“"Analyse des données de systèmes éducatifs"”

Projet 8 : Data Analyst

Réalisé par : Pouria Forouzesh

Mentor: Alexendre Gazagnes

Avril – 2022

Contents

Contexte

J'ai choisi le projet 2 de Data scientist d'OpenClassrooms pour avoir l'occasion à répéter tout ce que j'ai apprise pendant le 7 projet précédents.

Le lien du projet: https://openclassrooms.com/fr/paths/164-data-scientist#main_content

Scenario:

Data Scientist dans une 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 client?
- Quels pays l'entreprise doit-elle opérer en priorité?

Les données de la Banque mondiale sont disponibles à l'adresse suivante :

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

Ou en téléchargement direct dans le dossier dataset.

Analyse des données

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

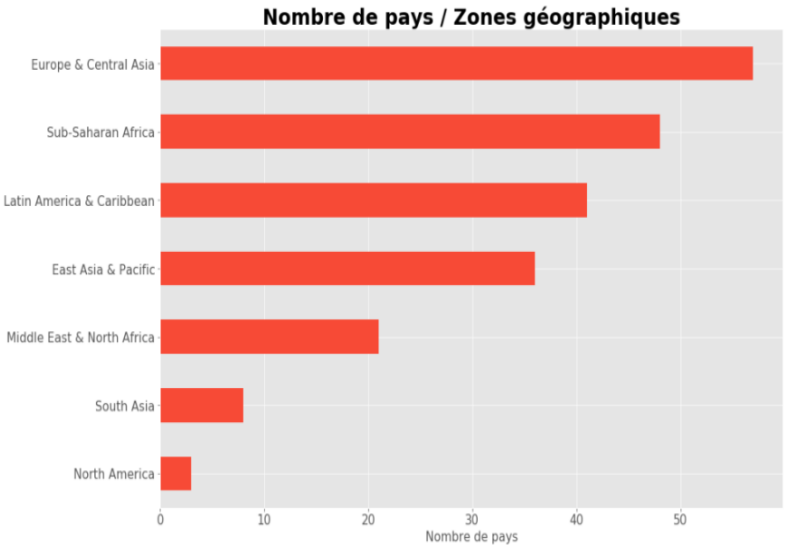
Description rapide des données

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

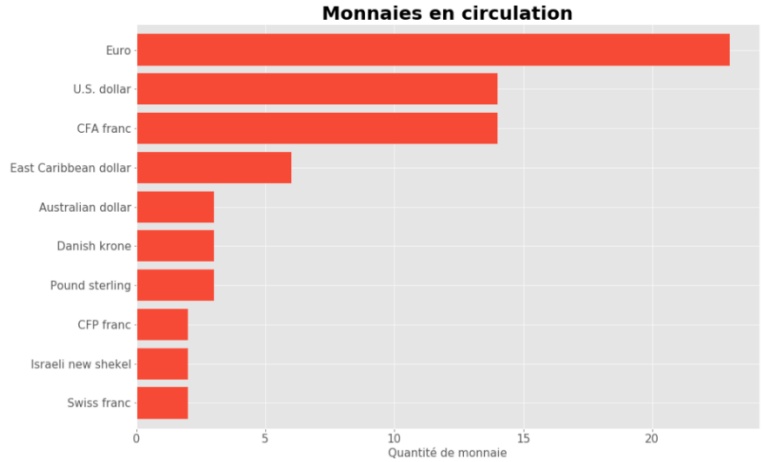
Analyses et graphiques, Pays, zones géographiques, devises monétaires, mots clés, indicateurs statistiques

Pays et zones géographiques



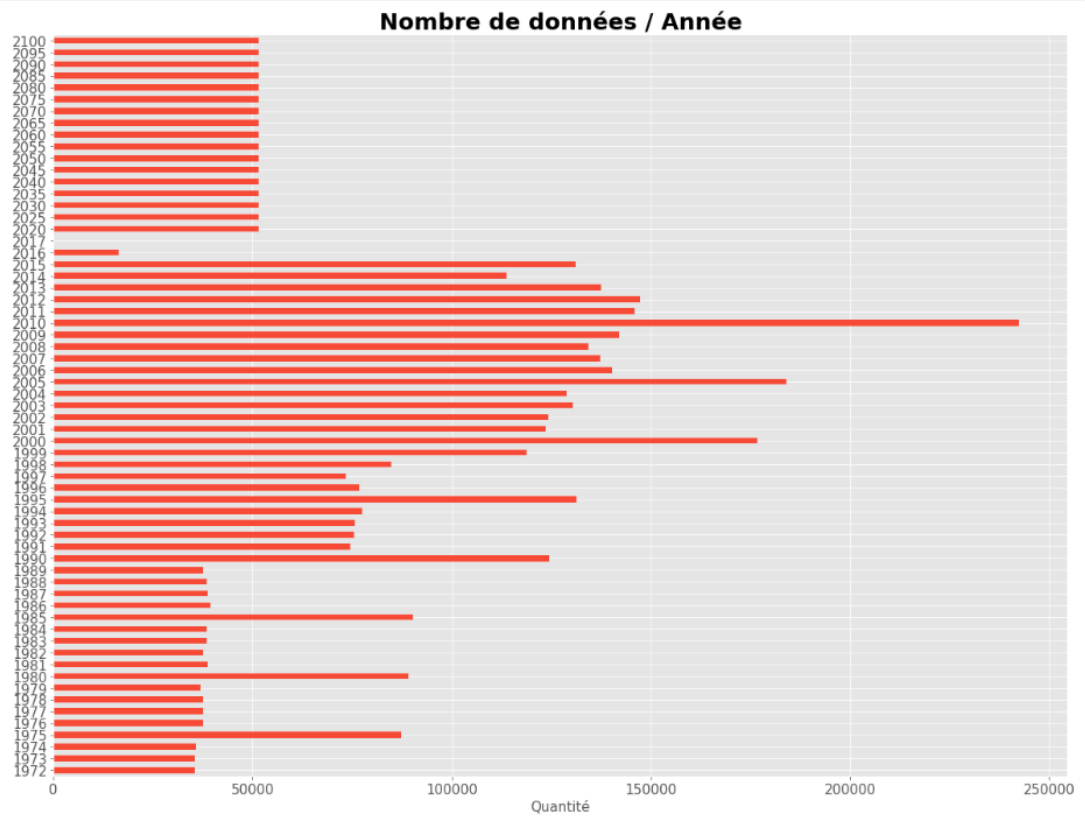
- Extrait du fichier EdStatsCountry.csv
- La plupart de pays sont situés en Europe (plus de 50 pays)
- Moins d 10 pays sont situés en Amérique du Nord

