

# **Analyse des données de systèmes éducatifs**

**Projet 2 Parcours Data Scientist  
Python - Support Jupyter Notebook**

Nalron Septembre 2020  
OpenClassrooms - Centrale Supélec

# **Academy start-up de la EdTech**

**Formation en ligne  
pour un public de  
niveau lycée et  
université**



# Projet de l'entreprise

**« Développement à l'International »**

**Décision d'ouverture de la plateforme  
vers de nouveau pays...**

# Problématique

- **Quels sont les pays à fort potentiel pour nos services?**
- **Quelle évolution du potentiel clients?**
- **Quels pays l'entreprise doit-elle opérer en priorité?**

# **Données sur l'éducation de la banque mondiale**

**3665 indicateurs niveau International**

**L'accès à l'éducation, l'obtention de diplômes et des informations relatives aux professeurs, aux dépenses liées à l'éducation...**

**Cycle de l'enseignement primaire à l'enseignement supérieur**

**<https://datacatalog.worldbank.org/dataset/education-statistics>**

***ou en téléchargement direct à ce lien***

# Analyse Pré-exploratoire

**Qualité du jeu de données, informations descriptives,  
analyse des indicateurs, des années et des zones  
géographiques**

# Généralités sur les données

- 5 jeux de données à degré d'utilité très variable
- Observations à forte granularité géographique : pays et/ou zones
- Variables disponibles sur notre contexte métier de l'éducation
- Variables disponibles avec des notions plus larges (population, richesse, technologique, ...)
- Historique des données à partir de 1970, mais sans grande utilité causée par les valeurs manquantes
- 2000 à 2015 sont des années recommandées pour le traitement des objectifs attendus

# Description rapide des données

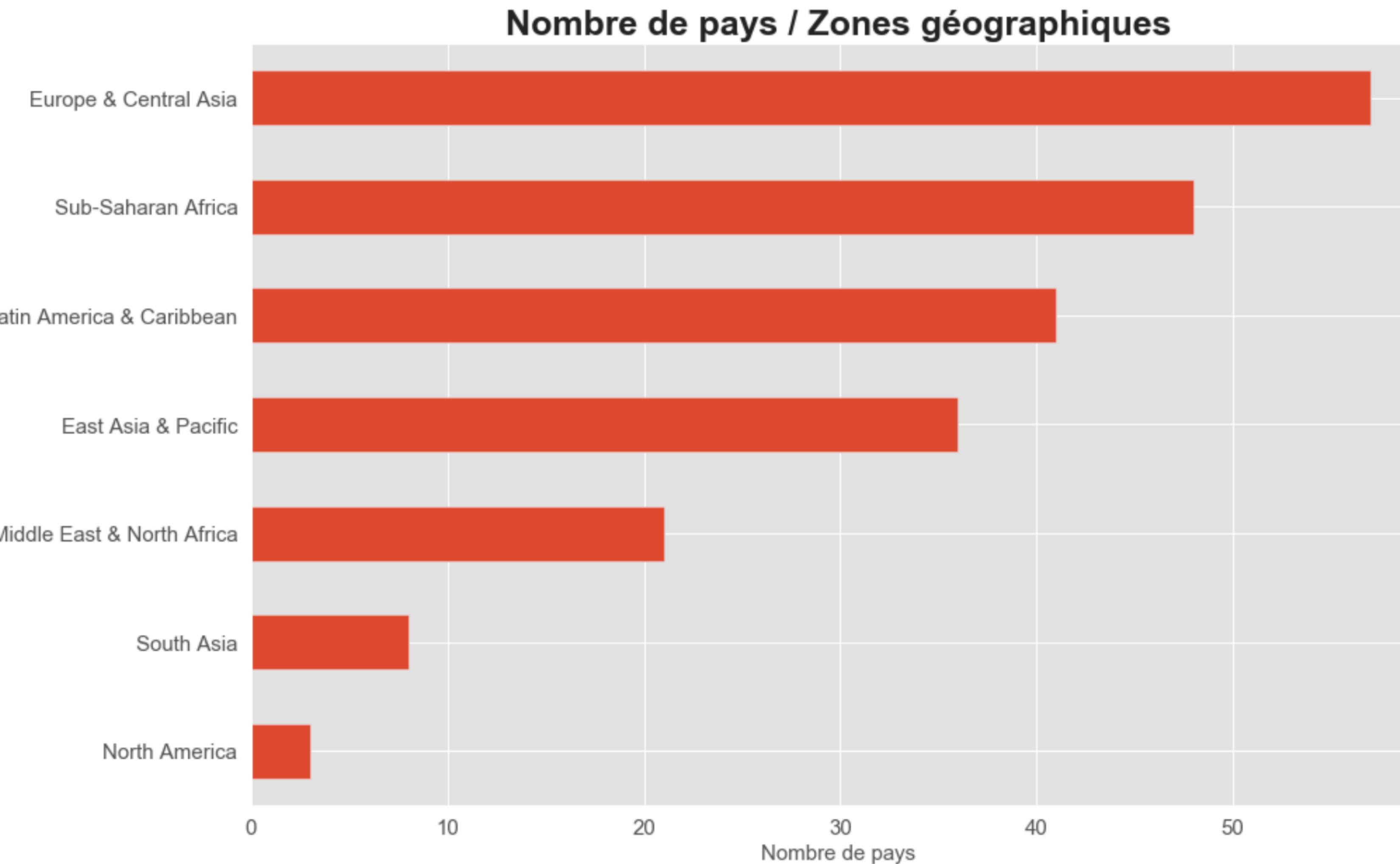
Fichiers	Observations	Variables	Valeurs manquantes	Doublon	Suppression variable	Variable la plus importante
EdStatsData.csv	886930	70	53455179	0	Unnamed: 69	Indicator Code
EdStatsSeries.csv	3665	21	55203	0	Unnamed: 20	Topic
EdStatsCountry.csv	241	32	2354	0	Unnamed: 31	Region
EdStatsFootNote.csv	643638	5	0	0	Unnamed: 4	DESCRIPTION
EdStatsCountry-Series.csv	613	4	0	0	Unnamed: 3	DESCRIPTION

