academy

Analyse exploratoire de données

Agenda

1. Contexte
2. Approche
3. Présentation du jeu de données
4. Sélection des données
5. Analyse
6. Prochaines étapes
7. Annexe : l’environnement technique

1) Contexte

Expansion à l'international

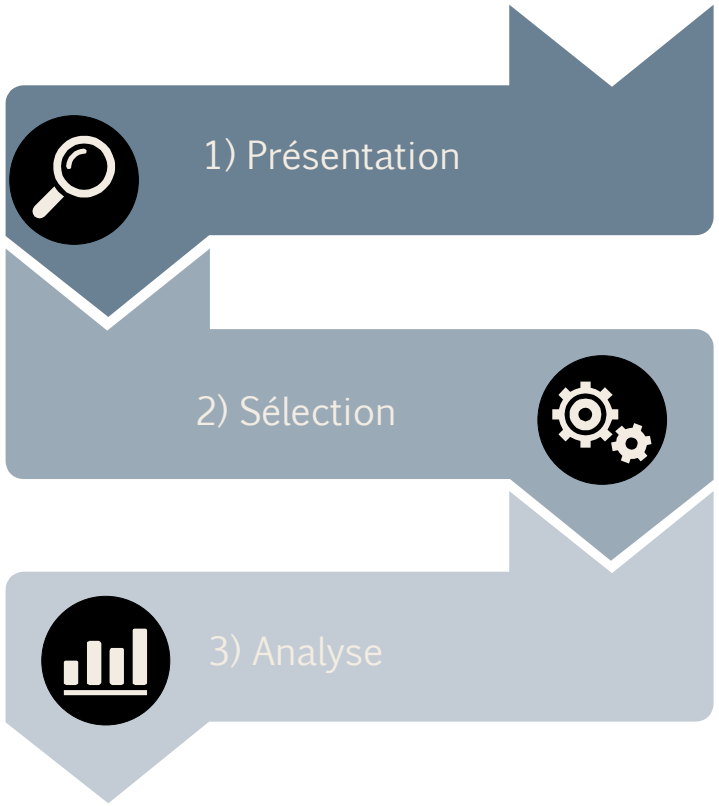
Academy est un EdTech qui propose des contenus de formation en ligne pour un public de niveau lycée et université. Elle cherche à s'étendre à l'international.

Analyse

La banque mondiale met à disposition des données sur l 'éducation, l'objectif est de voir si ces dernières permettent de répondre aux questions suivantes :

1. Quels sont les pays avec un fort potentiel de clients pour nos services ?
2. Pour chacun de ces pays, quelle sera l'évolution de ce potentiel de clients ?
3. Dans quels pays l'entreprise doit-elle opérer en priorité ?

2) Approche

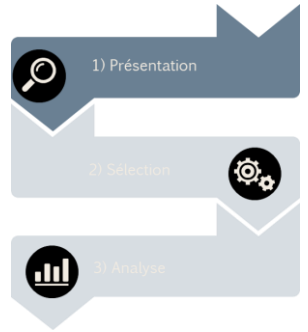


Présentation du jeu de données (Taille du jeu de données, Type de données, nombre de lignes, nombre de colonnes, ...)

Définition de la méthodologie appliquée pour sélectionner les données pertinentes afin d'arriver à des données exploitables (SmartData).

Processus iteratif sur base de la SmartData permettant de répondre aux questions demandées.

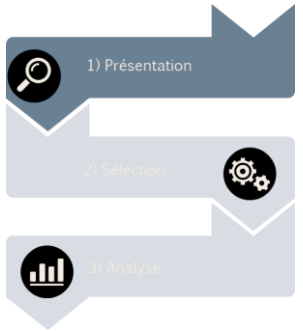
3) Présentation du jeu de données (1/2)

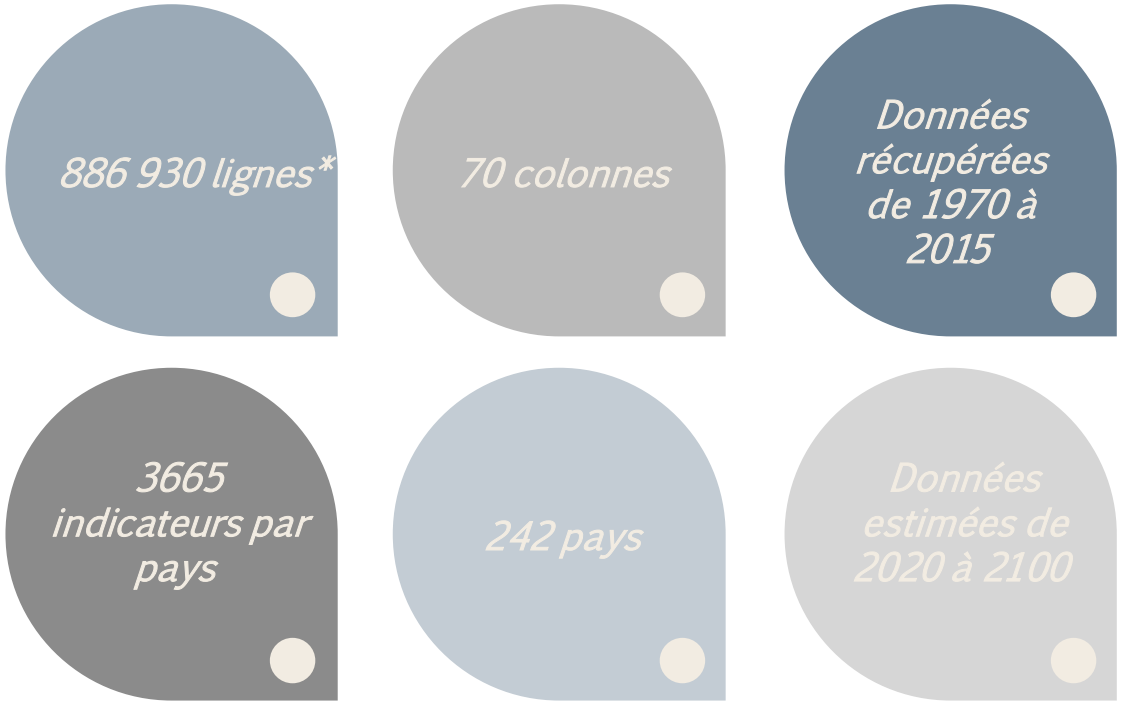


	Country Name	Country Code	Indicator Name	Indicator Code	1970	1971	1972	1973	1974	1975	...	2060	2065	2070	2075	2080	2085	2090	2095
880000	Zambia	ZMB	Barro-Lee: Percentage of population age 25+ wi...	BAR.TER.CMPT.25UP.ZS	0.39	NaN	NaN	NaN	NaN	0.27	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
880001	Zambia	ZMB	Barro-Lee: Percentage of population age 25+ wi...	BAR.TER.ICMP.25UP.ZS	0.60	NaN	NaN	NaN	NaN	0.42	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
880002	Zambia	ZMB	Barro-Lee: Percentage of population age 25-29 ...	BAR.NOED.2529.ZS	35.40	NaN	NaN	NaN	NaN	32.47	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
880003	Zambia	ZMB	Barro-Lee: Percentage of population age 25-29 ...	BAR.PRM.CMPT.2529.ZS	12.30	NaN	NaN	NaN	NaN	22.42	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
880004	Zambia	ZMB	Barro-Lee: Percentage of population age 25-29 ...	BAR.PRM.ICMP.2529.ZS	61.20	NaN	NaN	NaN	NaN	55.52	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Extrait du jeu de données

3) Présentation du jeu de données (2/2)





