

Analysez des données de systèmes éducatifs

OPENCLASSROOMS

Présentation de la mission



academy

Start-up EdTech qui propose des formations en ligne niveau lycée et université

Projet:

Expansion des formations à l'international

Objectif de la mission:

Conduire une analyse exploratoire afin de déterminer si les données de la banque mondiale peuvent apporter des informations dans le cadre du projet

I. Pré-analyse

Présentation des données

EdStatsCountry.csv

```
['Country Code', 'Short Name', 'Table Name', 'Long Name', '2-alpha code',  
'Currency Unit', 'Special Notes', 'Region', 'Income Group', 'WB-2 code',  
'National accounts base year', 'National accounts reference year',  
'SNA price valuation', 'Lending category', 'Other groups',  
'System of National Accounts', 'Alternative conversion factor',  
'PPP survey year', 'Balance of Payments Manual in use',  
'External debt Reporting status', 'System of trade',  
'Government Accounting concept', 'IMF data dissemination standard',  
'Latest population census', 'Latest household survey',  
'Source of most recent Income and expenditure data',  
'Vital registration complete', 'Latest agricultural census',  
'Latest industrial data', 'Latest trade data',  
'Latest water withdrawal data', 'Unnamed: 31'],
```

EdStatsSeries.csv

```
['Series Code', 'Topic', 'Indicator Name', 'Short definition',  
'Long definition', 'Unit of measure', 'Periodicity', 'Base Period',  
'Other notes', 'Aggregation method', 'Limitations and exceptions',  
'Notes from original source', 'General comments', 'Source',  
'Statistical concept and methodology', 'Development relevance',  
'Related source links', 'Other web links', 'Related indicators',  
'License Type', 'Unnamed: 20'],
```

EdStatsData.csv

```
['Country Name', 'Country Code', 'Indicator Name', 'Indicator Code',  
'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', '2005',  
'2006', '2007', '2008', '2009', '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'],
```

EdStatsCountry-Series.csv

	CountryCode	SeriesCode	DESCRIPTION	Unnamed: 3
0	ABW	SP.POP.TOTL	Data sources : United Nations World Population...	NaN
1	ABW	SP.POP.GROW	Data sources: United Nations World Population ...	NaN
2	AFG	SP.POP.GROW	Data sources: United Nations World Population ...	NaN
3	AFG	NY.GDP.PCAP.PP.CD	Estimates are based on regression.	NaN
4	AFG	SP.POP.TOTL	Data sources : United Nations World Population...	NaN

EdStatsFootNote.csv

	CountryCode	SeriesCode	Year	DESCRIPTION	Unnamed: 4
0	ABW	SE.PRE.ENRL.FE	YR2001	Country estimation.	NaN
1	ABW	SE.TER.TCHR.FE	YR2005	Country estimation.	NaN
2	ABW	SE.PRE.TCHR.FE	YR2000	Country estimation.	NaN
3	ABW	SE.SEC.ENRL.GC	YR2004	Country estimation.	NaN
4	ABW	SE.PRE.TCHR	YR2006	Country estimation.	NaN

Présentation de la table Country

- Shape: 241 lignes et 32 colonnes
- Colonne à éliminer : 'Unnamed: 31'
- Colonnes à retenir pour l'analyse: 'Country Code', 'Region', 'Income Group'

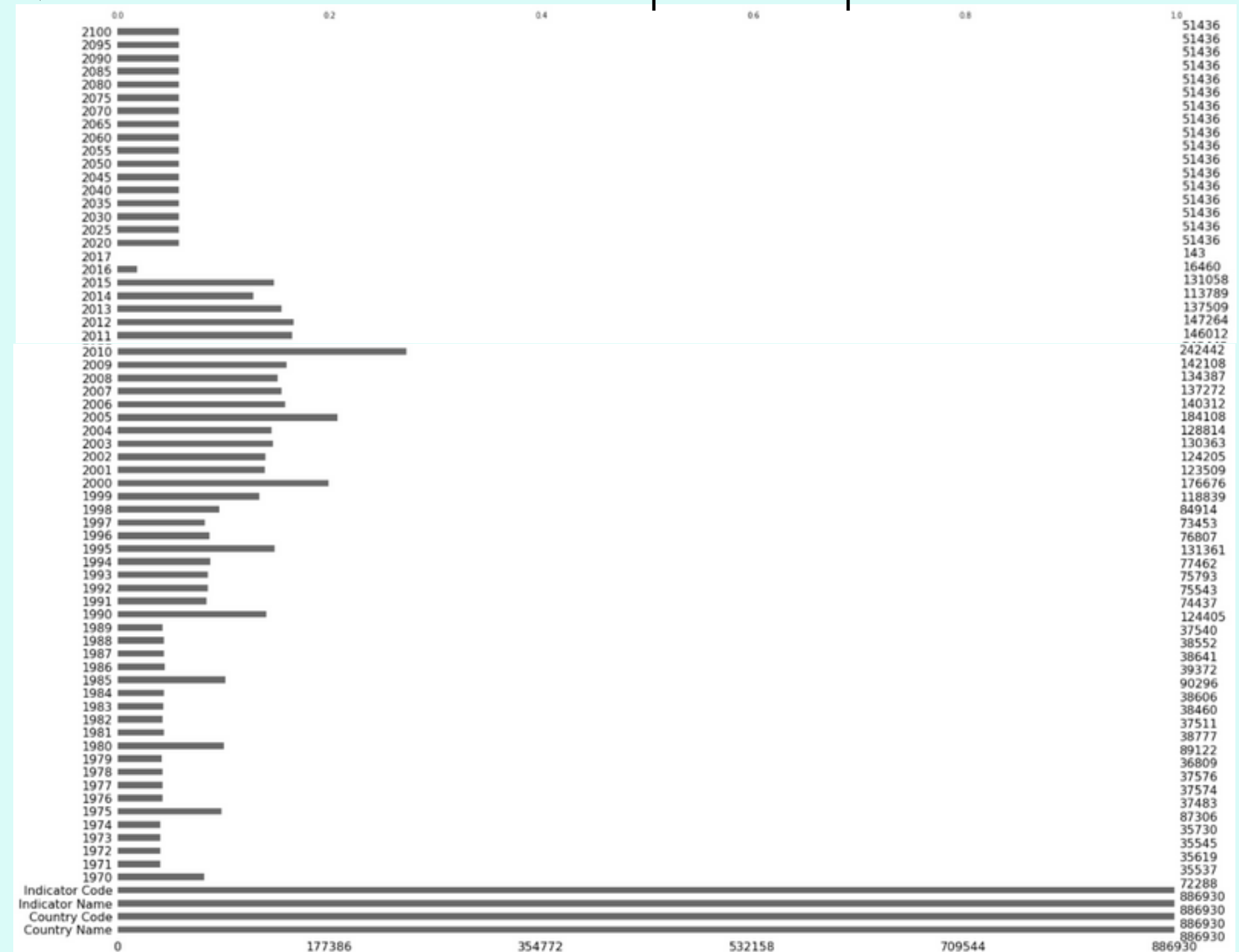
Présentation de la table Series

- Shape: 3665 lignes et 21 colonnes
- Colonne à éliminer : 'Unnamed: 20'
- Colonnes à retenir pour l'analyse: 'Series Code', 'Indicator Name', 'Long Definition'

Présentation de la table Data

- Shape: 886 930 lignes et 70 colonnes
- Colonne à éliminer : 'Unnamed: 69'
- Nombre de pays: 242
- Nombre d'indicateurs: 3665
- None values →

Nombre de données non-nulles pour chaque colonne



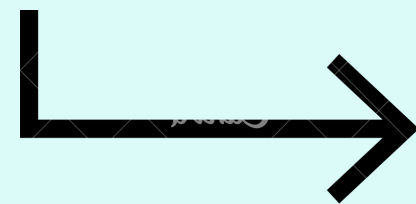
Nettoyer les données – Pays

Nombre de pays: **242** (195 pays indépendants selon l'ONU)

```
['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', 'Afghanistan', 'Albania',
```

=

- Beaucoup de pays sont des **régions** (Europe & Central Asia) ou des **groupes de pays** (Lower Middle Income) .
- Un pays sans Région



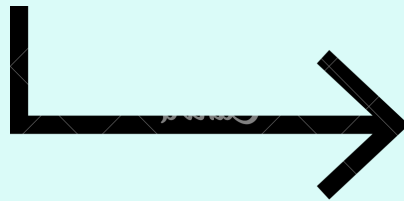
Après le filtre: **214** pays

Nettoyer les données – Indicateurs

Nombre d'indicateurs: **3665**

Series Code	Indicator Name
IT.NET.USER.P2	Internet users (per 100 people)
NY.GDP.PCAP.CD	GDP per capita (current US\$)
SE.TER.ENRL	Enrolment in tertiary education, all programmes, both sexes (number)
SP.POP.1564.TO	Population ages 15-64, total
SP.SEC.UTOT.IN	Population of the official age for upper secondary education, both sexes (number)
SP.TER.TOTL.IN	Population of the official age for tertiary education, both sexes (number)
UIS.E.3	Enrolment in upper secondary education, both sexes (number)
UIS.E.4	Enrolment in post-secondary non-tertiary education, both sexes (number)
UIS.SAP.4	Population of the official age for post-secondary non-tertiary education, both sexes (number)
UIS.XUNIT.GDPCAP.3.FSGOV	Government expenditure per upper secondary student as % of GDP per capita (%)