# **Informations clés identifiées**

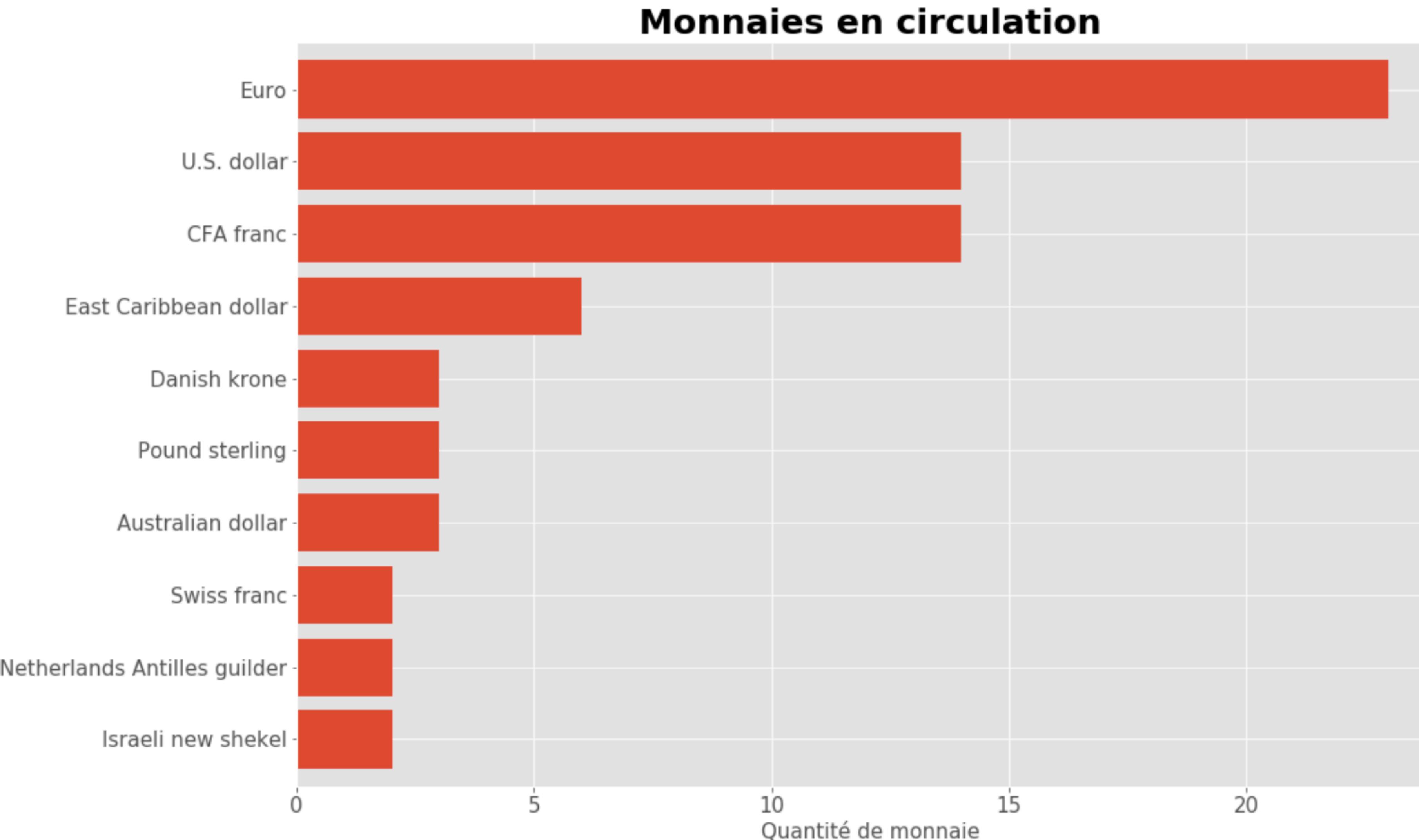
**« Analyses et graphiques »**

**Pays, zones géographiques, devises monétaires, mots clés, indicateurs statistiques, projections futures...**

# Pays et zones géographiques



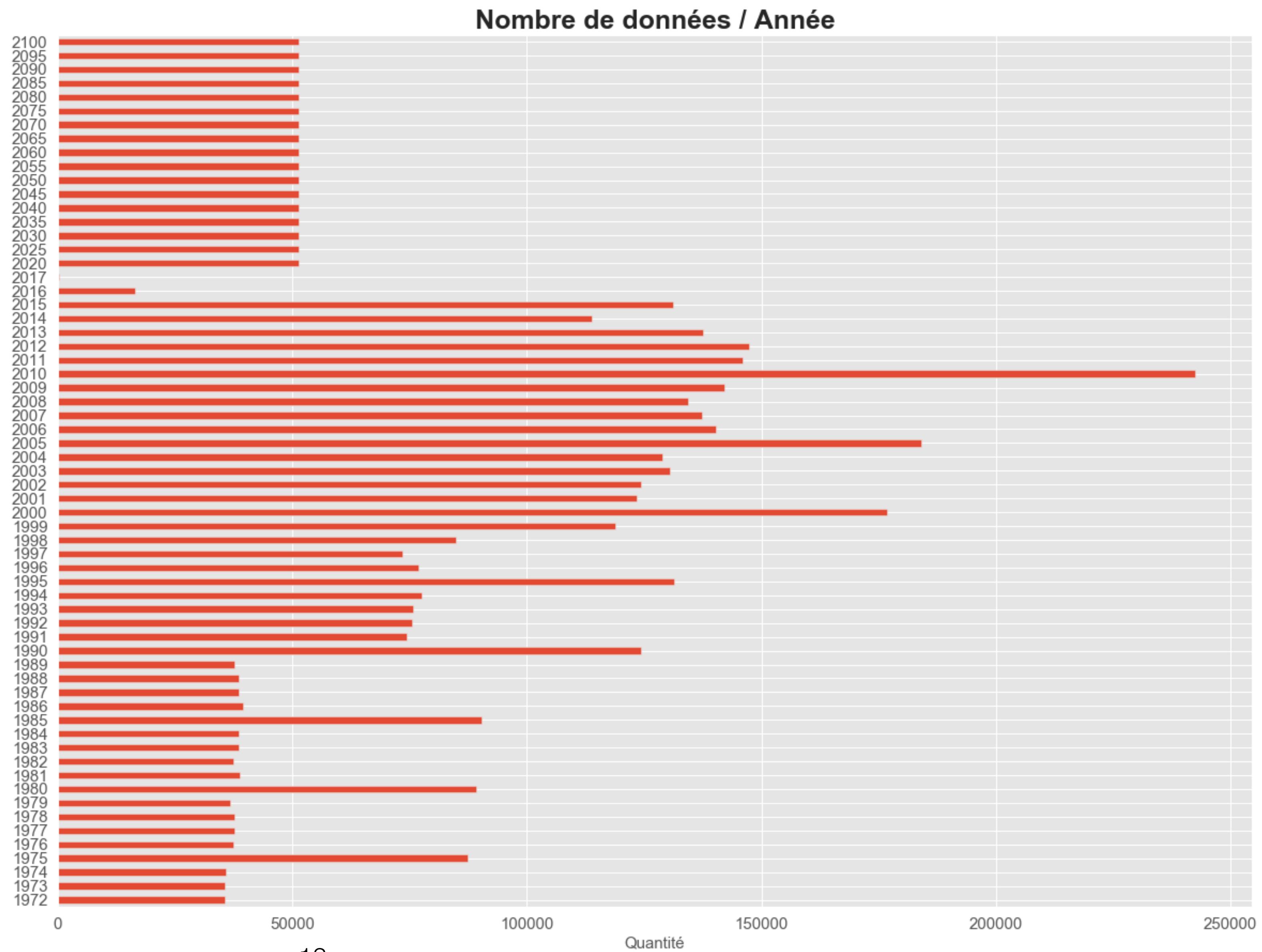
# Devises monétaires



# Années cibles 2000 à 2015

**Choix des années :**

- Contexte métier
- Données disponibles



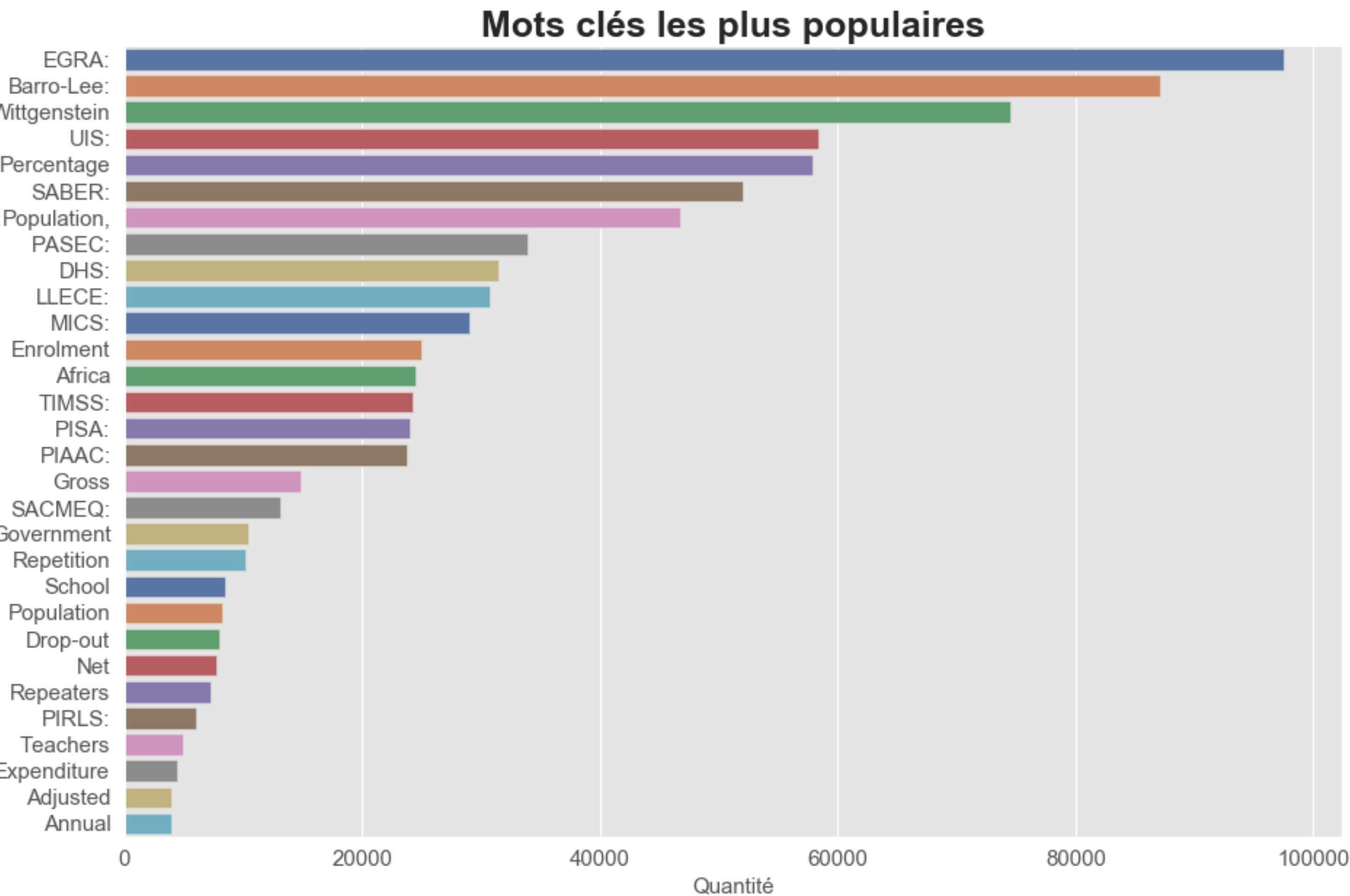
# Catégories d'indicateurs

**3665 indicateurs classés selon les catégories suivantes**

```
[ 'Attainment', 'Education Equality',
  'Infrastructure: Communications', 'Learning Outcomes',
  'Economic Policy & Debt: National accounts: US$ at current prices: Aggregate indicators',
  'Economic Policy & Debt: National accounts: US$ at constant 2010 prices: Aggregate indicators',
  'Economic Policy & Debt: Purchasing power parity',
  'Economic Policy & Debt: National accounts: Atlas GNI & GNI per capita',
  'Teachers', 'Education Management Information Systems (SABER)',
  'Early Child Development (SABER)',
  'Engaging the Private Sector (SABER)',
  'School Health and School Feeding (SABER)',
  'School Autonomy and Accountability (SABER)',
  'School Finance (SABER)', 'Student Assessment (SABER)',
  'Teachers (SABER)', 'Tertiary Education (SABER)',
  'Workforce Development (SABER)', 'Literacy', 'Background',
  'Primary', 'Secondary', 'Tertiary', 'Early Childhood Education',
  'Pre-Primary', 'Expenditures', 'Health: Risk factors',
  'Health: Mortality',
  'Social Protection & Labor: Labor force structure', 'Laber',
  'Social Protection & Labor: Unemployment',
  'Health: Population: Structure', 'Population',
  'Health: Population: Dynamics', 'EMIS',
  'Post-Secondary/Non-Tertiary'], dtype=object)
```

# Mots clés liés aux indicateurs

- **EGRA :**  
**Early Grade Reading Assessment**
- **Barro-lee :**  
**Dataset relatif à l'éducation**
- **Wittgenstein :**  