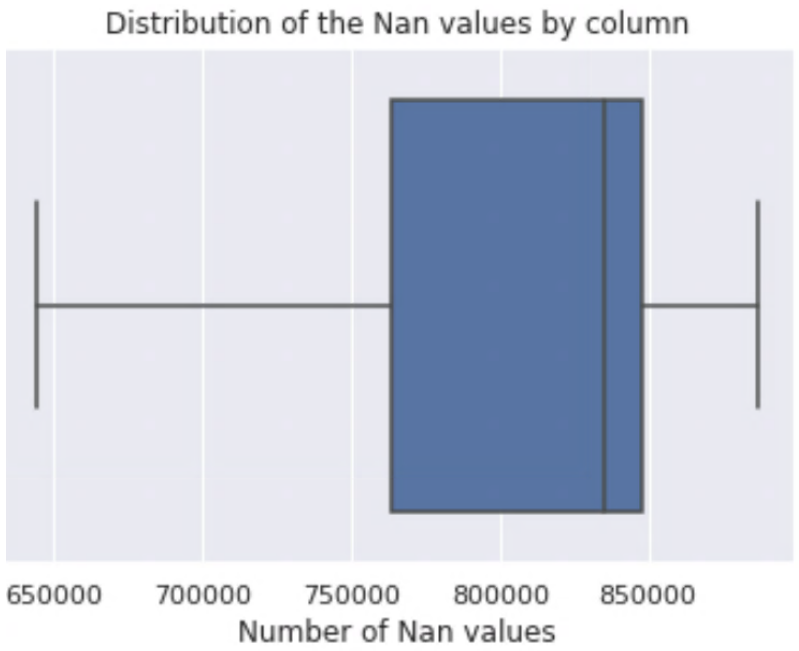
Indicateurs clés du jeu de données

* Le nombre de ligne étant important au regard de la mémoire affectée à la Virtual Machine Ubuntu, le fichier a été lu de manière itérative (chunk)

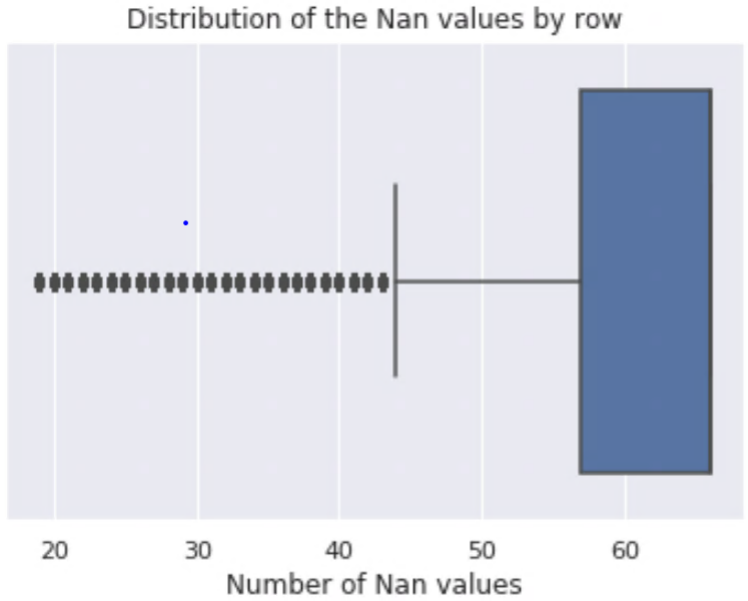
3) Présentation du jeu de données

Analyse de la complétude des données

Par colonne



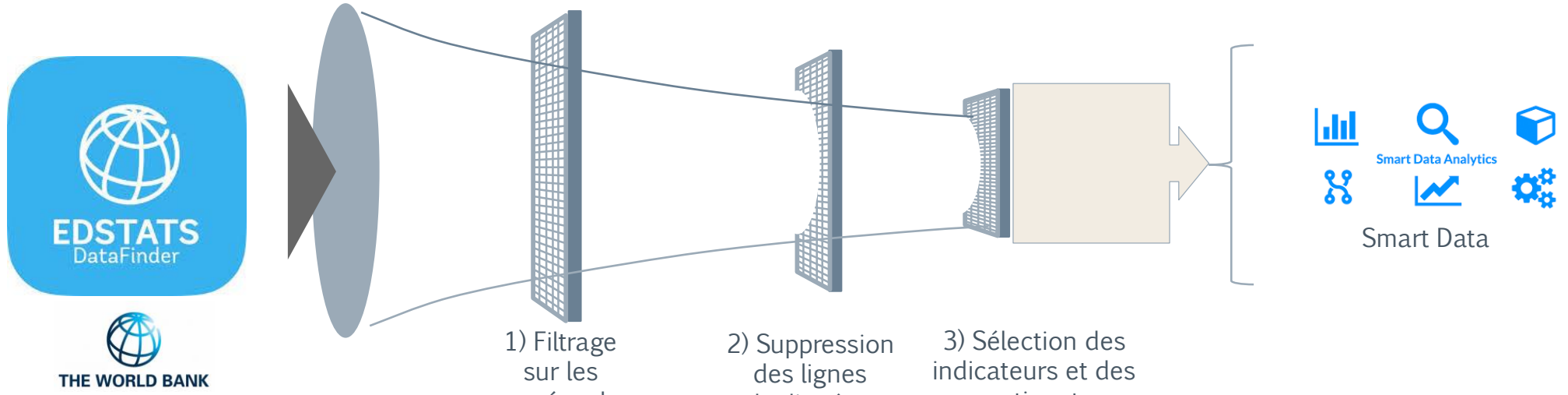
Par ligne



En moyenne,:
- 809 926 informations sur 886 930 lignes par colonne sont manquantes,
- 60 informations sur 66 par ligne sont manquantes

Il est necessaire de supprimer les lignes et les colonnes comprenant le plus de données manquantes

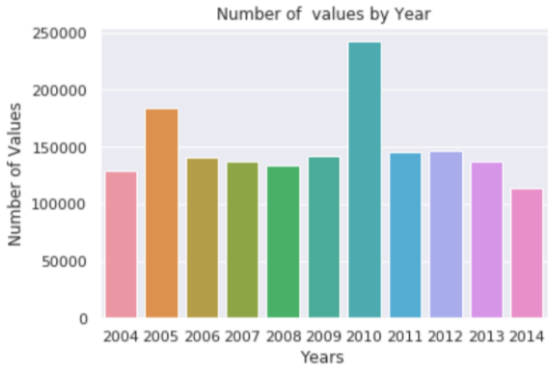
4) Sélection

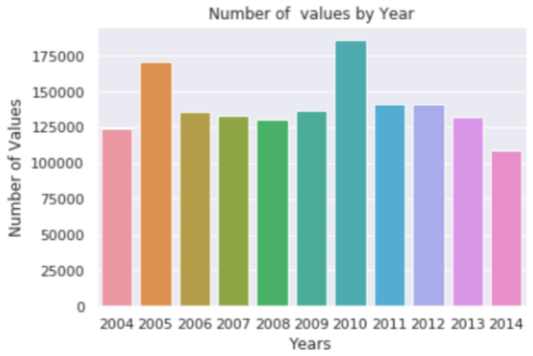


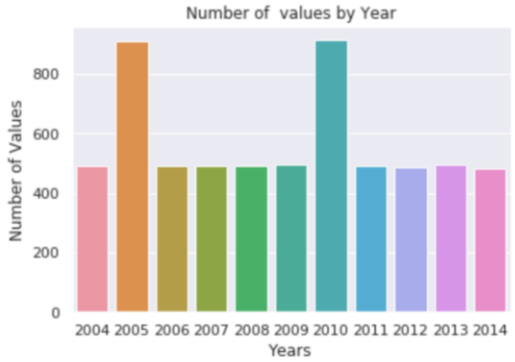
1) Filtrage sur les années de 2004 à 2014