Devises monétaires



- Extrait du fichier EdStatsCountry.csv
- La monnaie officielle de plus de 20 pays est l'euro

Choix des années



- Les années ne sont pas toutes significatives, elles restent en grande majorité inexploitables du fait des valeurs manquantes.
- Notons que les indicateurs après 2015 ne sont pas exploitables, à exclure de l'analyse. .
- Suivant le contexte de l'analyse et les objectifs souhaités, on peut distinguer également les années avant 2000 comme étant moins qualitatives

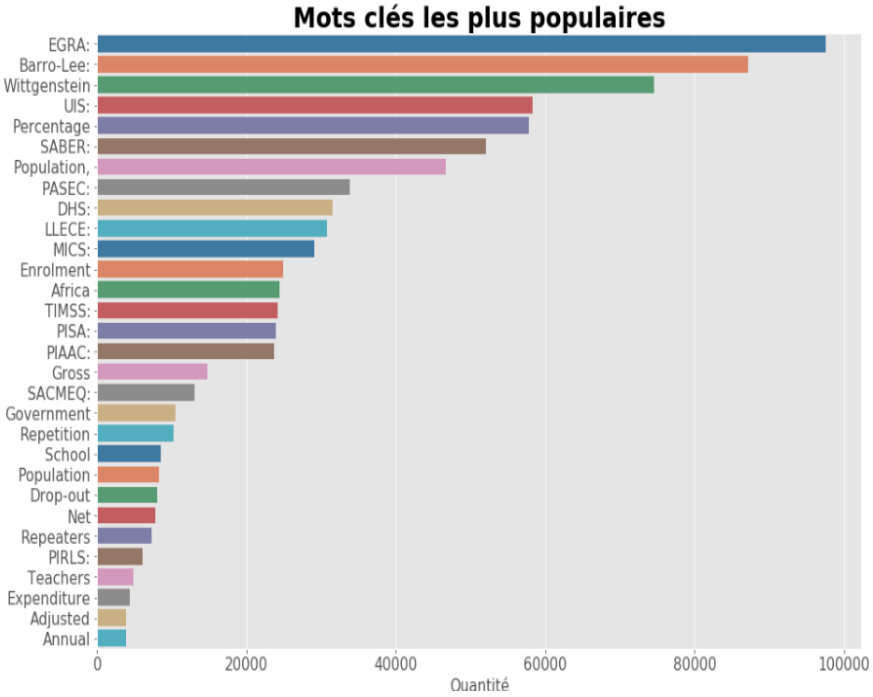
Catégories d'indicateurs

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)
```

- Extrait du fichier EdStatsSeries.csv

Mots clés liés aux indicateurs



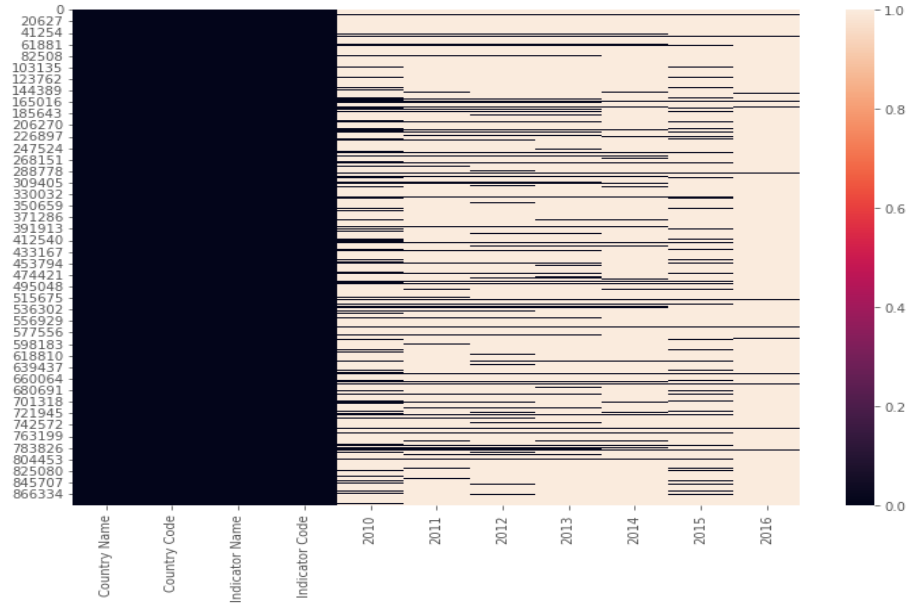
Les mots clés les plus représentatifs des 3665 indicateurs sont évocateurs du secteur de l'éducation.

- 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, School, etc…

Cet échantillon est d'une importance capitale pour répondre aux attentes de l'entreprise. Un historique sur plusieurs années semble être exploitable, avec des indicateurs propres à l'éducation et des observations selon l'ensemble des pays de Monde.

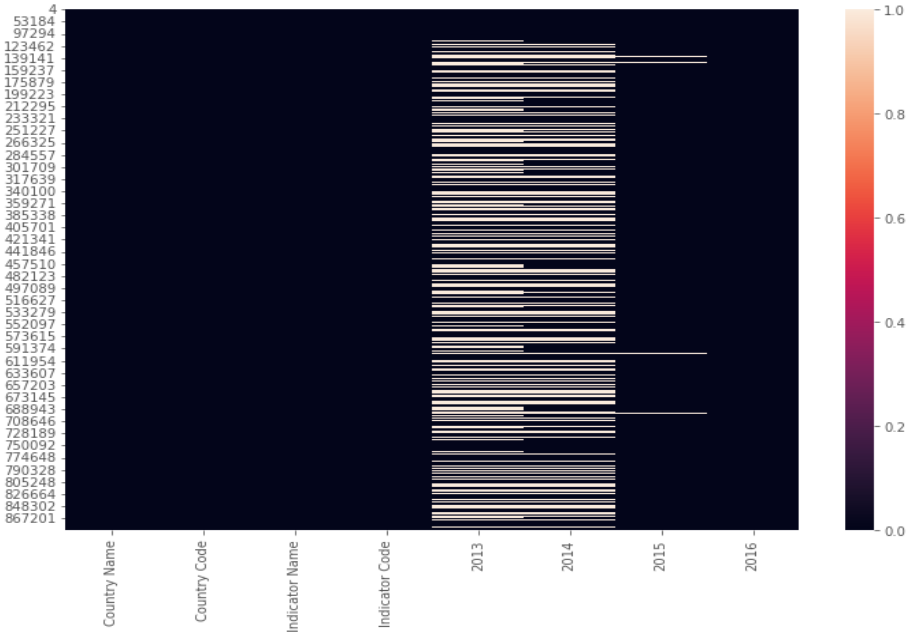
Analyse des données

Préparation de DataFrame



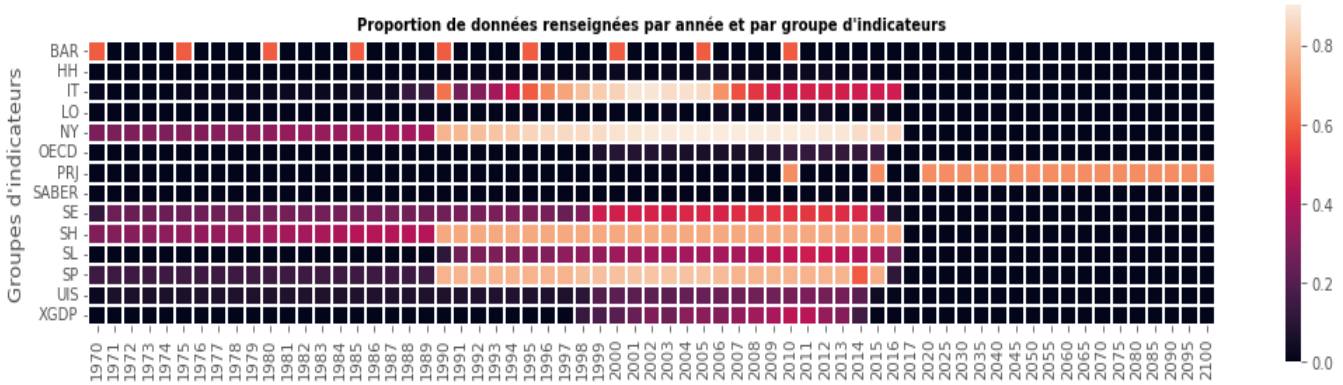
- On remplit artificiellement les valeurs manquantes des colonnes par les valeurs de colonne la plus proche par la méthode ffill.