Shape de la DataFrame après le
filtre sur les indicateurs et les pays:
2150 lignes et 69 colonnes

 **10** indicateurs

Nettoyer les données – Années

Filtre sur les années:

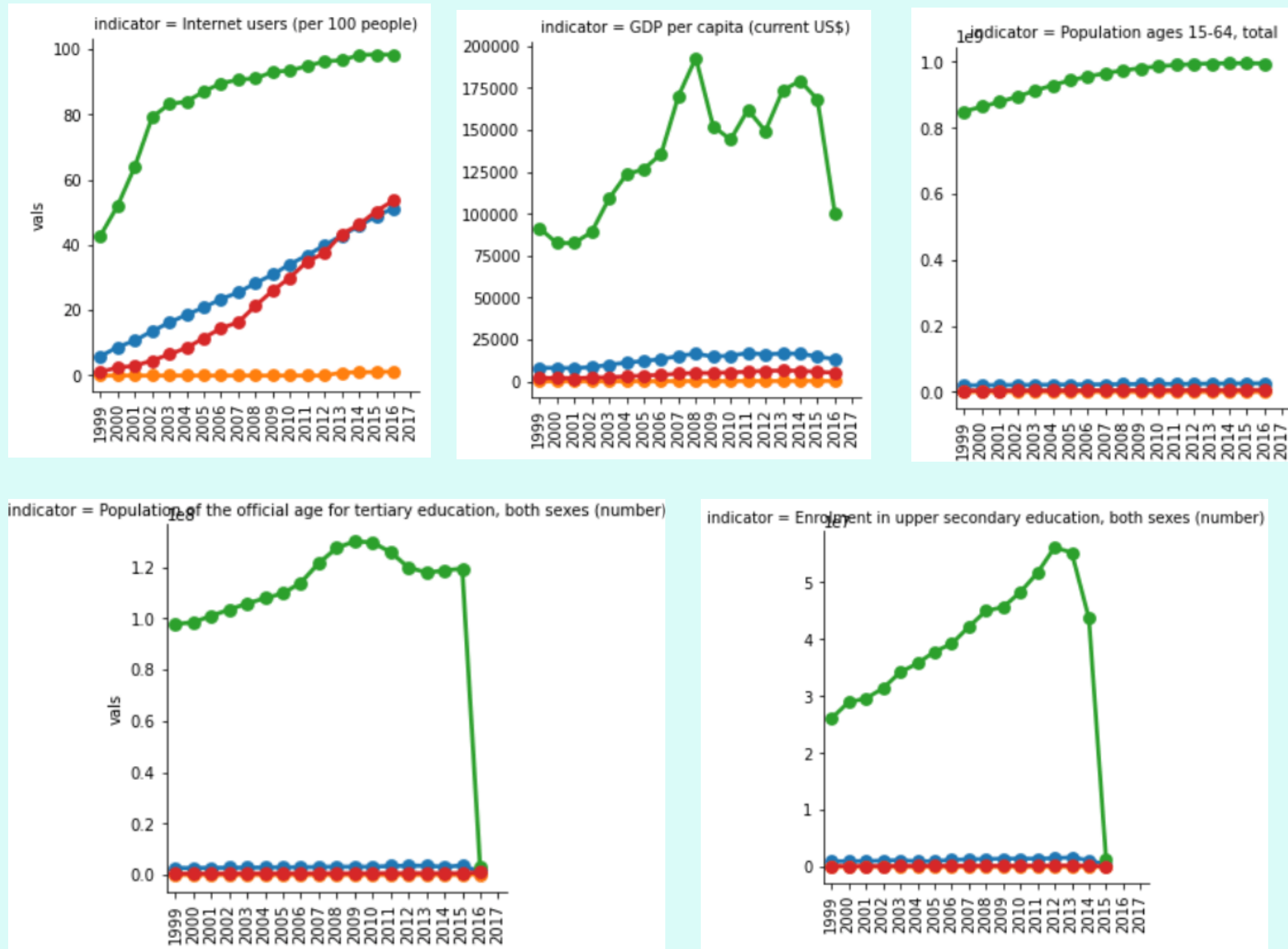


Filtre sur les lignes:

- **85%** de None Values
- Aucune valeur sur les **10 dernières années**

Shape: **1686 , 23**

Analyse des indicateurs dans le temps



→ Trop de données pour pouvoir en tirer des conclusions intéressantes

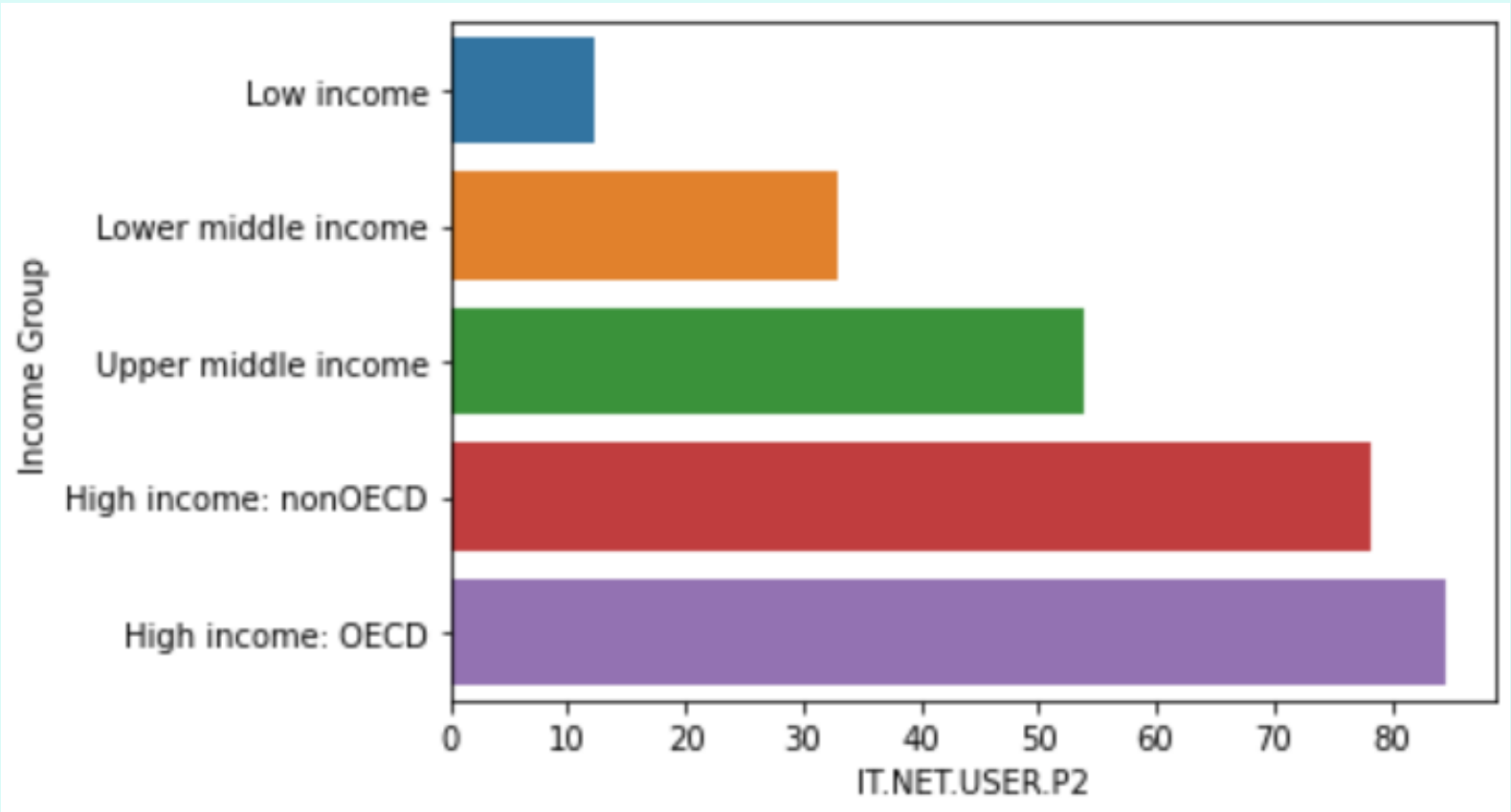
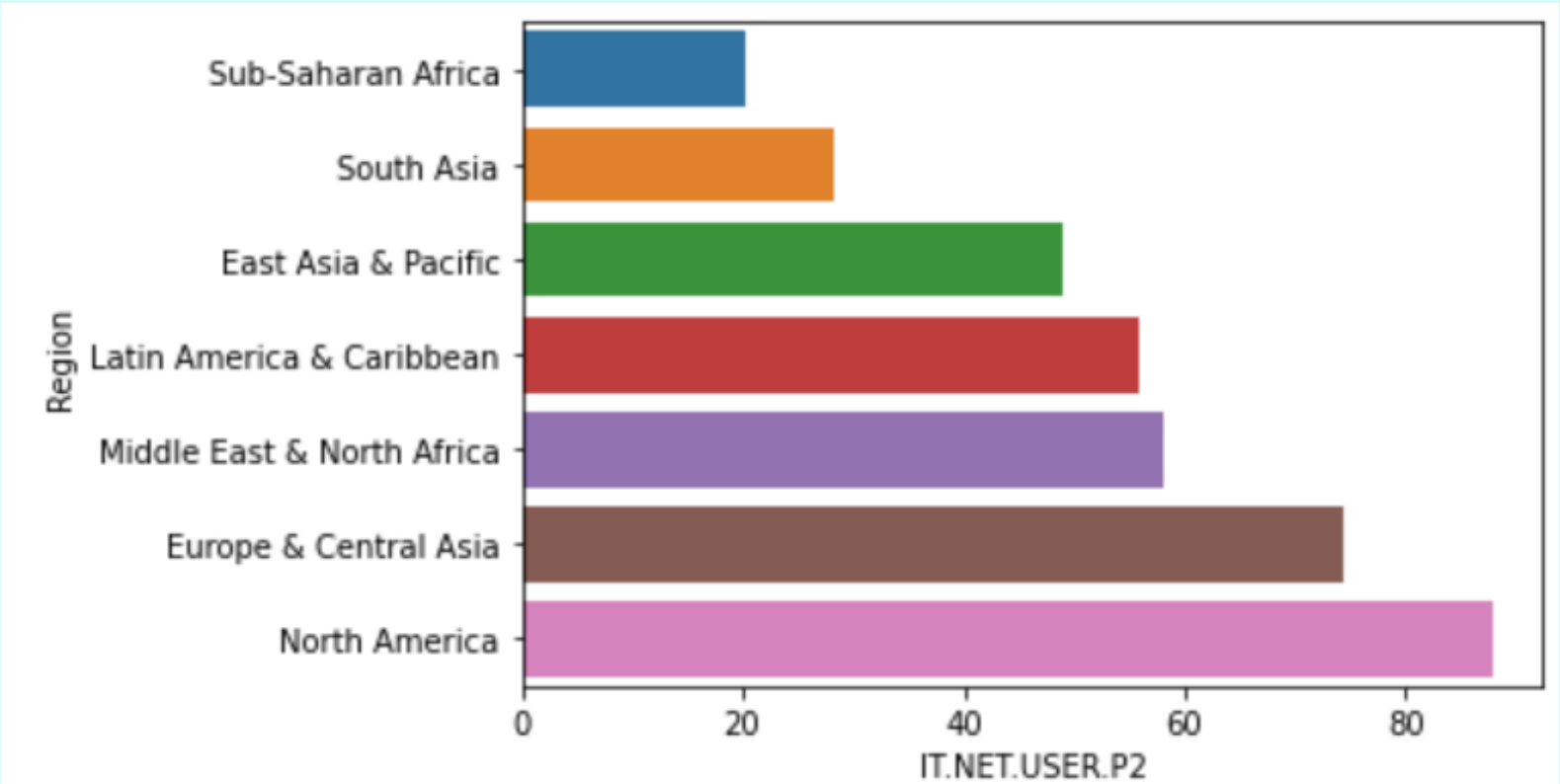
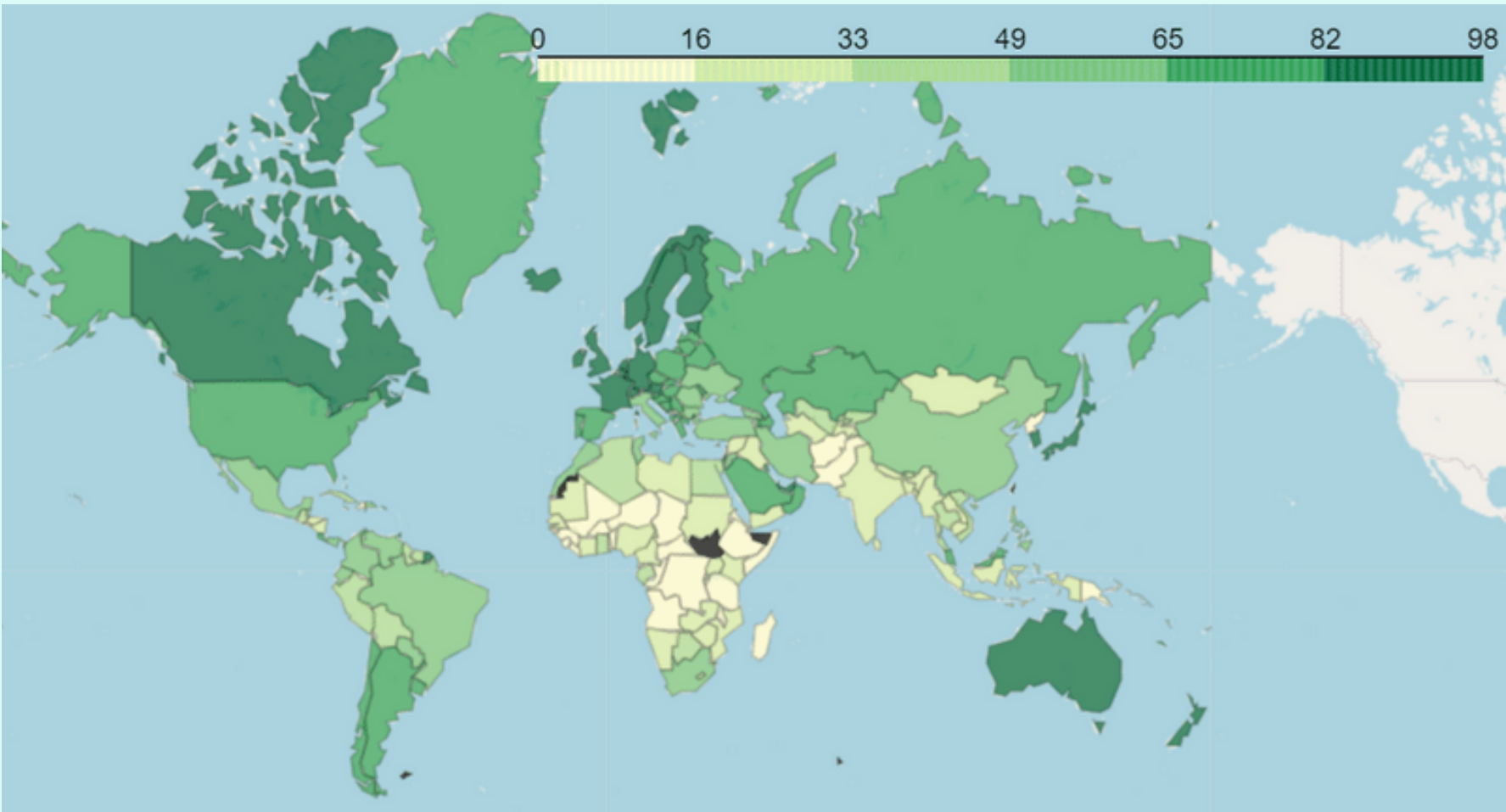
Stratégie:

Garder pour chaque indicatuer/pays la donnée la plus récente (max 10 ans)