2) Suppression des lignes dupliquées

3) Sélection des indicateurs et des pays pertinents pour l'analyse



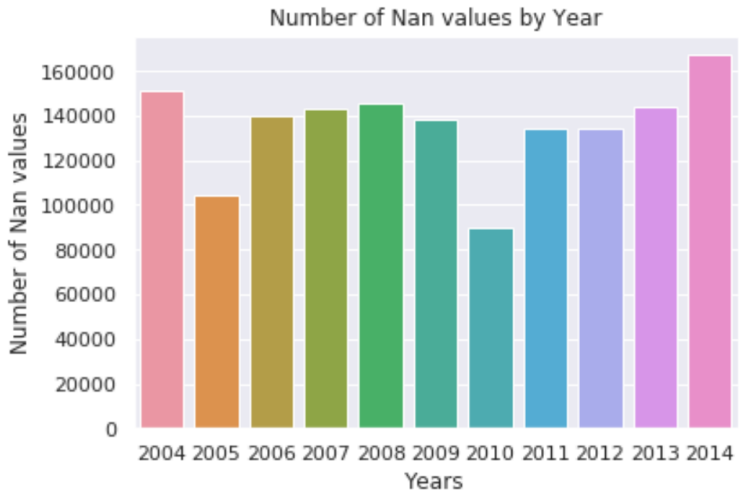




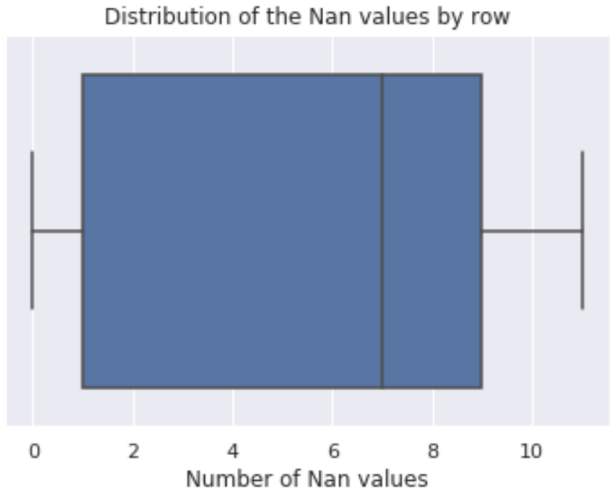
4) Sélection – Suppression des lignes dupliquées

Analyse de la complétude des données

Par colonne

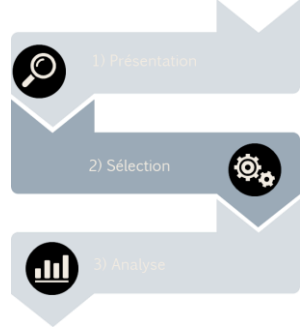


Par ligne



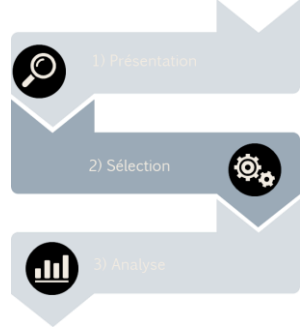
En moyenne,:
- 135 686 informations sur 275740 lignes par colonne sont manquantes,
- 5,4 informations sur 10 années sont manquantes

La qualité du jeu de données s'est améliorée mais elle n'est toujours pas suffisante pour pouvoir faire des analyses.



4) Sélection – Sélection des indicateurs (1/3)

Code de l'indicateur	Description	Raison
SP.POP.TOTL	Population totale d'un pays	Indicateur du potentiel du marché
SP.POP.1564.TO	Population des 15-64 ans	Indicateur du potentiel du marché
SP.SEC.TOTL.IN	Population en âge de faire du secondaire	Indicateur du potentiel du marché
SP.TER.TOTL.IN	Population en âge de faire du supérieur	Indicateur du potentiel du marché
BAR.SEC.CMPT.2024.ZS	Pourcentage de la population des 20-24 ans ayant validé le secondaire	Indicateur du potentiel du marché
BAR.SEC.CMPT.2529.ZS	Pourcentage de la population des 25-29 ans ayant validé le secondaire	Indicateur du potentiel du marché
BAR.SEC.CMPT.3034.ZS	Pourcentage de la population des 30-34 ans ayant validé le secondaire	Indicateur du potentiel du marché
BAR.TER.CMPT.2024.ZS	Pourcentage de la population des 20-24 ans ayant validé le supérieur	Indicateur du potentiel du marché
BAR.TER.CMPT.2529.ZS	Pourcentage de la population des 25-29 ans ayant validé le supérieur	Indicateur du potentiel du marché
BAR.SEC.CMPT.3034.ZS	Pourcentage de la population des 30-34 ans ayant validé le supérieur	Indicateur du potentiel du marché



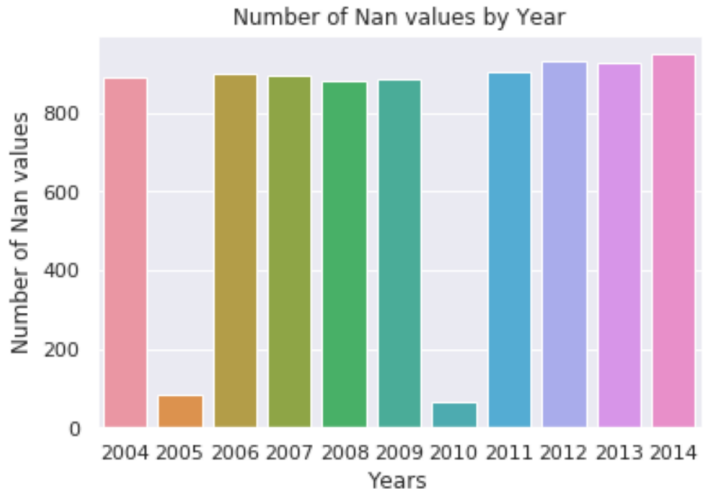
4) Sélection – Sélection des indicateurs (2/3)

Code de l'indicateur	Description	Raison
LO.EGRA.READ.ENG.ADV.2GRD	Part des étudiants ayant eu 80% en anglais au test	Possibilité de faire des supports en anglais
LO.EGRA.READ.ENG.ADV.3GRD	Part des étudiants ayant eu 80% en anglais au test	Possibilité de faire des supports en anglais
LO.EGRA.READ.ENG.ADV.4GRD	Part des étudiants ayant eu 80% en anglais au test	Possibilité de faire des supports en anglais
IT.NET.USER.P2	Pourcentage des foyers disposant d'un accès Internet	Indicateur du potentiel du marché
NY.GDP.PCAP.CD	PIB / habitant	Indicateur de richesse
NY.GDP.MKTP.CD	PIB	Indicateur de richesse
SE.XPD.TOTL.GB.ZSBAR.TER.CMPT.2024.ZS	Pourcentage de dépense à l'éducation par rapport au PIB	Indicateur de l'importance de l'éducation dans le pays

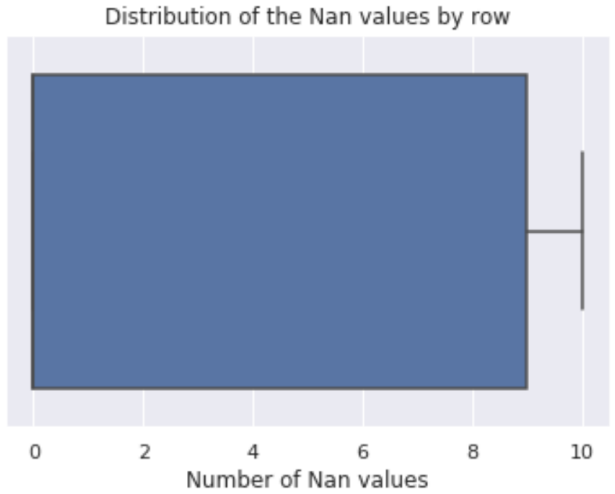
4) Sélection – Sélection des indicateurs (3/3)

Analyse de la complétude des données

Par colonne



Par ligne



En moyenne,:
- 693 informations sur 2586 lignes par colonne sont manquantes,
- 3,2 informations sur 10 années sont manquantes

La qualité du jeu de données s'est nettement améliorée, mais des ameliorations sont encore possible au niveau du nombre de pays.

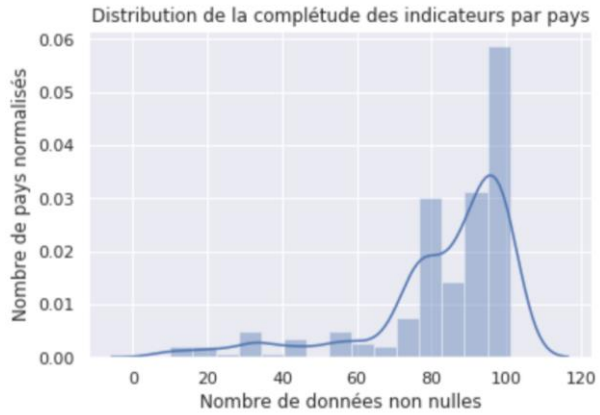
4) Sélection – Sélection des pays

Méthode

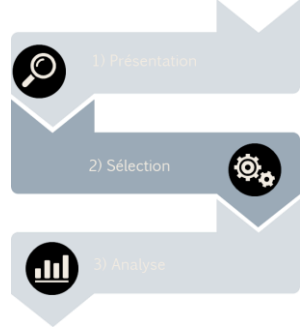
Comptabilisation par pays d'éléments Non Nuls. Les pays disposant du niveau de complétude d'information le plus important sont conservés pour la suite de l'analyse.