- On constate qu'il y a plus de valeur manquante dans la colonne 2016



Choix des indicateurs

Heatmap des données présentes par groupe d'indicateurs



- **SP.SEC.TOTL.IN:** Population d'âge officiel pour l'enseignement secondaire
- **SP.TER.TOTL.IN**: Population d'âge officiel pour l'enseignement supérieur
- **SP.POP.GROW**: Croissance démographique (% annuel)
- **SE.SEC.ENRL:** Inscription dans l'enseignement secondaire
- **SE.TER.ENRL**: Inscription dans l'enseignement supérieur
- **NY.GDP.PCAP.CD**: PIB par habitant (USD courants)
- **IT.NET.USER.P2**: Internautes (pour 100 personnes)
- **SP.POP.TOTL**: Population totale
- **PRJ.POP.ALL.4.MF**: l'indicateur donne une projection sur l'évolution de la réussite en enseignement supérieur

- On sélectionne uniquement les indicateurs qui sont renseignés pour plus de 100 pays
- On sélectionne uniquement les pays ayant plus d 500 mille habitants

Vérification des valeurs disponibles

```
Nan --> 58
```

	IT.NET.USER.P2	NY.GDP.PCAP.CD	SE.SEC.ENRL	SE.TER.ENRL	SP.POP.GROW	SP.POP.TOTL	SP.SEC.TOTL.IN	SP.TER.TOTL.IN
0	66.363445	4124.982390	315079.0	160527.0	-0.159880	2876101.0	311514.0	276247.0
1	42.945527	3916.881571	NaN	1289474.0	1.825463	40606052.0	4056674.0	3492401.0
2	13.000000	3308.700233	NaN	221037.0	3.367572	28813463.0	3691322.0	2374694.0

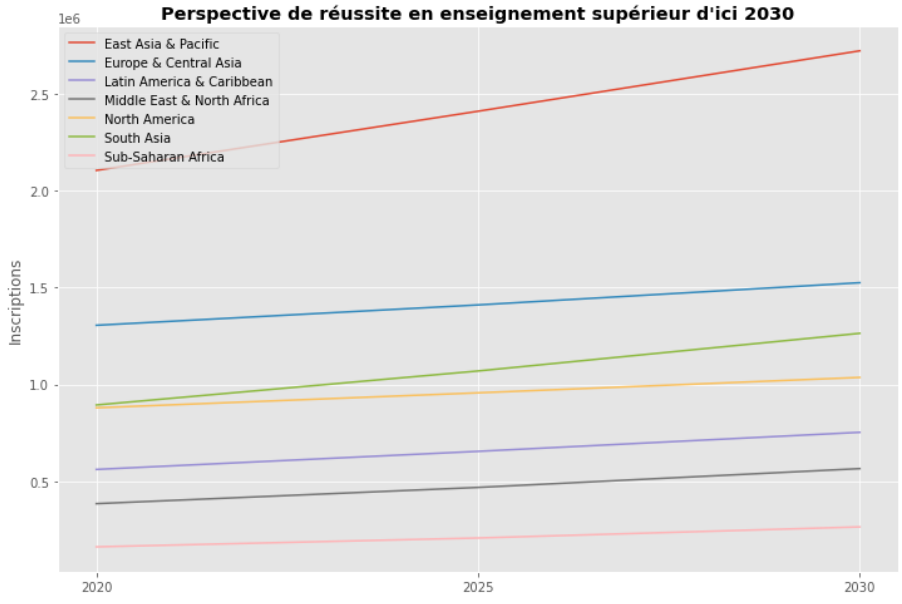
- Il y a de la valeur manquant dans DataFrame, on remplit par la moyenne de leurs 2 voisins plus proches. (KNNImputer)