II. Analyse

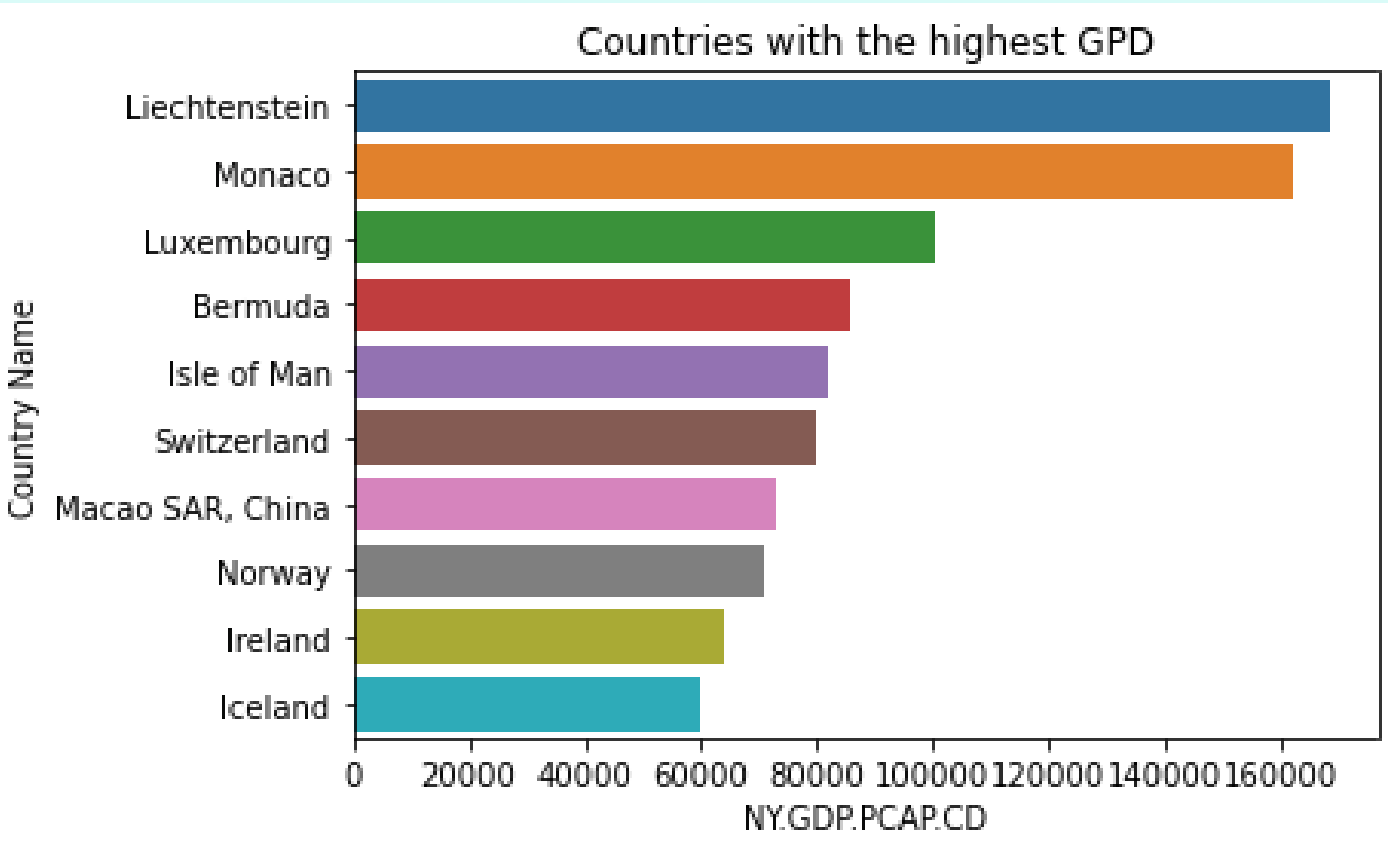
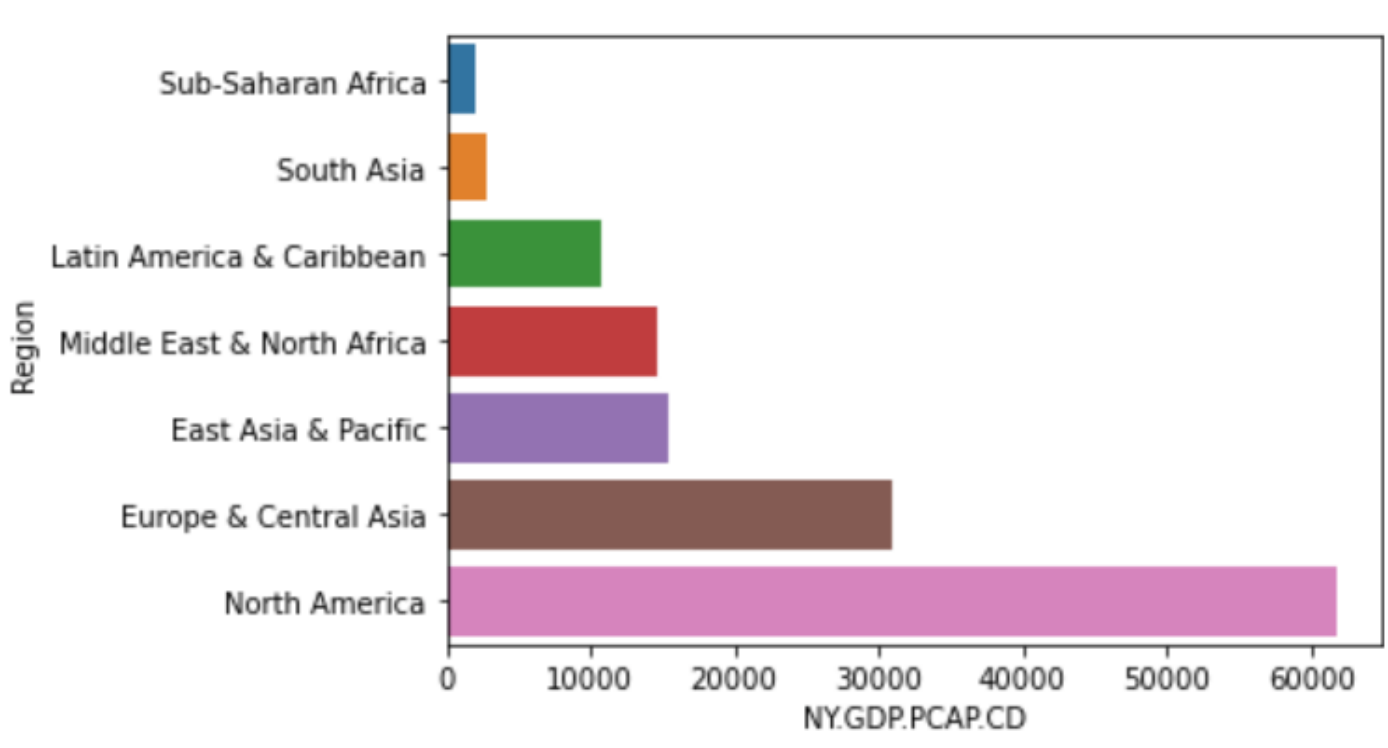
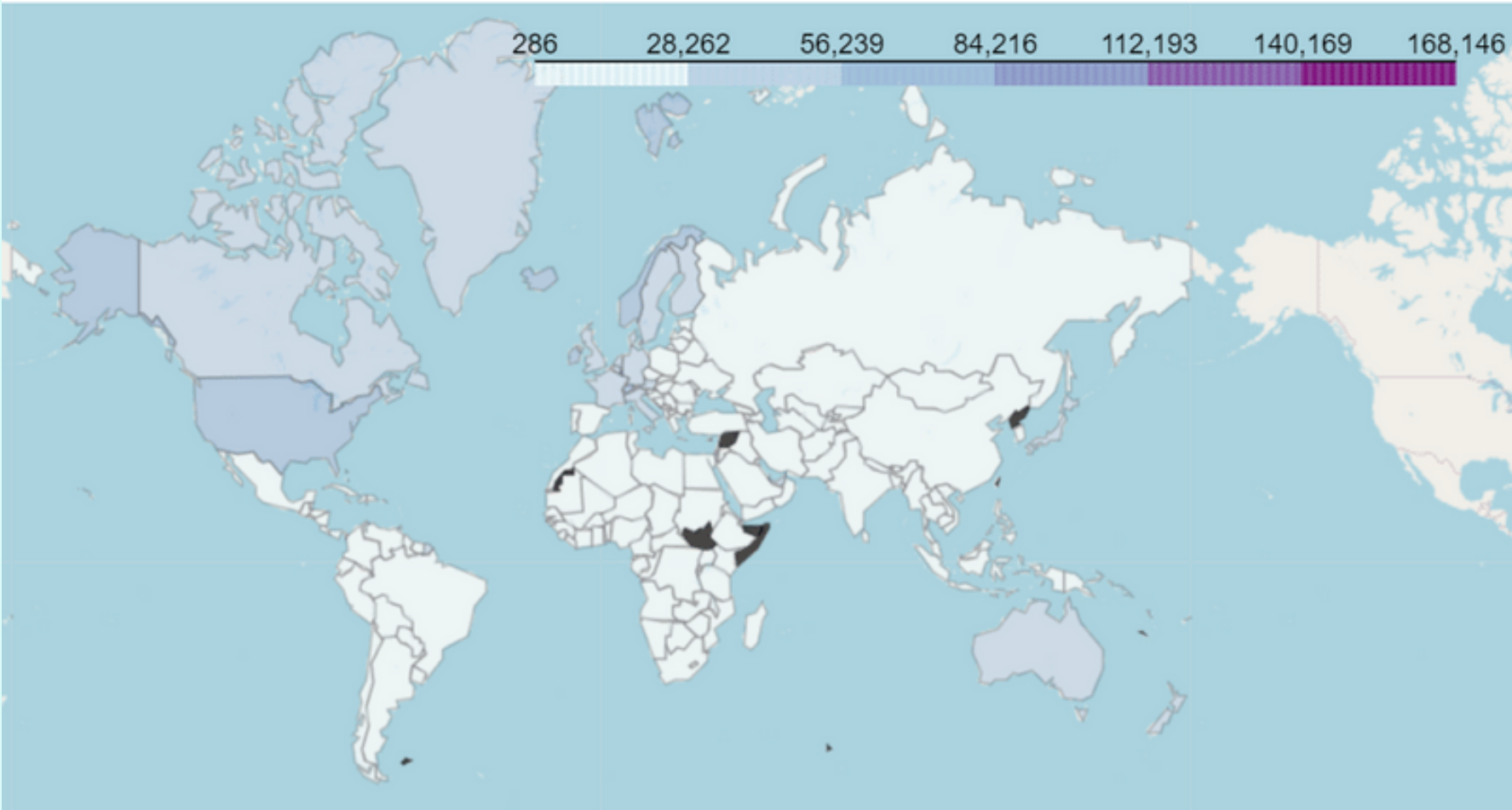
Indicateur: Internet user (100 personnes)

count	743995.000000
mean	51.051608
std	28.467729
min	0.000000
25%	25.246250
50%	53.226773
75%	76.176737
max	98.240016