Résultats obtenus

Caractéristiques	Valeur
Nombre de pays	242
Moyenne du nombre de valeurs	83.181818
Ecart type	19.649628
Min de valeurs	10
25%	77
50%	90
75%	97
Max	101



Afin de limiter le nombre de pays et de se baser sur ceux disposant du plus d'information, les pays dont le nombre de valeurs renseignées est supérieur à 97 sont conservés. (71 pays)



4) Sélection – Eléments complémentaires

Corrélation entre l'indicateur population et population des 15/64

Année	Niveau de corrélation
2004	0.999602
2005	0.999637
2006	0.999672
2007	0.999700
2008	0.999726
2009	0.999753
2010	0.999782
2011	0.999802
2012	0.999820
2013	0.999835
2014	0.999843

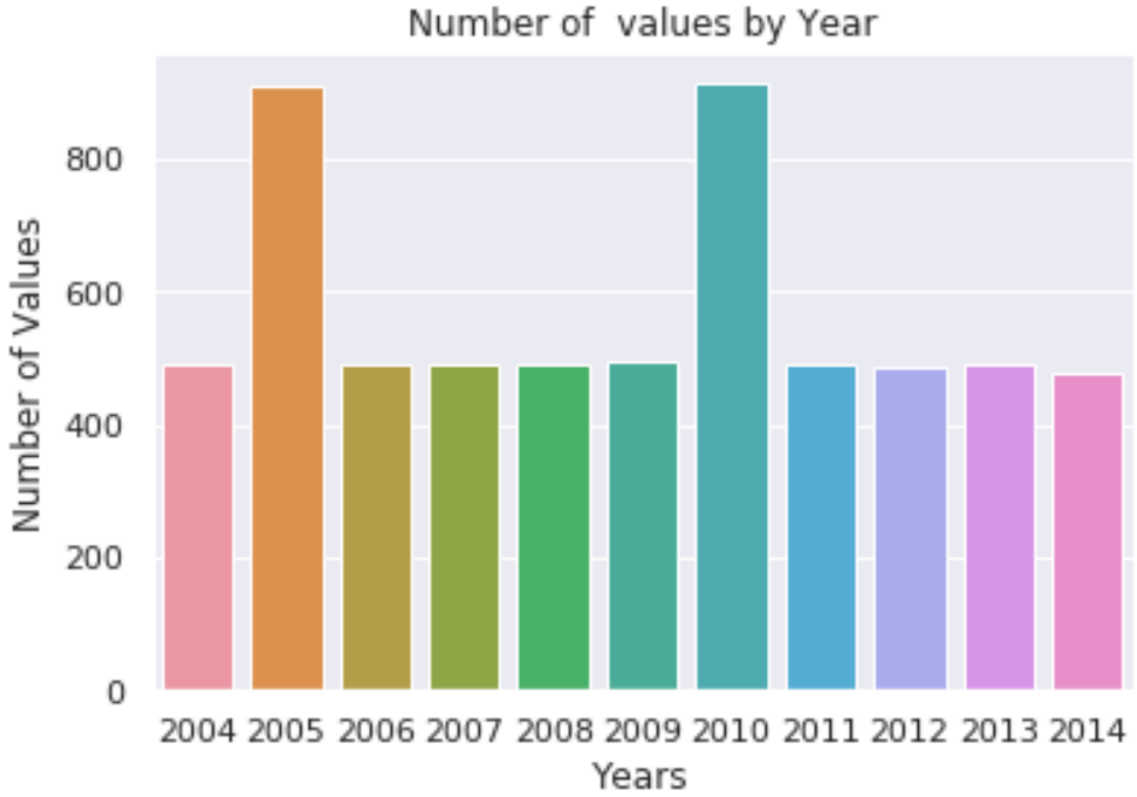
De part une forte correlation, l'indicateur de la population 15/64 est supprimé

Nombre d'occurrence par indicateur

Indicateur	Nombre d'occurence
BAR.SEC.CMPT.2024.ZS	70
BAR.SEC.CMPT.2529.ZS	71
BAR.SEC.CMPT.3034.ZS	71
BAR.TER.CMPT.2024.ZS	68
BAR.TER.CMPT.2529.ZS	71
BAR.TER.CMPT.3034.ZS	70
IT.NET.USER.P2	71
LO.EGRA.READ.ENG.ADV.2GRD	1
LO.EGRA.READ.ENG.ADV.3GRD	1
LO.EGRA.READ.ENG.ADV.4GRD	0
NY.GDP.MKTP.CD	71
NY.GDP.PCAP.CD	71
SE.XPD.TOTL.GB.ZS	71
SP.POP.TOTL	71
SP.SEC.TOTL.IN	71
SP.TER.TOTL.IN	71

De part leurs faibles occurrences, les indicateurs EGRA sont supprimés

4) Sélection – effet Barro Lee



Les deux pics de données en 2005 et 2010 sont liés aux indicateurs Barro Lee qui sont réalisés tous les 5 ans

4) Sélection – indicateurs clés

918 lignes
(vs 886 930)

17 colonnes
(vs 70)

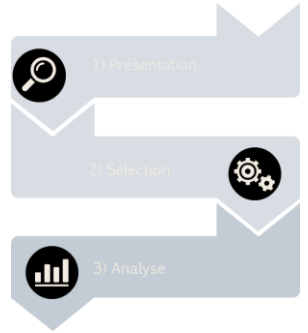
2004 à 2014
(vs 1970 à 2015)

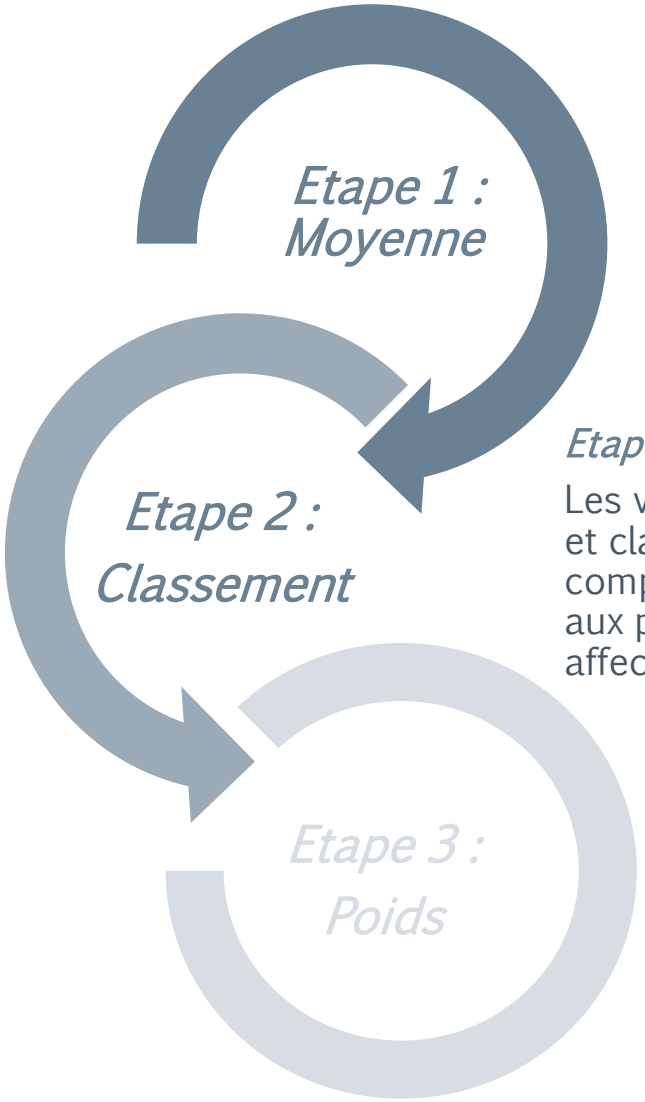
13 indicateurs par pays
(vs 3665)

71 pays
(vs 242)

Indicateurs clés de la SmartData

5) Analyse – modèle





Etape 1 : Moyenne

Une moyenne est calculée chaque indicateur de chaque pays sur l'ensemble des données entre 2004 et 2014

Etape 2 : Classement

Les valeurs de chaque indicateur sont comparées et classées parmi les différents pays. Un score compris entre 0 et 10 est affecté. 0 étant affecté aux pays ayant les plus petites valeurs et 10 étant affecté au pays ayant la plus grande valeur.