**Wittgenstein Centre Human Capital Data Explore**
- **UIS :**  
**UNESCO Institut de Statistiques**
- **PISA :**  
**Tests comparatifs de compétences pour les élèves**
- **Teachers**



# Choix des indicateurs

- **IT.NET.USER.P2** : Taux d'accès à internet (pour 100 personnes)
- **NY.GDP.PCAP.C** : PIB par habitant
- **SP.POP.TOTL** : Population Totale
- **SP.POP.1524.TO.UN** : Population âges 15-24
- **UIS.E.4** : Inscription dans l'enseignement post-secondaire non tertiaire H/F
- **UIS.E.3** : Inscription dans l'enseignement secondaire supérieur H/F
- **SE.TER.ENRL** : Inscription dans l'enseignement supérieur, full programmes H/F
- **PRJ.POP.ALL.4.MF** : Potentiel d'évolution, projection sur les années à venir

# Vérification des valeurs disponibles

	Indicator Name	Indicator Code	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
0	Enrolment in post-secondary non-tertiary education	UIS.E.4	111	113	103	106	103	100	95	101	100	98	95	96	99	108	88	2
1	Enrolment in tertiary education, all programme levels	SE.TER.ENRL	149	148	158	159	159	154	154	154	159	162	165	167	166	156	149	116
2	Enrolment in upper secondary education, both sexes	UIS.E.3	173	174	175	171	182	182	175	180	177	177	174	179	177	165	143	7
3	GDP per capita (current US\$)	NY.GDP.PCAP.CD	224	224	228	228	229	229	230	229	228	227	228	228	224	225	219	218
4	Internet users (per 100 people)	IT.NET.USER.P2	221	222	224	218	221	223	222	229	228	227	227	229	227	226	223	223
5	Population, ages 15-24, total	SP.POP.1524.TO.UN	190	191	192	192	191	191	187	181	181	181	181	181	181	181	181	181
6	Population, total	SP.POP.TOTL	240	240	240	240	240	240	240	240	240	240	240	240	239	239	232	232

Corpus des variables métier exploitables sans difficultés apparentes.

---

2015 année non prise en compte

# Visualisation descriptive des indicateurs

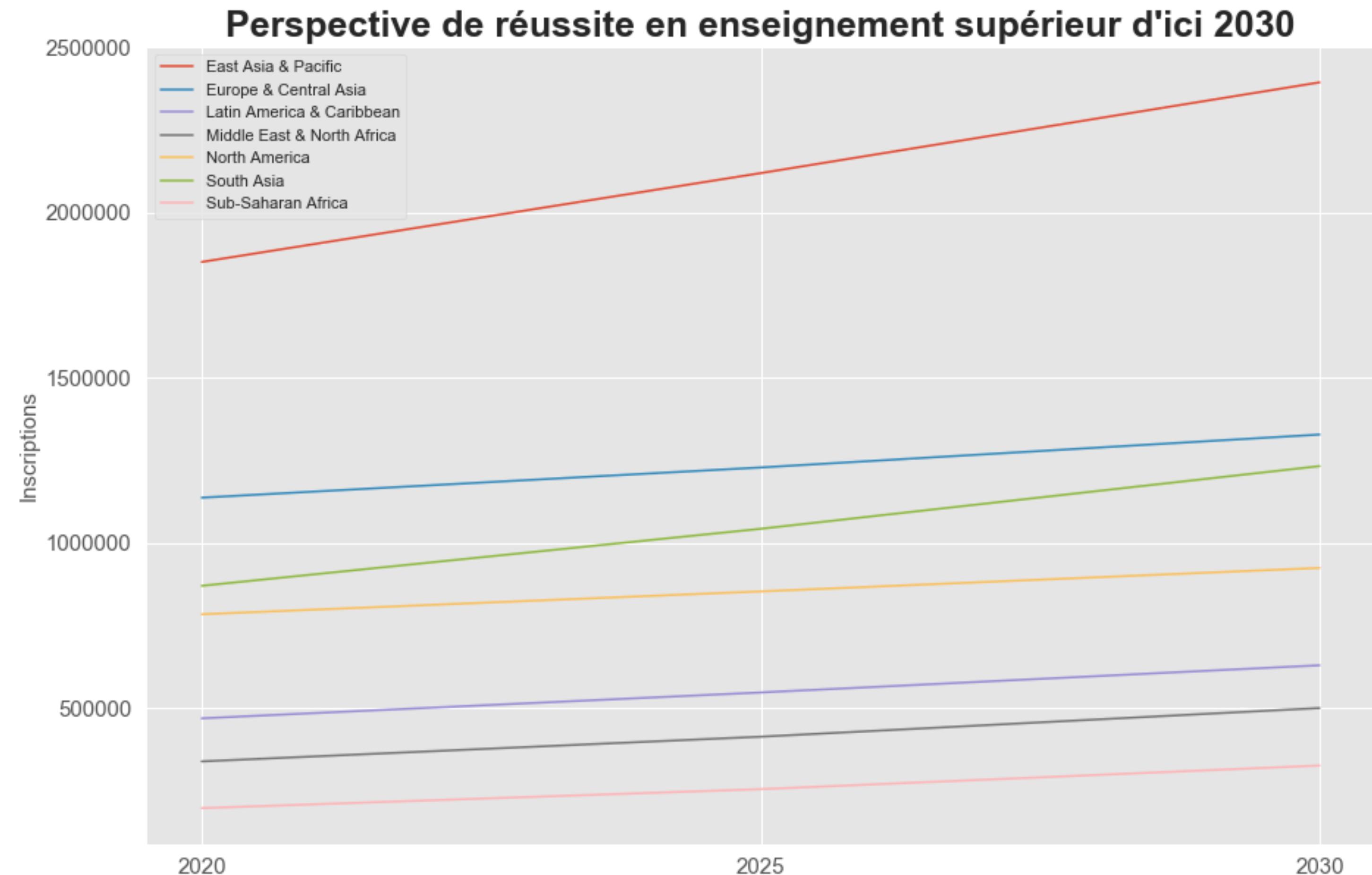
« Première approche par zones géographiques »