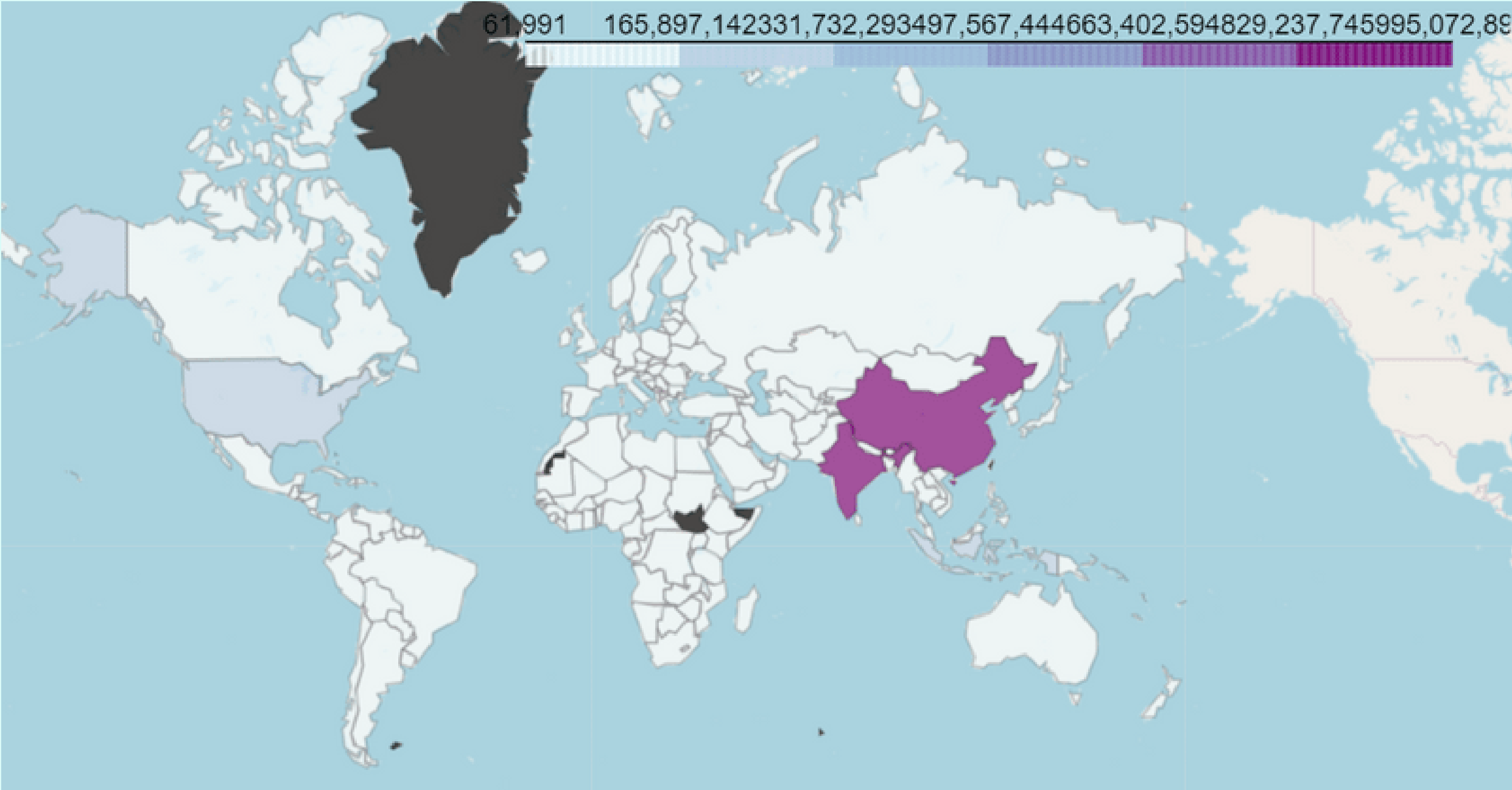
Indicateur: PIB

count	203.000000
mean	15888.812275
std	24294.301541
min	285.727442
25%	2058.065758
50%	5602.549434
75%	18211.052170
max	168146.015281



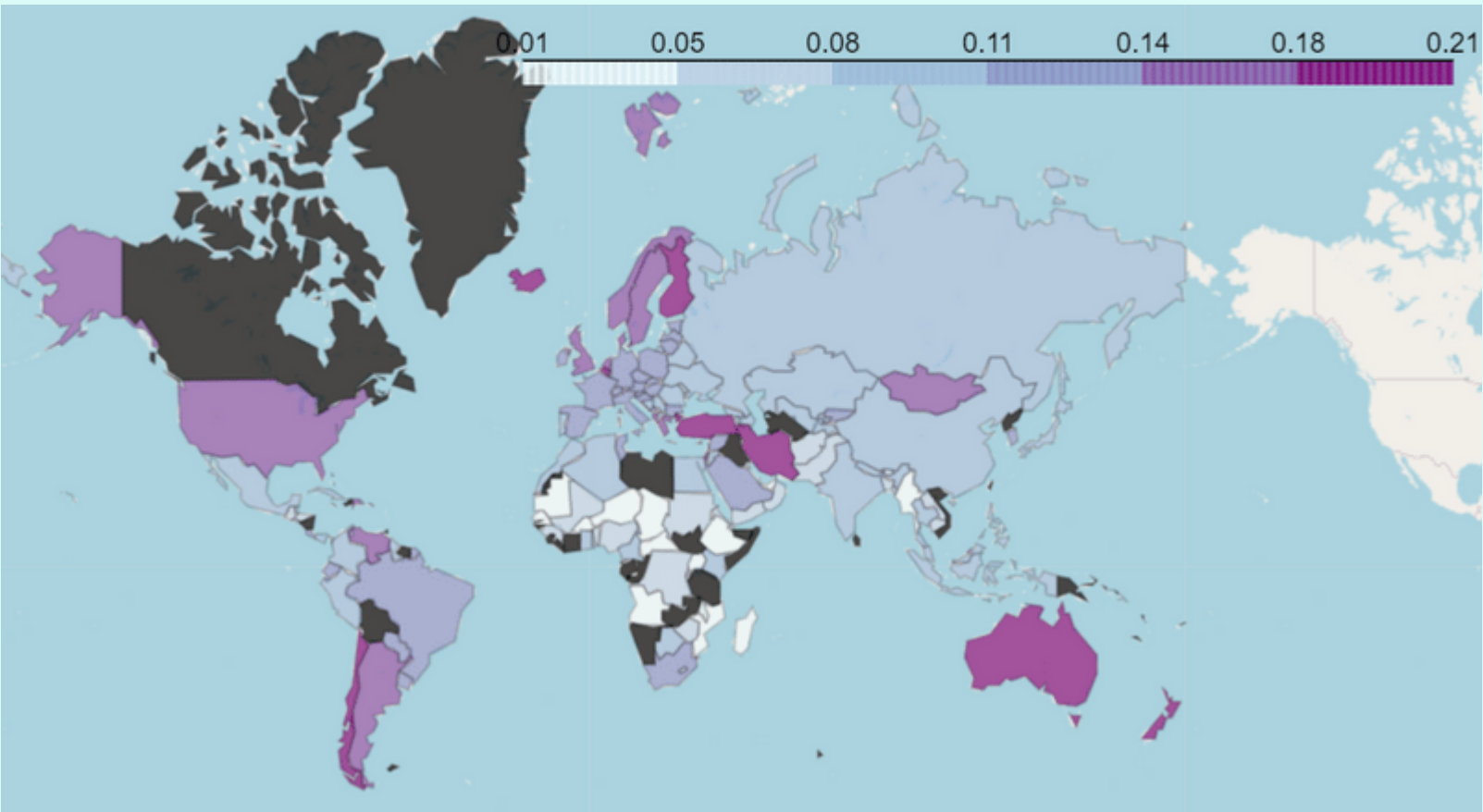
Indicateur: Population entre 15 et 64 ans

count	1.940000e+02
mean	2.500610e+07
std	9.770916e+07
min	6.199100e+04
25%	1.335739e+06
50%	5.366097e+06
75%	1.583602e+07
max	9.950729e+08