Etape 3 : Poids

Chaque indicateur est affecté d'un poids sur le score obtenu précédemment. Ensuite une moyenne pondérée est calculée afin de pouvoir obtenir un scoring par pays et ainsi pouvoir les comparer

5) Analyse – 1ère version (1/2)

1) Présentation
2) Sélection
3) Analyse

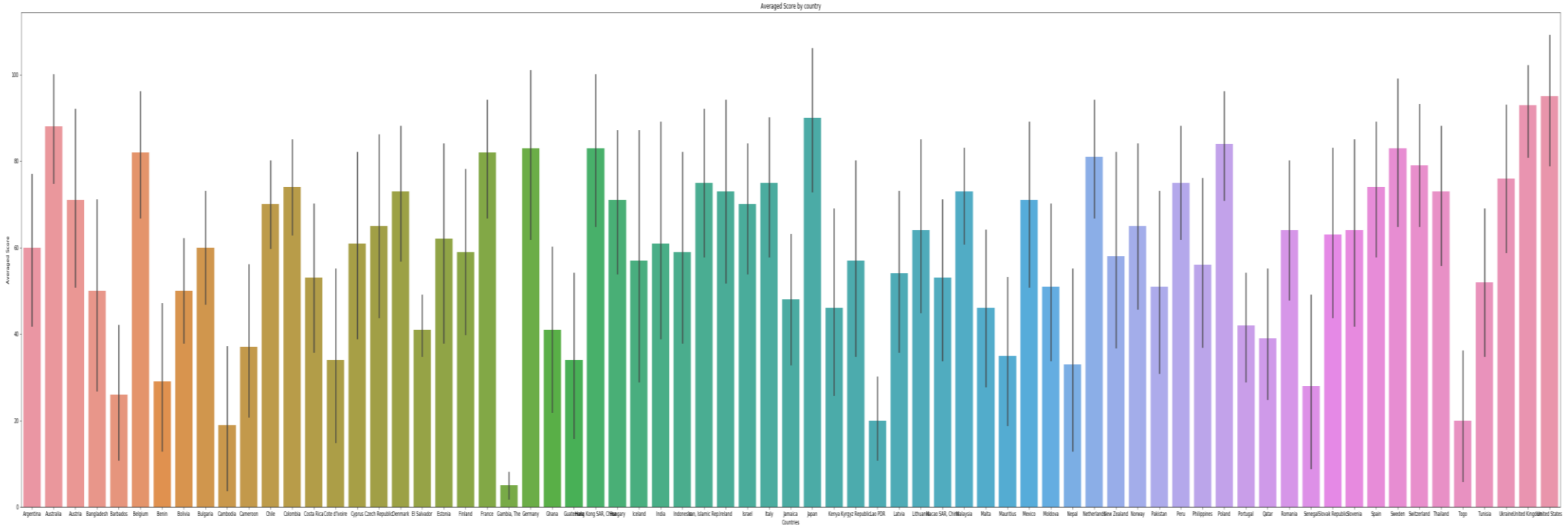
Indicateur	Poids
BAR.SEC.CMPT.2024.ZS	1
BAR.SEC.CMPT.2529.ZS	1
BAR.SEC.CMPT.3034.ZS	1

Indicateur	Poids
BAR.TER.CMPT.2024.ZS	1
BAR.TER.CMPT.2529.ZS	1
BAR.TER.CMPT.3034.ZS	1

Indicateur	Poids
IT.NET.USER.P2	1
NY.GDP.MKTP.CD	1
NY.GDP.PCAP.CD	1

Indicateur	Poids
SE.XPD.TOTL.GB.ZS	1
SP.POP.TOTL	1
SP.SEC.TOTL.IN	1

Indicateur	Poids
SP.TER.TOTL.IN	1



Le nombre de pays est trop important, les pays disposant d'un score supérieur à 50 sont conservés pour la suite de l'analyse.

5) Analyse – 1ère version (2/2)

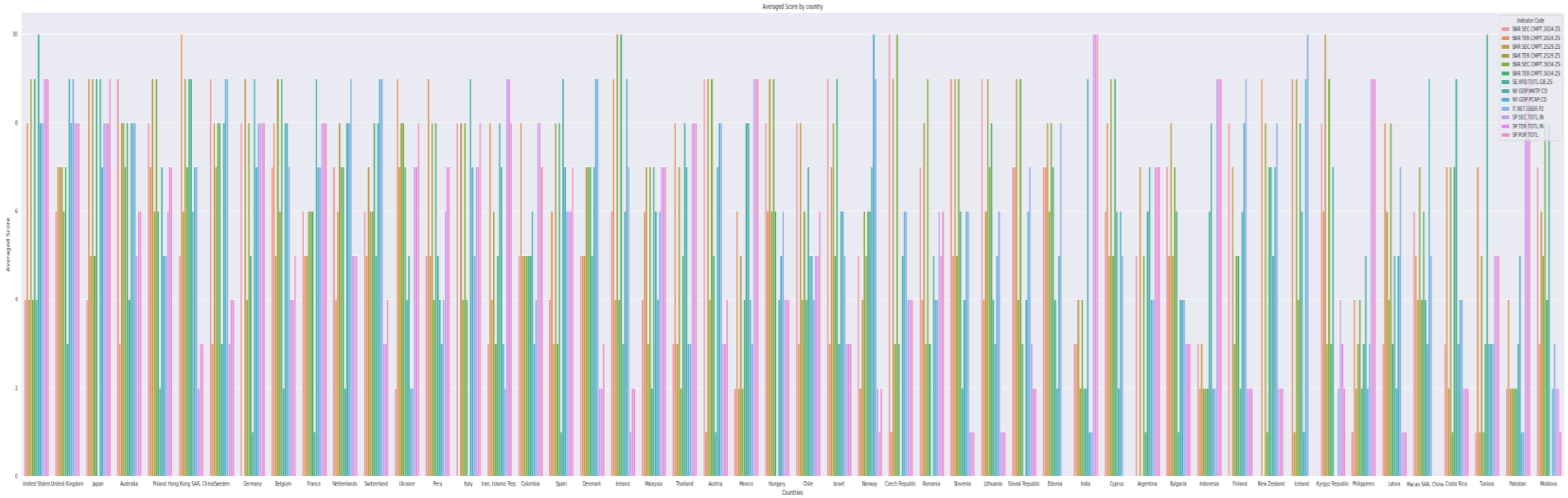
Indicateur	Poids
BAR.SEC.CMPT.2024.ZS	1
BAR.SEC.CMPT.2529.ZS	1
BAR.SEC.CMPT.3034.ZS	1

Indicateur	Poids
BAR.TER.CMPT.2024.ZS	1
BAR.TER.CMPT.2529.ZS	1
BAR.TER.CMPT.3034.ZS	1

Indicateur	Poids
IT.NET.USER.P2	1
NY.GDP.MKTP.CD	1
NY.GDP.PCAP.CD	1

Indicateur	Poids
SE.XPD.TOTL.GB.ZS	1
SP.POP.TOTL	1
SP.SEC.TOTL.IN	1

Indicateur	Poids
SP.TER.TOTL.IN	1



Le GDP et les indicateurs relatifs à la population ont des impacts significatifs sur les premiers pays. De plus, la population totale, les populations secondaire et supérieur semblent être corrélées pour certains pays. Pour ces raisons, dans la 2ème version, le GDP, les indicateurs relatifs à la populations auront un poids de 1, les autres indicateurs auront un poids de 2.