```
(127, 9)
Nan --> 0
```

	Country Name	Region	IT.NET.USER.P2	NY.GDP.PCAP.CD	SE.SEC.ENRL	SE.TER.ENRL	SP.POP.GROW	SP.SEC.TOTL.IN	SP.TER.TOTL.IN
124	Vietnam	East Asia & Pacific	46.500000	2214.387662	7166669.25	2466643.0	1.071293	9377680.0	8554144.0
125	West Bank and Gaza	Middle East & North Africa	61.178385	2943.404534	721414.00	221018.0	2.884693	869213.0	499101.0
126	Zambia	Sub-Saharan Africa	25.506579	1269.573537	1576502.50	476389.5	3.002816	1882524.0	1497656.0

Visualisation descriptive des indicateurs

Potentiel d'évolution à 10 ans



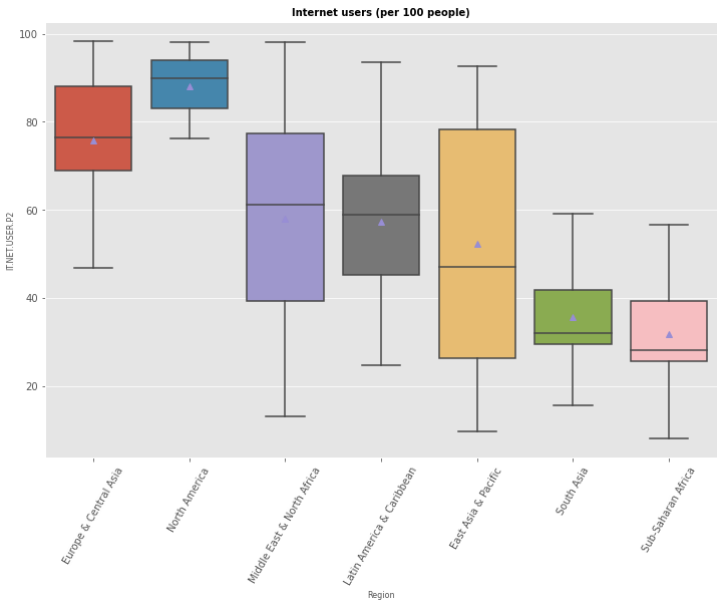
- L'Asie et l'Europe répondent à un besoin éducatif plus « marqué » d'ici 2030

Indicateur sur les ressources informatiques

Première approche par zones géographiques

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

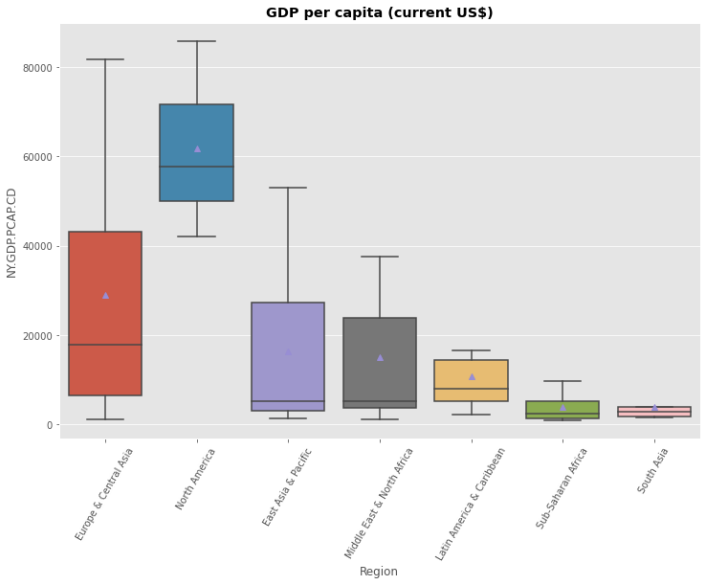
Accès Internet



- Il y a de fortes inégalités entre régions géographiques mais également au sein d'une même région