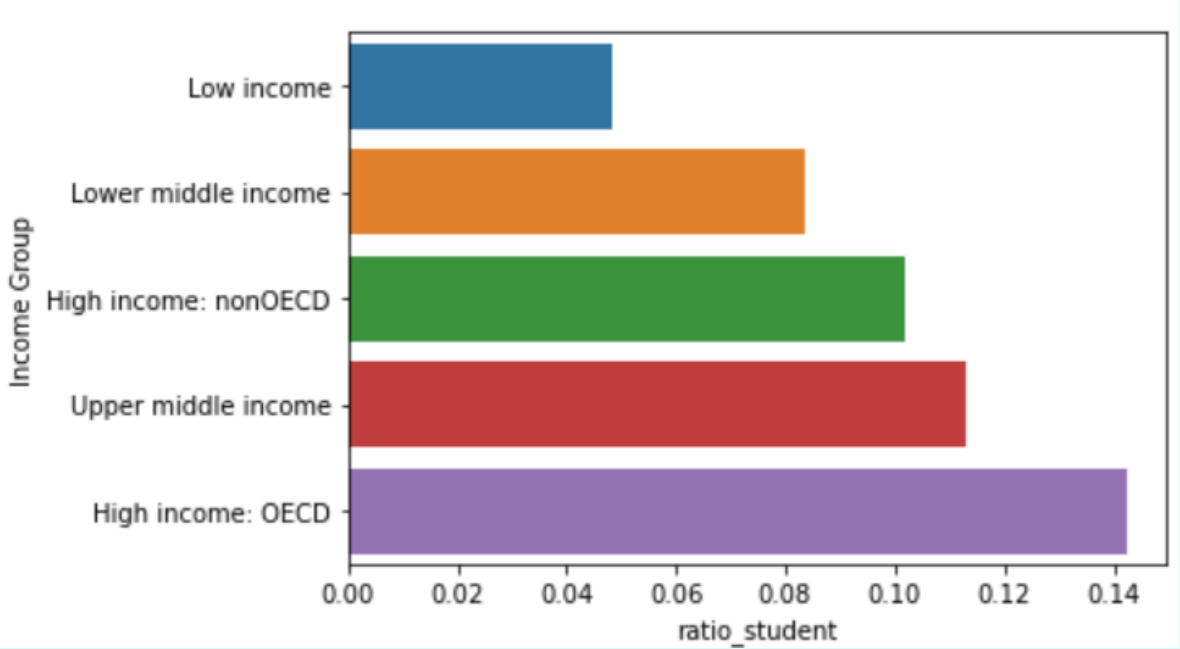
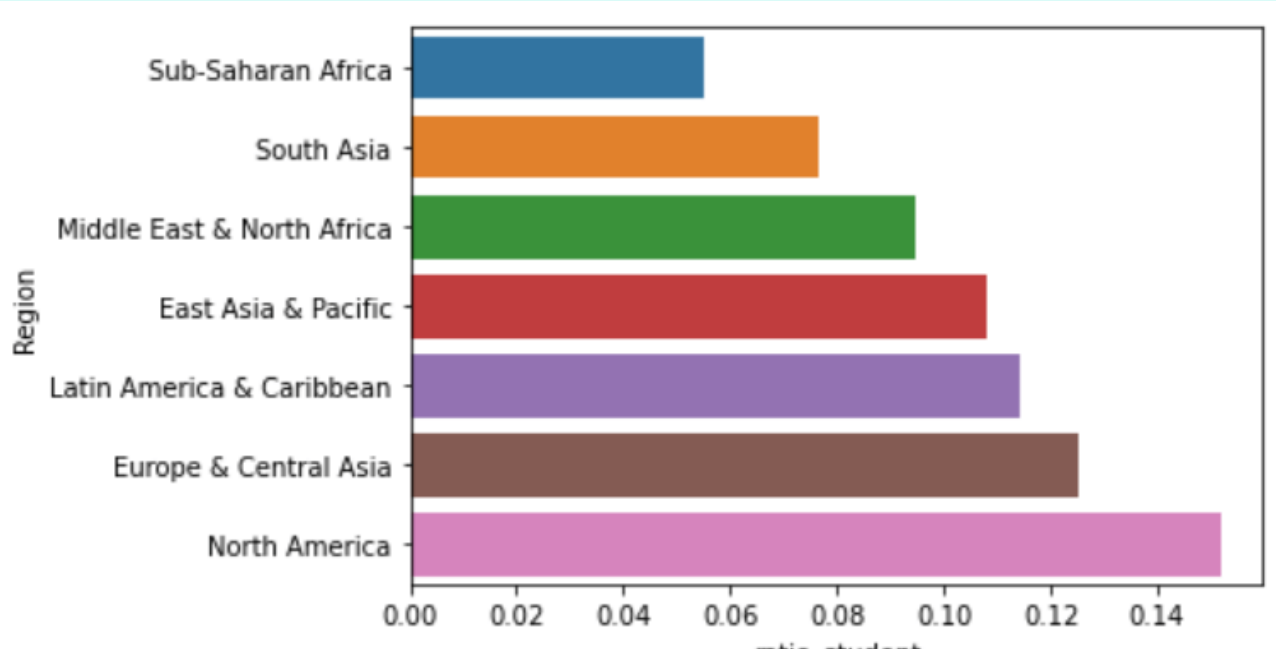


Indicateur: Ratio des étudiants

$$\text{RatioStudent} = \frac{n\text{Secondary} + n\text{Tertiary} + n\text{NonTertiary}}{\text{Population15/64}}$$



count	148.000000
mean	0.096216
std	0.041322
min	0.013131
25%	0.069468
50%	0.098269
75%	0.119330
max	0.208409



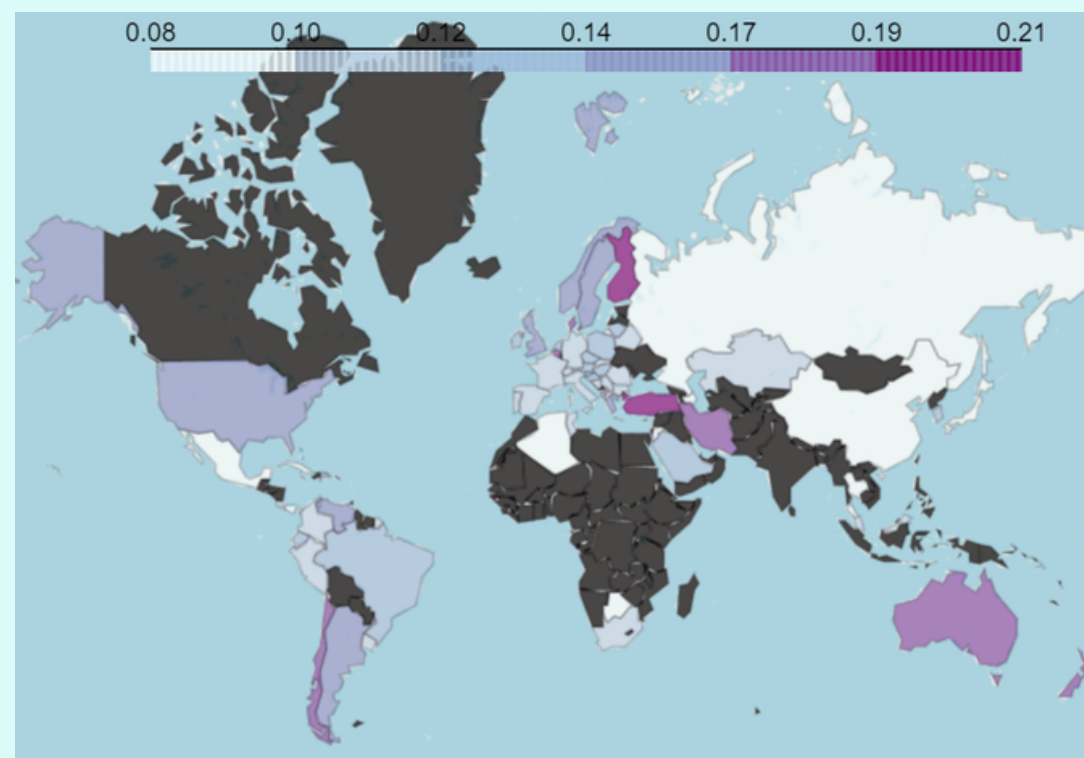
Filterer les données

Filtre sur les pays qui ont:

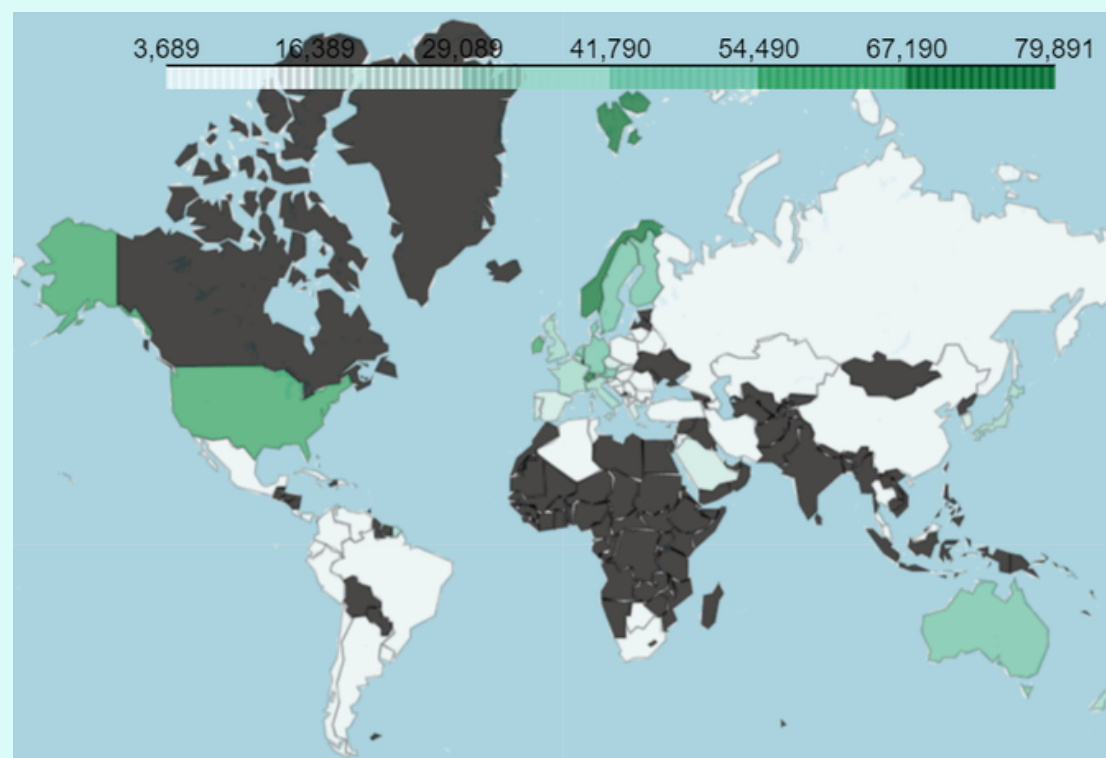
- Un revenu faible ou moyen faible
- Une population faible (pays en dessous du premier quartile)
- Un ratio faible de personnes potentiellement intéressées par le produit

➔ **68** pays potentiellement intéressants

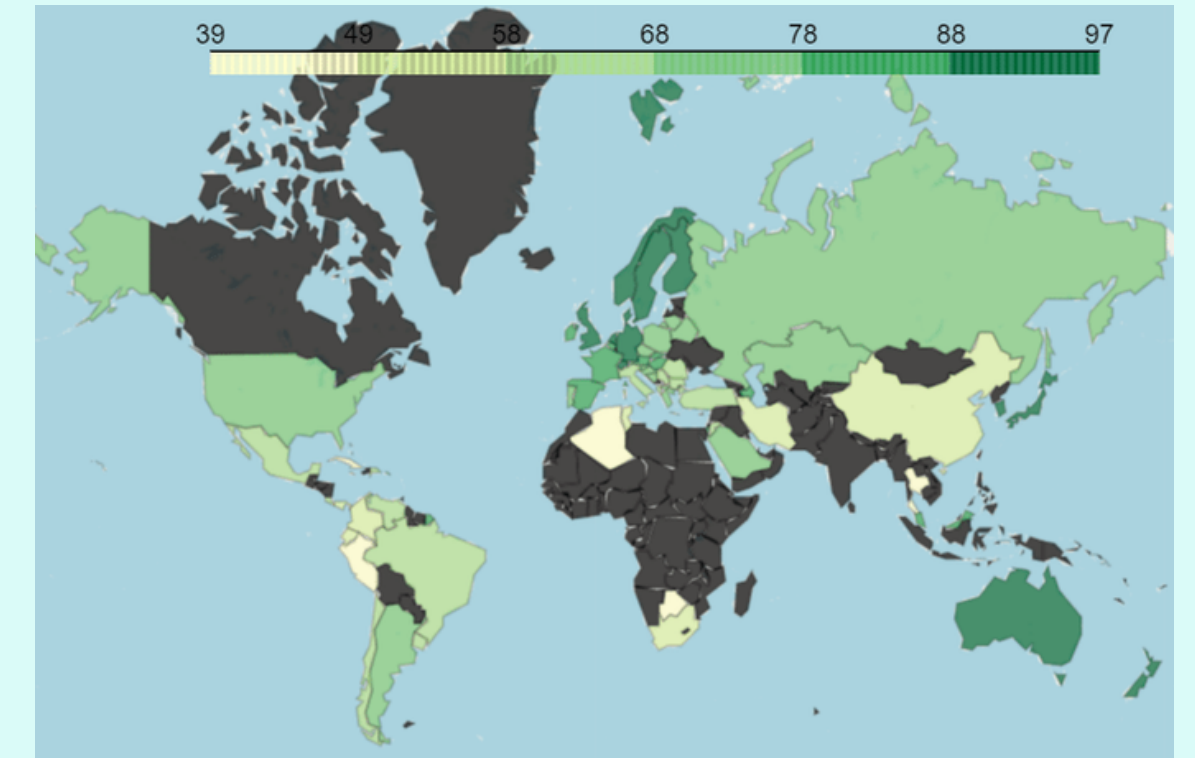
Ratio of student



PIB



Interent Users

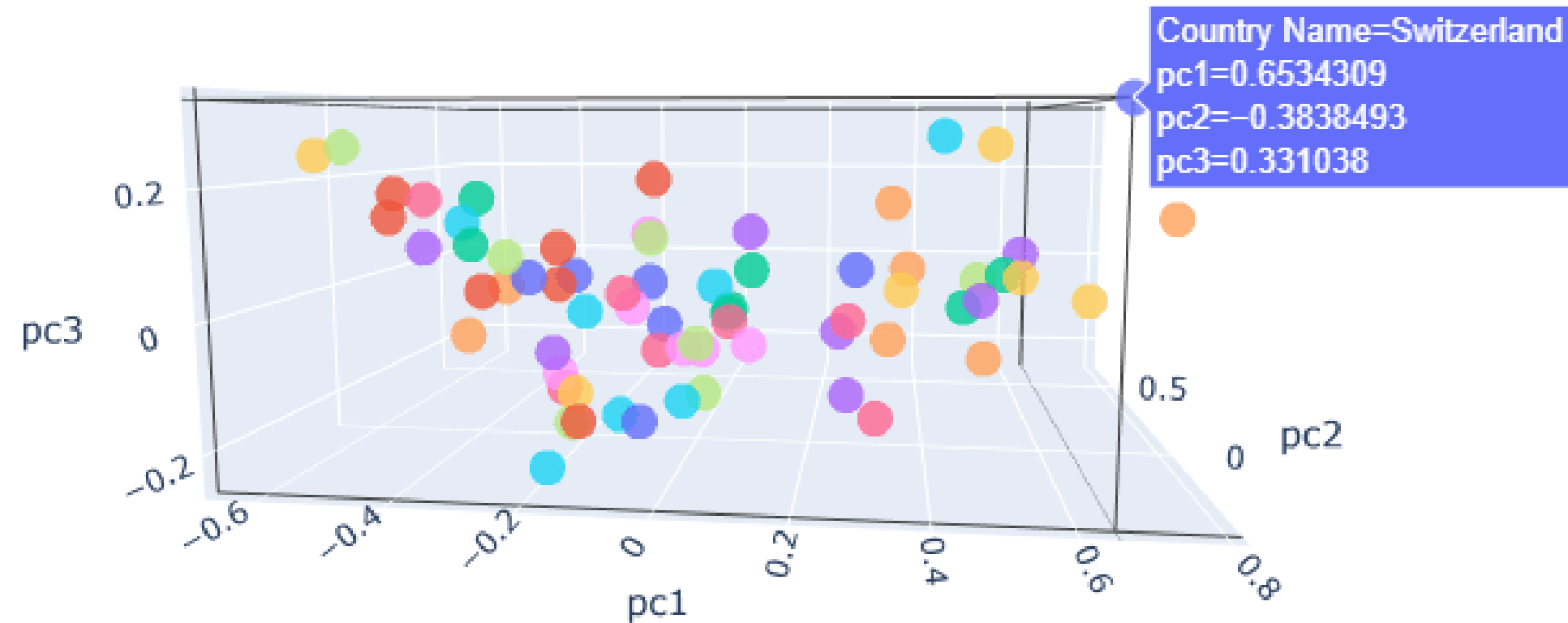


Comparer les pays

Indicateurs: Internet Users, PIB, Ratio des étudiants

Outil: PCA (**3 Composantes** ce qui nous permet de garder **100% de l'information initiale**)