5) Analyse – 2ème version (1/2)

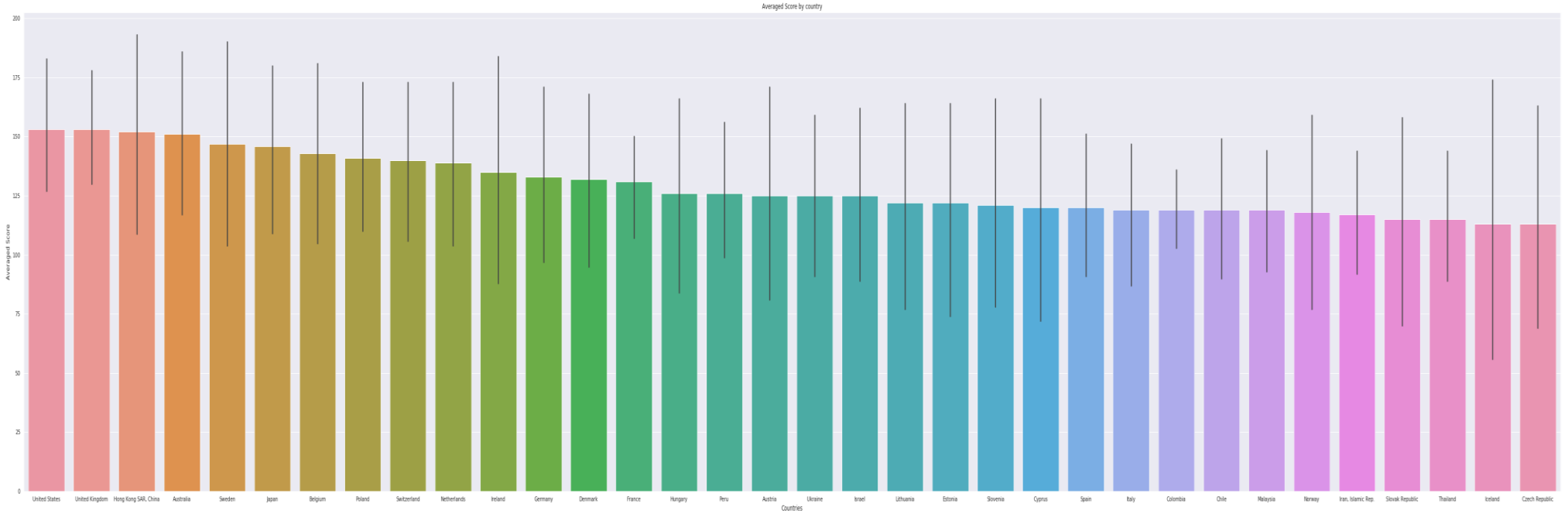
Indicateur	Poids
BAR.SEC.CMPT.2024.ZS	2
BAR.SEC.CMPT.2529.ZS	2
BAR.SEC.CMPT.3034.ZS	2

Indicateur	Poids
BAR.TER.CMPT.2024.ZS	2
BAR.TER.CMPT.2529.ZS	2
BAR.TER.CMPT.3034.ZS	2

Indicateur	Poids
IT.NET.USER.P2	2
NY.GDP.MKTP.CD	1
NY.GDP.PCAP.CD	2

Indicateur	Poids
SE.XPD.TOTL.GB.ZS	2
SP.POP.TOTL	1
SP.SEC.TOTL.IN	1

Indicateur	Poids
SP.TER.TOTL.IN	1



Le nombre de pays est trop important, les pays disposant d'un score supérieur à 110 ont été conservés pour la suite de l'analyse.

5) Analyse – 2ème version (2/2)

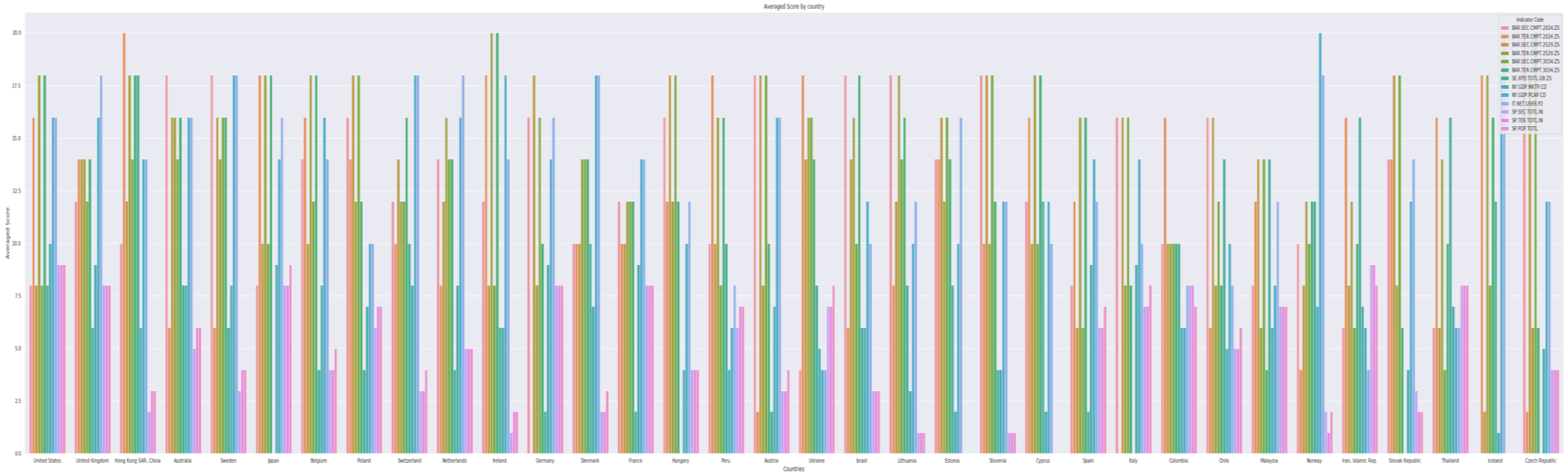
Indicateur	Poids
BAR.SEC.CMPT.2024.ZS	2
BAR.SEC.CMPT.2529.ZS	2
BAR.SEC.CMPT.3034.ZS	2

Indicateur	Poids
BAR.TER.CMPT.2024.ZS	2
BAR.TER.CMPT.2529.ZS	2
BAR.TER.CMPT.3034.ZS	2

Indicateur	Poids
IT.NET.USER.P2	2
NY.GDP.MKTP.CD	1
NY.GDP.PCAP.CD	2

Indicateur	Poids
SE.XPD.TOTL.GB.ZS	2
SP.POP.TOTL	1
SP.SEC.TOTL.IN	1

Indicateur	Poids
SP.TER.TOTL.IN	1



Les opportunités sont meilleures où le PIB/habitant, le taux de connexion internet et le niveau de dépenses dans l'education sont importants. Pour ces derniers, le poids est mis à 4.

5) Analyse – 3ème version (1/2)

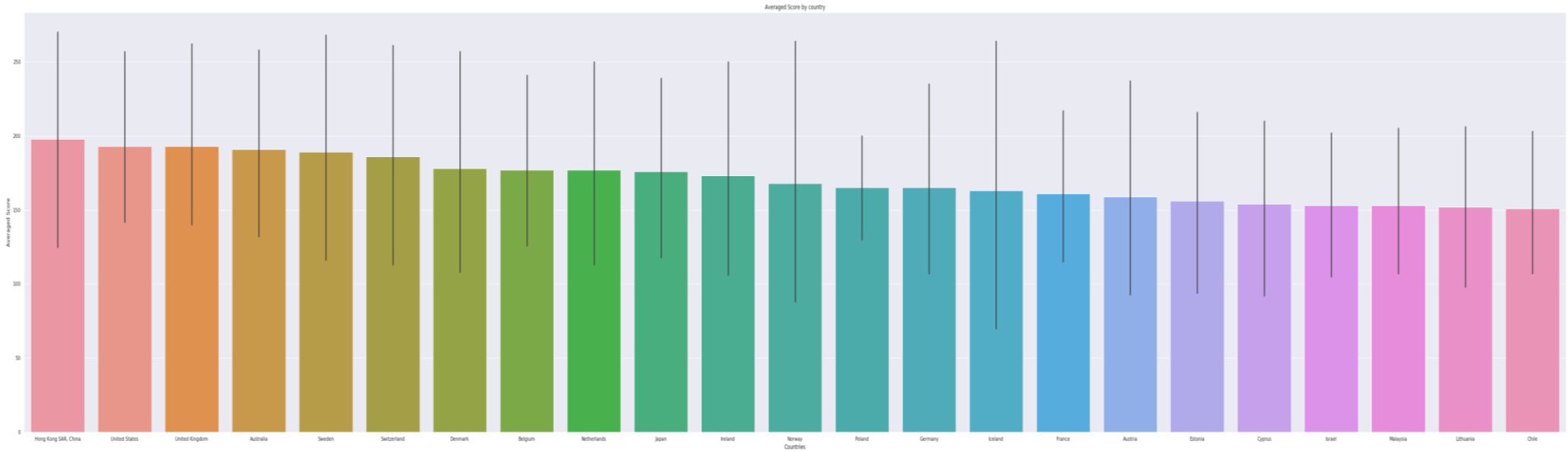
Indicateur	Poids
BAR.SEC.CMPT.2024.ZS	2
BAR.SEC.CMPT.2529.ZS	2
BAR.SEC.CMPT.3034.ZS	2