L'objectif ici est de comprendre comment se comportent nos indicateurs selon les blocs pays

*Une sélection de pays pourra être réalisée a posteriori*

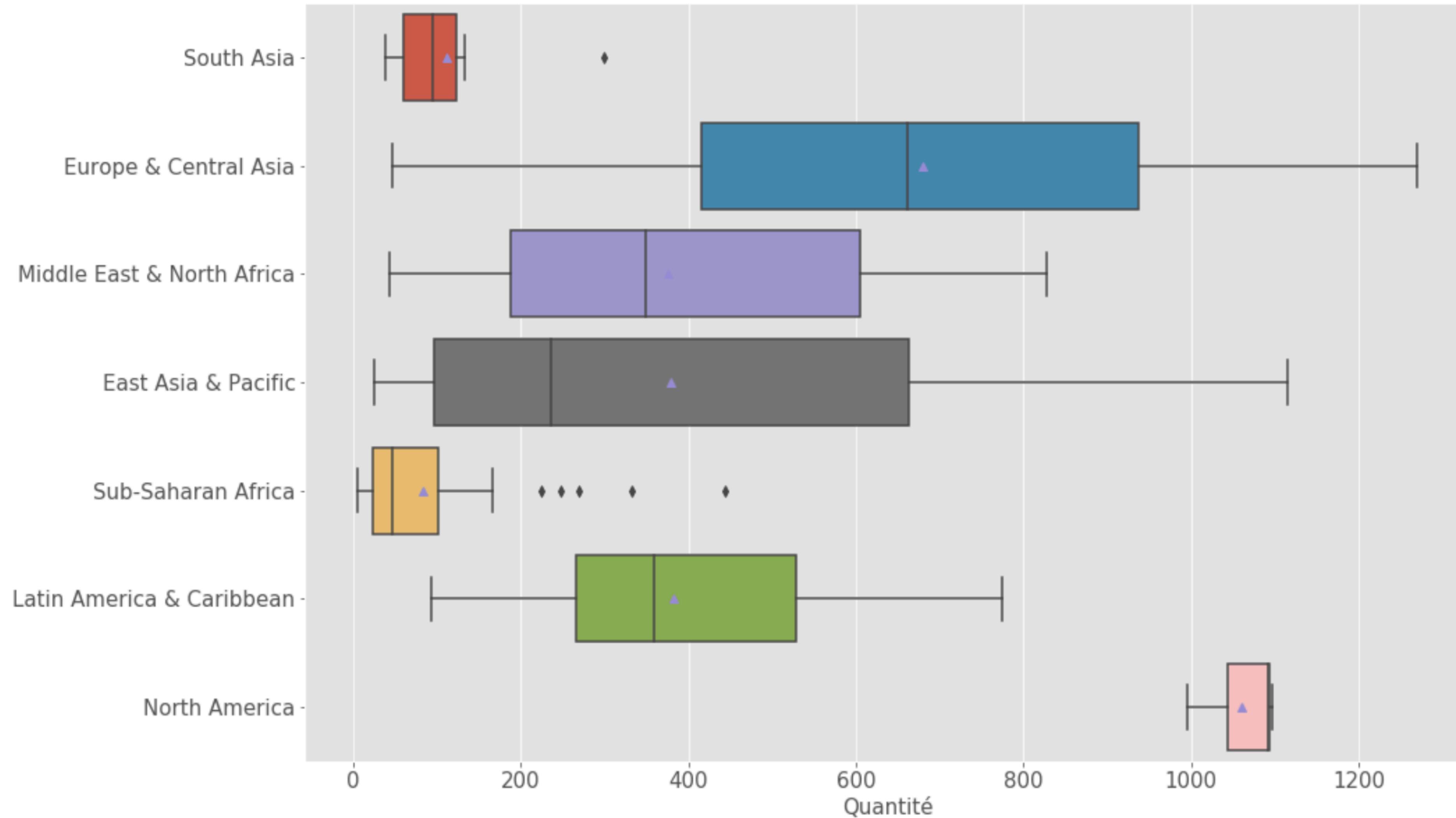
# Potentiel d'évolution à 10 ans

L'Asie et l'Europe  
répondent à un  
besoin éducatif plus  
« marqué » d'ici 2030