Indicateur sur la Richesse des pays

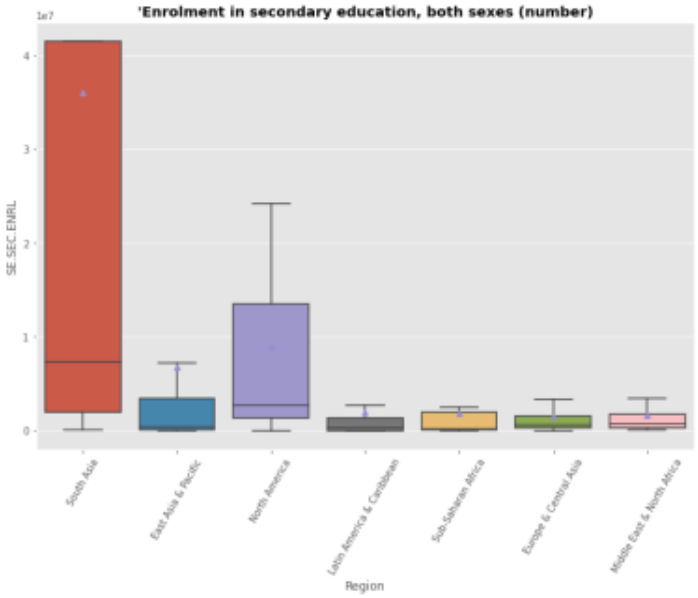
PIB / habitant



- Les inégalités de PIB sont très marquées d'une région à l'autre

Indicateurs relatifs à l'éducation

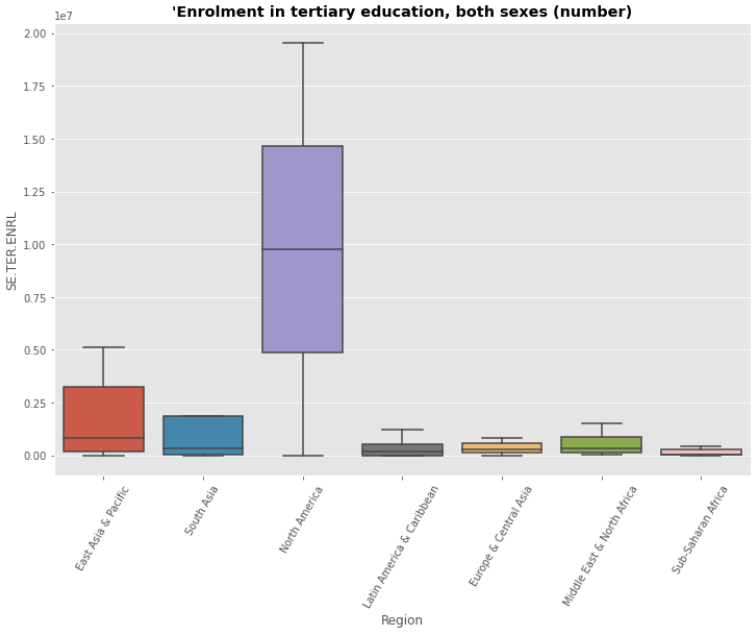
Inscription dans l'enseignement post-secondaire non tertiaire H/F



- Les dispersions sont très marquées d'Asie du Sud

- Asie du Sud a une moyenne très élevée par rapport aux autres régions

Inscription dans l'enseignement secondaire supérieur H/F

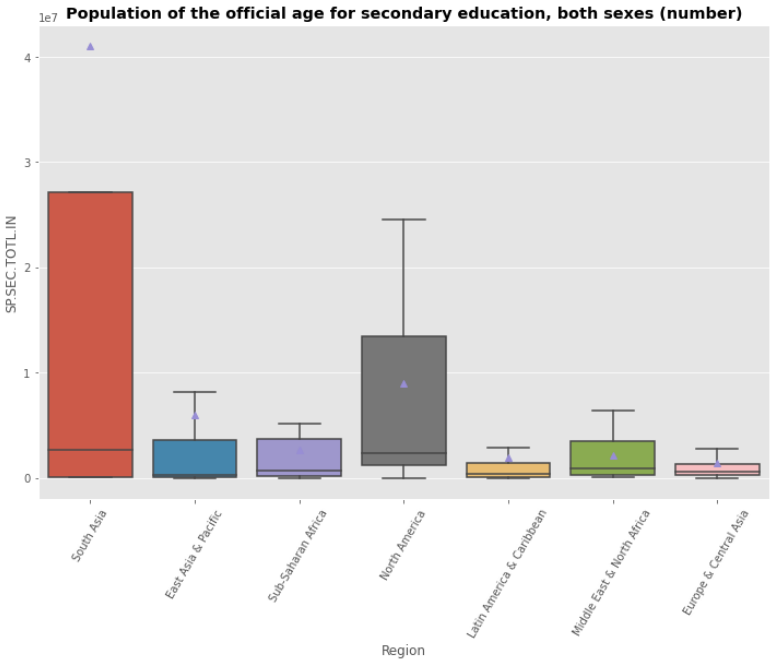


- Les dispersions sont très marquées d'Amérique du Sud

- Pour les autres régions, la valeur du médian est proche d'un à l'autre

Indicateurs sur la population

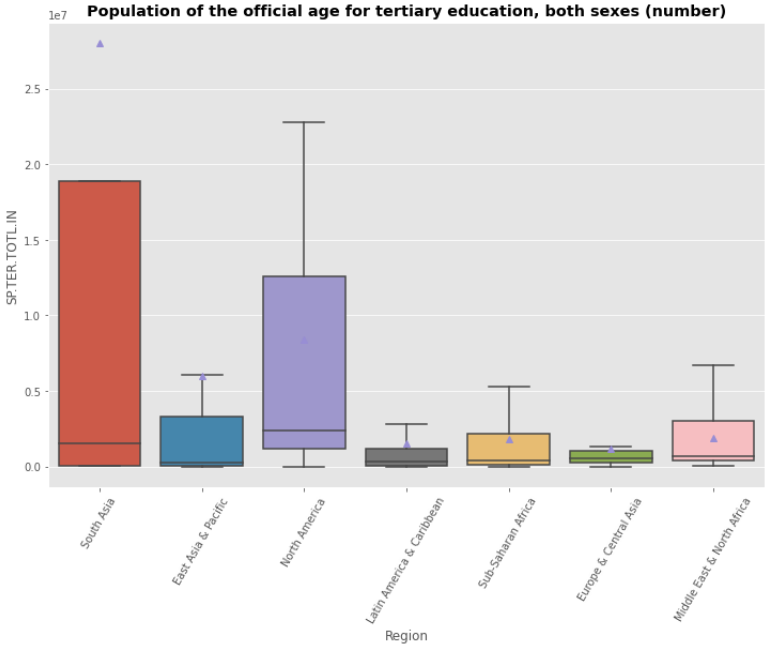
Population / l'enseignement secondaire



- Les médians de régions ne sont pas très différents.

- Asie du Sud a une moyenne très élevée par rapport aux autres régions.

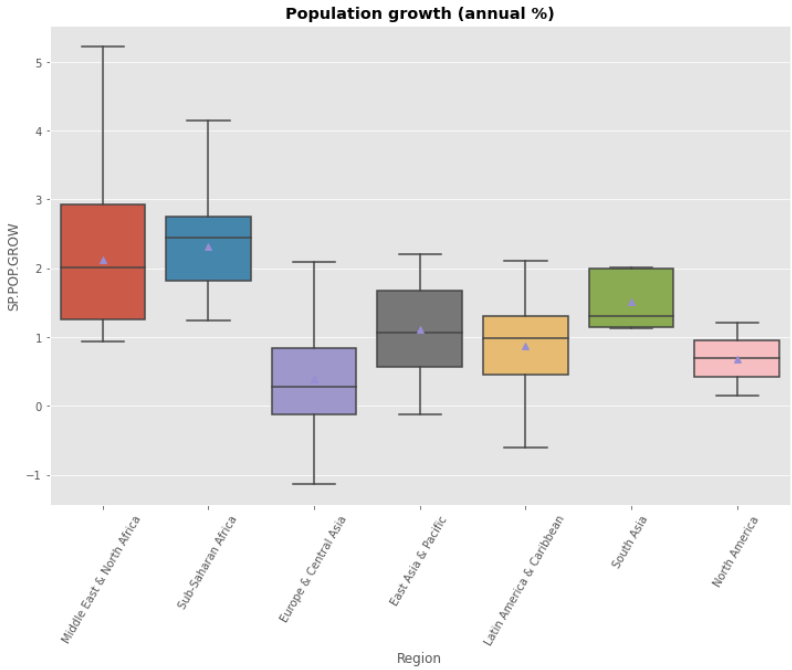
Population / l'enseignement supérieur