Indicateur	Poids
BAR.TER.CMPT.2024.ZS	2
BAR.TER.CMPT.2529.ZS	2
BAR.TER.CMPT.3034.ZS	2

Indicateur	Poids
IT.NET.USER.P2	4
NY.GDP.MKTP.CD	1
NY.GDP.PCAP.CD	4

Indicateur	Poids
SE.XPD.TOTL.GB.ZS	4
SP.POP.TOTL	1
SP.SEC.TOTL.IN	1

Indicateur	Poids
SP.TER.TOTL.IN	1



Le nombre de pays est trop important, les pays disposant d'un score supérieur à 150 ont été conservés pour la suite de l'analyse.

5) Analyse – 3ème version (2/2)

1) Présentation
2) Sélection
3) Analyse

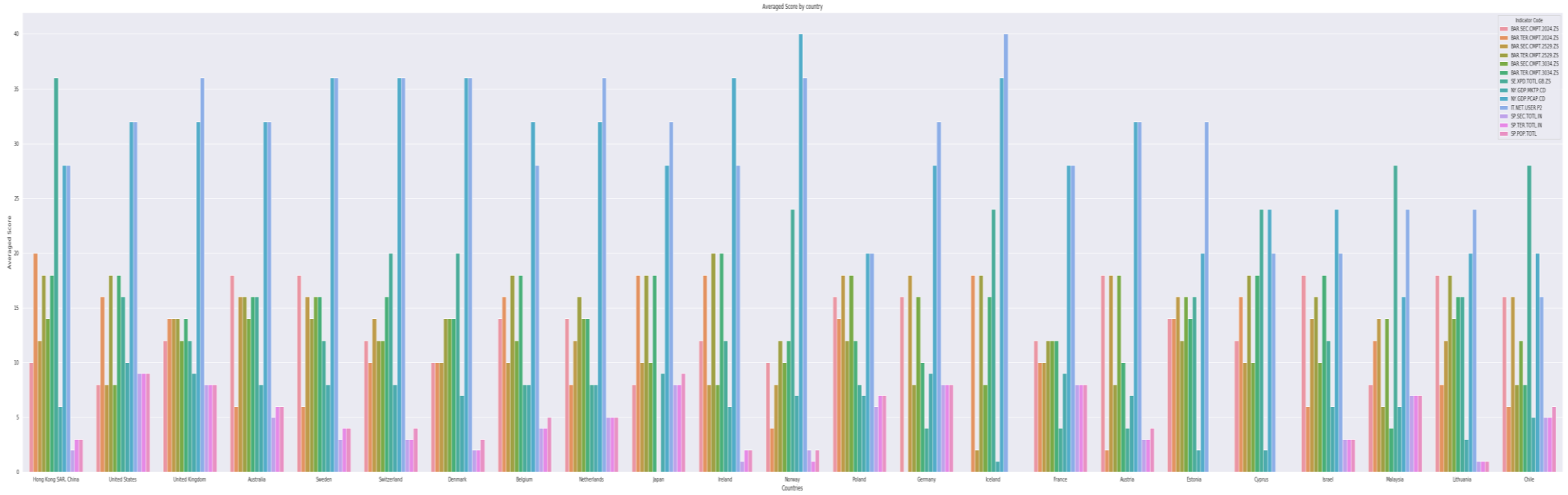
Indicateur	Poids
BAR.SEC.CMPT.2024.ZS	2
BAR.SEC.CMPT.2529.ZS	2
BAR.SEC.CMPT.3034.ZS	2

Indicateur	Poids
BAR.TER.CMPT.2024.ZS	2
BAR.TER.CMPT.2529.ZS	2
BAR.TER.CMPT.3034.ZS	2

Indicateur	Poids
IT.NET.USER.P2	4
NY.GDP.MKTP.CD	1
NY.GDP.PCAP.CD	4

Indicateur	Poids
SE.XPD.TOTL.GB.ZS	4
SP.POP.TOTL	1
SP.SEC.TOTL.IN	1

Indicateur	Poids
SP.TER.TOTL.IN	1



Les opportunitées sont meilleures dans les pays où la population dispose d'un niveau supérieur. Ces niveaux passent à un poids de 3.

5) Analyse – 4ème version (1/2)

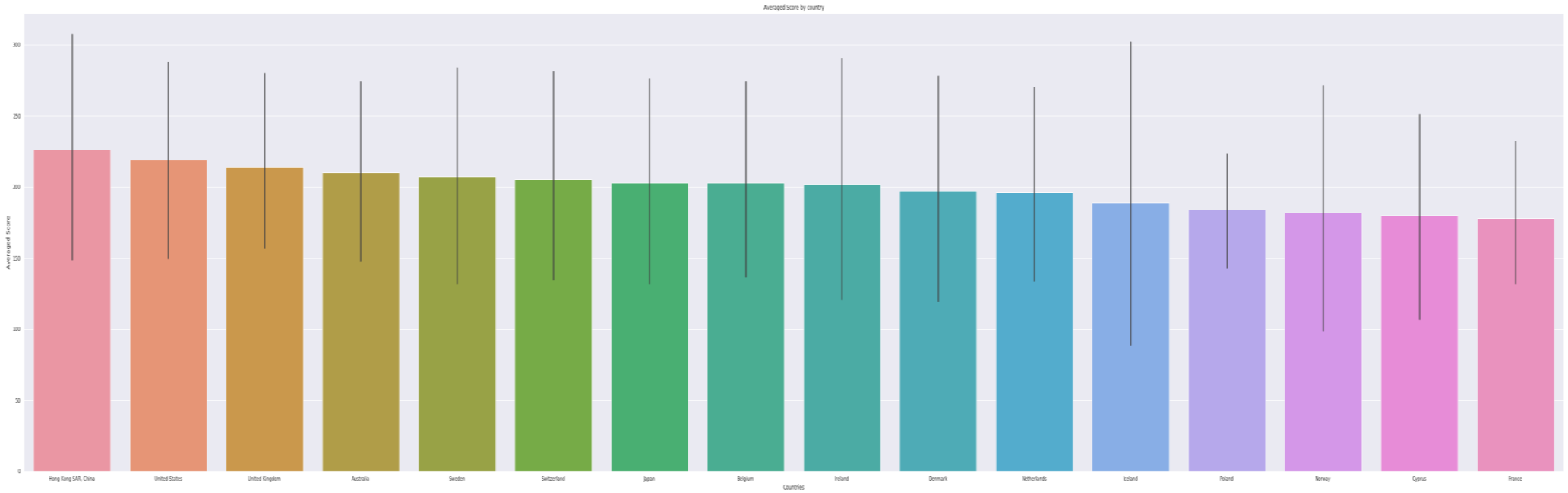
Indicateur	Poids
BAR.SEC.CMPT.2024.ZS	2
BAR.SEC.CMPT.2529.ZS	2
BAR.SEC.CMPT.3034.ZS	2

Indicateur	Poids
BAR.TER.CMPT.2024.ZS	3
BAR.TER.CMPT.2529.ZS	3
BAR.TER.CMPT.3034.ZS	3

Indicateur	Poids
IT.NET.USER.P2	4
NY.GDP.MKTP.CD	1
NY.GDP.PCAP.CD	4

Indicateur	Poids
SE.XPD.TOTL.GB.ZS	4
SP.POP.TOTL	1
SP.SEC.TOTL.IN	1

Indicateur	Poids
SP.TER.TOTL.IN	1



Le nombre de pays est trop important, les pays disposant d'un score supérieur à 178 ont été conservés.