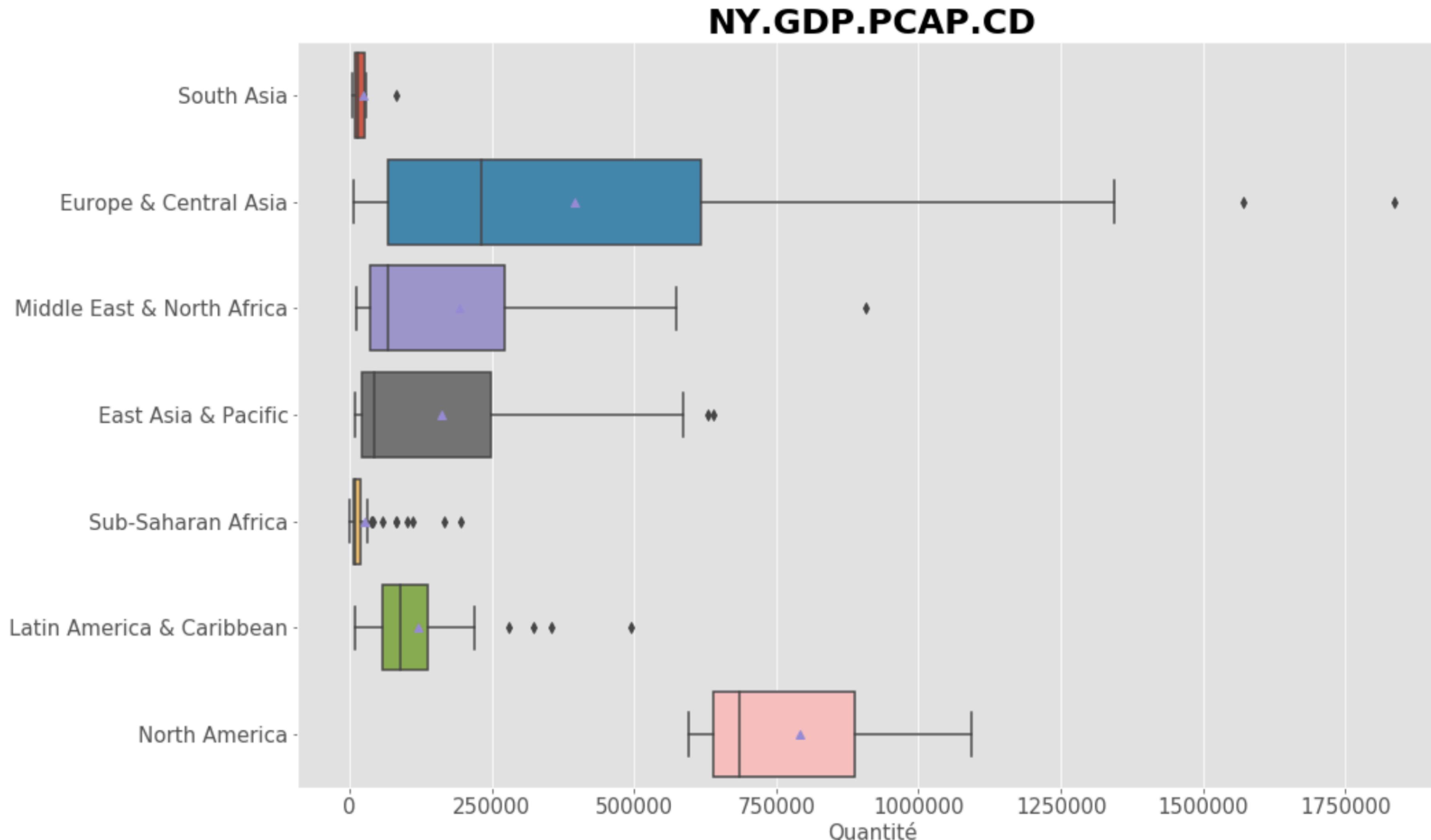


# Accès Internet

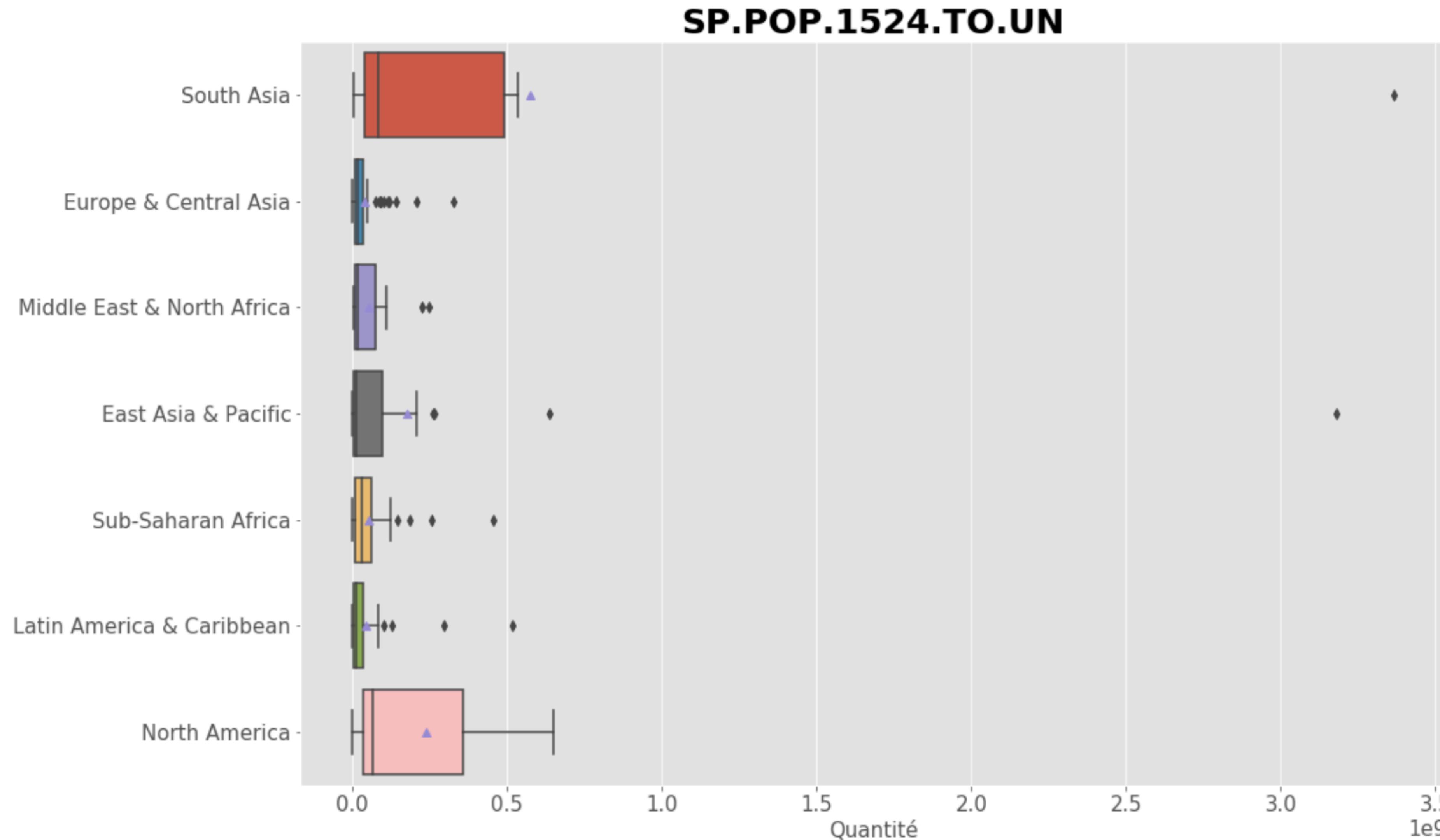
**IT.NET.USER.P2**



# PIB / habitant