- Les médians de régions ne sont pas très différents.

- Asie du Sud a une moyenne très élevée par rapport aux autres régions.

Croissance démographique (% annuel)



- La croissance démographique pour l'Afrique et le Moyen-Orient est fortement positive.

- Certains pays en l'Europe et Asie centrale et Asie de l'Est ont la croissance démographique négative.

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.

Les figures sur TABLEAU:

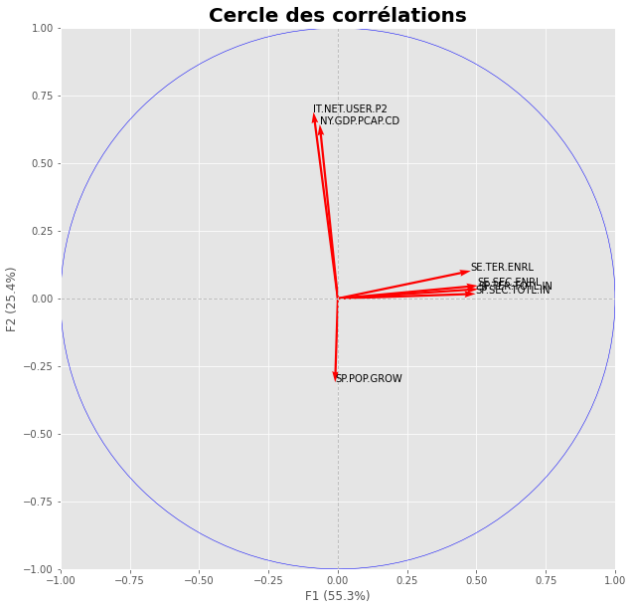
https://public.tableau.com/app/profile/pouria.forouzesh2812

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

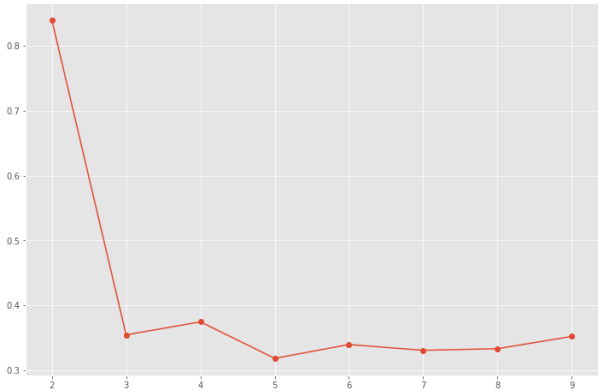
➔ Nouvelle base orthonormée

➔ Variance maximale 80%

➔ 3 types de profils pays

➔ Prise en compte des corrélations entre nos variables

Caractéristiques Clusters K-means

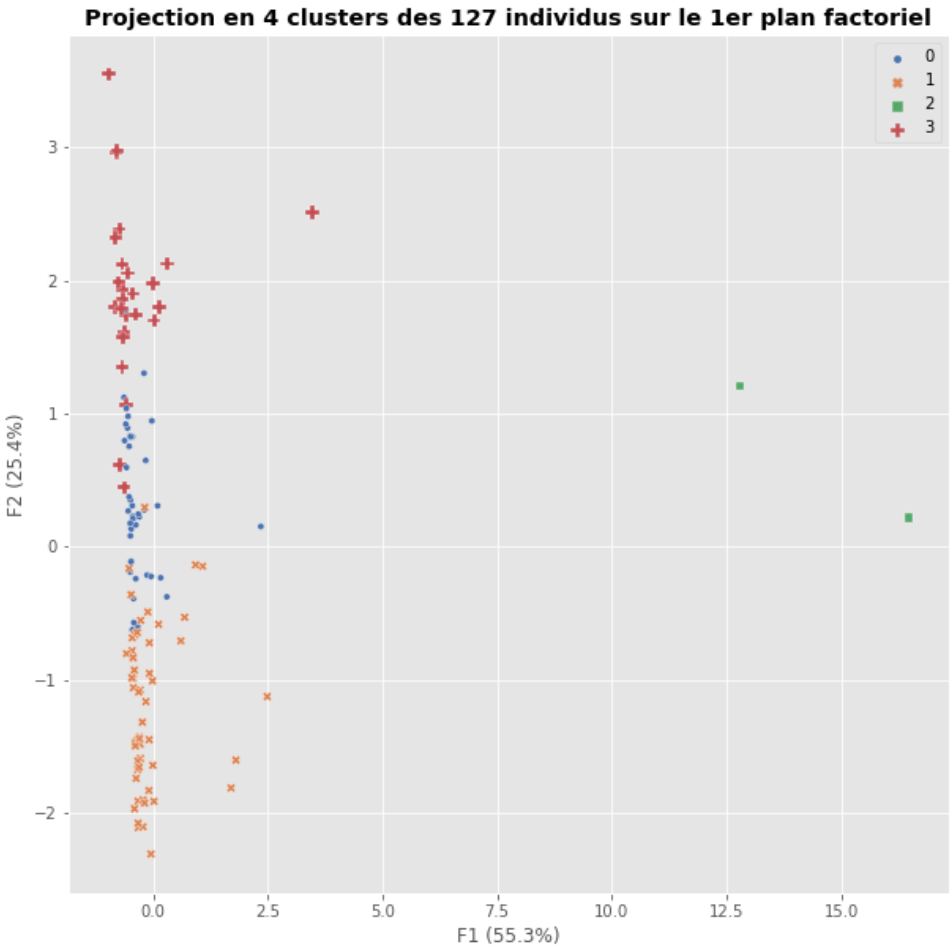


Découpage réalisé en 4 Clusters

La métrique Silhouette expose un bon équilibre pour ce Clustering…