5) Analyse – 4ème version (2/2)

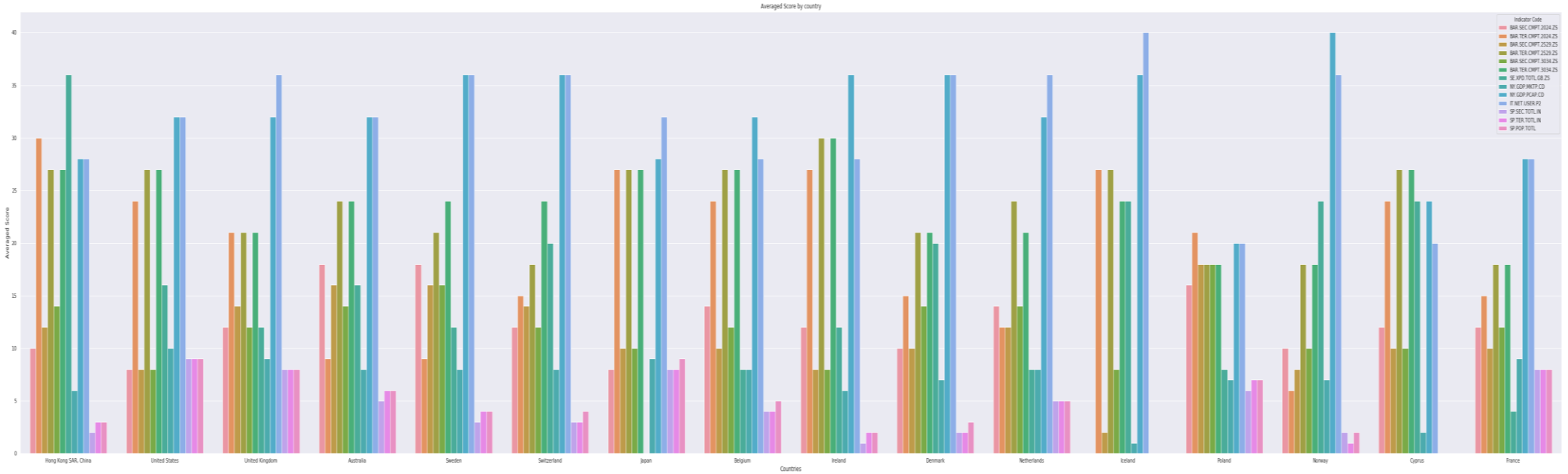
Indicateur	Poids
BAR.SEC.CMPT.2024.ZS	2
BAR.SEC.CMPT.2529.ZS	2
BAR.SEC.CMPT.3034.ZS	2

Indicateur	Poids
BAR.TER.CMPT.2024.ZS	3
BAR.TER.CMPT.2529.ZS	3
BAR.TER.CMPT.3034.ZS	3

Indicateur	Poids
IT.NET.USER.P2	4
NY.GDP.MKTP.CD	1
NY.GDP.PCAP.CD	4

Indicateur	Poids
SE.XPD.TOTL.GB.ZS	4
SP.POP.TOTL	1
SP.SEC.TOTL.IN	1

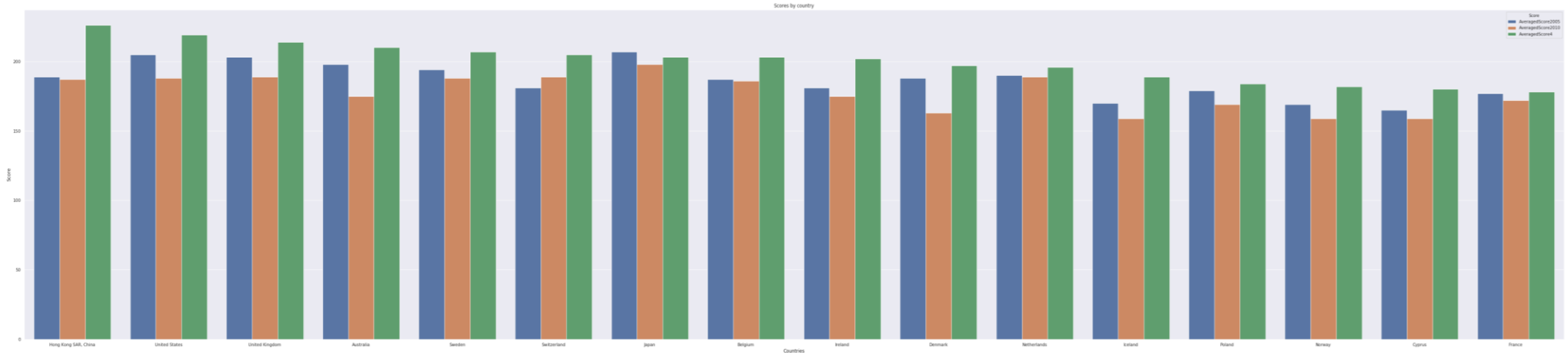
Indicateur	Poids
SP.TER.TOTL.IN	1



Les 16 pays ci-dessous ressortent de l'analyse.

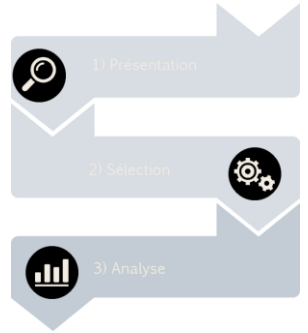
5) Analyse – Possibilité d'evolution du marché

Méthode : le processus de scoring est réitéré pour les années 2005 et 2010 (ces années ont été sélectionnées car elles disposent de plus de données : voir effet Barro Lee) . Ces résultats peuvent ainsi être comparés aux résultats obtenus précédemment.



En 2010, tous les scores ont baissé, ceci peut s'expliquer lors l'étape de classement, les pays sélectionnés ont été moins bien positionnés par rapport aux autres pays. Les résultats de 2005 et 2010 sont inférieurs à ceux obtenus précedemment pour l'ensemble des pays. Il y a donc une possibilité de croissance pour l'ensemble de ces pays.

5) Analyse – Résultats



1

Quels sont les pays avec un fort potential de clients pour nos services ?

Les pays sont :

Hong Kong	Etats-Unis	Royaume-Uni	Australie	Suède	Suisse
Japon	Belgique	Irlande	Danemark	Pays-Bas	Islande
Pologne	Norvège	Chypre	France		

2

Pour chacun de ces pays, quelle sera l'evolution de ce potentiel de clients ?

L'ensemble de ces pays dispose d'un potentiel de croissance.

3

Dans quels pays l'entreprise doit-elle opérer en priorité ?

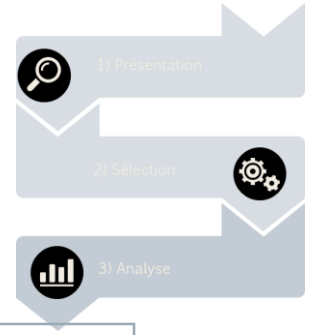
Les pays à opérer en priorité sont :

Etats-Unis	Royaume-Uni	Australie	Suède	Suisse

Les raisons sont :

- PIB / Habitant important
- Taux de population ayant un niveau d'étude supérieur
- Taille de la population
- Maîtrise de l'anglais

6) Prochaines étapes



Amélioration du modèle

- Enrichir le jeu de données avec des résultats à des examens internationaux en anglais (EF English Proficiency Index).
- Estimer les résultats des indicateurs Barro Lee sur les 4 dernières années.
- Intégrer des moyennes pondérées pour les années en affectant un poids plus important aux données récentes.

Etudes de marché

Des études de marché sont à mener dans les 5 pays recommandés afin d'avoir une vue sur :

- Les besoins des clients et du marché en terme de contenu
- Un état des lieux de la concurrence
- Le réglementaire

7) Environnement technique