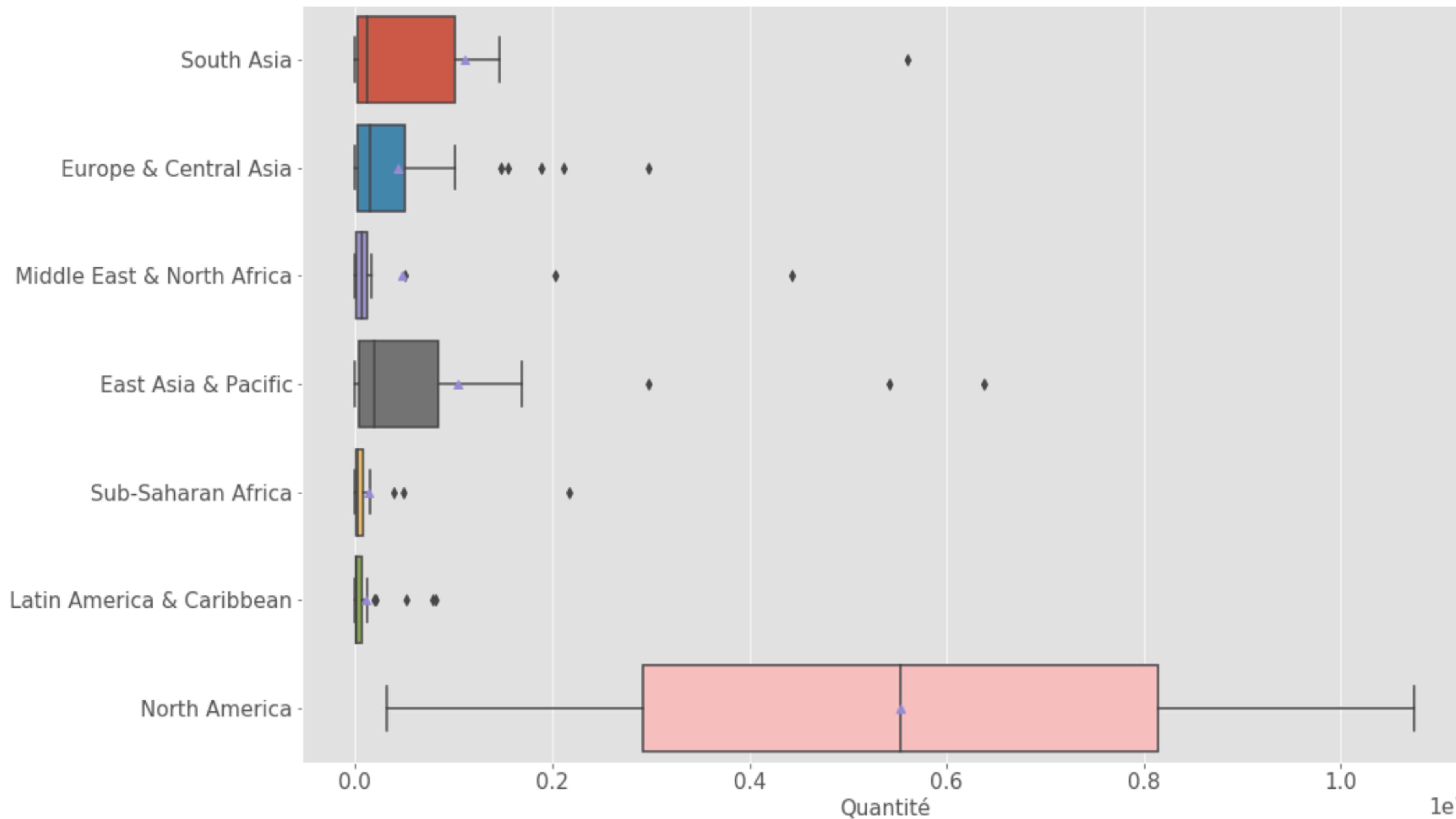


# Populations



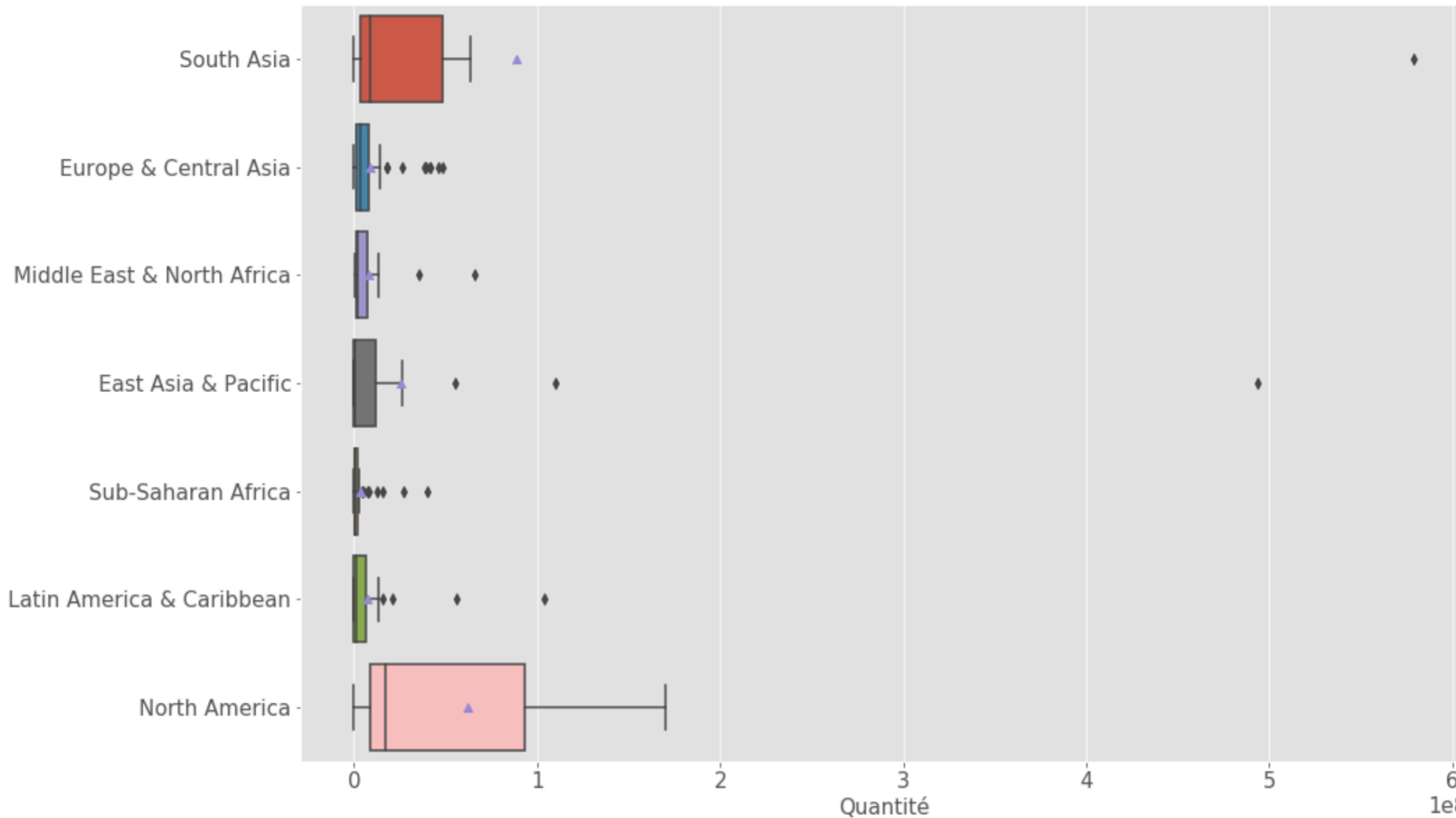
# Inscription dans l'enseignement post-secondaire non tertiaire H/F