```
44 pays dans le cluster 0
55 pays dans le cluster 1
2 pays dans le cluster 2
26 pays dans le cluster 3
```

Projection en 4 Clusters des individus sur le 1er plan factoriel



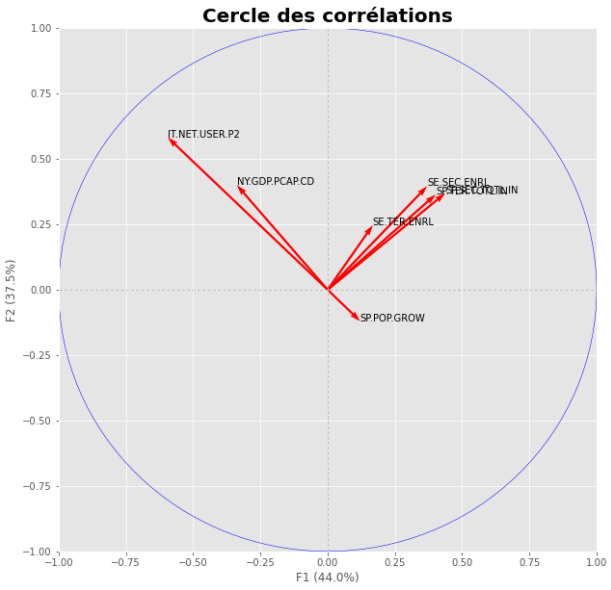
- Cluster 3 est pertinent

	Country Name	Region	Clusters	potentiel	rank
121	United States	North America	3	9.926093	1.0
67	Luxembourg	Europe & Central Asia	3	5.824735	2.0
93	Qatar	Middle East & North Africa	3	4.595495	3.0
110	Switzerland	Europe & Central Asia	3	3.545182	4.0

Analyse des pays sans les deux pays très dispersés

On supprime la Chine et L'inde pour avoir des pays plus harmonisé

Cercle des corrélations



Projection des variables sur le premier plan factoriel ACP

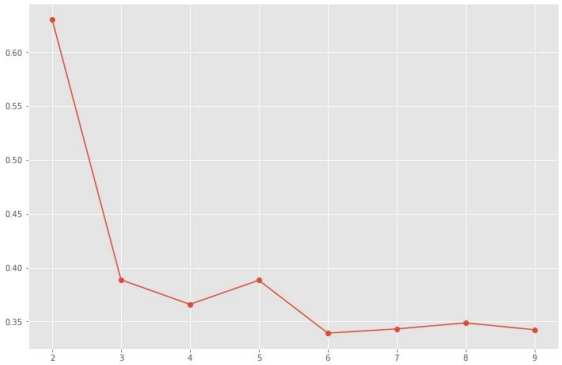
➔ Nouvelle base orthonormée

➔ Variance maximale 81%

➔ 3 types de profils pays

➔ Prise en compte des corrélations entre nos variables

Caractéristiques Clusters K-means

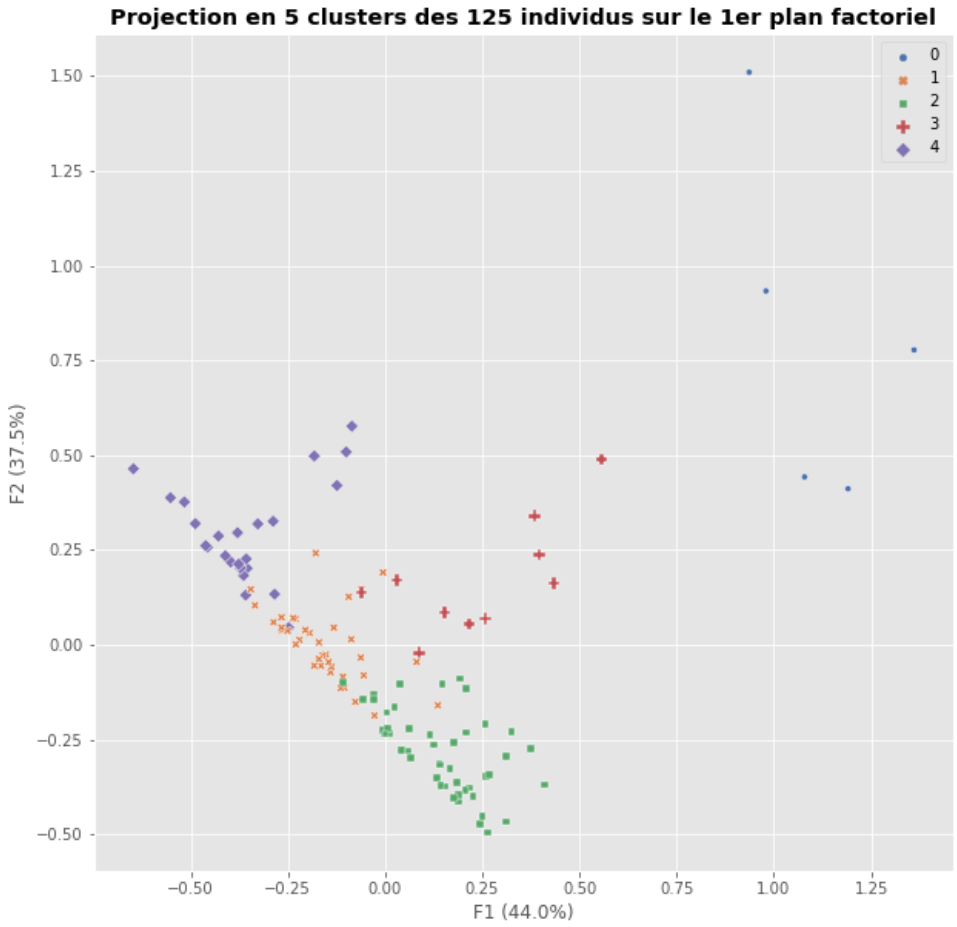


Découpage réalisé en 5 Clusters

La métrique Silhouette expose un bon équilibre pour ce Clustering…