Features weighted:
{'IT.NET.USER.P2': 25.69, 'NY.GDP.PCAP.CD': 35.72, 'ratio_student': 38.6}



→ A partir des coordonnées de chaque pays et des poids de chaque composante principale:

Calcul d'un score

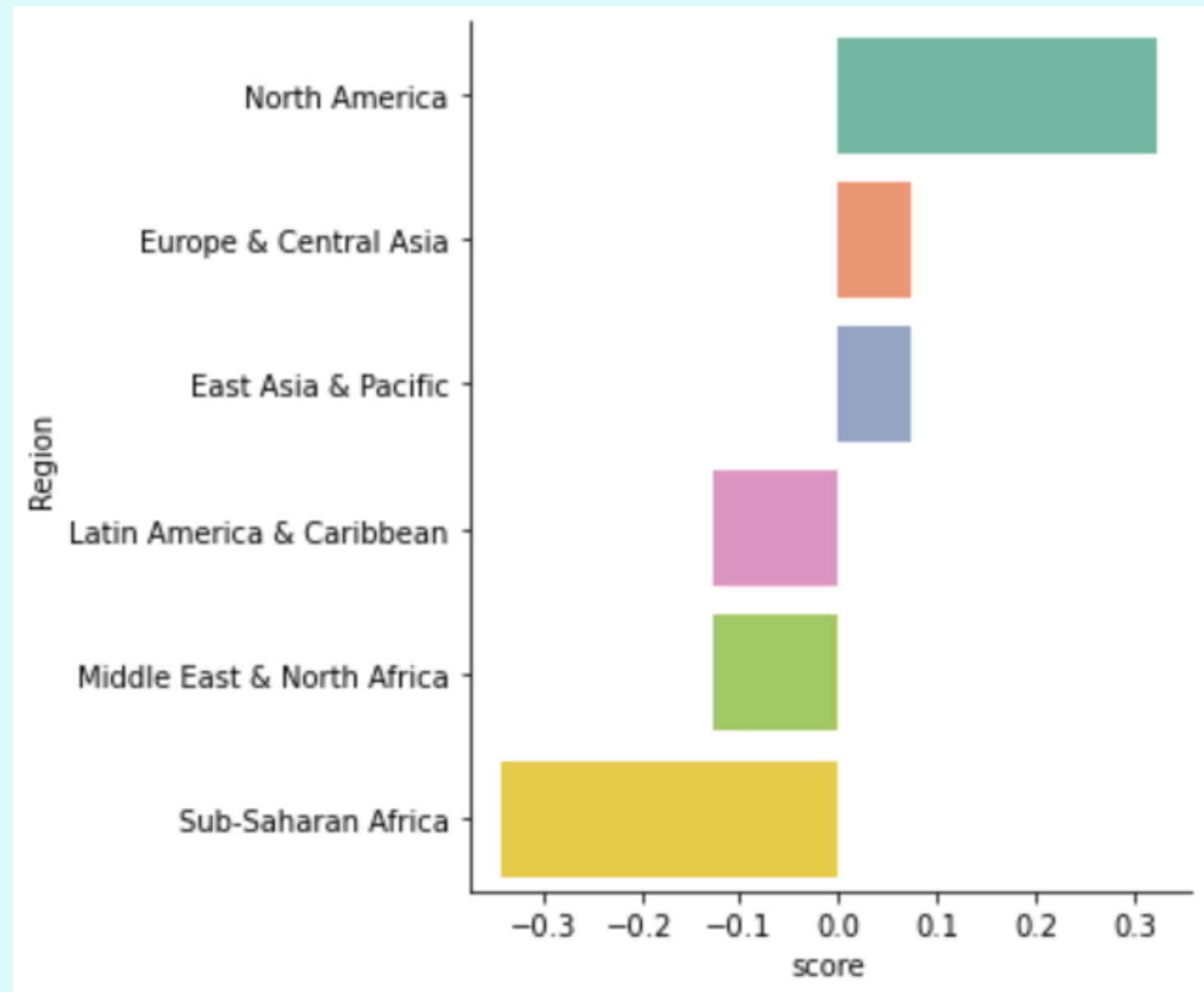
Poid pc1: 0.68877254

Poid pc2: 0.21989843

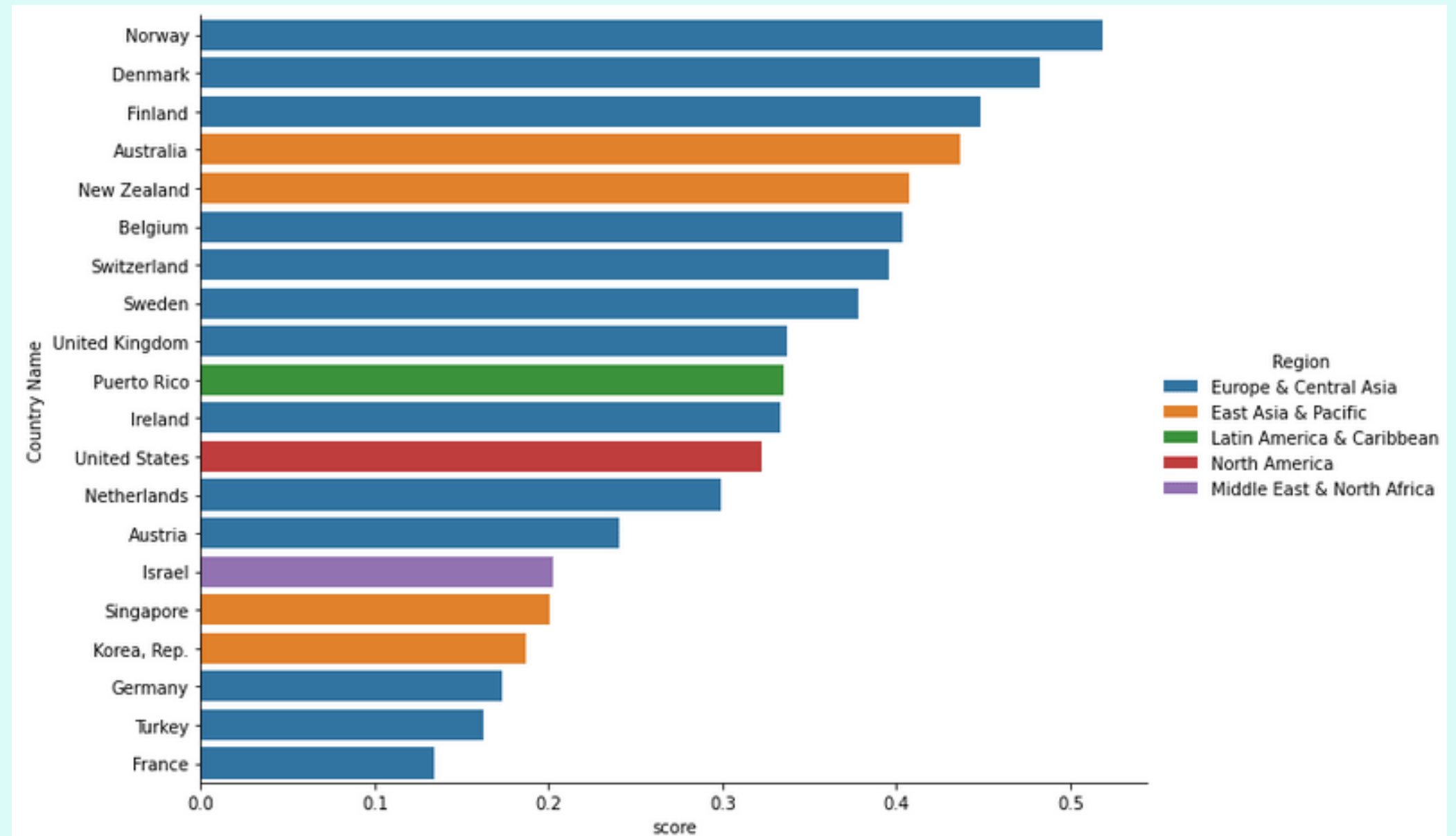
Poid pc3: 0.09132904

Comparer les pays

Top Regions:



Top 20 pays:



III. Conclusion

Limites

Des données:

- Peu d'indicateurs qui ont des données de prédictions
- Beaucoup de valeurs manquantes (ex:Canada)
- Il manque des indicateurs intéressants
- Incohérence entre certains indicateurs

Informations sur le projet:

- Informations sur les pays dans lesquels Academy est déjà présent
- Informations sur les types de cours
- Informations sur les potentiels concurrents
- Informations sur les prix des cours

Conclusion

Les données permettent de:

- Identifier les régions géographiques et les groupes de pays les plus intéressants
- Avoir une idée des pays qui pourraient être attrayant pour academy

Next steps:

- Valider les indicateurs avec une personne business
- Prédire les données des indicateurs sélectionnés grâce à une régression

Questions?

Merci pour votre attention