**UIS.E.4**

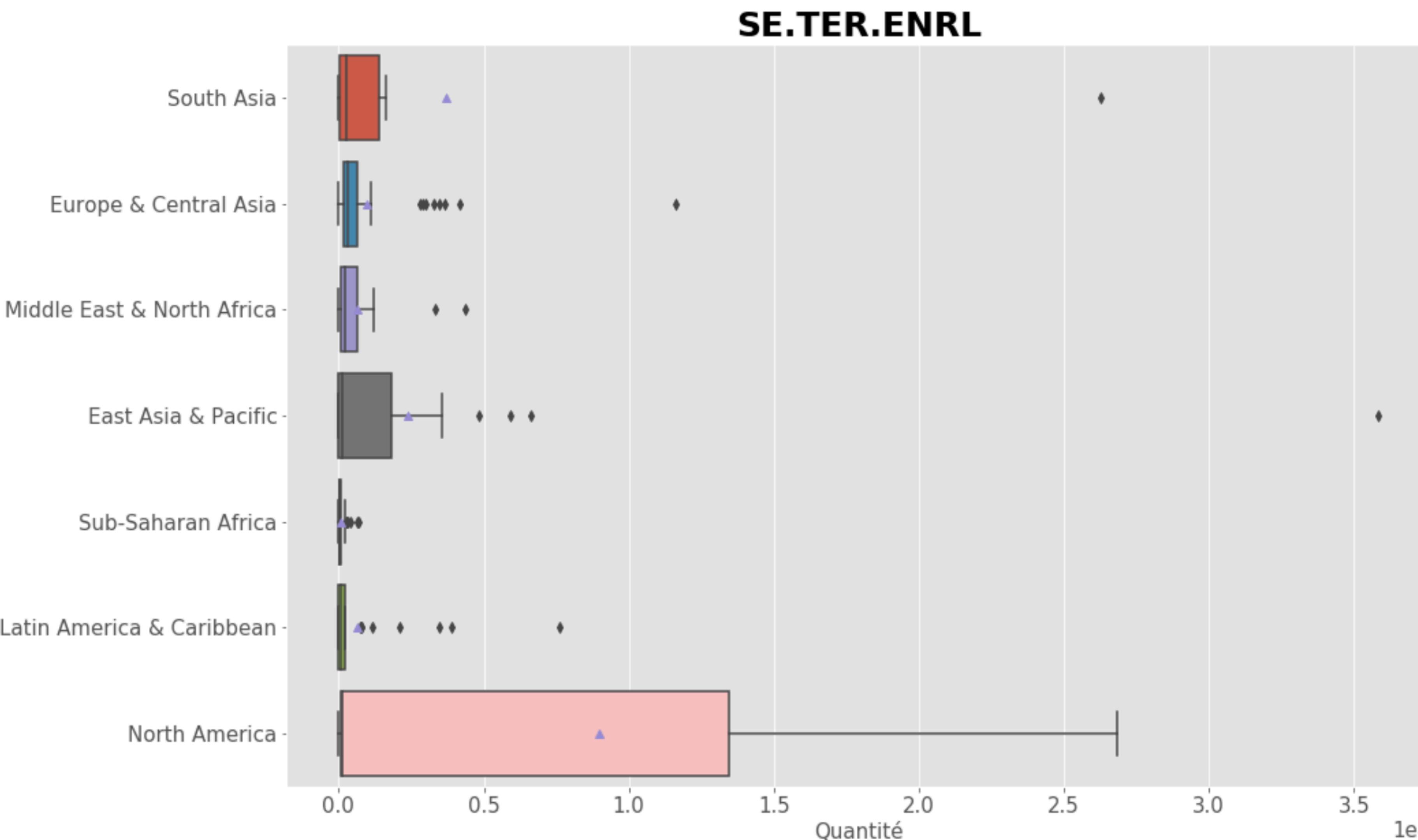


# Inscription dans l'enseignement secondaire supérieur H/F

**UIS.E.3**



# Inscription dans l'enseignement supérieur tous les programmes H/F



# Aide à l'analyse comparative des moyennes

Indicator Code	Region	IT.NET.USER.P2	NY.GDP.PCAP.CD	SE.TER.ENRL	SP.POP.1524.TO.UN	SP.POP.TOTL	UIS.E.3	UIS.E.4
0	East Asia & Pacific	379.299863	161901.941107	2.395051e+07	1.749074e+08	8.898658e+08	2.590999e+07	1.037147e+06
1	Europe & Central Asia	679.502809	396098.640789	9.538724e+06	3.858897e+07	2.315592e+08	8.933412e+06	4.377339e+05
2	Latin America & Caribbean	383.083264	120049.957562	6.188257e+06	4.449309e+07	2.103232e+08	7.411199e+06	1.122108e+05
3	Middle East & North Africa	375.564801	193611.084823	6.606025e+06	5.282638e+07	2.615392e+08	8.225265e+06	4.735776e+05
4	North America	1060.915132	790842.553716	8.981394e+07	2.381603e+08	1.669911e+09	6.241189e+07	5.531707e+06
5	South Asia	111.238647	23166.697887	3.715507e+07	5.751262e+08	2.921328e+09	8.870475e+07	1.114222e+06
6	Sub-Saharan Africa	82.825092	26654.910487	9.000697e+05	5.254422e+07	2.538349e+08	3.201480e+06	1.405623e+05

Version centrée réduite - facilité de comparaison

Region	IT.NET.USER.P2	NY.GDP.PCAP.CD	SP.POP.TOTL	SP.POP.1524.TO.UN	UIS.E.3	UIS.E.4	SE.TER.ENRL
East Asia & Pacific	-0.189944	-0.328872	-0.032084	0.037653	-0.031213	-0.108538	-0.127365
Europe & Central Asia	0.766497	0.602274	-0.530115	-0.715502	-0.718425	-0.659083	-0.464079
Latin America & Caribbean	-0.177890	-0.495272	-0.645897	-0.682882	-0.740593	-0.708447	-0.646938
Middle East & North Africa	-0.201844	-0.202799	-0.631461	-0.636841	-0.687128	-0.682047	-0.443944
North America	1.981669	2.171741	2.243971	0.387123	0.783083	1.075207	2.397405
South Asia	-1.043982	-0.880471	0.424228	2.248850	2.089447	1.927876	-0.084069
Sub-Saharan Africa	-1.134507	-0.866602	-0.828642	-0.638400	-0.695171	-0.844967	-0.631011

# Brèves remarques

Grande variance des variables propres à l'éducation

Valeurs atypiques en queue de distribution...

Vérification des pays concernés (*Chine, Inde, USA, etc*),  
aucune valeur aberrante identifiée

**Nos groupements de pays actuels ne permettent pas une « sélection pertinente » de pays, trop de disparités ressortent dans l'approche visuelle par boxplot.**

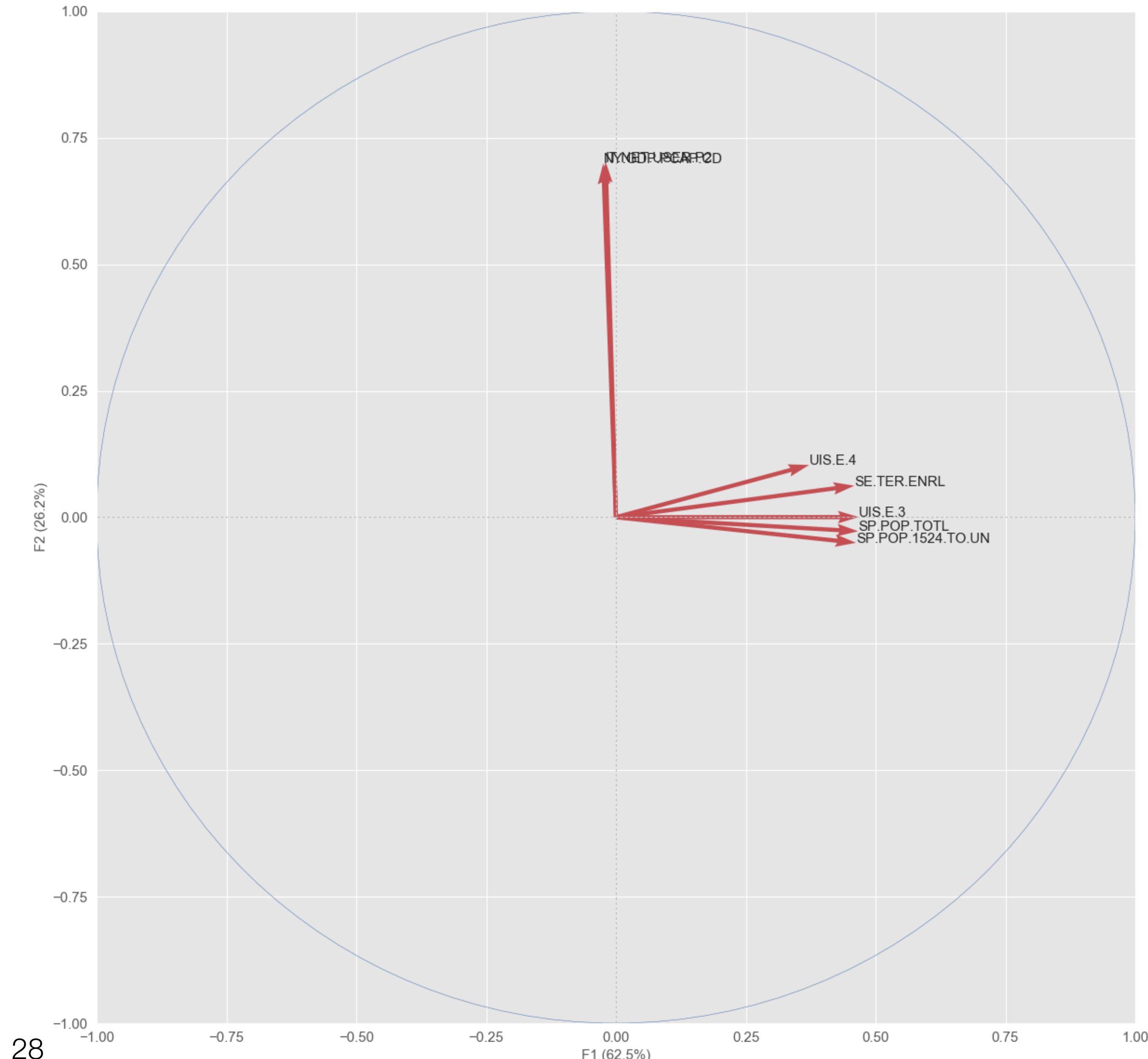
# **Analyse des pays**

**Identifier des patterns de pays aux propriétés similaires**  
**Exploration par apprentissage non-supervisé via une**  
**ACP et un clustering K-means**

# Cercle des corrélations

## Projection des variables sur le premier plan factoriel ACP

- Nouvelle base orthonormée
- Variance maximale 88%
- 2 types de profils pays
- Prise en compte des corrélations entre nos variables