```
5 pays dans le cluster 0
38 pays dans le cluster 1
47 pays dans le cluster 2
10 pays dans le cluster 3
25 pays dans le cluster 4
```

Projection en 5 Clusters des individus sur le 1er plan factoriel



- Cluster 4 est pertinent

	Country Name	Region	Clusters	potentiel_2	rank
119	United Arab Emirates	Middle East & North Africa	4	5.555846	1.0
79	New Zealand	East Asia & Pacific	4	3.191098	2.0
82	Norway	Europe & Central Asia	4	3.005169	3.0
36	Finland	Europe & Central Asia	4	2.258755	4.0

Recommandations des pays à cibler…

Recommandations basées sur des critères sociodémographiques et liés au contexte métier de l'entreprise.

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

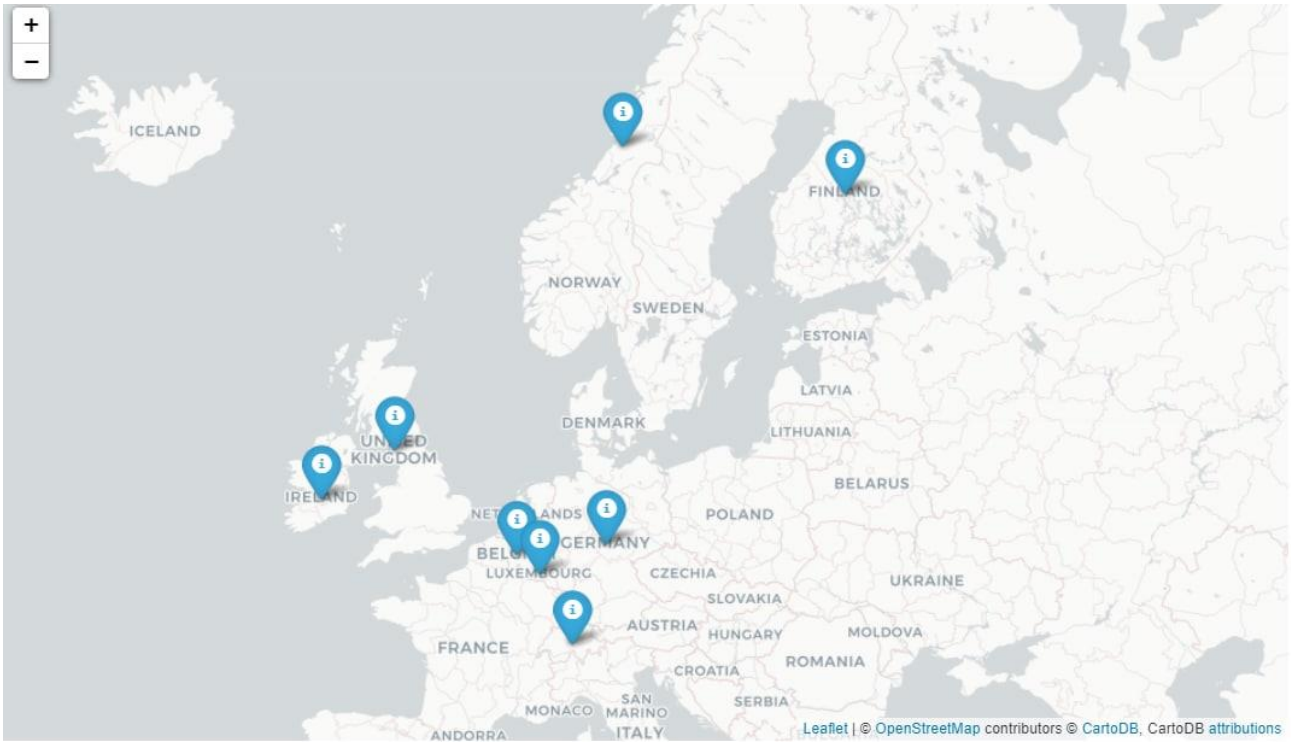
Pays prioritaires de l'U.E

Première solution, un développement "plus rapide" tourné vers l'Union Européenne :

A partir des hypothèses proposées et de leur représentation dans l'inertie totale, les pays cibles prioritaires seraient :

Luxembourg, Norvège, Finlande, Suisse, Belgique, Irlande, Allemagne, (Royaume-Uni)

Ces pays sont non seulement dans l'Union Européenne (facilité monétaire, culturelle, etc…), et répondent à une position dominante en terme de pouvoir d'achat de la population.



Vision élargie

Seconde solution, un développement en dehors de l'Union Européenne :

A partir des hypothèses proposées et de leur représentation dans l'inertie totale, les pays cibles seraient :

Canada, Australie, Nouvelle-Zélande, Qatar, Japon.

A noter également, un potentiel marché à affiner pour **la Chine, l'Inde, et les USA**.

Ces pays sont plus difficiles d'accès de part leur différence culturelle, une barrière linguistique possible, mais répondent à un potentiel métier intéressant.

Autre solution prenant en compte une perspective d'évolution sur d'ici 2030 :

Très fort potentiel pour **la Chine, l'Inde, et les USA** . Ceci étant, les pays suivants, avec une pénétration marché plus facile, seraient à travailler pour un positionnement d'ici 2030 : **Allemagne, Royaume-Uni, Canada, 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