP')
plt.xticks(rotation=65)
plt.show
```



7-correlation



Les variables sont très correlée entre elles ;

SCORING ET RECOMMANDATIONS

Pour le scoring nous allons attribuer un coefficient à chaque indicateur selon le tableau suivant :

INDICATEURS AVEC LEUR PONDERATION

 # SE.SEC.ENRR:3

 # SE.TER.ENRR:3

 # NY.GDP.PCAP.CD:5

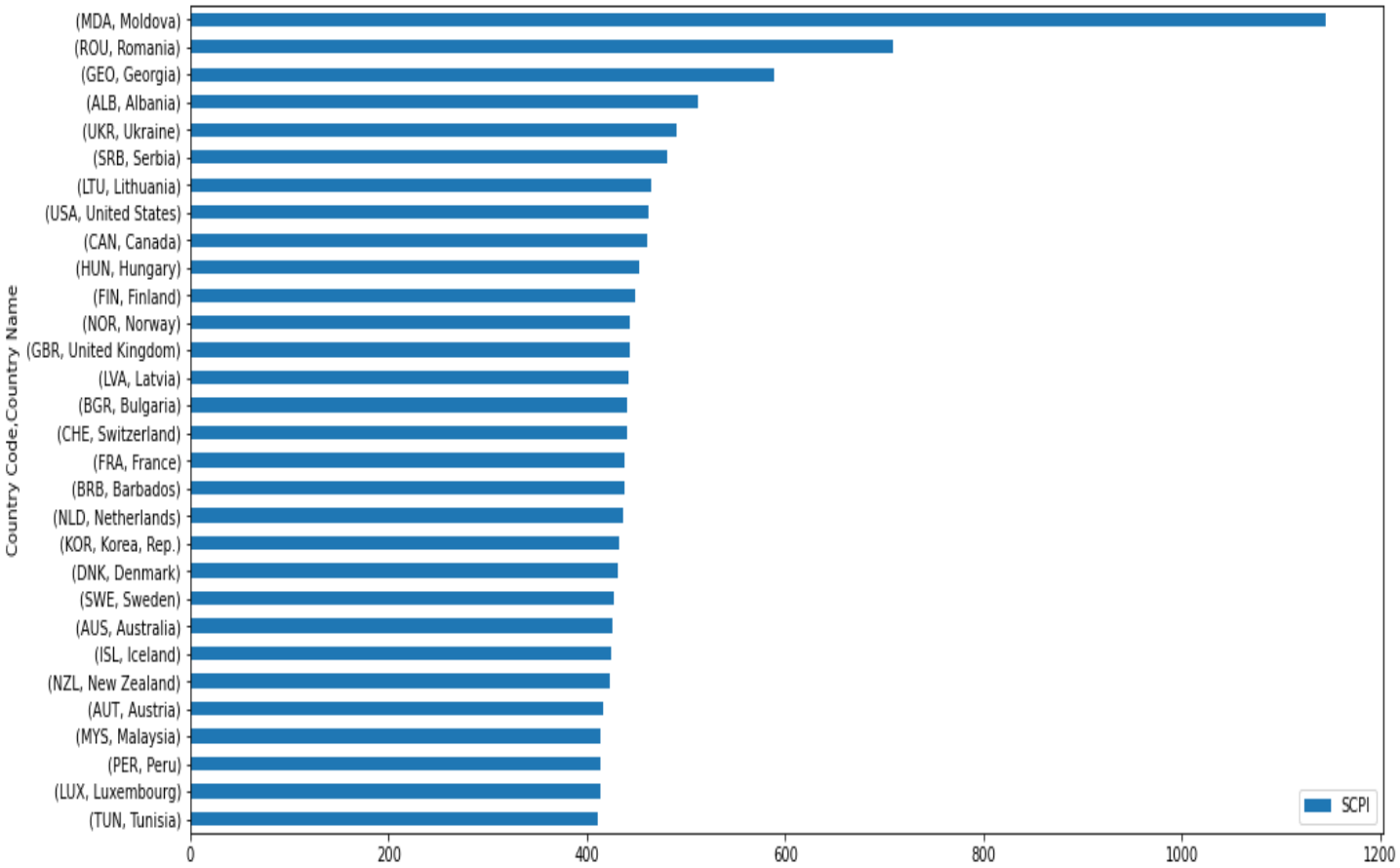
 # IT.NET.USER.P2:8

 # SP.POP.GROW':4

 # SP.POP.1524.TO.UN':8

Classement des pays par rapport SCORE Final

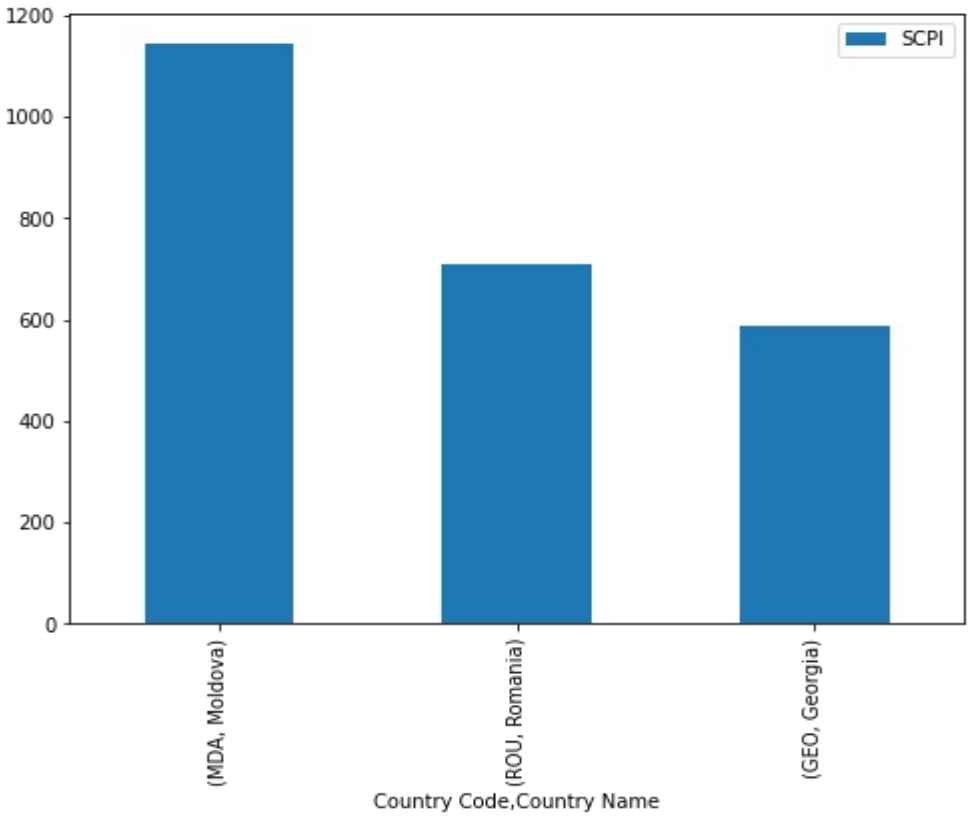
```
df1.head(30).sort_values('SCPI', ascending=True).plot(kind='barh',figsize=(16,8))
```



TOP 3 des pays

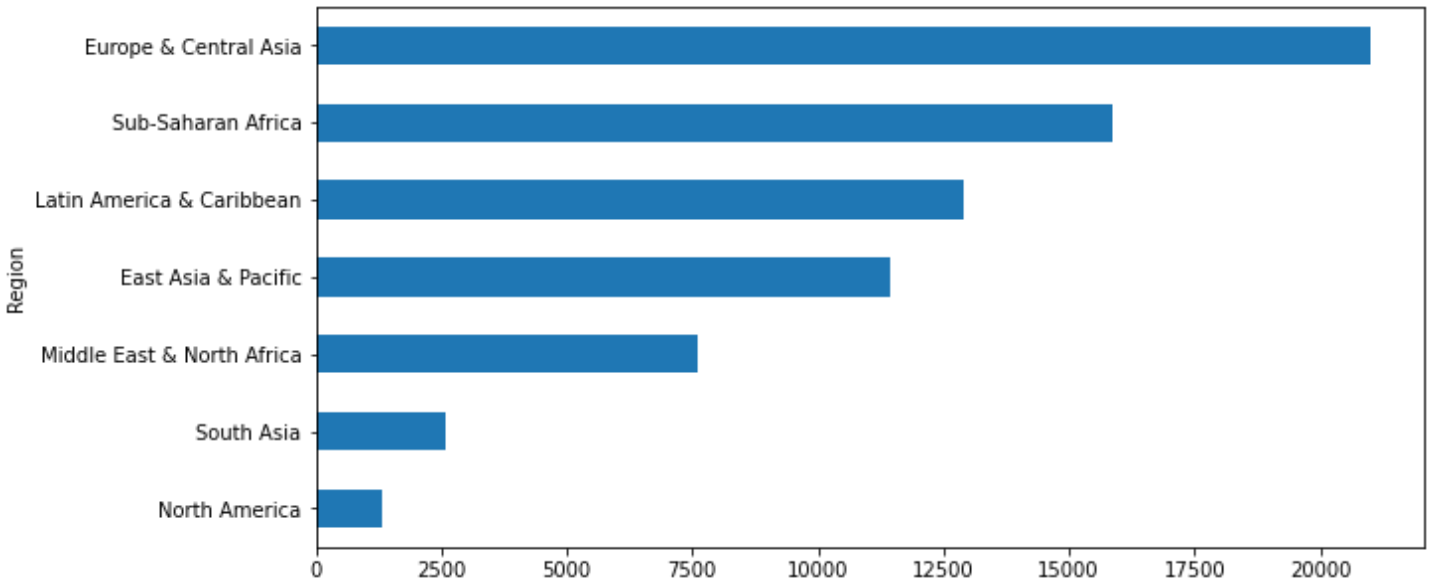
		SCPI
Country Code	**Country Name**	
MDA	**Moldova**	1145.65
ROU	**Romania**	708.60
GEO	**Georgia**	589.21

```
df1.head(3).plot(kind='bar',figsize=(8,6))
```



CLASSEMENT DES REGIONS

```
df_gby_region.sort_values(ascending=True).plot(kind='barh',figsize=(10,5))
```



CONCLUSION

3 pays se degagent les plus innattendus

Resultat a prendre avec precaution.

Pour l etude je conseille d integrer des donnees d autres sources telles que les donnees de data-population (https://www.populationdata.net/palmares/idh/description/) et d autres indicateurs comme l indice de developpement humain.

Les regions en tete sont celle ayant une population elevée. Il serait judicieus de faire une analyse plus poussée sur le choix des coefficients.