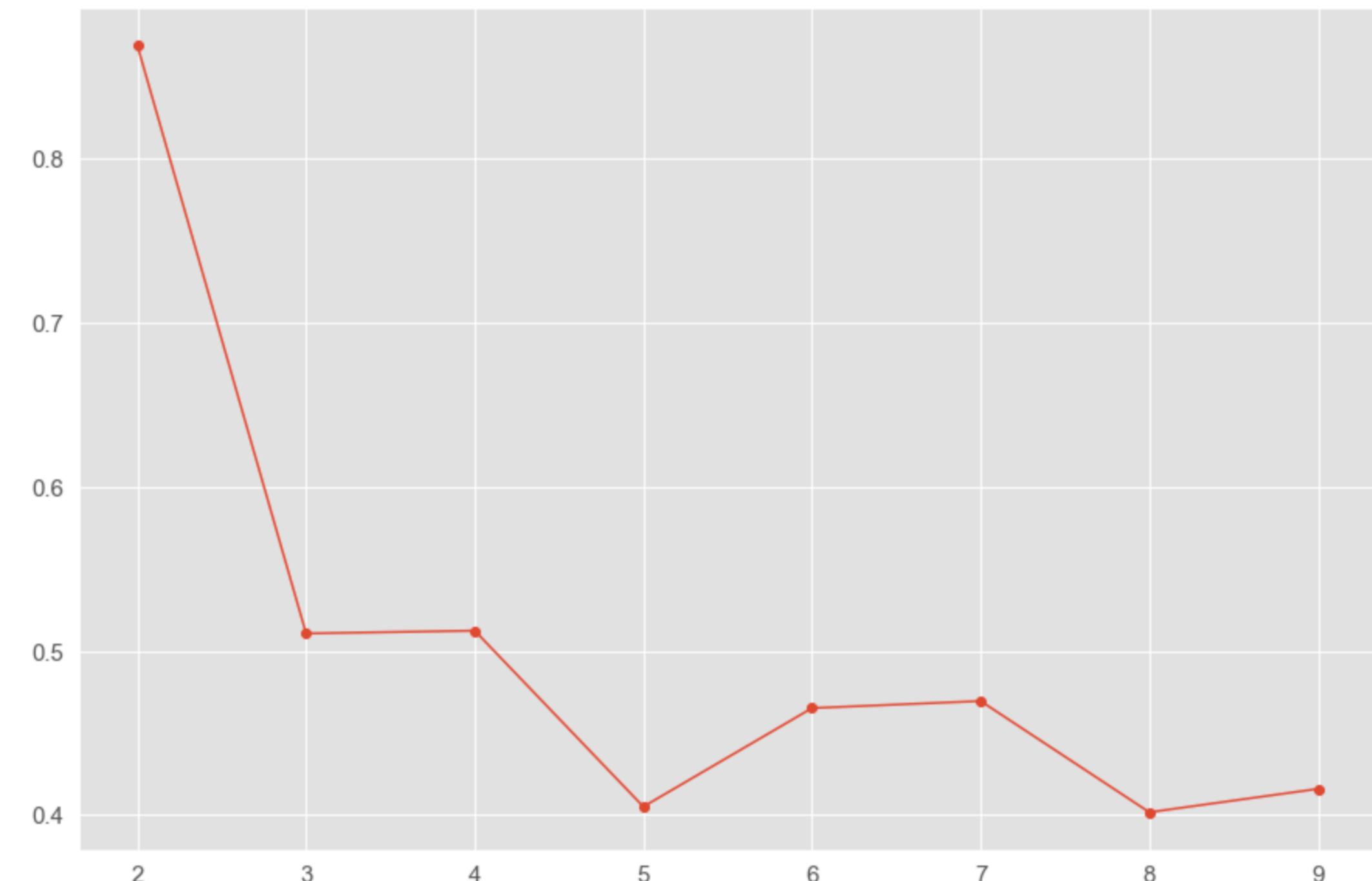
# Caractéristiques clusters K-means

	IT.NET.USER.P2	NY.GDP.PCAP.CD	SP.POP.TOTL	SP.POP.1524.TO.UN	UIS.E.3	UIS.E.4	SE.TER.ENRL
cluster0	1.258389	1.171855	-0.136778	-0.188526	-0.158090	-0.122061	-0.140692
cluster1	-0.664201	-0.609345	6.374361	7.724582	7.700869	7.583737	3.920354
cluster2	-0.560206	-0.523231	-0.146370	-0.107063	-0.123830	-0.143614	-0.110156
cluster3	1.549379	1.560635	5.466174	1.324915	1.660805	2.230974	7.340002

39 pays dans le cluster 0  
2 pays dans le cluster 1  
88 pays dans le cluster 2  
1 pays dans le cluster 3

## Découpage réalisé en 4 clusters

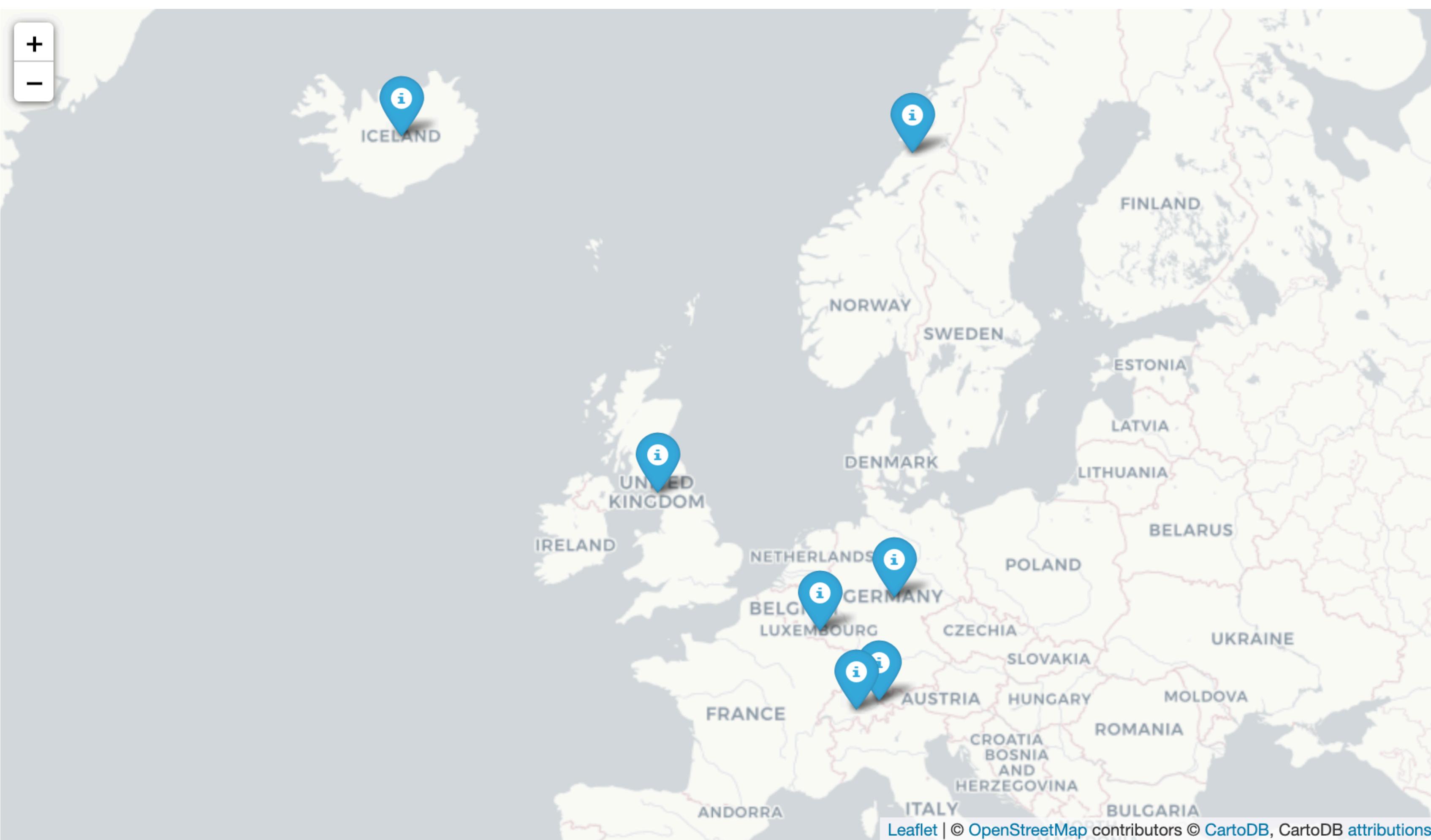
la métrique Silhouette expose un bon équilibre pour ce clustering...



# **Pays cibles**

**Recommandations basées sur des critères socio-démographiques et liés au contexte métier de l'entreprise**

**Sélection de pays tenant compte des corrélations entre les variables de l'échantillon...**



# Pays prioritaires de l'U.E

Liechtenstein, Luxembourg, Norvège, Suisse, Islande, Suisse, Allemagne, Royaume-Uni



# Vision élargie

Pays « non prioritaires » de couleur orange :  
Canada, Australie, Nouvelle-Zélande, Qatar, Japon

# Conclusion

**Les données sur l'éducation de la banque mondiale permettent une première orientation pour le projet d'expansion.**

---

**Il est désormais impératif d'avoir une approche Benchmark des zones choisies : observer, analyser, comparer... la concurrence et les leaders du marché.**